

7 Attach the hot wire to the breaker screw terminal.

Circuit breakers come in various configurations. In the project shown here, a 20-amp twin breaker is being installed to conserve space. (A *twin* is two slim single-pole breakers in a standard single-pole breaker space.) A typical single-pole breaker is installed in exactly the same way. **NOTE:** Twin breakers can *not* be used to supply 240 volts.

When adding a circuit to an existing panel, cut the power to the panel and test to be sure it's off. Remove the cover, then use needle-nose pliers to jab out a knockout ❶. Install a cable connector into the knockout, pull the cable through the connector, and tighten the connector to prevent strain on electrical connections ❷. Cut enough extra cable to reach the farthest point inside the panel. Staple cable within 12 in. of the panel.

Score the cable sheathing by making two diagonal passes and sliding the sheathing off. Or score the sheathing lightly down the middle ❸, peel back the sheathing, and use diagonal cutters to snip it off. Leave about ½ in. of sheathing sticking into the panel.

Separate the ground, neutral, and hot wires. Connect the ground wire first ❹. In a service panel, the ground will attach to a ground/neutral bus; in a subpanel, the ground will attach to a separate ground bar. Insert only one ground wire beneath each lug screw. Next strip ½ in. of insulation from the neutral and connect it to the neutral bus ❺. Making right-angle turns in individual neutrals toward the bus creates a neat, orderly arrays of wires.

Snap the breaker onto the hot bus ❻ and press down to seat the breaker securely on the hot bus pole(s). Strip ½ in. of insulation from the hot wire, then connect it to the lug screw on the end of the breaker ❼. *Note:* Before attaching the hot wire, slide a circuit tag onto the hot wire to identify the location it feeds.

After connections in the panel are complete, make sure all splices, switches, receptacles, fixtures, and/or appliances on the new circuit are installed and there are no open wires or shorts in the circuit. Replace the cover, turn on the breaker, and test the devices for proper function.

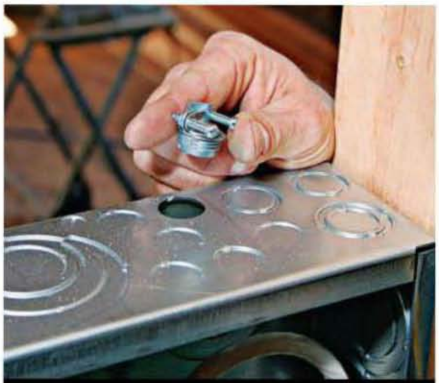

2 Insert the threaded shaft of the cable connector (clamp) into the knockout.

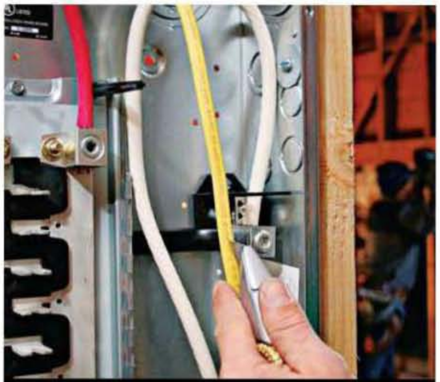

3 Score Romex cable sheathing lightly to avoid damaging the wire insulation.

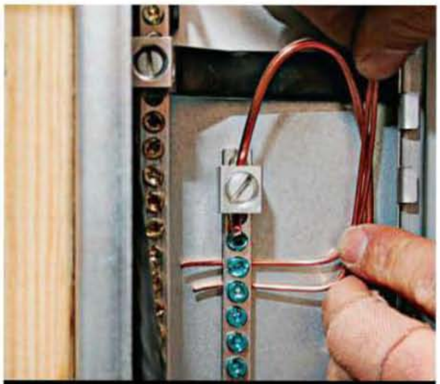

4 Connect the ground wire to the ground bus of the subpanel.

GFCI VS. AFCI BREAKERS: WHAT'S THE DIFFERENCE?

Installing GFCI and AFCI breakers is essentially the same, but these two Code-required breakers offer protection from different hazards.

Ground fault circuit interrupter (GFCI) breakers appeared first. A ground fault is any failure of the electrical system that leaks current from a hot wire. GFCIs are highly sensitive devices that can detect minuscule (5 milliamp) current leaks and shut off power almost instantaneously—typically, within 1/40 second. A GFCI's primary function is to protect people from electrical shocks, especially when there is moisture present. The NEC requires GFCI protection on all bathroom receptacles; kitchen receptacles within 4 ft. of a sink; all receptacles serving kitchen counters; outdoor receptacles; accessible basement or garage receptacles; and receptacles near pools, hot tubs, and so on.

Arc-fault circuit interrupter (AFCI) breakers, on the other hand, help prevent house fires. AFCIs detect minute fluctuations in current associated with arcing and de-energize the circuits before a fire can start. Arcing most commonly occurs when a nail or screw punctures a cable, when insulation on an extension cord becomes frayed, or connections become loose at an outlet or a switch. The NEC requires AFCI protection on all 120v bedroom circuits.

DOUBLE-POLE CIRCUIT BREAKERS

The back of a double-pole breaker has two pressure clips that snap onto two hot bus poles—thus providing 240v of power. With double-pole breakers, both red and black hot wires are attached to the breaker. Energy will flow whether you connect a red or black hot wire to either terminal, but electricians routinely alternate colored wires: black, red, black, red.

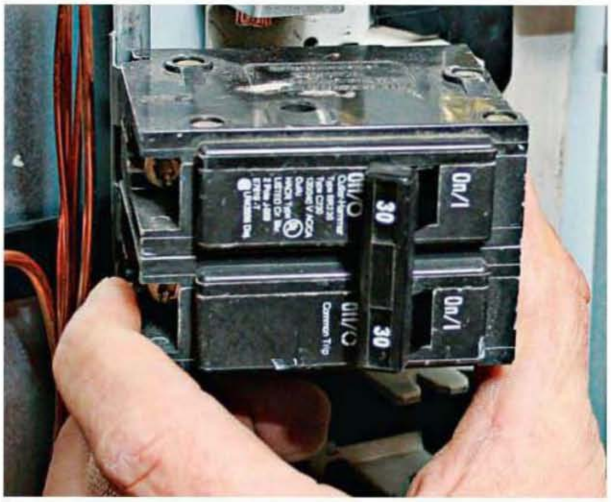

INSTALLING A SUBPANEL

There are many reasons to install a subpanel:

- Increasingly, main service panels are installed outside so firefighters can disconnect power before going into the house—in this case, a subpanel inside contains all the branch circuits.
- When a system has unused capacity but the service panel has no available slots for more circuit breakers, a subpanel allows for expansion.
- A subpanel distributes power to a separate building.
- A subpanel can offset voltage drops on circuits that are too distant (70 ft. or 80 ft.) from the service panel. In this case, the larger gauge wires that feed subpanels suffer less voltage drop than the smaller gauge wires of branch circuits.

Size subpanels based on anticipated loads. If you're adding a subpanel in the same building as the main panel so you can add lighting and general-use circuits, install a 60-amp subpanel with at least 12 breaker slots. If the subpanel is distributing power to a distant building that you want to use as a workshop or office, or perhaps expand later, install a 100-amp or 150-amp subpanel.

People frequently add subpanels when planning a major kitchen remodel, because kitchens have a lot of appliances. Locate the subpanel as close as possible to the kitchen.

When adding a subpanel to an existing system, an electrician first shuts off power at the service panel and tests to be sure the power is off.

If studs are spaced 16 in. on-center, install a standard, 14½-in.-wide panel. Use four 1¼ in. by 10 screws to mount the panel ❶; screws should embed at least 1 in. into studs. Install the panel at a comfortable height, so it will be easy to wire and access. Remove a concentric knockout in the panel and install a connector appropriate to the cable or conduit that feeds the panel. Also install a protective cable bushing. Here, the feed is a #2 Romex cable. Pull 3 ft. of cable into the panel, tighten the connector, and strip the cable sheathing. (Note the cross-brace between the studs, just below the panel.) ❷

Code requires that you secure the cable within 12 in. of the panel, but 2/0 cable is stiff. Strapping cable to a stud would force it to enter the panel at a sharp angle, which could stress the knockout. It is far easier to strap the cable to a brace in the middle of the stud bay and run the cable straight into the panel.

Secure the ground wire to the main lug of the subpanel ground bus ❸. Cut the three insulated feed wires, allowing enough extra to loop them gently. Avoid sharp bends, which can damage wire. Strip 1 in. of insulation from the feed ends. Use an Allen wrench or torque wrench to connect the feed wires to their respective lugs ❹. Loop the hot feeds around the perimeter of the panel, distancing them from the neutral feed so there will be a open area through which to run smaller circuit wires ❺. Looping the wire generously also ensures that there will be enough extra cable in case you need to strip and reconnect feed ends later ❻.

ACCORDING TO CODE

The NEC requires a main breaker in any subpanel in a separate building.

LOCATING A SUBPANEL

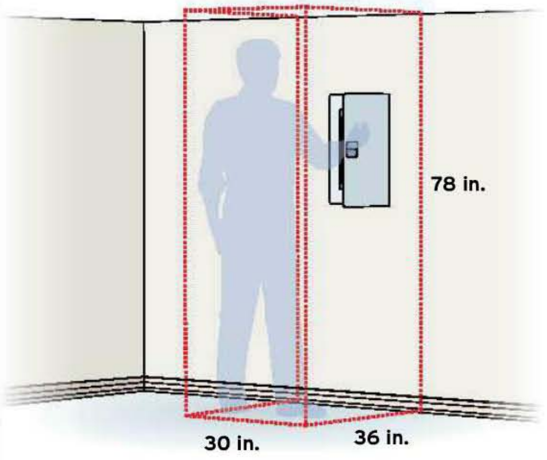
78 in.
30 in.
36 in.

Any panel—whether a main or a subpanel—must have sufficient clearance around it. NEC minimums require 78 in. of headroom, 36 in. free space in front of the panel, and 30 in. across the face of the panel. The panel should be installed at a comfortable height, meaning that no breaker handle may be higher than 72 in.

The area must also be dry and easily accessed. Do not install a subpanel near a bathroom or near flammable materials. If you install a subpanel in a common area between living spaces and an attached garage, wrap the panel enclosure with two layers of ⅝ in. drywall on all sides. However, if you place the panel in a dry interior wall you can install a standard two-lug panel, whose 14½ in. width fits neatly between two studs spaced 16 in. on-center.

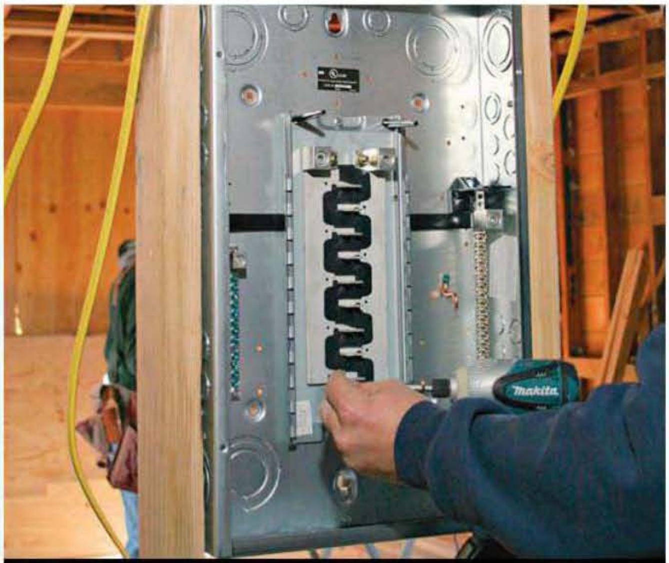

1 When possible, install a subpanel between studs spaced 16 in. on-center. Secure it in place using at least four screws.

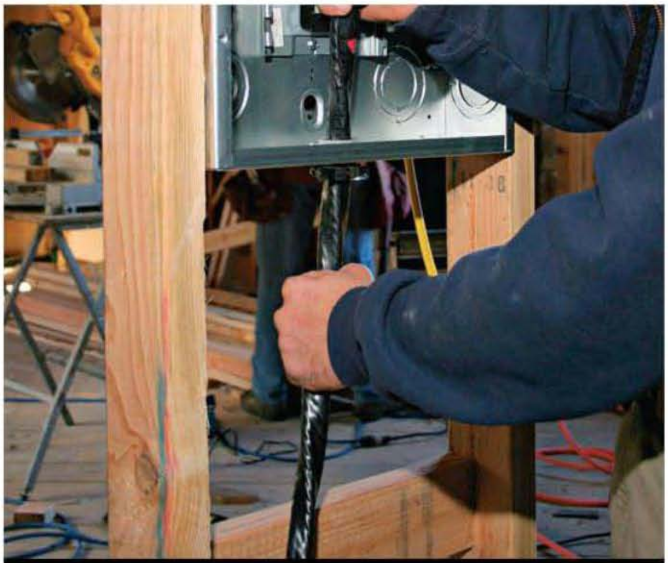

2 Remove a knockout, install a cable connector, and then pull the feeder cable into the subpanel. Strap the cable to a cross-brace.

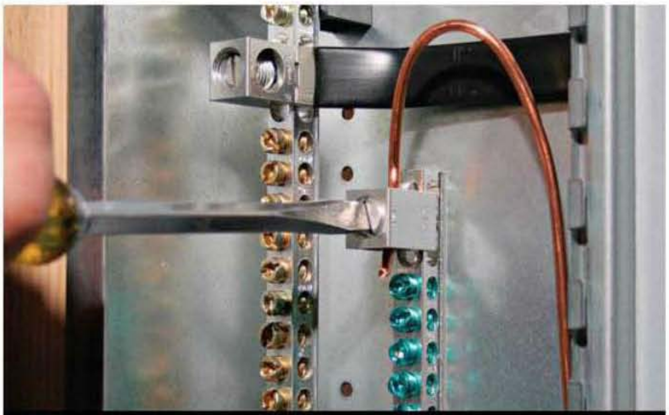

3 First connect the ground wire from the feeder cable to the lug on the ground bus. The ground bus is bonded to the metal panel.

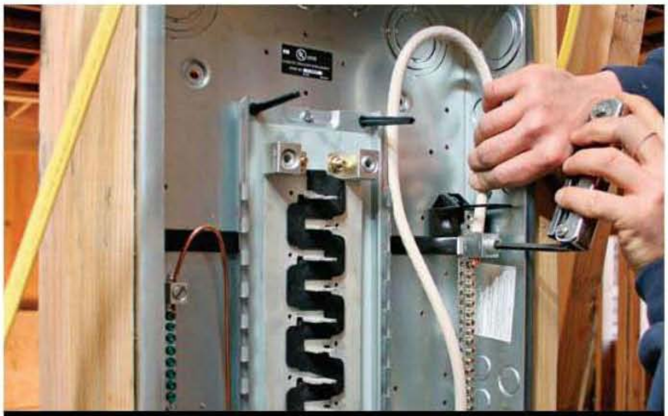

4 Connect the neutral wire to the neutral bus lug. Plastic mounting brackets isolate—and insulate—neutral buses from the panel.

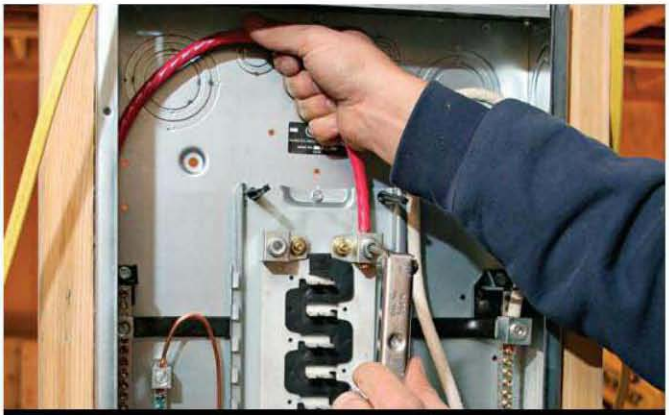

5 Run the two hot feeds around the other side of the panel. Strip the hot feed ends and attach them to hot bus lugs.

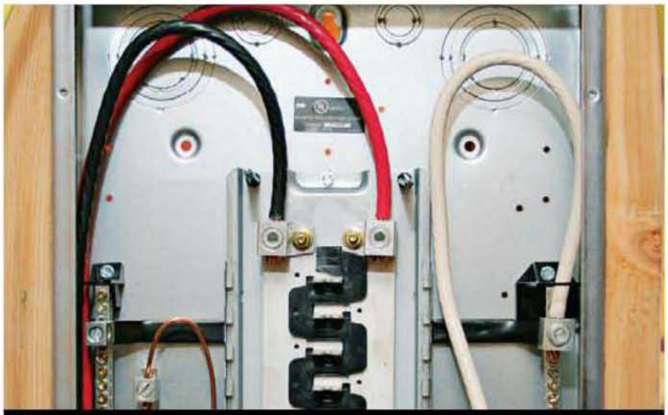

6 In the completed subpanel, all feeds terminate into lugs.

GLOSSARY

AC Armored cable. Also, Alternating current.

ACCESSIBLE Not permanently concealed by the structure or finishes. Able to be accessed without damaging the building.

AMPACITY (AMPS) The amount of current a circuit or conductor can safely carry (conduct). Measured in amperes (amps).

AMPS The measure of the volume of electrons flowing through a system (current).

ARC FAULT CIRCUIT INTERRUPTER (AFCI) A circuit breaker designed to de-energize a circuit within a period of time if it senses arcing. An arc is a spark between conductors or connections.

BONDING JUMPER (MAIN) Conductor connecting the neutral conductor to the grounding electrode conductor in a service panel.

CABLE An assembly of several conductors. Most often refers to plastic- or metal-sheathed cable that contains several individual wires.

CIRCUIT (BRANCH) Circuit originating at a circuit breaker or fuse in an electrical panel and feeding utilization equipment (lights, switches, receptacles, appliances).

CIRCUIT BREAKER A device intended to de-energize a circuit if current exceeds specified parameters.

CONDULET A conduit body with a cover designed to act as a pulling point in a conduit run, to change conduit direction, or as an intersecting point of multiple conduit runs.

CONDUCTOR Technically, anything that conducts electricity. Most often, it denotes a wire.

CURRENT The flow of electrons in a system. Current is measured in amperes (amps). There are two types of current: DC (direct current) and AC (alternating current). Typically AC is found in homes and buildings.

ENERGIZED Live. The presence of voltage in a circuit or conductor.

FEEDER A large ampacity circuit supplying a service panel, subpanel, or other high current piece of equipment.

FIXTURE A light fixture (table lamp, sconce, chandelier, recessed can, etc.)

FUSE BOX A metal box designed to house fuses installed for circuit protection.

GROUND A connection to earth (the ground). To make such a connection.

GROUND BUSS A piece of metal designed to connect multiple wires to a grounding electrode conductor.

GROUNDING CONDUCTOR A wire in an electrical system designed to bond metal parts of the electrical system to the earth (ground).

GROUNDING ELECTRODE CONDUCTOR A wire of a gauge determined by the ampacity of a service panel that connects the service panel (and ground bus) to a grounding electrode (ground rod or equivalent).

GROUND FAULT A fault situation in which an energized conductor or piece of equipment comes in contact with grounded metal parts.

GROUND FAULT CIRCUIT INTERRUPTER (GFCI) A special receptacle or circuit breaker intended to protect people by de-energizing a circuit within a specific period of time if current to ground exceeds specified parameters.

GROUND SCREW A green colored screw used to connect ground conductors to boxes and devices.

LISTED Equipment and materials included on a list published by an organization, acceptable to the authority having jurisdiction, that states that the equipment or materials meet specific design criteria or are suitable for the use intended. Most commonly UL (Underwriter's Laboratory) listed.

LOCATION, DAMP Protected from weather and not subject to saturation (under eves, canopies, etc.)

LOCATION, DRY Not subject to moisture (interior, protected from weather).

LOCATION, WET Underground; exposed to weather; subject to saturation.

MAIN CIRCUIT BREAKER A large ampacity circuit breaker that protects and acts as a means to de-energize a service panel or subpanel.

MC Metal clad cable.

NEUTRAL The neutral or grounded conductor (not ground conductor) is the return path for current in an electrical system; designated by a white or light gray color. The neutral is bonded to ground at the service panel.

NEUTRAL BUS A piece of metal designed to connect multiple wires to the neutral conductor.

OHMS The measure of resistance to the flow of electrons (current).

OVERLOAD A situation in which the current flowing through a circuit or circuit conductor exceeds the safe operating ampacity of the conductor.

RECEPTACLE Device designed for the connection of a plug.

ROMEX® Type of NM cable, sheathed with nonmetallic material.

SERVICE PANEL (MAIN PANEL) The primary electrical panel located immediately after the meter socket and housing the main circuit breaker or main fuses and/or other devices designed to de-energize and protect circuits.

SERVICE DROP Conductors from the utility pole to the point of connection (weather head).

SERVICE-ENTRANCE CONDUCTORS (OVERHEAD) Conductors between the point of connection (meter lugs) and a point outside (weather head) connected to the service drop.

SERVICE-ENTRANCE CONDUCTORS (UNDERGROUND) The conductors between the point of connection (meter socket) and the utility service lateral.

SCREW SHELL The metal conductive interior body of a light socket or fuse holder with large threads (or ridges) that allow a bulb or fuse to be screwed in.

SPLICES The point at which two wires or conductors are joined or connected together. For branch circuits, splices are commonly made with pressure connectors (wire nuts).

SUBPANEL An electrical panel that is supplied from a service panel and is installed either to provide additional circuit breakers or to distribute branch circuits at a distance from the service panel.

SWITCH A device that opens and closes a circuit, controlling the operation of a light fixture, fan, other equipment, or appliances.

VOLTAGE, NOMINAL (VOLTS) A value assigned to a circuit or system to designate its class. Modern single family residences in the U.S. typically have 120/240V systems.

VOLTAGE TO GROUND (VOLTS) The difference in potential (voltage) between a single energized conductor and ground (neutral or equipment ground).

WATERTIGHT Enclosures built to prevent the intrusion of moisture.

WATTAGE (WATTS) A measure of Power consumed.

WEATHERPROOF Equipment or enclosures built so that weather will not interfere with their operation.

INDEX

A

AFCIs (arc-fault circuit interrupters), 164, 243
Aluminum wiring, 14, 30, 160
Amps, volts, and watts, 6, 161
Analyzers, 21, 24
Appliances, 210–21
  baseboard heaters, 220
  basics of, 212–13
  common requirements, 212
  dishwashers, 215
  garbage disposers, 214–15
  laundry setups, 218–20
  oven outlets/connections, 216–18
  small, circuits, 162, 165
  unfinished cords (preparing), 213
  water heaters, 221
Arcing, 14, 15

B

Back-wired devices, 40, 42
Baseboard heaters, 220
Bathroom circuits/lighting, 162, 165
Bathroom fans, 141–47
Boxes
  capacity of, 25
  ceiling, 85, 99, 187, 195
  for ceiling fans, 152–53
  choosing, 25–26
  code requirements, 163, 183
  connectors, clamps and clips, 32
  covering all connections, 54, 163
  cut-in, 26
  cutting holes for, 194
  fill worksheet, 25
  grounding, 163
  installing, 27, 85, 99, 152–53, 186–87, 194–95
  locating, 27, 47, 184, 185, 186–87
  metal vs. plastic, 25–26
  mounting between joists, 29, 182
  removing knockouts, 28
  retrofitting, 195, 197
  surface, metal, 208-09
  symbols for, 168
  wall, 27, 186
Building codes, 163–64
  boxes, 163, 183
  cable, 163, 183
  checking, 3, 4, 165, 224
  circuit breakers, 183
  circuit wiring, 163
  common requirements, 163
  general-use circuit requirements, 164
  grounding, 163, 183
  homeowner work allowed?, 165
  outdoor wiring, 224
  receptacles, 27, 164
  rough-in, 183
  wire connections, 163, 183

C

Cable rippers, 21, 22, 35
Cables. See also Multimedia
  AC, installing, 203
  armored, stripping, 22
  code requirements, 163, 183
  connectors, 32, 182, 202
  corner feeds, 190
  defined, 241
  drilling for, 188
  evaluating condition of, 14–15, 160
  fishing, 191–93, 207, 228, 229
  flexible metal (installing), 202
  home run, marking, 189
  low-voltage. See Low-voltage (Lo Vo or LV) fixtures
  MC, installing, 204
  metal, 202–04
  nail plates protecting, 189, 197
  ordering, 183
  pulling, 189, 197
  reading codes on, 31
  securing to framing, 33, 189
  service panel attachment, 8
  standoffs/clips for, 33, 189
  stapling, 33, 189
  stripping, 34-35
  supporting, requirements, 183
  12/3, 12/2, 14/3, and 14/2, 30, 31
  types of, 30, 31
  from utility service, 10, 11, 161, 239
Ceiling fans, 151–55, 168
Chandeliers, 75–77, 85–87
Circuit breakers. See also Service panels
  adding, 242–43
  AFCI, 164, 243
  amp ratings, 8
  code requirements, 183
  cutting power at, 16
  double-pole, 243
  GFCI, 243
  lockouts, 179, 238
  repeated tripping of, 161
  types of, 243
  wire gauges matching size of, 15
Circuits
  adding, 162
  bathroom and kitchen, 162, 165
  capacities, 161
  code requirements, 163, 164
  dedicated, 162, 165, 212, 219
  designated, 219
  load calculation, 162
  overloaded, 161
  requirements by room, 165
  wiring requirements, 163
Clamps and clips, 32, 182
Conductors, 240, 241
Conduit, 30, 31, 202, 205–09, 218–19, 227, 231, 232
Connectors, cable, 32, 182, 202
Connectors, wire, 36, 37, 182. See also Clamps and clips
Connectors/jacks, multimedia, 122, 133–37
Continuity testers, 24
Cordless power tools, 21, 23
Current, defined, 6
Cut-in boxes, 26
Cutters and pliers, 21, 22

D

Dedicated circuits, 162, 165, 212, 219
Designated circuits, 219
Diagonal pliers, 21, 22
Dishwashers, 215
Doorbells, 110–15
Drill bits, 23, 181
Drill extension, 181
Drilling for cable, 188
Drills, 23, 180, 181
Drywall saws, 22
Dust masks, 7, 180

E

Electrical notation (systems), 168
Electrical systems. See also Planning new work; Rough-in wiring
  components, illustrated, 10–11 (see also specific components)
  inspecting/evaluating, 14–15, 158–62
  service drops/entrances, 10, 11, 161, 239
  sizing of, 161–62
Electricity, understanding, 6

F

Fans, 138–55
  bathroom, 141–47
  integral junction boxes in, 140
  range hoods, 148–50
  roof vents for, 143–44
  wiring schematics, 140
Feeds, defined, 241
Fish tape, 21, 22, 181
Fishing cable, 191–93, 207, 228, 229

- Floor plan, developing, 165-67
- Fluorescent bulbs, 69
- Fluorescent fixtures, 92-93
- Furnace transformers, 118-19
- Fuse boxes, 9, 16, 158-59, 161-62
- Fuses
  - amp ratings, 8
  - blown, 9, 161
  - "pennying," 159
  - types of, 9

G

- Garbage disposers, 214-15
- GFCI breakers, 243
- GFCIs (ground-fault circuit interrupters), 10, 13, 40, 47, 165, 171, 226
- Glossary, 248
- Grounding. See also GFCIs (ground-fault circuit interrupters)
  - basics of, 12-13, 15
  - cable and conduit systems, 209
  - code requirements, 163, 183
  - equipment, 212
  - service panels, 12-13, 15, 159, 163
  - ungrounded outlets, 47, 160

H

- Halogen bulbs, 69, 84
- Hammers, 22
- Hand tools, 21, 22, 180
- Hickey, defined, 68
- Holes, cutting, 180-81, 194, 195

I

- Incandescent bulbs, 69
- Inductance testers, 18-19, 21, 24, 43
- Inspections, 3, 167, 178-79
- Inspectors, reviewing plans, 166, 167

K

- Kitchen circuits/lighting, 162, 164, 165. See also Lighting
- Knob-and-tube wiring, 30, 160

L

- Lamps, 68, 70-74
  - cleaning and rewiring, 70-73
  - replacing plugs, 74
  - safety precaution, 71
  - testing, 70-71
  - tinning wires for, 73
- Lasers, 180, 184, 185
- Laundry setups, 218-20
- Leads, defined, 68
- Levels, 22, 180
- Light box, defined, 68. See also Boxes
- Lighting, 66-107. See also Lamps; Low-voltage (Lo Vo or LV) fixtures; Outdoor wiring
  - bathroom, 165
  - bulb types, 69
  - ceiling box for, 85, 99
  - ceiling fixture elements, 78
  - chandeliers, 75-77, 85-87
  - fan-light combinations, 155
  - general, 164
  - installing bulbs, 84
  - kitchen, 164, 165
  - load calculation, 162
  - mounting fixtures, 78-79
  - porcelain fixtures (installing), 80-81
  - recessed, 88-91, 199
  - removing fixtures, 82
  - safety precautions, 71, 81, 82
  - sample plan, 168-69
  - symbols for fixtures, 168
  - task, 164
  - terms, 68
  - undercabinet fluorescent, 92-93
  - wall sconces, 82-84
  - wiring fixtures, 68 (see also Switches, wiring; specific fixture type)
- Line voltage, defined, 68
- Lineman's pliers, 21, 22
- Load calculation, 162
- Load ratings, 41
- Low-voltage (Lo Vo or LV) fixtures, 68, 94-107
  - attaching feed rods and fixtures, 102
  - canopies, 95, 98, 101
  - installing cable lighting system, 95-102
  - installing light fixtures, 107
  - installing monorail track lighting, 103-07
  - mounting box for, 99
  - outdoor path lights, 233-34
  - running and tensioning cables, 96-97
  - schematics, 95, 103
  - standoffs, 95-96, 104
  - systems overview, 94
  - transformers, 68, 95, 99-100, 105, 115, 118-19
- Low-voltage controls, 108-19
  - doorbells, 110-15
  - furnace transformers, 118-19
  - thermostats, 116-17
- Lumens, defined, 68

M

- Main breaker, 8-9
- Media distribution panels, 127-32
- Meter-main combos, 9
- Meters, 10, 11
- Motion detector, wiring, 63
- Motion-sensor lights, 225, 235
- Multimedia, 120-37
  - cable, 122-23
  - connecting cables at panel, 131
  - connectors and jacks, 122, 133-37
  - faceplate connections and options, 136-37
  - installing cable connectors, 133
  - media distribution panels, 127-32
  - outlets, locating/installing, 132
  - phone lines and jacks, 124-25
  - signal strength, 122
  - stripping cable, 123
  - twisted wires, 123
- Multimeters, 24

N

- National Electric Code® (NEC), 3, 163
- Needle-nose pliers, 21, 22
- Neon voltage testers, 17, 24
- Nipple, defined, 68
- Notation, electrical, 168
- Nut-driver bits, 180, 181

O

- Ohms, defined, 6
- Omnidirectional flare, defined, 68
- Outdoor wiring, 222-35
  - adding receptacles, 229
  - choosing lights, 225
  - getting power, 224
  - holiday lights, 229
  - low-voltage path lights, 233-34
  - motion-sensor lights, 225, 235
  - outlets, 230
  - planning, 224
  - power to outbuilding, 231-32
  - safety precautions, 224
  - tapping into existing outlet, 228-29
  - tools and materials, 226-27
  - weatherproof covers, 226-27
- Outlets. See Receptacles
- Oven outlets/connections, 216-18

P

Phone lines and jacks, 124-25
Pilot light switches, 62
Planning new work, 156-75. See also Building codes
  adding outlets and circuits, 162
  assessing wiring condition, 160
  developing floor plan, 165-67
  electrical notation (symbols) for, 168
  electrical system overview, 10-11
  getting feedback, 166
  inspecting fuse box/breaker panel, 158-59
  load calculation, 162
  materials lists, 166, 167
  on-site inspections, 167
  pro's take on rewiring, 160
  remodeling kitchens, 162
  requirements by room, 165
  reviewing plans with inspector, 166, 167
  sample plans (lighting/switches, and power), 168-69
  sizing system, 161-62
  wiring portfolio, 170-75 (see also Receptacles, wiring; Switches, wiring)
Pliers and strippers, 21, 22
Plug-in circuit analyzers, 21, 24
Plumb laser, 184
Polarized fixtures, 74
Polarized receptacles, 41, 47
Power, defined, 6
Power tools, 23, 180-81
Primary wires, defined, 68

R

Range hoods, 148-50
Receptacles, 38, 40-49
  adding, 162, 229
  amp ratings, 40-41
  back-wired, 40, 42
  choosing, 40-41
  circuit capacities, 161
  code requirements, 164
  GFCIs, 10, 13, 40, 47, 165, 171, 226
  grounding, 15, 47, 160, 163
  load calculation, 162
  locating, 27, 47, 184, 185, 186
  making flush to drywall, 47
  matching load ratings, 41
  neutral wires attached to, 43
  for 120/240 range, 175
  outdoor, 230
  polarized, 41, 47
  spacers (caterpillars) building up, 47
  surface metal boxes for, 208-09
  symbols for, 168
  testing, 15, 17-19, 43
  three-prong, 10
  two-slot ungrounded, 15, 47, 160
Receptacles, wiring, 198
  controlled by switch, 174-75
  duplex receptacles, 44-46, 170-71, 174-75
  at end of circuit, 144, 170
  feeding wires through receptacle, 144, 170
  ganged receptacles, 171
  GFCIs, 47, 48, 171
  in midcircuit, 144, 170
  portfolio of options, 170-71, 174-75
  split-tab receptacles, 49, 174-75, 215
  switch/receptacle combo, 62
Recessed lighting, 88-91, 199
Reciprocating saws, 23
Roof vents, 143-44
Rotary cutters, 23
Rough-in wiring, 176-209
  box locations, 184, 185, 186-87
  code requirements, 183 (see also Building codes)
  conduit systems, 205-09
  corner cable runs, 190
  creating wiring trench, 196
  cutting box holes, 194, 195
  drilling for cable, 188
  fishing cable, 191-93, 207
  inspection, 178-79
  installing boxes, 186-87
  laying out job, 184-88
  materials for, 182-83
  metal cable, 202-04
  ordering materials, 183
  outlet box, 198
  pulling cable, 189, 197
  recessed light can, 199
  retrofitting boxes, 195, 197
  starting and organizing, 178
  switch boxes, 200-201
  tools for, 180-81
Running thread, defined, 68

S

Safety. See also Grounding
  avoiding shocks, 12
  evaluating system, 14-15, 158-62
  gear, 7, 180
  general guidelines, 7
  lighting fixtures, 71, 81, 82
  respecting power, ii, 4, 7
  service panels, 7, 179, 238
  wiring in orderly way, 45
Safety glasses, 7
Saws, 22, 23, 181
Sconces, 82-84
Screwdrivers, 21, 22, 180
Secondary wires, defined, 68
Service conductors, 10, 11
Service drops/entrances, 10, 11, 161, 239
Service panels, 8-9, 236-45. See also Circuit breakers
  cable attachments, 8
  components, illustrated, 8, 9, 240
  cutting power at, 16
  grounding, 12-13, 15, 159, 163
  how it fits into system, 11
  inspecting/evaluating, 14-15, 158-59
  internal workings, 8-9
  key terms, 241
  main breaker, 8-9
  meter-main combos, 9
  minimum amps for, 14
  safety precautions, 7, 179, 238
  sizing, 161-62, 240-41
  subpanels, 9, 236, 238-41, 244-45
  understanding, 238-39
  7-in-1 tool, 22
Soldering wires, 73
Solenoid voltage testers, 24
Source (supply) wires, 68
Splicing wires, 37, 64, 180, 183
Staples, 33, 189
Stepladders, 180
Strippers and pliers, 21, 22
Stripping cable, 34-35, 123
Stripping wires, 37
Subpanels, 9, 236, 238-41, 244-45
Switches, 38, 40-43, 50-65
  back-wired, 40, 42
  choosing, 40-41
  code requirements, 164
  convertible, 55
  dimmer, 40, 56-57
  double, 63
  four-way, 51, 61
  locating, 184, 185, 186
  matching load ratings, 41
  motion detector, 63
  neutral wires attached to, 43
  pilot light, 62
  replacing (single-pole), 54-55
  roughing in (three-way), 58
  sample plan, 168-69
  single-pole, 50, 52-55
  specialty, 40, 56-57, 63
  symbols for, 168

  - testing, 43, 50–51, 52
  - three-way, 51, 58–60, 172–74
  - timer, 64–65
- Switches, wiring, 200–201
  - with back-fed wiring, 53, 60, 172
  - controlling receptacles, 174–75
  - double switches, 63
  - with fixture at end of cable run, 172
  - four-way, 61
  - ganged switches, 172, 201
  - motion detector, 63
  - pilot light, 62
  - portfolio of options, 172–75
  - single-pole, 52–53, 200
  - slide dimmers, 56–57
  - switch/receptacle combo, 62
  - three-way, 59–60, 172–74
  - timers, 64–65
- Symbols, common, 168

T

- Tape measures, 22, 180
- Terms, 6, 68, 241. See also Glossary
- Testing
  - receptacles, 15, 17–19, 43
  - switches, 43, 50–51, 52
  - testers, 50
  - testers for, 21, 24
- Thermostats, 116–17
- Timer switches, 64–65
- Tinning wires, 73
- Tool belts, 22, 181
- Tools, 20–25, 180–81, 226. See also specific tools
- Transformers, 68, 95, 99–100, 105, 115, 118–19

U

- Undercabinet fixtures, 92–93
- Utility knife, 21, 34

V

- Volts, amps, and watts, 6, 161

W

- Water heaters, 221
- Watts, volts, and amps, 6, 161
- Wire
- aluminum, 14, 30, 160
- choosing, 36
- defined, 241
- splicing, 37
- strippers/crimpers, 21
- stripping, 37
- types of, 36
- Wire connector driver, 180
- Wire connectors, 36, 37
- Wire reels, 180, 181
- Wiring. See also Receptacles, wiring; Switches, wiring; specific components
- assessing condition of, 160
- circuit, types of, 30
- code requirements, 163
- new work. See Planning new work; Rough-in wiring
- safety precaution, 45
- Wiring trench, 196

HOUSE & HOME

Let the experts show you how

Get your wiring project done right the first time with site-tested advice from the pros. The most comprehensive reference for homeowners available, *Wiring Complete* is filled with trade secrets only a pro would know, as well as detailed instructions and over 800 photos and drawings that show every step.

REVISED EDITION

WIRING COMPLETE

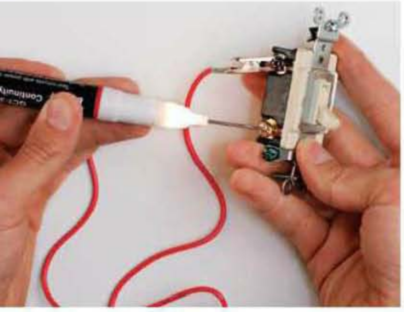

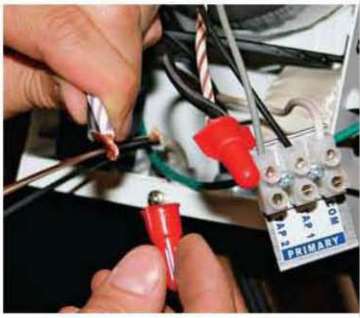

TABLE OF CONTENTS

- » UNDERSTANDING ELECTRICITY
- » TOOLS, MATERIALS, AND TECHNIQUES
- » RECEPTACLES AND SWITCHES
- » LIGHTING
- » REPAIRING LOW-VOLTAGE CONTROLS
- » MULTI-MEDIA
- » FANS
- » PLANNING
- » ROUGH-IN
- » APPLIANCES
- » OUTDOOR WIRING
- » PANELS AND SUBPANELS

Michael Litchfield is best known as the author of the home renovation bible, *Renovation: A Complete Guide* (Taunton, 2005).

Michael McAlister is a licensed electrician working in Berkeley, CA.

The Taunton Press also publishes *Fine Homebuilding* and finehomebuilding.com, where today's builders and architects share what they know about building and remodeling.

Look for other Taunton Press books wherever books are sold or visit our website at www.taunton.com

The Taunton Press
Inspiration for hands-on living®

Taunton Direct
63 South Main Street
P.O. Box 5507
Newtown, CT 06470-5507
(800) 888-8286

US $21.95
ISBN 978-1-60085-256-5
52195
9 781600 852565

Taunton Product #071303

BLACK&DECKER®

NEW FOR 2009

The Complete Photo Guide to
HOME IMPROVEMENT

More than 200 Value-Adding Remodeling Projects

BLACK&DECKER®

THE COMPLETE PHOTO GUIDE TO

HOME IMPROVEMENT

Creative Publishing International

MINNEAPOLIS, MINNESOTA
www.creativepub.com

Creative Publishing International

Copyright © 2009
Creative Publishing international, Inc.
400 First Avenue North, Suite 300
Minneapolis, Minnesota 55401
1-800-328-0590
www.creativepub.com
All rights reserved

Printed in China

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

The complete photo guide to home improvement.
p. cm.
"Black & Decker."
Includes index.
Summary: "Features more than 200 common do-it-yourself remodeling projects"--Provided by publisher.
ISBN-13: 978-1-58923-452-9 (hard cover)
ISBN-10: 1-58923-452-9 (hard cover)
1. Dwellings--Maintenance and repair--Amateurs' manuals. 2. Dwellings--Remodeling--Amateurs' manuals. I. Black & Decker Corporation (Towson, Md.) II. Title.

TH4817.3.C654 2008
643'.7--dc22

2008045755

President/CEO: Ken Fund
VP for Sales & Marketing: Kevin Hamric

Home Improvement Group

Publisher: Bryan Trandem
Managing Editor: Tracy Stanley
Senior Editor: Mark Johanson
Editor: Jennifer Gehlhar

Creative Director: Michele Lanci-Altomare
Senior Design Managers: Jon Simpson, Brad Springer
Design Managers: James Kegley

Lead Photographer: Joel Schrell
Photo Coordinator: Joanne Wawra
Shop Manager: Bryan McLain
Shop Assistant: Cesar Fernandez Rodriguez

Production Managers: Linda Halls, Laura Hokkanen

Page Layout Artist: Danielle Smith

The Complete Photo Guide to Home Improvement
Created by: The Editors of Creative Publishing international, Inc., in cooperation with Black & Decker.
Black & Decker® is a trademark of The Black & Decker Corporation and is used under license.

NOTICE TO READERS

For safety, use caution, care, and good judgment when following the procedures described in this book. The publisher and Black & Decker cannot assume responsibility for any damage to property or injury to persons as a result of misuse of the information provided.

The techniques shown in this book are general techniques for various applications. In some instances, additional techniques not shown in this book may be required. Always follow manufacturers' instructions included with products, since deviating from the directions may void warranties. The projects in this book vary widely as to skill levels required: some may not be appropriate for all do-it-yourselfers, and some may require professional help.

Consult your local building department for information on building permits, codes, and other laws as they apply to your project.

Contents

The Complete Photo Guide to Home Improvement

Introduction 7

HOME IMPROVEMENT SKILLS

Anatomy of a House 10
Building Codes & Permits 16
Working with Drawings 20
Making Construction Plans 22
Working Safely 24

Techniques. 27

Demolition 28
Building Walls 40
Installing Wallboard 44
Understanding Plumbing 52
Working with Copper Pipe 56
Working with Plastic Pipe 64
Working with PEX Pipe. 68
Planning New Plumbing. 76
Understanding Wiring. 82
Working with Wire & Cable 90
Working with Conduit 102
Working with Boxes 108
Working with Switches 116
Working with Receptacles 122
Planning New Wiring 128
Using Circuit Maps 138

HOME IMPROVEMENT BASICS

Flooring 149

Installing Underlayment. 150
Installing Radiant Floor Mats 156
Installing Hardwood Strip Flooring . . . 162
Installing Laminate Floors 166
Installing Parquet Flooring 170
Installing Sheet Vinyl 174
Evaluating an Existing Floor. 176
Installing Resilient Tile 180
Installing Bamboo Flooring 186
Installing Ceramic Tile 191
Installing Mosaic Tile 198
Installing Carpet Squares. 200
Refinishing Wood Floors. 204

Walls & Ceilings 209

Installing Paneling 210
Installing Beadboard Wainscoting . . . 214
Building Glass Block Walls 220
Applying Veneer Plaster. 224
Paneling Ceilings 228
Installing Suspended Ceilings 232
Installing Acoustical Ceiling Tiles. . . . 236
Installing Metal Ceilings 239
Installing Base Molding 242
Installing Picture Rail 246
Installing Chair Rail. 248
Installing Crown Molding 250

Windows & Doors 255

Framing & Installing Windows 256
Installing Garden Windows 264
Installing Bay Windows 268
Installing Skylights 278
Installing Tubular Skylights. 288
Installing Interior Doors 290

Installing Entry Doors 306
Installing Storm Doors 310
Installing Patio Doors 312
Installing Door & Window Casing. . . . 318

Lighting 323
Installing Ceiling Lights. 324
Installing Recessed Ceiling Lights . . . 327
Installing Track Lighting 330

MAJOR REMODELING

Kitchen Remodeling 337
Planning & Designing Kitchens 338
Installing Kitchen Cabinets. 350
Building Kitchen Islands 358
Building Tiled Islands 360
Installing Postform Countertops 364
Installing Laminate Countertops 368
Installing Tile on Countertops 376
Installing Tile Backsplashes 386
Installing Kitchen Sinks 390
Replacing Kitchen Faucets. 392
Connecting Drains 398
Hooking Up Dishwashers 402
Installing Food Disposers 407
Installing Undercabinet Lights 412

Bathroom Remodeling. . . . 421
Planning & Designing Bathrooms . . . 422
Installing Shower Kits. 430
Installing Alcove Bathtubs 438
Installing 3-piece Tub Surrounds 446
Installing Sliding Tub Doors 450
Tiling Bathroom Walls. 454
Installing Toilets 458
Installing Wall-hung Vanities 462
Installing Vessel Sinks 464
Installing Pedestal Sinks. 468
Installing Vanity Cabinets 470
Installing Bathroom Sink Faucets & Drains 472
Installing Tub Spouts 480
Installing Vanity Lights 484
Installing Grab Bars 486
Installing Glass Shelving. 488
Installing Towel Warmers 490

Basements & Attics 493
Evaluating Your Attic 494
Evaluating Your Basement 496
Planning the Framing 498
Installing Basement Floors. 500
Installing Attic Floors 506
Installing Basement Walls 510
Installing Attic Walls 514
Finishing Attic Ceilings. 516
Adding Basement Egress Windows . . . 518
Installing Baseboard Heaters 524

Exterior Improvements . . . 529
Estimating Roofing, Siding & Trim . . . 530
Preparing the Roof for Shingles 532
Shingling a Roof 538
Installing Vinyl Siding 546

Resources/Photo Credits. . . 554

Conversion Charts 555

Index 556

Introduction

Home improvement is a huge subject that demands a big book. *The Complete Photo Guide to Home Improvement* fits the bill. With over 200 projects and more than 2,000 color photos, all of the most popular home remodeling subjects and projects are included in this single resource. Here's one way to look at it: If a picture is indeed worth a thousand words, it would require an 8,000-page volume to convey in written form the information packed into the photos in this book. And unlike the tiny pictures found in other home improvement books, the photos within these pages are large and clear and very easy to follow.

Doing your own home remodeling work takes a combination of skills, planning, and specific information. For example, if you decide to remodel your kitchen (the most frequently remodeled room), you should start the process by defining the scale of the project and identifying the tasks within the larger job that you will perform yourself. This phase corresponds to the first section of this book, "Home Improvement Basics." If you love getting your hands dirty this won't be your favorite part, but it is quite important.

Once you have some background information under your belt and a rough plan in your head, it's time to tackle the skillbuilding most of us need. In "Techniques" you will find clear how-to information for just about every task, from demolition to stripping cable. Thorough information on home wiring and home plumbing fill out a large part of the section, followed by step-by-step instructions for installing flooring, finishing walls and ceilings, working with windows and doors, and—finally—upgrading ventilation and lighting.

In the final section of the book you will move from the general to the more specific. Individual chapters on kitchen remodeling and bathroom remodeling deal with projects, planning and information that are specific to each subject. Here you'll find a host of options for installing cabinets and countertops, for example, to supplement the information you've encountered in the Techniques portion of the book. Converting attics and basements also gets its own section of targeted information. And for a strong finish, we offer extended how-to sequences for the two most common exterior improvements—roofing and installing siding.

Whether you are simply freshening up the appearance of your house or embarking on a down-to-the studs remodel of your bathroom, you'll find the information you need. Even if you have special remodeling needs, such as creating accessible kitchens and baths, you'll see that the subjects are covered. More than a quarter of a million homeowners have relied on *The Complete Photo Guide to Home Improvement* as their principal guidebook through the remodeling process. And now, with this new 3rd edition, you'll be pleased that you chose to join them.

shatter resistant

HOME IMPROVEMENT SKILLS

Pionite
See Larger Sample
AC381
Suede
Storm
Cirrus
AV741
Suede
AC361
Suede
Graphite
Talc
10
11 MADE IN CHINA 12
FL02

Anatomy of a House

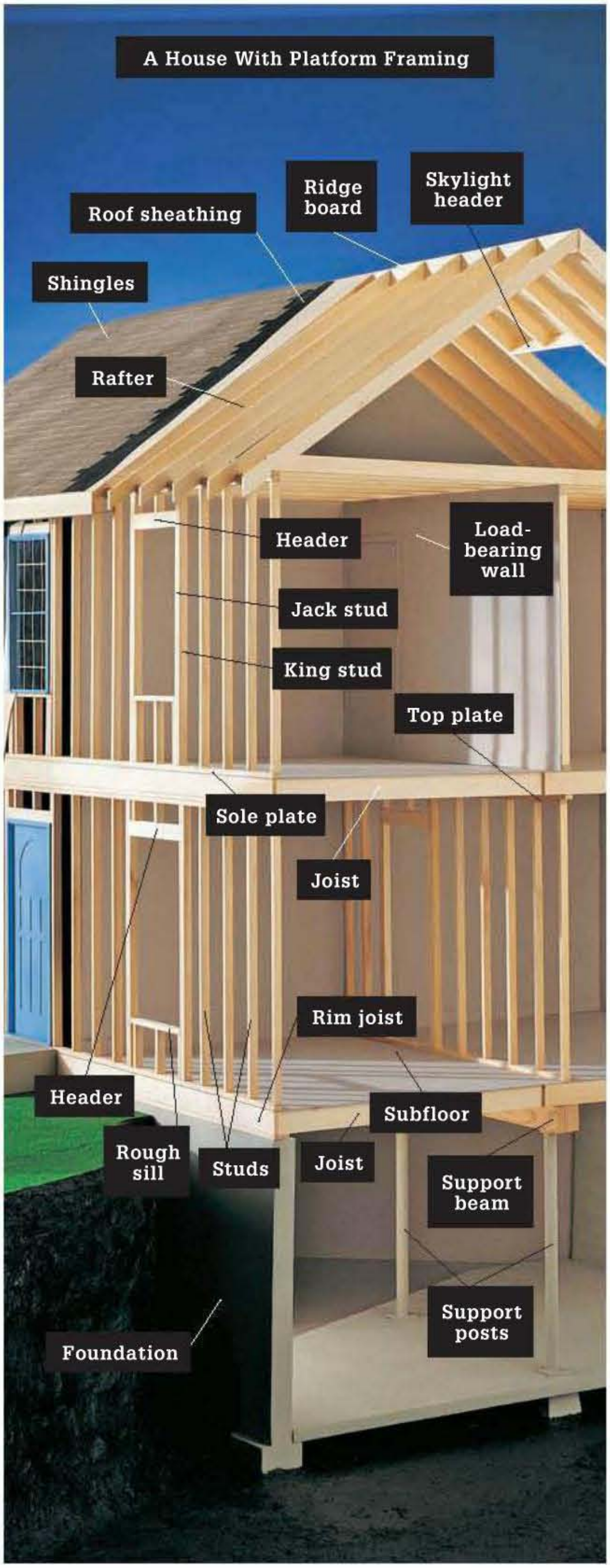
A House With Platform Framing
Roof sheathing
Ridge board
Skylight header
Shingles
Rafter
Header
Load-bearing wall
Jack stud
King stud
Top plate
Sole plate
Joist
Rim joist
Header
Subfloor
Rough sill
Studs
Joist
Support beam
Support posts
Foundation

Before you start a do-it-yourself home improvement project, you should familiarize yourself with a few basic elements of home construction and remodeling. Take some time to get comfortable with the terminology of the models shown on the next few pages. The understanding you will gain in this section will make it easier to plan your project, buy the right materials, and clear up any confusion you might have about the internal design of your home.

If your project includes modifying exterior or load-bearing walls, you must determine if your house was built using platform- or balloon-style framing. The framing style of your home determines what kind of temporary supports you will need to install while the work is in progress. If you have trouble determining what type of framing was used in your

ANATOMY OF A HOUSE WITH PLATFORM FRAMING

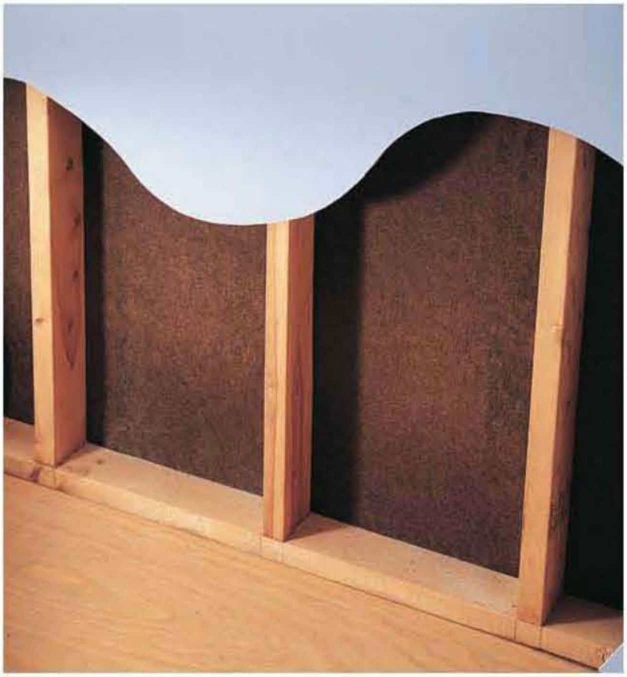

Platform framing (photos, left and above) is identified by the floor-level sole plates and ceiling-level top plates to which the wall studs are attached. Most houses built after 1930 use platform framing. If you do not have access to unfinished areas, you can remove the wall surface at the bottom of a wall to determine what kind of framing was used in your home.

home, refer to the original blueprints, if you have them, or consult a building contractor or licensed home inspector.

Framing in a new door or window on an exterior wall normally requires installing a header. Make sure that the header you install meets the requirements of your local building code, and always install cripple studs where necessary.

Floors and ceilings consist of sheet materials, joists, and support beams. All floors used as living areas must have joists with at least 2 × 8 construction. For modification of smaller joists see page 14.

There are two types of walls: load-bearing and partition. Load-bearing walls require temporary supports during wall removal or framing of a door or window. Partition walls carry no structural load and do not require temporary supports.

ANATOMY OF A HOUSE WITH BALLOON FRAMING

Balloon framing (photos, right and above) is identified by wall studs that run uninterrupted from the roof to a sill plate on the foundation, without the sole plates and top plates found in platform-framed walls (page opposite). Balloon framing was used in houses built before 1930, and it is still used in some new home styles, especially those with high vaulted ceilings.

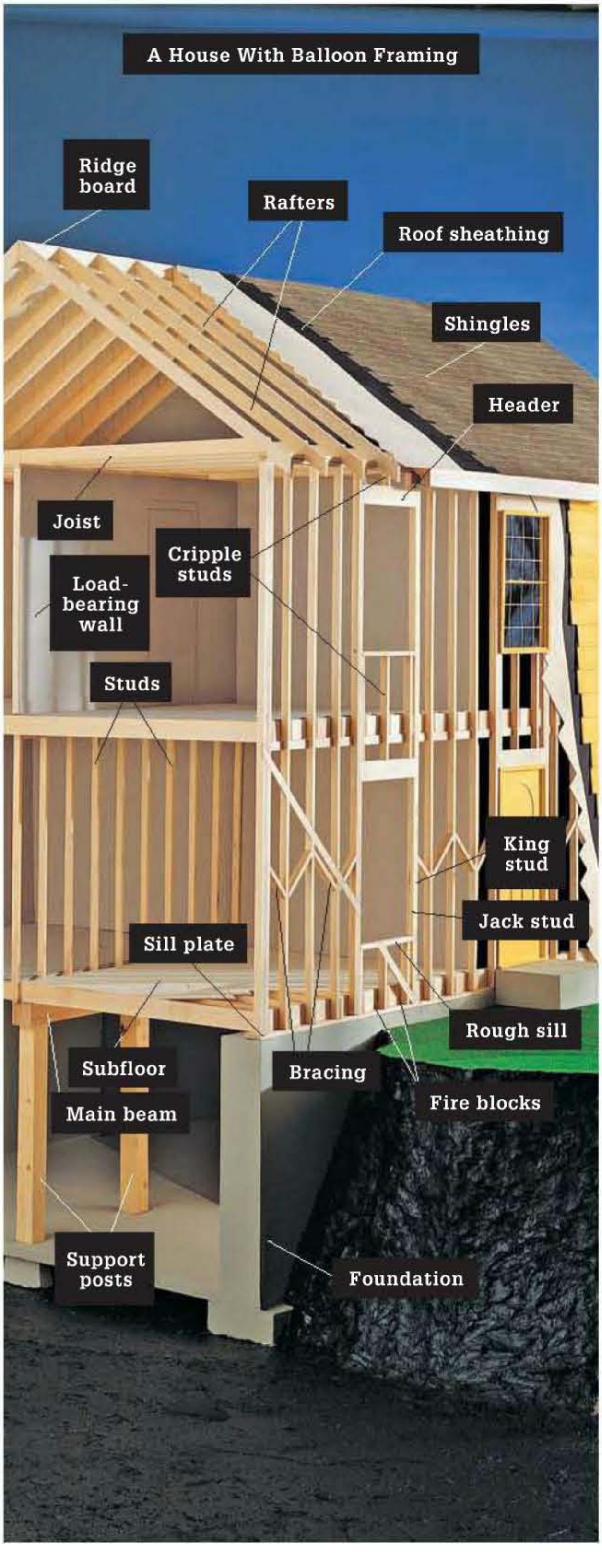
A House With Balloon Framing
Ridge board
Rafters
Roof sheathing
Shingles
Header
Joist
Cripple studs
Load-bearing wall
Studs
King stud
Jack stud
Sill plate
Rough sill
Subfloor
Bracing
Fire blocks
Main beam
Support posts
Foundation

Anatomy Details

Many remodeling projects, like adding new doors or windows, require that you remove one or more studs in a load-bearing wall to create an opening. When planning your project, remember that new openings require a permanent support beam called a header, above the removed studs, to carry the structural load directly.

The required size for the header is set by local building codes and varies according to the width of the rough opening. For a window or door opening, a header can be built from two pieces of 2" dimensional lumber sandwiched around ⅜" plywood (chart, right). When a large portion of a load-bearing wall (or an entire wall) is removed, a laminated beam product can be used to make the new header.

If you will be removing more than one wall stud, make temporary supports to carry the structural load until the header is installed.

Recommended Header Sizes

Rough Opening Width	Recommended Header Construction
Up to 3 ft.	⅜" plywood between two 2 × 4s
3 ft. to 5 ft.	⅜" plywood between two 2 × 6s
5 ft. to 7 ft.	⅜" plywood between two 2 × 8s
7 ft. to 8 ft.	⅜" plywood between two 2 × 10s

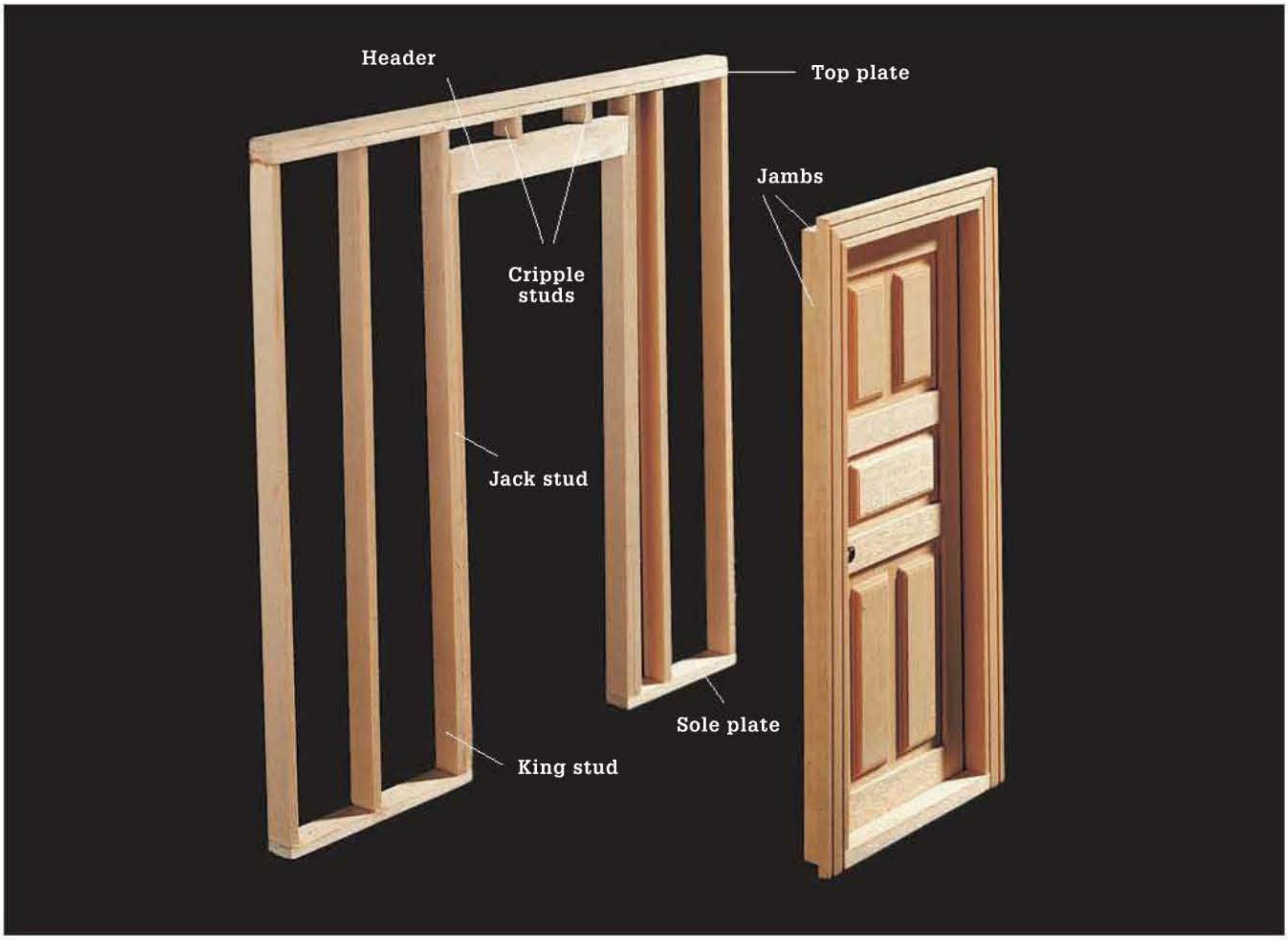

Door opening: The structural load above the door is carried by cripple studs that rest on a header. The ends of the header are supported by jack studs (also known as trimmer studs) and king studs that transfer the load to the sole plate and the foundation of the house. The rough opening for a door should be 1" wider and ½" taller than the dimensions of the door unit, including the jambs. This extra space lets you adjust the door unit during installation.

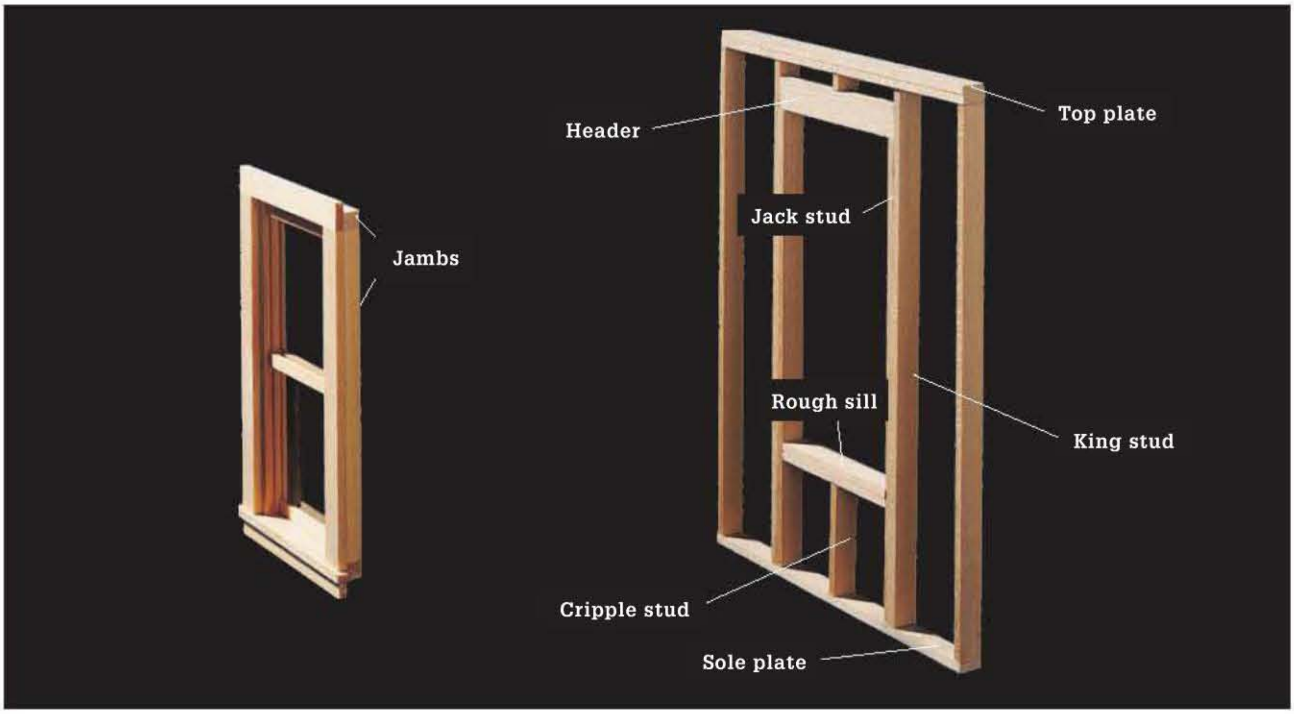

Window opening: The structural load above the window is carried by cripple studs resting on a header. The ends of the header are supported by jack studs and king studs, which transfer the load to the sole plate and the foundation of the house. The rough sill, which helps anchor the window unit but carries no structural weight, is supported by cripple studs. To provide room for adjustments during installation, the rough opening for a window should be 1" wider and ½" taller than the window unit, including the jambs.

Framing Options for Window & Door Openings (new lumber shown in yellow)

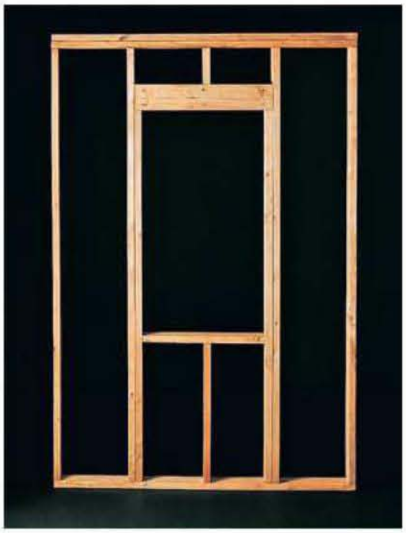

Using an existing opening avoids the need for new framing. This is a good option in homes with masonry exteriors, which are difficult to alter. Order a replacement unit that is 1" narrower and ½" shorter than the rough opening.

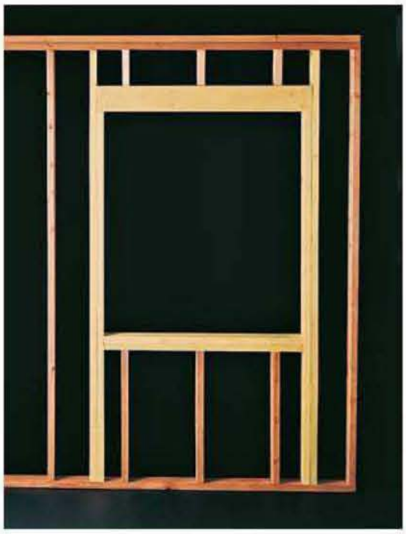

Framing a new opening is the only solution when you're installing a window or door where none existed or when you're replacing a unit with one that is much larger.

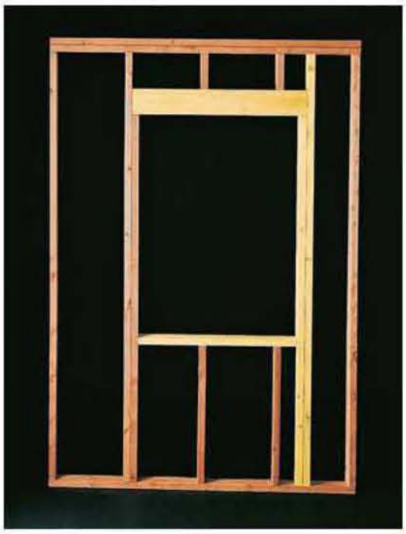

Enlarging an existing opening simplifies the framing. In many cases, you can use an existing king stud and jack stud to form one side of the new opening.

Floor & Ceiling Anatomy

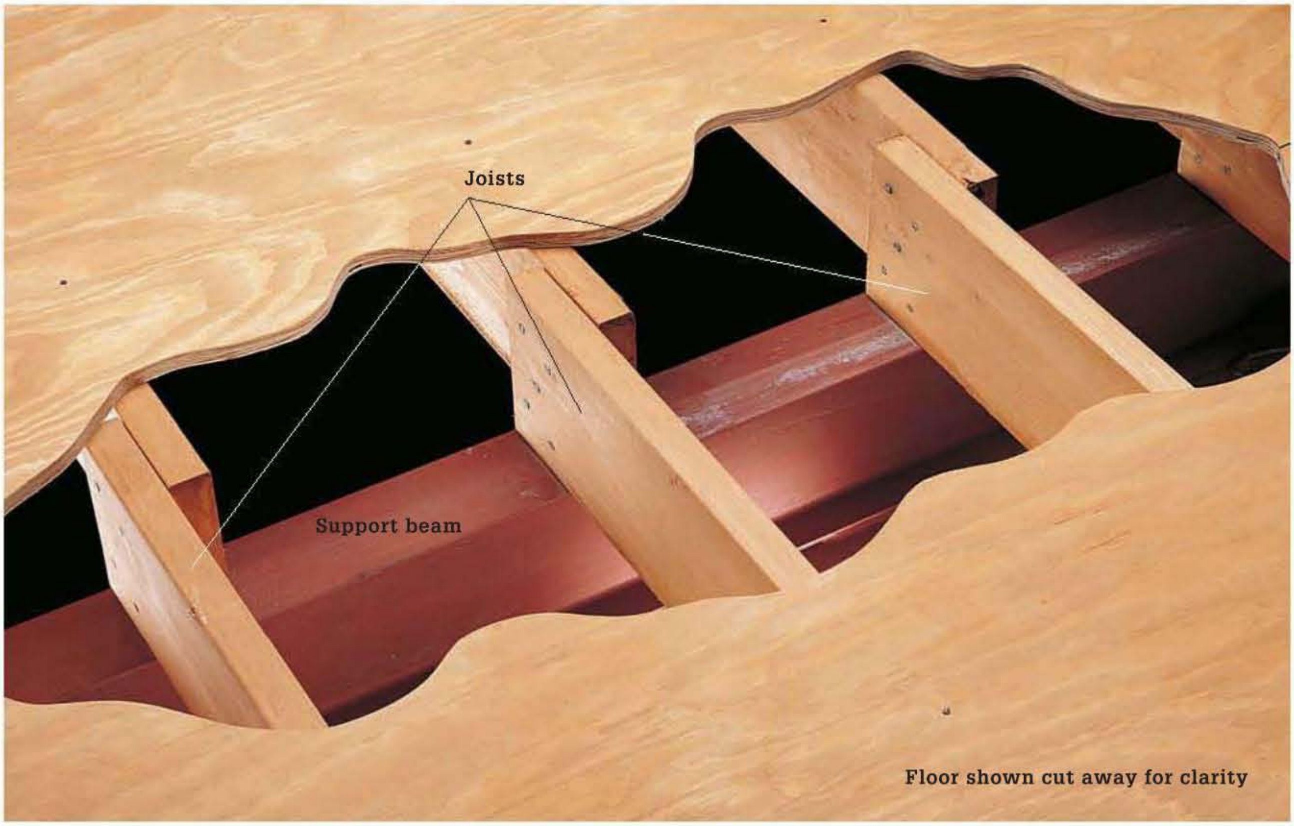
Joists
Support beam
Floor shown cut away for clarity

Joists carry the structural load of floors and ceilings. The ends of the joists rest on support beams, foundations, or load-bearing walls. Rooms used as living areas must be supported by floor joists that are at least 2 × 8" in size. Floors with smaller joists can be reinforced with sister joists (photos, below).

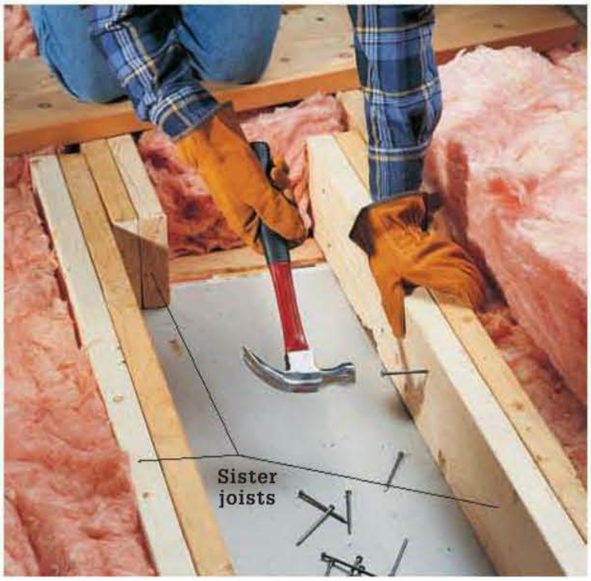
Sister joists

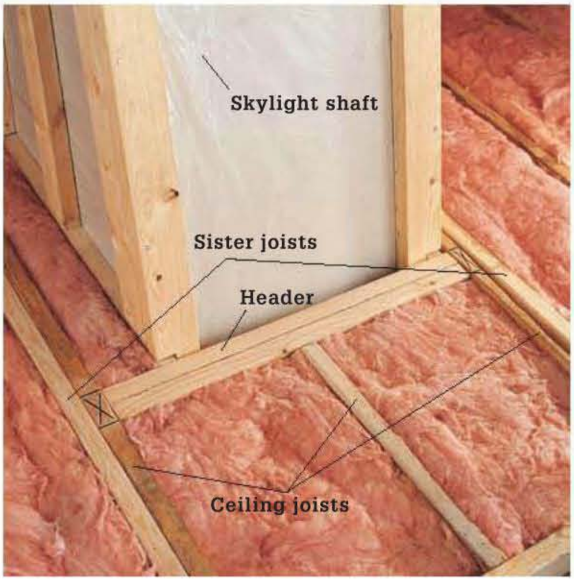
Skylight shaft
Sister joists
Header
Ceiling joists

Floors with 2 × 6 joists, like those sometimes found in attics, cannot support living areas unless a sister joist is attached alongside each original joist to strengthen it (above, left). This often is necessary when an attic is converted to a living area. Sister joists also are used to help support a header when ceiling joists must be cut, such as when framing a skylight shaft (above, right).

Roof Anatomy

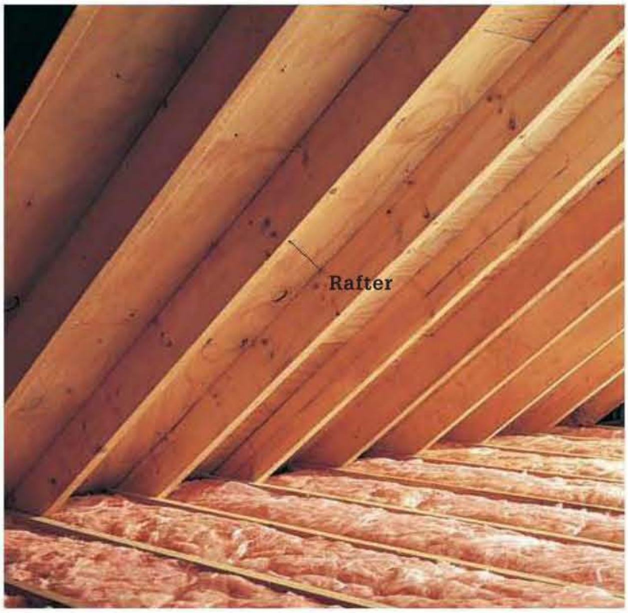

Rafters made from 2 × 4s or 2 × 6s spaced every 16" or 24" are used to support roofs in most houses built before 1950. If necessary, rafters can be cut to make room for a large skylight. Check in your attic to determine if your roof is framed with rafters or roof trusses (right).

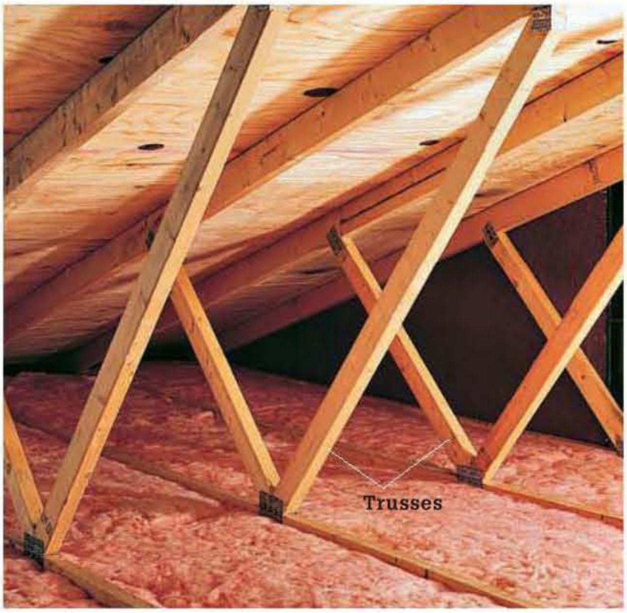

Trusses are prefabricated "webs" made from 2" dimensional lumber. They are found in many houses built after 1950. Never cut through or alter roof trusses. If you want to install a skylight in a house with roof trusses, buy a unit that fits in the space between the trusses.

Wall Anatomy

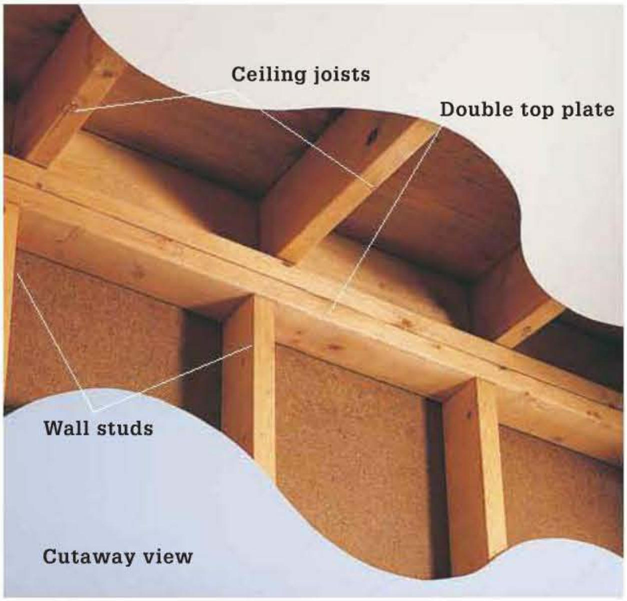
Ceiling joists
Double top plate
Wall studs
Cutaway view

Load-bearing walls carry the structural weight of your home. In platform-framed houses, load-bearing walls can be identified by double top plates made from two layers of framing lumber. Load-bearing walls include all exterior walls and any interior walls that are aligned above support beams.

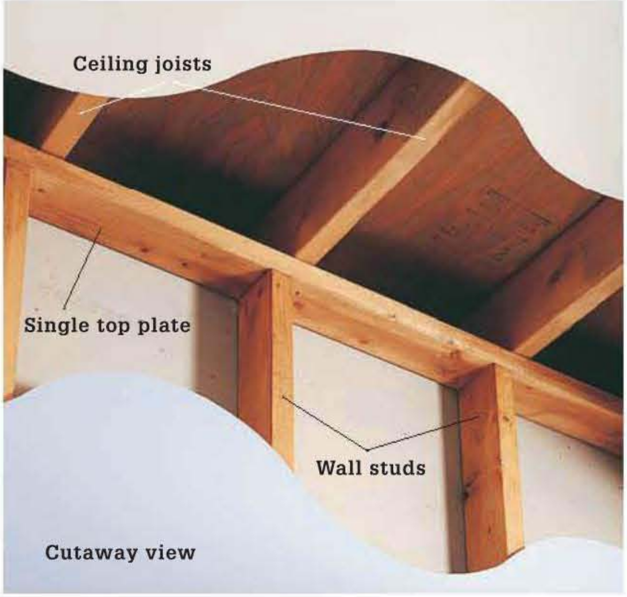
Ceiling joists
Single top plate
Wall studs
Cutaway view

Partition walls are interior walls that do not carry the structural weight of the house. They have a single top plate and can be perpendicular to the floor and ceiling joists but are not aligned above support beams. Any interior wall that is parallel to floor and ceiling joists is a partition wall.

Building Codes & Permits

Building permits are required for any remodeling project that involves a change or addition to your home's structure or mechanical systems. Building permits are issued to ensure your remodeling project meets local building codes, which establish material standards, structural requirements, and installation guidelines for your project. In short, they ensure that your (or your contractor's) work is done properly.

The areas outlined on pages 17 and 18—room dimensions, exits and openings, light and ventilation, and fire protection—are usually covered by general building permits. If your project involves major changes to your plumbing, electrical, or HVAC systems, you may be required to obtain separate permits from the respective administration departments.

Building permits are required by law, and getting caught without them can result in fines from the city and possible trouble with your insurance company. Also, work done without permits can cause problems if you try to sell your house.

Most local building codes follow the national codes, such as the National Electrical Code, but are adapted to meet the demands of local conditions and legislation. Keep in mind that local codes always supersede national codes. Always check with your local building department before finalizing your plans.

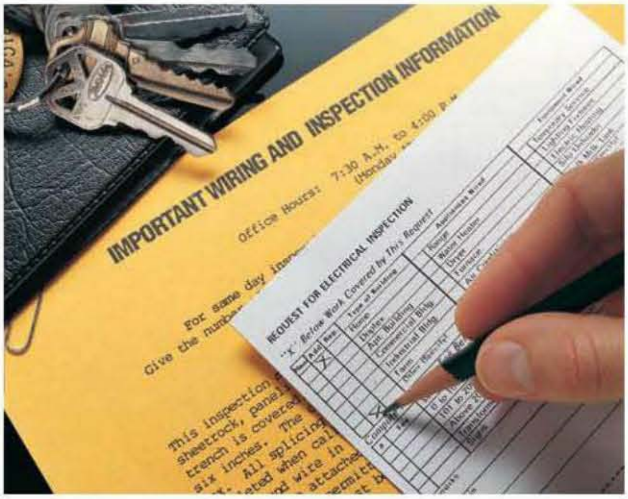
IMPORTANT WIRING AND INSPECTION INFORMATION
REQUEST FOR ELECTRICAL INSPECTION

Before issuing permits, your local building department will require plans and cost estimates for your project. After your plans have been approved, you must pay permit fees, which are based on the cost of the project. You'll also learn what inspections are required and when you should call for inspections.

Once issued, a building permit typically is good for 180 days. You can apply for an extension by submitting a written request showing justifiable cause for the delay.

Tip ▸

Here are some tips to help you prepare for the permit process:

- To obtain a building permit, you must fill out a form from your local building department that includes a description of the project; your home's address, legal description, and occupancy; and an estimate of the project cost.
- The building department may require two to four sets of construction documents or drawings of your project—including floor and elevation plans—to be submitted for inspection and approval.
- A building inspector will examine all construction plans and stamp or send written notification of approval and acceptance.
- One set of approved documents is kept by the building official, one set is sent to the applicant, and one set is displayed at the site until the project is completed.
- Some permits are granted by phase of construction. After the work for one phase is completed and inspected, a permit for the next phase is issued. However, building officials will not guarantee issuance of subsequent permits.
- All work is inspected by a building official to ensure compliance with codes and permits.
- Your project is complete only after the local building inspector makes a final inspection and gives approval of your site.

Room Dimensions

- Habitable rooms must be at least 7 ft. wide and 7 ft. deep.
- Ceilings in all habitable rooms, hallways, corridors, bathrooms, toilet rooms, laundry rooms, and basements must be at least 7 ft., 6" high, measured from the finished floor to the lowest part of the ceiling.
- Beams, girders, and other obstructions that are spaced more than 4 ft. apart can extend 6" below the required ceiling height.
- In nonhabitable rooms, such as unfinished basements, ceilings may be 6 ft., 8" from the floor, and beams, girders, and ducts may be within 6 ft., 4" of the floor.
- Habitable rooms cannot have more than 50% of their floor area under sloped ceilings less than 7 ft., 6" high, and no portion of a floor area can be under a ceiling less than 5 ft. high.
- Finished floor is not considered measurable floor area when it is below sloped ceilings less than 5 ft. high or beneath furred ceilings less than 7 ft., 6" high.
- One habitable room in a home must have at least 120 square feet of gross floor area. Other habitable rooms can have gross floor space of 70 sq. ft. minimum.

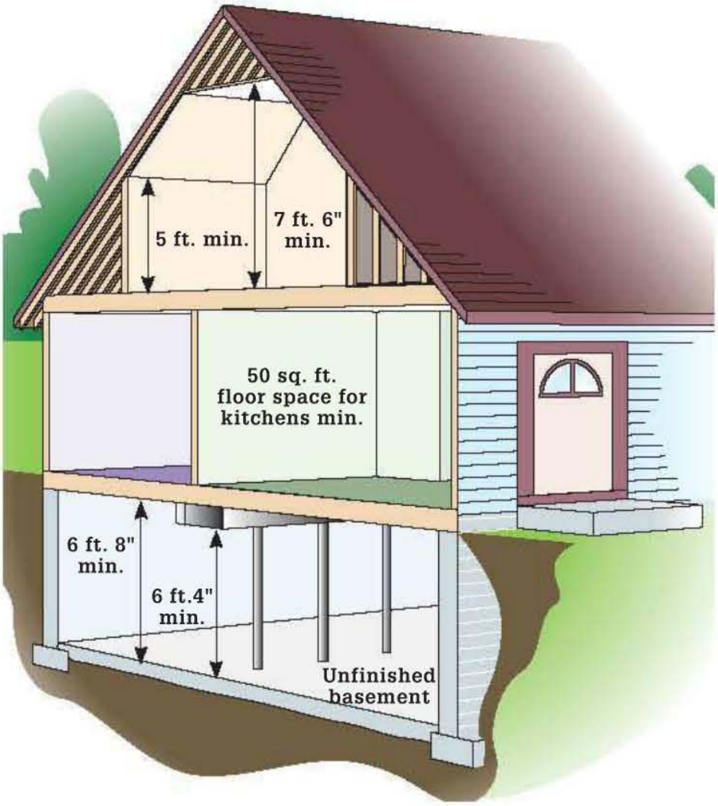
5 ft. min.
7 ft. 6" min.
50 sq. ft. floor space for kitchens min.
6 ft. 8" min.
6 ft.4" min.
Unfinished basement

- Kitchens cannot have less than 50 sq. ft. of gross floor area.
- Hallways must be at least 3 ft. wide.

Exits & Openings

- Sleeping rooms and habitable basements must have at least one egress window or exterior door for emergency escape. Occupants must be able to open the exit from inside the home, without a key or tool.
- An egress window must have a net clear opening of at least 5.7 sq. ft., with a minimum height of 24" and a minimum width of 20".
- Window sills on egress windows cannot be more than 44" above the floor.
- Egress windows below ground level must have window wells. If the wells are deeper than 44", they must have permanent ladders or steps. The steps can project up to 6" into the well but must be usable when the window is fully opened. Steps must be at least 12" wide and project at least 3" from the wall. Ladder rungs must be less than 18" apart.
- Screens, bars, grills, and covers on emergency exits must open easily and be removable from inside the home, without tools or keys.
- Exit doors must be at least 3 ft. wide and 6 ft., 8" high. They must provide direct outside access and operate without special knowledge or tools.
- Bulkhead enclosures may serve as emergency exits in habitable basements if they provide direct access to the basement and meet the dimension requirements for emergency exits.

Natural Light & Ventilation

- Ventilation includes windows, doors, louvers, and other approved openings or mechanical systems.
- Windows must equal at least 8% of the floor area in habitable rooms. The minimum openable area of a window must equal at least 4% of the room's floor area.
- In bathrooms, windows must be at least 3 sq. ft., and at least half of the window must open.
- Windows must open and operate from inside the room, and they must exit to a street, alley, yard, court, or porch.
- Window light can be replaced by an artificial light if it produces 6.46 lux from 30" above the floor.
- Mechanical ventilation can replace operable windows. In bedrooms, ventilation must supply outside air at a rate of 15 cubic ft. per minute (cfm) for each occupant. In primary bedrooms, the rate is based on two occupants. In additional bedrooms, the rate is based on one occupant per room.
- In bathrooms, intermittent mechanical ventilation rates must be 50 cfm, and continuous rates must be 20 cfm. Bathroom ventilation must exhaust to the outside.

Fire Protection

- All concealed and interconnected spaces, such as soffits, drop and cove ceilings, stair stringers, and areas around vents, pipes, ducts, chimneys, and fireplaces must be fireblocked to prevent fire spread.
- Exterior walls must be constructed to resist fire for at least one hour, with exposure from both sides.
- Batts or blankets of fiberglass, mineral wool, or other approved material must be secured between wall studs and partitions, at the ceiling and floor level, and at 10-ft. intervals both vertically and horizontally.
- Foam insulation installed in interior walls covered with ½" wallboard or other approved material must have a flame-spread rating of 75 or less and a smoke-developing index of 450 or less.
- Other insulation, including facings, vapor barriers, and breather papers, must have a flame-spread index of 25 or less and a smoke-developing index of 450 or less.
- Loose-fill insulation mounted with screens or supports must have a flame-spread rating of 25 or less and a smoke-developing index of 450 or less.

- Wall and ceiling finishes must have a flame-classification rating of 200 or less and a smoke-developing index of 450 or less.
- Smoke alarms must be installed in bedrooms, in hallways near bedrooms, and on each full story of a home. Multiple alarms must be wired together so one activation triggers all alarms.

Universal Design ▸

Universal design is intended for all people. While standard home and product designs are based on the "average" person—that is, the average adult male—not everyone fits into that category. Some people are short, some tall; some have difficulty walking, while others walk ably but find bending difficult. And physical abilities change constantly, as do family situations. By incorporating universal design into your remodeling plans, you can create spaces that work better for everyone who lives in or visits your home, regardless of their size, age, or ability.

Universal design is simply good design that improves everyday situations. For example, wide doorways make passage easier for a person carrying a load of laundry as well as for someone in a wheelchair; a lowered countertop enables a child to help prepare dinner and allows a person who tires easily to sit while cooking. More a way of thinking than a set of rules, universal design can be applied to any area of your home—from room layouts to light fixtures to door hardware. In all cases, universal design encourages independence by creating a safe, comfortable environment.

Many people take on remodeling projects to accommodate changes in their households. Perhaps you are remodeling because your aging parents are coming to live with you or your grown children or grandchildren are coming for an extended visit. Or you may be preparing your home for your own retirement years. Considering both your current and future needs is an essential part of a fundamental universal design concept: creating a lifespan home—one that accommodates its residents throughout their lives. A lifespan home enables your aging parents to live comfortably with you now and will allow you to stay in your home as you grow older. And, while universal design makes your everyday life easier, it will also make your home more appealing to a wide range of potential buyers if you choose to sell.

Much of the universal design information in this book comes from universal design specialists, kitchen and bath designers, physical therapists, specialty builders and manufacturers, and organizations such as the National Kitchen and Bath Association (NKBA). Some suggestions are ADA (Americans with Disabilities Act) requirements; while these generally apply to public spaces, they often are used as guidelines for residential design. As always, be sure that all aspects of your project meet local code requirements.

For more help with planning with universal design, contact a qualified professional. Many kitchen and bath designers, home builders, and product manufacturers specialize in universal design.

Working with Drawings

Drawings are necessary for any remodeling project that involves construction, enlargement, alteration, repair, demolition, or change to any major system within your home. There are two basic types of construction drawings: floor plans and elevation drawings.

Floor plans show a room as seen from above. These are useful for showing overall room dimensions, layouts, and the relationships between neighboring rooms. Elevation drawings show a side view of a room, showing one wall per drawing. Elevations are made for both the interior and exterior of a house and generally show more architectural detail than floor plans.

Both floor plans and elevation drawings provide you with a method for planning and recording structural and mechanical systems for your project. They also help the local building department to ensure your project meets code requirements.

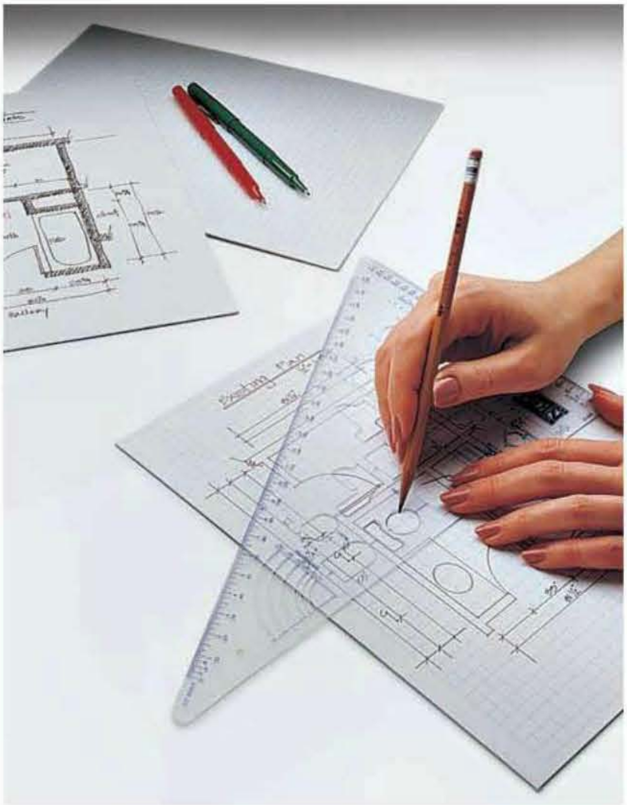

Use existing blueprints of your home, if available, to trace original floor plans and elevation drawings onto white paper. Copy the measurement scale of the original blueprints onto the traced drawings. Make photocopies of the traced drawings, then use the photocopies to experiment with remodeling ideas.

Before you draw up new plans, check with your home's architect, builder, or your local building department. These places often have copies of your home's floor plans on file. If your home is a historic building, your plans may be on file at a state or local historic office or university library.

If you are unable to obtain a copy of your home's floor plans, you can draw your own. This process provides you with a wealth of information about your home. Drawings let you see how changes will affect your home's overall layout and feel. They also help you plot your ideas, list materials, and solve design problems.

Follow the steps on page 21 to create floor plans and elevation drawings. Keep in mind that your plans may change as your ideas develop; until you have worked out all the elements of your design, consider your plans to be drafts. When you have arrived at a plan that meets your needs, draw up final floor and elevation plans to submit to your local building department.

To create floor plans, draw one story at a time. First, measure each room on the story from wall to wall. Transfer the rooms' dimensions to ¼" grid paper, using a scale of ¼" = 1 ft. Label each room for its use and note its overall dimensions. Include wall thicknesses, which you can determine by measuring the widths of door and window jambs—do not include the trim.

Next, add these elements to your drawings:

- Windows and doors; note which way the doors swing.
- Stairs and their direction as it relates to each story.
- Permanent features, such as plumbing fixtures, major appliances, countertops, built-in furniture, and fireplaces.
- Overhead features, such as exposed beams, or wall cabinets—use dashed lines.
- Plumbing, electrical, and HVAC elements. You may want a separate set of drawings for these mechanical elements and service lines.
- Overall dimensions measured from outside the house. Use these to check the accuracy of your interior dimensions.

To create elevation drawings, use the same ¼" = 1 ft. scale, and draw everything you see on one wall (each room has four elevations). Include:

- Ceiling heights and the heights of significant features such as soffits and exposed beams.
- Doors, including the heights (from the floor to the top of the opening) and widths.
- Windows, including the height of the sills and tops of the openings, and widths.
- Trim and other decorative elements.

When your initial floor plans and elevations are done, use them to sketch your remodeling layout options. Use overlays to show hidden elements or proposed changes to a plan. Photographs of your home's interior and exterior may also be helpful. Think creatively, and draw many different sketches; the more design options you consider, the better your final plans will be.

When you have completed your remodeling plans, draft your final drawings and create a materials list for the project.

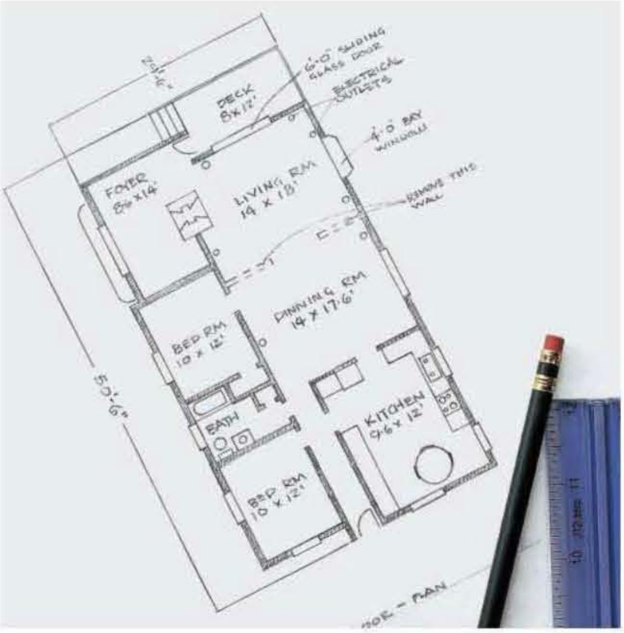
20'-6"
DECK 8x12'
6'-0" SLIDING GLASS DOOR
ELECTRICAL OUTLETS
4'-0" BAY WINDOW
FOYER 8½x14'
LIVING RM 14 x 18'
REMOVE THIS WALL
DINNING RM 14 x 17'6"
BED RM 10 x 12'
BATH
KITCHEN 9'6 x 12'
BED RM 10 x 12'
50'-6"
FLOOR - PLAN

Draft a detailed floor plan showing the layout of the area that will be remodeled, including accurate measurements. Show the location of new and existing doors and windows, wiring, and plumbing fixtures.

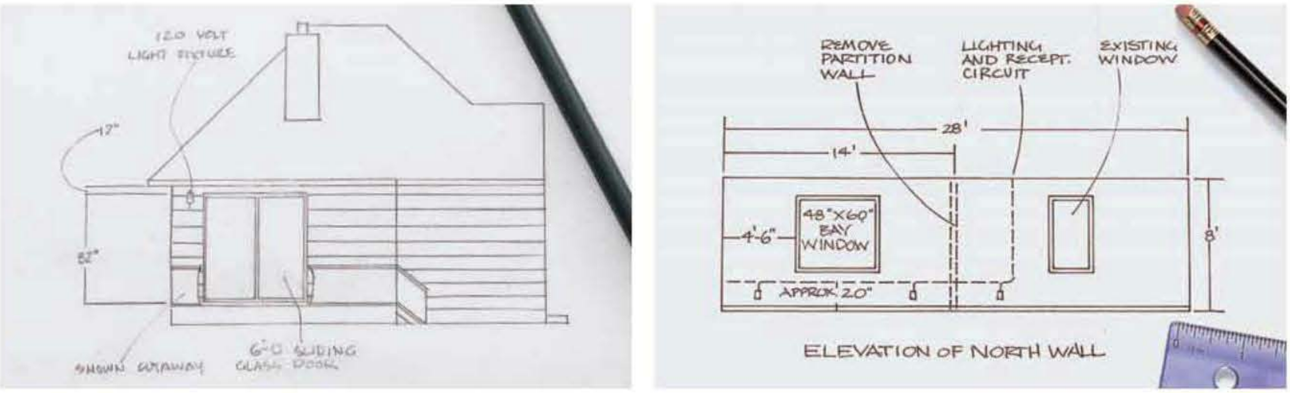
120 VOLT LIGHT FIXTURE
12"
82"
SHOWN CUTAWAY
6'-0 SLIDING GLASS DOOR
REMOVE PARTITION WALL
LIGHTING AND RECEPT. CIRCUIT
EXISTING WINDOW
28'
14'
48"X60" BAY WINDOW
4'-6"
8'
APPROX 20"
ELEVATION OF NORTH WALL

Create elevation drawings showing a side-view layout of windows and doors, as viewed from both inside and outside the home. Indicate the size of windows and doors, ceiling heights, and the location of wiring and plumbing fixtures.

Making Construction Plans

The best way to prepare for a remodel project is to create a construction plan. Having a complete construction plan enables you to view your entire project at a glance. It helps you identify potential problems, provides sense of the time involved, and establishes a logical order of steps. Without a construction plan, it's easy to make costly mistakes, such as closing up a wall with wallboard before the rough-ins have been inspected.

The general steps shown here follow a typical construction sequence. Your plan may differ at several points, but thinking through each of these steps will help you create a complete schedule.

1. CONTACT THE BUILDING DEPARTMENT

To avoid any unpleasant—and expensive—surprises, discuss you project with a building official. Find out about the building codes in your area and what you'll need to obtain the applicable permits. Explain how much of the work you plan to do yourself. In some states, plumbing, electrical, and HVAC work must be done by licensed professionals. Also determine what types of drawings you'll need to get permits and whether you'll need engineer's drawings and calculations.

2. CREATE YOUR DRAWINGS

Make your floor plans and elevation drawings (see pages 20 to 21). This step also involves most of the design work for your project; you may want to get help from a professional for this phase.

3. GET THE PERMITS

Have your final plans reviewed by the building inspector, and make any necessary adjustments required to obtain all of the permits for your project. This is also the time to schedule inspections. Find out what work must be inspected and when to call for inspections.

4. HIRE CONTRACTORS

If you're getting help with your project, it's best to find and hire the contractors early in the process, as their schedules will affect yours. It may be necessary for some contractors to obtain their own work permits from the building department. To avoid problems, make sure all of the contractors know exactly what work they are being hired to do and what work you will be doing yourself. Always check contractors' references and make sure any contractors are licensed and insured before hiring them. This is also the time to order materials and arrange for delivery.

5. COMPLETE THE FRAMING AND MAJOR MECHANICAL CHANGES

Begin the construction work with any major structural or mechanical changes. Move mechanical elements and reroute major service lines. Complete any rough-ins that must happen before the framing goes up, such as adding ducts, installing under-floor drains, or replacing old plumbing. Complete the new framing. Build the rough openings for windows and doors, and install the windows.

6. COMPLETE THE ROUGH-INS

Run drain, waste, and vent (DWV), water, and gas supply lines. Install electrical boxes, and run the wiring. Complete the HVAC rough-ins. Jot down measurements of pipes and locations of wiring, for future reference. Have the building inspector approve your work before you close up the walls. Install any fixtures that go in during the rough-in stage (others will come after the wall surfaces are installed).

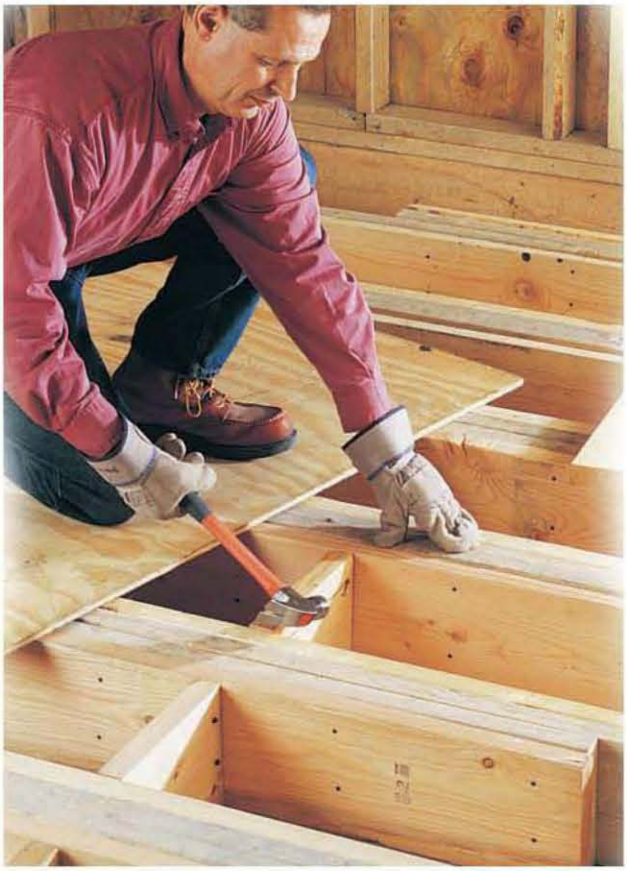

7. FINISH THE WALLS AND CEILINGS

After your work has passed inspection, insulate the walls, ceilings, and pipes. Install fiberglass insulation used as fireblocking. Make sure protector plates for pipes and wires running through framing are in place. Add vapor barriers as required by local code.

Make sure everything is in place before you cover up the framing, then finish the walls and ceilings. If you're installing wallboard, do the ceilings first, then the walls. Tape and finish the wallboard. Install other finish treatments. Texture, prime, and paint the wallboard when it's convenient. If you are installing a suspended ceiling, do so after you finish the walls.

8. ADD THE FINISHING TOUCHES

Install doors, moldings, woodwork, cabinets, and built-in shelving, and lay the floor coverings. The best order for these tasks will depend on the materials you're using and the desired decorative effects.

Install any new plumbing fixtures you have chosen for bathrooms, and complete the drain and supply hookups. Make electrical connections, and install all fixtures, devices, and appliances. Get a final inspection from the building inspector.

Joint Compound

Working Safely

Your personal safety when working on carpentry projects depends greatly on what safety measures you take. The power tools sold today offer many safety features, such as blade guards, locks to prevent accidental starts, and double insulation to reduce the risk of shock in the event of a short circuit. It's up to you to take advantage of these safety features. For example, never operate a saw with the blade guard removed. You risk injury from flying debris as well as from being cut by the blade.

Follow all precautions outlined in the owner's manuals for your tools and make sure you protect yourself with safety glasses, earplugs, and a dust mask or respirator to filter out dust and debris.

Keep your work environment clean. A cluttered work area is more likely to result in accidents. Clean your tools and put them away at the end of every work period, and sweep up dust and debris.

Some materials emit dangerous fumes or particles. Keep such materials stored away from heat sources and out of the reach of children; always use these products in a well-ventilated area.

Maintaining safety is an ongoing project. Take the time to update your first-aid kit and evaluate your workspace, tools, and safety equipment on a regular basis. To avoid accidents, repair and replace old and worn-out parts before they break.

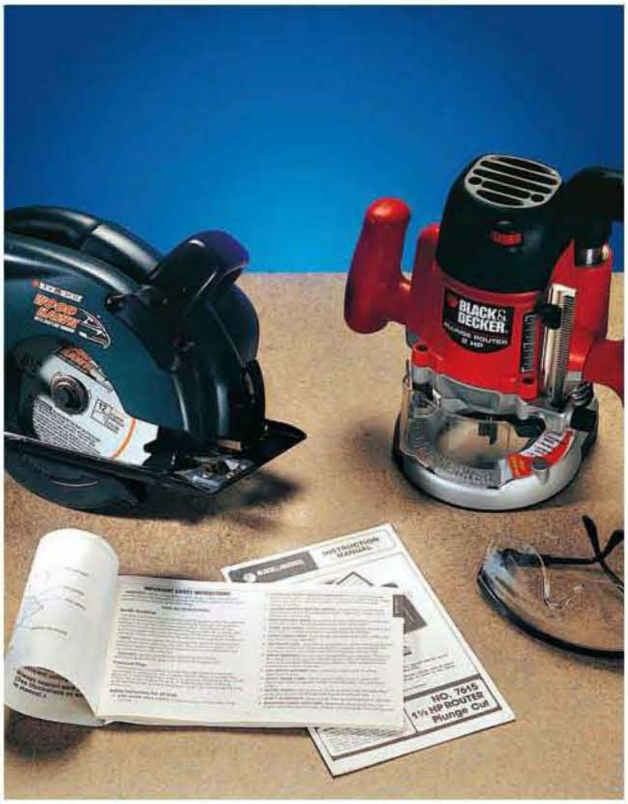

Read the owner's manual before operating any power tool. Your tools may differ in many ways from those described in this book, so it's best to familiarize yourself with the features and capabilities of the tools you own. Always wear eye and ear protection when operating a power tool. Wear a dust mask when the project will produce dust.

Some walls may contain asbestos. Many homes built or remodeled between 1930 and 1950 have older varieties of insulation that included asbestos. Consult a professional for removal of hazardous pollutants like asbestos, and if you find asbestos or materials that may contain asbestos, do not attempt to remove them on your own. Even if you determine that no asbestos is present, it is a good idea to wear a particle mask and other safety gear when doing demolition.

First-Aid Kits

Assemble a first-aid kit. Cuts from hand or power tools can be serious and require prompt and thoughtful attention. Be prepared for such situations with a well-equipped first-aid kit that is easy to find. Record any emergency telephone numbers on the first-aid kit or by the nearest phone so they are available in an emergency.

Equip your kit with a variety of items (photo right), including bandages, needles, tweezers, antiseptic ointment, cotton swabs, cotton balls, eye drops, a first-aid handbook, a chemical-filled cold pack, elastic bandages, first-aid tape, and sterile gauze.

For puncture wounds, cuts, burns, and other serious injuries, always seek medical attention as soon as first aid—such as washing and wrapping of cuts—has been provided.

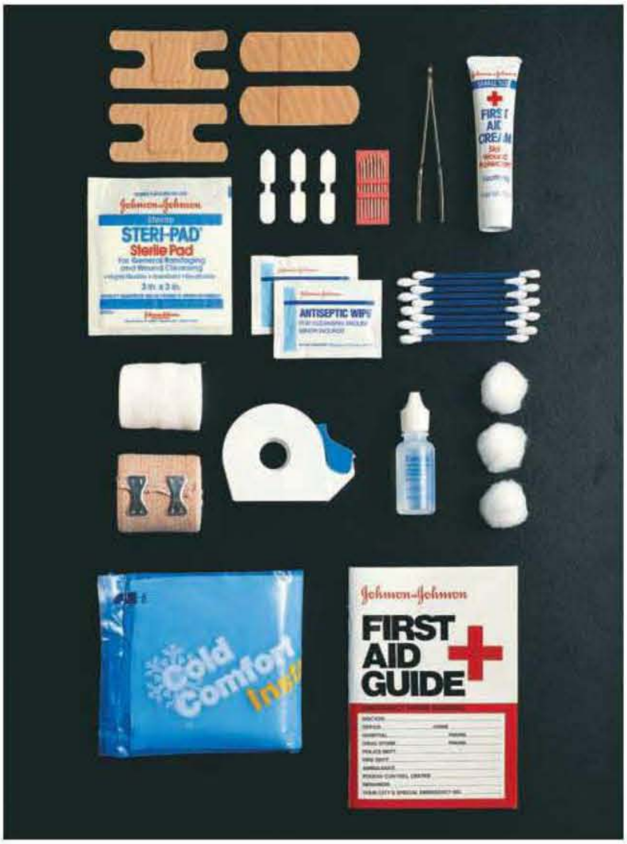
FIRST AID CREAM
Johnson & Johnson
STERI-PAD
Sterile Pad
3 in x 3 in
ANTISEPTIC WIPE
Cold Comfort Instant
Johnson & Johnson
FIRST AID GUIDE

Safe Practices

Keep your tools sharp and clean. Accidents are more likely when blades are dull and tools are filled with sawdust and dirt.

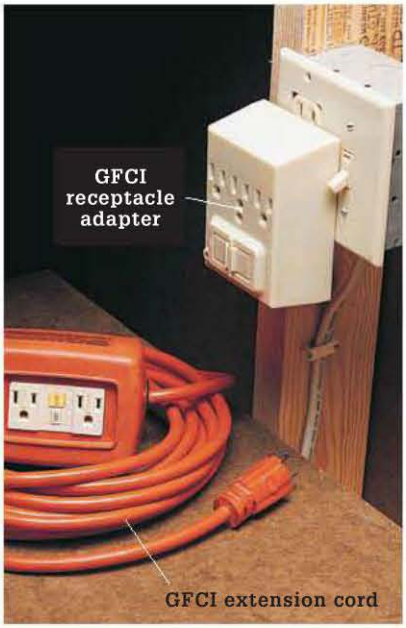
GFCI receptacle adapter
GFCI extension cord

Use a GFCI receptacle, adapter, or extension cord to reduce the risk of shock while operating a power tool outdoors or in wet conditions.

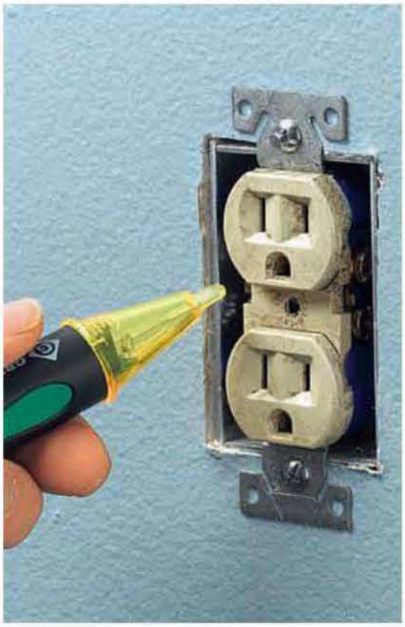

Check with a neon circuit tester to make sure the power is off before removing cover plates, exposing wires, or drilling or cutting into walls that contain wiring.

Techniques

While every home improvement project is different from the others in some way, the basic skills you need come from a common set of DIY techniques. Building walls, running plumbing lines, and adding electrical circuits are some of the chief techniques you'll need to call upon. In this chapter you'll find a condensed version of the essential information you need to put these valuable techniques into practice and become an accomplished do-it-yourselfer.

Once the new walls are up (but before they're covered) you need to fill them with pipes and wires. In the plumbing section that follows you'll learn how to plan new lines and hookups as well as how to work with basic plumbing materials. In the wiring section you'll find useful instructions on working with electrical cables and devices. A handy series of circuit maps shows you exactly how your remodeled room should be wired.

In this chapter:

- Demolition
- Building Walls
- Installing Wallboard
- Understanding Plumbing
- Working with Copper Pipe
- Working with Plastic Pipe
- Working with PEX Pipe
- Planning New Plumbing
- Understanding Wiring
- Working with Wire & Cable
- Working with Conduit
- Working with Boxes
- Working with Switches
- Working with Receptacles
- Planning New Wiring
- Using Circuit Maps

Demolition

Many home remodeling projects actually begin with demolition. When you're remodeling, it's often necessary to cut or enlarge openings for new doors or windows or even to remove entire walls. The basic procedures for this type of demolition are the same whether you're working with doors and windows on exterior walls or altering interior walls.

Your first step will be to determine how your house was framed. House framing variations will dictate the proper procedures for creating openings in walls or removing walls altogether. Then, you'll need to inspect the walls for hidden mechanicals—wiring, plumbing, and HVAC lines.

After you've rerouted any utility lines, you're ready to remove the interior wall surfaces. If you're replacing old windows and doors, now is the time to remove them as well. Where necessary, you can now remove exterior wall surfaces, but don't remove any framing members yet.

The next step will depend on the nature of your project. If you are removing a load-bearing wall or creating a new or enlarged opening in one, you'll need to build temporary supports to brace the ceiling while the work is being done. This step won't be necessary if you are removing a non-loadbearing wall. Then you can remove any wall framing members, following the applicable procedures for load-bearing or non-loadbearing walls.

With the removal steps of the project completed, you'll be ready to install your new windows and doors.

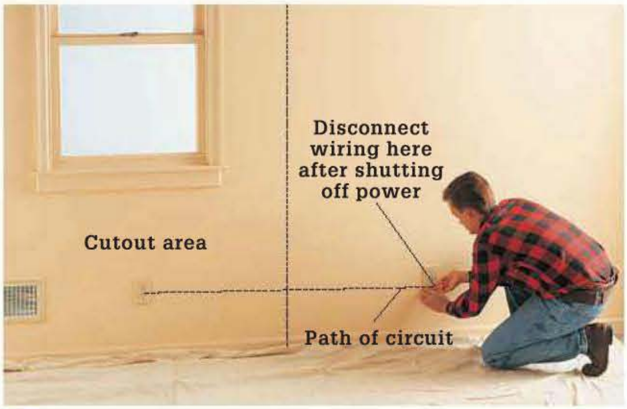
Disconnect wiring here after shutting off power
Cutout area
Path of circuit

Disconnect electrical wiring before you cut into walls. Trace the wiring back to a fixture outside the cutout area, then shut off the power and disconnect the wires leading into the cutout area. Turn the power back on and check for current with a circuit tester before cutting into the walls.

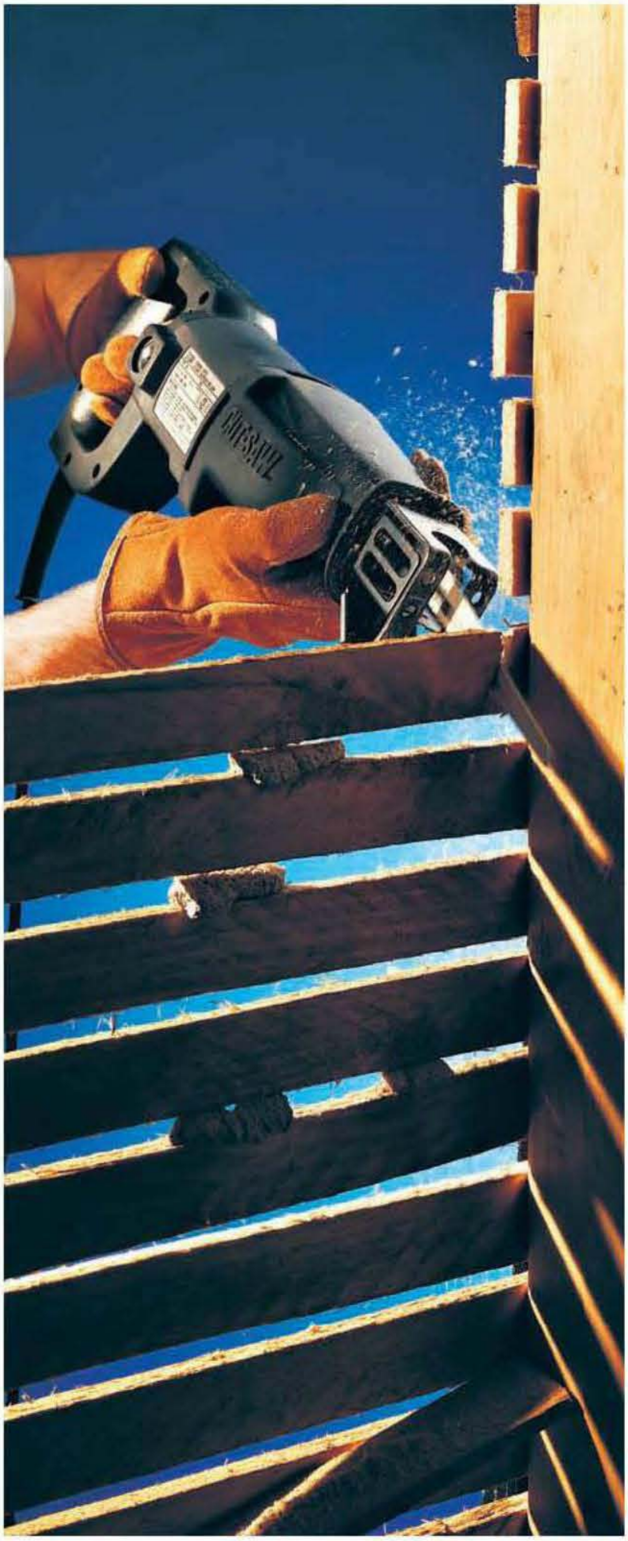

Demolition is the starting point of most projects where a window or door opening is enlarged.

How to Install Temporary Support (perpendicular joists)

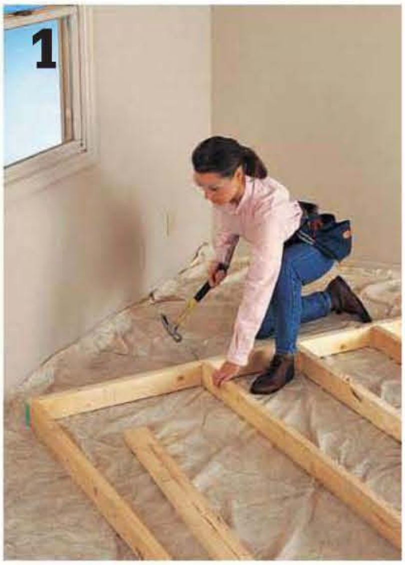

Build a 2 × 4 stud wall that is 4 ft. wider than the planned wall opening and 1¾" shorter than the distance from floor to ceiling.

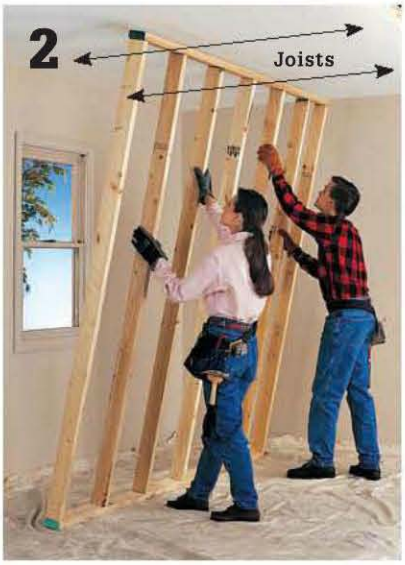

Raise the stud wall up and position it 3 ft. from the wall, centered on the planned rough opening.

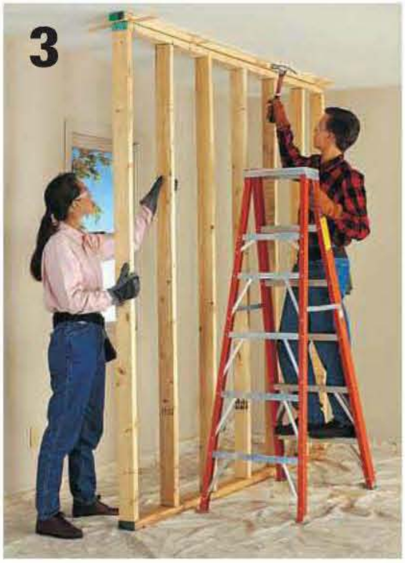

Slide a 2 × 4 top plate between the temporary wall and the ceiling. Check to make sure the wall is plumb, and drive shims under the top plate at 12" intervals until the wall is wedged tightly in place.

How to Install Temporary Support (parallel joists)

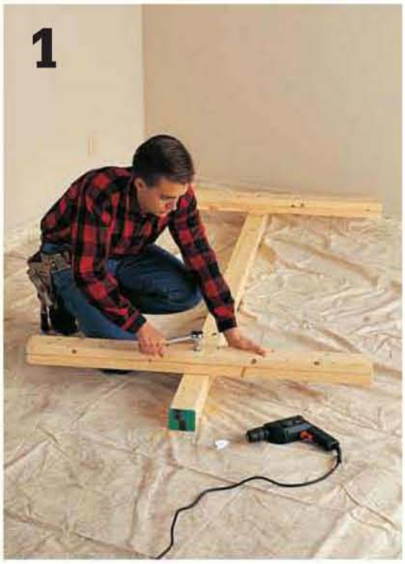

Build two 4-ft.-long cross braces, using pairs of 2 × 4s nailed together. Attach the cross braces to the double top plate, 1 ft. from the ends, using countersunk lag screws.

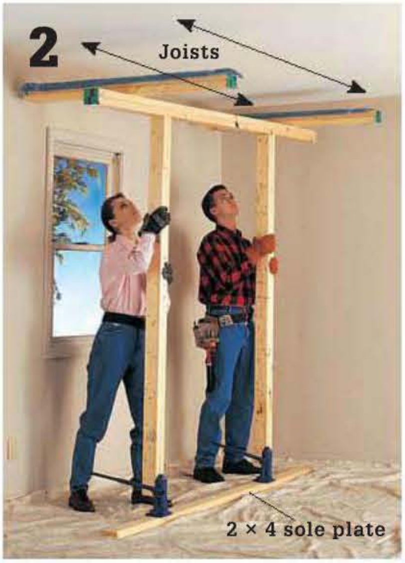

Place a 2 × 4 sole plate directly over a floor joist, then set hydraulic jacks on the sole plate. For each jack, build a post 8" shorter than the jack-to-ceiling distance. Nail the posts to the top plate, 2 ft. from the ends. Cover the braces with the cloth, and set the support structure on the jacks.

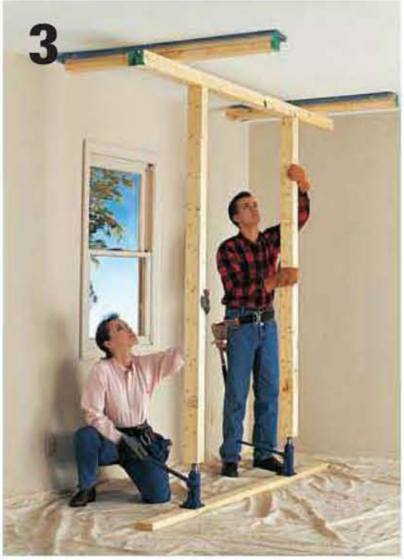

Adjust the support structure so the posts are exactly plumb, and pump the hydraulic jacks until the cross braces just begin to lift the ceiling. Do not lift too far or you may damage the ceiling or floor.

How to Remove Wallboard

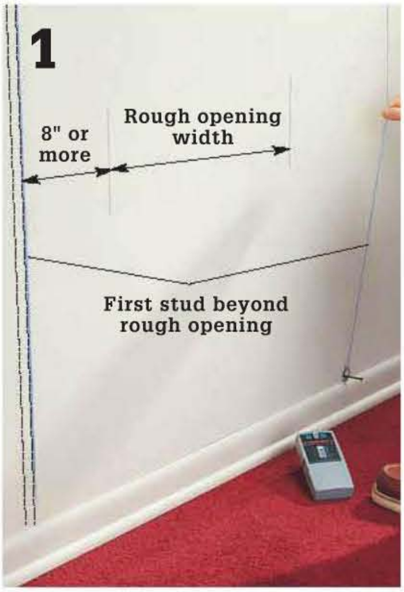

Mark the width of the rough opening on the wall and locate the first stud on either side of the planned rough opening. If the rough opening is more than 8" from the next stud, use a chalk line to mark a cutting line on the inside edge of the stud.

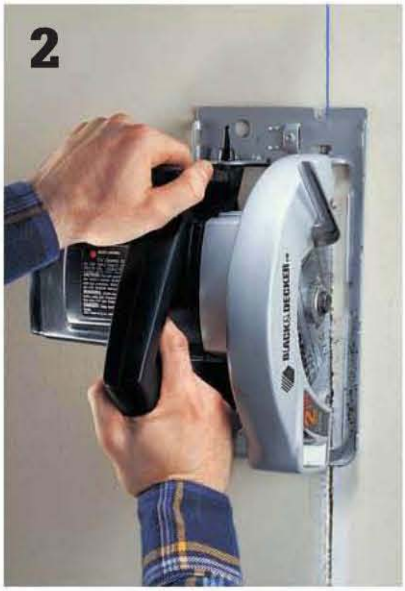

Remove the baseboards and other trim, and prepare the work area. Make a ¾"-deep cut from floor to ceiling along both cutting lines, using a circular saw. Use a utility knife to finish the cuts at the top and bottom and to cut through the taped horizontal seam where the wall meets the ceiling surface.

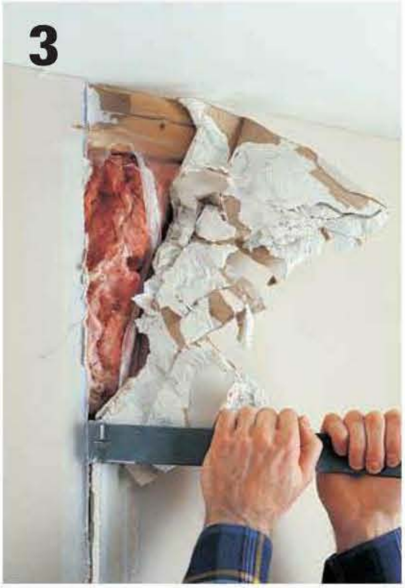

Insert the end of a pry bar into the cut near a corner of the opening. Pull the pry bar until the wallboard breaks, then tear away the broken pieces. Take care to avoid damaging the wallboard outside the project area.

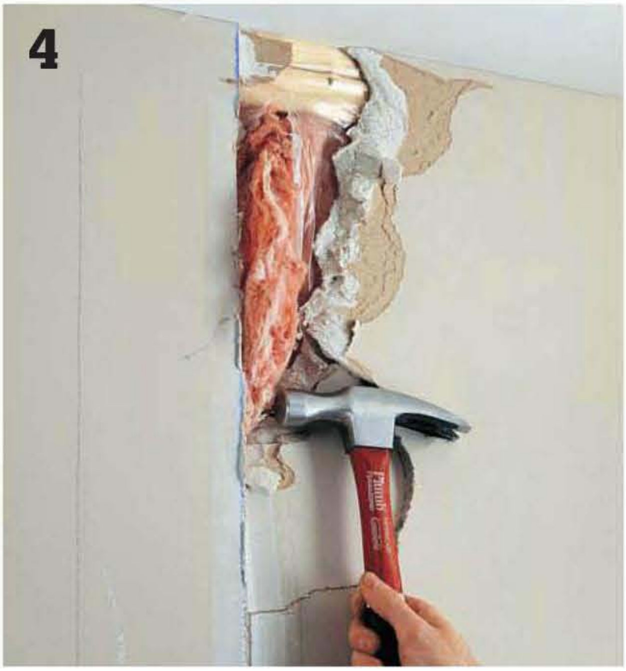

Continue removing the wallboard by striking the surface with the side of a hammer and pulling the wallboard away from the wall with the pry bar or your hands.

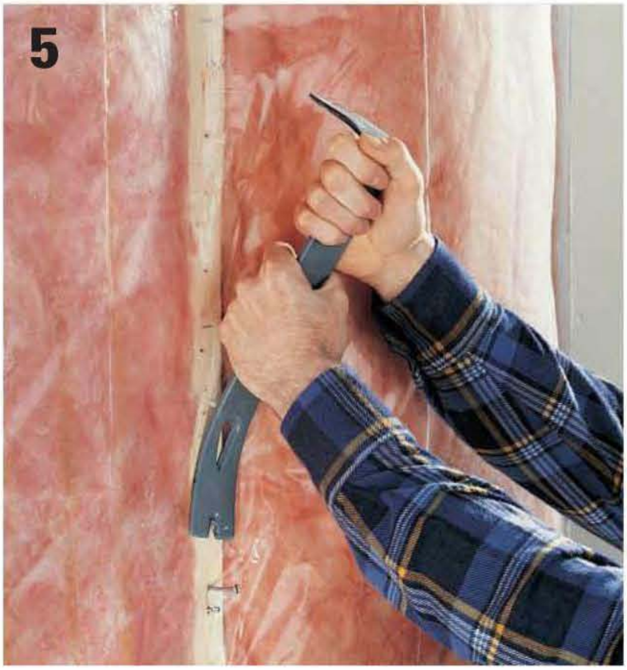

Remove nails, screws, and any remaining wallboard from the framing members, using a pry bar. Remove any vapor barrier and insulation.

How to Remove Plaster Walls

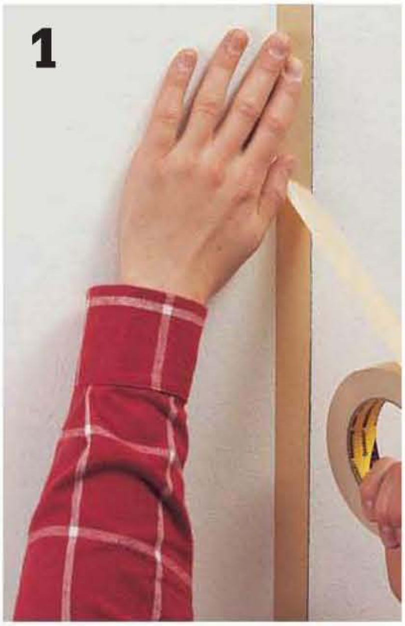

Shut off the power and inspect the wall for wiring and plumbing. Mark the wall area to be removed by following the directions on page 30. Apply a double layer of masking tape along the outside edge of each cutting line.

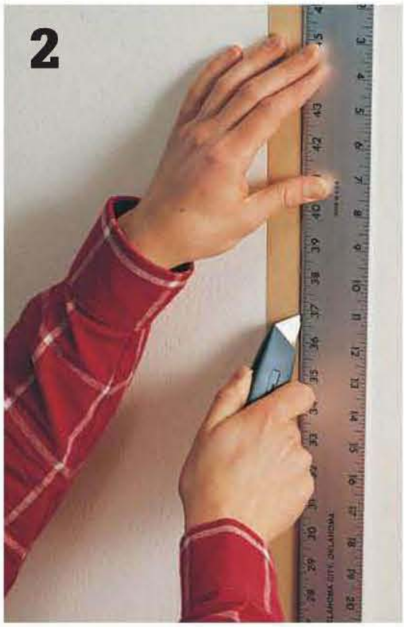

Score each line several times with a utility knife, using a straightedge as a guide. Scored lines should be at least ⅛" deep.

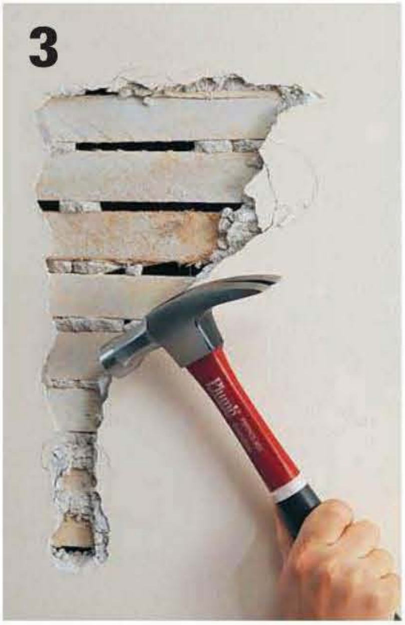

Beginning at the top of the wall in the center of the planned opening, break up the plaster by striking the wall lightly with the side of a hammer. Clear away all plaster from floor to ceiling to within 3" of the marked lines.

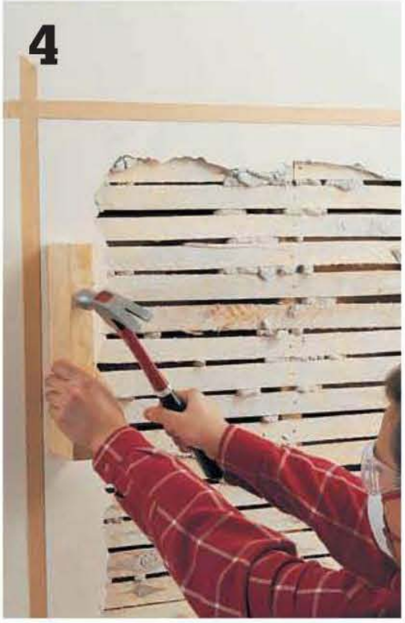

Break the plaster along the edges by holding a scrap piece of 2 × 4 on edge just inside the scored line and rapping it with a hammer. Use a pry bar to remove the remaining plaster.

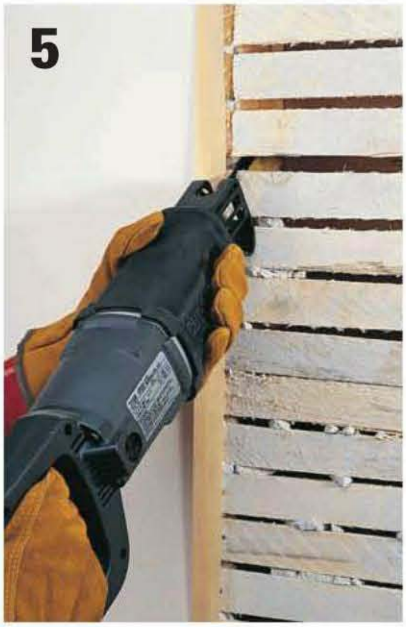

Cut through the lath along the edges of the plaster, using a reciprocating saw or jigsaw.

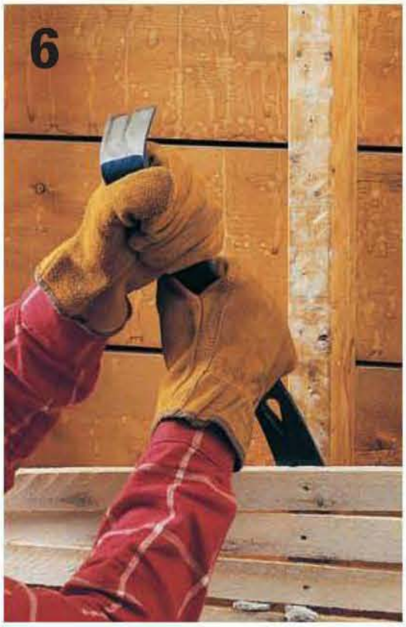

Remove the lath from the studs, using a pry bar. Pry away any remaining nails, and remove any vapor barrier and insulation.

How to Make an Opening in an Exterior Wall

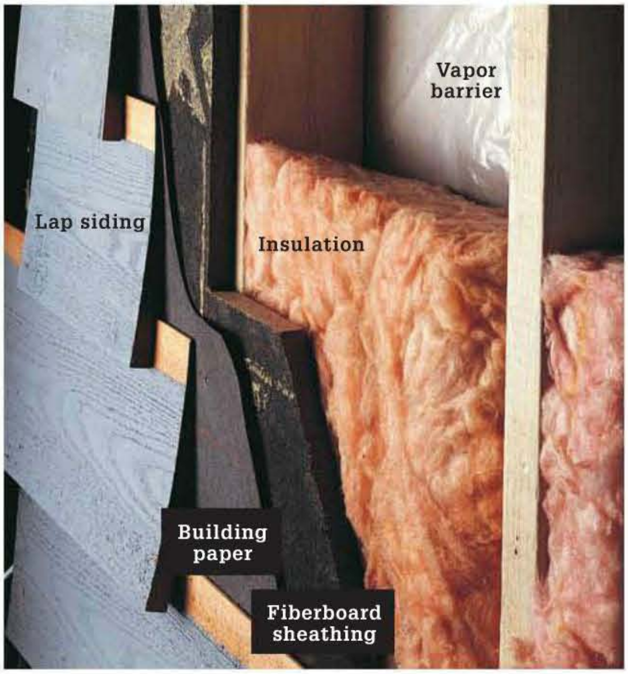

A cross-section view of a typical exterior wall with lap siding.

From inside the house, drill through the wall at the corners of the framed opening. Push casing nails through the holes to mark their location. For round-top windows, drill holes around the curved outline (see variation, page 195).

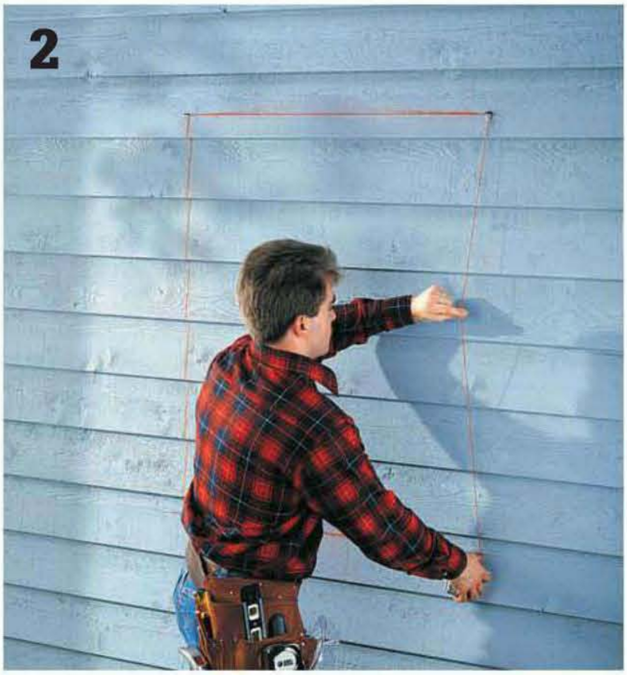

Measure the distance between the nails on the outside of the house to make sure the dimensions are accurate. Mark the cutting lines with a chalk line stretched between the nails. Push the nails back through the wall.

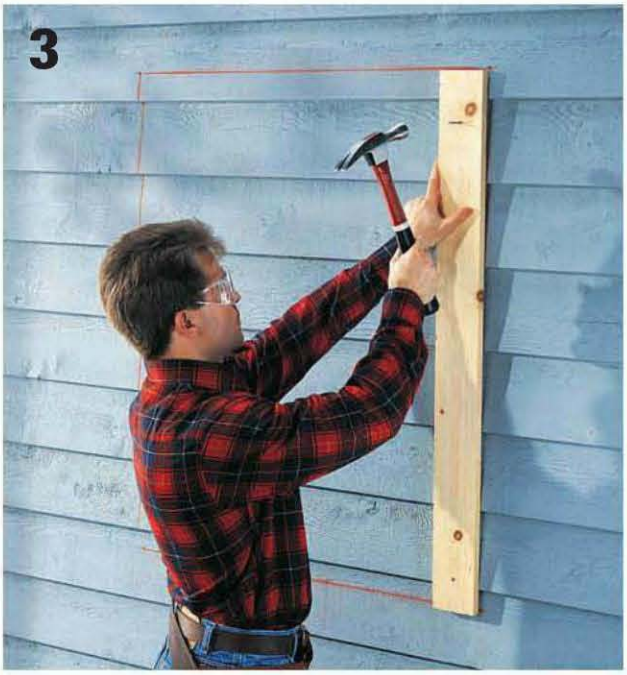

Nail a straight 1 × 4 flush with the inside edge of the right cutting line. Sink the nail heads with a nail set to prevent scratches to the foot of the saw. Set the circular saw to its maximum blade depth.

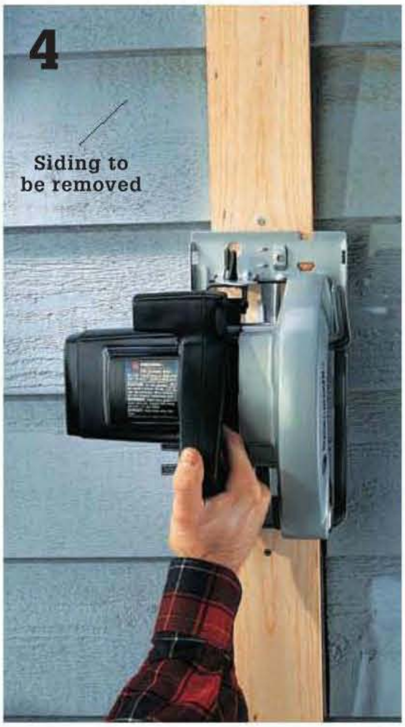

Rest the saw on the 1 × 4, and cut along the marked line, using the edge of the board as a guide. Stop the cuts about 1" short of the corners to keep from damaging the framing members.

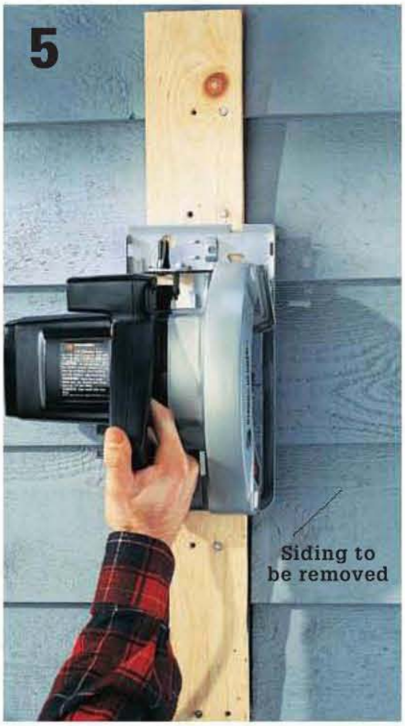

Reposition the 1 × 4, and make the remaining straight cuts. Drive nails within 1½" of the inside edge of the board, because the siding under this area will be removed to make room for door or window brick moldings.

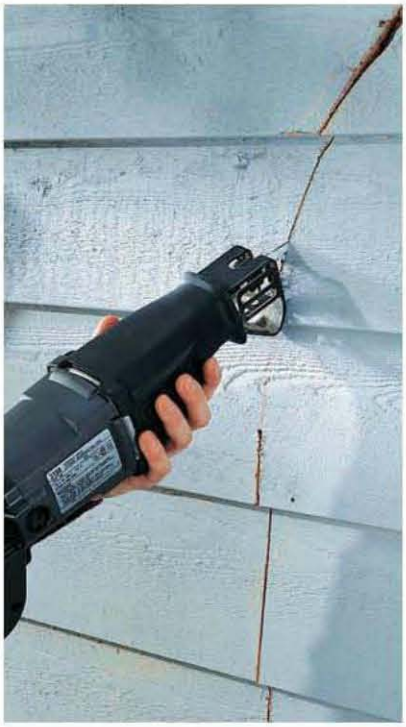

Variation: For round-top windows, make curved cuts using a reciprocating saw or jigsaw. Move the saw slowly to ensure smooth, straight cuts. To draw an outline for round-top windows, use a cardboard template.

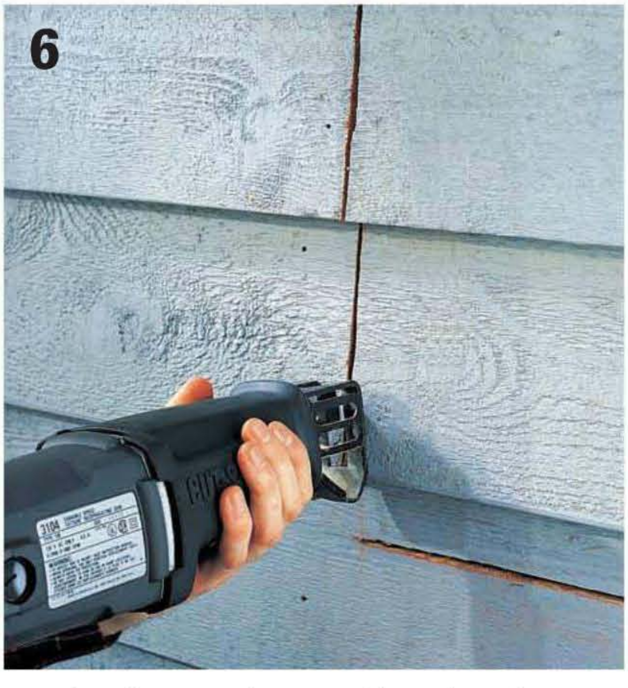

Complete the cuts at the corner with a reciprocating saw or jigsaw.

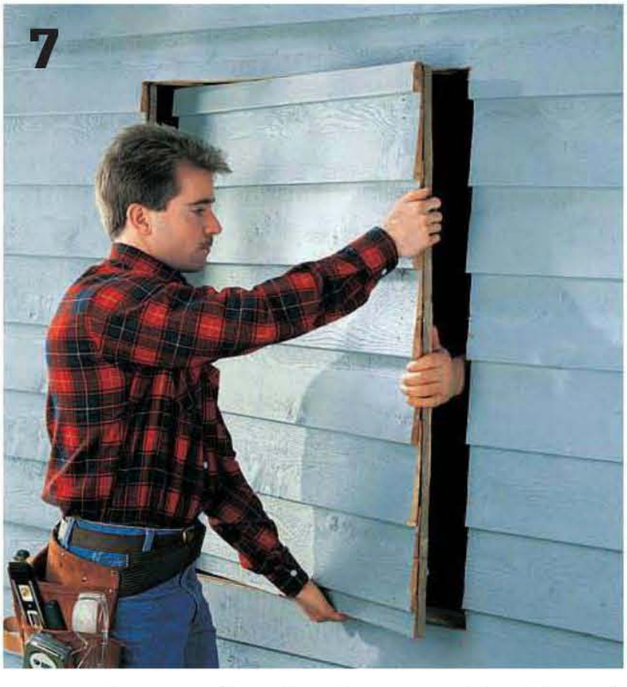

Remove the cut wall section. If you are working with metal siding, wear work gloves. If you wish, remove the siding pieces from the sheathing and save them for future use.

How to Make an Opening in a Stucco Wall

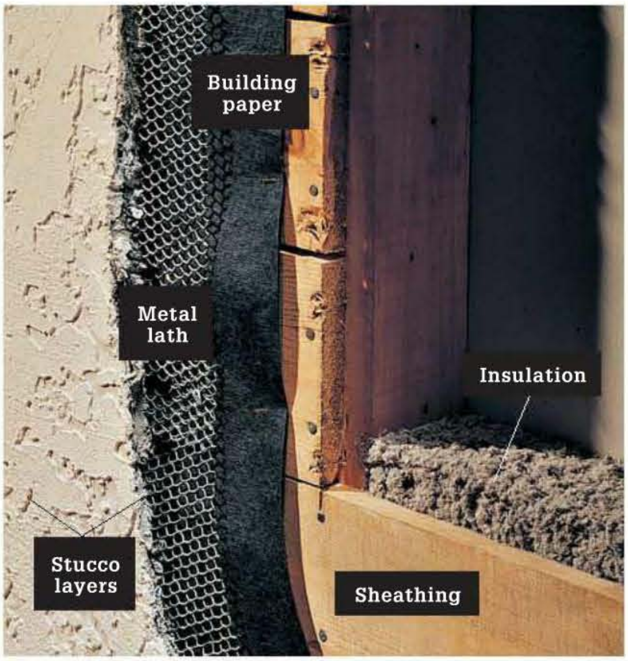

Stucco is a multiple-layer cement product applied to metal lath. Building paper is sandwiched between the metal lath and the sheathing to create a waterproof barrier. Stucco is extremely durable due to its cement base. But if you don't do the removal carefully, it's easy to crack the stucco past the outline for the new window or door.

From inside the house, drill through the wall at the corners of the framed opening. Use a twist bit to drill through the sheathing, then use a masonry bit to finish the holes. Push casing nails through the holes to mark their locations.

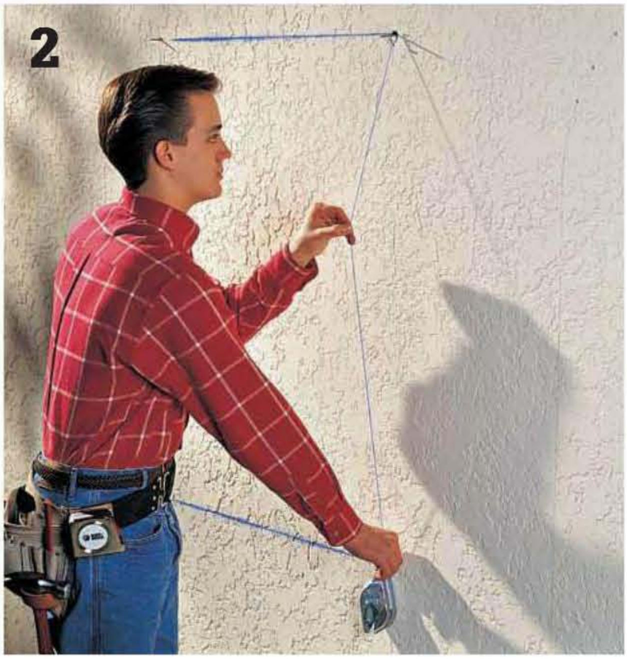

On the outside wall, measure the distance between the nails to make sure the rough opening dimensions are accurate. Mark cutting lines between the nails, using a chalk line.

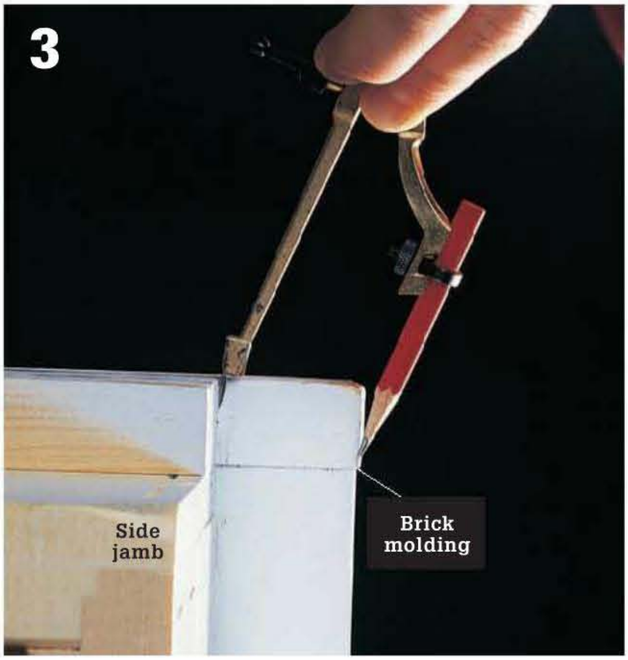

Match the distance between the side jambs and the edge of the brick molding on a window or door with the legs of a compass.

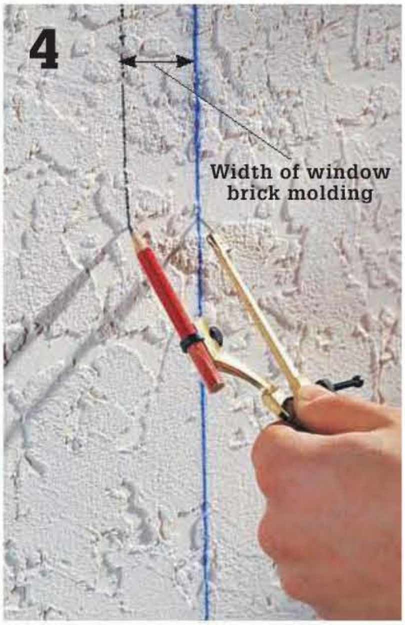

Scribe a cutting line on the stucco by moving the compass along the outline, with the compass point held on the marked line. This added margin will allow the brick molding to fit tight against the wall sheathing.

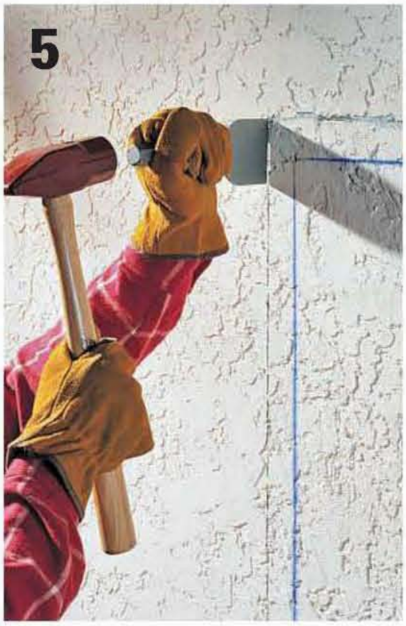

Score the stucco surface around the outside edge of the scribed line, using a masonry chisel and masonry hammer. The scored grooves should be at least ⅛" deep to serve as a guide for the circular saw blade.

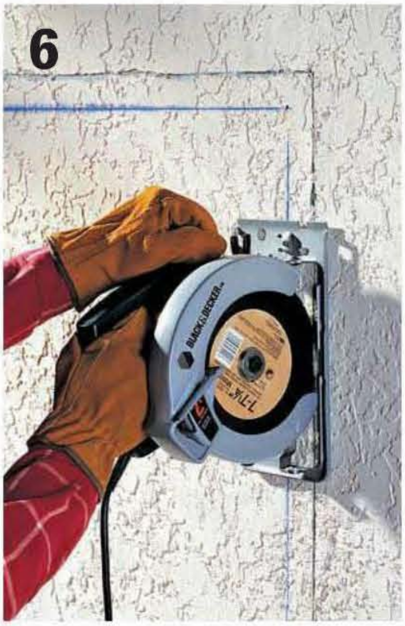

Make straight cuts using a circular saw and masonry-cutting blade. Make several passes with the saw, gradually deepening the cuts until the blade just cuts through the metal lath, causing sparks to fly. Stop cuts just ahead of the corners to avoid damaging the stucco past the cutting line; complete the cuts with a masonry chisel.

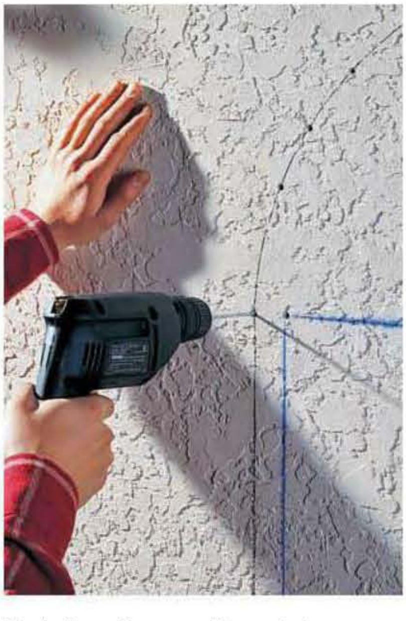

Variation: For round-top windows, mark the outline on the stucco, using a cardboard template, and drill a series of holes around the outline, using a masonry bit. Complete the cut with a masonry chisel.

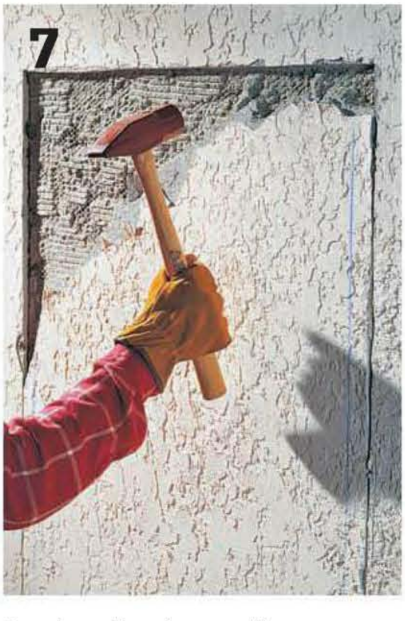

Break up the stucco with a masonry hammer or sledgehammer, exposing the underlying metal lath. Use aviation snips to cut through the lath around the opening. Use a pry bar to pull away the lath and attached stucco.

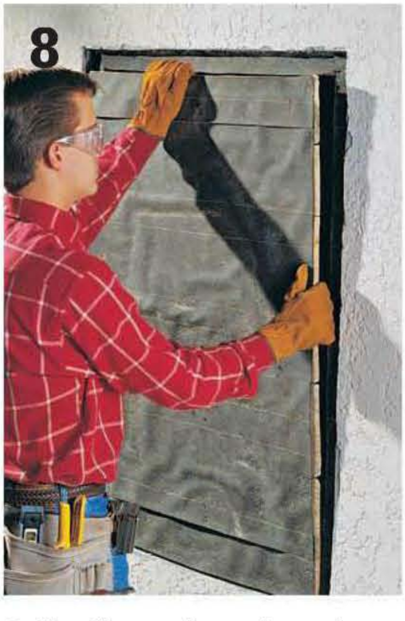

Outline the rough opening on the sheathing, using a straightedge as a guide. Cut the rough opening along the inside edge of the framing members, using a circular saw or reciprocating saw. Remove the cut section of sheathing.

How to Remove a Partition Wall

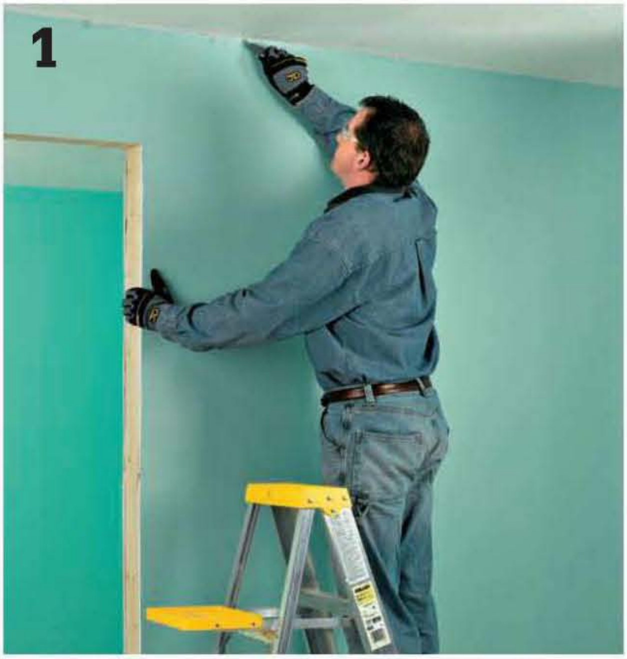

Use a utility knife to score the intersections where the wall you're removing meets the ceiling to keep from damaging it during wall removal. Pry away baseboard trim and remove receptacle plates and switch covers to prepare for demolition.

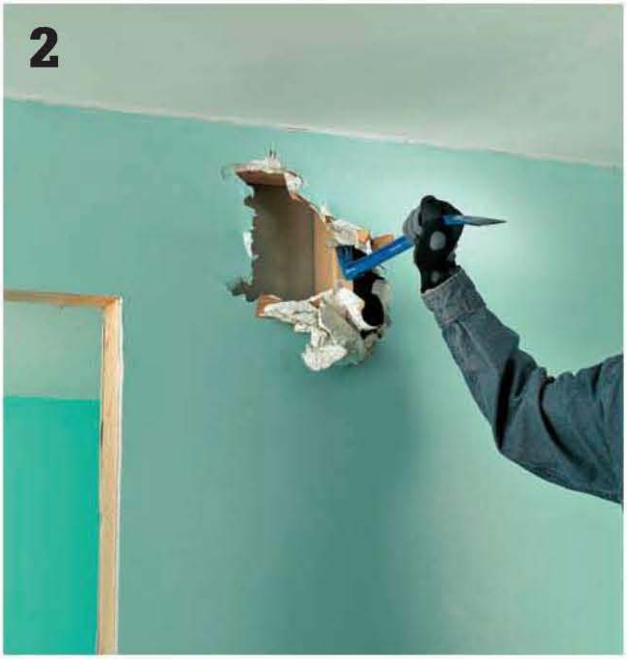

Use the side of a hammer to punch a starter hole in the wallboard, then carefully remove the wallboard with a pry bar. Try to pull off large sections at a time to minimize dust. Remove any remaining wallboard nails or screws from the wall studs.

Reroute outlets, switches, plumbing, or ductwork. Have professionals do this for you if you are not experienced with these systems or confident in your skills. This work should be inspected after it is completed.

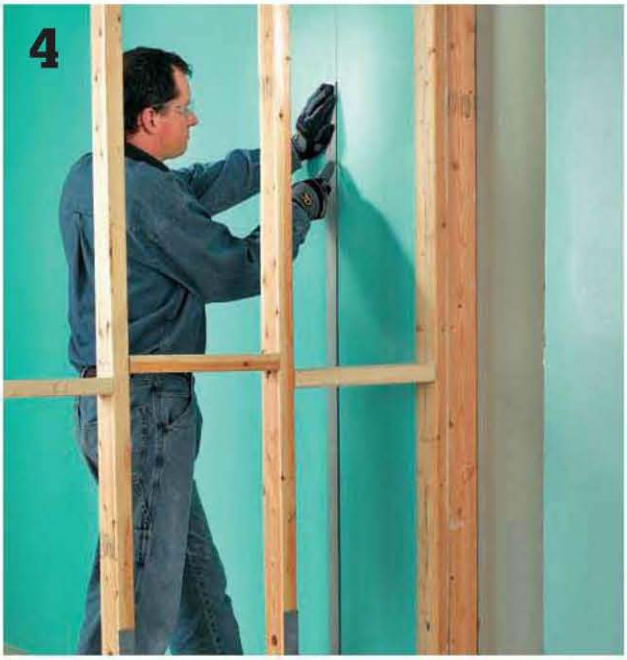

Locate the closest permanent studs on the adjacent wall or walls with a stud finder, and carefully remove the wallboard up to these studs. Score the wallboard first with a utility knife, then cut through it with a circular saw.

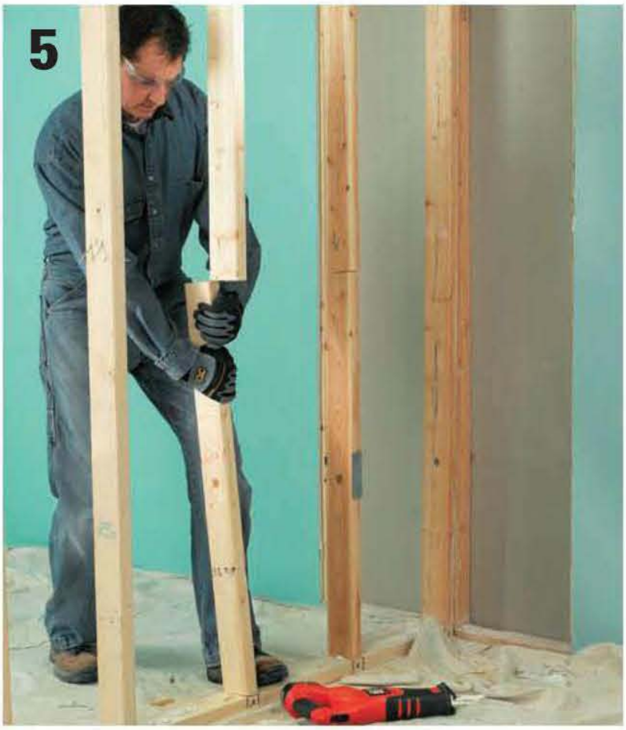

Remove the wall studs by cutting through them in the middle with a reciprocating saw and prying out the upper and lower sections. Remove the endmost studs where the wall meets an adjacent wall or walls.

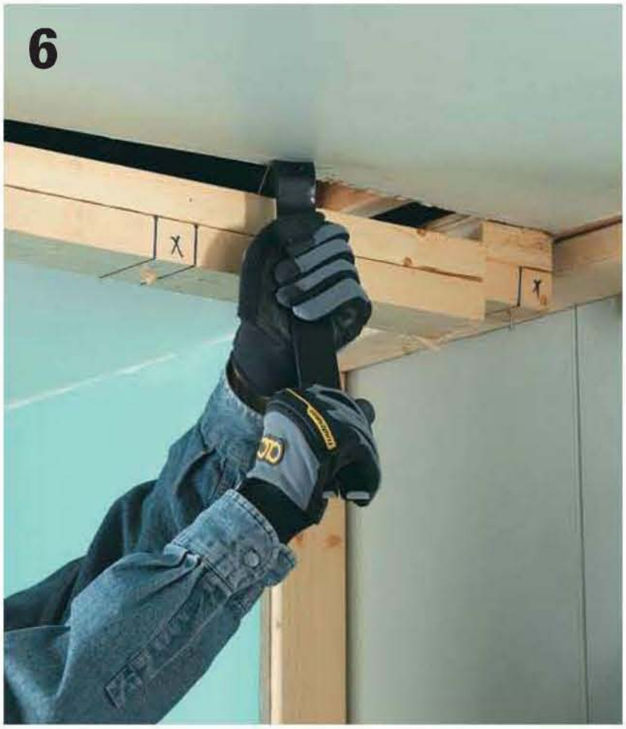

Cut through the wall's top plate with a circular saw or reciprocating saw. Pry out the top plate sections carefully to avoid damaging the ceiling.

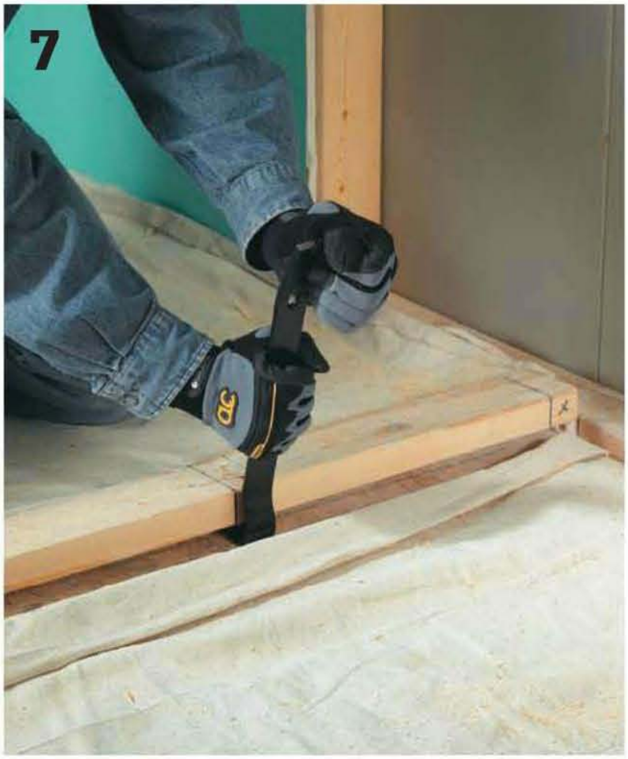

Remove the sole plate just as you did the top plate by cutting through it and prying up the long pieces.

Patch the walls and ceiling with strips of wallboard, and repair the floor as needed with new floor coverings.

Removing Trim

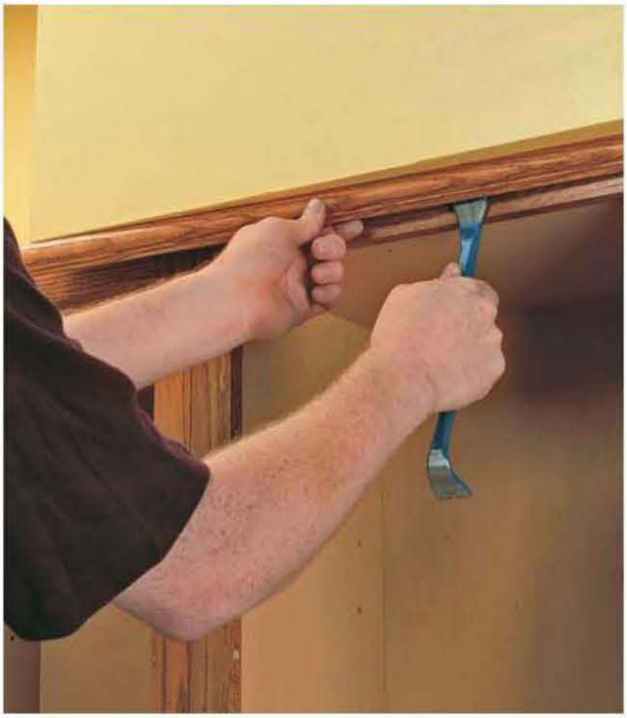

Remove trim moldings at the edges and tops of the cabinets with a flat pry bar or putty knife.

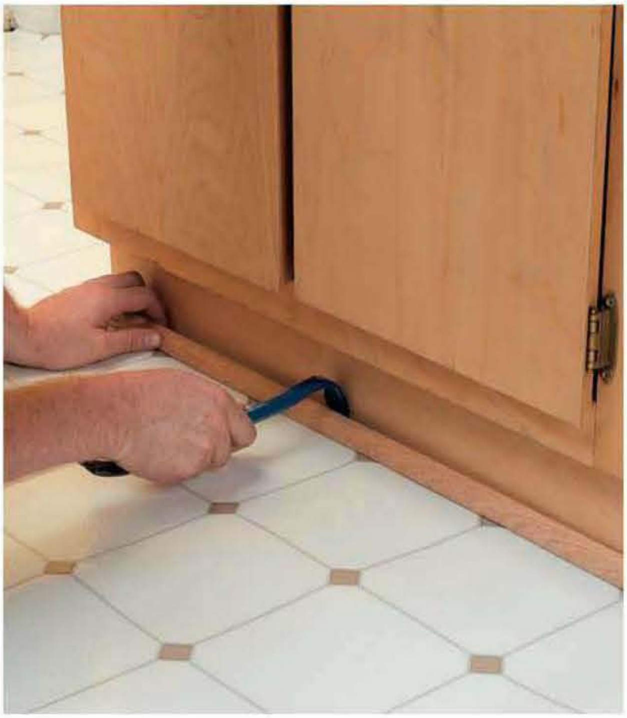

Remove base shoe from cabinet base if the molding is attached to the floor.

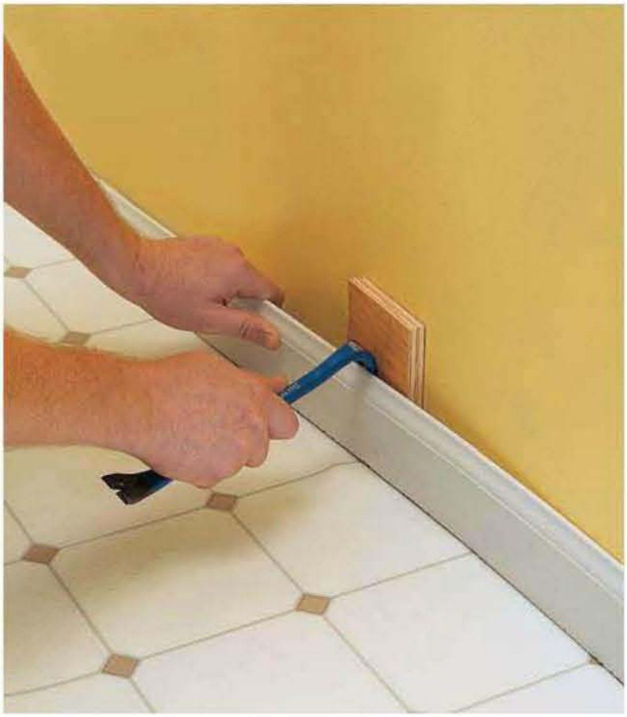

Remove baseboards and other trim moldings with a pry bar. Protect wall surfaces with scraps of wood.Label the trim boards on the back side so you can replace them correctly.

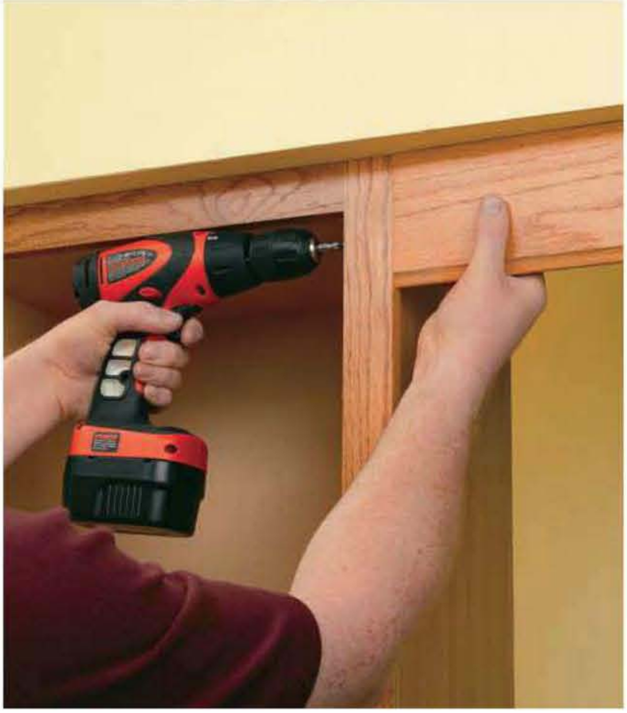

Remove valances above cabinets. Some valances are attached to the cabinets or soffits with screws. Others are nailed and must be pried loose.

How to Remove Cabinets

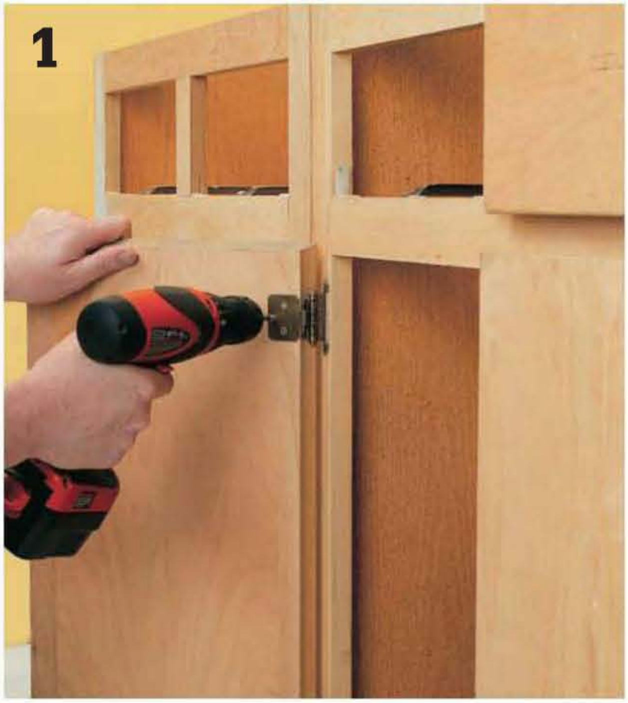

Remove doors and drawers to make it easier to get at interior spaces. You may need to scrape away old paint to expose hinge screws.

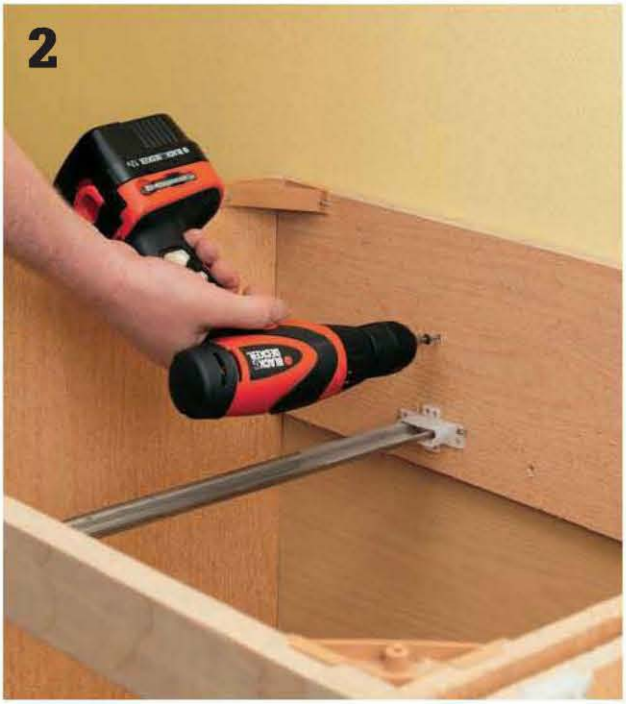

At the backs of cabinets, remove any screws holding the cabinet to the wall. Cabinets can be removed as a group, or can be disassembled.

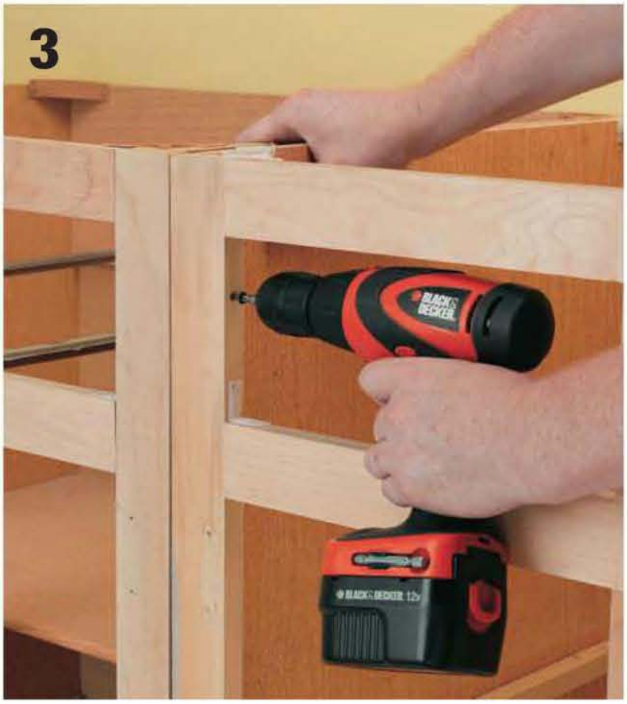

Detach individual cabinets by removing screws that hold face frames together.

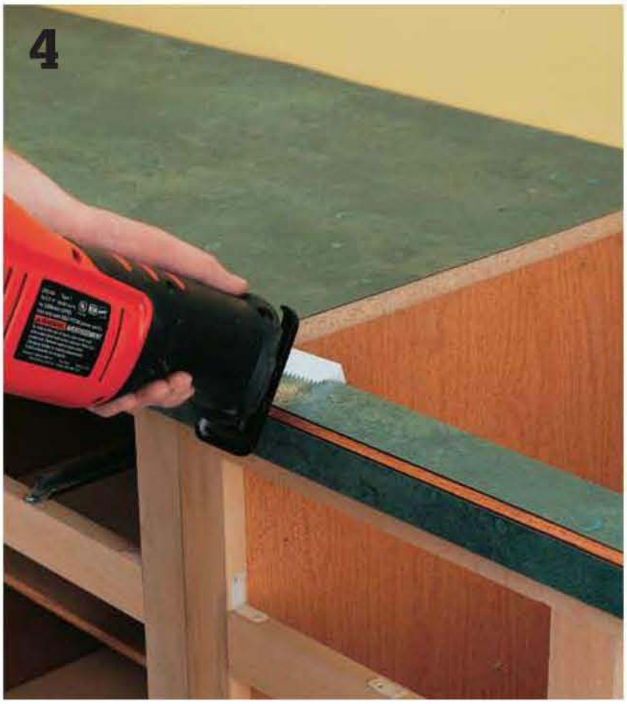

Countertops are usually not salvageable. Cut them into manageable pieces with a reciprocating saw, or take them apart piece by piece with a hammer and pry bar.

Building Walls

Partition walls are constructed between load-bearing walls to divide space. They should be strong and well made, but their main job is to house doors and to support wall coverings. The sole plate at the bottom of the wall should be fastened securely to the subfloor and joists with panel adhesive and 16d common nails. Use masonry fasteners on concrete basement floors.

Anchoring New Partition Walls

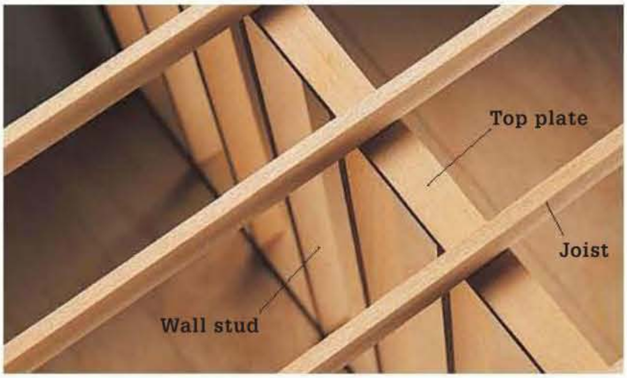

When a new wall is perpendicular to the ceiling or floor joists above, attach the top plate directly to the joists, using 16d nails.

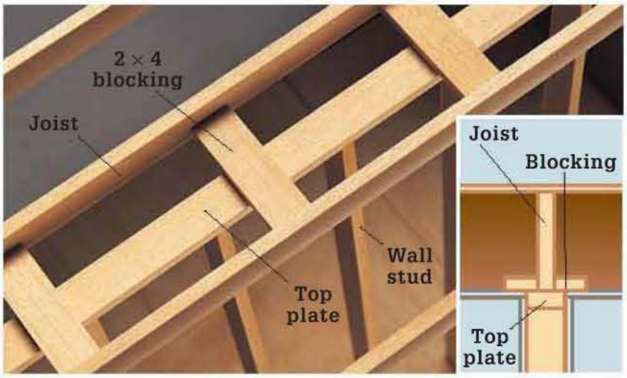

When a new wall falls between parallel joists, install 2 × 4 blocking between the joists every 24". If the new wall is aligned with a parallel joist, install blocks on both sides of the wall, and attach the top plate to the joist (inset).

How to Build a Partition Wall

Mark the location of the new wall on the ceiling, then snap two chalk lines or use a scrap piece of 2× lumber as a template to mark layout lines for the top plate. Use a stud finder to locate floor joists or roof framing above the ceiling, and mark these locations with tick marks or tape outside the layout lines.

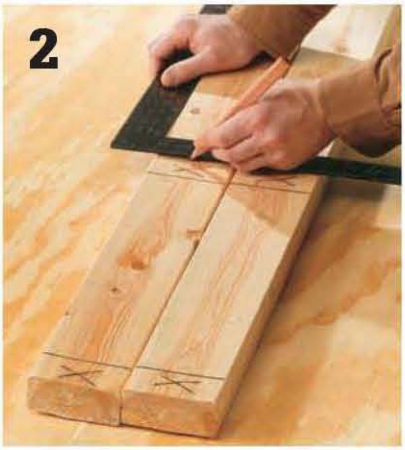

Cut the top and sole plates to length and lay them side by side. Use a combination or framing square to draw pairs of lines across both plates to mark the stud locations. Space the studs at 16" intervals, on center.

Mark the location of any door framing on the top and sole plates. Refer to the door's rough opening specifications when marking the layout. Draw lines for both the king and jack studs.

4

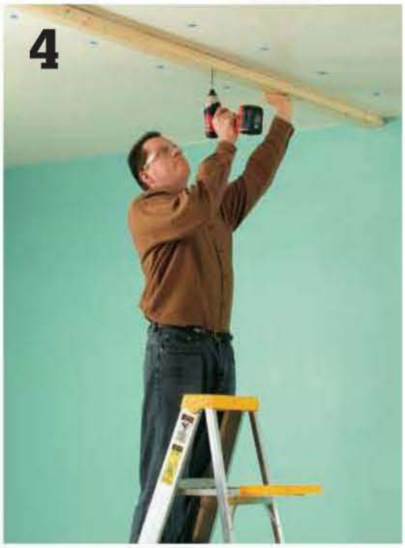

Fasten the top plate to the ceiling using 3" deck screws or 10d nails. Be sure to orient the plate so the stud layout faces down.

5

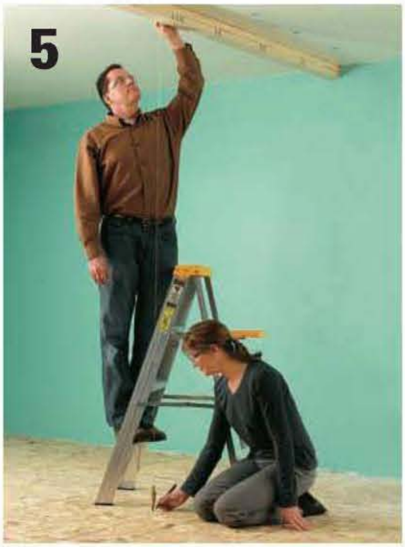

Hang a plumb bob from the edge of the top plate at several points along its length to find the sole plate location on the floor. The tip of the plumb bob should almost touch the floor. Wait until it stops moving before marking the sole plate reference point. Connect the points with a line to establish one edge of the sole plate. Use a piece of scrap 2× material as a template for marking the other edge.

6

Cut away the portion of the sole plate where the new door will be, and nail or screw the two sections to the floor between the sole plate layout lines. Use the cutaway door section as a spacer for the door when fastening the plates. Drive the fasteners into the floor framing. For concrete floors, attach the sole plate with a powder-actuated nail gun or with hardened masonry screws.

7

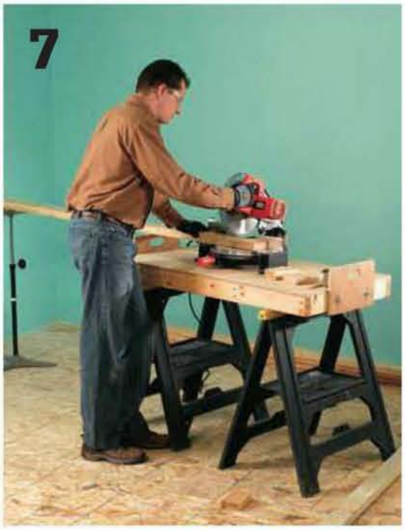

Measure the distance between the top and sole plates at several places along the wall to determine the stud lengths. The stud length distance may vary, depending on structural settling or an out-of-flat floor. Add $^1/_8$" to the stud length(s), and cut them to size. The extra length will ensure a snug fit between the wall plates.

8

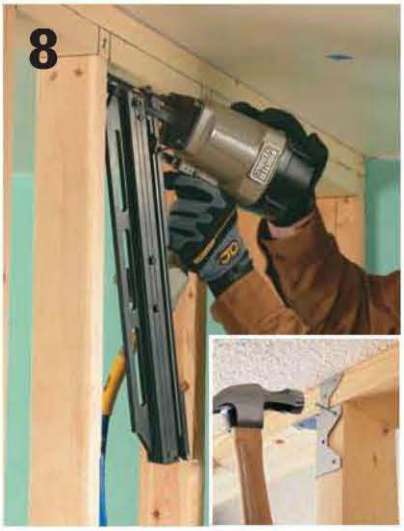

Install the framing members. Use a hammer to tap each stud into position, then toenail the studs to the top and sole plates. Fit, trim, and nail the studs, one at a time. *Tip: If the studs tend to shift during nailing, drive pilot holes for the nails first, or use 3" deck screws instead of nails. (Inset) An option for attaching wall studs to plates is to use metal connectors and 4d nails.*

9

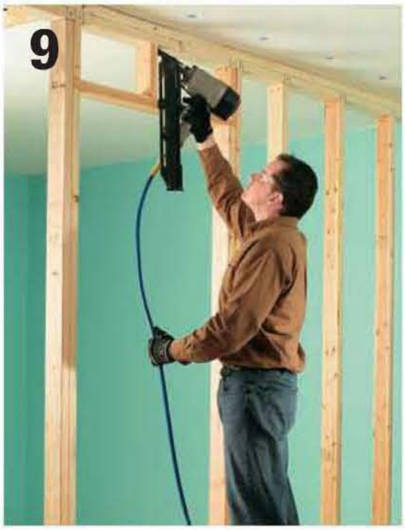

Nail the king studs, jack studs, a header and a cripple stud in place to complete the rough door framing. See page 12 for more information on framing a door opening.

(continued)

If building codes in your area require fire blocking, install 2× cutoff scraps between the studs, 4 ft. from the floor, to serve this purpose. Stagger the blocks so you can endnail each piece.

Drill holes through the studs to create raceways for wiring and plumbing. When this work is completed, fasten metal protector plates over these areas to prevent drilling or nailing through wiring and pipes later. Have your work inspected before proceeding with wallboard.

Cover the wall with wallboard. Plan the layout wisely to minimize waste and to avoid butted joints with untapered seams. If you are installing wallboard on a ceiling as well, do that first. For more on installing wallboard, see pages 44 to 51.

Tools & Materials for Framing with Steel

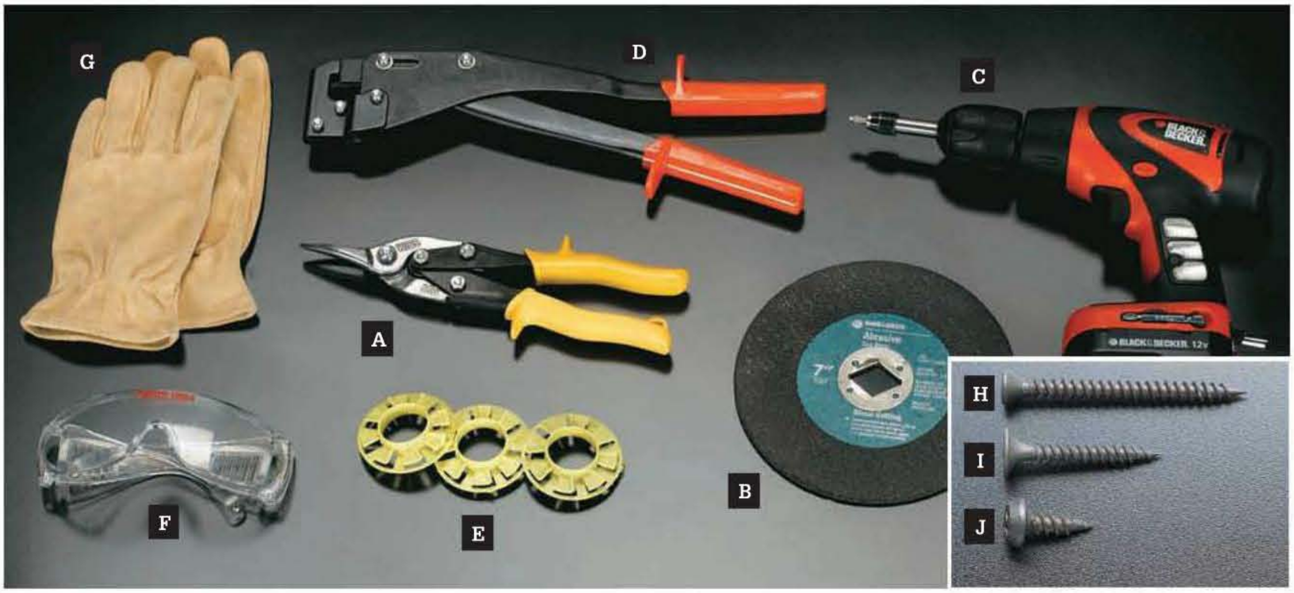

Steel framing requires a few specialty tools and materials. Aviation snips (A) are needed to cut tracks and studs, though a miter saw outfitted with a steel-cutting abrasive blade (B) can speed up the process. A drill or screw gun (C) is required for fastening framing. Handy for large projects, a stud crimper (D) creates mechanical joints between tracks and studs. Plastic grommets (E) are placed in knockouts to help protect utility lines. Protective eyewear and heavy work gloves (F, G) are necessities when working with the sharp edges of hand-cut steel framing. Use self-tapping screws (inset) to fasten steel components. To install wood trim, use type S trim-head screws (H); to fasten wallboard, type S wallboard screws (I); and to fasten studs and tracks together, 7⁄16" type S panhead screws (J).

Joining Sections Using Steel Studs ▸

Steel studs and tracks have the same basic structure—a web that spans two flanged sides—however, studs also contain a ¼" lip to improve their rigidity.

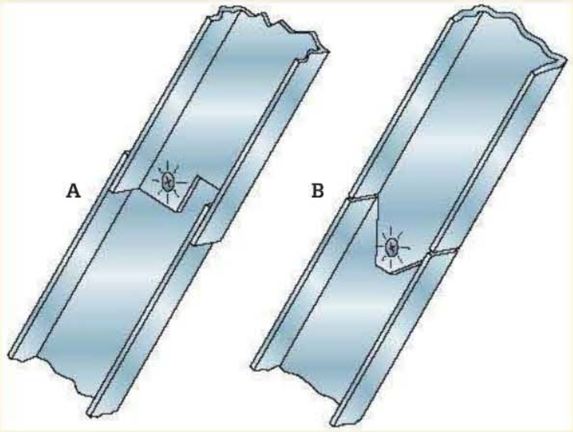

Join sections with a spliced joint (A) or notched joint (B). Make a spliced joint by cutting a 2" slit in the web of one track. Slip the other track into the slit and secure with a screw. For a notched joint, cut back the flanges of one track and taper the web so it fits into the other track; secure with a screw.

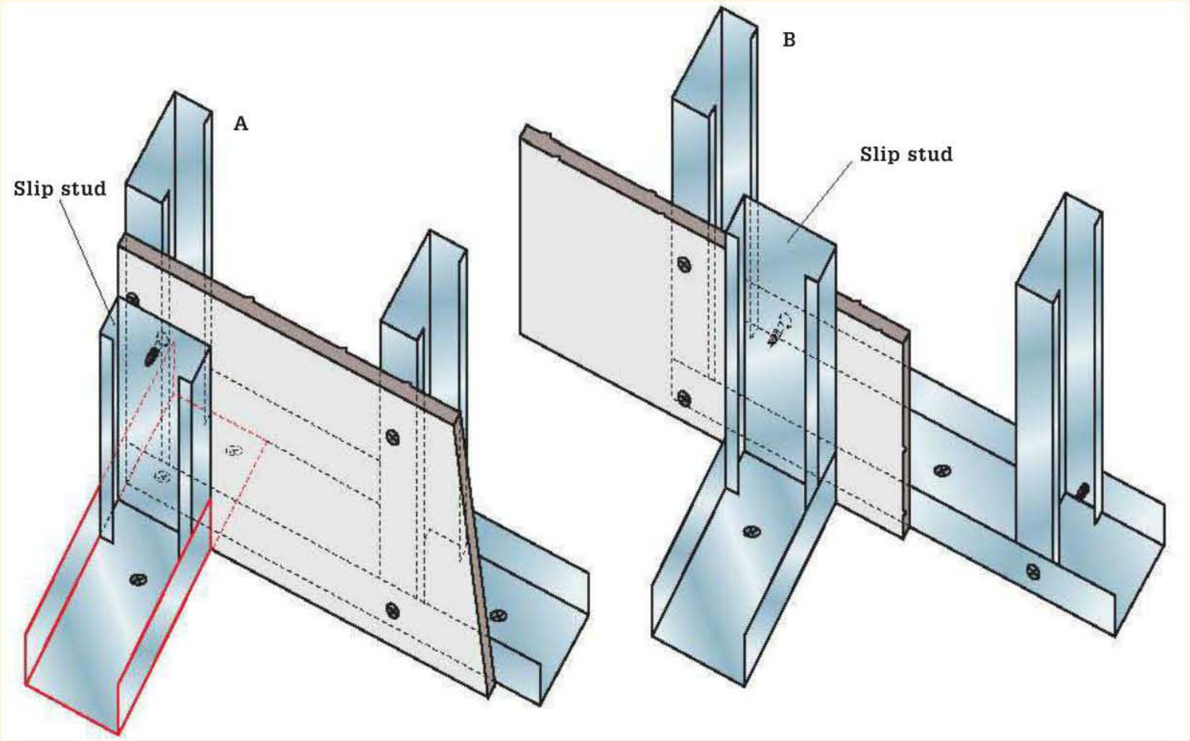

Build corners using a slip stud: A slip stud is not fastened until the adjacent drywall is in place. Form L-shaped corners (A) by overlapping the tracks. Cut off the flange on one side of one track, removing enough to allow room for the overlapping track and drywall. Form a T-shaped corner (B) by leaving a gap between the tracks for the drywall. Secure each slip stud by screwing through the stud into the tracks of the adjacent wall. Also screw through the back side of the drywall into the slip stud, if possible. Where there's no backing behind the slip stud, drive screws at a 45° angle through the back corners of the slip stud and into the drywall.

Installing Wallboard

Hanging wallboard is a project that can be completed quickly and easily with a little preplanning and a helping hand.

Planning the layout of panels will help you reduce waste and deal with problem areas. Where possible, install full panels perpendicular to the framing to add strength and rigidity to walls and ceilings. To save yourself time and trouble during the finishing process, avoid joints where two untapered panel ends are butted together. These are difficult to finish because there's no recess for the compound and tape. In small areas, install long sheets horizontally that run the full length of the walls. Or hang the panels vertically, which produces more seams that need taping but eliminates butted end joints. If butted joints are unavoidable, stagger the seams and locate them away from the center of the wall, or install back blocking to help mask unflattering effects.

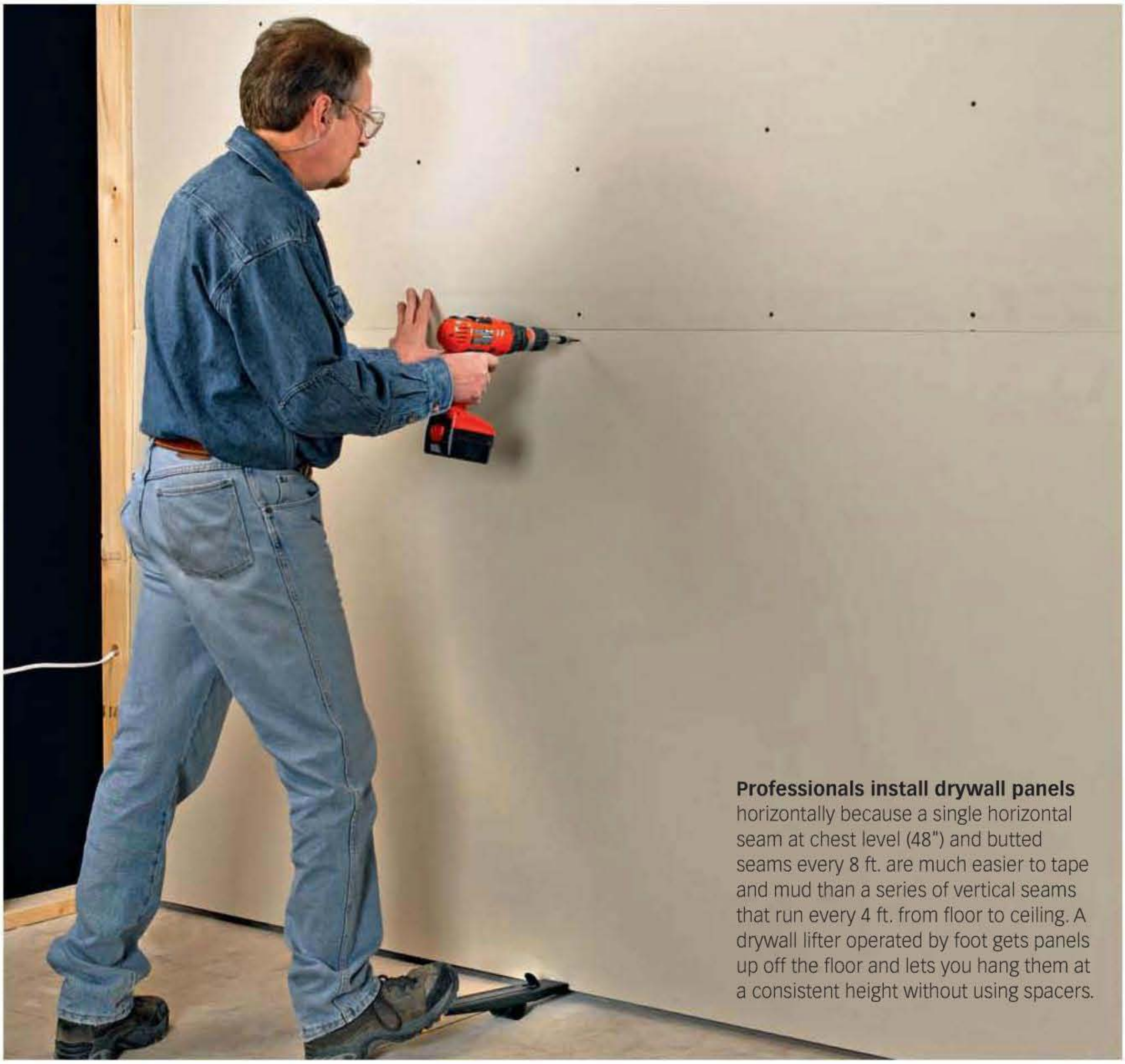

Professionals install drywall panels horizontally because a single horizontal seam at chest level (48") and butted seams every 8 ft. are much easier to tape and mud than a series of vertical seams that run every 4 ft. from floor to ceiling. A drywall lifter operated by foot gets panels up off the floor and lets you hang them at a consistent height without using spacers.

Tools & Prep Work

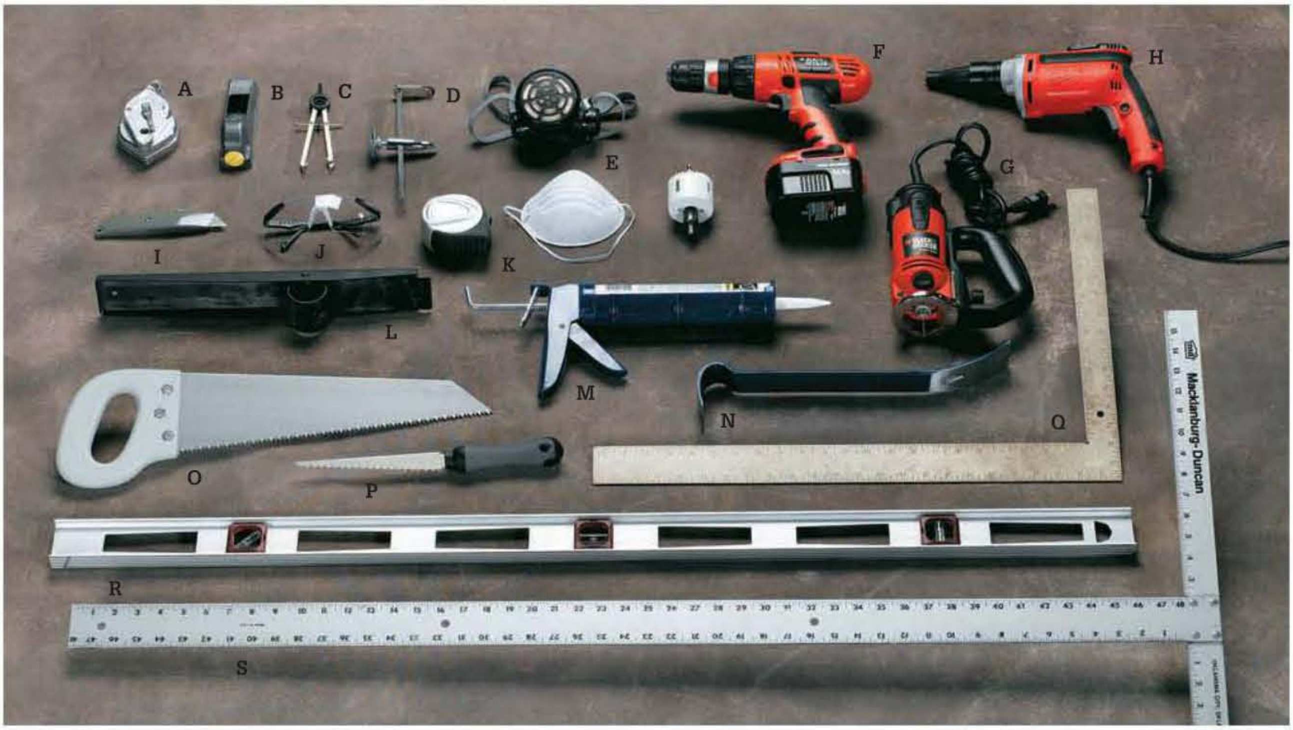

Tools for installing wallboard include: chalkline (A), surform plane (B), compass (C), wallboard compass (D), protective masks (E), drill with hole saw (F), trim router (G), drywall gun (H), utility knife (I), eye protection (J), tape measure (K), wallboard lifter (L), caulk gun (M), pry bar (N), wallboard saw (O), keyhole saw (P), framing square (Q), level (R), wallboard T-square (S).

Preparing for Wallboard Installation

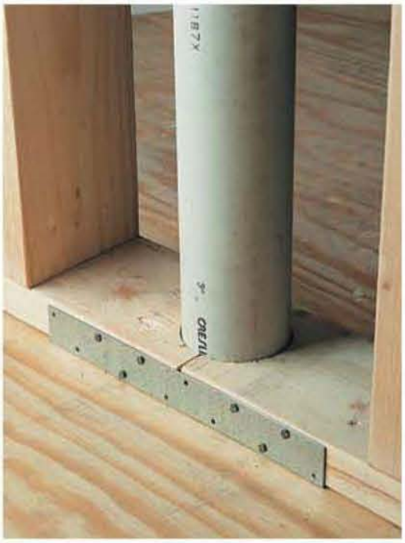

Use protector plates where wires or pipes pass through framing members less than 1¼" from the face. The plates prevent wallboard screws from puncturing wires or pipes.

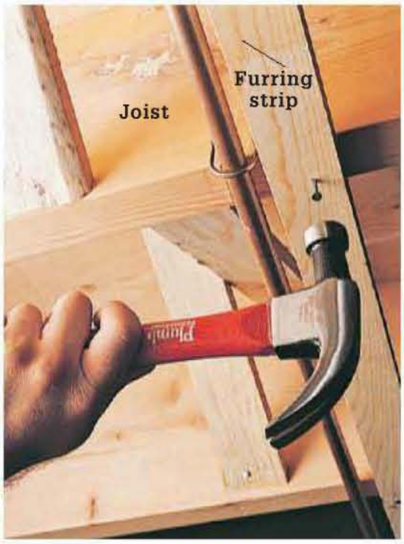

Nail furring strips to the framing to extend the wall surface beyond any obstructions such as water pipes or heating ducts.

Mark the locations of the studs on the floor with a carpenter's pencil or masking tape. After wallboard covers the studs, the marks indicate the stud locations.

How to Cut Wallboard

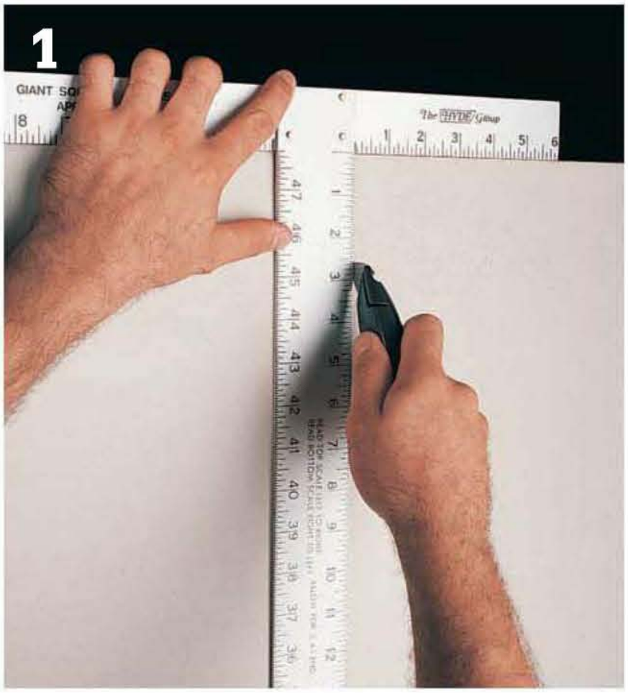

Position the wallboard T-square with the short arm flush against the edge. Use a utility knife to score the wallboard face paper along the arm of the square at the cutting point.

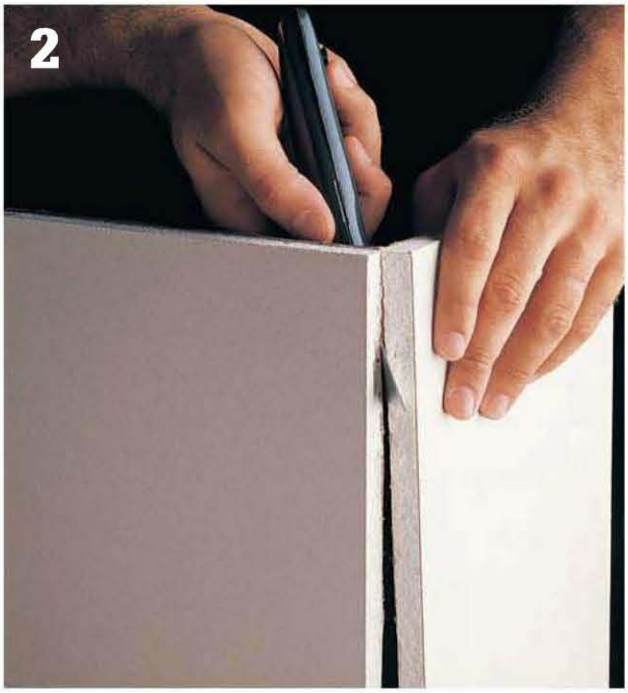

Bend the scored section with both hands to break the plaster core of the wallboard. Fold back the unwanted piece and cut through the back paper to separate the pieces.

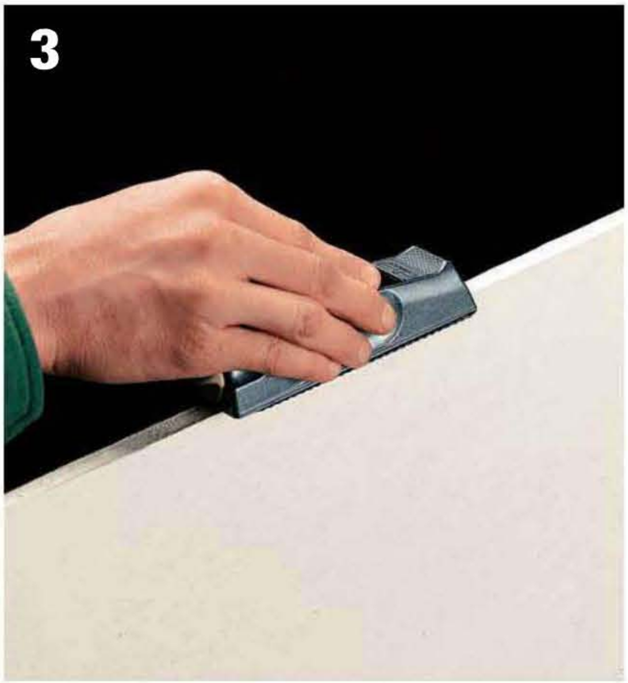

Smooth rough edges with a drywall rasp. One or two passes with the rasp should be sufficient. To help fit a piece into a tight space, bevel the edge slightly toward the back of the panel.

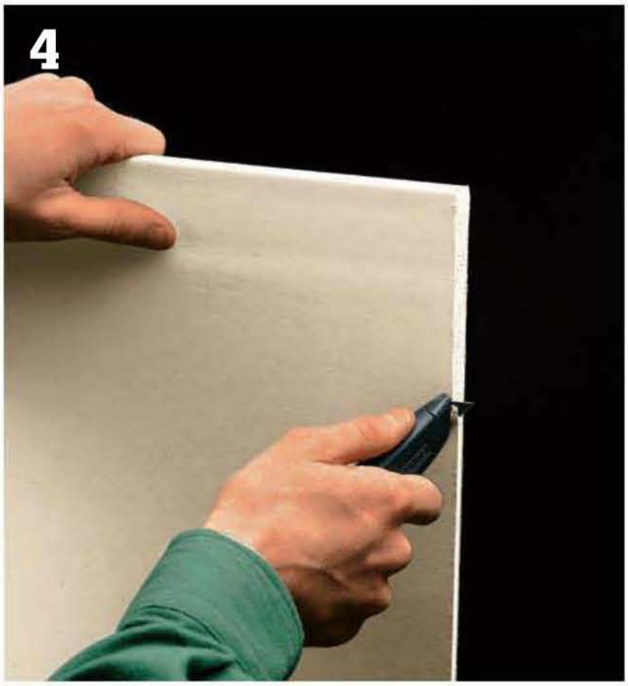

Where untapered panel ends will be butted together, bevel-cut the outside edges of each panel at 45°, removing about ⅛" of material. This helps prevent the paper from creating a ridge along the seam. Peel off any loose paper from the edge.

How to Install Wallboard on Flat Ceilings

1

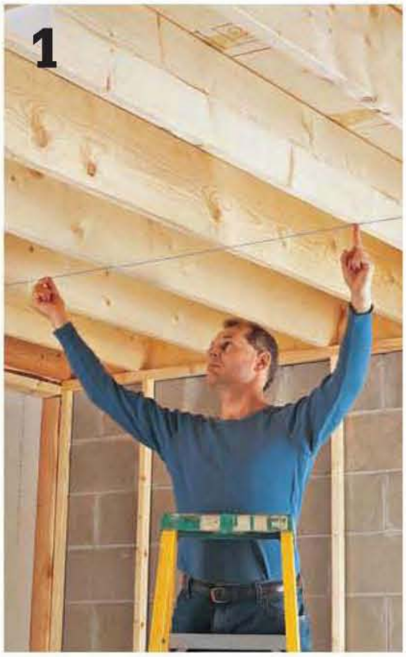

Snap a chalk line perpendicular to the joists, 48⅛" from the starting wall.

2

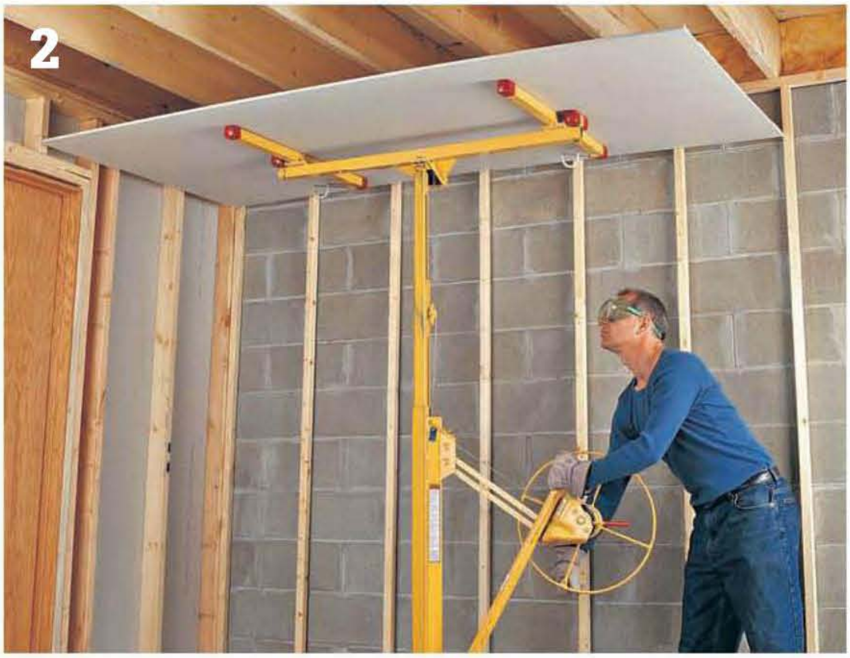

Measure to make sure the first panel will break on the center of a joist. If necessary, cut the panel on the end that abuts the side wall so the panel breaks on the next farthest joist. Load the panel onto a rented drywall lift, or use a helper, and lift the panel flat against the joists.

3

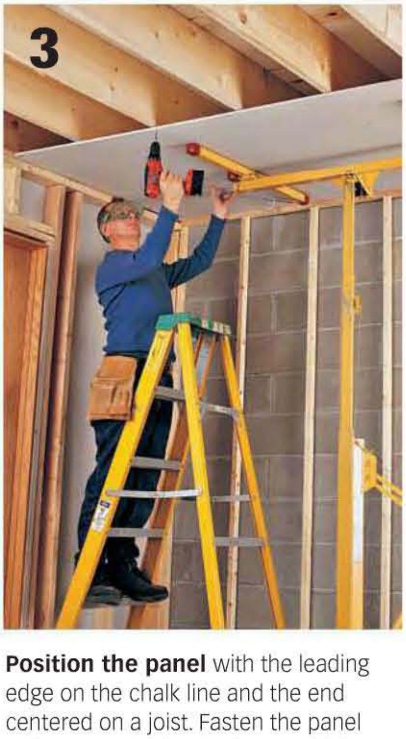

Position the panel with the leading edge on the chalk line and the end centered on a joist. Fasten the panel with appropriately sized screws driven every 12" along the joists.

4

After the first row of panels is installed, begin the next row with a half panel. This ensures that the butted end joints will be staggered between rows.

Tip ▸

Drywall stilts bring you within reach of ceilings, so you can fasten and finish the drywall without a ladder. Stilts are commonly available at rental centers and are surprisingly easy to use.

How to Install Wallboard on Walls

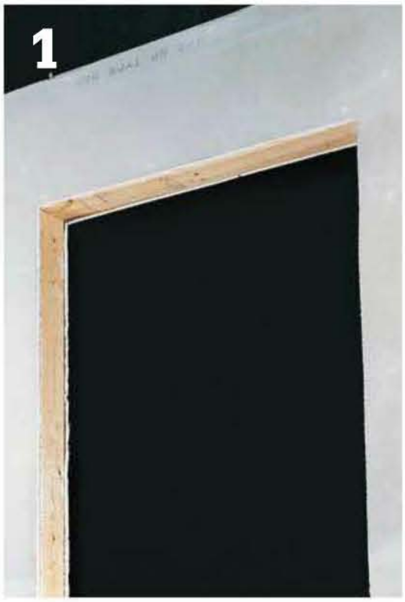

Plan wallboard placement so joints do not fall at the corners of doors or windows. Wallboard joints at corners often crack and cause bulges that interfere with miter joints in window or door trim.

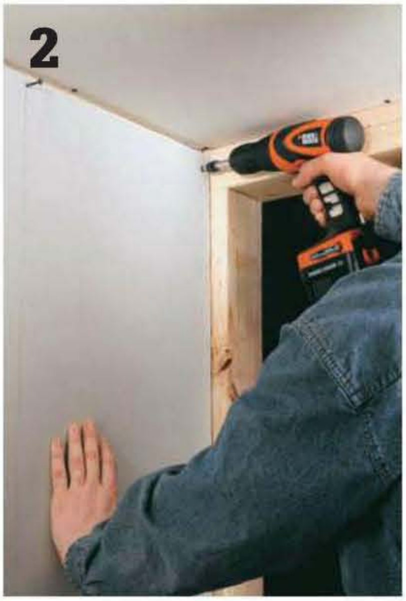

With a helper or a wallboard lift, hoist the first panel tight against the ceiling, making sure the side edge is centered on a stud. Push the panel flat against the framing and drive the starter screws to secure the panel. Make any cutouts, then fasten the field of the panel with drywall screws driven every 12".

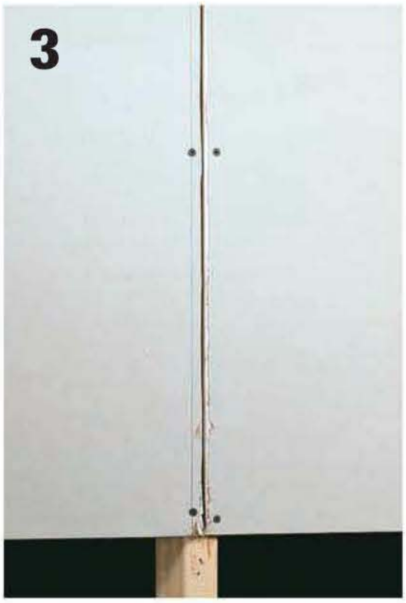

Measure, cut, and install the remaining panels along the upper wall. Bevel panel ends slightly, leaving a ⅛" gap between them at the joint. Butt joints can also be installed using back blocking to create a recess.

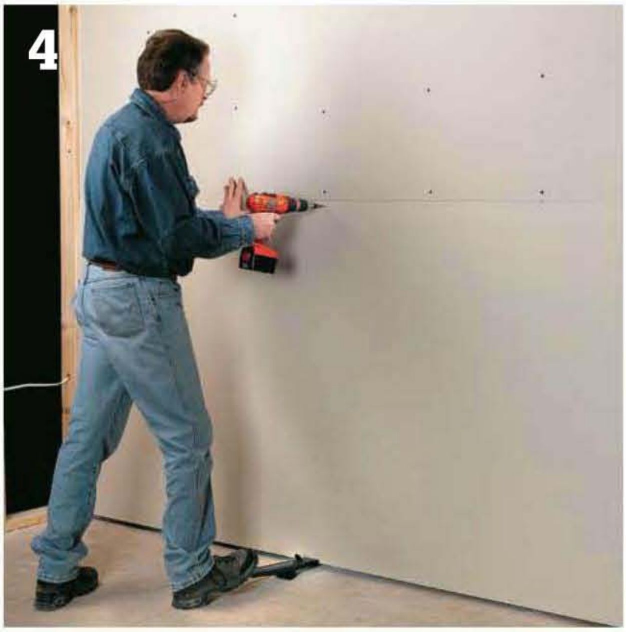

Measure, cut, and install the bottom row, butting the panels tight to the upper row and leaving a ½" gap at the floor. Secure to the framing along the top edge using the starter screws, then make all cutouts before fastening the rest of the panel.

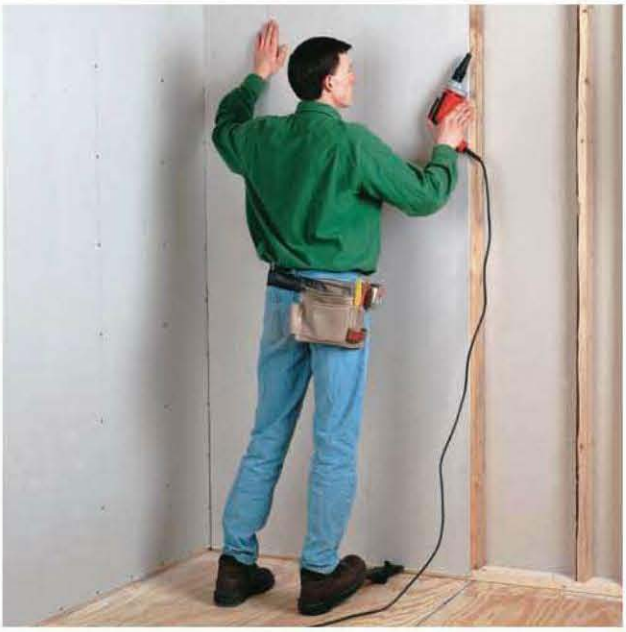

Variation: When installing wallboard vertically, cut each panel so it's ½" shorter than the ceiling height to allow for expansion. (The gap will be covered by base molding.) Avoid placing tapered edges at outside corners; they are difficult to finish.

How to Install Wallboard at Inside Corners

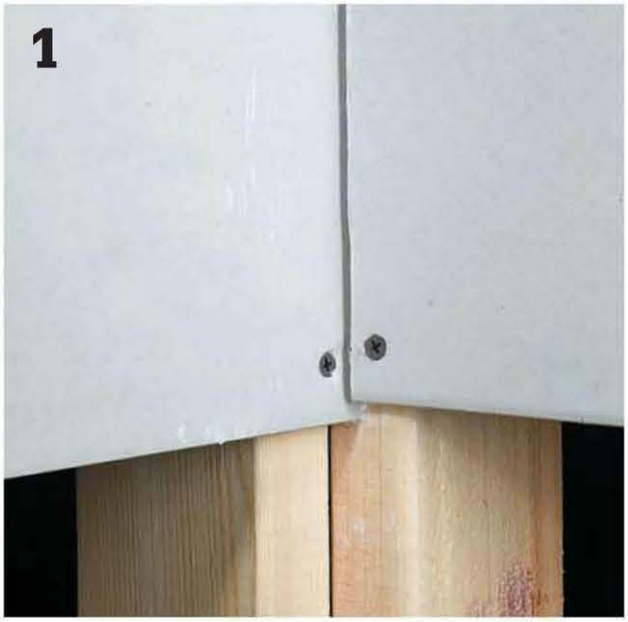

Standard 90° inside corners are installed with the first panel butted against the framing and the adjacent panel butted against the first. The screw spacing remains the same as on a flat wall (see page 48).

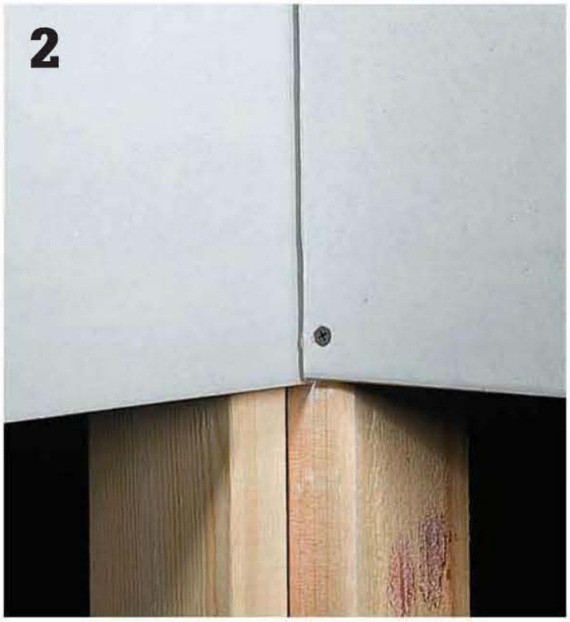

Use a "floating corner" to reduce the chances of popped fasteners and cracks. Install the first panel, fastening only to within one stud bay of the corner. Push the leading edge of the adjacent panel against the first to support the unfastened edge. Fasten the second panel normally, including the corner.

How to Install Wallboard at Outside Corners

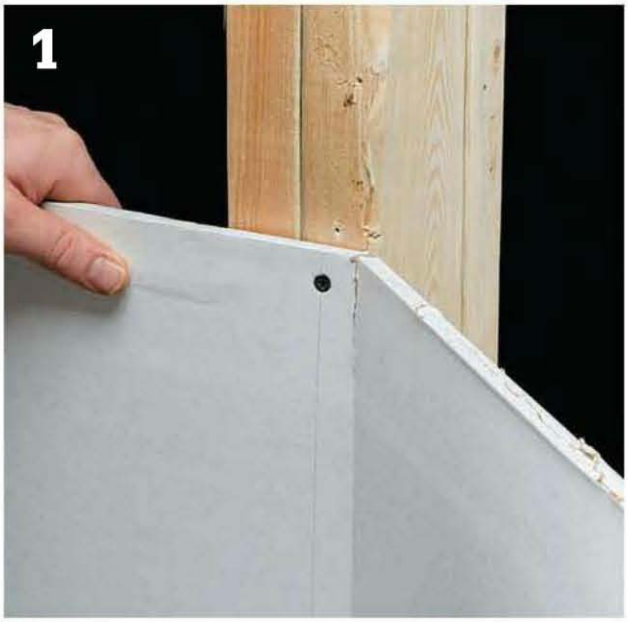

Cut the first panel so it is longer than the opening and extends past the corner framing. Fasten the panel in place, then score the backside and snap cut to remove the waste piece.

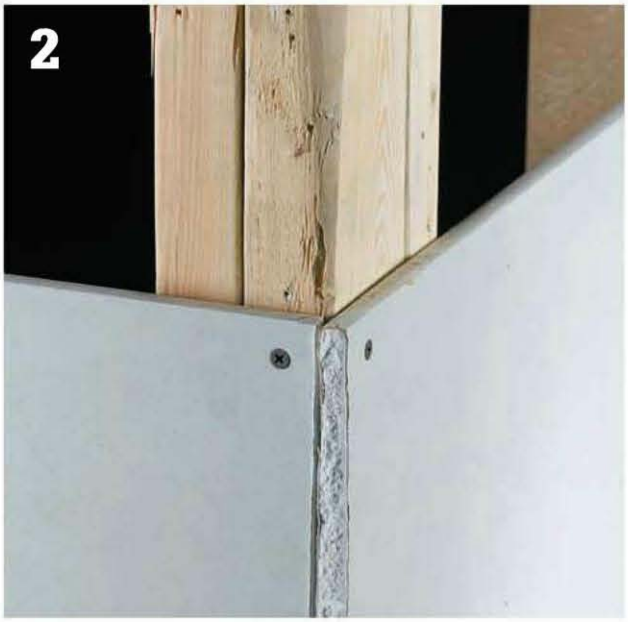

Install the second panel on the adjacent wall so it overlaps the end of the first panel slightly. Score and snap-cut so end is flush with first panel wall surface.

How to Tape Wallboard Joints

1

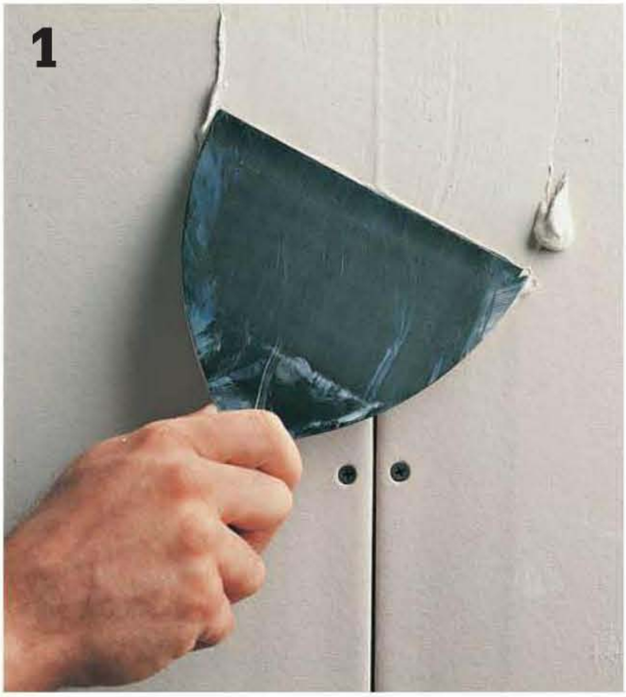

Apply a thin layer of wallboard compound over the joint with a 4" or 6" wallboard knife. To load the knife, dip it into a pan filled with wallboard compound.

2

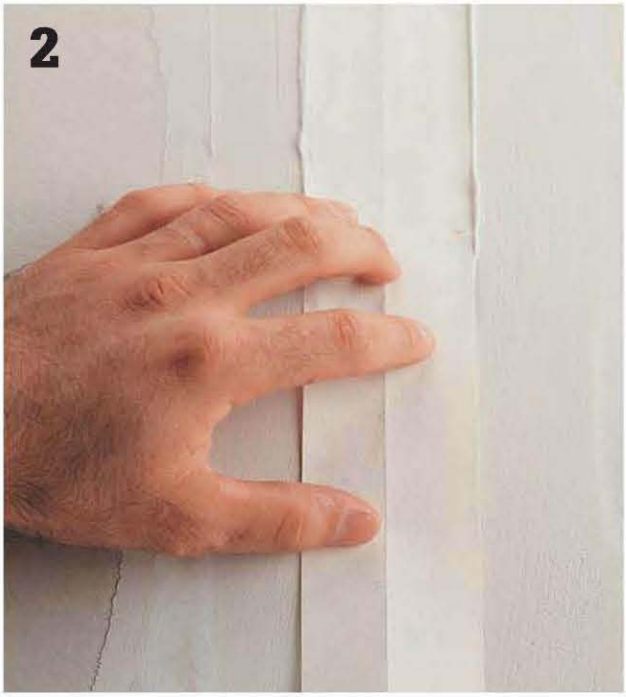

Press the wallboard tape into the compound immediately, centering the tape on the joint. Wipe away the excess compound and smooth the joint with a 6" knife. Let dry overnight.

3

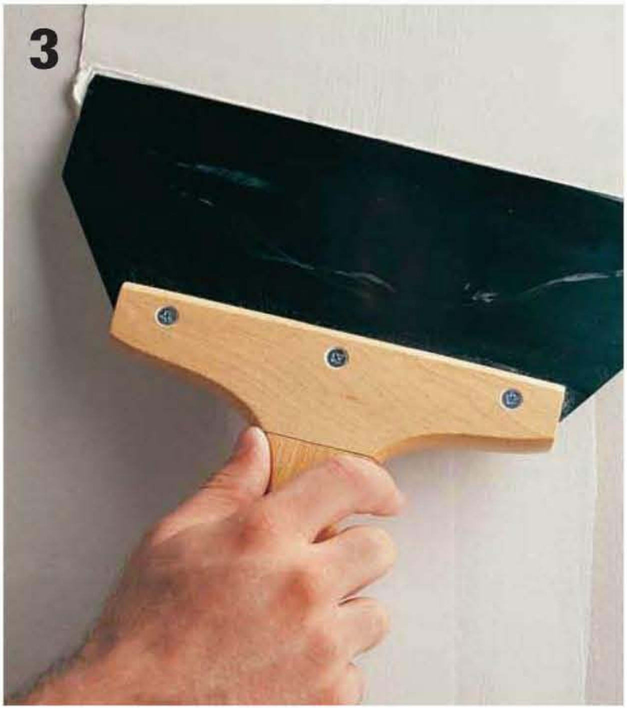

Apply a thin finish coat of compound with a 10" wallboard knife. Allow the second coat to dry and shrink overnight. Apply the last coat and let it harden slightly before wet-sanding.

Tip ▸

TRIKO PRODUCTS
TAPING • TOPPING • TEXTURE
REDI-MIX - ALL PURPOSE
JOINT COMPOUND

Use premixed wallboard compound for most taping and finishing jobs to eliminate mixing. Use paper wallboard tape when using premixed wallboard compound.

How to Finish Inside Corners

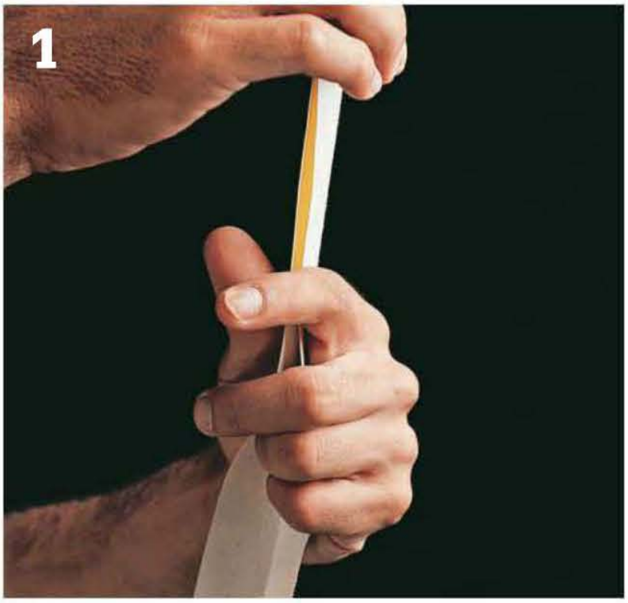

Fold a strip of paper wallboard tape in half by pinching the strip and pulling it between thumb and forefinger. Apply a thin layer of wallboard compound to both sides of the inside corner, using a 4" wallboard knife.

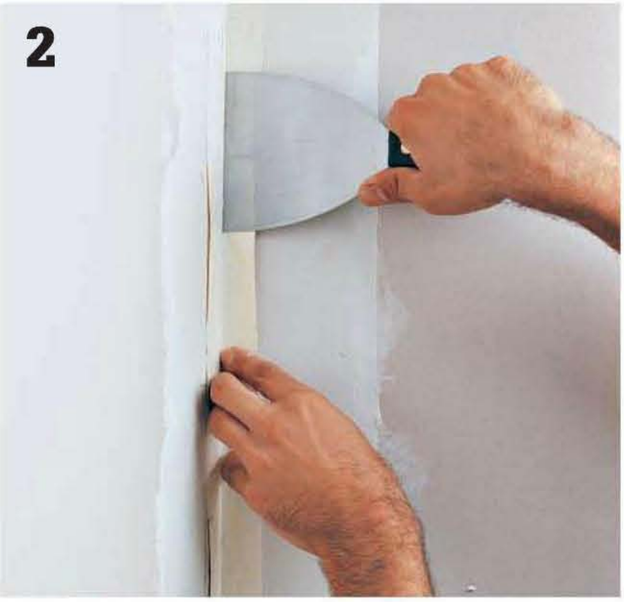

Position the end of the folded tape strip at the top of the joint and press the tape into the wet compound with the knife. Smooth both sides of the corner. Finish as described in step 3, preceding page.

How to Finish Outside Corners

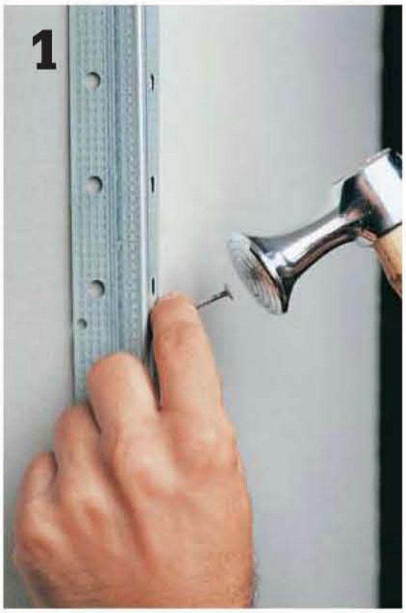

Position corner bead on outside corners. Using a level, adjust the bead so corner is plumb. Attach with 1¼" wallboard nails or screws spaced at 8" intervals. (Some corner beads are fastened with wallboard compound.)

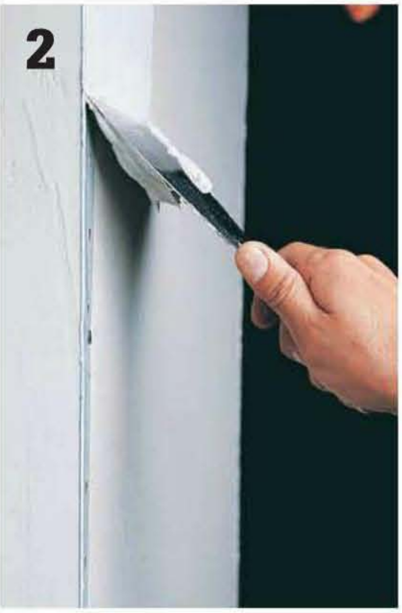

Cover the corner bead with three coats of wallboard compound, using a 6" or 10" wallboard knife. Let each coat dry and shrink overnight before applying the next coat. Smooth the final coat with a wet sander.

Tip ▸

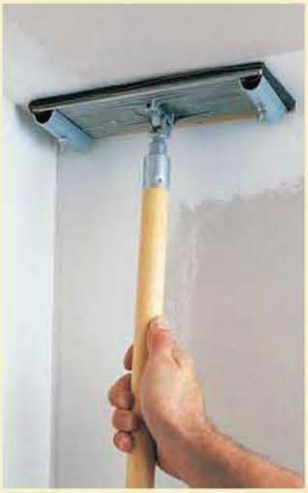

Sand joints lightly after wallboard compound dries. Use a pole sander to reach high areas without a ladder. Wear a dust mask when dry-sanding.

Understanding Plumbing

Because most of a plumbing system is hidden inside walls and floors, it may seem to be a complex maze of pipes and fittings. In fact, home plumbing is simple and straightforward. Understanding how home plumbing works is an important first step toward doing routine maintenance and money-saving repairs.

A typical home plumbing system includes three basic parts: a water supply system, a fixture and appliance set, and a drain system. These three parts can be seen clearly in the photograph of the cut-away house on the opposite page.

Fresh water enters a home through a main supply line (1). This fresh water source is provided by either a municipal water company or a private underground well. If the source is a municipal supplier, the water passes through a meter (2) that registers the amount of water used. A family of four uses about 400 gallons of water each day.

Immediately after the main supply enters the house, a branch line splits off (3) and is joined to a water heater (4). From the water heater, a hot water line runs parallel to the cold water line to bring the water supply to fixtures and appliances throughout the house. Fixtures include sinks, bathtubs, showers, and laundry tubs. Appliances include water heaters, dishwashers, clothes washers, and water softeners. Toilets and exterior sillcocks are examples of fixtures that require only a cold water line.

The water supply to fixtures and appliances is controlled with faucets and valves. Faucets and valves have moving parts and seals that eventually may wear out or break, but they are easily repaired or replaced.

Waste water then enters the drain system. It first must flow past a drain trap (5), a U-shaped piece of pipe that holds standing water and prevents sewer gases from entering the home. Every fixture must have a drain trap.

The drain system works entirely by gravity, allowing waste water to flow downhill through a series of large-diameter pipes. These drain pipes are attached to a system of vent pipes. Vent pipes (6) bring fresh air to the drain system, preventing suction that would slow or stop drain water from flowing freely. Vent pipes usually exit the house at a roof vent (7).

All waste water eventually reaches a main waste and vent stack (8). The main stack curves to become a sewer line (9) that exits the house near the foundation. In a municipal system, this sewer line joins a main sewer line located near the street. Where sewer service is not available, waste water empties into a septic system.

TRIDENT
METER
GALLONS

Water meters and main shutoff valves are located where the main water supply pipe enters the house. The water meter is the property of your local municipal water company. If the water meter leaks, or if you suspect it is not functioning properly, call your water company for repairs.

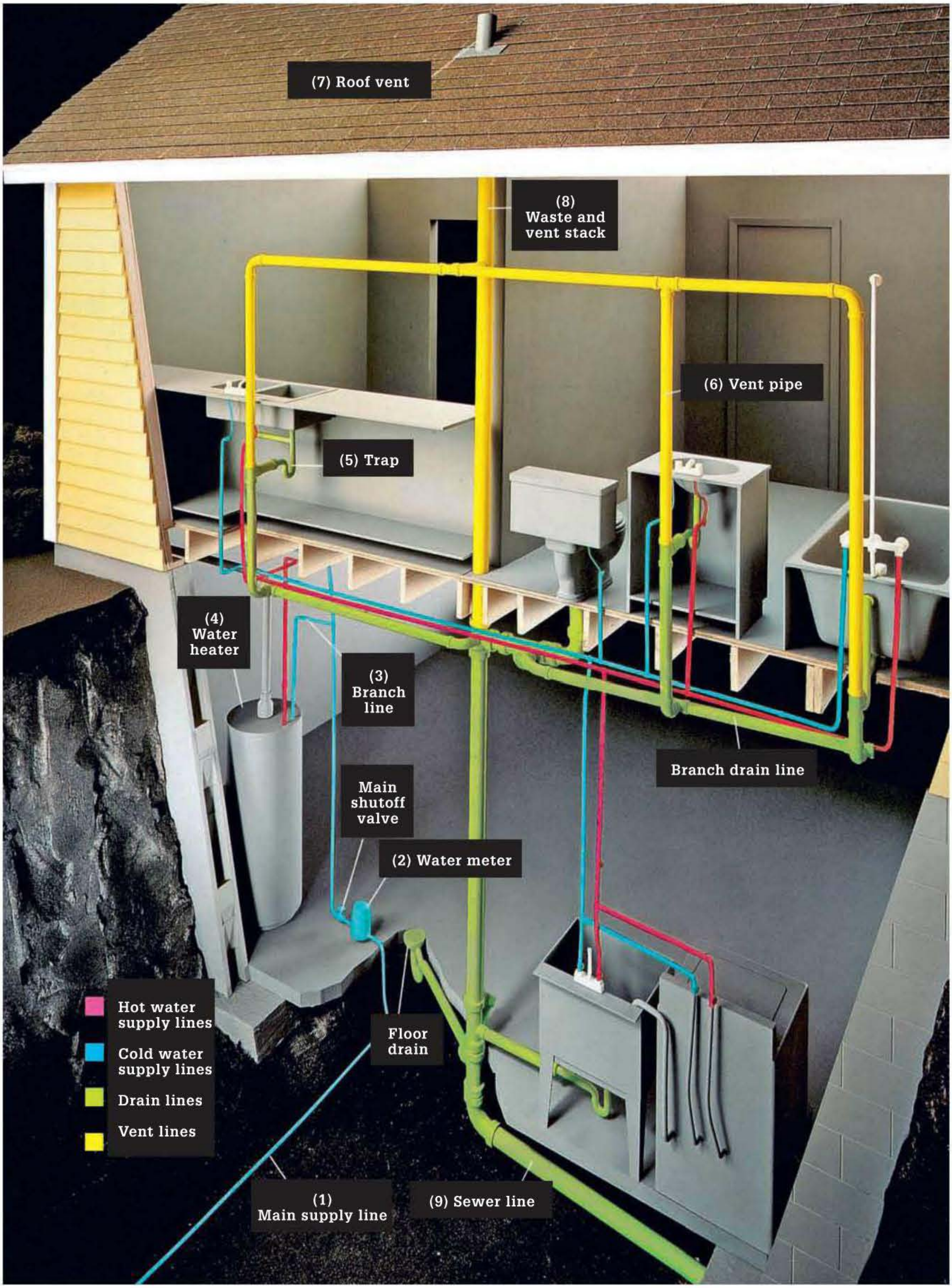
(7) Roof vent
(8) Waste and vent stack
(6) Vent pipe
(5) Trap
(4) Water heater
(3) Branch line
Branch drain line
Main shutoff valve
(2) Water meter
Hot water supply lines
Cold water supply lines
Drain lines
Vent lines
Floor drain
(1) Main supply line
(9) Sewer line

Water Supply System

Water supply pipes carry hot and cold water throughout a house. In homes built before 1960, the original supply pipes were usually made of galvanized iron. Newer homes have supply pipes made of copper. In most areas of the country, supply pipes made of rigid plastic or PEX are accepted by local plumbing codes.

Water supply pipes are made to withstand the high pressures of the water supply system. They have small diameters, usually ½" to 1", and are joined with strong, watertight fittings. The hot and cold lines run in tandem to all parts of the house. Usually, the supply pipes run inside wall cavities or are strapped to the undersides of floor joists.

Hot and cold water supply pipes are connected to fixtures or appliances. Fixtures include sinks, tubs, and showers. Some fixtures, such as toilets or hose bibs, are supplied only by cold water. Appliances include dishwashers and clothes washers. A refrigerator icemaker uses only cold water. Tradition says that hot water supply pipes and faucet handles are found on the left-hand side of a fixture, with cold water on the right.

Because it is pressurized, the water supply system is prone to leaks. This is especially true of galvanized iron pipe, which has limited resistance to corrosion.

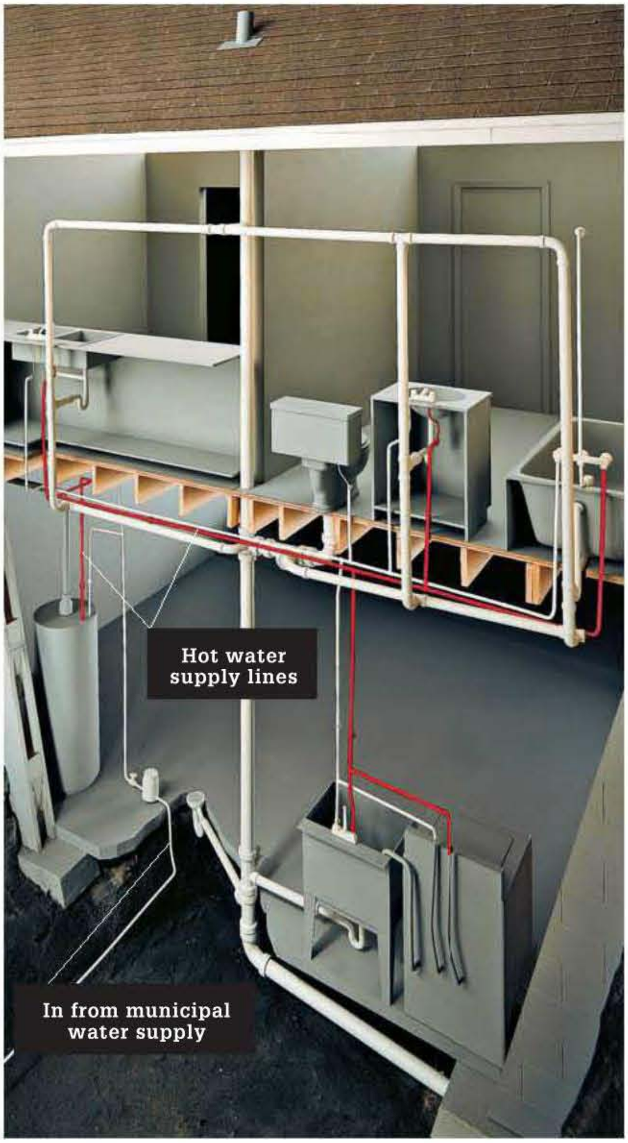

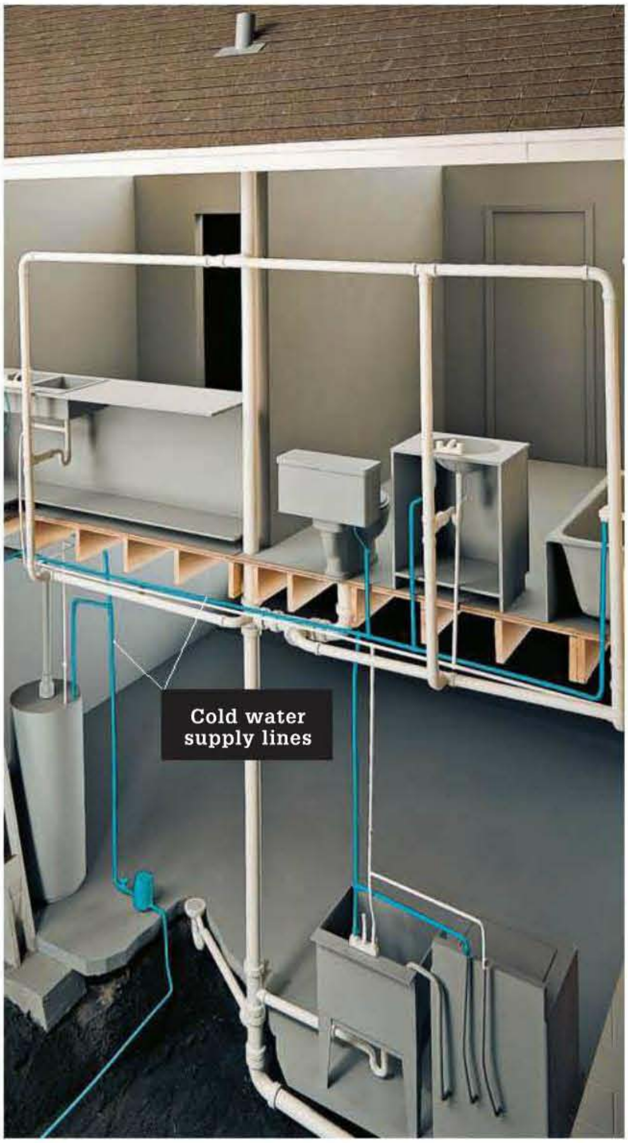

Drain-Waste-Vent System

Drain pipes use gravity to carry waste water away from fixtures, appliances, and other drains. This waste water is carried out of the house to a municipal sewer system or septic tank.

Drain pipes are usually plastic or cast iron. In some older homes, drain pipes may be made of copper or lead. Because they are not part of the supply system, lead drain pipes pose no health hazard. However, lead pipes are no longer manufactured for home plumbing systems.

Drain pipes have diameters ranging from 1¼" to 4". These large diameters allow waste to pass through easily.

Traps are an important part of the drain system. These curved sections of drain pipe hold standing water, and they are usually found near any drain opening. The standing water of a trap prevents sewer gases from backing up into the home. Each time a drain is used, the standing trap water is flushed away and is replaced by new water.

In order to work properly, the drain system requires air. Air allows waste water to flow freely down drain pipes.

To allow air into the drain system, drain pipes are connected to vent pipes. All drain systems must include vents, and the entire system is called the drain-waste-vent (DWV) system. One or more vent stacks, located on the roof, provide the air needed for the DWV system to work.

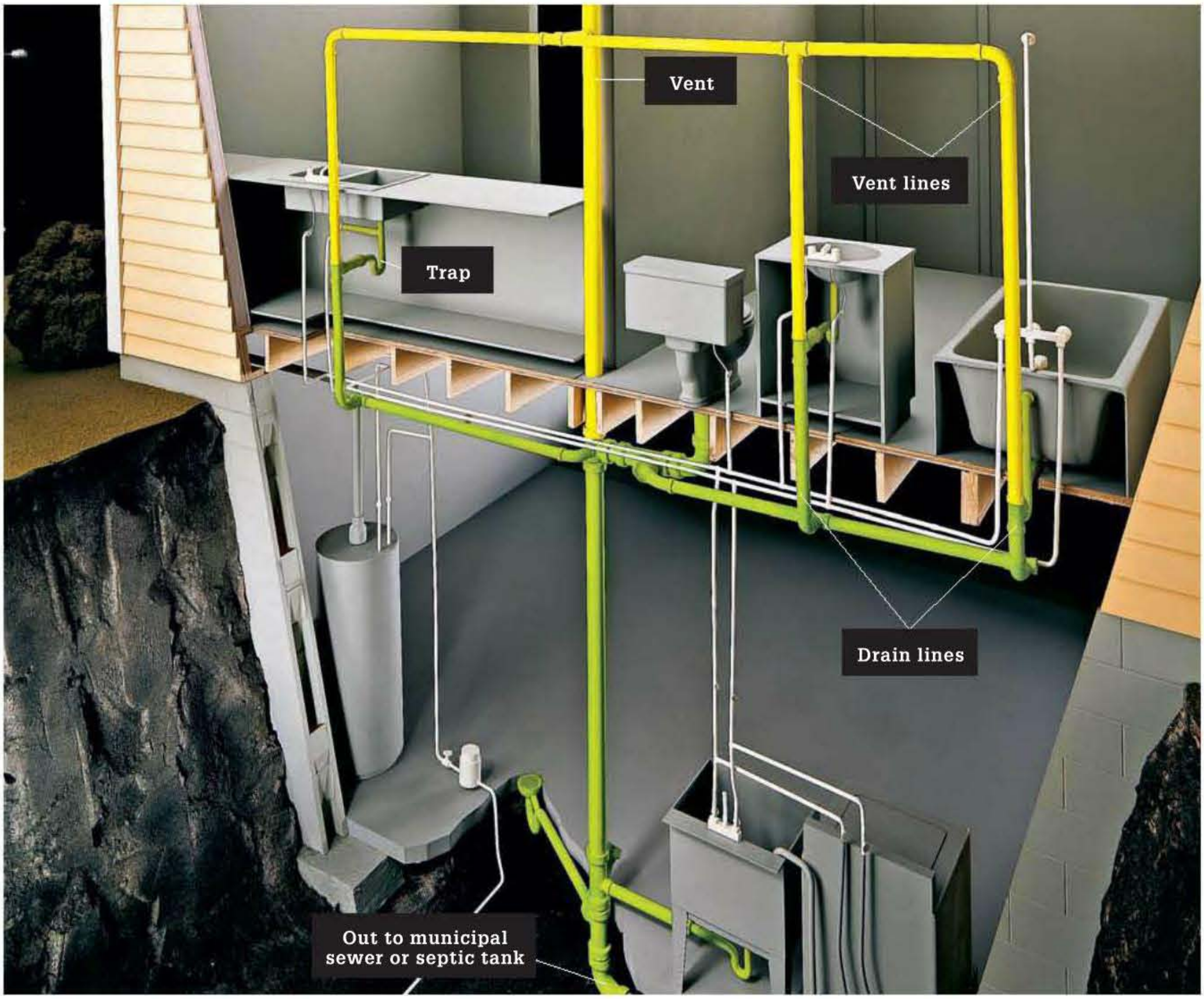

Working with Copper Pipe

Copper is the ideal material for water supply pipes. It resists corrosion and has smooth surfaces that provide good water flow. Copper pipes are available in several diameters (page 314), but most home water supply systems use ½" or ¾" pipe. Copper pipe is manufactured in rigid and flexible forms.

Rigid copper, sometimes called hard copper, is approved for home water supply systems by all local codes. It comes in three wall-thickness grades: Types M, L, and K. Type M is the thinnest, the least expensive, and a good choice for do-it-yourself home plumbing.

Rigid Type L usually is required by code for commercial plumbing systems. Because it is strong and solders easily, Type L may be preferred by some professional plumbers and do-it-yourselfers for home use. Type K has the heaviest wall thickness and is used most often for underground water service lines.

Flexible copper, also called soft copper, comes in two wall-thickness grades: Types L and K. Both are approved for most home water supply systems, although flexible Type L copper is used primarily for gas service lines. Because it is bendable and will resist a mild frost, Type L may be installed as part of a water supply system in unheated indoor areas, like crawl spaces. Type K is used for underground water service lines.

A third form of copper, called DWV, is used for drain systems. Because most codes now allow low-cost plastic pipes for drain systems, DWV copper is seldom used.

Copper pipes are connected with soldered, compression, or flare fittings (see chart below). Always follow your local code for the correct types of pipes and fittings allowed in your area.

Soldered fittings, also called sweat fittings, often are used to join copper pipes. Correctly soldered fittings (pages 278 to 280) are strong and trouble-free. Copper pipe can also be joined with compression fittings or flare fittings. See chart below.

Copper Pipe & Fitting Chart ▸

	Rigid Copper			Flexible Copper		
Fitting Method	**Type M**	**Type L**	**Type K**	**Type L**	**Type K**	**General Comments**
Soldered	yes	yes	yes	yes	yes	Inexpensive, strong, and trouble-free fitting method. Requires some skill.
Compression	yes	no	no	no	no	Makes repairs and replacement easy. More expensive than solder. Best used on flexible copper.
Flare	no	no	yes	yes	yes	Use only with flexible copper pipes. Usually used as a gas-line fitting. Requires some skill.

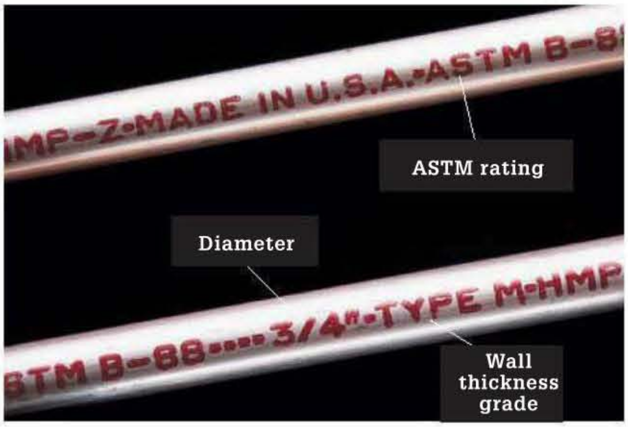
MP-Z-MADE IN U.S.A.-ASTM B-8
ASTM rating
Diameter
STM B-88----3/4"-TYPE M-HMP
Wall thickness grade

Grade stamp information includes the pipe diameter, the wall-thickness grade, and a stamp of approval from the ASTM (American Society for Testing and Materials). Type M pipe (shown) is identified by red lettering, Type L by blue lettering.

Bend flexible copper pipe with a coil-spring tubing bender to avoid kinks. Select a bender that matches the outside diameter of the pipe. Slip bender over pipe using a twisting motion. Bend pipe slowly until it reaches the correct angle, but not more than 90º.

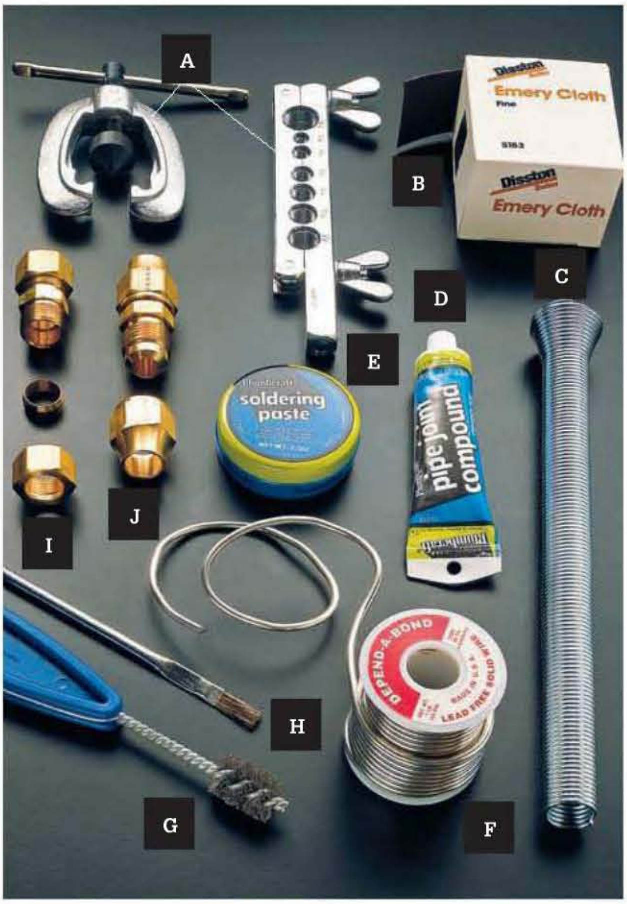
A
B
C
D
E
F
G
H
I
J
Disston
Emery Cloth
soldering paste
pipejoint compound
DEPEND-A-BOND
LEAD FREE SOLID WIRE

Specialty tools and materials for working with copper include: flaring tools (A), emery cloth (B), coil-spring tubing bender (C), pipe joint compound (D), soldering paste (flux) (E), lead-free solder (F), wire brush (G), flux brush (H), compression fitting (I), flare fitting (J).

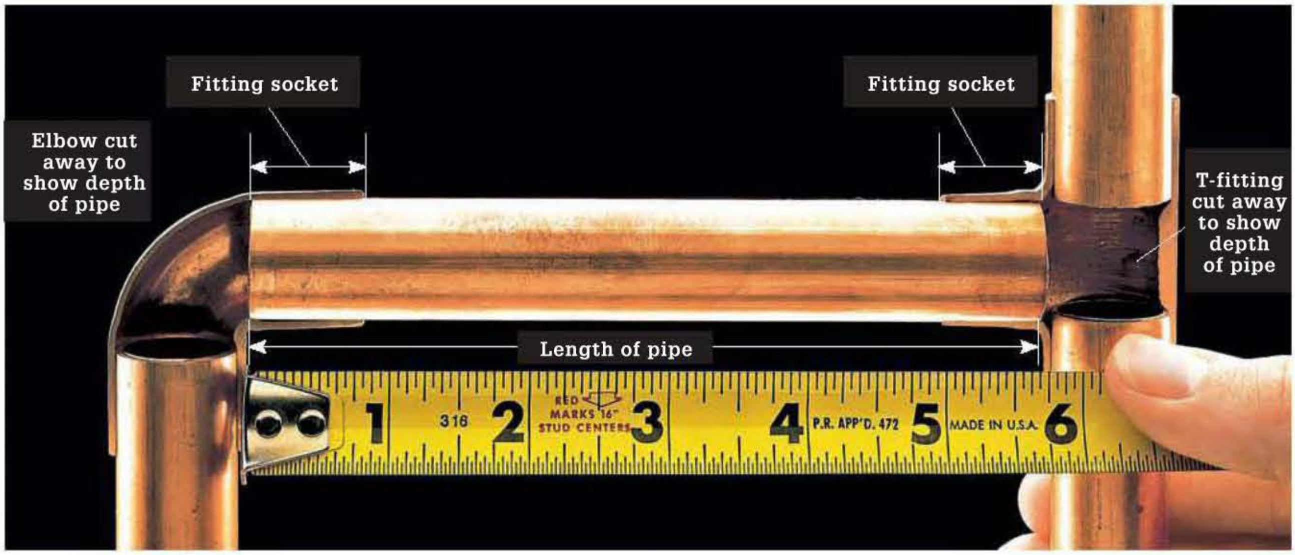
Fitting socket
Fitting socket
Elbow cut away to show depth of pipe
T-fitting cut away to show depth of pipe
Length of pipe
1
316
2
RED MARKS 16" STUD CENTERS
3
4
P.R. APP'D. 472
5
MADE IN U.S.A.
6

Find the length of copper pipe needed by measuring between the bottom of the copper fitting sockets (fittings shown in cutaway). Mark the length on the pipe with a felt-tipped pen.

Cutting & Soldering Copper

The best way to cut rigid and flexible copper pipe is with a tubing cutter. A tubing cutter makes a smooth, straight cut, an important first step toward making a watertight joint. Remove any metal burrs on the cut edges with a reaming tool or round file.

Copper can be cut with a hacksaw. A hacksaw is useful in tight areas where a tubing cutter will not fit. Take care to make a smooth, straight cut when cutting with a hacksaw.

A soldered pipe joint, also called a sweated joint, is made by heating a copper or brass fitting with a propane torch until the fitting is just hot enough to melt metal solder. The heat draws the solder into the gap between the fitting and pipe to form a watertight seal. A fitting that is overheated or unevenly heated will not draw in solder. Copper pipes and fittings must be clean and dry to form a watertight seal.

Tools & Materials ▸

- Tubing cutter with reaming tip (or hacksaw and round file)
- Wire brush
- Flux brush
- Propane torch
- Spark lighter (or matches)
- Round file
- Cloth
- Adjustable wrench
- Channel-type pliers
- Copper pipe
- Copper fittings
- Emery cloth
- Soldering paste (flux)
- Sheet metal
- Lead-free solder
- Rag

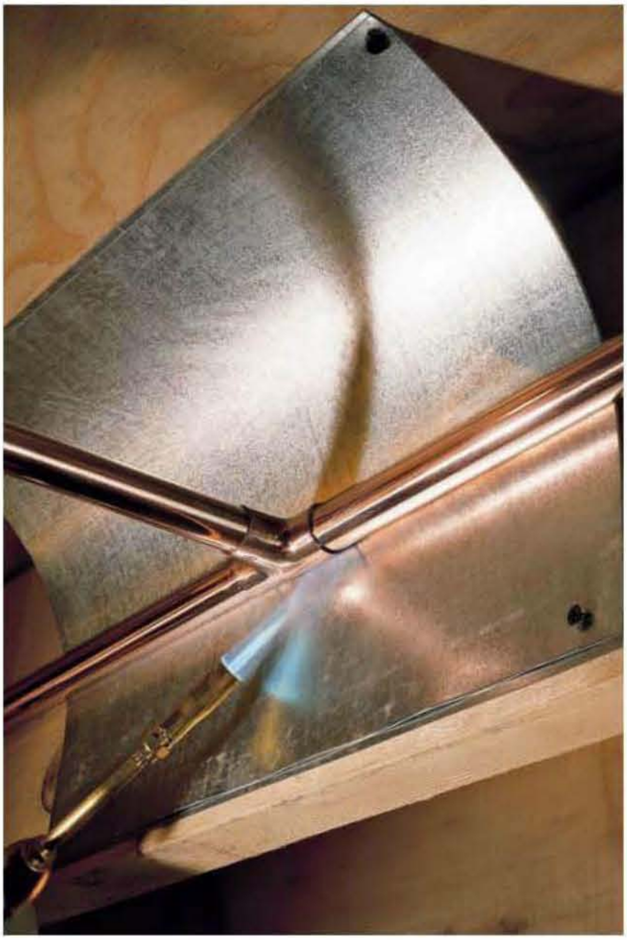

Protect wood from the heat of the torch flame while soldering, using a double layer (two 18" × 18" pieces) of 26-gauge sheet metal. Buy sheet metal at hardware stores or building supply centers and keep it to use with all soldering projects.

Soldering Tips ▸

Use caution when soldering copper. Pipes and fittings become very hot and must be allowed to cool before handling.

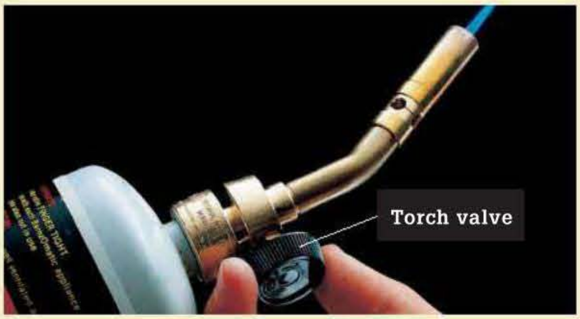

Prevent accidents by shutting off propane torch immediately after use. Make sure valve is closed completely.

How to Cut Rigid & Flexible Copper Pipe

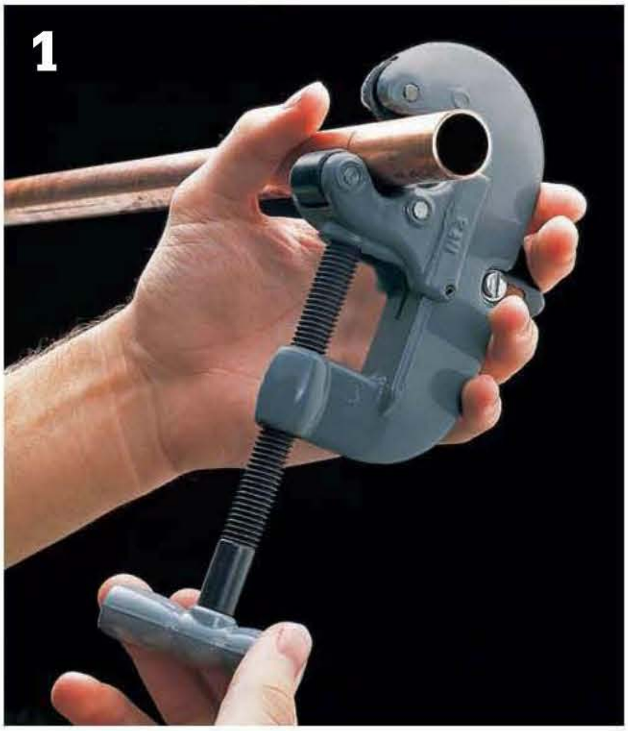

Place the tubing cutter over the pipe and tighten the handle so that the pipe rests on both rollers, and the cutting wheel is on the marked line.

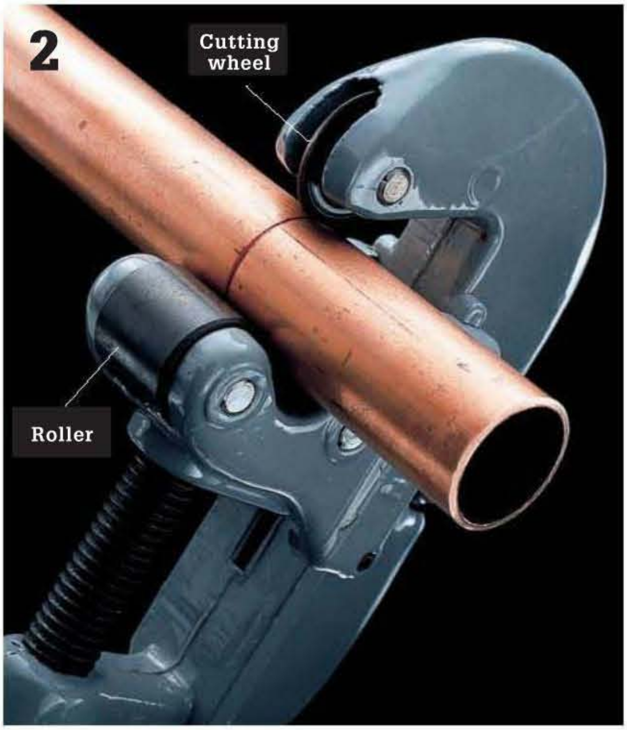

Turn the tubing cutter one rotation so that the cutting wheel scores a continuous straight line around the pipe.

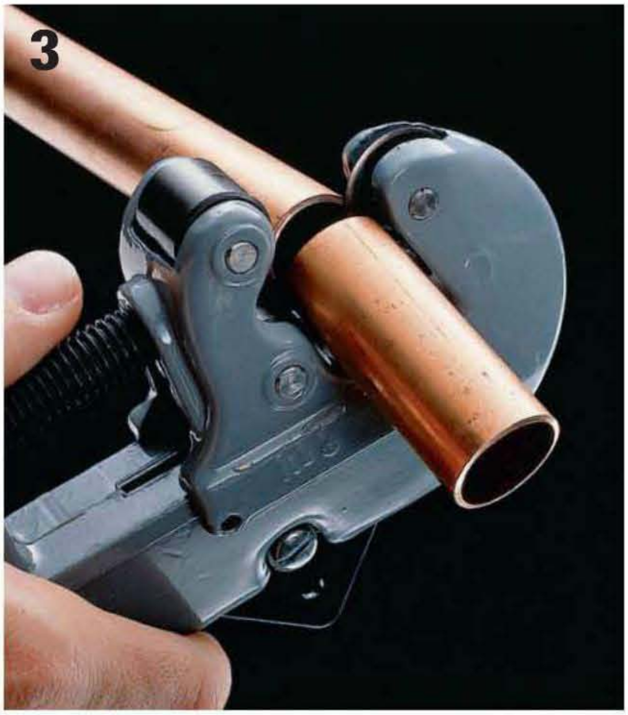

Rotate the cutter in the opposite direction, tightening the handle slightly after every two rotations, until the cut is complete.

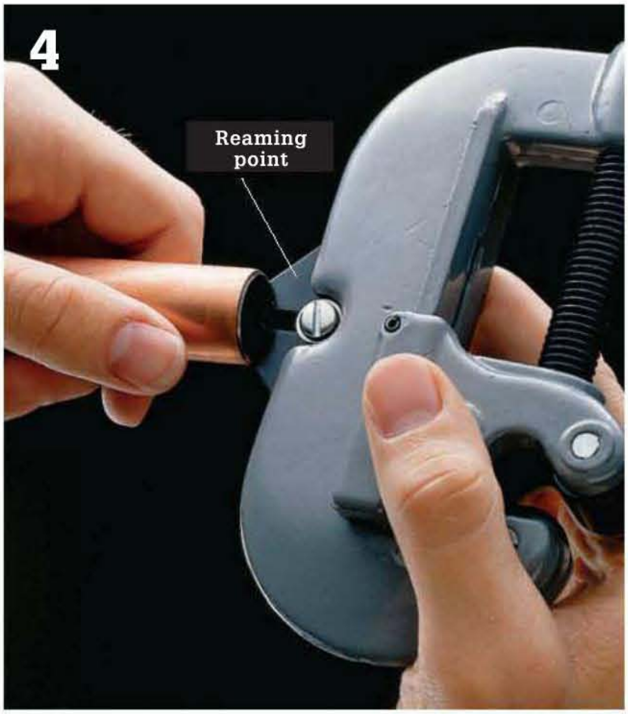

Remove sharp metal burrs from the inside edge of the cut pipe, using the reaming point on the tubing cutter, or a round file.

How to Solder Copper Pipes & Fittings

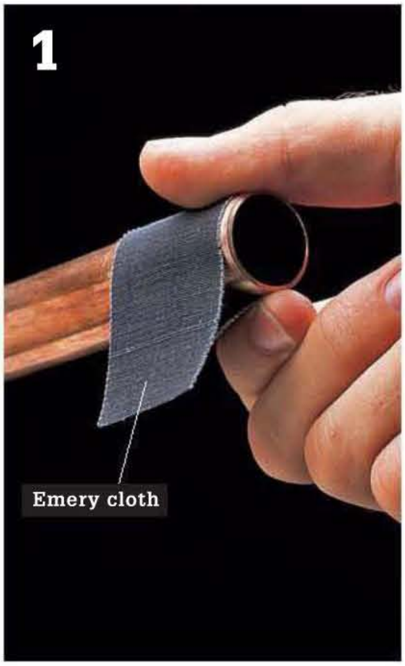
1
Emery cloth

Clean the end of each pipe by sanding with emery cloth. Ends must be free of dirt and grease to ensure that the solder forms a good seal.

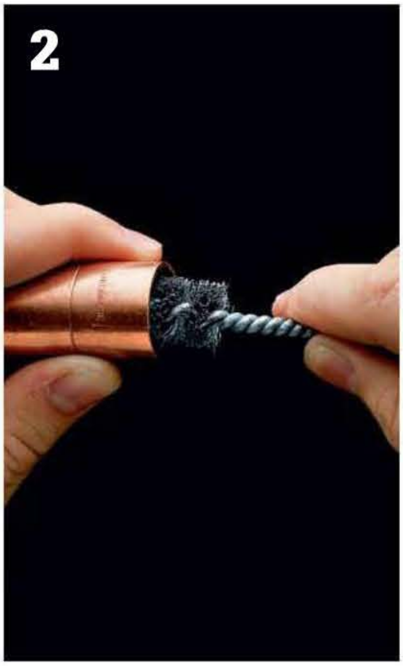

Clean the inside of each fitting by scouring with a wire brush or emery cloth.

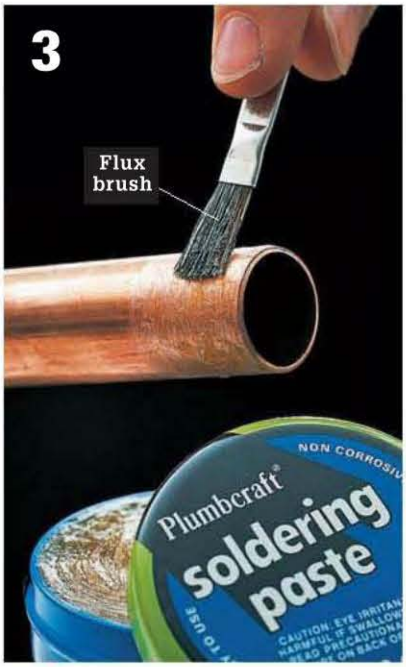
3
Flux brush
Plumbcraft
soldering paste
NON CORROSIVE

Apply a thin layer of soldering paste (flux) to end of each pipe, using a flux brush. Soldering paste should cover about 1" of pipe end.

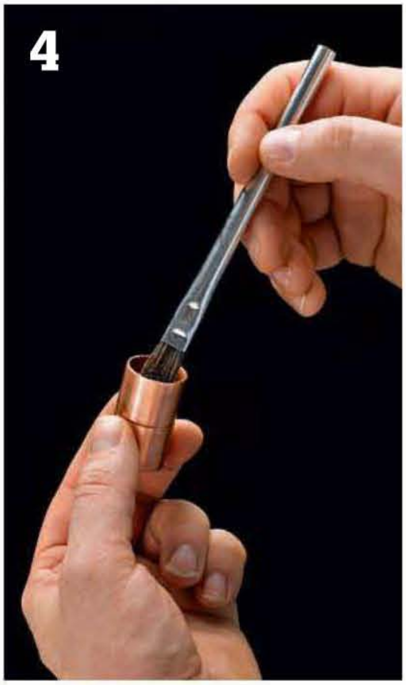

Apply a thin layer of flux to the inside of the fitting.

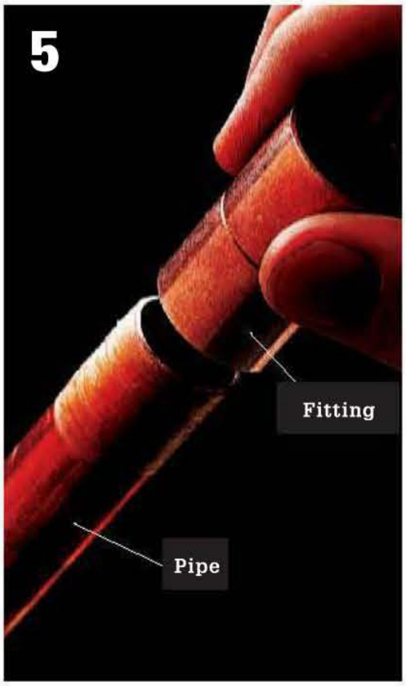
5
Fitting
Pipe

Assemble each joint by inserting the pipe into the fitting so it is tight against the bottom of the fitting sockets. Twist each fitting slightly to spread soldering paste.

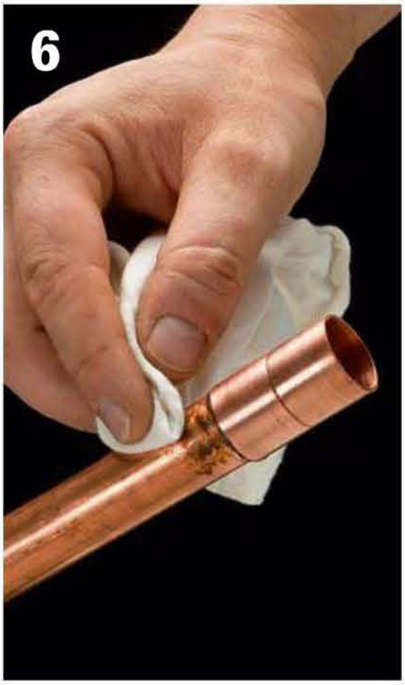

Use a clean dry cloth to remove excess flux before soldering the assembled fitting.

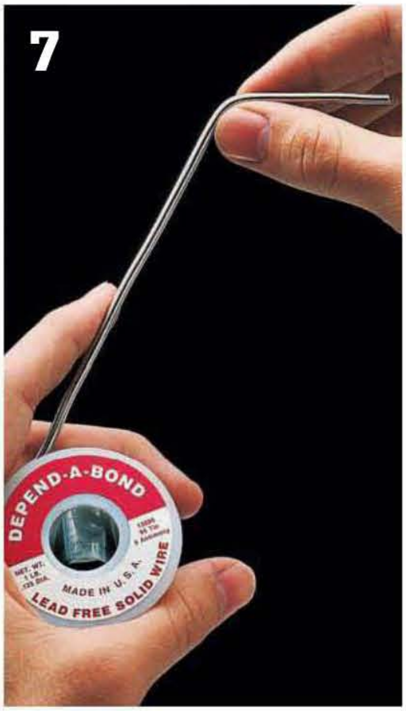
7
DEPEND-A-BOND
MADE IN U.S.A.
LEAD FREE SOLID WIRE

Prepare the wire solder by unwinding 8" to 10" of wire from spool. Bend the first 2" of the wire to a 90º angle.

Open the gas valve and trigger the spark lighter to ignite the torch.

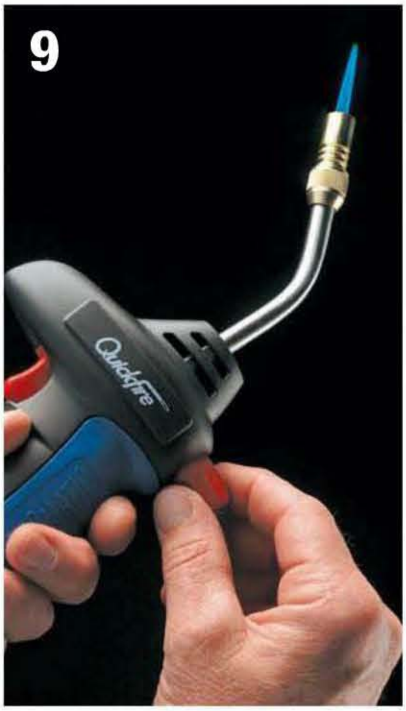
9
Quickfire

Adjust the torch valve until the inner portion of the flame is 1" to 2" long.

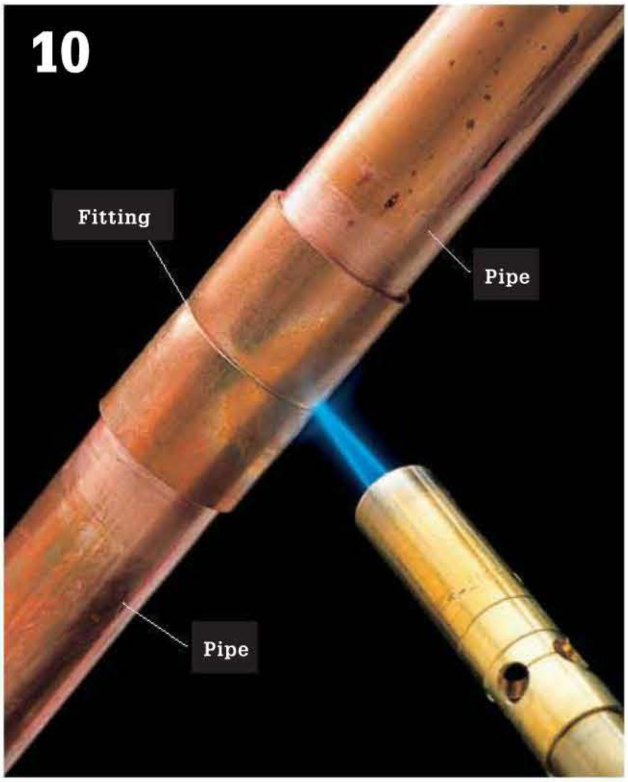
10
Fitting
Pipe
Pipe

Move the torch flame back and forth and around the pipe and the fitting to heat the area evenly.

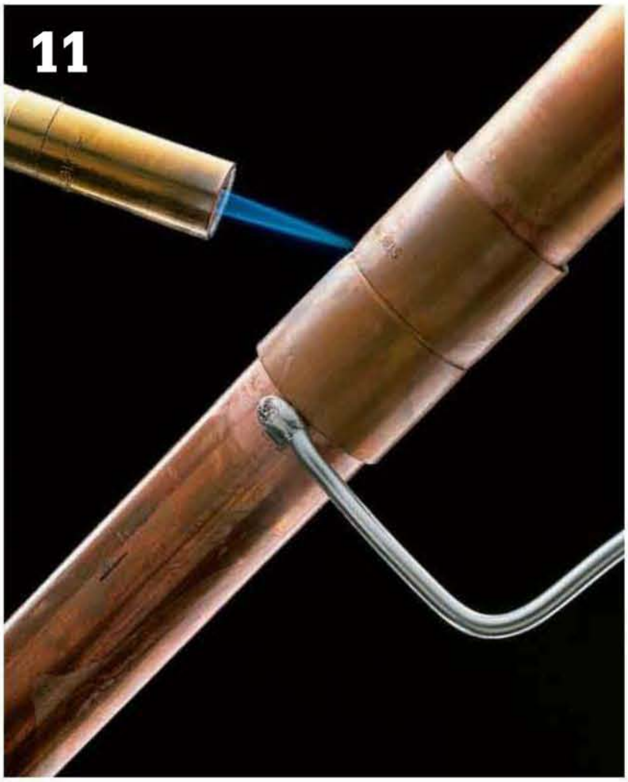

Heat the other side of the copper fitting to ensure that heat is distributed evenly. Touch solder to pipe. Solder will melt when the pipe is at the right temperature.

(continued)

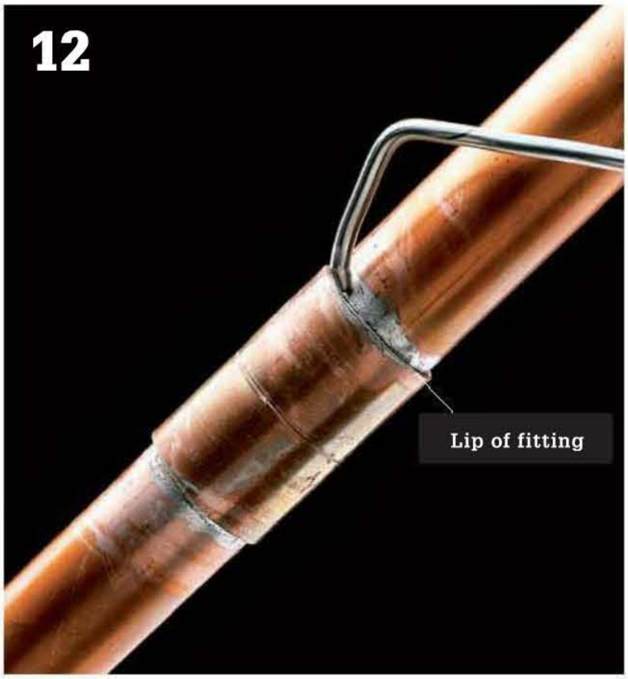

When solder melts, remove the torch and quickly push ½" to ¾" of solder into each joint. Capillary action fills the joint with liquid solder. A correctly soldered joint should show a thin bead of solder around the lips of the fitting.

Allow the joint to cool briefly, then wipe away excess solder with a dry rag. *Caution: Pipes will be hot.* If joints leak after water is turned on, disassemble and resolder.

How to Solder Brass Valves

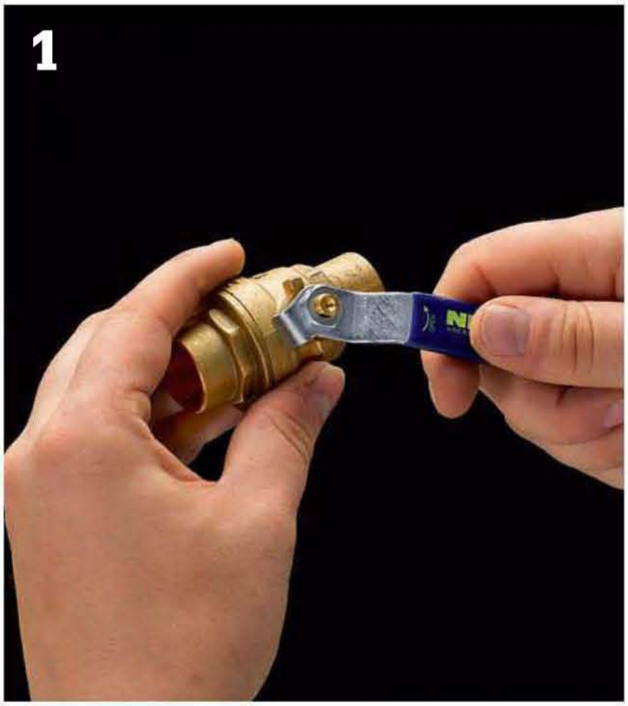

Valves should be fully open during all stages of the soldering process.

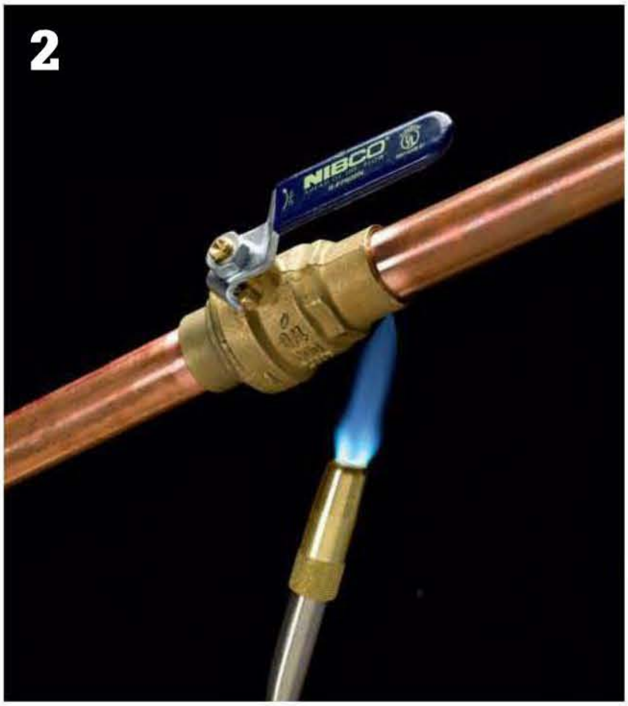

To prevent valve damage, quickly heat the pipe and the flanges of the valve, not the valve body. After soldering, cool the valve by spraying with water.

How to Take Apart Soldered Joints

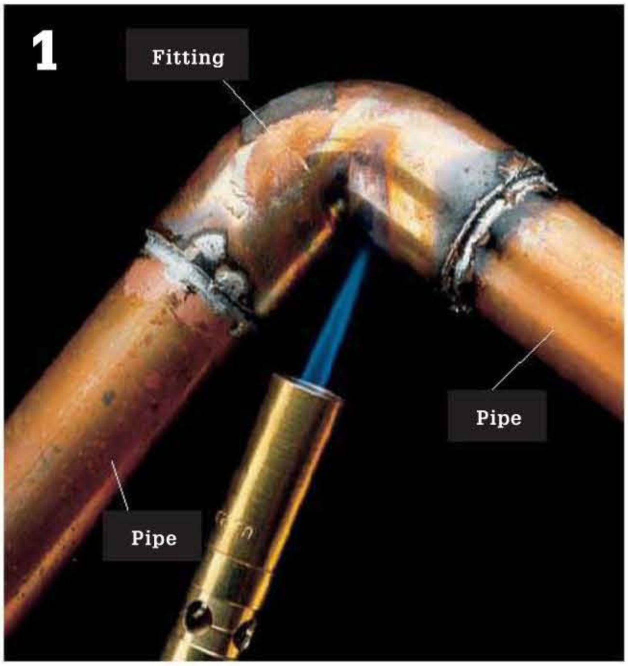

Turn off the water and drain the pipes by opening the highest and lowest faucets in the house. Light your torch. Hold the flame tip to the fitting until the solder becomes shiny and begins to melt.

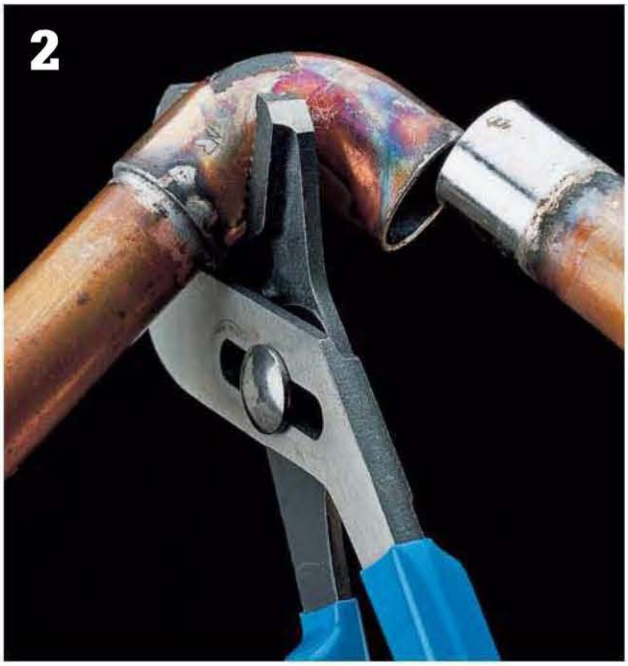

Use channel-type pliers to separate the pipes from the fitting.

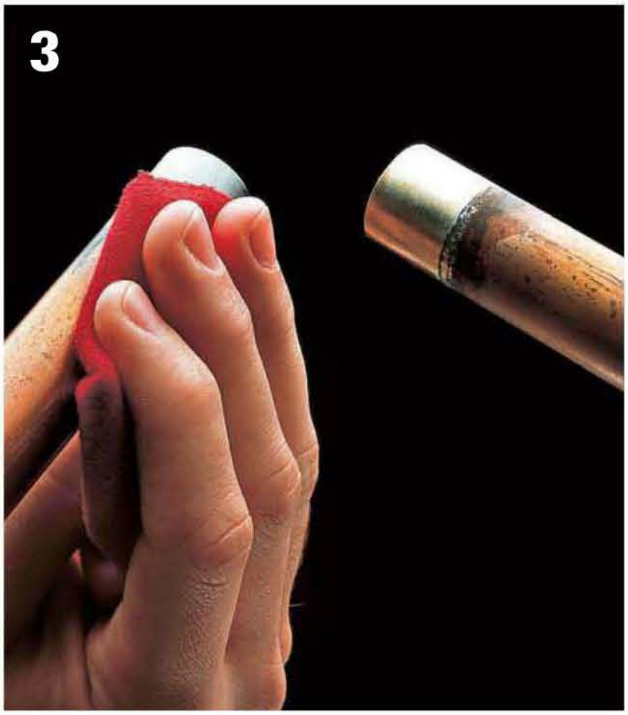

Remove old solder by heating the ends of the pipe with your torch. Use a dry rag to wipe away melted solder quickly. *Caution: Pipes will be hot.*

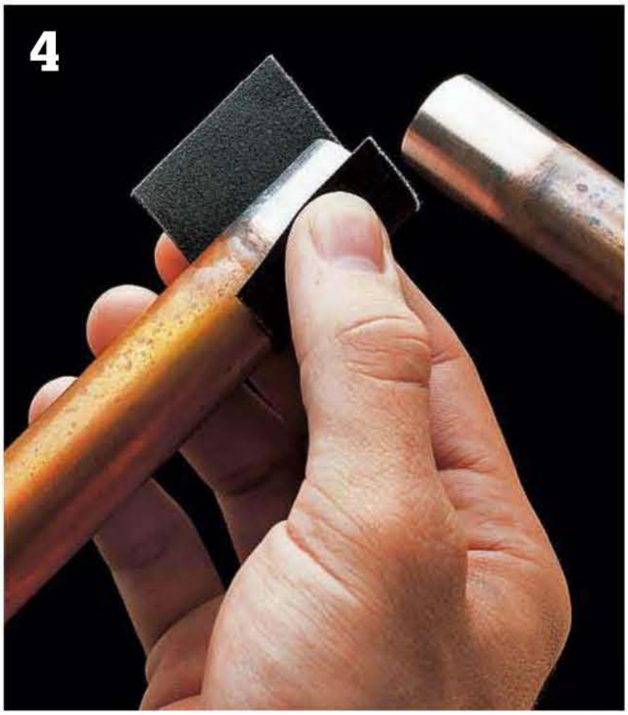

Use emery cloth to polish the ends of the pipe down to bare metal. Never reuse fittings.

Working with Plastic Pipe

Cut rigid ABS, PVC, or CPVC plastic pipes with a tubing cutter or with any saw. Cuts must be straight to ensure watertight joints.

Rigid plastics are joined with plastic fittings and solvent glue. Use a solvent glue that is made for the type of plastic pipe you are installing. For example, use ABS solvent on ABS pipe. Some solvent glues, called "all-purpose" or "universal" solvents, may be used on all types of plastic pipe, but create an inferior bond.

Solvent glue hardens in about 30 seconds, so test-fit all plastic pipes and fittings before gluing the first joint. For best results, the surfaces of plastic pipes and fittings should be dulled with emery cloth and liquid primer before they are joined.

Liquid solvent glues and primers are toxic and flammable. Provide adequate ventilation when fitting plastics, and store the products away from any source of heat.

Plastic grip fittings can be used to join rigid or flexible plastic pipes to copper plumbing pipes.

Tools & Materials ▸

- Tape measure
- Felt-tipped pen
- Tubing cutter
  (or miter box
  or hacksaw)
- Utility knife
- Channel-type pliers
- Gloves
- Plastic pipe
- Fittings
- Emery cloth
- Plastic pipe primer
- Solvent glue
- Rag
- Petroleum jelly

Solvent welding is a chemical bonding process used to permanently join PVC pipes and fittings.

Oatey
CPVC CEMENT
ABS CEMENT
FLEXIBLE PVC CEMENT
CLEANER
PURPLE PRIMER

Primer and solvent glue are specific to the plumbing material being used. Do not use all-purpose or multi-purpose products. Light to medium body glues are appropriate for DIYers as they allow the longest working time and are easiest to use. The products work best when fresh, so buy small containers and dispose of any unused product after a few months.

How to Cut Rigid Plastic Pipe

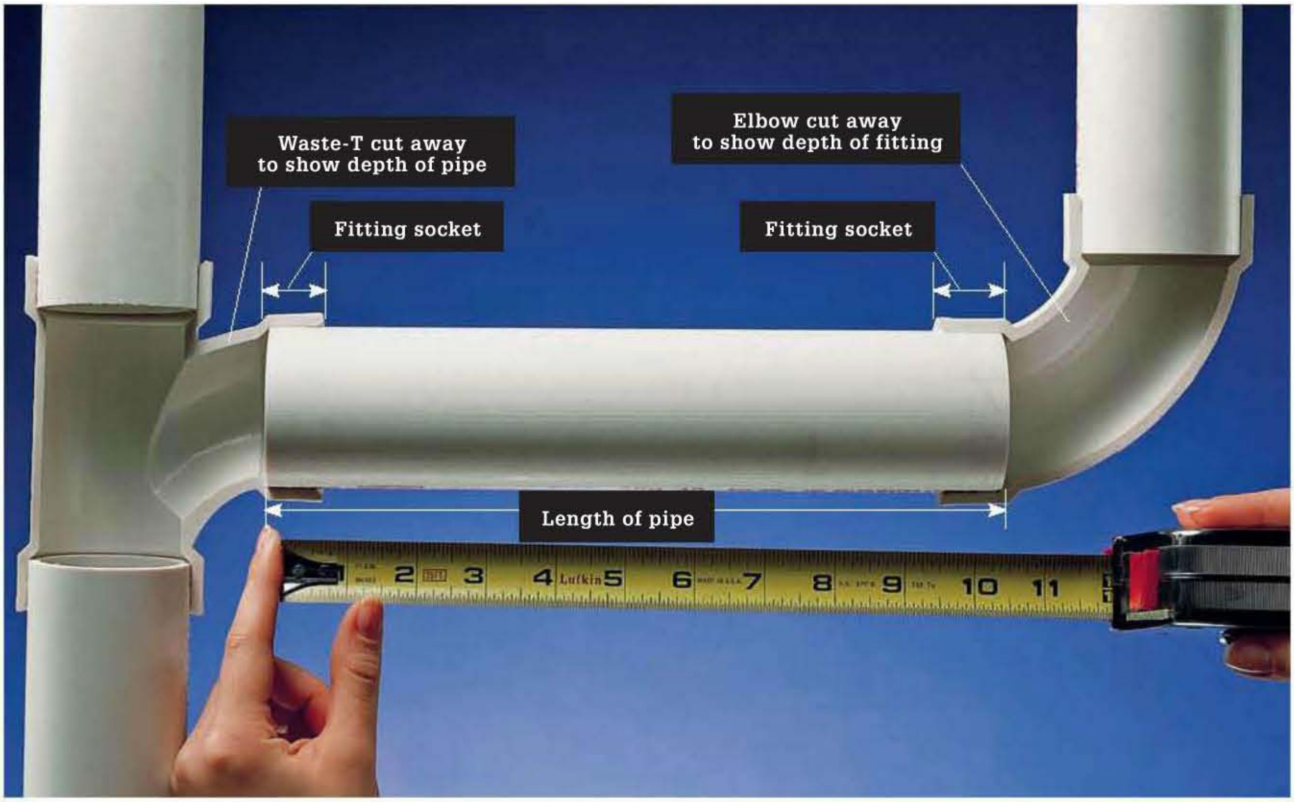
Waste-T cut away to show depth of pipe
Elbow cut away to show depth of fitting
Fitting socket
Fitting socket
Length of pipe

Find the length of plastic pipe needed by measuring between the bottoms of the fitting sockets (fittings shown in cutaway). Mark the length on the pipe with a felt-tipped pen.

The best cutting tool for plastic pipe is a power miter saw with a fine tooth woodworking blade or a plastic-specific blade.

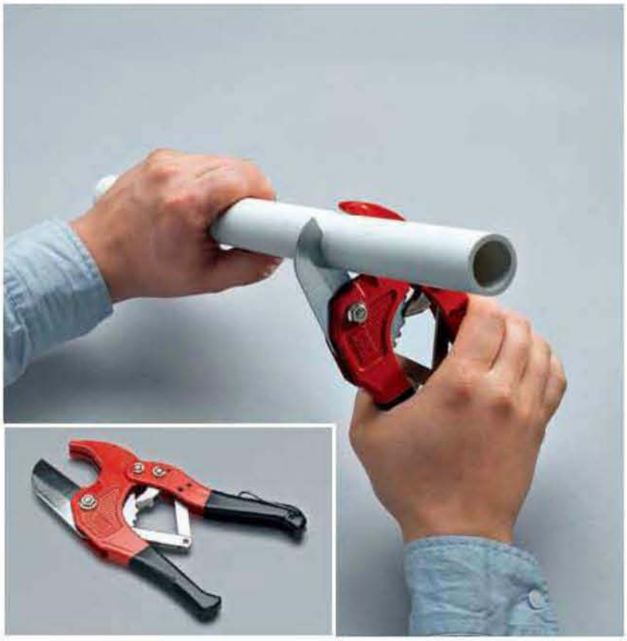

A ratcheting plastic-pipe cutter can cut smaller diameter PVC and CPVC pipe in a real hurry. If you are plumbing a whole house you may want to consider investing in one. They also are sold only at plumbing supply stores.

How to Solvent-Glue Rigid Plastic Pipe

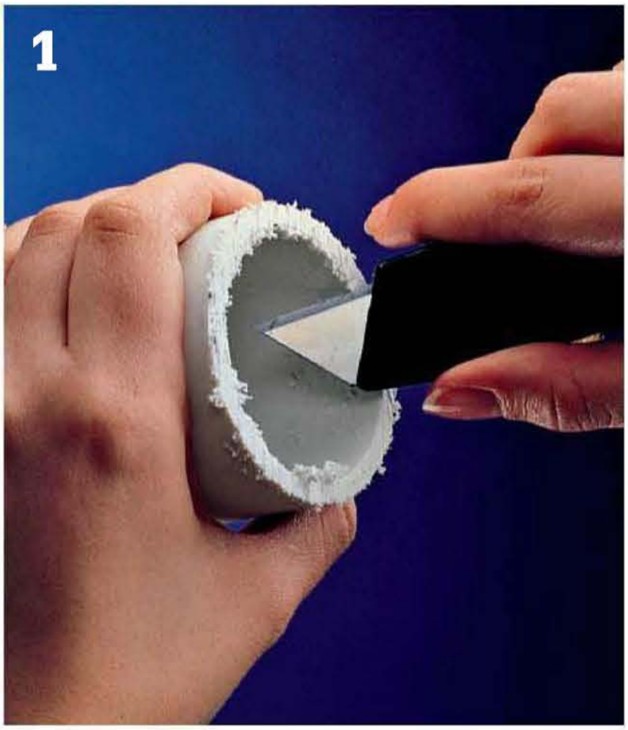

Remove rough burrs on cut ends of plastic pipe, using a utility knife.

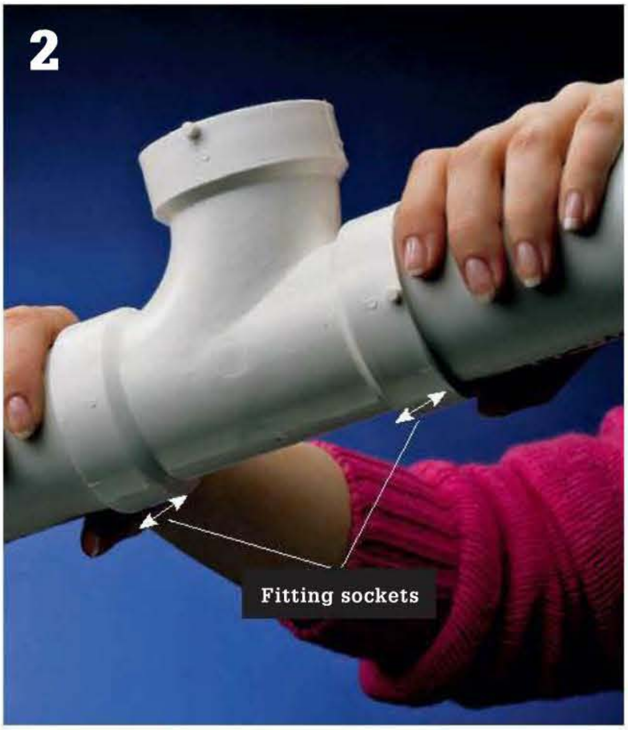

Test-fit all pipes and fittings. Pipes should fit tightly against the bottom of the fitting sockets.

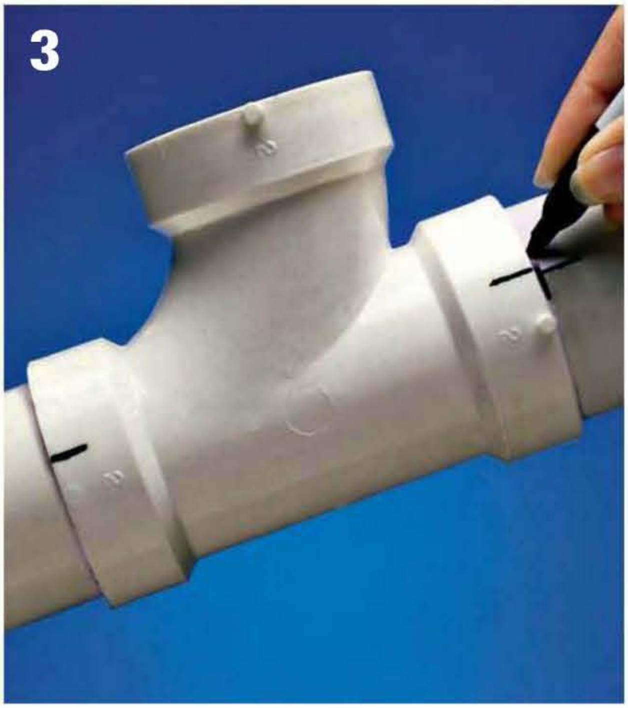

Mark the depth of the fitting sockets on the pipes. Take pipes apart. Clean the ends of the pipes and fitting sockets with emery cloth.

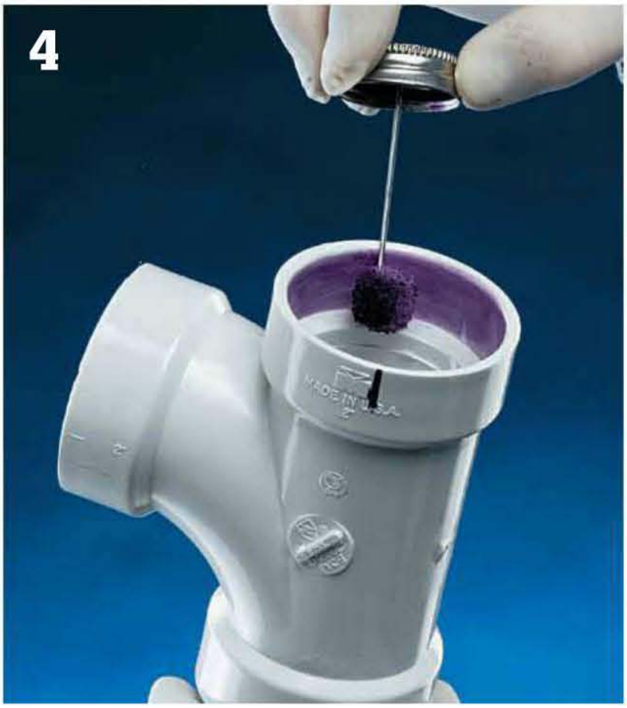

Apply a light coat of plastic pipe primer to the ends of the pipes and to the insides of the fitting sockets. Primer dulls glossy surfaces and ensures a good seal.

5

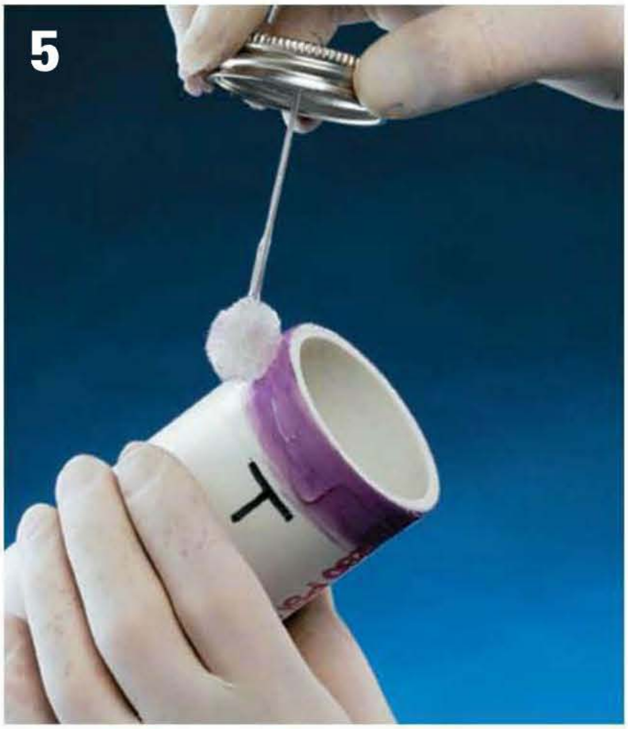

Solvent-glue each joint by applying a thick coat of solvent glue to the end of the pipe. Apply a thin coat of solvent glue to the inside surface of the fitting socket. Work quickly: solvent glue hardens in about 30 seconds.

6

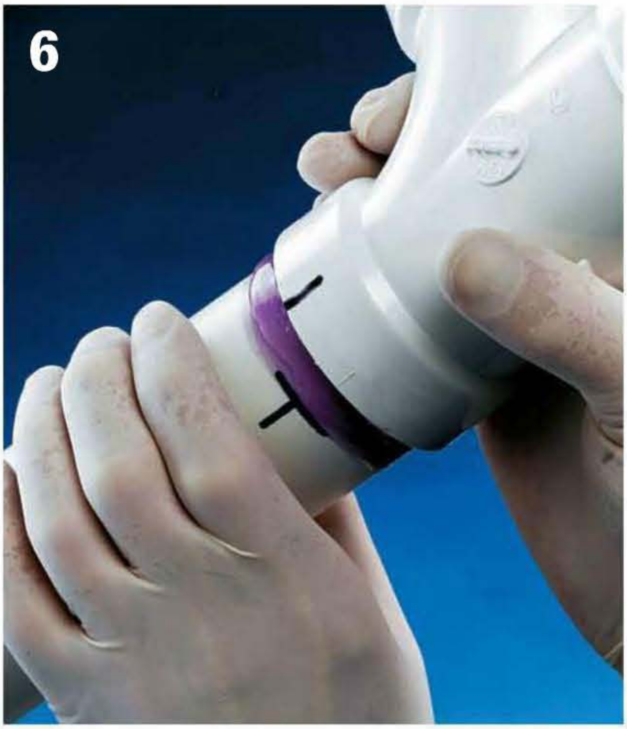

Quickly position the pipe and fitting so that the alignment marks are offset by about 2". Force the pipe into the fitting until the end fits flush against the bottom of the socket.

7

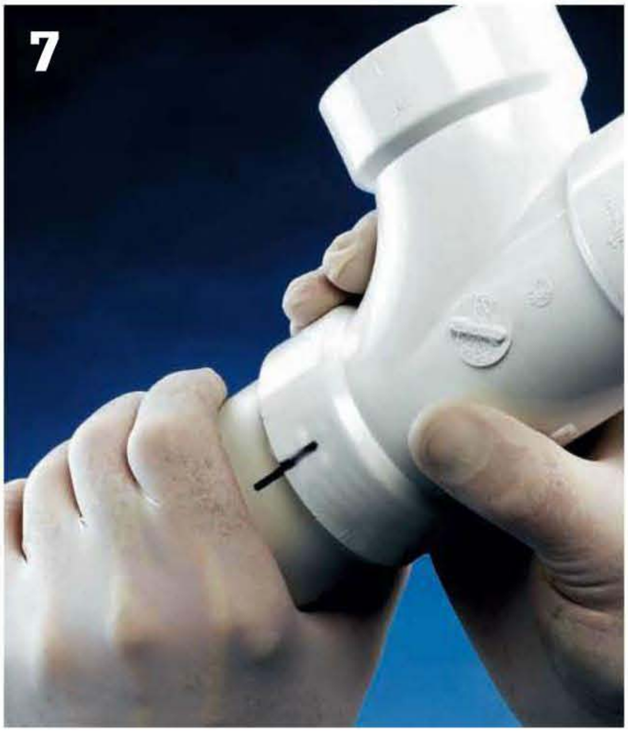

Spread solvent by twisting the pipe until the marks are aligned. Hold the pipe in place for about 20 seconds to prevent the joint from slipping.

8

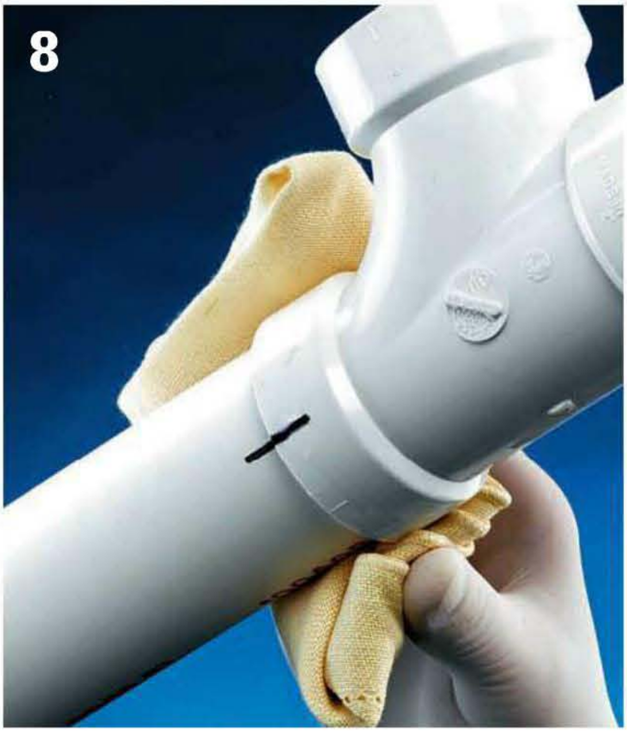

Wipe away excess solvent glue with a rag. Do not disturb the joint for 30 minutes after gluing.

Working with PEX Pipe

Cross-linked polyethylene (PEX) is growing quickly in acceptance as a supply pipe for residential plumbing. It's not hard to understand why. Developed in the 1960s but relatively new to the United States, this supply pipe combines the ease of use of flexible tubing with the durability of rigid pipe. It can withstand a wide temperature range (from subfreezing to 180° F); it is inexpensive; and it's quieter than rigid supply pipe.

PEX is flexible plastic (polyethylene, or PE) tubing that's reinforced by a chemical reaction that creates long fibers to increase the strength of the material. It has been allowed by code in Europe and the southern United States for many years, but has won approval for residential supply use in most major plumbing codes only recently. It's frequently used in manufactured housing and recreational vehicles and in radiant heating systems. Because it is so flexible, PEX can easily be bent to follow corners and make other changes in direction. From the water main and heater, it is connected into manifold fittings that redistribute the water in much the same manner as a lawn irrigation system.

For standard residential installations, PEX can be joined with very simple fittings and tools. Unions are generally made with a crimping tool and a crimping ring. You simply insert the ends of the pipe you're joining into the ring, then clamp down on the ring with the crimping tool. PEX pipe, tools, and fittings can be purchased from most wholesale plumbing suppliers and at many home centers. Coils of PEX are sold in several diameters from ¼" to 1". PEX tubing and fittings from different manufacturers are not interchangeable. Any warranty coverage will be voided if products are mixed.

Tools & Materials ▸

- Tape measure
- Felt-tipped pen
- Full-circle crimping tool
- Go/no-go gauge
- Tubing cutter
- PEX pipe
- Manifolds
- Protector plates
- PEX fittings
- Utility knife
- Plastic hangers
- Crimp ring

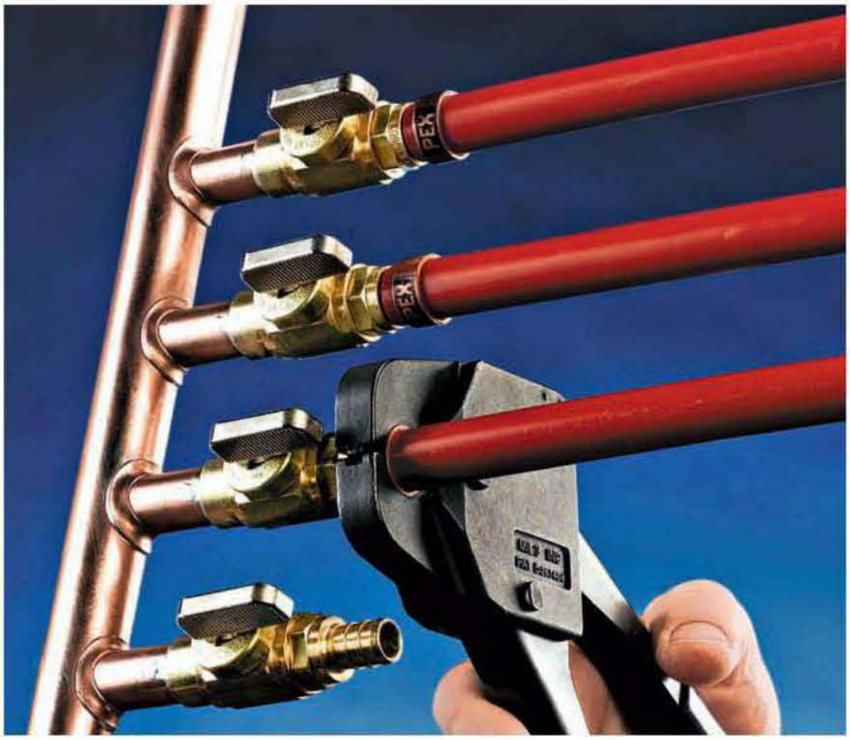

PEX pipe is a relatively new water supply material that's growing in popularity, in part because it can be installed with simple mechanical connections. In this example, we see a standard 4-outlet copper manifold, with branch water lines being attached with a crimping tool. To ensure leak-free joints, each joint is carefully measured with a go/no-go tool before the water is turned on.

PEX Tools & Materials

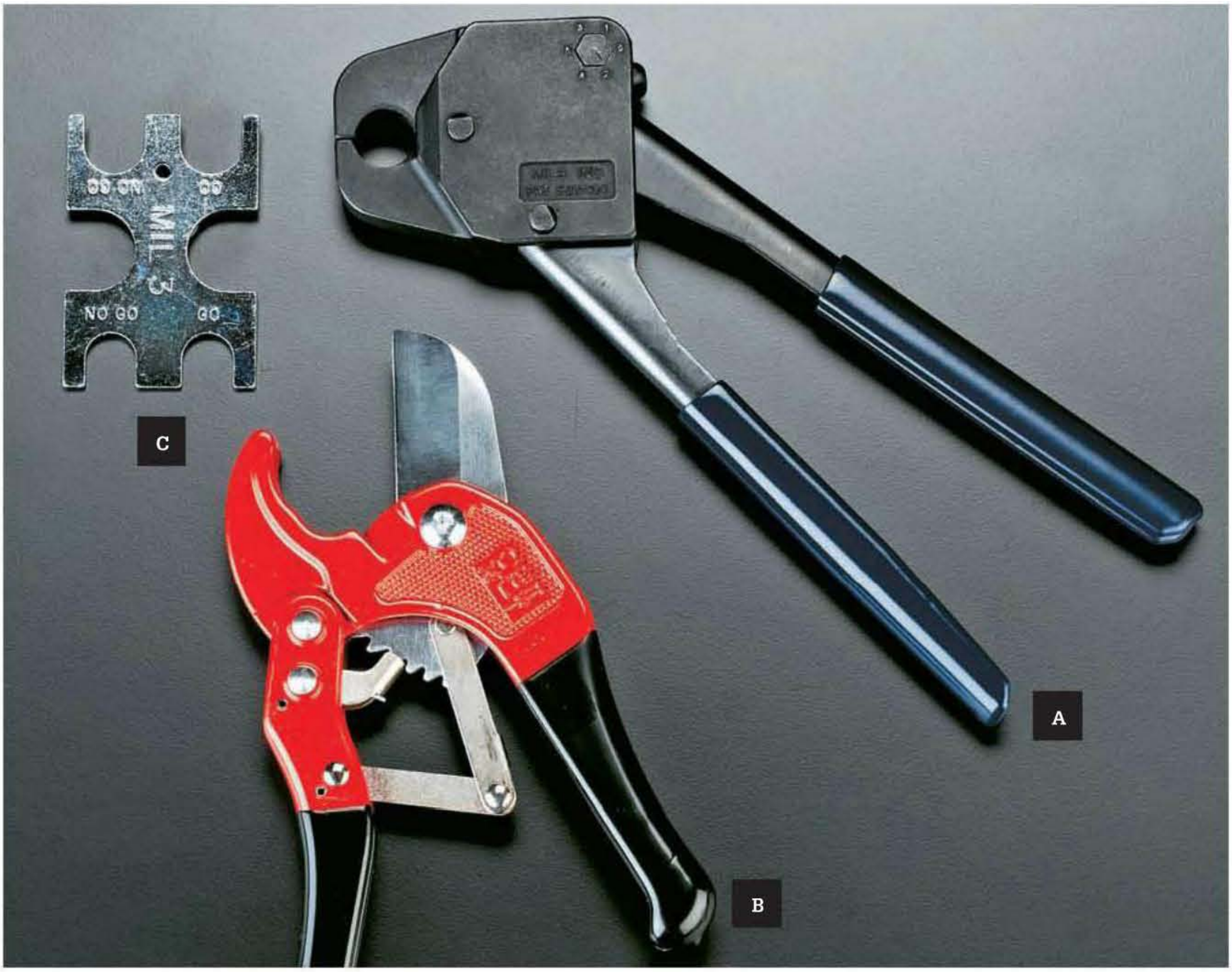

Specialty tools for installing PEX are available wherever PEX is sold. The basic set includes a full-circle crimping tool (A), a tubing cutter (B), and a go/no-go gauge (C) to test connections after they've been crimped.

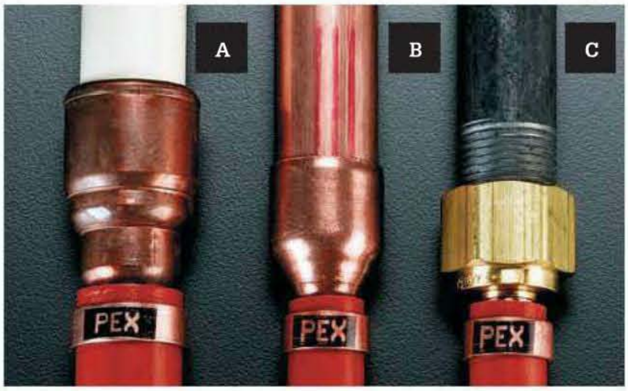

PEX is connected to other water supply materials with transition fittings, including CPVC-to-PEX (A), copper-to-PEX (B), and iron-to-PEX (C).

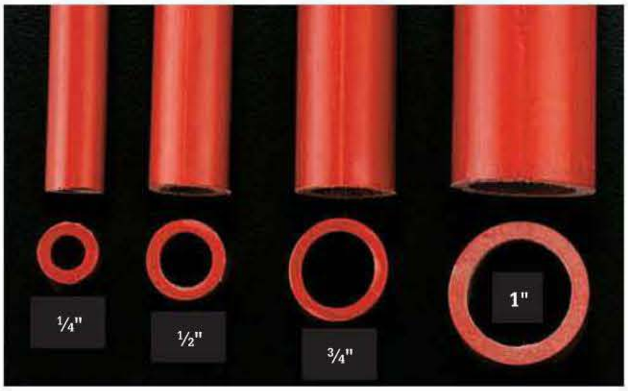

Generally, you should use the same diameter PEX as is specified for rigid supply tubing, but in some "home run" installations (see page 72) you can use ⅜" PEX where ½" rigid copper would normally be used.

PEX Installation

Check with your local plumbing inspector to verify that PEX is allowed in your municipality. PEX has been endorsed by all major plumbing codes in North America, but your municipality may still be using an older set of codes. Follow the guidelines below when installing PEX:

- Do not install PEX in above-ground exterior applications because it degrades quickly from UV exposure.
- Do not use PEX for gas lines.
- Do not use plastic solvents or petroleum-based products with PEX (they can dissolve the plastic).
- Keep PEX at least 12" away from recessed light fixtures and other potential sources of high heat.
- Do not attach PEX directly to a water heater. Make connections at the heater with metallic tubing (either flexible water-heater connector tubing or rigid copper) at least 18" long; then join it to PEX with a transition fitting.
- Do not install PEX in areas where there is a possibility of mechanical damage or puncture. Always fasten protective plates to wall studs that house PEX.
- Always leave some slack in installed PEX lines to allow for contraction and in case you need to cut off a bad crimp.
- Use the same minimum branch and distribution supply-pipe dimensions for PEX that you'd use for copper or CPVC, according to your local plumbing codes.

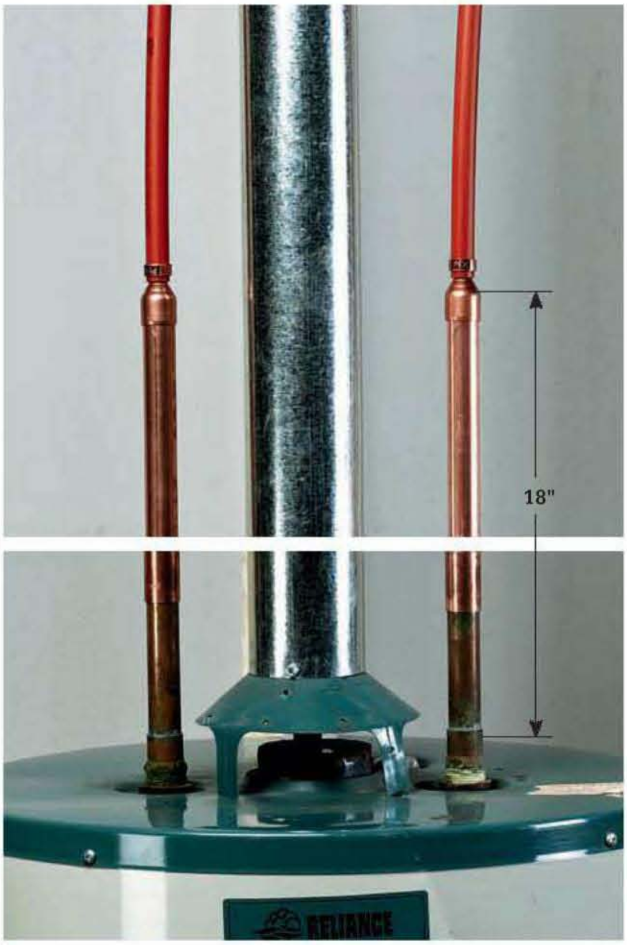

Do not connect PEX directly to a water heater. Use metal connector tubes. Solder the connector tubes to the water heater before attaching PEX. Never solder metal tubing that is already connected to PEX lines.

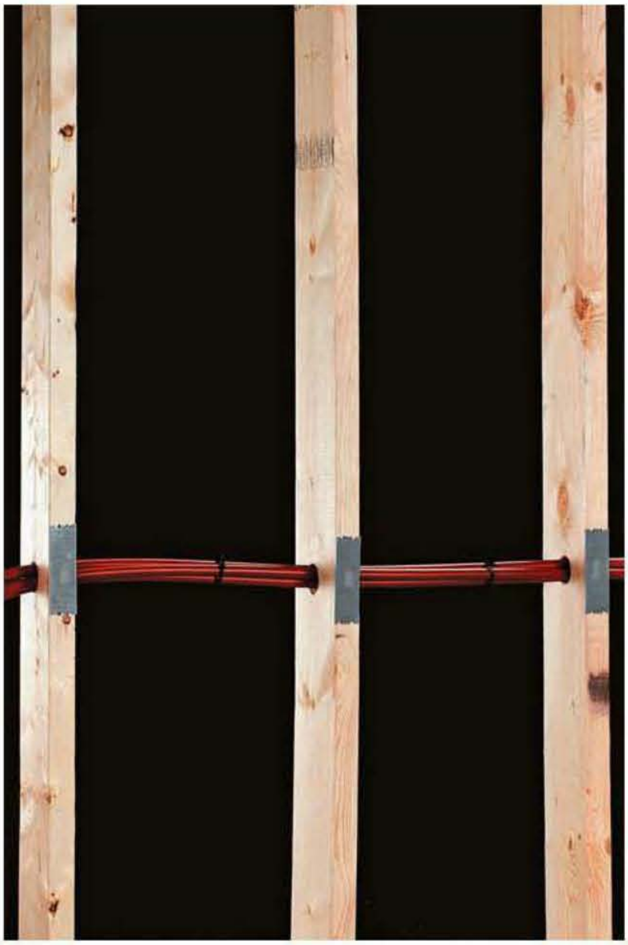

Bundle PEX together with plastic ties when running pipe through wall cavities. PEX can contract slightly, so leave some slack in the lines.

Buying PEX

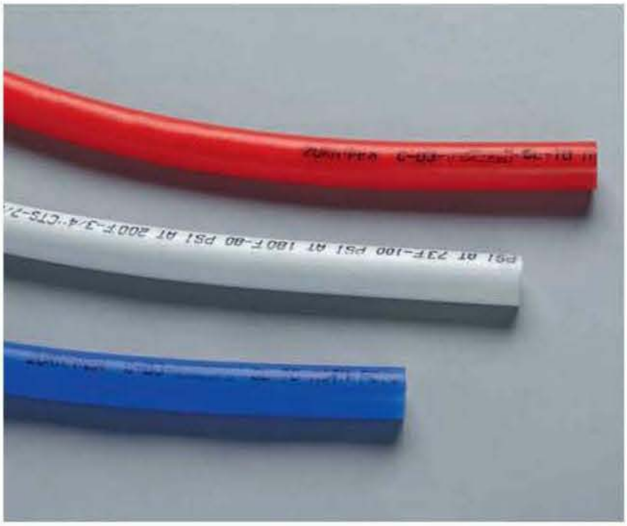

Color coding is a practice many PEX manufacturers have embraced to make identification easier. Because the material is identical except for the color, you can buy only one color (red is more common) and use it for both hot and cold supply lines.

PEX combines the flexibility of plastic tubing with the durability of rigid supply pipe. It is sold in coils of common supply-pipe diameters.

The PEX Advantage ▸

PEX supply tubing offers a number of advantages over traditional rigid supply tubing:

- Easy to install. PEX does not require coupling joints for long runs or elbows and sweeps for turns. The mechanical connections do not require solvents or soldering.
- Easy to transport. Large coils are lightweight and much easier to move around than 10-ft. lengths of pipe.
- Good insulation. The PEX material has better thermal properties than copper for lessened heat loss.
- Quiet. PEX will not rattle or clang from trapped air or kinetic energy.
- Good for retrofit jobs. PEX is easier to snake through walls than rigid supply tubing and is compatible with copper, PVC, or iron supply systems if the correct transition fittings are used. If your metal supply tubes are used to ground your electrical system, you'll need to provide a jumper if PEX is installed in midrun. Check with a plumber or electrician.
- Freeze resistance. PEX retains some flexibility in sub-freezing conditions and is less likely to be damaged than rigid pipe, but it is not frostproof.

General Codes for PEX ▸

PEX has been endorsed for residential use by all major building codes, although some municipal codes may be more restrictive. The specific design standards may also vary, but here are some general rules:

- For PEX, maximum horizontal support spacing is 32" and maximum vertical support spacing is 10 ft.
- Maximum length of individual distribution lines is 60 ft.
- PEX is designed to withstand 210° F water for up to 48 hours. For ongoing use, most PEX is rated for 180 degree water up to 100 pounds per square inch of pressure.
- Directional changes of more than 90 degrees require a guide fitting (see page 75).

System Designs

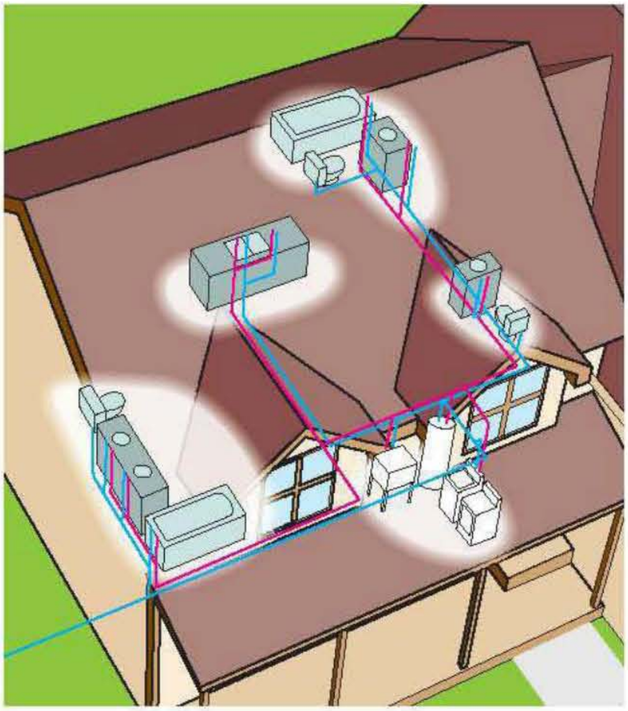

Trunk-and-branch systems are configured in much the same way as a traditional rigid copper or PVC supply systems. A main supply line (the trunk line) carries water to all of the outlets via smaller branch lines that tie into the trunk and serve a few outlets in a common location.

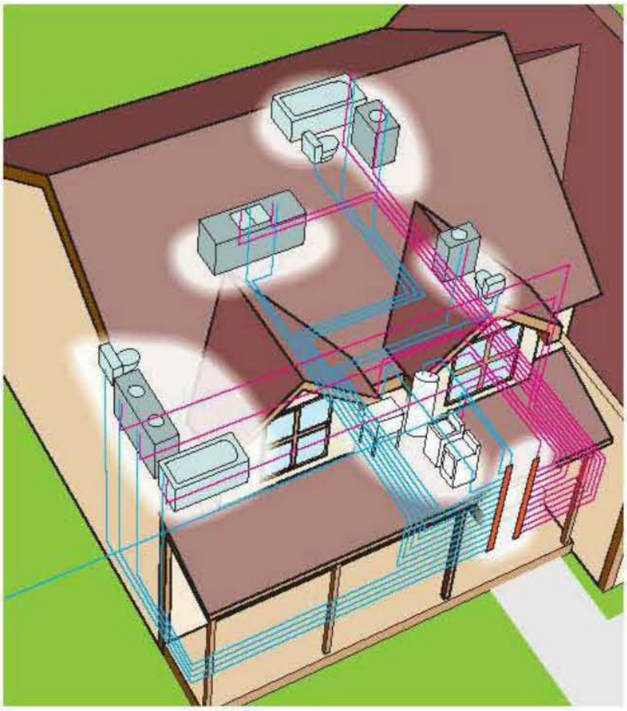

Home run systems rely on one or two central manifolds to distribute the hot and cold water very efficiently. Eliminating the branch fittings allows you to use thinner supply pipe in some situations.

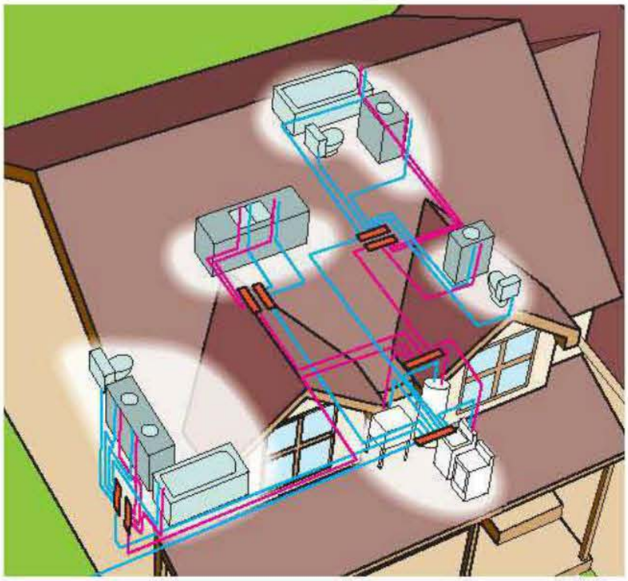

Remote manifold systems are a hybrid between traditional trunk-and-branch systems and home run systems. Instead of relying on just one or two manifolds, they employ several smaller manifolds downline from a larger manifold. Each smaller manifold services a group of fixtures, as in a bathroom or kitchen.

Choosing a PEX system ▸

- For maximum single-fixture water pressure: Trunk and branch
- For economy of materials: Trunk and branch or remote manifold
- For minimal wait times for hot water (single fixture): Home run
- For minimal wait times for hot water (multiple fixtures used at same approximate time): Trunk and branch or remote manifold
- For ease of shutoff control: Home run
- For lowest number of fittings and joints: Home run

How to Make PEX Connections

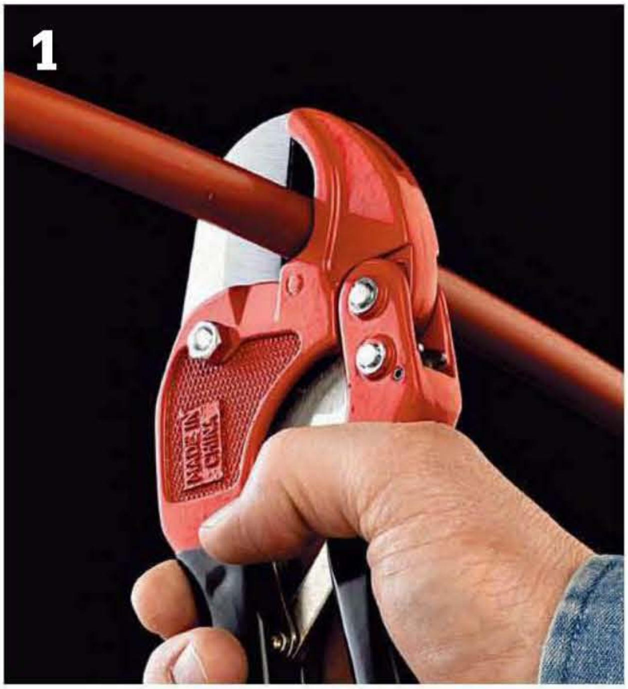

Cut the pipe to length, making sure to leave enough extra material so the line will have a small amount of slack once the connections are made. A straight, clean cut is very important. For best results, use a tubing cutter.

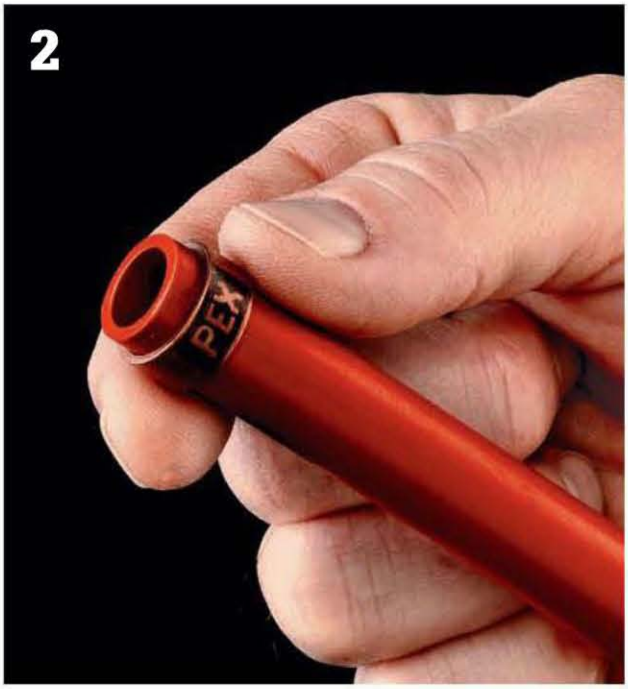

Inspect the cut end to make sure it is clean and smooth. If necessary, deburr the end of the pipe with a sharp utility knife. Slip a crimp ring over the end.

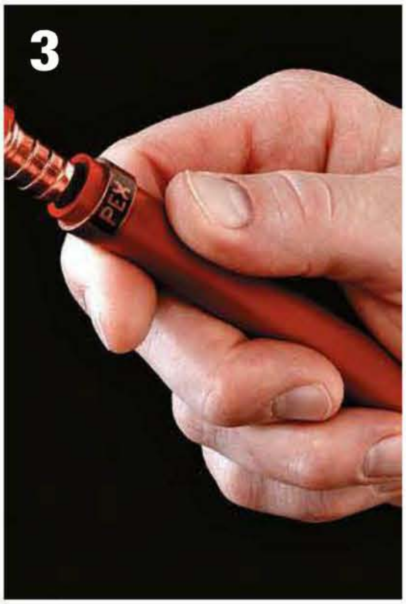

Insert the barbed end of the fitting into the pipe until it is snug against the cut edges.Position the crimp ring so it is ⅛" to ¼" from the end of the pipe, covering the barbed end of the fitting. Pinch the fitting to hold it in place.

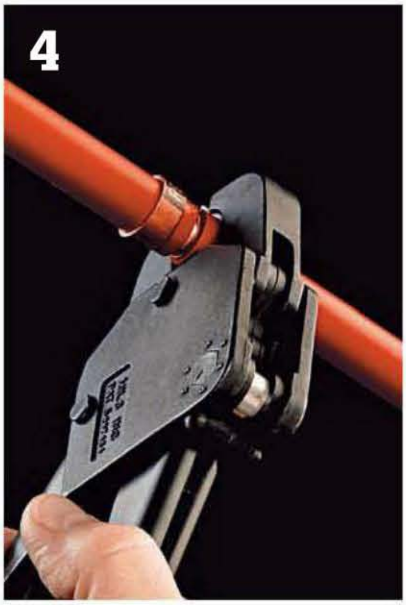

Align the jaws of a full-circle crimping tool over the crimp ring and squeeze the handles together to apply strong, even pressure to the ring.

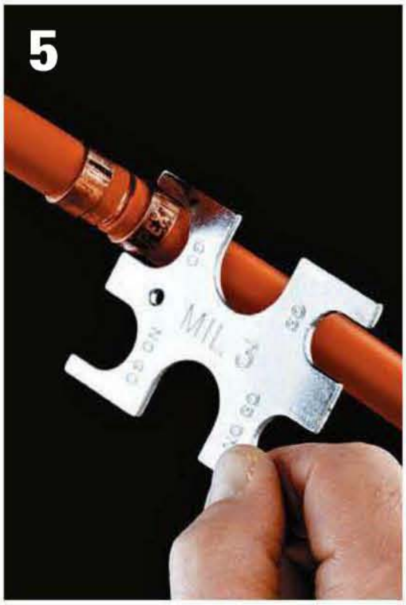

Test the connection to make sure it is mechanically acceptable, using a go/no-go gauge. If the ring does not fit into the gauge properly, cut the pipe near the connection and try again.

How to Plumb a PEX Water-Supply System

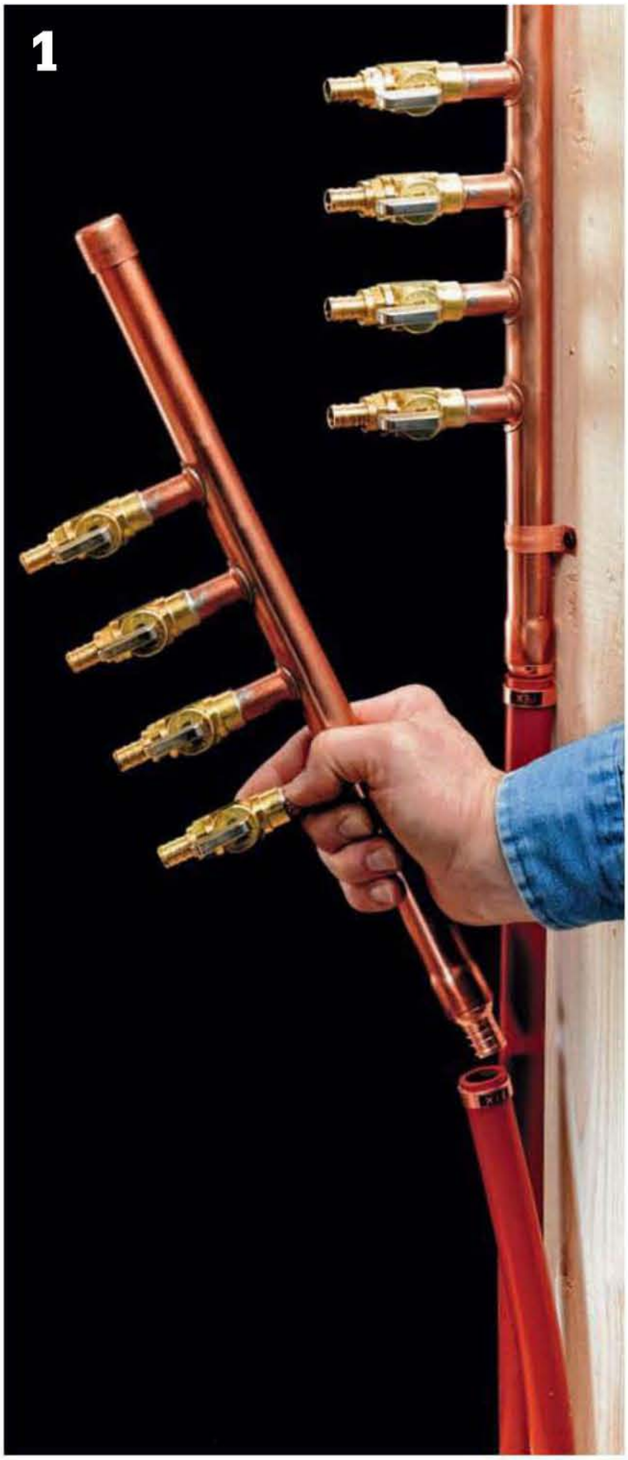

Install copper manifolds (one for hot and one for cold) in an accessible location central to the fixtures. The manifold should have one outlet for each supply line it will support (fixtures that require hot and cold supply will need a separate outlet for each). Run supply lines from the water heater and water main to the copper manifolds. Connect the supply pipes to the manifolds with crimp fittings.

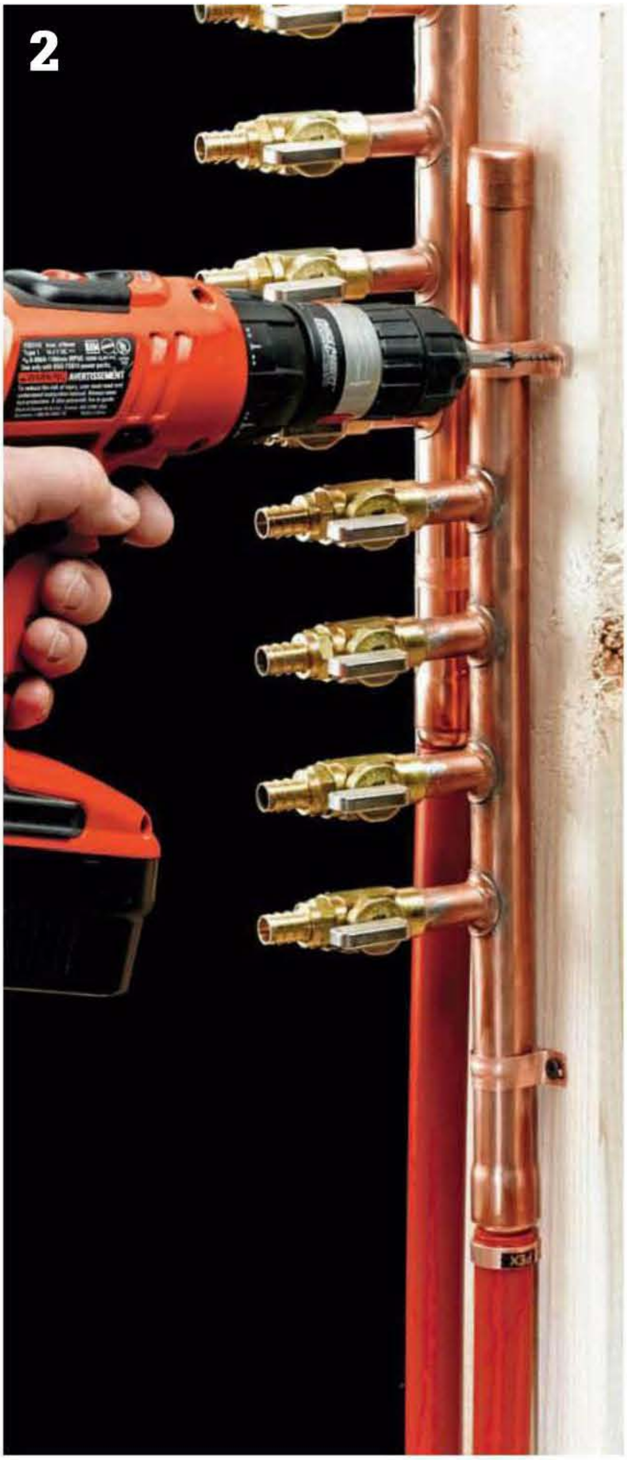

A manifold may be attached vertically or horizontally, but it must be anchored with correctly sized hangers screwed to the framing members.

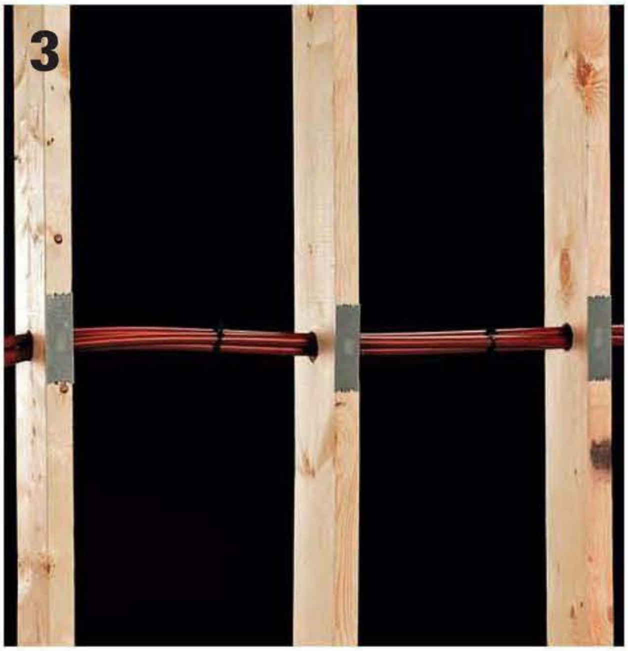

Starting at each fixture (and leaving at least 12" of extra pipe exposed), run appropriately sized PEX through holes in the framing to the manifolds. Pipes may be bundled together loosely with plastic ties. Protect the line with a nailing plate at each stud location. Be sure to leave some slack in the supply lines.

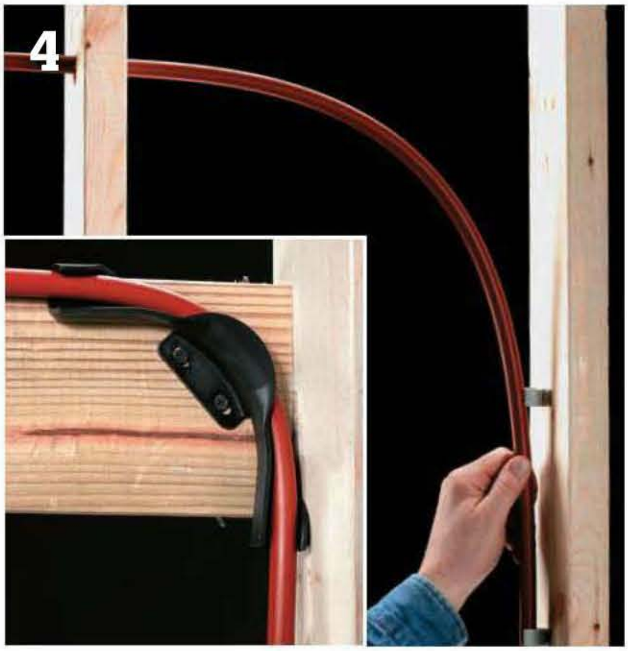

Support the pipe with a plastic hanger near every floor or ceiling and midway up vertical runs. Also use hangers to guide pipe near the beginnings and ends of curves and near fittings. Use a plastic guide for sharp curves (inset). Do not bend PEX so sharply that it kinks.

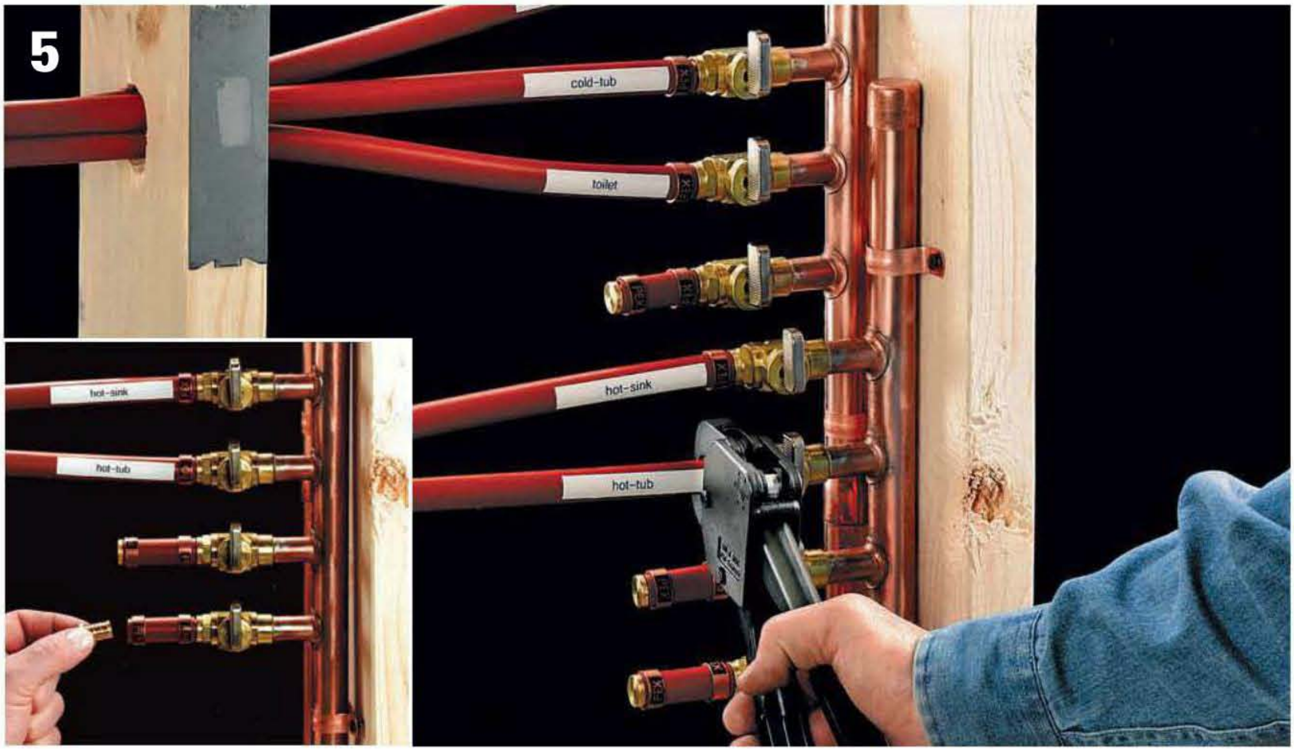

Cut each branch supply line to length (leave some extra in case you need to recrimp). Install shutoff valves for each outlet (most manifolds come with preattached valves). Connect the PEX branch supply lines to the shutoff valves. Label each pipe. Use a short length of PEX and a plug to seal any unused outlets (inset).

Planning New Plumbing

Start planning by drawing maps. Mapping your home's plumbing system is a good way to familiarize yourself with the plumbing layout and can help you when planning plumbing renovation projects. With a good map, you can envision the best spots for new fixtures and plan new pipe routes more efficiently. Maps also help in emergencies, when you need to locate burst or leaking pipes quickly.

Draw a plumbing map for each floor on tracing paper, so you can overlay floors and still read the information below. Make your drawings to scale and have all plumbing fixtures marked. Fixture templates and tracing paper are available at drafting supply stores.

Snoop around your basement for clues about the locations of supply, drain, vent, and gas pipes in your walls.

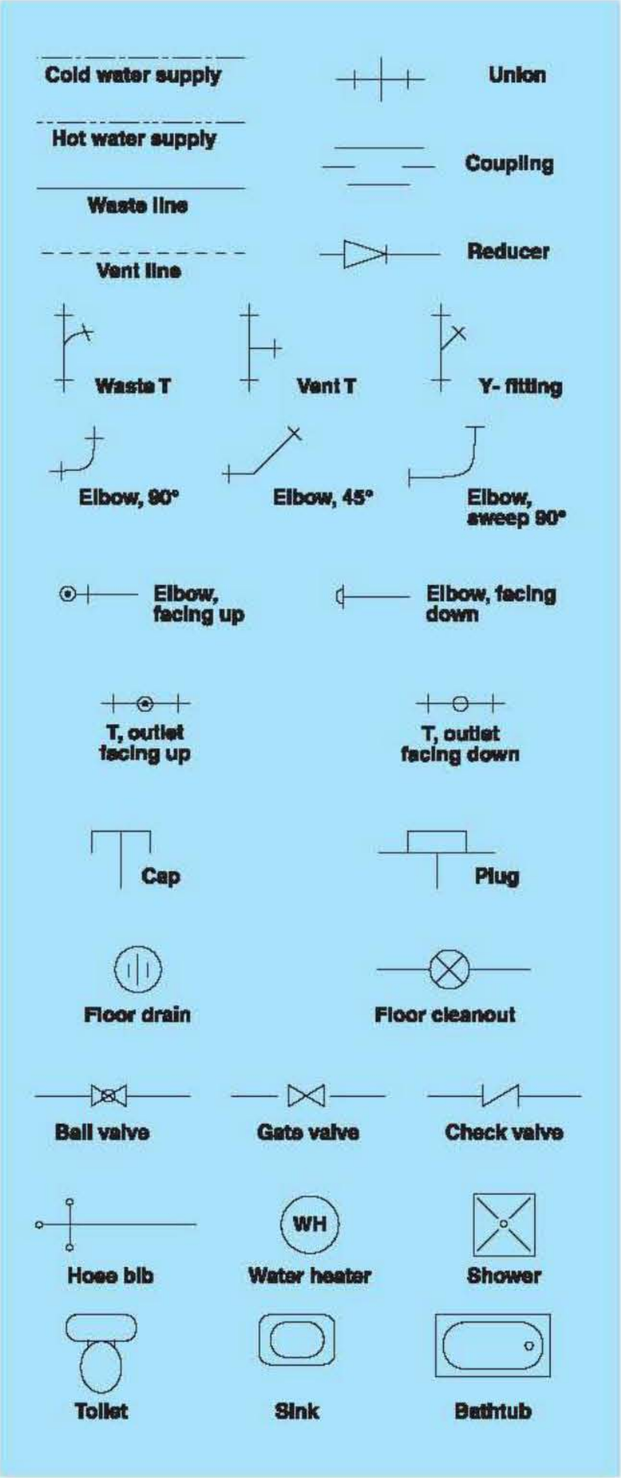
Cold water supply
Union
Hot water supply
Coupling
Waste line
Reducer
Vent line
Waste T
Vent T
Y- fitting
Elbow, 90°
Elbow, 45°
Elbow, sweep 90°
Elbow, facing up
Elbow, facing down
T, outlet facing up
T, outlet facing down
Cap
Plug
Floor drain
Floor cleanout
Ball valve
Gate valve
Check valve
Hose bib
WH
Water heater
Shower
Toilet
Sink
Bathtub

Use standard plumbing symbols on your map to identify the components of your plumbing system. These symbols will help you and your building inspector follow connections and transitions more easily.

Understanding Plumbing Codes

The plumbing code is the set of regulations that building officials and inspectors use to evaluate your project plans and the quality of your work. Codes vary from region to region, but most are based on the National Uniform Plumbing Code, the authority we used in the development of this book.

Code books are available for reference at bookstores and government offices. However, they are highly technical, difficult-to-read manuals. More user-friendly for do-it-yourselfers are the variety of code handbooks available at bookstores and libraries. These handbooks are based on the National Uniform Plumbing Code but are easier to read and include many helpful diagrams and photos.

Plumbing code handbooks sometimes discuss three different plumbing "zones" in an effort to accommodate variations in regulations from state to state. The states included in each zone are listed below.

Zone 1: Washington, Oregon, California, Nevada, Idaho, Montana, Wyoming, North Dakota, South Dakota, Minnesota, Iowa, Wisconsin, Nebraska, Kansas, Utah, Arizona, Colorado, New Mexico, Indiana, parts of Texas.

Zone 2: Alabama, Arkansas, Louisiana, Tennessee, North Carolina, Mississippi, Georgia, Florida, South Carolina, parts of Texas, parts of Maryland, parts of Delaware, parts of Oklahoma, parts of West Virginia.

Zone 3: Virginia, Kentucky, Missouri, Illinois, Michigan, Ohio, Pennsylvania, New York, Connecticut, Massachusetts, Vermont, New Hampshire, Rhode Island, New Jersey, parts of Delaware, parts of West Virginia, parts of Maine, parts of Maryland, parts of Oklahoma.

Remember that your local plumbing code always supersedes the national code. Local codes may be more restrictive than the national code. Your local building inspector is a valuable source of information and may provide you with a convenient summary sheet of the regulations that apply to your project.

INSPECTION RECORD
NOTICE
NATIONAL PLUMBING CODES HANDBOOK
UNIFORM BUILDING CODE

The plumbing inspector is the final authority when it comes to evaluating your work. By visually examining and testing your new plumbing, the inspector ensures that your work is safe and functional.

GETTING A PERMIT

To ensure public safety, your community requires that you obtain a permit for most plumbing projects, including most of the projects demonstrated in this book.

When you visit your city building inspection office to apply for a permit, the building official will want to review three drawings of your plumbing project: a site plan, a water supply diagram, and a drain-waste-vent diagram. These drawings are described on this page. If the official is satisfied that your project meets code requirements, he or she will issue you a plumbing permit, which is your legal permission to begin work. The building official also will specify an inspection schedule for your project. As your project nears completion, you will be asked to arrange for an inspector to visit your home while the pipes are exposed to review the installation and ensure its safety.

Although do-it-yourselfers often complete complex plumbing projects without obtaining a permit or having the work inspected, we strongly urge you to comply with the legal requirements in your area. A flawed plumbing system can be dangerous, and it can potentially threaten the value of your home.

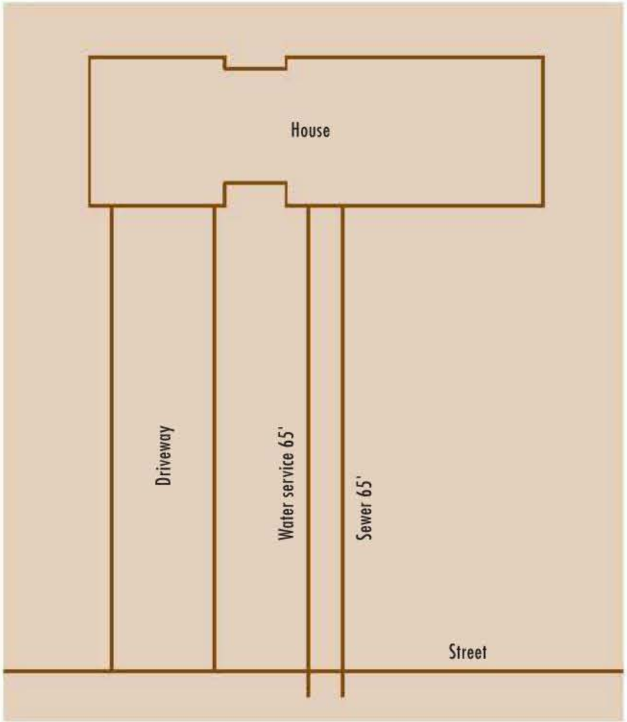
House
Driveway
Water service 65'
Sewer 65'
Street

The site plan shows the location of the water main and sewer main with respect to your yard and home. The distances from your foundation to the water main and from the foundation to the main sewer should be indicated on the site plan.

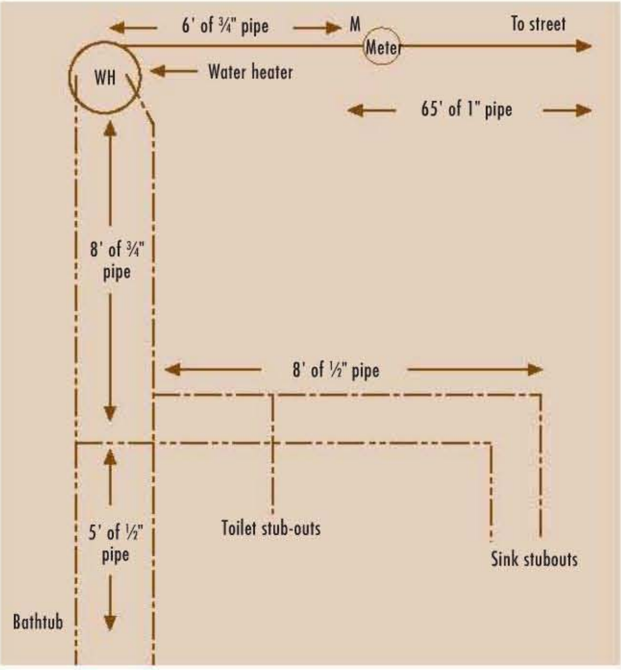
6' of ¾" pipe
M
Meter
To street
WH
Water heater
65' of 1" pipe
8' of ¾" pipe
8' of ½" pipe
Toilet stub-outs
Sink stubouts
5' of ½" pipe
Bathtub

The supply riser diagram shows the length of the hot and cold water pipes and the relation of the fixtures to one another. The inspector will use this diagram to determine the proper size for the new water supply pipes in your system.

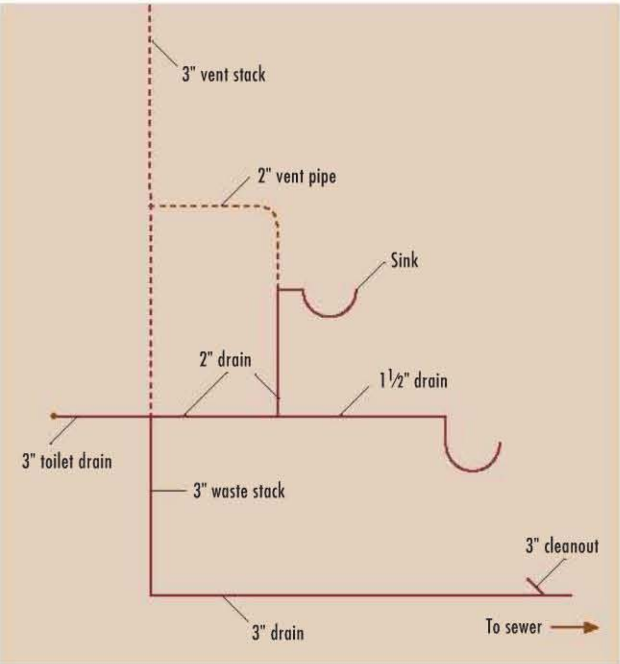
3" vent stack
2" vent pipe
Sink
2" drain
1½" drain
3" toilet drain
3" waste stack
3" cleanout
3" drain
To sewer

A DWV diagram shows the routing of drain and vent pipes in your system. Make sure to indicate the lengths of drain pipes and the distances between fixtures. The inspector will use this diagram to determine if you have properly sized the drain traps, drain pipes, and vent pipes in your project.

Sizing for Water Distribution Pipes

Fixture	Unit rating
Toilet	3
Vanity sink	1
Shower	2
Bathtub	2
Dishwasher	2
Kitchen sink	2
Clothes washer	2
Utility sink	2
Sillcock	3

Size of service pipe from street	Size of distribution pipe from water meter	Maximum length (ft.)—total fixture units					
		40	60	80	100	150	200
¾"	½"	9	8	7	6	5	4
¾"	¾"	27	23	19	17	14	11
¾"	1"	44	40	36	33	28	23
1"	1"	60	47	41	36	30	25
1"	1¼"	102	87	76	67	52	44

Water distribution pipes are the main pipes extending from the water meter throughout the house, supplying water to the branch pipes leading to individual fixtures. To determine the size of the distribution pipes, you must first calculate the total demand in "fixture units" (above, left) and the overall length of the water supply lines, from the street hookup through the water meter and to the most distant fixture in the house. Then, use the second table (above, middle) to find the minimum size for the water distribution pipes. Note that the fixture unit capacity depends partly on the size of the street-side pipe that delivers water to your meter.

Sizes for Branch Pipes & Supply Tubes

Fixture	Min. branch pipe size	Min. supply tube size
Toilet	½"	⅜"
Vanity sink	½"	⅜"
Shower	½"	½"
Bathtub	½"	½"
Dishwasher	½"	½"
Kitchen sink	½"	½"
Clothes washer	½"	½"
Utility sink	½"	½"
Sillcock	¾"	N.A.
Water heater	¾"	N.A.

Branch pipes are the water supply lines that run from the distribution pipes toward the individual fixtures. Supply tubes are the vinyl, chromed copper, or braided tubes that carry water from the branch pipes to the fixtures. Use the chart above as a guide when sizing branch pipes and supply tubes.

Valve Requirements

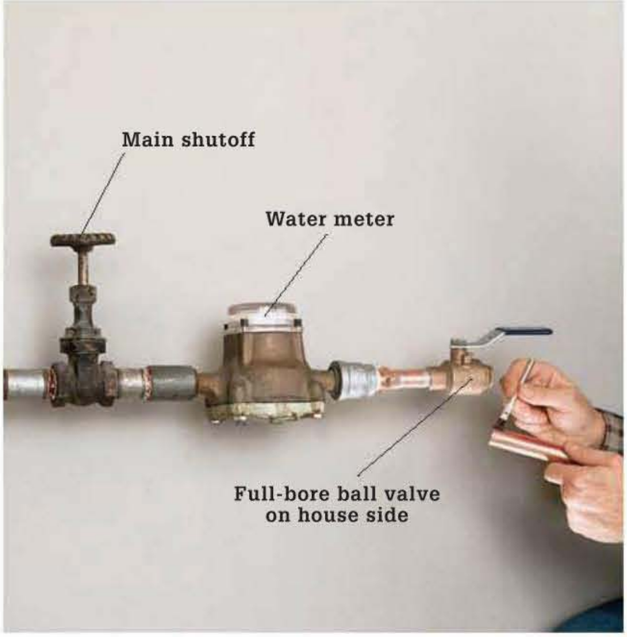

Full-bore gate valves or ball valves are required in the following locations: on both the street side and house side of the water meter; on the inlet pipes for water heaters and heating system boilers. Individual fixtures should have accessible shutoff valves, but these need not be full-bore valves. All sillcocks must have individual control valves located inside the house.

Drain cleanouts make your DWV system easier to service. In most areas, the plumbing code requires that you place cleanouts at the end of every horizontal drain run. Where horizontal runs are not accessible, removable drain traps will suffice as cleanouts.

Fixture Units & Minimum Trap Size

Fixture	Fixture units	Min. trap size
Shower	2	2"
Vanity sink	1	1¼"
Bathtub	2	1½"
Dishwasher	2	1½"
Kitchen sink	2	1½"
Kitchen sink*	3	1½"
Clothes washer	2	1½"
Utility sink	2	1½"
Floor drain	1	2"

*Kitchen sink with attached food disposer

Minimum trap size for fixtures is determined by the drain fixture unit rating, a unit of measure assigned by the plumbing code. *Note: Kitchen sinks rate 3 units if they include an attached food disposer, 2 units otherwise.*

Pipe Support Intervals

Type of pipe	Vertical-run support interval	Horizontal-run support interval
Copper	10 ft.	6 ft.
PEX	5 ft.	3 ft.
CPVC	10 ft.	3 ft.
PVC	10 ft.	4 ft.
Steel	12 ft.	10 ft.
Iron	15 ft.	5 ft.

Minimum intervals for supporting pipes are determined by the type of pipe and its orientation in the system. Remember that the measurements shown above are minimum requirements; local code may require supports at closer intervals.

Sizes for Horizontal & Vertical Drain Pipes

Pipe size	Maximum fixture units for horizontal branch drain	Maximum fixture units for vertical drain stacks
1¼"	1	2
1½"	3	4
2"	6	10
2½"	12	20
3"	20	30
4"	160	240

Drain pipe sizes are determined by the load on the pipes, as measured by the total fixture units. Horizontal drain pipes less than 3" in diameter should slope ¼" per foot toward the main drain. Pipes 3" or more in diameter should slope ⅛" per foot. *Note: Horizontal or vertical drain pipes for a toilet must be 3" or larger.*

Vent Pipe Sizes, Critical Distances

Size of fixture drain	Minimum vent pipe size	Maximum critical distance
1¼"	1¼"	2½ ft.
1½"	1¼"	3½ ft.
2"	1½"	5 ft.
3"	2"	6 ft.
4"	3"	10 ft.

Vent pipes are usually one pipe size smaller than the drain pipes they serve. Code requires that the distance between the drain trap and the vent pipe fall within a maximum "critical distance," a measurement that is determined by the size of the fixture drain. Use this chart to determine both the minimum size for the vent pipe and the maximum critical distance.

Wet Venting

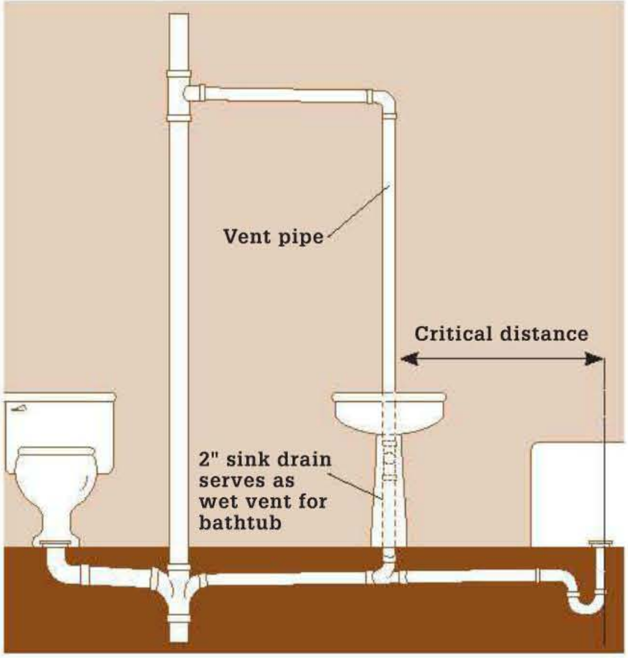

Wet vents are pipes that serve as a vent for one fixture and a drain for another. The sizing of a wet vent is based on the total fixture units it supports (opposite page): a 3" wet vent can serve up to 12 fixture units; a 2" wet vent is rated for 4 fixture units; a 1½" wet vent, for only 1 fixture unit. *Note: The distance between the wet-vented fixture and the wet vent itself must be no more than the maximum critical distance (chart above).*

Vent Pipe Orientation to Drain Pipe

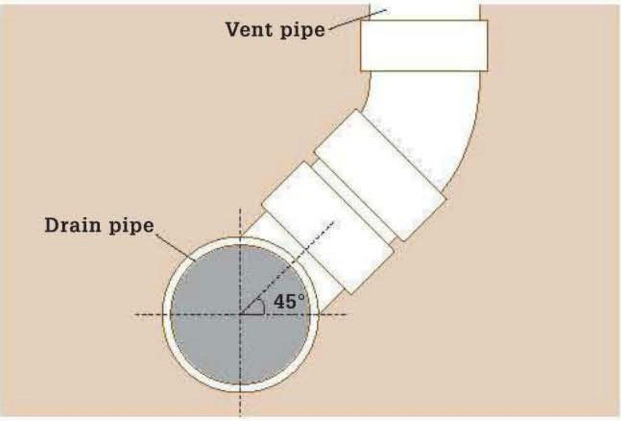

Vent pipes must extend in an upward direction from drains, no less than 45° from horizontal. This ensures that waste water cannot flow into the vent pipe and block it. At the opposite end, a new vent pipe should connect to an existing vent pipe or main waste-vent stack at a point at least 6" above the highest fixture draining into the system.

Auxiliary Venting

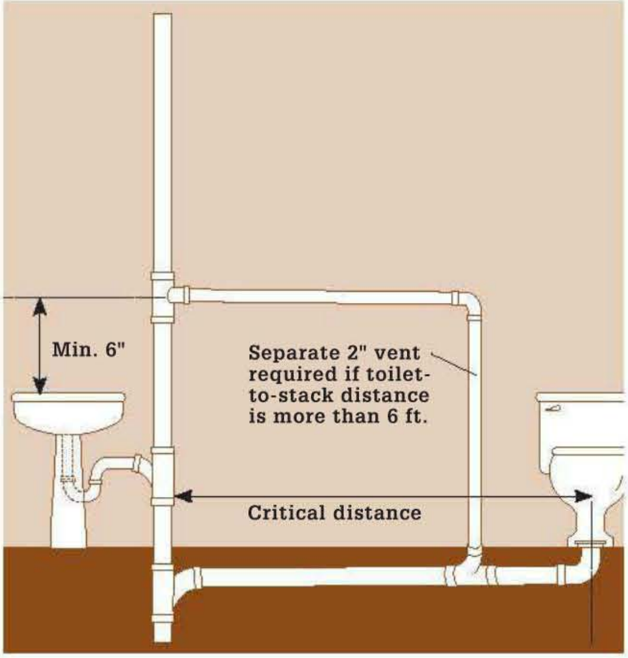
Min. 6"
Separate 2" vent required if toilet-to-stack distance is more than 6 ft.
Critical distance

Fixtures must have an auxiliary vent if the distance to the main waste-vent stack exceeds the critical distance (illustration above). A toilet, for example, should have a separate vent pipe if it is located more than 6 ft. from the main waste-vent stack. This secondary vent pipe should connect to the stack or an existing vent pipe at a point at least 6" above the highest fixture on the system.

Understanding Wiring

A household electrical system can be compared with a home's plumbing system. Electrical current flows in wires in much the same way that water flows inside pipes. Both electricity and water enter the home, are distributed throughout the house, do their "work," and exit.

In plumbing, water first flows through the pressurized water supply system. In electricity, current first flows along hot wires. Current flowing along hot wires also is pressurized. The pressure of electrical current is called *voltage.*

Large supply pipes can carry a greater volume of water than small pipes. Likewise, large electrical wires carry more current than small wires. This current-carrying capacity of wires is called *amperage.*

Water is made available for use through the faucets, spigots, and showerheads in a home. Electricity is made available through receptacles, switches, and fixtures.

Water finally leaves the home through a drain system, which is not pressurized. Similarly, electrical current flows back through neutral wires. The current in neutral wires is not pressurized and is said to be at zero voltage.

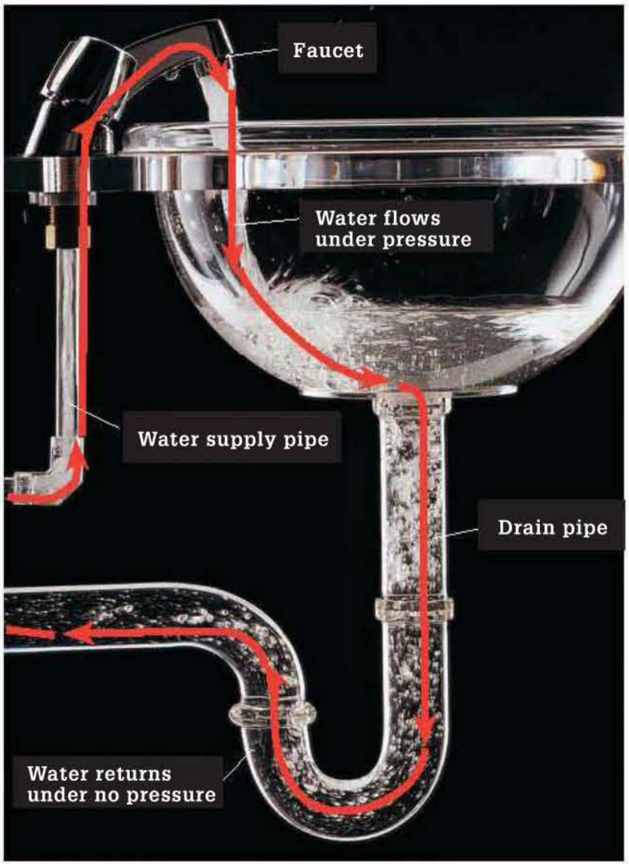
Faucet
Water flows under pressure
Water supply pipe
Drain pipe
Water returns under no pressure

Water and electricity both flow. The main difference is that you can see water (and touching water isn't likely to kill you). Like electricity, water enters a fixture under high pressure and exits under low pressure.

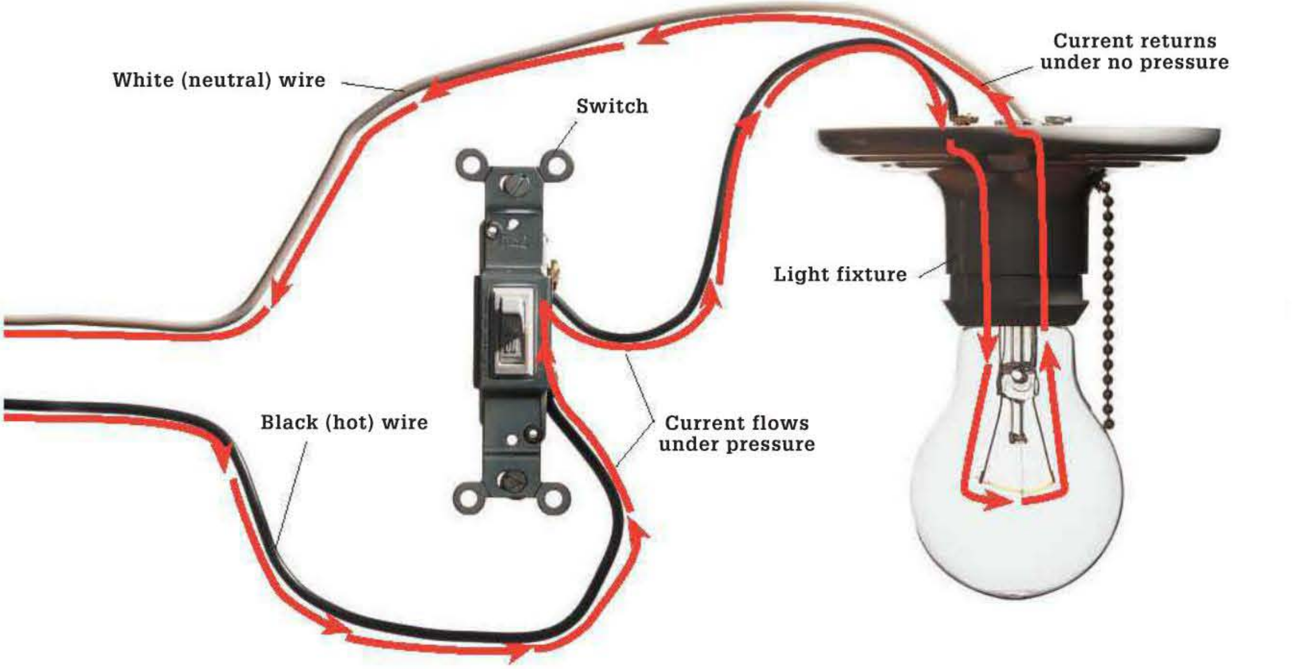
White (neutral) wire
Switch
Current returns under no pressure
Light fixture
Black (hot) wire
Current flows under pressure

The Delivery System

Electrical power that enters the home is produced by large power plants. Power plants are located in all parts of the country and generate electricity with turbines that are turned by water, wind, or steam. From these plants electricity enters large "step-up" transformers that increase voltage to half a million volts or more.

Electricity flows easily at these large voltages and travels through high-voltage transmission lines to communities that can be hundreds of miles from the power plants. "Step-down" transformers located at substations then reduce the voltage for distribution along street lines. On utility power poles, smaller transformers further reduce the voltage to ordinary 120-volt current for household use.

Lines carrying current to the house either run underground or strung overhead and attached to a post called a service mast. Most homes built after 1950 have three wires running to the service head: two power lines, each carrying 120 volts of current, and a grounded neutral wire. Power from the two 120-volt lines may be combined at the service panel to supply current to large 240-volt appliances like clothes dryers or electric water heaters.

Incoming power passes through an electric meter that measures power consumption. Power then enters the service panel, where it is distributed to circuits that run throughout the house. The service panel also contains fuses or circuit breakers that shut off power to the individual circuits in the event of a short circuit or an overload. Certain high-wattage appliances, like microwave ovens, are usually plugged into their own individual circuits to prevent overloads.

Voltage ratings determined by power companies and manufacturers have changed over the years. Current rated at 110 volts changed to 115 volts, then 120 volts. Current rated at 220 volts changed to 230 volts, then 240 volts. Similarly, ratings for receptacles, tools, light fixtures, and appliances have changed from 115 volts to 125 volts. These changes do not affect the performance of new devices connected to older wiring. For making electrical calculations, use a rating of 120 volts or 240 volts for your circuits.

Power plants supply electricity to thousands of homes and businesses. Step-up transformers increase the voltage produced at the plant, making the power flow more easily along high-voltage transmission lines.

Substations are located near the communities they serve. A typical substation takes current from high-voltage transmission lines and reduces it for distribution along street lines.

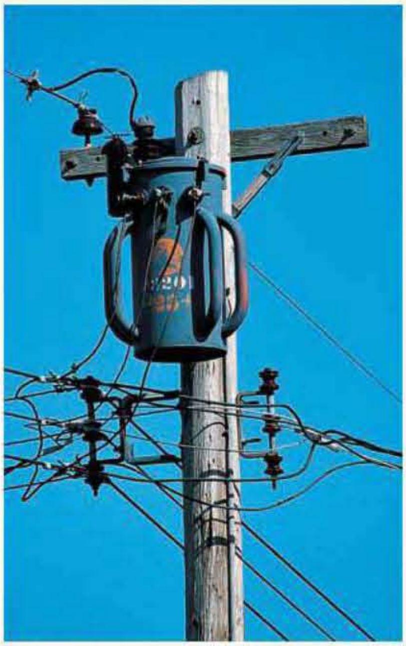

Utility pole transformers reduce the high-voltage current that flows through power lines along neighborhood streets. A utility pole transformer reduces voltage from 10,000 volts to the normal 120-volt current used in households.

Parts of the Electrical System

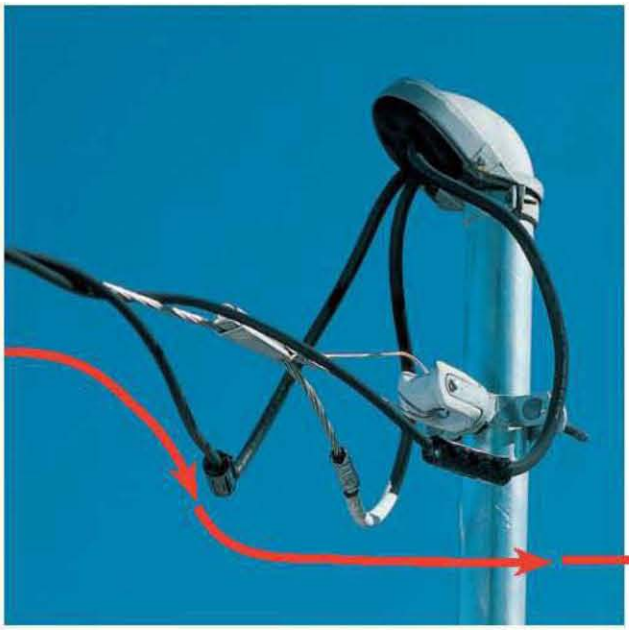

The service mast is the metal pole and weatherhead that create the entry point for electricity into your home. The mast is supplied with two wires carrying 120 volts and a third grounded neutral wire. The current runs to the service mast from the nearest transformer.

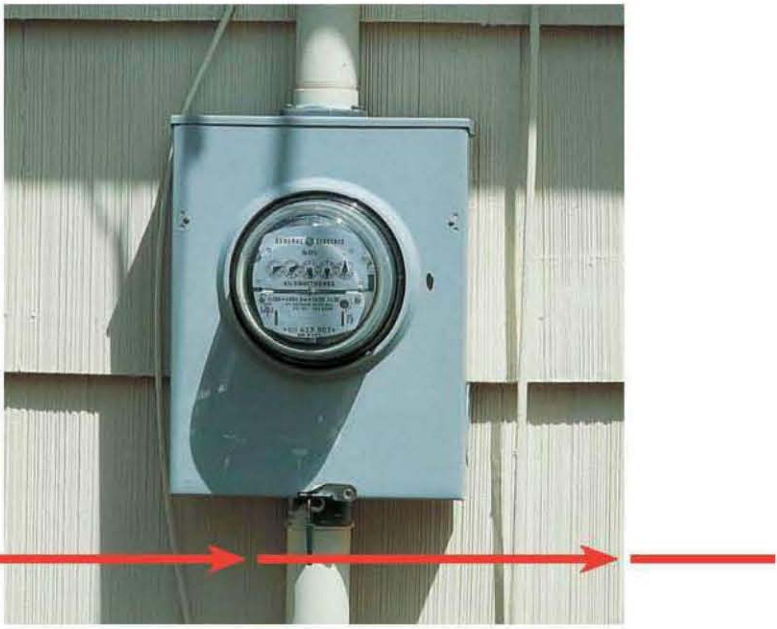

The electric meter measures the amount of electrical power consumed. It is usually attached to the side of the house, and connects to the service mast. A thin metal disc inside the meter rotates when power is used. The electric meter belongs to your local power utility company. If you suspect the meter is not functioning properly, contact the power company.

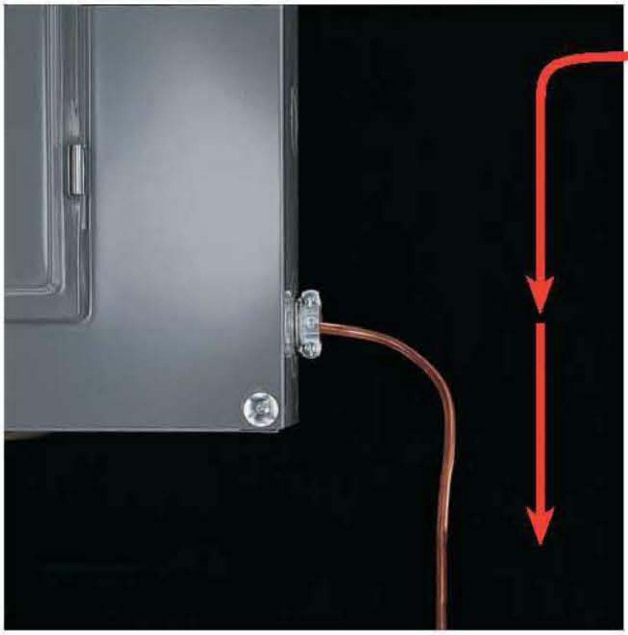

Grounding wire connects the electrical system to the earth through grounding rods or, in older systems, through a cold water pipe. In the event of an overload or short circuit, the grounding wire allows excess electrical power to find its way harmlessly to the earth.

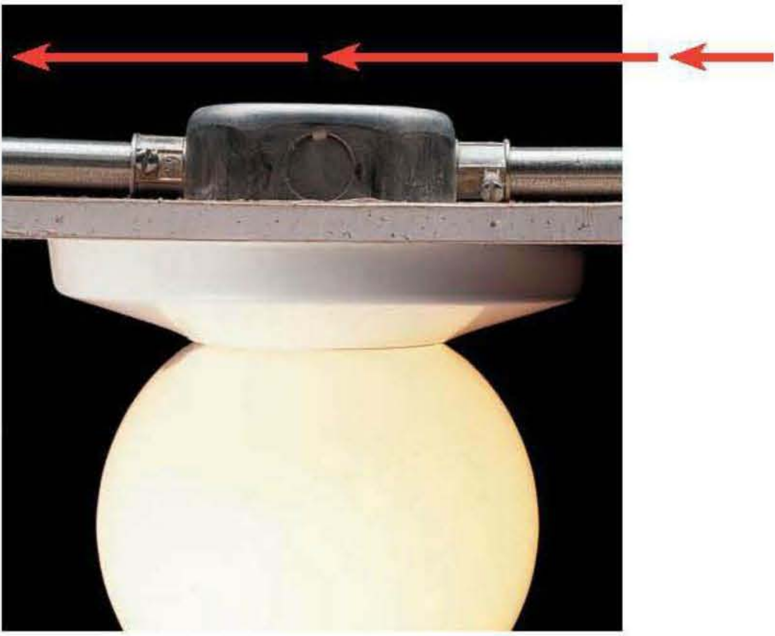

Light fixtures attach directly to a household electrical system. They are usually controlled with wall switches. The two most common types of light fixtures are incandescent and fluorescent.

The main service panel, in the form of a fuse box or breaker box, distributes power to individual circuits. Fuses or circuit breakers protect each circuit from short circuits and overloads. Fuses and circuit breakers also are used to shut off power to individual circuits while repairs are made.

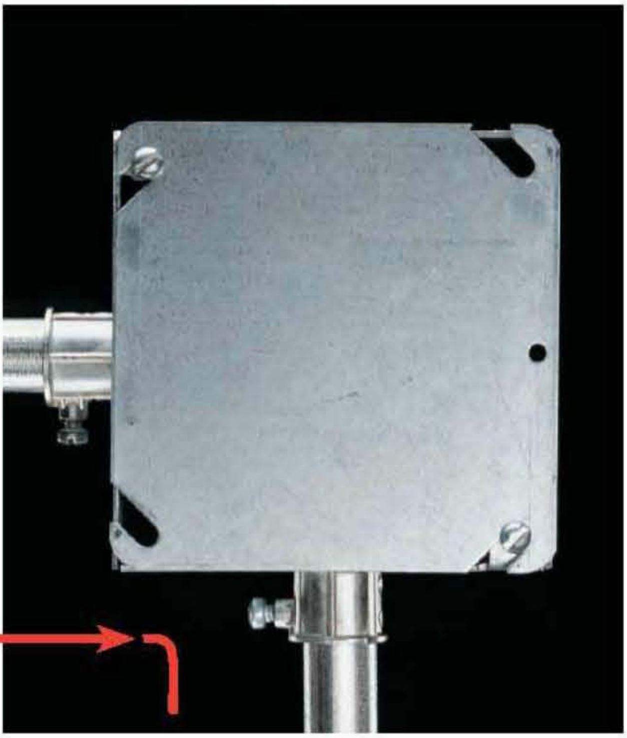

Electrical boxes enclose wire connections. According to the National Electrical Code, all wire splices or connections must be contained entirely in a covered plastic or metal electrical box.

Switches control electrical current passing through hot circuit wires. Switches can be wired to control light fixtures, ceiling fans, appliances, and receptacles.

Receptacles, sometimes called outlets, provide plug-in access to electrical power. A 120-volt, 15-amp receptacle with a grounding hole is the most typical receptacle in wiring systems installed after 1965. Most receptacles have two plug-in locations and are called *duplex receptacles.*

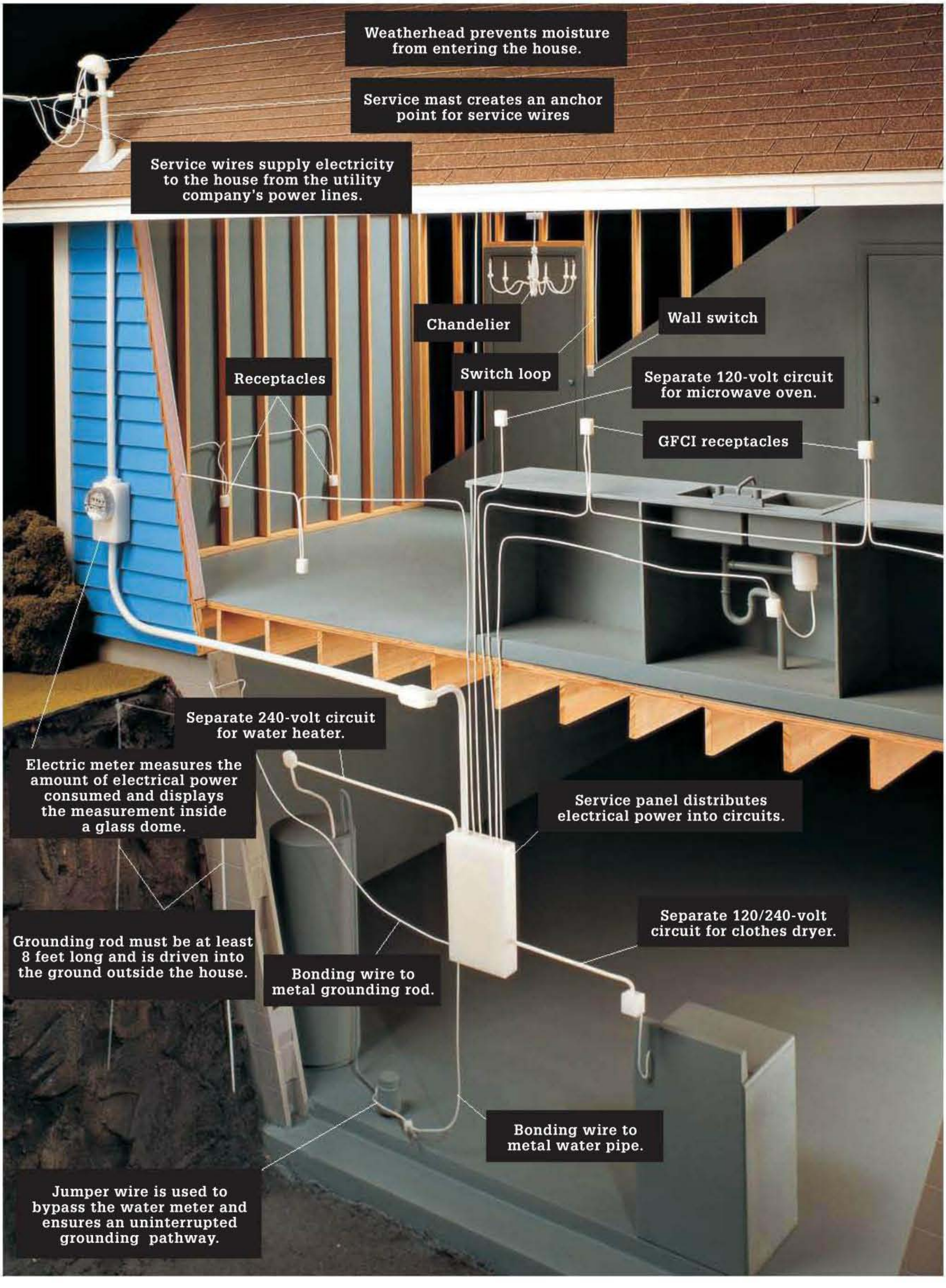
Weatherhead prevents moisture from entering the house.
Service mast creates an anchor point for service wires
Service wires supply electricity to the house from the utility company's power lines.
Chandelier
Wall switch
Receptacles
Switch loop
Separate 120-volt circuit for microwave oven.
GFCI receptacles
Separate 240-volt circuit for water heater.
Electric meter measures the amount of electrical power consumed and displays the measurement inside a glass dome.
Service panel distributes electrical power into circuits.
Separate 120/240-volt circuit for clothes dryer.
Grounding rod must be at least 8 feet long and is driven into the ground outside the house.
Bonding wire to metal grounding rod.
Bonding wire to metal water pipe.
Jumper wire is used to bypass the water meter and ensures an uninterrupted grounding pathway.

Glossary of Electrical Terms

Ampere (or amp): Refers to the rate at which electrical power flows to a light, tool, or appliance.

Armored cable: Two or more wires that are grouped together and protected by a flexible metal covering.

Box: A device used to contain wiring connections.

BX: See armored cable (Bx is the older term).

Cable: Two or more wires that are grouped together and protected by a covering or sheath.

Circuit: A continuous loop of electrical current flowing along wires or cables.

Circuit breaker: A safety device that interrupts an electrical circuit in the event of an overload or short circuit.

Conductor: Any material that allows electrical current to flow through it. Copper wire is an especially good conductor.

Conduit: A metal or plastic pipe used to protect wires.

Continuity: An uninterrupted electrical pathway through a circuit or electrical fixture.

Current: The movement of electrons along a conductor.

Duplex receptacle: A receptacle that provides connections for two plugs.

Feed wire: A conductor that carries 120-volt current uninterrupted from the service panel.

Fuse: A safety device, usually found in older homes, that interrupts electrical circuits during an overload or short circuit.

Greenfield: Materials used in flexible metal conduit. See armored cable.

Grounded wire: See neutral wire.

Grounding wire: A wire used in an electrical circuit to conduct current to the earth in the event of a short circuit. The grounding wire often is a bare copper wire.

Hot wire: Any wire that carries voltage. In an electrical circuit, the hot wire usually is covered with black or red insulation.

Insulator: Any material, such as plastic or rubber, that resists the flow of electrical current. Insulating materials protect wires and cables.

Junction box: See box.

Meter: A device used to measure the amount of electrical power being used.

Neutral wire: A wire that returns current at zero voltage to the source of electrical power. Usually covered with white or light gray insulation. Also called the grounded wire.

Outlet: See receptacle.

Overload: A demand for more current than the circuit wires or electrical device was designed to carry. Usually causes a fuse to blow or a circuit breaker to trip.

Pigtail: A short wire used to connect two or more circuit wires to a single screw terminal.

Polarized receptacle: A receptacle designed to keep hot current flowing along black or red wires, and neutral current flowing along white or gray wires.

Power: The result of hot current flowing for a period of time. Use of power makes heat, motion, or light.

Receptacle: A device that provides plug-in access to electrical power.

Romex: A brand name of plastic-sheathed electrical cable that is commonly used for indoor wiring. Commonly known as NM cable.

Screw terminal: A place where a wire connects to a receptacle, switch, or fixture.

Service panel: A metal box usually near the site where electrical power enters the house. In the service panel, electrical current is split into individual circuits. The service panel has circuit breakers or fuses to protect each circuit.

Short circuit: An accidental and improper contact between two current-carrying wires, or between a current-carrying wire and a grounding conductor.

Switch: A device that controls electrical current passing through hot circuit wires. Used to turn lights and appliances on and off.

UL: An abbreviation for Underwriters Laboratories, an organization that tests electrical devices and manufactured products for safety.

Voltage (or volts): A measurement of electricity in terms of pressure.

Wattage (or watt): A measurement of electrical power in terms of total energy consumed. Watts can be calculated by multiplying the voltage times the amps.

Wire connector: A device used to connect two or more wires together. Also called a *wire nut.*

Wiring Safety

Safety should be the primary concern of anyone working with electricity. Although most household electrical repairs are simple and straightforward, always use caution and good judgment when working with electrical wiring or devices. Common sense can prevent accidents.

The basic rule of electrical safety is: Always turn off power to the area or device you are working on. At the main service panel, remove the fuse or shut off the circuit breaker that controls the circuit you are servicing. Then check to make sure the power is off by testing for power with a current tester. Restore power only when the repair or replacement project is complete.

Follow the safety tips shown on these pages. Never attempt an electrical project beyond your skill or confidence level. Never attempt to repair or replace your main service panel or service entrance head. These are jobs for a qualified electrician and require that the power company shuts off power to your house.

Shut power OFF at the main service panel or the main fuse box before beginning any work.

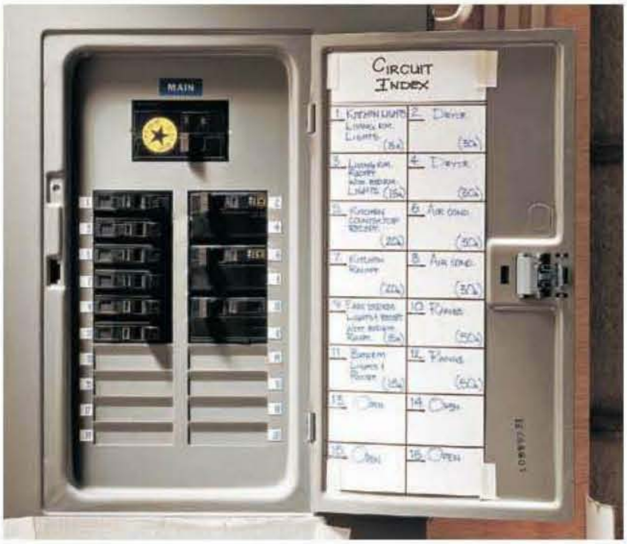

Create a circuit index and affix it to the inside of the door to your main service panel. Update it as needed.

Confirm power is OFF by testing at the outlet, switch, or fixture with a current tester.

Electracraft®
Quik-connect
double switch
QUIET Single Pole
UL

Use only UL-approved electrical parts or devices. These devices have been tested for safety by Underwriters Laboratories.

Wear rubber-soled shoes while working on electrical projects. On damp floors, stand on a rubber mat or dry wooden boards.

Use fiberglass or wood ladders when making routine household repairs near the service mast.

Extension cords are for temporary use only. Cords must be rated for the intended usage.

Electracraft®
plug
fuses
TIME DELAY FUSE
20

Breakers and fuses must be compatible with the panel manufacturer and match the circuit capacity.

Never alter the prongs of a plug to fit a receptacle. If possible, install a new grounded receptacle.

Do not penetrate walls or ceilings without first shutting off electrical power to the circuits that may be hidden.

Working with Wire & Cable

Wires are made of copper, aluminum, or aluminum covered with a thin layer of copper. Solid copper wires are the best conductors of electricity and are the most widely used. Aluminum and copper-covered aluminum wires require special installation techniques and are rarely used in new installation, due to safety concerns.

A group of two or more wires enclosed in a metal, rubber, or plastic sheath is called a *cable* (photo, opposite page). The sheath protects the wires from damage. Metal conduit also protects wires, but it is not considered a cable.

Individual wires are covered with rubber or plastic vinyl insulation. An exception is a bare copper grounding wire, which does not need an insulation cover. The insulation is color coded (chart, below left) to identify the wire as a hot wire, a neutral wire, or a grounding wire.

In most wiring systems installed after 1965, the wires and cables are insulated with plastic vinyl. This type of insulation is very durable and can last as long as the house itself.

Before 1965, wires and cables were insulated with rubber. Rubber insulation has a life expectancy of about 25 years. Old insulation that is cracked or damaged can be reinforced temporarily by wrapping the wire with plastic electrical tape. However, old wiring with cracked or damaged insulation should be inspected by a qualified electrician to make sure it is safe.

Wires must be large enough for the amperage rating of the circuit (chart, below right). A wire that is too small can become dangerously hot. Wire sizes are categorized according to the American Wire Gauge (AWG) system. To check the size of a wire, use the wire stripper openings of a combination tool (page 30) as a guide.

Wire Color Chart ▸

Wire color		Function
	White	Neutral wire carrying current at zero voltage.
	Black	Hot wire carrying current at full voltage.
	Red	Hot wire carrying current at full voltage.
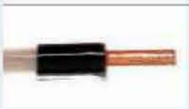	White, black markings	Hot wire carrying current at full voltage.
	Green	Serves as a grounding pathway.
	Bare copper	Serves as a grounding pathway.

Individual wires are color-coded to identify their function. In some circuit installations, the white wire serves as a hot wire that carries voltage. If so, this white wire may be labeled with black tape or paint to identify it as a hot wire.

Wire Size Chart ▸

Wire gauge		Wire capacity & use
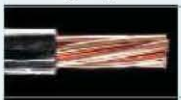	#6	60 amps, 240 volts; central air conditioner, electric furnace.
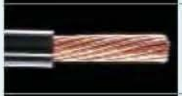	#8	40 amps, 240 volts; electric range, central air conditioner.
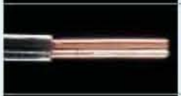	#10	30 amps, 240 volts; window air conditioner, clothes dryer.
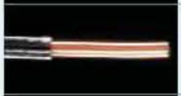	#12	20 amps, 120 volts; light fixtures, receptacles, microwave oven.
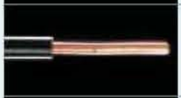	#14	15 amps, 120 volts; light fixtures, receptacles.
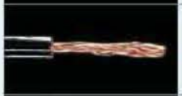	#16	Light-duty extension cords.
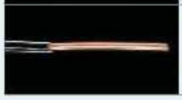	#18 to 22	Thermostats, doorbells, security systems.

Wire sizes (shown actual size) are categorized by the American Wire Gauge system. The larger the wire size, the smaller the AWG number.

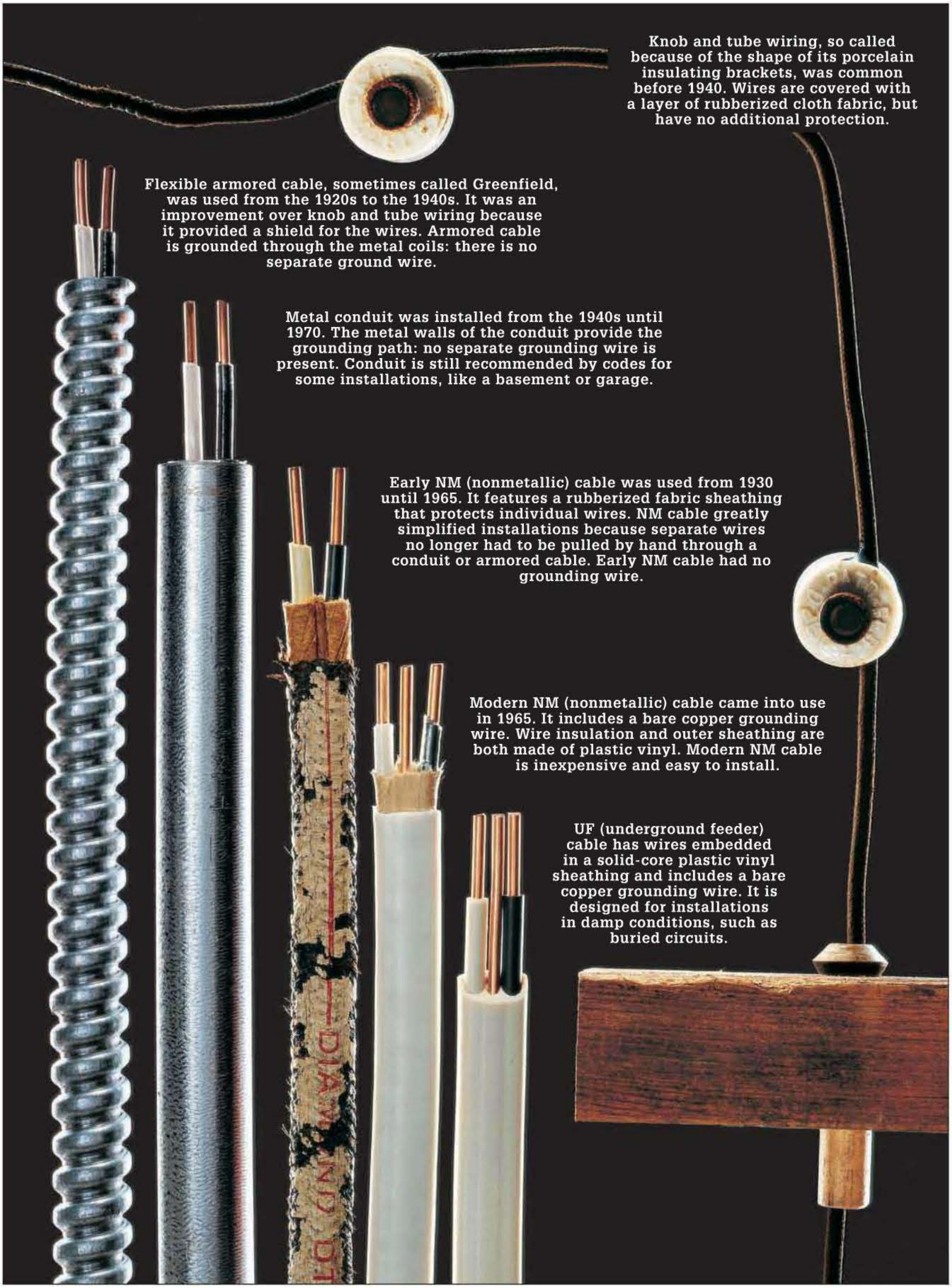

Knob and tube wiring, so called because of the shape of its porcelain insulating brackets, was common before 1940. Wires are covered with a layer of rubberized cloth fabric, but have no additional protection.

Flexible armored cable, sometimes called Greenfield, was used from the 1920s to the 1940s. It was an improvement over knob and tube wiring because it provided a shield for the wires. Armored cable is grounded through the metal coils: there is no separate ground wire.

Metal conduit was installed from the 1940s until 1970. The metal walls of the conduit provide the grounding path: no separate grounding wire is present. Conduit is still recommended by codes for some installations, like a basement or garage.

Early NM (nonmetallic) cable was used from 1930 until 1965. It features a rubberized fabric sheathing that protects individual wires. NM cable greatly simplified installations because separate wires no longer had to be pulled by hand through a conduit or armored cable. Early NM cable had no grounding wire.

Modern NM (nonmetallic) cable came into use in 1965. It includes a bare copper grounding wire. Wire insulation and outer sheathing are both made of plastic vinyl. Modern NM cable is inexpensive and easy to install.

UF (underground feeder) cable has wires embedded in a solid-core plastic vinyl sheathing and includes a bare copper grounding wire. It is designed for installations in damp conditions, such as buried circuits.

SHEATHED CABLE TYPE NM

NM (nonmetallic) sheathed cable should be used for most indoor wiring projects in dry locations. NM cable is available in a wide range of wire sizes, and in either "2-wire with ground" or "3-wire with ground" types. NM cable is sold in boxed rolls that contain from 25 to 250 ft. of cable.

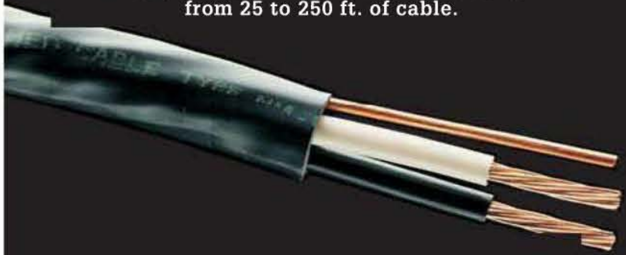

Large-appliance cable, also called SER cable, is used for kitchen ranges and other 50-amp or 60-amp appliances that require 8-gauge or larger wire. It is similar to NM cable, but each individual conducting wire is made from fine-stranded copper wires. Large-appliance cable is available in both 2-wire and 3-wire types.

GROUND TYPE UF

UF (underground feeder) cable is used for wiring in damp locations, such as in an outdoor circuit. It has a white or gray solid-core vinyl sheathing that protects the wires inside. It also can be used indoors wherever NM cable is allowed.

Coaxial cable is used to connect cable television jacks. Coaxial cable is available in lengths up to 25 ft. with preattached F-connectors (A). Or you can buy bulk cable (B) in any length.

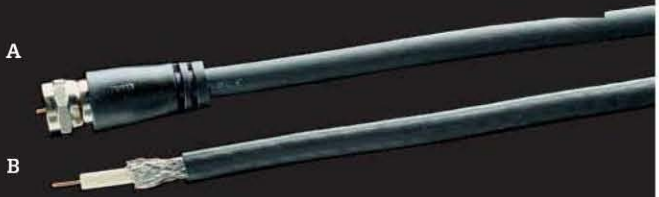

THHN/THWN wire can be used in all conduit applications. Each wire, purchased individually, is covered with a color-coded thermoplastic insulating jacket. Make sure the wire you buy has the THHN/THWN rating. Other wire types are less resistant to heat and moisture than THHN/THWN wire.

Telephone cable is used to connect telephone outlets. Your phone company may recommend four-wire cable (shown below) or eight-wire cable, sometimes called four-pair. Eight-wire cable has extra wires that are left unattached. These extra wires allow for future expansion of the system.

Tips for Working With Wire ▸

Wire gauge		Ampacity	Maximum wattage load
	14-gauge	15 amps	1440 watts (120 volts)
	12-gauge	20 amps	1920 watts (120 volts) 3840 watts (240 volts)
	10-gauge	30 amps	2880 watts (120 volts) 5760 watts (240 volts)
	8-gauge	40 amps	7680 watts (240 volts)
	6-gauge	50 amps	9600 watts (240 volts)

Wire "ampacity" is a measurement of how much current a wire can carry safely. Ampacity varies according to the size of the wires, as shown at left. When installing a new circuit, choose wire with an ampacity rating matching the circuit size. For dedicated appliance circuits, check the wattage rating of the appliance and make sure it does not exceed the maximum wattage load of the circuit.

Reading NM (Nonmetallic) Cable

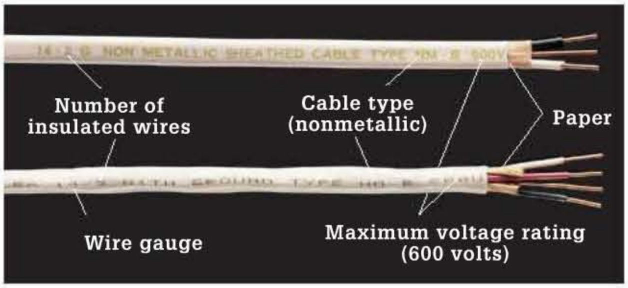
14/2 G NON METALLIC SHEATHED CABLE TYPE NM B 600V
Number of insulated wires
Cable type (nonmetallic)
Paper
Wire gauge
Maximum voltage rating (600 volts)

NM (nonmetallic) cable is labeled with the number of insulated wires it contains. The bare grounding wire is not counted. For example, a cable marked 14/2 G (or 14/2 WITH GROUND) contains two insulated 14-gauge wires, plus a bare copper grounding wire. Cable marked 14/3 WITH GROUND has three 14-gauge wires plus a grounding wire. NM cable also is stamped with a maximum voltage rating, as determined by Underwriters Laboratories (UL).

Reading Unsheathed, Individual Wire

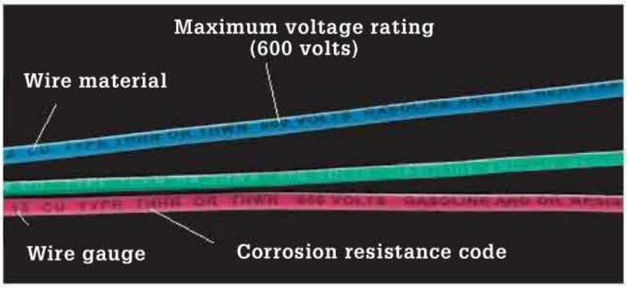
Maximum voltage rating (600 volts)
Wire material
Wire gauge
Corrosion resistance code

Unsheathed, individual wires are used for conduit and raceway installations. Wire insulation is coded with letters to indicate resistance to moisture, heat, and gas or oil. Code requires certain letter combinations for certain applications. T indicates thermoplastic insulation. H stands for heat resistance and two Hs indicate high resistance (up to 194° F). W denotes wire suitable for wet locations. Wire coded with an N is impervious to damage from oil or gas.

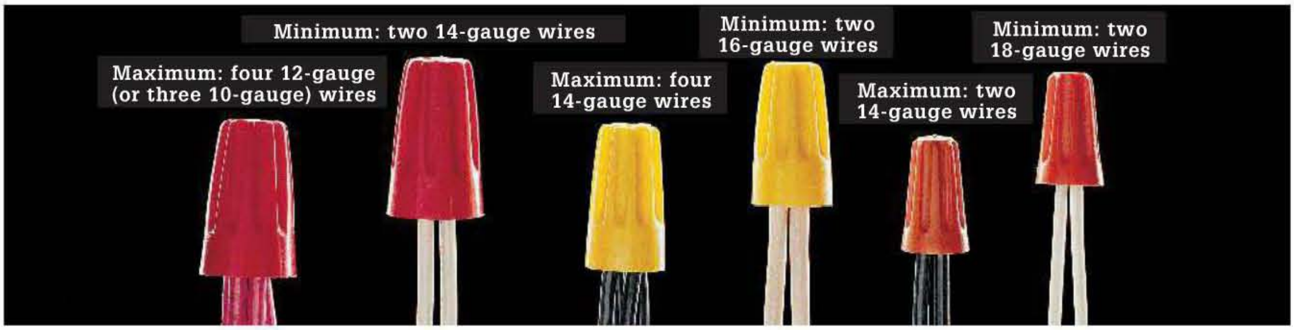
Minimum: two 14-gauge wires
Maximum: four 12-gauge (or three 10-gauge) wires
Maximum: four 14-gauge wires
Minimum: two 16-gauge wires
Maximum: two 14-gauge wires
Minimum: two 18-gauge wires

Use wire connectors rated for the wires you are connecting. Wire connectors are color-coded by size, but the coding scheme varies according to manufacturer. The wire connectors shown above come from one major manufacturer. To ensure safe connections, each connector is rated for both minimum and maximum wire capacity. These connectors can be used to connect both conducting wires and grounding wires. Green wire connectors are used only for grounding wires.

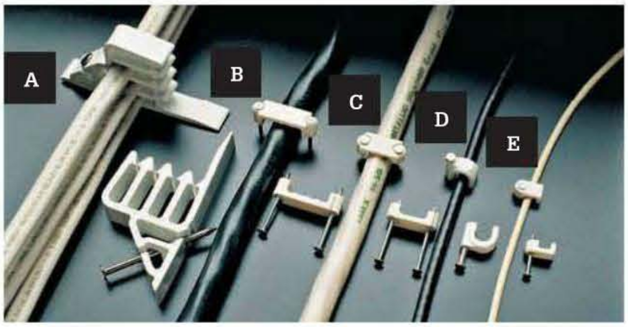

Use plastic cable staples to fasten cables. Choose staples sized to match the cables. Stack-It® staples (A) hold up to four 2-wire cables; ¾" staples (B) for 12/2, 12/3, and all 10-gauge cables; ½" staples (C) for 14/2, 14/3, or 12/2 cables; coaxial staples (D) for anchoring television cables; bell wire staples (E) for attaching telephone cables.

WAGO
100
773–164
4x 12 AWG Max. Cu
50
773–166
6x 12 AWG Max. Cu

Push-in connectors are a relatively new product for joining wires. Instead of twisting the bare wire ends together, you strip off about ¾" of insulation and insert them into a hole in the connector. The connectors come with two to four holes sized for various gauge wires. These connectors are perfect for inexperienced DIYers because they do not pull apart like a poorly twisted connection can.

How to Strip NM Sheathing & Insulation

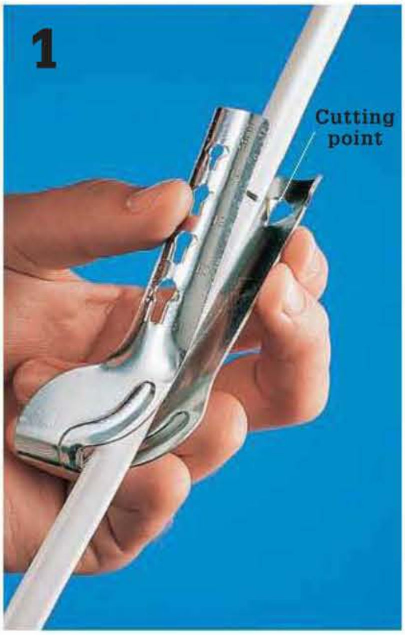

Measure and mark the cable 8 to 10" from end. Slide the cable ripper onto the cable, and squeeze tool firmly to force cutting point through plastic sheathing.

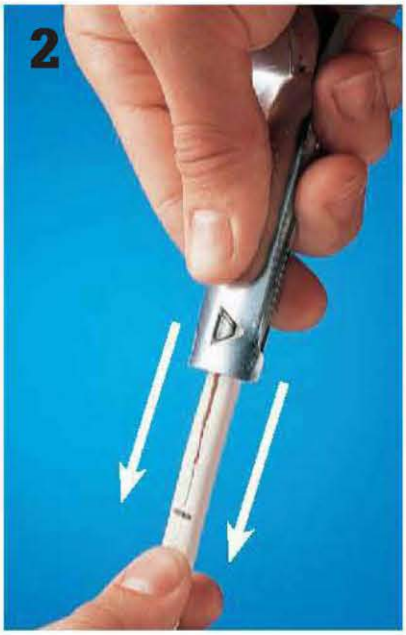

Grip the cable tightly with one hand, and pull the cable ripper toward the end of the cable to cut open the plastic sheathing.

Peel back the plastic sheathing and the paper wrapping from the individual wires.

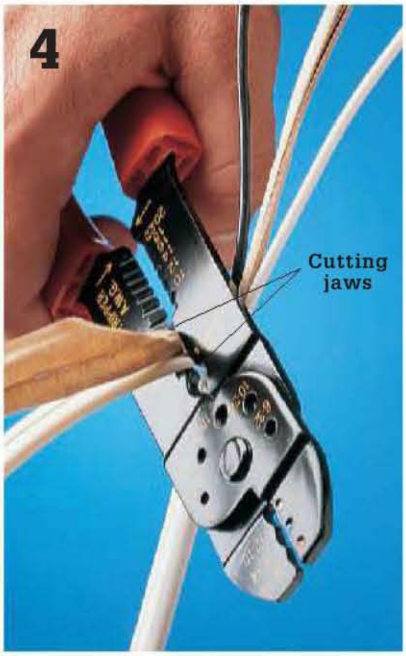

Cut away the excess plastic sheathing and paper wrapping, using the cutting jaws of a combination tool.

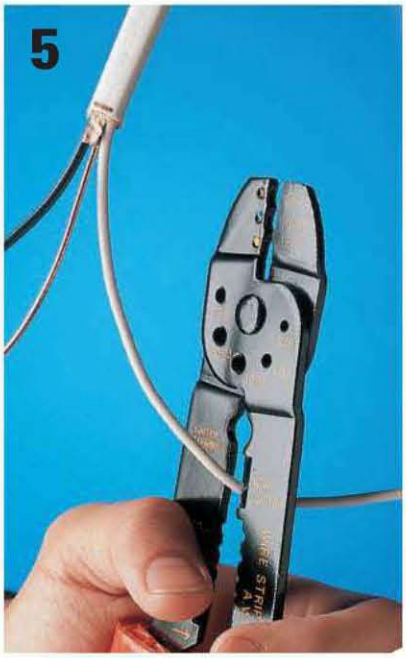

Cut individual wires as needed using the cutting jaws of the combination tool.

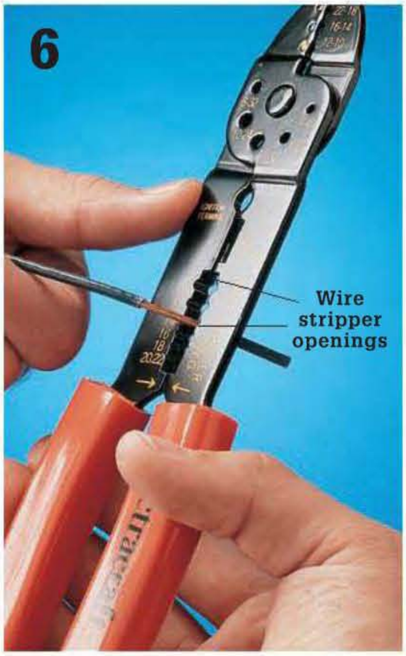

Strip insulation for each wire, using the stripper openings. Choose the opening that matches the gauge of the wire, and take care not to nick or scratch the ends of the wires.

How to Connect Wires to Screw Terminals

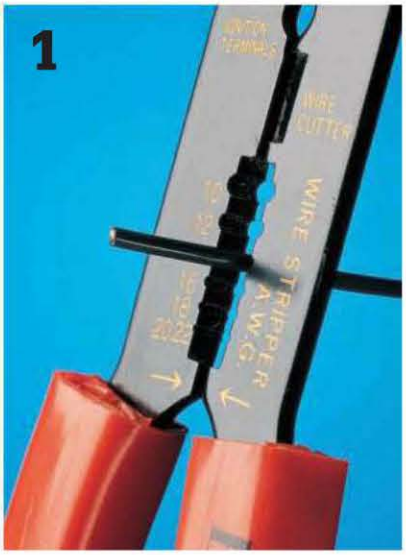
1
WIRE STRIPPER
A.W.G.
WIRE CUTTER

Strip about ¾" of insulation from each wire using a combination tool. Choose the stripper opening that matches the gauge of the wire, then clamp the wire in the tool. Pull the wire firmly to remove plastic insulation.

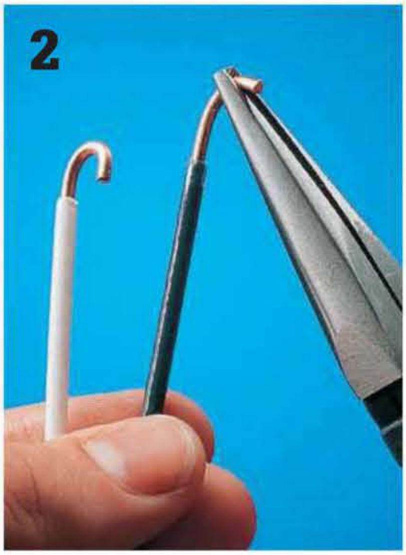

Form a C-shaped loop in the end of each wire using a needlenose pliers. The wire should have no scratches or nicks.

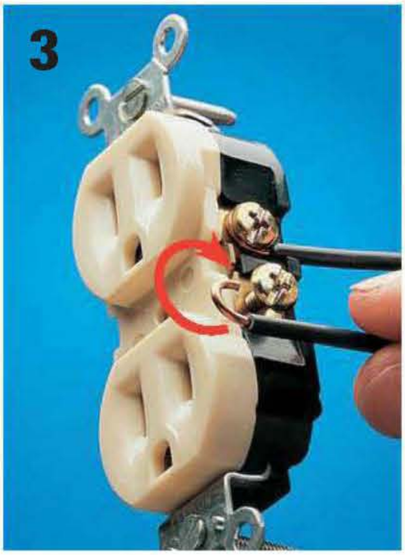

Hook each wire around the screw terminal so it forms a clockwise loop. Tighten screw firmly. Insulation should just touch head of screw. Never place the ends of two wires under a single screw terminal. Instead, use a pigtail wire (page 97).

How to Connect Wires with Push-ins

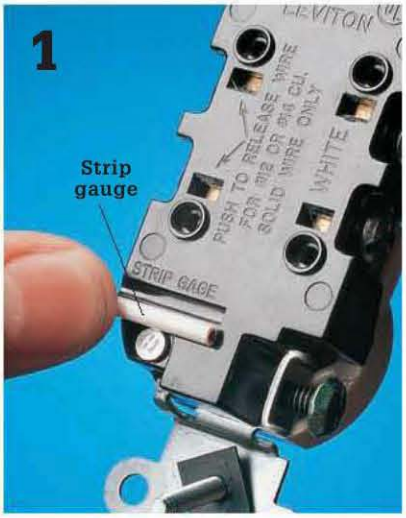
1
Strip gauge
LEVITON
PUSH TO RELEASE WIRE FOR #12 OR #14 CU. SOLID WIRE ONLY
WHITE
STRIP GAGE

Mark the amount of insulation to be stripped from each wire using the strip gauge on the back of the switch or receptacle. Strip the wires using a combination tool (step 1, above). Never use push-in fittings with aluminum wiring.

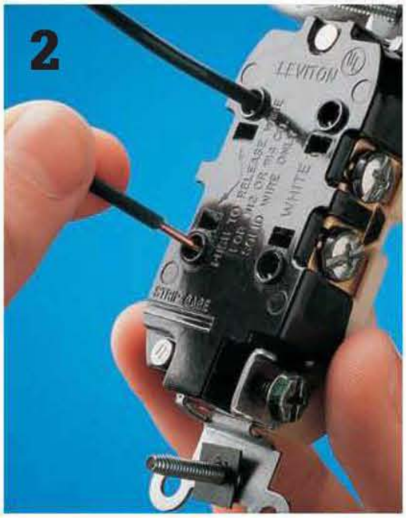
2
LEVITON
WHITE
STRIP GAGE

Insert the bare copper wires firmly into the push-in fittings on the back of the switch or receptacle. When inserted, wires should have no bare copper exposed. *Note: Although push-in fittings are convenient, most experts believe screw terminal connections (above) are more dependable.*

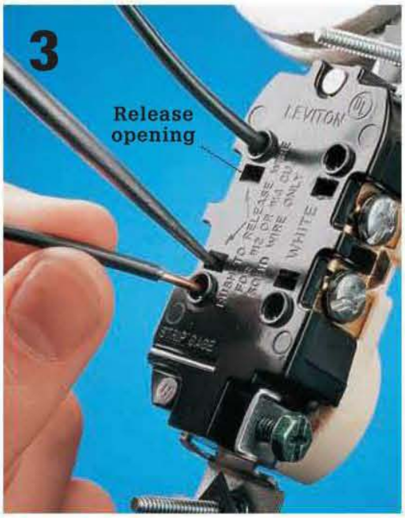
3
Release opening
LEVITON
WHITE
STRIP GAGE

Remove a wire from a push-in fitting by inserting a small nail or screwdriver in the release opening next to the wire. Wire will pull out easily.

How to Join Wires with a Wire Connector

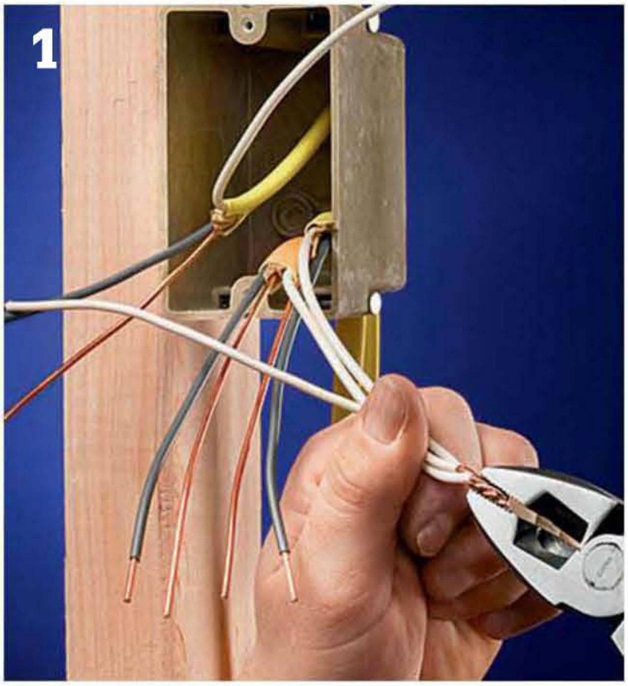

Ensure power is off and test for power. Grasp the wires to be joined in the jaws of a pair of linesman's pliers. The ends of the wires should be flush and they should be parallel and touching. Rotate the pliers clockwise two or three turns to twist the wire ends together.

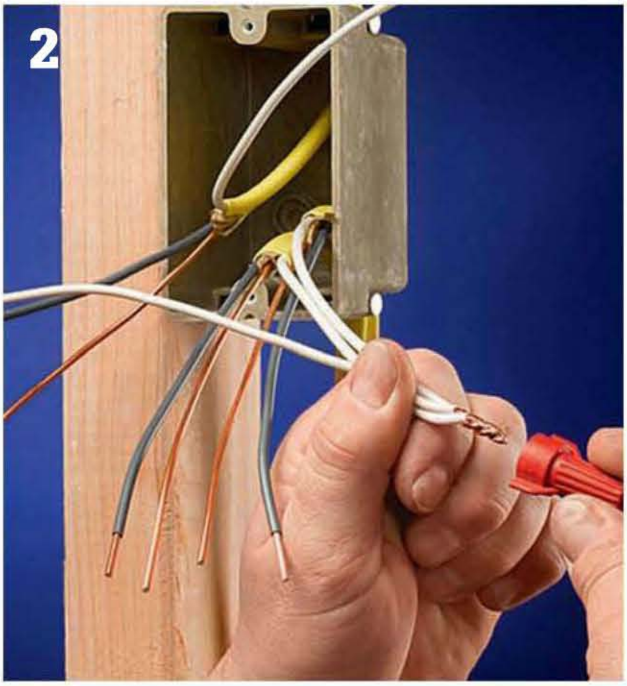

Twist a wire connector over the ends of the wires. Make sure the connector is the right size (see page 29). Hand-twist the connector as far onto the wires as you can. There should be no bare wire exposed beneath the collar of the connector. Do not overtighten the connector.

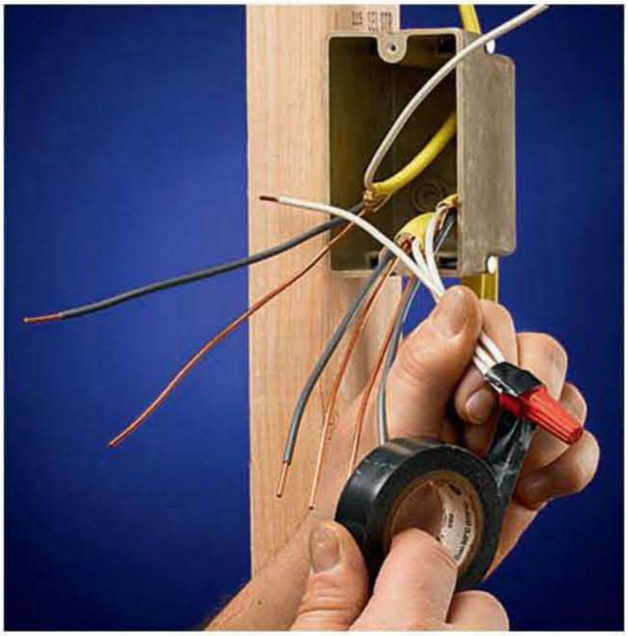

Option: Reinforce the joint by wrapping it with electrician's tape. By code, you cannot bind the wire joint with tape only, but it can be used as insurance. Few professional electricians use tape for purposes other than tagging wires for identification.

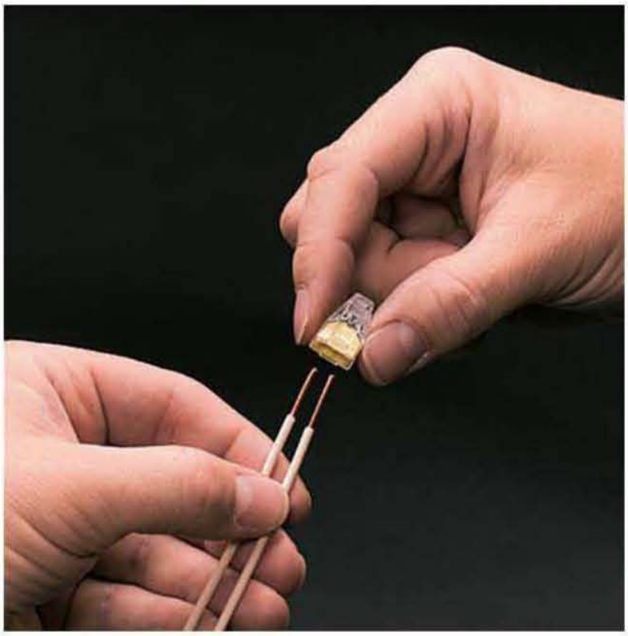

Option: Strip ¾" of insulation off the ends of the wires to be joined, and insert each wire into a push-in connector. Gently tug on each wire to make sure it is secure.

How to Pigtail Wires

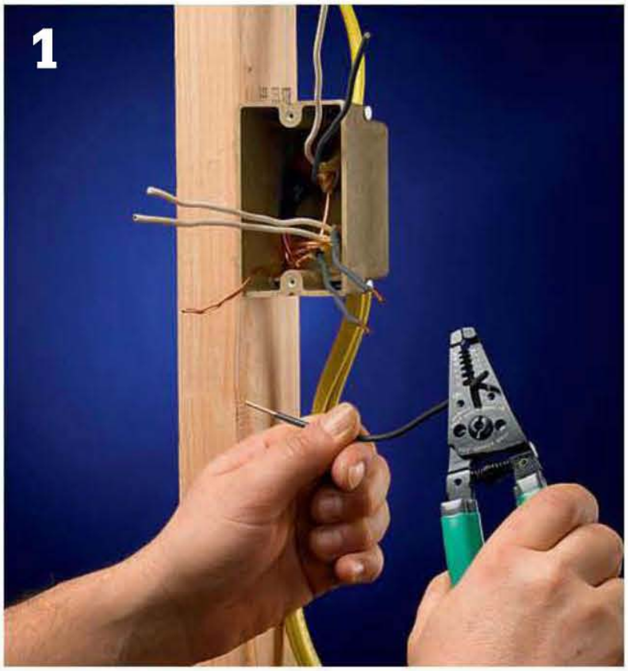

Cut a 6" length from a piece of insulated wire the same gauge and color as the wires it will be joining. Strip ¾" of insulation from each end of the insulated wire. *Note: Pigtailing is done mainly to avoid connecting multiple wires to one terminal, which is a code violation.*

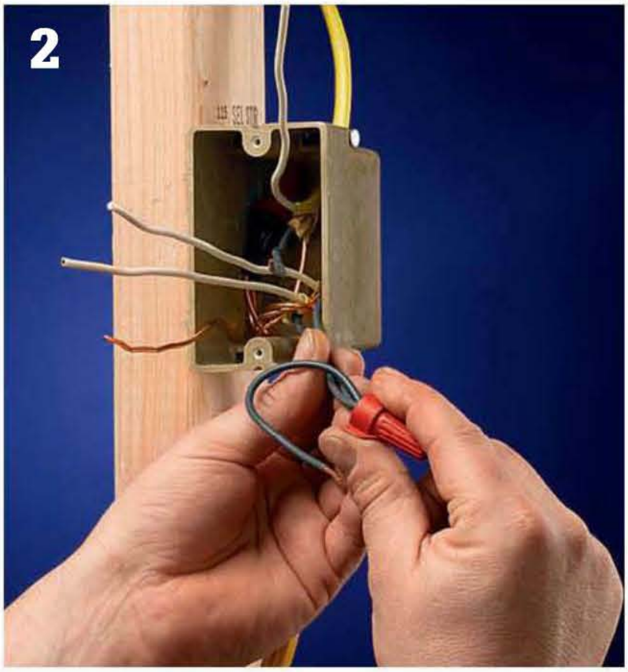

Join one end of the pigtail to the wires that will share the connection using a wire nut (see previous page).

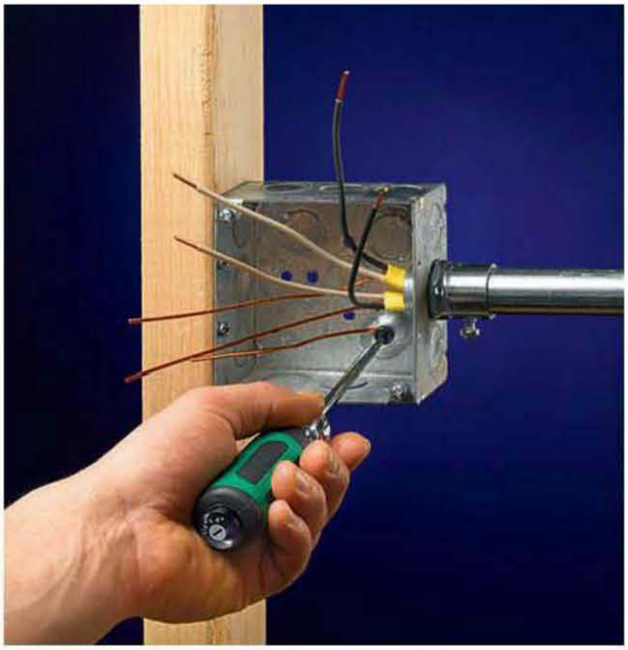

Alternative: If you are pigtailing to a grounding screw or grounding clip in a metal box, you may find it easier to attach one end of the wire to the grounding screw before you attach the other end to the other wires.

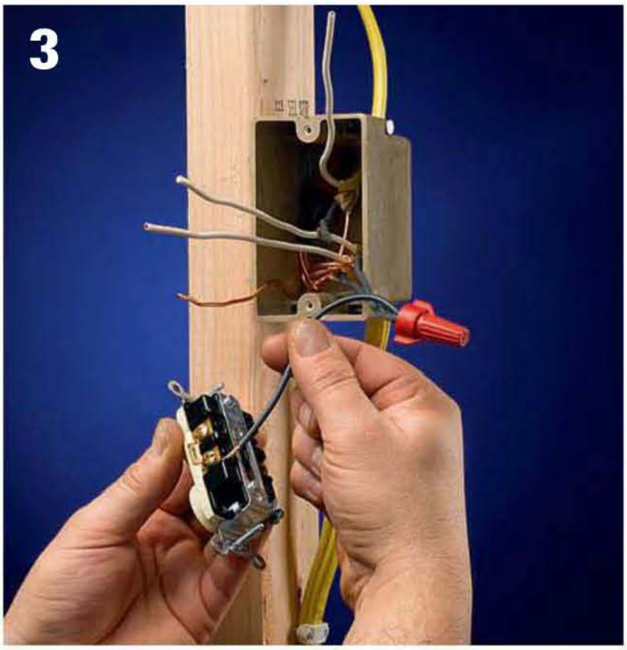

Connect the pigtail to the appropriate terminal on the receptacle or switch. Fold the wires neatly and press the fitting into the box.

How to Install NM Cable

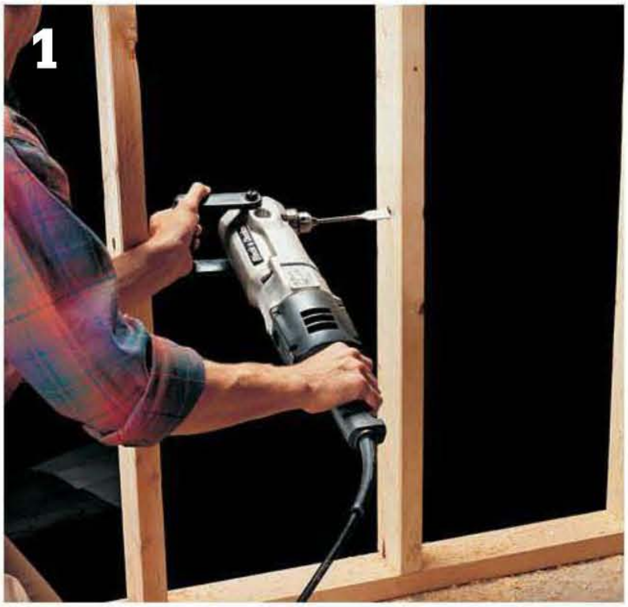

Drill ⅝" holes in framing members for the cable runs. This is done easily with a right-angle drill, available at rental centers. Holes should be set back at least 1¼" from the front face of the framing members.

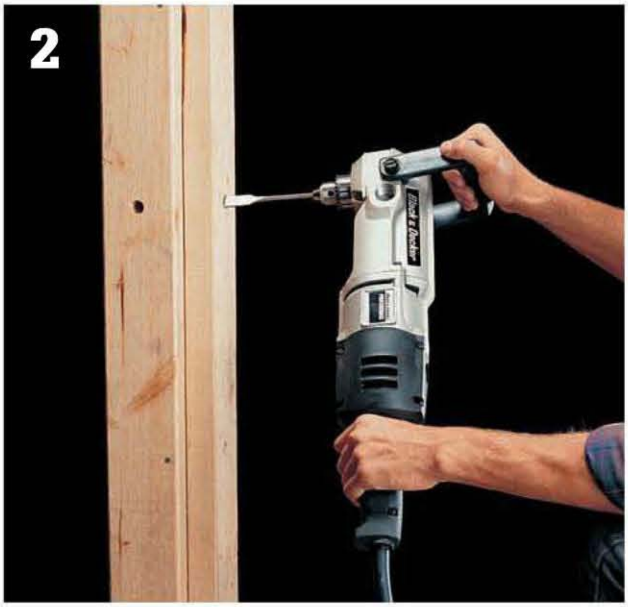

Where cables will turn corners (step 6, page 99), drill intersecting holes in adjoining faces of studs. Measure and cut all cables, allowing 2 ft. extra at ends entering the breaker panel and 1 foot for ends entering the electrical box.

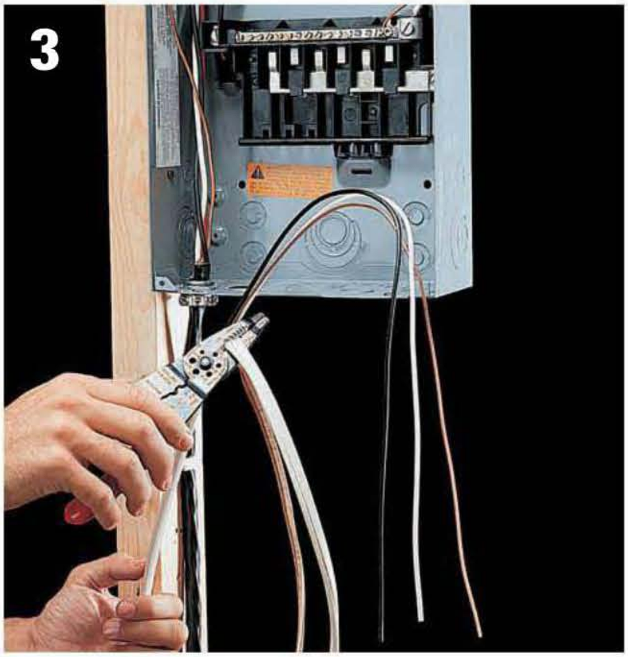

Shut off power to circuit breaker panel. Use a cable ripper to strip cable, leaving at least ¼" of sheathing to enter the circuit breaker panel. Clip away the excess sheathing.

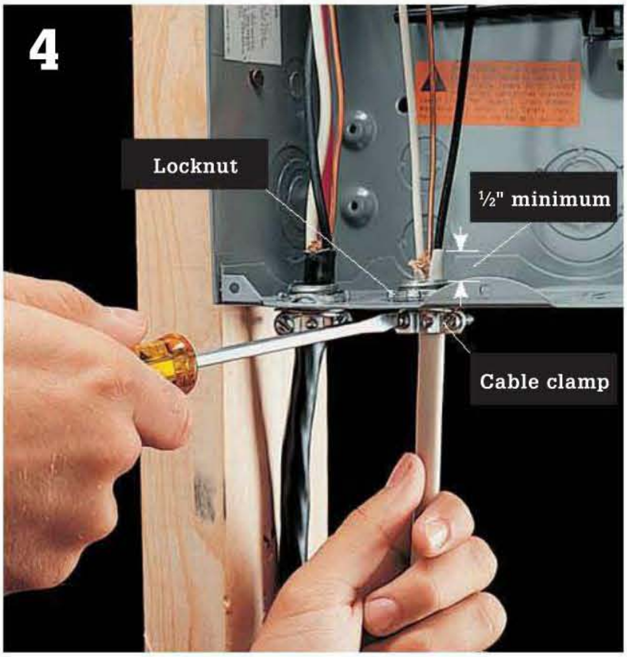

Open a knockout in the circuit breaker panel using a hammer and screwdriver. Insert a cable clamp into the knockout, and secure it with a locknut. Insert the cable through the clamp so that at least ½" of sheathing extends inside the circuit breaker panel. Tighten the mounting screws on the clamp so the cable is gripped securely but not so tightly that the sheathing is crushed.

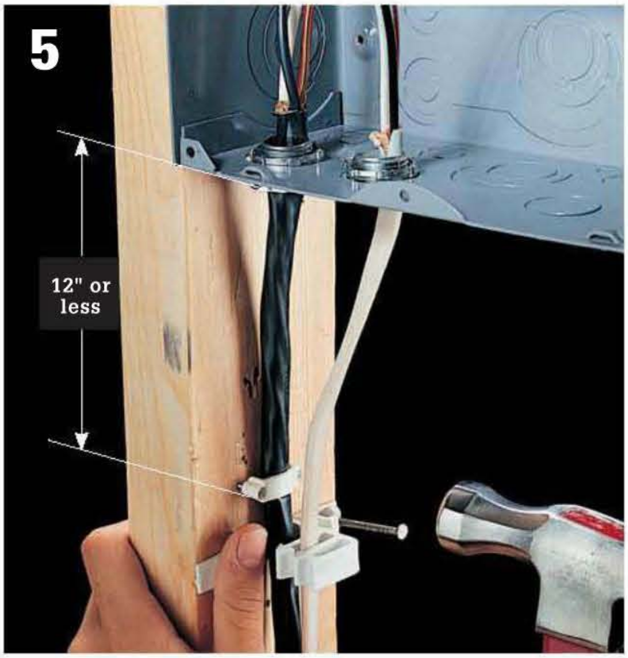

Anchor the cable to the center of a framing member within 12" of the circuit breaker panel using a cable staple. Stack-It® staples work well where two or more cables must be anchored to the same side of a stud. Run the cable to the first electrical box. Where the cable runs along the sides of framing members, anchor it with cable staples no more than 4 ft. apart.

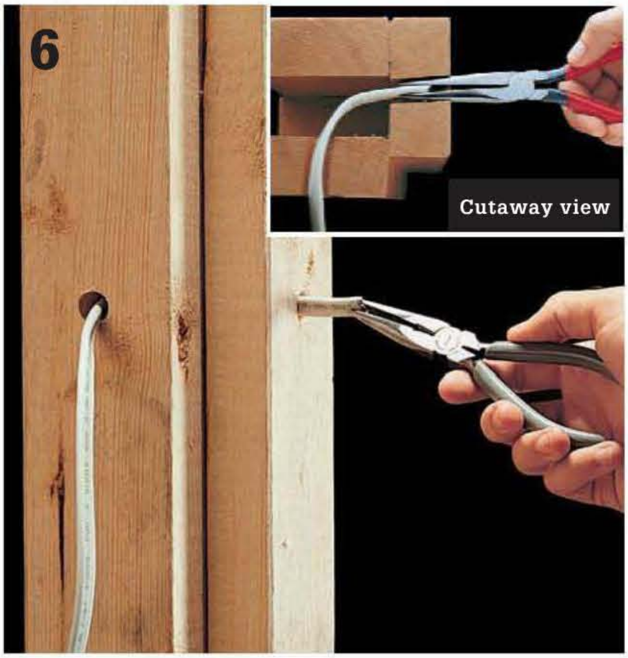

At corners, form a slight L-shaped bend in the end of the cable and insert it into one hole. Retrieve the cable through the other hole using needlenose pliers (inset).

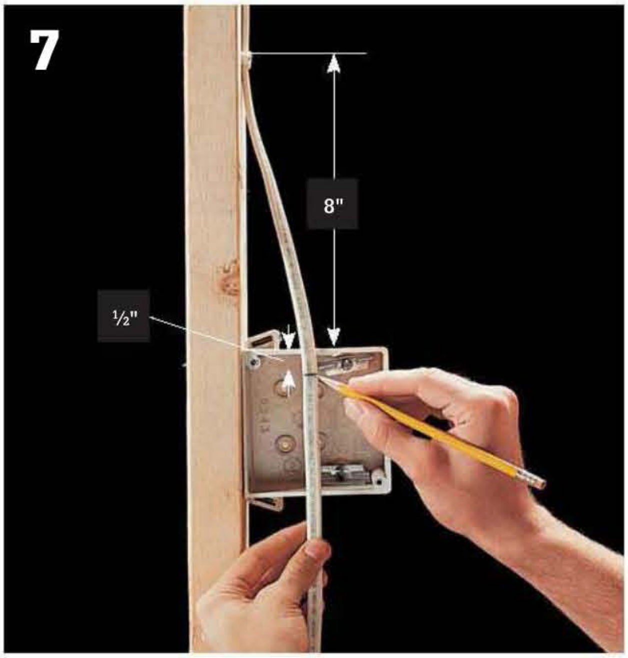

Staple the cable to a framing member 8" from the box. Hold the cable taut against the front of the box, and mark a point on the sheathing ½" past the box edge. Remove sheathing from the marked line to the end using a cable ripper, and clip away excess sheathing with a combination tool. Insert the cable through the knockout in the box.

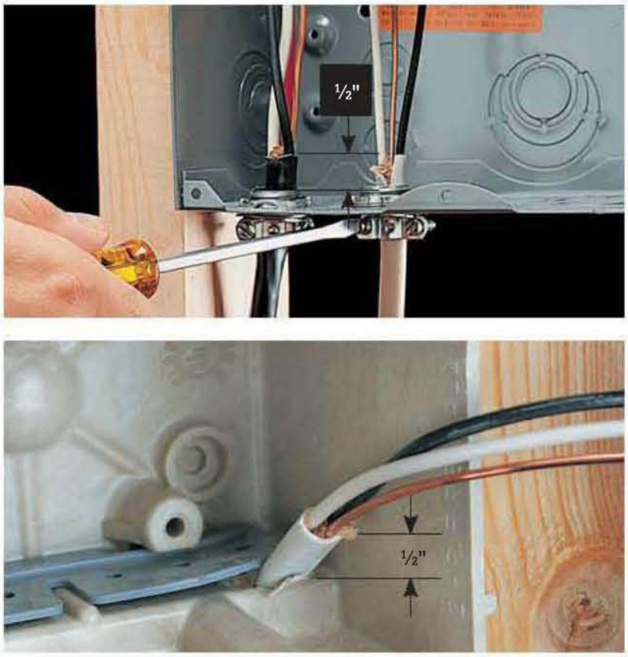

Variation: Different types of boxes have different clamping devices. Make sure cable sheathing extends ½" past the edge of the clamp to ensure that the cable is secure and that the wire won't be damaged by the edges of the clamp.

(continued)

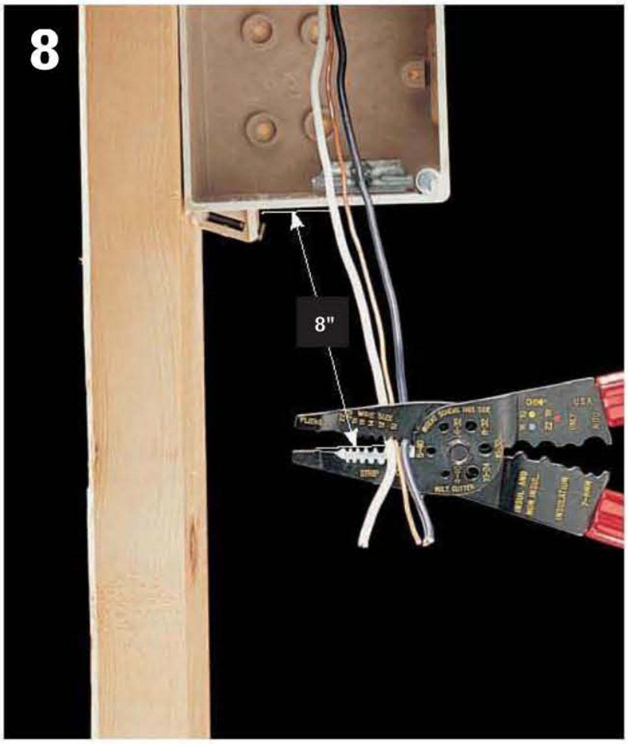

As each cable is installed in a box, clip back each wire so that 8" of workable wire extends past the front edge of the box.

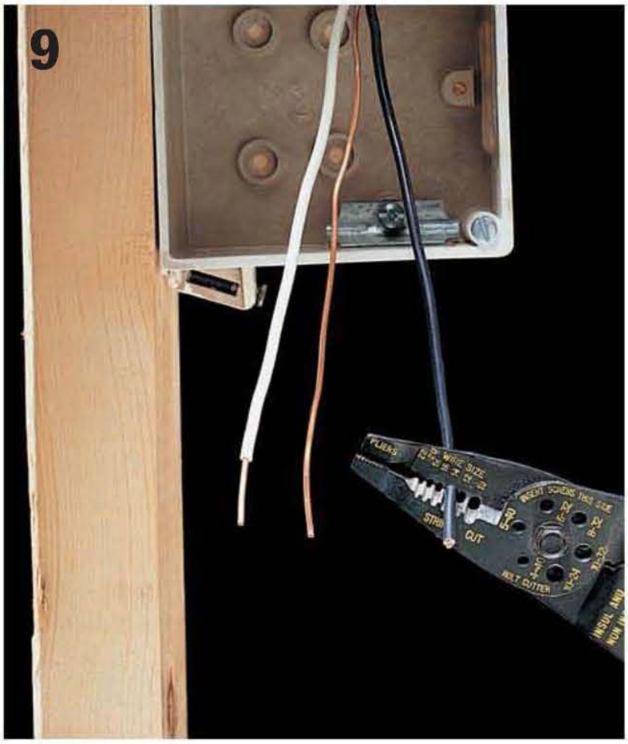

Strip ¾" of insulation from each circuit wire in the box using a combination tool. Take care not to nick the copper.

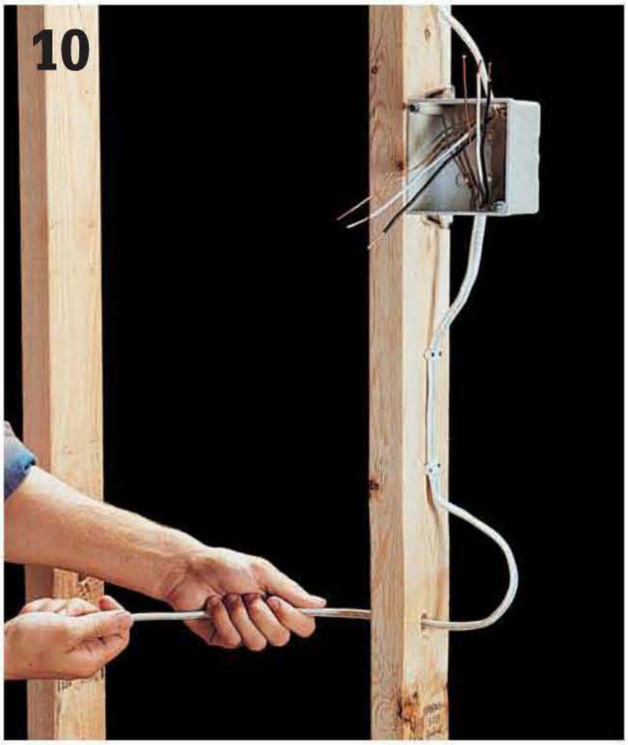

Continue the circuit by running cable between each pair of electrical boxes, leaving an extra 1 ft. of cable at each end.

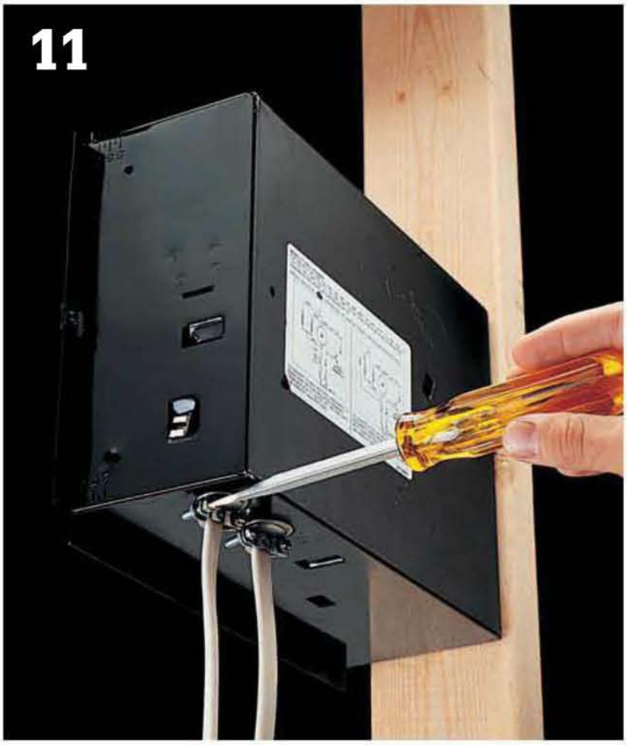

At metal boxes and recessed fixtures, open knockouts, and attach cables with cable clamps. From inside fixture, strip away all but ¼" of sheathing. Clip back wires so there is 8" of workable length, then strip ¾" of insulation from each wire.

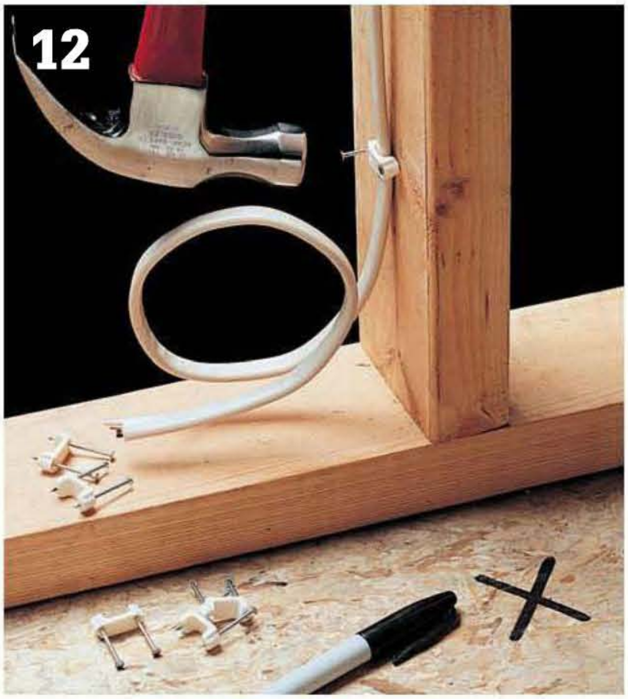

For a surface-mounted fixture like a baseboard heater or fluorescent light fixture, staple the cable to a stud near the fixture location, leaving plenty of excess cable. Mark the floor so the cable will be easy to find after the walls are finished.

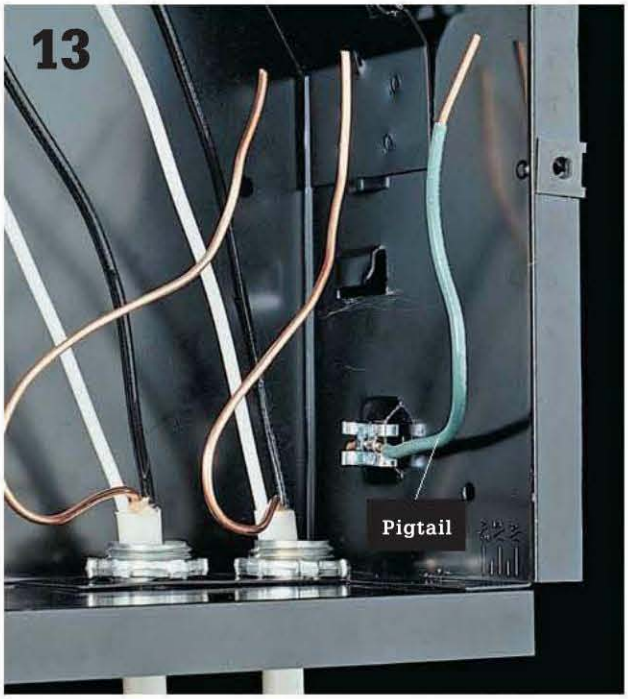

At each recessed fixture and metal electrical box, connect one end of a grounding pigtail to the metal frame using a grounding clip attached to the frame (shown above) or a green grounding screw.

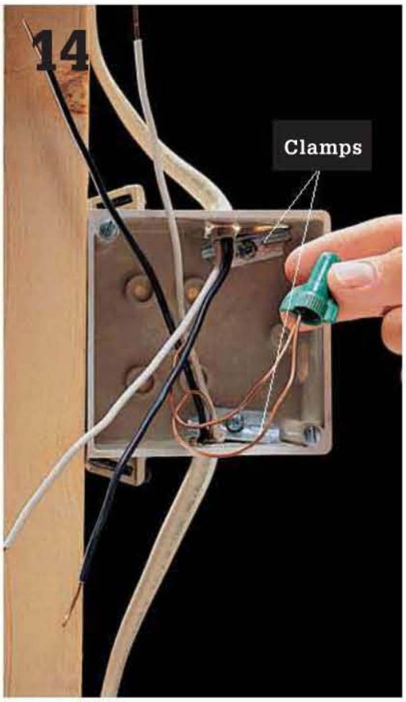

At each electrical box and recessed fixture, join grounding wires together with a wire connector. If the box has internal clamps, tighten the clamps over the cables.

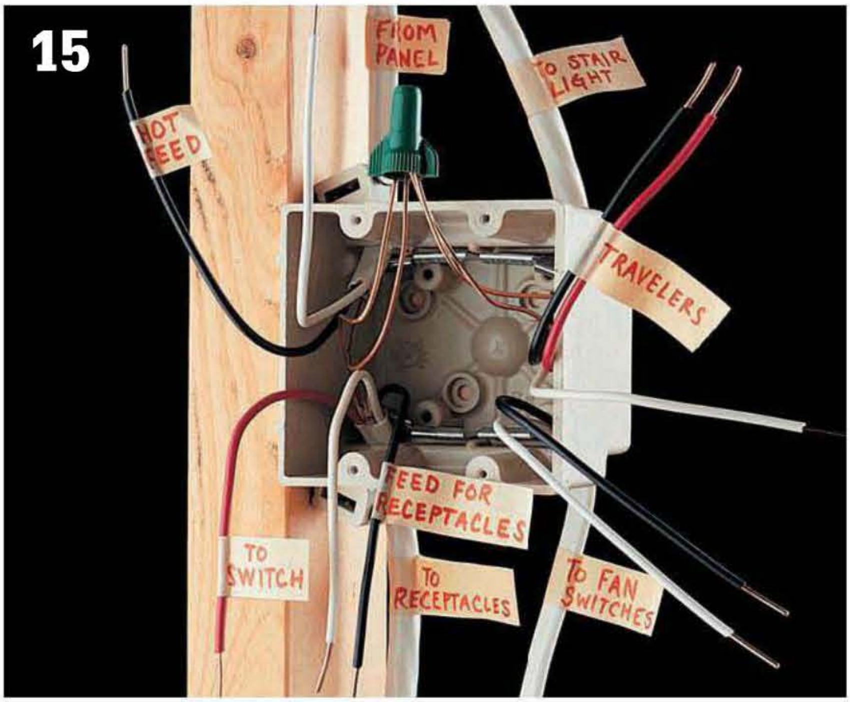
15
FROM PANEL
TO STAIR LIGHT
HOT FEED
TRAVELERS
FEED FOR RECEPTACLES
TO SWITCH
TO RECEPTACLES
TO FAN SWITCHES

Label the cables entering each box to indicate their destinations. In boxes with complex wiring configurations, also tag the individual wires to make final hookups easier. After all cables are installed, your rough-in work is ready to be reviewed by the electrical inspector.

Working with Conduit

Electrical wiring that runs in exposed locations must be protected by rigid tubing called conduit. For example, conduit is used for wiring that runs across masonry walls in a basement laundry and for exposed outdoor wiring. THHN/THWN wire (page 92) normally is installed inside conduit, although UF or NM cable can also be installed in conduit.

There are several types of conduit available, so check with your electrical inspector to find out which type meets code requirements in your area. Conduit installed outdoors must be rated for exterior use. Metal conduit should be used only with metal boxes, never with plastic boxes.

At one time, conduit could only be fitted by using elaborate bending techniques and special tools. Now, however, a variety of shaped fittings are available to let a homeowner join conduit easily.

Electrical Grounding in Metal Conduit

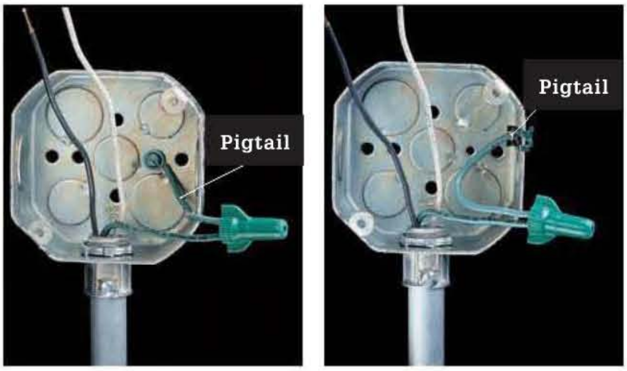
Pigtail
Pigtail

Install a green insulated grounding wire for any circuit that runs through metal conduit. Although code allows the metal conduit to serve as the grounding conductor, most electricians install a green insulated wire as a more dependable means of grounding the system. The grounding wires must be connected to metal boxes with a pigtail and grounding screw (left) or grounding clip (right).

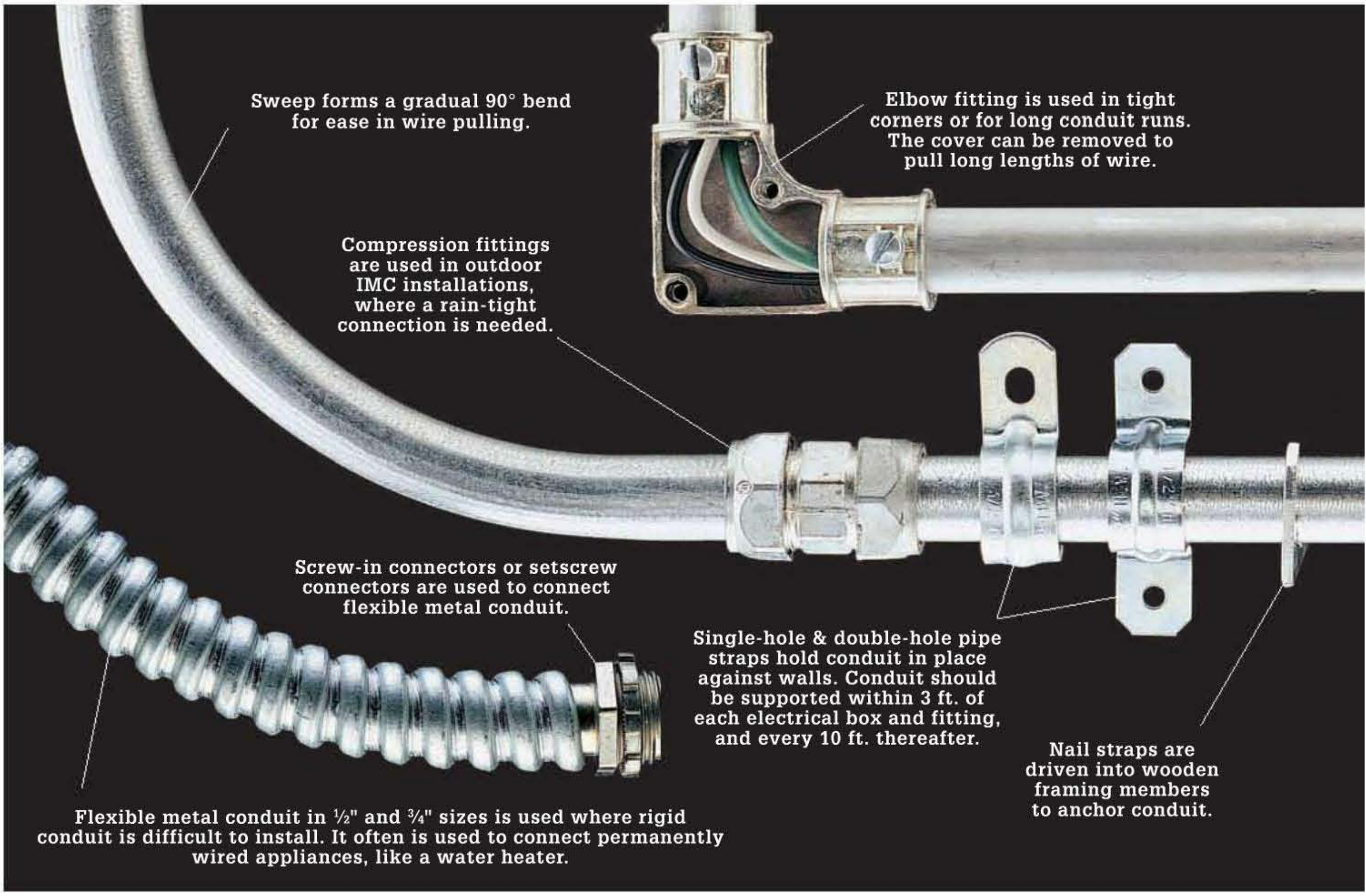
Sweep forms a gradual 90° bend for ease in wire pulling.
Elbow fitting is used in tight corners or for long conduit runs. The cover can be removed to pull long lengths of wire.
Compression fittings are used in outdoor IMC installations, where a rain-tight connection is needed.
Screw-in connectors or setscrew connectors are used to connect flexible metal conduit.
Single-hole & double-hole pipe straps hold conduit in place against walls. Conduit should be supported within 3 ft. of each electrical box and fitting, and every 10 ft. thereafter.
Nail straps are driven into wooden framing members to anchor conduit.
Flexible metal conduit in ½" and ¾" sizes is used where rigid conduit is difficult to install. It often is used to connect permanently wired appliances, like a water heater.

Fill Capacity

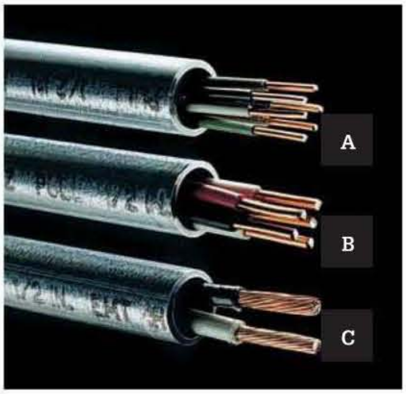

Conduit ½" in diameter can hold up to six 14-gauge or 12-gauge THHN/THWN wires (A), five 10-gauge wires (B), or two 8-gauge wires (C). Use ¾" conduit for greater capacity.

Metal Conduit

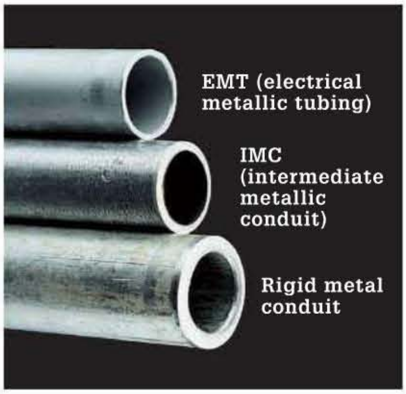
EMT (electrical metallic tubing)
IMC (intermediate metallic conduit)
Rigid metal conduit

EMT is lightweight and easy to install but should not be used where it can be damaged. IMC has thicker galvanized walls and is a good choice for exposed outdoor use. Rigid metal conduit provides the greatest protection for wires, but it is more expensive and requires threaded fittings.

Plastic Conduit

Plastic PVC conduit is allowed by many local codes. It is assembled with solvent glue and PVC fittings that resemble those for metal conduit. When wiring with PVC conduit, always run a green grounding wire.

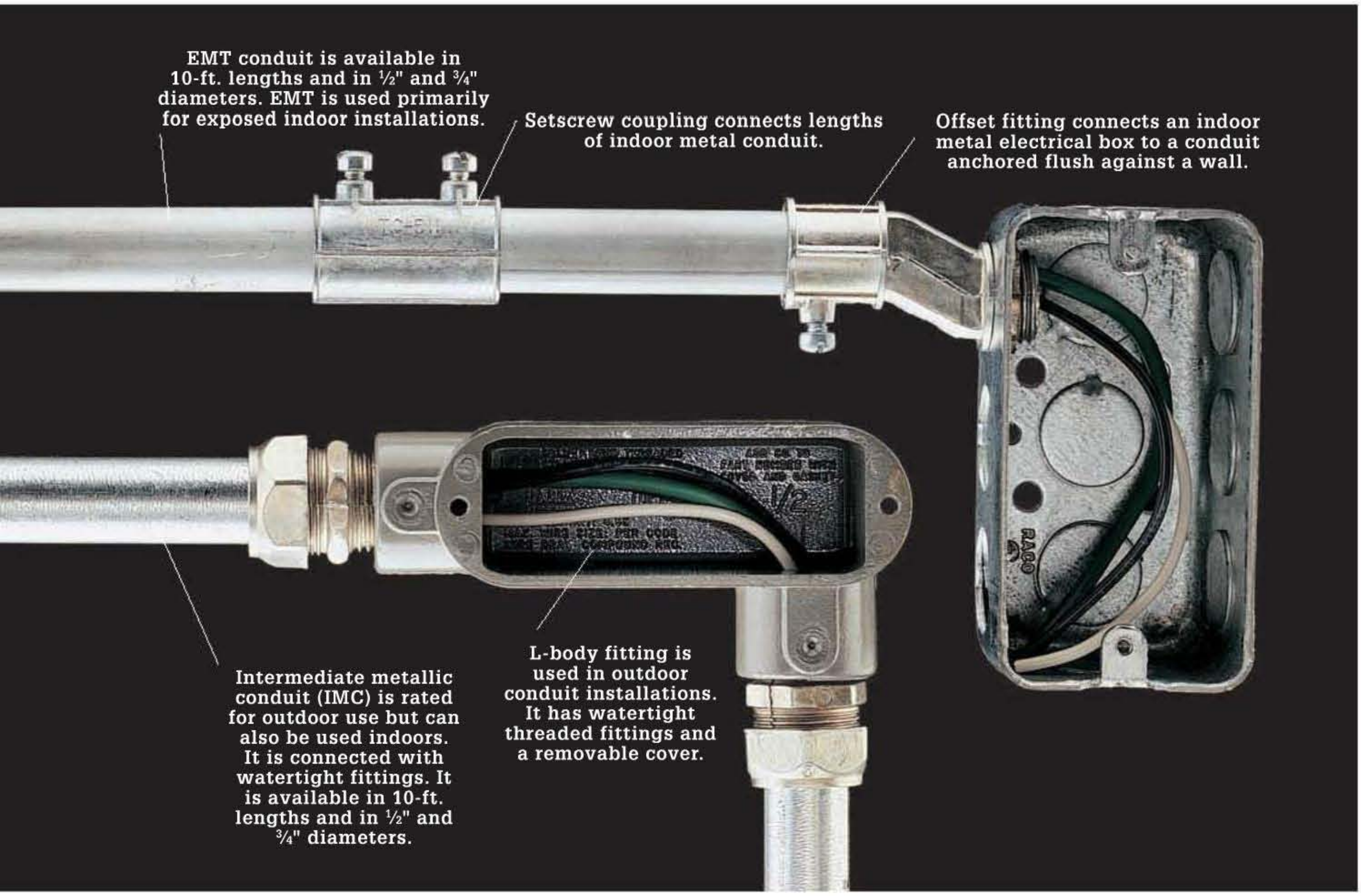

EMT conduit is available in 10-ft. lengths and in ½" and ¾" diameters. EMT is used primarily for exposed indoor installations.

Setscrew coupling connects lengths of indoor metal conduit.

Offset fitting connects an indoor metal electrical box to a conduit anchored flush against a wall.

Intermediate metallic conduit (IMC) is rated for outdoor use but can also be used indoors. It is connected with watertight fittings. It is available in 10-ft. lengths and in ½" and ¾" diameters.

L-body fitting is used in outdoor conduit installations. It has watertight threaded fittings and a removable cover.

Conduit Materials & Tools

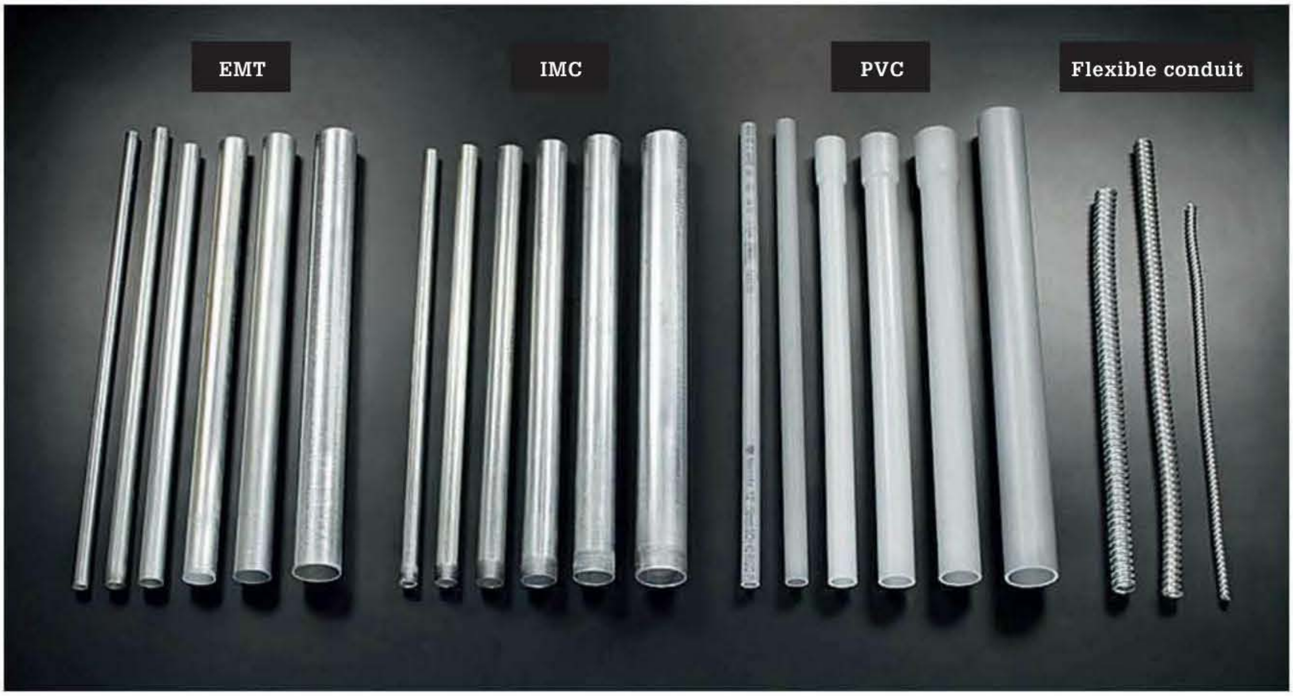
EMT
IMC
PVC
Flexible conduit

Conduit types used most in homes are EMT (electrical metallic tubing), IMC (intermediate metallic conduit), RNC (rigid nonmetallic conduit), and flexible metal conduit. The most common diameters by far are ½" and ¾", but larger sizes are stocked at most building centers.

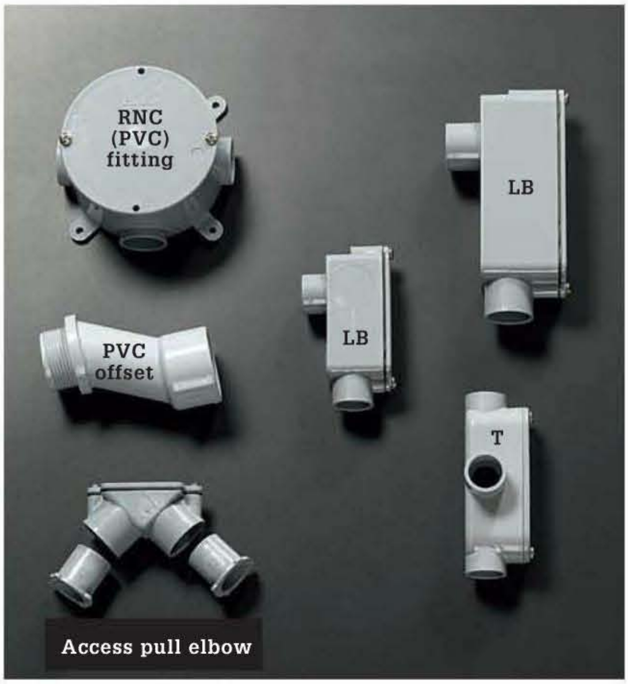
RNC (PVC) fitting
LB
LB
PVC offset
T
Access pull elbow

Nonmetallic conduit fittings typically are solvent welded to nonmetallic conduit, as opposed to metal conduit, which can be threaded and screwed into threaded fittings or attached with setscrews or compression fittings.

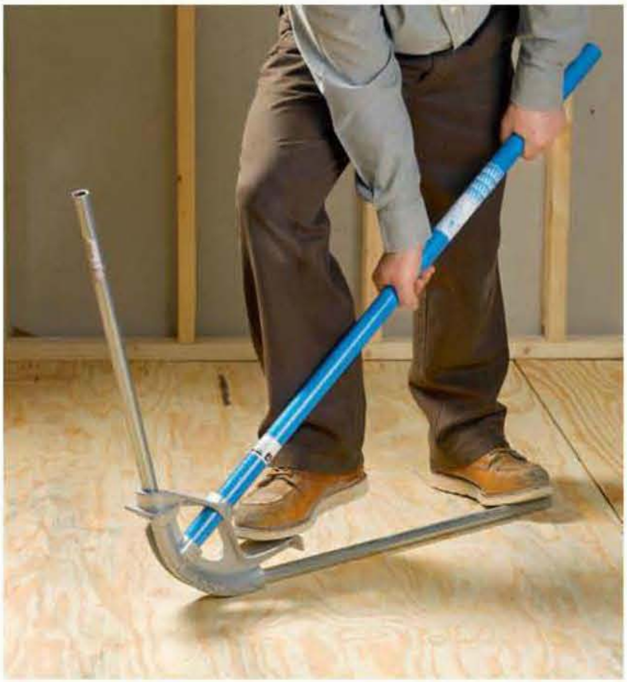

A thin-wall conduit bender is used to bend sweeps into EMT or IMC conduit.

How to Make Nonmetallic Conduit Connections

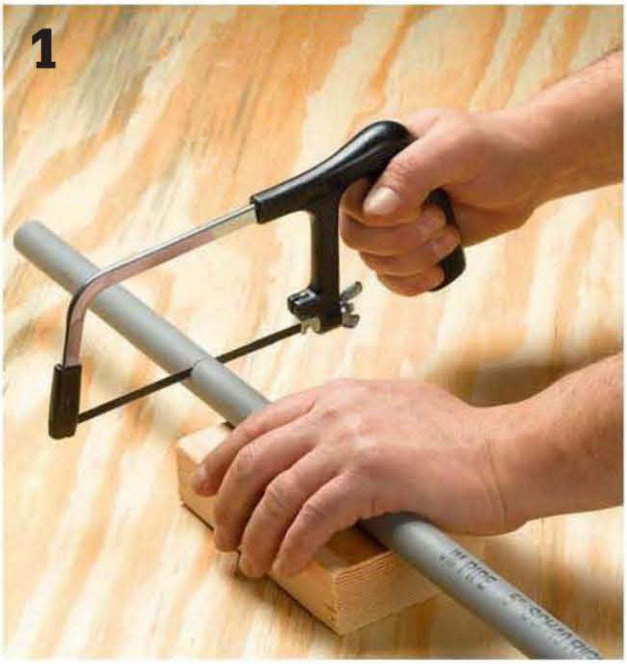

Cut the rigid nonmetallic conduit (RNC) to length with a fine-tooth saw, such as a hacksaw. For larger diameter (1½" and above), use a power miter box with a fine-tooth or plastic cutting blade.

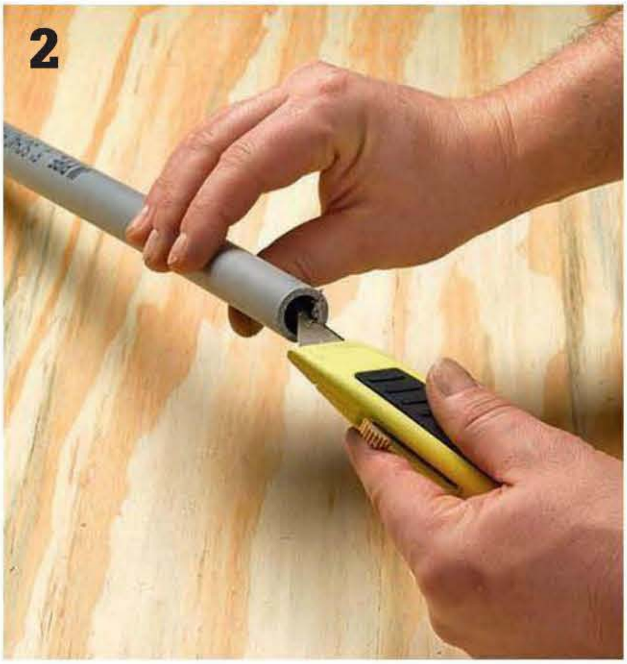

Deburr the cut edges with a utility knife or fine sandpaper such as emery paper. Wipe the cut ends with a dry rag. Also wipe the coupling or fitting to clean it.

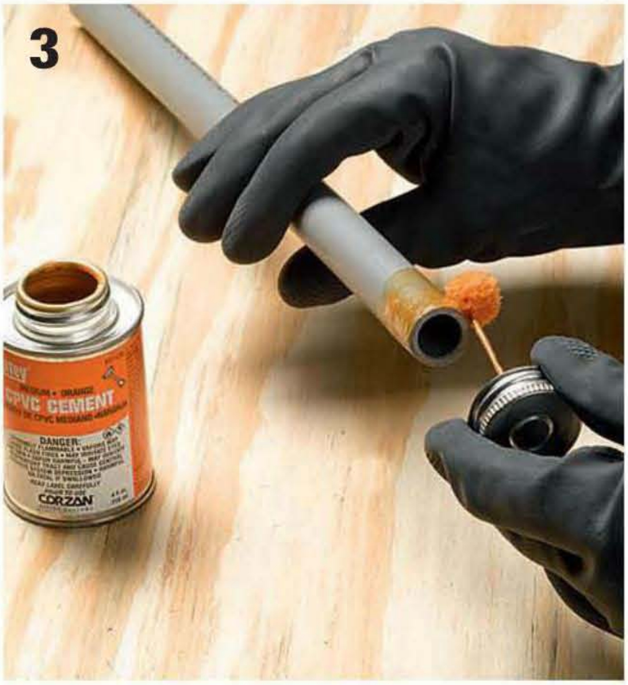

Apply a coat of PVC cement to the end of the conduit. Wear latex gloves to protect your hands. The cement should be applied past the point on the conduit where it enters the fitting or coupling.

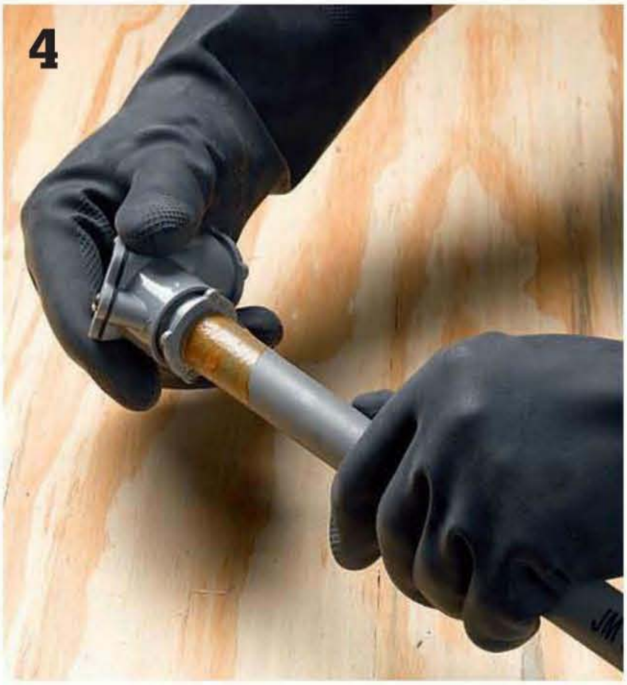

Insert the conduit into the fitting or coupling and spin it a quarter turn to help spread the cement. Allow the joint to set undisturbed for 10 minutes.

How to Install Conduit & Wires on a Concrete Wall

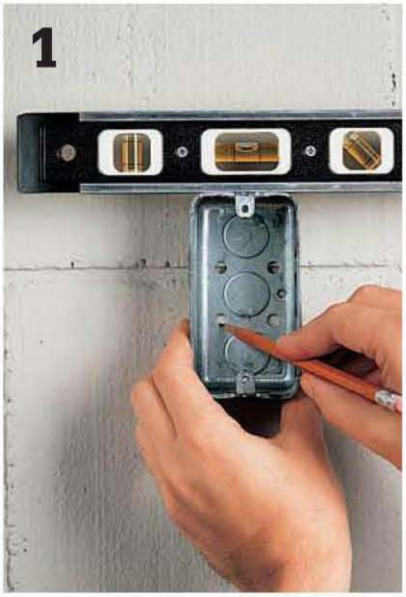

Measure from the floor to position electrical boxes on the wall, and mark location for mounting screws. Boxes for receptacles in an unfinished basement or other damp area are mounted at least 2 ft. from the floor. Laundry receptacles usually are mounted at 48".

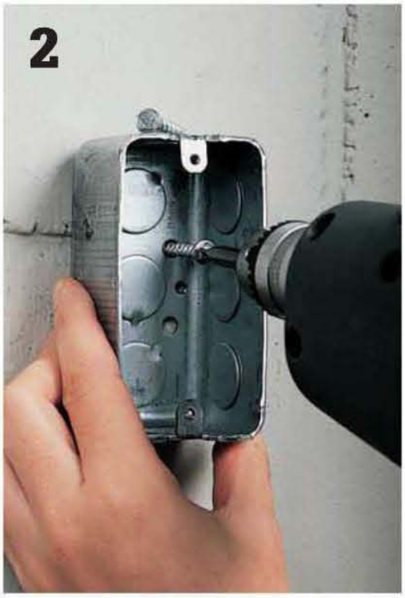

Drill pilot holes with a masonry bit, then mount the box against a masonry wall with masonry anchors. Or use masonry anchors and panhead screws.

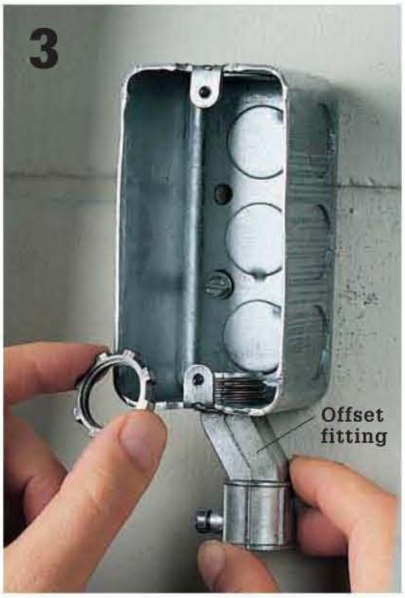

Open one knockout for each length of conduit that will be attached to the box. Attach an offset fitting to each knockout using a locknut.

Measure the first length of conduit and cut it with a hacksaw. Remove any rough inside edges with a pipe reamer or a round file. Attach the conduit to the offset fitting on the box, and tighten the setscrew.

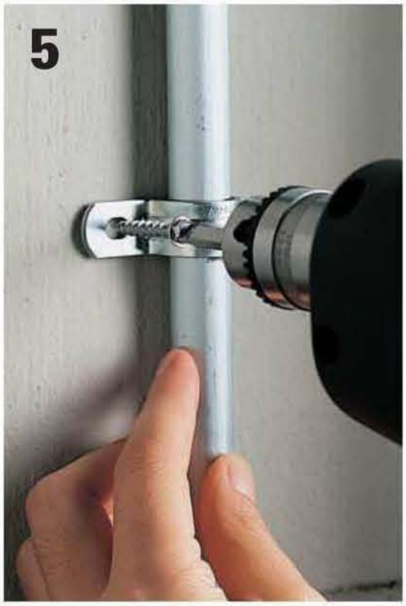

Anchor the conduit against the wall with pipe straps and masonry anchors. Conduit should be anchored within 3 ft. of each box and fitting, and every 10 ft. thereafter.

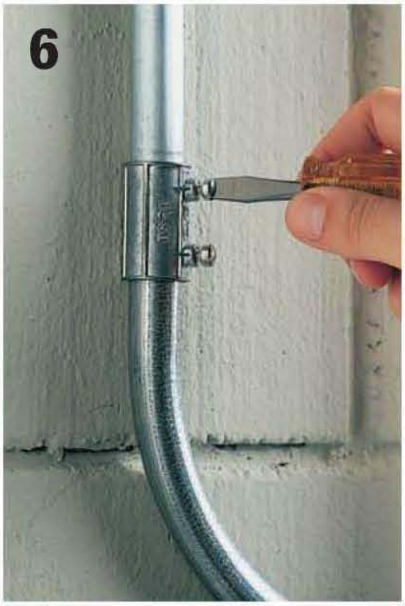

Make conduit bends by attaching a sweep fitting using a setscrew fitting or compression fitting. Continue conduit run by attaching additional lengths using setscrew or compression fittings.

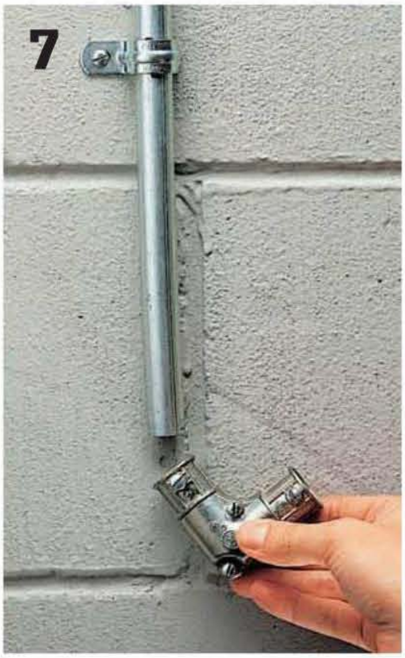
7

Use an elbow fitting in conduit runs that have many bends, or in runs that require very long wires. The cover on the elbow fitting can be removed to make it easier to extend a fish tape and pull wires.

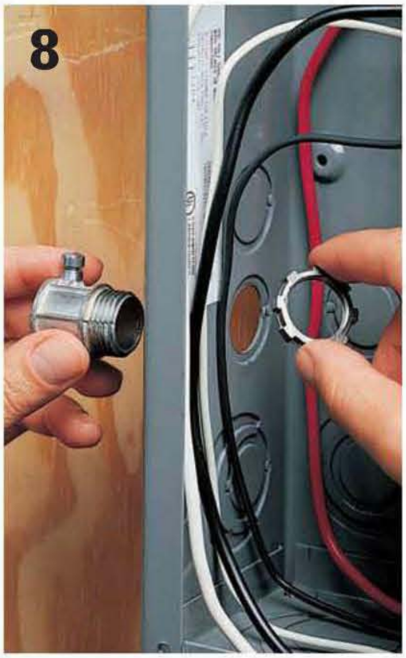
8

At the service breaker panel, turn the power off, then remove the cover and test for power. Open a knockout in the panel, then attach a setscrew fitting, and install the last length of conduit.

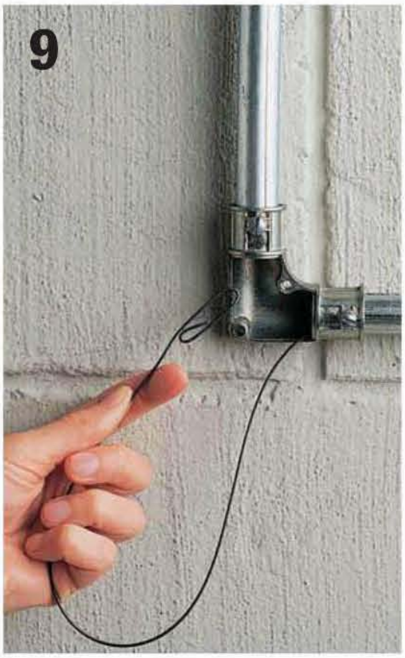
9

Unwind the fish tape and extend it through the conduit from the circuit breaker panel outward. Remove the cover on an elbow fitting when extending the fish tape around tight corners.

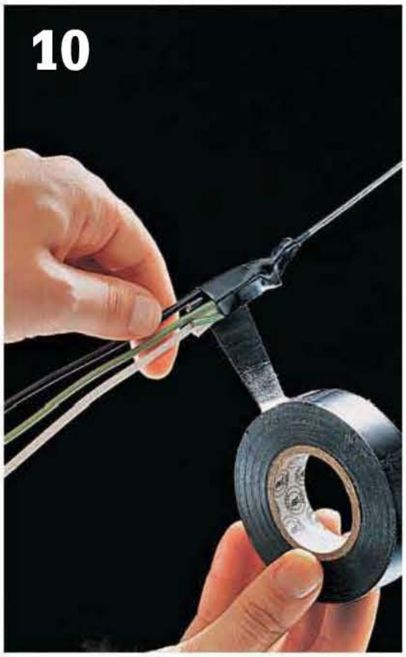
10

Trim back 2" of outer insulation from the end of the wires, then insert the wires through the loop at the tip of the fish tape.

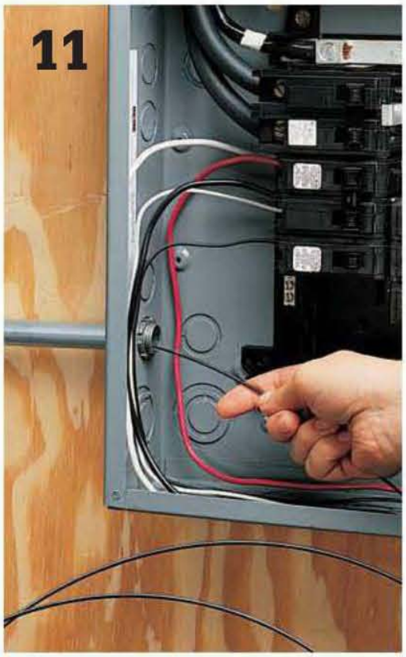
11

Retrieve the wires through the conduit by pulling on the fish tape with steady pressure. *Note: Use extreme care when using a metal fish tape inside a circuit breaker panel, even when the power is turned off.*

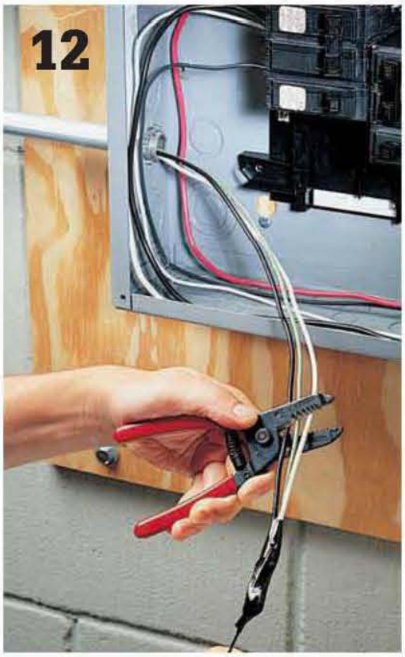
12

Clip off the taped ends of the wires. Leave at least 2 ft. of wire at the service panel and 8" at each electrical box.

Working with Boxes

The National Electrical Code requires that wire connections and cable splices be contained inside an approved metal or plastic box. This shields framing members and other flammable materials from electrical sparks.

Electrical boxes come in several shapes. Rectangular and square boxes are used for switches and receptacles. Rectangular (2 × 3") boxes are used for single switches or duplex receptacles. Square (4 × 4") boxes are used any time it is convenient for two switches or receptacles to be wired, or "ganged," in one box, an arrangement common in kitchens or entry hallways. Octagonal electrical boxes contain wire connections for ceiling fixtures.

All electrical boxes are available in different depths. A box must be deep enough so a switch or receptacle can be removed or installed easily without crimping and damaging the circuit wires. Replace an undersized box with a larger box using the Electrical Box Chart (right) as a guide. The NEC also says that all electrical boxes must remain accessible. Never cover an electrical box with drywall, paneling, or wallcoverings.

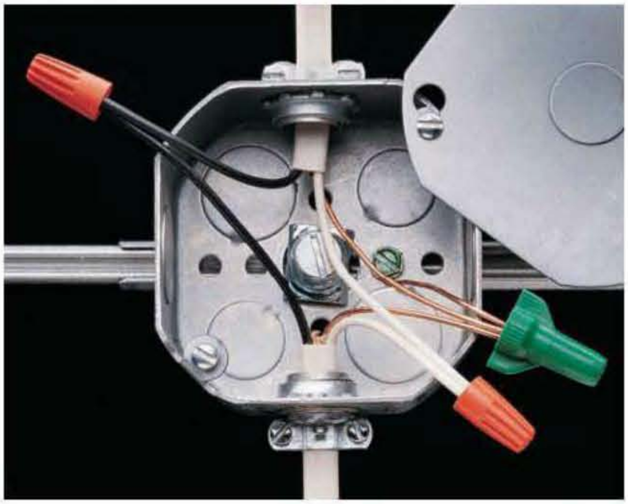

Octagonal boxes usually contain wire connections for ceiling fixtures. Cables are inserted into the box through knockout openings and are held with cable clamps. Because the ceiling fixture attaches directly to the box, the box should be anchored firmly to a framing member. Often, it is nailed directly to a ceiling joist. However, metal braces are available that allow a box to be mounted between joists or studs. A properly installed octagonal box can support a ceiling fixture weighing up to 35 pounds. Any box must be covered with a tightly fitting cover plate, and the box must not have open knockouts.

Electrical Box Fill Chart ▸

Box size and shape	Maximum number of conductors permitted (see Notes below)			
	18 AWG	16 AWG	14 AWG	12 AWG
Junction boxes				
4 × 1¼" R or O	8	7	6	5
4 × 1½" R or O	10	8	7	6
4 × 2⅛" R or O	14	12	10	9
4 × 1¼" S	12	10	9	8
4 × 1½" S	14	12	10	9
4 × 2⅛" S	20	17	15	13
4 11/16 × 1¼" S	17	14	12	11
4 11/16 × 1½" S	19	16	14	13
4 11/16 × 2⅛" S	28	24	21	18
Device boxes				
3 × 2 × 1½"	5	4	3	3
3 × 2 × 2"	6	5	5	4
3 × 2 × 2¼"	7	6	5	4
3 × 2 × 2½"	8	7	6	5
3 × 2 × 2¾"	9	8	7	6
3 × 2 × 3½"	12	10	9	8
4 × 2⅛ × 1½"	6	5	5	4
4 × 2⅛ × 1⅞"	8	7	6	5
4 × 2⅛ × 2⅛"	9	8	7	6

Notes:

- R = Round; O = Octagonal; S = Square or rectangular
- Each hot or neutral wire entering the box is counted as one conductor.
- Grounding wires are counted as one conductor in total—do not count each one individually.
- Raceway fittings and external cable clamps do not count. Internal cable connectors and straps count as either half or one conductor, depending on type.
- Devices (switches and receptacles mainly) each count as two conductors.
- Straps (yokes) from mounting devices each count as two conductors.
- When calculating total conductors, any nonwire components should be assigned the gauge of the largest wire in the box.
- For wire gauges not shown here, contact your local electrical inspections office.

Common Electrical Boxes

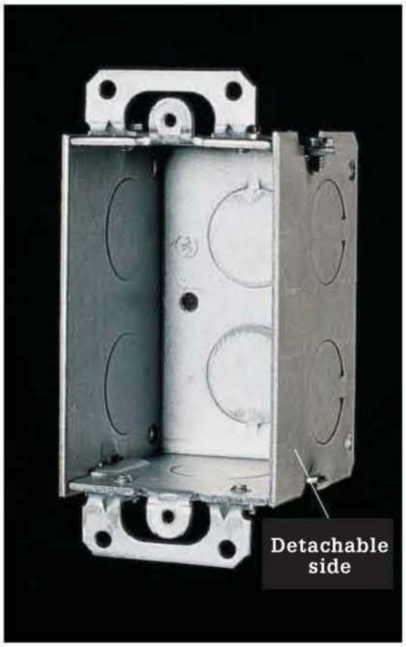

Rectangular boxes are used with wall switches and duplex receptacles. Single-size rectangular boxes (shown above) may have detachable sides that allow them to be ganged together to form double-size boxes.

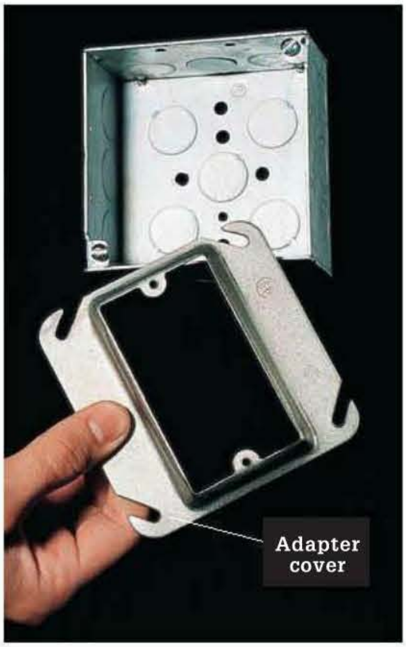

Square 4 × 4" boxes are large enough for most wiring applications. They are used for cable splices and ganged receptacles or switches. To install one switch or receptacle in a square box, use an adapter cover.

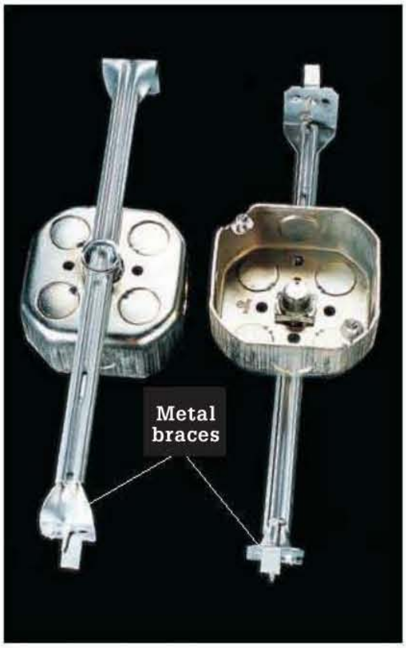

Braced octagonal boxes fit between ceiling joists. The metal braces extend to fit any joist spacing and are nailed or screwed to framing members.

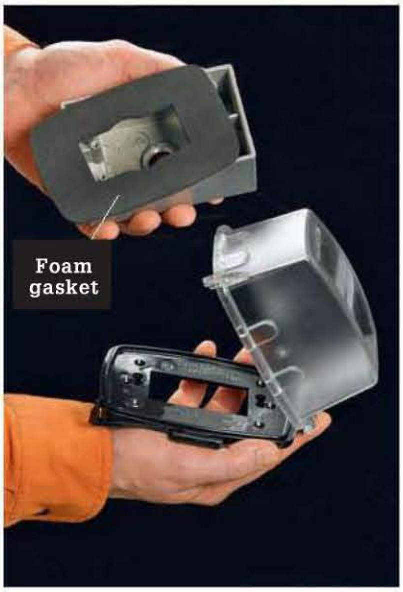

Outdoor boxes have sealed seams and foam gaskets to guard a switch or receptacle against moisture. Corrosion-resistant coatings protect all metal parts. Code-compliant models include a watertight hood.

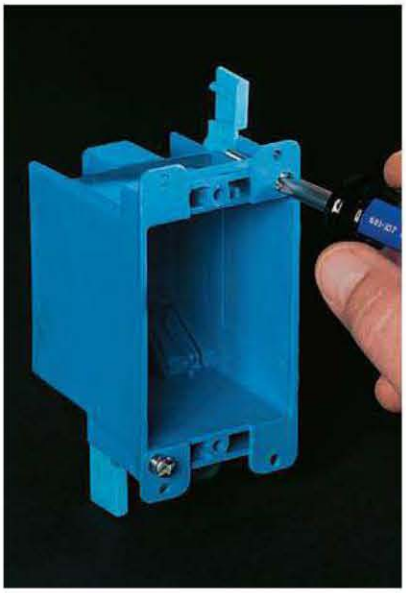

Retrofit boxes upgrade older boxes to larger sizes. One type (above) has built-in clamps that tighten against the inside of a wall and hold the box in place.

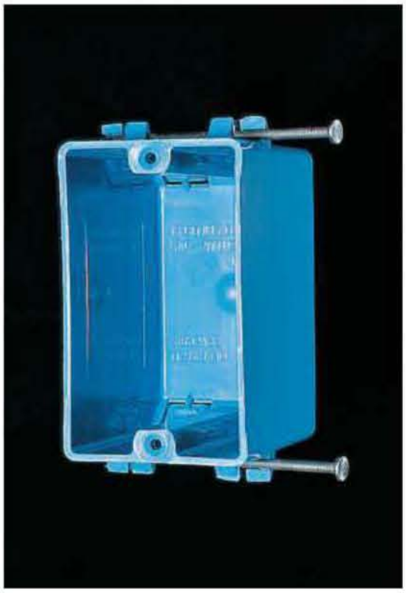

Plastic boxes are common in new construction. They can be used only with NM (nonmetallic) cable. The box may include preattached nails for anchoring it to framing members. Wall switches must have grounding screws if installed in plastic boxes.

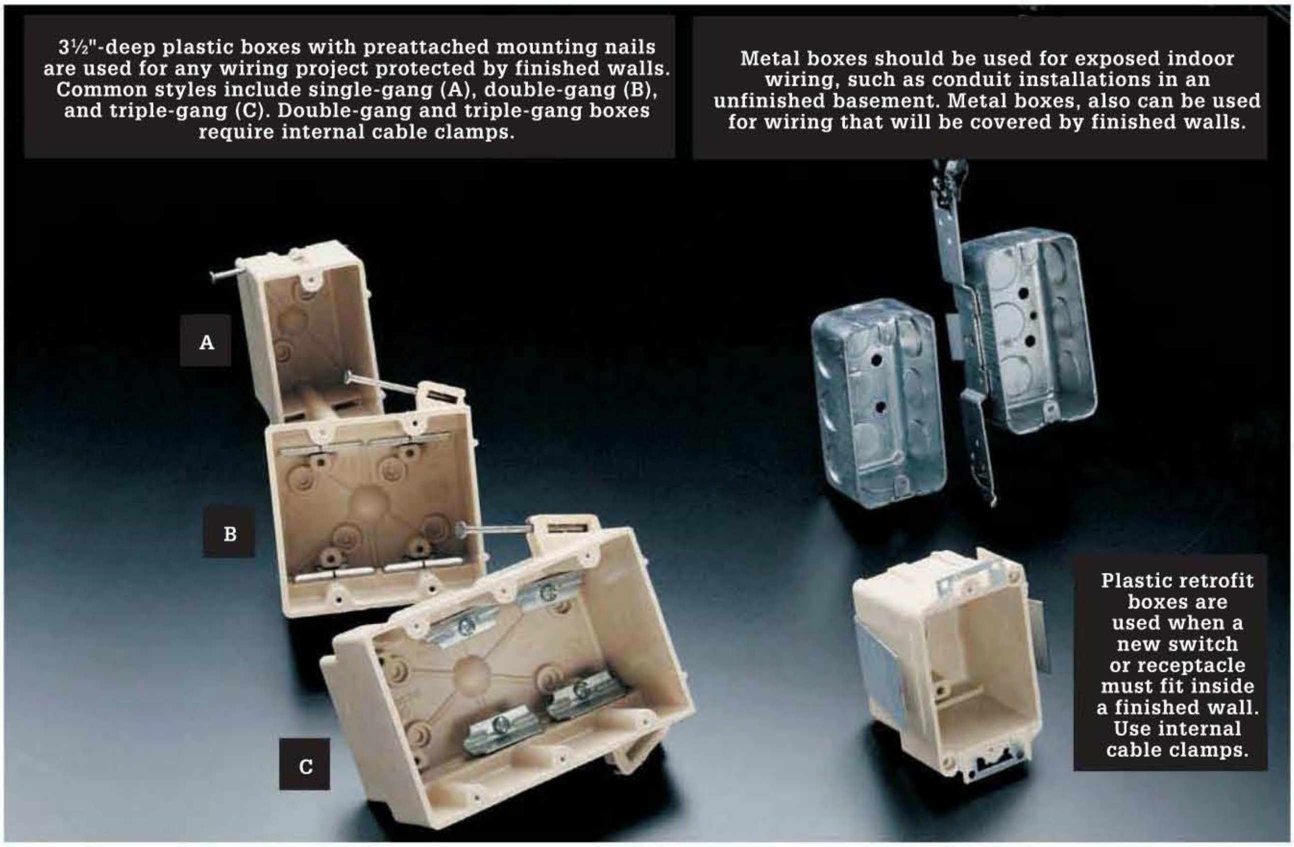

3½"-deep plastic boxes with preattached mounting nails are used for any wiring project protected by finished walls. Common styles include single-gang (A), double-gang (B), and triple-gang (C). Double-gang and triple-gang boxes require internal cable clamps.

Metal boxes should be used for exposed indoor wiring, such as conduit installations in an unfinished basement. Metal boxes, also can be used for wiring that will be covered by finished walls.

Plastic retrofit boxes are used when a new switch or receptacle must fit inside a finished wall. Use internal cable clamps.

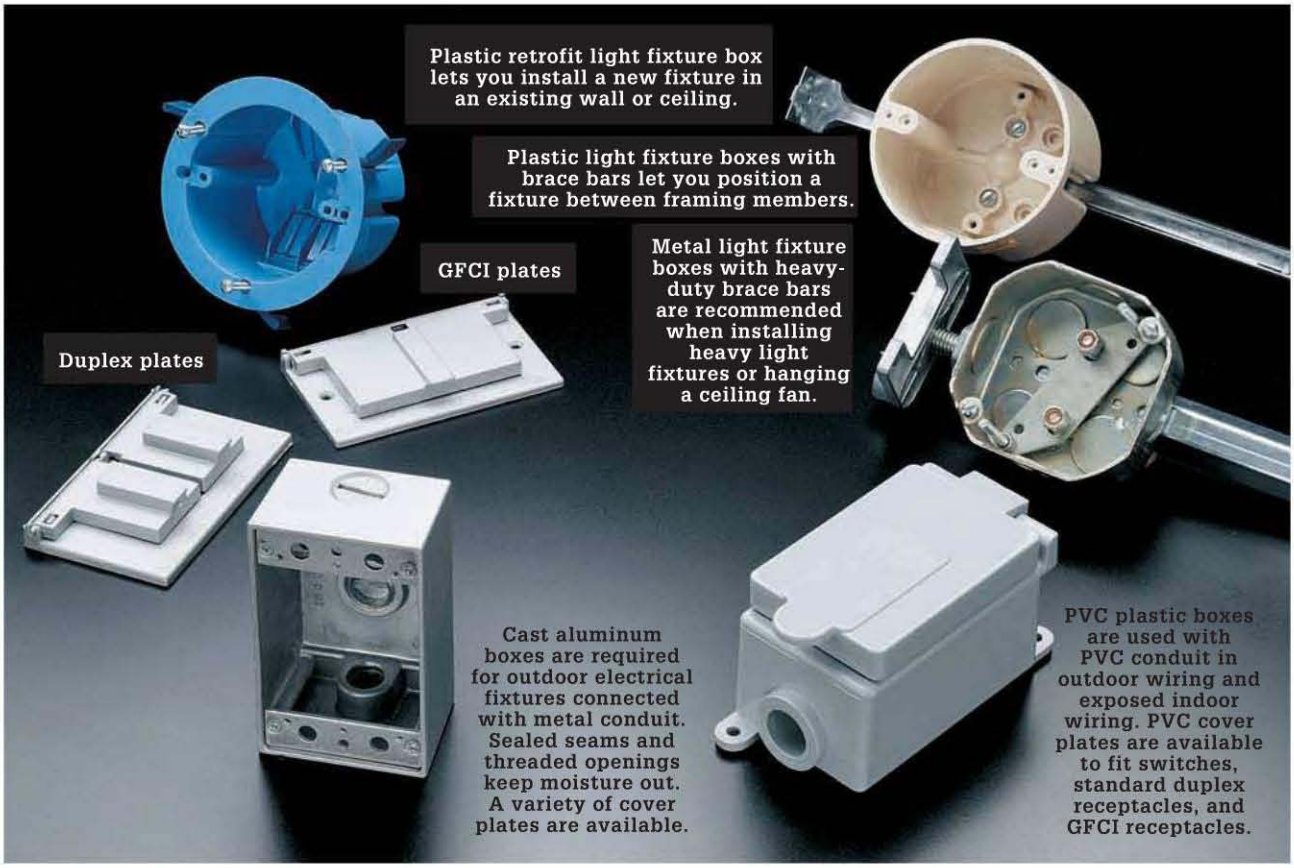

Plastic retrofit light fixture box lets you install a new fixture in an existing wall or ceiling.

Plastic light fixture boxes with brace bars let you position a fixture between framing members.

Metal light fixture boxes with heavy-duty brace bars are recommended when installing heavy light fixtures or hanging a ceiling fan.

GFCI plates

Duplex plates

Cast aluminum boxes are required for outdoor electrical fixtures connected with metal conduit. Sealed seams and threaded openings keep moisture out. A variety of cover plates are available.

PVC plastic boxes are used with PVC conduit in outdoor wiring and exposed indoor wiring. PVC cover plates are available to fit switches, standard duplex receptacles, and GFCI receptacles.

How to Locate Electrical Boxes

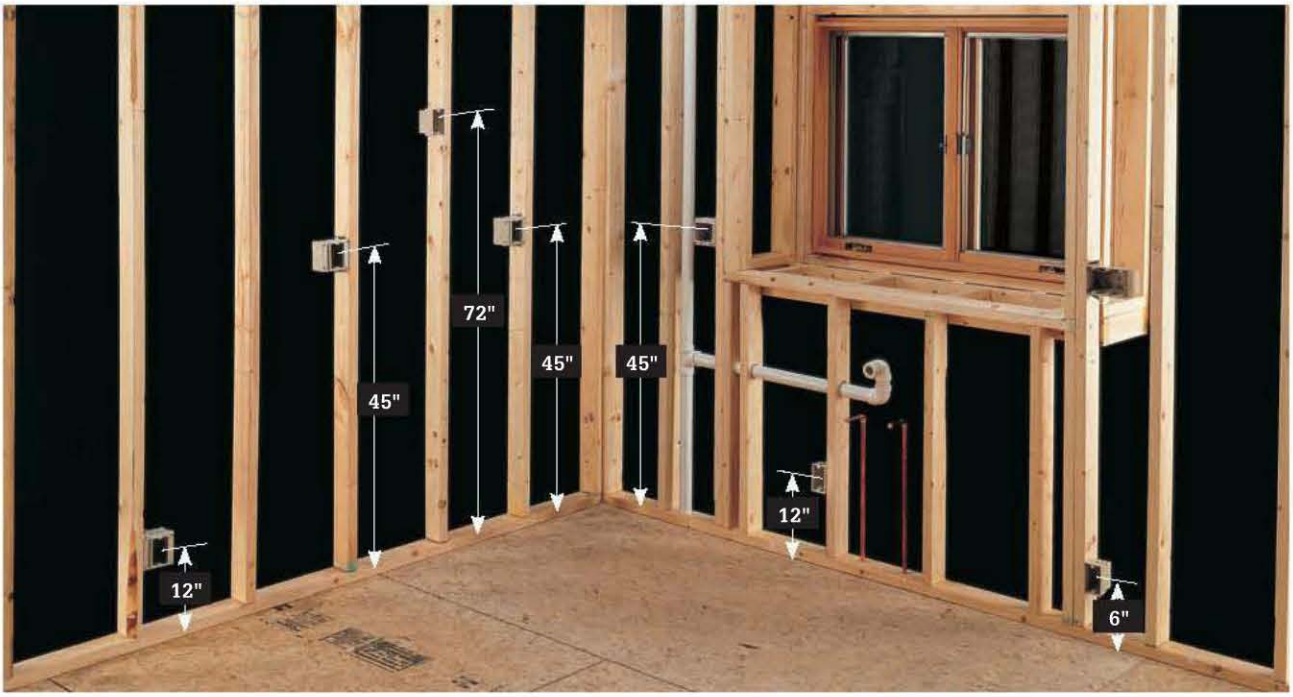

Heights of electrical boxes vary depending on use. In the kitchen shown here, boxes above the countertop are 45" above the floor, in the center of 18" backsplashes that extend from the countertop to the cabinets. All boxes for wall switches also are installed at this height. The center of the box for the microwave receptacle is 72" off the floor. The centers of the boxes for the range and food disposer receptacles are 12" off the floor, but the center of the box for the dishwasher receptacle is 6" off the floor.

Typical Wallcovering Thickness

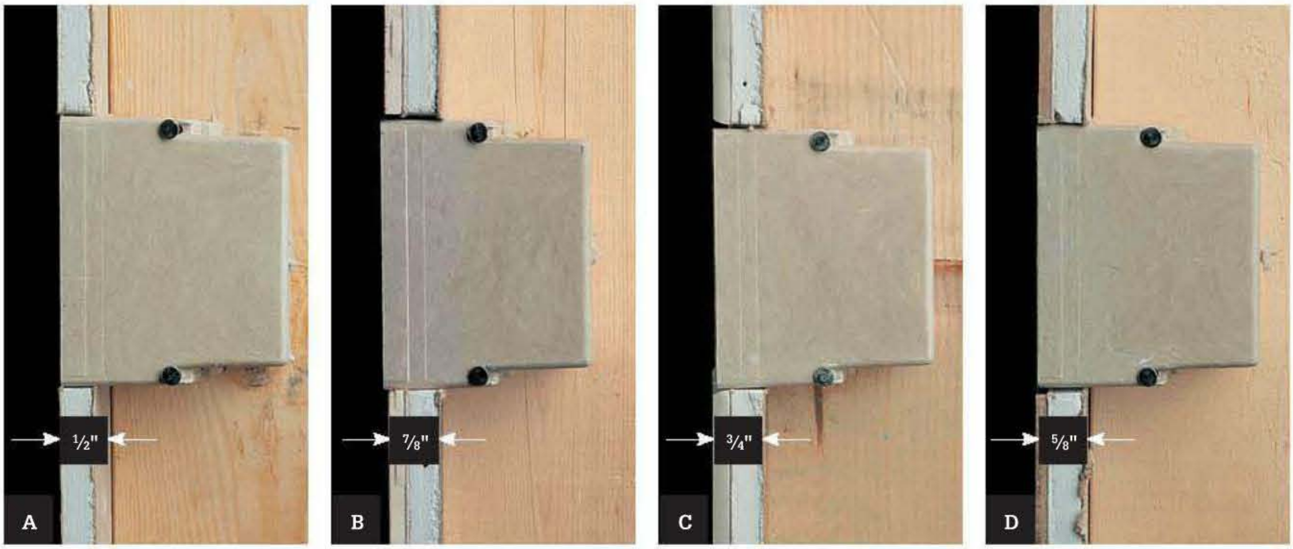

Consider the thickness of finished walls when mounting electrical boxes against framing members. Code requires that the front face of boxes be flush with the finished wall surface, so how you install boxes will vary depending on the type of wall finish that will be used. For example, if the walls will be finished with ½" wallboard (A), attach the boxes so the front faces extend ½" past the front of the framing members. With ceramic tile and wall board (B), extend the boxes ⅞" past the framing members. With ¼" Corian® over wallboard (C), boxes should extend ¾"; and with wallboard and laminate (D), boxes extend ⅝".

How to Install Electrical Boxes for Receptacles

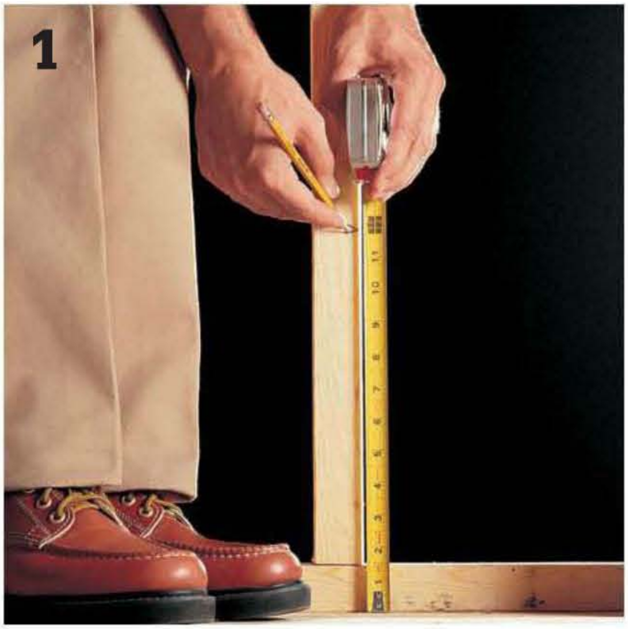

Mark the location of each box on studs. Standard receptacle boxes should be centered 12" above floor level. GFCI receptacle boxes in a bathroom should be mounted so they will be about 10" above the finished countertop.

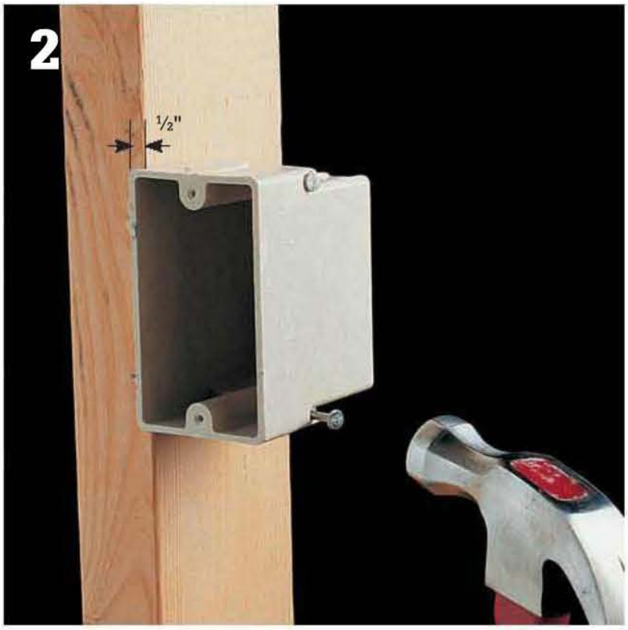

Position each box against a stud so the front face will be flush with the finished wall. For example, if you will be installing ½" wallboard, position the box so it extends ½" past the face of the stud. Anchor the box by driving the mounting nails into the stud.

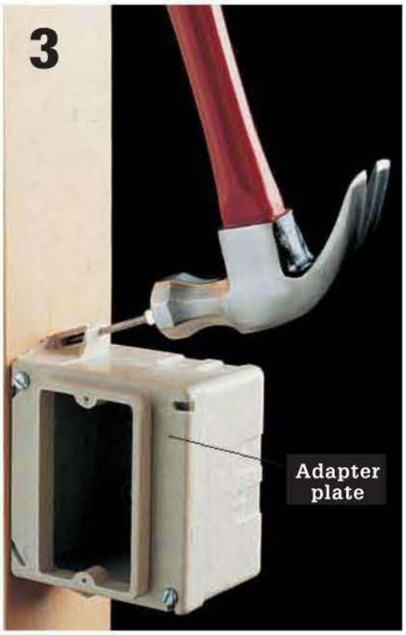

If installing square boxes, attach the adapter plates before positioning the boxes. Use adapter plates that match the thickness of the finished wall. Anchor the box by driving the mounting nails into the stud.

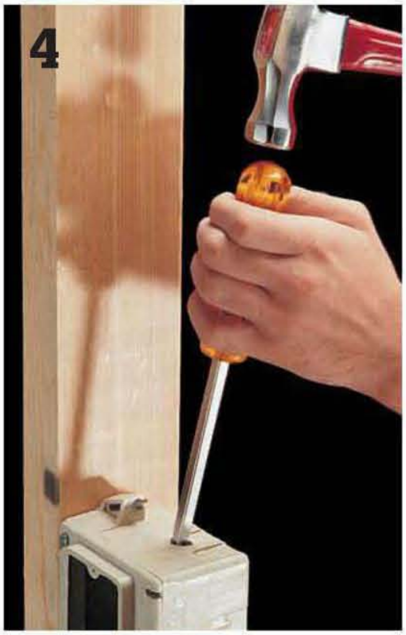

Open one knockout for each cable that will enter the box using a hammer and screwdriver.

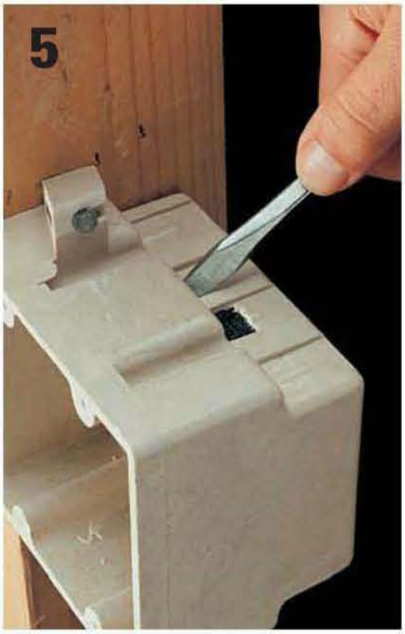

Break off any sharp edges that might damage vinyl cable sheathing by rotating a screwdriver in the knockout.

How to Install Electrical Boxes for Light Fixtures

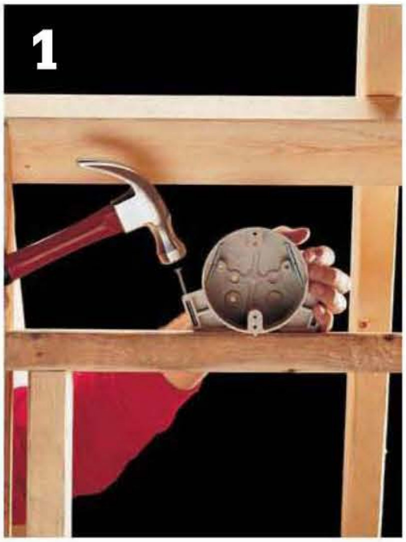

Position the light fixture box for a vanity light above the frame opening for a mirror or medicine cabinet. Place the box for a ceiling light fixture in the center of the room. Position each box against a framing member so the front face will be flush with the finished wall or ceiling, then anchor the box by driving the mounting nails into the framing.

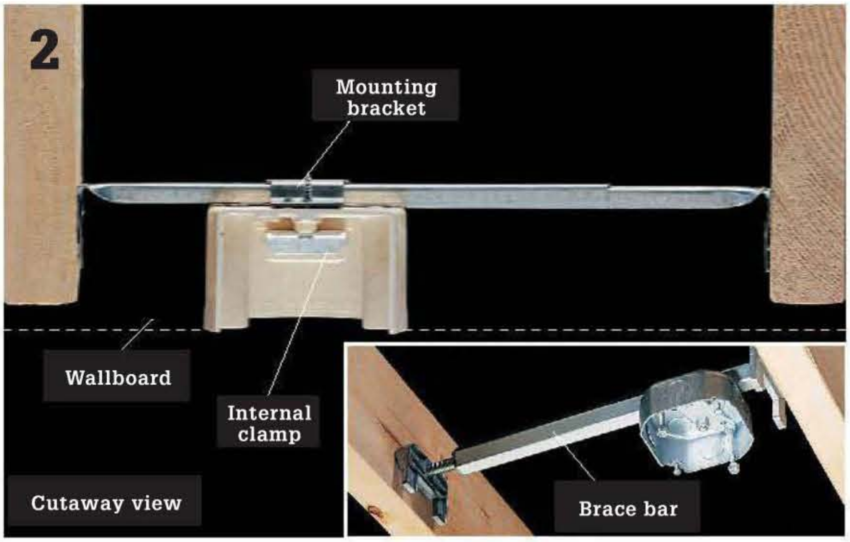

To position a light fixture between joists, attach an electrical box to an adjustable brace bar. Nail the ends of the brace bar to joists so the face of the box will be flush with the finished ceiling surface. Slide the box along the brace bar to the desired position, then tighten the mounting screws. Use internal cable clamps when using a box with a brace bar. *Note: For ceiling fans and heavy fixtures, use a metal box and a heavy-duty brace bar rated for heavy loads (inset photo).*

How to Install Electrical Boxes for Switches

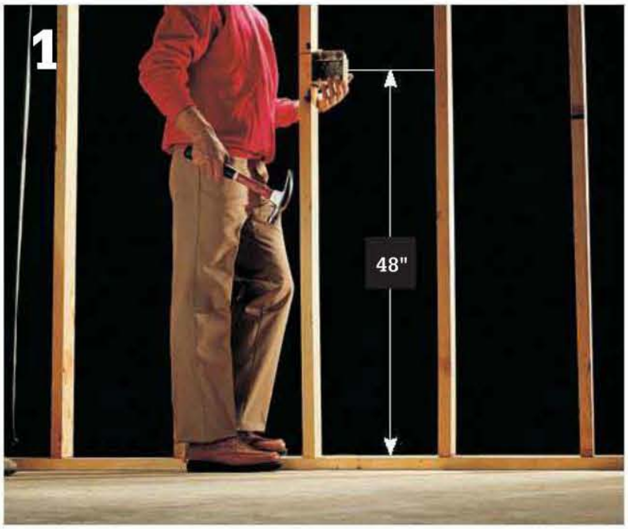

Install switch boxes at accessible locations, usually on the latch side of a door, with the center of the box 48" from the floor. The box for a thermostat is mounted at 48 to 60". Position each box against the side of a stud so the front face will be flush with the finished wall, and drive the mounting nails into the stud.

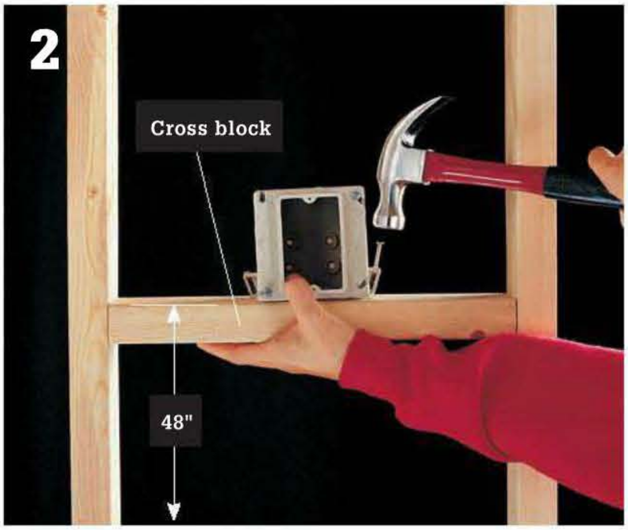

To install a switch box between studs, first install a cross block between studs, with the top edge 46" above the floor. Position the box on the cross block so the front face will be flush with the finished wall, and drive the mounting nails into the cross block.

Ceiling Boxes

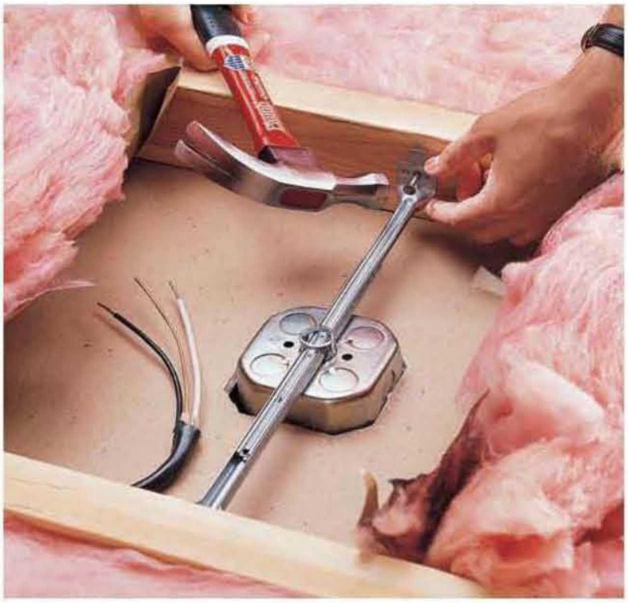

Ceiling boxes for lights are generally round or octagonal in shape to fit typical lamp mounting plates. The easiest way to install one is by nailing the brace to open ceiling joists from above. If the ceiling is insulated, pull the insulation away from the box if the fixture you're installing is not rated IC for insulation contact.

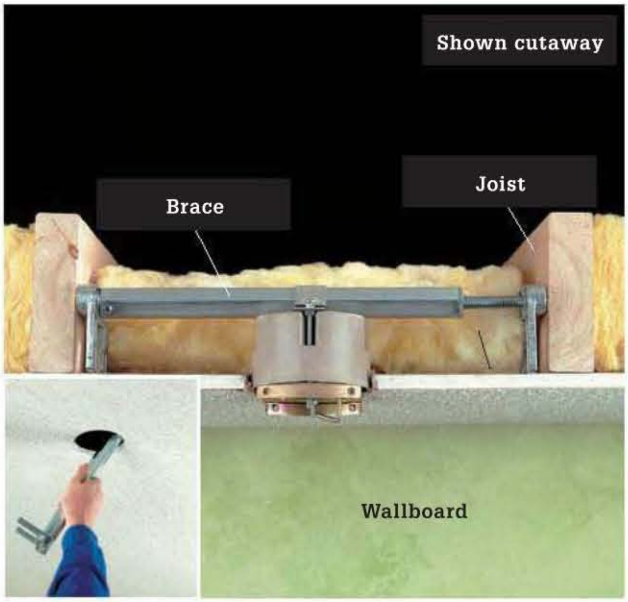

A heavy-duty brace is required for anchoring boxes that will support heavy chandeliers and ceiling fans. A remodeling brace like the one seen here is designed to install through a small cutout in the ceiling (inset photo).

How to Install a Junction Box

1

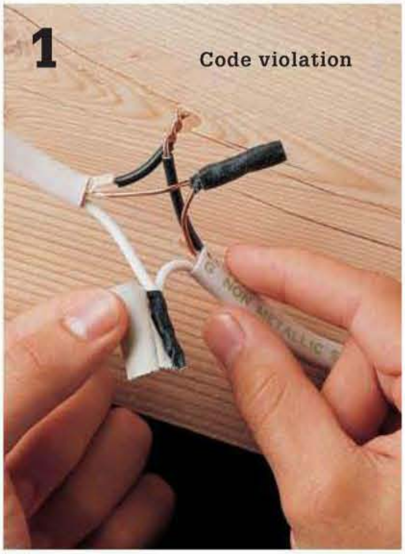

Turn off power to circuit wires at the main service panel. Carefully remove any tape or wire connectors from the exposed splice. Avoid contact with the bare wire ends until the wires have been tested for power.

2

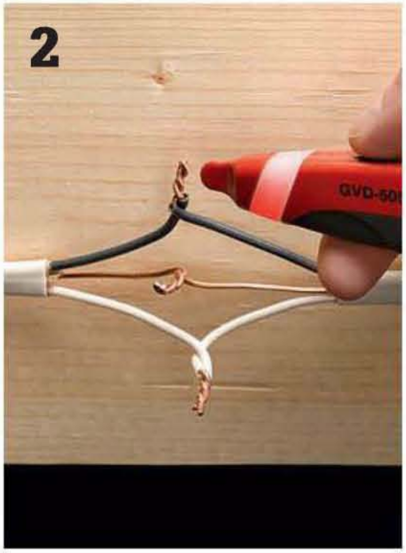

Test for power. The tester should not glow. If it does, the wires are still hot. Shut off power to correct circuit at the main service panel. Disconnect the illegally spliced wires.

3

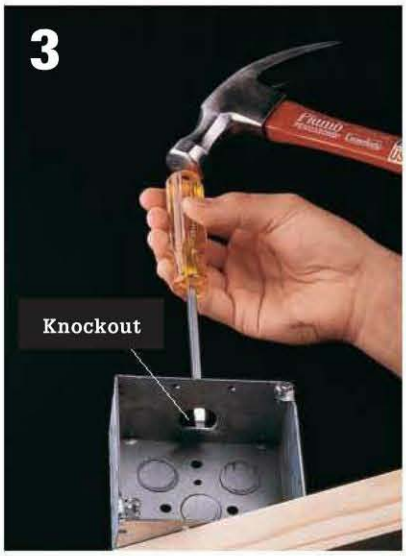

Open one knockout for each cable that will enter the box using a hammer and screwdriver. Any unopened knockouts should remain sealed.

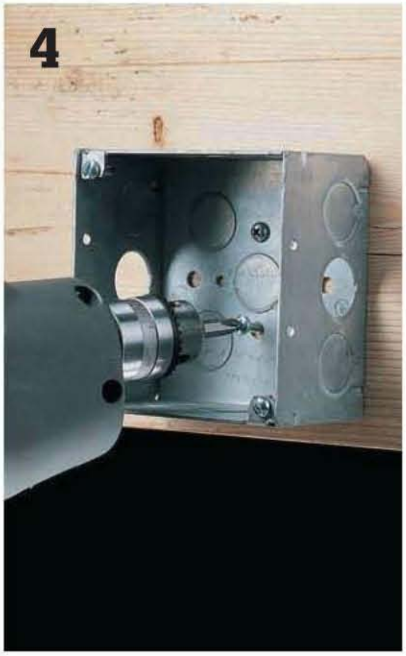

Anchor the electrical box to a wooden framing member using screws or nails.

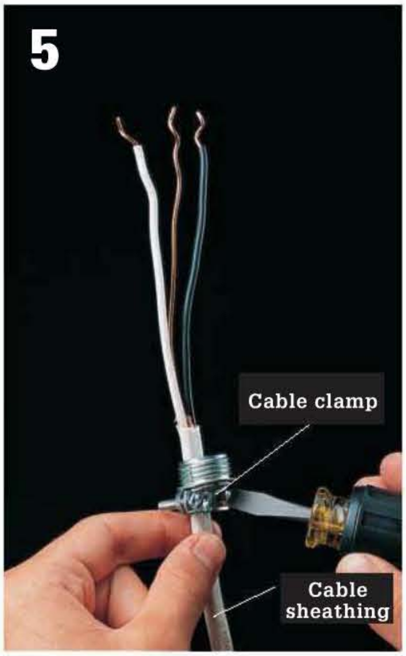
5
Cable clamp
Cable sheathing

Thread each cable end through a cable clamp. Tighten the clamp with a screwdriver. See if there is any slack in the cables so you can gain a little extra cable to work with.

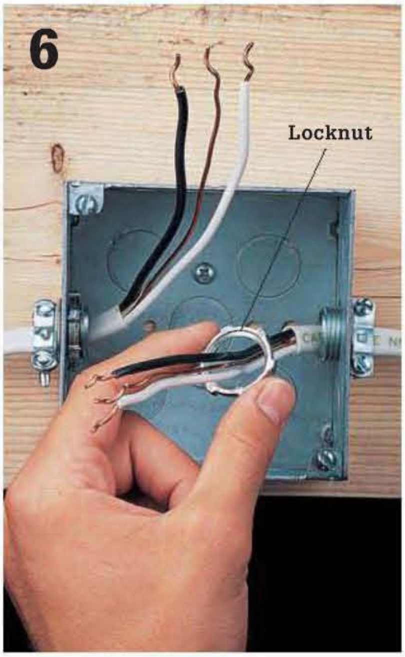

Insert the cables into the electrical box, and screw a locknut onto each cable clamp.

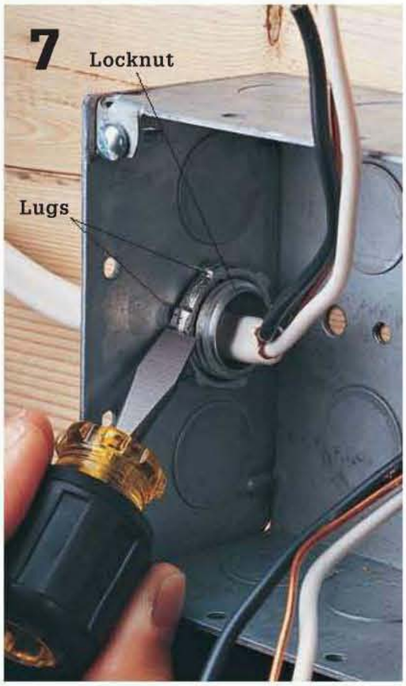
7
Locknut
Lugs

Tighten the locknuts by pushing against the lugs with the blade of a screwdriver.

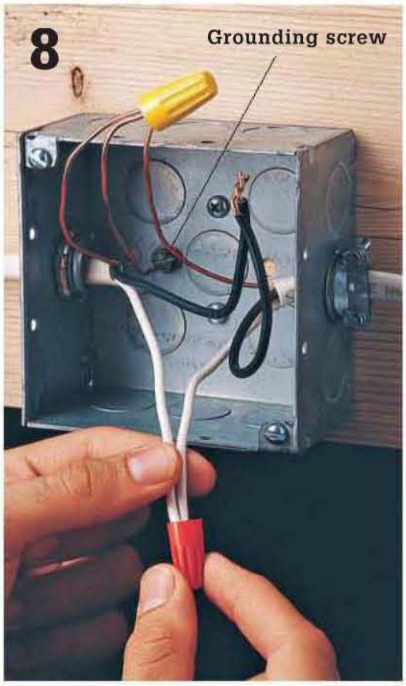
8
Grounding screw

Use wire connectors to reconnect the wires. Pigtail the copper grounding wires to the green grounding screw in the back of the box.

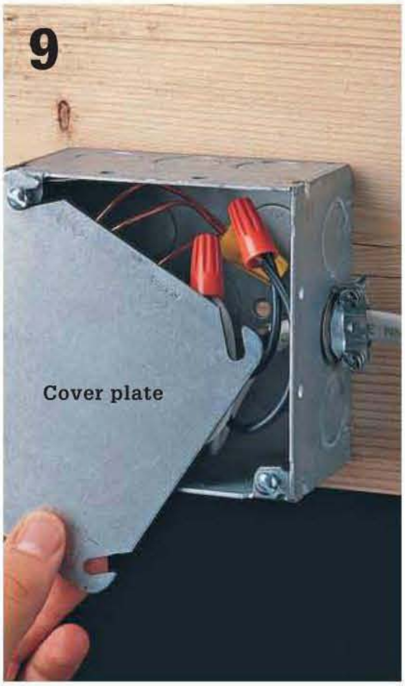
9
Cover plate

Carefully tuck the wires into the box, and attach the cover plate. Turn on the power to the circuit at the main service panel. Make sure the box remains accessible and is not concealed by finished walls or ceilings.

Working with Switches

Wall switches are available in three general types. To repair or replace a switch, it is important to identify its type.

Single-pole switches are used to control a set of lights from one location. Three-way switches are used to control a set of lights from two different locations and are always installed in pairs. Four-way switches are used in combination with a pair of three-way switches to control a set of lights from three or more locations.

Identify switch types by counting the screw terminals. Single-pole switches have two screw terminals, three-way switches have three screw terminals, and four-way switches have four. Most switches include a grounding screw terminal, which is identified by its green color.

When replacing a switch, choose a new switch that has the same number of screw terminals as the old one. The location of the screws on the switch body varies depending on the manufacturer, but these differences will not affect the switch operation.

Whenever possible, connect switches using the screw terminals rather than push-in fittings. Some specialty switches have wire leads instead of screw terminals. They are connected to circuit wires with wire connectors.

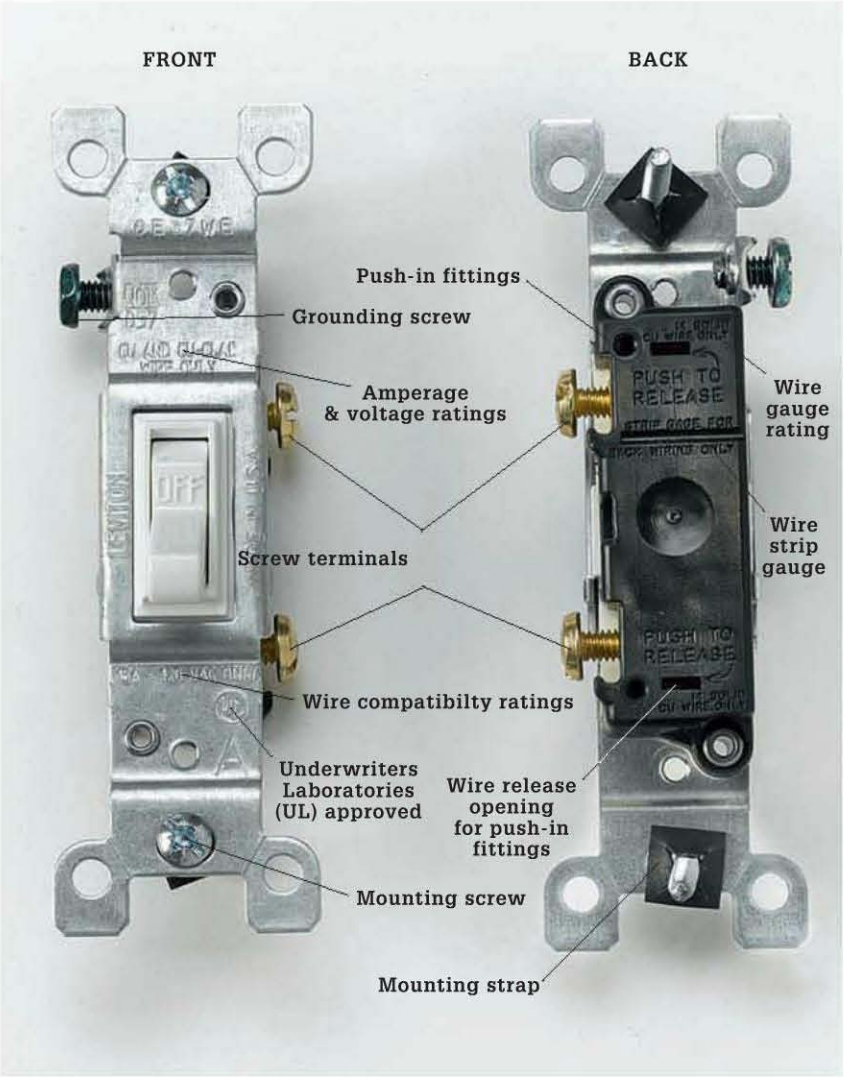
FRONT
BACK
Push-in fittings
Grounding screw
Amperage & voltage ratings
Wire gauge rating
Screw terminals
Wire strip gauge
Wire compatibilty ratings
Underwriters Laboratories (UL) approved
Wire release opening for push-in fittings
Mounting screw
Mounting strap
PUSH TO RELEASE
OFF

A wall switch is connected to circuit wires with screw terminals or with push-in fittings on the back of the switch. A switch may have a stamped strip gauge that indicates how much insulation must be stripped from the circuit wires to make the connections.

The switch body is attached to a metal mounting strap that allows it to be mounted in an electrical box. Several rating stamps are found on the strap and on the back of the switch. The abbreviation UL or UND. LAB. INC. LIST means that the switch meets the safety standards of the Underwriters Laboratories. Switches also are stamped with maximum voltage and amperage ratings. Standard wall switches are rated 15A or 125V. Voltage ratings of 110, 120, and 125 are considered to be identical for purposes of identification.

For standard wall switch installations, choose a switch that has a wire gauge rating of #12 or #14. For wire systems with solid-core copper wiring, use only switches marked COPPER or CU. For aluminum wiring, use only switches marked CO/ALR. Switches marked AL/CU can no longer be used with aluminum wiring, according to the National Electrical Code.

Single-pole Wall Switches

A single-pole switch is the most common type of wall switch. It has ON-OFF markings on the switch lever and is used to control a set of lights, an appliance, or a receptacle from a single location. A single-pole switch has two screw terminals and a grounding screw. When installing a single-pole switch, check to make sure the ON marking shows when the switch lever is in the up position.

In a correctly wired single-pole switch, a hot circuit wire is attached to each screw terminal. However, the color and number of wires inside the switch box will vary, depending on the location of the switch along the electrical circuit.

If two cables enter the box, then the switch lies in the middle of the circuit. In this installation, both of the hot wires attached to the switch are black.

If only one cable enters the box, then the switch lies at the end of the circuit. In this installation (sometimes called a switch loop), one of the hot wires is black, but the other hot wire usually is white. A white hot wire sometimes is coded with black tape or paint.

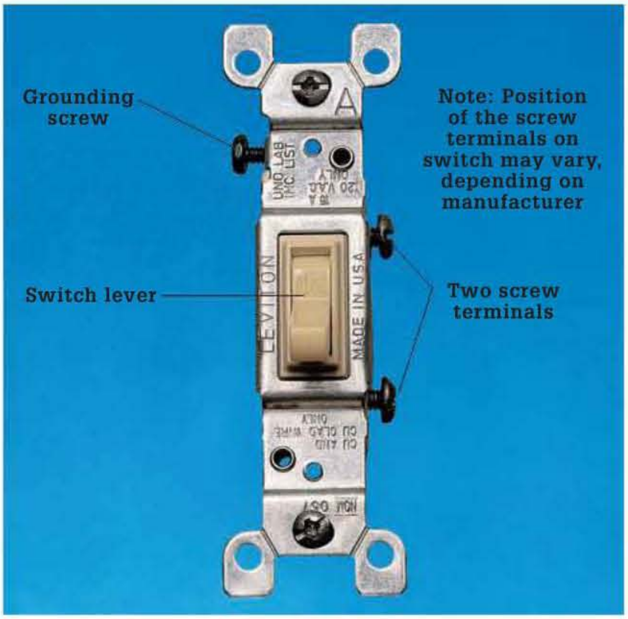
Grounding screw
Note: Position of the screw terminals on switch may vary, depending on manufacturer
Switch lever
Two screw terminals

A single-pole switch is essentially an interruption in the black power supply wire that is opened or closed with the toggle. Single-pole switches are the simplest of all home wiring switches.

Typical Single-pole Switch Installations

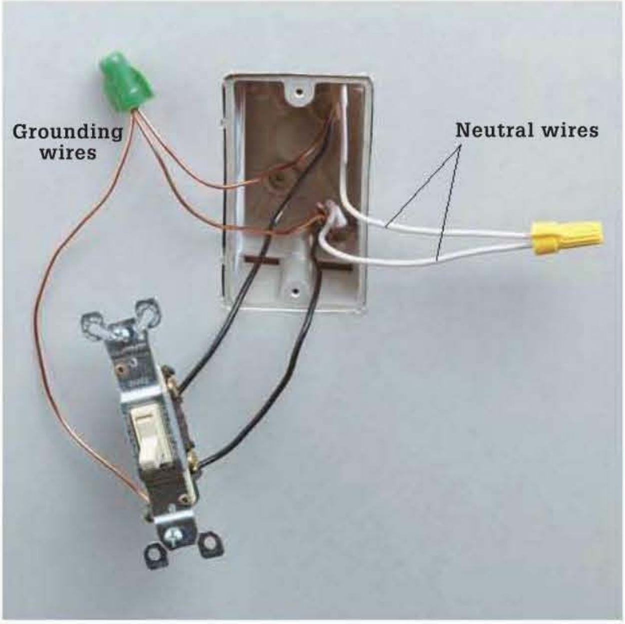

Two cables enter the box when a switch is located in the middle of a circuit. Each cable has a white and a black insulated wire, plus a bare copper grounding wire. The black wires are hot and are connected to the screw terminals on the switch. The white wires are neutral and are joined together with a wire connector. Grounding wires are pigtailed to the switch.

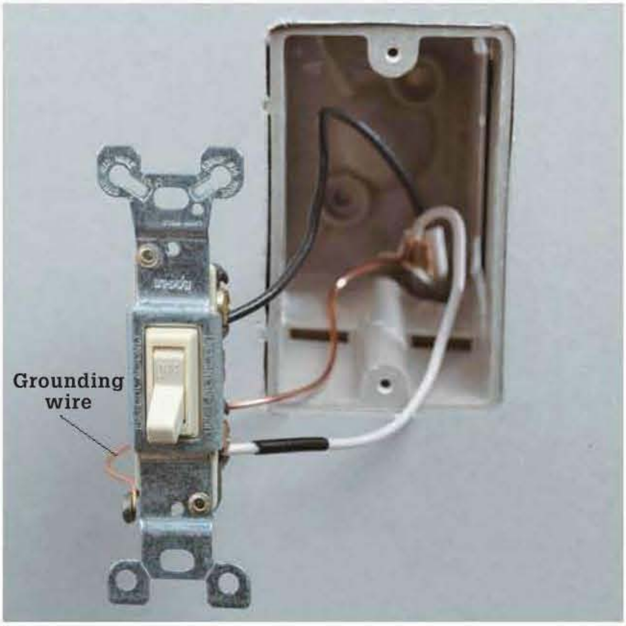

One cable enters the box when a switch is located at the end of a circuit. The cable has a white and a black insulated wire, plus a bare copper grounding wire. In this installation, both of the insulated wires are hot. The white wire may be labeled with black tape or paint to identify it as a hot wire. The grounding wire is connected to the switch grounding screw.

Three-way Wall Switches

Three-way switches have three screw terminals and do not have ON-OFF markings. Three-way switches are always installed in pairs and are used to control a set of lights from two locations.

One of the screw terminals on a three-way switch is darker than the others. This screw is the common screw terminal. The position of the common screw terminal on the switch body may vary, depending on the manufacturer. Before disconnecting a three-way switch, always label the wire that is connected to the common screw terminal. It must be reconnected to the common screw terminal on the new switch.

The two lighter-colored screw terminals on a three-way switch are called the traveler screw terminals. The traveler terminals are interchangeable, so there is no need to label the wires attached to them.

Because three-way switches are installed in pairs, it sometimes is difficult to determine which of the switches is causing a problem. The switch that receives greater use is more likely to fail, but you may need to inspect both switches to find the source of the problem.

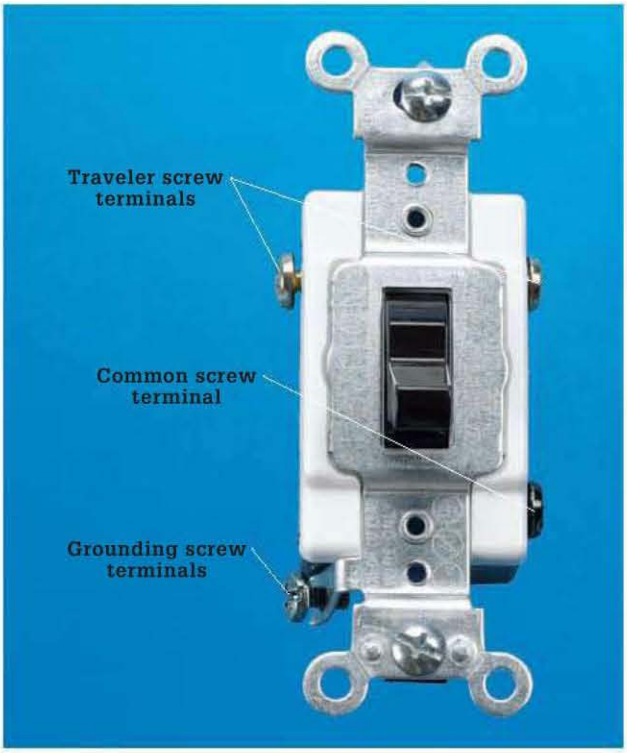
Traveler screw terminals
Common screw terminal
Grounding screw terminals

Typical Three-way Switch Installations

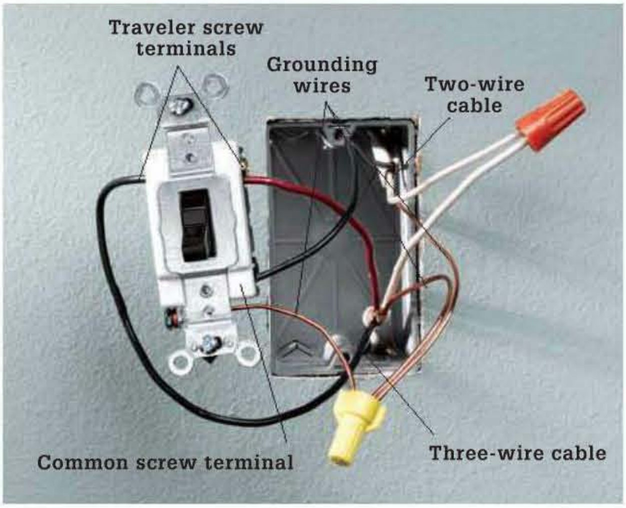
Traveler screw terminals
Grounding wires
Two-wire cable
Common screw terminal
Three-wire cable

Two cables enter the box if the switch lies in the middle of a circuit. One cable has two wires, plus a bare copper grounding wire; the other cable has three wires, plus a ground. The black wire from the two-wire cable is connected to the dark common screw terminal. The red and black wires from the three-wire cable are connected to the traveler screw terminals. The white neutral wires are joined together with a wire connector, and the grounding wires are pigtailed to the grounded metal box.

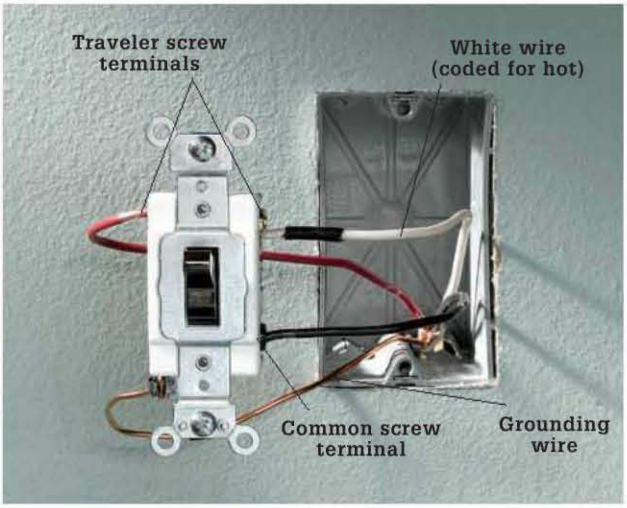
Traveler screw terminals
White wire (coded for hot)
Common screw terminal
Grounding wire

One cable enters the box if the switch lies at the end of the circuit. The cable has a black wire, red wire, and white wire, plus a bare copper grounding wire. The black wire must be connected to the common screw terminal, which is darker than the other two screw terminals. The white and red wires are connected to the two traveler screw terminals. The white wire is taped to indicate that it is hot. The bare copper grounding wire is connected to the grounded metal box.

Four-way Wall Switches

Four-way switches have four screw terminals and do not have ON-OFF markings. Four-way switches are always installed between a pair of three-way switches. This switch combination makes it possible to control a set of lights from three or more locations. Four-way switches are common in homes where large rooms contain multiple living areas, such as a kitchen opening into a dining room. Switch problems in a four-way installation can be caused by loose connections or worn parts in a four-way switch or in one of the three-way switches (facing page).

In a typical installation, there will be a pair of three-way cables that enter the box for the four-way switch. With most switches, the white and red wires from one cable should be attached to the bottom or top pair of screw terminals, and the white and red wires from the other cable should be attached to the remaining pair of screw terminals. However, not all switches are configured the same way, and wiring configurations in the box may vary, so always study the wiring diagram that comes with the switch.

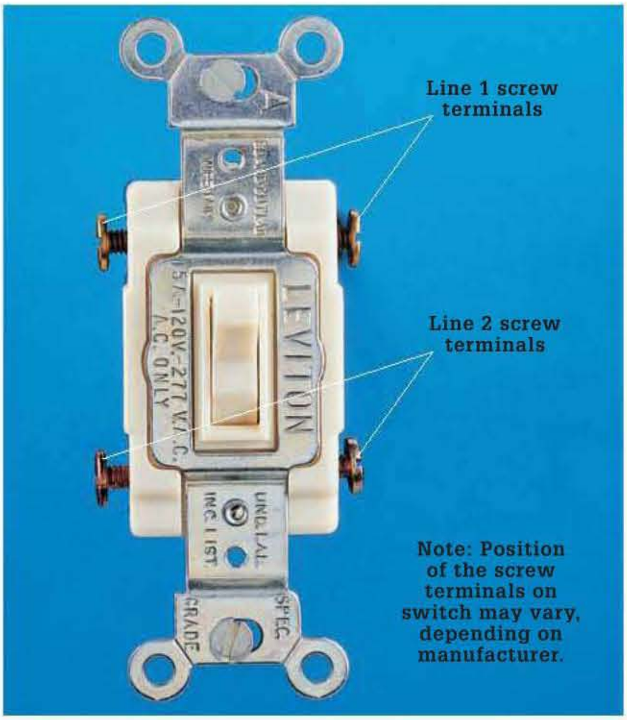

Note: Position of the screw terminals on switch may vary, depending on manufacturer.

Typical Four-way Switch Installation

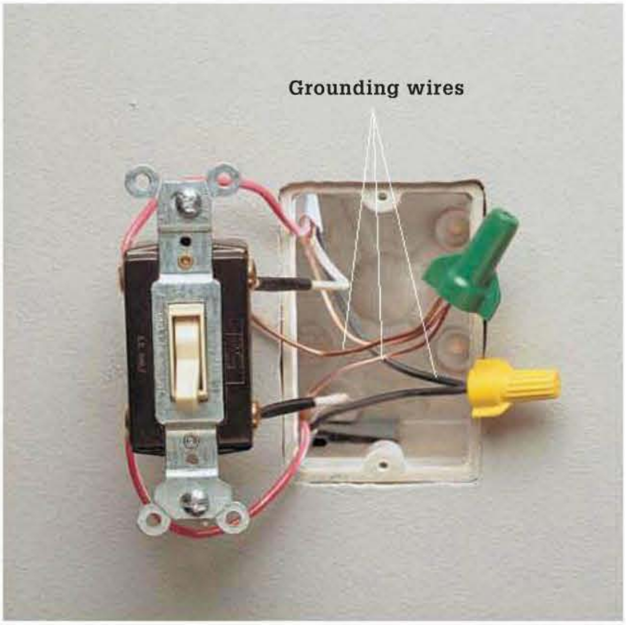

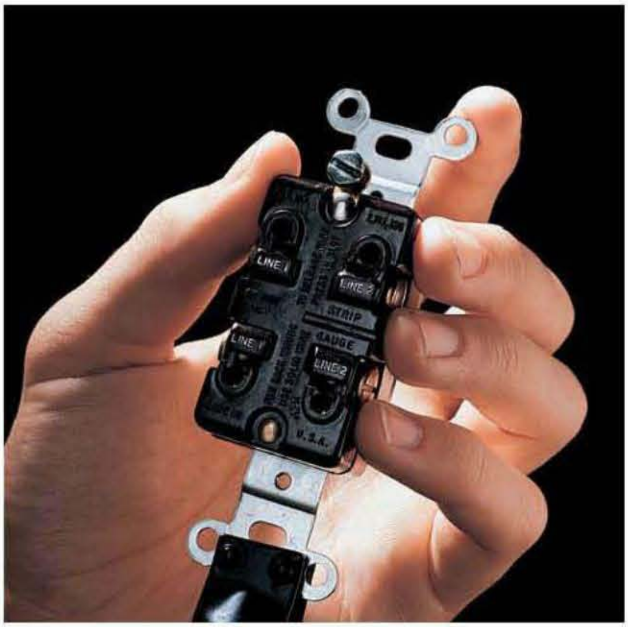

Four wires are connected to a four-way switch. The red and white wires from one cable are attached to the top pair of screw terminals, while the red and white wires from the other cable are attached to the bottom screw terminals.

Switch variation: Some four-way switches have a wiring guide stamped on the back to help simplify installation. For the switch shown above, one pair of color-matched circuit wires will be connected to the screw terminals marked LINE 1, while the other pair of wires will be attached to the screw terminals marked LINE 2.

Testing Switches

A switch that does not work properly may have worn or broken internal parts. Test switches with a battery-operated continuity tester. The continuity tester detects any break in the metal pathway inside the switch. Replace the switch if the continuity tester shows the switch to be faulty.

Never use a continuity tester on wires that might carry live current. Always shut off the power and disconnect the switch before testing for continuity.

Some specialty switches, like dimmers, cannot be tested for continuity. Electronic switches can be tested for manual operation using a continuity tester, but the automatic operation of these switches cannot be tested.

How to Test a Single-pole Wall Switch

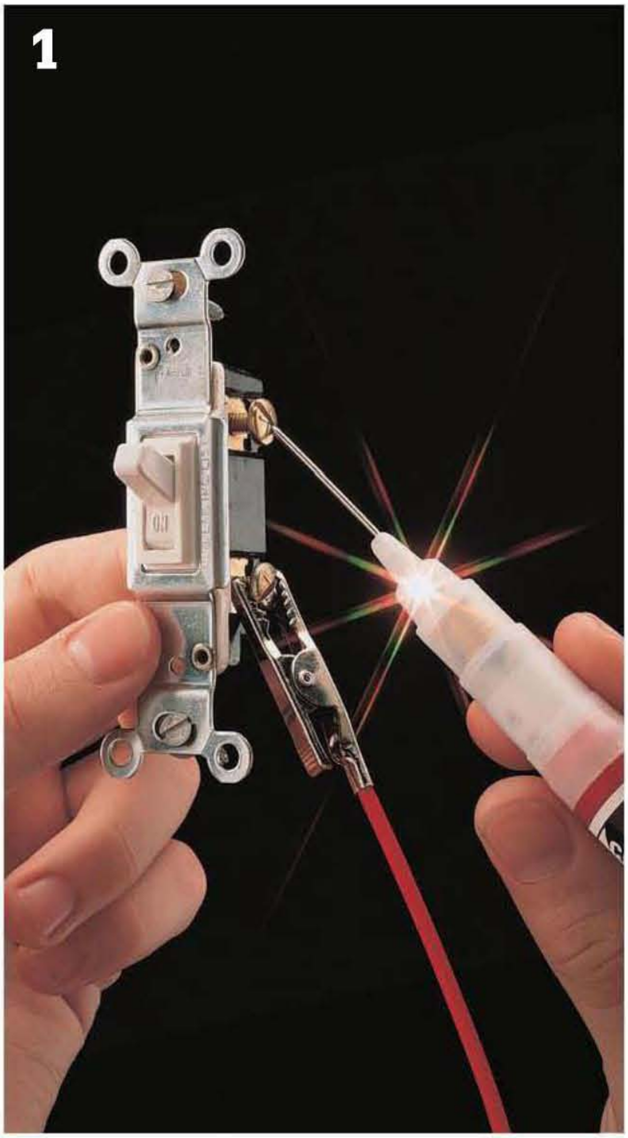

Attach clip of tester to one of the screw terminals. Touch the tester probe to the other screw terminal. Flip switch lever from ON to OFF. If switch is good, tester glows when lever is ON, but not when OFF.

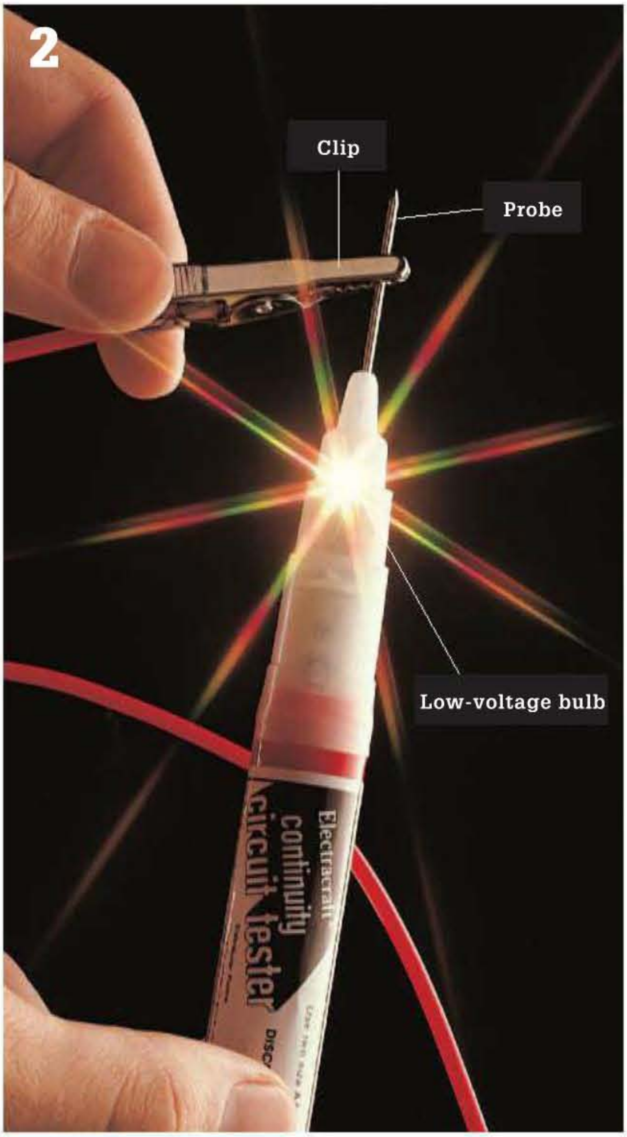

Continuity tester uses battery-generated current to test the metal pathways running through switches and other electrical fixtures. Always "test" the tester before use. Touch the tester clip to the metal probe. The tester should glow. If not, then the battery or lightbulb is dead and must be replaced.

How to Test a Three-way Wall Switch

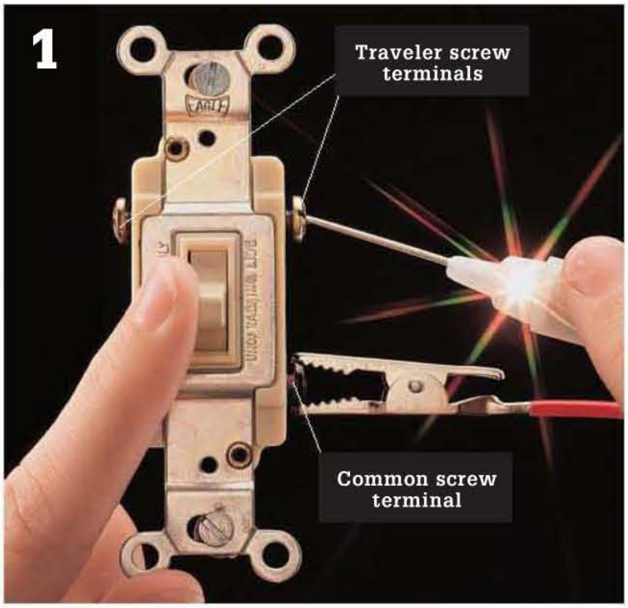

Attach tester clip to the dark common screw terminal. Touch the tester probe to one of the traveler screw terminals, and flip switch lever back and forth. If switch is good, the tester should glow when the lever is in one position, but not both.

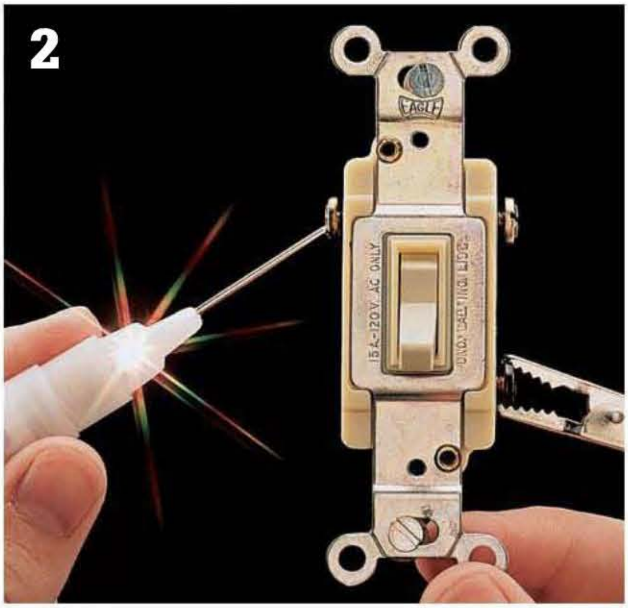

Touch probe to the other traveler screw terminal, and flip the switch lever back and forth. If switch is good, the tester will glow only when the switch lever is in the position opposite from the positive test in step 1.

How to Test a Four-way Wall Switch

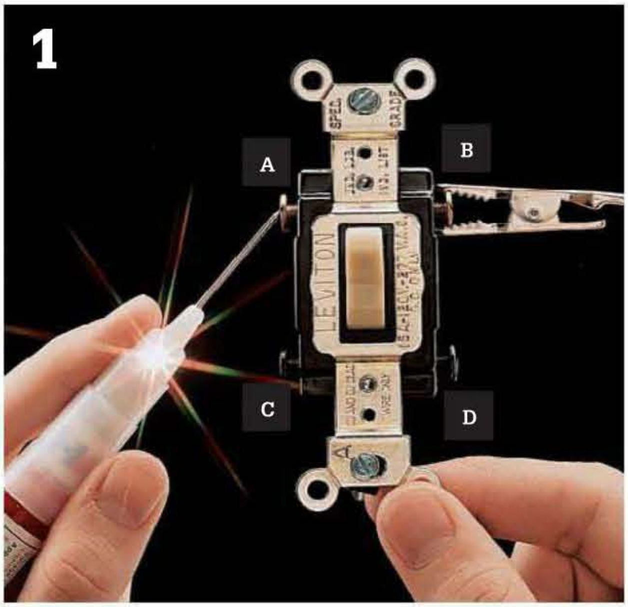

Test switch by touching probe and clip of continuity tester to each pair of screw terminals (A-B, C-D, A-D, B-C, A-C, B-D). The test should show continuous pathways between two different pairs of screw terminals. Flip lever to opposite position, and repeat test. Test should show continuous pathways between two different pairs of screw terminals.

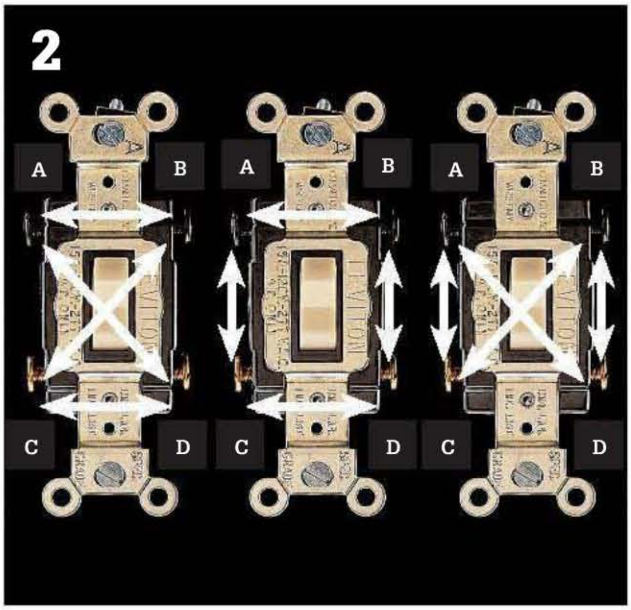

If switch is good, test will show a total of four continuous pathways between screw terminals—two pathways for each lever position. If not, then switch is faulty and must be replaced. (The arrangement of the pathways may differ, depending on the switch manufacturer. The photo above shows the three possible pathway arrangements.)

Working with Receptacles

A 120-volt duplex receptacle can be wired to the electrical system in a number of ways. The most common are shown on these pages.

Wiring configurations may vary slightly from these photographs, depending on the kind of receptacles used, the type of cable, or the technique of the electrician who installed the wiring. To make dependable repairs or replacements, use masking tape and label each wire according to its location on the terminals of the existing receptacle.

Receptacles are wired as either end-of-run or middle-of-run. These two basic configurations are easily identified by counting the number of cables entering the receptacle box. End-of-run wiring has only one cable, indicating that the circuit ends. Middle-of-run wiring has two cables, indicating that the circuit continues on to other receptacles, switches, or fixtures.

A split-circuit receptacle is shown on the next page. Each half of a split-circuit receptacle is wired to a separate circuit. This allows two appliances of high wattage to be plugged into the same receptacle without blowing a fuse or tripping a breaker. This wiring configuration is similar to a receptacle that is controlled by a wall switch. Code requires a switch-controlled receptacle in any room that does not have a built-in light fixture operated by a wall switch.

Split-circuit and switch-controlled receptacles are connected to two hot wires, so use caution during repairs or replacements. Make sure the connecting tab between the hot screw terminals is removed.

Two-slot receptacles are common in older homes. There is no grounding wire attached to the receptacle, but the box may be grounded with armored cable or conduit.

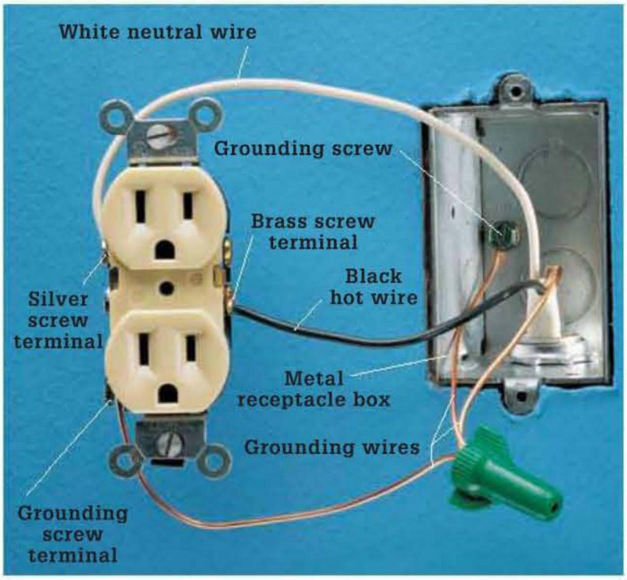
White neutral wire
Grounding screw
Brass screw terminal
Black hot wire
Silver screw terminal
Metal receptacle box
Grounding wires
Grounding screw terminal

Single cable entering the box indicates end-of-run wiring. The black hot wire is attached to a brass screw terminal, and the white neutral wire is connected to a silver screw terminal. If the box is metal, the grounding wire is pigtailed to the grounding screws of the receptacle and the box. In a plastic box, the grounding wire is attached directly to the grounding screw terminal of the receptacle.

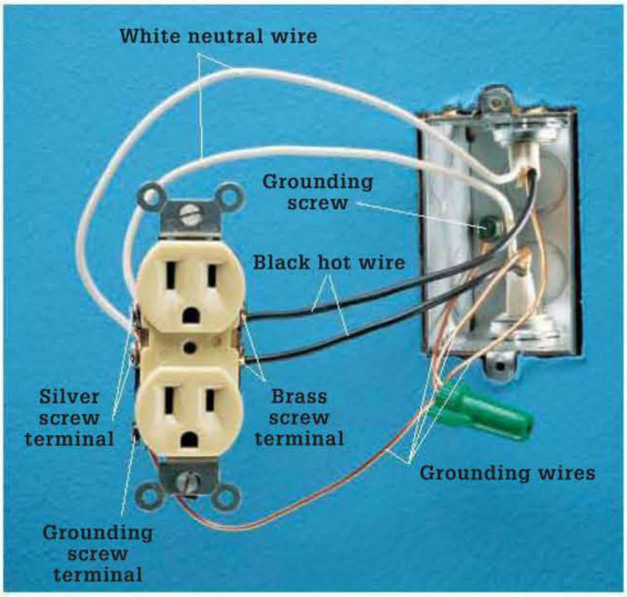
White neutral wire
Grounding screw
Black hot wire
Silver screw terminal
Brass screw terminal
Grounding wires
Grounding screw terminal

Two cables entering the box indicate middle-of-run wiring. Black hot wires are connected to brass screw terminals, and white neutral wires to silver screw terminals. The grounding wire is pigtailed to the grounding screws of the receptacle and the box.

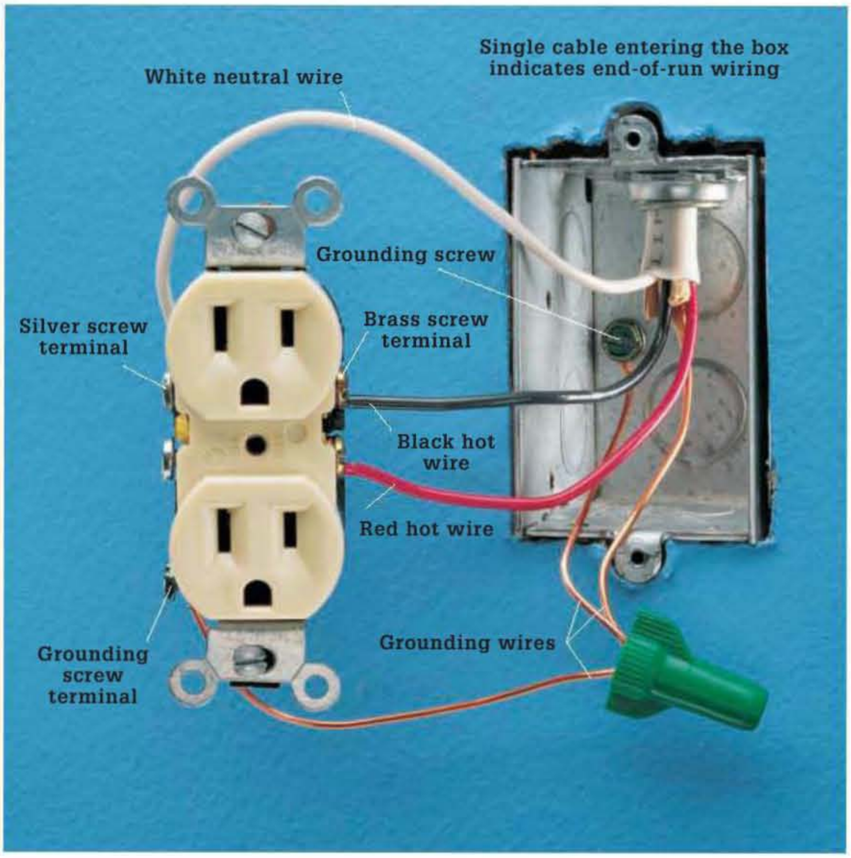
White neutral wire
Single cable entering the box indicates end-of-run wiring
Grounding screw
Silver screw terminal
Brass screw terminal
Black hot wire
Red hot wire
Grounding wires
Grounding screw terminal

A split-circuit receptacle is attached to a black hot wire, a red hot wire, a white neutral wire, and a bare grounding wire. The wiring is similar to a switch-controlled receptacle.
The hot wires are attached to the brass screw terminals, and the connecting tab or fin between the brass terminals is removed. The white wire is attached to a silver screw terminal, and the connecting tab on the neutral side remains intact. The grounding wire is pigtailed to the grounding screw terminal of the receptacle and to the grounding screw attached to the box.

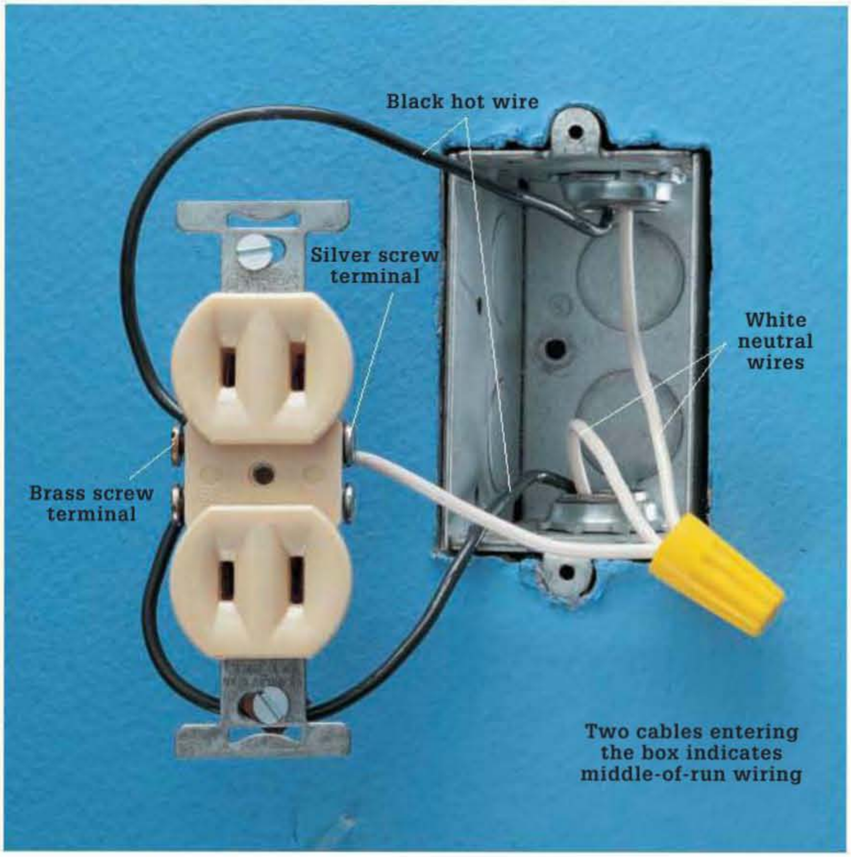
Black hot wire
Silver screw terminal
White neutral wires
Brass screw terminal
Two cables entering the box indicates middle-of-run wiring

A two-slot receptacle is often found in older homes. The black hot wires are connected to the brass screw terminals, and the white neutral wires are pigtailed to a silver screw terminal. Two-slot receptacles may be replaced with three-slot types, but only if a means of grounding exists at the receptacle box. In some municipalities, you may replace a two-slot receptacle with a GFCI receptacle as long as the receptacle has a sticker that reads "No equipment ground."

How to Install a New Receptacle

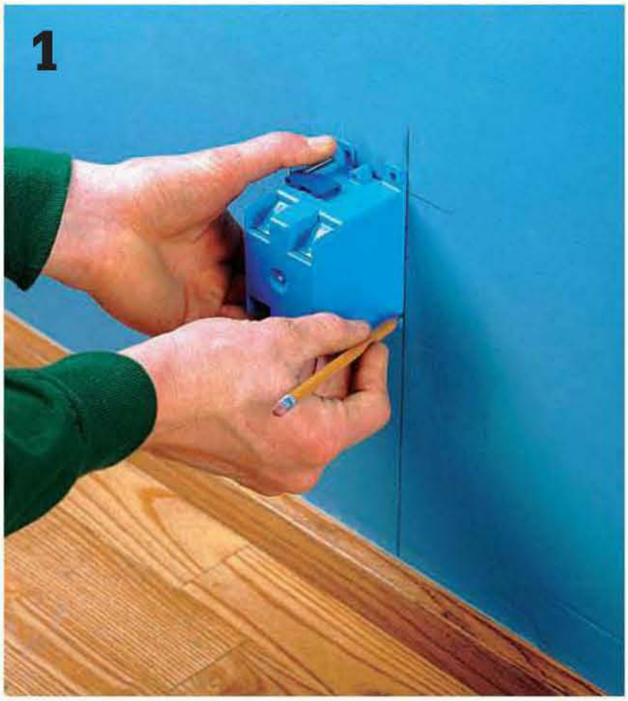

Position the new pop-in box on the wall and trace around it. Consider the location of hidden utilities within the wall before you cut.

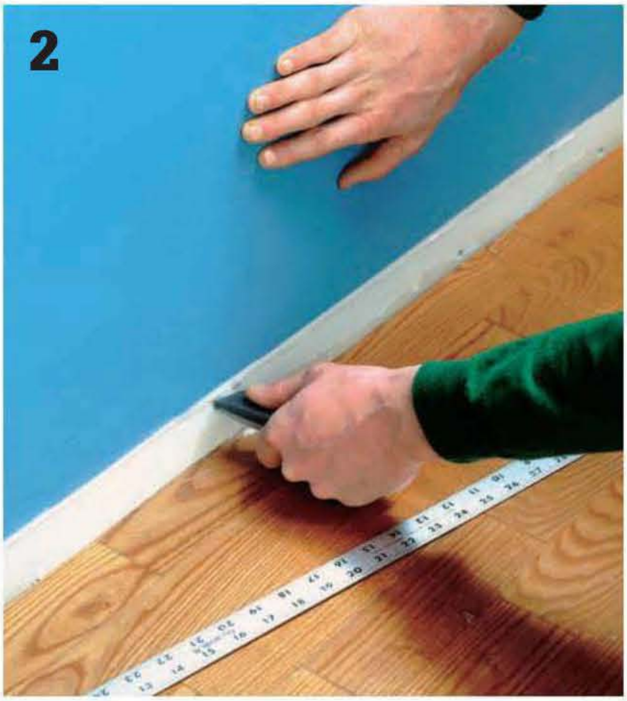

Remove baseboard between new and existing receptacle. Cut away the drywall about 1" below the baseboard with a jigsaw, wallboard saw, or utility knife.

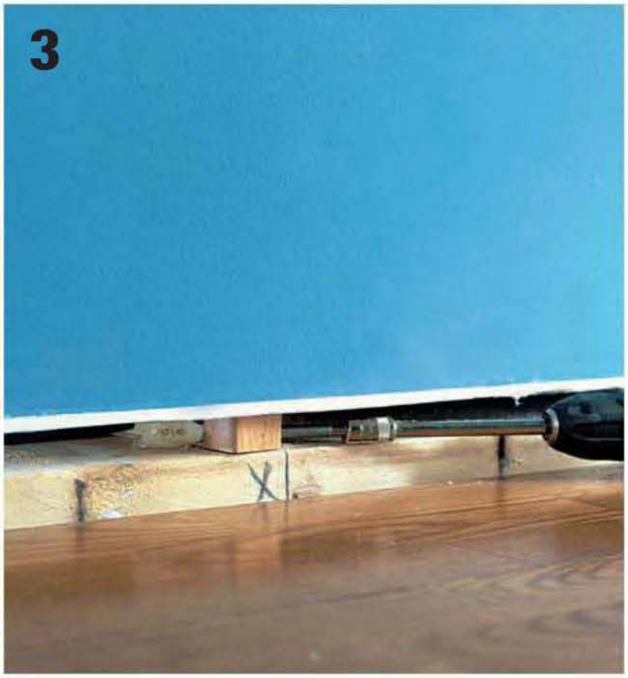

Drill a ⅝" hole in the center of each stud along the opening between the two receptacles. A drill bit extender or a flexible drill bit will allow you a better angle and make drilling the holes easier.

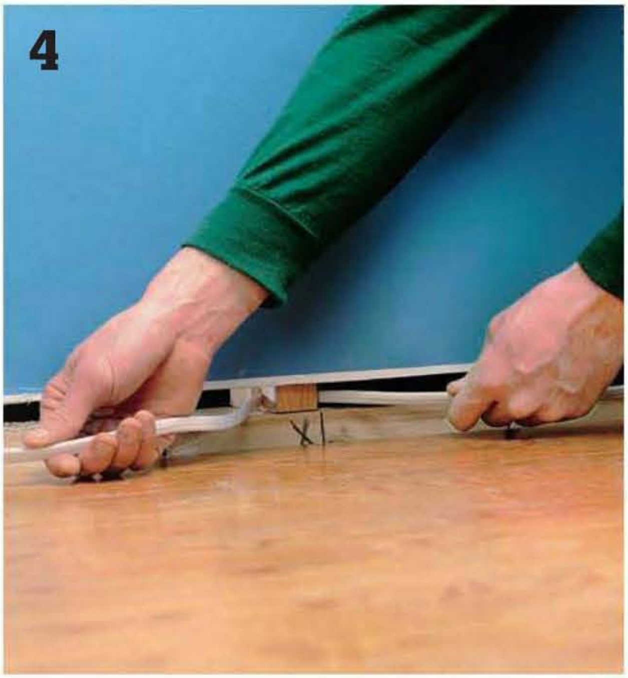

Run the branch cable through the holes from the new location to the existing receptacle. Staple the cable to the stud below the box. Install a metal nail plate on the front edge of each stud that the cable routes through.

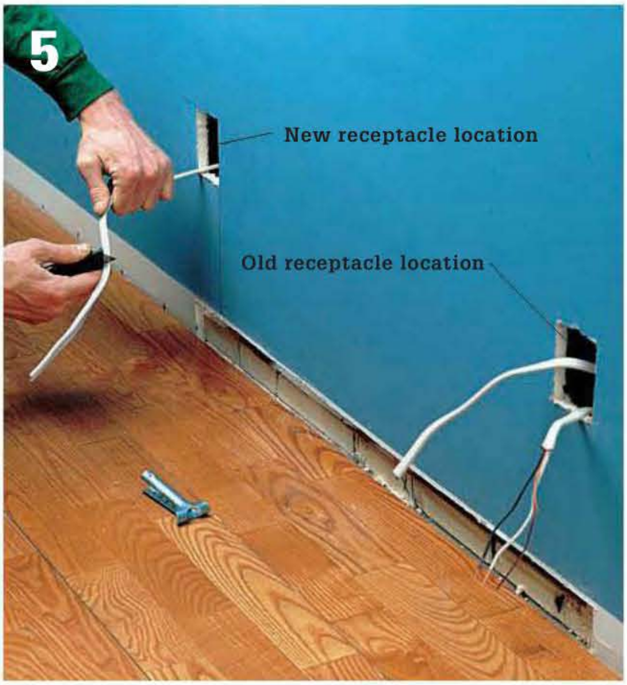

Turn off the power at the main panel and test for power. Remove the old receptacle and its box, and pull the new branch cable up through the hole. Remove sheathing and insulation from both ends of the new cable.

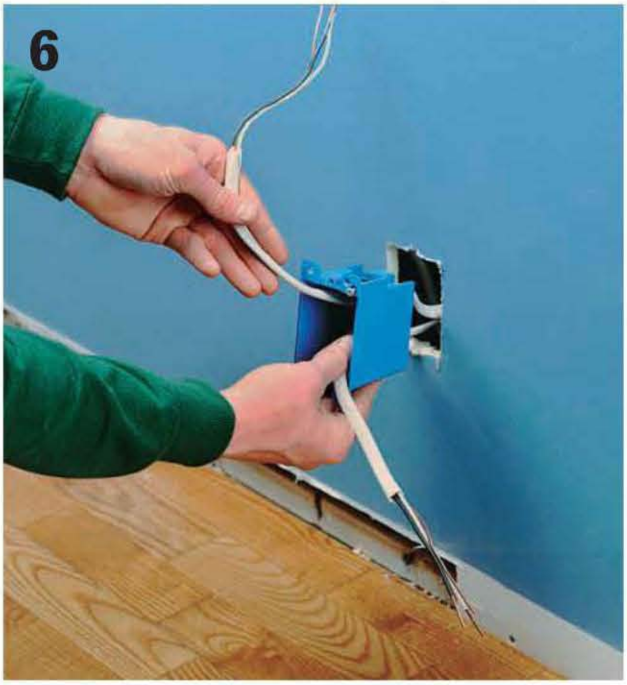

Thread the new and old cables into a pop-in box large enough to contain the added wires and clamp the cables. Fit the box into the old hole and attach it.

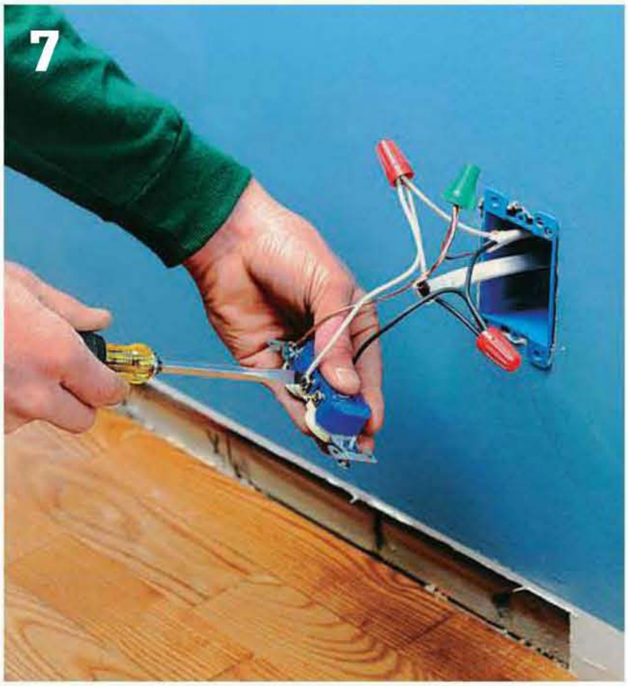

Reconnect the old receptacle by connecting its neutral, hot, and grounding screws to the new branch cable and the old cable from the panel with pigtails.

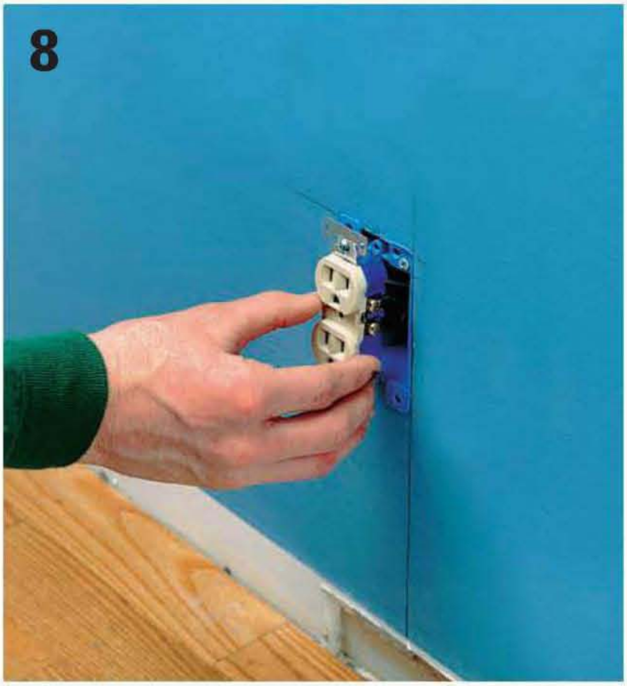

Pull the cable through another pop-in box for the new receptacle. Secure the cable and install the box. Connect the new receptacle to the new branch cable. Insert the receptacle into the box and attach the receptacle and cover plate with screws. Patch the opening with ½"-thick wood strips or drywall. Reattach the baseboard to the studs.

GFCI Receptacles

The ground-fault circuit-interrupter (GFCI) receptacle protects against electrical shock caused by a faulty appliance, or a worn cord or plug. It senses small changes in current flow and can shut off power in as little as 1/40 of a second.

GFCIs are now required in bathrooms, kitchens, garages, crawl spaces, unfinished basements, and outdoor receptacle locations. Consult your local codes for any requirements regarding the installation of GFCI receptacles. Most GFCIs use standard screw terminal connections, but some have wire leads and are attached with wire connectors. Because the body of a GFCI receptacle is larger than a standard receptacle, small crowded electrical boxes may need to be replaced with more spacious boxes.

The GFCI receptacle may be wired to protect only itself (single location), or it can be wired to protect all receptacles, switches, and light fixtures from the GFCI "forward" to the end of the circuit (multiple locations).

Because the GFCI is so sensitive, it is most effective when wired to protect a single location. The more receptacles any one GFCI protects, the more susceptible it is to "phantom tripping," shutting off power because of tiny, normal fluctuations in current flow.

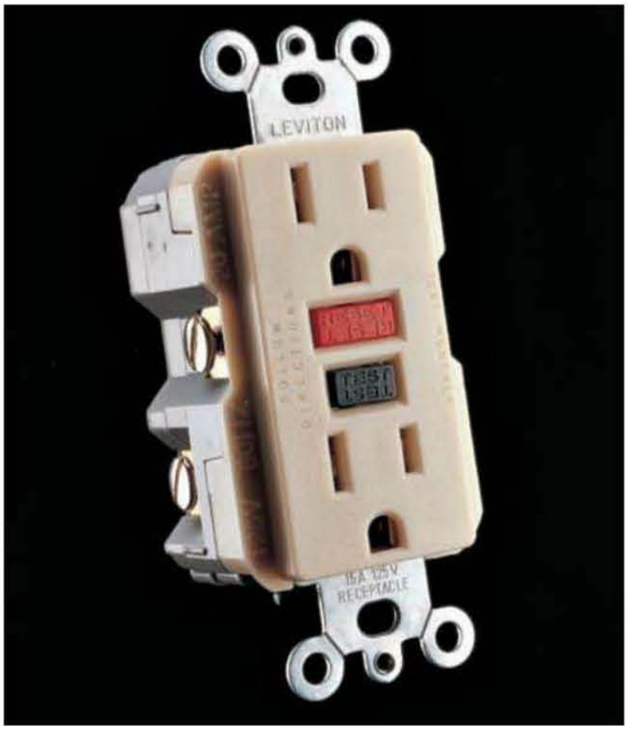
LEVITON
RESET
TEST
15A 125V RECEPTACLE

Tools & Materials ▸

- Circuit tester
- Screwdriver
- Wire connectors
- Masking tape

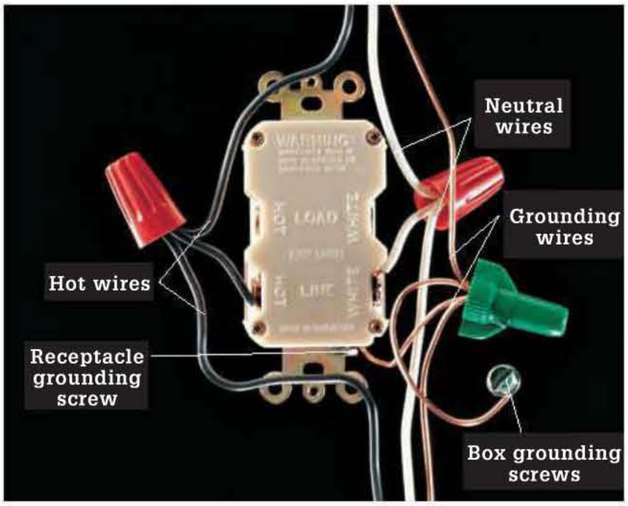
Neutral wires
Grounding wires
Hot wires
Receptacle grounding screw
Box grounding screws
WARNING
LOAD
LINE
HOT
WHITE

A GFCI wired for single-location protection (shown from the back) has hot and neutral wires connected only to the screw terminals marked LINE. A GFCI connected for single-location protection may be wired as either an end-of-run or middle-of-run configuration (page 110).

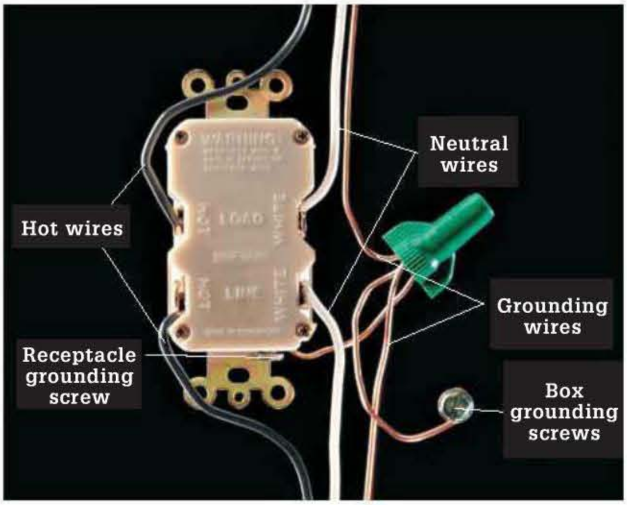
Neutral wires
Hot wires
Grounding wires
Receptacle grounding screw
Box grounding screws
WARNING
LOAD
LINE
HOT
WHITE

A GFCI wired for multiple-location protection (shown from the back) has one set of hot and neutral wires connected to the LINE pair of screw terminals, and the other set connected to the LOAD pair of screw terminals. A GFCI receptacle connected for multiple-location protection may be wired only as a middle-of-run configuration.

How to Install a Single-location GFCI

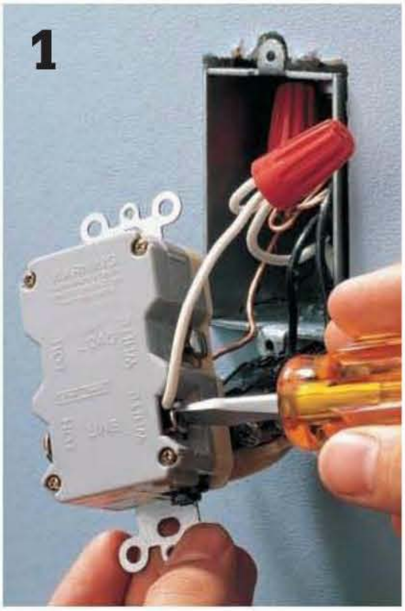

Pigtail all the white neutral wires together, and connect the pigtail to the terminal marked WHITE LINE on the GFCI (see photo on opposite page).

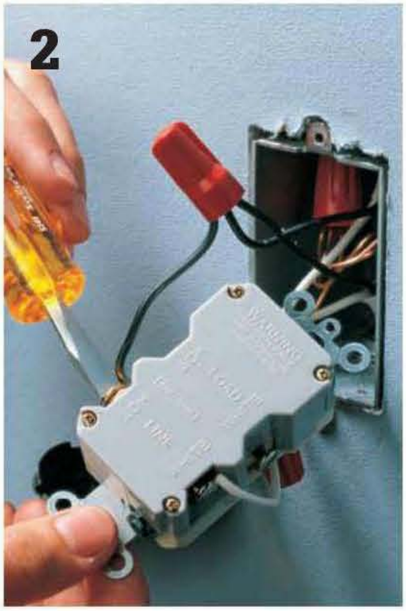

Disconnect all black hot wires from the brass screw terminals of the old receptacle. Pigtail these wires together, and connect them to the terminal marked HOT LINE on the GFCI.

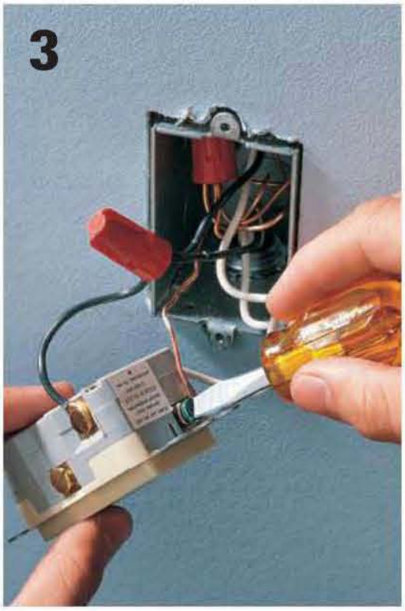

If a grounding wire is available, connect it to the green grounding screw terminal of the GFCI. Mount the GFCI in the receptacle box, and reattach the cover plate. Restore power, and test the GFCI according to the manufacturer's instructions.

How to Install Multiple-location GFCIs

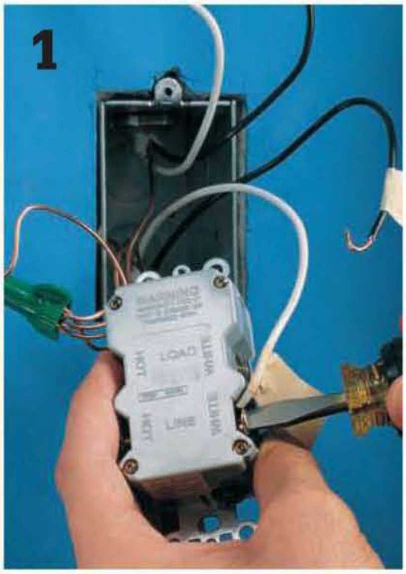

Connect the white feed wire to the terminal marked WHITE LINE on the GFCI. Connect the black feed wire to the terminal marked HOT LINE on the GFCI.

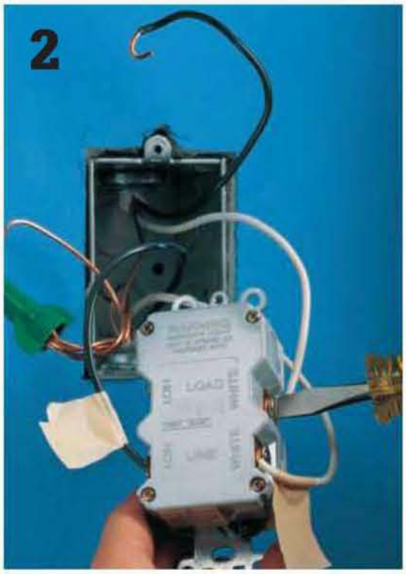

Connect the other white neutral wire to the terminal marked WHITE LOAD on the GFCI.

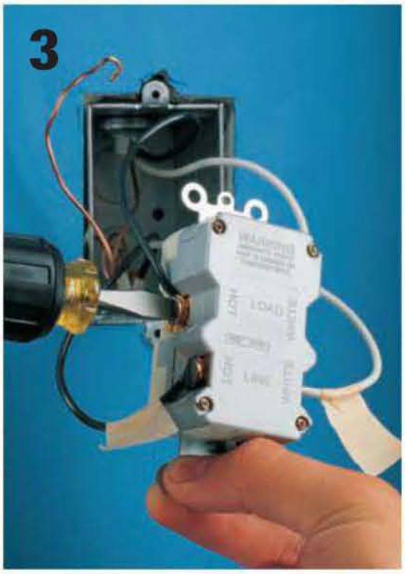

Connect the other black hot wire to the terminal marked HOT LOAD on the GFCI.

Planning New Wiring

Careful planning of a wiring project ensures you will have plenty of power for present and future needs. Whether you are adding circuits in a room addition, wiring a remodeled kitchen, or adding an outdoor circuit, consider all possible ways the space might be used, and plan for enough electrical service to meet peak needs.

For example, when wiring a room addition, remember that the way a room is used can change. In a room used as a spare bedroom, a single 15-amp circuit provides plenty of power, but if you ever choose to convert the same room to a family recreation space, you will need additional circuits.

When wiring a remodeled kitchen, it is a good idea to install circuits for an electric oven and countertop range, even if you do not have these electric appliances. Installing these circuits now makes it easy to convert from gas to electric appliances at a later date.

A large wiring project adds a considerable load to your main electrical service. In about 25 percent of all homes, some type of service upgrade is needed before new wiring can be installed. For example, many homeowners will need to replace an older 60-amp electrical service with a new service rated for 100 amps or more. This is a job for a licensed electrician but is well worth the investment. In other cases, the existing main service provides adequate power, but the main circuit breaker panel is too full to hold any new circuit breakers. In this case it is necessary to install a circuit breaker subpanel to provide room for hooking up added circuits.

This chapter gives an easy five-step method for determining your electrical needs and planning new circuits.

Five Steps for Planning a Wiring Project

1

Examine your main service panel (page 130). The amp rating of the electrical service and the size of the circuit breaker panel will help you determine if a service upgrade is needed.

2

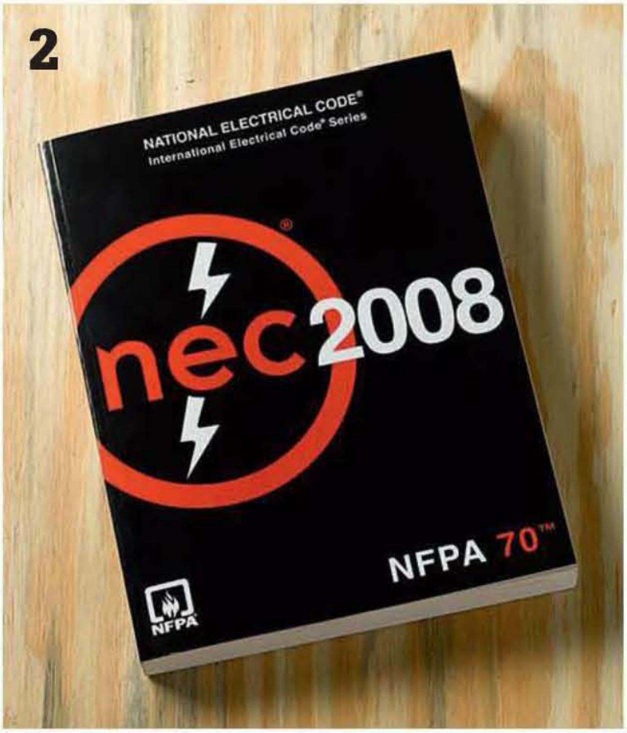
NATIONAL ELECTRICAL CODE®
International Electrical Code® Series
nec 2008
NFPA 70™
NFPA

Learn about codes (pages 131 to 135). The National Electrical Code (NEC), and local electrical codes and building codes, provide guidelines for determining how much power and how many circuits your home needs. Your local electrical inspector can tell you which regulations apply to your job.

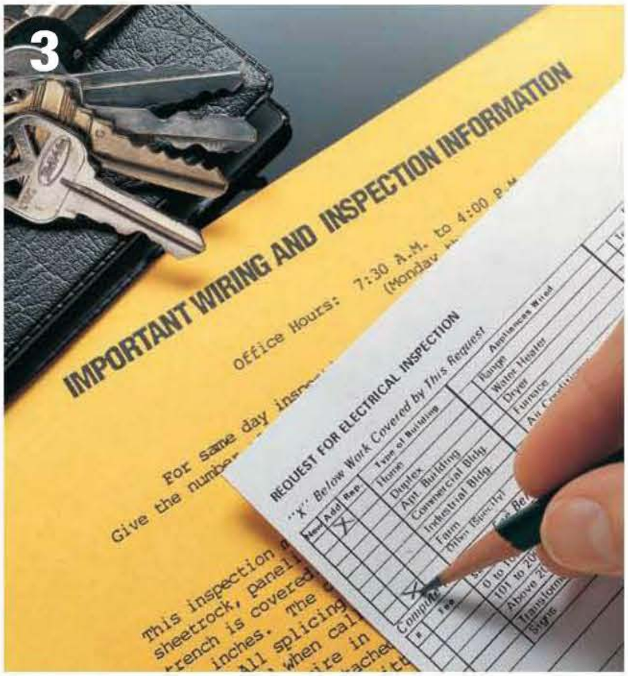
3
IMPORTANT WIRING AND INSPECTION INFORMATION
Office Hours: 7:30 A.M. to 4:00 P.M.
REQUEST FOR ELECTRICAL INSPECTION
"X" Below Work Covered by This Request

Prepare for inspection. Remember that your work must be reviewed by your local electrical inspector. When planning your wiring project, always follow the inspector's guidelines for quality workmanship.

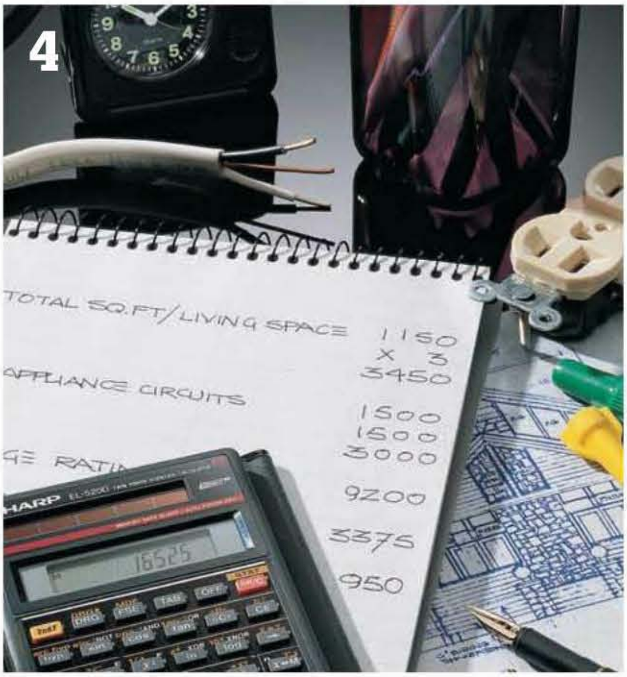
4
TOTAL SQ.FT/LIVING SPACE 1150
x 3
3450
APPLIANCE CIRCUITS
1500
1500
3000
9200
3375
950
SHARP
16525

Evaluate electrical loads. New circuits put an added load on your electrical service. Make sure that total load of the existing wiring and the planned new circuits does not exceed the main service capacity.

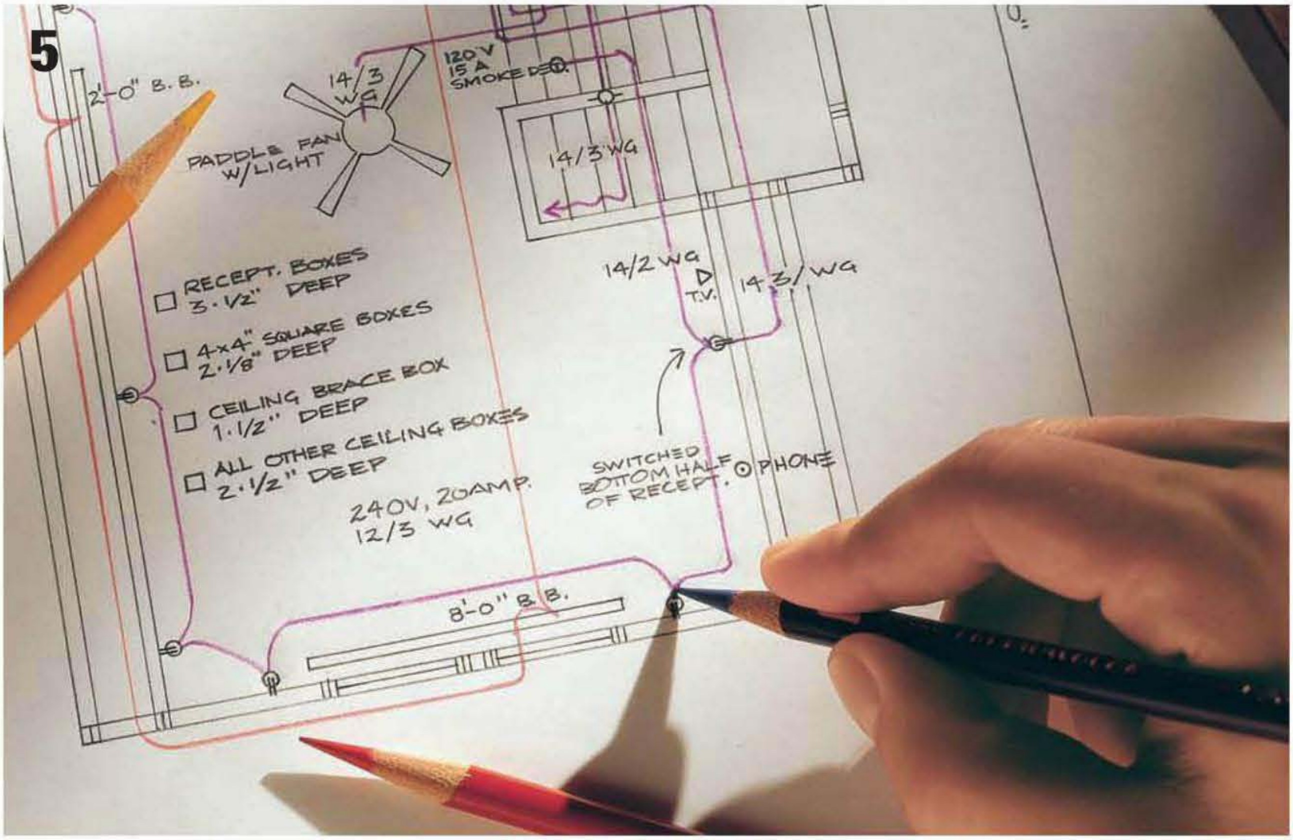
5
2'-0" B.B.
14/3 WG
120 V 15 A SMOKE DET.
PADDLE FAN W/LIGHT
14/3 WG
☐ RECEPT. BOXES 3·1/2" DEEP
☐ 4x4" SQUARE BOXES 2·1/8" DEEP
☐ CEILING BRACE BOX 1·1/2" DEEP
☐ ALL OTHER CEILING BOXES 2·1/2" DEEP
14/2 WG
T.V.
14/3/ WG
SWITCHED BOTTOM HALF OF RECEPT.
PHONE
240V, 20AMP. 12/3 WG
8'-0" B.B.

Draw a wiring diagram and get a permit (pages 136 to 137). This wiring plan will help you organize your work.

Examine Your Main Service Panel

The first step in planning a new wiring project is to look in your main circuit breaker panel and find the size of the service by reading the amperage rating on the main circuit breaker. As you plan new circuits and evaluate electrical loads, knowing the size of the main service helps you determine if you need a service upgrade.

Also look for open circuit breaker slots in the panel. The number of open slots will determine if you need to add a circuit breaker subpanel.

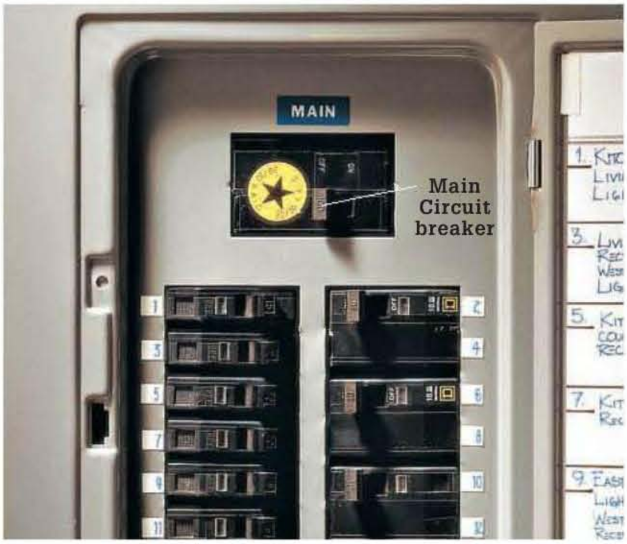

Find the service size by opening the main service panel and reading the amp rating printed on the main circuit breaker. In most cases, 100-amp service provides enough power to handle the added loads of projects like the ones shown in this book. A service rated for 60 amps or less may need to be upgraded.

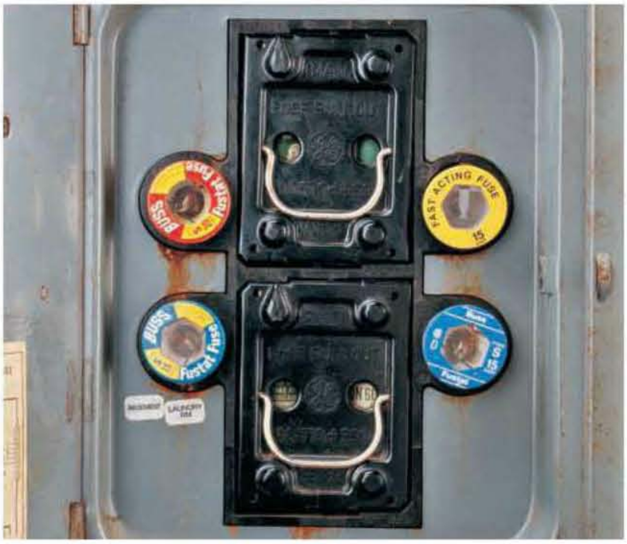

Older service panels use fuses instead of circuit breakers. Have an electrician replace this type of panel with a circuit breaker panel that provides enough power and enough open breaker slots for the new circuits you are planning.

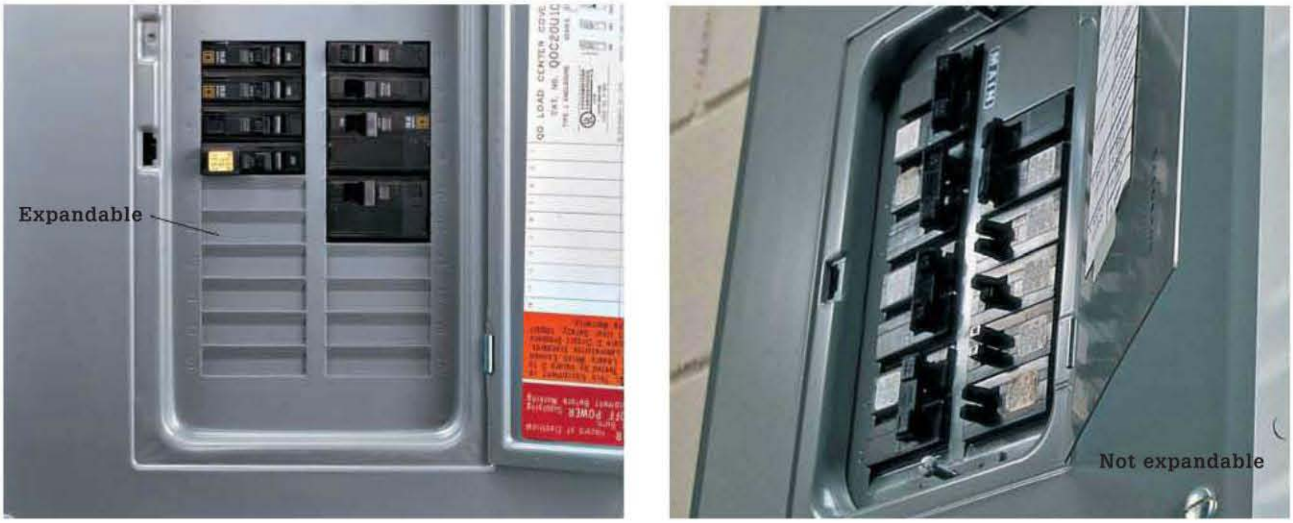

Look for open circuit breaker slots in the main circuit breaker panel or in a circuit breaker subpanel, if your home already has one. You will need one open slot for each 120-volt circuit you plan to install and two slots for each 240-volt circuit. If your main circuit breaker panel has no open breaker slots, install a subpanel to provide room for connecting new circuits.

Learn About Codes

To ensure public safety, your community requires that you get a permit to install new wiring and have the completed work reviewed by an appointed inspector. Electrical inspectors use the National Electrical Code (NEC) as the primary authority for evaluating wiring, but they also follow the local Building Code and Electrical Code standards.

As you begin planning new circuits, call or visit your local electrical inspector and discuss the project with him. The inspector can tell you which of the national and local code requirements apply to your job, and may give you a packet of information summarizing these regulations. Later, when you apply to the inspector for a work permit, he will expect you to understand the local guidelines as well as a few basic National Electrical Code requirements.

The National Electrical Code is a set of standards that provides minimum safety requirements for wiring installations. It is revised every three years. The national code requirements for the projects shown in this book are thoroughly explained on the following pages. For more information, you can find copies of the current NEC, as well as a number of excellent handbooks based on the NEC, at libraries and bookstores.

In addition to being the final authority of code requirements, inspectors are electrical professionals with years of experience. Although they have busy schedules, most inspectors are happy to answer questions and help you design well-planned circuits.

Basic Electrical Code Requirements

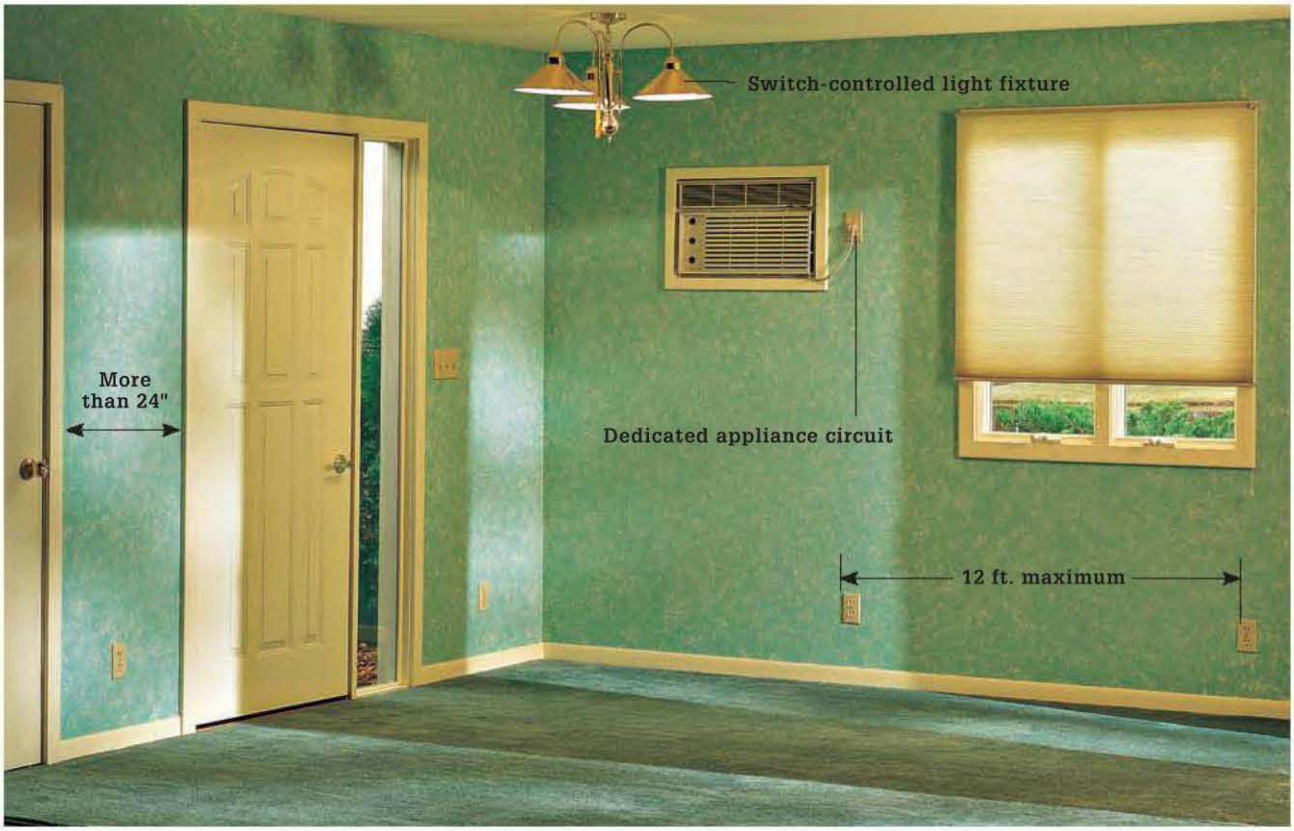

Electrical Code requirements for living areas: Living areas need at least one 15-amp or 20-amp basic lighting/receptacle circuit for each 600 sq. ft. of living space and should have a dedicated circuit for each type of permanent appliance, like an air conditioner, computer, or a group of baseboard heaters and within 6 ft. of any door opening. Receptacles on basic lighting/receptacle circuits should be spaced no more than 12 ft. apart. Many electricians and electrical inspectors recommend even closer spacing. Any wall more than 24" wide also needs a receptacle. Every room should have a wall switch at the point of entry to control either a ceiling light or plug-in lamp. Kitchens and bathrooms must have a ceiling-mounted light fixture.

Selected NEC Standards & Tips

Measure the living areas of your home, excluding closets and unfinished spaces. A sonic measuring tool gives room dimensions quickly and contains a built-in calculator for figuring floor area. You will need a minimum of one basic lighting/receptacle circuit for every 600 sq. ft. of living space. The total square footage also helps you determine heating and cooling needs for new room additions.

Three-way switches

Stairways with six steps or more must have lighting that illuminates each step. The light fixture must be controlled by three-way switches at the top and bottom landings.

Kitchen and bathroom receptacles must be protected by a ground-fault circuit-interrupter (GFCI). Also, all outdoor receptacles and general-use receptacles in an unfinished basement or crawl space and garages must be protected by a GFCI.

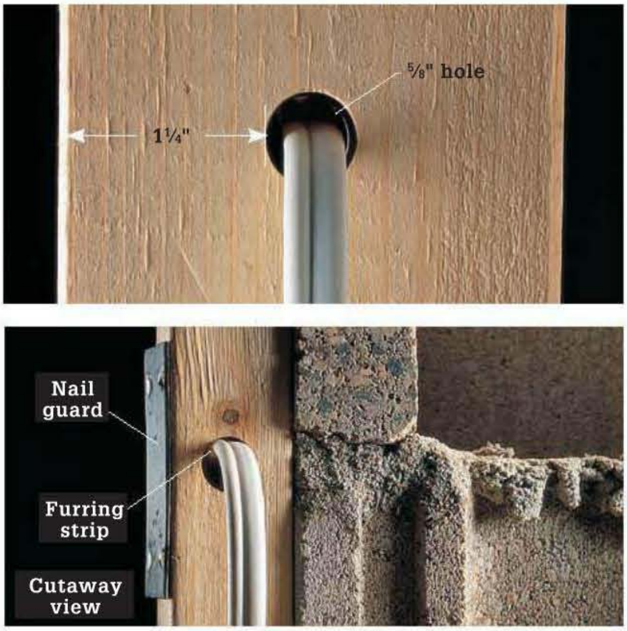

Cables must be protected against damage by nails and screws by at least 1¼" of wood (top). When cables pass through 2 × 2 furring strips (bottom), protect the cables with metal nail guards.

Closets and other storage spaces need at least one light fixture that is controlled by a wall switch near the entrance. Prevent fire hazards by positioning the light fixtures so the outer globes are at least 12" away from all shelf areas.

Hallways more than 10 ft. long need at least one receptacle. All hallways should have a switch-controlled light fixture.

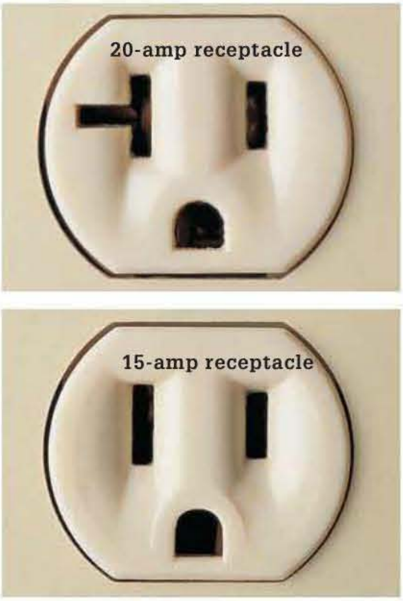
20-amp receptacle
15-amp receptacle

Amp ratings of receptacles must match the size of the circuit. A common mistake is to use 20-amp receptacles (top) on 15-amp circuits—a potential cause of dangerous circuit overloads because it allows you to plug in appliances that draw over 15 amps.

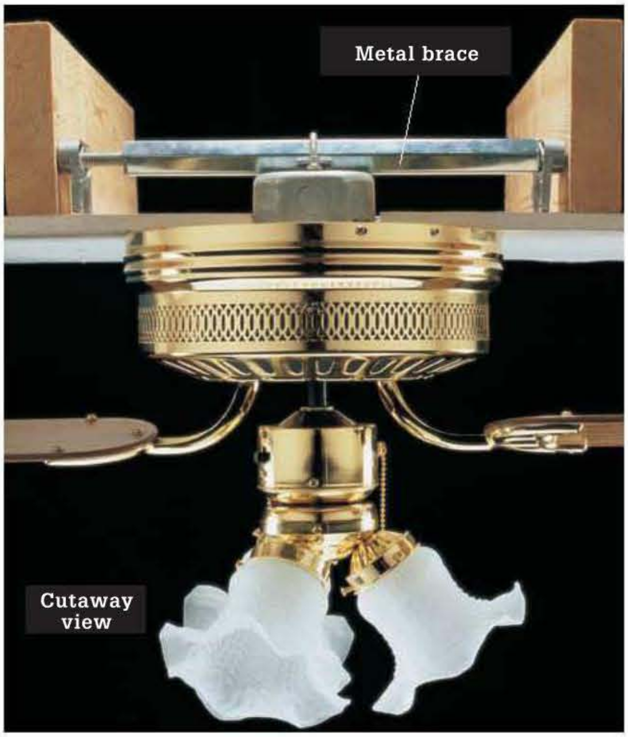
Metal brace
Cutaway view

A metal brace attached to framing members is required for ceiling fans and large light fixtures that are too heavy to be supported by an electrical box.

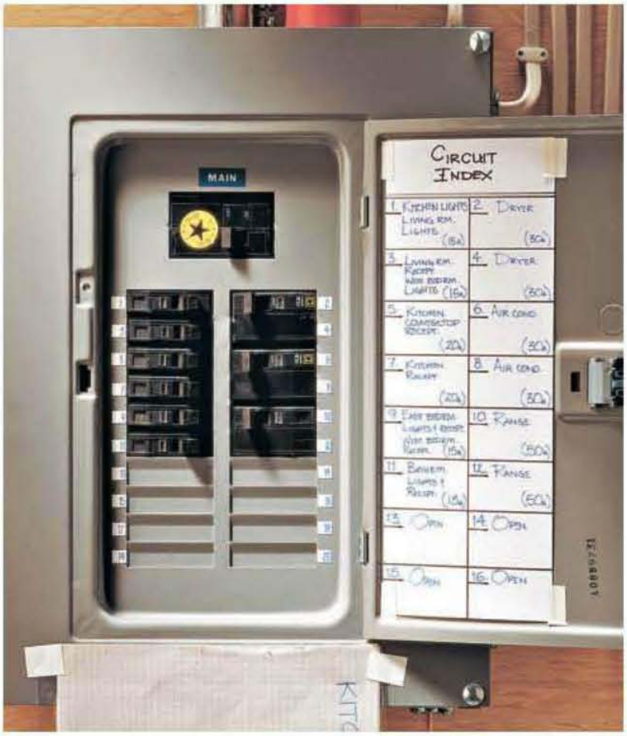
MAIN
CIRCUIT INDEX
1. KITCHEN LIGHTS LIVING RM. LIGHTS (15A)
2. DRYER (30A)
4. DRYER (30A)
5. KITCHEN COUNTERTOP RECPT. (20A)
6. AIR COND (30A)
7. KITCHEN RECPT (20A)
8. AIR COND (30A)
10. RANGE (50A)
12. RANGE (50A)
13. OPEN
14. OPEN
15. OPEN
16. OPEN

Label new circuits on an index attached to the circuit breaker panel. List the rooms and appliances controlled by each circuit. Make sure the area around the panel is clean, well lit, and accessible.

Highlights of the National Electrical Code ▸

BY MATERIAL

Service Panel (page 130)

- Maintain a minimum 30 × 36" of clearance in front of the service panel.
- Ground all 120-volt and 240-volt circuits.
- Match the amperage rating of the circuit when replacing fuses.
- Locate service panels and subpanels a maximum of 79" above floor level.
- Use handle-tie breakers for 240-volt loads (line to line).
- Close all unused service panel openings.
- Label each fuse and breaker clearly on the panel.

Electrical Boxes (page 108)

- Use boxes that are large enough to accommodate the number of wires entering the box.
- Locate all receptacle boxes 12" above the finished floor (standard).
- Locate all switch boxes 48" above the finished floor (standard). For special circumstances, inspectors will allow switch and location measurements to be altered, such as a switch at 36" above the floor in a child's bedroom or receptacles at 24" above the floor to make them more accessible for someone in a wheelchair.
- Install all boxes so they remain accessible.
- Leave no gaps greater than ⅛" between wallboard and front of electrical boxes.
- Place receptacle boxes flush with combustible surfaces.
- Leave a minimum of 8" of usable cable or wire extending past the front of the electrical box.

Wires & Cables (page 90)

- Use wires that are large enough for the amperage rating of the circuit (see Wire Size Chart, page 90).
- Drill holes at least 2" back from the exposed edge of joists to run cables through. Do not attach cables to the bottom edge of joists.
- Do not run cables diagonally between framing members.
- Run cable between receptacles 20" above the floor.
- Use nail plates to protect cable that is run through holes drilled or cut into studs less than 1¼" from front edge of stud.
- Do not crimp cables sharply.
- Contain spliced wires or connections entirely in a plastic or metal electrical box.
- Use wire connectors to join wires.
- Use staples to fasten cables within 8" of an electrical box and every 48" along its run.
- Leave a minimum ¼" (maximum 1") of sheathing where cables enter an electrical box.
- Clamp cables and wires to electrical boxes with approved NM clamp. No clamp is necessary for one-gang plastic boxes if cables are stapled within 8".
- Label all cables and wires at each electrical box to show which circuits they serve for the rough-in inspection.
- Connect only a single wire to a single screw terminal. Use pigtails to join more than one wire to a screw terminal.

Switches (page 116)

- Use a switch-controlled receptacle in rooms without a built-in light fixture operated by a wall switch.
- Use three-way switches at the top and bottom on stairways with six steps or more.
- Use switches with grounding screw with plastic electrical boxes.
- Locate all wall switches within easy reach of the room entrance.

Receptacles (page 122)

- Match the amp rating of a receptacle with the size of the circuit.
- Include receptacles on all walls 24" wide or greater.
- Include receptacles so a 6-ft. cord can be plugged in from any point along a wall or every 12-ft. along a wall.
- Include receptacles in any hallway that is 10-ft. long or more.
- Use three-prong, grounded receptacles for all 15- or 20-amp, 120-volt branch circuits.
- Include a switch-controlled receptacle in rooms without a built-in light fixture operated by a wall switch.
- Install GFCI-protected receptacles in bathrooms, kitchens, garages, crawl spaces, unfinished basements, and outdoor receptacle locations.
- Install an isolated-ground circuit to protect sensitive equipment, like a computer, against tiny power fluctuations. Computers should also be protected by a standard surge protector.

Light Fixtures

- Use mounting straps that are anchored to the electrical boxes to mount ceiling fixtures.
- Keep non-IC-rated recessed light fixtures 3" from insulation and ½" from combustibles.
- Include at least one switch-operated lighting fixture in every room.

Grounding

- Ground all receptacles by connecting receptacle grounding screws to the circuit grounding wires.
- Use switches with grounding screws whenever possible. Always ground switches installed in plastic electrical boxes and all switches in kitchens, bathrooms, and basements.

BY ROOM

Kitchens/Dining Rooms

- Install a dedicated 40- or 50-amp, 120/240-volt circuit for a range (or two circuits for separate oven and countertop units).
- Install two 20-amp small appliance circuits.
- Install dedicated 15-amp, 120-volt circuits for dishwashers and food disposals (required by many local codes).
- Use GFCI receptacles for all accessible countertop receptacles; receptacles behind fixed appliances do not need to be GFCIs.
- Position receptacles for appliances that will be installed within cabinets, such as microwaves or food disposals, according to the manufacturer's instructions.
- Include receptacles on all counters wider than 12".
- Space receptacles a maximum of 48" apart above countertops and closer together in areas where many appliances will be used.
- Locate receptacles 18" above countertop. If backsplash is more than the standard 4" or the bottom of cabinet is less than 18" from countertop, center the box in space between countertop and bottom of wall cabinet.
- Mount one receptacle within 12" of the countertop on islands and peninsulas that are 12 × 24" or greater.
- Do not put lights on small appliance circuits.
- Install additional lighting in work areas at a sink or range for convenience and safety.

Bathrooms

- Install a separate 20-amp circuit.
- Ground switches in bathrooms.
- Use GFCI-protected receptacles.
- Install at least one ceiling-mounted light fixture.
- Place blower heaters in bathrooms well away from the sink and tub.

Utility/Laundry Rooms

- Install a separate 20-amp circuit for a washing machine.
- Install a minimum feed 30-amp #10 THHN wire for the dryer powered by a separate 120/240-volt major appliance circuit.
- Install metal conduit for cable runs in unfinished rooms.
- Use GFCI-protected receptacles, except for fixed appliances, such as freezers or dryers.

Living, Entertainment, Bedrooms

- Install a minimum of two 15-amp circuits in living rooms.
- Install a minimum of one 15- or 20-amp basic lighting/receptacle circuit for each 600 sq. ft. of living space.
- Install a dedicated circuit for each permanent appliance, like an air conditioner, computer, or group of electric baseboard heaters.
- Do not use standard electrical boxes to support ceiling fans.
- Include receptacles on walls 24" wide or more.
- Space receptacles on basic lighting/receptacle circuits a maximum of 12 ft. apart. For convenience you can space them as close as 6 ft.
- Position permanent light fixtures in the center of the room's ceiling.
- Install permanently wired smoke alarms in room additions that include sleeping areas.

Outdoors

- Check for underground utilities before digging.
- Use UF cable for outdoor wiring needs.
- Run cable in schedule 80 PVC plastic, as required by local code.
- Bury cables housed in conduit at least 18" deep; cable not in conduit must be buried at least 24" deep.
- Use weatherproof electrical boxes with watertight covers.
- Use GFCI-protected receptacles.
- Install receptacles a minimum of 12" above ground level.
- Anchor freestanding receptacles not attached to a structure by embedding the schedule 80 PVC plastic conduit in a concrete footing, so that it is at least 12" but no more than 18" above ground level.
- Plan on installing a 20-amp, 120-volt circuit if the circuit contains more than one light fixture rated for 300 watts, or more than four receptacles.

Stairs/Hallways

- Use three-way switches at the top and bottom on stairways with six steps or more.
- Include receptacles in any hallway that is 10 ft. long or more.
- Position stairway lights so each step is illuminated.

Draw a Diagram & Obtain a Permit

Drawing a wiring diagram is the last step in planning a circuit installation. A detailed wiring diagram helps you get a work permit, makes it easy to create a list of materials, and serves as a guide for laying out circuits and installing cables and fixtures. Use the circuit maps on pages 138 to 145 as a guide for planning wiring configurations and cable runs. Bring the diagram and materials list when you visit electrical inspectors to apply for a work permit.

Never install new wiring without following your community's permit and inspection procedure. A work permit is not expensive, and it ensures that your work will be reviewed by a qualified inspector to guarantee its safety. If you install new wiring without the proper permit, an accident or fire traced to faulty wiring could cause your insurance company to discontinue your policy and can hurt the resale value of your home.

When electrical inspectors look over your wiring diagram, they will ask questions to see if you have a basic understanding of the electrical code and fundamental wiring skills. Some inspectors ask these questions informally, while others give a short written test. Inspectors may allow you to do some, but not all, of the work. For example, they may ask that all final circuit connections at the circuit breaker panel be made by a licensed electrician, while allowing you to do all other work.

A few communities allow you to install wiring only when supervised by an electrician. This means you can still install your own wiring but must hire an electrician to apply for the work permit and to check your work before inspectors review it. The electrician is held responsible for the quality of the job.

Remember that it is the inspectors' responsibility to help you do a safe and professional job. Feel free to call them with questions about wiring techniques or materials.

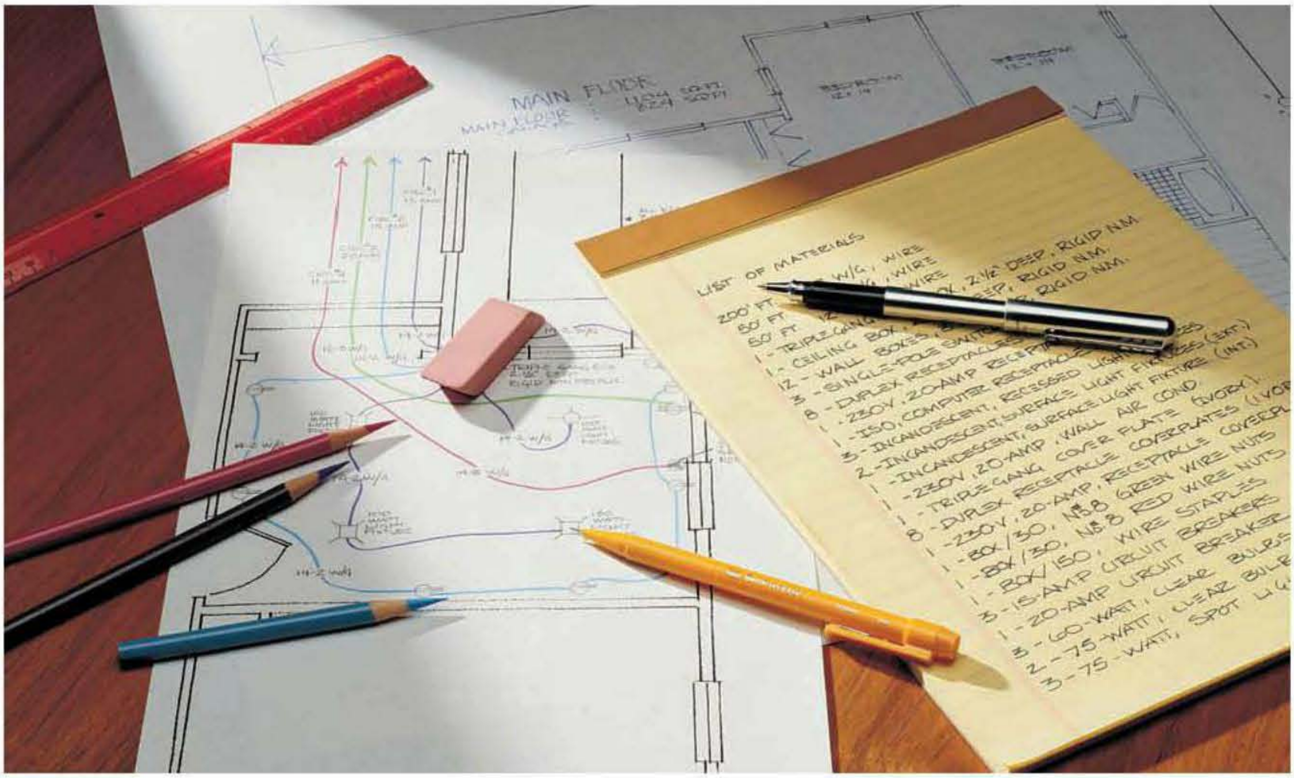
MAIN FLOOR
LIST OF MATERIALS
200' FT
50' FT
50' FT
W/G, WIRE
WIRE
WIRE
1 - TRIPLE GANG BOX, 2½" DEEP, RIGID NM.
1 - CEILING BOX, DEEP, RIGID NM.
12 - WALL BOXES, RIGID NM.
3 - SINGLE-POLE SWITCHES
8 - DUPLEX RECEPTACLES
1 - 230V, 20-AMP RECEPTACLE
1 - 150, COMPUTER RECEPTACLE
3 - INCANDESCENT, RECESSED LIGHT FIXTURES (EXT.)
2 - INCANDESCENT, SURFACE LIGHT FIXTURE (INT.)
1 - 230V, 20-AMP, WALL AIR COND.
1 - TRIPLE GANG COVER PLATE (IVORY)
8 - DUPLEX RECEPTACLE COVERPLATES (IVOR
1 - 230V, 20-AMP, RECEPTACLE COVERPL
1 - BOX/30, No.8 GREEN WIRE NUTS
1 - BOX/30, No.8 RED WIRE NUTS
1 - BOX/150, WIRE STAPLES
3 - 15-AMP CIRCUIT BREAKERS
1 - 20-AMP CIRCUIT BREAKER
3 - 60-WATT, CLEAR BULBS
2 - 75-WATT, CLEAR BULBS
3 - 75-WATT, SPOT LIG

A detailed wiring diagram and a list of materials is required before electrical inspectors will issue a work permit. If blueprints exist for the space you are remodeling, start your electrical diagram by tracing the wall outlines from the blueprint. Use standard electrical symbols (next page) to clearly show all the receptacles, switches, light fixtures, and permanent appliances. Make a copy of the symbol key, and attach it to the wiring diagram for the inspectors' convenience. Show each cable run, and label its wire size and circuit amperage.

How to Draw a Wiring Plan

Draw a scaled diagram of the space you will be wiring, showing walls, doors, windows, plumbing pipes and fixtures, and heating and cooling ducts. Find the floor space by multiplying room length by width, and indicate this on the diagram. Do not include closets or storage areas when figuring space.

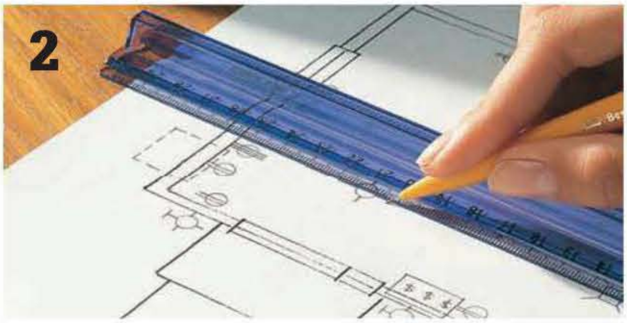

Mark the location of all switches, receptacles, light fixtures, and permanent appliances using the electrical symbols shown below. Where you locate these devices along the cable run determines how they are wired. Use the circuit maps on pages 138 to 145 as a guide for drawing wiring diagrams.

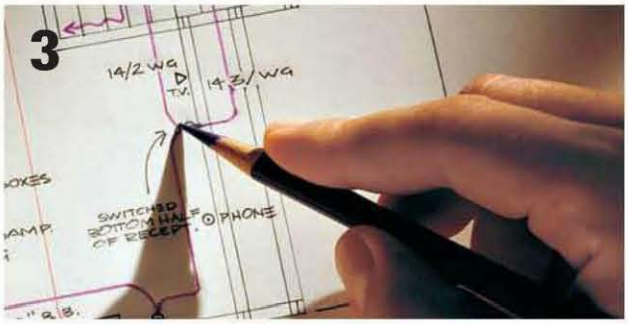
3
14/2 WG
14-3/WG
SWITCHED BOTTOM HALF OF RECEPT.
PHONE

Draw in cable runs between devices. Indicate cable size and type, and the amperage of the circuits. Use a different-colored pencil for each circuit.

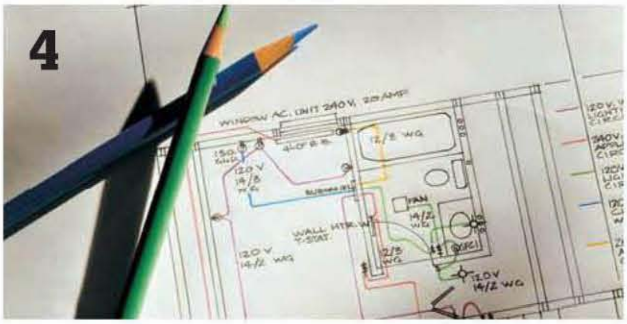
4
WINDOW AC. UNIT 240V, 20 AMP
WALL HTR. W/ T-STAT
120 V 14/2 WG

Identify the wattages for light fixtures and permanent appliances, and the type and size of each electrical box. On another sheet of paper, make a detailed list of all materials you will use.

Electrical Symbol Key ▸

(copy this key and attach it to your wiring plan)

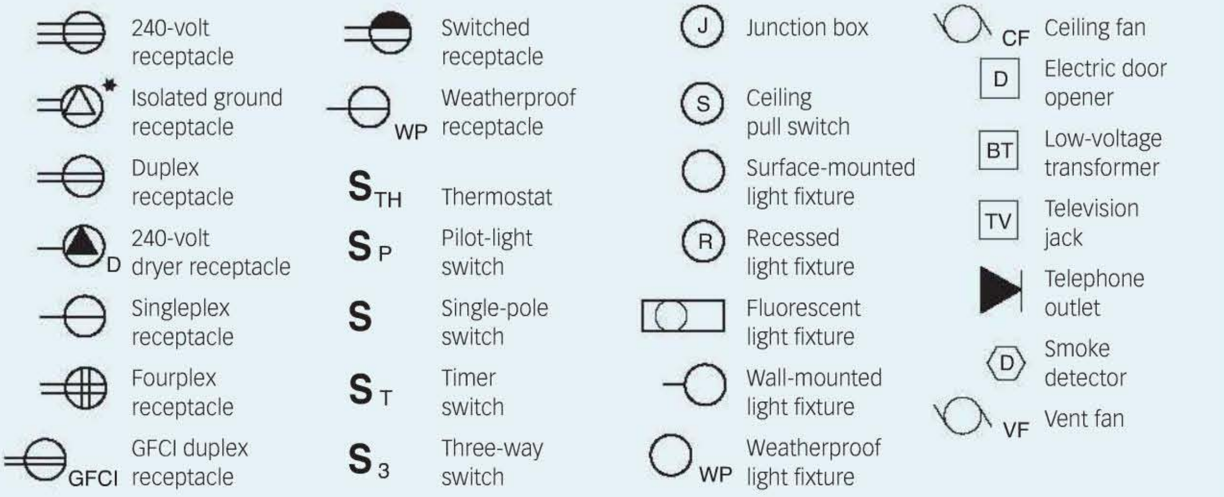
240-volt receptacle
Isolated ground receptacle
Duplex receptacle
D 240-volt dryer receptacle
Singleplex receptacle
Fourplex receptacle
GFCI GFCI duplex receptacle
Switched receptacle
WP Weatherproof receptacle
S TH Thermostat
S P Pilot-light switch
S Single-pole switch
S T Timer switch
S 3 Three-way switch
J Junction box
S Ceiling pull switch
Surface-mounted light fixture
R Recessed light fixture
Fluorescent light fixture
Wall-mounted light fixture
WP Weatherproof light fixture
CF Ceiling fan
D Electric door opener
BT Low-voltage transformer
TV Television jack
Telephone outlet
D Smoke detector
VF Vent fan

Using Circuit Maps

The arrangement of switches and appliances along an electrical circuit differs for every project. This means that the configuration of wires inside an electrical box can vary greatly, even when fixtures are identical.

The circuit maps on the following pages show the most common wiring variations for typical electrical devices. Most new wiring you install will match one or more of the maps shown. Find the maps that match your situation and use them to plan your circuit layouts.

The 120-volt circuits shown on the following pages are wired for 15 amps using 14-gauge wire and receptacles rated at 15 amps. If you are installing a 20-amp circuit, substitute 12-gauge cables and use receptacles rated for 20 amps.

In configurations where a white wire serves as a hot wire instead of a neutral, both ends of the wire are coded with black tape to identify it as hot. In addition, each of the circuit maps shows a box grounding screw. This grounding screw is required in all metal boxes, but plastic electrical boxes do not need to be grounded.

Note: For clarity, all grounding conductors in the circuit maps are colored green. In practice, the grounding wires inside sheathed cables usually are bare copper.

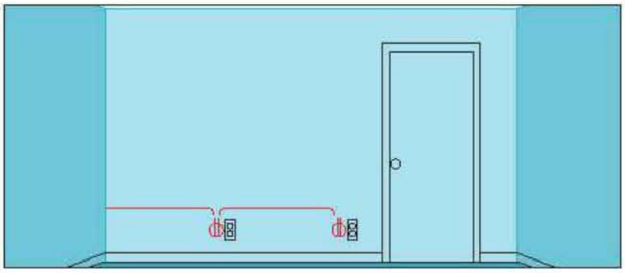

1. 120-VOLT DUPLEX RECEPTACLES WIRED IN SEQUENCE

Use this layout to link any number of duplex receptacles in a basic lighting/receptacle circuit. The last receptacle in the cable run is connected like the receptacle shown at the right side of the circuit map below. All other receptacles are wired like the receptacle shown on the left side. Requires two-wire cables.

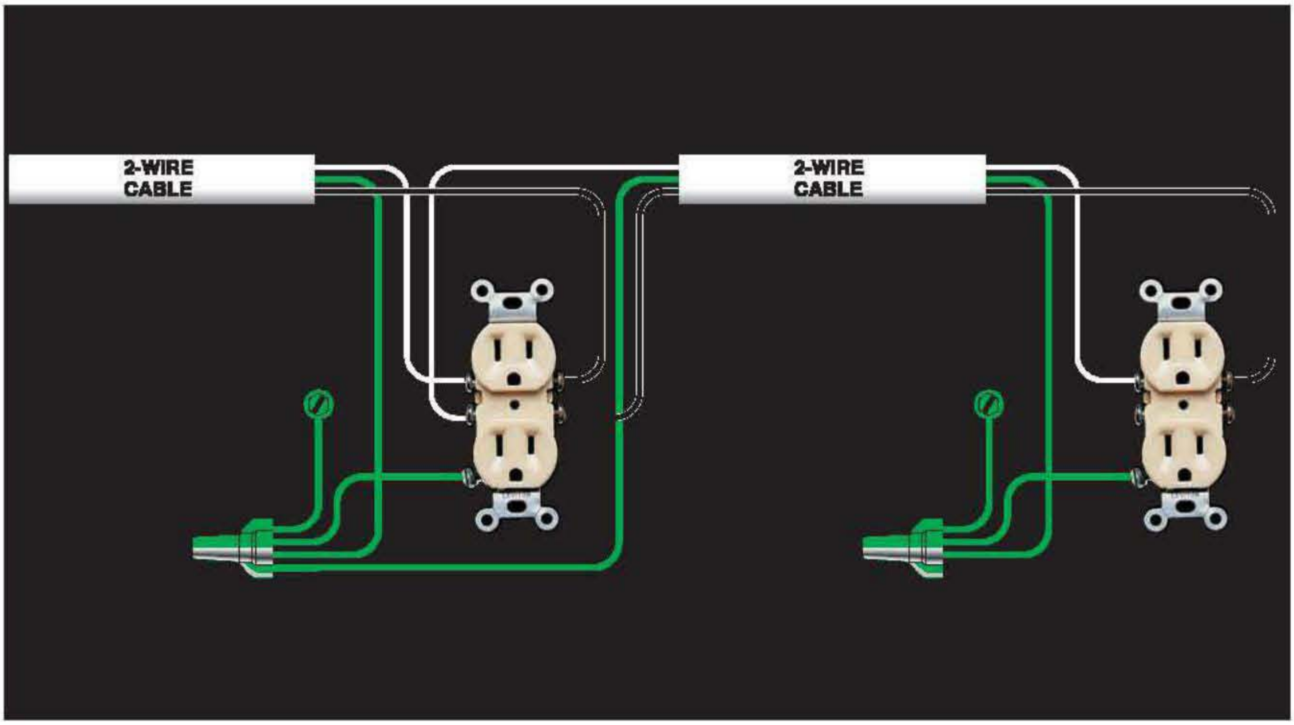

2. GFCI RECEPTACLES (Single-location Protection)

Use this layout when receptacles are within 6 ft. of a water source, like those in kitchens and bathrooms. To prevent nuisance tripping caused by normal power surges, GFCIs should be connected only at the line screw terminal so they protect a single location, not the fixtures on the load side of the circuit. Requires two-wire cables. Where a GFCI must protect other fixtures, use circuit map 3.

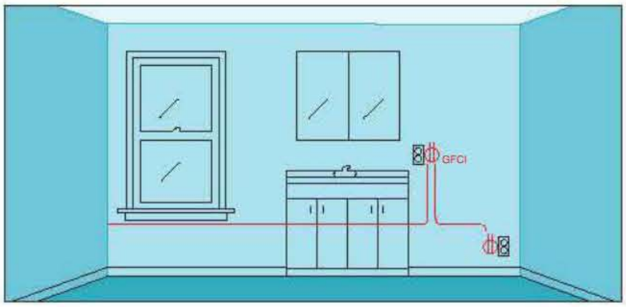

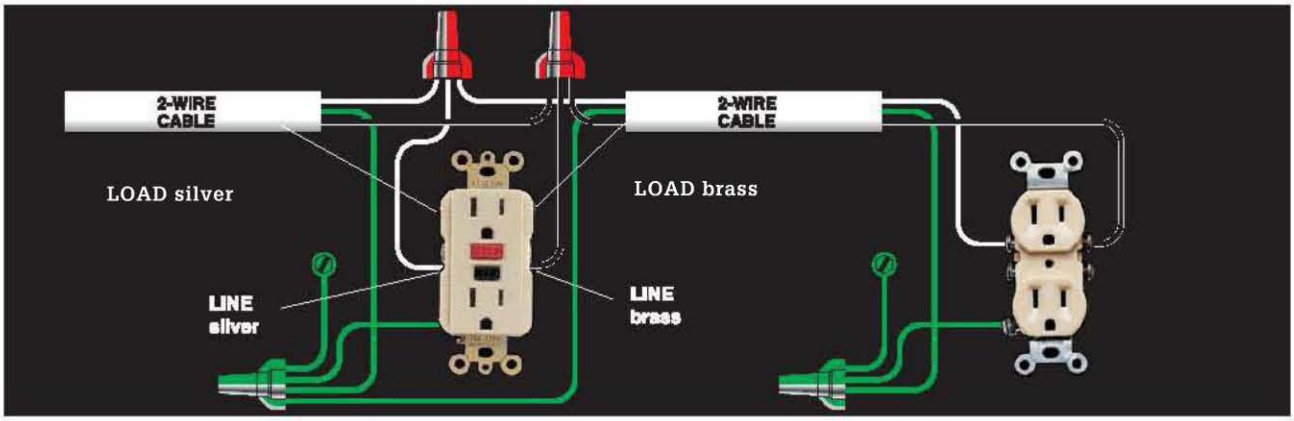
2-WIRE CABLE
2-WIRE CABLE
LOAD silver
LOAD brass
LINE silver
LINE brass

3. GFCI RECEPTACLE, SWITCH & LIGHT FIXTURE (Wired for Multiple-location Protection)

In some locations, such as an outdoor circuit, it is a good idea to connect a GFCI receptacle so it also provides shock protection to the wires and fixtures that continue to the end of the circuit. Wires from the power source are connected to the line screw terminals; outgoing wires are connected to load screws. Requires two-wire cables.

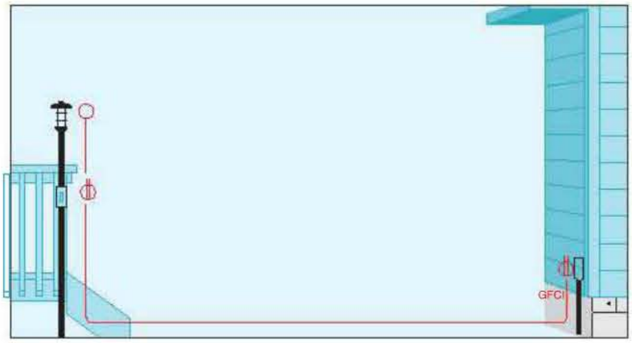

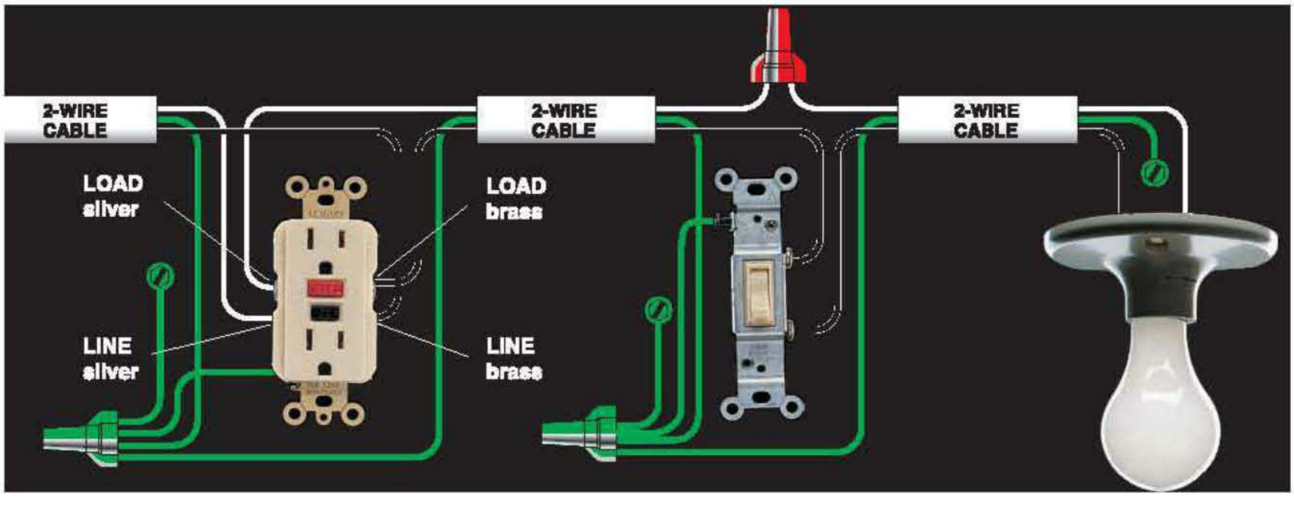
2-WIRE CABLE
2-WIRE CABLE
2-WIRE CABLE
LOAD silver
LOAD brass
LINE silver
LINE brass

4. SINGLE-POLE SWITCH & LIGHT FIXTURE (Light Fixture at End of Cable Run)

Use this layout for light fixtures in basic lighting/receptacle circuits throughout the home. It is often used as an extension to a series of receptacles (circuit map 1). Requires two-wire cables.

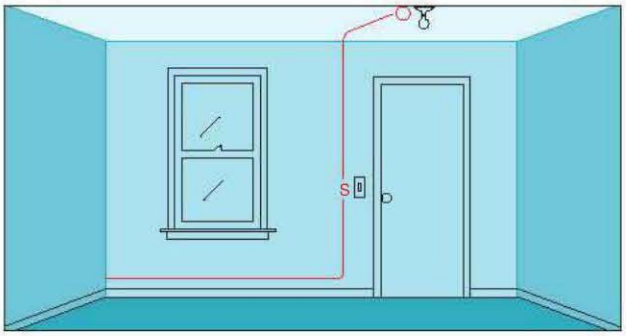

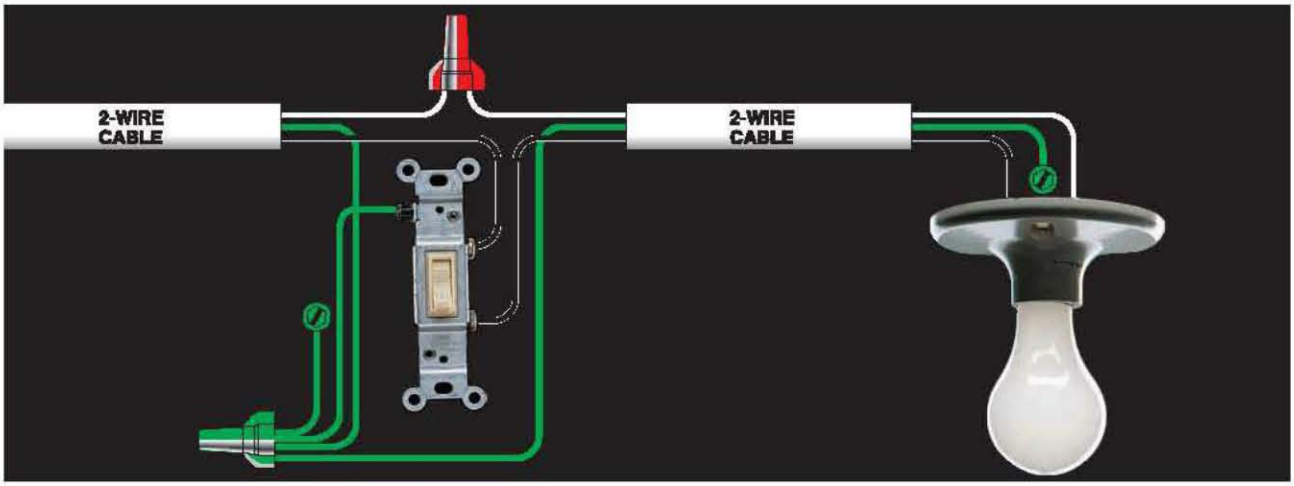
2-WIRE CABLE
2-WIRE CABLE

5. SINGLE-POLE SWITCH & LIGHT FIXTURE (Switch at End of Cable Run)

Use this layout, sometimes called a switch loop, where it is more practical to locate a switch at the end of the cable run. In the last length of cable, both insulated wires are hot; the white wire is tagged with black tape at both ends to indicate it is hot. Requires two-wire cables.

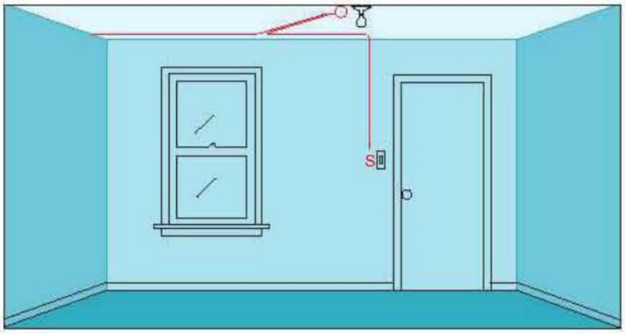

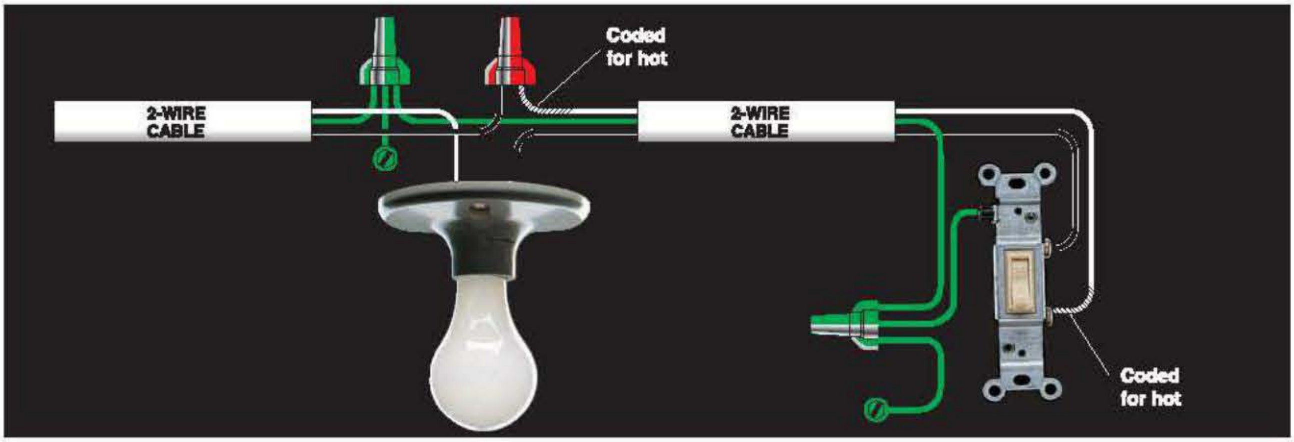
Coded for hot
2-WIRE CABLE
2-WIRE CABLE
Coded for hot

6. SINGLE-POLE SWITCH & TWO LIGHT FIXTURES (Switch Between Light Fixtures, Light at Start of Cable Run)

Use this layout when you need to control two fixtures from one single-pole switch and the switch is between the two lights in the cable run. Power feeds to one of the lights. Requires two-wire and three-wire cables.

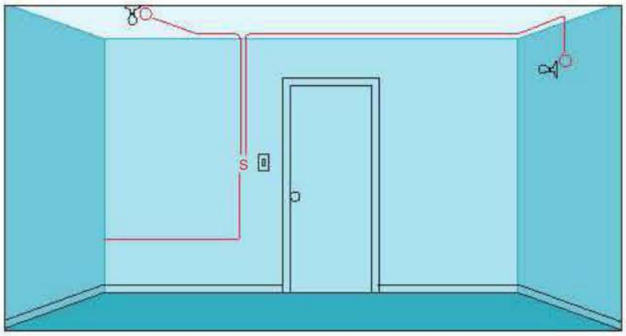

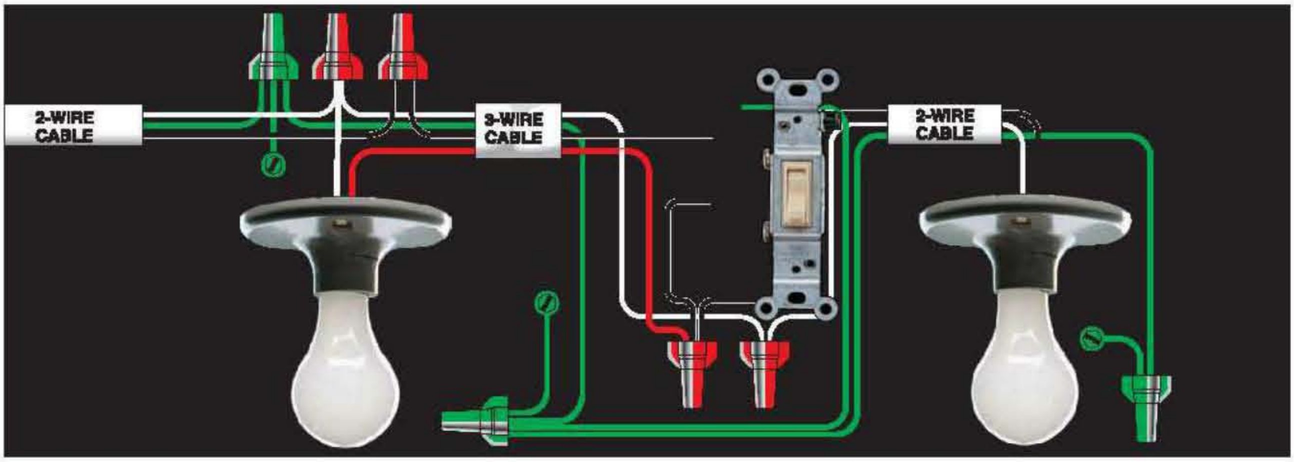
2-WIRE CABLE
3-WIRE CABLE
2-WIRE CABLE

7. SINGLE-POLE SWITCH & LIGHT FIXTURE, DUPLEX RECEPTACLE (Switch at Start of Cable Run)

Use this layout to continue a circuit past a switched light fixture to one or more duplex receptacles. To add multiple receptacles to the circuit, see circuit map 1. Requires two-wire and three-wire cables.

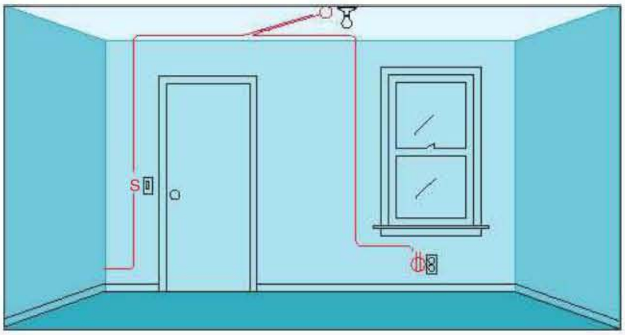

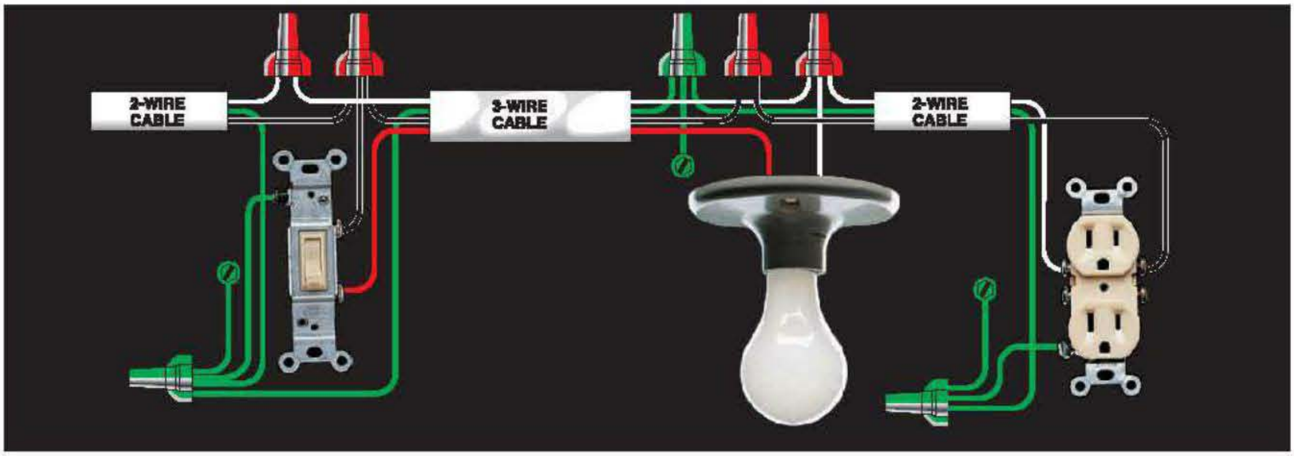
2-WIRE CABLE
3-WIRE CABLE
2-WIRE CABLE

8. DOUBLE RECEPTACLE SMALL-APPLIANCE CIRCUIT WITH GFCIs & SHARED NEUTRAL WIRE

Use this layout to wire a double receptacle circuit when code requires that some of the receptacles be GFCIs. The GFCIs should be wired for single-location protection (see circuit map 2). Requires three-wire and two-wire cables.

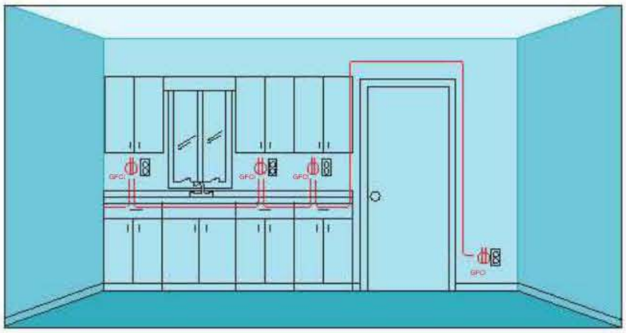
GFCI
GFCI
GFCI
GFCI

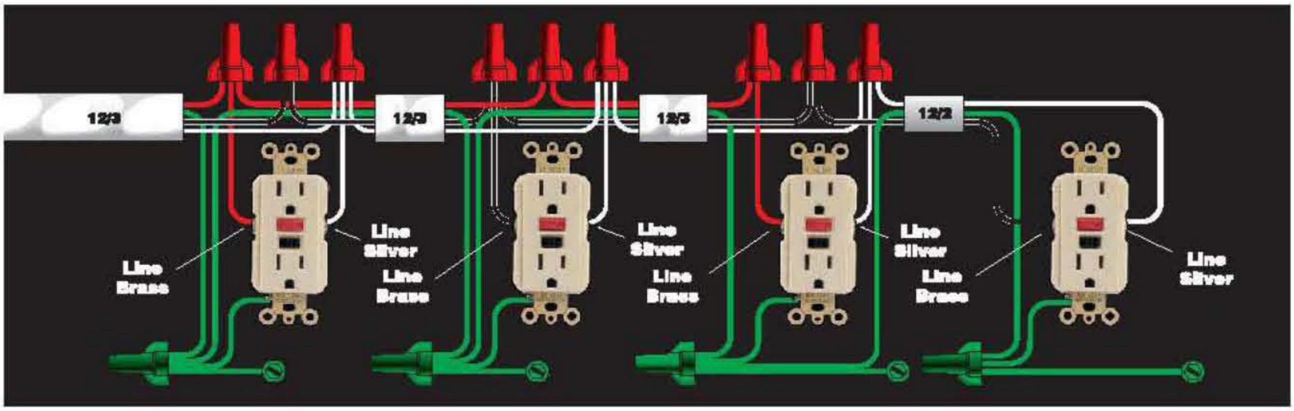
12/3
12/3
12/3
12/2
Line Brass
Line Silver
Line Brass
Line Silver
Line Brass
Line Silver
Line Brass
Line Silver

9. DOUBLE RECEPTACLE SMALL APPLIANCE CIRCUIT WITH GFCIs & SEPARATE NEUTRAL WIRES

If the room layout or local codes do not allow for a shared neutral wire, use this layout instead. The GFCIs should be wired for single-location protection (see circuit map 2). Requires two-wire cable.

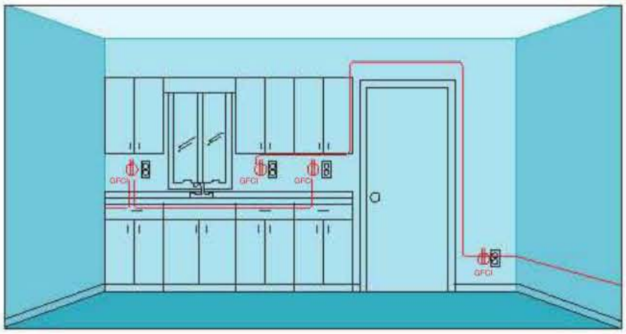
GFCI
GFCI
GFCI
GFCI

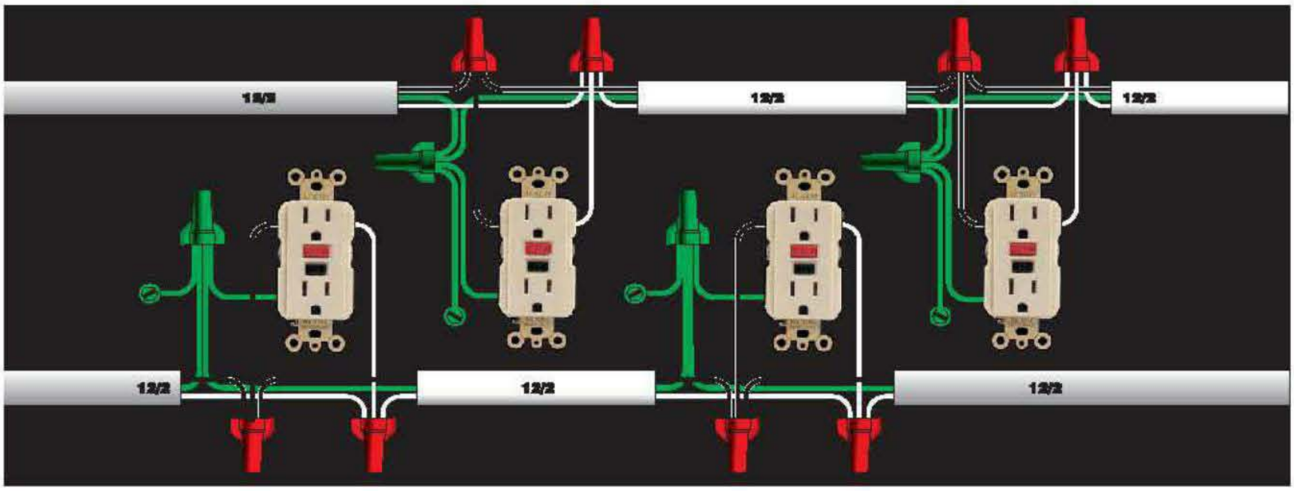
12/2
12/2
12/2
12/2
12/2
12/2

10. 120/240-VOLT RANGE RECEPTACLE

This layout is for a 50- or 60-amp, 120/240-volt dedicated appliance circuit wired with 6/3 cable, as required by code for a large kitchen range. The black and red circuit wires, connected to a double-pole circuit breaker in the circuit breaker panel, each bring 120 volts of power to the setscrew terminals on the receptacle. The white circuit wire attached to the neutral bus bar in the circuit breaker panel is connected to the neutral setscrew terminal on the receptacle.

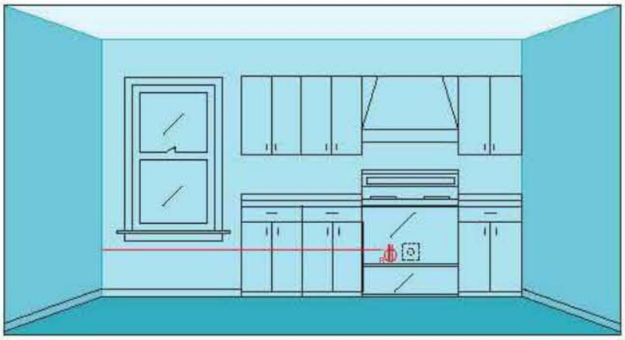

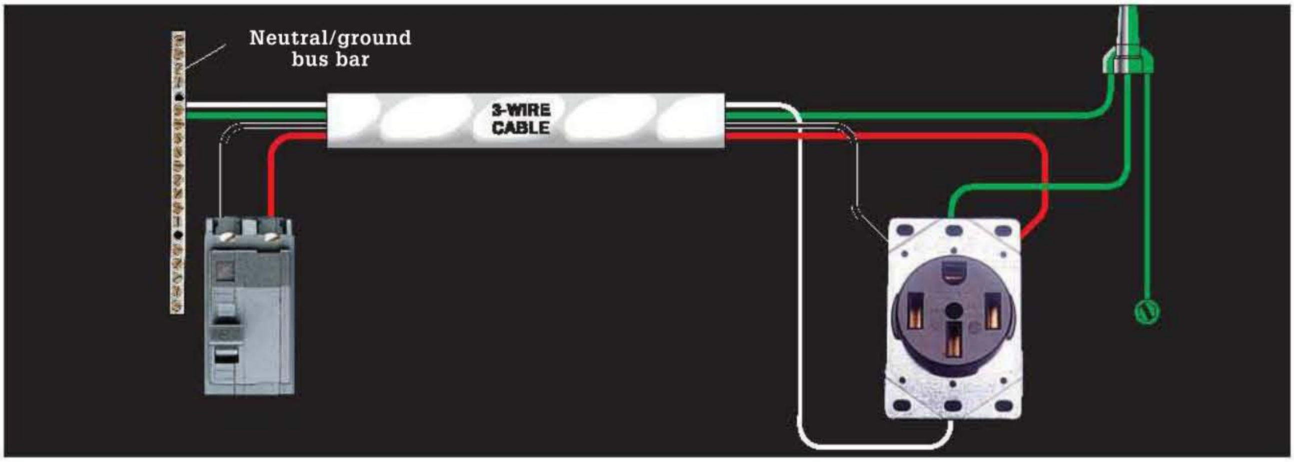
Neutral/ground bus bar
3-WIRE CABLE

11. 240-VOLT BASEBOARD HEATERS, THERMOSTAT

This layout is typical for a series of 240-volt baseboard heaters controlled by a wall thermostat. Except for the last heater in the circuit, all heaters are wired as shown below. The last heater is connected to only one cable. The size of the circuit and cables are determined by finding the total wattage of all heaters. Requires two-wire cable.

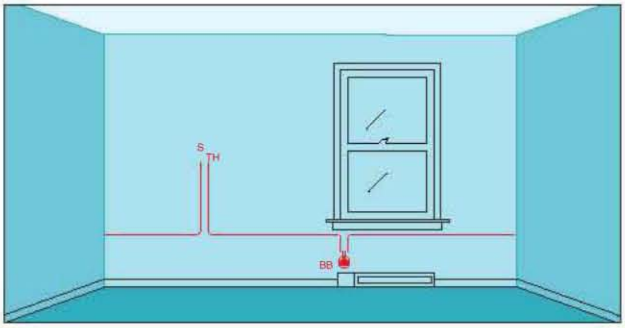

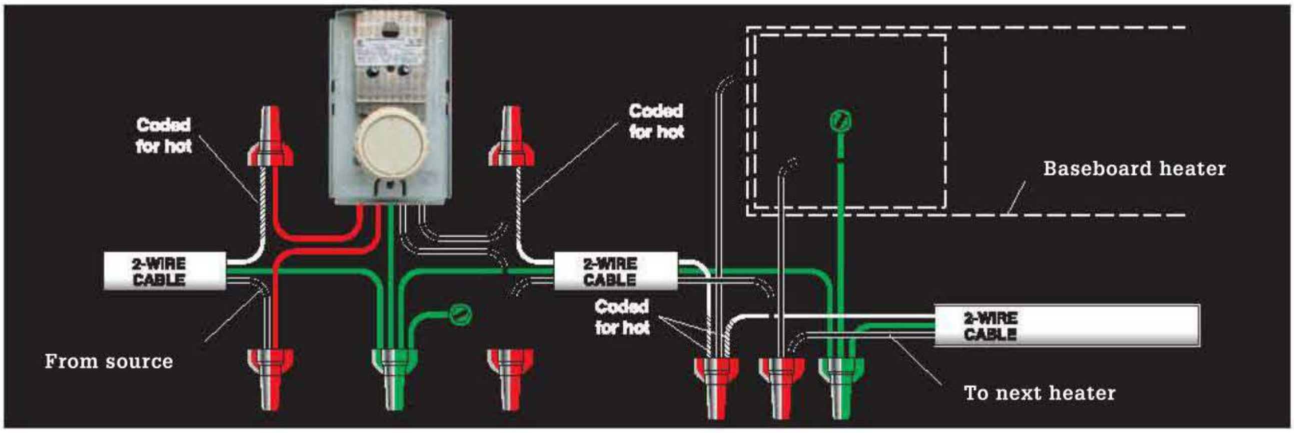
Coded for hot
Coded for hot
2-WIRE CABLE
2-WIRE CABLE
Coded for hot
Baseboard heater
2-WIRE CABLE
From source
To next heater

12. THREE-WAY SWITCHES & LIGHT FIXTURE (Fixture Between Switches)

This layout for three-way switches lets you control a light fixture from two locations. Each switch has one common screw terminal and two traveler screws. Circuit wires attached to the traveler screws run between the two switches, and hot wires attached to the common screws bring current from the power source and carry it to the light fixture. Requires two-wire and three-wire cables.

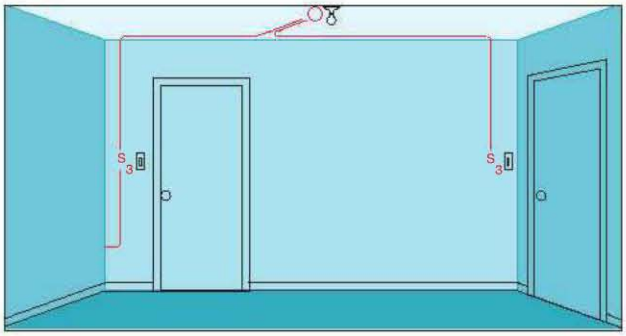

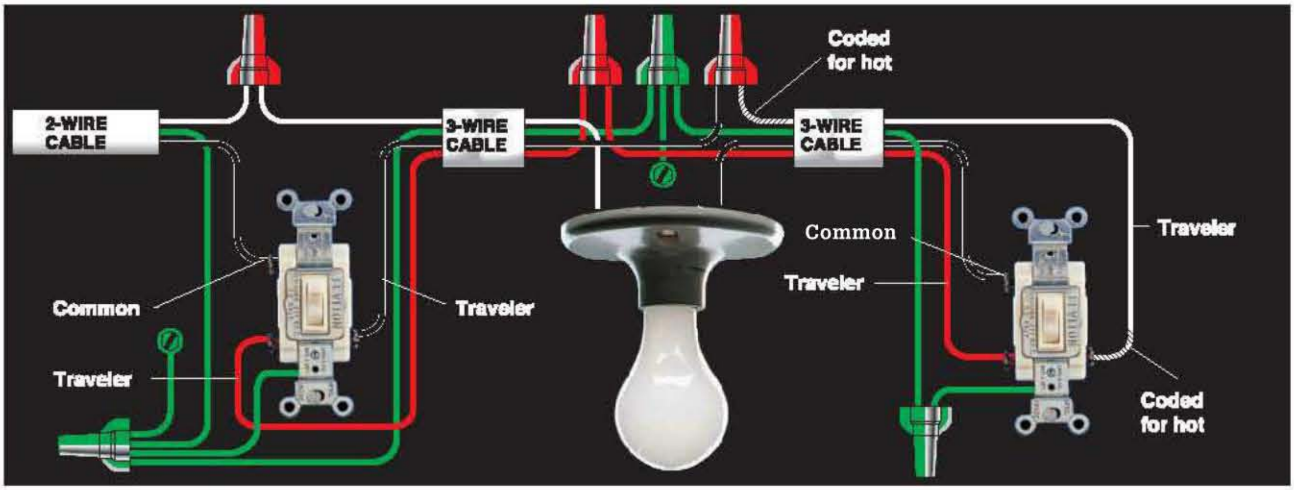
2-WIRE CABLE
3-WIRE CABLE
Coded for hot
3-WIRE CABLE
Common
Traveler
Common
Traveler
Traveler
Common
Traveler
Coded for hot

13. THREE-WAY SWITCHES & LIGHT FIXTURE (Fixture at Start of Cable Run)

Use this layout when it is more convenient to locate the fixture ahead of the three-way switches in the cable run. Requires two-wire and three-wire cables.

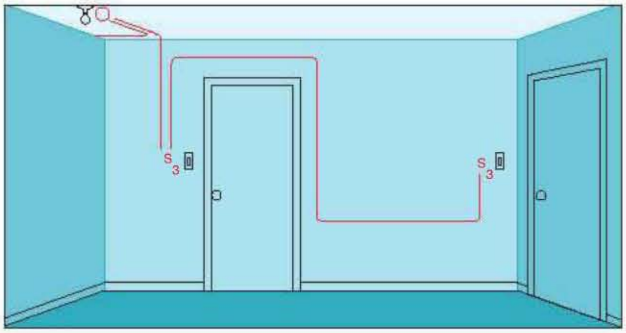

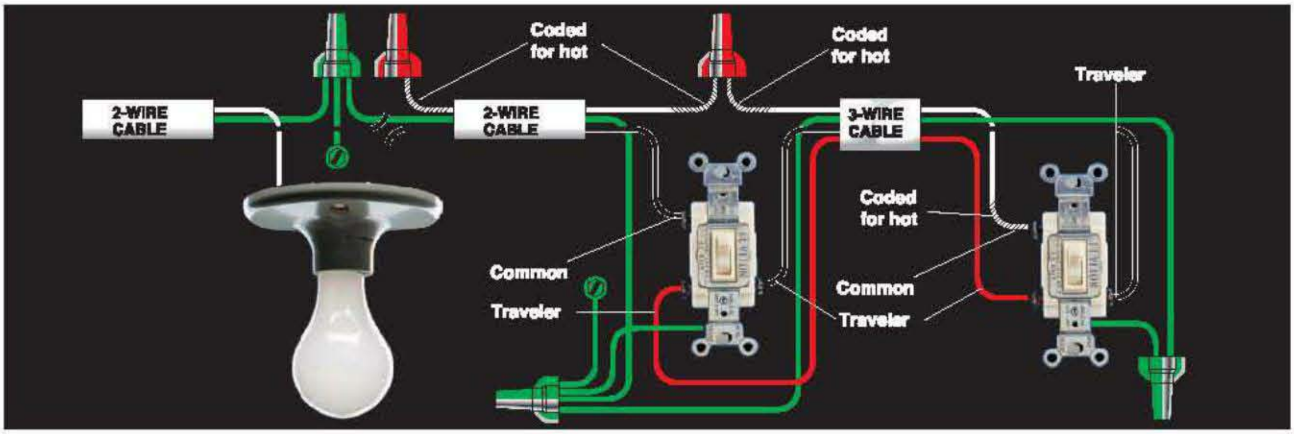
Coded for hot
Coded for hot
Traveler
2-WIRE CABLE
2-WIRE CABLE
3-WIRE CABLE
Coded for hot
Common
Traveler
Common
Traveler

14. FOUR-WAY SWITCH & LIGHT FIXTURE (Fixture at Start of Cable Run)

This layout lets you control a light fixture from three locations. The end switches are three-way and the middle is four-way. A pair of three-wire cables enter the box of the four-way switch. The white and red wires from one cable attach to the top pair of screw terminals (line 1) and the white and red wires from the other cable attaches to the bottom screw terminals (line 2). Requires two three-way switches and one four-way switch and two-wire and three-wire cables.

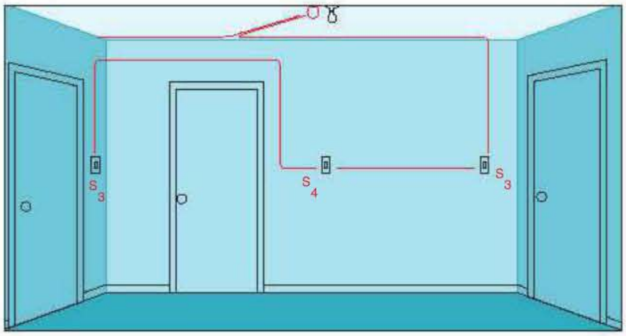

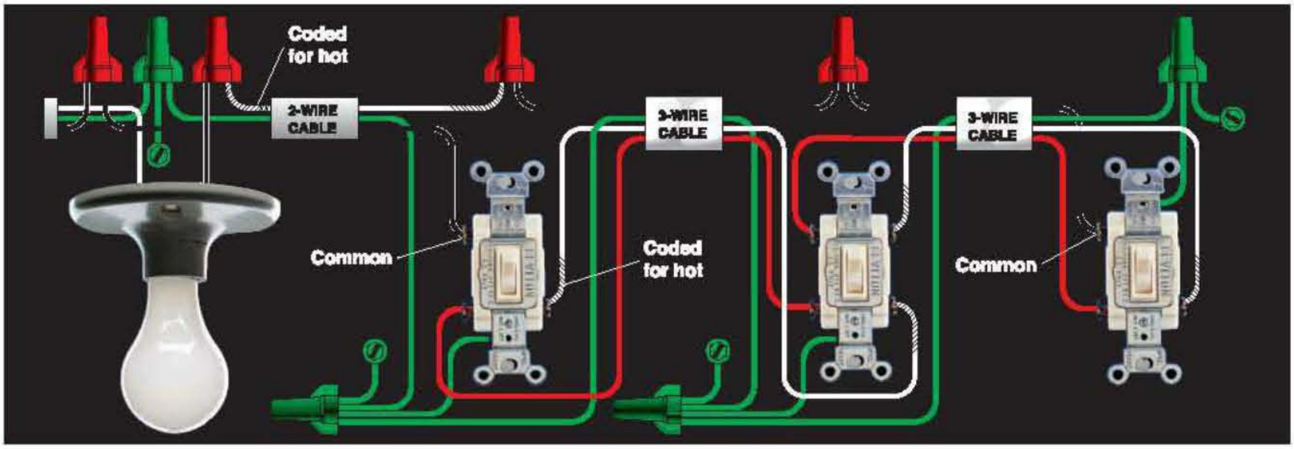

Coded for hot
2-WIRE CABLE
3-WIRE CABLE
3-WIRE CABLE
Common
Coded for hot
Common

15. FOUR-WAY SWITCH & LIGHT FIXTURE (Fixture at End of Cable Run)

Use this layout when it is more practical to locate the fixture at the end of the cable run. Requires two three-way switches and one four-way switch and two-wire and three-wire cables.

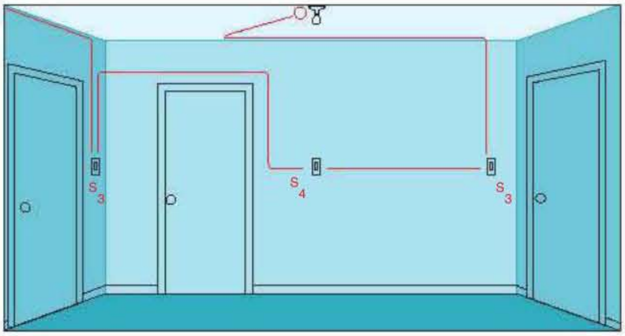

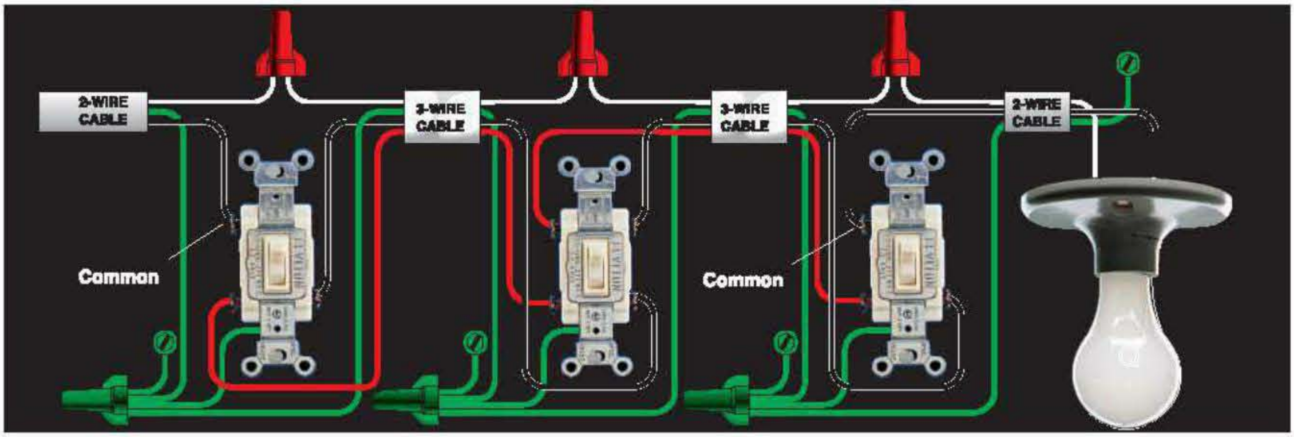

2-WIRE CABLE
3-WIRE CABLE
3-WIRE CABLE
2-WIRE CABLE
Common
Common

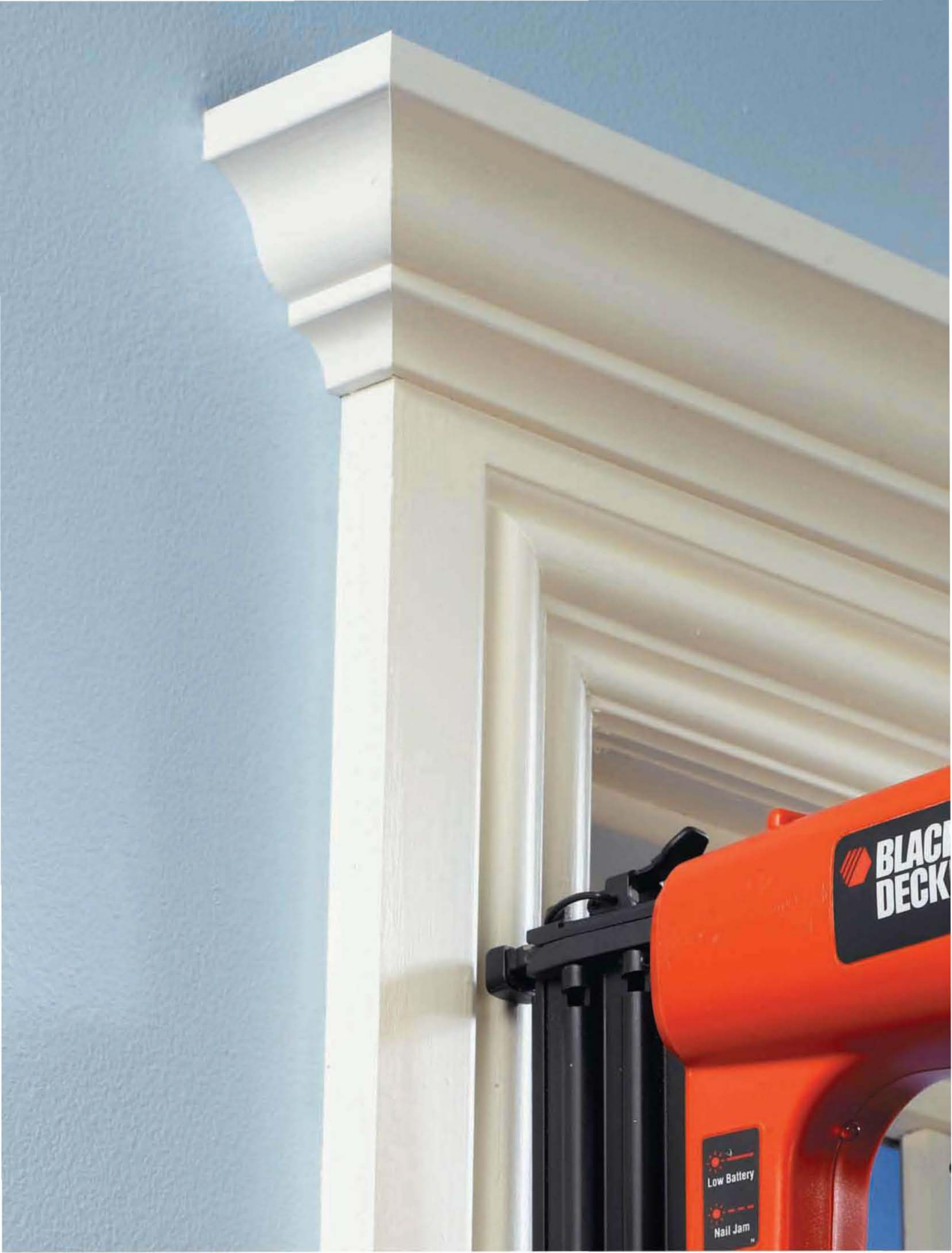
BLACK
DECK
Low Battery
Nail Jam

HOME IMPROVEMENT BASICS

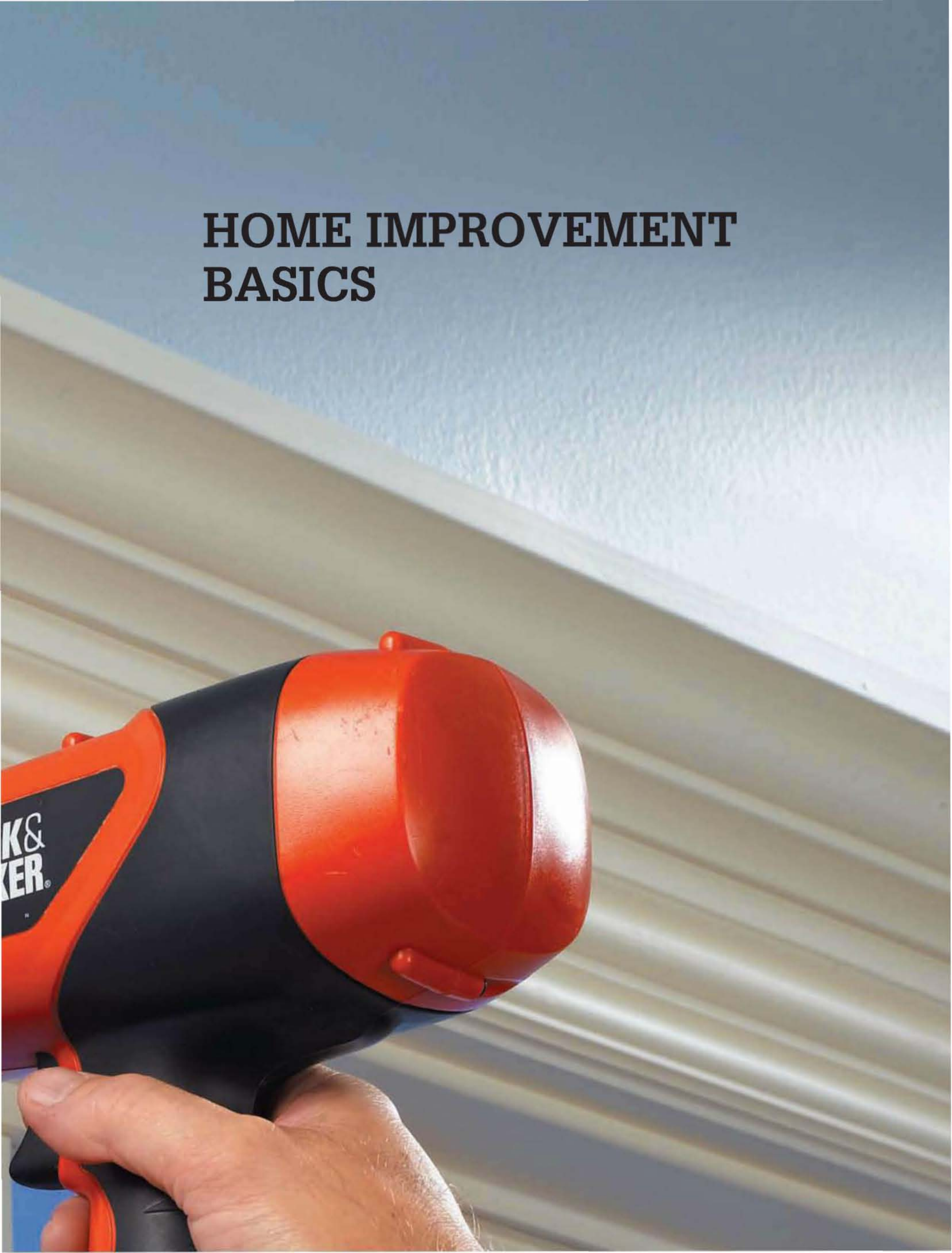

Flooring

Like any successful remodeling project, replacing your floor covering requires detailed planning and attention to design. Flooring is not separate from the rest of the room; it should fit into the overall design to create a desired effect. A floor can create excitement and become a focal point or it can serve as a background for the rest of the room.

Through careful planning, you can choose flooring that can be used successfully in multiple rooms, or select a pattern or design that is repeated throughout the room or in adjacent rooms.

Keep in mind that your flooring design will last a long time, especially if you install ceramic tile or wood. In most cases, the only way to change the design of your floor is to install a new floor covering.

In this chapter:

- Installing Underlayment
- Installing Radiant Floor Mats
- Installing Hardwood Strip Flooring
- Installing Laminate Floors
- Installing Parquet Flooring
- Installing Sheet Vinyl
- Evaluating an Existing Floor
- Installing Resilient Tile
- Installing Bamboo Flooring
- Installing Ceramic Tile
- Installing Mosaic Tile
- Installing Carpet Squares
- Refinishing Wood Floors

Installing Underlayment

Underlayment is a layer of sheeting screwed or nailed to the subfloor to provide a smooth, stable surface for the floor covering. The type of underlayment you choose depends in part on the type of floor covering you plan to install. Ceramic and natural stone tile floors usually require an underlayment that stands up to moisture, such as cementboard. For vinyl flooring, use a quality-grade plywood; most warranties are void if the flooring is installed over substandard underlayments. If you want to use your old flooring as underlayment, apply an embossing leveler to prepare it for the new installation. Most wood flooring and carpeting do not require underlayment and are often placed directly on a plywood subfloor.

When you install new underlayment, attach it securely to the subfloor in all areas, including under movable appliances. Notch the underlayment to fit the room's contours. Insert the underlayment beneath door casings and moldings. Once the underlayment is installed, use a latex patching compound to fill gaps, holes, and low spots. This compound is also used to cover screw heads, nail heads, and seams in underlayment. Some compounds include dry and wet ingredients that need to be mixed, while others are premixed. The compound is applied with a trowel or wallboard knife.

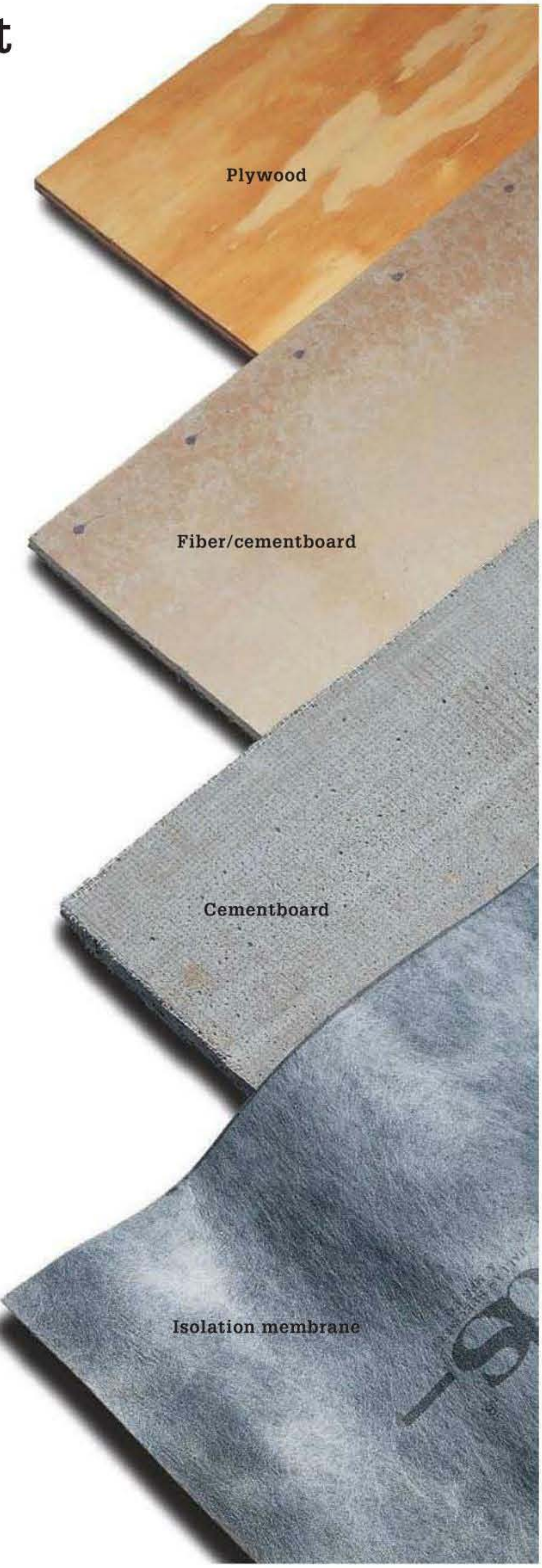

Tools & Materials

Drill
Circular saw
Wallboard knife
Power sander
¼" notched trowel
Straightedge
Utility knife
Jigsaw with carbide-tipped blade
⅛" notched trowel
Flooring roller
Underlayment
1" deck screws
Floor-patching compound
Latex additive
Thin-set mortar
1½" galvanized deck screws
Fiberglass-mesh wallboard tape

How to Install Plywood Underlayment

Plywood is the most common underlayment for vinyl flooring and some ceramic tile installations. For vinyl, use ¼" exterior-grade, AC plywood. This type has one smooth side for a quality surface. Wood-based floor coverings, like parquet, can be installed over lower-quality exterior-grade plywood. For ceramic tile, use ½" AC plywood. When installing plywood, leave ¼" expansion gaps at the walls and between sheets.

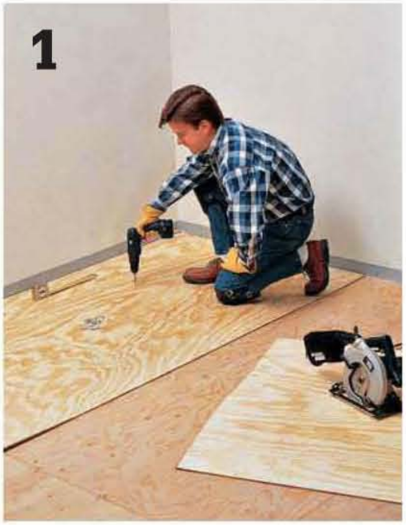

Install a full sheet of plywood along the longest wall, making sure the underlayment seams are not aligned with the subfloor seams. Fasten the plywood to the subfloor using 1" deck screws driven every 6" along the edges and at 8" intervals in the field of the sheet.

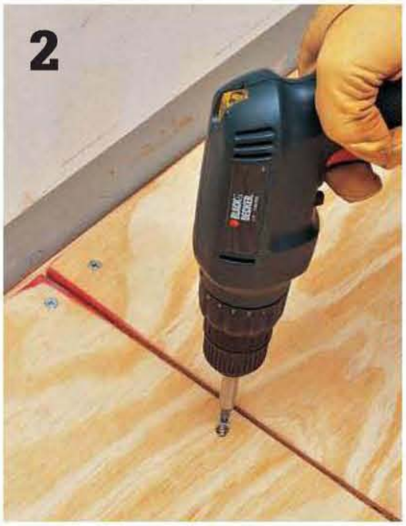

Continue fastening sheets of plywood to the subfloor, driving the screw heads slightly below the underlayment surface. Leave ¼" expansion gaps at the walls and between sheets. Offset seams in subsequent rows.

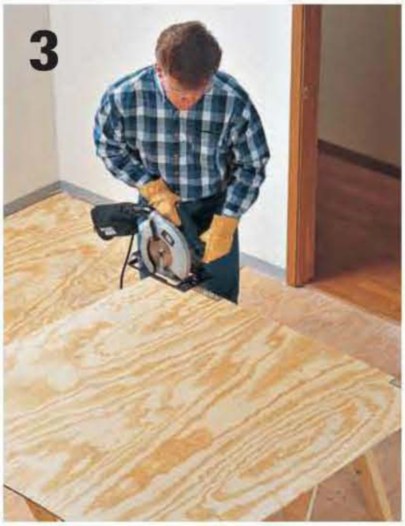

Using a circular saw or jigsaw, notch the plywood to meet the existing flooring in doorways. Fasten the notched sheets to the subfloor.

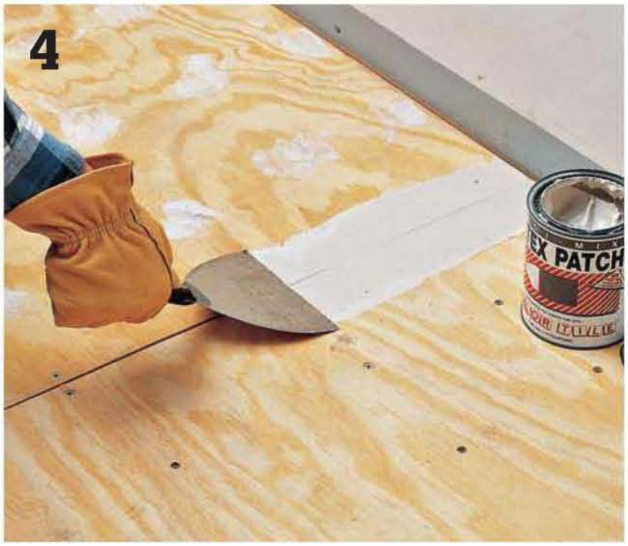

Mix floor-patching compound and latex or acrylic additive following the manufacturer's directions. Spread it over seams and screw heads, using a wallboard knife.

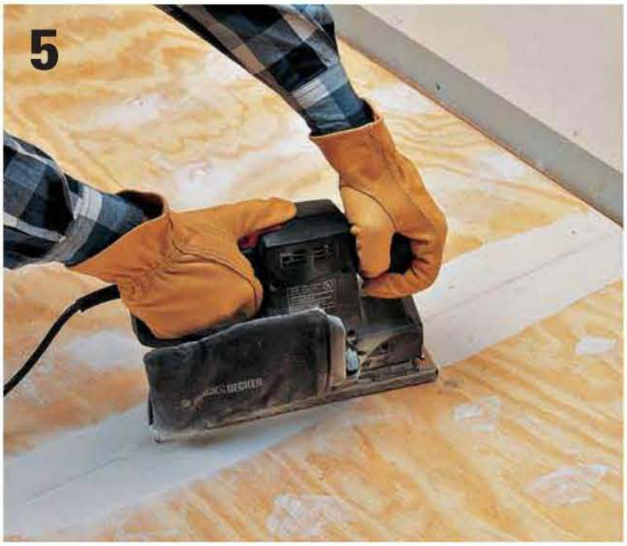

Let the patching compound dry, then sand the patched areas, using a power sander.

How to Install Cementboard

Ceramic and natural stone tile floors usually require an underlayment that stands up to moisture, such as cementboard. Fiber/cementboard is a thin, high-density underlayment used under ceramic tile and vinyl flooring in situations where floor height is a concern. Cementboard is used only for ceramic tile or stone tile installations. It remains stable even when wet, so it is the best underlayment to use in areas that are likely to get wet, such as bathrooms. Cementboard is more expensive than plywood, but a good investment for a large tile installation.

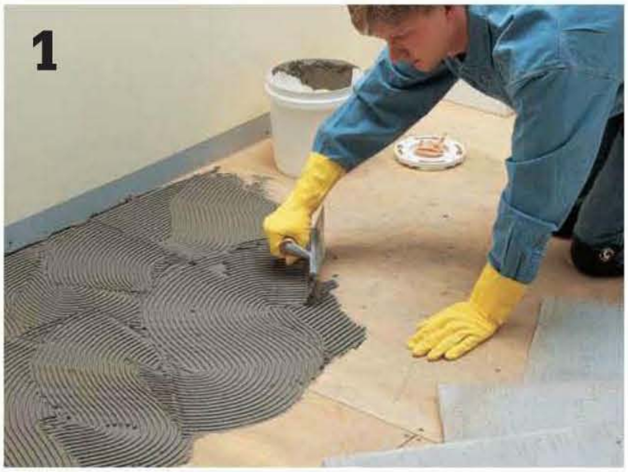

Mix thin-set mortar according to the manufacturer's directions. Starting at the longest wall, spread the mortar on the subfloor in a figure eight pattern using a ¼" notched trowel. Spread only enough mortar for one sheet at a time. Set the cementboard on the mortar with the rough side up, making sure the edges are offset from the subfloor seams.

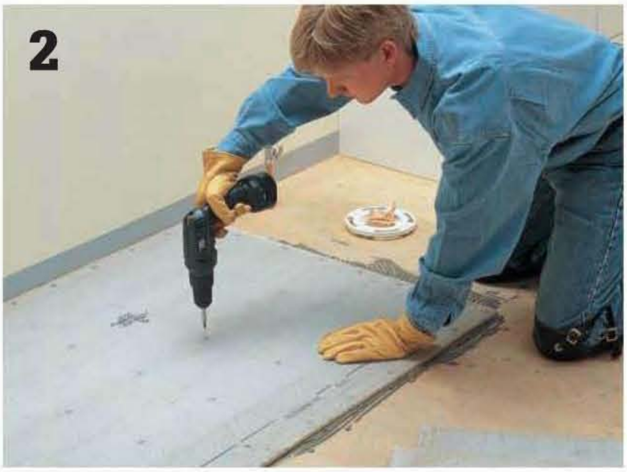

Fasten the cementboard to the subfloor using 1¼" cementboard screws driven every 6" along the edges and 8" throughout the sheet. Drive the screw heads flush with the surface. Continue spreading mortar and installing sheets along the wall. *Option: If installing fiber/cementboard underlayment, use a 3⁄16" notched trowel to spread the mortar, and drill pilot holes for all screws.*

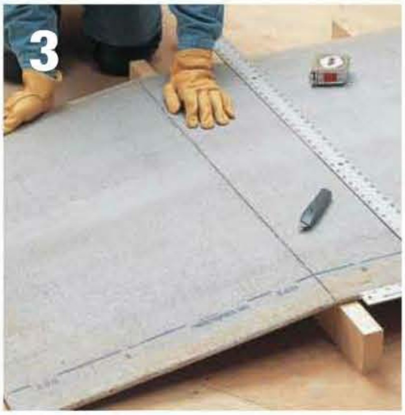

Cut cementboard pieces as necessary, leaving an ⅛" gap at all joints and a ¼" gap along the room perimeter. For straight cuts, use a utility knife to score a line through the fiber-mesh layer just beneath the surface, then snap the board along the scored line.

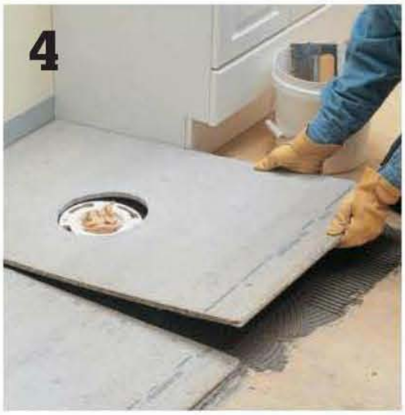

To cut holes, notches, or irregular shapes, use a jigsaw with a carbide-tipped blade. Continue installing cementboard sheets to cover the entire floor.

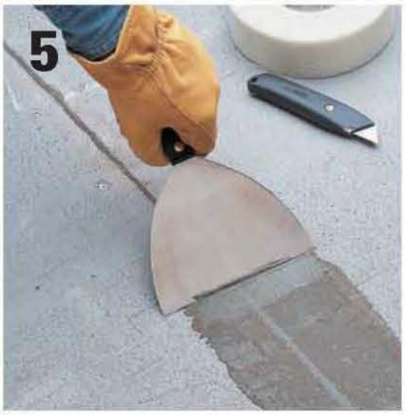

Place fiberglass-mesh wallboard tape over the seams. Use a wallboard knife to apply thin-set mortar to the seams, filling the gaps between sheets and spreading a thin layer of mortar over the tape. Allow the mortar to set for two days before starting the tile installation.

How to Install Isolation Membrane

Isolation membrane is used to protect ceramic tile installations from movement that may occur on cracked concrete floors. This product is used primarily for covering individual cracks, but it can be used over an entire floor. Isolation membrane is also available in a liquid form that can be poured over the project area.

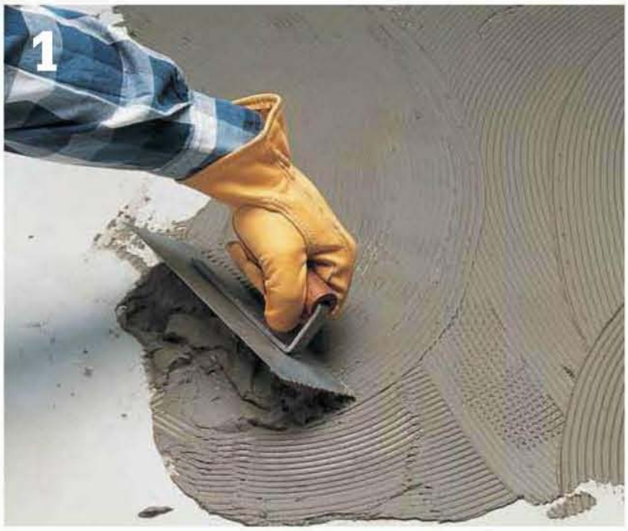

Thoroughly clean the subfloor, then apply thin-set mortar with a ⅛" notched trowel. Start spreading the mortar along a wall in a section as wide as the membrane and 8 to 10 ft. long. *Note: For some membranes, you must use a bonding material other than mortar. Read and follow manufacturer's directions.*

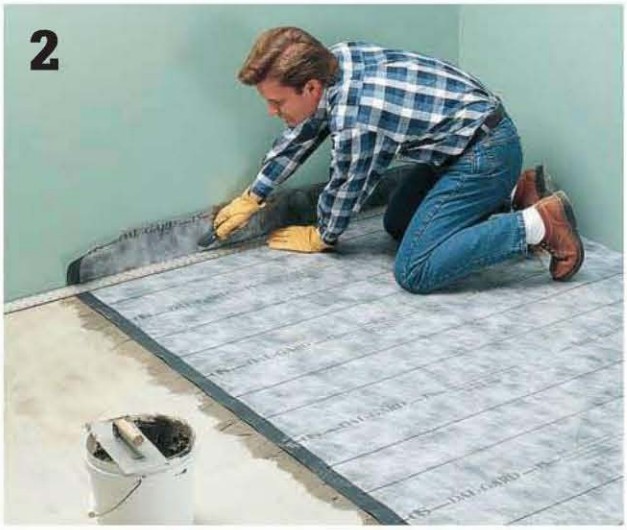

Roll out the membrane over the mortar. Cut the membrane to fit tightly against the walls, using a straightedge and utility knife.

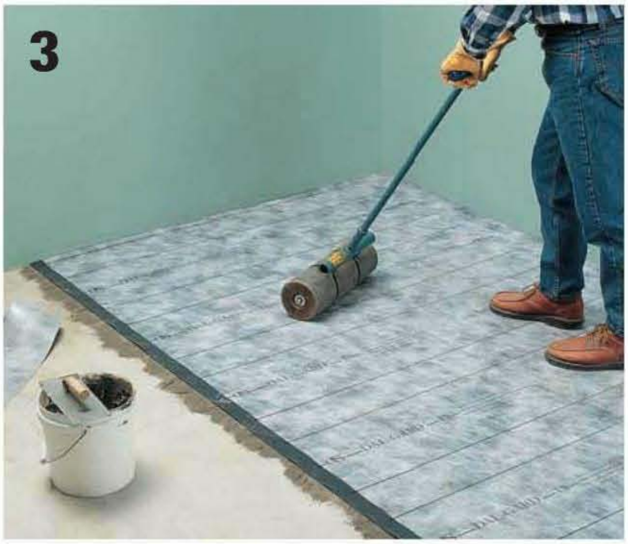

Starting in the center of the membrane, use a heavy floor roller to smooth out the surface toward the edges. This frees trapped air and presses out excess bonding material.

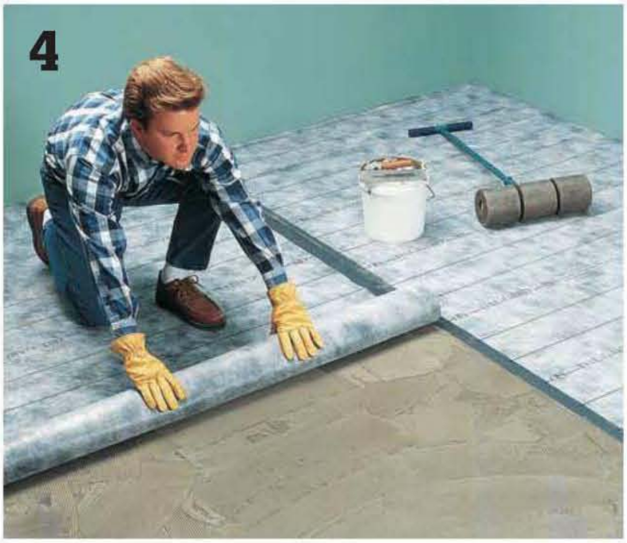

Repeat steps 1 through 3, cutting the membrane as necessary at the walls and obstacles, until the floor is completely covered with membrane. Do not overlap the seams, but make sure they're tight. Allow the mortar to set for two days before installing the tile.

Installing Raised Underlayment Panels

Concrete floors are practical and durable—and generally cold and uncomfortable. For a fast and easy makeover, you can now find raised underlayment panels that simply rest on the concrete and provide a surface for other flooring materials. The tongue-and-groove plywood panels have dimpled plastic on the bottom. This allows air to circulate underneath so that the concrete stays dry, and insulates the flooring above. The assembled panels can support laminates and resilient sheets or tiles. And you can install them in a weekend.

How to Install Raised Underlayment Panels

Start at one corner, and measure the length and width of the room from that starting point. Calculate the number of panels you will need to cover the space in both directions. If the starting corner is not square, trim the first row of panels to create a straight starting line.

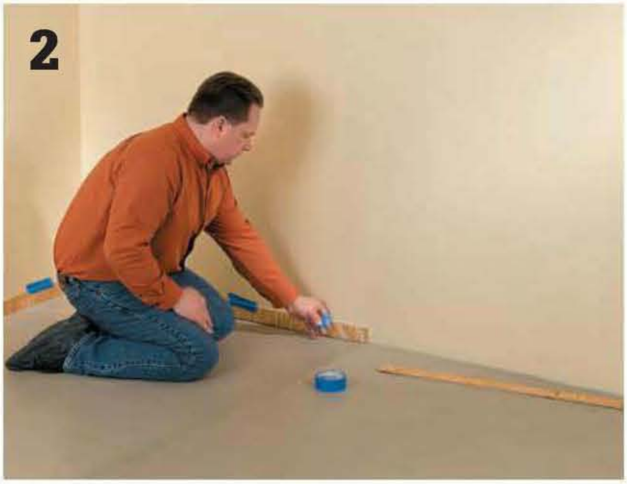

Create an expansion gap around the edges. Place ¼" spacers at all walls, doors, and other large obstacles. To make your own spacers, cut sheets of ¼" plywood to the thickness of the panels, and hold them in place temporarily with masking tape.

Dry-lay a row of panels across the room. If the last row will be less than 6" wide, balance it by trimming the first panel or the starting row, if necessary, to account for the row end pieces.

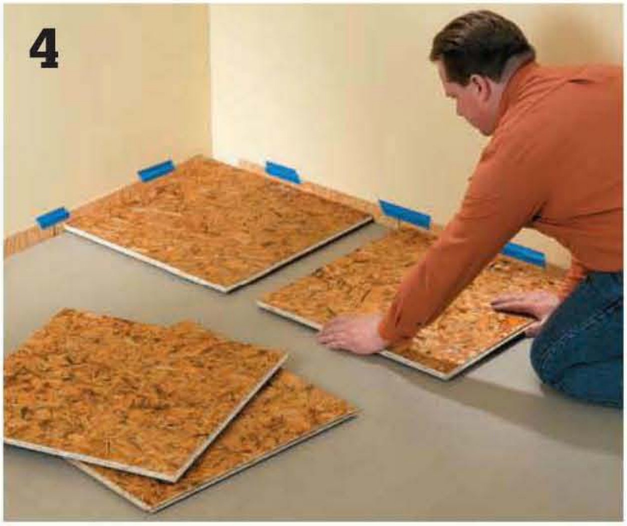

Starting in the corner, lay the first panel with the grooved side against the ¼" spacers. Slide the next panel into place and press-fit the groove of the second panel into the tongue of the first. Check the edges against the wall.

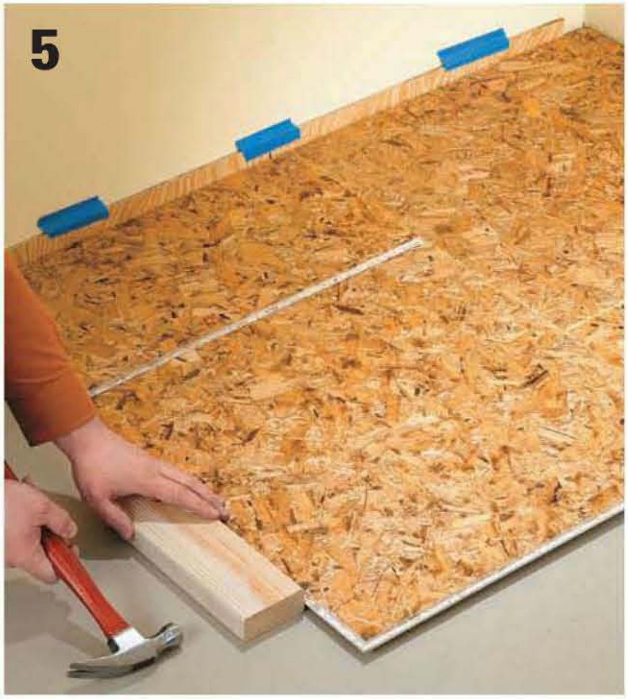

Repeat these steps to complete the first row. If necessary, tap the panels into place with a scrap piece of lumber and a rubber mallet or hammer—just be careful not to damage the tongue or groove edges. Starting with the second row, stagger the seams so that the panels interlock.

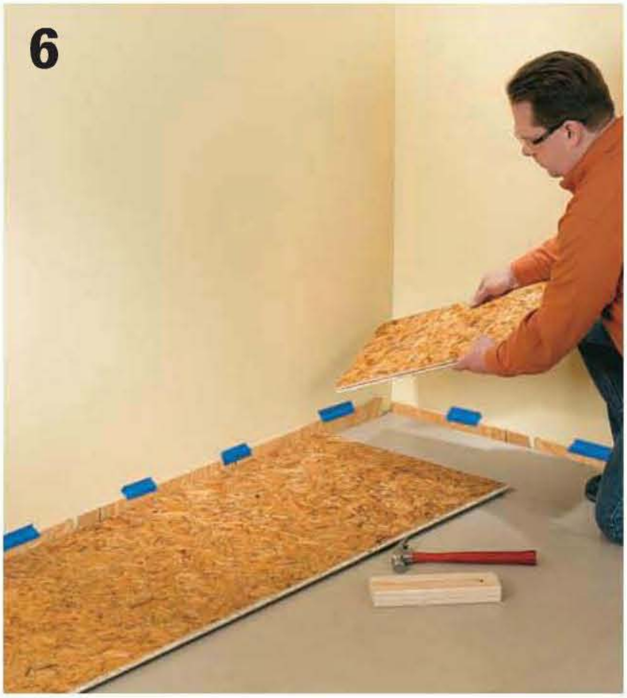

Cut the last panel to fit snugly between the next-to-last panel and the ¼" spacer on the far wall. Install the last panel at an angle and tap it down. Continue working from the starting point, checking after each row to be sure the panels are square and level.

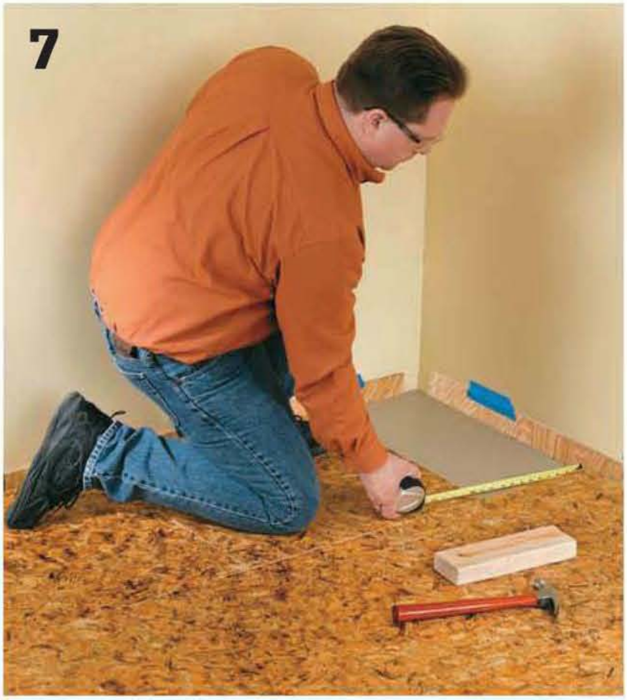

When you reach the last row and last panel to complete your installation, you may have to cut the panel to fit. Measure for fit, allowing for the ¼" expansion gap from the wall. Cut the panel and fit it into place.

When all the panels are in place and the finished floor is installed, remove the spacers from around the perimeter of the room.

Installing Radiant Floor Mats

Floor-warming systems require very little energy to run and are designed to heat floors only; they generally are not used as sole heat sources for rooms.

A typical floor-warming system consists of one or more thin mats containing electric resistance wires that heat up when energized like an electric blanket. The mats are installed beneath the tile and are hardwired to a 120-volt GFCI circuit. A thermostat controls the temperature, and a timer turns the system off automatically.

The system shown in this project includes two plastic mesh mats, each with its own power lead that is wired directly to the thermostat. Radiant mats may be installed over a plywood subfloor, but if you plan to install floor tile you should put down a base of cementboard first, and then install the mats on top of the cementboard.

A crucial part of installing this system is to use a multimeter to perform several resistance checks to make sure the heating wires have not been damaged during shipping or installation.

Electrical service required for a floor-warming system is based on size. A smaller system may connect to an existing GFCI circuit, but a larger one will need a dedicated circuit; follow the manufacturer's requirements.

To order a floor-warming system, contact the manufacturer or dealer (see Resources, page 554). In most cases, you can send them plans and they'll custom-fit a system for your project area.

Tools & Materials ▸

Vacuum cleaner
Multimeter
Tape measure
Scissors
Router/rotary tool
Marker
Trowel
  or rubber float
Notched trowel
Staple gun
Hot glue gun
Electric wire fault
  indicator (optional)
Radiant floor mats
12/2 NM cable
Conduit
Wire connectors
Thinset mortar
Thermostat with sensor
Junction box(es)
Tile or stone
  floorcovering

A radiant floor-warming system employs electric heating mats that are covered with floor tile to create a floor that's cozy under foot.

Installation Tips ▸

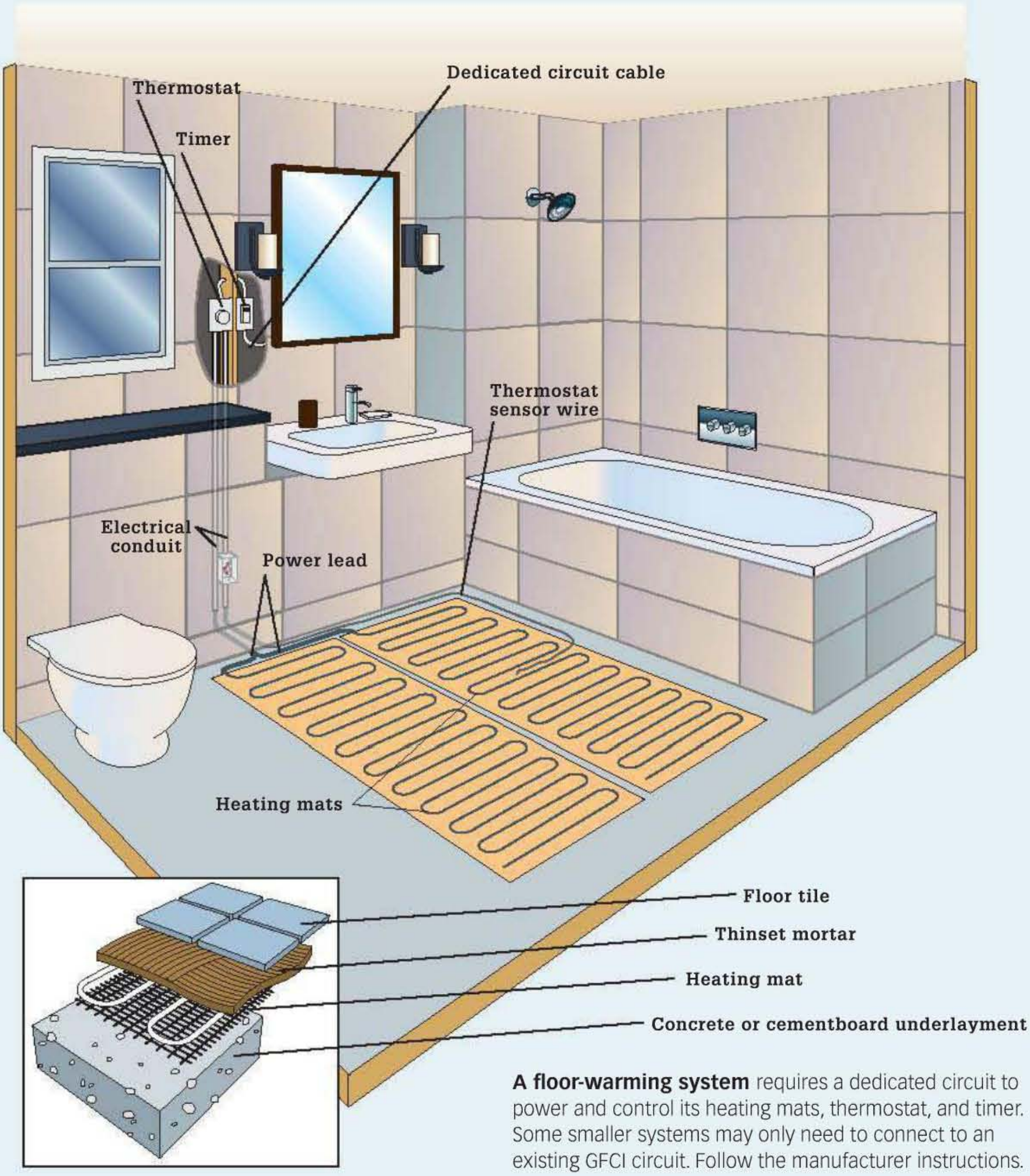
Thermostat
Timer
Dedicated circuit cable
Thermostat sensor wire
Electrical conduit
Power lead
Heating mats
Floor tile
Thinset mortar
Heating mat
Concrete or cementboard underlayment

A floor-warming system requires a dedicated circuit to power and control its heating mats, thermostat, and timer. Some smaller systems may only need to connect to an existing GFCI circuit. Follow the manufacturer instructions.

- Each radiant mat must have a direct connection to the power lead from the thermostat, with the connection made in a junction box in the wall cavity. Do not install mats in series.
- Do not install radiant floor mats under shower areas.
- Do not overlap maps or let them touch.
- Do not cut heating wire or damage heating wire insulation.
- The distance between wires in adjoining mats should equal the distance between wire loops measured center to center.

Installing a Radiant Floor-warming System

10,000 AIC
OFF
OFF
ON
PUSH TO TEST

Floor-warming systems must be installed on a circuit with adequate amperage and a GFCI breaker. Smaller systems may tie into an existing circuit, but larger ones need a dedicated circuit. Follow local building and electrical codes that apply to your project.

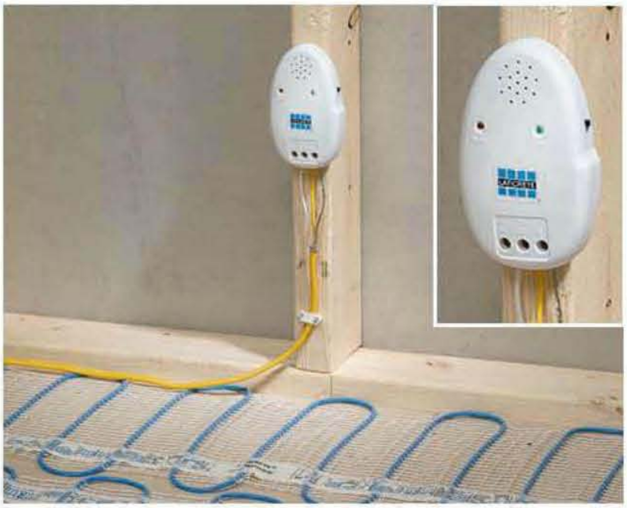

An electric wire fault indicator monitors each floor mat for continuity during the installation process. If there is a break in continuity (for example, if a wire is cut) an alarm sounds. If you choose not to use an installation tool to monitor the mat, test for continuity frequently using a multimeter.

How To Install a Radiant Floor-warming System

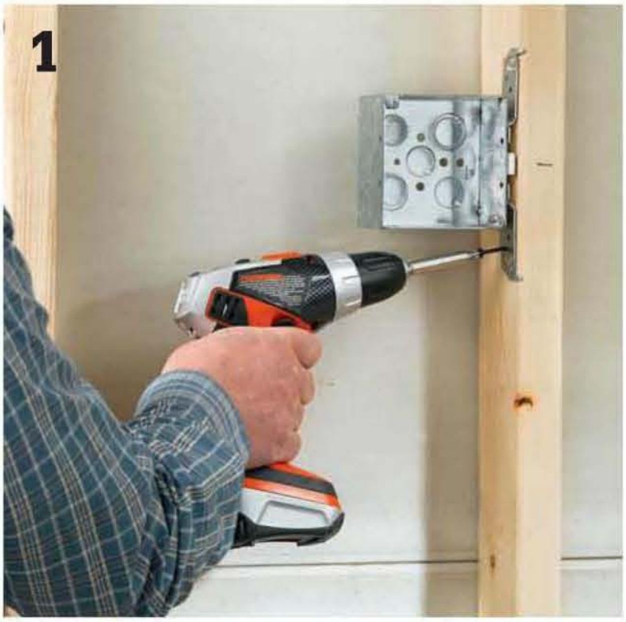

Install electrical boxes to house the thermostat and timer. In most cases, the box should be located 60" above floor level. Use a 4"-deep × 4"-wide double-gang box for the thermostat/timer control if your kit has an integral model. If your timer and thermostat are separate, install a separate single box for the timer.

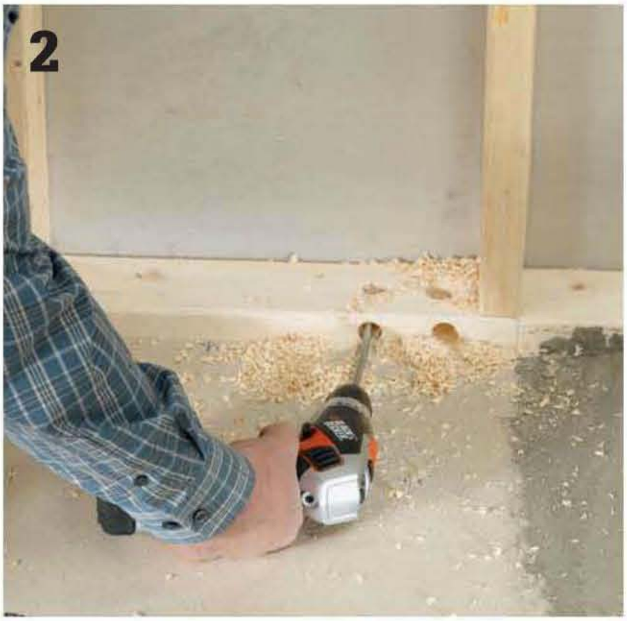

Drill access holes in the sole plate for the power leads that are preattached to the mats (they should be over 10 ft. long). The leads should be connected to a supply wire from the thermostat in a junction box located in a wall near the floor and below the thermostat box. The access hole for each mat should be located directly beneath the knockout for that cable in the thermostat box. Drill through the sill plate vertically and horizontally so the holes meet in an L-shape.

3

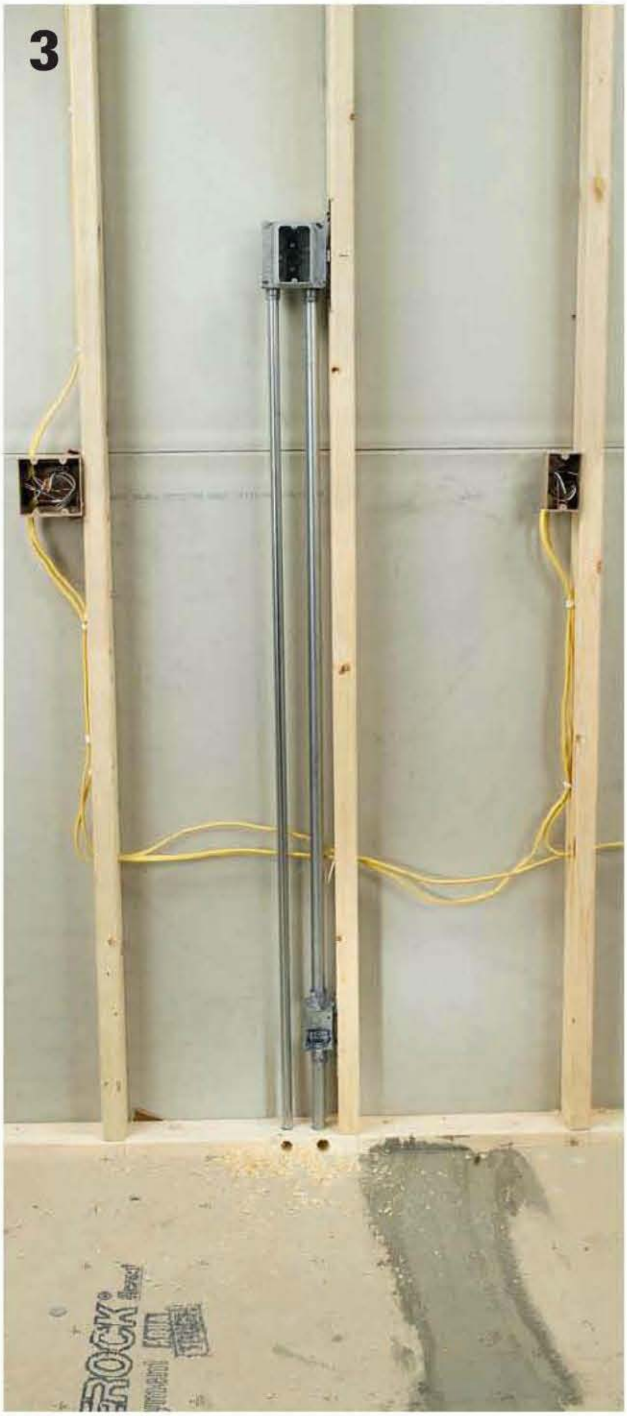

Run conduit from the electrical boxes to the sill plate. The line for the supply cable should be ¾" conduit. If you are installing multiple mats, the supply conduit should feed into a junction box about 6" above the sill plate and then continue into the ¾" hole you drilled for the supply leads. The sensor wire needs only ½" conduit that runs straight from the thermostat box via the thermostat. The mats should be powered by a dedicated 20-amp GFCI circuit of 12/2 NM cable run from your main service panel to the electrical box (this is for 120-volt mats—check your instruction manual for specific circuit recommendations).

4

Clean the floor surface thoroughly to get rid of any debris that could potentially damage the wire mats. A vacuum cleaner generally does a more effective job than a broom.

5

Test for resistance using a multimeter set to measure ohms. This is a test you should make frequently during the installation, along with checking for continuity. If the resistance is off by more than 10% from the theoretical resistance listing (see manufacturer's chart in installation instructions), contact a technical support operator for the kit manufacturer. For example, the theoretical resistance for the 1 × 50 ft. mat seen here is 19, so the ohms reading should be between 17 and 21.

6

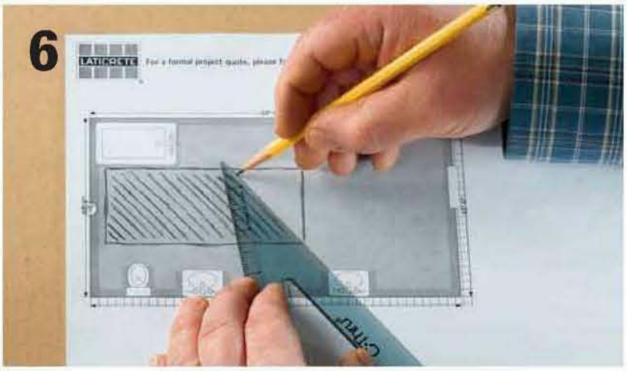

Finalize your mat layout plan. Most radiant floor warming mat manufacturers will provide a layout plan for you at the time of purchase, or they will give you access to an online design tool so you can come up with your own plan. This is an important step to the success of your project, and the assistance is free.

(continued)

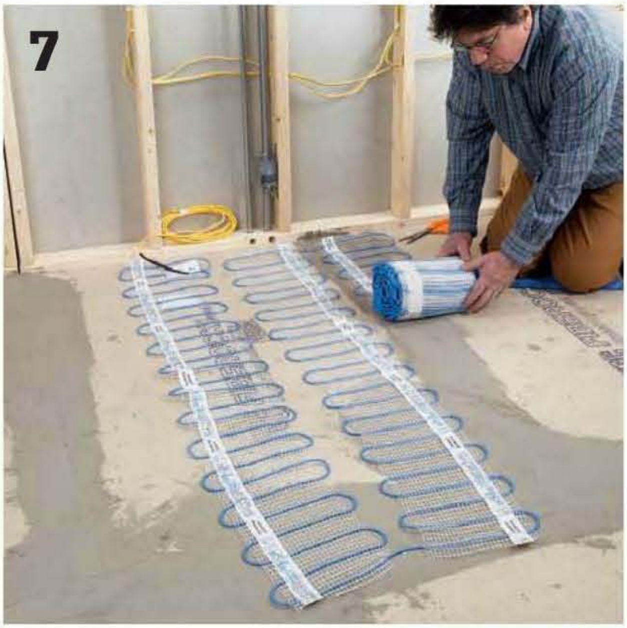

Unroll the radiant mat or mats and allow them to settle. Arrange the mat or mats according to the plan you created. It's okay to cut the plastic mesh so you can make curves or switchbacks, but do not cut the heating wire under any circumstances, not even to shorten it.

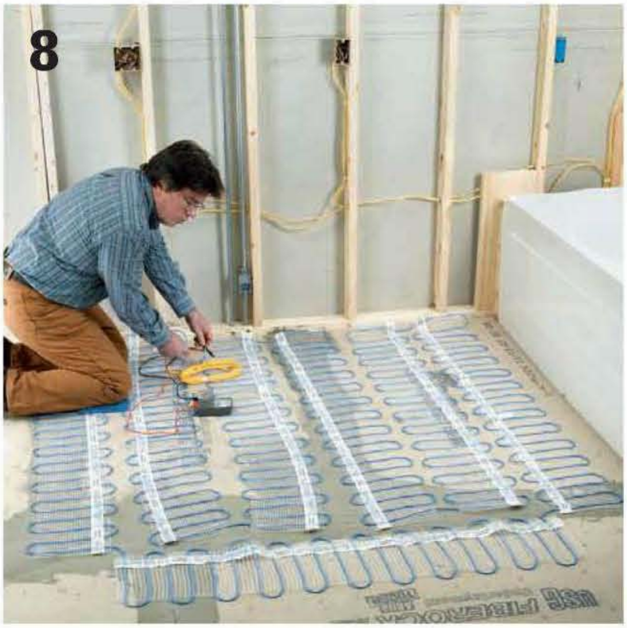

Finalize the mat layout and then test the resistance again using a multimeter. Also check for continuity in several different spots. If there is a problem with any of the mats, you should identify it and correct it before proceeding with the mortar installation.

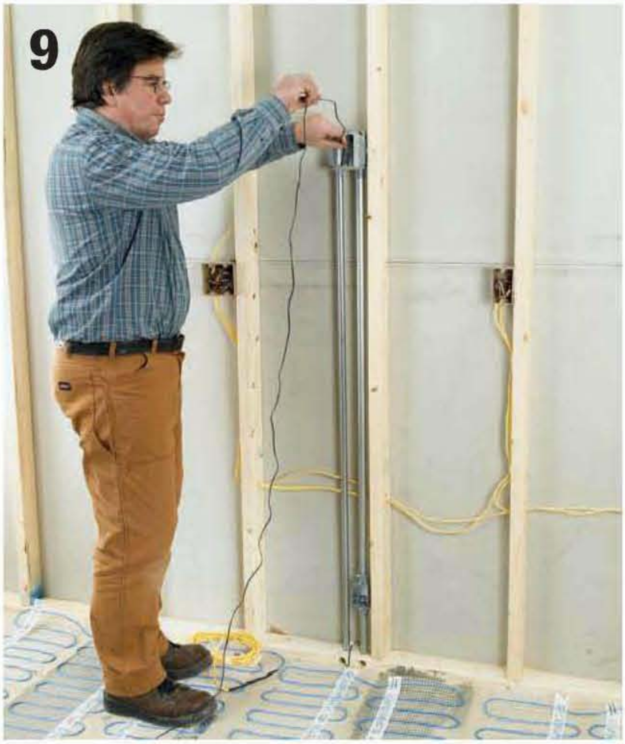

Run the thermostat sensor wire from the electrical box down the ½" conduit raceway and out the access hole in the sill plate. Select the best location for the thermostat sensor and mark the location onto the flooring. Also mark the locations of the wires that connect to and lead from the sensor.

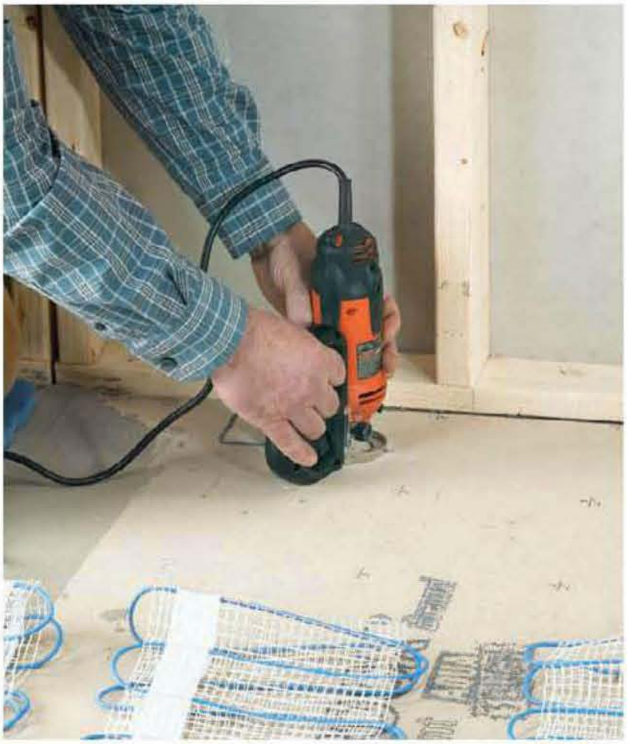

Variation: If your local codes require it, roll the mats out of the way and cut a channel for the sensor and the sensor wires into the floor or floor underlayment. For most floor materials, a spiral cutting tool does a quick and neat job of this task. Remove any debris.

10

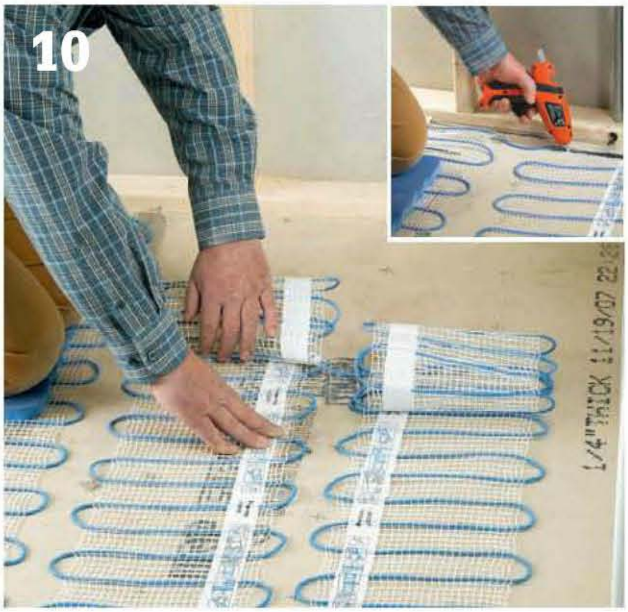

Bond the mats to the floor. If the mats in your system have adhesive strips, peel off the adhesive backing and roll out the mats in the correct position, pressing them against the floor to set the adhesive. If your mats have no adhesive, bind them with strips of double-sided carpet tape. The thermostat sensor and the power supply leads should be attached with hot glue (inset photo) and run up into their respective holes in the sill plate if you have not done this already. Test all mats for resistance and continuity.

11

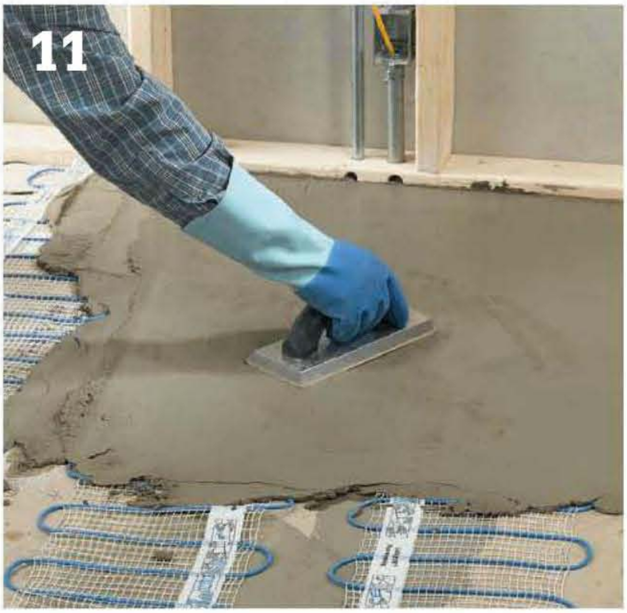

Cover the floor installation areas with a layer of thinset mortar that is thick enough to fully encapsulate all the wires and mats (usually around ¼" in thickness). Check the wires for continuity and resistance regularly and stop working immediately if there is a drop in resistance or a failure of continuity. Allow the mortar to dry overnight.

12

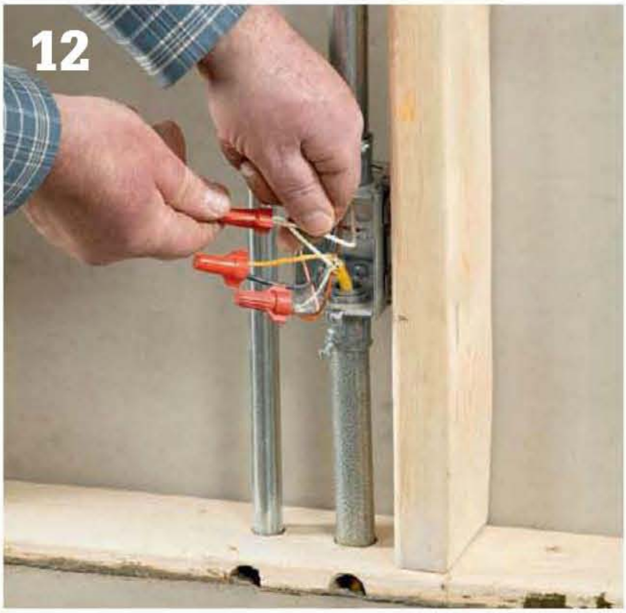

Connect the power supply leads from the mat or mats to the NM cable coming from the thermostat inside the junction box near the sill. Power must be turned off. The power leads should be cut so about 8" of wire feeds into the box. Be sure to use cable clamps to protect the wires.

13

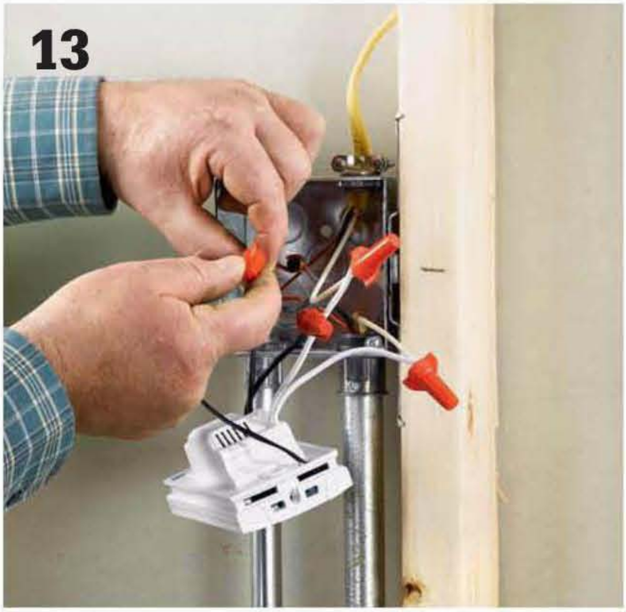

Connect the sensor wire and the power supply lead (from the junction box) to the thermostat/timer according to the manufacturer's directions. Attach the device to the electrical box, restore power, and test the system to make sure it works. Once you are convinced that it is operating properly, install flooring and repair the wall surfaces.

Installing Hardwood Strip Flooring

Tongue-and-groove hardwood flooring has always been popular with homeowners. It offers an attractive look, is one of the longest lasting floor coverings, and can be stripped and refinished to look like new.

Oak has been the most common type of strip flooring because of its durability and wood graining, and it's the species most people think of when hardwood is mentioned. Other woods, such as maple, cherry, and birch, are also becoming popular.

Exotic species of wood from around the world are now finding their way into American homes as people want a premium strip or plank floor that is unique and stylish, and expresses their personalities. The more than sixty exotic hardwoods include Brazilian cherry, Australian cypress, Honduran mahogany, tobaccowood, teak, zebrawood, and bamboo—which is not really wood but a type of grass.

This section describes how to install nailed-down tongue-and-groove flooring, how to install a decorative medallion, and how to install tongue-and-groove strip flooring over troweled-on adhesive. Customizing your floor with borders, accents, and medallions is easier than you may think. A number of manufacturers produce a variety of decorative options made to match the thickness of your floor.

Tools & Materials ▸

- Rosin paper
- Utility knife
- Chalk line
- 8d finish nails
- Drill
- Staple
- Nail set
- Hammer
- Pry bar
- Wood mallet
- Power nailer
- Pull bar

Real hardwood flooring has depth of beauty and warmth underfoot that even the highest quality laminate imitations struggle to match.

How to Install Tongue-and-Groove Hardwood Flooring

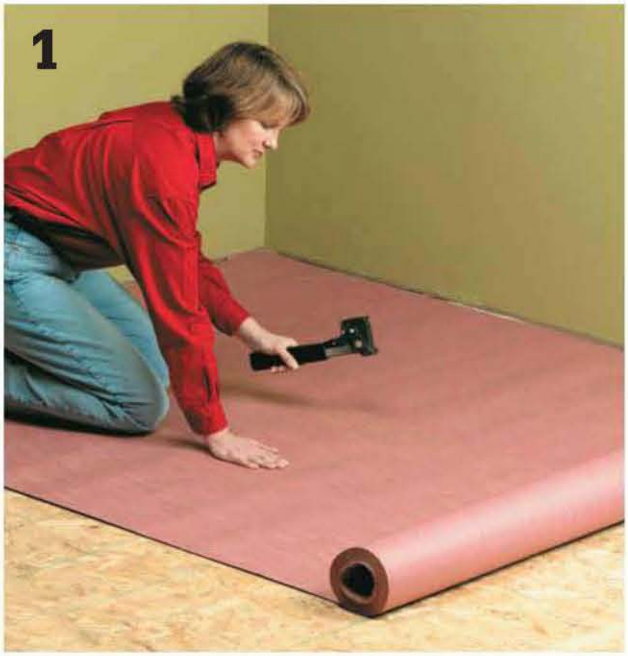

Cover the entire subfloor with rosin paper. Staple the paper to the subfloor, overlapping edges by 4". Cut the paper with a utility knife to butt against the walls.

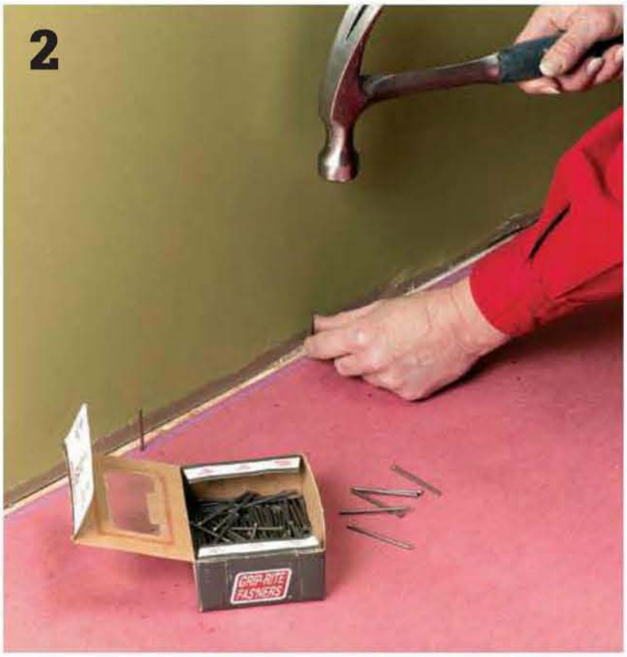

Make a mark on the floor ½" from the starter wall at both ends of the wall. Snap a chalk line between the marks. Nail 8d finish nails every 2" to 3" along the chalk line to mark the location for your first row.

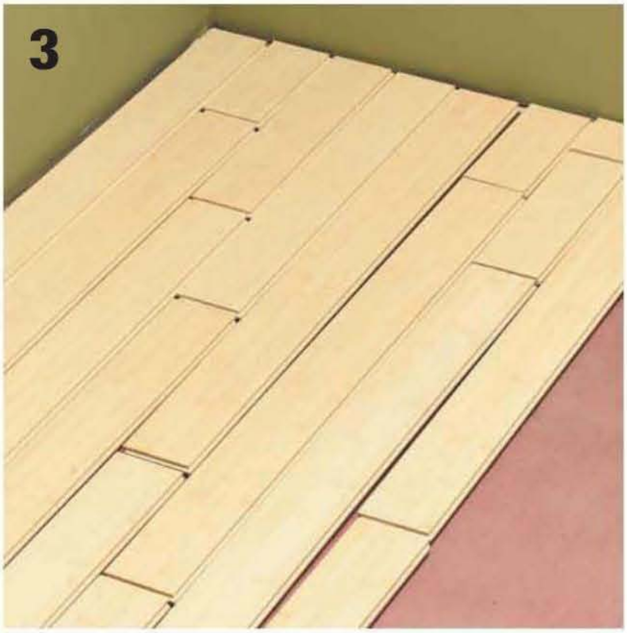

Lay out the first 8 rows of flooring in a dry run with the groove side facing the wall. Make sure the first row of boards is straight. Arrange the boards to get a good color and grain mix. Offset the ends by at least 6".

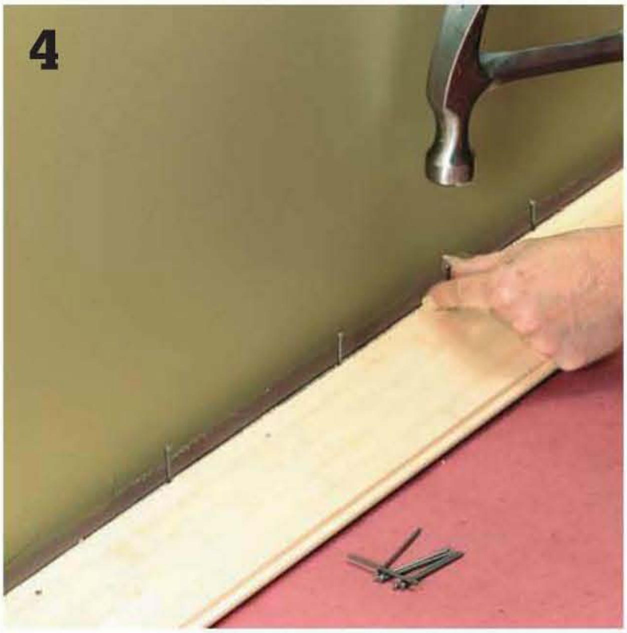

Place the starter row against the nails on the chalk line. Drill pilot holes in the flooring every 6" to 8", about ½" from the groove edge. Face nail the first row until the nail heads are just above the boards, then sink them using a nail set. (Be careful not to hit the boards with your hammer or you'll mar the surface.)

(continued)

5

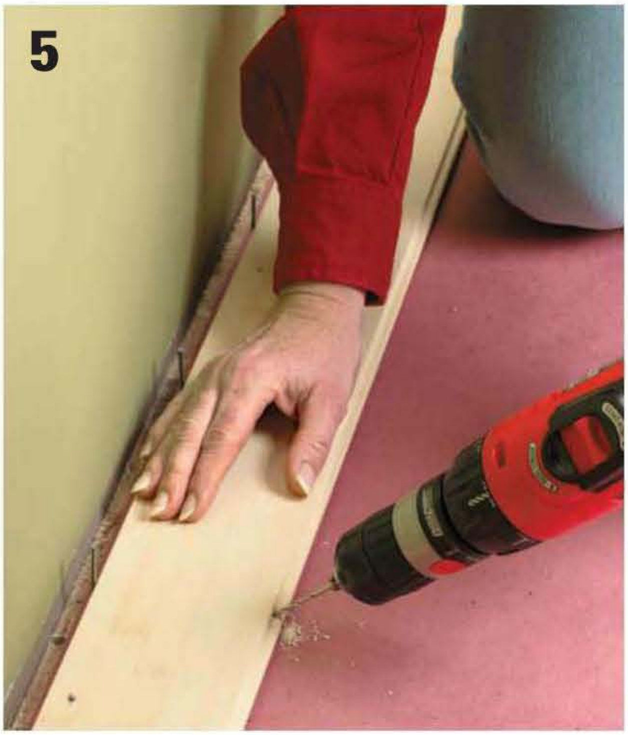

Drill pilot holes every 6" to 8" directly above the tongue, keeping the drill at a 45° angle.

Tip ▸

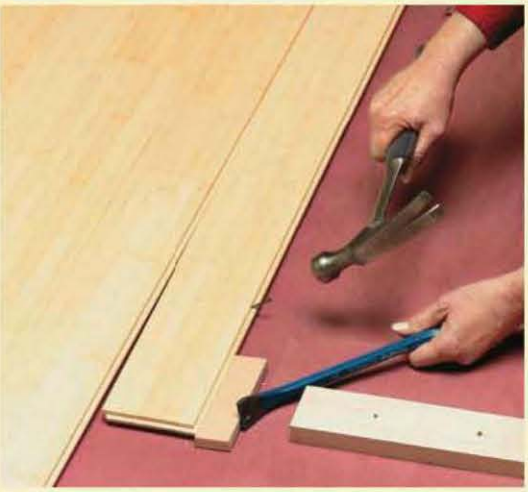

To install crooked boards, drill pilot holes above the tongue and insert nails. Fasten a scrap board to the subfloor using screws. Force the floor board straight using a pry bar and a scrap board placed in front of the flooring. With pressure on the floor board, blind nail it into place.

6

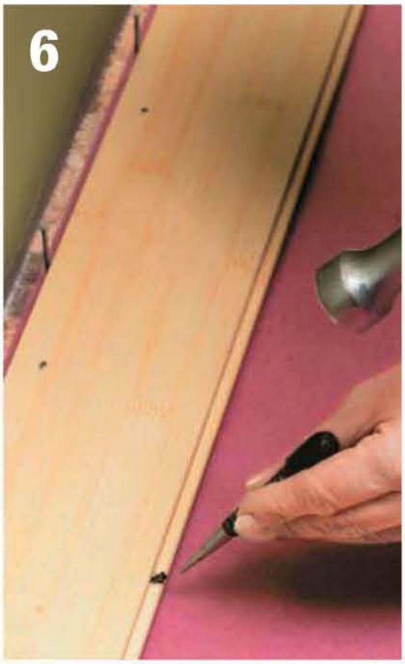

Blind nail a nail into each pilot hole. Keep the nail heads ½" out, then set them just below the surface, using a nail set.

7

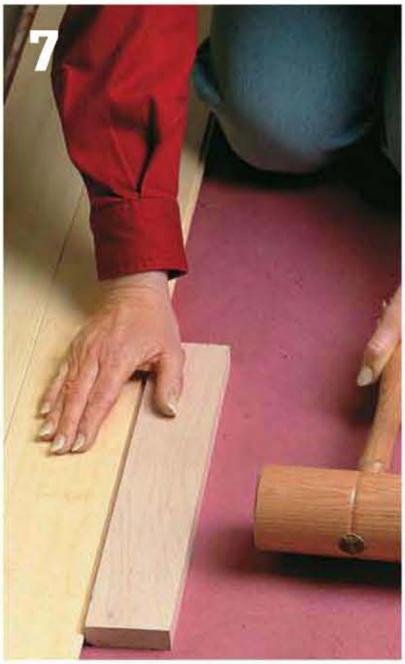

Set the second row of boards in place against the starter row, fitting together the tongue and groove connections. Use a scrap board and wood mallet to tap the floor boards together. Drill pilot holes and blind nail the boards. Do this for the next few rows.

8

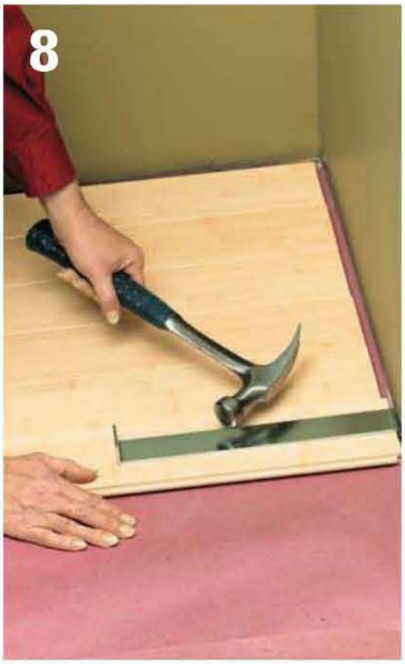

To install the last board in a row, place the tongue and groove joints together, then place a flooring pull bar over the end of the board. Hit the end of the pull bar with a hammer until the board slides into place. Stay ½" away from the walls.

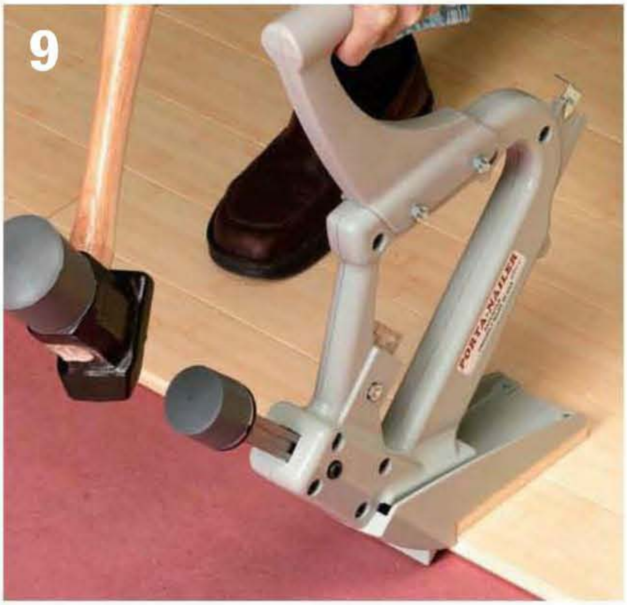

Once you have several rows installed and have enough room, use a power nailer. Place the nailer lip over the edge of the board and strike it with a mallet. Drive a nail 2" from the end of each board and about every 8" in the field. Keep a few rows of flooring laid out ahead of you as you work, and keep the joints staggered.

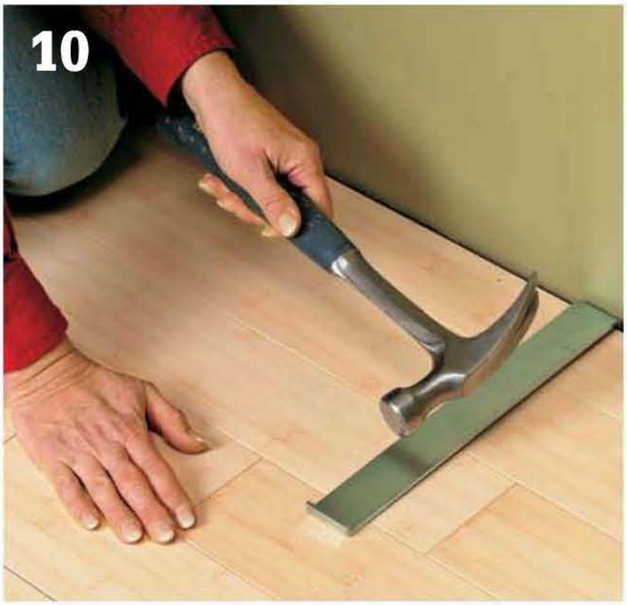

When you're out of room for the power nailer, drill pilot holes and blind nail the boards. For the last rows, drill pilot holes in the top of the boards, ½" from the tongue, and face nail them. The last row may need to be ripped to size. Pull the last row into place using the flooring pull bar, leaving a ½" gap along the wall. Drill pilot holes and face nail.

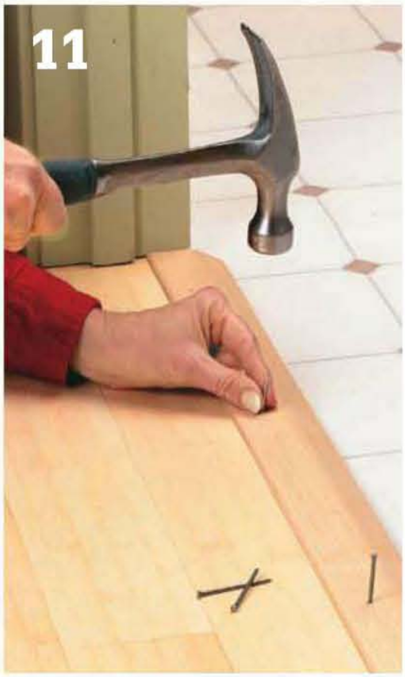

Install a reducer strip or transition strip between the wood floor and an adjoining room. Cut the strip to size. Fit the strip's groove over the floor board's tongue, then drill pilot holes and face nail. Set the nails with a nail set. Fill all visible nail holes with wood putty.

Tip ▸

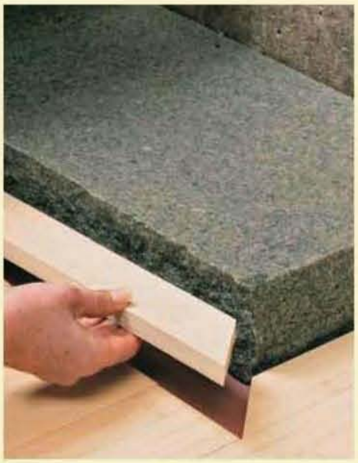

To install around an object, cut a notch in the board. For larger obstacles, cut 45° miters in boards so grooves face away from the object. Rip tongues off the boards. Set the boards against the object and flooring, fitting mitered ends together. Drill pilot holes and face nail in place. Apply silicone caulk between the floor and board edge.

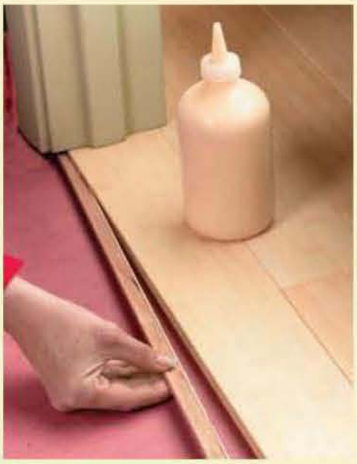

To reverse directions of the tongue and groove for doorways, glue a spline into the groove of the board. Fit the groove of the next board onto the spline, then nail the board in place.

Installing Laminate Floors

Laminate flooring comes in a floating system that is simple to install, even if you have no experience with other home-improvement projects. You may install a floating laminate floor right on top of plywood, concrete slab, sheet vinyl, or hardwood flooring. Just be sure to follow the manufacturer's instructions.

The pieces are available in planks or squares in a variety of different sizes, colors, and faux finishes—including wood and ceramic. The part you see is really a photographic print. Tongue-and-groove edges lock pieces together, and the entire floor floats on the underlayment. At the end of this project there are a few extra steps to take if your flooring manufacturer recommends using glue on the joints.

The rich wood tones of beautiful laminate planks may cause you to imagine hours of long, hard installation work, but this is a DIY project that you can do in a single weekend. Buy the manufactured planks at a home-improvement or flooring store and install laminate flooring with the step-by-step instructions offered in the following pages.

Tools & Materials

- Circular saw
- Underlayment
- ½" spacers
- Tapping block
- Scrap foam
- Speed square
- Manufacturer glue
- Painter's tape
- Chisel
- Rubber mallet
- Drawbar
- Finish nails
- Nail set strap clamps
- Threshold and screws

Laminate strip floors install quickly, wear well and are among the cheapest floor coverings you can find. Overall they have improved in appearance but the lower quality products continue to be fairly unconvincing imitations of natural materials.

How to Install a Floating Floor

To install the underlayment, start in one corner and unroll the underlayment to the opposite wall. Cut the underlayment to fit, using a utility knife or scissors. Overlap the second underlayment sheet according to the manufacturer's recommendations, and secure the pieces in place with adhesive tape.

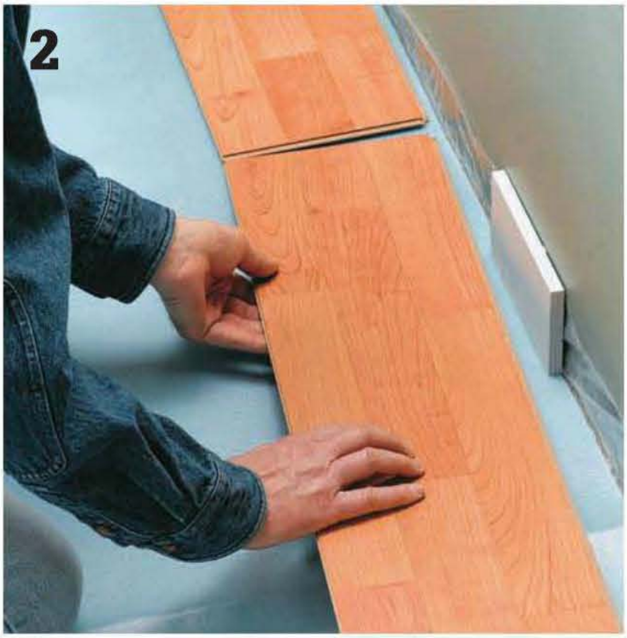

Working from the left corner of the room to right, set wall spacers and dry lay planks (tongue side facing the wall) against the wall. The spacers allow for expansion. If you are flooring a room more than 26 ft. long or wide, you need to buy appropriate-sized expansion joints. *Note: Some manufacturers suggest facing the groove side to the wall.*

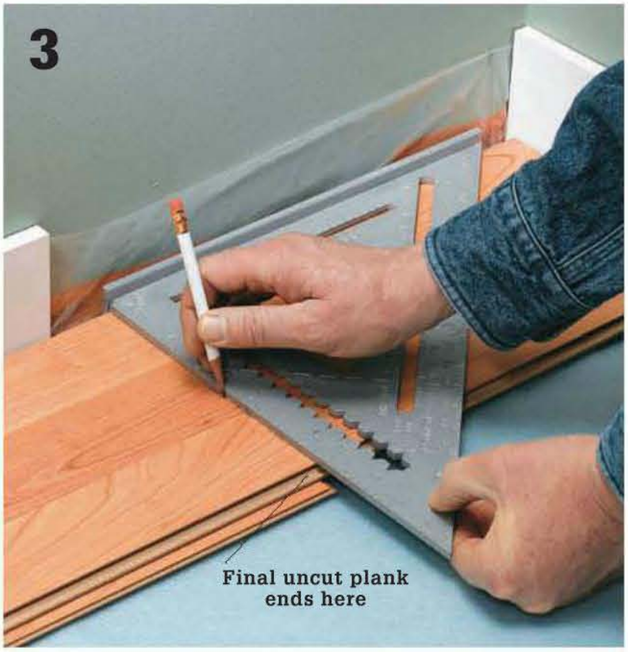
3
Final uncut plank ends here

Set a new plank right side up, on top of the previously laid plank, flush with the spacer against the wall at the end run. Line up a speed square with the bottom plank edge and trace a line. That's the cutline for the final plank in the row.

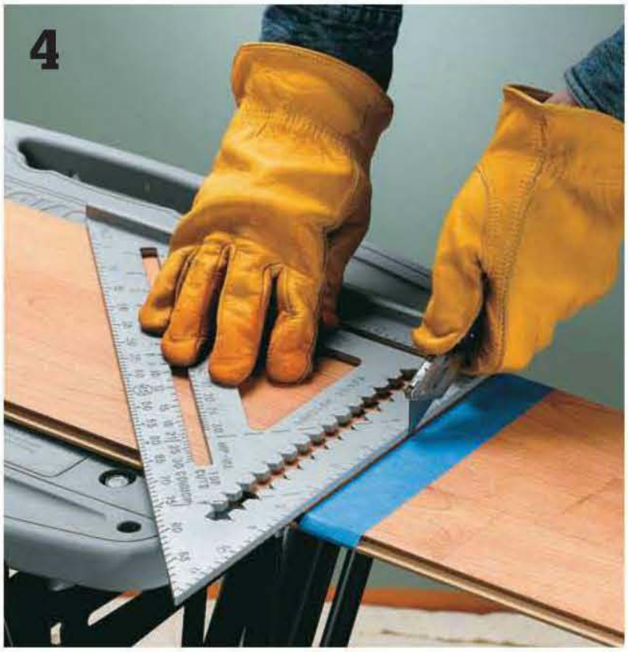

Press painter's tape along the cutline on the top of the plank to prevent chips when cutting. Score the line drawn in Step 3 with a utility knife. Turn the plank over and extend the pencil line to the backside.

(continued)

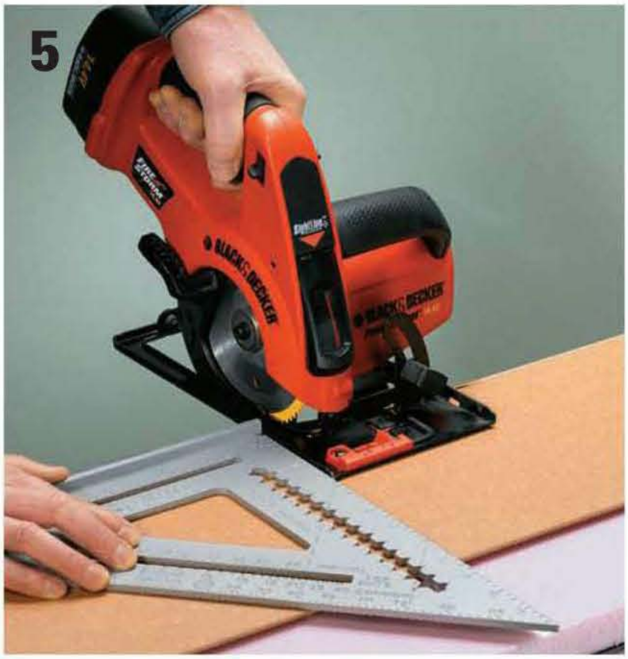
5

Clamp the board (face down) and rigid foam insulation or plywood to a work table. The foam reduces chipping. Clamp a speed square on top of the plank, as though you are going to draw another line parallel to the cutline—use this to make a straight cut. Place the circular saw's blade on the waste side of the actual cutline.

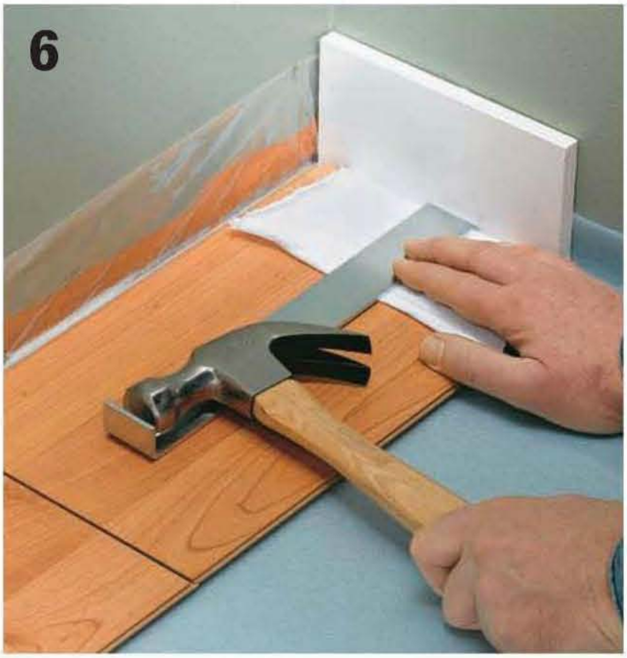
6

To create a tight fit for the last plank in the first row, place a spacer against the wall and wedge one end of a drawbar between it and the last plank. Tap the other end of the drawbar with a rubber mallet or hammer. Protect the laminate surface with a thin cloth.

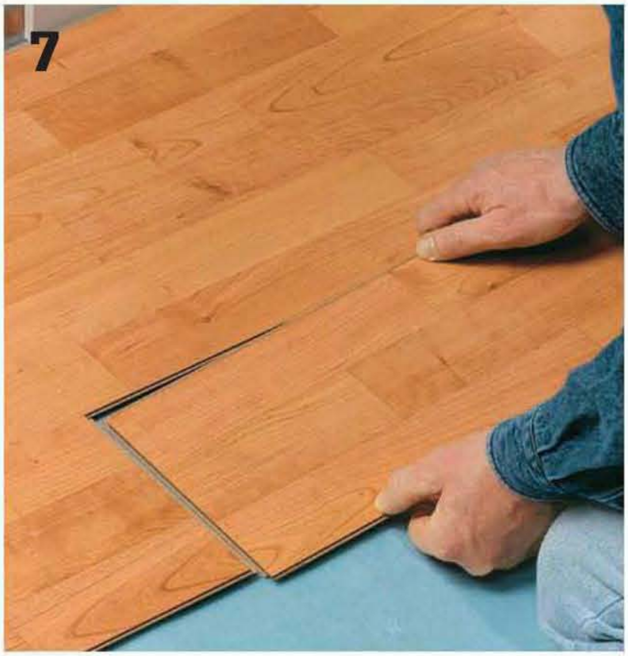
7

Continue to lay rows of flooring, making sure the joints are staggered. This prevents the entire floor from relying on just a few joints, and keeps the planks from lifting. Staggering also stengthens the floor, because the joints are shorter and more evenly distributed.

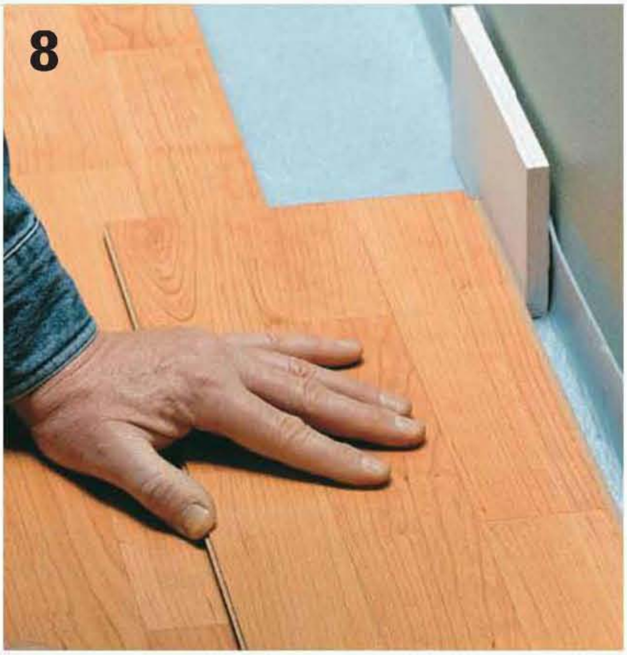
8

To fit the final row, place two planks on top of the last course; slide the top plank up against the wall spacer. Use the top plank to draw a cutline lengthwise on the middle plank. Cut the middle plank to size using the same method as in Step 3, just across the grain. The very last board must be cut lengthwise and widthwise to fit.

How to Work Around Obstacles

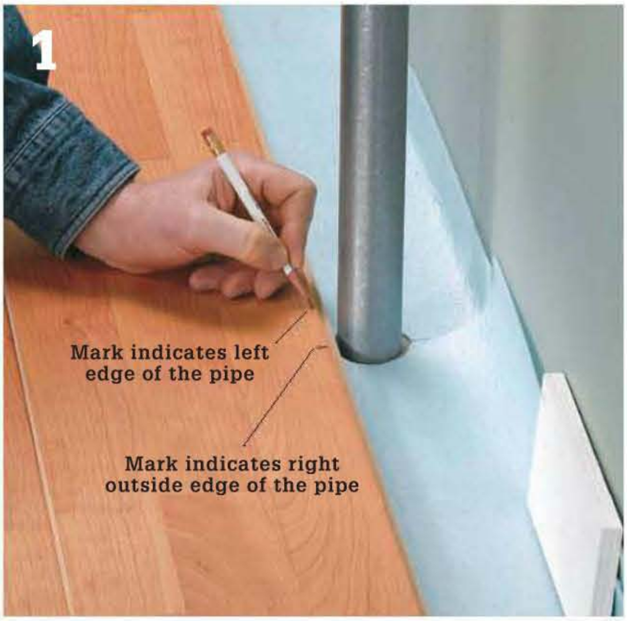
1
Mark indicates left edge of the pipe
Mark indicates right outside edge of the pipe

Position a plank and use a pencil to make two marks along the length of the plank, indicating the points where the obstacle begins and ends.

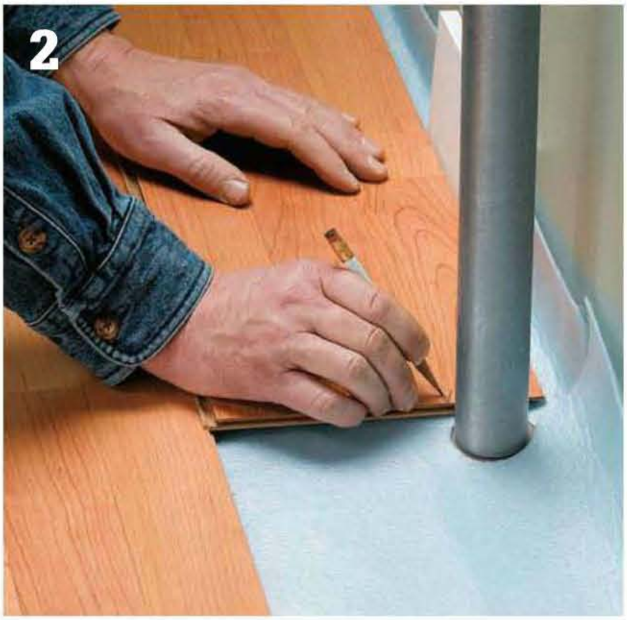

Position the plank end against the obstacle. Make two marks with a pencil, this time on the end of the plank to indicate where the obstacle falls along the width of the board.

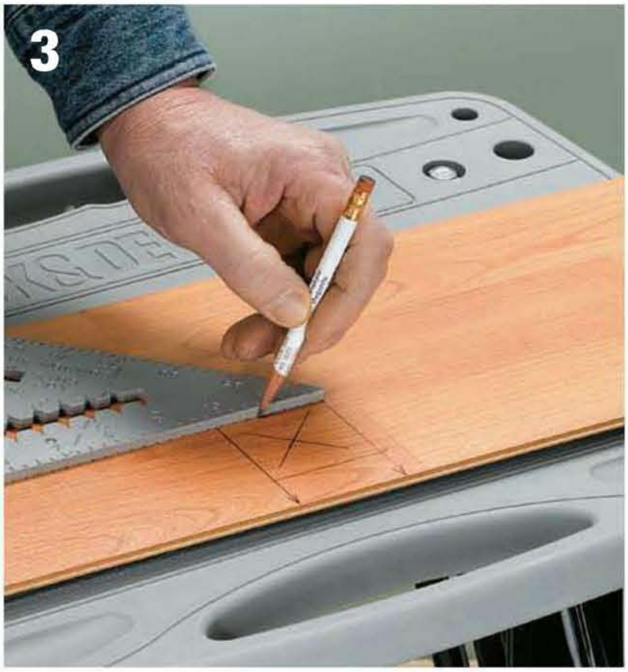

Use a speed square to extend the four lines. The space at which they intersect is the part of the plank that needs to be removed to make room for the obstacle to go through it. Use a drill with a Forstner bit, or a hole saw the same diameter as the space within the intersecting lines, and drill through the plank at the X. You'll be left with a hole; extend the cut to the edges with a jigsaw.

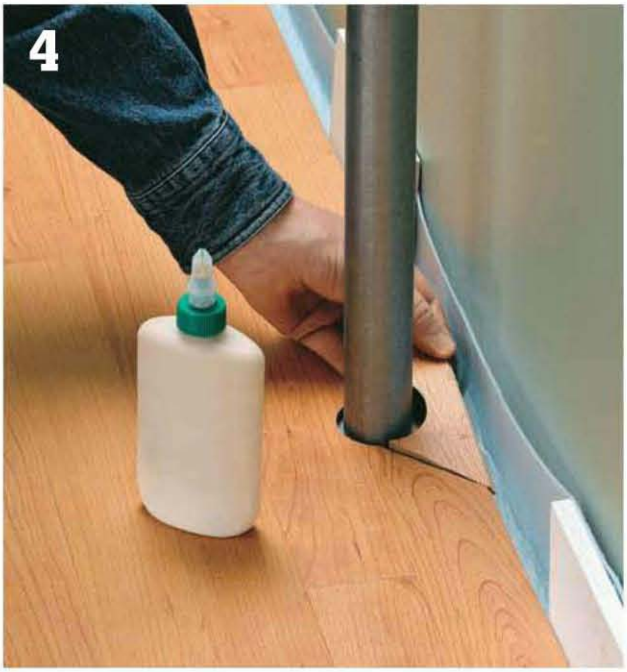

Install the plank by locking the tongue-and-groove joints with the preceding board. Fit the end piece in behind the pipe or obstacle. Apply manufacturer-recommended glue to the cut edges, and press the end piece tightly against the adjacent plank. Wipe away excess glue with a damp cloth.

Installing Parquet Flooring

For a hardwood floor with great design appeal, consider installing a parquet floor. It offers more visual interest than strip flooring without sacrificing the beauty and elegance of wood. Parquet comes in a variety of patterns and styles to create geometric designs. It can range from elaborate, custom-designed patterns on the high end, to the more common herringbone pattern, to the widely available and less expensive block design.

Parquet has experienced a radical transformation over the years. A few years ago, each individual piece of parquet was hand-cut and painstakingly assembled piece by piece. Today, parquet is prefabricated so the individual pieces making up the design are available as single tiles, which not only has reduced the cost, but has made the flooring easier to install.

Many types and designs of parquet floors are available, from custom-made originals to standard patterns, but they are all installed the same way—set in adhesive on a wood subfloor. The effort can be very rewarding: Parquet can be used to create shapes not possible with other wood flooring.

The finger block pattern is one of the most widely available parquet coverings and also one of the least expensive. The configuration of perpendicular strips of wood emphasizes the different grains and natural color variations.

Tools & Materials ▸

- Tape measure
- Chalk line
- Carpenter's square
- Parquet flooring
- Adhesive
- Notched trowel
- Putty knife
- Rubber mallet
- 100- to 150-pound floor roller
- Jigsaw
- Solvent

Parquet flooring tiles are made with real hardwood and can be a great vehicle for introducing exotic wood species at a reasonable cost. The custom parquet floor seen here is expensive, but a basic parquet floor with stock tiles is an affordable option and an easy DIY project.

How to Install Parquet Flooring

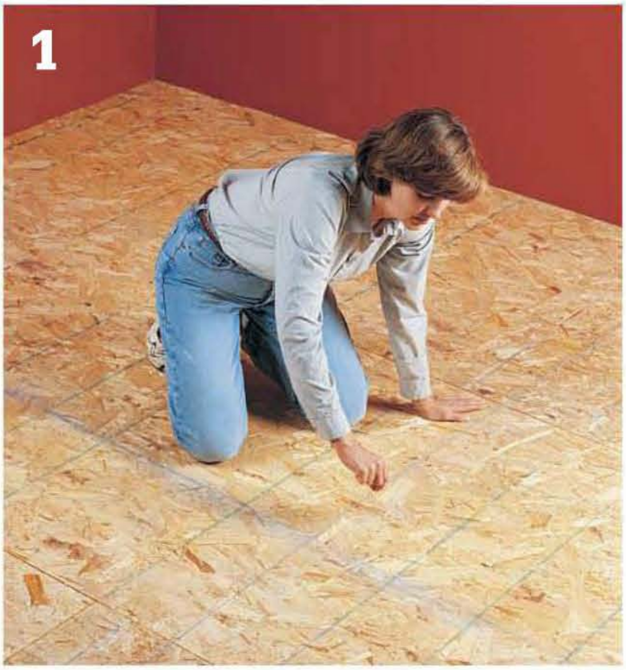

Mark the centerpoint of each wall. Snap chalk lines between the marks on opposite walls to establish your reference lines. Use the 3-4-5 triangle method to check the lines for squareness (see page 181).

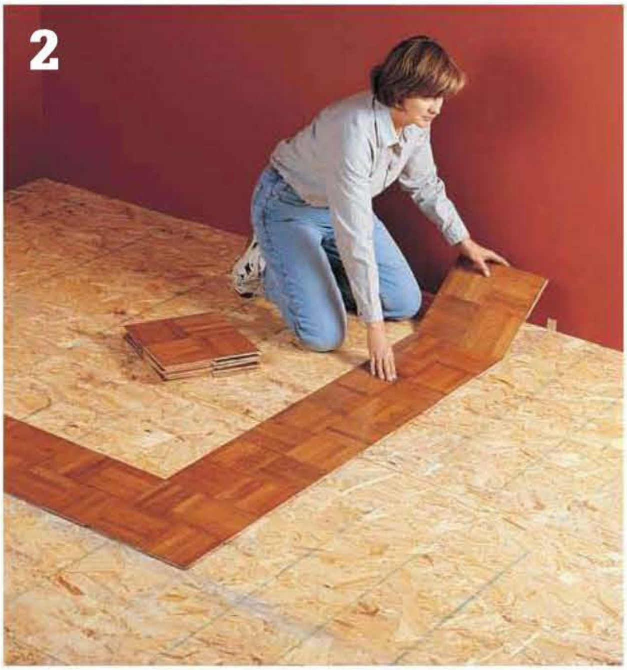

Lay out a dry run of panels from the center point along the reference lines to adjacent walls. Place ¼" spacers along the walls to allow for expansion. If more than half of the last panel needs to be cut off, adjust the lines by half the width of the panel. Snap new working lines, if necessary.

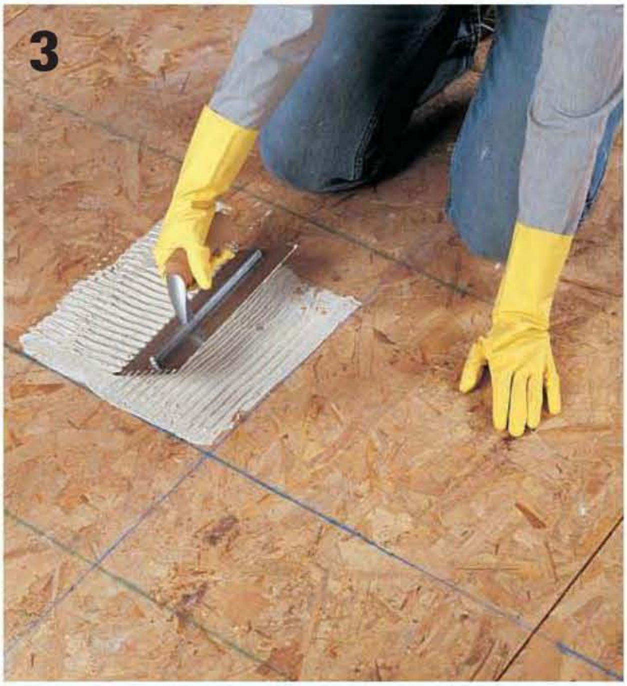

Put enough adhesive on the subfloor for your first panel, using a putty knife. Spread the adhesive into a thin layer with a notched trowel held at a 45° angle. Apply the adhesive right up to the working lines, but do not cover them.

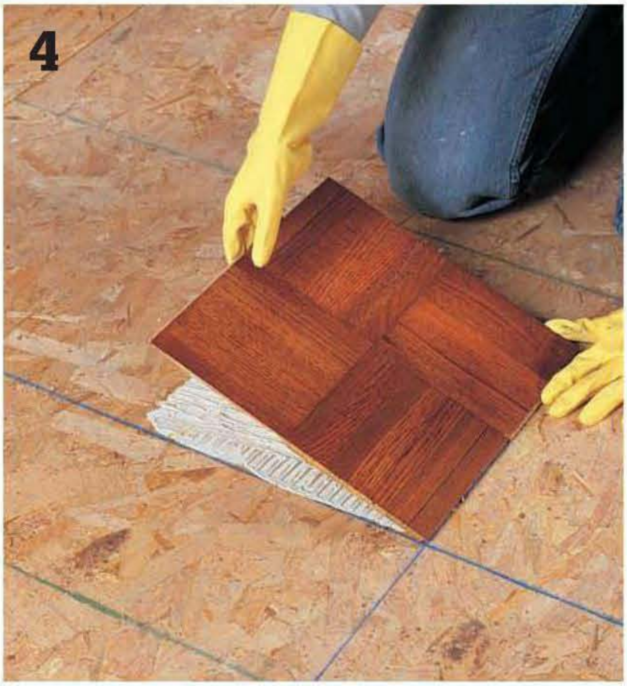

Place the first panel on the adhesive so two sides are flush with the working lines. Take care not to slide or twist the panel when setting it into place. This panel must be positioned correctly to keep the rest of your floor square.

(continued)

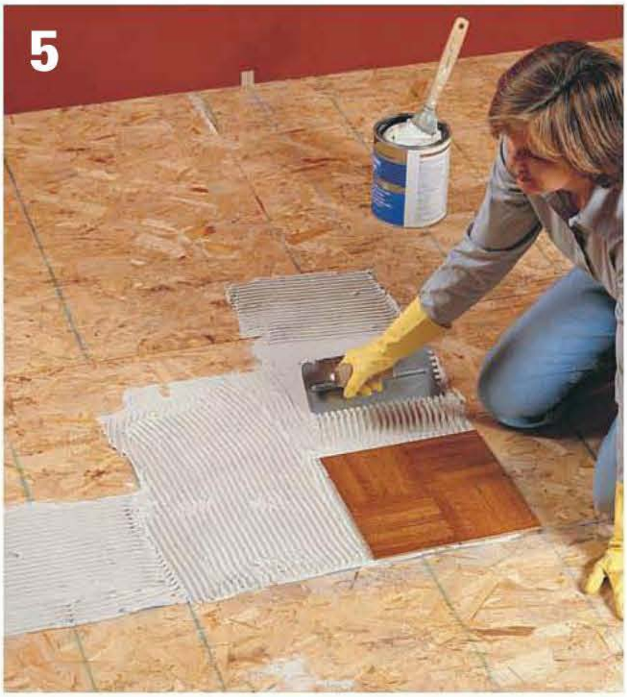

Apply enough adhesive for six to eight panels and spread it with a notched trowel.

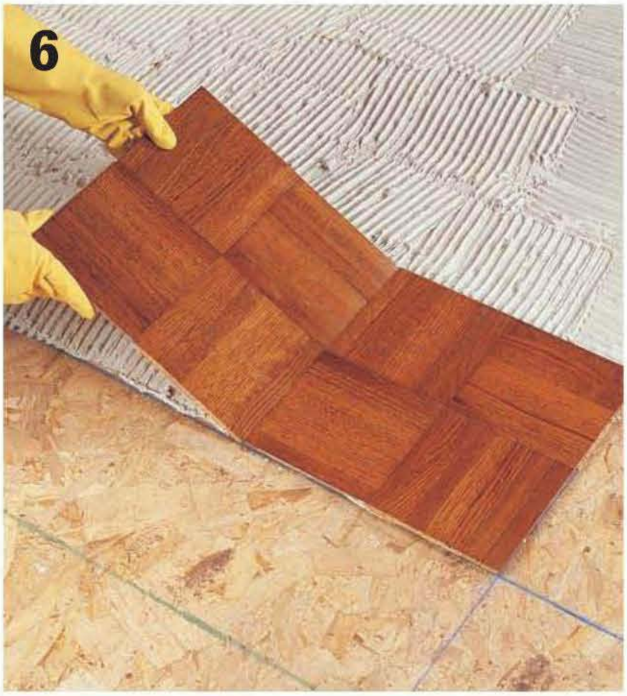

Set the next panel in place by holding it at a 45° angle and locking the tongue-and-groove joints with the first panel. Lower the panel onto the adhesive without sliding it. Install remaining panels the same way.

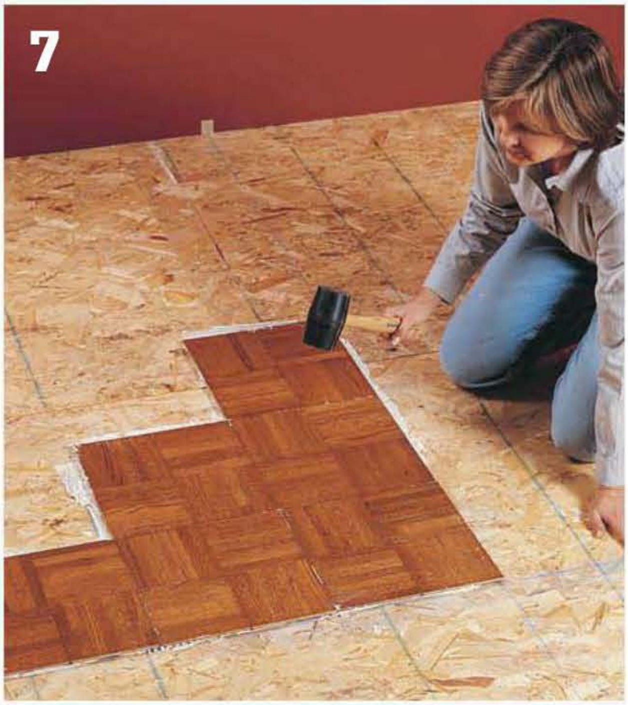

After every six to eight panels are installed, tap them into the adhesive with a rubber mallet.

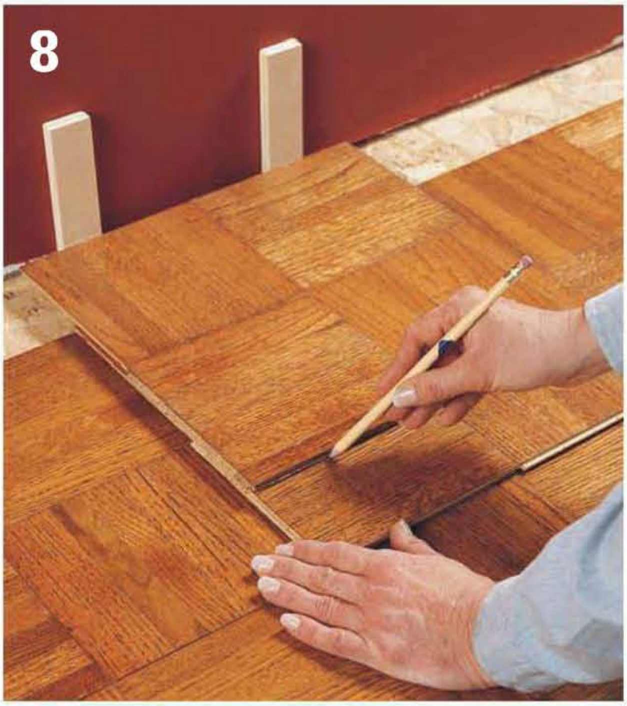

For the last row, align panels over the top of the last installed row. Place a third row over the top of these, with the sides butted against ½" spacers along the wall. Draw a line along the edge of the third panels onto the second row, cut the panels at the marks, and install.

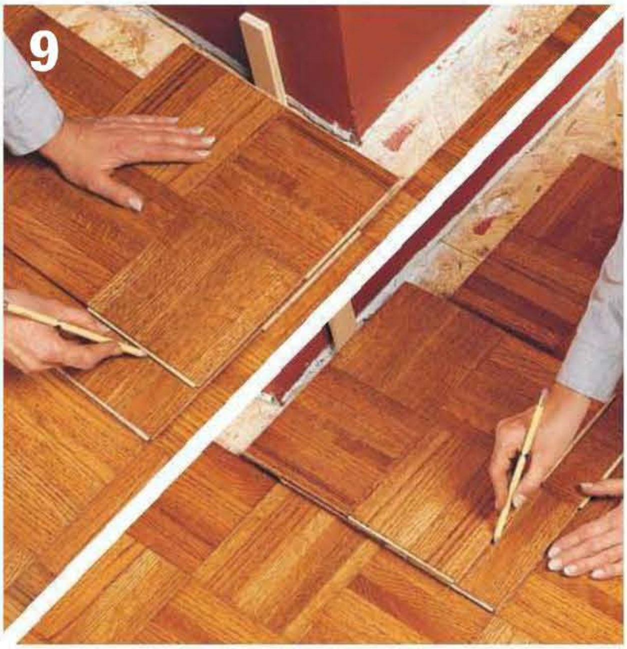

To work around corners or obstacles, align a panel over the last installed panel, then place another panel on top of it as in step 8. Keep the top panel ½" from the wall or obstacle and trace along the opposite edge onto the second panel (top). Move the top two panels to the adjoining side, making sure not to turn the top panel. Make a second mark on the panel the same way (bottom). Cut the tile with a jigsaw and install.

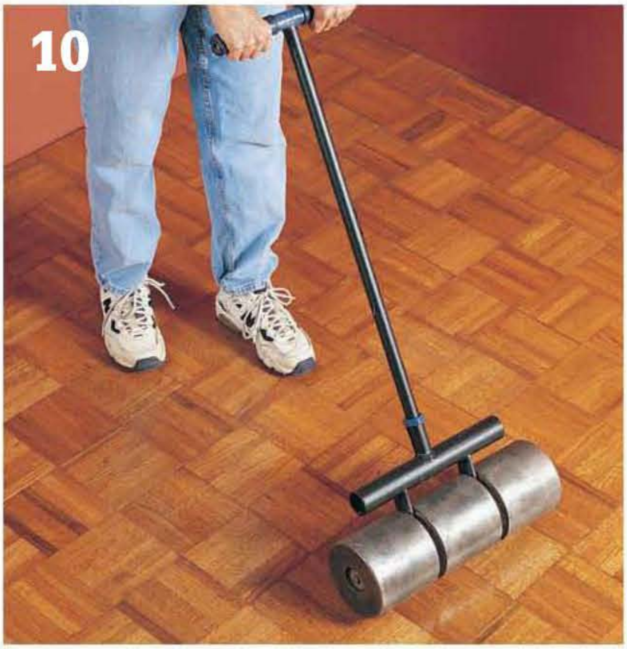

Within 4 hours of installing the floor, roll the floor with a 100- to 150-pound floor roller. Wait at least 24 hours before walking on the floor again.

How to Install Parquet in a Diagonal Layout

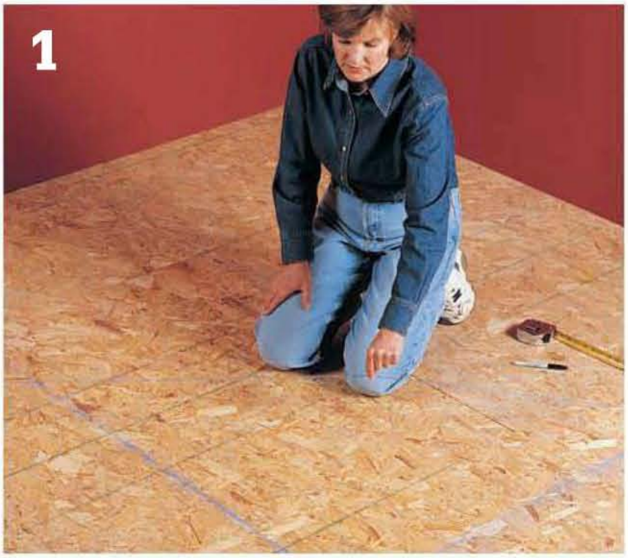

Establish perpendicular working lines following Step 1 on page 171. Measure 5 ft. from the centerpoint along each working line and make a mark. Snap chalk lines between the 5 ft. marks. Mark the centerpoint of these lines, then snap a chalk line through the marks to create a diagonal reference line.

Lay out a dry run of tiles along a diagonal line. Adjust your starting point as necessary. Lay the flooring along the diagonal line using adhesive, following the steps for installing parquet (pages 171 to 173). Make paper templates for tile along walls and in corners. Transfer the template measurements to tiles, and cut to fit.

Installing Sheet Vinyl

Preparing a perfect underlayment is the most important phase of resilient sheet vinyl installation. Cutting the material to fit the contours of the room is a close second. The best way to ensure accurate cuts is to make a cutting template. Some manufacturers offer template kits, or you can make your own. Be sure to use the recommended adhesive for the sheet vinyl you are installing. Many manufacturers require that you use their glue for installation. Use extreme care when handling the sheet vinyl, especially felt-backed products, to avoid creasing and tearing.

Tools & Materials ▸

Linoleum knife
Framing square
Compass
Scissors
Non-permanent
  felt-tipped pen
Utility knife
Straightedge
¼" V-notched trowel
J-roller
Stapler
Flooring roller
Chalk line
Heat gun
1⁄16" V-notched trowel
Straightedge
Vinyl flooring
Masking tape
Heavy butcher
  or brown
  wrapping paper
Duct tape
Flooring adhesive
⅜" staples
Metal threshold bars
Nails

Sheet vinyl has no seams (or at least very few seams), which makes it a practical choice for wet areas such as bathrooms. Advancing technology has made a host of new patterns and styles available, including the mosaic tile look-alike pattern seen here.

Tools for Resilient Floors

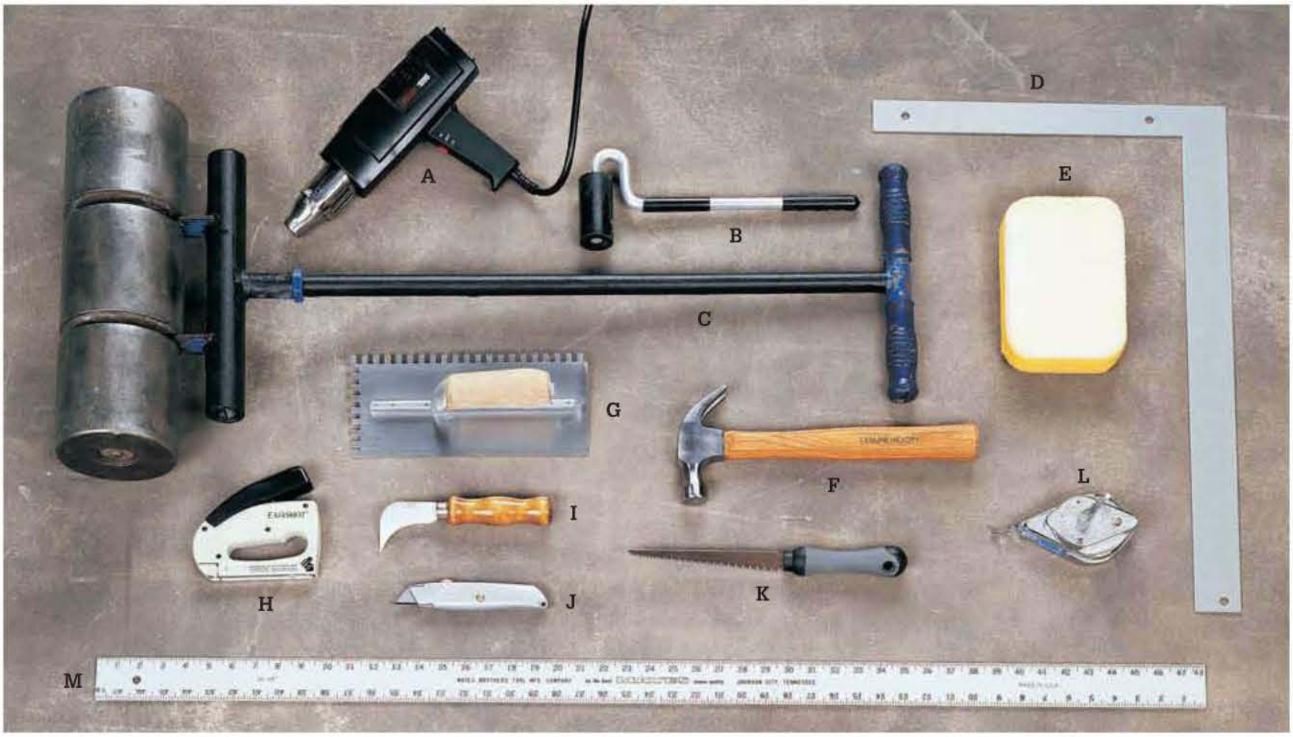

Tools for resilient flooring include: a heat gun (A), J-roller (B), floor roller (C), framing square (D), sponge (E), hammer (F), notched trowel (G), stapler (H), linoleum knife (I), utility knife (J), wallboard knife (K), chalk line (L), straightedge (M).

Buying & Estimating

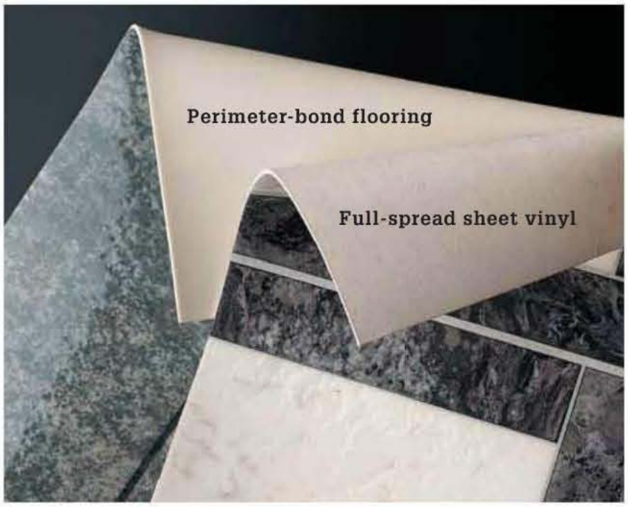

Resilient sheet vinyl comes in full-spread and perimeter-bond styles. Full-spread sheet vinyl has a felt-paper backing and is secured with adhesive that is spread over the floor before installation. Perimeter-bond flooring, identifiable by its smooth, white PVC backing, is laid directly on underlayment and is secured by a special adhesive spread along the edges and seams.

Dry-back tile
Self-adhesive tile

Resilient tile comes in self-adhesive and dry-back styles. Self-adhesive tile has a pre-applied adhesive protected by wax paper backing that is peeled off as the tiles are installed. Dry-back tile is secured with adhesive spread onto the underlayment before installation. Self-adhesive tile is easier to install than dry-back tile, but the bond is less reliable. Don't use additional adhesives with self-adhesive tile.

Evaluating an Existing Floor

The first step in preparing for a new floor covering is evaluating your old floor. A careful examination can help you decide whether to repair damaged areas, or replace the flooring altogether.

Evaluating your floor is a three-step process. Begin by identifying the existing floor material and the installation method used. Is your sheet vinyl attached using the full-spread method or the perimeter-bond method? Is your carpet glued down or stretched? Next, check the condition of the floor. Is it securely attached or is it loose in spots? Is it chipped or cracked? Finally, note the height of the existing floor in relation to adjoining floor surfaces. Is it significantly higher than surrounding floors?

A new floor covering or underlayment can often be installed on top of existing flooring. If the existing flooring is not sound or smooth, however, you will have to do some preparation work. Applying a floor leveler is one way to make your existing floor easier to use. More complex preparations may involve removing and replacing the underlayment or making spot repairs to the subfloor.

Warning ▸

Resilient flooring manufactured before 1986 may contain asbestos, which can cause severe lung problems if inhaled. The recommended method for dealing with asbestos-laden flooring is to cover it with an underlayment. If the flooring must be removed, do not do the work yourself. Instead, consult a certified asbestos-abatement contractor.

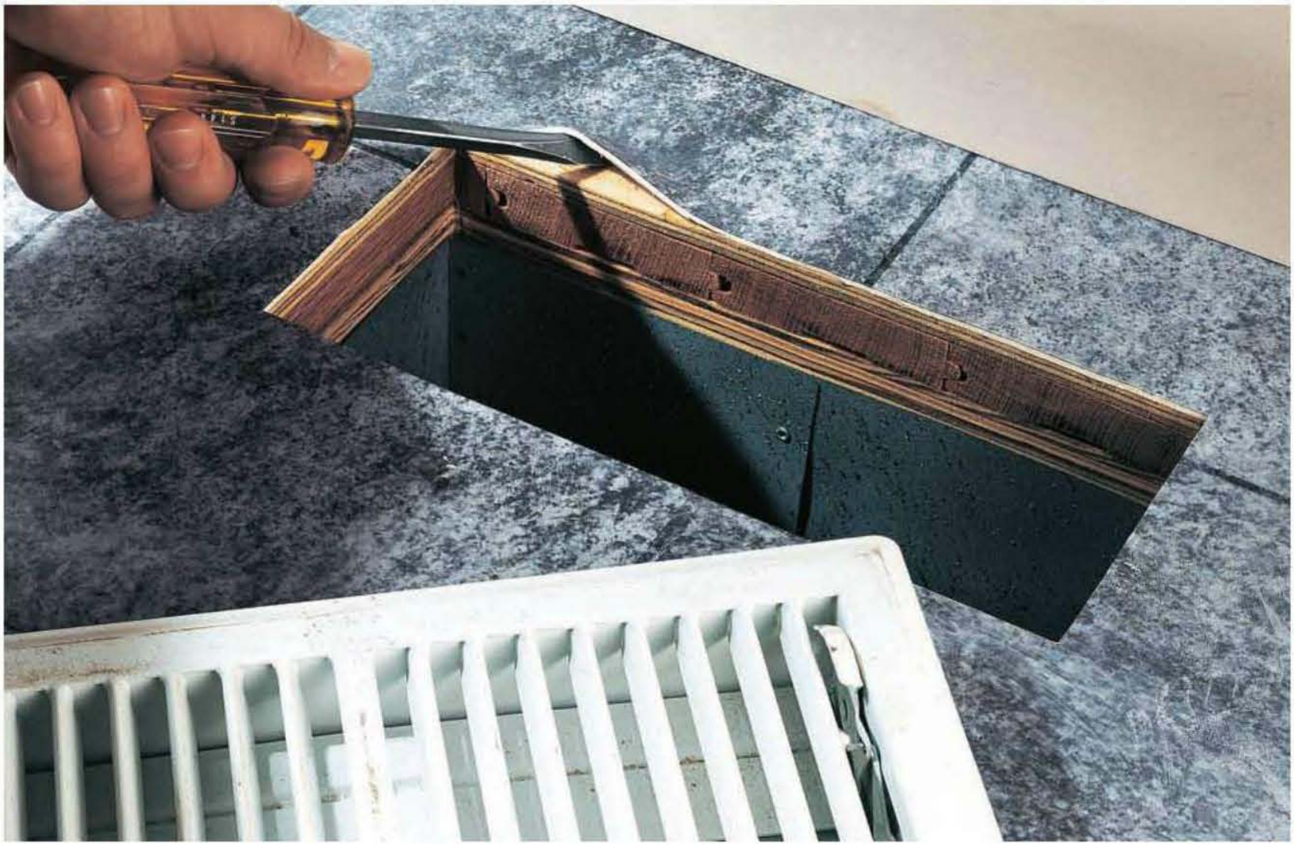

Determining the number and type of coverings already on your floor is an important early evaluation step. Too many layers of flooring and underlayment can stress floor joists and ultimately cause a new floor to fail. An easy way to check for old flooring is to remove floor vents.

How to Install Perimeter-bond Sheet Vinyl

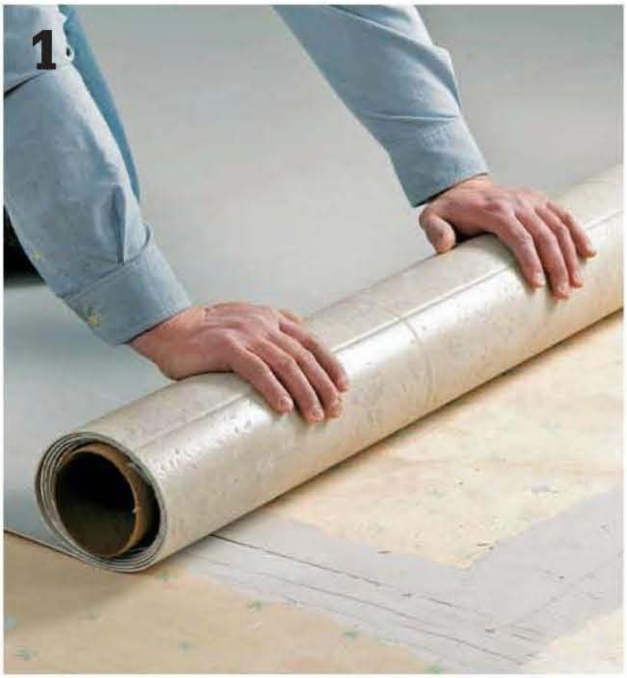

Unroll the flooring on any large, flat, clean surface. To prevent wrinkles, sheet vinyl comes from the manufacturer rolled with the pattern-side out. Unroll the sheet and turn it pattern-side up for marking.

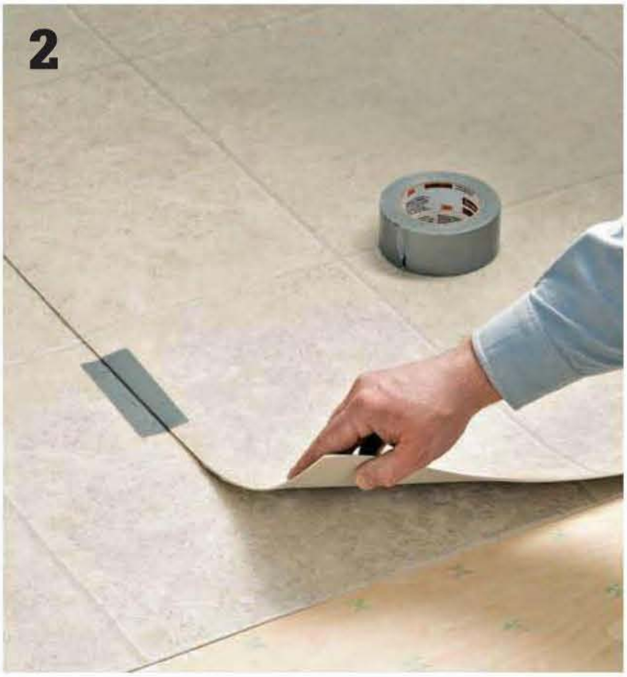

For two-piece installations, overlap the edges of the sheets by at least 2". Plan to have the seams fall along the pattern lines or simulated grout joints. Align the sheets so the pattern matches, then tape the sheets together with duct tape.

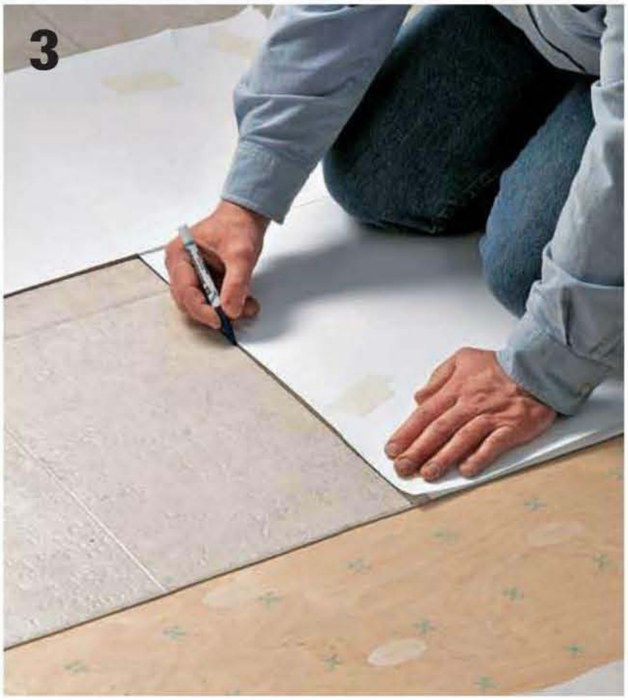

Position the paper template over the sheet vinyl and tape it in place. Trace the outline of the template onto the flooring using a non-permanent felt-tipped pen.

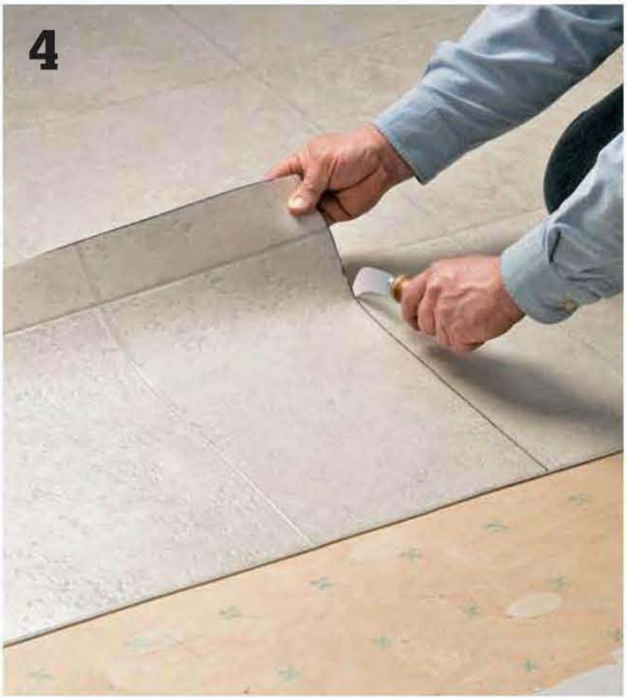

Remove the template. Cut the sheet vinyl with a sharp linoleum knife or a utility knife with a new blade. Use a straightedge as a guide for making longer cuts.

(continued)

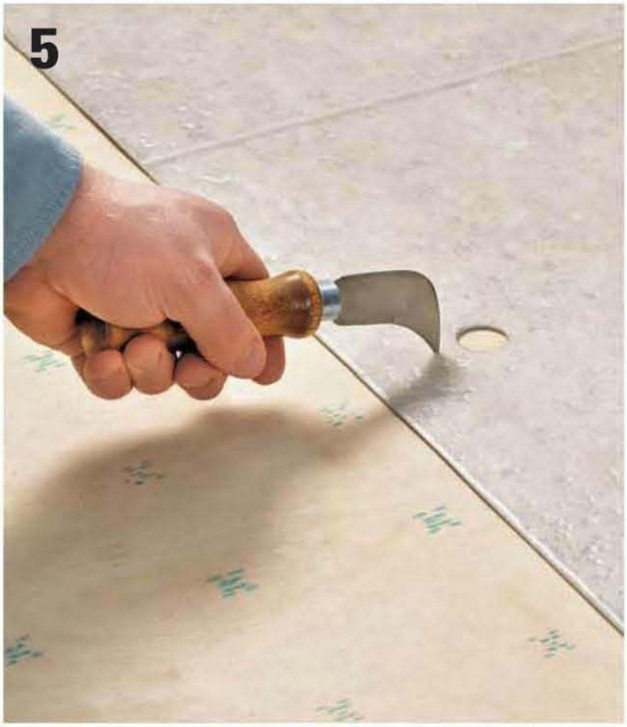

Cut holes for pipes and other permanent obstructions. Cut a slit from each hole to the nearest edge of the flooring. Whenever possible, make slits along pattern lines.

Roll up the flooring loosely and transfer it to the installation area. Do not fold the flooring. Unroll and position the sheet vinyl carefully. Slide the edges beneath door casings.

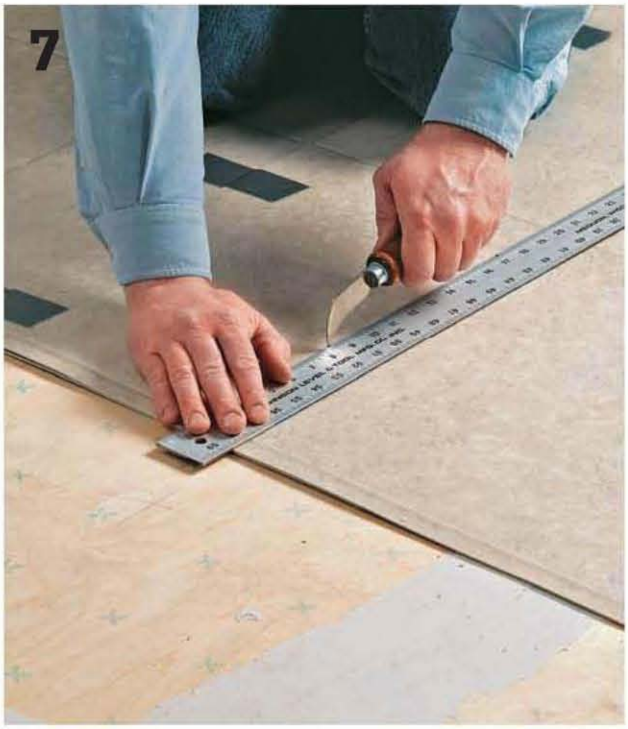

Cut the seams for two-piece installations using a straightedge as a guide. Hold the straightedge tightly against the flooring, and cut along the pattern lines through both pieces of vinyl flooring.

Remove both pieces of scrap flooring. The pattern should now run continuously across the adjoining sheets of flooring.

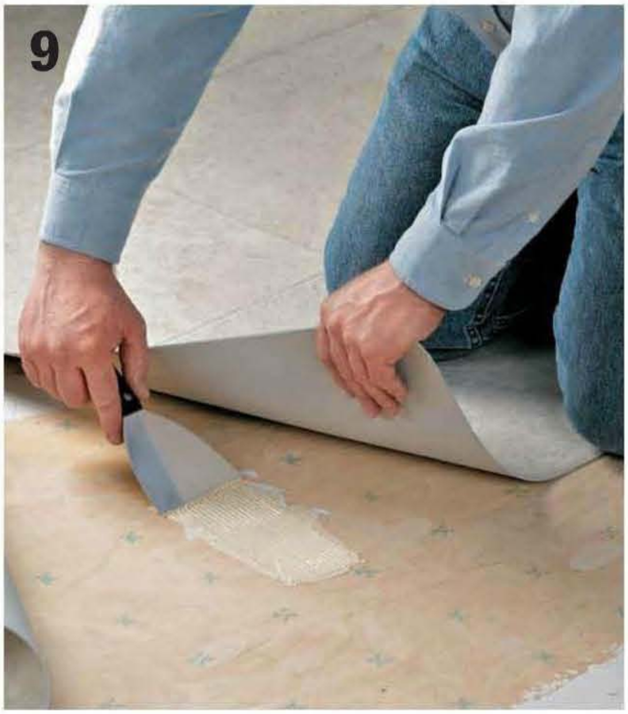

Fold back the edges of both sheets. Apply a 3" band of multipurpose flooring adhesive to the underlayment or old flooring, using a ¼" V-notched trowel or wallboard knife.

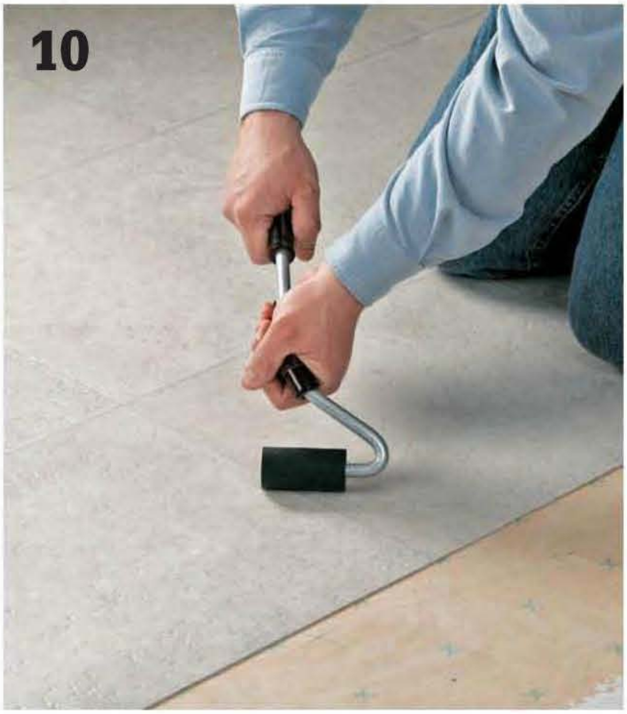

Lay the seam edges one at a time onto the adhesive. Make sure the seam is tight, pressing the gaps together with your fingers, if needed. Roll the seam edges with a J-roller or wallpaper seam roller.

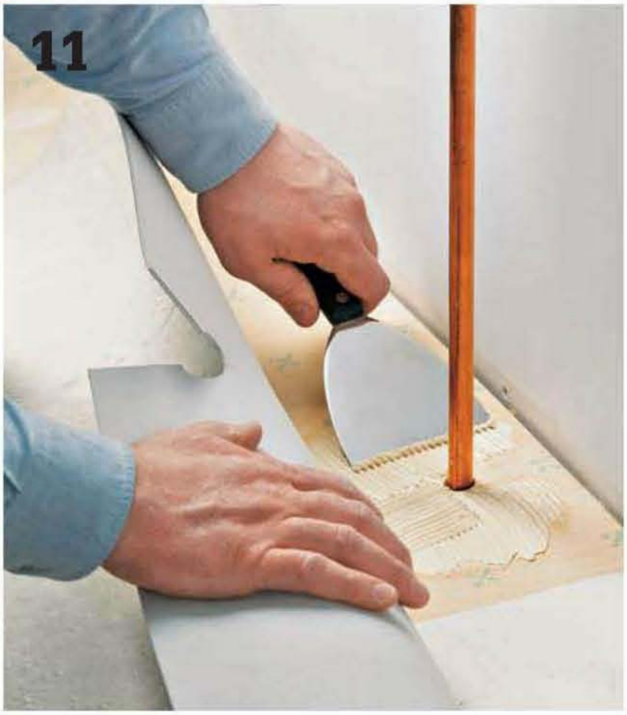

Apply flooring adhesive underneath flooring cuts at pipes or posts and around the entire perimeter of the room. Roll the flooring with the roller to ensure good contact with the adhesive.

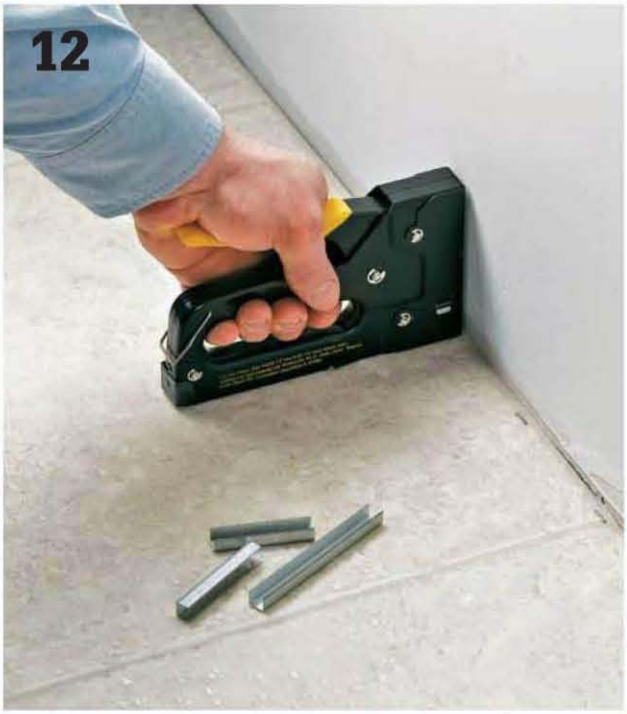

If you're applying flooring over a wood underlayment, fasten the outer edges of the sheet with ⅜" staples driven every 3". Make sure the staples will be covered by the base molding.

Installing Resilient Tile

As with any tile installation, resilient tile requires carefully positioned layout lines. Before committing to any layout and applying tile, conduct a dry run to identify potential problems.

Keep in mind the difference between reference lines (see opposite page) and layout lines. Reference lines mark the center of the room and divide it into quadrants. If the tiles don't lay out symmetrically along these lines, you'll need to adjust them slightly, creating layout lines. Once layout lines are established, installing the tile is a fairly quick process. Be sure to keep joints between the tiles tight and lay the tiles square.

Tiles with an obvious grain pattern can be laid so the grain of each tile is oriented identically throughout the installation. You can also use the quarter-turn method, in which each tile has its pattern grain running perpendicular to that of adjacent tiles. Whichever method you choose, be sure to be consistent throughout the project.

Tools & Materials ▸

- Tape measure
- Chalk line
- Framing square
- Utility knife
- 1⁄16" notched trowel
- Heat gun
- Resilient tile
- Flooring adhesive (for dry-back tile)

Resilient tiles have a pattern layer that is bonded to a vinyl base and coated with a transparent wear layer. Some come with adhesive pre-applied and covered by a paper backing, others have dry backs and are designed to be set into flooring adhesive.

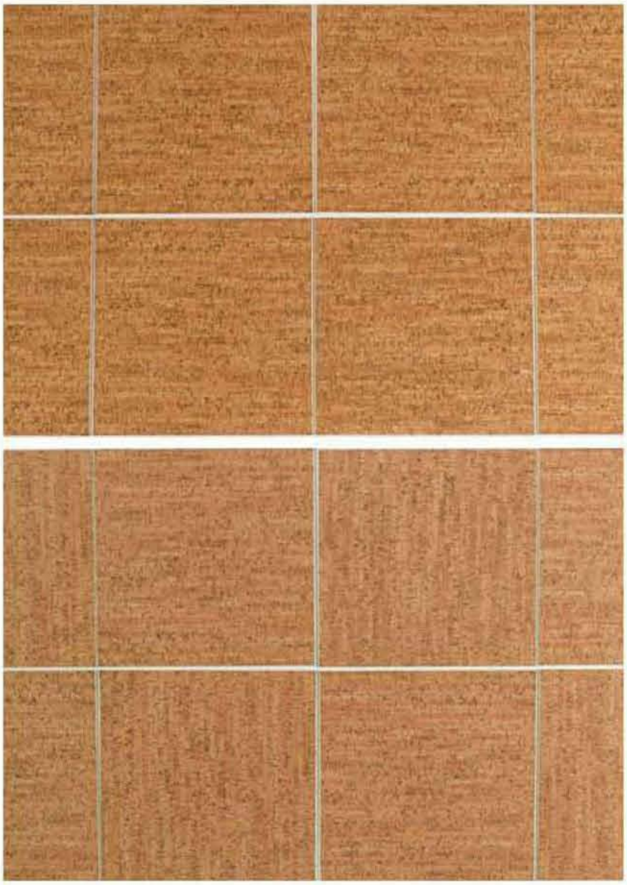

Check for noticeable directional features, like the grain of the vinyl particles. You can set the tiles in a running pattern so the directional feature runs in the same direction (top), or in a checkerboard pattern using the quarter-turn method (bottom).

How to Make Reference Lines for Tile Installation

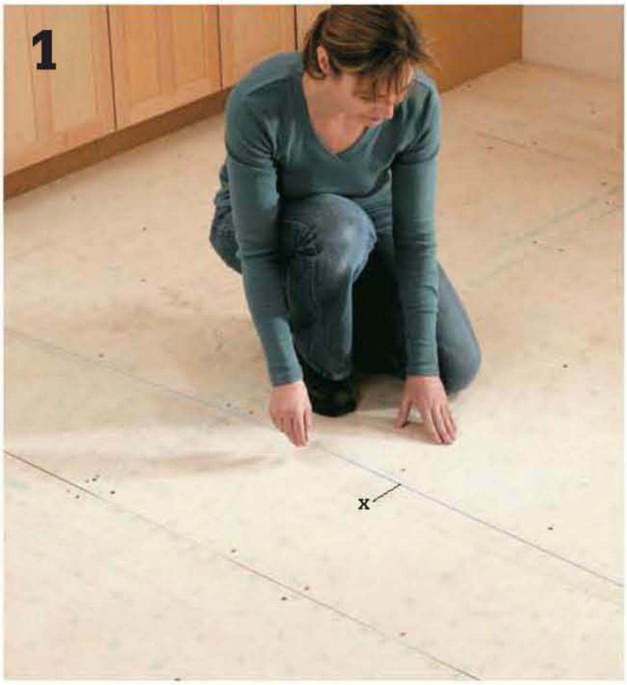

Position a reference line (X) by measuring along opposite sides of the room and marking the center of each side. Snap a chalk line between these marks.

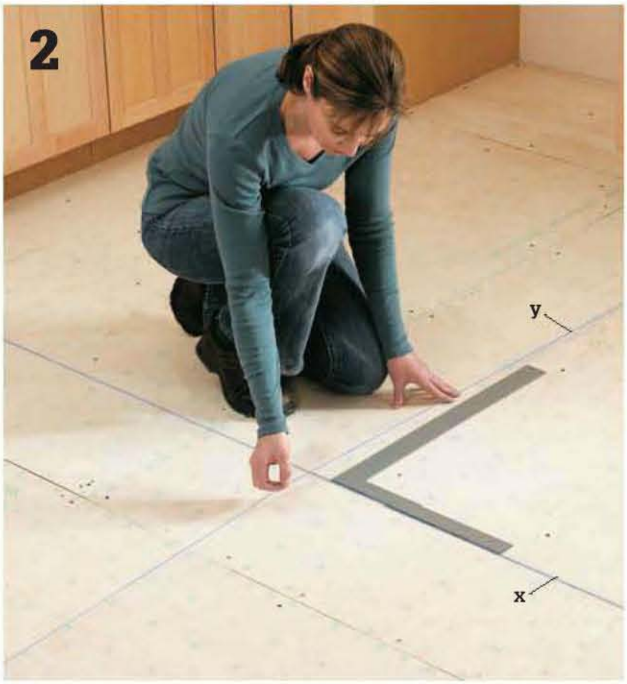

Measure and mark the centerpoint of the chalk line. From this point, use a framing square to establish a second reference line perpendicular to the first one. Snap the second line (Y) across the room.

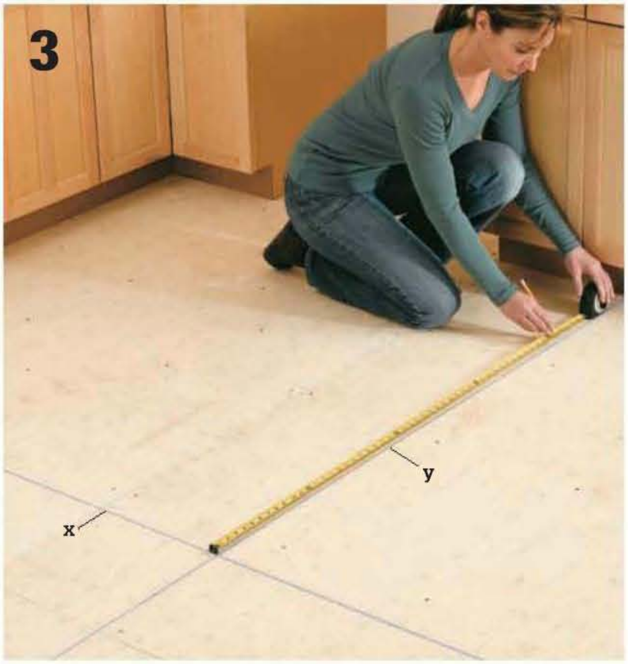

Check the reference lines for squareness using the 3-4-5 triangle method. Measure along reference line X and make a mark 3 ft. from the centerpoint. Measure from the centerpoint along reference line Y and make a mark at 4 ft.

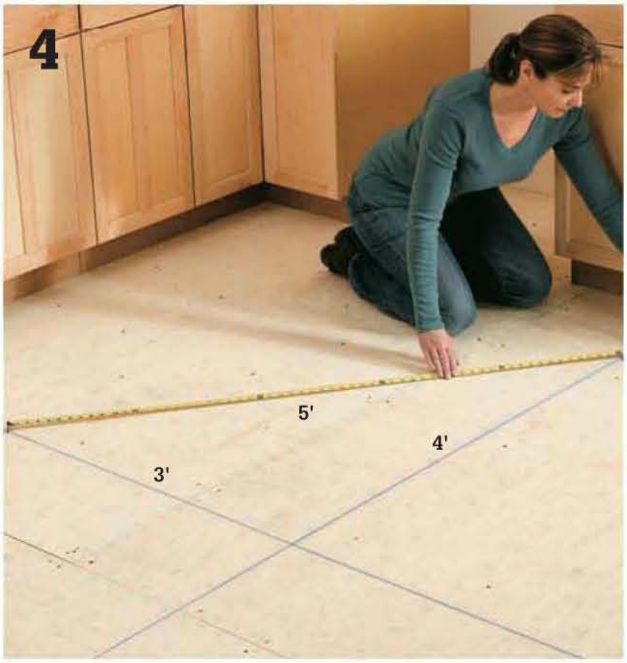

Measure the distance between the marks. If the reference lines are perpendicular, the distance will measure exactly 5 ft. If not, adjust the reference lines until they're exactly perpendicular to each other.

How to Install Dry-backed Resilient Tile

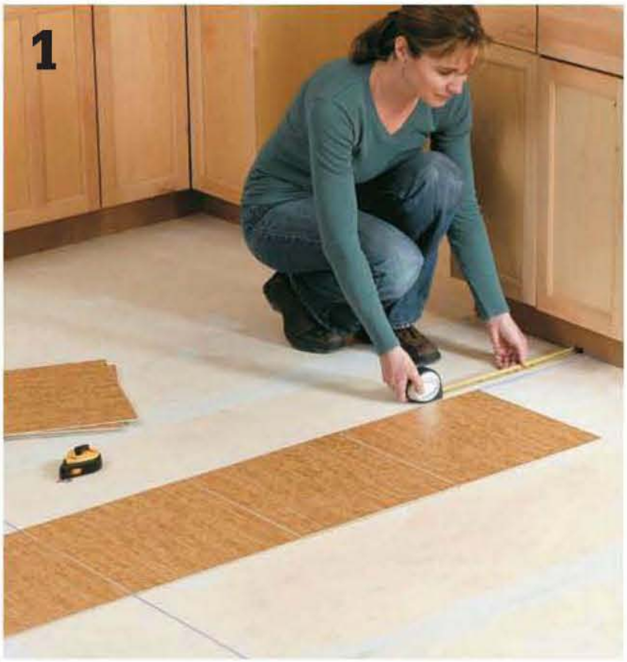

Snap perpendicular reference lines with a chalk line. Dry-fit tiles along layout line Y so a joint falls along reference line X. If necessary, shift the layout to make the layout symmetrical or to reduce the number of tiles that need to be cut.

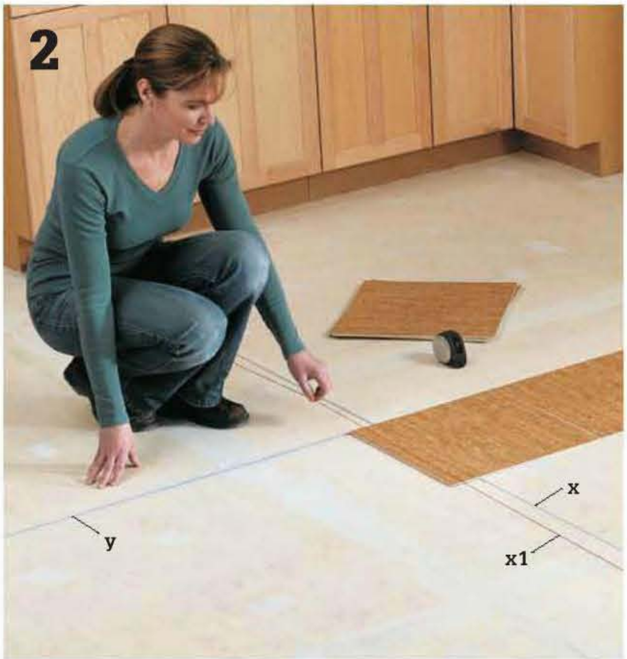

If you shift the tile layout, create a new line that is parallel to reference line X and runs through a tile joint near line X. The new line, X1, is the line you'll use when installing the tile. Use a different colored chalk to distinguish between lines.

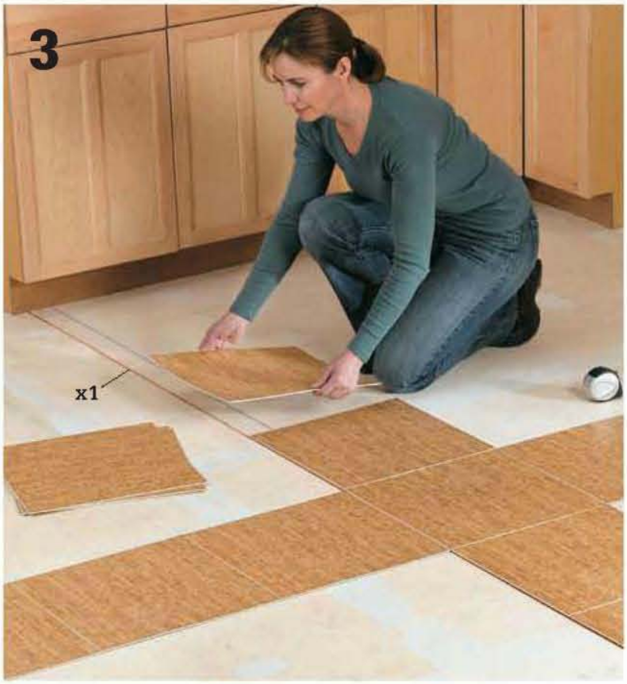

Dry-fit tiles along the new line, X1. If necessary, adjust the layout line as in steps 1 and 2.

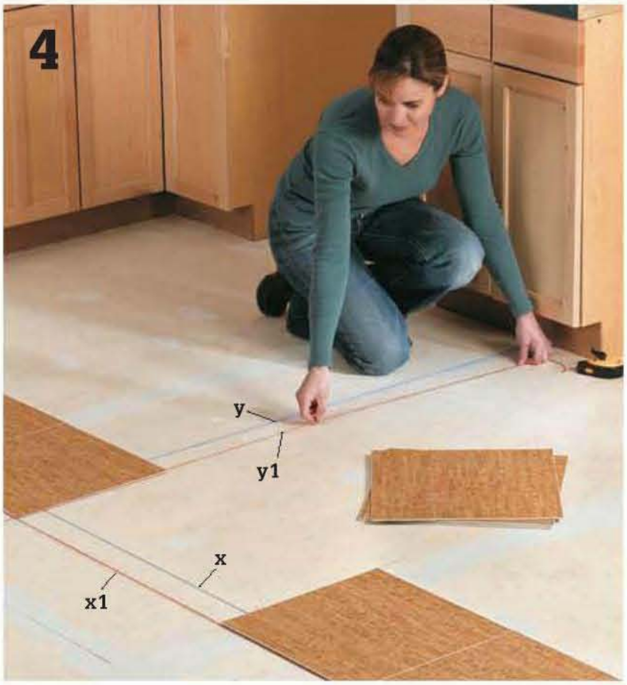

If you adjusted the layout along X1, measure and make a new layout line, Y1, that's parallel to reference line Y and runs through a tile joint. Y1 will form the second layout line you'll use during installation.

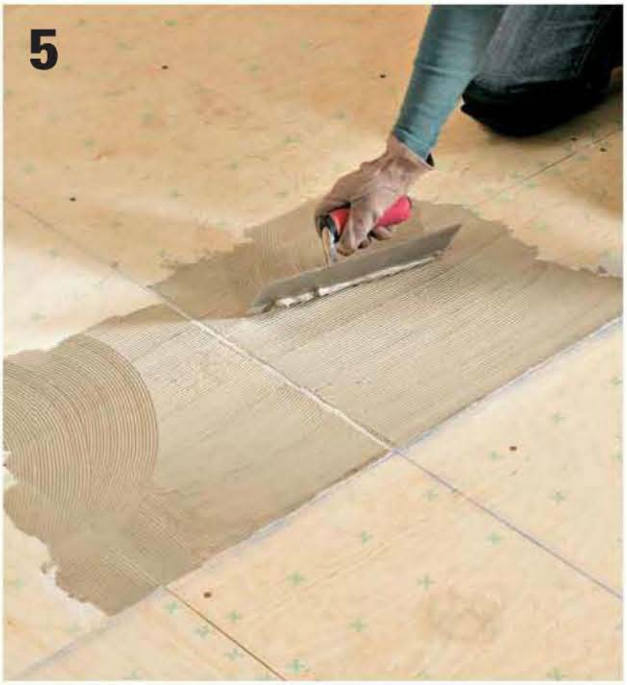

Apply adhesive around the intersection of the layout lines using a trowel with 1⁄16" V-shaped notches. Hold the trowel at a 45° angle and spread adhesive evenly over the surface.

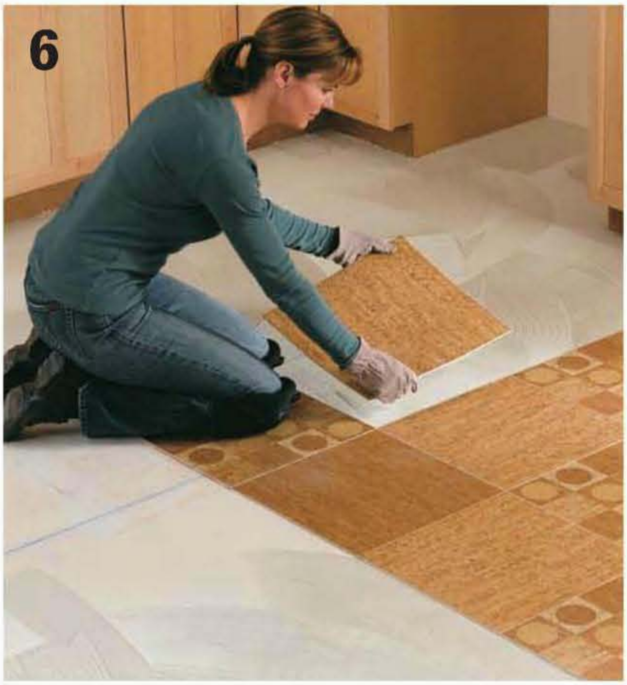

Spread adhesive over most of the installation area, covering three quadrants. Allow the adhesive to set according to the manufacturer's instructions, then begin to install the tile at the intersection of the layout lines. You can kneel on installed tiles to lay additional tiles.

When the first three quadrants are completely tiled, spread adhesive over the remaining quadrant, then finish setting the tile.

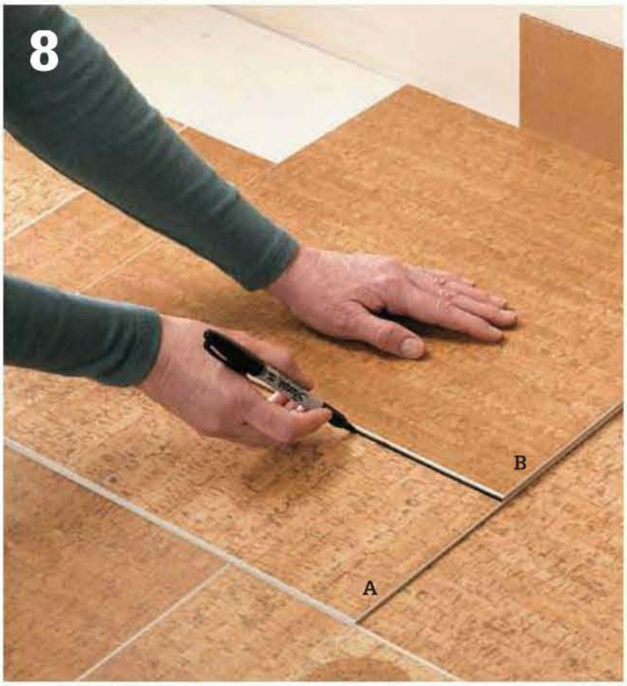

To cut tiles to fit along the walls, place the tile to be cut (A) face up on top of the last full tile you installed. Position a 1⁄8"-thick spacer against the wall, then set a marker tile (B) on top of the tile to be cut. Trace along the edge of the marker tile to draw a cutting line.

(continued)

Tip ▸

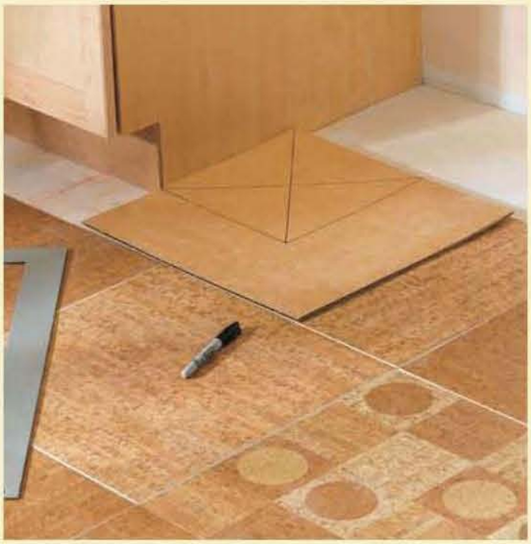

To mark tiles for cutting around outside corners, make a cardboard template to match the space, keeping a ⅛" gap along the walls. After cutting the template, check to make sure it fits. Place the template on a tile and trace its outline.

9

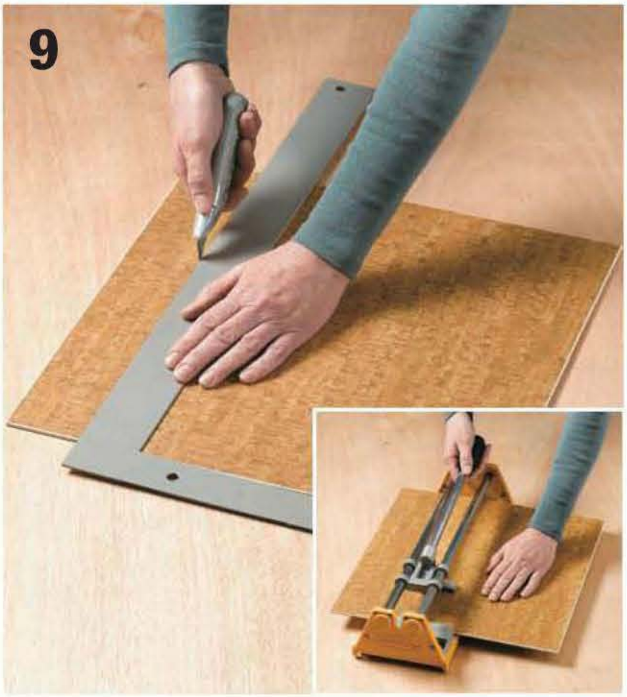

Cut tile to fit using a utility knife and straightedge. Hold the straightedge securely against the cutting line to ensure a straight cut. *Option: You can use a ceramic-tile cutter to make straight cuts in thick vinyl tiles (see inset).*

10

Install cut tiles next to the walls. If you're precutting all tiles before installing them, measure the distance between the wall and install tiles at various points in case the distance changes.

11

Continue installing tile in the remaining quadrants until the room is completely covered. Check the entire floor. If you find loose areas, press down on the tiles to bond them to the underlayment. Install metal threshold bars at room borders where the new floor joins another floor covering.

How to Install Self-adhesive Resilient Tile

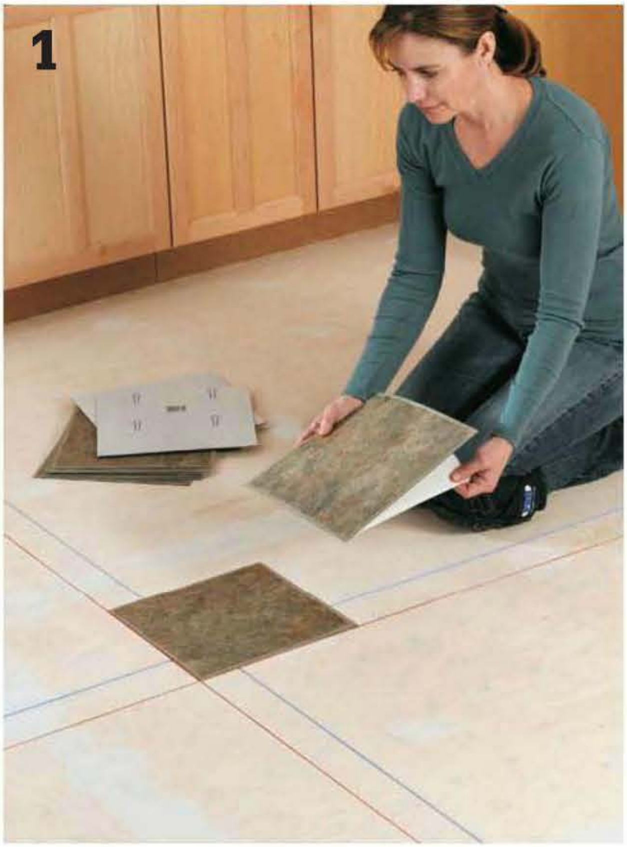

Once your reference lines are established, peel off the paper backing and install the first tile in one of the corners formed by the intersecting layout lines. Lay three or more tiles along each layout lines in the quadrant. Rub the entire surface of each tile to bond the adhesive to the floor underlayment.

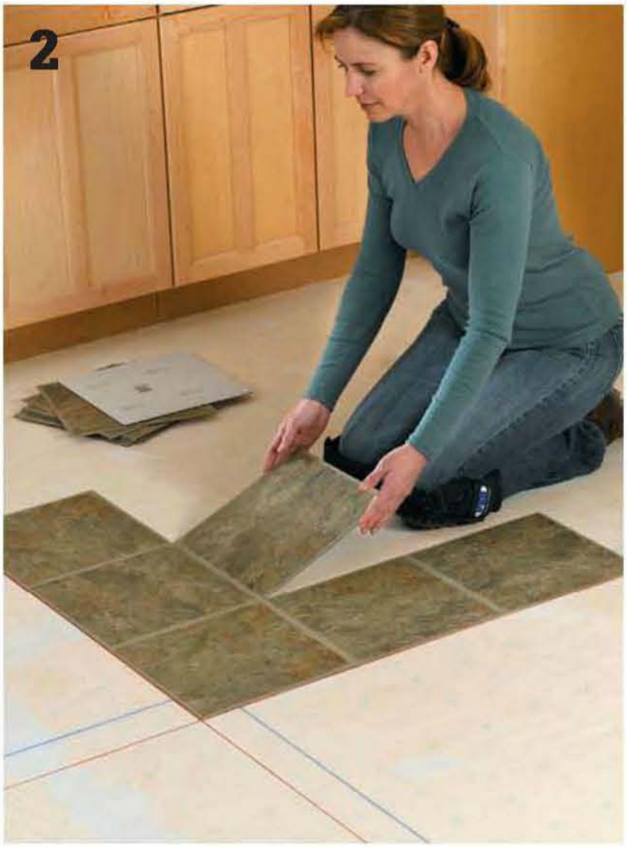

Begin installing tiles in the interior area of the quadrant. Keep the joints tight between tiles..

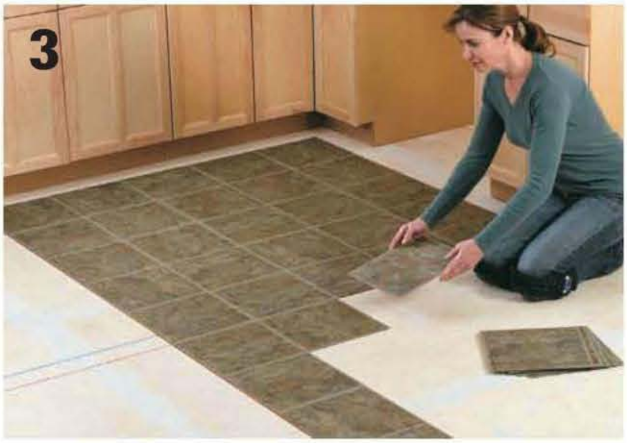

Finish setting full tiles in the first quadrant, then set the full tiles in an adjacent quadrant. Set the tiles along the layout lines first, then fill in the interior tiles.

Continue installing the tile in the remaining quadrants until the room is completely covered. Check the entire floor. If you find loose areas, press down on the tiles to bond them to the underlayment. Install metal threshold bars at room border where the new floor joins another floor covering.

Installing Bamboo Flooring

It looks like hardwood and is available in traditional tongue-and-groove form and in laminate planks. But bamboo is not wood. It's really a grass—and one of the most popular flooring materials today.

Bamboo flooring is made by shredding stalks of the raw material, then pressing them together with a resin that holds the shreds in their finished shape. Not only is bamboo a fast-growing and renewable crop, the companies that make bamboo flooring use binders with low emissions of volatile organic compounds (VOCs). The result is tough, economical, and ecologically friendly. In other words, it's just about perfect for flooring.

If you choose tongue-and-groove bamboo, the installation techniques are the same as for hardwoods. Bamboo is also available as a snap-fit laminate for use in floating floors. In this project we show Teragren Synergy Strand in Java (see Resources, page 554): thin, durable planks that are glued to the underlayment.

Tools & Materials ▸

Adhesive
Carpenter's level
Carpenter's square
Chalk line
Cleaning supplies
Flat-edged trowel
Marking pen or pencil
Measuring tape
Moisture level meter
Notched trowel
Rubber mallet
Scrap lumber
Shims
Straightedge
Weighted roller

Tips for a Successful Installation ▸

60°
70°
RECOMMENDED TEMPERATURE RANGE
40%
60%
RECOMMENDED HUMIDITY RANGE

Bamboo plank flooring should be one of the last items installed on any new construction or remodeling project. All work involving water or moisture should be completed before floor installation. Room temperature and humidity of installation area should be consistent with normal, year-round living conditions for at least a week before installation. Room temperature of 60 to 70° F and humidity range of 40 to 60% is recommended.

About radiant heat: The subfloor should never exceed 85° F. Check the manufacturer's suggested guidelines for correct water temperature inside heating pipes. Switch on the heating unit three days before flooring installation. Room temperature should not vary more than 15° F year-round. For glue-down installations, leave the heating unit on for three days following installation.

How to Install Bamboo Planks

SYNERGY
STRAND BAMBOO FLOORING
STRAND BAMBOO FLOORING

1

Give the bamboo time to adjust to installation conditions. Store it for at least 72 hours in or near the room where it will be installed. Open the packages for inspection, but do not store the planks on concrete or near outside walls.

2

Even though thin-plank bamboo is an engineered material, it can vary in appearance. Buy all planks from the same lot and batch number. Then visually inspect the planks to make sure they match. Use the same lighting as you will have in the finished room.

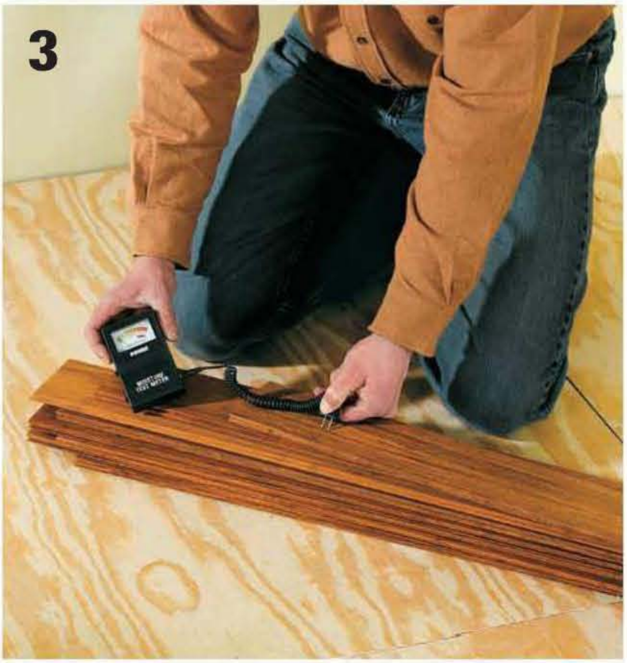

3

Inspect wood surfaces. The planks and underlayment should have no more than 12% moisture. Bamboo planks can be installed on plywood or oriented strandboard at least ¾" thick. The underlayment must be structurally sound.

4

Make sure the underlayment is level. It should not change by more than ⅛" over 10 feet. If necessary, apply a floor leveler to fill any low places, and sand down any high spots. Prevent squeaks by driving screws every 6" into the subfloor below.

(continued)

5

Sweep and vacuum the floor surface, then measure all room dimensions.

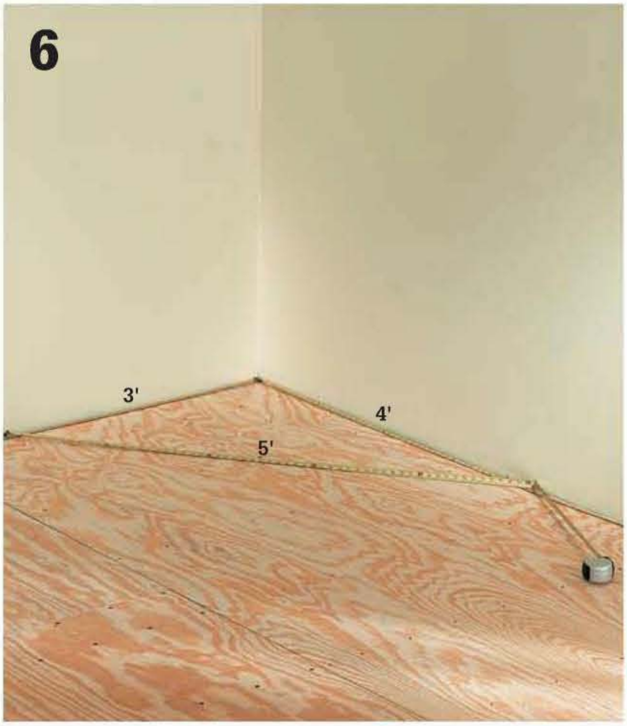

6

Check corners for squareness using the 3-4-5 triangle method.

7

The planks should be perpendicular to the floor joists below. Adjust your starting point if necessary. Snap a chalk line next to the longest wall. The distance from the wall should be the same at both ends, leaving ½" for expansion.

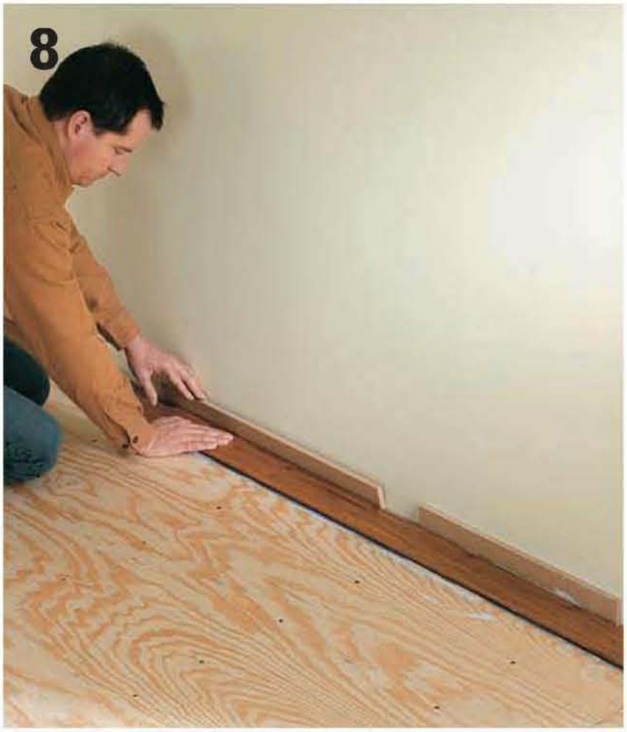

8

Lay the first course of planks with the tongue edge toward the wall. Align the planks with the chalk line. Hold the edge course in place with wedges or by nailing through the tongue edge. This row will anchor the others, so make sure it stays securely in place.

9

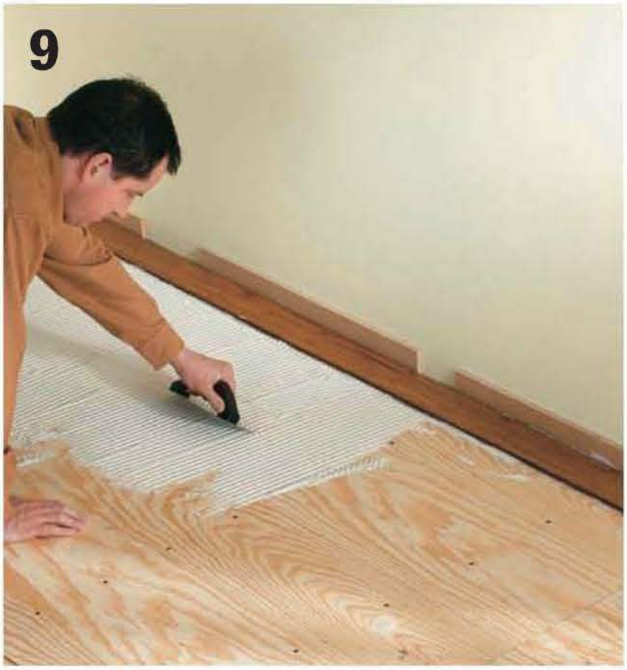

Once the starter row is in place, install the planks using a premium wood-flooring adhesive. Be sure to follow the manufacturer's instructions. Begin at the chalk line and apply enough adhesive to lay down one or two rows of planks. Spread the adhesive with a V-notched trowel at a 45° angle. Let the adhesive sit for the specified time.

10

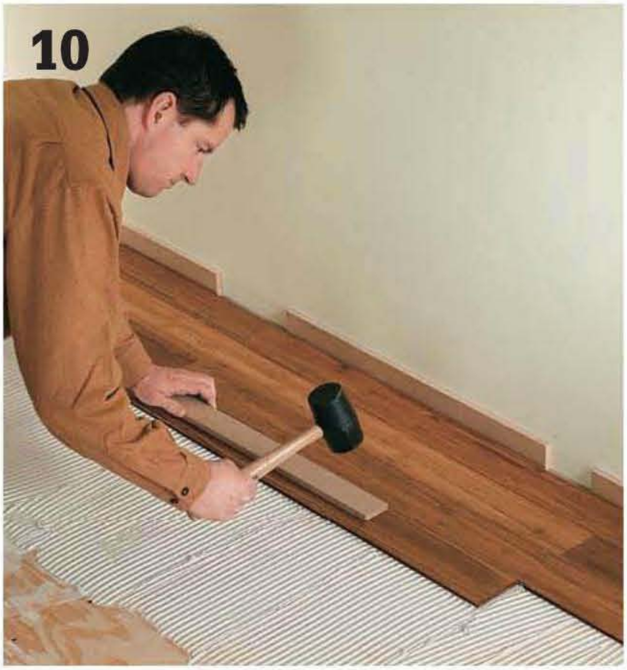

When the adhesive is tacky and ready to use, lay the first section of bamboo planks. Set each plank in the adhesive by placing a clean piece of scrap lumber on top and tapping it down with a rubber mallet. Check the edge of each section to make sure it keeps a straight line.

11

After you finish the first section, cover the next area with adhesive and give it time to become tacky. This slows down the project, but it prevents you from using more adhesive than you can use—and it allows the section you just finished to set up.

12

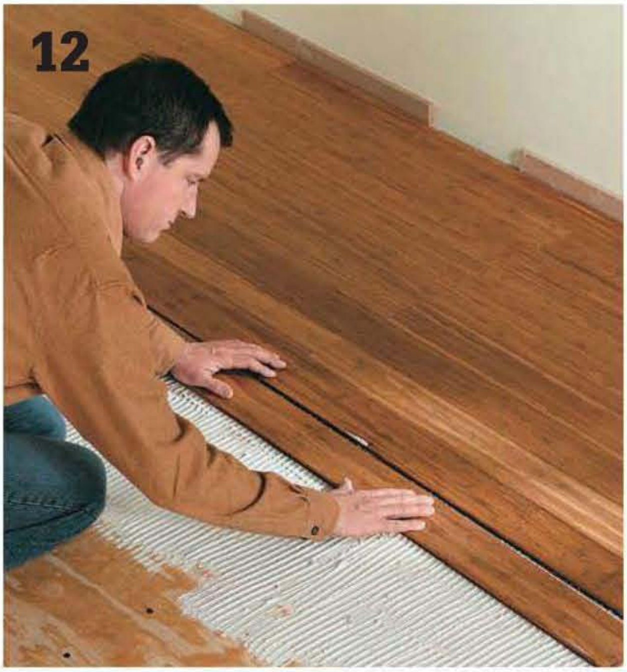

When the adhesive is ready, lay down the next section of planks. Fit the new planks tightly against the previous section, taking care not to knock the finished section out of alignment. If the planks have tongue-and-groove edges, fit them carefully into place.

(continued)

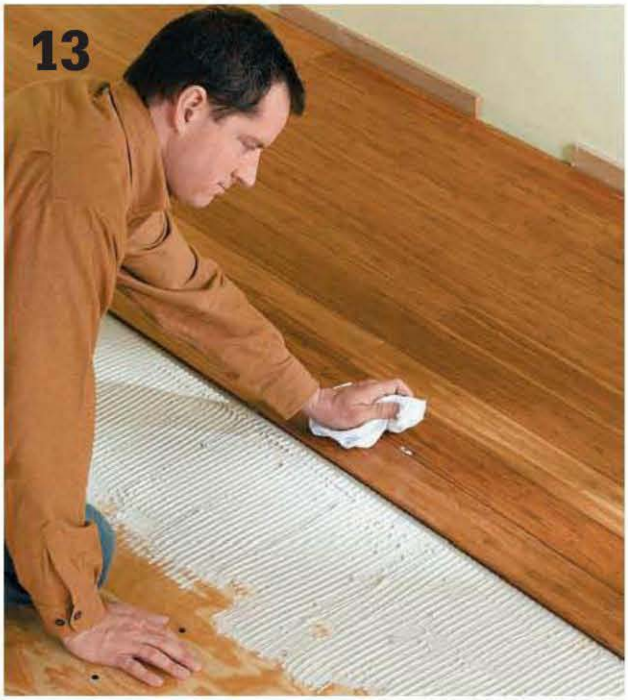

13

Continue applying adhesive and installing planks, one section at a time, to cover the entire floor. When adhesive gets on the flooring surface, wipe it off quickly.

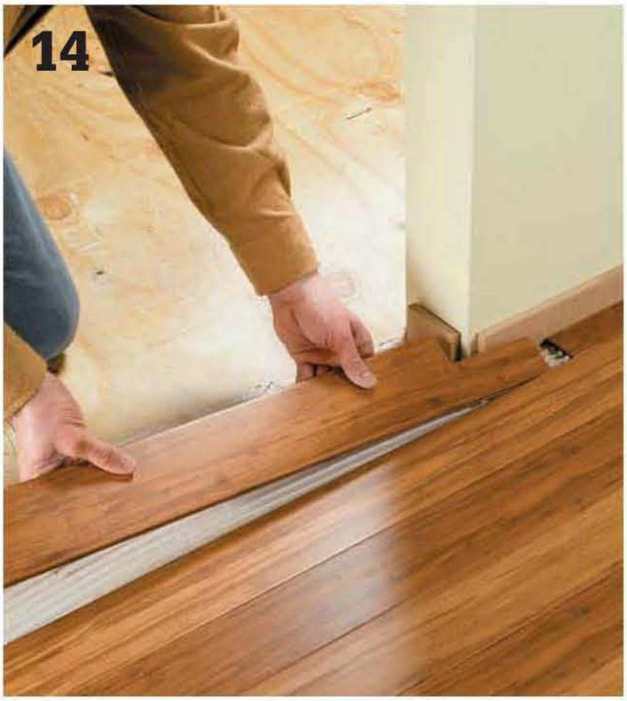

14

At the edges and around any fixed objects, such as doorways or plumbing pipes, leave a ½" gap for expansion. Use shims to maintain the gaps if needed. These spaces can be covered with baseboards, base shoes, and escutcheons.

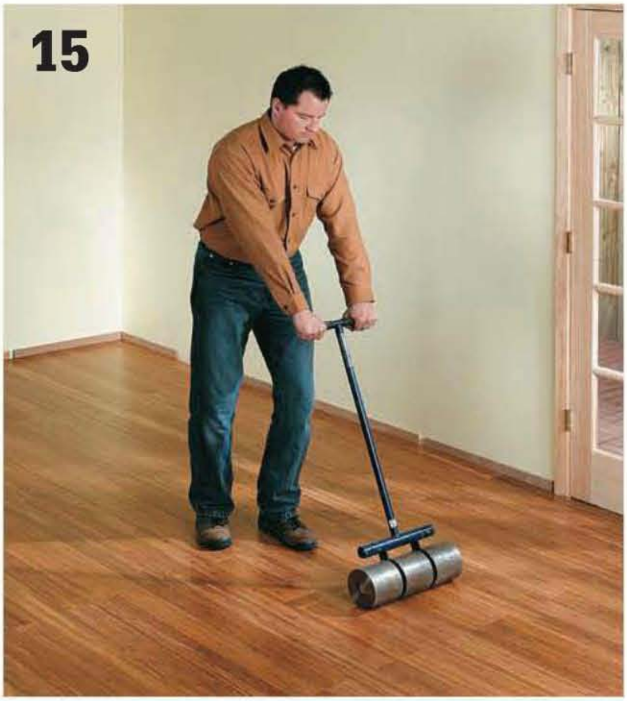

15

As you finish each section, walk across it a few times to maximize contact between the planks and the adhesive. When all the planks are in place, clean the surface and use a clean weighted roller. Push the roller in several directions, covering the entire surface many times.

16

In places that are difficult to reach with a roller, lay down a sheet of protective material, such as butcher paper, and stack weights on the paper. Let the finished floor sit for at least 24 hours, then clean the surface and remove any spacers from the expansion gaps. Finally, install the finishing trim.

Installing Ceramic Tile

Ceramic tile installation starts with the same steps as installing resilient tile. You snap perpendicular reference lines and dry-fit tiles to ensure the best placement.

When setting tiles, work in small sections so the mortar doesn't dry before the tiles are set. Use spacers between tiles to ensure consistent spacing. Plan an installation sequence to avoid kneeling on set tiles. Be careful not to kneel or walk on tiles until the designated drying period is over.

Tools & Materials ▸

- ¼" square trowel
- Rubber mallet
- Tile cutter
- Tile nippers
- Hand-held tile cutter
- Needlenose pliers
- Grout float
- Grout sponge
- Soft cloth
- Small paint brush
- Thin-set mortar
- Tile
- Tile spacers
- Grout
- Latex grout additive
- Wall adhesive
- 2 × 4 lumber
- Grout sealer
- Tile caulk
- Sponge brush

Ceramic floor tile continues to be one of the most popular DIY floorcoverings. It is relatively easy to install, is very durable and, depending on the style you choose, can be quite affordable.

How to Install Ceramic Tile

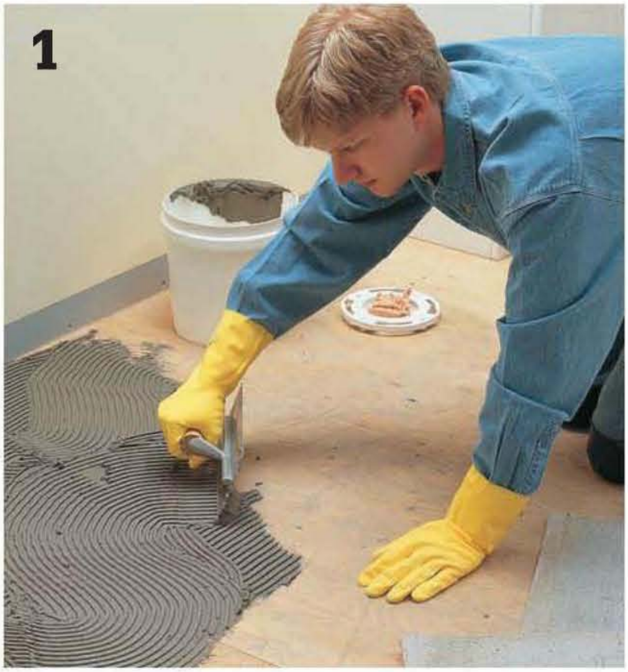

Make sure the subfloor is smooth, level, and stable. Spread thin-set mortar on the subfloor for one sheet of cementboard. Place the cementboard on the mortar, keeping a ¼" gap along the walls.

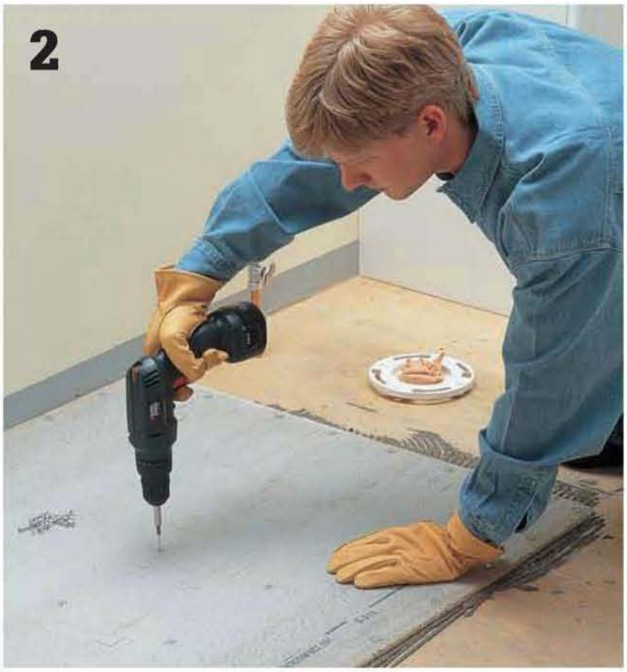

Fasten it in place with 1¼" cementboard screws. Place fiberglass-mesh wallboard tape over the seams. Cover the remainder of the floor, following the steps on page 152.

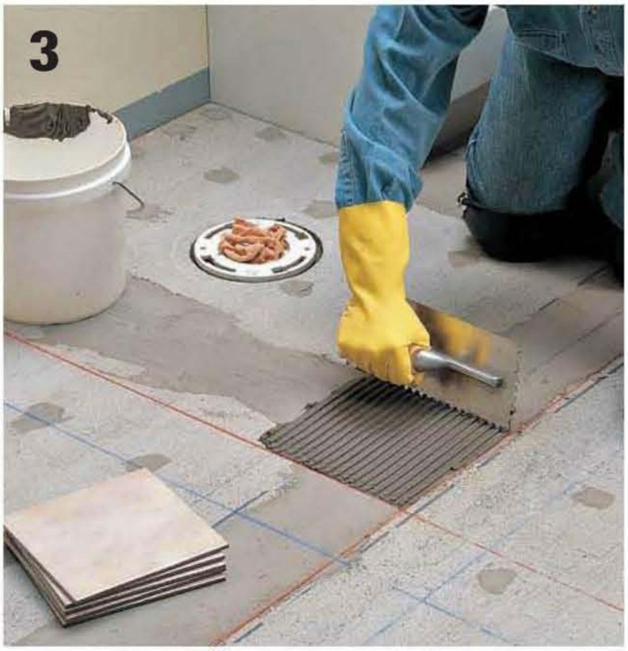

Draw reference lines and establish the tile layout (see page 181). Mix a batch of thin-set mortar, then spread the mortar evenly against both reference lines of one quadrant, using a ¼" square-notched trowel. Use the notched edge of the trowel to create furrows in the mortar bed.

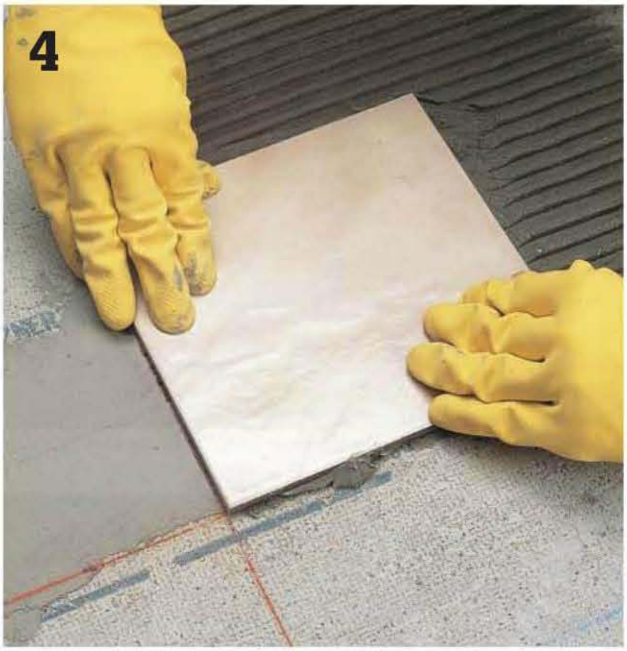

Set the first tile in the corner of the quadrant where the reference lines intersect. When setting tiles that are 8" square or larger, twist each tile slightly as you set it into position.

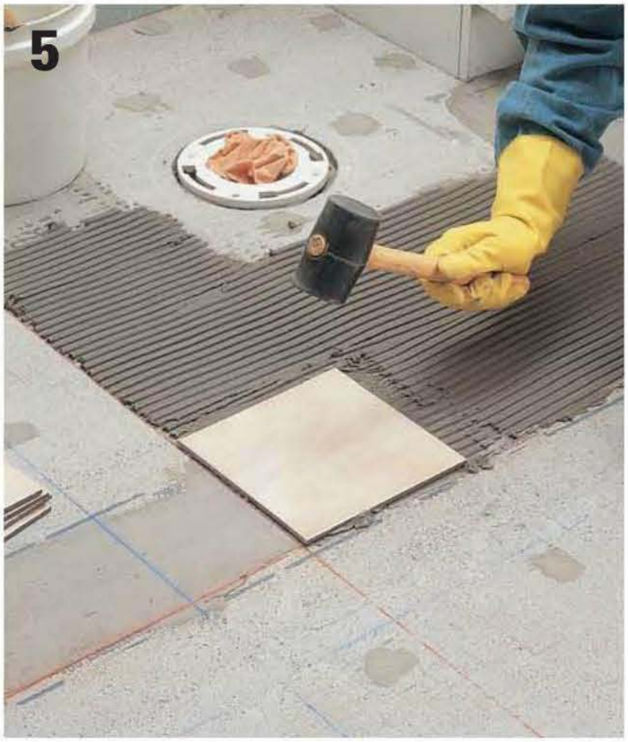

Using a soft rubber mallet, gently tap the central area of each tile a few times to set it evenly into the mortar.

Variation: For large tiles or uneven stone, use a larger trowel with notches that are at least ½" deep.

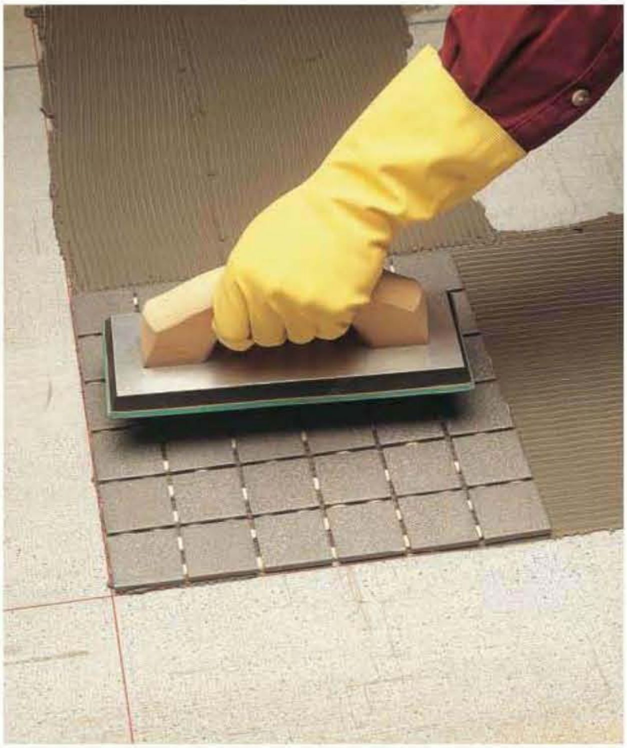

Variation: For mosaic sheets, use a $\frac{3}{16}$" V-notched trowel to spread the mortar and a grout float to press the sheets into the mortar. Apply pressure gently to avoid creating an uneven surface.

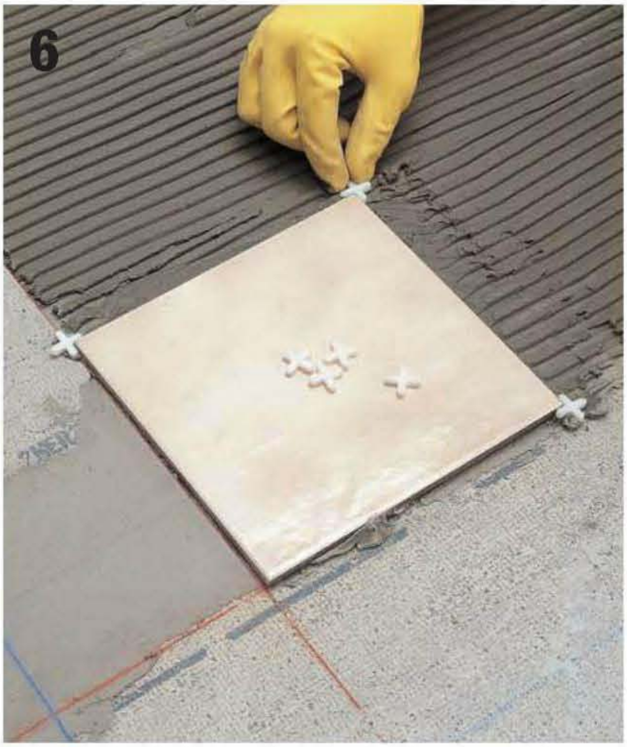

To ensure consistent spacing between tiles, place plastic tile spacers at the corners of the set tile. With mosaic sheets, use spacers equal to the gaps between tiles.

(continued)

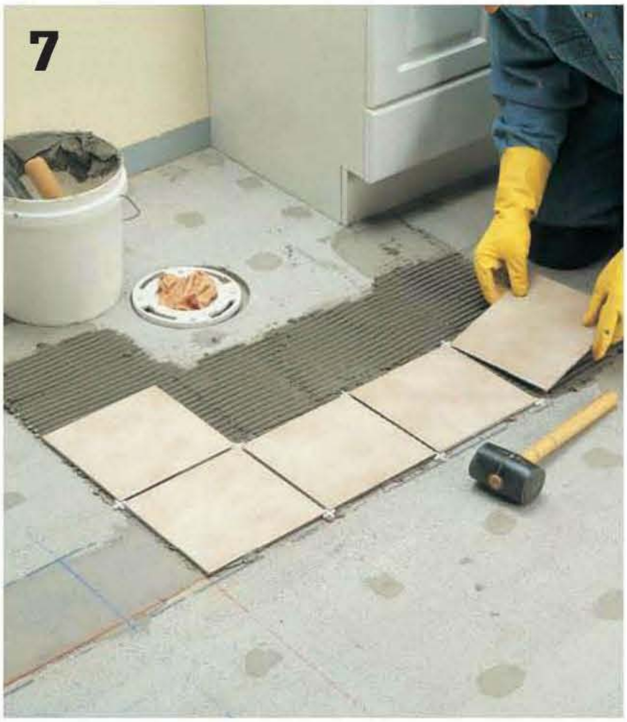

Position and set adjacent tiles into the mortar along the reference lines. Make sure the tiles fit neatly against the spacers.

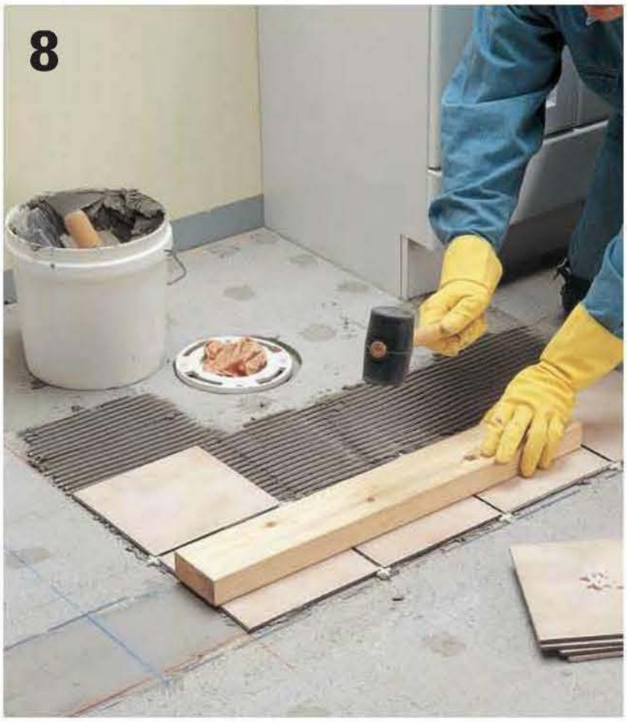

To make sure the tiles are level with one another, place a straight piece of 2 × 4 across several tiles, then tap the board with a mallet.

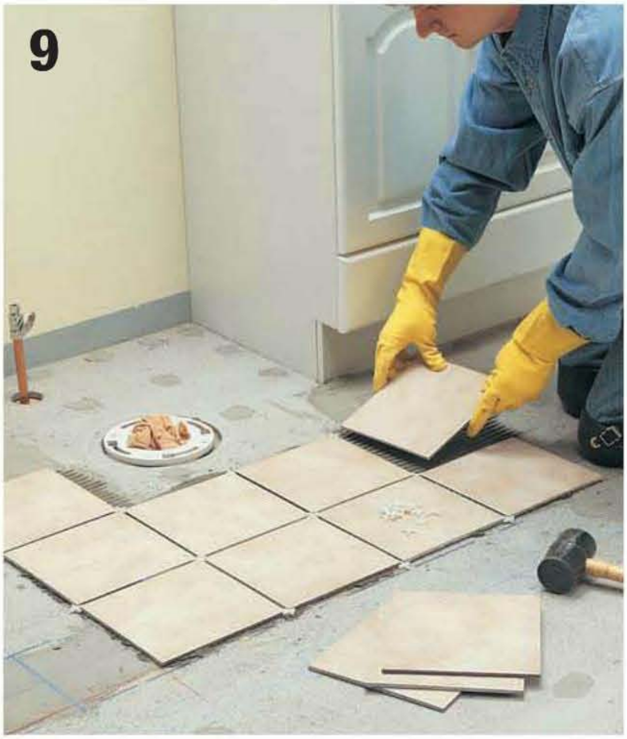

Lay tile in the remaining area covered with mortar. Repeat steps 2 to 7, continuing to work in small sections, until you reach walls or fixtures.

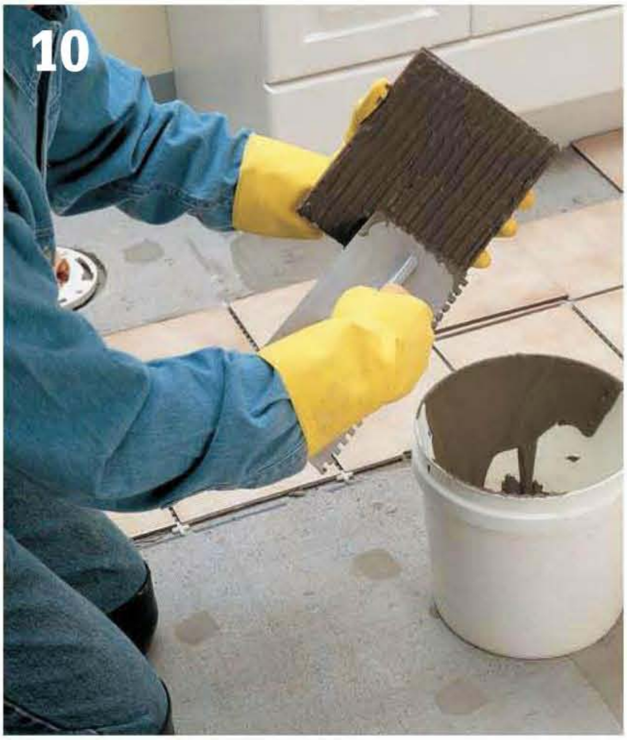

Measure and mark tiles to fit against walls and into corners (see pages 183 to 184). Cut the tiles to fit. Apply thin-set mortar directly to the back of the cut tiles, instead of the floor, using the notched edge of the trowel to furrow the mortar

11

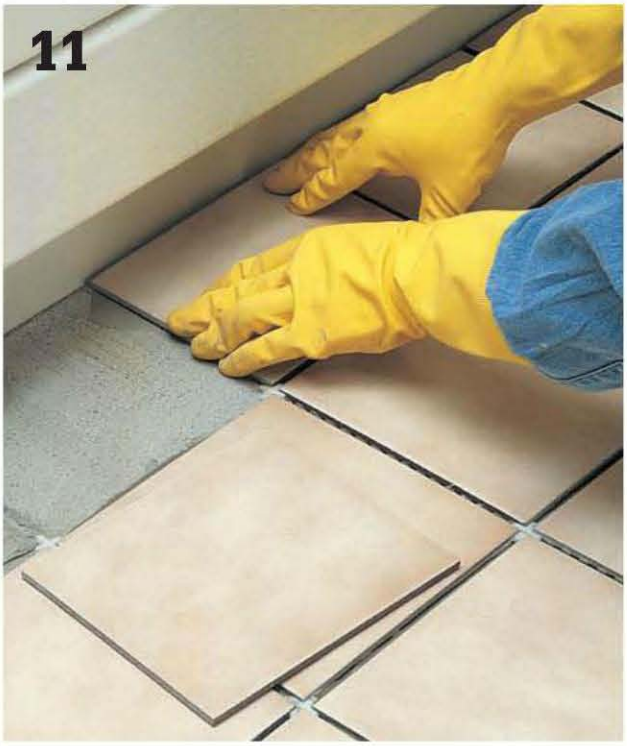

Set the cut pieces of tile into position. Press down on the tile until each piece is level with adjacent tiles.

12

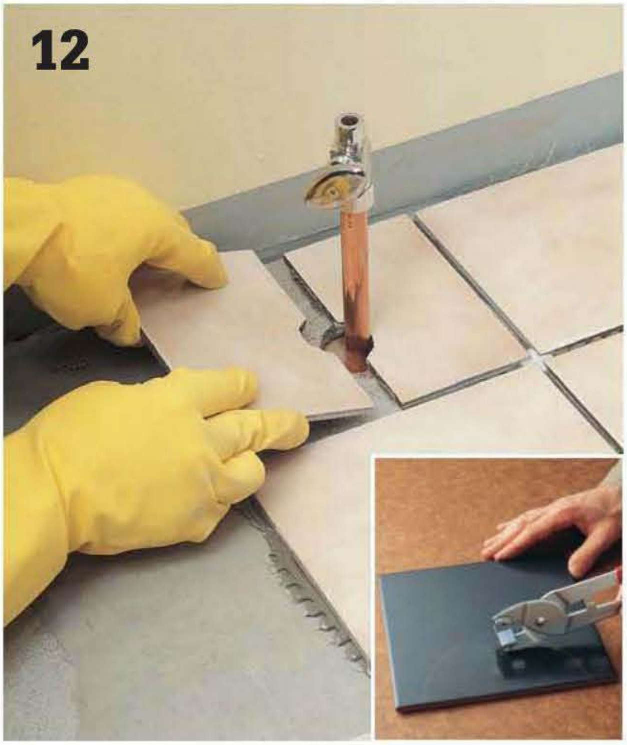

Measure, cut, and install tiles that require notches or curves to fit around obstacles (inset), such as exposed pipes or toilet drains.

13

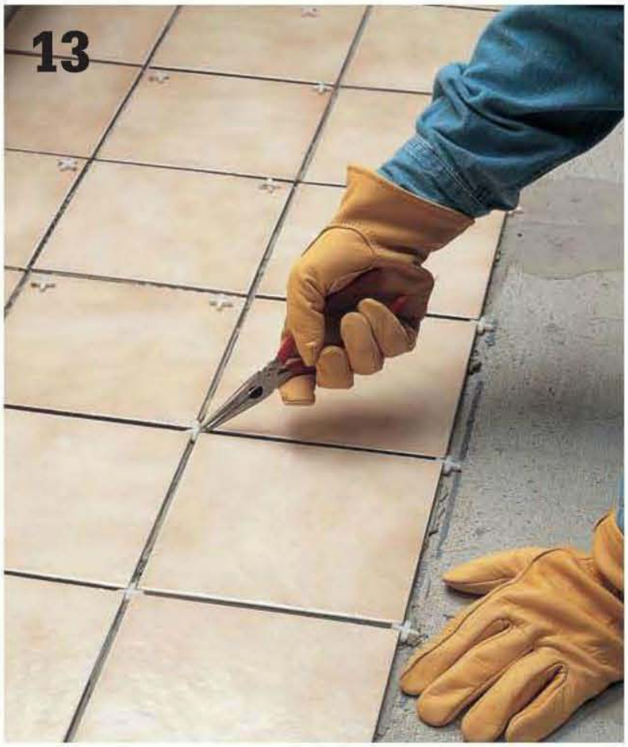

Carefully remove the spacers with needlenose pliers before the mortar hardens.

14

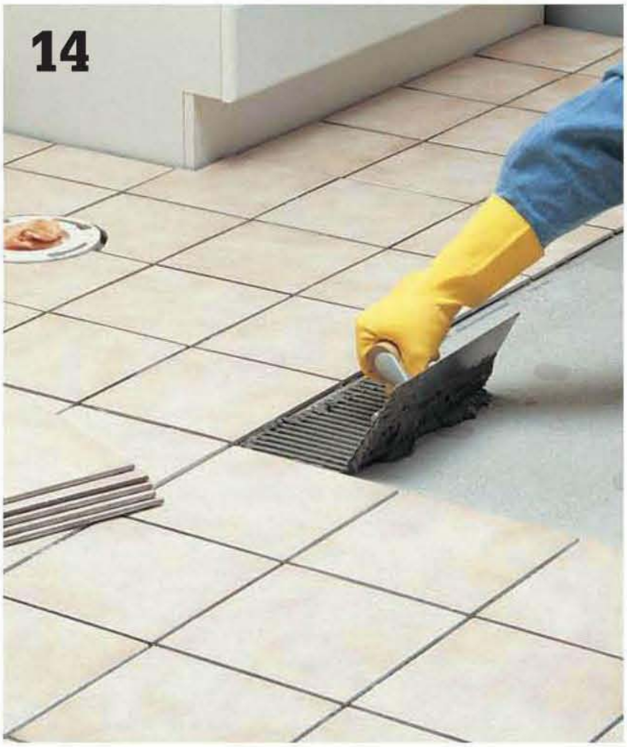

Apply mortar and set tiles in the remaining quadrants, completing one quadrant before starting the next. Inspect all of the tile joints and use a utility knife or grout knife to remove any high spots of mortar that could show through the grout.

(continued)

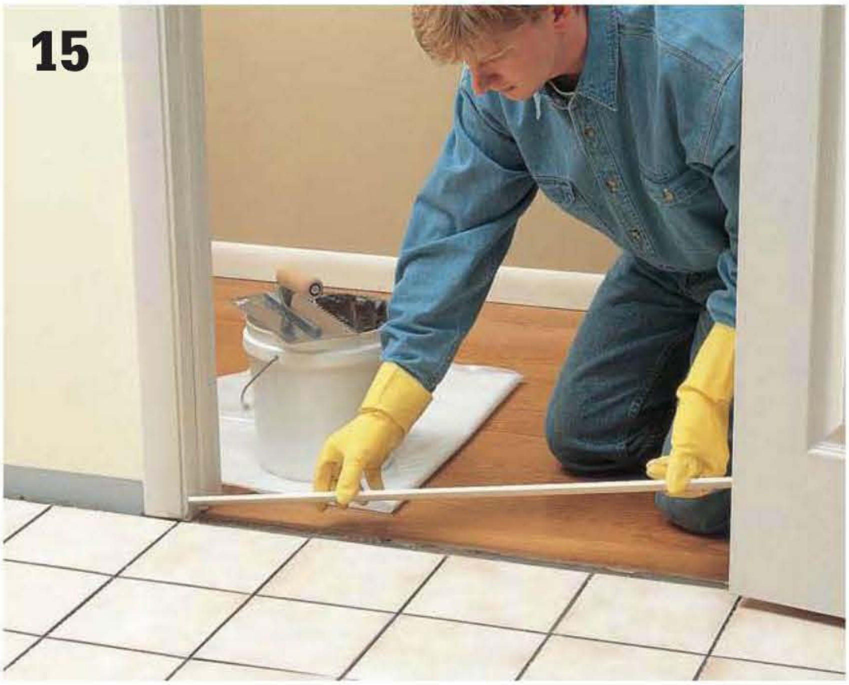

Install threshold material in doorways. If the threshold is too long for the doorway, cut it to fit with a jigsaw or circular saw and a tungsten-carbide blade. Set the threshold in thin-set mortar so the top is even with the tile. Keep the same space between the threshold as between tiles. Let the mortar set for at least 24 hours.

16
LATICRETE
TEC
sanded
floor grout

Prepare a small batch of floor grout to fill the tile joints. When mixing grout for porous tile, such as quarry or natural stone, use an additive with a release agent to prevent grout from bonding to the tile surfaces.

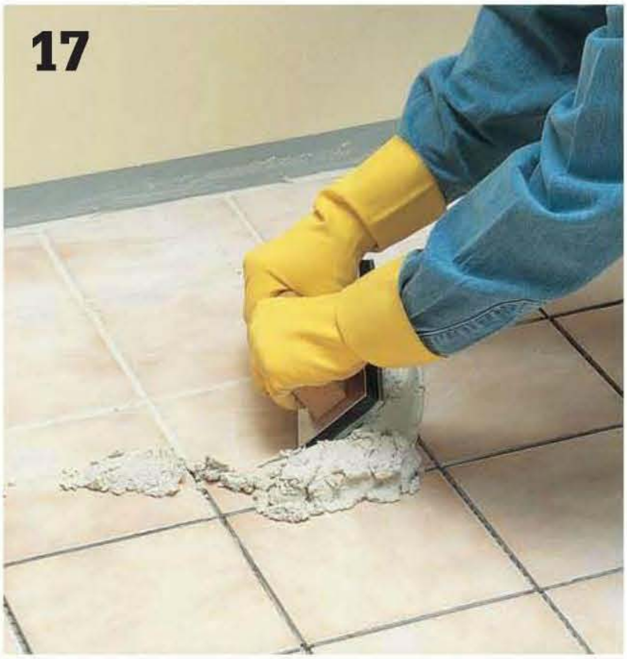

Starting in a corner, pour the grout over the tile. Use a rubber grout float to spread the grout outward from the corner, pressing firmly on the float to completely fill the joints. For best results, tilt the float at a 60° angle to the floor and use a figure eight motion.

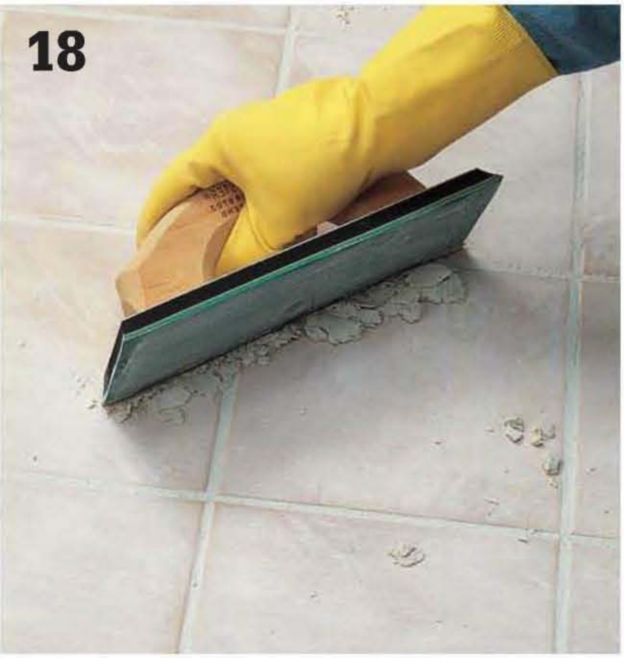

Use the grout float to remove excess grout from the surface of the tile. Wipe diagonally across the joints, holding the float in a near-vertical position. Continue applying grout and wiping off excess until about 25 square feet of the floor has been grouted.

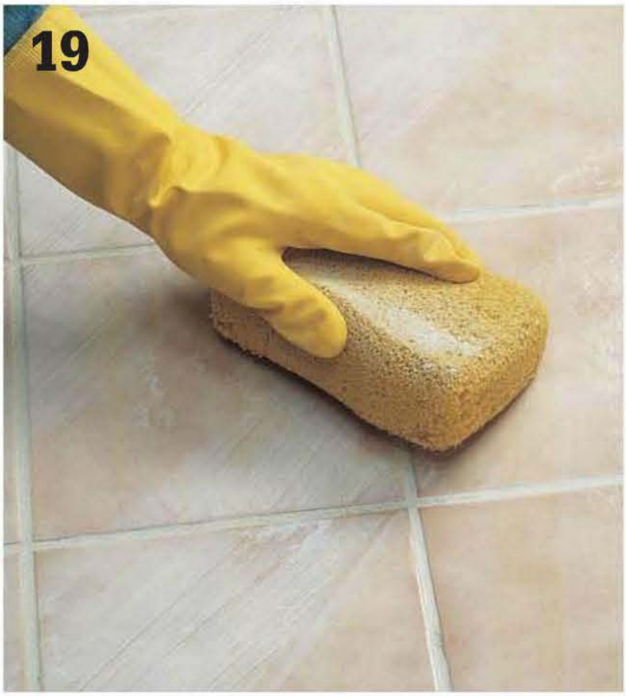

Wipe a damp grout sponge diagonally over about 2 square feet of the floor at a time. Rinse the sponge in cool water between wipes. Wipe each area only once since repeated wiping can pull grout back out of joints. Repeat steps 17 to 18 to apply grout to remainder of floor.

Allow the grout to dry for about 4 hours, then use a soft cloth to buff the tile surface and remove any remaining grout film.

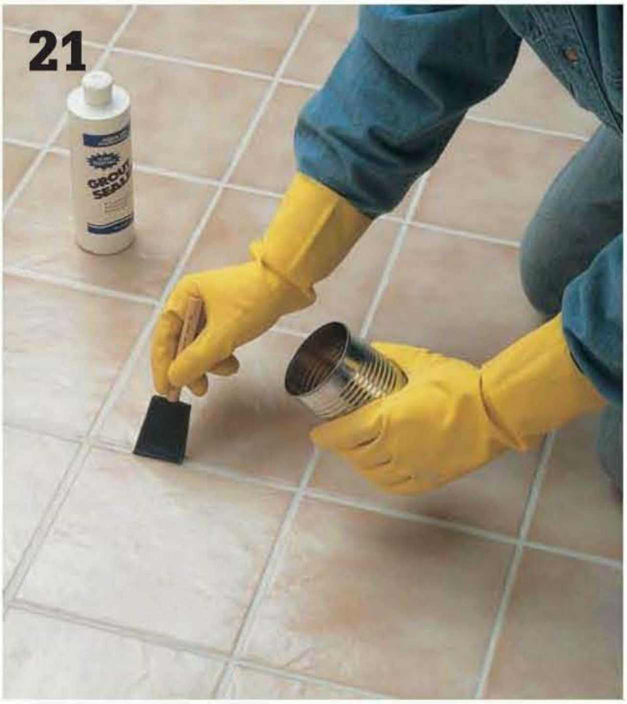

Apply grout sealer to the grout lines, using a small sponge brush or sash brush. Avoid brushing sealer on to the tile surfaces. Wipe up any excess sealer immediately.

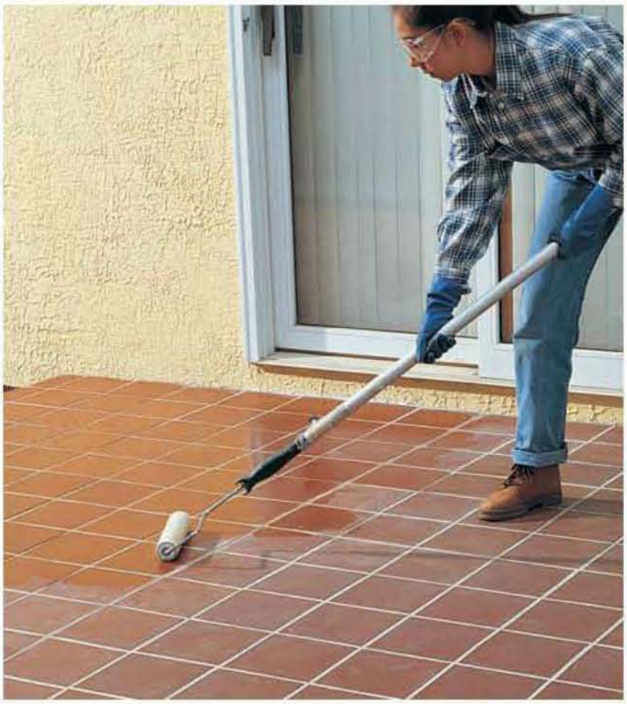

Variation: Use a tile sealer to seal porous tile, such as quarry tile or unglazed tile. Following the manufacturer's instructions, roll a thin coat of sealer over the tile and grout joints, using a paint roller and extension handle.

Installing Mosaic Tile

Mosaic tile is an excellent choice for smaller areas. It requires the same preparation and handling as larger tiles, with a few differences. Sheets of mosaic tile are held together by a fabric mesh backing. This makes them more difficult to hold, place, and move. They may not be square with your guidelines when you first lay them down. And mosaic tiles require many more temporary spacers and much more grout.

A few cautions: Variations in color and texture are just as likely among mosaic tile as individual tiles, so buy all your tile from the same lot and batch. Mortar or mastic intended for ceramic tile may not work with glass mosaic tile. Finally, in projects where the finished project will be exposed to the elements, make sure you have adhesive and grout suitable for outdoor use.

Tools & Materials ▸

- Carpenter's square
- Chalk line
- Cleaning supplies
- Coarse sponge
- Craft/utility knife
- Grout sealer
- Marking pen or pencil
- Measuring tape
- Notched trowel
- Recommended adhesive
- Rubber mallet
- Sanded grout
- Scrap lumber
- Straightedge
- Tile nippers
- Tile spacers

Mosaic tiles come in sheets (usually 12 × 12") and can be made from ceramic, porcelain, glass or any number of designer materials. Normally installed for their appearance, mosaics are relatively high maintenance and prone to cracks because of all the grout lines.

How to Install Mosaic Tile

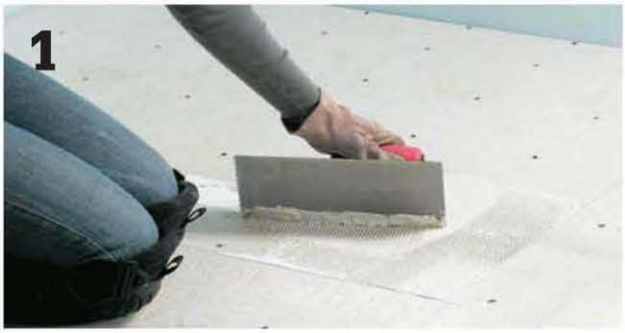

Clean and prepare the area and then draw reference lines (see page 181). Beginning at the center intersection, apply the recommended adhesive to one quadrant. Spread it outward evenly with a notched trowel. Lay down only as much adhesive as you can cover in 10 to 15 minutes.

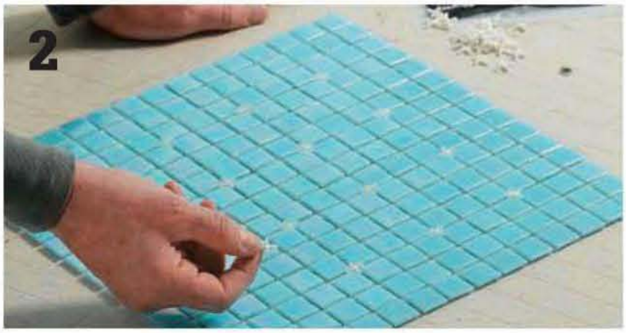

Select a sheet of mosaic tile. Place several plastic spacers within the grid so that the sheet remains square. Pick up the sheet of tiles by diagonally opposite corners. This will help you hold the edges up so that you don't trap empty space in the middle of the sheet.

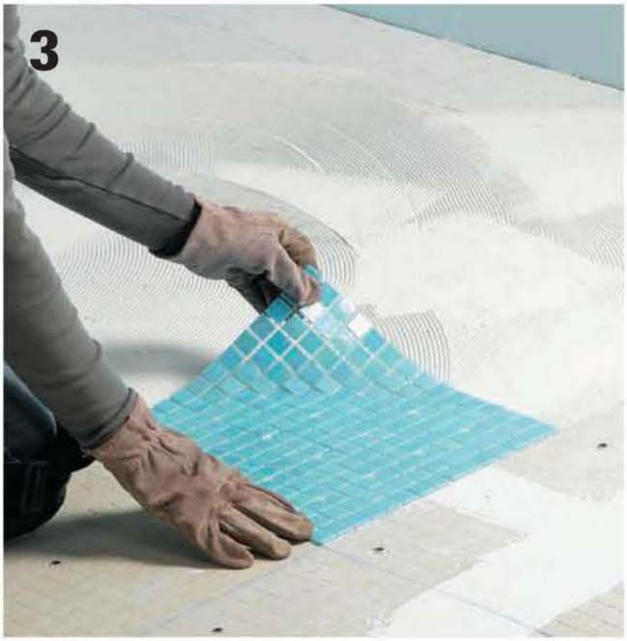

Gently press one corner into place on the adhesive. Slowly lower the opposite corner, making sure the sides remain square with your reference lines. Massage the sheet into the adhesive, being careful not to press too hard or twist the sheet out of position. Insert a few spacers in the outside edges of the sheet you have just placed. This will help keep the grout lines consistent.

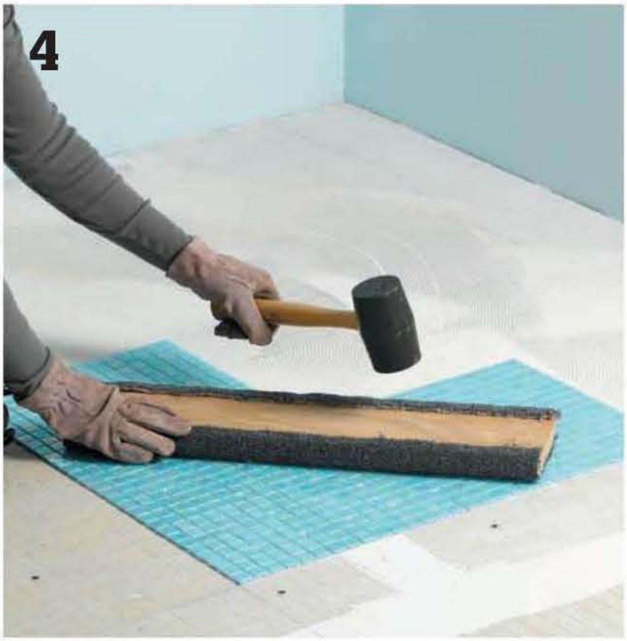

When you have placed two or three sheets, lay a scrap piece of flat lumber across the tops and tap the wood with a rubber mallet to set the fabric mesh in the adhesive, and to force out any trapped air.

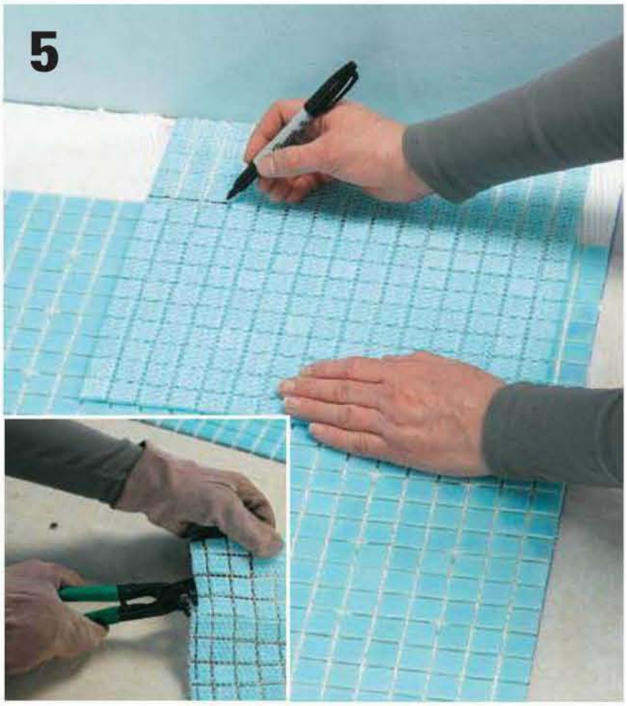

At the outer edges of your work area, you will probably need to trim one or more rows from the last sheet. If the space left at the edge is more than the width of a regular grout line, use tile nippers to trim the last row that will fit. Save these leftover tiles for repairs.

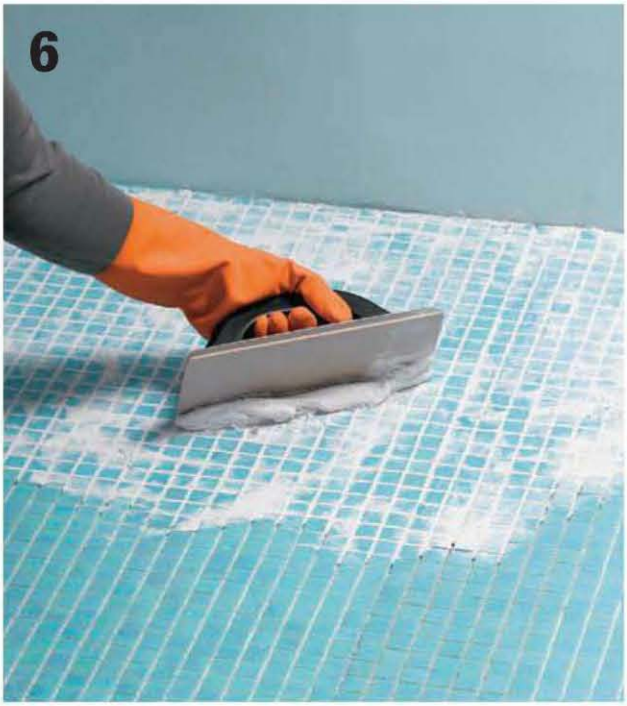

After the adhesive has cured, usually 24 to 48 hours, apply grout (see pages 196 to 197). With many more spaces, mosaic tiles will require more grout. Follow the manufacturer's instructions for spreading and floating the grout. Clean up using the instructions for individual tiles (see page 197).

Installing Carpet Squares

Most carpeting has a single design and is stretched from wall to wall. It covers more square feet of American homes than any other material. But if you want a soft floor covering that gives you more options, carpet squares are an excellent choice.

Manufacturers have found ways to create attractive new carpet using recycled fibers. This not only reuses material that would otherwise become landfill, it reduces waste in manufacturing as well. So, instead of adding to problems of resource consumption and pollution, carpet squares made from recycled materials help reduce them.

The squares are attached to each other and to the floor with adhesive dots. They can be installed on most clean, level, dry underlayment or existing floor. If the surface underneath is waxed or varnished, check with the manufacturer before you use any adhesives on it.

Tools & Materials ▸

- Adhesive
- Aviator's snips
- Carpenter's square
- Chalk line
- Cleaning supplies
- Craft/utility knife
- Flat-edged trowel
- Marking pen or pencil
- Measuring tape
- Notched trowel
- Straightedge

Carpet tiles combine the warmth and comfort of carpet with do-it-yourself installation, custom designs, and easy replacement. They can be laid wall-to-wall or in an area rug style, as shown above.

How to Install Carpet Squares

Take the squares out of the package. Be sure the room is well ventilated. Carpet squares should be at room temperature for at least 12 hours before you lay them down.

2
FLOR
FLOR
FLOR

Check the requirements for the recommended adhesive. You can install carpet squares over many other flooring materials, including hardwood, laminates, and resilient sheets or tiles. The carpet squares shown here are fastened with adhesive dots, so almost any existing floor provides a usable surface.

Make sure the existing floor is clean, smooth, stable, and dry. Use floor leveler if necessary to eliminate any hills or valleys. If any part of the floor is loose, secure it to the subfloor or underlayment before you install the carpet squares. Vacuum the surface and wipe it with a damp cloth.

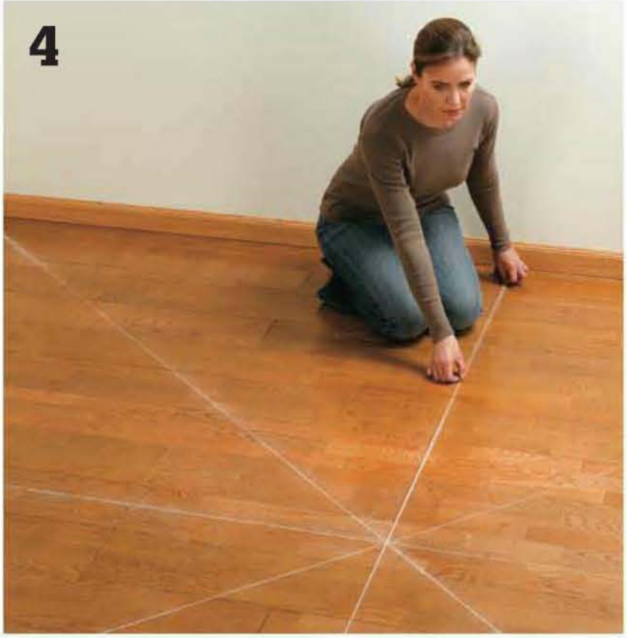

Snap chalk lines between diagonally opposite corners to find the center point for the room. In rooms with unusual shapes, determine the visual center and mark it. Next, snap chalk lines across the center and perpendicular to the walls. This set of guidelines will show you where to start.

(continued)

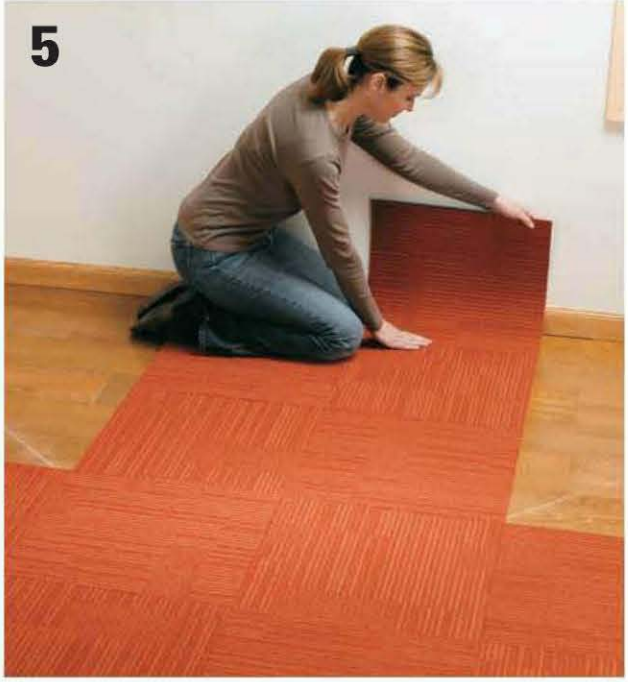

Lay a base row of carpet squares on each side of the two guidelines. When you reach the walls, make note of how much you will need to cut. You should have the same amount to cut on each side. If not, adjust the center point and realign the squares.

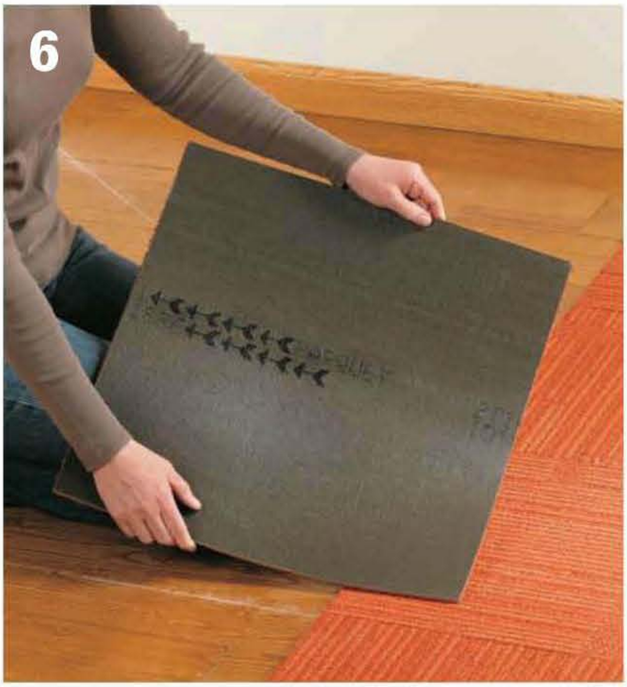

Check the backs of the squares before you apply any adhesive. They should indicate a direction, using arrows or other marks, so that the finished pile has a consistent appearance. If you plan to mix colors, this is the time to establish your pattern.

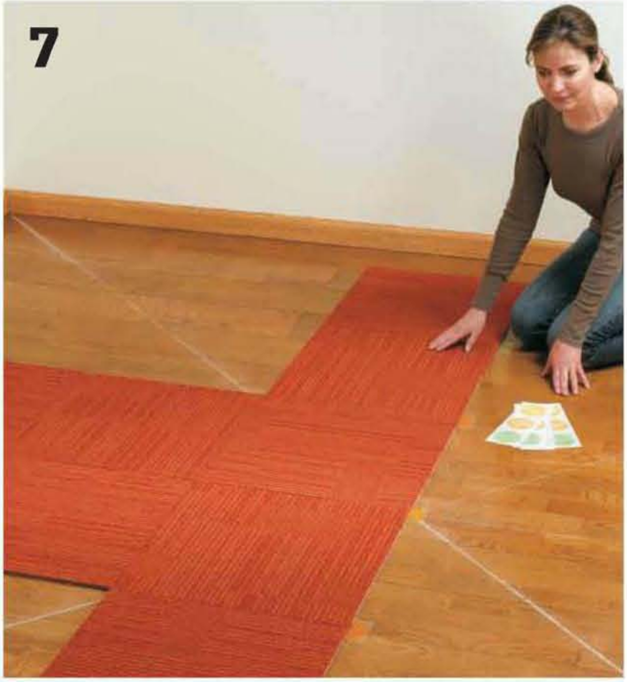

Fasten the base rows in place using the manufacturer's recommended adhesive. This installation calls for two adhesive dots per square. As you place each square, make sure it is aligned with the guidelines and fits tightly against the next square.

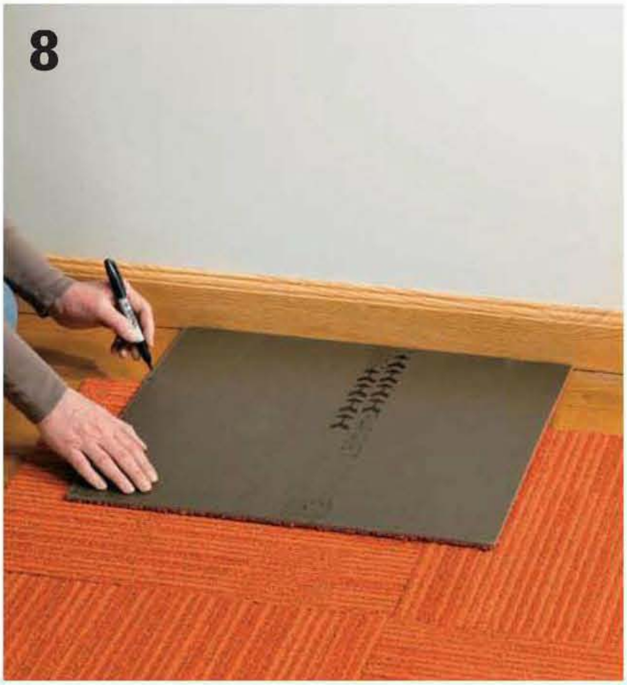

When you reach a wall, flip the last square over. Push it against the wall until it is snug. If you are planning a continuous pattern, align the arrows with the existing squares. If you are creating a parquet pattern, turn the new square 90 degrees before marking it.

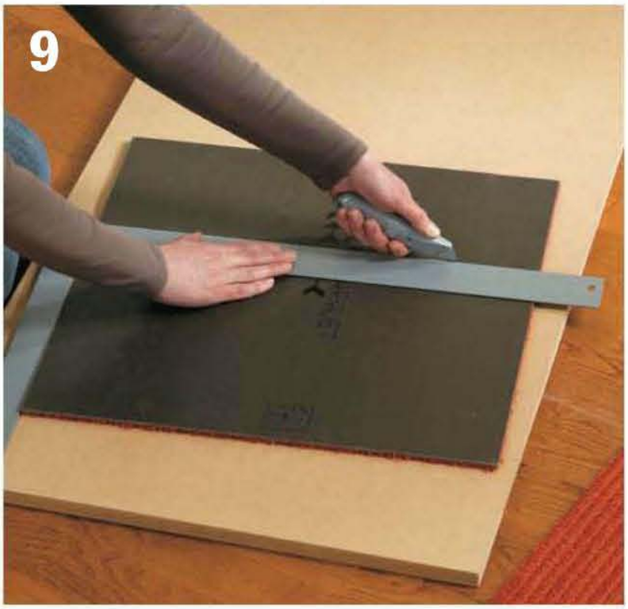

9

Mark notches or draw a line across the back where the new square overlaps the next-to-last one. Using a sharp carpet knife, a carpenter's square, and a tough work surface, cut along this line. The cut square should fit neatly in the remaining space.

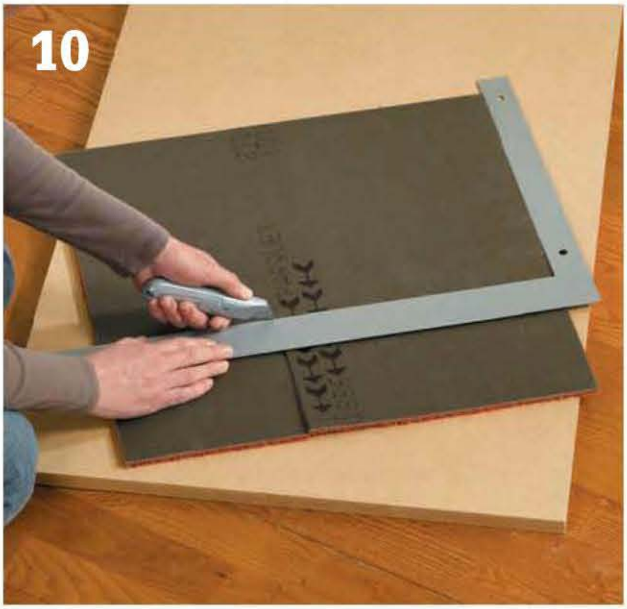

10

At a door jamb, place a square face up where it will go. Lean the square against the jamb and mark the point where they meet. Move the square to find the other cutline, and mark that as well. Flip the square over, mark the two lines using a carpenter's square, and cut out the corner.

11

Finish all four base rows before you fill in the rest of the room. As you work, check the alignment of each row. If you notice a row going out of line, find the point where the direction changed, then remove squares back to that point and start again.

12

Work outward from the center so that you have a known reference for keeping rows straight. Save the cut pieces from the ends. They may be useful for patching odd spaces around doorways, heat registers, radiator pipes, and when you reach the corners.

Refinishing Wood Floors

Refinishing hardwood floors is one of the first major home improvement projects homeowners attempt. The tools you need are readily available for rent, but perhaps the main reason so many DIYers attempt floor refinishing is that it is a very invasive project that requires a lot of planning and moving of furniture. Doing it yourself allows you to work gradually on your own schedule.

Tools & Materials ▸

Staple gun
Zip door
Painter's tape
Plastic
Fan
Pry bar
2 × 4 scrap
Wood shims
Nail set
Drum sander
Sandpaper
Power edge sander
Rotary buffer
Paint scraper
Sanding block
Tack cloth

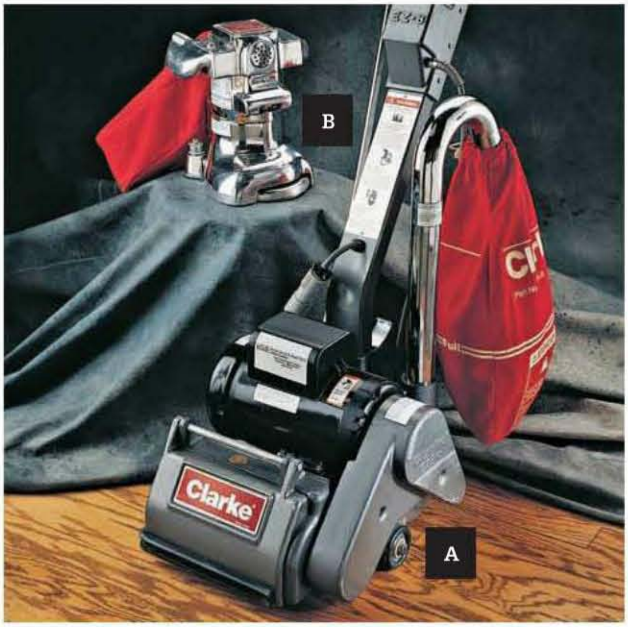

A drum sander (A) and an edge sander (B) are rental tools that can tackle just about any floorsanding project.

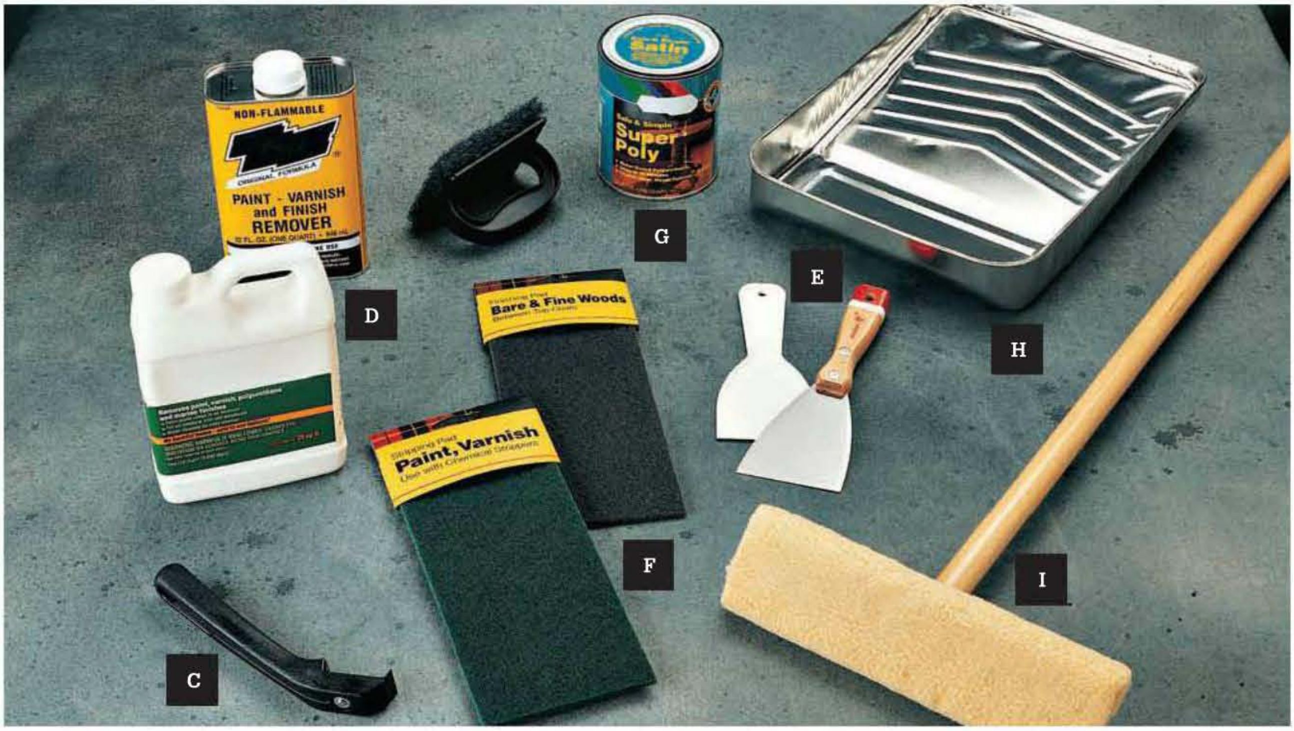

Other products and tools for resurfacing and refinishing floors: Paint scrapers (C) are helpful for removing old finish in corners and other areas that cannot be reached by sanders. When refinishing floors, chemical stripping products (D) are often a more efficient method that yields better results. This is especially true for floors that are uneven, or for parquet and veneered floors, which cannot be sanded. Stripper knives (E) and abrasive pads (F) are used with the stripping products. For the final finish, water-based polyurethane (G) is poured into a paint tray (H) and applied using a wide painting pad with a pole extension (I).

How to Refinish Wood Floors

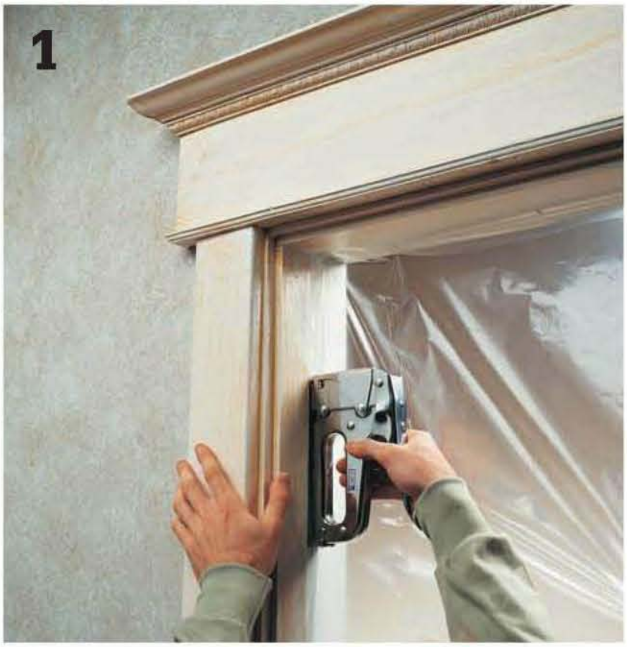
1

Staple plastic on all doorways. Place a zip door over the entryway you plan to use for the duration of the project. Use painter's tape and plastic to cover heating and cooling registers, ceiling fans, and light fixtures. Finally, place a fan in a nearby window to blow the circulating dust outside.

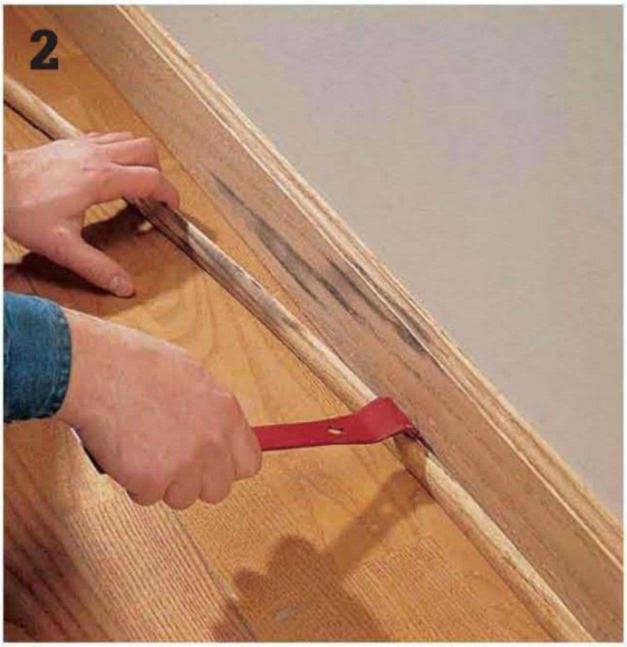
2

Wedge a prybar between the shoe molding and baseboards. Move along the wall as nails loosen. Once removed, place a scrap 2 × 4 board against the wall and, with a pry bar, pry out baseboard at nails. Maintain the gap with wood shims. Drive protruding nails in floor ⅛" below the surface with a nail set.

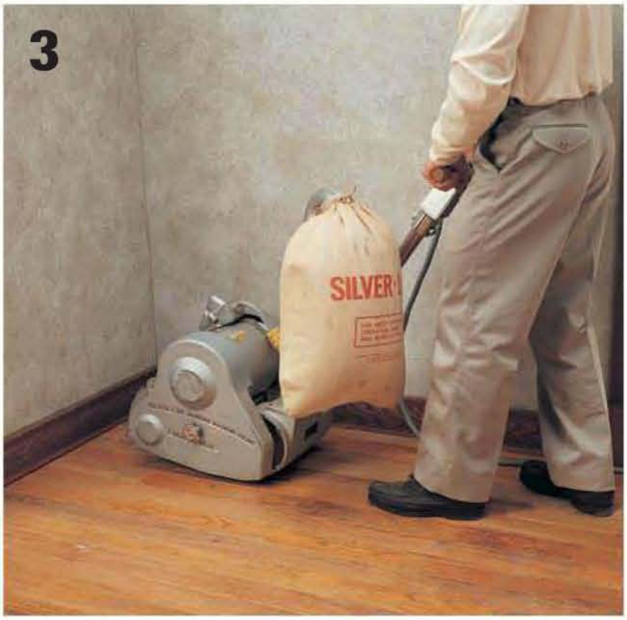
3

Practice with the drum sander turned off. Move forward and backward; tilt or raise it off the floor a couple of times. A drum sander is difficult to maneuver. Once it touches the floor it walks forward; if you stop it, it gouges the floor.

4

For the initial pass with the drum sander, sand with the grain, using 40- or 60-grit sandpaper. For large scratches use 20 or 30. Start two-thirds down the room length on the right side; work your way to the left. Raise drum. Start motor. Slowly lower drum to the floor. Lift the sander off the floor as you approach the wall. Move to the left 2 to 4" and then walk it backwards the same distance you just walked forward. Repeat.

(continued)

5

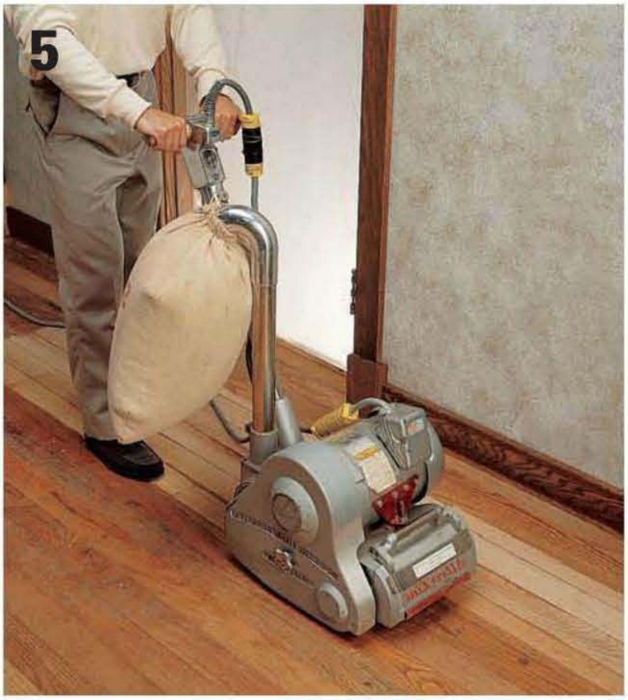

When you get to the far left side of the room, turn the machine around and repeat the process. Overlap the sanded two-thirds to feather out the ridgeline. Repeat this drum sanding process 3 or 4 times using 120-grit paper. For the final pass or two use finer paper (150-grit).

6

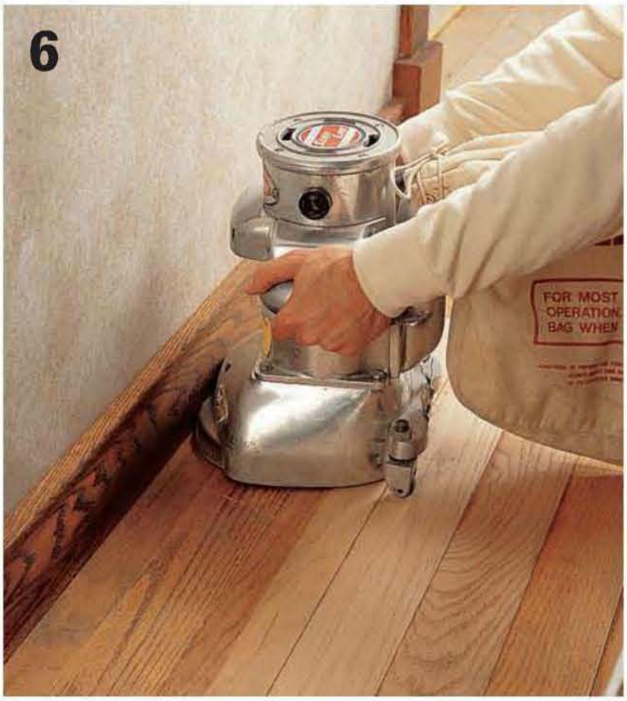

Use a power edge sander along the walls, using the same grit that you last used with the drum sander. Make a succession of overlapping half-circles as you move along the entire perimeter of the room. Next, run a rotary buffer over the floor twice: first with an 80-grit screen and then with a 100-grit screen. Finally, use a random orbital sander to smooth out the floor.

7

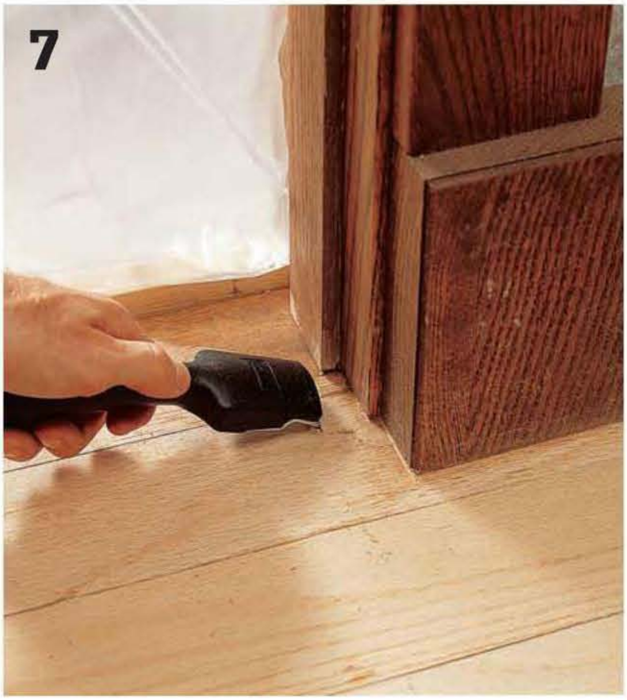

Use a paint scraper to get to corners and hard-to-reach nooks. Pull the scraper toward you with a steady downward pressure. Pull with the grain. Next, sand with a sanding block.

8

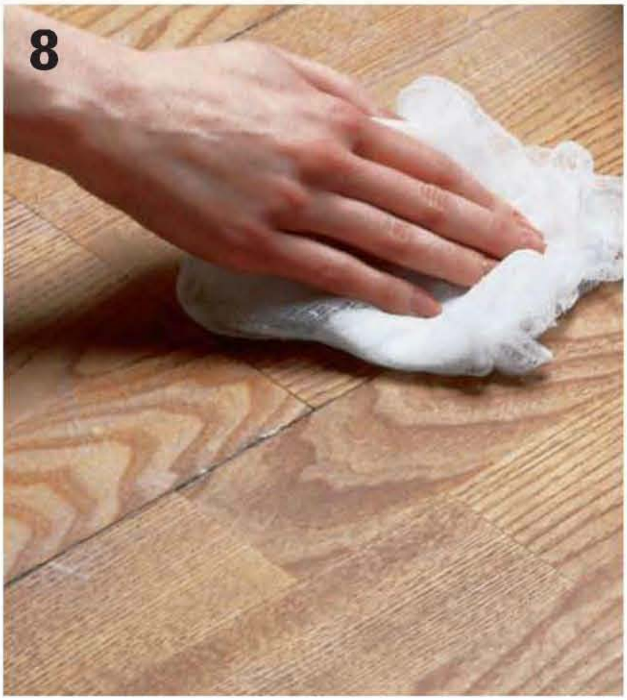

Remove all plastic on doors, windows, and fixtures and then sweep and vacuum to prepare the room for finish. Wipe up fine particles with a tack cloth.

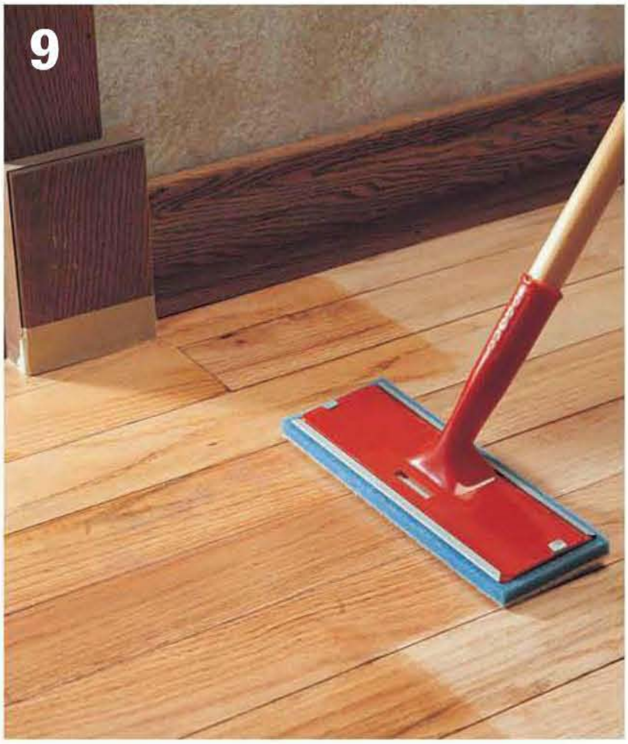

Seal the sanded wood with a 1:1 mixture of water-based polyurethane and water, using a painting pad and pole.

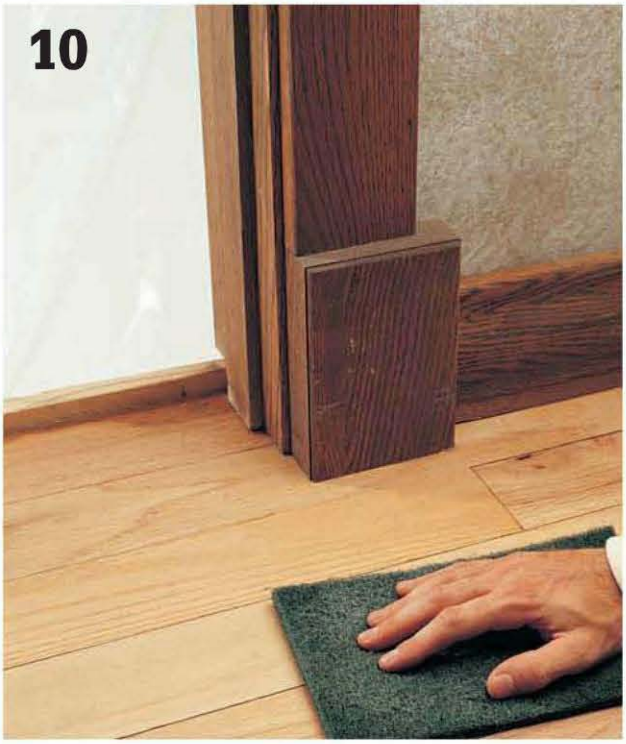

Allow the floor to dry. Buff the surface lightly to remove any raised wood grain, using a medium abrasive pad. Vacuum the surface, using a bristle attachment, then wipe with a tack cloth.

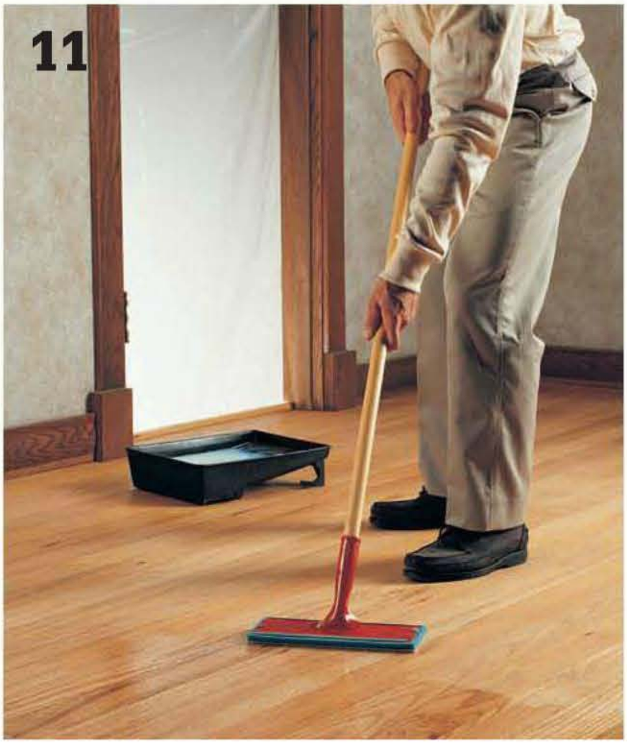

Apply at least two coats of undiluted polyurethane finish to get a hard, durable finish. Allow the finish to dry; repeat step 10 and then add a final coat. Do not overbrush.

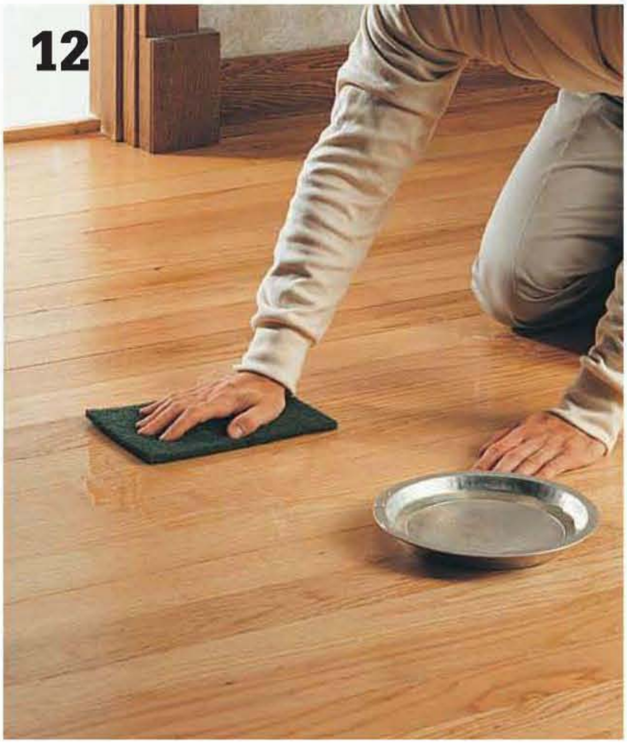

When the final coat is dry, buff the surface with water and fine abrasive pad. Wait at least 72 hours before replacing the shoe molding.

BDBN1202 Type 1 12V DC
Use only with B&D FSB12 or FS120BX battery pack.
5/8" - 2" (16-51mm) x 18 GA. FINISH NAILS
ACCEPTS ONLY STRAIGHT 18 GA FASTENERS
US Pat. No. 6705503, 6604666. Other patents pending.
Questions / Comments?: 1-800-54-HOW-TO
Black & Decker (U.S.) Inc.,
Towson, MD 21286 U.S.A.
Made in China
PRESS TO UNLOCK
SLIDE TO ADJUST
WARNING
AVERTISSEMENT

Walls & Ceilings

Walls and ceilings are much more than structural barriers dividing your house into separate areas. They are canvases awaiting an inspired treatment that can set mood, distinguish space, and reflect your personality within each individual room.

Through creative design and a thoughtful selection of colors, textures, and materials, nearly any wall and ceiling design you dream up can become a reality. From basic painted wallboard and traditional wood paneling to elegant glass block and distinct veneer plaster finishes, there is an ever-increasing variety of wall and ceiling finish materials available to you. Most home centers and lumberyards keep an impressive selection in stock and can easily accommodate special orders.

In this chapter:

- Installing Paneling
- Installing Beadboard Wainscoting
- Building Glass Block Walls
- Applying Veneer Plaster
- Paneling Ceilings
- Installing Suspended Ceilings
- Installing Acoustical Ceiling Tiles
- Installing Metal Ceilings
- Installing Base Molding
- Installing Picture Rail
- Installing Chair Rail
- Installing Crown Molding

Installing Paneling

Paneling is a versatile wall-surfacing material that comes in a wide range of styles, colors, and prices. Paneling sheets are made from a variety of materials for numerous applications:

Solid and veneer wood paneling is durable and easy to clean. Available in finished and unfinished sheets, wood paneling brings a warm, rich tone to any room. It is often used as wainscoting (pages 214 to 219) and also can be used as an inexpensive cover-up for damaged plaster.

Laminate panels are sheets of MDF, particleboard, or plywood faced with paper, print or vinyl. Laminates are available in hundreds of colors, styles, and patterns, providing a durable alternative to paint or wallcoverings.

FRP (fiberglass reinforced plastic), extruded plastic, and vinyl panels contain solid material throughout the panel, creating a low-maintenance, water-resistant wall surface for bathrooms, utility rooms, garages and workshops, as well as numerous commercial applications.

Tileboard is moisture-resistant hardboard coated with melamine, providing a durable, easy-to-clean plastic finish. It's designed to replicate the appearance of ceramic tile, for use in bathrooms, laundry rooms, and kitchens.

Bamboo paneling is gaining in popularity due to its unique look and green-friendly specs. Panels are constructed of strips of bamboo laminated to a fabric backing, which allows it to conform to any type of surface, flat or curved.

Most paneling is available in 4 × 8, 4 × 9, and 4 × 10 sheets. Some manufacturers also offer sheets in sixty-inch widths. Paneling that is ¼" or less in thickness requires a solid backer of at least ½" wallboard; paneling ⅜" thick or more is rigid enough to be fastened directly to framing with sixteen-inch O.C. spacing. Installation typically involves a panel adhesive, either applied in beads along the wall or framing, or troweled onto the back surface of the panel. Make sure to check the manufacturer's instructions for the product you purchase.

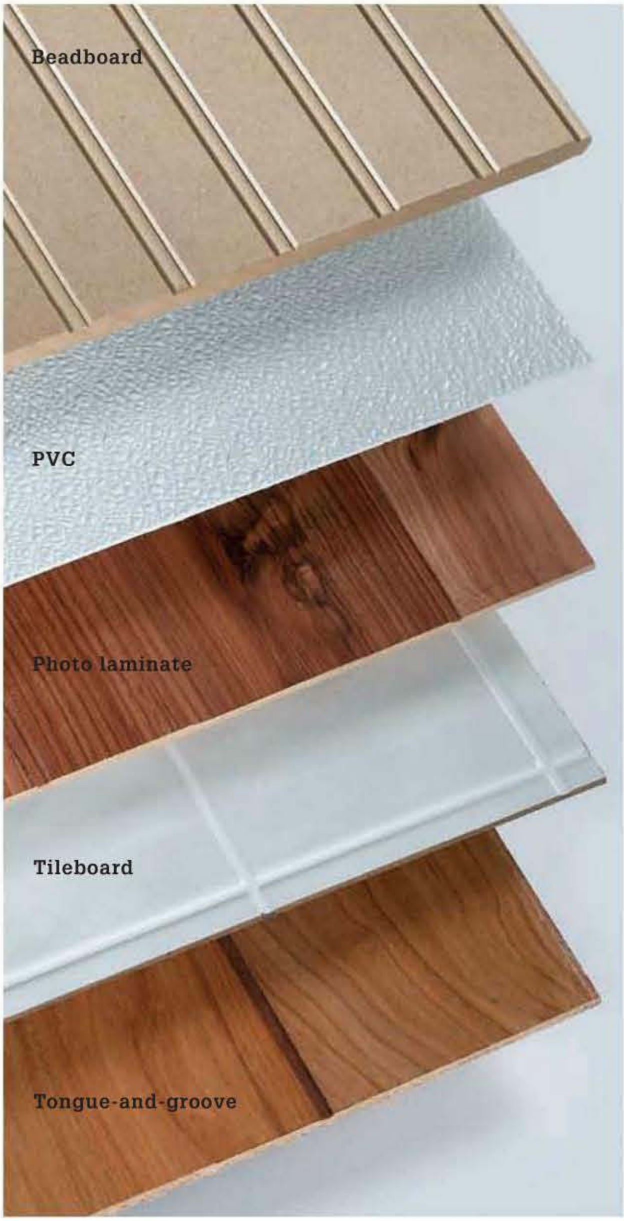

Tools & Materials ▸

- Pry bar
- Stud finder
- Tape measure
- Plumb bob
- Circular saw
- Straightedge
- Hammer
- Carpenter's level
- Compass
- Jigsaw
- Caulk gun
- Paneling sheets
- 4d finish nails
- Wood stain
- Panel adhesive
- Powdered chalk

How to Install Wood Paneling

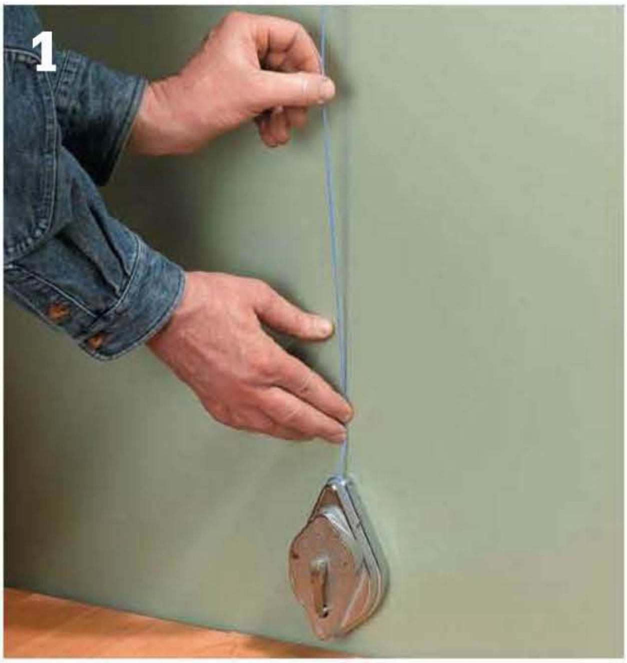

Starting in the corner farthest from the entry, use a stud finder to locate the center of the stud closest to, but less than 48" from, the corner. Find and mark stud centers every 48" from this first stud. Snap a plumb chalk line down the wall at each location. Paneling seams will fall along these lines.

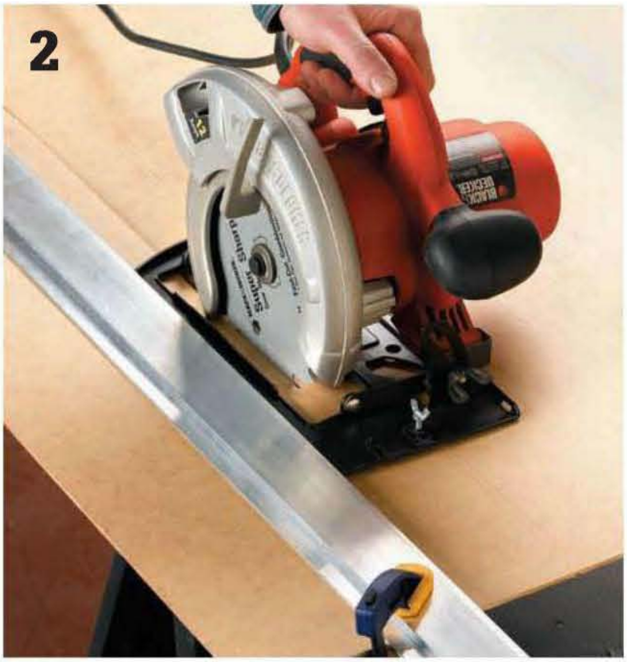

Lay the first paneling sheet face-side down. Measure the distance from corner to the first plumb mark and add 1" to allow for scribing. Use a circular saw and clamped straightedge to cut paneling to this measurement.

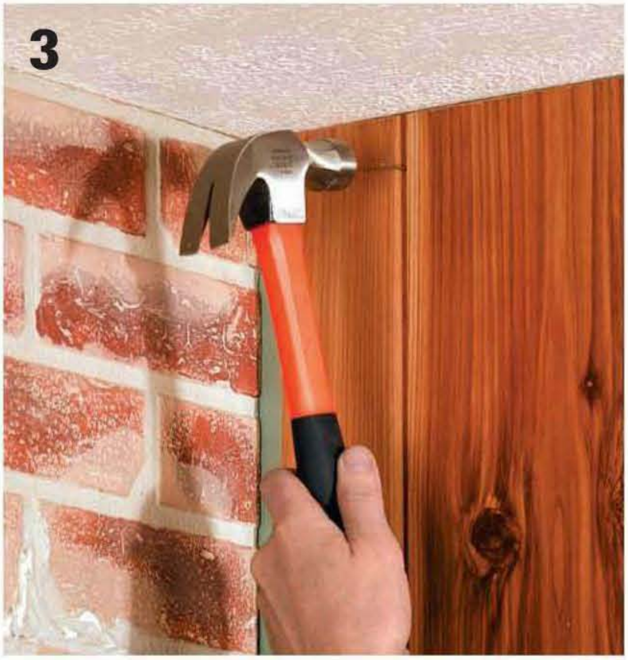

Position the first sheet of paneling against the wall so that the cut edge is 1" away from the corner, and the opposite, finished edge is plumb. Temporarily tack the top of the paneling to the wall.

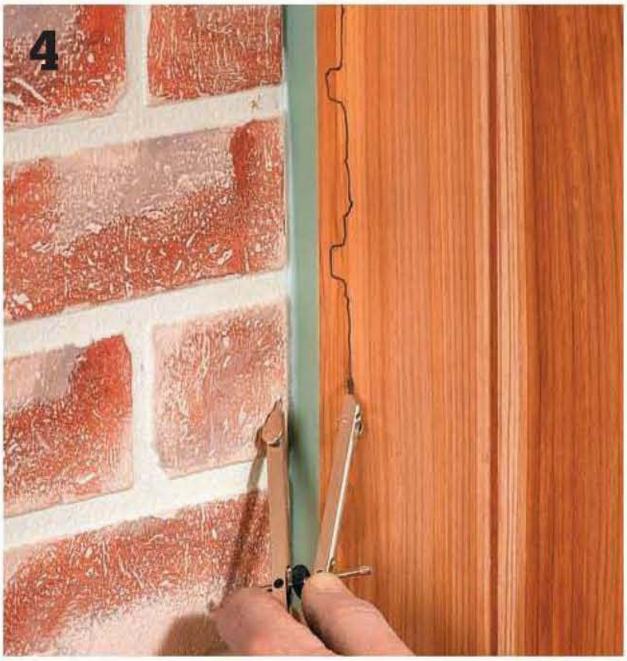

Spread the legs of a compass to 1¼", then run the compass down the full height of the wall to scribe the corner irregularities onto the face of the paneling. Remove paneling from wall.

(continued)

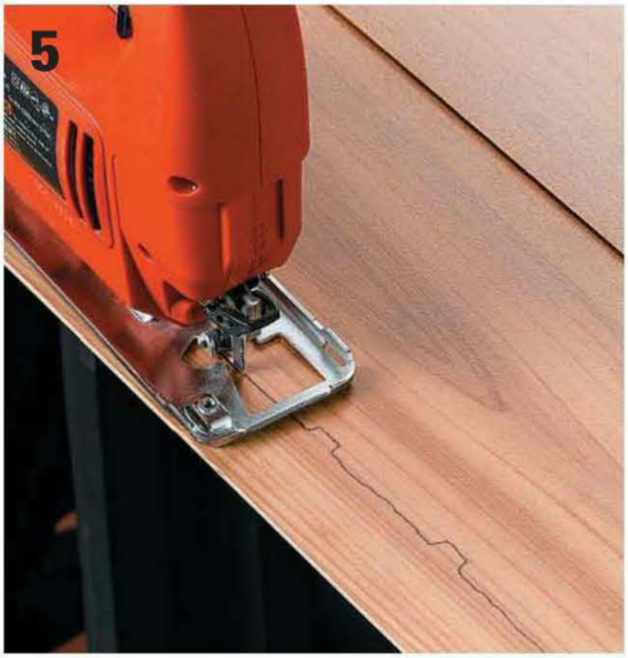

Lay the paneling face-side up and cut along the scribe line with a jigsaw. To prevent splintering, use a fine-tooth woodcutting blade. The scribed edge will fit perfectly against the wall corner.

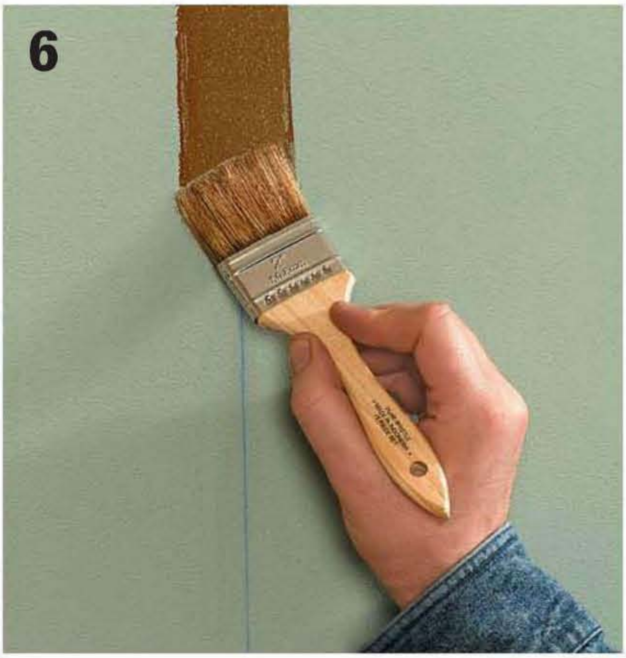

Apply stain or paint to the wall at the plumb lines so the backer will not show through the slight gaps at joints. Select a color that matches the color of the paneling edges, which may be darker than the paneling surface.

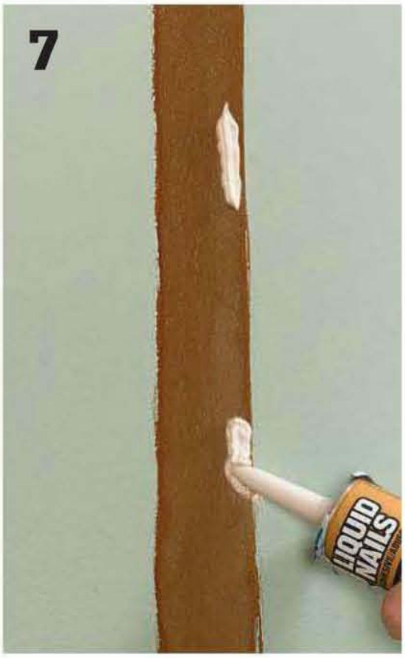

Use a caulk gun to apply 2"-long beads of panel adhesive to the wall at 6" intervals and in a continuous, wavy bead about 1" back from plumb lines (to prevent adhesive from seeping out through the joints). For new construction, apply adhesive directly to the studs.

Attach the paneling to the top of the wall, using 4d finishing nails driven every 16". Press the paneling against the adhesive, then pull it away from the wall. Press the paneling back against the wall when the adhesive is tacky, about 2 minutes.

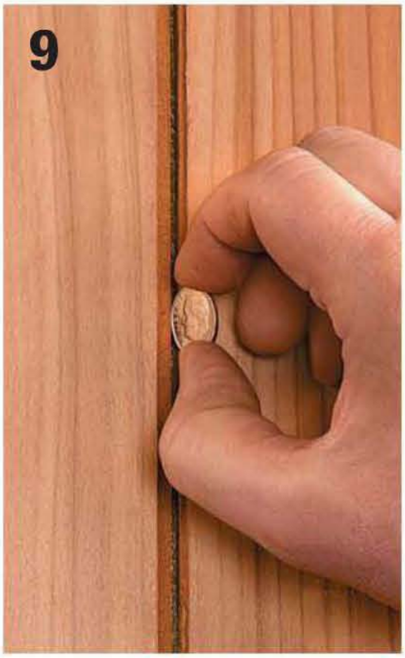

Hang the remaining paneling so that there is a slight space at the joints. This space allows paneling to expand in damp weather. Use a dime as a spacing gauge.

How to Cut Openings in Paneling

1

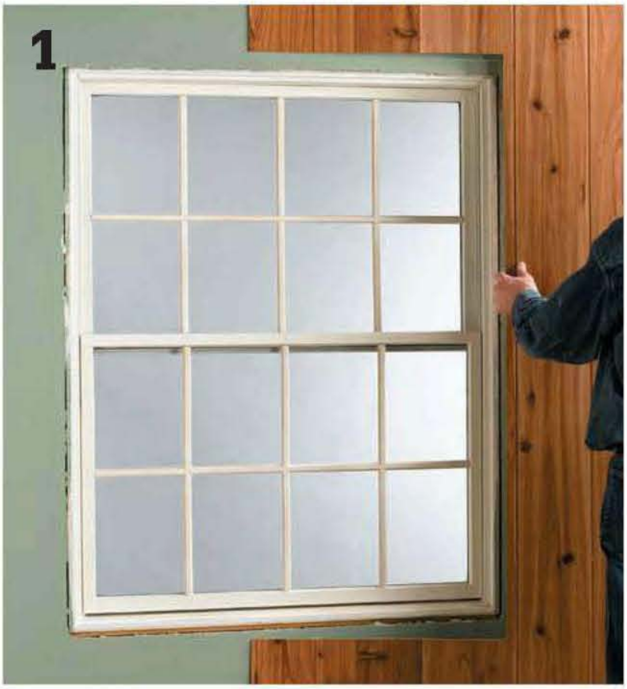

For window, door and other openings, measure the opening and mark the outline on the backside of the paneling. Cut to size using a circular saw and straightedge. Install as you would a full sheet of paneling (page 212).

2

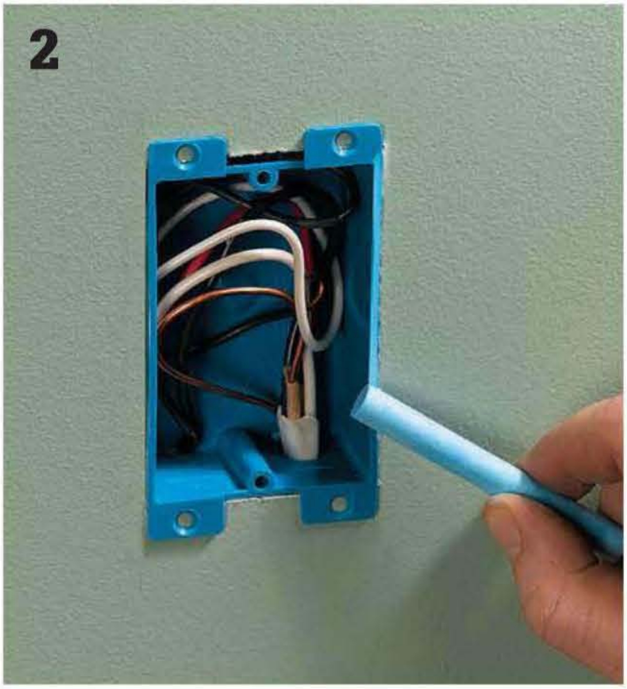

For receptacles, switches, fixtures, and heating vents, coat the edges of electrical boxes and ductwork with chalk.

3

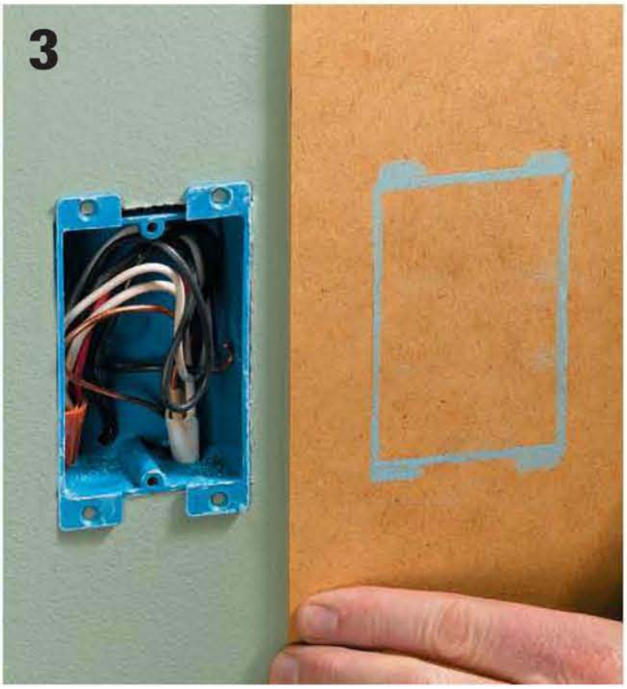

Press the paneling against the wall, so the backside presses against the outlet or vent—the chalk outline will transfer to the paneling.

4

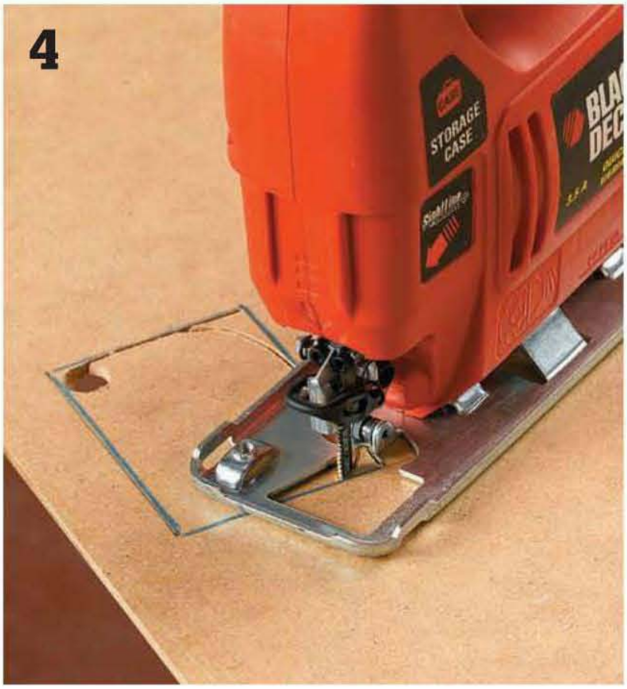

Lay the paneling face-side down. Drill a hole at one corner of each outline, then use a jigsaw with a fine-tooth woodcutting blade to make the cutouts.

Installing Beadboard Wainscoting

Wainscoting refers to virtually any specialized treatment of the lower three to four feet of interior walls. The form demonstrated here, using tongue-and-groove boards, first gained popularity in the early twentieth century. Recently, it has re-emerged as a stylish way to dress up a room.

Typical tongue-and-groove boards for wainscoting are made of pine, fir, or other softwoods and measure ¼" to ¾" thick. Each board has a tongue on one edge, a groove on the other, and usually a decorative bevel or bead on each edge. Boards are cut to length, then attached with nails, most of which are driven through the tongues of the boards. This technique, known as blindnailing, hides the nails from view.

Once installed, the wainscoting is capped at a height of 30" to 36" with a molding called a *cap rail.* The exact height of the wainscoting is a matter of personal preference. When installed to the height of the furniture in the room, wainscoting provides visual symmetry. It also allows the cap rail to double as a chair rail, protecting the lower portion of the walls from damage.

When installed over finished wallboard, wainscoting usually requires that nailers be fastened to the wall studs to provide a reliable backing for nailing. You can skip this step if you know there is consistent blocking between the studs to substitute for this backing. However, this is usually difficult to confirm unless the walls were framed with tongue-and-groove wainscoting in mind.

Wainscoting can be painted or stained. Oil-based stains can be applied before or after installation, since most of the stain will be absorbed into the wood and won't interfere with the tongue-and-groove joints. If you're painting, choose a latex-based paint; it will resist cracking as the joints expand and contract with changes in the weather.

Tools & Materials ▸

- Pencil
- Level
- Circular saw
- Miter saw
- Miter box
- Hammer
- Nail set
- Plane
- Circuit tester
- Pry bar
- Tape measure
- Paintbrush
- Tongue-and-groove boards
- Finish nails
- 1 × 3 furring strips
- 2" 10d finish nails
- Receptacle box extenders, as required
- Paint or stain

Tongue-and-groove wainscoting boards are milled with smooth faces or contoured to add additional texture to your walls. For staining, choose a wood species with a pronounced grain. For painting, poplar is a good choice, since it has few knots and a consistent, closed grain that accepts paint evenly.

How to Prepare for a Wainscoting Project

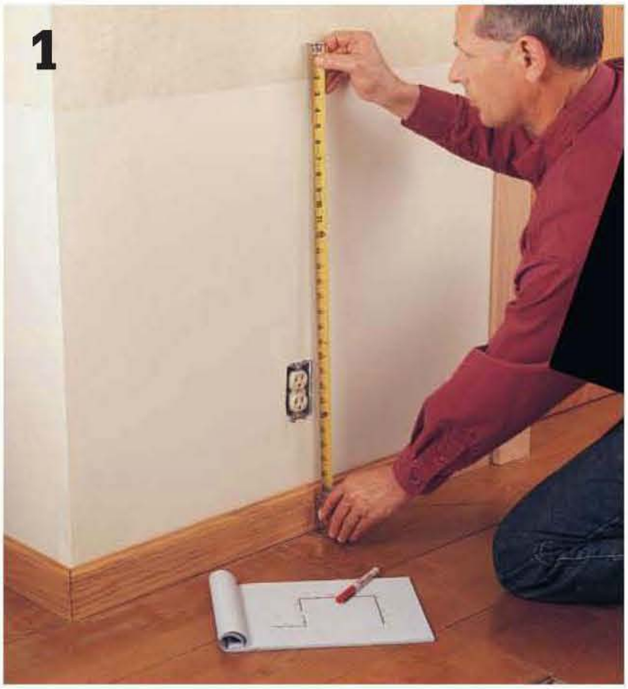

Measure to make a plan drawing of each wall in your project. Indicate the locations of fixtures, receptacles, and windows. Use a level to make sure the corners are plumb. If not, mark plumb lines on the walls to use as reference points.

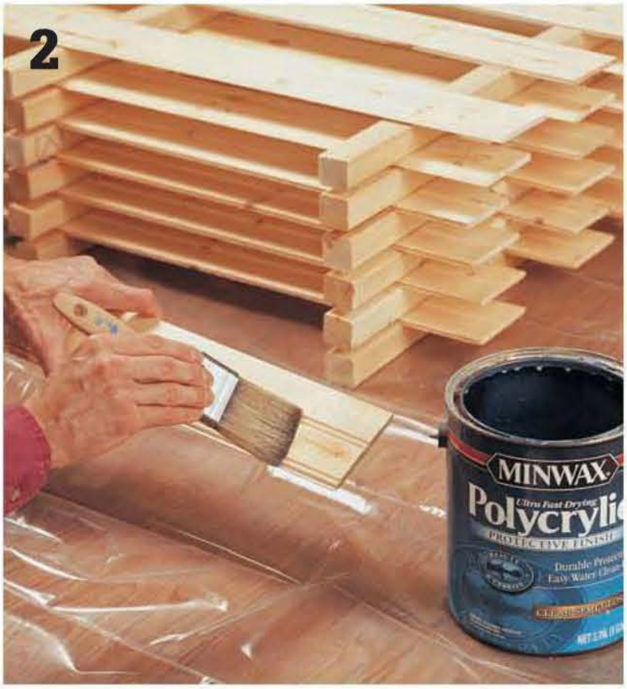

Condition the planking by stacking it in the room where it will be installed. Place spacers between the planks to let air circulate around each board, allowing the wood to adjust to the room's temperature and humidity. Wait 72 hours before staining or sealing the front, back, and edges of each plank.

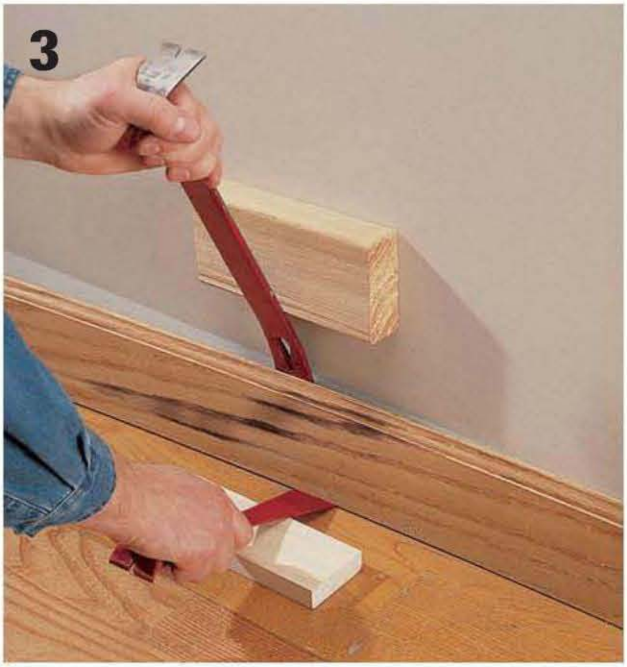

Remove the baseboard moldings, along with any receptacle cover plates, vent covers, or other wall fixtures within the area you plan to cover. Before you begin, turn off the electricity to the circuits in the area.

Mark the walls with level lines to indicate the top of the wainscoting. Mark a line ¼" from the floor to provide a small gap for expansion at the floor.

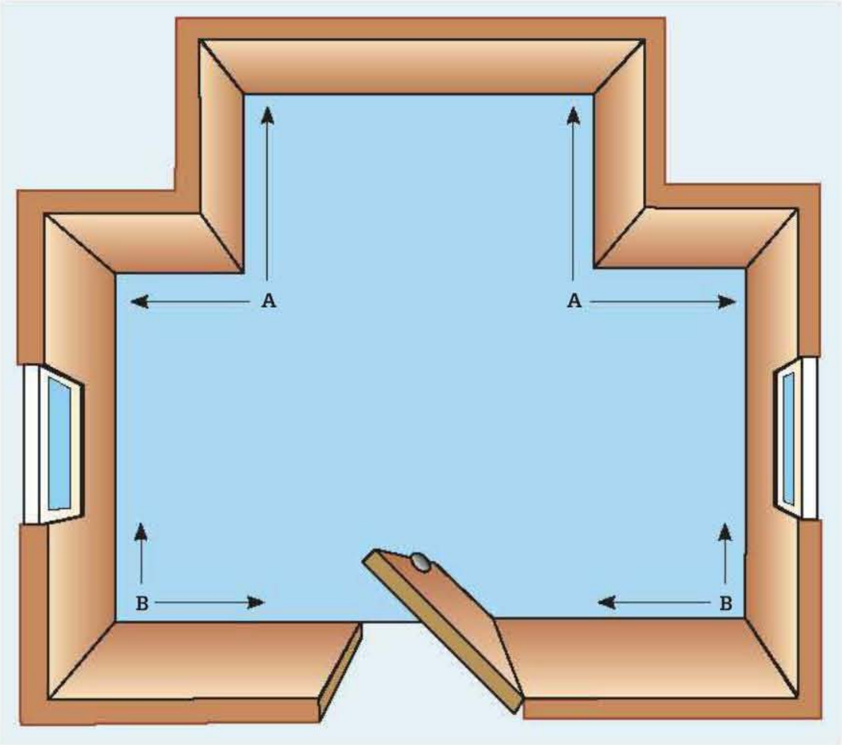

Begin installation at the corners. Install any outside corners (A) first, working your way toward the inside corners. In sections of a room that have no outside corners, start at the inside corners (B), and work your way toward the door and window casings. Calculate the number of boards required for each wall, using the measurements on the drawing you created earlier (length of wall divided by width of one plank). When making this calculation, remember that the tongues are removed from the corner boards. If the total number of boards for a wall includes a fraction of less than ½ of a board, plan to trim the first and last boards to avoid ending with a board cut to less than half its original width.

How to Install Wainscoting at Outside Corners

1

Cut a pair of boards to the widths indicated in the calculations you developed during the planning process.

2

Position the boards at the corner, butting them to create a plumb corner. Facenail the boards in place, then nail the joint, using 6d finish nails. Drive the nails to within ⅛" of the face of the boards, then finish with a nail set.

3

Position a piece of corner trim and nail it in place, using 6d finish nails. Install the remaining boards (opposite, steps 5 and 6).

How to Install Wainscoting at Inside Corners

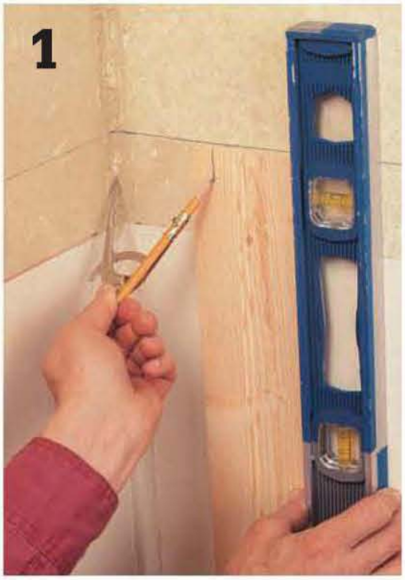

Hold a level against the first board and hold the board flush with the corner. If the wall is out of plumb, trim the board to compensate: Hold the board plumb, position a compass at the inside corner of the wall, and use it to scribe a line down the board.

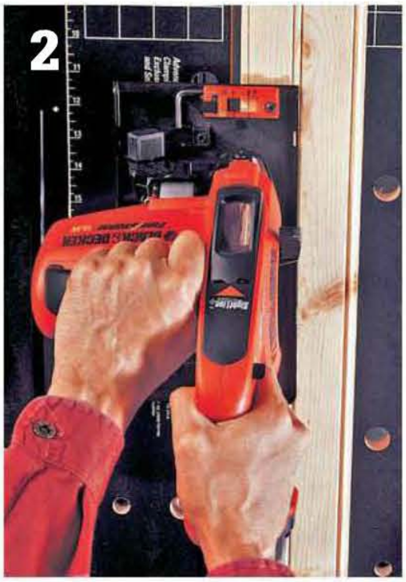

Cut along the scribed line with a circular saw. Subsequent boards may require minor tapering with a plane to adjust for plumb.

Hold the first board in the corner, leaving a ¼" gap for expansion, and facenail into the center of the board at each nailer location, using 6d finish nails. Drive the top nails roughly ½" from the edge so they'll be hidden from view once the cap rail is attached.

Install a second board at the corner by butting it against the first one, then facenailing in at least two locations. Nail to within ⅛" of the face of the board, then use a nail set to finish.

Position subsequent boards. Leave a 1⁄16" gap at each joint to allow for seasonal expansion. Use a level to check every third board for plumb. If the wainscoting is out of plumb, adjust the fourth board, as necessary, to compensate.

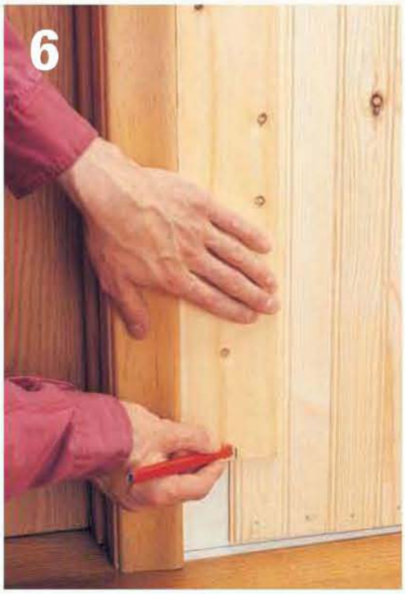

Mark and cut the final board to fit. If you're at a door casing, cut the board to fit flush with the casing (trim off at least the tongue). If you're at an inside corner, make sure it is plumb. If not, scribe and trim the board to fit.

How to Make a Cutout

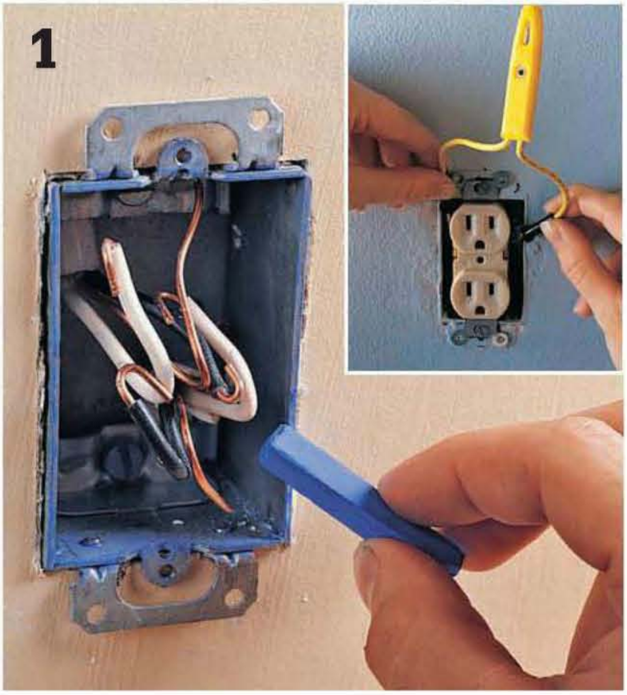

Test the receptacle (inset) to make sure the power is off. Then, unscrew and remove the receptacle from the box. Coat the edges of the electrical box with bright colored chalk.

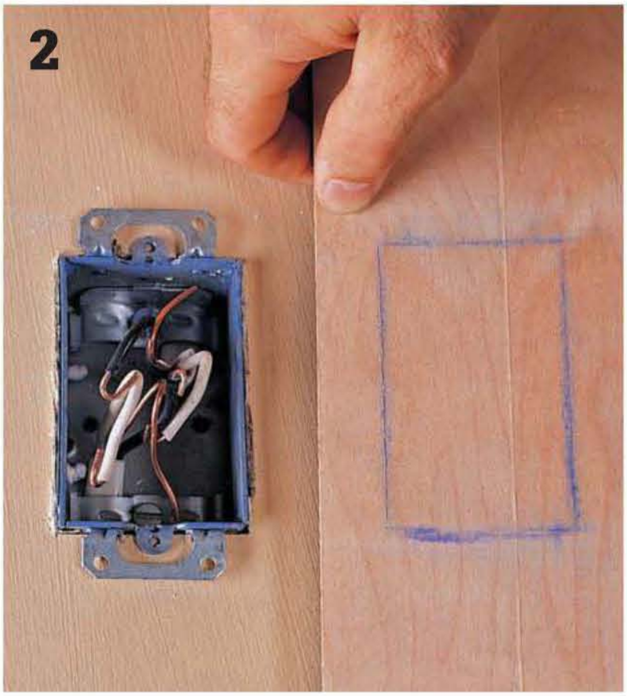

Press the back of the board that will be installed over the receptacle, directly against the electrical box, to create a cutting outline.

Lay the board face down and drill a large pilot hole near one corner of the outline. Use a jigsaw fitted with a fine-tooth woodcutting blade to make the cutout. Be careful not to cut outside the lines.

Facenail the wainscoting to the wall, then reattach the receptacle with the tabs overlapping the wainscoting so the receptacle is flush with the opening. You may need longer screws.

Tip ▸

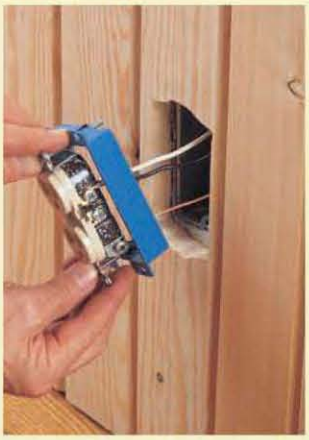

When paneling around a receptacle with thick stock, you will need to attach a receptacle box extender to the inside of the box, then reconnect the receptacle so it is flush with the opening in the paneling.

How to Install Wainscoting Around a Window

On casement windows, install wainscoting up to the casings on the sides and below the window. Install ½" cove molding, quarter round, or other trim to finish the edges.

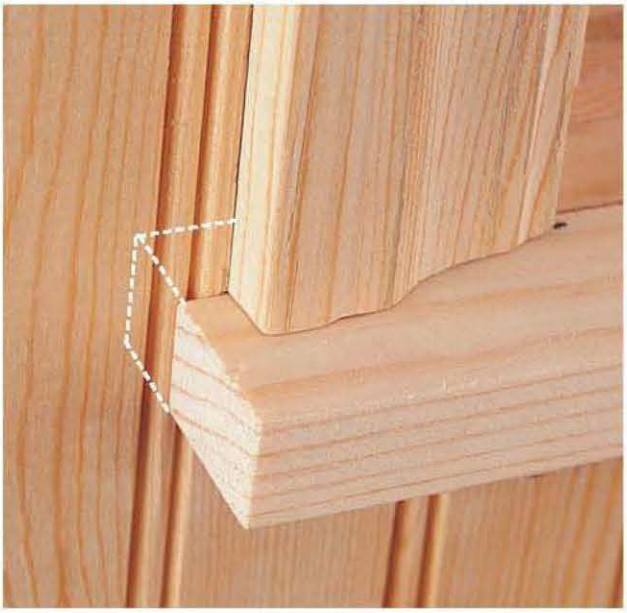

On double-hung windows, remove any window trim and install wainscoting up to the jambs on the sides and below the window. Cut the stool to fit over the wainscoting, then reinstall the apron.

How to Finish a Wainscoting Project

Cut baseboard moldings (pages 242 to 245) to fit over the wainscoting and attach them by nailing 6d finishing nails at the stud locations. If you plan to install a base shoe, leave a small gap at the floor.

Cut cap rail to fit as you would contoured chair rail (page 249). At doors and windows, install the cap rail so its edge is flush with the side casings.

Attach the cap rail by nailing 4d finish nails through the flats of the moldings at the stud locations so that nails enter both the studs and the wainscoting. Set the nails with a nail set.

Building Glass Block Walls

With its ability to transmit light, a glass block partition wall defines separate living areas while maintaining a sense of openness. You can find glass block at specialty distributors and home centers in a variety of patterns, shapes, and sizes, along with all the products needed for the installation.

You can build your wall to any height. Top a low wall with a course of bullnose blocks to give it a finished rounded edge, or with flat block to create a shelf. To build a full-height wall, calculate the number of courses of block you'll have, then frame-in a header to fill the remaining space between the finished block and the ceiling.

Because of its weight, a glass block wall requires a sturdy foundation. A four-inch-thick concrete basement floor should be strong enough, but a wood floor may need to be reinforced. Contact the local building department for requirements in your area. Also bear in mind that glass block products and installation techniques vary by manufacturer—ask a glass block retailer or manufacturer for advice about the best products and methods for your project.

Tools & Materials ▸

Chalk line
Circular saw
Jigsaw
Paintbrush
Drill
Mixing box
Trowel
Level
Pliers
Jointing tool
Nylon- or natural-bristle brush
Sponge
2 × 6 lumber
16d common nails
Water-based asphalt emulsion
Panel anchors
2½" drywall screws
Foam expansion strips
Glass block mortar
8" glass blocks
¼" T-spacers
Board
Reinforcement wire
16-gauge wire
Caulk or wall trim
Baseboard

Glass block has many uses in modern up-scale homes and today's installation systems allow homeowners to easily work with it.

How to Build a Glass Block Wall

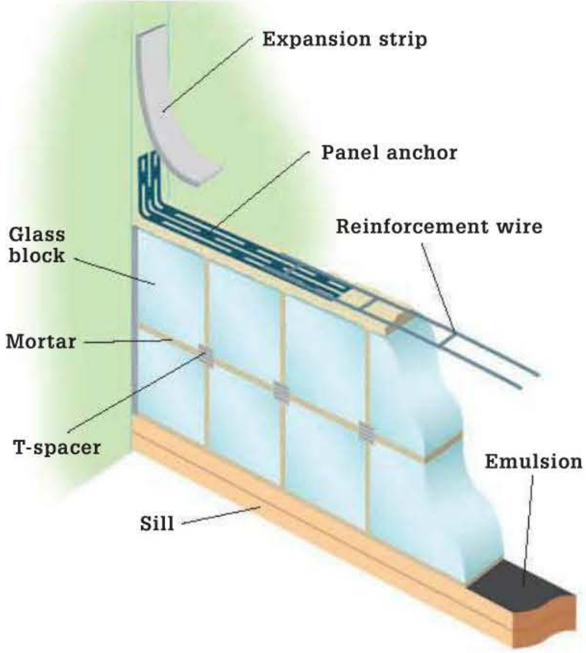

The block wall in this project has a sill made of two 2 × 6s cut to the exact width of the block. This provides a stable base to help resist floor movement and protect the lower courses. The block wall is secured to an anchor stud in an adjoining wall, by means of metal panel anchors. Expansion strips between the two walls allow for movement.

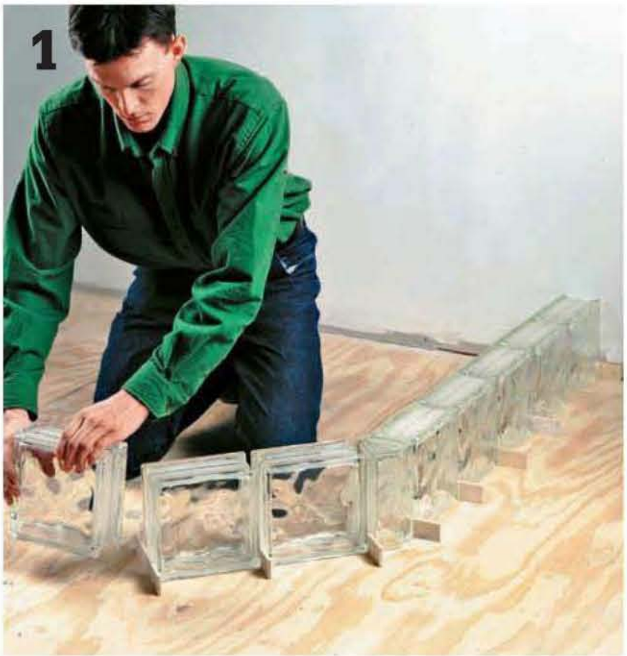

Dry-lay the first course of glass block, using a ⅜" wood spacer between the wall and the first block, and ¼" spacers between the remaining blocks, to set the gaps for the mortar joints. Mark the wall position onto the floor, then remove the blocks. Snap chalk lines along the marks to create the sill outline.

Determine the sill thickness based on the size of your baseboard and thickness of the floorcovering. Rip 2 × 6 lumber to the width of the block. If the end blocks are shaped, trim the sill pieces to match, using a jigsaw. Fasten the sill to the subfloor and framing below with 16d common nails. Apply asphalt emulsion to the sill, using a paintbrush.

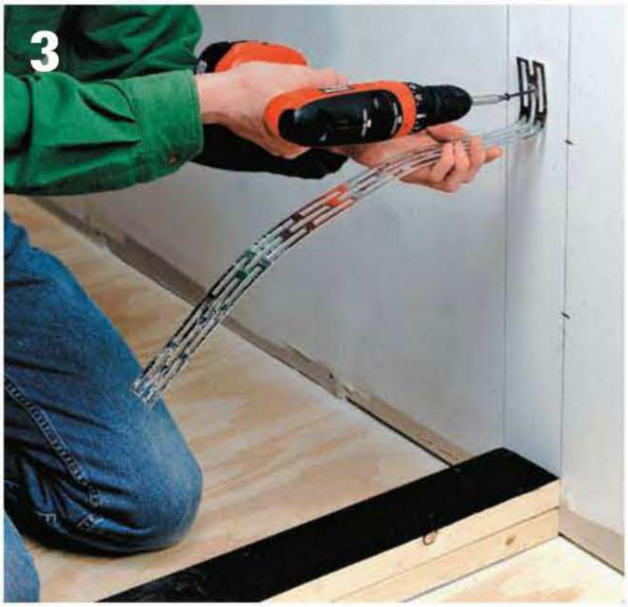

Mark plumb lines on the adjoining wall, straight up from sides of the sill. Mark the finished height of each course along the lines. Fasten a panel anchor to the anchor stud at the top of every second course, using 2½" drywall screws. Cut expansion strips to size and adhere them to the wall between the anchors.

(continued)

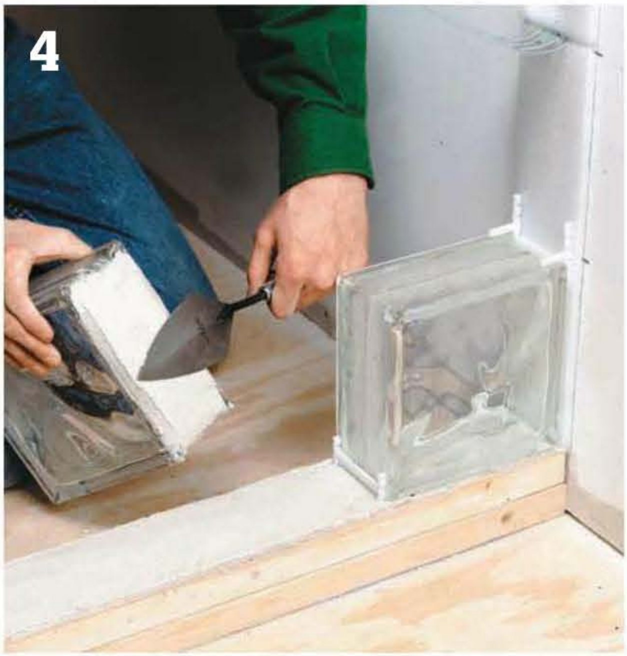

Mix only as much mortar as you can apply in about 30 minutes. Lay a ⅜"-thick mortar bed on the sill, enough for three or four blocks. Set the first block, using ¼" T-spacers at the mortar joint locations (follow the manufacturer's directions for modifying T-spacers at the bottom and sides of the wall). Do not place mortar between blocks and expansion strips. Butter the trailing edge of each subsequent block with enough mortar to fill the sides of both blocks.

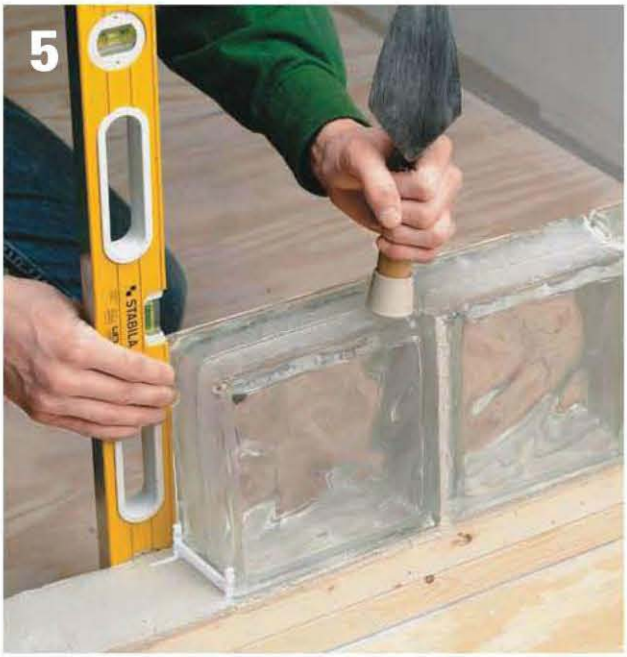

Lay the remainder of the course. If the wall has a corner, work from both ends toward the center, and install the corner piece last. Use ¼" T-spacers between blocks to maintain proper spacing. Plumb and level each block as you work, then check the entire course, using a flat board and a level. Tap blocks into place using a rubber mallet—do not strike them with a metal tool.

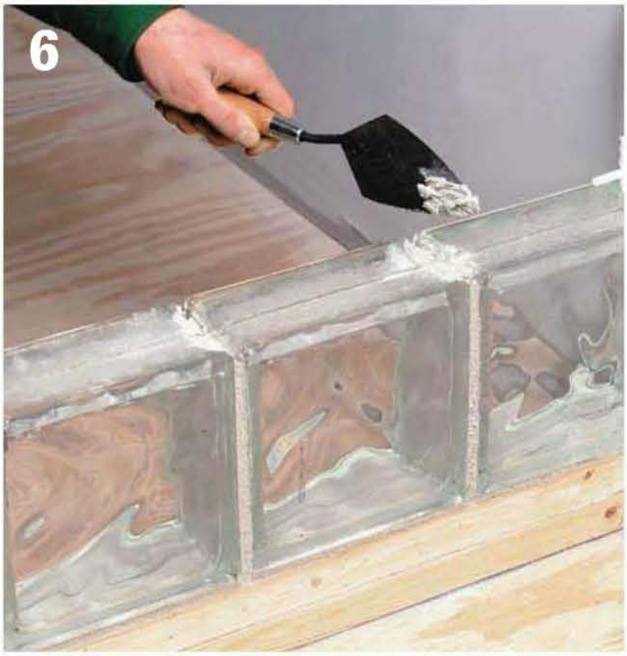

At the top of the course, fill the joints with mortar, and then lay a ¼" bed of mortar for the second course. Lay the block for the second course, checking each block for level and plumb as you work.

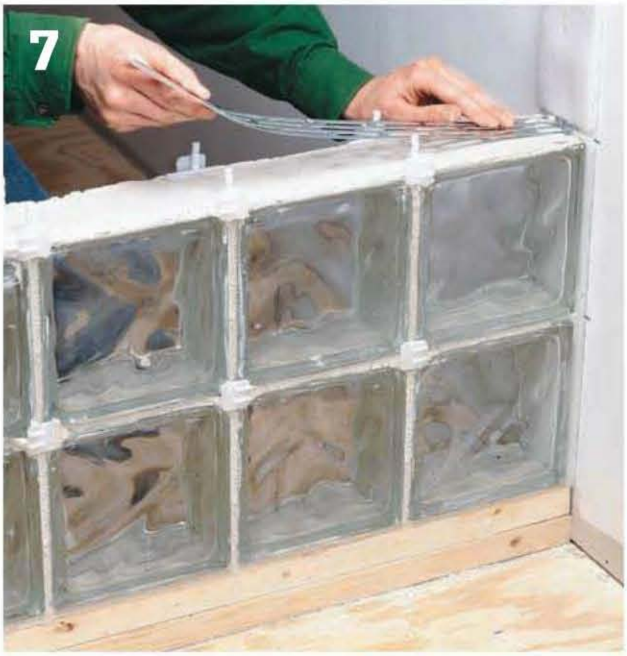

Apply a ⅛" bed of mortar over the second course, then press the panel anchor into the mortar. Repeat this process at each anchor location.

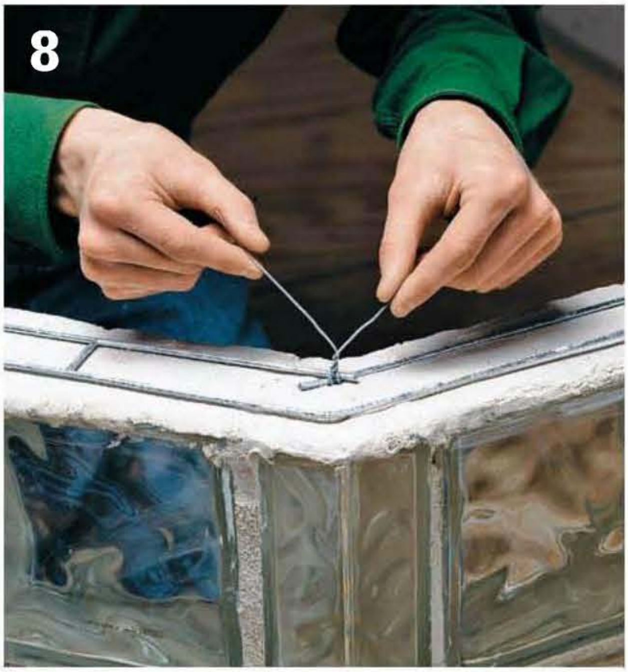

Add reinforcement wire in the same joints as the panel anchors, overlapping the anchors by 6". Also overlap the wire by 6" where multiple pieces are needed. At corners, cut the inner rail of the wire, bend the outer rail to follow the corner, then tie the inner rail ends together with 16-gauge wire. Add another ⅛" mortar bed, then lay the next course of block.

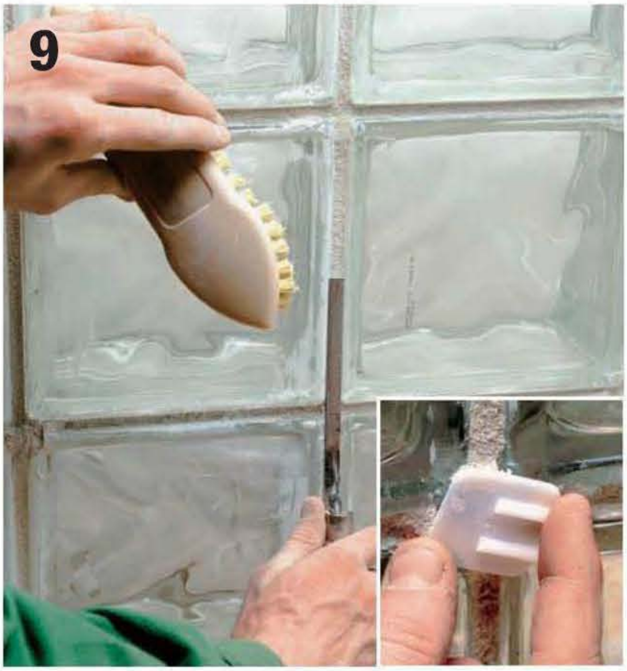

Build the wall in complete courses, checking the mortar after each course. When it is hard enough to resist light finger pressure (usually within 30 minutes), twist off the T-spacer tabs (inset) and pack mortar in the voids. Then, tool all of the joints with a jointing tool. Remove excess mortar from the glass, using a brush or damp sponge.

Clean the glass block thoroughly, using a wet sponge and rinsing it often. Allow the surface to dry, then remove cloudy residue with a clean, dry cloth. After the mortar has cured for two weeks, apply a sealant. Caulk the seam between the glass block and the adjoining wall, or cover the gap with trim.

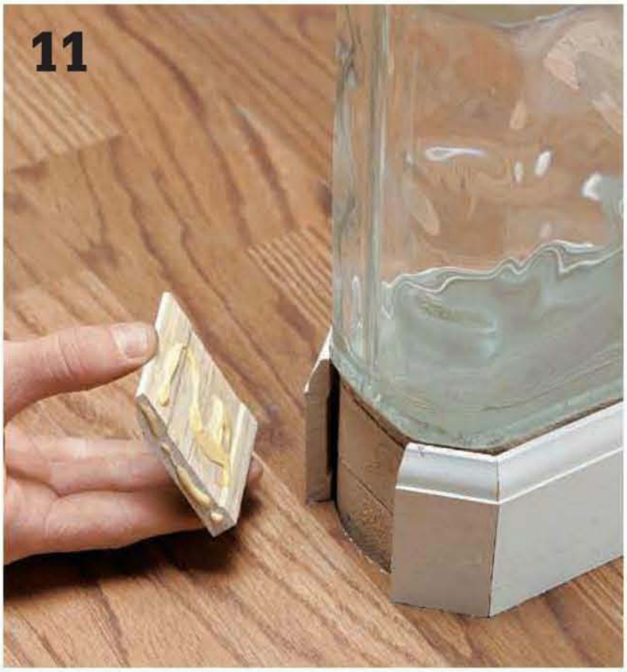

Reinstall the flooring, if necessary, then cut baseboard to fit around the sill (see pages 242 to 245). If the end of your wall has curved (bullnose) block, wrap the end with three pieces of trim.

Applying Veneer Plaster

While gypsum wallboard all but wiped out traditional plaster and lath in the 1940s, a new generation of plaster products now make plaster easier and cheaper to apply, leading to renewed popularity for this classic material.

Veneer plaster systems provide a solid, uniform wall surface that is highly resistant to nail pops, cracks and surface damage. A skim coat of plaster is troweled onto a gypsum wallboard base that has a distinctive blue color, commonly called blueboard. While blueboard is installed like standard wallboard, it has a highly absorptive face paper to which the wet-mix plaster bonds. Blueboard joints do not need to be taped as precisely as standard wallboard joints, and seams and fastener heads do not show through the finished plaster surface, a common problem with standard wallboard.

Veneer systems are available in one-coat and two-coat systems. One-coat systems have a single layer of finish plaster applied directly to the blueboard base; two-coat systems, a rough basecoat for the finish plaster to mechanically "key" or bond to, providing a more rigid surface. Finish plaster can be troweled smooth or tooled for a texture. Sand and other additives can be used to create coarser textures.

Applying veneer plaster effectively does take some time to master, but no more so than that of any masonry technique that requires troweling. The key is to apply the plaster in quick, short strokes, called "scratching in," and then to immediately trowel it over with a steady, even stroke to smooth the plaster to a consistent thickness, typically 1⁄16" to 1⁄8".

Tools & Materials ▸

Stapler
Hammer
Heavy-duty ½" drill with mixing paddle
16-gal. drum
Mortar hawk
12" trowel
Fine-wire rake or broom (for basecoat)
Spray water bottle
Metal corner bead with mesh flanges
1¼" wallboard screws
Non-adhesive fiberglass mesh tape
¼" staples
Clean potable water
Dry-mix veneer basecoat plaster (for two-coat application)
Dry-mix veneer finish plaster

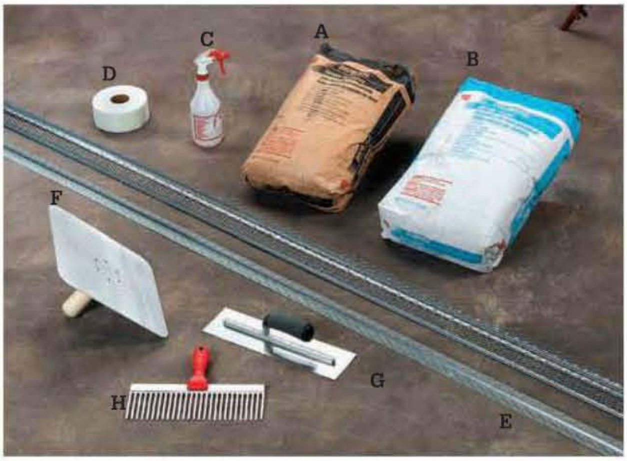

Tools and materials for installing veneer plaster include: dry-mix veneer plaster basecoats (A); finish plaster (B), available for smooth or textured applications; spray bottle for moistening surfaces (C); non-adhesive fiberglass mesh tape for covering blueboard panel seams and inside corners (D); outside corner beads with metal beads and mesh flanges (E); mortar hawk (F); 12" trowel (G); thin-wire rake for roughening the base coat (H).

ONE- AND TWO-COAT VENEER PLASTER SYSTEMS

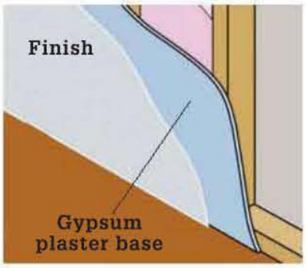

One-coat veneer plaster systems use a single, 1⁄16"- to 3⁄32"-thick coat of finish plaster applied directly to a blueboard base. The coat can be troweled smooth or textured, resulting in a hard, monolithic surface.

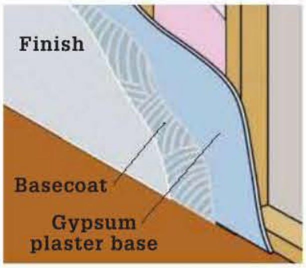

Two-coat veneer plaster systems are composed of a 1⁄16" to 1⁄8" basecoat plaster applied to blueboard, followed by a 1⁄16"- to 3⁄32"-thick coat of finish plaster. The finish coat bonds with the scratched basecoat surface, forming a more uniform and monolithic surface than that of a one-coat system.

How to Apply a One-Coat Veneer Plaster System

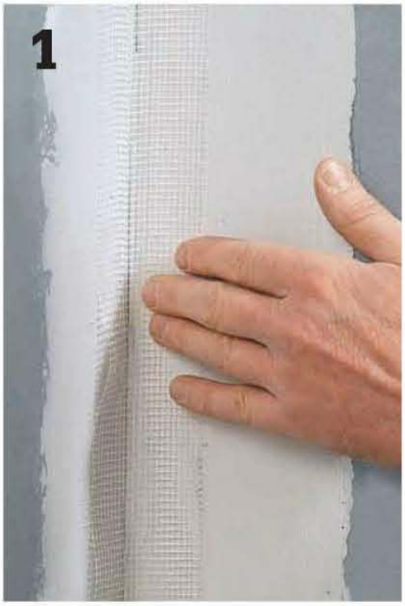

Cover all seams first. Apply a thin layer of plaster along all flat seams and corner bead, feathering out the edges by 6". For inside corners, apply a thin bed of plaster and embed the loose tape, then cover with another thin layer. Allow all taped seams to set.

Variation: Blueboard joints can also be reinforced with paper tape. Embed the tape in a thin plaster bed, then and cover with another thin layer to conceal the tape fully. *Note: Some manufacturers recommend setting-type compound for embedding paper tape; always follow the manufacturer's directions for the products you use.*

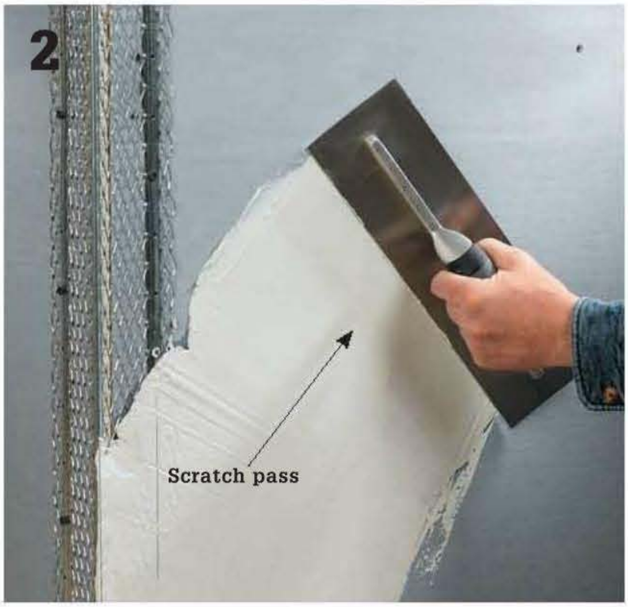

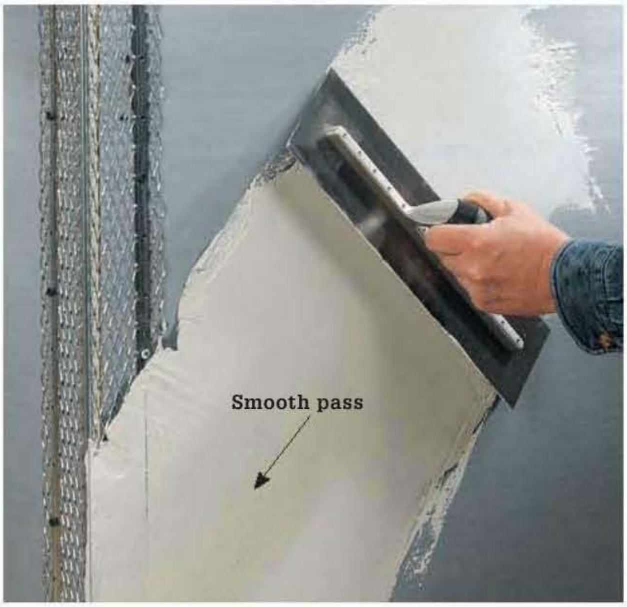

After the seams have set, begin plastering the surface, beginning at one corner and moving to the opposite. Start with ceilings and then do the walls, completing one entire surface before moving on to the next. To apply the plaster, tightly scratch in the material up the wall (photo left), then immediately double-back over it, smoothing over the material to a thickness of $\frac{1}{16}$ to $\frac{3}{32}$", as specified by the manufacturer. Use tight, quick strokes to apply the plaster during the "scratch pass" and long, even strokes to achieve consistency during the "smooth pass."

(continued)

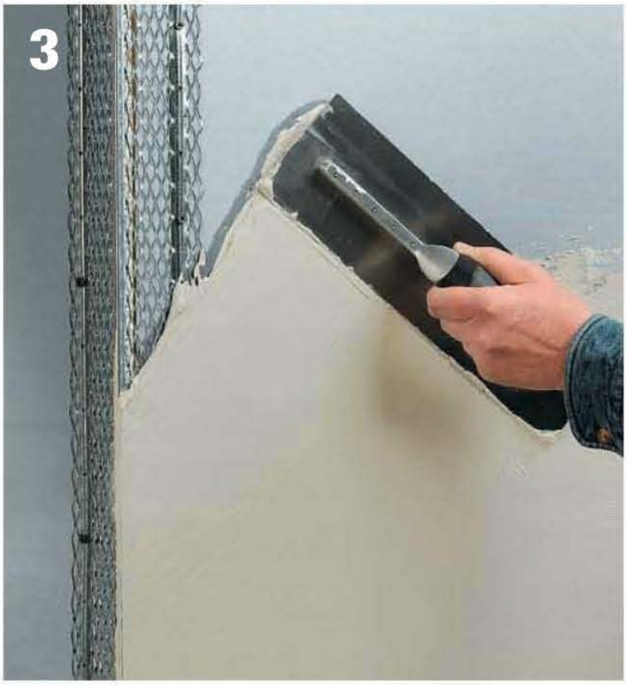

Continue to apply plaster by scratching in and smoothing over the surface. Don't worry about uniformity and trowel ridges at this point. Rather, make sure the entire surface is completely concealed with a relatively even plaster coat, 1⁄16" to 3⁄32"-thick.

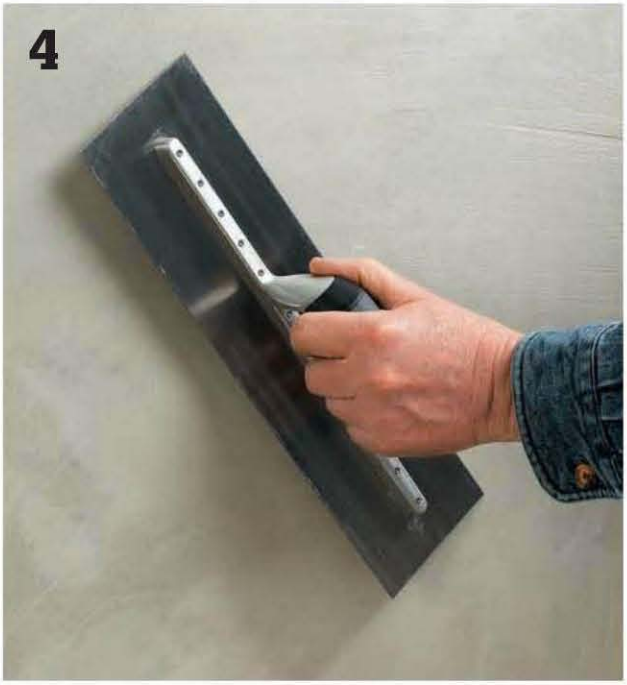

Once the plaster begins to firm, trowel the surface to fill any voids and remove tooling marks and imperfections, integrating the surface into a uniform smoothness.

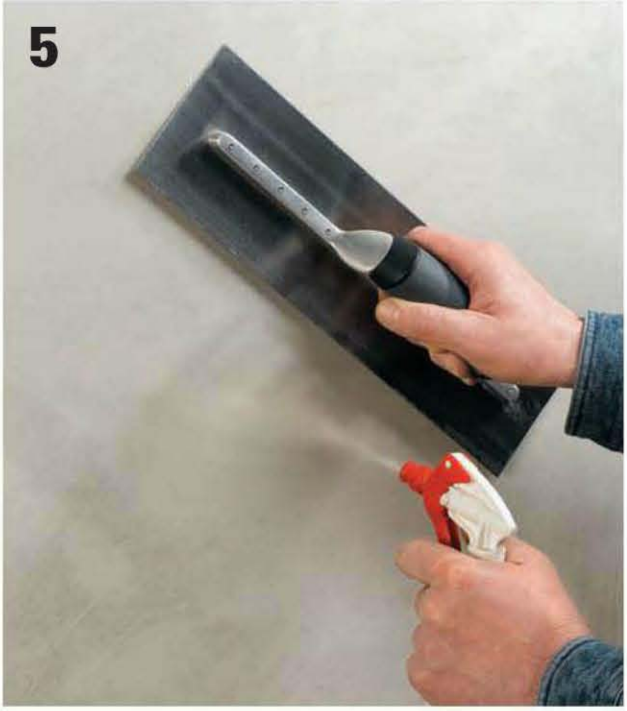

Prior to the plaster setting, make a final pass with the trowel to smooth the surface, using water sparingly. Do not over trowel; stop before the plaster begins to darken and sets.

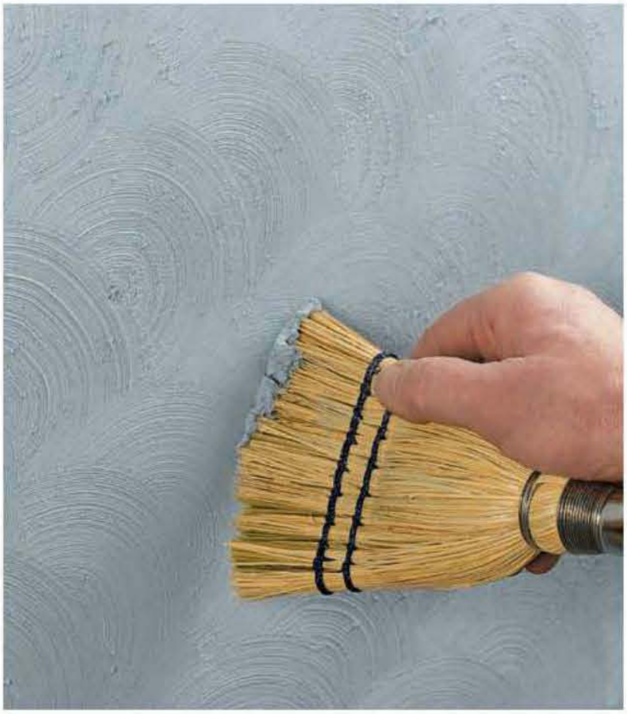

Variation: For textured surfaces, skip the final troweling and work the surface with a texturing tool to achieve the desired results. *Note: Sand or texture added to the plaster mixture does not require tooling.*

How to Apply Basecoat in a Two-Coat Veneer Plaster System

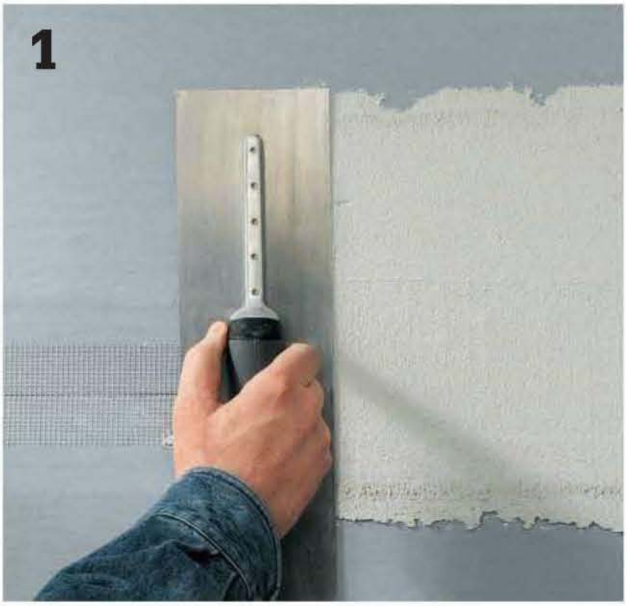

Apply a thin layer of basecoat along all flat seams and corner bead, feathering out the edges by 6". For inside corners, apply a thin bed of basecoat and embed the loose tape, then cover with another thin layer. Allow all taped seams to set.

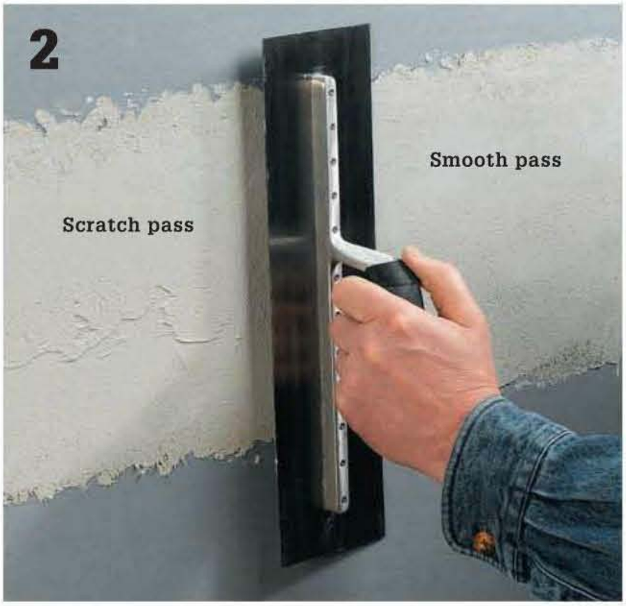

After the seams have set, tightly scratch in basecoat, then immediately double-back over it, smoothing over the material to a thickness of 1⁄16" to 1⁄8", as specified by the manufacturer. Use tight, quick strokes to apply basecoat for the "scratch pass" and long, even strokes to achieve consistency for the "smooth pass."

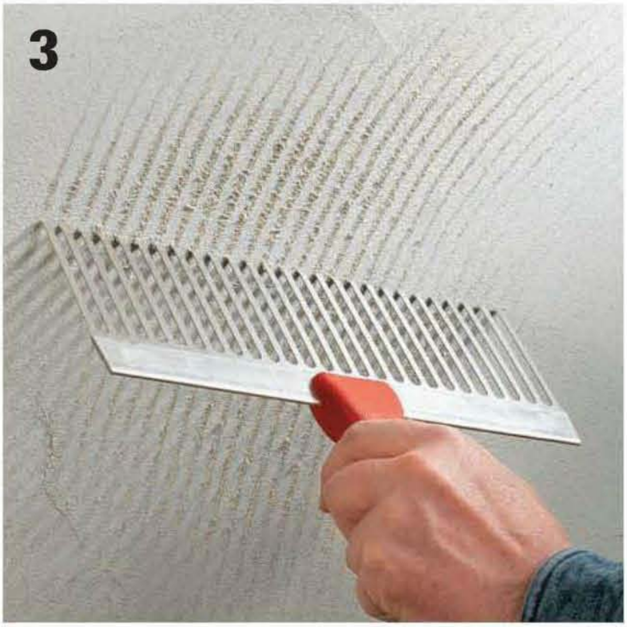

Once the plaster begins to firm or "take up," trowel the surface to fill any voids and remove tooling marks and imperfections, integrating the surface into a reasonably uniform surface—do not over-trowel to a smooth surface. Create keys for the final coat, using a thin-wire rake to roughen the basecoat.

Approximately two hours after the basecoat has set, the finish coat can be applied using the same techniques as for a one-coat veneer plaster system (see pages 225 to 226).

Paneling Ceilings

Tongue-and-groove paneling is a pleasing alternative to a wallboard ceiling, particularly in a knee-wall attic. Pine paneling is most common, but any tongue-and-groove material can be used. These materials are typically ⅜" to ¾" thick and are attached directly to ceiling joists and rafters (over faced insulation, when required). Most codes require you to install ⅜" wallboard as a fire stop under ceiling material thinner than ¼".

Allow for waste by purchasing 15% more material than the square footage of the ceiling; add more for waste if the ceiling requires many angled cuts. Since the tongue portion on most pieces slips into the groove on an adjacent piece, square footage is based on the exposed face (called the reveal) once the boards are installed. A compound miter saw is the best tool for ensuring clean cuts. This is especially important if the ceiling includes non-90° angles.

Tongue-and-groove boards are attached with flooring nails driven through the shoulder of the tongue into each rafter (called blindnailing because the nail heads are covered by the next board). Nailing through the board face is only necessary on the first and last course and on scarf joints.

Layout is very important to the success of a paneled surface, because the lines clearly reveal flaws such as pattern deviations, misaligned walls, and installation mistakes. Before beginning the installation, measure to see how many boards will be installed (using the reveal measurement). If the final board will be less than 2 inches wide, trim the first, or starter board, by trimming the long edge that abuts the wall.

If the angle of the ceiling peak is not parallel to the wall, you must compensate for the difference by ripping the starter piece at an angle so that the leading edge, and every piece thereafter, is parallel to the peak.

Tools & Materials ▸

Chalk line
Compound miter saw
Circular saw
Drill
Nail set
Tongue-and-groove paneling
1¾" spiral flooring nails
Trim molding

Tongue-and-groove paneling can be installed directly over rafters or joists or over wallboard. In attic installations, it's important to insulate first, adding a separate vapor barrier if required by building codes. Local code may also require that paper-faced insulation behind a kneewall be covered with drywall or other material.

How to Panel a Ceiling

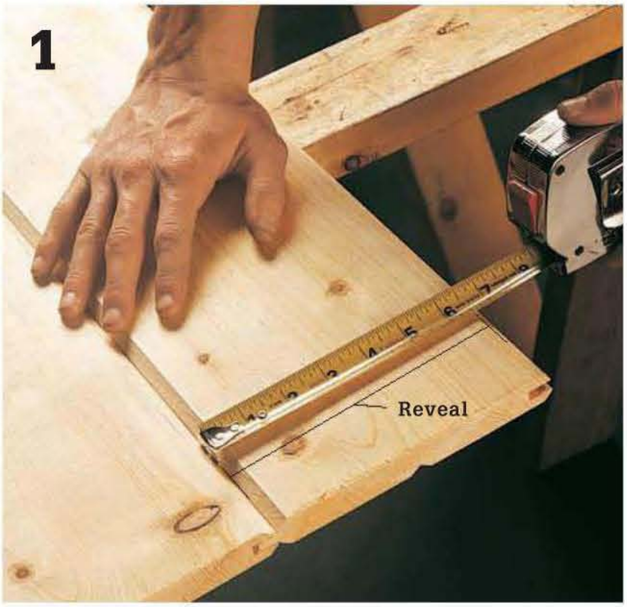

To plan your layout, first measure the reveal of the boards. Fit two pieces together and measure from the bottom edge of the upper board to the bottom edge of the lower board. Calculate the number of boards needed to cover one side of the ceiling by dividing the reveal dimension into the overall distance between the top of the wall and the peak.

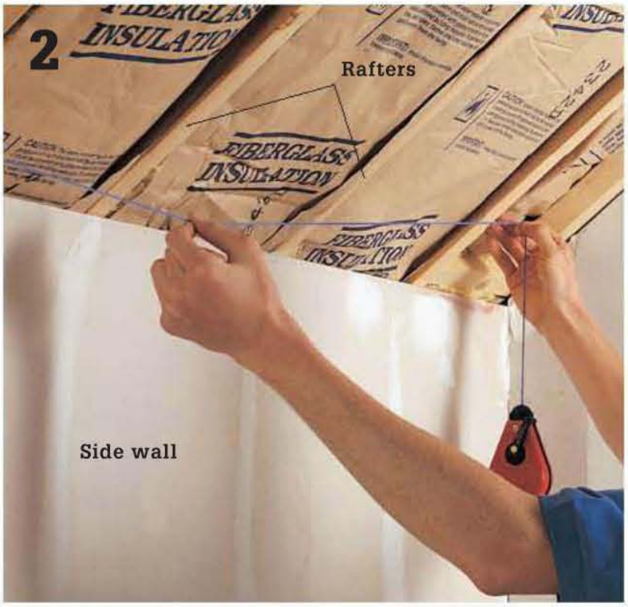

Use the calculation from step 1 to make a control line for the first row of panels—the starter boards. At both ends of the ceiling, measure down from the peak an equal distance, and make a mark to represent the top (tongue) edges of the starter boards. Snap a chalk line through the marks.

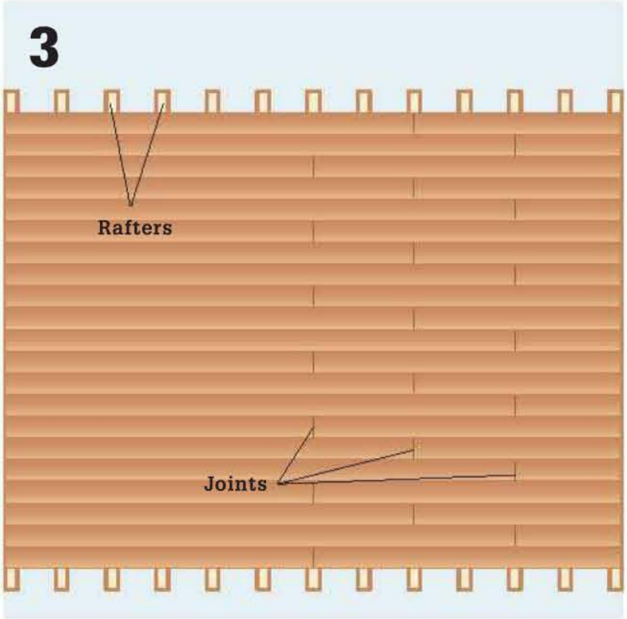

If the boards aren't long enough to span the entire ceiling, plan the locations of the joints. Staggering the joints in a three-step pattern will make them less conspicuous. Note that each joint must fall over the middle of a rafter. For best appearance, select boards of similar coloring and grain for each row.

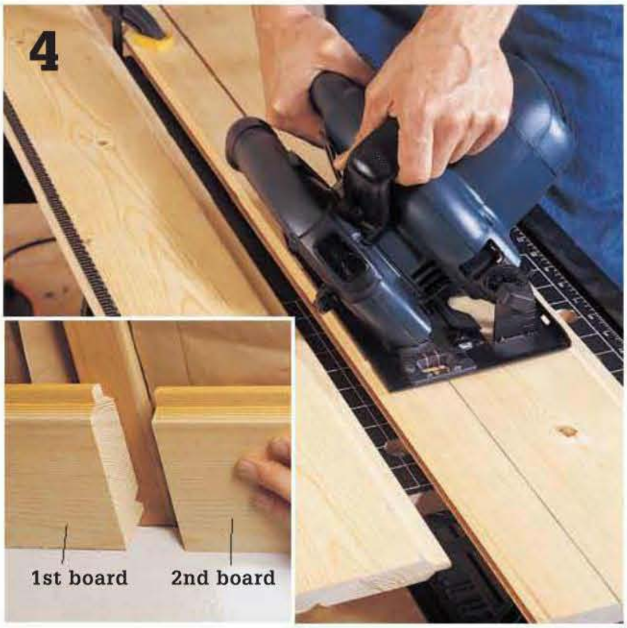

Rip the first starter board to width by bevel-cutting the bottom (grooved) edge. If the starter row will have joints, cut the board to length using a 30° bevel cut on the joint end only. Two beveled ends joined together form a scarf joint (inset), which is less noticeable than a butt joint. If the board spans the ceiling, square-cut both ends.

(continued)

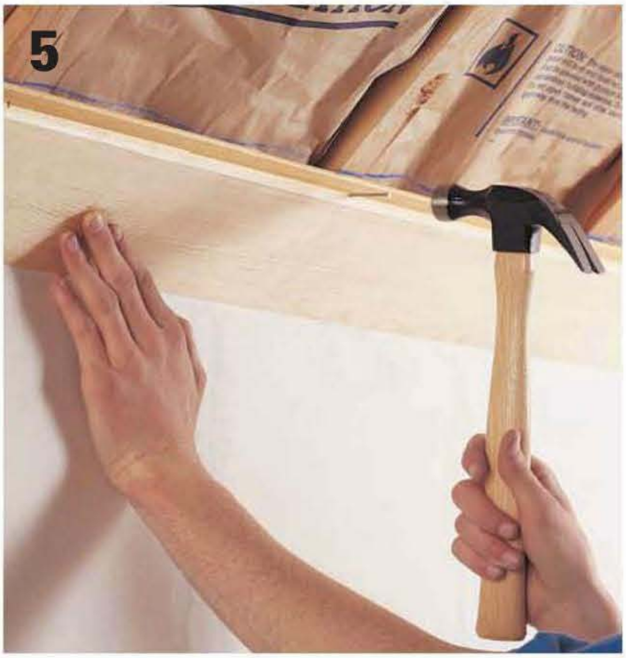

Position the first starter board so the tongue is on the control line. Leave a ⅛" gap between the square board end and the end wall. Fasten the board by nailing through its face about 1" from the grooved edge and into the rafters. Then, blind-nail through the base of the tongue into each rafter, angling the nail backward at 45°. Drive the nail heads beneath the wood surface, using a nail set.

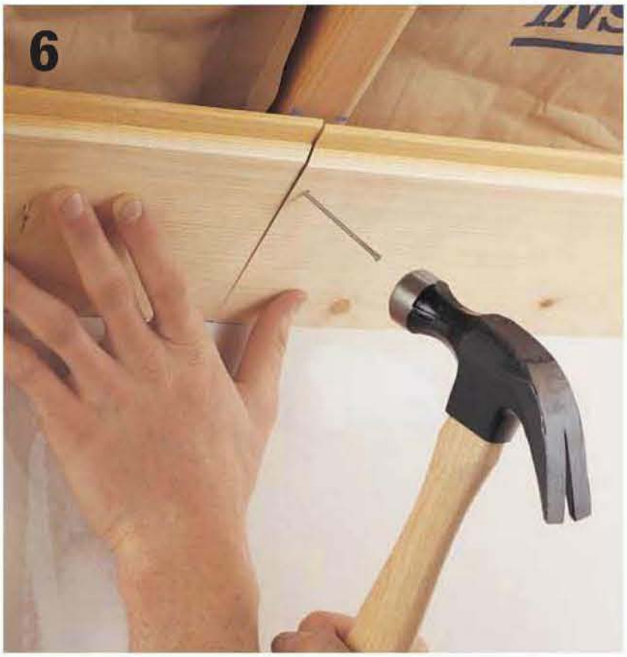

Cut and install any remaining boards in the starter row one at a time, making sure the scarf joints fit together tightly. At each scarf joint, drive two nails through the face of the top board, angling the nail to capture the end of the board behind it. If necessary, predrill the nail holes to prevent splitting.

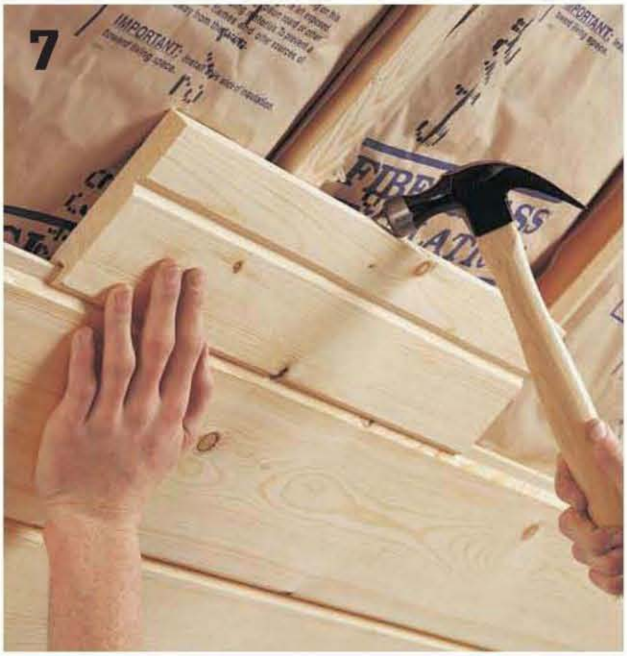

Cut the first board for the next row, then fit its grooved edge over the tongue of the board in the starter row. Use a hammer and a scrap piece of paneling to drive downward on the tongue edge, seating the grooved edge over the tongue of the starter board. Fasten the second row with blind-nails only.

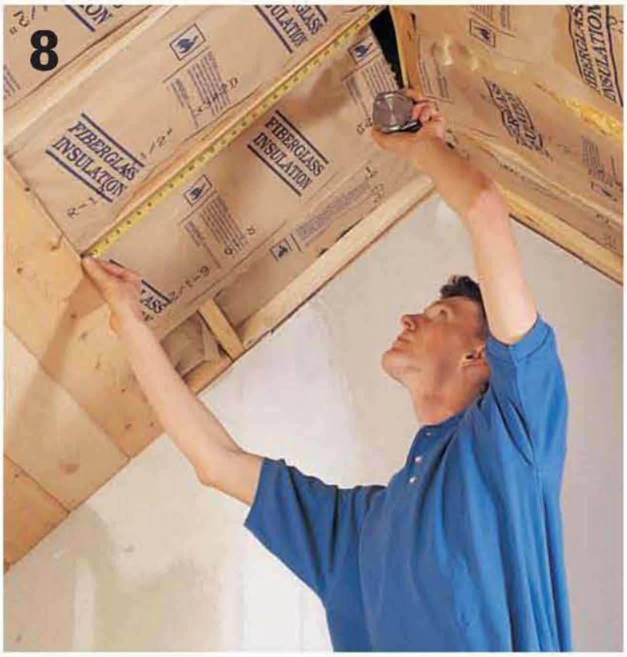

As you install successive rows, measure down from the peak to make sure the rows remain parallel to the peak. Correct any misalignment by adjusting the tongue-and-groove joint slightly with each row. You can also snap additional control lines to help align the rows.

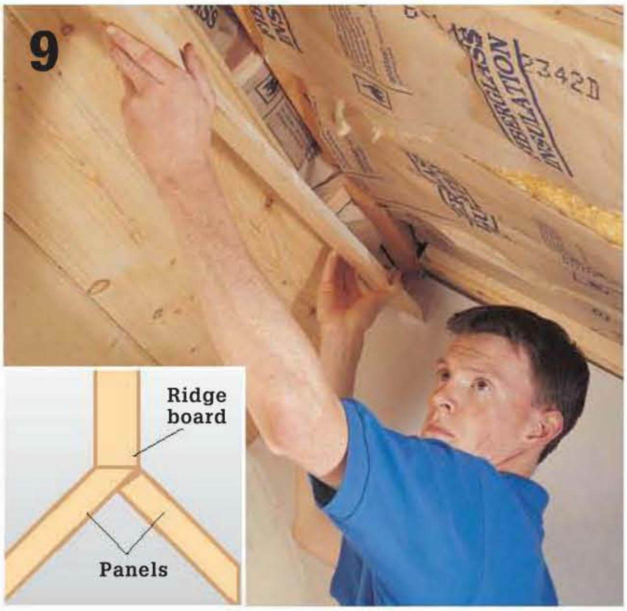

Rip the boards for the last row to width, beveling the top edges so they fit flush against the ridge board. Facenail the boards in place. Install paneling on the other side of the ceiling, then cut and install the final row of panels to form a closed joint under the ridge board (inset).

Install trim molding along walls, at joints around obstacles, and along inside and outside corners, if desired. (Select-grade 1 × 2 works well as trim along walls.) Where necessary, bevel the back edges of the trim or miter-cut the ends to accommodate the slope of the ceiling.

Tips for Paneling an Attic Ceiling ▸

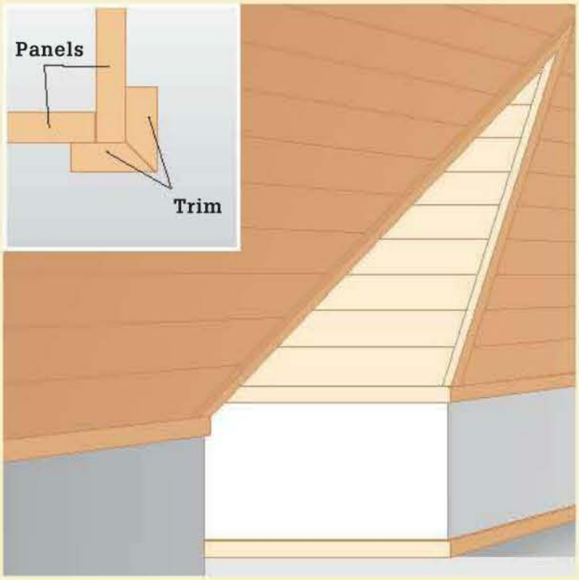

Use mitered trim to cover joints where panels meet at outside corners. Dormers and other roof elements create opposing ceiling angles that can be difficult to panel around. It may be easier to butt the panels together and hide the butt joints with custom-cut trim. The trim also makes a nice transition between angles.

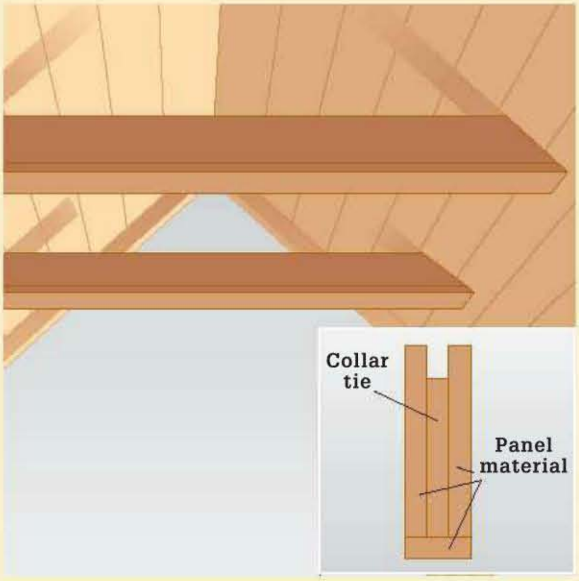

Wrap collar ties or exposed beams with custom-cut panels. Install the paneling on the ceiling first. Then, rip-cut panels to the desired width. You may want to include a tongue-and-groove joint as part of the trim detail. Angle-cut the ends of the trim so it fits tight to the ceiling panels.

Installing Suspended Ceilings

Suspended ceilings are traditionally popular ceiling finishes for basements and utility areas, particularly because they hang below pipes and other mechanicals while providing easy access to them. However, the commercial appearance and grainy texture of basic ceiling tiles make them an unlikely choice for formal areas such as living rooms. Basic tiles are not your only option.

Suspended ceiling tile manufacturers have a wide array of ceiling tiles to choose from that go above and beyond traditional institutional tiles. Popular styles mimic historical tin tiles and add depth to the ceiling while minimizing sound and vibration noise.

A suspended ceiling is a grid framework made of lightweight metal brackets hung on wires attached to ceiling or floor joists. The frame consists of T-shaped main beams (mains), cross-tees (tees), and L-shaped wall angles. The grid supports ceiling panels, which rest on the flanges of the framing pieces. Panels are available in 2 × 2-ft. or 2 × 4-ft. sizes, in a variety of styles. Special options include insulated panels, acoustical panels that absorb sound, and light-diffuser screens for use with fluorescent lights. Generally, metal-frame ceiling systems are more durable than ones made of plastic.

To begin your ceiling project, devise the panel layout based on the size of the room, placing equally sized trimmed panels on opposite sides to create a balanced look. Your ceiling must also be level. For small rooms, a 4-foot or 6-foot level will work, but a water level is more effective for larger jobs. You can make a water level with two water-level ends (available at hardware stores and home centers) attached to flexible plastic tubing.

Tools & Materials ▸

- Water level
- Chalk line
- Drill
- Aviation snips
- String
- Lock-type clamps
- Screw-eye driver
- Pliers
- Straightedge
- Utility knife
- Suspended ceiling kit (frame)
- Screw eyes
- Hanger wires
- Ceiling panels
- 1½" drywall screws or masonry nails

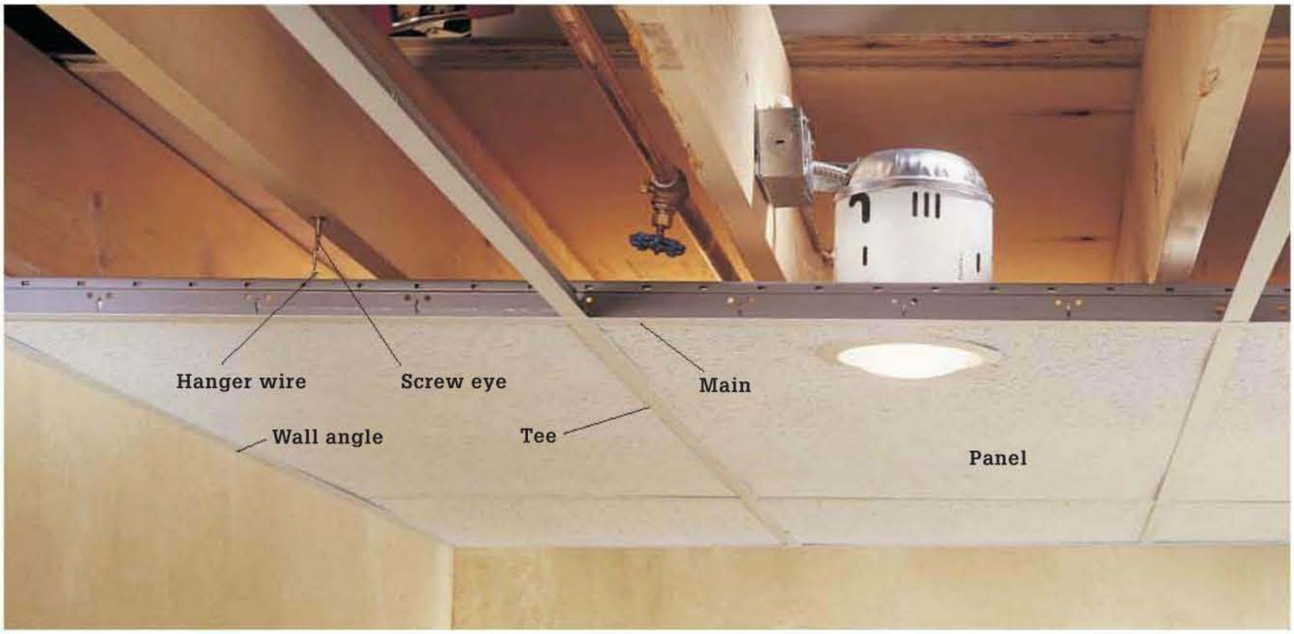

Suspended ceilings are the most common ceiling choice in areas, such as basements, where access to the joist cavities needs to be maintained. They're also installed as a way to conceal problem ceilings.

Installing a Suspended Ceiling

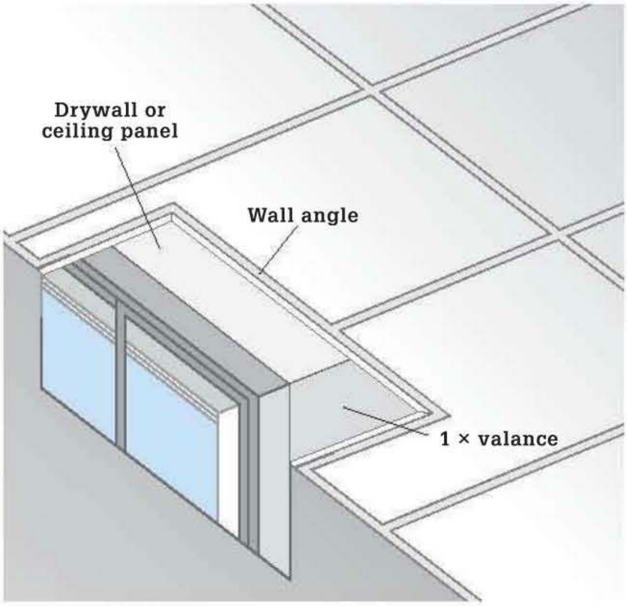
Drywall or ceiling panel
Wall angle
1 × valance

Build a valance around basement awning windows so they can be opened fully. Attach 1 × lumber of an appropriate width to joists or blocking. Install drywall (or a suspended-ceiling panel trimmed to fit) to the joists inside the valance.

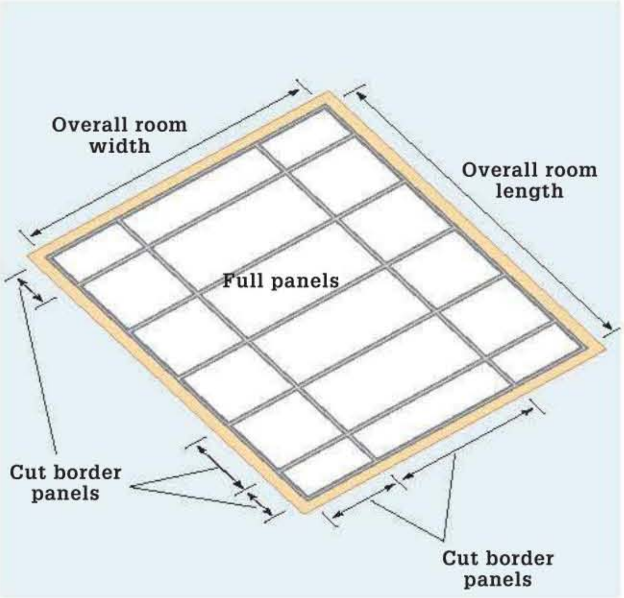
Overall room width
Overall room length
Full panels
Cut border panels
Cut border panels

Draw your ceiling layout on paper, based on the exact dimensions of the room. Plan so that trimmed border panels on opposite sides of the room are of equal width and length (avoid panels smaller than ½-size). If you include lighting fixtures in your plan, make sure they follow the grid layout.

How to Install a Suspended Ceiling

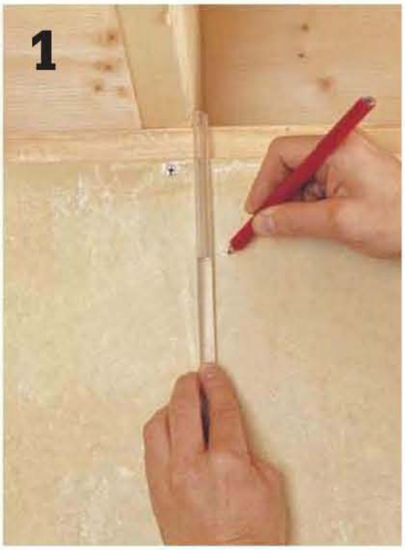

Make a mark on one wall that represents the ceiling height plus the height of the wall angle. Use a water level to transfer that height to both ends of each wall. Snap a chalk line to connect the marks. This line represents the top of the ceiling's wall angle.

Attach wall angle pieces to the studs on all walls, positioning the top of the wall angle flush with the chalk line. Use 1½" drywall screws (or short masonry nails driven into mortar joints on concrete block walls). Cut angle pieces using aviation snips.

Tip ▸

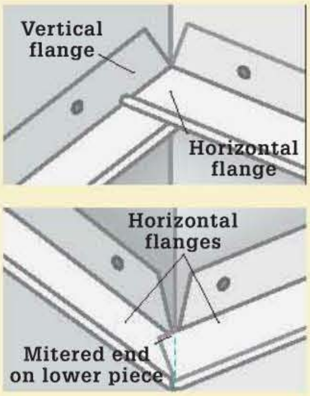
Vertical flange
Horizontal flange
Horizontal flanges
Mitered end on lower piece

Trim wall angle pieces to fit around corners. At inside corners (top), back-cut the vertical flanges slightly, then overlap the horizontal flanges. At outside corners (bottom), miter-cut one horizontal flange, and overlap the flanges.

(continued)

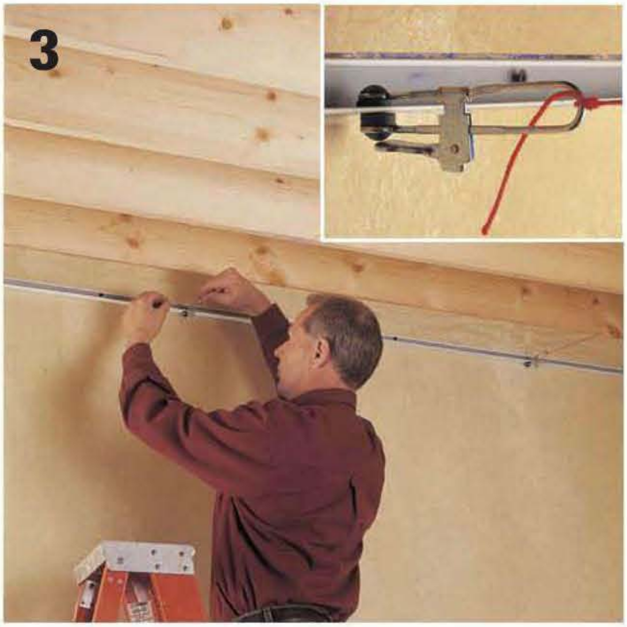

Mark the location of each main on the wall angles at the ends of the room. The mains must be parallel to each other and perpendicular to the ceiling joists. Set up a guide string for each main, using a thin string and lock-type clamps (inset). Clamp the strings to the opposing wall angles, stretching them very taut so there's no sagging.

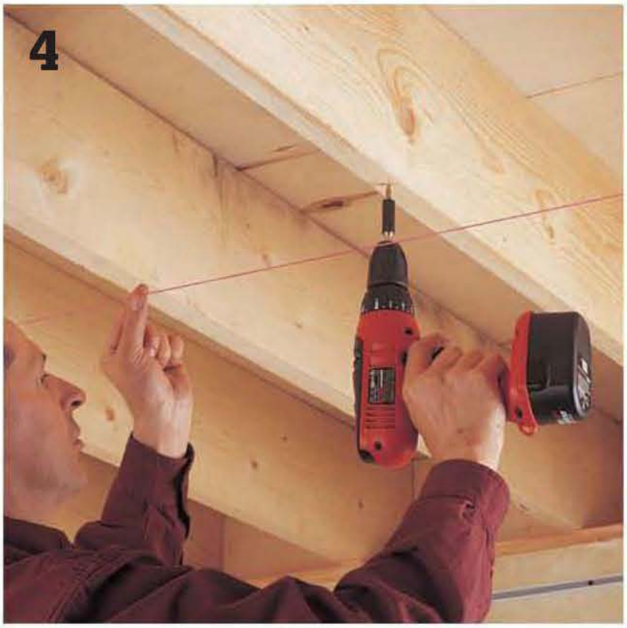

Install screw eyes for hanging the mains, using a drill and screw-eye driver. Drill pilot holes and drive the eyes into the joists every 4 ft., locating them directly above the guide strings. Attach hanger wire to the screw eyes by threading one end through the eye and twisting the wire on itself at least three times. Trim excess wire, leaving a few inches of wire hanging below the level of the guide string.

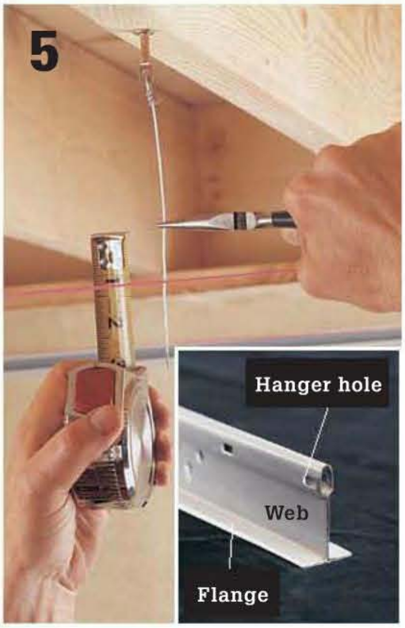

Measure the distance from the bottom of a main's flange to the hanger hole in the web (inset). Use this measurement to prebend each hanger wire. Measure up from the guide string and make a 90° bend in the wire, using pliers.

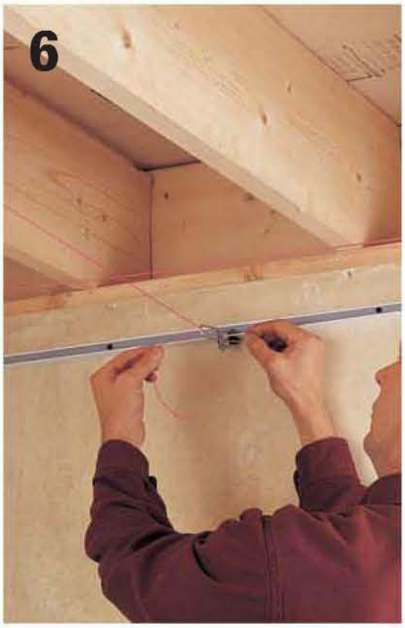

Following your ceiling plan, mark the placement of the first tee on opposite wall angles at one end of the room. Set up a guide string for the tee, using a string and clamps, as before. This string must be perpendicular to the guide strings for the mains.

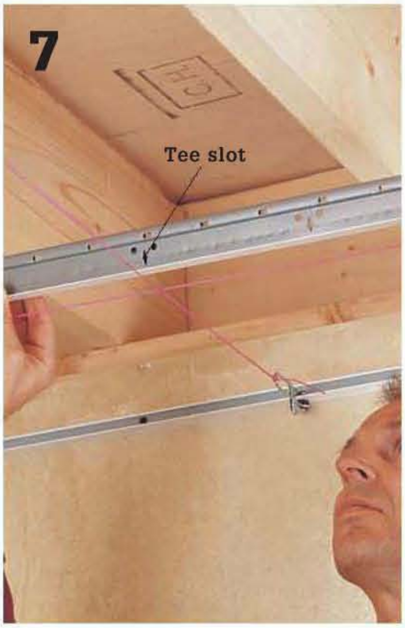

Trim one end of each main so that a tee slot in the main's web is aligned with the tee guide string, and the end of the main bears fully on a wall angle. Set the main in place to check the alignment of the tee slot with the string.

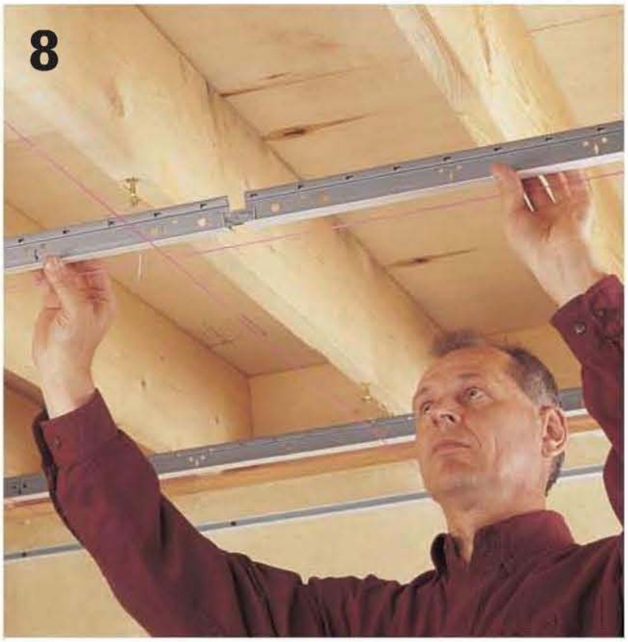

Cut the other end of each main to fit, so that it rests on the opposing wall angle. If a single main cannot span the room, splice two mains together, end-to-end (the ends should be fashioned with male-female connectors). Make sure the tee slots remain aligned when splicing.

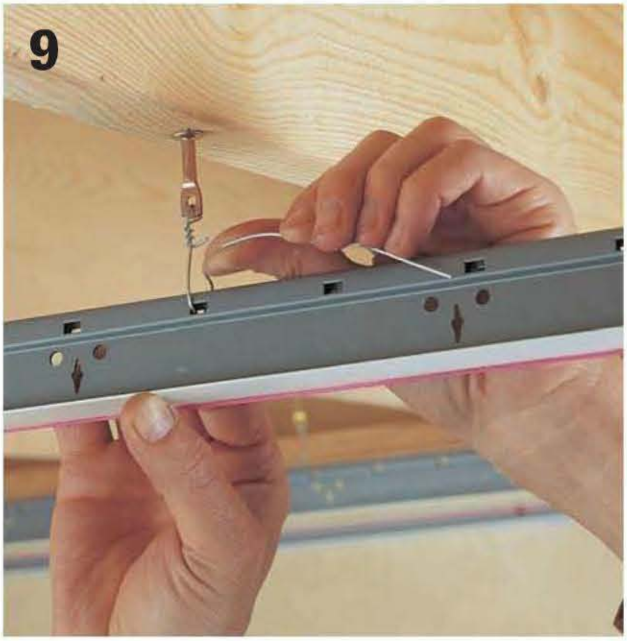

Install the mains by setting the ends on the wall angle and threading the hanger wires through the hanger holes in the webs. The wires should be as close to vertical as possible. Wrap each wire around itself three times, making sure the main's flange is level with the main guide string. Also install a hanger near each main splice.

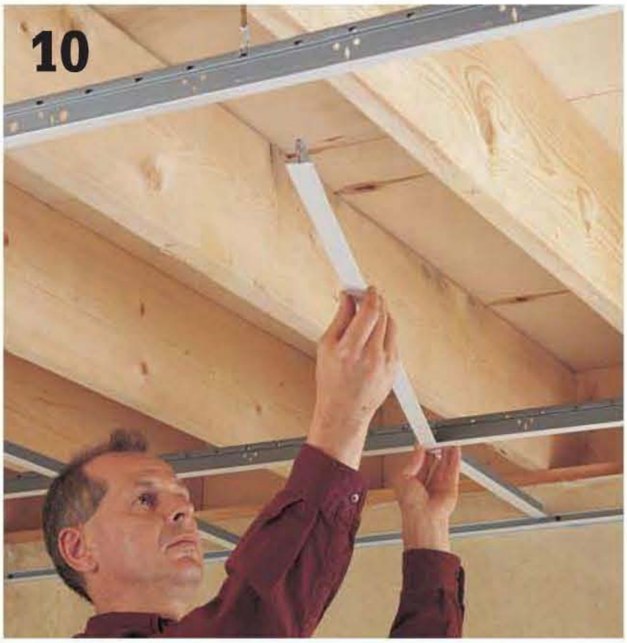

Attach tees to the mains, slipping the tabbed ends into the tee slots on the mains. Align the first row of tees with the tee guide string; install the remaining rows at 4-ft. intervals. If you're using 2 × 2-ft. panels, install 2-ft. cross-tees between the midpoints of the 4-ft. tees. Cut and install the border tees, setting the tee ends on the wall angles. Remove all guide strings and clamps.

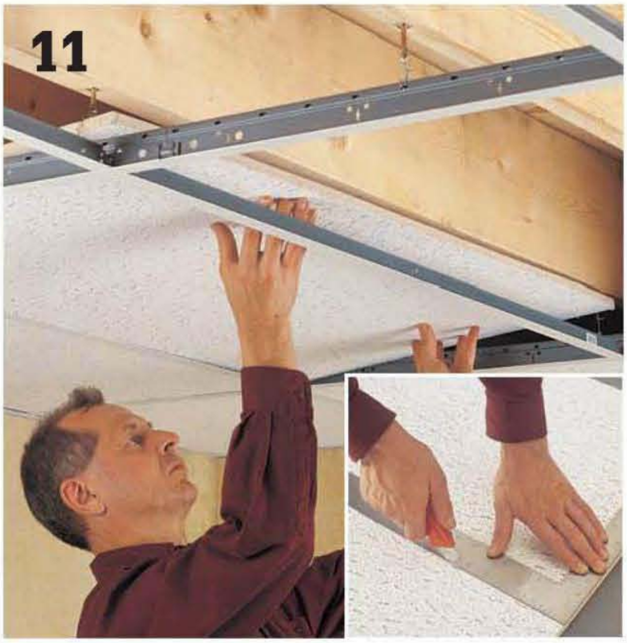

Place full ceiling panels into the grid first, then install the border panels. Lift the panels in at an angle, and position them so they rest on the frame's flanges. Reach through adjacent openings to adjust the panels, if necessary. To trim the border panels to size, cut them face-up, using a straightedge and utility knife (inset).

Installing Acoustical Ceiling Tiles

Easy-to-install ceiling tile can lend character to a plain ceiling or help turn an unfinished basement or attic into beautiful living space. Made of pressed mineral and fiberboard, ceiling tiles are available in a variety of styles. They also provide moderate noise reduction.

Ceiling tiles typically can be attached directly to a drywall or plaster ceiling with adhesive. If your ceiling is damaged or uneven, or if you have an unfinished joist ceiling, install 1 × 2 furring strips as a base for the tiles, as shown in this project. Some systems include metal tracks for clip-on installation.

Unless your ceiling measures in even feet, you won't be able to install the 12-inch tiles without some cutting. To prevent an unattractive installation with small, irregular tiles along two sides, include a course of border tiles along the perimeter of the installation. Plan so that tiles at opposite ends of the room are cut to the same width and are at least half the width of a full tile.

Most ceiling tile comes prefinished, but it can be painted to match any decor. For best results, apply two coats of paint using a roller with a ¼" nap. Wait 24 hours between coats.

Tools & Materials ▸

- 4-ft. level
- Stepladder
- Chalk line
- Utility knife
- Straightedge
- Hammer or drill
- Handsaw
- Stapler
- 1 × 2 furring strips
- 8d nails or 2" screws
- String
- Ceiling tiles
- Staples
- Trim molding

Acoustic tile improves the sound quality inside a home theater and reduces the transmission of sound to the rooms surrounding it.

How to Install Ceiling Tile

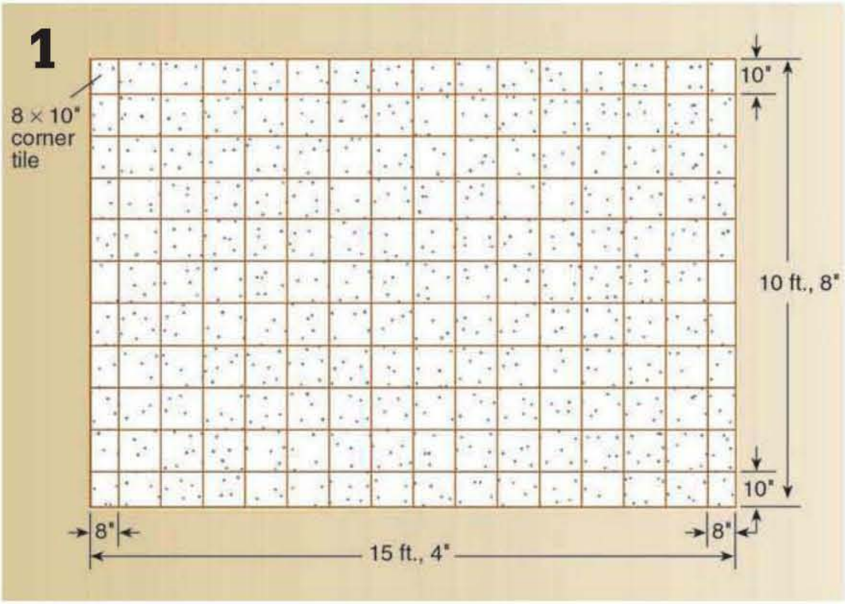

Measure the ceiling and devise a layout. If the length (or width) doesn't measure in even feet, use this formula to determine the width of the border tiles: add 12 to the number of inches remaining and divide by 2. The result is the width of the border tile. (For example, if the room length is 15 ft., 4", add 12 to the 4, then divide 16 by 2, which results in an 8" border tile.)

Install the first furring strip flush with the wall and perpendicular to the joists, fastening with two 8d nails or 2" screws at each joist. Measure out from the wall a distance equal to the border tile width minus ¾", and snap a chalk line. Install the second furring strip with its wall-side edge on the chalk line.

Install the remaining strips 12" on-center from the second strip. Measure from the second strip and mark the joist nearest the wall every 12". Repeat along the joist on the opposite side of the room, then snap chalk lines between the marks. Install the furring strips along the lines. Install the last furring strip flush against the opposite side wall. Stagger the butted end joints of strips between rows so they aren't all on the same joist.

Check the strips with a 4-ft. level. Insert wood shims between the strips and joists as necessary to bring the strips into a level plane.

(continued)

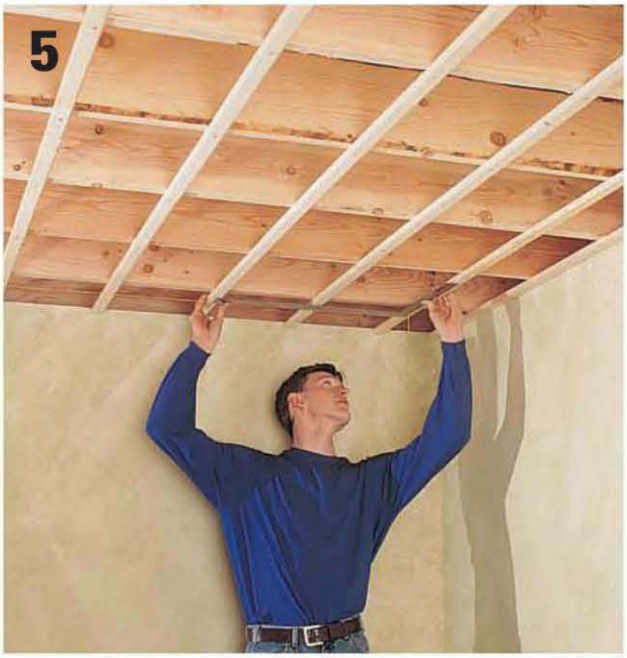

Set up taut, perpendicular string lines along two adjacent walls to help guide the tile installation. Inset the strings from the wall by a distance that equals that wall's border tile width plus ½". Use a framing square to make sure the strings are square.

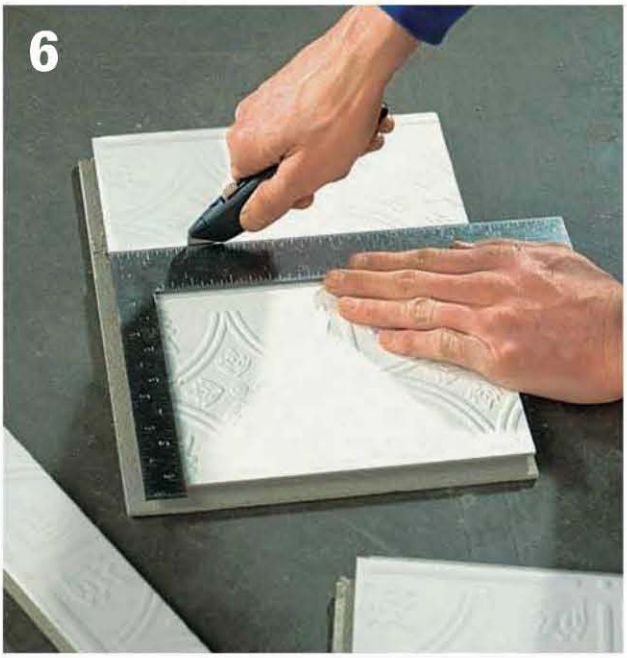

Cut the corner border tile to size with a utility knife and straightedge. Cutting the border tiles ¼" short will ease fitting them. The resulting gap between the tile and wall will be covered by trim. Cut only on the edges without the stapling flange.

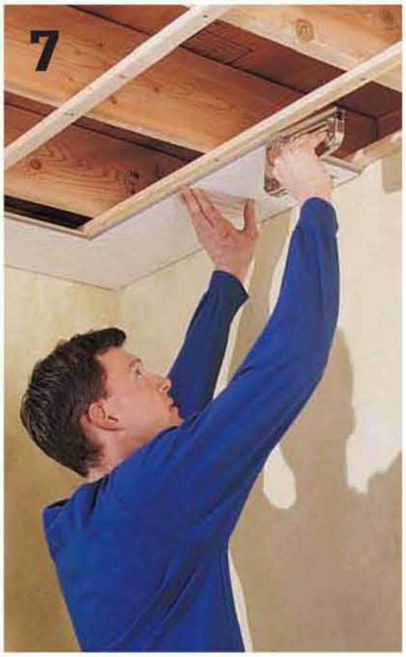

Position the corner tile with the flange edges aligned with the two string lines and fasten it to the furring strips with four ½" staples. Cut and install two border tiles along each wall, making sure the tiles fit snugly together.

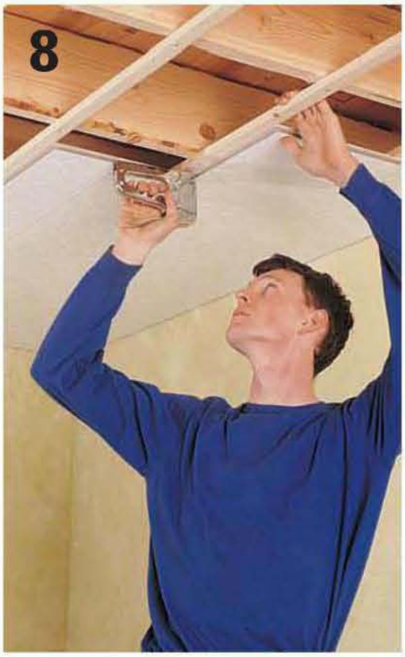

Fill in between the border tiles with full-size tiles. Continue working diagonally in this manner, toward the opposite corner. For the border tiles along the far wall, trim off the flange edges and staple through the faces of the tiles, close to the wall.

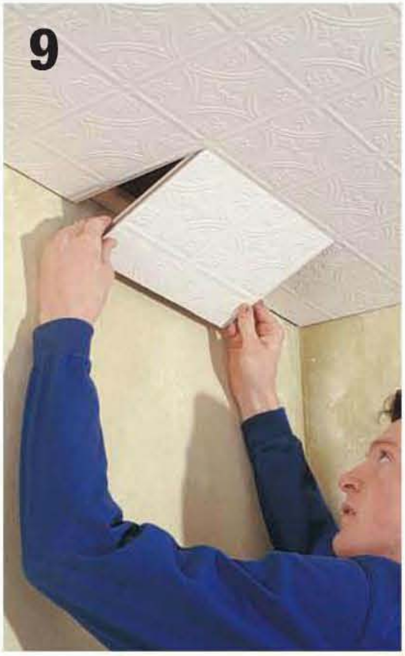

Install the final row of tiles, saving the far corner tile and its neighbor for last. Cut the last tile to size, then remove the tongue and nailing flange along the side edges. Finish the job by installing trim along the edges.

Installing Metal Ceilings

Today's metal ceilings offer the distinctive elegance of 19th-century tin tile in a durable, washable ceiling finish. Available at home centers and specialty distributors, metal ceiling systems include field panels (in 2 × 2-, 2 × 4-, and 2 × 8-ft. sizes), border panels that can be cut to fit your layout, and cornice molding for finishing the edges. The panels come in a variety of materials and finishes ready for installation, or they can be painted.

To simplify installation, the panels have round catches, called nailing buttons, that fit into one another to align the panels where they overlap. The buttons are also the nailing points for attaching the panels. Use 1" decorative conehead nails where nail heads will be exposed, and ½" wire nails where heads are hidden.

Install your metal ceiling over a smooth layer of ⅜" or ½" plywood, which can be fastened directly to the ceiling joists with drywall screws, or installed over an existing finish. The plywood provides a flat nailing surface for the panels. As an alternative, some manufacturers offer a track system for clip-on installation.

Begin your installation by carefully measuring the ceiling and snapping chalk lines to establish the panel layout. For most tile patterns, it looks best to cover the center of the space with full tiles only, then fill in along the perimeter with border panels, which are not patterned. Make sure your layout is square.

Tools & Materials ▸

- Chalk line
- Level
- Tin snips
- Drill with ⅛" metal bit
- Compass
- Metal file
- ⅜" or ½" plywood
- 2" drywall screws
- Field panels
- Border panels with molding edge
- Cornice molding
- Masking tape
- ½" wire nails
- 1" conehead nails
- Wood block

Real metal ceilings have traditional embossed patterns with an unmistakable luxurious quality. But, they are expensive and a bit unwieldy to install. Acoustic panels with an embossed vinyl pattern layer are much cheaper and easier to install, but are also less authentic.

How to Install a Metal Tile Ceiling

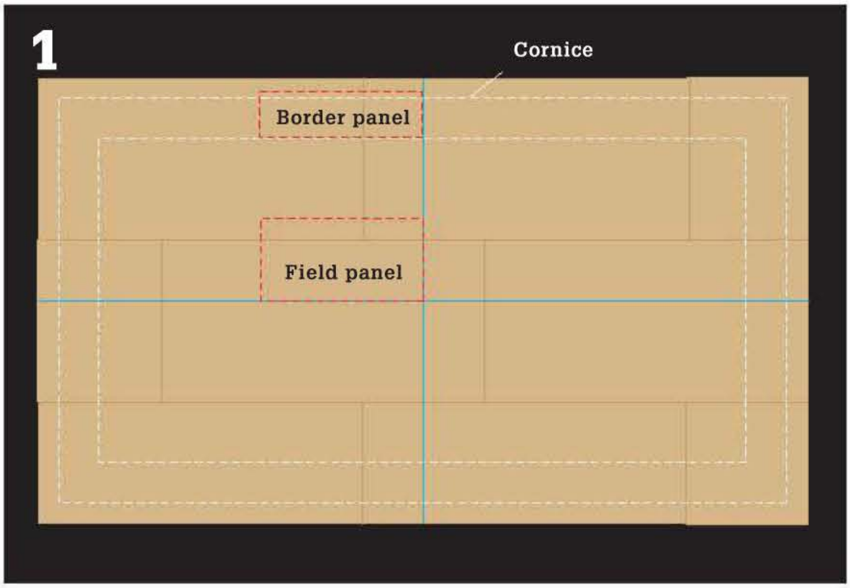

Measure to find the center of the ceiling, then snap perpendicular chalk lines intersecting the center. On the walls, mark a level reference line representing the bottom edges of the cornice molding. Where possible, plan to install the panels so they overlap toward the room's entrance, to help conceal the seams.

Align the first field panel with the chalk lines at the ceiling's center, and attach it with ½" wire nails along the edges where another panel will overlap it. Drive the nails beside the nailing buttons—saving the buttons for nailing the overlapping panel.

Continue to install field panels, working along the length of the area first, then overlapping the next row. Make sure the nailing buttons are aligned. Underlap panels by sliding the new panel into position beneath the installed panel, then fasten through both panels at the nailing buttons, using 1" conehead nails. Where field panels meet at corners, drill ⅛" pilot holes for the conehead nails.

Cut the border panels to width so they will underlap the cornice by at least 1". Use sharp tin snips, and cut from the edge without edge molding. Install the panels so the nailing buttons on the molding align with those on the field panels. Fasten through the buttons with conehead nails, and along the cut edge with wire nails. At corners, miter-cut the panels, and drive conehead nails every 6" along the seam.

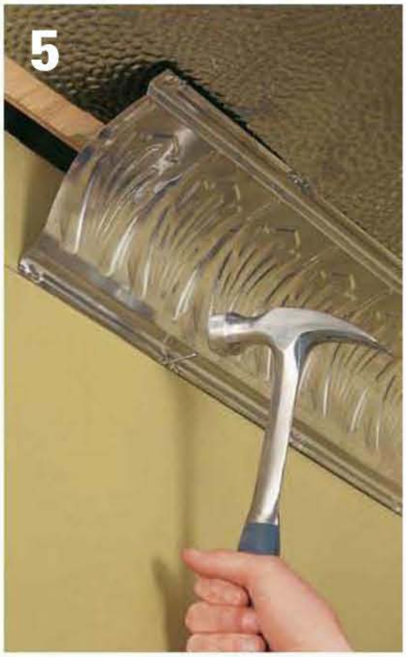

Install each cornice piece with its bottom edge on the level line. Drive 1" conehead nails through the nailing buttons and into the wall studs. Don't nail the ends until the succeeding piece is in place. Fasten the top edges to the ceiling.

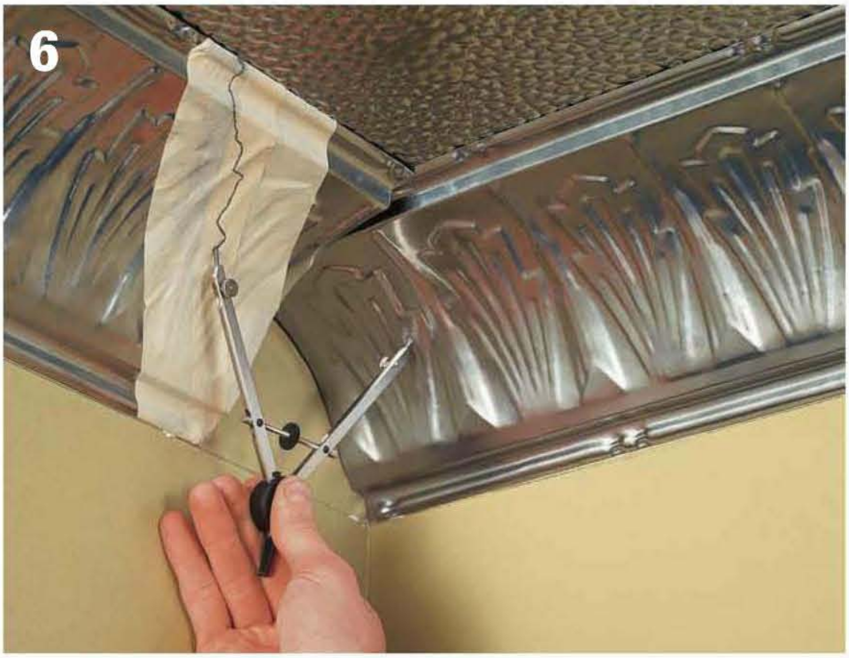

At inside corners, install one cornice piece tightly into the corner, then scribe the mating piece to fit, using masking tape and a compass. Cut along the scribed line with tin snips, and make minor adjustments with a metal file. You may have to cut the mating piece several times, so start with plenty of length. If you have several corners, use this technique to cut templates for the corner pieces.

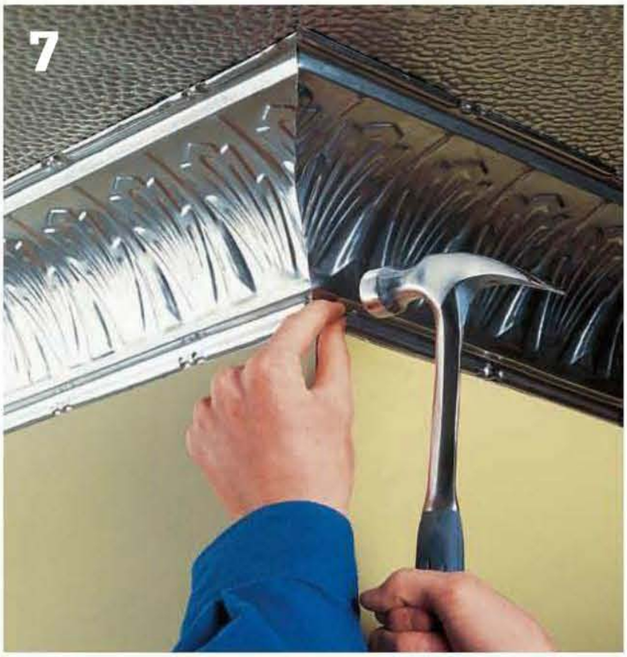

At outside corners, cut the ends of two scrap pieces at a 33° angle. Fit the pieces together at the corner, then trim and mark each piece in turn, making minor adjustments until they fit well. Use the scrap pieces as templates for marking the workpieces. Fasten near the corner only when both mating pieces are in place.

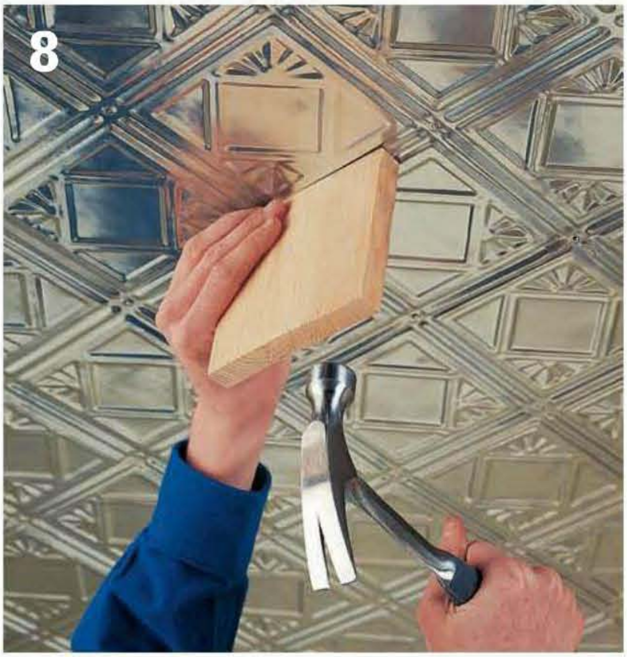

Using a hammer and a piece of wood, carefully tap any loose joints to tighten them. If the cornice will be left unpainted, file the joints for a perfect fit. If you're painting the ceiling, seal the seams with paintable silicone caulk, then apply two coats of paint using a roller with a ¼" nap. Allow the first coat to dry for 24 hours before applying the second coat.

Installing Base Molding

Baseboard trim is installed to conceal the joint between the finished floor and the wallcovering. It also serves to protect the wallboard at the floor. Installing plain, one-piece baseboard such as ranch-style base or cove base is a straightforward project. Outside corner joints are mitered, inside corners are coped, and long runs are joined with scarf cuts.

The biggest challenge to installing base is dealing with out-of-plumb and nonsquare corners. However, a T-bevel makes these obstacles easy to overcome.

Plan the order of your installation prior to cutting any pieces and lay out a specific piece for each length of wall. It may be helpful to mark the type of cut on the back of each piece so you don't have any confusion during the install.

Locate all studs and mark them with painter's tape, 6 inches higher than your molding height. If you need to make a scarf joint along a wall, make sure it falls on the center of a stud. Before you begin nailing trim in place, take the time to pre-finish the moldings. Doing so will minimize the cleanup afterward.

Tools & Materials ▸

- Pencil
- Tape measure
- Power miter saw
- T-bevel
- Coping saw
- Metal file set
- Pneumatic finish nail gun & compressor
- Moldings
- Pneumatic fasteners
- Carpenter's glue
- Finishing putty

How to Install One-piece Base Molding

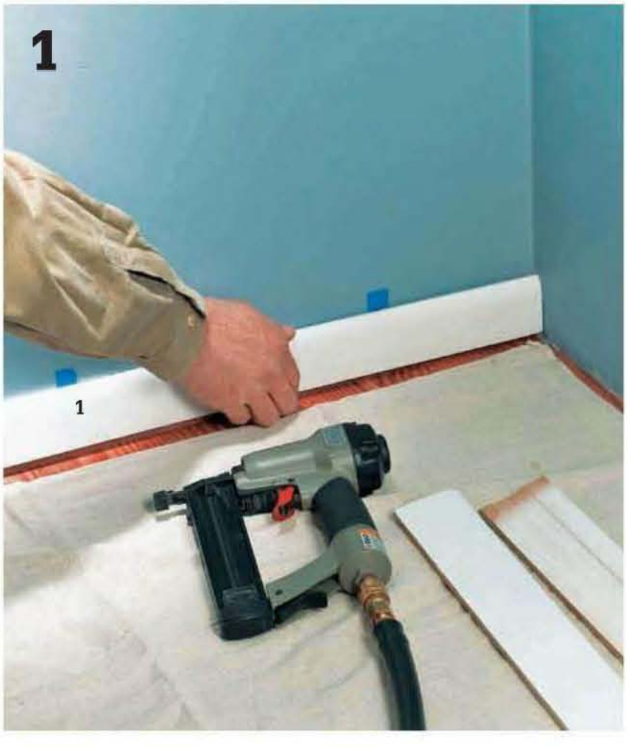

Measure, cut, and install the first piece of baseboard. Butt both ends into the corners tightly. For longer lengths, it is a good idea to cut the piece slightly oversized (up to 1/16" on strips over 10 ft. long) and "spring" it into place. Nail the molding in place with two nails at every stud location.

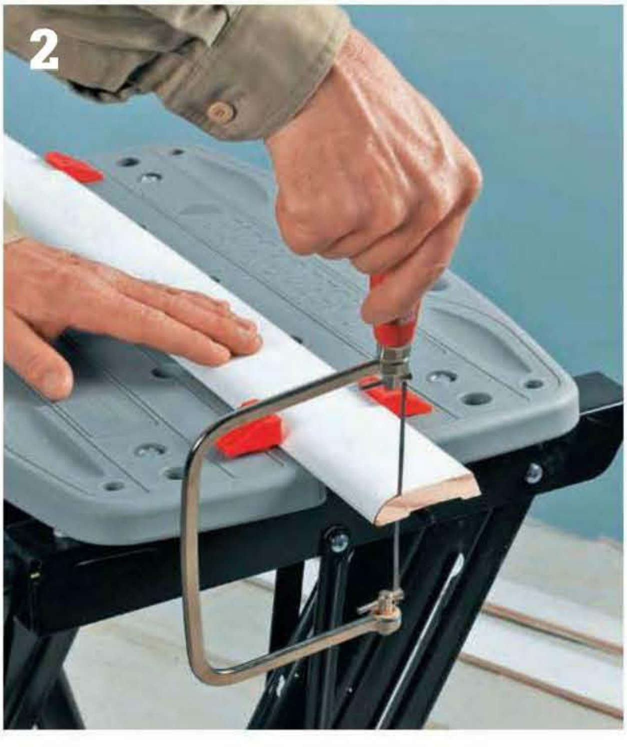

Cut the second piece of molding oversized by 6" to 10" and cope-cut the adjoining end to the first piece. Fine-tune the cope with a metal file and sandpaper. Dry-fit the joint, adjusting it as necessary to produce a tight-fitting joint.

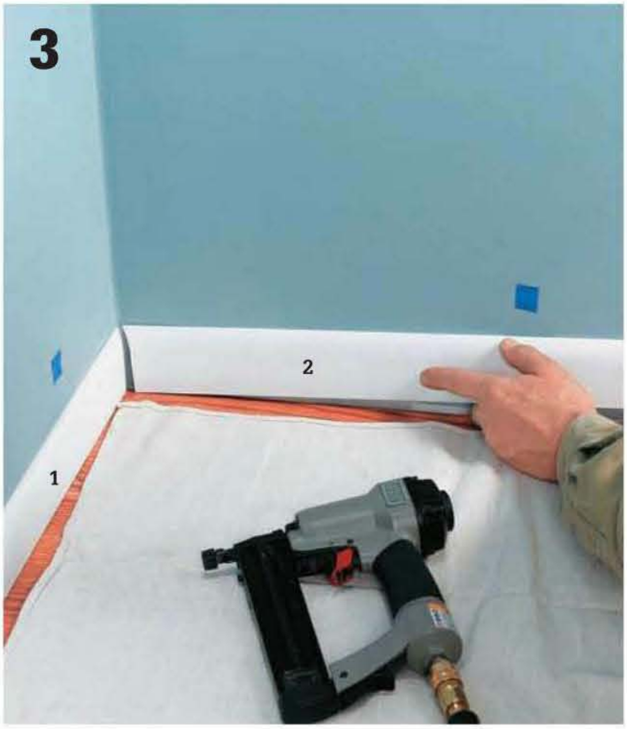

Check the corner for square with a framing square. If necessary, adjust the miter cut of your saw. Use a T-bevel to transfer the proper angle. Cut the second piece (coped) to length and install it with two nails at each stud location.

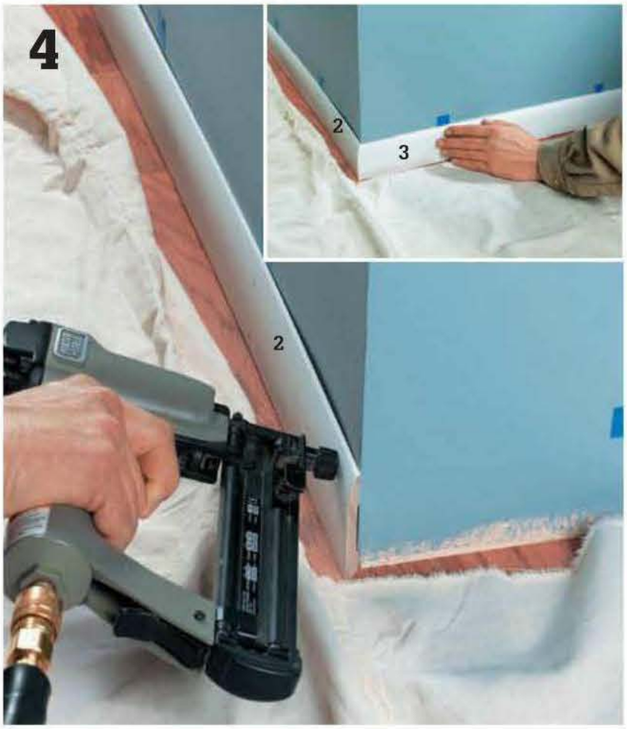

Adjust the miter angle of your saw to cut the adjoining outside corner piece (3). Test-fit the cut to ensure a tight joint (inset photo). Remove the mating piece of trim and fasten the first piece for the outside corner joint.

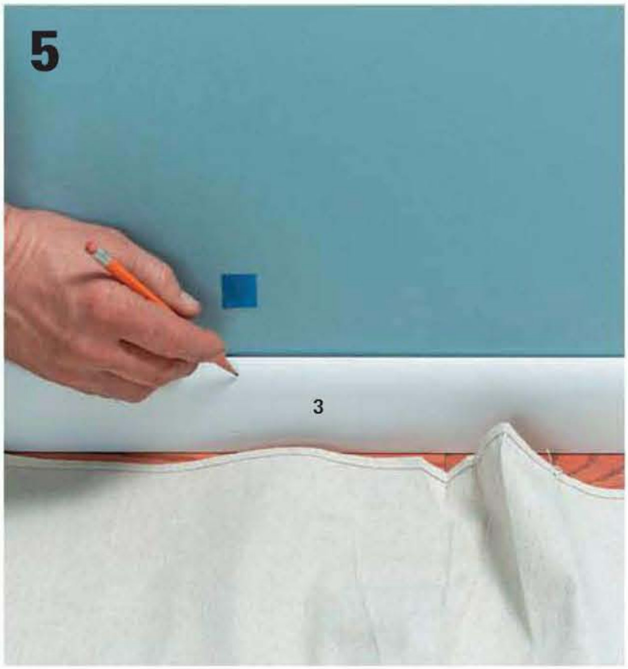

Lay out any scarf joints by placing the piece in position so that the previous joint is tight and then marking the center of a stud location nearest the opposite end. Set the angle of your saw to a 30° angle and cut the molding at the marked location.

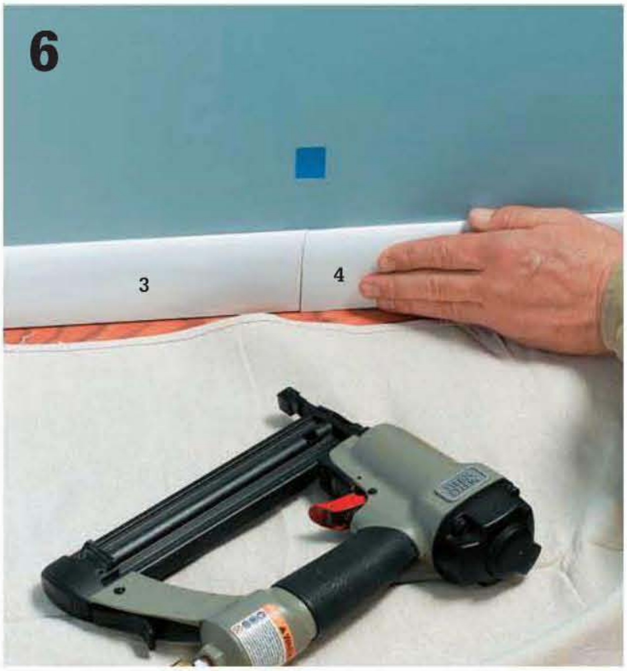

Nail the third piece in place, making sure the outside corner joint is tight. Cut the end of the fourth piece to match the scarf joint angle and nail it in place with two nails at each stud location. Add the remaining pieces of molding, fill the nail holes with putty, and apply a final coat of finish.

How to Install Built-up Base Molding

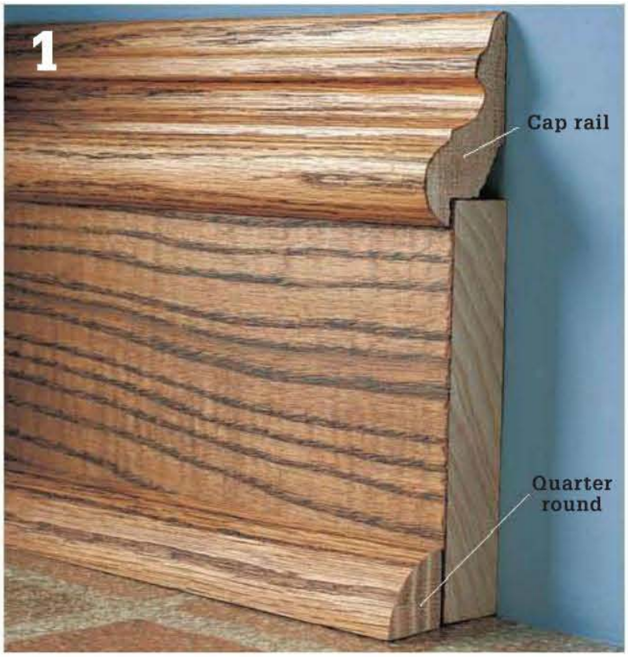
1
Cap rail
Quarter round

Dress up simple baseboard stock with cap moldings and base shoe or quarter round. The baseboard can be made of solid wood, as shown above, or from strips of veneered plywood, as shown at right.

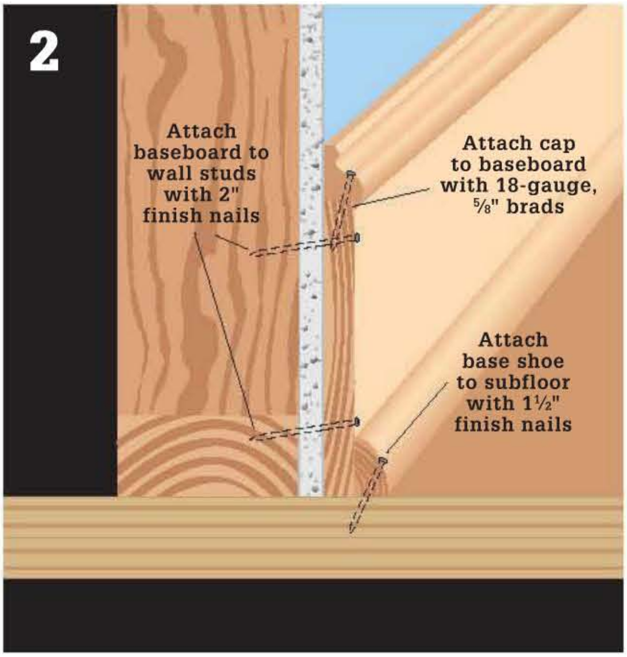
2
Attach baseboard to wall studs with 2" finish nails
Attach cap to baseboard with 18-gauge, ⅝" brads
Attach base shoe to subfloor with 1½" finish nails

Built-up baseboard requires more attention to the nailing schedule than simple one-piece baseboards. The most important consideration (other than making sure your nails are all driven into studs or other solid wood) is that the base shoe must be attached to the floor, while the baseboard is attached to the wall. This way, as the gap between the wall and floor changes, the parts of the built-up molding can change with them.

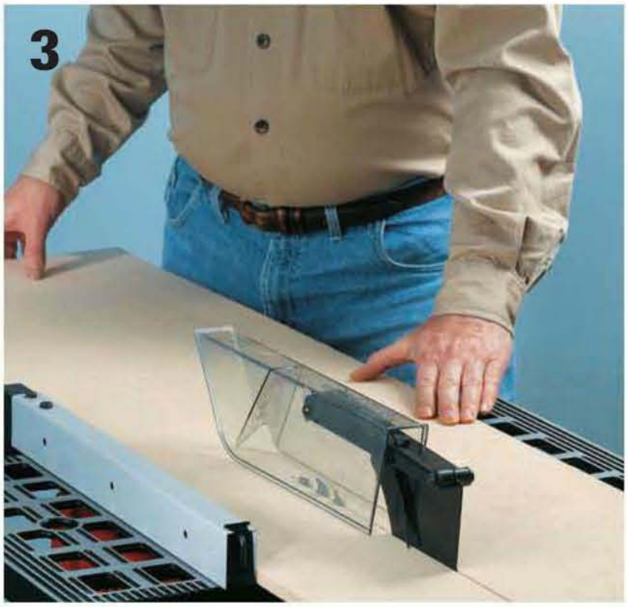

Cut the plywood panel into 6" strips with a table saw or a straightedge guide and a circular saw. Lightly sand the strips, removing any splinters left from the saw. Then, apply the finish of your choice to the moldings and the plywood strips.

Tip ▸

Baseboard can be built up on the back with spacer strips so it will project further out from the wall. This can allow you to match existing casings or to create the impression of a thicker molding. However, the cap rail needs to be thick enough to cover the plywood edge completely, or the core of the panel may be visible.

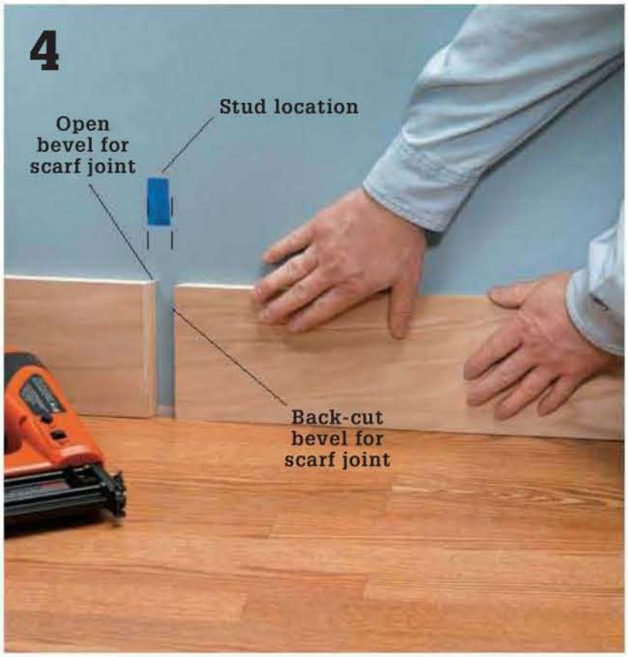
4
Open bevel for scarf joint
Stud location
Back-cut bevel for scarf joint

Install the plywood strips with 2" finish nails driven at stud locations. Use scarf joints on continuous runs, driving pairs of fasteners into the joints. Cut and install moldings so that all scarf joints fall at stud locations.

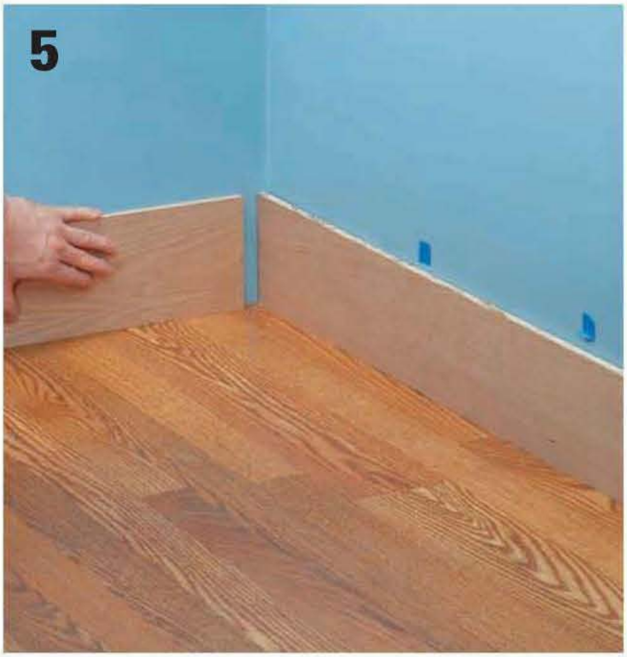

Test-fit inside corner butt joints before cutting a workpiece. If the walls are not square or straight, angle or bevel the end cut a few degrees to fit the profile of the adjoining piece. The cap molding will cover any gaps at the top of the joint.

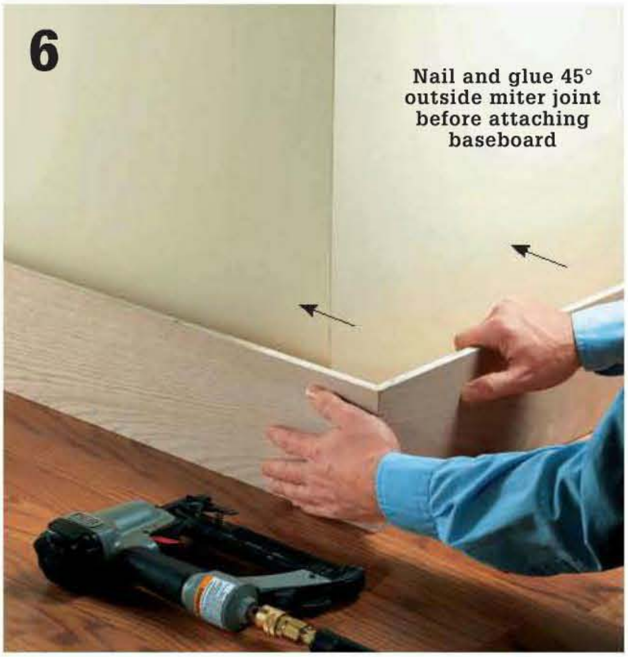
6
Nail and glue 45° outside miter joint before attaching baseboard

Miter outside corners squarely at 45°. Use wood glue and 1¼" brad nails to pull the mitered pieces tight, and then nail the base to the wall at stud locations with 2" finish nails. Small gaps at the bottom or top of the base molding will be covered with cap or base shoe.

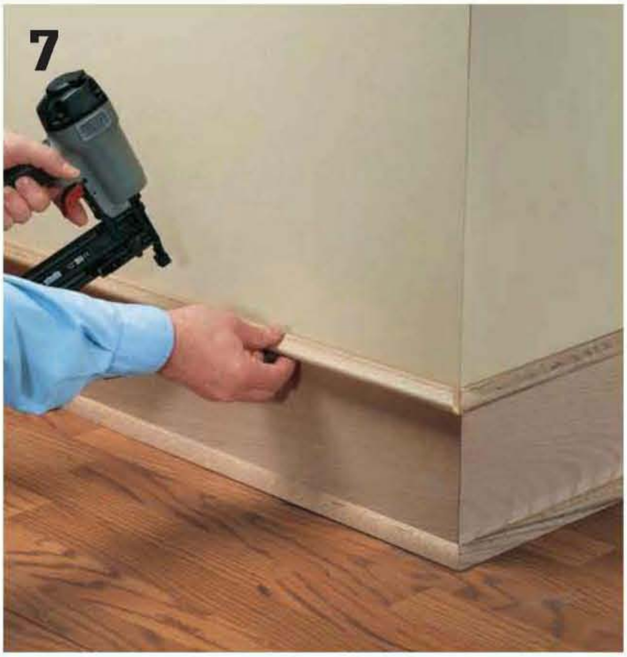

Use a brad nailer with 18-gauge, ⅝" brads to install the cap and base shoe moldings along the edges of the plywood base. Fit scarf joints on longer lengths, coped joints on inside corners, and miter joints on outside corners. Stagger the seams so that they do not line up with the base molding seams, following the suggested nailing pattern (right). Set any protruding nails with a nail set and fill all nail holes with putty.

Installing Picture Rail

Picture rail molding is a specialty molding that was installed in many older homes so the homeowners could avoid making nail holes in the finished walls. Picture rail molding is a simple but elegant way to add style to any room. Special picture hanging hooks slide over the molding and artwork may be hung with a cord over the hook. Picture rail molding also provides its own decorative touch, breaking up the vertical lines from floor to ceiling. For this reason, it is also installed as a decorative touch by itself.

Picture rail molding is easy to install but should be reinforced with screws, not brads or nails, especially if you are hanging large, heavy items. Depending upon the style of your home, picture rail can be hung anywhere from 1 ft. to a few inches down from the ceiling. In some homes, picture rail is added just below the cornice or crown molding to add an additional layer of depth. When applied this way, it is commonly referred to as a frieze board.

In the example shown, the picture rail is installed using a level line to maintain height. If your ceiling is uneven, you may choose to install picture rail a constant distance from the ceiling to avoid an uneven appearance.

Traditionally, picture rail (see top of photo) is used in conjunction with special hooks to hang artwork in formal rooms without penetrating the wall with fasteners.

Tools & Materials ▸

- Ladder
- Pencil
- Stud finder
- Tape measure
- Power miter saw
- T-bevel
- Pneumatic finish nail gun & compressor
- 4-ft. level or laser level
- Drill with bits
- Painter's tape
- Moldings
- Pneumatic fasteners
- 1⅝" wallboard screws
- Hole filler

How to Install Picture Rail Molding

1

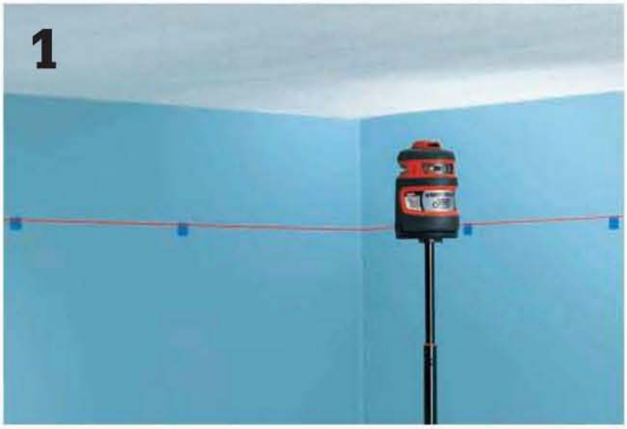

Measure down the desired distance from the ceiling and draw a level reference line around the room using a pencil and a 4-ft. level (or, take advantage of modern technology and use a laser level). While you are up there, use a stud finder to locate the framing members, and mark the locations on the walls with blue painter's tape.

2

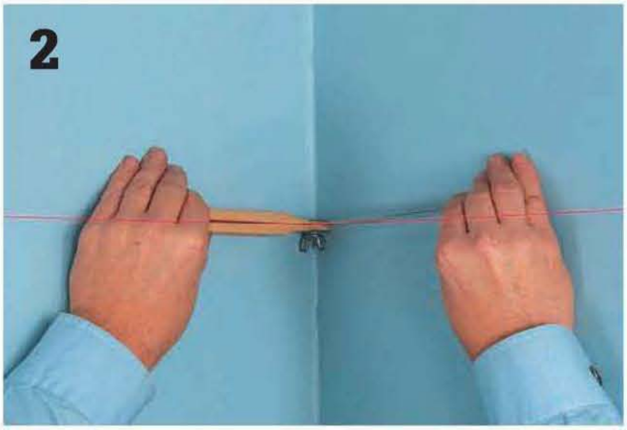

Most corners are close to 90°, but to cut a tight inside corner, the actual angle must be divided exactly in half. Use a T-bevel to measure the angle of the corner, tightening the lock nut with the blade and the handle on the reference line.

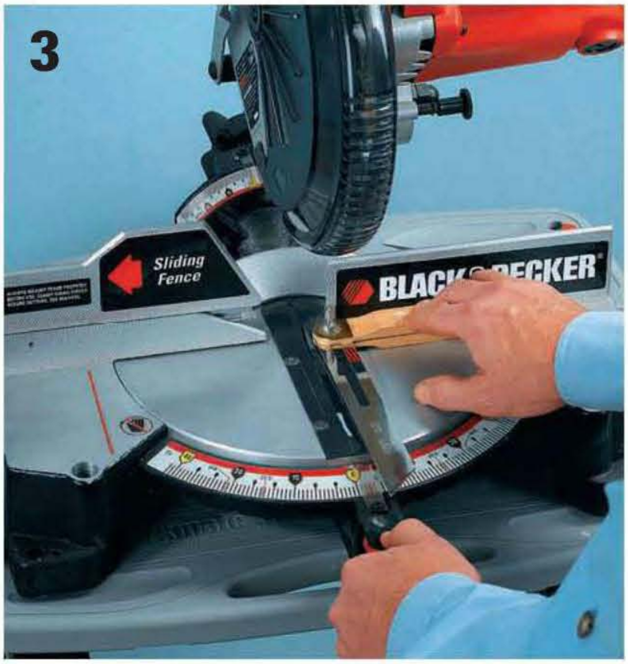

Place the T-bevel on the table of your power miter saw and adjust the miter blade so that it matches the angle. With the T-bevel tight to the fence, read the angle the saw is set to when it aligns with the T-bevel. If the blade is angled to the right of zero degrees the angle is larger than 90; to the left, smaller.

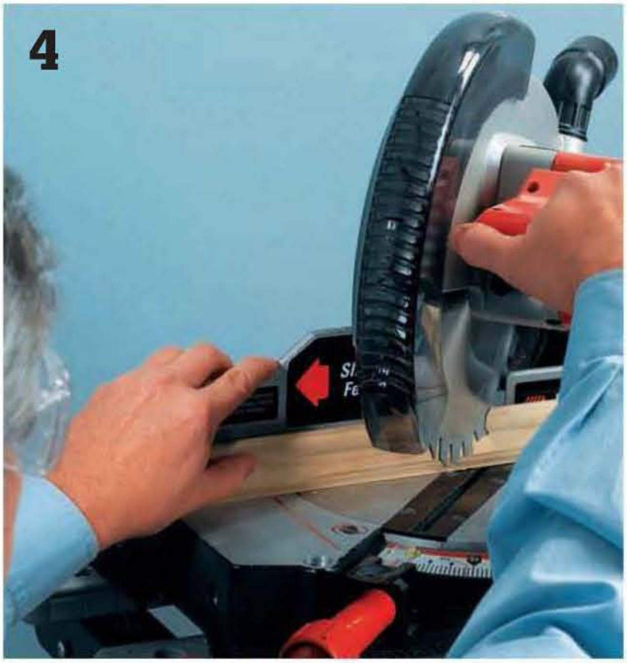

Read the angle from the miter saw table, divide the number by two, and add or subtract that number from 45 degrees to find the proper cutting angle for each corner. Cut each molding slightly longer than the measured length.

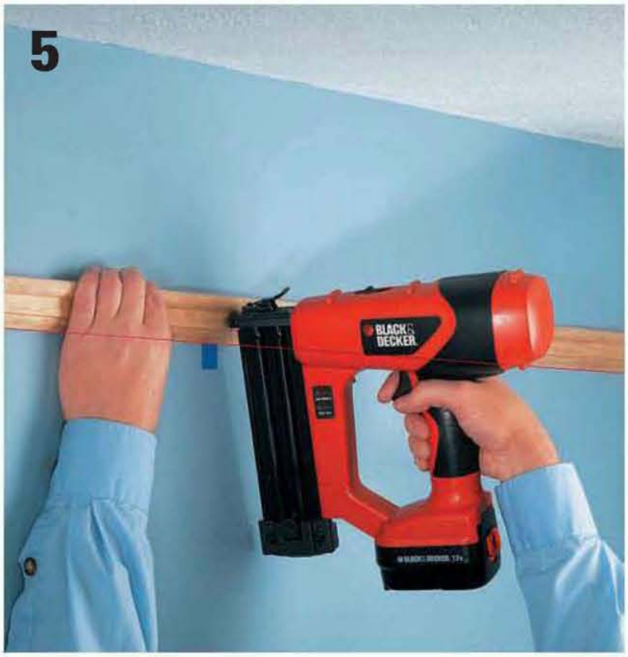

Nail the molding at the stud locations, covering the level line around the room (if you're using a laser level, you simply keep it in position and turned on to cast a reference line you can follow). After each molding is completely nailed in place, go back to each stud location and drive 1⅝" wallboard screws into the molding through counter-bored pilot holes.

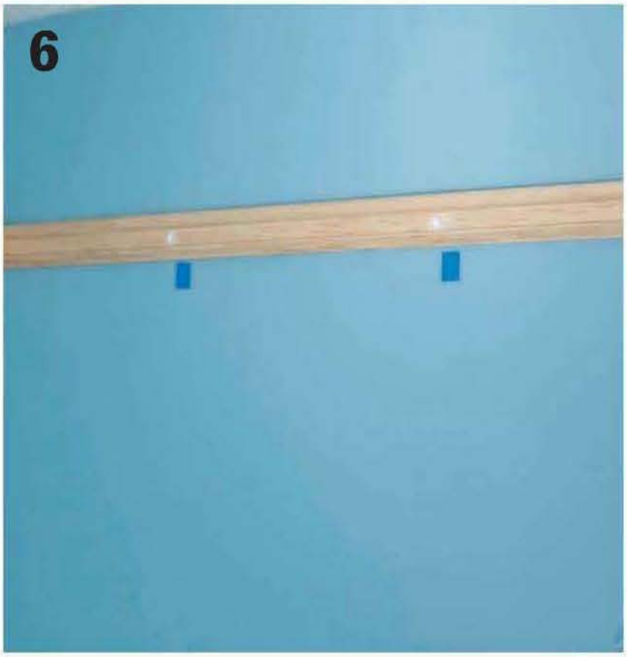

Fill nail holes with wood filler. Let the filler dry and sand it smooth. Then apply a final coat of paint over the molding face.

Installing Chair Rail

Chair rail molding typically runs horizontally along walls at a height of around 36" (the rule of thumb is to install it one-third of the way up the wall). Originally installed to protect walls from collisions with chair backs, today chair rail is commonly used to divide a wall visually. Chair rail may cap wainscoting, serve as a border for wallpaper, or divide two different colors on a wall. Interesting chair rail profiles can be effective alone on a one-color wall.

Stock chair rail moldings are available at most lumberyards and home centers. However, more intricate and elaborate chair rails can be created by combining multiple pieces of trim. Keep in mind the height of your existing furnishings when installing a chair rail. It would be disappointing to discover that the new molding has a bad visual effect with your couch or chair backs when the project is completed.

Chair rail once was installed to protect fragile walls from chair backs, but today it is mainly installed as a decorative accent that breaks up dull walls visually.

Tools & Materials ▸

- Pencil
- Stud finder
- Tape measure
- Power miter saw
- 4-ft. level
- Air compressor
- Finish nail gun
- Metal file set
- Moldings
- Pneumatic fasteners
- Painter's tape
- Carpenter's glue
- Finishing putty
- Finishing materials

How to Install Chair Rail

1

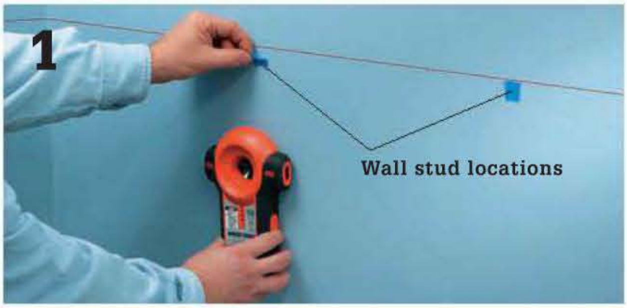

On the starting wall of your installation, measure up to the height at which you plan to install the chair rail, minus the width of the molding. Mark a level line at this height around the room. Locate all studs along the walls and mark their locations with painter's tape below the line.

2

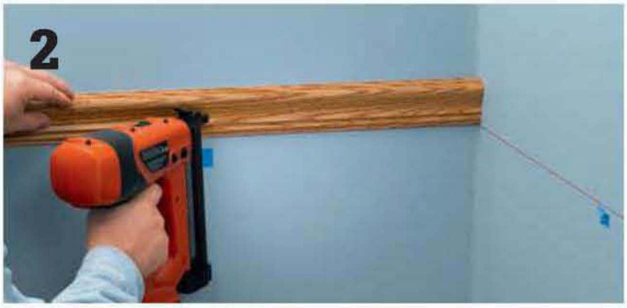

Measure, cut, and install the first piece of chair rail with the ends cut squarely, butting into both walls (in a wall run with two inside corners). Nail the molding in place with two 2" finish nails at each stud location.

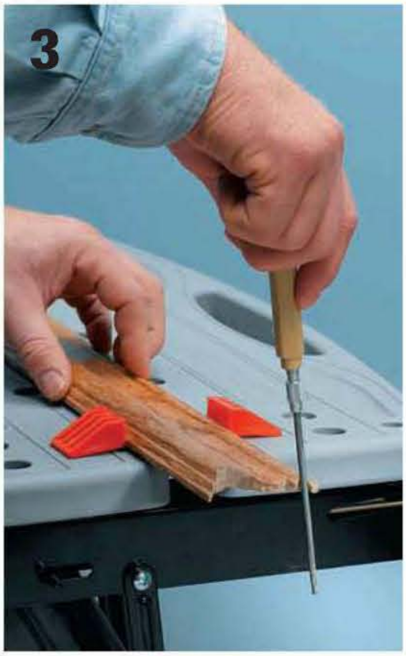

Miter-cut the second piece of molding with a power miter saw and then cope the end with a coping saw. Clean up the edge of the cope cut with a metal file to ensure a tight fit. Dry-fit the piece to check for any gaps in the joint.

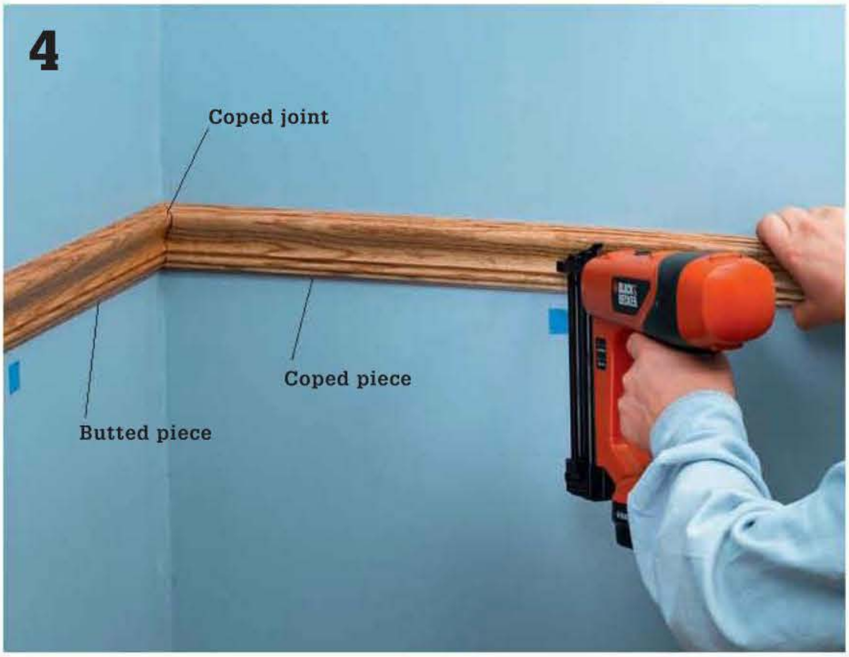

When the coped joint fits tightly, measure, mark, and cut the opposing end of the second piece of trim squarely with a miter saw. Nail the second piece in place with two nails at each stud location. Follow the level line with the bottom edge of the molding.

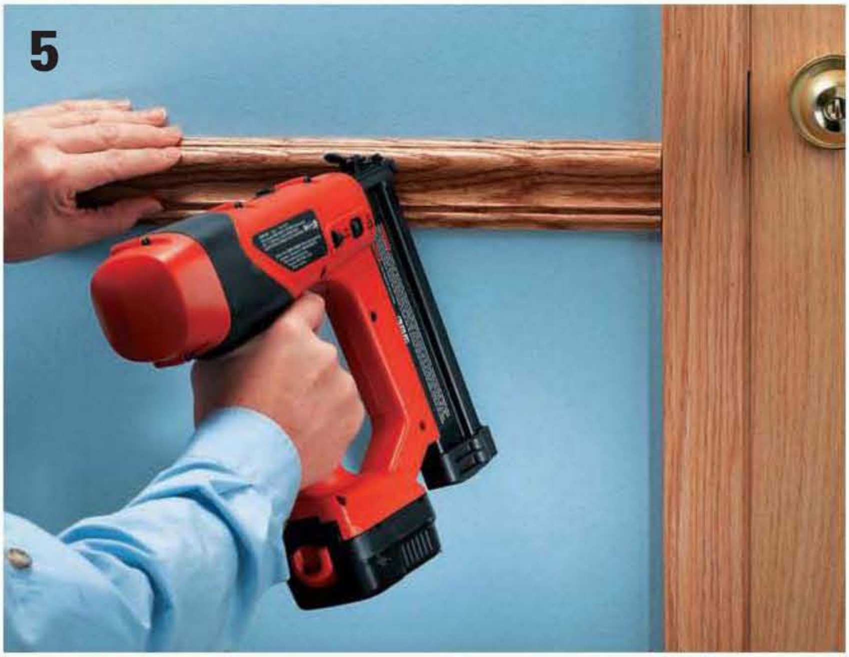

Install the third piece of chair rail with a cope cut at one end. Use a butt joint where the molding runs into door and window casings. Fill all nail holes with putty and apply a final coat of finish to the molding.

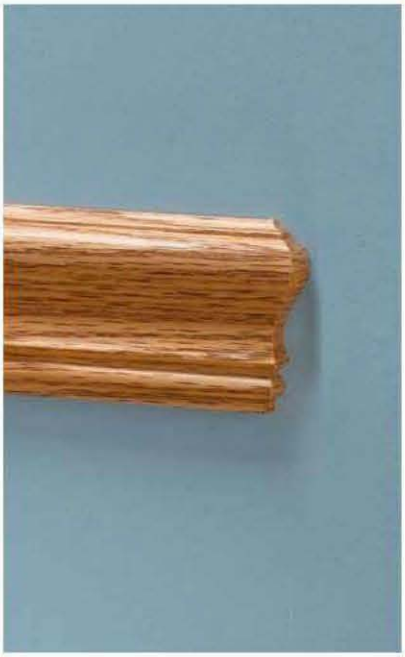

Option: Cut a mitered return for the chair rail in areas where it will end without joining into another molding. Cut the return with a miter saw and glue it in place, using painter's tape to hold it until the glue dries.

Installing Crown Molding

Simply put, crown molding is angled trim that bridges the joint between the ceiling and the wall. In order to cover this joint effectively, crown moldings are "sprung." This means that the top and bottom edges of the molding have been beveled, so when the molding is tilted away from the wall at an angle the tops and bottoms are flush on the wall and ceiling surfaces. Some crown moldings have a 45° angle at both the top and the bottom edges; another common style ("38° crown") has a 38° angle on one edge and a 52° angle on the other edge.

Installing crown molding can be a challenging and sometimes confusing process. Joints may be difficult for you to visualize before cutting, and wall and ceiling irregularities can be hard to overcome. If you have not worked on crown molding joints before, it is recommended that your first attempt be made with paint-grade materials. Stain-grade crown is commonly made of solid hardwood stock, which makes for expensive cutting errors, and difficulty concealing irregularities in joints.

Inside corner joints of crown molding should be cope cut, not mitered, except in the case of very intricate profile crown that is virtually impossible to cope (and must therefore be mitered). While mitering inside corners may appear to save time and produce adequate results, after a few changing seasons the joints will open up and be even more difficult to conceal.

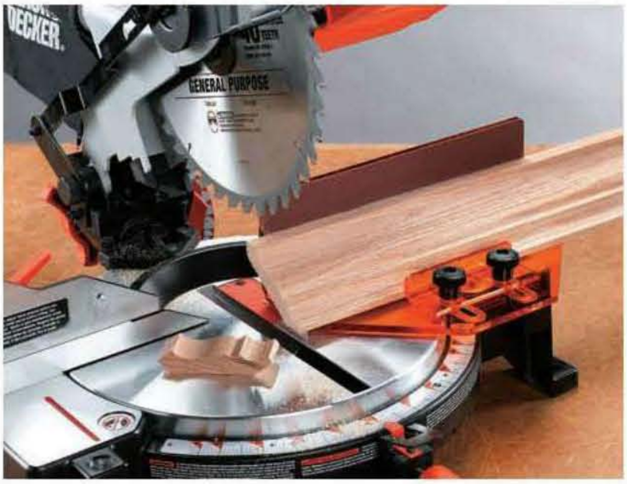

Cutting compound miters is tricky. Here, crown molding is cut with the workpiece held against a fence or fence extension. This hand-held approach is quick and effective, but takes some getting used to. A practically foolproof option is to use an adjustable jig, such as the compound miter jig shown here.

Tools & Materials ▸

- Pencil
- Tape measure
- Circular saw
- Straightedge guide
- Drill with bits
- Coping saw
- Power miter saw
- Pneumatic finish nail gun
- Framing square or combination square
- Nail set
- Hammer
- Metal files
- 2 × 4 material for backing
- 3" wallboard screws
- Carpenter's glue
- Crown molding
- 2", 1½" finish nails
- Fine-grit sandpaper
- Hole filler
- Paint and brushes

Basic crown molding softens the transitions between walls and ceilings. If it is made from quality hardwood crown molding can be quite beautiful when installed and finished with a clear topcoat. But historically, it is most often painted—either the same color as the ceiling (your eye tends to see it as a ceiling molding, not a wall molding) or with highly elaborate painted and carved details.

How to Install Basic Crown Molding

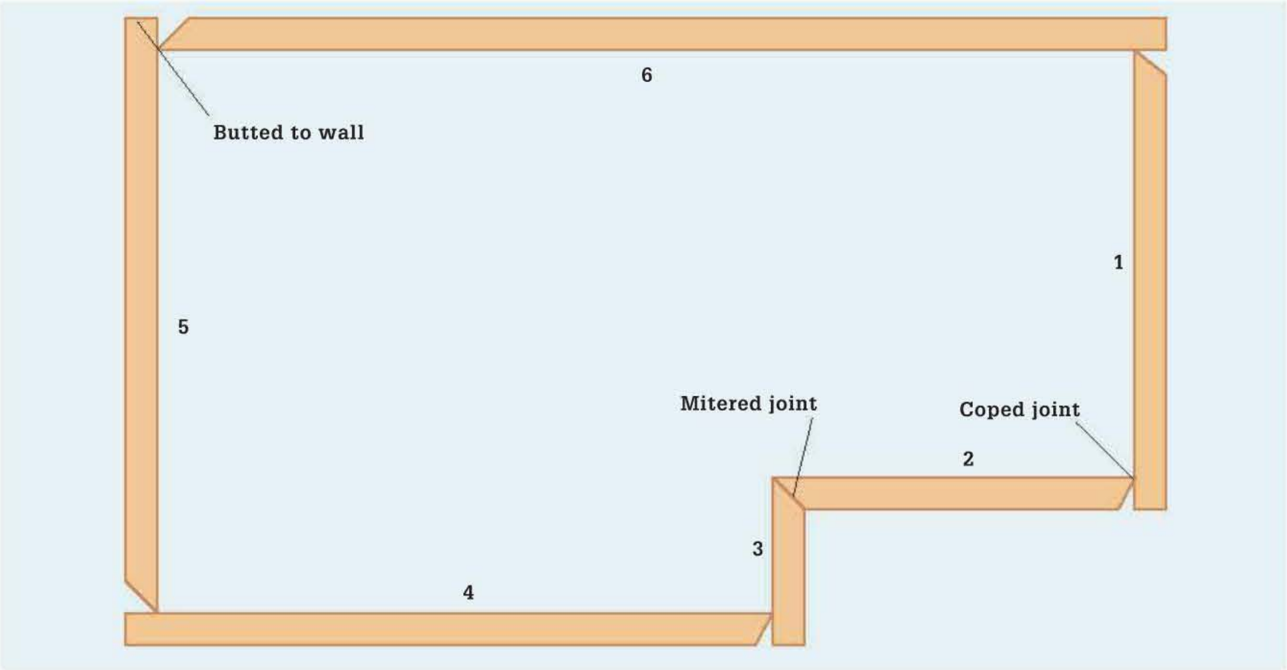

Plan the order of the installation to minimize the number of difficult joints on each piece and use the longest pieces for the most visible sections of wall. Notice that the left end of first piece is cope-cut rather than butted into the wall. Cope-cutting the first end eliminates the need to cope-cut both ends of the final piece, and places the cuts in the same direction. This simplifies your installation, making the method to cut each piece similar.

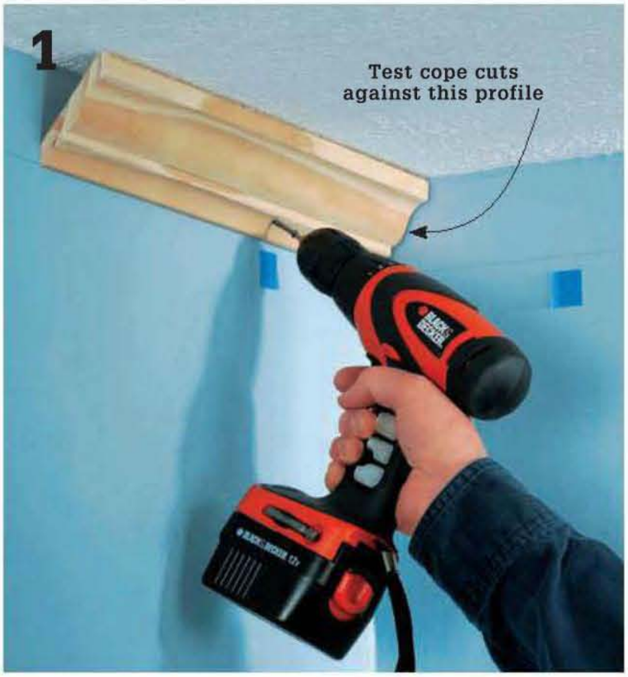

Cut a piece of crown molding about 1-ft. long with square ends. Temporarily install the piece in the corner of the last installation wall with two screws driven into the blocking. This piece serves as a template for the first cope cut on the first piece of molding.

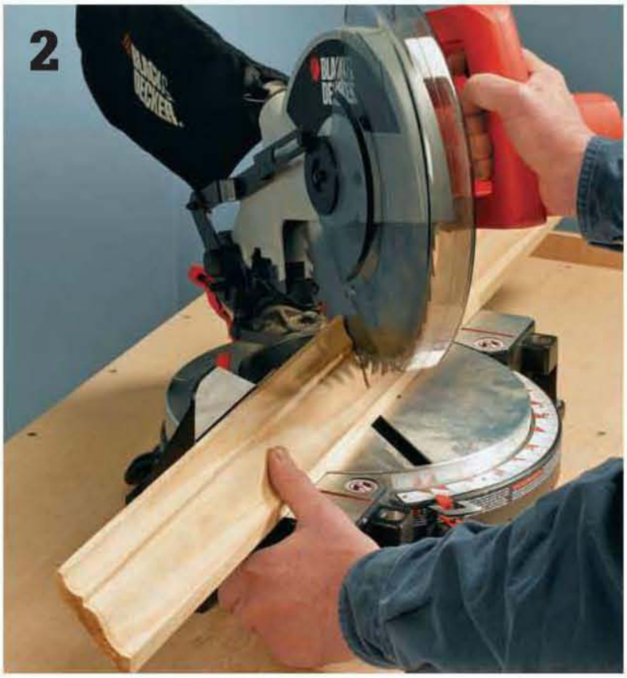

Place the first piece of molding upside down and sprung against the fence of the miter saw. Mark a reference line on the fence for placement of future moldings, and cut the first coped end with an inside miter cut to reveal the profile of the piece.

(continued)

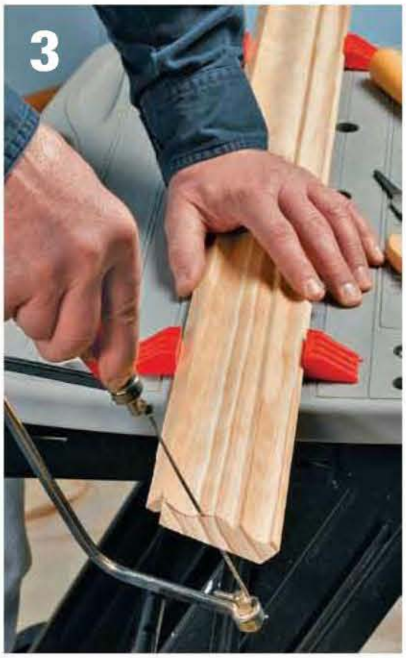

Cope-cut the end of the first piece with a coping saw. Carefully cut along the profile, angling the saw as you cut to back-bevel the cope. Test-fit the coped cut against the temporary scrap from Step 1. Fine-tune the cut with files and fine-grit sandpaper.

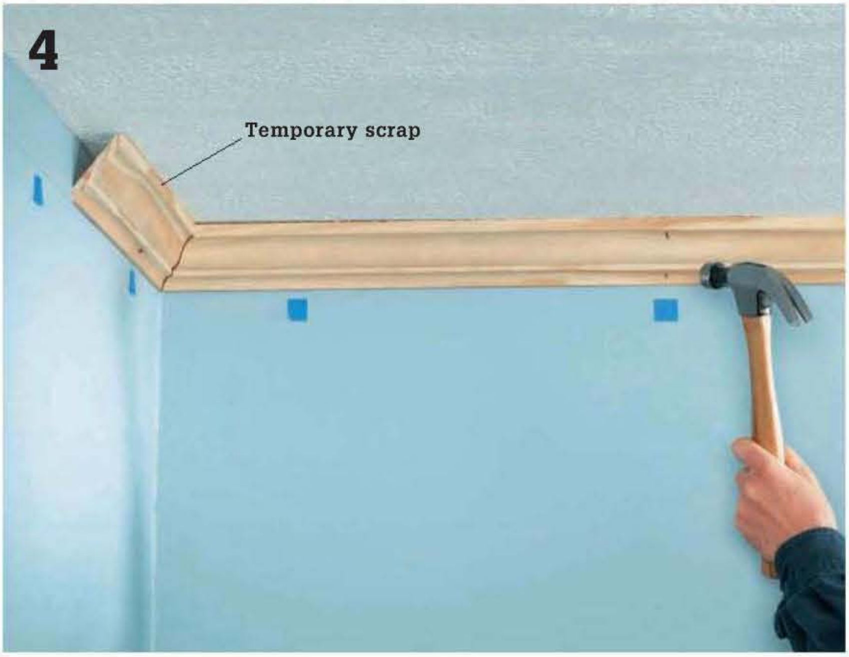

Measure, cut to length, and install the first piece of crown molding, leaving the end near the temporary scrap loose for final fitting of the last piece. Nail the molding at the top and bottom of each stud location.

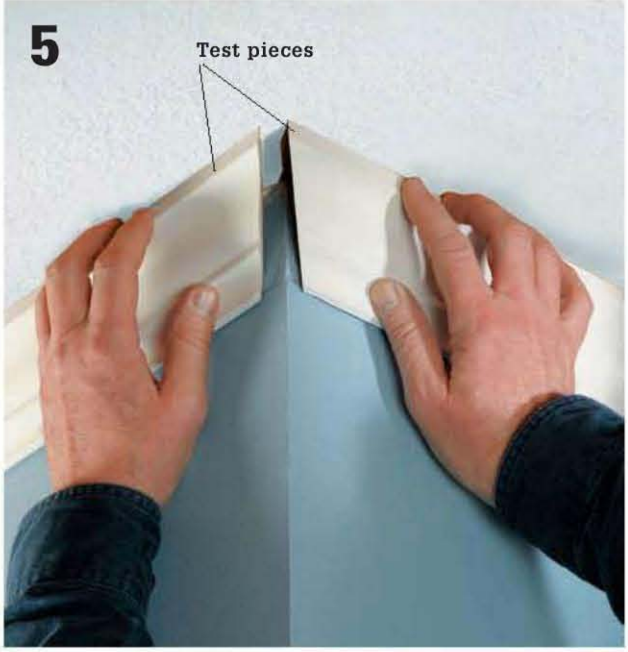

Cut two test pieces to check the fit of outside corners. Start with each molding cut at 45°, adjusting the angles larger or smaller until the joints are tight. Make sure the test moldings are properly aligned and are flush with the ceiling and walls. Make a note of your saw settings once the joint fits tightly.

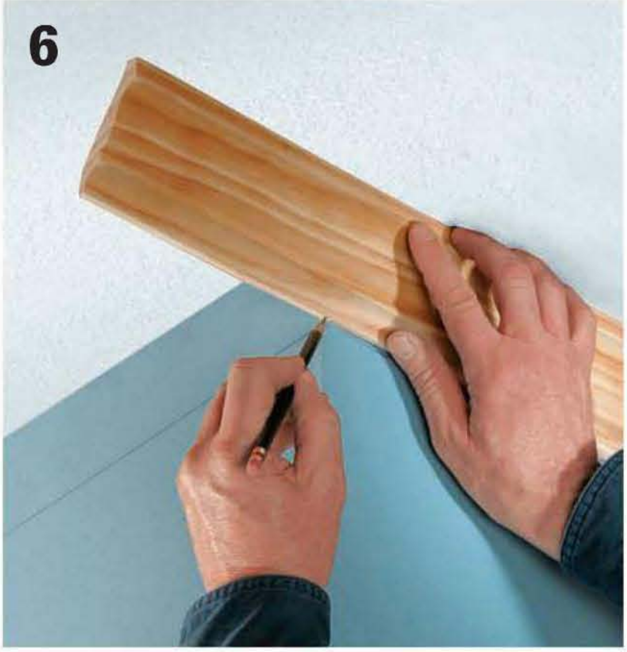

Position the actual stock so a cut end is flush against the wall at one end and, at the other end, mark the outside corner on the back edge of the molding. Miter-cut the piece at the mark, according to the angles you noted on the test pieces.

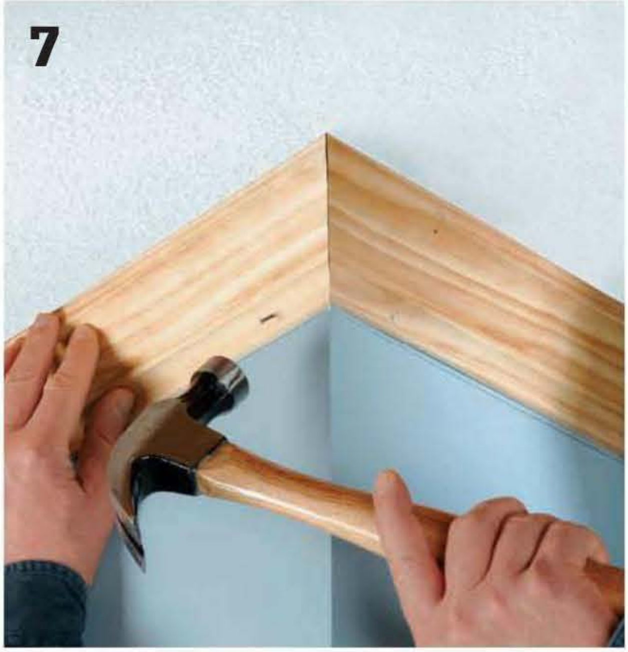

Measure and cut the third piece with an outside corner miter to match the angle of your test pieces. Cut the other end squarely, butting it into the corner. Install the piece with nails driven at stud locations. Install the subsequent pieces of crown molding, coping the front end and butting the other as you work around the room.

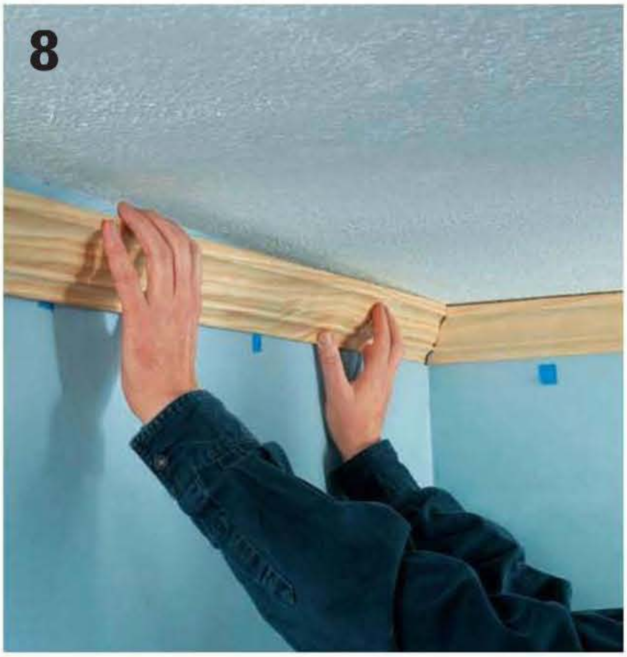

To fit the final piece, cope the end and cut it to length. Remove the temporary scrap piece from Step 3, and slide the last molding into position. Nail the last piece at the stud locations when the joints fit well, and finish nailing the first piece.

Fill all nail holes (use spackling compound if painting; wait until the finish is applied and fill with tinted putty for clear finishes). Use a putty knife to force spackling compound or tinted wood putty into loose joints and caulk gaps ⅛" or smaller between the molding and the wall or ceiling with flexible, paintable, latex caulk.

Lightly sand the filled nail holes and joint gaps using fine sandpaper. Sand the nail holes flush with the surface of the moldings and paint the entire project.

ELEVATION
1/2"=1'-0"
ELEVATION
1/2"=1'-0"
3-SEASON PORCH
DESIGN

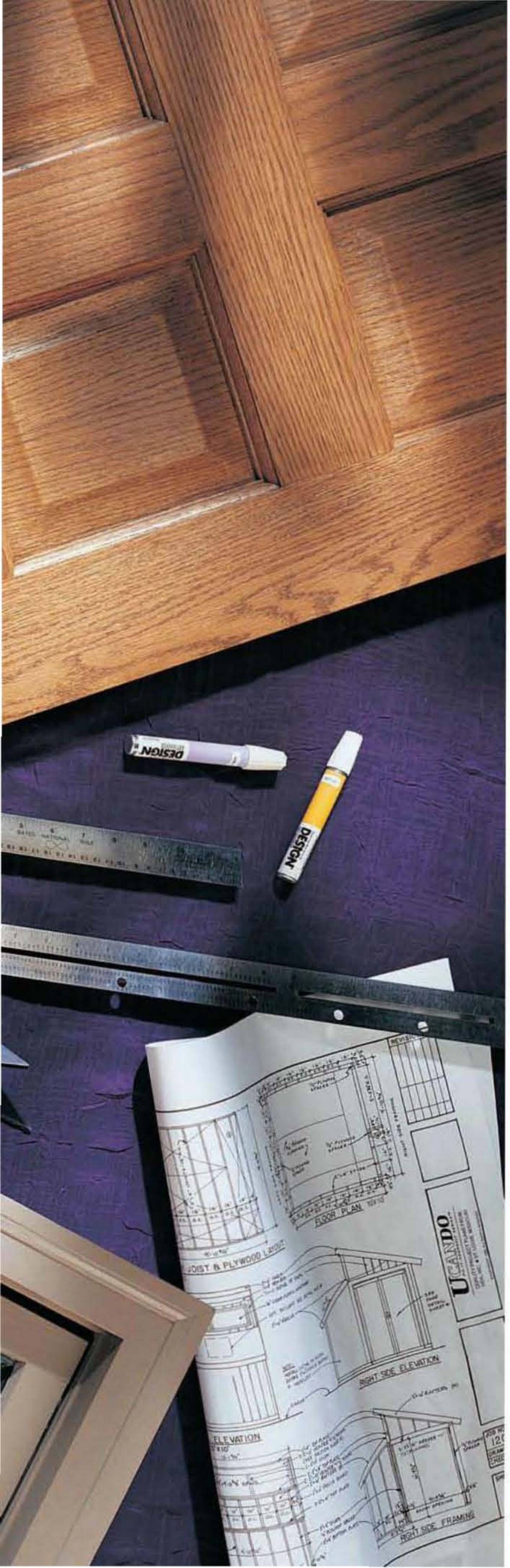

Windows & Doors

Once you've got a feel for the style of windows or doors you'd like to add to your home, the next step involves careful planning. You'll need to consider how a new door or window meets your needs for energy efficiency, durability, ease of maintenance, and security.

There's also the equally important matter of how a new window or door will impact your home's livability. Will it face a direction that takes advantage of sunlight and prevailing breezes or make the room stifling hot in mid July? Adding a window to a living space may also bring matters of egress into consideration. Building codes mandate that your project window must satisfy these requirements.

In this chapter:

- Framing & Installing Windows
- Installing Garden Windows
- Installing Bay Windows
- Installing Skylights
- Installing Tubular Skylights
- Installing Interior Doors
- Installing Entry Doors
- Installing Storm Doors
- Installing Patio Doors
- Installing Door & Window Casing

Framing & Installing Windows

Many windows must be custom-ordered several weeks in advance. To save time, you can complete the interior framing before the window unit arrives, but be sure you have the exact dimensions of the window unit before building the frame. Do not remove the outside wall surface until you have the window and accessories and are ready to install them.

Follow the manufacturer's specifications for rough opening size when framing for a window. The listed opening is usually 1" wider and ½" taller than the actual dimensions of the window unit. The following pages show techniques for wood-frame houses with platform framing.

If your house has balloon framing where wall studs pass continuously from one floor to the next, use the method shown on page 271 to install a header. Consult a professional to install a window on the second story of a balloon-framed house.

If your home's exterior has siding or is stucco, see pages 34 to 35 for tips on removing these surfaces and making the opening.

Tools & Materials ▸

- Tape measure
- Pencil
- Combination square
- Hammer
- Level
- Circular saw
- Handsaw
- Pry bar
- Nippers
- Drill
- Reciprocating saw
- Stapler
- Nail set
- Caulk gun
- 10d common nails,
- 1" galvanized roofing nails
- Shims
- 2× lumber
- ⅛" plywood
- Building paper
- Drip edge
- 10d galvanized casing nails
- 8d casing nails
- Fiberglass insulation
- Paintable silicone caulk

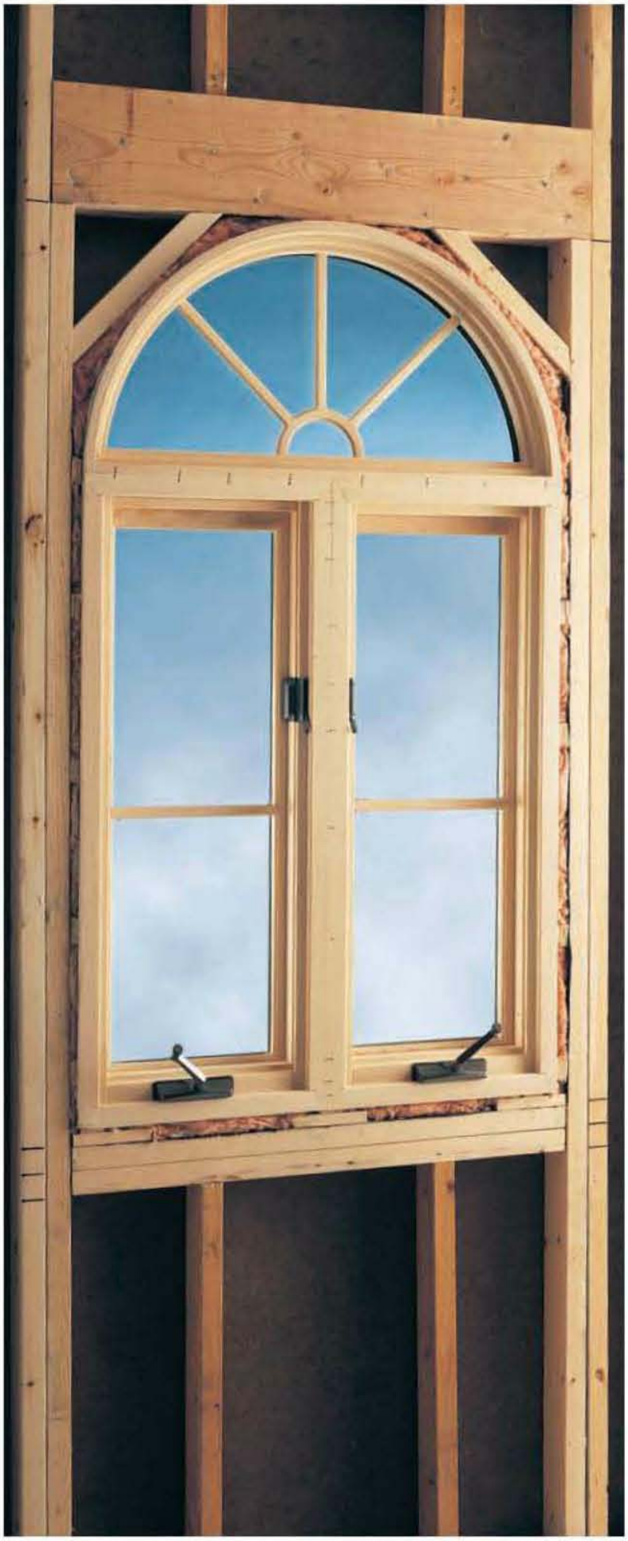

Correct framing techniques will ensure ease of installation and keep your windows operating smoothly.

How to Frame a Window Opening

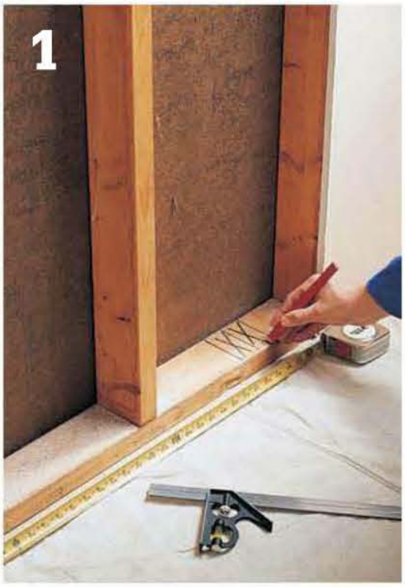

Prepare the project site and remove the interior wall surfaces (pages 30 to 31). Measure and mark the rough opening width on the sole plate. Mark the locations of the jack studs and king studs on the sole plate. Where practical, use the existing studs as king studs.

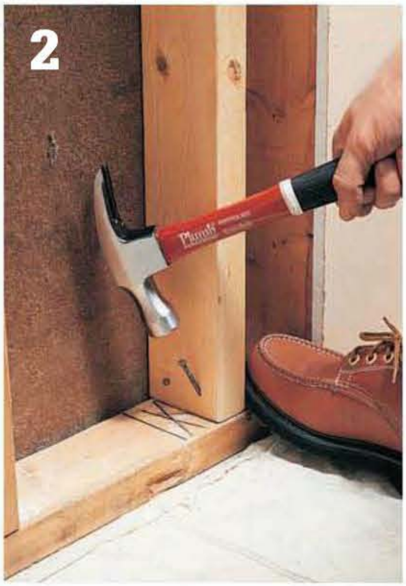

Measure and cut the king studs, as needed, to fit between the sole plate and the top plate. Position the king studs and toenail them to the sole plate with 10d nails.

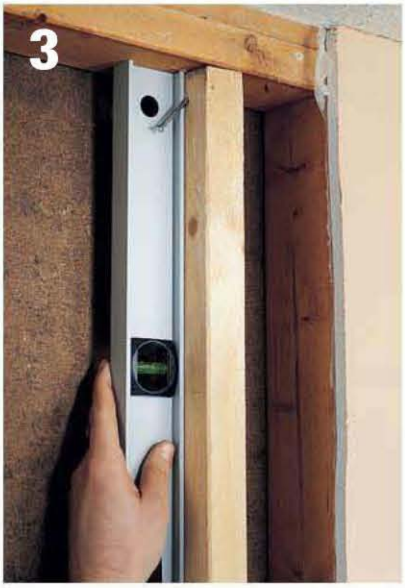

Check the king studs with a level to make sure they are plumb, then toenail them to the top plate with 10d nails.

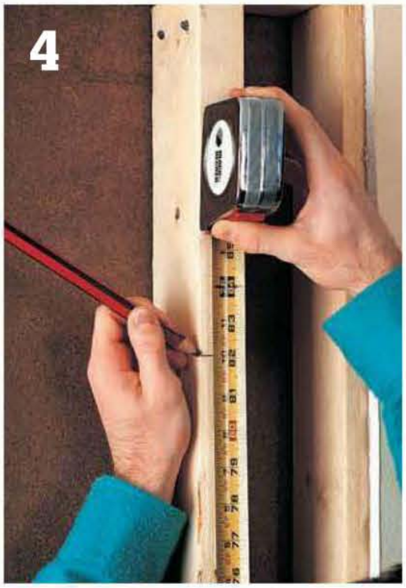

Measuring from the floor, mark the top of the rough opening on one of the king studs. This line represents the bottom of the window header. For most windows, the recommended rough opening is ½" taller than the height of the window frame.

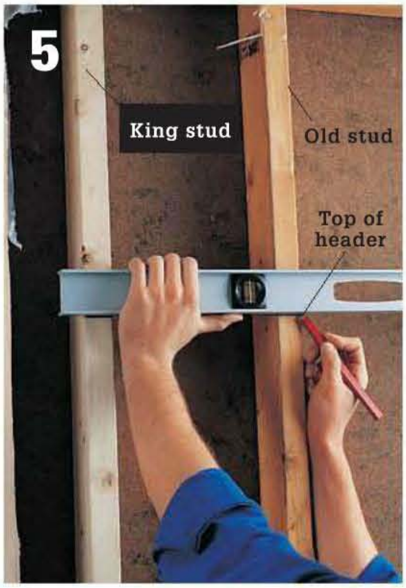

Measure and mark where the top of the window header will fit against the king studs. The header size depends on the distance between the king studs. Use a carpenter's level to extend the lines across the old studs to the opposite king stud.

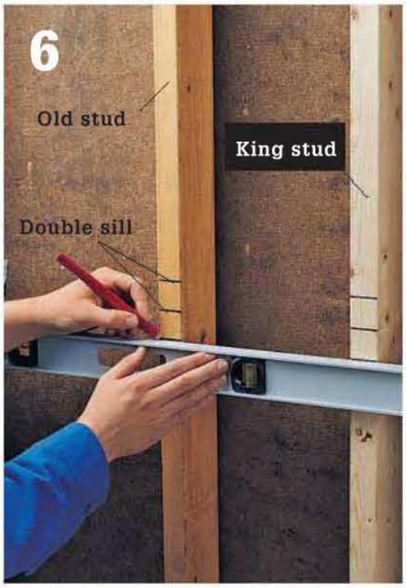

Measure down from the header line and mark the double rough sill on the king stud. Use a carpenter's level to extend the lines across the old studs to the opposite king stud. Make temporary supports (page 29) if removing more than one stud.

(continued)

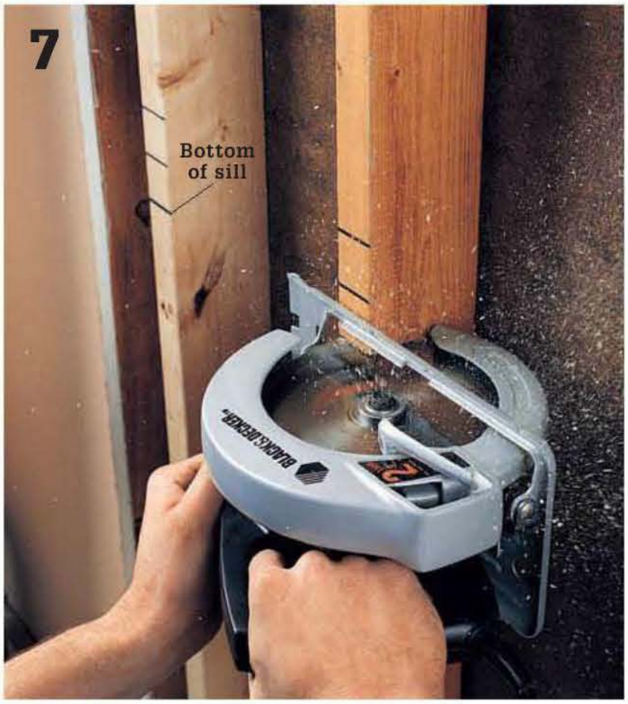
7
Bottom of sill

Set a circular saw to its maximum blade depth, then cut through the old studs along the lines marking the bottom of the rough sill and along the lines marking the top of the header. Do not cut the king studs. On each stud, make an additional cut about 3" above the sill cut. Finish the cuts with a handsaw.

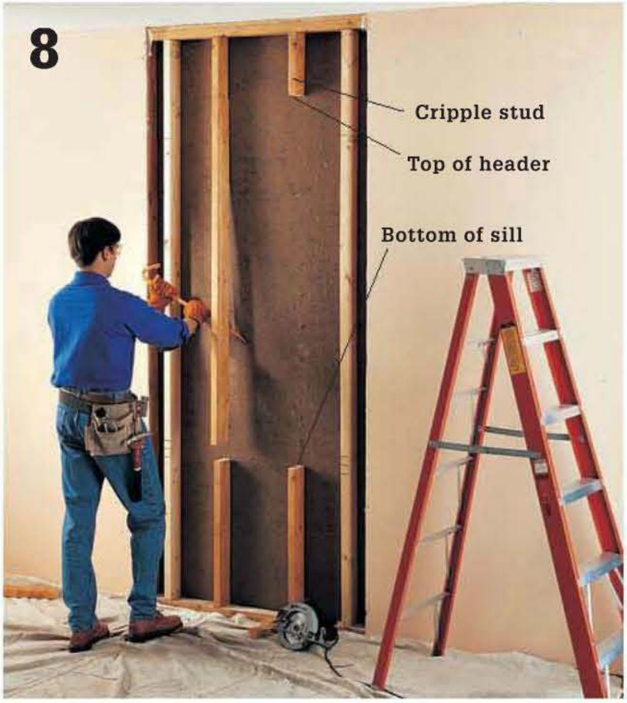
8
Cripple stud
Top of header
Bottom of sill

Knock out the 3" stud sections, then tear out the old studs inside the rough opening, using a pry bar. Clip away any exposed nails, using nippers. The remaining sections of the cut studs will serve as cripple studs for the window.

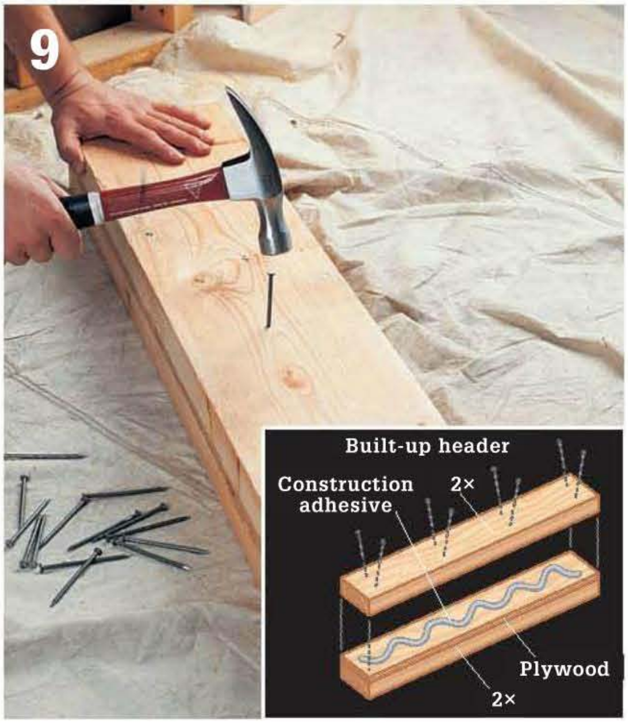
9
Built-up header
Construction adhesive
2×
Plywood
2×

Build a header to fit between the king studs on top of the jack studs, using two pieces of 2× lumber sandwiched around ½" plywood.

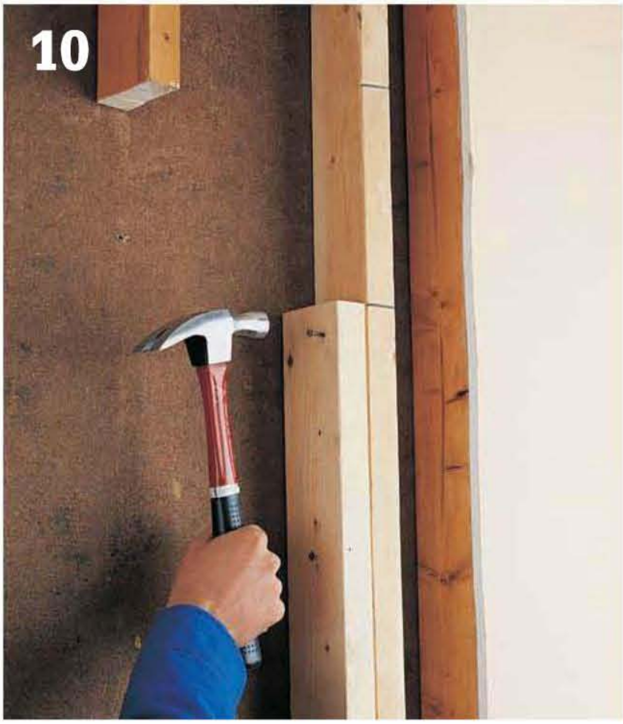

Cut two jack studs to reach from the top of the sole plate to the bottom header lines on the king studs. Nail the jack studs to the king studs with 10d nails driven every 12".

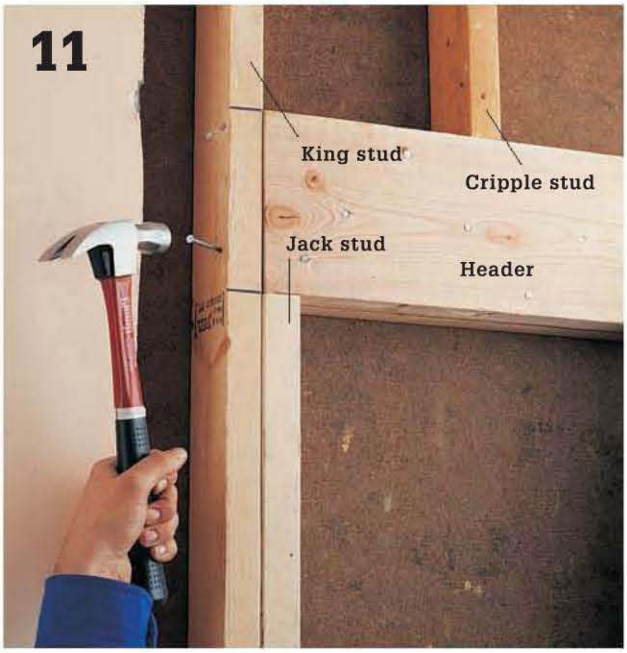

Position the header on the jack studs, using a hammer if necessary. Attach the header to the king studs, jack studs, and cripple studs, using 10d nails.

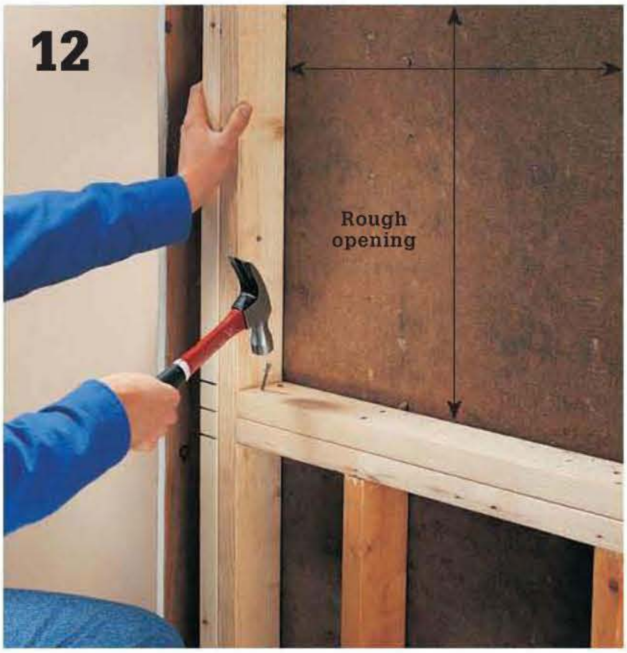

Build the rough sill to reach between the jack studs by nailing a pair of 2 × 4s together. Position the rough sill on the cripple studs, and nail it to the jack studs and cripple studs with 10d nails.

How to Install a Replacement Window with a Nailing Flange

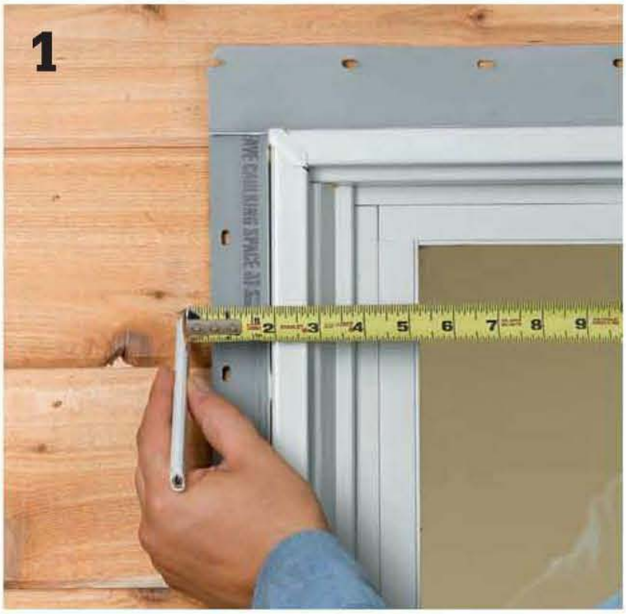

Remove the existing window and set the new window into the rough opening. Center it left to right, and shim beneath the sill to level it. On the exterior side, measure out from the window on all sides, and mark the siding for the width of the brick molding you'll install around the new window. Extend layout lines to mark where you'll cut the siding.

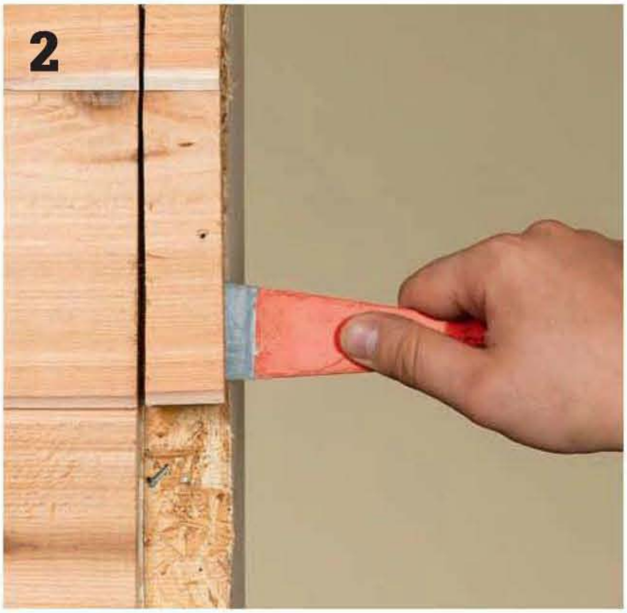

Remove exterior siding around the window area to expose the wall sheathing. Use a zip tool to separate vinyl siding for removal or use a pry bar and hammer to remove wood clapboard. For more on removing exterior surfaces, (see pages 32 to 33).

(continued)

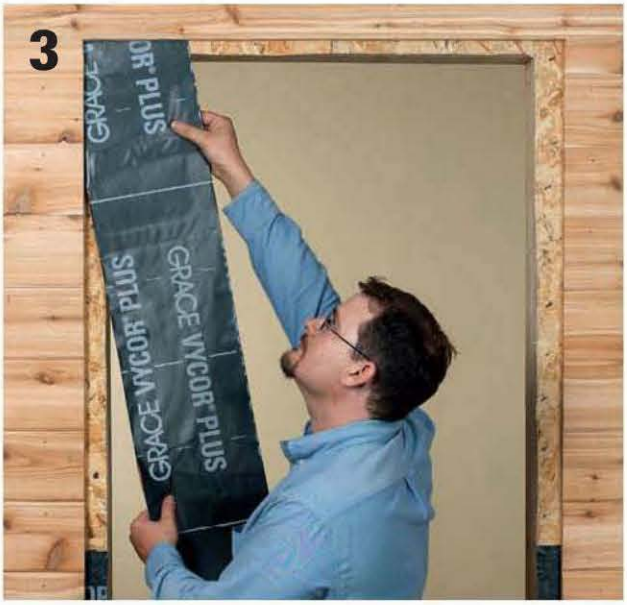

Cover the sill and rough opening framing members with self-adhesive, rolled flashing. Apply additional strips of flashing behind the siding and up the sill flashing. Finish flashing with a strip along the header. The flashing should cover the front edges and sides of the opening members.

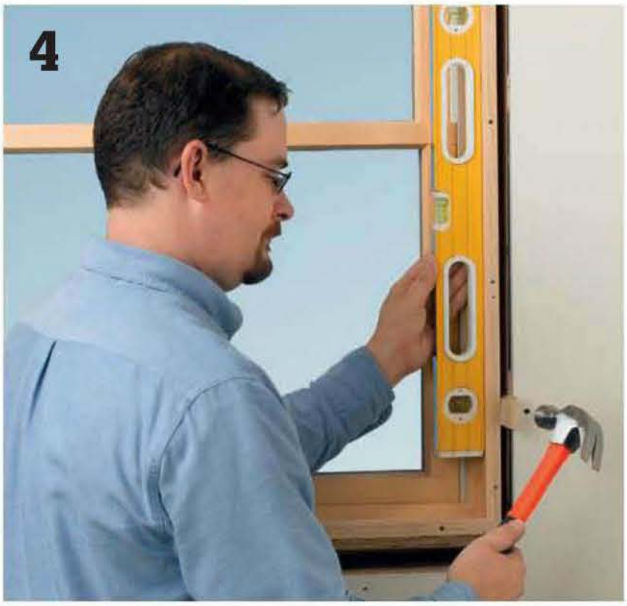

Apply a bead of silicone caulk around the back face of the window flange, then set it into the rough opening, centering it side-to-side in the opening. Tack the window in place by driving one roofing nail partway through the top flange. On the interior side, level and plumb the window, using shims to make any necessary adjustments.

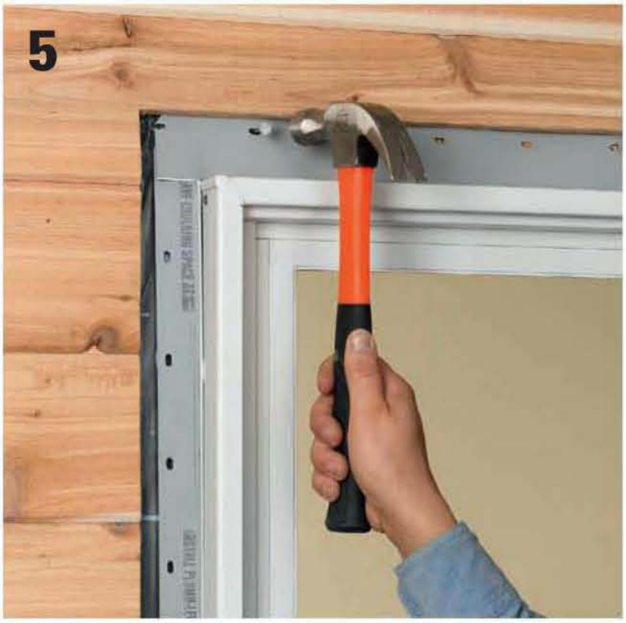

Tack the window to the header at one end of the nailing flange, using a 1" galvanized roofing nail. Drive a roofing nail through the other top corner of the flange to hold the window in place, then secure the flange all around the window with more roofing nails. Apply strips of rolled, self-adhesive flashing to cover the window flanges. Start with a strip that covers the bottom flange, then cover the side flanges, overlapping the bottom flashing and extending 8 to 10" above the window. Complete the flashing with a strip along the top, overlapping the side flashing.

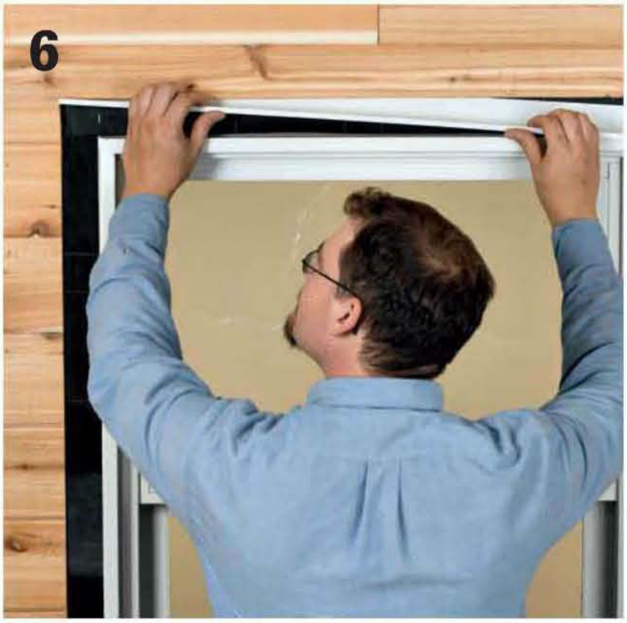

Install a piece of metal drip edge behind the siding and above the window. Secure it with silicone caulk only.

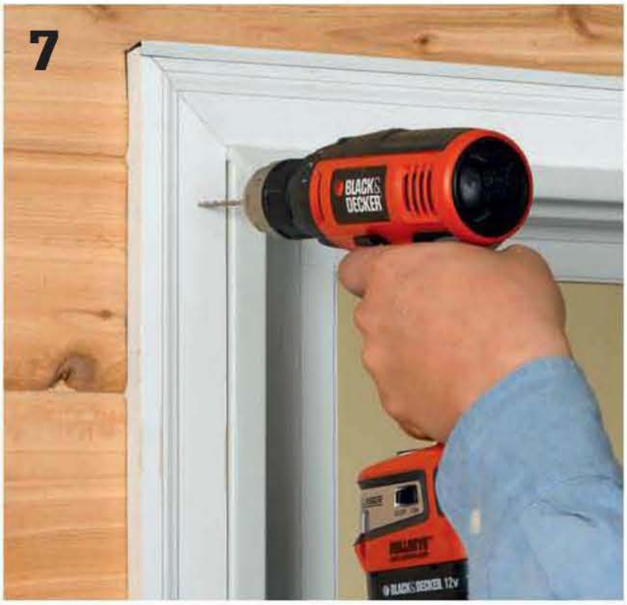

Cut and attach brick molding around the window, leaving a slight gap between the brick molding and the window frame. Use 8d galvanized casing nails driven into pilot holes to secure the brick molding to the rough framing. Miter the corner joints. Reinstall the siding in the window installation area, trimming as needed.

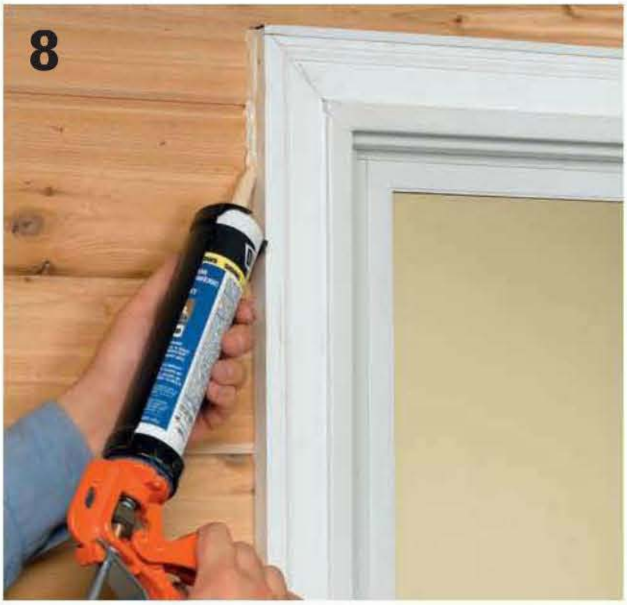

Use high-quality caulk to fill the gap between the brick molding and the siding. On the interior side, fill gaps between the window frame and surrounding framing with foam backer rod, low-expansion foam, or fiberglass insulation. Install the interior casing.

Tip Installation Variation: Masonry Clips ▸

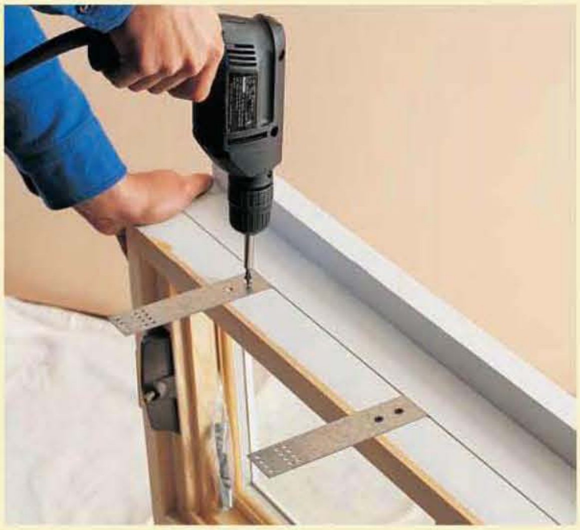

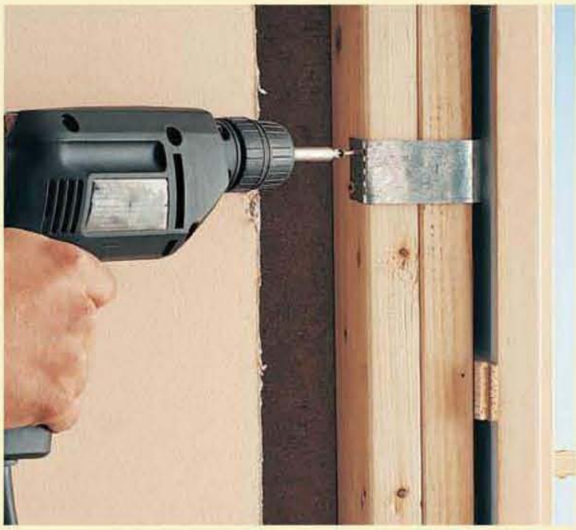

Use metal masonry clips when the brick molding on a window cannot be nailed because it rests against a masonry or brick surface. The masonry clips hook into precut grooves in the window jambs (above, left) and are attached to the jambs with screws. After the window unit is positioned in the rough opening, the masonry clips are bent around the framing members and anchored with screws (above, right). *Note: Masonry clips can also be used in ordinary lap siding installations if you want to avoid making nail holes in the smooth surface of the brick moldings. For example, windows that are precoated with polymer-based paint can be installed with masonry clips so that the brick moldings are not punctured with nails.*

How to Install a Round-Top Window

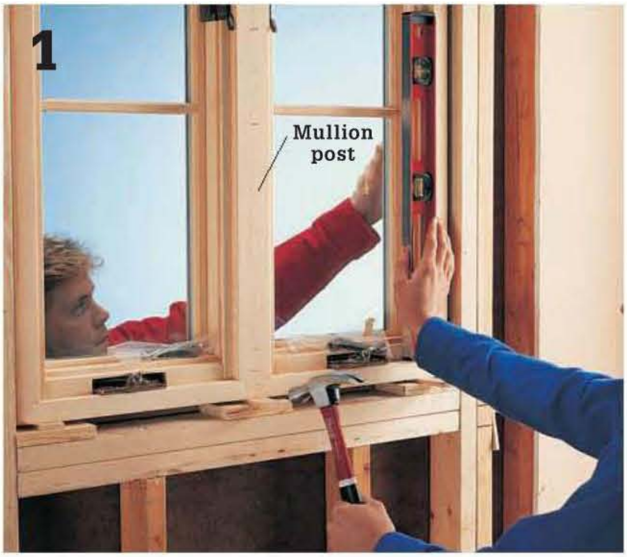

Remove the exterior wall surface as directed on pages 32 to 33, then test-fit the window, centering it in the rough opening. Support the window with wood blocks and shims placed under the side jambs and mullion post. Check to make sure the window is plumb and level, and adjust the shims, if necessary.

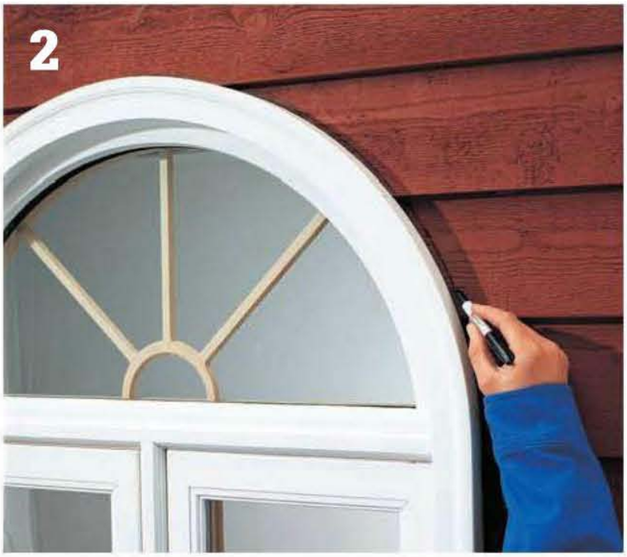

Trace the outline of the brick molding on the wood siding. Remove the window after finishing the outline. *Note: If you have vinyl or metal siding, you should have enlarged the outline to make room for the extra J-channel moldings required by these sidings.*

Tips for Framing a Round-top Window ▸

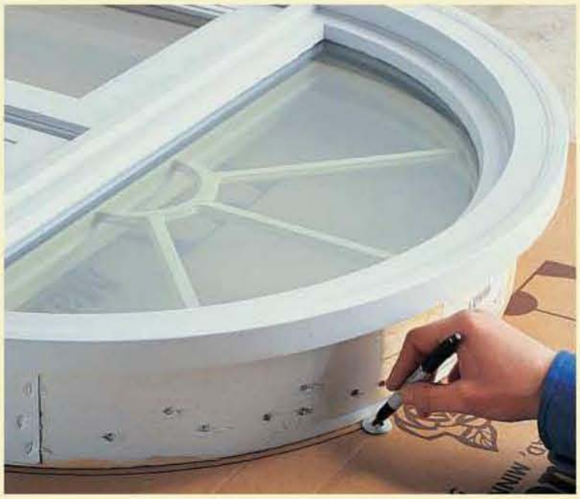

Create a template to help you mark the rough opening on the sheathing. Scribe the outline of the curved frame on cardboard, allowing an extra ½" for adjustments within the rough opening. A ¼ × 1¼" metal washer makes a good spacer for scribing the outline. Cut out the template along the scribed line.

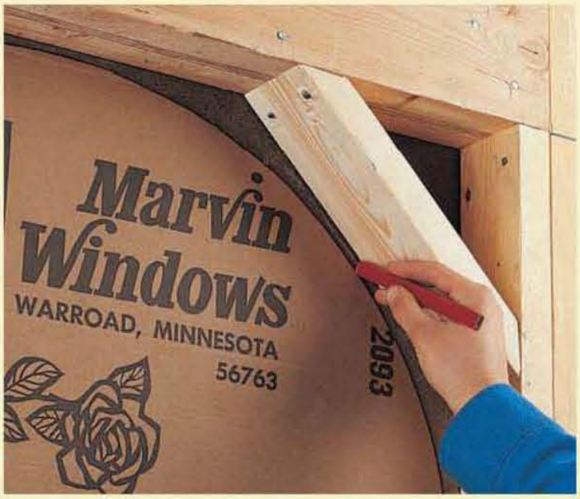

Tape the template to the sheathing, with the top flush against the header. Use the template as a guide for attaching diagonal framing members across the top corners of the framed opening. The diagonal members should just touch the template. Outline the template on the sheathing as a guide for cutting the exterior wall surface.

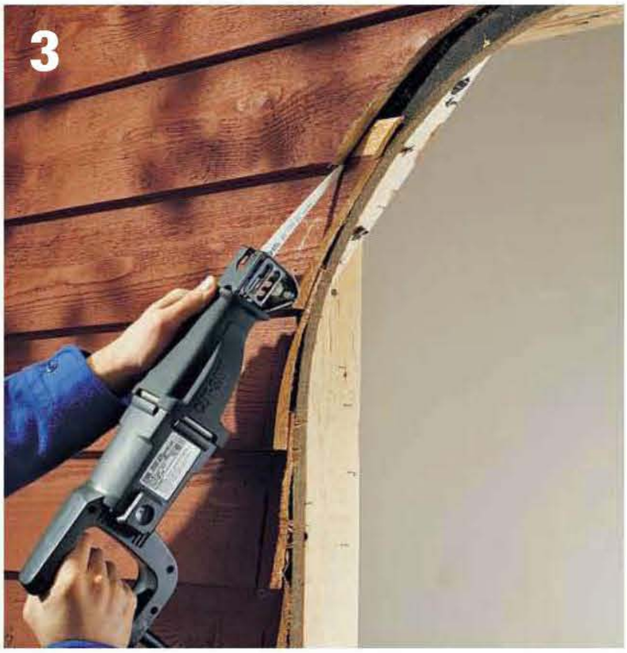

Cut the siding along the outline just down to the sheathing. For a round-top window, use a reciprocating saw held at a low angle. For straight cuts, use a circular saw adjusted so the blade cuts through only the siding. Use a sharp chisel to complete the cuts at the corners.

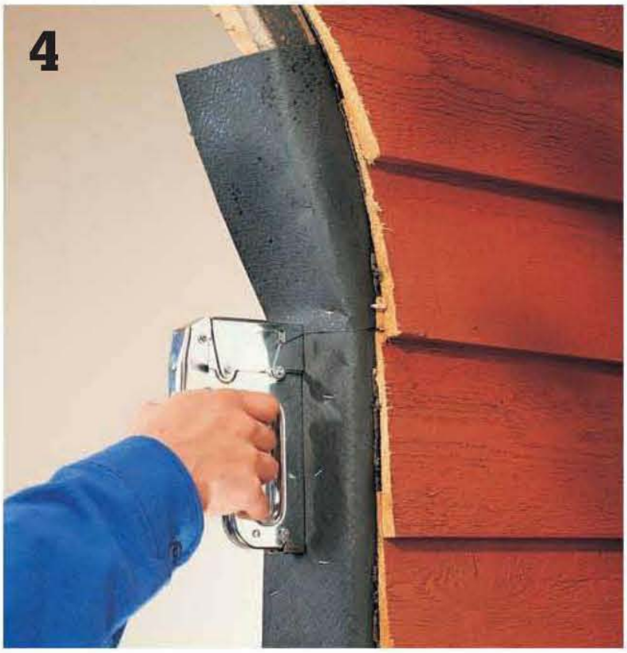

Cut 8"-wide strips of building paper and slide them between the siding and sheathing around the entire window opening. Bend the paper around the framing members and staple it in place. Work from the bottom up, so each piece overlaps the piece below. *Note: You can also use adhesive-backed, rolled flashing instead of building paper.*

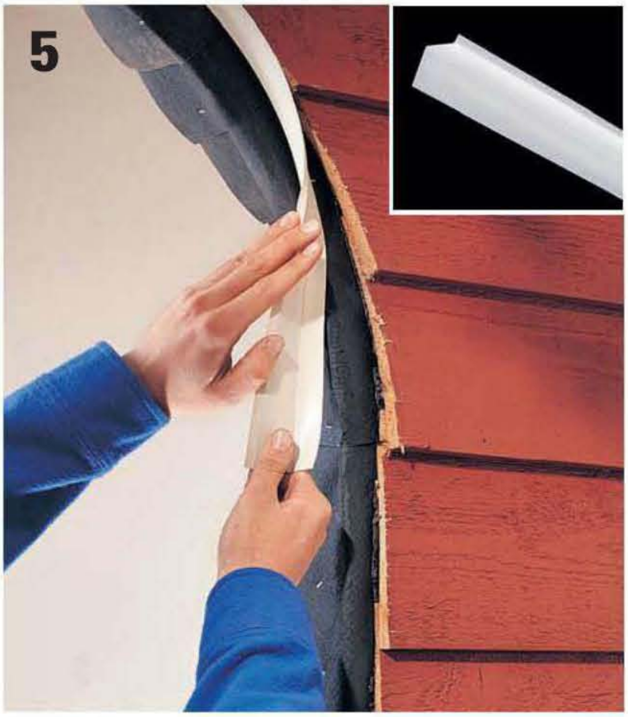

Cut a length of drip edge to fit over the top of the window, then slide it between the siding and building paper. For round-top windows, use flexible vinyl drip edge; for rectangular windows, use rigid metal drip edge (inset).

Insert the window in the opening, and push the brick molding tight against the sheathing. Nail through the brick molding, as usual, to secure the window in the opening.

Installing Garden Windows

Although often found in kitchens, a garden window is an attractive option for nearly any room in your home. Projecting out from the wall 16 to 24", garden windows add space to a room, making it feel larger. The glass roof and box-like design make them ideal growing environments for plants or display areas for collectibles. Garden windows also typically include front- or side-opening windows. These allow for ventilation and are usually available in either awning or casement style.

Home stores often stock garden windows in several common sizes. However, it may be difficult to locate a stock window that will fit in your existing window rough opening. In cases like this you must rebuild the rough opening to the proper size. It may be worth the added expense to custom-order your garden window to fit into the existing rough opening.

The large amount of glass in a garden window has a direct effect on the window's energy efficiency. When purchasing a garden window, as a minimum, look for double-pane glass with low-emissivity (low-E) coatings. More expensive super-efficient types of glass are available for severely cold climates.

Installation methods for garden windows vary by manufacturer. Some units include a nailing flange that attaches to the framing and holds the window against the house. Other models hang on a separate mounting frame that attaches to the outside of the house. In this project, the garden window has a built-in mounting sleeve that slides into the rough opening and is attached directly to the rough framing.

Tools & Materials ▸

- Tape measure
- Hammer
- Level
- Framing square
- Circular saw
- Wood chisel
- Stapler
- Drill and bits
- Caulking gun
- Utility knife
- Garden window kit
- Wood strips
- 2 × 4s
- Shims
- Exterior trim
- Building paper
- 3" screws
- Drip edge
- Construction adhesive
- 4d siding nails
- 8d galvanized casing nails
- Interior trim
- Paintable silicone caulk

A garden window's glass roof makes it an ideal sun spot for houseplants. It can also help a room feel larger.

How to Install a Garden Window

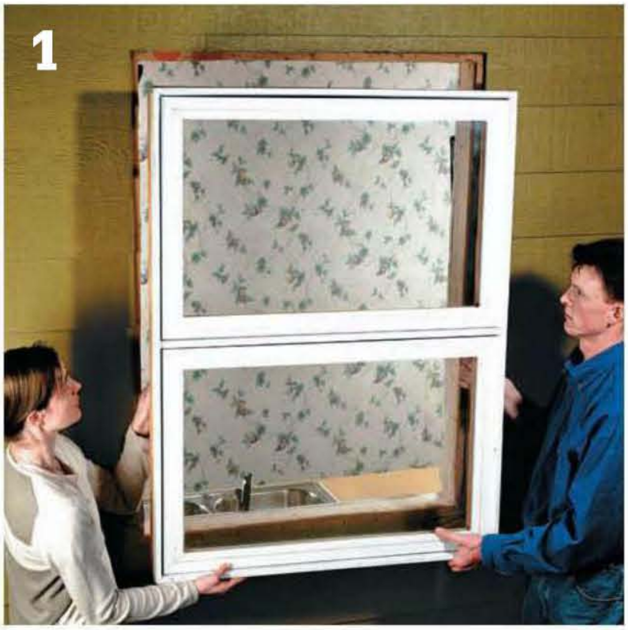

Prepare the project site and remove the interior and exterior trim, then remove the existing window.

Check the rough opening measurements to verify the correct window sizing. The rough opening should be about ½" larger than the window height and width. If necessary, attach wood strips to the rough framing as spacers to bring the opening to the required size.

Use a level to check that the sill of the rough opening is level and the side jambs are plumb. Use a framing square to make sure each corner is square. The rough framing must be in good condition in order to support the weight of the garden window. If the framing is severely deteriorated or out of plumb or square, you may need to reframe the rough opening (pages 40 to 43).

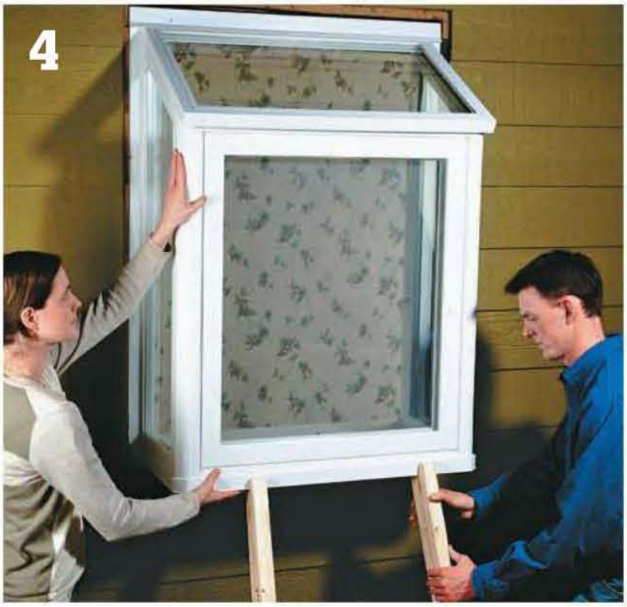

Insert the garden window into the opening, pressing it tight against the framing. Support the unit with notched 2 × 4s under the bottom edge of the window until it has been fastened securely to the framing.

(continued)

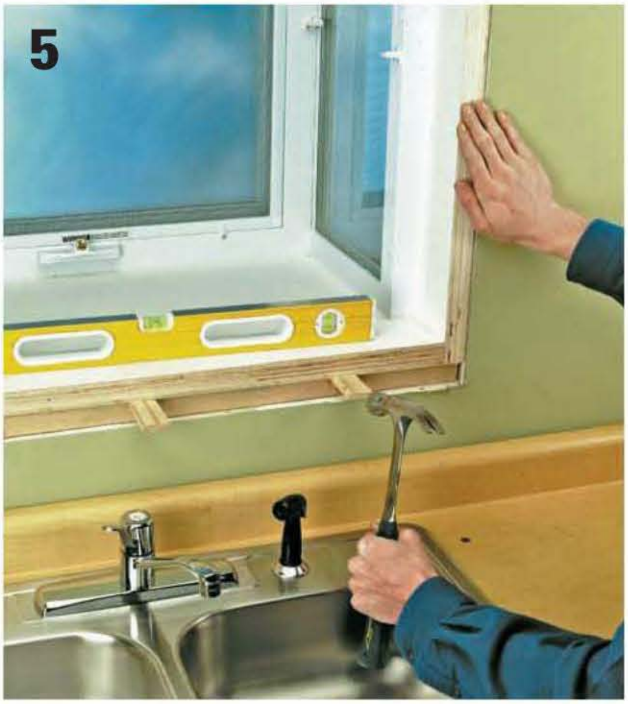

The inside edge of the window sleeve should be flush with the interior wall surface. Check the sill of the garden window for level. Shim beneath the lower side of the sill, if necessary, to make it level.

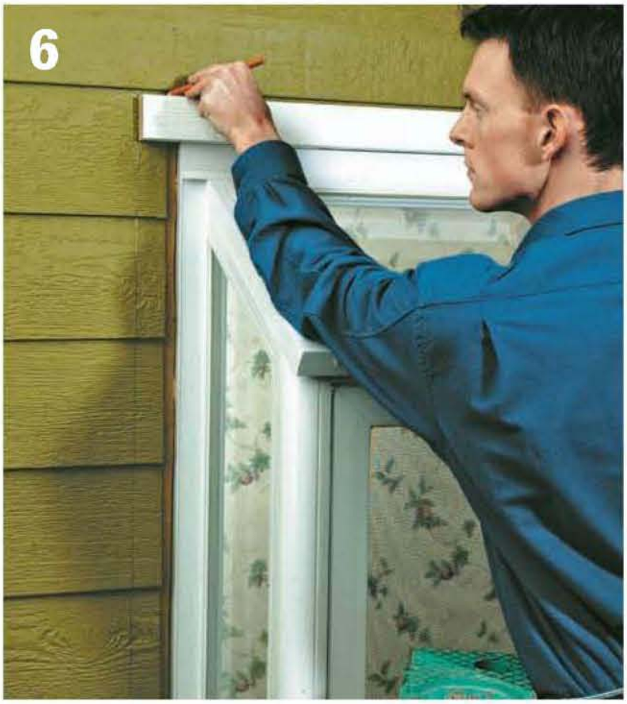

Once the garden window is in place and level, hold a piece of window trim in place along the exterior of the window and trace the outline onto the siding. Remove the window. Cut the siding down to the sheathing using a circular saw.

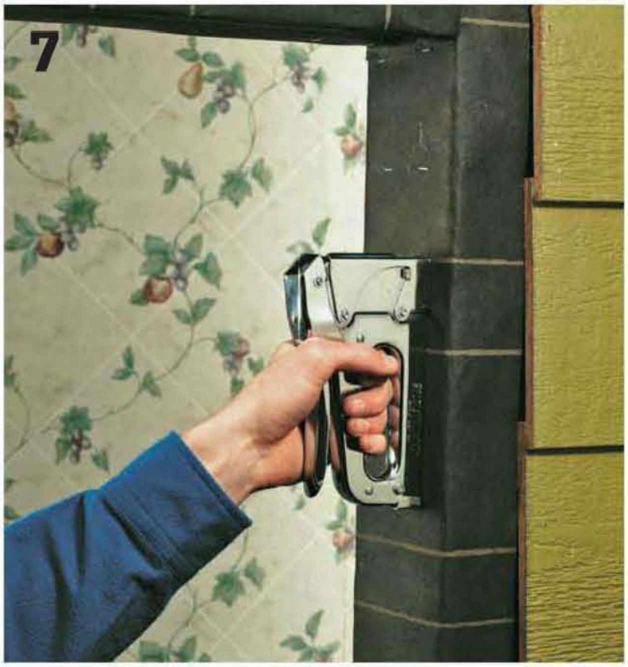

Install strips of building paper between siding and sheathing. Wrap them around the framing and staple them in place. On the sides, work from the bottom up so each piece overlaps the piece below. Reposition the window and reshim. Make sure the space between the window and the siding is equal to the width of the trim on all sides.

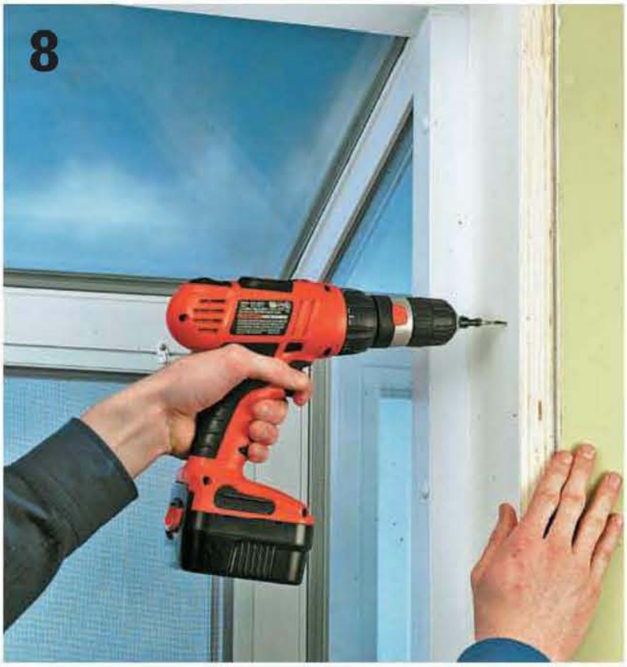

Drill countersunk pilot holes every 12" to 16" through the window sleeve into the rough header, jack studs, and sill.

9

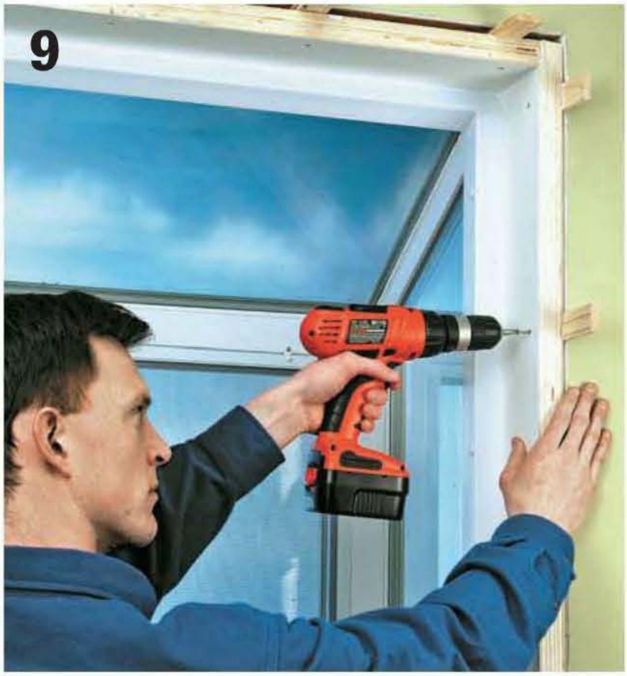

Insert shims between the window sleeve and rough frame at each hole location along the top and sides to prevent bowing of the window frame. Fasten the window to the framing using 3" screws. Continue checking for level, plumb, and square as the screws are tightened.

10

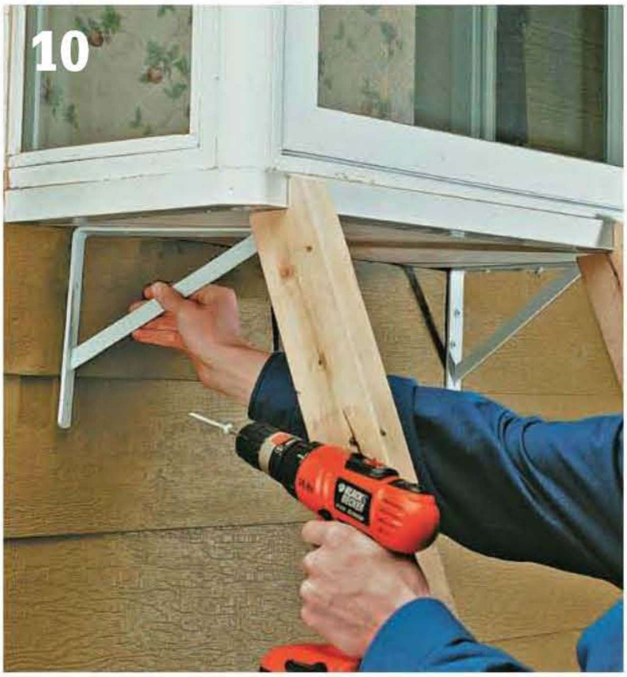

Locate and mark the studs nearest the edges of the window using a stud finder. Cut two pieces of siding to fit behind the brackets, and tack them in place over the marked studs with 4d siding nails. Position the support brackets with the shorter side against the siding and the longer side beneath the window. Fasten the brackets to the window and the studs using the included screws.

11

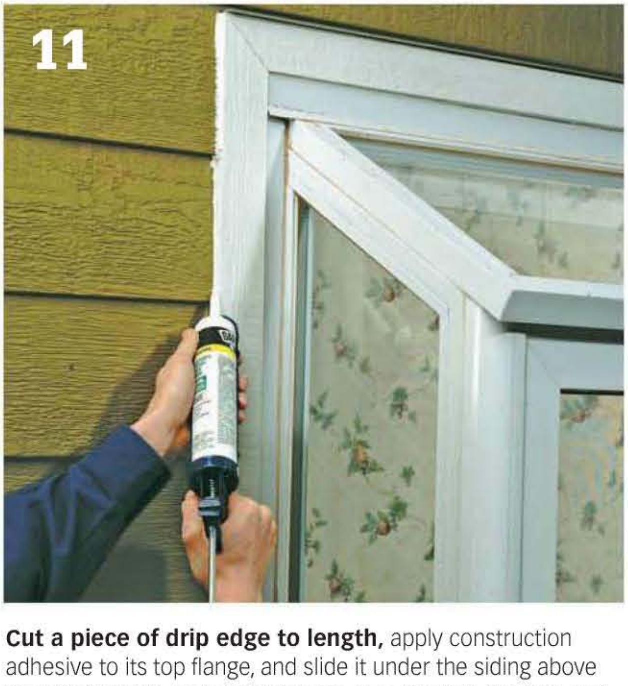

Cut a piece of drip edge to length, apply construction adhesive to its top flange, and slide it under the siding above the window. Cut each trim piece to size. Position the trim and attach it using 8d galvanized casing nails driven through pilot holes. Seal the edges of the trim with a bead of paintable silicone caulk, approximately ⅜" wide.

12

Cut all protruding shims flush with the framing using a utility knife or handsaw. Insulate or caulk gaps between the window sleeve and the wall. Finish the installation by reinstalling the existing interior trim or installing new trim.

Installing Bay Windows

Modern bay windows are preassembled for easy installation, but it will still take several days to complete an installation. Bay windows are large and heavy, and installing them requires special techniques. Have at least one helper to assist you, and try to schedule the work when there's little chance of rain. Using prebuilt bay window accessories will speed your work (see next page).

A large bay window can weigh several hundred pounds, so it must be anchored securely to framing members in the wall and supported by braces attached to framing members below the window. Some window manufacturers include cable-support hardware that can be used instead of metal support braces.

Before purchasing a bay window unit, check with the local building department regarding the code requirements. Many local codes require large windows and low bay windows with window seats to be glazed with tempered glass for safety.

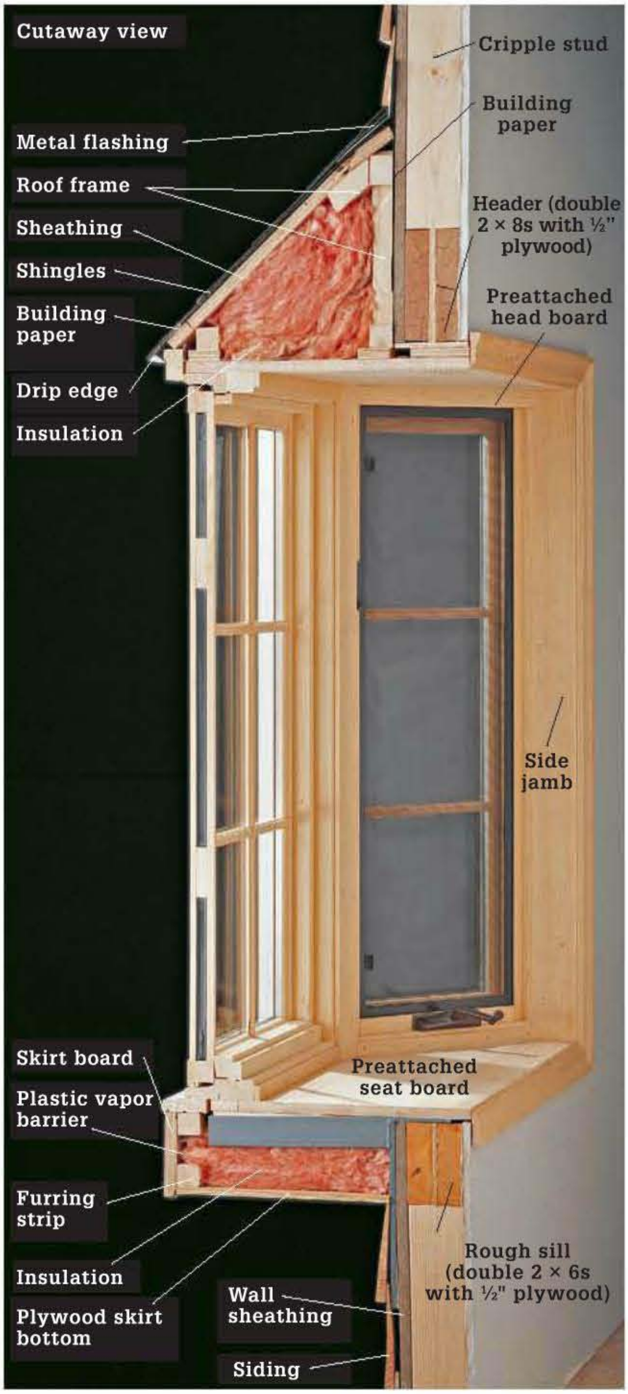
Cutaway view
Metal flashing
Roof frame
Sheathing
Shingles
Building paper
Drip edge
Insulation
Cripple stud
Building paper
Header (double 2 × 8s with ½" plywood)
Preattached head board
Side jamb
Skirt board
Plastic vapor barrier
Furring strip
Insulation
Plywood skirt bottom
Wall sheathing
Siding
Preattached seat board
Rough sill (double 2 × 6s with ½" plywood)

Tools & Materials ▸

Straightedge
Circular saw
Wood chisel
Pry bar
Drill
Level
Nail set
Stapler
Aviation snips
Roofing knife
Caulk gun
Utility knife
T-bevel
Bay window unit
Prebuilt roof frame kit
Metal support brackets
2× lumber
16d galvanized common nails
16d and 8d galvanized casing nails
3" and 2" galvanized utility screws
16d casing nails
Tapered wood shims
Building paper
Fiberglass insulation
6-mil polyethylene sheeting
Drip edge
1" roofing nails
Step flashing
Shingles
Top flashing
Roofing cement
2 × 2 lumber
5½" skirt boards
¾" exterior-grade plywood
Paintable silicone caulk

Tips for Installing a Bay Window ▸

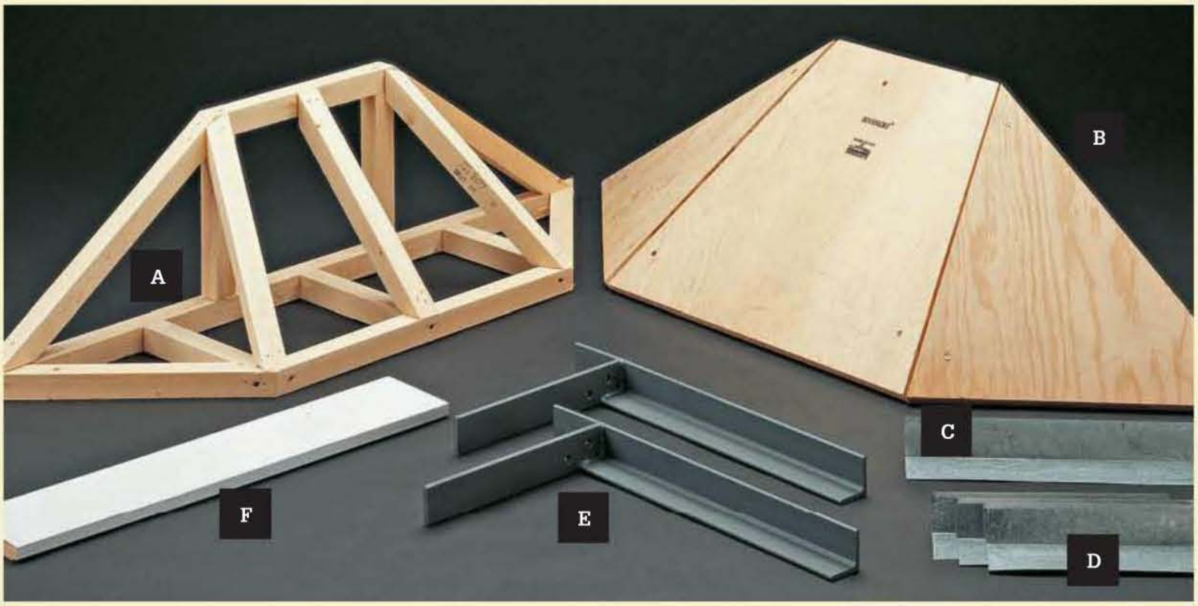

Use prebuilt accessories to ease installation of a bay window. Roof frames (A) come complete with sheathing (B), metal top flashing (C), and step flashing (D) and can be special-ordered at most home centers. You will have to specify the exact size of your window unit and the angle (pitch) you want for the roof. You can cover the roof inexpensively with building paper and shingles or order a copper or aluminum shell. Metal support braces (E) and skirt boards (F) can be ordered at your home center if not included with the window unit. Use two braces for bay windows up to 5 ft. wide and three braces for larger windows. Skirt boards are clad with aluminum or vinyl and can be cut to fit with a circular saw or miter saw.

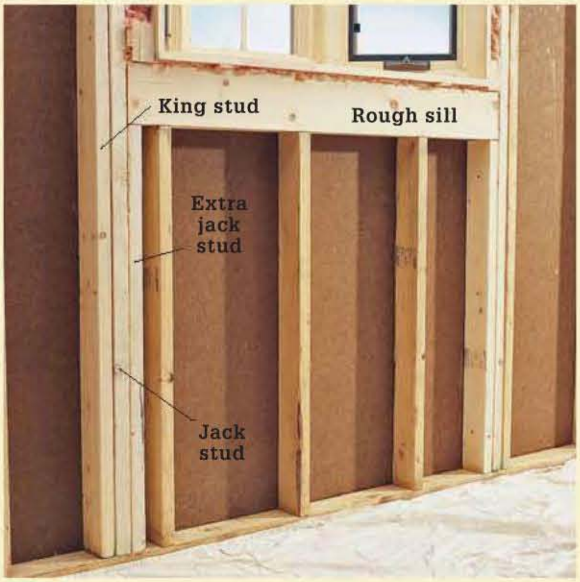

Construct a bay window frame similar to that for a standard window (see page 257) but use a built-up sill made from two 2 × 6s sandwiched around ½" plywood. Install extra jack studs under the sill ends to help carry the window's weight.

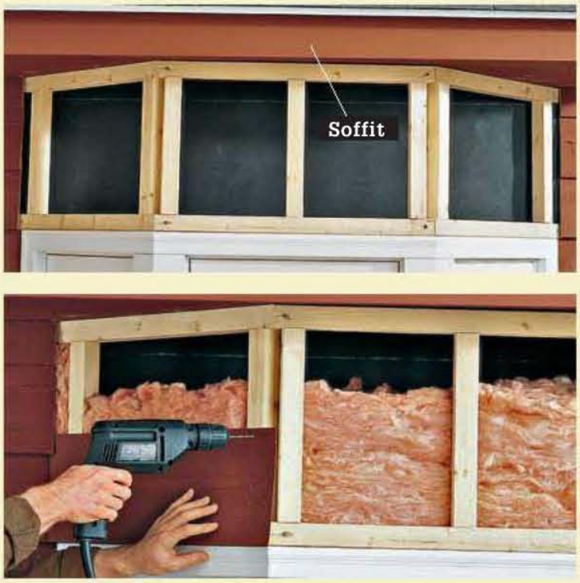

Build an enclosure above the bay window if the roof soffit overhangs the window. Build a 2 × 2 frame (top) to match the angles of the bay window, and attach the frame securely to the wall and overhanging soffit. Install a vapor barrier and insulation, then finish the enclosure so it matches the siding (bottom).

How to Install a Bay Window

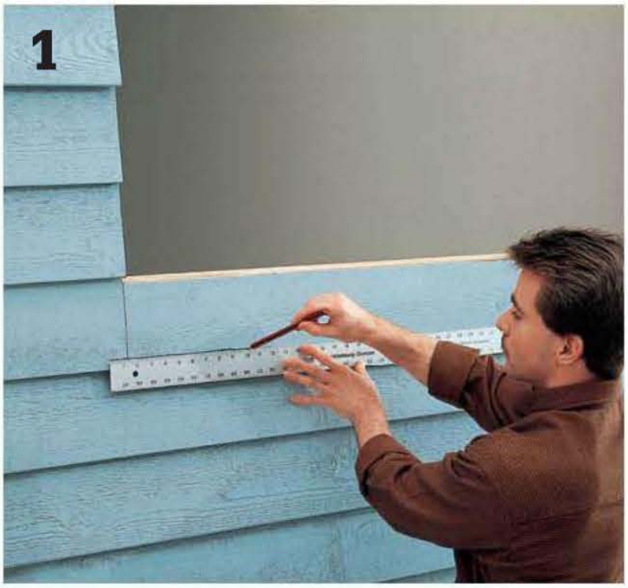

Prepare the project site and remove interior wall surfaces (pages 32 to 35), then frame the rough opening. Remove the exterior wall surfaces. Mark for removal a section of siding directly below the rough opening. The width of the marked area should equal that of the window unit and the height should equal that of the skirt board.

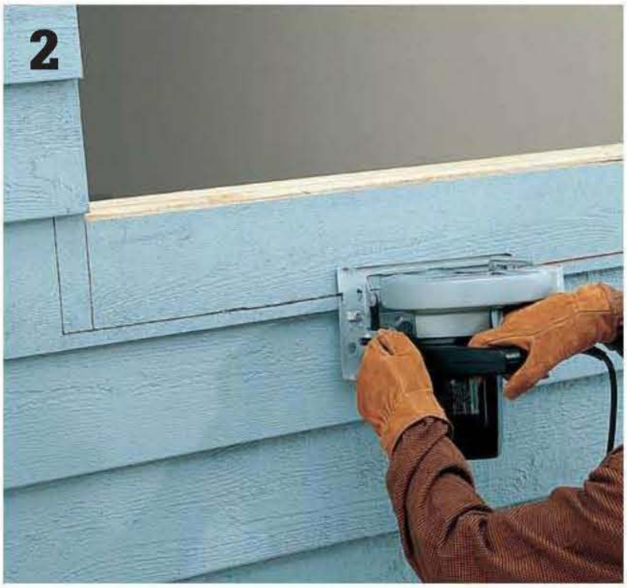

Set the blade on a circular saw just deep enough to cut through the siding, then cut along the outline. Stop just short of the corners to avoid damaging the siding outside the outline. Use a sharp chisel to complete the corner cuts. Remove the cut siding inside the outline.

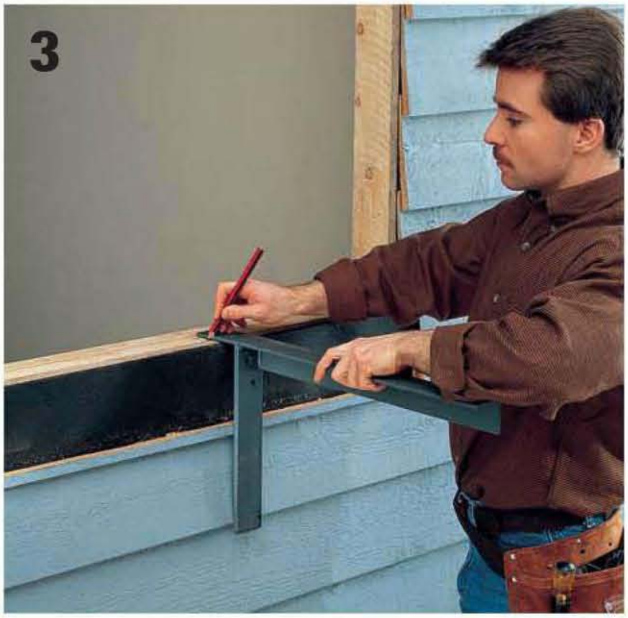

Position the support braces along the rough sill within the widest part of the bay window and above the cripple stud locations. Add cripple studs to match the support brace locations, if necessary. Draw outlines of the braces on the top of the sill. Use a chisel or circular saw to notch the sill to a depth equal to the thickness of the top arm of the support braces.

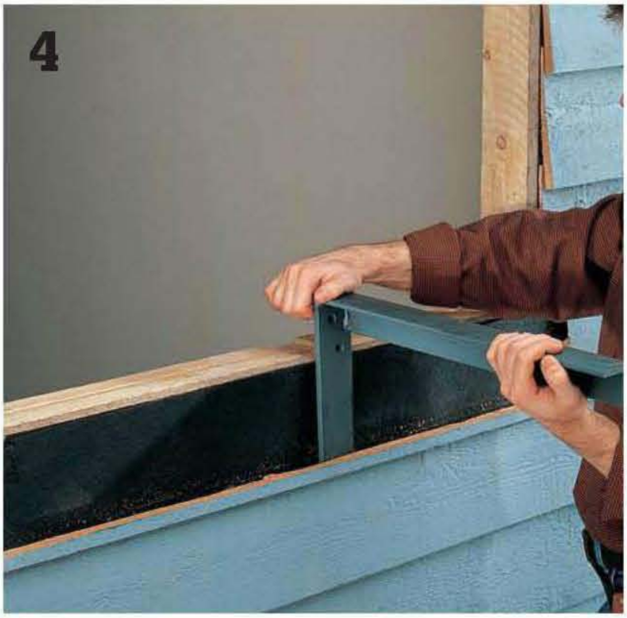

Slide the support braces down between the siding and the sheathing. Pry the siding material away from the sheathing slightly to make room for the braces, if necessary. *Note: On stucco, you will need to chisel notches in the masonry surface to fit the support braces.*

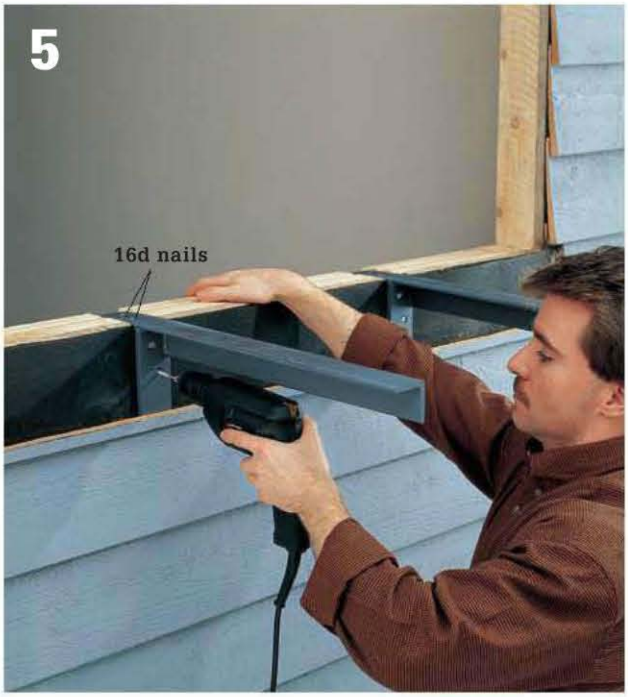

Attach the braces to the rough sill with galvanized 16d common nails. Drive 3" utility screws through the front of the braces and into the rough sill to prevent twisting.

Lift the bay window onto the support braces and slide it into the rough opening. Center the unit within the opening.

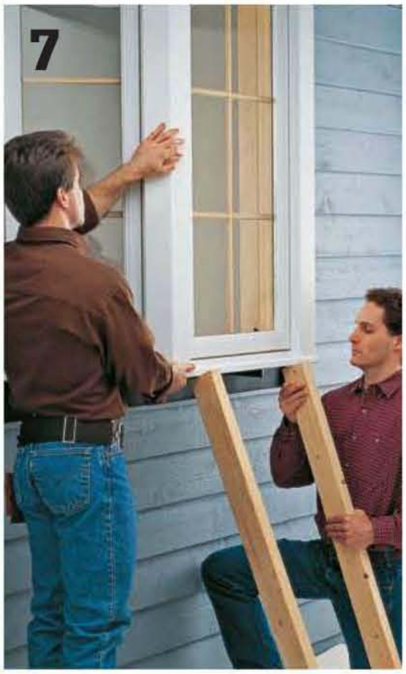

Check the window unit to make sure it is level. If necessary, drive shims under the low side to level the window. Temporarily brace the outside bottom edge of the unit with 2 × 4s to keep it from moving on the braces.

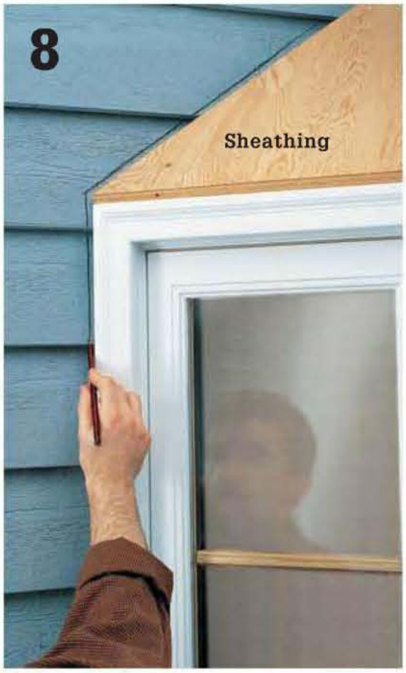

Set the roof frame on top of the window, with the sheathing loosely tacked in place. Trace the outline of the window and roof unit onto the siding. Leave a gap of about ½" around the roof unit to allow room for flashing and shingles.

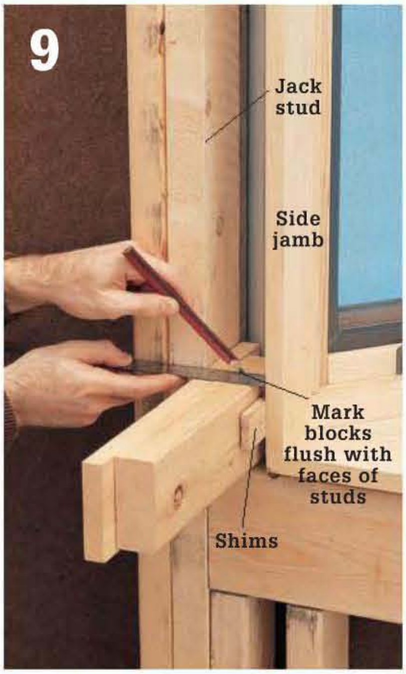

If the gap between the side jambs and jack studs is more than 1" wide, mark and cut wood blocks to bridge the gap (smaller gaps require no blocks). Leave a small space for inserting wood shims. Remove the window, then attach blocks every 12" along studs.

(continued)

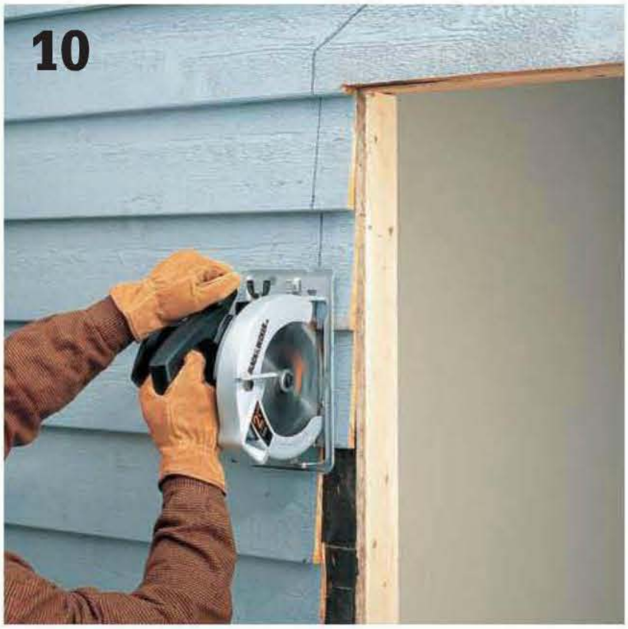

Cut the siding just down to the sheathing along the outline using a circular saw. Stop just short of the corners, then use a wood chisel to complete the corner cuts. Remove the cut siding. Pry the remaining siding slightly away from the sheathing around the roof outline to allow for easy installation of the metal flashing. Cover the exposed sheathing with 8"-wide strips of building paper (step 3, page 260).

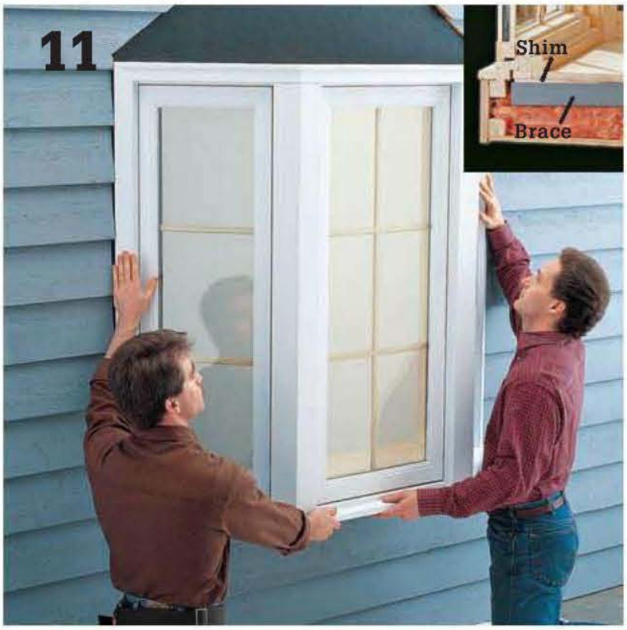

Set the bay window unit back on the braces, and slide it back into the rough opening until the brick moldings are tight against the sheathing. Insert wood shims between the outside end of the metal braces and the seat board (inset). Check the unit to make sure it is level, and adjust the shims, if necessary.

Anchor the window by drilling pilot holes and driving 16d casing nails through the brick molding and into the framing members. Space nails every 12", and use a nail set to drive the nail heads below the surface of the wood.

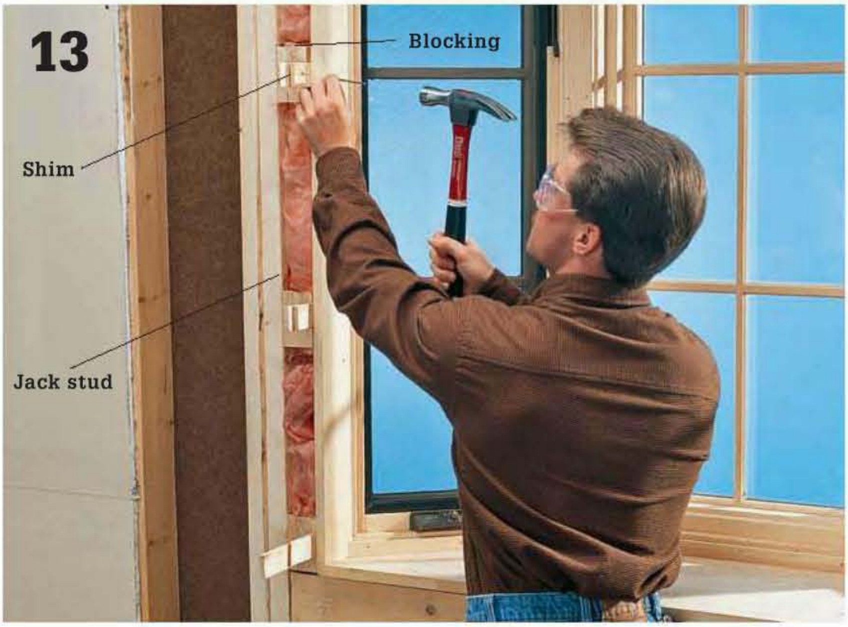

Drive wood shims into the spaces between the side jambs and the blocking or jack studs and between the headboard and header, spacing the shims every 12". Fill the spaces around the window with loosely packed fiberglass insulation. At each shim location, drive 16d casing nails through the jambs and shims and into the framing members. Cut off the shims flush with the framing members using a handsaw or utility knife. Use a nail set to drive the nail heads below the surface. If necessary, drill pilot holes to prevent splitting the wood.

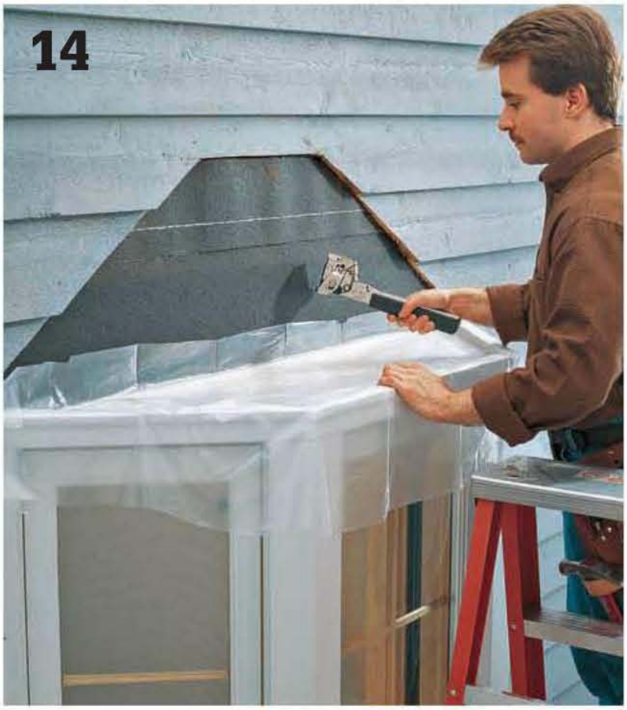

Staple sheet plastic over the top of the window unit to serve as a vapor barrier. Trim the edges of the plastic around the top of the window using a utility knife.

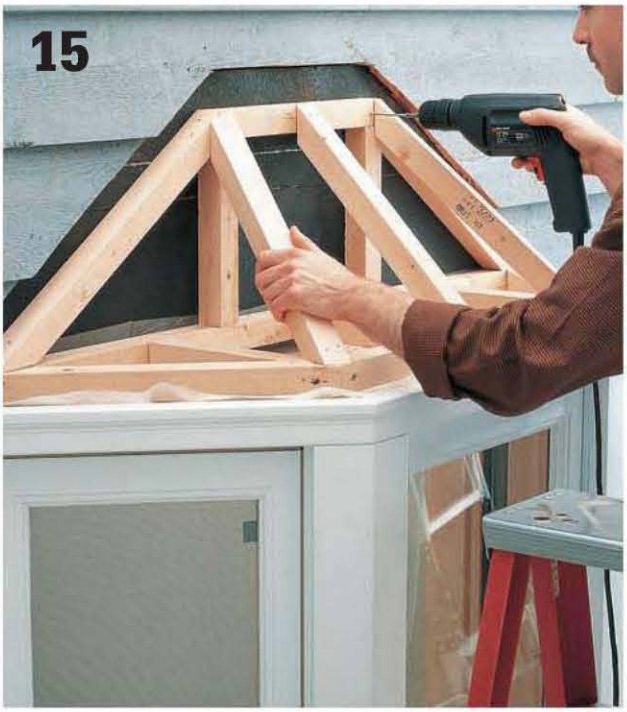

Remove the sheathing pieces from the roof frame, then position the frame on top of the window unit. Attach the roof frame to the window and to the wall at stud locations using 3" utility screws.

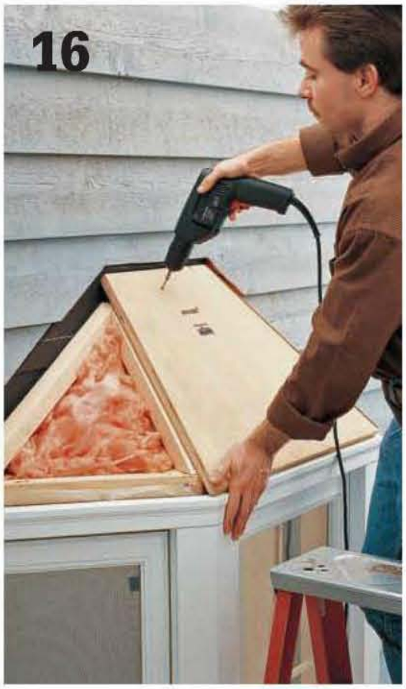

Fill the empty space inside the roof frame with loosely packed fiberglass insulation. Screw the sheathing back onto the roof frame using 2" utility screws.

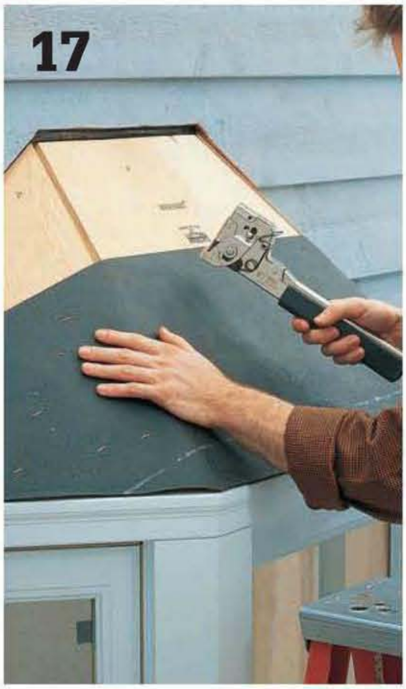

Staple asphalt building paper over the roof sheathing. Make sure each piece of building paper overlaps the one below by at least 5".

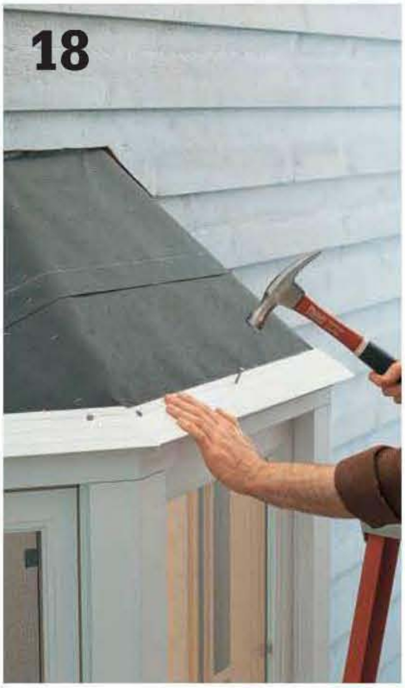

Cut drip edges with aviation snips, then attach them around the edge of the roof sheathing using roofing nails.

(continued)

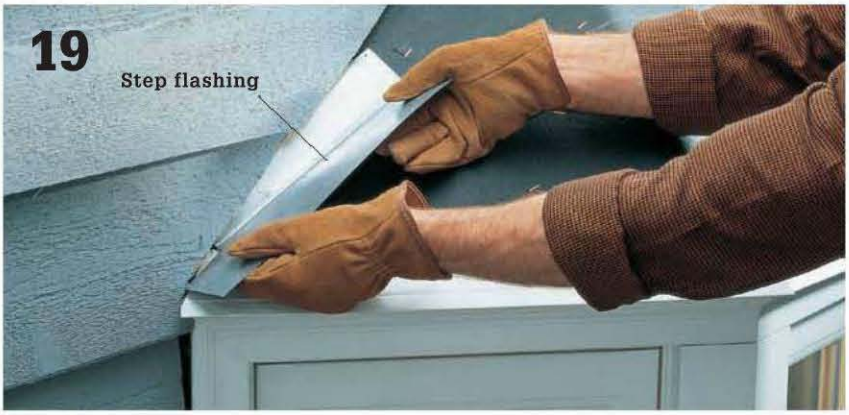

Cut and fit a piece of step flashing on each side of the roof frame. Adjust the flashing so it overhangs the drip edge by ¼". Flashings help guard against moisture damage.

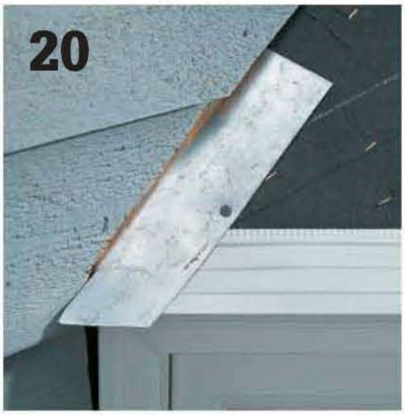

Trim the end of the flashing to the same angle as the drip edge. Nail the flashing to the sheathing with roofing nails.

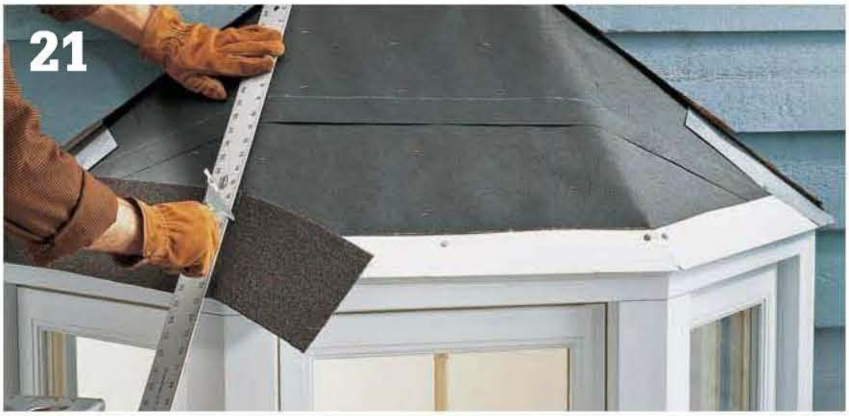

Cut 6"-wide strips of shingles for the starter row. Use roofing nails to attach the starter row shingles so they overhang the drip edge by about ½". Cut the shingles along the roof hips using a straightedge and a roofing knife.

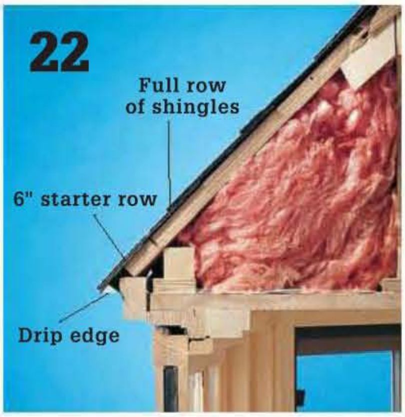

Nail a full row of shingles over the starter row, aligning the bottom edges with the bottom edge of the starter row. Make sure shingle notches are not aligned.

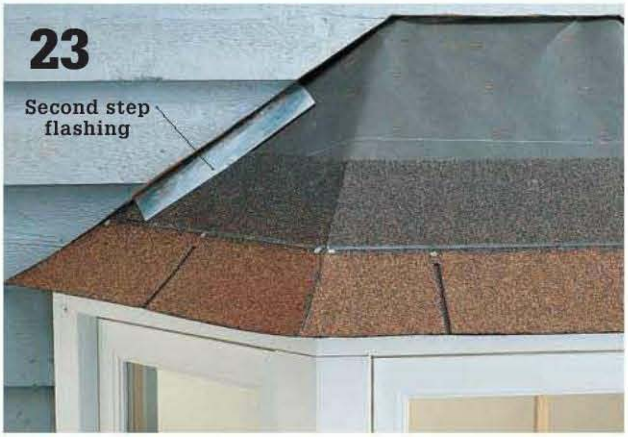

Install another piece of step flashing on each side of the roof, overlapping the first piece of flashing by about 5".

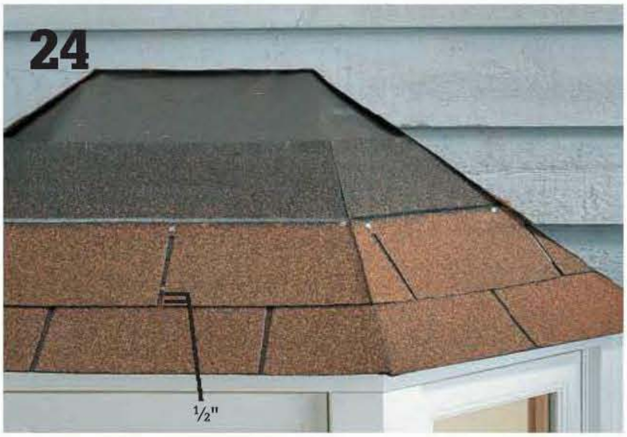

Cut and install another row of full shingles. The bottom edges should overlap the tops of the notches on the previous row by ½". Attach the shingles with roofing nails driven just above the notches.

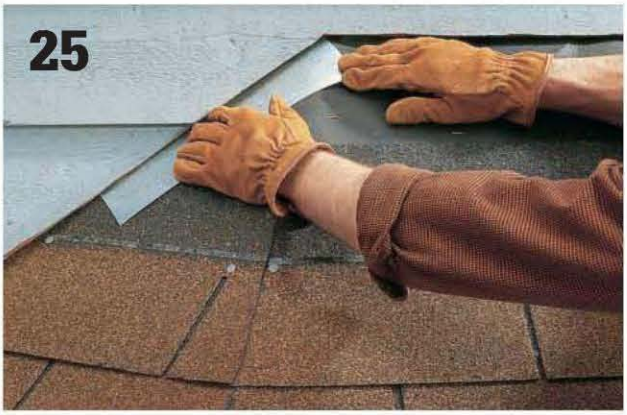

Continue installing alternate rows of step flashing and shingles to the top of the roof. Bend the last pieces of step flashing to fit over the roof hips.

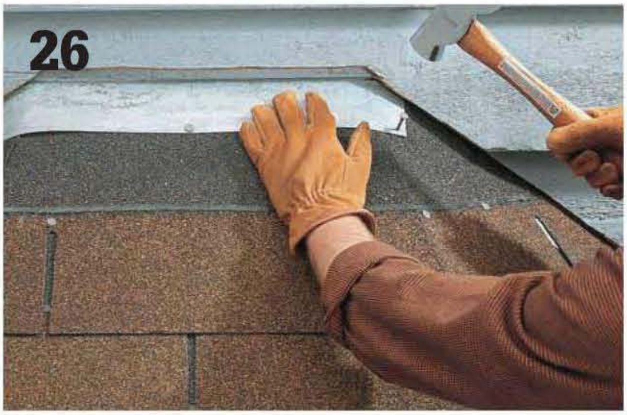

When the roof sheathing is covered with shingles, install the top flashing. Cut and bend the ends over the roof hips, and attach it with roofing nails. Attach the remaining rows of shingles over the top flashing.

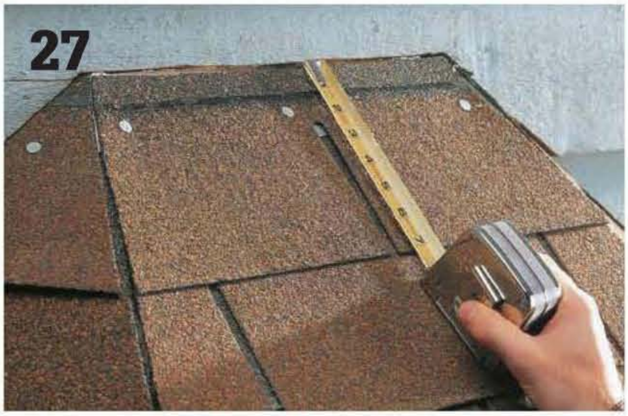

Find the height of the final row of shingles by measuring from the top of the roof to a point ½" below the top of the notches on the last installed shingle. Trim the shingles to fit.

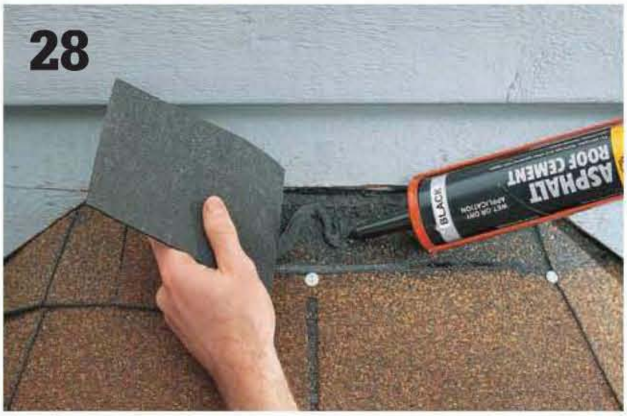

Attach the final row of shingles with a thick bead of roofing cement—not nails. Press firmly to ensure a good bond.

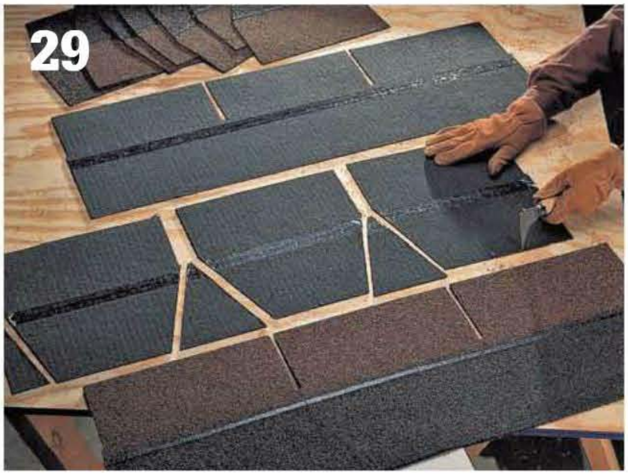

Make ridge caps by cutting shingles into 1-ft.-long sections. Use a roofing knife to trim off the top corners of each piece, so the ridge caps will be narrower at the top than at the bottom.

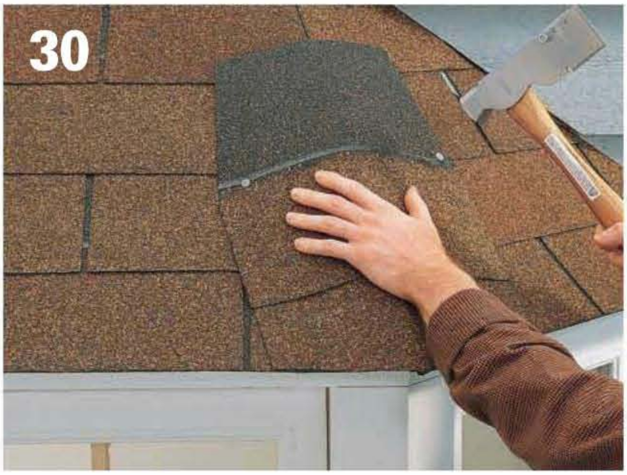

Install the ridge caps over the roof hips, beginning at the bottom of the roof. Trim the bottom ridge caps to match the edges of the roof. Keep the same amount of overlap with each layer.

(continued)

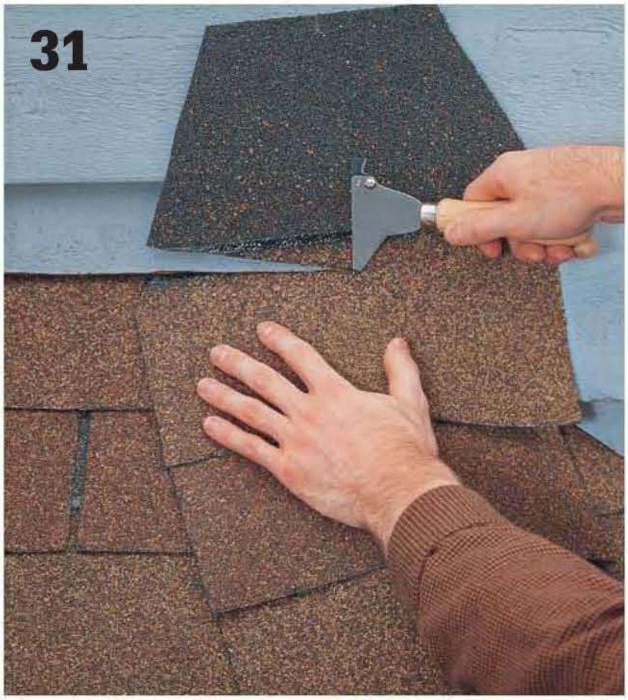

31

At the top of the roof hips, use a roofing knife to cut the shingles to fit flush with the wall. Attach the shingles with roofing cement—do not use any nails.

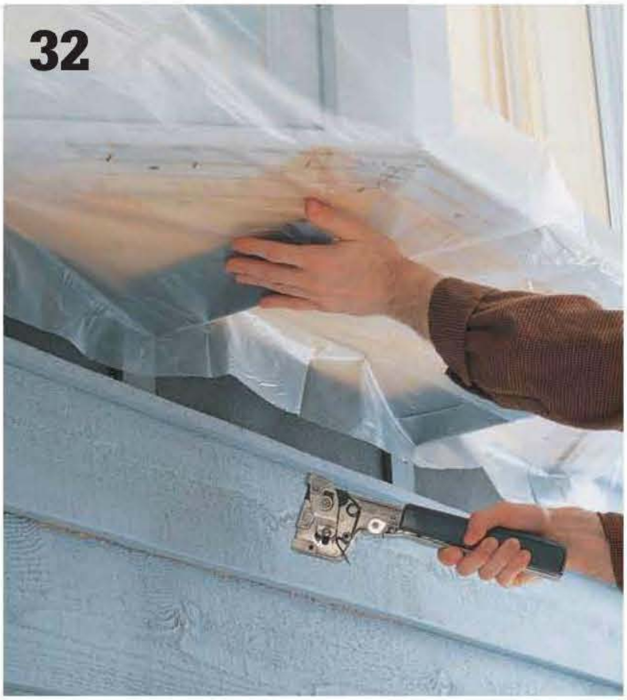

32

Staple sheet plastic over the bottom of the window unit to serve as a vapor barrier. Trim the plastic around the bottom of the window.

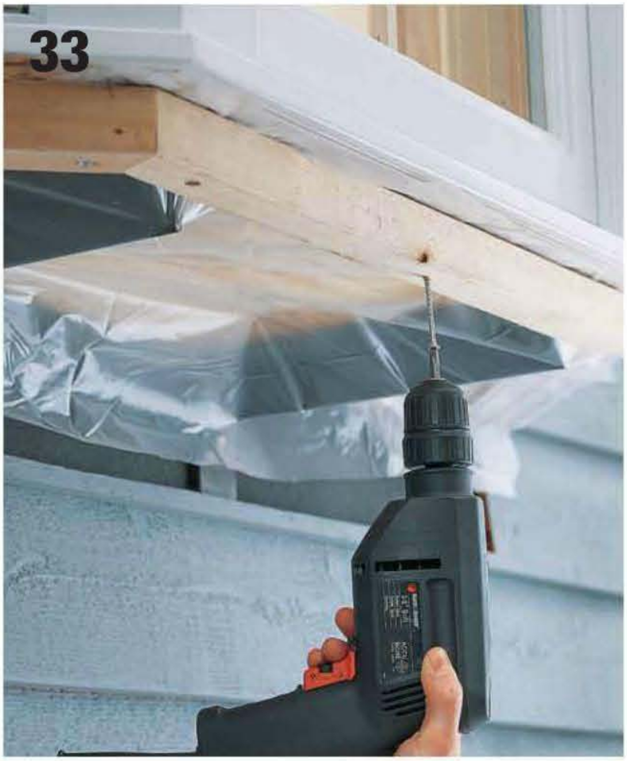

33

Cut and attach a 2 × 2 skirt frame around the bottom of the bay window using 3" galvanized utility screws. Set the skirt frame back about 1" from the edges of the window.

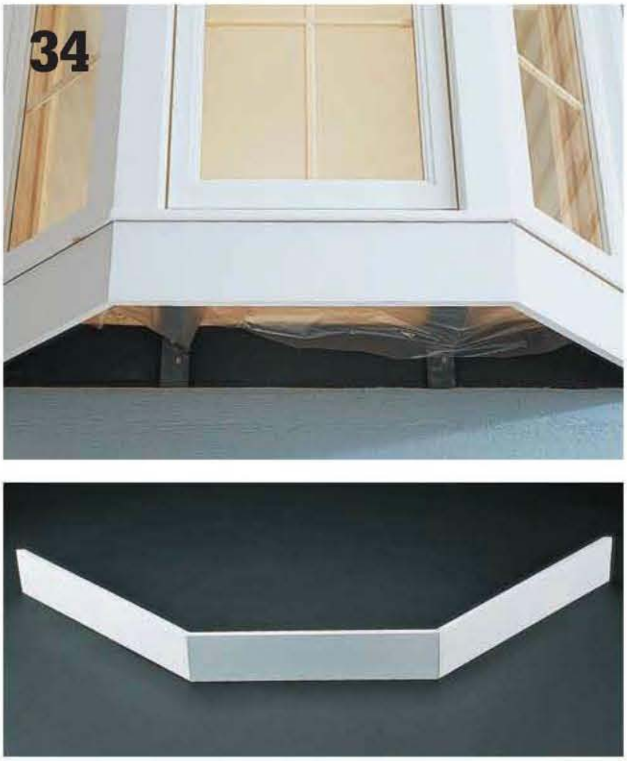

34

Cut skirt boards to match the shape of the bay window bottom, mitering the ends to ensure a tight fit. Test-fit the skirt board pieces to make sure they match the bay window bottom.

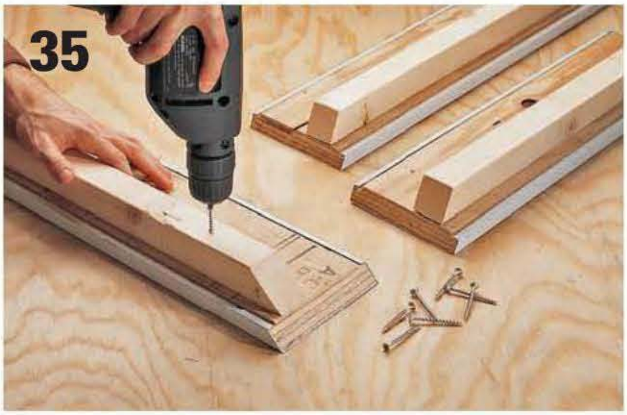

Cut a 2 × 2 furring strip for each skirt board. Miter the ends to the same angles as the skirt boards. Attach the furring strips to the back of the skirt boards, 1" from the bottom edges, using 2" galvanized utility screws.

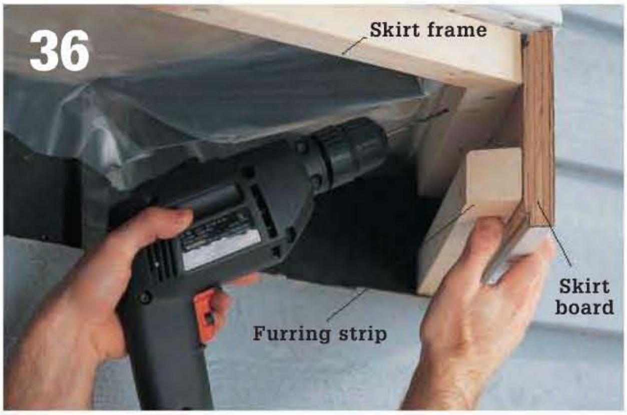

Attach the skirt board pieces to the skirt frame. Drill ⅛" pilot holes every 6" through the back of the skirt frame and into the skirt boards, then attach the skirt boards with 2" galvanized utility screws.

Measure the space inside the skirt boards using a T-bevel to duplicate the angles. Cut a skirt bottom from ¾" exterior-grade plywood to fit this space.

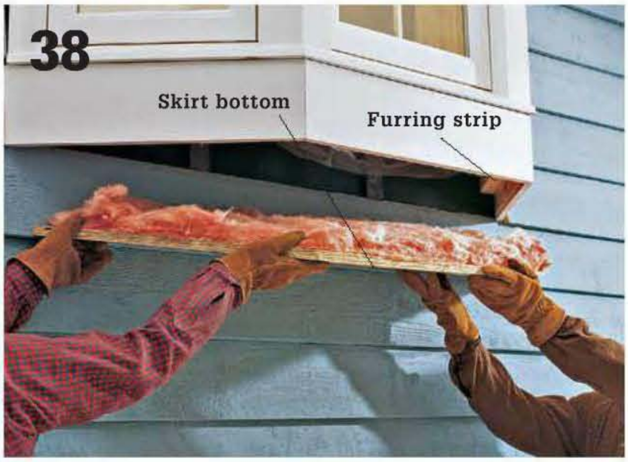

Lay fiberglass insulation on the skirt bottom. Position the skirt bottom against the furring strips and attach it by driving 2" galvanized utility screws every 6" through the bottom and into the furring strips.

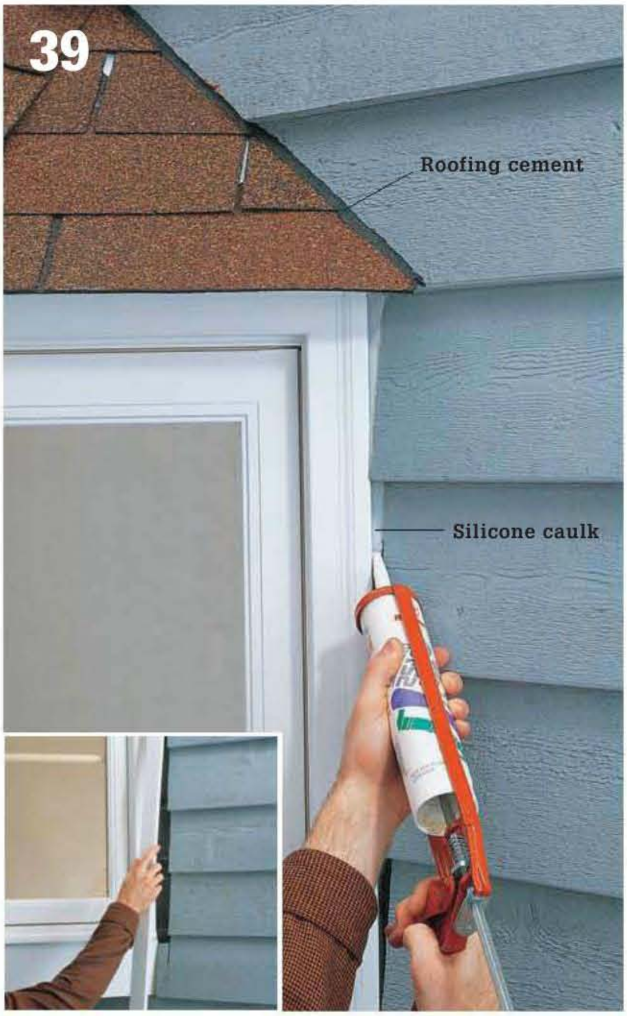

Install any additional trim pieces (inset) specified by your window manufacturer using 8d galvanized casing nails. Seal the roof edges with roofing cement, and seal around the rest of the window with paintable silicone caulk.

Installing Skylights

Since skylights let in so much light, their sizing and placement are important considerations. A skylight that's too big can quickly overheat a space, especially in an attic. The same is true of using too many skylights in any one room. For that reason it's often best to position a skylight away from the day's brightest sun. You may want an operable skylight that opens and closes to vent warm air.

When a skylight is installed above an unfinished attic space, a special skylight shaft must be constructed to channel light directly to the room below.

Skylights can offer warmth in the winter, cooling ventilation in the summer, and a view of the sky or the treetops around your house during any season. And, of course, skylights provide natural light.

Installing a skylight above finished space involves other considerations. First, the ceiling surface must be removed to expose the rafters. To remove wall and ceiling surfaces, see pages 28 to 31.

A skylight frame is similar to a standard window frame. It has a header and sill, like a window frame, but it has king rafters rather than king studs. Skylight frames also have trimmers that define the sides of the rough opening. Refer to the manufacturer's instructions to determine what size to make the opening for the skylight you select.

With standard rafter-frame roof construction, you can safely cut into one or two rafters as long as you permanently support the cut rafters, as shown in the following steps. If your skylight requires alteration of more than two rafters or if your roofing is made with unusually heavy material, such as clay tile or slate, consult an architect or engineer before starting the project.

Today's good-quality skylight units are unlikely to leak, but a skylight is only as leakproof as its installation. Follow the manufacturer's instructions, and install the flashing meticulously, as it will last a lot longer than any sealant.

Tools & Materials ▸

- 4-ft. level
- Circular saw
- Drill
- Combination square
- Reciprocating saw
- Pry bar
- Chalk line
- Stapler
- Caulk gun
- Utility knife
- Aviation snips
- Plumb bob
- Jigsaw
- Wallboard tools
- 2× lumber
- 1 × 4
- 16d and 10d common nails
- Building paper
- Roofing cement
- Skylight flashing
- 2", 1¼", and ¾" roofing nails
- Finish nails
- Fiberglass insulation
- ½" wallboard
- Twine
- Wallboard screws
- 6-mil polyethylene sheeting
- Finishing materials

How to Install a Skylight

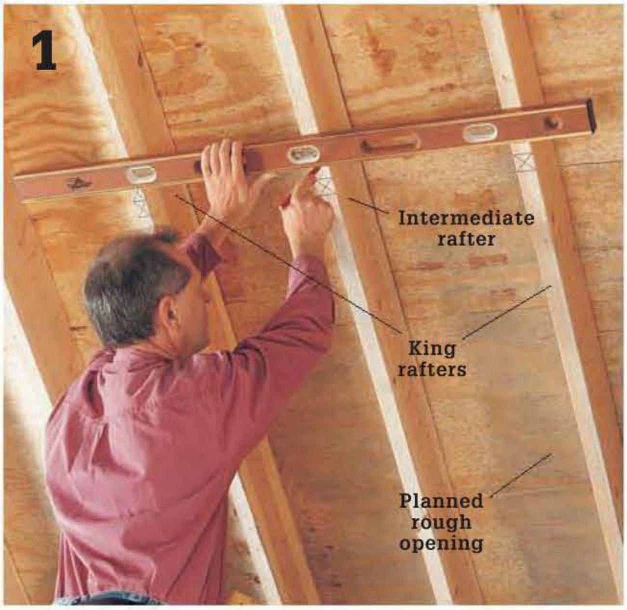

Use the first rafter on each side of the planned rough opening as a king rafter. Measure and mark where the double header and sill will fit against the king rafters. Then, use a level as a straightedge to extend the marks across the intermediate rafter.

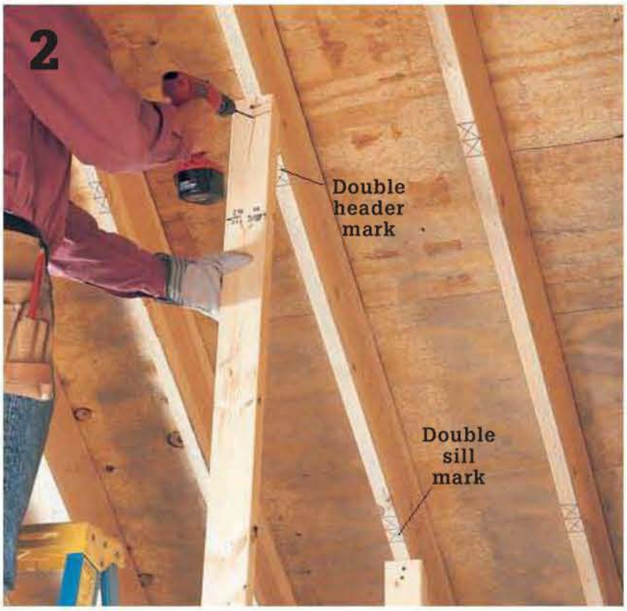

Brace the intermediate rafter by installing two 2 × 4s between the rafter and the attic floor. Position the braces just above the header marks and just below the sill marks. Secure them temporarily to the rafter and subfloor (or joists) with screws.

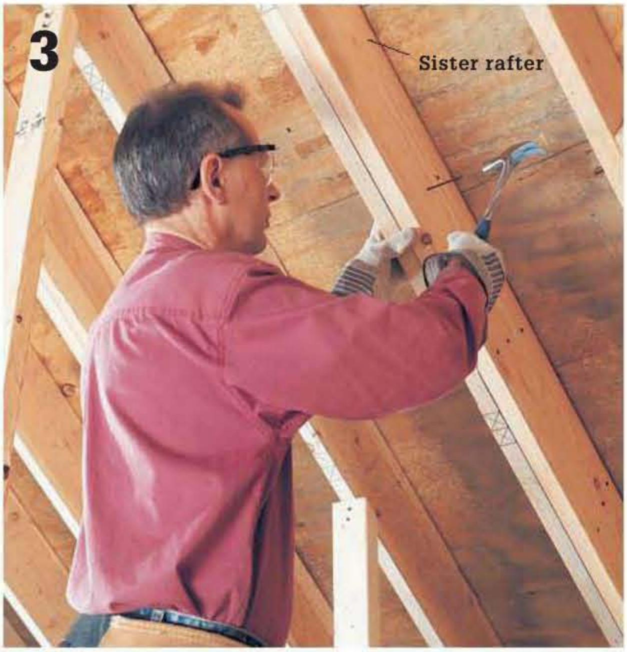

Reinforce each king rafter by attaching a full-length "sister" rafter against its outside face. Cut sister rafters from the same size of lumber as existing rafters, matching lengths and end cuts exactly. Work each one into position, flush against the outside face of the king rafters, then nail the sisters to the kings with pairs of 10d common nails spaced 12" apart.

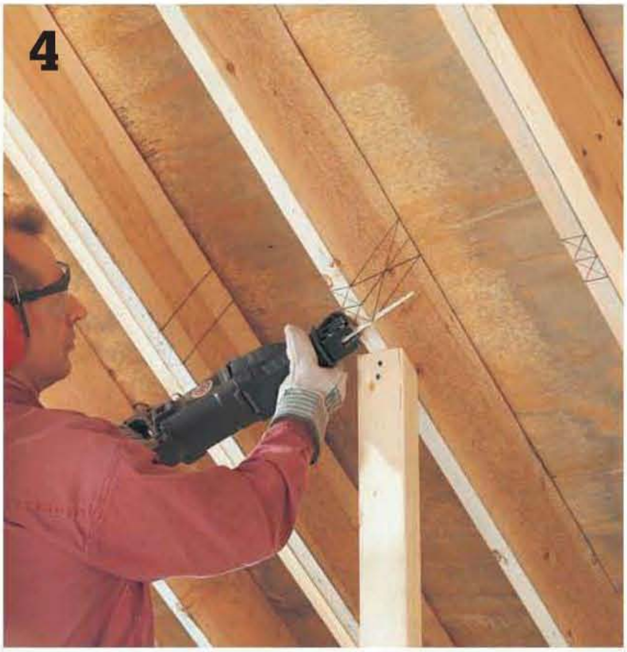

Use a combination square to transfer the sill and header marks across the face of the intermediate rafter, then cut along the outermost lines with a reciprocating saw. Do not cut into the roof sheathing. Carefully remove the cutout section with a pry bar. The remaining rafter portions will serve as cripple rafters.

(continued)

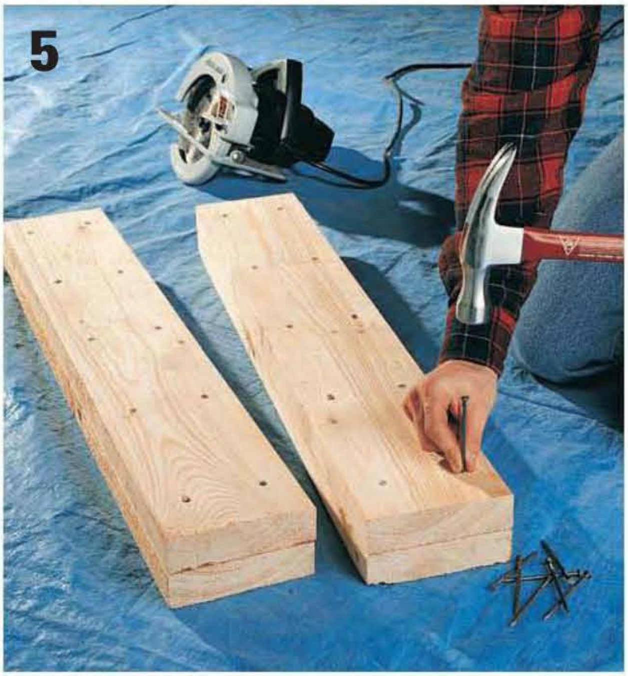

Build a double header and double sill to fit snugly between the king rafters, using 2× lumber that is the same size as the rafters. Nail the header pieces together using pairs of 10d nails spaced 6" apart.

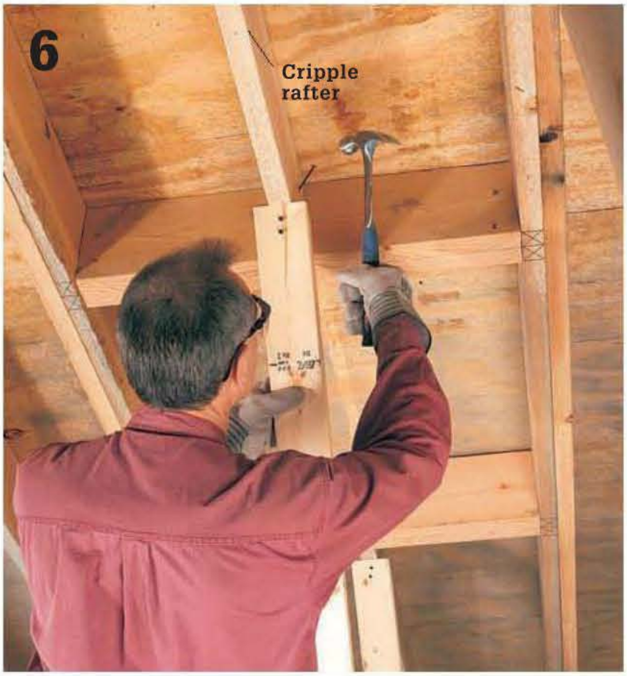

Install the header and sill, anchoring them to the king rafters and cripple rafters with 16d common nails. Make sure the ends of the header and sill are aligned with the appropriate marks on the king rafters.

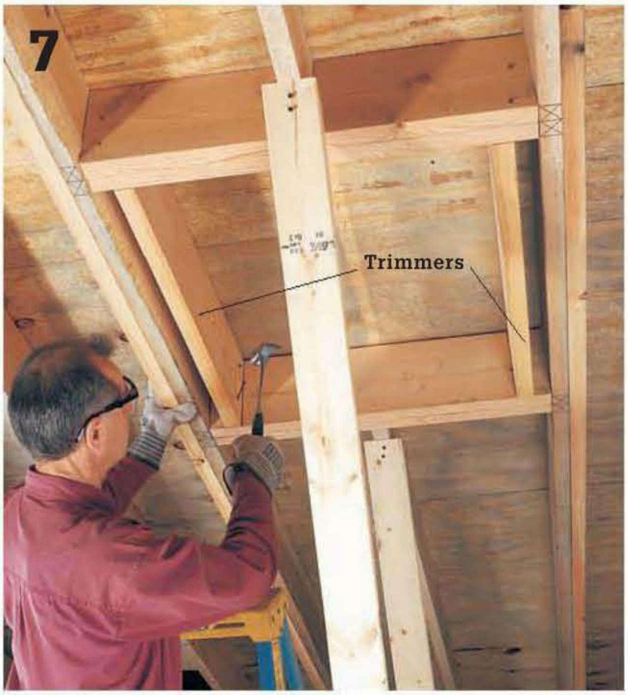

If your skylight unit is narrower than the opening between the king studs, measure and make marks for the trimmers: They should be centered in the opening and spaced according to the manufacturer's specifications. Cut the trimmers from the same 2× lumber used for the rest of the frame, and nail them in place with 10d common nails. Remove the 2 × 4 braces.

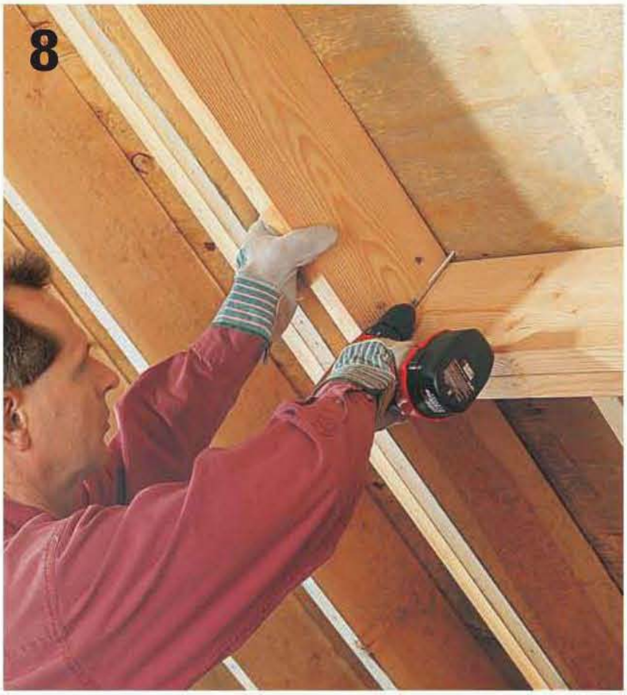

Mark the opening for the roof cutout by driving a screw through the sheathing at each corner of the frame. Then, tack a couple of scrap boards across the opening to prevent the roof cutout from falling and causing damage below.

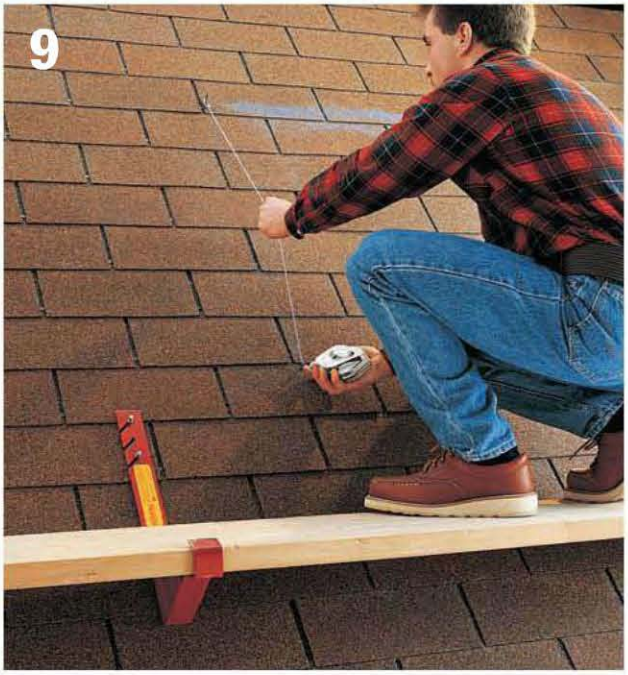

From the roof, measure between the screws to make sure the rough opening dimensions are accurate. Snap chalk lines between the screws to mark the rough opening, then remove the screws.

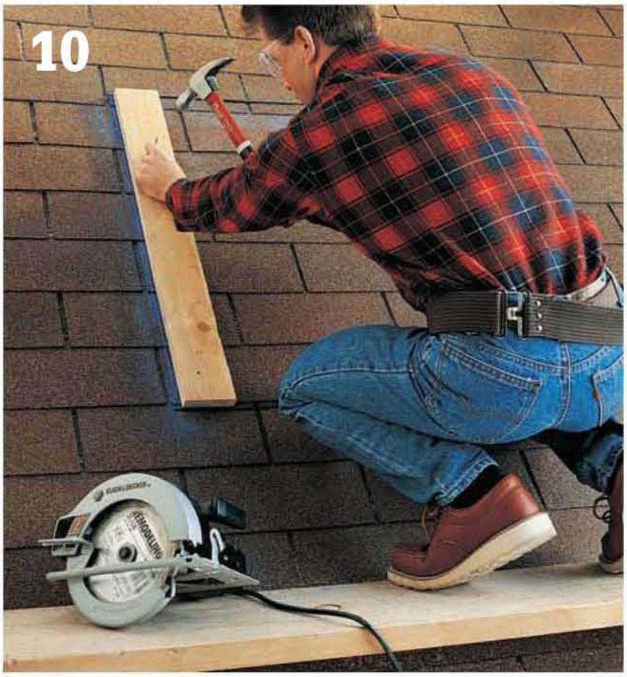

Tack a straight 1 × 4 to the roof, aligned with the inside edge of one chalk line. Make sure the nail heads are flush with the surface of the board.

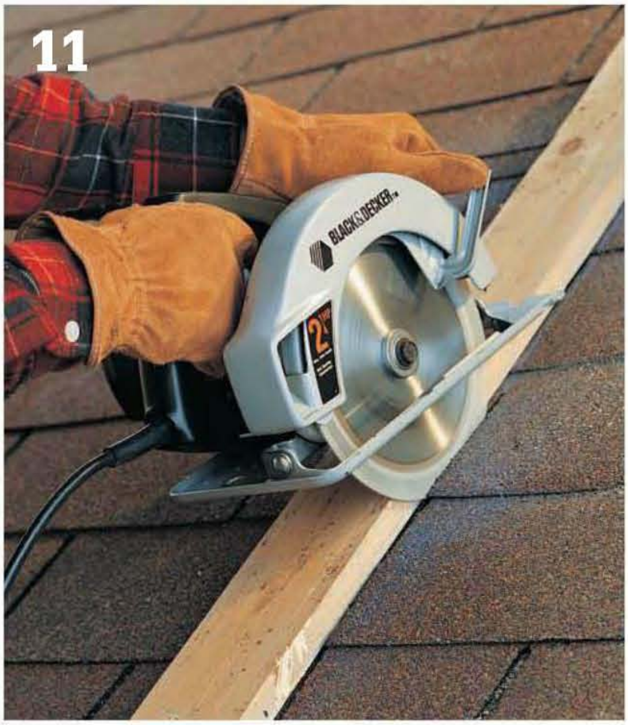

Cut through the shingles and sheathing along the chalk line using a circular saw and an old blade or a remodeling blade. Rest the saw foot on the 1 × 4, and use the edge of the board as a guide. Reposition the 1 × 4, and cut along the remaining lines. Remove the cutout roof section.

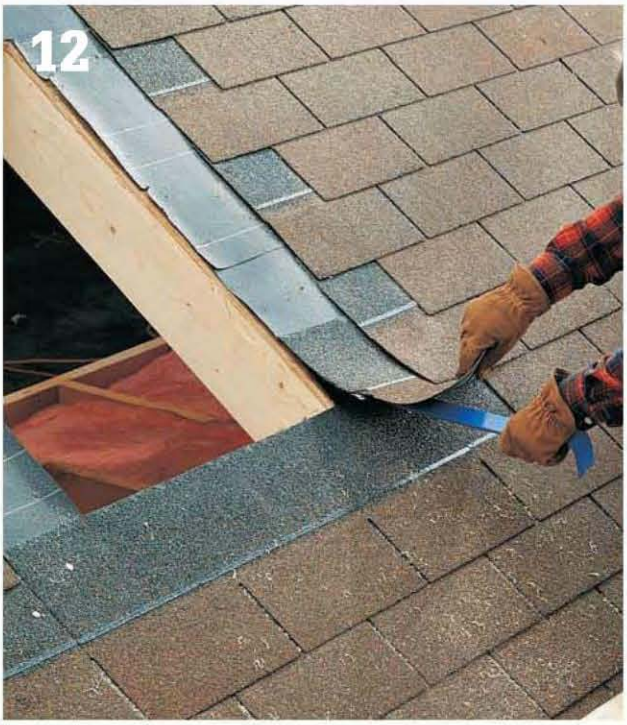

Remove the shingles around the rough opening with a flat pry bar, exposing at least 9" of building paper on all sides of the opening. Remove whole shingles rather than cutting them.

(continued)

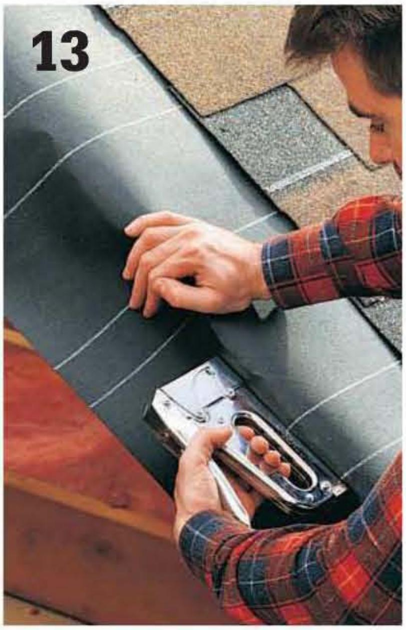

Cut strips of building paper and slide them between the shingles and existing building paper. Wrap the paper so that it covers the faces of the framing members, and staple it in place.

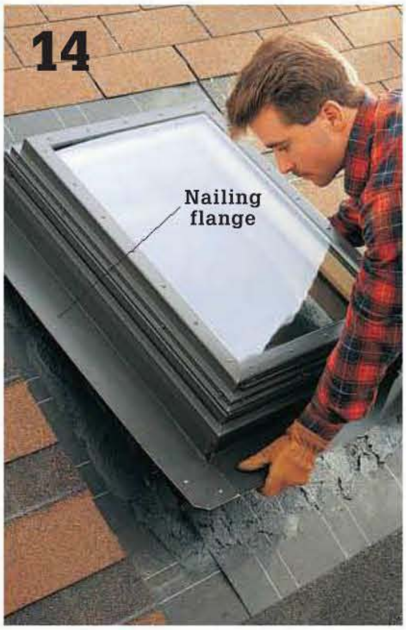

Spread a 5"-wide layer of roofing cement around the roof opening. Set the skylight into the opening so that the nailing flange rests on the roof. Finally, adjust the unit so it sits squarely in the opening.

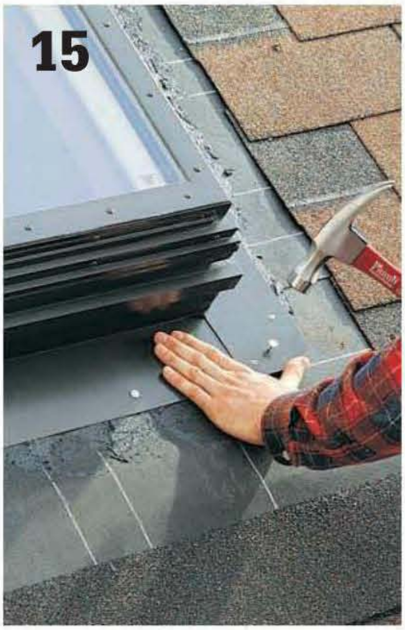

Nail through the flange and into the sheathing and framing members with 2" galvanized roofing nails spaced every 6". *Note: If skylight uses L-shaped brackets instead of a nailing flange, follow manufacturer's instructions.*

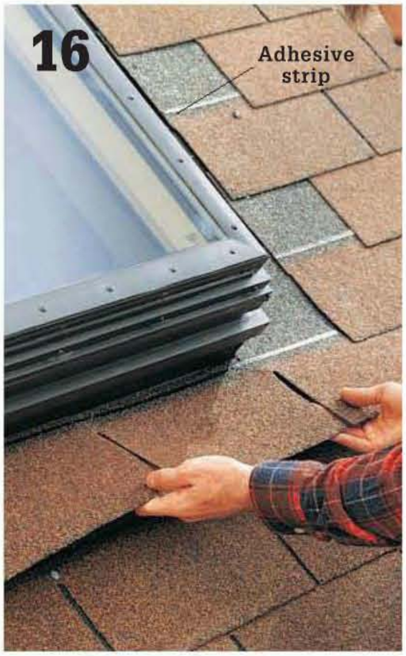

Patch in shingles up to the bottom edge of the skylight unit. Attach the shingles with 1¼" roofing nails driven just below the adhesive strip. If necessary, cut the shingles with a utility knife so they fit against the bottom of the skylight.

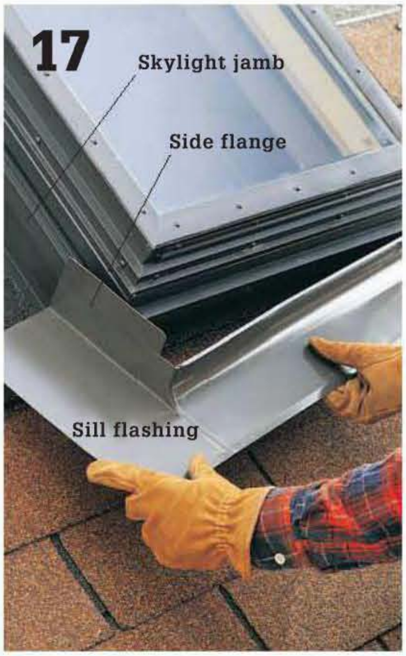

Spread roofing cement on the bottom edge of the sill flashing, then fit the flashing around the bottom of the unit. Attach flashing by driving ¾" galvanized roofing nails through the vertical side flange (near the top of the flashing) and into the skylight jambs.

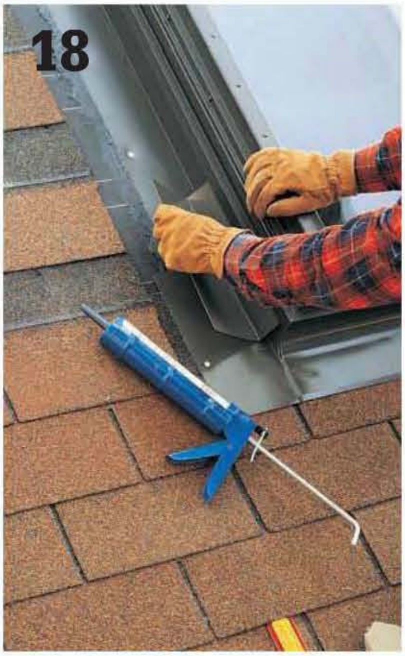

Spread roofing cement on the bottom of a piece of step flashing, then slide flashing under the drip edge on one side of the skylight. Step flashing should overlap sill flashing by 5". Press the step flashing down to bond it. Repeat on the opposite side of the skylight.

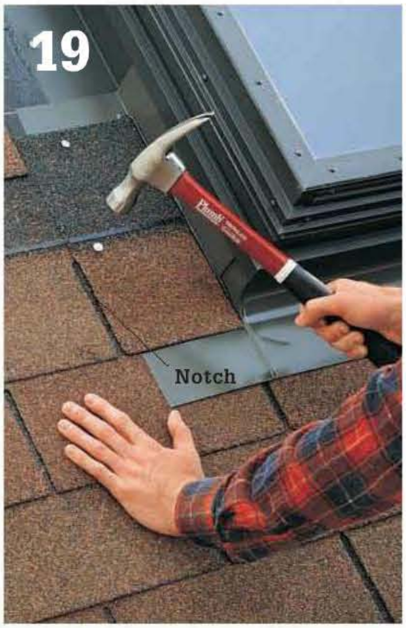

Patch in the next row of shingles on each side of the skylight, following the existing shingle pattern. Drive a 1¼" roofing nail through each shingle and the step flashing and into the sheathing. Drive additional nails just above the notches in the shingles.

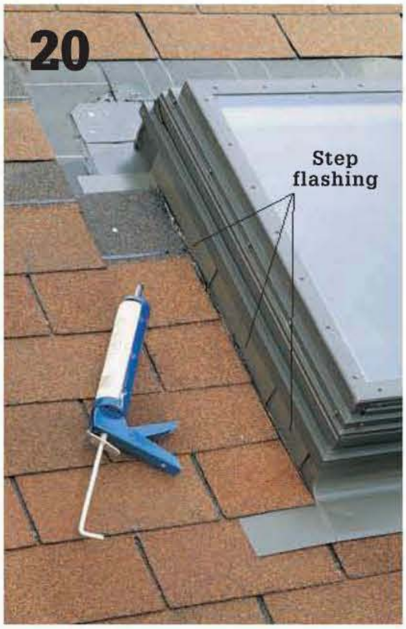

Continue applying alternate rows of step flashing and shingles using roofing cement and roofing nails. Each piece of flashing should overlap the preceding piece by 5".

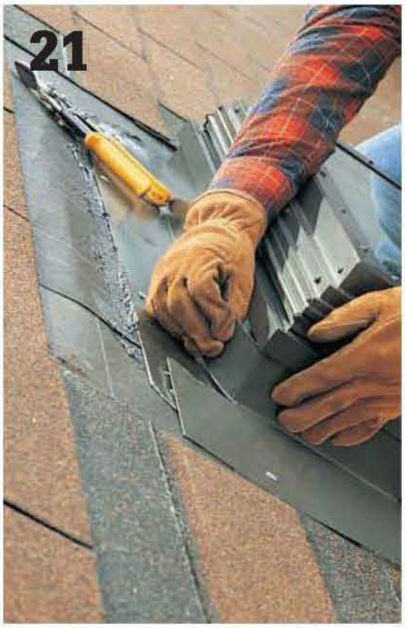

At the top of the skylight, cut and bend the last piece of step flashing on each side, so the vertical flange wraps around the corner of the skylight. Patch in the next row of shingles.

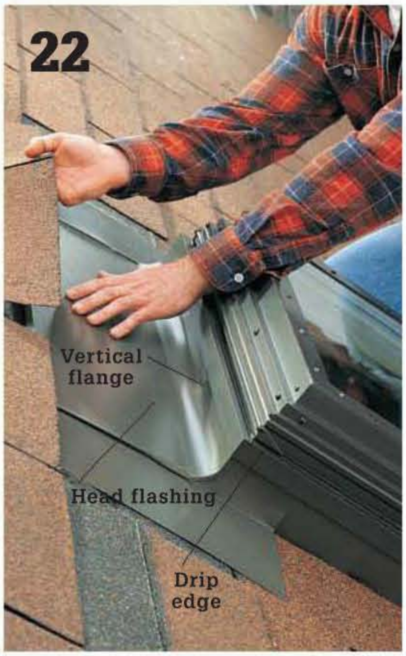

Spread roofing cement on the bottom of the head flashing to bond it to the roof. Place the flashing against the top of the skylight so the vertical flange fits under the drip edge and the horizontal flange fits under the shingles above the skylight.

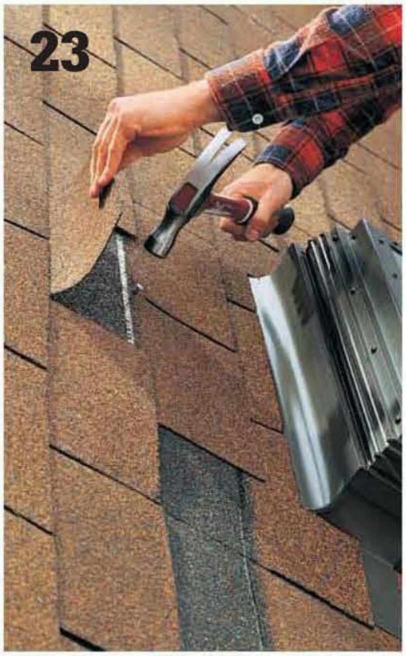

Fill in the remaining shingles, cutting them to fit, if necessary. Attach the shingles with roofing nails driven just above the notches.

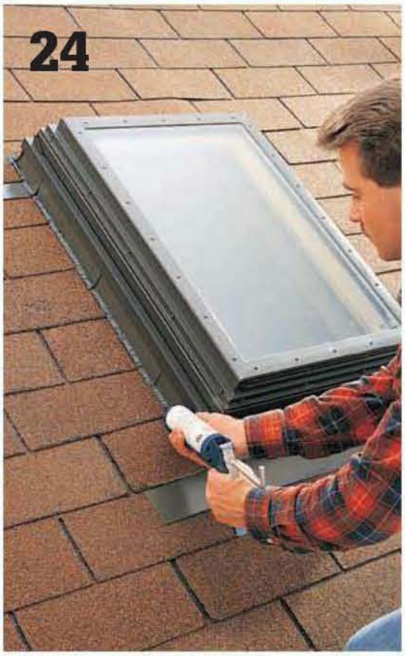

Apply a continuous bead of roofing cement along the joint between the shingles and the skylight. Finish the interior of the framed opening as desired.

How to Build a Skylight Shaft

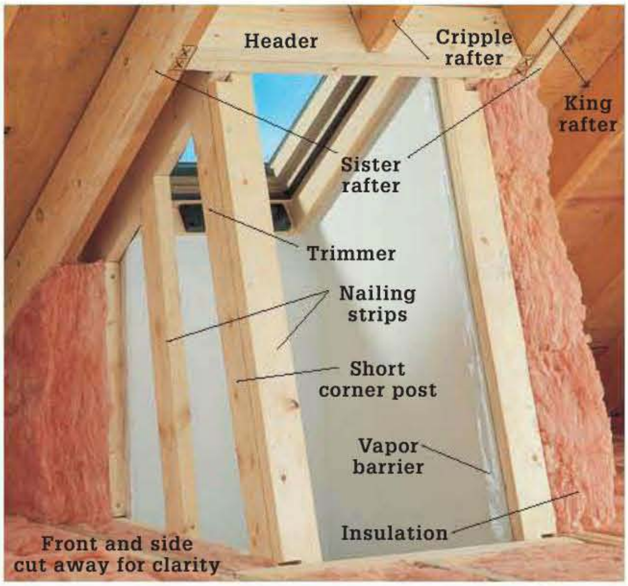
Header
Cripple rafter
King rafter
Sister rafter
Trimmer
Nailing strips
Short corner post
Vapor barrier
Insulation
Front and side cut away for clarity

A skylight shaft is made with 2 × 4 lumber and wallboard and includes a vapor barrier and fiberglass insulation. You can build a straight shaft with four vertical sides or an angled shaft that has a longer frame at ceiling level and one or more sides set at an angle. Since the ceiling opening is larger, an angled shaft lets in more direct light than a straight shaft.

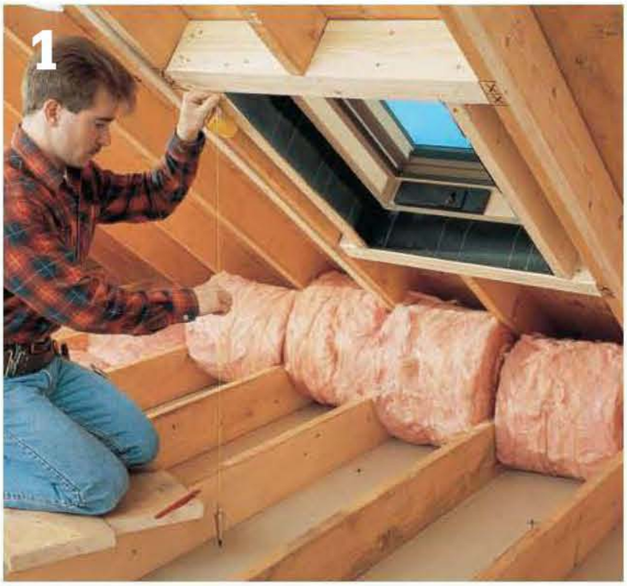

Remove any insulation in the area where the skylight will be located; turn off and reroute electrical circuits as necessary. Use a plumb bob as a guide to mark reference points on the ceiling surface, directly below the inside corners of the skylight frame.

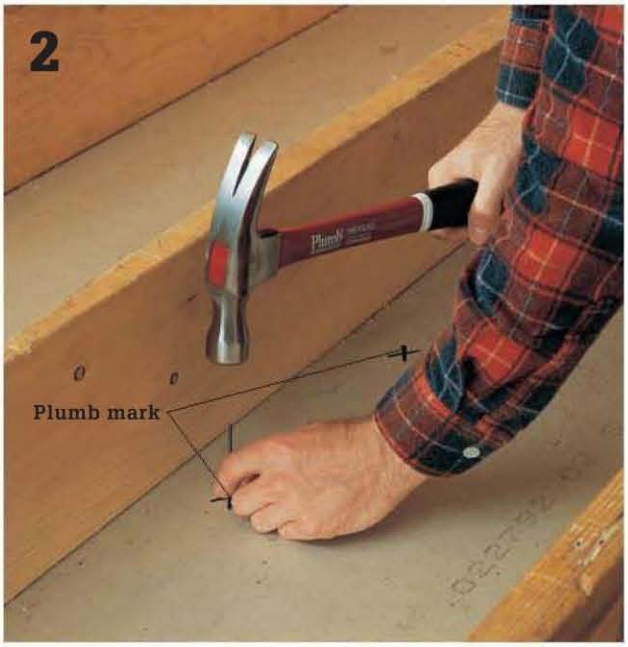
2
Plumb mark

If you are installing a straight shaft, use the plumb marks made in step 1 to define the corners of the ceiling opening; drive a finish nail through the ceiling surface at each mark. If you are installing an angled shaft, measure out from the plumb marks and make new marks that define the corners of the ceiling opening; drive finish nails at the new marks.

From the room below, mark cutting lines, then remove the ceiling surface.

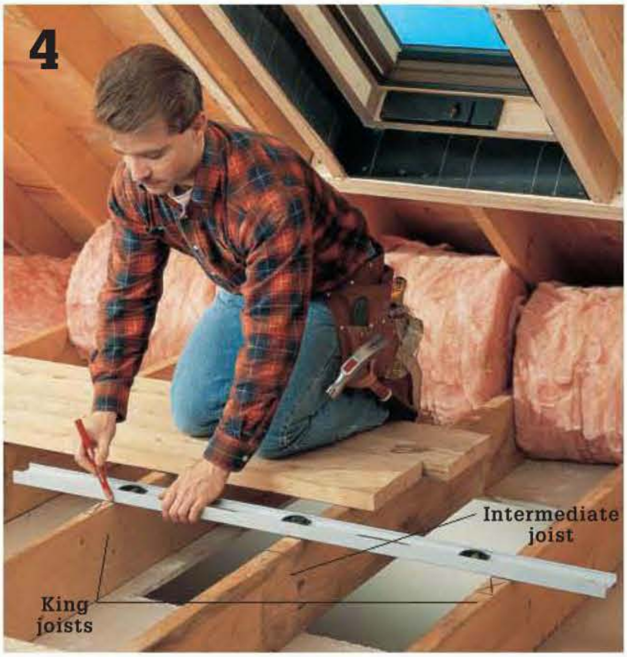

Use the nearest joists on either side of the ceiling opening to serve as king joists. Measure and mark where the double header and double sill will fit against the king joists and where the outside edge of the header and sill will cross any intermediate joists.

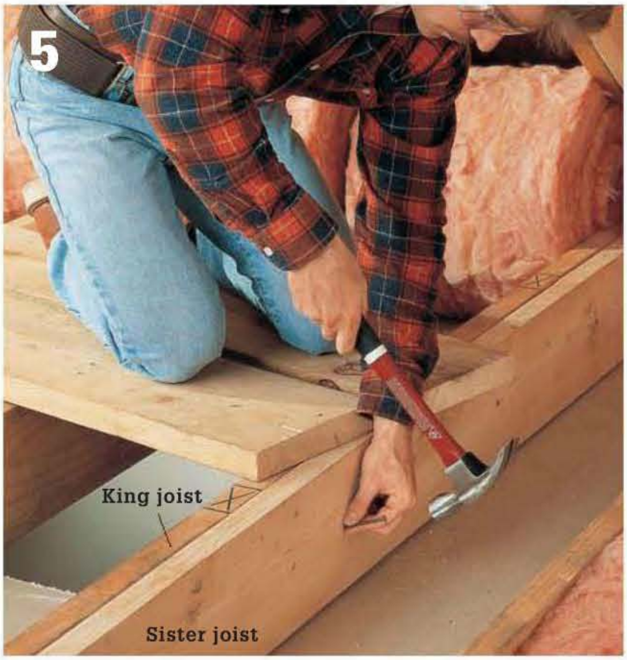

If you will be removing a section of an intermediate joist, reinforce the king joists by nailing full-length "sister" joists to the outside faces of the king joists using 10d nails.

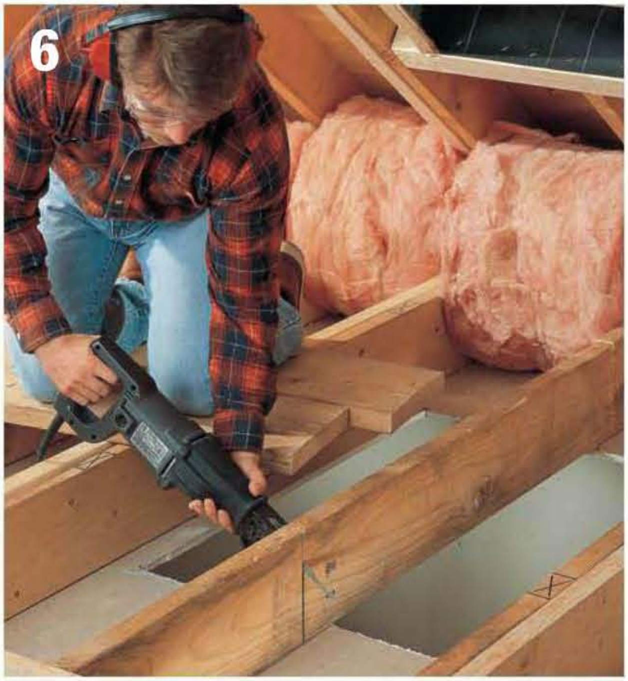

Install temporary supports below the project area to support the intermediate rafter on both sides of the opening. Use a combination square to extend cutting lines down the sides of the intermediate joist, then cut out the joist section with a reciprocating saw. Pry loose the cutout portion of the joist, being careful not to damage the ceiling surface.

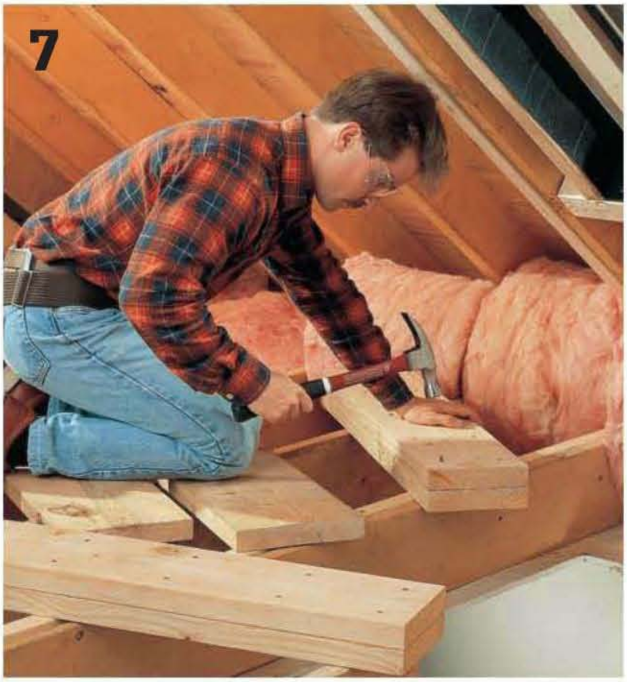

Build a double header and double sill to span the distance between the king joists using 2× dimensional lumber the same size as the joists.

(continued)

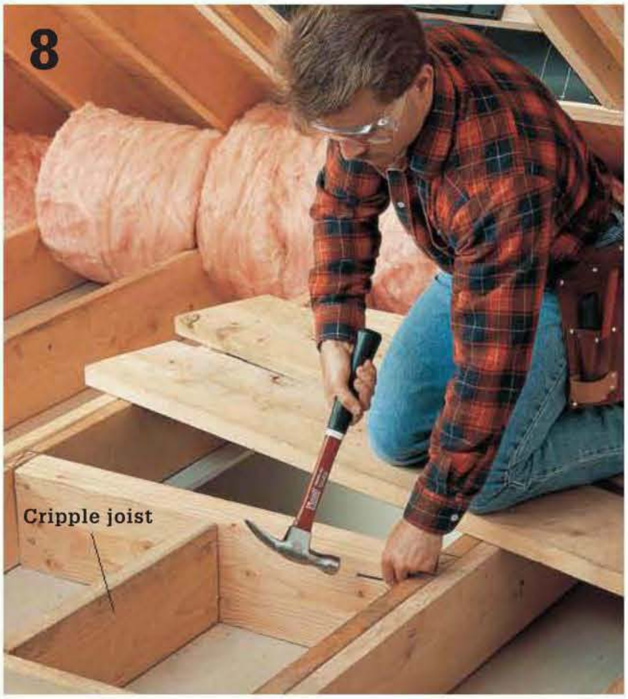
8
Cripple joist

Install the double header and double sill, anchoring them to the king joists and cripple joists with 10d nails. The inside edges of the header and sill should be aligned with the edge of the ceiling cutout.

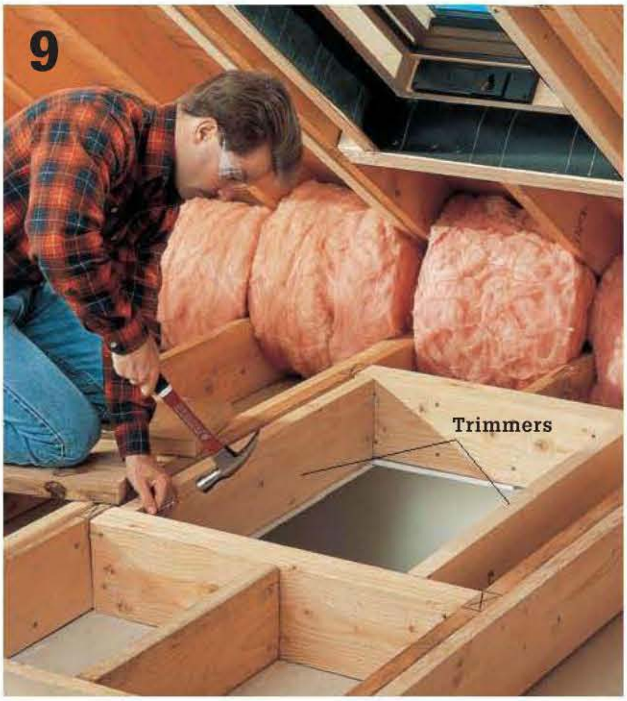

Complete the ceiling opening by cutting and attaching trimmers, if required, along the sides of the ceiling cutout between the header and sill. Toenail the trimmers to the header and sill with 10d nails.

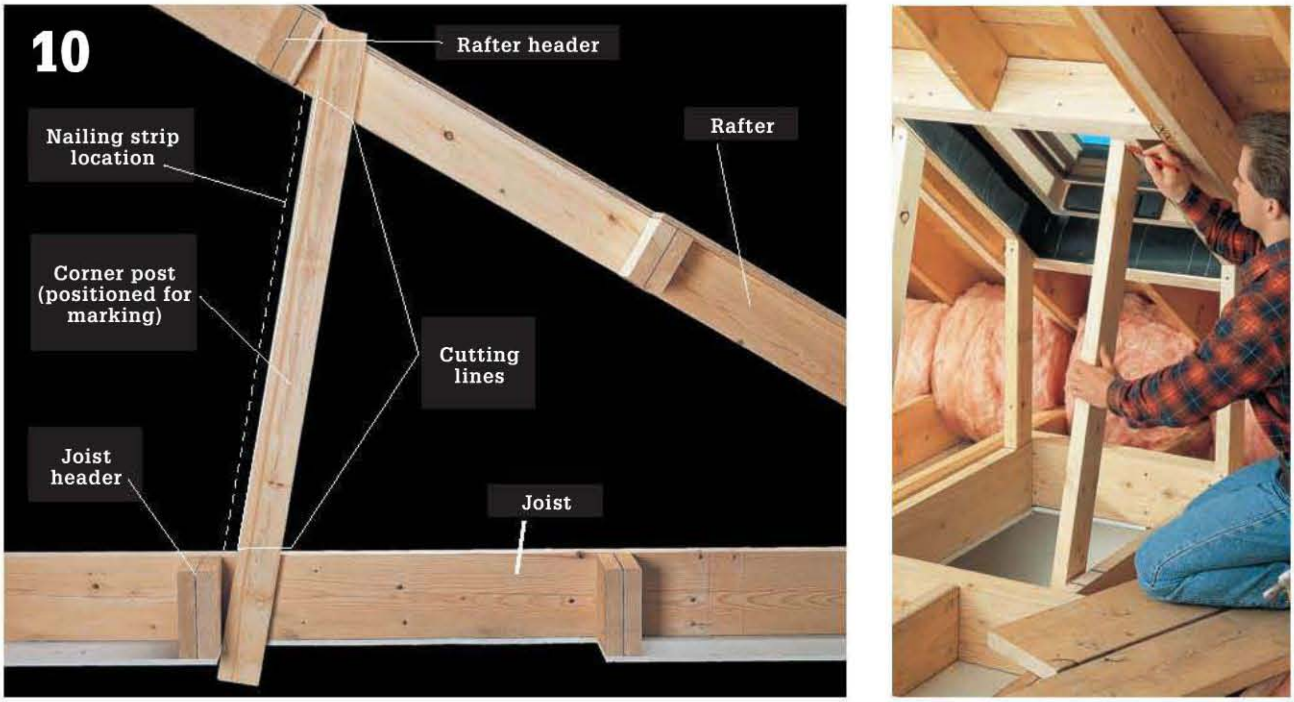
10
Rafter header
Nailing strip location
Rafter
Corner post (positioned for marking)
Cutting lines
Joist header
Joist

Install 2 × 4 corner posts for the skylight shaft. To measure for the posts, begin with a 2 × 4 that is long enough to reach from the top to the bottom of the shaft. Hold the 2 × 4 against the inside of the framed openings, so it is flush with the top of the rafter header and the bottom of the joist header (left photo). Mark cutting lines where the 2 × 4 meets the top of the joist or trimmer and the bottom of the rafter or trimmer (right photo). Cut along the lines, then toenail the posts to the top and bottom of the frame with 10d nails.

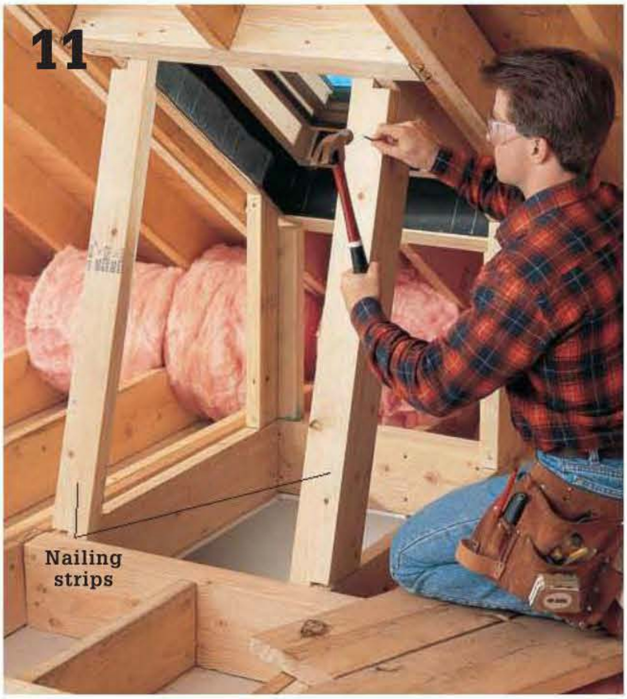
11
Nailing strips

Attach a 2 × 4 nailing strip to the outside edge of each corner post to provide a nailing surface for attaching the wallboard. Notch the ends of the nailing strips to fit around the trimmers; a perfect fit is not necessary.

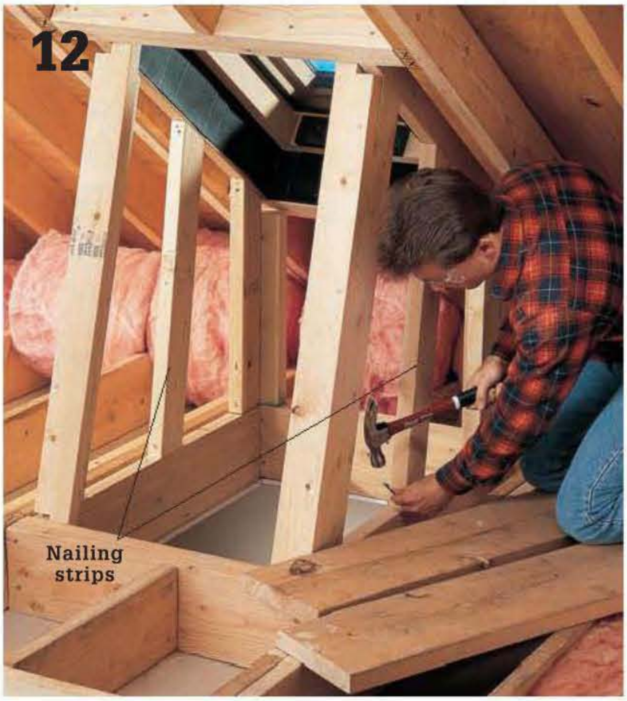
12
Nailing strips

Install additional 2 × 4 nailing strips between the corner posts if the distances between posts are more than 24". Miter the top ends of the nailing strips to fit against the rafter trimmers.

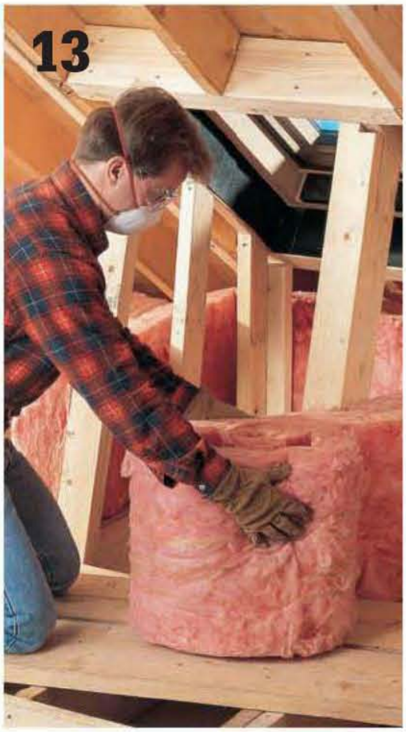

Wrap the skylight shaft with fiberglass insulation. Secure the insulation by wrapping twine around the shaft and insulation.

14
Insulation removed for clarity

From inside the shaft, staple a plastic vapor barrier of 6-mil polyethylene sheeting over the insulation.

15
Wallboard and insulation removed for clarity

Finish the inside of the shaft with wallboard (pages 48 and 51). *Tip: To reflect light, paint the shaft interior with a light-colored, semigloss paint.*

Installing Tubular Skylights

Any interior room can be brightened with a tubular skylight. Tubular skylights are quite energy-efficient and are relatively easy to install, with no complicated framing involved.

The design of tubular skylights varies among manufacturers, with some using solid plastic reflecting tubes and others using flexible tubing. Various diameters are also available. Measure the distance between the framing members in your attic before purchasing your skylight to be sure it will fit.

This project shows the installation of a tubular skylight on a sloped, asphalt-shingled roof. Consult the dealer or manufacturer for installation procedures on other roof types.

A tubular skylight is an economical way to introduce more sunlight into a room without embarking on a major framing project.

Tools & Materials ▸

Pencil
Drill
Tape measure
Wallboard saw
Reciprocating saw
Pry bar
Screwdriver
Hammer
Wire cutters
Utility knife
Chalk
Tubular skylight kit
Stiff wire
2" roofing nails or flashing screws
Roofing cement

How to Install a Tubular Skylight

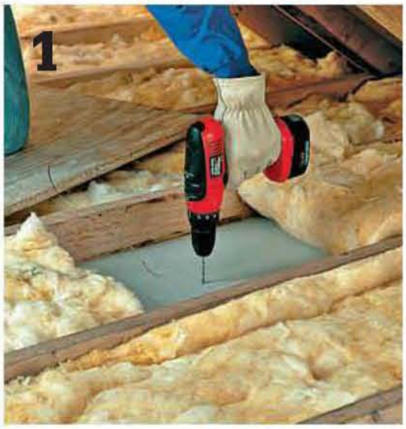

Drill a pilot hole through the ceiling at the approximate location for your skylight. Push a stiff wire up into the attic to help locate the hole. In the attic, make sure the space around the hole is clear of any insulation. Drill a second hole through the ceiling at the centerpoint between two joists.

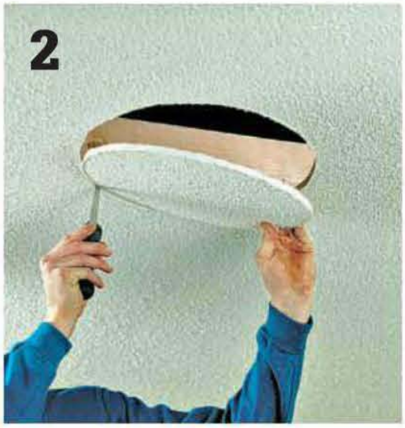

Center the ceiling ring frame over the hole and trace around it with a pencil. Carefully cut along the pencil line with a wallboard saw or reciprocating saw. Save the wallboard ceiling cutout to use as your roof-hole pattern. Attach the ceiling frame ring around the hole with the included screws.

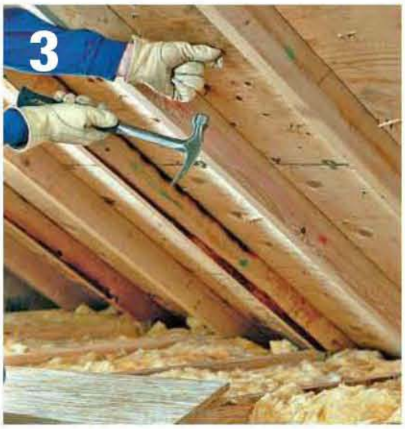

In the attic, choose the most direct route for the tubing to reach the roof. Find the center between the appropriate rafters and drive a nail up through the roof sheathing and shingles.

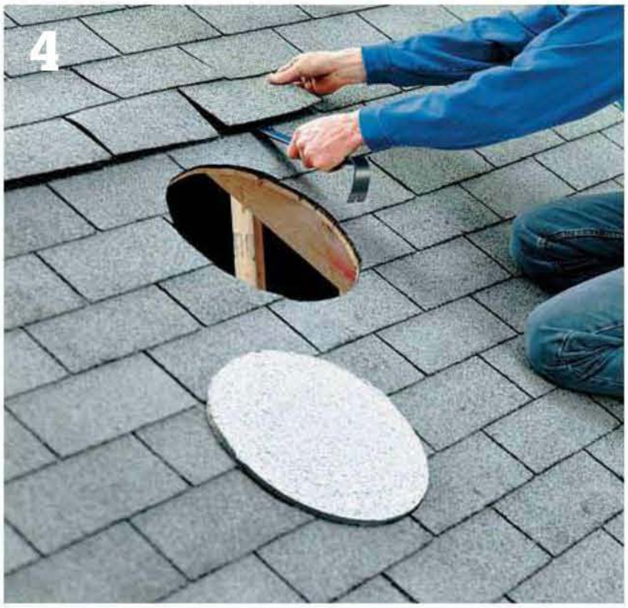

Use the wallboard ceiling cutout, centered over the nail hole, as a template for the roof opening. Trace the cutout onto the roof with chalk. Drill a starter hole to insert the reciprocating saw blade, then cut out the hole in the roof. Pry up the lower portion of the shingles above the hole. Remove any staples or nails around the hole edge.

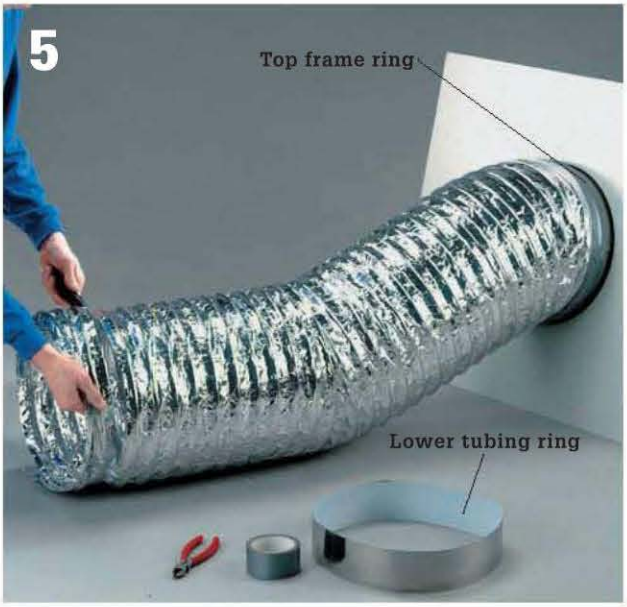

Pull the tubing over the top frame ring. Bend the frame tabs out through the tubing, keeping two or three rings of the tubing wire above the tabs. Wrap the junction three times around with included PVC tape. Then, in the attic, measure from the roof to the ceiling. Stretch out the tubing and cut it to length with a utility knife and wire cutters. Pull the loose end of tubing over the lower ring and wrap it three times with PVC tape.

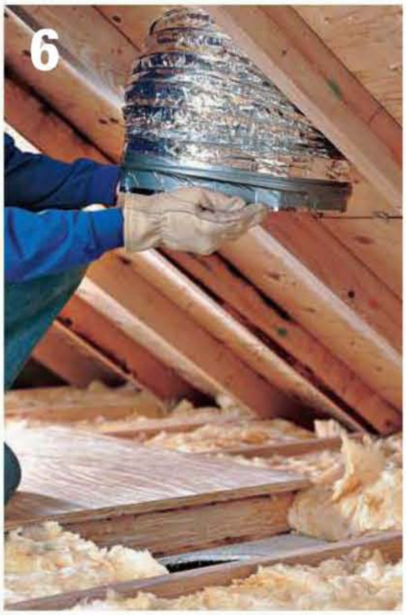

Lower the tubing through the roof hole and slide the flashing into place with the upper portion of the flashing underneath the existing shingles. This is easier with two people, one on the roof and one in the attic.

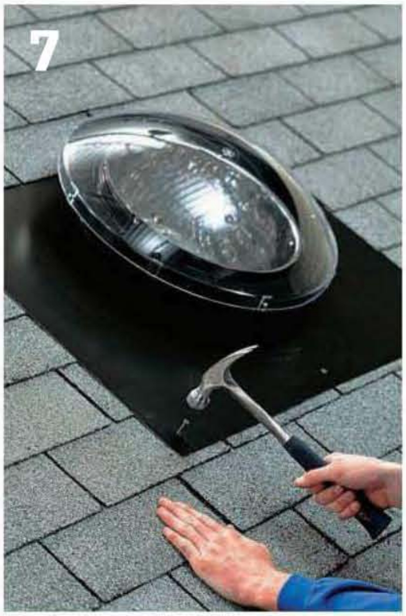

Secure the flashing to the roof with 2" roofing nails or flashing screws. Seal under the shingles and over all the nail heads with roofing cement. Attach the skylight dome and venting to the frame with the included screws.

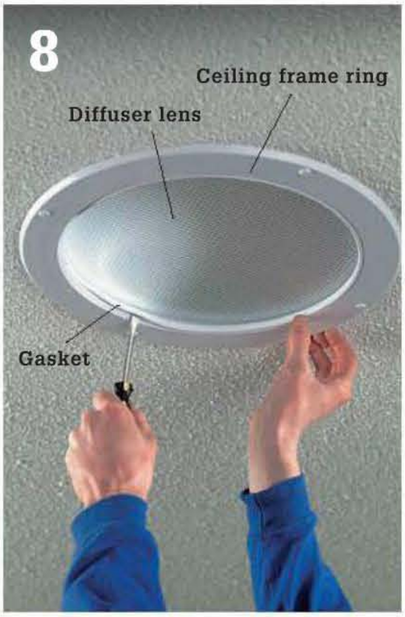

Pull the lower end of the tubing down through the ceiling hole. Attach the lower tubing ring to the ceiling frame ring and fasten it with screws. Attach the gasket to the diffuser lens and work the gasket around the perimeter of the ceiling frame. Repack any insulation around the tubing in the attic.

Installing Interior Doors

Creating an opening for a door in a wall involves building a framework about 1" wider and ½" taller than the door's jamb frame. This oversized opening, called a *rough opening,* will enable you to position the door easily and shim it plumb and level. Before framing a door, it's always a good idea to buy the door and refer to the manufacturer's recommendations for rough opening size.

Door frames consist of a pair of full-length king studs and two shorter jack studs that support the header above the door. A header provides an attachment point for wallboard and door casings. On load-bearing walls, it also helps to transfer the building's structural loads down into the wall framework and eventually the foundation.

Door framing requires flat, straight, and dry framing lumber, so choose your king, jack, and header pieces carefully. Sight down the edges and ends to look for warpage, and cut off the ends of pieces with splits.

Tools & Materials ▸

- Tape measure
- Framing square
- Hammer or nail gun
- Handsaw or reciprocating saw
- Framing lumber
- 10d or pneumatic framing nails
- ⅜" plywood (for structural headers)
- Construction adhesive

Creating a square, properly sized opening for a door is the most important element of a successful door installation project.

How to Frame a Rough Opening for an Interior Prehung Door

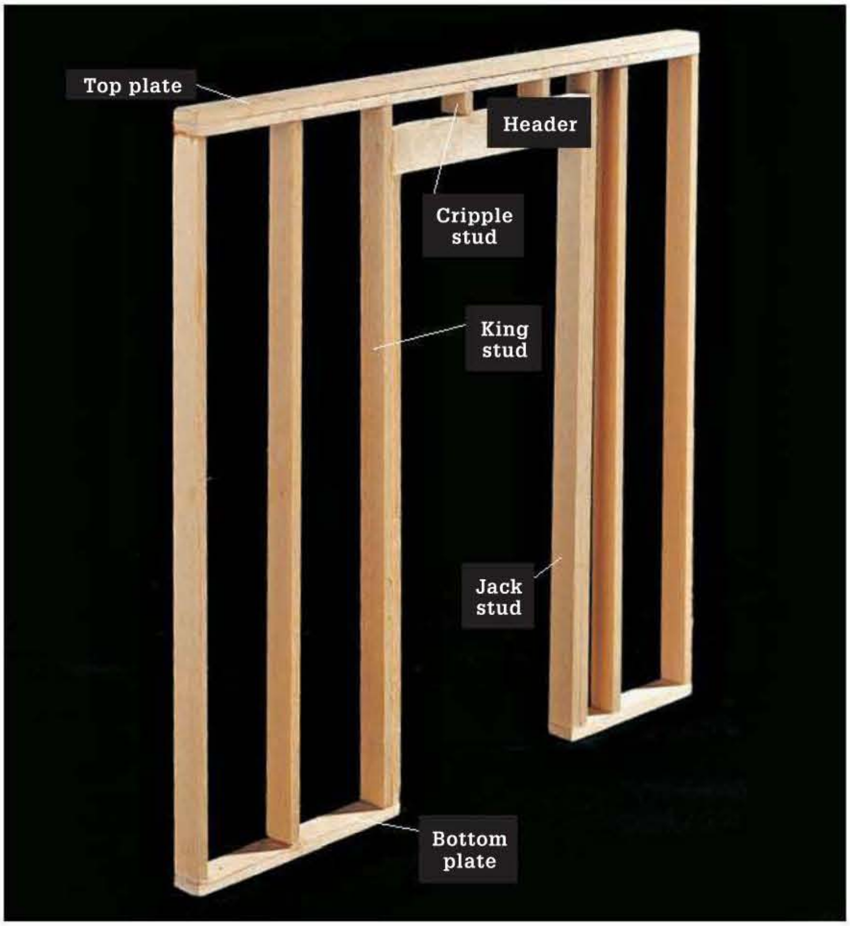
Top plate
Header
Cripple stud
King stud
Jack stud
Bottom plate

Door frames for prehung doors (left) start with king studs that attach to the top and bottom plates. Inside the king studs, jack studs support the header at the top of the opening. Cripple studs continue the wall-stud layout above the opening. In non-loadbearing walls, the header may be a 2 × 4 laid flat or a built-up header (below). The dimensions of the framed opening are referred to as the rough opening.

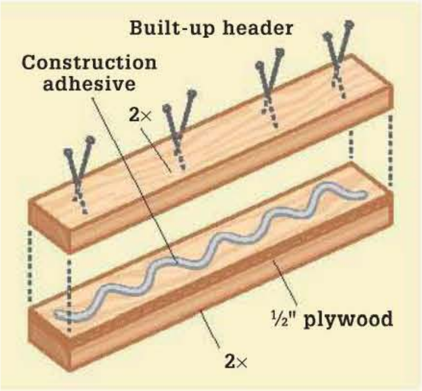
Built-up header
Construction adhesive
2×
½" plywood
2×

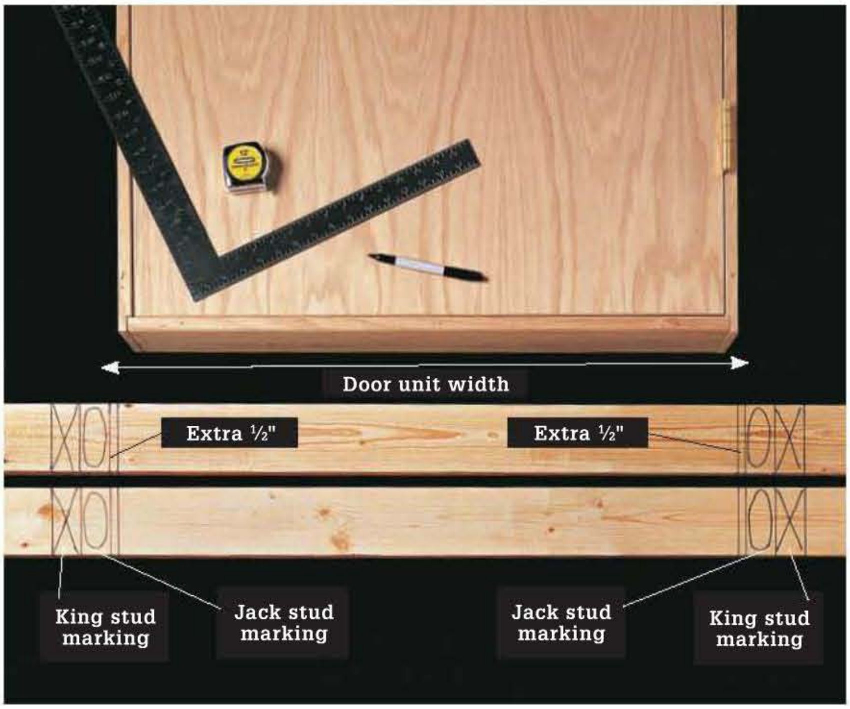
Door unit width
Extra ½"
Extra ½"
King stud marking
Jack stud marking
Jack stud marking
King stud marking

To mark the layout for the door frame, measure the width of the door unit along the bottom. Add 1" to this dimension to determine the width of the rough opening (the distance between the jack studs). This gives you a ½" gap on each side for adjusting the door frame during installation. Mark the top and bottom plates for the jack and king studs.

How to Frame a Prehung Door Opening

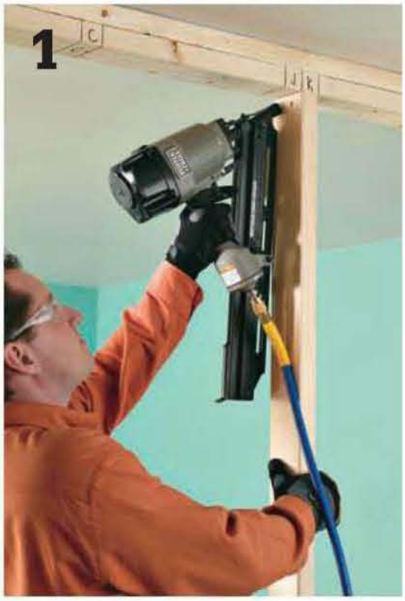

Mark layout lines for the king and jack studs on the wall's top and sole plates. Cut the king studs slightly longer than the distance between the wall plates, and toenail them in place with 10d nails or 3" pneumatic nails.

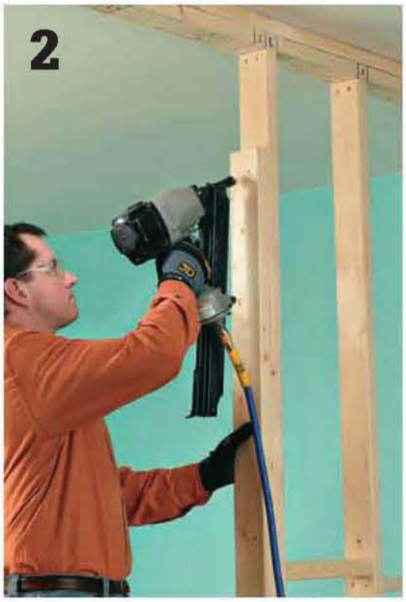

Cut the jack studs to length (they should rest on the sole plate). The height of a jack stud for a standard interior door is 83½", or ½" taller than the door. Nail the jack studs to the king studs.

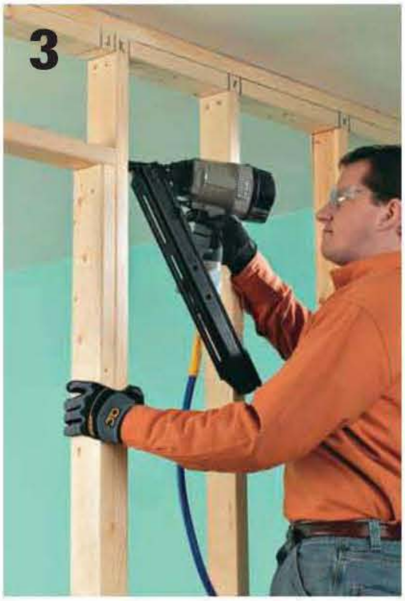

In a partition wall, the header can be a piece of 2× framing lumber that lays flat on top of the jack studs. Cut it to length, and install by endnailing through the king studs or down into the jack studs.

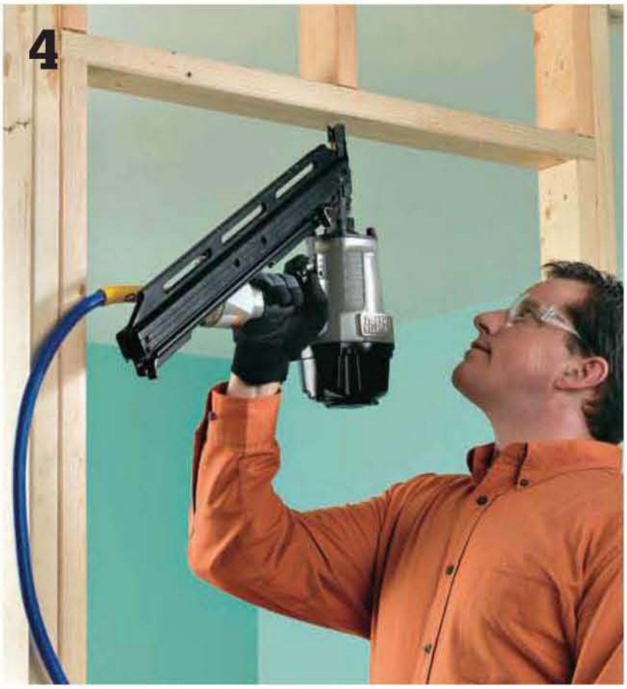

Fasten a cripple stud above the header, halfway between the king studs. It will prevent the header from warping. Toenail it into the top plate, and drive nails into it through the header.

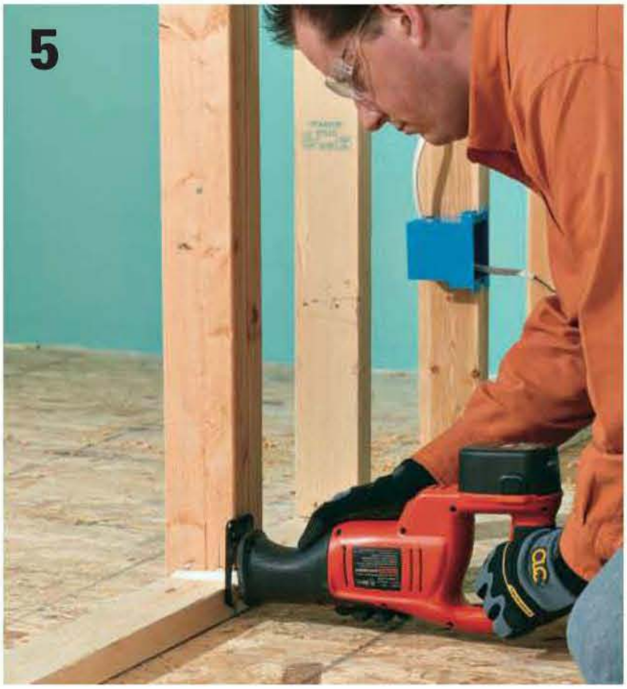

If you haven't cut a sole plate opening for the door yet, do that now with a reciprocating saw or handsaw. Trim the sole plate flush with the jack studs.

How to Frame an Opening for a Load-bearing Wall

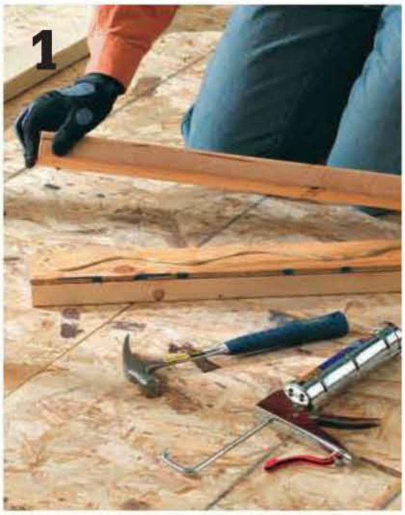

Door framing on load-bearing walls will require a structural header that transfers loads above the wall into the jack studs, sole plate, and down into the house foundation. Build it by sandwiching a piece of $^{3}/_{8}$" plywood between two 2 × 4s. Use construction adhesive and nails to fasten the header together.

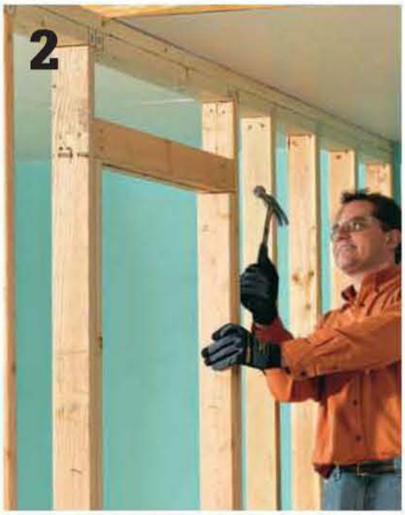

Install the built-up header by resting it on the jack studs and endnailing through the king studs. Use 10d nails or 3" pneumatic nails.

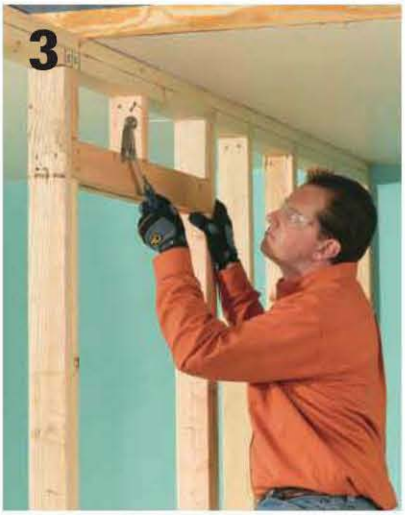

Toenail a cripple stud between the top plate and header, halfway between the king studs. It transfers structural loads into the header.

Option: Framing Openings for Sliding & Folding Doors

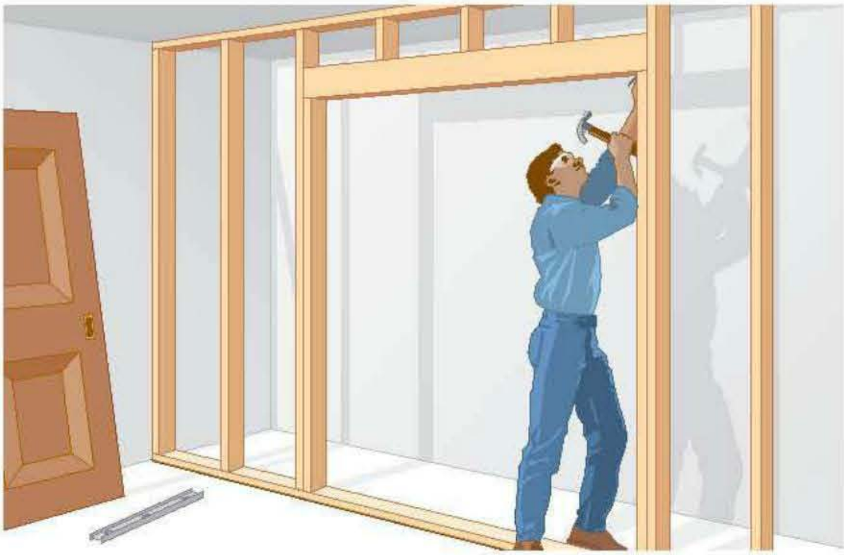

The same basic framing techniques are used, whether you're planning to install a sliding, bifold, pocket, or prehung interior door. The different door styles require different frame openings. You may need to frame an opening 2 to 3 times wider than the opening for a standard prehung door. Purchase the doors and hardware in advance, and consult the hardware manufacturer's instructions for the exact dimensions of the rough opening and header size for the type of door you select.

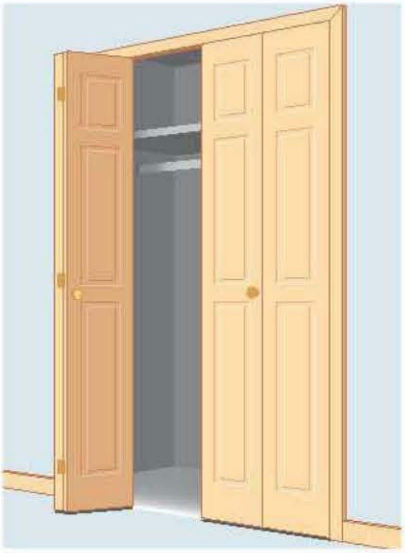

Most bifold doors are designed to fit in a 80"-high finished opening. Wood bifold doors have the advantage of allowing you to trim the doors, if necessary, to fit openings that are slightly shorter.

Installing a Prehung Interior Door

Install prehung interior doors after the framing work is complete and the wallboard has been installed. If the rough opening for the door has been framed accurately, installing the door takes about an hour.

Standard prehung doors have 4½"-wide jambs and are sized to fit walls with 2 × 4 construction and ½" wallboard. If you have 2 × 6 construction or thicker wall surface material, you can special-order a door to match, or you can add jamb extensions to a standard-sized door (photo, below).

Tools & Materials ▸

Level
Hammer
Handsaw
Prehung interior door
Wood shims
8d casing nails

Tip ▸

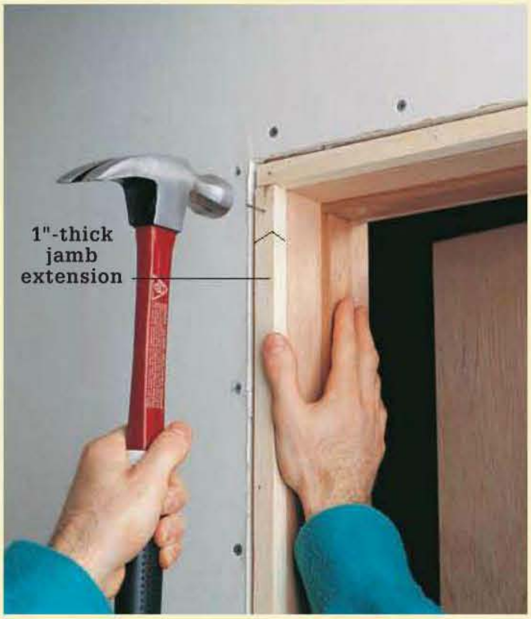

If your walls are built with 2 × 6 studs, you'll need to extend the jambs by attaching 1"-thick wood strips to the edges of the jamb after the door is installed. Use glue and 4d casing nails when attaching jamb extensions. Make the strips from the same wood as the jamb.

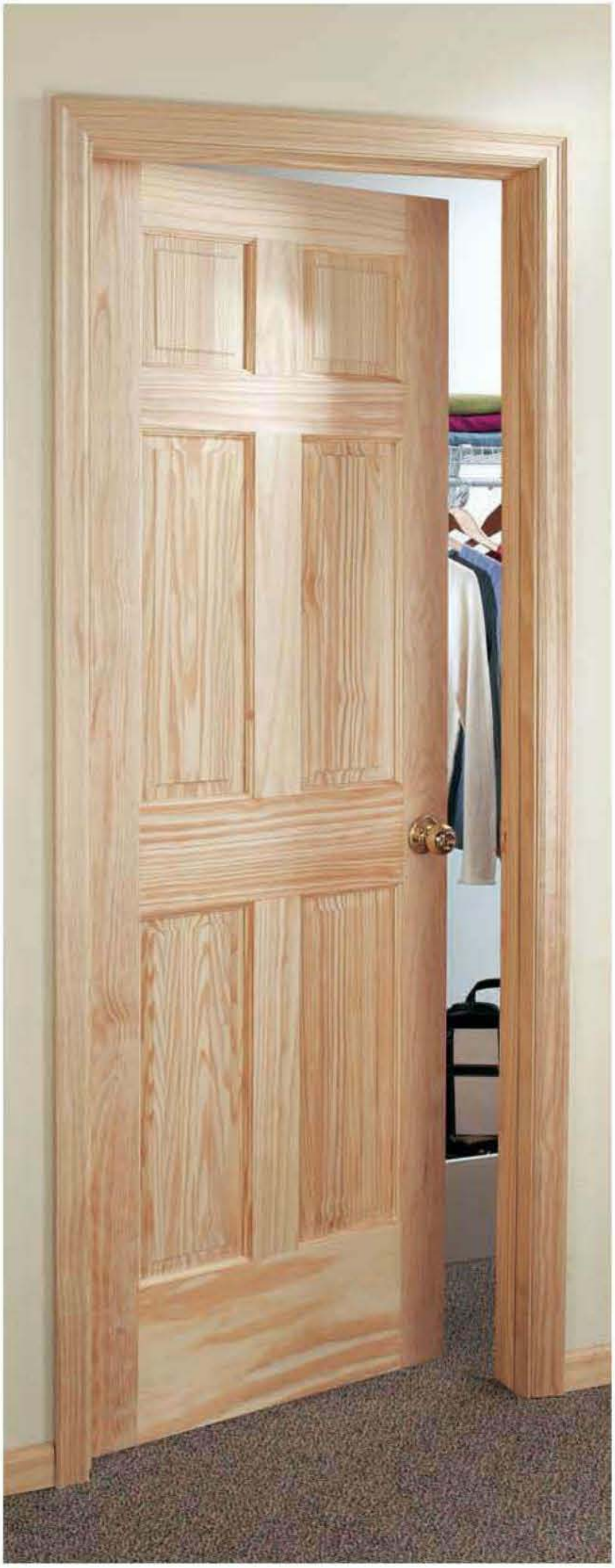

Prehung doors are shipped as single units with the door already hung on hinges attached to pre-installed jambs.

How to Install a Prehung Interior Door

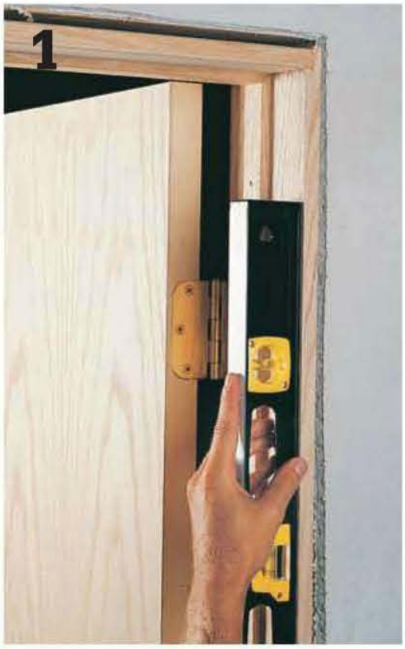

Slide the door unit into the framed opening so the edges of the jambs are flush with the wall surface and the hinge-side jamb is plumb.

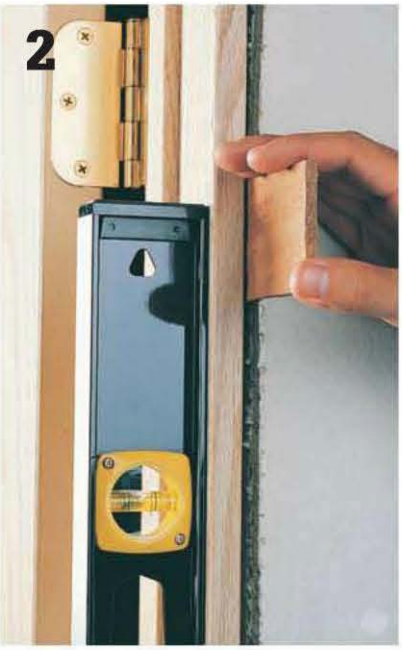

Insert pairs of wood shims driven from opposite directions into the gap between the framing members and the hinge-side jamb, spaced every 12". Check the hinge-side jamb to make sure it is still plumb and does not bow.

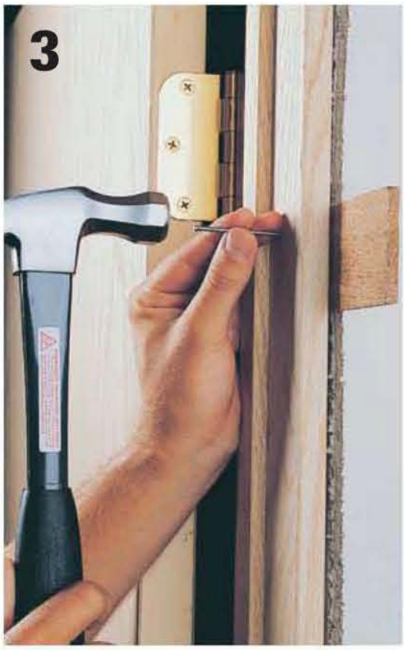

Anchor the hinge-side jamb with 8d casing nails driven through the jamb and shims and into the jack stud.

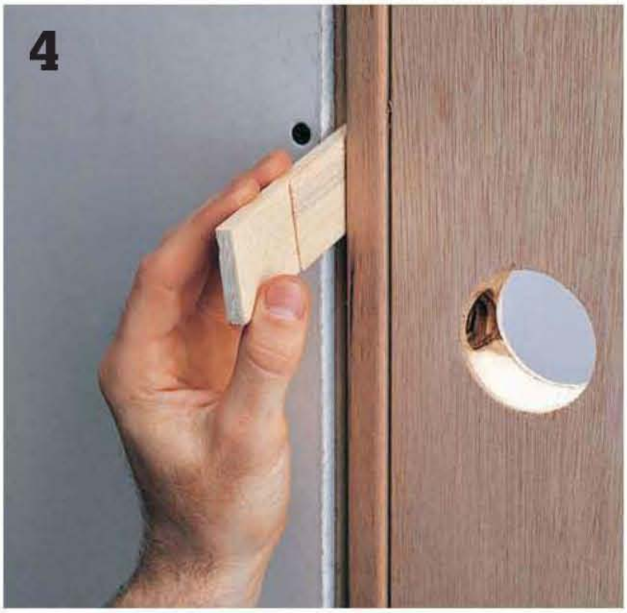

Insert pairs of shims in the gap between the framing members and the latch-side jamb and top jamb, spaced every 12". With the door closed, adjust the shims so the gap between door edge and jamb is ⅛" wide. Drive 8d casing nails through the jambs and shims, into the framing members.

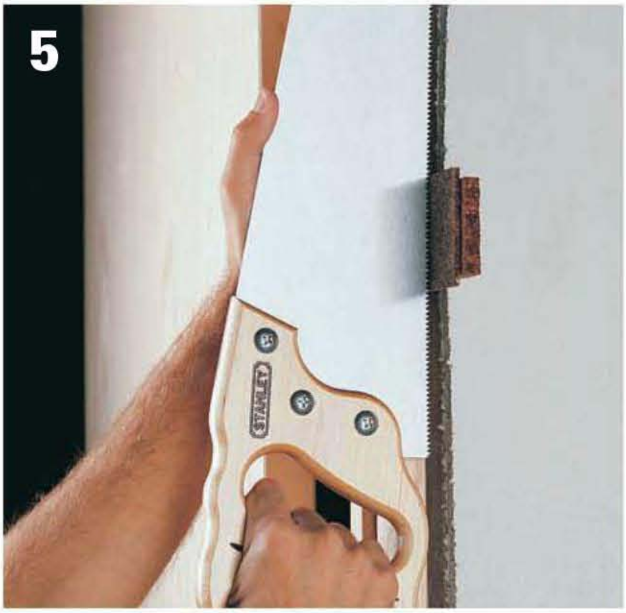

Cut the shims flush with the wall surface, using a handsaw. Hold the saw vertically to prevent damage to the door jamb or wall. Finish the door and install the lockset as directed by the manufacturer. See pages 318 to 321 to install trim around the door.

Pocket Doors

Pocket doors are a space-saving alternative to traditional hinged interior doors. Swinging doors can monopolize up to 16 square feet of floor space in a room, which is why pocket doors are a perfect choice for tight spaces, like small bathrooms. Installed in pairs, pocket doors can divide large rooms into more intimate spaces and can still be opened to use the entire area.

Pocket door hardware kits generally are universal and can be adapted for almost any interior door. In this project, the frame kit includes an adjustable track, steel-clad split studs, and all the required hanging hardware. The latch hardware, jambs, and the door itself are all sold separately. Pocket door frames can also be purchased as preassembled units that can be easily installed into a rough opening.

Framing and installing a pocket door is not difficult in new construction or a major remodel. But retrofitting a pocket door in place of a standard door or installing one in a wall without an existing door, is a major project that involves removing the wall material, framing the new opening, installing and hanging the door, and refinishing the wall. Hidden utilities, such as wiring, plumbing, and heating ducts, must be rerouted if encountered.

The rough opening for a pocket door is at least twice the width of a standard door opening. If you are installing the pocket door in a partition wall, see page 292 to learn how to frame the opening. If the wall is load bearing, you will need to install an appropriately sized header (page 293).

Because pocket doors are easy to open and close and require no threshold, they offer increased accessibility for wheelchair or walker users, provided the handles are easy to use (page 299). If you are installing a pocket door for this purpose, be aware that standard latch hardware may be difficult to use for some individuals.

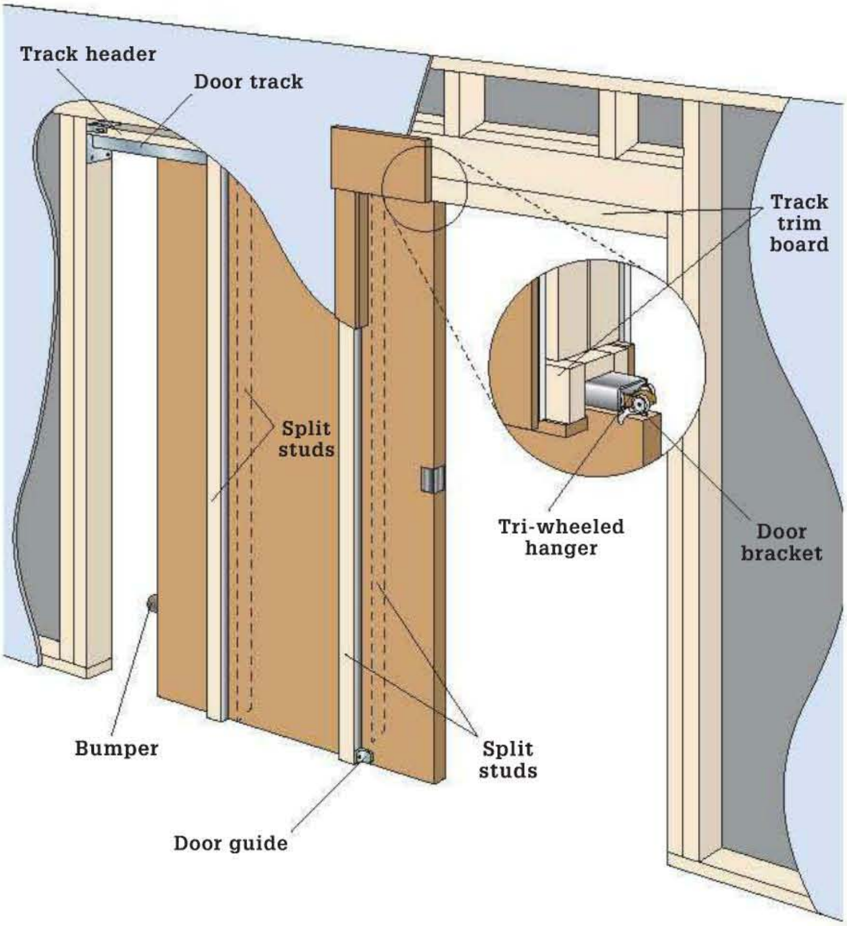

Tools & Materials ▸

- Tape measure
- Circular saw
- Hammer, nail set
- Screwdriver
- Level
- Drill
- Handsaw
- Hacksaw
- Wallboard tools
- 2 × 4 lumber
- 16d, 8d & 6d common nails
- Pocket door frame kit
- Door
- 1¼" wallboard screws
- Wallboard materials
- Manufactured pocket door jambs (or build jambs from 1× material)
- 8d & 6d finish nails
- 1½" wood screws
- Door casing

How to Install a Pocket Door

Prepare the project area and frame the rough opening to the manufacturer's recommended dimensions. Measuring from the floor, mark each jack stud at the height of the door plus ¾ to 1½" (depending on the door clearance above the floor) for the overhead door track. Drive a nail into each jack stud, centered on the mark. Leave about ⅛" of the nail protruding.

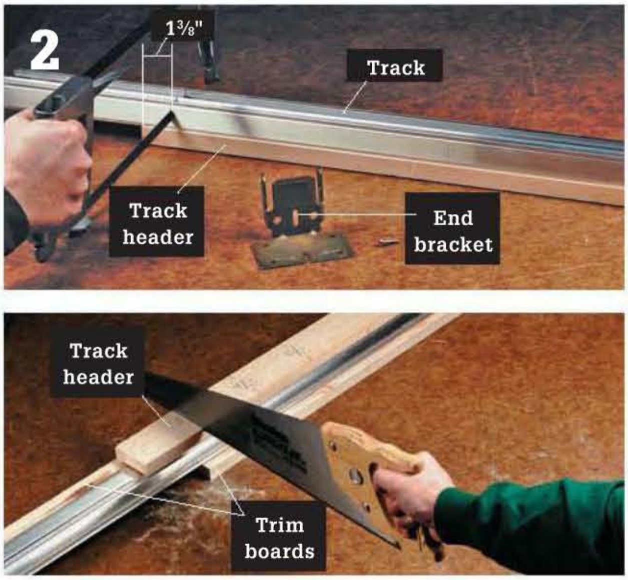

Remove the adjustable end bracket from the overhead door track. Cut the wooden track header at the mark that matches your door size. Turn the track over and cut the metal track 1⅜" shorter than the wooden track header using a hacksaw (top). Replace the end bracket. Cut the side trim boards along the marks corresponding to your door size, being careful not to cut the metal track (bottom).

Set end brackets of the track on the nails in the jack studs. Adjust the track to level and set the nails. Then drive 8d common nails through the remaining holes in the end brackets.

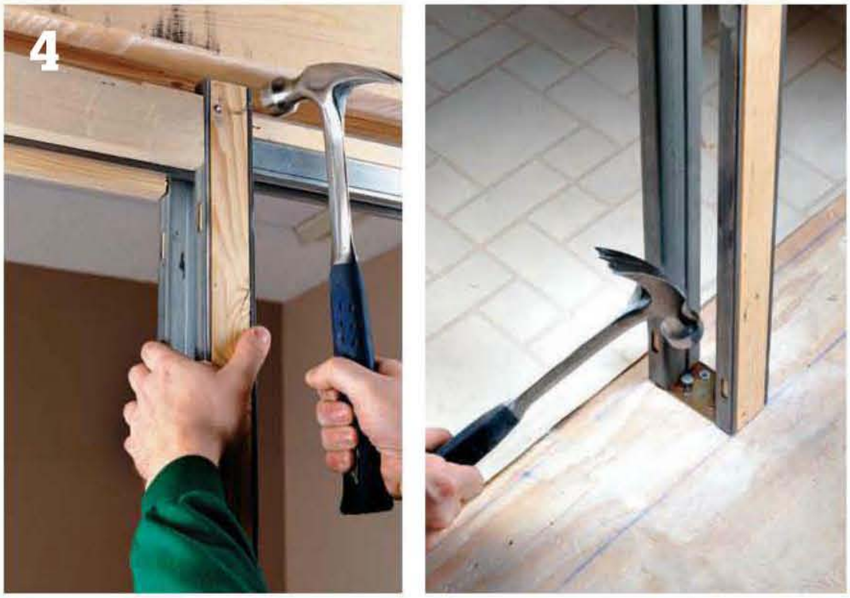

Snap chalk lines on the floor across the opening, even with the sides of the rough opening. Tap the floor plate spacers into the bottom ends of the pairs of steel-clad split studs. Butt one split stud pair against the door track trim board, check it for plumb, and fasten it to the track header using 6d common nails (left). Center the other split stud pair in the "pocket" and fasten it to the track header. Plumb the split studs again and attach them to the floor with 8d common nails or 2" screws driven through spacer plates (right).

(continued)

Cover the open framing with wallboard to the edge of the opening. You may want to leave the wallboard off one side of the wall to allow for door adjustment. Use 1¼" wallboard screws, which will not protrude into the pocket.

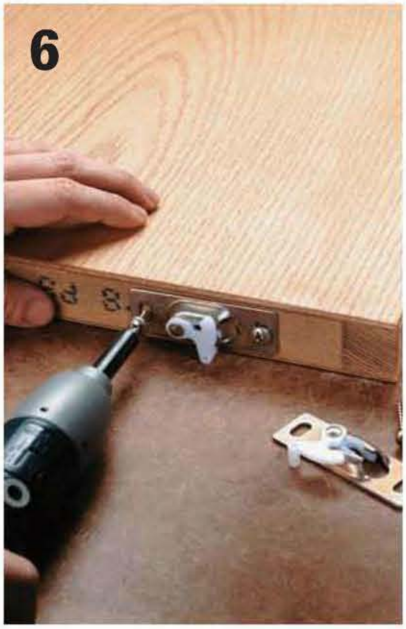

Paint or stain the door as desired. When the door has dried, attach two door brackets to the top of the door, using included screws driven through pilot holes. Install the rubber bumper to the rear edge of the door with its included screw.

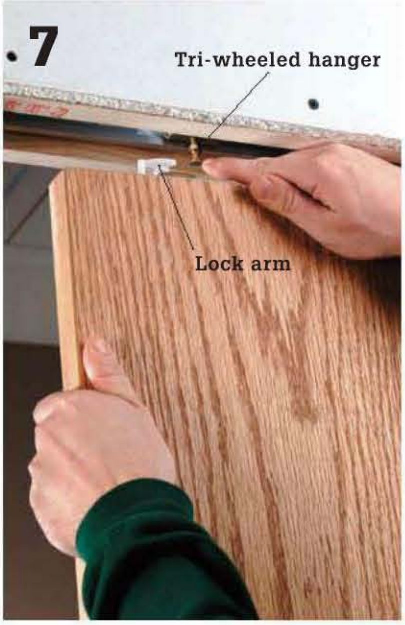

Slide two tri-wheeled hangers into the overhead door track. Set the door in the frame, aligning the hangers with the door brackets. Then raise the door and press each hanger into the door bracket until it snaps into place. Close the lock arm over the hanger.

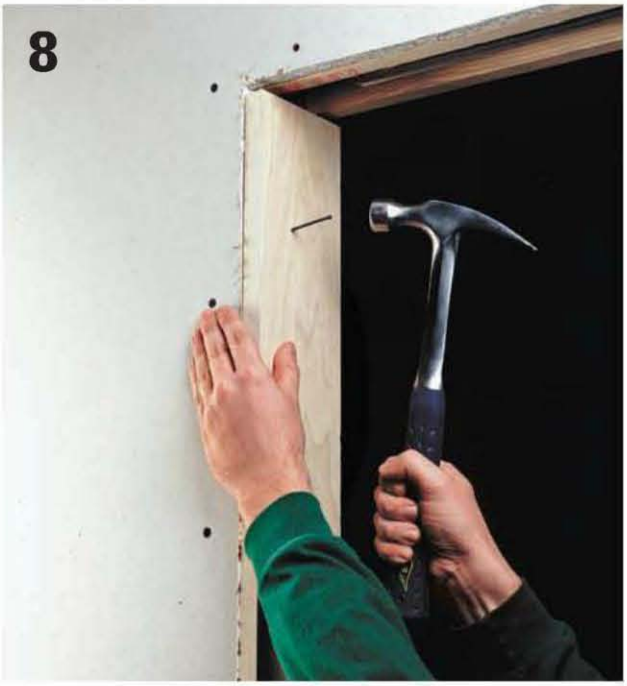

Cut the strike-side jamb to length and width. Fasten it to the jack stud using 8d finish nails, shimming the jamb to plumb as necessary. Close the door and adjust the hanger nuts to fine-tune the door height so the door is parallel with the jamb from top to bottom.

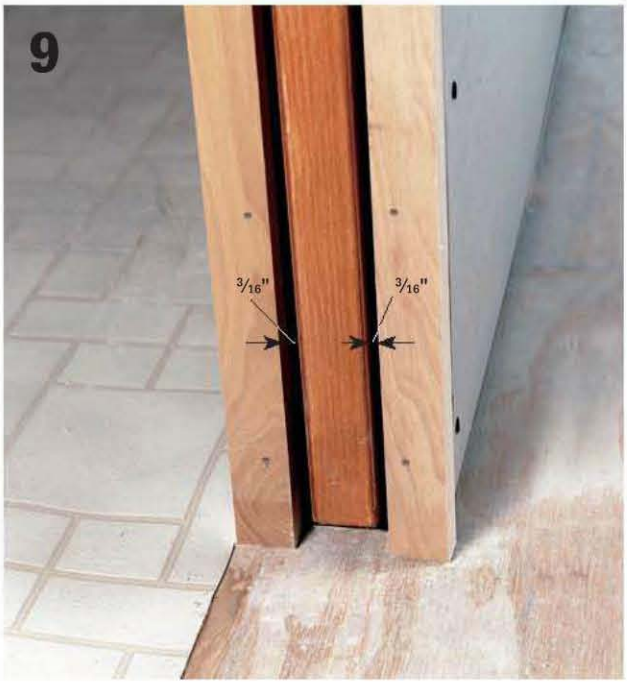

Measure and cut the split jambs to size. Fasten each split jamb to the front edge of the split stud using 8d finish nails. Maintain 3⁄16" clearance on both sides of the door. If necessary, shim between the bumper and door until the door is flush with the jambs when open.

Measure and cut the split head jambs to size. Use 1½" wood screws driven through countersunk pilot holes to attach the head jamb on the side that has access to the lock arm of the hangers to allow for easy removal of the door. Attach the other head jamb using 6d finish nails. Maintain 3⁄16" clearance on each side of the door.

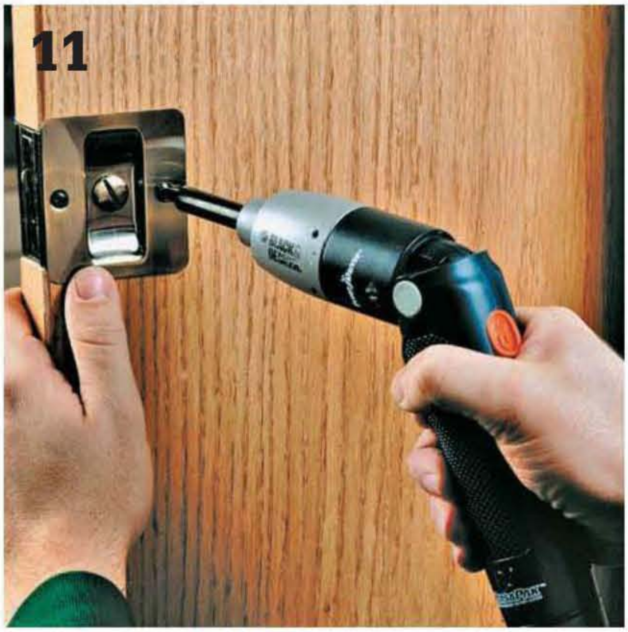

Install the included door guides on both sides of the door near the floor at the mouth of the pocket. Install the latch hardware according to the manufacturer's directions. Finish the wallboard and install casing around the door. Fill all nail holes, then paint or stain the jambs and casing as desired.

Improving Pocket Door Accessibility ▸

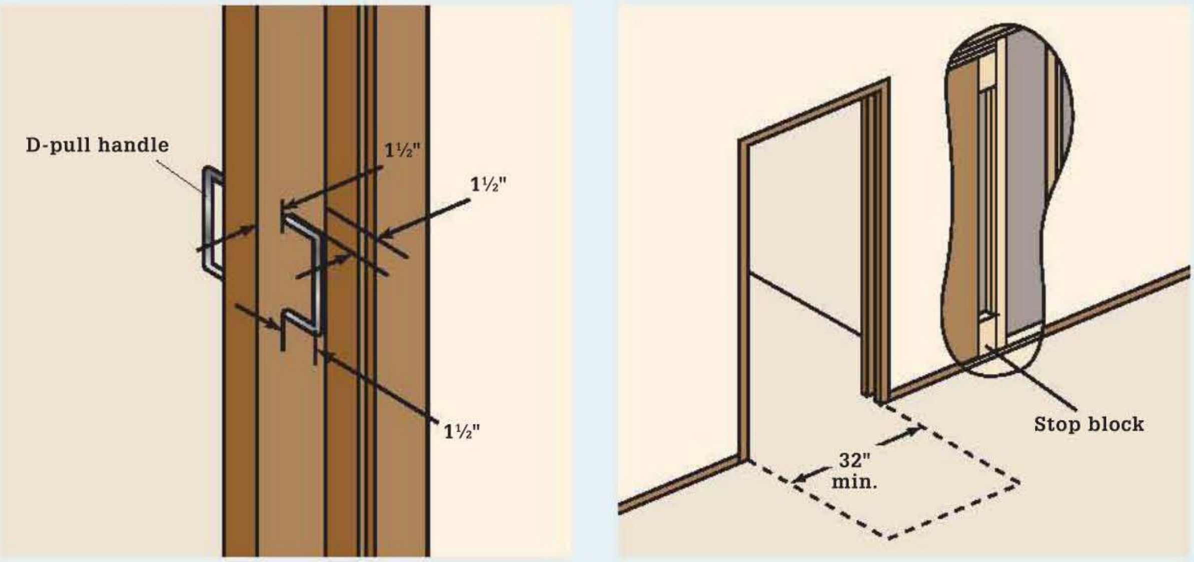

D-pull handles are easier to use than standard recessed hardware. Choose pulls that project at least 1½" from the door. Mount the pulls 1½" from the edge of the door to provide room for fingers when the door is closed (left). Install a stop block at the back of the frame (right), so the door stops 1½" short of the D-pull to provide room for fingers when the door is open. Because this design reduces the width of the door opening by 3", you must use a 36"-wide door to maintain the recommended doorway width of 32".

Bifold Doors

Bifold doors provide easy access to a closet without requiring much clearance for opening. Most home centers stock kits that include two pairs of prehinged doors, a head track, and all the necessary hardware and fasteners. Typically, the doors in these kits have predrilled holes for the pivot and guide posts. Hardware kits are also sold separately for custom projects. There are many types of bifold door styles, so be sure to read and follow the manufacturer's instructions for the product you use.

Tools & Materials ▸

Tape measure
Level
Circular saw
Straightedge (optional)
Drill
Plane
Screwdriver
Hacksaw
Prehinged bifold doors
Head track
Mounting hardware
Panhead screws
Flathead screws

A variety of designer bifold doors are available for installation between rooms and closets. They provide the same attractive appearance as French doors but require much less floor space.

How to Install Bifold Doors

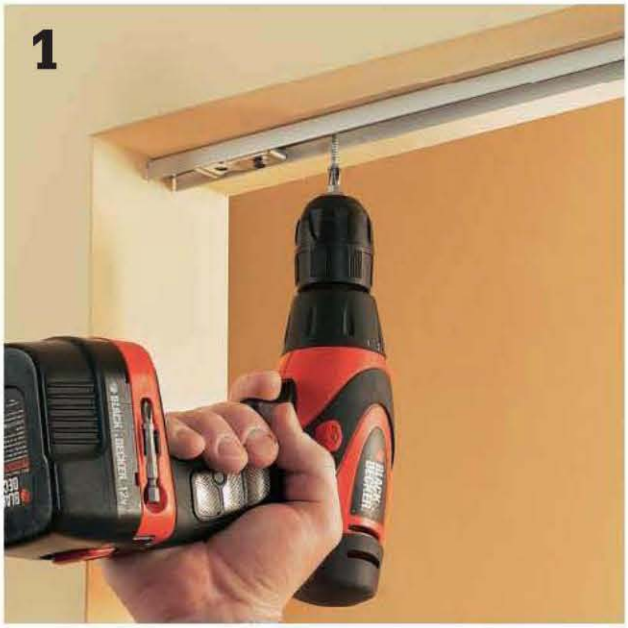

Cut the head track to the width of the opening using a hacksaw. Insert the roller mounts into the track, then position the track in the opening. Fasten it to the header using panhead screws.

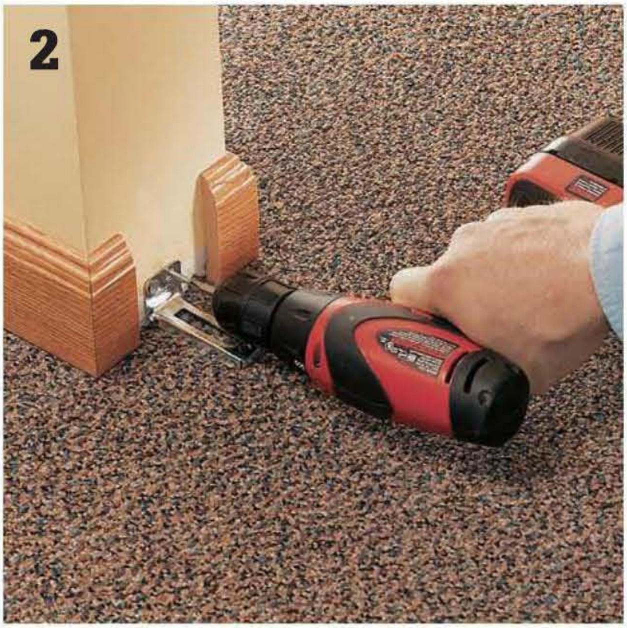

Measure and mark each side jamb at the floor for the anchor bracket so the center of the bracket aligns exactly with the center of the head track. Fasten the brackets in place with flathead screws.

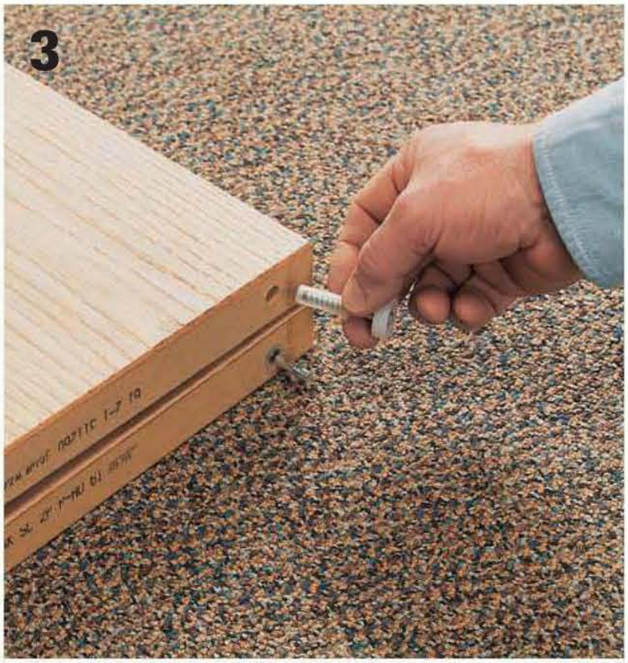

Check the height of the doors in the opening, and trim if necessary. Insert pivot posts into predrilled holes at the bottoms and tops of the doors. Insert guide posts at the tops of the leading doors. Make sure all posts fit snugly.

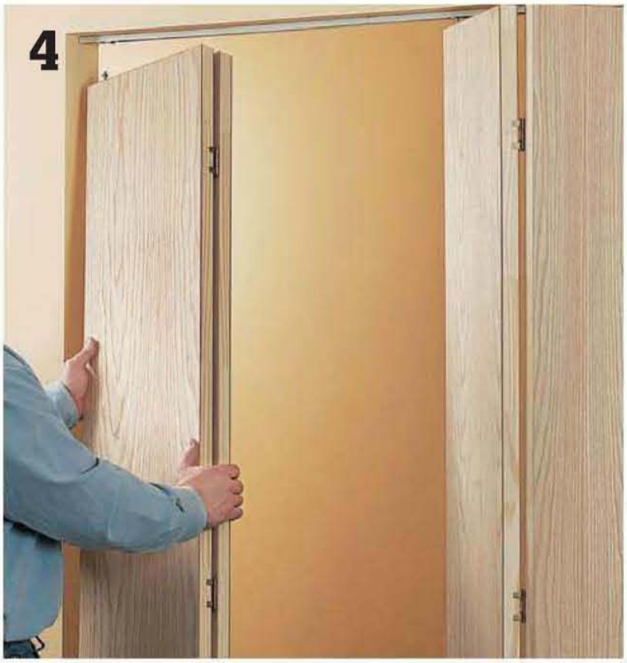

Fold one pair of doors closed and lift into position, inserting the pivot and guide posts into the head track. Slip the bottom pivot post into the anchor bracket. Repeat for the other pair of doors. Close the doors and check alignment along the side jambs and down the center. If necessary, adjust the top and bottom pivots following the manufacturer's instructions.

French Doors

French doors are made up of two separate doors, hinged on opposing jambs of a doorway. The doors swing out from the center of the doorway and into or out from a room. Like most doors, French doors are typically sold in prehung units, but are also available separately. They are generally available only in wood with a variety of designs and styles to choose from.

Before purchasing a prehung French door unit, determine the size of doors you will need. If you are planning to install the doors in an existing doorway, measure the dimensions of the rough opening from the unfinished framing members, then order the unit to size—the manufacturer or distributor will help you select the proper unit.

You can also pick the prehung unit first, then alter an existing opening to accommodate it (as shown in this project). In this case, build the rough opening a little larger than the actual dimensions of the doors to accommodate the jambs. Prehung units typically require adding 1" to the width and ½" to the height.

If the doorway will be in a load-bearing wall, you will need to make temporary supports (page 29) and install an appropriately sized header. Sizing the header (depth) is critical: it's based on the length of the header, the material it's made from, and the weight of the load it must support. For actual requirements, consult your local building department.

When installing French doors, it is important to have consistent reveals between the two doors and between the top of the doors and the head jamb. This allows the doors to close properly and prevents the hinges from binding.

Tools & Materials ▸

- Tape measure
- Circular saw
- 4-ft. level
- Hammer
- Handsaw
- Drill
- Utility knife
- Nail set
- Prehung French door unit
- 2 × 4 and 2 × 6 lumber
- ½" plywood
- 10d & 16d common nails
- Wood shims
- 8d finish nails

Traditionally, French doors open onto the patio or lush garden of a backyard. But you can create stylish entrances inside your home by bringing French doors to formal dining rooms, sitting rooms, dens, and master suites.

How to Install French Doors

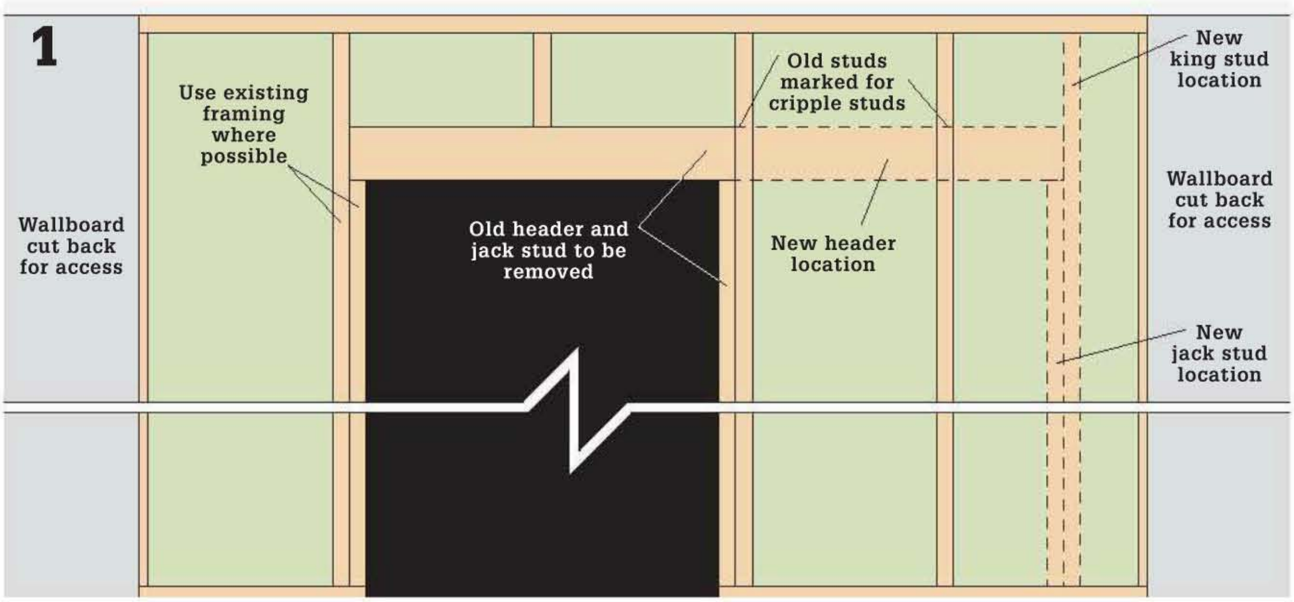
1
Use existing framing where possible
Wallboard cut back for access
Old header and jack stud to be removed
Old studs marked for cripple studs
New header location
New king stud location
Wallboard cut back for access
New jack stud location

Shut off power and water to the area. Remove the wall surfaces from both sides of the wall (pages 276 to 279), leaving one stud bay open on each side of the new rough opening. Also remove or reroute any wiring, plumbing, or ductwork. Lay out the new rough opening, marking the locations of all new jack and king studs on both the top and bottom plates. Where practical, use existing framing members. To install a new king stud, cut a stud to size and align with the layout marks; toenail to the bottom plate with 10d common nails, check for plumb, then toenail to the top plate to secure. Finally, mark both the bottom and top of the new header on one king stud, then use a level to extend the lines across the intermediate studs to the opposite king stud. If using existing framing, measure and mark from the existing jack stud.

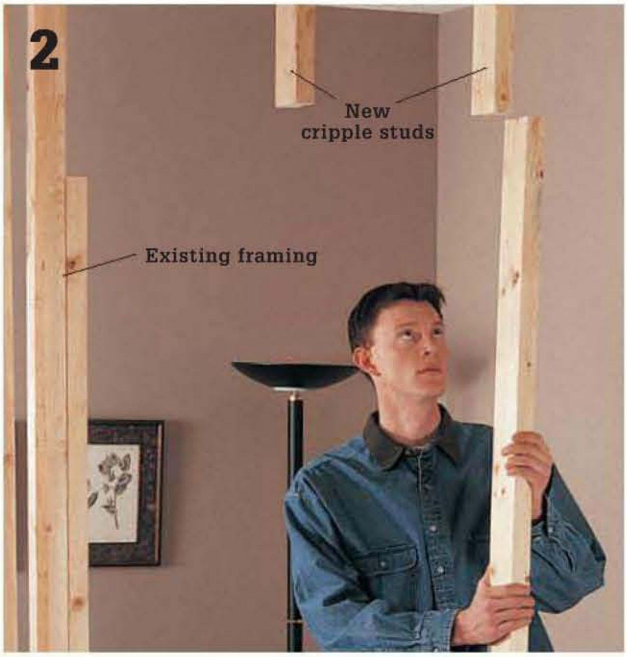
2
New cripple studs
Existing framing

Cut the intermediate studs at the reference marks for the top of the header using a reciprocating saw. Pry the studs away from the sole plates and remove—the remaining top pieces will be used as cripple studs.

To install a jack stud, cut the stud to fit between the sole plate and the bottom of the header, as marked on the king stud. Align it at the mark against the king stud, then fasten it in place with 10d common nails driven every 12".

(continued)

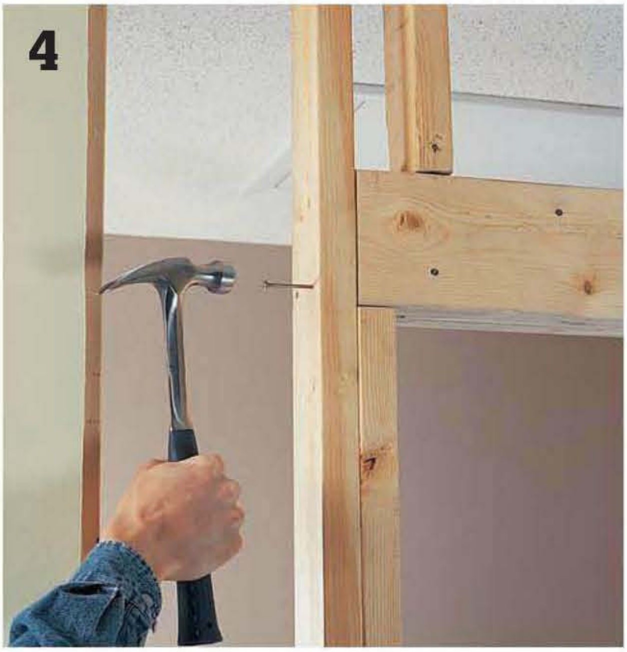

Build the header to size (pages 292 to 293) and install, fastening it to the jack studs, king studs, and cripple studs using 16d common nails. Use a handsaw to cut through the bottom plate so it's flush with the inside faces of the jack studs. Remove the cutout portion.

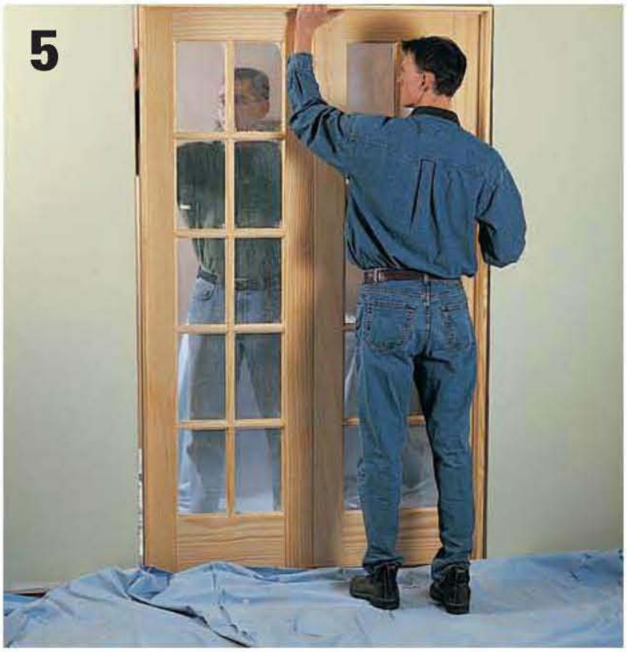

Finish the walls (for wallboard installation, see pages 44 to 51) before installing the doors, then set the prehung door unit into the framed opening so the jamb edges are flush with the finished wall surfaces and the unit is centered from side to side.

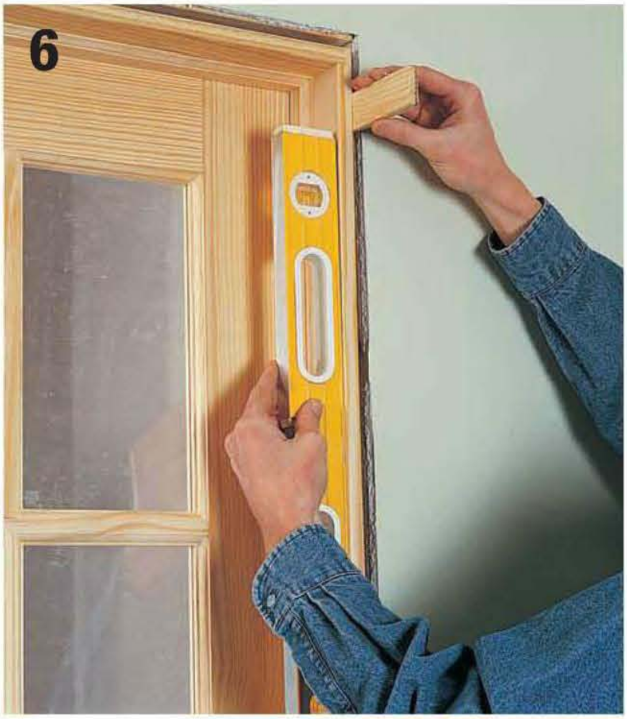

Using a level, adjust the unit to plumb one of the side jambs. Starting near the top of the door, insert pairs of shims driven from opposite directions into the gap between the framing and the jamb, sliding the shims until they are snug. Check the jamb to make sure it remains plumb and does not bow inward.

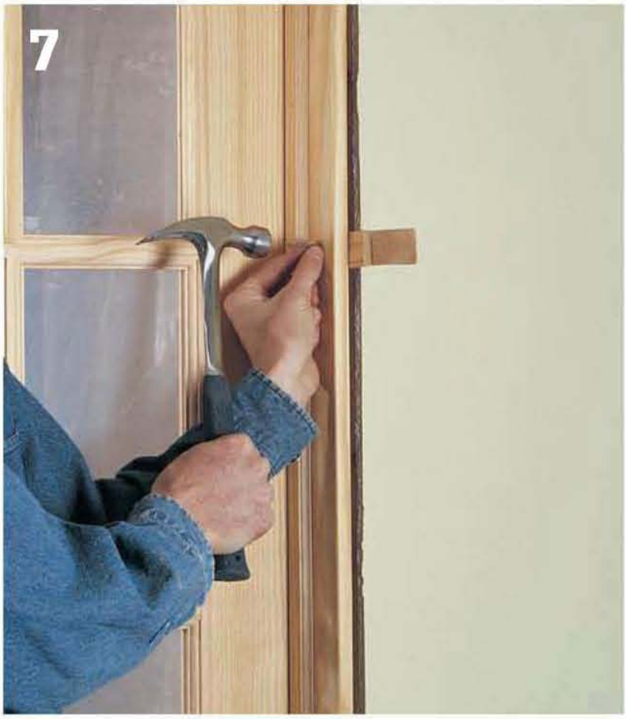

Working down along the jamb, install shims near each hinge and near the floor. Make sure the jamb is plumb, then anchor it with 8d finish nails driven through the jamb and shims and into the framing. Leave the nail heads partially protruding so the jamb can be readjusted later if necessary.

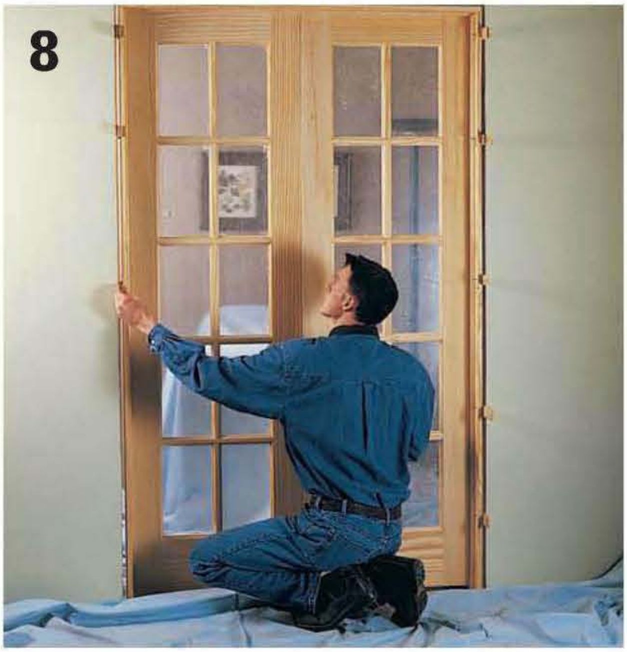

Install shims at the other side jamb, aligning them roughly with the shims of the first jamb. With the doors closed, adjust the shims so the reveal between the doors is even and the tops of the doors are aligned.

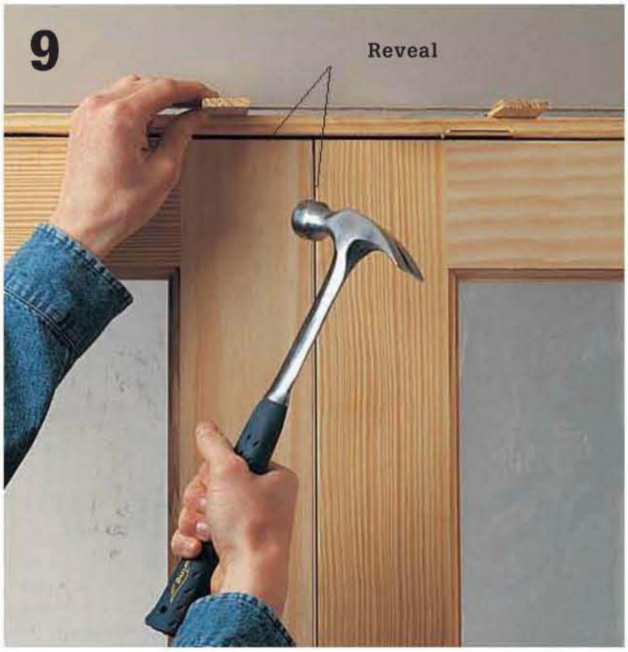

Shim the gap between the header and the head jamb to create a consistent reveal along the top when the doors are closed. Insert pairs of shims every 12". Drive 8d finish nails through the jambs and shims and into the framing members.

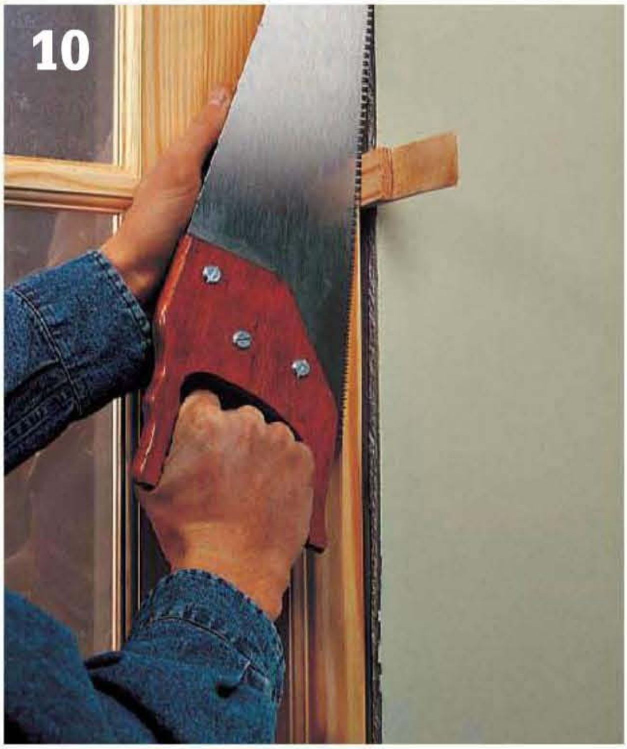

Drive all the nails fully, then set them below the surface of the wood with a nail set. Cut off the shims flush with the wall surface using a handsaw or utility knife. Hold the saw vertically to prevent damage to the door jamb or wall. Install the door casing.

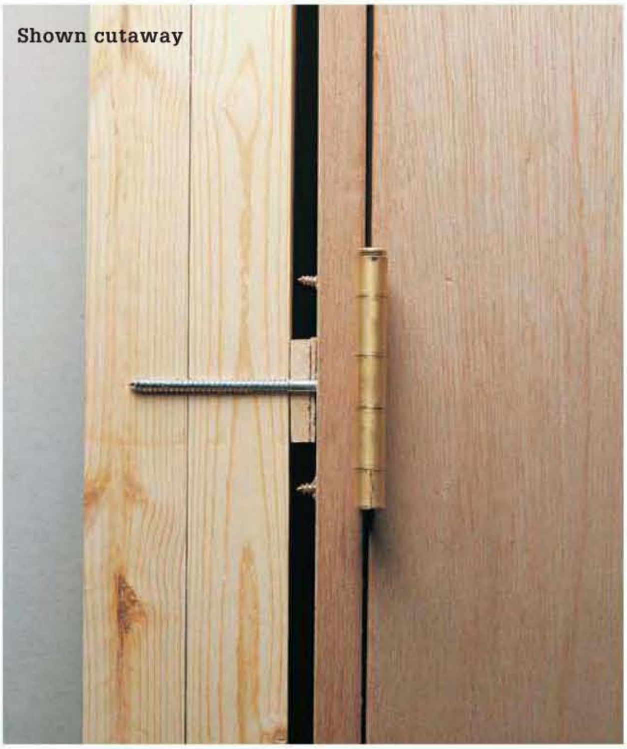

Option: Replace the center mounting screw on each hinge with a 3" wood screw to provide extra support for door hinges and jambs. These long screws extend through the side jambs and deep into the framing members. Be careful not to overtighten screws, which will cause the jambs to bow.

Installing Entry Doors

Few parts of a house have a more dramatic effect on the way your home is perceived than the main entry door. A lovely, well-maintained entryway that is tastefully matched architecturally to the house can utterly transform a home's appearance. In fact, industry studies have suggested that upgrading a plain entry door to a higher-end entry door system can pay back multiple times in the resale of your house. But perhaps more importantly, depending on your priorities, it makes a great improvement in how you feel about your home. Plus, it usually pays benefits in home security and energy efficiency as well.

If you are replacing a single entry door with a double door or a door with a sidelight or sidelights, you will need to enlarge the door opening (see pages 32 to 35). Be sure to file your plans with your local building department and obtain a permit. You'll need to provide temporary support from the time you remove the wall studs in the new opening until you've installed and secured a new door header that's approved for the new span distance.

The American Craftsman style door with sidelights (see Resources, page 554) installed in this project has the look and texture of a classic wood door, but it is actually created from fiber-glass. Today's fiberglass doors are quite convincing in their ability to replicate wood grain, while still offering the durability and low-maintenance of fiberglass.

Tools & Materials ▸

Tape measure
Level
Reciprocating saw
Caulk & caulk gun
Hammer
Shims
Framing nails
Finish nails
Nail set
Finishing materials

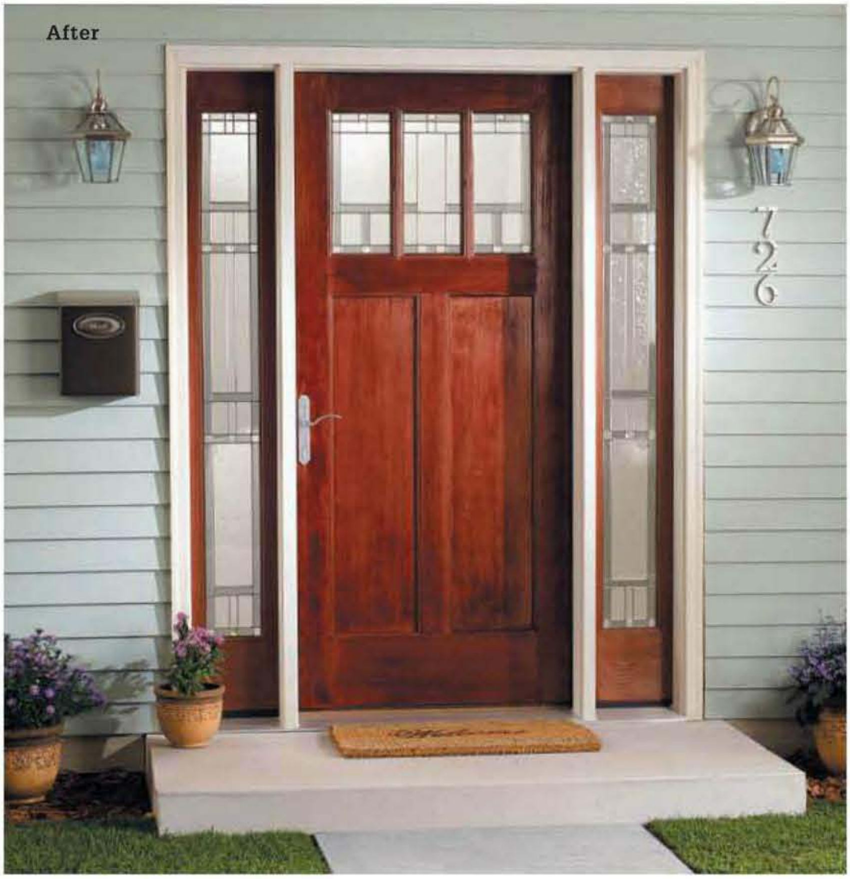

Replacing an ordinary entry door with a beautiful new upgrade has an exceptionally high payback in increased curb appeal and in perceived home value, according to industry studies.

How to Replace an Entry Door

1

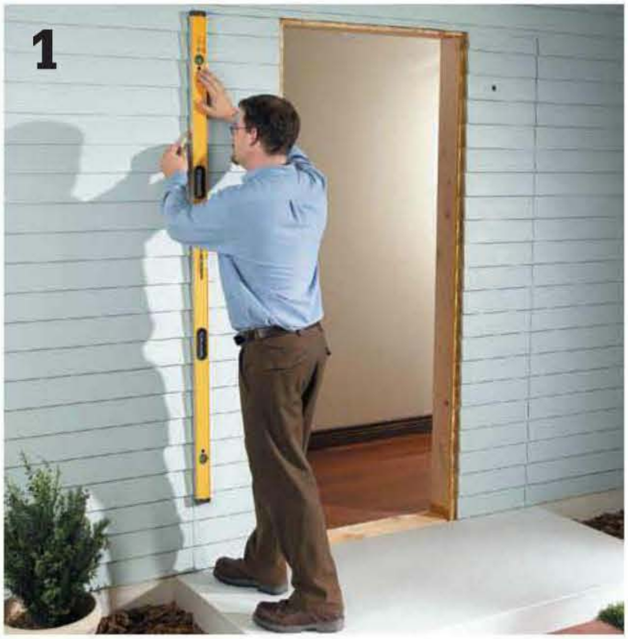

Remove the old entry door by cutting through the fasteners driven into the jamb with a reciprocating saw (see pages 32 and 35). If the new door or door system is wider, mark the edges of the larger rough opening onto the wall surface. If possible, try to locate the new opening so one edge will be against an existing wall stud. Be sure to include the thickness of the new framing you'll need to add when removing the wall coverings.

2

Frame in the new rough opening for the replacement door (see pages 257 to 261). The instructions that come with the door will recommend a rough opening size, which is usually sized to create a ½" gap between the door and the studs and header. Patch the wall surfaces.

3

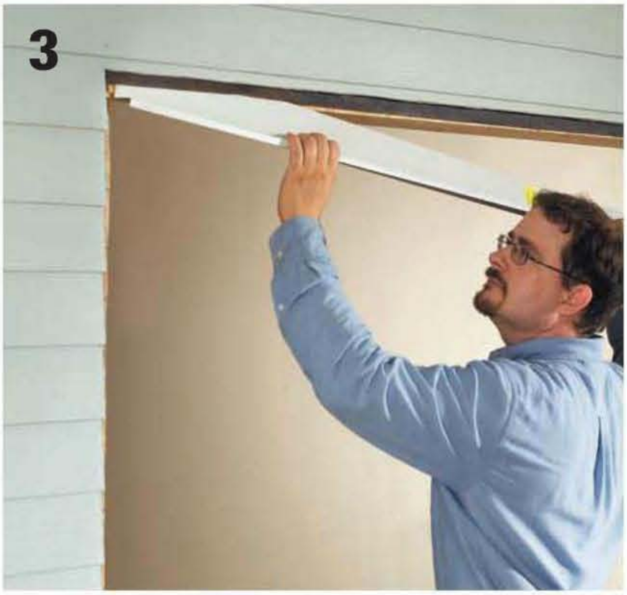

Cut metal door dripcap molding to fit the width of the opening and tuck the back edge up behind the wallcovering at the top of the door opening. Attach the dripcap with caulk only–do not use nails or screws.

4

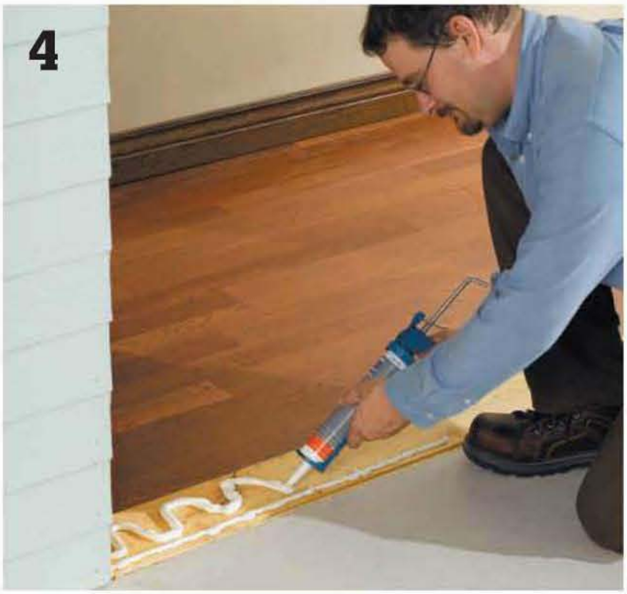

Unpack the door unit and set it in the rough opening to make sure it fits correctly. Remove it. Make sure the subfloor is clean and in good repair, and then apply heavy beads of caulk to the underside of the door sill and to the subfloor in the sill installation area. Use plenty of caulk.

(continued)

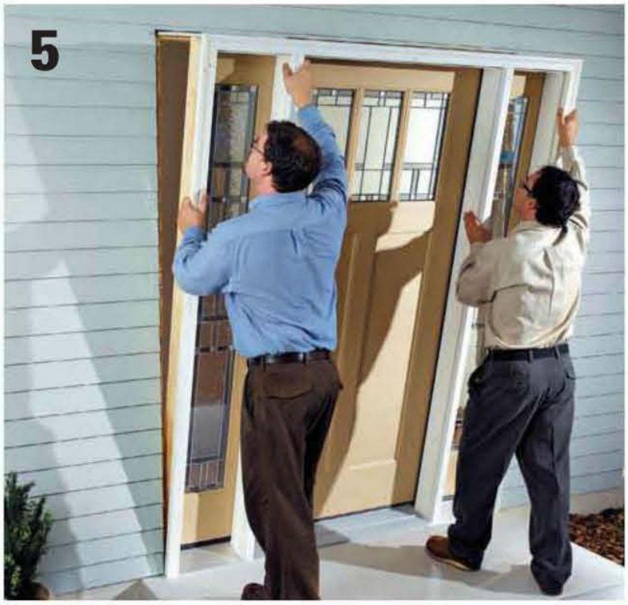

Set the door sill in the threshold and raise the unit up so it fits cleanly in the opening, with the exterior trim flush against the wall sheathing. Press down on the sill to seat it in the caulk and wipe up any squeeze-out with a damp rag

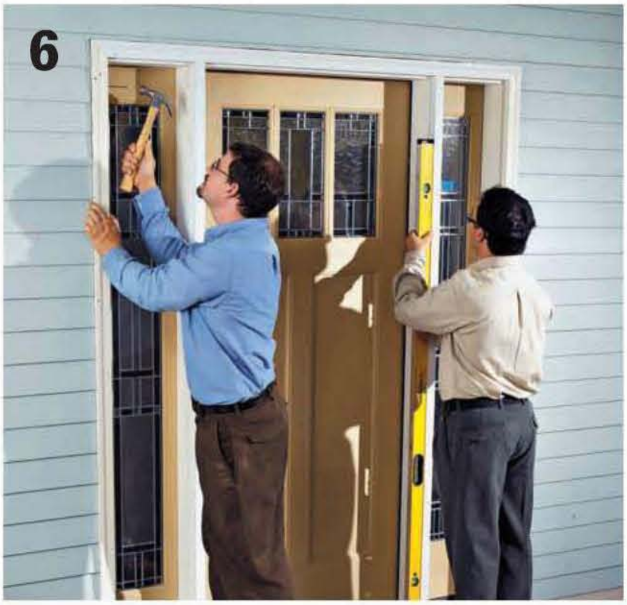

Use a 6-ft. level to make sure the unit is plumb and then tack it to the rough opening stud on the hinge side, using pairs of 10d nails driven partway through the casing on the weatherstripped side of the door (or the sidelight). On single, hinged doors, drive the nails just above the hinge locations. *Note: Many door installers prefer deck screws over nails when attaching the jambs. Screws offer more gripping strength and are easier to adjust, but covering the screw heads is more difficult than filling nail holes.*

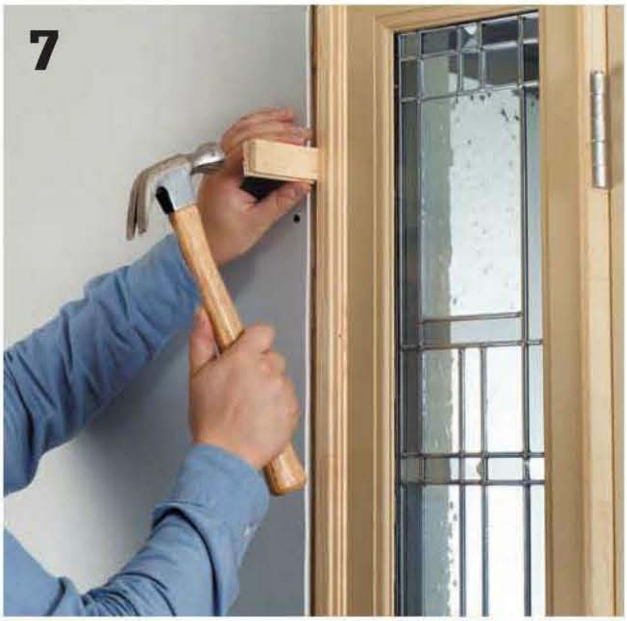

Drive wood shims between the jamb and the wall studs to create an even gap. Locate the shims directly above the pairs of nails you drove. Doublecheck the door with the level to make sure it is still plumb.

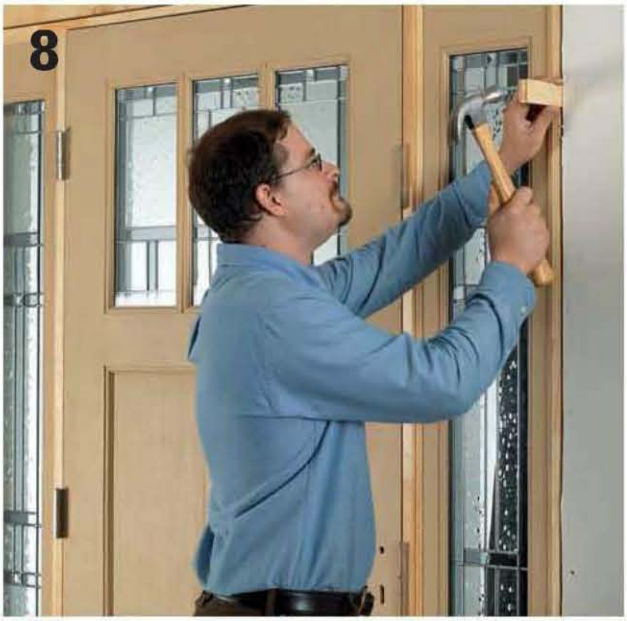

Drive shims between the jamb on the latch side of the unit and into the wall stud. Only drive the nails part way. Test for plumb again and then add shims at nail locations (you may need to double-up the shims, as this gap is often wider than the one on the hinge side). Check to make sure the door jamb is not bowed.

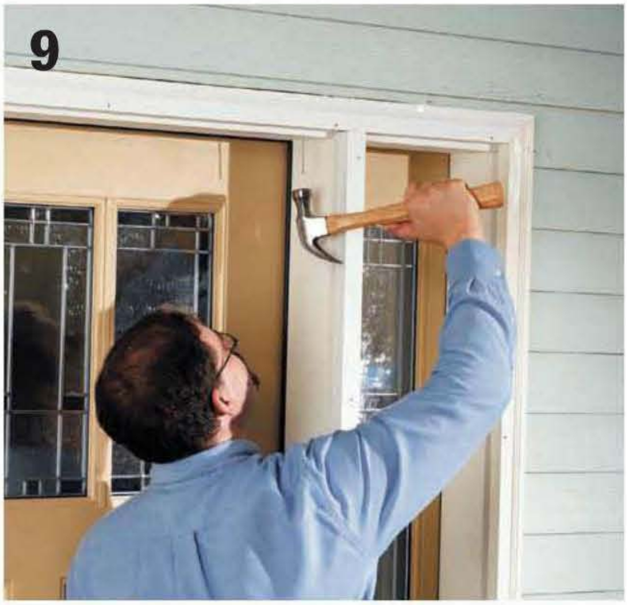

Drive finish nails at all remaining locations, following the nailing schedule in the manufacturer's installation instructions.

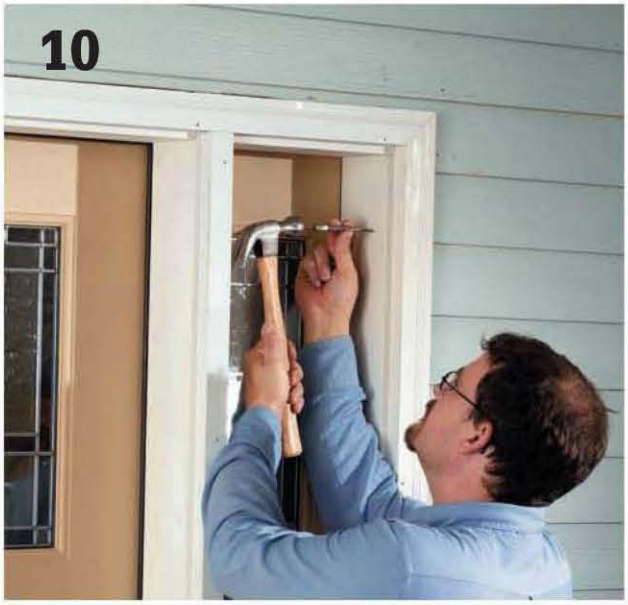

Use a nail set to drive the nail heads below the wood surface. Fill the nail holes with wood putty (you'll get the best match if you apply putty that's tinted to match stained wood after the finish is applied). The presence of the wood shims at the nail locations should prevent the jamb from bowing as you nail.

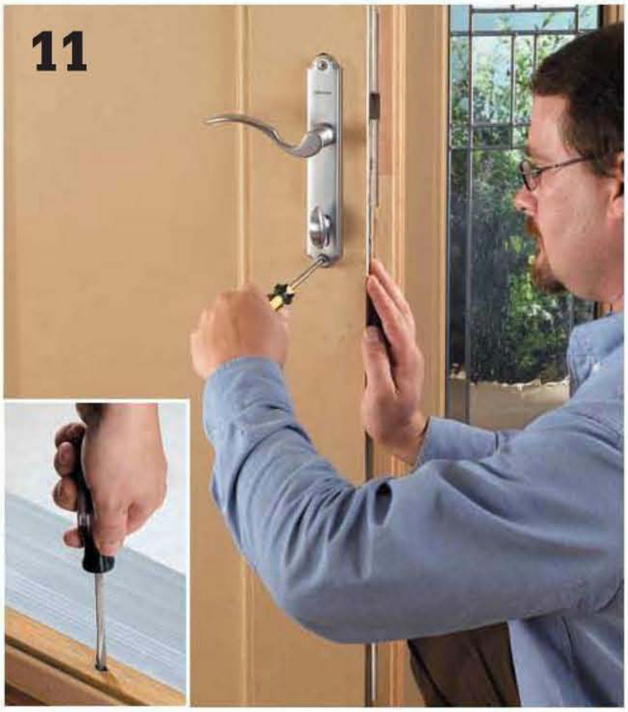

Install the lockset, strikeplates, deadbolts or multipoint locks, and any other door hardware. If the door finish has not been applied, you may want to do so first, but generally it makes more sense to install the hardware right away so the door can be operated and locked. Attach the door sill to the threshold and adjust it as needed, normally using the adjustment screws (inset).

Apply your door finish if it has not yet been applied. Read the manufacturer's suggestions for finishing very closely and follow the suggested sequences. Some manufacturers offer finish kits that are designed to be perfectly compatible with their doors. Install interior case molding and caulk all the exterior gaps after the finish dries.

Installing Storm Doors

Install a storm door to improve the appearance and weather resistance of an old entry door or to protect a newly installed door against weathering. In all climates, adding a storm door can extend the life of an entry door.

When buying a storm door, look for models that have a solid inner core and seamless outer shell construction. Carefully note the dimensions of your door opening, measuring from the inside edges of the entry door's brick molding. Choose a storm door that opens from the same side as your entry door.

Adjustable sweeps help make storm doors weathertight. Before installing the door, attach the sweep to the bottom of the door. After the door is mounted, adjust the height of the sweep so it brushes the top of the sill lightly when the door is closed.

Tools & Materials ▸

Tape measure
Pencil
Plumb bob
Hacksaw
Hammer
Drill and bits
Screwdrivers
Storm door unit
Wood spacer strips
4d casing nails

How to Cut a Storm Door Frame to Fit a Door Opening

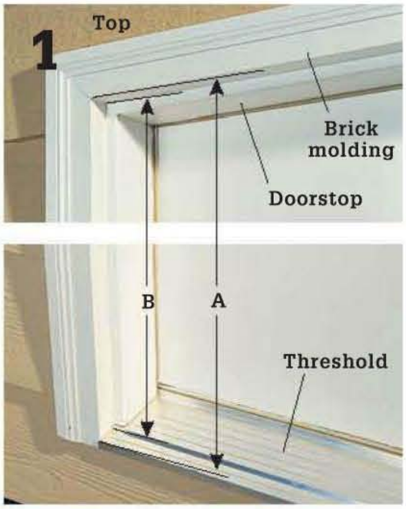

Because entry door thresholds are slanted, the bottom of the storm door frame needs to be cut to match the threshold angle. Measure from the threshold to the top of the door opening along the corner of the brick molding (A), then measure along the front edge of the entry doorstop (B).

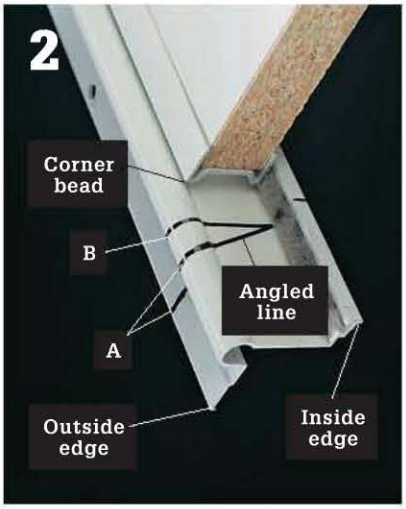

Subtract ⅛" from measurements A and B to allow for small adjustments when the door is installed. Measuring from the top of the storm door frame, mark the adjusted points A and B on the corner bead. Draw a line from point A to the outside edge of the frame and from point B to the inside edge. Draw an angled line from point A on the corner bead to point B on the inside edge.

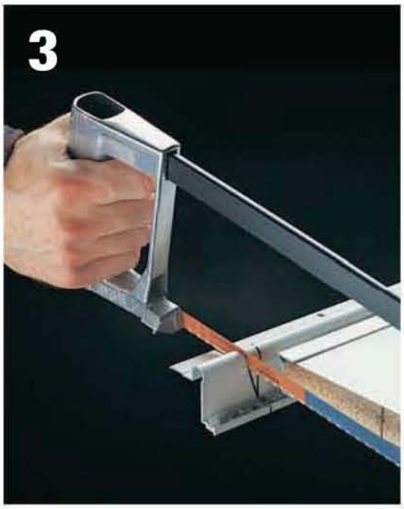

Use a hacksaw to cut down through the bottom of the storm door frame, following the angled line. Make sure to hold the hacksaw at the same slant as the angled line to ensure that the cut will be smooth and straight.

How to Fit & Install a Storm Door

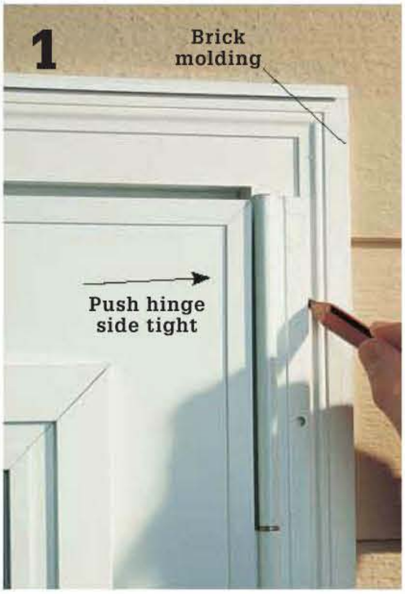

Position the storm door in the opening and push the frame tight against the brick molding on the hinge side of the storm door, then draw a reference line on the brick molding, following the edge of the storm door frame.

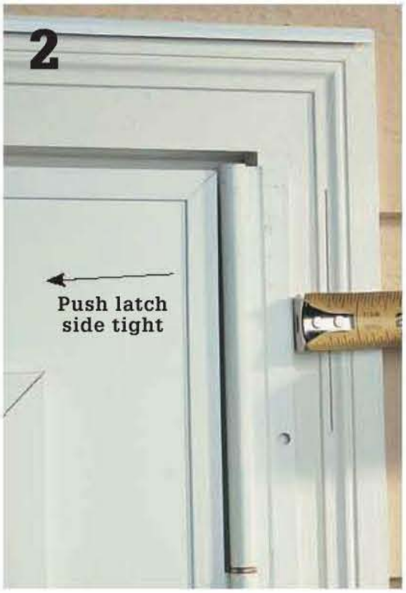

Push the storm door frame tight against the brick molding on the latch side, then measure the gap between the reference line and the hinge side of the door frame. If the distance is greater than ⅜", spacer strips must be installed to ensure the door will fit snugly.

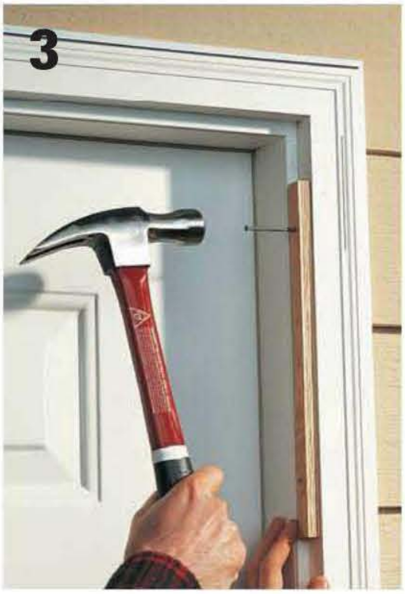

To install spacers, remove the door, then nail thin strips of wood to the inside of the brick molding at storm door hinge locations. The thickness of the wood strips should be ⅛" less than the gap measured in step 2.

Replace the storm door and push it tight against the brick molding on the hinge side. Drill pilot holes through the hinge-side frame of the storm door and into the brick molding spaced every 12". Attach the frame with mounting screws.

Remove any spacer clips holding the frame to the storm door. With the storm door closed, drill pilot holes and attach the latch-side frame to the brick molding. Use a coin to keep an even gap between the storm door and the storm door frame.

Center the top piece of the storm door frame on top of the frame sides. Drill pilot holes and screw the top piece to the brick molding. Adjust the bottom sweep, then attach the locks and latch hardware as directed by the manufacturer.

Installing Patio Doors

For easy installation, buy a patio door with the door panels already mounted in a preassembled frame. Try to avoid patio doors sold with frame kits that require complicated assembly.

Because patio doors have very long bottom sills and top jambs, they are susceptible to bowing and warping. To avoid these problems, be very careful to install the patio door so it is level and plumb and to anchor the unit securely to framing members. Yearly caulking and touch-up painting helps prevent moisture from warping the jambs.

Tools & Materials ▸

- Pencil
- Hammer
- Circular saw
- Handsaw
- Wood chisel
- Stapler
- Caulk gun
- Level
- Pry bar
- Cordless screwdriver
- Drill and bits
- Nail set
- Shims
- Drip edge
- Building paper
- Silicone and latex caulk
- 10d casing nails
- 3" wood screws
- Sill nosing
- Fiberglass insulation
- Patio door kit

Patio doors offer the best qualities of both windows and doors—plenty of natural light, a great view, wide room access, and reasonable security.

If not included with the unit, screen doors can be ordered from most patio door manufacturers. Screen doors have spring-mounted rollers that fit into a narrow track on the outside of the patio door threshold.

Installing Sliding Patio Doors

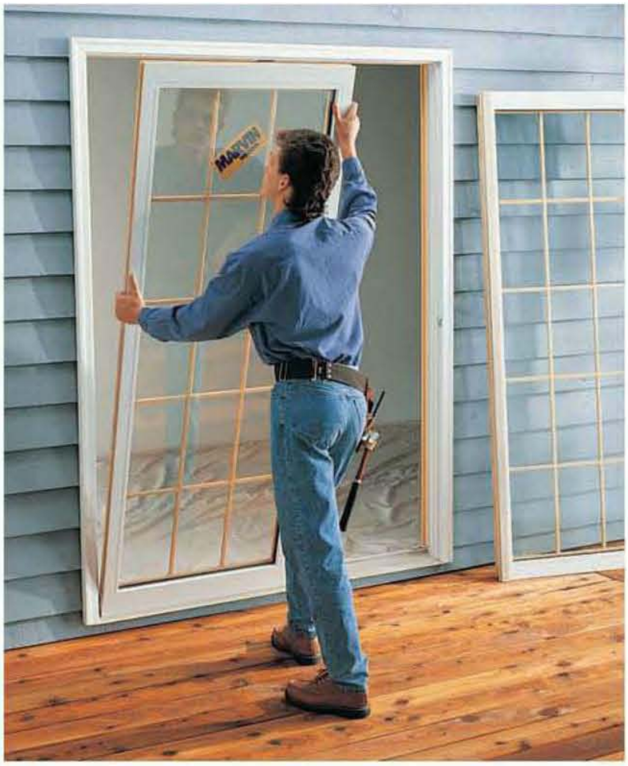

Remove heavy glass panels if you must install the door without help. Reinstall the panels after the frame has been placed in the rough opening and nailed at opposite corners. To remove and install the panels, remove the stop rail found on the top jamb of the door unit.

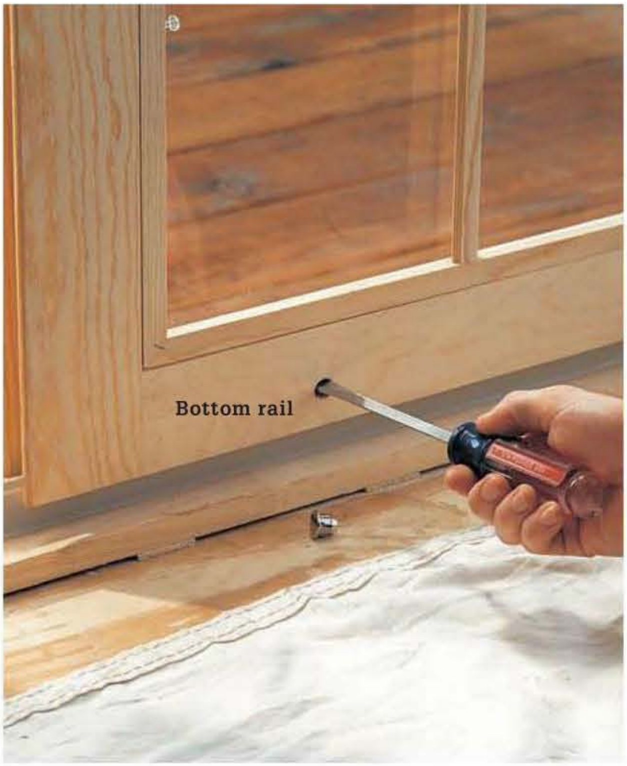

Adjust the bottom rollers after installation is complete. Remove the coverplate on the adjusting screw, found on the inside edge of the bottom rail. Turn the screw in small increments until the door rolls smoothly along the track without binding when it is opened and closed.

Tips for Installing Hinged Patio Doors ▸

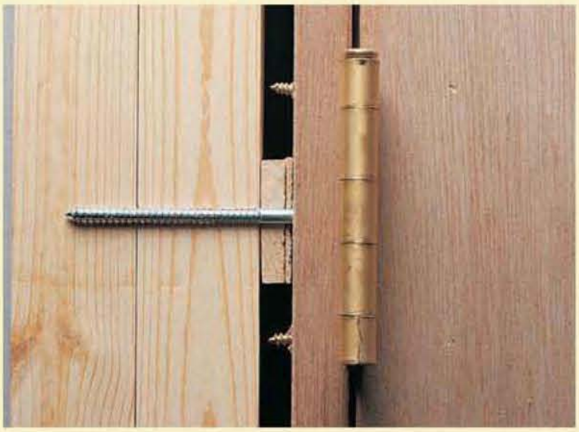

Provide extra support for door hinges by replacing the center mounting screw on each hinge with a 3" wood screw. These long screws extend through the side jambs and deep into the framing members.

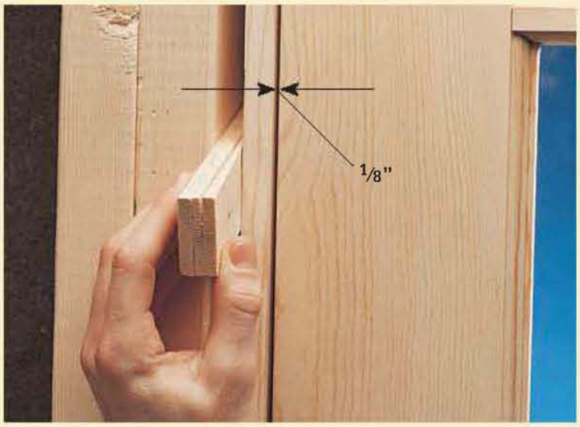

Keep a uniform ⅛" gap between the door, side jambs, and top jamb to ensure that the door will swing freely without binding. Check this gap frequently as you shim around the door unit.

How to Install a Patio Door

Prepare the work area and remove the interior wall surfaces, then frame the rough opening for the patio door. Remove the exterior surfaces inside the framed opening.

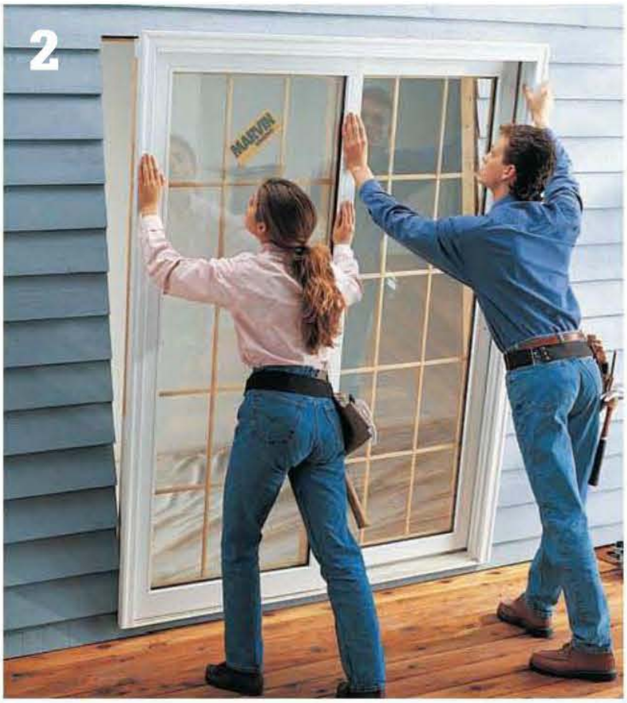

Test-fit the door unit, centering it in the rough opening. Check to make sure the door is plumb. If necessary, shim under the lower side jamb until the door is plumb and level. Have a helper hold the door in place while you adjust it.

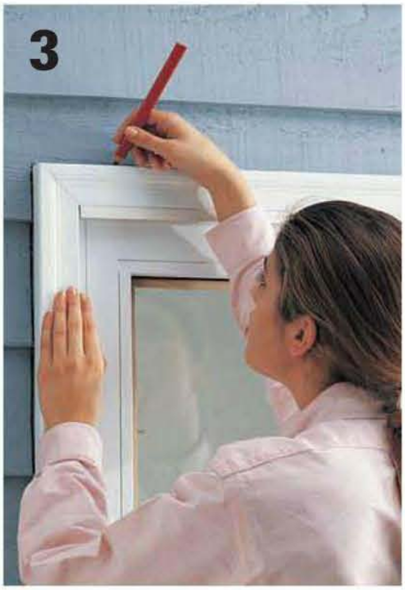

Trace the outline of the brick molding onto the siding, then remove the door unit. *Note: If you have vinyl or metal siding, see page 282 for advice on removing the siding.*

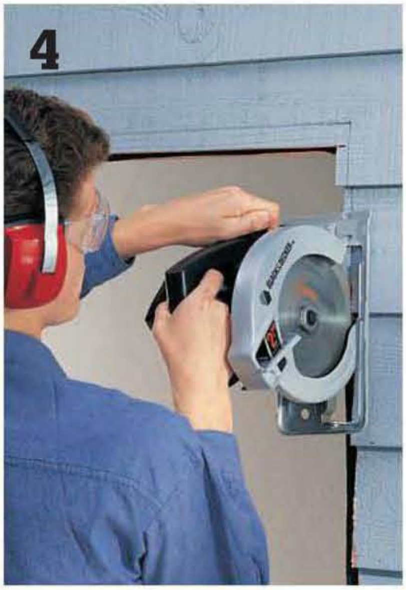

Cut the siding along the outline, just down to the sheathing using a circular saw. Stop just short of the corners to prevent damage to the remaining siding. Finish the cuts at the corners with a sharp wood chisel.

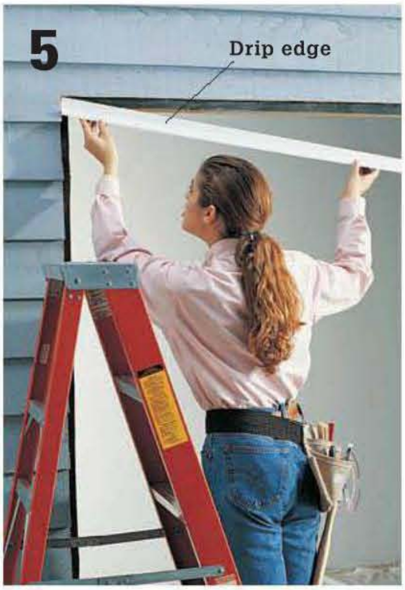

To provide an added moisture barrier, cut a piece of drip edge to fit the width of the rough opening, then slide it between the siding and the existing building paper at the top of the opening. Do not nail the drip edge.

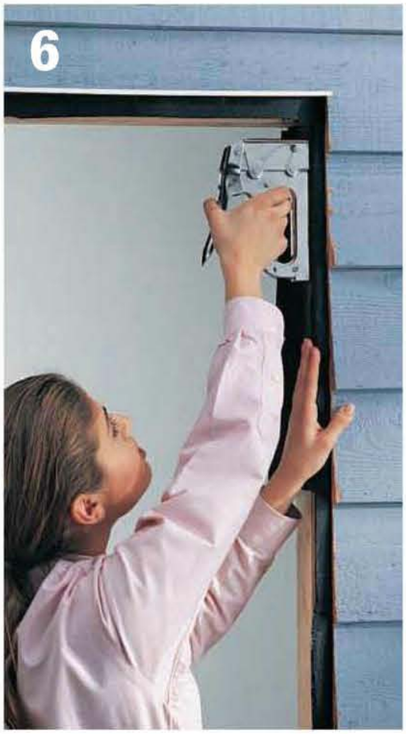

Cut 8"-wide strips of building paper and slide them between the siding and sheathing. Bend the paper around the framing members and staple it in place. Each piece overlaps the piece below it.

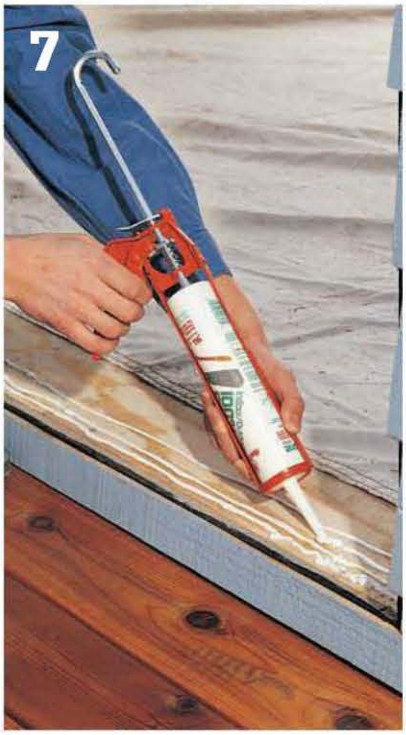

Apply several thick beads of silicone caulk to the subfloor at the bottom of the door opening.

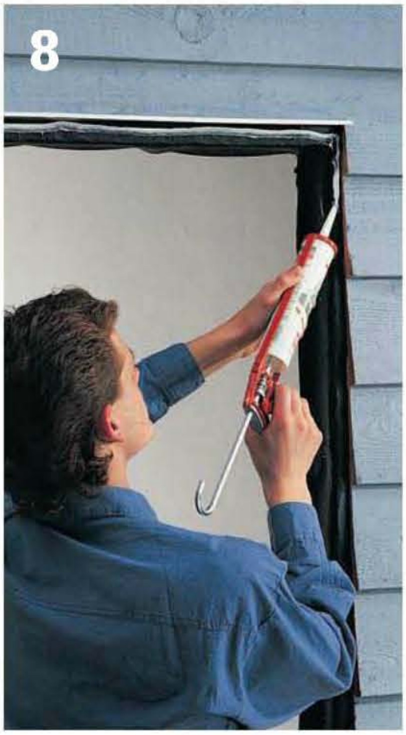

Apply silicone caulk around the front edge of the framing members where the siding meets the building paper.

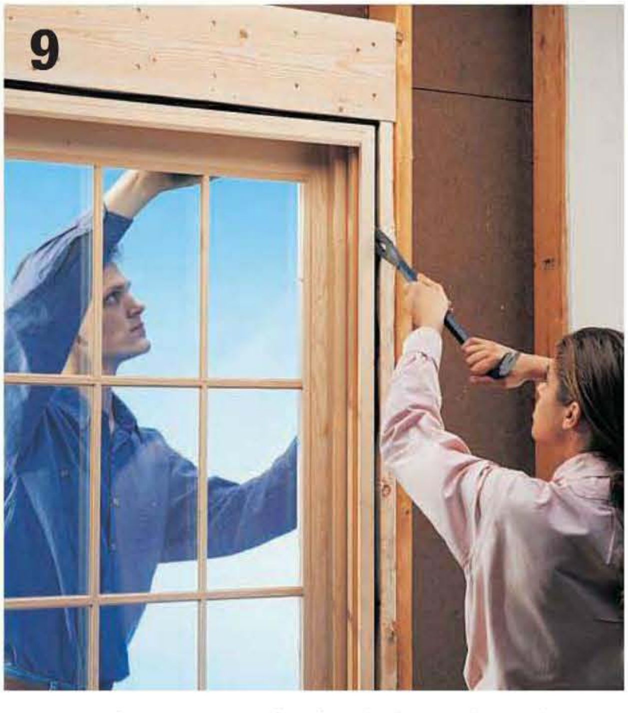

Use a pry bar to center the door in the rough opening so the brick molding is tight against the sheathing. Have a helper hold the door unit from outside.

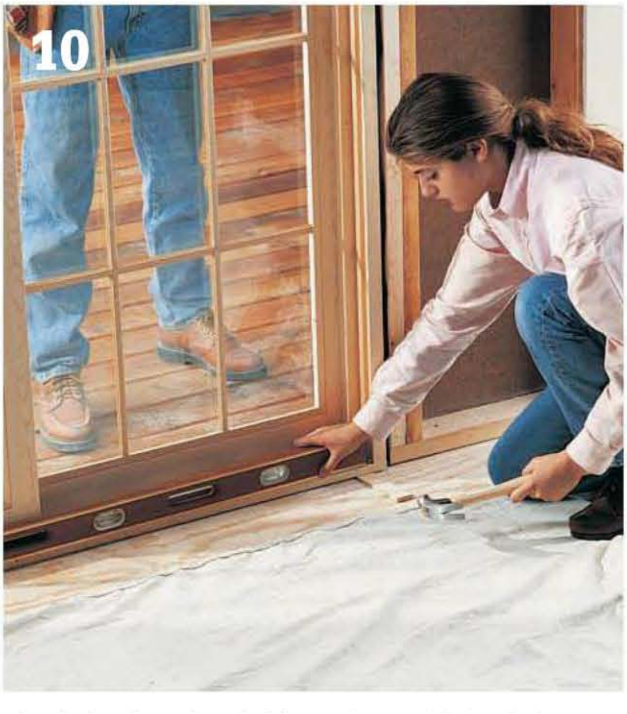

Check the door threshold to make sure it is level. If necessary, shim under the lower side jamb until the patio door unit is level.

(continued)

11

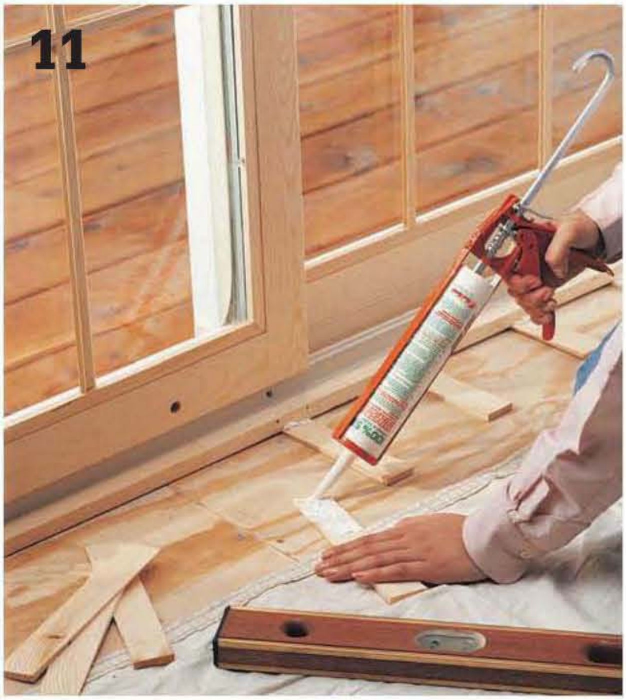

If there are gaps between the threshold and subfloor, insert shims coated with caulk into the gaps, spaced every 6". Shims should be snug, but not so tight that they cause the threshold to bow. Clear off excess caulk immediately.

12

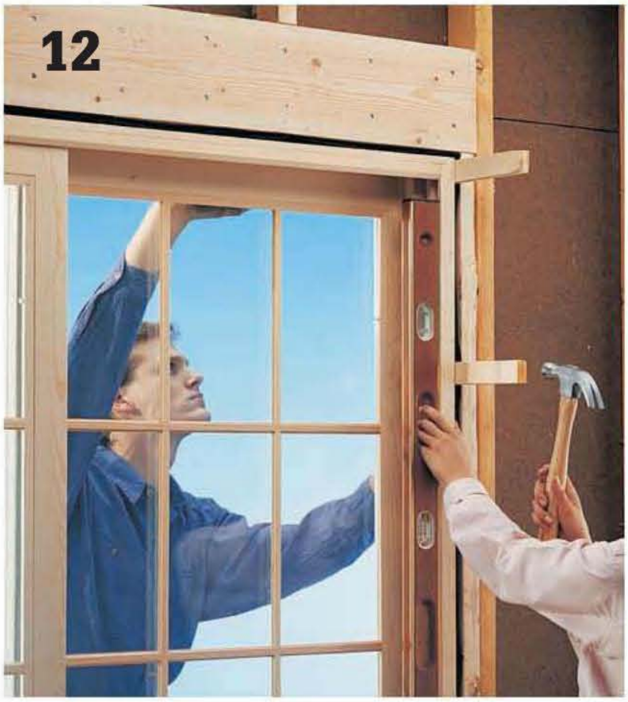

Place pairs of hardwood wedge shims together to form flat shims. Insert the shims every 12" into the gaps between the side jambs and the jack studs. For sliding doors, shim behind the strike plate for the door latch.

13

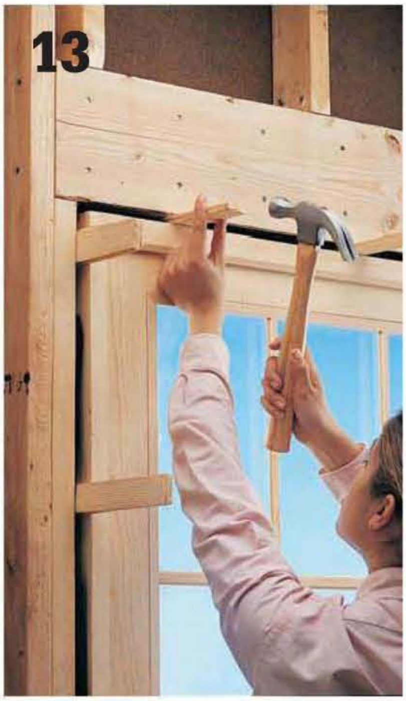

Insert shims every 12" into the gap between the top jamb and the header.

14

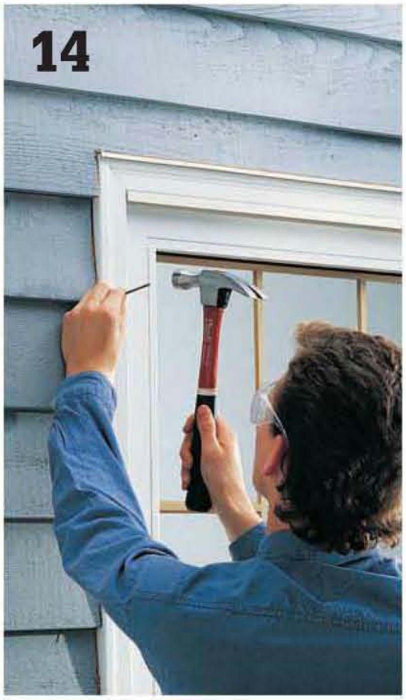

From outside, drive 10d casing nails, spaced every 12", through the brick molding and into the framing members. Use a nail set to drive the nail heads below the surface of the wood.

15

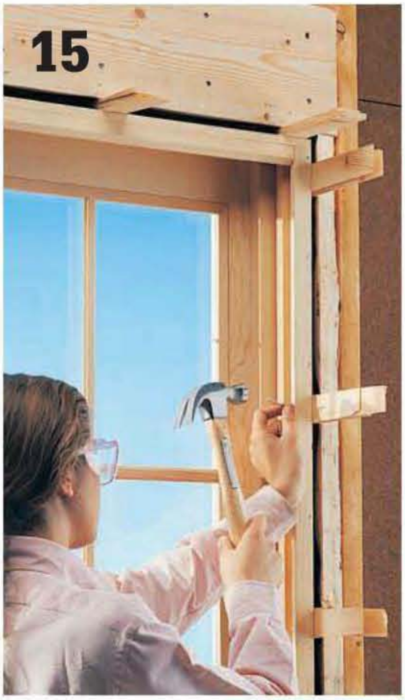

From inside, drive 10d casing nails through the door jambs and into the framing members at each shim location. Use a nail set to drive the nail heads below the surface of the wood.

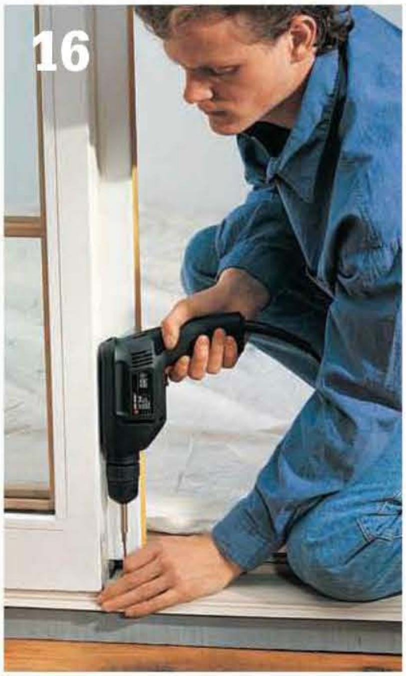

Remove one of the screws and cut the shims flush with the stop block found in the center of the threshold. Replace the screw with a 3" wood screw driven into the subfloor as an anchor.

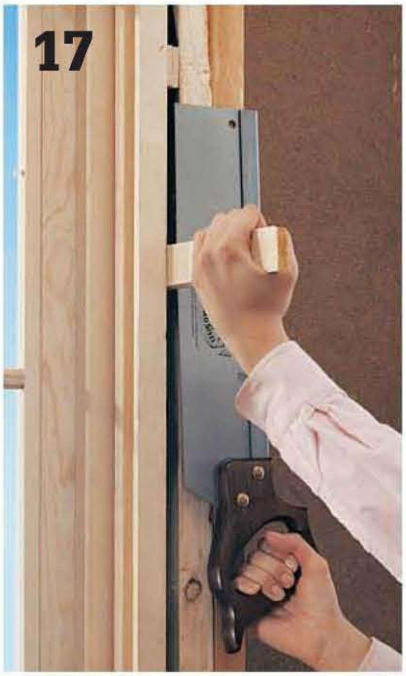

Cut off the shims flush with the face of the framing members using a handsaw. Fill gaps around the door jambs and beneath the threshold with loosely packed fiberglass insulation.

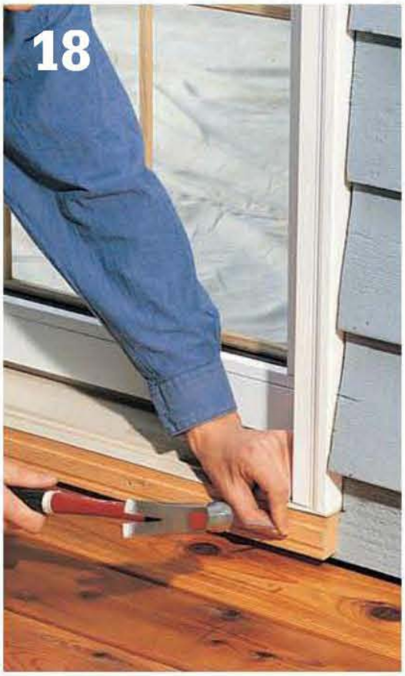

Reinforce and seal the edge of the threshold by installing sill nosing under the threshold and against the wall. Drill pilot holes and attach the sill nosing with 10d casing nails.

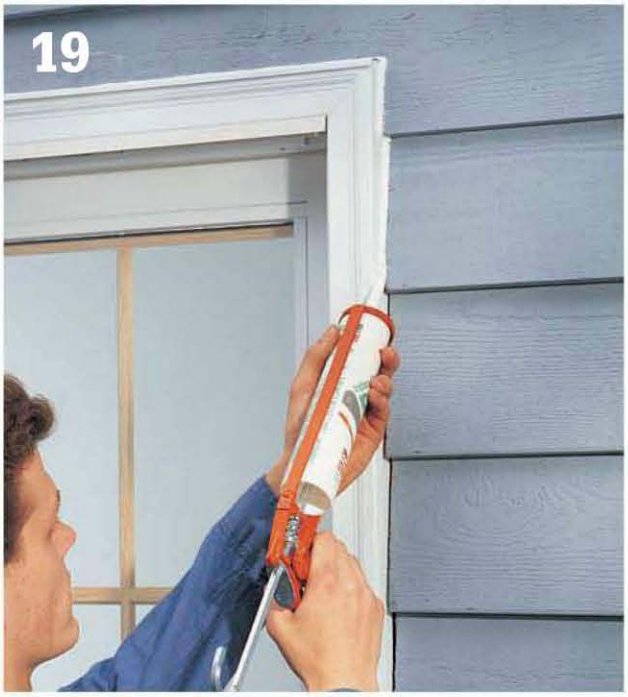

Make sure the drip edge is tight against the top brick molding, then apply paintable silicone caulk along the top of the drip edge and along the outside edge of the side brick moldings. Fill all exterior nail holes with caulk.

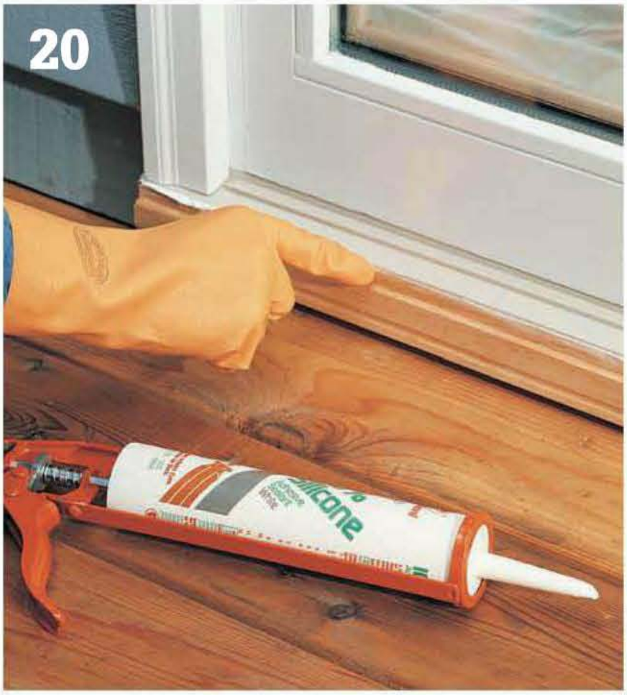

Caulk completely around the sill nosing using your finger to press the caulk into any cracks. As soon as the caulk is dry, paint the sill nosing. Finish the door and install the lockset as directed by the manufacturer.

Installing Door & Window Casing

Door and window casings provide an attractive border around doors and windows. They also cover the gaps between door or window jambs and the surfaces of surrounding walls.

Install door and window casings with a consistent reveal between the inside edges of the jambs and casings, making sure the casings are level and plumb.

In order to fit casings properly, the jambs and wallcoverings must lie in the same plane. If either one protrudes, the casings will not lie flush. To solve this problem, you'll need to remove some material from whichever surface is protruding.

Use a block plane to shave protruding jambs or a surface forming rasp to shave a protruding wallboard edge. Wallboard screws rely on the strength of untorn facing paper to support the wallboard. If the paper around the screws is damaged, drive additional screws nearby where the paper is still intact.

Tools & Materials ▸

Tape measure
Pencil
Combination square
Nail set
Level
Straightedge
Power miter saw
Hammer or pneumatic nailer
Casing material
Plinths and corner blocks (optional)
4d and 6d finish nails
Wood putty

Case molding is installed around windows and doors to conceal the gaps between jambs and the wall. Venturing beyond the very common ranch-style casing offers some high design payback.

How to Install Mitered Casing on Doors & Windows

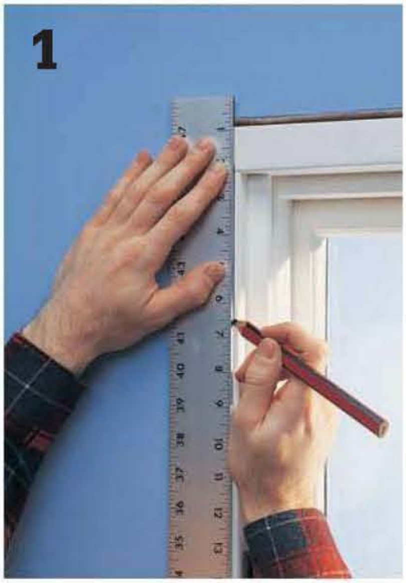

On each jamb, mark a reveal line ⅛" from the inside edge. The casings will be installed flush with these lines. *Note: On double-hung windows, the casings are usually installed flush with the edge of the jambs, so no reveal line is needed.*

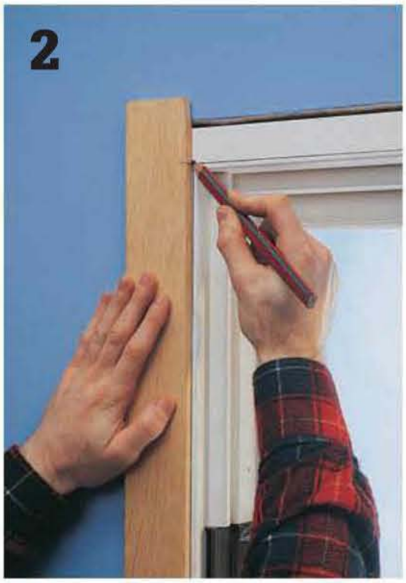

Place a length of casing along one side jamb, flush with the reveal line. At the top and bottom of the molding, mark the points where horizontal and vertical reveal lines meet. (When working with doors, mark the molding at the top only).

Make 45° miter cuts on the ends of the moldings. Measure and cut the other vertical molding piece, using the same method.

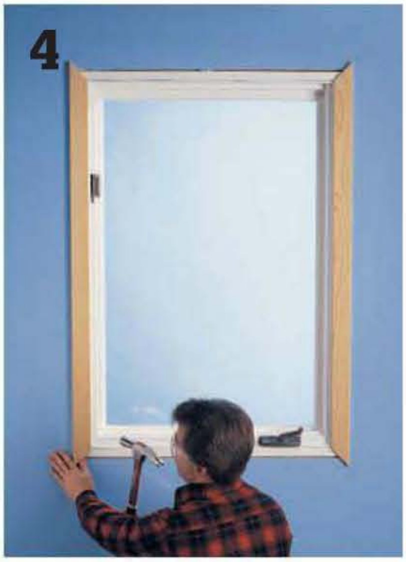

Drill pilot holes spaced every 12" to prevent splitting and attach the vertical casings with 4d finish nails driven through the casings and into the jambs. Drive 6d finish nails into framing members near the outside edge of the casings.

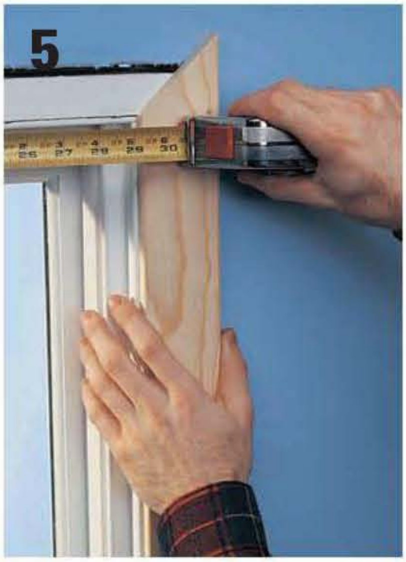

Measure the distance between the side casings, and cut top and bottom casings to fit, with ends mitered at 45°. If window or door unit is not perfectly square, make test cuts on scrap pieces to find the correct angle of the joints. Drill pilot holes and attach with 4d and 6d finish nails.

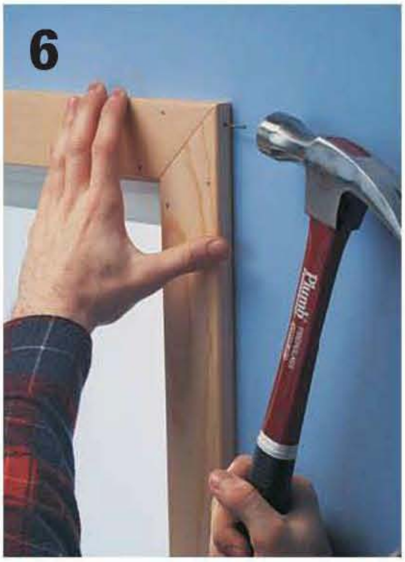

Locknail the corner joints by drilling pilot holes and driving 4d finish nails through each corner, as shown. Drive all nail heads below the wood surface, using a nail set, then fill the nail holes with wood putty.

How to Install Butted Door Casings

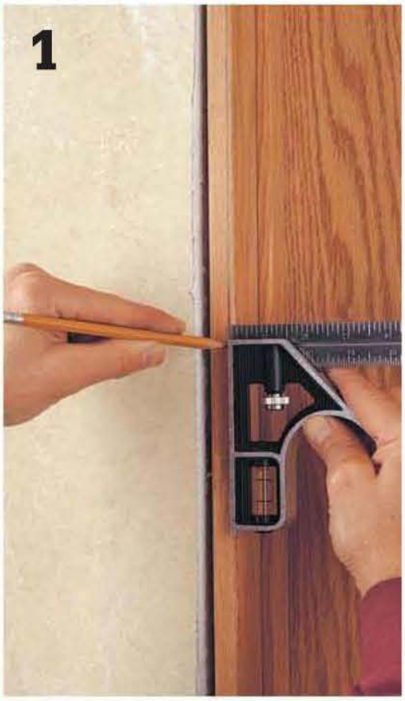

On each jamb, mark a reveal line ⅛" from the inside edge. The casings will be installed flush with these lines.

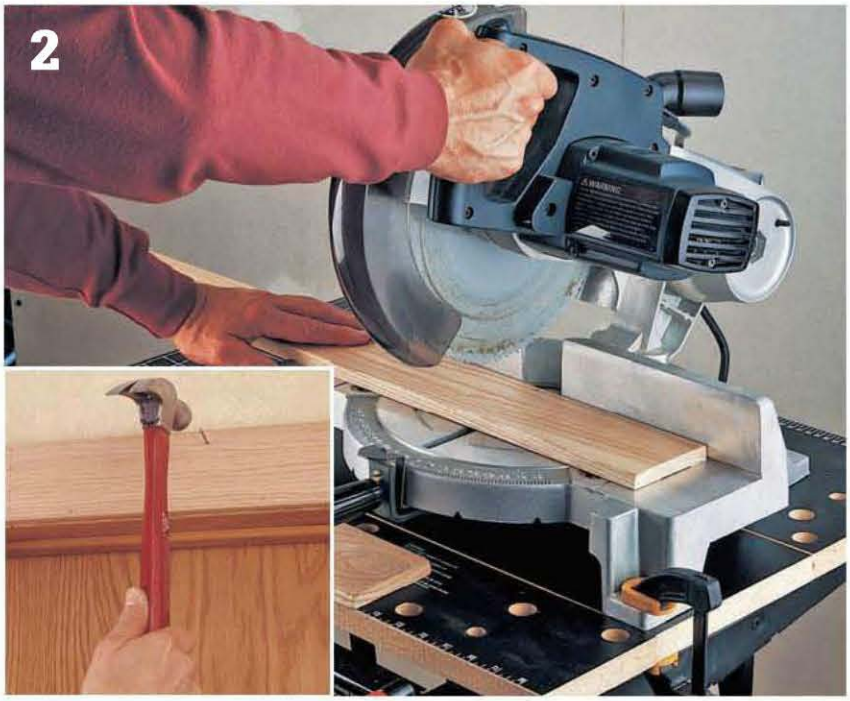

Cut the head casing to length. Mark the centerpoint of the head casing and the centerpoint of the head jamb. Align the casing with the head jamb reveal line, matching the centerpoints so that the head casing extends evenly beyond both side jamb casings. Nail the casing to the wall at stud locations and at the jamb (inset).

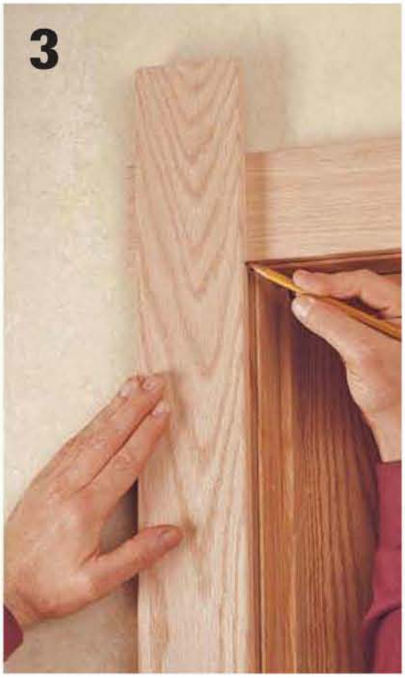

Hold the side casings against the head casing and mark them for cutting, then cut the side casings to fit.

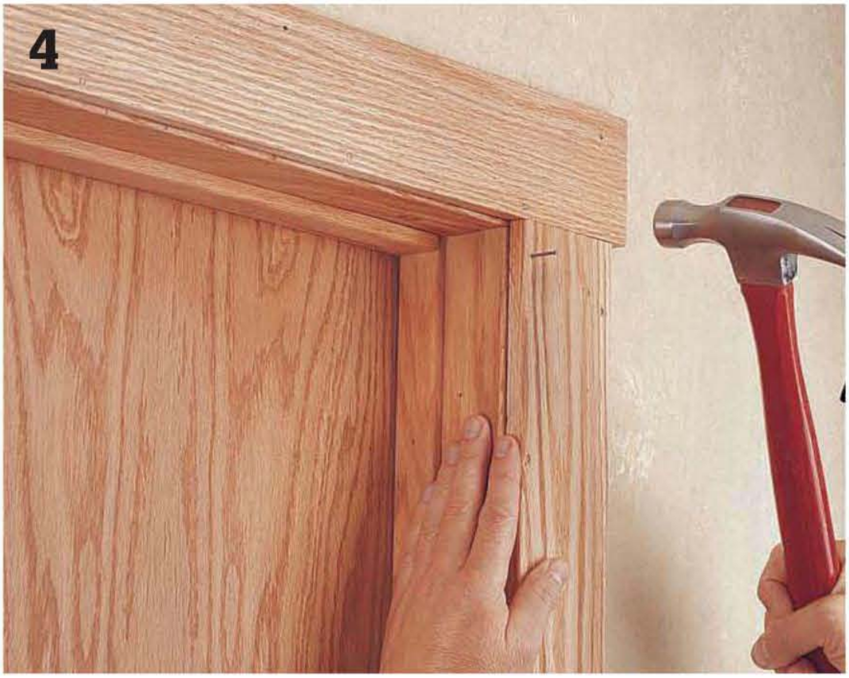

Align the side casings with the side jamb reveal lines, then nail the casings to the jambs and framing members. Set the nails, using a nail set. Fill the nail holes with wood putty.

Options for Installing Door & Window Casings

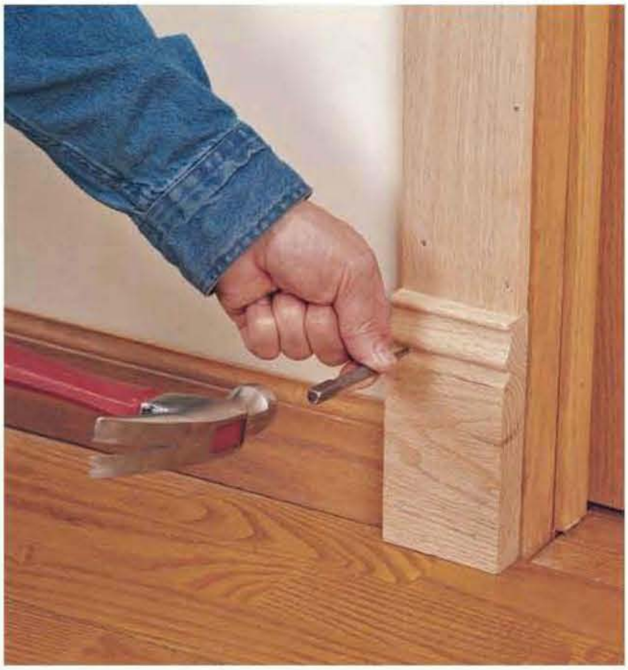

Dress up door casings by adding plinths. Cut the plinths from 1× stock and bevel one edge. Nail the plinths to the jambs with 2" 10d finish nails so the beveled edges are aligned with the reveal lines for the casings. Measure and cut the casings to fit.

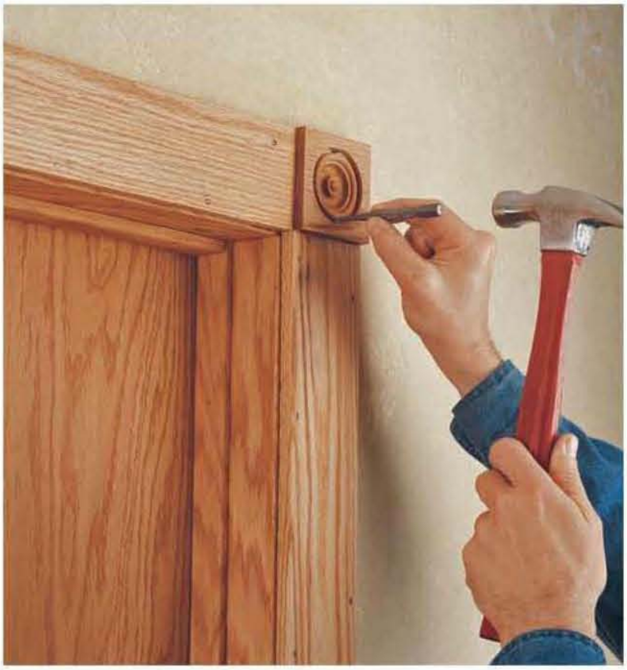

Add corner blocks, also known as rosettes, at the ends of the head casing. Attach the corner blocks once the side casings are in place, then cut the head casing to fit. Set the nails, using a nailset, after all pieces are installed.

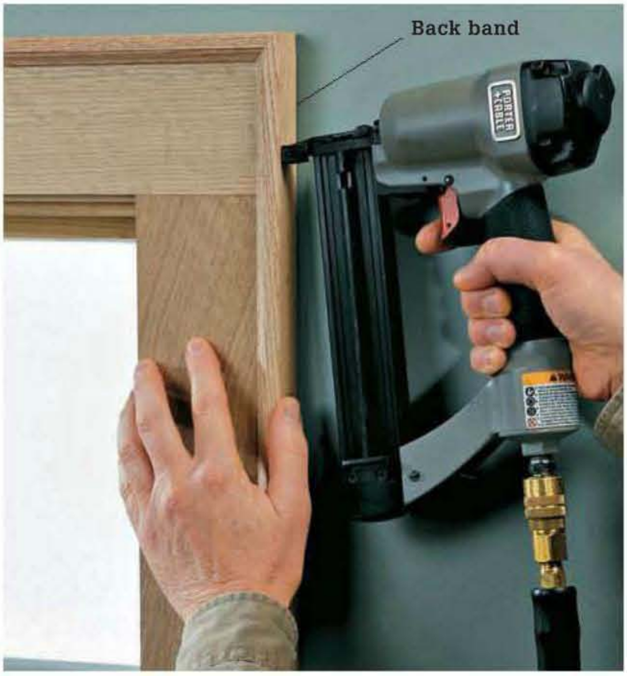

Backband molding can dress up butted window casings. Install the back band around the perimeter of the window, mitering the joints at the corners. Nail the back band in place with 4d finish nails or pneumatic brads.

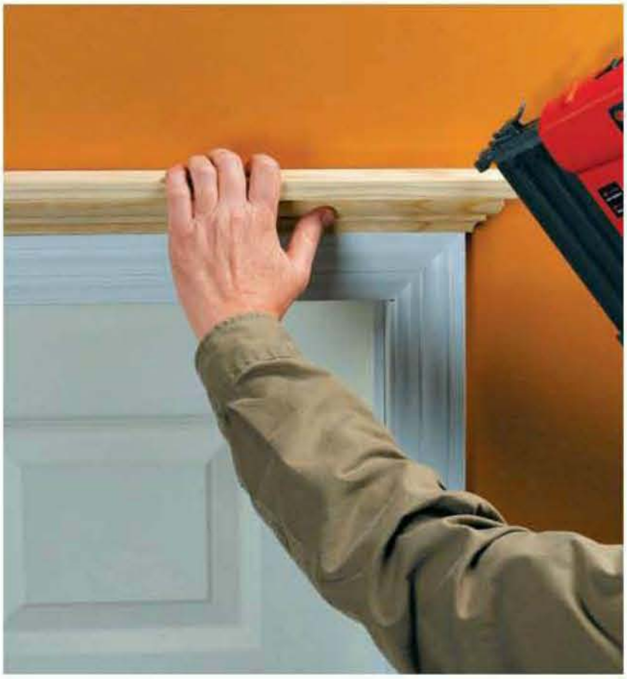

Create a decorative door header by nailing a combination of bed and lattice moldings over the top casing. Size the header to overhang the side casings.

Lighting

Artificial lighting is a necessary component of any room, but beyond simply providing illumination, lighting fixtures are an important part of a room's design. From the low-key but modern appearance of recessed canister lights to the elegance of a crystal chandelier, your choice of fixture will go a long way toward defining your remodeled room.

The easiest lighting improvement is to simply make a one-for-one replacement of an old fixture. The electrical connections are simple and, depending on the style you choose, you may not have to do any alterations to the room at all. Just make sure the light the new fixture provides is adequate for your room. In other words, don't replace an old fixture that holds three 60-watt bulbs with a single 25-watt pendant unless you are providing additional new light.

This chapter shows how to hook up the three types of light fixture you're most likely to include in your remodeling plan: an overhead ceiling light, recessed canister lights, and track lighting. If you need to add new wiring to supply the light, see the information on pages 82 through 147. Always shut off power at the main service panel before beginning even the simplest wiring project. Call a professional if you are unsure about your wiring job.

In this chapter:

- Installing Ceiling Lights
- Installing Recessed Ceiling Lights
- Installing Track Lighting

Installing Ceiling Lights

Ceiling fixtures don't have any moving parts and their wiring is very simple, so, other than changing bulbs, you're likely to get decades of trouble-free service from a fixture. This sounds like a good thing, but it also means that the fixture probably won't fail and give you an excuse to update a room's look with a new one. Fortunately, you don't need an excuse. Upgrading a fixture is easy and can make a dramatic impact on a room. You can substantially increase the light in a room by replacing a globe-style fixture by one with separate spot lights, or you can simply install a new fixture that matches the room's décor.

Tools & Materials ▸

- Replacement light fixture
- Wire stripper
- Voltage sensor
- Insulated screwdrivers
- Wire connectors

Installing a new ceiling fixture can provide more light to a space, not to mention an aesthetic lift. It's one of the easiest upgrades you can do.

How to Replace a Ceiling Light

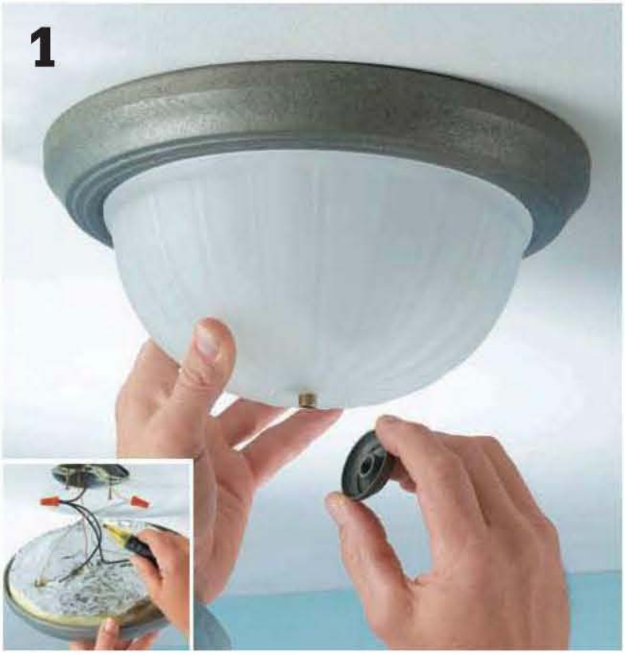

Shut off power to the ceiling light and remove the shade or diffuser. Loosen the mounting screws and carefully lower the fixture, supporting it as you work (do not let light fixtures hang by their electrical wires alone). Test with a voltage sensor to make sure no power is reaching the connections.

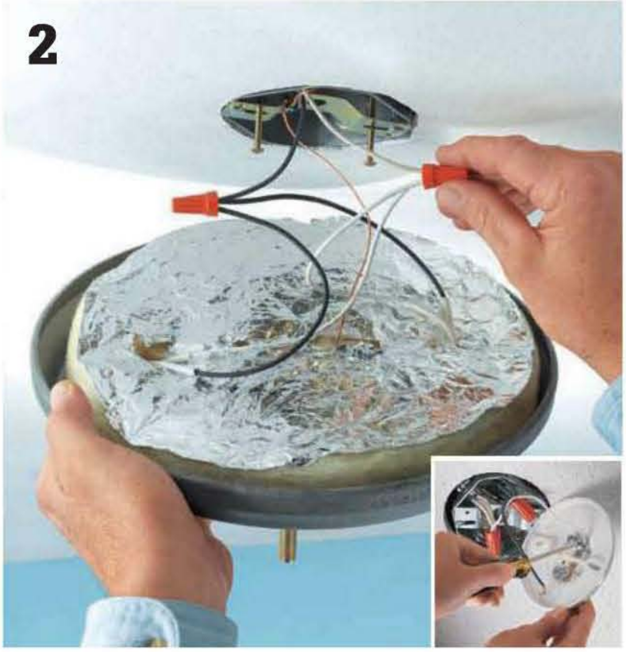

Remove the twist connectors from the fixture wires or unscrew the screw terminals and remove the white neutral wire and the black lead wire (inset).

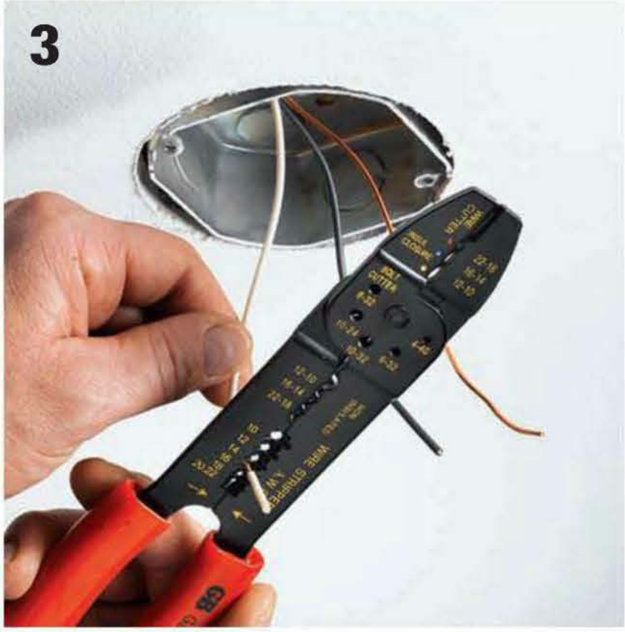

Before you install the new fixture, check the ends of the wires coming from the ceiling electrical box. They should be clean and free of nicks or scorch marks. If they're dirty or worn, clip off the stripped portion with your combination tool. Then strip away about ¾" of insulation from the end of each wire.

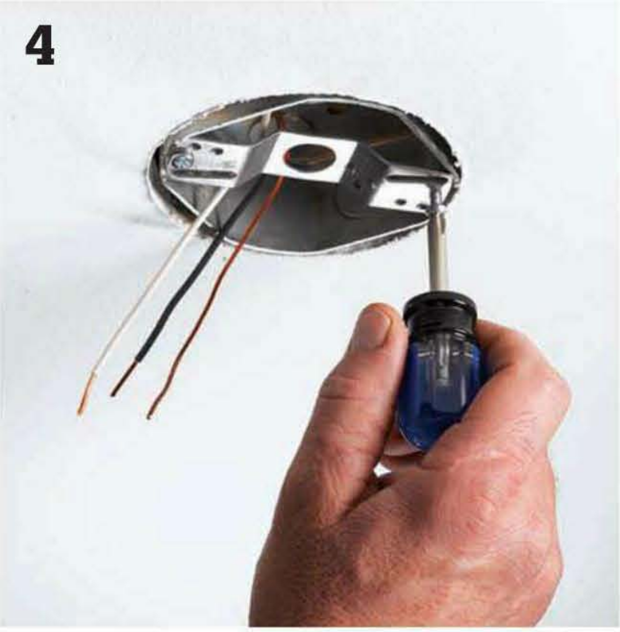

Attach a mounting strap to the ceiling fixture box if there is not one already present. Your new light may come equipped with a strap, otherwise you can find one for purchase at any hardware store.

(continued)

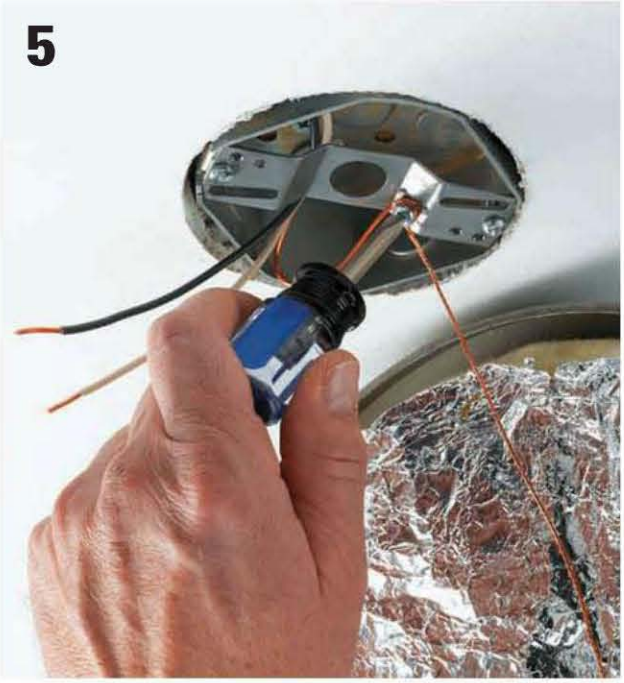
5

Lift the new fixture up to the ceiling (you may want a helper for this) and attach the bare copper ground wire from the power supply cable to the grounding screw or clip on the mounting strap. Also attach the ground wire from the fixture to the screw or clip.

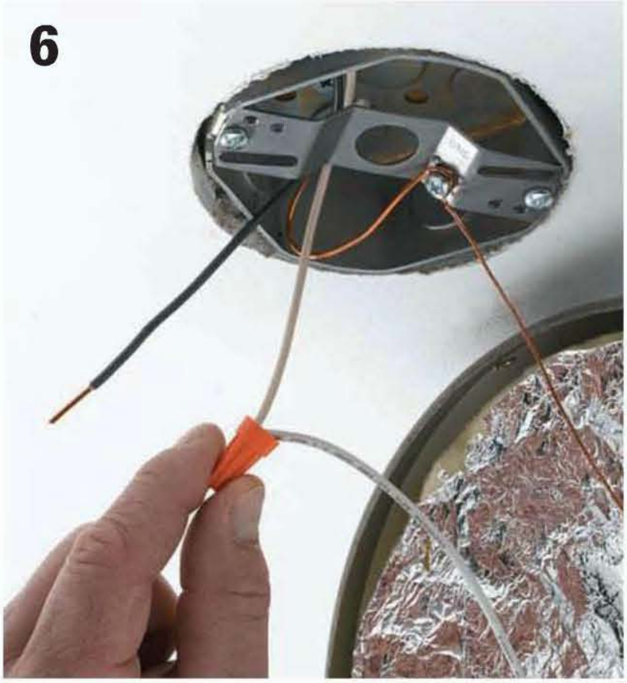
6

With the fixture supported by a ladder or a helper, join the white wire lead and the white fixture wire with a wire connector (often supplied with the fixture).

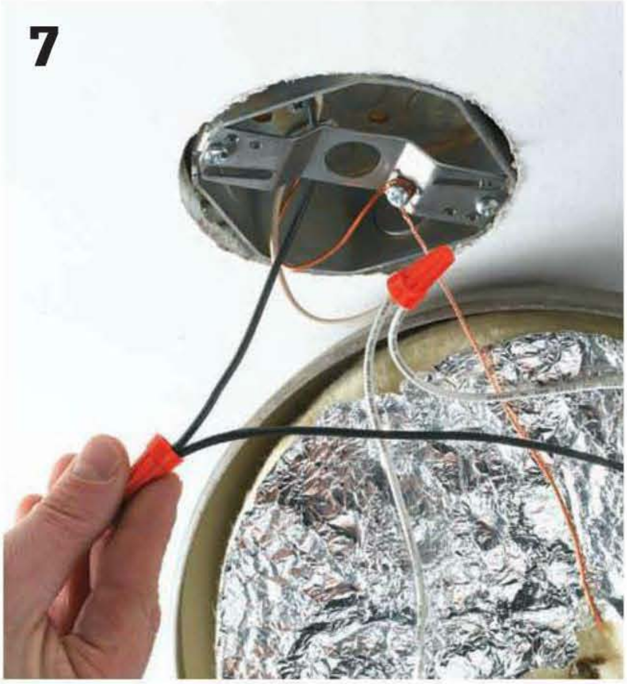
7

Connect the black power supply wire to the black fixture wire with a wire connector.

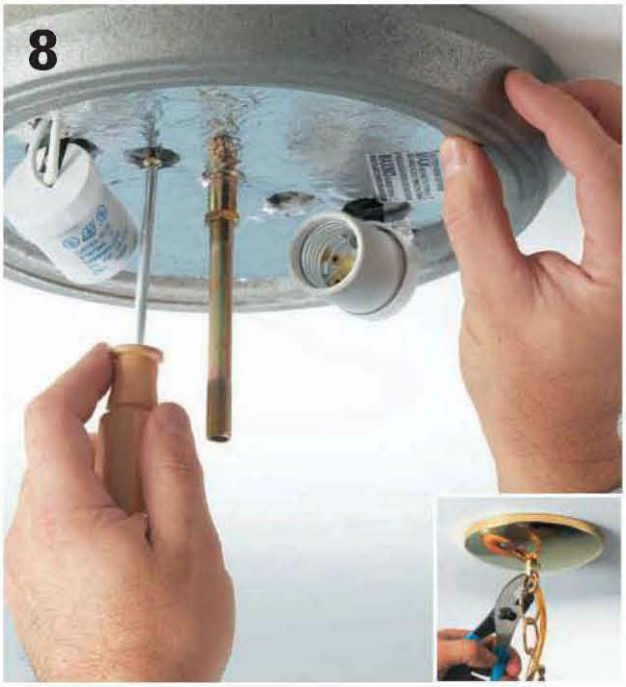
8

Position the new fixture mounting plate over the box so the mounting screw holes align. Drive the screws until the fixture is secure against the ceiling. *Note: Some fixtures are supported by a threaded rod or nipple in the center that screws into a female threaded opening in the mounting strap (inset).*

Installing Recessed Ceiling Lights

Recessed lights are versatile fixtures suited for a variety of situations. Fixtures rated for outdoor use can also be installed in roof soffits and overhangs for accent and security lighting. Recessed fixtures can also be installed over showers or tubs. Be sure to use fixture cans and trims rated for bathroom use.

There are recessed lighting cans in all shapes and sizes for almost every type of ceiling or cabinet. Cans are sold for unfinished ceilings (new construction) or for finished ceilings (retrofit installation). Cans are also rated as insulation compatible or for uninsulated ceilings. Be sure to use the correct one for your ceiling to prevent creating a fire hazard.

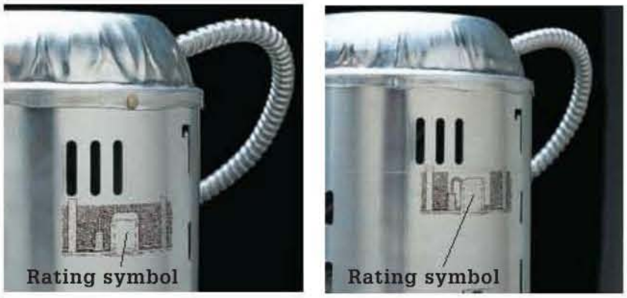

Choose the proper type of recessed light fixture for your project. There are two types of fixtures: those rated for installation within insulation (left), and those which must be kept at least 3" from insulation (right). Self-contained thermal switches shut off power if the unit gets too hot for its rating. A recessed light fixture must be installed at least ½" from combustible materials.

Tools & Materials ▸

- Recessed-lighting can for new construction or remodeling and trim
- Circuit tester
- Cable ripper
- Combination tool
- Pliers
- Fish tape
- Drywall saw
- NM cable

Recessed ceiling lights often are installed in series to provide exacting control over the amount and direction of light. Spacing the canisters in every other ceiling joist bay is a common practice.

How to Install Recessed Ceiling Lights

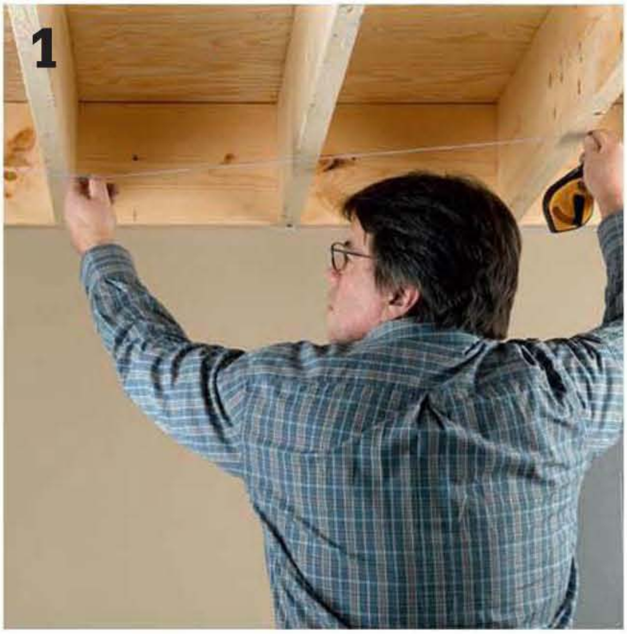

Mark the location for the light canister. If you are installing multiple lights, measure out from the wall at the start and end of the run, and connect them with a chalkline snapped parallel to the wall. If the ceiling is finished with a surface (wallboard), see next page.

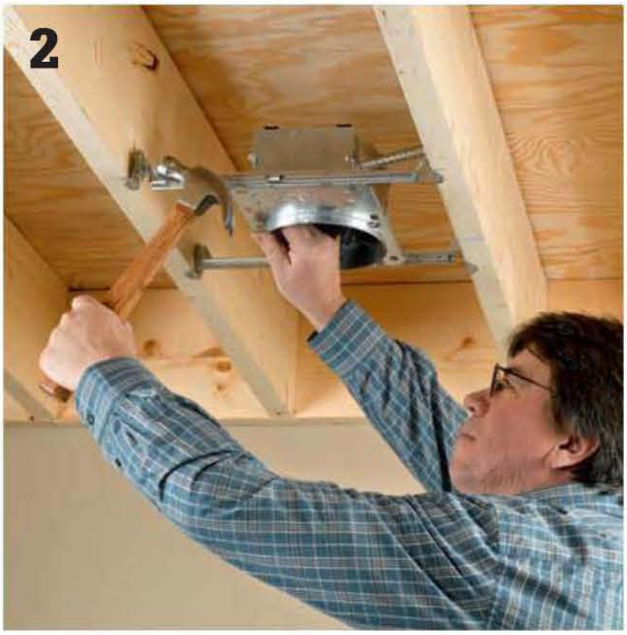

Install the housing for the recessed fixture. Housings for new construction (or remodeling installations where the installation area is fully accessible from either above or below) have integral hanger bars that you attach to the each joist in the joist bay.

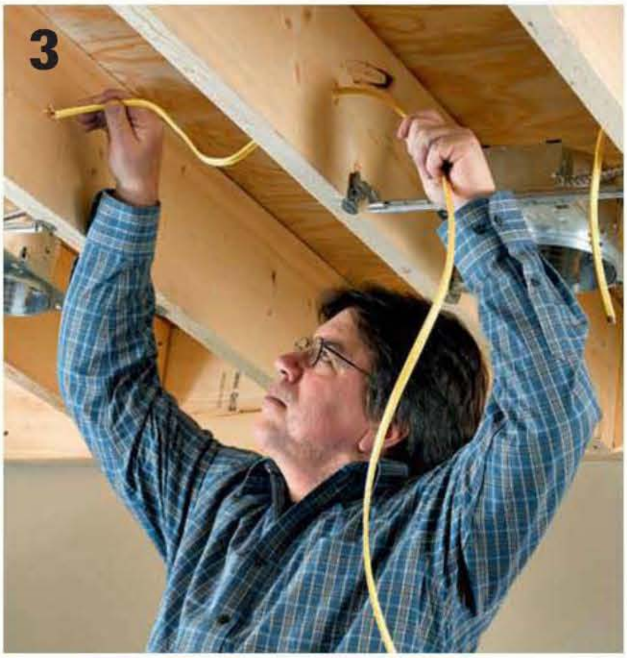

Run electric cable from the switch to each canister location. Multiple lights are generally installed in series so there is no need to make pigtail connections in the individual boxes. Make sure to leave enough extra cable at each location to feed the wire into the housing and make the connection.

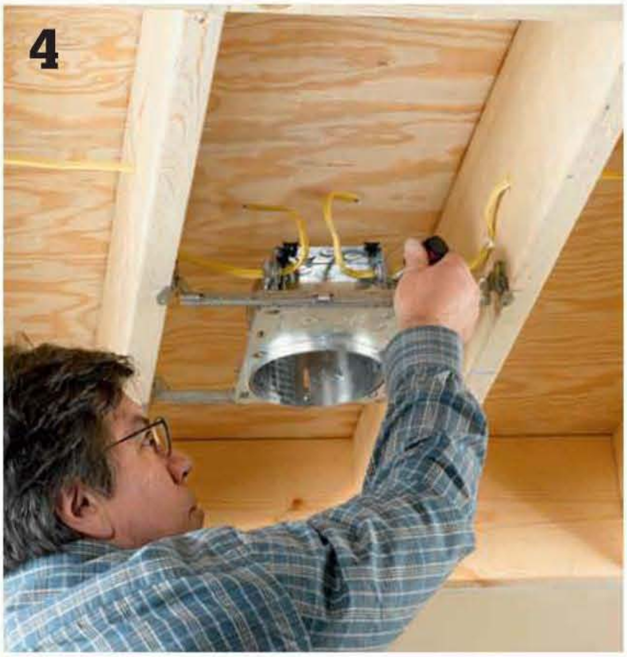

Run the feeder cables into the electrical boxes attached to the canister housings. You'll need to remove knockouts first and make sure to secure the cable with a wire staple within 8" of the entry point to the box.

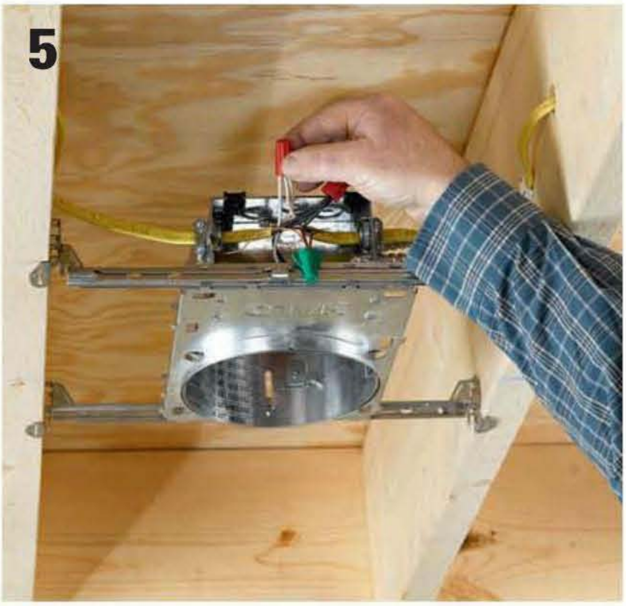

Connect the feeder wires to the fixture wires inside the junction box. Twist the hot lead together with the black fixture wire, as well as the black lead to other fixtures further downline. Also connect the neutral white wires. Join the ground wires and pigtail them to the grounding screw or clip in the box. Finish the ceiling, as desired.

Attach your trim kit of choice. Normally, these are hung with torsion spring clips from notches or hooks inside the canister. This should be done after the ceiling is installed and finished for new construction projects. With certain types of trim kits, such as eyeball trim, you'll need to install the light bulb before the trim kit.

How to Connect a Recessed Fixture Can in a Finished Ceiling

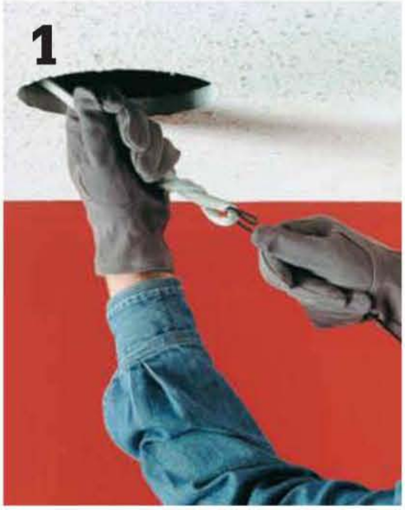

Make the hole for the can. Most fixtures will include a template for sizing the hole. Fish 14/2 cable from the switch location to the hole. Pull about 16" of cable out of the hole for making the connection.

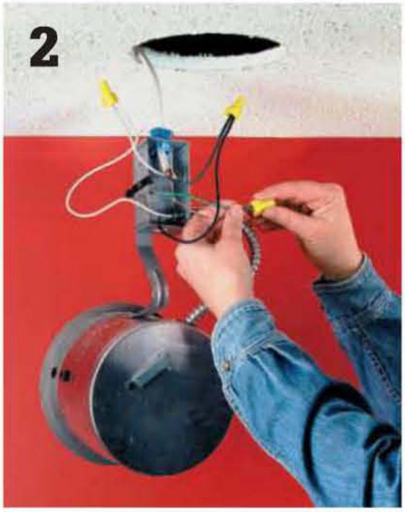

Remove a knockout from the electrical box attached to the can. Thread the cable into the box; secure it with a cable clamp. Remove sheathing insulation. Connect the black fixture wire to the black circuit wire, the white fixture wire to the white circuit wire, and then connect the ground wire to the grounding screw or grounding wire attached to the box.

Retrofit cans secure themselves in the hole with spring-loaded clips. Install the can in the ceiling by depressing the mounting clips so the can will fit into the hole. Insert the can so that its edge is tight to the ceiling. Push the mounting clips back out so they grip the drywall and hold the fixture in place. Install the trim piece.

Installing Track Lighting

Track lighting offers a beautiful and functional way to increase the amount of light in a room or simply to update its look. A variety of fixture and lamp options lets you control the shape, color, and intensity of the light. Installing track lighting in place of an existing ceiling-mounted light fixture involves basic wiring and hand-tool skills, but the connections are even easier to make than with traditional light fixtures. Once installed, the system is very easy to upgrade or expand in the future.

Tools & Materials ▸

- Drill/driver and bits
- Wire stripper
- Screwdriver
- Voltage sensor
- Toggle bolts
- Prewired track and fittings
- Track light heads
- Wire connector
- Ceiling box

If you currently have a ceiling-mounted light fixture that is not meeting your lighting needs, it's simple to replace it with a track-lighting fixture. With track lighting you can easily change the type and number of lights, their position on the track, and the direction they aim. These fixtures come in many different styles, including short 3-ft. track systems with just one or two lights up to 12-ft. systems with five or more lights.

How to Install Track Lighting

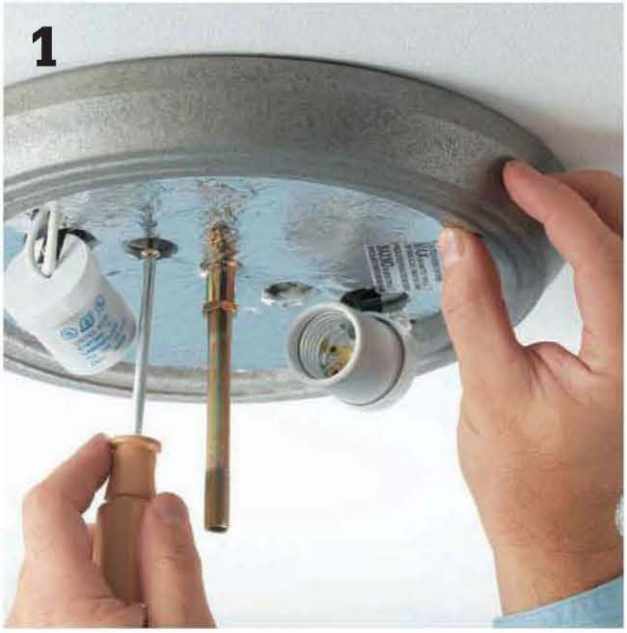

Disconnect the old ceiling light fixture (for remodeling projects) after shutting off power to the circuit at the main service panel. The globe or diffuser and the lamps should be removed before the fixture mounting mechanism is detached.

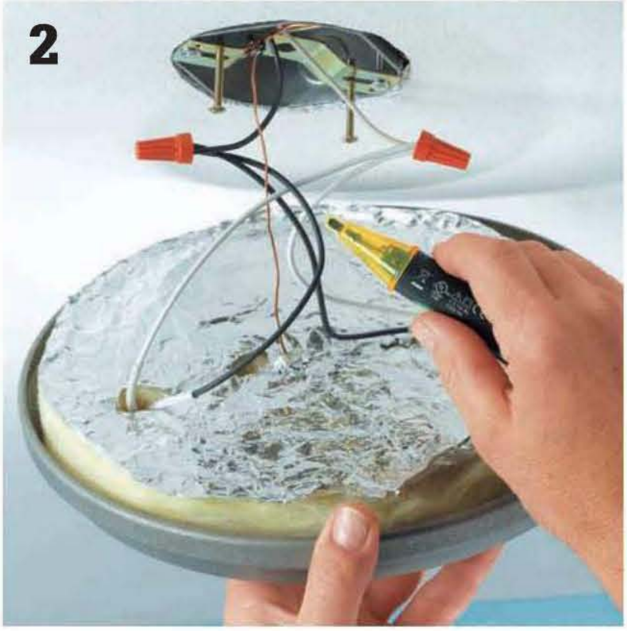

Test the fixture wires with a voltage sensor to make sure the circuit is dead. Support the fixture from below while you work—never allow a light fixture to hang by its electrical wires alone. Remove the wire connectors and pull the wires apart. Remove the old light fixture.

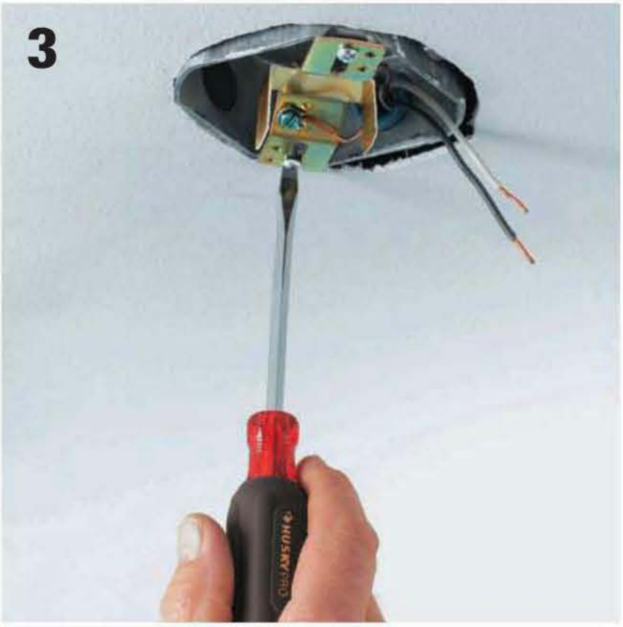

Attach the mounting strap for the new track light to the old ceiling box. If the mounting strap has a hole in the center, thread the circuit wires through the hole before screwing the strap to the box. The green or bare copper ground from the circuit should be attached to the grounding screw or clip on the strap or box.

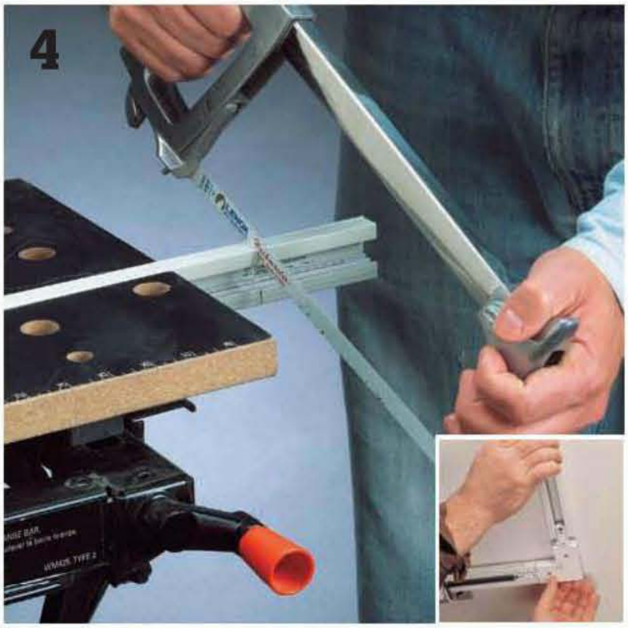

Cut the track section to length, if necessary, using a hack saw. Deburr the cut end with a metal file. If you are installing multiple sections of track, assemble the sections with the correct connector fittings (sold separately from your kit). You can also purchase T-fittings or L-fittings (inset photo) if you wish to install tracks in either of these configurations.

(continued)

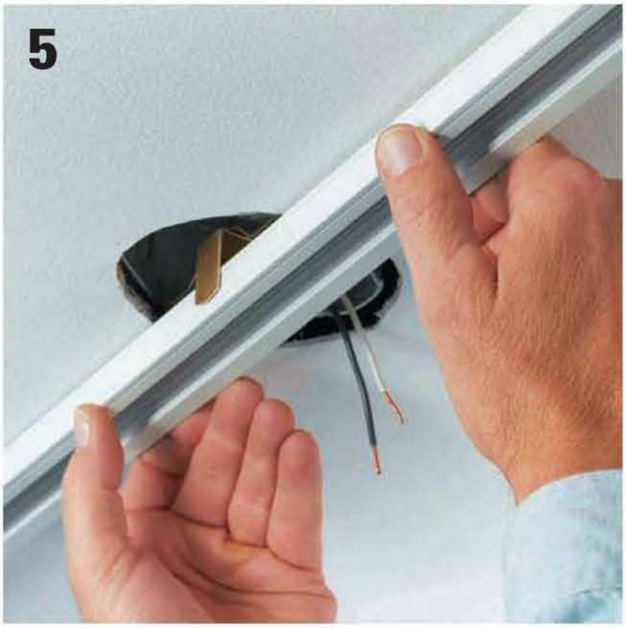

Position the track section in the mounting saddle on the mounting strap and hold it temporarily in place in the location where it will be installed. The track section will have predrilled mounting holes in the back. Draw a marking point on the ceiling at each of these locations. If your track does not have predrilled mounting holes, remove it and drill a 3⁄16" hole in the back every 16".

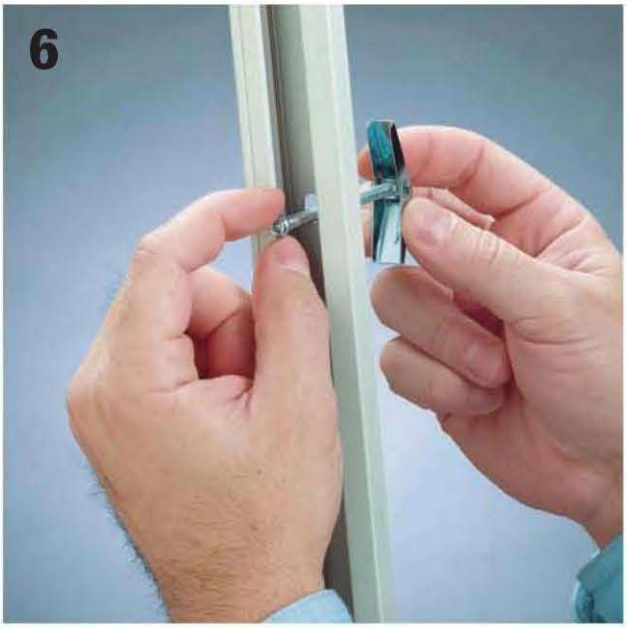

Insert the bolt from a toggle bolt or molly bolt into each predrilled screw location and twist the toggle or molly back onto the free end. These types of hardware have greater holding power than anchor sleeves. Drill a 5⁄8" dia. access hole in the ceiling at each of the mounting hole locations you marked on the ceiling in step 5.

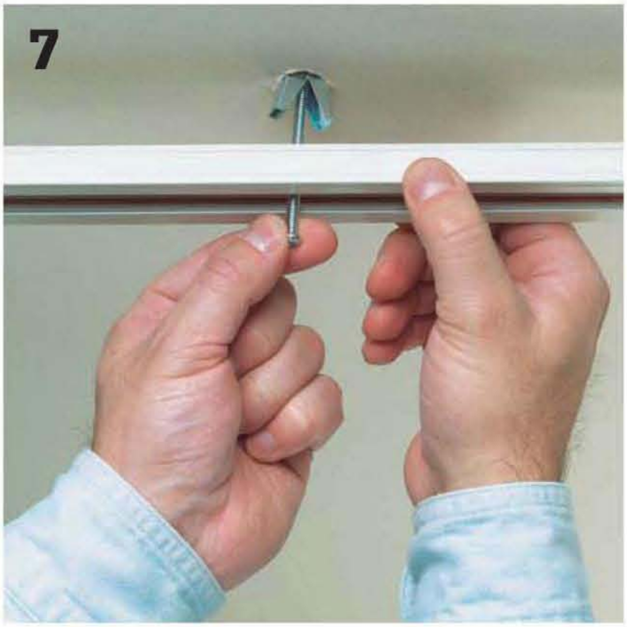

Insert the toggle or molly into the access hole far enough so it clears the top of the hole and the wings snap outward. Then tighten each bolt so the track is snug against the ceiling. If the mounting hole happens to fall over a ceiling joint, simply drive a wallboard screw at that hole location.

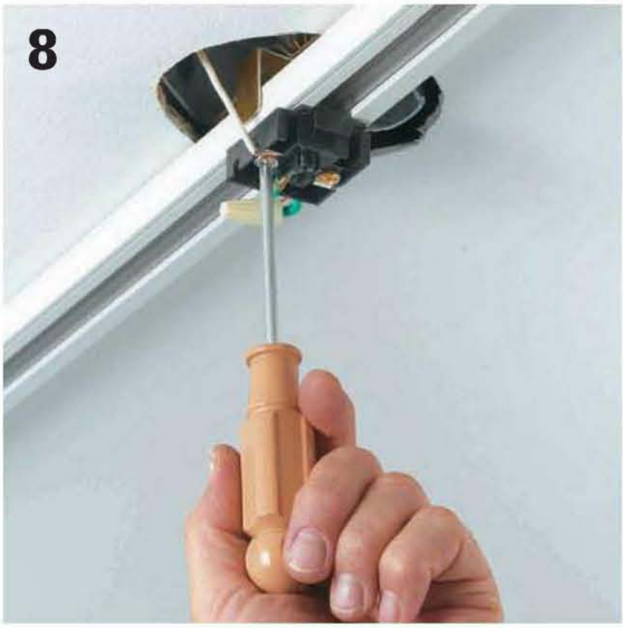

Hook up wires from the track's power supply fitting to the circuit wires. Connect black to black and white to white. The grounding wire from the power supply fitting can either be pigtailed to the circuit ground wire and connected to the grounding screw or clip, or it can be twisted together with the circuit grounding wire at the grounding terminal. Snap the fitting into the track if you have not already done so.

9

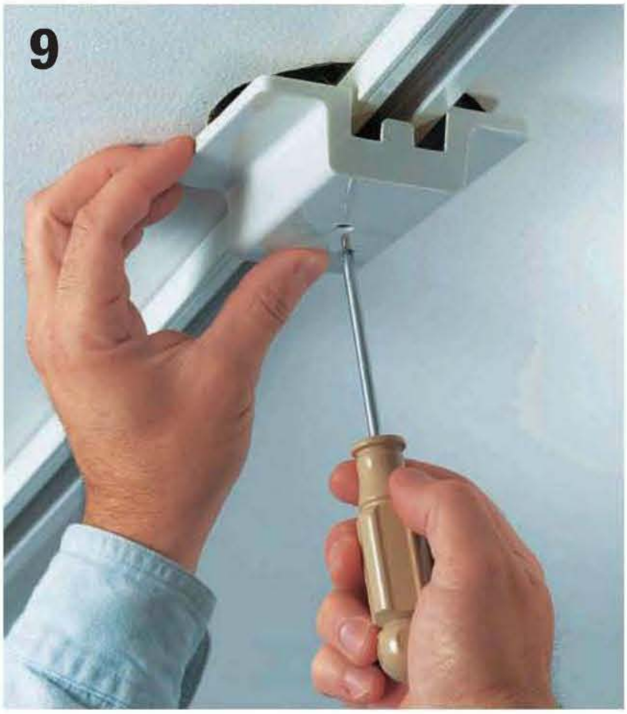

Attach the protective cover that came with your kit to conceal the ceiling box and the electrical connections. Some covers simply snap in place, others require a mounting screw.

10

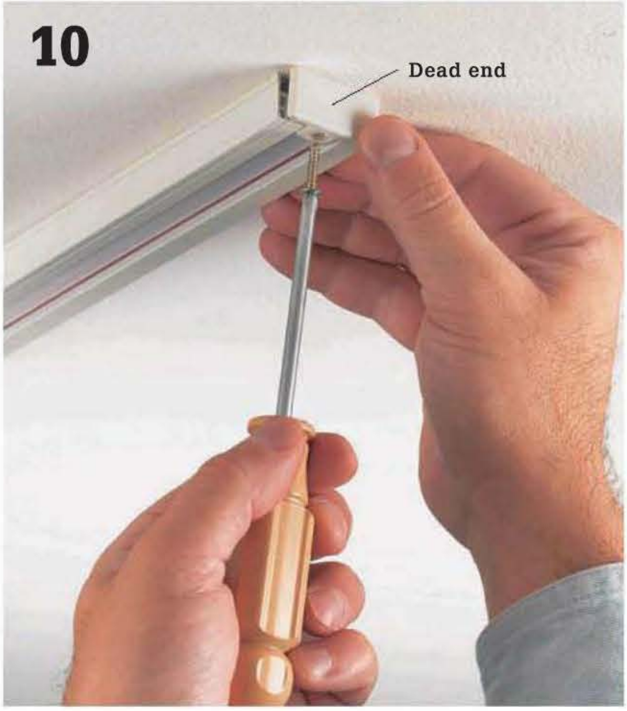

Cap the open ends of the track with a dead end cap fitting. These also may require a mounting screw. Leaving track ends open is a safety violation.

11

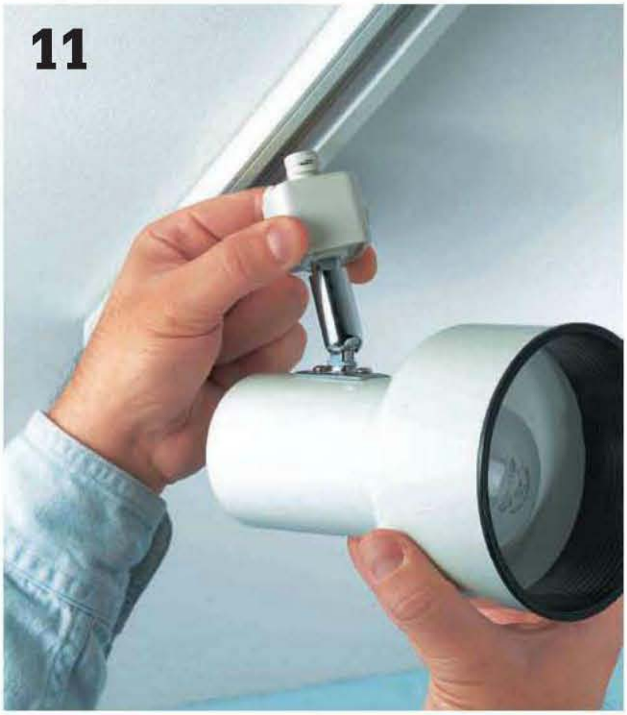

Insert the light heads into the track by slipping the stem into the track slot and then twisting it so the electrical contact points on the head press against the electrified inner rails of the track slot. Tug lightly on the head to make sure it is secure before releasing it.

12

Arrange the track light heads so their light falls in the manner you choose, and then depress the locking tab on each fixture to secure it in position. Restore power and test the lights.

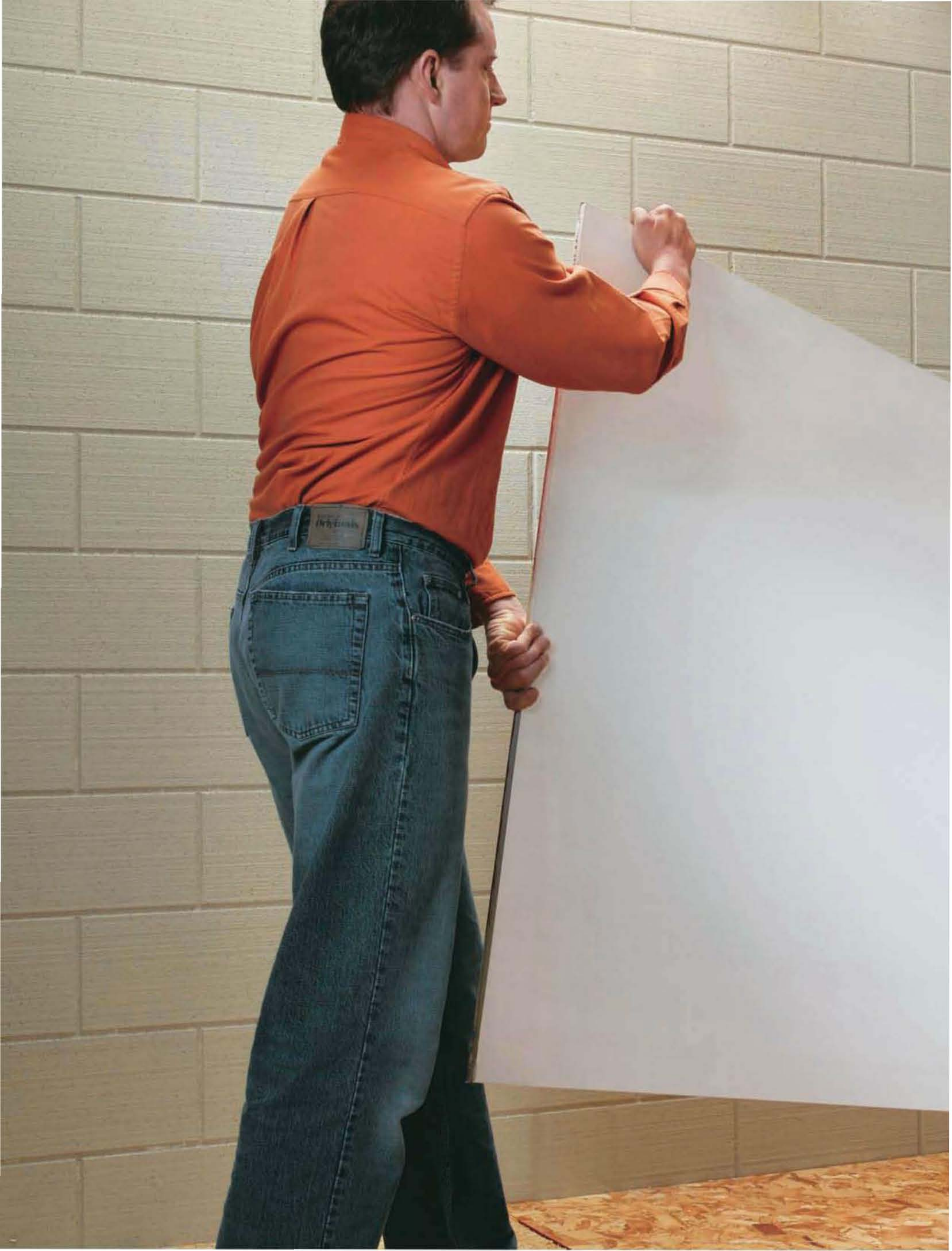

MAJOR REMODELING

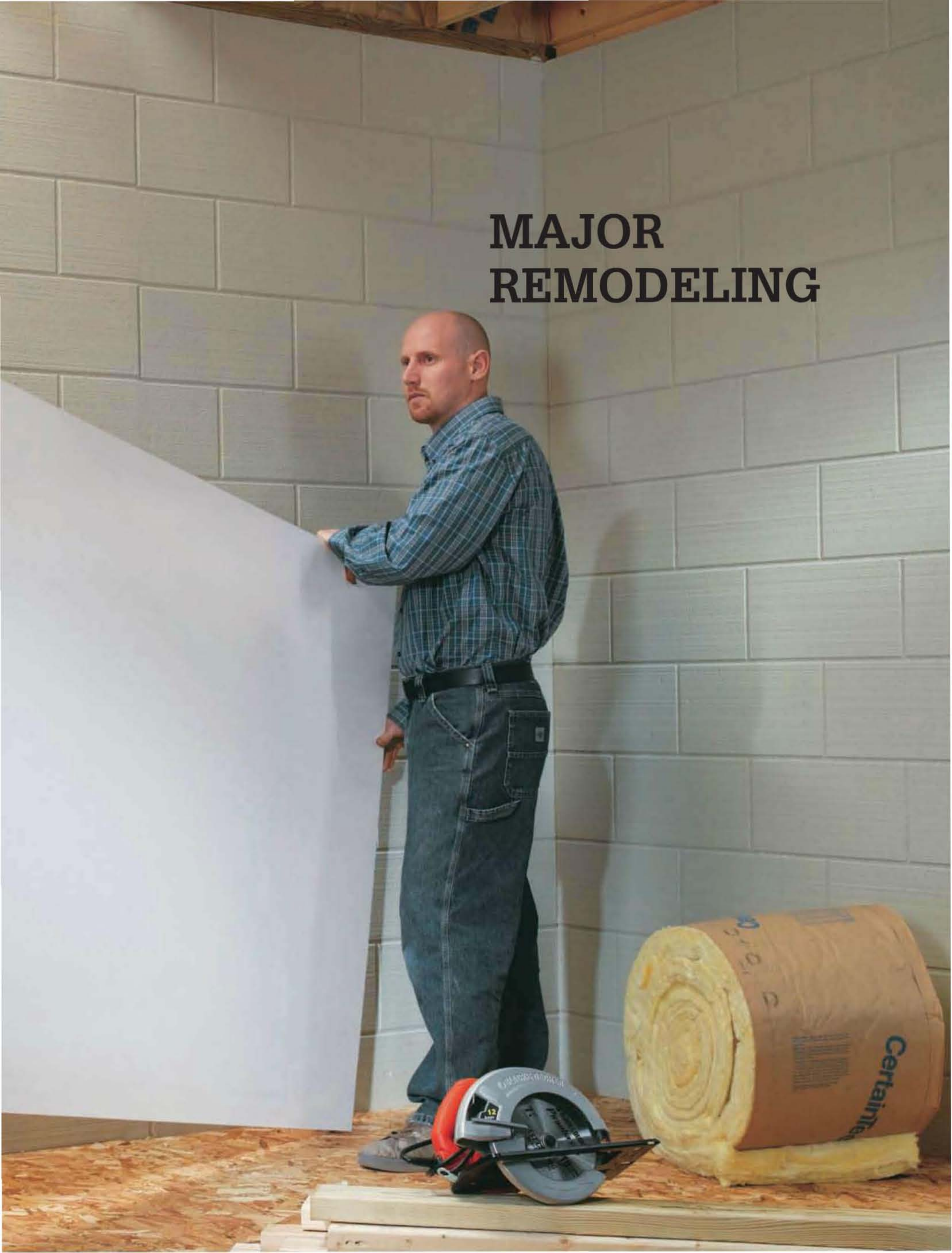

Kitchen Remodeling

Kitchens are remodeled with greater frequency than any other room in the house. This happens often for reasons of style: keeping up with design trends is important to many homeowners. But it also happens for more practical reasons. Kitchens receive a vast amount of wear and tear that results from high usage. Appliances stop working, flooring wears out, cabinets fail, even windows and doors suffer from constant wear. High humidity levels are also an issue in kitchens, where steam is a necessary byproduct of daily life.

The information in this chapter is unique to kitchen remodeling projects. But in practice, you will need to refer to the other chapters in this book to complete all of the design, planning and installation work that go into creating a new kitchen.

In this chapter:

- Planning & Designing Kitchens
- Installing Kitchen Cabinets
- Building Kitchen Islands
- Building Tiled Islands
- Installing Postform Countertops
- Installing Laminate Countertops
- Installing Tile on Countertops
- Installing Tile Backsplashes
- Installing Kitchen Sinks
- Replacing Kitchen Faucets
- Connecting Drains
- Hooking Up Dishwashers
- Installing Food Disposers
- Installing Undercabinet Lights

Planning & Designing Kitchens

Your motives for remodeling your kitchen probably fall into one of two categories: efficiency or appearance. In other words, either your kitchen is inconvenient for your family to use, or you just don't like the way it looks. Or maybe it's both dysfunctional and ugly. A logical place to start, then, is by documenting the elements that need improvement.

Begin by simply taking the time to observe how you now use your kitchen. Keep a notepad handy and jot down any major or minor problems and annoyances that prevent you from being as efficient or as comfortable as you'd like to be when cooking or eating. Also consider how your kitchen is used for special occasions and what issues arise at those times. Don't forget to make note of the positives. If you love having that window over your kitchen sink or the skinny cabinet next to the stove, jot down those observations as well.

Once you have documented the problems with your kitchen, give your imagination permission to roam. Don't worry about money yet. This is your dream kitchen, after all, and there will be plenty of time to bring your fantasies down to earth as you begin the planning stages. Now is the time to consider every possibility. Look at friends' kitchens, at magazine kitchens and at model kitchens. When looking at photographs like the ones included in this book, look at both the overall effect and the individual components. Don't let an odd color scheme scare you away from a faucet that you love. Use a file folder to collect pictures of kitchens or items you like. When it comes down to business, you'll have a refined idea of what you're looking for.

If your kitchen needs remodeling, you will know it. The trick in coming up with a workable design for your space is to get past the "ick" factor and really examine what you have to work with. When you assess your cabinets, you may notice immediately that they are very dark and outdated in appearance. But go past that and ask yourself if they are laid out efficiently. Imagine them as brand new cabinets in the same arrangement, and then try and think of ways you could improve the layout. Whether it's with cabinets or countertops or lighting or appliances, consider the components of your kitchen as generic elements.

Special Considerations

If you're going through the trouble of remodeling your kitchen—and the remodel is not solely for resale value—consider making room for some of your own interests and idiosyncrasies. A pull out recycling center with bins for separating different categories is always a smart idea. A built-in beer refrigerator with a custom tap is great for a home-brew enthusiast. A pull out shelf custom-sized to display a selection of tea is perfectly located next to the kettle. Another, perhaps more practical, investment is the careful measurement of counter heights and seat heights to ergonomically suit your personal needs. Finally, consider how your kitchen will be used over the next five to twenty years. The needs of a family with young children are much different than the needs of empty nesters.

It's okay to hang onto elements you love. Even if your favorite pullout shelf or work table or stool doesn't immediately appear to go with the new elements you hope to introduce, as the owner of the kitchen you're allowed to hang onto the things you're used to working with.

Design Standards

While you may imagine that the answer to every cramped kitchen is to knock down walls and add more space, that's often not the easiest or best solution. One alternative to tearing down the walls, is to add more windows for added light. An extra door, or even a pass-through window from the kitchen to an adjacent room, can also help make the space feel more airy. In short, the cramped feeling may only be one of perception.

Most kitchens fall into one of four categories: Galley, L-shaped, U-shaped and Open Plan. Whether they are small or large, old or new, these floor plans have proven to be popular models for efficient kitchens, though that doesn't mean they'll necessarily be the most efficient for your needs.

Galley: In small homes or city apartments, the galley kitchen is a space-saving choice that is ideal for one or two users. The components may all be lined up along one wall or divided between two parallel walls. For this floor plan to work best, the central galley space should be large enough to allow for all appliances to be open at the same time, with enough space remaining for someone to walk through the middle.

L-shape: This corner kitchen layout can feel roomier than it is in reality, because of the L-shape floor plan. However, the two "arms" might also create an awkward workstation, with little room to set items down mid-way through a meal. Consider taking advantage of the lost central space with a counter-height dining table that can double as a prep station or even a freestanding central island.

U-shape: This layout takes the benefits of a galley kitchen—space-saving solution plus accessibility for one user—and adds a third wall to create the ultimate triangular floor plan. In a compact space, placing the sink at the far end with the refrigerator and range on opposing walls creates a simple workstation. On a larger scale, a sizable island can anchor the center of the room and provide more storage.

Open Plan: Whether the kitchen is located in the middle of a larger great room or off to one side, an open plan layout works best with the addition of a central island or a dividing counter of some type. Perfect for family kitchens or entertaining spaces where the cooks can easily visit with guests, it's important for this floor plan to be well organized so that all necessities are close at hand.

Whatever layout plan you opt for, the most fundamental principle that you don't want to violate is to maintain plenty of free area between appliances, sinks and other elements of the work triangle. In a galley kitchen like the one seen here, the corridor between kitchen walls should be at least 4 ft. wide, and preferably wider.

Common Kitchen Layouts

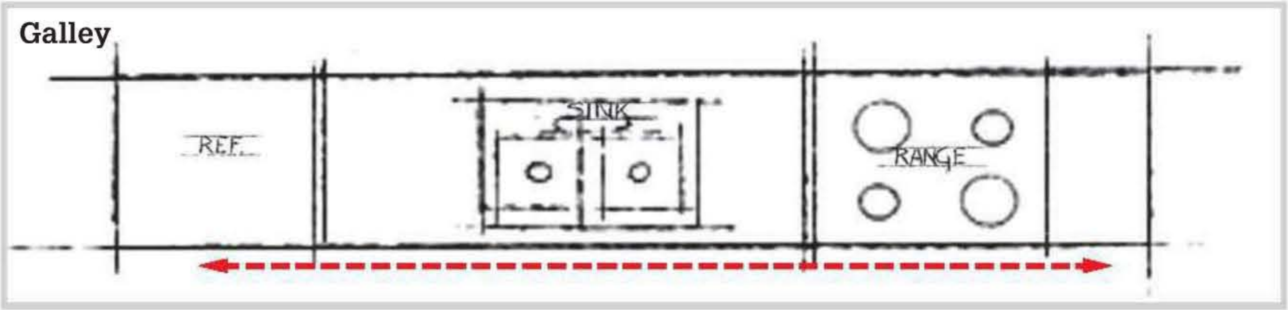
Galley

REF.
SINK
RANGE

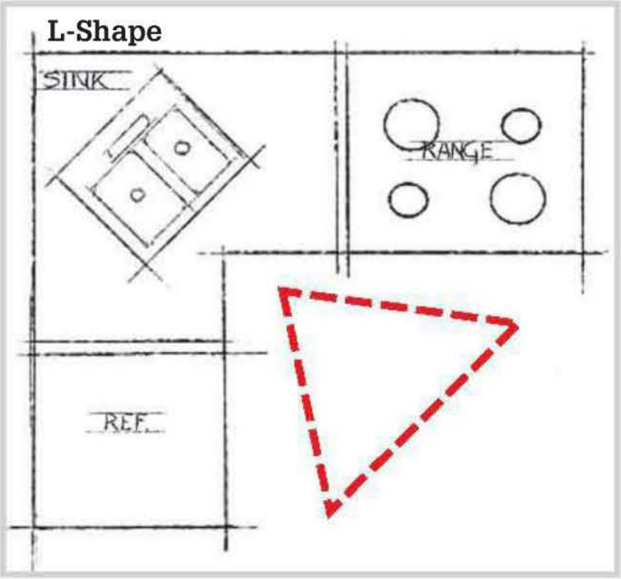
L-Shape

SINK
RANGE
REF.

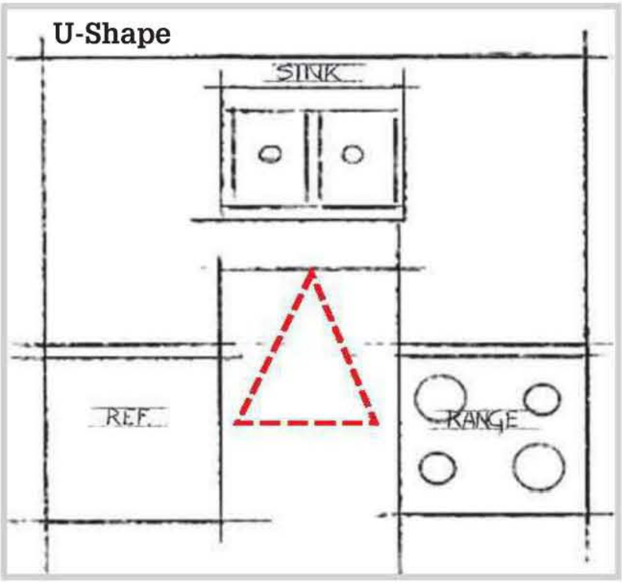
U-Shape

SINK
REF.
RANGE

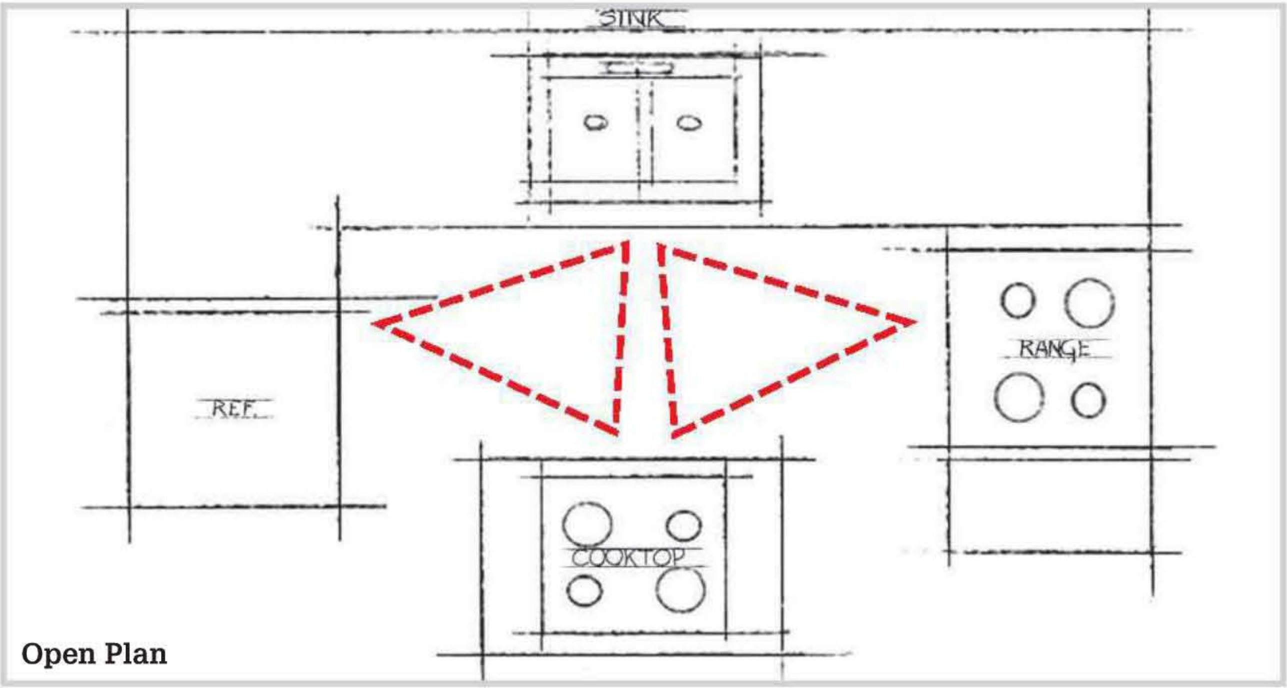
Open Plan

SINK
REF.
RANGE
COOKTOP

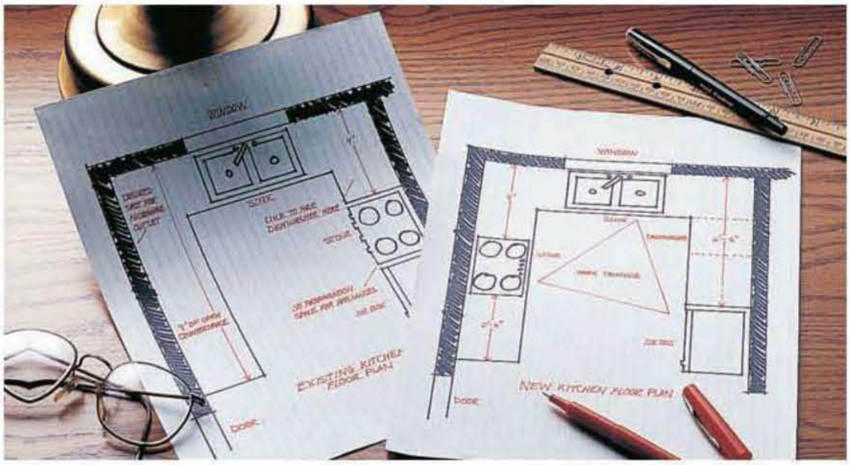

The work triangle is a layout concept that lets you develop a convenient arrangement of the range, sink and refrigerator in the kitchen.

Whether you are doing the work yourself or hiring others, once you have a good idea of the features you want in your new kitchen, it's time to create detailed plan drawings. Good planning drawings will help you in several phases of the planning process:

- Selecting cabinets and appliances to fit your kitchen layout.
- Soliciting accurate work bids when negotiating with plumbers, electricians and other subcontractors.
- Obtaining a building permit at your local Building Inspections office.
- Scheduling the stages of a remodeling project.
- Evaluating the work of contractors. If a carpenter or cabinetmaker fails to meet your expectations, your plan drawings serve as proof that the contractor did not complete the work as agreed.

CODES & STANDARDS

Creating plans for a kitchen can seem like an overwhelming challenge, but fortunately there are guidelines available to help you. Some of these guidelines are legal regulations specified by your local Building Code and must be followed exactly. Most codes have very specific rules for basic construction, as well as for plumbing and electrical installations.

Another set of guidelines, known as standards, are informal recommendations developed over time by kitchen designers, cabinetmakers and appliance manufacturers. These design standards suggest parameters for good kitchen layout, and following them helps ensure that your kitchen is comfortable and convenient to use.

GUIDELINES FOR LAYOUT

The goal of any kitchen layout is to make the cook's work easier and, where possible, to allow other people to enjoy the same space without getting in the way. Understanding the accepted design standards can help you determine whether your present layout is sufficient or if your kitchen needs a more radical layout change or expansion.

Work triangle & traffic patterns. A classic kitchen design concept, the work triangle theory proposes that the sink, range and refrigerator be arranged in a triangular layout according to the following guidelines:

- Position of the triangle should be such that traffic flow will not disrupt the main functions of the kitchen.
- Total distance between the corners of the triangle should be no more than 26 ft. and no less than 12 ft.
- Each side of the triangle should be between 4 and 9 ft. in length.

If two people frequently work in the kitchen simultaneously, the layout should include two work triangles. In a two-triangle kitchen, the triangles may share one side, but they should not cross one another.

Don't fret too much if you can't make the triangle layout work perfectly. Some kitchens, for example, may have four workstations instead of three, and others may not have enough space to accommodate the classic triangle.

For general traffic design, it is recommended to leave 4-ft. "corridors" between all stationary items for walking comfort. Some designers will allow this standard to be reduced to 3 feet in smaller kitchens.

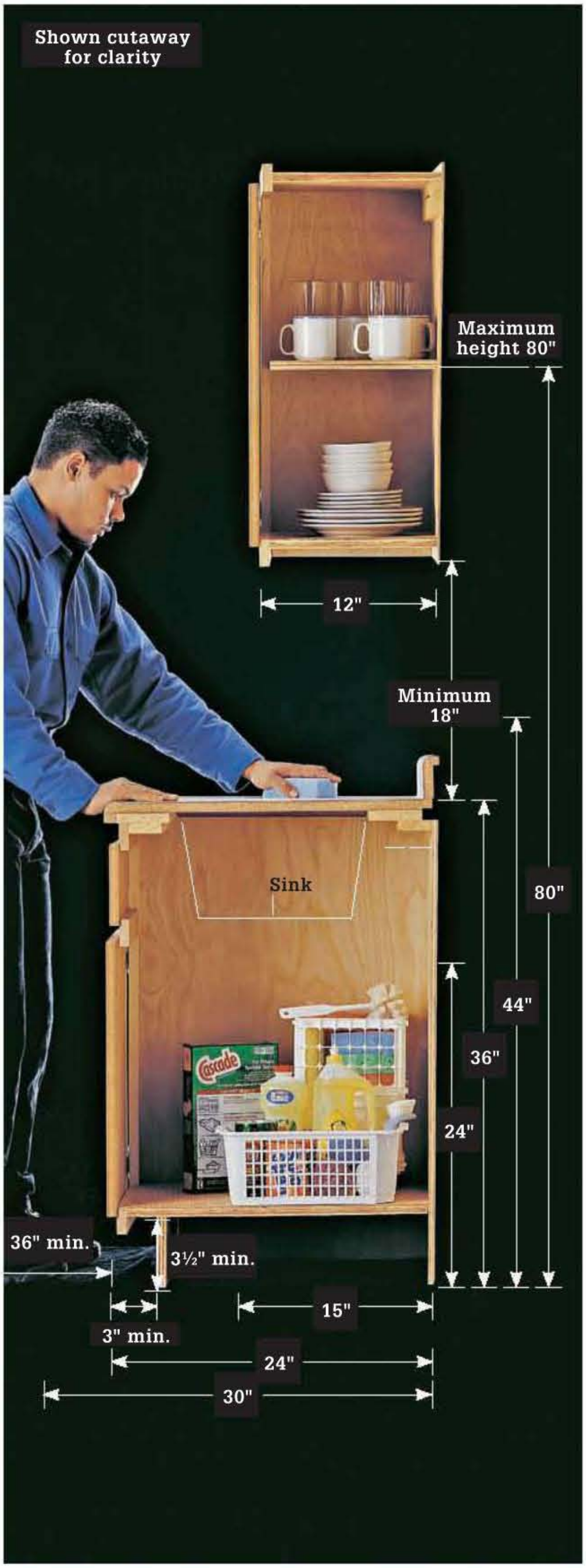

Dimensions and positions of cabinets follow accepted design standards, as shown here.

The sizes of base cabinets and wall cabinets are fairly uniform among manufacturers, and unless you have them custom-built in unusual sizes, they will conform to the following standards:

- Base cabinets: height—34½"; depth—23" to 24"; width—9" to 48", in 3" increments.
- Wall cabinets: height—12", 15", 18", 24", 30", 33", 42"; depth—12"; width—24", 30", 33", 36", 42", 48".
- Oven cabinets: height—84", 96"; depth—24"; width—27", 30", 33".
- Utility cabinets: height—84"; depth—12", 24"; width—18", 24", 36".

Not every manufacturer will offer all these sizes and styles, so it's a good idea to obtain product catalogs when planning the layout of cabinets. Some other tips:

- Use functional corner cabinets rather than "blind" cabinets that provide no access to the corner area.
- Include at least five storage/organizing units, such as swing-out pantry units, appliance garages and specialized drawers or shelves.

Eating areas. Kitchen tabletops and countertops used for dining are generally positioned 30", 36", or 42" above the floor, and the recommended space for each person varies according to the height of the surface.

Islands. A kitchen island should be positioned so there is at least 36" of clear space between the edges of its countertop and surrounding walls or cabinets

GUIDELINES FOR BASIC CONSTRUCTION

Plans for a major remodeling project that involves moving or adding walls, or building a new room addition must accurately show the locations and dimensions of the new walls and all doors and windows. This will allow the construction carpenter to give you an accurate bid on the work and will allow him to obtain the necessary building permits. If you will be moving walls or adding windows or doors, you must identify load-bearing walls and provide appropriate support during removal and rebuilding.

WINDOWS

Most Building Codes require that kitchens have at least one window, with at least 10 sq. ft. of glass area. Some local Building Codes, however, will allow windowless kitchens, so long as they have proper venting. Kitchen designers recommend that kitchens have windows, doors or skylights that together have a total glass surface area equal to at least 25% of the total floor area.

DOORS

Exterior entry doors should be at least 3 ft. wide and 6½ ft. high. Interior passage doors between rooms must be at least 2½ ft. wide. A kitchen must have at least two points of entry, arranged so traffic patterns don't intrude on work areas.

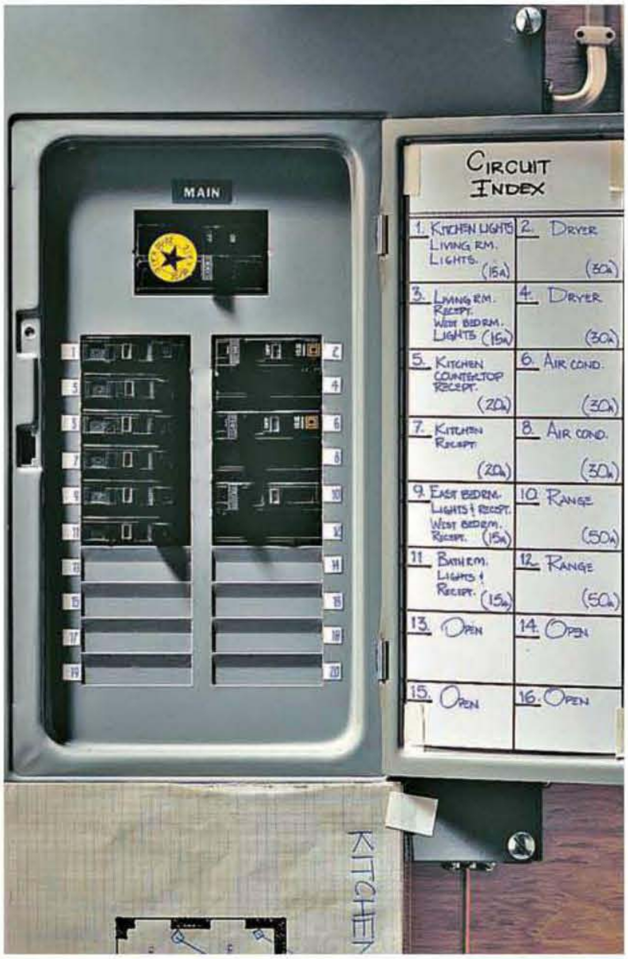
MAIN
CIRCUIT INDEX
1. KITCHEN LIGHTS LIVING RM. LIGHTS. (15A)
2. DRYER (30A)
3. LIVING RM. RECEPT. WEST BEDRM. LIGHTS (15A)
4. DRYER (30A)
5. KITCHEN COUNTERTOP RECEPT. (20A)
6. AIR COND. (30A)
7. KITCHEN RECEPT (20A)
8. AIR COND. (30A)
9. EAST BEDRM. LIGHTS & RECEPT. WEST BEDRM. RECEPT. (15A)
10. RANGE (50A)
11. BATHRM. LIGHTS & RECEPT. (15A)
12. RANGE (50A)
13. OPEN
14. OPEN
15. OPEN
16. OPEN
KITCHEN

Examine your circuit breaker panel. It may have an index that identifies circuits serving the kitchen. If your service panel has open slots, an electrician can add additional kitchen circuits relatively easily. If your service panel is full, he may have to install a new service panel at additional cost.

GUIDELINES FOR ELECTRICAL SERVICE & LIGHTING

Nearly any kitchen-remodeling project will require some upgrading of the electrical service. While your old kitchen may be served by a single 120-volt circuit, it's not uncommon for a large modern kitchen to require as many as seven individual circuits. And in a few cases, the extra demands of the new kitchen may require that the main electrical service for your entire house be upgraded by an electrician. By comparing the electrical service in your present kitchen with the requirements described below, you'll get an idea of how extensive your electrical service improvements will need to be. Your plan drawings should indicate the locations of all the outlets, lighting fixtures and electrical appliances in your new kitchen.

The National Electric Code requires the following for kitchens:

- Two small-appliance circuits (120-volt, 20-amp) to supply power for the plug-in countertop appliances.
- Wall outlets spaced no more than 12 ft. apart.
- Countertop outlets spaced no more than 4 ft. apart.
- GFCI (ground-fault circuit interrupter), protected receptacles installed in any general use outlet, whether above counter or at floor level.
- Dedicated circuits for each major appliance. Install a 20-amp, 120-volt circuit for a built-in microwave, a 15-amp circuit for the dishwasher and food disposer. An electrical range, cooktop or wall oven requires a dedicated 50-amp, 240-volt circuit.

The Electric Code only requires that a kitchen have some form of lighting controlled by a wall switch, but kitchen designers have additional recommendations:

- A general lighting circuit (120-volt, 15-amp) that operates independently from plug-in outlets.
- Plentiful task lighting, usually mounted under wall cabinets or soffits, to illuminate each work area.
- Decorative lighting fixtures to highlight attractive cabinets or other features of the kitchen.

GUIDELINES FOR PLUMBING

If your new kitchen layout changes the location of the sink, or if you are planning to add an additional sink or dishwasher, the water supply and drain pipes will need to be upgraded. Your plan drawings should indicate these intended changes.

Extending plumbing lines for a new kitchen is often fairly easy and surprisingly inexpensive, but there are some exceptions you should note:

Old pipes. If your present plumbing is more than 25 years old, there is a good chance the plumber will recommend replacing the pipes before installing the kitchen fixtures. Depending on circumstances, this can be an expensive proposition, but if you're faced with this decision, we strongly urge you to take a deep breath and do what the plumber suggests. Those corroded old pipes will need to be replaced someday, and this work is easier and cheaper if you're already in the process of remodeling the kitchen.

Outdated systems. Older plumbing systems may have drain trap and vent arrangements that violate modern Code requirements. If your plumber needs to run all-new vent pipes, this will increase the costs.

Island sinks. If your new kitchen will include an island sink, your plumber will need to run vent pipes beneath the floor. For this reason, plumbing an island sink is more expensive than plumbing a wall sink.

GUIDELINES FOR HEATING, VENTILATION & AIR-CONDITIONING

Your plan drawings should also show the locations of heating/air-conditioning registers or fixtures in your proposed kitchen. If you're planning a cosmetic make-over or a simple layout change, there is a pretty good chance you can get by with the same registers, radiators or heaters found in your present kitchen.

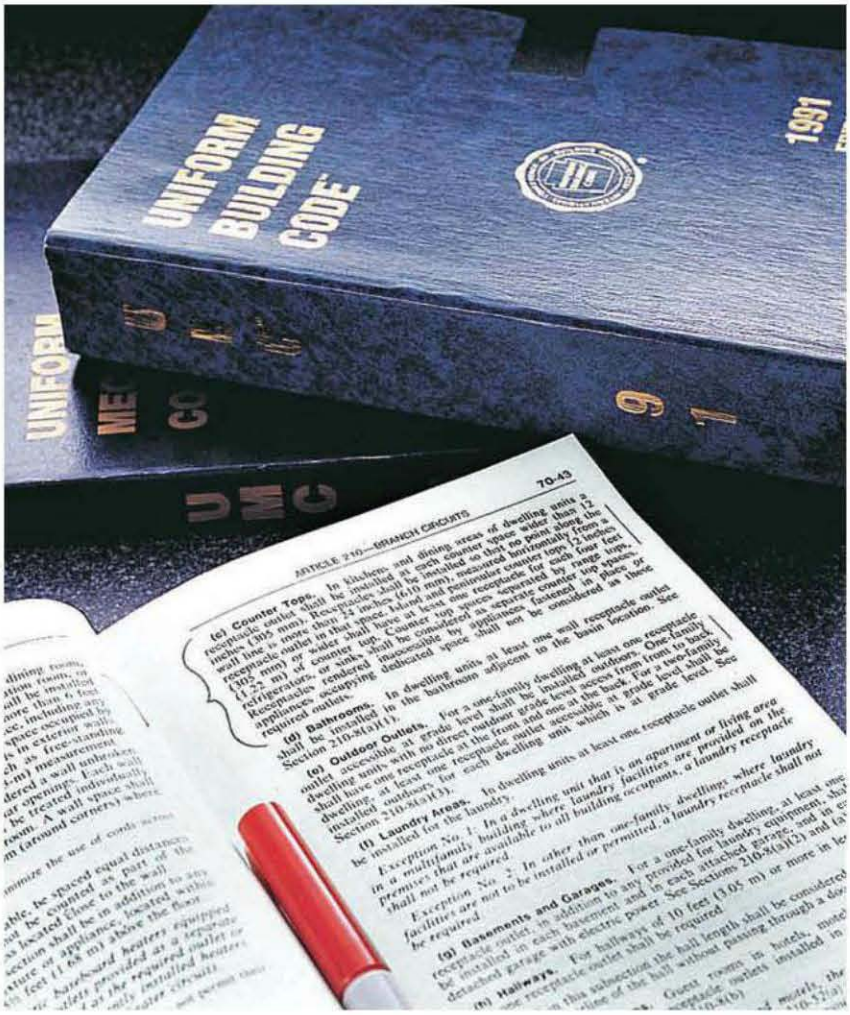

Code books can help you understand the structural, electrical and plumbing requirements for kitchens. In addition to the formal Code books, which are written for professional tradesmen, there are many Code handbooks available that are written for homeowners. Bookstores and libraries carry both the formal Code books and Code handbooks.

But if your new kitchen will be substantially larger than it is now, or if the ratio of wall space filled by glass windows and doors will be greater, it's possible that you'll need to expand its heating and cooling capacity.

Increasing your kitchen's heating and cooling can be as simple as extending ducts by a few feet, or as complicated as installing a new furnace. When installing a large room addition, for instance, you may learn that the present furnace is too small to adequately heat the increased floor space of your home.

How do you determine what your kitchen needs in the way of expanded heating and cooling? Unless you happen to be a mechanical engineer, you'll need to consult a professional to evaluate your heating/ventilation/air-conditioning (HVAC) system. The Code requirements for room heating are quite simple, but the methods used to calculate required energy needs of a room are fairly complex.

The Building Code requires simply that a room must be able to sustain a temperature of 70°F, measured at a point 3 ft. above the floor. HVAC contractors use a complicated formula to calculate the most efficient way to meet this Code requirement. You can make this estimation more accurate by providing the following information:

- The exact dimensions of your kitchen.
- The thickness and amount of insulation in the walls.
- The number of doors and windows, including their size and their energy ratings.
- The total square footage of your house.
- The heating and cooling capacity of your furnace and central air-conditioner, measured in BTUs.

This information, usually printed on the unit's access panel, will help the HVAC contractor determine if the system can adequately serve your new kitchen.

Finally, your cooktop should be equipped with an electric vent hood to exhaust cooking fumes and moisture from the kitchen. The volume of air moved by a vent fan is restricted by Code, so you should always check with a Building Inspector before selecting a vent hood.

Metal ductwork for the vent hood must be run through an exterior wall or through the ceiling. If your cooktop is located in an island cabinet, a special island vent fan is necessary.

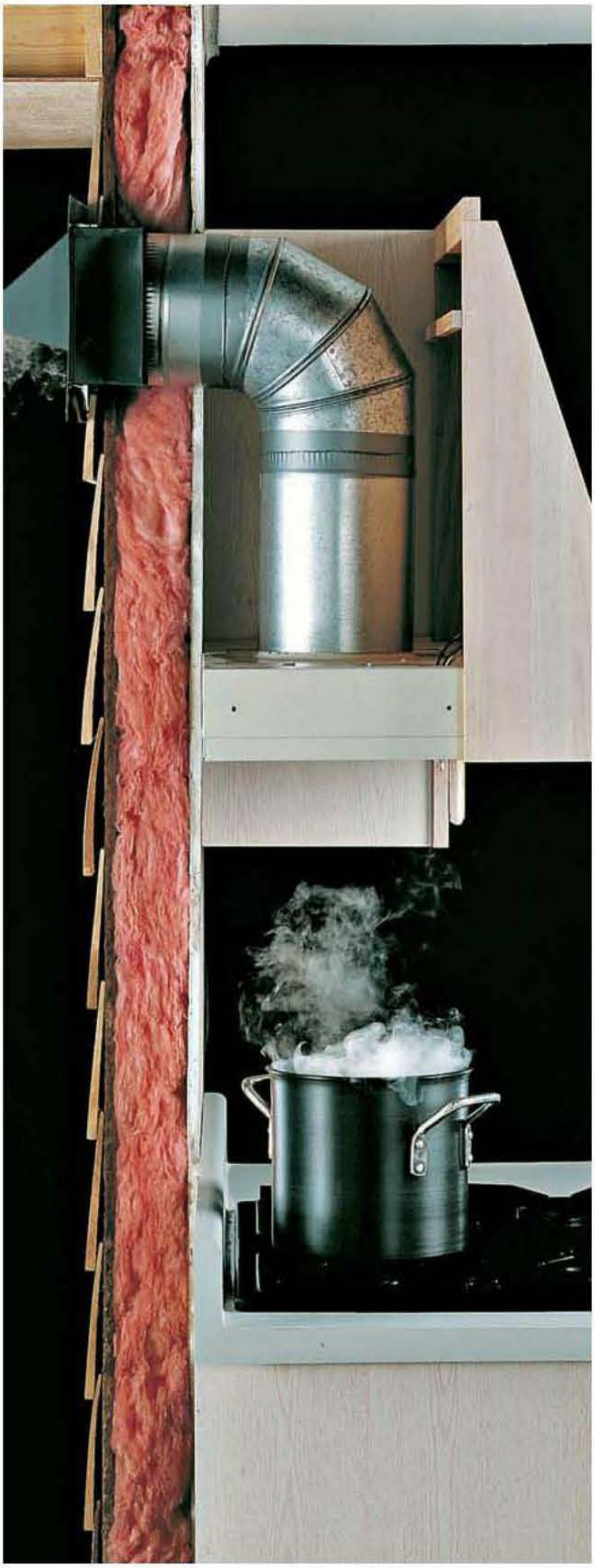

Vent hoods are required by some local codes on all ranges and cooktops. The vent fan exhausts cooking fumes and moisture to the outdoors.

Energy Efficiency

For the sake of both environmental resources and your own budget, selecting a kitchen design that incorporates energy- and water-saving appliances is a good idea. Most of the new appliances on the market are more energy efficient than older models, but the relative quality of their performance varies. Do some comparison-shopping when it comes to appliances and you are sure to outfit your kitchen with fixtures that save energy and money.

Lighting is another way to save energy. Consider the aesthetic benefits of adding a skylight or window to your kitchen; now think of the impact that natural lighting will have on your energy budget. Less artificial light will be needed, thus lowering electric bills.

Another way to cut down on energy requirements is to use fluorescent lighting or new innovations like LED lighting in the kitchen. These products not only reduce glare better than incandescent bulbs, they also use a fraction of the energy used by incandescent bulbs and need to be replaced less frequently. Newer versions of fluorescent bulbs produce a more true-to-life light color, rather than the old buzzing overhead tubes that gave everything a greenish hue. Some local building codes require a certain percentage of a kitchen's lighting to be fluorescent, so consider the benefits before lamenting this ordinance.

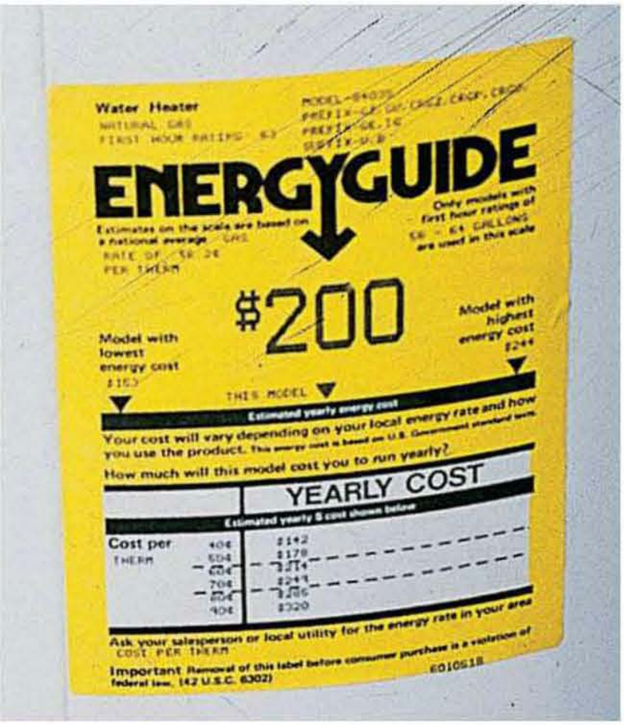

Energy usage labels can be found on most major appliances you'll encounter at the appliance store. More energy efficient models may qualify for a rebate from your local public utility company. Ask the sales staff for information on these programs.

As a general rule, newer appliances are more efficient when it comes to energy and water consumption. For example, this dishwasher has a feature that allows the user to wash only one of the racks, using less water and energy when smaller loads are being cleaned.

Accessible Kitchens ▸

Universal Design and Accessible Design are the terms used by architects, engineers, and product designers to describe making accommodations for the needs of every type of person, regardless of age and physical ability. For example, large handled tools, contrasting typeface, wide doorways, and smooth flooring are all results of Universal Design research. In the kitchen, the consideration of accessibility must be more extensive than almost anywhere else in the home. From safety lighting, to pull out shelves, to the height of counters, planning a kitchen that everyone can easily use requires careful forethought. There are tips on Universal Design throughout this book, offering advice on accessibility for many of the standard kitchen elements.

Universal design for kitchen cabinets puts the majority of items in the comfortable reach zone between 2 and 5 feet above the floor. Using pop-up and pull-down shelves can extend this area. Full-extension hardware on drawers and pull-out shelves eliminates reaching and fumbling for unseen items at the back of drawers and cabinets. Base cabinets installed at various heights serve users of different heights. Hanging some sections of upper cabinets at 12" to 15" above the countertop—rather than 18"—makes it easier to see and access items.

A full-length pantry also increases the amount of storage space in the comfortable reach zone. A pull-out pantry allows for easier viewing of the contents, but make sure the hardware is top quality and operates easily and smoothly. An appliance garage, a cabinet with a tambour door that sits on the countertop, is an excellent way to efficiently use corner space and store heavy or frequently used appliances out of sight. Make sure an electrical outlet or two are inside the garage.

UNIVERSAL TIPS

- Magnetic touch latches or c-shaped handles are the easiest to open.
- Avoid gloss cabinet finishes to reduce glare for the vision impaired.
- Countertops at a variety of heights should be ergonomically correct for a range of different users.

Countertop height and ease of fixture operation are two of the most important aspects of universal design. If your kitchen remodeling project needs to accommodate a person with special needs, most states have agencies that can offer you specific advice for your situation.

Universal Design

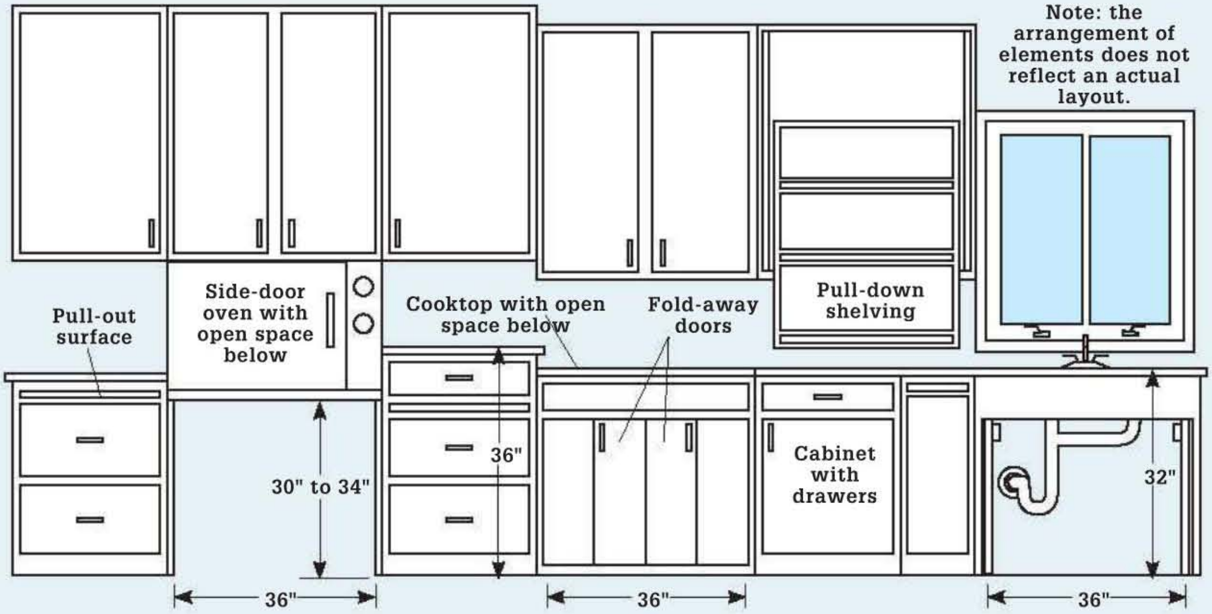
Note: the arrangement of elements does not reflect an actual layout.
Pull-out surface
Side-door oven with open space below
Cooktop with open space below
Fold-away doors
Pull-down shelving
30" to 34"
36"
Cabinet with drawers
32"
36"
36"
36"

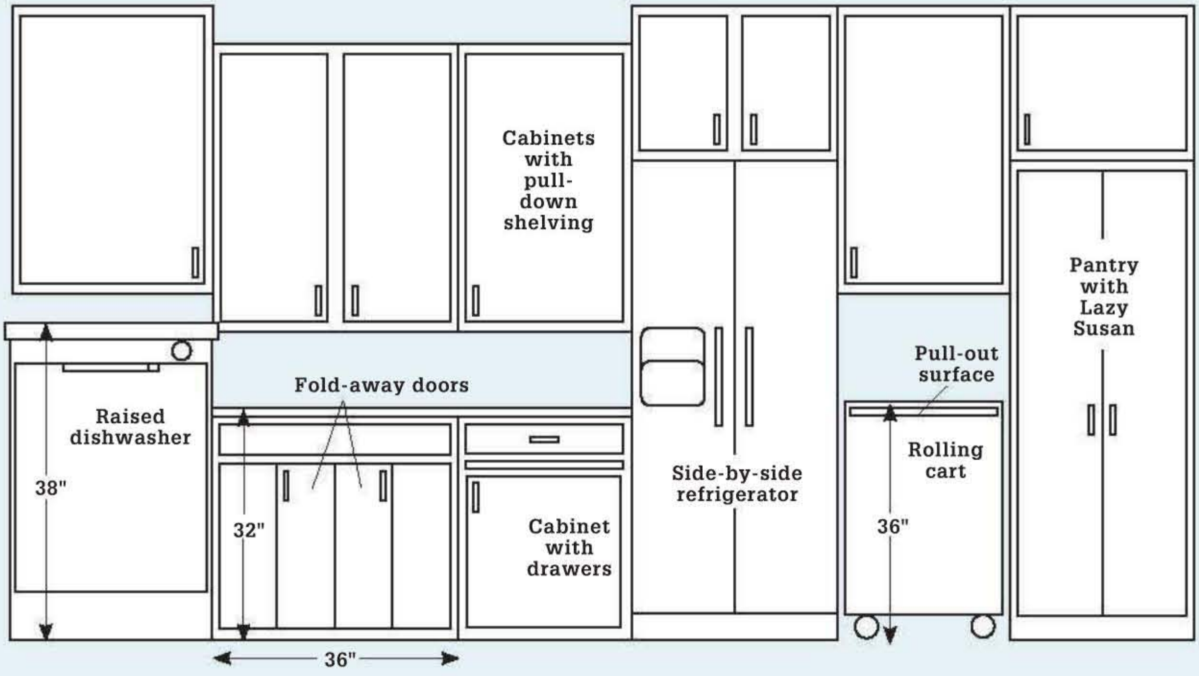
Cabinets with pull-down shelving
Pantry with Lazy Susan
Pull-out surface
Fold-away doors
Raised dishwasher
Rolling cart
Side-by-side refrigerator
38"
32"
Cabinet with drawers
36"
36"

Design your kitchen around a clear, circular space of at least 5 ft. in diameter to provide room for a wheelchair. If your kitchen doesn't have 60" of clear space, allow 48" for pathways. Plan for 30 to 48" of clear approach space in front of all appliances and workstations.

Installing Kitchen Cabinets

Cabinets must be firmly anchored to wall studs, and they must be plumb and level when installed. The best way to ensure this is by attaching a ledger board to the wall to assist in the installation. As a general rule, install the upper cabinets first so your access is not impeded by the base cabinets. (Although some pros prefer to install the base cabinets first so they can be used to support the uppers during installation.) It's also best to begin in a corner and work outward from there.

Tools & Materials ▸

Handscrew clamps
Level
Hammer
Utility knife
Nail set
Stepladder
Drill
Counterbore drill bit
Cordless screwdriver
Jigsaw

Cabinets
Trim molding
Toe-kick molding
Filler strips
Valance
6d finish nails
Finish washers
#10 × 4" wood screws
#8 × 2½" screws
3" drywall screws

Stock cabinets are sold in boxes that are keyed to door and drawer packs (you need to buy these separately). It is important that you realize this when you are estimating your project costs at the building center (often a door pack will cost as much or more than the cabinet). Allow plenty of time for assembling the cabinets out of the box. It can take an hour or more to put some more complex cabinets together.

How to Fit a Corner Cabinet ▸

Before installation, test-fit corner and adjoining cabinets to make sure doors and handles do not interfere with each other. If necessary, increase the clearance by pulling the corner cabinet away from the side wall by no more than 4". To maintain even spacing between the edges of the doors and the cabinet corner, cut a filler strip and attach it to the corner cabinet or the adjoining cabinet. Filler strips should be made from material that matches the cabinet doors and face frames.

How to Prepare Walls

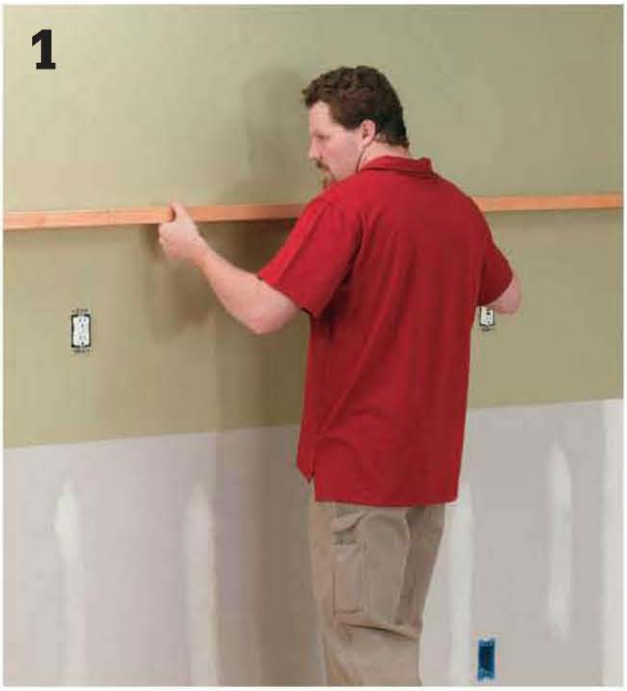

Find high and low spots on wall surfaces, using a long, straight 2 × 4. Sand down any high spots.

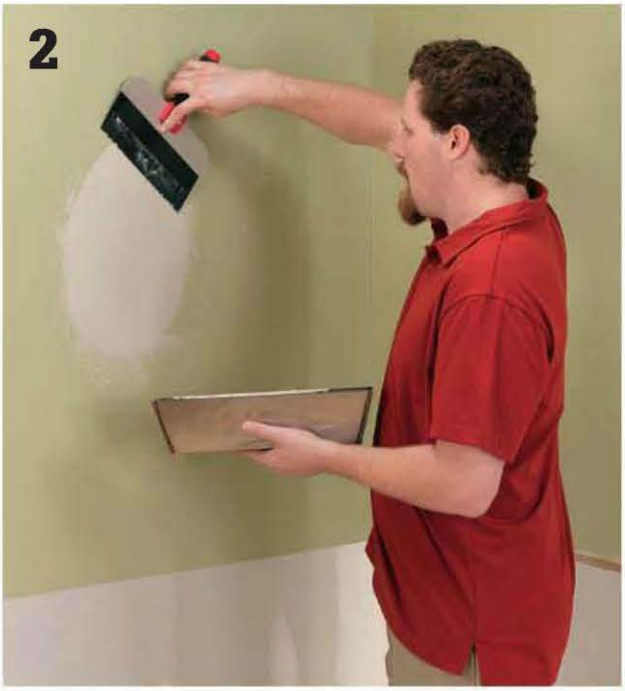

Fill in low spots of wall by applying wallboard compound using a taping knife. Let the compound dry, and then sand it lightly.

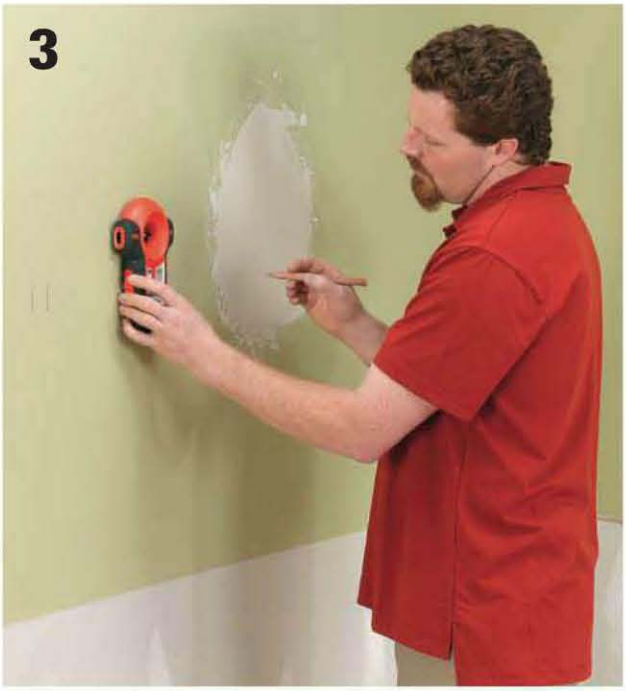

Locate and mark wall studs, using an electronic stud finder. Cabinets normally will be hung by driving screws into the studs through the back of the cabinets.

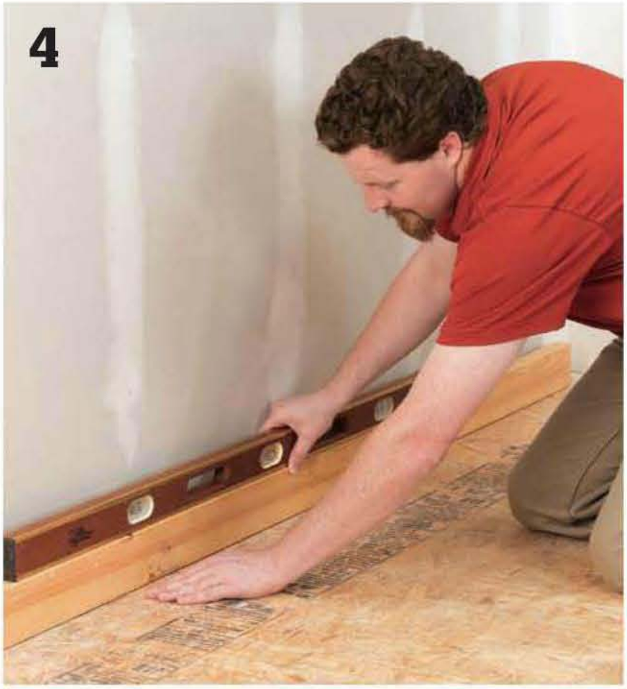

Find the highest point along the floor that will be covered by base cabinets. Place a level on a long, straight 2 × 4, and move the board across the floor to determine if the floor is uneven. Mark the wall at the high point.

(continued)

5

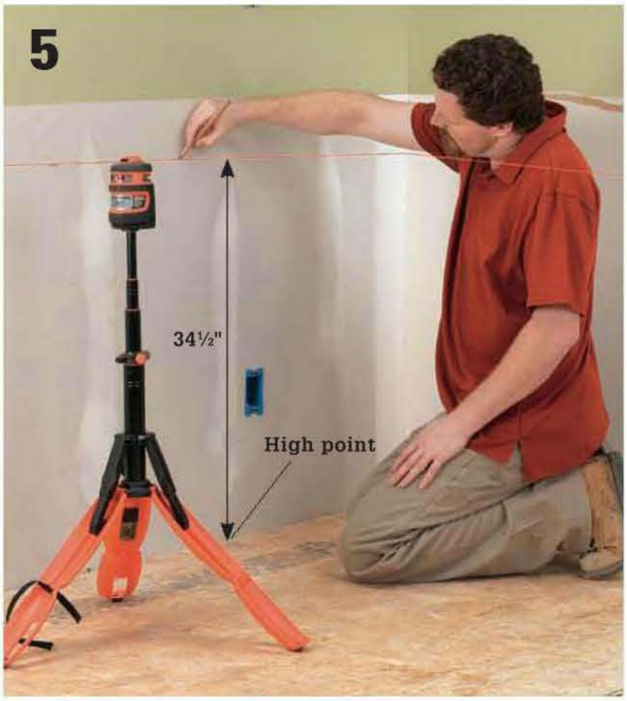

Measure up 34½" from the high-point mark (for standard cabinets). Use a level (a laser level is perfect) to mark a reference line on walls. Base cabinets will be installed with top edges flush against this line.

6

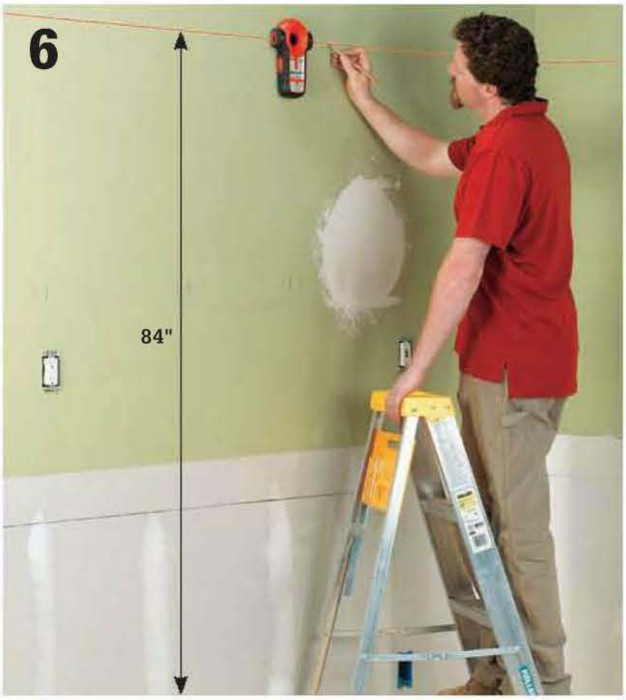

Measure up 84" from the high-point mark and draw a second reference line. Wall cabinets will be installed with their top edges flush against this line.

7

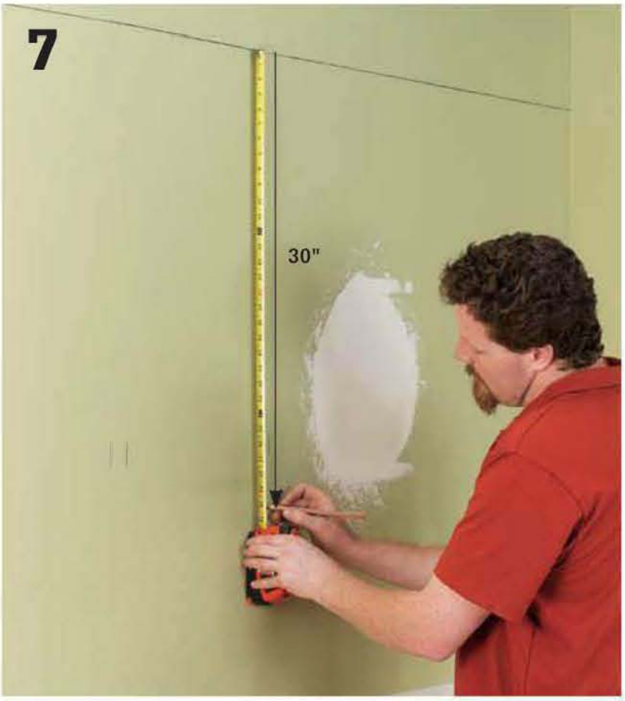

Measure down 30" from the wall-cabinet reference line and draw another level line where the bottoms of the cabinets will be. Temporary ledgers will be installed against this line.

8

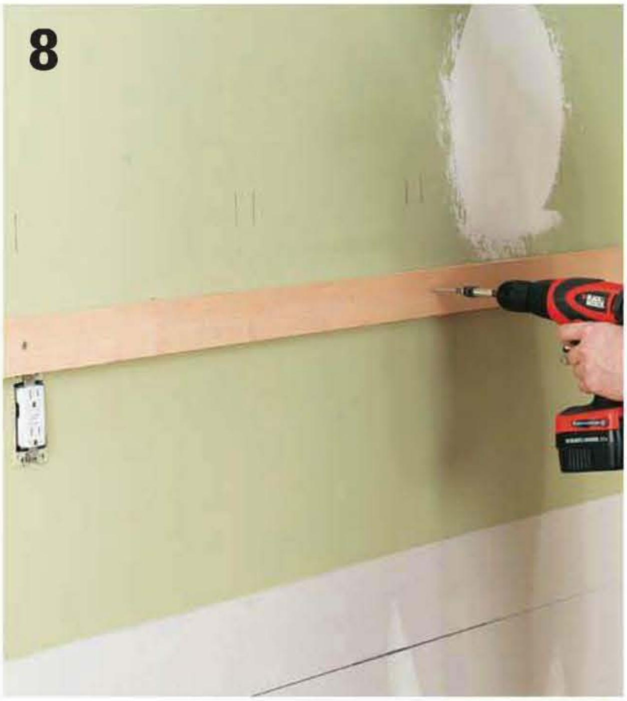

Install 1 × 3 temporary ledgers with top edges flush against the reference lines. Attach ledgers with 2½" wallboard screws driven into every other wall stud. Mark stud locations on ledgers. Cabinets will rest temporarily on ledgers during installation (the ledgers alone will not support them, however).

How to Hang Wall Cabinets

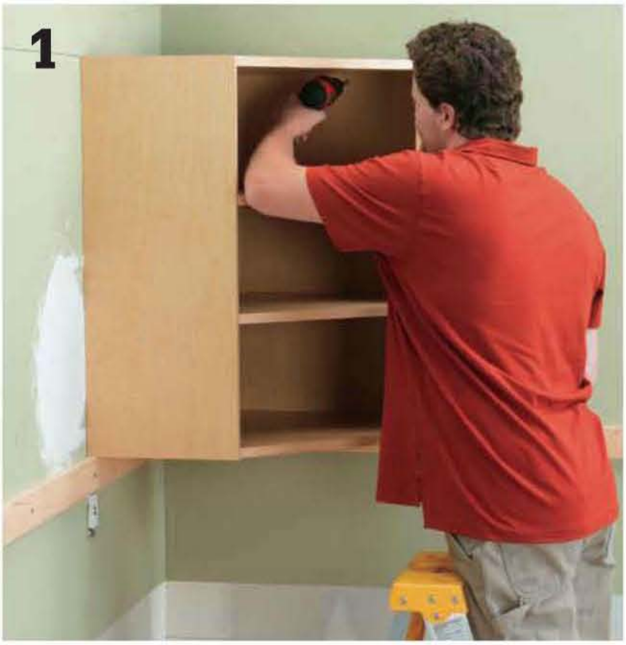

Position a corner upper cabinet on a ledger (see page 352) and hold it in place, making sure it is resting cleanly on the ledger. Drill 3⁄16" pilot holes into the wall studs through the hanging strips at the top, rear of cabinet. Attach the cabinet to the wall with 2½" screws. Do not tighten fully until all cabinets are hung.

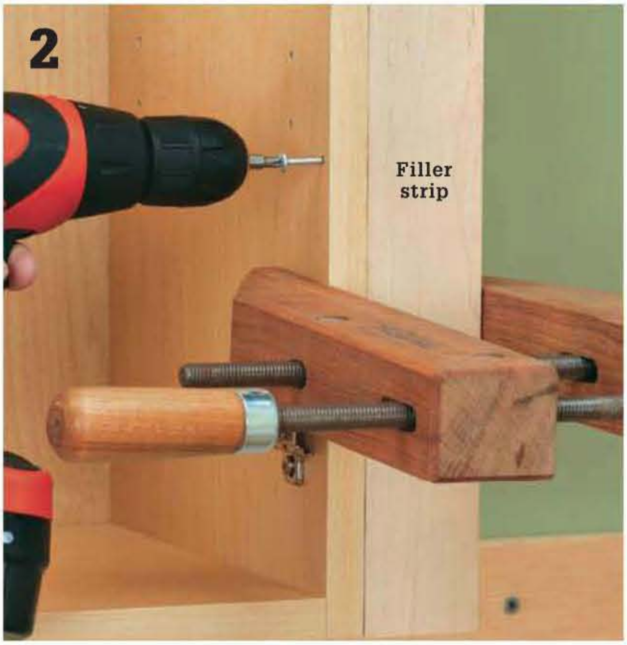

Attach a filler strip to the front edge of the cabinet, if needed (see page 350). Clamp the filler in place, and drill counterbored pilot holes through the cabinet face frame, near hinge locations. Attach filler to cabinet with 2½" cabinet screws or flathead wood screws.

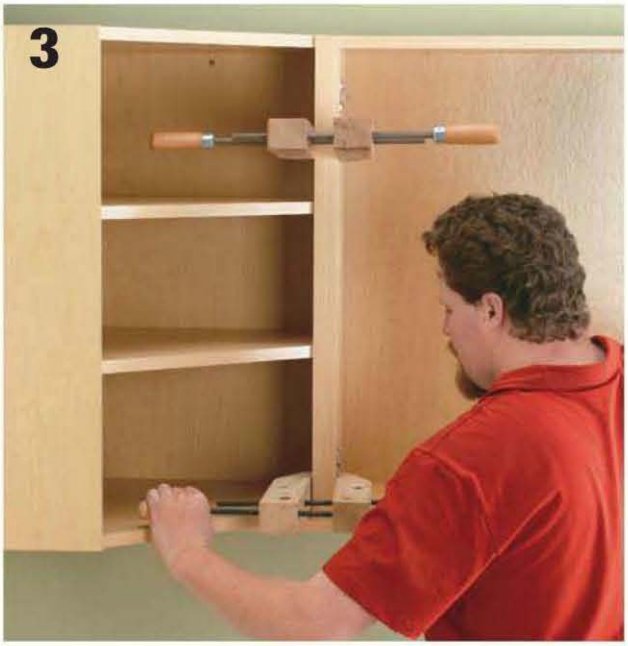

Position the adjoining cabinet on the ledger, tight against the corner cabinet or filler strip. Clamp the corner cabinet and the adjoining cabinet together at the top and bottom. Handscrew clamps will not damage wood face frames.

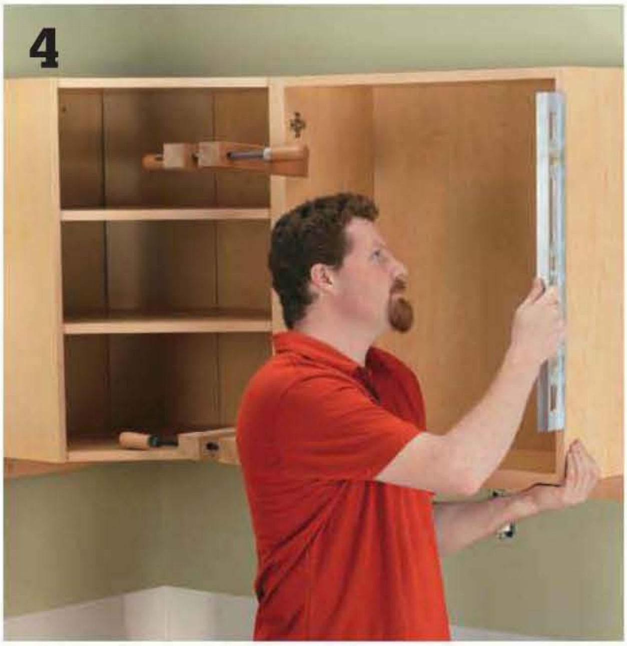

Check the front cabinet edges or face frames for plumb. Drill 3⁄16" pilot holes into wall studs through hanging strips in rear of cabinet. Attach cabinet with 2½" screws. Do not tighten wall screws fully until all cabinets are hung.

(continued)

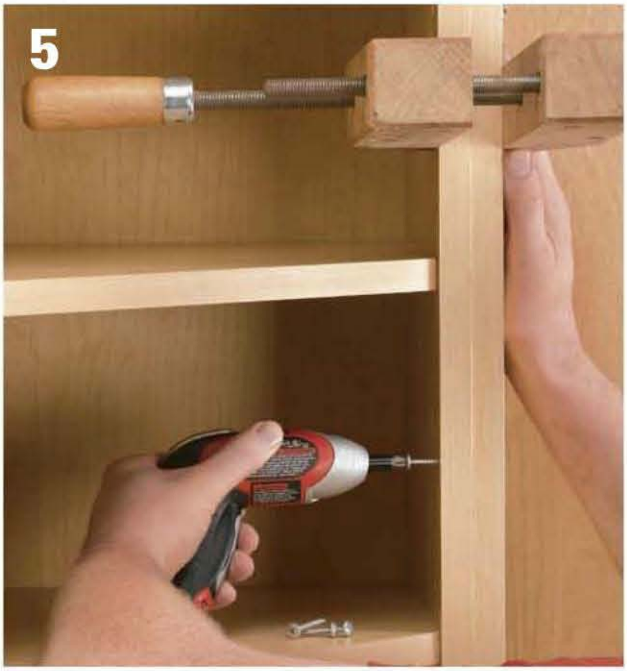

Attach the corner cabinet to the adjoining cabinet. From inside corner cabinet, drill pilot holes through face frame. Join cabinets with sheet-metal screws.

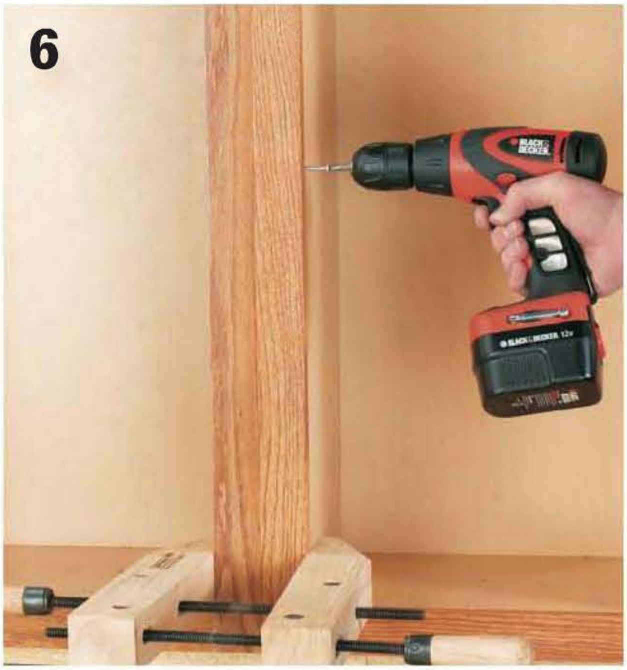

Position and attach each additional cabinet. Clamp frames together, and drill counterbored pilot holes through side of face frame. Join cabinets with wood screws. Drill 3⁄16" pilot holes in hanging strips, and attach cabinet to studs with wood screws.

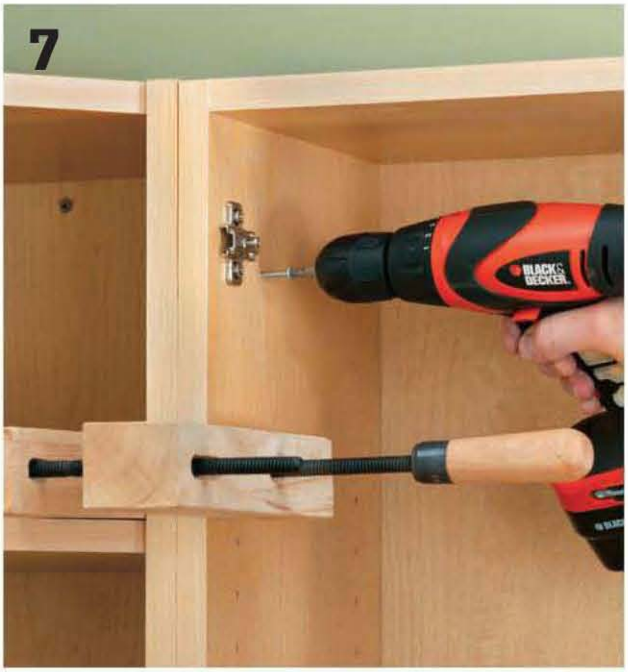

Join frameless cabinets with #8 × 1¼" panhead wood screws or wood screws with decorative washers. Each pair of cabinets should be joined by at least four screws.

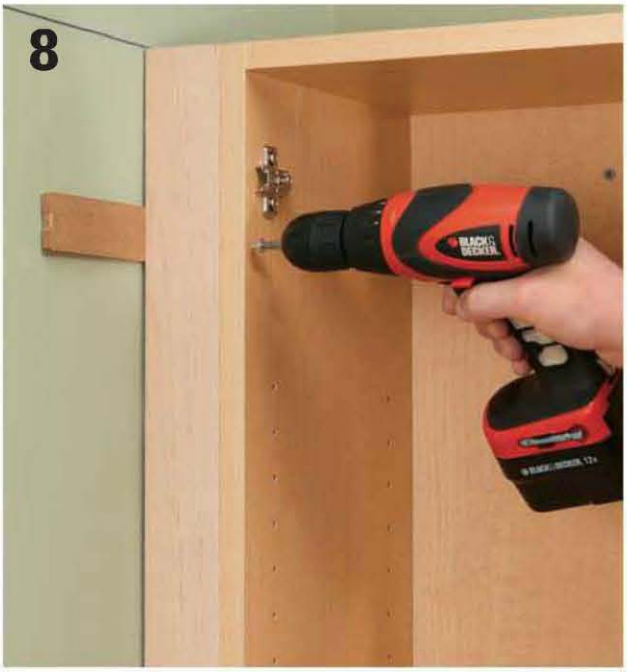

Fill gaps between the cabinet and wall or neighboring appliance with a filler strip. Cut the filler strip to fit the space, then wedge wood shims between the filler and the wall to create a friction fit that holds it in place temporarily. Drill counterbored pilot holes through the side of the cabinet (or the edge of the face frame) and attach filler with screws.

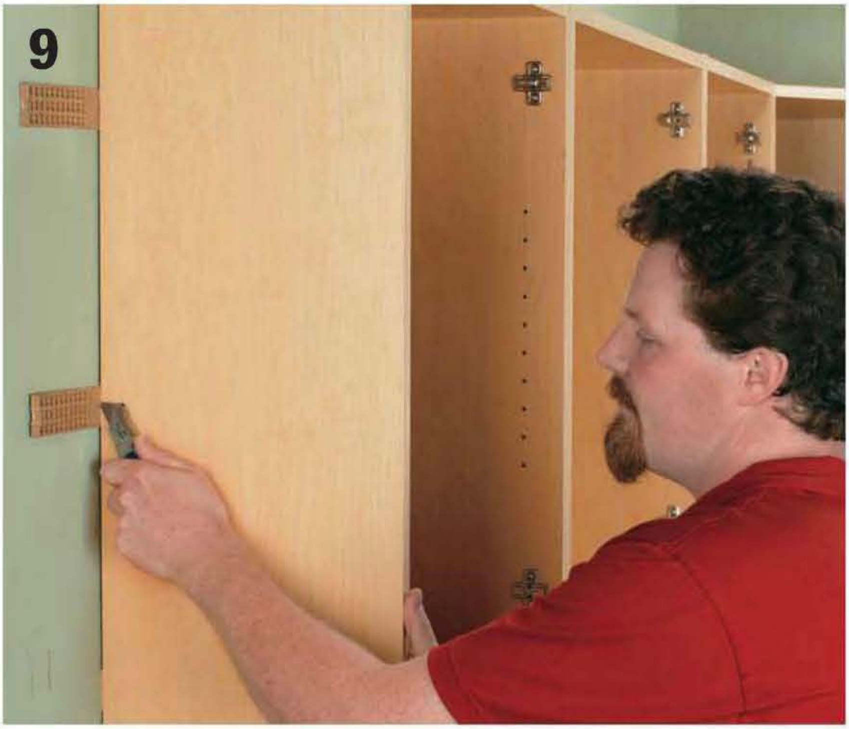

Remove the temporary ledger. Check the cabinet run for plumb, and adjust if necessary by placing wood shims behind cabinet, near stud locations. Tighten wall screws completely. Cut off shims with utility knife.

Use trim moldings to cover any gaps between cabinets and walls. Stain moldings to match cabinet finish.

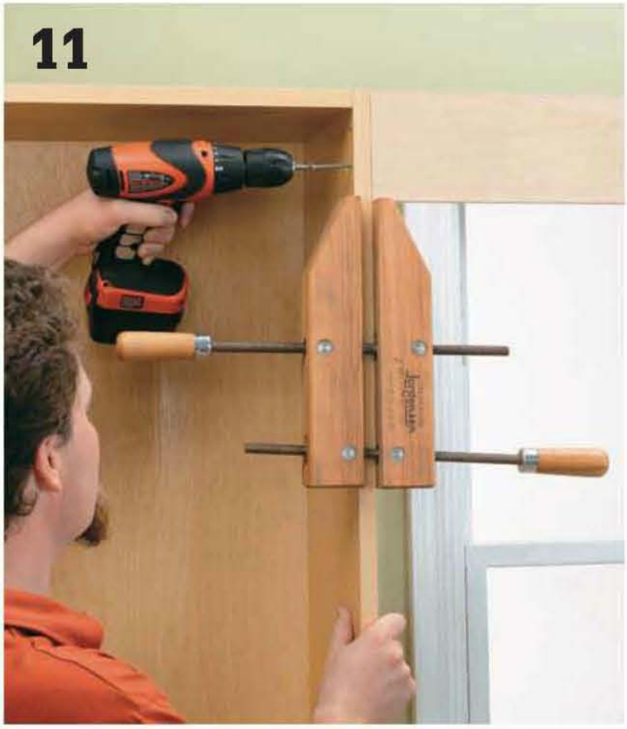

Attach decorative valance above sink. Clamp valance to edge of cabinet frames, and drill counterbored pilot holes through cabinet frames into end of valance. Attach with sheet-metal screws.

Install the cabinet doors. If necessary, adjust the hinges so that the doors are straight and plumb.

How to Install Base Cabinets

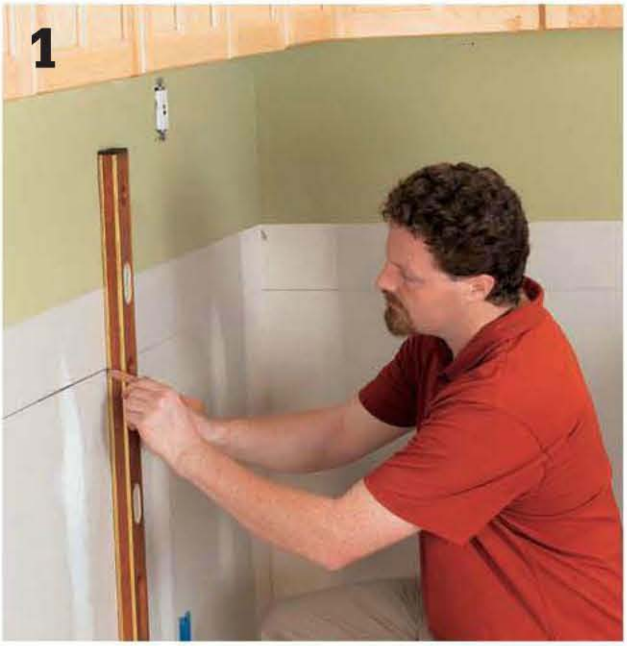

Begin the installation with a corner cabinet. Draw plumb lines that intersect the 34½" reference line (measured from the high point of the floor—see page 351) at the locations for the cabinet sides.

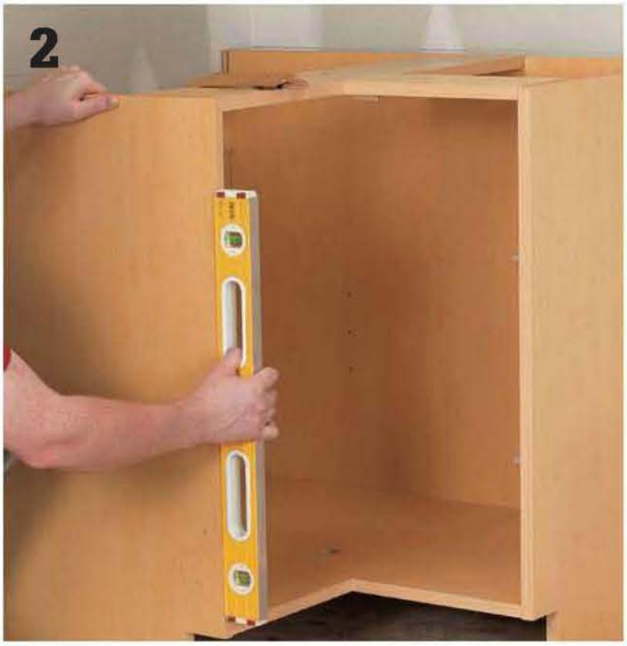

Place cabinet in corner. Make sure the cabinet is plumb and level. If necessary, adjust by driving wood shims under cabinet base. Be careful not to damage flooring. Drill 3⁄16" pilot holes through the hanging strip and into wall studs. Tack the cabinet to the wall with wood screws or wallboard screws.

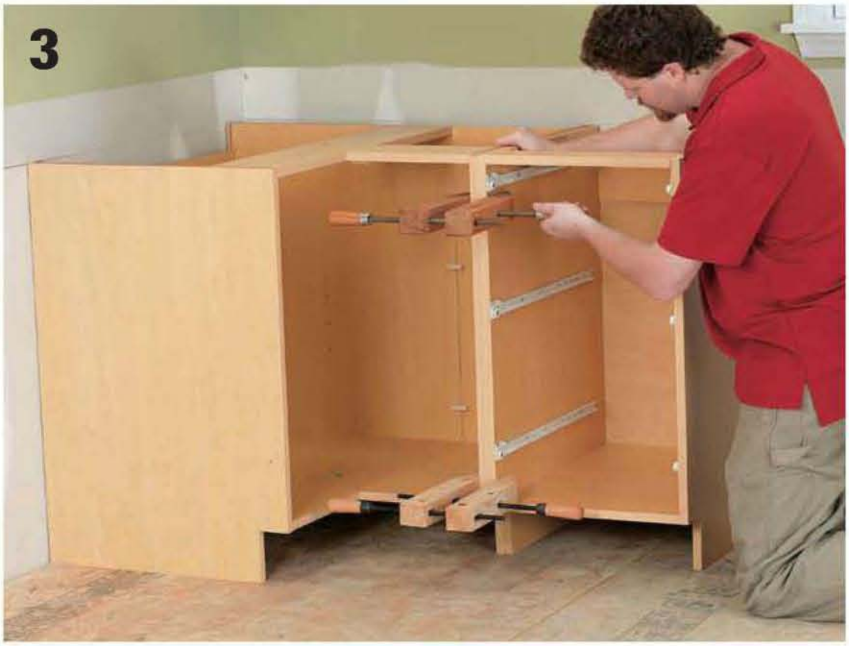

Clamp the adjoining cabinet to the corner cabinet. Make sure the new cabinet is plumb, then drill counterbored pilot holes through the cabinet sides or the face frame and filler strip. Screw the cabinets together. Drill 3⁄16" pilot holes through hanging strips and into wall studs. Tack the cabinets loosely to the wall studs with wood screws or wallboard screws.

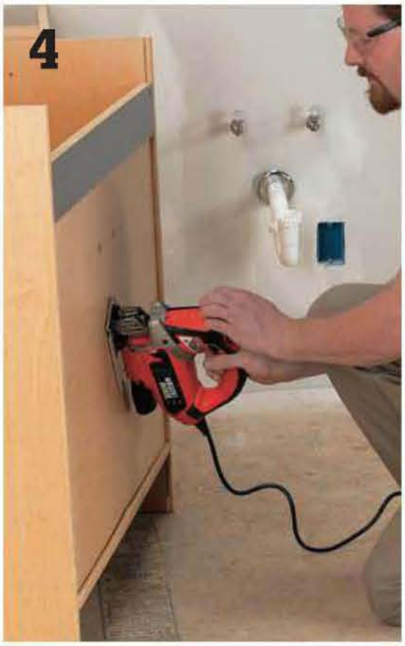

Use a jigsaw to cut any cabinet openings needed in the cabinet backs (for example, in the sink base seen here) for plumbing, wiring or heating ducts.

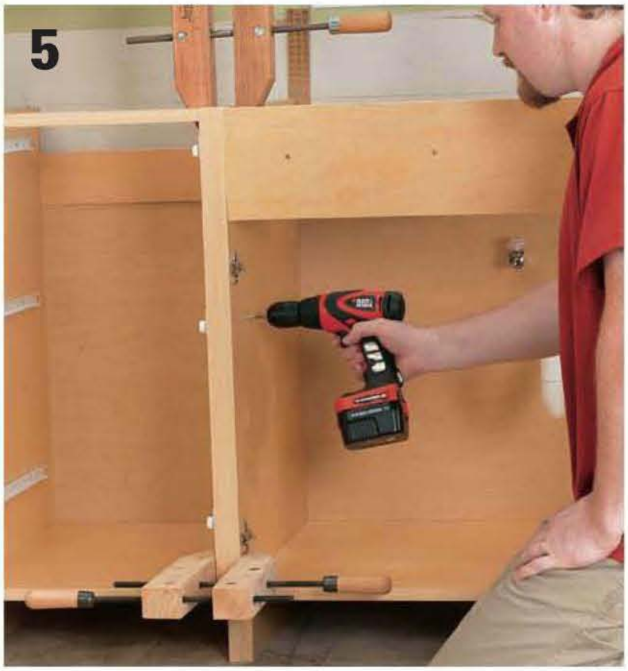

Position and attach additional cabinets, making sure the frames are aligned and the cabinet tops are level. Clamp cabinets together, then attach the face frames or cabinet sides with screws driven into pilot holes. Tack the cabinets to the wall studs, but don't drive screws too tight—you may need to make adjustments once the entire bank is installed.

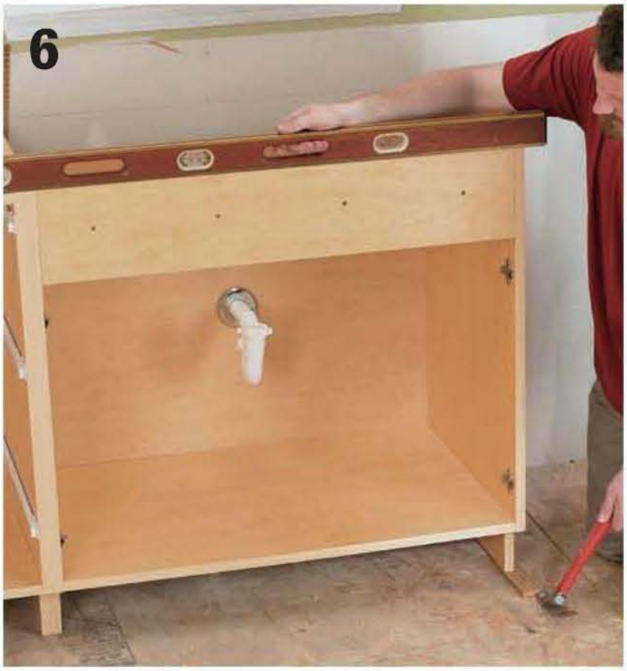

Make sure all cabinets are level. If necessary, adjust by driving shims underneath cabinets. Place shims behind the cabinets near stud locations to fill any gaps. Tighten wall screws. Cut off shims with utility knife.

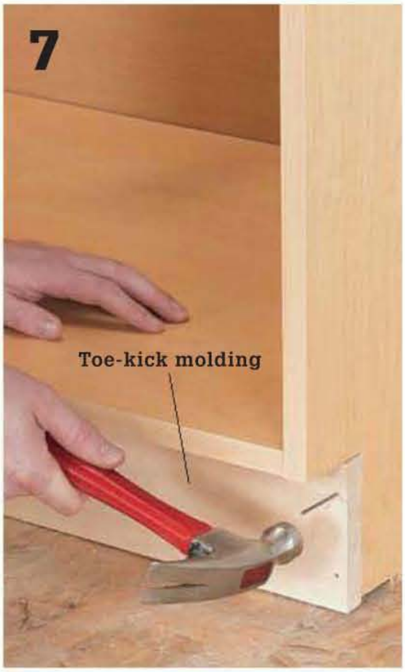

Use trim moldings to cover gaps between the cabinets and the wall or floor. The toe-kick area is often covered with a strip of wood finished to match the cabinets or painted black.

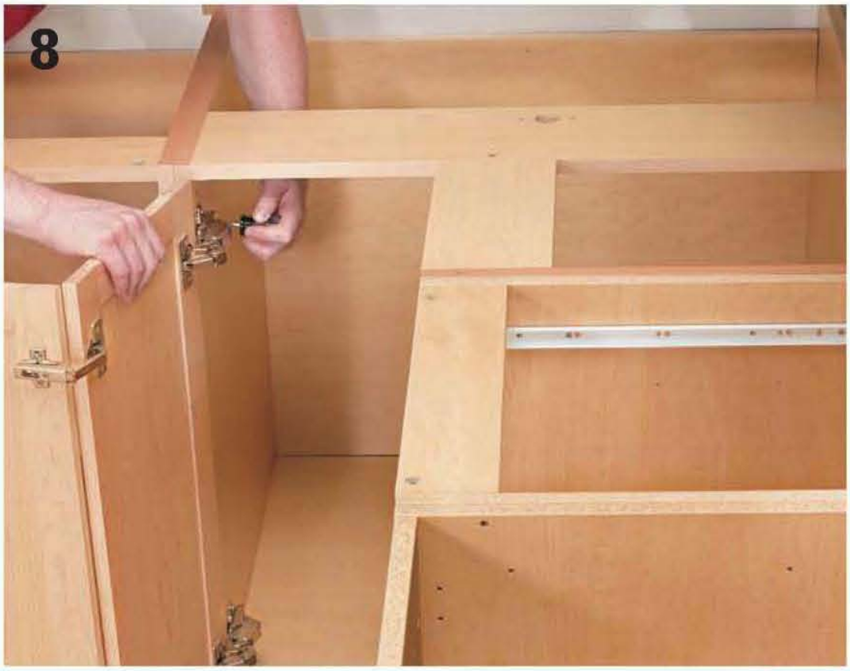

Hang cabinet doors and mount drawer fronts. Test to make sure the drawers close smoothly and the doors fit evenly and flush. Self-closing cabinet hinges (by far the most common type installed today) have adjustment screws that allow you to make minor changes to the hardware to correct any problems.

Building Kitchen Islands

Kitchen islands can be created using a whole range of methods, from repurposing an old table to fine, custom woodworking. But perhaps the easiest (and most failsafe) way to add the conveniences and conviviality of a kitchen island is to make one from stock base cabinets. The cabinets and countertops don't have to match your kitchen cabinetry, but that is certainly an option you should consider. When designing and positioning your new island, be sure to maintain a minimum distance of 3 ft. between the island and other cabinets (4 ft. or more is better).

Tools & Materials ▸

- Marker
- Drill/driver
- 2 × 4 cleats
- Pneumatic nailer and 2" finish nails or hammer and 6d finish nails
- 2 base cabinets (approx. 36" wide × 24" deep)
- Countertop
- Wallboard screws

Two base cabinets arranged back-to-back make a sturdy kitchen island base that's easy to install. When made with the same style cabinets and countertops as the rest of the kitchen, the island is a perfect match.

How to Create a Stock-cabinet Island

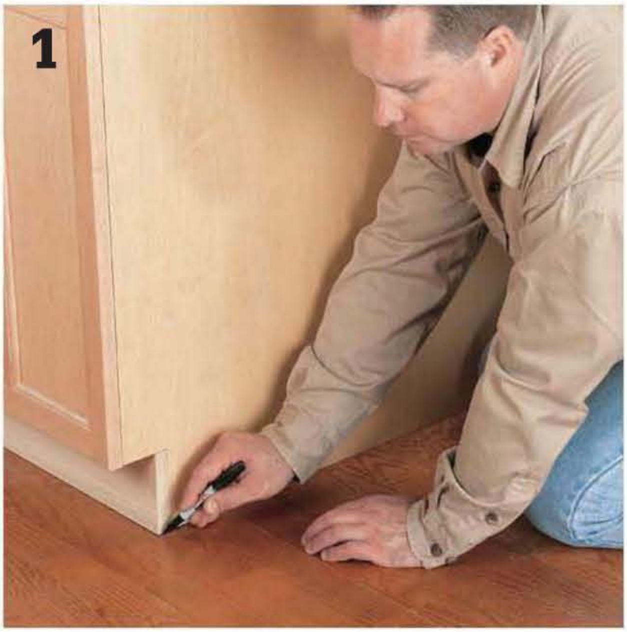

Set two base cabinets back-to-back in position on the floor and outline the cabinet corners onto the flooring. Remove the cabinets and draw a new outline inside the one you just created to allow for the thickness of the cabinet sides (usually ¾").

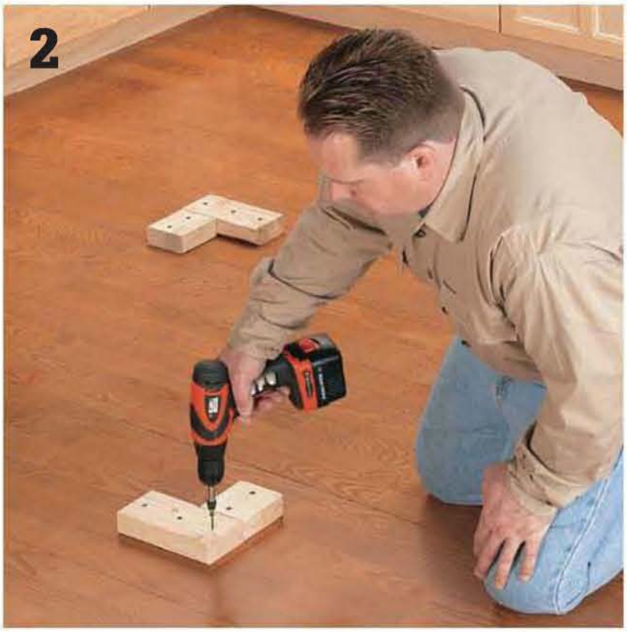

Cut 2 × 4 cleats to fit inside the inner outline to provide nailing surfaces for the cabinets. Attach the cleats to the floor with screws or nails. *Tip: Create an L-shape cleat for each inside corner.*

Join the two base cabinets by driving 1¼" wallboard screws through the nailing strips on the backs of the cabinets from each direction. Make sure the cabinet sides are flush and aligned. Lower the base cabinets over the cleats. Check the cabinets for level, and shim underneath the edges of the base if necessary.

Attach the cabinets to the floor cleats using 6d finish nails. Drill pilot holes for nails, and recess nail heads with a nail set. Install a countertop on top of the cabinets (see pages 362 to 385).

Building Tiled Islands

Islands are one of the most requested kitchen features. People love them for many reasons, including their value as bi-level counter space. In most cases, the lower level is used as work space and the upper as casual dining space. The upper level provides a little camouflage for the work space, something that's especially welcome in open-plan kitchens where meal preparation areas can be seen from other areas.

When planning casual dining space, remember that designers suggest at least 24" per person. For the work space, remember that standard design guidelines recommend at least 36" of uninterrupted work space to the side of a sink or cooktop.

On work surfaces, mosaic and other small tile is rarely the best choice. Larger tile requires fewer grout lines, always a good idea when it comes to cleaning and maintenance. But there is no rule that all three elements of a bi-level island have to use the same material. In fact, projects like this offer wonderful opportunities to mix materials or colors or textures. Choose floor tile or tile made especially for counters for the horizontal surfaces, and then branch out when it comes to the backsplash. Wall tile and mosaics work beautifully.

Tools & Materials ▸

- Tape measure
- Circular saw
- Drill
- Utility knife
- Straightedge
- Stapler
- Drywall knife
- Framing square
- Notched trowel
- Tile cutter
- Carpeted 2 × 4
- Mallet
- Rubber grout float
- Sponge
- Foam brush
- Caulk gun
- Birch plywood
- 1 × 2 hardwood
- 2 × 4 lumber
- Ceramic tile
- Tile spacers
- ¾" exterior-grade (CDX) plywood
- 4-mil polyethylene sheeting
- Packing tape
- ½" cementboard
- 1¼" deck screws
- 3" deck screws
- Fiberglass mesh tape
- Thinset mortar
- Grout with latex additive
- Silicone caulk
- Silicone grout sealer
- L-brackets
- 6d finish nails
- Drywall screws
- Glue

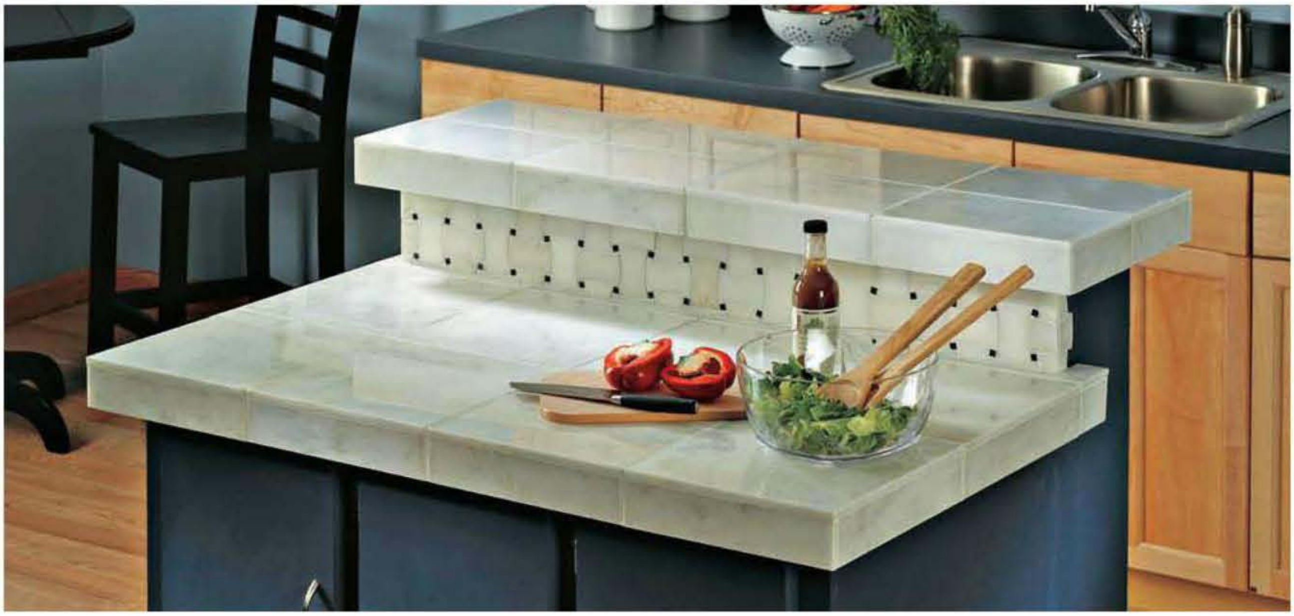

Attractive backsplashes complete tile contertops. Here, individual tiles seamlessly stack side by side to create one cohesive, interwoven pattern. Using the same material and tones as the surrounding countertop allows the pattern to add interest without cluttering the small island, resulting in a sophisticated finished design.

How to Build a Tiled Bi-level Island

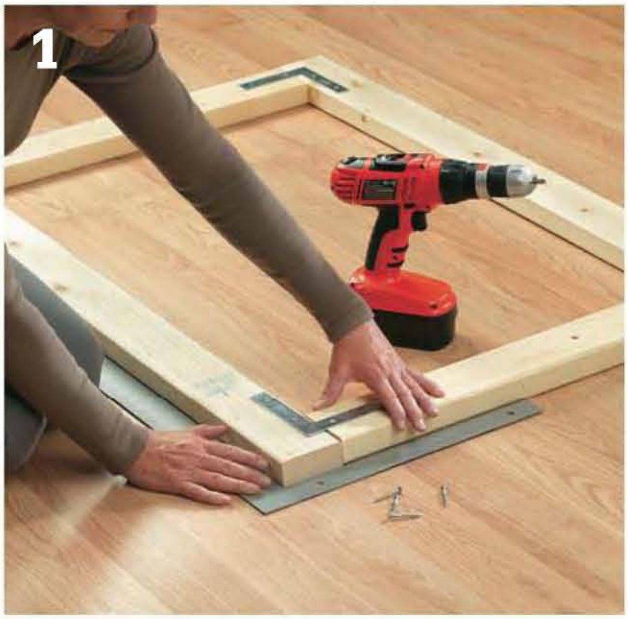

Build a 2 × 4 base for the island cabinet by cutting the 2 × 4s to length and joining them in a square frame that lays flat (wide sides down) on the floor. Use metal L-brackets to reinforce the joints. If you don't wish to move the island, fasten the frame to the floor in position with construction adhesive and/or deck screws.

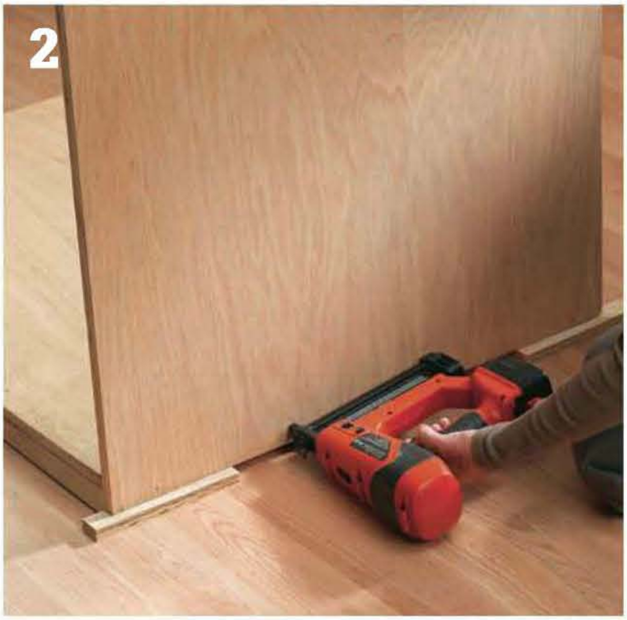

Cut bottom panels the same dimensions as the base frame from ¾" birch plywood. Attach it to the frame with finish nails. Then, cut the side panels to size and shape and fasten them to the edges of the curb with 6d finish nails and adhesive. Slip ¾" shims (scrap plywood works well) beneath the side panels before fastening them.

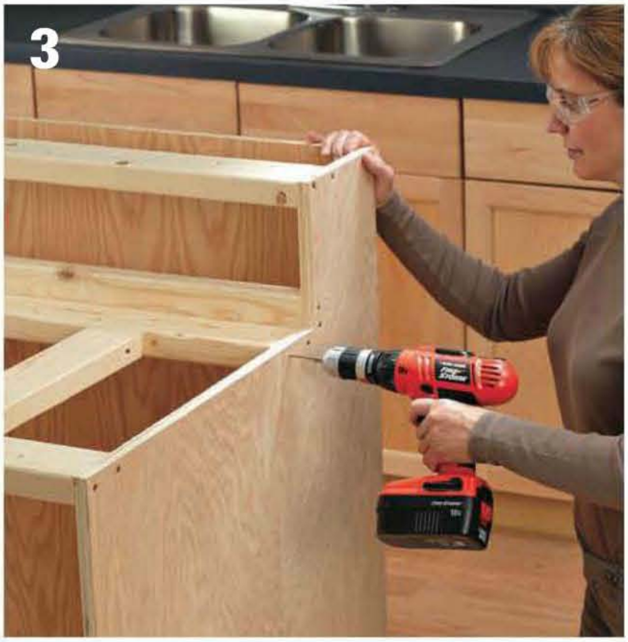

Cut the 2 × 4 cross supports to length and install them between the side panels at every corner, including the corners created by the L-shape cutout. Use 3" deck screws driven through the side panels and into the ends of the cross supports.

Lightly sand the cabinets and then clean off the dust. Prime and paint the cabinet interior and exterior.

(continued)

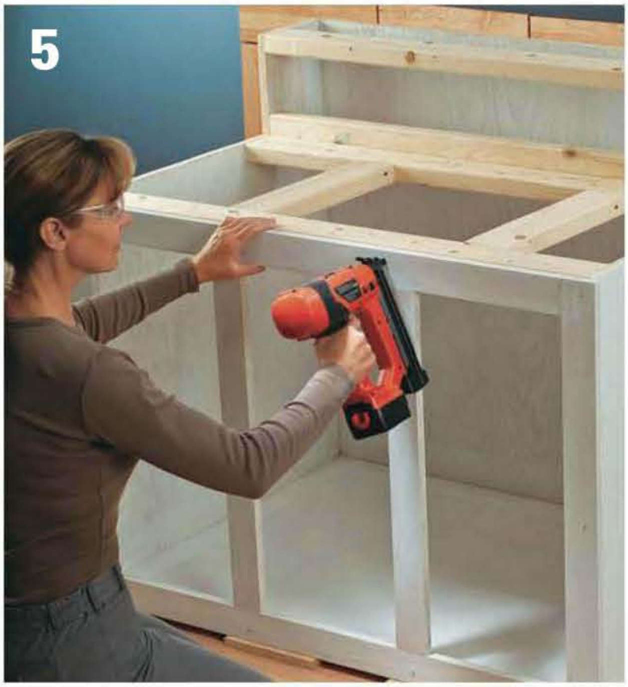

Build a face frame from 1 × 2 hardwood to fit the cabinet front. Attach it to the cabinet with 6d finish nails and hang cabinet doors (we installed three 13"-wide overlay doors).

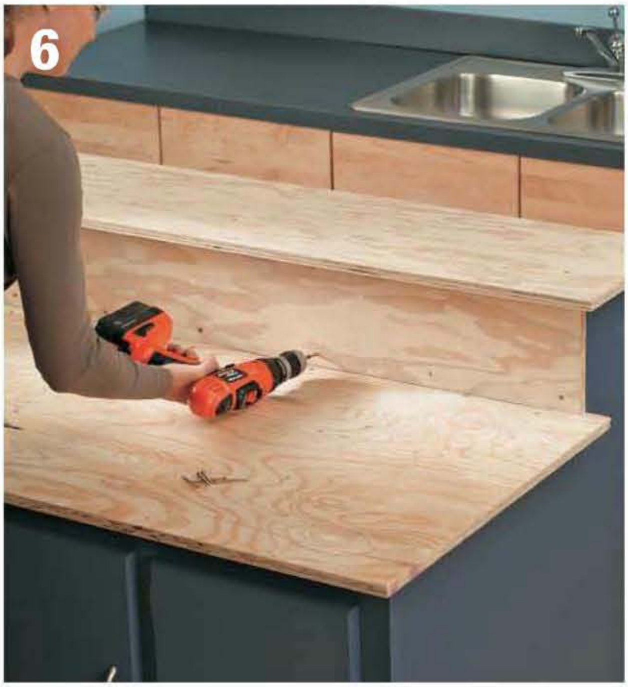

Cut strips of ¾" exterior plywood to make the subbases for the countertops and a backer for the backsplash. The lower counter subbase should over hang by 2" on the front and sides. The upper should overhang 2" on the sides and be centered on the cabinet front to back. Attach the backer and subbases with drywall screws driven down into the 2 × 4 cross supports.

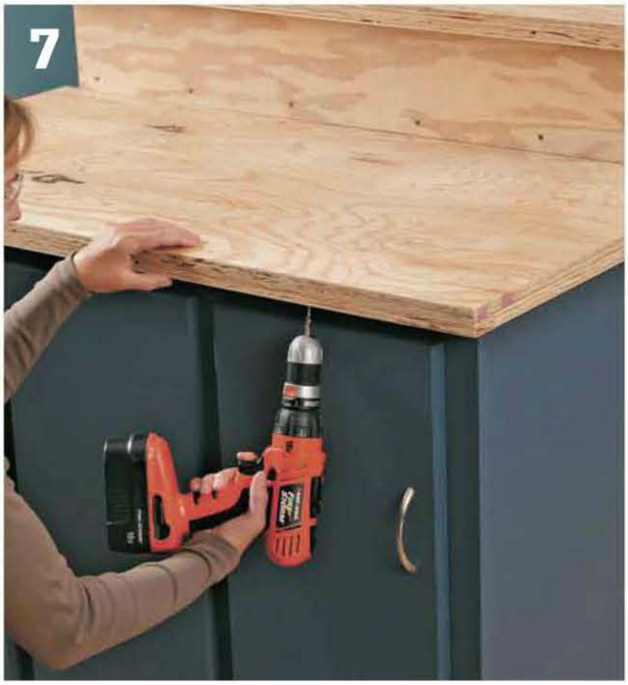

Cut 2" wide strips of plywood for buildup strips and attach to the undersides of the subbases with glue and screws.

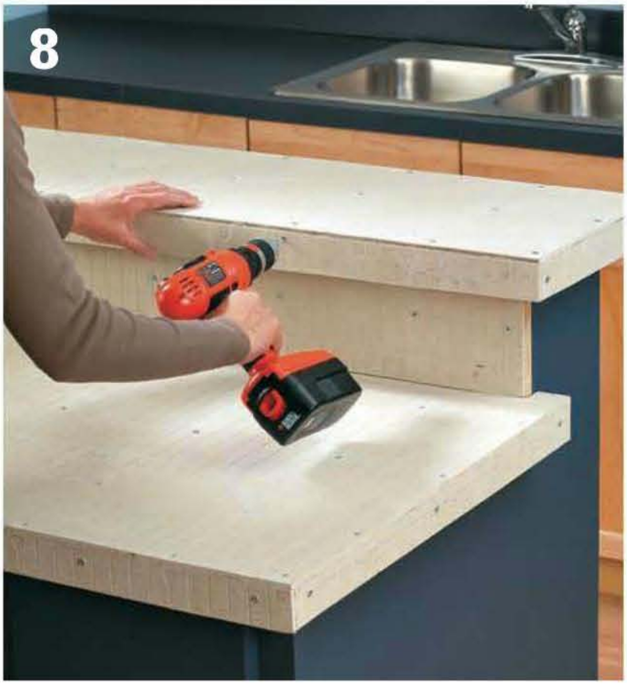

Attach tile backerboard to the counter subbases and the backsplash and tape seams and cover screws heads with compound (see page 380).

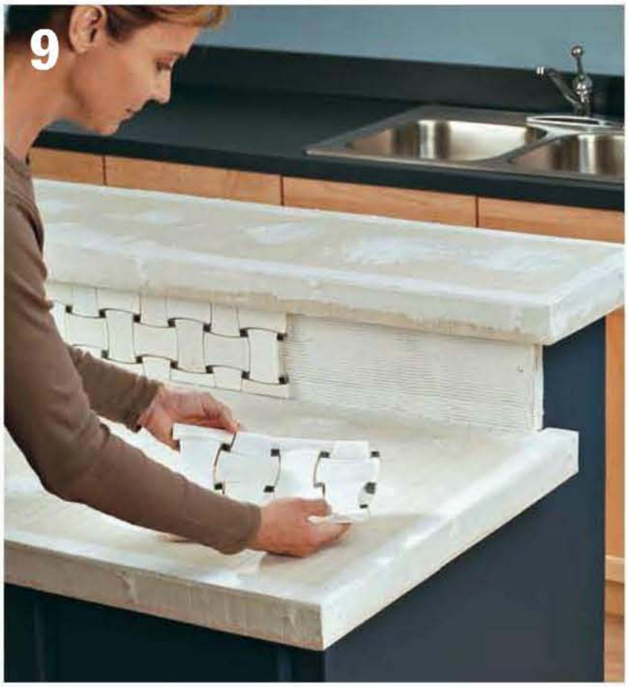

Cut mosaic sheets to fit the backsplash area and attach them with thinset adhesive (see Tiling a Backsplash, page 387).

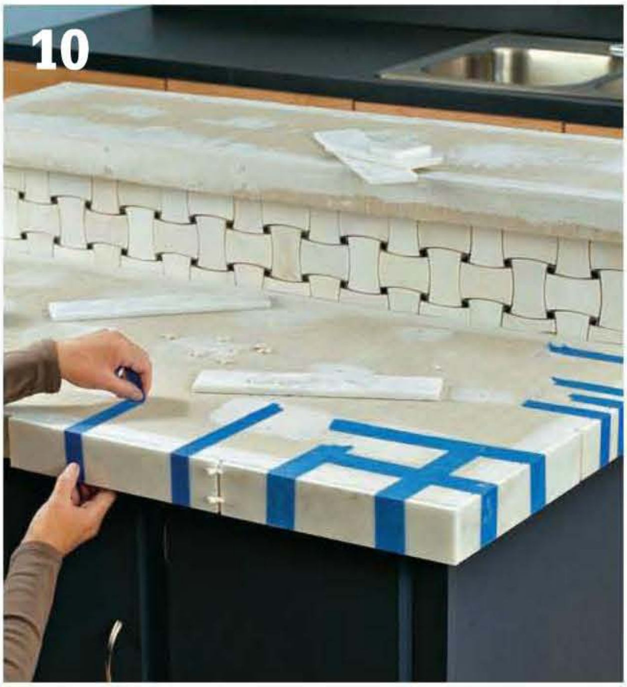

Cut edge tiles and fasten them around the perimeter of the subbase with thinset adhesive. Hold tiles in place with tape until adhesive sets up. The tiles should be flush or slightly below the bottoms of the buildup strips, and project past the top surfaces so they will be level with the field tiles.

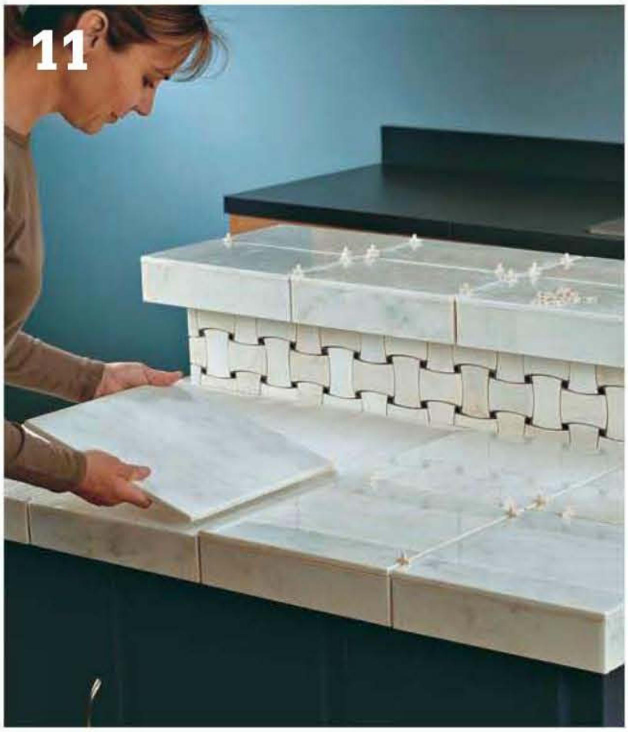

Install the field tiles for the countertops last (see pages 376 to 385).

Choose a suitable grout color and apply it to the tile with a grout float (page 385). Buff off excess once it has dried. Seal the grout with grout sealer.

Installing Postform Countertops

Postform laminate countertops are available in stock and custom colors. Pre-mitered sections are available for two- or three-piece countertops that continue around corners. If the countertop has an exposed end, you will need an endcap kit that contains a preshaped strip of matching laminate. Postform countertops have either a waterfall edge or a no-drip edge. Stock colors are typically available in 4-, 6-, 8-, 10- and 12-foot straight lengths and 6- and 8-foot mitered lengths.

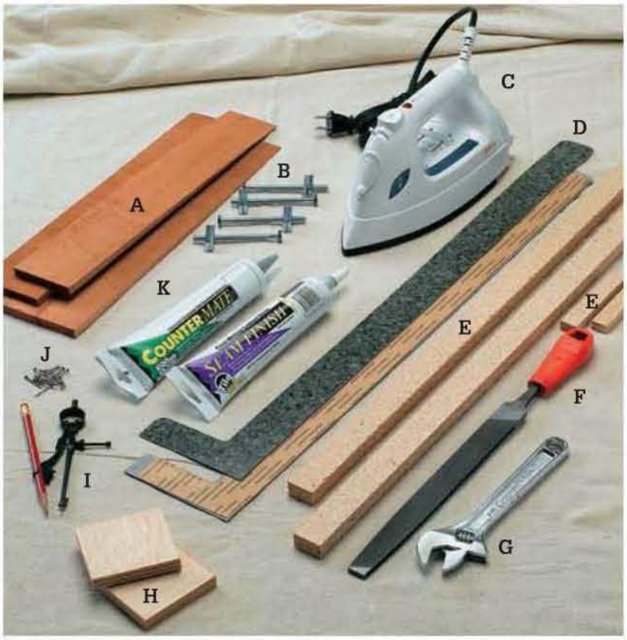

Materials and tools for installing a post-form countertop include: Wood for shimming (A), take-up bolts for drawing miters together (B), household iron (C), endcap laminate to match countertop (D), endcap battens (E), file (F), adjustable wrench (G), buildup blocks (H), compass (I), fasteners (J), silicone caulk and sealer (K).

Post-form countertops are among the easiest and cheapest to install. They are a good choice for beginning DIYers, but the design and color options are fairly limited.

How to Install a Postform Countertop

Tools & Materials ▸

- Tape measure
- Framing square
- Pencil
- Straightedge
- C-clamps
- Hammer
- Level
- Caulking gun
- Jigsaw
- Compass
- Adjustable wrench
- Belt sander
- Drill and spade bit
- Cordless screwdriver
- Postform countertop
- Wood shims
- Take-up bolts
- Drywall screws
- Wire brads
- Endcap laminate
- Silicone caulk
- Wood glue

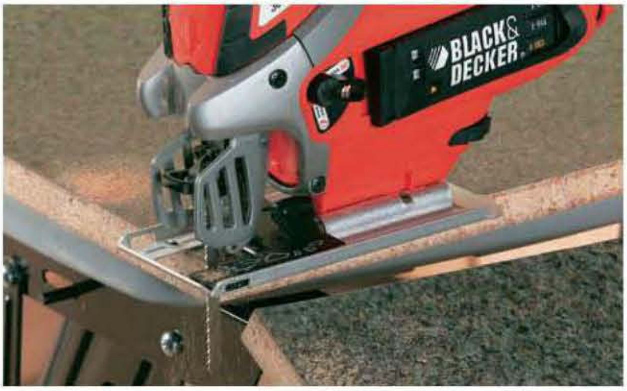

Option: If the saw foot must rest on the good surface of the postform, use a jigsaw fitted with a downstroke blade to cut the postform. If you are unable to locate a downstroke blade, you can try applying tape over the cutting lines, but you are still likely to get tear-out from a normal upstroke jigsaw blade.

1

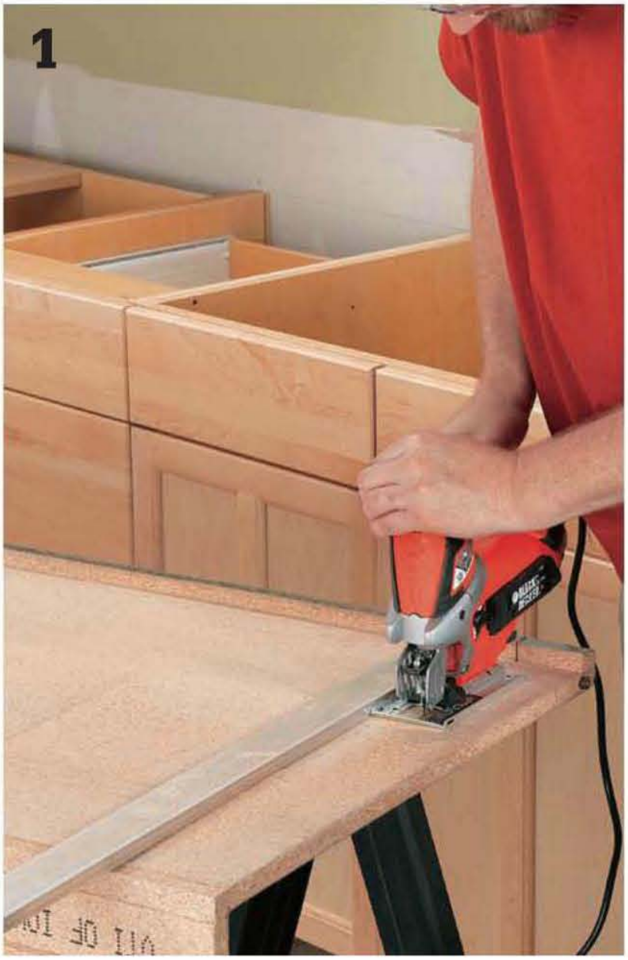

Use a framing square to mark a cutting line on the bottom surface of the countertop. Cut off the countertop with a jigsaw, using a clamped straight-edge as a guide.

2

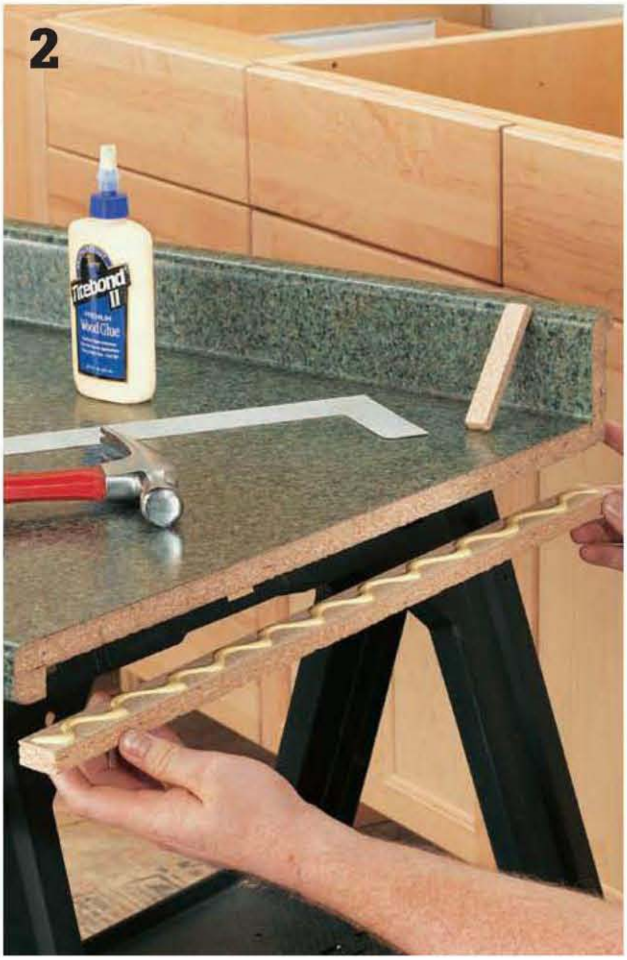

Attach the battens from the endcap kit to the edge of the countertop, using carpenter's glue and small brads. Sand out any unevenness with a belt sander.

(continued)

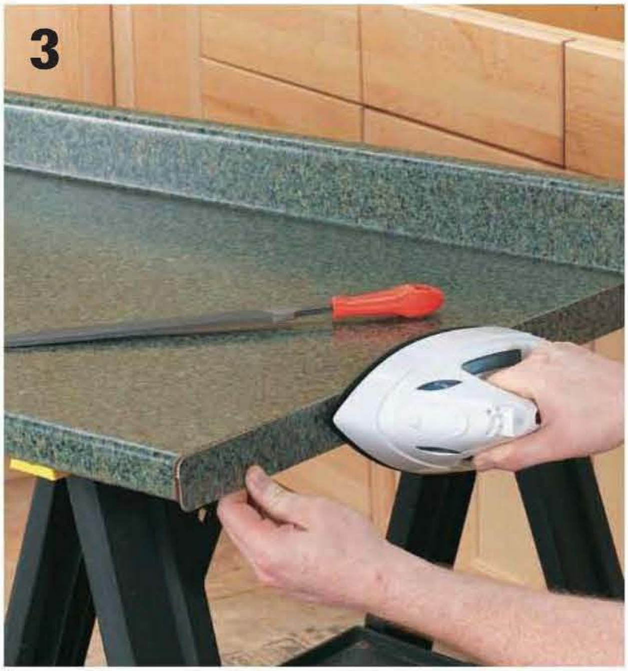

Hold the endcap laminate against the end, slightly overlapping the edges. Activate adhesive by pressing an iron set at medium heat against the endcap. Cool with a wet cloth, then file the endcap laminate flush with the edges of the countertop.

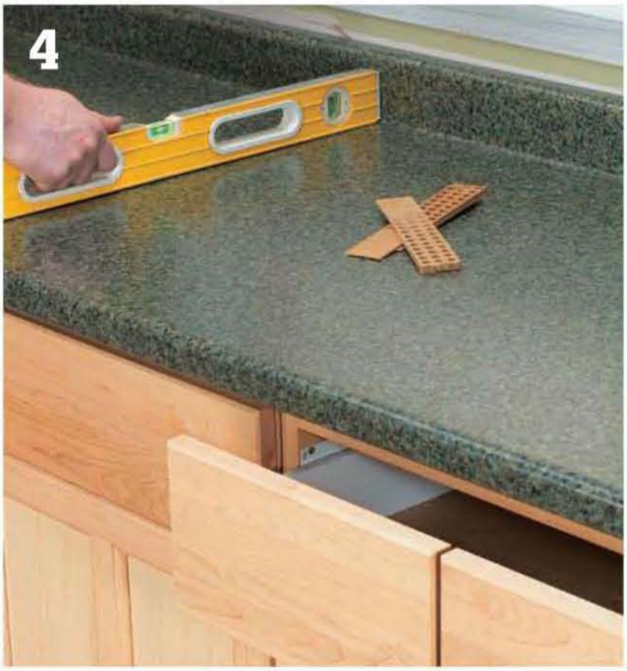

Position the countertop on base cabinets. Make sure the front edge of the countertop is parallel to the cabinet faces. Check the countertop for level. Make sure that drawers and doors open and close freely. If needed, adjust the countertop with shims.

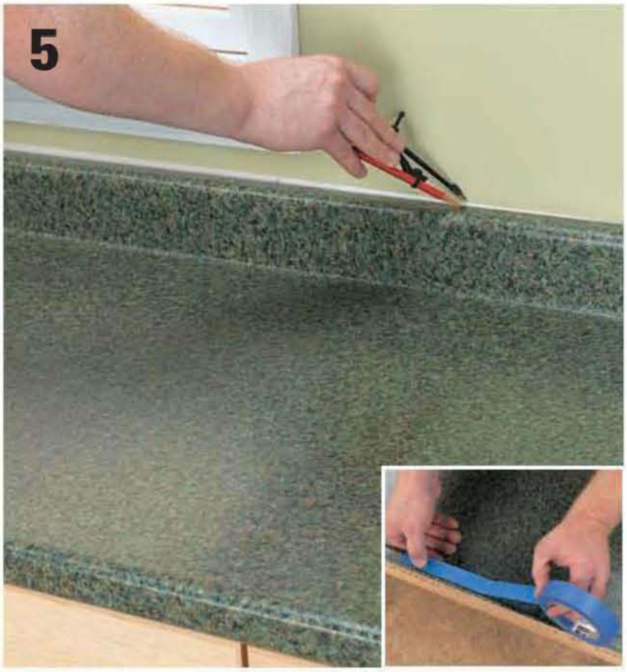

Because walls are usually uneven, use a compass to trace the wall outline onto the backsplash. Set the compass arms to match the widest gap, then move the compass along the length of the wall to transfer the outline to the top of the backsplash. Apply painter's tape to the top edge of the backsplash, following the scribe line (inset).

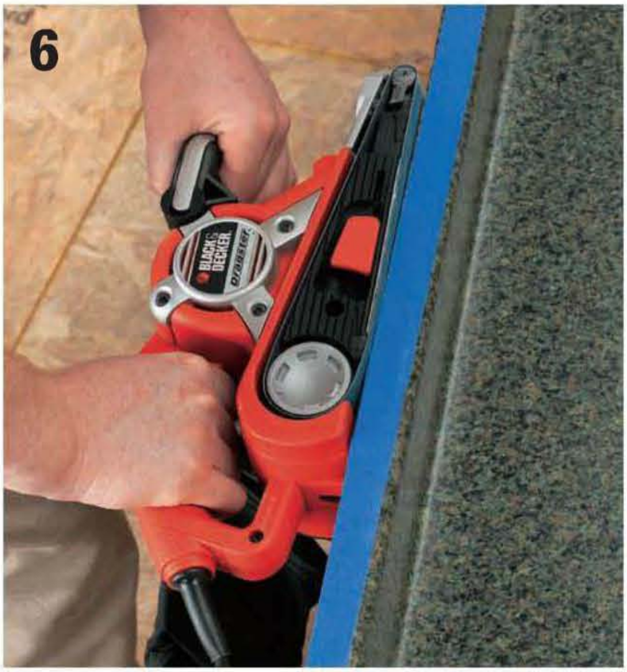

Remove the countertop. Use a belt sander to grind the backsplash to the scribe line.

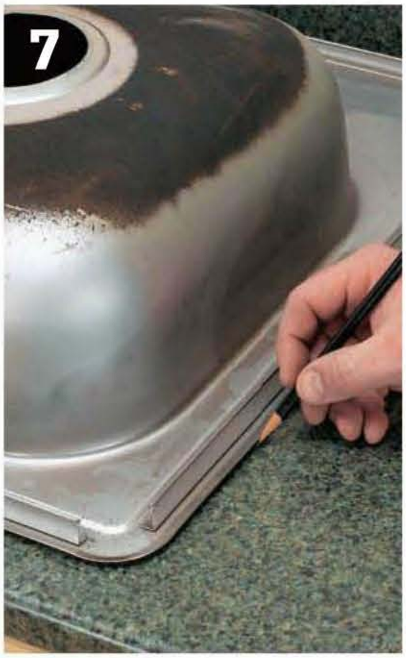
7

Mark cutout for self-rimming sink. Position the sink upside down on the countertop and trace its outline. Remove the sink and draw a cutting line ⅝" inside the sink outline.

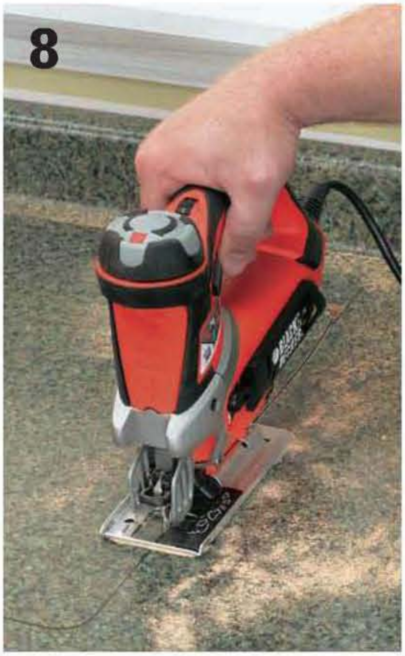
8

Drill a starter hole just inside the cutting line. Make sink cutouts with a jigsaw. Support the cutout area from below so that the falling cutout does not damage the cabinet.

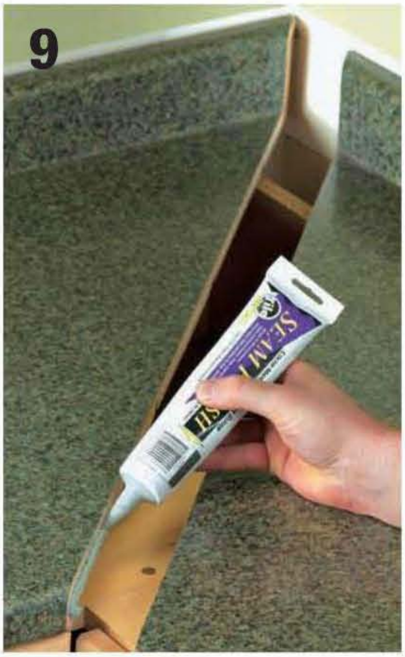
9

Apply a bead of silicone caulk to the edges of the mitered countertop sections. Force the countertop pieces tightly together.

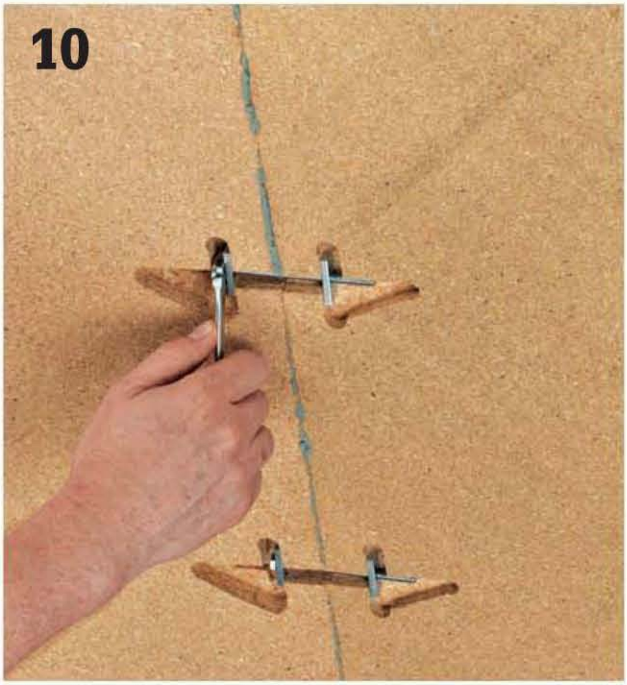
10

From underneath the countertop, install and tighten miter take-up bolts. Position countertop tightly against the wall and fasten it to the cabinets by driving wallboard screws up through corner brackets and into the countertop. Screws should be long enough to provide maximum holding power, but not long enough to puncture the laminate surface.

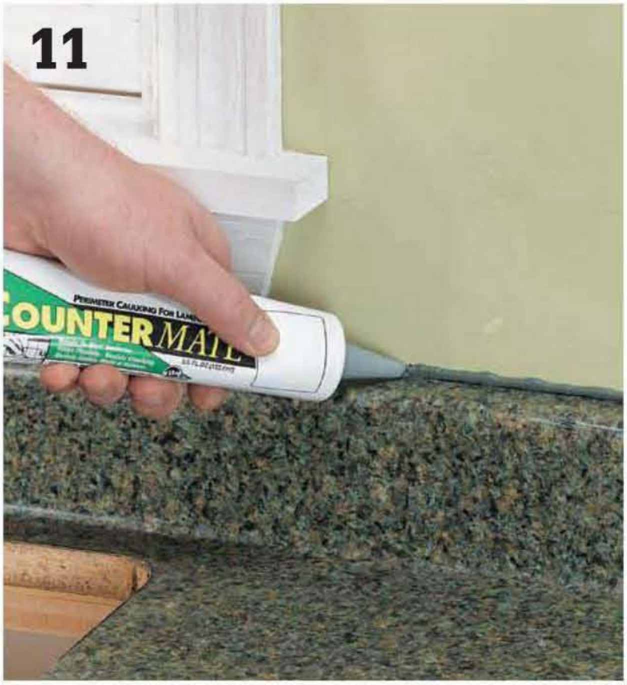
11

Seal the seam between the backsplash and the wall with silicone caulk. Smooth the bead with a wet fingertip. Wipe away excess caulk.

Installing Laminate Countertops

Building your own custom laminate countertop using sheets of plastic laminate and particleboard offers two advantages: the countertop you get will be less expensive than a custom-ordered countertop, and it will allow you more options in terms of colors and edge treatments. A countertop made with laminates also can be tailored to fit any space, unlike premade countertop material that is a standard width (usually 25").

Laminate commonly is sold in 8-ft. or 12-ft. lengths that are about 1/20" thick. In width, they range from 30" strips to 48" sheets. The 30" strips are sized specifically for countertops, allowing for a 25"-wide countertop, a 1½" wide front edge strip and a short backsplash.

The plastic laminate is bonded to the particle-board or MDF substrate with contact cement (although most professional installers use products that are available only to the trades). Water-base contact cement is nonflammable and nontoxic, but solvent-base contact cement (which requires a respirator and is highly flammable) creates a much stronger, more durable bond.

Tools & Materials ▸

- Tape measure
- Framing square
- Straightedge
- Scoring tool
- Paint roller
- 3-way clamps
- Caulk gun
- J-roller
- Miter saw
- Scribing compass
- Circular saw
- Screwdriver
- Belt sander
- File
- Router
- ¾" particleboard
- Sheet laminate
- Contact cement and thinner
- Wood glue
- Drywall screws

Fabricating your own custom countertop from particleboard and plastic laminate is not exactly an easy DIY project, but it gives you unlimited options and the results can be very satisfying.

Tips for Working with Laminate ▸

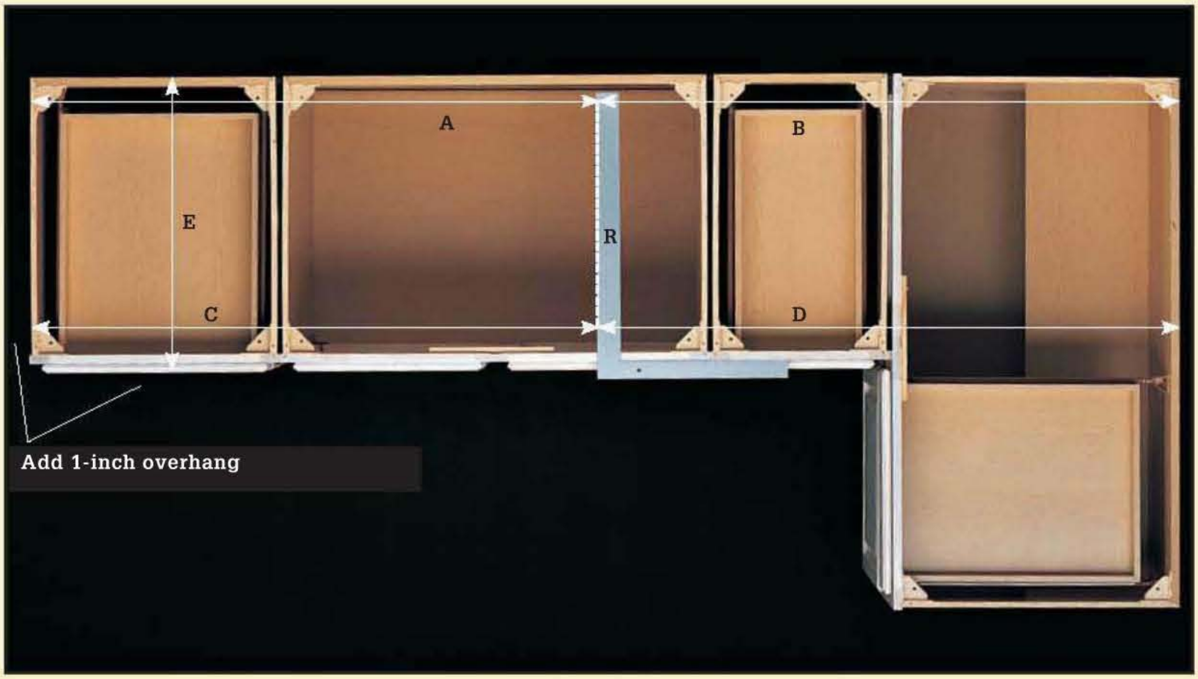

Measure along tops of base cabinets to determine the size of the countertop. If wall corners are not square, use a framing square to establish a reference line (R) near the middle of the base cabinets, perpendicular to the front of the cabinets. Take four measurements (A, B, C, D) from the reference line to the cabinet ends. Allow for overhangs by adding 1" to the length for each exposed end, and 1" to the width (E).

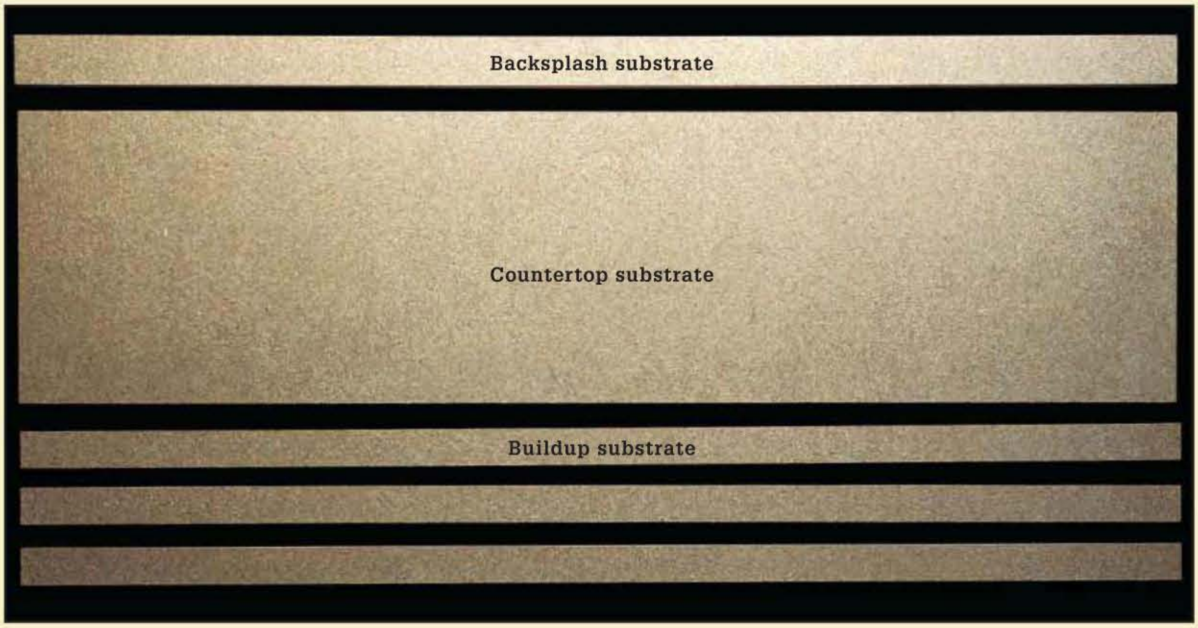

Lay out cutting lines on the particleboard so you can rip-cut the substrate and build-up strips to size, using a framing square to establish a reference line. Cut core to size using a circular saw with clamped straightedge as a guide. Cut 4" strips of particleboard for backsplash, and for joint support where sections of countertop core are butted together. Cut 3" strips for edge buildups.

How to Create a Laminate Countertop

1

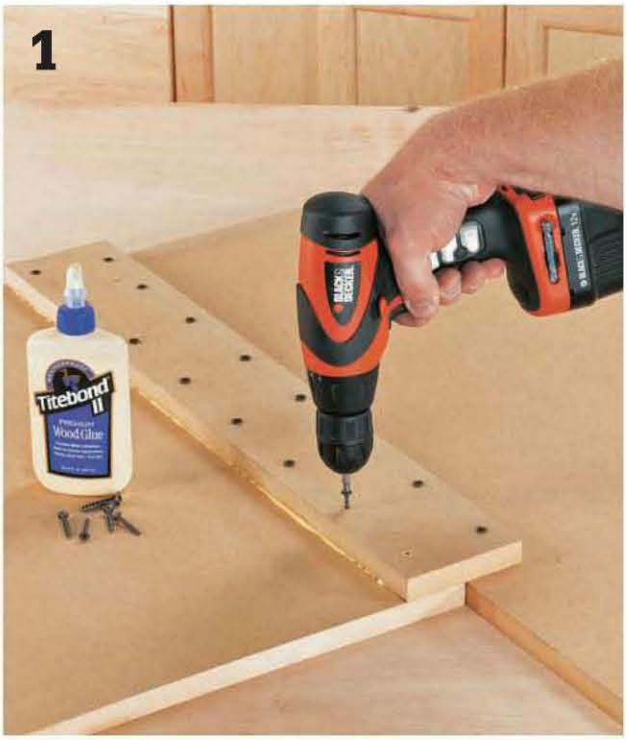

Join the countertop substrate pieces on the bottom side (see page 369 for sizing and cutting substrate). Attach a 4" particleboard joint support across the seam, using carpenter's glue and 1¼" wallboard screws.

2

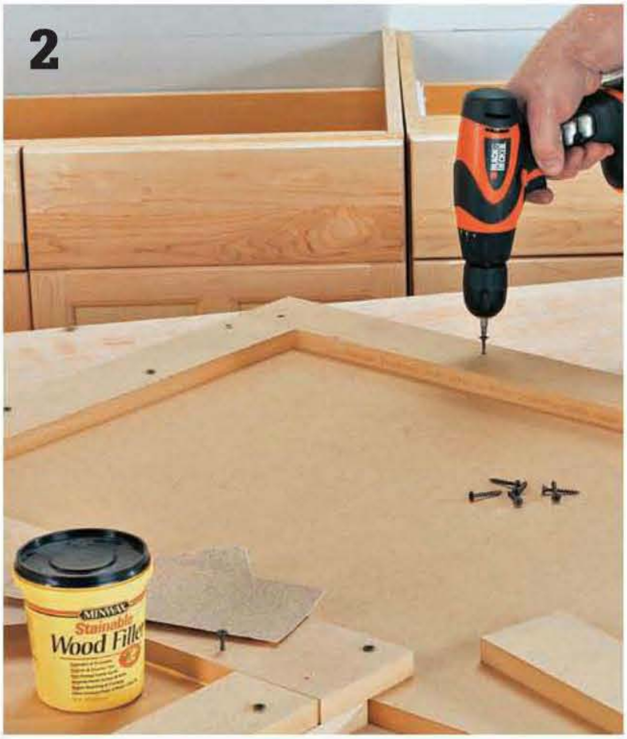

Attach 3"-wide edge buildup strips to the bottom of the countertop, using 1¼" wallboard screws. Fill any gaps on the outside edges with latex wood filler, and then sand the edges with a belt sander.

3

To determine the size of the laminate top, measure countertop substrate. Laminate seams should not overlap substrate. Add ½" trimming margin to both the length and width of each piece. Measure laminate needed for face and edges of backsplash, and for exposed edges of countertop substrate. Add ½" to each measurement.

4

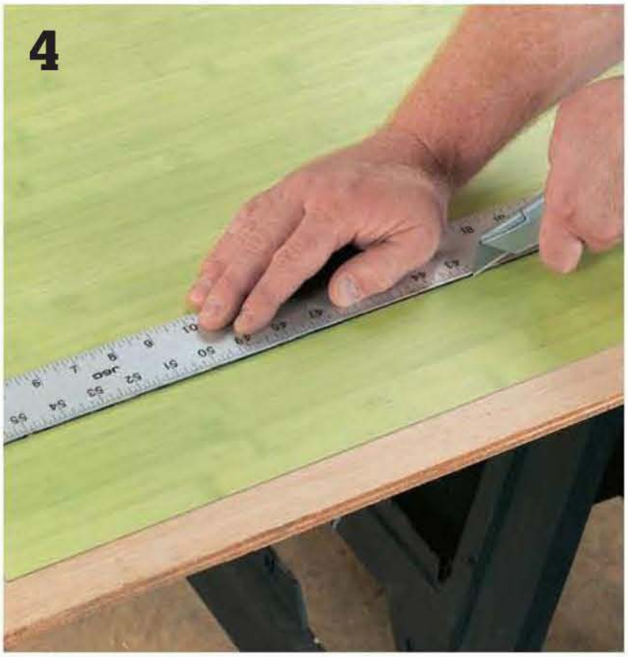

Cut laminate by scoring and breaking it. Draw a cutting line, then etch along the line with a utility knife or other sharp cutting tool. Use a straightedge as a guide. Two passes of scoring tool will help laminate break cleanly.

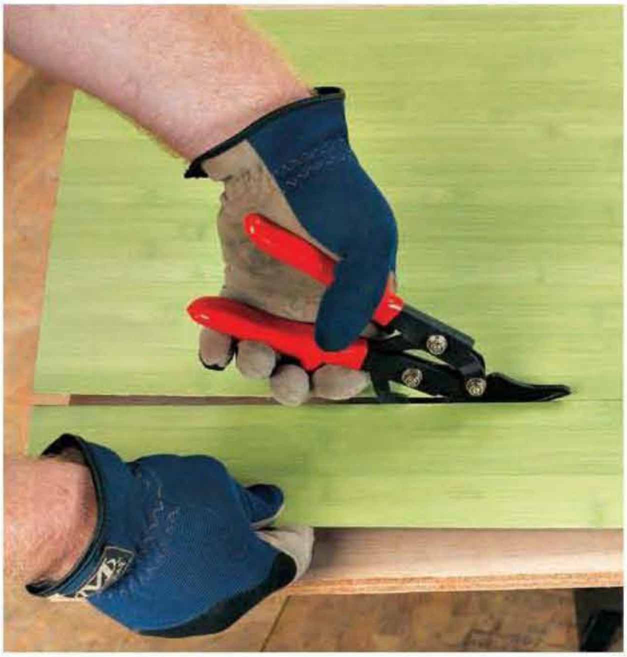

Option: Some laminate installers prefer to cut laminate with special snips that resemble avaiator snips. Available from laminate suppliers, the snips are faster than scoring and snapping, and less likely to cause cracks or tears in the material. You'll still need to square the cut edges with a trimmer or router.

5

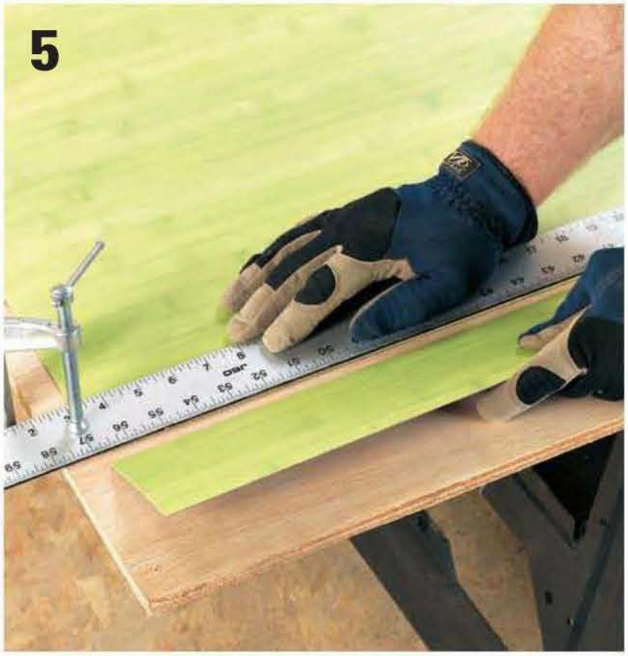

Bend laminate toward the scored line until the sheet breaks cleanly. For better control on narrow pieces, clamp a straightedge along scored line before bending laminate. Wear gloves to avoid being cut by sharp edges.

6

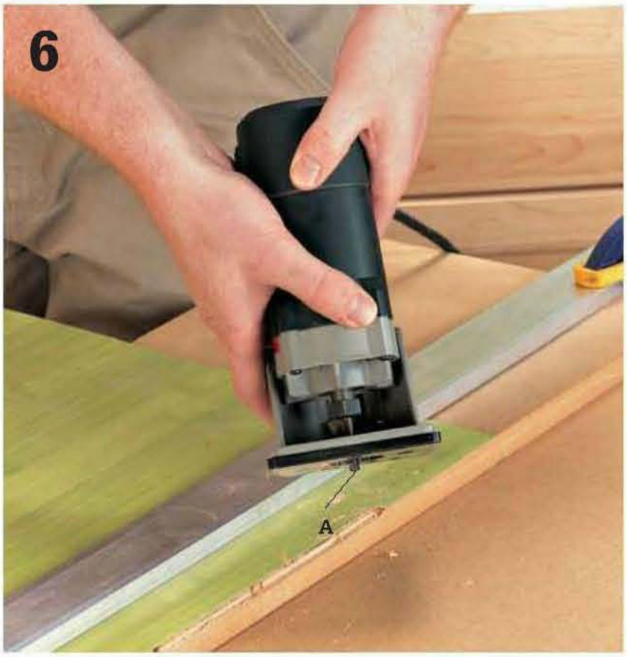

Create tight-piloted seams by using a router and a straight bit to trim edges that will butt together. Measure from cutting edge of the bit to edge of the router baseplate (A). Place laminate on scrap wood and align edges. To guide the router, clamp a straightedge on the laminate at distance A plus ¼", parallel to laminate edge. Trim laminate.

(continued)

7

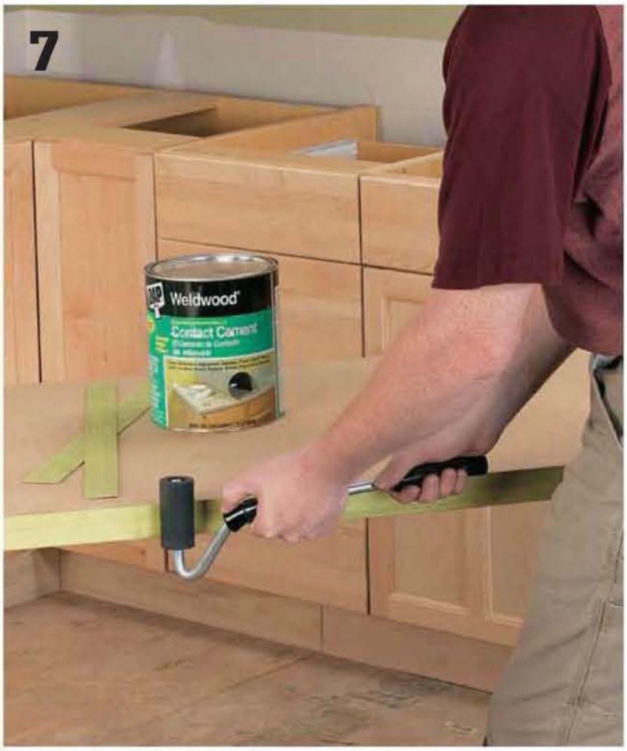

Apply laminate to sides of countertop first. Using a paint roller, apply two coats of contact cement to the edge of the countertop and one coat to back of laminate. Let cement dry according to manufacturer's directions. Position laminate carefully, then press against edge of countertop. Bond with J-roller.

8

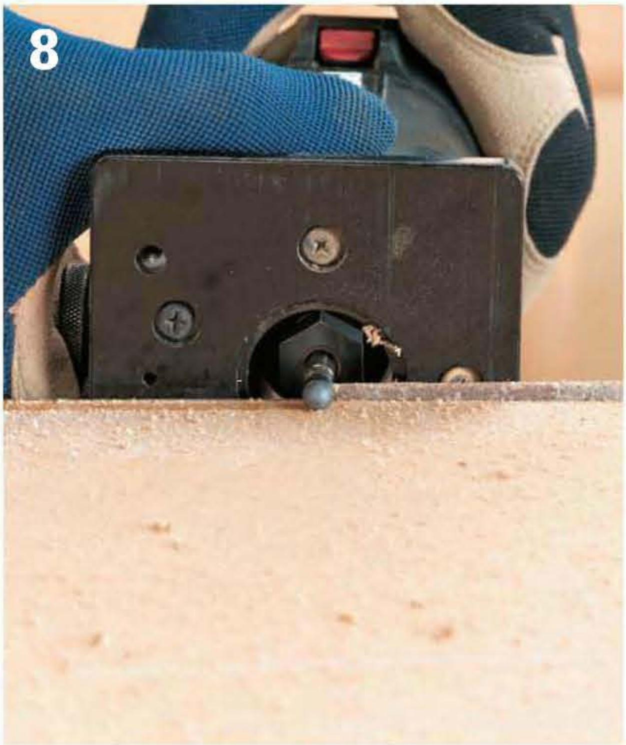

Use a router and flush-cutting bit to trim edge strip flush with top and bottom surfaces of countertop substrate. At edges where router cannot reach, trim excess laminate with a file. Apply laminate to remaining edges, and trim with router.

9

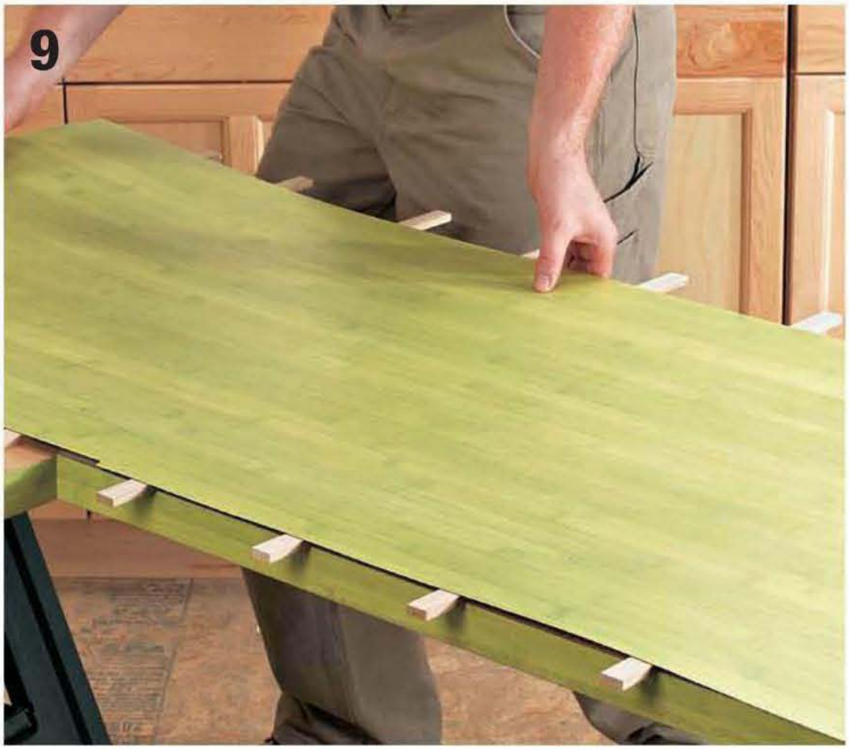

Test-fit laminate top on countertop substrate. Make sure laminate overhangs all edges. At seam locations, draw a reference line on core where laminate edges will butt together. Remove laminate. Make sure all surfaces are free of dust, then apply one coat of contact cement to back of laminate and two coats to substrate. Place spacers made of ¼"-thick scrap wood at 6" intervals across countertop core. Because contact cement bonds instantly, spacers allow laminate to be positioned accurately over core without bonding. Align laminate with seam reference line. Beginning at one end, remove spacers and press laminate to countertop core.

10

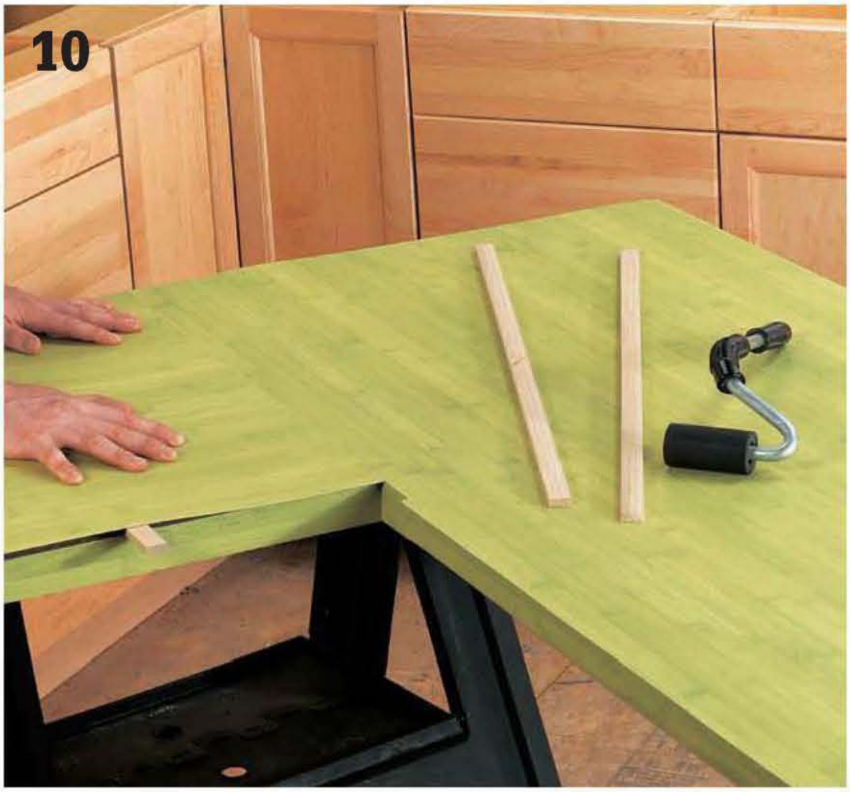

Apply contact cement to remaining substrate and next piece of laminate. Let cement dry, then position laminate on spacers, and carefully align butt seam. Beginning at seam edge, remove spacers and press laminate to the countertop substrate.

11

Roll the entire surface with a J-roller to bond the laminate to the substrate. Clean off any excess contact cement with a soft cloth and mineral spirits.

12

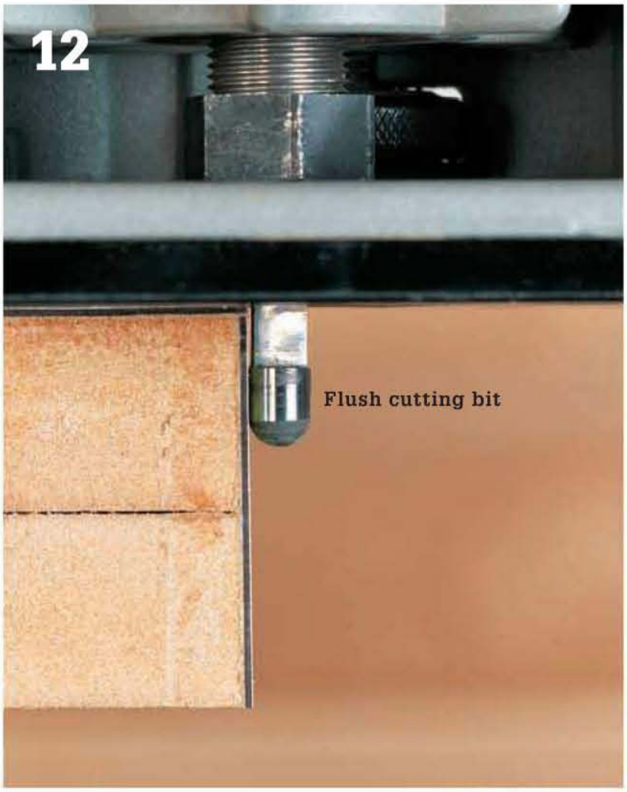

Remove excess laminate with a router and flush-cutting bit. At edges where router cannot reach, trim excess laminate with a file. Countertop is now ready for final trimming with bevel-cutting bit.

(continued)

13

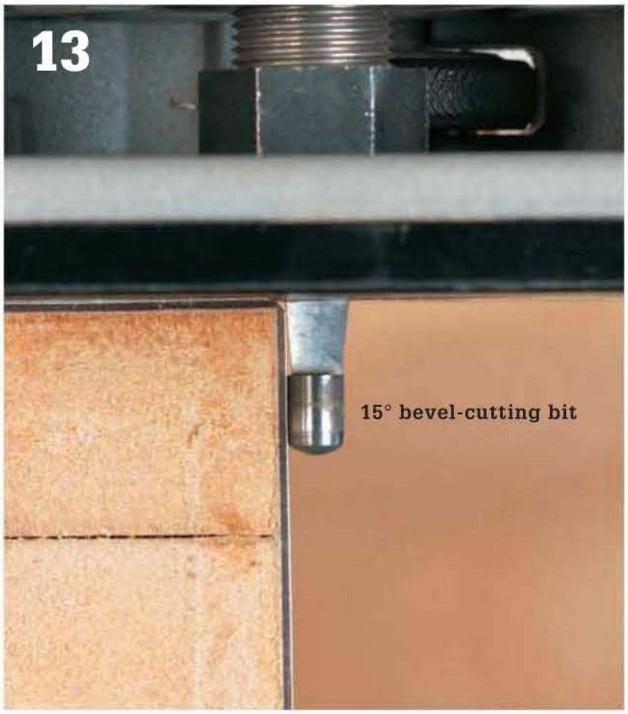

Finish-trim the edges with router and 15° bevel-cutting bit. Set bit depth so that the bevel edge is cut only on top laminate layer. Bit should not cut into vertical edge surface.

Tip ▸

File all edges smooth. Use downward file strokes to avoid chipping the laminate.

14

Cut 1¼"-wide strips of ¼" plywood to form an overhanging scribing strip for the backsplash. Attach to the top and sides of the backsplash substrate with glue and wallboard screws. Cut laminate pieces and apply to exposed sides, top and front of backsplash. Trim each piece as it is applied.

15

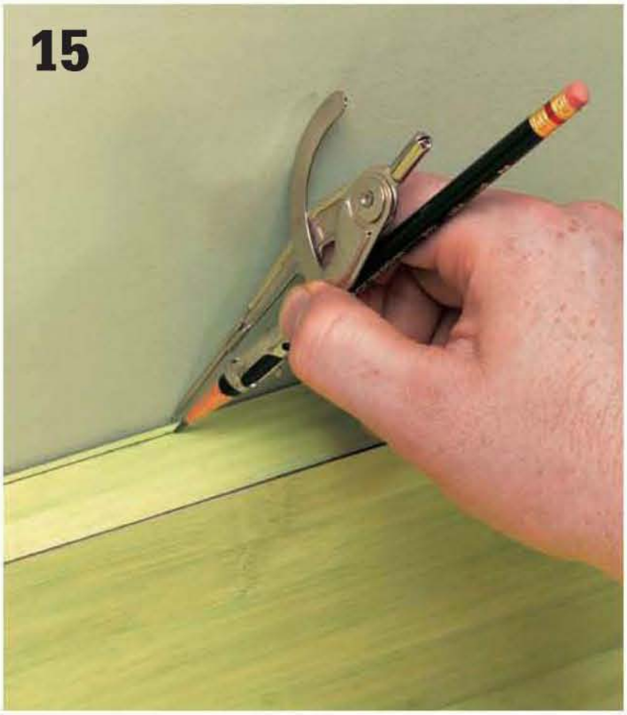

Test-fit the countertop and backsplash. Because your walls may be uneven, use a compass to trace the wall outline onto the backsplash scribing strip. Use a belt sander to grind backsplash to scribe line (see page 366).

16

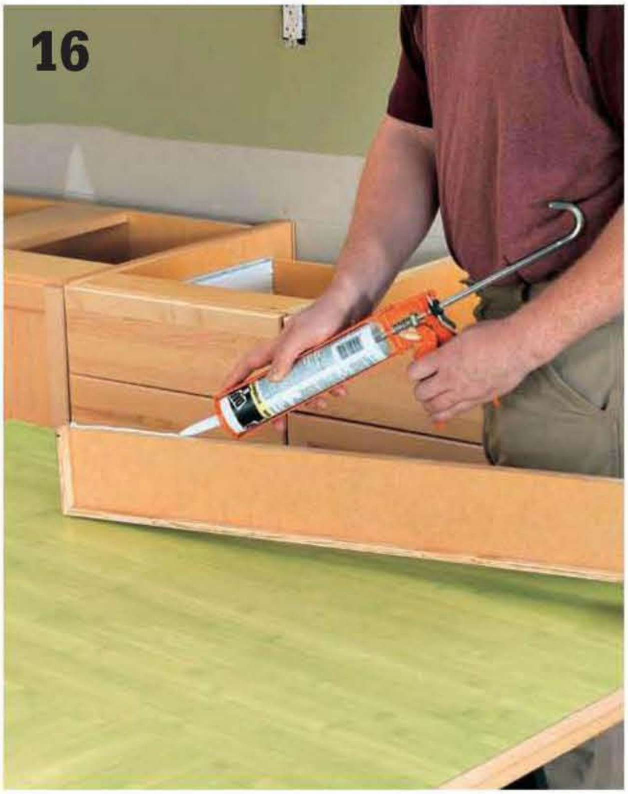

Apply a bead of silicone caulk to the bottom edge of the backsplash.

17

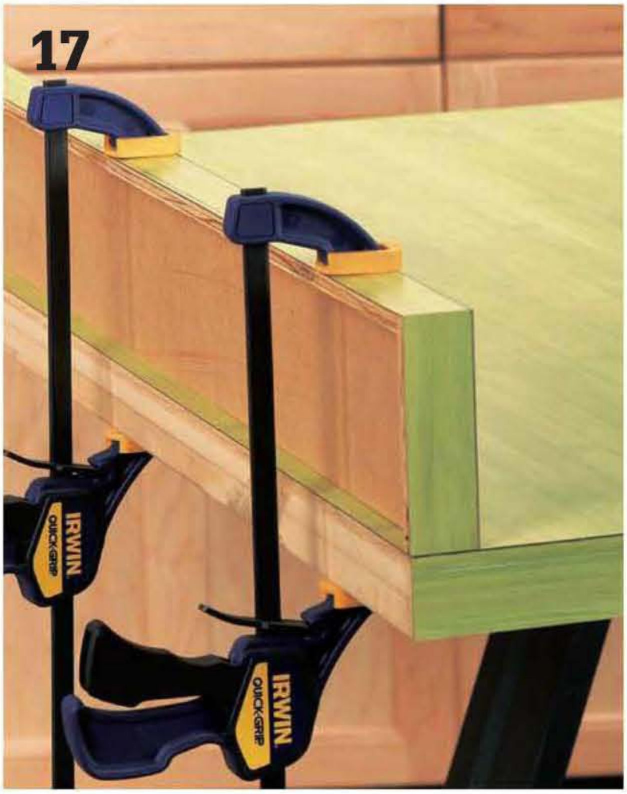

Position the backsplash on the countertop, and clamp it into place with bar clamps. Wipe away excess caulk, and let dry completely.

18

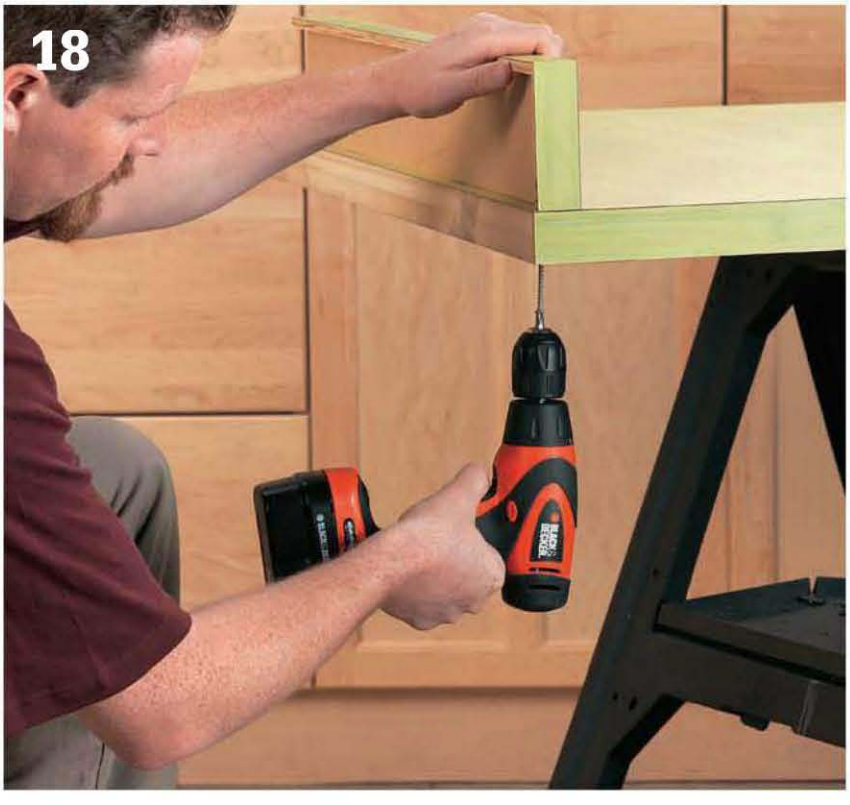

Drive 2" wallboard screws through the countertop and into the backsplash core. Make sure screw heads are countersunk completely for a tight fit against the base cabinet. Install countertops.

Installing Tile on Countertops

Ceramic and porcelain tile remain popular choices for countertops and backsplashes for a number of reasons: It's available in a vast range of sizes, styles and colors; it's durable and repairable; and some tile—not all—is reasonably priced. With careful planning, tile is also easy to install, making a custom countertop a good do-it-yourself project.

The best tile for most countertops is glazed ceramic or porcelain floor tile. Glazed tile is better than unglazed because of its stain resistance, and floor tile is better than wall tile because it's thicker and more durable.

While glazing protects tile from stains, the grout between tiles is still vulnerable because it's so porous. To minimize staining, use a grout that contains a latex additive, or mix the grout with a liquid latex additive. After the grout cures fully, apply a quality grout sealer, and reapply the sealer once a year thereafter. Choosing larger tiles reduces the number of grout lines. Although the selection is a bit limited, if you choose 13" × 13" floor tile, you can span from the front to the back edge of the countertop with a single seam.

The countertop in this project has a substrate of ¾" exterior-grade plywood that's cut to fit and fastened to the cabinets. The plywood is covered with a layer of plastic (for a moisture barrier) and a layer of ½"-thick cementboard. Cementboard is an effective backer for tile because it won't break down if water gets through the tile layer. The tile is adhered to the cementboard with thin-set adhesive. The overall thickness of the finished countertop is about 1½". If you want a thicker countertop, you can fasten an additional layer of plywood (of any thickness) beneath the substrate. Two layers of ¾" exterior-grade plywood without cementboard is also an acceptable substrate.

You can purchase tiles made specifically to serve as backsplashes and front edging. While the color and texture may match, these tiles usually come in only one length, making it difficult to get your grout lines to align with the field tiles. You can solve this problem by cutting your own edging and backsplash tiles from field tiles.

Tools & Materials ▸

Tape measure
Circular saw
Drill
Utility knife
Straightedge
Stapler
Drywall knife
Framing square
Notched trowel
Tile cutter
Carpeted 2 × 4
Mallet
Rubber grout float
Sponge
Foam brush
Caulk gun
Ceramic tile
Tile spacers
¾" exterior-grade (CDX) plywood
4-mil polyethylene sheeting
Packing tape
½" cementboard
1¼" galvanized deck screws
Fiberglass mesh tape
Thin-set mortar
Grout with latex additive
Silicone caulk
Silicone grout sealer

Ceramic or porcelain makes a durable countertop that is heat-resistant and relatively easy for a DIYer to create. By using larger tiles, you minimize the grout lines (and the cleaning that goes with them).

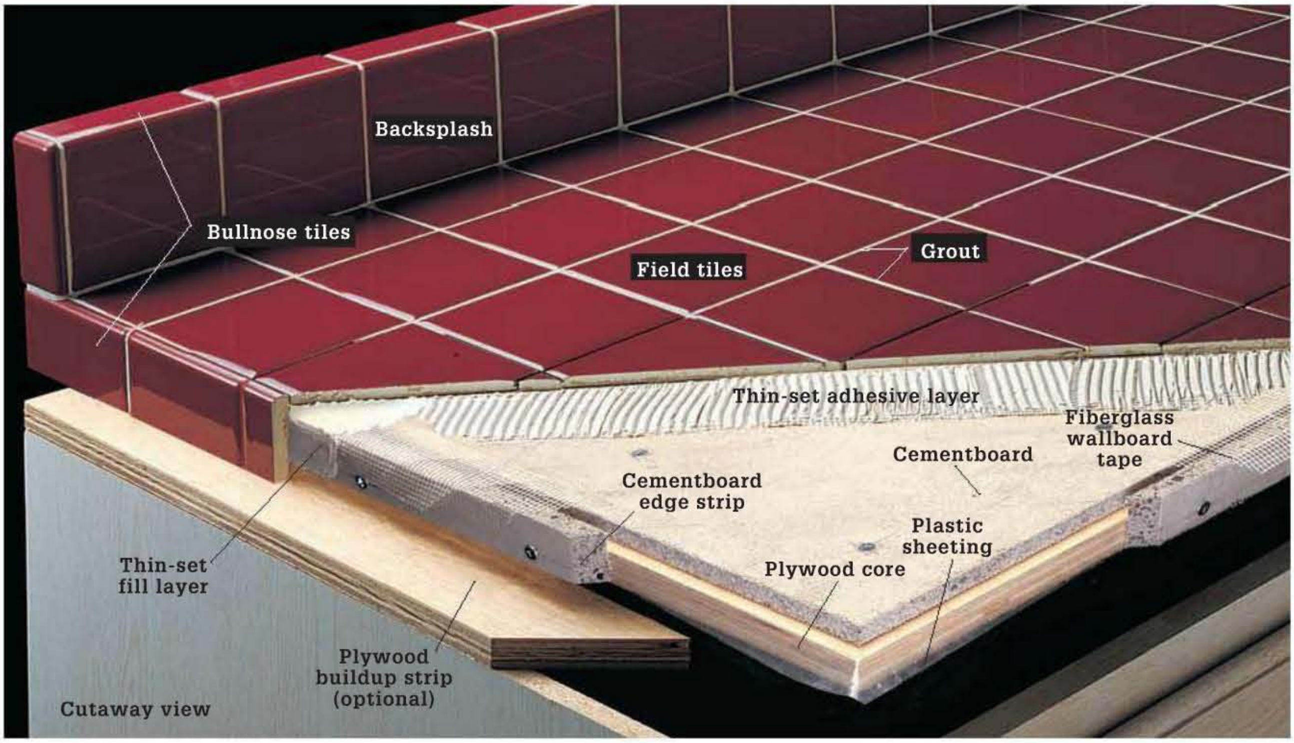
Backsplash
Bullnose tiles
Field tiles
Grout
Thin-set adhesive layer
Fiberglass wallboard tape
Cementboard
Cementboard edge strip
Plastic sheeting
Plywood core
Thin-set fill layer
Plywood buildup strip (optional)
Cutaway view

A ceramic tile countertop starts with a core of ¾" exterior-grade plywood that's covered with a moisture barrier of 4-mil polyethylene sheeting. Half-inch cementboard is screwed to the plywood, and the edges are capped with cementboard and finished with fiberglass mesh tape and thin-set mortar. Tiles for edging and backsplashes may be bullnose or trimmed from the factory edges of field tiles.

Options for Backsplashes & Countertop Edges

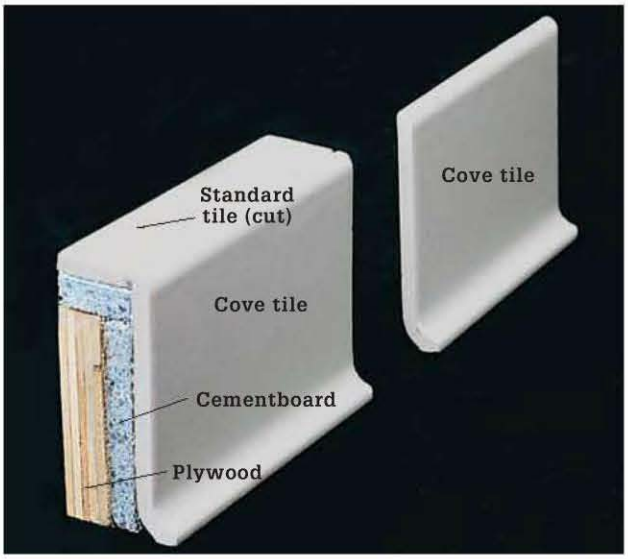
Standard tile (cut)
Cove tile
Cove tile
Cementboard
Plywood

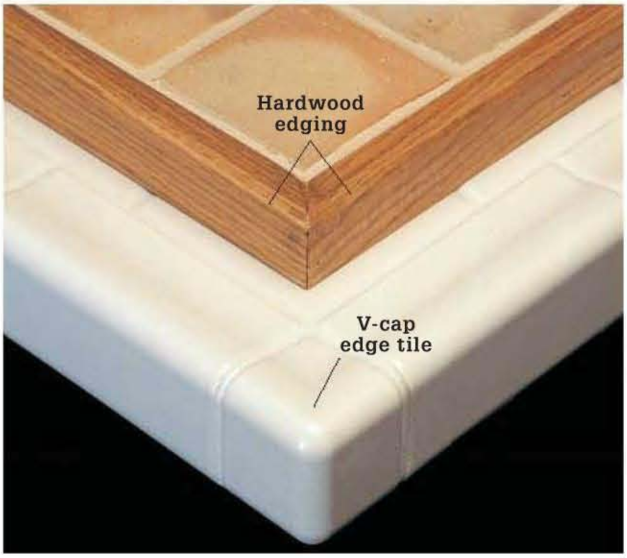
Hardwood edging
V-cap edge tile

Backsplashes can be made from cove tile attached to the wall at the back of the countertop. You can use the tile alone or build a shelf-type backsplash, using the same construction as for the countertop. Attach the plywood backsplash to the plywood core of the countertop. Wrap the front face and all edges of the plywood backsplash with cementboard before laying tile.

Edge options include V-cap edge tile and hardwood strip edging. V-cap tiles have raised and rounded corners that create a ridge around the countertop perimeter—good for containing spills and water. V-cap tiles must be cut with a tile saw. Hardwood strips should be prefinished with at least three coats of polyurethane finish. Attach the strips to the plywood core so the top of the wood will be flush with the faces of the tiles.

Tips for Laying Out Tile ▸

- You can lay tile over a laminate countertop that's square, level and structurally sound. The laminate cannot have a no-drip edge. Use a belt sander with 60- or 80-grit sandpaper to rough up the surface before setting the tiles. If you're using a new substrate and need to remove your existing countertop, make sure the base cabinets are level front to back, side to side and with adjoining cabinets. If necessary, unscrew a cabinet from the wall and use shims on the floor or against the wall to level it.
- Installing battens along the front edge of the countertop helps ensure the first row of tile is perfectly straight. For V-cap tiles, fasten a 1 × 2 batten along the reference line, using screws. The first row of field tile is placed against this batten. For bullnose tiles, fasten a batten that's the same thickness as the edging tile, plus ⅛" for mortar thickness, to the face of the countertop so the top is flush with the top of the counter. The bullnose tiles are aligned with the outside edge of the batten. For wood edge trim, fasten a 1 × 2 batten to the face of the countertop so the top edge is above the top of the counter. The tiles are installed against the batten.
- Before installing any tile, lay out the tiles in a dry run, using spacers. If your counter is L-shaped, start at the corner and work outward. Otherwise, start the layout at a sink to ensure equal-sized cuts on both sides of the sink. If necessary, shift your starting point so you don't end up cutting very narrow tile segments.

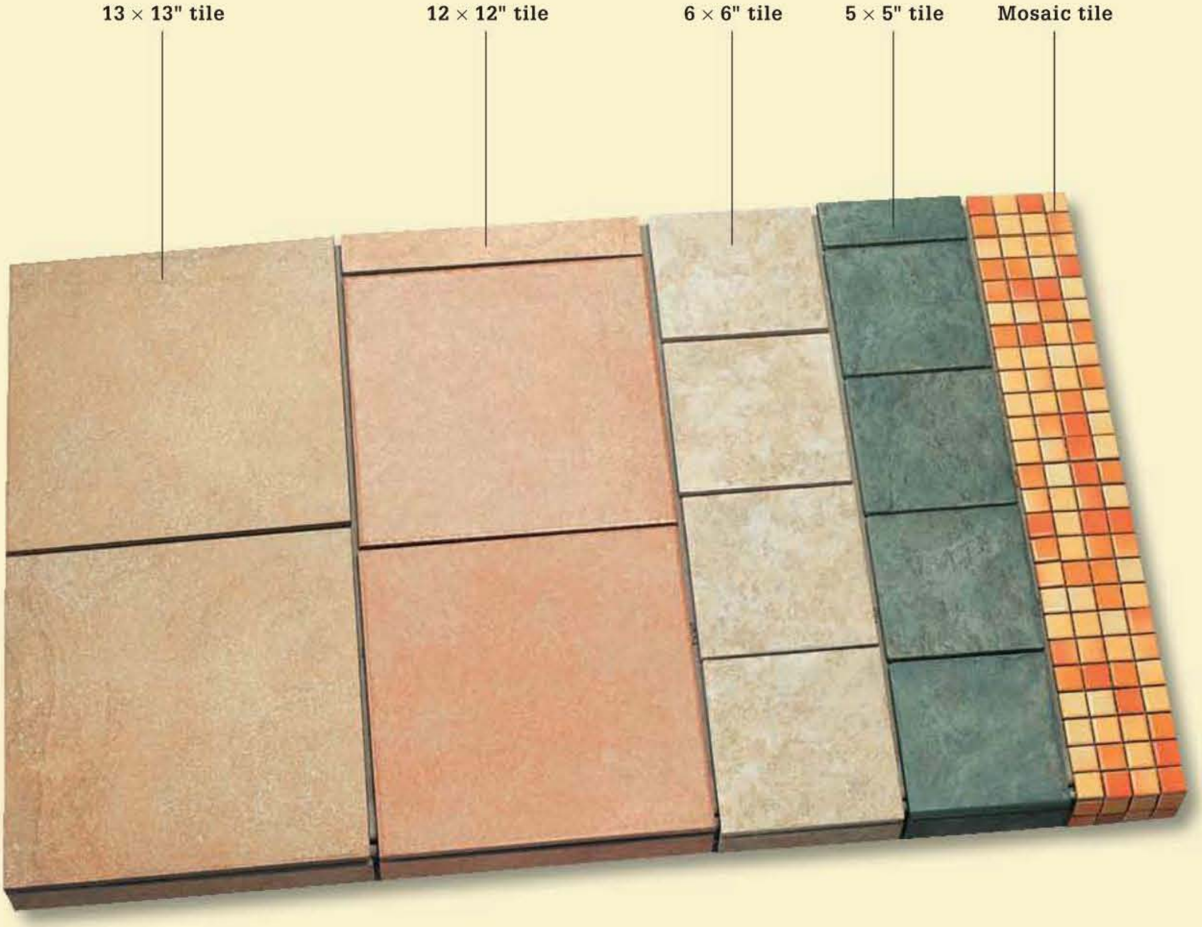

The bigger the tile, the fewer the grout lines. If you want a standard 25"-deep countertop, the only way to get there without cutting tiles is to use mosaic strips or 1" tile. With 13 × 13" tile, you need to trim 1" off the back tile but have only one grout line front to back. As the size of the tile decreases, the number of grout lines increases.

How to Build a Tile Countertop

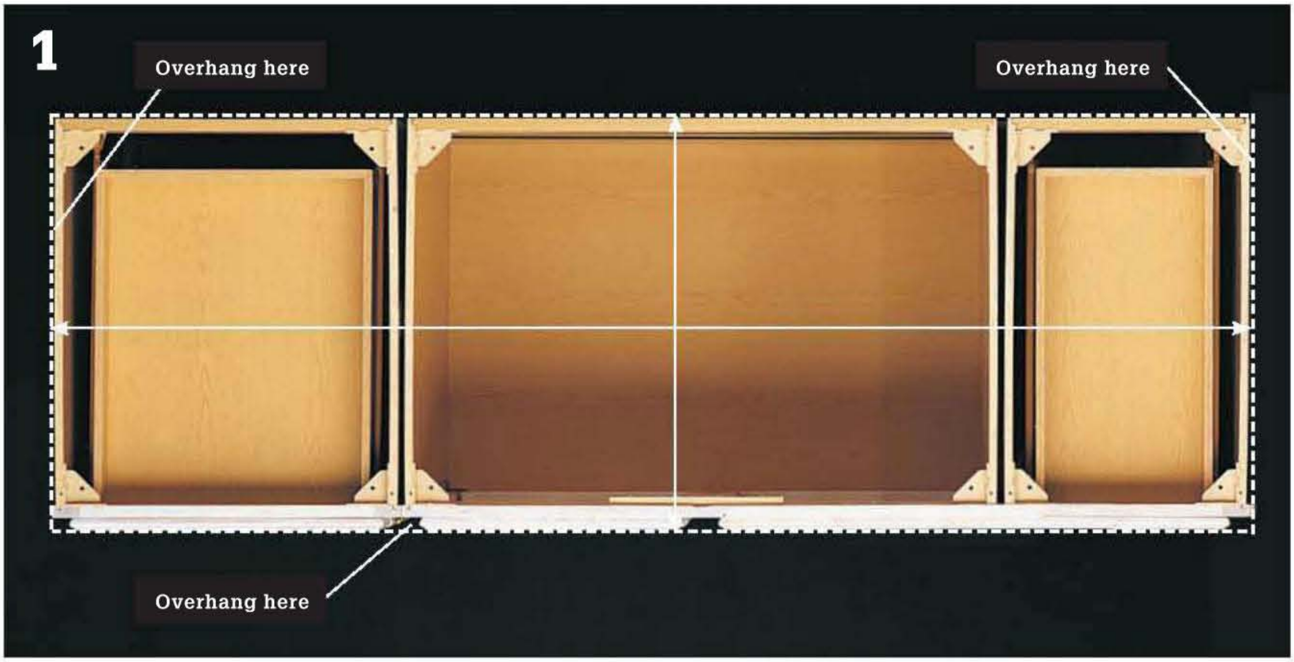

Determine the size of the plywood substrate by measuring across the top of the cabinets. The finished top should overhang the drawer fronts by at least ¼". Be sure to account for the thickness of the cementboard, adhesive and tile when deciding how large to make the overhang. Cut the substrate to size from ¾" plywood, using a circular saw. Make cutouts for sinks and other fixtures.

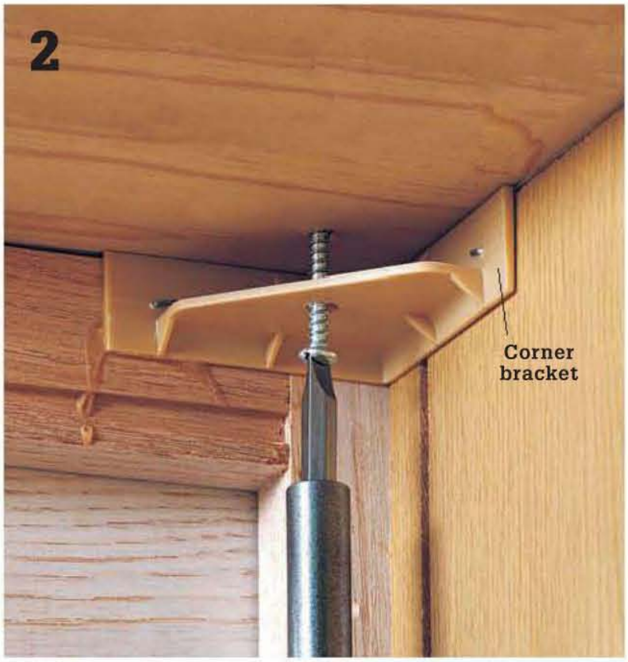

Set the plywood substrate on top of the cabinets, and attach it with screws driven through the cabinet corner brackets. The screws should not be long enough to go through the top of the substrate.

Cut cementboard using a straightedge and utility knife or a cementboard cutter with a carbide tip. Hold the straightedge along the cutting line, and score the board several times with the knife. Bend the piece backward to break it along the scored line. Back-cut to finish.

(continued)

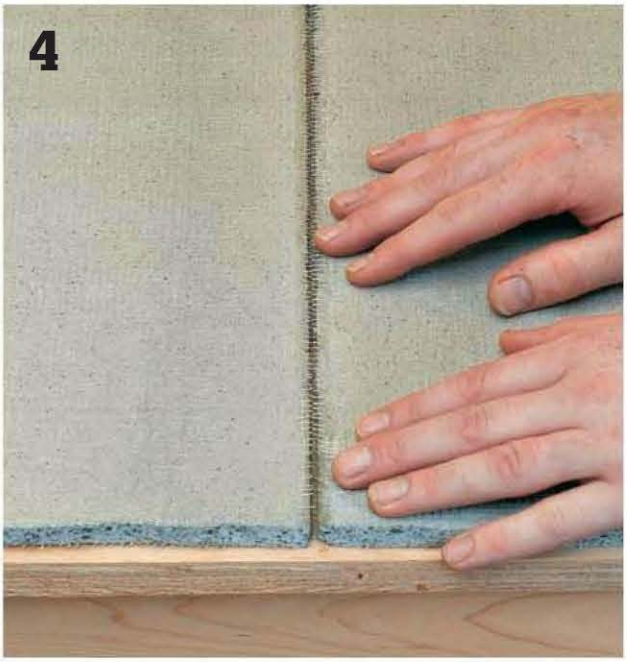

4

Dry-fit cementboard pieces on the plywood core with the rough sides of the panels facing up. Leave a ⅛" gap between the cementboard sheets and a ¼" gap along the perimeter.

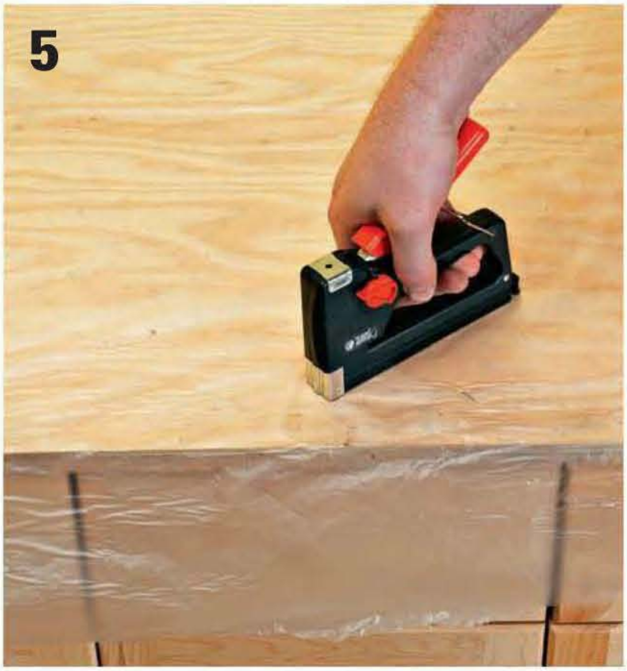

5

Lay the 4-mil plastic moisture barrier over the plywood substrate, draping it over the edges. Tack it in place with a few staples. Overlap seams in the plastic by 6", and seal them with packing tape.

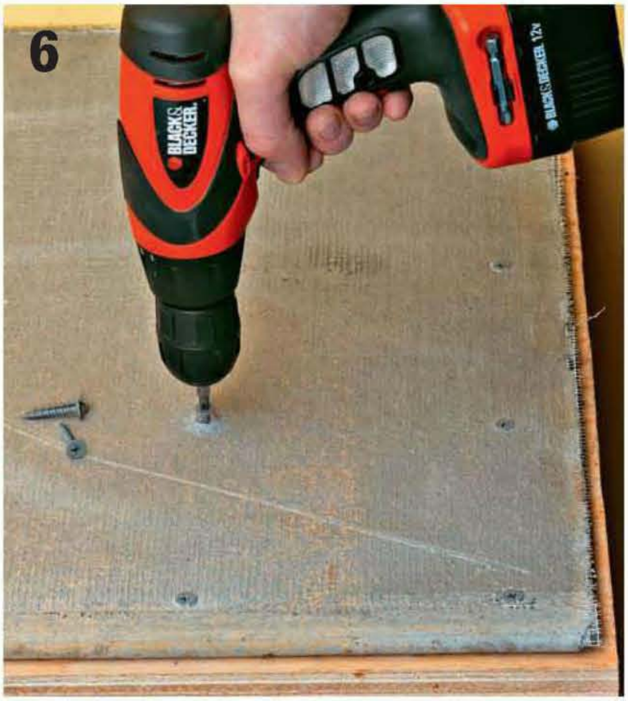

6

Lay the cementboard pieces rough-side up on the plywood and attach them with cementboard screws driven every 6". Drill pilot holes using a masonry bit, and make sure all screw heads are flush with the surface. Wrap the countertop edges with 1¼"-wide cementboard strips, and attach them to the core with cementboard screws.

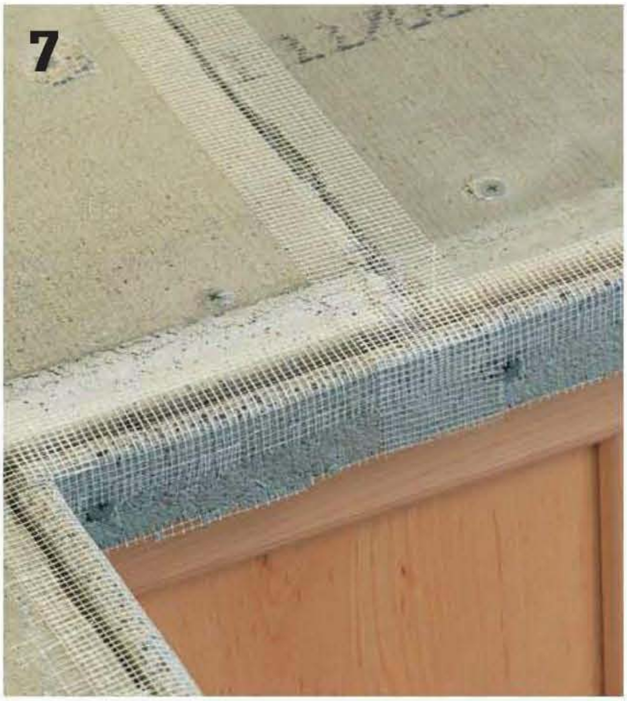

7

Tape all cementboard joints with fiberglass mesh tape. Apply three layers of tape along the front edge where the horizontal cementboard sheets meet the cementboard edging.

Fill all gaps and cover all of the tape with a layer of thin-set mortar. Feather out the mortar with a drywall knife to create a smooth, flat surface.

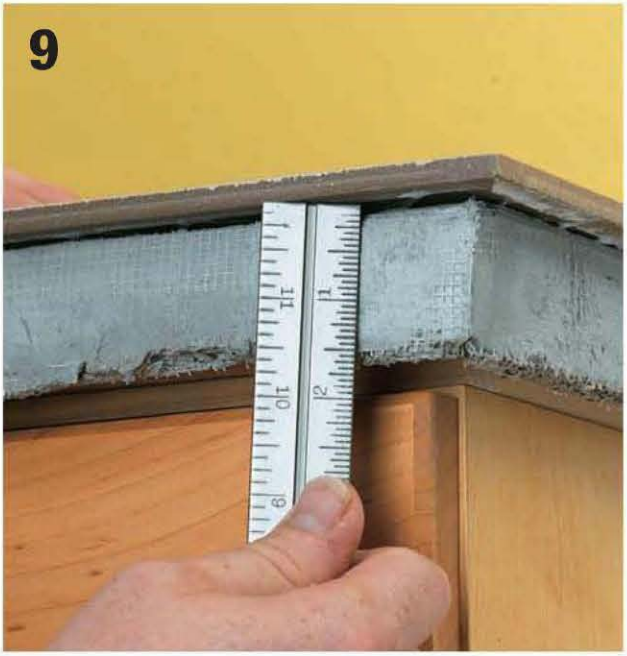

Determine the required width of your edge tiles. Lay a field tile onto the tile base so it overhangs the front edge by ½" or so. Then, hold a metal rule up to the underside of the tile and measure the distance from the tile to the bottom of the subbase. Your edge tiles should be cut to this width (the gap for the grout line will cause the edge tile to extend past the subbase, concealing it completely).

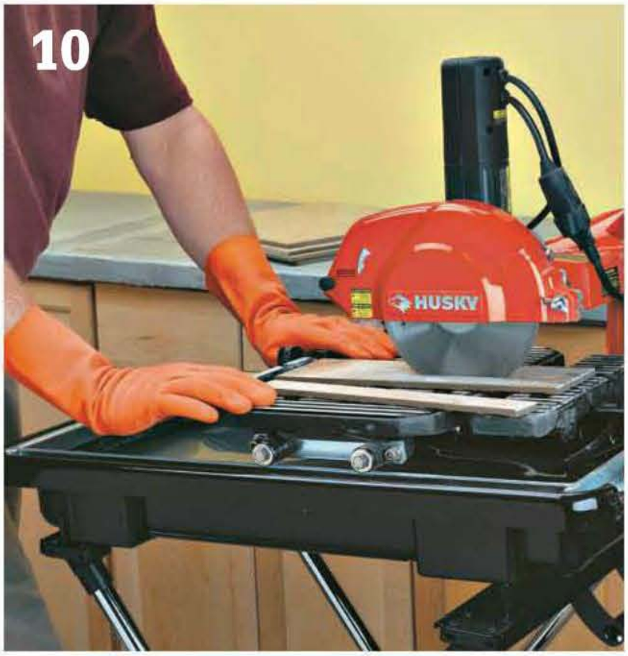

Cut your edge tiles to the determined width, using a tile saw. It's worth renting a quality wet saw for tile if you don't own one. Floor tile is thick and difficult to cut with a hand cutter (especially porcelain tiles).

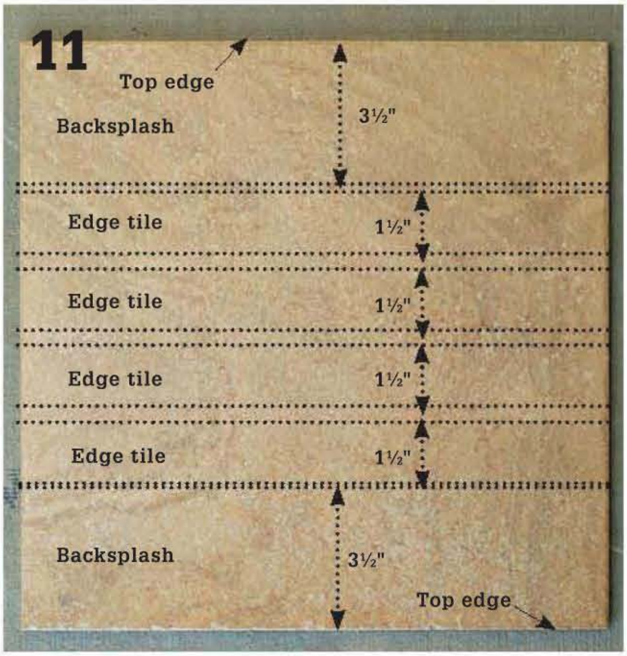

Cut tiles for the backsplash. The backsplash tiles (3½" wide in our project) should be cut with a factory edge on each tile that will be oriented upward when they're installed. You can make efficient use of your tiles by cutting edge tiles from the center area of the tiles you cut to make the backsplash.

(continued)

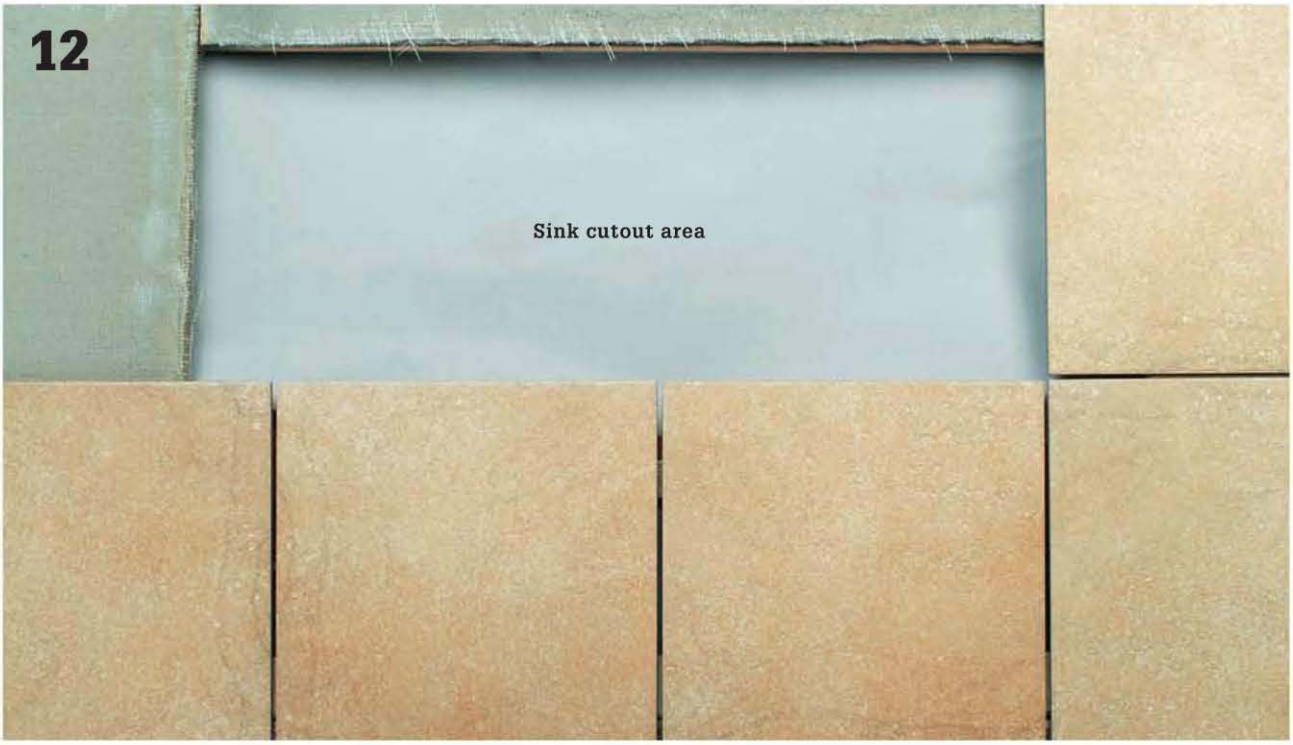

Dry-fit tiles on the countertop to find the layout that works best. Once the layout is established, make marks along the vertical and horizontal rows. Draw reference lines through the marks and use a framing square to make sure the lines are perpendicular.

Variation: Laying Out with Small Floor Tiles and Bullnose Edging ▸

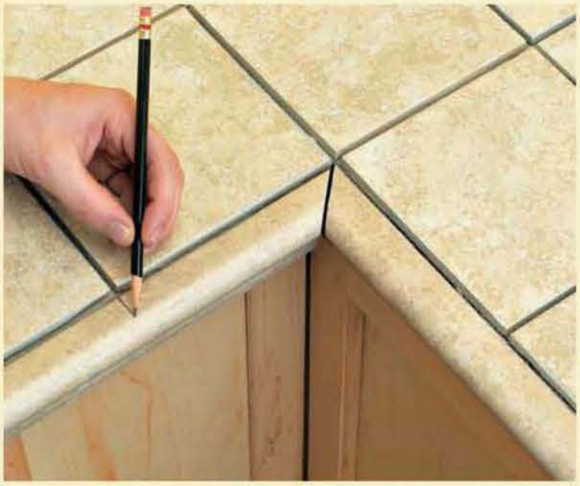

Lay out tiles and spacers in a dry run. Adjust starting lines, if necessary. If using battens, lay the field tile flush with the battens, then apply edge tile. Otherwise, install the edging first. If the countertop has an inside corner, start there by installing a ready-made inside corner or cutting a 45° miter in edge tile to make your own inside corner.

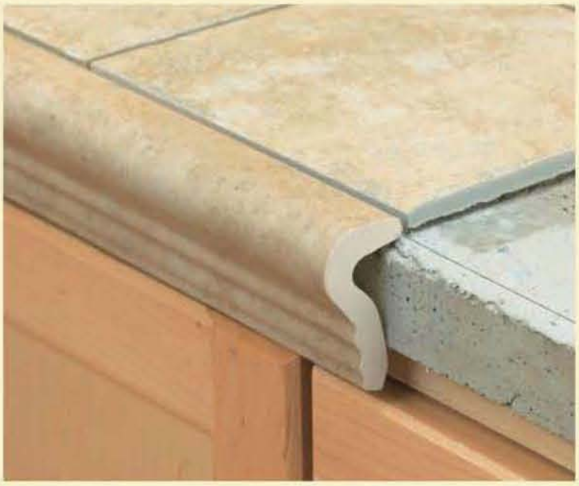

Place the first row of field tile against the edge tile, separating the tile with spacers. Lay out the remaining rows of tile. Adjust starting lines if necessary to create a layout using the least number of cut tiles.

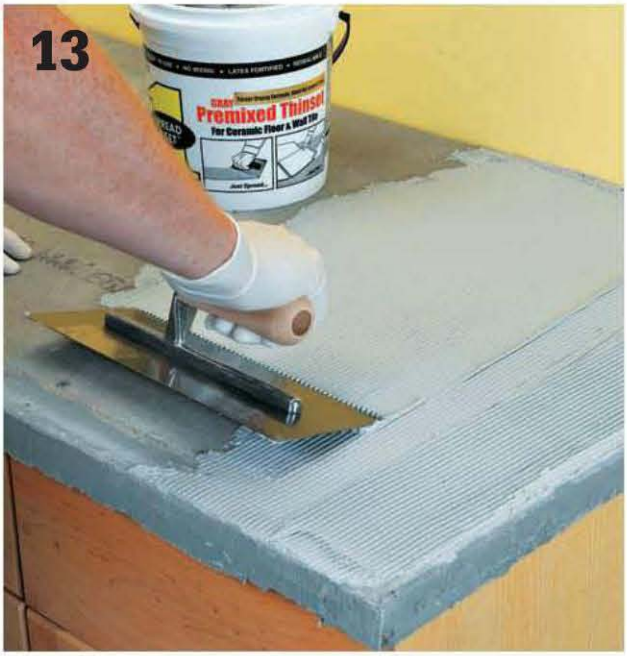

Use a ⅜" square notched trowel to apply a layer of thinset adhesive to the cementboard. Apply enough for two or three tiles, starting at one end. Hold the trowel at roughly a 30-degree angle and try not to overwork the adhesive or remove too much.

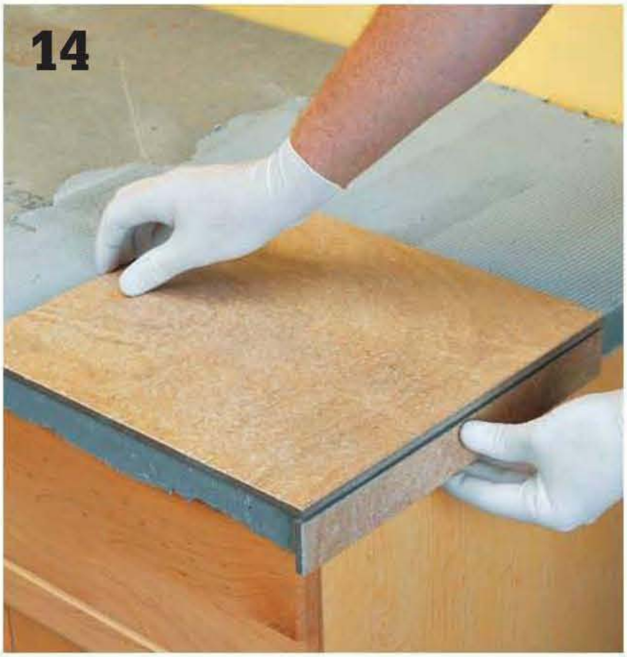

Set the first tile into the adhesive. Hold a piece of the edge in tile against the countertop edge as a guide to show you exactly how much the tile should overhang the edge.

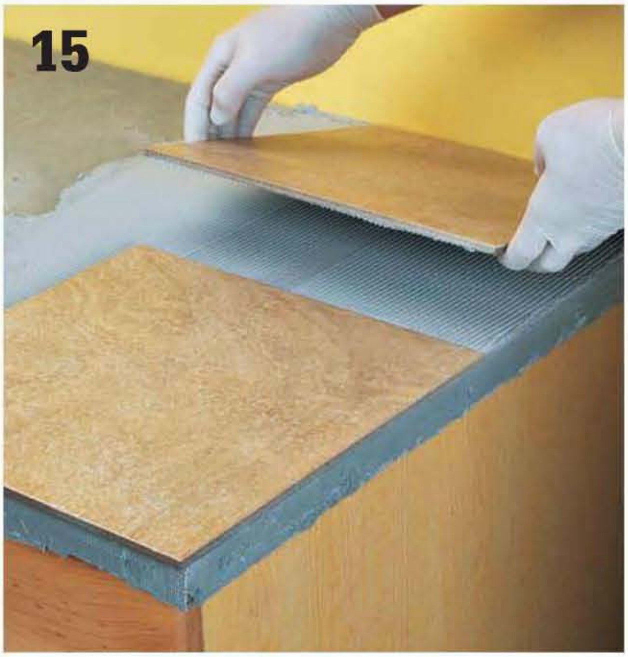

Cut all the back tiles for the layout to fit (you'll need to remove about 1" of a 13 × 13" tile) before you begin the actual installation. Set the back tiles into the thinset, maintaining the gap for groutlines created by the small spacer nubs cast into the tiles. If your tiles have no spacer nubs, see next step.

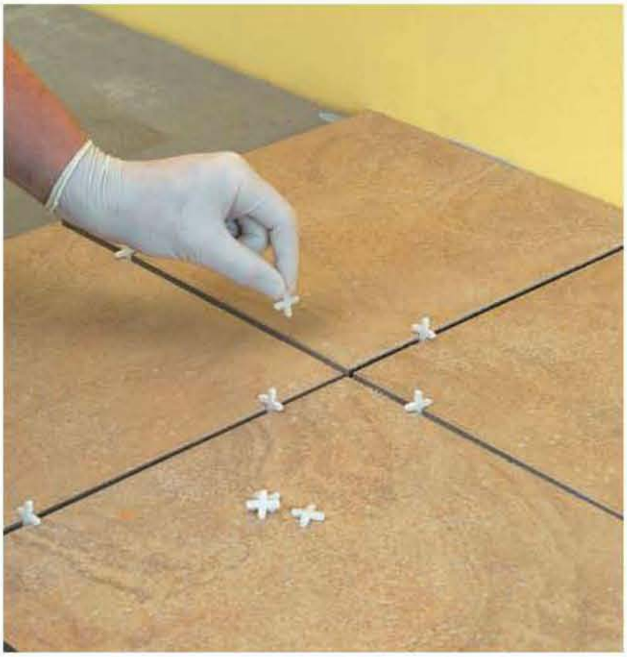

Option: To maintain even grout lines, some beginning tilers insert x-shaped plastic spacers at the joints. This is less likely to be useful with large tiles like those shown here, but it is effective with smaller tile. Many tiles today feature built-in spacing lugs, so the spacers are not necessary. If you use spacers, remove them before the thinset sets. If you leave them in place they will corrupt your grout lines.

(continued)

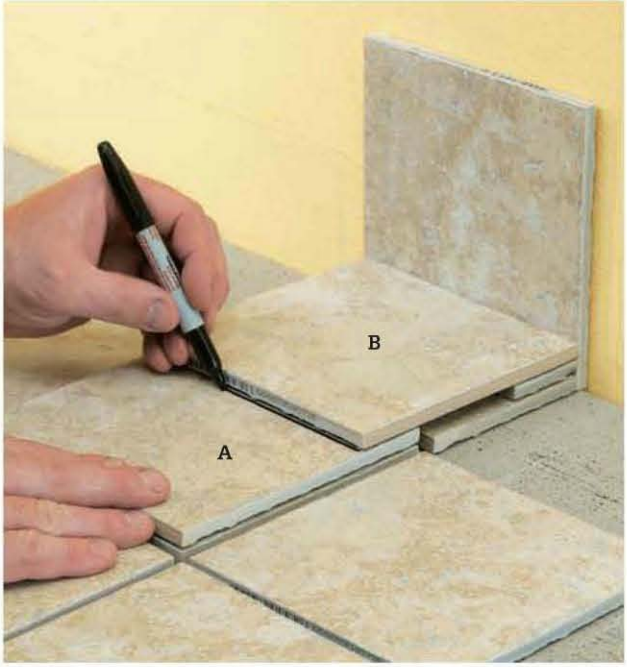

Variation: To mark border tiles for cutting, allow space for backsplash tiles, grout and mortar by placing a tile against the back wall. Set another tile (A) on top of the last full tile in the field, then place a third tile (B) over tile A and hold it against the upright tile. Mark and cut tile A and install it with the cut edge toward the wall. Finish filling in your field tiles.

16

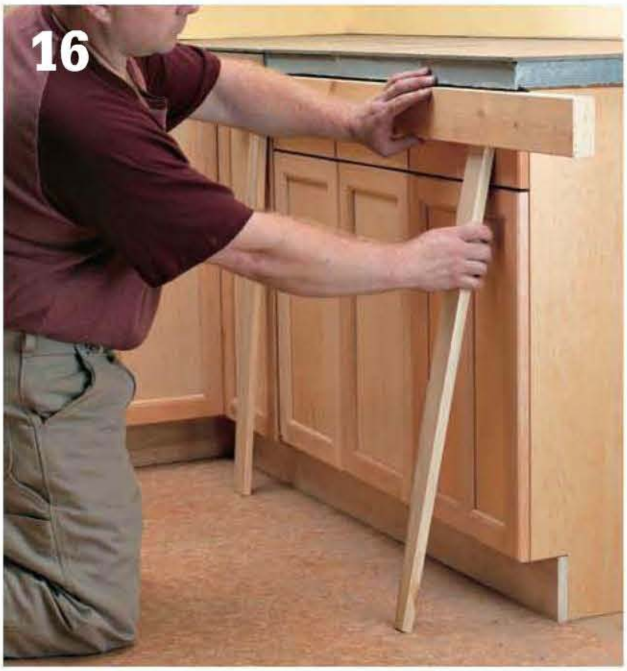

To create a support ledge for the edge tiles, prop pieces of 2 × 4 underneath the front edge of the substrate overhang, using wood scraps to prop the ledge tightly up against the substrate.

17

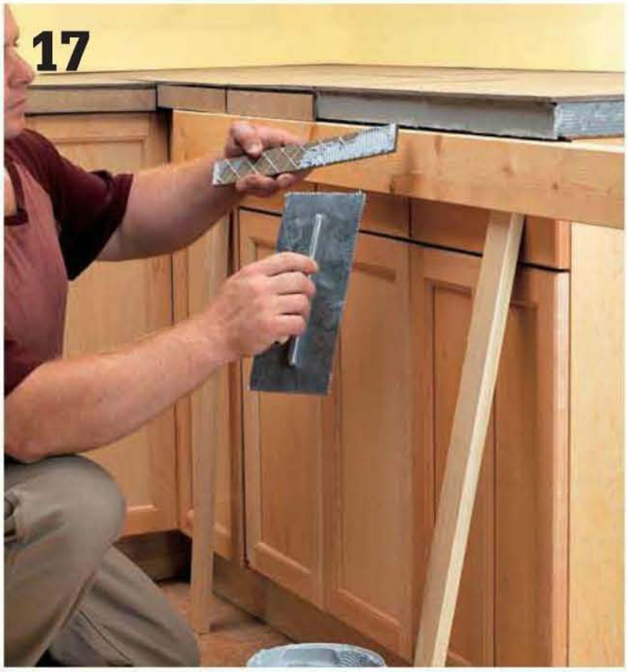

Apply a thick layer of thinset to the backside of the edge tile with your trowel. This is called "buttering" and it is easier and neater than attempting to trowel adhesive onto the countertop edge. Press the tiles into position so they are flush with the leading edges of the field tiles.

18

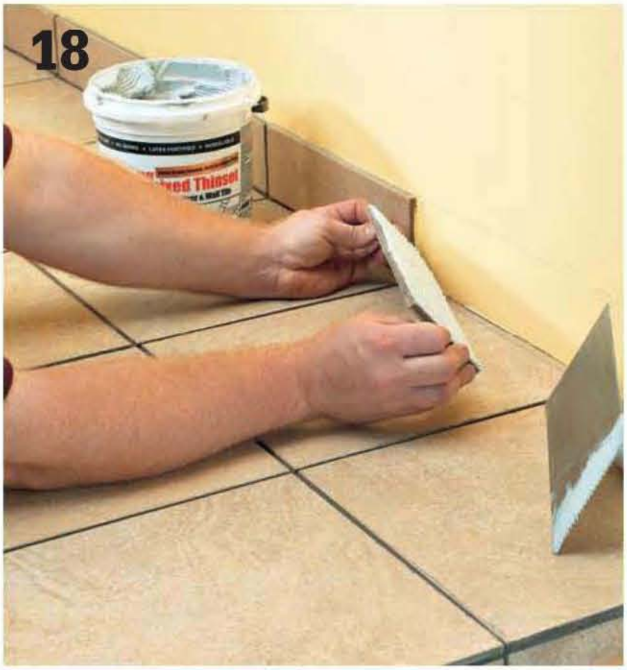

Butter each backsplash tile and press it into place, doing your best to keep all of the grout lines aligned.

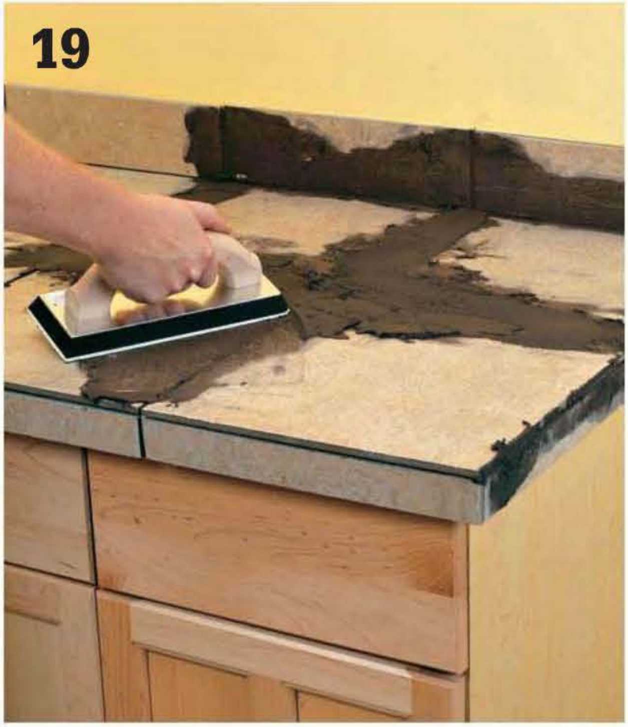

Mix a batch of grout to complement the tile (keeping in mind that darker grout won't look dirty as quickly as lighter grout). Apply the grout to the grout line areas with a grout float.

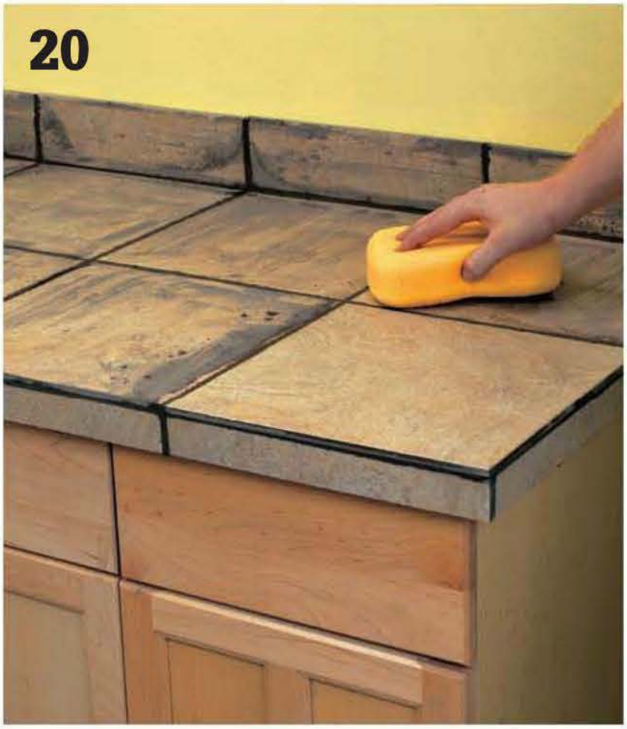

Let the grout dry until a light film is created on the countertop surface and then wipe off the excess grout using a sponge and warm, clean water.

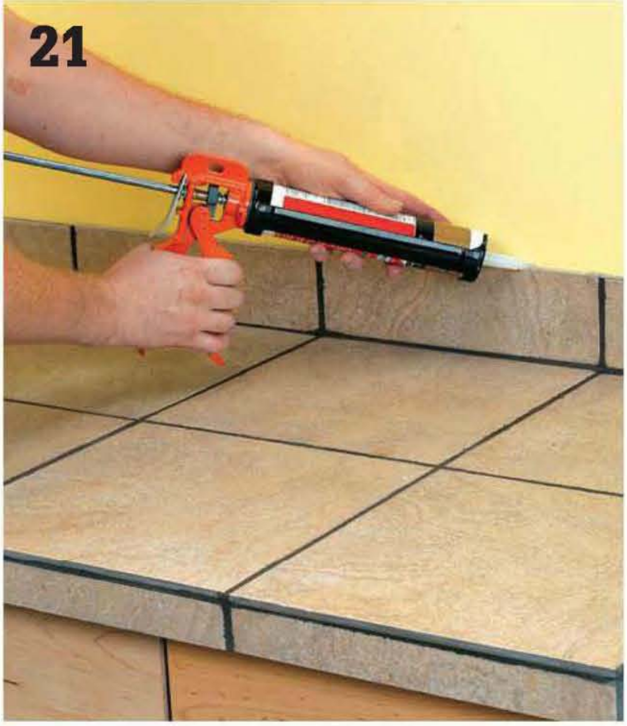

After the grout has dried run a bead of clear silicone caulk along the joint between the backsplash and the wall. Install your sink and faucet.

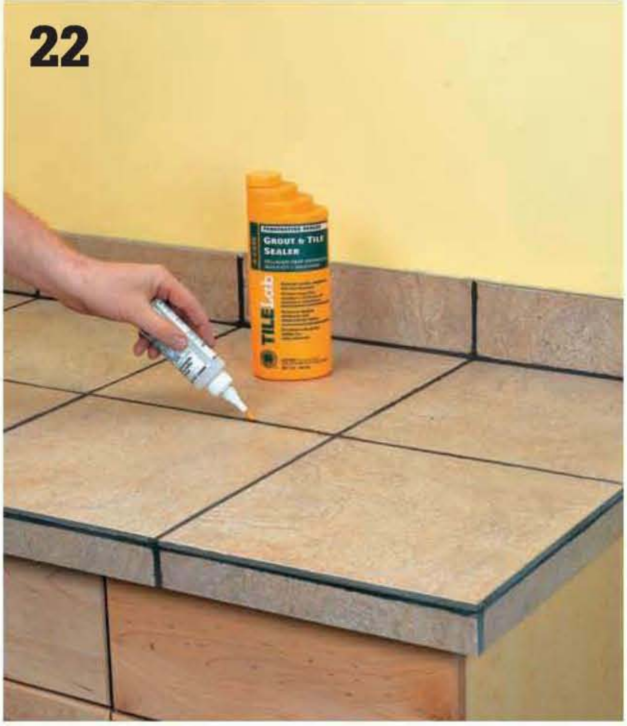

Wait at least one week and then seal the grout lines with a penetrating grout sealer. This is important to do. Sealing the tiles themselves is not a good idea unless you are using unglazed tiles (a poor choice for countertops, however).

Installing Tile Backsplashes

There are few spaces in your home with as much potential for creativity and visual impact as the space between your kitchen countertop and cupboards. A well-designed backsplash can transform the ordinary into the extraordinary.

Tiles for the backsplash can be attached directly to wallboard or plaster and do not require backerboard. When purchasing the tile, order 10 percent extra to cover breakage and cutting. Remove switch and receptacle coverplates and install box extenders to make up for the extra thickness of the tile. Protect the countertop from scratches by covering it with a drop cloth.

A tiled backsplash normally extends all the way from countertop to the bottoms of the wall cabinets. The tile pattern also can be extended to the wall underneath your range hood.

Tools & Materials ▸

- Level
- Tape measure
- Pencil
- Tile cutter
- Rod saw
- Notched trowel
- Rubber grout float
- Beating block
- Rubber mallet
- Sponge
- Bucket
- Straight 1 × 2
- Wall tile
- Tile spacers (if needed)
- Bullnose trim tile
- Mastic tile adhesive
- Masking tape
- Grout
- Caulk
- Drop cloth
- Grout sealer

Tips for Planning Tile Layouts ▸

Gather planning brochures and design catalogs to help you create decorative patterns and borders for the backsplash.

Break tiles into fragments and make a mosaic backsplash. Always use a sanded grout for joints wider than ⅛".

Add painted mural tiles to create a focal point. Mixing various tile styles adds an appealing contrast.

How to Tile a Backsplash

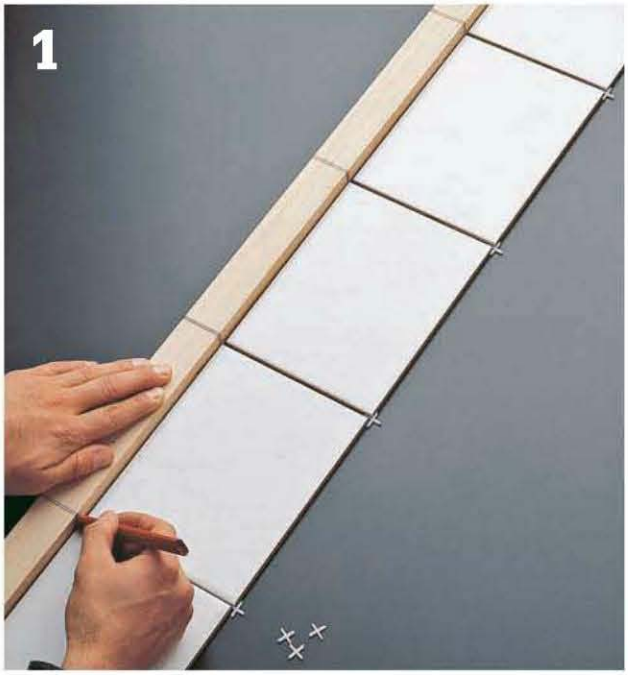

Make a story stick by marking a board at least half as long as the backsplash area to match the tile spacing.

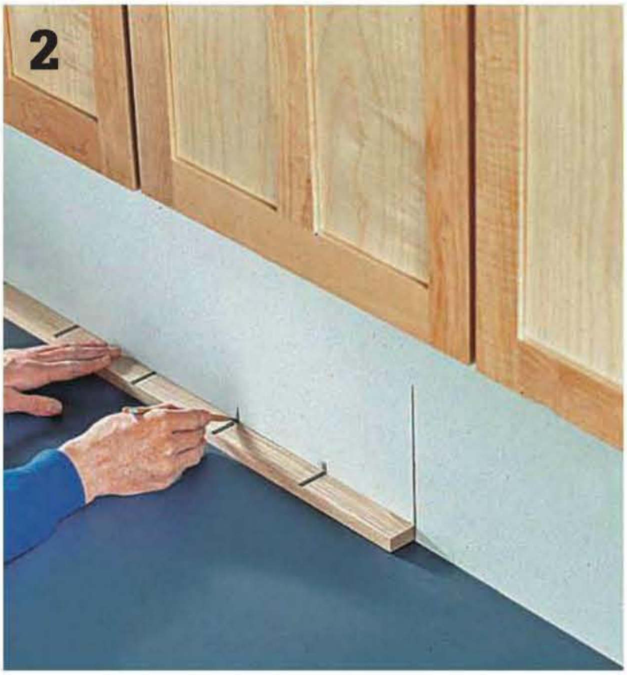

Starting at the midpoint of the installation area, use the story stick to make layout marks along the wall. If an end piece is too small (less than half a tile), adjust the midpoint to give you larger, more attractive end pieces. Use a level to mark this point with a vertical reference line.

While it may appear straight, your countertop may not be level and therefore is not a reliable reference line. Run a level along the counter to find the lowest point on the countertop. Mark a point two tiles up from the low point and extend a level line across the entire work area.

(continued)

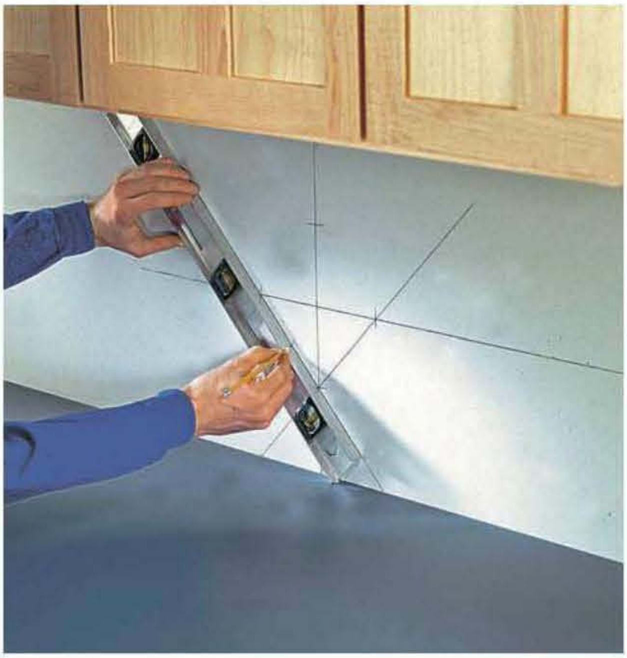

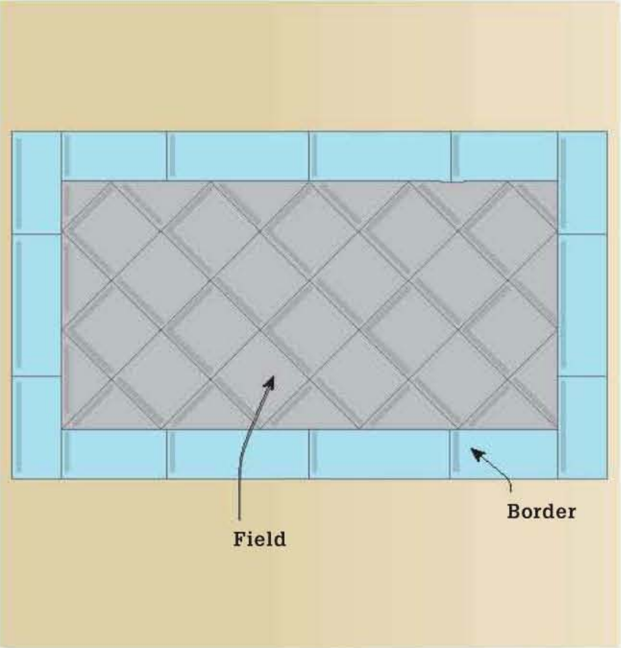

Variation: Diagonal Layout. Mark vertical and horizontal reference lines, making sure the angle is 90°. To establish diagonal layout lines, measure out equal distances from the crosspoint, then connect the points with a line. Additional layout lines can be extended from these as needed. To avoid the numerous, unattractive perimeter cuts common to diagonal layouts, try using a standard border pattern as shown. Diagonally set a field of full tiles, then cut enough half tiles to fill out the perimeter. Finally, border the diagonal field with tiles set square to the field.

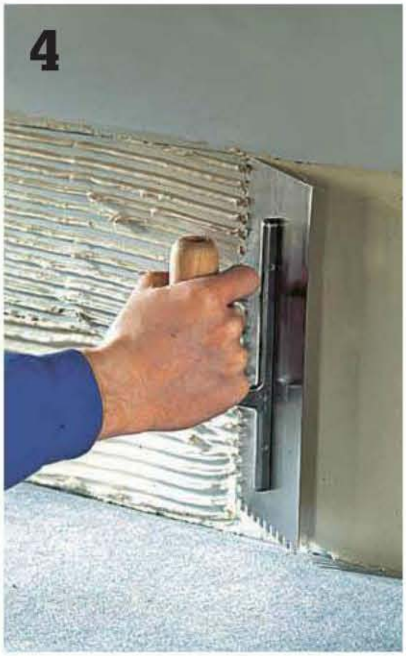

Apply mastic adhesive evenly to the area beneath the horizontal reference line, using a notched trowel. Comb the adhesive horizontally with the notched edge.

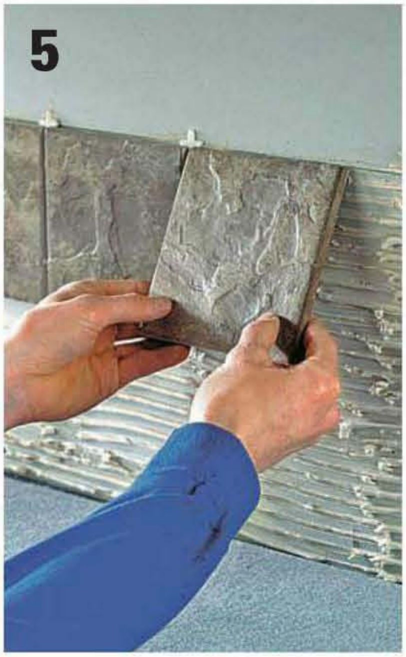

Starting at the vertical reference line, press tiles into the adhesive using a slight twisting motion. If the tiles are not self-spacing, use plastic spacers to maintain even grout lines. If the tiles do not hang in place, use masking tape to hold them in place until the adhesive sets.

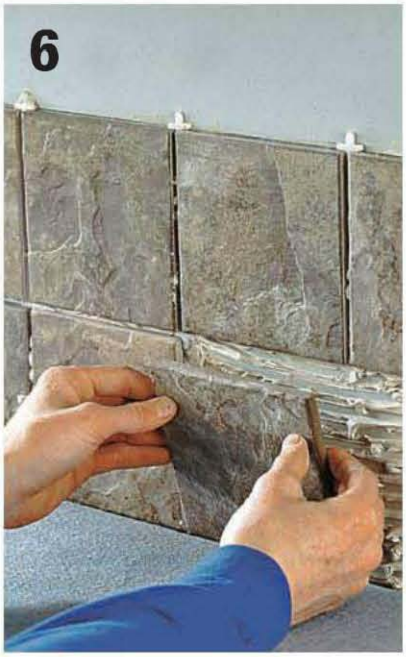

Install a whole row along the reference line, checking occasionally to make sure the tiles are level. Continue installing tiles below the first row, trimming tiles that butt against the countertop as needed.

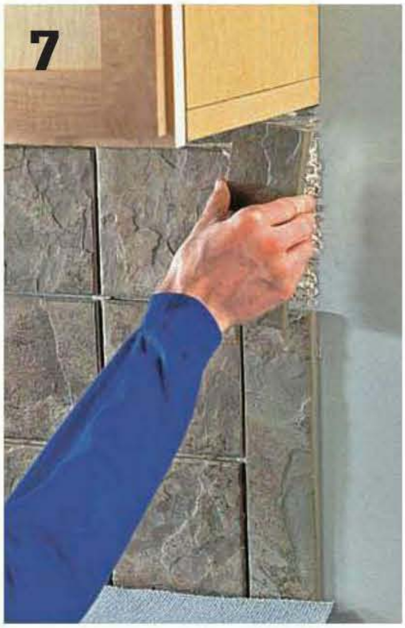

Apply adhesive to an area above the line and continue placing tiles, working from the center to the sides. Install trim tile, such as bullnose tile, to the edges of the rows.

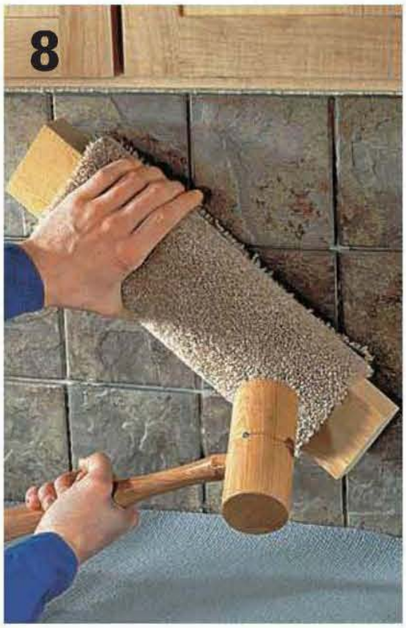

When the tiles are in place, make sure they are flat and firmly embedded by laying a beating block against the tile and rapping it lightly with a mallet. Remove the spacers. Allow the mastic to dry for at least 24 hours, or as directed by the manufacturer.

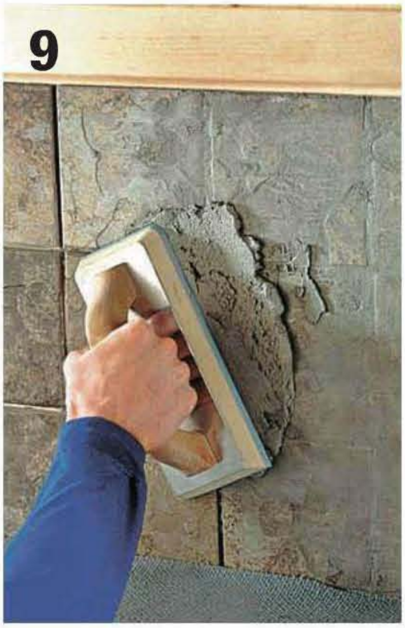

Mix the grout and apply it with a rubber grout float. Spread it over the tiles, keeping the float at a low 30° angle, pressing the grout deep into the joints. *Note: For grout joints ⅛" and smaller, be sure to use a non-sanded grout.*

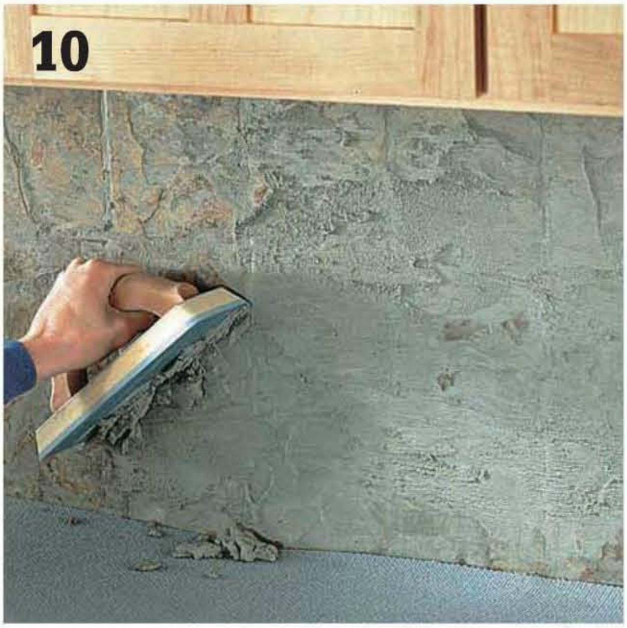

Wipe off excess grout, holding the float at a right angle to the tile, working diagonally so as not to remove grout from the joints. Clean any remaining grout from the tiles with a damp sponge, working in a circular motion. Rinse the sponge thoroughly and often.

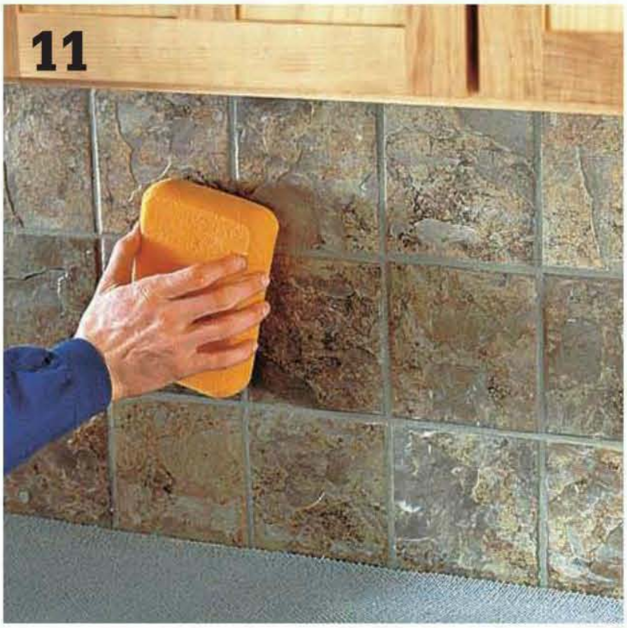

Shape the grout joints by making slow, short passes with the sponge, shaving down any high spots; rinse the sponge frequently. Fill any voids with a fingerful of grout. When the grout has dried to a haze, buff the tile clean with a soft cloth. Apply a bead of caulk between the countertop and tiles. Reinstall any electrical fixtures you removed. After the grout has completely cured, apply grout sealer.

Installing Kitchen Sinks

Most drop-in, self-rimming kitchen sinks are easily installed.

Drop-in sinks for do-it-yourself installation are made from cast iron coated with enamel, stainless steel, enameled steel, acrylic, fiberglass or resin composites. Because cast-iron sinks are heavy, their weight holds them in place and they require no mounting hardware. Except for the heavy lifting, they are easy to install. Stainless steel and enameled-steel sinks weigh less than cast-iron and most require mounting brackets on the underside of the countertop. Some acrylic and resin sinks rely on silicone caulk to hold them in place.

If you are replacing a sink but not the countertop, make sure the new sink is the same size or larger. All old silicone caulk residue must be removed with acetone or denatured alcohol, or the new caulk will not stick.

Shopping Tips ▸

- When purchasing a sink you also need to buy strainer bodies and baskets, sink clips and a drain trap kit.
- Look for basin dividers that are lower than the sink rim—this reduces splashing.
- Drain holes in the back or to the side leave more usable space under the sink.
- When choosing a sink, make sure the predrilled openings will fit your faucet.

Tools & Materials ▸

Caulk gun
Spud wrench
Screwdriver
Sink
Sink frame
Plumber's putty or silicone caulk
Mounting clips
Jigsaw
Pen or pencil

Drop-in sinks, also known as self-rimming sinks, have a wide sink flange that extends beyond the edges of the sink cutout. They also have a wide back flange to which the faucet is mounted directly.

How to Install a Self-rimming Sink

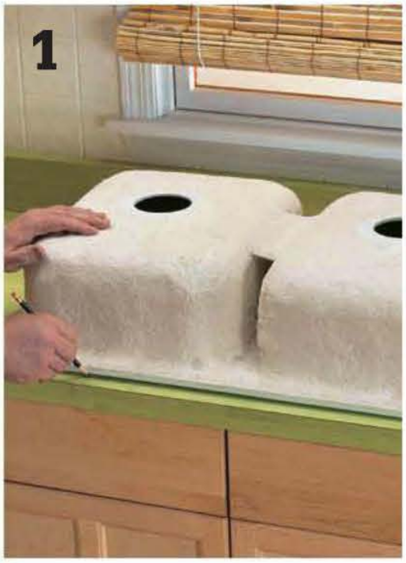

Invert the sink and trace around the edges as a reference for making the sink cutout cutting lines, which should be parallel to the outlines, but about 1" inside of them to create a 1" ledge. If your sink comes with a template for the cutout, use it.

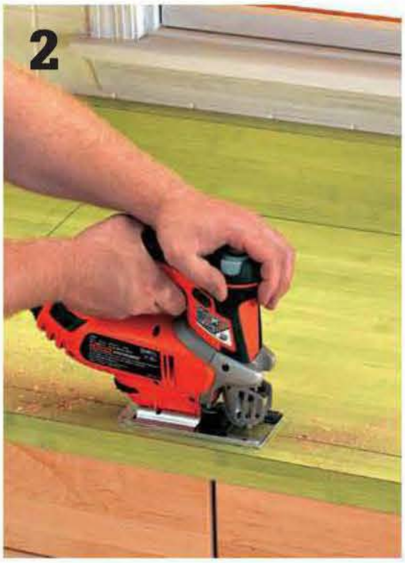

Drill a starter hole and cut out the sink opening with a jigsaw. Cut right up to the line. Because the sink flange fits over the edges of the cutout, the opening doesn't need to be perfect, but as always you should try to do a nice, neat job.

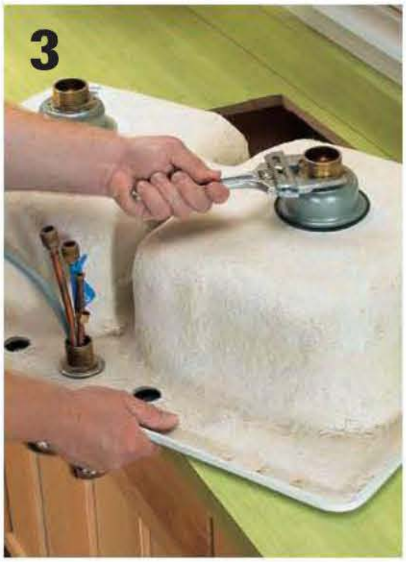

Attach as much of the plumbing as makes sense to install prior to setting the sink into the opening. Having access to the underside of the flange is a great help when it comes to attaching the faucet body, sprayer and strainer, in particular.

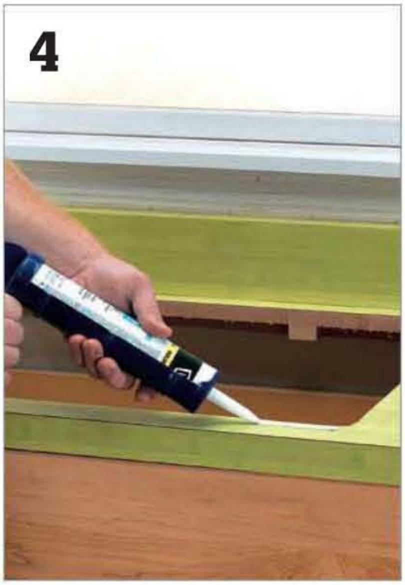

Apply a bead of silicone caulk around the edges of the sink opening. The sink flange most likely is not flat, so try and apply the caulk in the area that will make contact with the flange.

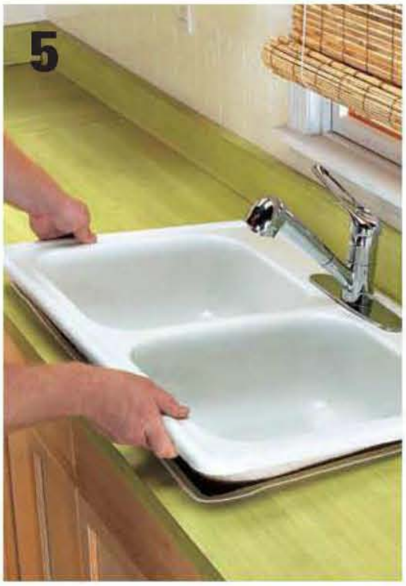

Place the sink in the opening. Try and get the sink centered right away so you don't need to move it around and disturb the caulk, which can break the seal. If you are installing a heavy cast-iron sink, it's best to leave the strainers off so you can hold the sink at the drain openings.

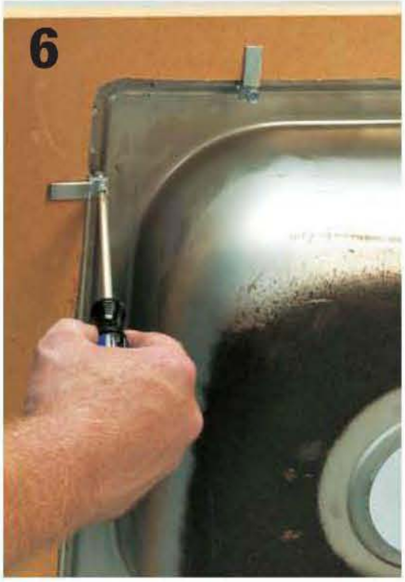

For sinks with mounting clips, tighten the clips from below using a screwdriver or wrench (depending on the type of clip your sink has). There should be at least three clips on every side. Don't overtighten the clips—this can cause the sink flange to flatten or become warped.

Replacing Kitchen Faucets

Tools & Materials ▸

- Adjustable wrench
- Basin wrench or channel-type pliers
- Hacksaw
- Faucet
- Putty knife
- Screwdriver
- Silicone caulk
- Scouring pad
- Scouring cleaner
- Plumber's putty
- Flexible vinyl or braided steel supply tubes
- Drain components
- Penetrating oil

Most new kitchen faucets feature single-handle control levers and washerless designs that rarely require maintenance. Additional features include brushed metallic finishes, detachable spray nozzles, or even push-button controls.

Connect the faucet to hot and cold water lines with easy-to-install flexible supply tubes made from vinyl or braided steel. If your faucet has a separate sprayer, install the sprayer first. Pull the sprayer hose through the sink opening and attach to the faucet body before installing the faucet.

Where local codes allow, use plastic tubes for drain hookups. A wide selection of extensions and angle fittings lets you easily plumb any sink configuration. Manufacturers offer kits that contain all the fittings needed for attaching a food disposer or dishwasher to the sink drain system.

Modern kitchen faucets tend to be single-handle models, often with useful features such as a pull-out head that functions as a sprayer. This Price Pfister™ model comes with an optional mounting plate that conceals sink holes when mounted on a predrilled sink flange.

Choosing a New Kitchen Faucet

When choosing a new kitchen faucet, you'll find many options. The best place to start the process is with your sink. In the past, most faucets were mounted directly to the sink deck, which had three or four predrilled holes to accommodate the faucets, spout, sprayer and perhaps a liquid soap dispenser or an air gap for your dishwasher. Modern kitchen faucets don't always conform to this set-up, with many of them designed to be installed in a single hole in the sink deck or in the countertop. If you plan to keep your old sink, look for a faucet that won't leave empty holes in the deck. You can buy snap-in plugs to fill unfilled stainless steel sink holes, or you can add a soap dispenser, but in general you'll have the best luck if you replace like for like.

The two most basic kitchen faucet categories are one-handle and two-handle. One handled models are much more popular now because you can adjust the water temperature easily with just one hand. Another difference is in the faucet body. Some faucets have the taps and the spout mounted onto a faucet body so the spacing between the tailpieces is preset. Others, called widespread faucets, have independent taps and spouts that can be configured however you please, as long as the tubes connecting the taps to the spout reach. This type is best if you are installing the faucet in the countertop (a common way to go about it with new countertops such as solid surface, quartz or granite).

In the past, kitchen faucets almost always had remote pull-out sprayers. The sprayer was attached to the faucet body with a hose directly below the mixing valve. While this type of sprayer is still fairly common, many faucets today have an integral pull-out spout that is very convenient and less prone to failure than the old-style sprayers.

A single-handle, high arc faucet with traditional remote sprayer. The mounting plate is decorative and optional.

Single-handle faucets may require four holes, as does this model with its side sprayer and matching soap/lotion dispenser.

Two-handled faucets are less common, but remain popular choices for traditional kitchens. The gooseneck spout also has a certain elegance, but avoid this type if you have a shallow sink that's less than 8" deep.

A single-handle faucet with pull-out spray head requires only one hole in your sink deck or countertop—a real benefit if your sink is not predrilled or if it is an undermount model.

How to Remove an Old Faucet

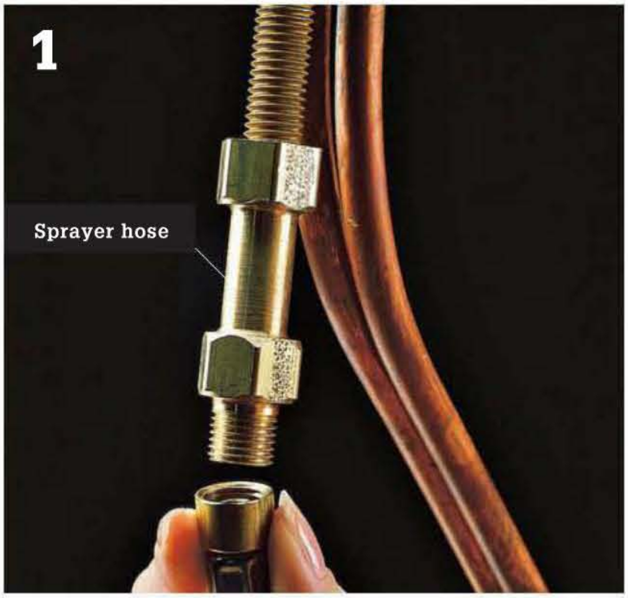

To remove the old faucet, start by clearing out the cabinet under the sink and laying down towels. Turn off the hot and cold stop valves and open the faucet to make sure the water is off. Detach the sprayer hose from the faucet sprayer nipple and unscrew the retaining nut that secures the sprayer base to the sink deck. Pull the sprayer hose out through the sink deck opening.

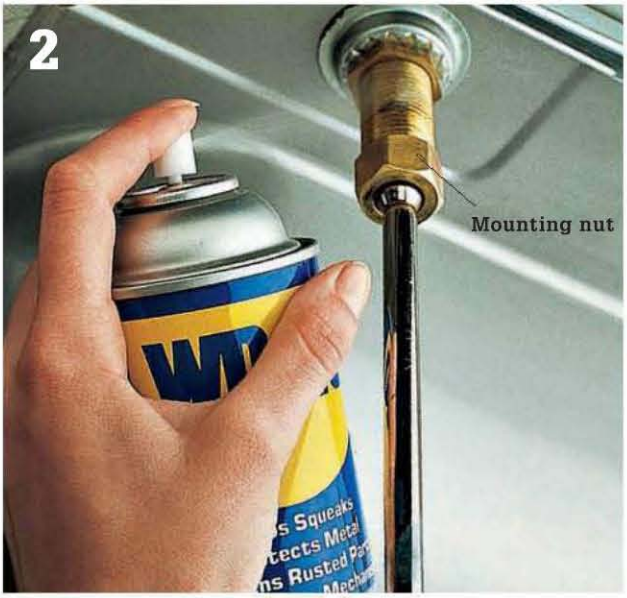

Spray the mounting nuts that hold the faucet or faucet handles (on the underside of the sink deck) with penetrating oil for easier removal. Let the oil soak in for a few minutes.

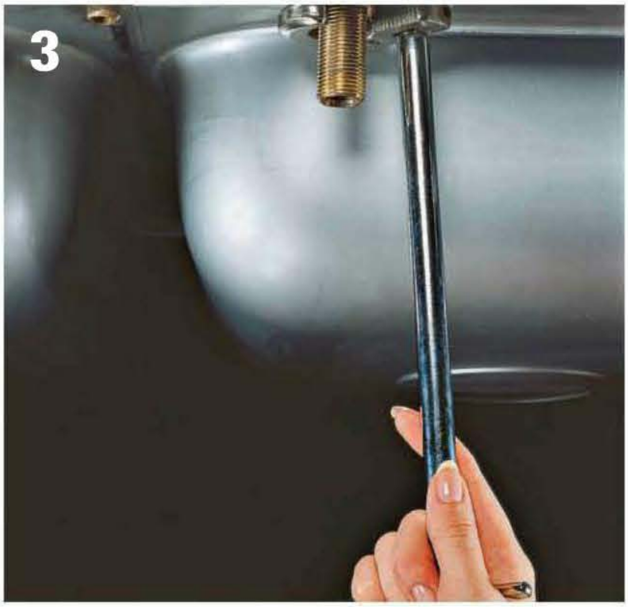

Unhook the supply tubes at the stop valves. Don't reuse old chrome supply tubes. If the stops are missing or unworkable, replace them. Then remove the coupling nuts and the mounting nuts on the tailpieces of the faucet with a basin wrench or channel-type pliers.

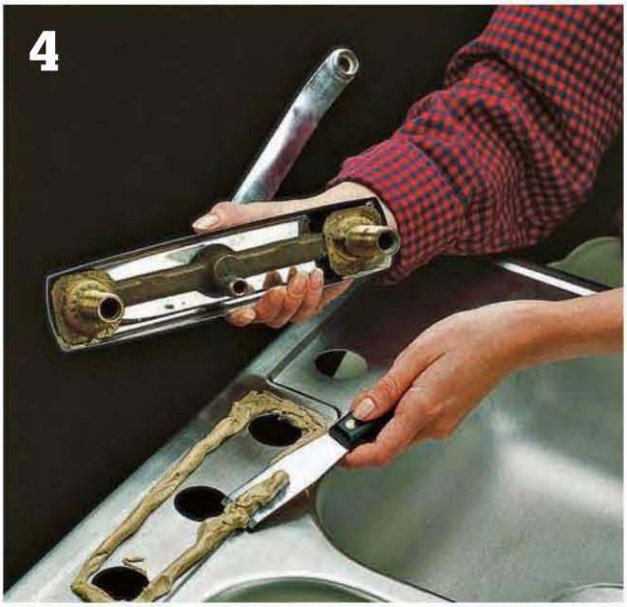

Pull the faucet body from the sink. Remove the sprayer base if you wish to replace it. Scrape off old putty or caulk with a putty knife and clean off the sink with a scouring pad and an acidic scouring cleaner like Bar Keeper's Friend®. *Tip: Scour stainless steel with a back and forth motion to avoid leaving unsightly circular markings.*

How to Install a Kitchen Sink Faucet

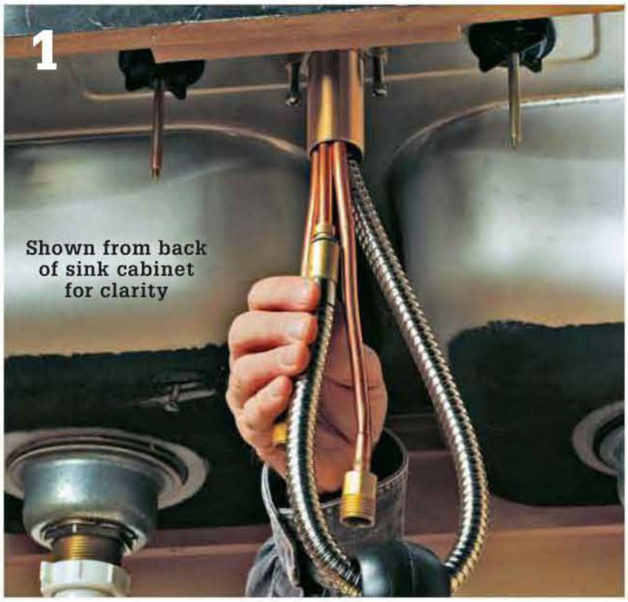

Shut off hot and cold water at the faucet stop valves. Assemble the parts of the deck plate that cover the outer mounting holes in your sink deck (unless you are installing a two-handle faucet, or mounting the faucet directly to the countertop, as in an undermount sink situation). Add a ring of plumber's putty in the groove on the underside of the base plate.

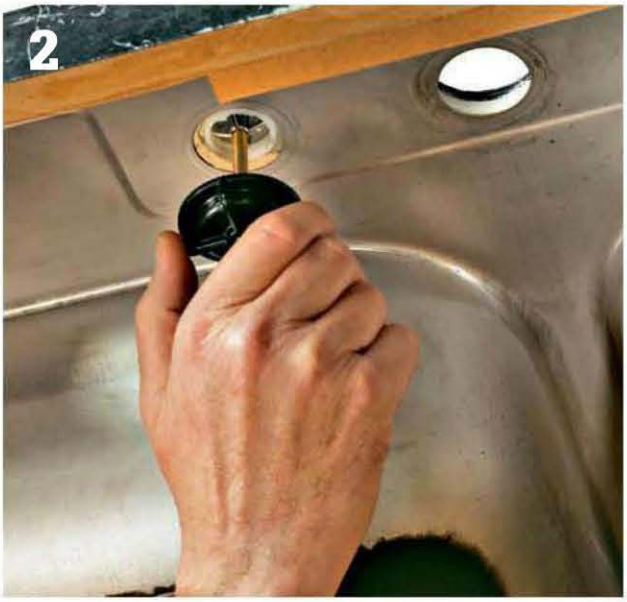

Set the base plate onto the sink flange so it is correctly aligned with the predrilled holes in the flange. From below, tighten the wing nuts that secure the deck plate to the sink deck.

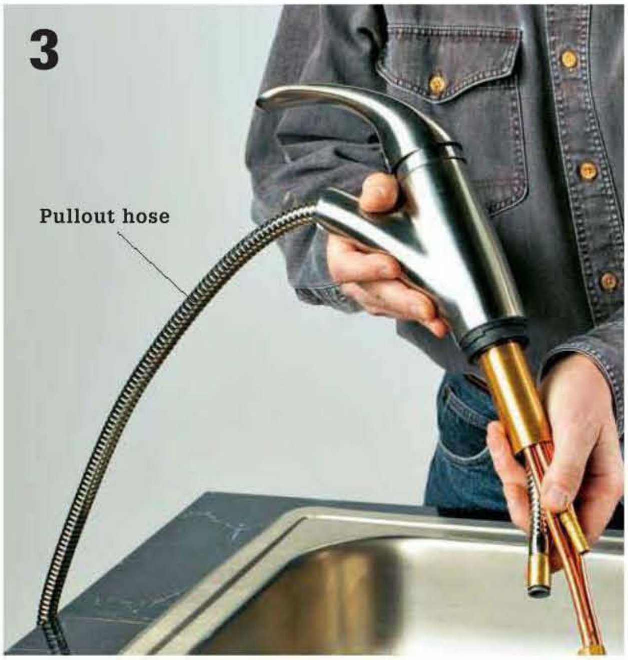

Retract the pullout hose by drawing it out through the faucet body until the fitting at the end of the hose is flush with the bottom of the threaded faucet shank. Insert the shank and the supply tubes down through the top of the deck plate.

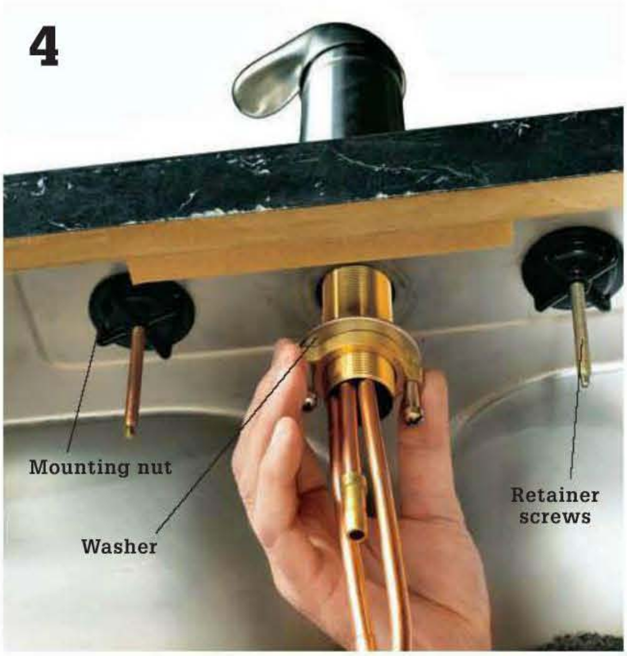

Slip the mounting nut and washer over the free ends of the supply tubes and pullout hose, then thread the nut onto the threaded faucet shank. Hand tighten. Tighten the retainer screws with a screwdriver to secure the faucet.

(continued)

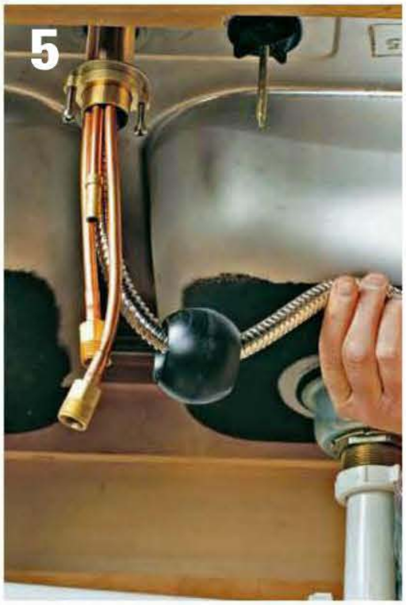

Slide the hose weight onto the pullout hose (the weight helps keep the hose from tangling and makes it easier to retract the hose).

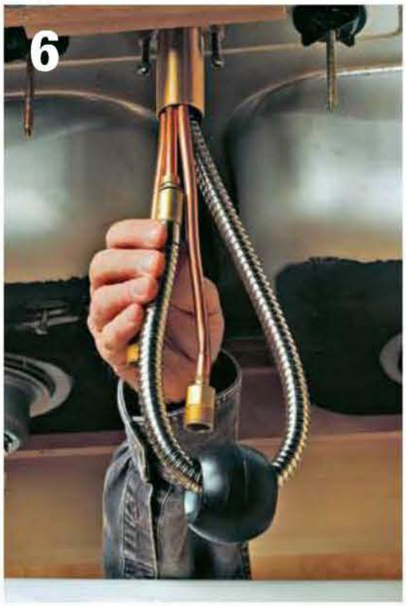

Connect the end of the pullout tube to the outlet port on the faucet body using a quick connector fitting.

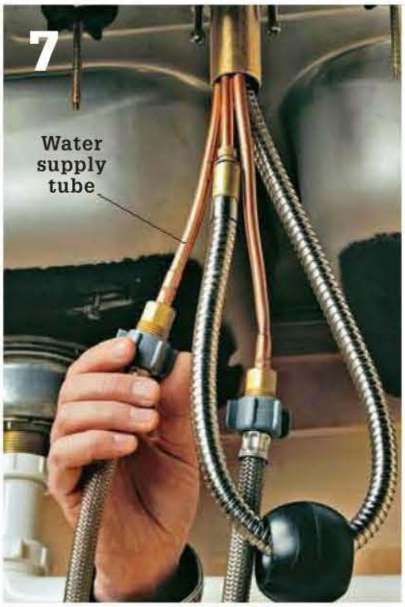

Hook up the water supply tubes to the faucet inlets. Make sure the lines are long enough to reach the supply risers without stretching or kinking.

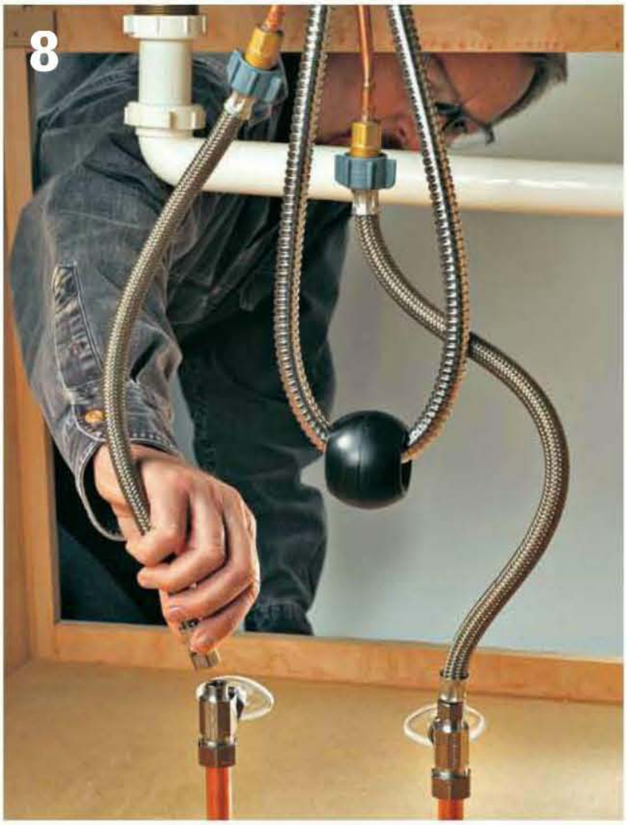

Connect the supply lines to the supply risers at the stop valves. Make sure to get the hot lines and cold lines attached correctly.

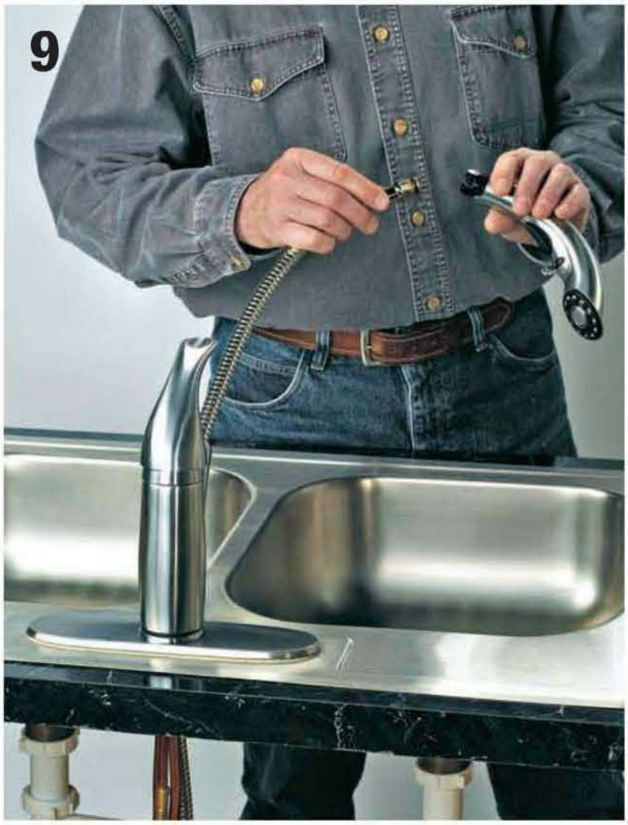

Attach the spray head to the end of the pullout hose and turn the fitting to secure the connection. Turn on water supply and test. *Tip: Remove the aerator in the tip of the spray head and run hot and cold water to flush out any debris.*

Variation: One-piece Faucet with Sprayer ▸

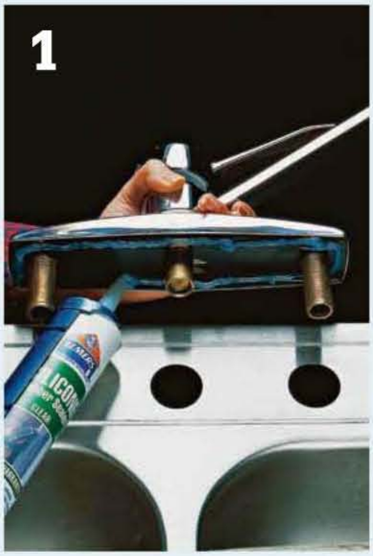

Apply a thick bead of silicone caulk to the underside of the faucet base, then insert the tailpieces of the faucet through the appropriate holes in the sink deck. Press down lightly on the faucet to set it in the caulk.

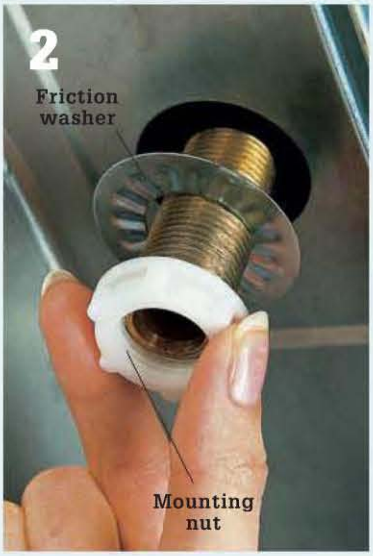

Slip a friction washer onto each tailpiece and then hand-tighten a mounting nut. Tighten the mounting nut with channel-type pliers or a basin wrench. Wipe up any silicone squeeze-out on the sink deck with a wet rag before it sets up.

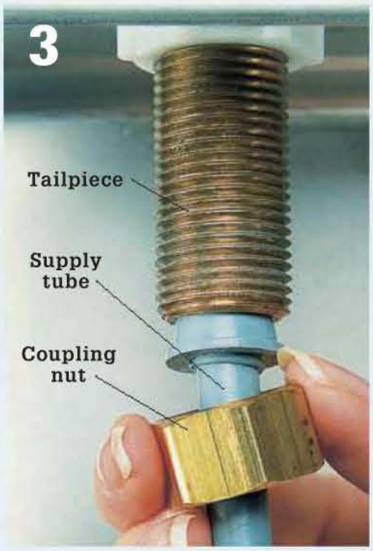

Connect supply tubes to the faucet tailpieces—make sure the tubes you buy are long enough to reach the stop valves and that the coupling nuts will fit the tubes and tailpieces.

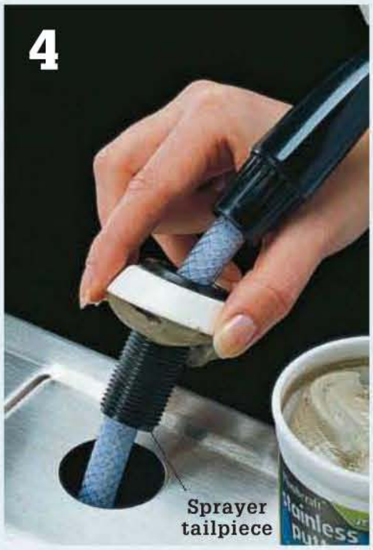

Apply a ¼" bead of plumber's putty or silicone caulk to the underside of the sprayer base. With the base threaded onto the sprayer hose, insert the tailpiece of the sprayer through the opening in the sink deck.

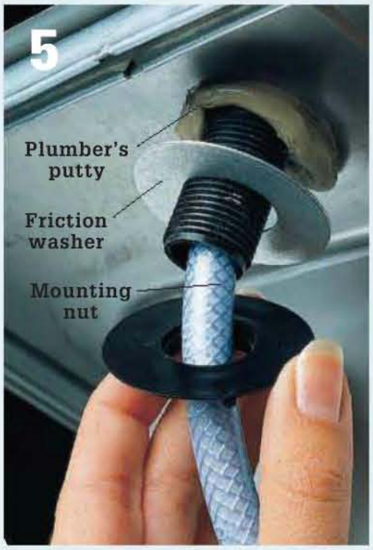

From beneath, slip the friction washer over the sprayer tailpiece and then screw the mounting nut onto the tailpiece. Tighten with channel-type pliers or a basin wrench. Wipe any excess putty or caulk on the sink deck from around the base.

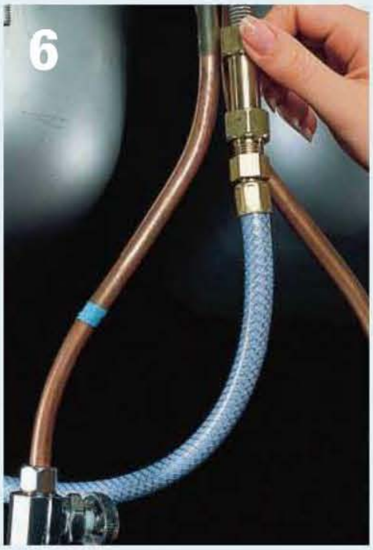

Screw the sprayer hose onto the hose nipple on the bottom of the faucet. Hand-tighten and then give the nut one quarter turn with pliers or a basin wrench. Turn on the water supply at the shutoff, remove the aerator and flush debris from the faucet.

Connecting Drains

Kitchen sink drains don't last forever, but they're very easy and inexpensive to replace. The most common models today are made of PVC plastic pipe and fittings held together with slip fittings. In addition to making the installation fairly forgiving, the slip fitting makes the drain easy to disassemble if it gets clogged. The project shown here is a bit unusual by today's standards, in that it does not include either a dishwasher drain or a garbage disposer. But you will see how to add each of these drain systems to your kitchen sink in the following two sections.

You can buy the parts for the kitchen drain individually (you can usually get better quality materials this way) or in a kit (see photo, next page). Because most kitchen sinks have two bowls, the kits include parts for plumbing both drains into a shared trap, often with a baffle in the T-fitting where the outlet line joins with the tailpiece from the other bowl. If you are installing a disposer, consider installing individual traps to eliminate the baffle, which reduces the flow capacity by half.

Tools & Materials ▸

Flat screwdriver
Spud wrench
Trap arm
Mineral spirits
Cloth
Strainer kit
Plumber's putty
Teflon tape
Washers
Waste-T fitting
S- or P-trap

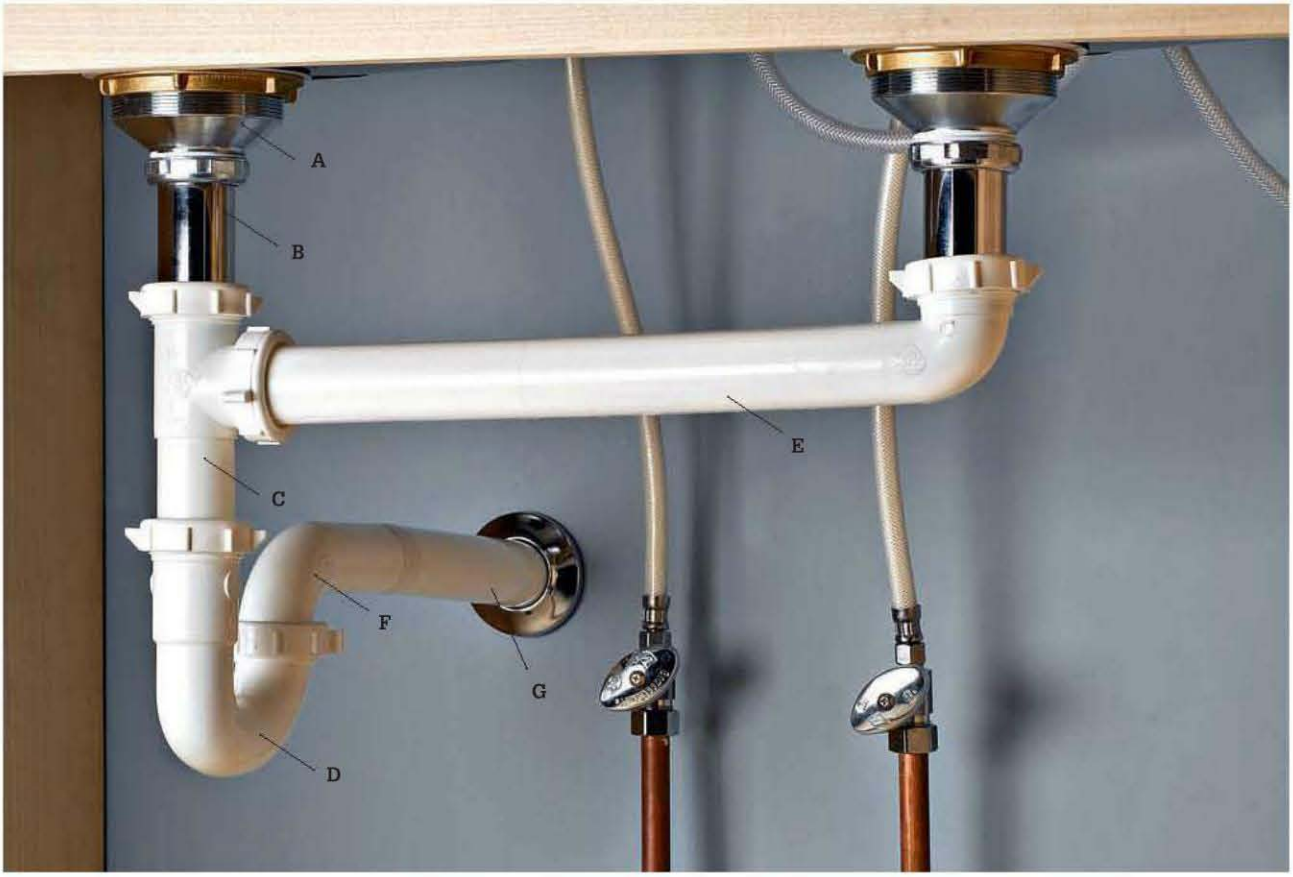

Kitchen sink drains include a strainer basket (A), tailpiece (B), continuous waste T (C), P- or S-trap (D), outlet drain lines (E), trap arm (F), and wall stubout (G).

Drain Kits ▸

Kits for installing a new sink drain include all the pipes, slip fittings, and washers you'll need to get from the sink tailpieces (most kits are equipped for a double bowl kitchen sink) to the trap arm that enters the wall or floor. For wall trap arms, you'll need a kit with a P-trap. For floor drains, you'll need an S-trap. Both drains normally are plumbed to share a trap. Chromed brass or PVC with slip fittings let you adjust the drain more easily and pull it apart and then reassemble if there is a clog. Kitchen sink drains and traps should be 1½" o.d. pipe—the 1¼" pipe is for lavatories and doesn't have enough capacity for a kitchen sink.

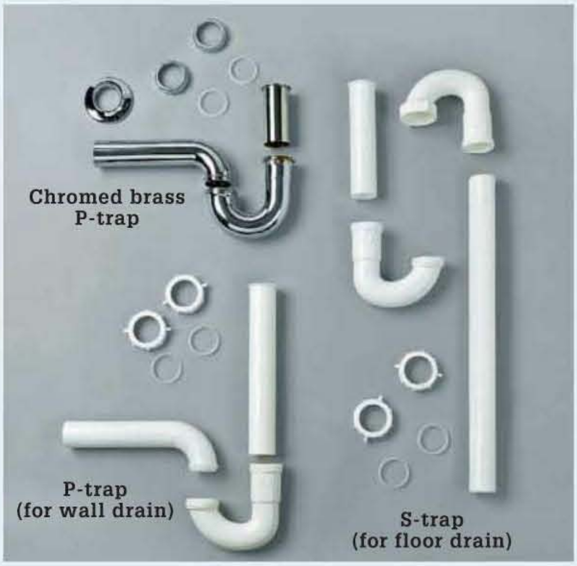

Tips for Choosing Drains ▸

Wall thickness varies in sink drain pipes. The thinner plastic material is cheaper and more difficult to obtain a good seal with than the thicker, more expensive tubing. The thin product is best reserved for lavatory drains, which are far less demanding.

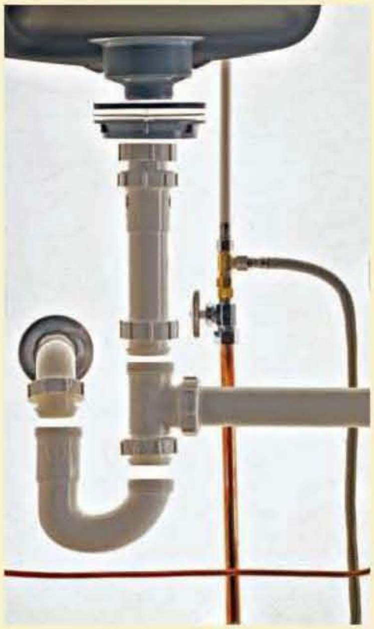

Slip joints are formed by tightening a male-threaded slip nut over a female-threaded fitting, trapping and compressing a beveled nylon washer to seal the joint.

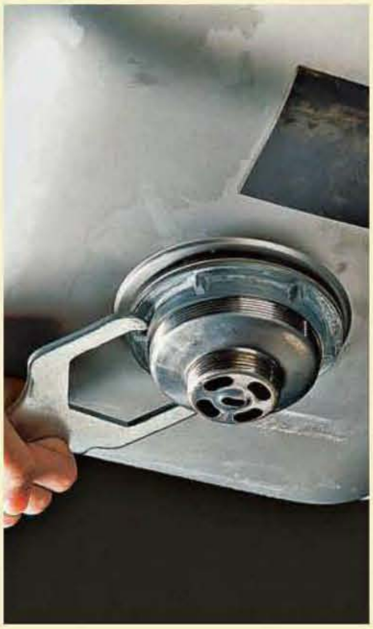

Use a spud wrench to tighten the strainer body against the underside of the sink bowl. Normally, the strainer flange has a layer of plumber's putty between it and the sink drain, and a pair of washers (one rubber, one fibrous) below.

How to Hook Up a Kitchen Sink Drain

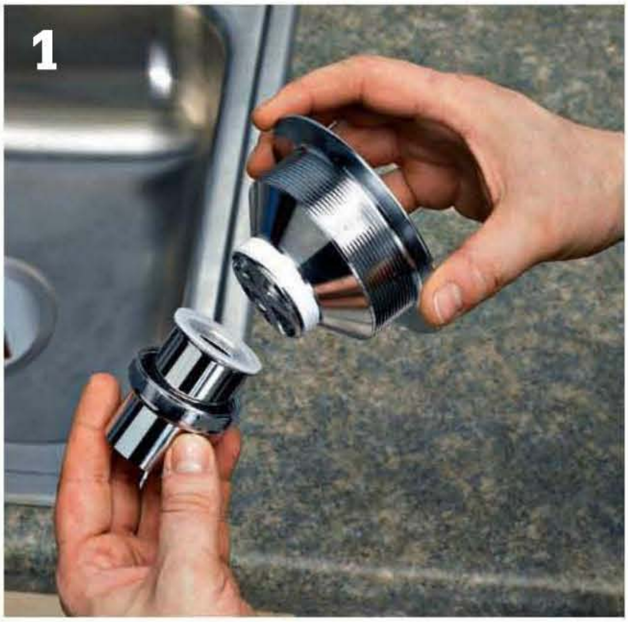

If you are replacing the sink strainer body, remove the old one and clean the top and bottom of the sink deck around the drain opening with mineral spirits. Attach the drain tailpiece to the threaded outlet of the strainer body, inserting a nonbeveled washer between the parts, if your strainer kits include one. Lubricate the threads or apply Teflon tape so you can get a good, snug fit.

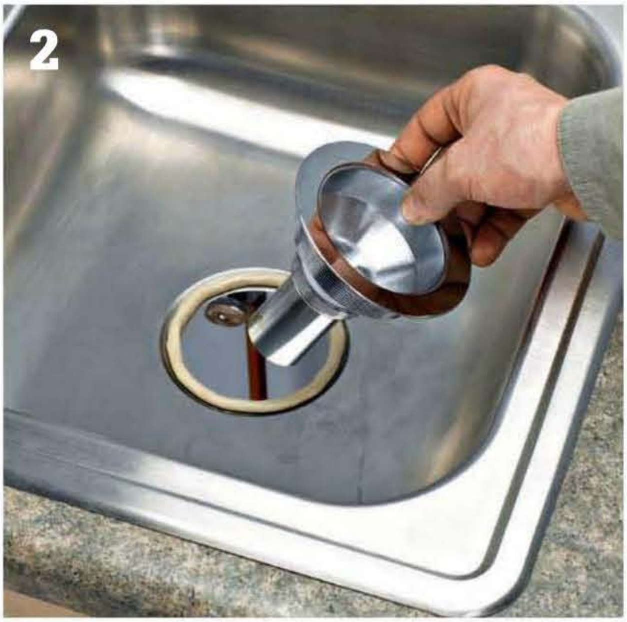

Apply plumber's putty around the perimeter of the drain opening and seat the strainer assembly into it. Add washers below as directed and tighten the strainer locknut using a spud wrench (see photo, previous page) or striking the mounting nubs at the top of the body with a flat screwdriver.

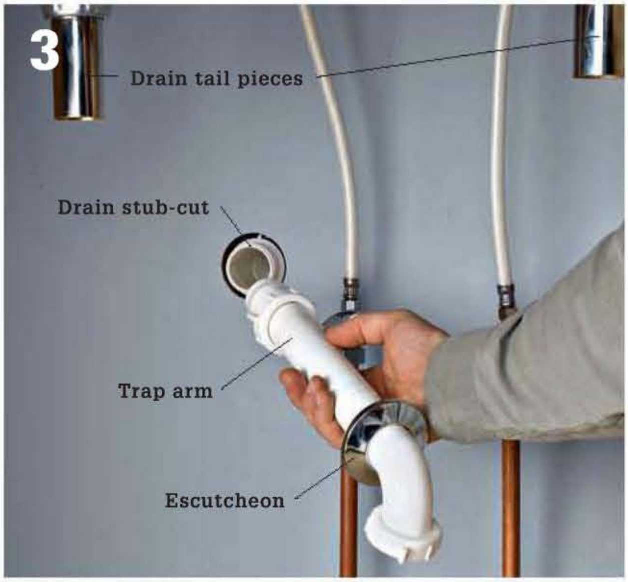

Attach the trap arm to the male-threaded drain stubout in the wall, using a slip nut and beveled compression washer. The outlet for the trap arm should point downward. *Note: The trap arm must be higher on the wall than any of the horizontal lines in the set-up, including lines to dishwasher, disposer, or the outlet line to the second sink bowl.*

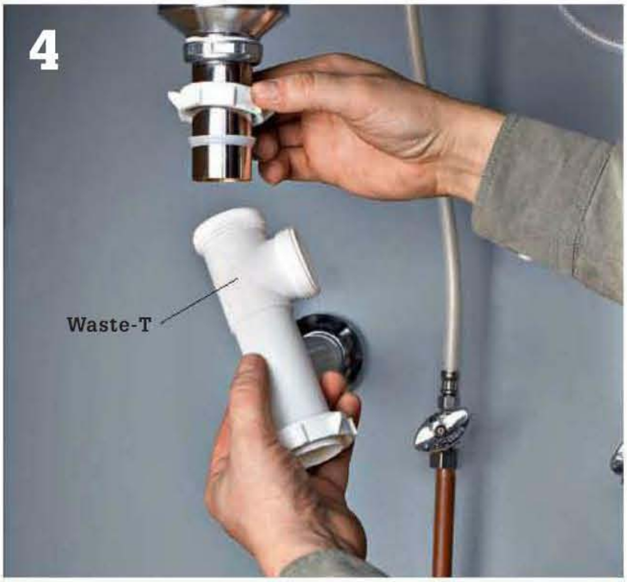

Attach a waste-T-fitting to the drain tailpiece, orienting the opening in the fitting side so it will accept the outlet drain line from the other sink bowl. If the waste-T is higher than the top of the trap arm, remove it and trim the drain tailpiece.

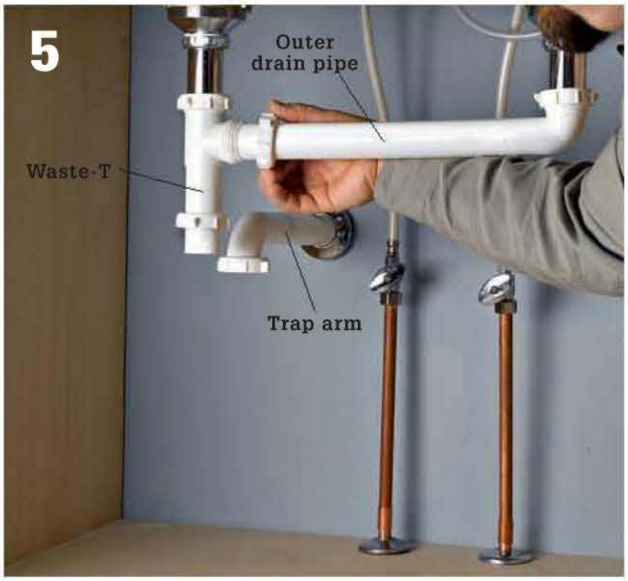
5
Outer drain pipe
Waste-T
Trap arm

Joint the short end of the outlet drain pipe to the tailpiece for the other sink bowl and then attach the end of the long run to the opening in the waste-T. The outlet tube should extend into the T ½" or so—make sure it does not extend far enough to block water flow from above.

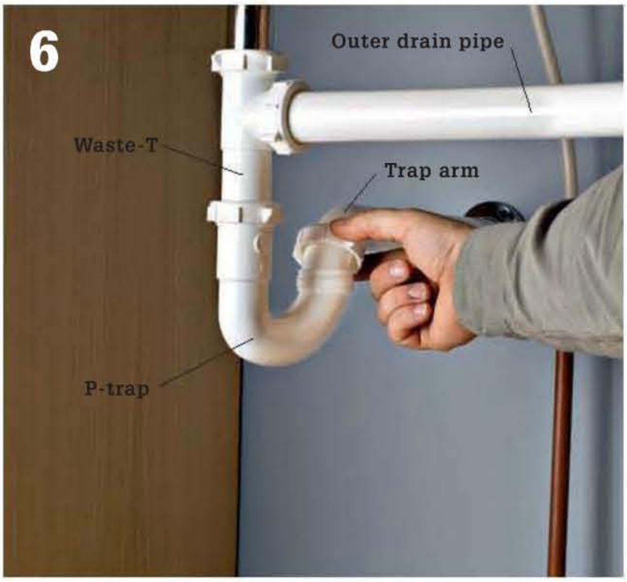
6
Outer drain pipe
Waste-T
Trap arm
P-trap

Attach the long leg of a P-trap to the waste-T and attach the shorter leg to the downward-facing opening of the trap arm. Adjust as necessary and test all joints to make sure they are still tight, and then test the system.

Variation: Drain in Floor ▸

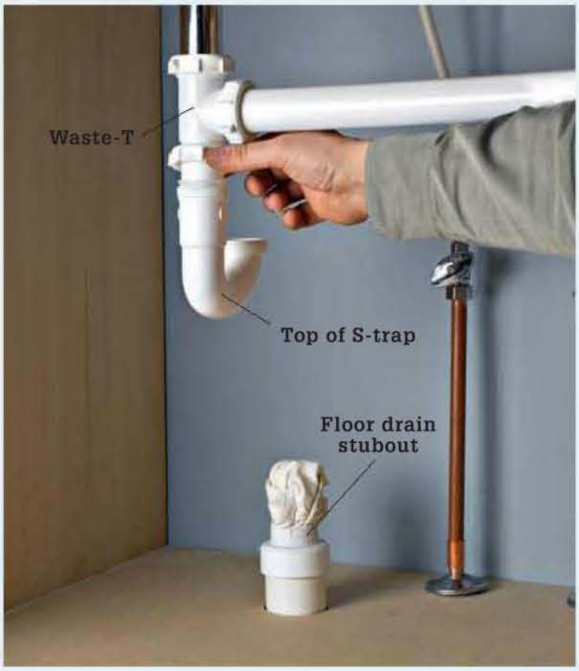
Waste-T
Top of S-trap
Floor drain stubout

If your drain stubout comes up out of the floor instead of the wall, you'll need an S-trap to tie into it instead of a P-trap. Attach one half of the S-trap to the threaded bottom of the waste-T.

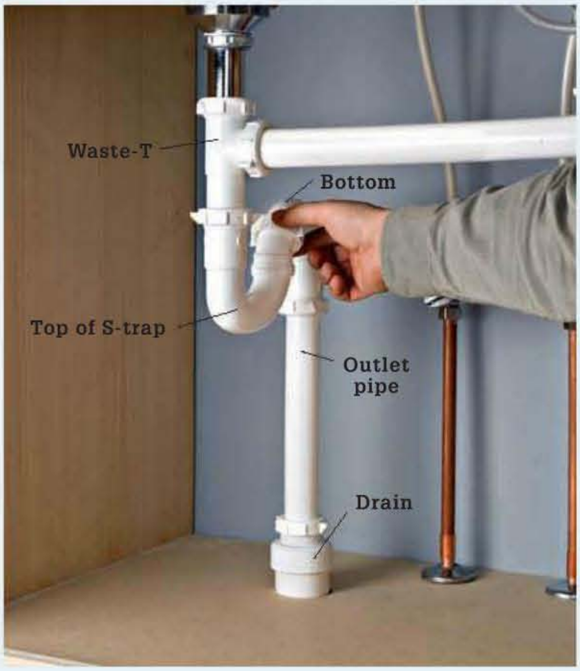
Waste-T
Bottom
Top of S-trap
Outlet pipe
Drain

Attach the other half of the S-trap to the stubout using a slip fitting. This should result in the new fitting facing downward. Join the halves of the S-trap with a slip nut, trimming the unthreaded end if necessary.

Hooking Up Dishwashers

A dishwasher that's past its prime may be inefficient in more ways than one. If it's an old model, it probably wasn't designed to be very efficient to begin with. But more significantly, if it no longer cleans effectively, you're probably spending a lot of time and hot water pre-rinsing the dishes. This alone can consume more energy and water than a complete wash cycle on a newer machine. Even if your old dishwasher still runs, replacing it with an efficient new model can be a good green upgrade.

In terms of sizing and utility hookups, dishwashers are generally quite standard. If your old machine is a built-in and your countertops and cabinets are standard sizes, most full-size dishwashers will fit right in. Of course, you should always measure the dimensions of the old unit before shopping for a new one to avoid an unpleasant surprise at installation time. Also, be sure to review the manufacturer's instructions before starting any work.

Tools & Materials ▸

- Screwdrivers
- Adjustable wrench
- 2-ft. level
- ⅝" automotive heater hose
- Automotive heater hose
- 4"-length of ½" copper tubing
- Cable connector
- Teflon tape
- Hose clamps
- Wire connectors
- Carpet scrap
- Bowl

Replacing an old, inefficient dishwasher is a straightforward project that usually takes just a few hours. The energy savings begin with the first load of dishes and continue with every load thereafter.

How to Replace a Dishwasher

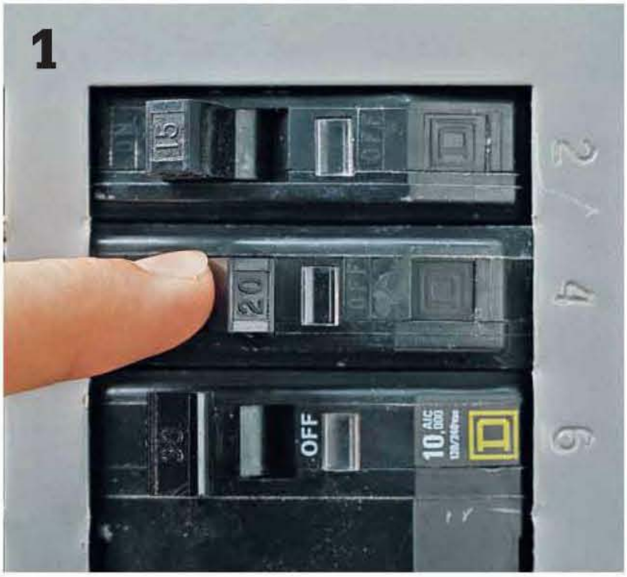

Start by shutting off the electrical power to the dishwasher circuit at the service panel. Also, turn off the water supply at the shutoff valve, usually located directly under the floor.

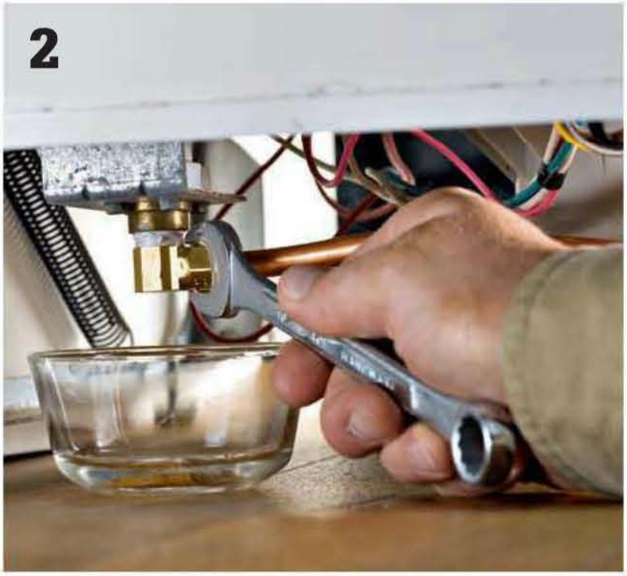

Disconnect old plumbing connections. First unscrew the front access panel. Once the access panel is removed, disconnect the water supply line from the L-fitting on the bottom of the unit. This is usually a brass compression fitting, so just turning the compression nut counterclockwise with an adjustable wrench should do the trick. Use a bowl to catch any water that might leak out when the nut is removed.

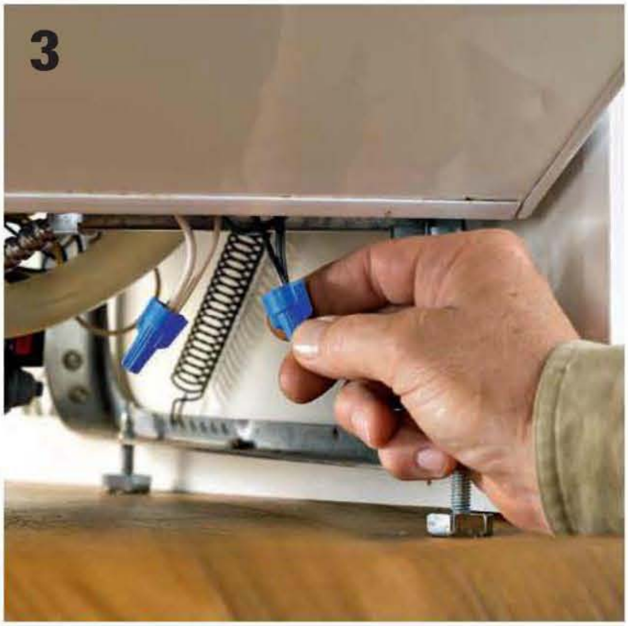

Disconnect old wiring connections. The dishwasher has an integral electrical box at the front of the unit where the power cable is attached to the dishwasher's fixture wires. Take off the box cover and remove the wire connectors that join the wires together.

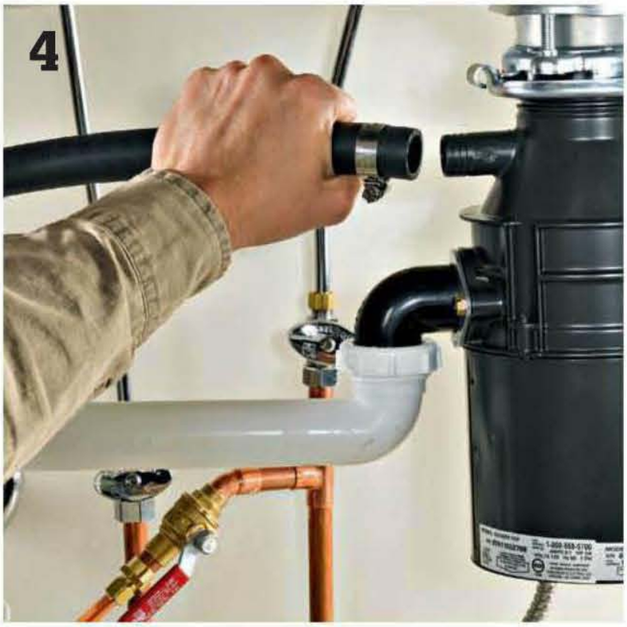

Disconnect the discharge hose, which is usually connected to the dishwasher port on the side of the garbage disposer. To remove it, just loosen the screw on the hose clamp and pull it off. You may need to push this hose back through a hole in the cabinet wall and into the dishwasher compartment so it won't get caught when you pull the dishwasher out.

(continued)

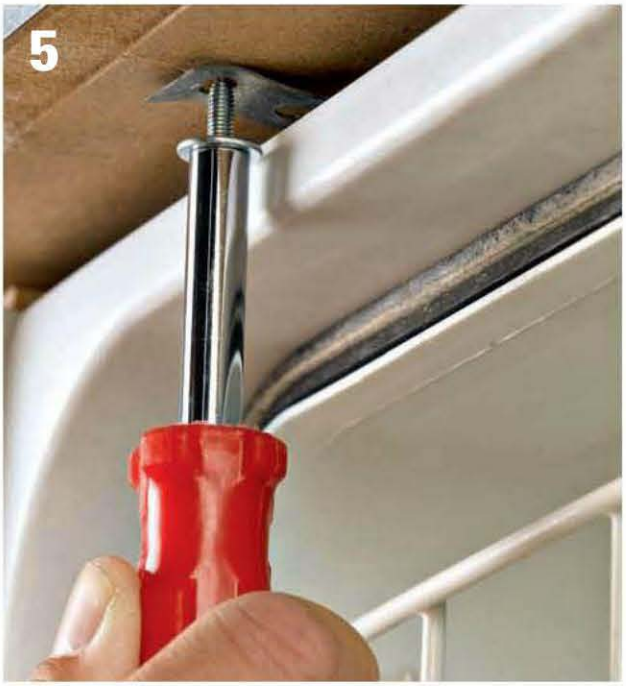

Detach the unit from surrounding cabinets before you pull it out. Remove the screws that hold the brackets to the underside of the countertop. Then put a piece of cardboard or old carpet under the front legs to protect the floor from getting scratched, and pull the dishwasher out.

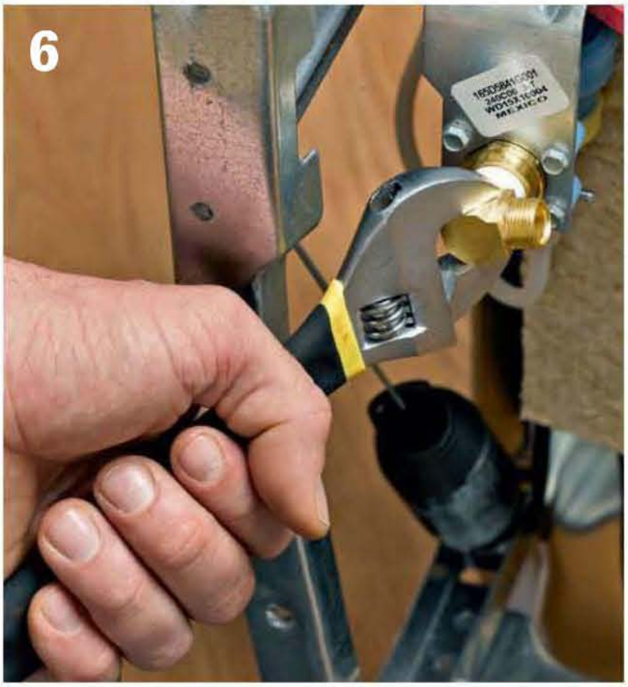

Prepare the new dishwasher. Tip it on its back and attach the new L-fitting into the threaded port on the solenoid. Apply some Teflon tape or pipe sealant to the fitting threads before tightening it, to prevent possible leaks.

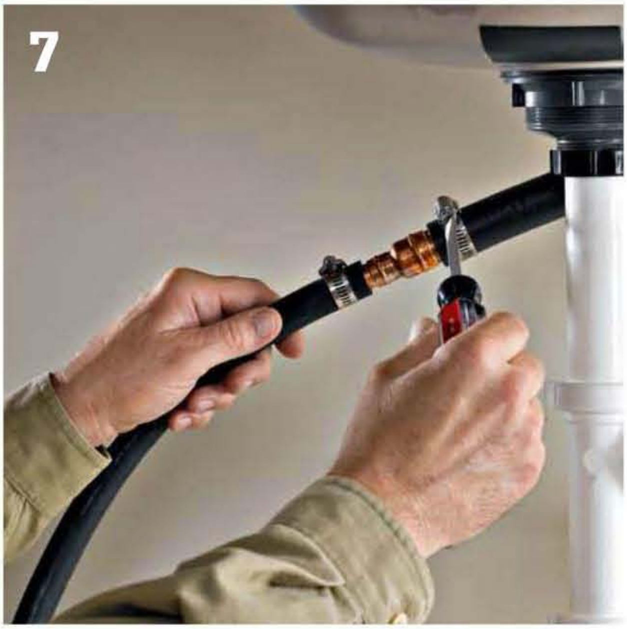

Attach a length of new automotive heater hose, usually ⅝" diameter, to the end of the dishwasher's discharge hose nipple with a hose clamp. The new hose you are adding should be long enough to reach from the discharge nipple to the port on the side of the kitchen sink garbage disposer.

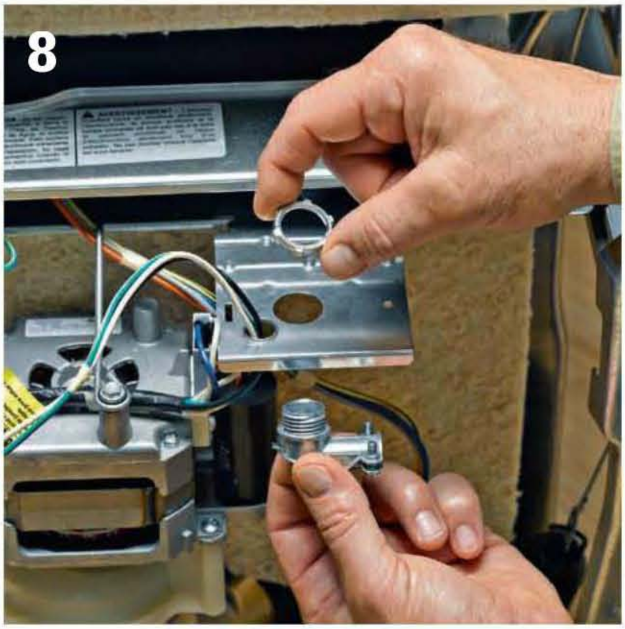

Prepare for the wiring connections. Like the old dishwasher, the new one will have an integral electrical box for making the wiring connections. To gain access to the box, just remove the box cover. Then install a cable connector on the back of the box and bring the power cable from the service panel through this connector. Power should be shut off at the main service panel at all times.

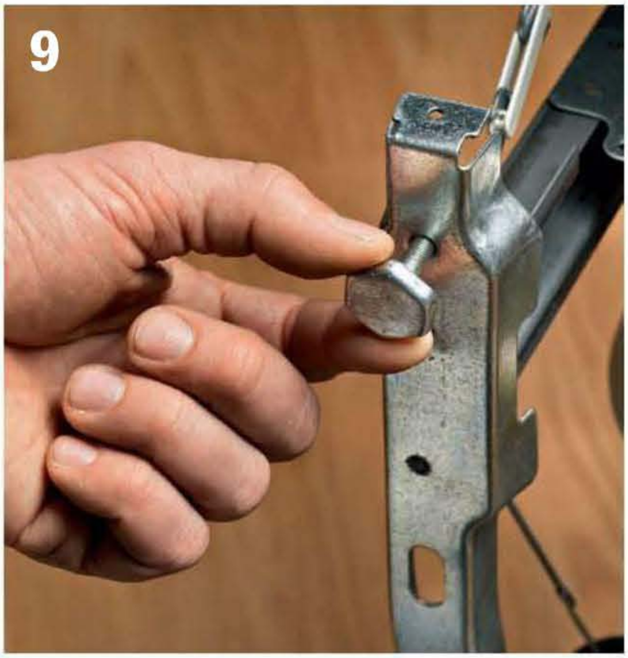

Install a leveling leg at each of the four corners while the new dishwasher is still on its back. Turn the legs into the threaded holes designed for them. Leave about ½" of each leg projecting from the bottom of the unit. (These will have to be adjusted later to level the appliance.) Tip the appliance up onto the feet and slide it into the opening. Check for level in both directions, and adjust the feet as required.

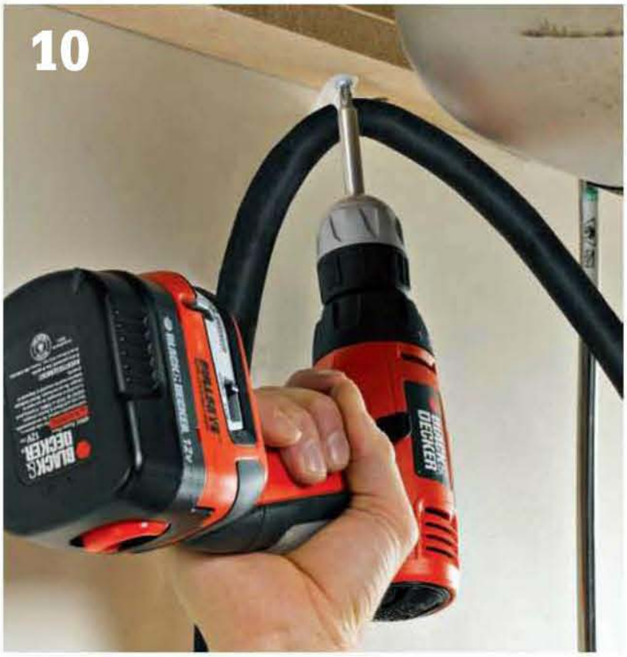

Once the dishwasher is level, attach the brackets to the underside of the countertop to keep it from moving. Then pull the discharge hose into the sink cabinet and install it so there's a loop that is attached with a bracket to the underside of the countertop. (This loop prevents waste water from flowing from the disposer back into the dishwasher.)

Lengthening a Discharge Hose ▸

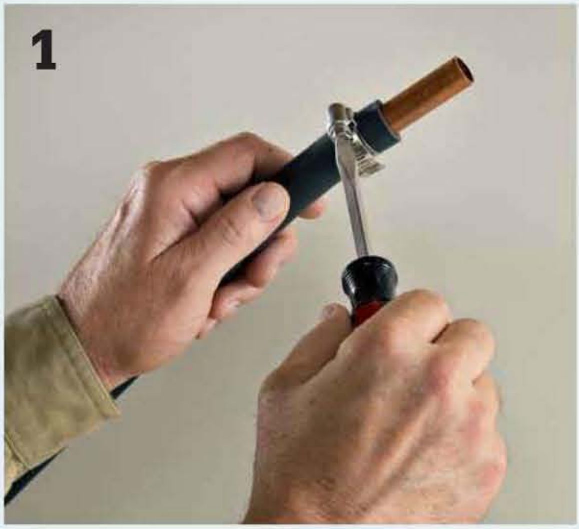

If the discharge hose has to be modified to fit onto the disposer port, insert a 4"-long piece of ½" copper tubing into the hose and hold it in place with a hose clamp. This provides a nipple for the rubber adapter that fits onto the disposer.

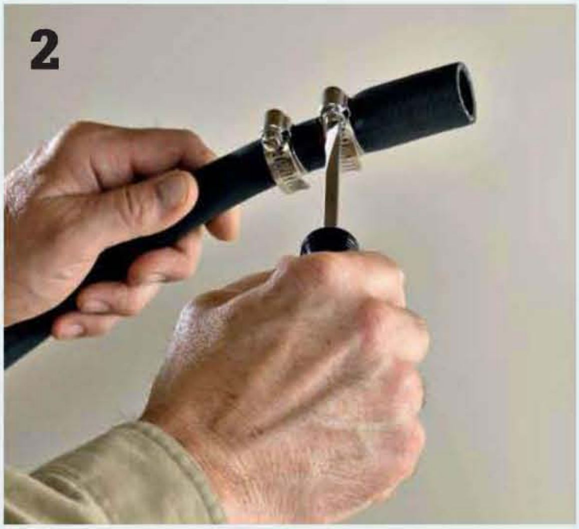

Clamp the rubber disposer adapter to the end of the copper tubing nipple. Then tighten the hose clamp securely.

(continued)

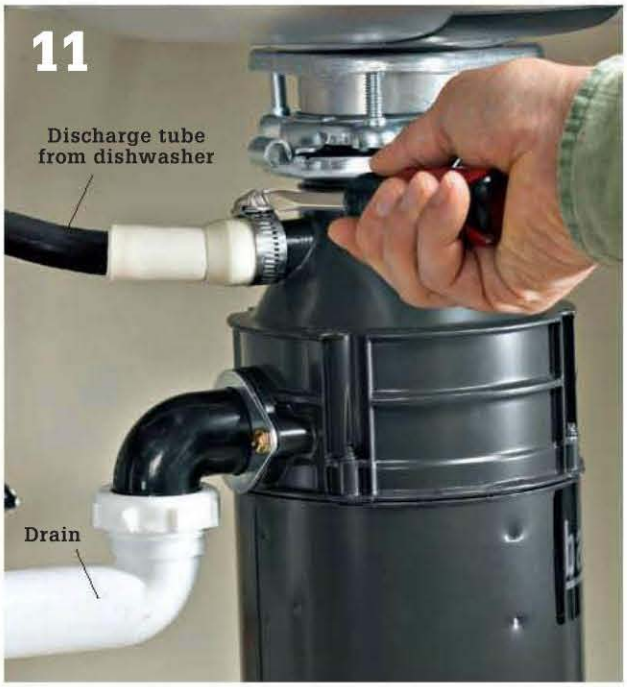

Push the adapter over the disposer's discharge nipple and tighten it in place with a hose clamp. If you don't have a disposer, this discharge hose can be clamped directly to a modified sink tailpiece that's installed below a standard sink strainer.

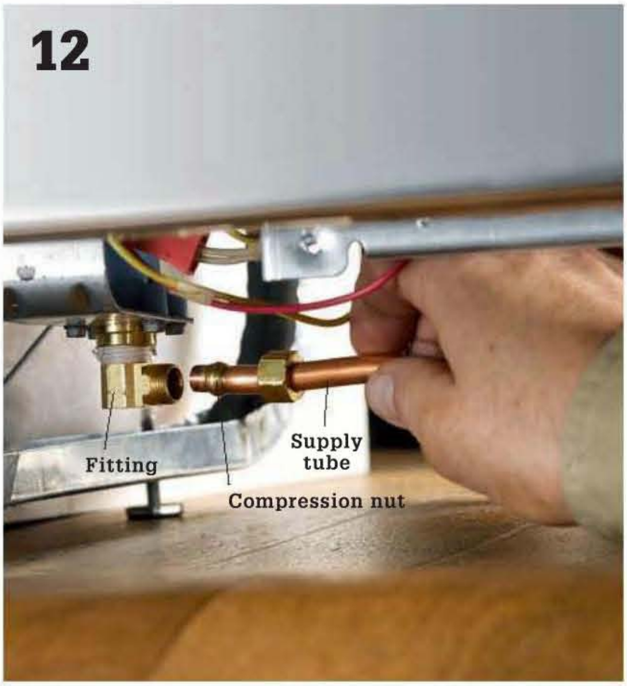

Adjust the L-fitting on the dishwasher's water inlet valve until it points directly toward the water supply tubing. Then lubricate the threads slightly with a drop of dishwashing liquid and tighten the tubing's compression nut onto the fitting. Use an adjustable wrench and turn the nut clockwise.

Complete the electrical connections by tightening the connector's clamp on the cable. Join the power wires to the fixture wires with wire connectors. Attach the ground wire (or wires) to the grounding screw on the box, and replace the cover.

Install the access panel, usually by hooking it on a couple of prongs just below the dishwasher's door. Install the screws (if any) that hold it in place, and turn on the water and power supplies. Replace the toe-kick panel at the bottom of the dishwasher.

Installing Food Disposers

Food disposers are standard equipment in the modern home, and most of us have come to depend on them to macerate our plate leavings and crumbs so they can exit the house along with waste water from the sink drain. If your existing disposer needs replacing, you'll find that the job is relatively simple, especially if you select a replacement appliance that is the same model as the old one. In that case, you can probably reuse the existing mounting assembly, drain sleeve, and drain plumbing.

Most food disposers are classified as "continuous feed" because they can only operate when an ON/OFF switch on the wall is being actively held down. Let go of the switch, and the disposer stops. Each appliance has a power rating between ⅓ and 1 HP (horsepower). More powerful models bog down less under load and the motors last longer because they don't have to work as hard. They are also costlier.

A disposer is hardwired to a switch mounted in an electrical box in the wall above the countertop. If your kitchen is not equipped for this, consult a wiring guide or hire an electrician. The actual electrical hookup of the appliance is quite simple (you only have to join two wires) but you can hire an electrician if you are not comfortable with the job.

Tools & Materials ▸

- Screwdriver
- Channel-type pliers
- Spud wrench (optional)
- Hammer
- Hacksaw or tubing cutter
- Kitchen drain supplies
- Drain auger
- Putty knife
- Mineral spirits
- Plumber's putty
- Wire caps
- Hose clamps
- Threaded Y-fitting
- Electrical tape

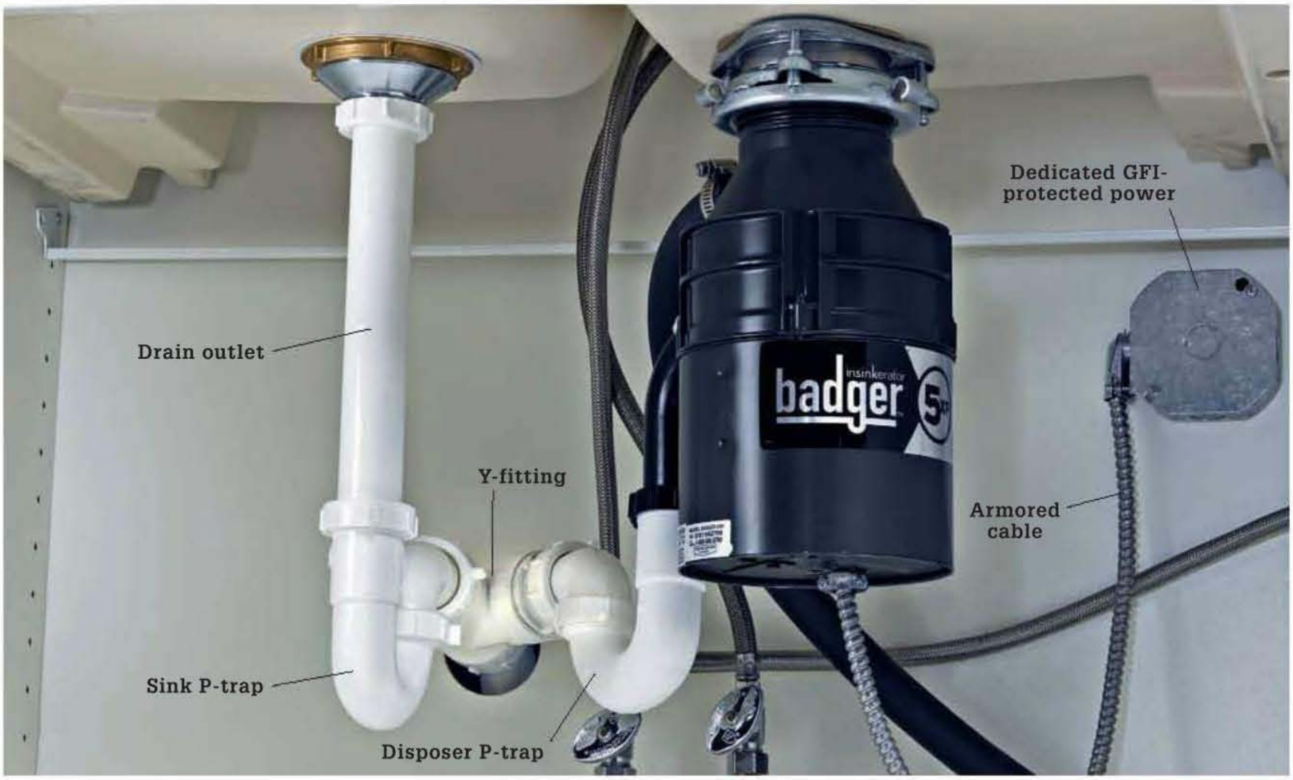

A properly functioning food disposer that's used correctly can actually help reduce clogs by ensuring that large bits of organic matter don't clog the drain system. Many plumbers suggest using separate P-traps for the disposer and the drain outlet tube as shown here.

How to Install a Food Disposer

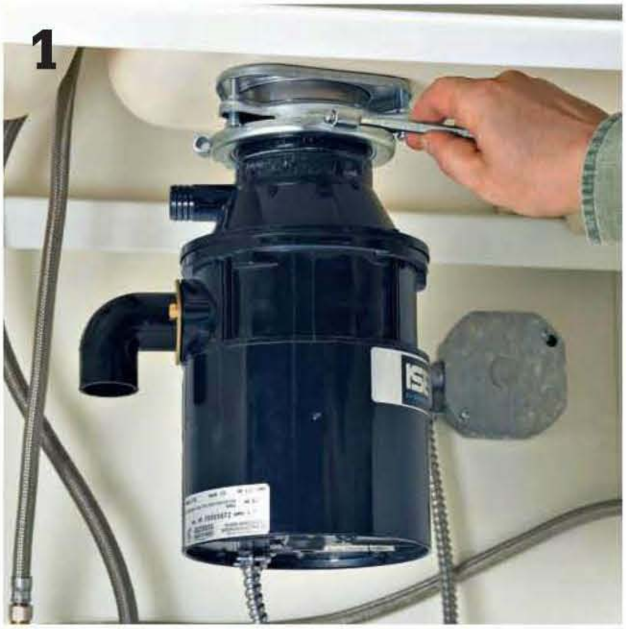

Remove the old disposer if you have one. You'll need to disconnect the drain pipes and traps first. If your old disposer has a special wrench for the mounting lugs, use it to loosen the lugs. Otherwise, use a screwdriver. If you do not have a helper, place a solid object directly beneath the disposer to support it before you begin removal. *Important: Shut off electrical power at the main service panel before you begin removal. Disconnect the wire leads, cap them, and stuff them into the electrical box.*

Tip ▸

Alternate: If you are installing a disposer in a sink that did not previously have one, remove the old sink strainer and drain tailpiece. Scrape up any old plumbers putty and clean the sink thoroughly around the drain opening with mineral spirits.

Clear the drain lines all the way to the branch drain before you begin the new installation. Remove the trap and trap arm first.

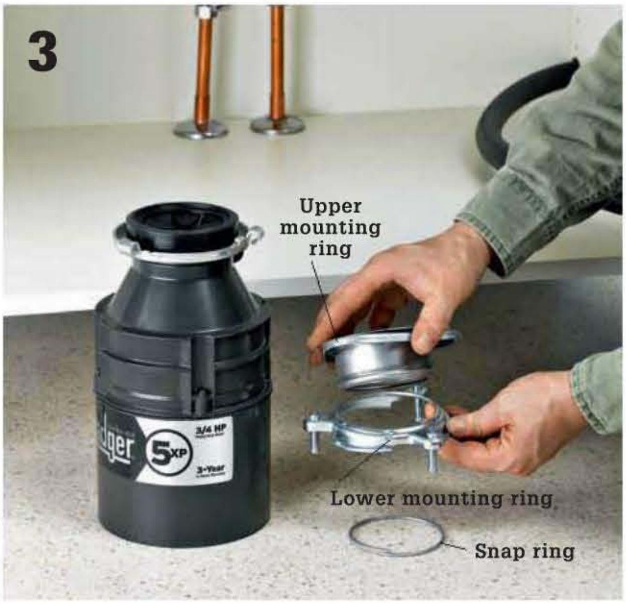

Disassemble the mounting assembly and then separate the upper and lower mounting rings and the backup ring. Also remove the snap ring from the sink sleeve.

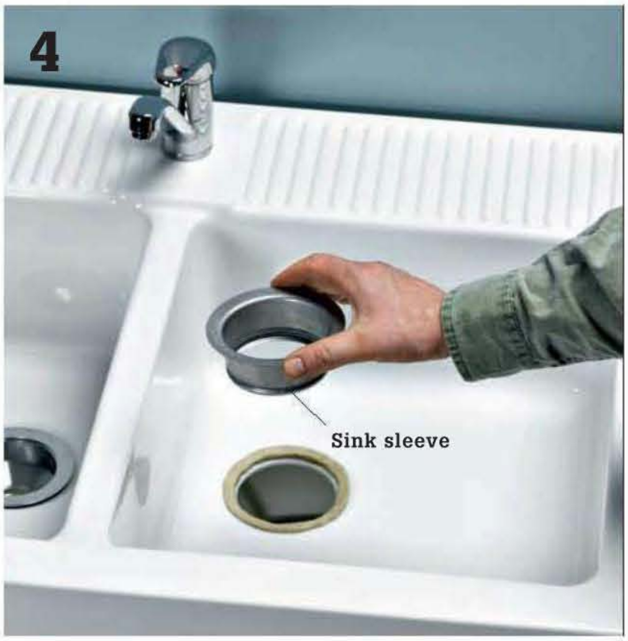

Press the flange of the sink sleeve for your new disposer into a thin coil of plumber's putty that you have laid around the perimeter of the drain opening. The sleeve should be well-seated in the putty.

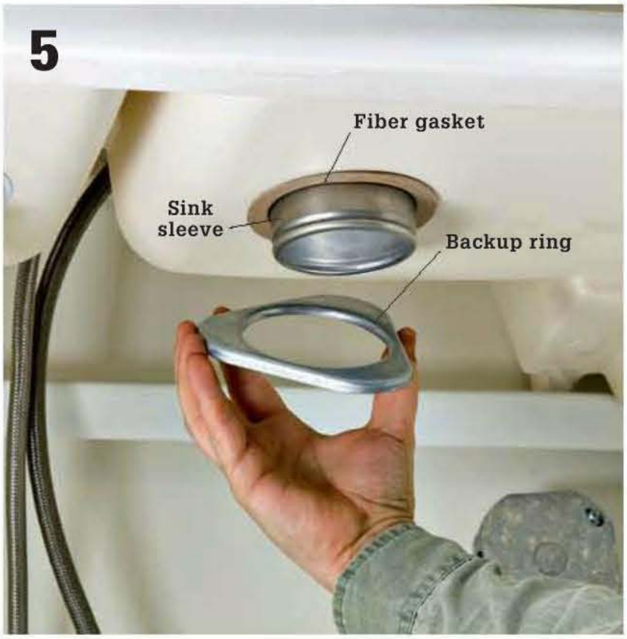

Slip the fiber gasket and then the backup ring onto the sink sleeve, working from inside the sink base cabinet. Make sure the backup ring is oriented the same way it was before you disassembled the mounting assembly.

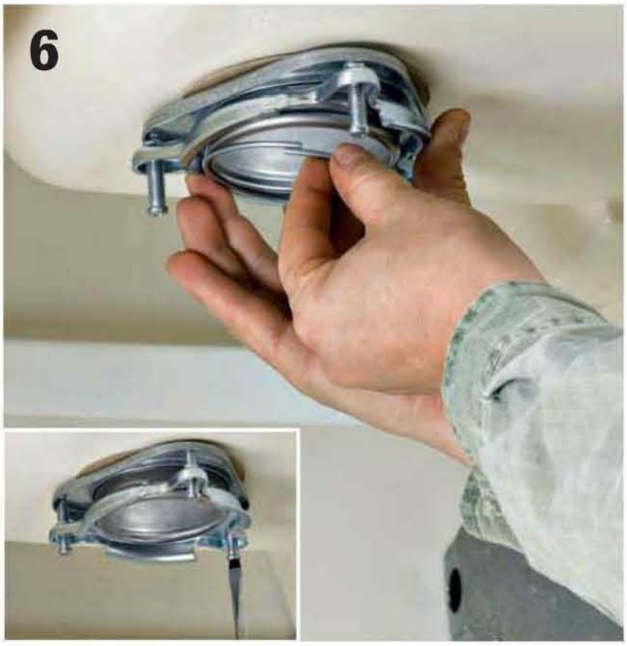

Insert the upper mounting ring onto the sleeve with the slotted ends of the screws facing away from the backup ring so you can access them. Then, holding all three parts at the top of the sleeve, slide the snap ring onto the sleeve until it snaps into the groove. Tighten the three mounting screws on the upper mounting ring until the tips press firmly against the backup ring (inset photo). It is the tension created by these screws that keeps the disposer steady and minimizes vibrating.

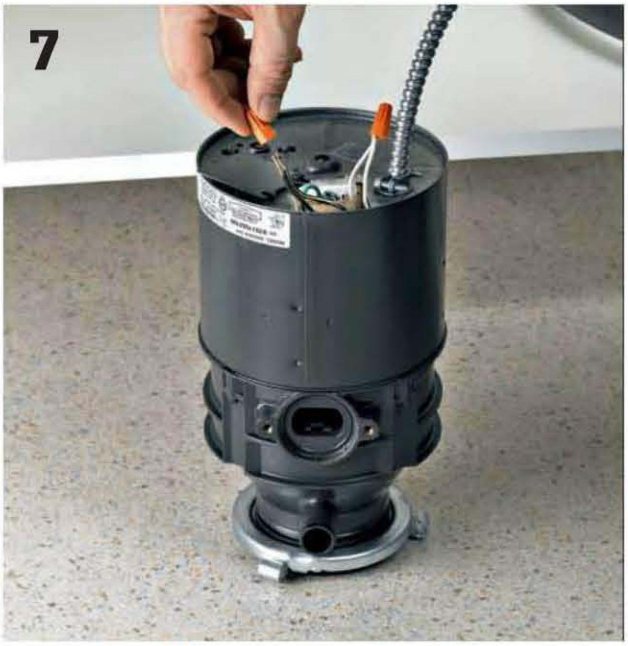

Make electrical connections before you mount the disposer unit on the mounting assembly. Make sure the power is still turned off at the service panel. Remove the access plate from the disposer. Attach the white and black feeder wires from the electrical box to the white and black wires (respectively) inside the disposer. Twist a small wire cap onto each connection and wrap it with electrical tape for good measure. Also attach the green ground wire from the box to the grounding terminal on your disposer.

(continued)

8

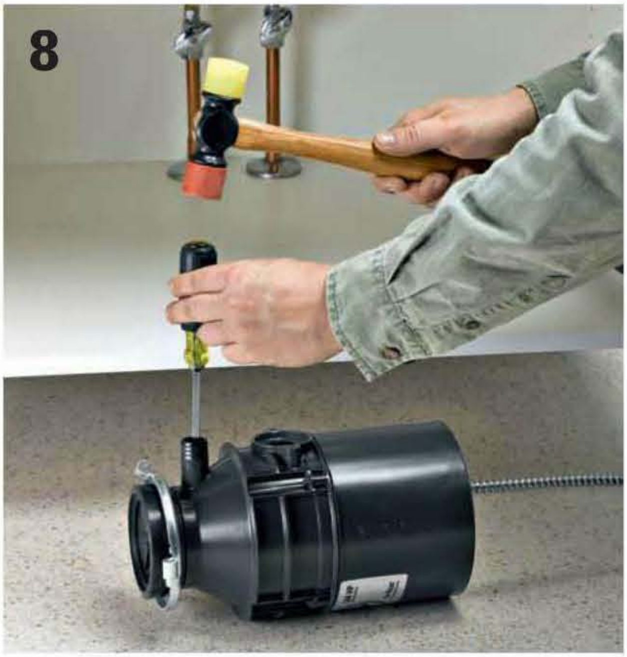

Knock out the plug in the disposer port if you will be connecting your dishwasher to the disposer. If you have no dishwasher, leave the plug in. Insert a large flathead screwdriver into the port opening and rap it with a mallet. Retrieve the knock plug from inside the disposer canister.

9

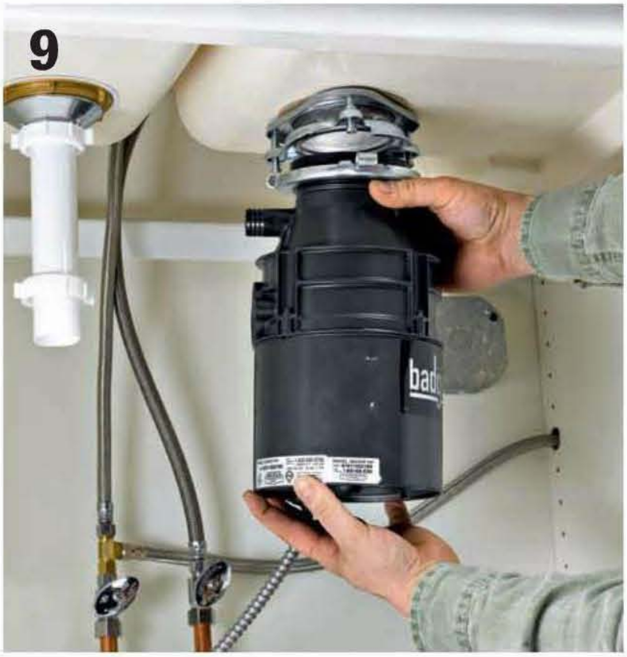

Hang the disposer from the mounting ring attached to the sink sleeve. To hang it, simply lift it up and position the unit so the three mounting ears are underneath the three mounting screws and then spin the unit so all three ears fit into the mounting assembly. Wait until after the plumbing hookups have been made to lock the unit in place.

10

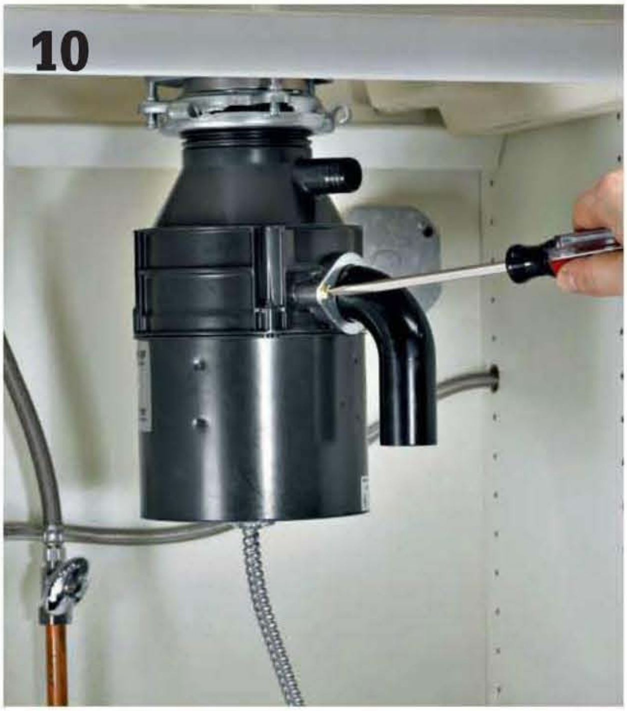

Attach the discharge tube to the disposer according to the manufacturer's instructions. It is important to get a very good seal here, or the disposer will leak. Go ahead and spin the disposer if it helps you access the discharge port.

11

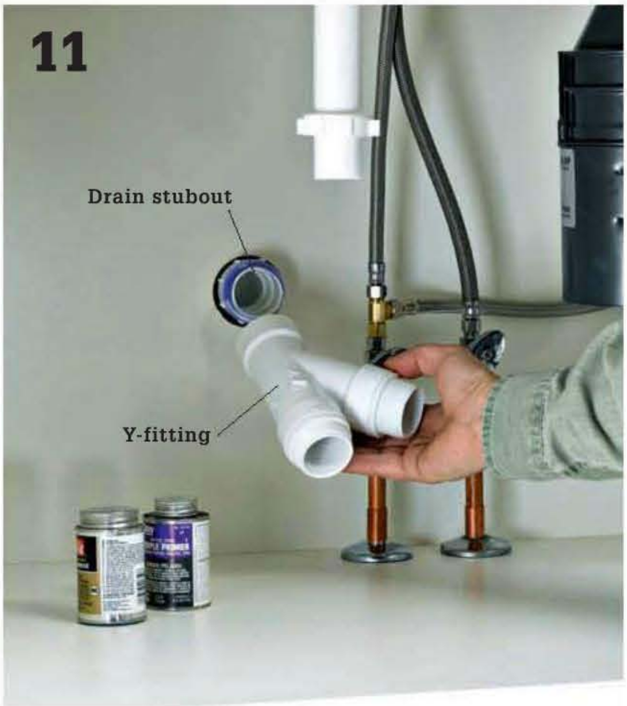

Attach a Y-fitting at the drain stubout. The Y-fitting should be sized to accept a drain line from the disposer and another from the sink. Adjust the sink drain plumbing as needed to get from the sink P-trap to one opening of the Y.

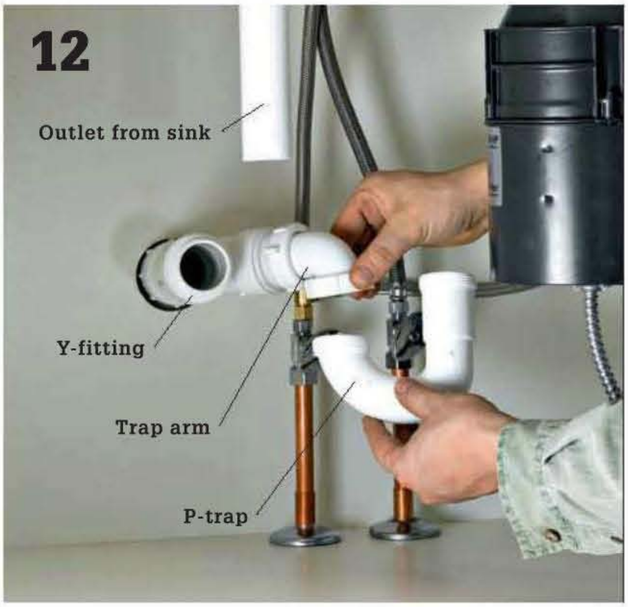

Install a trap arm for the disposer in the open port of the Y-fitting at the wall stubout. Then, attach a P-trap or a combination of a tube extension and a P-trap so the low end of the trap will align with the bottom of the disposer discharge tube.

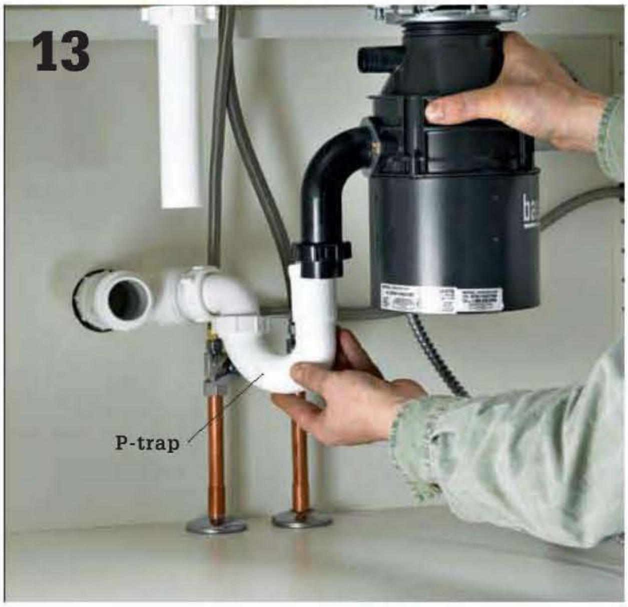

Spin the disposer so the end of the discharge tube is lined up over the open end of the P-trap and confirm that they will fit together correctly. If the discharge tube extends down too far, mark a line on it at the top of the P-trap and cut through the line with a hacksaw. If the tube is too short, attach an extension with a slip joint. You may need to further shorten the discharge tube first to create enough room for the slip joint on the extension. Slide a slip nut and beveled compression washer onto the discharge tube and attach the tube to the P-trap.

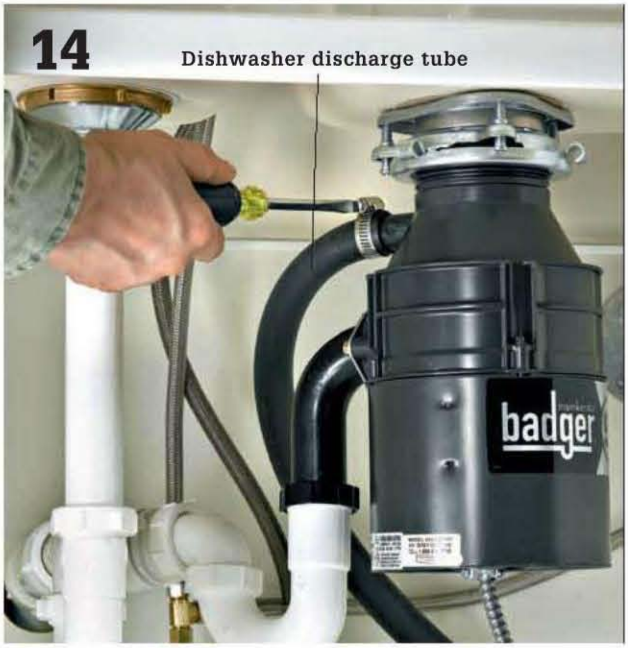

Connect the dishwasher discharge tube to the inlet port located at the top of the disposer unit. This may require a dishwasher hookup kit.

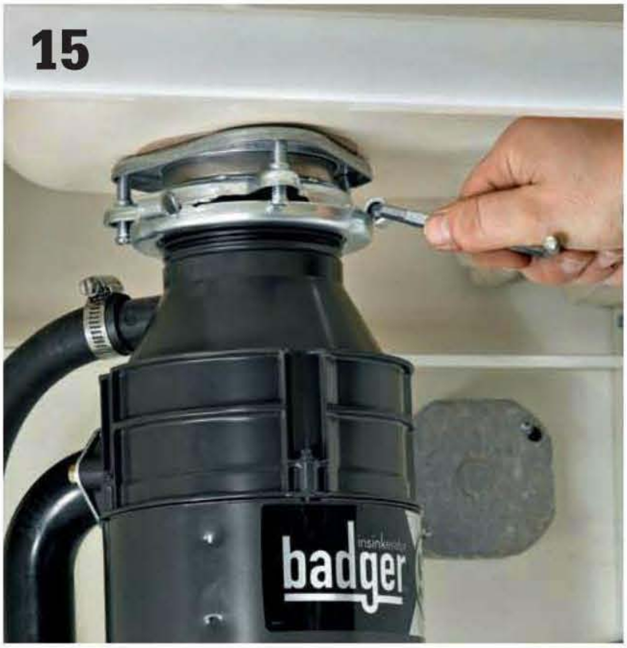

Lock the disposer into position on the mounting ring assembly once you have tested to make sure it is functioning correctly and without leaks. Lock it by turning one of the mounting lugs with a screwdriver until it makes contact with the locking notch.

Installing Undercabinet Lights

Hardwired undercabinet lights illuminate the kitchen countertop and sink areas that fall in the shadow of ceiling lights. Most of these light fixtures, which are often called *strip lights,* utilize fluorescent, halogen, or xenon bulbs that emit very low levels of heat and are therefore very efficient.

If you are doing a kitchen remodel with all-new cabinets, run the new light circuit wiring before the cabinets are installed. For a retrofit, you'll need to find an available power source to tie into. Options for this do not include the dedicated 20-amp small-appliance circuits that are required in kitchens. The best bet is to run new circuit wire from a convenient ceiling light switch box, but this will mean cutting into the walls to run cable. Another option is to locate a receptacle that's on the opposite side of a shared wall, preferably next to a location where a base cabinet is installed in the kitchen. By cutting an access hole in the cabinet back you can tie into the receptacle box and run cable through the wall behind the cabinets, up to the upper cabinet location, and out the wall to supply the fixture that's mounted to the underside of the upper cabinet.

You can purchase undercabinet lights that are controlled by a wall switch, but most products have an integral on/off button so you can control lights individually.

Tools & Materials ▸

Circuit tester
Utility knife
Wallboard saw
Hammer
Screwdriver
Drill and hole saw
Jigsaw

Wire stripper
Undercabinet lighting kit
14/2 NM cable
Wire connectors
Switch box
Switch

Undercabinet lights provide directed task lighting that bring sinks and countertop work surfaces out from the shadows. Hardwired lights may be controlled either by a wall switch or an onboard on/off switch located on the fixture. *Note: Do not supply power for lights from a small-appliance circuit.*

How to Install a Hardwired Undercabinet Light

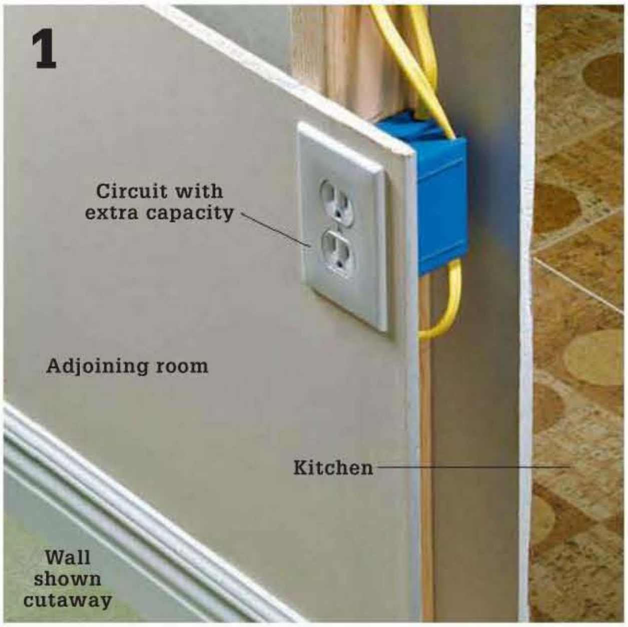

Look in an adjoining room for a usable power source in the form of a receptacle that has a box located in the wall behind your base cabinets. Unlike the small-appliance circuit with outlets in your backsplash area, these typically are not dedicated circuits (which can't be expanded). Make sure that the receptacle's circuit has enough capacity to support another load. Shut the power to the receptacle off at the main service panel and test for power.

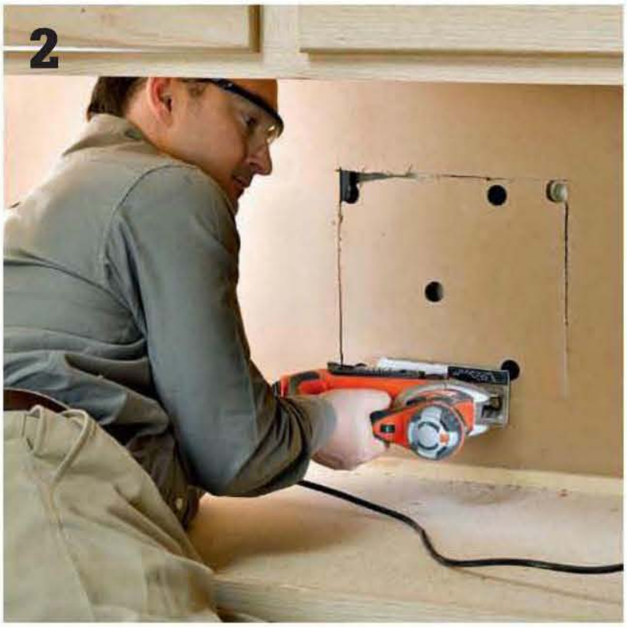

Cut a hole in the base cabinet back panel, roughly the area where you know the usable receptacle to be. Use a keyhole saw or drywall saw and make very shallow cuts until you have positively identified the locations of the electrical box and cables. Then finish the cuts with a jigsaw.

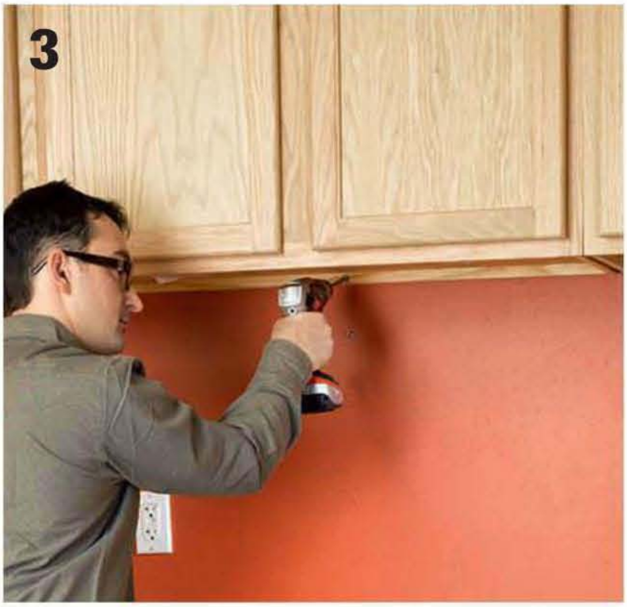

Drill an access hole through the cabinet frame and wall for the cable that will feed the undercabinet light. A ½" dia. hole should be about the right size if you are using 12-ga. or 14-ga. sheathed NM cable.

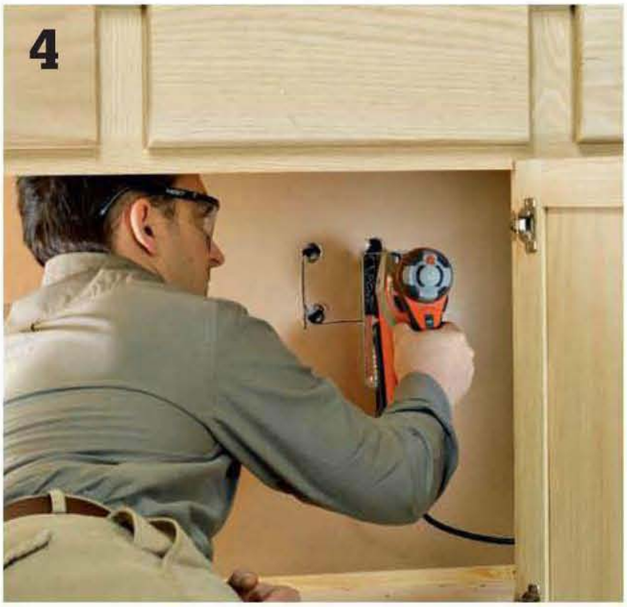

Cut a small access hole (4 × 4" or so) in the back panel of the base cabinet directly below the undercabinet light location.

(continued)

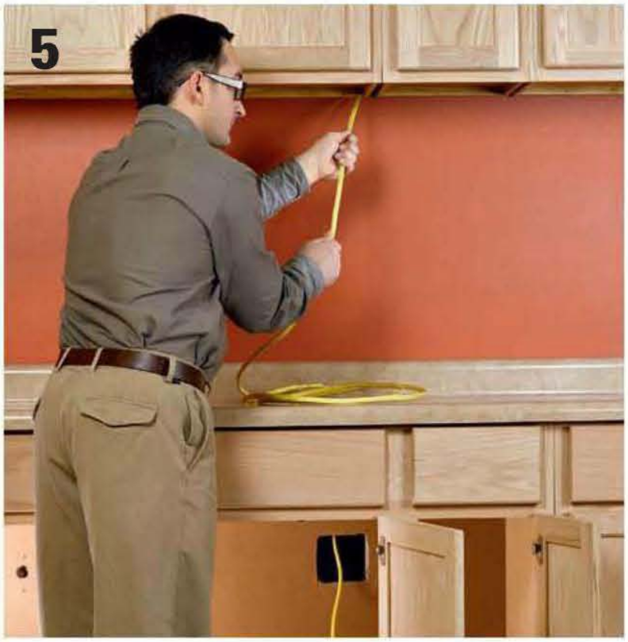

Feed cable into the access hole at the light location until the end reaches the access hole below. Don't cut the cable yet. Reach into the access hole and feel around for the free cable end and then pull it out through the access hole once you've found it. Cut the cable, making sure to leave plenty of extra on both ends.

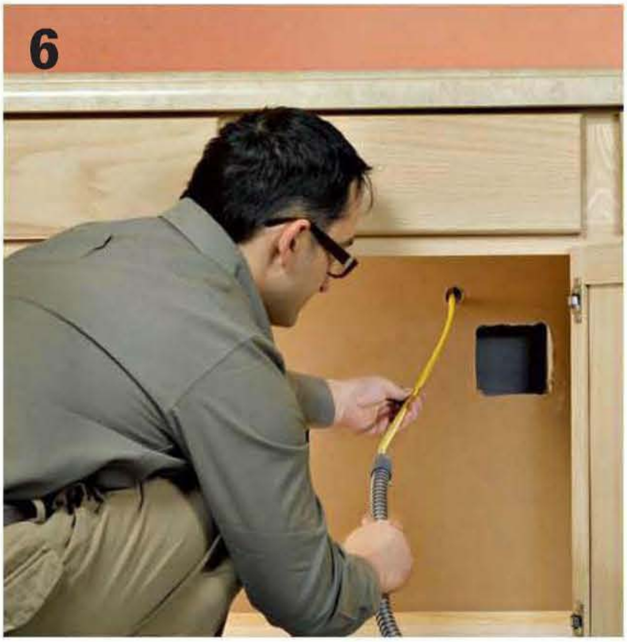

String the cable into a piece of flexible conduit that's long enough to reach between the two access holes in the base cabinets. To make patching the cabinet back easier, drill a new access hole for the cable near the square access hole. Attach a connector to each end of the conduit to protect the cable sheathing from the sharp edges of the cut metal.

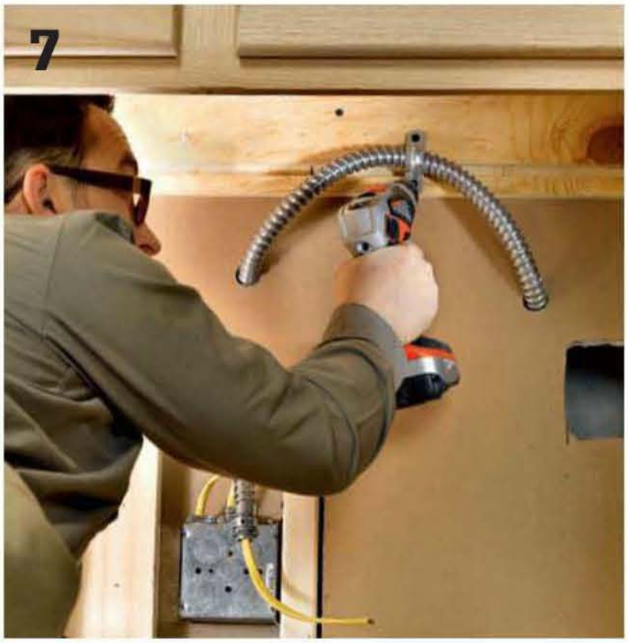

Hang the conduit with hanger straps attached to the base cabinet frame or back panel, drilling holes in the side walls of the cabinet where necessary to thread the conduit through. On back panels, use small screws to hang the straps instead of brads or nails. Support the conduit near both the entrance and the exit holes (the conduit should extend past the back panels by a couple of inches).

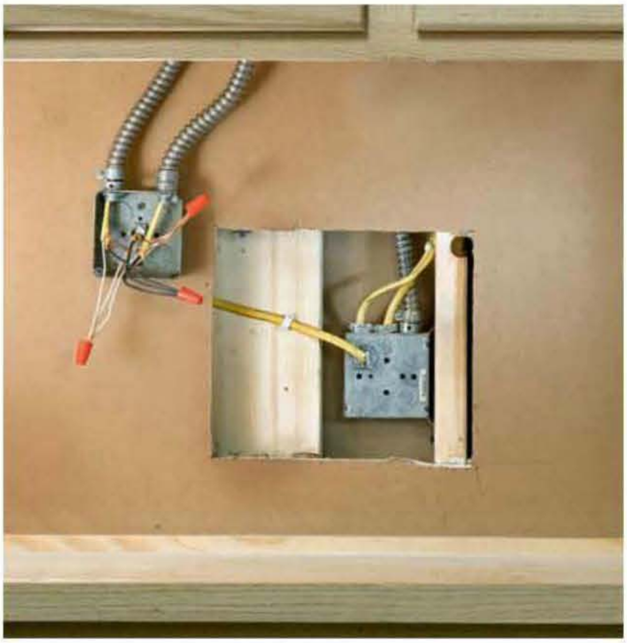

Variation: If you are installing more than one undercabinet light, run cable down from each installation point as you did for the first light. Mount an electrical junction box to the cabinet back near the receptacle providing the power. Run the power cables from each light through flexible conduit and make connections inside the junction box. Be sure to attach the junction box cover once the connections are made.

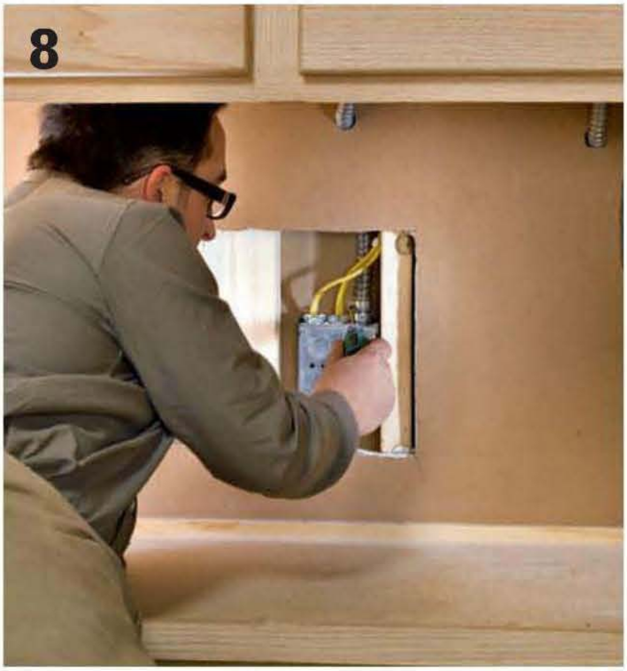

Remove the receptacle from the box you are tying into and insert the new circuit cable into one of the knockouts using a cable clamp. Check a wire capacity chart (see page 90) to make sure the box is big enough for the new conductors. Replace it with a larger box if necessary. Reinstall the receptacle once the connections are made.

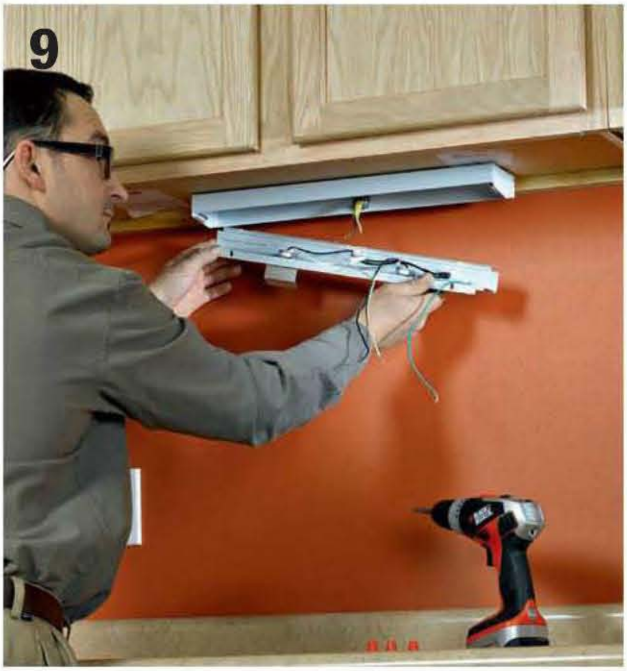

Install the undercabinet light. Some models have a removable diffuser that allows access to the fixture wires, and these should be screwed to the upper cabinet prior to making your wiring hookups. Other models need to be connected to the circuit wires before installation. Check your manufacturer's installations.

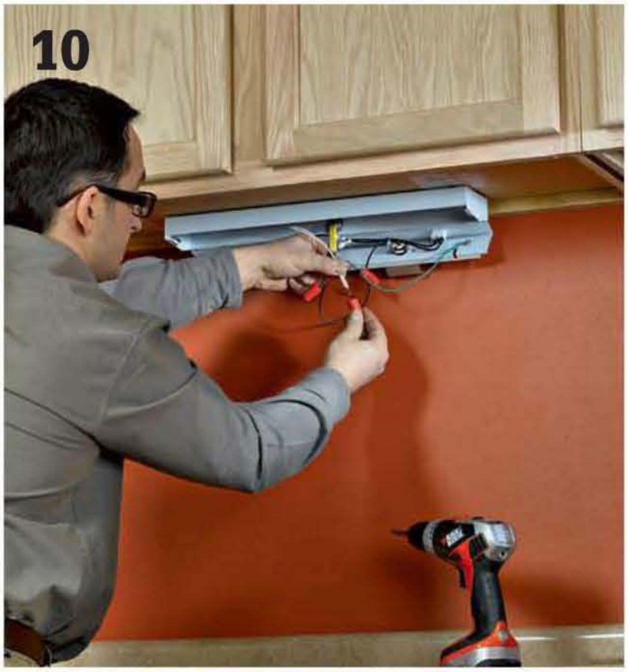

Connect wires inside the light fixture according to the light manufacturer's directions. Make sure the incoming cable is stapled just before it enters the light box and that a cable clamp is used at the knockout in the box to protect the cable. Restore power and test the light.

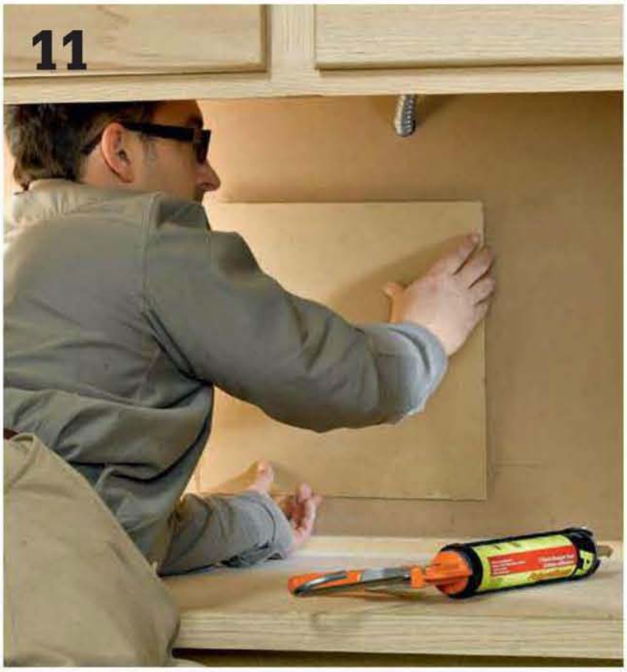

Cut patches of hardboard and fit them over the access holes, overlapping the edges of the cutouts. Adhere them to the cabinet backs with panel adhesive.

Green Kitchens ▸

In many ways, a green kitchen is just like any well-designed kitchen. There are windows for natural lighting and air circulation. Surfaces are durable, hygienic, and low-maintenance. The room is well lighted, with overhead fixtures and plenty of task lighting. And the overall plan is designed for maximum work efficiency and tailored to the way the homeowners use the kitchen on a daily basis. In terms of aesthetics, a green kitchen can be just as beautiful, if not more so, than a conventional kitchen.

So what's different about a green kitchen? For starters, it has highly efficient appliances, particularly the refrigerator. It contains sustainable and non-toxic materials, like formaldehyde-free cabinet cases and linoleum flooring. It's also ready to deal with the variety of waste products surrounding food preparation, including organic garbage, recyclable materials, and gray water.

Many green kitchens are further defined by what they don't have: trendy upgrades and extras that do little to improve anyone's daily experience in the kitchen. Depending on the type of cooking you do, you might get a lot of use out of special features, like a pot filler faucet, an extra prep sink, warming drawers, or a commercial-style range. But in reality, most cooks don't get full value from these pricey additions, which take up space and are often over-engineered and energy-inefficient (not to mention the environmental impact of producing them).

Whether you're planning a major kitchen remodel or simply replacing an item or two, here are some of the main elements you should consider for making the space more useful, healthful, and resource-efficient.

Efficiency and common sense are two of the hallmarks of any green room, but this is especially true in the kitchen. Keep the appliances as small as you can and look for natural materials, like this butcherblock countertop made from maple cutoffs.

PORTRAIT OF A GREEN KITCHEN

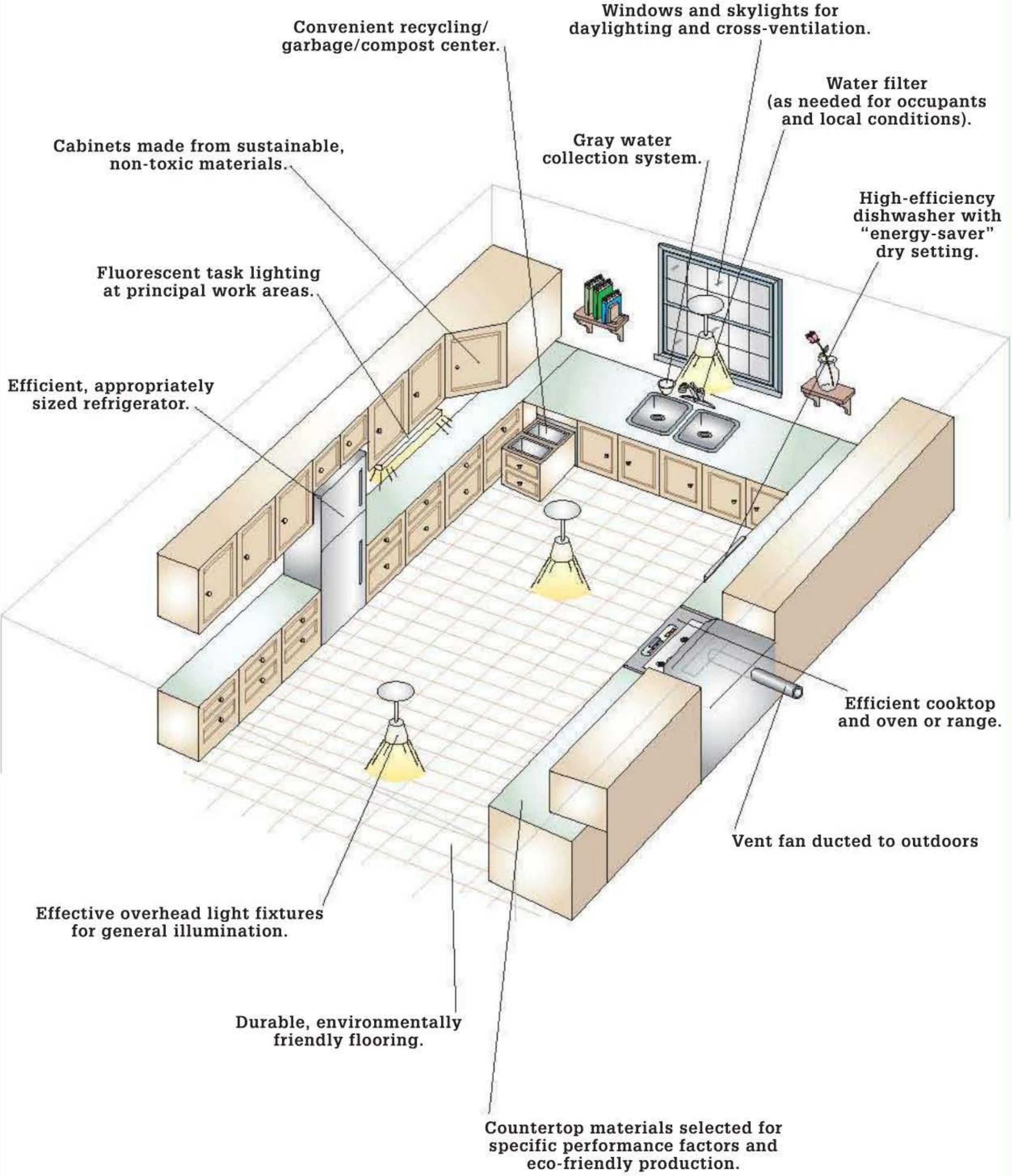
Convenient recycling/ garbage/compost center.
Windows and skylights for daylighting and cross-ventilation.
Water filter (as needed for occupants and local conditions).
Cabinets made from sustainable, non-toxic materials.
Gray water collection system.
High-efficiency dishwasher with "energy-saver" dry setting.
Fluorescent task lighting at principal work areas.
Efficient, appropriately sized refrigerator.
Efficient cooktop and oven or range.
Vent fan ducted to outdoors
Effective overhead light fixtures for general illumination.
Durable, environmentally friendly flooring.
Countertop materials selected for specific performance factors and eco-friendly production.

Green Kitchen Appliances ▸

Because it's full of hardworking appliances, the kitchen uses more energy than any other room in the house. Upgrading your old machines with more efficient models is one of the simplest ways to help make your kitchen green. Refrigerators are the big consumers, accounting for up to 15% of the average home's total electricity usage. Fortunately, you don't have to sacrifice quality or style to save energy. Highly efficient appliances, like many other machines, are simply better designed and better built than their hoggish counterparts. And you can't tell a miserly model by its looks.

Lots of do-it-yourselfers install their own appliances; see pages 402 to 407 for steps on installing a high-efficiency dishwasher. A refrigerator is the easiest to install: Just plug it in, slide it in place, and level it. If yours has an icemaker, you'll have to connect a small water line in the back of the unit (and check it for leaks in the beginning).

If you're replacing an existing appliance, find out where you can recycle the old unit. Appliances are a major source of recyclable steel and other materials, and besides, dumping an operable machine in the landfill just to make room for a new one isn't a very green move. Appliances that still work can find a great home through a local charity organization (this means an old, inefficient appliance is still in circulation, but it postpones the production of a new product, and it helps out those in need). To scrap an ailing or dead appliance, contact your local recycling authority or Earth 911, online at www.earth911.org.

Whatever you do, don't keep your old refrigerator and move it into the garage for backup storage. Keeping a second fridge plugged in will undo all energy-saving efforts from your kitchen upgrades many times over.

European sized refrigerators are becoming increasingly popular as a way to control energy usage. A typical model like the one above has a refrigerator compartment of 7-to-8 cubic feet in capacity, with a freezer that's 3 to 4 cu. ft.

Scale is important and many kitchen appliance manufacturers have begun marketing to this. Narrow dishwashers with smaller compartments offer ample space for some users, and they do their job with much less water and energy.

ENERGY STAR & ENERGYGUIDE LABELS

If you're in the market for a new appliance, it's time to start paying attention to those yellow-and-black ENERGYGUIDE labels you see pasted to the front of many new products. The labels, required by the Federal Trade Commission, help consumers quickly compare models for their energy efficiency. The box in the middle of each label provides the "energy use" of all models in a given class, with an arrow showing where that particular model falls within the range. The label also gives you an estimated annual operating cost for the appliance.

Energy Star is the Department of Energy's wide-reaching program that awards Energy Star status to the most efficient products in a given class. The program applies to major appliances, windows, doors, HVAC systems, and other household items. In general, Energy Star products exceed the federal government's energy-efficiency standards and perform within the top 25% of their category.

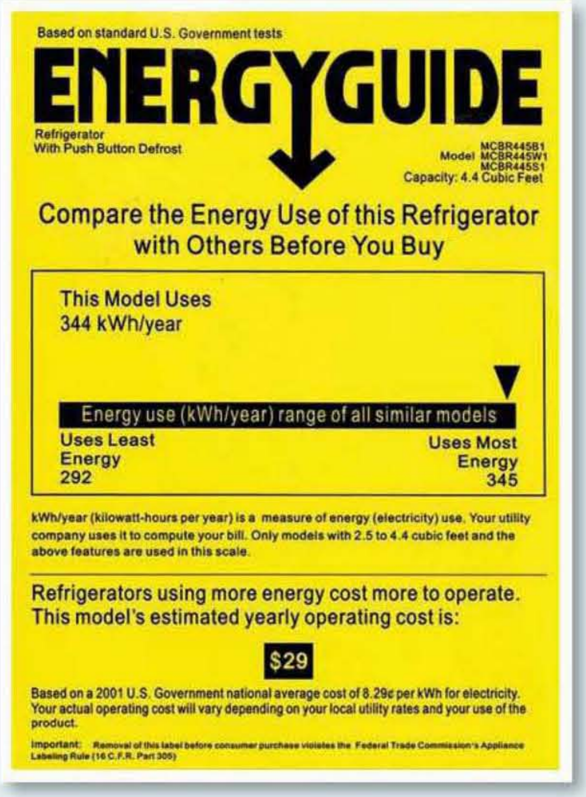

Based on standard U.S. Government tests

ENERGYGUIDE

Refrigerator
With Push Button Defrost

Model MCBR445B1
MCBR445W1
MCBR445S1
Capacity: 4.4 Cubic Feet

Compare the Energy Use of this Refrigerator with Others Before You Buy

This Model Uses
344 kWh/year

Energy use (kWh/year) range of all similar models

Uses Least Energy 292

Uses Most Energy 345

kWh/year (kilowatt-hours per year) is a measure of energy (electricity) use. Your utility company uses it to compute your bill. Only models with 2.5 to 4.4 cubic feet and the above features are used in this scale.

Refrigerators using more energy cost more to operate. This model's estimated yearly operating cost is:

$29

Based on a 2001 U.S. Government national average cost of 8.29¢ per kWh for electricity. Your actual operating cost will vary depending on your local utility rates and your use of the product.

Important: Removal of this label before consumer purchase violates the Federal Trade Commission's Appliance Labeling Rule (16 C.F.R. Part 305)

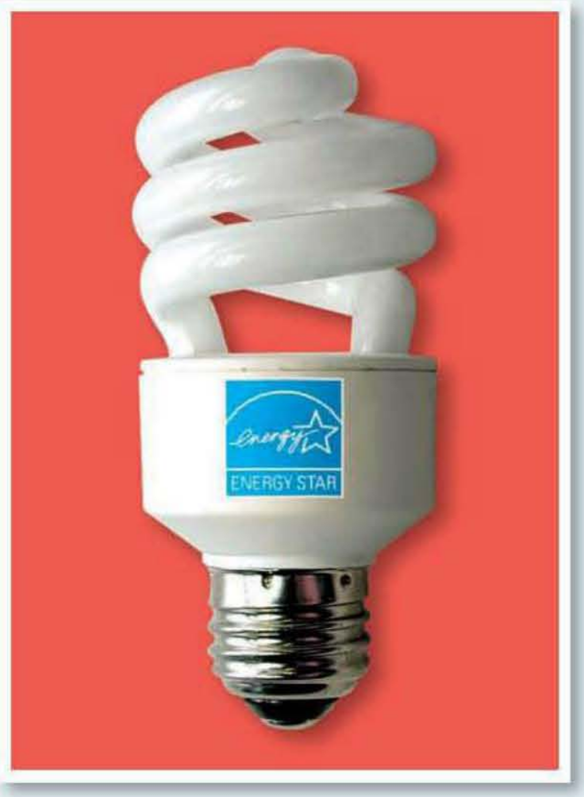

ADDITIONAL RESOURCES FOR CHOOSING & USING KITCHEN APPLIANCES

- U.S. Department of Energy (DOE): www.eere.energy.gov
- Energy Star: www.energystar.gov
- Consumer Reports, for unbiased analysis of appliance performance and other shopping considerations, online at www.consumerreports.org
- American Council for an Energy-Efficient Economy: www.aceee.org
- Your local utility provider
- Your state's energy authority; some states offer rebates on high-efficiency appliances
- Association of Home Appliance Manufacturers: www.aham.org
- The Green Guide: www.thegreenguide.com

Bathroom Remodeling

Bathroom fixtures and materials have changed dramatically in recent years. Tubs can rival healthclub spas, toilets and showerheads are more water efficient, and sinks run the gamut from stainless steel to wood basins.

Before you jump into a bathroom remodeling project, decide what's most important to you. Budget determines how luxurious you can get, but maybe it's worthwhile to splurge on that one item that would make you feel most pampered.

Home centers and kitchen and bath showrooms have a variety of bathroom setups you can browse through to see how the products look and feel. It's always a good idea to see a product in person before doling out the cash.

In this chapter:

- Planning & Designing Bathrooms
- Installing Shower Kits
- Installing Alcove Bathtubs
- Installing 3-piece Tub Surrounds
- Installing Sliding Tub Doors
- Tiling Bathroom Walls
- Installing Toilets
- Installing Wall-hung Vanities
- Installing Vessel Sinks
- Installing Pedestal Sinks
- Installing Vanity Cabinets
- Installing Bathroom Sink Faucets & Drains
- Installing Tub Spouts
- Installing Vanity Lights
- Installing Grab Bars
- Installing Glass Shelving
- Installing Towel Warmers

Planning & Designing Bathrooms

Once you've drawn up your plans and created a materials list, you'll need to have them reviewed by your local building department. Getting approval early in the process can save you time and expense later. To help ensure success, here are some design standards for you to follow:

The National Kitchen and Bath (NKBA) publishes a list of bathroom design standards to help people plan rooms that are safe and accessible to all users (see Resources, page 554).

Your bathroom probably won't conform to all of the recommended standards, but they can help guide your overall plan. What your plan must include is everything prescribed by the local building codes, including plumbing and wiring codes.

Bathroom Design Standards ▸

Codes and permits are necessary to ensure safety in any remodel. They're not the most fun to focus on—not like choosing just the right floor covering or deciding between granite or marble countertops—but they are important.

- Plan doorways with a clear floor space equal to the door's width on the push side and greater than the door's width on the pull side. *Note: Clear floor spaces within the bathroom can overlap.*
- Design toilet enclosures with at least 36" × 66" of space; include a pocket door or a door that swings out toward the rest of the bathroom.
- Install toiletpaper holders approximately 26" above the floor, toward the front of the toilet bowl.
- Place fixtures so faucets are accessible from outside the tub or shower. Add antiscald devices to tub and sink faucets (they are required for shower faucets).
- Avoid steps around showers and tubs, if possible.
- Fit showers and tubs with safety rails and grab bars.
- Install shower doors so they swing open into the bathroom, not the shower.
- Use tempered glass or another type of safety glass for all glass doors and partitions.
- Include storage for soap, towels, and other items near the shower, located within 15" to 48" above the floor. These should be accessible to a person in the shower or tub.
- Provide natural light equal to at least 10% of the floor area in the room.
- Illuminate all activity centers in the bathroom with task and ambient lighting.
- Provide a minimum clearance of 15" from the centerline of sinks to any sidewalls. Double-bowl sinks should have 30" clearance between bowls from centerline to centerline.
- Provide access panels for all electrical, plumbing, and HVAC systems connections.
- Include a ventilation fan that exchanges air at a rate of 8 air changes per hour.
- Choose countertops and other surfaces with edges that are smoothed, clipped, or radiused.

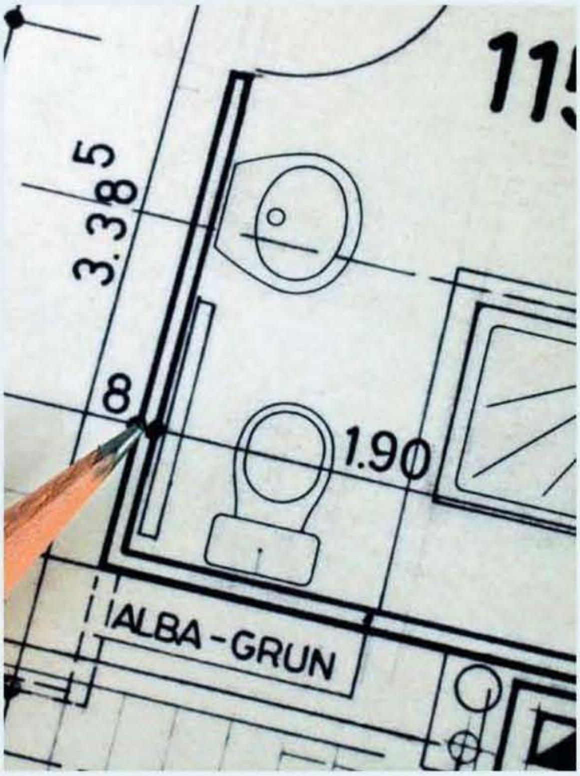

Building Codes for Bathrooms ▸

The following are some of the most common building codes for bathrooms. Contact your local building department for a list of all codes enforced in your area.

- The minimum ceiling height in bathrooms is 7 ft. Minimum floor area is determined by clearances around fixtures.
- Sinks must be at least 4" from side walls and have 21" of clearance in front.
- Sinks must be spaced 4" away from neighboring sinks and toilets, and 2" away from bathtubs.
- Toilets must be centered 15" from side walls and tubs, with 21" clearance in front.
- New and replacement toilets must be low-flow models (1.6 gal./flush).
- Shower stalls must be at least 30" × 30", with 24" of clearance in front of shower openings.
- Steps must be at least 10" deep and no higher than 7¼".
- Faucets for showers and combination tub/showers must be equipped with antiscald devices.
- Supply lines that are ½" in diameter can supply a single fixture, or one sink and one toilet.
- A ¾"-diameter supply line must be used to supply two or more fixtures.
- Waste and drain lines must slope ¼" per foot toward the main DWV stack to aid flow and prevent blockage.
- Each bathroom must be wired with at least one 20-amp circuit for GFCI-protected receptacles, and one 15-amp (minimum) circuit for light fixtures and vent fans without heating elements.
- All receptacles must be GFCI-protected.
- There must be at least one permanent light fixture controlled by a wall switch.
- Wall switches must be at least 60" away from bathtubs and showers.
- Toilet, shower, vanity, or other bathroom compartments must have adequate lighting.
- Light fixtures over bathtubs and showers must be vaporproof, with a UL rating for wet areas.
- Vanity light fixtures with built-in electrical receptacles are prohibited.
- Whirlpool motors must be powered by dedicated GFCI-protected circuits.
- Bathroom vent ducts must terminate no less than 10 ft. horizontally or 3 ft. vertically above skylights.

**Note: Codes for accessible bathrooms may differ (see page 429).*

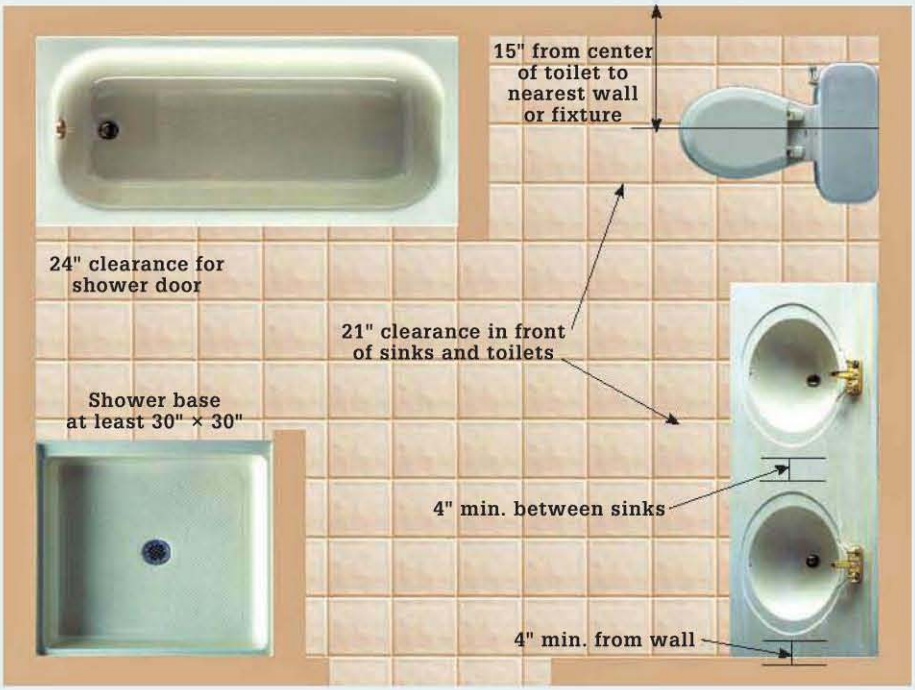
15" from center of toilet to nearest wall or fixture
24" clearance for shower door
21" clearance in front of sinks and toilets
Shower base at least 30" × 30"
4" min. between sinks
4" min. from wall

Follow minimum clearance and size guidelines when planning locations of bathroom fixtures. Easy access to fixtures is fundamental to creating a bathroom that is comfortable, safe, and easy to use.

Accessible Bathrooms (Universal Design)

The safety and accessibility of nearly all aspects of a bathroom can be improved by following universal design guidelines.

For safer floors, add a slip-resistant glaze to ceramic tile, and add nonslip adhesive strips or decals to shower floors. If you're replacing your floor, look for mosaic tiles, vinyl, and cork materials. Matte finishes tend to be less slippery than polished surfaces, and they reduce glare.

Toilets, faucets, sinks, cabinets, tubs, and showers can all be adapted or changed for increased usability by people who have experienced inflexibility or loss of strength.

Toilet height can be adjusted with the installation of a height adapter that raises a standard toilet seat 2 to 5". You can also consider an adjustable-height toilet or a model with a power-lift seat.

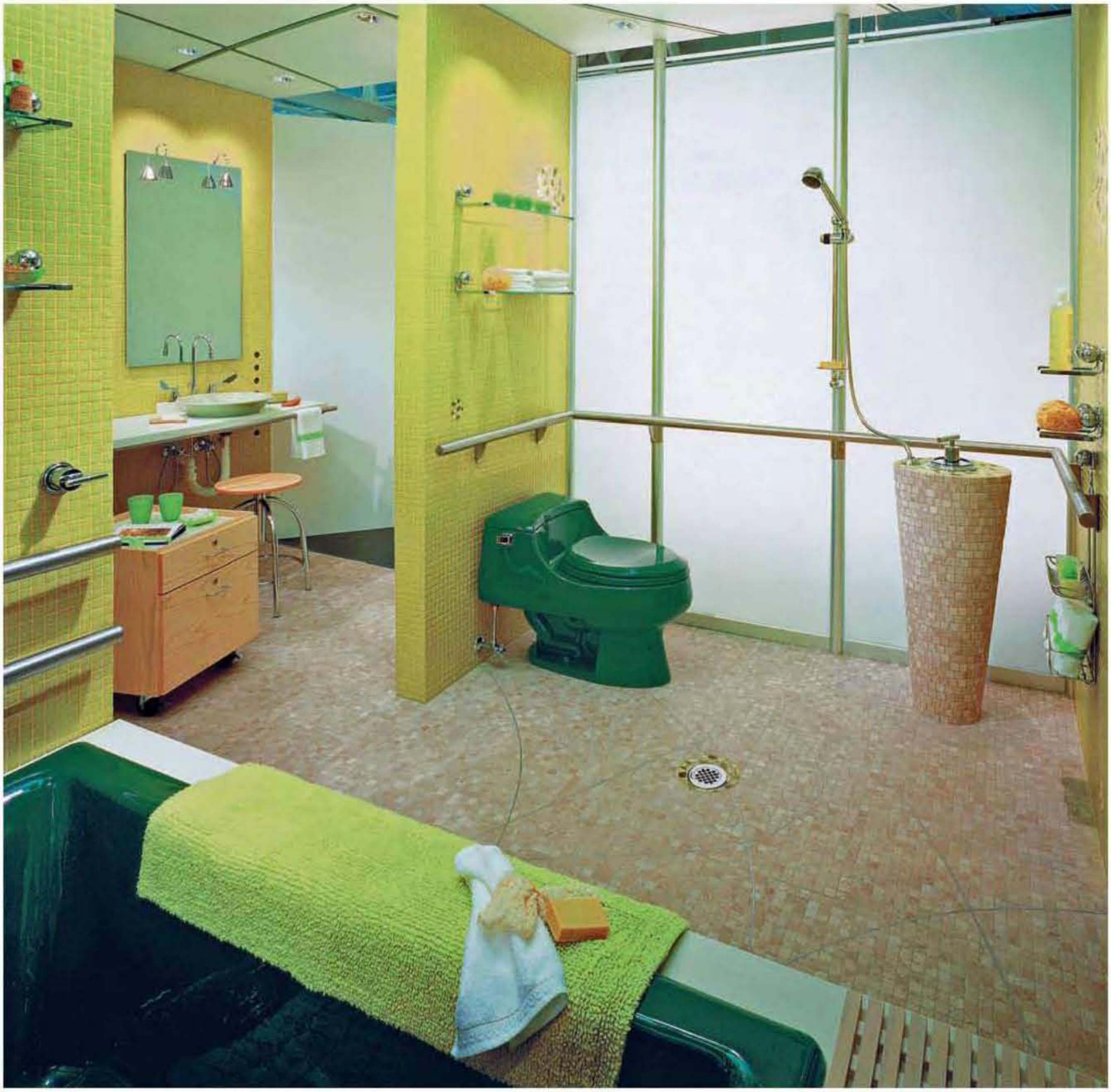

A bathroom designed for accessibility is an increasingly important element found in today's homes.

An integral personal hygiene system can be installed to help people with physical disabilities maintain independence.

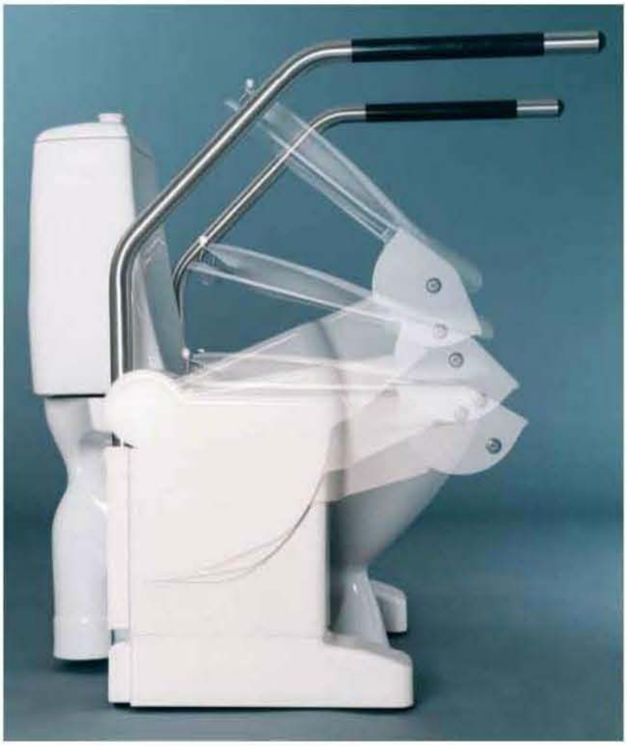

Power-lifts make toilet use easier for people with limited leg or joint strength.

Roll-out base cabinets provide the option of seating space at the countertop.

(continued)

Consider replacing your toilet with a wall-hung style that can be installed at any height, providing additional clear space for maneuvering a wheelchair or walker. To help people with physical disabilities maintain independent personal hygiene, install either a bidet or a toilet with an integral personal hygiene system. Grab bars are a must on walls around the toilet.

Install antiscald guard and volume-control devices on faucets and showerheads. Replace double-handle faucets with single-lever models, which are easier to use. A faucet with motion-sensor operation for hands-free use is a plus.

Change a standard showerhead to an adjustable showerhead mounted on a vertical slide-bar to accommodate people of all heights. Look for handheld models for best control.

Vanities and cabinets with C-shaped pulls or magnetic touch latches are easy to use. Add pull-down hardware to cabinets to bring items within reach for seated people or those with limited mobility.

Install your countertops at varied heights to serve both seated and standing users. Also consider mounting the faucet controls at the side of the sink, rather than the back, for easier access.

If you are replacing your sink, choose a style that is shallower at the front and deeper at the drain. Or install a wall-mounted sink positioned at 30 to 32". Providing a clear space under sinks and low sections of countertop allows seated users to comfortably reach the vanity. Use fold-away doors, remove face frames on base cabinets, or install roll-out base cabinets to gain clear space. Always finish the floor under fold-away or roll-out cabinets. Then insulate hot-water supply pipes or install a protective panel to prevent burns to seated users.

Side-mounted faucet and drain pulls in combination with a wall-mounted sink are easily accessible.

A tilted mirror allows seated or short users to easily see themselves.

Install grab bars in and around the shower and tub. Adding a shower seat or installing a pull-down or permanent seat in the shower allows elderly or disabled family members to sit while bathing. Install a permanently mounted shower seat at 18" high, and be sure the seat is at least 15" deep.

Water controls and faucets on tubs repositioned toward the outside edge at a height of 38 to 48" allows the water to be turned on and adjusted before getting in to bathe. If you're buying a new bathtub, consider one with a side-access door.

Consider replacing a combination tub and shower with a stand-alone shower with a floor that slopes gently toward the drain, rather than a curb to contain water. When fitted with a pull-down seat and an adjustable handheld showerhead, roll-in showers can accommodate people with a wide range of abilities.

Reverse door hinges so the doors open out. That way, the door swing won't be blocked if someone falls.

If needed, widen doorways to 32 to 36" so wheelchair and walker users can enter the bathroom easily. Or, replace a swing door with a pocket door to gain clear space.

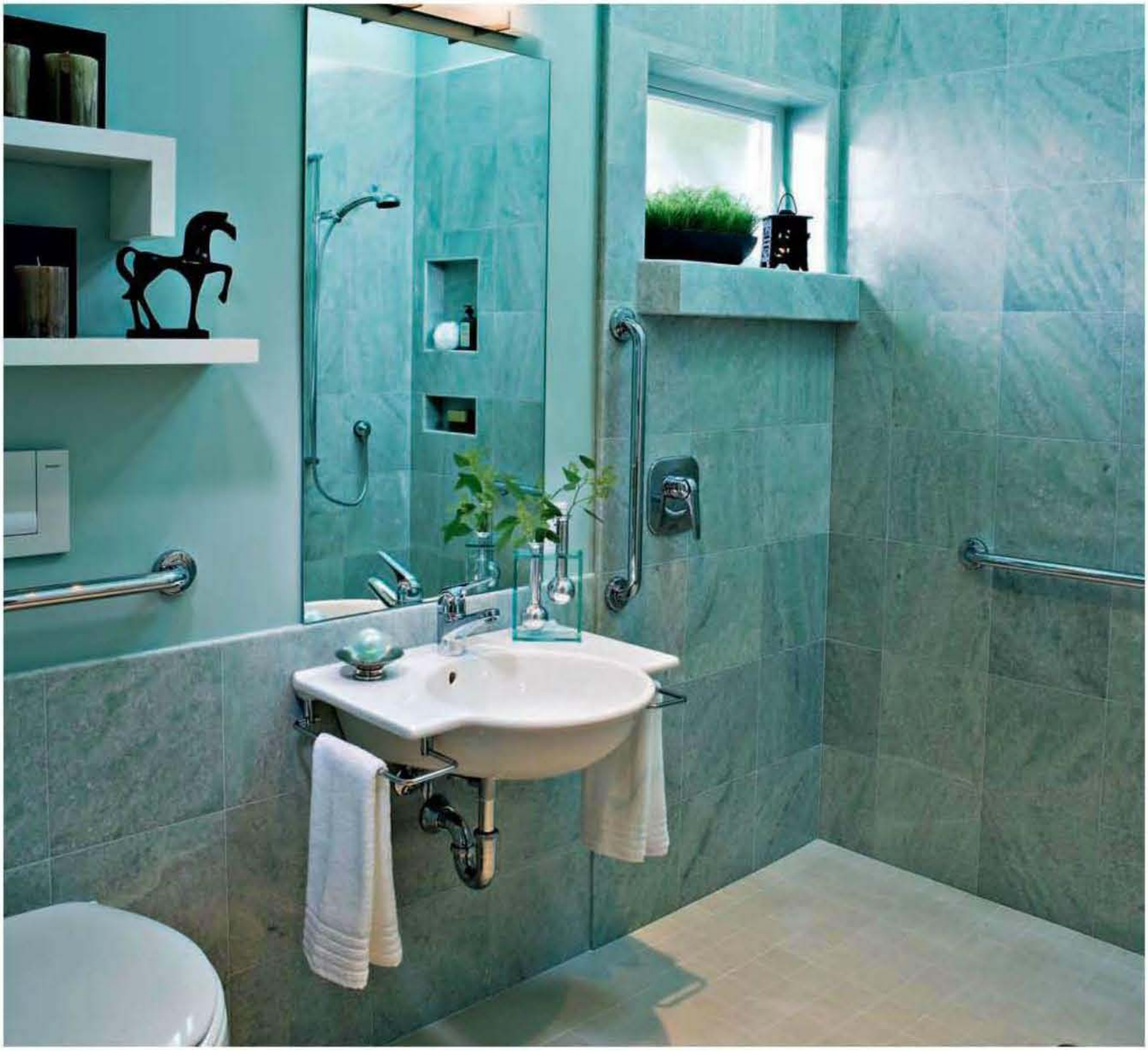

Roll-in shower designs, roll-under sinks, grab bars, adjustable slide-bar showers, and open floor space make bathing easier for people with disabilities.

(continued)

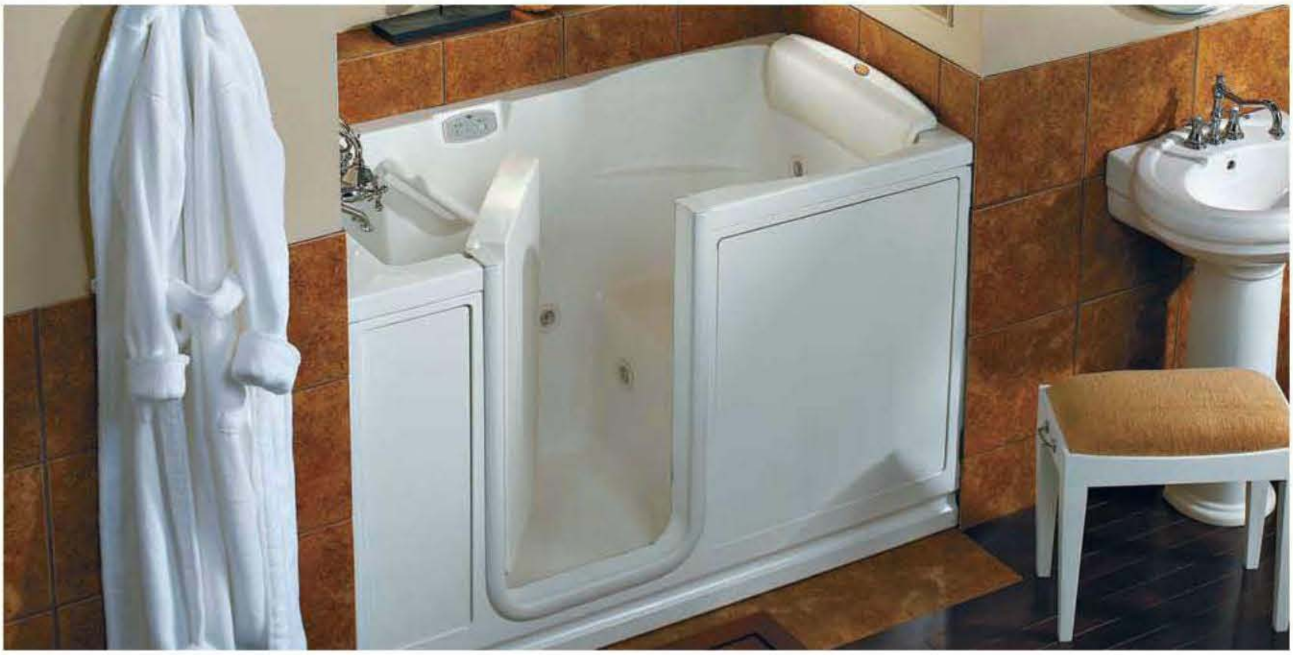

A swing-in door makes bathtubs accessible to practically anyone. The door creates a tight seal when closed and cannot be opened if the tub is full.

Install a pocket door in your bathroom to gain more clearance room for wheelchairs and walkers. Pocket doors do not require door stop molding, allowing for some additional clearance.

Recommended Clearances ▸

A bathroom should be planned with enough approach space and clearance room to allow a wheelchair or walker user to enter and turn around easily. The guidelines for approach spaces (patterned areas) and clearances shown here include some ADA guidelines and recommendations from universal design specialists.

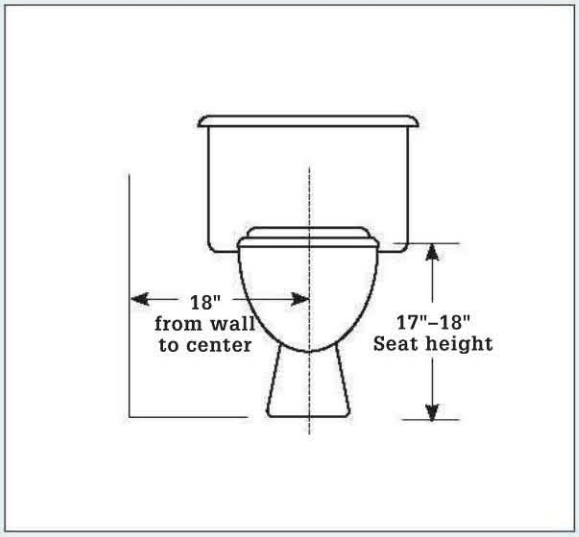
18" from wall to center
17"–18" Seat height

Toilet

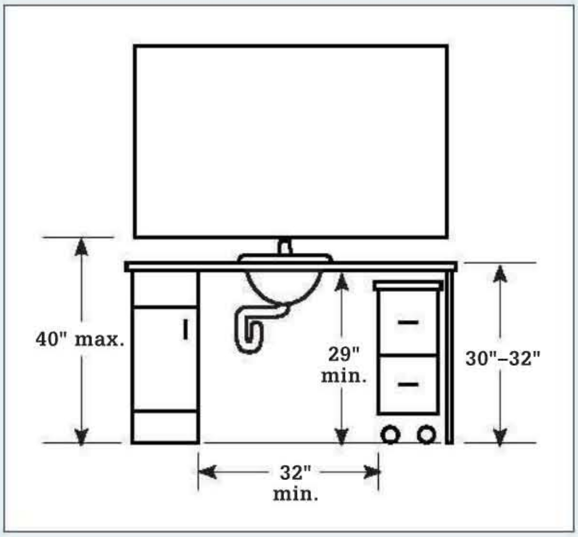
40" max.
29" min.
30"–32"
32" min.

Sink & Vanity

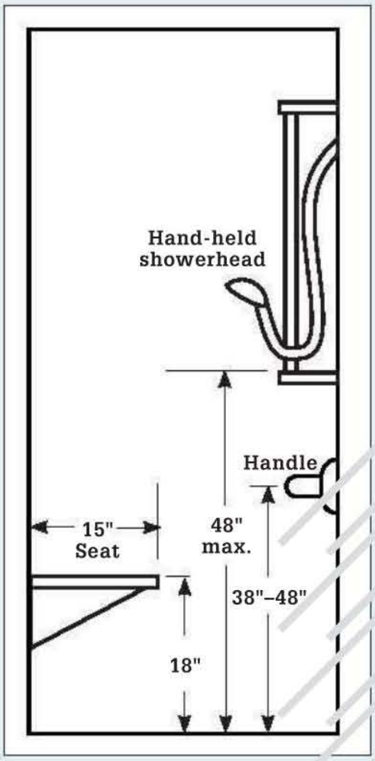
Hand-held showerhead
Handle
15" Seat
48" max.
38"–48"
18"

Shower

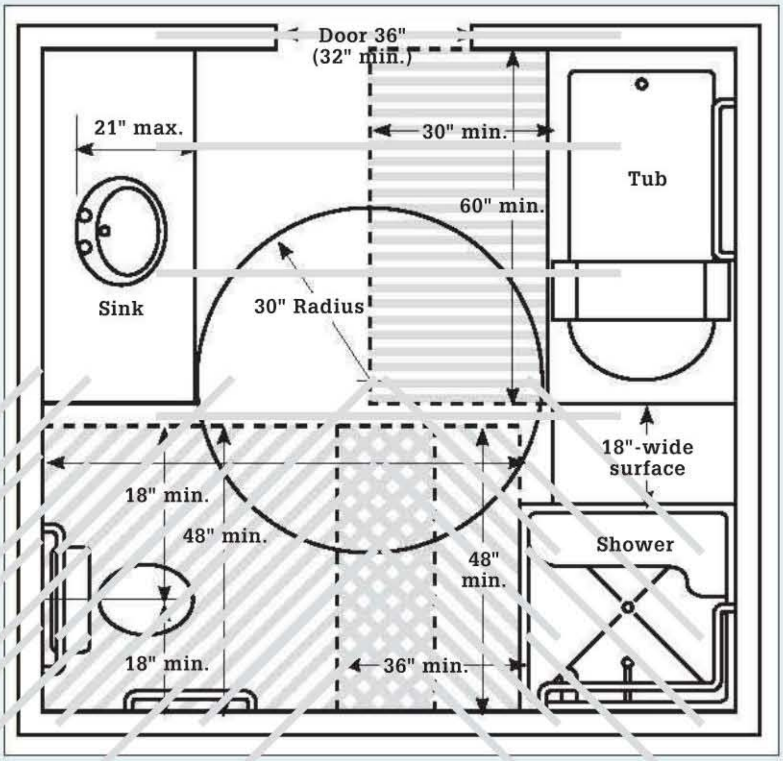
Door 36" (32" min.)
21" max.
30" min.
Tub
60" min.
Sink
30" Radius
18"-wide surface
18" min.
48" min.
48" min.
Shower
18" min.
36" min.

Floor Plan

Installing Shower Kits

The fastest and easiest way to create a new shower in your bathroom is to frame in the stall area with lumber and wallboard and then install a shower enclosure kit. Typically consisting of three fiberglass or plastic walls, these enclosure kits snap together at the corners and nestle inside the flanges of the shower pan to create nearly foolproof mechanical seals. Often, the walls are formed with shelves, soap holders, and other conveniences.

If you are on a tight budget, you can find extremely inexpensive enclosure kits to keep costs down. You can even create your own custom enclosure using waterproof beadboard panels and snap-together connectors. Or, you can invest in a higher grade kit made from thicker material that will last much longer. Some kits are sold with the receptor (and perhaps even the door) included. The kit shown here is designed to be attached directly to wall studs, but others require a backer wall for support. The panels are attached to the backer with high-tack panel adhesive.

Tools & Materials ▸

- Tape measure
- Pencil
- Hammer
- Carpenter's square
- Screwdrivers
- Pipe wrench
- Level
- Strap wrench
- Adjustable wrench
- Pliers
- Drill/driver
- Center punch
- File
- Utility knife
- Hacksaw
- Masking tape
- Silicone caulk and caulk gun
- Shower enclosure kit
- Shower door
- Showerhead
- Faucet
- Plumbing supplies
- Panel adhesive
- Spud wrench
- Large-head roofing nails
- Jigsaw
- Duct tape
- Miter box

A paneled shower surround is inexpensive and easy to install. Designed for alcove installations, they often are sold with matching shower pans (called *receptors*).

How to Install a Shower Enclosure

1

Remove the wallcovering and floor covering in the installation area. Dispose of the materials immediately and thoroughly clean the area. *Note: Most kits can be installed over wallboard, but more professional-looking results can be achieved if it is removed.*

2

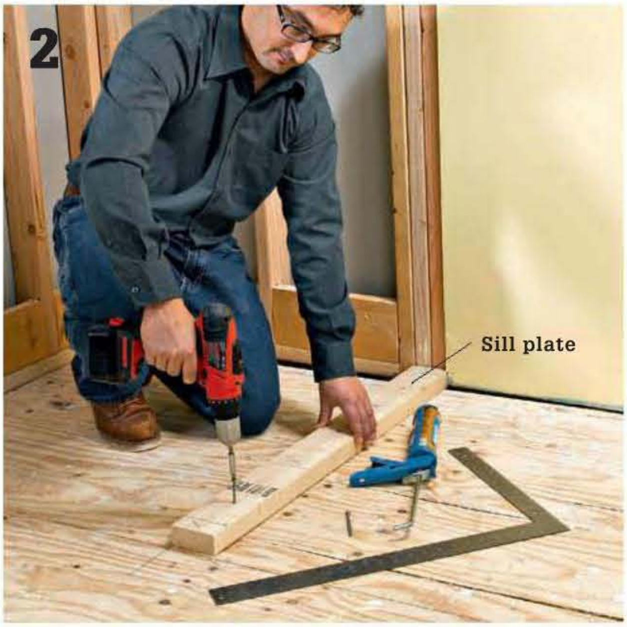

If you are adding a wall to create the alcove, lay out the locations for the studs and plumbing on the new wood sill plate. Also lay out the stud locations on the cap plate that will be attached to the ceiling. Refer to the enclosure kit instructions for exact locations and dimensions of studs. Attach the sill plate to the floor with deck screws and panel adhesive, making sure it is square to the back wall and the correct distance from the side wall.

3

Align a straight 2 × 4 right next to the sill plate and make a mark on the ceiling. Use a level to extend that line directly above the sill plate. Attach the cap plate at that point.

4

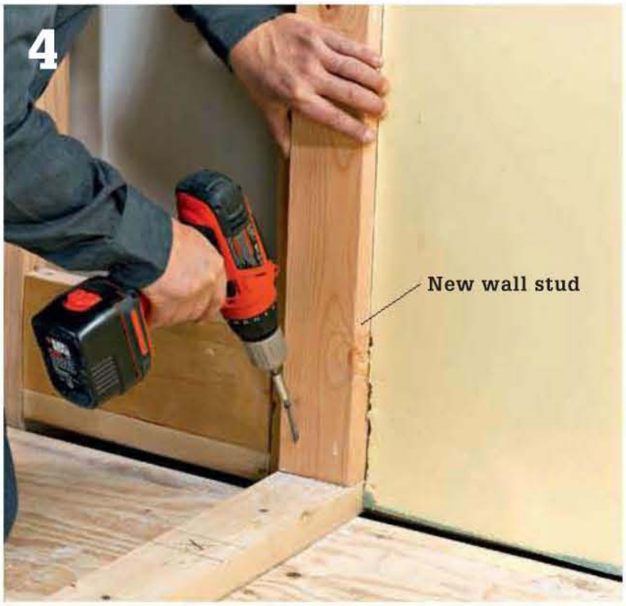

Install the 2 × 4 studs at the outlined locations. Check with a level to make sure each stud is plumb, and then attach them by driving deck screws toenail style into the sill plate and cap plate.

(continued)

5

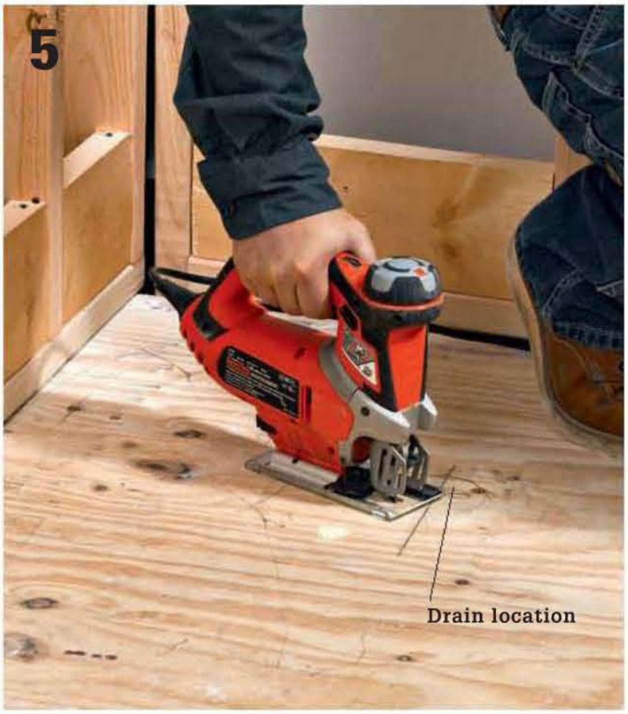

Cut an access hole in the floor for the drain, according to the installation manual instructions. Drill openings in the sill plate of the wet wall (the new wall in this project) for the supply pipes, also according to the instructions.

6

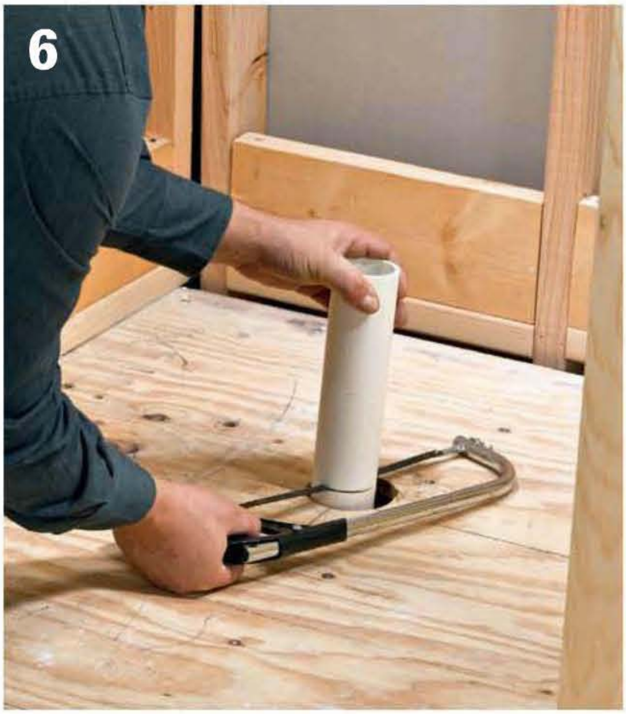

Install a drain pipe and branch line and then trim the drain pipe flush with the floor. If you are not experienced with plumbing, hire a plumber to install the new drain line.

7

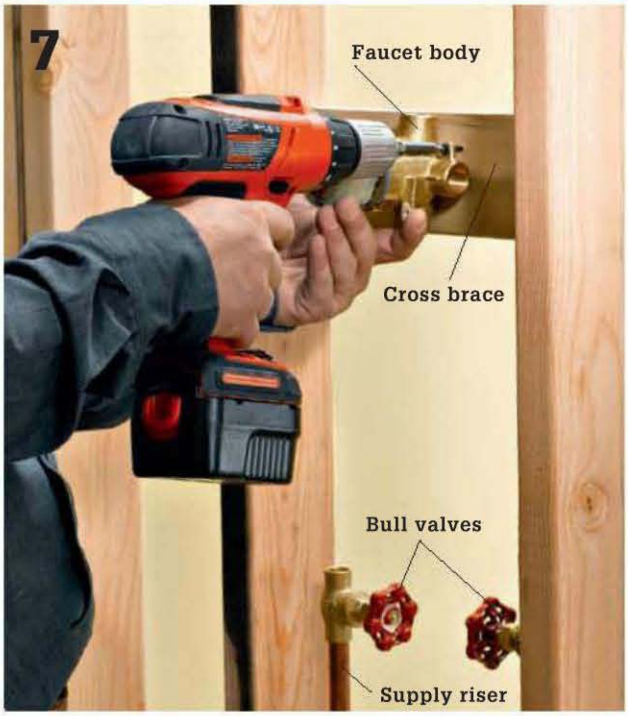

Install new supply risers as directed in the instruction manual (again, have a plumber do this if necessary). Also install cross braces between the studs in the wet wall for mounting the faucet body and shower arm.

8

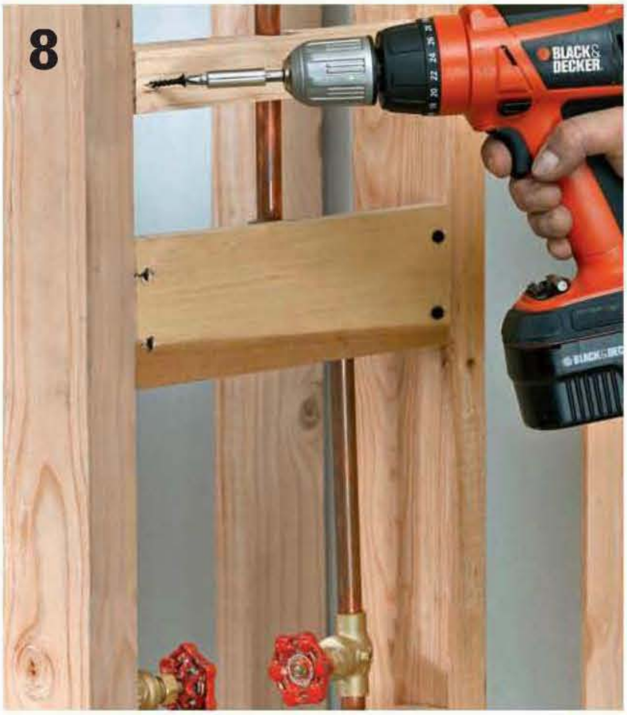

If the supply plumbing is located in a wall (old or new) that is accessible from the non-shower side, install framing for a removable access panel.

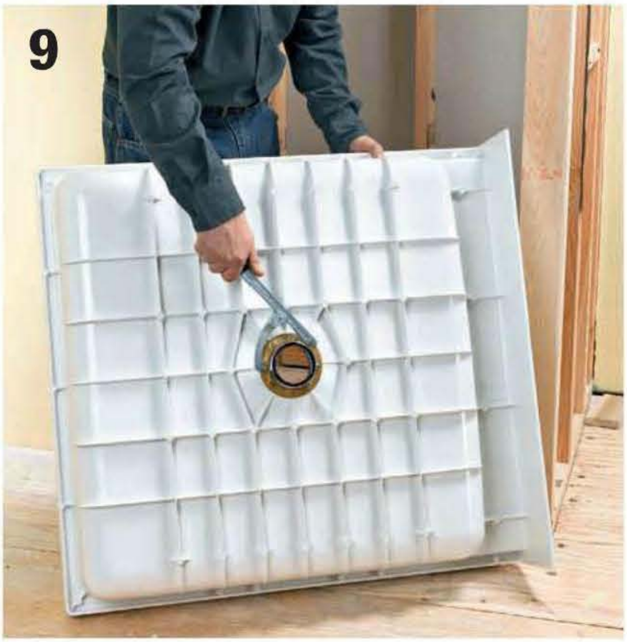

Attach the drain tailpiece that came with your receptor to the underside of the unit, following the manufacturer's instructions precisely. Here, an adjustable spud wrench is being used to tighten the tailpiece.

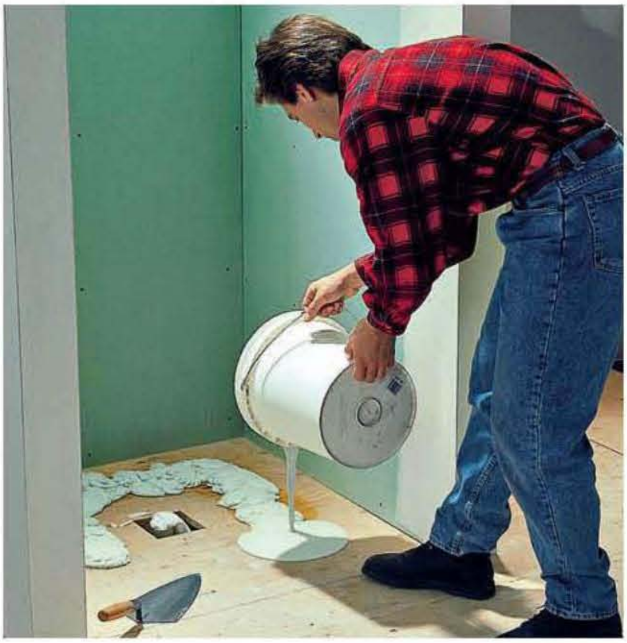

Option: To stabilize the receptor, especially if the floor is uneven, pour or trowel a layer of thinset mortar into the installation area, taking care to keep the mortar out of the drain access hole. Do not apply mortar in areas where the receptor has feet that are intended to make full contact with the floor.

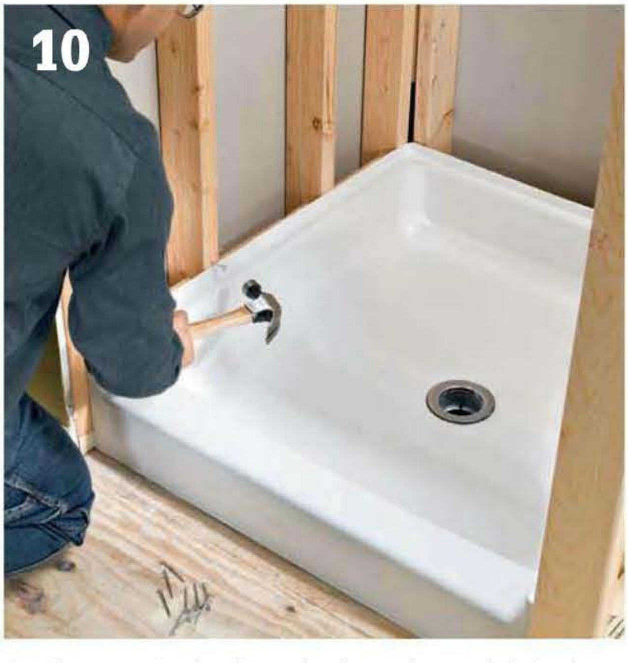

Set the receptor in place, check to make sure it is level, and shim it if necessary. Secure the receptor with large-head roofing nails driven into the wall stud so the nailheads pin the flange against the stud. Do not overdrive the nails.

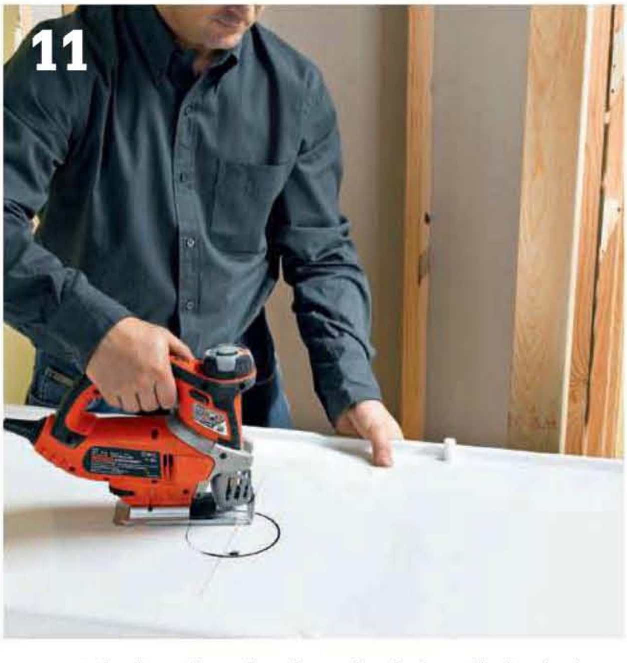

Lay out the locations for the valve hole or holes in the end wall panel that will be installed on the wet wall. Check your installation instructions. Some kits come with a template marked on the packaging carton. Cut the access hole with a hole saw and drill or with a jigsaw and fine-tooth blade. If using a jigsaw, orient the panel so the finish surface is facing down.

(continued)

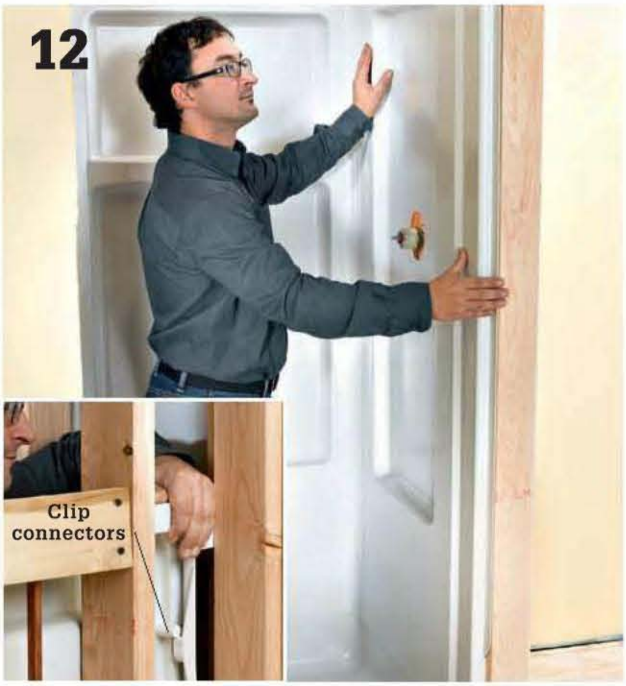

Position the back wall so there is a slight gap (about 1⁄32") between the bottom of the panel and the rim of the receptor—set a few small spacers on the rim if need be. Tack a pair of roofing nails above the top of the back panel to hold it in place (or, use duct tape). Position both end walls and test the fits. Make clip connections between panels (inset) if your kit uses them.

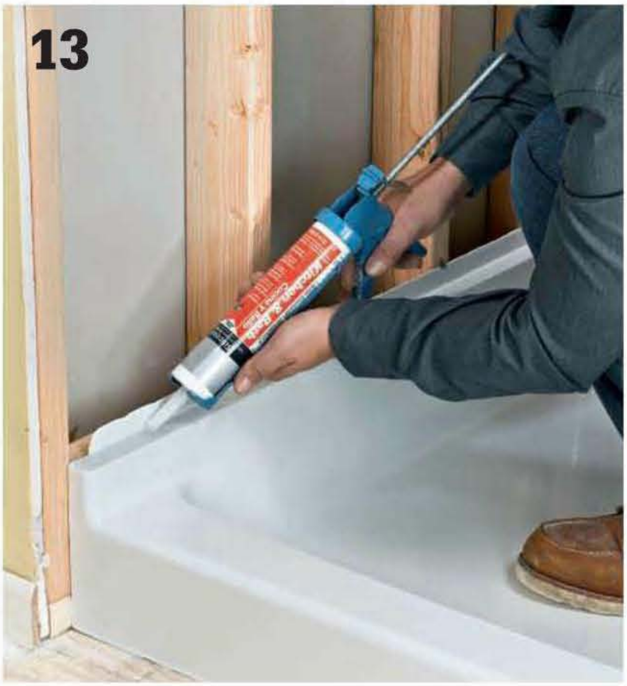

Remove the end walls so you can prepare the installation area for them. If your kit recommends panel adhesive, apply it to the wall or studs. In the kit shown here, only a small bead of silicone sealant on the receptor flange is required.

Reinstall the end panels, permanently clipping them to the back panel according to the kit manufacturer's instructions. Make sure the front edges of the end panels are flush with the front of the receptor.

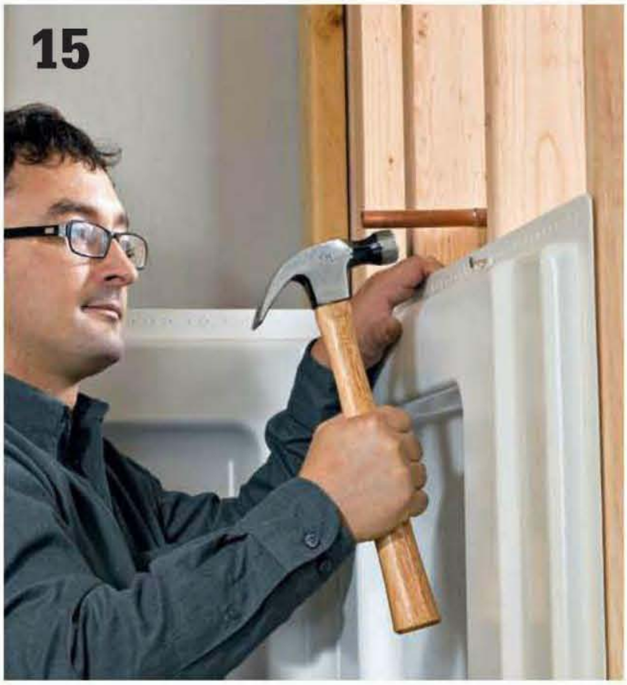

Once the panels are positioned correctly and snapped together, fasten them to the wall studs. If the panels have predrilled nail holes, drive roofing nails through them at each stud at the panel tops and every 4" to 6" along vertical surfaces.

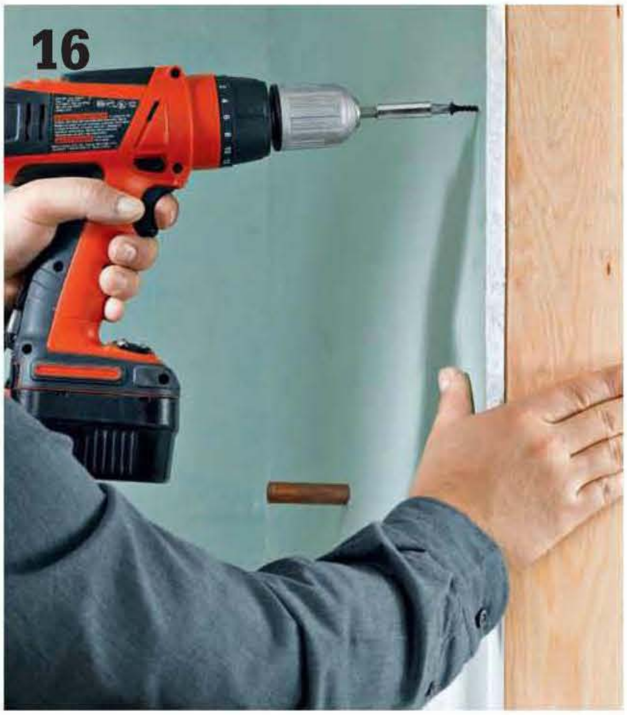

Install wallcovering material above the enclosure panels and anywhere else it is needed. Use moisture-resistant materials, and maintain a gap of ¼" between the shoulders of the top panel flanges and the wallcovering.

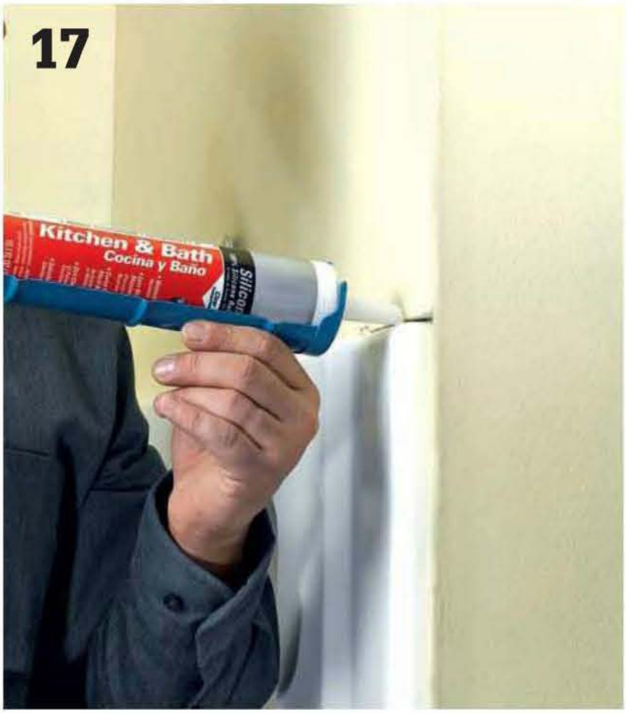

Finish the walls and then caulk between the enclosure panels and the wallcoverings with silicone caulk.

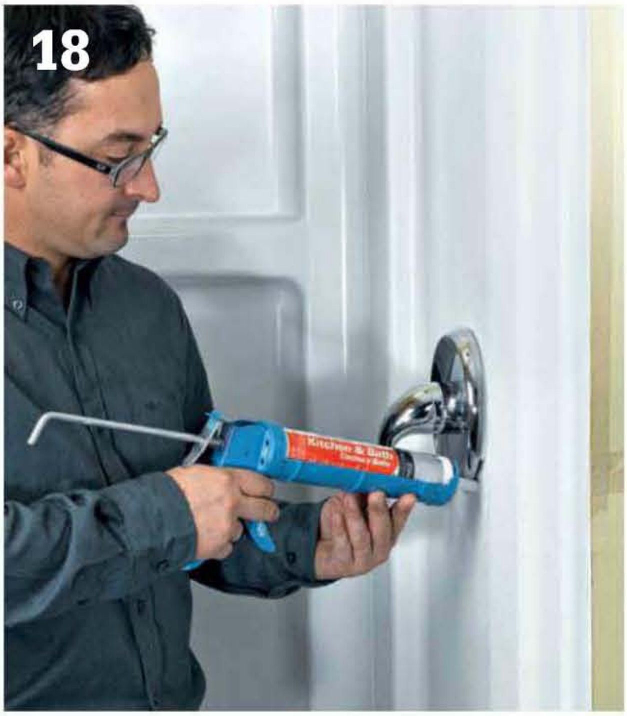

Install the faucet handles and escutcheon and caulk around the escutcheon plate. Install the shower arm escutcheon and showerhead.

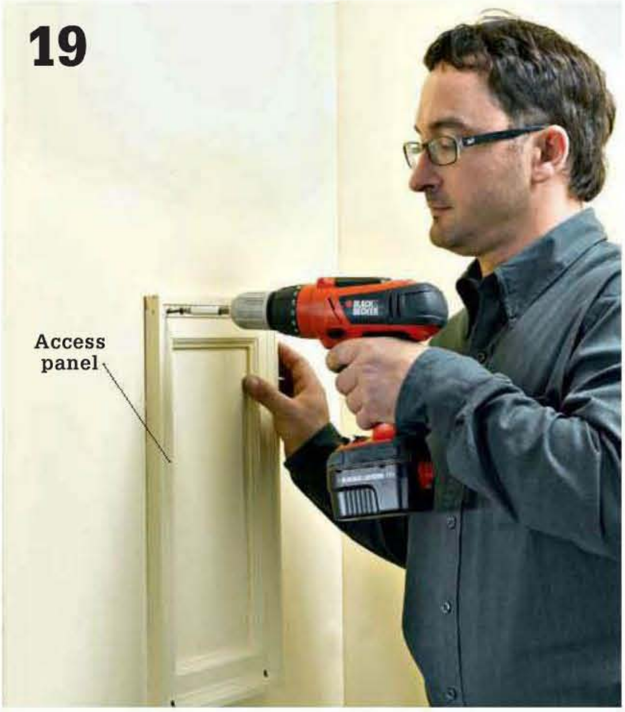

Make an access panel and attach it at the framed opening created in step 8. A piece of ¼" plywood framed with mitered case molding and painted to match the wall is one idea for access panel covers.

How to Install a Hinged Shower Door

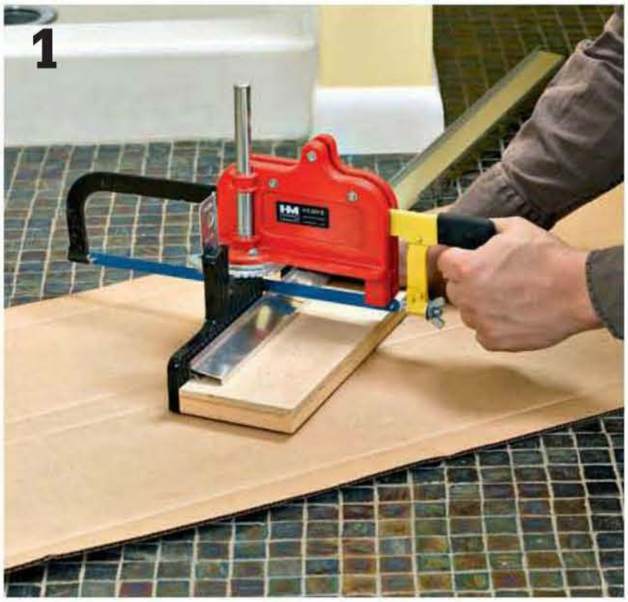

Measure the width of the shower opening. If the walls of the shower slope inward slightly before meeting the base, take your measurement from a higher point at the full width of the opening so you don't cut the door base too short. Cut the base piece to fit using a hacksaw and a miter box. File the cut ends if necessary to deburr them.

Identify which side jamb will be the hinge jamb and which will be the strike jamb according to the direction you want your hinged door to swing—an outward swing is preferred. Prepare the jambs for installation as directed in your instructions.

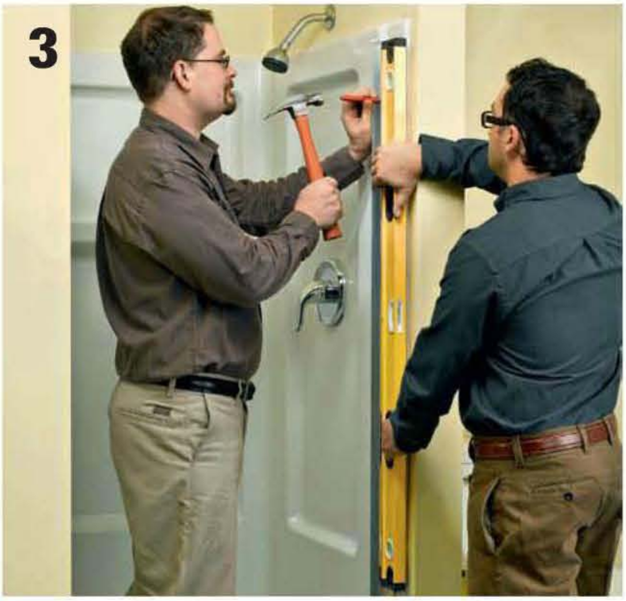

Place the base jamb on the curb of the shower base. If the joint where the wall meets the curb is sloped, you'll need to trim the corners of the base piece to follow the profile. Place a jamb carefully onto the base and plumb it with a level. Then, mark a drilling point by tapping a centerpunch in the middle of each nail hole in each jamb. Remove the jambs, drill pilot holes, and then attach the jambs with the provided screws.

Remove the bottom track and prepare the shower base curb for installation of the base track, following the manufacturer's directions. Permanently install the bottom track. Bottom tracks (not all doors have them) are usually attached to the side jambs or held in place with adhesive. Never use fasteners to secure them to curb.

5

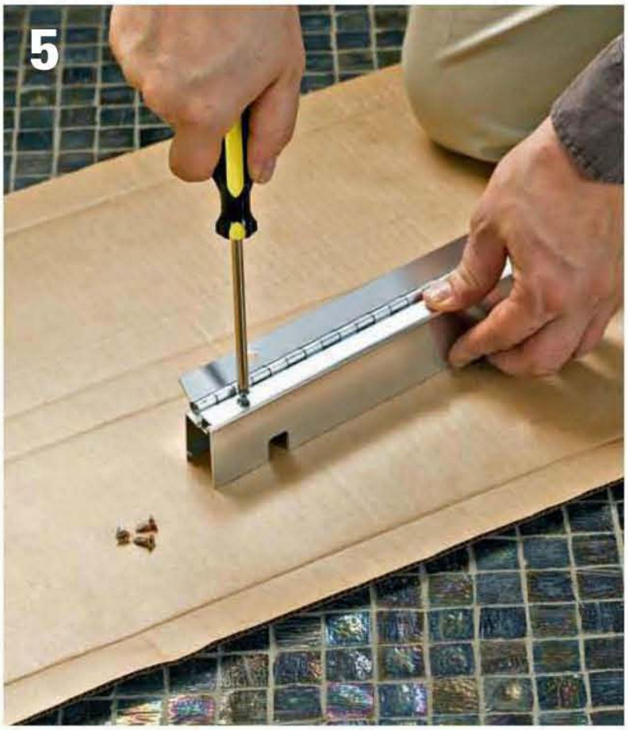

Working on the floor or another flat surface, attach the door hinge to the hinge jamb, if required. In most systems, the hinge is fitted over the hinge jamb after you attach it to the wall.

6

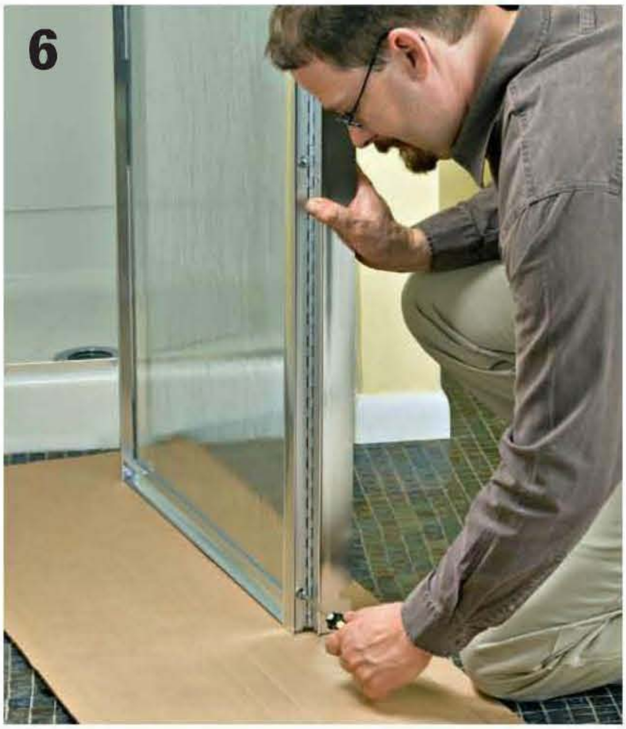

Attach the hinge to the door panel, according to the manufacturer's instructions. Attach any cap fitting that keeps water out of the jamb.

7

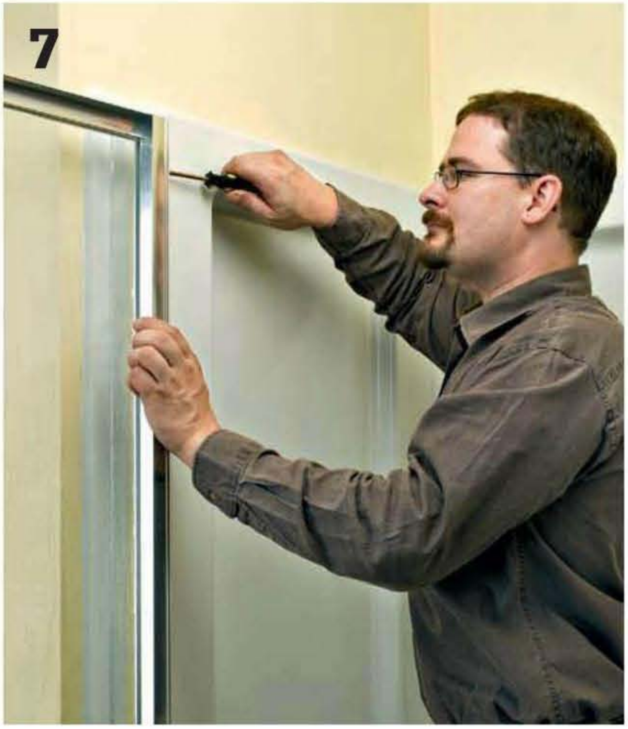

Fit the hinge jamb over the side jamb and adjust it as directed in your instruction manual. Once the clearances are correct, fasten the jambs to hang the door.

8

Install the magnetic strike plate and any remaining caps or accessories such as towel rods. Also attach the sweep that seals the passage, if provided.

Installing Alcove Bathtubs

Many of our homes are equipped with an alcove tub that includes a tub surround and shower feature. By combining the tub and the shower in one fixture, you conserve precious bathroom floorspace and simplify the initial installation. Plus, you only have one bathing fixture that needs cleaning.

But because tub/showers are so efficient, they do get a lot of use and tend to have fairly limited lifespans. The fact that the most inexpensive tubs on the market are designed for alcove use also reduces the average tub/shower lifespan. Pressed steel tubs have enamel finishes that crack and craze; plastic and fiberglass tubs get grimy and stained; even acrylic and composite tubs show wear eventually (and as with other fixtures, styles, and colors change too).

Plumbing an alcove tub is a relatively difficult job because getting access to the drain lines attached to the tub and into the floor is often very awkward. Although an access panel is required by most codes, the truth is that many tubs were installed without them or with panels that are too small or hard to reach to be of much use. If you are contemplating replacing your tub, the first step in the decision process should be to find the access panel and determine if it is sufficient. If it is not (or there is no panel at all), consider how you might enlarge it. Often, this means cutting a hole in the wall on the adjoining room and also in the ceiling below. This creates more work, of course, but compared to the damage caused by a leaky drain from a subpar installation, making an access opening is a small inconvenience.

Tools & Materials ▸

- Channel-type pliers
- Hacksaw
- Carpenter's level
- Pencil
- Tape measure
- Saw
- Screwdriver
- Drill
- Adjustable wrench
- Trowel
- Shims
- Galvanized deck screws
- Drain-waste-overflow kit
- 1 × 3, 1 × 4, 2 × 4 lumber
- Galvanized roofing nails
- Galvanized roof flashing
- Thinset mortar
- Tub & tile caulk
- Propane torch

By replacing a dingy old alcove tub with a fresh new one, you can make the tub and shower area as pleasant to use as it is efficient.

Tips for Installing Bathtubs ▸

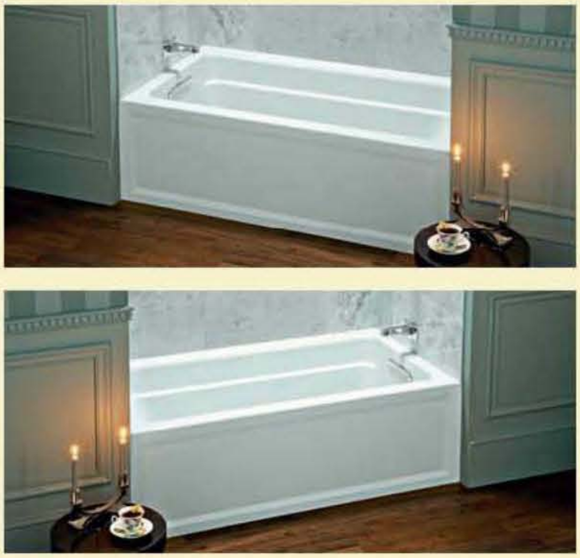

Choose the correct tub for your plumbing setup. Alcove-installed tubs with only one-sided aprons are sold as either "left-hand" or "right-hand" models, depending on the location of the predrilled drain and overflow holes in the tub. To determine which type you need, face into the alcove and check whether the tub drain is on your right or your left.

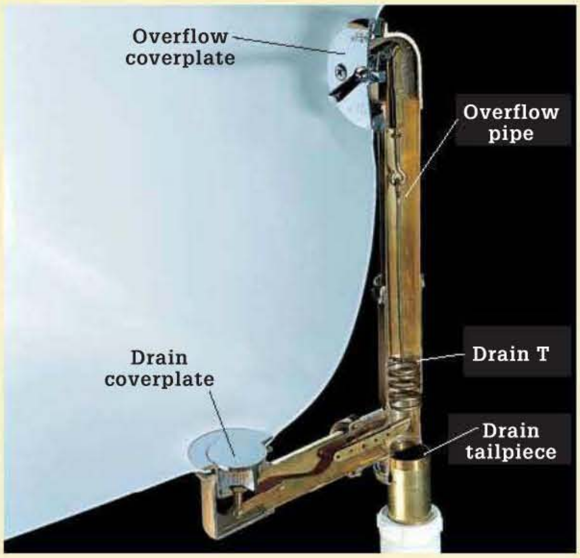
Overflow coverplate
Overflow pipe
Drain coverplate
Drain T
Drain tailpiece

A Drain-waste-overflow kit with stopper mechanism must be purchased separately and attached after the tub is set. Available in both brass and plastic types, most kits include an overflow coverplate, an overflow pipe that can be adjusted to different heights, a drain T-fitting, an adjustable drain tailpiece, and a drain coverplate that screws into the tailpiece.

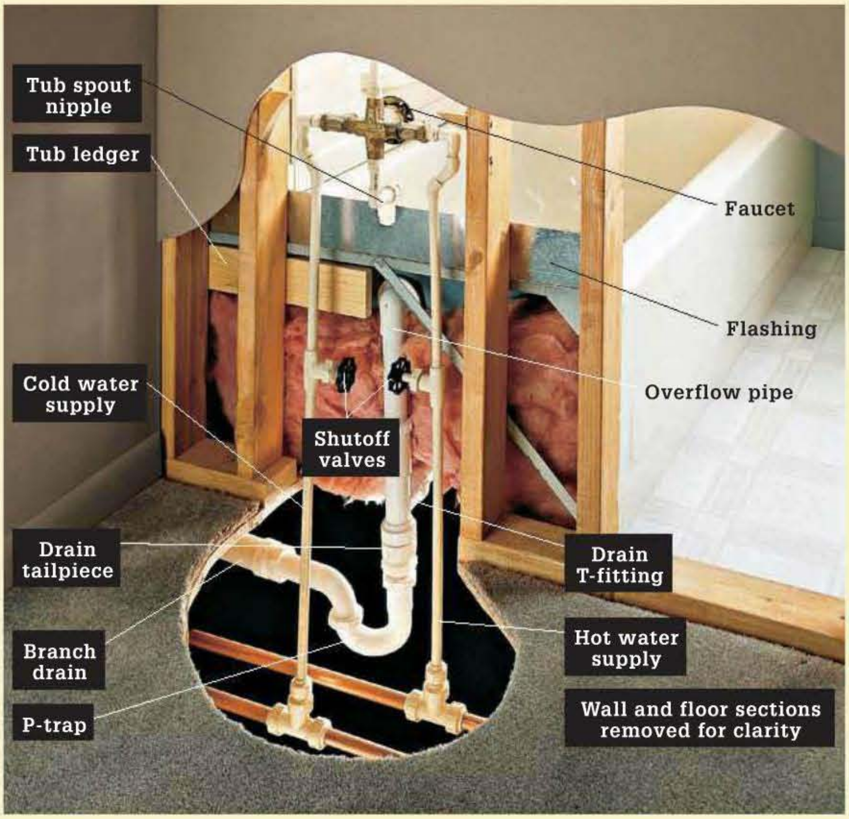
Tub spout nipple
Tub ledger
Faucet
Flashing
Cold water supply
Overflow pipe
Shutoff valves
Drain tailpiece
Drain T-fitting
Branch drain
Hot water supply
P-trap
Wall and floor sections removed for clarity

The supply system for a bathtub includes hot and cold supply pipes, shutoff valves, a faucet and handle(s), and a spout. Supply connections can be made before or after the tub is installed.

The drain-waste-overflow system for a bathtub includes the overflow pipe, drain T, P-trap, and branch drain. The overflow pipe assembly is attached to the tub before installation.

How To Remove an Alcove Bathtub

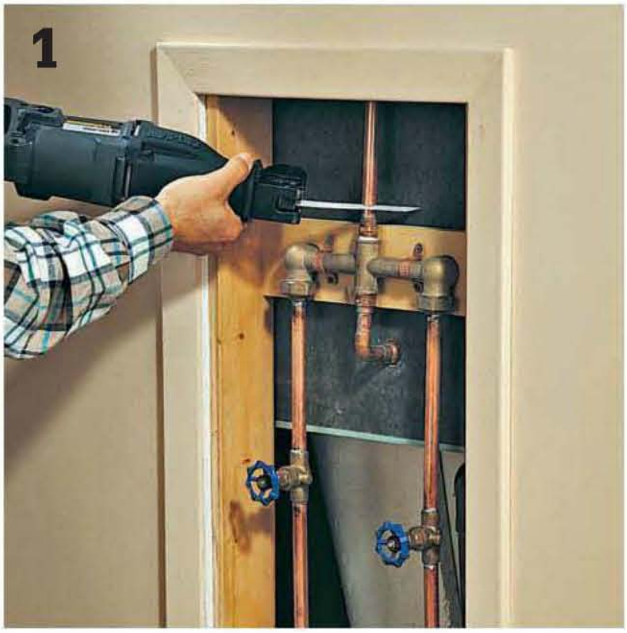

Cut the old supply tubes, if you have access to them, with a reciprocating saw and metal cutting blade or with a hacksaw. Be sure to shut off the water supply at the stop valves first. Cut the shower pipe just above the faucet body and cut the supply tubes just above the stop valves.

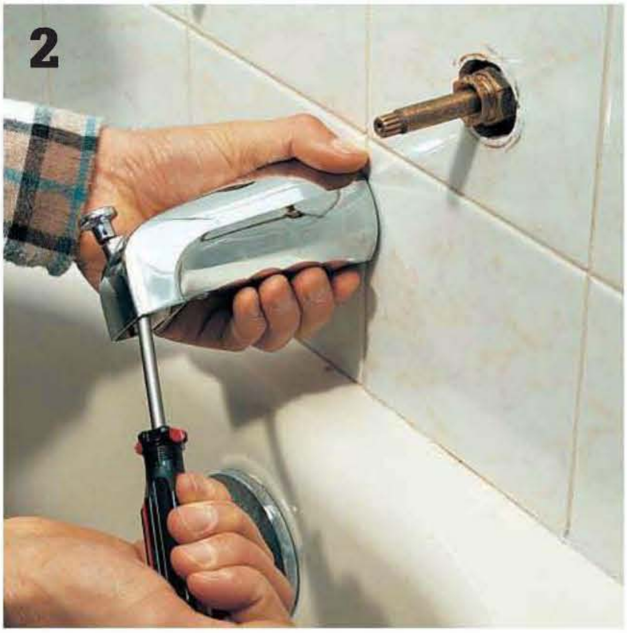

Remove the faucet handles, tub spout, shower head and escutcheon, and arm. For the spout, check the underside for a set screw and loosen it if you find one. Then, insert a long screwdriver into the spout and turn the spout counterclockwise.

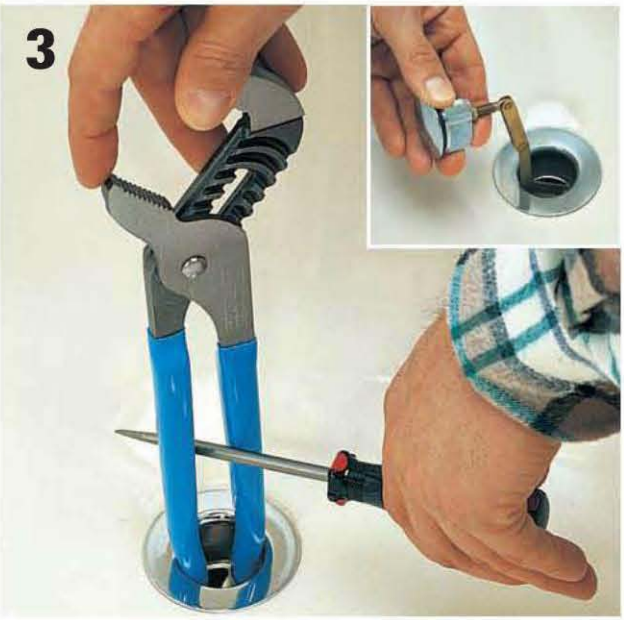

Remove the drain plug, working from the tub side. If the tub has a pop-up drain with linkage, twist the plug to disengage the linkage and remove the plug (inset). Then, insert the handles of a channel-type pliers into the drain opening and past the drain crosspiece. Twist the pliers counterclockwise to remove the plug.

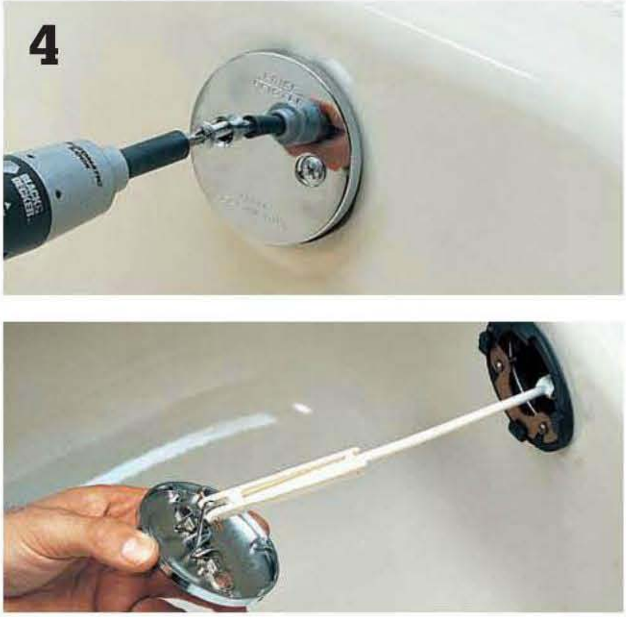

Remove the overflow coverplate (top photo) and then withdraw the pop-up drain linkage through the overflow opening (lower photo).

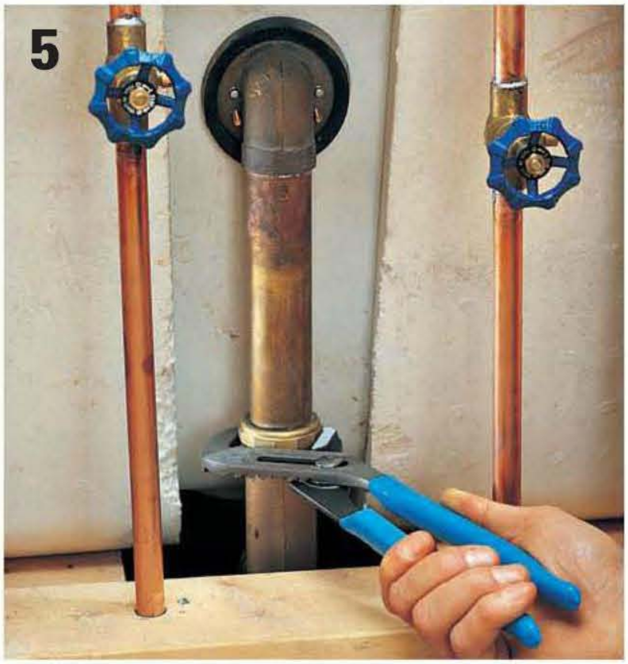

Disconnect the overflow pipe from the drain assembly and remove both parts (your access may not be as unrestricted as seen here). If you need to cut the pipes, go ahead and do it. In most cases, it is difficult to maneuver the tub out with the DWO assembly still attached.

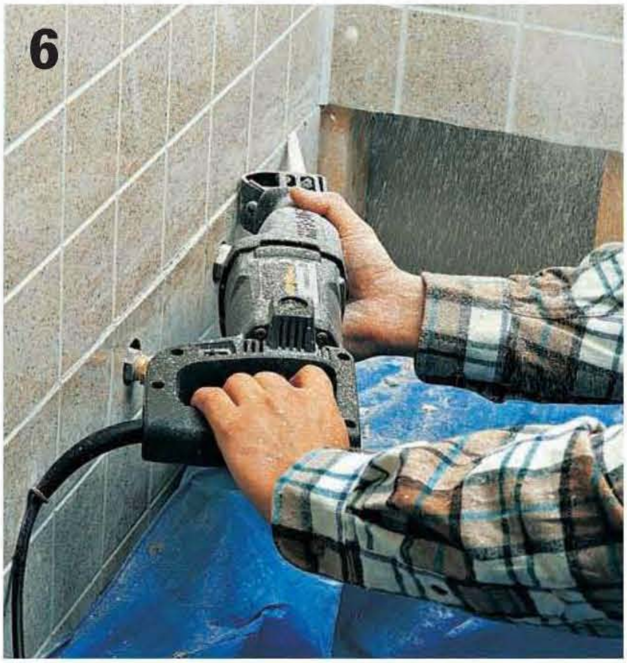

Cut the wall to a line about 6" above the tub rim. Alcove tubs are fastened to the wall studs with nails driven through or above a flange that extends up from the rim. You'll need to remove a bit of the wall covering so you can remove the fasteners.

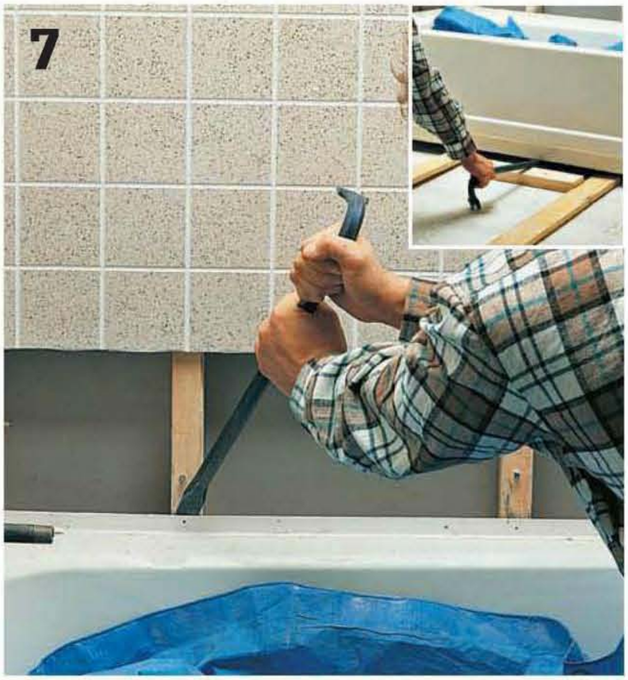

If you can, pry out fasteners and then pull the tub away from the walls by levering between the back rim of the tub and the back wall of the alcove. If it resists, check for adhesive caulk or even flooring blocking the bottom of the apron. If needed, raise the tub and slide a pair of 1 × 4 runners under the skirt edge (inset photo) to make it easier to slide out.

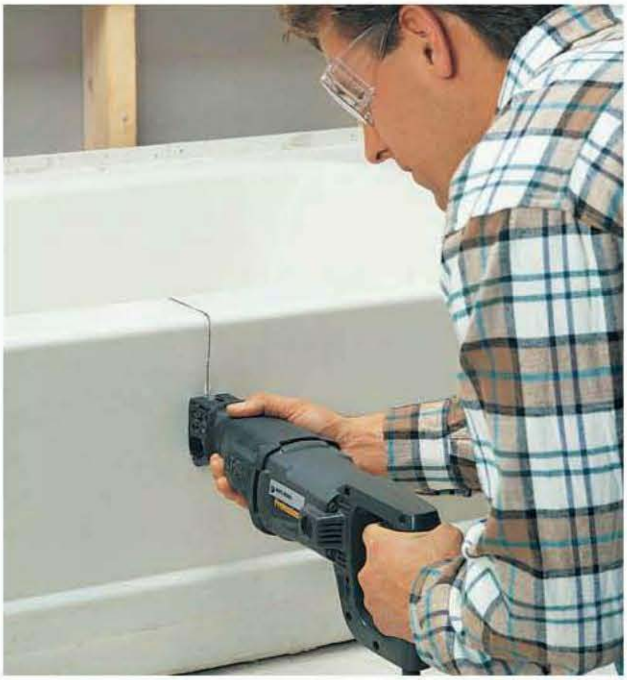

Option: Cut stubborn tubs in half to wrangle them out of the alcove. This has the added benefit of making the tubs easier to get out the door, down the stairs, and into the dumpster.

How to Install a New Alcove Tub

1

Prepare for the new tub. Inspect and remove old or deteriorated wall surfaces or framing members in the tub area. With today's mold-resistant wallboard products, it makes extra sense to go ahead and strip off the old alcove wallboard down to the studs so you can replace them. This also allows you to inspect for hidden damage in the wall and ceiling cavities.

2

Check the subfloor for level—if it is not level, use pour-on floor leveler compound to correct it (ask at your local flooring store). Make sure the supply and drain pipes and the shutoff valves are in good repair and correct any problems you encounter. If you have no bath fan in the alcove, now is the perfect time to add one.

3

Check the height of the crossbraces for the faucet body and the showerhead. If your family members needed to stoop to use the old shower, consider raising the brace for the showerhead. Read the instructions for your new faucet/diverter and check to see that the brace for the faucet body will conform to the requirements (this includes distance from the surround wall as well as height). Adjust the brace locations as needed.

4

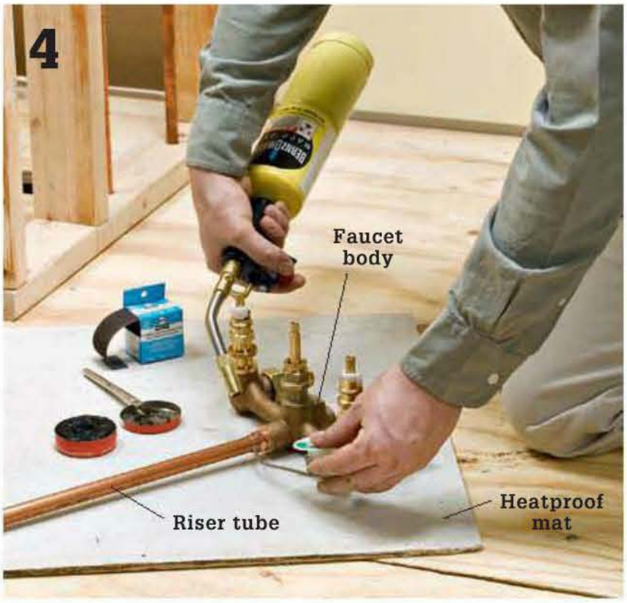

Begin by installing the new water supply plumbing. Measure to determine the required height of your shower riser tube and cut it to length. Solder the bottom of the riser to the faucet body and the top to the shower elbow.

5

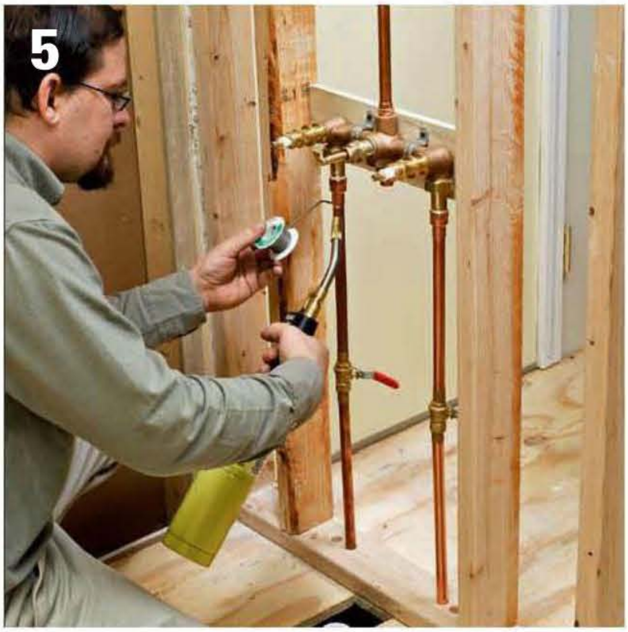

Attach the faucet body to the cross brace with pipe hanger straps. Then, attach supply tubing from the stop valves to the faucet body, making sure to attach the hot water to the left port and cold to the right port. Also secure the shower elbow to its cross brace with a pipe strap. Do not attach the shower arm yet.

6

Slide the bathtub into the alcove. Make sure tub is flat on the floor and pressed flush against the back wall. If your tub did not come with a tub protector, cut a piece of cardboard to line the tub bottom, and tape pieces of cardboard around the rim to protect the finish from shoes and dropped tools.

7

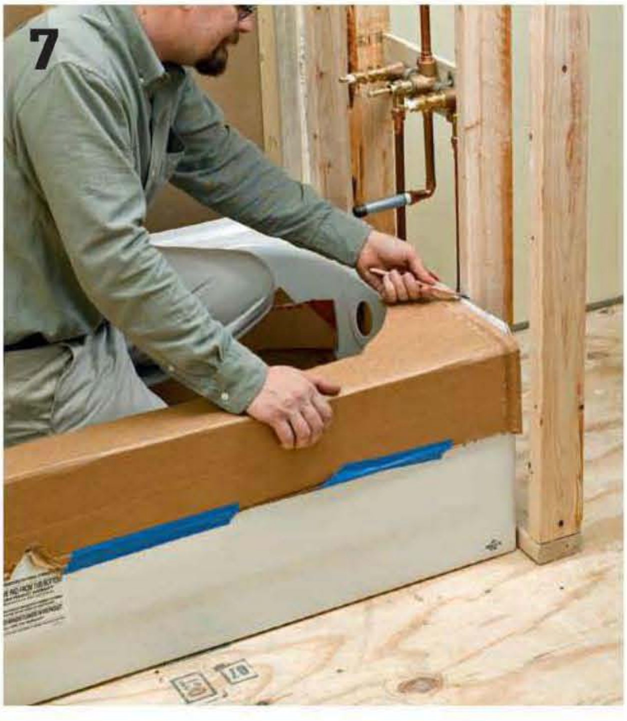

Mark locations for ledger boards. To do this, trace the height of the top of the tub's nailing flange onto the wall studs in the alcove. Then remove the tub and measure the height of the nailing flange. Measure down this same amount from your flange lines and mark new ledger board location.

8

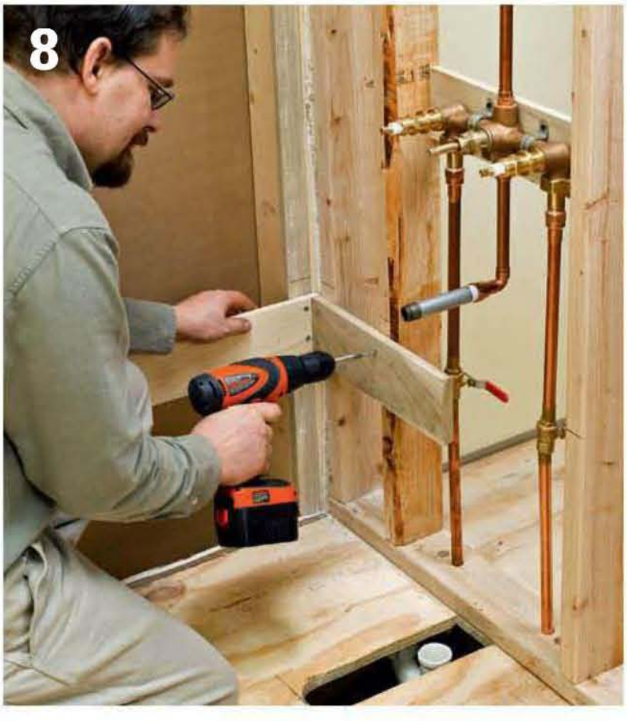

Install 1 × 4 ledger boards. Drive two or three 3"-galvanized deck screws through the ledger board at each stud. All three walls should receive a ledger. Leave an open space in the wet wall to allow clearance for the DWO kit.

(continued)

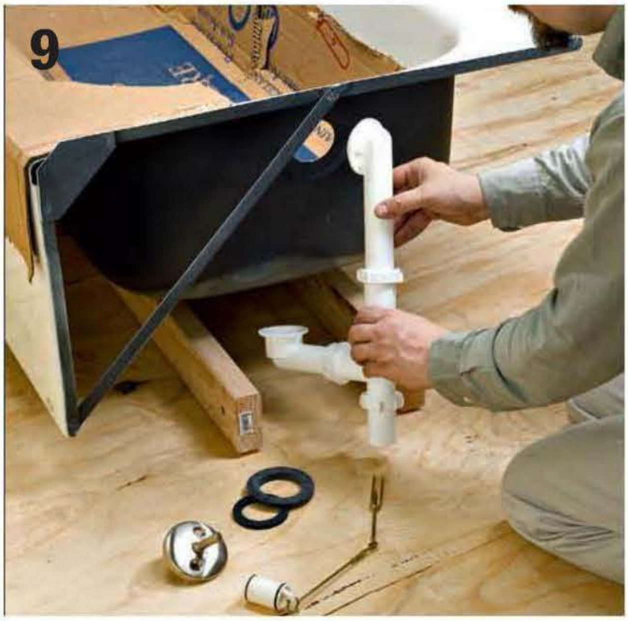

Install the drain-waste-overflow (DWO) pipes before you install the tub. Make sure to get a good seal on the slip nuts at the pipe joints. Follow the manufacturer's instructions to make sure the pop-up drain linkage is connected properly. Make sure rubber gaskets are positioned correctly at the openings on the outside of the tub.

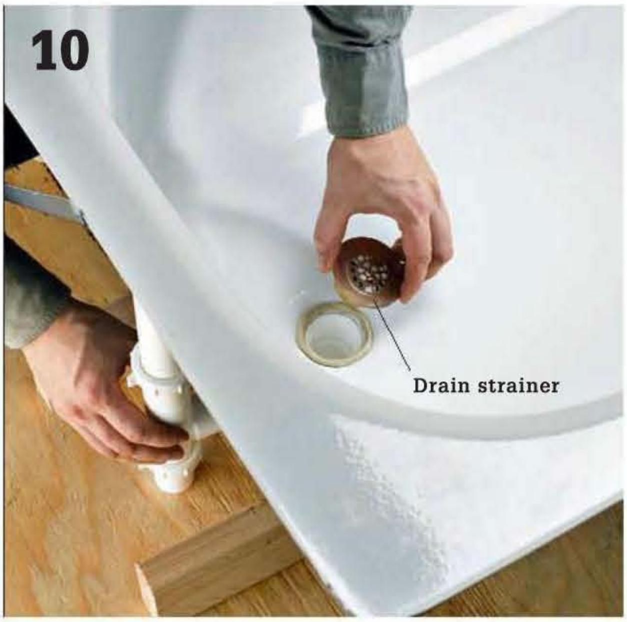

Thread the male-threaded drain strainer into the female-threaded drain waste elbow. Wrap a coil of plumber's putty around the drain outlet underneath the plug rim first. Hand tighten only.

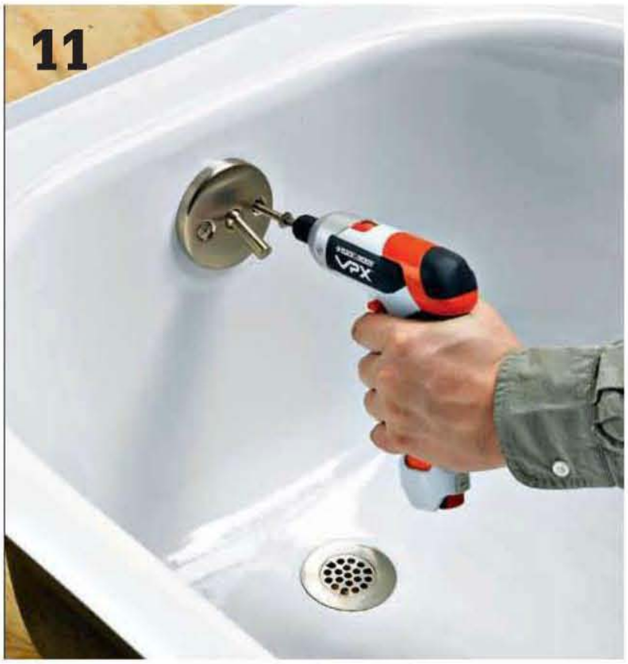

Attach the overflow coverplate, making sure the pop-up drain controls are in the correct position. Tighten the mounting screws that connect to the mounting plate to sandwich the rubber gasket snugly between the overflow pipe flange and the tub wall. Then, finish tightening the drain strainer against the waste elbow by inserting the handle of a pair of pliers into the strainer body and turning.

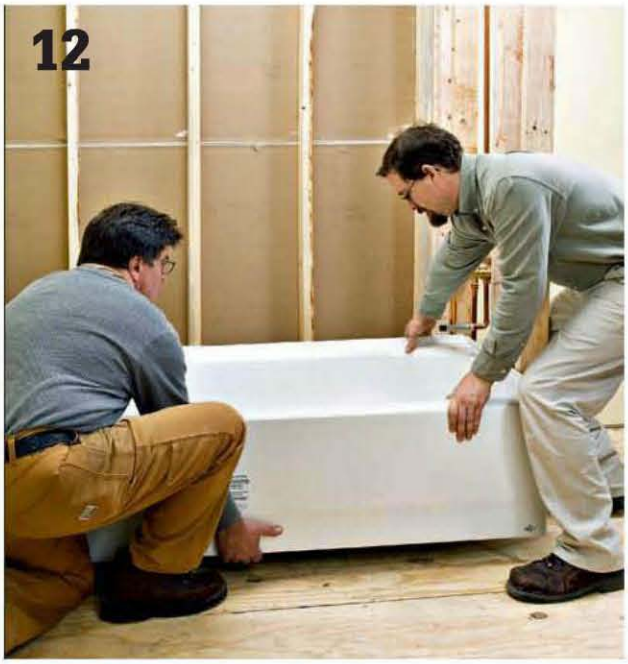

Place the tub back into the alcove, taking care not to bump the DWO assembly and disturb the connections. You definitely will want a helper for this job. If the drain outlet of the DWO assembly is not directly over the drain pipe when the tub is in position, you'll need to remove it and adjust the drain line location.

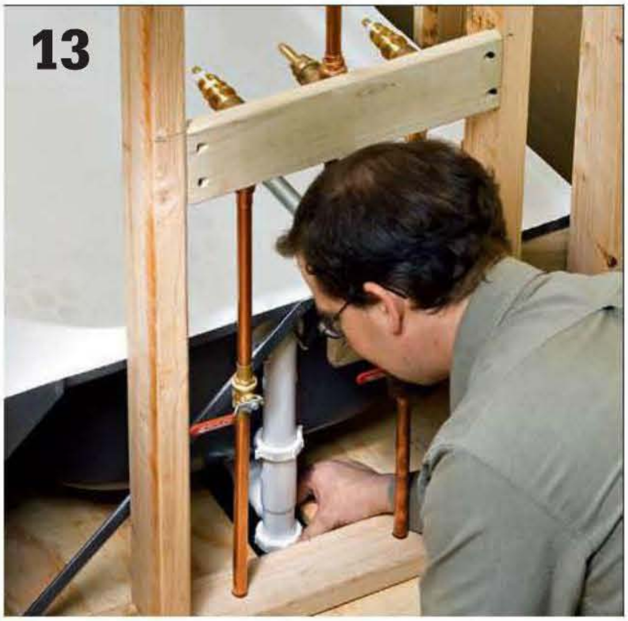

Attach the drain outlet from the DWO assembly to the drain P-trap. This is the part of the job where you will appreciate the time you spent creating a roomy access panel for the tub plumbing. Test the drain and overflow to make sure they don't leak. Also test the water supply plumbing, temporarily attaching the handles, spout, and shower arm so you can operate the faucet and the diverter.

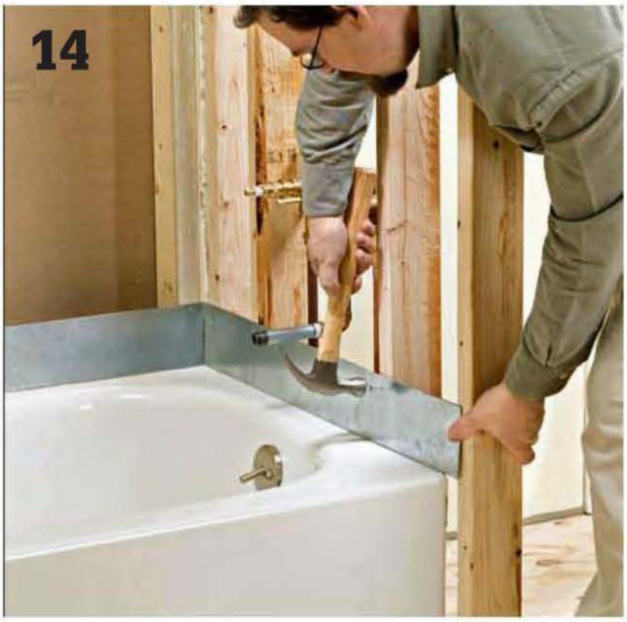

Drive a 1½" galvanized roofing nail at each stud location, just over the top of the tub's nailing flange. The nail head should pin the flange to the stud. Be careful here—an errant blow or overdriving can cause the enameled finish to crack or craze. *Option: You may choose to drill guide holes and nail through the flange instead.*

Install the wallcoverings and tub surround (see pages 446 to 449 for a 3-piece surround installation). You can also make a custom surround from tileboard or cementboard and tile.

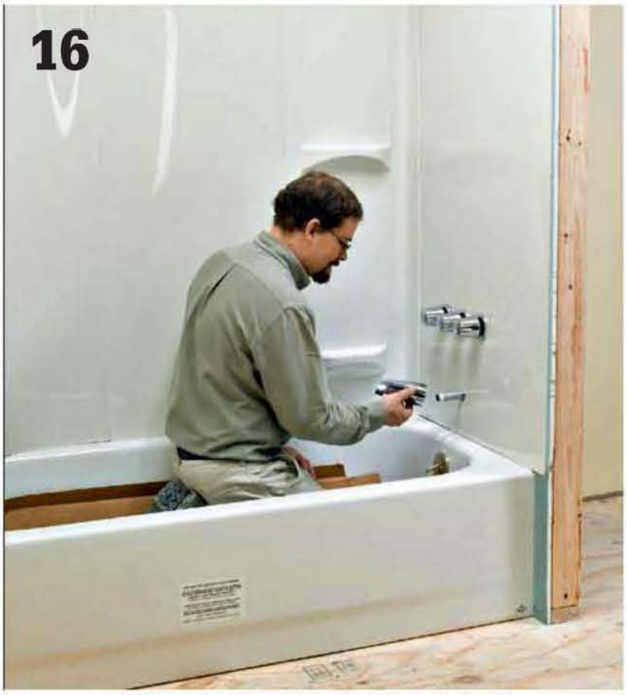

Install fittings. First, thread the shower arm into the shower elbow and attach the spout nipple to the valve assembly. Also attach the shower head and escutcheon, the faucet handle/diverter with escutcheon, and the tub spout. Use thread lubricant on all parts.

Installing 3-piece Tub Surrounds

No one wants bathroom fixtures that are aging or yellowed from years of use. A shiny new tub surround can add sparkle and freshness to your dream bath.

Tub surrounds come in many different styles, materials, and price ranges. Choose the features you want and measure your existing bathtub surround for sizing. Surrounds typically come in three or five pieces. A three-panel surround is being installed here, but the process is similar for five-panel systems.

Surface preparation is important for good glue adhesion. Plastic tiles and wallpaper must be removed and textured plaster must be sanded smooth. Surrounds can be installed over ceramic tile that is well attached and in good condition, but it must be sanded and primed. All surfaces must be primed with a water-based primer.

Tools & Materials ▸

- Jigsaw
- Hole saw
- Drill
- Measuring tape
- Level
- Caulking gun
- Primer
- Adhesive
- Screwdriver
- Adjustable wrench
- Pry bar
- Hammer
- 3-piece tub surround

Three-piece tub surrounds are inexpensive and come in many colors and styles. The typical unit has two end panels and a back panel that overlap in the corners to form a watertight seal. They are formed from fiberglass, PVC, acrylic, or proprietary resin-based polymers. Five piece versions are also available and typically have more features such as integral soap shelves and even cabinets.

How to Install a 3-piece Tub Surround

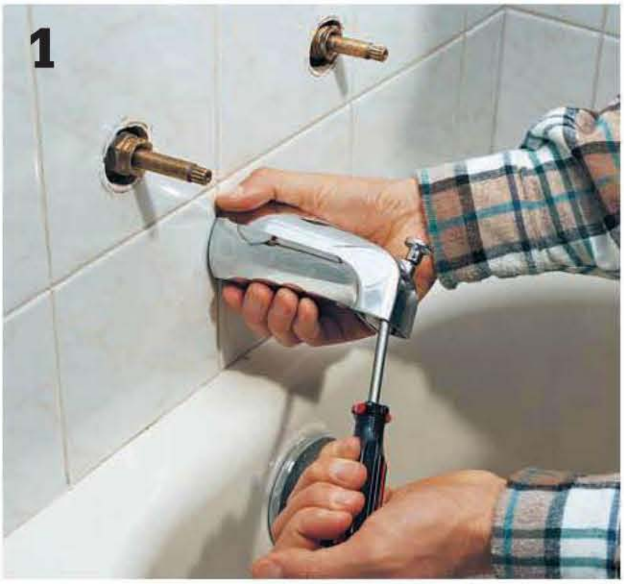

Remove the old plumbing fixtures and wallcoverings in the tub area. In some cases you can attach surround panels to old tileboard or even tile, but it is generally best to remove the wallcoverings down to the studs if you can, so you may inspect for leaks or damage.

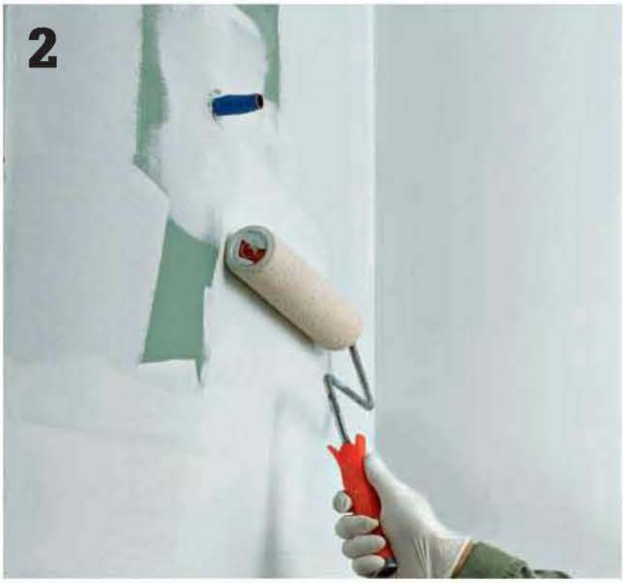

Replace the wallcoverings with appropriate materials, such as water and mold-resistant wallboard or cementboard (for ceramic tile installations). Make sure the new wall surfaces are smooth and flat. Some surround kit manufacturers recommend that you apply a coat of primer to sheet goods such as greenboard to create a better bonding surface for the panel adhesive.

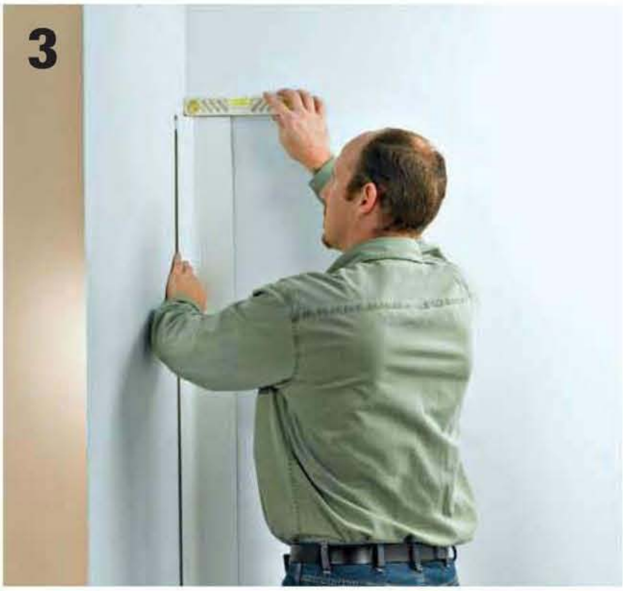

Test-fit the panels before you start; the tub may have settled unevenly or the walls may be out of plumb. Check the manufacturer's directions for distinguishing right and left panels. Place a panel in position on the tub ledge. Use a level across the top of the panel to determine if it is level. Create a vertical reference line to mark the edge of the panel on the plumbing end.

Test-fitting Tip ▸

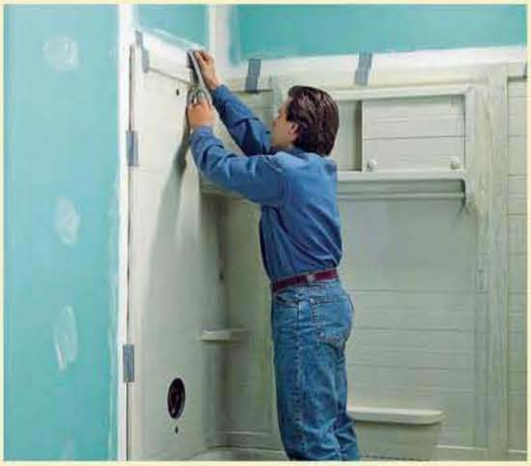

Ensure a perfect fit by taping the surround panels to the walls in the tub area. Make sure the tops are level when the overlap seams are aligned and that you have a consistent ⅛" gap between the panel bottoms and the tub flange. Mark the panels for cutting if necessary and, once the panels have been removed, make necessary adjustments to the walls.

(continued)

4

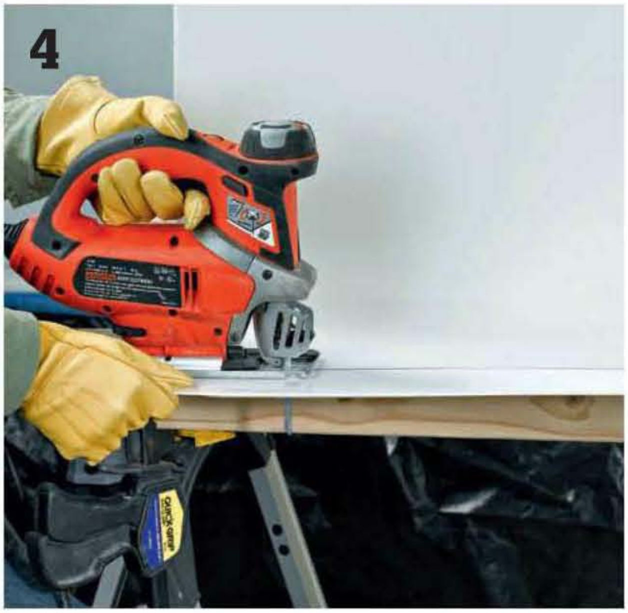

Some kits are created to fit a range of bathtub dimensions. After performing the test fit, check the fitting instructions to see if you need to trim any of the pieces. Follow the manufacturer's instructions for cutting. Here, we had to cut the corner panels because the instructions advise not to overlap the back or side panel over the corner panels by more than 3". Cut panels using a jigsaw and a fine-tooth blade that is appropriate for cutting fiberglass or acrylic tileboard. The cut panels should be overlapped by panels with factory edges.

5

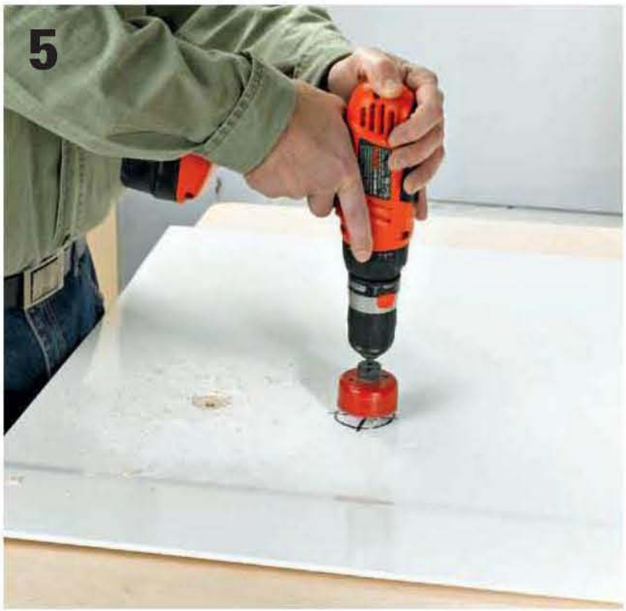

Measure and mark the location of the faucets, spout, and shower outlets. Measure in from the vertical reference line (made in step 3) and up from the top of the tub ledge. Re-measure for accuracy, as any cuts to the surround are final. Place the panel face-up on a sheet of plywood. Mark the location of the holes. Cut the holes ½" larger than the pipe diameter. If your faucet has a recessed trim plate (escutcheon), cut the hole to fit the recess. Using a hole saw or a jigsaw, cut out the plumbing outlets.

6

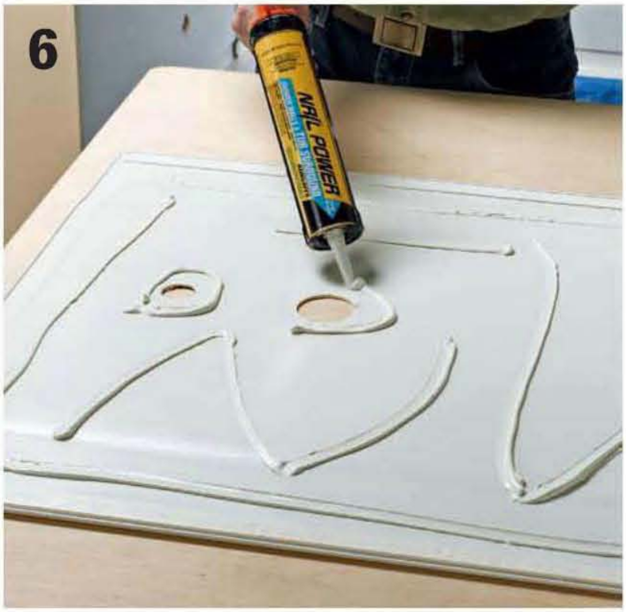

Install the plumbing end panel, test-fitting first. In this surround, the end panels are installed first. Apply adhesive to the back of the plumbing panel. Circle the plumbing outlet holes 1" from the edge. Do not apply adhesive closer than 1" to the double-sided tape (if your kit has tape) or to the bottom edge of the panel.

7

Remove the protective backing from the tape. Carefully lift the panel by the edges and place against the corner and top of the tub ledge. Press firmly from top to bottom in the corner, then throughout the panel.

8

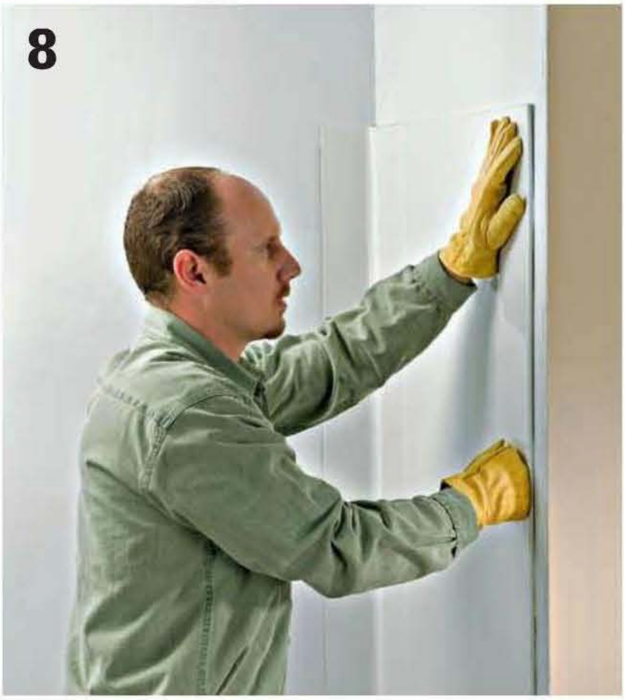

Test-fit the opposite end panel and make any necessary adjustments. Apply the adhesive, remove the protective backing from the tape, and put in place. Apply pressure to the corner first from top to bottom, and then apply pressure throughout.

9

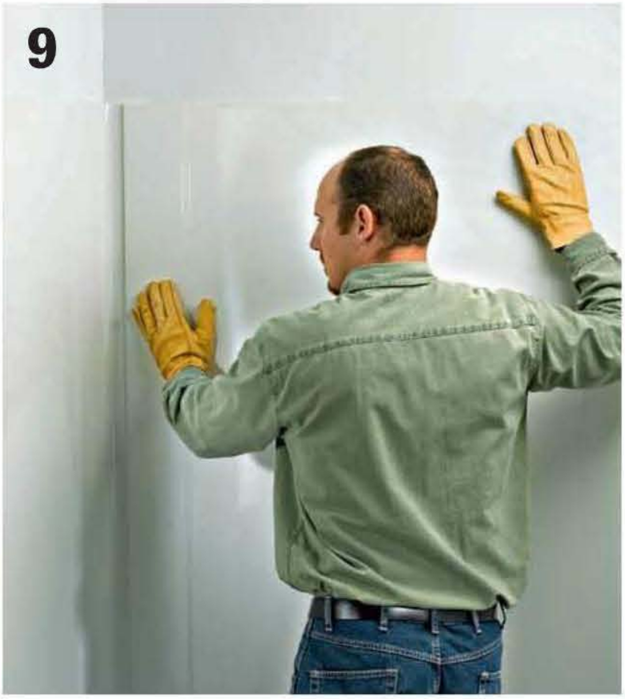

Apply adhesive to the back panel following the manufacturer's instructions. Remove protective backing from the tape. Lift the panel by the edges and carefully center between the two end panels. When positioned, firmly press in place from top to bottom.

10

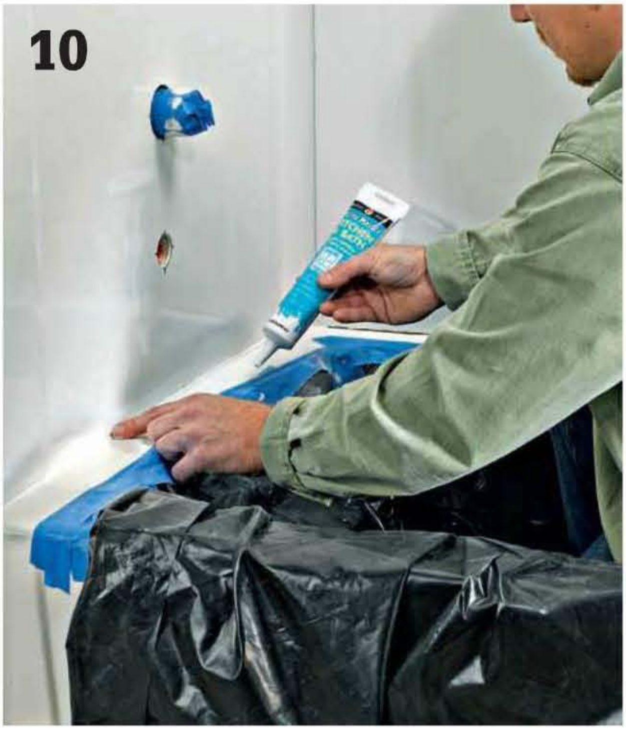

Apply caulk to the bottom and top edges of the panels and at panel joints. Dip your fingertip in water and use it to smooth the caulk to a uniform bead.

11

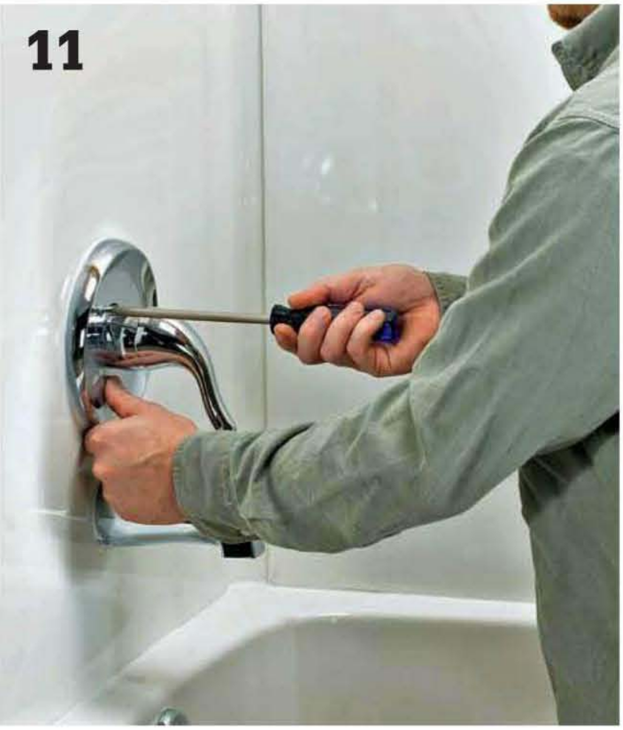

Apply silicone caulk to escutcheons or trim plates and reinstall them. Allow a minimum of 24 hours for caulk and adhesive to dry thoroughly before using the shower or tub.

Installing Sliding Tub Doors

Curtains on your bathtub shower are a hassle. If you forget to tuck them inside the tub, water flows freely onto your bathroom floor. If you forget to slide them closed, mildew sets up shop in the folds. And every time you brush against them, they stick to your skin. Shower curtains certainly don't add much elegance or charm to a dream bath. Neither does a deteriorated door. Clean up the look of your bathroom, and even give it an extra touch of elegance, with a new sliding tub door.

When shopping for a sliding tub door, you have a choice of framed or frameless. A framed door is edged in metal. The metal framing is typically aluminum but is available in many finishes, including those that resemble gold, brass, or chrome. Glass options are also plentiful. You can choose between frosted or pebbled glass, clear, mirrored, tinted, or patterned glass. Doors can be installed on ceramic tile walls or through a fiberglass tub surround.

Tools & Materials ▸

Measuring tape
Pencil
Hacksaw
Miter box
Level
Drill
Center punch
Razor blade
Marker
Masonry bit for tile wall
Phillips screwdriver
Caulk gun
Masking tape
Silicone sealant & remover
Tub door kit
Masking tape

A sliding tub door framed in aluminum gives the room a sleek, clean look and is just one of the available options.

How to Install Sliding Tub Doors

Remove the existing door and inspect the walls. Use a razor blade to cut sealant from tile and metal surfaces. Do not use a razor blade on fiberglass surfaces. Remove remaining sealant by scraping or pulling. Use a silicone sealant remover to remove all residue. Remove shower curtain rods, if present. Check the walls and tub ledge for plumb and level.

Measure the distance between the finished walls along the top of the tub ledge. Refer to the manufacturer's instructions for figuring the track dimensions. For the product seen here, 3⁄16" is subtracted from the measurement to calculate the track dimensions.

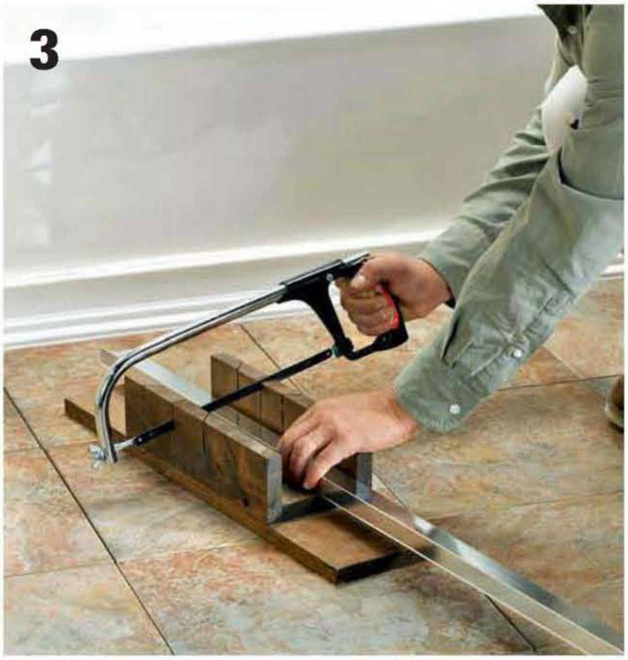

Using a hacksaw and a miter box, carefully cut the track to the proper dimension. Center the track on the bathtub ledge with the taller side out and so the gaps are even at each end. Tape into position with masking tape.

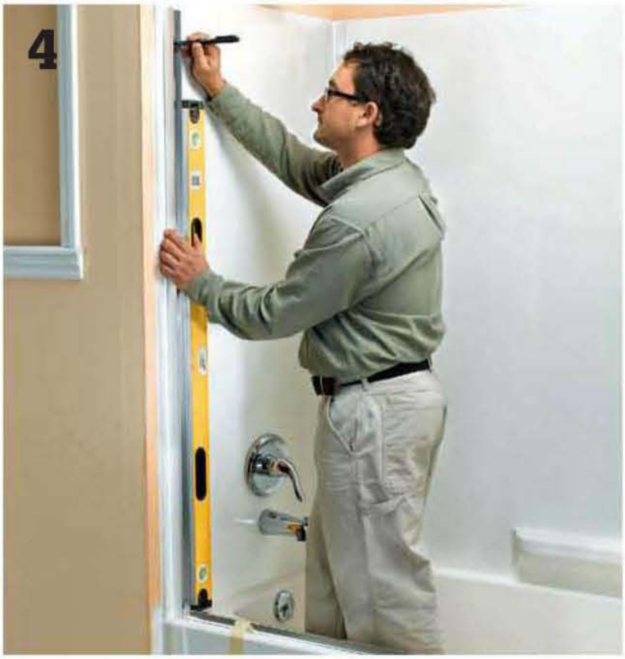

Place a wall channel against the wall with the longer side out and slide it into place over the track so they overlap. Use a level to check the channel for plumb, and then mark the locations of the mounting holes on the wall with a marker. Repeat for the other wall channel. Remove the track.

(continued)

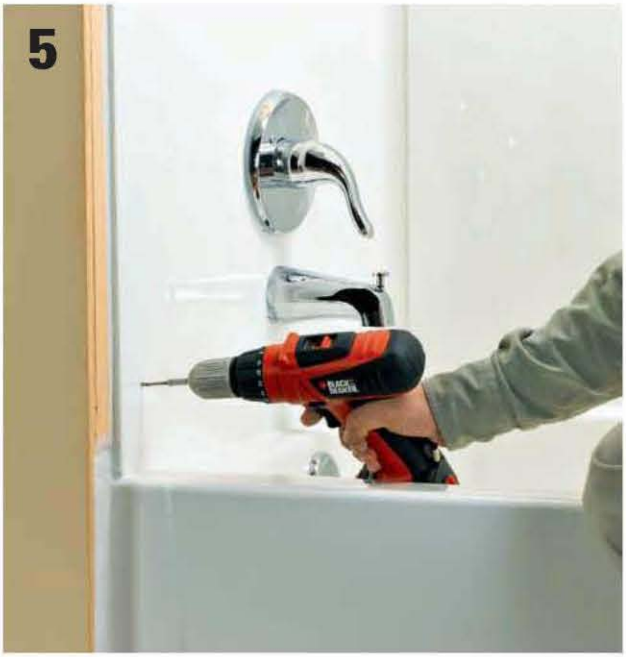
5

Drill mounting holes for the wall channel at the marked locations. In ceramic tile, nick the surface of the tile with a center punch, use a ¼" masonry bit to drill the hole, and then insert the included wall anchors. For fiberglass surrounds, use a ⅛" drill bit; wall anchors are not necessary.

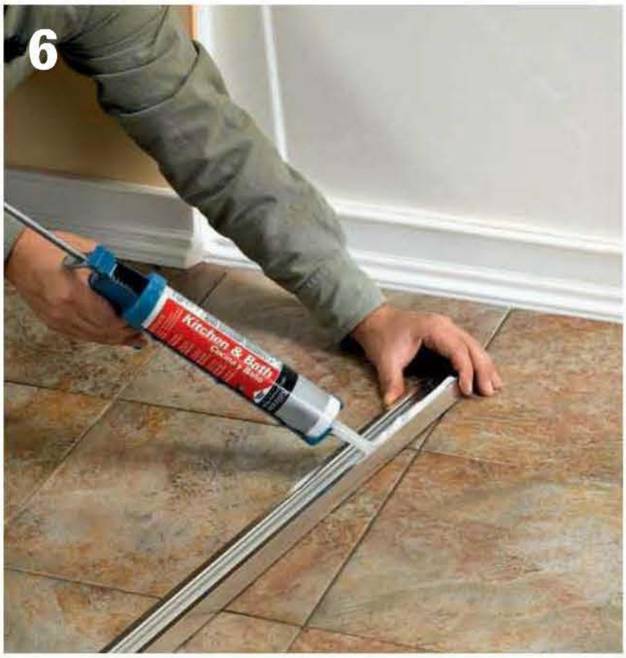
6

Apply a bead of silicone sealant along the joint between the tub and the wall at the ends of the track. Apply a minimum ¼" bead of sealant along the outside leg of the track underside.

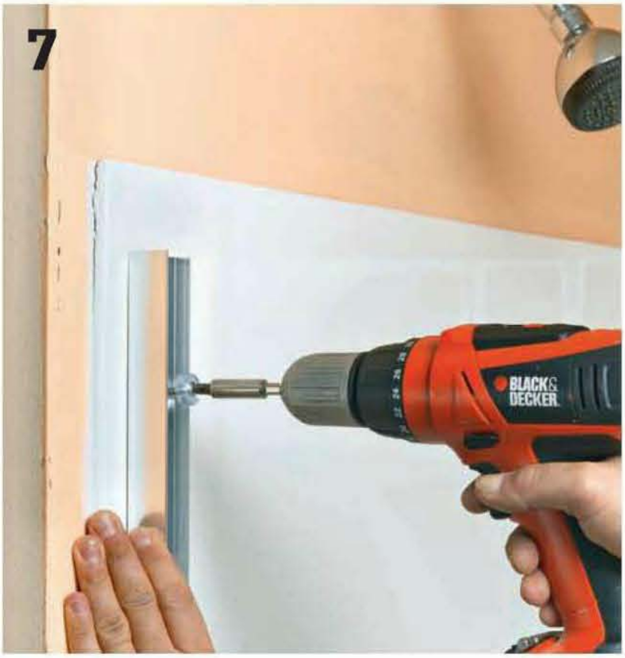
7

Position the track on the tub ledge and against the wall. Attach the wall channels using the provided screws. Do not use caulk on the wall channels at this time.

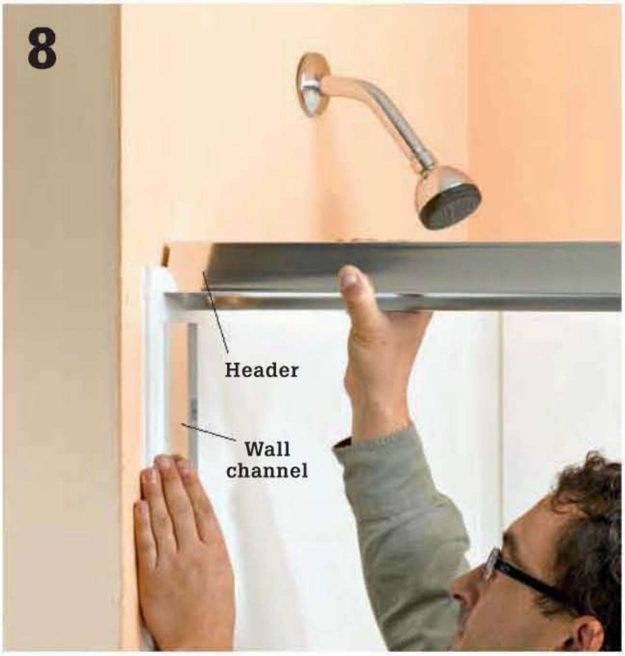
8

Cut and install the header. At a location above the tops of the wall channels, measure the distance between the walls. Refer to the manufacturer's instructions for calculating the header length. For the door seen here, the length is the distance between the walls minus 1⁄16". Measure the header and carefully cut it to length using a hacksaw and a miter box. Slide the header down on top of the wall channels until seated.

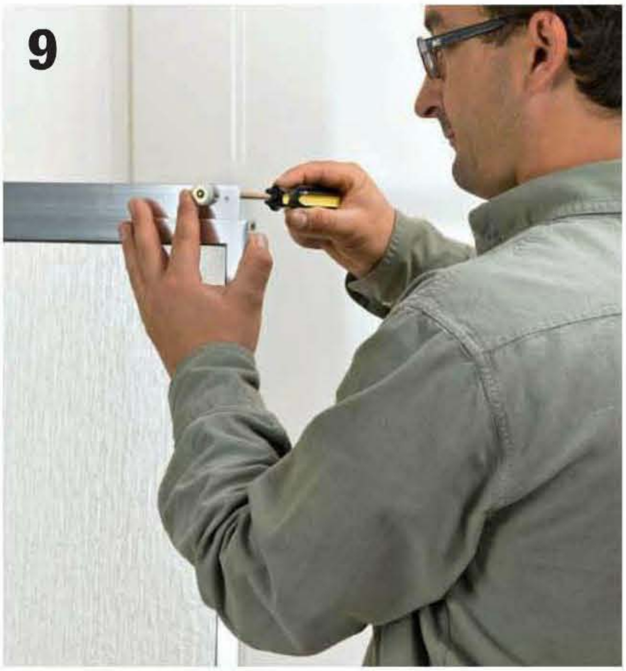

Mount the rollers in the roller mounting holes. To begin, use the second-from-the-top roller mounting holes. Follow the manufacturer's instructions for spacer or washer placement and orientation.

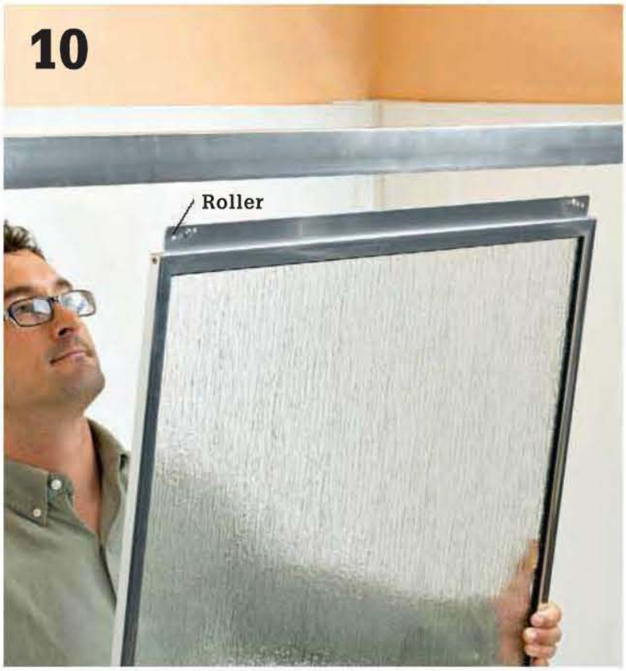

Carefully lift the inner panel by the sides and place the rollers on the inner roller track. Roll the door toward the shower end of the tub. The edge of the panel should touch both rubber bumpers. If it doesn't, remove the door and move the rollers to different holes. Drive the screws by hand to prevent overtightening.

Lift the outer panel by the sides with the towel bar facing out from the tub. Place the outer rollers over the outer roller track. Slide the door to the end opposite the shower end of the tub. If the door does not contact both bumpers, remove the door and move the rollers to different mounting holes.

12
Kitchen & Bath

Apply a bead of clear silicone sealant to the inside seam of the wall and wall channel at both ends and to the U-shaped joint of the track and wall channels. Smooth the sealant with a fingertip dipped in water.

Tiling Bathroom Walls

Tile is an ideal covering for walls in kitchens and bathrooms, but there's no reason to limit its use to those rooms. It's not as common in North American homes, but in Europe tile has been used in rooms throughout the house for generations. And why not? Beautiful, practical, easy to clean and maintain, tile walls are well suited to many spaces. On the preceding pages, you've seen some design ideas for tile walls.

When shopping for tile, keep in mind that tiles that are at least 6" × 6" are easier to install than small tiles, because they require less cutting and cover more surface area. Larger tiles also have fewer grout lines that must be cleaned and maintained. Check out the selection of trim and specialty tiles and ceramic accessories that are available to help you customize your project.

Most wall tile is designed to have narrow grout lines (less than ⅛" wide) filled with unsanded grout. Grout lines wider than ⅛" should be filled with sanded floor-tile grout. Either type will last longer if it contains, or is mixed with, a latex additive. To prevent staining, it's a good idea to seal your grout after it fully cures, then once a year thereafter.

You can use standard drywall or water-resistant drywall (called "greenboard") as a backer for walls in dry areas. In wet areas, install tile over cementboard. Made from cement and fiberglass, cementboard cannot be damaged by water, though moisture can pass through it. To protect the framing, install a waterproof membrane, such as roofing felt or polyethylene sheeting, between the framing members and the cementboard. Be sure to tape and finish the seams between cementboard panels before laying the tile.

See page 45 for information on planning and laying out tile walls.

Tools & Materials ▸

Tile-cutting tools
Marker
Tape measure
4-ft. level
Notched trowel
Mallet
Grout float
Grout sponge
Soft cloth
Small paintbrush or foam brush
Caulk gun
Straight 1 × 2
Dry-set tile mortar with latex additive
Ceramic wall tile
Ceramic trim tile (as needed)
2 × 4
Carpet scrap
Tile grout with latex additive
Tub & tile caulk
Alkaline grout sealer
Cardboard
Story stick/pole

Wall tile is very popular and practical in bathrooms, where it can be used on all surfaces or primarily as an accent. Installing custom-tiled tub and shower surrounds is a popular DIY project.

How to Install Wall Tile in a Bathtub Alcove

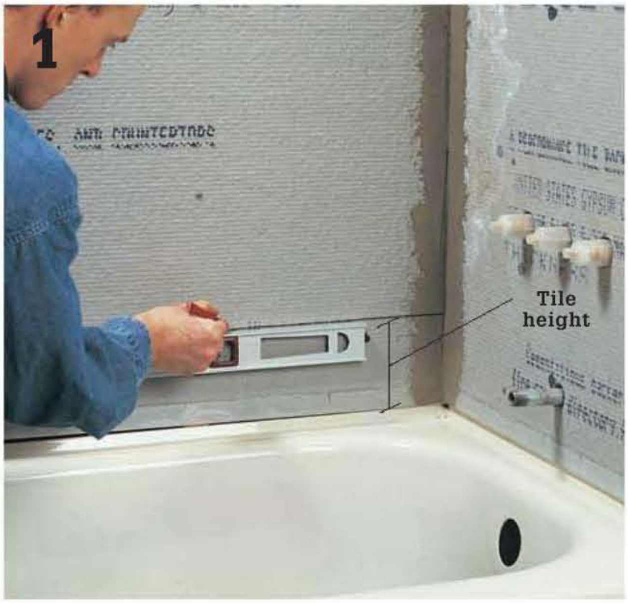

Beginning with the back wall, measure up and mark a point at a distance equal to the height of one ceramic tile (if the tub edge is not level, measure up from the lowest spot). Draw a level line through this point, along the entire back wall. This line represents a tile grout line and will be used as a reference line for making the entire tile layout.

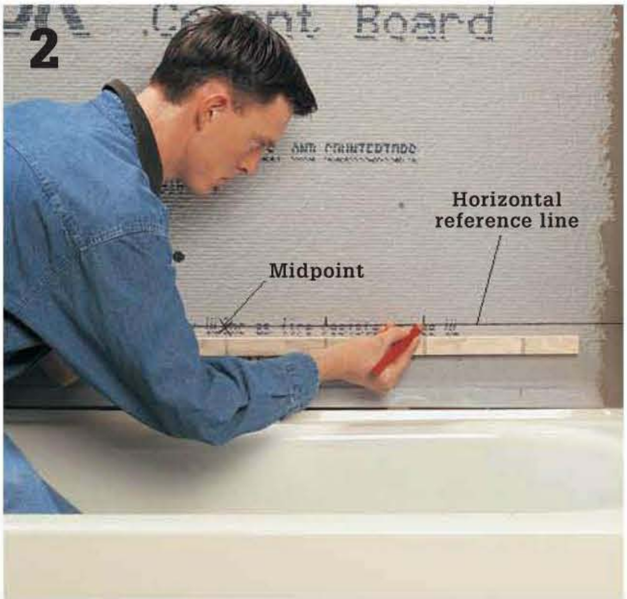

Measure and mark the midpoint on the horizontal reference line. Using a story stick, mark along the reference line where the vertical grout joints will be located. If the story stick shows that the corner tiles will be less than half of a full tile width, move the midpoint half the width of a tile in either direction and mark (shown in next step).

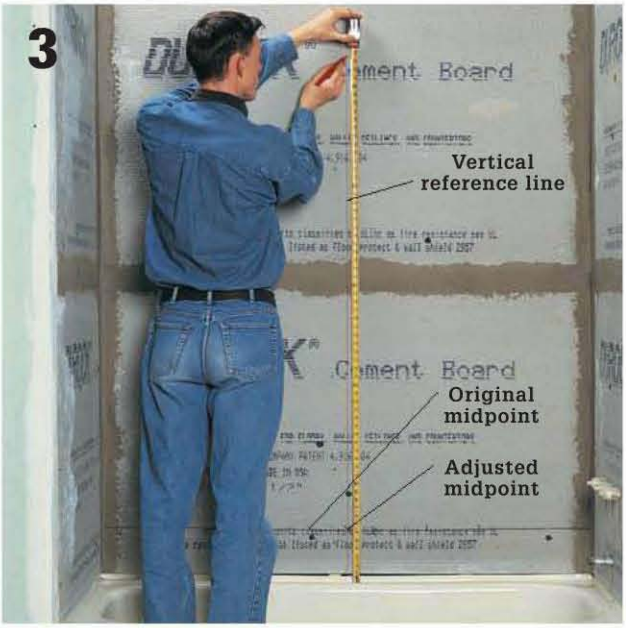

Use a level to draw a vertical reference line through the adjusted midpoint from the tub edge to the ceiling. Measure up from the tub edge along the vertical reference line and mark the rough height of the top row of tiles.

Use the story stick to mark the horizontal grout joints along the vertical reference line, beginning at the mark for the top row of tiles. If the cut tiles at the tub edge will be less than half the height of a full tile, move the top row up half the height of a tile. *Note: If tiling to a ceiling, evenly divide the tiles to be cut at the ceiling and tub edge, as for the corner tiles.*

(continued)

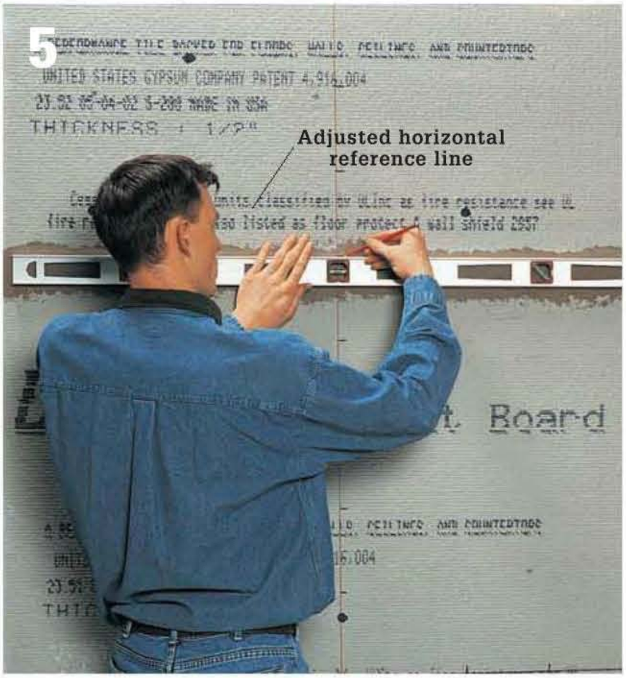

Use a level to draw an adjusted horizontal reference line through the vertical reference line at a grout joint mark close to the center of the layout. This splits the tile area into four workable quadrants.

Use a level to transfer the adjusted horizontal reference line from the back wall to both side walls, then lay out both side walls. Adjust the layout as needed so the final column of tiles ends at the outside edge of the tub. Use only the adjusted horizontal and vertical reference lines for ceramic tile installation.

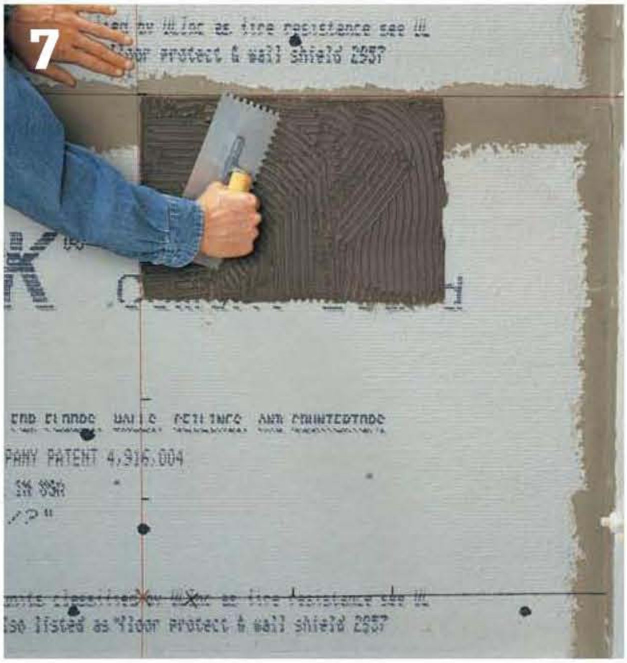

Mix a small batch of thinset mortar containing a latex additive. (Some mortar has additive mixed in by the manufacturer and some must have additive mixed separately.) Spread adhesive on a small section of the wall, along both legs of one quadrant, using a ¼" notched trowel.

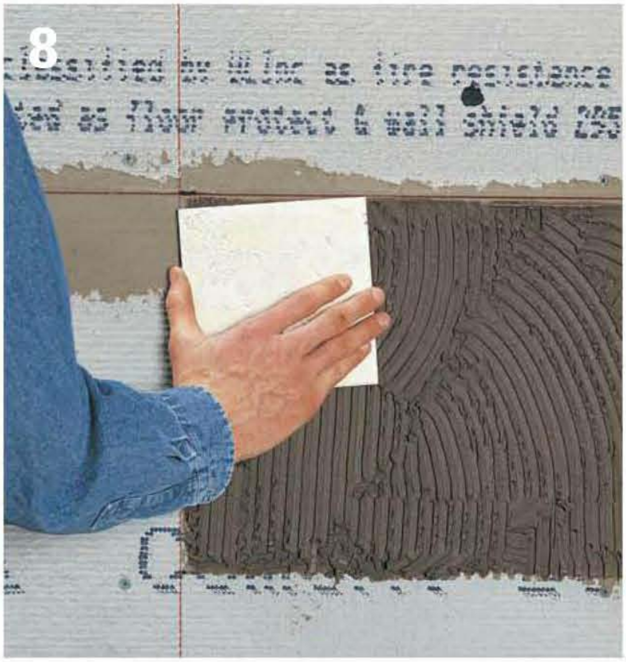

Use the edge of the trowel to create furrows in the mortar. Set the first tile in the corner of the quadrant where the lines intersect, using a slight twisting motion. Align the tile exactly with both reference lines. When placing cut tiles, position the cut edges where they will be least visible.

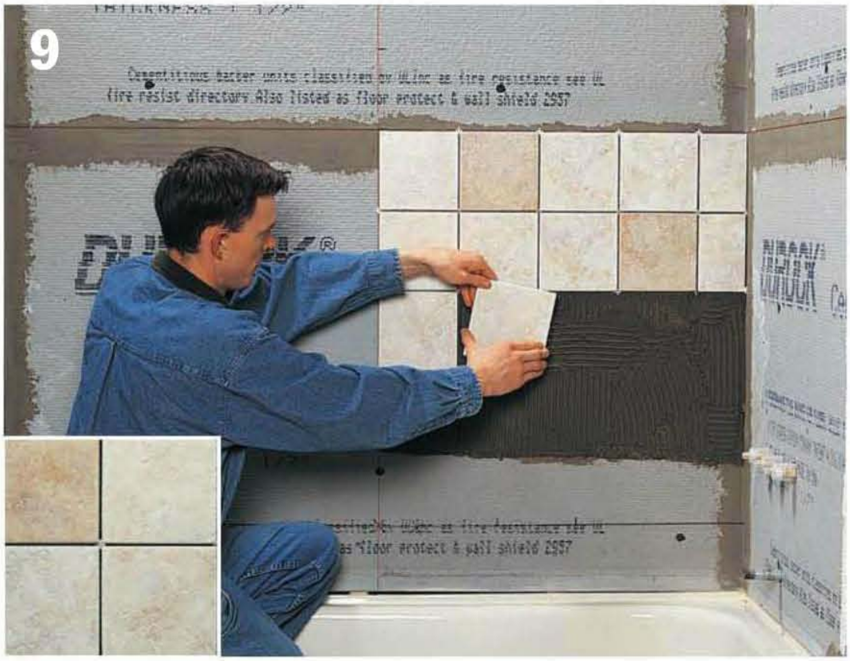

Continue installing tiles, working from the center out into the field of the quadrant. Keep the tiles aligned with the reference lines and tile in one quadrant at a time. If the tiles are not self-spacing, use plastic spacers inserted in the corner joints to maintain even grout lines (inset). The base row against the tub edge should be the last row of tiles installed.

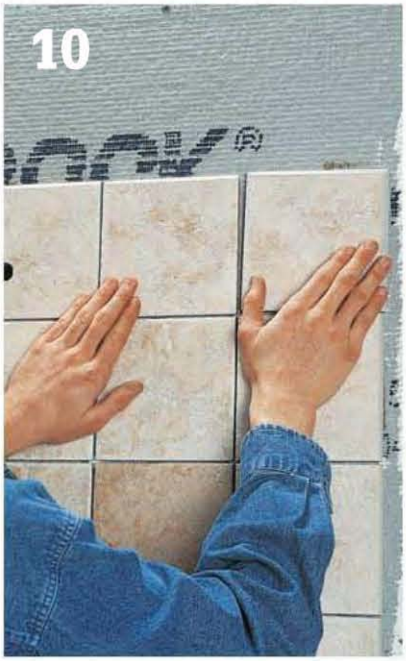

Install remaining tiles, including border areas. Wipe away excess mortar along the top edges of the edge tiles.

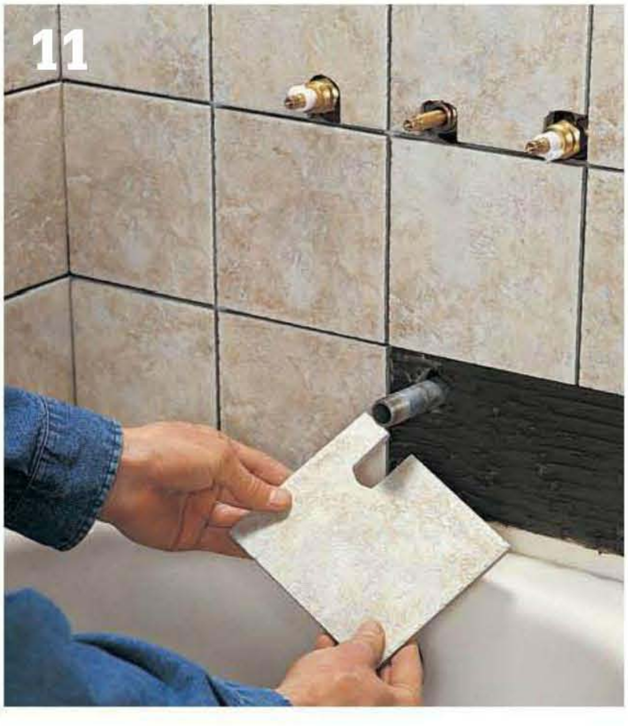

Mark and cut tiles to fit around all plumbing accessories or plumbing fixtures.

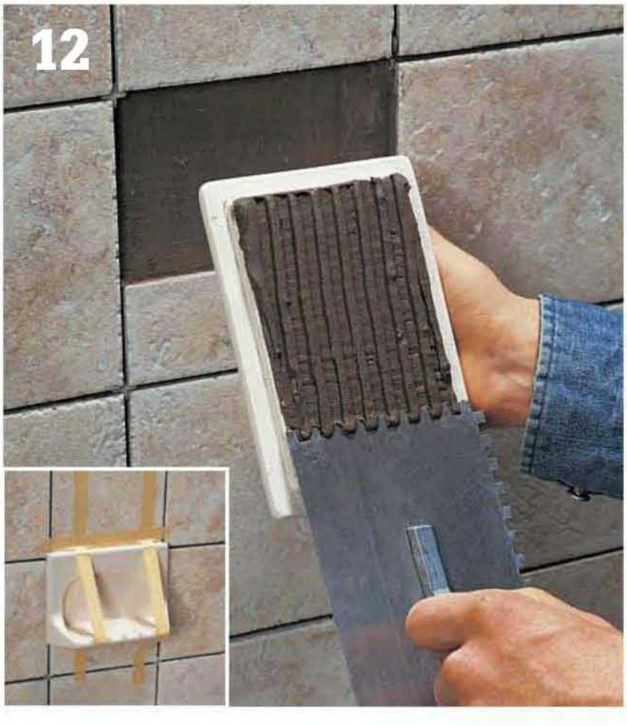

Install any ceramic accessories by applying thinset mortar to the back side, then pressing the accessory into place. Use masking tape to support the weight until the mortar dries (inset). Fill the tub with water, then seal expansion joints around the bathtub, floor, and corners with silicone caulk.

Installing Toilets

You can replace a poorly functioning or inefficient toilet with a high-efficiency, high-quality new toilet in just a single afternoon. All toilets made since 1996 have been required to use 1.6 gallons or less per flush, which has been a huge challenge for the industry. Today, the most evolved 1.6-gallon toilets have wide passages behind the bowl and wide (3") flush valve openings—features that facilitate short, powerful flushes. This means fewer second flushes and fewer clogged toilets. These problems were common complaints of the first generation of 1.6-gallon toilets and continue to beleaguer inferior models today. See what toilets are available at your local home center in your price range, then go online and see what other consumers' experiences with those models have been. New toilets often go through a "de-bugging" stage when problems with leaks and malfunctioning parts are more common. Your criteria should include ease of installation, good flush performance, and reliability. With a little research, you should be able to purchase and install a high-functioning economical gravity-flush toilet that will serve you well for years to come.

Tools & Materials ▸

Adjustable wrench
Bucket and sponge
Channel-type pliers
Hacksaw
Penetrating oil
Pliers
Putty knife
Rubber gloves
Screwdriver
Supply tube
Teflon tape
Toilet seat bolts
Toilet seat
Towels
Utility knife
Wax ring with flange

Replacing a toilet is simple, and the latest generation of 1.6-gallon water-saving toilets has overcome the performance problems of earlier models.

How to Install a Toilet

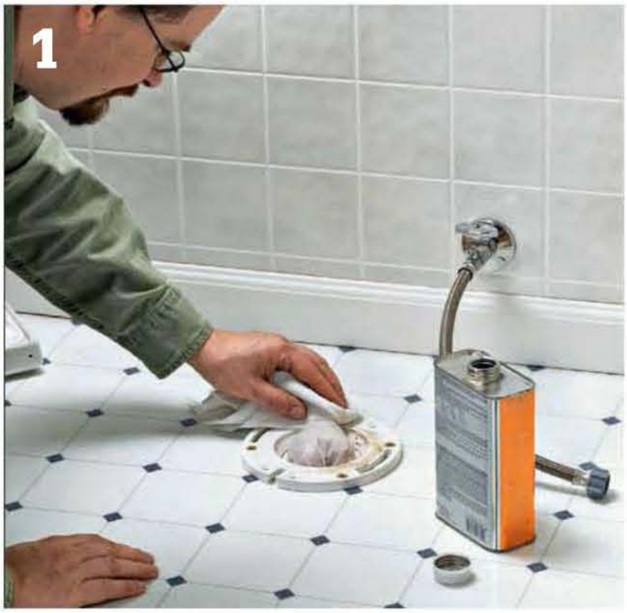

Clean and inspect the old closet flange. Look for breaks or wear. Also inspect the flooring around the flange. If either the flange or floor is worn or damaged, repair the damage. Use a rag and mineral spirits to completely remove residue from the old wax ring. Place a rag into a plastic bag and use it to block toxic sewer gas in the drain opening.

Tip ▸

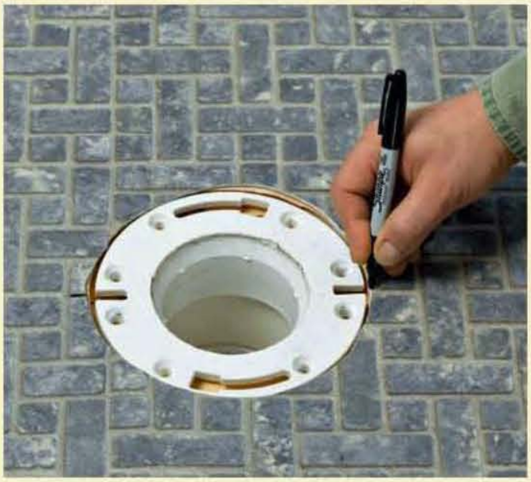

If you will be replacing your toilet flange or if your existing flange can be unscrewed and moved, orient the new flange so the slots are parallel to the wall. This allows you to insert bolts under the slotted areas, which are much stronger than the areas at the ends of the curved grooves.

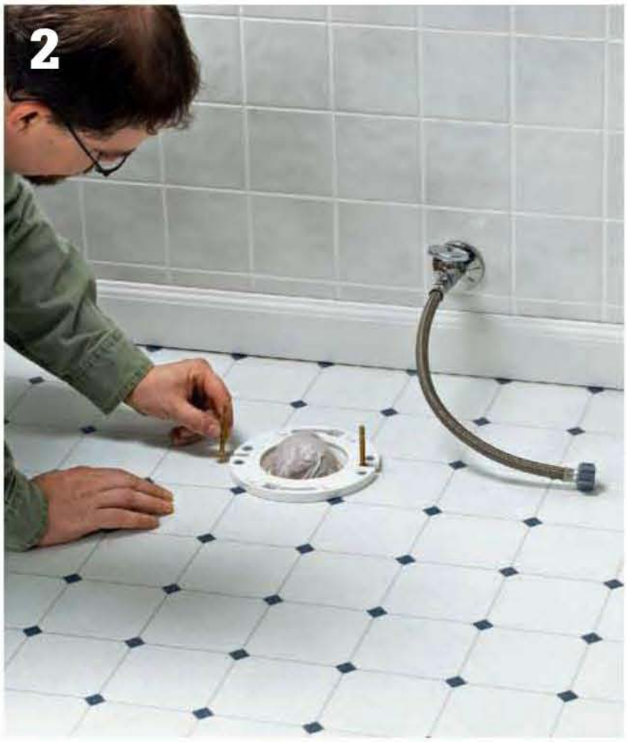

Insert new tank bolts (don't reuse old ones) into the openings in the closet flange. Make sure the heads of the bolts are oriented to catch the maximum amount of flange material.

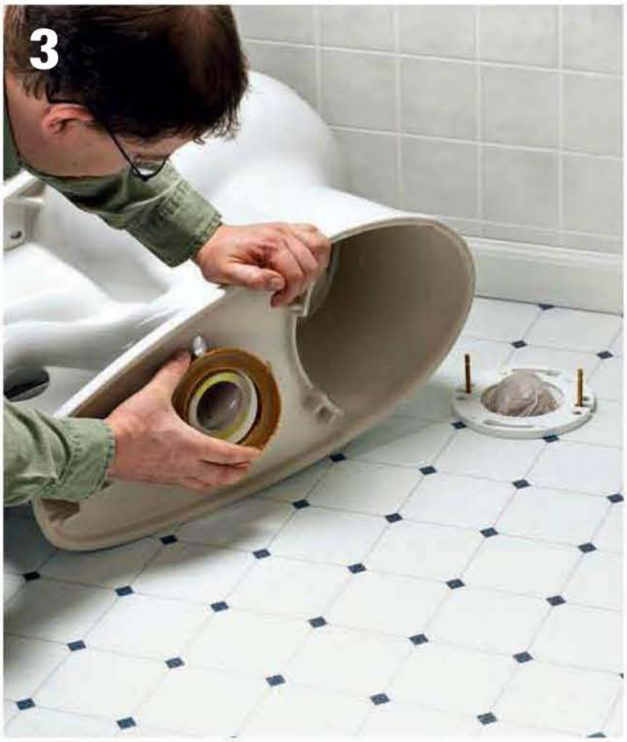

Remove the wax ring and apply it to the underside of the bowl, around the horn. Remove the protective covering. Do not touch the wax ring. It is very sticky. Remove the plastic bag.

(continued)

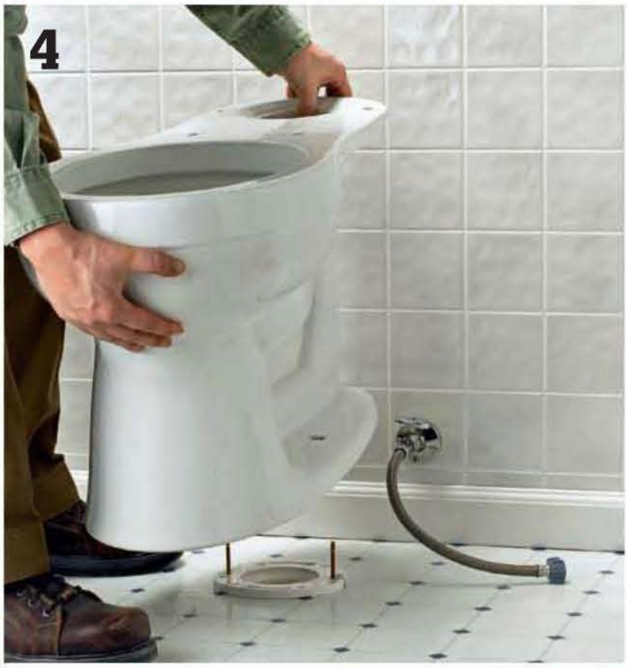

Lower the bowl onto the flange, taking care not to disturb the wax ring. The holes in the bowl base should align perfectly with the tank bolts. Add a washer and tighten a nut on each bolt. Hand tighten each nut and then use channel-type pliers to further tighten the nuts. Alternate back and forth between nuts until the bowl is secure. *Do not overtighten.*

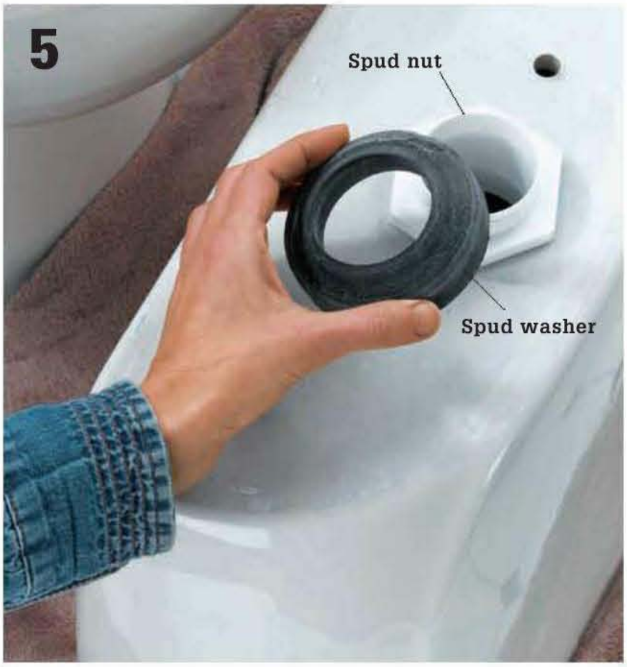

Attach the toilet tank. Some tanks come with a flush valve and a fill valve preinstalled. For models that do not have this, insert the flush valve through the tank opening and tighten a spud nut over the threaded end of the valve. Place a foam spud washer on top of the spud nut.

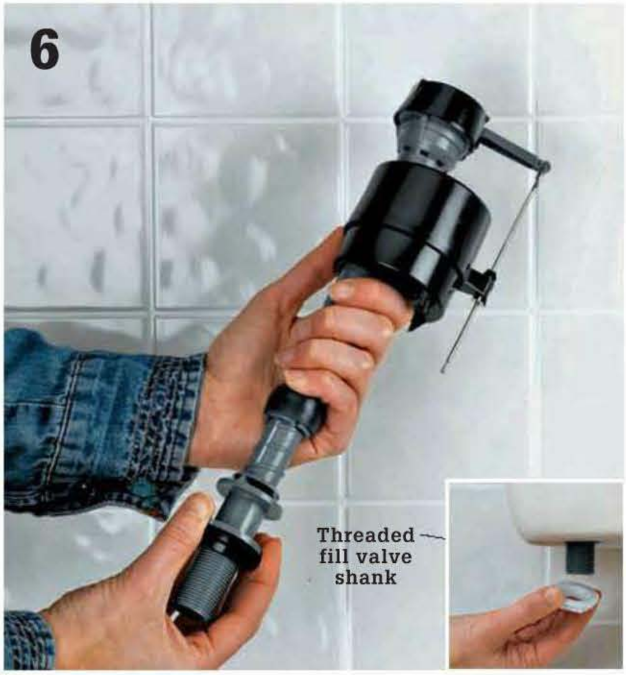

Adjust the fill valve as directed by the manufacturer to set the correct tank water level height and install the valve inside the tank. Hand tighten the nylon lock nut that secures the valve to the tank (inset photo) and then tighten it further with channel-type pliers.

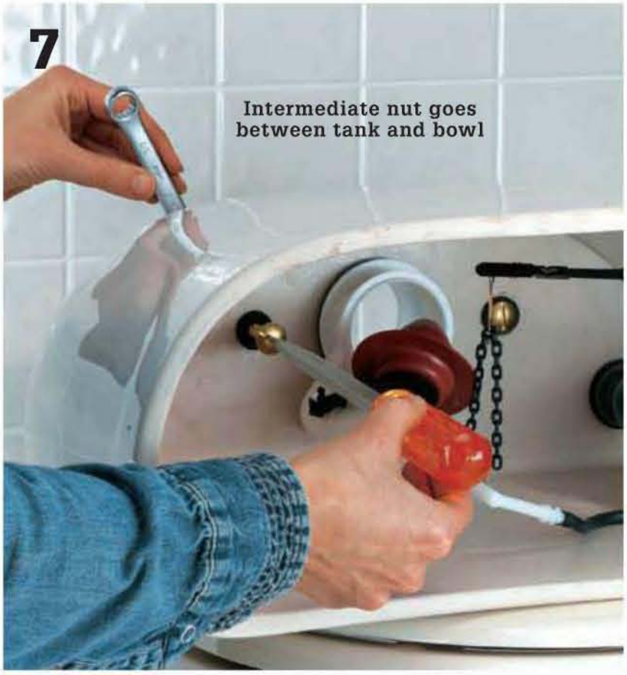

With the tank lying on its back, thread a rubber washer onto each tank bolt and insert it into the bolt holes from inside the tank. Then, thread a brass washer and hex nut onto the tank bolts from below and tighten them to a quarter turn past hand tight. Do not overtighten.

8

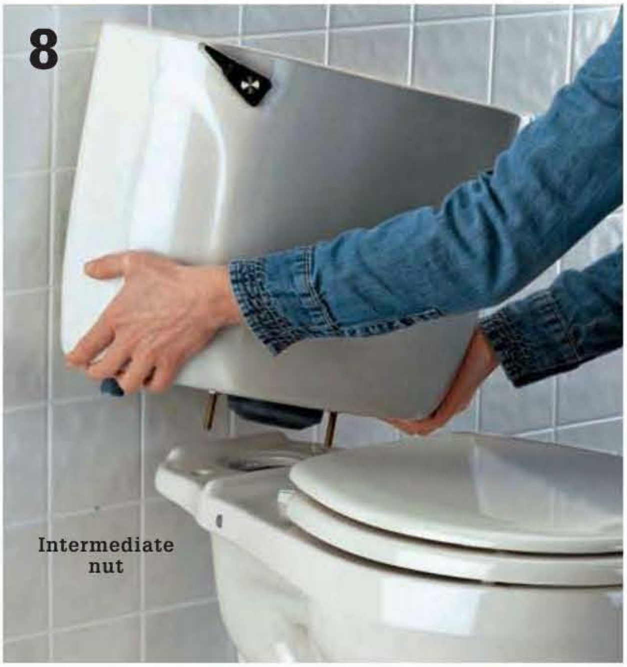

Position the tank on the bowl, spud washer on opening, bolts through bolt holes. Put a rubber washer, followed by a brass washer and a wing nut, on each bolt and tighten these up evenly.

9

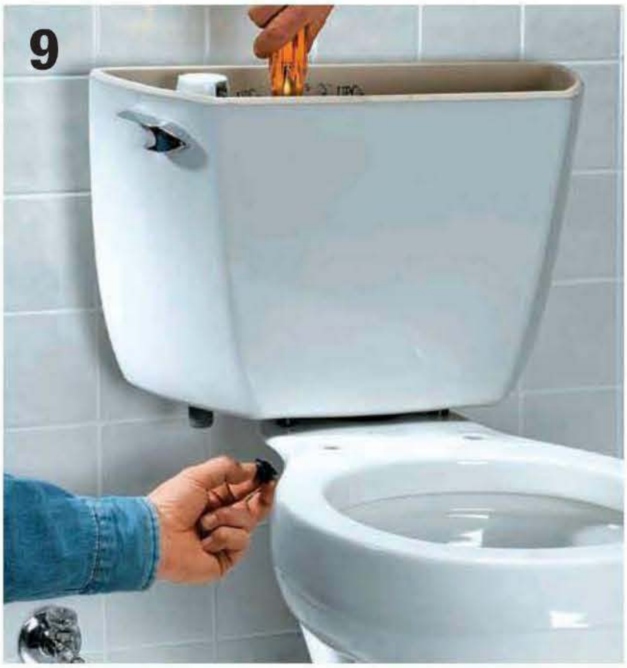

You may stabilize the bolts with a large slotted screwdriver from inside the tank, but tighten the nuts, not the bolts. You may press down a little on a side, the front, or the rear of the tank to level it as you tighten the nuts by hand. Do not overtighten and crack the tank. The tank should be level and stable when you're done.

10

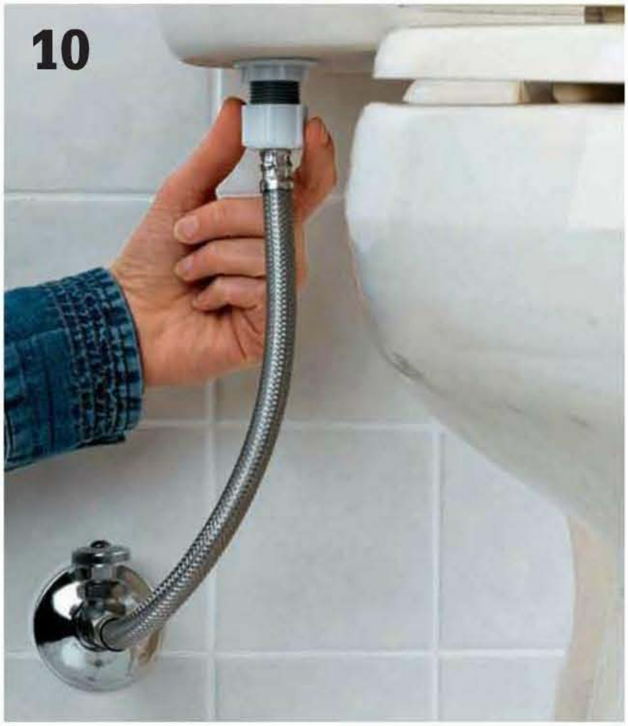

Hook up the water supply by connecting the supply tube to the threaded fill valve with the coupling nut provided. Turn on the water and test for leaks. Do not overtighten.

11

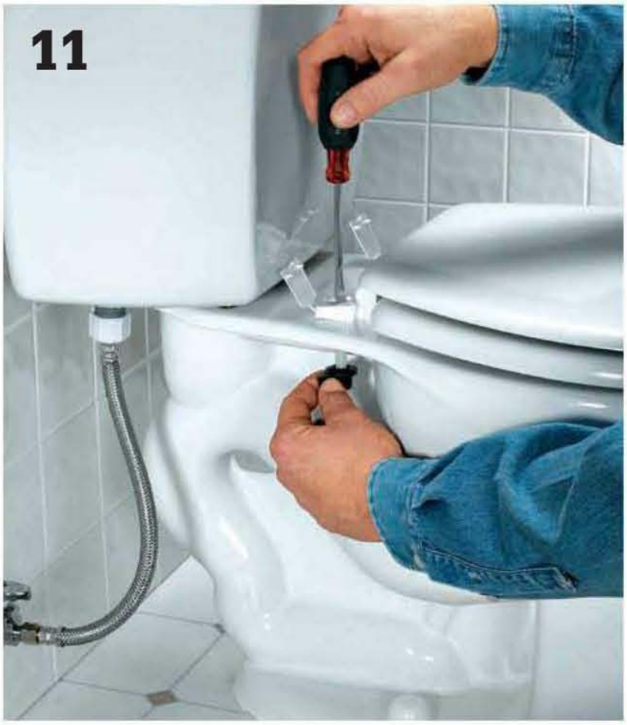

Attach the toilet seat by threading the plastic or brass bolts provided with the seat through the openings on the back of the rim and attaching the nuts.

Installing Wall-hung Vanities

Think of a wall-mounted sink or vanity cabinet and you're likely to conjure up images of public restrooms where these conveniences are installed to improve access for floor cleaning. However, wall-hung sinks and vanities made for home use are very different from the commercial installations.

Often boasting high design, beautiful modern vanities and sinks come in a variety of styles and materials, including wood, metal, and glass. Some attach with decorative wall brackets that are part of the presentation; others look like standard vanities without legs. Install wall-hung sinks and vanities by attaching them securely to studs or wood blocking.

Tools & Materials ▸

Studfinder
Drill
Level
Vanity

Today's wall-hung sinks are stylish and attractive. They must be mounted into studs or added blocking to keep them secure.

How to Install a Wall-hung Vanity Base

Remove the existing sink or fixture and inspect the wall framing. Also determine if plumbing supply and waste lines will need to be moved to accommodate the dimensions of the new fixture. Locate the studs in the sink location with a stud finder.

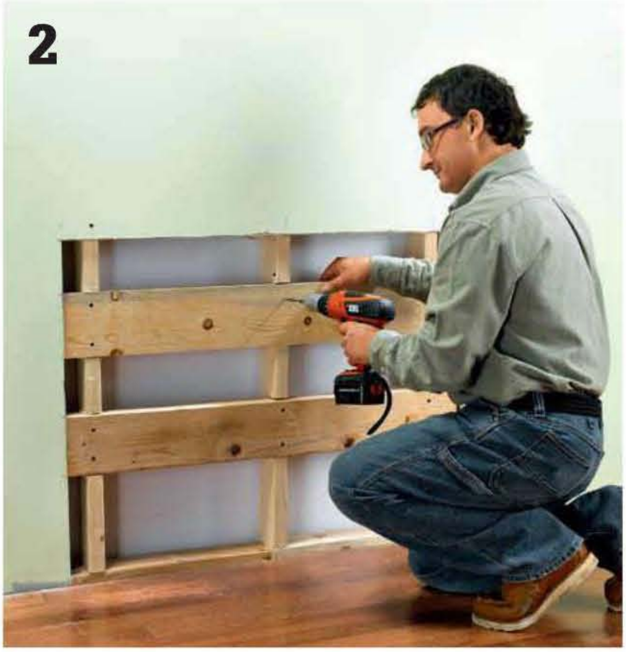

Hold the sink or cabinet in the installation area and check to see if the studs align with the sink or sink bracket mounting holes. If they do, skip to step 3. If the studs do not align, remove the wallboard behind the mounting area. Install 2 × 6 blocking between studs at the locations of the mounting screws. Replace and repair wallboard.

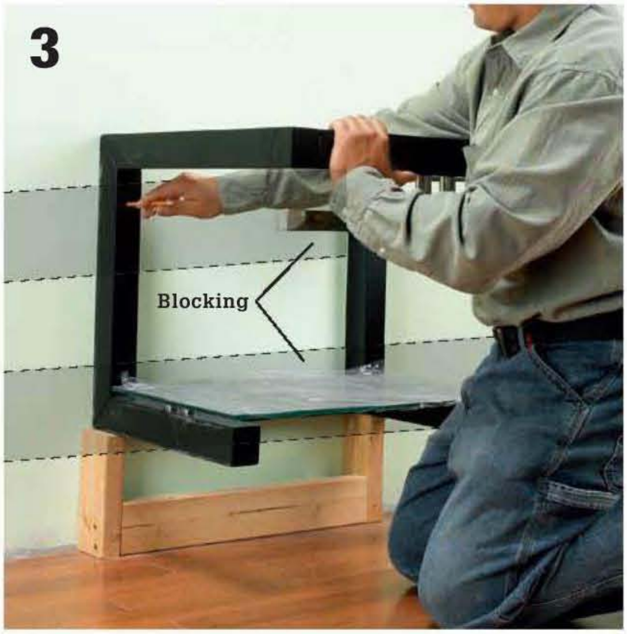

Mark the locations of the mounting holes on the wall using a template or by supporting the sink or vanity against the wall with a temporary brace (made here from scrap 2 × 4s) and marking through the mounting holes.

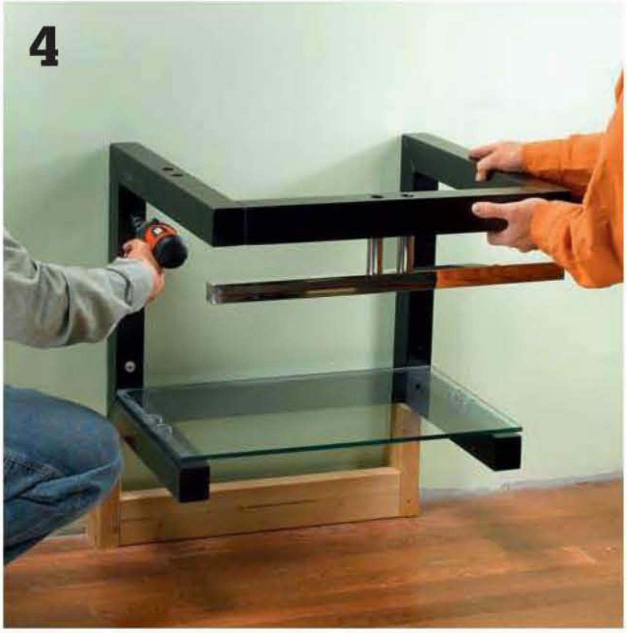

Drill pilot holes at the marks. Have a helper hold the vanity in place while you drive the mounting screws. Hook up the plumbing.

Installing Vessel Sinks

The vessel sink harkens back to the days of washstands and washbowls. Whether it's round, square, or oval, shallow or deep, the vessel sink offers great opportunity for creativity and proudly displays its style. Vessel sinks are a perfect choice for a powder room, where they will have high visibility.

Most vessel sinks can be installed on any flat surface—from a granite countertop to a wall-mounted vanity to an antique dresser. Some sinks are designed to contact the mounting surface only at the drain flange. Others are made to be partially embedded in the surface. Take care to follow the manufacturer's instructions for cutting holes for sinks and faucets.

A beautiful vessel sink demands an equally attractive faucet. Select a tall spout mounted on the countertop or vanity top or a wall-mounted spout to accommodate the height of the vessel. To minimize splashing, spouts should directly flow to the center of the vessel, not down the side. Make sure your faucet is compatible with your vessel choice. Look for a centerset or single-handle model if you'll be custom drilling the countertop—you only need to drill one faucet hole.

Tools & Materials ▸

Jigsaw
Trowel
Pliers
Wrench
Caulk gun and caulk
Sponge
Drill
Vanity or countertop
Vessel sink
Pop-up drain
P-trap and drain kit
Faucet
Phillips screwdriver

Vessel sinks are available in countless styles and materials, shapes, and sizes. Their one commonality is that they all need to be installed on a flat surface.

Vessel Sink Options

This glass vessel sink embedded in a "floating" glass countertop is a stunning contrast to the strong and attractive frame anchoring it to the wall.

The natural stone vessel sink blends elegantly into the stone countertop and is enhanced by the sleek faucet and round mirror.

The stone vessel sink is complemented by the wall-hung faucet. The rich wood vanity on which it's perched adds warmth to the room.

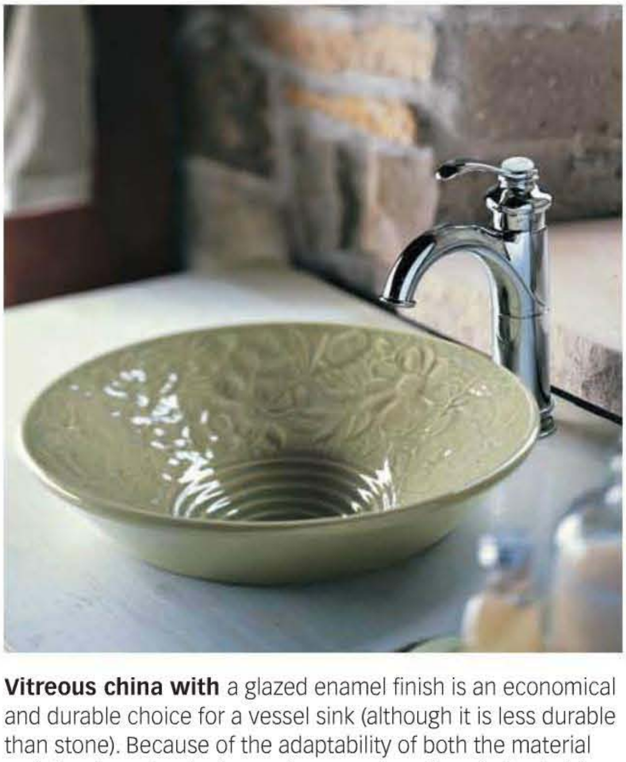

Vitreous china with a glazed enamel finish is an economical and durable choice for a vessel sink (although it is less durable than stone). Because of the adaptability of both the material and the glaze, the design options are virtually unlimited with vitreous china.

How to Install a Vessel Sink

Secure the vanity cabinet or other countertop that you'll be using to mount the vessel sink (see pages 464 to 465).

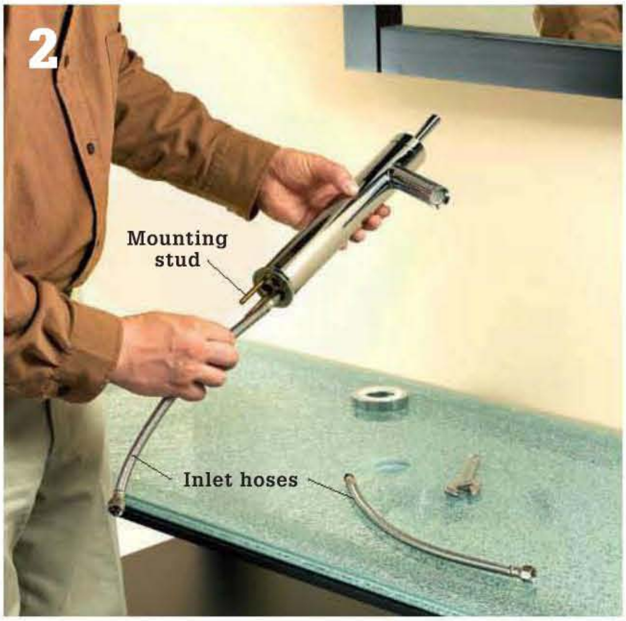

Begin hooking up the faucet. Insert the brass mounting stud into the threaded hole in the faucet base with the slotted end facing out. Hand tighten, and then use a screwdriver to tighten another half turn. Insert the inlet hoses into the faucet body and hand tighten. Use an adjustable wrench to tighten another half turn. Do not overtighten.

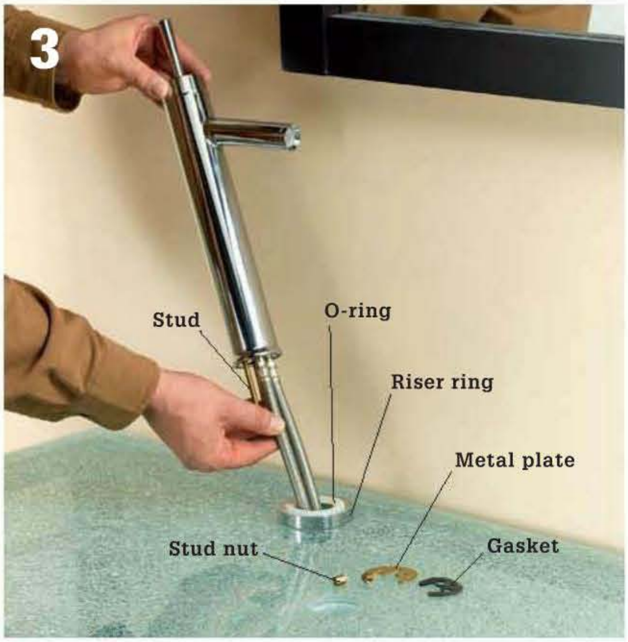

Place the O-ring on top of the riser-ring over the faucet cutout in the countertop. From underneath, slide the rubber gasket and the metal plate over the mounting stud. Thread the mounting stud nut onto the mounting stud and hand tighten. Use an adjustable wrench to tighten another half turn.

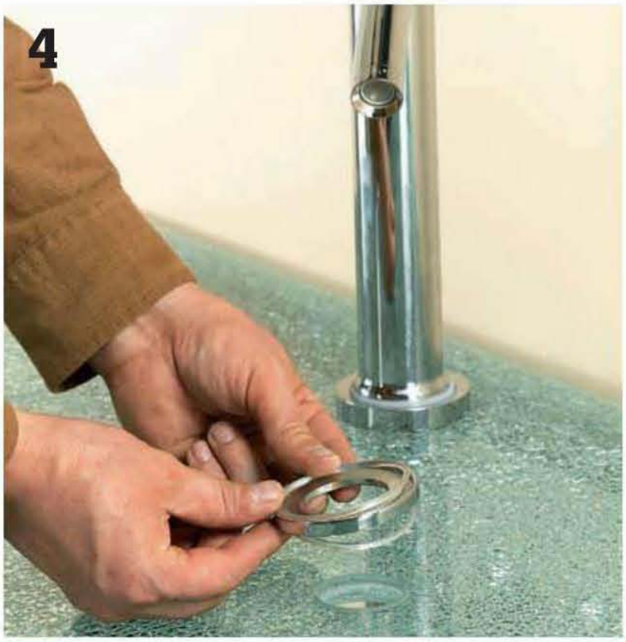

To install the sink and pop-up drain, first place the small metal ring between two O-rings and place over the drain cutout.

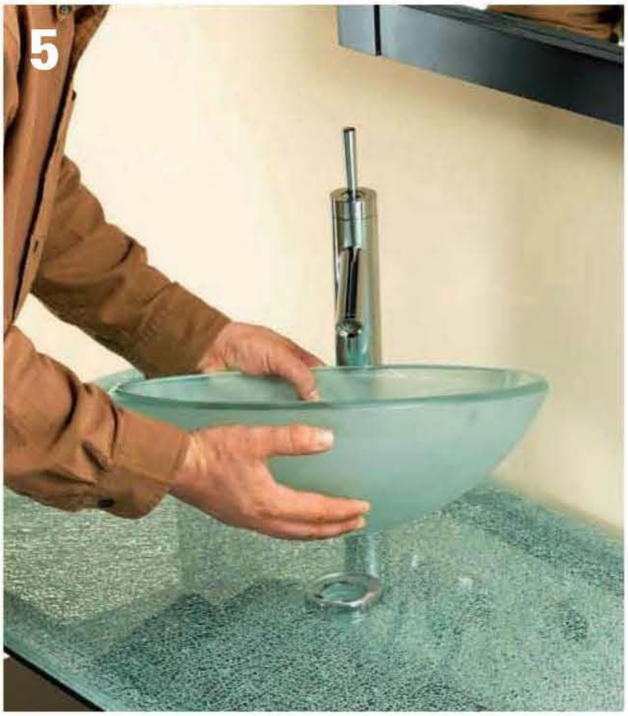

Place the vessel bowl on top of the O-rings. In this installation, the vessel is not bonded to the countertop.

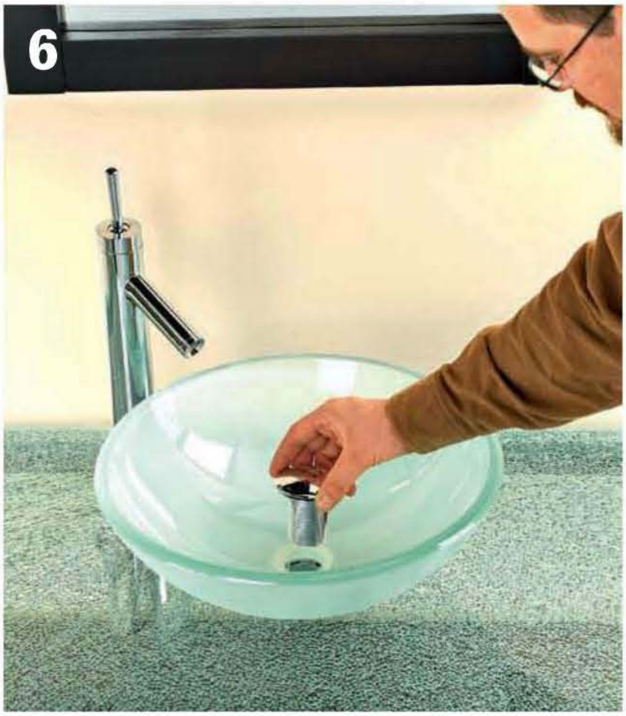

Put the small rubber gasket over the drain hole in the vessel. From the top, push the pop-up assembly through the drain hole.

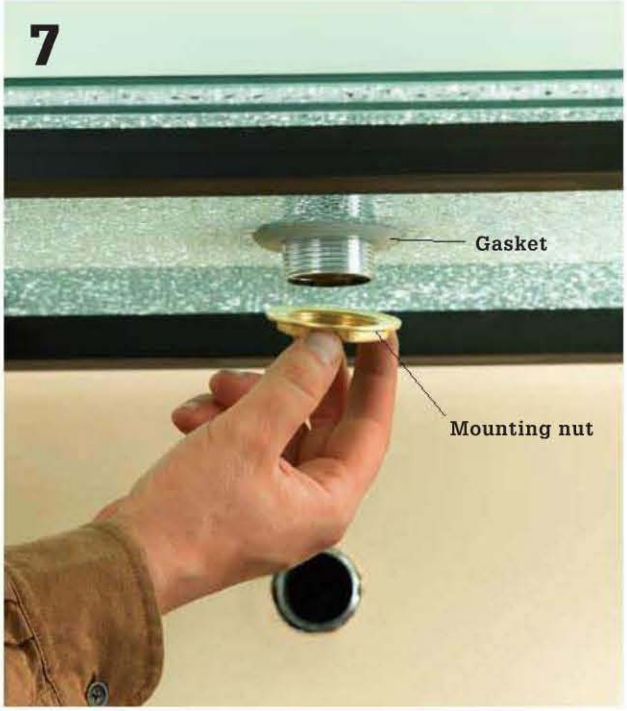

From underneath, push the large rubber gasket onto the threaded portion of the pop-up assembly. Thread the nut onto the pop-up assembly and tighten. Use an adjustable wrench or basin wrench to tighten an additional half turn. Thread the tailpiece onto the pop-up assembly.

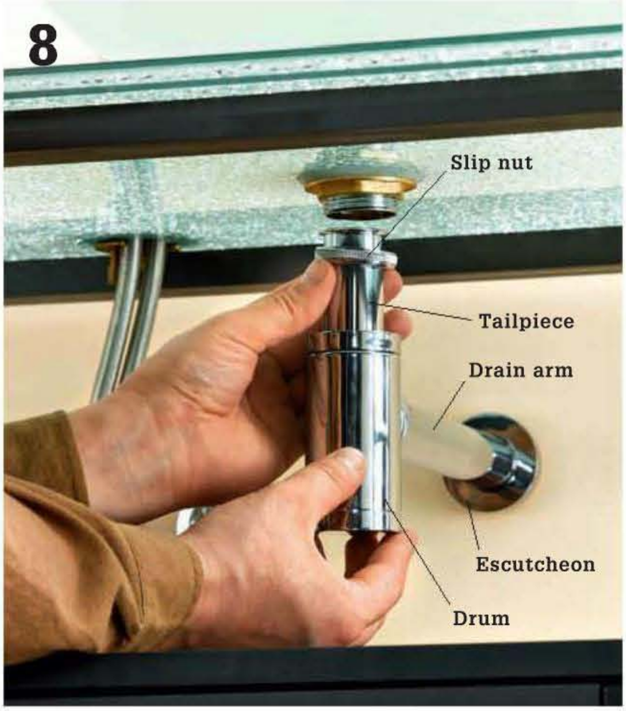

Install the drum trap. Loosen the rings on the top and outlet of the drum trap. Slide the drum trap top hole over the tailpiece. Slide the drain arm into the side outlet, with the flat side of the rubber gasket facing away from the trap. Insert the drain arm into the wall outlet. Hand tighten the rings.

Installing Pedestal Sinks

Pedestal sinks move in and out of popularity more frequently than other sink types, but even during times they aren't particularly trendy, they retain fairly stable demand. You'll find them most frequently in half baths, where their small footprint makes them an efficient choice. Designers are also discovering the appeal of tandem pedestal sinks of late, where the smaller profiles allow for his-and-hers sinks that don't dominate visually.

The primary drawback to pedestal sinks is that they don't offer any storage. Their chief practical benefit is that they conceal plumbing some homeowners would prefer not to see.

Pedestal sinks are mounted in two ways. Most of the more inexpensive ones you'll find at home stores are hung in the manner of wall-hung sinks. The pedestal is actually installed after the sink is hung and its purpose is only decorative. But other pedestal sinks (typically on the higher end of the design scale) have structurally important pedestals that bear most or all of the weight of the sink.

Pedestal sinks are available in a variety of styles and are a perfect fit for small half baths. They keep plumbing hidden, lending a neat, contained look to the bathroom.

How to Install a Pedestal Sink

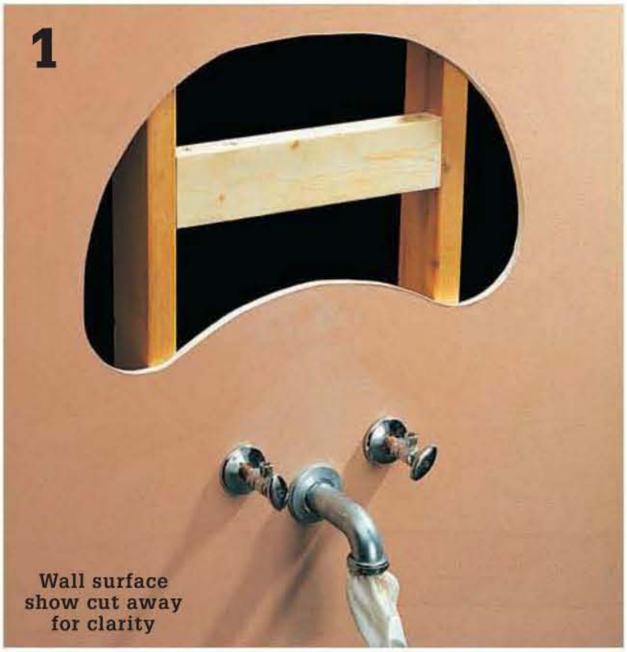

Install 2 × 4 blocking between the wall studs, behind the planned sink location. Cover the wall with water-resistant drywall. Waste and supply lines may need to be moved, depending on the sink.

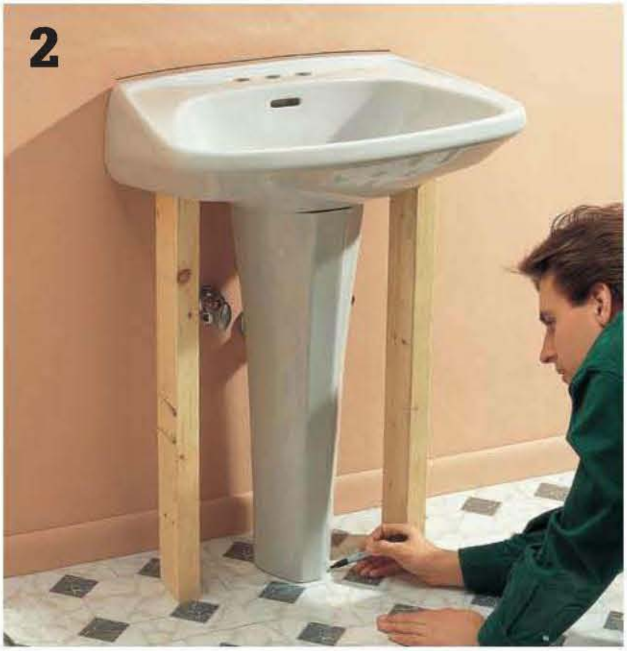

Set the basin and pedestal in position and brace it with 2 × 4s. Outline the top of the basin on the wall, and mark the base of the pedestal on the floor. Mark reference points on the wall and floor through the mounting holes found on the back of the sink and the bottom of the pedestal.

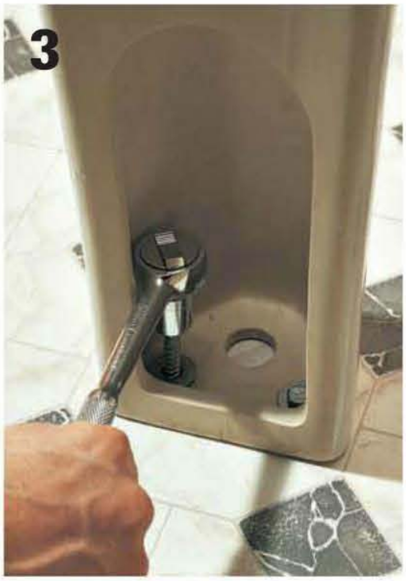

Set aside the basin and pedestal. Drill pilot holes in the wall and floor at the reference points, then reposition the pedestal. Anchor the pedestal to the floor with lag screws.

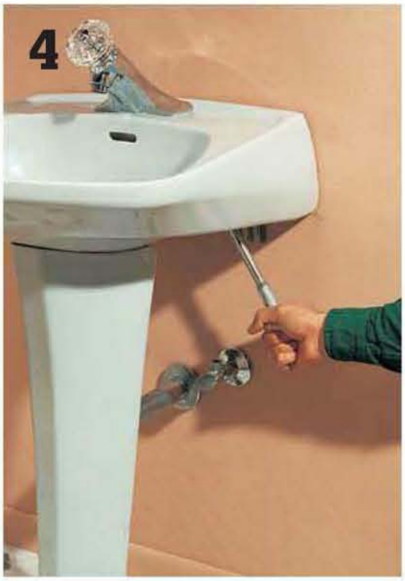

Attach the faucet, then set the sink on the pedestal. Align the holes in the back of the sink with the pilot holes drilled in the wall, then drive lag screws and washers into the wall brace using a ratchet wrench. Do not overtighten the screws.

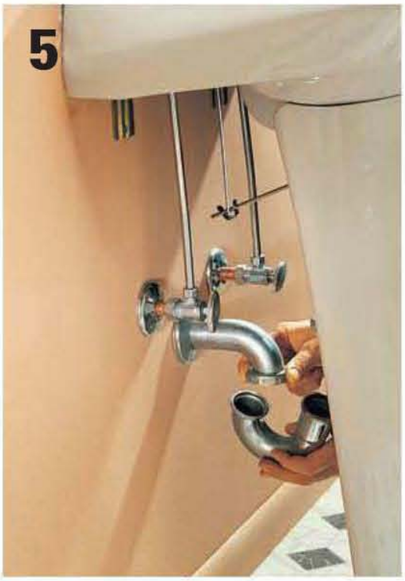

Hook up the drain and supply fittings. Caulk between the back of the sink and the wall when installation is finished.

Installing Vanity Cabinets

Most bathroom countertops installed today are integral (one-piece) sink-countertop units made from cultured marble or other solid materials, like solid surfacing. Integral sink-countertops are convenient, and many are inexpensive, but style and color options are limited.

Some remodelers and designers still prefer the distinctive look of a custom-built countertop with a self-rimming sink basin, which gives you a much greater selection of styles and colors. Installing a self-rimming sink is very simple.

Integral sink-countertops are made in standard sizes to fit common vanity widths. Because the sink and countertop are cast from the same material, integral sink-countertops do not leak, and do not require extensive caulking and sealing.

Tools & Materials ▸

- Pencil
- Scissors
- Carpenter's level
- Screwdriver
- Channel-type pliers
- Ratchet wrench
- Basin wrench
- Cardboard
- Masking tape
- Plumber's putty
- Lag screws
- Tub and tile caulk
- Plumber's putty

How to Install a Vanity Cabinet

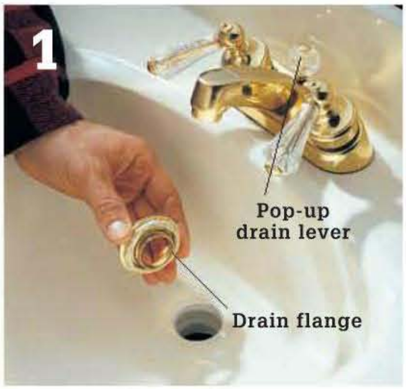

Set the sink-countertop unit onto sawhorses. Attach the faucet and slip the drain lever through the faucet body. Place a ring of plumber's putty around the drain flange, then insert the flange in the drain opening.

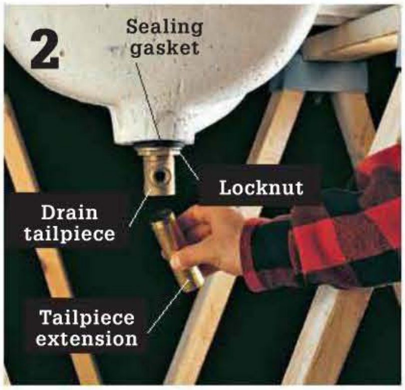

Thread the locknut and sealing gasket onto the drain tailpiece, then insert the tailpiece into the drain opening and screw it onto the drain flange. Tighten the locknut securely. Attach the tailpiece extension. Insert the pop-up stopper linkage.

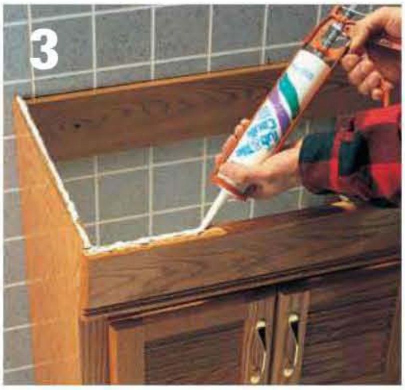

Apply a layer of tub and tile caulk (or adhesive, if specified by the countertop manufacturer) to the top edges of the cabinet vanity, and to any corner braces.

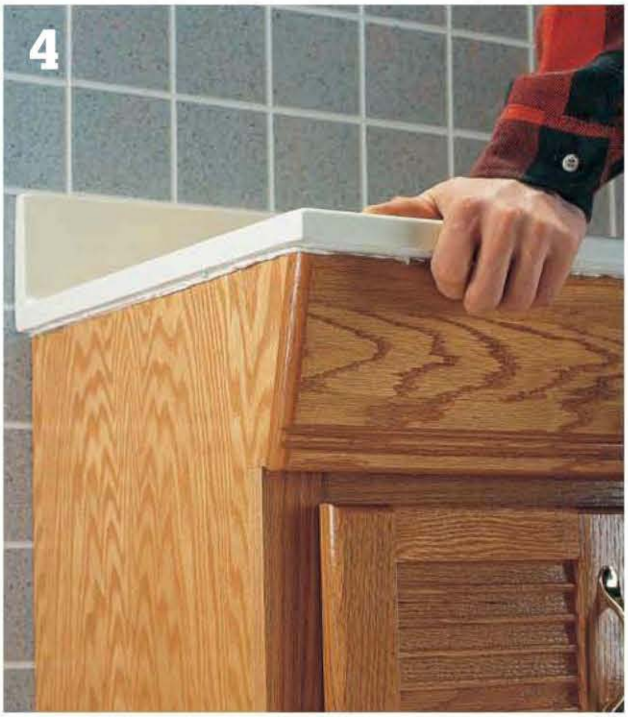

Center the sink-countertop unit over the vanity so the overhang is equal on both sides and the backsplash of the countertop is flush with the wall. Press the countertop evenly into the caulk.

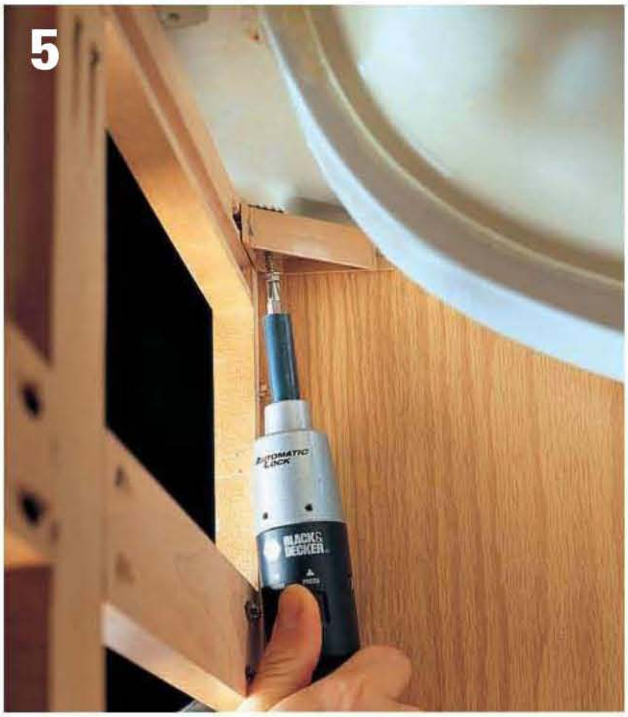
5
AUTOMATIC LOCK
BLACK & DECKER

Cabinets with corner braces: Secure the countertop to the cabinet by driving a mounting screw through each corner brace and up into the countertop. *Note: Cultured marble and other hard countertops require predrilling and a plastic screw sleeve.*

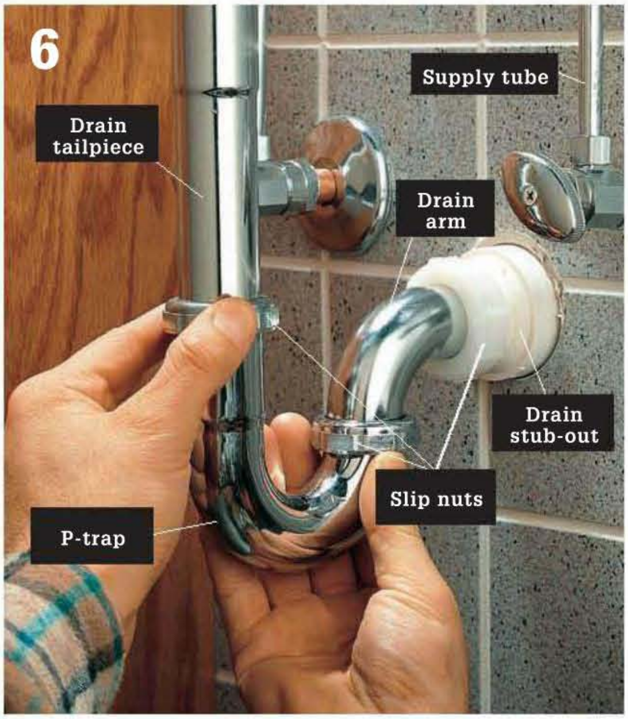
6
Supply tube
Drain tailpiece
Drain arm
Drain stub-out
Slip nuts
P-trap

Attach the drain arm to the drain stub-out in the wall, using a slip nut. Attach one end of the P-trap to the drain arm, and the other to the tailpiece of the sink drain, using slip nuts. Connect supply tubes to the faucet tailpieces.

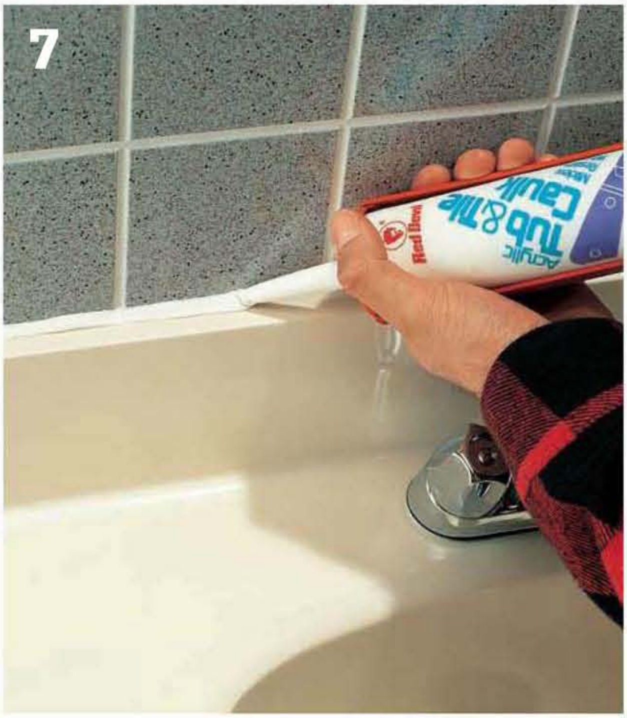
7
Red Devil
Acrylic Tub & Tile Caulk

Seal the gap between the backsplash and the wall using tub and tile caulk.

Installing Bathroom Sink Faucets & Drains

One-piece faucets, with either one or two handles, are the most popular fixtures for bathroom installations.

"Widespread" faucets with separate spout and handles are being installed with increasing frequency, however. Because the handles are connected to the spout with flex tubes that can be 18" or longer, widespread faucets can be arranged in many ways.

Tools & Materials ▸

- Hacksaw or tin snips
- Channel-type pliers
- Pliers
- Basin wrench
- Adjustable wrench
- Screwdriver
- Plumber's putty
- Teflon tape
- Faucet kit
- Pipe joint compound
- Flexible supply tubes
- Heat-proof grease
- Loctite (adhesive)

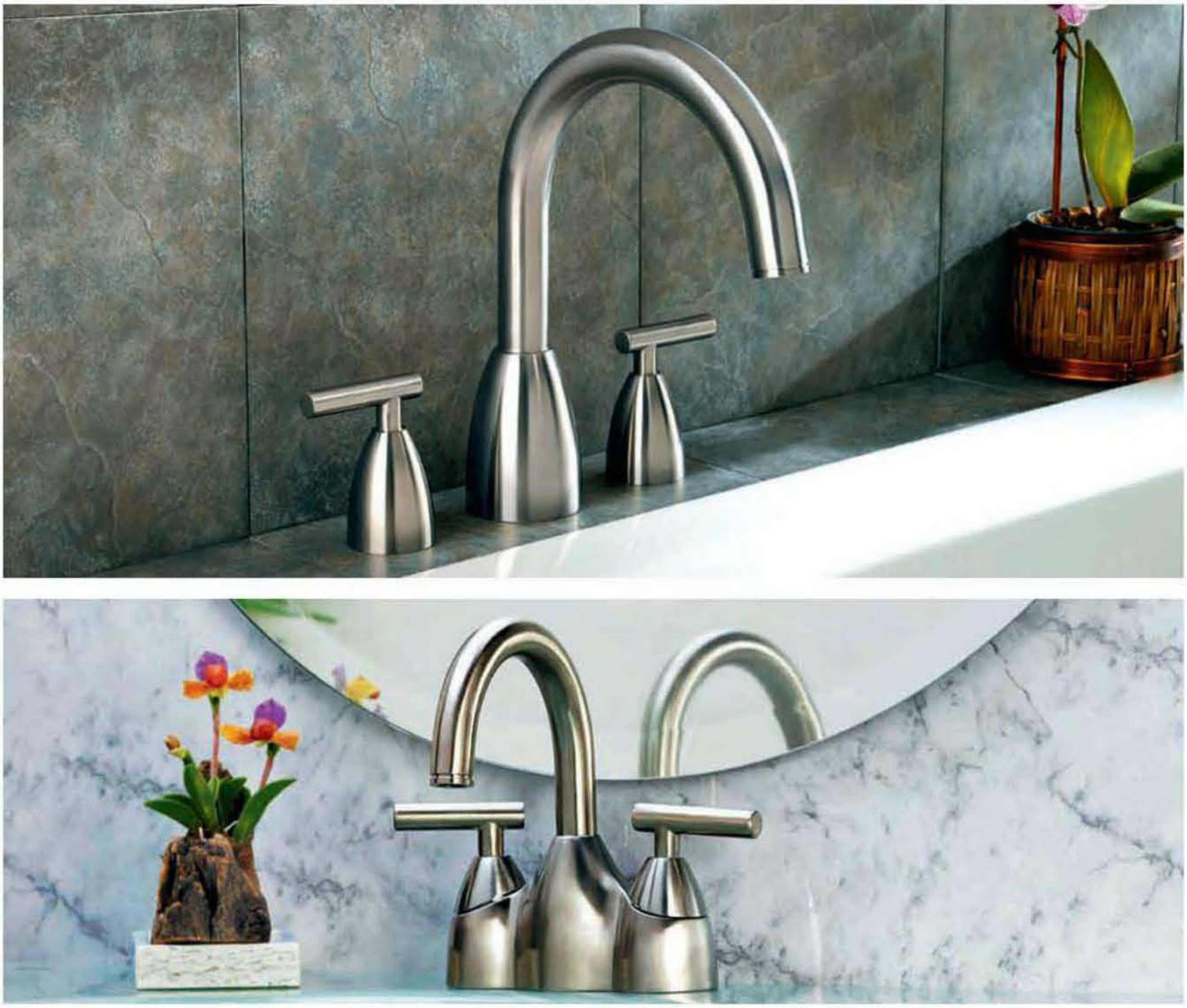

Bathroom sink faucets come in two basic styles: the widespread with independent handles and spout (top); and the single-body, deck-mounted version (bottom).

Bathroom Faucet & Drain Hookups

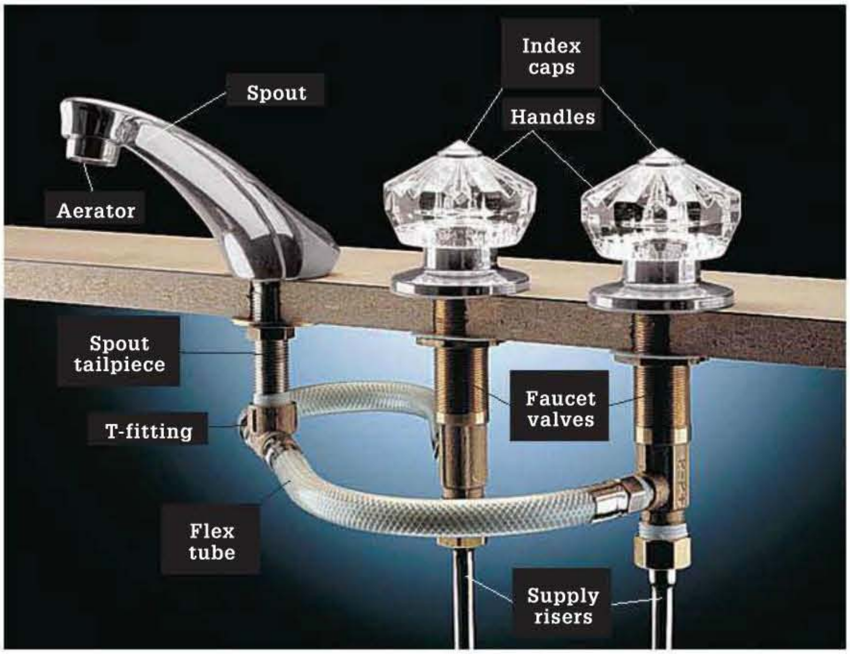
Index caps
Spout
Handles
Aerator
Spout tailpiece
Faucet valves
T-fitting
Flex tube
Supply risers

Widespread lavatory faucets have valves that are independent from the spout so they can be configured however you choose, provided that your flex tube connectors are long enough to span the distance.

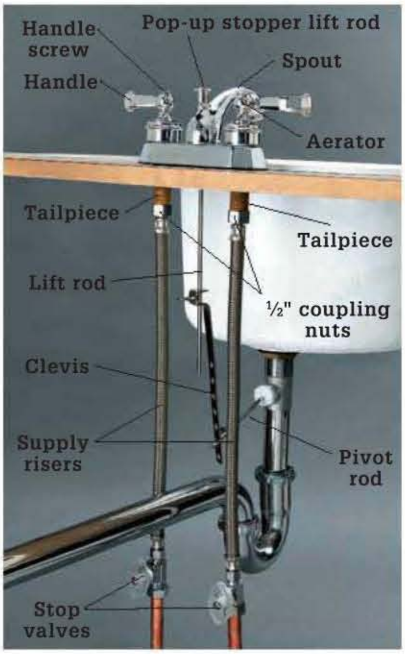
Handle screw
Pop-up stopper lift rod
Spout
Handle
Aerator
Tailpiece
Tailpiece
Lift rod
½" coupling nuts
Clevis
Supply risers
Pivot rod
Stop valves

Single-body lavatory faucets have both valves and the spout permanently affixed to the faucet body. They do not offer flexibility in configurations, but they are very simple to install.

The pop-up stopper fits into the drain opening so the stopper will close tightly against the drain flange when the pop-up handle is lifted up.

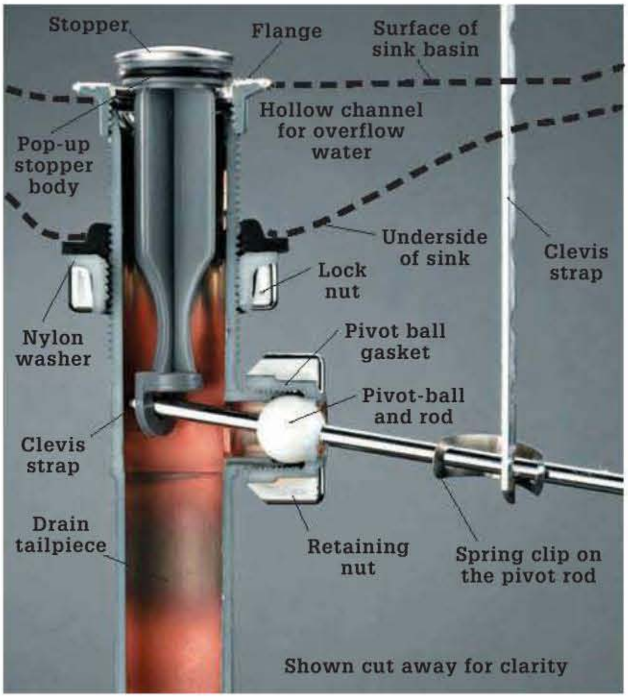
Stopper
Flange
Surface of sink basin
Hollow channel for overflow water
Pop-up stopper body
Underside of sink
Lock nut
Clevis strap
Nylon washer
Pivot ball gasket
Pivot-ball and rod
Clevis strap
Drain tailpiece
Retaining nut
Spring clip on the pivot rod
Shown cut away for clarity

The linkage that connects the pop-up stopper to the pop-up handle fits into a male-threaded port in the drain tailpiece. Occasionally the linkage will require adjustment or replacement.

How to Install a Widespread Faucet

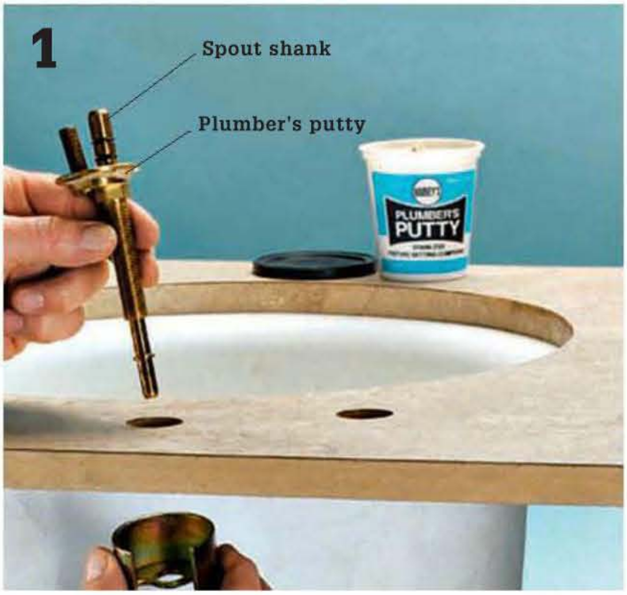
1
Spout shank
Plumber's putty
PLUMBER'S PUTTY

Insert the shank of the faucet spout through one of the holes in the sink deck (usually the center hole but you can offset it in one of the end holes if you prefer). If the faucet is not equipped with seals or O-rings for the spout and handles, pack plumber's putty on the undersides before inserting the valves into the deck. *Note: If you are installing the widespread faucet in a new sink deck, drill three holes of the size suggested by the faucet manufacturer.*

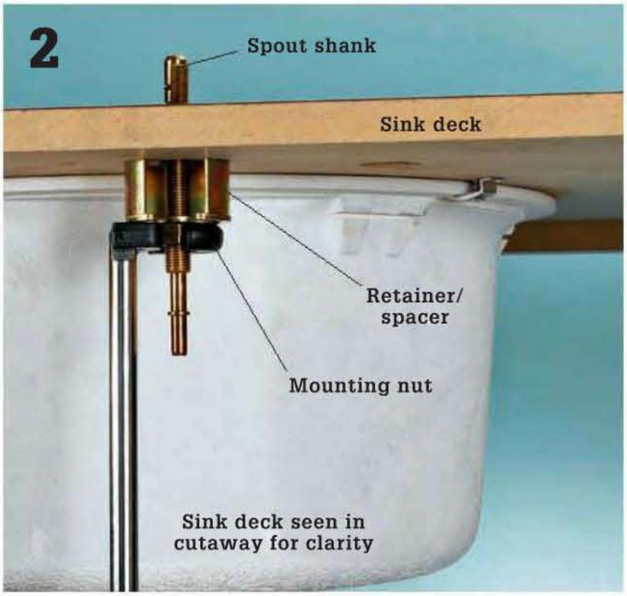
2
Spout shank
Sink deck
Retainer/
spacer
Mounting nut
Sink deck seen in
cutaway for clarity

In addition to mounting nuts, many spout valves for widespread faucets have an open-retainer fitting that goes between the underside of the deck and the mounting nut. Others have only a mounting nut. In either case, tighten the mounting nut with pliers or a basin wrench to secure the spout valve. You may need a helper to keep the spout centered and facing forward.

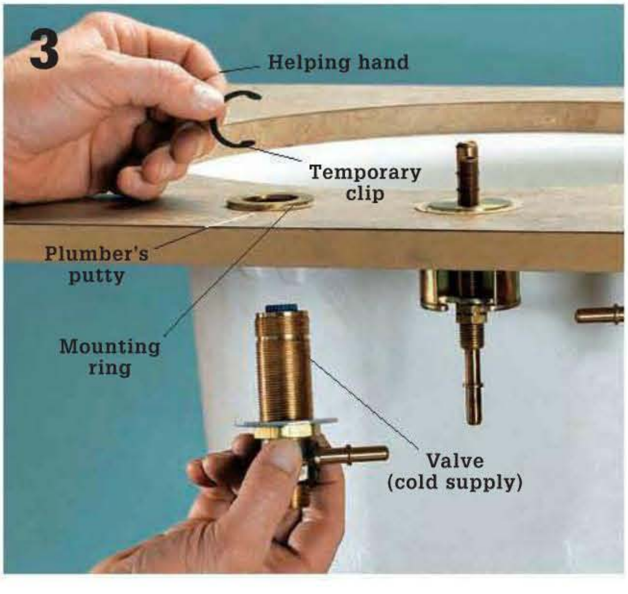
3
Helping hand
Temporary
clip
Plumber's
putty
Mounting
ring
Valve
(cold supply)

Mount the valves to the deck using whichever method the manufacturer specifies (it varies quite a bit). In the model seen here, a mounting ring is positioned over the deck hole (with plumber's putty seal) and the valve is inserted from below. A clip snaps onto the valve from above to hold it in place temporarily (you'll want a helper for this).

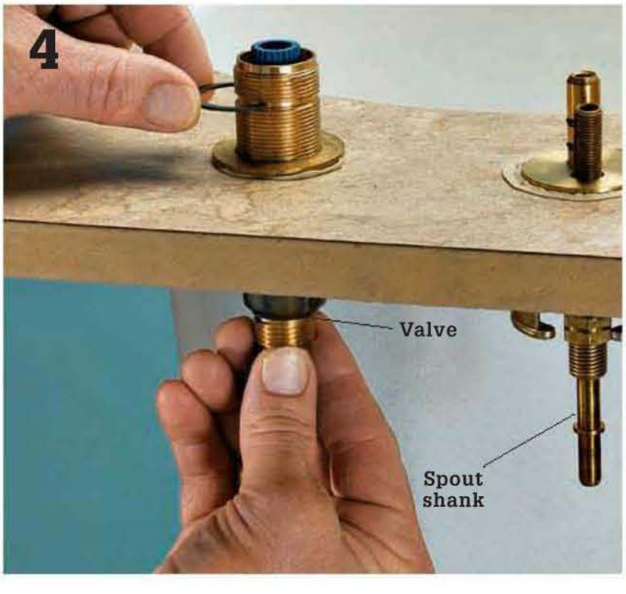
4
Valve
Spout
shank

From below, thread the mounting nuts that secure the valves to the sink deck. Make sure the cold water valve (usually has a blue cartridge inside) is in the right-side hole (from the front) and the hot water valve (red cartridge) is in the left hole (from the front). Install both valves.

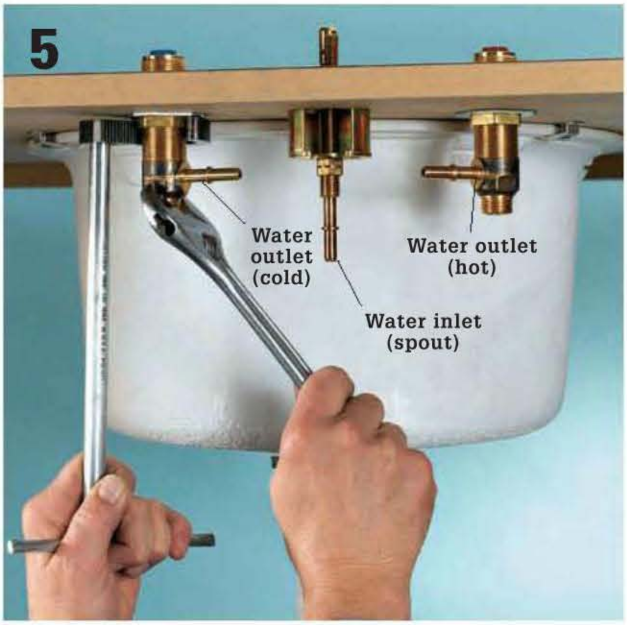
5
Water outlet (cold)
Water outlet (hot)
Water inlet (spout)

Once you've started the nut on the threaded valve shank, secure the valve with a basin wrench, squeezing the lugs where the valve fits against the deck. Use an adjustable wrench to finish tightening the lock nut onto the valve. The valves should be oriented so the water outlets are aimed at the inlet on the spout shank.

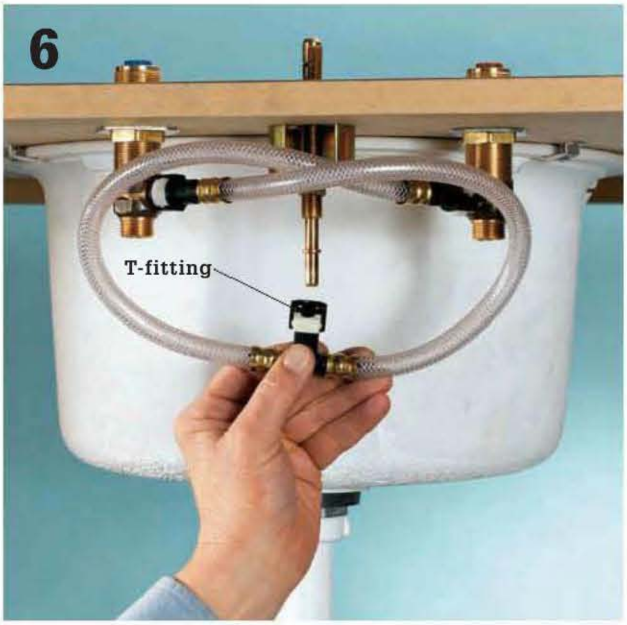

Attach the flexible supply tubes (supplied with the faucet) to the water outlets on the valves. Some twist onto the outlets, but others (like the ones above) click into place. The supply hoses meet in a T-fitting that is attached to the water inlet on the spout.

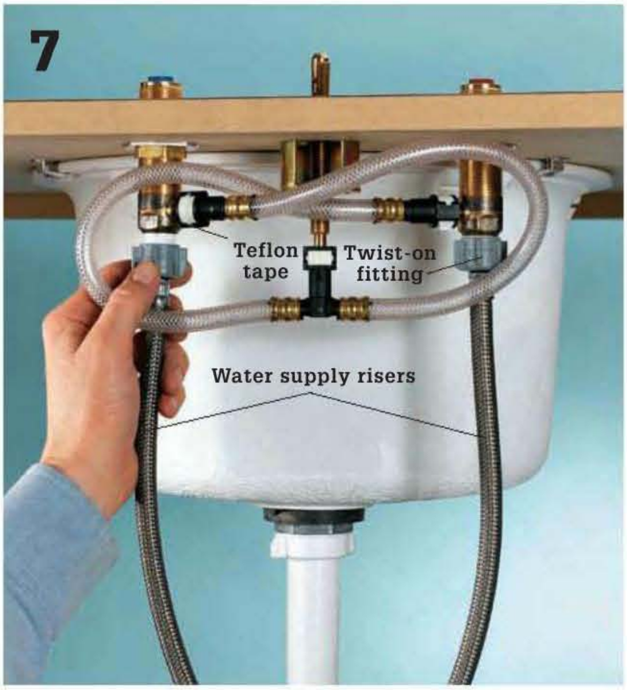
7
Teflon tape
Twist-on fitting
Water supply risers

Attach flexible braided-metal supply risers to the water stop valves and then attach the tubes to the inlet port on each valve (usually with Teflon tape and a twist-on fitting at the valve end of the supply riser).

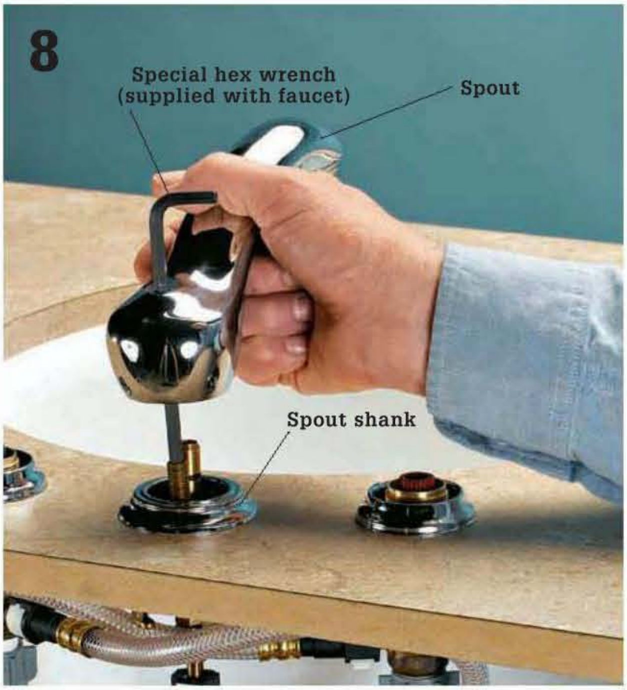
8
Special hex wrench (supplied with faucet)
Spout
Spout shank

Attach the spout. The model shown here comes with a special hex wrench that is threaded through the hole in the spout where the lift rod for the pop-up drain will be located. Once the spout is seated cleanly on the spout shank, tighten the hex wrench to secure the spout. Different faucets will use other methods to secure the spout to the shank.

(continued)

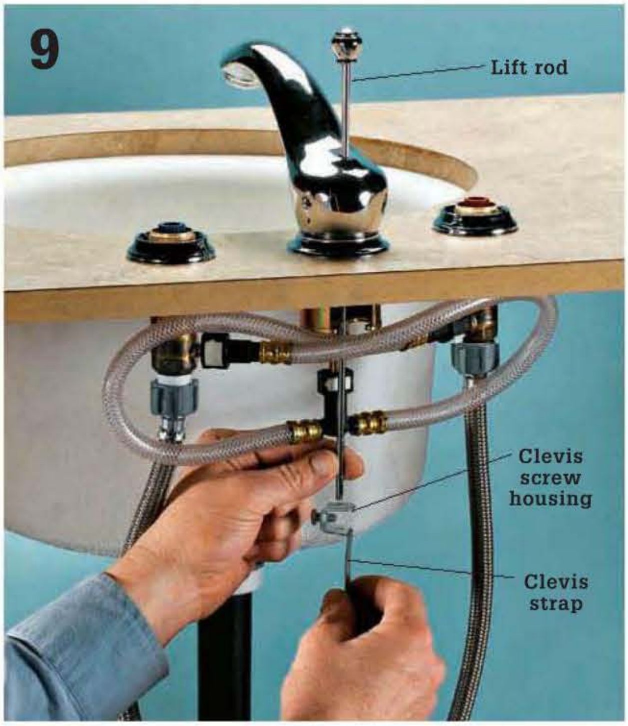

If your sink did not have a pop-up stopper, you'll need to replace the sink drain tailpiece with a pop-up stopper body (often supplied with the faucet). See page 473. Insert the lift rod through the hole in the back of the spout and, from below, thread the pivot rod through the housing for the clevis screw.

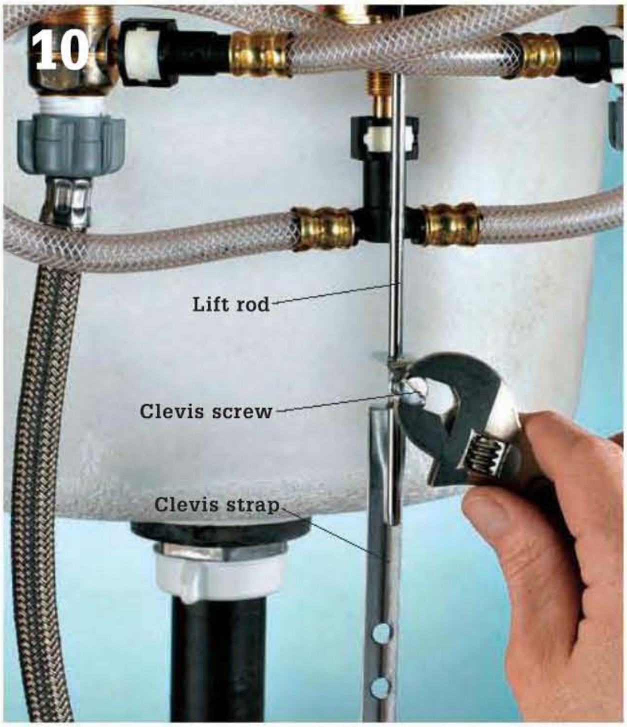

Attach the clevis strap to the pivot rod that enters the pop-up drain body, and adjust the position of the strap so it raises and lowers properly when the lift rod is pulled up. Tighten the clevis screw at this point. It's hard to fit a screwdriver in here, so you may need to use a wrench or pliers.

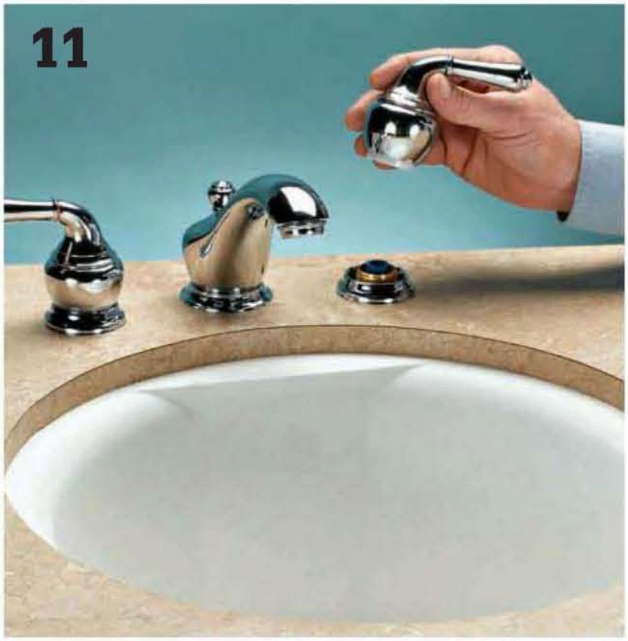

Attach the faucet handles to the valves using whichever method is required by the faucet manufacturer. Most faucets are designed with registration methods to ensure that the handles are symmetrical and oriented in an ergonomic way once you secure them to the valves.

Turn on the water supply and test the faucet. Remove the faucet aerator so any debris in the lines can clear the spout.

Variation: How to Install a Single-body Faucet ▸

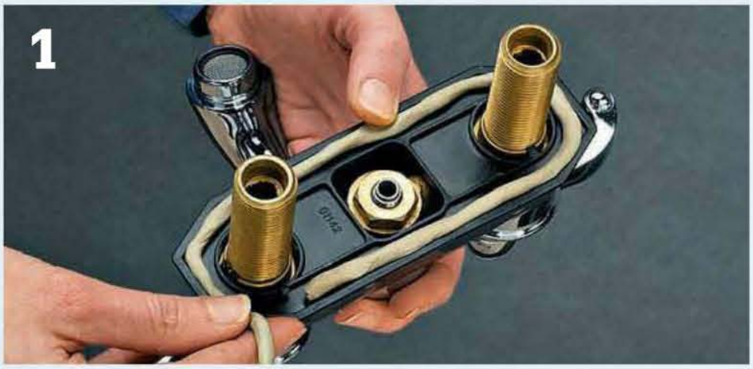

Most faucets come with a plastic or foam gasket to seal the bottom of the faucet to the sink deck. These gaskets will not always form a watertight seal. If you want to ensure no water gets below the sink, discard the seal and press a ring of plumber's putty into the sealant groove built into the underside of the faucet body.

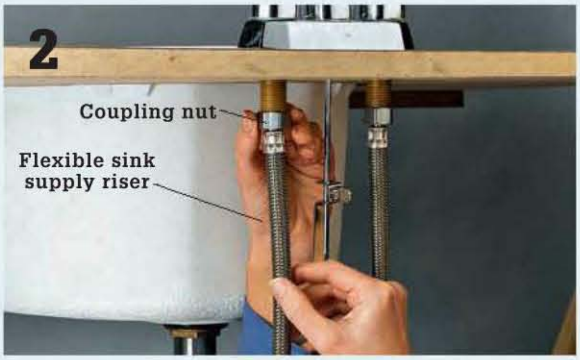

Insert the faucet tailpieces through the holes in the sink. From below, thread washers and mounting nuts over the tailpieces, then tighten the mounting nuts with a basin wrench until snug. Put a dab of pipe joint compound on the threads of the stop valves and thread the metal nuts of the flexible supply risers to these. Wrench tighten about a half-turn past hand tight. Overtightening these nuts will strip the threads. Now tighten the coupling nuts to the faucet tailpieces with a basin wrench.

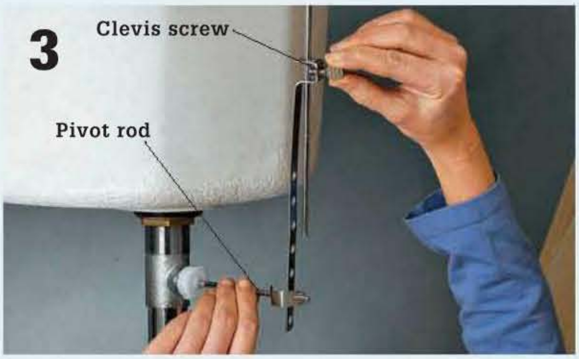

Slide the lift rod of the new faucet into its hole behind the spout. Thread it into the clevis past the clevis screw. Push the pivot rod all the way down so the stopper is open. With the lift rod also all the way down, tighten the clevis to the lift rod.

Grease the fluted valve stems with heatproof grease, then put the handles in place. Put a drop of Loctite on each handle screw before tightening it on. (This will keep your handles from coming loose.) Cover each handle screw with the appropriate index cap—Hot or Cold.

Unscrew the aerator from the end of the spout. Turn the hot and cold water taps on full. Turn the water back on at the stop valves and flush out the faucet for a couple of minutes before turning off the water at the faucet. Check the riser connections for drips. Tighten a compression nut only until the drip stops.

How to Install a Pop-up Drain

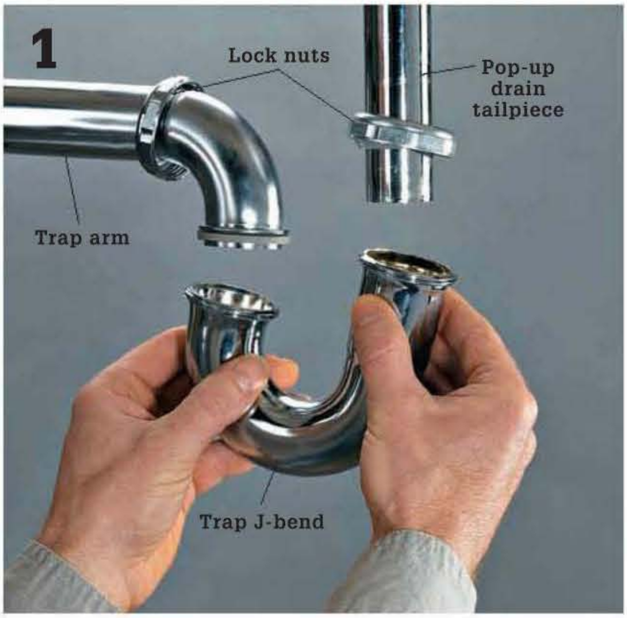
1
Lock nuts
Pop-up drain tailpiece
Trap arm
Trap J-bend

Put a basin under the trap to catch water. Loosen the nuts at the outlet and inlet to the trap J-bend by hand or with channel-type pliers and remove the bend. The trap will slide off the pop-up body tailpiece when the nuts are loose. Keep track of washers and nuts and their up/down orientation by leaving them on the tubes.

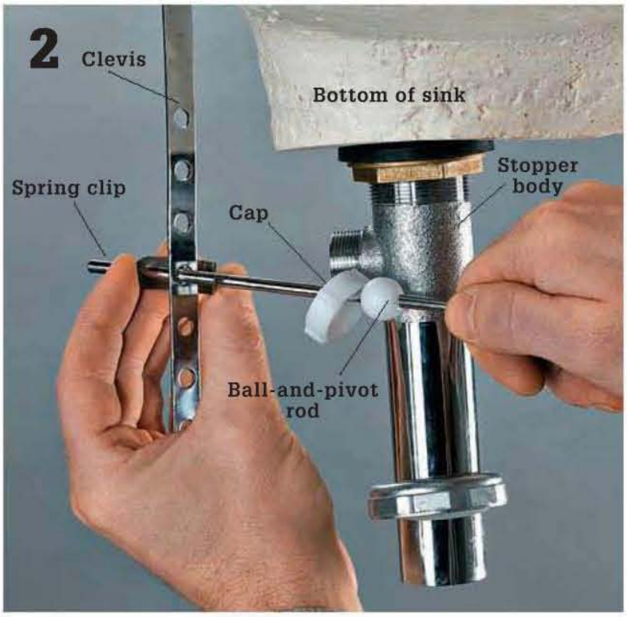
2
Clevis
Bottom of sink
Stopper body
Spring clip
Cap
Ball-and-pivot rod

Unscrew the cap holding the ball-and-pivot rod in the pop-up body and withdraw the ball. Compress the spring clip on the clevis and withdraw the pivot rod from the clevis.

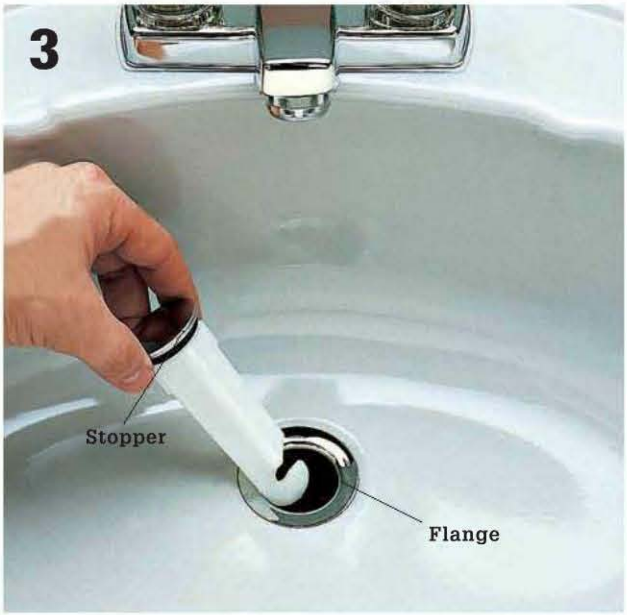
3
Stopper
Flange

Remove the pop-up stopper. Then, from below, remove the lock nut on the stopper body. If needed, keep the flange from turning by inserting a large screwdriver in the drain from the top. Thrust the stopper body up through the hole to free the flange from the basin, and then remove the flange and the stopper body.

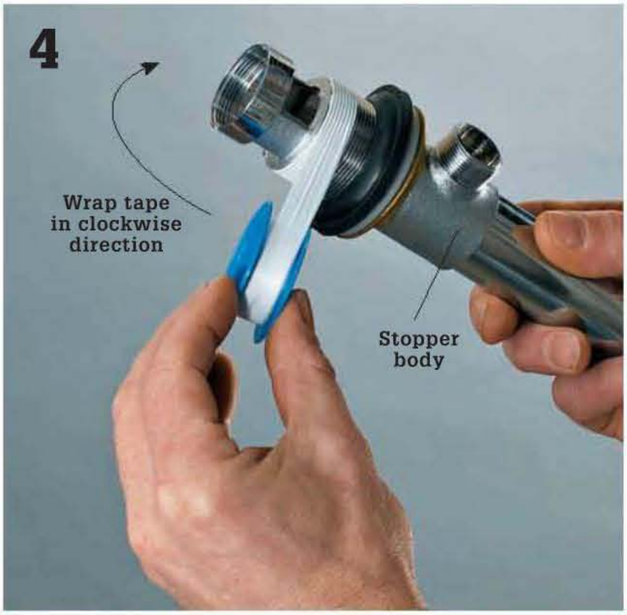
4
Wrap tape in clockwise direction
Stopper body

Clean the drain opening above and below, and then thread the locknut all the way down the new pop-up body, followed by the flat washer and the rubber gasket (beveled side up). Wrap three layers of Teflon tape clockwise onto the top of the threaded body. Make a ½"-dia. snake from plumber's putty, form it into a ring, and stick the ring underneath the drain flange.

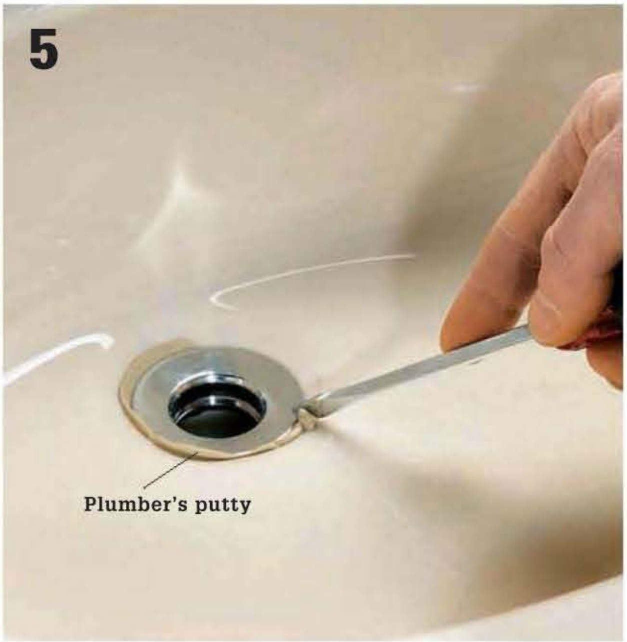

From below, face the pivot rod opening directly back toward the middle of the faucet and pull the body straight down to seat the flange. Thread the locknut/washer assembly up under the sink, then fully tighten the locknut with channel-type pliers. Do not twist the flange in the process, as this can break the putty seal. Clean off the excess plumber's putty from around the flange.

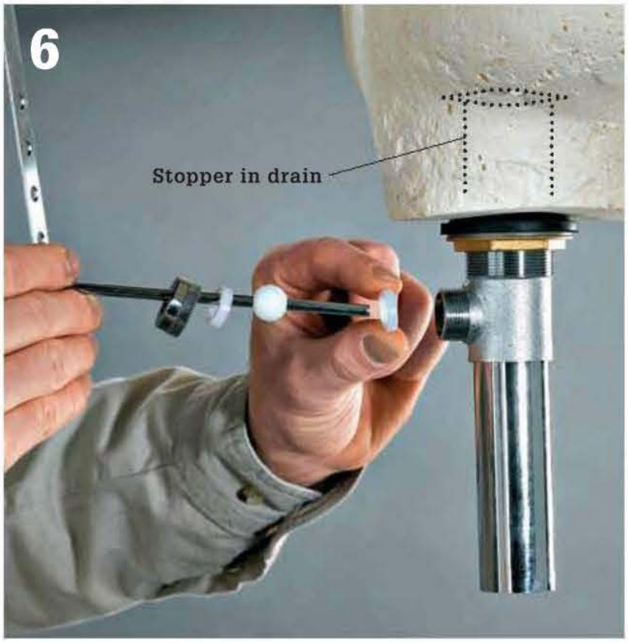

Drop the pop-up stopper into the drain hole so the hole at the bottom of its post is closest to the back of the sink. Put the beveled nylon washer into the opening in the back of the pop-up body with the bevel facing back.

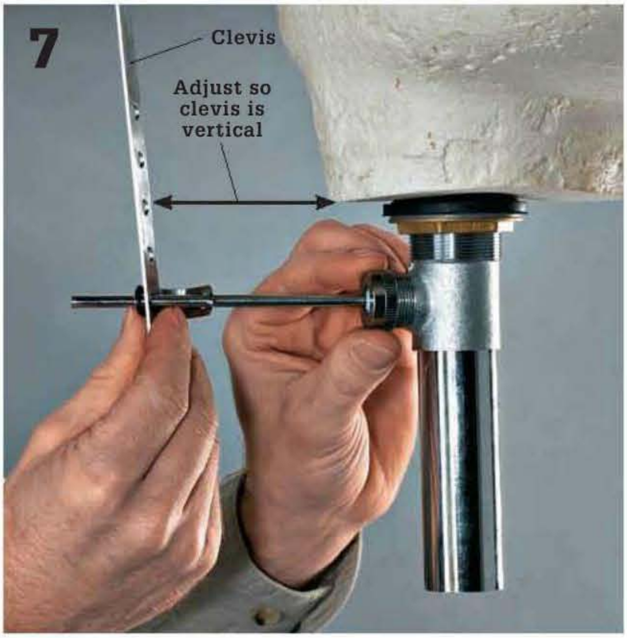

Put the cap behind the ball on the pivot rod as shown. Sandwich a hole in the clevis with the spring clip and thread the long end of the pivot rod through the clip and clevis. Put the ball end of the pivot rod into the pop-up body opening and into the hole in the the stopper stem. Screw the cap on to the pop-up body over the ball.

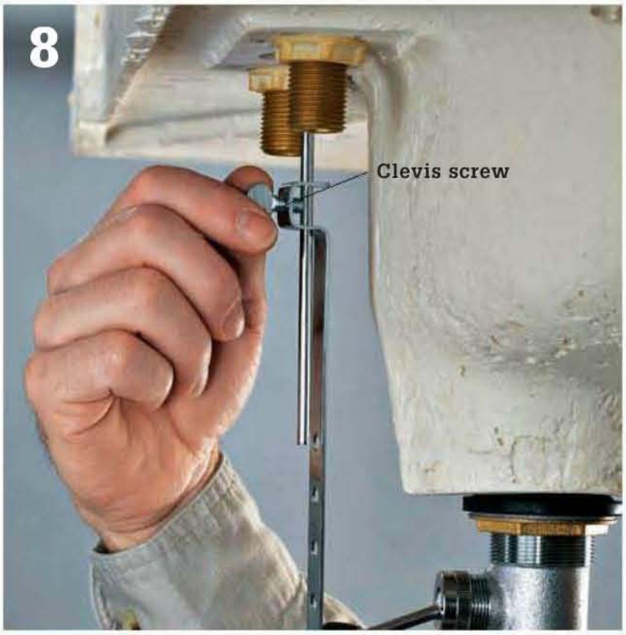

Loosen the clevis screw holding the clevis to the lift rod. Push the pivot rod all the way down (which fully opens the pop-up stopper). With the lift rod also all the way down, tighten the clevis screw to the rod. If the clevis runs into the top of the trap, cut it short with your hacksaw or tin snips. Reassemble the J-bend trap.

Installing Tub Spouts

In many situations, replacing a bathtub spout can be almost as easy as hooking up a garden hose to an outdoor spigot. There are some situations where it is a bit more difficult, but still pretty simple. The only time it's a real problem is when the spout is attached to a plain copper supply nipple, rather than a threaded nipple. You'll know this is the case if the spout has a setscrew on the underside where it meets the wall. Many bathtub spouts are sold in kits with a matching showerhead and handle or handles. But for a simple one-for-one replacement, spouts are sold separately. You just need to make sure the new spout is compatible with the existing nipple.

Tub spouts can be relatively complicated plumbing fittings, often performing three or four important functions. The spout itself is simple enough, since its only function is to deliver bathwater to the tub. But the diverter network and pop-up drain contain multiple moving parts that require precise adjustment and occasional repair or replacement (see photo, next page). The diverter is basically a stop valve that's activated by a lever or knob to block flow of water from the spout, forcing it up to a showerhead or out through a handheld showerhead, as seen here.

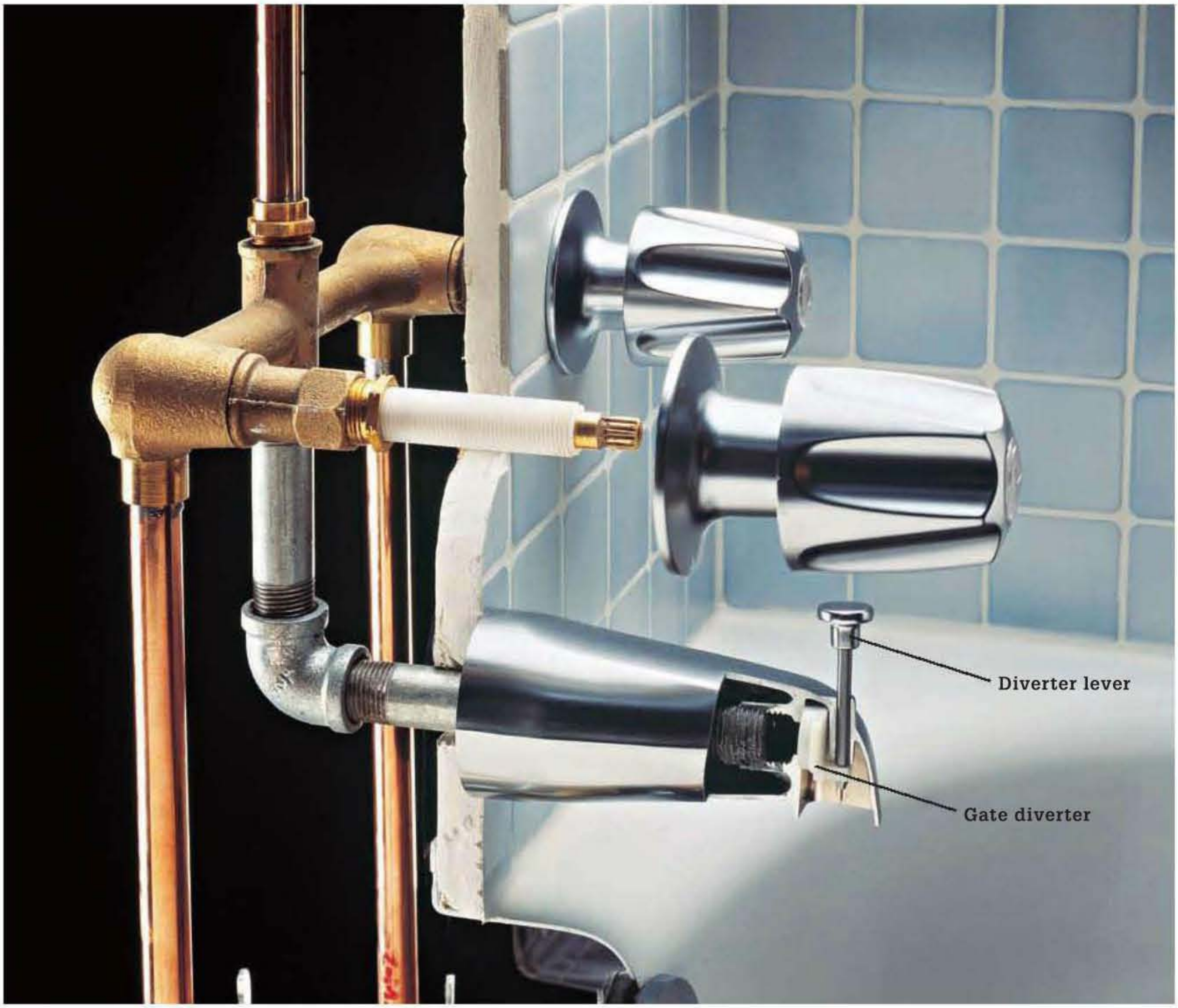

In many bathtub/shower plumbing systems, the spout has the important job of housing the diverter—a gate inside the spout that is operated by a lever with a knob for pulling. An open gate allows water to come out of the spout when the faucet is turned on. When the diverter is pulled shut, the water is redirected up a riser pipe and to the showerhead. Failure of the diverter is one of the most common reasons for replacing a spout.

Tools & Materials ▸

- Pliers
- Adjustable pliers
- Channel-type pliers
- Rags
- Masking tape
- Screwdrivers
- Allen wrenches
- Replacement spout
- Measuring tape
- Teflon tape
- Utility knife

Installation Tip ▸

If you are installing a brand new tub/shower faucet, it likely contains an antiscald device. Most of these devices work by sensing a decrease in cold water flow and compensating by stemming the hot water flow too. As you install the new plumbing, be aware that faucets with antiscald protection will not function at all if both water supply tubes are not turned on at the stop valve. In other words: you can't test the hot and cold supply independently.

How to Install a Slip-fit Spout

1

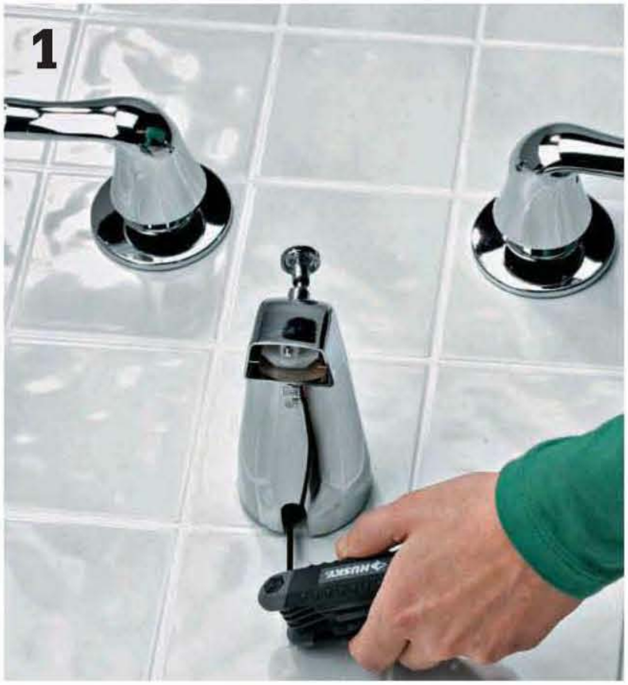

Slip fitting: Check underneath the tub spout to look for an access slot or cutout, which indicates the spout is a slip-fit style that is held in place with a setscrew and mounted on a copper supply nipple. Loosen the screw with a hex (Allen) wrench. Pull off the spout.

2

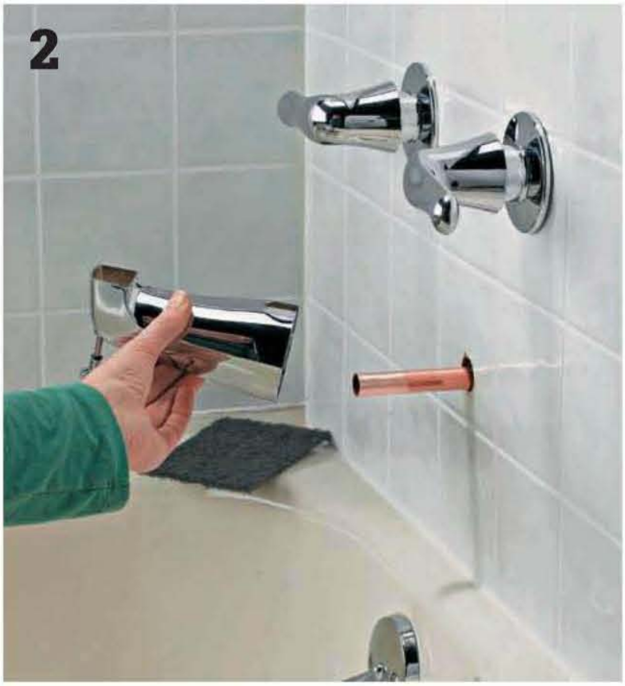

Clean the copper nipple with steel wool. If you find any sharp edges where the nipple was cut, smooth them out with emery paper. Then, insert the O-ring that comes with the spout onto the nipple (see the manufacturer's instructions) and slide the spout body over the nipple in an upside-down position.

3

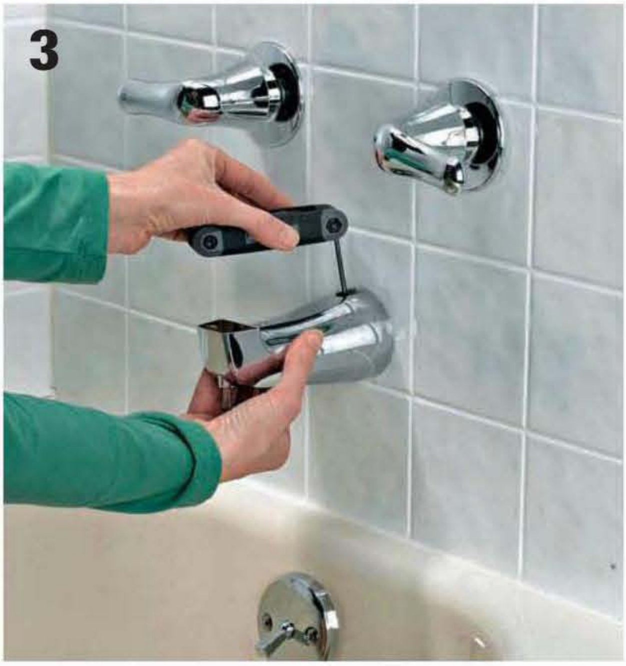

With the spout upside down for ease of access, tighten the setscrews on the clamp, working through the access slot or cutout, until you feel resistance.

4

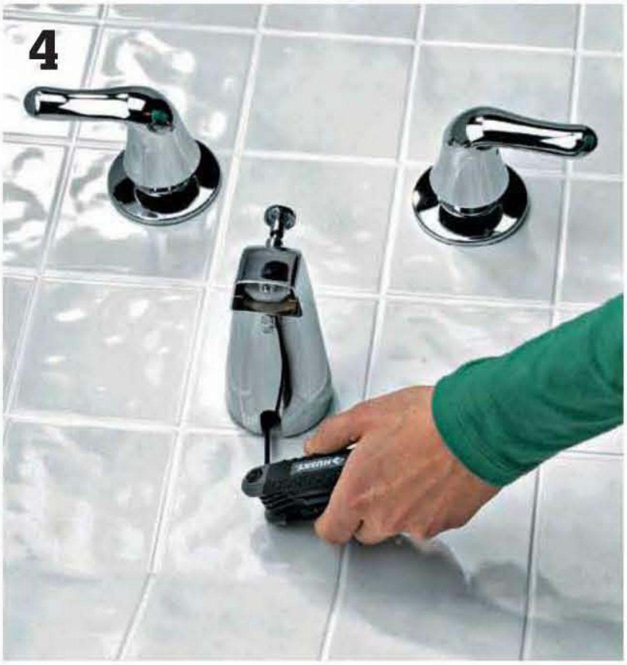

Spin the spout so it's right-side up and then tighten the setscrew from below, making sure the wall end of the spout is flush against the wall. Do not overtighten the setscrew.

How to Install a Threaded Spout

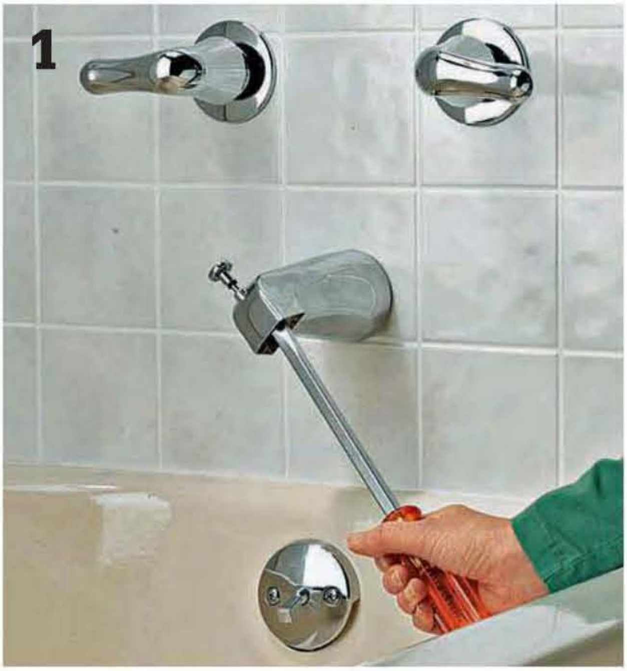

If you see no setscrew or slot on the underside of the spout, it is attached to a threaded nipple. Unscrew the tub spout by inserting a heavy-duty flat screwdriver into the spout opening and spinning it counterclockwise.

Tool Tip ▸

Alternatively, grip the spout with a padded pipe wrench or channel-type pliers. Buy a compatible replacement spout at a home center or hardware store.

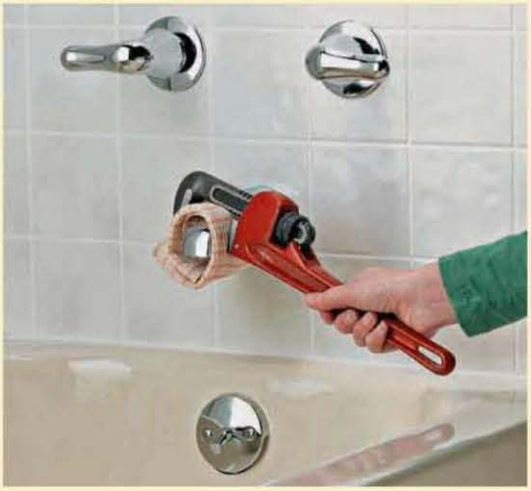

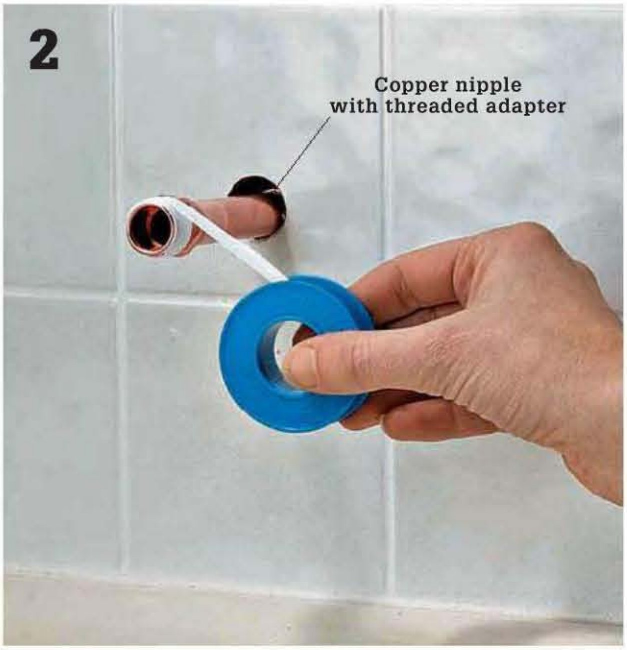

Wrap several courses of Teflon tape clockwise onto the pipe threads of the nipple. Using extra Teflon tape on the threads creates resistance if the spout tip points past six o'clock when tight.

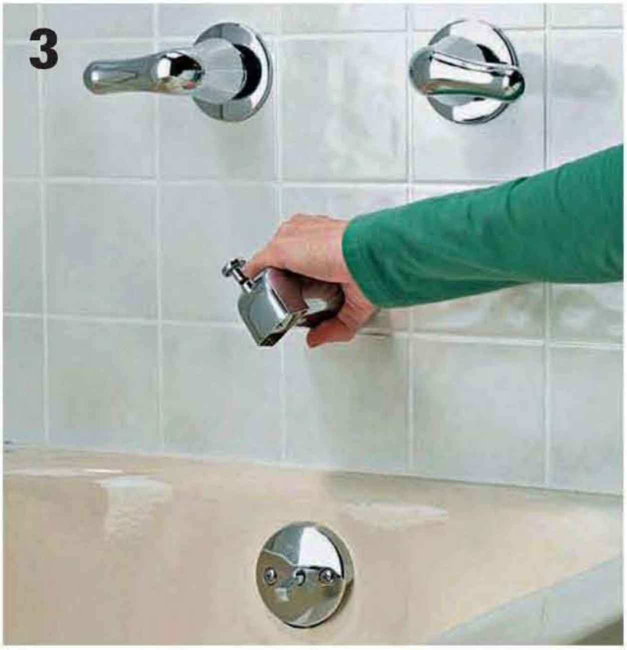

Twist the new spout onto the nipple until it is flush against the wall and the spout is oriented properly. If the spout falls short of six o'clock, you may protect the finish of the spout with tape and twist it a little beyond hand tight with your channel-type pliers—but don't overdo it; the fitting can crack.

Installing Vanity Lights

Many bathrooms have a single fixture positioned above the vanity, but a light source in this position casts shadows on the face and makes grooming more difficult. Placing light fixtures on either side of the mirror is a better arrangement.

For a remodel, mark the mirror location, run cable, and position boxes before drywall installation. You can also retrofit by installing new boxes and drawing power from the existing fixture.

The light sources should be at eye level; 66" is typical. The size of your mirror and its location on the wall may affect how far apart you can place the sconces, but 36 to 40" apart is a good guideline.

Tools & Materials ▸

Drywall saw
Drill
Combination tool
Circuit tester
Screwdrivers
Hammer
Electrical boxes and braces
Vanity light fixtures
NM cable
Wire connectors

Vanity lights on the sides of the mirror provide even lighting.

How to Replace Vanity Lights in a Finished Bathroom

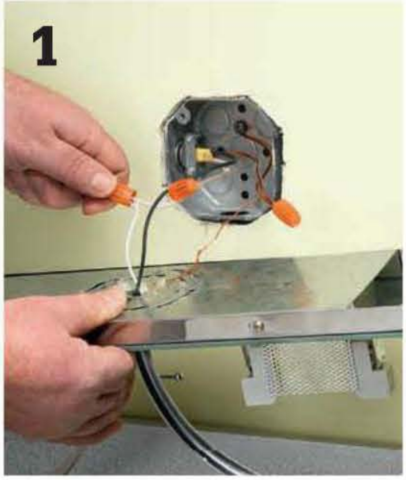

Turn off the power at the service panel. Remove the old fixture from the wall and test to make sure that the power is off. Then remove a strip of drywall from around the old fixture to the first studs beyond the approximate location of the new fixtures. Make the opening large enough that you have room to route cable from the existing fixture to the boxes.

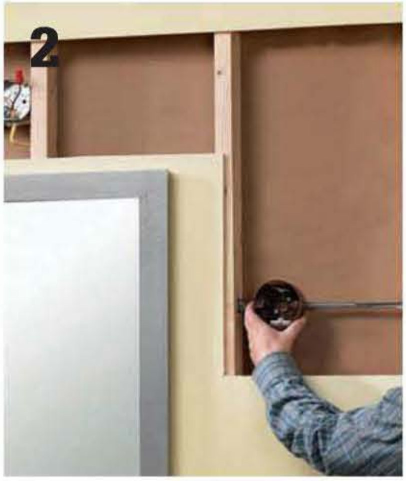

Mark the location for the fixtures and install new boxes. Install the boxes about 66" above the floor and 18 to 20" from the centerline of the mirror (the mounting base of some fixtures is above or below the bulb, so adjust the height of the bracing accordingly). If the correct location is on or next to a stud, you can attach the box directly to the stud, otherwise you'll need to install blocking or use boxes with adjustable braces (shown).

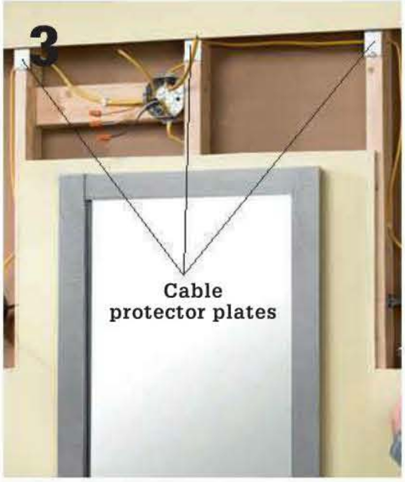

Open the side knockouts on the electrical box above the vanity. Then drill ⅝" holes in the centers of any studs between the old fixture and the new ones. Run two NM cables from the new boxes for the fixtures to the box above the vanity. Protect the cable with metal protector plates. Secure the cables with cable clamps, leaving 11" of extra cable for making the connection to the new fixtures. Remove sheathing and strip insulation from the ends of the wires.

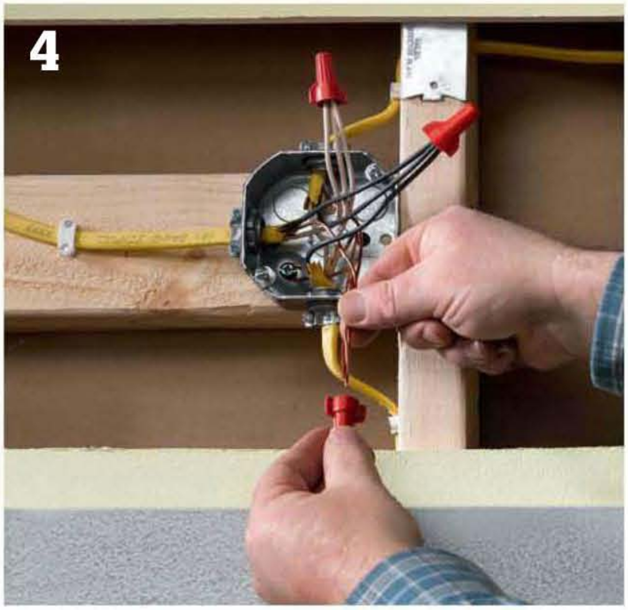

Connect the white wires from the new cables to the white wire from the old cable, and connect the black wires from the new cables to the black wire from the old cable. Connect the ground wires. Cover all open boxes and then replace the drywall, leaving openings for the fixture and the old box. (Cover the old box with a solid junction box cover plate.)

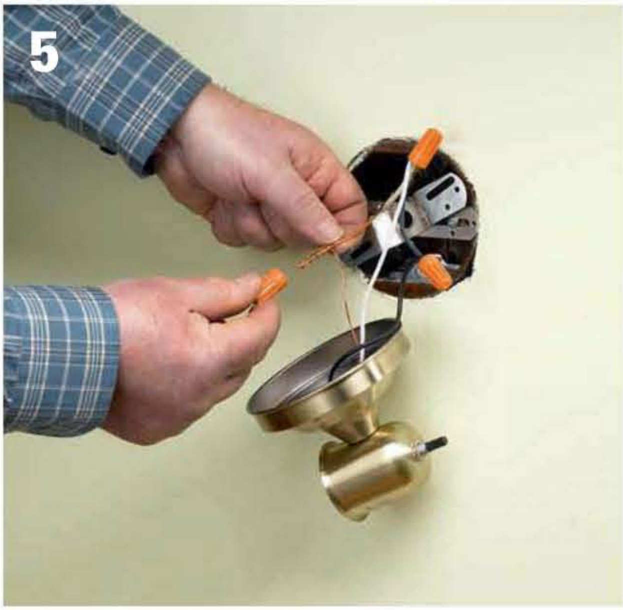

Install the fixture mounting braces on the boxes. Attach the fixtures by connecting the black circuit wire to the black fixture wire, and connecting the white circuit wire to the white fixture wire. Connect the ground wires. Position each fixture over each box, and attach with the mounting screws. Restore power and test the circuit.

Installing Grab Bars

Bathrooms are beautiful with their shiny ceramic tubs, showers, and floors, but add water and moisture to the mix and you've created the perfect conditions for a fall. The good news is many falls in the bathroom can be avoided by installing grab bars at key locations.

Grab bars help family members steady themselves on slippery shower, tub, and other floor surfaces. Plus, they provide support for people transferring from a wheelchair or walker to the shower, tub, or toilet.

Grab bars come in a variety of colors, shapes, sizes, and textures. Choose a style with a 1¼ to 1½" diameter that fits comfortably between your thumb and fingers. Then properly install it 1½" from the wall with anchors that can support at least 250 pounds.

The easiest way to install grab bars is to screw them into wall studs or into blocking or backing attached to studs. Blocking is a good option if you are framing a new bathroom or have the wall surface removed during a major remodel (see Illustration A). Use 2 × 6 or 2 × 8 lumber to provide room for adjustments, and fasten the blocks to the framing with 16d nails. Note the locations of your blocking for future reference.

As an alternative, cover the entire wall with ¾" plywood backing secured with screws to the wall framing, so you can install grab bars virtually anywhere on the wall (see Illustration B).

Grab bars can be installed in areas without studs. For these installations, use specialized heavy-duty hollow-wall anchors designed to support at least 250 pounds.

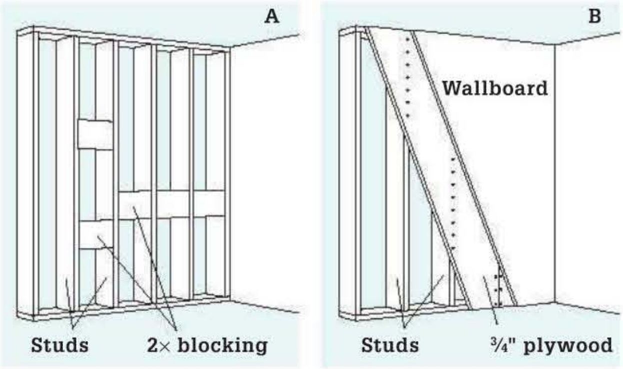

Blocking or backing is required for secure grab bars. If you know where the grab bars will be located, add 2× blocking between studs (Illustration A). You also can cover the entire wall with ¾" plywood backing, which allows you to install grab bars virtually anywhere on the wall.

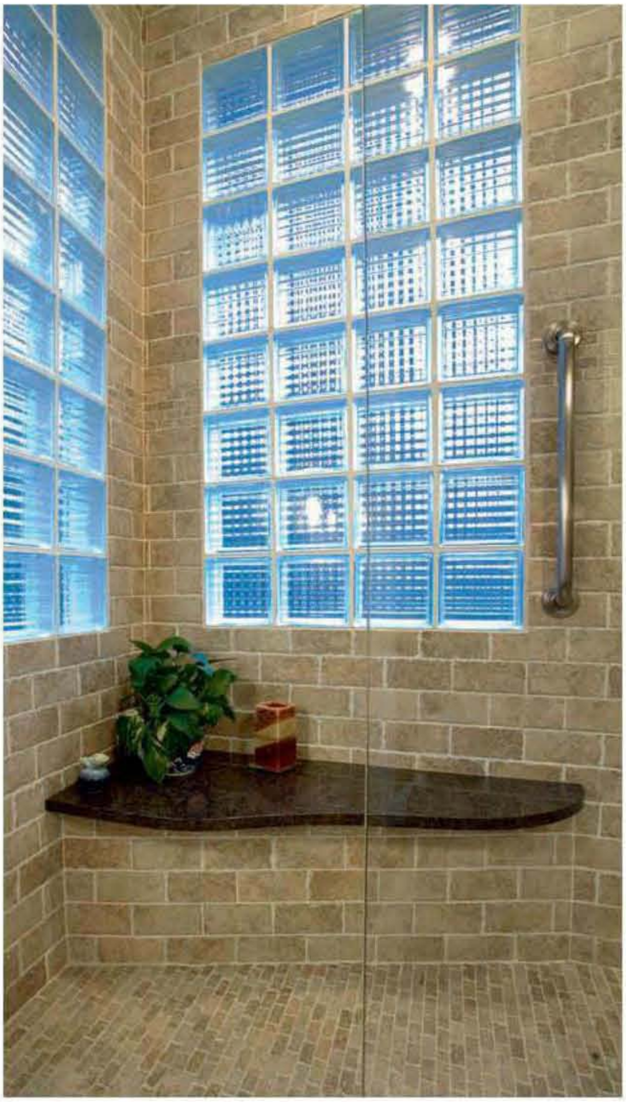

Grab bars promote independence in the bathroom, where privacy is especially important. Grab bars not only help prevent slips and falls, they also help people steady themselves in showers and lower themselves into tubs.

Tools & Materials ▸

Measuring tape
Pencil
Stud finder
Level
Drill
Masonry bit
Grab bar
Hollow-wall anchors
#12 stainless steel screws
Silicone caulk

How to Install Grab Bars

Locate the wall studs in the installation area, using a stud finder. If the area is tiled, the stud finder may not detect studs, so try to locate the studs above the tile, if possible, then use a level to transfer the marks lower on the wall. Otherwise, you can drill small, exploratory holes through grout joints in the tile, then fill the holes with silicone caulk to seal them. Be careful not to drill into pipes.

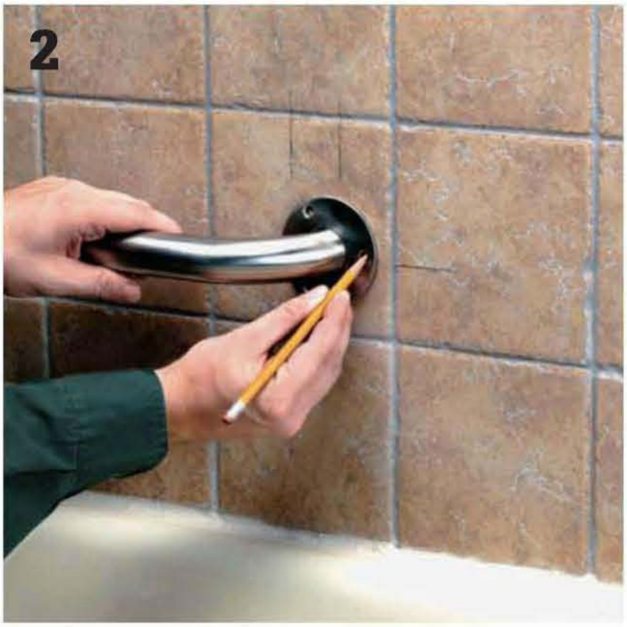

Mark the grab bar height at one stud location, then use a level to transfer the height mark to the stud that will receive the other end of the bar. Position the grab bar on the height marks so at least two of the three mounting holes are aligned with the stud centers. Mark the mounting hole locations onto the wall.

Drill pilot holes for the mounting screws. If you are drilling through tile, start with a small bit (about ⅛"), then redrill the hole with the larger bit. For screws that won't hit studs, drill holes for wall anchors, following the manufacturer's directions for sizing. Install anchors, if necessary.

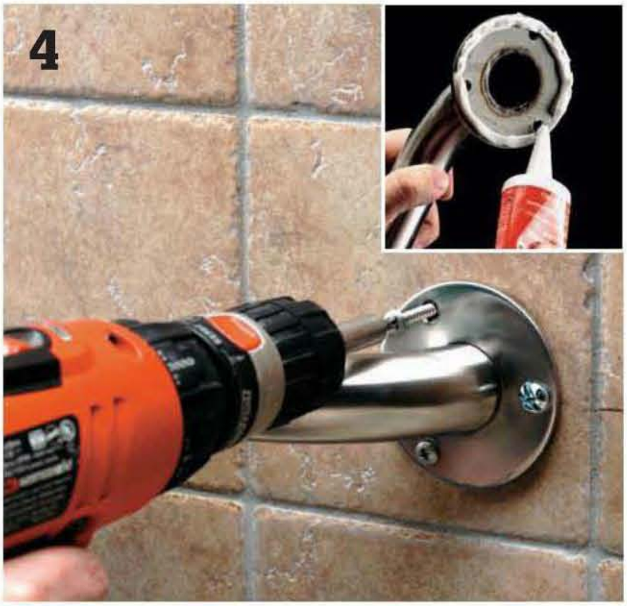

Apply a continuous bead of silicone caulk to the back side of each bar end (inset). Secure the bar to the studs using #12 stainless steel screws (the screws should penetrate the stud by at least 1"). Install a stainless steel screw or bolt into the wall anchors. Test the bar to make sure it's secure.

Installing Glass Shelving

Glass shelving is unobtrusive, so it can fit many styles of bathrooms—from sleek modern to Victorian. You can find a wide variety of shelving available in home stores and online.

Most glass shelves are held in place with metal mounts. How the shelves are secured to the mounts and how the mounts are attached to the wall differs. Most shelves have a hidden bracket that is secured to the wall. The mount then slips over the bracket and is secured with a setscrew. The most basic models may have mounts that are screwed directly into the wall with exposed screws. The directions here are for shelving that uses hidden brackets.

If you are installing shelves on a tiled wall, mount the brackets in grout lines if at all possible to minimize the possibility of cracking the tiles. Many glass shelves have some variability in the distance between the mounts.

Tools & Materials ▸

Level	Pencil
Drill	Glass shelves

Glass shelves fit any style and size of bathroom. They are held in place with metal mounts attached to the walls.

How to Install Glass Shelves

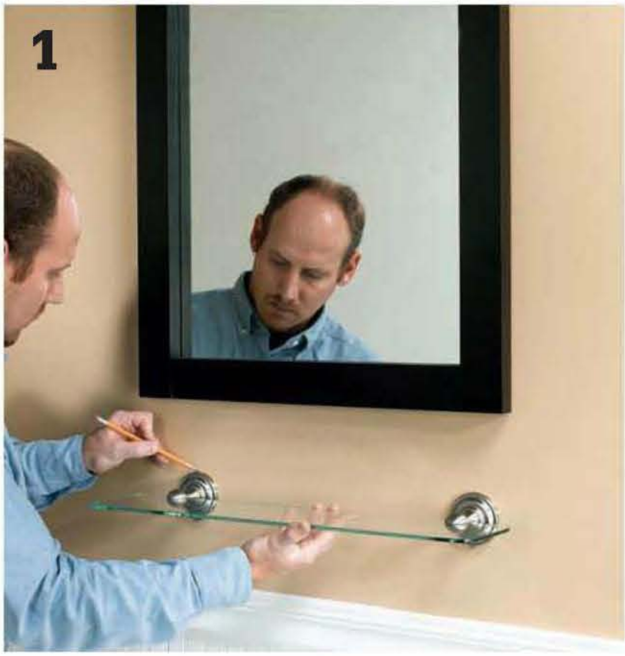

Assemble the shelf and shelf holders (not the brackets). Hold the shelf against the wall in the desired location. On the wall, mark the center point of each holder, where the setscrew is. Remove the shelf from the holders and set aside.

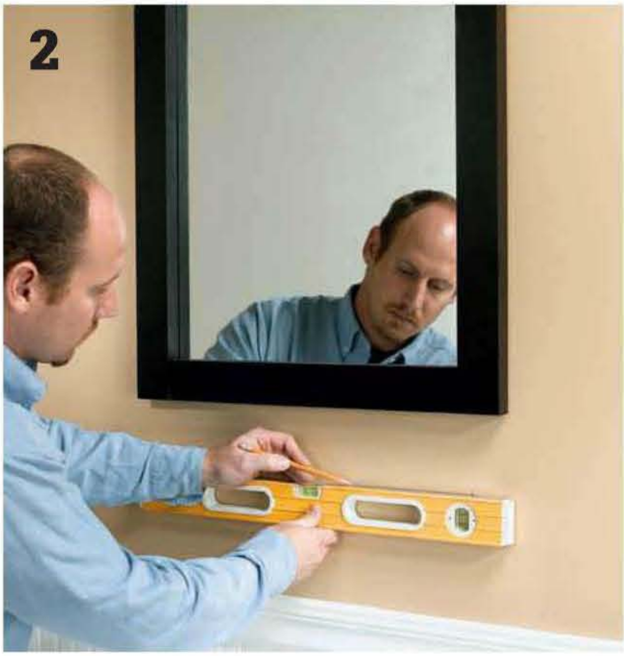

Remove the shelves and use the level to extend the mark into a 3" vertical line. Use the level to mark a horizontal line across the centers of these lines.

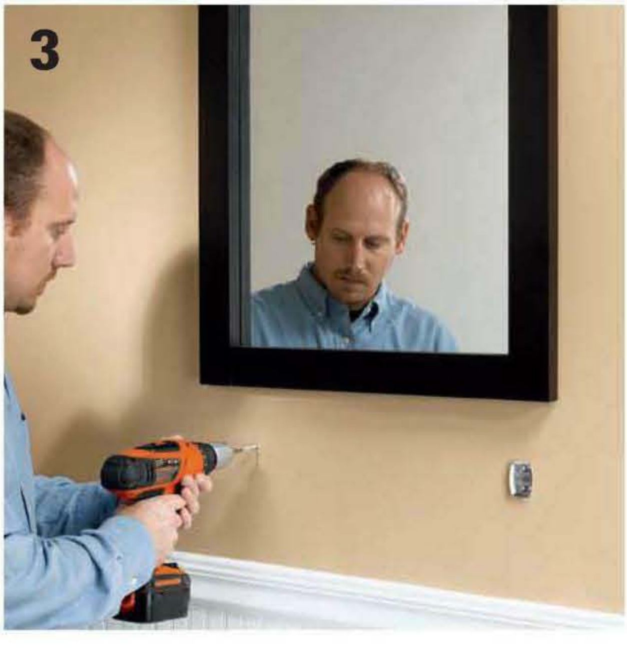

Center the middle round hole of the bracket over the intersection of the vertical and horizontal lines. Mark the center of each of the oblong holes. Put the bracket aside and drill a ¼" hole at each mark. Insert the included wall anchors in the holes. Replace the bracket and insert the screws into the wall anchors and drive the screws. Repeat for the second bracket.

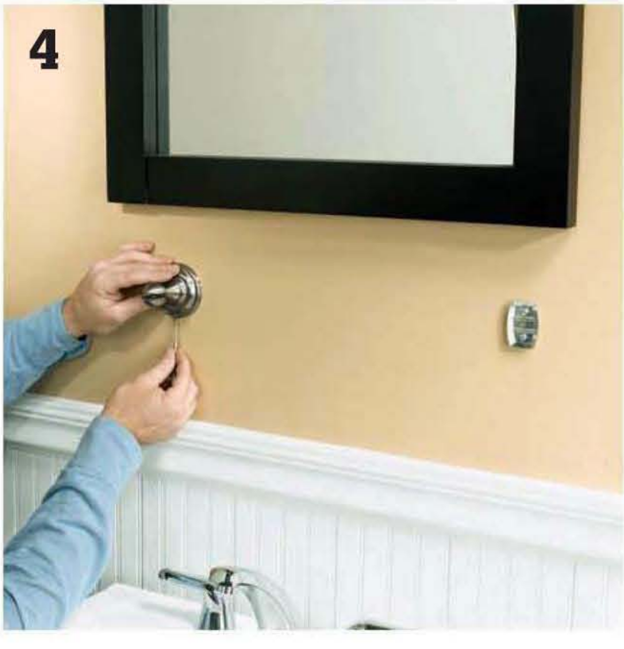

Slide a holder over a bracket, check that the shelf mount is level, and tighten the setscrew. Repeat with the other holder. Insert the shelf and fix in place. Check the shelf for level. If it's not level, remove one holder and loosen the bracket screws. Slide the bracket up or down to make the unit level. Replace the holder and shelf.

Installing Towel Warmers

Here's a little bit of luxury that need not be limited to high-end hotel stays. You can have toasty towels in your own bathroom with an easy-to-install towel warmer. In a relatively cold room, this can make stepping out of the shower a much more pleasant experience.

Heated towel racks are available in a wide range of styles and sizes. Freestanding floor models as well as door- and wall-mounted versions can be plugged in for use when desired. Hardwired wall-mounted versions can be switched on when you enter the bathroom so your towels are warm when you step out of the shower. Although installing them requires some electrical skills, the hardwired models do not need to be located near wall receptacles and they do not have exposed cords or extension cords hanging on the wall. However, if you locate the warmer directly above an existing receptacle, you can save a lot of time and mess by running cable up from the receptacle to the new electrical box for the warmer.

Before installing hardwired models, check your local electrical codes for applicable regulations. If you are not experienced with home wiring, have an electrician do this job for you or opt for a plug-in model.

Tools & Materials ▸

Drill
Level
Keyhole saw
Wiring tools
Phillips screwdriver
Stud finder
Wire connectors
Retrofit electrical outlet box
NM cable
Towel warmer
Pencil
Masking tape

A hard-wired towel warmer offers the luxury of heated towels without the safety concerns of a plug-in device.

How to Install a Hardwired Towel Warmer

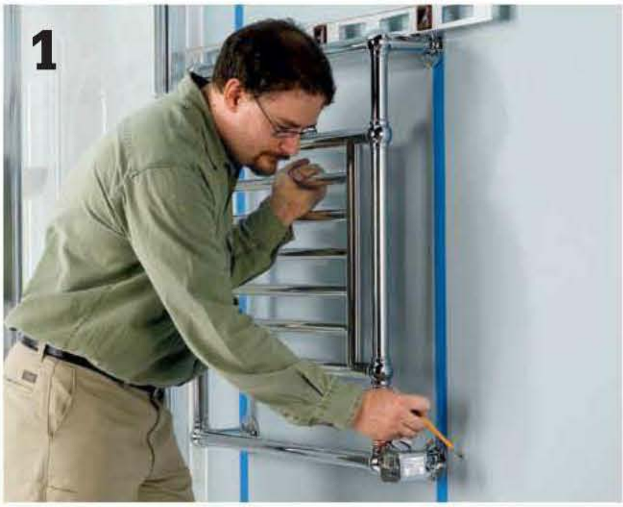

Use a stud finder to locate the studs in the area you wish to place the towel warmer. Mark the stud locations with masking tape or pencil lines. Attach the wall brackets to the towel warmer and hold the unit against the wall at least 7" from the floor and 3" from the ceiling or any overhang. Mark the locations of the wall bracket outlet plate (where the electrical connection will be made) and the mounting brackets.

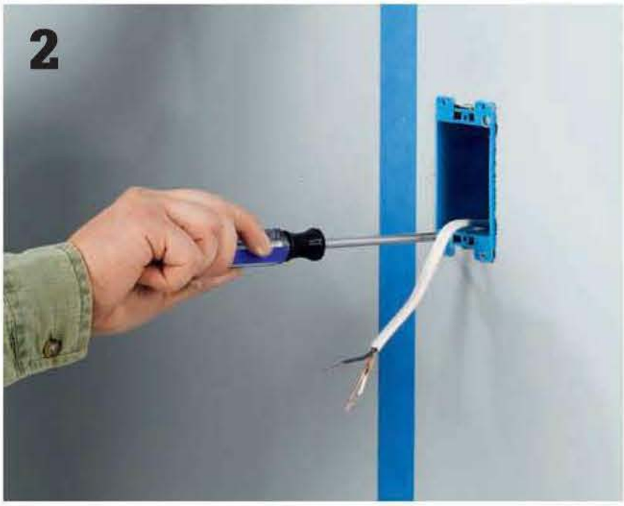

Shut off electrical power at the main service panel. At the mark for the wall bracket outlet, cut a hole in the wallcovering for a retrofit electrical box. Run NM cable from the opening to a GFCI-protected circuit (here, we ran cable down to a receptacle directly beneath it), or install a separate GFCI-protected circuit (you'll need to consult a wiring book or an electrician). Pull the cable through the hole in the retrofit box, and then tighten the cable clamp. Place the box in the hole flush with the wall surface and tighten the mounting screw in the rear of the box. Cut the wires so about 5" extends into the box and strip the insulation off at ⅜" from the end of each wire.

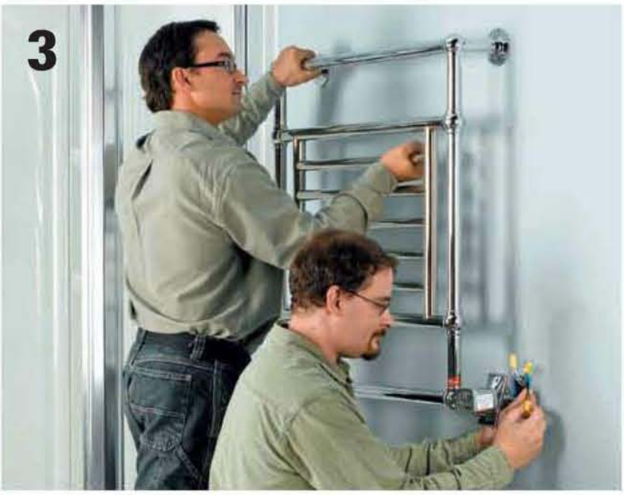

Position the towel warmer over the outlet box and mark the locations of the screw holes for the wall brackets. Make sure the appliance is level. Remove the warmer and drill ¼" pilot holes at the marked locations. If the marks are located over studs, drill ⅛" pilot holes. If not, push wall anchors into the holes. Thread the mounting screws through the brackets. Have a helper hold the towel rack in place and use wire connectors to connect the wires, including the ground wire, according to the instructions.

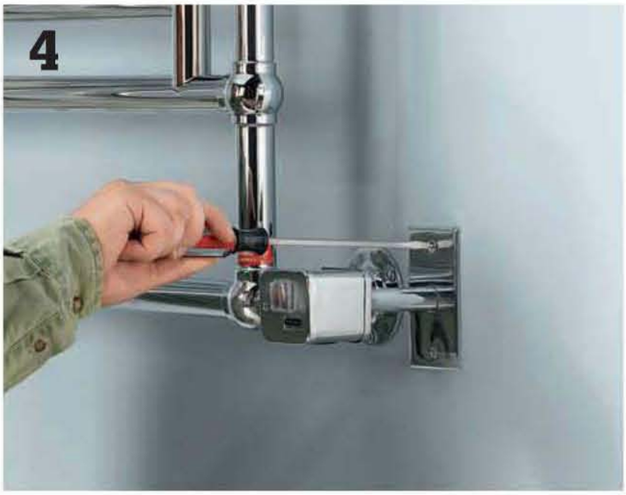

After the electrical connections are made, fasten the towel rack brackets to the wall. Turn on power and test the towel warmer. Finally, attach the electrical cover plate with integral on/off switch.

Basements & Attics

Is your house lacking just a little more space but you love it too much to think about moving? Has your life changed since you bought the house and now you find you need a home office for your freelance work or telecommuting, a place for the children in your life to play, or an extra bedroom for a relative? Or maybe now that your obligations have changed, you find yourself fantasizing about a space dedicated to your hobbies?

This chapter can help you determine how best to use your tantalizingly empty attic or basement, inspire you with great design ideas, and give you the practical how-to to tackle the projects on your own.

Over half of houses in North America are more than 30 years old. The attics and basements in these older homes are often significantly underutilized spaces. Even in new construction, attics and basements often are left unfinished so that the new homeowners can determine for themselves how best to incorporate the space.

In this chapter:

- Evaluating Your Attic
- Evaluating Your Basement
- Planning the Framing
- Installing Basement Floors
- Installing Attic Floors
- Installing Basement Walls
- Installing Attic Walls
- Finishing Attic Ceilings
- Adding Basement Egress Windows
- Installing Baseboard Heaters

Evaluating Your Attic

Start your attic evaluation with a quick framing inspection. If the roof is framed with rafters, you can continue to the next test. If it's built with trusses, however, consider remodeling your basement instead. The problem is that the internal supports in trusses leave too little space to work with, and trusses cannot be altered.

The next step is to check for headroom and overall floor space. Most building codes call for 7½ ft. of headroom over 50% of the "usable" floor space, which is defined as any space with a ceiling height of at least 5 ft. Remember that these minimums apply to the finished space—after the flooring and ceiling surfaces are installed. Other things can affect headroom, as well, such as reinforcing the floor frame, and increasing rafter depth for strength or insulation.

You may also find various supports in your attic that are there to strengthen your roof but may limit your space. Collar ties are horizontal boards that join two rafters together in the upper third of the rafter span. They prevent rafter uplift in high winds. Often collar ties can be moved up a few inches but cannot be removed. Rafter ties join rafters in the lower third of their span to prevent spreading. In most attics, the ceiling or floor joists serve as rafter ties. Purlins are horizontal boards that run at right angles to the rafters and are supported by struts. These systems shorten the rafter span, allowing the use of smaller lumber for the rafters. You may be allowed to substitute kneewalls for purlins and struts. If you'll need to have any support system altered or moved, consult an architect or engineer.

The rafters themselves also need careful examination. Inspect them for signs of stress or damage, such as cracks, sagging, and insect infestation. Look for dark areas indicating roof leaks. If you find leaks or you know your roofing is past its useful life, have it repaired or replaced before you start the finishing process. And even if the rafters appear healthy, they may be too small to support the added weight of finish materials. Small rafters can also be a problem if they don't provide enough room for adequate insulation.

At this point, it's a good idea to have a professional check the structural parts of your attic, including the

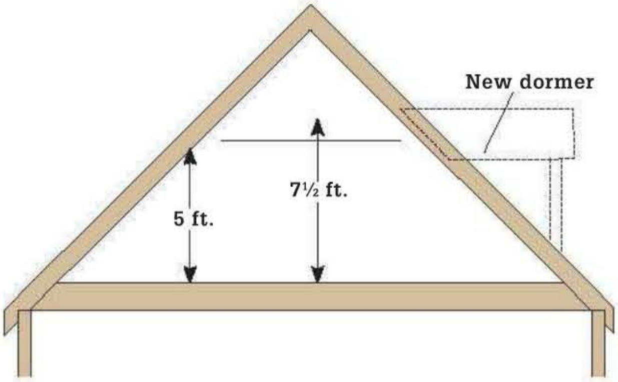

Habitable rooms must be at least 70 sq. ft. total and measure at least 7 ft. in any one direction. To meet headroom requirements, 50% of the usable floor space must have a ceiling height of 7½ ft. You can add to your floor space and headroom by adding protruding windows called dormers. In addition to space, dormers add light and ventilation to your attic.

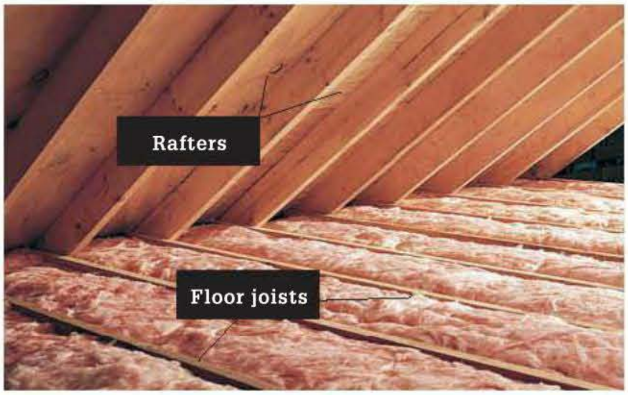

Rafter framing creates open space in an attic because the rafters carry most of the roof's weight.

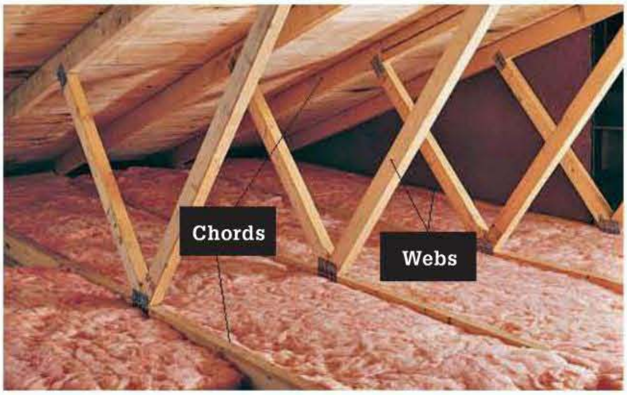

Trusses are made of interconnected cords and webs, which close off most of the attic space.

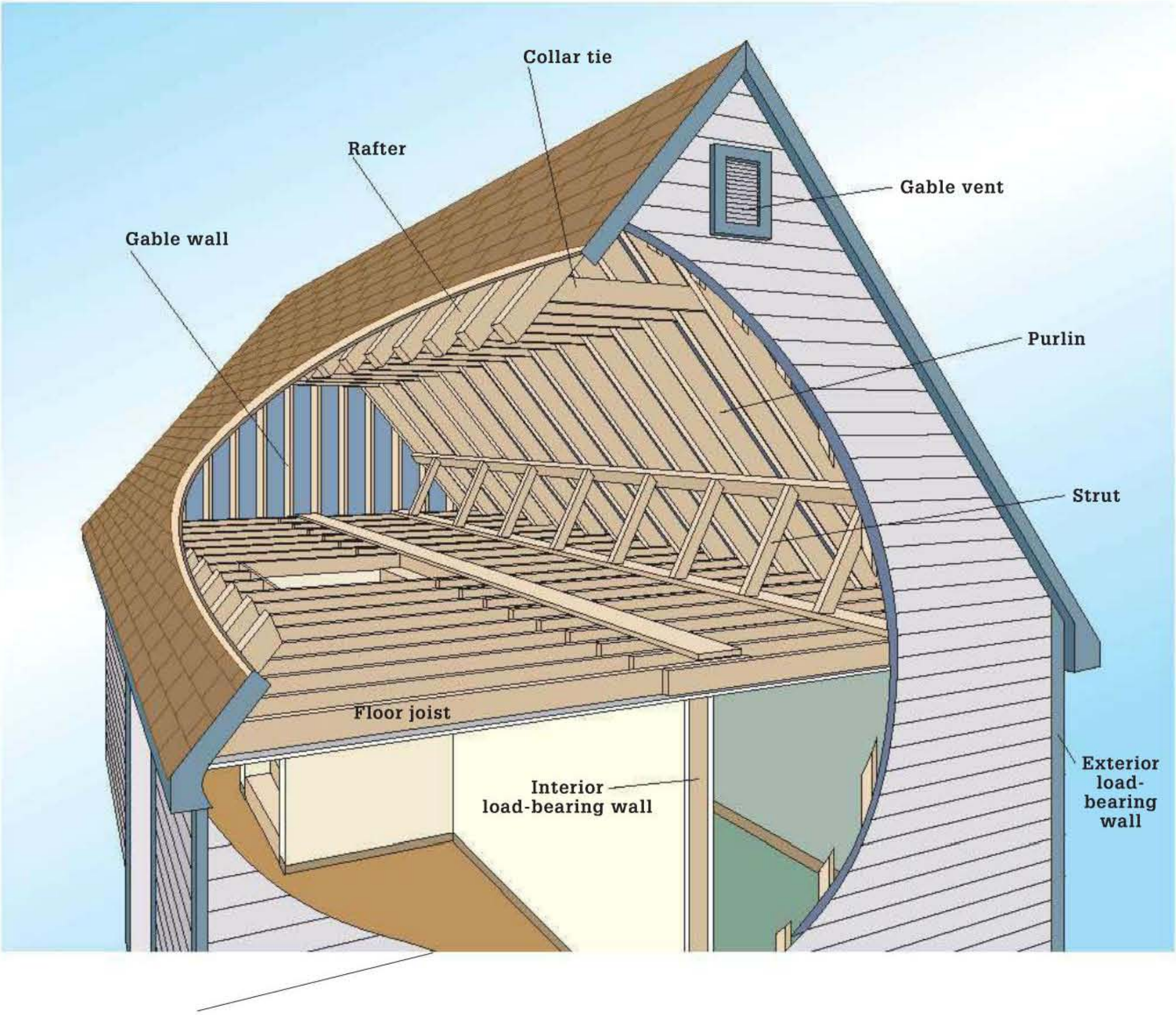

rafters and everything from the floor down. In some cases, finishing an attic is like adding a story to your home, which means that the structure must have adequate support for the new space. Attic floors are often built as ceiling frames for the level below and are not intended to support living space. Floors can be strengthened with additional joists, known as sister joists or with new joists installed between the existing ones.

Support for the attic floor is provided by the load-bearing walls below and, ultimately, by the foundation. If these elements can't support the finished attic, they'll need to be reinforced. This may be as simple as strengthening the walls with plywood panels or as complicated as adding support posts and beams or reinforcing the foundation.

In addition to these structural matters, there are a few general code requirements you should keep in mind as you inspect your attic. If you plan to add a bedroom, it will need at least one exit to the outside. This can be a door leading to an outside stairwell or an egress window. Most codes also have minimum requirements for ventilation and natural light, which means you may have to add windows or skylights.

One of the largest expenses of finishing an attic is in providing access: You'll need a permanent stairway at least 36" wide, with room for a 36" landing at the top and bottom. This is an important planning issue because adding a stairway affects the layout and traffic patterns of the attic as well as the floor below.

Finally, take an inventory of existing mechanicals in your attic. While plumbing and wiring runs can be moved relatively easily, other features, such as chimneys, must be incorporated into your plans. This is a good time to have your chimney inspected by a fire official and to obtain the building code specifications for framing around chimneys.

Evaluating Your Basement

The two things that put an end to most basement finishing plans are inadequate headroom and moisture. Begin your evaluation by measuring from the basement floor to the bottom of the floor joists above. Most building codes require habitable rooms to have a finished ceiling height of 7½ ft., measured from the finished floor to the lowest part of the finished ceiling. However, obstructions, such as beams, soffits, and pipes, (spaced at least 4 ft. on center) usually can hang down 6" below that height. Hallways and bathrooms typically need at least 7-ft. ceilings.

While it's impractical to add headroom in a basement, there are some ways of working around the requirements. Ducts and pipes often can be moved, and beams and other obstructions can be incorporated into walls or hidden in closets or other uninhabitable spaces. Also, some codes permit lower ceiling heights in rooms with specific purposes, such as recreation rooms. If headroom is a problem, talk to the local building department before you dash your dreams.

A well-built basement is structurally sound and provides plenty of support for finished space, but before you cover up the walls, floor, and ceiling, check for potential problems. Inspect the masonry carefully. Large cracks may indicate shifting of the soil around the foundation; severely bowed or out-of-plumb walls may be structurally unsound. Small cracks usually cause moisture problems rather than structural woes, but they should be sealed to prevent further cracking. Contact an engineer or foundation contractor for help with foundation problems. If you have an older home, you may find sagging floor joists overhead or rotted wood posts or beams; any defective wood framing will have to be reinforced or replaced.

Your basement's mechanicals are another important consideration. The locations of water heaters, pipes, wiring, circuit boxes, furnaces, and ductwork can have a significant impact on the cost and difficulty of your project. Can you plan around components, or will they have to be moved? Is there enough headroom to install a suspended ceiling so mechanicals can remain accessible? Or, will you have to reroute pipes and ducts to increase headroom? Electricians and HVAC contractors can assess your systems and suggest modifications.

Aside from being dark and scary places, unfinished basements often harbor toxic elements. One of the most common is radon, a naturally occurring radioactive gas that is odorless and colorless. It's believed that prolonged exposure to high levels of radon can cause lung cancer. The Environmental Protection Agency has free publications to help you test for radon and take steps to reduce the levels in your house. For starters, you can perform a "short-term" test using a kit from a hardware store or home center. Look for the phrase "Meets EPA Requirements" to ensure the test kit is accurate. Keep in mind that short-term tests are not as conclusive as professional, long-term tests. If your test reveals high levels of radon, contact a radon specialist.

Another basement hazard is insulation containing asbestos, which was commonly used in older homes for insulating ductwork and heating pipes. In most cases, this insulation can be left alone provided it's in good condition and is protected from damage. If you

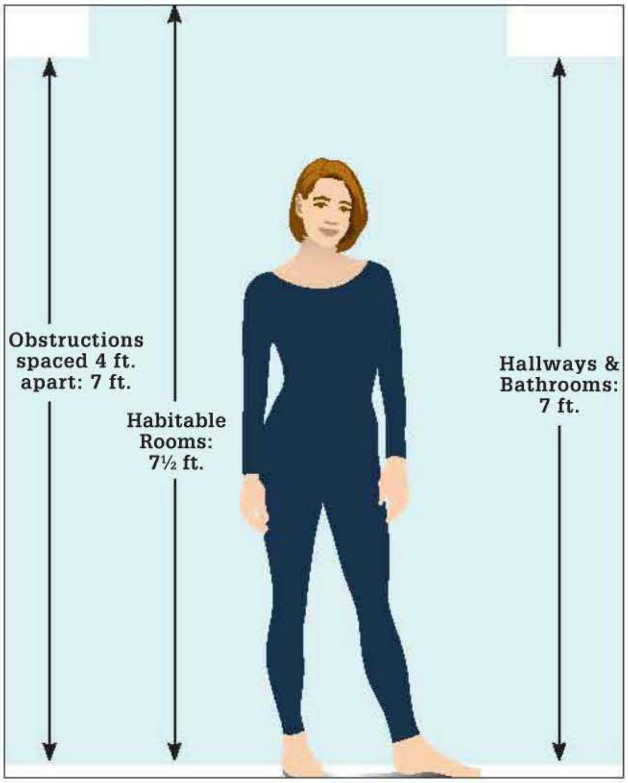
Obstructions spaced 4 ft. apart: 7 ft.
Habitable Rooms: 7½ ft.
Hallways & Bathrooms: 7 ft.

Basement headroom is often limited by beams, ducts, pipes, and other elements. Typical minimums for ceiling height are shown here: 7½ ft. for habitable rooms; 7 ft. for bathrooms and hallways; 7 ft. for obstructions spaced no less than 4 ft. apart.

fear the insulation in your basement poses a hazard, contact an asbestos abatement contractor to have it evaluated or safely removed.

Also check the local codes for exits from finished basements—most codes require two. The stairway commonly serves as one exit, while the other can be a door to the outside, an egress window, or a code-compliant bulkhead (an exterior stairway with cellar doors). Each bedroom will also need an egress window or door for escape.

Stairways must also meet the local code specifications. If yours doesn't, you'll probably have to hire someone to rebuild it.

As a final note, if you're planning to finish the basement in a new house, ask the builder how long you should wait before starting the project. Poured concrete walls and floors need time to dry out before they can be covered. Depending on where you live, you may be advised to wait up to two years, just to be safe.

Rerouting service lines and mechanicals adds quickly to the expense of a project, so consider your options carefully.

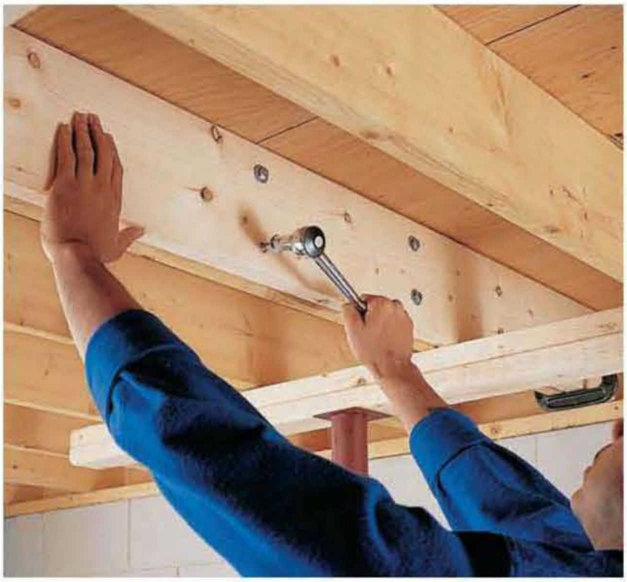

Weakened or undersized joists and other framing members must be reinforced or replaced.

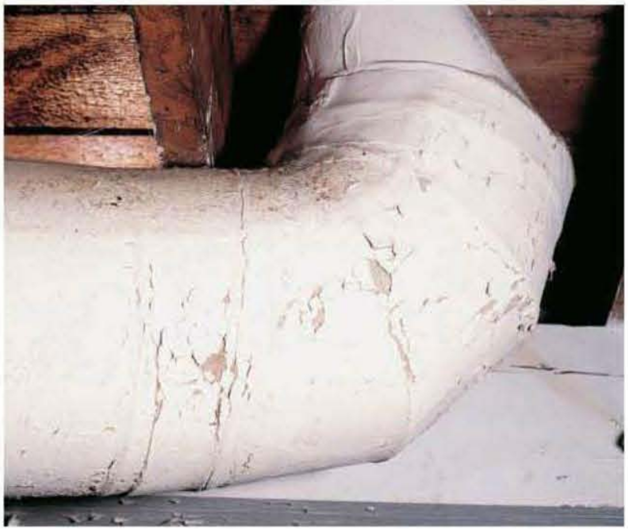

Old insulation containing asbestos poses a serious health risk if it is deteriorating or is disturbed.

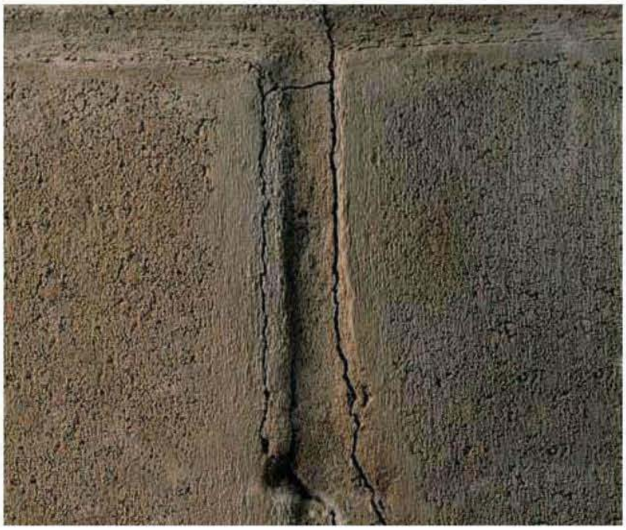

Minor cracks such as these in masonry walls and floors usually can be sealed and forgotten, while severe cracking may indicate serious structural problems.

Planning the Framing

Use walls to define your new spaces. Walls can create quiet private retreats or comfortable bathrooms or serve as barriers between formal living areas and dusty, unfinished storage spaces. To determine where your walls should go, start with a thorough investigation of the unfinished space. All obstacles, such as mechanical systems, service lines, floor drains, support columns, chimneys, and roof framing, must be considered. As you work with different layouts, think about which of these elements can be enclosed by walls, which can be hidden within a wall or concealed by a soffit or chase, and which, if any, can be moved.

One technique to help you get started is to draw full-scale "walls" onto your basement floor, using children's sidewalk chalk (on wood attic floors, use wide masking tape instead of chalk). This can help you visualize the planned spaces and give you a better sense of room sizes. Complete the proposed layout in chalk, then walk through the rooms to test the traffic patterns. As you plan your rooms, keep in mind that most building codes require habitable rooms to have at least 70 sq. ft. of floor space and measure a minimum of 7 ft. in any direction.

The next step is to draw floor plans. This doesn't require drafting skills—just a tape measure, a ruler, graph paper, and some pencils. Simply measure your basement or attic floor space, then scale down the dimensions and transfer them to the graph paper. Add all obstacles, windows, doors, and other permanent fixtures. When everything is in place, start experimenting with different layouts. If you have your home's original blueprints, trace the floor plans onto tracing paper and work on new layouts from there.

Creating a successful layout takes time and often requires some creative problemsolving. To help generate ideas for your remodel, study the before-and-after drawings on page 499. While these floor plans may not look like your basement, they include many of the common elements and obstacles involved in a finishing project. They also show how carefully placed walls can transform an unfinished space into several livable areas that still leave room for storage and mechanical elements.

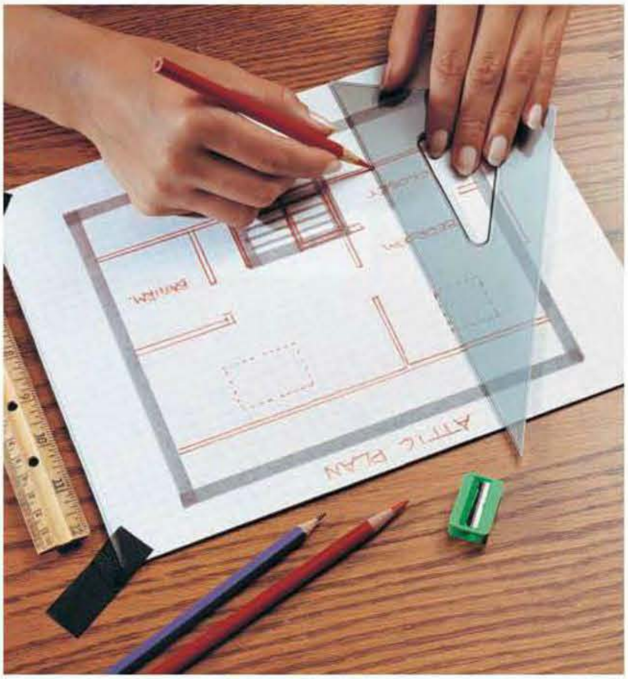

Use graph paper to sketch your wall layouts. Scale your floor plan (aerial view) drawings at ¼" equals 1 ft. For elevation drawings (wall details as viewed from the side), use a scale of ½" equals 1 ft.

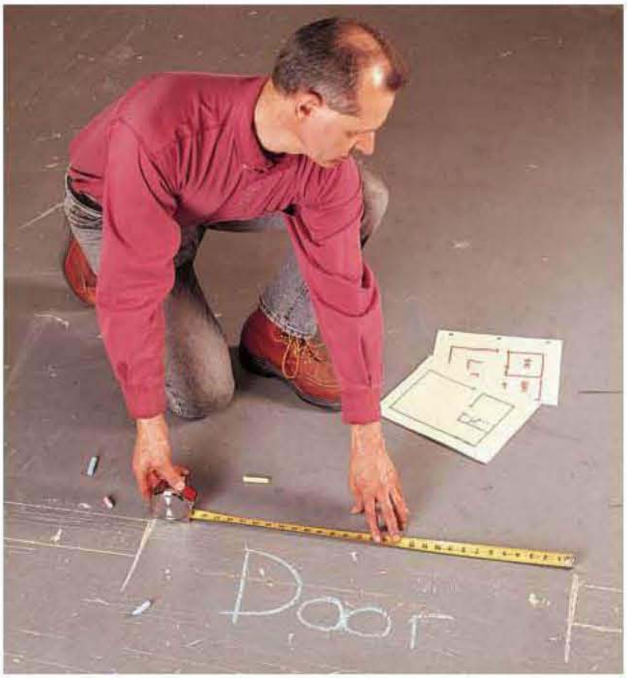

Draw layouts onto your basement floor with sidewalk chalk. Use different colors to represent elements other than walls, such as doors, windows, and ceiling soffits. Remove the chalk with a damp rag.

Basement Layouts: Before and After

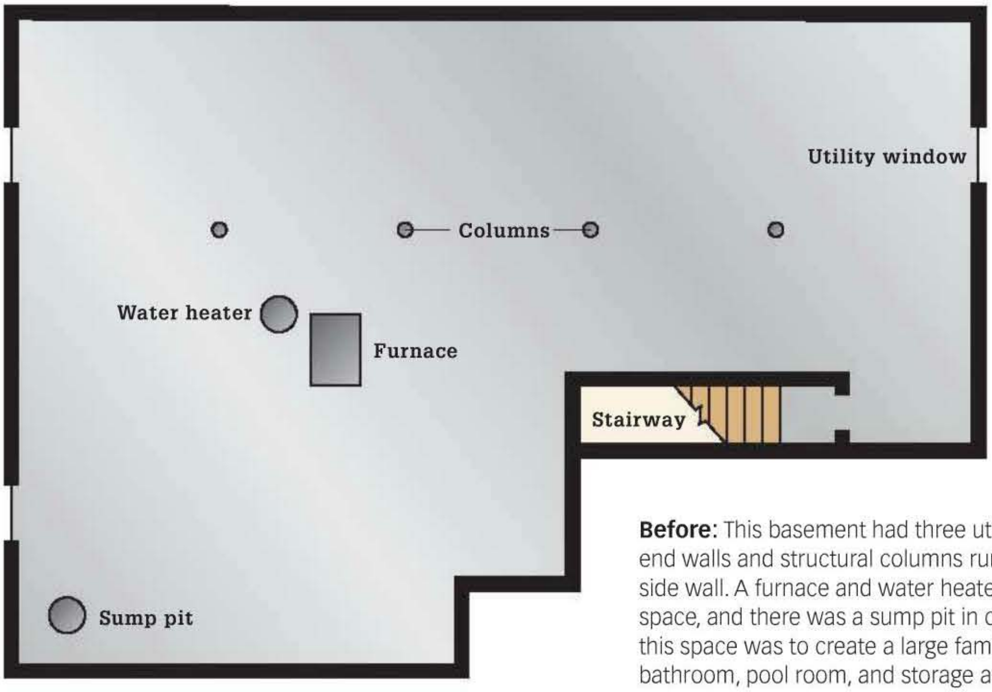
Utility window
Columns
Water heater
Furnace
Stairway
Sump pit

Before: This basement had three utility windows along the end walls and structural columns running parallel to the long side wall. A furnace and water heater broke up the central space, and there was a sump pit in one corner. The goal for this space was to create a large family room, home office, bathroom, pool room, and storage area.

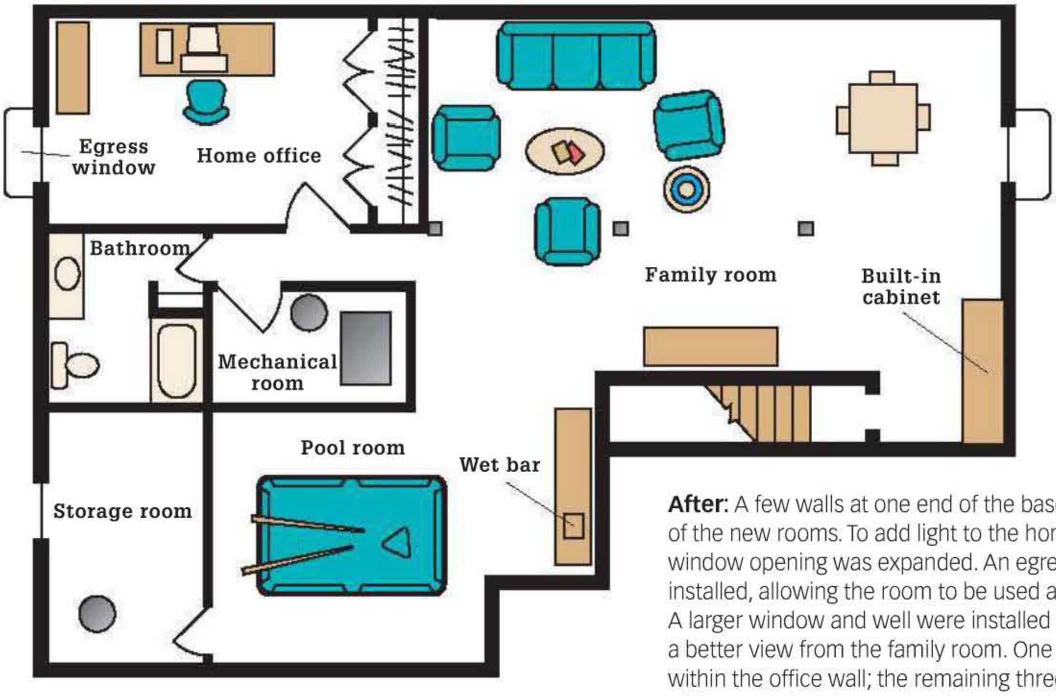
Egress window
Home office
Bathroom
Mechanical room
Family room
Built-in cabinet
Pool room
Wet bar
Storage room

After: A few walls at one end of the basement define several of the new rooms. To add light to the home office, the existing window opening was expanded. An egress window was installed, allowing the room to be used as a bedroom as well. A larger window and well were installed to provide light and a better view from the family room. One column was hidden within the office wall; the remaining three were wrapped with wood trim. The mechanical room contains the furnace and water heater, with plenty of space for servicing the units. Next to the office is a full bathroom, designed with a square layout that leaves a comfortable amount of space between the fixtures. The pool room occupies a well-defined space, where games won't disrupt activity in the family room. A wet bar can easily be accessed from both the pool room and family room. The stairway needed only a new handrail to become code-compliant. At the bottom of the stairs, a built-in cabinet provides storage and adds a decorative touch to the basement entrance.

Installing Basement Floors

Preparing a concrete floor—for carpet, laminate, vinyl, or wood flooring—has changed for the better in dramatic fashion, thanks to new subfloor products that have built-in vapor barriers and cleats that create a slight air gap between the subfloor and the concrete slab. This system allows air to circulate, protecting the finished flooring from any slab moisture. The older method of laying ¾" plywood over a frame of 2 × 4 "sleepers" was difficult, time-consuming, and raised the floor level by 2" or more—a significant drawback in basements where ceiling levels may already be too low. The new dry-floor subfloor systems are less than 1" thick and are very easy to install. There are several types of these dry-floor systems available, but the one most readily available and easiest to use is a product sold in 2 × 2-ft. tongue-and-groove squares.

Although subfloor panels can be adjusted for slight irregularities in the concrete slab, they can't overcome problems with a floor that is badly cracked and heaved. Nor is the built-in air gap beneath the system a solution to a basement that has serious water problems. A badly heaved slab will need to be leveled with a cement-based leveling compound, and serious water problems will need to be rectified before you consider creating finished living space in a basement.

Allow the subfloor panel squares to acclimate in the basement for at least 24 hours with the plastic surfaces facing down before installing them. In humid summer months, the squares—as well as the finished wood flooring product, if that's what you'll be installing—should be allowed to acclimate for a full two weeks before installation.

To estimate the number of 2-ft.-square subfloor panels you'll need, calculate the size of the room in square feet (multiply the width times the length of the room), then divide by 3.3 to determine the number of panels required.

Tools & Materials ▸

- Long straightedge
- Straightedge trowel
- Circular saw or jigsaw
- Hammer
- Carpenter's square
- Flat pry bar
- Dust mask
- Eye protection
- Portland-cement–based leveling compound
- Dry-floor subfloor squares
- Leveling shims
- Flooring spacers

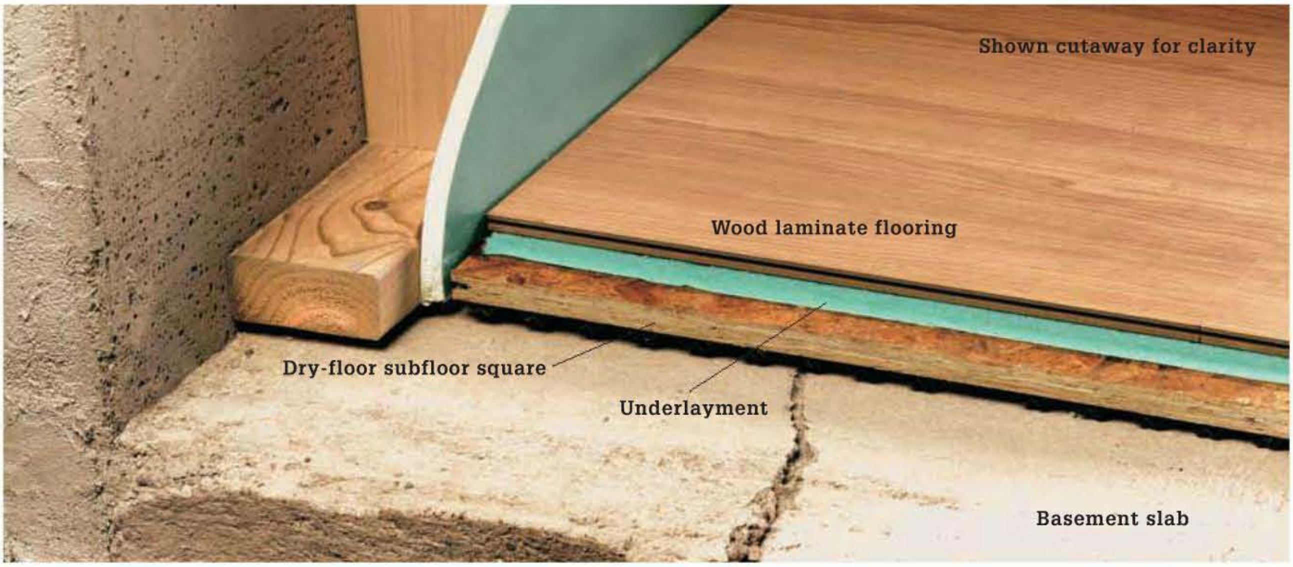

Most basement floors need some preparation before flooring can be laid. Patching compound and floor leveler can smooth rough concrete, while a subfloor system creates a new surface that is safe from moisture and feels like a framed wood floor.

Preparing Concrete Floors

Vacuum the floor thoroughly, then use a long straightedge to look for areas of the floor with serious dips or heaves. *Note: Any old floor coverings that may trap moisture should be removed before installing subfloor panels.*

Mix a batch of cement-based mortar and apply the compound to low areas with a straightedge trowel. After the patch dries, scrape the edges to feather the patch into the surrounding floor.

How to Install Subfloor Panels

1

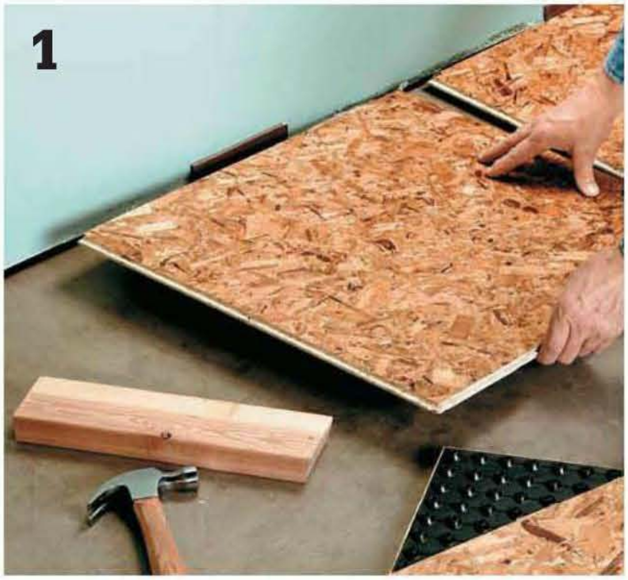

Beginning with the longest straight wall in the room, check one corner for square. If necessary, cut the first panel to match the angle of the corner. Position the panel in the corner, using ¼" spacers to create a gap between the panel and the walls, with the grooves of the panel against the spacers. Slide in the next panel, and use a piece of wood and a hammer to tap it firmly against the first panel. Repeat this placement along the entire wall. Cut the last panel to size.

2

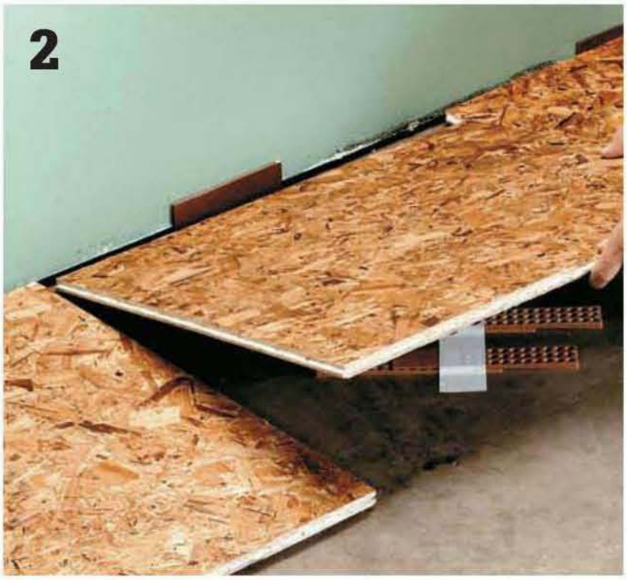

By hand, press along the entire row of panels, looking for low spots. At points where the subfloor flexes, lift the panels and place leveling shims on the floor. In some areas, you may find it necessary to stack two, three, or more shims to create the correct thickness to fill the low spot. Tape the shims to the floor to keep them in place as you reposition the panels.

(continued)

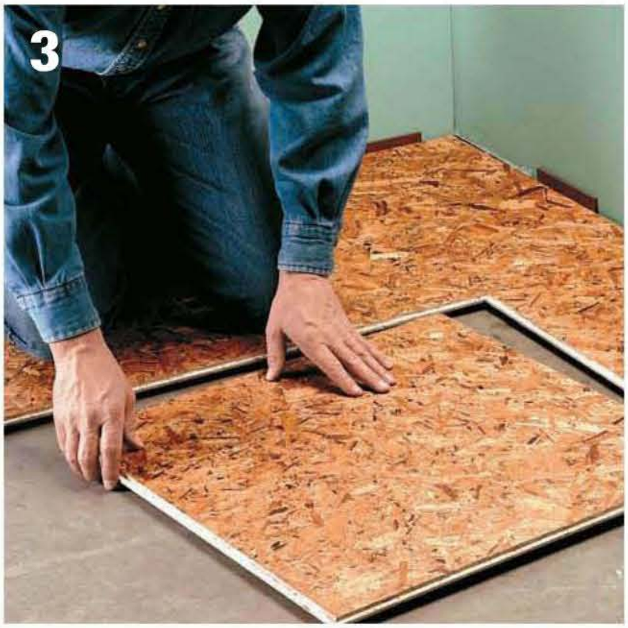

Begin the next row of panels by cutting a half-wide panel, creating offset joints between the rows. Finish the second row, test it for flatness, and shim if necessary. Move on to the third row, this time beginning with a full panel. Work across the entire room in this fashion, testing each row for flatness, shimming where necessary, and making sure joints are offset between rows.

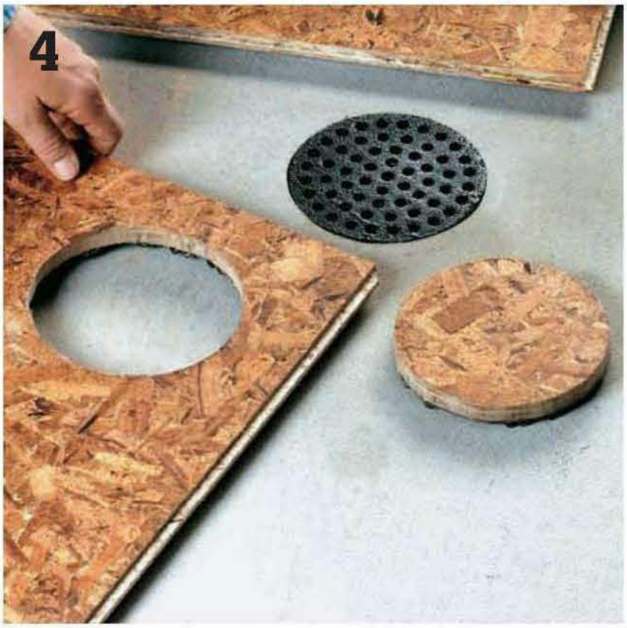

If the room has a floor drain, building code requires that you cut a round patch in the panel that falls directly over the floor drain. This patch can be removed to allow access to the floor drain should it ever be necessary.

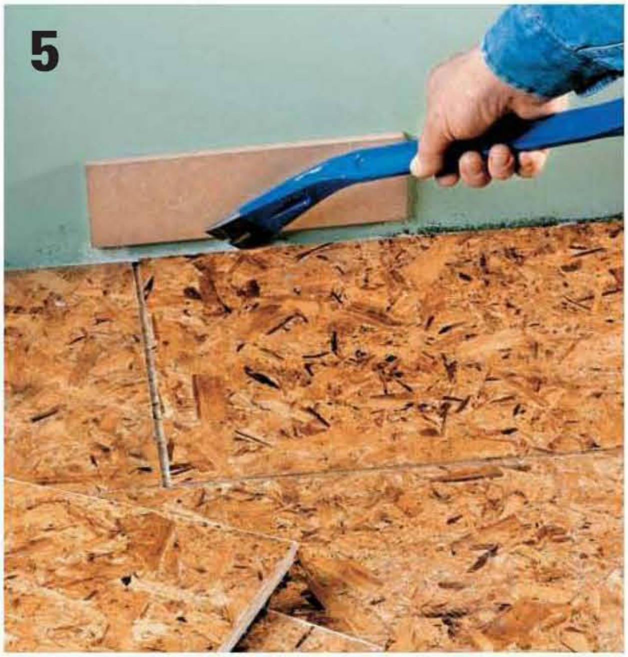

For the last row of panels, measure and cut panels to fit, maintaining a ¼" gap between the panels and the wall. Work each panel into place with a flat pry bar. When all panels are in place, you can remove all spacers and continue with the installation of the finished flooring. Laminate flooring can be installed directly over the subfloor panels.

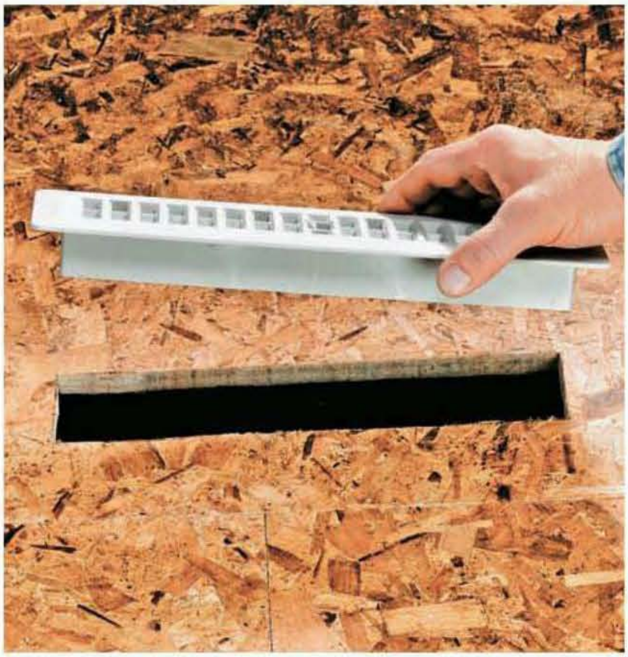

Variation: Where slab moisture is known to be a problem, you can cut vent openings spaced every 8 to 12", sized to match standard metal floor vent covers. Trace the vent opening onto the panel, and cut the opening with a jigsaw.

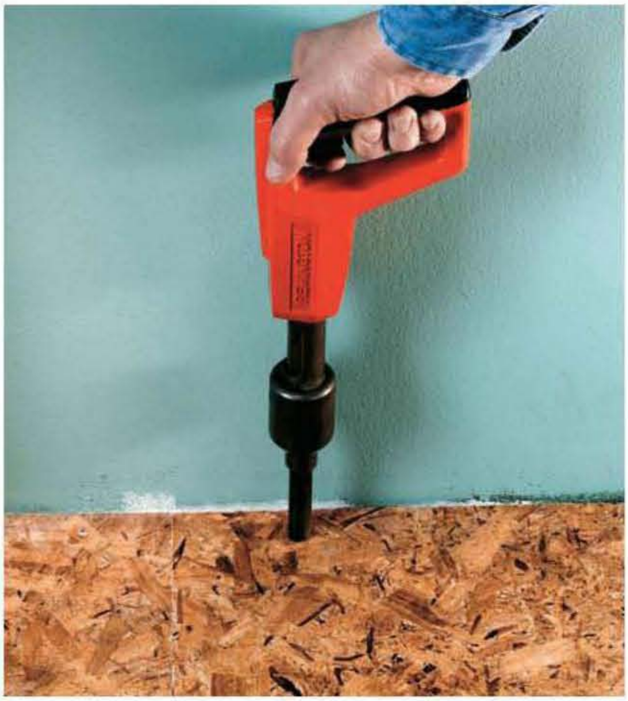

Variation: If carpeting will be stretched over the dry-floor panels, anchor the panels to the floor around the perimeter of the room and at the center of the room using concrete masonry anchors. Use a powder-actuated nailer to drive 2" nails through the panels and into the concrete slab.

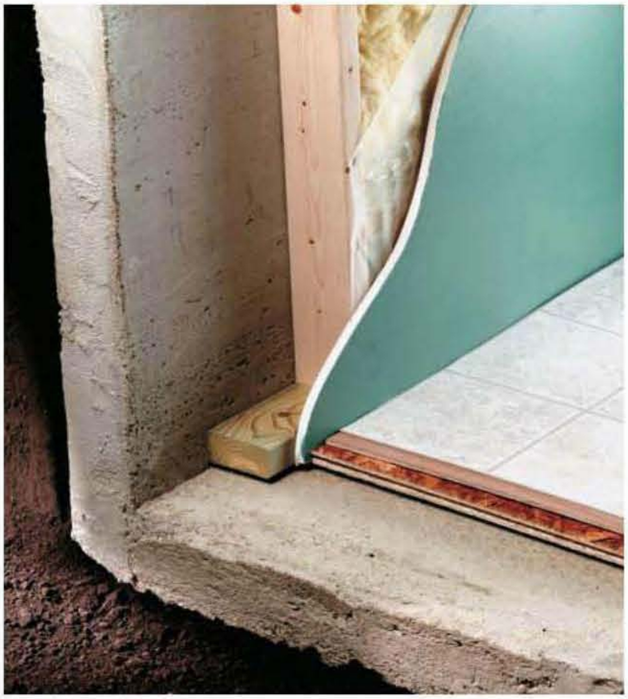

Variation: For vinyl flooring or for engineered wood flooring, install ¼" plywood underlayment over the dry-floor panels, using ½" screws or nails, which won't penetrate the moisture barrier on the underside of the panels. Don't glue the plywood to the subfloor panels, and never glue vinyl flooring directly to the subfloor panels.

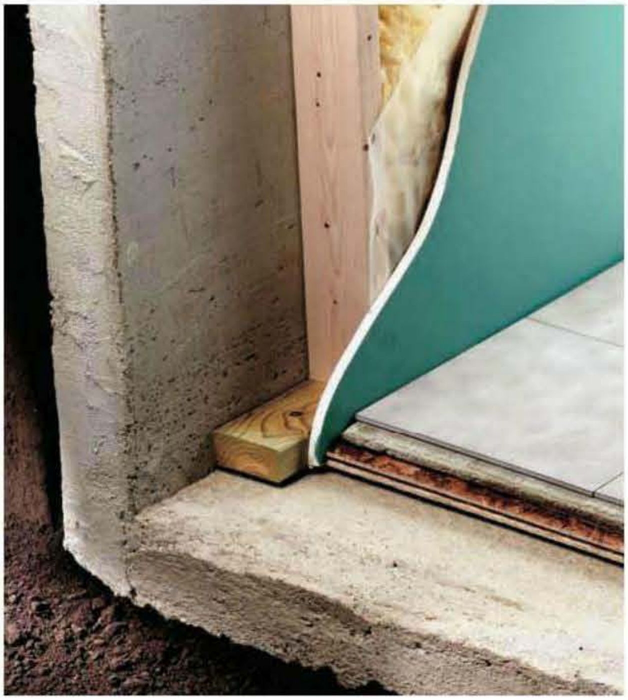

Variation: For ceramic tile, install ½" cementboard over the subfloor panels, and attach it using ¾" screws.

6

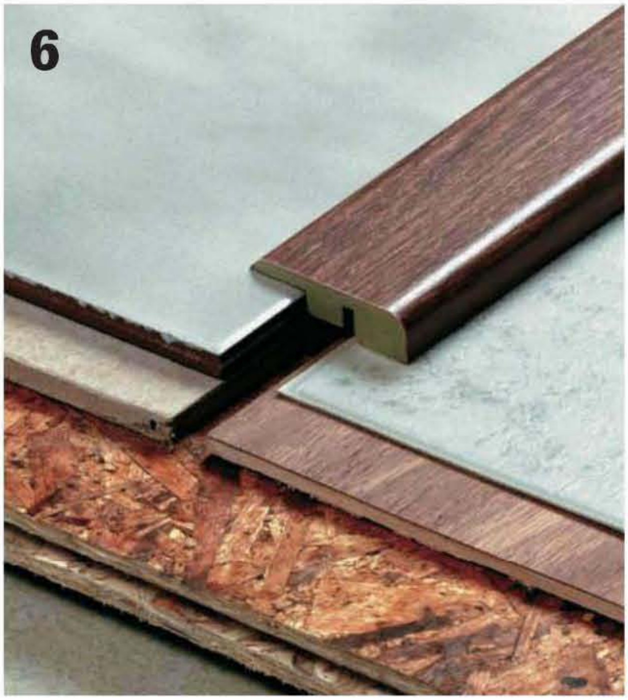

Where floor heights change, install transition strips or reducers, available in flooring departments of home improvement centers, to bridge the changes in floor heights.

How to Install a Sleeper Floor

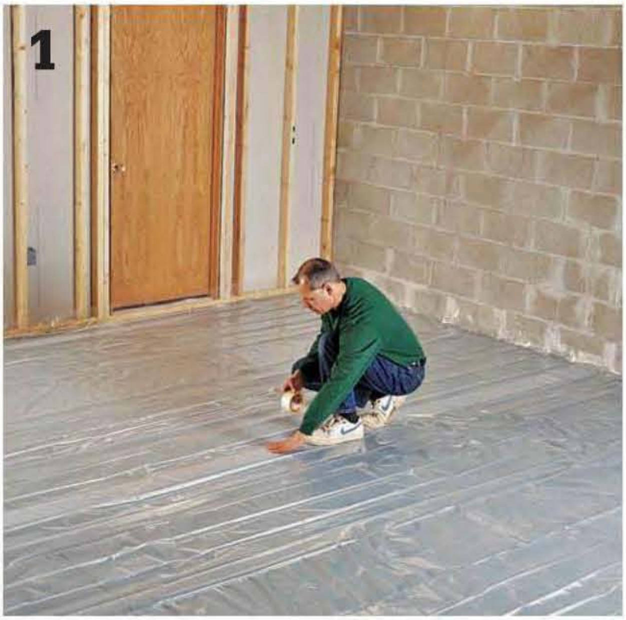

Chip away loose or protruding concrete with a masonry chisel and hammer, then vacuum the floor. Roll out strips of 6-mil polyethylene sheeting, extending them 31" up each wall, Overlap strips by 6", then seal the seams with packing tape. Temporarily tape the edges along the walls. Be careful not to damage the sheeting.

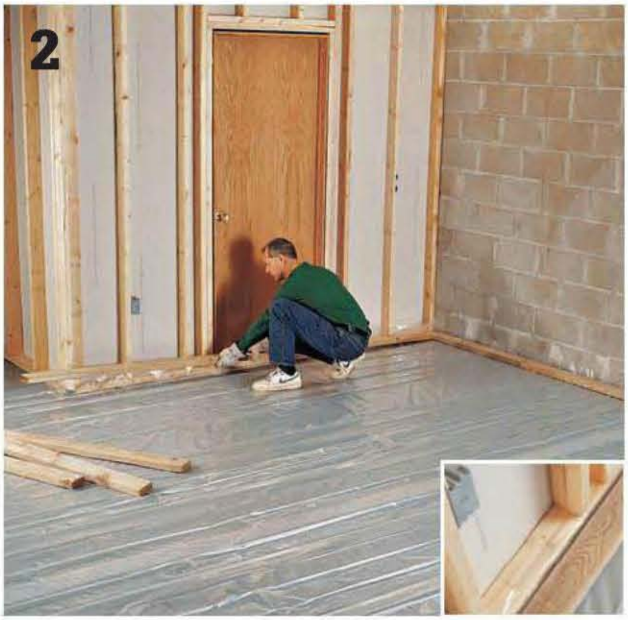

Lay out pressure-treated 2 x 4s along the perimeter of the room. Position the boards ½" in from all walls (inset). *Note: Before laying out the sleepers, determine where the partition walls will go. If a wall will fall between parallel sleepers, add an extra sleeper to support the planned wall.*

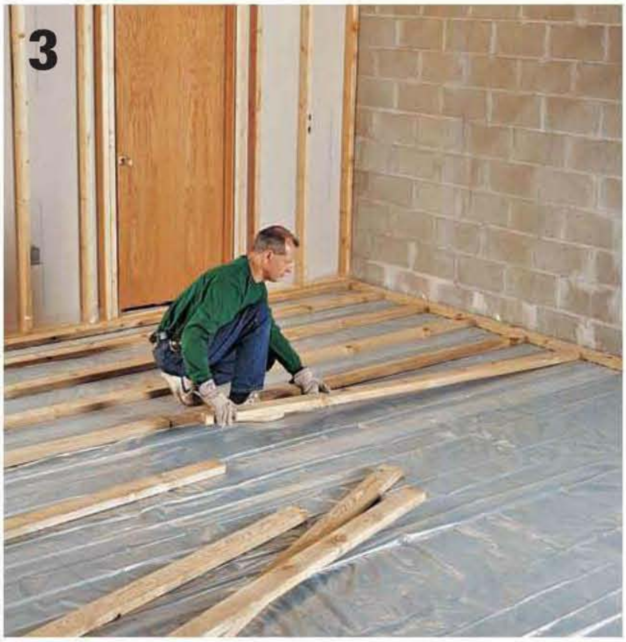

Using a circular saw, cut the sleepers to fit between the perimeter boards, leaving a ¼" gap at each end. Position the first sleeper so its center is 16" from the outside edge of the perimeter board. Lay out the remaining sleepers, using 16"-on-center spacing.

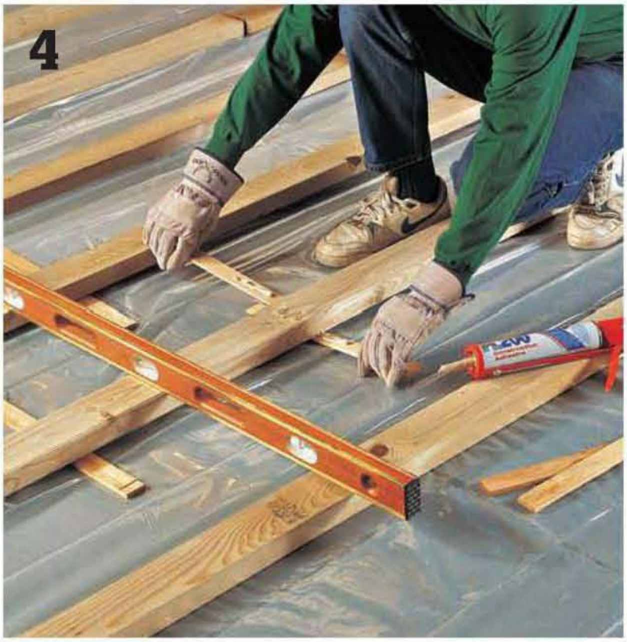

Where necessary, use tapered cedar shims to compensate for dips and variations in the floor. Place a 4-ft. level across neighboring sleepers. Apply construction adhesive to two wood shims. Slide the shims under the board from opposite sides until the board is level with adjacent sleepers.

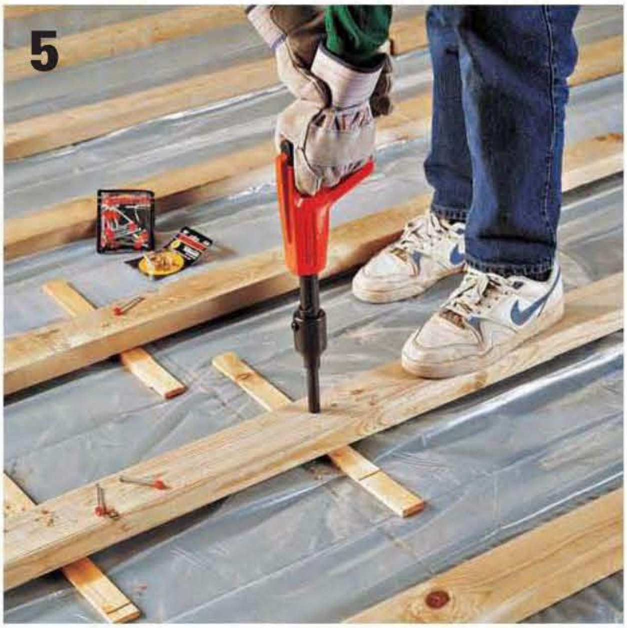

Fasten the perimeter boards and sleepers to the floor using a powder-actuated nailer or masonry screws. Drive a fastener through the center of each board at 16" intervals. Fastener heads should not protrude above the board's surface. Place a fastener at each shim location, making sure the fastener penetrates both shims.

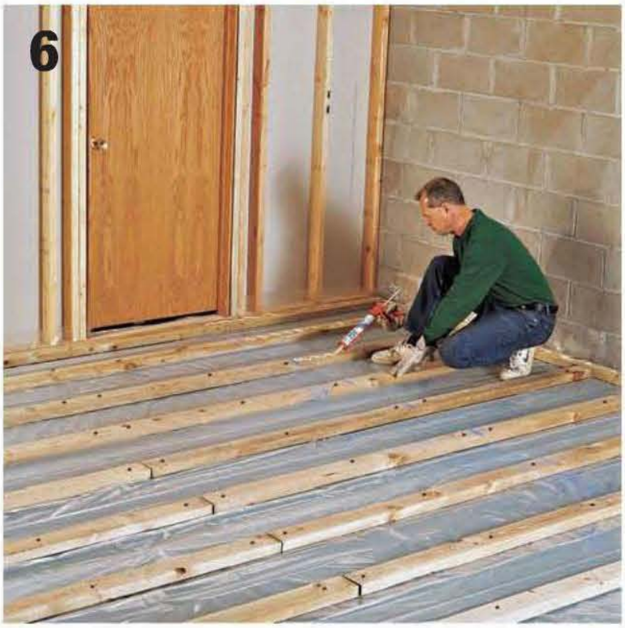

Establish a control line for the first row of plywood sheets by measuring 49" from the wall and marking the outside sleeper at each end of the room. Snap a chalk line across the sleepers at the marks. Run a ¼"-wide bead of adhesive along the first six sleepers, stopping just short of the control line.

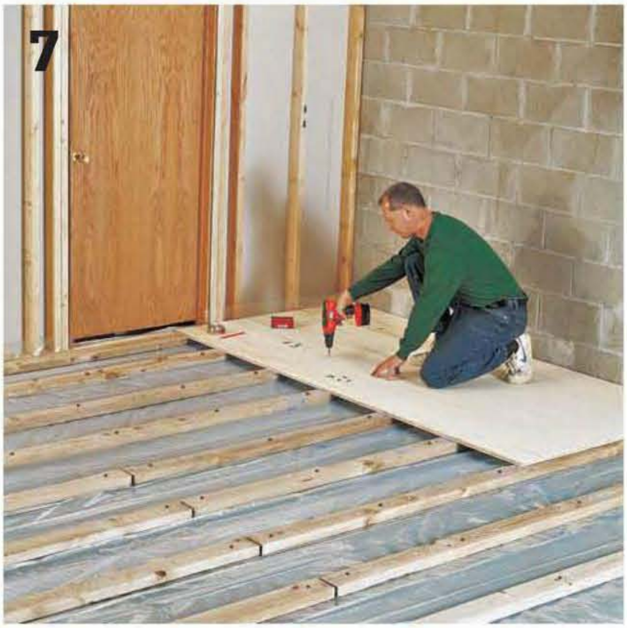

Position the first sheet of plywood so the end is ½" away from the wall and the grooved edge is flush with the control line. Fasten the sheet to the sleepers using 2" wallboard screws. Drive a screw every 6" along the edges and every 8" in the field. Don't drive screws along the grooved edge until the next row of sheeting is in place.

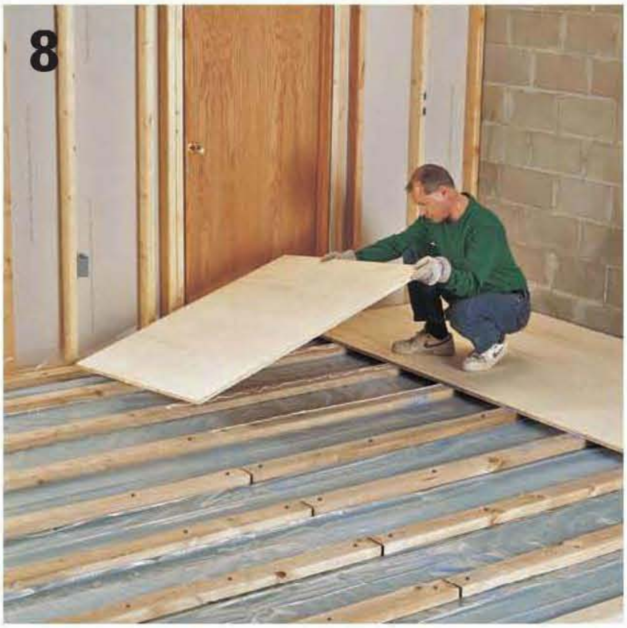

Install the remaining sheets in the first row, maintaining an ⅛" gap between ends. Begin the second row with a half-sheet (4 ft. long) so the end joints between rows are staggered. Fit the tongue of the half sheet into the groove of the adjoining sheet. If necessary, use a sledgehammer and wood block to help close the joint. After completing the second row, begin the third row with a full sheet. Alternate this pattern until the subfloor is complete.

Installing Attic Floors

Before you build the walls that will define the rooms in your attic, you'll need a sturdy floor beneath it all. Existing floors in most unfinished attics are merely ceiling joists for the floor below and are too small to support a living space.

There are several options for strengthening your attic's floor structure. The simplest method is to install an additional, identically sized joist next to each existing joist, connecting the two with nails. This process is known as *sistering*.

Sistering doesn't work when joists are smaller than 2 × 6s, are spaced too far apart, or where there are obstructions, such as plaster keys, from the ceiling below. An alternative is to build a new floor by placing larger joists between the existing ones. By resting the joists on 2 × 4 spacers, you avoid obstructions and minimize damage to the ceiling surface below. However, be aware that the spacers will reduce your headroom by 1½", plus the added joist depth.

To determine the best option for your attic, consult an architect, engineer, or building contractor, as well as a local building inspector. Ask what size of joists you'll need and which options are allowed in your area. Joist sizing is based on the span (the distance between support points), the joist spacing (typically 16" or 24" on-center), and the type of lumber used. In most cases, an attic floor must be able to support 40 pounds per sq. ft. of live load (occupants, furniture) and 10 psf dead load (wallboard, floor covering).

The floor joist cavities offer space for concealing the plumbing, wiring, and ductwork servicing your attic, so consider these systems as you plan. You'll also need to locate partition walls to determine if any additional blocking between joists is necessary.

When the framing is done, the mechanical elements and insulation are in place, and everything has been inspected and approved, complete the floor by installing ¾" tongue-and-groove plywood. If your remodel will include kneewalls, you can omit the subflooring behind the kneewalls, but there are good reasons not to: A complete subfloor will add strength to the floor, and will provide a sturdy surface for storage.

Tools & Materials ▸

Circular saw
Rafter square
Drill
Caulk gun
2 × joist lumber
2 × 4 lumber
16d, 10d, and 8d common nails
¾" T&G plywood
Construction adhesive
2¼" wallboard screws

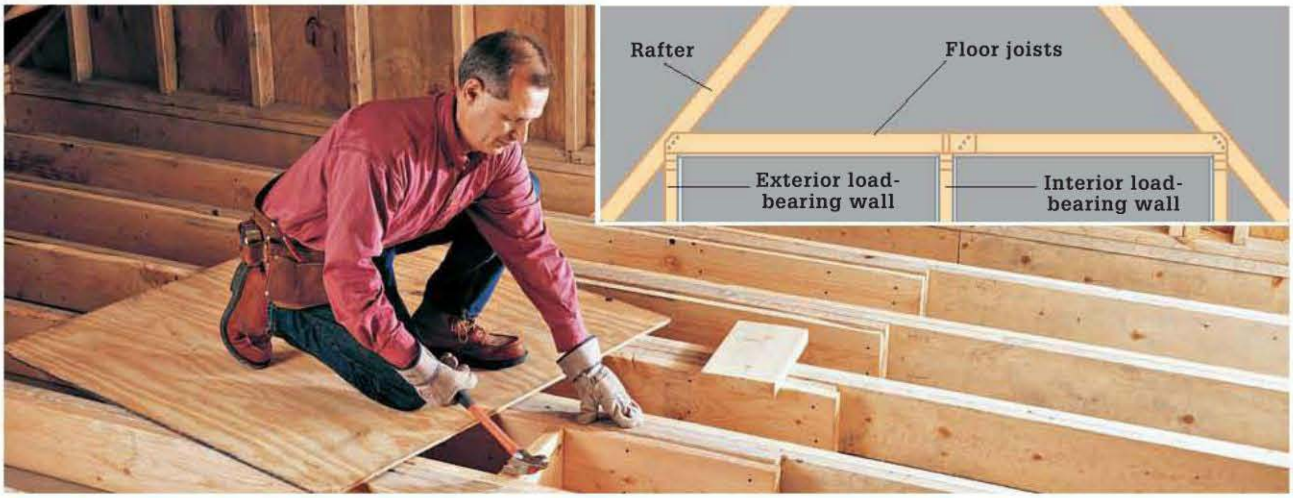

Attic joists typically rest on top of exterior walls and on an interior load-bearing wall, where they overlap from side to side and are nailed together. Always use a sheet of plywood as a platform while working over open joists.

How to Add Sister Joists

Remove all insulation from the joist cavities and carefully remove any blocking or bridging between the joists. Determine the lengths for the sister joists by measuring the existing joists. Also measure at the outside end of each joist to determine how much of the top corner was cut away to fit the joist beneath the roof sheathing. *Note: Joists that rest on a bearing wall should overlap each other by at least 3".*

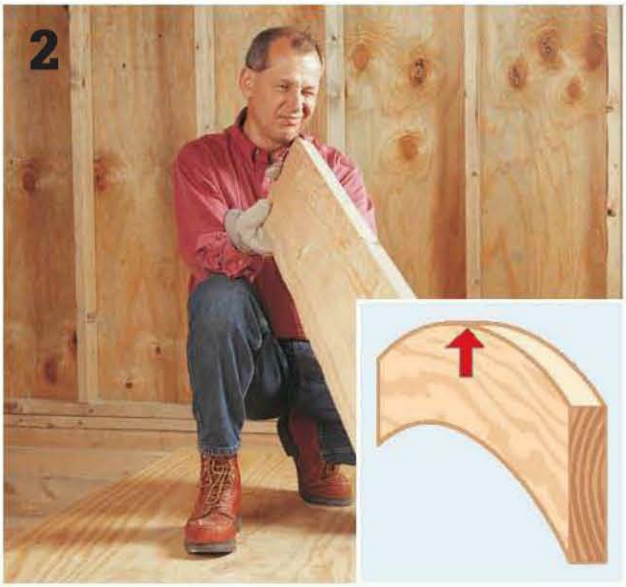

Before cutting, sight down both narrow edges of each board to check for crowning—upward arching along the length of the board. Draw an arrow that points in the direction of the arch. Joists must be installed "crown-up;" this arrow designates the top edge. Cut the board to length, then clip the top, outside corner to match the existing joists.

Set the sister joists in place, flush against the existing joists and with their ends aligned. Toenail each sister joist to the top plates of both supporting walls, using two 16d common nails.

Nail the joists together using 10d common nails. Drive three nails in a row, spacing the rows 12 to 16" apart. To minimize damage to the ceiling surface below caused by the hammering (such as cracking and nail popping), you can use an air-powered nail gun (available at rental stores), or 3" lag screws instead of nails. Install new blocking between the sistered joists, as required by the local building code.

How to Build a New Attic Floor

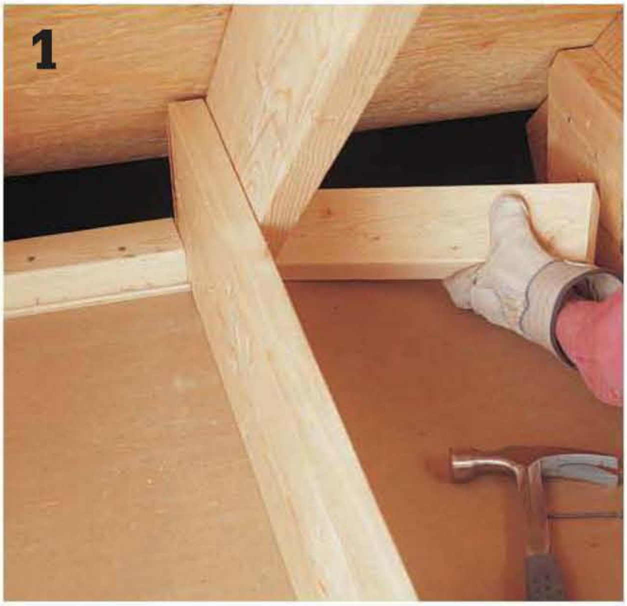

Remove any blocking or bridging from between the existing joists, being careful not to disturb the ceiling below. Cut 2 × 4 spacers to fit snugly between each pair of joists. Lay the spacers flat against the top plate of all supporting walls, and nail them in place with 16d common nails.

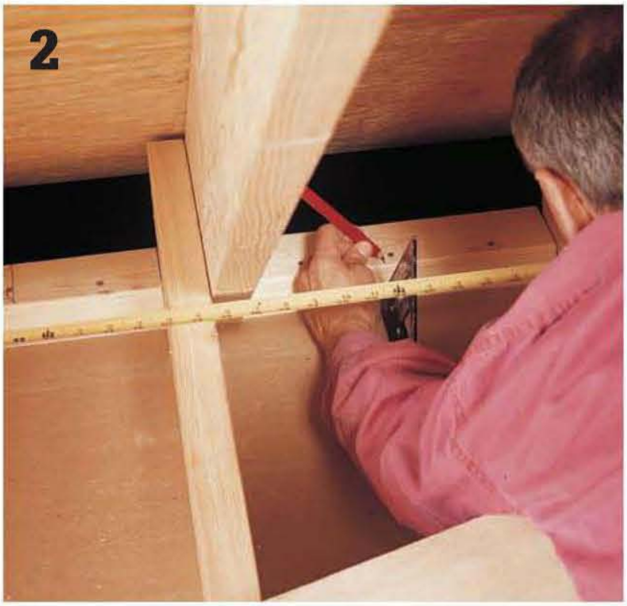

Create a layout for the new joists by measuring across the tops of the existing joists and using a rafter square to transfer the measurements down to the spacers. Following 16"-on center spacing, mark the layout along one exterior wall, then mark an identical layout onto the interior bearing wall. Note that the layout on the opposing exterior wall will be offset 1½", to account for the joist overlap at the interior wall.

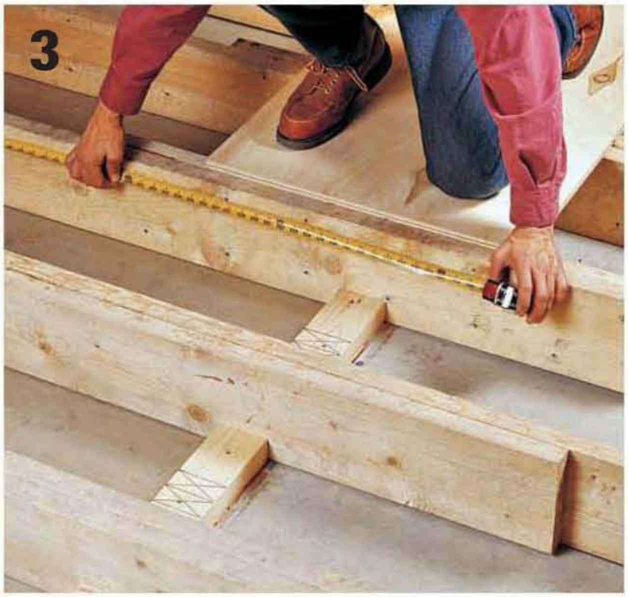

To determine joist length, measure from the outer edge of the exterior wall to the far edge of the interior bearing wall. The joists must overlap each other above the interior wall by 3". Before cutting, mark the top edge of each joist. Cut the joists to length, then clip the top outside corners so the ends can fit under the roof sheathing.

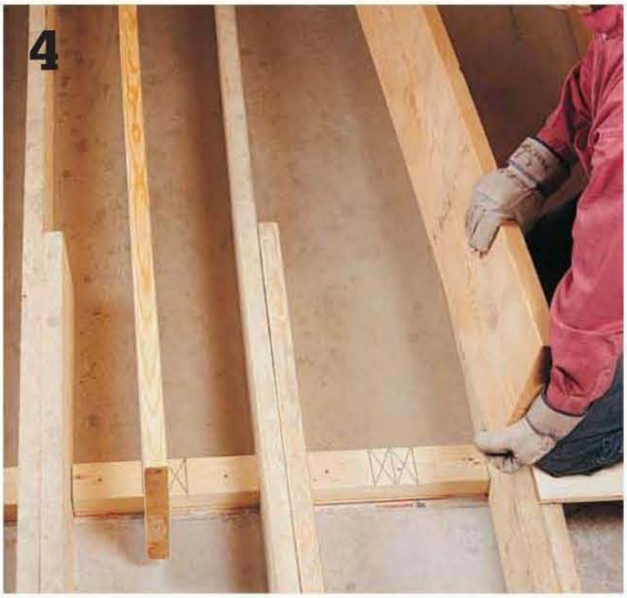

Set the joists in place on their layout marks. Toenail the outside end of each joist to the spacer on the exterior wall, using three 8d common nails.

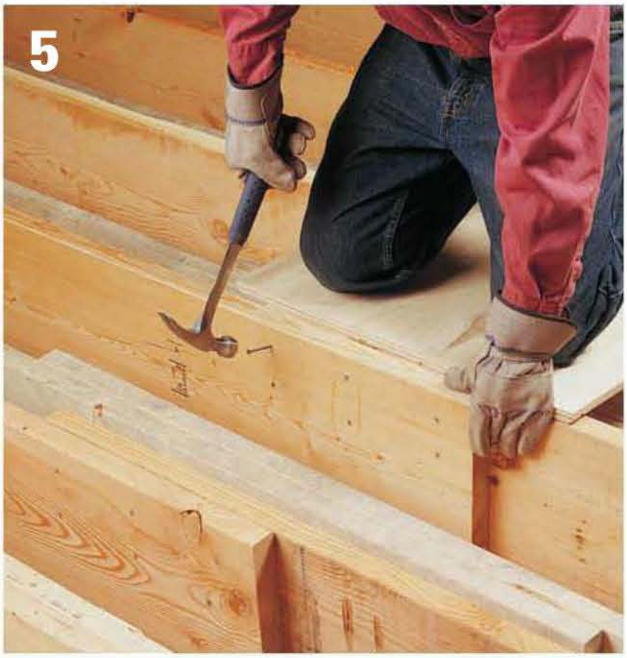

Nail the joists together where they overlap atop the interior bearing wall, using three 10d nails for each. Toenail the joists to the spacers on the interior bearing wall, using 8d nails.

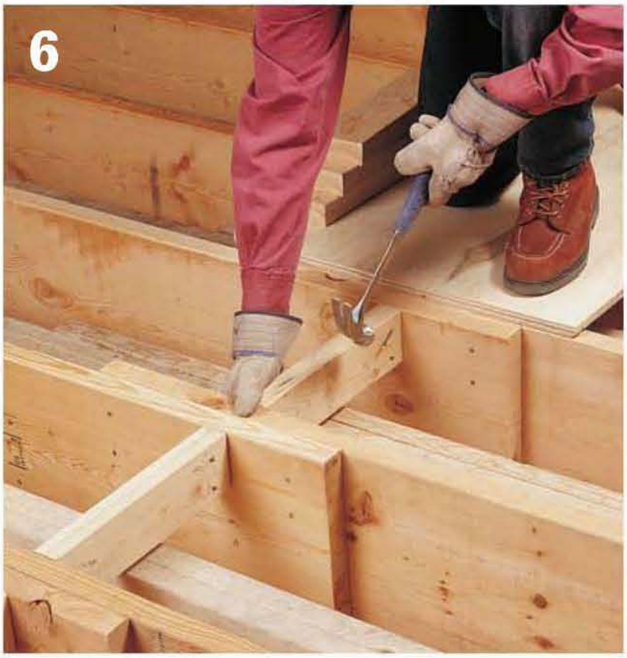

Install blocking or bridging between the joists, as required by the local building code. As a suggested minimum, the new joists should be blocked as close as possible to the outside ends and where they overlap at the interior wall.

How to Install Subflooring

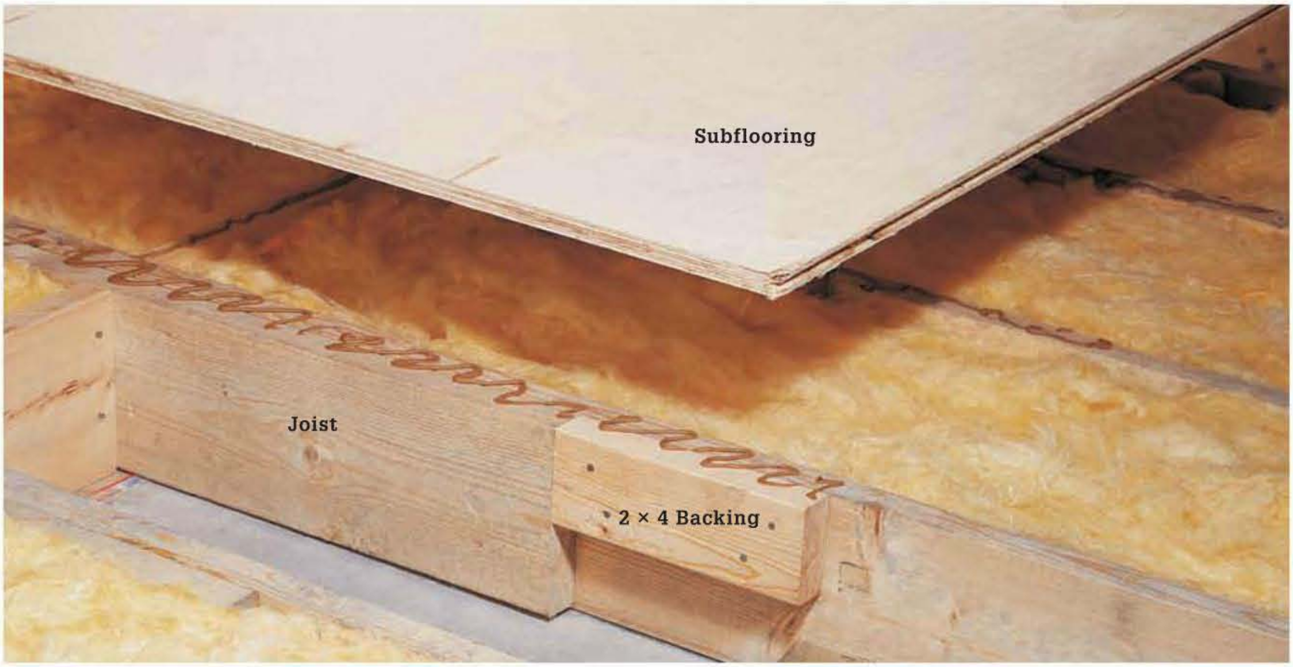

Install the subflooring only after all framing, plumbing, wiring, and ductwork is completed and has received the required building inspections. Also install any insulation and complete any caulking necessary for soundproofing. Fasten the sheets with construction adhesive and 2¼" wallboard or deck screws, making sure the sheets are perpendicular to the joists and the end joints are staggered between rows. Where joists overlap at an interior bearing wall, add backing as needed to compensate for the offset in the layout. Nail a 2 × 4 or wider board to the face of each joist to support the edges of the intervening sheets.

Installing Basement Walls

There are two common methods for covering foundation walls. Because it saves space, the more popular method is to attach 2 × 2 furring strips directly to the masonry wall. These strips provide a 1½"-deep cavity between strips for insulation and service lines, as well as a framework for attaching wallboard. The other method is to build a complete 2 × 4 stud wall just in front of the foundation wall. This method offers a full 3½" for insulation and lines, and it provides a flat, plumb wall surface, regardless of the foundation wall's condition.

To determine the best method for your project, examine the foundation walls. If they're fairly plumb and flat, you can consider furring them. If the walls are wavy or out of plumb, however, it may be easier to build stud walls. Also check with the local building department before you decide on a framing method. There may be codes regarding insulation minimums and methods of running service lines along foundation walls.

A local building official can also tell you what's recommended—or required—in your area for sealing foundation walls against moisture. Common types of moisture barriers include masonry waterproofers that are applied like paint and plastic sheeting installed between masonry walls and wood framing. The local building code will also specify whether you need a vapor barrier between the framing and the wallboard.

Before you shop for materials, decide how you'll fasten the wood framing to your foundation walls and floor. If you're covering a large wall area, it will be worth it to buy or rent a powder-actuated nailer for the job.

Tools & Materials ▸

- Caulk gun
- Trowel
- Paint roller
- Circular saw
- Drill
- Powder-actuated nailer
- Plumb bob
- Silicone caulk
- Paper-faced insulation
- Hydraulic cement
- Masonry waterproofer
- 2 × 2 and 2 × 4 lumber
- 2½" wallboard screws
- Construction adhesive
- Concrete fasteners
- Insulation

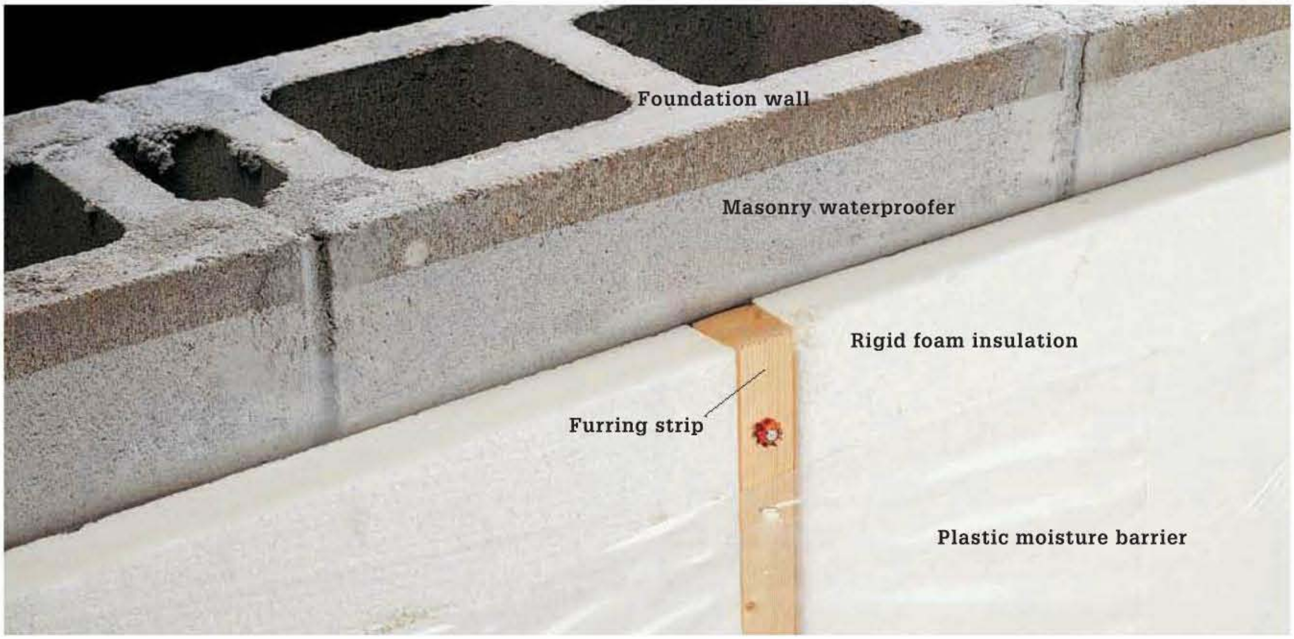

Basement walls can be made more livable by installing rigid foam insulation board between wood furring strips and then adding mold-resistant wallcovering. Although codes and practices vary dramatically, the current recommendation is to avoid the use of vapor barriers on the interior side of the wall.

How to Seal and Prepare Masonry Walls

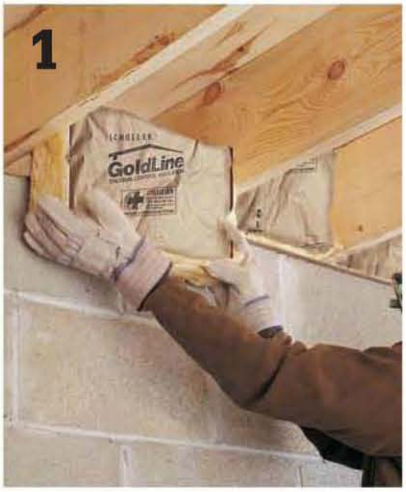

Insulate the rim-joist cavities (above the foundation walls) with solid pieces of paper-faced fiberglass insulation. Make sure the paper, which serves as a vapor barrier, faces the room.

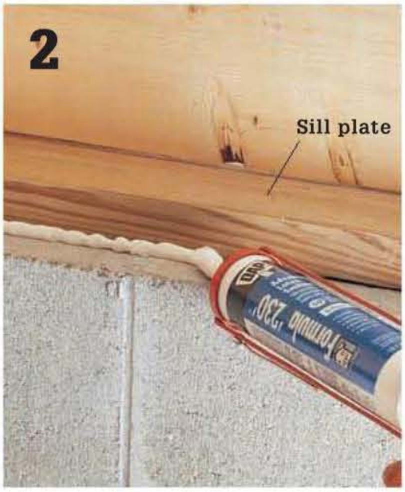

Apply silicone caulk to the joints between the sill plates and the foundation walls.

Fill small cracks with hydraulic cement or masonry caulk, and smooth the excess with a trowel. Ask the building department whether masonry waterproofer or a plastic moisture barrier is required in your area. Apply waterproofer as directed by the manufacturer, or install plastic sheeting following code specifications.

Options for Attaching Wood to Masonry

Masonry nails are the cheapest way to attach wood to concrete block walls. Drive the nails into the mortar joints for maximum holding power and to avoid cracking the blocks. Drill pilot holes through the strips if the nails cause splitting. Masonry nails are difficult to drive into poured concrete.

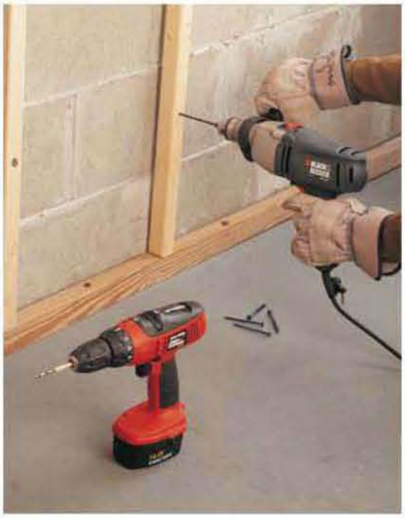

Self-tapping masonry screws hold well in block or poured concrete, but they must be driven into predrilled holes. Use a hammer drill to drill holes of the same size in both the wood and the concrete after the wood is positioned. Drive the screws into the web portion of the blocks.

(continued)

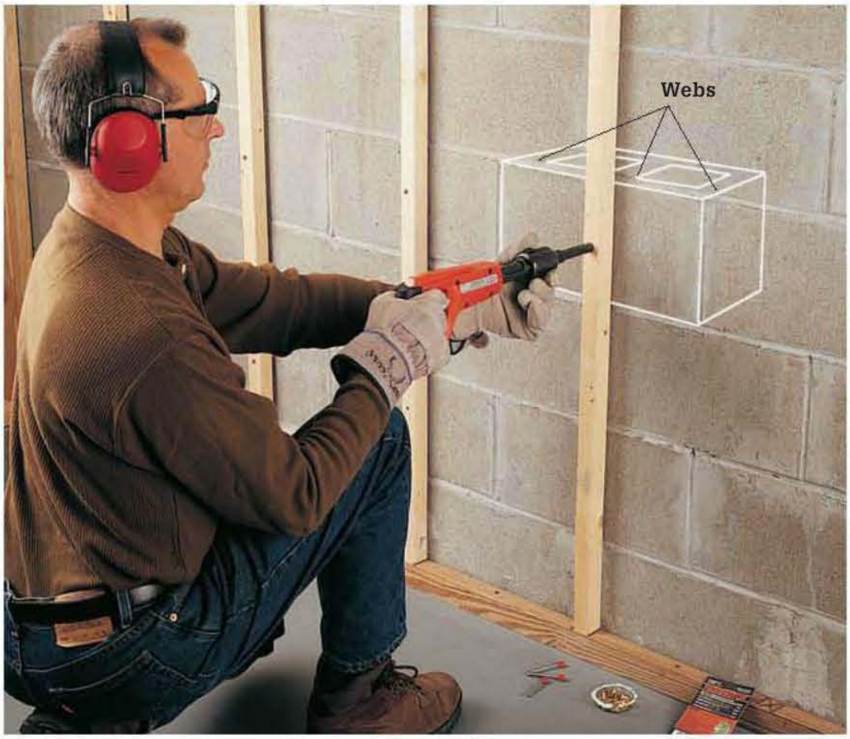

Powder-actuated nailers offer the quickest and easiest method for fastening framing to block, poured concrete, and steel. They use individual caps of gunpowder—called loads—to propel a piston that drives a hardened-steel nail (pin) through the wood and into the masonry. The loads are color-coded for the charge they produce, and the pins come in various lengths. *Note: Always drive pins into the solid web portions of concrete blocks, not into the voids. Trigger-type nailers, like the one shown here, are easiest to use, but hammer-activated types are also available. You can buy nailers at home centers and hardware stores, or rent them from rental centers. (Ask for a demonstration at the rental center.) Always wear hearing and eye protection when using these extremely loud tools.*

How to Install Furring Strips on Masonry Walls

1

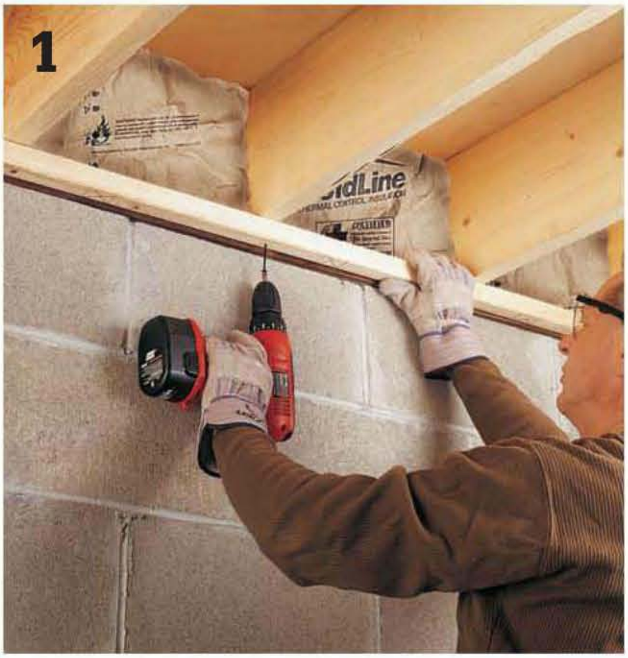

Cut a 2 × 2 top plate to span the length of the wall. Mark the furring-strip layout onto the bottom edge of the plate, using 16"-on-center spacing. Attach the plate to the bottom of the joists with 2½" wallboard screws. The back edge of the plate should line up with the front of the blocks.

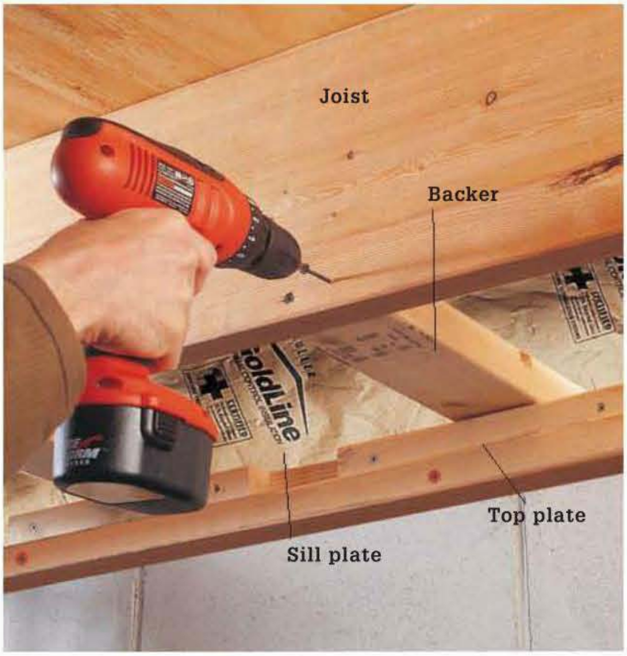

Note: If the joists run parallel to the wall, you'll need to install backers between the outer joist and the sill plate to provide support for ceiling wallboard. Make T-shaped backers from short 2 × 4s and 2 × 2s. Install each so the bottom face of the 2 × 4 is flush with the bottom edge of the joists. Attach the top plate to the foundation wall with its top edge flush with the top of the blocks.

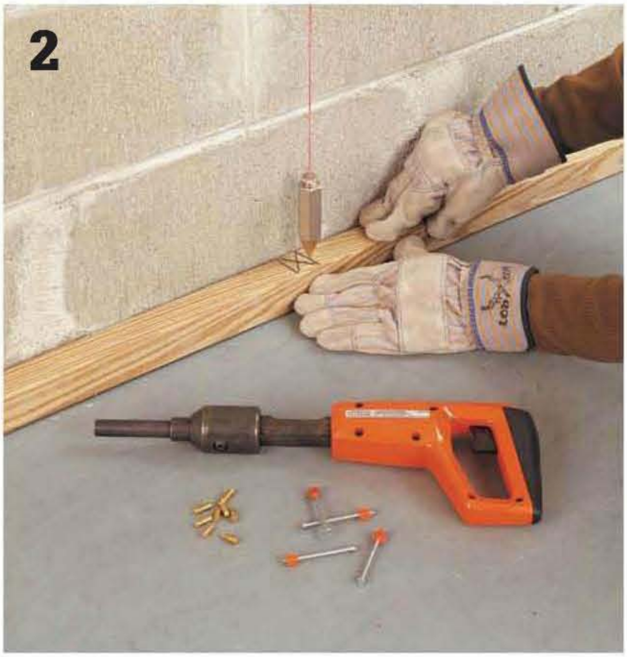

Install a bottom plate cut from pressure-treated 2 × 2 lumber so the plate spans the length of the wall. Apply construction adhesive to the back and bottom of the plate, then attach it to the floor with a nailer. Use a plumb bob to transfer the furring-strip layout marks from the top plate to the bottom plate.

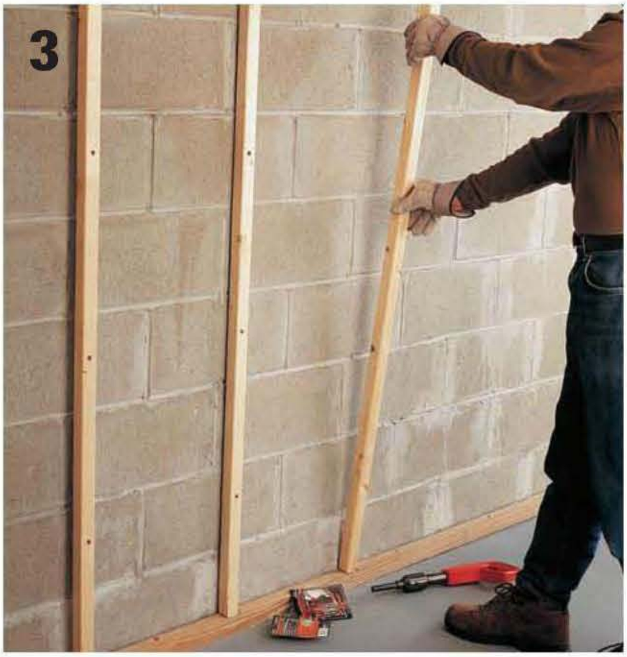

Cut 2 × 2 furring strips to fit between the top and bottom plates. Apply construction adhesive to the back of each furring strip, and position it on the layout marks on the plates. Nail along the length of each strip at 16" intervals.

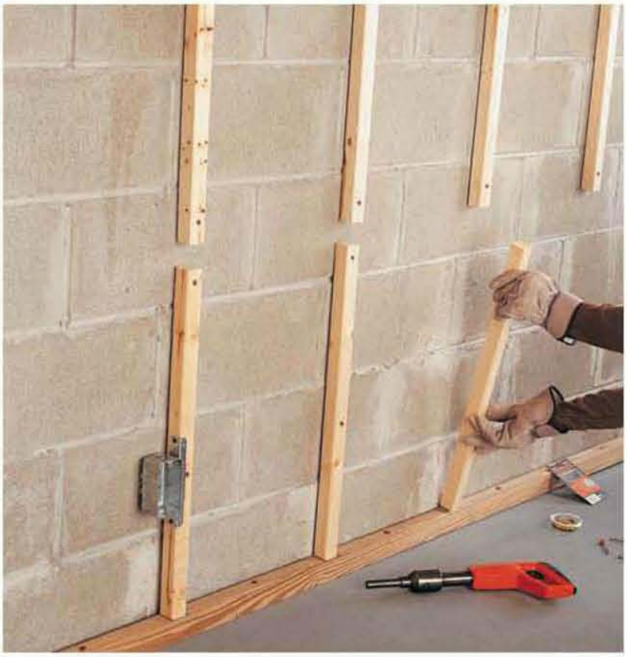

Option: Leave a channel for the installation of wires or supply pipes by installing pairs of vertically aligned furring strips with a 2" gap between each pair. *Note: Consult local codes to ensure proper installation of electrical or plumbing materials.*

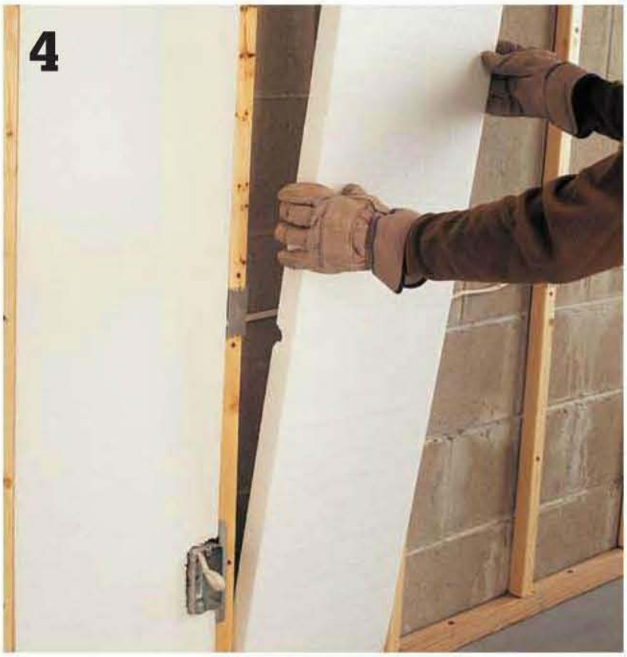

Fill the cavities between furring strips with rigid insulation board. Cut the pieces so they fit snugly within the framing. If necessary, make cutouts in the insulation to fit around mechanical elements, and cover any channels with metal protective plates before attaching the wall surface. Add a vapor barrier if required by local building code.

Installing Attic Walls

Attic kneewalls are short walls that extend from the attic floor to the rafters. They provide a vertical dimension to attic rooms, and without them, attics tend to feel cramped. Kneewalls are typically 5 ft. tall, for a couple of reasons: That's the minimum ceiling height for usable floor space according to most building codes, and it defines a comfortable room without wasting too much floor space. The unfinished space behind kneewalls doesn't have to go to waste: It's great for storage and for concealing service lines. To provide access to this space, create a framed opening in the wall during the framing process.

Kneewalls are similar to partition walls, except they have beveled top plates and angle-cut studs that follow the slope of the rafters. The added stud depth created by the angled cut requires a 2 × 6 top plate. Before starting on your kneewall project, it may help to review the techniques for building a partition wall.

Tools & Materials ▸

Circular saw
Level
Chalk line
T-bevel

2 × 4 and 2 × 6 lumber
16d
8d common nails

Attic kneewalls are just the right height to be backdrops for furniture, and they make a perfect foundation for built-in storage units.

How to Build a Kneewall

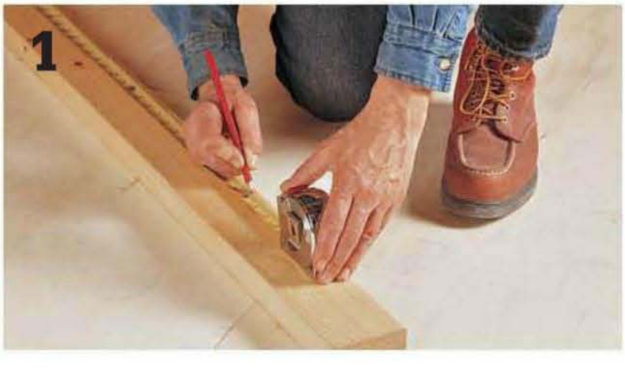
1

Create a storyboard using a straight 2 × 4. Cut the board a few inches longer than the planned height of the wall. Measure from one end and draw a line across the front edge of the board at the exact wall height.

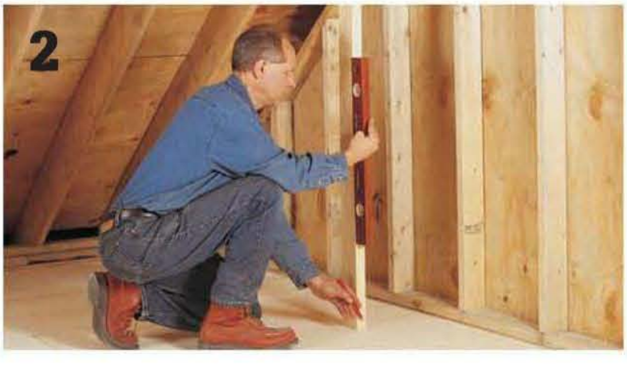
2

At one end of the room, set the storyboard flat against the outer rafter. Plumb the storyboard with a level while aligning the height mark with the bottom edge of the rafter. Transfer the height mark onto the rafter edge, then make a mark along the front edge of the storyboard onto the subfloor. These marks represent the top and bottom wall plates.

3

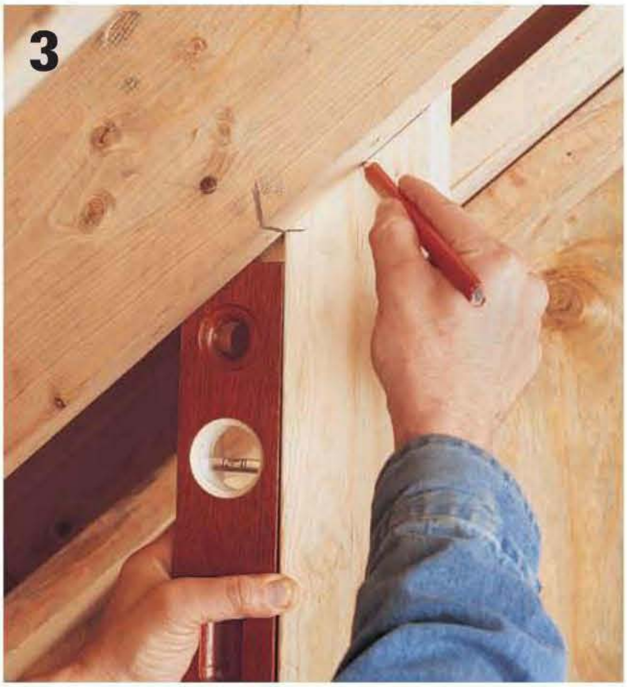

Holding the storyboard perfectly plumb, trace along the bottom edge of the rafter to transfer the rafter slope onto the face of the storyboard.

4

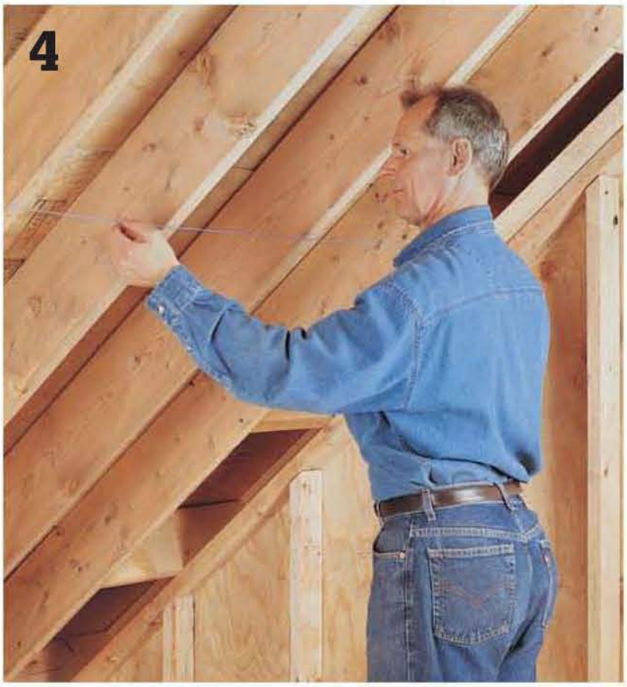

Repeat the wall-plate marking process on the other end of the room. Snap a chalk line through the marks—across the rafters and along the subfloor. If necessary, add backing for fastening the top plate to the gable wall.

5

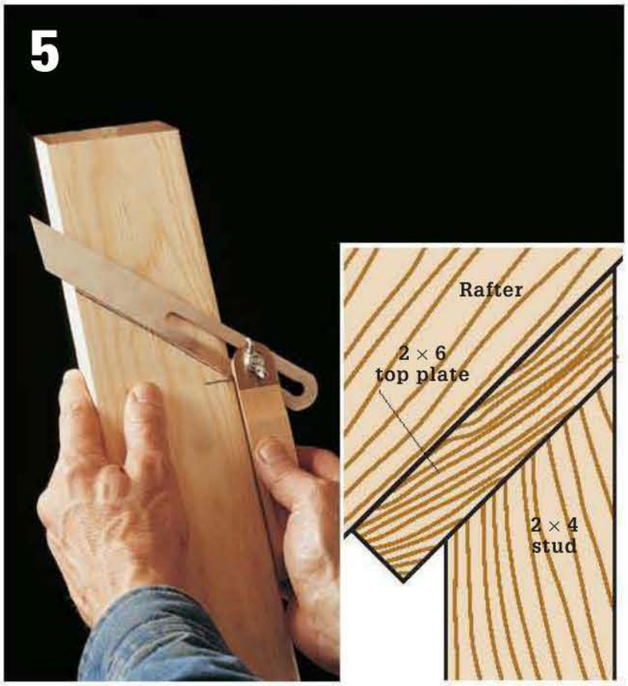

To cut a beveled edge on the top wall plate, set a T-bevel to match the rafter-slope line on the storyboard. Use the T-bevel to adjust the blade of a circular saw or table saw to the proper angle. Then, bevel-cut one edge of the 2 × 6 top plate. *Note: When the top plate is laid flat across the rafters, the front edge should be perpendicular to the floor.*

6

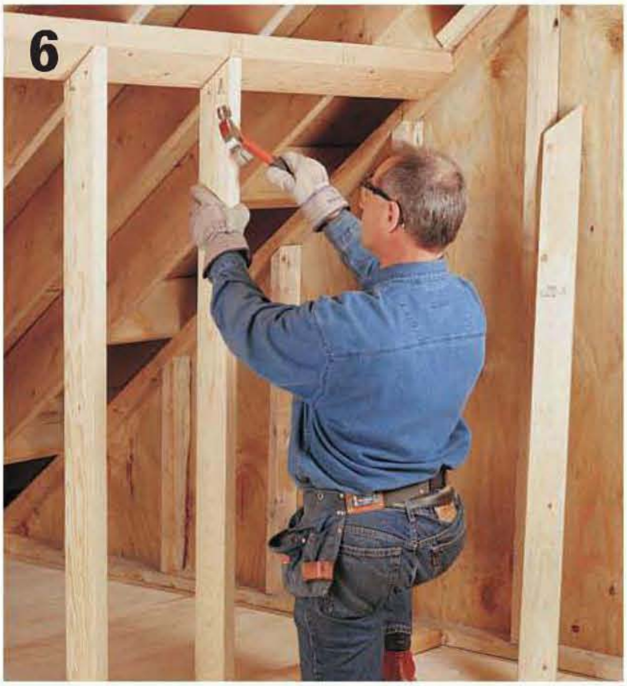

Mark the stud locations on the wall plates. Install the plates along the chalk lines, fastening them to the rafters and floor joists, respectively, using 16d nails. Measure and cut each stud to fit, angle-cutting the top end so that it meets flush with the top plate. Toenail each stud in place with three 8d nails.

Finishing Attic Ceilings

By virtue of sloping roofs, most attics naturally have "cathedral" ceilings. It's up to you whether to leave the peaks intact—and apply a finish surface all the way up to the ridge—or to frame-in a horizontal ceiling, creating a flat surface that's more like a standard ceiling. Before deciding, consider the advantages and disadvantages of each treatment.

If your attic has collar ties—horizontal braces installed between opposing rafters—your planning should start with those. Are the ties high enough to meet the code requirements for attic headroom? If not, consult an architect or engineer to see if you can move them up a few inches (do not move or remove them without professional guidance). If the ties are high enough, you can incorporate them into a new ceiling or leave them exposed and wrap them with a finish material, such as wallboard or finish-grade lumber. Do not use collar ties as part of your ceiling frame.

A peaked ceiling is primarily an aesthetic option. Its height expands the visual space of the room, and its rising angles provide a dramatic look that's unique in most homes. Because a peaked ceiling encloses the rafter bays all the way up to the ridge, this treatment may require additional roof vents to maintain proper ventilation.

By contrast, a flat ceiling typically offers a cleaner, more finished appearance closer to that of a conventional room, and flat ceilings offer some practical advantages over peaked styles. First, they provide a concealed space above the ceiling, great for running service lines. If there are vents high on the gable walls, this open space can help ventilate the roof (make sure to insulate above the ceiling). The ceiling itself can hold recessed lighting fixtures or support a ceiling fan. And if your plans call for full-height partition walls, you may want a ceiling frame to enclose the top of the wall.

When determining the height of flat-ceiling framing, be sure to account for the floor and ceiling finishes. And remember that most building codes require a finished ceiling height of at least 90".

Tools & Materials ▸

4-ft. level
Chalk line
Circular saw
2 × 4 and 2 × 6 lumber
10d common nails

Flat attic ceilings provide space for recessed light fixtures, vents, and speakers.

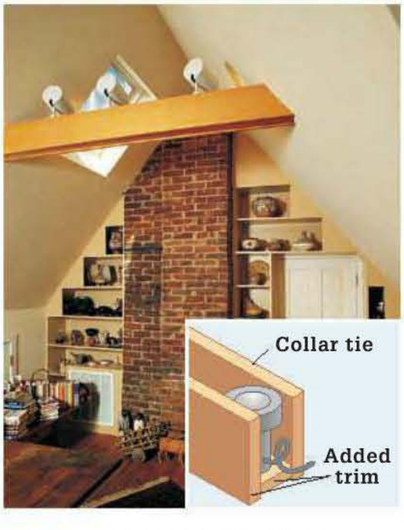

Exposed collar ties can add an interesting architectural element to a peaked ceiling. By adding to the existing ties, you can create a channel for holding small light fixtures.

How to Frame a Flat Ceiling

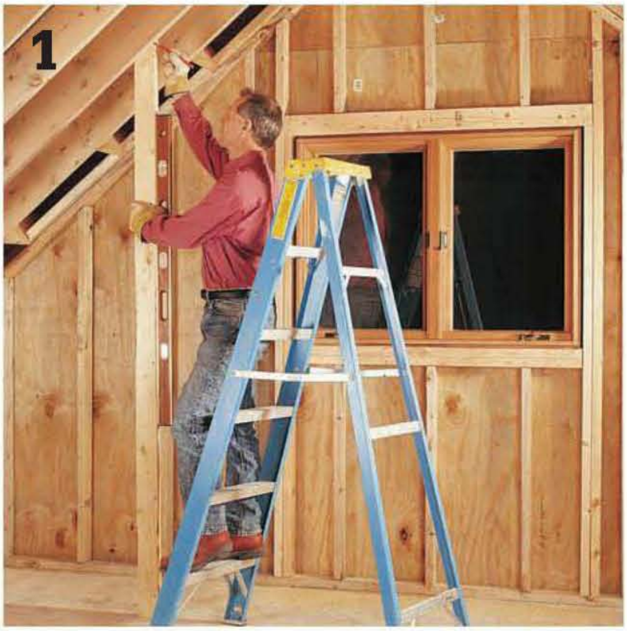

Make a storyboard for the planned height of the ceiling frame. At one end of the attic, hold the storyboard plumb and align the height mark with the bottom edge of a rafter. Transfer the mark to the rafter. Repeat at the other end of the attic, then snap a chalk line through the marks. This line represents the bottom edge of the ceiling frame.

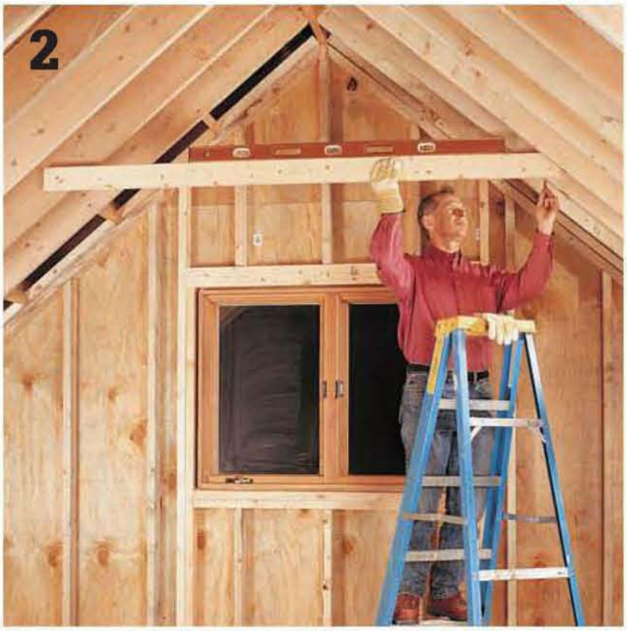

Using a level and the storyboard, level over from the chalk line and mark two outside rafters on the other side of the attic. Snap a chalk line through the marks. *Note: The storyboard is used merely as a straightedge for this step.*

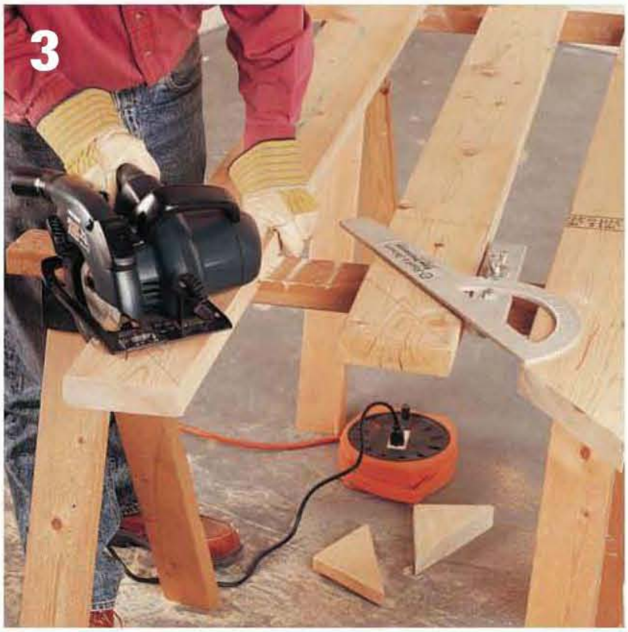

Cut 2 × 6 joists to span across the rafters, angle-cutting the ends to follow the roof pitch. Check each joist for crowning to make sure you're cutting it so it will be installed with the crowned edge up. Make the overall length about ½" short so the ends of the joists won't touch the roof sheathing.

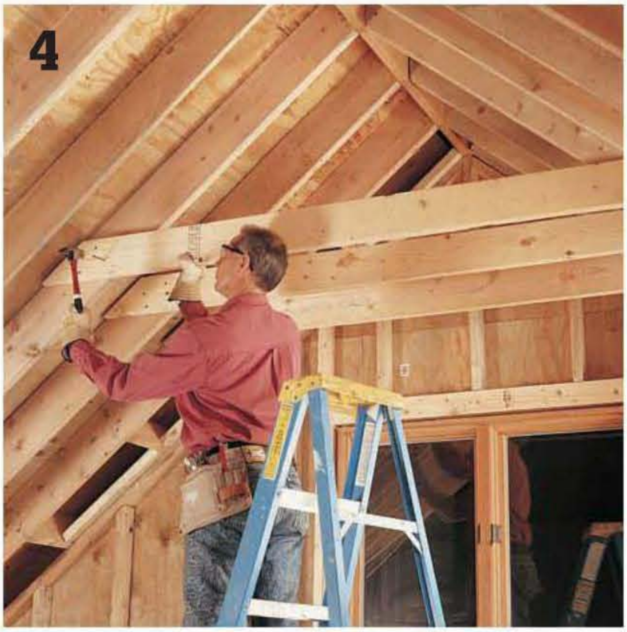

Nail each joist to the rafters with three 10d common nails at each end. Be sure to maintain 16"- or 24"- on-center spacing between joists to provide support for attaching wallboard or other finish material.

Adding Basement Egress Windows

An egress window brings a pleasant source of natural light and ventilation to a dark, dank basement. More importantly, it can provide a life-saving means of escape in the event of a fire. Before you proceed with this project, read more about building code issues regarding basement egress on page 17. Contact your local building department to apply for the proper permits and to learn more about the code requirements for your area.

As long as the window opens wide enough to meet minimum standards for egress, the particular window style is really up to you. Casement windows are ideal, because they crank open quickly and provide unobstructed escape. A tall, double-hung window or wide sliding window can also work. Select a window with insulated glass and clad with vinyl or aluminum for durability; it will be subject to humidity and temperature fluctuations just like any other above-grade window in your home.

The second fundamental component of a basement egress window project is the subterranean escape well you install outside the foundation. There are several options to choose from: prefabricated well kits made of lightweight plastic that bolt together and are easy to install; corrugated metal wells are a lower-cost option; or, you can build a well from scratch using concrete, stone or landscape timber.

Installing an egress window involves four major steps: digging the well, cutting a new or larger window opening in the foundation, installing the window, and finally, installing the well. You'll save time and effort if you hire a backhoe operator to excavate the well. In most cases, you'll also need a large concrete saw (available at most rental stores) to cut the foundation wall.

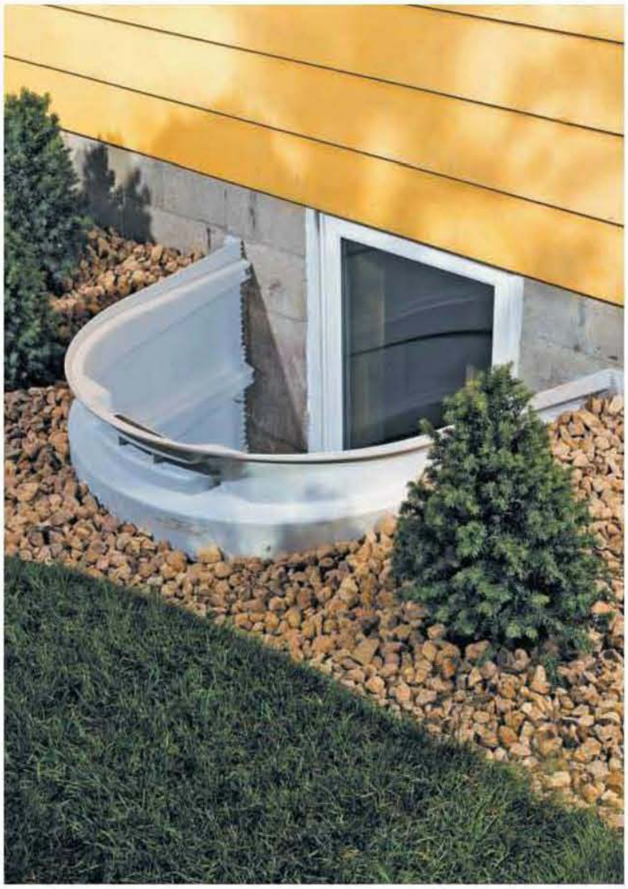

Replacing a small basement window with an egress window is a big job, but it is required if you want to convert part of a basement into livable space, especially a bedroom.

Tools & Materials ▸

- Tape measure
- 4-ft. level
- Stakes and string line
- Shovel
- Colored masking tape
- Hammer drill with ½". dia. × 12- to 16"-long masonry bit
- Concrete saw
- Hand maul
- Wide masonry chisel
- Trowel, miter saw
- Drill/driver, hammer
- Caulk and caulk gun
- Gloves
- Window well and window
- Pea gravel
- Plastic sheeting
- Self-tapping masonry screws
- 2× pressure-treated lumber
- Shims
- Insulation materials
- Concrete sleeve anchors

How to Install an Egress Window & Window Well

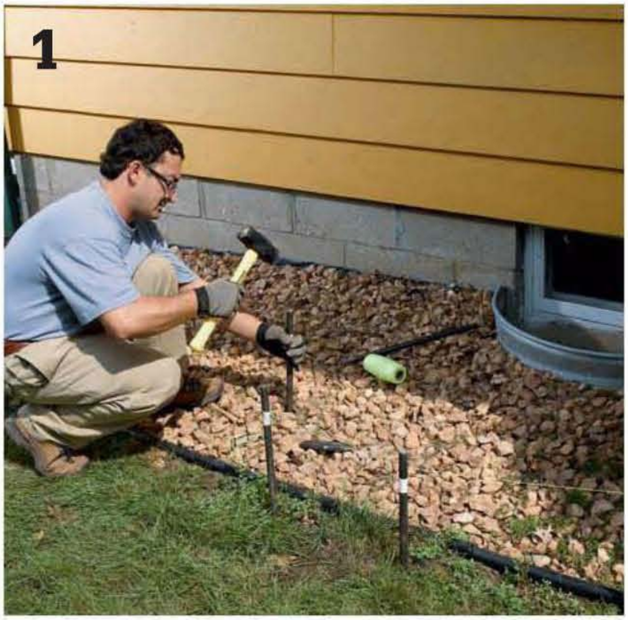

Lay out the border of the window well area with stakes and string. Plan the length and width of the excavation to be several feet larger than the window well's overall size to provide extra room for installation and adjustment.

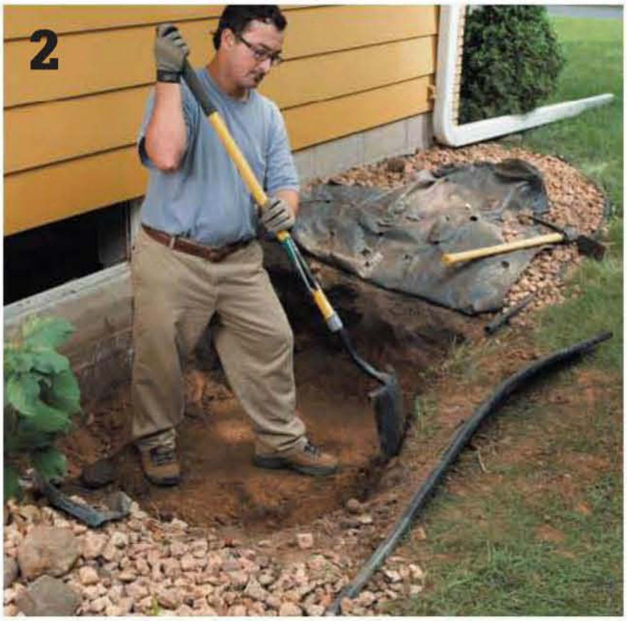

Excavate the well to a depth 6 to 12" deeper than the well's overall height to allow room for drainage gravel. Make sure to have your local public utilities company inspect the well excavation area and okay it for digging before you start.

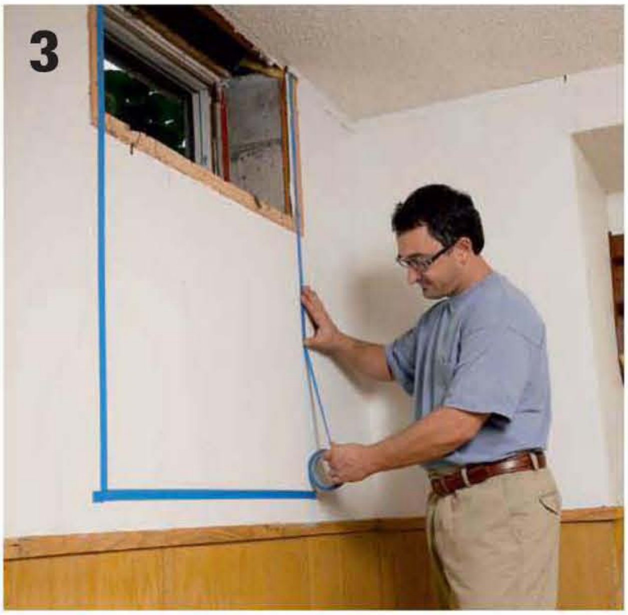

Measure and mark the foundation wall with brightly colored masking tape to establish the overall size of the window's rough opening (here, we're replacing an existing window). Be sure to take into account the window's rough opening dimensions, the thickness of the rough framing (usually 2x stock) and the width of the structural header you may need to build. Remember also that sill height must be within 44" of the floor. Remove existing wall coverings inside the layout area.

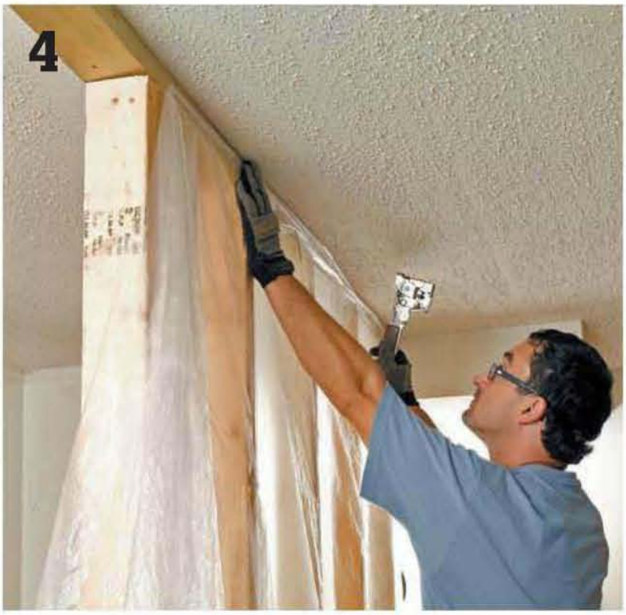

If the floor joists run perpendicular to your project wall, build a temporary support wall parallel to the foundation wall and 6 to 8 ft. from it (for more on building support walls, see page 29). Staple sheet plastic to the wall and floor joists to form a work tent that will help control concrete dust.

(continued)

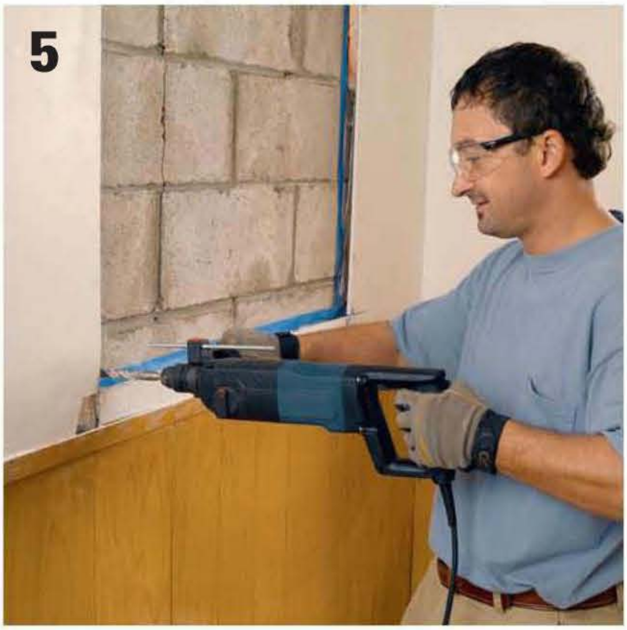

You'll need to cut through the foundation from both the inside and outside so the blocks will break cleanly. Drill a hole through the wall at each bottom corner with a hammer drill and long masonry bit to give you reference points for marking the outside cuts.

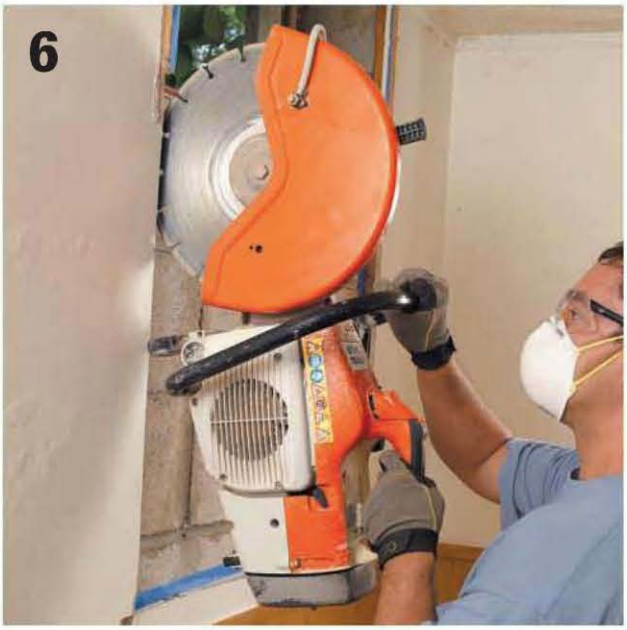

Equip a masonry cutting saw (or large angle grinder) with a diamond blade and set it for a ½" cut to score the blocks first, then reset the saw to full depth and make the final bottom and side cuts through the blocks. Wear a tight-fitting particle mask, ear and eye protection, and gloves for all of this cutting work; the saw will generate a tremendous amount of thick dust. Feed the saw slowly and steadily. Stop and rest periodically so the dust can settle.

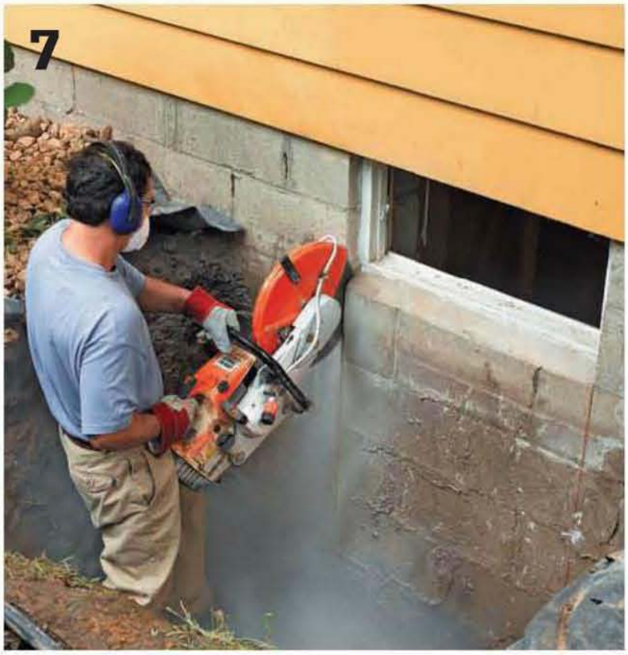

Connect the holes you drilled previously with a level and plumb lines to mark the outside foundation wall for cutting. Measure the opening between the holes to make sure it is the correct size and that it is square. Score the cuts first, then make the full-depth cuts.

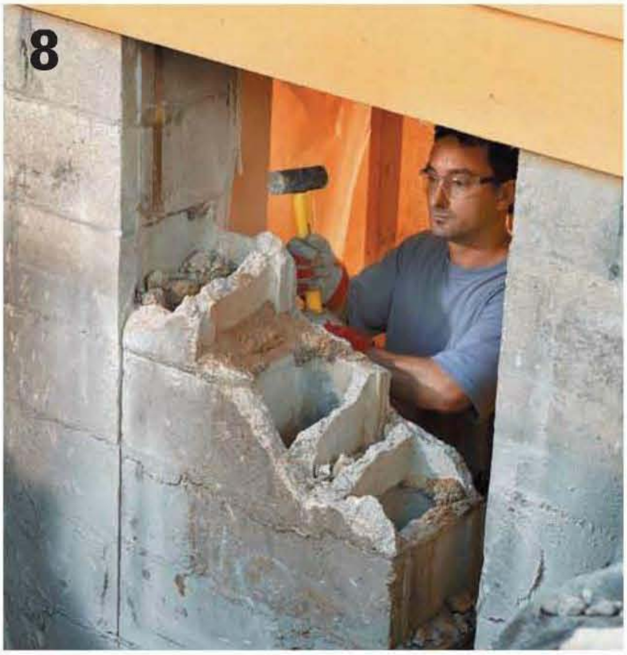

Strike the blocks with a hand maul to break or loosen the block sections. When all the blocks are removed, carefully chip away remaining debris with a cold chisel to create flat surfaces.

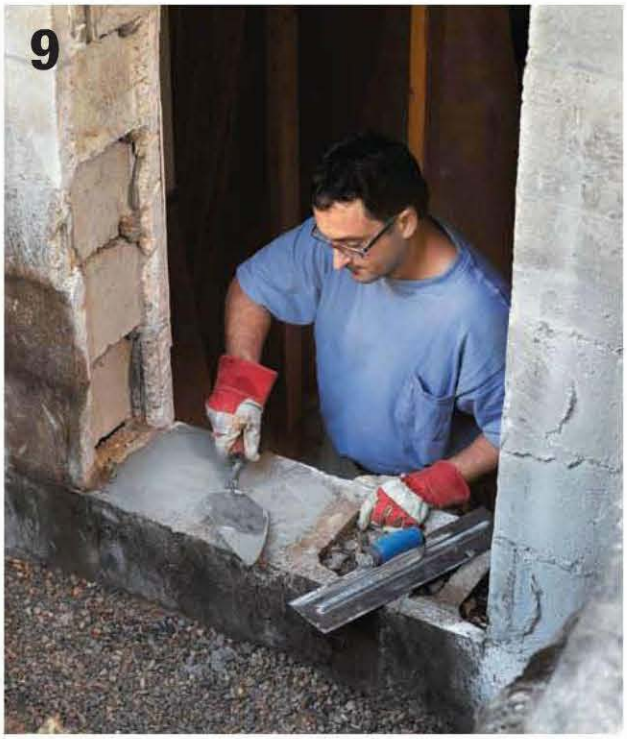
9

In concrete block walls, fill the hollow voids with broken pieces of block, then level and smooth the voids by trowelling on a fresh layer of quick-curing concrete. Flatten the surfaces, and allow the concrete to dry overnight.

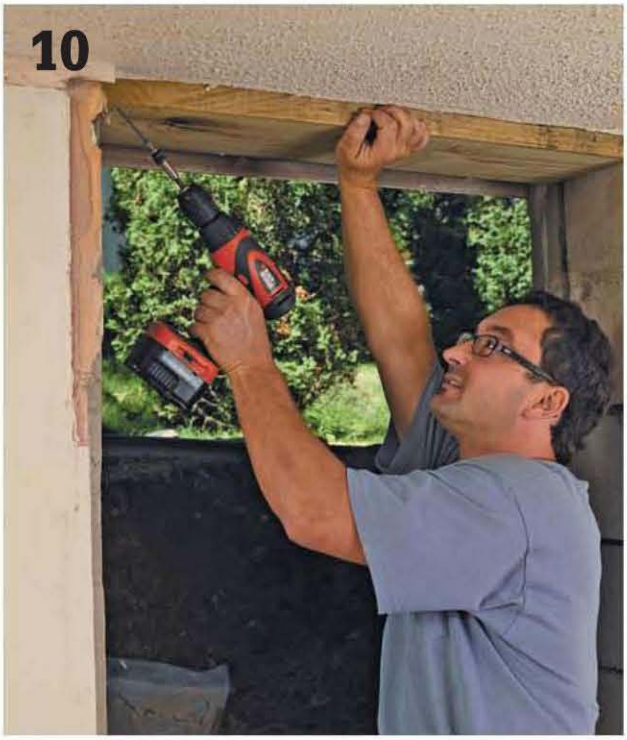
10

If your project requires a new header above the new window, build it from pieces of 2× lumber sandwiching some ½" plywood and fastened together with construction adhesive and 10d nails. Slip it into place and tack it temporarily to the mudsill with 3½" deck screws driven toenail style.

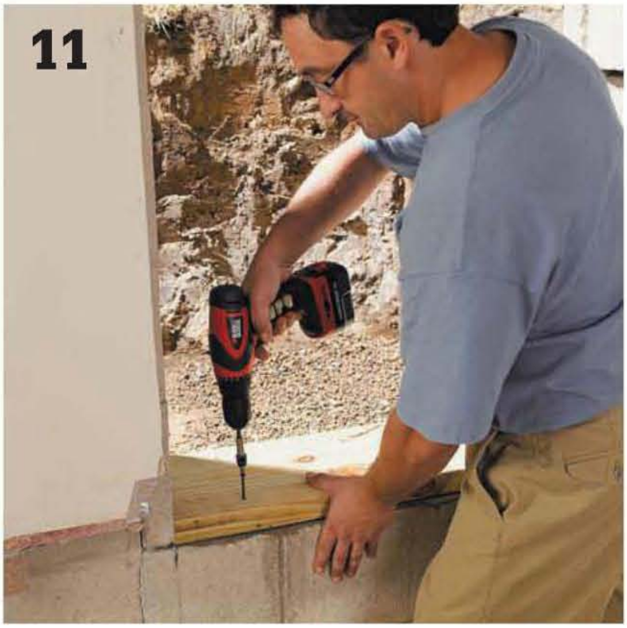
11

Cut the sill plate for the window's rough frame from 2× treated lumber that's the same width as the thickness of the foundation wall. Fasten the sill to the foundation with 3/16 × 3¼" countersunk masonry screws. Drill pilot holes for the screws first with a hammer drill.

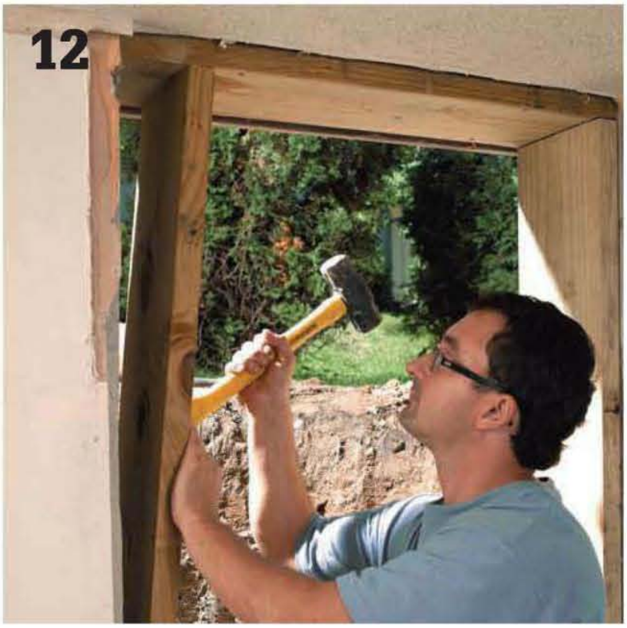
12

Cut two pieces of treated lumber just slightly longer than the opening so they'll fit tightly between the new header and sill. Tap them into place with a maul. Adjust them for plumb and fasten them to the foundation with countersunk masonry screws or powder-actuated fasteners.

(continued)

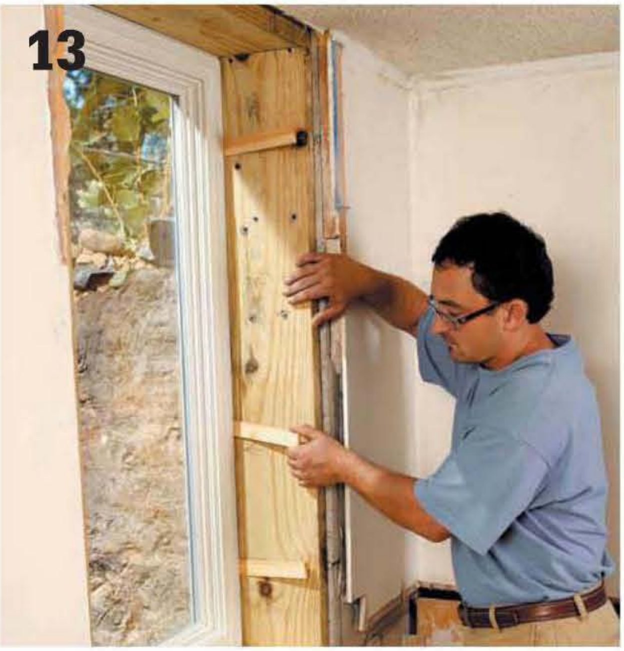

Apply a thick bead of silicone caulk around the outside edges of the rough frame and set the window in its opening, seating the nailing flanges into the caulk. Shim the window so the frame is level and plumb. Test the action of the window to make sure the shims aren't bowing the frame.

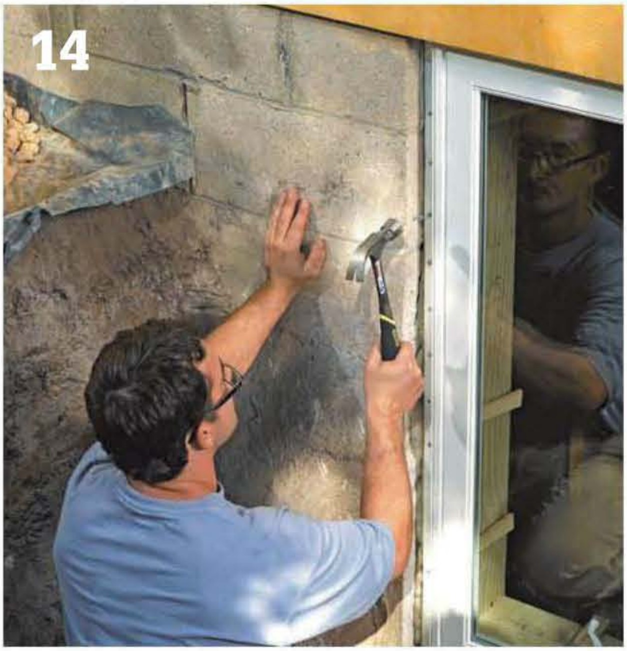

Attach the window's nailing flanges to the rough frame with screws or nails, as specified by the manufacturer. Check the window action periodically as you fasten it to ensure that it still operates smoothly.

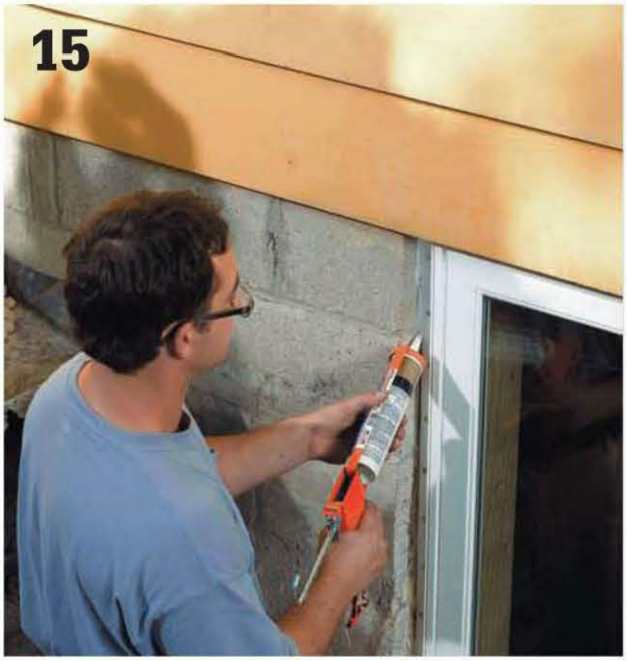

Seal gaps between the rough frame and the foundation with a bead of exterior silicone or polyurethane caulk. If the gaps are wider than ¼", insert a piece of backer rod first, then cover it with caulk. On the interior, fill gaps around the window shims with strips of foam backer rod, fiberglass insulation, or a bead of minimally expanding spray foam. Do not distort the window frame.

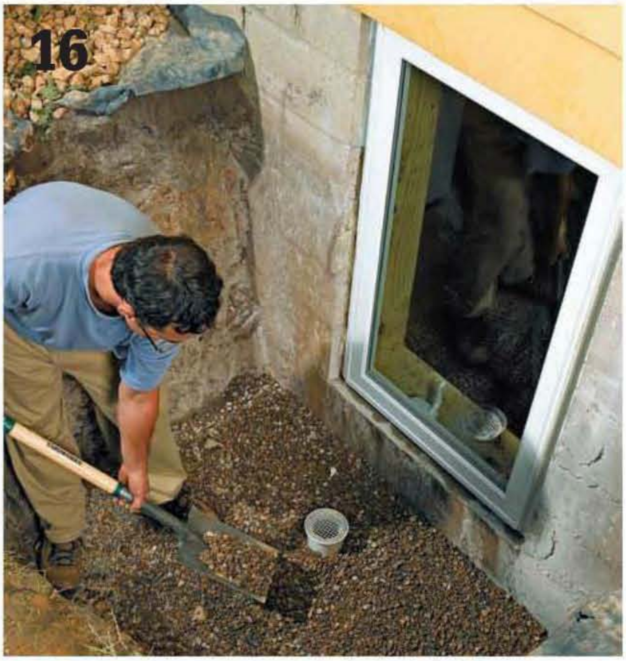

Fill the well excavation with 6 to 12" of pea gravel. This will serve as the window's drain system. Follow the egress well kit instructions to determine the exact depth required; you may need to add more gravel so the top of the well will be above the new window. *Note: We added a drain down to the foundation's perimeter tile for improved drainage as well.*

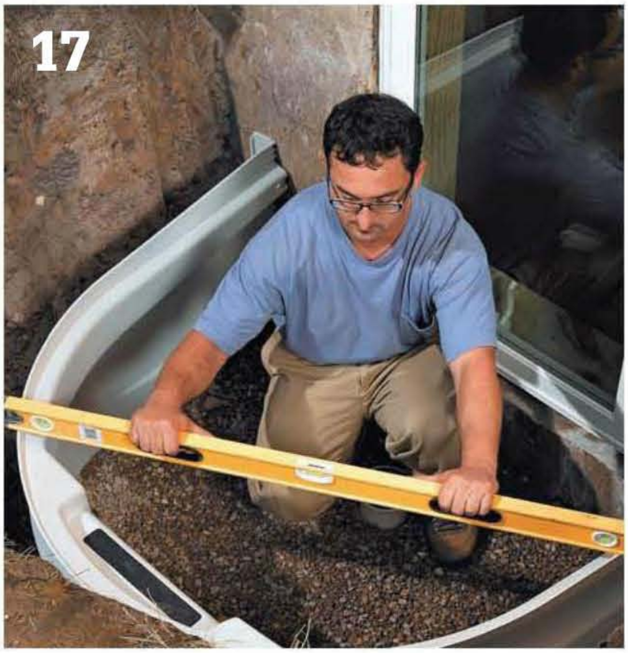

Set the bottom section of the well into the hole, and position it evenly from left to right relative to the window. Adjust the gravel surface to level the well section carefully.

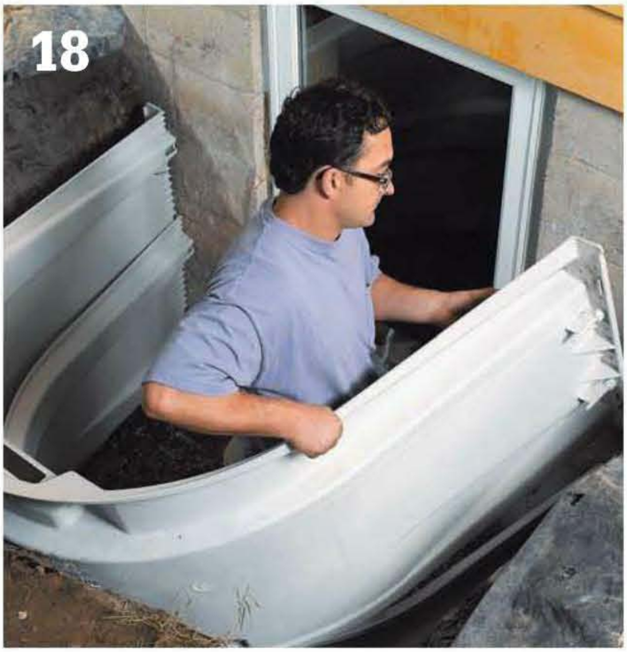

Stack the second well section on top of the first, and connect the two with the appropriate fasteners.

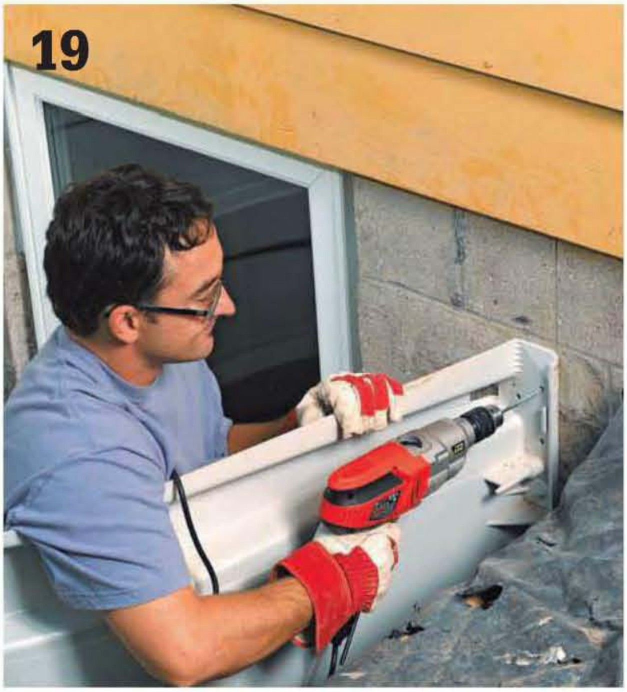

Fasten the window well sections to the foundation wall with concrete sleeve anchors driven into pre-bored pilot holes. You could also use masonry nails driven with a powder actuated nailer.

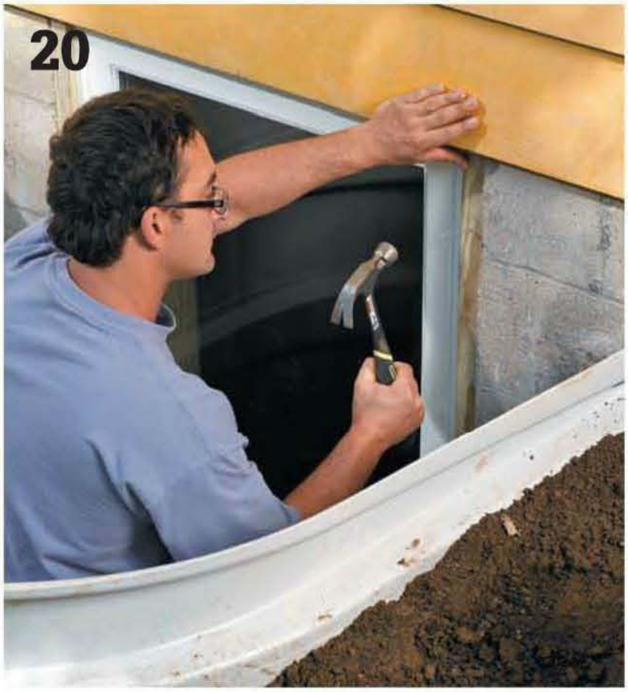

When all the well sections are assembled and secured, nail pieces of treated lumber trim around the window frame to hide the nailing flange. Complete the well installation by using excavated dirt to backfill around the outside of the well. Pack the soil with a tamper, creating a slope for good drainage. If you are installing a window well cover, set it in place and fasten it according to the manufacturer's instructions. (The cover must be removable.)

Installing Baseboard Heaters

Baseboard heaters are a popular way to provide additional heating for an existing room or primary heat to a converted attic or basement.

Heaters are generally wired on a dedicated 240-volt circuit controlled by a thermostat. Several heaters can be wired in parallel and controlled by a single thermostat (see circuit map 11, page 143).

Baseboard heaters are generally surface-mounted without boxes, so in a remodeling situation, you only need to run cables before installing wallboard. Be sure to mark cable locations on the floor before installing drywall. Retrofit installations are also not difficult. You can remove existing baseboard and run new cable in the space behind.

Tools & Materials ▸

Drill/driver
Wire stripper
Cable ripper
Wallboard saw
Baseboard heater or heaters
240-thermostat (in-heater or in-wall)
12/2 NM cable
Electrical tape
Basic wiring supplies

Baseboard heaters can provide primary or supplemental heat for existing rooms or additions. Install heaters with clear space between the heater and the floor.

Baseboard Thermostats

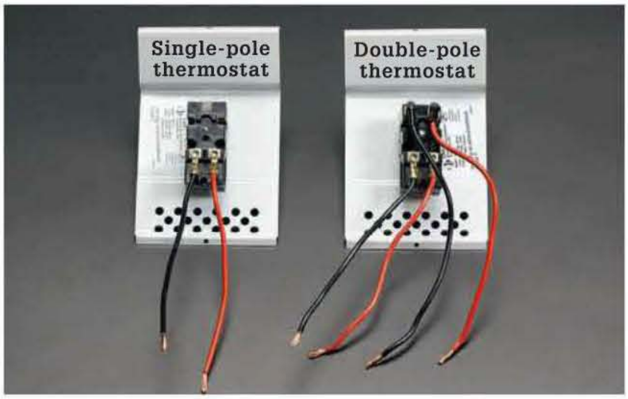

Single-pole and double-pole thermostats work in a similar manner, but double-pole models are safer. The single-pole model will open the circuit (causing shutoff) in only one leg of the power service. Double-pole models have two sets of wires to open both legs, lessening the chance that a person servicing the heater will contact a live wire.

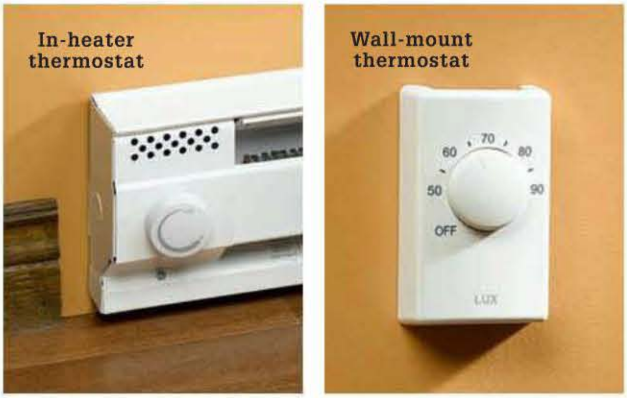

In-heater and wall-mount are the two types of baseboard thermostats you can choose from. If you are installing multiple heaters, a single wall-mount thermostat is more convenient. Individual in-heater thermostats give you more zone control, which can result in energy savings.

How Much Heater Do You Need?

If you don't mind doing a little math, determining how many lineal feet of baseboard heater a room requires is not hard.

1. Measure the area of the room in square feet (length × width): ________
2. Divide the area by 10 to get the baseline minimum wattage: ________
3. Add 5% for each newer window or 10% for each older window: ________
4. Add 10% for each exterior wall in the room: ________
5. Add 10% for each exterior door: ________
6. Add 10% if the space below is not insulated: ________
7. Add 20% if the space above is not well insulated: ________
8. Add 10% if ceiling is more than 8 ft. high: ________
9. Total of the baseline wattage plus all additions: ________
10. Divide this number by 250 (the wattage produced per foot of standard baseboard heater): ________
11. Round up to a whole number. This is the minimum number of feet of heater you need. ________

Note: It is much better to have more feet of heater than is required than fewer. Having more footage of heater does not consume more energy; it does allow the heaters to work more efficiently.

Planning Tips for Baseboard Heaters

- 240-volt heaters are much more energy efficient than 120-volt heaters.
- Baseboard heaters require a dedicated circuit. A 20-amp, 240-volt circuit of 12-gauge copper wire will power up to 16 ft. of heater.
- Do not install a heater beneath a wall receptacle. Cords hanging down from the receptacle are a fire hazard.
- Do not mount heaters directly on the floor. You should maintain at least 1" of clear space between the baseboard heater and the floor covering.
- Installing heaters directly beneath windows is a good practice.
- Locate wall thermostats on interior walls only, and do not install directly above a heat source.

How to Install a 240-volt Baseboard Heater

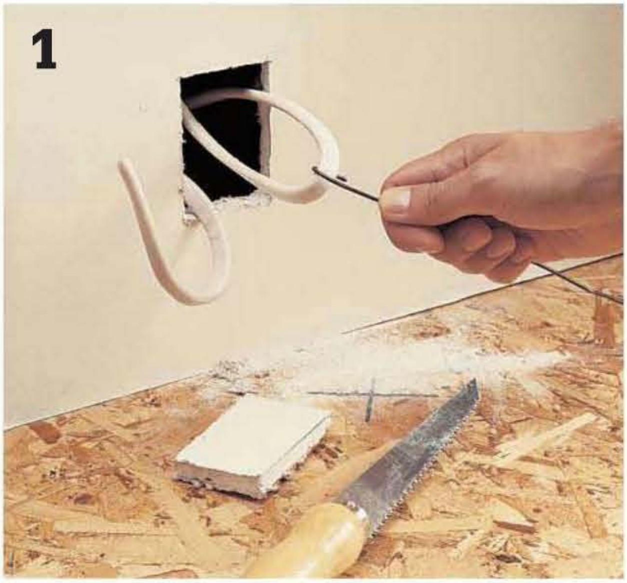

At the heater locations, cut a small hole in the drywall 3 to 4" above the floor. Pull 12/2 NM cables through the first hole: one from the thermostat, the other to the next heater. Pull all the cables for subsequent heaters. Middle-of-run heaters will have two cables, while end-of-run heaters have only one cable.

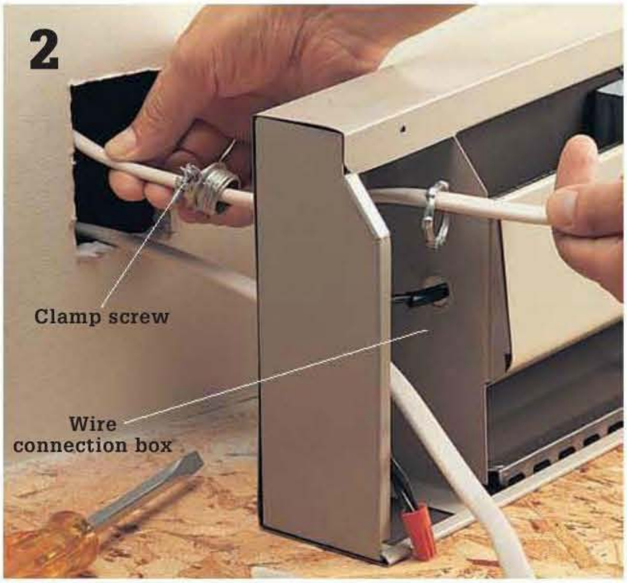

Remove the cover on the wire connection box. Open a knockout for each cable that will enter the box, then feed the cables through the cable clamps and into the wire connection box. Attach the clamps to the wire connection box, and tighten the clamp screws until the cables are gripped firmly.

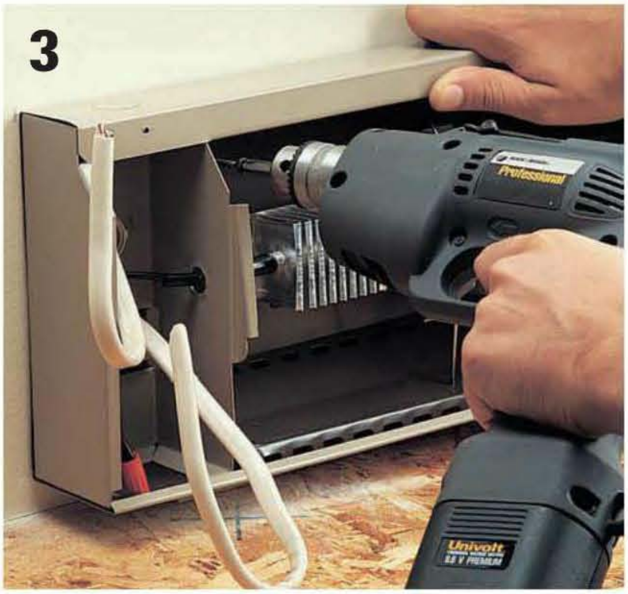

Anchor heater against wall about 1" off floor by driving flathead screws through back of housing and into studs. Strip away cable sheathing so at least ½" of sheathing extends into the heater. Strip ¾" of insulation from each wire using a combination tool.

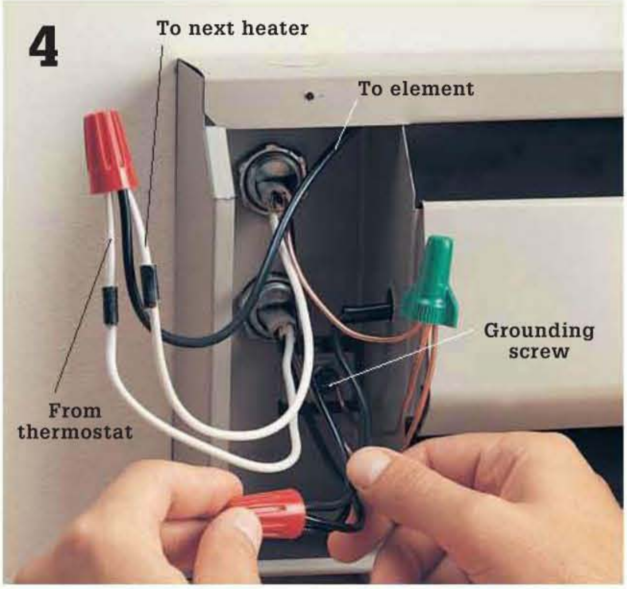

Make connections to the heating element if the power wires are coming from a thermostat or another heater controlled by a thermostat. See next page for other wiring schemes. Connect the white circuit wires to one of the wire leads on the heater. Tag white wires with black tape to indicate they are hot. Connect the black circuit wires to the other wire lead. Connect a grounding pigtail to the green grounding screw in the box, then join all grounding wires with a wire connector. Reattach cover.

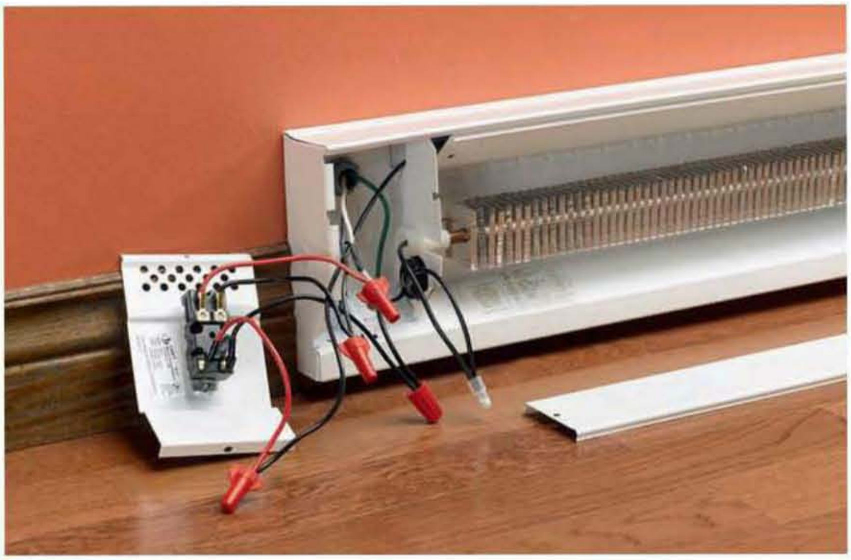

One heater with end-cap thermostat. Run both power leads (black plus tagged neutral) into the connection box at either end of the heater. If installing a single-pole thermostat, connect one power lead to one thermostat wire and connect the other thermostat wire to one of the heater leads. Connect the other hot LINE wire to the other heater lead. If you are installing a double-pole thermostat, make connections with both legs of the power supply.

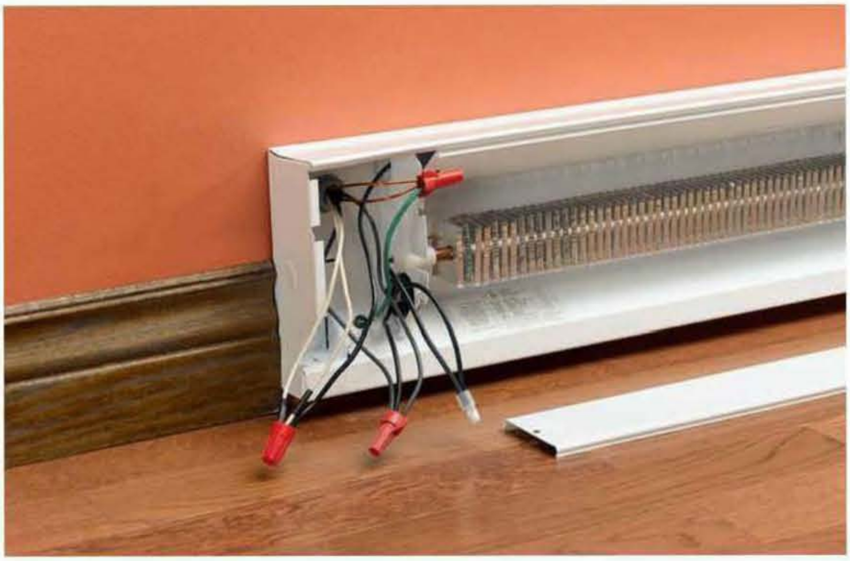

Multiple heaters. At the first heater, join both hot wires from the thermostat to the wires leading to the second heater in line. Be sure to tag all white neutrals hot. Twist copper ground wires together and pigtail them to the grounding screw in the baseboard heater junction box. This parallel wiring configuration ensures that power flow will not be interrupted to the downstream heaters if an upstream heater fails.

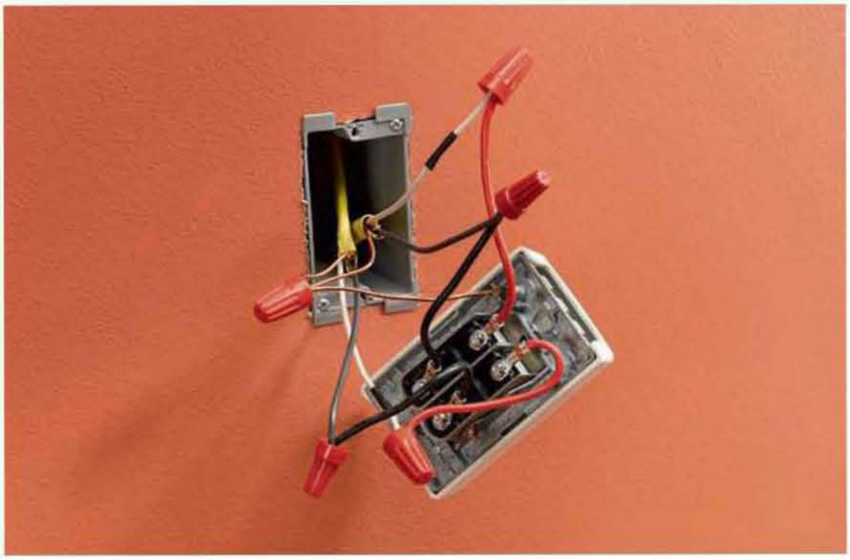

Wall-mounted thermostat. If installing a wall-mounted thermostat, the power leads should enter the thermostat first and then be wired to the individual heaters singly or in series. Hookups at the heater are made as shown in step 4. Be sure to tag the white neutral as hot in the thermostat box as well as in the heater box.

Exterior Improvements

Roofing, siding, and exterior trim details create the outer envelope of your home, safeguarding it from the elements. Aside from their protective purposes, these three systems should also blend together aesthetically to improve your home's appearance. Roofing and siding replacement, in particular, are among the more costly projects you'll face as a homeowner, and generally they're investments you'll make only every few decades or so. For these reasons, it's wise to plan your projects carefully, so you can keep all the relevant factors in mind before you begin.

This chapter will help you get started on the right foot. You'll begin the planning process by evaluating your needs and estimating both price and material quantities required for the job. Then, you'll take a more in-depth look at the wide variety of product options available. New forms and styles of roofing and siding continue to come to market, and there may be some you simply aren't aware of yet. A roofing, siding, or trim project will usually require at least some work at heights. You'll learn important tips for working safely, how to set up scaffolds, and how to prepare your job site to minimize damage, manage debris, and work efficiently.

In this chapter:

- Estimating Roofing, Siding & Trim
- Preparing the Roof for Shingles
- Shingling a Roof
- Installing Vinyl Siding

Estimating Roofing, Siding & Trim

Ordering Roofing Materials

Roofing materials are ordered in squares, with one square equaling 100 square feet. To determine how many squares are needed, first figure out the square footage of your roof. The easiest way to make this calculation is to multiply the length by the width of each section of roof, and then add the numbers together.

For steep roofs and those with complex designs, do your measuring from the ground and multiply by a number based on the slope of your roof. Measure the length and width of your house, include the overhangs, then multiply the numbers together to determine the overall square footage. Using the chart at the lower right, multiply the square footage by a number based on the roof's slope. Add 10 percent for waste, then divide the total square footage by 100 to determine the number of squares you need. Don't spend time calculating and subtracting the areas that won't be covered, such as skylights and chimneys. They're usually small enough that they don't impact the number of squares you need. Besides, it's good to have extra materials for waste, mistakes, and later repairs.

To determine how much flashing you'll need, measure the length of the valley to figure valley flashing, the lengths of the eaves and rakes to figure drip edge, and the number and size of vent pipes to figure vent flashing.

TIMBERLINE
TIMBERLINE

Asphalt shingles come in packaged bundles weighing around 65 pounds each. For typical three-tab shingles, three bundles will cover one square (100 sq. ft.) of roof.

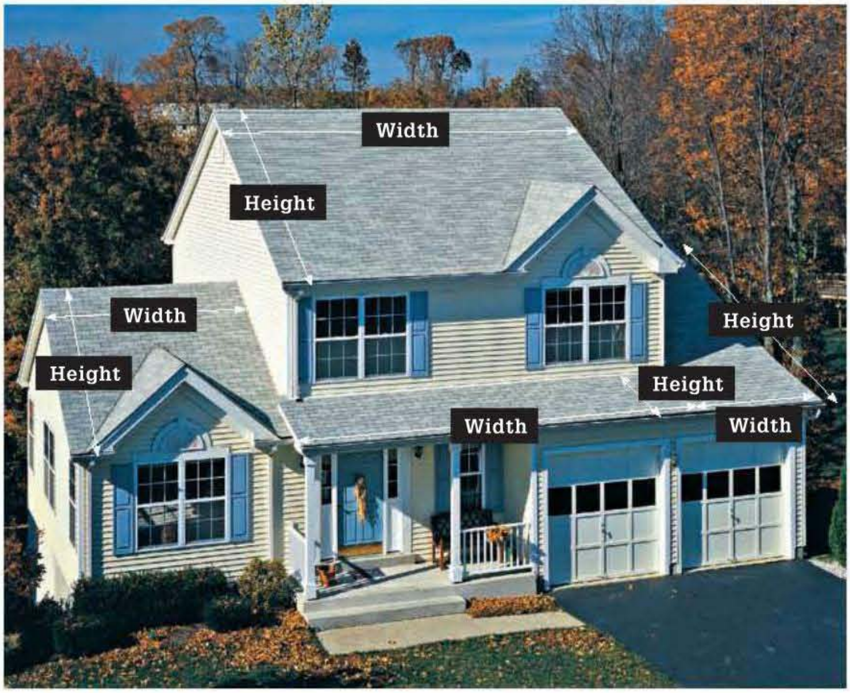

Calculate the roof's surface by multiplying the height of the roof by the width. Do this for each section, then add the totals together. Divide that number by 100, add 10 percent for waste, and that's the number of squares of roofing materials you need.

Conversion Chart

Slope	Multiply by	Slope	Multiply by
2 in 12	1.02	8 in 12	1.20
3 in 12	1.03	9 in 12	1.25
4 in 12	1.06	10 in 12	1.30
5 in 12	1.08	11 in 12	1.36
6 in 12	1.12	12 in 12	1.41
7 in 12	1.16		

Ordering Siding Materials

Siding is sold by the linear foot, square foot, board foot, or square, depending on the type of material. To determine the amount of siding materials you'll need, calculate the square footage of each wall, then add them together to get your total surface area.

To figure the square footage for a wall, multiply the length of the wall by the height. Calculate the area on the gable end of a wall by using the formula for triangles, which is one-half the length of the base multiplied by the height of the triangle. You won't need to subtract for areas covered by windows and doors, and you won't need to add an extra 10 percent for waste. The window and door areas are roughly equal to the amount of waste you'll need.

Use your height and length measurements from the walls to determine how many feet of starter strip, channels, corner posts, and trim you'll need. Keep in mind you'll probably also need to apply trim or channels around all of your doors and windows.

Depending on the siding material you're ordering, you may need extra material to allow for overlap. For example, if you want to install 10" lap siding with an 8" exposure, you'll need to account for a 2" overlap for each board. Likewise, the exposure rate for wood shakes or shingles will determine how many squares you'll need. Depending on the exposure rate, one square of shingles could cover 80 sq. ft, 100 sq. ft, or even 120 sq. ft. Most manufacturers have charts that show how much material is needed to cover a specified number of square feet at various rates of exposure. If you have trouble estimating how much material you need to purchase, ask your supplier for help.

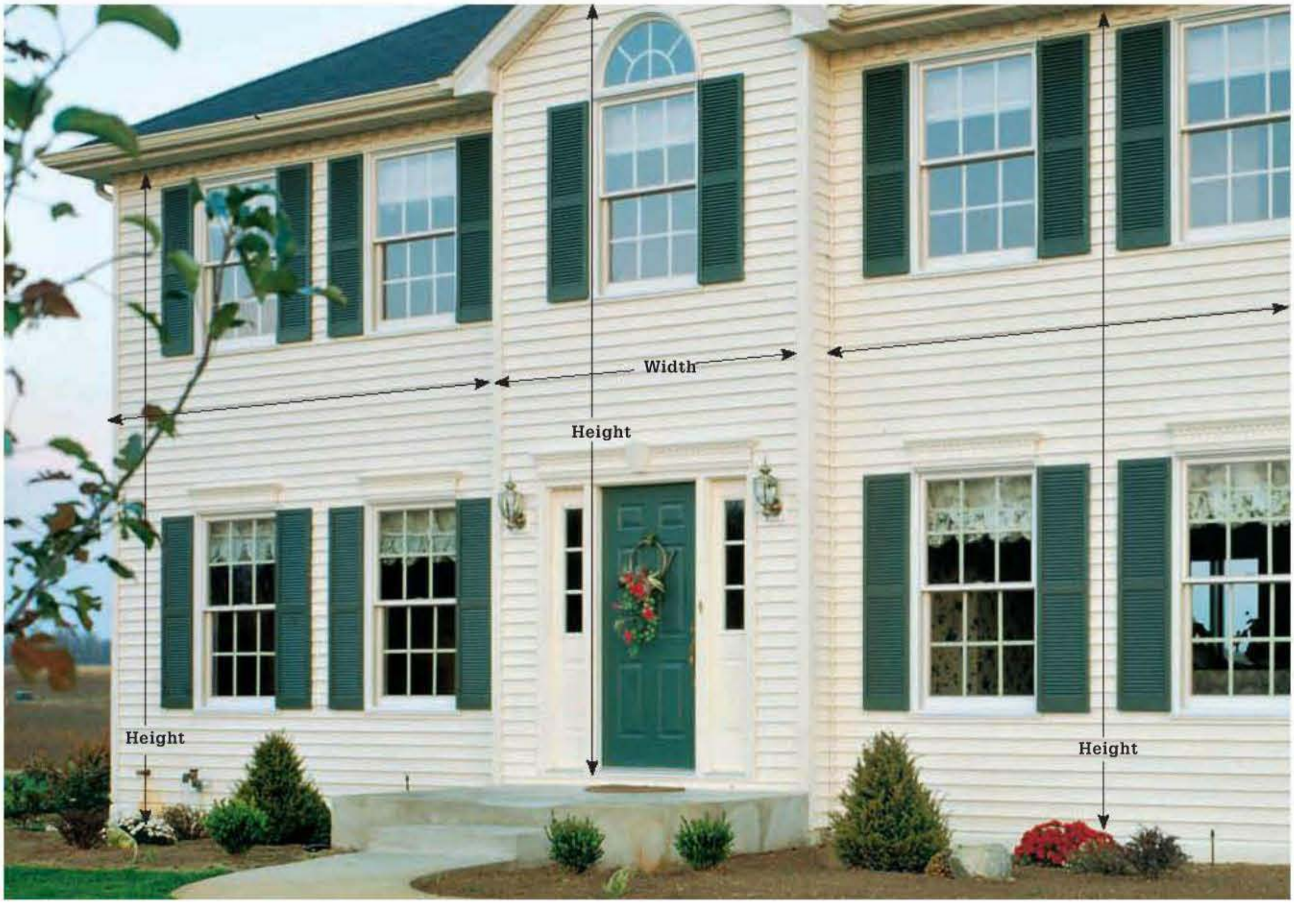

Estimate the amount of siding you'll need by calculating the square footage of each wall, then adding the numbers together. To determine the square footage of a wall, multiply the wall length by its height. Subtract the square footage of all windows and doors, and then add 10 percent for waste to get square footage. Don't forget about corner trim, J-channel and other trim pieces.

Preparing the Roof for Shingles

Working conditions on a roof can be arduous, so make the job as easy as possible by gathering the right tools and equipment before you begin.

Some of these tools, such as a pneumatic nailer and roofer's hatchet, are specific to roofing projects. If you don't have them and don't want to buy them, you can rent them from a rental center.

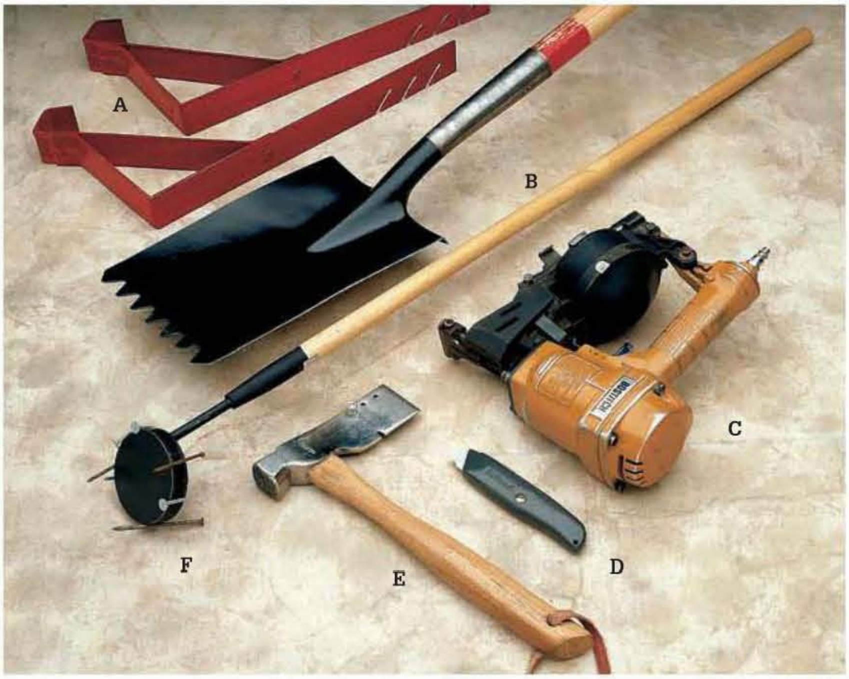

Specialty roofing tools include roof jacks (A) for use on steep roofs, roofing shovel (B) for tearoff work, pneumatic nailer (C), utility knife with hooked blade (D) for trimming shingles, roofing hammer with alignment guides and hatchet blade (E) for shingle installation, and a release magnet for site cleanup (F).

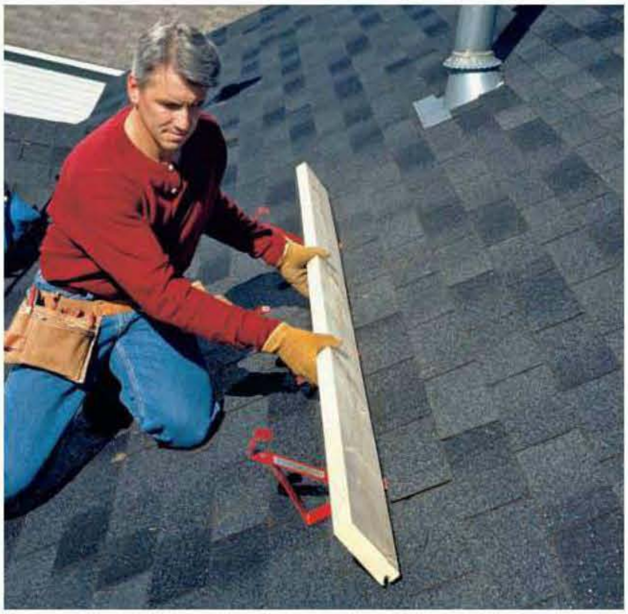

Use roof jacks on steep roofs. Nail the supports at the fourth or fifth course of shingles, and add the widest board the supports will hold.

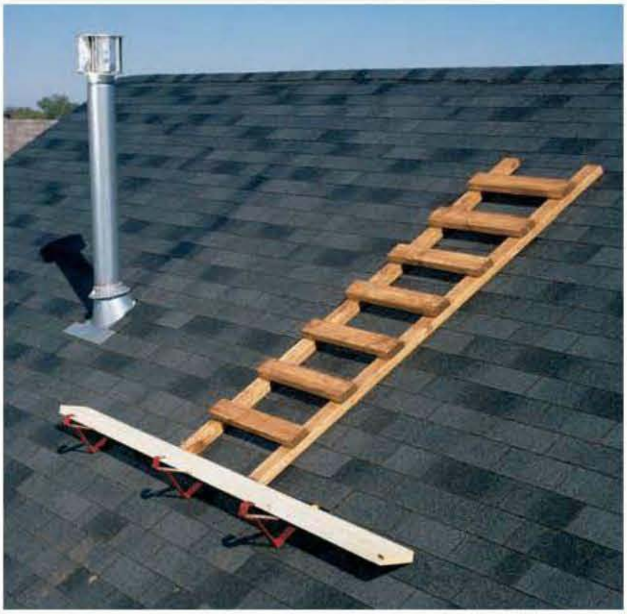

For more secure footing, fashion a roofing ladder by nailing wood strips across a pair of 2 × 4s. Secure the ladder to the roof jacks, and use it to maintain your footing.

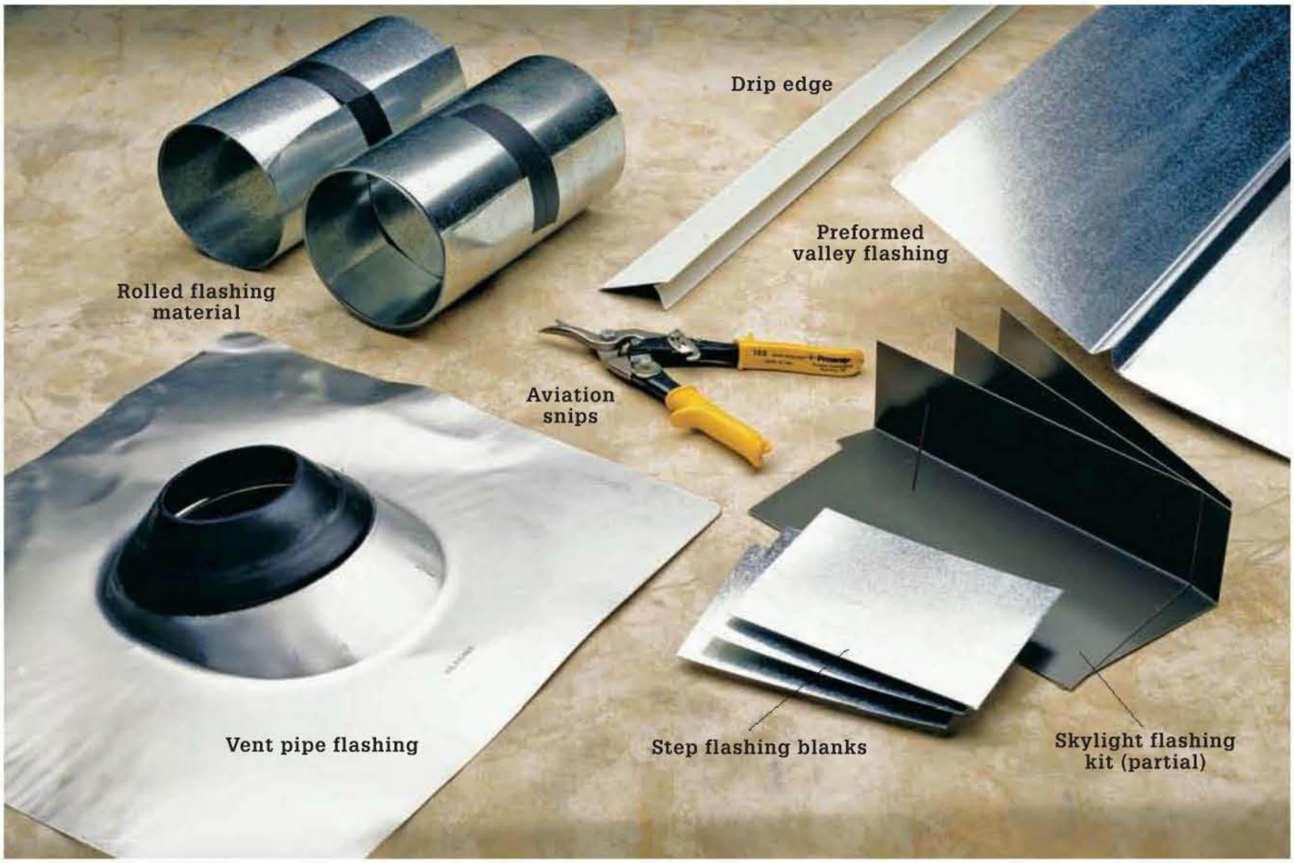
Drip edge
Preformed valley flashing
Rolled flashing material
Aviation snips
Vent pipe flashing
Step flashing blanks
Skylight flashing kit (partial)

Roof flashing can be hand cut or purchased in preformed shapes and sizes. Long pieces of valley flashing, base flashing, top saddles, and other nonstandard pieces can be cut from rolled flashing material using aviation snips. Step flashing blanks can be bought in standard sizes and bent to fit. Drip edge and vent pipe flashing are available preformed. Skylight flashing usually comes as a kit with the window. Complicated flashings, such as chimney crickets, can be custom fabricated by a metalworker.

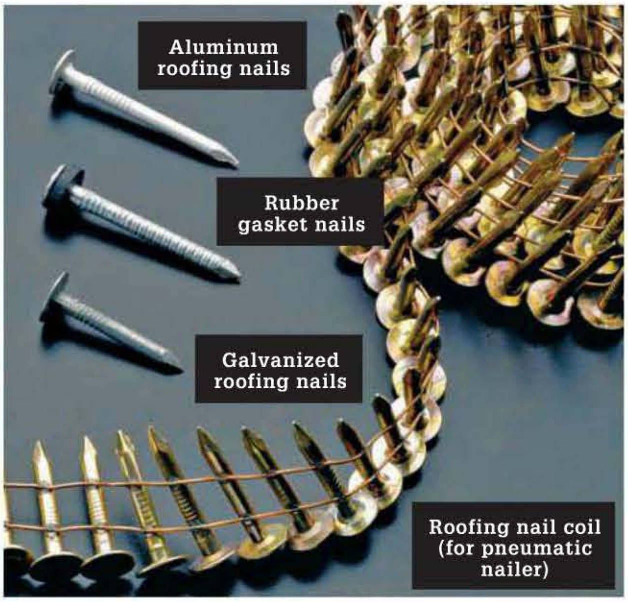
Aluminum roofing nails
Rubber gasket nails
Galvanized roofing nails
Roofing nail coil (for pneumatic nailer)

Different fasteners are specially developed for different jobs. Use galvanized roofing nails to hand nail shingles; use aluminum nails for aluminum flashing; use rubber gasket nails for galvanized metal flashing; and use nail coils for pneumatic nailers.

Felt paper (30#)
Roofing cement
Ice-guard membrane

Common roofing materials include 30# felt paper for use as underlayment; ice-guard membrane for use as underlayment in cold climates; and tubes of roofing cement for sealing small holes, cracks, and joints.

How to Tear Off Old Shingles

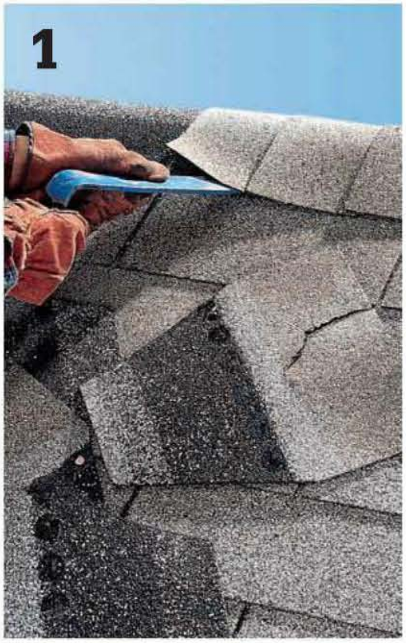

Remove the ridge cap using a flat pry bar. Pry up the cap shingles at the nail locations.

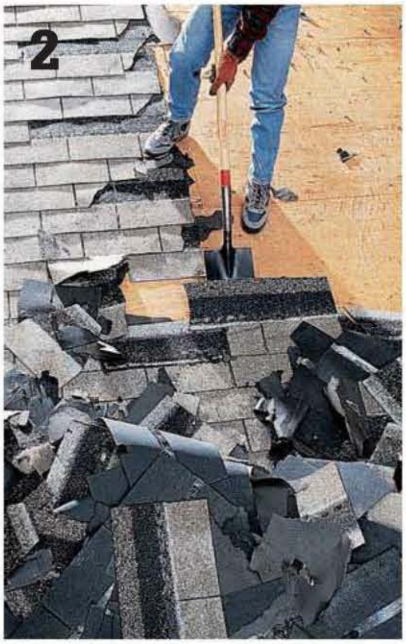

Working downward from the peak, tear off the felt paper and old shingles with a roofing shovel or pitchfork.

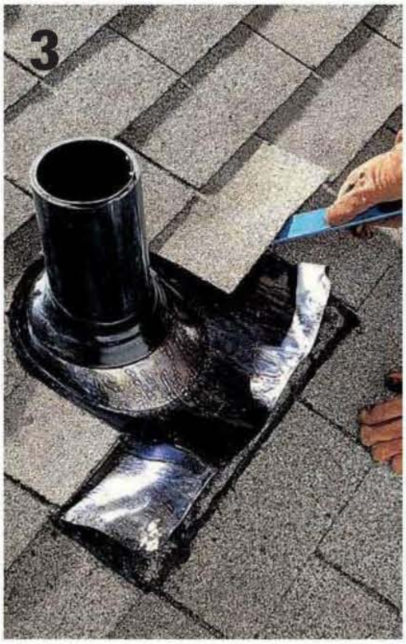

Unless flashing is in exceptional condition, remove it by slicing through the roofing cement that attaches it to the shingles. You may be able to salvage flashing pieces, such as chimney saddles and crickets, and reuse them.

After removing the shingles, felt paper, and flashing from the entire tear-off section, pry out any remaining nails and sweep the roof with a broom.

If an unexpected delay keeps you from finishing a section before nightfall, cover any unshingled sections, using tarps weighted down with shingle bundles.

How to Replace Damaged Sheathing

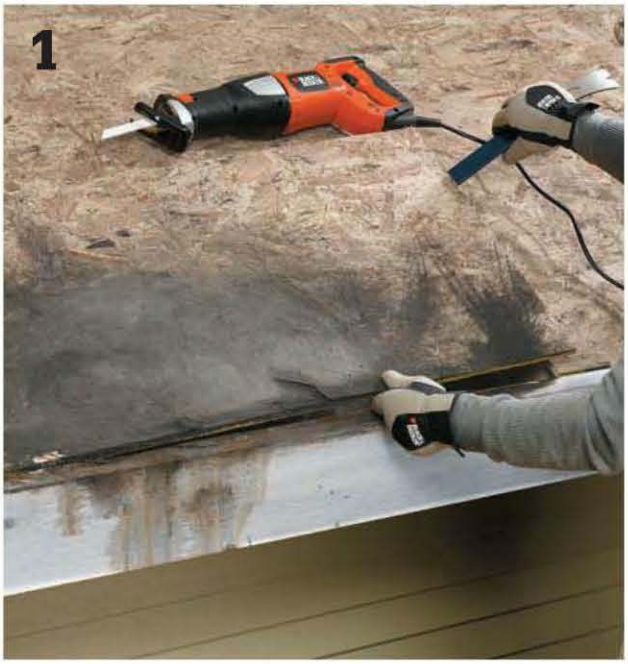

Use a reciprocating saw to cut next to the rafters in an area that extends well beyond the damaged area. Pry out the damaged sections using a pry bar.

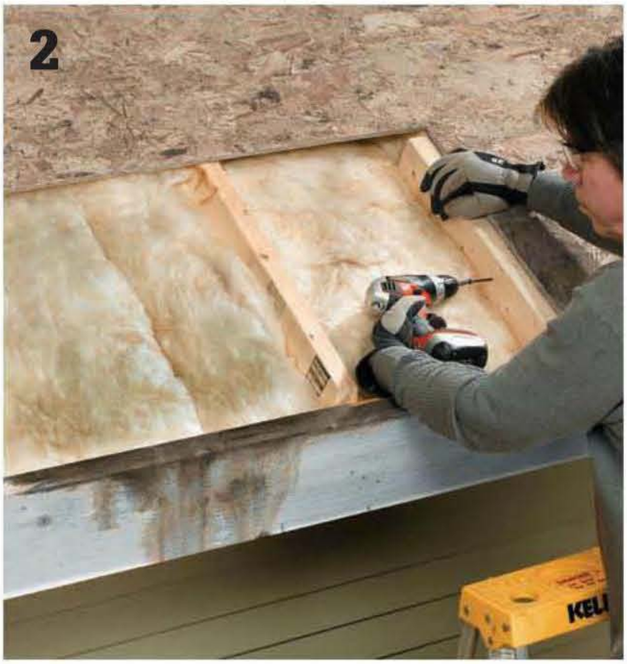

Attach 2 × 4 nailing strips to the inside edges of the rafters using 3" deck screws

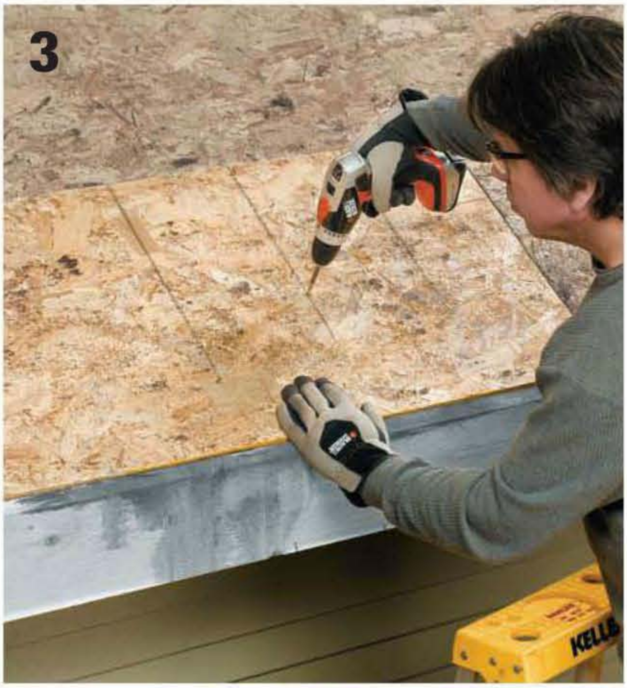

Use exterior grade plywood to make a patch. Measure the cutout area, allow for a ⅛" gap on all sides for expansion, and cut the patch to size. Attach the patch to the rafters and nailing strips using 2¼" deck screws or 8d ring-shank siding nails.

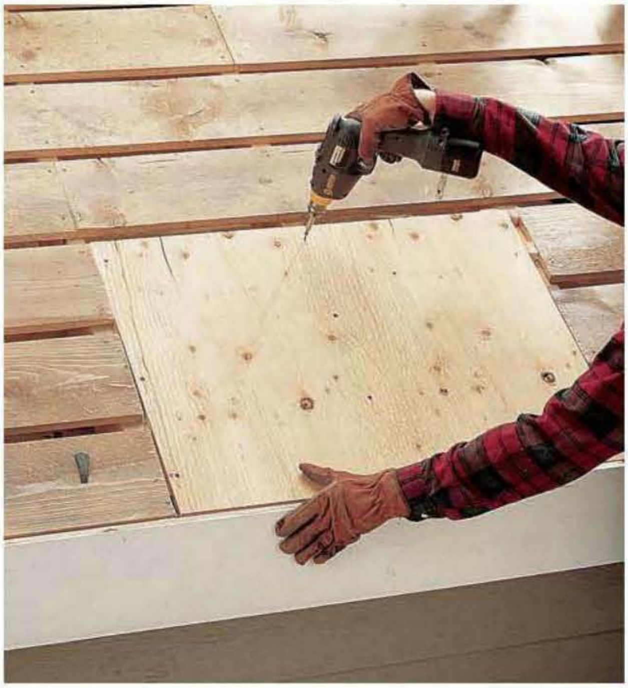

Option: If your existing roof deck is made of boards (1 × 6 was common before plywood took over the market), it is perfectly acceptable to use plywood when replacing a section of the deck. The plywood should be the same thickness as the boards, generally, ¾".

How to Install Underlayment

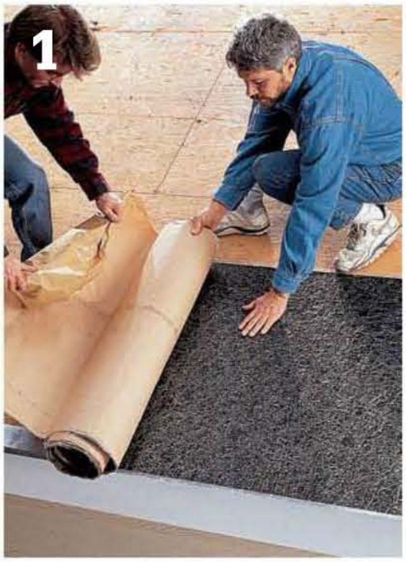

Snap a chalk line 35⅝" up from the eaves, so the first course of the 36"-wide membrane will overhang the eaves by ⅜". Install a course of ice and water shield, using the chalk line as a reference, and peeling back the protective backing as you unroll it.

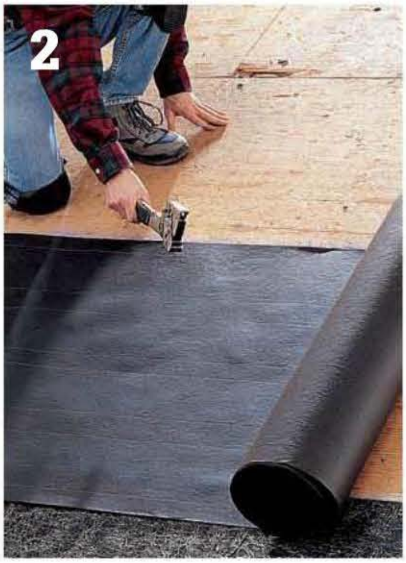

Measuring up from the eaves, make a mark 32" above the top of the last row of underlayment, and snap another chalk line. Roll out the next course of felt paper (or ice guard, if required) along the chalk line, overlapping the first course by 4". *Tip: Drive staples every 6 to 12" along the edges of felt paper, and one staple per sq. ft. in the field area.*

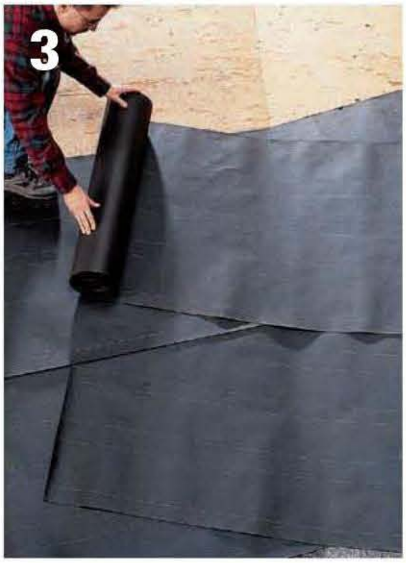

At valleys, roll felt paper across from both sides, overlapping the ends by 36". Install felt paper up to the ridge—ruled side up—snapping horizontal lines every two or three rows to check alignment. Overlap horizontal seams by 4", vertical seams by 12", and hips and ridges by 6". Trim the courses flush with the rake edges.

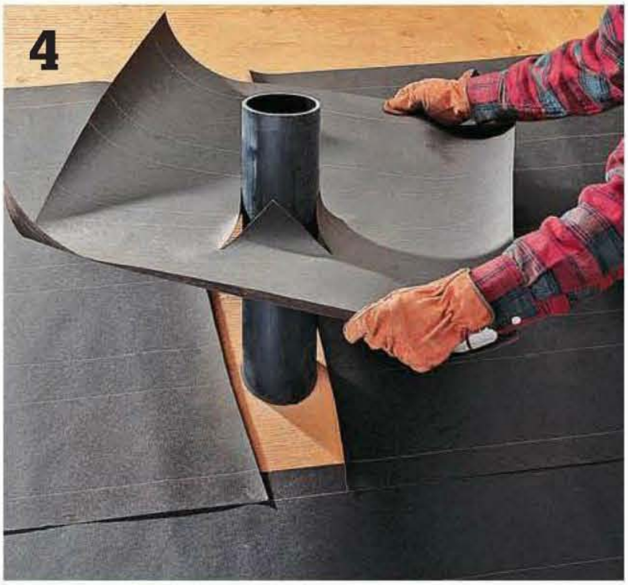

Apply felt paper up to an obstruction, then resume laying the course on the opposite side (make sure to maintain the line). Cut a patch that overlaps the felt paper by 12" on all sides. Make a crosshatch cutout for the obstruction. Position the patch over the obstruction, staple it in place, then caulk the seams with roofing cement.

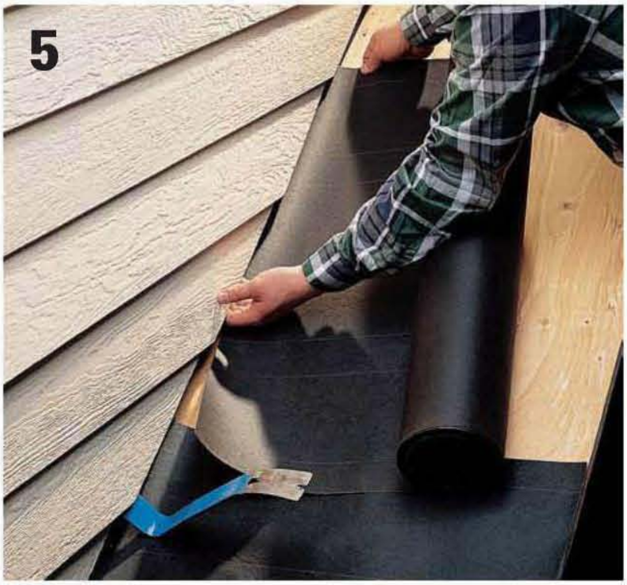

At the bottom of dormers and sidewalls, tuck the felt paper under the siding where it intersects with the roof. Carefully pry up the siding and tuck at least 2" of paper under it. Also tuck the paper under counter flashing or siding on chimneys and skylights. Leave the siding or counter flashing or siding unfastened until after you install the step flashing.

How to Install Drip Edge

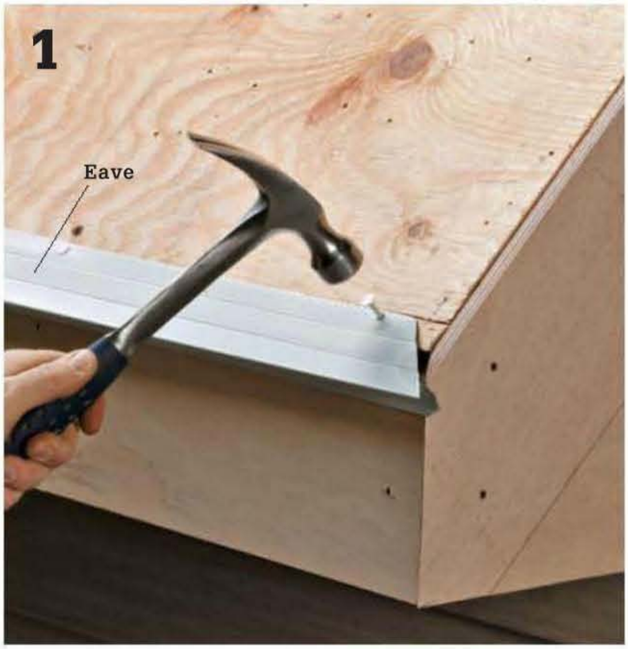

Cut a 45° miter at one end of the drip edge using aviation snips. Place the drip edge along the eaves end of the roof, aligning the mitered end with the rake edge. Nail the drip edge in place every 12".

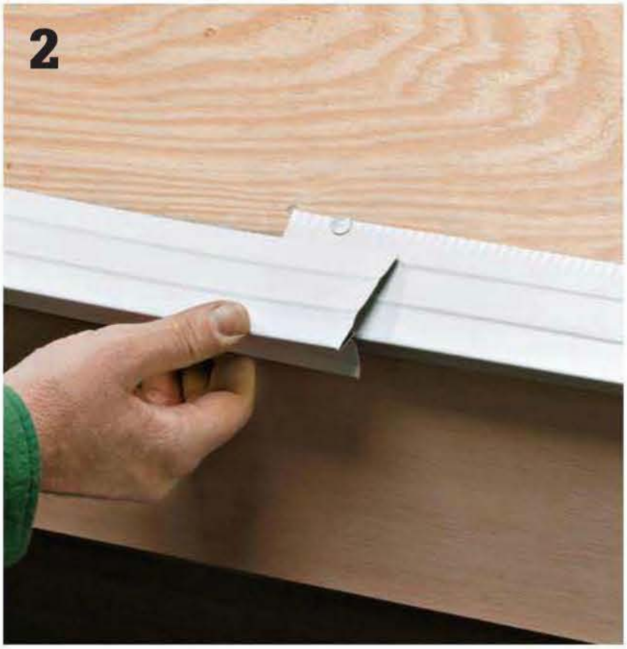

Overlap pieces of drip edge by 2". Install drip edge across the entire eaves, ending with a mitered cut on the opposite corner.

Apply felt paper, and ice guard if needed, to the roof, overhanging the eaves by ⅜".

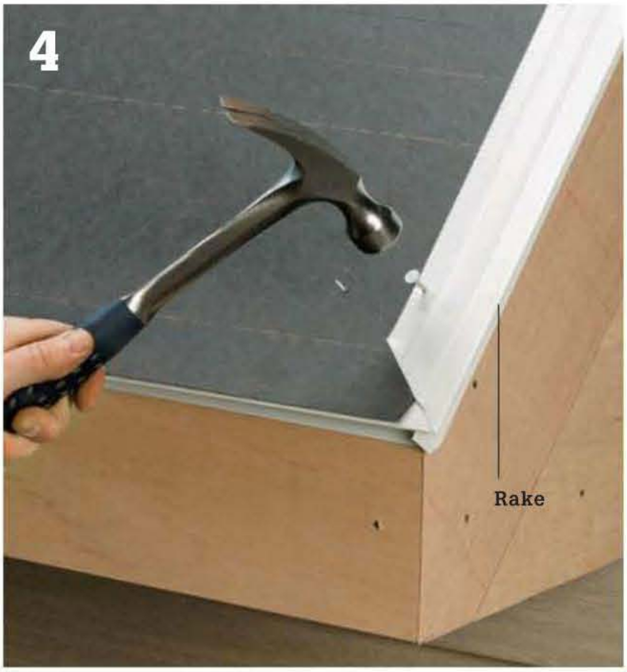

Cut a 45° miter in a piece of drip edge and install it along the rake edge, forming a miter joint with the drip edge along the eaves. Overlap pieces by 2", making sure the higher piece is on top at the overlap. Apply drip edge all the way to the peak. Install drip edge along the other rake edges the same way.

Shingling a Roof

If you want to install asphalt shingles on your roof, then you're in good company. Asphalt shingles, also known as composition shingles, are the roofing of choice for nearly four out of five homeowners in America. They perform well in all types of climate, are available in a multitude of colors, shapes, and textures to complement every housing design, and are less expensive than most other roofing products.

Asphalt shingles are available as either fiberglass shingles or organic shingles. Both types are made with asphalt, the difference being that one uses a fiberglass reinforcing mat, while the other uses a cellulose-fiber mat. Fiberglass shingles are lighter, thinner, and have a better fire rating. Organic shingles have a higher tear strength, are more flexible in cold climates, and are used more often in northern regions.

Although the roofing market has exploded with innovative new asphalt shingle designs, such as the architectural or laminated shingle that offers a three-dimensional look, the standard three-tab asphalt shingle is still the most common, which is the project we're featuring here. The tabs provide an easy reference for aligning shingles for installation.

To help the job get done faster, rent an air compressor and pneumatic roofing gun. This will greatly reduce the time you spend nailing.

Tools & Materials ▸

- Aviation snips
- Carpenter's square
- Chalk line
- Flat bar
- Roofer's hatchet or pneumatic nailer
- Utility knife
- Straightedge
- Tape measure
- Chalk gun
- Flashing
- Shingles
- Nailing cartridges
- Roofing cement
- Roofing nails (⅞", 1¼")
- Rubber gasket nails

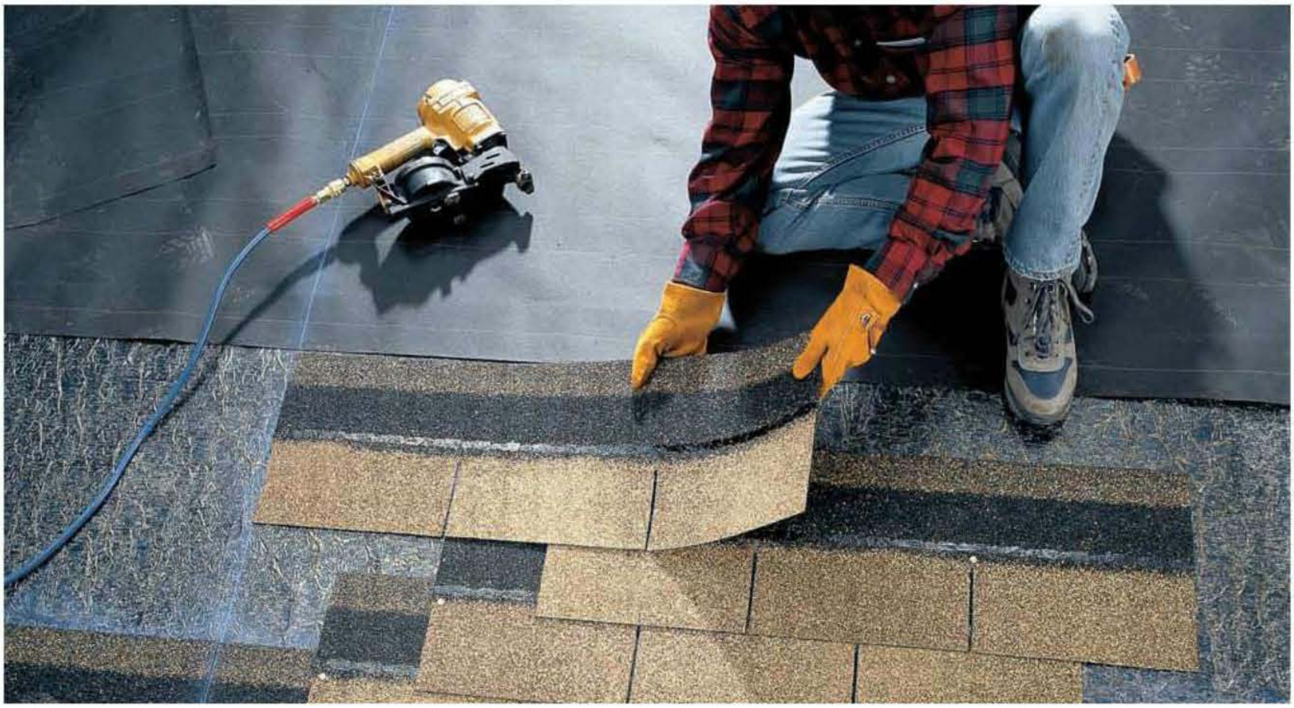

Stagger shingles for effective protection against leaks. If the tab slots are aligned in successive rows, water forms channels, increasing erosion of the mineral surface of the shingles. Creating a 6" offset between rows of shingles—with the three-tab shingles shown above—ensures that the tab slots do not align.

How to Install Three-tab Shingles

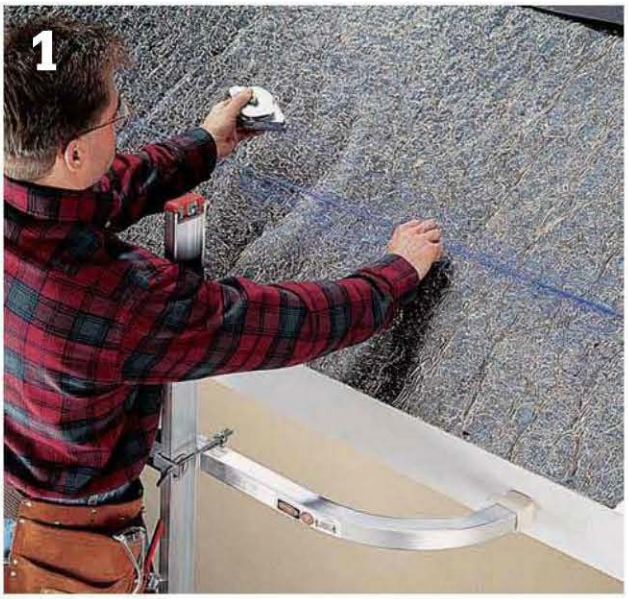

Cover the roof with felt paper (page 536) and install drip edge (page 537). Snap a chalk line onto the felt paper or ice guard 11½" up from the eaves edge, to mark the alignment of the starter course. This will result in a ½" shingle overhang for standard 12" shingles. *Tip: Use blue chalk rather than red. Red chalk will stain roofing materials.*

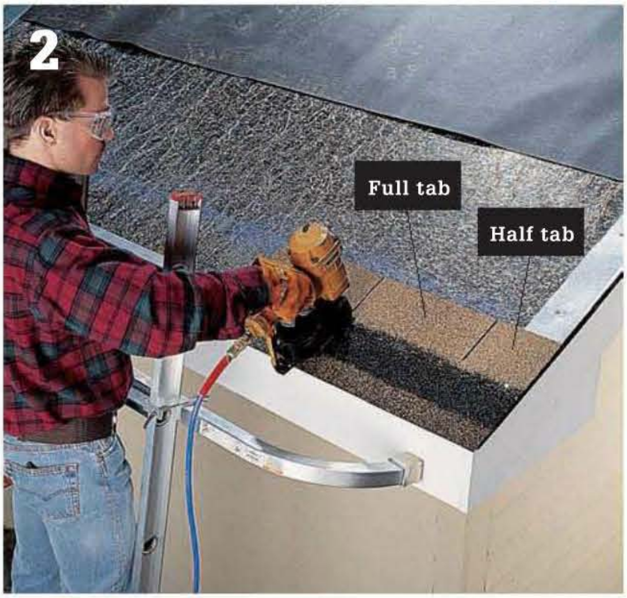

Trim off one-half (6") of an end tab on a shingle. Position the shingle upside down, so the tabs are aligned with the chalk line and the half-tab is flush against the rake edge. Drive ⅞" roofing nails near each end, 1" down from each slot between tabs. Butt a full upside-down shingle next to the trimmed shingle, and nail it. Fill out the row, trimming the last shingle flush with the opposite rake edge.

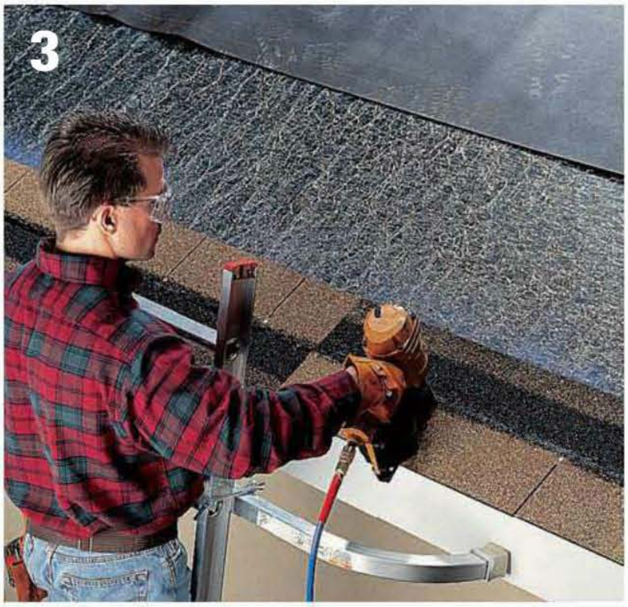

Apply the first full course of shingles over the starter course with the tabs pointing down. Begin at the rake edge where you began the starter row. Place the first shingle so it overhangs the rake edge by ⅜" and the eaves edge by ½". Make sure the top of each shingle is flush with the top of the starter course, following the chalk line.

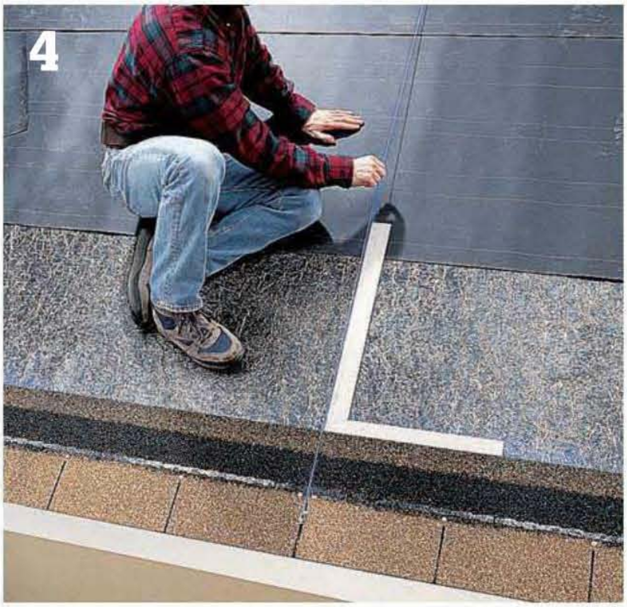

Snap a chalk line from the eaves edge to the ridge to create a vertical line to align the shingles. Choose an area with no obstructions, as close as possible to the center of the roof. The chalk line should pass through a slot or a shingle edge on the first full shingle course. Use a carpenter's square to establish a line perpendicular to the eaves edge.

(continued)

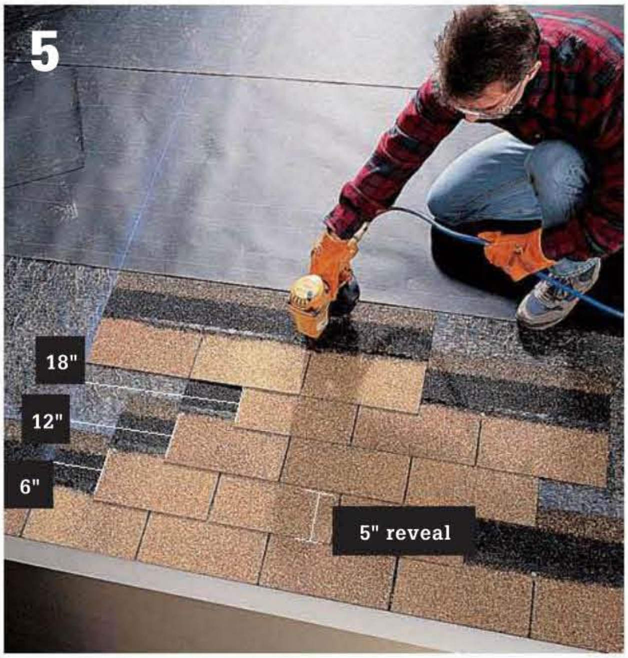

Use the vertical reference line to establish a shingle pattern with slots that are offset by 6" in succeeding courses. Tack down a shingle 6" to one side of the vertical line, 5" above the bottom edge of the first-course shingles to start the second row. Tack down shingles for the third and fourth courses, 12" and 18" from the vertical line. Butt the fifth course against the line.

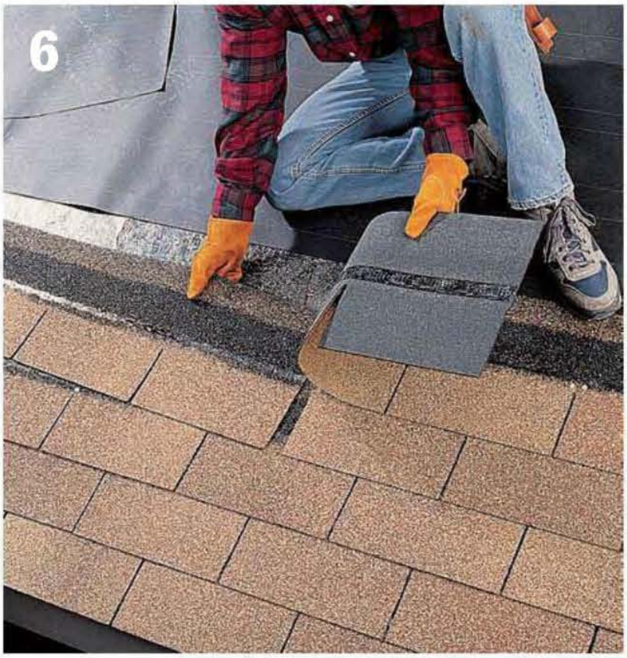

Fill in shingles in the second through fifth courses, working upward from the second course and maintaining a consistent 5" reveal. Slide lower-course shingles under any upper-course shingles left partially nailed, and then nail them down. *Tip: Install roof jacks, if needed, after filling out the fifth course.*

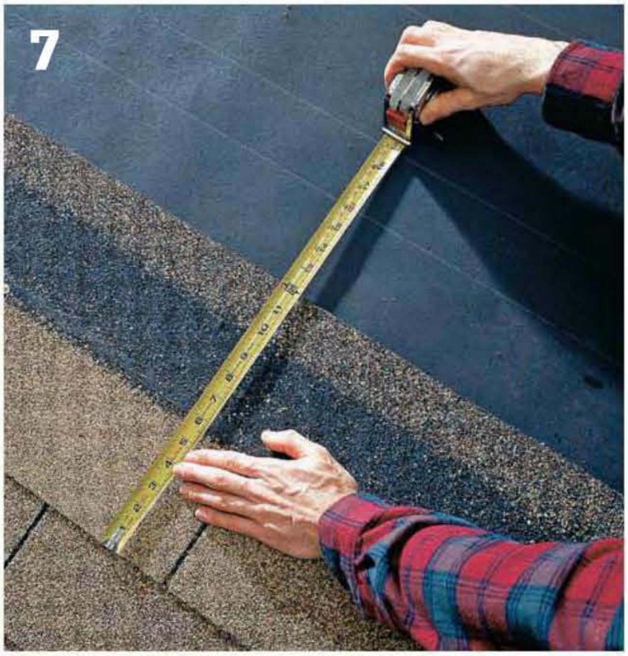

Check the alignment of the shingles after each four-course cycle. In several spots on the last installed course, measure from the bottom edge of a shingle to the nearest felt paper line. If you discover any misalignment, make minor adjustments over the next few rows until it's corrected.

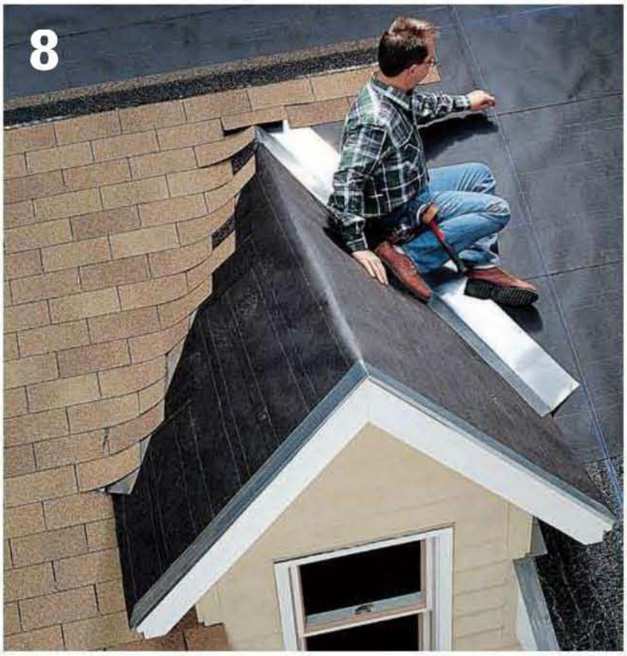

When you reach obstructions, such as dormers, install a full course of shingles above them so you can retain your shingle offset pattern. On the unshingled side of the obstruction, snap another vertical reference line using the shingles above the obstruction as a guide.

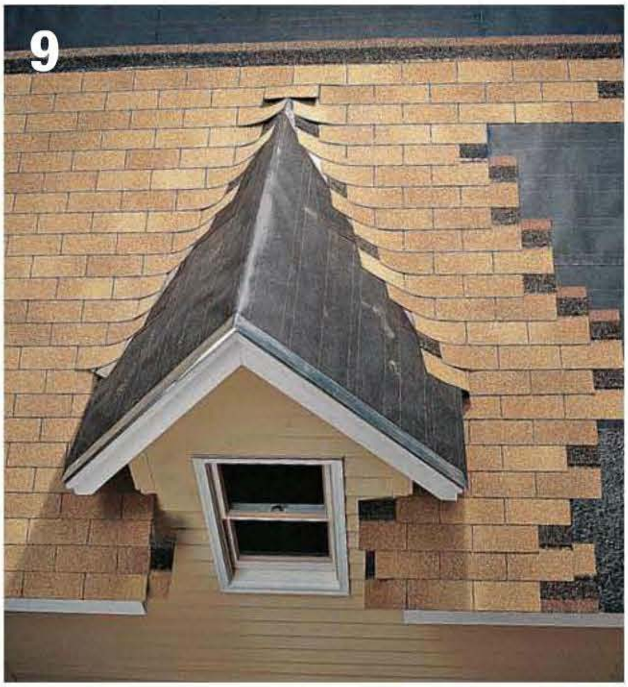

Shingle upward from the eaves on the unshingled side of the obstruction using the vertical line as a reference for re-establishing your shingle slot offset pattern. Fill out the shingle courses past the rake edges of the roof, then trim off the excess.

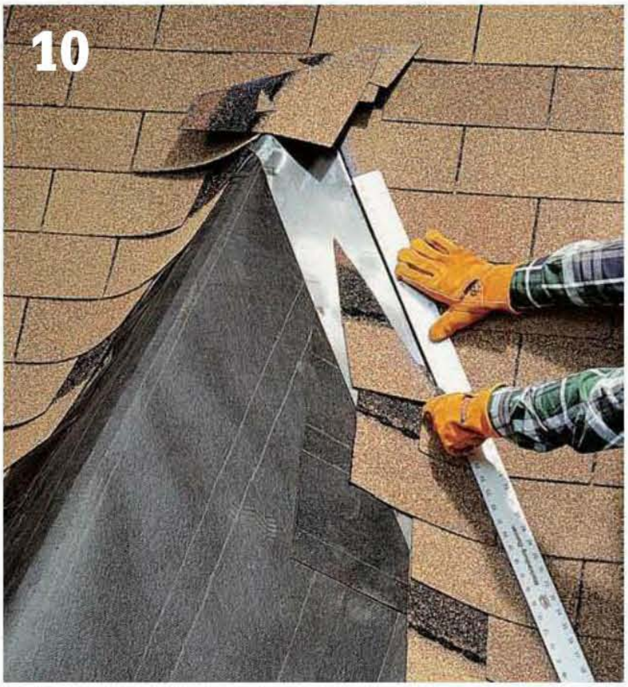

Trim off excess shingle material at the V in the valley flashing using a utility knife and straightedge. Do not cut into the flashing. The edges will be trimmed back farther at a slight taper after both roof decks are completely shingled.

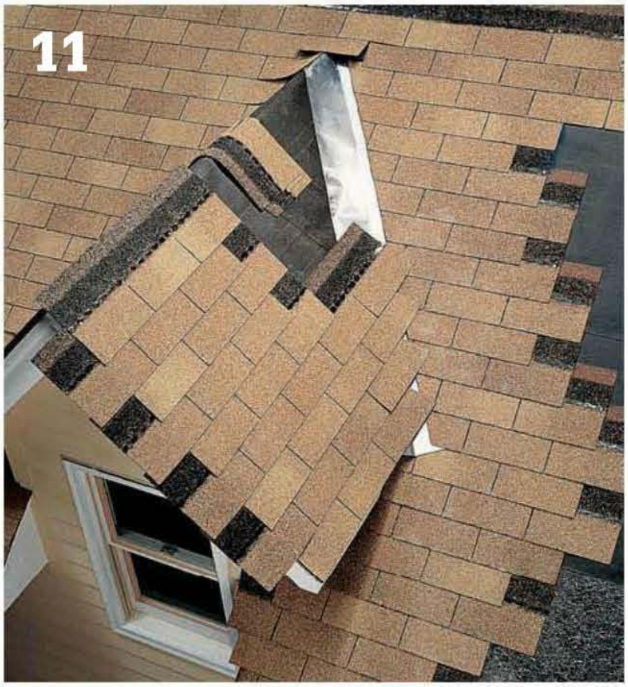

Install shingles on adjoining roof decks, starting at the bottom edge using the same offset alignment pattern shown in steps 1 to 6. Install shingles until courses overlap the center of the valley flashing. Trim shingles at both sides of the valley when finished.

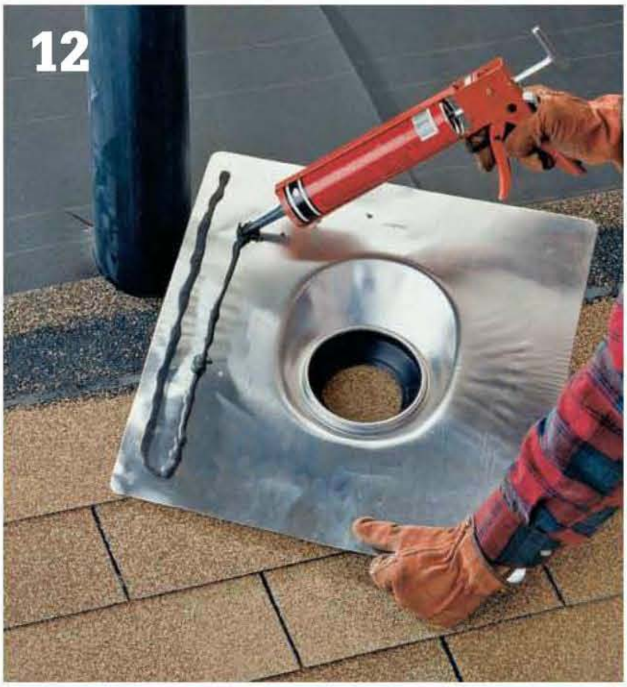

Install shingles up to the vent pipe so the flashing rests on at least one row of shingles. Apply a heavy double bead of roofing cement along the bottom edge of the flange.

(continued)

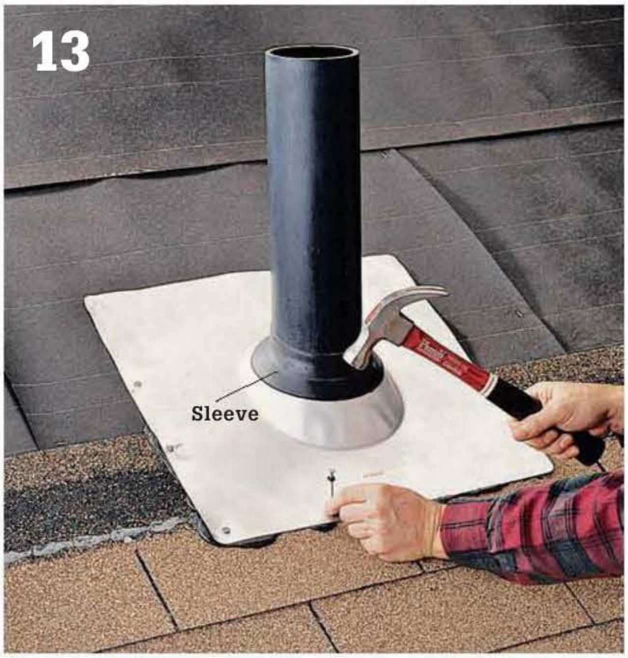

Place the flashing over the vent pipe. Position the flashing collar so the longer portion of the tapered neck slopes down the roof and the flange lies over the shingles. Nail the perimeter of the flange using rubber gasket nails.

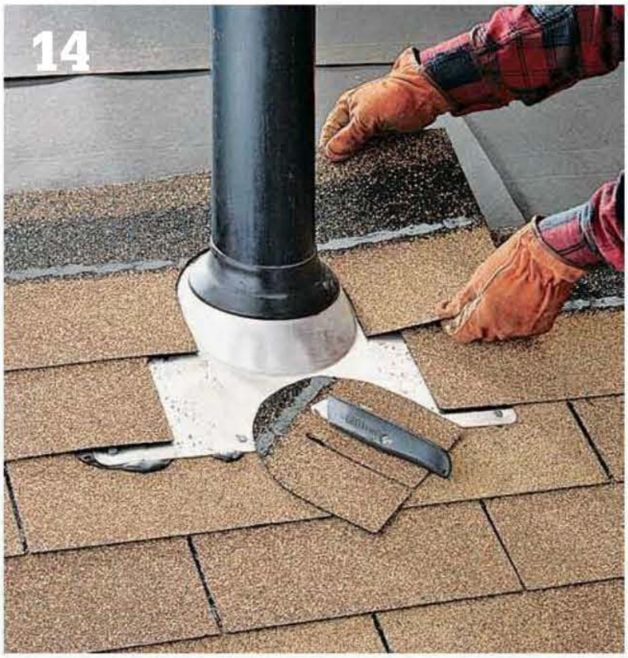

Cut shingles to fit around the neck of the flashing so they lie flat against the flange. Do not drive roofing nails through the flashing. Instead, apply roofing cement to the back of shingles where they lie over the flashing.

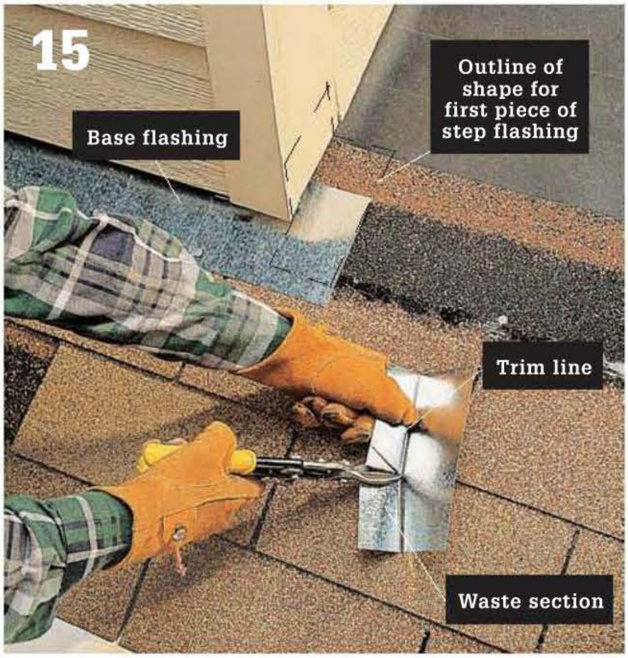

Shingle up to an element that requires flashing so the top of the reveal areas are within 5" of the element. Install base flashing using the old base flashing as a template. Bend a piece of step flashing in half and set it next to the lowest corner of the element. Mark a trim line on the flashing, following the vertical edge of the element. Cut the flashing to fit.

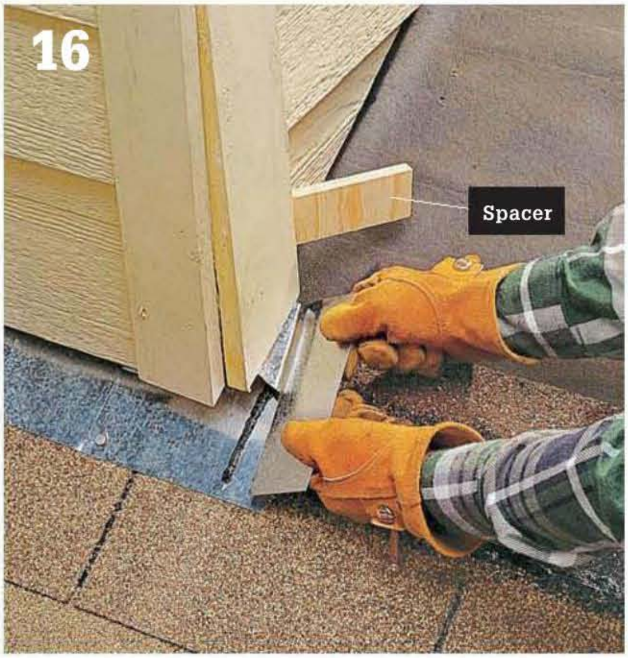

Pry out the lowest courses of siding and any trim at the base of the element. Insert spacers to prop the trim or siding away from the work area. Apply roofing cement to the base flashing in the area where the overlap with the step flashing will be formed. Tuck the trimmed piece of step flashing under the propped area, and secure the flashing. Fasten the flashing with one rubber gasket nail driven near the top and into the roof deck.

17

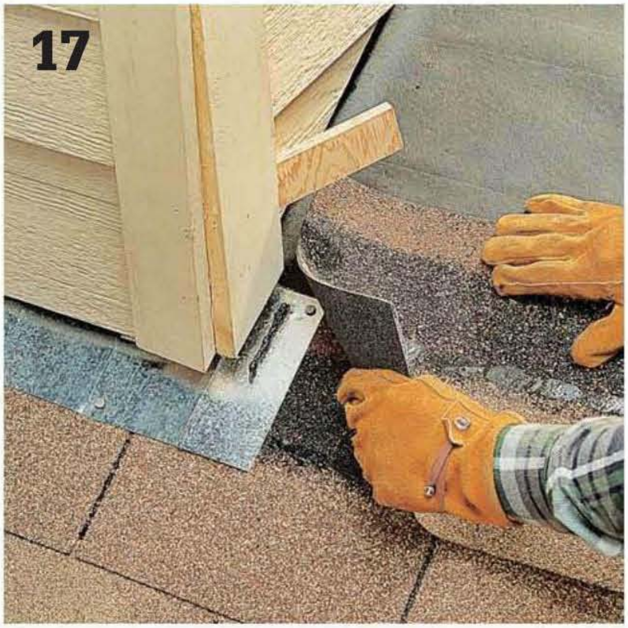

Apply roofing cement to the top side of the first piece of step flashing where it will be covered by the next shingle course. Install the shingle by pressing it firmly into the roofing cement. Do not nail through the flashing underneath.

18

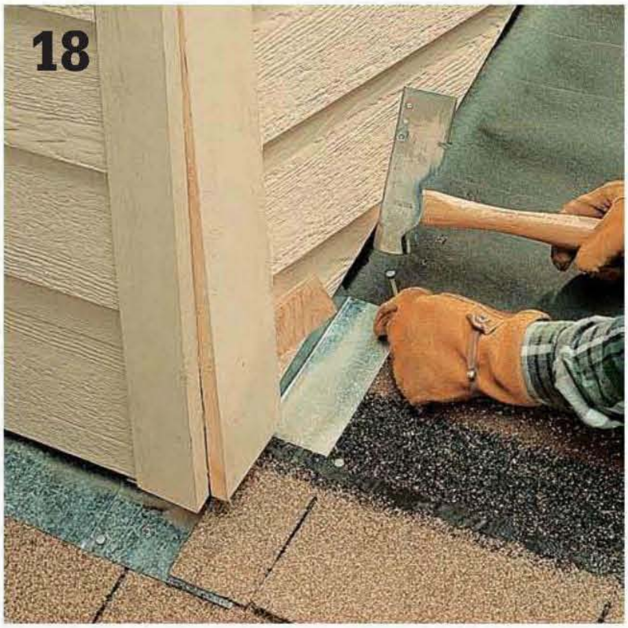

Tuck another piece of flashing under the trim or siding, overlapping the first piece of flashing at least 2". Set the flashing into roofing cement applied on the top of the shingle. Nail the shingle in place without driving nails through the flashing. Install flashing up to the top of the element the same way. Trim the last piece of flashing to fit the top corner of the element. Reattach the siding and trim.

19

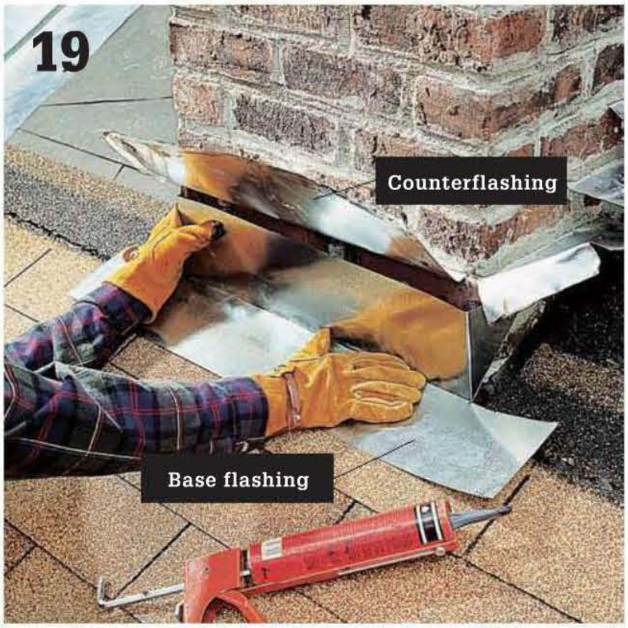

Shingle up to the chimney base. Use the old base flashing as a template to cut new flashing. Bend up the counter flashing. Apply roofing cement to the base of the chimney and the shingles just below the base. Press the base flashing into the roofing cement and bend the flashing around the edges of the chimney. Drive rubber gasket nails through the flashing flange into the roof deck.

20

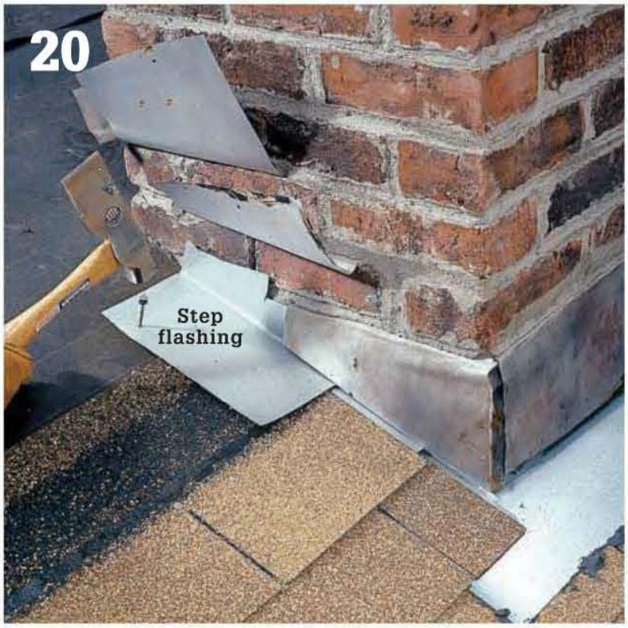

Install step flashing and shingles, working up to the high side of the chimney. Fasten flashing to the chimney with roofing cement. Fold down the counter flashing as you go.

(continued)

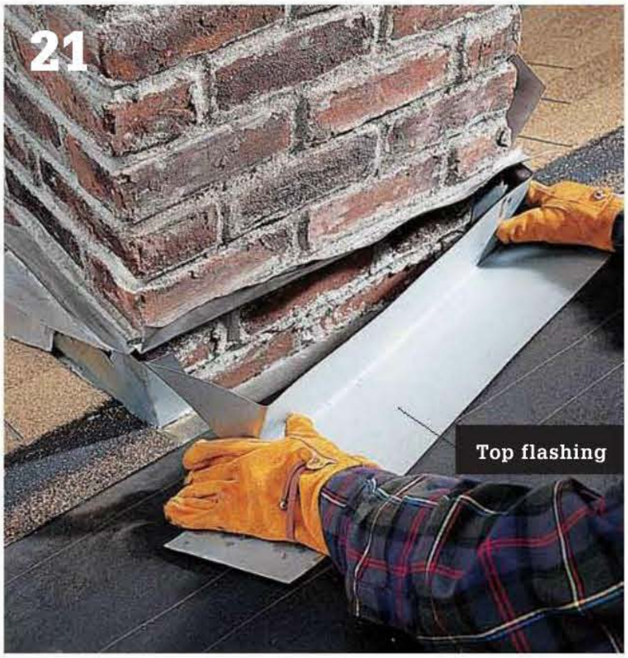

Cut and install top flashing (also called a saddle) around the high side of the chimney. Overlap the final piece of flashing along each side. Attach the flashing with roofing cement applied to the deck and chimney and with rubber gasket nails driven through the flashing base into the roof deck. Shingle past the chimney using roofing cement (not nails) to attach shingles over the flashing.

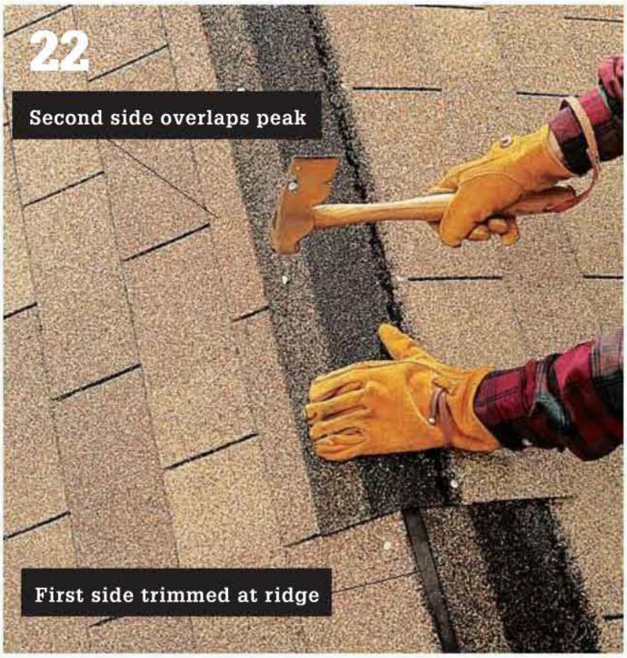

When you reach a hip or ridge, shingle up the first side until the top of the uppermost reveal area is within 5" of the hip or ridge. Trim the shingles along the peak. Install shingles on the opposite side of the hip or ridge. Overlap the peak no more than 5".

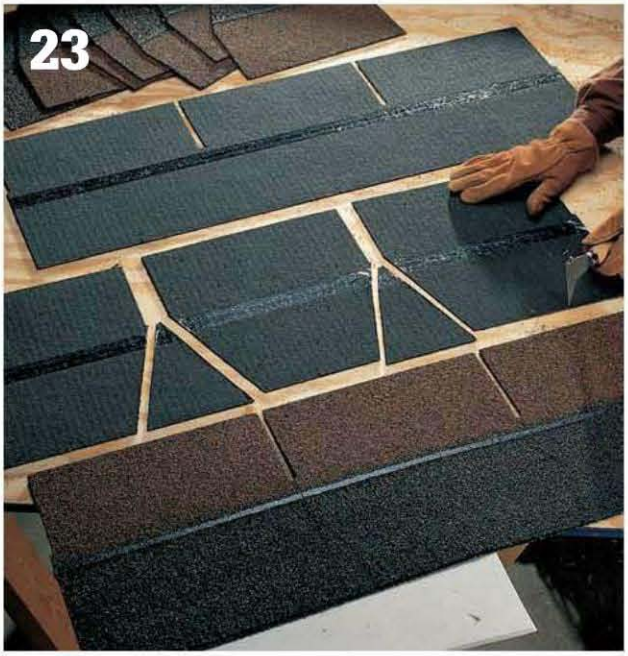

Cut three 12"-sq. cap shingles from each three-tab shingle. With the back surface facing up, cut the shingles at the tab lines. Trim the top corners of each square with an angled cut, starting just below the seal strip to avoid overlaps in the reveal area.

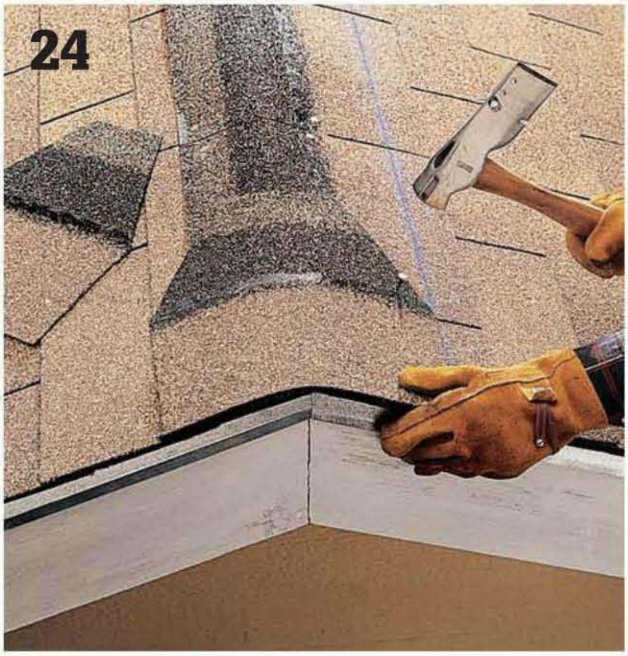

Snap a chalk line 6" down from the ridge, parallel to the peak. Attach cap shingles, starting at one end of the ridge, aligned with the chalk line. Drive two 1¼" roofing nails per cap about 1" from each edge, just below the seal strip.

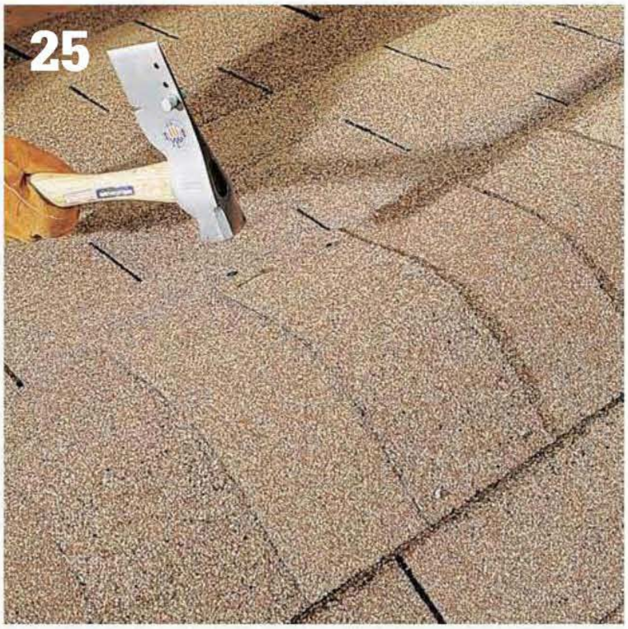

Following the chalk line, install cap shingles halfway along the ridge, creating a 5" reveal for each cap. Then, starting at the opposite end, install caps over the other half of the ridge to meet the first run in the center. Cut a 5"-wide section from the reveal area of a shingle tab, and use it as a "closure cap" to cover the joint where the caps meet.

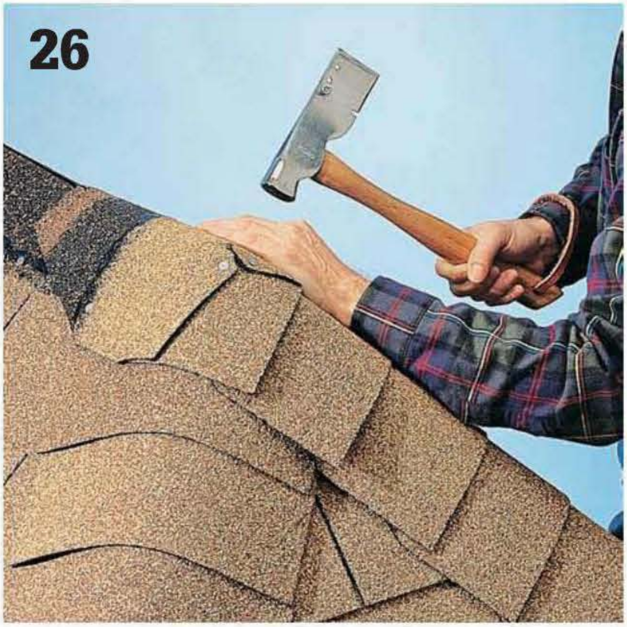

Shingle the hips in the same manner using a chalk reference line and cap shingles. Start at the bottom of each hip and work to the peak. Where hips join with roof ridges, install a custom shingle cut from the center of a cap shingle. Set the cap at the end of the ridge and bend the corners so they fit over the hips. Secure each corner with a roofing nail, and cover the nail heads with roofing cement.

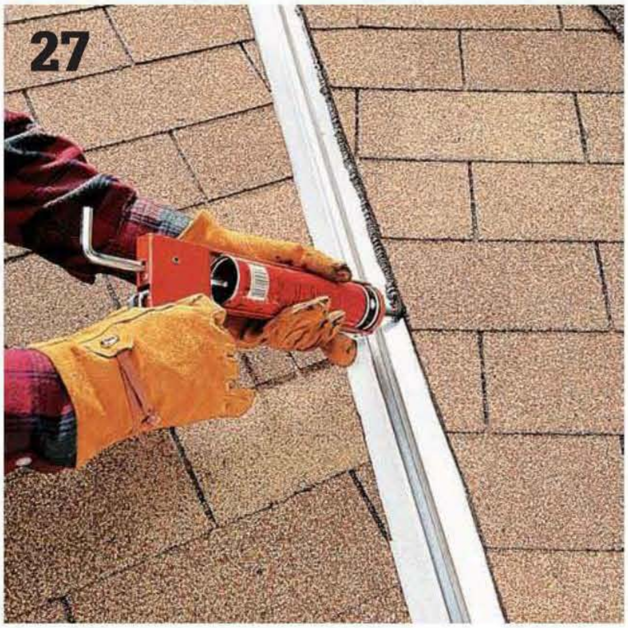

After all shingles are installed, trim them at the valleys to create a gap that's 3" wide at the top and widens at a rate of ⅛" per foot as it moves downward. Use a utility knife and straightedge to cut the shingles, making sure not to cut through the valley flashing. At the valleys, seal the undersides and edges of shingles with roofing cement. Also cover exposed nail heads with roofing cement.

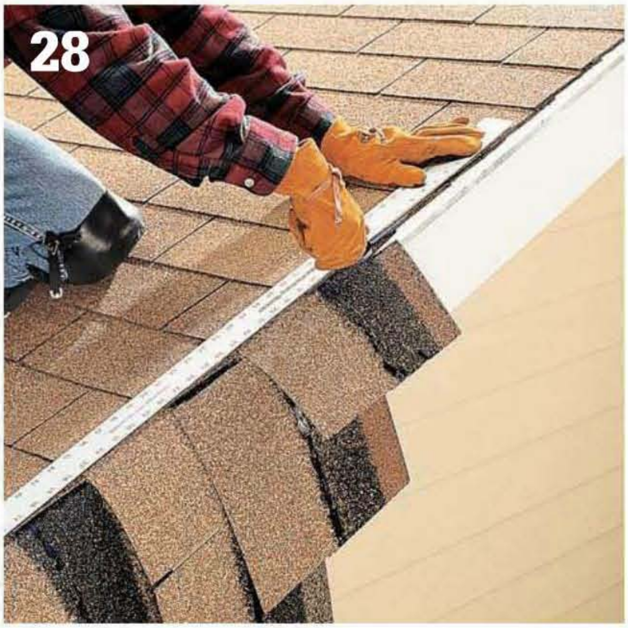

Mark and trim the shingles at the rake edges of the roof. Snap a chalk line ⅜" from the edge to make an overhang, then trim the shingles.

Installing Vinyl Siding

Vinyl has become one of the most popular sidings due to its low cost, uniform appearance, and maintenance-free durability. Installation is fairly simple, with each row locking onto the lip of the underlying course, then nailed along the top.

There are a couple of key factors that will make or break your siding project. First, the sheathing must be straight and solid before the siding is applied. The siding will only look as straight and smooth as the wall it's on. Second, determine how the siding should overlap to hide the seams from the main traffic patterns. This usually means starting in the back and working toward the front of the house.

Do not nail the siding tight to the house. The panels need to slide back and forth as they expand and contract with changes in the temperature. If the siding can't move, it will bow and need to be reinstalled. Keep a 1⁄32" gap between the head of the nail and the siding.

Vinyl siding is available in a wide variety of colors and styles, and with a lot of accessories such as trim, fluted lineals, vertical columns, crown molding, and band boards. The most common vinyl siding is horizontal lap siding, which is shown starting on page 548. This project shows a foam underlayment, which reduces outside noise, protects the siding from dents, and adds an insulation value. Vertical vinyl siding is also available. It's specifically made for vertical applications.

Tools & Materials ▸

Hammer
Circular saw
  or radial-arm saw
Clamps
Tape measure
String
Straightedge
Utility knife
Aviation snips
Level
Chalk line
Framing square
Nail slot punch
Zip-lock tool
Snap lock punch
Caulk gun
Vinyl siding
J-channel
Corner posts
Undersill
Starter strip
Nails
Cutting table
Safety glasses
Silicone caulk

Vinyl siding can look very similar to wood lap siding, but it doesn't require regular upkeep. Vinyl can be installed on any type and style of house.

Cutting Vinyl Siding

Use a fine-tooth blade installed backward in the saw to cut vinyl siding. Use a radial-arm saw, power miter saw, or a circular saw, and move the blade slowly through the siding. Always wear safety glasses when cutting siding.

Support the siding on a cutting table when cutting. Vinyl siding is too flimsy to be placed across sawhorses without support. You can build a cutting table by fastening a long piece of scrap plywood between two sawhorses.

Using Specialty Tools

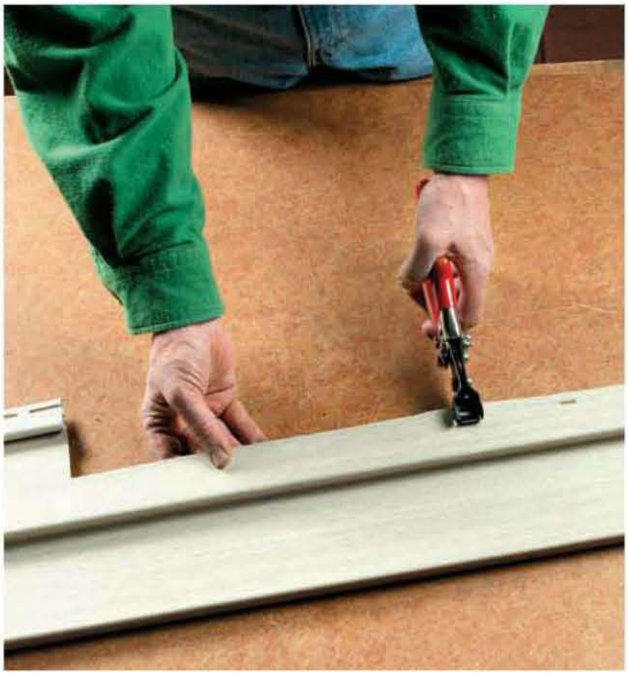

A snap lock punch is used to make raised tabs, or dimples, in a cut edge of siding where the nailing hem has been removed. This eliminates the need to facenail the panel.

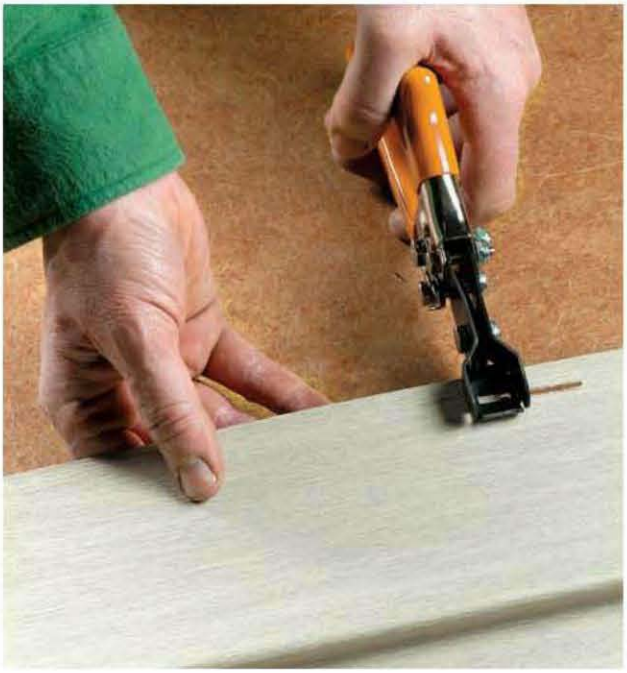

A nail slot punch is used to make horizontal nail slots in the face of panels. It can also be used to add or elongate the opening of an existing nail slot to match irregular stud spacing.

How to Install Vinyl Siding

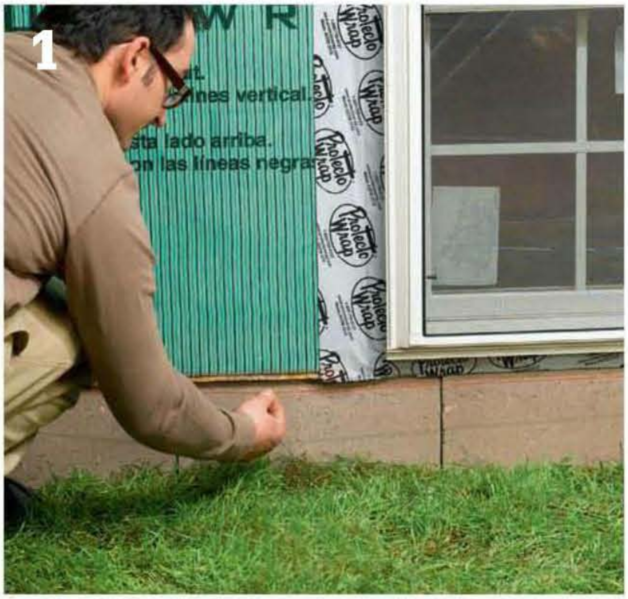

Install housewrap following manufacturer instructions. Identify the lowest corner of the house that has sheathing, and partially drive a nail 1½" above the bottom edge of the sheathing. Run a level string to the opposite corner of the wall and partially drive a nail. Do this around the entire house. Snap chalk lines between the nails.

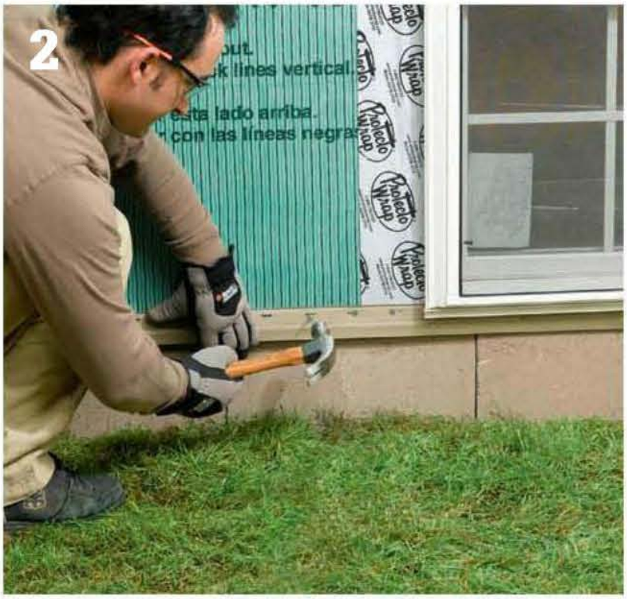

Place the top edge of the starter strip along the chalk line and nail every 10". Nail in the center of the slots and don't nail tight to the house. Keep a ¼" gap between strips, and leave space at the corners for a ½" gap between starter strips and corner posts.

Option: Install foam vinyl siding underlayment on the house using cap nails. Align the bottom of the underlayment with the starting strip. To cut panels to size, score them with a utility knife, then break them over your cutting table. Some panels need to be taped at the seams. Follow manufacturer's recommendations.

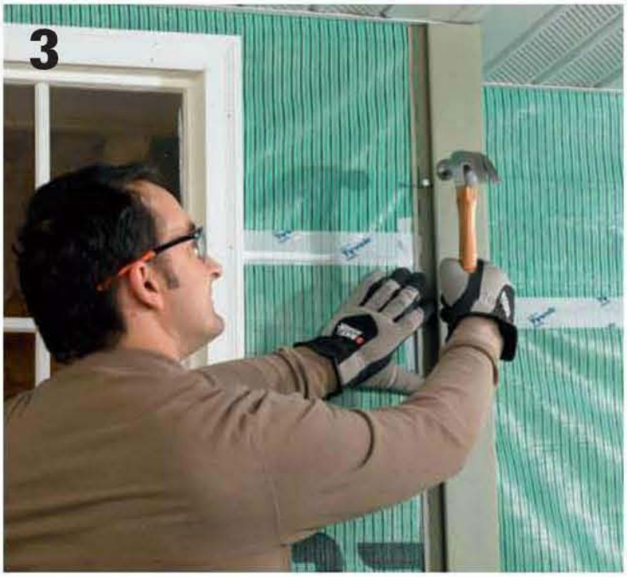

Install a corner post, keeping a ¼" gap between the top of the post and the soffit. Extend the bottom of the post ¼" below the bottom of the starter strip. Drive a nail at the top end of the uppermost slot on each side of the post (the post hangs from these nails). Make sure the post is plumb on both sides using a level. Secure the post by driving nails every 8 to 12" in the center of the slots. Do not nail the post tight. Install the other posts the same way.

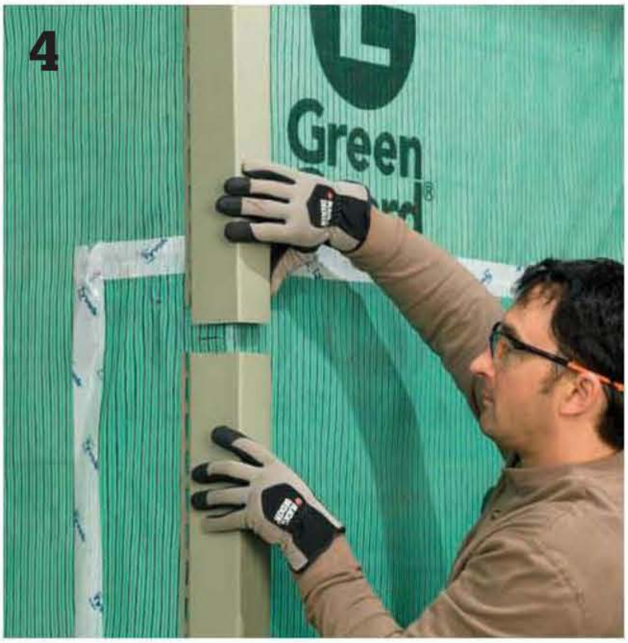

4

If more than one corner post is needed to span the length of a corner, the upper post overlaps the lower post. For an outside corner post, cut off 1" from the nailing flanges on the bottom edge of the top post. For an inside corner post, cut off 1" from the nailing flange on the upper edge of the bottom post. Overlap the posts by ¾", leaving ¼" for expansion.

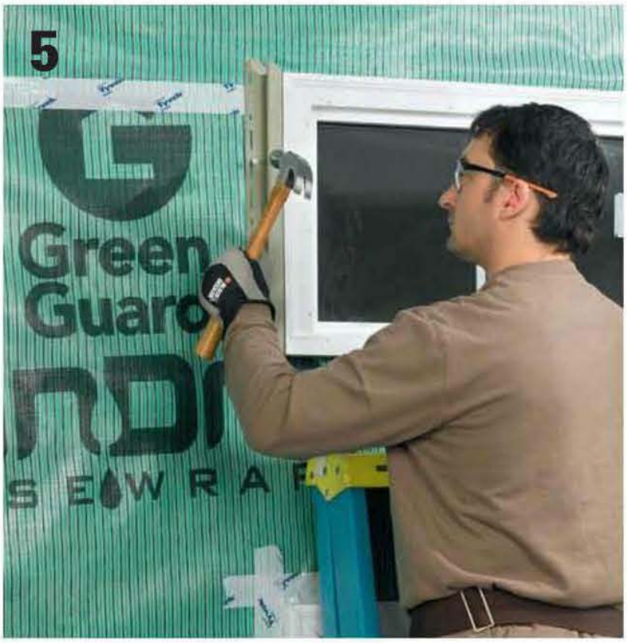

5

Measure and cut two J-channels that are the length of a window plus the width of the J-channel. Place one of the J-channels against the side of the window, aligning the bottom edge with the bottom edge of the window. Nail the channel in place. Nail the second J-channel against the opposite side of the window the same way.

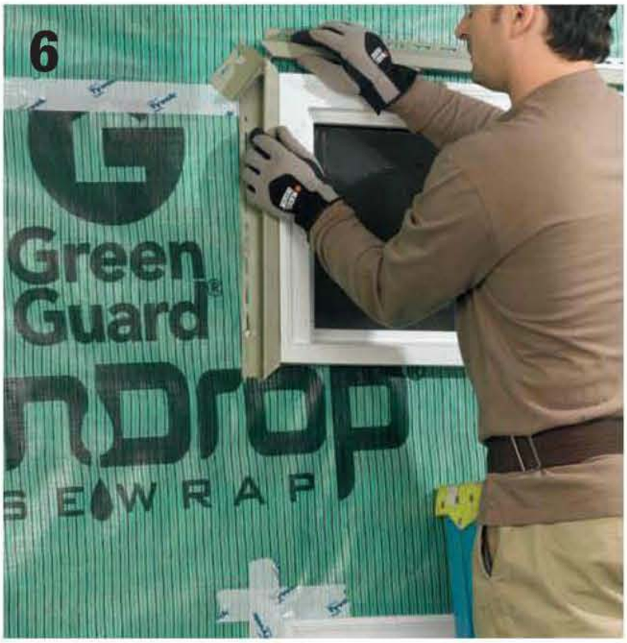

6

At the top of the window, measure between the outside edges of the side J-channels and cut a piece of J-channel to fit. Cut a ¾" tab at each end. Bend the tabs down to form a drip edge. Miter cut the face at each end at 45°. Center the J-channel over the window and nail it in place. The top J-channel overlaps the side pieces, and the drip edges fit inside the side pieces. Do this for each window and door.

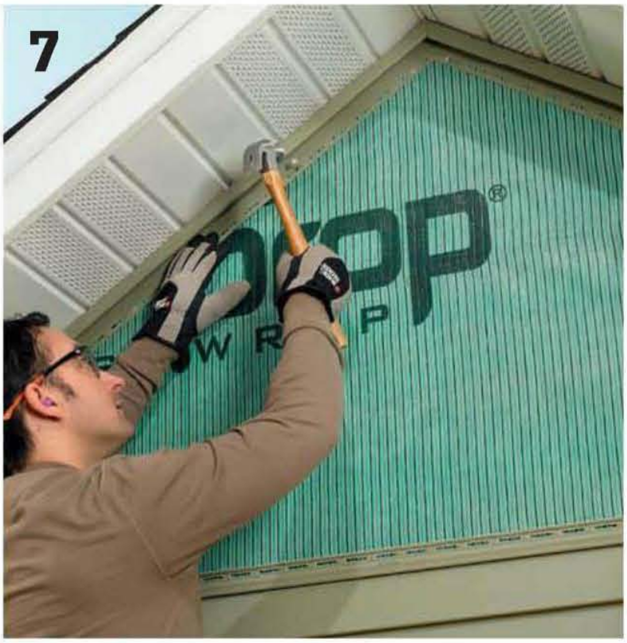

7

Measure, cut, and install J-channel along the gable ends. Nail the channels every 8 to 12". To overlap J-channels, cut 1" from the nailing hem. Overlap the channels ¾", leaving ¼" for expansion. At the gable peak, cut one channel at an angle to butt against the peak. Miter the channel on the opposite side to overlap the first channel.

(continued)

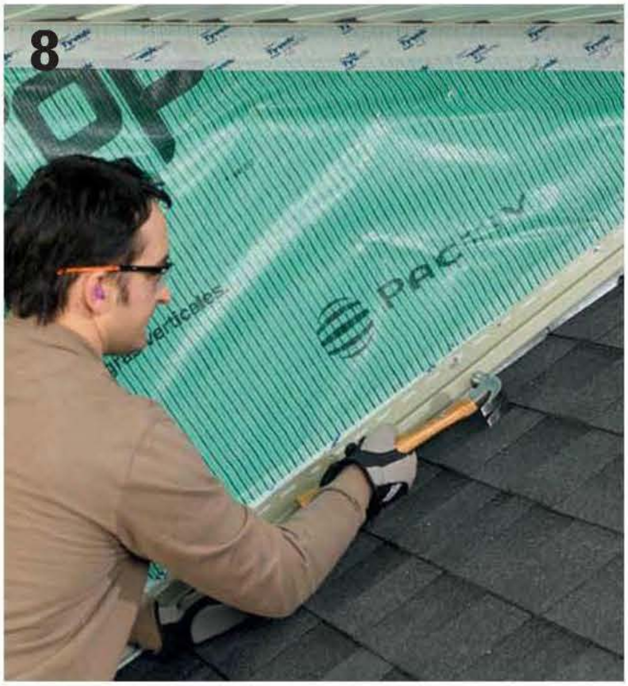

To install J-channel over a roof line, snap a chalk line along the roof flashing ½" above the roof. Align the bottom edge of the J-channel along the chalk line, and nail the channel in place. Make sure the channel does not make direct contact with the shingles.

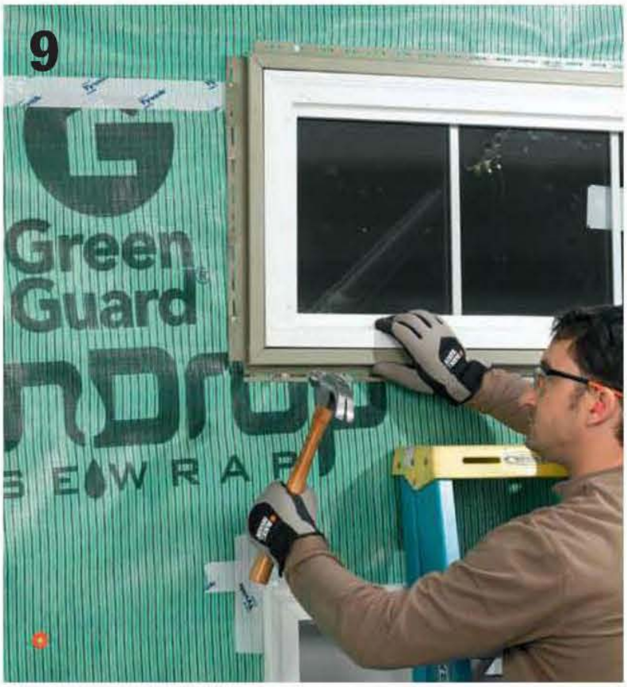

Measure, cut, and install undersill beneath each window. The undersill should be flush with the outside lip of the side channels.

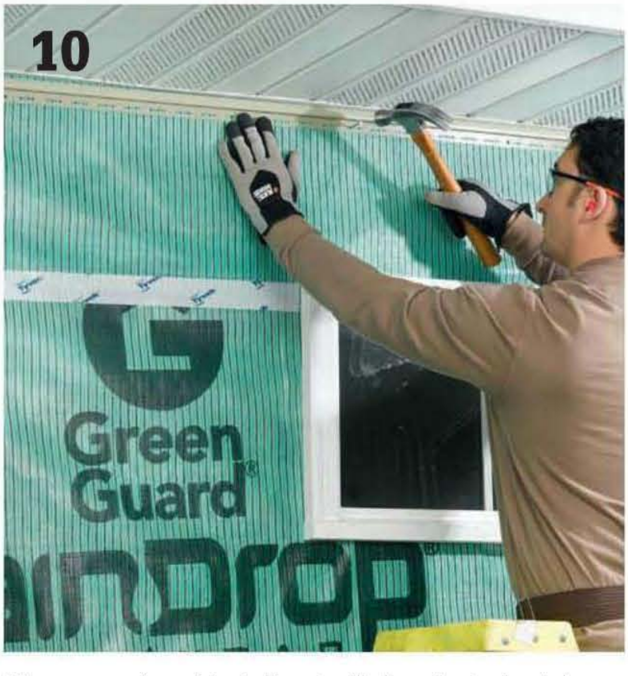

Measure, cut, and install undersill along the horizontal eaves on the house. If more than one undersill is needed, cut the nailing hem 1¼" from the end of one undersill. Overlap the undersills by 1".

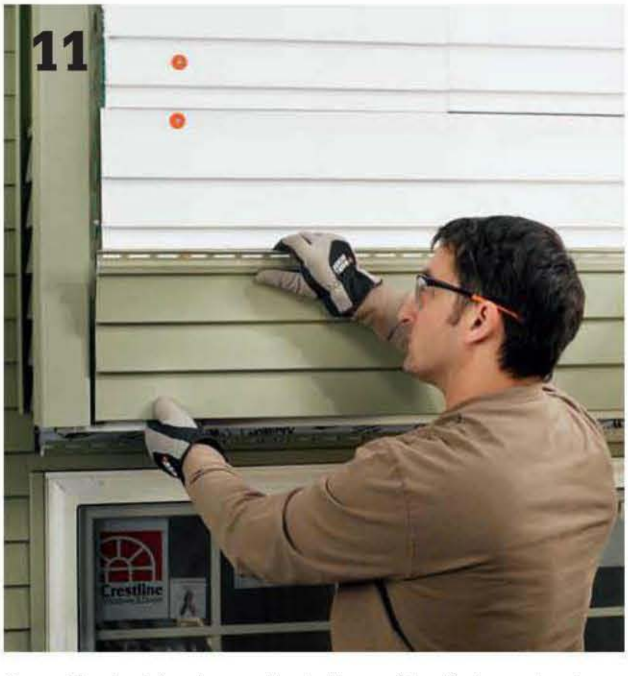

Snap the locking leg on the bottom of the first panel onto the starter strip, making sure it's securely locked in place. Keep a ¼" gap between the end of the panel and the corner post. Nail the panel a minimum of every 16" on center. Don't drive the nails tight. *Note: This installation shows a vinyl siding underlayment in place.*

12

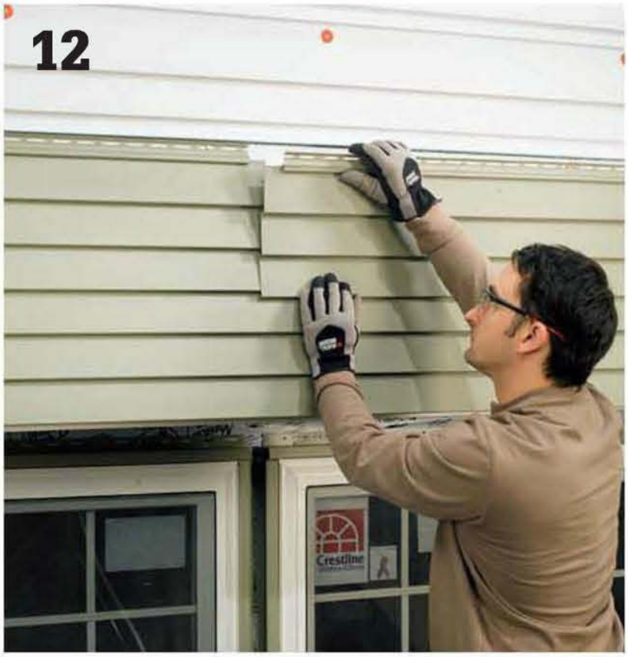

Overlap panels by 1". Cut panels so the factory cut edge is the one that's visible. Keep nails at least 6" from the end of panels to allow for smooth overlap. Do not overlap panels directly under a window.

13

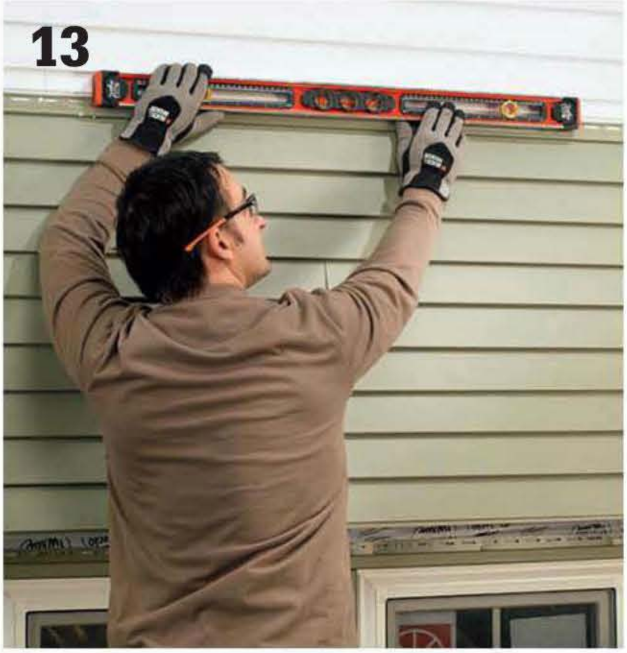

Place the second row over the first, snapping the locking leg into the lock of the underlying panels. Leave ¼" gap at corners and J-channels. Install subsequent rows, staggering seams at least 24" unless separated by more than three rows. Check every several rows for level. Make adjustments in slight increments, if necessary.

14

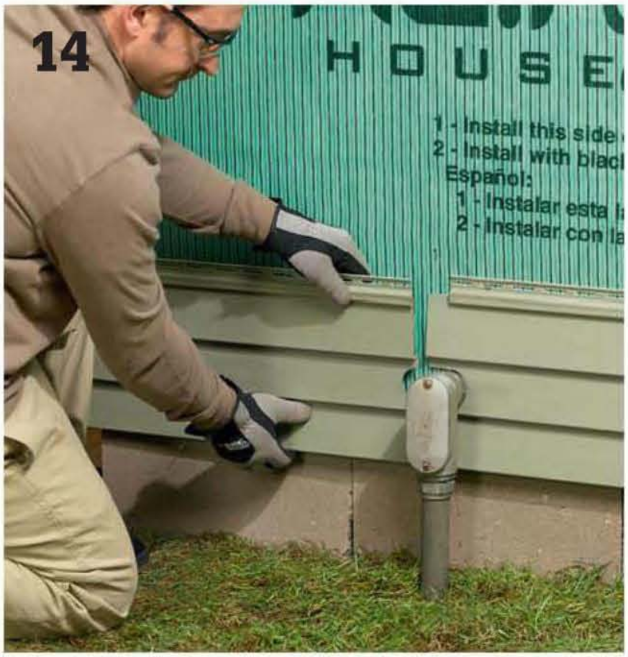

For hose spigots, pipes, and other protrusions, create a seam at the obstacle. Begin with a new panel to avoid extra seams. Cut an opening ¼" larger than the obstacle, planning for a 1" overlap of siding. Match the shape and contour as closely as possible. Fit the panels together around the obstruction and nail in place.

15

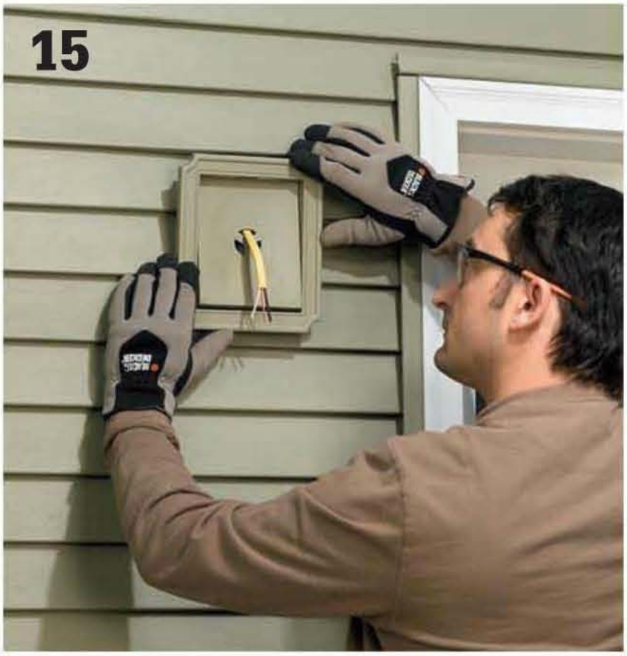

Place mounting blocks around outlets, lights, and doorbells. Assemble the base around the fixture, making sure it's level, and nail in place. Install siding panels, cutting them to fit around the mounting block with a ¼" gap on each side. Fasten the cover by snapping it over the block.

(continued)

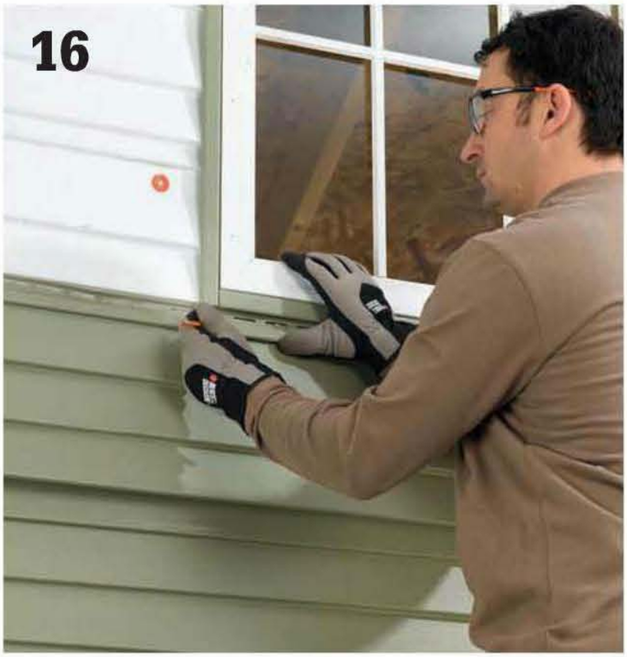

Where panels must be notched to fit below a window, position the panel below the window and mark the edges of the window, allowing for a ¼" gap. Place a scrap piece of siding alongside the window and mark the depth of the notch, keeping a ¼" gap. Transfer the measurement to the panel, mark the notch, and cut it out. Create tabs on the outside face every 6" using a snap lock punch. Install the panel, locking the tabs into the undersill.

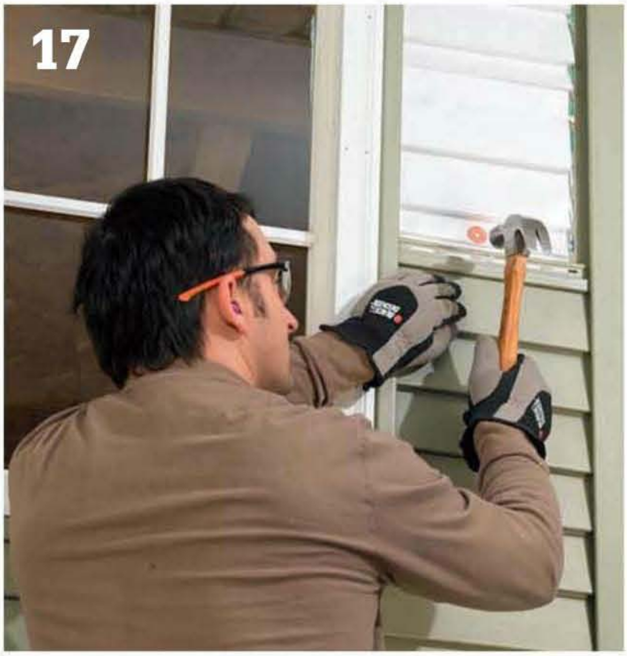

Install cut panels between windows and between windows and corners as you would regular panels. Avoid overlapping panels and creating seams in small spaces. The panels need to align with panels on the opposite side of the window.

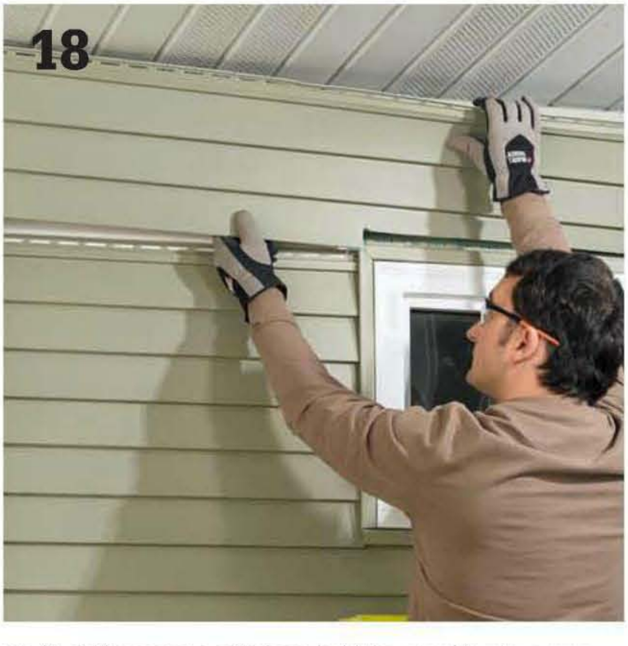

To fit siding over a window, hold the panel in place over the window and mark it. Use a scrap piece of siding to mark the depth of the cut. Transfer the measurement to the full panel and cut the opening. Fit the cut edge into the J-channel above the window, lock the panel in place, and nail it.

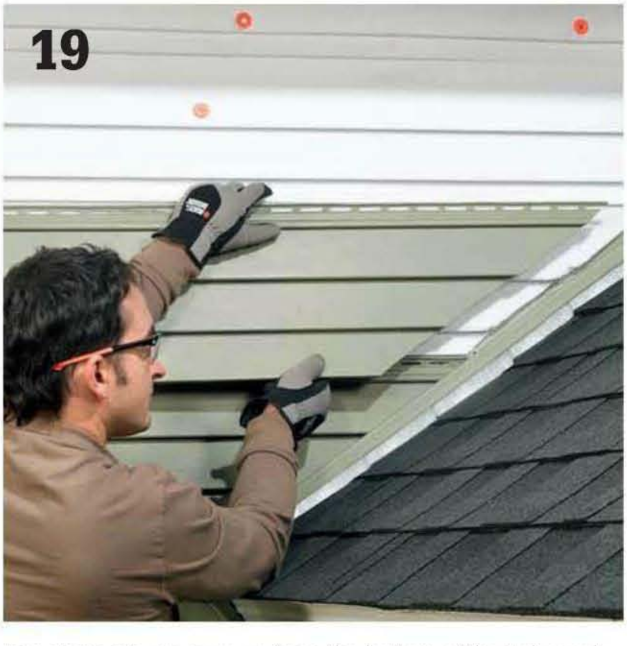

For dormers, measure up from the bottom of the J-channel the height of a panel and make a mark. Measure across to the opposite J-channel. Use this measurement to mark and cut the panel to size. Cut and install panels for the rest of the dormer the same way.

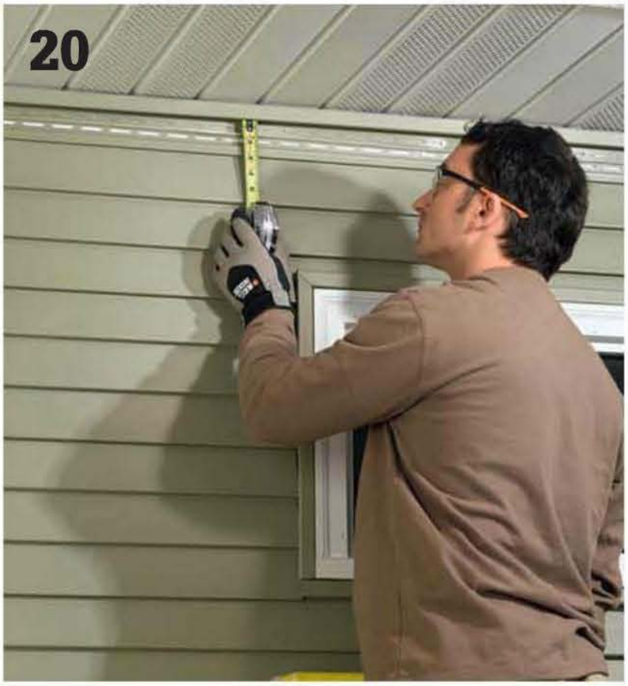

20

Measure the distance between the lock on the last fully installed panel and the top of the undersill under the horizontal eaves. Subtract ¼", then mark and rip a panel to fit. Use a snap lock punch to punch tabs on the outside face every 6". Install the panel, locking the tabs into the undersill.

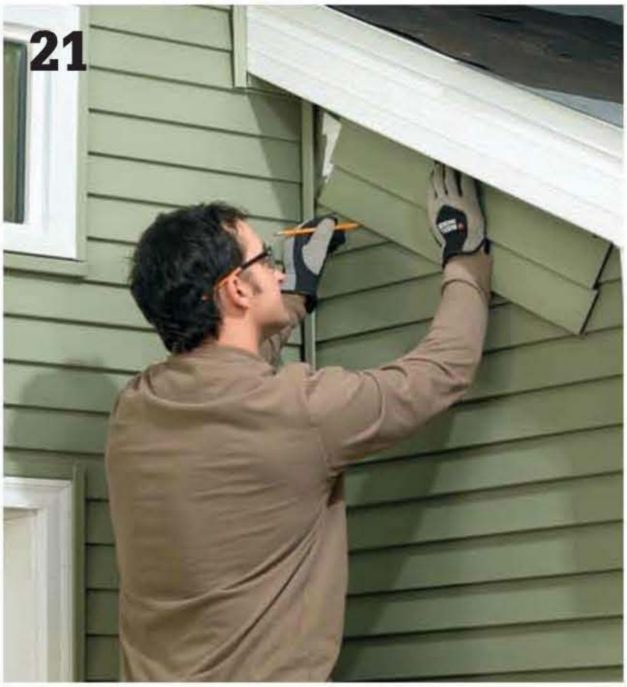

21

Place a scrap panel in the J-channel along the gable end of the house. Place another scrap over the last row of panels before the gable starts, slide it under the first scrap, and mark the angle where they intersect. Transfer this angle to full panels. Make a similar template for the other side. Cut the panels and set the cut edge into the J-channel, leaving a ¼" gap.

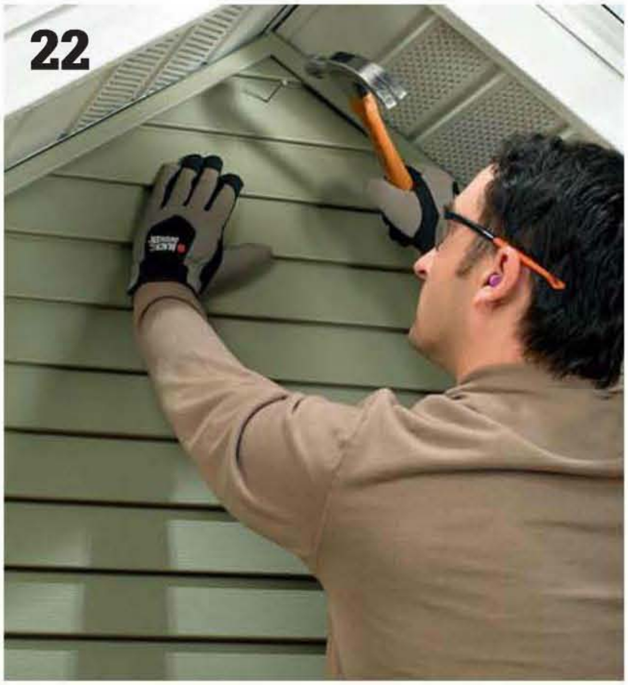

22

Cut the last piece of siding to fit the gable peak. Drive a single aluminum or stainless steel finish nail through the top of the panel to hold it in place. This is the only place where you will facenail the siding.

23

Apply caulk between all windows and J-channel, and between doors and J-channel.

Resources

American Institute of Architects
800-364-9364
www.aiaonline.com

American Lighting Association
800-724-4484
www.americanlightingassoc.com

American Society of Interior Designers
202-546-3480
www.asid.org

Association of Home Appliance Manufacturers
202-872-5955
www.aham.org

Certified Forest Products Council
503-224-2205
www.certifiedwood.org

Construction Materials Recycling Association
630-548-4510
www.cdrecycling.org

Energy & Environmental Building Association
952-881-1098
www.eeba.org

International Residential Code (book)
International Conference of Building Officials
800-284-4406
www.icbo.com

Kohler Co.
800-456-4537
www.kohler.com

Laticrete International
(radiant floor mats)
800-243-4788
www.laticrete.com

Light Gauge Steel Engineers Association
615-279-9251
www.lgsea.com

National Association of the Remodeling Industry (NARI)
847-298-9200
www.nari.org

National Fire Protection Agency
617-770-3000
www.nfpa.org

National Kitchen & Bath Association (NKBA)
800-843-6522
www.nkba.org

National Wood Flooring Association
800-422-4556
www.woodfloors.org

North American Insulation Manufacturers Association
703-684-0084
www.naima.org

Tile Council of America
864-646-8453
www.tileusa.com

U.S. Environmental Protection Agency—Indoor Air Quality
www.epa.gov/iaq/

Photo Credits

Alcoa Home Exteriors, Inc.
p. 531, 546

American Standard
p. 438

Armstrong World Industries
p. 174, 180 (left), 236

Brian Greer's Tin-Ceilings
519 570 1447
p. 239

Broan Nutone
p. 327

Ceramic Tiles of Italy
p. 454

CertainTeed, Corp.
p. 530 (lower)

Cherry Tree Design
p. 428 (lower)

Daltile
p. 386

Economic Mobility, Inc.
p. 425 (lower left)

Elkay
p. 348, 426 (left)

The Energy Conservatory
www.energyconservatory.com
p. 218 (right) Energy Star Program

Fotolia / www.fotolia.com
p. 480

General Electric
p. 418 (right)

Ginger
p. 426 (right)

Harrell Remodeling, Inc.
p. 427, 486

IKEA Home Furnishings
p. 416

ISTOCK Photo / www.istock.com,
p. 324 George Peters, 420, 422, 468
Nicola Gavin, 528

Jacuzzi
p. 428 (top)

Kentucky Wood Floors
p. 170

Kohler
p. 19, 392, 424, 425 (lower right), 446, 484

Kraftmaid
p. 339

Marvin Windows
p. 318

MIRAGE Prefinished Hardwood Floors
p. 162

Oasis Montana
p. 418 (left)

Robert Perron, photographer / www.bobperron.com
p. 516 (left & right)

Price Pfister
p. 393 (all), 472 (top & lower)

Seattle Glass Block
p. 220

SieMatic
p. 336

Swanstone
p. 470 (top)

Toto
p. 425 (top)

Urban Homes, NY
p. 340

Brian Vanden Brink, photographer / www.brianvandenbrink.com
p. 492

VELUX-America
p. 278

Weather Shield
p. 18 (top)

Western Red Cedar Lumbar Association
p. 228

Conversion Charts

Metric Equivalent

Inches (in.)	1/64	1/32	1/25	1/16	1/8	1/4	3/8	2/5	1/2	5/8	3/4	7/8	1	2	3	4	5	6	7	8	9	10	11	12	36	39.4
Feet (ft.)																								1	3	3 1/12
Yards (yd.)																									1	1 1/12
Millimeters (mm)	0.40	0.79	1	1.59	3.18	6.35	9.53	10	12.7	15.9	19.1	22.2	25.4	50.8	76.2	101.6	127	152	178	203	229	254	279	305	914	1,000
Centimeters (cm)							0.95	1	1.27	1.59	1.91	2.22	2.54	5.08	7.62	10.16	12.7	15.2	17.8	20.3	22.9	25.4	27.9	30.5	91.4	100
Meters (m)																								.30	.91	1.00

Converting Measurements

To Convert:	To:	Multiply by:
Inches	Millimeters	25.4
Inches	Centimeters	2.54
Feet	Meters	0.305
Yards	Meters	0.914
Miles	Kilometers	1.609
Square inches	Square centimeters	6.45
Square feet	Square meters	0.093
Square yards	Square meters	0.836
Cubic inches	Cubic centimeters	16.4
Cubic feet	Cubic meters	0.0283
Cubic yards	Cubic meters	0.765
Pints (U.S.)	Liters	0.473 (Imp. 0.568)
Quarts (U.S.)	Liters	0.946 (Imp. 1.136)
Gallons (U.S.)	Liters	3.785 (Imp. 4.546)
Ounces	Grams	28.4
Pounds	Kilograms	0.454
Tons	Metric tons	0.907

To Convert:	To:	Multiply by:
Millimeters	Inches	0.039
Centimeters	Inches	0.394
Meters	Feet	3.28
Meters	Yards	1.09
Kilometers	Miles	0.621
Square centimeters	Square inches	0.155
Square meters	Square feet	10.8
Square meters	Square yards	1.2
Cubic centimeters	Cubic inches	0.061
Cubic meters	Cubic feet	35.3
Cubic meters	Cubic yards	1.31
Liters	Pints (U.S.)	2.114 (Imp. 1.76)
Liters	Quarts (U.S.)	1.057 (Imp. 0.88)
Liters	Gallons (U.S.)	0.264 (Imp. 0.22)
Grams	Ounces	0.035
Kilograms	Pounds	2.2
Metric tons	Tons	1.1

Converting Temperatures

Convert degrees Fahrenheit (F) to degrees Celsius (C) by following this simple formula: Subtract 32 from the Fahrenheit temperature reading. Then mulitply that number by 5/9. For example, 77°F - 32 = 45. 45 × 5/9 = 25°C.

To convert degrees Celsius to degrees Fahrenheit, multiply the Celsius temperature reading by 9/5, then add 32. For example, 25°C × 9/5 = 45. 45 + 32 = 77°F.

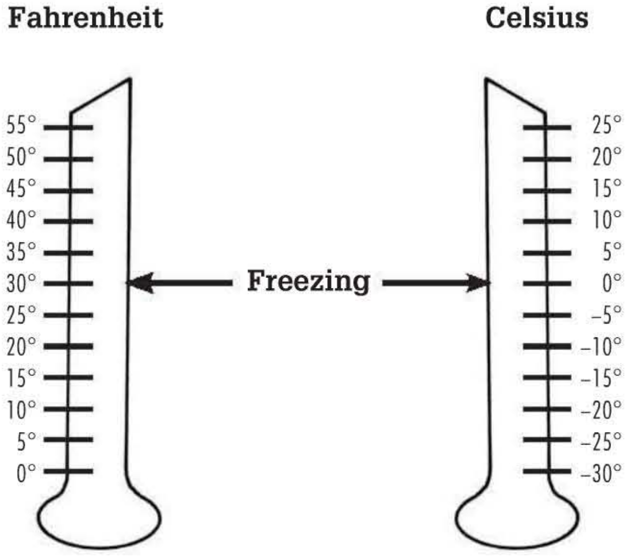

Contributing Editors, Art Directors, Set Builders, and Photographers

Phil Aarrestad
Dan Anderson
Glenn Austin
Randy Austin
Kim Bailey
Mark Biscan
Stewart Block
Charlie Boldt
Scott Boyd
Rose Brandt
Gary Branson
Greg Breining
Timothy Bro
Dave Brus
Keith Bruzelius
Ron Bygness
Peter Caley
Jennifer Caliandro
Rudy Calin
Tom Carpenter
Julie Caruso
Dan Cary
Janice Cauley
Marcia Chambers
Scott Christiansen
Tom Cooper
Paul Currie
Cy DeCosse
Jim Destiche
Doug Deutsche
Diane Dreon
Arthur Durkee
Barbara Falk
Jerri Farris
Lyle Ferguson
Mary Firestone
Steve Galvin
Kim Gerber
Abby Gnagey
Patricia Goar
Paul Gorton
John Haglof
Lynne Hanauer
Mark Hardy
David Hartley
Carol Harvatin
Rebecca Hawthorne
Tom Heck
Jon Hegge
Mike Hehner
Tami Helmer
Paul Herda
John Hermansen
Tim Himsel
Jonathan Hinz
Sara Holle
Lori Holmberg
Jim Huntley
Rex Irmen
Troy Johnson
Kari Johnston
Rob Johnstone
William B. Jones
Phil Juntti
Andrew Karre
Patrick Kartes
Geoffrey Kinsey
Tony Kubat
Karl Larson
John Lauenstein
Tom Lemmer
Bill Lindner
Earl Lindquist
Daniel London
Barbara Lund
Curtis Lund
Sarah Lynch
Mark Macemon
Bernice Maehren
Dave Mahoney
Paul Markert
Chris Marshall
Brett Martin
Jamey Mauk
Jeanette Moss McCurdy
John Nadeau
Paul Najlis
Bill Nelson
Charles Nields
Mette Nielsen
Kristen Olson
Carol Osterhus
Brad Parker
Mike Parker
Christian Paschke
Matthew Paymar
Mike Peterson
Greg Pluth
Anne Price-Gordon
John Riha
Joe Robillard
Mary Rohl
Tom Rosch
Susan Roth
Andrea Rugg
Gary Sandin
Joel Schmarje
Philip Schmidt
Mark Scholtes
Ned Scubic
Gina Seeling
Cathleen Shannon
Mike Shaw
Hugh Sherwood
Dane Smith
Danielle Smith
Steve Smith
Angela Spann
Dick Sternberg
Ruth Strother
Lori Swanson
Andrew Sweet
Dianne Talmage
Glenn Terry
Keith Thompson
Adrianne Truthe
Gregory Wallace
Kevin Walton
Robert Weaver
John Webb
Brad Webster
Wayne Wendland
John Whitman
Dan Widerski
Steve Willson
Christopher Wilson
Mike Woodside

Index

A

- ABS pipes. *See* Plastic pipes (ABS, PVC, CPVC)
- Acoustical ceiling tiles, installing, 236–238
- Air-conditioning, guidelines for kitchen, 345–346
- Ampacity (amperage) of wires, 90, 91
- Appliances
  - circuit maps for, 142–143
  - dishwashers, hooking up, 402–406
  - energy efficient, 347, 418–419
  - food disposers, installing, 407–411
  - minimum trap sizes for, 80
- Asbestos, 24, 176
- Attics
  - finishing ceilings, 516–517
  - installing
    - baseboard heaters, 524–527
    - floors, 506–509
    - walls, 514–515
  - planning remodeling, 494–495, 498

B

- Balloon framing, elements of, 11
- Bamboo flooring, installing, 186–190
- Bamboo panels, 210
- Baseboard heaters
  - circuit maps for, 143
  - installing, 524–527
- Basements
  - adding egress windows, 518–523
  - installing
    - baseboard heaters, 524–527
    - floors, 500–505
    - walls, 510–513
  - planning remodeling, 496–499
- Base molding, installing, 242–245
- Bathrooms
  - accessibility of, 424–429, 486–487
  - codes & standards, 135, 422–423
  - installing
    - alcove bathtubs, 438–445
    - drains, 472–473, 478–479
    - faucets, 472–477
    - flooring
      - ceramic tile, 191–197
      - mosaic tile, 198–199
    - floor-warming systems, 156–161
    - glass shelving, 488–489
    - grab bars, 486–487
    - showers
      - enclosure kits, 430–435
      - hinged doors for, 436–437
    - sinks
      - pedestal, 468–469
      - vessel, 464–467
    - sliding tub doors, 450–453
    - 3-piece tub surrounds, 446–449
    - toilets, 458–461
    - towel warmers, 490–491
    - tub spouts, 480–483
    - vanity cabinets, 470–471
    - vanity lights, 484–485
    - wall-hung vanities, 462–463
  - tiling walls, 454–457
- Bay windows, installing, 268–277
- Beadboard wainscoting, installing, 214–219
- Bedrooms, highlights of National Electrical Code about, 135
- Bifold doors, installing interior, 300–301
- Branch pipes, sizing for, 79
- Brass valves, soldering, 62
- Building codes & permits, 16–19
  - bathrooms, 135, 422–423
  - electrical
    - basic requirements, 131–131
    - highlights, 134–135
    - learning about, 128, 131
  - kitchens, 135, 342–343
  - for PEX, 71
  - for plumbing, 77–78
- Built-up base molding, installing, 244–245
- Butted casings on doors, installing, 319

C

- Cabinets
  - installing bathroom vanity, 470–471
  - installing kitchen
    - base cabinets, 356–357
    - hanging cabinets, 353–355
    - overview of, 350
    - preparing walls, 351–352
  - removing, 39
- Cables
  - entering receptacles, 122
  - highlights of National Electrical Code about, 134
  - installing NM (Nonmetallic), 98–101
  - overview of, 90–92
- Carpet squares, installing, 200–203
- Ceilings
  - finishing attic, 516–517
  - installing
    - acoustical tiles, 236–238
    - crown molding, 250–253
    - electrical boxes on, 114
    - light fixtures on, 324–326
    - metal, 239–241
    - panels on, 228–231
    - recessed light fixtures on, 327–329
    - skylights, 278–283
      - building shafts for, 284–287
      - tubular, 288–289
    - suspended, 232–235
    - track lighting on, 330–333
    - wallboard on flat, 47
- Cementboard underlayment, installing, 152
- Ceramic tile flooring, installing, 191–197
- Chair rails, installing, 248–249
- Circuit maps, using, 138–145
- Codes. *See* Building codes & permits
- Color chart of wires, 90
- Concrete walls, installing wires on, 106–107
- Conductors, maximum number permitted per electrical box, 108
- Conduit
  - connecting nonmetallic, 105
  - installing on concrete walls, 106–107
  - overview of, 102–104
- Construction plans, 22–23
- Copper pipes
  - cutting, 59
  - soldering, 56, 58, 60–62
  - working with, 56–57
- Countertops, installing
  - laminate, 368–375
  - postform, 364–367
  - tiled, 376–385
- CPVC pipes. *See* Plastic pipes (ABS, PVC, CPVC)
- Crown molding, installing, 250–253

D

- Demolition
  - making opening for exterior doors, 32–35
  - overview of, 28
  - removing
    - cabinets, 39
    - partition walls, 36–37
    - plaster walls, 31
    - trim, 38
    - wallboard, 30
  - support, installing temporary, 29
- Dining rooms, highlights of National Electrical Code about, 135
- Dishwashers, hooking up, 402–406

Doors
  codes for exit, 17
  exterior
    making opening for in lap siding, 32–33
    making opening for in stucco, 34–35
  framing, 12–13
  framing interior, 290–291
    for load-bearing walls, 293
    prehung, 292
    sliding & folding doors, 293
  guidelines for kitchen, 344
  installing casings, 318–321
  installing entry, 306–309
  installing interior
    bifold doors, 300–301
    French doors, 302–305
    overview of, 290
    pocket doors, 296–299
    prehung, 294–295
  installing paneling around, 213
  installing patio, 312–317
  installing storm, 310–311
Drains
  hooking up kitchen sink, 398–401
  installing bathroom sink, 472–473, 478–479
Drain-waste-vent system
  cleanouts for, 80
  drain pipe sizes, 80
  overview of, 55
  vent pipe sizes & orientation, 80
Drawings, 20–21
Dry-backed resilient flooring tiles, installing, 175, 176, 180–184

E
Electrical boxes
  common types of, 108, 109, 110
  highlights of National Electrical Code about, 134
  installing
    on ceilings, 114
    junction boxes, 114–115
    for light fixtures, 113
    for receptacles, 112
    for switches, 113
  locating, 111
  maximum number of conductors permitted per box, 108
  overview of, 85, 108
  wall covering thickness and, 111
Electrical symbols, 137
Electrical system
  delivery from power plants, 83
  glossary of terms, 87
  guidelines for kitchen service, 344
  meters described, 84
  overview of home, 86
  parts of, 84–85
  safety, 25, 28
Entertainment rooms, highlights of National Electrical Code about, 135
Entry doors, installing, 306–309
Exits & openings, codes for, 17

F
Faucets
  installing bathroom, 472–477
  installing kitchen
    overview of, 392–393
    putting in new faucet, 395–397
    removing old faucet, 394
Fire protection, codes for, 18
First-aid kits, 25
Fittings
  methods used with copper pipes, 56
  soldering copper pipes &, 56, 58, 60–62
Fixtures
  installing panels & wainscoting around, 213, 218
  minimum trap sizes for plumbing, 80
  See also Light fixtures
Flexible copper pipes
  cutting, 59
  overview of, 56, 57
Flooring
  elements of, 14
  installing
    bamboo planks, 186–190
    carpet squares, 200–203
    ceramic tile, 191–197
    hardwood strip, 162–165
    laminate, 166–169
    mosaic tile, 198–199
    parquet, 170–173
    radiant floor-warming systems, 156–161
    resilient tiles, 175, 176, 180–185
    sheet vinyl, 174–179
  installing attic, 506–509
  installing basement, 500–505
  refinishing wood, 204–207
  underlayment for
    installing
      cementboard, 152
      isolation membrane, 153
      panels, 154
      plywood, 151
      raised panels, 154–155
    overview of, 150
Food disposers, installing, 407–411
Four-way switches
  circuit maps for, 145
  installing of, 119
  overview of, 119
  testing, 121
Framing
  doors, 12–13, 290–293
  elements of, 10–11
  with steel, 42–43
  windows, 12–13, 256–259
French doors, installing interior, 302–305
FRP panels, 210

G
Garden windows, installing, 264–267
GFCI (ground-fault circuit-interrupter) receptacles
  circuit maps for, 139, 142
  installing, 127
  overview of, 126
Glassblock walls, building, 220–223
Glass shelving, installing, 488–489
Grab bars, installing, 486–487
Green kitchens, 347, 416–419
Grounding
  highlights of National Electrical Code about, 135
  wires
    described, 84
    installing through metal conduits, 102

H
Hallways, highlights of National Electrical Code about, 135
Handicapped persons. *See* Universal design
Hard copper pipes. *See* Rigid copper pipes
Hardwood strip flooring, installing, 162–165
Headers, overview of, 12–13
Heating
  guidelines for kitchen, 345–346
  installing baseboard heaters, 524–527
  installing panels & wainscoting around vents, 213, 218
  installing towel warmers, 490–491

I
Inspections, 16
Isolation membrane underlayment, installing, 153

J
Joints, taking apart soldered, 63
Joists, installing temporary, 29
Junction boxes, installing, 114–115

K
Kitchens
  accessibility of, 348–349
  appliances
    circuit maps for, 142–143
    dishwashers, hooking up, 402–406
    energy efficient, 347, 418–419
    food disposers, installing, 407–411
    minimum trap sizes for, 80
  building islands
    stock-cabinet, 358–359
    tiled, 360–363
  codes & standards
    electrical, 135, 344
    heating, ventilation & air-conditioning, 345–346
    plumbing, 345
  designing
    general guidelines, 342–344
    layouts, 340–343
    overview of, 338–339
  dishwashers, hooking up, 402–406
  drains, hooking up sink, 398–401
  green, 347, 416–419
  installing backsplashes, 386–389
  installing cabinets
    base, 356–357
    hanging, 353–355
    overview of, 350
    preparing walls, 351–352
  installing countertops
    laminate, 368–375
    postform, 364–367
    tiled, 376–385
  installing faucets
    overview of, 392–393
    putting in new faucet, 395–397
    removing old faucet, 394
  installing food disposers, 407–411
  installing sinks, 390–391
  installing undercabinet lights, 412–415

L
Laminate flooring, installing, 166–169
Laminate panels, 210
Lap siding, making opening for exterior doors in, 32–33
Laundry rooms, highlights of National Electrical Code about, 135
Light fixtures
  circuit maps for, 139–141, 144–145
  described, 84
  highlights of National Electrical Code about, 134
  installing
    bathroom vanity, 484–485
    on ceilings, 324–329
    electrical boxes for, 113, 114
    track on ceilings, 330–333
    undercabinet in kitchens, 412–415
Lighting
  codes for natural, 18
  guidelines for kitchen, 344
Living rooms, highlights of National Electrical Code about, 135

M
Materials
  for applying veneer plaster, 224
  for framing with steel, 42
  for wood floors, refinishing, 204
  for working with conduit, 104
  for working with PEX pipes, 68
Mitered casings on doors & windows, installing, 319
Molding
  installing base, 242–243
  installing built-up base, 244–245
  installing crown, 250–253
Mosaic tile flooring, installing, 198–199

N
National Electrical Code highlights, 134–135
NM (Nonmetallic) cable
  about, 91, 92
  installing, 98–101
  reading, 93
  stripping sheathing & insulation, 94

O
One-coat veneer plaster system, applying, 224–226
Outdoors
  highlights of National Electrical Code about, 135
  installing ceramic tiles, 191–197
Outlets. *See* Receptacles

P
Paneling
  installing
    on ceilings, 228–231
    metal on ceilings, 239–241
    on walls, 211–213
  types, 210
Parquet flooring, installing, 170–173
Partition walls
  building, 40–42
  removing, 36–37
Patio doors, installing, 312–317
Perimeter-bond sheet vinyl, institution, 174–175, 177–179
Permits. *See* Building codes & permits
PEX pipes
  buying, 71
  connecting, 73
  installing, 70
  overview of, 68
  plumbing water-supply system with, 74–75
  system designs for, 72
  tools & materials for, 68
Picture rails, installing, 246–247
Pipes
  copper
    cutting, 59
    soldering, 56, 58, 60–62
    working with, 56–57
  PEX
    buying, 71
    connecting, 73
    installing, 70
    overview of, 68
    plumbing water-supply system with, 74–75
    system designs for, 72
    tools & materials for, 68
  plastic
    cutting, 65
    joining with solvent-glue, 66–67
    overview of, 64
  sizing for, 79
  supports for, 80
Plaster, applying veneer, 224–227
Plaster walls, removing, 31
Plastic pipes (ABS, PVC, CPVC)
  cutting, 65
  joining with solvent-glue, 66–67
  overview of, 64
Platform framing, elements of, 10
Plumbing
  codes & permits, 77–78, 423
  guidelines for kitchen, 345
  mapping system, 76
  overview of, 52–55
  symbols, 76
  water supply system with PEX pipes, 74–75
Plywood underlayment, installing, 151
Pocket doors, installing interior, 296–299
Prehung doors, installing interior, 294–295
PVC pipes. *See* Plastic pipes (ABS, PVC, CPVC)

R
Radiant floor mats
  bamboo flooring and, 186
  installing, 156–161
Raised underlayment panels, installing, 154–155
Ranges, circuit maps for, 143

- Receptacles
  - circuit maps for, 138–139, 142
  - highlights of National Electrical Code about, 134
  - installing
    - electrical boxes for, 112
    - GFCI, 127
    - new, 124–125
    - panels & wainscoting around, 213, 218
  - overview of, 85, 122–123
- Recessed ceiling lights, installing, 327–329
- Remodeling steps, 20–23
- Resilient flooring tiles, installing, 175, 176, 180–185
- Resilient sheet vinyl flooring, installing, 174–179
- Rigid copper pipes
  - cutting, 59
  - overview of, 56
- Rigid plastic pipes. *See* Plastic pipes (ABS, PVC, CPVC)
- Roofs
  - elements of, 15
  - estimating material needed, 531
  - shingling
    - installing drip edge, 537
    - installing three-tab shingles, 538–545
    - installing underlayment, 536
    - preparation, 532–533
    - removing old shingles, 534–535
- Room dimensions, codes for, 17
- Round-top windows, installing, 262–263

S

- Safety, 24–25
  - asbestos and, 24, 176
  - wiring, 88–89
- Self-adhesive resilient flooring tiles, installing, 175, 176, 180, 185
- Self-rimming sinks, installing, 390–391
- Service masts, described, 84
- Service panels
  - described, 85
  - highlights of National Electrical Code about, 134
  - new wiring and main, 128, 130
- Sheet vinyl flooring, installing, 174–179
- Showers, installing
  - enclosure kits, 430–435
  - hinged doors for, 436–437
- Siding
  - estimating material needed, 532
  - installing vinyl, 548–553
    - overview of, 546
    - tools for, 547
  - making opening for exterior doors in lap, 32–33
- Single-pole wall switches
  - circuit maps for, 140–141
  - installing, 117
  - overview of, 117
  - testing, 120
- Sinks
  - hooking up drains, 398–401
  - installing
    - bathroom pedestal, 468–469
    - bathroom vessel, 464–467
    - self-rimming, 390–391
- Skylights
  - building shafts, 284–287
  - installing, 278–283
    - tubular, 288–289
- Sliding patio doors, installing, 312–317
- Small appliances, circuit maps for, 142
- Soft copper pipes. *See* Flexible copper pipes
- Soldered joints, taking apart, 63
- Soldering copper pipes, 56, 58, 60–62
- Solvent-glue, joining plastic pipes with, 66–67
- Split-circuit receptacles, 122, 123
- Stairs, highlights of National Electrical Code about, 135
- Steel, framing with, 42–43
- Storm doors, installing, 310–311
- Stucco, making opening for exterior doors in, 34–35
- Studs, joining sections using steel, 43
- Supply tubes, sizing for, 79
- Suspended ceilings, installing, 232–235
- Switches
  - four-way
    - circuit maps for, 145
    - installing of, 119
    - overview of, 119
    - testing, 121
  - highlights of National Electrical Code about, 134
  - installing electrical boxes for, 113
  - installing panels & wainscoting around, 213, 218
  - overview of, 85, 116
  - single-pole wall
    - circuit maps for, 140–141
    - installing, 117
    - overview of, 117
    - testing, 120
  - testing, 120–121
  - three-way
    - circuit maps for, 144
    - installing, 118
    - overview of, 118
    - testing, 121
- Symbols
  - electrical, 137
  - plumbing, 76

T

- Thermostats, circuit maps for, 143
- Three-way switches
  - circuit maps for, 144
  - installing, 118
  - overview of, 118
  - testing, 121
- Tileboard panels, 210
- Tiles
  - building tiled kitchen islands, 360–363
  - installing
    - acoustical ceiling, 236–238
    - backsplashes, 386–389
    - on bathroom walls, 454–457
    - flooring
      - ceramic, 191–197
      - mosaic, 198–199
    - metal ceiling, 239–241
    - tiled countertops, 376–385
- Toilets, installing, 458–461
- Tongue-and groove hardwood flooring, installing, 162–165
- Tools
  - for applying veneer plaster, 224
  - for framing with steel, 42
  - installing vinyl siding, 547
  - for installing wallboard, 45
  - for resilient floors, 175
  - safety and, 25
  - stilts, 47
  - for wood floors, refinishing, 204
  - for working with conduit, 104
  - for working with PEX pipes, 68
- Towel warmers, installing, 490–491
- Track lighting, installing on ceilings, 330–333
- Trim, removing, 38
- Tubs, installing
  - alcove, 438–445
  - sliding doors, 450–453
  - spouts, 480–483
  - 3-piece surrounds, 446–449
- Two-coat veneer plaster system, applying basecoat in, 224, 226
- Two-slot receptacles, 122, 123

U

- Undercabinet kitchen lights, installing, 412–415
- Underlayment for floors
    - installing
      - cementboard, 152
      - isolation membrane, 153
      - panels, 154

    - plywood, 151
    - raised panels, 154–155
  - overview of, 150
- Underlayment for roofs, installing, 536
- Universal design
  - in bathrooms, 424–429, 486–487
  - in kitchens, 348–349
  - overview of, 19
- Utility rooms, highlights of National Electrical Code about, 135

V

- Valve requirements, 79
- Vanities
  - installing lights for, 484–485
  - installing wall-hung, 462–463
- Veneer plaster, applying, 224–227
- Ventilation
  - codes for, 18
  - guidelines for kitchen, 345–346
- Vent system. *See* Drain-waste-vent system
- Vinyl siding, installing, 546–553
- Voltage ratings, 83

W

- Wainscoting, installing, 214–219
- Wallboard
  - cutting, 46
  - installing
    - at corners, 49, 51
    - finishing & taping, 50–51
    - on flat ceilings, 47
    - overview of, 44
    - tools & prep work for, 45
    - on walls, 48
  - removing, 30
- Wall-hung vanities, installing, 462–463
- Walls
  - applying veneer plaster, 224–227
  - building
  - glassblock, 220–223
    - partition, 40–42
  - cabinet installation preparation of, 351–352
  - elements of, 15
  - installing
    - attic, 514–515
    - basement, 510–513
    - base molding, 242–245
    - beadboard wainscoting, 214–219
    - chair rails, 248–249
    - crown molding, 250–253
    - picture rails, 246–247
    - wallboard on, 48
    - wires on concrete, 106–107
    - wood paneling, 210, 211–213
  - removing partition, 36–37
  - tiling bathroom, 454–457
- Waste system. *See* Drain-waste-vent system
- Water supply system
  - copper pipes, 56–57
  - overview of, 54
  - plumbing with PEX pipes, 74–75
  - sizing for distribution pipes, 79
- Windows
  - adding basement egress, 17, 518–523
  - framing, 12–13, 256–259
  - guidelines for kitchen, 344
  - installing
    - bay, 268–277
    - casings, 318–321
    - garden, 264–267
    - paneling around, 213
    - replacements with masonry clips, 261
    - replacements with nailing flange, 259–261
    - round-top, 262–263
    - wainscoting around, 219
- Wires/wiring
  - connecting
    - with push-ins, 95
    - to screw terminals, 95
    - using wire connector, 96
  - connector types, 93
  - grounding
    - described, 84
    - installing through metal conduits, 102
  - installing on concrete walls, 106–107
  - new
    - drawing plans, 136–137
    - examining main service panel, 130
    - National Electrical Code
      - basic requirements, 131–133
      - highlights, 134–135
    - planning steps, 128–129
  - overview of, 82, 90–92
  - pigtailing, 97
  - reading unsheathed individual, 93
  - safety, 88–89
- Wood floors, refinishing, 204–207

Also From **CREATIVE PUBLISHING international**

Complete Guide to A Green Home
Complete Guide to Attics & Basements
Complete Guide to Basic Woodworking
Complete Guide Build Your Kids a Treehouse
Complete Guide to Contemporary Sheds
Complete Guide to Creative Landscapes
Complete Guide to Custom Shelves & Built-Ins
Complete Guide to Decorating with Ceramic Tile
Complete Guide to Decks
Complete Guide to DIY Projects for Luxurious Living
Complete Guide to Dream Bathrooms
Complete Guide to Dream Kitchens
Complete Guide to Finishing Walls & Ceilings
Complete Guide to Floor Décor
Complete Guide to Gazebos & Arbors
Complete Guide to Home Carpentry
Complete Guide to Landscape Construction
Complete Guide Maintain Your Pool & Spa
Complete Guide to Masonry & Stonework
Complete Guide to Outdoor Wood Projects
Complete Guide to Painting & Decorating
Complete Guide to Patios
Complete Guide to Plumbing
Complete Guide to Roofing & Siding
Complete Guide to Trim & Finish Carpentry
Complete Guide to Windows & Entryways
Complete Guide to Wiring
Complete Guide to Wood Storage Projects
Complete Guide to Yard & Garden Features
Complete Outdoor Builder
Complete Photo Guide to Home Repair
Complete Photo Guide to Home Improvement
Complete Photo Guide to Homeowner Basics

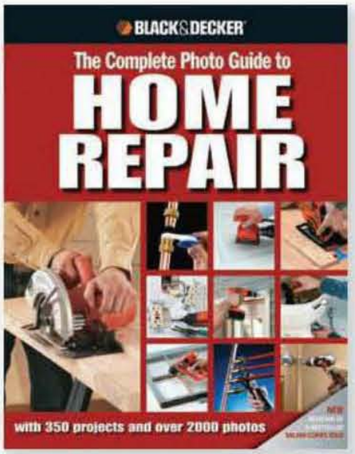
BLACK&DECKER
The Complete Photo Guide to
HOME REPAIR
with 350 projects and over 2000 photos

ISBN 1-58923-417-0

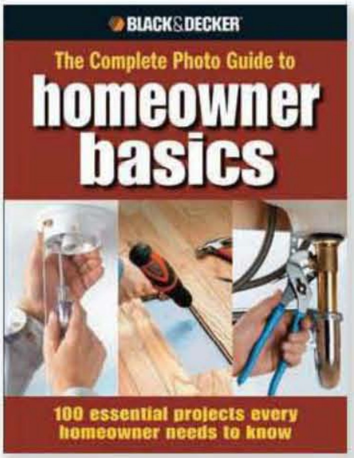
BLACK&DECKER
The Complete Photo Guide to
homeowner basics
100 essential projects every homeowner needs to know

ISBN 1-58923-376-X

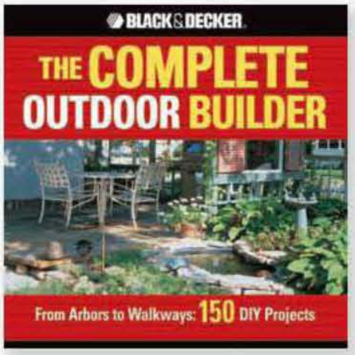
BLACK&DECKER
THE COMPLETE OUTDOOR BUILDER
From Arbors to Walkways: 150 DIY Projects

ISBN 1-58923-264-X

Creative Publishing
international

400 First Avenue North • Suite 300 • Minneapolis, MN 55401 • www.creativepub.com

Rainwater Harvesting 101

RAINWATER HARVESTING 101

GROWNYC

Robert J. Kafin, Esq., Chairman

Marcel Van Ooyen, Executive Director

Prepared by:
Julia Leung, DEC Summer Intern
for
GrowNYC
51 Chambers Street, Room 228
New York, NY 10007
(212) 788-7900

August 2008

PREFACE

GrowNYC is a hands-on non-profit that has been improving New York City's environment for over thirty years. GrowNYC's dedicated staff green our neighborhoods, create the environmental leaders of the future, promote waste prevention and recycling, and run the largest farmers market program in the country. GrowNYC achieves its mission through the following projects and programs: The Open Space Greening Program, Greenmarket, The New Farmer Development Project, Environmental Education Training Student Organizers, Learn It, Grow It, Eat It and Office of Recycling Outreach and Education.

The Open Space Greening Program (OSG) empowers people in neighborhoods throughout the city to create, manage and sustain community gardens and park/playgrounds. OSG provides best practices workshops, services, tools, donated plant material, and open space planning/mapping information and other services. Grow Truck provides tools, donated supplies, plants and horticultural advice and assistance to gardening groups all over New York City. The Plant-A-Lot Project gives substantial material and technical assistance to several new gardens each year and helps the 45 existing gardens created in prior years.

Since 2002, OSG staff has taken the lead in building and maintaining rainwater harvesting systems in community gardens across the five boroughs, leading educational workshops, assisting community gardeners in identifying alternate water resources and building low cost systems. GrowNYC is an active member of the Water Resources Group, a network of NYC greening and environmental organizations promoting sustainable water conservation practices in NYC. The New York State Department of Environmental Conservation (DEC) recognized GrowNYC's efforts with The Environmental Excellence Award in 2006. This manual was created to disseminate to a wide audience the design parameters and building techniques used by GrowNYC staff.

Summer intern was able to prepare this document thanks to a grant from The New York State Department of Environmental Conservation in 2008 for Pollution Prevention. Ms. Leung provided many of the photos, charts, and drawings used. OSG staff, Lars Chellberg, and Lenny Librizzi provided invaluable guidance, advice, and editing of the manual. GrowNYC Executive Director Marcel Van Ooyen and Assistant Executive Director Julie Walsh edited the text. Additional information and assistance was provided by NYC community gardeners and members of the Water Resources Group.

Table of Contents

Introduction of Rain Water Harvesting 2
- History of Rain Water Harvesting 5
- Benefits of Rain Water Harvesting in Urban Areas 6
- Current Efforts 7

Materials 10
- Where to obtain supplies 11
- Cost Estimates 12
- PVC Applications 14
- Sizing of the tank and roof washer 18

Construction 21
- Tank 23
- Leaders 26
- Roof Washer 30
- Overflow 33
- Anchoring 36
- Drip Irrigation 39
- Filtration Systems 41
- Treatment 42
- Aesthetics 43
- Roofing Material 45

Seasonal Maintenance 46

Living with the RWH system 47

Existing RWH Systems 48

10 Ways to Conserve Water 49

Bibliography 50

Feedback Form 51

Introduction

"It isn't easy to come up with 'one size fits all' instructions for building rainwater harvesting systems because of variations in styles of roofs, downspouts, storage tanks, and garden layouts. You have to use a combination of research, common sense, ingenuity, and dumb luck to design and build your system."

-Lenny Librizzi, Assistant Director of Open Space Greening at GrowNYC

Rain water harvesting (RWH) is the means of collecting and storing rain water in large, durable containers, usually, collecting from rooftop gutters. RWH systems come in a variety of shapes and sizes (see pictures below).

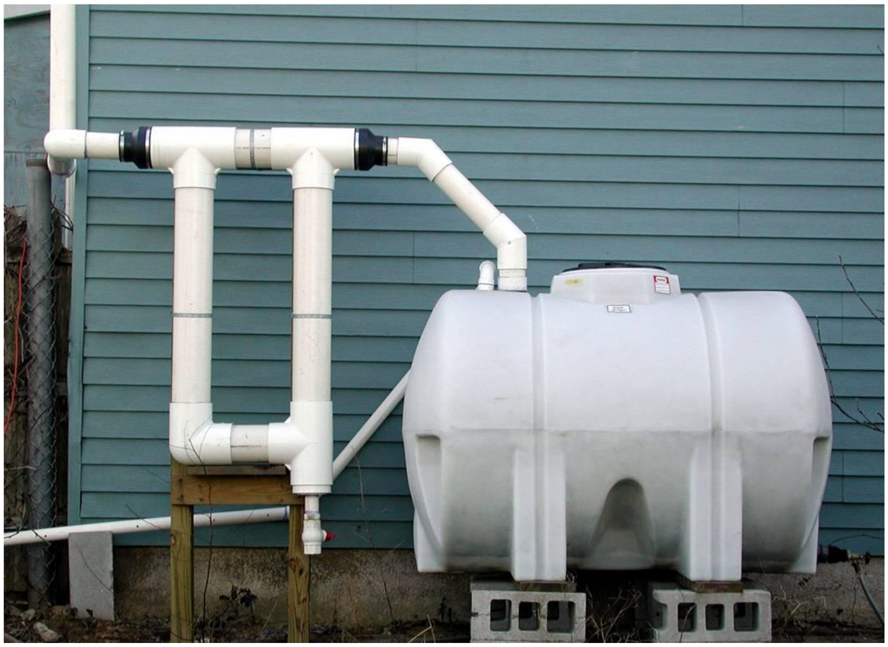

Rain water harvesting systems are fairly easy to construct. Tanks in NYC community gardens range in size from 300 to 1000 gallons but can be as small as 55 gallons and as big as 10,000 gallons. Rain water collected from the downspout of an adjacent building or shed in the garden is redirected to the water storage area in the garden. The RWH system includes 3 parts (see drawing): the tank **(1),** the first flush **(2)** and the overflow pipe **(3).**

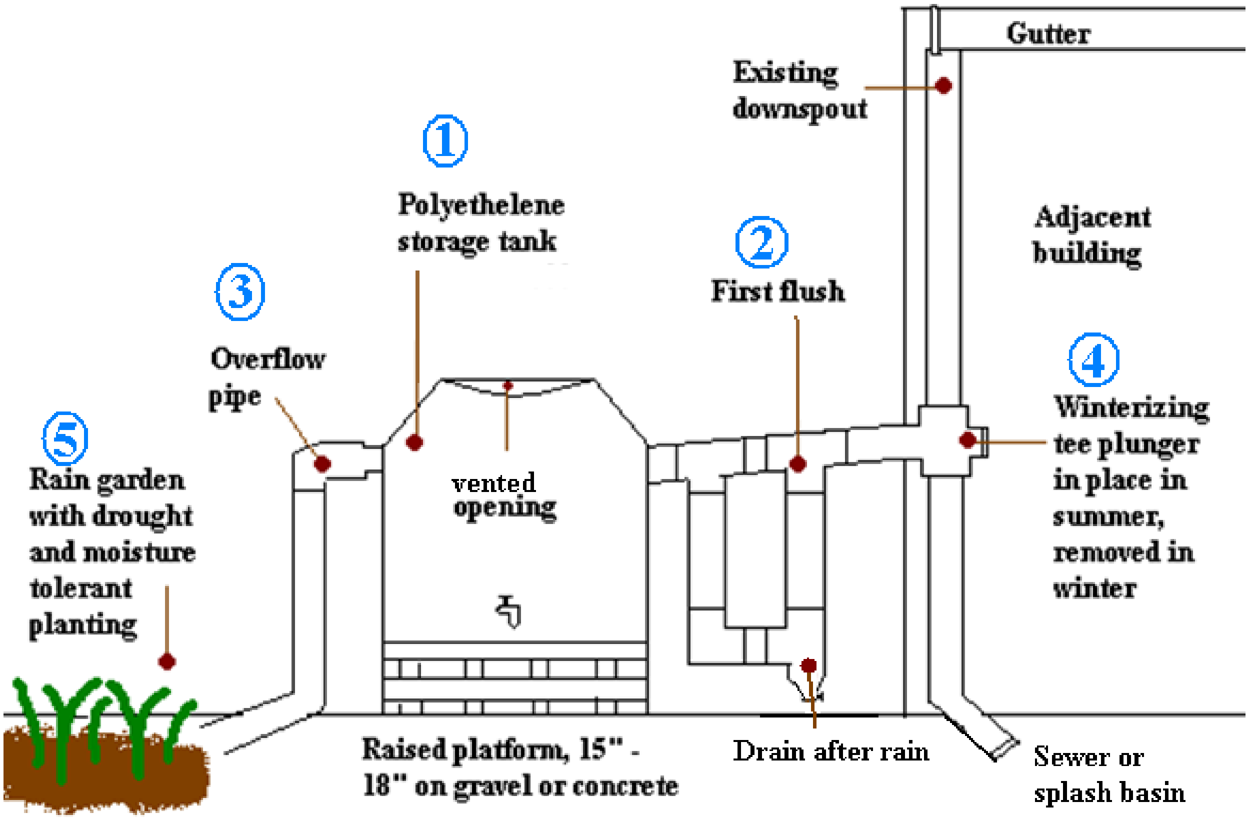
Gutter
Existing downspout
①
Polyethelene storage tank
②
First flush
Adjacent building
③
Overflow pipe
④
Winterizing tee plunger in place in summer, removed in winter
⑤
Rain garden with drought and moisture tolerant planting
vented opening
Raised platform, 15" - 18" on gravel or concrete
Drain after rain
Sewer or splash basin

During a rainfall event water from the gutter flows into the downspout. Instead of the water going into the sewer system, the rainwater harvesting system diverts the water into pipes. This diverter consists of a 3 way tee with a plunger in place during the summer. This plunger keeps the water from entering the downspout and forces it to flow into the harvesting system. It is taken out in the winter when rainwater is no longer collected **(4).**

Here is a closer view of how the plunger works.

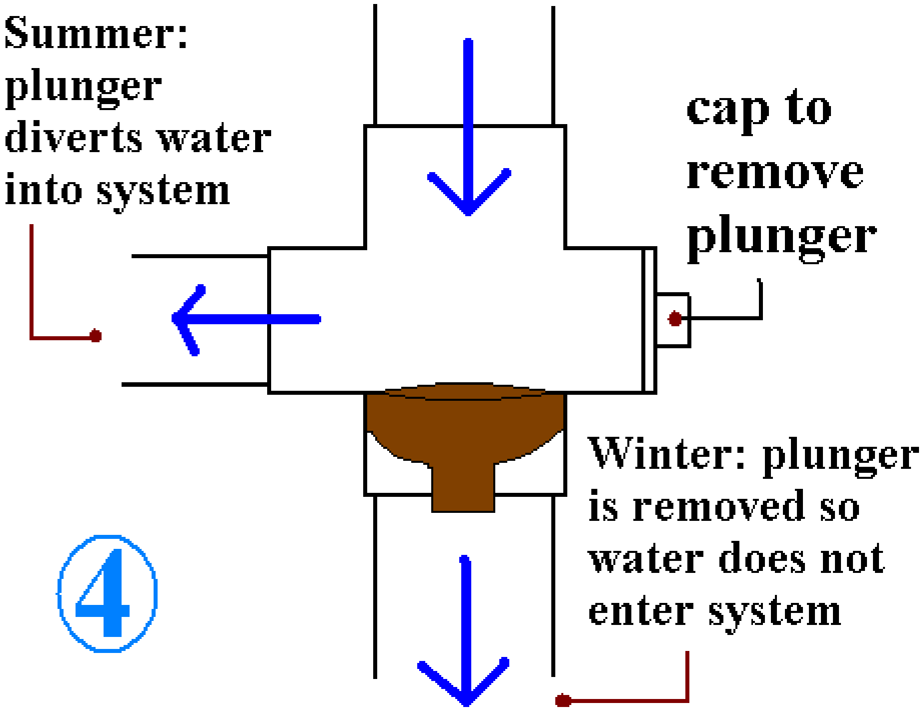
Summer: plunger diverts water into system
cap to remove plunger
Winter: plunger is removed so water does not enter system
④

The pipes lead to a roof washer system which is a containment area for the first few gallons of water. Since the initial flushes of water contain rooftop debris and leaves, the roof washer acts as a filtering system by separating the dirty water from the cleaner water. Once the roof washer is full, the cleaner water enters the rain tank. As soon as the tank is full, excess water flows into the overflow pipe which leads to an adjacent rain garden **(5),** is directed back to where it originally flowed or piped underground. A rain garden is a plot containing hardy plants that can survive with both saturated and dry soil.

History of rainwater harvesting

Rainwater harvesting is an ancient practice. Many different cultures have used this technology for agricultural purposes. The Philippines have been using rain water for rice terraces for thousands of years now. Indian history indicates that rain water systems have been in use since 3000 BC. Usage can also be traced back 2000 years ago in Thailand and other parts of Southeast Asia, where simple gutters were used to fill jars and pots. The earth dams of ancient Egypt were used to control runoff. Ancient rain water cisterns can still be seen on the islands of Capri and Malta.

In addition to rainwater harvesting, the Romans also used their systems as air conditioners. When the water evaporated, it created a cooling effect in the microclimate. As their populations started to increase, the Romans developed underground cisterns. This way, less water would be lost due to evaporation. They connected these cisterns to above ground pools as a means of water filtration. When these pools overflowed, the cleaner water would enter the cisterns. This design was an inspiration for modern day rain barrels. The Roman's shallow pool mirrors the modern day roof washer or first flush system.

The world's largest cistern is the Yerebatan Sarayi, built by Caesar Justinian in 527 AD. It is located in modern day Turkey and is a popular tourist attraction. It is 140 meters by 70 meters big and can store up to 80,000 meters^3 of water. This huge structure is completely underground and involves a series of intersecting vaults. Binbirdik is another cistern in Istanbul, created by Caesar Constantine in 389 AD that can store 80,000 meters^3. However construction of these cisterns stopped due to the difficulty of building underground and the outbreak of human fecal contamination in large cities. [1]

In the 1970's, a new technology of rain water harvesting was developed where storm water was stored in well storage tanks in the form of different sized ponds. A thin layer of red clay was used to line the bottom of the pond to prevent seepage and trees were planted around the pond to prevent excessive evaporation. Over 40,000 of these systems were built in the Loess Plateaus of China. [2]

Although rain water harvesting was a significant and successful design in the past, its popularity has declined over the centuries. Urbanization demanded a more centralized water supply system. Watersheds and pipelines came into use and running water became one of the world's greatest inventions. However, due to modern day water pollution and drought, rainwater harvesting techniques have come back into practice again.

Benefits of rainwater harvesting in urban areas

By keeping storm water out of the sewage system, gardeners help keep their local water bodies clean. Because most of NYC is paved over, rainwater can't be absorbed by the ground. Instead, it runs across the pavement, picking up oil, street debris, animal feces and other waste as it moves. This runoff then enters the sewer system.

Unfortunately, not all runoff makes it to a treatment plant before entering the city's rivers and harbor. This is because most of NYC is on a combined sewer system, whereby the same pipes are used to transport both storm water runoff and household waste to sewage treatment plants. If these treatment plants overflow, the contaminated water then enters the water bodies surrounding NYC.

During dry weather the sewage treatment system generally works well, but these treatment plants can only handle about twice their dry weather volume. This means that heavy rains often result in combined sewer overflows (CSOs), which release untreated wastewater, storm water and street debris into local waterways. Untreated sewage can carry disease causing pathogens and nutrient rich organic material, which can choke the Harbor's ecosystem.

Not only can rainwater harvesting prevent water pollution, it also conserves water. Typical systems can store up to 1000 gallons of water. Harvesting during rainy days in spring and summer provides a source of water for the dry spells between rain events. Instead of running the hose to water your garden, you can use the rainwater stored in your tank. This will relieve the strain on our reservoirs and also save you money on your water bill. Forty percent of household water consumption is used on domestic irrigation! [3]

Rainwater harvesting has proven to be successful. The 35 rain water harvesting systems working in NYC's community gardens collect 422,900 gallons of water every year. That's enough for 264,313 toilet flushes!

Of course small scale construction of rainwater systems is not enough to significantly impact our environment. However, if rain barrels were used more extensively and rain gardens became a larger part of city landscape planning, water conservation would have a greater environmental impact. We would experience cleaner water bodies, less flooding and lower water costs.

Rain water harvesting has great potential to reduce rainwater flow into storm drains by collecting rainwater in gardens and can reduce the dependence on the NYC watershed. Every drop of water counts and conservation is the most sustainable, cost-effective source of water supply for our region.

Current efforts

Community gardeners across North America and around the world have revived the practice of using rain barrels in their gardens out of necessity.

In **Seattle**, budget constraints and rising water prices spawned the construction of rainwater harvesting systems. Some of Seattle's rainwater harvesting systems are works of art. Here gate pillars support the flow of rainwater from the building on the right into the cistern on the left. Seattle Public Utilities have recommended changes to the land use and building codes that encourage water conservation. SPU has also sponsored some wildly successful sales of rain barrels to Seattle residents.

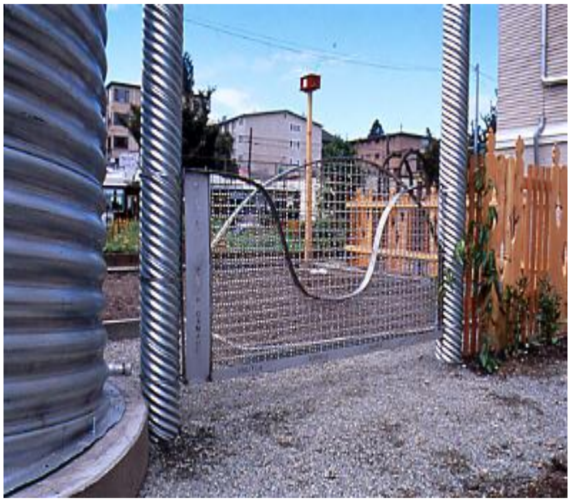

The special curves on this gate represent the monthly rain fall in Seattle and the amount of water collected year round.

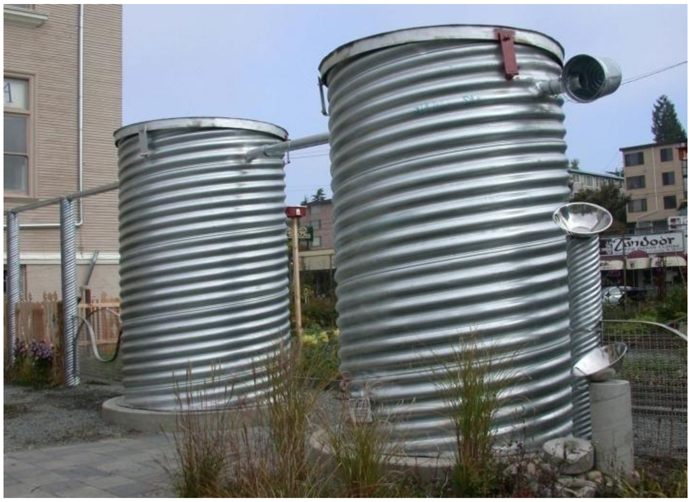

The overflow pipe of this Seattle cistern creates a mini waterfall.

A cat enjoying the rain garden in Kansas City.

The 10,000 Rain Gardens Program in **Kansas City** is an example of regional efforts to educate and introduce water conservation to the Kansas City community. Rain gardens are areas of vegetation on porous ground that can filter and drain excess storm water. Kansas City officials encourage their residents to plant these gardens on their own property.

The City of **Vancouver** designs and manufactures rain barrels for residents to use for irrigation (see picture on right). Vancouver subsidizes the cost of the rain barrels by 50%. Over 2000 barrels have already been sold.

Portland, Oregon, granted a permit for a household to harvest rainwater for use indoors during all but the dry summer months (see sketch and picture below). The water undergoes enough filtering to meet EPA's standards for drinking water. In 1998 this system cost less than $1,500 to install.

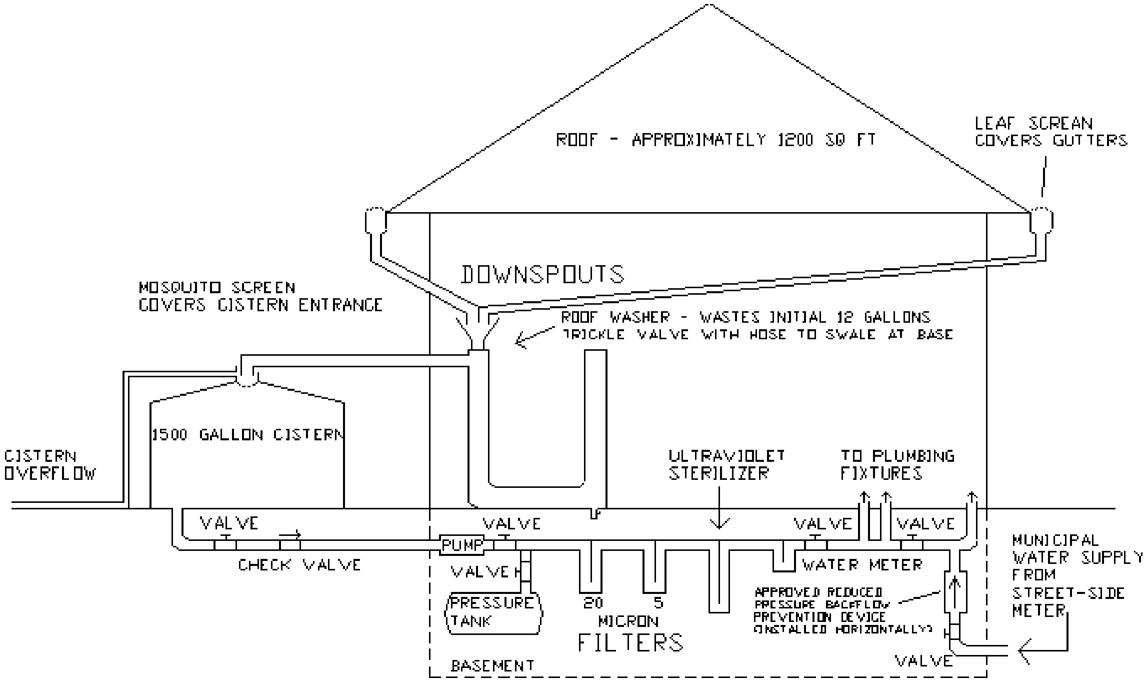
ROOF - APPROXIMATELY 1200 SQ FT
LEAF SCREAN COVERS GUTTERS
DOWNSPOUTS
MOSQUITO SCREEN COVERS CISTERN ENTRANCE
ROOF WASHER - WASTES INITIAL 12 GALLONS TRICKLE VALVE WITH HOSE TO SWALE AT BASE
1500 GALLON CISTERN
CISTERN OVERFLOW
ULTRAVIOLET STERILIZER
TO PLUMBING FIXTURES
VALVE
CHECK VALVE
VALVE
PUMP
VALVE
PRESSURE TANK
20
5
MICRON
FILTERS
VALVE
VALVE
WATER METER
APPROVED REDUCED PRESSURE BACKFLOW PREVENTION DEVICE (INSTALLED HORIZONTALLY)
MUNICIPAL WATER SUPPLY FROM STREET-SIDE METER
VALVE
BASEMENT

Maplewood, Minnesota has forged a coalition between its water and sanitation departments and residential landowners to plant large-scale curbside rain gardens that reduce storm-water sewage (see picture on left).

The Minnesota Arboretum uses its parking lots to demonstrate best landscape practices such as planting rain gardens and using permeable pavers to reduce contamination caused by storm water runoff (see picture on right).

grass absorbing areas

The City of **Toronto** Downspout Disconnection Program offers a free service to homeowners to disconnect downspouts from the sewer system and install rain barrels which are available at a discount.

Chicago also encourages its citizens to disconnect downspouts. In addition, Chicago's Water Agenda 2003 included a rooftop garden initiative, a pilot program for permeable alleys, and rain gardens planted in the City's rights-of-way.

Austin offers rebates of up to $30 for newly installed rain barrels and of up to $500 for installation of rainwater harvesting systems, following design approval by the City.

In 2001, the first Drought Emergency to be announced since 1989 was declared in **New York City**. The reservoir levels were 40% below normal which resulted in mandatory water use restrictions. The Water Resources Group (WRG) was founded as a response to this single drought year. Since then WRG and GrowNYC have built 35 rainwater harvesting systems in New York City, collecting over 422, 900 gallons of storm water per year.

"New York City currently lags behind other cities like Boston and Chicago, which capture 90% of their combined sewage overflow; New York City only captures about 70%, according to the city"[4]

Materials

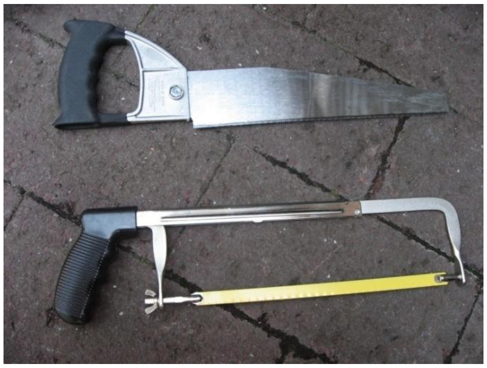

Every rain water harvesting system is different but they do have similarities. Here is a comprehensive list of tools and supplies. Your system will use many of these but may not require all of them.

Tools:

- Drill
- Hole saw attachment or jigsaw
- Screw drivers
- Hammer
- Level
- PVC saw (see top saw in picture) and metal hacksaw (bottom saw)
- Tin snips or sheet metal shears
- Crimping tool (see picture on right) and cable cutting tools

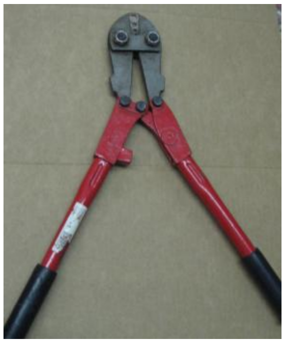

Gutters:

- Gutter lengths
- Leaders and bends
- Pre-fabricated gutter hangers
- Plumbing strap (for securing pipes to wall)
- Flexible, accordion style expandable plastic pipe
- Tube of gutter sealant and caulk gun to apply it
- PVC cleaner and cement
- Teflon tape (for creating a water tight seal on threaded bushings)

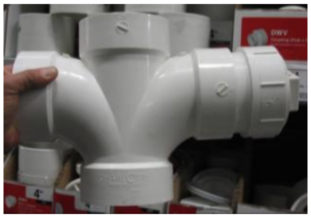

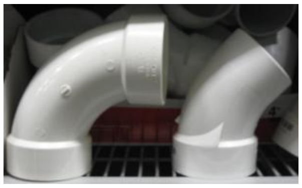

Barrel storage system:

- Tanks and PVC parts or other materials to construct the manifold. Here we have a capped 4 way tee (see leftmost picture) and a 90° and 45° elbow respectively (see rightmost picture)
- Window screen or screened vents for mosquito proof vent
- Bulk head fittings (An elongated compression fitting, which will allow pipe, to run through a bulkhead)
- Platform material such as treated lumber, plastic lumber or concrete blocks
- Eyebolts (for securing tank)
- Spigot
- Metal flanges of corresponding size to spigot and hardware
- Rope or cabling supplies to secure the tank

Where to obtain supplies

Polyethylene tanks are best purchased locally because of high shipping costs. Search online for local suppliers. You can personally ship smaller tanks and save on extra expenses. Here are some websites that sell tanks:

- http://www.liquidlogictanks.com/index.html
- http://www.tank-depot.com/
- http://www.rainbarrelsource.com/
- http://www.rainbarrelsandmore.com/
- http://www.aquabarrel.com/

Drip irrigation suppliers include:

- http://www.dripdepot.com/
- http://www.dripirrigation.com/
- http://www.rainbird.com/drip/index.htm
- http://www.dripworksusa.com/
- http://www.netafim.com/
- http://www.chapindrip.com/
- http://www.farmtek.com

Downspout filters and a variety of other specialized parts for systems can be purchased from:

- http://www.starkenvironmental.com/a-1-filtration.html
- http://www.braewater.com/
- http://www.rainharvesting.com.au/default.asp
- http://downspoutfilter.com/index.php
- http://rainharvest.com/shop/default.asp

Tools and PVC gutter pieces can be purchased at your local hardware stores. Lowes or Home Depot also sells these items. For locations near you, visit their websites at http://www.lowes.com/ or http://www.homedepot.com.

Cost Estimates

Note that the prices listed indicate the values in 2008. Adjust for a 15% – 20% annual increase.

Polyethylene tanks: Generally, tanks cost $1 per gallon but it can sometimes be too big to transport on your own. Shipping costs are about an additional $1 per gallon.

Leaders and Gutters: The number of leaders and gutters needed per site will vary depending on the roof configuration. A 10-foot long gutter is $15.00, a 10 foot long 2”x 3” leader is $9.00, and a 10 foot long 3”x 4” leader is $16.00. Connecting parts, corners and bends are anywhere from $2.00 to $10.00.

Platforms: Platforms for containers can be made out of cinder blocks, plywood or any other sturdy material. While the tank is light when it is empty, remember that a full tank of water can weigh thousands of pounds. A small 55 gallon tank weighs over 400 pounds when full! Be sure that your platform is strong enough to withstand large weights. Cement blocks can be scavenged from construction sites. Should you have to purchase them, 8” x16” cement blocks are $ 1.50 each. The lumber prices are $16.00 for a pressure treated 8 foot long 2”x 10” piece or $13.00 for a pine or fir 8-foot long 2”x 10” piece. An 8 foot long 4” x 4” piece costs $12.00. 4” x 4” lumber laid lengthwise in alternating directions placed on a gravel base makes a sturdy platform (see picture on page 25). Using poured concrete pilings and a built wooden deck for a platform may cost about $500.

Miscellaneous Hardware and specialized parts: This category includes accordion connector pieces, screens, sealants, sheet metal, screws, nuts, bolts, gravel, filtering parts and supplies, roof washers, weed fabric and other specialized parts. These parts will vary by site and type of barrel.

Plumbing Supplies: The amount of plumbing you need varies depending on the system you are making. This category includes overflow pieces, PVC pipes, connector pieces, rubber fittings, spigots, bulkhead fittings and the piping needed to draw the water away from the system. Upgrades will include soaker hoses or other type of overflow disperser and roof washers. A sophisticated drip irrigation system could add $500 – 1000 to the cost.

Tool Kit: Some specialty tools are needed such as drill bits, screw bits, hacksaw blades, tin snips, hole saws, caulk gun, caulking or silicon, jig saw, pliers and tape measures.

Cost analysis for a 300 gallon rainwater system in 2008

Tank	$300
Shipping of Tank	$300
Leaders and Gutters	$100
Platform	$300
Hardware	$100
Plumbing	$100
Tool Kit	$50
Total Cost	$1250

Cost analysis for a 1000 gallon rainwater system in 2008

Tank	$1000
Shipping of Tank	$1000
Leaders and Gutters	$200
Platform	$600
Hardware	$200
Plumbing	$200
Tool Kit	$50
Total Cost	$3250

If there is not already an existing roof or shed nearby, a shade structure can be constructed. One possible design is the winged structure seen here:

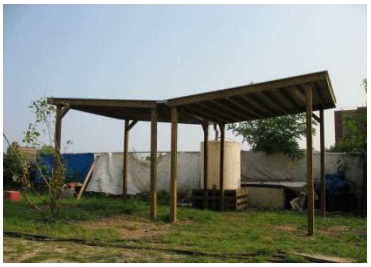

Cost analysis for a 300 gallon rainwater system with winged structure in 2008

Tank	$300
Shipping of Tank	$300
Leaders and Gutters	$100
Platform	$300
Hardware	$100
Plumbing	$100
Tool Kit	$50
Materials for Shade Structure	$2500
Total Cost	$3750

PVC Applications

Up to now PVC or polyvinyl chloride has been the key piping material in the rainwater harvesting systems in NYC community gardens. This hard plastic is inexpensive, durable, easy to use and readily available. Tees, elbows, bushings and couplings are all made out of PVC. Although PVC is fairly stable, there are environmental hazards in manufacturing and burning PVC. We are currently looking for other sources that are more environmentally benign but also having the same flexibility with parts as PVC does. Although metal can be an alternative, it can be costly. If there is a material that you think can be a replacement for PVC, please add your information or comment to this wiki: http://www.waterresourcesgroup.org/wiki/index.php?title=Main_Page.

In order to join PVC pipes together, use PVC specific cleaner primer and cement. PVC cleaner primer removes dirt and melts the surface for ultimate adhesion. After the pipe is clean, you can use cement to join the two pieces together. Only apply cleaner and cement to the outside part of the male piece (see picture on left) and the inside part of the female piece (see picture on right). It is important to hold the two pieces together for 20 seconds after applying the cement because they may move apart. After 20 seconds, the two pieces are permanently joined (see bottommost picture). Wipe off excess glue.

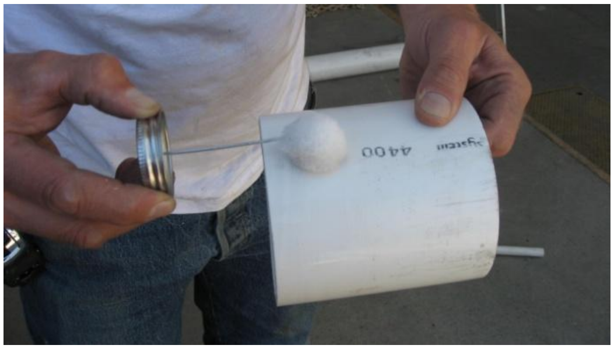

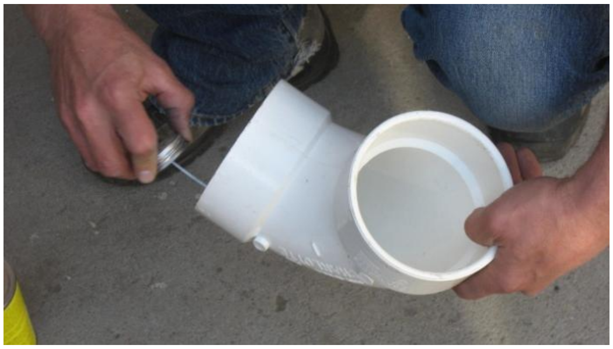

Use bushings and reducer fittings when it is necessary to change the size of your pipes. These come in a wide variety of glue (slip) and threaded designs.

This 2”x 1 ¼” slip bushing connects a 2” (diameter) piece of pipe with a 1 ¼ “pipe. Slip bushings can be adhered using the cleaner and cement method mentioned above.

This 2”x 1” slip and thread bushing connects a 2” slip pipe with a 1” threaded pipe. Thread bushings can be adhered tightly using Teflon tape. Roll the tape starting at the end of the male thread bushing towards the middle. Two or three wraps are sufficient.

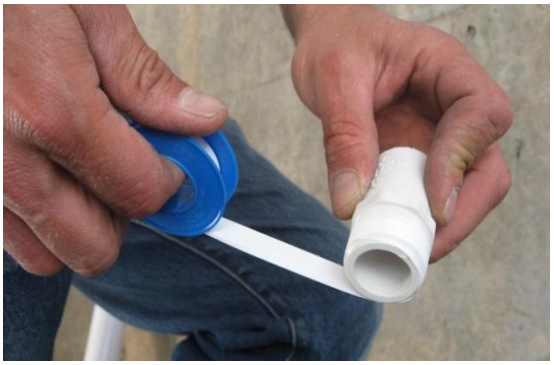

PVC parts are very versatile since no two rainwater harvesting systems are alike. Many different combinations of PVC parts can be used for the same function. Here are 5 different ways to connect a 2” pipe with a ¾ “pipe.

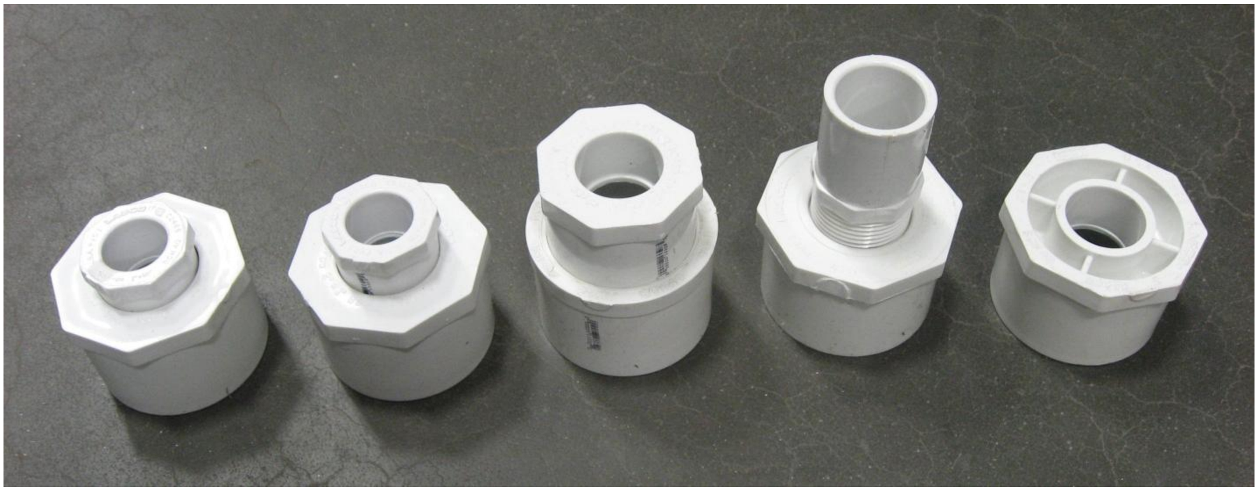

It is very important to align adjacent pipes. The two tees in the first picture (see picture below) must align exactly in order to connect to the same wall. Careless glueing will lead to problems. Use a hard, level surface such as a concrete floor to help you align pipes.

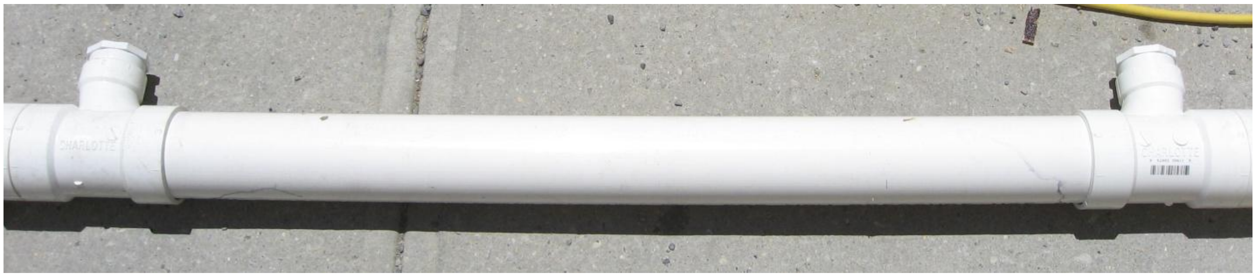

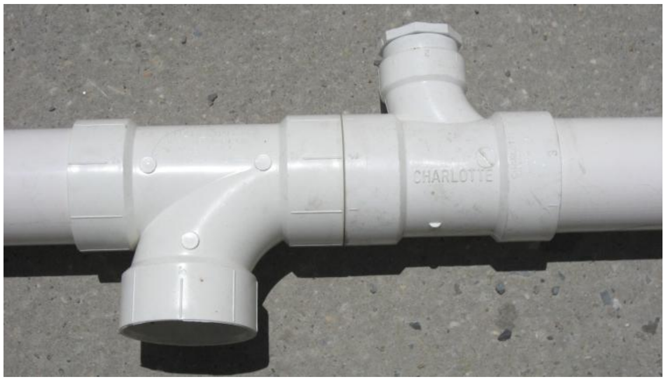

The pipes in the second picture (see picture on left) must be exactly perpendicular to each other. The tee on the left will connect to the barrel. The tee on the right will connect to existing pipes. Use a marker to indicate the points where the one pipe must meet the other.

The pictures above also point out the use of piping. Notice how the two tees are joined in each picture. In the second picture, the two tees are joined by using a small piece of pipe in between them, called a sleeve. The interior coupling is too short to be seen in this picture. In a situation where you want the tees to be further apart, such as in the first picture, just extend the size of the pipe to the desired length.

Although every rainwater system must be carefully planned before actual construction begins, it is very easy to miscalculate or measure incorrectly. Often times, errors are not realized at first. To fix a measurement or connection problem where the 2 pipes that were intended to connect do not match up, use two 45° degree elbows. A combination of two 45° elbows will produce almost any twist or angle desired.

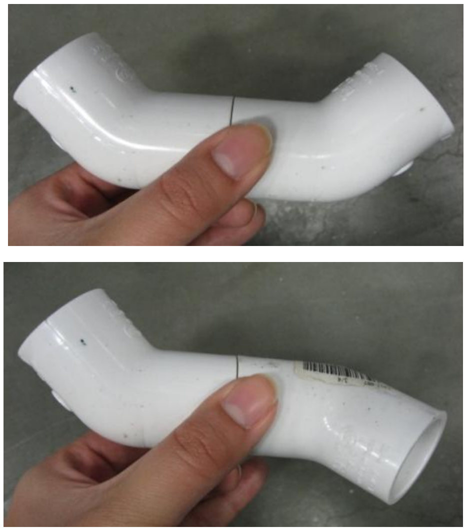

Two 45° elbows can make a 90° degree twist or a 180° degree twist.

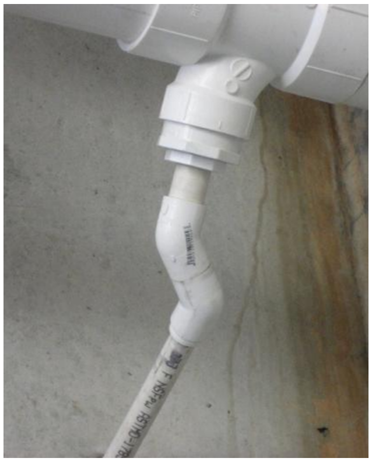

Two pipes that did not connect before can now be connected with the help of two 45° elbows.

It is also possible to shorten the PVC pipe by cutting and reconnecting it with a rubber coupling. Notice that the pipe below is too long to connect with the smaller pipes attached to the wall (see top picture). It was shortened and reconnected using a rubber coupling (see bottom picture).

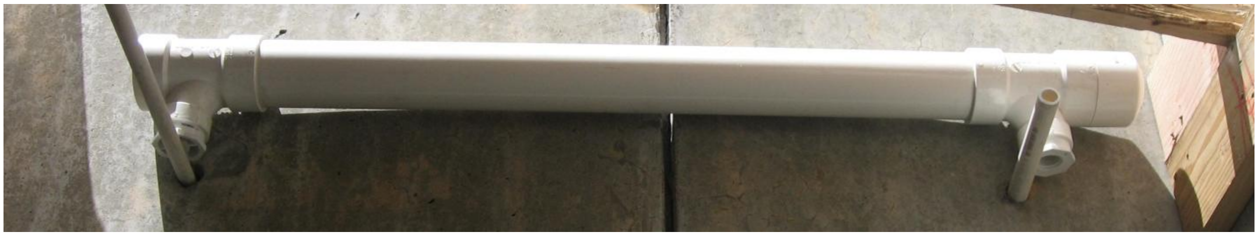

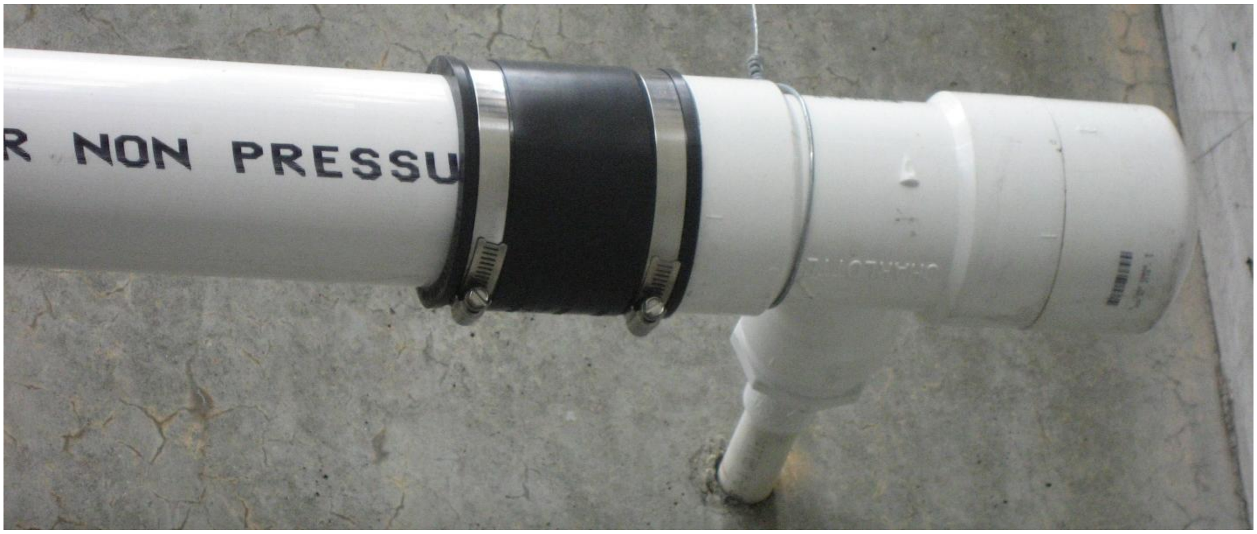

Sizing of the tank and roof washer

A variety of rainwater systems can be constructed for the same downspout and roof. Here are two different systems that have been installed over the years at the 1100 Block Bergen Garden in Brooklyn. It is generally better to have a single barrel rather than multiple barrels because of possible leakage from the numerous connections. The 9 barrels in the original system had 17 connections. There were 17 locations where leakage could occur as opposed to 1 location in the current system.

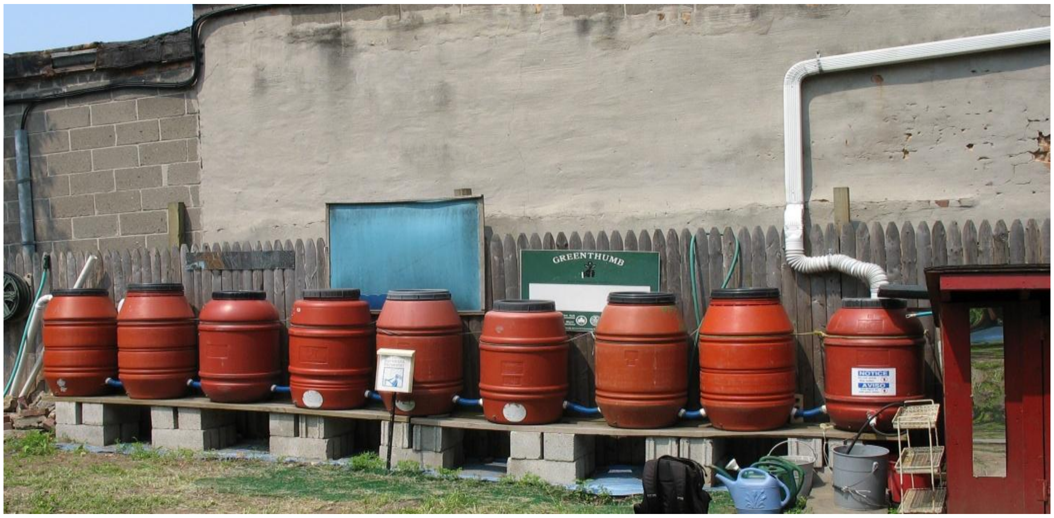

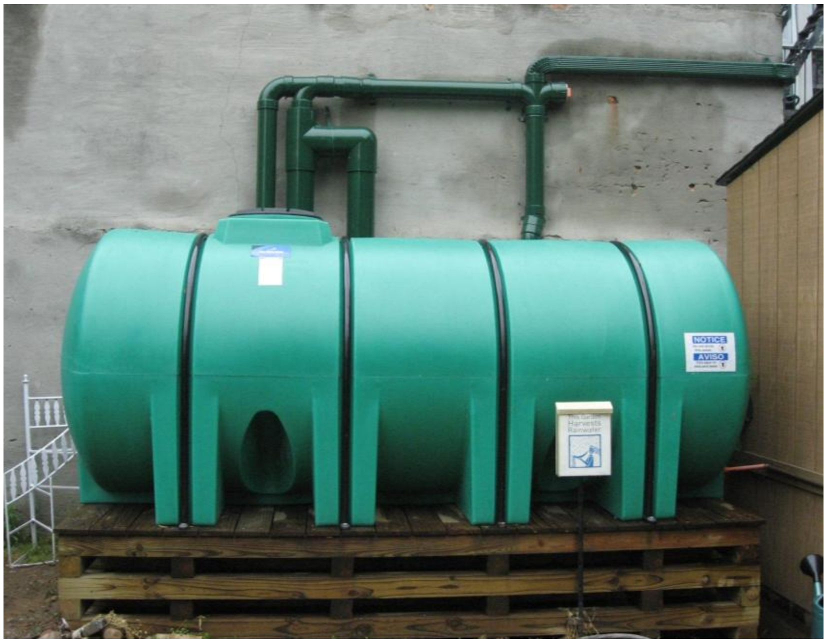

The capacity of the system should be determined by two factors: the size of the roof, which determines how much water can be collected and the size of the garden, which measures how much water is needed. A 300 gallon vertical cylindrical storage tank is approximately 42 inches in diameter and 51 inches tall. A 1000 gallon polyethylene tank is approximately 92 inches in diameter by 60 inches tall or 61 inches in diameter by 100 inches tall.

The amount of water you can harvest varies depending on the size of your roof. You will be surprised how much water you will be able to collect from even the smallest roof. **The rule of thumb is 600 gallons of water per inch of rain per thousand square feet of catchment area.** Not all the rain that falls can actually be collected. Efficiency is usually presumed to be 75% depending on system design and capacity. Here is the basic formula for calculating the potential amount that can be collected:

$$\frac{\text{(Catchment area)} \times \text{(inches of rain)} \times \text{(600 gallons)} \times (.75)}{\text{1000 square feet}}$$

Remember that a roof can have several downspouts. Your catchment area only consists of the region where the downspout that is connected to the rain barrel collects from. Pay attention to the gutter and slope of the roof to determine which part is your catchment area. In order to calculate the catchment area, use measuring tape to determine the width and length of your roof in square feet. Multiply the length and width of your catchment area for the area of your roof.

The sample roof shown below (see picture below[5]) has a catchment area that is 40 feet wide and 30 feet long. Hence, it has a 1200 square feet roof (40 feet wide x 30 feet long). Assume that it rains 2 inches. We can now plug this information into our general formula (see equation above).

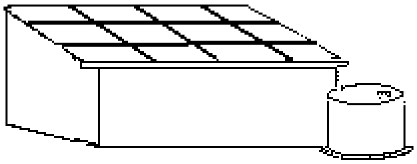

Catchment Area = 1200 square feet
Amount of Rain = 2 inches
Gallons of water collected per inch of rain per 1000 square feet = 600 gallons
Percent Efficiency = 75% or .75

$$\frac{\text{(1200 square feet)} \times \text{(2 inches of rain)} \times \text{(600 gallons)} \times (.75)}{\text{1000 square feet}} = \text{1080 gallons}$$

Therefore a 1200 square foot roof will collect 1080 gallons of water on a day with 2 inches of rainfall. Using this formula and your region's average rainfall numbers combined with an estimated water need, you should be able to calculate the approximate size of your tank or barrel system.

Roof washer sizes also differ according to the size of the roof. **The rule of thumb is one to two gallons of roof washer capacity for every 100 square feet of catchment area.** A 1 foot length of 6 inch diameter PVC pipe holds 1.5 gallons. A 1 foot length of 4 inch diameter PVC pipe holds .66 gallons.

We will use the same 1200 square foot roof shown above as an example. Since the rule of thumb suggests one gallon of roof washer capacity for every 100 square feet of catchment area, the sample roof will need 12 gallons of roof washer capacity.

If you are using 6 inch diameter pipe, use the following formula:

$$\frac{\text{Number of gallons of roof washer capacity}}{\text{1.5 gallons/ feet}}$$

The size of our roof washer on our sample roof can be calculated by:

$$\frac{\text{12 gallons of roof washer capacity}}{\text{1.5 gallons / feet}} = \text{8 feet of 6" pipe}$$

If you are using 4 inch diameter pipe, use the following formula:

$$\frac{\text{Number of gallons of roof washer capacity}}{\text{.66 gallons / feet}} = \frac{\text{12 gallons}}{\text{.66 gallons/ feet}} = \text{18 feet of 4" pipe}$$

Therefore a 1200 square foot roof will need 8 feet of 6” diameter pipe or 18 feet of 4” diameter pipe. Remember it is possible to separate long lengths of roof washers into several, continuous pipes instead of just one long pipe.

Construction

The first step of the construction project is to make a plan drawing. The plan drawing considers the location of the downspout and the amount of space available. Each system must include the tank, the first flush and the overflow. Here are some issues that should be considered in your planning:

- The most stable place to position your barrel is against a stable wall on level ground as close to the downspout as possible.
- Gravity moves water downhill. Be sure there is available space for a downward pitch in all pipes.
- The barrel on its platform is the highest point of the garden but the lowest point of the system.
- The overflow pipe should be directed toward a rain garden not toward pathways or structures.
- The overflow pipe should flow from the barrel's highest point.
- The spigot should be at the barrel's lowest point.

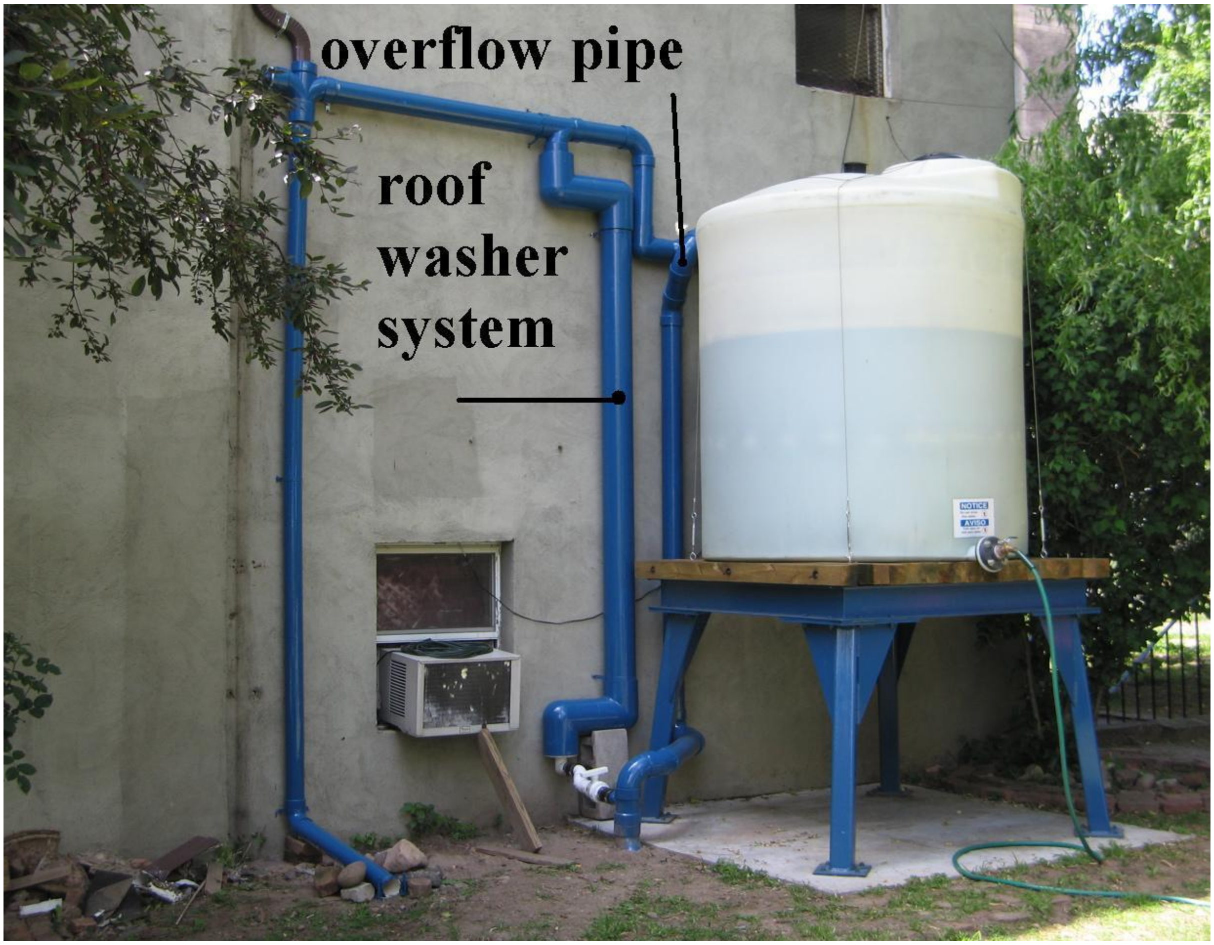

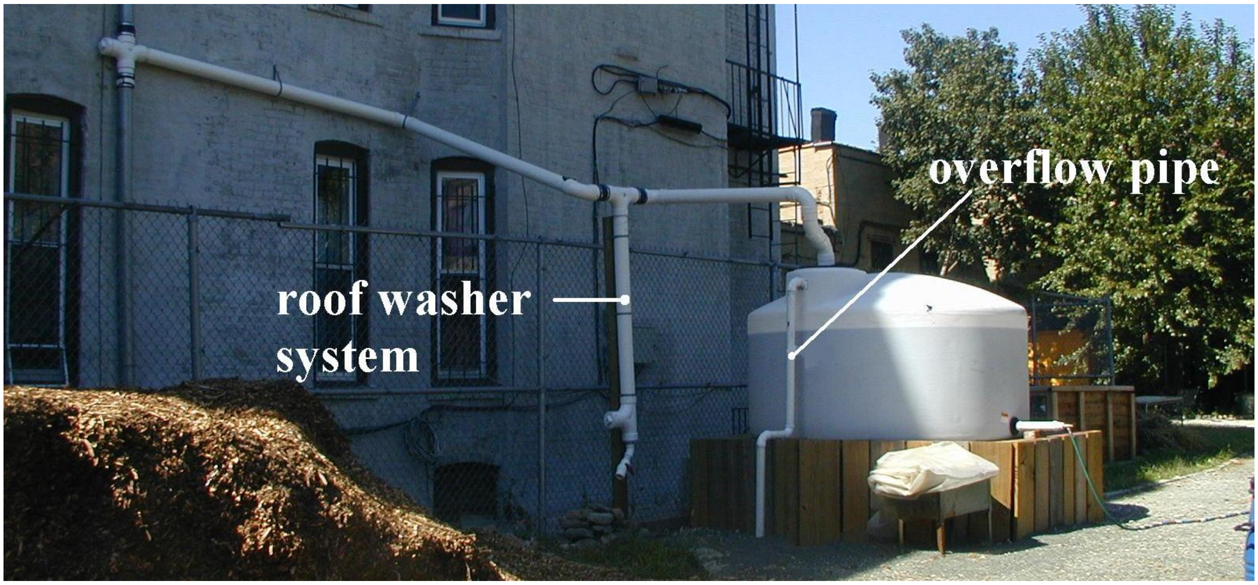
overflow pipe
roof washer system

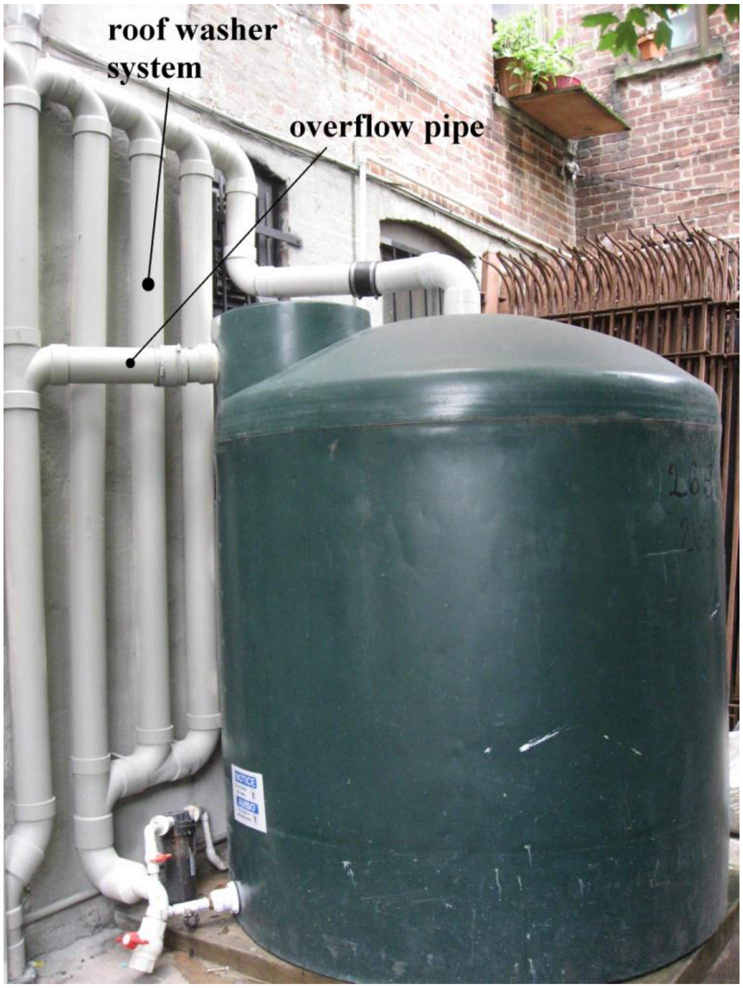
roof washer system
overflow pipe

Tanks

Tanks can be made from all sorts of materials such as cement (see picture on left[6]), metal, ceramic and wood (see picture on right[7]). In tropical countries, a terra cotta tank can be used. No wintering tee is necessary for these tanks because there is no danger of freezing.

For design purposes, bladders can also be used to store water instead of tanks. Bladders are large, flexible bags functioning in a way similar to the water balloon (see picture on left[8]). This way, engineers can create storage tanks in a variety of shapes instead of the traditional cylinder.

The tank system includes the spigot, the inflow pipe, the overflow pipe, the lid and the platform it rests on.

Spigot: A variety of bushings can be used to connect the spigot. It is preferable to use a single reducer bushing but that is often not possible. For example, the tank's spigot hole is 2 inches in diameter. The spigot pictured here is a ¾ "ball valve hose bib. In order to reduce the 2" opening to the ¾"spigot, the following 3 bushings were used:

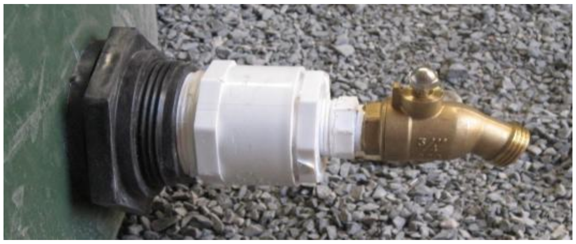

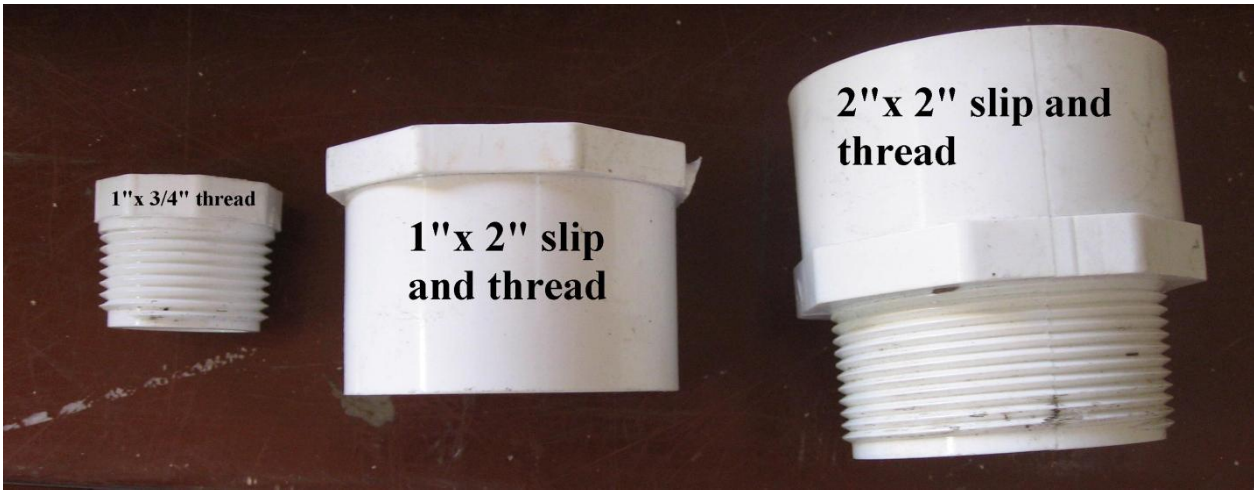

Lid: The lid should remain closed at all times. Still water is a haven for mosquito larvae. Inspect the lid and any vents annually and clean as necessary

Inflow and Overflow pipes: Some tanks come with bulkhead fittings installed for the inflow and overflow pipes. For those that don't, a hole saw is used to drill a large hole in the barrel. Two bushing pieces, like the ones pictured below, are used to fashion a bulkhead fitting.

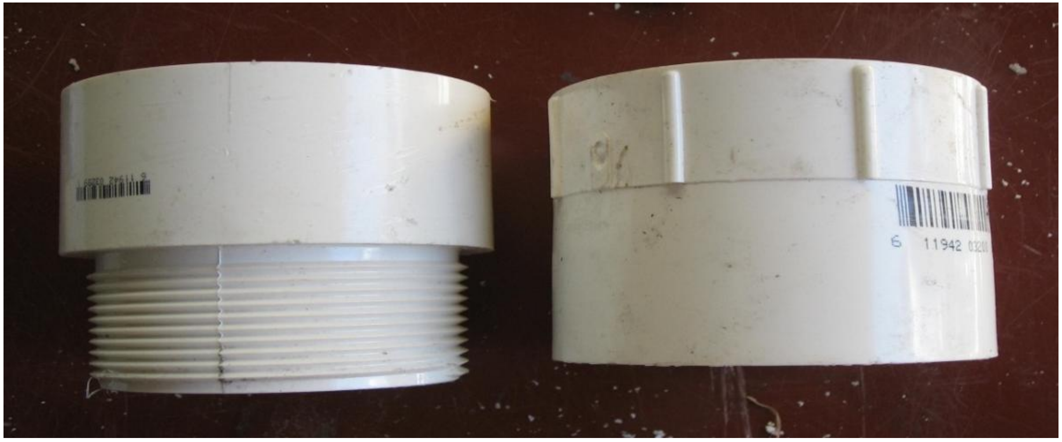

Use a marker to trace the hole you will cut for the threaded bushings. Make the hole tight to the threads. Once the hole is prepared, thread the first piece in from the outside.

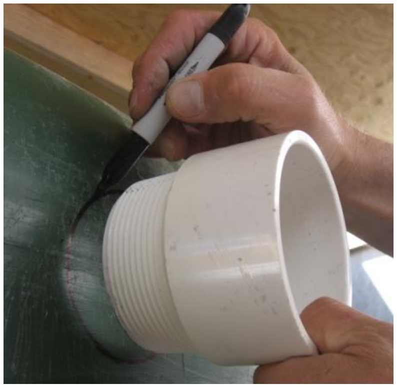

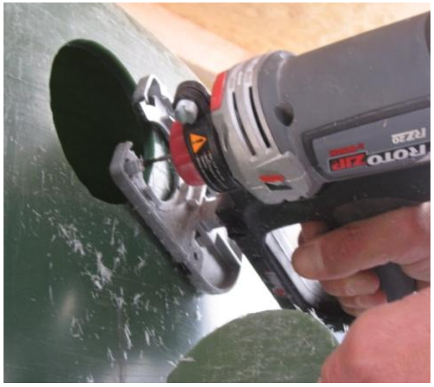

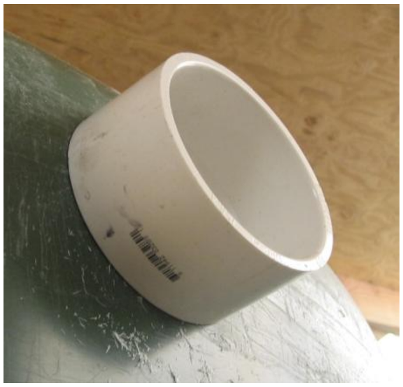

Thread the second piece from the inside of the tank using a flexible wrench. Use gutter sealant to create a watertight seal.

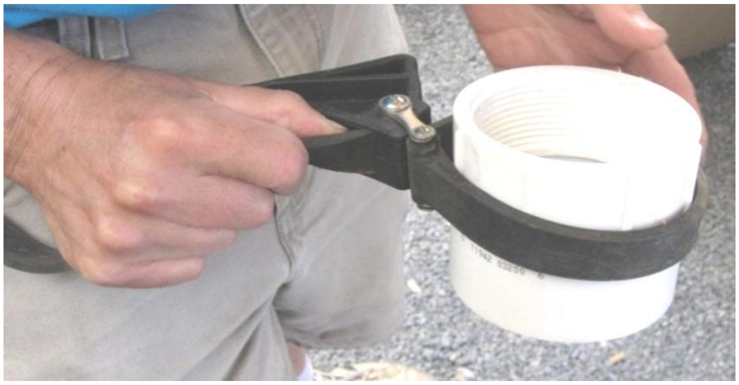

Platform: The purpose of the platform is to raise the tank enough to get your watering can under the spigot and to create pressure flow for a hose. Spread a thick gravel base over an area slightly larger than the platform. This will promote good drainage and allow for final leveling once the platform is built.

A stable arrangement of cement cinder blocks, like this one, makes a great platform for small 55 gallon tanks but is not ideal for larger tanks. Since the blocks easily shift, rot resistant 4”X 4” or 6”X 6” lumber laid out in alternating rows is the best option for a platform. Steel platforms are also a possibility but are more costly. Be sure to level and compact the ground well where the tank will sit.

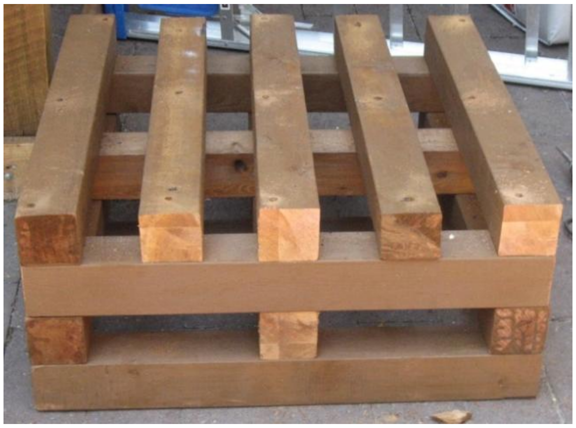

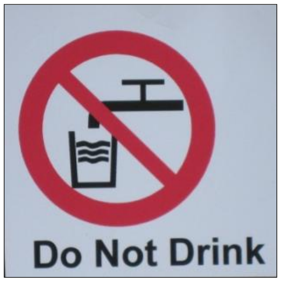

Warning Signs: Do not drink the water collected. Only use it to irrigate your garden. Installed PVC is inert and it is used as a water supply pipe for house trailers and other homes. The possible contamination is almost exclusively from harmful bacteria. For water to meet drinking water standards it must be treated, usually with chlorine, often filtered and sometimes treated with ultraviolet light. A warning sticker or sign should be placed on your rain barrel to avoid the possibility of anyone mistakenly drinking the water.

Leaders

The leaders of the rain water harvesting system refer to the series of pipes that lead to the tank. This includes the pipe that connects the downspout, the wintering tee, the roof washer system and the pipe that connects the barrel. Gravity keeps the water flowing. Remember to put a downward pitch on all pipes. A quarter inch down for every linear foot will create an adequate pitch. Use a level (see picture below). It may be useful for someone to stand back to visually check the pitch and help you adjust the pipes as you are pitching them.

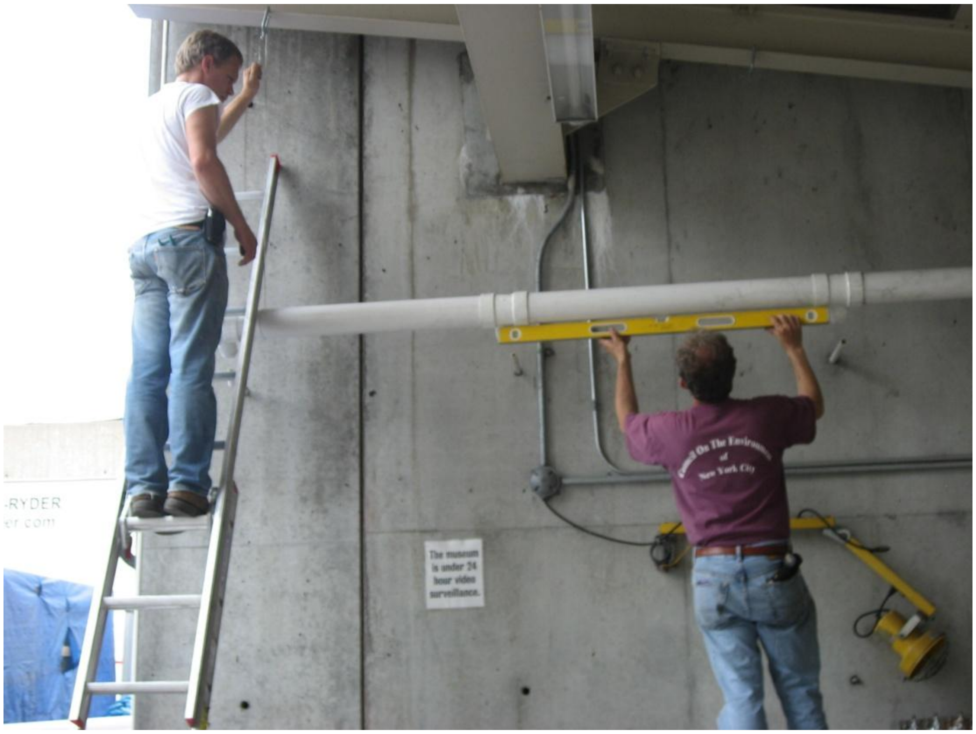
RYDER
er.com
The museum is under 24 hour video surveillance.
Council On The Environment of New York City

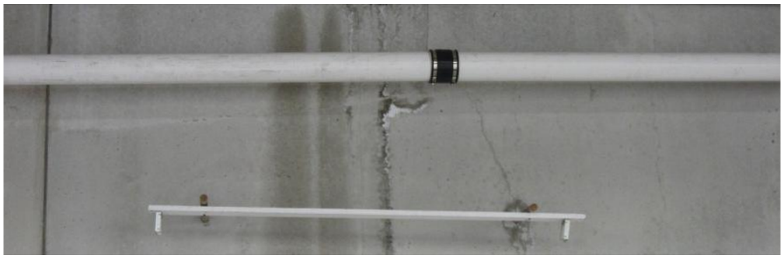

Keep pipes pitched downward.

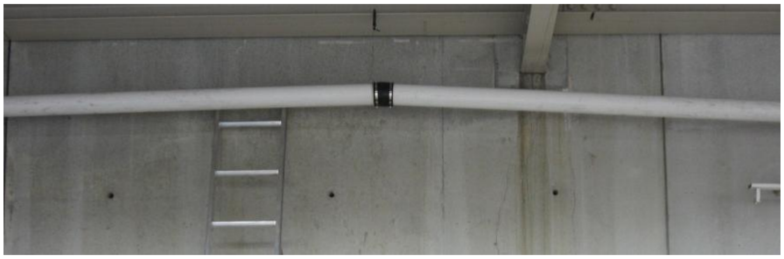

DO NOT pitch pipes upward.

Keep the pitch of your pipes even and downhill. The picture on top is correct. The picture on the bottom has an upward pitch and will create problems.

Connecting the downspout: Our intent with rain water harvesting is to divert storm water, store it for future use and keep it out of the sewers. There are several options to divert the rain water: insert a 3 way tee into the downspout (see picture on left) or connect the existing downspout directly to the rain barrel (see picture on right). You can design your own custom diverter as well.

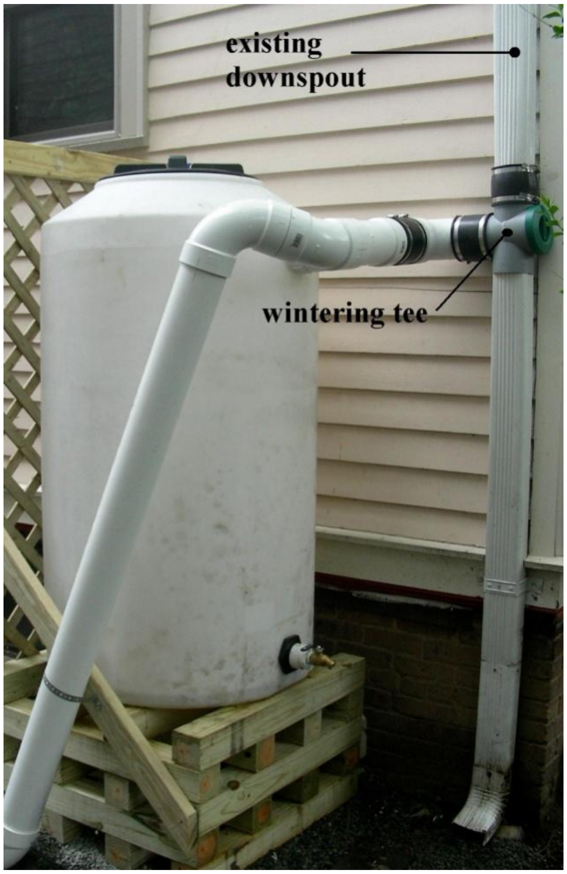
existing
downspout
wintering tee

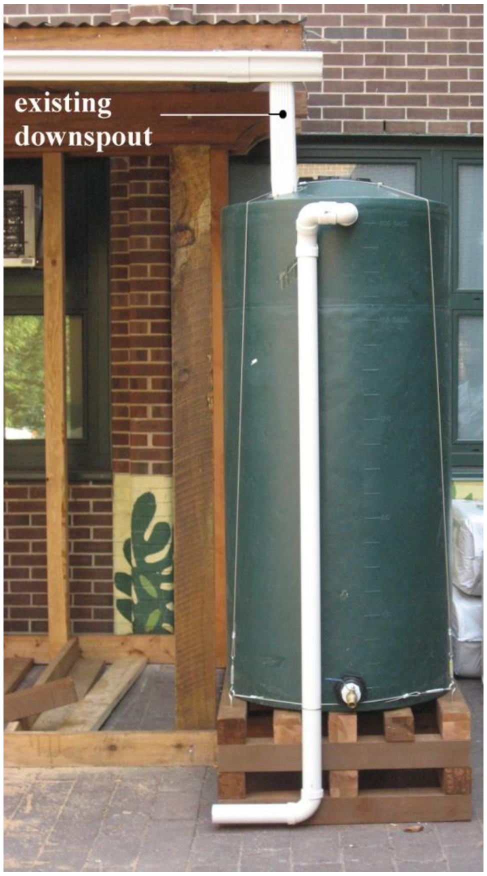
existing
downspout

Another possibility is to replace the downspout entirely with pipes leading to your rain water harvesing system (see picture below and on the next page).

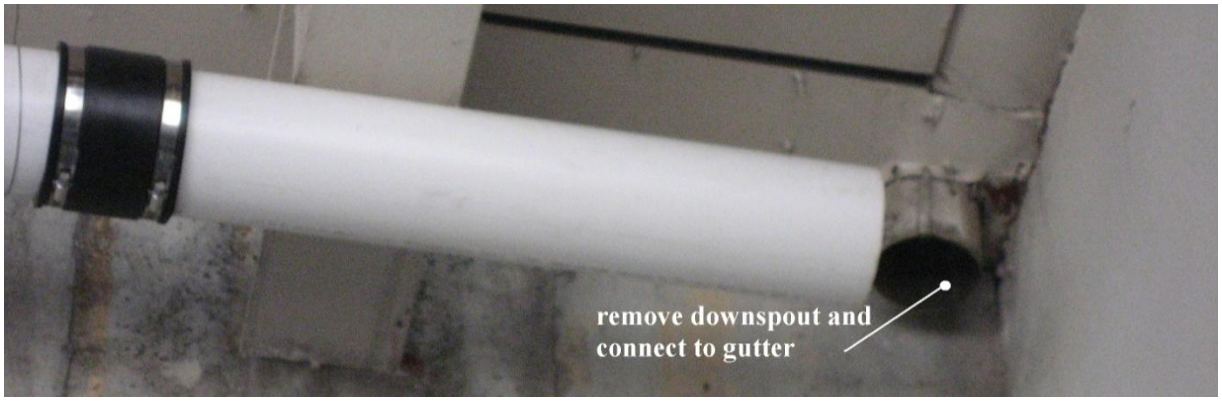
remove downspout and
connect to gutter

Connecting to the downspout is one of the last steps in construction of your system. You do not want to disconnect your downspout and start collecting rainwater with an unfinished system.

Wintering tee: The purpose of the wintering tee is to allow the water to be directed back into the sewers during freezing temperatures. Rain water should not be collected at that time because it may freeze and cause breaks in the system. A simple wintering tee utilizes a removable plunger head. The plunger head blocks water flow into the sewer and forces it to enter the system. It is removed during the winter and rain water will just flow straight down into the sewer (see picture on left and diagram on page 4). Complicated wintering tees include a filtering system. It operates with the same principle but instead of a removable plunger head, there is a removable filter which catches debris (see picture on right).

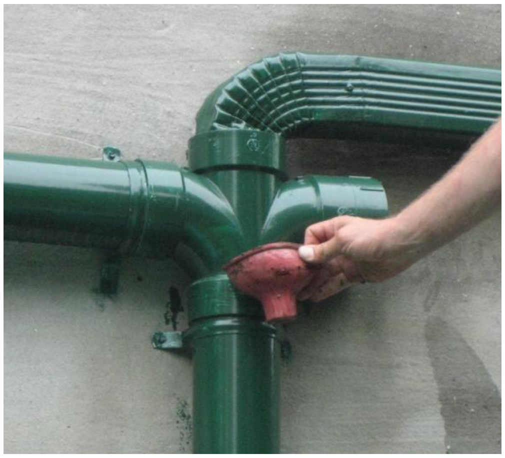

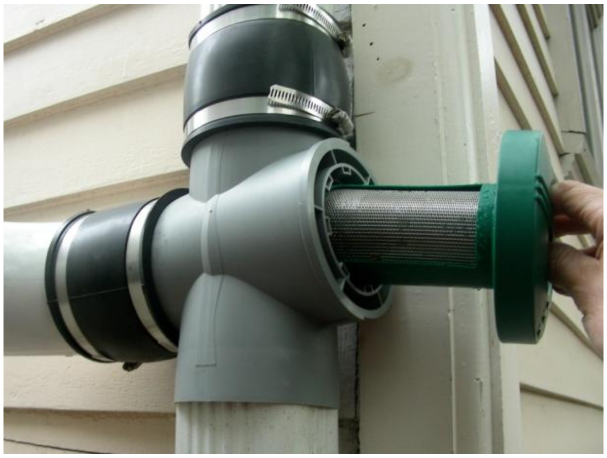

Roof washer system: Refer to ‘Roof washers’ on page 30.

Connecting the barrel: Refer to ‘Tanks’ on page 23.

Here are some ideas to help you with the design of your leader.

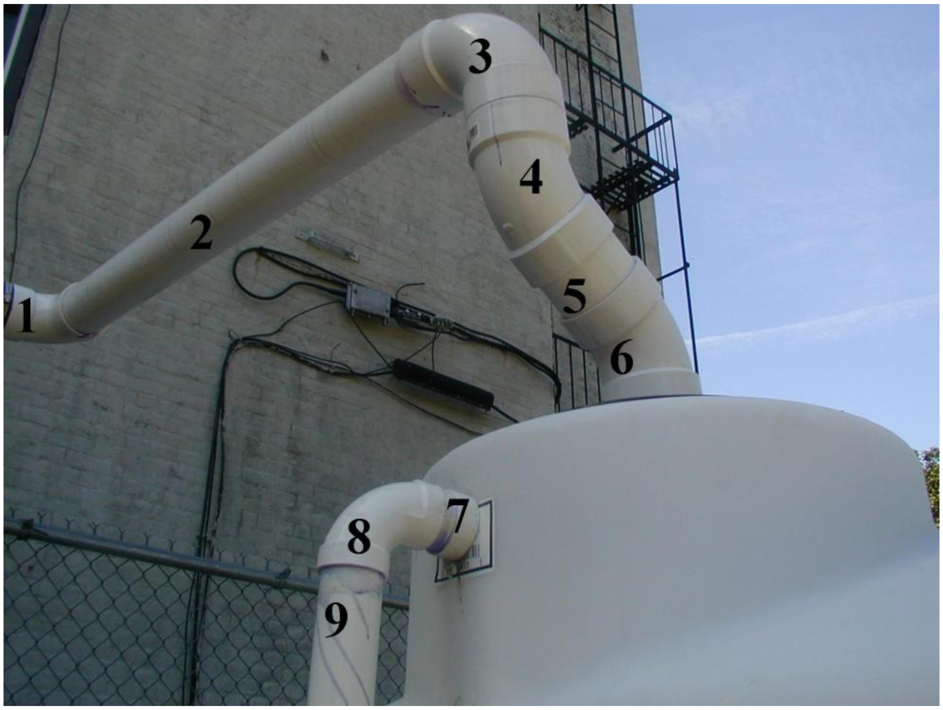

INFLOW
1: 90° elbow
2: 4"pipe
3: 90° elbow
4: 45° elbow
5: 4"pipe
6: 45° elbow

OVERFLOW
7: 2"bushing
8: 90° elbow
9: 2"pipe

This design uses a separate inflow and overflow pipe.

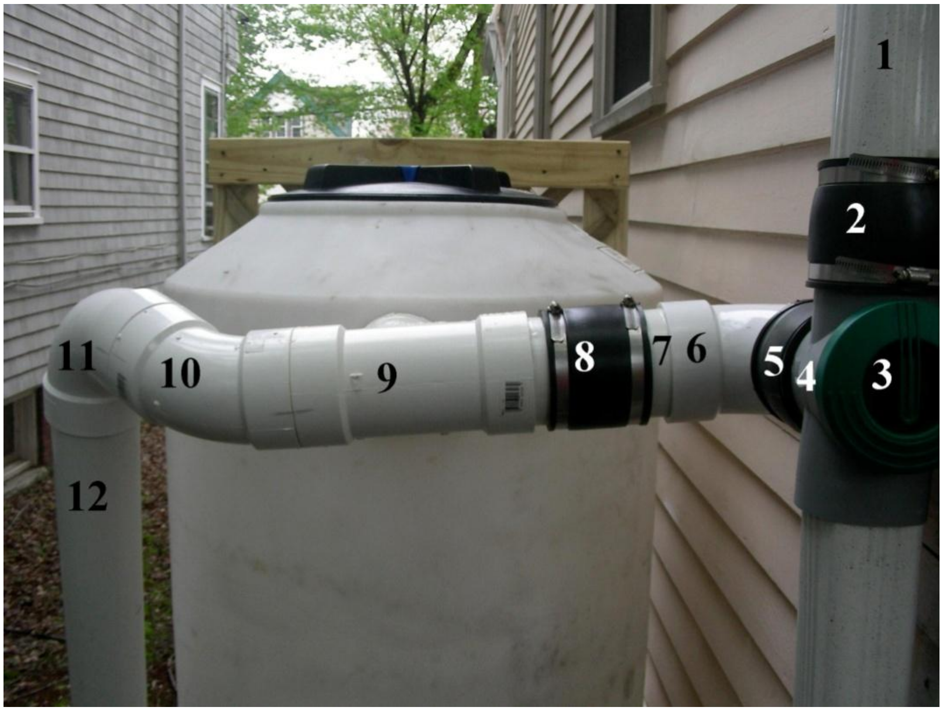

INFLOW
1: downspout
2: 4"rubber coupling
3: cap
4: 4 way tee
5: rubber coupling
6: 90° elbow
7: 4"sleeve
8: 4"rubber coupling
9: 3 way tee

OVERFLOW
10: 45° elbow
11: 90° elbow
12: 4"pipe

This design uses the same tank opening for both the inflow and overflow.

Roof Washers

Roof tops are prone to collect leaves and dirt. The rain carries all this debris with it as it enters the gutter and downspout. The roof washer or first flush system is a simple way to filter the water you collect. It is a series of pipes that storm water flows into before entering the rain barrel. Once the roof washer is full, water will begin to fill the rain barrel without mixing with the dirty water contained in the first flush. The appropriate size of the roof washer varies depending on the size of the roof. Refer to 'Sizing of the tank and roof washer system' on page 18 for details on how to calculate roof washer size. Small first flush systems can be a single downward pipe. However, to accommodate larger roofs, first flush systems can be constructed of multiple downward pipes.

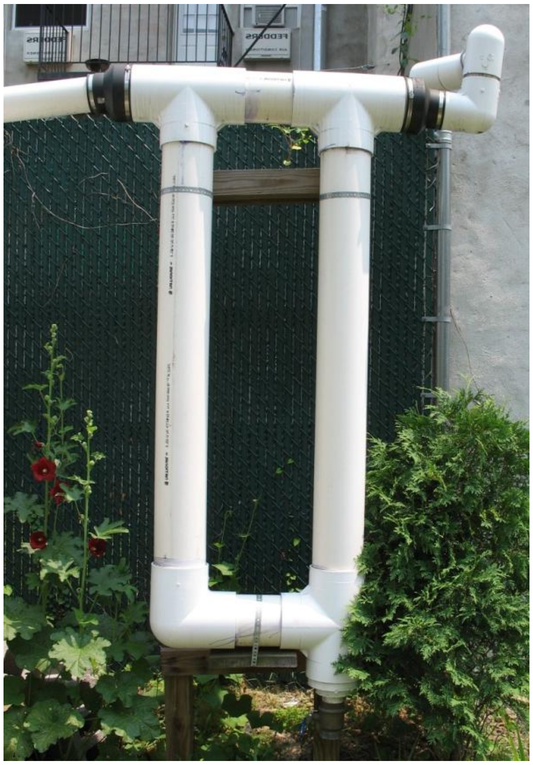

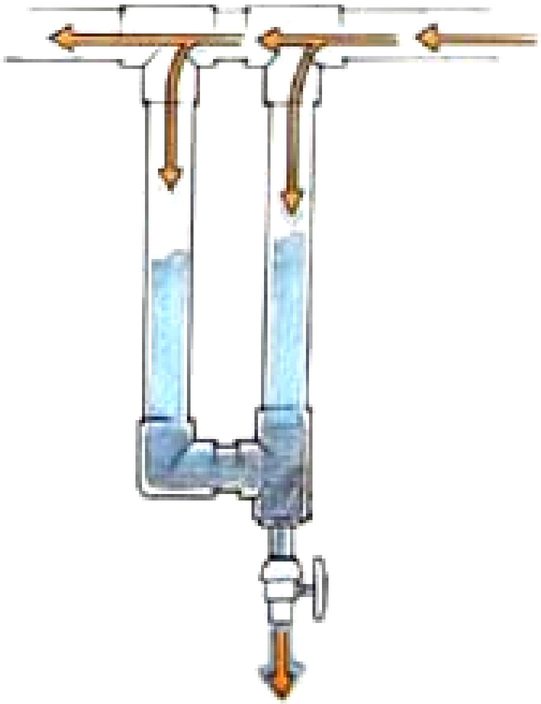

The roof washer must be emptied promptly after a rain event and valve reset in the closed position to be ready for the next rain. Drain the roof washer to a safe location. Include another spigot or valve in your design to drain the first flush.

An alternative to the container method of the roof washer is a device like the Flow-Rate Diverter by Safe Rain. It consists of a plastic ball attached to a spring. The top of the ball contains a screen and a movable tab. During a rainfall event, water that enters the pipe will either flow into the ball or on its sides (see picture on left). The purpose of the screen on top is to gradually allow water to enter the ball. Once the ball is filled, it will sink down and block the drain (see picture on right). Water can then start filling up the pipe and start entering the tank. The moveable tab controls the amount of water that enters the ball by covering or exposing the screen as desired. For example, larger barrels that require larger roof washers require a screen that is less exposed. Larger amounts of water will have to pass through the pipe before the ball can be filled. The ball will have a small drip hole to slowly empty once the rain has ended to then return to its original position to be ready for the next rainfall.

Entering rainwater will flow into this pipe. Some will enter the ball, making it heavier and some will flow past the ball and out of the pipe.

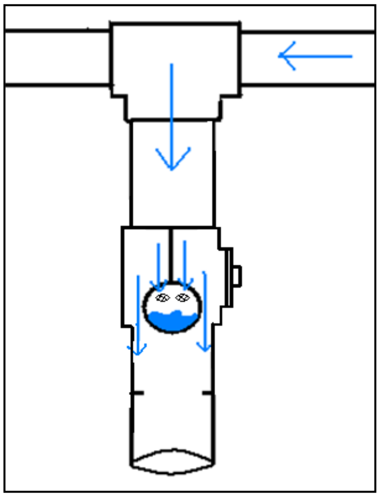

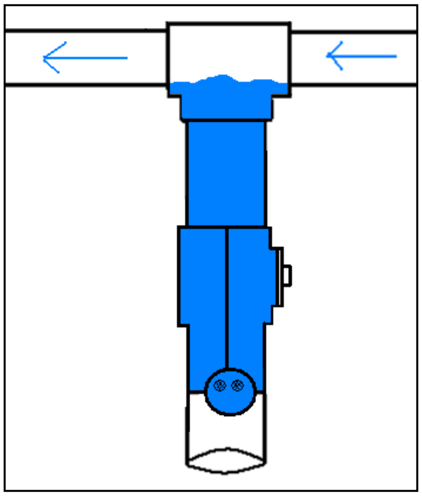

Once the ball is saturated with water, it will sink down and block the pipe. Water will fill up the pipe and finally enter the tank.

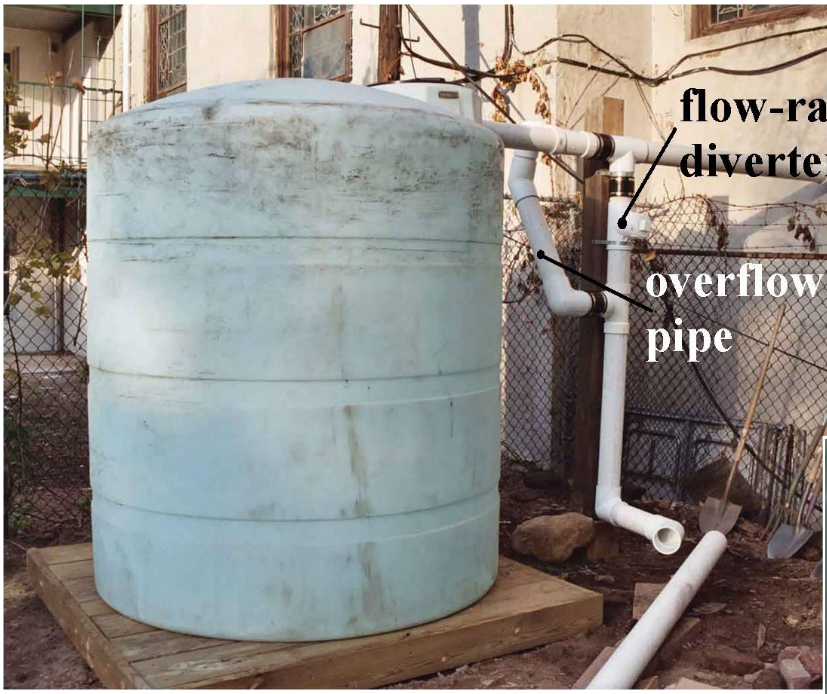

This device allows the overflow pipe to be connected to the roof washer pipe.

gradually enters the ball through this screen.

Here is an idea to help you design a roof washer of your own.

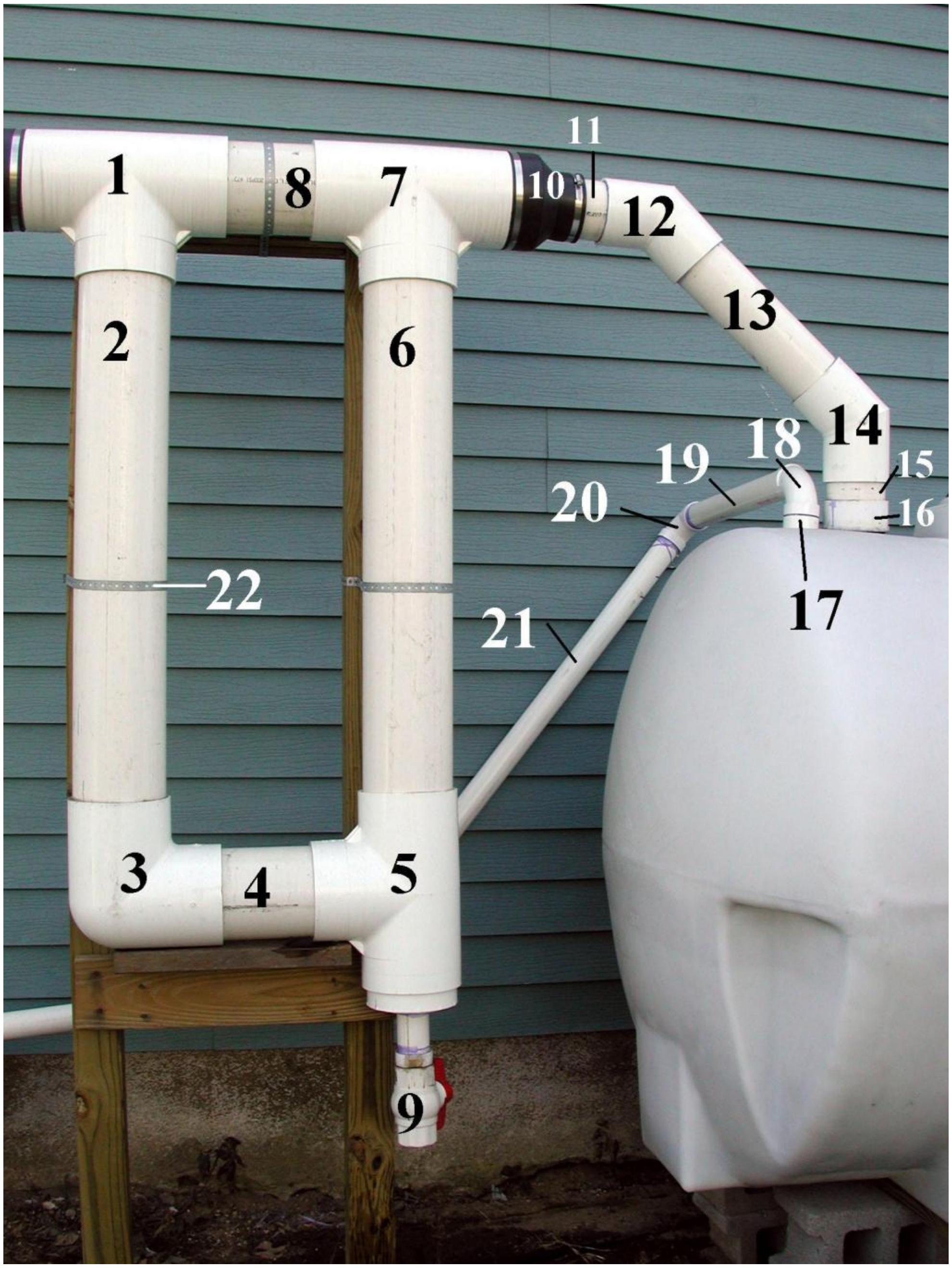

ROOF WASHER
1: 3 way tee
2: 6"pipe
3: 90° elbow
4: 6"pipe
5: 3 way tee
6: 6"pipe
7: 3 way tee
8: 6" pipe
9: drain valve

INFLOW
10: 6"x4" rubber coupling
11: 4" pipe
12: 45° elbow
13: 4"pipe
14: 45° elbow
15: 4"pipe
16: bushing

OVERFLOW
17: bushing
18: 90° elbow
19: 2"pipe
20: 90° elbow
21: 2" pipe

ANCHORING
22: plumber's strap

Although this rainwater harvesting system has a good roof washer design, it does not have a good overflow design. Excess water will back up into the inflow pipe before the overflow pipe starts working. The overflow pipe should also be the same diameter as the inflow pipe to avoid a bottleneck situation.

Overflow pipes

When the rain barrel reaches its capacity, the overflow pipe discharges the excess water so that water won’t start spilling out from around the lid. Cut a hole near the top of the barrel and connect an overflow pipe there. Be sure that the overflow pipe is not directed toward water-sensitive structures or areas where water can collect and do damage. The overflow pipe should be the same size as the inflow pipe so that a bottleneck situation is avoided during heavy rainfalls.

Here are some ideas to help you design an overflow pipe of your own.

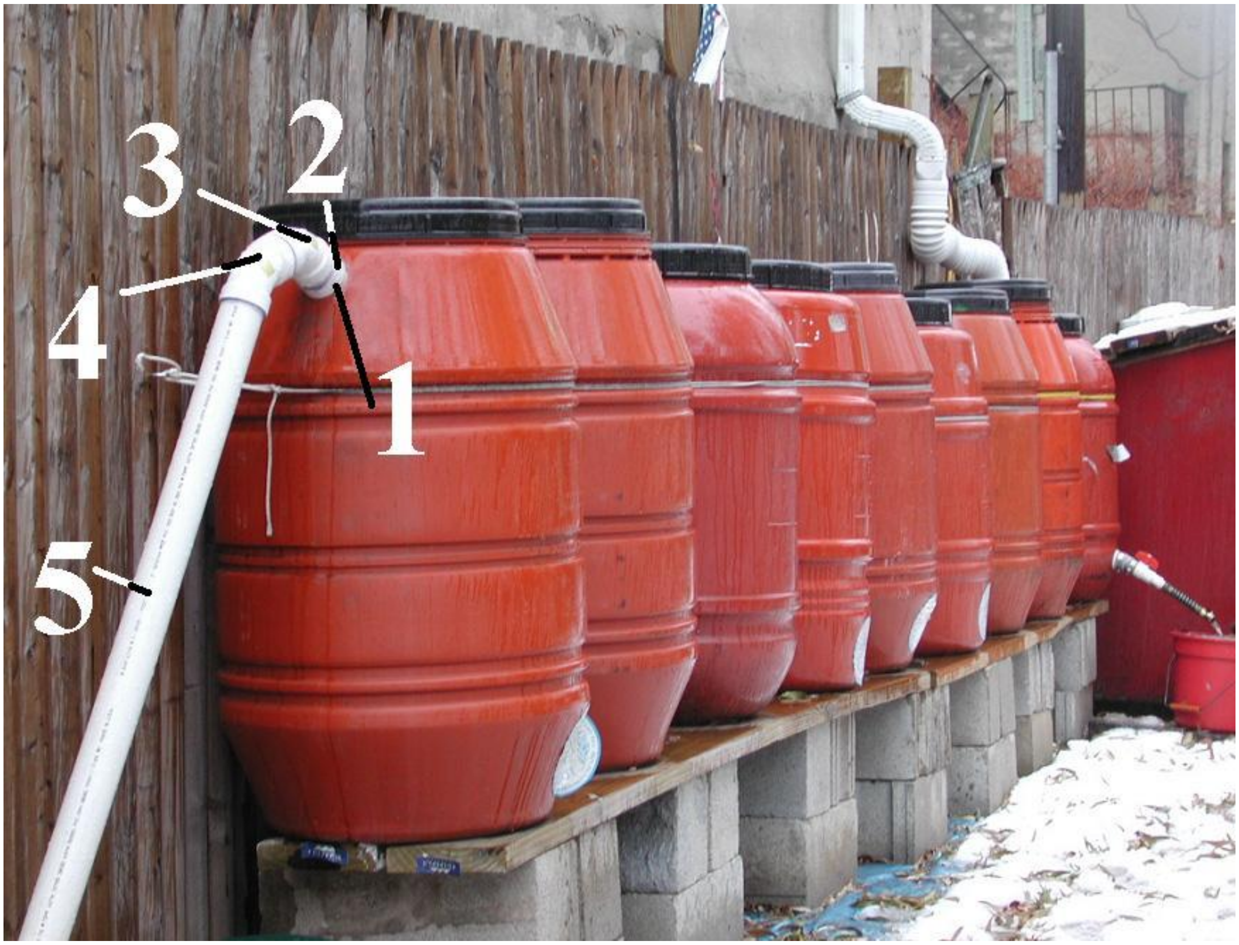

OVERFLOW
1: bushing
2: 2”pipe
3: 45° elbow
4: 45° elbow
5: 2” pipe

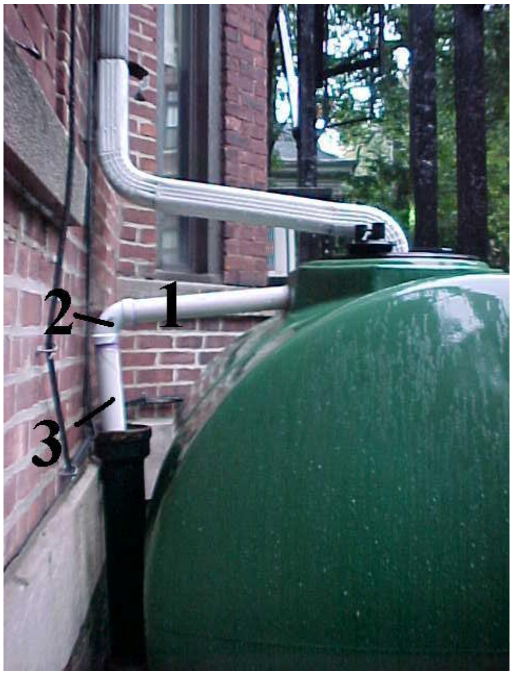

The overflow pipe is connected back to the sewer. Building systems close to the downspout requires less piping.

OVERFLOW
1: 2” pipe
2: 90° elbow
3: 2” pipe

Instead of directing water back into the sewer, you can direct the water into a rain garden, gravel filled trench or connect it to a drip irrigation system. Overflow can be turned into works of art. Here the overflow is turned into a small water course (see picture on below), which ultimately ends in a rain garden.

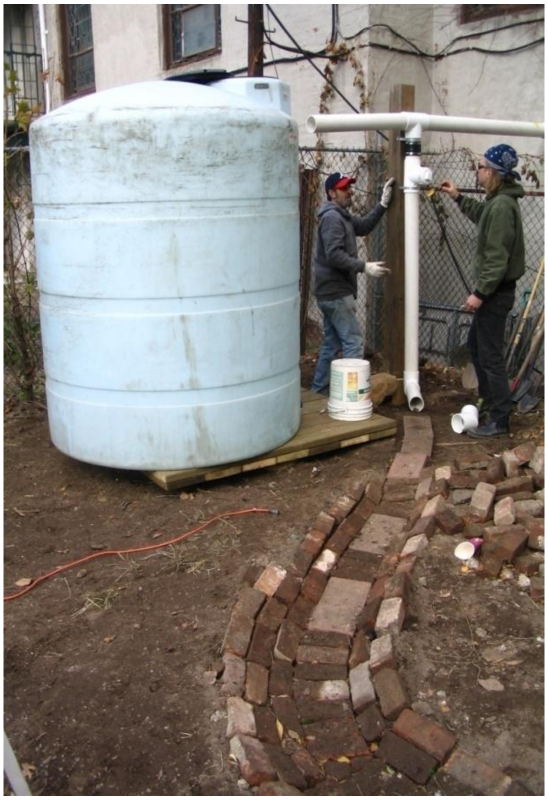

Rain gardens are depressed plots of land (see pictures below). It should be at the lowest point in the garden so that water can flow there easily. Dig a deep trench and fill it with 5 inches of gravel for seepage. Cover this area with burlap netting so that the soil will not sink down into the gravel. Place a couple of inches of soil on top and start planting. Only plants able to survive both dry and saturated soil should be grown. Native plants are also encouraged because they are more tolerant of local climate and soil conditions. These plants include wildflowers, ferns, shrubs and small trees. [9] Brooklyn Botanic Garden provides a thorough list of rain garden plants for different regions on their website at:
http://www.bbg.org/gar2/topics/design/2004sp_raingardens.html

Bioswales are alternatives to rain gardens. They are depressions in the landscape with collections of rock, gravel and vegetation that act as a filter for water. They effectively strain silt, inorganic contaminants, organic chemicals and pathogens. The longer the water is trapped inside the bioswale, the easier it is for pollutants to be trapped. They are most commonly found surrounding a parking lot so that oil runoff can be filtered before entering the sewer system.[10]

Anchoring

Because the rain barrel is fairly light and may tip or fall when it is empty, it is important to anchor the system to the platform or a structure. The easiest method to tie down your system is to use nylon rope. Steel cable is also a solution. Since cabling is more complicated, we will illustrate these steps.

Cut a piece of steel aircraft cable and slip 5 cable swedges in it Use the first piece of cable swedge **(see picture 1)** and a crimping tool to connect it into a circle **(2).** Its circumference should be a less than the circumference of the rain barrel but more than that of the lid. Cut two more long pieces of steel cable. They should be at least twice the height of the rain barrel. Use one cable and slip it through two of the metal cable swedges that are already on the circle. Making sure both lengths of cable are even; use the crimping tool to secure the metal swedges in place. Do the same with the other cable and two metal swedges **(3).**

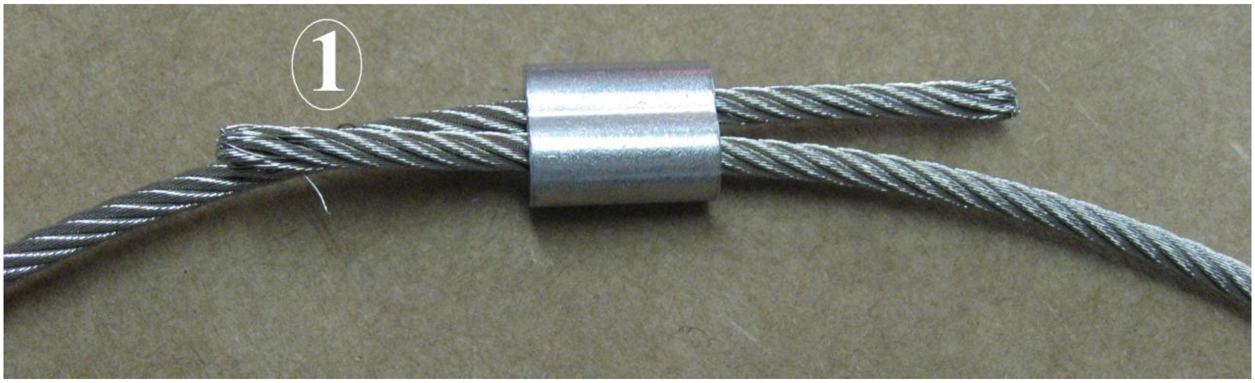

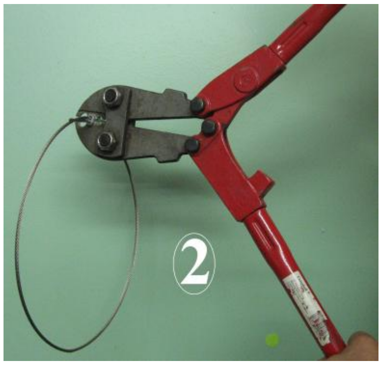

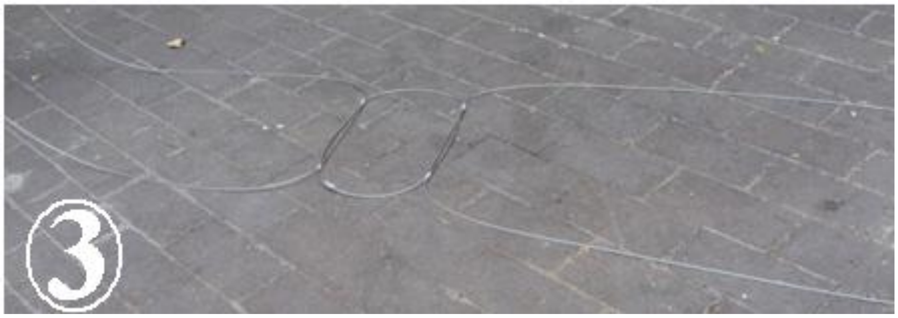

Screw a metal eyebolt into each of the four corners of your platform. Connect a turn buckle to each metal eyebolt **(4).** Place the series of cables onto your rain barrel. The circle should be on top of the rain barrel **(5).** The four hanging ends should be facing each corner of the platform. Loop the ends of the cables into the turn buckle **(6).** Use a metal swedge to secure the loop. Do this for all four corners. Finally, twist the turn buckles until the metal cables are taut **(7)** to finish securing your rain barrel **(8).**

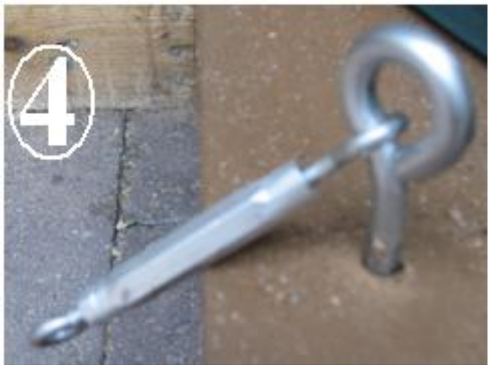

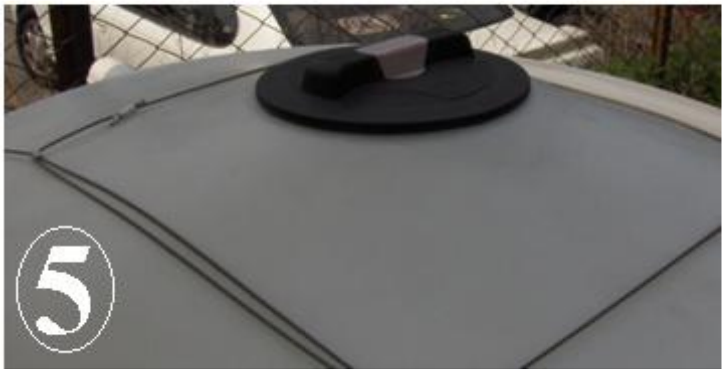

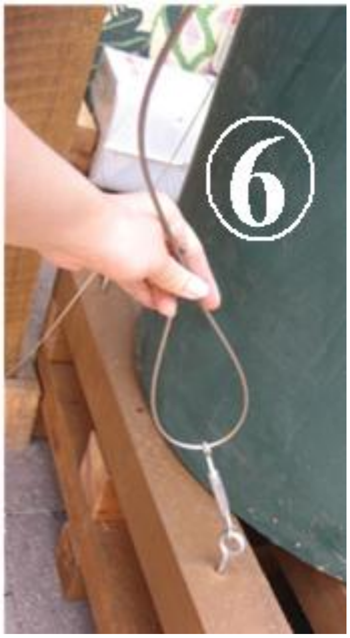

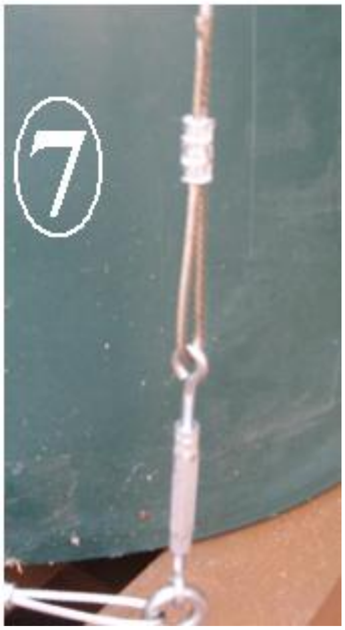

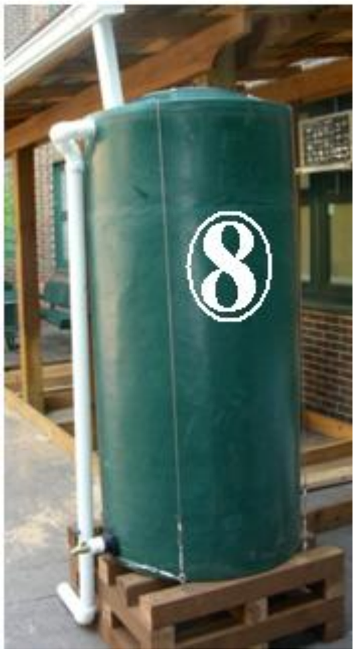

For cinder block platforms, nail a thick wooden plank on top and use the eyebolt method mentioned above.

Pre-fabricated steel brackets are generally used to anchor horizontal tanks.

NOTICE
AVISO

It is a good idea to anchor pipes as well. If the pipe is hugging the rain barrel, anchor the pipe to that platform using cable **(see picture 1),** perforated plumber's strap **(2)** or other methods.

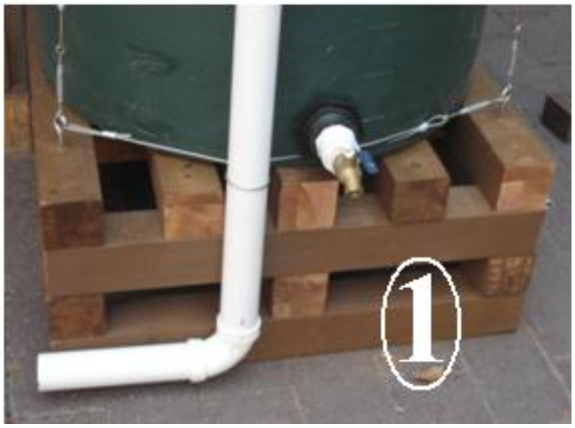

Pipes that are directly against the wall of a building can be secured to the wall using plumber's strap attached with tapcon screws, which seal the hole drilled in the wall to avoid leaks **(2).**

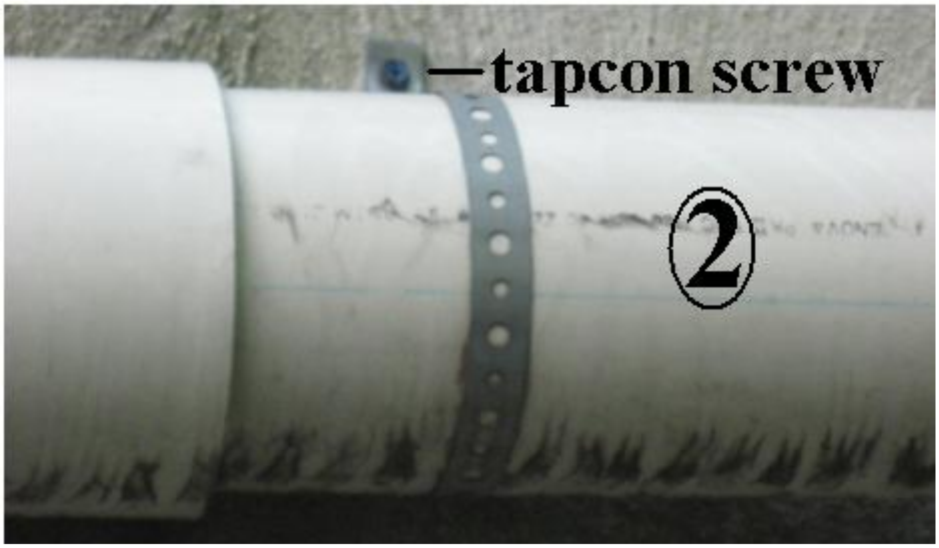

Drip Irrigation

Drip irrigation systems reduce water use. Small amounts of water are supplied to the base of plants. Since the water is applied directly to the soil, rather than onto the plant, evaporation from leaf surfaces is eliminated. The water is also placed where it is needed rather than sprayed over the entire garden.

With a drip system, water flows into a series of thin flexible hose lines. These irrigation hoses have tiny holes at even intervals. These hoses are placed on the surface or beneath a layer of mulch. When water enters the system, it slowly trickles out of these holes and waters the roots of plants (see pictures below[11]). The overflow of rain water harvesting systems is sometimes connected to these smaller pipes.

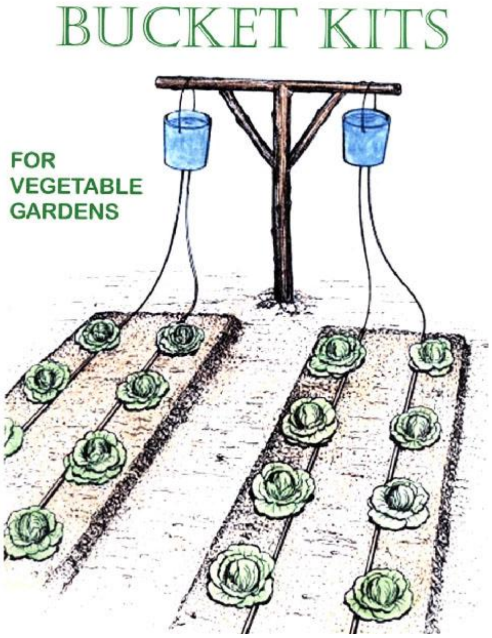

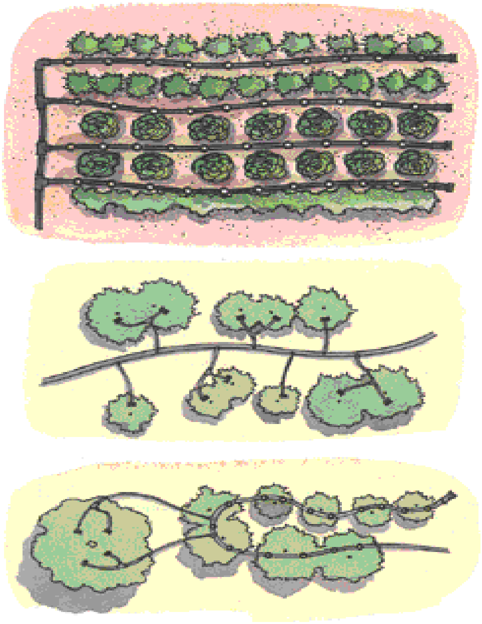

Tubing comes in many sizes (see pictures below[12]). It is used to get the water from the source to the garden. A variety of fittings are available to go around corners and to connect pieces.

In most cases, no special tools or skills are needed. Plastic pipe is punched with an inexpensive tube punch that assures the proper hole size. Spaghetti tubes snap into the hole. No gluing is required. Because the holes are small, they can easily be plugged if you put one in the wrong place.

Gravity Feed Drip Irrigation Kits are used to connect to rain barrels and rainwater harvesting systems where the water pressure is low (see picture below[13]). This system uses small ¼ inch valves instead of emitters to avoid the problem of clogged emitters

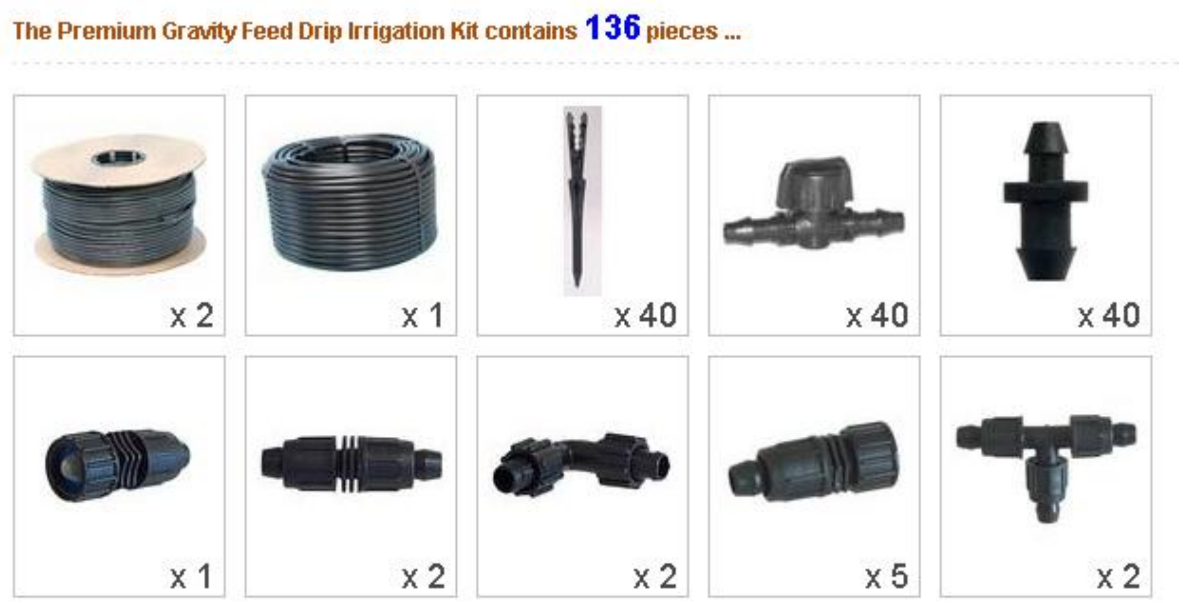
The Premium Gravity Feed Drip Irrigation Kit contains 136 pieces ...
x 2
x 1
x 40
x 40
x 40
x 1
x 2
x 2
x 5
x 2

Drip Irrigation systems need planning, but are neither expensive nor difficult to install. Most drip irrigation suppliers will help you design a system to best meet your gardening needs. Drip systems require periodic maintenance. Check regularly for leaks and broken connections.

Filtration Systems

This rain water harvesting system already has two means of filtration, one at the gutter **(see picture 1)** and the other at the roof washer **(2).**

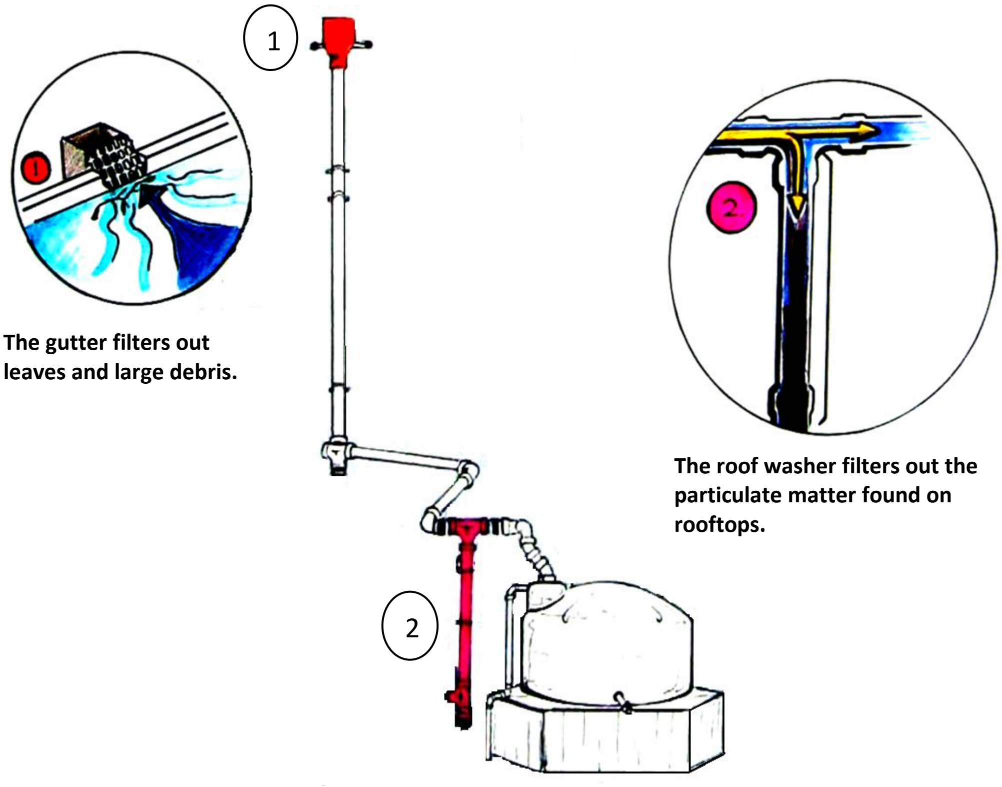

The gutter filters out leaves and large debris.

The roof washer filters out the particulate matter found on rooftops.

This gutter screen that is put at the mouth of the downspout (see picture on right) prevents leaves and other large debris from entering the system. If you want to buy a filter, two choices are Rain Keeper Downspout filter for smaller roofs and the Rain Keeper Downspout collection filter for smaller systems. Refer to “Where to obtain supplies” on page 11 for more information.

Treatment

During warm weather, bacteria may grow inside the storage tank. Organic matter also poses a problem as algae or other contaminants may grow in storage tanks.WRG recommends adding a small amount of chlorine, in the form of unscented household bleach, to your water storage tank. Do not use bleach with any additives like fragrances or softeners. A log should be kept to ensure the proper addition of the chemical. Store log with bleach container. Please remember to wear gloves when handling bleach, to store bleach in a cool, dry place and to label bottle clearly to avoid improper use. On the 1st of each month, add a small amount of bleach to *each* water storage container. If the tank is emptied and refilled in less than one month's time, an application of bleach should be added to the tank when it refills

Tank Size	Amount of Bleach to Add
1000 gallons	1/3 cup
750 gallons	1/4 cup
500 gallons	3 tablespoons
50 gallons	2 teaspoons

Regularly emptying the roof washer and the addition of a small amount of bleach to your tank will ensure the water is safe for watering vegetables. The water stored in the tanks is for ***irrigation purposes only*****.** It is safe to use on vegetables and other plants in the garden, but it is not safe to drink. The quality of this water does not comply with local and national drinking water standards.

Aesthetics

“The most beautiful system is carefully constructed and incorporates a carefully thought out, well designed and installed plumbing system.”

-Lars Chellberg, GrowNYC staff member

Although a rain barrel is considered by some to be unsightly, it is easy to beautify your rain water harvesting system. PVC pipes and tanks can be covered by plants or painted (see pictures below).

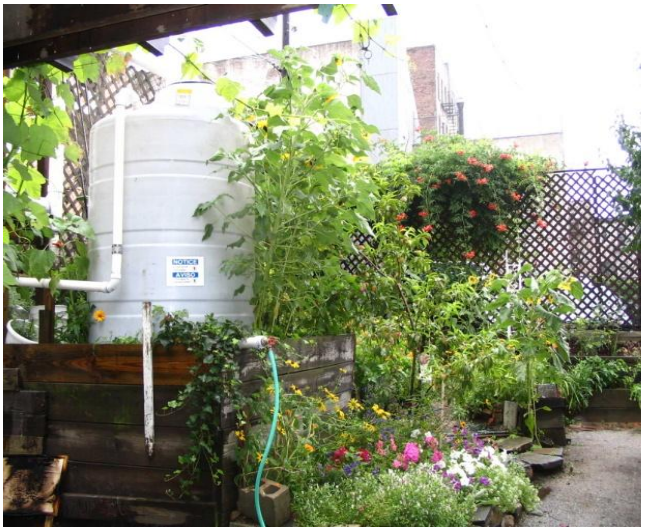

A gazebo at this Seattle community garden covers an underground cistern that stores 5,000 gallons of collected rainwater (see picture below).

Promoting the sustainable practice of rain water harvesting is also important. For example, NYC community gardens use a “This garden harvests rain water” sign (see picture on right), a good idea to put in front of your system. Have printed information for people to learn more about rain water harvesting like this Water Resources Group brochure (see picture on left).

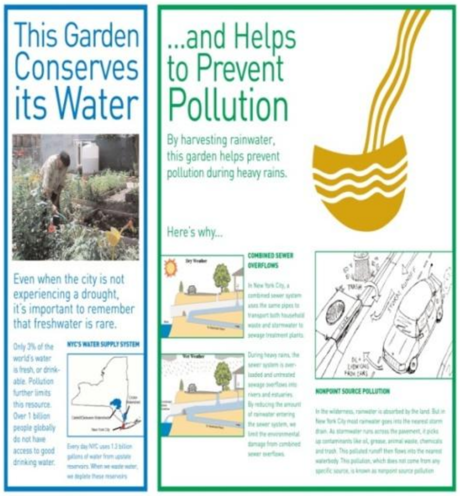

This Garden Conserves its Water

Even when the city is not experiencing a drought, it's important to remember that freshwater is rare.

Only 3% of the world's water is fresh, or drinkable. Pollution further limits this resource. Over 1 billion people globally do not have access to good drinking water.

NYC'S WATER SUPPLY SYSTEM

Every day NYC uses 1.3 billion gallons of water from upstate reservoirs. When we waste water, we deplete these reservoirs.

...and Helps to Prevent Pollution

By harvesting rainwater, this garden helps prevent pollution during heavy rains.

Here's why...

COMBINED SEWER OVERFLOWS

In New York City, a combined sewer system uses the same pipes to transport both household waste and stormwater to sewage treatment plants.

During heavy rains, the sewer system is overloaded and untreated sewage overflows into rivers and estuaries. By reducing the amount of rainwater entering the sewer system, we limit the environmental damage from combined sewer overflows.

NONPOINT SOURCE POLLUTION

In the wilderness, rainwater is absorbed by the land. But in New York City most rainwater goes into the nearest storm drain. As stormwater runs across the pavement, it picks up contaminants like oil, grease, animal waste, chemicals and trash. This polluted runoff then flows into the nearest waterbody. This pollution, which does not come from any specific source, is known as nonpoint source pollution.

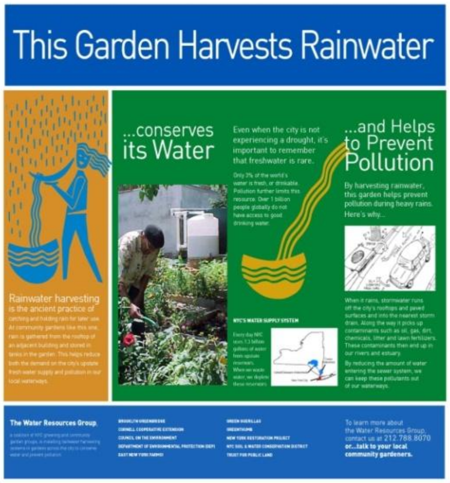

This Garden Harvests Rainwater

Rainwater harvesting is the ancient practice of catching and holding rain for later use. At community gardens like this one, rain is gathered from the rooftop of an adjacent building and stored in tanks in the garden. This helps reduce both the demand on the city's upstate fresh water supply and pollution in our local waterways.

...conserves its Water

Even when the city is not experiencing a drought, it's important to remember that freshwater is rare.

Only 3% of the world's water is fresh, or drinkable. Pollution further limits this resource. Over 1 billion people globally do not have access to good drinking water.

NYC'S WATER SUPPLY SYSTEM

Every day NYC uses 1.3 billion gallons of water from upstate reservoirs. When we waste water, we deplete these reservoirs.

...and Helps to Prevent Pollution

By harvesting rainwater, this garden helps prevent pollution during heavy rains. Here's why...

When it rains, stormwater runs off the city's rooftops and paved surfaces and into the nearest storm drain. Along the way it picks up contaminants such as oil, gas, dirt, chemicals, litter and lawn fertilizers. These contaminants then end up in our rivers and estuary.

By reducing the amount of water entering the sewer system, we can keep these pollutants out of our waterways.

The Water Resources Group

To learn more about the Water Resources Group, contact us at 212.788.8070 or...talk to your local community gardeners.

Sometimes it is necessary to break some basic RWH rules for aesthetic reasons. This rain water harvesting system has an inflow pipe that is pitched upward which means that it must rain enough for the water to fill up the roof washer and the inflow pipe before it can enter the tank. This was done so that the roof washer could be positioned below the brick wall and not interfere with the tree limb or the neighbor’s views.

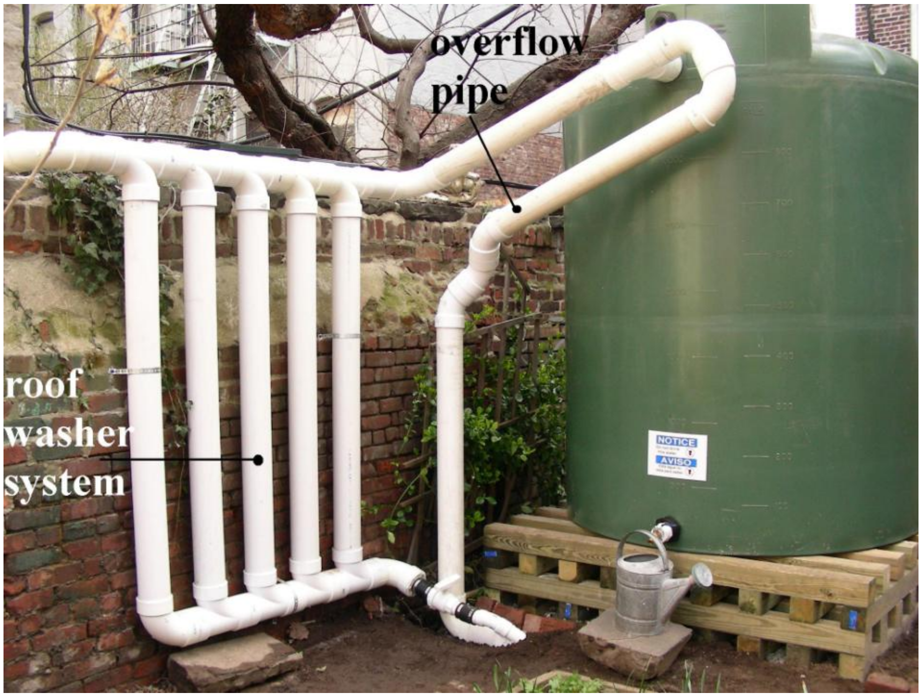

Roofing Material

Roof tops are essential to any rain water harvesting system. Popular types of roofing include corrugated metal decking, shingles, rolled asphalt and cedar shake.

Corrugated metal decking can come in two forms- zinc coated or hot dipped galvanized. Because, zinc coated roofs rust easily, hot dipped galvanized roofs are more popular. These roofs are constructed simply by nailing down with sealing washers to your structure or screwing. Apply gutter sealant on any nail or screw to make sure water doesn't leak from these spots. Metal roofs provide great structural support for long amounts of time. It has the highest efficiency rate when it comes to rain water harvesting because of its waterproof surface. Its only downfall is the noise factor. It gets very loud when rain water clashes with the rooftop. It is an ideal roofing material for non residential structures.

Shingles are small pieces of overlapping wood or slate. Place a ½ inch wide piece of plywood on top of the structure. Start nailing shingles in a horizontal, overlapping fashion starting from the gutter line.

Work your way up to the top of your roof, making sure your next row overlaps the previous row (see picture on left[14]).

Rolled asphalt or modified bitumen roofs occur most commonly on flatter rooftops. A large sheet of asphalt is simply glued down with tar or plastic. Asphalt roofs can either have a granular surface or a smooth surface. Although granular surfaces are slightly less efficient in catching rain water because it is easily trapped in the pieces of ground glass, it is preferred over smoother roofs. Granular roofs wear slowly, resist cutting and keep UV rays from degrading it. Typically, a few years after the granular roof is installed, it is painted with aluminum to lock in grains that may have fallen loose (see picture on right[15]).

Cedar shake or wooden roofs are installed in a vertically overlapping fashion. Wood is the least efficient in catching rain water because it is porous. However, once the wood is saturated, it will no longer absorb water (see picture on left[16]).

Seasonal Maintenance

To ensure your health and safety the Water Resources Group recommends the following seasonal maintenance schedule for keeping your rainwater free of contaminants.

Spring

- Close spigot
- Re-direct roof water from the drain pipe back into tank storage system by replacing the in-line plunger or changing the position of the valve
- Clean any winter debris from gutters, leader inlets and roof
- Repair any leaks in barrel or tanks
- Inspect and clean barrel tops
- Fill system with a few inches of water to check hose connections for leaks
- Inspect rope/cable retainers to assure that barrels are secure
- Be sure that your tank is labeled with a “do not drink the water” sign in all appropriate languages
- To help ensure that children do not drink the water remove the valve from spigot and store with tools

Summer

- Keep the roof, gutters and leader inlets clear of debris, check monthly

- Inspect vents at top of each barrel to insure that they are clean and intact
- Visit your system during a heavy rain or shortly after to check for leaks and overflow problems.
- Any water from the overflow system should drain within 24 hours of a rainfall; if puddles form, you should move the outflow pipe to a more porous site or consider installing a small rain garden
- Inspect pipes and connectors regularly for any damage or disrepair
- Check man-way hatch on top of your tank to make sure it is securely closed

Fall

- Remove plunger from the PVC joint to redirect water into the drain pipe to the sewer
- **Empty water from the entire system- roof washer and tank**
- Open main valve of storage tank and rain barrels
- Open valve on roof washer. Valves should remain open all winter to keep water from freezing in system
- Disconnect leader from system and re-route water as necessary for particular system
- Cover any openings in leaders

Living with the RWH System

Now that you have built your rain water harvesting system, you have helped to conserve one of the Earth's most valuable resources. A rain barrel is not something you build and walk away from. It is important to make sure there are no leaks in your barrel and to drain the first flush after every storm.

Remember to drain your roof washer and tank during the winter and to take out your wintering tee! Freezing water will BREAK the system. The cracks seen in the two pictures below are from left over water freezing and expanding.

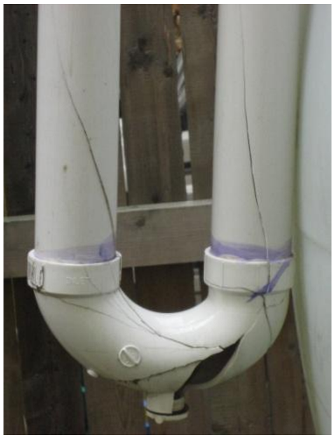

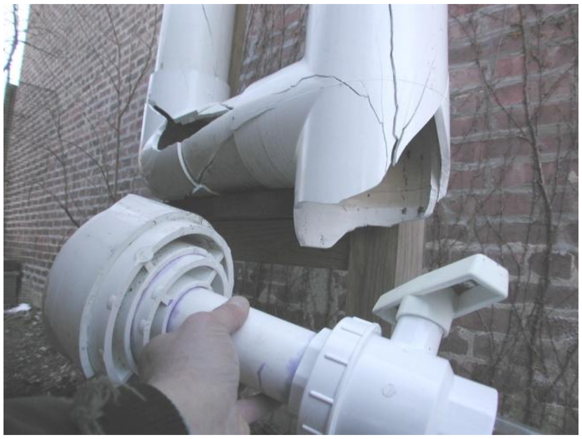

If properly built and maintained, your rain water harvesting system will last many years. Even the smallest tanks will save thousands of gallons of water.

It is essential to learn how to manage your water usage. Seventy percent of residential water use goes to outdoor activities. One third of that water is wasted (see diagram on right[17]). Using pumps or sprinklers to water plants are one of the most inefficient ways to irrigate. Plants only need to be watered at the roots. Sprinklers relentlessly waste water by randomly shooting out water to tree branches or leaves. Drip irrigation systems with outlets placed strategically next to plant roots are the most efficient ways to irrigate.

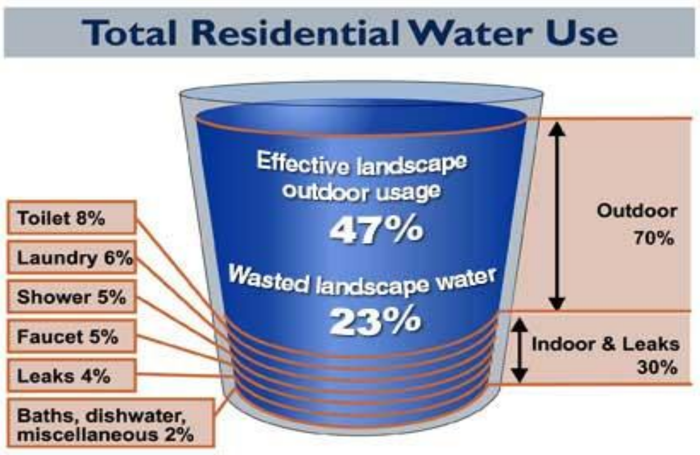

Existing RWH Systems

Organic Gardening Magazine has been sponsoring construction of rain water harvesting systems in community gardens since 2007. Their Waterworks Project funded 30 systems in the past 2 years. Looking at existing rain water harvesting systems and talking to community gardeners may assist you in building your own. Locations in the USA and Canada include:

Alemany Farm - San Francisco, CA
Ashview Community Garden - Atlanta, GA
Aspen Farms - Philadelphia, PA
Boyd Street Urban Farm - Portland, ME
Bradner Gardens Park – Seattle, WA
Brentwood Community Garden - Portland, OR
Children Garden - Camden, NJ
City Seeds Urban Farm - St. Louis, MO
Dias y Flores – New York, NY
The Farm Garden at the Early Childhood Education Center - Columbus, OH
Fremont Community Garden - Sacramento, CA
The Garden of Dreams - New York, NY
Global Gardens - Tulsa, OK
Gloryland Community Garden, Detroit, MI
Growing Green Youth Garden, Buffalo, NY
Guadalupe Montessori School, Silver City, NM
Hope Community Garden - Toronto, ON Canada
Marigold Meadows - Phoenix, AZ
Master Peace Youth and Community Garden - Riverdale, MD
Our Saviour Community Garden - Dallas, TX
The 1100 Block Bergen Street Garden - Brooklyn, NY
Urban Ministry Center - Charlotte, NC
Wasatch Community Gardens - Salt Lake City, UT
Woodlawn Garden - Portland, OR
Xochiquetzal Peace Garden - Chicago, IL

For more information on these gardens, visit their website at:
http://www.organicgardening.com/feature/0,7518,s1-2-10-1531-1-1X2X3-4,00.html

This guide can be found on GrowNYC’s website at:
http://www.GrowNYC.org/openspace/rainwater

10 Ways to Conserve Water

1. Turn off the faucet when brushing your teeth, washing your face and shaving.

2. Fix any leaky sinks, toilets or showerheads.

3. Wash only full loads- for clothing and dishes.

4. Do not run water to thaw meats.

5. Water plants at the roots.

6. Take showers instead of baths.

7. Replace showerheads and sink aerators with low flow ones.

8. Use a pitcher to store cold water in the refrigerator instead of running the tap every time.

9. Water plants during early mornings and late afternoons to reduce evaporation.

10. Do not flush the toilet unnecessarily. Dispose of bugs, cigarette butts and tissues another way. [18] [19] [20]

The Water Resources Group (WRG), a coalition of NYC greening & community garden groups is dedicated to the preservation of NYC water resources through gardening, ecological design and education programs. WRG is installing rainwater harvesting systems in gardens across NYC to conserve water and prevent pollution. For more information on how you can help conserve water in New York City, visit the Water Resources Group website at www.waterresourcesgroup.org.

If you have any questions regarding the construction of your rain water harvesting system e-mail Lenny Librizzi, Assistant Director of Open Space Greening at GrowNYC at llibrizzi@GrowNYC.org.

Endnotes

---

[1] Hasse, Rolf. “Rain water reservoirs above ground structures for roof catchment.” Welcome to Energy. 19 November 2003. 20 June 2008 <http://ces.iisc.ernet.in/energy/water/paper/drinkingwater/rainwater/introduction.html>

[2] “An introduction to rain water harvesting.” The WWW Virtual Library on Urban Environmental Management. 9 May 2008. 20 June 2008 <http://www.gdrc.org/uem/water/rainwater/introduction.html>

[3] 10,000 Rain Gardens. 2008. 20 June 2008 < http://www.rainkc.com/>

[4] “Water.” The New York Observer. 21 April 2008: G 10.

[5] “An introduction to rain water harvesting.” The WWW Virtual Library on Urban Environmental Management. 9 May 2008. 20 June 2008 <http://www.gdrc.org/uem/water/rainwater/introduction.html>

[6] “Rhizome Collective.” Eco Warrior. 28 Nov 2007. 11 July 2008 <http://hurricanecandice.wordpress.com/2007/11/28/rhizome-collective/>

[7] “Oak Whiskey Barrel.” Rain Barrel by Aaron. 2003. 11 July 2008 <www.ne-design.net/oak-whiskey-barrel.html>

[8] “Products.” RainDrops Cisterns Rainwater Harvesting and Storage. 23 June 2008. 11 July 2008 <http://www.raindropscisterns.com/cat_photo_3.php>

[9] “Rain Garden.” Wikipedia: The Free Encyclopedia. 12 June 2008. 20 June 2008 <http://en.wikipedia.org/wiki/Rain_garden>

[10] “Bioswales.” Wikipedia: The Free Encyclopedia. 3 April 2008. 20 June 2008 <http://en.wikipedia.org/wiki/Bioswales>

[11] Chapin Watermatics Inc. 11 July 2008 <http://www.chapindrip.com/index.php>

[12] Drip depot. 11 July 2008 <http://www.dripdepot.com/>

[13] Drip depot. 11 July 2008 <http://www.dripdepot.com/>

[14] “Roofing.” Sam Siding Home Improvement. 20 June 2008 <www.samsiding.com/Roofing.aspx>

[15] Lexis Coatings. 20 June 2008 <http://www.lexiscoatings.com/>

[16] “Cedar.” Hynes Home Improvement: The Sign of Quality. 11 July 2008. <http://www.hyneshomeimprovement.com/roofing-cedar.html>

[17] “Hydrapure Water Solutions.” Hydrapure. 11 July 2008. <http://www.hydrapure.net>

[18] Lenz, Ericka. “10 easy ways to conserve water.” Gaiam Life- Your Guide to a Better Life. 20 June 2008 <http://life.gaiam.com/gaiam/p/10SimpleWaystoConserveWater.html>

[19] “10 ways to conserve water.” Etowah Water and Sewer Authority. 2 February 2008. 2008 June 2008 <http://www.etowahwater.org/FAQs/10_Ways_to_Conserve_Water/10_ways_to_conserve_water.html>

[20] “Ten Ways To Conserve Water Around The House And Help The Environment.” The Good Human. 2 July 2007. 20 June 2008. <http://www.thegoodhuman.com/2007/07/02/ten-ways-to-conserve-water-around-the-house-and-help-the-environment/>

This manual was put together by Julia Leung, Lenny Librizzi, Assistant Director of Open Space Greening at GrowNYC and Lars Chellberg.

After reading this manual, please fill out this feedback form and mail it to:

GrowNYC
51 Chambers Street, Room 228
New York, NY 10007

On a scale of 1 – 10, 10 being the most agreeable, 1 being the least, assess the following statements:

This manual is easy to understand.
1 2 3 4 5 6 7 8 9 10

This manual contains clear instructions.
1 2 3 4 5 6 7 8 9 10

This manual helped you make your rain water harvesting system.
1 2 3 4 5 6 7 8 9 10

You understand how a rain water harvesting system works.
1 2 3 4 5 6 7 8 9 10

You are no longer interested in rain water harvesting.
1 2 3 4 5 6 7 8 9 10

You would recommend this manual to your friends.
1 2 3 4 5 6 7 8 9 10

You know the importance of water conservation.
1 2 3 4 5 6 7 8 9 10

What are the strong points of this manual?

What are the weak points of this manual?

Chemtrails Destroyed !

How to Build A Cloudbuster to disperse Chemtrails and Improve the Weather

Deutsch

Editor's Note: Since this article was first written around the end of 2001, a grassroots network of thousands of people has grown who are using and developing this powerful technology in a number of ways to convert the negative energies that we are being subjected to, into positive ones. Many thanks to genius inventor Don Croft.

ORGANITE: This is the new word for the basis of Don's creation. It's a mixture of metal filings (in-organic substance) and resin (organic substance) and a crystal. It converts negative orgone energy to positive orgone energy. To be effective over a period of time, it needs to be in earth or running water to cleanse itself of the negative energies it collects.

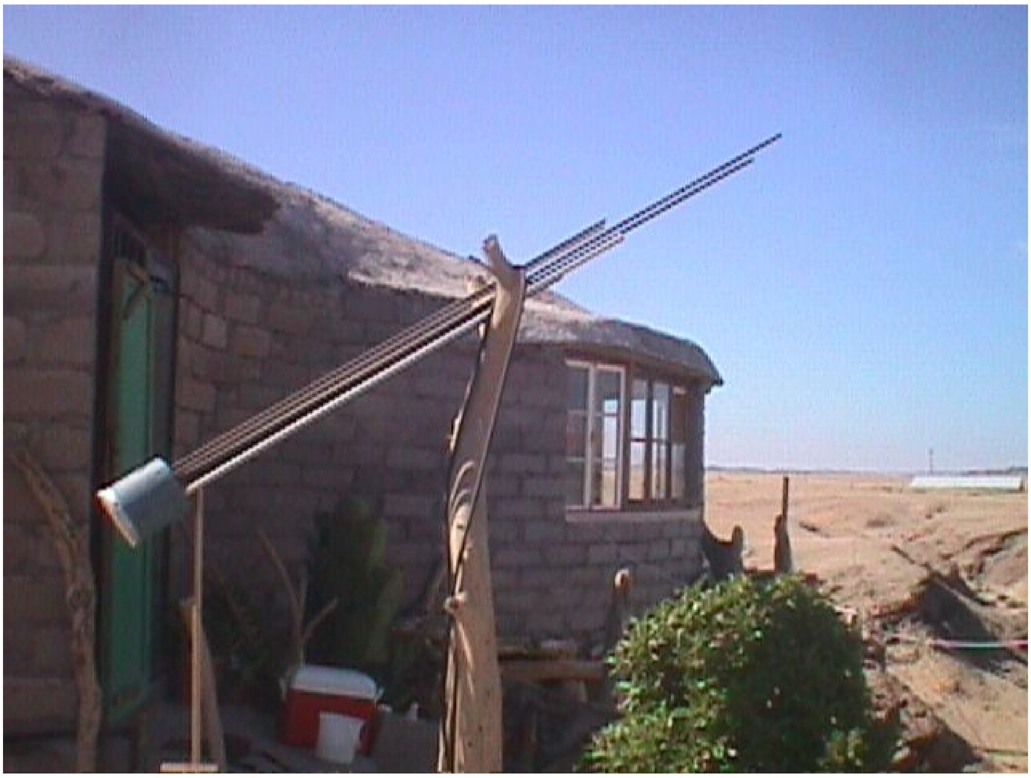

This CB was made by Gert Botha, of Namibia, Africa, from scavenged materials, so the result is slightly modified from the instructions. This cloudbuster was used in Africa's Namib Desert and brought several days of continuous thunderstorms after a 12- year drought.

Definition of a contrail: A contrail occurs when a plane travels at a high altitude (about 30,000 ft. or more) and compresses the air into a water vapor or ice crystals through jet engines or the wing tips pushing through the air. This trail disappears after a minute or so due to evaporation, because it's water.

Definition of a chemtrail: The term "chemtrail" is a relatively new word, which appeared in the last few years along with the appearance of chemtrails.

A chemtrail is very different from a contrail. At first a chemtrail might look a bit like a contrail. However, instead of disappearing like a contrail does, a chemtrail just keeps spreading out and forming a hazy cloud bank. These trails traverse the whole sky and stay for up to around five or even eight hours. They have been known to turn what was originally a clear blue sky into a grey haze. (Note: As of Nov 2001 there are already at least several dozen cloudbusters in seven countries, so if you see chemtrails that don't last anywhere near this long, there may already be a cloudbuster in the area).

It is believed chemtrails are composed at least in part by an oil based substance of unknown content that is actually being sprayed from jets on populated areas. It has also been found that some chemtrails contain unknown biological components. There are many reports of increased diseases, especially lung diseases, in areas where there are chemtrails.

Chemtrails are often seen to be sprayed by white planes with no markings. Sometimes you will see these planes give off no contrail at all, and immediately after the same kind of plane in the same area will start to leave a thick chemtrail. At other times, distinct short gaps in chemtrails have been photographed, as though the plane ran out of chemicals and then had to start spraying with a new supply (as in the photo).

For information on Chemtrails see the sites listed at www.relfe.com/weblinks.html#chemtrails

See also satellite photos of chemtrails at
www.rense.com/general4/chmgb.htm
www.rense.com/general3/massive.htm

History of the Cloudbuster

The Wilhelm Reich cloudbuster (CB) was well known for its ability to create rain. But a problem existed with its continued use, in that the deadly orgone energy that it absorbed could in fact hurt the operator if it was not correctly "drained off". This "draining off" of deadly orgone energy was accomplished by 1. Connecting the cloudbuster to a body of running water (difficult) or 2. Connecting the cloudbuster output to an orgone accumulator. However, the orgone accumulator would soon become saturated with the deadly orgone energy which in turn caused more problems for the operator. So the cloudbuster, as good a technology as it is, has lain dormant for many years only being used by such professional weather engineers as Trevor James Constable, among others.

"There is a principle which is proof against all information, which is proof against all arguments, which cannot fail to keep a man in everlasting ignorance; that principle is contempt, prior to investigation." -HERBERT SPENCER

2001 - The Cloudbuster Re-Invented

Don Croft, a patriot, alternative scientist, researcher, inventor and freedom fighter has succeeded in striking the first blow against the ongoing chemtrails program that is assisting in the destruction of our country and the enslavement of its citizens! . By utilizing the research of Wilhelm Reich, assisted by his partner Carol and having help from God almighty, Don has created a cheap, portable and easy to build device that consistently destroys Chemtrails and heals the atmosphere. The Reich/Croft " chembuster" is the answer to these ongoing attacks by the forces of darkness. Unlike the original cloudbuster, the chembuster actually changes the deadly orgone energy to good orgone energy and so does NOT become saturated or dangerous to the operator.

Join The Battle - Now !

It is human nature to sit on the sidelines and think that one person cannot change the things that are happening to our country. You now have no excuse for inaction. For the cost of a few dollars and a few hours of your time YOU can make a difference. With this device a mere few thousand patriots can SHUT DOWN the multi-billion dollar chemtrail program. The Reich/Croft chembuster cannot do any damage. It can only help and heal. Its up to you to join the fight to save America.

Chembuster Construction Details

The following was written by Don Croft, who lives in vagabondage
www.worldwithoutparasites.org

Don is using Orgone energy, which was the name given to ch'i, ether, life force, etc., Dr Wilhelm Reich.

He does this out of respect and gratitude for Reich's life's work and sacrifice.

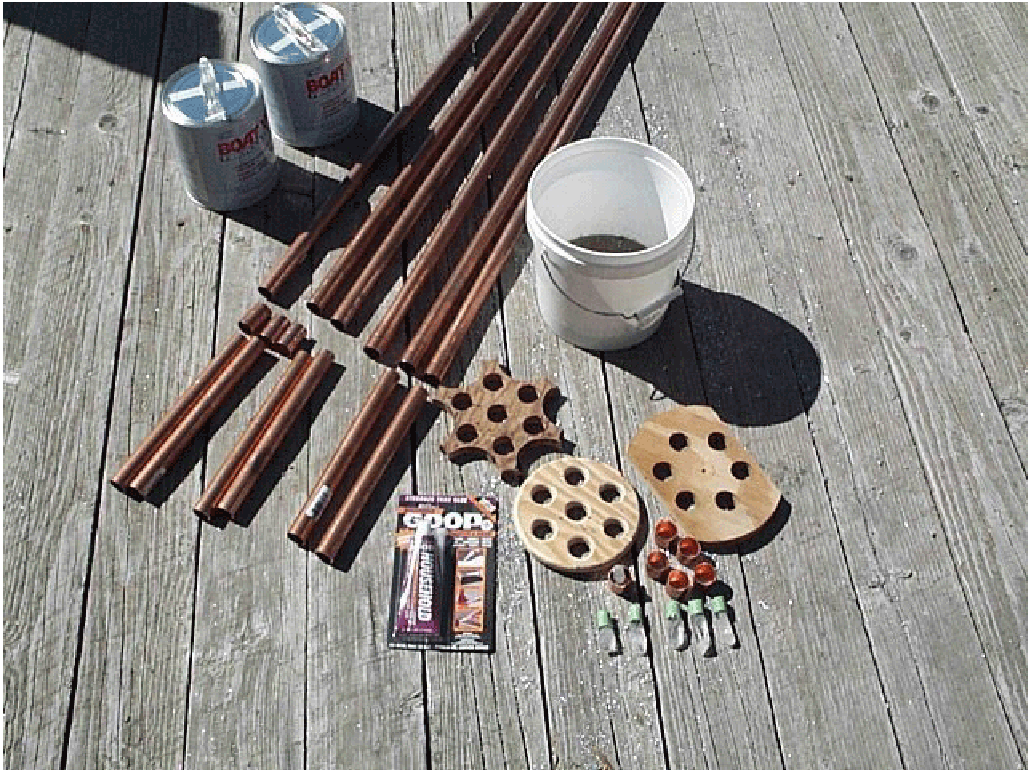

Cloudbuster Materials

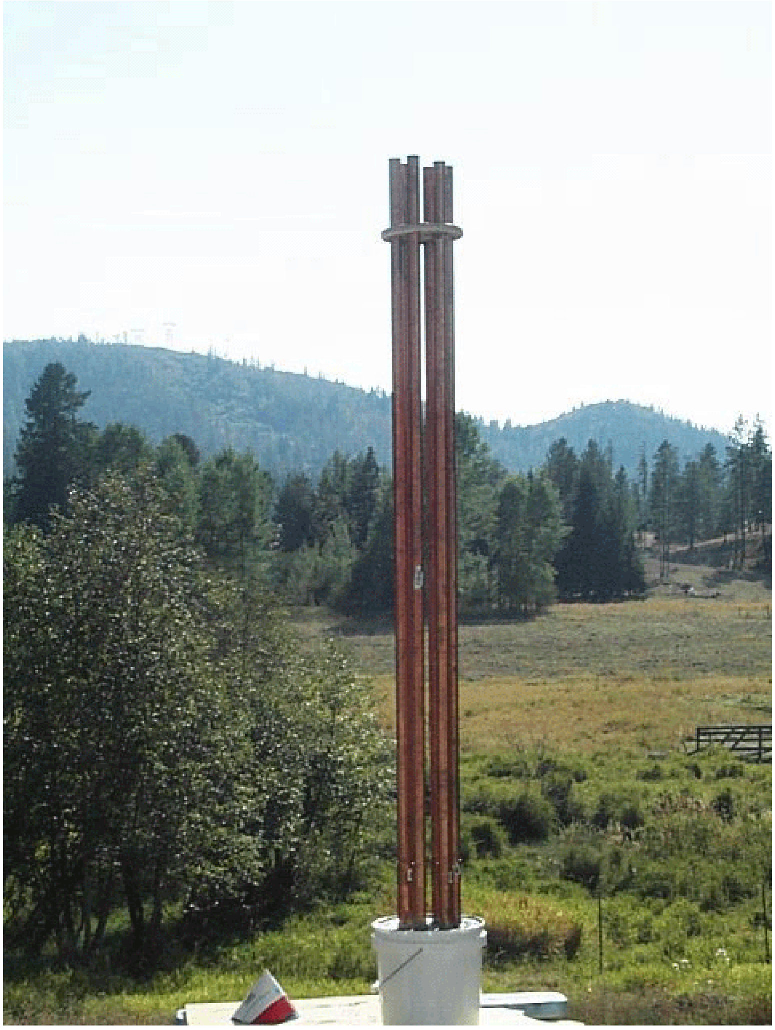

Finished Cloudbuster

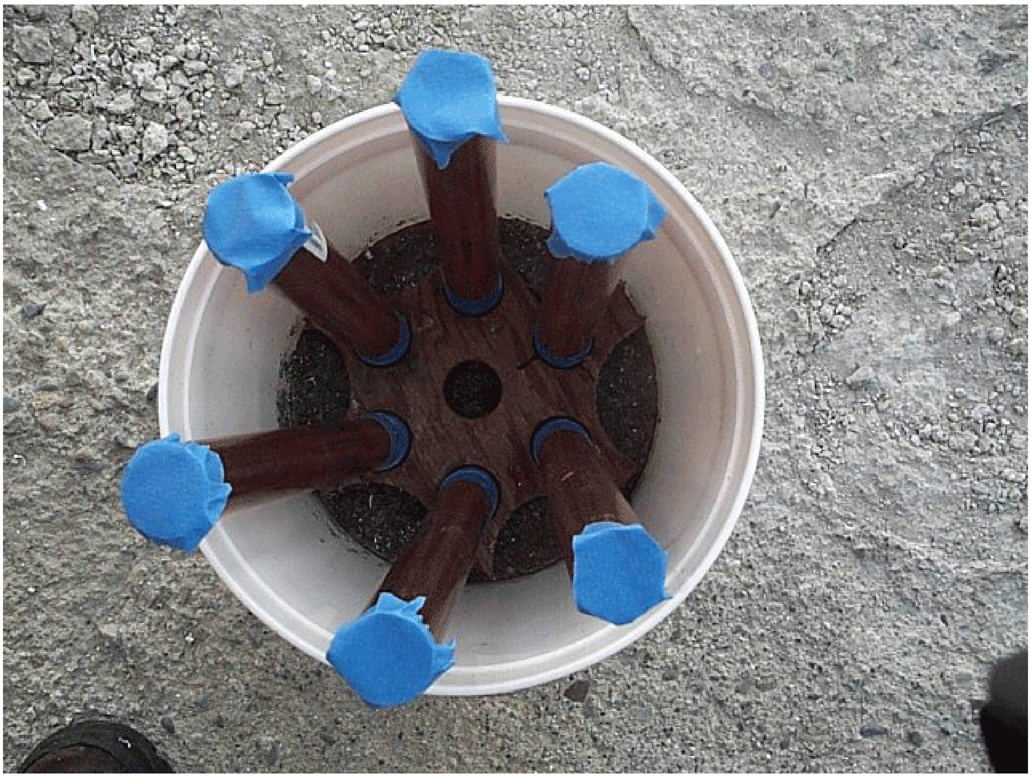

Set pipes in bottom spacer

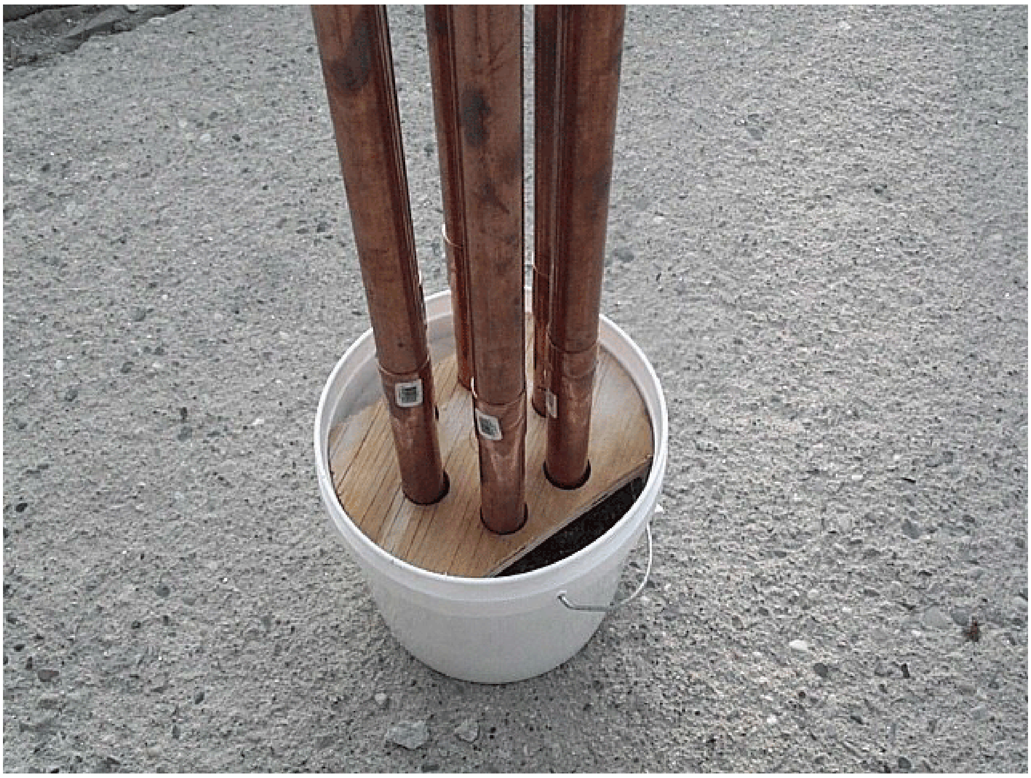

Middle Plywood Spacer

CLOUDBUSTER DIRECTIONS

Carol and I got these specs by dowsing, and they work very, very well, though may not be the last word, if you get my drift. Let us know at zapper16@earthlink.net if you can successfully modify it.

BUCKET: Base: 9" in dia., 9" deep. I use two-gallon plastic buckets from the Home Depot paint department, but two-gallon food buckets, available from restaurants, or two-gallon spackle buckets from drywall contractors are fine. I leave the buckets on, as the handles make it easier to move the finished cloudbuster.

COPPER PIPES: Six standard 1" copper pipes, 6' long, open at one end, the capped ends submerged to within an inch or so of the bottom of the bucket before casting. For portability and shipping, you can use 12” sections of pipe in the base, adding five-foot lengths by putting couplers between the 12” and five-foot pieces. You won’t need
to buy as much copper pipe this way, too.

CRYSTALS: I glued a double-terminated quartz crystal, about 2" long, into a .75” section of garden hose, and one of these devices is glued in turn inside each of the six copper end caps, pointing in the same direction as the pipe will be after the cap is glued on. I use 'Goop' glue. I taped around the cap where it joined the pipe to prevent resin from leaking into the cap during casting.

METAL PARTICLES: Get some metal particles from a machine shop, recycling yard or a place where they saw a lot of aluminum (sign shops and aluminum fabrication shops are good places to look). I use aluminum for weight considerations, but ANY metal will do. Very fine particles (such as filings from a key-making machine)
aren’t quite as good as ones that will easily go through a window screen & slightly bigger ones are okay. There really is a wide margin of suitability! .

PLYWOOD: Template 1 - 3/4” exterior plywood is used to make the base for the copper endcaps/ pipes. I arrange the six pipes evenly around a 2 1/2” radius circle, which makes the pipes 2 1/2” apart on centers. A flat, 1 1/4” auger bit drills a suitable hole for each end cap.

Template 2 - Make another piece to fit inside the rim of the bucket. Remember to cut a section away so you can fit your fingers under it to remove it from the bucket. This piece is not part of the finished CB, just used to get the spacing for the pipes right during construction. Cut the holes in this with a 1 1/8” flat bit so that it will fit well over the six pipes. This piece will hold the pipes in position while the second batch of resin hardens.

Template 3 - I make a circular plywood spacer for the top ends of the pipes, 4” radius with 1 1/8” holes on the same 2 1/2” radius circles for the pipe ends. This keeps the pipes parallel with each other and fairly rigid so you can point it at unholy things like predatory entities and City Hall ;-)

CLICK HERE TO VIEW THE TEMPLATE

RESIN: I use polyester resin (used to build fiberglass boats) but epoxy also works, as does envirotech resin from www.eti-usa.com (this sets up very quickly, so you need to be proficient with it). (If you are going to construct many of these units, try to purchase your resin in 5 gallon buckets over the internet, possibly from a surfboard or boat manufacturing supply company. ACE hardware for $12.95 per gallon. Home depot gets $24 per gallon for resin.

On a level surface, pour an inch or so of catalyzed resin into the bottom of the bucket and stir in metal particles until the surface is uniform and flat.

The ratio that seems to work is approx. one part resin to one part metal particles.

Put the base wood Template 1 onto that before it hardens, so that it just touches the surface and doesn't sink in much. Pour a half gallon of catalyzed resin into the bucket after the six pipes are set into the holes in the base template. Use the dowel to stir in handfuls of metal particles in the same ratio as before. Note that the wood Template 1 should not be pushed down to the bottom of the bucket with the resin squishing over the top.

Expeditiously move the Template 2 down on to the six pipes until it is snugly within the rim of the bucket. Put the top Template 3 onto the top of the pipes and move them until the pipes are parallel. After the resin has hardened, remove the second template and set it aside. This is only used during construction, as I mentioned.
You can finish off putting the resin/metal mixture in now in two stages if you want, up to within a half inch of the rim of the bucket. Now it’s finished (wait until it hardens before you shoot anything ;-)

A MONEY SAVING SUGGESTION, ALSO FOR BETTER PORTABILITY:

Cut six pipes, 12" long, and put the crystals/endcaps on them for the base. Buy six joints which allow a pipe to come into them from both directions and stop at the middle. This allows you to make the rest of the pipe assembly from three pipes, 10' long, cut in half. The finished product leaves you with only one piece of pipe, four feet long leftover.

When positioning the 12" pipes for pouring the second layer of mix, temporarily tape the top ends of the 12" pipes so no metal particles are dropped in them by mistake.

After the mix is made, take off the tape and push the rim plywood piece down onto the pipes so that it fits into the rim of the bucket.

Put the spacers, then the 5' pipes onto the 12"base pipes, then put the third (top) plywood piece onto the tips.

The pipes will look skewed, most likely, so just twist the top plywood piece around until the pipes line up parallel. As you do that, the rim plywood piece will move around. You need to do this expeditiously, though not in a hurry, so that the resin doesn't harden before you're done.

We experimented with 5' extensions to the pipes to increase the range. It had a strong effect on the moon and on the upper atmosphere of the planet for hundreds of miles. We also apparently shut down the weather warfare facility under Homestead Air Force Base while we were in the Keys by leaving the extended cloudbuster pointing at the dead orgone field above it for a week.

We found that we don't need to point it in any particular direction in order to clear the atmosphere of bad orgone, and the longer it's left standing, the nicer it feels for a longer distance from the cloudbuster (many miles). We just stand it on its bottom, pointing straight up. We did once eliminate a belt of smog which was just off the coast by aiming at it for a few minutes, so direction can be a factor.

We also severely alarmed the operators of a ufo that was in the vicinity. It instantaneously shot out into space, where it remained in the same position relative to us, so the basic 6' long pipe configuration apparently only affects the atmosphere.

Do be discriminating which UFO’s you aim at, though. If they are not predatory and you make them mad, they may pay you back by messing with your carburetor or something, as they did to us once. The really advanced ones are not susceptible to the cloudbuster, as they apparently don’t need a deadly orgone field to operate. Our take is that all but these use nuclear fusion or some other noxious, old-paradigm energy source. It’s open season on the B Sirian craft and other predatory ones, as far as we’re concerned. They can’t hurt you, by the way, if you aren’t afraid of them. Universal law restricts them more than it does us. We can pay later for breaking it, but they are just unable to break it, apparently.

Cloudbusters destroy chemtrails consistently. It may take a few days for your cloudbuster to activate the environment enough to do this, but after that point the spew will disappear within seconds of leaving the UN jets and dissipated trails that drift within range of the cloudbuster will also disperse, but more slowly. The ones that are made up of solids take longer to dissipate, but we’ve found that these only make up about 10% or so of the spewing episodes.

The evil intent of the UN is most apparent to me in the fact that Lucis Trust (formerly Lucifer Trust) publishes all of its official literature and dictates its philosophy. Only this belief system, which closely resembles the doctrine of the Nazi party, could condone the rampant genocide which is now being carried out by the UN.

Dr. Reich’s cloudbusters were extremely dangerous when in operation since he had no orgone generator attached to them. Bad orgone does carry a charge which can quickly drain one’s life force away when it’s concentrated in one spot. It's slightly similar to a strong positively charges static electrical field. When cloudcover is unproductive or unbalanced (too much lightning, for instance, or too much rain) the cloudbuster busts it up. We plowed a furrow through a line of artificially created thunderstorm in June, 2001. Carol said the square-shaped thunderheads were chock full of predatory ET's, bent on mutilating animals at the time.

It also stops droughts. Overall, it just balances the weather, I think it sort of mediates between the ground and the atmosphere somehow. The one in Namibia finally brought heavy rain to the desert there. It happened 9/5/01, and continued for several days. Gert, the fellow who made the cloudbuster and reported the occurrence, told me that this had simply never happened in the Namib desert before to the best of anyone's knowledge. It's arguably the driest region on the planet.

WARNING: If you touch the pipes when they’re drawing in strong unbalanced or dead orgone and don’t put your hand on the base after that, you may experience headaches, even into the next day, as a friend of ours did recently.

Mind you that our efforts, though somewhat pioneering, should be seen mainly as a springboard for further developments rather than the last word.

Somebody's starting a list for cloudbuster aficionados shortly.

SUGGESTION: If you want even more rain, try putting water on top of the material in the bucket and/or hosing down the whole apparatus daily.

The HHG (Holy Hand Grenade)

The HHG (Holy Hand Grenade) is a small orgone generator that will correct imbalances in energy fields.

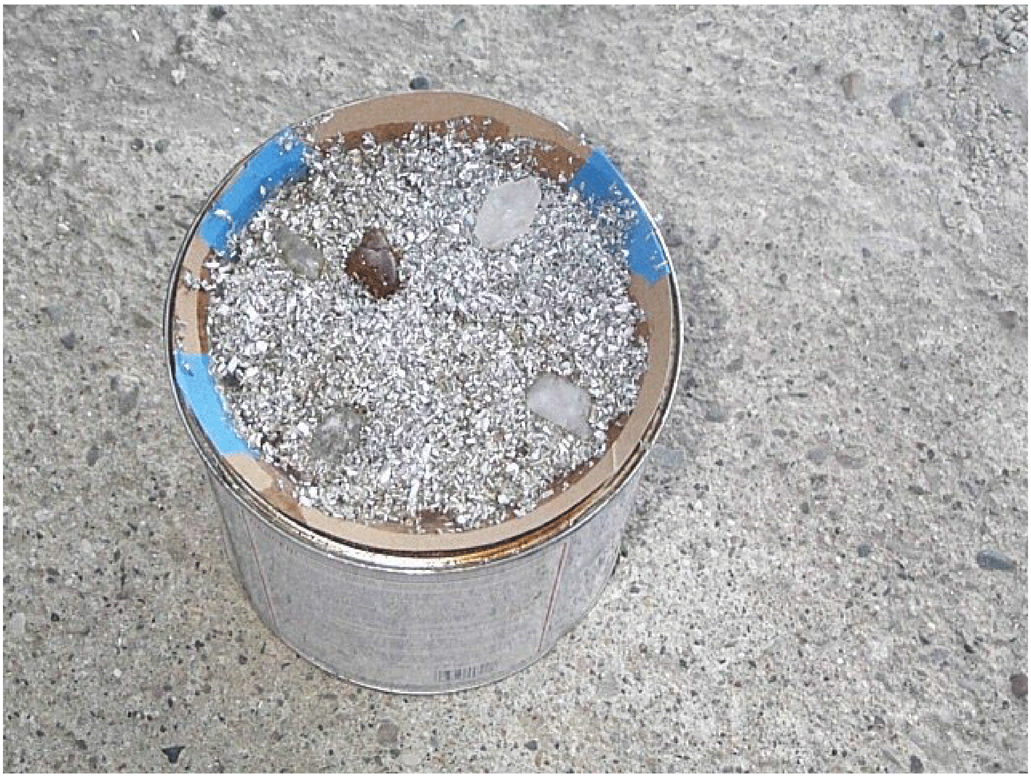

HHG almost done

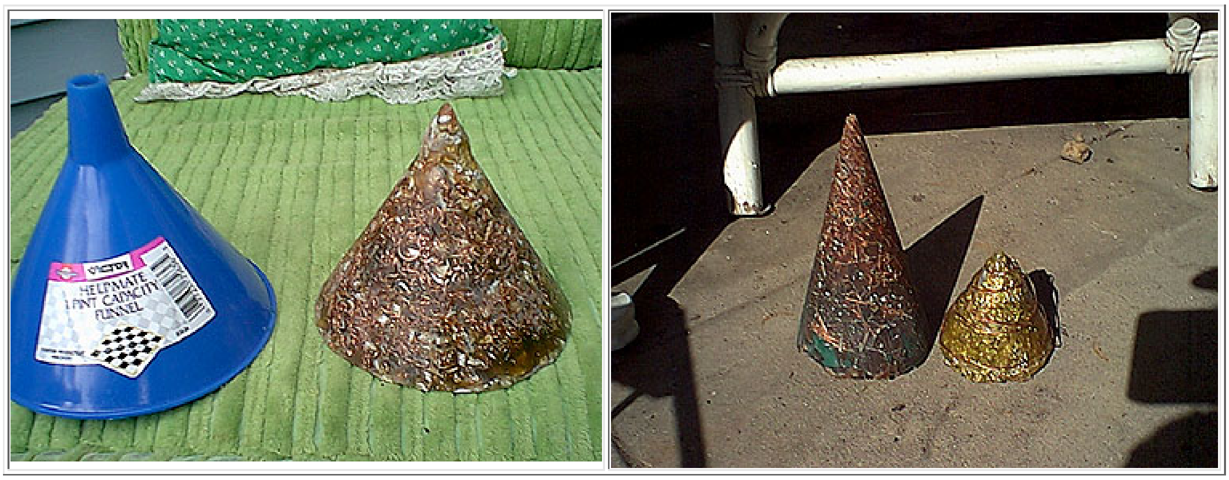

HHGs

HHG Construction

If you want to clear out the unpleasant entities from your immediate environment, expel parasitic/ predatory ET's from your auric field, neutralize the energetic effects of chemtrails, etc., just make one or more of these and put it on the ground or in your house, according to your intuitive promptings.

My partner, Carol, and I have found ! them very effective for minimizing the deadly orgone fields around nuke plants, even the ones built on natural vortices (sick puppies did that intentionally). They also heal and strengthen all natural vortices and man-made ones, like the Caribbean Blue Holes. Life doesn't have to be difficult and years of training and discipline are no longer necessary if you want to do big magic, to which I can bear witness. In the new (present) paradigm one's character and intention count more than training and institutional acclaim.

There are plenty of people who are nice, but these aren't generally smart. There are about as many people who are really smart, but they tend not to be very nice. Have you noticed that more nice people have been 'smartening up' lately? I would say that that's happening more than smart people are getting nice. Does this mean that the 'meek
are inheriting the earth'?

You might think I'm digressing, but the Y factor is very important when working with orgone. 'Y' stands for 'you.' Orgone is attracted to nice people and is repelled and even changed into deadly and/or dead orgone by unpleasant, selfish, fearful people. For example, a Joe Cell might not work for Joe Stalin, Count de Sade or Henry Kissinger. (I really think it's okay to draw attention to predators). Orgone accumulators are susceptible to depletion and pollution by unhealthy orgone, but HHG's (orgone generators) draw in the bad stuff, transform it into healthy orgone, and send it back out. Miraculous! Astonishing! Fabulous! Thank you, JESUS! Predatory people are challenged to lighten up or leave the vicinity of orgone generators. We're about to see the effect of one in a federal kangaroo courtroom shortly (we won't actually be present). Speaking of predators…[update: this device was getting the desired effect, but the judge and prosecutor contrived to keep it out of the courtroom after that and managed to railroad the defendant into prison, unfortunately]

Most people will be content to have a HHG in the home to keep the energy clean and healthy there. It doesn't matter where you place it to do this. A few people have the skill and inclination to find vortices, so here's some advice based on what we've been doing:

If you can dowse with a pendulum or rods, etc., that may be the best way to find the right spot to put a HHG. It's especially effective when placed on a ley line and especially on the junction of two or more of these, but this isn't necessary. Every room and building has its own dynamics, so you could dowse for the correct spot based on that.

If your house is built on one of these leyline junctions, though, lucky you! More often than not these days the vortices produced by these nodes are distorted and even backward-spinning. The HHG very quickly heals and redirects these on its own. You don't need to do anything else (don't hate me because I've given you an easy, inclusive way to do this ;-). If you're a weird science guru or wannabe you're probably not reading this anyway). Most of us just want to have a nice feeling in the house and immediate vicinity. You can put a HHG anywhere for this.

One thing that has worked for us in locating major vortices: think of the place from which the most degenerate, unspeakable things have been promulgated and you may find that it has been built over a vortex. Examples: The Federal Reserve Bank of New York/World Trade Center; the Pentagon; NSA headquarters, Montauk, New York;
Jekyll Island Hotel, Georgia. We already did those, but there are a few more of these left to do (as of May 12, 2001) such as the city reservoir of San Diego, Bohemian Grove, near San Francisco, Denver Airport (the new underground facilities have nothing to do with airplanes ;-) a spot in Salt Lake City, 3 in LA, etc. This really is fun but if you do one of these, make sure you can have a day or so afterward to recuperate. Also be sure to ask for divine protection ;-) We'll be visiting those spots shortly. We develop sort of energetic callouses when we do it often enough, so that we don't miss a beat afterward. [update: we finished them at Bohemian Grove, California, which is a satanic gathering site for all of the top world order 'leaders,' on June 18, 2001]

If there are places in your area at which evil has been done on a large scale, put an HHG there, too, to correct the imbalance. We put one in the witch cemetery in Salem, Mass, for instance (because of the evil done TO the witches, of course). There are plenty of places where atrocities were committed by the alleged gov't on people of color, unfortunately, and every populated area (except in Utah) has been violated with chemtrails on a daily basis lately. Pollution = deadly/dead orgone and is the sure
sign that something is being done to violate/destabilize that part of the planet. It's so easy to turn that around with a HHG and/or a cloudbuster (let me know if you need cloudbuster directions). Zapper16@earthlink.net.

ENOUGH EXTEMPORIZATION HERE IT IS:

INGREDIENTS FOR THE HHG:

A plastic funnel or cardboard party hat (tape the seam so the resin doesn't leak out) To form a point at the end of the funnel, saw off the spout and form a point with 2" wide masking tape, being sure to prevent leaks. The funnel will allow you to make several from this mold. Party hats are better if you just want to make one. When I make them with party hats, I leave the hats on the finished HHg.

A couple of cups of aluminum particles (OR a pint of BB's). I use aluminum from the crosscut saws in shops that use extruded aluminum a lot, such as large sign shops, fabrication plants, etc. Some plastic and/or wood in the sweepings is fine. I pick out the cigarette butts for aesthetic considerations ;-) Larger towns and cities have recycling businesses that have the aluminum particles in large bins, which they sell cheaply.

A quart of polyester resin or epoxy resin (this will encourage you to make at least two ;-) More politically correct resin is available from Envirotech, at www.eti.com.

5 double-terminate quarts crystals, at least an inch long. Cloudy and funky are fine, even if an evil person has handled them ;-) The mere proximity of an HHG to any crystal will clear it faster and better than all the combined resources of the most powerful new age gurus or shamans on the planet. I'm just the messenger here, so don't hate me. The meek are inheriting the earth, after all. Others have told me that orgone generators remove the cloudiness of quartz crystals. Mine are already clear, so I haven't seen it
happen, myself. One of mine did turn clear when I went to the Bahamas. I'm pretty sure the Atlantean energy at the Mars Bay blue hole did that.

Global Light Network will sell you 5 appropriate crystals from Arkansas for $7, including shipping. Contact David for these at davdart@texas.net

The plastic funnel needs to be big enough to make a 4" high cone. The bigger the cone, the more powerful the HHG. I encourage you to experiment with different configurations. Let me know what you find out, please, since you'll likely surpass my own efforts thus far. We made one 7" high for the vortex at the secret underground base near San Diego from which a few investigators never returned. We deployed that one on 6/16/01. Boy, are they mad at us! We'd never been surveilled so much before that day.

Set the funnel, point down, in a can or jar.

Put some aluminum particles or BB's in the bottom, enough to make the vertical crystal stand on its own. I put the better defined end of the double terminated crystal toward the point of the cone.

Pour some resin on top of that and let it soak down through the metal particles.

Put enough aluminum particles or BB's to cover the rest of the vertical crystal and pour enough resin in to cover that.

Place the other four double terminated crystals horizontally, aligned 90 degrees apart so that the four compass directions will be covered when you're ready to deploy the HHG.

Make a mark on the funnel to indicate where the point of one of the bottom four crystals is if you want. This is marginally important. I never do that, myself.

Cover the four crystals with aluminum particles or BB's and cover that with resin. At this point, make the funnel level, either visually or with a spirit level (the ruler with the bubble thing built into it). You don't need to get it perfect, but perfect is a little better than not. I never fill it to the edge of the funnel, since this makes it harder to finish off and is pretty messy.

If you want to make a really big one, make a paper cone (taping the seam) out of a big piece of thin poster board and tape that into a large funnel. We'll be putting a big one in the reservoir in San Diego [done, June 01] and at the huge vortex southeast of Bimini soon. Those vortices are particularly big and distorted for some reason.
Remember that lining up with the compass points is not critical. I put small parachutes on the ones I use in deep water locations so that they'll be upright when they reach bottom.

Make a mark on the flat part of the hardened material now so you'll know how the four bottom crystals are aligned, since you can't see them any more.

Take the now-usable but crude HHG out of the funnel now and file the material at the little end down to create a point on the cone. I use a rough belt on my belt sander for this, but feel free to do it the hard way. You can file the edges of the bottom a little if you don't want to cut yourself with it later.

Having made the mark on the bottom, you can line it up to the compass wherever you put it. Put another mark near the bottom of the cone surface to make this easier ;-) This is optional, as it works fine if not lined up with the compass, too.

You can make it a little better by getting fancy crystals if you want. You can probably improve it in other ways, in which case, please inform me at zapper16@earthlink.net. I won't make these for money but feel free to do so yourself. I've capitalized this way on the pioneering work of other people.

To juice it up even more, make a copper coil on the outside of the funnel or party hat, being sure to wind the wire clockwise toward the point (as viewed from the bottom of the cone toward the point. Do 7 turns or so. Install this in the cone before putting the other ingredients in. This suggestion! is from (through ;-) our masterful friend, Linda
at www.spiritherbs.com, who is also a Wingmakers aficionado.

HHG's are also available from www.powerpyramids.com

IMPORTANT UPDATE March 15 2003: The cloudbuster and HHGs need to be in touch with the ground to be most effective. If you cannot place them on the ground, connect them to the ground with copper wire.

THERAPEUTIC HHG (not for everyone, but body workers find them helpful, as do meditators and energy experimenters)

Do all of the above in a smaller funnel . The resulting cone should be under 2" or so, simply because this is easier to use. You'll need non-ferrous metal, preferably aluminum because it's easier to tool.

Instead of quartz, get an amethyst gemstone for the top and a garnet gemstone for the bottom, center (we use 9mm amethysts and 6mm garnets). In between, put a strong neodymium magnet. North side down is customary, but when a single magnet is used, we've found that it doesn't matter which pole is toward the body. This is true only when an array is used which causes one pole to be dominant in the field. Be sure to get some aluminum material around the magnet before pouring more resin in so it doesn't jump toward the can holding the funnel. BB's aren't appropriate for this, since they'll just ball up around the magnet and drive you berserk.

You need to allow enough space in the pointy end to round off the point to about a half inch diameter without assaulting the amethyst during the rounding process. I use a belt sander to round off the top but you can do it with a file or a rough rock if you're a Luddite.

Rounding the top makes it better than the big HHG for directing energy through the bottom into the human body, especially at chakras. This one is fun to wear over the crown chakra or sixth chakra. A new meaning for the term,'party hat.' It does seriously get rid of pain and will speed healing to any part of the body. It also prevents nightmares and promotes lucid dreaming and meditation. When I'm meditating, some accuse me of sleeping. The occasional snoring is done just to jerk them around - self righteous twits!

If you want to put a copper, cone-shaped spiral in this one, contact me for further instructions (thanks, Linda!).

I WISH I KNEW HOW THIS WORKS

Thanks a lot for the account of your farmer neighbor, Ber. I saw recently that ALL of the madrone and myrtle trees along I-5 in Oregon are dead now. I saw a bright green fungus on all of the deciduous trees in that area during the winter of 99-00, which I attributed at that time to the mycoplasma of the ever-present chemtrails. I read that nearly all of the foals born in Kentucky this year were DOA and/or malformed.

Here's our take on how the cloudbuster works, and also how the UN chemtrails, HAARP, and electronic mind control are able to function. It all has to do with non-electromagnetic energy in the form of deadly and/or dead orgone. I use Dr. Reich's terms for this energy, since he's the only scientist who has done extensive research with it.

I used to make orgone accumulators. I made them in many shapes and sizes and recognized the benefits and advantages from the beginning. The only drawback to these is that they absorb harmful orgone as readily as they do healthy orgone.

Orgone is omnipresent. It's ether, life force, prana, ch'i, etc., and makes up the energy and matter potential in all of creation. In its balanced state it promotes life, growth, equilibrium, but in its unbalanced state (deadly orgone is overactive orgone, dead orgone is, well, DEAD) it has exactly the opposite effect. When you see smog, you're seeing dead orgone in which particulate matter is suspended, sort of like a colloid. Very strong wind, tornadoes, hurricanes, cloudcover which yields no rain, drought, are all manifestations of harmful orgone.

In order for any parasitic activity to take place, an energy field needs to first be developed. Parasites do that in the body by excreting acids into their immediate environment. Infections, chronic pain, and a host of symptoms are manifestations of the activity of these parasites. Healthy orgone results when the polarity of the parasites' environment is reversed from positive (acidic) to negative (slightly alkaline) by the introduction of weak electric current through the skin. This is the business we're currently engaged in, so we get plenty of data and confirmation regarding this process.

In the atmosphere, agents of the present day order create massive fields of unbalanced orgone energy by introducing very powerful, specific frequencies of radio waves, spraying large volumes of acidic petrochemicals above the clouds, and by other, less believable means. Enter the orgone generator.

My frustration with the orgone accumulator's fickle behavior was answered three years ago, when one of my customers directed my attention to http://www.orgone.net he was actually experiencing some frustration, since he paid $500 for $5 worth of epoxy and metal particles, and apparently nothing else.

My daughter, Nora, was living with me then. She sees energy the way my new wife does (we've been together a little over a year) so as soon as I put the two materials together, she confirmed that it was, indeed putting out a very strong, constant field of healthy blue orgone. Adding crystals and copper pipes made it possible for me to direct the energy and alter the characteristics of the field.

One of my first thoughts was that I could create a cloudbuster that didn't need to be connected to running water, or any water, for that matter, as Dr. Reich's needed to be, and I could set it up to aim at city hall in the town I was living in, Ashland, Oregon. Theyr'e as crooked as any other local government, so I expected to get some good, observable results.

It was some time before I got around to making a cloudbuster, though. Nora could see energy, but I needed some input from somebody older than twelve. A little over a year ago, Carol and I started working together. We began experimenting with small orgone generators, mostly around natural vortices, which Carol is able to see. Every one we've ever encountered since, except the one the Navajos nurture at Four Corners, has been distorted and misdirected by the present day order.

In each case, putting an orgone generator within the vortex brought it back to vibrant health and balance, sometimes instantaneously, other times, as at the Federal Reserve Bank of New York, and at the secret facilities under Montauk, NY, over a period of hours and even days, depending on the nature of the activities carried out within the vortices previously... We developed the Holy Hand Grenade (HHG) during this time, which is a cone shaped orgone generator, 5'' or so high, with a vertical crystal inside near the point and four horizontal crystals arrayed within the bottom surface.

We've put these in vortices along the east coast of the US and across the country to the west coast, in Canada and the Bahamas. We're just getting started. Some intrepid Canadians are about to put one each at the Vatican and Assisi. I'm recommending that people put one at every major masonic temple in the world, since any time you put one in a focal point of error, the whole region gets the benefit and the people of error get deprived of the earth energy they had been abusing.

I wrote a full account of our activities over several months, which anyone can have if they email me. It's also being published serial fashion in Open Line, a monthly metaphysical newspaper. There's no physics like metaphysics, after all.

In late March, this year, we had planned to make a sort of pilgrimage to Orgonon, Wilhelm Reich's home and lab in Maine.

I set it up at our site in the RV campground in Ft Pierce, Florida for several hours while we packed, then took it apart and put it in the car. I didn't assemble it again for another week, when we arrived in Maine. We looked at the weather channel in the motel, and noticed that the whole state of Florida was still experiencing a drought, but there was a circle of heavy rainfall for a twenty mile or so radius around Ft. Pierce.

The next morning, on a whim, I assembled the cloudbuster in our motel room. There had been very heavy overcast for days, and another blizzard was forecast for that day. We went outside at about 8:30 to get breakfast and saw a neat blue hole right above us, and several people were standing around, staring at it. It was surrounded by the same, very dark clouds that had been there for days. The hole grew gradually for several hours, even after we left that afternoon. The people at the Reich Museum had made it clear that we were not welcome, sad to say, though I had emailed them a couple of weeks before to let them know that we'd be in the area. I think they'll take things like that more seriously in the future ;-)

Here's how we think it works--when metal particles are put in an organic matrix, orgone is apparently generated instead of just accumulated. I don't know why. When unbalanced orgone is present, it's drawn into the orgone generating material and transmuted into healthy orgone. I don't have a clue how this happens, either.

Making an orgone generator in a two gallon bucket and sticking copper pipes straight up out of it draws harmful orgone into the generator from many miles around and it goes out as healthy orgone in all directions, though Carol sees it happening in a sort of lens shape, parallel to the ground more or less, then moving up into the atmosphere. This gradual effect may explain why it sometimes takes a cloudbuster up to a week to get enough good stuff up to the 20,000' + altitude to disappear the chemtrails. The average radius for effective chemtrail destruction by a cloudbuster seems to be about thirty miles. Good weather effects extend a great deal farther. The rest of the region has had a drought, but the area around Spokane has been getting rain every week or so all summer. As we drive along I-90 toward Settle, there's a lot of new green growth and flowers all summer long until we reach the Columbia Gorge, 150 miles west. Anyone who lives here can tell you that this is not characteristic of the semi-desert environment west of Spokane.

We have concluded that it takes three things to make any macroparasitic/predatory activity to succeed: A field of unbalanced orgone, technology of some sort, and an operator who directs his/ her intention through both. Just as fire requires three things, oxygen, heat and fuel, and removing any one of these stops the fire, removing any of the components of parasitic/predatory activity stops that process cold. We're focusing on removing the harmful orgone medium, which puts us squarely on the side of universal law. No need to go into the other-than-human entities' involvement here, though it does add another dimension, so to speak.

Parasites are as parasites do, after all. Most people are not parasitic or predatory, so by simply neutralizing the effects of those who are by removing their harmful energies, the whole planet will come up a notch or two in awareness, freedom, etc. Another benefit of a healthy, strong orgone field is that fear and anger become diminished, and whatever diminishes fear increases courage; whatever diminishes anger increases love and acceptance..

There are two cloudbusters in Spokane this summer. Now that the smog is gone and the orgone level has increased, the whole ambiance of the city has improved. I personally used to dread having to go there, but now, after a whole day in the city (including traffic) I arrive back home feeling refreshed and relaxed. I had once vowed that I would never again spend the night in any US city because I was so badly affected by the concentration of harmful orgone found in all cities.

This really is what magic is all about. The possibilities are limitless.

I've been encouraging people to build holy hand grenades for their homes, too. I've gotten several unsolicited reports from people who did that which said that they no longer had trouble being in the presence of the overbearing in-laws and others who came to visit them. This needs to be experienced to be appreciated, I think.

DISCLAIMER & MISSION STATEMENT FOR DON CROFT AND CAROL DERRY

This is not the same as Dr. Reich's original cloudbuster. The only resemblance is the use of pipes. The former could cause changes in the atmosphere by drawing the unbalanced orgone out of the targeted area and putting it into flowing water.

This one apparently draws in the unbalanced orgone, revitalizes it as balanced orgone, and sends it right back out into the atmosphere. Don't expect the same results as Dr. Reich and others have gotten, but do expect the elimination of chemtrails, the elimination or reduction in smog and general malaise over a very wide area.
It's not necessary to point this model to remove the chemtrails. It will do that for up to a fifty mile radius if it's just left standing in the garden or on the deck.

The base is a large orgone generator, not an accumulator, and has very different characteristics from an accumulator, one of which is that it apparently can't be polluted by unbalanced orgone fields, even when they're produced by a nuclear reactor. This is why we feel confident to share this with anyone. The production of balanced orgone increases in direct proportion to the strength of the unbalanced field.

Carol and I got these specs by dowsing, and they work very, very well, though may not be the last word, if you get my drift. Let us know at zapper16@earthlink.net if you can successfully modify it.

Carol and I have been engaged in a project to (1) balance some earth energies, (2) provide a reliable cure for, among other things, the rather serious illnesses (now endemic) resulting from the chemtrail program (our cloudbusters also get rid of the chemtrails and cancel weather warfare effects), (3) interact in a productive way with ET's and native non-homo sapiens sentient races who have been doing business on, in and around our planet, outside of the awareness of! most people.

We sincerely believe that what I've been writing about is true. I can't guarantee that any of it is true, though, and no doubt some of it is colored by our own limited understanding and perception.

Due to the almost total absence of true journalism on the planet, I'm offering this work as a less-than-adequate substitute in our fields of interest. The Internet has replaced newspapers, television and radio (Art Bell's program excepted, along with a few others). There are journalists who have the expertise and even courage to meet the
standards of objectivity and scholarship needed to present a clear picture of what's really going on regarding the criminal activities of members of the alleged New World Order and their business and occult dealings with parasitic/predatory ET's, but the media are owned by these same predatory humans, so public access through
the media to true journalism is on hold for the time being.

The Internet is, we believe, a manifestation of the emerging awareness that is causing the disintegration of the institutions and political modus operendi of the alleged order mentioned above. It's the only bastion for free speech and every effort to control and regulate it by these redundant planetary institutions has failed. Our contention is that the Internet itself is giving birth to a more appropriate world commonwealth and that simply discussing this principle by email is speeding up the organic process that will bring it to full fruition. There is no need, nor is it possible, to control this process.

The impetus for writing was our own feeling that we could be risking our own safety by pursuing these subjects, a feeling that has been justified historically by the alleged order's persecution and even murder of people who have pursued similar paths. Our instincts tell us that there is inherent protection in notoriety, as the alleged order is afraid of nothing but the collective awakening of humanity and the exposure of their own misdeeds.

Since the circulation of our stories by Al McClain, Russ Michael, Sue Potter and others (thank you very much!) we've felt a considerable reduction in hostile attention by the alleged authorities, many of whom are still people of conscience at any rate, however misplaced their loyalty may be.

Most remarkable events and processes go relatively un-noticed in their own times. Many of us have seen UFO's up close. Notice that when you see them you aren't generally struck with wonder at the time. Also, when we participate in a firewalk it doesn't seem that remarkable when we're in the process. The telling of it increases
our astonishment in both cases. I use these as examples only.

I don't advise anyone to believe what I'm writing. Our own belief in these matters is based (besides a sense of "knowing") on corroborating evidence from a variety of sources, not least of which are www.wingmakers.com and www.albielek.com. You need to find your own way to check things out. Most people will resonate with aspects of what I'm writing, but we all have to know things from our own efforts, not through those of other people, however spiritual, 'connected,' or competent we believe them to be

Carol has had pronounced psychic abilities, including telepathy, energy sensitivity and the ability to astral travel, her whole life. Years ago, she began applying self discipline and training to hone her skills and reduce the tendency to color her perception with her own prejudices, which, I feel, are minor. She is also adept at using ordinary playing cards to interpret synchronous processes. I met her when I traded an orgone accumulator for a Cards of Destiny reading years ago. I knew at that time that orgone devices increase psychic perception. Carol's a Jack of Clubs, by the way, and I'm an Ace of Spades.

I don't have any special psychic abilities (beyond what other ordinary people have) but I have found that if I follow my intuitive promptings good things happen and I get more information. My metaphysical path began when I bought my first zapper in April, 1996 and cured my lifelong depression. I had already been on a spiritual path and knew the value of integrity and of reigning in the 'insistent self.'

Both of us feel that service to humanity is more important than self-advancement, though we also believe that any work deserves compensation in some form. We've been paying for our "research" and for sharing information about the holy hand grenades and cloudbuster by selling Terminators, our only product. The long term reward for sharing these devices, though will be attained when the true World Commonwealth that is foretold by every prophet and seer in human history, has supplanted the parasitic shadow order that currently operates our economies, governments, educational institutions, agriculture, medical/drug cartel, etc. Under the mandate of the emerging order, these will all be decentralized and more efficiently and fairly administered at the local level. This is not new or unexperienced. There are many cultures, especially in less centralized nations, which have been organized this way for thousands of years.

There are plenty of pitfalls and lots of trickery aimed at diverting us from investigating reality. The only defense is to be detached from the process and look for the overview and the unifying principle patterns that indicate the signature of the new paradigm and the underlying unity of the human race.

For instance, we now tend to turn away from whatever uses secrecy, elitism, complexity, hedonism, and exploitation. What is true doesn't need to be protected any more and no individual has the right or responsibility to tell others what to believe. The new paradigm allows us all to find our own level and fakers are no longer able to deceive people who are basically aware of the parameters of the new paradigm. Happily, there aren't many professional obfuscators now, at least in the metaphysical community. The bottom seems to have dropped out of the obfuscator market now that a critical mass of 'knowers' has been reached, and knowledge has ceased to be institutional property.

Gregg Braden had written that a study was done on what portion of any population actually determines the course of the progress for the rest. They came up with this figure: the square root of 2%, which amounts to a few thousand in the US. I don't remember the rest of it, but I have the impression that these few thousand are not generally the experts who are on the payrolls of institutions, and some live in Beverly Hills; some live in Harlem and rural Alabama.

The real pioneers now are mostly online and correspond with each other and most are self-financed. This is a sharp departure from the former paradigm in which Nicola Tesla, Wilhelm Reich, and many others needed the support and/or approval of the alleged governments and monetary establishment to do their work. There's something to contemplate.

We're on the threshold of having free energy. I personally know one of the inventors and he is aware that there are many, many different technologies that provide free energy now, most of which can be marketed. We hope to be the first ones to actively promote a demonstrable free energy device, the Joe Cell. If someone else beats us to it, that's fine, too. I think we can demonstrate that it's now safe to promote these things. I've been openly promoting the benefits of the zapper on the Internet for years and haven't been harassed by the alleged gov't, though, according to their rules, I'm breaking any number of alleged laws. I've made it clear that opposition by these unlawful agencies is an endorsement, so I'm not hiding from them. The 'law of rules' is not and will never be the rule of law. We will continue to trust in our intuition and divine providence for personal protection and instruction, as anyone can.

Rules are made to be broken, which is why everyone 'cheats' on taxes. Integrity dictates that if one participates in any system, one should abide by the rules or get out of the system. I paid off my bank debts and sent their SSN number back to them over five years ago. I simply couldn't bear to be considered the chattel of the Federal Reserve Corporation any longer, as anyone who uses an SSN is, de facto, despite the Constitutions clear opposition to those disgusting practices.

Getting approval from the alleged gov't for innovative devices and procedures is no longer necessary, we feel, and is even counterproductive. There are real laws on the books that cover fraud and other treacheries. We don't need the Uniform Commercial Code to regulate our personal lives. Pretty soon the UCC will be seen as an unfortunate lapse into tyranny that lasted from 1935 to around the year 2,000, after which a world commonwealth was gradually established based on the freedom and responsibility of the individual, as the American Republic had been. We simply believe it's no longer necessary to wait to exercise these birthrights, and there are entities and supernormal agencies that want us to succeed and are protecting us. Why would anyone believe something and not be willing to take risks for it? Naturally, when one makes this commitment, it's understood that he/she will persevere alone, if necessary, and until one's last breath. The brainwashing of the baby boomers has dictated that this is unfashionable and even unbalanced, but happily the children of this generation feel otherwise, as they have shown by opposing the alleged world order itself.

Lastly, I know we're on a path that doesn't jive with what's come to be seen as conventional metaphysics (is that an oxymoron?). Still, we encourage everyone to pursue metaphysical research in whatever way is comfortable and to use whatever terminology and cosmology that seems appropriate. We're happy to correspond and work with anyone, but we tend to shy away from publicly assessing the integrity or viability of others. As I mentioned, we all find our own level in the new paradigm and we
don't really need to institutionalize spirituality or science any more. I do think professional obfuscators and tyrannical agencies are fair game, though.

NOTE: The cloudbuster is synergistic with Slim Spurling's Harmonizer.

by Don Croft www.worldwithoutparasites.org
Email: zapper16@earthlink.net

Edited by Michael and Stephanie Relfe
www.metatech.org

Please distribute far and wide, so long as (1) no alterations are made without the approval of Don Croft & (2) you mention and create a link to www.metatech.org

Please save this article and photos on your own computer, in case this site ever goes down.

REPORTS FROM USERS OF THE CLOUDBUSTER

They are going totally nuts in Los Angeles today (late March 2002). We're talking chemtrails over a hundred miles long going from one end of the city to the other. I wish I had a wide angle lens so I could illustrate just how effective these CBs are. There is as I type a clear area maybe 30 miles across directly over my house. They have been pounding the same area all day and nothing has stuck. Unfortunately, I can't say the same for the areas of the city not protected. At least 50-75 trails have merged into a huge purple/brown mass to the East. I really feel for anyone under this canopy. They are going all out, obviously in desperation.
Rick Moors**,** Los Angeles
P.S. I would like to thank the pilots of the black helicopter that buzzed my house last night at about 100' and the several that have made passes directly over my house at low altitude today. Your validation of the effectiveness of my CB is very heartwarming. Soon you'll have a lot more to keep track of.

"***When we observed chemtrails overhead we directed the cloudbuster at them and saw them breakup. The same was repeated the following day. I am now in the same position using the***

cloudbuster as I was three years ago when I first stated using Slim Spurlings "Harmonizer". Too many coincidences. This simple technology is affecting our environment in a very positive manner. It is truly grass roots and I feel blessed to be a part of it."

Frederick Fischl, Spokane WA
email: fredfischl@aol.com

I ***witnessed Don`s cloudbuster in action in Seattle this past summer and was impressed with the speed of response. Fred Fischl did the demo and noted that the response was much faster than he had previously observed, due to the fact that we had a Harmonizer field running at the time in close proximity. I have built and used several cloudbusters similar in principle to Reich`s starting in 1974, with universally good results.***

Don`s innovation on the technology is a superb simplification which makes it easily available to all and may have some very far reaching benefits in the immediate future as we contemplate chemical and bio terrorism from many sources. I believe that the nuclear terrorism has had its fangs pulled due to the many Harmonizer units currently in the field in the major population centers. However more are needed in the country to strengthen the overall field.

Slim Spurling 38969WCR ,Roggen CO, 80652 Tel:303-849-5411
www.slimspurling.com

MORE REPORTS FROM USERS OF THE CLOUDBUSTER AND HHG

HOW TO OBTAIN PARTS FOR THE CB, OR COMPLETED CBS

1) http://ctbusters.com

2) Spacers are available from Alteronics, 372 Lodgeview Drive, Oroville, CA 95966,
Email: orovilledave@yahoo.com for prices

Here are photos of the bottom wooden spacers, and a top spacer made of plastic:

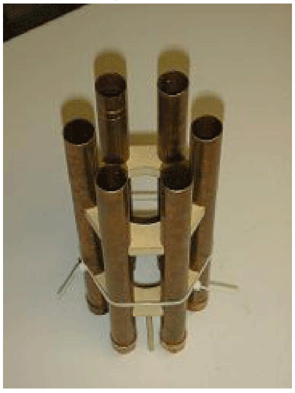

3) CBs and crystals from Michelle Ridgley
P.O. Box 66491
Baltimore, MD 21239
email her for a price list: mridgley@aol.com

Neither Don Croft or the webmaster of this site, Stephanie Relfe, receive any money from the sale of Michelle's products.

(David at davdart@texas.net also sells suitable crystals for cloudbusters)

HHGS are available from www.powerpyramids.com/

A Good Article sorting out the confusion with: Cloudbusters, Chembuster, Trevor Constable, Jeff Rense, Don Croft and Reich is at http://educate-yourself.org/cn/chembustersexplained24feb03.shtml

Read Also:

Dolphin Rings: A request relayed in a vision

CLOUDBUSTERS EGROUPS!

Please join if you intend to have a cloudbuster!!!

1)http://groups.yahoo.com/group/cloudbusters/. However, that group got hacked too often by a number of parties, including we presume operatives for the secret government.

2) Don Croft now has an egroup at http://www.ethericwarriors.com

NEW! The Tower Buster! Negative mind-controlling energies sent from cell phone towers can be reduced with Don' s new Tower Buster! Don believes making Tower Busters is currently more important than making Chembusters!

www.metatech.org
Email:info@metatech.org

had existed only a few minutes earlier, a result of intensified orgone energy (http://educate-yourself.org/dc/dcwhatisorgone.shtml) accumulating in that area.

Almost simultaneously, we wondered aloud whether a group of people could use this technique to break up chemtrails! I tried it a number of times on my own, but I couldn't get chemtrail 'clouds' to break up and disperse the same way normal clouds would and I wondered silently whether a *Cloudbuster*, a simple invention of Wilhelm Reich.to direct orgone energy in the atmosphere, could tackle the job.

Over the course of 35 years, Trevor Constable had modified Reich's original long pipe arrayed Cloudbuster down to a much smaller, compact, and simpler design that could 'make' or 'break' weather in any direction he chose: He could make rain or stop it; he could relieve drought and break up smog or create a dry spell and intensify smog conditions. The manner in which you manipulate the atmospheric ether in a given area is the key.

Trevor discovered that success depends on how you apply his Cloudbuster design to direct the *flow* of etheric energy. Its normal movement is from west to east in northern latitudes. If you reverse its flow and build up clouds, you'll get rain. If the flow of orgone is stagnated and bottled up, you'll get drought or smog. If you release the etheric flow after being 'bottled up' or stuck during a dry spell, you will get torrential rains. There is likely no civilian on the face of this earth who has accumulated more expertise about weather engineering than Trevor James Constable

However, the hero of this article is Don Croft, a man who was already familiar with Reich's observations of orgone energy and his invention of the Cloudbuster. Don was talking one day with a man who told him about a radionics type device that he obtained over the internet. He said that paid over $500 for a plastic box which contained some dials and mostly metal shavings and epoxy resin.

Don was intrigued and read more information about the device. The maker of the device had capitilized on an observation that Wilhelm Reich had written about in one of his books on orgone energy, *The Cancer Biopathy*. Reich found that any organic material, when juxtapositioned with metal, will act as an absorber of orgone energy which is first attracted to and then reflected away by the metal. The device maker was calling his device an *orgone generator* (abbreviated "orgen" here). Apparently, an orgen can generate the positive form of orgone ("OR") energy by transmuting or neutralizing the deadly or negative form of orgone energy dubbed "DOR" . Orgone a*ccumulators* , on the other hand, can accumulate (and intensify) both the positive *and* deadly polarities of orgone energy. Orgone accumulators have traditionally been abbreviated "oracs" and Don sometimes uses that term himself, but Don's modified version of a Cloudbuster employs the mixture of metal shavings and plastic resin of the orgone generator and is *not* the same thing as a traditional Reichean orogne accumulator or Reichean Cloudbuster. The marrying of metal shavings and plastic resin of the orgone generator with Reich's Cloudbuster pipes has resulted in a different atmospheric instrument, with *new* and *inique* attributes from the traditional Reich Cloudbuster.

It's probably more appropo to call it a "Chembuster" to distinguish it from Reich's Cloudbuster and avoid misunderstandings among Reich aficionados and pedagogues. The orgone generator portion of Don's hybrid devices consists of metal shavings, magnets, crystals, and sometimes a Mobius coil or spiral coils that are rigidly held in place by epoxy or polyester resin. With psychic inisights gained form his wife Carol, and a little help from dowsing, Don had also put together a much smaller, cone shaped orgone generator using the above mentioned materials and called it a *Holy Handgrenade* (HHg) which you can make yourself (http://educate-yourself.org/dc/HHGconstruction10jan02.shtml) or obtain one ready made (http://educate-yourself.org/dc/orgonegenindex.shtml#chembusters)

A number of people have reported unusual and positive emotional/physical effects produced by a HHg, even from a considerable distance. Don also incorporated the orgone generator into his Terminator, his latest hybrid modification of a Dr. Hulda Clark type 'zapper.'

Five Crystal Orgone Generators

Orgone generators have the ability to focus life-positive orgone energy and break up areas of noxious, life-draining energies of DOR, a word coined by Wilhelm Reich for Deadly ORgone. Don found that his orgone generator could neutralize DOR and seemed to only generate positive, life-enhancing energies. Don and his wife began planting these five crystal orgone generators at outdoor locations near nuclear power plants and other areas of noxious, life depleting energies (secret underground and underwater bases, etc.) and started to have a definite effect on improving the energy fields radiating from these zones.

Since he already was thinking about using the Reich Cloudbuster to try and break up chemtrails, Don took it a step further and combined a large version of his HHg with the rigid copper pipes of a Cloudbuster and a 2 gallon plastic bucket to hold the whole enchilada together. This modification of Reich's Cloudbuster into a *Chembuster* is a unique and original prototype initiated by Don & Carol Croft and produces some rather 'unique' effects.

Don originally had set up his Chembuster outside on a chemtrail-overcast day with the pipes pointing straight up and sure enough, after a while, the chemtrails started to break up revealing blue sky above. The longer he left the Chembuster in place, the clearer the sky got, eventually covering an area out to 90 miles. He later improved the Chembuster design and got the coverage to extend out to 120 miles. A man named Rick from the Los Angeles area timed how long it took his Chembuster to disperse overhead Chemtrails and posted his photos to the Don's original Cloudbuster chat forum (now defunct). Amazingly, the area immediately above Rick's Chembuster transformed from Chemtrail laden overcast to CLEAR blue sky in *less than 10 minutes*. You can see the photo sequence here: (http://educate-yourself.org/dc/dcricks10minctclearingjob.shtml).

Don had contacted me by E mail last year and was sending me reports of his success with the Chembuster both in this country and abroad. He had sent the instructions to an African man named Gert Botha living in Namibia who built a modified version of the device using locally scavenged materials (read Africa's First Cloudsbuster (http://educate-yourself.org/ct/africasfirstcloudbusterjun01.shtml)

Initially, Gert found that his Chembuster was having an effect on the incessant sand storms, causing them to constantly change direction, but eventually he found that aiming the Chembuster south at an angle of 30 degrees to the horizon forced the sandstorms to stop. After being parked in this position for two months, his Chembuster brought on torrential rains for the first time in many dry months to one of the driest regions of Africa.

Your Weekend Project

Complete instructions to make your own Chembuster are included in this article. It will cost you about $150 US dollars if you buy all of the materials new, but if you're a scrounger, you can probably do it for the cost of the plastic resin. The instructions were written by Don Croft, but the original aritlce has assembled, edited and organized by Stephanie Relfe who has a couple of great web sites that everyone should take a look at: www.metatech.org and www.relfe.com .

Anyone with modest do-it-yourself skills can build these devices, but be aware that a Reichean style Cloudbuster is not a toy. It directs orgone energy, both the good type and the harmful DOR type. Don's modification of the traditional CB has removed most of the concern and potential dangers of DOR, since the orgen neutralizes the DOR, but one individual reported getting a headache after touching the pipes of a Chembuster and NOT touching the base to ground himself out. Always touch the base after handling the pipes of an active Chembuster.

A traditional Cloudbuster patterened after Reich's design has the potential to be harmful if you don't know what you are doing. Trevor Constable told me first hand a story about the late Robert McCullough, Reich's friend and assistant in the Arizona desert in the 1950's. Dr Reich and Bob were running pioneering experiments with Reich's Cloudbuster in the desert when Bob had allowed a big (functioning) Cloudbuster to accidentally touch the side of his bare leg-which instantly paralyzed his leg and caused him excruciating pain. It took weeks to recover. Bob walked with a limp and had to use a cane for the rest of his life. The desert environment contains the highest ratio of atmospheric DOR energy than any other place on Earth, so it was the worst possible location to have such an accident. You don't have that sort of concern with Don's Chembuster, but touch the base (just to be sure) after handling the pipes to avoid the potential of getting a headache. In general, avoid touching the pipes of a Chembuster after it has been set up and running for a few days. If you need to touch the pipes of an active Chembuster, it's a good idea to take a heavy copper conductor wire, like an battery jumper cable, and touch the jumper cable to each pipe in turn while holding the opposite end of the cable against the earth (or at least against the resin/metal matrix of the base if you can't get at the soil).

Below, you'll find the edited reprint of Don's paper on Chemtrail Busting and complete instructions for making your own Chembuster or "CB".

While written some time ago, Don's Introductory Notes offer important insights into the Croft world view and their mission. I believe you will find it highly informative and in sync with the emerging paradigm of the new millennium.

CB Parts, Kits, & Finished Chembusters

Don explains in the paper where to locate most supplies for the CB such as Home Depot. You can obtain quartz crystals for the CB or five crystal orgone generator from Gladys Bridges<crystals87@sbcglobal.net> in Arkansas

If you wish to obtain a ready made Chembuster in either a 2 gallon or 5 gallon version, using either copper pipes or less costly zinc plated iron pipes, or a complete Chembuster kit for making the 2 or 5 gallon chembuster, then visit this link, http://educate-yourself.org/dc/orgonegenindex.shtml Scroll down to *Chembusters*.

© Copyright 2002 - 2008 Educate-Yourself.org All Rights Reserved.

How to Build A Chembuster to Disperse Chemtrails and Improve the Weather

Written by Don Croft <terminator3@turbonet.com>
Edited, assembled, and originally posted by Stephanie Relfe at http//www.metatech.org

Definition of a contrail: A contrail occurs when a plane travels at a high altitude (about 30,000 ft. or more) and compresses the air into a water vapor or ice crystals through jet engines or the wing tips pushing through the air. This trail disappears after anywhere from a few seconds to 15 minutes, as the water vapor evaporates.

Definition of a chemtrail: The term "chemtrail" is a relatively new word, which appeared in the last few years along with the appearance of chemtrails. A chemtrail is very different from a contrail. At first a chemtrail might look a bit like a contrail. However, instead of disappearing like a contrail does, a chemtrail just keeps spreading out and forming a hazy cloud bankout and forming a hazy cloud bank These trails traverse the whole sky and stay for up to around five or even eight hours. . .

They have been known to turn what was originally a clear blue sky into a grey haze. (Note: As of Nov 2001 there are already several dozen Chembusters in seven countries, so if you see chemtrails that don't last anywhere near this long, there may already be a Chembuster in the area). It is believed chemtrails are composed at least in part by an oil based substance of unknown content that is actually being sprayed from jets on populated areas. It has also been found that some chemtrails contain red blood cells, aluminum,

mycoplasma, viruses and other unknown biological components. There are many reports of increased diseases, especially respiratory ailments, in areas where there are chemtrails.

Chemtrails are often seen to be sprayed by white planes with no markings. Sometimes you will see these planes give off no contrail at all, and immediately after the same kind of plane in the same area will start to leave a thick chemtrail. At other times, distinct short gaps in chemtrails have been photographed, as though the plane ran out of chemicals and then had to start spraying with a new supply.

For information on Chemtrails see the sites listed at http://www.carnicom.com/contrails.htm

History of the Cloudbuster

The Wilhelm Reich Cloudbuster was well known for its ability to create rain. But a problem existed with its continued use, in that the deadly orgone energy that it absorbed could in fact hurt the operator if it was not correctly "drained off". This "draining off" of deadly orgone energy was accomplished by 1. Connecting the Cloudbuster to a body of running water (difficult) or 2. Connecting the Cloudbuster output to an orgone accumulator. However, the orgone accumulator would soon become saturated with the deadly orgone energy which in turn caused more problems for the operator. So the Cloudbuster, as good a technology as it is, has lain dormant for many years.

The Cloudbuster Re-Invented

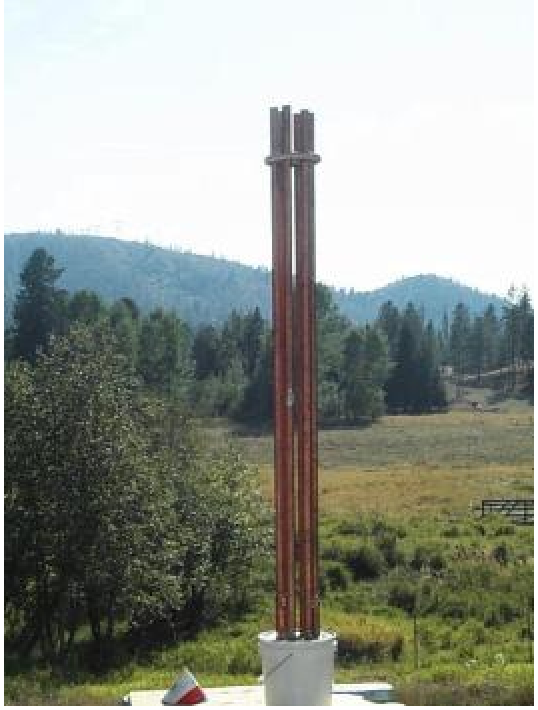

By utilizing the research of Wilhelm Reich, we have created a cheap, portable and easy-to-build device that consistently destroys Chemtrails and heals the atmosphere. The “Chembuster” is the answer to these ongoing attacks. Unlike the original Cloudbuster, the CB changes the deadly orgone energy to good orgone energy and does NOT become saturated or dangerous to the operator. You can build one for about $150.

Join The Battle - Now!

It is human nature to sit on the sidelines and think that one person cannot change the things that are happening to our country. You now have no excuse for inaction. For a few dollars and a few hours of your time YOU can make a difference.

A properly constructed CB unit should be able to greatly reduce the effects of chemtrails for about 45 miles in all directions (90 mile diameter) from the unit. If used with Slim Spurling’s environmental tools, and a few extra crystals, the diameter can be increased to about 120 miles in all directions.

With this device a mere few hundred concerned citizens can SHUT DOWN the multi-billion dollar chemtrail program. The Chembuster cannot do any damage. It can only help and heal. It’s up to you to join the fight to save America.

Chembuster Construction Details

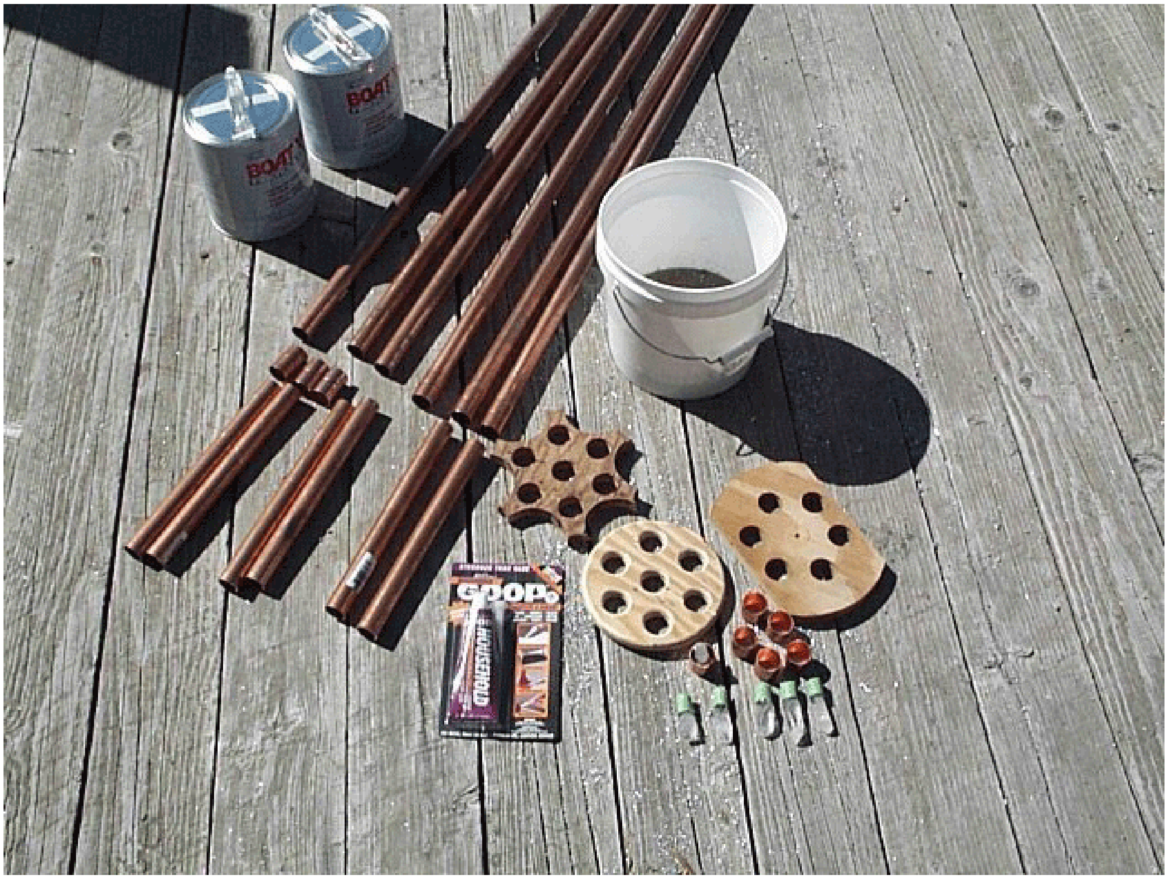

Chembuster Materials

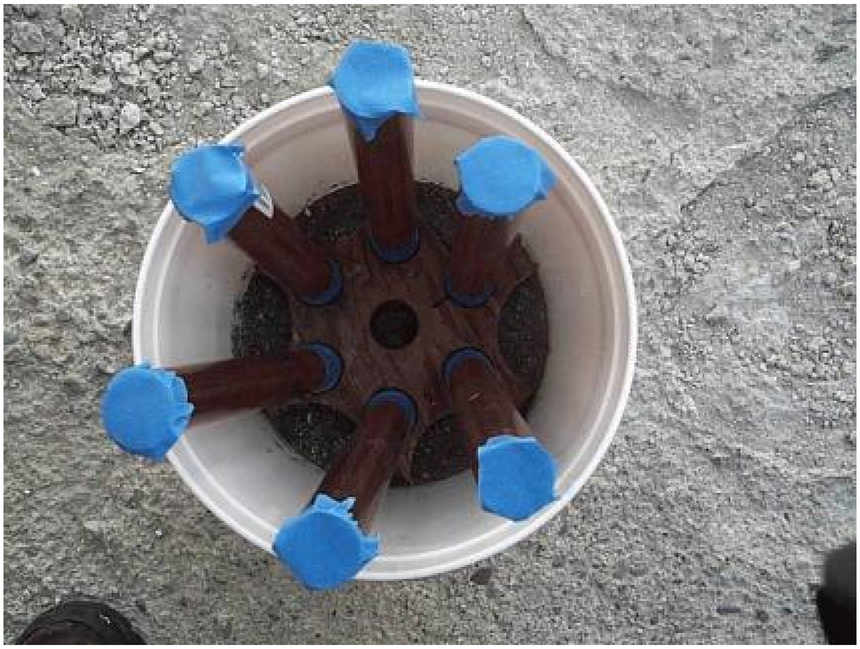

Set pipes in bottom spacer

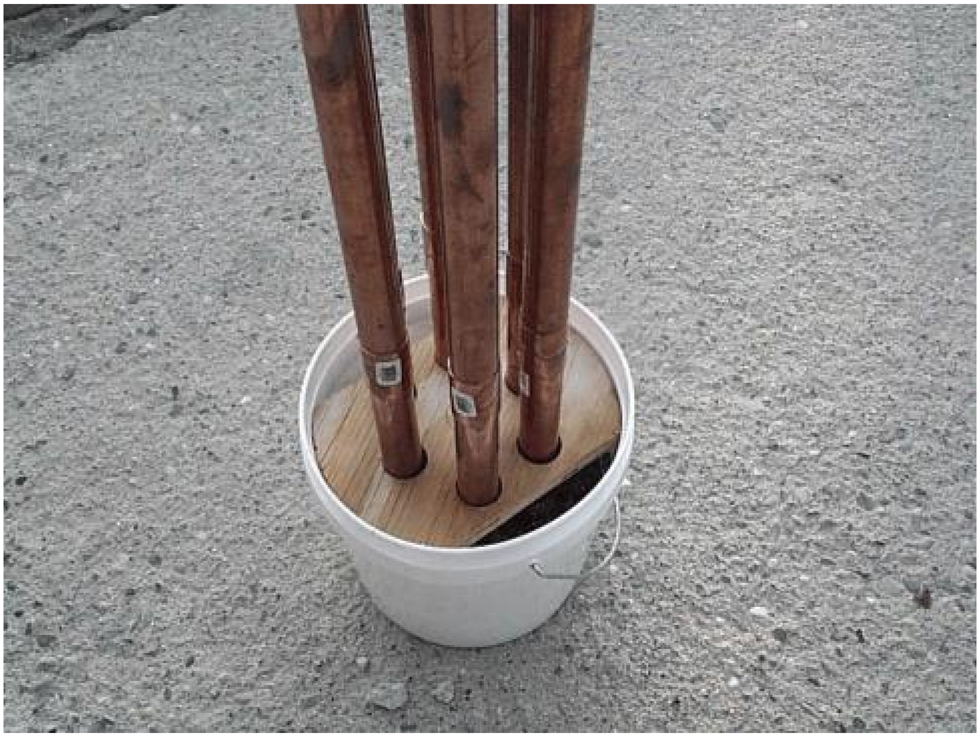

Middle Plywood Spacer
CHEMBUSTER DIRECTIONS

BUCKET: Base: 9" in dia., 9" deep. I use two-gallon plastic buckets from the Home Depot paint department, but two-gallon food buckets, available from restaurants, or two-gallon spackle buckets from drywall contractors are fine. Leave the buckets on, as the handles make it easier to move the finished Chembuster.

COPPER PIPES: Six standard Type M, 1" copper pipes, 6' long, open at one end, the capped ends submerged to within an inch or so of the bottom of the bucket before casting. For portability and shipping, you can use 12” sections of pipe in the base, adding five-foot lengths by putting couplers between the 12” and five-foot pieces. You won’t need to buy as much copper pipe this way, too.

CRYSTALS: I glued a double-terminated (two pointed ends) quartz crystal, about 2" long, into a .75” section of garden hose, (or use electrical tape) then, glue to the inside of each of the six copper end caps, pointing in the same direction as the pipe will be after the cap is glued on. I use 'Goop' glue. Tape, or solder around the cap where it joins the pipe to prevent resin from leaking into the cap during casting. (Note) Another option is to add one citrine gemstone to each pipe before adding the crystal, this smoothes the CB’s energy and helps to transmute negative energies. However, citrine does not seem to make the CB any more effective at eliminating chemtrails.

METAL PARTICLES: Get some metal particles from a machine shop, recycling yard or a place where they saw a lot of aluminum (sign shops and aluminum fabrication shops are good places to look). I use aluminum for weight considerations, but ANY metal will do. Very fine particles (such as filings from a key-making machine) aren’t quite as good as ones that will easily go through a ½” screen are okay. There really is a wide margin of suitability!

PLYWOOD: Template 1 - 3/4” exterior plywood is used to make the base for the copper endcaps/pipes. I arrange the six pipes evenly around a 2 1/2” radius circle, which makes the pipes 2 1/2” apart on centers. A flat, 1 1/4” auger bit drills a suitable hole for each end cap.

Template 2 - Make another piece to fit inside the rim of the bucket. Remember to cut a section away so you can fit your fingers under it to remove it from the bucket. This piece is not part of the finished CB, just used to get the spacing for the pipes right during construction. Cut the holes in this with a 1 1/8” flat bit so that it will fit well over the six pipes. This piece will hold the pipes in position while the second batch of resin hardens.

Template 3 - I make a circular plywood spacer for the top ends of the pipes, 4” radius with 1 1/8” holes on the same 2 1/2” radius circles for the pipe ends. This keeps the pipes parallel with each other and fairly rigid.

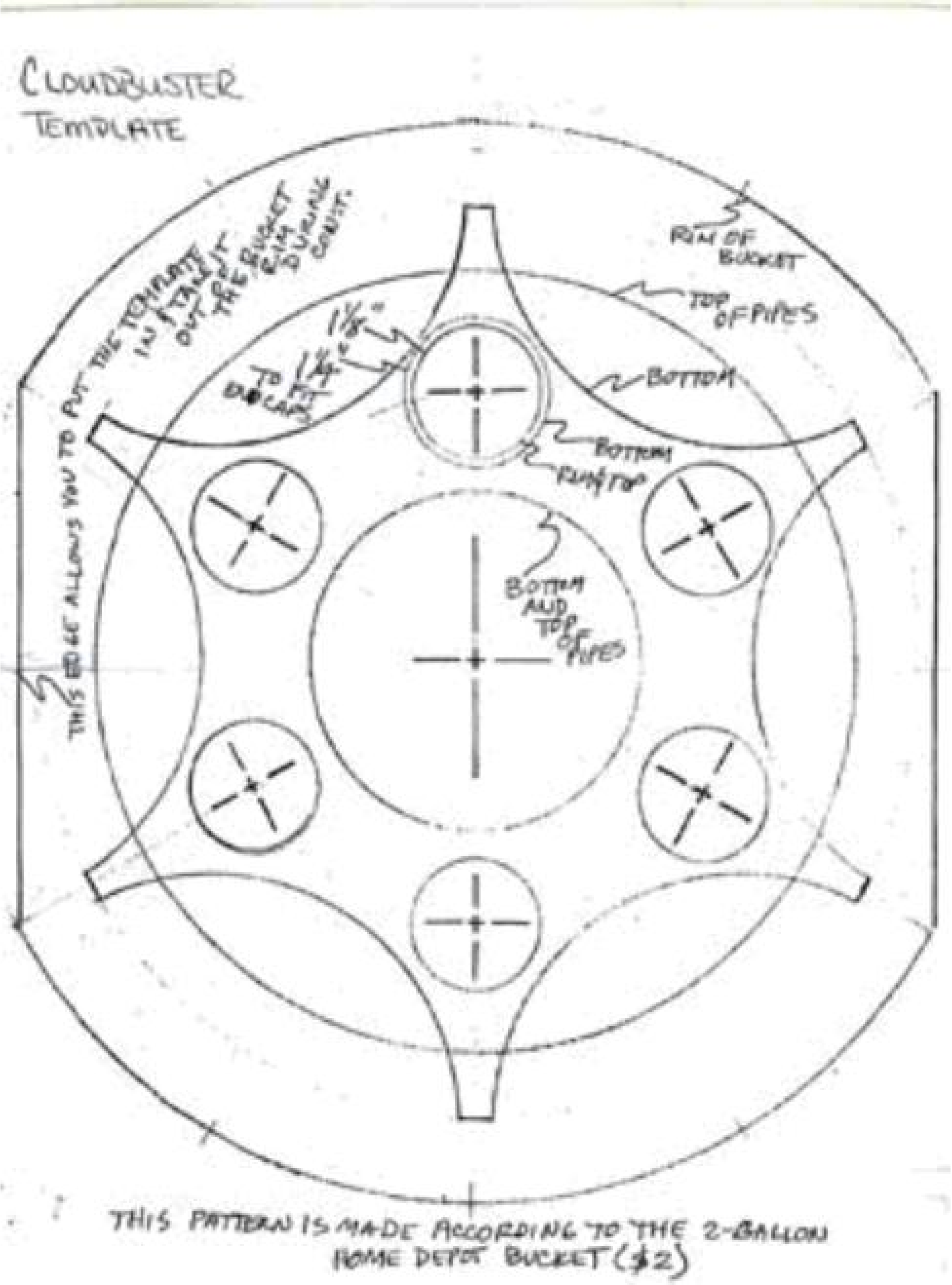
CLOUDBUSTER
TEMPLATE
THIS EDGE ALLOWS YOU TO PUT THE TEMPLATE IN & TAKE IT OUT OF THE BUCKET RIM DURING CONST.
RIM OF BUCKET
TOP OF PIPES
1 1/8"
1 1/4"
TO FIT ENDCAPS
BOTTOM
BOTTOM
RIM/TOP
BOTTOM AND TOP OF PIPES
THIS PATTERN IS MADE ACCORDING TO THE 2-GALLON HOME DEPOT BUCKET ($2)

Not to scale Widest circle is 9 1/8”

RESIN: I use polyester resin (used to build fiberglass boats), but epoxy also works, as does Envirotech resin from www.eti-usa.com (this sets up very quickly, so you need to be proficient with it). If you are going to construct many of these units, try to purchase your resin in 5 gallon buckets over the internet, possibly from a surfboard or boat manufacturing supply company. Home Depot gets $24 per gallon for resin.

On a level surface, pour an inch or so of catalyzed resin into the bottom of the bucket and stir in metal particles until the surface is uniform and flat. CAUTION: Resin fumes are flammable; use in a warm, but well ventilated area.

The ratio that seems to work is approx. one part resin to one part metal particles (equal volume of resin to metal particles). Put the base Template 1 onto that before it hardens, so that it just touches the surface and doesn't sink in much. Pour a half gallon of catalyzed resin into the bucket after the six pipes are set into the holes in the base template. Use the dowel to stir in handfuls of metal particles in the same ratio as before. Note that the wood Template 1 should not be pushed down to the bottom of the bucket with the resin squishing over the top.

Carefully move the Template 2 down on to the six pipes until it is snugly within the rim of the bucket. Put the top Template 3 onto the top of the pipes and move them until the pipes are parallel. After the resin has hardened, remove the second template and set it aside. This is only used during construction, as I mentioned.

You can finish off putting the resin/metal mixture in now in two stages if you want, up to within a half inch of the rim of the bucket. Now it’s finished (wait until it hardens before you move it. You will actually use about one gallon of metal shavings and 1 ½ gal. of resin.

A MONEY SAVING SUGGESTION, ALSO FOR BETTER PORTABILITY:

Cut six pipes, 12" long, and put the crystals/endcaps on them for the base. Buy six coupler joints which allow a pipe to come into them from both directions and stop at the middle. This allows you to make the rest of the pipe assembly from three pipes, 10' long, cut in half. The finished product leaves you with only one piece of pipe, four feet long leftover.

When positioning the 12" pipes for pouring the second layer of mix, temporarily tape the top ends of the 12" pipes so no metal particles are dropped in them by mistake.

After the mix is made, take off the tape and push the rim plywood piece down onto the pipes so that it fits into the rim of the bucket.

Put the spacers (soldered if desired), then the 5' pipes onto the 12" base pipes, then put the third (top) plywood piece onto the 5’ copper pipes.

The pipes will look skewed, most likely, so just twist the top plywood piece around until the pipes line up parallel. As you do that, the rim plywood piece will move around. You need to do this expeditiously, though not in a hurry, so that the resin doesn't harden before you're done.

We experimented with 5' extensions to the pipes to increase the range. It had a strong effect on the moon and on the upper atmosphere of the planet for hundreds of miles. We also apparently shut down the weather warfare facility under Homestead Air Force Base by leaving the extended Chembuster pointing at the dead orgone field above it for a week.

We found that we don't need to point it in any particular direction in order to clear the atmosphere of bad orgone, and the longer it's left standing, the nicer it feels for a longer distance from the Chembuster (many miles). We just stand it on its bottom, pointing straight up. We did once eliminate a belt of smog, which was just off the coast by aiming at it for a few minutes, so direction can be a factor.

Chembusters destroy chemtrails consistently. It may take a few days for your Chembuster to activate the environment enough to do this, but after that point the spew will disappear within seconds of leaving the UN jets and dissipated trails that drift within range of the Chembuster will also disperse, but more slowly. The ones that are made up of solids take longer to dissipate, but we've found that these only make up about 10% or so of the spewing episodes.

Dr. Reich's Cloudbusters were extremely dangerous when in operation since he had no orgone generator attached to them. Bad orgone does carry a charge which can quickly drain one's life force away when it's concentrated in one spot. It's similar to a strong positively charges static electrical field. When cloud cover is unproductive or unbalanced (too much lightning, for instance, or too much rain) the Chembuster busts it up.

It also may help stop drought. Overall, it just balances the weather, I think it sort of mediates between the ground and the atmosphere somehow. The one in Namibia, Africa finally brought heavy rain to the desert there. It happened 9/5/01, and continued for several days. Gert, the fellow who made the Chembuster and reported the occurrence (http://educate-yourself.org/africasfirstcloudbusterjun01.html), told me that this had simply never happened in the Namib desert before to the best of anyone's knowledge. It's arguably the driest region on the planet.

WARNING: If you touch the pipes when they're drawing in strong unbalanced or dead orgone and don't put your hand on the base after that, you may experience headaches, even into the next day, as a friend of ours did recently.

Mind you that our efforts, though somewhat pioneering, should be seen mainly as a springboard for further developments rather than the last word.

SUGGESTION: If you want even more rain, try putting water on top of the material in the bucket and/or hosing down the whole apparatus daily. In freezing climates, cover the top of the six pipes with mini umbrellas so water can't freeze and shatter the crystals. The CB appears to effect chemtrails equally well whether it's placed indoors or out. Up to seven crystals can be placed in each pipe to improve performance.

CB plowing through chemtrails in the midwest

Free Newsletter

THE ORGONE ACCUMULATOR HANDBOOK

Construction Plans, Experimental Use and Protection Against Toxic Energy

by James DeMeo, Ph.D.

With a Foreword by Eva Reich, M.D.

The Orgone Accumulator Handbook

by James DeMeo, Ph.D., with a foreword by Eva Reich, M.D., and an appendix examining new experimental evidence demonstrating the effectiveness of the orgone energy accumulator.

Learn how to concentrate and work with *orgone* (life) energy using simple, readily-available materials. A limitless supply is freely available from the atmosphere. These experimental devices have been shown to stimulate plant growth and human physiology in controlled studies.

The orgone accumulator is based on discoveries by the late **Dr. Wilhelm Reich**, whose work remains vital and of wide interest, despite unprecedented hostility to his scientific research. In a court decision unique in American history, his books were ordered *burned* by the U.S. Federal government in 1957. Dr. Reich reported medical benefits for his patients who used the accumulators.

In this edition Dr. DeMeo updates information on the construction of orgone energy blankets, shooters, box-type accumulators, bion packs and "draw buckets" for atmospheric cleansing, with many instructional diagrams and photographs, as well as a complete bibliography and resource guide for research.

About the Author: James DeMeo has been researching the topic of the life energy, and specifically the works of Dr. Wilhelm Reich since 1970. He has authored or edited dozens of published articles, books and research compendiums (including *Saharasia, Heretic's Notebook, On Wilhelm Reich & Orgonomy*, and the German-language *Nach Reich*) which subjected Reich's ideas to rigorous testing. DeMeo formally studied Environmental Science and Geography at Florida International Univ., and the Univ. of Kansas, where he earned his Ph.D. He served on the Faculty of Geography at Illinois State Univ. and the Univ. of Miami, and founded the Orgone Biophysical Research Laboratory in 1978.

ISBN 0-9621855-0-7
51295
9 780962 185502

List Price: $12.95

Books by James DeMeo

- * *Saharasia: The 4000 BCE Origins of Child Abuse, Sex-Repression, Warfare and Social Violence In the Deserts of the Old World*

- * (Editor) *Heretic's Notebook: Emotions, Protocells, Ether-Drift and Cosmic Life-Energy, with New Research Supporting Wilhelm Reich*

- * *The Orgone Accumulator Handbook: Construction Plans, Experimental Use, and Protection Against Toxic Energy (Second revised edition)*

- * (Editor) *On Wilhelm Reich and Orgonomy*

- * (Co-Editor with Bernd Senf) *Nach Reich: Neue Forschungen zur Orgonomie: Sexualökonomie, Die Entdeckung der Orgonenergie*

- * *Bibliography on Orgone Biophysics: 1934 to 1986*

THE ORGONE ACCUMULATOR HANDBOOK

Construction Plans, Experimental Use and Protection Against Toxic Energy

by

James DeMeo, Ph.D.

Natural Energy Works
Greensprings, PO Box 1148
Ashland, Oregon 97520 USA

Published by

Natural Energy Works
Greensprings, PO Box 1148
Ashland, Oregon 97520
United States of America

Tel/Fax: 541-552-0118 Email: demeo@mind.net

The ORGONE ACCUMULATOR HANDBOOK, Copyright © 1989, 1999 by V. James DeMeo. All Rights Reserved. Printed in the United States of America. No part of this book may be used or reproduced in any manner whatsoever without written permission, except in the case of brief quotations embodied in critical articles or reviews. For information, send inquiries to Natural Energy Works.

ISBN 0-9621855-0-7

Second Revised Edition, with Postscript 1999

7 8 9 10 11 12 13 14 15

Library of Congress Catalog Card No. 89-90975

Cover: NASA photo of Apollo 12 astronaut walking on the surface of the moon (see Life Magazine, 12 Dec. 1969). His body orgone energy field is softly glowing a blue color in the lunar vacuum, possibly due to excitation by his high-frequency radio communications equipment. This blue coloration of the energy field in the photo, which has been seen in a few other lunar astronaut images, has been systematically ignored or explained away as an effect of "lunar dust" or "water vapor". In fact, it is a visible expression of the human orgone (life) energy field.

Acknowledgements

This book is the product of many years of study of the prior works of Wilhelm Reich, and other dedicated scientists, without whose research findings it would not have been possible. The reader will find their names and research papers listed in the reference section of this book. I have corresponded with, and learned from many of these researchers over the years. In particular, I thank Eva Reich and Jutta Espanca for their constructive criticisms of the first edition of this Handbook. Additionally, I thank my own mentors, Robert Morris, Robert Nunley, and Richard Blasband, from each of whom I learned different things about life energy. I thank Theirrie Cook and Don Bill, loyal friends who have helped with the progress of my research in many ways, and also Jim Martin for the initial ideas and encouragement to revise my original small Handbook into the present edition. Jim also prepared the typesetting and most of the graphics in this edition, providing many helpful points along the way. Thanks to Stephan Müschenich and Rainer Gebauer for the courage they showed in doing their double-blind, controlled study of the human physiological response to accumulator treatment, at the University of Marburg, and for the Appendix document. Thanks also to the various researchers and doctors in Germany, who today are openly working with the accumulator in a manner presently impossible in the USA. In particular, I thank Heiko Lassek, Dorothea Opferman-Fuckert and Manfred Fuckert from whom I learned about the possiblities and limitations of physical orgone therapy. Also thanks to Vince Wiberg, for the simple and inexpensive methods for detecting electromagnetic disturbances, presented here. And finally, thank you Wilhelm Reich, for the discovery of the orgone energy and the orgone accumulator.

Acknowledgments

[illegible]

"We consider the discovery of the orgone energy among the greatest events in human history."– from a letter to the American Medical Association, signed by 17 medical doctors in 1949.

"THE ORGONE ACCUMULATOR IS THE MOST IMPORTANT SINGLE DISCOVERY IN THE HISTORY OF MEDICINE, BAR NONE" – Theodore P. Wolfe, M.D., from *Emotional Plague Versus Orgone Biophysics*, 1948.

"It is justifiable that the discovery of orgone energy and its medical applications by means of the orgone accumulator, the orgone shooter, bionous earth, and orgone water have opened up an abundance of new and, it appears, amazingly good prospects." – Wilhelm Reich, M.D. from *The Cancer Biopathy (Discovery of the Orgone, Volume 2)*, 1948.

"What would you say of the leading philosophers here to whom I have offered a thousand times of my own accord to show my studies, but who with the lazy obstinacy of a serpent who has eaten his full have never consented to look at the planets, the moon, or telescope? To such people philosophy is a kind of book... where the truth is to be sought not in the Universe or in nature, but (I use their own words) by comparing texts." – Galileo Galilei, Italian astronomer of the 1600s who proved that the Earth moved in the heavens, shortly before being prosecuted and tortured by the Catholic Church.

"...the orgone energy does not exist."– Judge John D. Clifford, from a 1954 U.S. court ruling in which all of Dr. Reich's books and research journals were banned, and ordered burned in incinerators; Reich was sent to a Federal Penitentiary, where he died.

CONTENTS

Page

1. Foreword, by Eva Reich, M.D. 1
2. Author's Preface 2

Part I:
The Biophysics of the Orgone Energy

3. What Is the Orgone Energy? 11
4. Wilhelm Reich's Discovery of the Orgone Energy and Invention of the Orgone Accumulator 13
5. Objective Demonstration of the Orgone Energy 24
6. Discovery of an Unusual Energy by Other Scientists 30

Part II:
The Safe and Effective Use of Orgone Accumulating Devices

7. General Principles for Construction and Experimental Use of the Orgone Accumulator....... 42
8. Warnings and Dangers: Oranur and Dor 48
9. Cleaning Up Your Bioenergetic Environment 63
10. Draw Buckets, Bion Waters, Bion Packs 72
11. Physiological and Biomedical Effects 78
12. Personal Observations with the Accumulator 86
13. Some Simple and Not-So-Simple Experiments with the Orgone Accumulator 93
14. Questions and Answers 104

Page

Part III:
Construction Plans for Orgone Accumulating Devices

15. Construction of a 2-Ply Orgone Blanket 108
16. Construction of a 5-Ply Garden Seed Charger 114
17. Construction of a 10-Ply Orgone Accumulator 118
18. Construction of an Orgone Shooter Funnel 124
19. Construction of an Orgone Shooter Wand 125
20. Construction of a 3-Ply Human-Sized Orgone Energy Accumulator 126

Selected References 135

Sources of Additional Information:
On Wilhelm Reich and Orgonomy 145
On Alternative Health Clinics and Institutes 147

Appendix: "The Psycho-Physiological Effects of the Reich Orgone Energy Accumulator", by Stefan Müschenich and Rainer Gebauer 148

Postscript 1999 157

CONTENTS

PAGE

PART III

Contributions to Theory

[illegible]

Selected Bibliography [illegible]

PART IV

[illegible]

1. Foreword

At last, thirty two years after the 1957 death of Wilhelm Reich, human beings can begin to study orgonomy like any other body of knowledge, helped by the *Orgone Accumulator Handbook.* This concise and informative book contains in a nutshell a condensed, clear account of the discovery, made usable by all who are interested in the cosmic life energy. Herein is printed: the scientific definition of the orgone energy; the history of how the steps of observation, experimentation, and theoretical insight led Reich to practical applications; the principles for construction and experimental uses of the orgone energy accumulator, with detailed suggestions for needed material, layering, and dimensions; and finally a very useful reference list. Professor J. DeMeo shows his thorough knowledge of the subject, which as yet is banned and omitted from the 20th Century academic curriculum, except for a few pioneering lecture courses (in New York and West Berlin).

Wilhelm Reich said that even though the life energy had been known for thousands of years, he managed to make it concretely usable, and that the era of its applications has just started. However, this *Handbook* is the first printed material in recent years on specifically how to concentrate the energy from the Earth's atmosphere. It is usable for a laboratory course on the subject of cosmic life energy. This material could be comprehended by intelligent high school or college students. It answers my almost fifty year old hope, for the inclusion of life energy facts in the body of knowledge that all educated people on Earth should learn in their schooling. Thank you James DeMeo.

Eva Reich, M.D.
West Berlin, March 1989

2. Author's Preface

When I was 12 years old, a favorite uncle of mine died a suffering death from lung cancer. The doctors had removed one of his lungs, and for a few months he lingered, being incapable of talking or moving very much, and in a great deal of pain. My aunts would not allow the children to see him in such sorry shape, except for one time, when he was dressed up for the entire family, which had gathered to quietly say good-bye. I was very sad when he died. When I was 15, my mother was diagnosed as having breast cancer. I was by her bed in the hospital when she recovered from surgery, and when she was told that her breast had been amputated in a radical mastectomy. I will never forget the look on her face. She survived the surgery, but the emotional resignation which she carried, and which preceded her cancer by several decades, was never diagnosed or discussed. Friends of our family had urged us to look into some alternative treatments for cancer, but everybody believed that the doctors in the hospital knew best. Listed as a "survivor" in the cancer statistics, my mother progressively declined after the surgery, and died about eight years later, having refused to undergo additional surgery.

My experience with relatives dying of cancer is not unusual, as degenerative disease is now at epidemic levels. The statistics today demonstrate that the "war on cancer" has been lost, and that in spite of all the radical surgeries, drugs and radiation treatments, patients today survive no longer or more frequently than they did in the 1950s. Indeed, degenerative disorders have today spread into youthful age groups and populations where they once were rare. Scientific evidence does not exist to support the assertion that surgery, radiation, and chemotherapy are effective forms of treatment for cancer, and traditional medicine today gives hardly more than lip service to preventative considerations. These troubling facts become all the more disturbing when one begins to study the various alternative, non-invasive and non-toxic cancer therapies. Dismissed for decades as "quackery" by organized medicine, most of these therapies appear to be reasonably or even remarkably effective. Their advocates and practitioners have often taken great risks to bring what they believe are safe and effective treatments to sick people. The organized medical community, with financial links to the pharmaceutical industry, has not cared to seriously look into these techniques. Instead, the techniques have been unwarrantedly attacked, and pseudo investigations have been launched, with predictable outcomes: clinics have been shut by brute police force, through court orders; medical records and research protocols have been seized, and jail sentences have been handed out. Books have also been burned. In this context, a great fraud has been perpetrated upon the American people, and upon our courts and legal system, by the larger organized medical associations, and related government bureaucracies.

In this short *Handbook*, I cannot give a history of these antiscientific and unethical abuses, but a few articles and books on the question are listed in the reference section. Clearly, a major reason for the impotence of modern medicine in dealing with degenerative diseases lies in the fact that the organized medical community has used police-state tactics to suppress important new findings, and the unorthodox practitioner, irrespective of any scientific evidence that exists. In fact, the most well documented and effective unorthodox therapies have been the most hotly attacked. Many, many promising therapies have been suppressed over the years, through an emotional collusion, and some outright conspiracy for economic motives, between medical interests, drug companies, plague journalists, prejudiced academics, and puritanical government bureaucrats and judges. In this process, the truth has been trampled badly, and the methods of science, discarded.

The most clear and telling example of how these social forces combine to kill a new discovery, and its discoverer, is the case of Dr. Wilhelm Reich and his orgone energy accumulator. Reich had been one of Freud's younger co-workers, and was a prime mover in the early psychoanalytic movement in Vienna and Berlin. However, his ideas were more revolutionary than those of the older psychoanalysts. He forcefully argued that human misery and mental illness were partly the result of social conditions, which could be changed to prevent neurosis. He urged reform in laws regarding the abusive treatment of children, and in the economic plight of women. He formed a viable social and political movement in pre-Hitler Germany, which urged reforms in laws restricting divorce and access to contraception, which were both then illegal. He championed the rights of young unmarried people to a healthy sexual life, and a loosening of the father's power and dominance over the family. He wrote extensively about these matters in the 1920s and 1930s, and exposed the roots of the Nazi movement in the obedience-demanding, patriarchal, sex-negative German family structure. For his writings on genitality, and his anti-fascist activities, Reich was eventually expelled from Freud's inner circle, and also from the International Psychoanalytic Association. German psychoanalysis was at that time leaning towards appeasement of the Nazis, and some analysts, such as Carl Jung, even became spokespersons or apologists for National Socialism. Reich was eventually placed on both Hitler's and Stalin's death lists in the 1930s, and had to flee to Scandinavia, and later from there to the United States. His writings were condemned to flames in both Germany and Russia.

By the time Reich arrived in the United States, in 1939, he had made a number of major scientific discoveries, and quickly attracted a group of young, enthusiastic scientists and doctors to assist with his work. The American period of his research, which lasted until his death in 1957, was particularly productive, in spite of the later court actions against him. It was during this period that Reich experimentally clarified, and made practical use of the biological and atmospheric life energy, which he called the orgone energy. The orgone energy, Reich observed, was a real, physical energy, which radiated from microbes, animals, humans, and from inorganic materials as well. It charged living creatures, yet also existed in the atmosphere in a free form. It could be accumulated in special enclosures, where it could be observed, felt, and measured.

Reich's orgone energy experiments attracted the hostile criticisms of many in the medical community, and a smear campaign in the press triggered an investigation by the U.S. Food and Drug Administration (FDA). In its efforts to stop Reich's work, the FDA did not seek to responsibly or accurately reproduce his experimental findings. Instead, FDA bureaucrats relied upon gossip and rumor, ignored available published evidence, and secured the testimony of "expert" witnesses who had no real familiarity with or interest in the scientific facts involved. Many breaches of the scientific method, and of legal, moral, and scientific ethics occurred, as government officials, and a number of psychiatrists, analysts, and physicists, sought to put an end to his work.

We also know a lot more today about the inner workings of the FDA regarding the Reich case, given that several scholarly reviews of FDA files, using the *Freedom of Information Act*, have taken place. These are cited in the reference section. Reich originally offered to cooperate with the FDA investigators regarding the orgone energy. But they refused his offer, as they were hot to "get Reich" on whatever charges they could, and had been urged in this direction by various highly-placed individuals within the medical communities. The *Bulletin of the Menninger Clinic*, and the *Journal of the American Medical Association*, for example, participated in the spreading of false gossip and rumors regarding Reich's work. Reich was aware of these unethical actions, and a number of his co-workers had been professionally hurt by the lies, and by the actions of the FDA. At no time did any governmental or private scientific or medical organization attempt to seriously review or replicate his experiments.

These kinds of attacks made Reich understandably furious, such that when the FDA finally sought a Complaint for Injunction against his activities, he refused to appear in court, to act, as he put it, "as a 'defendant' in matters of basic natural scientific research". Instead, he wrote a compelling *Response* ("Motion to Dismiss") to the judge. Reich's refusal to grant authority to the courts regarding the validity of his orgone research prompted a legal over-reaction, the likes of which we usually associate with despotic societies. In a judicial ruling that is, to the best of my knowledge, unique in American history, the FDA sought and obtained a Federal Court Decree of Injunction, which ruled that the orgone energy "does not exist". The court ordered a ban on the interstate shipment of books bearing the word "orgone". This included books into which the forbidden word appeared only in the preface, or introductory remarks. Additionally, all books and research journals discussing the orgone energy in detail were *ordered destroyed*, and devices using the energy, dismantled or destroyed. (Case #1056, March 19, 1954, US District Court, Portland, Maine, Judge John D. Clifford, Jr.)

"BANNED, until expunged of all references to the orgone energy:
- *The Discovery of the Orgone*
  - *Vol. I, The Function of the Orgasm*
  - *Vol. II, The Cancer Biopathy*
- *The Sexual Revolution*
- *Ether, God and Devil*
- *Cosmic Superimposition*
- *Listen, Little Man*
- *The Mass Psychology of Fascism*
- *Character Analysis*
- *The Murder of Christ*
- *People in Trouble*

BANNED and ORDERED DESTROYED:
- *The Orgone Energy Accumulator, Its Scientific and Medical Use*
- *The Oranur Experiment*
- *The Orgone Energy Bulletin*
- *The Orgone Energy Emergency Bulletin*
- *International Journal of Sex-Economy and Orgone Research*
- *Internationale Zeitschrift fur Orgonomie*
- *Annals of the Orgone Institute*"

And so, in the late 1950s and early 1960s, Reich's books and research journals, even those which were "only" banned, were periodically seized by FDA agents and Federal Marshals, and burnt in incinerators in Maine and New York. No scientific or professional organizations, journalists or writer's unions publicly objected to the book burnings, or acted to help Reich, whose laboratory headquarters was invaded by FDA agents, who destroyed instruments with axes. In addition to the above actions, the court ordered Reich to cease "disseminating information" on the orgone energy, effectively censoring his writing and speaking on the subject. Several years later, Reich was charged with Contempt of Court when his assistant committed a technical violation of the Injunction, at a time when Reich was more than a thousand miles away, engaged in field work in the deserts of Arizona. Though he appealed all the way to the Supreme Court, Reich lost the case, and was incarcerated in Lewisburg Federal Penitentiary, where he died in 1957. His death in prison occurred two weeks prior to his parole date, at a time when he was happily anticipating his freedom, and a life in Switzerland with his new wife.

Whatever we may think of Reich's response to the court challenge, the principles upon which he stood were very important, and date back at least to Galileo's trial by fire with the Catholic Church. The lesson from Galileo's time was that no Court, Tribunal, or religious or scientific organization on Earth has the capacity to say, on the basis of textual comparisons or divine revelation, just what is or is not Natural Law. The results of an experiment cannot be judged by those who have never reproduced it, and the unresearched opinions of scientists are no better than the unresearched opinions of anyone else, be they members of the American Medical Association, the National Academy of Sciences, or the same Country Club attended by the President. Galileo urged his critics to "look into the telescope", to verify his observations in a most direct and simple way. But they refused to do so on moral principle, and derisively mocked him. Reich's critics have taken the same approach, in their adamant refusal to reproduce his experiments, and in most cases, to even review the published evidence. Today, over 30 years since Reich's death, his most vocal critics still take the same anti-scientific approach, and condemn what they have not personally read or investigated.

Based upon these historical facts, it is clear that **the FDA, and indeed, all courts, academic bodies, and governmental agencies of all kinds, have forever relinquished any right to say anything about what the average citizen may or may not do with respect to the orgone energy accumulator**. The discovery of the orgone is in much safer hands among the average citizen, than in the hands of the various politicians, academies of science, and medical organizations. This *Handbook* is therefore not primarily directed towards an academic or medical audience. Instead, the case of Dr. Wilhelm Reich and the orgone energy accumulator is taken directly to the general public. Like the sunlight, the air and water, the orgone energy is a part of nature, existing everywhere, and ought to be available to everyone, free of restrictive regulation and control. As an invention, the orgone accumulator is also now in the public domain, unpatentable, and cannot be dominated by any single individual or corporation. It is also still *perfectly legal* for citizens to build, own, and use orgone accumulators.

Of course, with this right comes a great deal of responsibility, as the proper use and maintenance of an accumulator makes both social and environmental demands upon its owner. The cosmic orgone energy ocean is, like our air, food, and water, fast becoming contaminated and poisoned, and swift action by concerned people is required if life on the planet is to continue. This *Handbook* will give a basic overview of the orgone energy, the accumulator, and the construction and safe use of orgone accumulating devices. For the more precise scientific details and data, the reader is encouraged to obtain and review the published materials listed in the Reference and Information sections.

Three years after Reich's death, the trustees of his estate arranged for republication of his works. Today, most of his books have been republished, or are available in libraries, and through other sources. A group of Reich's co-workers also founded the American College of Orgonomy, and by 1967 had published the first issue of the *Journal of Orgonomy*. Since that time, interest in Reich's works has gradually increased, and many new studies verifying his findings on the orgone energy, and the accumulator, have taken place. College courses focusing upon Reich's life and works now exist, and he has been the subject of many reviews, biographies, and films. A younger generation of scientists and health practitioners is rediscovering Reich, and new journals devoted to orgone research or bioenergetic functions have appeared in the USA, Germany, France, Italy, England, Canada, and Japan. One major new study on the accumulator from Germany is summarized in the Appendix.

The effort to kill the discovery of the orgone has failed.

Part I: The Biophysics of the Orgone Energy

3. What Is the Orgone Energy?

Orgone energy is cosmic life energy, the fundamental creative force long known to people in touch with nature, and speculated about by natural scientists, but now physically objectified and demonstrated. The orgone was discovered by Dr. Wilhelm Reich, who identified many of its basic properties. For instance, the orgone energy charges and radiates from all living and non-living substance. It also can readily penetrate all forms of matter, though with varying rates of speed. All materials affect the orgone energy, by attracting and absorbing it, or by repelling or reflecting it. The orgone can be seen, felt, measured and photographed. It is a real, physical energy, and not just some metaphorical, hypothetical force.

The orgone also exists in a free form in the atmosphere, and in the vacuum of space. It is excitable, compressible, and spontaneously pulsatile, capable of expanding and contracting. The orgone charge within a given environment, or within a given substance, will vary over time, usually in a cyclical manner. The orgone is most strongly attracted to living things, to water, and to itself. Orgone energy can lawfully stream or flow from one location to another in the atmosphere, but it generally maintains a west to east flow, moving with, but slightly faster than the rotation of the Earth. It is a ubiquitous medium, a cosmic ocean of dynamic, moving energy, which interconnects the whole physical universe; all living creatures, weather systems, and planets respond to its pulsations and movements.

The orgone is related to, but quite different from other forms of energy. It can, for instance, impart a magnetic charge to ferromagnetic conductors, but it is not magnetic itself. It can likewise impart an electrostatic charge to insulators, but neither is it fully electrostatic in nature. It reacts with great disturbance to the presence of radioactive materials, or to harsh electromagnetism, much in the manner of irritated protoplasm. It can be registered on specially adapted Geiger counters. The orgone also is the *medium* through which electromagnetic disturbances are transmitted, much in the manner of the older concept of *aether*, though it is not itself electromagnetic in nature.

Streamings of orgone energy within the Earth's atmosphere affect changes in air circulation patterns; atmospheric orgone functions underlie the buildup of storm potentials, and influence air temperature, pressure, and humidity. Cosmic orgone energy functions also appear to be at work in space, affecting gravitational and solar phenomena. Still, the mass-free orgone energy is not any one of these physico-mechanical factors, or even the sum of them. The properties of the orgone energy derive more from life itself, much in the manner of the older concept of a *vital force*, or *élan vital*; unlike those older concepts, however, the orgone also has been found to exist in a mass-free form, in the atmosphere and in space. It is primary, primordial cosmic *life energy*, while all other forms of energy are secondary in nature.

In the living world, orgone energy functions underlie major life processes; pulsation, streaming, and charge of the biological orgone determines the movements, actions, and behavior of protoplasm and tissues, as well as the strength of "bioelectrical" phenomena. Emotion is the ebb and flow, the charge and discharge of the orgone within the membrane of an organism, just as weather is the ebb and flow, the charge and discharge of the orgone in the atmosphere. Both organism and weather respond to the prevailing character and state of the life energy. Orgone energy functions appear across the whole of creation, in microbes, animals, stormclouds, hurricanes, and galaxies. Orgone energy not only charges and animates the natural world; we are immersed in a sea of it, much as a fish is immersed in water. More, it is the medium which communicates emotion and perception, through which we are connected to the cosmos, and made kin to all that is living.

4. Wilhelm Reich's Discovery of the Orgone Energy, and Invention of the Orgone Accumulator

Reich's initial work on the question of a biological energy began in the 1920's, when he was a student of Sigmund Freud, the originator of psychoanalysis. Freud's early theories on human behavior discussed in metaphorical terms the energy of the drives, which he termed the *libido*. While Freud and most other analysts eventually ceased using this term, Reich found it to be a very useful concept, and he continued to seek evidence for this force, which governed human emotion, behavior, and sexuality.

Reich's extensive clinical work led to the observation of *vegetative streamings* or *currents* of emotional energy in the body, which occurred in healthy individuals during states of great relaxation, as following a strong release of emotion, or after a very gratifying genital orgasm. The free, uninhibited expression of emotion, and natural sexual excitation and gratification during orgasm were identified by Reich as expressions of unimpeded energetic movement in the body. When the individual experienced great pain, as from childhood traumas, when the emotions were rigidly suppressed and held back ("big boys don't cry", "nice girls don't get angry"), or when chronic sexual stasis and starvation was experienced, the entire nervous system and musculature participated in the process of emotional suppression, or warding off of feeling. This "holding back" of feeling was also accompanied by a greater or lesser anxious retreat from pleasurable, or even potentially pleasurable situations, which would otherwise stir up suppressed and unpleasant feelings. Reich observed that when this kind of response to feeling and pleasure became chronic, so too did the individual experience a chronic stiffening and desensitization of the body, along with a reduction in respiration and contactfulness.

This chronic neuromuscular *armoring*, as Reich called it, was not a natural condition, though it had a certain rational survival value for situations of pain and trauma. When the armoring became chronic, however, as a *way of life*, it would thwart the individual's natural biological functioning, and affect their behavior even in circumstances where pain or trauma was not likely. The armoring effectively perpetuated the individual's pleasure-avoiding behaviors and emotion-censoring attitudes. Deep seated fears, and pressures to conform to the prevalent armored form of social life, usually prevented the individual from moving towards emotional health, or taking effective steps to change their situation. The bulk of Reich's early writings focused upon these social, sexual, and emotional concerns.

Reich also argued that the heterosexual genital orgasm played a central regulatory role in the energy economy of the individual, as a means to periodically discharge accumulated bioenergetic tension. The more intense the orgastic discharge of accumulated bioenergy, the more gratified, relaxed and pleasurably expansive one felt afterward. When sexual urges and other emotions were chronically frustrated, dammed-up, and repressed, however, great internal tension could build up to a bursting point, where neurotic symptoms or sadistic urges would appear. Reich developed therapeutic techniques for releasing dammed-up emotional energy within his patients, techniques which led to the release of long-buried feelings, and to a greater capacity for pleasure in life, particularly genital pleasure. As his patients became more healthy sexually, and as they reported an increase in genital gratification, he observed that their neurotic symptoms disappeared, as the quantity of dammed-up emotion and sexual tension was reduced. Some of Reich's early contributions to psychoanalytic theory and technique were at first welcomed. But later on, as he increasingly focused upon the consequences of child abuse and sexual repression, the more conservative analysts rejected and attacked him.

Reich eventually left psychoanalysis altogether, and he articulated his work under the new term, *Sex-Economy*.

Reich's early observations regarding human behavior, emotions, the orgasm, and vegetative streaming sensations strongly suggested a real, tangible nature to the emotional energy. He later used sensitive millivoltmeters to confirm this point of view, and to quantify bioelectrical energy currents, and their emotional correlates. However, he was convinced that the very low levels of observed bioelectrical activity could not fully explain the powerful energy forces observed in human behavior. This was particularly so regarding chronic immobilizing psychic disturbances in catatonic and other completely withdrawn mental patients. When their emotions were finally broken lose, these patients would experience a tremendous outpouring of sadness or rage. Afterward, they would also experience a dramatic relaxation of musculature, a spontaneous deepening of respiration, and a return to more contactful lucidity. In these cases, the patient's emotional energy was held down and bound up, until finally set free in the clinical setting. These observations of energy bound, and energy released, were reinforced by parallel observations regarding the discharge function of the orgasm. Based upon these kinds of observations, the question of exactly how and from where the organism acquired its emotional energy, and its exact nature, became increasingly important.

It was at this point in his research that Reich was forced to flee Germany for Scandinavia, following Hitler's rise to power. In Norway, Reich sought to find a way to confirm his model of human functioning. Pleasure, he observed, was identified by an increasing bioelectrical charge at the skin surface, while anxiety was accompanied by a loss of this same peripheral bioelectrical charge. Persons with a deep respiration and a relaxed posture would, he observed, regularly give stronger readings at the millivoltmeter than contracted, anxious, highly armored persons, who had a life-history of trauma, abuse, repressed emotion, and ungratified sexuality. As a child grew to adulthood, and became habituated or conditioned towards pleasure-seeking, or pleasure-avoiding (pain seeking) behaviors, so too would their skin charge, and other physiological measures, reflect a corresponding high or low energy charge. This movement of the organism, and their energy charge, in a direction "toward" or "away" from the world, he argued, was the result of one's life history. Life naturally moved towards pleasure, but retreated, and shrunk from pain. Chronic painful experience would eventually armor the organism, and make it difficult for them to reach out, towards the painful world. From this central set of observations, he postulated that a similar process could be duplicated and observed in lower organisms, such as the snail, earthworm, or even microscopic ameba.

Reich noted that the ameba had no "nervous system", or "brain", as with the higher animal, yet it expanded towards or contracted away from its environment in a manner similar to the higher animals. He believed that many of the functions attributed to the brain were really functions of whole-body processes, involving the participation of the autonomic nervous system, but primarily being the result of the energetic forces he had documented in a clinical and laboratory setting. These currents of biological energy, he argued, functioned the same in all living creatures, and he sought to test the idea by making millivoltmeter measurements of ameba during states of expansion and contraction. Reich went to the University of Oslo Microbiological Institute, and asked to obtain a culture of ameba. He was told that these kinds of simple organisms were never kept on hand in stored cultures, because they could be cultured directly from a moss or grass infusion. Reich was fully aware of the air germ theory, but was surprised to hear this, as the theory had not at that time been used to explain the genesis of more complex microbes, such as ameba and paramecium. These more complex microbes cannot be cultured directly from the air, for example.

Reich made the moss and grass infusions, but also made extended and careful microscopical observation of the process whereby the ameba developed. He did not see spores on the grass blades, swelling up to become new ameba. Instead, he observed that the moss and grass itself would disintegrate and break down into small blue-green vesicles. The tiny vesicles would, over a period of several days, develop and clump together, after which a new membrane would form around the clump; the clump of vesicles would roll and pulsate inside the membrane for a period, and eventually the whole thing would move away on its own, having *turned into a new ameba*. Moreover, Reich observed that a number of materials, both organic and inorganic, would, when allowed to disintegrate and swell in a sterile nutrient solution, form the tiny blue-green vesicles. These observations were greeted by the university microbiologists with skepticism, and Reich developed a series of stringent control tests to answer their objections, and to more clearly demonstrate the observable process. These control procedures involved lengthy autoclaving of nutrient solutions, and heating over flame, to incandescence, of the materials placed in the sterile nutrient medium. His control procedures and observations on this question were repeated and confirmed by other scientists of the day, and were presented to the French Academy of Science in 1938. But this did little to satisfy his critics, who shamelessly refused to reproduce the experiments, while simultaneously attacking him in the Norwegian newspapers.

Reich used very high magnifications, around 3500 to 4500 power, but not the usual microbiological stains or procedures which kill the life in the specimen. These facts made Reich's preparations very different from those of the average microbiologist, who to this day still kill and stain their preparations with a religious fervor, and see little value to observing living microbes in the light microscope above 1000 power. Standard electron microscope images, for example, can not be made of living specimens.

Reich gave a new name to the unusual microscopic vesicle he had discovered: the *bion*. Bions of similar size, shape, and motility would appear in the light microscope when various materials were subject to a process of slow swelling and disintegration, or when substances were heated to incandescence, and then immersed into sterile nutrient solutions.

Boiling, autoclaving, or heating samples to glowing incandescence would not eliminate the bions from the cultures, but could actually liberate them in greater numbers. Reich also studied the process of disintegration and decay of foodstuffs in the microscope, and noted that similar bionous processes were at work. The bions exhibited a *bluish* coloration, and radiant energy effects were likewise observed. It was during these microscopic observations of the bions that Reich first discovered the orgone radiation, and later, the principle of the orgone energy accumulator.

Like his findings on human behavior, Reich's bion experiments are far too intricate and important to be fully reviewed here, but it can be noted that they have been widely replicated by various scientists around the world. Classical microbiology of today has made confirming discoveries of very similar small vesicles, though Reich's priority has yet to be acknowledged. His findings on the bions also resolved two parallel riddles, the origins of protozoa from disintegrated dead plant tissue in the natural environment, and the origins of protozoan *cancer cells* from the energetically (emotionally) deadened tissues of the human body. Reich observed similar processes at work in both dead grass and deadened animal tissue: disintegration into bions, followed by a spontaneous reorganization of bions into protozoan forms. In both cases, of soil or tissues, Reich argued that the process was initiated by a *loss of life energy charge* of the tissues, followed by putrefaction and disintegration.

One special bion preparation, made from pulverized beach sand heated to incandescence and immersed into a sterile nutrient broth, yielded a powerful radiant energy phenomena. Lab workers developed conjunctivitis if they observed the preparations too long, while a skin inflammation could be developed by placing the bion solution close to the skin for a period. Working for extended hours in the laboratory, Reich developed a dark tan through his clothing, in the middle of winter. The radiation imparted a magnetic charge to nearby iron or steel implements, and a static charge to nearby insulators, such as rubber gloves. Film stored in nearby metal laboratory cabinets spontaneously fogged. He noted that whatever this bion radiation was, it was rapidly attracted to metals, but just as rapidly reflected away, or dissipated into the surrounding air. Organic materials, however, absorbed this radiation and held onto it. Attempts to identify the new radiation using traditional nuclear or electromagnetic radiation detectors failed.

Reich also noted that the air in rooms containing the special bion cultures would feel "heavy" or charged. When observed at night, in full darkness, the air would visibly scintillate and glow with a pulsing energy. He attempted to capture the energy radiating from his bion cultures inside a special cubical enclosure lined with sheet metal, which he felt would reflect and trap the radiations inside. As expected, the special metal-lined enclosure trapped and amplified the effects of the bion radiation. However, to his amazement, he found that the radiation was also present in the experimental enclosure *even when the bion cultures were removed*. In fact, there was nothing which could be done to make the observed radiation "go away". The special metal-lined enclosure appeared to pull the same form of radiation from the air which previously had been observed coming from the bion cultures.

Reich eventually became convinced that the special enclosures were capturing a free atmospheric form of the same energy that he also observed coming from living organisms. He called the newly discovered energy the *orgone*, and he developed ways to amplify the energy accumulating affects of the enclosure, mainly through multiple layering of the metallic and organic materials. No electricity, magnetism, electromagnetism, or nuclear radiations were employed in these accumulating structures, which were entirely passive in design. The special enclosures were thereafter called *orgone energy accumulators*.

The full sweep of Dr. Reich's clinical findings, his experiments with bioelectricality, the bions, on biogenesis and the origins of the cancer cell, and his discovery of the orgone energy and the orgone energy accumulator, cannot be given here, but a few points are summarized. The orgone accumulator was found to have specific life-positive effects upon plants and animals exposed to the concentrated life force inside it. A host of quantifiable effects upon the physical properties of the air, or other materials charged up inside the accumulators, were also discovered and documented. Reich and his co-workers published a host of research articles on the orgone energy accumulator, its unusual physical properties, and its life-positive biomedical effects. These effects have been repeatedly confirmed, and a research tradition in orgone biophysics continues to this day. Briefly, we may identify some of the known properties of the orgone energy, and effects of the orgone energy accumulator.

Properties of the Orgone Energy:

A) Ubiquitous, fills all space.
B) Mass-free; cosmic, primordial in nature.
C) Penetrates all matter, but at different rates of speed.
D) Spontaneously pulsates, expands and contracts, and flows with a spinning wave characteristic.
E) Directly observable and measurable.
F) Negatively entropic.
G) Strong mutual affinity and attraction to/by water.
H) Accumulated naturally by living organisms through foods, water, breath, and through the skin.
I) Mutual excitation and attraction of separate orgone energy streams, or of separate systems charged with orgone.
J) Excitability via secondary energies (nuclear, electromagnetism, electrical sparking, friction).

Physical effects of a strong orgone charge:

K) Slightly higher air temperature as compared to surroundings.
L) Higher electrostatic potential, with a slower electroscopical discharge rate as compared to surroundings.
M) Higher humidity and lower water evaporation rates as compared to surroundings.
N) Squelching of ionization effects inside gas-filled ionization Geiger-Muller tubes.
O) Development of ionization effects inside non-ionizable vacuum tubes (0.5 micron pressure or lower).
P) Ability to impede and absorb electromagnetism.

Biological effects of a strong orgone charge:

Q) General vagotonic, expansive effect on entire system.
R) Sensations of tingling and warmth at skin surface.
S) Increased core and skin temperature, flushing.
T) Moderation of blood pressure and pulse rate.
U) Increased peristalsis, deeper respiration.
V) Increased germination, budding, flowering and fruiting of plants.
W) Increased rates of tissue growth, repair, and wound healing, as determined through animal studies and human clinical trials.
X) Increased field strength, charge, integrity of tissues and immunity.
Y) Greater energy level, activity, and liveliness.

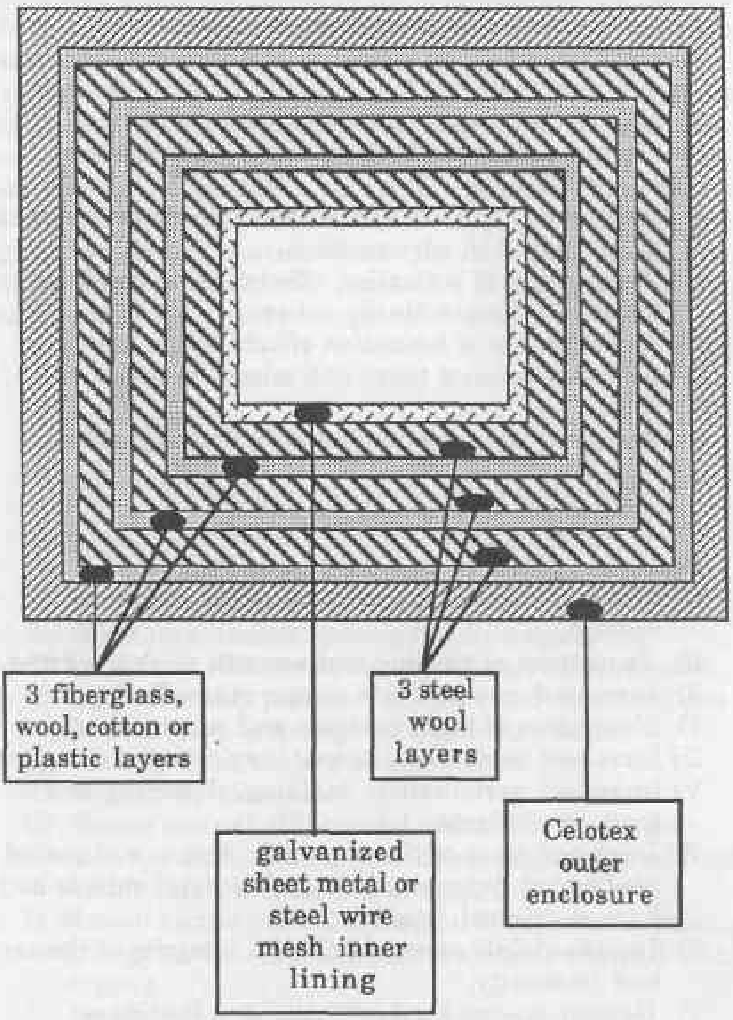
3 fiberglass, wool, cotton or plastic layers
3 steel wool layers
galvanized sheet metal or steel wire mesh inner lining
Celotex outer enclosure

Simplified Diagram of the Orgone Energy Accumulator

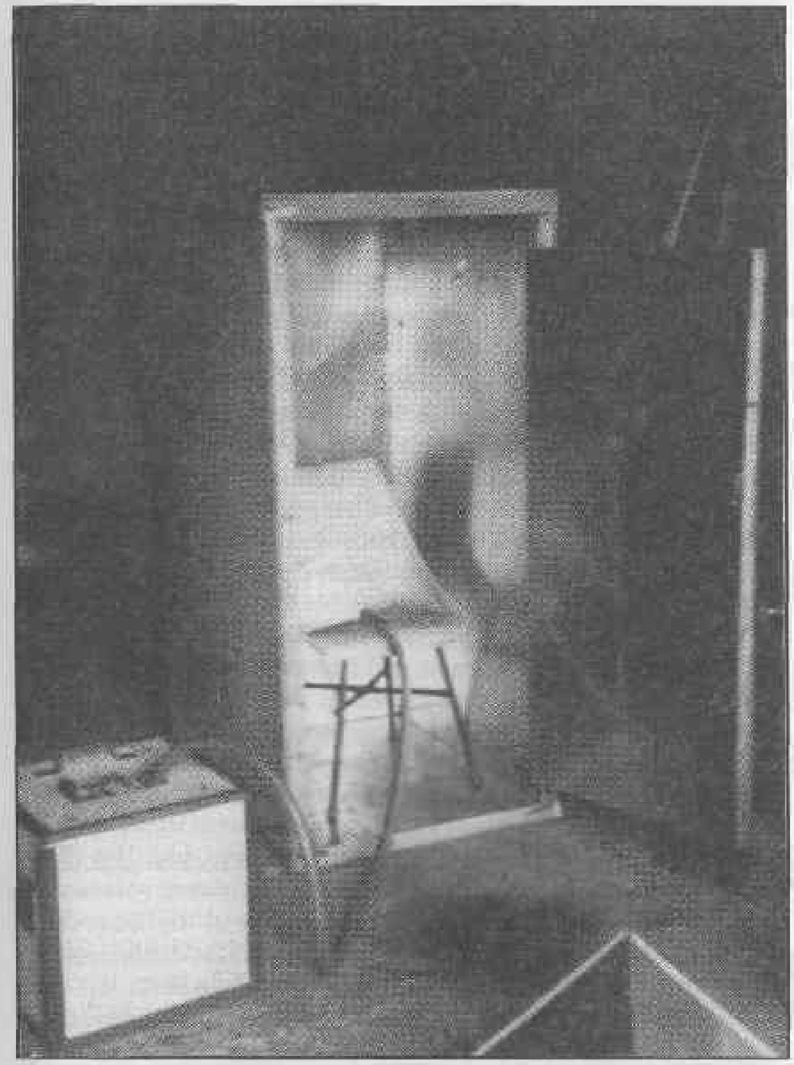

A three-ply accumulator in the author's laboratory. A ten-ply charger, with attached funnel shooter, is at the lower left. This charger normally sits inside the larger accumulator, under a wooden bench inside the accumulator.

5. Objective Demonstration of the Orgone Energy

A number of techniques have been developed over the years to document, measure, and objectify the orgone energy. These techniques are briefly listed, but the interested reader is directed to the references for more specific details.

A) Bioelectric Fields: Reich identified various bioelectrical phenomenon which he felt demonstrated a more powerful energy current at work in the body. The small millivolt currents of "bioelectricality", he argued, were only a small portion of this stronger energy current in the body, which he identified as being both emotional and sexual in nature, and which was later objectively identified as the orgone energy.

B) Radiant Effects from Bion Cultures: Special bion cultures derived from beach sand emitted a powerful radiation which could be felt and seen in darkrooms. This radiation did not register on instruments for detecting nuclear or electromagnetic energies. Additionally, the radiation could fog film, impart a static charge to insulators, and a magnetic charge to steel laboratory implements.

C) Darkroom and Atmospheric Observations, the Orgonoscope: Reich also observed and categorized various observable phenomena which could be seen by the dark-adjusted eye in the open air of darkrooms. Scintillating fog-like forms, and dancing, luminescent pin-points of light were observed, and numerous techniques were developed which demonstrated their real, objective nature. One of these techniques involved the development of a new instrument, the *orgonoscope*, which used hollow tubes, lenses, and a fluorescent screen for magnifying the various subjective light phenomena. Large room-sized orgone accumulators were also constructed, and observations made in these amplified and clarified many of the effects. A special corpuscular *orgone unit* was identified, whose lawful behavior changed according to cosmic and meteorological factors. These macroscopic particles were also observed in the daytime sky with the unaided eye, as a generally common phenomenon, visible to most people once they were pointed out. The Earth was observed to possess its own orgone energy envelope, or energy field, much as did individual living creatures.

D) X-Ray Photographs: Reich observed that the x-ray "ghost" phenomenon (spontaneous, unexplained fogging of x-ray films) could be explained as an effect of the orgone radiation, or life energy. He published several photos where the ghosts were purposefully created by excitation of orgone energy within the field of the x-ray machine.

E) Visible Light Photographs: Reich observed that his special radiating bion cultures would fog film stored in nearby metal cabinets. Culture dishes of radiating bions set directly upon the film would also render an image of the culture dish and its contents. More recently, Thelma Moss of UCLA has shown that life energy field photos can be made without electrical stimulation (as with Kirlian techniques), by energy field enhancement; living objects placed directly on a film for a few days within a darkened orgone accumulator will, under the proper conditions, render an image.

F) The Orgone Energy Field Meter: Reich developed this device to measure the strength of energy fields. Using a Tesla coil and special accumulator-like metal plates, the device could quantify the differences in energy level between people or objects.

G) The Orgone Energy Pulsation Demonstrator: Reich demonstrated that the energy field pulsations of a large metal sphere were capable of setting into motion a smaller metallic/organic pendulum suspended nearby.

H) The Accumulator Temperature Differential (To-T): An accumulator will spontaneously develop a slightly higher temperature than either its surroundings or a control enclosure, on those sunny and clear days when the orgone charge at the Earth's surface is strong. The effect vanishes during stormy, rainy weather, when the orgone charge at the

Earth's surface is weak (but strong in the Atmosphere). The results of this temperature experiment, which has been replicated many times, demonstrates that the orgone energy functions in opposition to the second law of thermodynamics.

I) The Accumulator Electrostatic Effects: An electroscope kept inside an orgone accumulator will dissipate its charge more slowly than an identical one kept in the open air, or inside of a control enclosure. A partly charged, or uncharged static electroscope which is kept inside an accumulator will sometimes spontaneously charge itself up. As with the temperature differential effect, the electrostatic effects vanished during rainy or overcast weather, when the orgone charge at the Earth's surface is weak.

J) The Accumulator Ionization Suppression/Amplification Effect: Geiger-Muller tubes and counters charged up inside a very strong accumulator for several weeks or months tend to go "dead" for a period, but may eventually yield an erratic count rate for background. Special vacuum tubes, which Reich had constructed and called vacor tubes (which mimic the design of the Geiger-Muller tube but are evacuated well below the level at which ionization could occur), will initially not yield any counts when hooked to a radiation detector. After charging up inside a very strong accumulator for weeks or months, however, these same vacor tubes will begin to yield very high counts per minute for background, even at very low exciting voltages. The results of this experiment run counter to the classical interpretation of the ionization effect inside the Geiger-Muller tube, and hence, to the classical particulate interpretation of radioactive decay.

K) The Accumulator Humidity/Water Evaporation Effect (EVo-EV): More recent studies have suggested that an accumulator tends to attract a slightly higher humidity into itself, and suppress the evaporation of water from an open vessel inside. As with other accumulator phenomena, this effect also diminishes or vanishes during rainy weather.

L) Atmospheric Energetic Pulsation, and the Reversed Orgonotic Potential: Based upon observations of the thermal, electroscopical, and ionization characteristics of the orgone accumulator, Reich identified a patterned and lawful energetic cycle at work within the atmosphere and energy field of the Earth. These observations likewise led to the identification of a reversed potential at work in the orgone energy, running counter to thermodynamic principles, which explained why natural orgonotic systems (organisms, weather systems, planets) maintained a higher concentration of energy than their surrounding environments. The stronger of two orgonotic systems will drain energy from the weaker system, and increase its own potential or charge, until the weaker system is drained, or some maximum capacity level is achieved. Discharge may occur thereafter. Under sunny, clear weather, the orgone charge at the Earth's surface is quite strong, and in a state of expansion, preventing any significant growth of clouds. Where overcast or stormy conditions prevail, the orgone charge at the Earth's surface is weak, while the charge in the atmosphere is strong and in a state of general contraction. This loss of charge at the Earth's surface during rainy weather slows down the activity of living creatures, and the accumulator will not function well at those times.

M) The Millivoltmeter: Virtually all objects and organisms within a given environment, to include the air, water, and the Earth itself, have a charge that will increase and decrease in a cyclical or pulsatory manner, timed to cosmic and meteorological factors. In living creatures, high potentials produce more active physical and emotional periods, while low potentials signal less active periods. In Nature, high atmospheric potentials signal cloudy periods with stronger storms, while high Earth potentials signal cloud-free conditions. The tiny currents are excellent predictors of powerful biological or environmental processes, but are themselves too slight and weak to be the causative agent. Reich, and other researchers who have examined these small voltage potentials (such as H.S. Burr), viewed them as indicative of a more powerful, ubiquitous natural phenomena which energetically linked together the Sun, Moon, Earth, weather systems, and all living creatures.

N) Plant Growth Enhancement Studies: Seeds and plants correctly charged up inside an accumulator demonstrate higher growth rates and yields of fruit. This is one of the more telling, and widely replicated experiments with the orgone accumulator. In my own trials, I have seen a six-fold increase in the length of mung bean sprouts inside a strong accumulator, as compared to a control group of sprouts. Germination rates, growth rates, budding, flowering and fruiting can be increased by charging seeds or growing plants directly inside the accumulator. The seeds can be sprouted directly inside the accumulator, or charged up for a few hours, days, or weeks prior to planting. Growth enhancement effects may also occur when the water alone is charged, and then externally given to the plants.

O) Non-Human Animal Studies: Controlled studies on the effects of orgone radiation from an accumulator on cancer mice and wounded mice have been performed. These studies generally confirm Reich's earlier arguments that tissues with a stronger energetic charge will heal more quickly, and develop tumors more slowly, or not at all, as compared to energetically weakened tissues. These findings invalidate many aspects of the DNA theory of cellular differentiation, which appears to be more directly under the structuring influence of the organism's own life energy field.

P) Human Studies: Aside from those clinical trials performed by Reich and his co-workers in the 1940s and 1950s, very little work has been done in the USA regarding the bioeffects of the accumulator on humans. All research into these questions was halted by medical police actions in the 1950s. Recent studies from Germany, however, have confirmed such bioeffects. In general, a person sitting inside an accumulator will feel a variety of warm, glowing, or sometimes tingling sensations at the skin surface; their body core temperature will rise and skin flush, while blood pressure and pulse rate will trend towards moderate levels, being neither too high or low. When properly used, it has a distinct vagotonic, enlivening effect. The chapter on

"Physiological and Biomedical Effects", as well as the Appendix article, will provide details on these questions.

6. Discovery of an Unusual Energy by Other Scientists

Studies by various natural scientists have, over the years, demonstrated energetic principles at work in the natural world which are similar to the orgone energy. Early Chinese medicine acknowledged the existence of such a force, called *Chi*, and the traditional method of acupuncture is based upon the existence of such an energy principle within the human body. Acupuncture points do not correspond directly with nerve endings, and the most able acupuncturists do not rely upon the Western models of physiology to explain its effects. Given the absence of a vital energy principle, Western medicine cannot explain acupuncture, and has resisted its adoption in the United States. Acupuncture furthermore works on animals, invalidating invocation of the placebo effect; ancient texts from India have also referred to the life energy, called *Prana*, and provide maps of *Nila points* (similar to acupuncture points) on elephants. The texts from ancient China and India speak about an energy that is taken in through the breath, and flows through the body along the various meridians. Health is constituted by a free, unimpeded flow of this energy, while sickness occurs when the flow of vital energy is blocked. This is very similar to Reich's ideas on the orgone energy, though the Asian sources say little about the free expression of emotion; they also often advocated a conscious control of emotions and sexual feeling (*orgasm avoidance*). In contrast, Reich demonstrated that such chronic restraint or self-control was the reason why the life energy became blocked or dammed up in the first place.

In the Western tradition, the vitalists of the eighteenth and nineteenth centuries also discussed the existence of a biological energy or life force, which was called *animal magnetism*, the *odic force, psychic force, élan vital,* and so on.

Indeed, Mesmer spoke of animal magnetism as an atmospheric fluid which surrounded, charged, and animated living creatures, and could be projected across a distance by a therapist. Mesmer was a teacher of Charcot, who was, in turn, the teacher of Freud, who was one of Reich's early mentors. Reich also studied with other vitalists, such as Kammerer and Bergson, and the vitalist tradition has persisted as a quietly spoken minority viewpoint in biology. Besides Reich, the more recent advocates of a vital or dynamic energy principle in nature included the late Harold S. Burr of Yale University. Burr argued for the existence of a powerful *electrodynamic field* at work in nature, affecting both weather and living creatures. The biologist Rupert Sheldrake has similarly developed a theory on *morphogenetic fields* which also takes from this tradition. Like Burr's work, Sheldrake's theory provides a dynamic, energetic explanation for inheritance, making the biochemical DNA theory unnecessary. Most recently, editors for the academic publication *New Scientist* called Sheldrake's book the "best candidate for burning" they had seen in a while.

The surgeon Robert O. Becker developed these prior principles to a most amazing state of advancement. His early research led to the development of a class of devices for the electrical stimulation of bone healing and pain relief. His later work took these principles, and developed them to the point that he could artificially stimulate the *regenerative growth of amputated limbs of laboratory mice*, in a manner similar to the way in which a salamander or spider will regrow a lost limb. This kind of regrowth is by nature limited only to less complex creatures, and does not generally exist among mammals, such as mice, rabbits, and humans. Regrowth of an amputated limb had never been previously demonstrated in a mouse, or any mammal for that matter. Becker's work was a severe blow for both the biochemical DNA theory of cellular regulation, and the theory that the bioelectrical field of a creature was just a meaningless "byproduct" of chemical metabolism, like the electric field surrounding a running automobile engine. His work proved that the energy field of the animal was a primary determinant of growth and repair, as was the case with Reich's work. Becker was preparing to replicate the limb regrowth experiments on humans, when the biomedical community reacted with severe outrage against him, pulling dirty tricks of all sorts to have his research funding cancelled, and his laboratory shut down.

Another vitalist of our era is Bjorn Nordenstrom, director of the Karolinska Radiological Institute in Sweden. Nordenstrom, like Reich, made a study of the x-ray "ghost" phenomena, which is an unusual spontaneous fogging of x-ray films. It appears as a wispy, smoke-like or blob-like form on the x-ray images of patients, and sometimes can be seen on the monitors of airport baggage x-ray equipment. It cannot be predicted, and most radiologists consider it to be a nuisance. However, Nordenstrom studied it, and observed distinct patterns correlated to his patient's bioelectrical fields. Like Reich, he also discovered and measured currents of bioelectricality in the body. His meticulous research was summarized in a book titled *Biologically Closed Electric Circuits: Clinical, Experimental, and Theoretical Evidence for an Additional Circulatory System*. After being heavily advertised in medical journals in the USA, it sold less than 200 copies, evidencing a contempt among mainstream medical doctors for any new findings that would support the principle of a life energy, even one of a purely bioelectrical nature. Unable to find support for his work in the West, Nordenstrom recently went to China to pursue his clinical research.

Other biological scientists have inferred the existence of such a vital energy principle, on the basis of their experimental work. When they provide good confirming evidence, they are hotly attacked. The French scientist Louis Kervran, for example, spent years developing very elegant and simple experiments demonstrating that the basic elements of chemistry were being *transmuted* by living creatures. Chickens fed a diet free of calcium, for example, would not lay mushy or fragile eggs, unless dietary silica was restricted. With a restricted silica intake, however, they laid mushy and fragile eggs, and it did not matter how much calcium they ate. Likewise, laboratory mice would heal broken bones very quickly when fed a diet high in organic silica, but not so fast when silica was minimized and only calcium was provided. These experiments strongly suggested that dietary silica was being transmuted into calcium in the bodies of animals. Kervran also experimentally demonstrated other likely transmutations, and other scientists in Europe and Japan confirmed his findings. He eventually came to the conclusion that there had to be some unknown form of powerful biological energy at work to drive the transmutations. But when he wrote to a prominent American scientist for assistance in obtaining equipment for an important experiment, he was impolitely told to go "read an introductory textbook on biology". In the United States, Kervran is better known among homeopathic doctors and organic farmers than by university professors. However, if Kervran is right–and the experimental evidence suggests that he is–then the textbooks of biochemistry will need to be rewritten. As Kervran pointed out, biology and biochemistry are two entirely different disciplines and should not be confused. Biology is concerned with observable fact, while biochemistry attempts to explain observed facts by a chemical theory which assumes elemental constancy. And it is with this basic assumption that part of the error lies.

Another French scientist, Jacques Benveniste, actually demonstrated such an energy principle at work in homeopathic dilutions. His experimental work was successfully replicated by independent laboratories in other countries, to satisfy his obstinate critics. But that was not good enough. For making this offending discovery, which lent some support to homeopathic physicians (who are often prosecuted and jailed in the United States), the science journal *Nature* dispatched a "hit squad" of fraud investigators, debunking magicians, and skeptical editors to his laboratory, under the guise of "evaluating" his laboratory procedures. The *Nature* science cops made a mess of Benveniste's lab, distracting laboratory workers, performing slight-of-hand tricks, and shouting, before finally being told to leave. *Nature* subsequently tried to smear Benveniste in their editorials, but did not factually refute his work through replication of experiments. Such is the warp and weave of traditional academic science in the 1980s.

In the atmospheric sciences, the tradition of dynamic energy forces which affect entire regions was preserved for a period by older weather forecasters, who used streamline, rather than frontal theory to predict weather. Streamline analysis more coherently focused upon streaming movements of air, or jet streams as they are called today. For example, when you look at the dynamic images of clouds, as seen from a satellite in space, you do not see "fronts". But you do see *streaming movements of clouds*. Reich independently discovered the basic configurations of these streams, years before the first weather satellites were launched. Likewise, the older atmospheric scientists often argued for a great interconnectivity in the atmosphere. Charles G. Abbot, head of the Smithsonian Astrophysical Observatory in the 1950s, used related energetic concepts to predict the weather months into the future. But he was ignored and ridiculed for his findings, in spite of their uncanny accuracy. Irving Langmuir, one of the originators of cloudseeding techniques, once objectively demonstrated that cloudseeding in New Mexico would trigger rain storms all the way into Ohio, and he warned his fellow workers about this danger. The cloudseeders of today, funded by millions in Federal dollars, act as if Langmuir's work never took place, and refuse to replicated his simple experiment. They deny the existence of long distance effects from cloudseeding, knowing that if such effects became public knowledge, they would be forced to stop.

Among the physical scientists, the idea of an energy in space was embodied in the concept of an *aether,* which dates back hundreds of years. The theologian/physicist Isaac Newton forcefully argued that this aether *had to be static*, in order to prevent it from directly participating in the movement and ordering of the heavens. That role, Newton argued, belonged only to the anthropomorphic God (who at that time was demanding that unbelievers be ruthlessly tortured and burned at the stake). And, over the years, a dead, unmoving aether has never been detected. However, an *aether with more* *dynamic properties* was objectively demonstrated by the physicist Dayton Miller. Miller also explained why prior attempts to measure the aether had failed. First, he observed that the aether is *entrained* at the Earth's surface, and moves faster at higher altitudes than lower altitudes. Prior attempts to measure its movement had taken place only at lower altitudes, or in heavy stone buildings or basement locations. Second, Miller's aether was *reflected by metals*, and prior attempts to measure it used instruments with the critical parts housed inside metal enclosures. Miller found that by doing the crucial aether-drift experiments on a mountain top, inside a flimsy building without metals or dense window materials, that it was readily detectable and measurable. He made over 200,000 separate measurements, over the course of 30 years of investigation. Contrast this to the famous Michelson-Morley experiment, which involved a grand total of six hours of actual measurement time, made over four days in 1887. The Michelson-Morley experiment is widely misquoted as having completely failed in the detection of the aether. It was a hinge-point in the sciences, after which the idea of the aether was given up entirely for the "empty space" theories of relativity and quantum dynamics.

Miller's extensive work on the aether question was never rebutted when he was alive, but his research was contemptuously compared with "searching for perpetual motion". After his death, the adherents of the empty space theory breathed a heavy sigh of relief. Today, every physics textbook starts out with the falsehood that "the aether was never measured or demonstrated". It should be pointed out that the theories of relativity and quantum dynamics, plus the expanding universe and "big bang" theories, are utterly shattered by the discovery of an energy in space, and many physicists, who cling to their theories religiously, simply refuse to look at this kind of evidence. Worse, the discipline of physics has become a military-oriented growth industry, with multi-billion dollar funding for the sustenance of nuclear power and bomb-making technology, for fusion plants, particle accelerators, and "star wars" experiments. This kind of research has not brought forth any real benefits or fruits for humankind, but has become a self-perpetuating "research industry", a multi-billion dollar edifice which, like the medico-pharmaceutical industry, is threatened to the core by these discoveries of a primary, cosmic life energy. The physics community has unfortunately reacted to these new findings with the same arrogance and viciousness that characterizes the medical communities' reaction to the life energy. Einstein's followers, for example, have recently been accused, in print, of very nasty stab-in-the-back tactics of censorship and suppression. An entire new journal, *Scientific Ethics*, for a short period at least began to expose the whole stinking mess.

Of great interest for Reich's work is that Miller's dynamic aether was *more active at higher altitudes*, and *reflected by metals*. The capacity to be reflected by metals, with a more active state at higher altitudes, are basic properties of the orgone energy, as independently discovered by Reich. The orgone also satisfies many other of the basic properties and functions of an aether, being ubiquitous and mass-free, and by providing a medium for the transmission of electromagnetic excitation. However, the orgone also spontaneously pulses, superimposes, and directly participates in the creation of both matter and life. But even without using the taboo word "aether", or the more offending word "orgone", another group of physicists have detected or inferred the existence of dynamic energy currents at work in deep space. For example, the American astrophysicist Halton Arp made so many photographs of energy/matter bridges between deep-space objects, where those energy/matter bridges should not have been there, that he was actually banned from using the big American telescopes. His simple photos demolished the theories of empty space, the expanding universe, and the "big bang" with a single shutter click. So great was the hatred against his work that he ultimately had to go to Germany to continue with his research. Hannes Alfven, another famous physicist, also deeply offended his contemporaries by suggesting, like Reich, that space was filled with streaming currents of plasmatic energy. The space scientists to this day refuse to send satellite probes where he says they should, as to do so might confirm that space is energetically rich. In fact, physics of today is in a state of turmoil, and is desperately trying to explain away the newer evidence for an energy in space, to preserve the big bang theory, relativity, quantum dynamics, and the billions of dollars of research money that supports the "empty space" religion, and its institutional priesthood.

Few of the above ideas, nor the findings on sunspot-weather correlations, are given much funding or investigation today. Science journals still routinely carry the false statement that "no mechanism" has been found for solar-terrestrial correlations, just as physics textbooks carry the falsehood that "the aether had never been detected". And it is true that these relationships cannot be true, nor do they make any sense, from the standpoint of the "empty space" theories of physics. They require a medium in the atmosphere and in space, through which excitations and influences can pass, independent from thermal or pressure phenomena, a force which propagates in the atmosphere faster than air currents, and which can likewise quickly propagate influences across the depths of space. Again, Reich's orgone energy fits such a description.

Other research has been done to show that living creatures, and the physical chemistry of water, are sensitive to weather or cosmic factors in a manner than cannot be explained according to simple mechanical phenomena, such as light, temperature, humidity, or pressure. Frank Brown, of Northwestern University, spent decades demonstrating that the biological clocks of various living creatures were sensitive to lunar cycles and other cosmic forces. Nobody could refute him when he was alive, but today, after his death, his findings are widely ignored. Likewise the works of the Italian chemist, Giorgio Piccardi, who demonstrated that the physical chemistry of water was changed by magnetism, sunspots, and other cosmic phenomena. His work helped to fuel an interest in the magnetic treatment of water in Europe, leading to new methods for reducing scale deposits in household plumbing, and in industrial boilers. Magnetism, correctly applied, can alter the solubility characteristics of water, allowing dissolved substances to remain in solution at concentrations higher than normal for a given temperature. In the USA, these findings have been greeted with derision, as every physics textbook says magnetism has no effect upon water. Also, almost every chemical laboratory uses magnetic stirring devices to mix their chemical solutions, instead of the "old-fashioned" hand-operated glass stirring rods; these magnetic stirring devices would, if Piccardi is correct (and he is) alter the chemistry, precipitate quantity, and titration curves for every chemical reaction exposed to them. And so, the new findings are ignored in the USA, while abroad, new products based upon the discovery are entering the marketplace. Simple magnetic water treatment systems for the home are now common in Europe, replacing in many cases the ion-exchange water softeners, with their bags and bags of salt. In the USA, meanwhile, the water softener industry, in collusion with dogmatic academics and politicians, has managed to have laws passed in a few states to forbid the sale of magnetic water treatment devices.

Piccardi's work extends beyond the issue of simple magnetic treatment of water, however. At one point he attempted to isolate an unknown cosmic energy which was affecting his chemical experiments, in a manner similar to strong magnetism. In order to block out the unknown radiation, which was correlated to sunspots, he constructed an electromagnetic shield around his experiments, in the form of an Earth-grounded metal box enclosure. Then, in order to stabilize the temperature inside the metal box, he placed a layer of wool around the outside. To his amazement, the metal box did not extinguish the cosmic phenomena, but amplified it. He and his co-workers spent decades performing chemical experiments inside similar enclosures, which mirror the construction of Reich's orgone energy accumulator. This independent corroboration of the orgone accumulator principle by Piccardi was also confirmed, though in a less direct manner, by the biologist Brown. Brown observed that hermetically sealed *metal enclosures*, with a constant pressure, temperature, light and humidity inside, would not extinguish cosmic influences upon biological clocks, but would instead allow them to be more clearly observed, or even add an unusual dimension to their behavior. For example, inside the metal box, the metabolism of potatoes followed a cycle that correlated with lunar, solar, and galactic parameters. Potato metabolism additionally demonstrated a correlation to local weather; *not the weather today, but the weather two days into the future!* In the enclosure, the energized potato would respond to external energetic factors in the environment which were also determinants of future weather events.

The above are just a few of the kinds of evidence that exist for an energetic principle similar, or identical to the orgone energy. In many cases, these researchers had not known of Reich's work. In a few, they hated Reich's guts, and would hardly tolerate mention of his name by their students! And yet, the facts speak powerfully for a corroboration of Reich's orgone energy. It must be stated, however, that Reich's discovery on the orgone energy is far more inclusive, comprehensive, and tangible than any of the above concepts. In addition to having been quantified, photographed, and measured, the orgone can be seen, felt, and, as noted in this book, accumulated within special experimental enclosures.

An additional word must also be given regarding the response of the scientific and academic communities to these new discoveries. The reader will note that most, if not all, of the above researchers were hotly attacked, or isolated and ignored for their findings, irrespective of their credentials, reputations, or the amount of evidence they provided. This emotional reaction, of running away from or attacking disturbing new ideas, was explained by Reich as being the results of a specific emotional disorder, which he called the *emotional plague*. Virtually all scientists who have verified portions of this common natural energy principle have been assaulted by special *emotional plague characters*, who build their reputations not upon work or research, but upon political power, and the number of scalps they have taken. Gossip, slander, political tactics, the sneak attack, and even manipulation of the courts and police are standard tactics of the plague. Their secret goal, like the Grand Inquisitors of the Church, is to kill disturbing new findings, and the men and women who make them. The history of science is filled with evidence for this kind of behavior. The reader is encouraged to read Reich's discussion on the emotional plague, in *Character Analysis (3rd Ed.), People in Trouble*, and *The Murder of Christ*, as it still constitutes the major obstacle in the way of human social progress, and scientific research.

Part II: The Safe and Effective Use of Orgone Accumulating Devices

7. General Principles for Construction and Experimental Use of the Orgone Energy Accumulator

A) The interior surface of all accumulators must be composed of bare metal. Paints, varnishes or coatings on the metal will interfere with the accumulating effect, though zinc galvanizing does not.

B) The outer, exterior surface of all accumulators must be composed of an orgone-absorbing, generally organic, non-metallic substance.

C) Metals and non-metallic materials may be alternated in multiple layers within the walls of the accumulator for stronger energy accumulation. The more layers, the more powerful the accumulator, though one does not simply double the strength by doubling the layers. A three ply accumulator will have about 70% of the strength of a ten ply accumulator (one "ply" consists of a layer of metal plus a layer of non-metal). Accumulators of different sizes may also be nested one inside another, to develop an even stronger charge. Points A and B above must be strictly followed, however. In multiple ply accumulators, you can double-up the final outer organic, non-metal layer, and the innermost metal layer, for additional energy accumulating capacity.

D) A major common error made by some who reproduce Reich's orgone accumulator experiments is the use of inappropriate accumulator materials. For accumulators used on living systems, and particularly for human use, copper, aluminum, and other non-ferrous materials must be completely avoided as they yield *toxic effects*. Similarly, certain types of polyurethane foams, rigid or soft, do not have a good effect upon the living system when used in an accumulator. Any type of material impregnated with formaldehyde, or made with other highly toxic glues or resins should not be used.

Good non-metals
wool, raw cotton
acrylic, styrene plastic
celotex (soundboard)
cork sheeting
glasswool, fibreglass
rock wool
bees wax, candle wax
soil, water

Poor or toxic non-metals
wood or plywood
urethane or polyurethane
pressboard (very hard)
organic materials containing
formaldehyde, asbestos,
or other toxic chemicals

Good metals
steel or iron, sheet, screen
galvanized steel
steel wool
stainless steel
steel/tin can alloy

Poor or toxic metals
aluminum, sheet or screen
lead
copper

E) Some individuals have experimented with accumulators composed of buried metal boxes, surrounded with rich dark soil, free of pesticides and herbicides. The larger of these kinds of accumulators give the appearance of a root cellar or "burial mound". Some authors familiar with ancient archaeological sites have even speculated that the life energy principles were known and used by ancient peoples. Certain ancient mounds and structures have a layered characteristic, using clay soils or stone of high iron content, covered over with other layers of organic-rich soils or peat.

F) An exceptionally powerful accumulator can be made by using bees wax or other dielectrical materials for the outer, non-metallic layers. These materials may be quite expensive for a larger accumulator, and are also fragile. If you use a fragile or crumbly material for the outer non-metal layer, you can coat the outer surface with clear shellac. This has been tried by many people, and does not appear to interfere with the accumulation or life enhancing qualities of the energy. Never use shellac on the interior surfaces, however.

G) Experiments have demonstrated that the shape of the accumulator is a factor of lesser importance than its material composition. However, accumulators made in the shapes of cones, pyramids, or tetrahedrons have yielded occasional unexplainable life-negative effects. Unless one is testing for such effects, accumulators should be constructed in rectangular, cubical, or cylindrical shapes. These have given the best results, and are also easier to construct. One anecdote here: in 1980, the author was in Egypt, and went into the Great Pyramid of Cheops. While inside, I was struck down by a most intense choking, and could not get a breath. The feeling was relieved by emptying my canteen of water over my head and chest. Later, I heard reports of whole groups of tourists being similarly stricken, to the point that some people had fainted away, and had to be resuscitated outside. I cannot say if this is an effect of poor ventilation or not, but in my own case, I was the only one out of 8 people in a tour group to be badly affected. Given my observations of stunted and killed seedlings within conical and pyramidal accumulators, it seems possible to me that these effects are the result of a toxic accumulation or overcharge effect. More work needs to be done to clarify these factors related to shape, as well as accumulator use in energetically stagnated environments, such as deserts. See the chapter on "Warnings and Dangers" for more details.

H) Corners of accumulators do not have to be precisely constructed, nor do layers have to be air tight or precisely fitted, though one certainly wishes to have as neat and clean of construction as possible. In some cases, I have seen metal boxes loosely wrapped with layered steel wool and cotton, felt, or wool. Also, some have used tin cans, as used for food preservation, wrapped with plastic, and then placed inside another larger can which was, in turn, wrapped with more plastic. These tin-cans were nested inside each other to make reasonably effective four or five ply accumulators, for seed charging or other purposes. They do not look especially neat or "scientific", but they do function.

I) Accumulators should be kept where fresh air can circulate. The door or lid to the accumulator should also be kept partly open when not in use. Its interior may be kept fresh and sparkling by keeping an open basin of water sitting inside when not in use. Periodically wipe the interior and exterior clean with a damp cloth.

J) Larger accumulators used by humans or farm animals are best kept outdoors under a sheltered area, where rain will not fall on it. Good air circulation and sunlight will assist with the accumulation affect. The best location for accumulator research would be inside a large wooden barn in the countryside, away from all kinds of electrical transmission lines, electromagnetic devices, and nuclear facilities. This finding on the best environment for the life energy is in full agreement with more recent findings on *house ecology*, wherein a constructed habitat is critically reviewed for toxic effects upon its inhabitants. See the "Warnings and Dangers" chapter for more details.

K) The accumulator will not develop a strong charge during wet, rainy weather. On such days, the orgone charge at the Earth's surface is very low, most of it being taken up into the storm clouds overhead or at a distance. The strongest orgone charge is found in the accumulator on clear, sunny days, when the orgone charge at the Earth's surface is also quite strong.

L) Orgone accumulators used at higher altitudes tend to yield stronger charges than at lower altitudes; lower latitudes may yield stronger charges than higher latitudes; lower humidity atmospheres tend to yield stronger charges than higher humidity atmospheres. Periods with many sunspots and solar flares coincide with periods of stronger orgone charge, as compared to periods with few sunspots and flares. Alignments between the Earth, Sun and Moon, during full and new moon periods, appears to yield a stronger, more excited charge in the atmosphere, and within the accumulator.

M) If you run a controlled experiment with the accumulator, do not place any relevant instruments immediately adjacent to it. Remember that the accumulator has an energy field, and will partially influence nearby objects in a manner similar to those kept inside it. The electrical or electromagnetic fields of various instruments might also disturb or otherwise affect an accumulator, making this caution doubly important for the research scientist.

N) Do not use any household electrical appliances connected to a wall plug inside or near to the accumulator. Neither should portable computers, TV sets, or other radiating devices be used. These will disturb the energy inside. The interior metal walls also conduct electricity, and there may be a danger of electrical shock. For human-sized accumulators, use a battery powered reading lamp if light is desired, or place a strong lamp just outside the door opening. Many people use such a light to read a book while sitting inside. Radio *receivers* do not appear to have a negative effect if used in the room, but the effects of "walkman" type headsets inside the accumulator are unknown.

O) For experimental accumulators, realize that any organic or moisture-bearing materials which are placed inside will absorb the orgone charge. Do not unnecessarily store or bring items inside an accumulator.

P) For human sized accumulators, one wishes the exterior walls to be no more than 2 to 4 inches from the skin surface. When sitting inside, it is best to partially or completely disrobe, as heavy clothing will interfere with absorption of the orgone radiation. A wooden chair or bench may be used as dry lumber is a relatively poor absorber of the orgone. Metal chairs are also OK, but may be uncomfortably cold to sit on.

Q) NOTE: A too-frequent or too-long use of the accumulator may lead to symptoms of overcharge, such as pressure in the head, slight nausea, general ill-feelings or dizziness. In such a case, leave the accumulator immediately and rest in the fresh air for a moment. Such symptoms will go away in a few minutes. However, Reich warned persons with a history of overcharged biopathies to use the accumulator only with caution, and then only for shorter periods. These overcharged biopathies include: hypertension, decompensated heart diseases, brain tumors, arteriosclerosis, glaucoma, epilepsy, heavy obesity, apoplexia, skin inflammations, and

conjunctivitis.

R) The question of "how much is enough" is related to one's own energy level, and is primarily a subjective determination, different for each individual. No one ever tells you how much water to drink to quench a thirst. You simply drink until you have the feeling of "having enough". The same is true regarding the use of the accumulator. When you have the feeling of "enough", then get out. With most people, this will be sometime after they have reached the point where their own energy field is gently *luminating*, or glowing with a warm excitement at their skin surface, and after sweating has commenced. If you are unsure about these kinds of feelings, be patient, as with some people, it may require many sessions before they can really feel the energetic effects. A good rule of thumb is to limit the sitting period to no longer than about 30 to 45 minutes. It may be used more than once a day, however. One should not attempt to "nap" inside for prolonged periods. Additional information on these bioeffects is given in the chapter on "Physiological and Biomedical Effects".

S) The qualitative state of the orgone, as well as its absolute charge, are constantly varying at any given location on the Earth's surface. Weather cycles cause the accumulator to vary its charge, and toxic environmental conditions (oranur and dor, see below) may periodically or chronically contaminate the accumulator, making its use potentially unsafe. Experimental use of the accumulator therefore demands that one learn about weather cycles and other environmental factors.

8. Warnings and Dangers: The Effects of Oranur and Dor

A common problem with use of the accumulator is the difficulty in finding an energetically clean environment in which to use it. The orgone energy in the atmosphere is very sensitive to certain kinds of disturbance and agitation. Much in the manner of living protoplasm, orgone energy can be *excited* or *irritated*, and certain environmental influences can drive it towards a toxic condition. If the energetic atmosphere in your home or neighborhood has been made toxic in this manner, use of an accumulator is disadvised, or advised only with great caution, as it will be very difficult to accumulate anything except a toxic charge.

For example, **orgone accumulators, particularly those intended for biological experiments or human use, should never be used in rooms with the following orgone-irritating devices:**

fluorescent lights
television sets
computer or microcomputer
other cathode ray tube devices
microwave ovens, or eddy-current stoves
electric blankets (even if only plugged in, and off)
diathermy, x-ray machines
sparking electric motors
induction devices or coils
other electromagnetic devices
ionization-type radioactive smoke detectors
clocks, wristwatches, or other devices containing radioactive, glow-in-the-dark materials (phospholuminescent materials, which work on the principle of absorbed visible light, are OK)
other radioactive materials, or strong chemical fumes

Orgone accumulators should not be used even in the same buildings where the more powerful of the above kinds of devices (such as x-ray machines) are used, or were recently used. Experiments by Reich, and more recently by others in large German hospitals, have demonstrated that x-ray equipment will destroy the life-enhancing effects of the orgone radiation. Additionally, there is a *persistence effect*, wherein toxic energetic conditions remain for a time after the irritating devices are shut down and removed from a room or building. **Orgone accumulators should likewise not be used in the immediate neighborhoods or vicinities where the following facilities are located:**

airport radar systems
cellular telephone or microwave relay towers
very high-tension power lines
AM, FM, or TV broadcast towers
nuclear power plants, storage facilities,
  or nuclear waste dumps
military installations with nuclear bomb storage
past or present nuclear bomb testing areas

Reich and others associated with him made warnings about these devices in the 1940s and 1950s, but only today do we see epidemiological studies corroborating their life-negative effects. Part of this problem has hinged upon the difficulty that, by simply demonstrating a correlation between two events, you do not prove causality. One has to show or demonstrate just what the mechanism is, and objectively demonstrate each step between the two correlated events, before cause and effect are proven. This is in most cases a very wise policy, but it is very unevenly applied in the world of the sciences. Orthodox theorems are rarely subjected to valid critical review based upon their failures to meet this strict criteria (eg, "bad genes", "viruses", etc.), while unorthodox theories are denied funding or simply thrown out or repressed for whatever weaknesses they may have. Industrial polluters may also raise this issue to avoid taking responsibility for the environmental damage they have done.

Regarding energetic questions, according to the best calculations of the physical scientists, low-level radiation *ought not* to have a deleterious effect upon the living system. The energy present in the low-level radiation, *as detected with conventional radiation detection instruments*, is simply not sufficient to do significant damage. And yet, the damage occurs. I emphasize the concern about "conventional radiation detection instruments", because a major fallacy of physics is that if an instrument does not measure an environmental disturbance, then no disturbance has occurred. The error here lies in the false assumption that their energy detection instruments must be detecting 100% of any disturbance. This unprovable assumption is, of course, challenged by biological or epidemiological evidence which demonstrates that an effect does exist. There is furthermore a great distrust of the body in the modern sciences, in that average people who are made sick by our modern energy radiating devices often are not believed, or are viewed suspiciously.

It is precisely here that Reich's findings on the orgone energy provide a clarification, as the life energy (and disturbances within it) *cannot be detected with ordinary nuclear or electromagnetic sensing instruments.* One has to modify the instrument to pick it up, or use entirely different methods. The orgone is also an interconnecting energy continuum, which provides a connection between the offending facility or appliance (nuclear plant, microwave tower, fluorescent light, TV set) and the living creature that is affected. As the local orgone energy field of the Earth, or the energy field of a home is badly disturbed and agitated by these devices, so too does the orgone energy field of a person in that environment become disturbed.

Modern physics partly acknowledges these connections, in that all nuclear bombs, nuclear reactors, and related facilities are said to radiate unshieldable, undetectable, theoretical *neutrinos* in tremendous quantities. These neutrinos race out from the facilities, penetrating all forms of radiation shielding, and impinge upon the bodies of everyone for miles around. Theoretically, they do not do any damage, but this is purely a speculative assumption. The major observed fact is that according to the best theories of classical physics, **significant energy is constantly being lost from the heart of the nuclear reactor, through the heavy reactor shielding into the surrounding countryside, that cannot be detected with ordinary radiation detectors**. From the perspective of Reich's findings, it appears that this lost energy is not discharged as "particles", but is discharged directly back into the cosmic orgone energy continuum, which becomes greatly agitated and overcharged. Given the ability of the orgone energy to penetrate material substance, this energetic disturbance is quickly propagated outward, through the reactor shielding to affect living creatures and weather in the surrounding area.

Likewise, the dilemma of illness provoked from low-level electromagnetism. Such radiation *ought not* be making people sick, but it does. The theoretical difficulty here is that physics says these electromagnetic waves are transmitted across the countryside and around the globe *without any medium of transmission*. This position is similar to someone studying and working with sound waves, or water waves, but denying the existence of the air or water. But the nuclear and electromagnetic "particle-waves" need a medium through which they can propagate. The big myth of modern physics is that this medium was never discovered, and this fallacy was discussed in a previous chapter.

Nuclear and electromagnetic devices and facilities do have deleterious effects upon the health of their workers, and of people who live nearby, whether or not one accepts the bioenergetic point of view outlined here. In general, the health hazards are not evenly distributed among a given population. Certain very high energy, or very low energy people, and in general the very young and very old, are more sensitive to these toxic energies, and will react to them more quickly and strongly. A few cases may be given. I met a secretary who was unable to make the transition from typewriter to word-processor computer. Every time she sat in front of the computer, she developed a metallic taste in her mouth, and became nauseous. This effect would last for weeks after a single exposure to the computer. Her boss and friends thought she was crazy, but she was simply more sensitive than they were. In another case I met a fellow who lived one mile from a large radio broadcasting tower, which was irradiating the entire area with a strong dose of electromagnetism at radiowave and microwave frequencies. By the time he realized the danger of his situation, his children had contracted leukemia, and he had to move out of the area. Medical practitioners who are confronted by such symptoms usually do not make a diagnosis based upon the energetic ecology of a patient's home and work environment. In a chapter below, I give a number of practical things people can do to protect themselves, and their accumulators, from these environmental hazards.

Reich's observations on the life-negative aspects of these devices and facilities cannot be explained according to conventional theories of physics and biology, but they have been corroborated to a greater or lesser extent. The various electromagnetic and nuclear devices and materials listed above are known to irritate the orgone energy, driving it into a wild, frenzied state which Reich identified as the *oranur effect*. Oranur was inadvertently discovered after small amounts of nuclear material were introduced into a strong orgone accumulator. Reich kept several large, 20-ply accumulators inside an even larger room-sized orgone accumulator at his laboratory in rural Maine. When the radioactive material was introduced into this highly-charged environment, the orgone energy field of the entire mountain region of his laboratory was driven into a state of wild agitation, which could be readily felt and seen. Laboratory workers became quite ill, and experimental mice kept in another building died in large numbers.

Reich found that the oranur effect persisted long after the nuclear materials were removed from the laboratory accumulators, making the facilities unusable for several years. Under such persisting oranur agitation, the orgone energy eventually becomes immobilized and "dead". Reich identified this deadened energetic state as *dor*, which was short for *deadly orgone*. A room charged with dor will feel unbearably stuffy, and it will be hard to get a good breath out of it. One also feels constantly dehydrated, given the water-hungry nature of dor. Some people react to dor with edema, and a particular extreme form of dor-sickness was identified by Reich and his co-workers. The organism responds with lethargy, immobilization, and emotional contactlessness. These effects are quite tangible, sensible, and measurable.

A great deal was learned about the atmospheric life energy from this experimental accident, and a 1951 work by Reich titled *The Oranur Experiment* describes the dramatic events. Reich later identified a number of other sources of mild to severe oranur, which could disturb the orgone energy within a home, office, building, neighborhood, or region.

In most people's homes, the most common orgone irritants are the TV set, the microwave oven, and fluorescent lights of any kind (the full-spectrum varieties reduce, but do not eliminate this problem). Fluorescent lights often produce hyperactive plants, with enlarged, oversized leaves, and so fool people into thinking that the lights are "good". Some studies have even shown that depressed people can be agitated to a greater activity or metabolism by exposure to fluorescent lights. Examples are wintertime emotional depression, depressed newborn infants, and even "depressed" office workers, all of whom are agitated to a temporary increase in activity under fluorescent oranur. In many cases, the increased activity was linked to light color or frequency, and this also has its influence. But the problem of fluorescent oranur excitation is usually not addressed as a factor in those studies. Oranur is, however, produced by all types of fluorescent lights, TV sets, and microwave ovens. It can be objectively measured through the disturbed electrical potential of a house plant exposed to such devices, through use of an orgone charged geiger counter, or by making extended measurements of accumulator functions, and observing the perturbation which occurs during oranur and dor conditions.

In any given city neighborhood, radio broadcast towers,

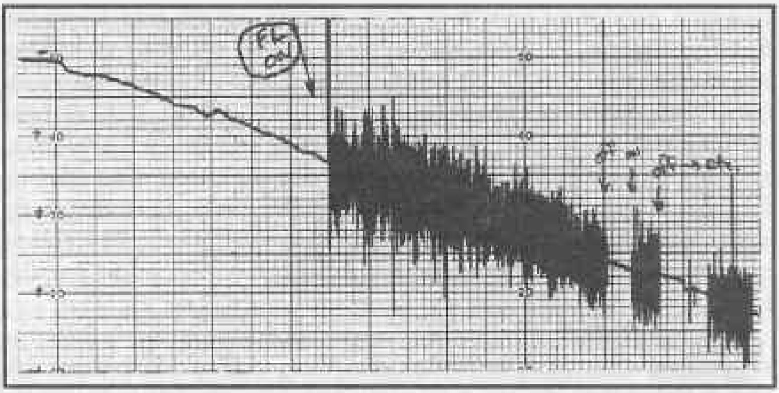

Above: Atmospheric energetic disturbance created by a fluorescent light, detected by measuring the bioelectrical field of a philodendron with a sensitive millivoltmeter (HP 412-A VTVM) both before and after the light was turned on. Below: A similar disturbance from a TV set. In both cases, no light from the apparatus reached the plant, which was shielded behind heavy cardboard.

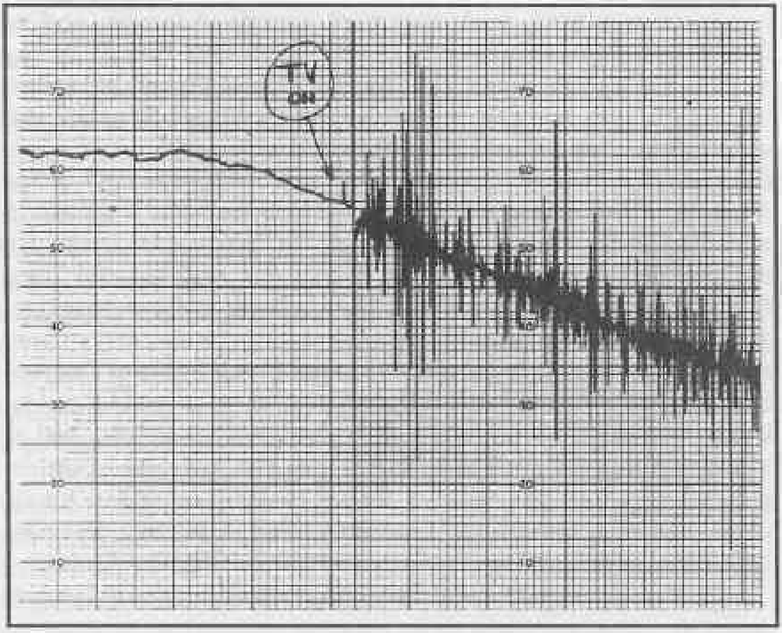

airport radars, and microwave telephone communications towers also constitute hazards, and producers of oranur. Like microwave ovens, and TV sets, they are allowed to leak relatively high levels of radiation into their local environments. Infrared automatic door opening sensors, or automatic light switches, also operate closer to the microwave end of the infrared spectrum, as do library or commercial inventory scanners, which are designed to stop petty larceny. These smaller sensors are powerful enough to trigger your automobile police radar detector at several hundred yards, and may pose a danger to workers who sit near to them day after day; the actual risk here is simply unknown. Like microwave ovens and TV sets, they are allowed to expose the "average" person with an "average" dose, which is irrationally assumed to be harmless. Until more is known about them, you should err on the side of caution. Do not site an accumulator near any of these devices.

Nuclear plants likewise are permitted to vent (or rather, dump) significant quantities of measurable radiation into the cooling water and ventilation air that passes through the facilities. Aside from the fact that the local population breathes and often drinks this waste, which can accumulate in the food chain, there is the problem of oranur and dor. Both are created by nuclear plants, and the atmospheric energy in these areas will be affected, with one qualitative state predominating over the other. Sensitive people can literally feel the difference in a region after a nuclear reactor has been operating for a period, and careful observations will sometimes reveal changes in weather patterns.

Underground atomic bomb tests are perhaps the worst offenders, as they badly shock and agitate the orgone energy field of the entire planet. There is some evidence suggesting that severe weather extremes, such as drought or severe storms, plus epidemics of "flu" symptoms, may be triggered by underground atomic bomb tests, through a widespread disturbance of the Earth's energy field. Some evidence, currently under review, has suggested that the entire Earth is disturbed in its rotational dynamics, and the upper atmosphere overheated and perturbed, by underground atomic bomb tests. These kinds of effects make no sense whatsoever from the viewpoint of classical biology and physics, which denies the existence of any vital energy principle, and assumes that space is "empty". From the standpoint of orgone biophysics, these effects do make sense.

Biophysically, the oranur effect can be felt by sensitive individuals as a jumpy, overexcited, possibly heated or slightly feverish feeling. One may feel a constant agitation; some people may mildly contract when it is present, while tempers may flare among others. Severe oranur affects each person at their weakest point, and tends to drive latent medical symptoms to the surface. The palms of the hands may exhibit a mottled characteristic, and sleep may be nearly impossible. There is a tendency towards inability to maintain a coherent focus upon work or other activities. The atmospheric expression of oranur is also one of overcharge. Skies may maintain a strong blue color, but significant haziness will appear on the horizon. Clouds do not coalesce or grow under oranur conditions, partly because the highly charged and agitated atmosphere cannot contract, and the charge within clouds cannot build beyond a certain point. Winds may be chaotic under oranur conditions, as if confused or agitated. Approaching rainstorms usually begin to fragment or dissipate as they approach an oranur-affected region. The atmosphere may have a "tense" or "strained" quality, reflecting the generally overcharged conditions. Rains will decrease, particularly as oranur is eventually replaced by deadened, immobilized dor conditions.

Dor has an atmospheric expression as well, and when sufficiently widespread, is associated with drought or desert conditions. It appears on the landscape as a steel-grey haze that reduces visibility, gives sunlight a burning or scorching quality, turns rains acidic, or completely blocks the rainfall. Clouds are rendered into a tattered state, similar to slightly dirty, shredded cotton, and never grow beyond a certain small size. Unusual small clouds with a black or dark grey coloration sometimes appear; in contrast to other small clouds in the area, these darkish clouds maintain their dull coloration even when directly illuminated by sunlight.

These are what Reich called *dor clouds.* They often form and reform continually over certain locations on the landscape, as if energetically attached to that spot.

While oranur and dor often exist together within a given region, one expression will generally predominate over the other. Being energetic phenomena, oranur and dor cannot be "blown away" by winds, though a good rainstorm may sequester and clean them out. Under exceptionally strong oranur or dor conditions, rainstorms are blocked and diverted, giving rise to prolonged drought. Desert locations are generally charged with great quantities of dor, particularly on the lower-lying portions of the topography. Regions with multiple nuclear power, refining, and waste storage facilities tend to be very highly charged with both dor and oranur. Sequential episodes of drought often occur in those areas, given that the life energy is rarely in a natural state, and is periodically overexcited, or deadened.

Contrast the above descriptions of oranur or dor conditions with that of the orgone energy in its normally sparkling and pulsing condition. When the orgone continuum maintains a state of healthy and vigorous atmospheric pulsation, regular cycles of rain-dry-rain-dry occur. The atmosphere is clean and transparent, sparkling and crisp, and does not have any notable atmospheric haze. The contrast between clouds and blue sky is apparent all the way down to the horizon. The open sky has a deep blue color, and cloud boundaries are sharp and well-defined. Clouds maintain a rounded shape, like buds of cauliflower, and build vertically without leaning sideways, or collapsing. Distant mountains maintain a bluish or purple coloration. The vegetation is also lush and crisp, full of life. Birds are active and soaring, and other animal life is also active. The sunlight warms, but does not scorch or burn as readily. The general subjective feeling during clear weather conditions is of great expansion, abounding energy, contactfullness, and liveliness. Breathing is so easy that the air literally pushes into your lungs. Most people feel exceptionally alive and alert, and more relaxed than usual. All of life is pushing upwards, against gravity, as an expression of the surging, gently pushing and expansive nature of the life energy. During rainy conditions, one may feel less energetic or sleepy, but still comfortable and at ease. Rains occur with a cyclical regularity.

Most older people are aware that this quality of atmosphere is increasingly rare. It was more commonly observed in the past than today. Hazy, stagnant dorish qualities are fast becoming the "norm", such that many young people, particularly in the big polluted cities, do not know what a really crisp and sparkling day is like. For example, older airline pilots remember when dorish haze was present only over a few industrial areas of the Northeast USA. Today, however, dorish haze can be observed in unbroken fashion, from coast to coast, and also a considerable distance out to sea as well! Likewise, with the continuing rampant deforestation and desertification that is occurring world wide, dorish desert conditions are spreading into regions that were once much more lush and wet. In wetter regions, as the atmosphere becomes dorish, regular rainstorms are replaced by foggy acidic drizzles. Naturalists report that the blue orgone glow over mountains vanishes about two years prior to the onset of massive forest death, a phenomena which is likewise associated with hazy and stagnant air pollution. Indeed, the blue orgone glow of the oceans, rivers, forests, and atmosphere is vanishing at a disturbing pace, foretelling a collapse of entire ecosystems. Just when the life energy has been documented and objectified, it is being polluted out of existence, and killed.

To describe the energetic states of oranur and dor, Reich used the example of a wild animal placed into a cage. At first, the animal reacts with fury, attempting to break free of the constraining enclosure. Later, the animal exhausts itself, and becomes inert and lethargic. This response, observed by zoo keepers around the world, has likewise been corroborated by John Ott in his work *Health and Light.* Ott demonstrated that laboratory mice exposed to agitating radiation from a TV set would at first become overexcited; later, the same mice would become inert and lethargic, eventually developing degenerative diseases. Ott gave many examples where aggressive behavior among breeding animals, such as mink and aquarium fish, were eliminated by removing oranur-producing fluorescent lights. Similar effects upon schoolchildren, from fluorescent lights in the classroom, are at work. Some school teachers have found that disruptive behavior in the classroom is often easily eliminated by simply turning off the fluorescent lights.

I have observed a similar reaction among children who are allowed to spend enormous amounts of time "watching" TV. During the early phases of a child's TV exposure, there is often little focus upon program content. The kids simply want the TV *on*, and they will often busy themselves with other things while sitting in front of it. One often sees this strange behavior among entire families, where the evening or weekend family activity revolves around the big color TV. Nobody seems to care what program is on, just so long as the TV *is* on. Like cocaine-eating laboratory mice, both children and adults may become *addicted* to the oranur effects of the TV set. Later, like Ott's mice, they may enter an inert, lethargic, or immobilized stage, what is popularly called the *couch-potato syndrome*, which may be a precursor to degenerative disease. Of course, there is an emotional component at work here, where emotionally contracted adults and children may use the TV as a means of escaping an unhappy social or family situation. But recall Reich's discovery that the orgone is the energy of the emotions. The effect of the TV is more than just a "cognitive" escape, and has distinct *bioenergetic* effects.

This bioenergetic form of electromagnetic/oranur addiction is most clearly seen when someone tries to turn the TV set off. Agitated or lethargic children, who were bathing themselves in the TV radiation, but paying scant attention to the program, may suddenly object quite vocally when attempts are made to shut it off. Even adults suffering from this syndrome will be discomforted by the thought of shutting the TV off, forcing them, as it would, out of a mildly catatonic state, into more direct emotional (bioenergetic) contact with other human beings. Of course, program content also plays a role, in that the more violent, cruel, and sexually titillating it is, the more it will touch upon the repressed sexual longing and bottled-up anger of people in our society, to further feed the syndrome. Couch potatoes, adults or children, rarely if ever watch the educational channel.

Another response that belongs in this category is the child in a video game parlor who has run out of quarters, or who is ordered to leave. Anguish and even violent outbursts may occur as the child or adult is weaned away. Ott has shown that these devices, TV and fluorescent lights in particular, are often a cause of childhood hyperactivity. Other researchers are today have observed similar behavioral disorders, reinforcing social isolation and emotional contraction among young children who are addicted to their microcomputers. Some of these computer-addicted kids are now getting *cataracts*, from constant exposure to the computer CRT, and leukemia may likewise be an associated disorder of the technology. The oranur in school computer rooms, and in video game parlors is so intense that it is readily sensible. The TV departments of large stores, which also have many fluorescent lights, also assault the casual passers-by with intense oranur.

I saw one clear case of TV radiation addiction with three hyperactive children who spent hours each day in front of the set, but paid little attention to program content. The minute they came home from school, the set *had to be on.* When the TV was finally shut down (the frustrated mother had to cut the power cord to defeat the clever kids), there was a wail of agonizing protest, and a period of even more agitated behavior. After about a week, however, the kids calmed down and began to develop new friendships and activities, and the hyperactivity disappeared completely. She got rid of the big color TV, and got a smaller black and white set. Even though the kids were later allowed to watch the black and white set as much as they wanted, they never fell back into the same trap, and the hyperactive syndrome did not occur again. In these cases, the energy system of the human had become addicted to the electromagnetic/oranur agitation, which took a clear and conscious effort to overcome.

When using the orgone accumulator in an oranur or dor environment, all of the above considerations are most important, as the accumulator will *amplify* whatever kinds of energy conditions are present in the local environment. If oranur or dor is present, an accumulator will amplify those tendencies, imparting a toxic, life-negative quality to its charge. In some cases, the effects of oranur and dor are both persistent and widespread, and cannot be affected by simply changing things in your own home. This is often the case with large polluted cities, and certainly in the regions near to nuclear power facilities. Regarding nuclear power facilities, a distance of around 30 to 50 miles is minimal for safety regarding both the biological effects of emitted low-level radiation, as well as for use of an accumulator. Similarly, if you are within a few miles of very high voltage electrical transmission wires, or large radio broadcasting towers, use of an accumulator is disadvised. Likewise, do not use an accumulator if your area has been recently subjected to a nuclear accident, and fallout is present.

One last set of considerations. Accumulators should never be used inside mobile homes or houses with aluminum skin or side paneling. Aluminum imparts a life-negative characteristic to the orgone energy, and it is advisable not to live inside such a structure even if an accumulator is not built or stored there. Mobile homes with wooden siding are safer, and no problem inherently exists with them. However, note that some mobile homes and buildings are insulated with fibreglass batting that has an aluminum backing. If this kind of insulation is widely used, it will act very much like aluminum siding, and turn the house into a large, aluminum accumulator. Also, houses with metal roofs, or new homes that use steel wall braces, instead of wooden components, will act somewhat like a large accumulator. I have lived inside such structures for short periods, and even without an accumulator, they tend to develop a very high charge inside. They can disturb a person's sleeping cycle, and lead to an amplification of the oranur effect if fluorescent lights, microwave ovens, or TV sets are used. Newer homes designed for energy efficiency often do not have adequate ventilation, which makes the energetic situations even worse. One does not wish to live inside an accumulator, and most sensitive people will be driven a bit crazy by the overcharge that spontaneously develops within such a structure.

Learn to recognize dor and oranur, such that if the feeling inside a building or accumulator is ever disturbed or uneasy, you can take necessary steps and precautions to eliminate those effects. Air the room out, and check to make sure that TV sets, fluorescent lights or other such apparatus are not in use in nearby rooms or on floors above and below. If an accumulator is already present in such a location, air it out also, and learn the steps for water-cleaning it. Follow the steps and guidelines given in the next chapter for "Cleaning Up Your Energetic Environment". The subjective feeling in an accumulator should be one of warmth, comfort and relaxation. This is where it becomes most important that you learn about your energetic environment, and to get in touch with your own body and organ sensations. It is clear that factors related to "civilized" living often have distinctly negative influences upon both the orgone energy field of the Earth, of our homes, and ourselves. Whatever can disturb or agitate the atmospheric orgone energy field of the Earth, or of your home, will likewise badly affect our own life energy. The next chapter focuses directly on what can be done to mitigate or eliminate these toxic effects.

9. Cleaning Up Your Bioenergetic Environment

The last chapter identified some potential problems regarding use of the orgone accumulator, or living in an energetically disturbed environment. The following points will help you to create a living environment in which the orgone accumulator will yield its strongest charge, with the most energetically soft and expansive characteristics. If you undertake as many of these steps as possible, it will not only protect your accumulator, but also yourself and family, whether you construct an accumulator or not.

A) The "Barn in the Woods": The best of all possible environments for locating an accumulator would be on the dry floor of a large, airy barn in the countryside. Most people will not have such a barn, but may have a covered outdoor porch, which meets these conceptual criteria. One should try to duplicate these conditions as closely as possible. The "barn in the woods" environment should ideally be located next to open fields and forest, at least 30 to 50 miles from any nuclear facility, and 5 miles from any large cross-country power transmission wires. It would also not be on the "path" between microwave transmission beams, nor within 5 miles of any large radio or TV broadcasting towers. It is best to have an open, airy structure, with good circulation and sunlight, but shielded from rains and high winds. No TVs, fluorescent lights, microwave ovens, radioactive smoke detectors, and so forth, would be present in the structure. Only a few electrical outlets, with overhead incandescent lights, should be present.

B) Plants, Fountains and Waterfalls: You can enhance the life-positive energy characteristics in a room by filling it with as many living plants as possible, and providing it with continuous, adequate ventilation. Green plants soften the effects of dor and oranur, and additionally oxygenate the air. The same is true regarding the cascading water of a fountain. Most people can feel these expansive and pleasing effects, and indoor plants, with waterfalls or fountains, are being increasingly used to improve the subjective aesthetics within both large and small structures.

C) Direct Water Cleaning: If your environment is polluted, or is of a dry, desert character, be prepared to routinely water-clean the accumulator by regularly wiping the insides and outside with a damp cloth. You can also keep an open bowl of water inside it during times of non-use, to draw off any stagnant energy.

D) House Building Materials: Books devoted to helping people find safe building materials free of toxic chemicals are now available from bookstores and libraries. These will inform you on the many new non-toxic products and building materials on the market. From a bioenergetic standpoint, a concern exists about living within, or building accumulators in, mobile homes or houses with aluminum or steel skins or siding. Aluminum siding makes a dwelling into an *aluminum accumulator*, which is known to have a toxic influence in and of itself. Any structure composed of metal walls, or even the newer homes using metal studs for interior wall partitions, can create an energy-accumulating effect. While you may want to "charge up" inside an accumulator on a periodic basis, you do not want to "live" inside one! Remember, the "barn in the woods" principle.

E) Lighting: Regarding lighting, all kinds of fluorescent lighting, even the "full spectrum" varieties, are disadvised. These all have an agitating electromagnetic ballast, and high-voltage cathodes, that excite and disturb the orgone energy continuum. None of these lighting fixtures, or the fluorescent tubes themselves, even the cathode-shielded variety, eliminate the unshieldable oranur effect. The best possible type of lighting is simple incandescent, with standard *unfrosted* bulbs, in which you can see the filament through clear glass. These bulbs closely duplicate the natural solar spectrum, and do not create oranur.

F) Cooking: Regarding your cooking, microwave ovens, and stoves that work on electromagnetic eddy-current principles, are not advised. While these ovens are certified as "safe" by the federal government, the standards used to set this criteria are quite out of date, and a collusion between oven manufacturers and the government exists. Ovens, stoves and toasters that work through electrical resistance heating are safer, but they do emit some electromagnetic disturbances in the extremely low frequency (ELF) range. Another drawback with electrical resistance cooking appliances is that they are not very efficient in their use of electrical energy. From a biological and energy efficiency standpoint, it is best to use pilotless, electric-spark, natural gas stoves and ovens for cooking.

G) Television: Regarding television, big color TVs are the worst offenders. They have three electron guns in the cathode ray tube (CRT), and operate at relatively high exciting voltages. The TV emits a wide spectrum of noxious energies, to include ELF, soft x-rays, radio frequencies and pulsed magnetic fields. They can rapidly build up the levels of oranur and dor in a room or house. As an alternative, the *projection* televisions, which do not have a cathode ray tube, appear safer from a bioenergetic viewpoint. They still have some disturbance upon the orgone energy, however, and should not be used near an accumulator. If you can't afford a projection TV set, try changing to a small black and white set, which has only one electron gun operating at a much lower voltage. Or take an even more radical step: dramatically cut back on, or stop watching TV. When I was a university professor, the first assignment I gave my students, to broaden their horizons and clean the mind of cobwebs, was to go out and get a big, chunky brick. The second assignment was to throw that brick through the front of their television set! Of course, no student ever did this, but the point is that you would be better off reading books, talking to friends and family, and more directly experiencing life, than playing couch-potato, or soaking up x-rays and propaganda in front of the tube.

H) Computers: Regarding computers, these are often worse than television sets, in that a worker sits quite close, and spends a lot more time in front of them. Like a TV set, they often employ the same CRT technology, and produce ELF, radiofrequencies, soft x-rays, and pulsed magnetic fields, and have been linked to fetal deformities and miscarriages. If you use a computer a lot, it is best to get one that does not use a CRT for display of the information. There are newer systems, mostly on the laptop portable microcomputers, that use *liquid crystal* technology. These do not foster the eye-irritation and headaches often associated with the CRT. *Plasma* screens are another technological alternative to the CRT, but they use more energy to operate than liquid crystal displays, and may therefore generate biologically disturbing radiations. Aside from the computer screen, the internal circuitry of the computer will also produce electromagnetic disturbances and oranur. For this reason, it is best to use the portable computers that run off rechargeable batteries and DC current. This kind of computer does not have a connection to the 60 cycle AC power outlet, nor does it have an internal AC transformer, both of which create significant ELF electromagnetism. Coupled with a liquid crystal display, they are probably the safest computers on the market, and additionally require a minimum amount of electrical energy to operate. If you use a CRT computer, consider to purchase a *gaussmeter* (available from Natural Energy Works) and evaluate the strength of the energy disturbance you are being exposed to. Anything more than *1 milligauss* or *1 kilovolt/meter*, in my view, is too much for a chronic long-term exposure, especially for pregnant women. There are all kinds of "protective devices" on the market for supposedly neutralizing CRT radiations, but if they do not reduce the measurable field, they probably do not offer any protection. Regardless, you should never use any kind of computer or television in or near an orgone accumulator.

I) Electric Blankets and Space Heaters: Regarding electric blankets, these have also been linked with increased miscarriages and spontaneous abortions among pregnant women. Even when off and only plugged into the wall socket, they emit a powerful ELF electrical field that can have a toxic influence. It is advised that you get rid of them, and go back to the wool blanket, down comforter, or heavy quilt. Electric blankets should never, never be used with an orgone blanket, or in an accumulator. By inference, this caution regarding electric blankets should be extended to cover portable electrical resistance space heaters.

J) Radio or TV Broadcasting Towers, and Power Line Emissions: Regarding radio or TV broadcast towers, and large overhead electrical transmission power lines, the environmental hazards are only now being documented. Do not be surprised if the local power company, or even environmental groups, have little information on these subjects. Educate yourself about the dangers, and make an assessment based upon your proximity, and what you have learned. According to my own study of the problem, 5 miles is a safe distance for both the larger, cross-country, very high voltage power lines, and for broadcast or telecommunication towers. However, in some special cases, very large electromagnetic broadcasting systems, as used by the military for global communications, may send a significant quantity of radiation through your area. In these cases, an independent assessment would have to be made.

K) Microwave Emissions from Radar and Telephone Towers: A major controversy is brewing about the biological effects of microwave radiation, which is finding increasing applications. In addition to use in household ovens, microwave frequencies are used for industrial drying and materials processing, and for weather, airport, and police radar systems. They are used for long distance and cellular telephone communications. Automatic door openers and motion detectors also use infrared frequencies bordering on the microwave portion of the spectrum. Accumulators should never be sited close to these appliances or facilities.

L) Smoke Detectors: Regarding smoke detectors, most of the cheaper ones use a bit of toxic radioactive waste as a source of ionization, as part of the operating mechanism. While they do function well for smoke detection, they should not be used in rooms with accumulators, nor where people are living or sleeping. The radioactive irritation constantly produces oranur, and can rapidly agitate the energy within a room or small apartment. As an alternative, there are several very good smoke detectors that use the *photovoltaic principle* as an alternative to ionization and radioactive waste. Photovoltaic smoke detectors meet or exceed all legal requirements and fire safety codes.

M) Nuclear Facilities: If you live in an area close to a nuclear power plant or waste storage facility, you should make a serious assessment of the danger to yourself and family. Get information on these facilities from local environmental groups. Usually, there are one or more of such groups trying to clean up or shut down nuclear facilities at any given location. These citizen's groups are the best informed regarding the health hazards of any given nuclear facility. Regardless, it is reasonable advise not to live or work in an area closer than 30, preferably 50 miles of such a facility. Accumulators should never be used close to nuclear facilities.

N) Simple Radiation Detection Instruments: Most of the instruments for detecting electromagnetic fields or ionizing nuclear radiation are quite expensive, and many of the cheaper instruments are not suitable for detecting low level radiation, as you would usually encounter in your home or neighborhood. For a very rough and quick determination, however, you can use a few of the following simple and inexpensive devices. Once you have approximated your exposure from these facilities and devices, and can precisely locate them, you might be able to get the local Health Department to make more quantitative measurements for you, to help you assess the exact nature of your exposure.

Microwaves: One way to assess stronger microwave emissions in your area is to get a microwave oven meter, and an automobile police radar detector. The meters on these devices often have a green "safe" area, and a red "unsafe" area. In fact, if your cheap microwave oven meter registers even a little bit, or if the radar detector "chirps" when used inside your home, you are likely getting an unsafe microwave exposure. If they register outdoors also, your location is obtaining a dosage of microwaves from some nearby source. The simple automobile radar detectors operate on microwave frequencies common to airport radars, automatic infrared door opening systems, and microwave telecommunications, which you will readily observe by driving around town with one. Using these simple detectors, you might be able to track down and locate local sources of emission in your neighborhood. Of course, you would not be able to make any real quantitative determination of risk by using these cheap and simple detectors. The dosages common for activating these detectors are significant, however, and should not be trivialized. At one apartment where I used to live, my police radar detector would start chirping whenever I parked my car facing a certain direction. The detector gave an indication similar to being only a hundred yards or so from an operating police radar. The activity was stronger higher up than at street level, and I later discovered that my apartment building was constructed along the path of a microwave telecommunications beam, that was being transmitted from one tower to another. The towers were at a relatively great distance from me, but the beam between them passed right overhead! The people on the upper floors were receiving a significant dose of microwaves in their apartments, and the oranur agitation in those upper floors was quite apparent. At other times, I have driven through entire towns or counties close to military facilities and airports, during which time the radar detector was constantly chirping away. No police radar was in use, and local people were being constantly bathed in the microwave energy coming from those facilities.

Electrical Fields: The electrical component of low frequency electromagnetic fields can be detected by using a cheap AM transistor radio, set to above 1600 KCS on the dial. With this setting, it will not pick up any radio broadcasts, but only background static. In fact, you can turn it up to full volume, and only hear a slight hissing sound. However, if you then hold it close to an electrical power outlet, power cord, electrical dim switch, telephone line or jack, computer, television set, or fluorescent light, you will find that the electrical disturbance will dramatically increase the level of interference and static in the radio. Your portable radio has thereby become sensitive to strong electrical fields in the low frequency range, and will give a clear audio signal when so exposed. By walking around your house and holding the radio close to devices, or even portions of your walls that you may suspect are giving off these toxic fields, you can locate safe and unsafe areas of your home. Use the cheapest possible radio, with a plastic case and no external antenna, available from Radio Shack and similar stores for under $10.

Magnetic Fields: The magnetic component of the electromagnetic field is likewise potentially toxic, and can be detected with a cheap magnetic acoustic coupler and amplifier. Radio Shack sells such a "telephone amplifier" (again, under $10) for attaching to your telephone. When not attached to a telephone and set to a high volume, the coupler is sensitive to stray magnetic fields from a variety of household sources. Static hiss will increase as the coupler is brought close to a magnetic field, and the amplifier will give an audio indication. Use this in the same manner as the transistor radio, discussed above, to map out toxic fields in your home. Do not locate an accumulator, or your bed, or your child's bed, near to any of these toxic energy sources.

Nuclear or Atomic Radiation: There are no cheap or simple methods known for detection of low level atomic (ionizing) radiation. Most Geiger counters are incapable of monitoring the weaker amounts of low level radiation that comes from computer CRTs or TV sets. Likewise, radiation from nuclear facilities is generally rendered dilute, but still dangerous, by mixing it with large volumes of air and water. Sophisticated methods for long term monitoring, or concentration of air and water samples, are required for proper measurement. Holding a simple Geiger counter in front of a TV set or computer screen, or holding one in the air near a nuclear plant, will rarely detect anything, and is generally a meaningless procedure. Likewise, the cheaper pocket dosimeters are generally made to detect fairly high levels of radiation, and will not register low level effects. Even so, I have seen physics professors hold a Geiger counter tube, made for detection of intense gamma radiation, in front of a

TV set, declaring it to be "completely safe". This, of course, is nonsense.

O) Environmental Action: From the above, it is clear that you can do a lot to eliminate toxic energetic disturbances inside your house. Dealing with such problems outside your home, in the local neighborhood, is a matter that requires concerted social action, and devotion of some of your time to working collectively with like-minded people. A lot of self education, and education of others, may be required to get even small problems resolved. Still, it is a myth to think that you cannot fight City Hall, or effectively take on a local industry or utility. Take the responsibility, and grab the bull by the horns. One thing is sure, the problem with radioactive and electromagnetic pollution, and the consequent dor and oranur associated with them, will only get worse in the foreseeable future. To get started, join a few national environmental groups, and begin networking with concerned people in your own area. Your local library should have a list of local and national environmental groups.

10. Draw Buckets, Bion Waters, and Bion Packs

Whenever we take a long soaking tub bath, or relax ourselves with a foot bath, we in part gain that feeling of relaxation because of the energy absorbing capacity of the water itself. Reich observed that water had a strong mutual affinity and attraction for the orgone energy. When we soak our body in a warm bath, our internal orgonotic charge, and bioenergetic tension, is reduced, and we relax. The effect may partly be explained by thermal warming of our bodies, but other considerations are clearly at work. By soaking in a tub of water, the energetic potential of the body will be reduced, while the energetic potential of the water will increase. We literally lose energy into the water, and relax, somewhat like a balloon that has lost a bit of air.

The energetic absorbing or *drawing* effect of water can be changed in character, into a combined *drawing and energizing* effect, through use of dissolved crystals, such as epsom salts, which increase the energy potential of the water, thereby making it a more powerful attractor and mobilizer of our own biological energy. A similar energizing and drawing effect can be brought about by soaking in a tub of water which contains 1 pound each of sea salt and baking soda. Salt and soda baths, lasting around 20 minutes, can be used to reduce tension and overcharge, or to drain off a toxic charge of energy.

The mineral baths at various natural springs where the water has observed healing properties appears to be based upon similar principles. Many resorts and health spas are built at locations where hot springs, or other waters or earth materials (muds, clays, ash) of an unusual nature exist. It is common for people to soak in these mineral waters, or in special mud or ash baths, after which they feel greatly relaxed, uncommonly energized, or even healed of chronic ailments. Medical symptoms may be relieved, temporarily at least, by partaking of these soaking baths. This tradition of the *healing waters* is more popular in Europe, given the hostility in the USA to noninvasive, natural kinds of treatment. The threat of criminal prosecution and jail often prevents clinics and health spas built on such mineral springs from openly competing with the hospitals.

Reich's work provides a basic explanation for these effects of natural waters and earth baths. Reich discovered the orgone energy, or life energy, during experiments which demonstrated that microscopic, energy-radiating vesicles could be derived from the disintegration of various organic and inorganic materials. Clay, soil, ground rock, beach sand, and iron filings were among the inorganic materials that, when allowed to disintegrate and swell in water or sterile nutrient broth solutions, would form the small radiating vesicles, which he later called *bions*. Certain sands from Scandinavian beaches were found to form bions of an exceptionally strong, bluish, radiating character. The blue bions from these preparations developed energy fields that could irradiate people and objects, and for a period, Reich experimentally used the energetic bion solutions for the treatment of various disease symptoms. Bion solutions were injected into experimental animals, where they had an immobilizing influence upon pathogenic bacteria and cancer cells. Later, *poultices* made from bions, or special *bion packs*, were made by which the energy directly liberated from the disintegrating substance could be used to irradiate the body. After the discovery of the orgone energy accumulator, which developed its charge directly from the atmosphere, Reich ceased experimental development of bion packs for such purposes. In later years, however, with the energetic and chemical poisoning of the atmosphere, and the consequent problem of contamination of the accumulator, interest in bion packs has been rekindled.

The following simple recipe for the bion pack was developed from a variety of sources. A bion pack can be made from clean beach sand or other earthen or clay materials known to have healing properties. A large handful of the earthen material is wrapped in a thick sock, or other heavy fabric container, like a sausage, about one foot in length and 1/2 foot in width. It should be tied or sewn shut so that the material cannot spill out. The bion pack is subsequently saturated and *boiled* in water, or in a pressure cooker, for about 15 minutes. After this, it is wrapped in wax paper or plastic and allowed to *freeze* solid, in the freezer. For the first use of the bion pack, the boiling and freezing process should be alternated several times. It should *not* be cooked in a microwave oven. The bion pack is used after one of the boilings, being allow to cool, with excess water drained off. The pack is then applied to the body, with additional cloth insulation in case it is too hot. As the beach sand disintegrates from the cooking and freezing, radiating microscopic blue bions will form. The radiation from such a bion pack should continue even after the pack has cooled off, and it can be rejuvenated after drying out by repeating the boiling. Orgone radiation can be obtained from this natural source, even in very dorish, polluted atmospheres, when use of an orgone blanket or accumulator would be disadvised. The effect was discovered by Reich early in his research, and both the existence and behavior of the bions has been confirmed by other scientists.

Before the modern era of drugs, health professionals used special kinds of warming, radiant clay or sand packs, or plasters, which were used to relieve aches and pains, or for healing wounds or infections. Many of these poultices or plasters were learned from native healers, who knew which muds or plants gave the best effects. The author does not know if these kinds of packs or poultices are still commercially available. There are many plastic or rubber "plasters" and "warmth packs" available from drug stores, but these are based solely upon thermal phenomena. However, the various health spas and mineral spring resorts, where people soak in special muds, clays, or ash baths, are using the principles of life energy radiation, liberated from those natural Earth substances through the principles of bionous disintegration. The people running such spas would know where to find the best materials for your bion pack. A similar bionous process may also be at work regarding the use of rock dust fertilizers, for reinvigoration of dying forests, and in the "mud facials" or "clay masques" applied to refresh the face, and tighten the sagging skin.

Another way to clean up the energetic atmosphere within a home or apartment is the use of *draw tubes* or *draw buckets.* Like the accumulator, these devices are very simple, passive instruments that function by virtue of basic energetic principles. Draw tubes are hollow metal tubes, made from 3/4" or 1" galvanized steel electrical conduit pipe, cut to around 2' lengths. The draw bucket is simply a plastic or metal pail, placed on a drain board, into which water is allowed to slowly circulate and overflow. The draw tubes are halfway inserted into the draw bucket, and are allowed to point off into various parts of the room or apartment which needs energetic cleansing.

As the water slowly circulates in the bucket, toxic forms of orgone energy are drained out of the room, and possibly from adjacent rooms as well. Dor tends to be exceptionally water-hungry, and will be removed from the room, assuming that additional quantities are not being created. Oranur will also be lessened, as the draw tubes and bucket gradually lower the energy level in the room, reducing agitation and overcharge. After the draw system has been in place and operating for awhile, you can hold your hand in front of the tubes and sometimes feel a slight tingling or "cool breeze" phenomenon. It is recommended that the tubes be positioned away from where people are resting or sleeping; they should likewise not be pointed to any part of the body longer than a few seconds. They can be set up for constant use in an office or working environment, to reduce oranur overcharge and agitation. On several occasions, I have seen such a system in use to lessen oranur in rooms with operating computer systems. In these cases, where a sink or drain is not near the areas to be treated, several lengths of hollow, flexible steel *greenfield* or *BX* cable can be used to extend the drawing effect from a sink or bathtub full of water into adjacent rooms. Point the open ends of the greenfield cable to the areas you want to clean. This kind of cable is used for electrical wiring, and can be purchased at most large hardware or electrical supply stores. Do not use aluminum greenfield cable, and do not run any wires inside it.

It is essential that the grounding water be clean, unpolluted, and circulating or moving. It must be constantly refreshed, even if only by a trickle of fresh water. A bucket, sink, or bathtub can be used. The tubes should be made of galvanized steel or stainless steel, though the former is cheaper and easier to cut. Silver is also very good, but quite expensive. The tubes must be hollow, without dust or dirt on the interior. One end of each tube must be completely immersed into the water, and a number of tubes should be used. They may have a plastic coating on the outside.

The principles of the draw tube and bucket are based upon Reich's findings that water has the capacity to strongly attract and absorb orgone energy, and that hollow metal tubes has the capacity to focus or extend the drawing influence of the water some distance outward. At one point in his research, Reich developed a device called the *medical dor-buster*, which was used experimentally on patients, to remove overcharge and dor from the body. Larger devices called *cloudbusters*, based upon this same principle, were later used to remove accumulations of dor and oranur from the atmosphere, to restore atmospheric sparkle and rainfall during oppressive, droughty conditions. Both of these devices pose considerable hazards and health risks for people who would experiment with them, and consequently they not discussed here. However, a good deal of research has gone into studying these devices. The author, for example, has engaged in research with the cloudbuster for more than a decade, and has used the apparatus to terminate a number of severe droughts; other experiments have demonstrated a clear ability to increase clouds and rain in deserts. The cloudbuster constitute as great of a breakthrough in the environmental sciences as does the accumulator for the health sciences. Space does not permit a discussion here of this remarkable technology, but the interested reader can obtain more information by contacting the author through the Orgone Biophysical Research Laboratory, listed in the Information section.

When draw tubes and buckets are allowed to work in a room for several hours or days, the room gains a softer feeling and sweeter smelling condition; stuffy or oppressive conditions generally vanish. The metal tubes amplify the natural drawing effects of the water, grounding out stale and toxic forms of the orgone energy, changing it from a life-negative to life-positive character.

11. Physiological and Biomedical Effects of the Accumulator

It will be useful to review the biological effects of the orgone accumulator, as given by various people who have actually worked with it, and know just what it can and cannot do. However, this chapter should not be considered as a definitive or comprehensive overview of Reich's findings on cancer, the biopathies, or even the biological effects of the accumulator. It is not, and only constitutes a bare-bones summary, to let the reader know what kinds of things to look for if the accumulator is ever used in a health-related context. A complete bibliography for the materials summarized below appears in the reference section.

The actual discovery of the orgone energy and accumulator was first announced by Reich in the 1942 issue (Volume 1) of the *International Journal of Sex-Economy and Orgone Research*, in a section on "The Construction of a Radiating Enclosure". That journal focused also upon the emotional aspects of the cancer biopathy, the relationship of cancer to emotional resignation, sexual starvation, and chronic energy depletion. Reich also published his findings on the spontaneous organization of cancer cells from the patient's own bionously disintegrating tissues. Additional information was later published in the *Cancer Biopathy*, the *Orgone Energy Bulletin*, and the *Orgonomic Diagnosis of Cancer Biopathy*. Reich's findings on cancer were confirmed by others, who likewise published in his journals. But he never saw the accumulator as a simple "cure" for cancer, and explicitly said so on a number of occasions. However, he did lay claim to the following discoveries:

1) Cancer is a systemic disorder, and not just a tumor.
2) The cancer biopathy starts in early life, with a major component related to early childhood trauma, and the consequent respiratory block and suppression of

emotions; later in adolescence, and in adulthood, the individual has great difficulty establishing a love life, and eventually resigns from sexual pleasure, and from joy or meaning in life.

3) The cancer patient possesses significant bioenergetic neuromuscular contraction and tension (armoring) that restricts circulation and oxygenation in certain body areas, notably the sexual organs.
4) The cancer patient suffers from a chronic loss, and gradual depletion of bioenergetic charge of body tissues.
5) Shortly before onset of tumor development, the individual experiences a powerful emotional blow, such as loss of a special loved one, which reinforces their emotional resignation.
6) The cancer cell originates from bionous processes, arising from the disintegration of the patients own energetically weakened tissues.
7) A specific *t-bacilli* is found in copious quantities in the tissues and blood of cancer patients; t-bacilli are culturable, and when inoculated into mice will cause tumors to form.
8) Use of the accumulator could not by itself reverse the deeper biopathic nature of the cancer disease; in a limited way, however, it could stimulate the bioenergy system to expand, recharge the tissues, and even disintegrate tumors.

While this last point may sound like a cure for cancer, Reich was cautious about this, though clearly optimistic. Of the case histories given in his writings, he emphasized the failures over the successes. He was constantly making careful evaluations of the patient's blood, and also developed a new bioenergetic blood test, that allowed even precancerous tendencies to be identified. He also observed that the gentle vagotonic, parasympathetic excitation the accumulator provided would often deepen the respiration of the patient, and help to bring long-buried feeling to the surface. Reich also worked with his patients characterologically, to overcome the emotional and respiratory blocking and sexual stasis associated with the cancer. The highly-charged blood would distribute new life energy from the accumulator throughout the body, into every organ and tissue, as patterns of emotional holding were likewise being relaxed, and the respiration deepened.

It was clear that the accumulator could recharge the organism, and even in a limited way help to overcome many secondary complications of the disorder. People would often regain lost organ functions and an increase in energy, for a few years, or sometimes in association with a complete remission of symptoms. But often, at least in the published accounts, a relapse would occur. In some cases, it was apparent that as the patient's tumors began to disintegrate, they would become debilitated by the toxic break-down products of the tumor, and die of secondary complications, such as kidney or liver failure. This was a particular problem when tumors deep inside the body were being broken down, and easy discharge of toxic tumor debris was not possible.

In some cases when the patient's bioenergy level was being recharged by the accumulator, they began to feel a welling-up of buried feelings, which they often did not want to deal with. In some cases, as they began to recover, they developed pains in the genital area or thighs, related to their sexual stasis. Reich found that almost all of his cancer patients had not had sexual intercourse for years, were trapped in a loveless, compulsive marriage, or were deeply religious. In such cases, overcoming the obstacle of sexual stasis and blocked emotion, and restoring their desire to live, was the key to a remission. In a few cases, as these emotional problems surfaced, his patients would discontinue further treatment with the accumulator, even though significant tumor reduction and restoration of body function had occurred.

For these reasons, and also to emphasize his interest in cancer *prevention*, Reich focused upon the central role of emotional and sexual resignation in the life histories of cancer patients. Where this resignation from life and feeling could be overcome, Reich observed that the prognosis was better than where the resignation was left untouched. This factor appears to explain the common observation that cancer patients who become *emotionally mobilized*, who learn to express their sadness, rage, and terror, and who regain a desire to live, will have a better prognosis.

Given Reich's findings on the emotional component of cancer, the following question has to be asked: What effect upon emotional and sexual resignation occurs when radical cancer surgery deforms or incapacitates the sexual organs, or other body areas? Or similarly, what happens emotionally when the body is so badly assaulted by caustic chemicals and radiation, such that visible, frightening deformity occurs, and normal body functions, such as eating, defecation, or sexual arousal, are no longer possible? Such horrific treatments of degenerative illness surely can only *increase* emotional resignation and sexual stasis. Having done so, they cannot help but to also *increase* the rate of degeneration, and likewise *increase* the rate of relapse and metastases. In this context, it is no wonder that the mutilating surgeries and toxic chemical treatments advocated by cancer specialists today do not have any greater benefit to patients than the treatments of 30 or even 50 years ago!

Of course, the better known unorthodox treatments, which are banned in the USA (but not in Europe, Canada, or Mexico) can do much better than this. They commonly offer the patient natural foods and herbal remedies that energize and detoxify, in a manner similar to the bionous baths and bion packs discussed above. Reich, unfortunately, had his hands full with the discovery of the life energy, and other matters, and spent little time focused upon methods for detoxification. In *The Cancer Biopathy*, he did demonstrate, with use of a special fluorophotometer, that honey had about *eight times* the orgone charge of refined sugar, and also that unpasteurized milk carried *twice* the charge of pasteurized milk. The implication here is that natural foods are highly charged with life energy, as compared with synthetic, devitalized, and refined food products. The treatments developed by Gerson, Hoxey, Livingstone, and others, appear to have independently discovered these kinds of nutritional differences through empirical means, and are clearly more advanced than Reich on the effects of diet and detoxification. (A list of clinics and research centers that specialize in these alternative treatments is given in the "Information" section.) These practitioners likewise employ special herbal or nutritional treatments that appear to have a significant bioenergetic component.

Without detracting anything from these alternative treatment methods, Reich's findings do more clearly provide a scientific basis for the origins of the cancer biopathy and cancer cell. His discussions on the emotional roots of cancer have been independently confirmed, and should assist in providing effective emotional invigoration of cancer patients. Reich's findings are also compatible with the various theorems on the causation of cancer from inadequate nutrition or environmental toxins, through the question of *energy level*. The measurable energy level of an individual appears functionally identical to the classical concept of *immunity*, or *disease resistance*, and is a key to understanding why one person gets sick, while another does not, under similar toxic environmental or dietary influences. Social and emotional factors, as well inherited factors, work a powerful influence upon the energy level, or charge of the tissues. Likewise, the discovery of viral/bacterial pleomorphism (capability of microbes to change form: viruses into bacteria, and vice versa), the independent observations of the t-bacilli, and the rediscovery of the bions by various biogenesis researchers, all confirm Reich's positions on the bionous, self-generated nature of the cancer cell. The following fact cannot be emphasized enough: **the causation, process of development, and reasonably effective, non-toxic therapies for cancer have existed for over 30 years.** The obstacle has not been a failure of science, but of the corrupting influence of politics, and the abuse of the courts and police by the orthodox medical establishment.

In spite of many difficulties, a lot of very clear and positive evidence has been gathered on the effectiveness of the accumulator for treatment of a variety of symptoms and disorders. Very effective pain relief for, and subsequent rapid healing of severe burns has been reported. Likewise, a great reduction in pain was reported when the accumulator was used by cancer patients with tumors, and by persons suffering from arthritis. Besides Reich, other doctors associated with his research effort published case studies on the treatment of cancer with the accumulator. These published accounts demonstrated a significant and promising therapy for the disease. Complete remissions were rare, but people always experienced a reduction in pain and other symptoms, with a lengthening of life by at least several months to years beyond the conventional prognosis. Other medical problems were experimentally approached, such as diabetes, arthritis, tuberculosis, rheumatic fever, anemia, abscesses, ulcers, and ichthyosis. In these cases, benefits from orgone radiation and therapy were suggested. Reich also wrote about the promising application of the therapy to leukemia. Additional benefits in the form of an immunity to flu and colds, the elimination of skin problems, and general increased vigor and energy level were also discussed in the pages of his research journals.

To my knowledge, no clinical studies on treatment of human disease with the accumulator have taken place in the USA since Reich's death in prison. Only animals studies have taken place, mainly the effects of the accumulator on cancer mice, and on wound healing in mice. These laboratory trials with mice do confirm the wound-healing and anti-cancer effects of the accumulator. More recent human clinical trials from hospitals in Germany have taken place, and prescription of the accumulator is a standard procedure for some German doctors. Several of the German physicians I met told me that *the somatic effects of the orgone energy accumulator were more powerful in the treatment of cancer than any other form of conventional or natural therapy they had tried.* They reported to me the following effects of the accumulator on cancer patients:

1. Pain was relieved, the appetite was stimulated, and the patients became more alert and active, often leaving the hospital bed, or the hospital itself, to resume activities that inter-

ested them.

2. The blood picture cleaned up, with red cells showing a stronger energetic charge, and fewer t-bacilli.

3. Tumors ceased growing, and in some cases, declined dramatically in size.

4. While patients often gave the outward appearance of a "cure", accumulator treatment alone could not touch the emotional aspect of the biopathy, which would continue to deplete the patient in a manner that could not be compensated for. While the accumulator would usually extend the life of the patient by months or even years, the patient would generally experience a relapse at some point, with a sudden recurrence of all symptoms, and a quick, less painful death.

5. The German doctors also stated that many cancer patients were presented to them that did not possess the character traits of the fully-blown cancer biopathy, as described by Reich in the 1940s. In particular, many younger people and children came to them with tumors and a very poor blood picture, and with evidence of a greatly depleted energy level; but they did not have the complete sexual stasis or emotional resignation typical of the disorder among older persons. They attributed this to the prior exposure of the patient to environmental toxins and pollutants, and the increasingly devitalized nature of common foods. These observations suggested that, under conditions of environmental and dietary stress, energetically weak individuals are prone to tissue disintegration and tumor formation, while energetically strong individuals are not. In such cases, accumulator treatment gave excellent results, with a much better prognosis for long-term recovery.

Other German scholars have evaluated the somatic effects of the orgone accumulator. One double-blind, controlled study, published as a dissertation through the University of Marburg, is titled the *Psycho-Physiological Effects of the Reich Orgone Energy Accumulator*. An extended abstract of that important study appears as an Appendix in this *Handbook*. Based upon these kinds of published findings, we may once again summarize the biological effects of a strong orgone charge:

- A) General vagotonic, expansive effect on the entire system.
- B) Sensations of tingling and warmth at skin surface.
- C) Increased core and skin temperature, flushing.
- D) Moderation of blood pressure and pulse rate.
- E) Increased peristalsis, deeper respiration.
- F) Increased germination, budding, flowering and fruiting of plants.
- G) Increased rates of tissue growth and repair, as determined through animal studies and human clinical trials.
- H) Increased field strength, charge, integrity of tissues and immunity.
- I) Greater energy level, activity, and liveliness.

Given these facts, it is not surprising that the accumulator might stimulate the recession of any symptom which is related to low energy charge in the blood or tissues, or to chronic overstimulation of the sympathetic nervous system. However, some medical problems are the result of chronic overcharge, and in those cases, use of an accumulator is disadvised, or advised only with caution, as previously mentioned. Again, Reich warned persons with a history of hypertension, decompensated heart diseases, brain tumors, arteriosclerosis, glaucoma, epilepsy, heavy obesity, apoplexia, skin inflammations or conjunctivitis not to use the accumulator, or to do so only with great caution and for shorter periods, due to the dangers of overcharge in those cases. Not all people suffer from a lack of energy, or even from "low energy". Quite often, people suffer more from a clamping down, or holding back of the emotional energy which they already have. In some cases, additional energy from an accumulator may simply give a person more energy to clamp down with. One should recognize this fact, and understand that regular accumulator use is not mandated for everyone, nor is it a panacea.

12. Personal Observations with the Orgone Accumulator

In the early 1970s, I met a young woman who had treated her ovarian cyst with an accumulator. Her doctor had urged surgery, but she did not have insurance or much money, and decided instead to try the accumulator. The woman had used the accumulator, a three-ply unit big enough to sit in, for about 45 minutes a day for two or three weeks. Around the middle of the third week, she had a vaginal discharge of blackish blood, which was the disintegrating tumor discharging into the uterine cavity. The woman felt completely healthy throughout the entire process, except for some discomfort during the time of the discharge. Some time after this, she went back to the doctor, who could not find a trace of the tumor. When told of the form of treatment, the doctor was derisive and uninterested.

Around this same time, I constructed a small but powerful accumulator, at a time when I was living only 8 miles away from the two Turkey Point nuclear power stations, in South Florida. I had been advised not to build accumulators that close to a nuclear plant, and had read Reich's account about oranur. Still, I remember thinking to myself, "It's just a small accumulator, and can't do much harm". The accumulator was left in a garage, along with a number of large metal appliances and objects, such as a clothes washer and dryer, refrigerator and filing cabinets. Within a week after doing so, the entire garage became so highly charged that it was impossible to stay in it for long. The sensible agitation and overcharge, which was provoked and amplified by the nuclear power plants, began to spread into the house, and the entire area often felt as if it were subtly resonating or vibrating. I still recall quite distinctly this phenomenon, which was most apparent at night, when winds ceased, and city noise was quiet. Meanwhile, plants inside the house began to die, and the white blood count of family members began to increase. A small Geiger counter began to yield erratic and racing counts for "background" radiation. In a bit of a panic, I dismantled the small accumulator, and removed other metal from the garage. A small draw-bucket was placed there, and the disturbance gradually quieted down. Still, the nuclear power plants were a constant worry, and we moved out of the area.

A few years later, I built another very powerful ten-ply accumulator, with shooter funnel, as described in the following chapters. One day when I was working outside, barefoot, I accidentally stepped on a hot soldering iron that had been carelessly left on the ground. My flesh was badly seared, and I was in a great deal of pain. However, the new accumulator and shooter were fortunately nearby, so I placed the burned foot into the shooter funnel. Within *seconds* the pain receded, and in a few minutes there was no pain at all! Without further discomfort, I could clean the severe burn, which had taken away all the layers of skin. The wound healed very rapidly after this, and I subsequently learned that pain relief from burns, and rapid healing of new skin, was one of the most powerful effects of the accumulator.

After constructing an accumulator that was large enough to sit in, I was able to confirm a number of subjective and objective measures that were first observed by Reich. It did indeed make one feel more invigorated and warmer, with a flushed skin. I no longer contracted colds or flu like before. I have never been sick in any major way, and so have no major "healing" of myself to report. Eventually I ceased sitting in the accumulator on a regular basis, as I just did not feel the need for it. More often, I use the orgone energy blanket. It is easier to store (on a hanger in an airy place) and can be retrieved for use very quickly. The most amazing effect of the blanket, I found, was its ability to stop a head cold, or at least to prevent it from developing into a chest cold. Prior to discovering the accumulator and blanket, all my colds or flu would spread from head to throat to chest. Since using the blanket, I rarely contract a head cold, and when I do, it can be prevented from spreading by simply resting with the blanket over my chest and throat. Over the years, I have also had a variety of small cuts and bruises, or toes cracked from smashing them into table legs (I still go barefoot a lot), all of which were treated with the shooter or blanket, with great pain relief and healing benefits.

Only on one occasion did the accumulator fail to help me with a health problem. I was bitten on the leg by a poisonous *brown recluse* "fiddleback" spider, the toxin from which killed a piece of skin on my calf about 3 inches in diameter. I did not know about the dangers of that kind of spider, and only began treating the bite after the skin had turned purple and became numb. The wound was treated several times per day with the shooter, while sitting inside the large accumulator. These treatments did not restore feeling or normal color, and the entire depth of killed skin eventually turned black and hard, falling out of my leg, leaving me with a gaping open wound for several weeks. A secondary blood infection was treated with antibiotics, and I was on crutches for weeks. The wound healed over, however, and the leg functions today without a problem. Only a small scar exists to mark the bite. A survey of medical literature on this kind of spider bite indicates that, short of questionable cortisone shots into the bite shortly after it occurs, there is no known remedy.

On several occasions, friends of mine who knew about my accumulators would ask if they or their friends could use them. In once such case, a 19 year old female had a disc-shaped encapsulated benign tumor of the breast, measuring about 1" in diameter. The tumor first developed after she became pregnant out of wedlock several years before. Her parents had badly mistreated her for this, and called her all sorts of names. The pregnancy was terminated, but the emotional abuse she had gone through led to a powerful bioenergetic contraction, and to the development of the tumor. She understandably did not tell her parents about the tumor, and had avoided doctors, being afraid of losing her breast. She had been treating the tumor with a vegetarian diet for several years, and it had not grown, nor gotten any smaller. After we discussed the matter, she began the orgone accumulator treatment by sitting inside for around 45 minutes a day, with a large shooter funnel over the breast. After three treatments, the tumor began to break apart, and disintegrate into smaller pieces. She became anxious at this point, however, and was openly agitated and upset about the accumulator, refusing to sit in it any more. Upset feelings related to the treatment she had received during her past pregnancy began to surface. She was also a student of the biological sciences, and, while she had a feeling of desperation about her situation, she had maintained a jocular surface attitude, saying she would try the accumulator only to "humor" her concerned friends. The fact that the accumulator actually appeared to *work*, when nothing else had, was a an intellectual confusion, and it simply became too much for her. She never sought additional treatment with the accumulator, but friends informed me shortly thereafter that the tumor had almost completely vanished. Here, it is important to point out Reich's observations that, in spite of the emotional components of the underlying cancer biopathy (which clearly emerged in the above case), certain kinds of superficial tumors, such as breast or skin cancer, could be effectively treated with orgone energy.

In another case, a 23 year old woman had been under conventional medical treatment for severe genital herpes for several years, but without any relief from the persisting genital lesions. She sat in the accumulator *once*, using a tube-type vaginal shooter wand. Within days of this, her lesions began to dry up and heal, leaving her symptom-free for the first time in years. She remained free of symptoms for at least several years thereafter.

I know of several cases where the orgone blanket was used for treatment, instead of a large accumulator. An elderly woman was given an orgone blanket to see if it would help her arthritis. She used it and found that it did provide relief from the discomfort and pain, and she regained a bit of movement in the affected areas. After this, she unfortunately used it with her electric blanket, after which all the arthritis symptoms intensified, back to their original condition. With great disappointment, she refused to have anything more to do with the orgone blanket.

In another case, a young woman treated her baby, which had a persisting slight fever and cold. She simply placed the child on top of the blanket in the crib, and left it there for around 15 or 20 minutes. When she returned, the child had a temperature of around 102°. She quickly removed the orgone blanket from the crib, and walked the child about for awhile. Its temperature soon dropped back to normal, but the cold symptoms also had vanished. Reich noted that orgone irradiation will increase a fever somewhat, even in adults, speeding the process of healing. Small children being treated for any kind of illness with a blanket or accumulator should obviously be watched closely. Also, no small child will feel comfortable being put inside a large accumulator all by themselves; but if mother will go with them, and make a game of it, they can sit on her lap, and this will be just as effective.

In another case, an elderly man with fibrosis of the lung, related to a life-time of smoking and emotional holding in the chest, was predicted to die within a few weeks. He was on oxygen, and could not speak more than a few words at a time or walk very far, given his inability to get a good breath. He began to use an orgone blanket and large box-type accumulator. Within a few weeks, he was up and about, rowing his small fishing boat. He reported that the only time he could get a good breath was when he was inside the accumulator, or when the blanket was on his chest. Many of his symptoms were relieved from the orgone therapy, and he remained active for many months thereafter. However, his condition worsened after he was put on an experimental medication by the doctors (prednazone). He died shortly thereafter. Again, no miracles were observed given his original terminal condition, but a good deal of comfort and relief, and an additional 6 months of life.

I once corresponded with a farmer who had a cow with a large gash in its side that had gotten badly infected and festered, refusing to heal. The veterinarians had tried all sorts of different treatments, but nothing seemed to help, and the poor beast was on the decline. Having tried everything else, the farmer made a four-ply orgone blanket, and secured it to the festering side of the cow with heavy duct tape. He left the blanket taped to the cow, not expecting to see any cure, and anticipated a sorry death for the animal. However, within a few days, the blanket had fallen off, revealing a large scab over the sore. He treated the cow a few more times with a new blanket, and says that today you can hardly find a scar on the lively beast.

Another farmer I met was diagnosed as having a fast-spreading form of liver cancer. The doctor told him to get his affairs in order, as he would be dead in 6 months. The farmer made an accumulator out of two steel oil drums, by removing both the tops and bottoms of the oil drums, sand-blasting the insides down to bare metal, and welding the two cylinders together, top to bottom. He then wrapped layers of steel wool and fibreglass around and around the steel tube he had constructed. With this tube accumulator laying on its side, he would go inside it, and take a nap from time to time. "Dr. DeMeo", he told me, "I object to your caution about not staying inside the accumulator for more than 30 or 45 minutes. I've stayed inside my accumulator for 7 hours at a stretch without problems, when I fell asleep inside it!" Well, I did not know what to make of this fellow, as when I met him, he was very weak and slow-moving. He seemed so low in energy that, in his case, the danger of overcharge did not exist. Still, he had lived for around a year beyond the terminal diagnosis of his doctor. About a year after my meeting with the man, I got a letter from him, saying he wanted to attend one of my workshops. When I finally met him again, I was absolutely amazed at his condition. He was about 40 pounds heavier, his face was ruddy and tanned, and he was literally bursting with energy. Sometimes, however, he would appear quite red in the face, as if he would explode, and once he started talking, you could not get him to shut up. Characterologically, he had gone from a situation of undercharge to overcharge. I pointed out this danger to him, and he did reduce his accumulator treatments. Anyhow, the story does not end here. It seems that he went back to his doctor, who saw his changed condition and could not find a trace of the liver cancer. The doctor got real mad at him, and accused him of going to some big city hospital for a "wonder drug". He told his doctor about the accumulator, but the doctor didn't believe him. Since this was in a small town in the Midwest, the fact that the farmer had survived the death sentence of the town's most reputable doctor, and had even thrived in spite of that death sentence, was the cause of considerable interest and discussion. Presently, I've been told, there's a shortage of steel oil drums, fibreglass, and steel wool in that town, as the man's friends and neighbors are very busy building their own accumulators!

13. Some Simple and Not-So-Simple Experiments with the Orgone Accumulator

After you construct one or more of the accumulator designs given in this *Handbook*, you can run a few simple experiments to confirm the effects for yourself. Make sure you monitor environmental conditions during the experiments, as per the factors listed above. Consult the various references given in this book for more information.

A) Confirmation of Subjective Sensations: If you are the type of person who works with your hands, who is generally relaxed, with a deep, full respiration, then you will most likely be able to confirm the following effects. Place your open relaxed hand inside the open end of an orgone accumulator, about an inch from the metal walls. You should feel a warm, penetrating radiant sensation, or a slight tingling sensation. The effect can also be confirmed by use of the funnel-type metal *orgone shooter*, which can move the orgone charge out of an attached accumulator in a directional manner, or the *shooter wand*, which is a pyrex glass test tube filled with steel wool, and charged up inside an accumulator. Holding these shooters close to your hand, upper lip, solar plexus, or other sensitive body area, will generally yield discernible sensations. Make sure you try this on clear, sunny days when the orgone charge at the Earth's surface is strong. On wet, rainy days, the effect will be minimal or absent. People with a shallow respiration, who work more with their brains than their hands, or those who carry a greater amount of emotional tension, will require more time and effort to confirm these sensations. One general rule of thumb: If you cannot feel the life-negative disturbances coming from TV sets, CRT displays on a computer, or fluorescent lights, it is not likely that you will be able to feel these subtle orgonotic effects.

B) Observations in Darkened Rooms: Many people remember from childhood the ability to see various foggy shapes or "dancing dot" luminescent phenomena in darkened rooms. Reich proved that these subjective phenomena were real, and not imaginary, nor located only "in the eye". To reproduce these observations, you must be able to discern energetic phenomena from the debris or "floaters" in or on the surface of your eye. Reich identified a *fog-like* form of the energy, and a *pointed* or *dot-like* form, which was a more highly-excited expression. Reports from the 1700s to the modern day have been made by sensitive people who could see radiant energy fields around living creatures and other objects in darkness or semi-darkness. Energy fields around magnets, or weakly-charged electrical wires, have also been observed in darkrooms by sensitive people. These effects are intensified by the presence of a strong orgone charge, as is the case when accumulators are present. Energetic phenomena inside accumulators is also directly observable. For proper viewing, allow your eyes to adjust to the full darkness for 30 minutes or so. To provide a scientific basis to these observations, the reader is guided to Reich's original accounts in *The Cancer Biopathy*.

C) Observations in the Daytime Sky: A dancing dot or *orgone unit* phenomenon is also observable in the daytime sky. This is best seen with a homogeneous background of solid cloud cover or solid blue sky. Trees often appear to be flaming this energy into the sky, or attracting it towards themselves, much in the manner of a Van Gogh painting. One must be relaxed when making these observations; one can also "soften" the eye focus, purposefully looking into the open space in between yourself and the distant infinity. Observing the sky through an open, hollow metal, plastic, or cardboard tube helps to facilitate these observations. The phenomenon is most apparent against plastic window panels and skylights, and especially when looking out of the plexiglass windows of a high flying jet aircraft. Remember that some of these phenomenon will be occurring within the eyeball, though most are not. Again, Reich's accounts of these subjective phenomena are most telling.

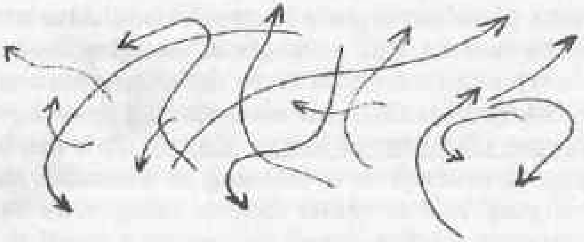

Observable luminating orgone units pulse and randomly move through the sky, with lifetimes of about one second.

D) Garden Plant Growth Enhancement Experiments: The life-positive effects of the accumulator may be observed in its charging up of seedlings, with subsequent increases in growth when they are planted. Take your garden seedlings and divide each type into two separate groups, labeled *A* and *B*. Place the seeds in group *A* inside an orgone accumulator for a day or two, up to a week, just before planting. Store the seeds in group *B* in a location away from the accumulator, but with similar temperature, moisture, and light conditions. You can keep the seeds in their plastic or paper packages while this is done, but make sure that neither group is near a TV, fluorescent light, microwave oven, computer, or other oranur-producing device. After charging, plant the various seeds in a way that you can identify the two groups. Monitor and measure the growth in both groups, taking notes and photographs. Count or otherwise measure the yields from each group. The accumulator group should have a greater growth and higher yield. Controlled studies by organic farmers, particularly those by Jutta Espanca of Portugal, have demonstrated very significant orgone charging effects. Espanca has found that garden seed charging works best if done for only a day, or even a few hours; but this must be done only on a very crisp, clear, and sparkling day, when the orgone charge at the Earth's surface, and in the accumulator, is quite strong and lively. Otherwise, charging may have to take place for slightly longer periods. Also note that seeds can be overcharged; attempts at charging seeds for 30 days or more often result in little difference between controls and charged seedlings, or even stunted growth.

E) Orgone Charging of Potted Plants: This can be done by charging up seeds prior to planting, as discussed above, or by charging up soil or water before using. One can also make a "wrap around" accumulator, using a metal can with the ends removed, and plastic and steel wool layers wrapped around the outside. Be sure that the final outside layer of plastic is fairly thick, and do not use aluminum materials. Leave the steel wool in a fluffy condition; don't compress it.

F) Home Seed Sprouting Experiments: The life-positive effects of the accumulator may also be observed in the way it enhances seed sprouting. Build an accumulator to house your seed sprouting apparatus. Store one sprout container in a darkened area away from the orgone accumulator, and another inside the darkened orgone accumulator. Make sure the temperature, ventilation, and light exposure of the two groups is identical and, once again, keep both groups away from oranur-producing devices. Measure the quantity of seeds going into each container, and make sure the quantity of water in each is about the same. Observe and record any subsequent differences in growth and taste. The accumulator group should have a greater growth and yield.

G) Laboratory Seed Sprouting Experiments: Obtain two small shallow glass dishes with a flat bottom, or two shallow glass laboratory culture dishes, about 4" diameter with a 1" lip. Into each dish place around 20 or 30 dry mung beans, to form a single layer of beans on the bottom of the dish. Add a measured quantity of water to each dish, which only halfway covers up the beans. The tops of the beans should remain exposed to the air, while the bottoms are wet. Place one dish of beans into a small but strong orgone accumulator, and the other dish into a control wood or cardboard enclosure of similar dimensions, but with no metals. Cover both the accumulator and control enclosure with a layer of black plastic, to seal them against any intruding light. Place the enclosures in a well-ventilated area, of equal temperature, and out of the direct sunlight. The enclosures should be in nearly identical environments with respect to light and temperature, but not placed closer to each other than about a meter. Again, no oranur-producing devices should be nearby. Each day, open the enclosures up and add as much water as is needed to keep the bean dishes wet to approximately the same height as originally was necessary to cover the beans halfway. If one dish begins to grow more quickly it will require more water, and this should be provided on demand. After one of the dishes of sprouts has reached about 4" high, record your observations of germination rate, length or weight of sprouts, general appearance, and other characteristics. Contrast the two groups. The accumulator group should have a greater growth and germination rate.

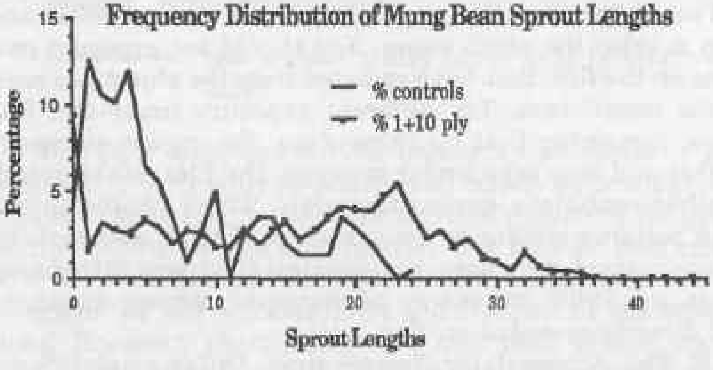

Contrast of mung bean sprout lengths between a group sprouted inside an orgone accumulator, and a group sprouted inside a control enclosure. (from DeMeo, J.: "Seed Sprouting Inside the Orgone Accumulator", *J. Orgonomy,* 12:253-258, 1978)

H) Passive Life Energy Photographs: Orgone energy field photographs can be made in a manner similar to that used to make *Kirlian* electrophotos, except that no electricity is needed. The techniques given here were developed by Thelma Moss, who pioneered the techniques of energy field electrophotography. From Dr. Moss' descriptions, you will need either a darkroom and photo lab, or access to a camera shop which can develop individual sheets of photo print paper. First, obtain some light-sensitive color or black and white print paper, making sure you do not open the package except in complete darkness. Obtain also a separate, empty, light-tight film box in which you can run your experiments. While in the dark, remove a piece of film paper and place it on the bottom of the empty film box. On top of the sheet of film, place a flower bud, leaf, piece of fruit, sprouting seedling, magnet, or other energized, living, or once-living object. Close the box and wrap it completely inside an orgone energy blanket, or place it inside a strong orgone accumulator. Make sure the film box is light-tight; cover it and the blanket with black plastic, or otherwise insure that no outside light will reach the film. Wait a day or two, or up to a week, and then develop the photo paper. You should see exposure patterns on the film that have radiated from the object you used in the experiment. Try different exposure times and film types; remember that on rainy days, the orgone charge is weaker and may take longer to expose the film. Make note of weather conditions during the trials. These photos should yield patterns similar to those seen on Kirlian electrophotographs, except that here, no electrical field was introduced. These are 100% *life energy photographs,* orgone-enhanced and directly recorded on film.

I) The Accumulator Temperature Differential Effect: Reich demonstrated that the warm glow felt inside the accumulator possessed an objective aspect which could be measured with a sensitive thermometer. An air-tight orgone accumulator will spontaneously warm up the air inside itself by a few tenths of a degree, up to several degrees. This temperature increase will make the accumulator interior slightly warmer than the surrounding air temperature, or

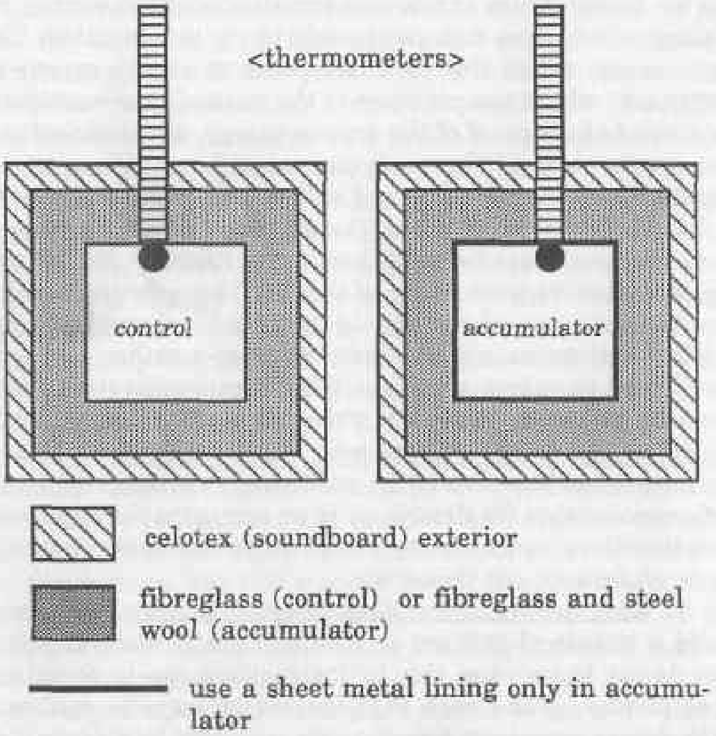
<thermometers>
control
accumulator
celotex (soundboard) exterior
fibreglass (control) or fibreglass and steel wool (accumulator)
use a sheet metal lining only in accumulator

The To-T Experiment: The orgone accumulator (right) and control (left) must be matched as closely as possible with respect to size, ventilation, and thermal dynamics for quantitative orgone energy experiments. The interior and exterior of the control enclosure should be the same size and dimensions as the accumulator, with identical lid dimensions. However, sheet metal and steel wool should not be used in the control enclosure. While metals are used in construction of the accumulator, they are strictly excluded from construction of the control. In the control, additional plastic or fibreglass layering is used in the place of steel wool. The thermal resistance *and* heat capacity *must be as closely matched as possible. A similar set of environmental conditions must also be maintained for the two enclosures.*

the air temperature inside a thermally-balanced control enclosure which does not use metals in its construction. This experiment, called the To-T (temperature in the orgone accumulator minus temperature of the control) was considered by Reich to be a proof of the orgone energy, and a violation of the second law of thermodynamics. Albert Einstein once replicated the experiment and called it a "bomb in physics"; a fascinating booklet titled *The Einstein Affair* documents the correspondence between Reich and Einstein on the matter. A definitive evaluation of the To-T experiment requires the construction of thermally balanced accumulator and control enclosures, careful monitoring of weather and environmental temperatures, sensitive thermometers capable of recording down to tenths of a degree, and prolonged systematic measurements. Those who wish to reproduce this experiment should consult the published reports given in the reference section for details. It is an area ripe for innovative investigation, and I strongly encourage the experimentalist to carefully seek out this effect.

J) The Accumulator Electrostatic Effects: Obtain or build a simple aluminum or gold leaf static electroscope. If you do not know what this is, instructions can be found in a good library. Make sure that the electroscope is calibrated with degree markings, from 0 to 90, such that its degree of deflection can be accurately measured. By running a plastic rod or comb through your dry hair, you can gather a significant static electrical charge and transfer it to the electroscope. Using a stopwatch, or a watch with a second hand, determine how long it takes for the electroscope to slowly lose its charge into the air, through a predetermined deflection angle. For example, you want to know how long it will take for the electroscope to discharge from a 50 to a 30 degree angle. You therefore should charge the electroscope up to a deflection angle *greater* than 50 degrees, waiting until it discharges to the 50 degree mark. Once this happens, you can count the number of elapsed seconds that pass until it reaches 30 degrees. The time elapsed is the *electroscopical discharge rate*. On sunny days, the discharge rate will be quite slow, while on rainy days, the discharge rate will be very quick, so quick that you might not even be able to measure it. If you time the electroscopical discharge rate inside an orgone accumulator, you will find that *it takes a longer time to discharge in the accumulator than in the open air*. The difference between the discharge rate inside the accumulator, and the discharge rate in the open air, is called the electroscopical discharge rate differential. This differential will be large on clear, sunny days, and minimal or zero on rainy overcast days. On rare occasions, an electroscope that is only weakly charged, or even completely discharged, may--if allowed to soak inside an orgone accumulator--spontaneously charge itself up to a higher level. All of these effects will vanish on rainy, overcast days. For more details, see the citations in the "Reference" section.

K) The Accumulator Evaporation Suppression Effect: This experiment requires a sensitive weighing scale that can measure to fractions of a gram. It also requires an accumulator and a thermally-balanced control enclosure of similar dimensions. For this control, do not use water-absorbent materials in the interior; instead, line the interior of the control with a non-metallic waterproofing material, such as plastic, enamel or varnish. Obtain and weigh two small, identically shaped and sized glass dishes of about 4" diameter and 1" high. Weight the dishes when empty, clean and dry. Next, add identical quantities of water to each vessel, filling them about halfway, and weigh again, calculating via subtraction the weight of water in each dish. Place one vessel of water inside the orgone accumulator, elevated on a small wood block, such that the bottom of the dish does not come into direct contact with the metal interior of the accumulator. The lid of the accumulator should be shut, but fixed open with a crack such that air may circulate. It should not be placed in a windy or sunlit area, however. Place the second vessel of water inside the control enclosure in a similar manner, on an identical wood block, and also prop its lid open. Place it in a location at least a meter away from the orgone accumulator, but with similar light, temperature, and wind characteristics. You may wish to drape a piece of black plastic over both the accumulator and control, in order to

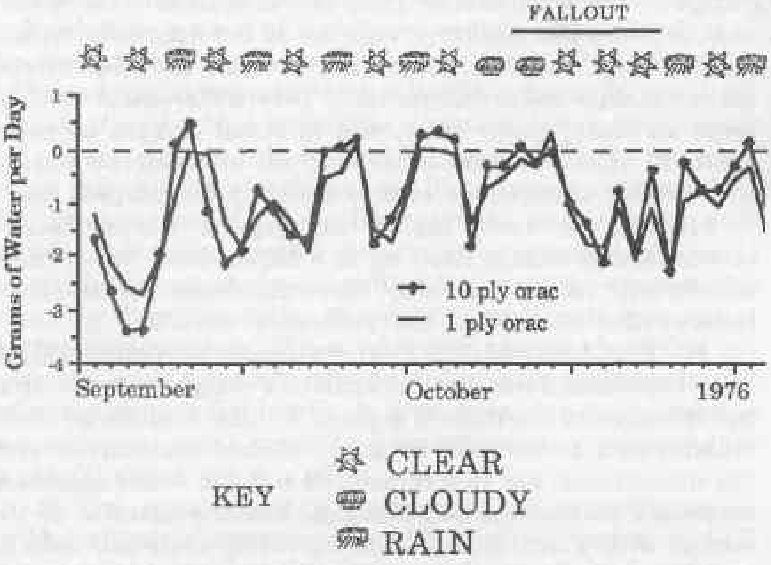

Water Evaporation Curve, EVo-EV: The curves identify the amount of water evaporated each day from a one cubic foot, ten ply orgone accumulator, and a similar one ply accumulator, minus the amount evaporated from a control enclosure. As the accumulator builds up its charge on clear, sunny days, it suppresses the evaporation of water inside itself. The control enclosure, however continues to evaporate moisture at a relatively high rate, up to several grams per day more than the accumulator. On wet, stormy days, the effect vanishes as orgone charge is lost at the Earth's surface, being taken up into the stormclouds. Disturbances in the regularity of the curve may also occur, such as (above) when radioactive fallout arrived at the laboratory site causing the accumulator to go temporarily "dead". (from DeMeo, J.: "Water Evaporation Inside the Orgone Accumulator", *J. Orgonomy*, 14:171-175, 1980) control for slight differences in light. Wait exactly 24 hours and remove the dishes of water, being careful not to spill any. Carefully weigh the dishes and compute the evaporative loss for the 24 hour period. Make this measurement once per day, preferably in the late evening, such that you can determine the amount of water evaporated each day from each container. You should find that the control enclosure evaporates significantly more moisture on clear, sunny days, when the accumulator suppresses water evaporation. On rainy days, when the accumulator no longer builds up a charge, the evaporation in the accumulator and control will be nearly identical. Subtract the quantity of water evaporated in the orgone accumulator from the quantity evaporated in the control for each 24 hour period. This quantity, called EVo-EV, will reveal the changing quantity of orgone energy charge in the local atmosphere, and in the accumulator. The evaporative values on any one day are less interesting than the dynamic manner in which the evaporation differential increases and decreases, according to the orgone energy charge at the Earth's surface.

14. Questions and Answers

- *Q: If the orgone energy really exists, why don't we hear about it from scientists working in the universities?*

A: Scientists working in the universities and research institutions have engaged in verifying research on the bions, orgone accumulator, and cloudbuster, and on the bioelectrical aspects of life, which Reich also pioneered. For example, Dr. James DeMeo, author of this *Handbook*, did research on the weather-related aspects of Reich's discoveries while he was a graduate student and Instructor at the University of Kansas. He continued with that research while a member of the faculty at Illinois State University and the University of Miami. Müschenich and Gebauer, of the University of Marburg in West Germany, recently completed a double-blind, controlled study on the physiological effects of the orgone accumulator on humans. Other scholars with a research or historical interest in the works of Wilhelm Reich have held positions at Harvard University, Temple University, the State University of New York, York University, McGill University, and elsewhere. Other research scientists have repeatedly verified and confirmed Reich's findings. Workshops and courses devoted to his works are now held at a few colleges and universities in the USA. Nonetheless, the history of science repeatedly shows that large institutions do not easily accommodate innovative research which may force radical changes in the major theories of science.

- *Q: Can an accumulator be used during wet or cloudy weather?*

A: Use of an accumulator during wet weather conditions will not be harmful, but it will be less effective, as the charge is significantly lower or absent at those times. It is best to use it during clear and sunny weather, when the atmospheric orgone energy continuum is strong and expansive, and the charge at the Earth's surface is greater.

- *Q: These accumulating devices are quite simple to construct. Aren't a lot of them accidentally constructed?*

A: A lot of "accumulators" are being constructed, without knowledge of the people involved. Every mobile home or house with a metal skin or side paneling will accumulate a charge, and a toxic one if aluminum is used. Oranur, and other toxic effects appear to readily occur in such houses, which are filled with all the modern orgone disturbing electromagnetic appliances, such as TV sets, microwave ovens, fluorescent lights, and so forth. Epidemiological studies that would address these general observations have never been performed.

- *Q: I have an old styrofoam beverage cooler. Can I line it with aluminum foil and make an accumulator?*

A: You can try this, but don't expect any firm results unless you comprehend and take into consideration virtually *all* of the procedures and warnings given in this *Handbook*. Styrofoam and aluminum are life-negative accumulating materials. If you run a biological experiment, you may wind up demonstrating only a life negative effect. For the scientist interested in orgone energy, these considerations are even more crucial and cannot be ignored.

- *Q: My accumulator gave very good charge the first months when it was in use, but now does not yield a good charge any more. Why is this?*

A: It is likely that the accumulator has been contaminated with dor. Some researchers have noted this effect, where the accumulator goes temporarily "dead", and hence keep their accumulators outdoors, sheltered from rain, but in the fresh air, with the lids or doors open so that air can circulate inside quite freely. You may be able to refresh a "dead" accumulator by wiping it inside and out with a damp cloth, every day for a week or so. Also, keep a bowl of water, or a draw bucket with draw tubes, inside it when it is not in use. Change the water in this bucket every day. Also, be certain that the accumulator is not near any of the oranur-producing devices previously identified, and that your neighborhood is as oranur-free as possible. The accumulator can also be sun-charged, by allowing it to sit in the direct sunlight for a few days. These steps should eliminate any dorish tendencies, and "rekindle" the charge.

- *Q: I have heard that sitting inside an accumulator will make a person more sexually potent. Is this true?*

A: This is mostly a false rumor propagated by Reich's enemies, who wrote smear articles in the 1940s and 1950s that branded Reich a lunatic, the accumulator a "sex box", and put false words into his mouth about the ability of the accumulator to restore lost sexual potency. However, Reich never made such a claim; in fact, he continually stressed the emotional and psychological foundations of sexual dysfunction, which could not be touched by treatment in an accumulator.

- *Q: Is the orgone energy accumulator legal? Can I get in trouble with the law for building or using one?*

A: There is no law against orgone energy, or the orgone energy accumulator. You may build, own, and use the blanket or accumulator in your own home, or elsewhere, as you choose. It can also be legally used for the self-treatment of any health-related condition, just as you can make very beneficial soups, purchase vitamins, or take soaking baths, without asking a doctor or the police. Understand, however, that forces within the medical community, pharmaceutical industry, and government are hard at work to make it illegal for you to do these things. If you are concerned about protecting your health freedoms, you should join forces with those social organizations that are working to preserve or extend those freedoms. *The price of liberty is eternal vigilance!*

Part III: Construction Plans for Orgone Accumulating Devices

15. Construction of a 2-Ply Orgone Energy Blanket

The orgone energy blanket is the most simple to construct of all the orgone charging and accumulating devices. They can be made at any size, and can be easily transported. You can use small ones while resting, or place large ones under and over a person who is immobilized in a bed. Like a standard accumulator, orgone blankets are not meant to be used for any prolonged period, though one can rest or nap with one if there is a need. In my experience, people will, even when sleeping, push off an uncomfortable orgone blanket just as they would a regular blanket. The following steps tell how to make an orgone energy blanket with final dimensions of 2' by 2'.

A) Obtain enough 100% wool fabric, or acrylic felt, to make three 2' x 2' squares. Also obtain several packages of very fine ("000" or "0000") steel wool pads.

B) Lay a 2' x 2' piece of fabric on a flat surface. Cover the exposed top surface of this fabric with a layer of steel wool, from unwrapped steel wool pads. Spread the steel wool out, so that it is not too thick. You should be able to see parts of the underlying fabric here and there.

C) On top of the steel wool, lay another 2' x 2' piece of fabric.

D) Cover the exposed top surface of this second fabric piece with another layer of steel wool.

E) Finish with another 2' x 2' piece of fabric, placed on top of the last steel wool layer. You now should have

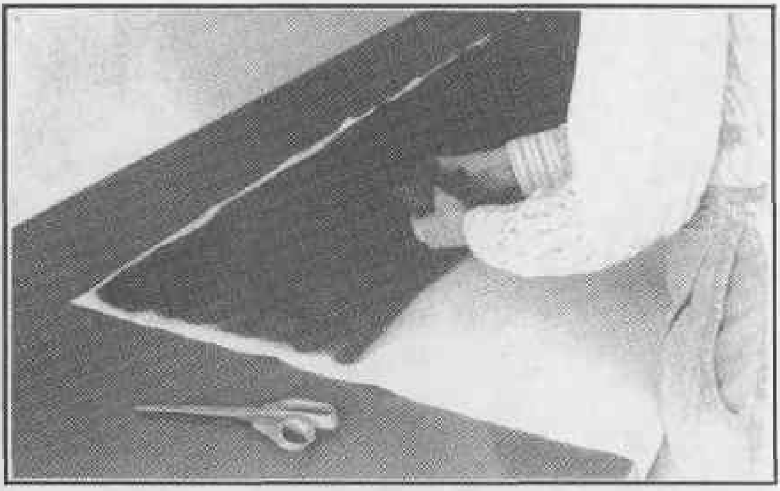

Spread steel wool loosely on section of cloth. Trim edges as shown.

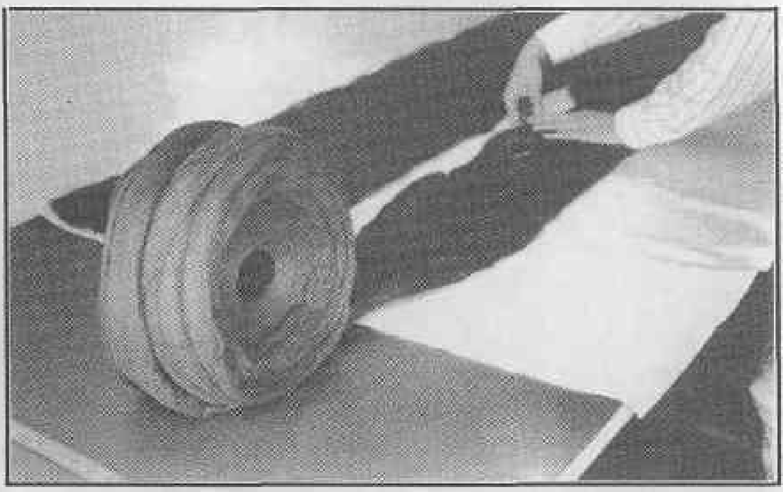

Rolls of steel wool can be purchased from manufacturers, or common steel wool pads, of "000" or "0000" grade, can be obtained at most hardware or paint stores.

Particle masks should be worn during construction to avoid breathing fine steel dust.

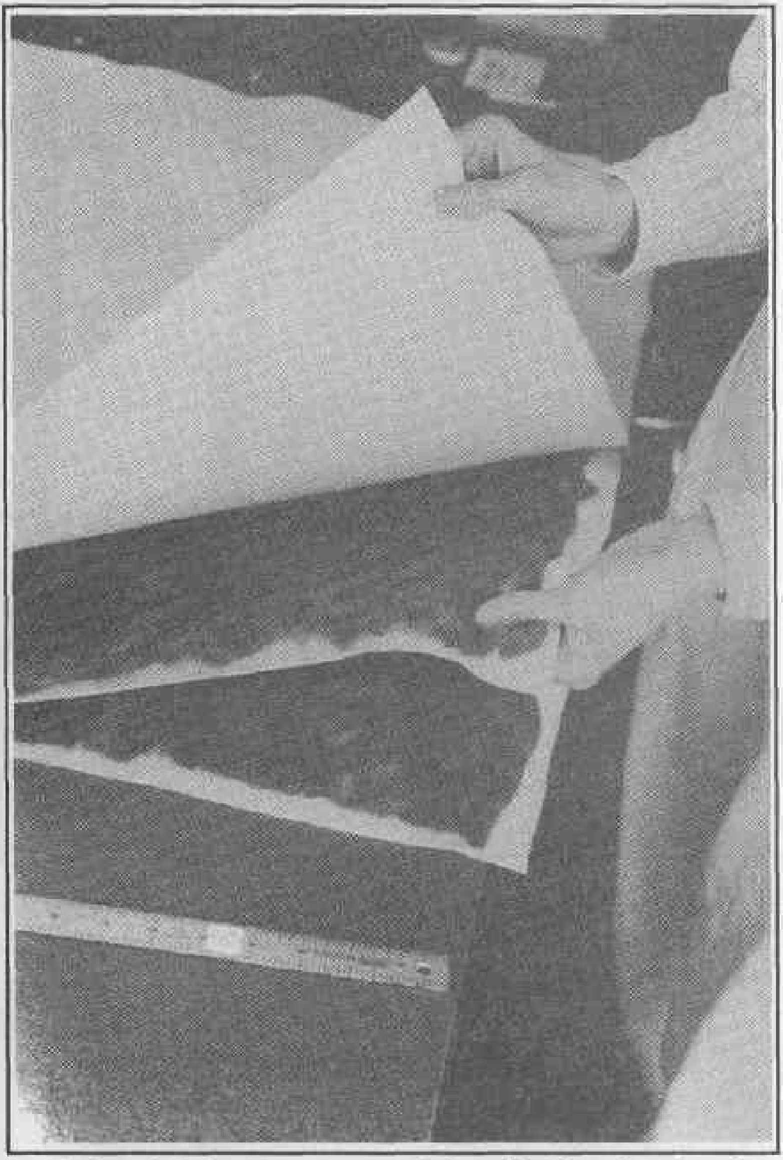

A functional orgone accumulating blanket showing the alternating layers of steel wool and cloth.

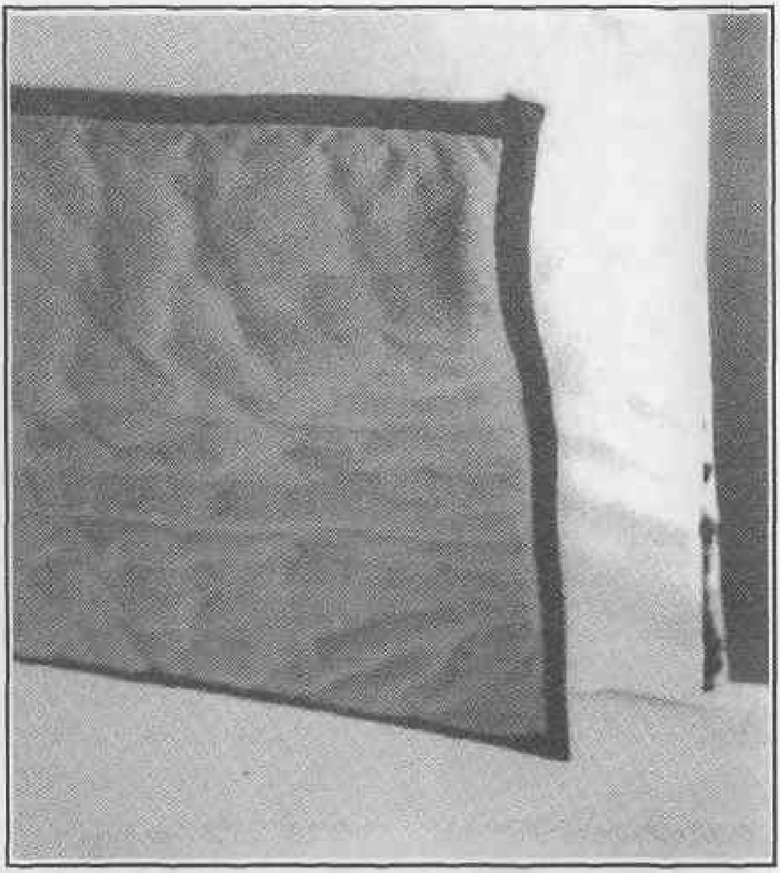

Finish off your blanket with a trim to keep everything together. Several quilt-stitches should be added to keep the insides from shifting around.

three pieces of fabric, with two layers of steel wool sandwiched in between.

F) Trim, stitch, and finish the borders, according to your own tastes and sewing skills.

G) Keep and use the blanket in an environment similar to that advised for a regular accumulator, away from any TV sets, microwave ovens, fluorescent lights, or other electromagnetic or radioactive gadgets. *Never use an orgone blanket with an electric blanket.* It can be stored on a hangar in the open air, or even inside a larger accumulator for greater charging.

H) Do not ever wash or dry clean your orgone blanket, as the steel wool will rust! Spot clean only with a *slightly* damp sponge.

I) Reich once made very heavy orgone blankets, composed of galvanized steel wire mesh, and alternating layers of wool and steel wool. While these work quite well, I find them to be uncomfortable and difficult to use. They do not appear to be any more effective than the simple design given here.

16. Construction of a 5-Ply Garden Seed Charger; The "Coffee Can" Accumulator

One can make a simple garden seed charger from a cleaned food container can or coffee can, using additional steel wool and fabric materials.

A) Empty a large coffee can or other steel or steel/tin food container (no aluminum!), clean it, remove all labels, and dry thoroughly. Make sure you save the cut-out metal lid, or make a replacement from another can, or from galvanized steel sheet metal. Use a can large enough to hold all the seeds you will be charging.

B) Obtain several yards of good 100% wool, or acrylic felt fabric. You will need enough fabric to go around the can about 5 times, plus enough for 5 round top pieces, and 5 similar bottom pieces.

C) Purchase several packages of very fine grade ("000" or "0000") steel wool. You will need enough unwrapped steel wool to cover an area equal to that of the fabric. Again, unwrap the steel wool pads as you need them, and spread it out.

D) Cut the fabric into a very long strip which is as wide as the can is tall. The length of this long strip should be about 6 times the circumference of the can. As you may not have a single strip of fabric this long. You can splice several pieces together.

E) Lay the long fabric strip out flat, and spread a thin layer of steel wool on top of it. Lay the empty can on one end of the steel wool/fabric strip and roll the can up inside the strip. Stop when the strip has been wrapped around the can about five times or more. Add a final layer of fabric to the outside, and stitch or tape it in place so that it will not unravel.

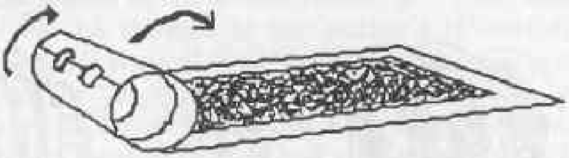

F) Measure the diameter across the top of the can, including the wrappings of fabric and steel wool. Cut out 10 circles of fabric material of this same diameter, 5 for the top, 5 for the bottom.

G) Sandwich steel wool between the fabric circles such that you have 4 layers of steel wool between 5 layers of fabric. Make two of these fabric/steel wool sandwiches, one of which will be used to cover the bottom of the can, and the other for the top.

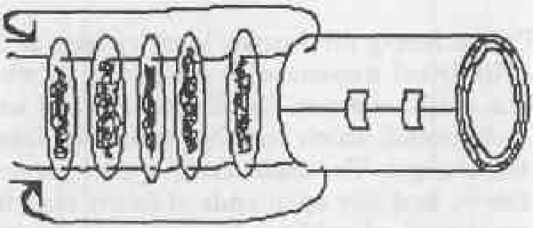

H) Take the metal disk cut from the top of the can and file smooth any jagged edges. Punch two small nail holes near the center, about 1/4" apart. Using a heavy upholstery or knitting needle, thread some heavy twine, wire, or yarn through the center of one of the steel wool/fabric sandwiches, and bring the twine through the holes in the metal can lid. Secure the can lid to the center of the steel wool/fabric sandwich. The steel wool/fabric sandwich should be about 2" wider in diameter than the metal can lid.

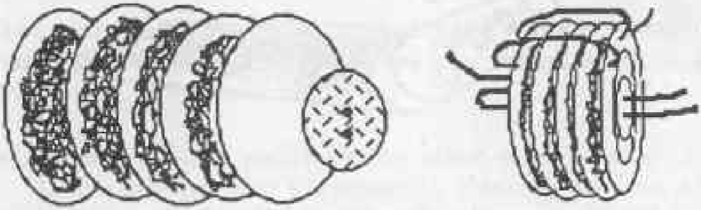

I) Using heavy thread, loosely stitch together the edges of the top steel wool/fabric sandwich (the one sewn to the metal can lid). Also loosely stitch the edges of the bottom steel wool/fabric sandwich together, and sew it to the bottom edge of the steel wool/fabric strip, wound around the can. Except for the top opening, the metal can should now be encased in the steel wool/fabric material.

J) Find a heavy pillowcase, laundry bag, or other larger cylindrical non-metallic container in which to keep the entire charger. Or, if you are good with a needle and thread, stitch together your own fabric cover for the charger. The main thing is that the outer layer of fabric, and any open ends of fabric showing pieces of steel wool, should not be subject to "knocking about" or moisture, such that it starts to fall apart or rust.

K) Review the section on seed charging, in the "Simple Experiments" chapter for instructions and additional ideas on the use of your charger. Or, as an alternative to constructing this accumulator, you could store your seeds inside a large metal cookie tin which would then be wrapped up inside the multiple folds of a very large orgone blanket, or be stashed inside one of the other, larger accumulators. Just realize that the greater the number of plys, and the greater the absolute quantity of materials going into the construction of the accumulator "pile", the stronger will be the charge. At the author's laboratory, for instance, the small 5- ply coffee can charger is stored inside the 10-ply, one-cubic foot accumulator, which in turn is stored inside the 3-ply large accumulator. This is a total of 18 plys, and yields a charge which is readily sensible.

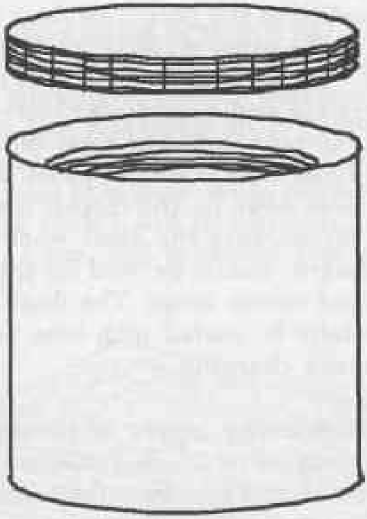

17. Construction of a 10-Ply Orgone Energy Accumulator

A very powerful, one-cubic foot, 10-ply accumulator can be made following the instructions below.

A) Cut six galvanized steel, 27 gauge sheet metal squares, measuring 1' x 1'. Use heavy tape, on the outer metal walls only, to construct a metal cube. Leave the top of the cube open, and do not tape it into place. The interior of the cube should remain bare metal, without any tape showing.

B) Use very fine ("000" or "0000") steel wool, and heavy clear plastic acrylic carpet protector material for construction of the plys. The clear plastic carpet protector is the same material used in model homes to protect carpets from wear, and is often sold on rolls in hardware and department stores. It is not cheap, but works very well. The clear plastic carpet runner should have rows of small tips on the side which normally faces down on the carpet; these tips work very nicely for holding the steel wool in place. Celotex (soundboard) should be used for the final outer layer, with wood corner strips. The final outer celotex may additionally be coated with bees wax and/or shellac to increase charging.

C) Ten alternating layers of plastic and steel wool roughly measure 2" in thickness. This being the case, you should construct the outer celotex casing in the shape of a cube, the interior dimensions of which will be 16" x 16" x 16". Cut six celotex panels with the following dimensions:

Celotex Panels:

top: 17" x 17"
bottom: 17" x 17"
2 sides: 17" x 16"
2 sides: 16" x 16"

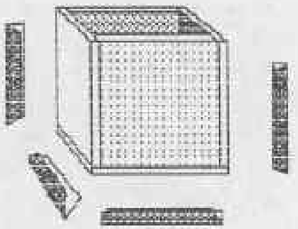

D) Use small nails and glue to fasten five of the six celotex panels together, to make a cube shape. Again, as with the metal box, do not attach the top. Add extra glue to the edges of the assembled celotex box, and allow it to dry before proceeding.

E) Using a mitre box, cut wood corner strips for the outside edges of the celotex box. Nail and glue these wood corner strips to the celotex box for added strength.

F) Cut 20 square pieces of plastic carpet runner, 16" x 16". Put ten of these squares aside for later use. Lay the other ten squares one at a time inside the bottom of the celotex box, tips facing up. In between each plastic square place a layer of steel wool, taken from unwrapped steel wool pads. When finished, the top of the last plastic square will face upward in the bottom of the celotex box, and it should also be covered with steel wool.

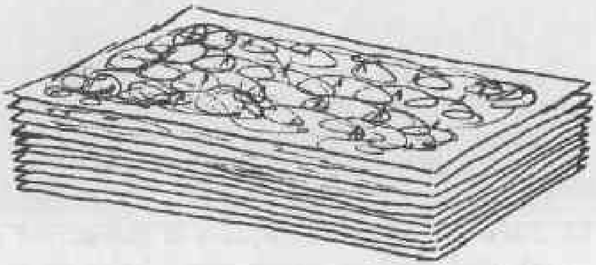

G) Place the galvanized steel sheet metal cube inside the celotex box, on top of the ten plys of plastic and steel wool. If you have constructed the celotex box correctly, the top of the metal cube should be about 2" below the top of the celotex box, and a space of about 2" should exist between the sides of the metal cube and the interior sides of the celotex box.

H) Cut 20 pieces of plastic, 12" x 16", and 20 more pieces, 12" x 12". These will be used to fill the side spaces between the metal cube and the celotex. Layer each plastic piece with steel wool, and stack into bunches of ten layers each. Do this on a flat surface before attempting to place them into a vertical position, in between the celotex and metal boxes.

I) Place the two stacks of 16" x 12" plastic/steel wool between the celotex and metal cube, on opposite sides of the metal cube. An outer layer of plastic should lay against the inner wall of the celotex box, while an inner layer of steel wool should rest against the outside of the metal box. The upper edge of the plastic should be nearly flush with the upper edge of the metal cube, both of which should remain about 2" below the upper edge of the celotex box.

J) Place the two stacks of 12" x 12" plastic/steel wool in the two remaining spaces between the celotex and metal cube, as given in the previous step.

K) Take the 10 remaining pieces of 16" x 16" plastic carpet runner and layer them with steel wool. Stack them up and set them aside. Unlike prior stacks, however, do not finish the final plastic layer with steel wool.

L) Take the remaining square of galvanized sheet metal, and drill or punch small holes in each corner, about 1/2" from each corner. The holes should be

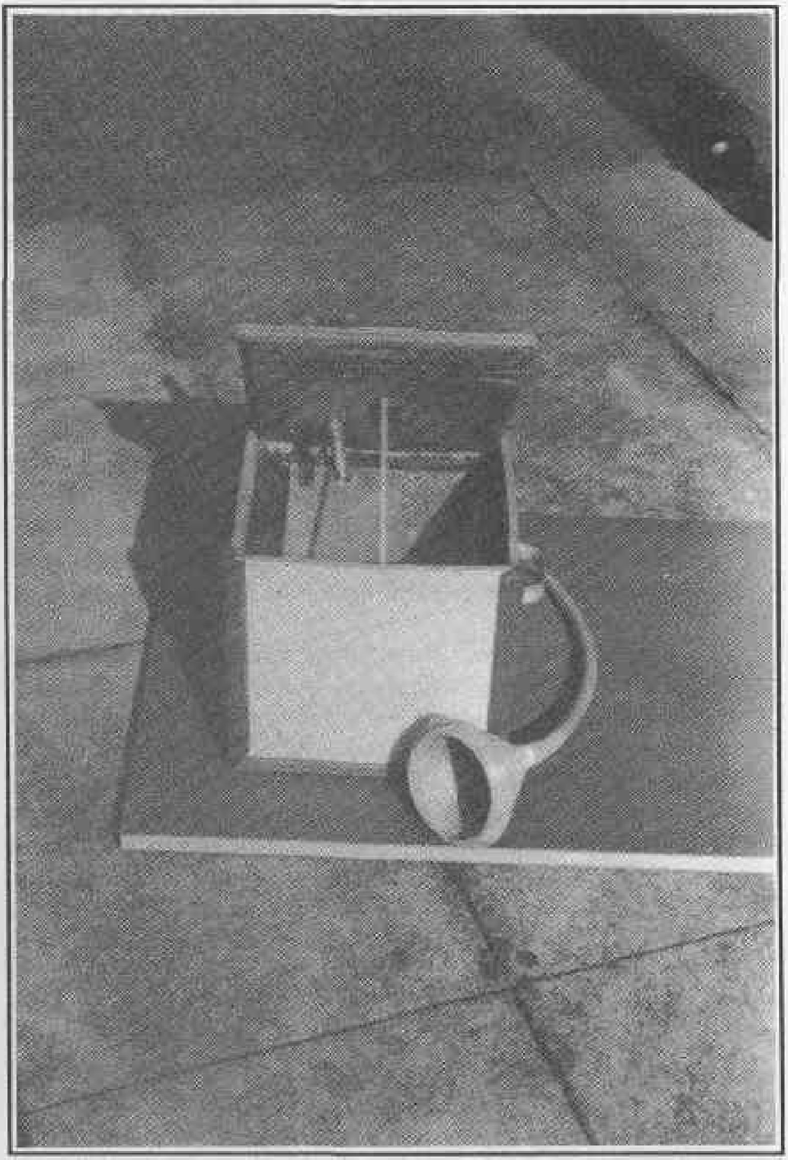

A one-cubic foot, 10-ply accumulator with attached funnel.

large enough to accommodate a long but narrow screw.

M) Find a rough wood or carpeted surface to work on. Lay the stack of plastic/steel wool squares on the last piece of celotex. Center it on the celotex; there should be about 1/2" of celotex appearing all around the stack. Now place the sheet metal square (with the holes) on top of the plastic/steel-wool/celotex stack, and center it also. There should be about 2" of plastic extending beyond the edge of the metal square on each side. Use some masking tape to temporarily hold the celotex/plastic/steel stack together.

N) Using an ice pick, carefully make four vertical holes through the plastic/steel wool stack, and through the celotex, using the corner holes in the sheet metal square as a guide. Do not use a drill, as the steel wool will spindle around the drill bit.

O) Using four skinny bolts, along with nuts and LARGE washers, secure the metal square and plastic/steel wool stack against the celotex square. Use a bolt no longer than necessary, such that the bolt ends will not protrude significantly. When completed, this whole lid assembly should fit snugly on top of the celotex box. The metal plate attached to the lid should closely, but not perfectly, align with the metal cube interior. With the lid in place, only bare metal should face to the interior of the accumulator.

P) For handles, first firmly glue a flat, wide and long wood strip to two outer sides of the celotex cube, near to the top. When completely dry, screw wood or metal handles to these wood strips. A handle should also be attached to the upper, outer side of the lid assembly, using a similar support mechanism. The celotex is simply too lightweight to accept screws or nails alone. You may likewise install a hinge between the lid assembly and the box, or coaster wheels to the bottom, but these are not necessary.

Q) For additional charging strength, the outer celotex walls may be coated with melted bees wax or paraffin (not really necessary), but in any case should be given several coats of protective natural shellac.

R) For additional charging strength, store this cubical accumulator in the bottom part of a larger, human-sized accumulator, under the bench that you normally sit upon. Be sure to maintain a clean and uncontaminated environment for your accumulator, as per the points given in the previous chapters. Prop the lid open when not in use, and store in a clean, dry place, without electromagnetic or nuclear contamination.

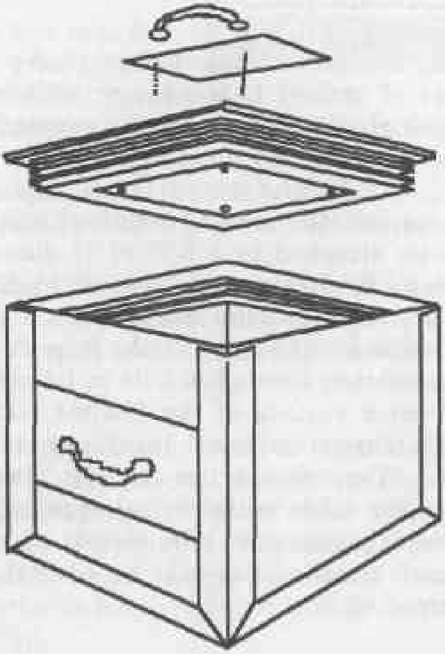

18. Construction of an Orgone Shooter Funnel

The shooter funnel is similar to other accumulating devices, but has an open face allowing for external irradiation of objects. It is often connected to a larger box accumulator, but this is not absolutely necessary.

A) Obtain a galvanized steel funnel with about 6" diameter at the large end, from a hardware, farm supply, or auto supply store. Auto supply stores sometimes sell these with a flexible metal hose already attached, for dispensing oil into a car engine, and this may help in the later attachment to a box accumulator.

B) Coat the outer metal surface of the funnel with a layer of melted bees-wax, or with strips of black plastic electrical tape, leaving exposed the bare metal surface of the funnel interior.

C) If desired, the small "draining" end of the funnel can be attached to a 3/4" or 1" diameter length of hollow, flexible metal "greenfield BX" cable (use galvanized steel, *not aluminum*). The other end of the cable is then run into the interior of a small box accumulator, through a hole in its side or lid. Wrap the outer surface of the flexible metal cable with black-plastic electrical insulation tape. (See page 121.) Your shooter funnel will then draw orgone down the cable to the funnel opening, increasing its radiating strength. Or, simply store the shooter funnel inside a box-type accumulator, to keep it charged up.

19. Construction of an Orgone Shooter Wand

The shooter wand is a very simple means for demonstrating the subjective sensations of the orgone radiation, and also for irradiation of orgone energy into body cavities. Simply store the completed shooter wand inside a well maintained accumulator, and remove it for use when necessary. Under relaxed conditions, most people can hold this wand in the hand, or place it on the solar plexus or upper lip, and readily feel the soft radiant glow of the orgone energy.

A) Obtain a pyrex test tube of about 3/4" to 1" diameter, and 6" to 9" long, from a laboratory or medical supply company.

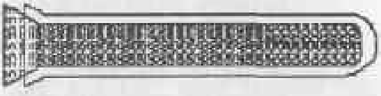

B) Fill the test tube full of fine grade ("000" or "0000") steel wool. Compress to a reasonable firmness.

C) Seal the open end of the test tube with a rubber stopper, and tape it shut with plastic electrical tape.

D) Place the shooter wand inside a small orgone accumulator for a period of several days or weeks before using. Keep it stored inside the accumulator in between uses.

E) If you use the shooter wand to irradiate the throat or other body cavities, or if the pyrex glass otherwise becomes soiled, wipe the glass clean with isopropyl alcohol prior to putting it back inside the accumulator. The wand should always be alcohol cleaned and air dried prior to being replaced into the accumulator for charging.

20. Construction of a Large 3-Ply Orgone Energy Accumulator

This accumulator is large enough to sit in, and is composed of 6 large rectangular panels. Each panel is made from a wood frame, galvanized steel sheet metal (27 gauge), steel wool, fibreglass (*without* aluminum backing), and celotex (soundboard). One side of each wood frame is faced with galvanized steel, the other with celotex, and the three alternating layers of fibreglass and steel wool are sandwiched in between.

A) First compute the size of the panels for an accumulator that will fit your personal needs, adding the necessary dimensions for overlap between the various panels. The side panels and back panel should physically sit on the edges of the bottom panel. The back panel should snuggle between the two side panels. The top panel should overlap and rest upon both side panels and the back panel, covering them. The door panel should, like the back panel, snuggle between the side panels when it is shut. I believe this arrangement is the simplest possible, and most efficient to build. Dimensions are given below for accumulators to accommodate people of varying sitting heights, but of average weight. As the distance of the body surface from the metal walls increases, there will be a reduction in effectiveness of the accumulator. The dimensions of the accumulator should be carefully selected to meet your needs. An additional 1/2" clearance is provided in the width dimension (1/4" each side), such that the door will open and shut freely.

Panel Dimensions:	*Large Size*	*Medium Size*	*Small Size*
Top:	29.5" x 35"	26.5" x 32"	23.5" x 28"
Bottom:	29.5" x 35"	26.5" x 32"	23.5" x 28"
Left side:	35" x 58"	32" x 54"	28" x 50"
Right side:	35" x 58"	32" x 54"	28" x 50"
Back:	25" x 58"	22" x 54"	19" x 50"
Door:	25" x 52"	22" x 48"	19" x 44"

Interior Dimensions:			
Height:	58"	54"	50"
Width:	25"	22"	19"
Depth:	31"	28"	24"

You will need:
Height = height sitting erect in a chair + about 3"
Width = shoulder width + about 4" (2" each side)
Breadth = sitting, knee to back distance + about 3"

B) Make wood frames of 3/4" x 1-1/2" pine (called "one by two" in lumber yards), like picture frames, such that the outer edges conform to the computed dimensions for your accumulator panels. Nail and glue all joints.

C) Arrange the open wood frames together, as they will be when the accumulator is complete, to make sure that all dimensions have been properly calculated and cut. If there is an error in your calculations, now is the time to find out, before the more expensive celotex and galvanized sheet metal have been cut.

D) Cut the celotex panels to size, from 1/2" or 3/4" thick celotex sheet, to cover fully one side of each wood frame. Nail and glue the celotex panels to each wood frame. Use 1/4" plywood instead of celotex for the bottom panel (and only the bottom panel).

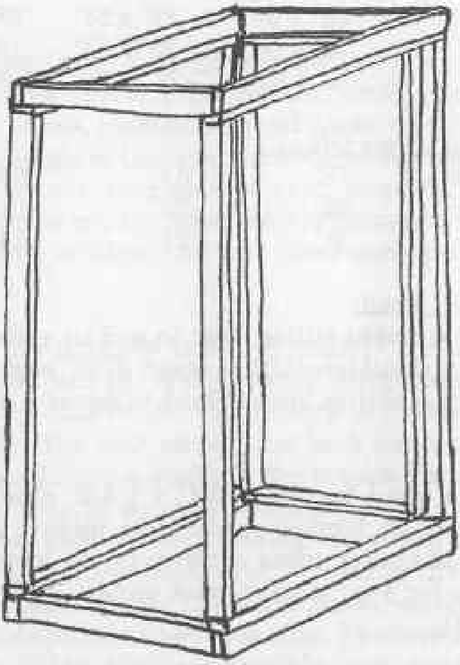

Assembled Wood Frames for Accumulator Panels

E) Cut batts of 1/4" thick fibreglass material to size and place a layer inside each of the open panel frames. Use gloves and a mask to protect yourself. Do not compress. Avoid lumps and holes. You can use wool, cotton, acrylic felt, or rock wool instead, if you wish, but for a large accumulator like this, the costs will be higher, and the accumulation will not be significantly stronger. These other materials may yield a different "feel" to the orgone charge, and if that is important to you, the cost may be justified.

F) Unroll very fine ("000" or "0000") steel wool pads and place a layer inside each of the open panel frames, on top of the fibreglass. Leave it fluffy, at around the same thickness as when unrolled, in as uniform a layer as possible. Some steel wool comes in large rolls which, if you can get it, will speed up construction for large accumulators.

G) Repeat steps E and F, placing a new layer of fibreglass on top of the previous steel wool layer, and another layer of steel wool on top of that.

H) Again repeat steps E and F, placing another new layer of fibreglass on top of the previous steel wool layer, and another layer of steel wool on top of that.

I) You should now have three alternating layers of fibreglass and steel wool contained in the open frame of each panel. The final layer facing you should be composed of steel wool. The panels should also be filling up, and may need to be slightly compressed before the final layer of galvanized steel sheet metal can be added. If you have used some other type of non-metallic material than fibreglass, and if the material lies loosely in the frame, you may have problems with slumping of the plys when the panels are finally sealed and placed in upright position. If this is the case, now is the time to do something to pre-

End Cross-Section of Side, Top, Back and Door Accumulator Panels

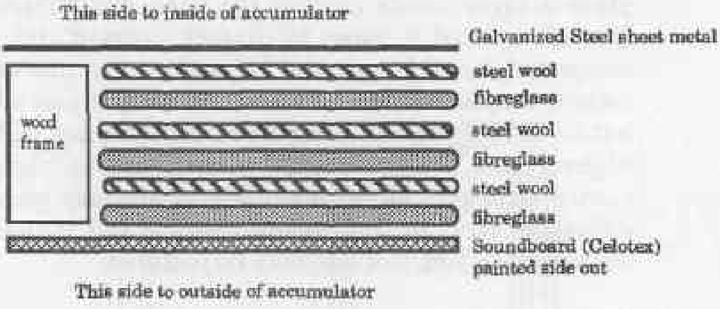
This side to inside of accumulator
Galvanized Steel sheet metal
steel wool
fibreglass
wood frame
steel wool
fibreglass
steel wool
fibreglass
Soundboard (Celotex) painted side out
This side to outside of accumulator

End Cross-Section of Bottom Accumulator Panel

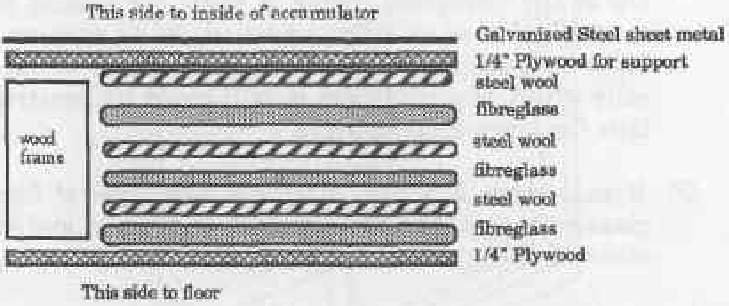
This side to inside of accumulator
Galvanized Steel sheet metal
1/4" Plywood for support
steel wool
fibreglass
wood frame
steel wool
fibreglass
steel wool
fibreglass
1/4" Plywood
This side to floor

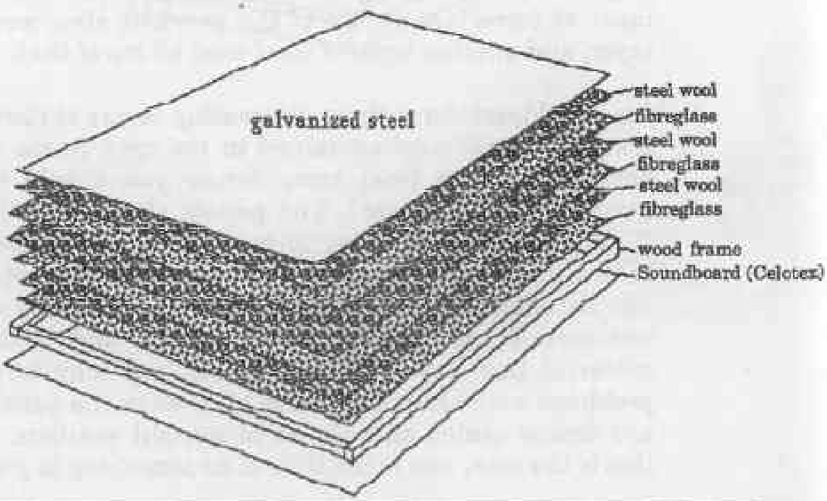
galvanized steel
steel wool
fibreglass
steel wool
fibreglass
steel wool
fibreglass
wood frame
Soundboard (Celotex)

Front View, full-size 3-ply accumulator

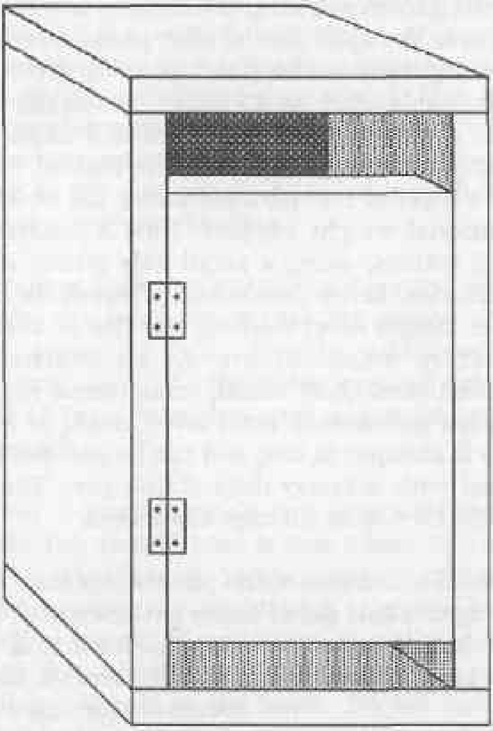

For Warmer, Tropical Climates: *Use door with 3" convection gaps at top and bottom, as per measurements on p.127. (Photo on p.23.) If needed, tack screen strips to door top and bottom, across gaps, to keep out insects.*

For Cooler or Cold Climates: *Use full height door with small window. (Photo on back cover.) Before cutting door materials, adjust door measurements (p.127) to only 1/4" less than full height of side panels. Before assembling door, cut matching 6" square openings in door sheet metal and celotex, centered at face height, then frame opening with 1x2" wood strips. Complete assembly as per instructions.*

Use hinges with removable hinge pins for easier assembly and take-down. Secure inside of door with hook and eye.

vent slumping.

J) Cut the galvanized steel sheet metal to size such that it fits over the open part of each panel, overlapping and covering the wood frames. Use the lightest sheet metal available, such as 27 gauge, which can be cut with hand shears for trimming, and still add to structural strength of the panels. For the bottom (floor) panel, add a layer of 1/4" plywood under the sheet metal, for additional weight support. Nail it securely onto the wood frames, using a small hole punch if necessary. Small steel tacks should easily penetrate through the sheet metal. After nailing, use file or shears to trim all sharp metal corners. As an alternative to galvanized steel sheet metal, some people have effectively used galvanized steel *wire mesh*, or *wire screen*. This is cheaper to use, and can be secured to the wood frame with a heavy duty staple gun. The steel wool should be visible through the screen.

K) Assemble and secure the panels together. Start by securing one side panel to the bottom panel, by using an "L" shaped metal brace at the front and rear of the side panel, near the bottom. Use screws, such that the accumulator can be taken apart later for easy moving. Secure the other side panel in a similar fashion, followed by the back panel. The back panel must be secured independently by small metal braces, placed between the wood frames of the bottom and back. Add the top panel, and secure it to the sides and back in a similar manner. The accumulator should now be quite sturdy, and is almost complete.

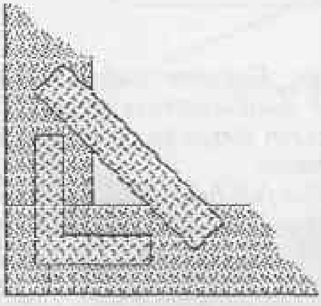

L) Carefully mark and drill holes for door hinges and attach them to the door panel and side panel. Make sure you center the door in a way that leaves equal space at the top and bottom, either for convection gaps (for the vented door style), or as clearance to allow the door to freely swing open and shut (for the full-height windowed door style). Use hinges with removable pins so that you can easily take the door off when moving the accumulator. After fixing the door to the side panel, it should open fully and close snugly, without binding.

M) Attach a hook and eye to the door and side panel opposite the hinge, such that the person sitting inside can secure it shut. Finally, add several coats of natural shellac to the outer exposed celetox, to protect it from humidity and add to its accumulating strength.

N) Your accumulator is now complete, except for a seat. You should have a seat which will allow you to place other accumulator chargers underneath. For this purpose, you might wish to specially construct a wooden bench. Wood is a good material to use, as it does not absorb the orgone energy significantly, and is not cold to the touch. Do not use woods that have been soaked in preservative or formaldehyde, however. Metal chairs are OK, but will be quite cold to sit on unless covered with a light fabric.

O) You may also wish to construct a *chest board,* or *orgone pillow,* for use inside the accumulator. As you sit inside, notice that there is a great distance from your chest to the front metal wall. This large distance inhibits orgone irradiation to your chest. An additional, small accumulator panel, similar to those used for the wall panels, could be constructed for use inside the larger box, to bring the radiation closer to the chest. However, a simpler way is to use a bundle of cotton, wool, or acrylic felt, rolled into a large pillow shape with equal layers of steel wool. The final outer layer should be composed of steel wool, and the entire bundle is then placed inside a thin cotton pillowcase. It should be large enough that it fits very snugly into the pillowcase. By holding this orgone pillow close to the chest, while you are in the accumulator, it will irradiate those frontal areas of your body that are not so well irradiated by the accumulator walls. Leave the orgone pillow inside the accumulator when not in use, to keep it charged up. You can also use this pillow outside of the accumulator, in a manner similar to the orgone blanket, with equally good results.

P) Do not connect electrical appliances to the accumulator. Follow the cautions given in the previous chapters. You can read a book while inside the accumulator, but should use either a strong *external* light source (to shine a beam of incandescent light into the accumulator), or you can use a battery-powered reading light inside. Again: *no fluorescent lights, television sets, heating blankets or pads, or other electrical or electromagnetic devices!!*

SELECTED REFERENCES

Books By Wilhelm Reich:
republished by Farrar Straus & Giroux:

The Bioelectrical Investigation of Sexuality and Anxiety
The Bion Experiments
The Cancer Biopathy (Discovery of the Orgone, Volume 2)
Character Analysis
Children of the Future: On the Prevention of Sexual Patholgy
Cosmic Superimposition: Man's Orgonotic Roots in Nature
The Early Writings of Wilhelm Reich
Ether, God and Devil
The Function of the Orgasm (Discovery of the Orgone, Volume 1)
Genitality in the Theory and Therapy of Neurosis
The Invasion of Compulsory Sex-Morality
Listen, Little Man!
The Mass Psychology of Fascism
The Murder of Christ (Emotional Plague of Mankind, Volume 2)
Passion of Youth
People in Trouble (Emotional Plague of Mankind, Volume 1)
Record of a Friendship, Correspondence of Wilhelm Reich and A.S. Neill
Reich Speaks of Freud
Selected Writings
The Sexual Revolution

Relevant Books and Special Reports by Wilhelm Reich:
never republished:

The Orgone Energy Accumulator, Its Scientific and Medical Use, Orgone Institute Press, Maine, 1951.

The Oranur Experiment, First Report (1947-1951), Wilhelm Reich Foundation, Maine, 1951.

The Einstein Affair, 1939-1952, Wilhelm Reich Biographical Material, History of the Discovery of the Life Energy, Documentary Volume A-IX-E, Orgone Institute Press, Rangeley, Maine, 1953.

Relevant Research Articles by Wilhelm Reich:

"Orgonotic Pulsation: The Differentiation of Orgone Energy from Electromagnetism", *Int. J. Sex-Economy & Orgone Research*, III:74-79, 1944.

"Orgone Biophysics, Mechanistic Science and 'Atomic Energy'", *Int. J. Sex-Economy & Orgone Research*, IV:200-201, 1945.

"Orgonotic Light Functions 1: Searchlight Phenomena in the Orgone Energy Envelope of the Earth", *Orgone Energy Bull.*, I(1):3-6, 1949.

"Orgonotic Light Functions 2: An X-Ray Photograph of the Excited Orgone Energy Field of the Palms", *Orgone Energy Bull.*, I(2):49-51, 1949.

"Orgonotic Light Functions 3: Further Characteristics of Vacor Lumination", *Orgone Energy Bull.*, I(3):97-99, 1949.

"Meteorological Functions in Orgone-Charged Vacuum Tubes", *Orgone Energy Bull.*, II(4):184-193, 1950.

"The Storm of November 25th and 26th, 1950", *Orgone Energy Bull.*, III(2):76-80, 1951.

"Three Experiments with Rubber at the Electroscope", *Orgone Energy Bull.*, III(3):144-145, 1951.

"The Anti-Nuclear Radiation Effect of Cosmic Orgone Energy", *Orgone Energy Bull.*, III(1):61-63, 1951.

"'Cancer Cells' in Experiment XX", *Orgone Energy Bull.*, III(1):1-3, 1951.

"The Leukemia Problem: Approach", *Orgone Energy Bull.*, III(2):139-144, 1951.

Selected Research Articles on the Orgone Energy by Others:

Anderson, W.A.: "Orgone Therapy in Rheumatic Fever", *Orgone Energy Bull.*, II(2):71-73, 1950.

Atkin, R.H.: "The Second Law of Thermodynamics and the Orgone Accumulator", *Orgone Energy Bull.*, I(2):52-60, 1949.

Atkin, R.H.: "A Space-Energy Continuum", *Orgone Energy Bull.*, IV(4):197-206, 1952.

Baker, C.F.: "The Orgone Energy Continuum", *J. Orgonomy*, 14(1):37-60, 1980.

Baker, C.F.: "The Orgone Energy Continuum: the Ether and Relativity", *J. Orgonomy*, 16(1):41-67, 1982.

Baker, C.F.: "Atmospheric Pulsation: Air and Water", *Annals, Inst. Orgonomic Science*, 2(1):25-32, 1985.

Baker, C.F., et al: "The Reich Blood Test", *J. Orgonomy*, 15(2):184-218, 1981.

Baker, C.F., et al: "The Reich Blood Test: 105 Cases", *Annals, Inst. Orgonomic Science*, 1(1):1-11, 1984.

Baker, C.F., et al.: "Wound Healing in Mice, Part I", "...Part II", *Annals, Inst. Orgonomic Science*, 1(1):24-32, 1984; 2(1):7-24, 1985.

Baker, C.F., et al.: "The Reich Blood Test: Clinical Correlation", *Annals, Inst. Orgonomic Science*, 2(1):1-6, 1985.

Baumann, J.: "Some Observations of the Atmospheric Orgone Energy", *Orgone Energy Bull.*, II(2):71-73, 1950.

Blasband, R.A.: "Thermal Orgonometry", *J. Orgonomy*, 5(2):175-188, 1971.

Blasband, R.A.: "The Orgone Energy Accumulator in the Treatment of Cancer Mice", *J. Orgonomy*, 7(1):81-85, 1973.

Blasband, R.A.: "Effects of the Orac on Cancer in Mice: Three Experiments", *J. Orgonomy*, 18(2):202-211, 1985.

Bremmer, K.M.: "Medical Effects of Orgone Energy", *Orgone Energy Bull.*, V(1-2):71-83, 1953.

Brenner, M.: "Bions and Cancer, A Review of Reich's Work", *J. Orgonomy*, 18(2):212-220, 1984.

Claymond, H.J.: "Preliminary Indications for an Energetic Concept of Soil Fertility", *Offshoots of Orgonomy*, 11:8-12, Fall 1985.

Cott, A.A.: "Orgonomic Treatment of Ichthyosis", *Orgone Energy Bull.*, III(3):163-166, 1951.

DeMeo, J.: "Effect of Fluorescent Lights and Metal Boxes on Growing Plants", *J. Orgonomy*, 9(1):95-99, 1975.

DeMeo, J.: "Seed Sprouting Inside the Orgone Accumulator", *J. Orgonomy*, 12(2):253-258, 1978.

DeMeo, J.: "Water Evaporation Inside the Orgone Accumulator", *J. Orgonomy*, 14(2):171-175, 1980.

Dew, R.A.: "Wilhelm Reich's Cancer Biopathy", in *Psychotherapeutic Treatment of Cancer Patients*, J.G. Goldberg, ed., Free Press, NY,

1980.

Eden, J.: "Personal Experiences with Oranur", *J. Orgonomy*, 5(1):88-95, 1971.

Eden, J.: *Orgone Energy, The Answer to Atomic Suicide*, Exposition Press, NY, 1972.

Eden, J.: *Animal Magnetism and the Life Energy*, Exposition Press, NY, 1974.

Espanca, J.: "The Effect of Orgone on Plant Life, Parts I - VII", *Offshoots of Orgonomy*, 3:23-28, Autumn 1981; 4:35-38, Spring 1982; 6:20-23, Spring 1983; 7:36-37, Autumn 1983; 8:35-43, Spring 1984; 11:30-32, Fall 1985; 12:45-48, Spring 1986.

Espanca, J.: "Orgone Energy Devices for the Irradiation of Plants", *Offshoots of Orgonomy*, 9:25-31, Fall 1984.

Grad, B.: "Wilhelm Reich's Experiment XX", *Cosmic Orgone Engineering*, VII(3-4):203-204, 1955.

Grossmann, W.: "Observations of Orgone Energy Lumination", *Orgone Energy Bull.*, IV(1):59-60, 1952.

Hamilton, A.E.: "Child's-Eye View of the Orgone Flow", *Orgone Energy Bull.*, IV(4):215-216, 1952.

Harman, R.A.: "Further Experiments with Negative To Minus T", *J. Orgonomy*, 20(1):67-74, 1986.

Hoppe, W.: "My First Experiences With the Orgone Accumulator", *Int. J. Sex-Economy & Orgone Research*, IV:200-201, 1945.

Hoppe, W.: "Sex-Economy and Orgone Research in Palestine", *Annals of the Orgone Institute*, I:73-81, 1947.

Hoppe, W.: "My Experiences With the Orgone Accumulator", *Orgone Energy Bull.*, I(1):12-22, 1949.

Hoppe, W.: "Further Experiences with the Orgone Accumulator", *Orgone Energy Bull.*, II(1):16-21, 1950.

Hoppe, W.: "Orgone Versus Radium Therapy in Skin Cancer, Report of a Case", *Orgonomic Medicine*, I(2):133-138, 1955.

Hoppe, W.: "The Treatment of a Malignant Melanoma with Orgone Energy", contained in *In the Wake of Reich*, D. Boadella, ed., Coventure Press, London, 1976.

Hughes, D.C.: "Some Geiger-Muller Counter Observations After Reich", *J. Orgonomy*, 16(1):68-73, 1982.

Konia, C.: "An Investigation of the Thermal Properties of the ORAC, Part I & II", *J. Orgonomy*, 8(1):47-64, 1974; 12(2):244-252, 1978.

Lance, L.: "Effects of the Orgone Accumulator on Growing Plants", *J. Orgonomy*, 11(1):68-71, 1977.

Lassek, H.: "Orgone Accumulator Therapy of Severely Diseased Persons", *Pulse of the Planet*, 3:39-47, 1991.

Lassek, H.: "Vegeto-Orgontherapie Nach Wilhelm Reich", *Natural Healing Methods for the European Community*, Doc. Reg. V.5, Min. of Tech., Niedersachen, Univ. Lueneburg, Germany 1991.

Lappert, P.: "Primary Bions Through Superimposition at Elevated Temperature and Pressure", *J. Orgonomy*, 19(1):92-112, 1985.

Levine, E.: "Treatment of a Hypertensive Biopathy with the Orgone Accumulator", *Orgone Energy Bull.*, III(1):53-58, 1951.

Müschenich, S. & Gebauer, R.: "The (Psycho-) Physiological Effects of the Reich Orgone Accumulator", Dissertation, University of Marburg, West Germany, 1985.

Opfermann-Fuckert, D.: "Reports on Treatments With Orgone Energy: 1. Physical Orgone Therapy with Orgone Energy Accumulator, Orgone Tunnel, Orgone Shooter, and Orgone Blanket. Ten Selected Cases", Unpublished report, presented at Symposia on Wilhelm Reich, Orgonon, Rangeley, Maine, 1986.

Raphael, C.M.: "Confirmation of Orgonomic (Reich) Tests for the Diagnosis of Uterine Cancer", *Orgonomic Med.* II(1):36-41, 1956.

Raphael, C.M. & MacDonald, H.E.: *Orgonomic Diagnosis of Cancer Biopathy*, Wilhelm Reich Foundation, Maine, 1952.

Rosenblum, C.F.: "The Red Shift", *J. Orgonomy*, 4:183-191, 1970.

Rosenblum, C.F.: "The Electroscope - Parts I - IV", *J. Orgonomy*, 3(2):188-197, 1969; 4(1):79-90, 1970; 10(1):57-80, 1976; 11(1):102-109, 1977.

Rosenblum, C.F.: "The Temperature Difference: An Experimental Protocol", *J. Orgonomy*, 6(1):61-71, 1972.

Sharaf, M.: "Priority of Wilhelm Reich's Cancer Findings", *Orgonomic Medicine*, I(2):145-150, 1955.

Seiler, H.: "New Experiments in Thermal Orgonometry", *J. Orgonomy*, 16(2):197-206, 1982.

Silvert, M.: "On the Medical Use of Orgone Energy", *Orgone Energy Bull.*, IV(1):51-54, 1952.

Sobey, V.M.: "Treatment of Pulmonary Tuberculosis with Orgone Energy", *Orgonomic Medicine*, I(2):121-132, 1955.

Sobey, V.M.: "A Case of Rheumatoid Arthritis Treated with Orgone Energy", *Orgonomic Medicine*, II(1):64-69, 1956.

Tropp, S.J.: "The Treatment of a Mediastinal Malignancy with the Orgone Accumulator", *Orgone Energy Bull.*, I(3):100-109, 1949.

Tropp, S.J.: "Orgone Therapy of an Early Breast Cancer", *Orgone Energy Bull.*, II(3):131-138, 1950.

Wevrick, N.: "Physical Orgone Therapy of Diabetes", *Orgone Energy Bull.*, III(2):110-112, 1951.

Introductory Works on Orgonomy, or about Wilhelm Reich:

Baker, E.F.: *Man in the Trap*, Macmillan, NY, 1967.

Bean, O.: *Me and the Orgone*, St. Martin's Press, NY, 1971.

DeMeo, J.: *Bibliography on Orgone Biophysics*, Natural Energy Works, El Cerrito, CA, 1986.

Ollendorff, I.: *Wilhelm Reich, A Personal Biography*, St. Martin's Press, NY, 1969.

Raknes, O.: *Wilhelm Reich and Orgonomy*, St. Martin's Press, NY, 1970.

Reich, P.: *A Book of Dreams*, Harper & Row, NY, 1973.

Sharaf, M.: *Fury on Earth*, St. Martin's-Marek, NY, 1983.

Wyckoff, J.: *Wilhelm Reich, Life Force Explorer*, Fawcett, Greenwich, CT, 1973.

Scholarly Works Addressing the 1950s Campaign Against Reich in the USA:

Blasband, D.: "United States of America v. Wilhelm Reich, Part I", *J. Orgonomy*, 1(1-2):56-130, 1967; "...Part II, the Appeal", *J. Orgonomy*, 2(1):24-67, 1968.

Blasband, R.A.: "An Analysis of the United States Food and Drug Administration's Scientific Evidence Against Wilhelm Reich, Part I, the Biomedical Evidence", *J Orgonomy*, 6(2):207-222, 1972.

DeMeo, J.: "Postscript on the F.D.A's. Experimental Evidence Against Wilhelm Reich", *Pulse of the Planet*, 1(1):18-23, 1989.

Greenfield, J.: *Wilhelm Reich Versus the USA*, W.W. Norton, NY, 1974.

Baker, C.F.: "An Analysis of the United States Food and Drug Administration's Scientific Evidence Against Wilhelm Reich, Part II, the Physical Concepts", *J. Orgonomy*, 6(2):222-231, 1972; "...Part III, Physical Evidence", *J. Orgonomy*, 7(2):234-245, 1973.

Wolfe, T.: *Emotional Plague Versus Orgone Biophysics: The 1947 Campaign*, Orgone Institute Press, NY, 1948.

Works by Other Researchers on Natural Forces Similar to Orgone:

Alfven, H.: *Cosmic Plasmas*, Kluwer, Boston, 1981.

Arp, H., et al: *The Redshift Controversy*, W.A. Benjamin, Reading, MA, 1973.

Arp, H.: *Quasars, Red Shifts, and Controversies*, Interstellar Media, Berkeley, CA, 1987.

Becker, R.O. & Selden, G.: *The Body Electric: Electromagnetism and the Foundation of Life*, Wm. Morrow, NY, 1985.

Bortels, V.H.: "Die hypothetische Wetterstrahlung als vermutliches Agens kosmo-meteoro-biologischer Reaktonen", *Wissenschaftliche Seitschrift der Humboldt-Universitat zu Berlin*, VI:115-124, 1956.

Bortels, V.H.: "Das Gefrieren kleiner Wassermengen als solar-meteoro-biologische Modellreaktion", *Naturwissenschaften*, 52:118, 1965.

Brown, F.A.: "Evidence for External Timing in Biological Clocks", contained in *An Introduction to Biological Rhythms*, J. Palmer, ed., Academic Press, NY, 1975.

Burr, H.S.: *Blueprint For Immortality*, Neville Spearman, London, 1971; *The Fields of Life*, Ballantine Books, NY, 1972.

Cope, F.W.: "Magnetic Monopole Currents in Flowing Water Detected Experimentally...", *Physiological Chemistry & Physics*, 12:21-29, 1980.

DeMeo, J.: "Evidence for the Existence of a Principle of Atmospheric Continuity", contained in *Preliminary Analysis of Changes in Kansas Weather Coincidental to Experimental Operations with a Reich Cloudbuster*, M.A. Thesis, University of Kansas, 1979; reprinted by Natural Energy Works, El Cerrito, CA, 1989.

Dewey, E.R., ed.: *Cycles, Mysterious Forces that Trigger Events*, Hawthorn Books, NY, 1971.

Dudley, H.C.: *Morality of Nuclear Planning*, Kronos Press, Glassboro, NJ, 1976.

Kervran, L.C.: *Biological Transmutations*, Beekman Press, Woodstock, NY, 1980.

Miller, D.: "The Ether-Drift Experiment and the Determination of the Absolute Motion of the Earth", *Reviews of Modern Physics*, 5:203-242, July, 1933.

Moss, T.: *The Body Electric, A Personal Journey into the Mysteries of Parapsychological Research, Bioenergy, and Kirlian Photography*, J. P. Tarcher, Los Angeles, 1979.

Nordenstrom, B.: *Biologically Closed Electric Circuits: Clinical, Experimental, and Theoretical Evidence for an Additional Circulatory System*, Nordic Medical Press, Stockholm, 1983.

Ott, J.: *Health and Light*, Devin Adair, Old Greenwich, CT, 1973.

Piccardi, G.: *Chemical Basis of Medical Climatology*, Charles Thomas, Springfield, IL, 1962.

Ravitz, L.J.: "History, Measurement, and Applicability of Periodic Changes in the Electromagnetic Field in Health and Disease", *Annals, NY Acad. Sciences*, 98:1144-1201, 1962.

Sheldrake, R.: *A New Science of Life: The Hypothesis of Causative Formation*, J.P. Tarcher, Los Angeles, 1981.

Nuclear and Electromagnetic Radiation Health Hazards:

Ahlbom, A. et al.: *Biological Effects of Power Line Fields*, NY State Power Lines Project, State of NY, Dept. of Health, 1987.

Bertell, R.: *No Immediate Danger, Prognosis for a Radioactive Earth*, Book Publishing Co., Summertown, TN, 1985.

Brodeur, P.: *The Zapping of America: Microwaves, Their Deadly Risk, and the Cover-Up*, W.W. Norton, NY, 1977.

Brodeur, P.: *Currents of Death: Power Lines, Computer Terminals, and the Attempt to Cover Up their Threat to your Health*, Simon & Schuster, NY, 1989.

Chaitow, L. & Kutter, E.: *How to Live With Low Level Radiation*, Healing Arts Press (One Park St., Rochester, VT 05767), 1988.

Curtis, R. & Hogan E.: *Perils of the Peaceful Atom*, Ballantine, NY 1969.

Dadd, D.L.: *Nontoxic and Natural, Guide for Consumers*, J.P. Tarcher, Los Angeles, 1984.

Dadd, D.L.: *The Nontoxic Home, Protecting Yourself and Your Family*, J.P. Tarcher, Los Angeles, 1986.

Fuller, J.G.: *The Day We Bombed Utah, America's Most Lethal Secret*, New American Library, NY, 1984.

Fuller, J.G.: *We Almost Lost Detroit*, Ballantine, NY, 1975.

Gofman, J.W.: *Radiation and Human Health*, Pantheon, NY, 1983.

Gofman, J.W. & Tamplin, A.R.: *Poisoned Power*, Rodale, Emmaus, PA 1971.

Lee, L.: *Radiation Protection Manual*, 2061 Hampton Ave. Redwood City, CA 94061.

Metzger, H.P.: *The Atomic Establishment*, Simon & Schuster, NY, 1972.

Miller, R.L.: *Under the Cloud, the Decades of Nuclear Testing*, Free Press, NY, 1986.

Pringle, P. & Spigelman, J.: *The Nuclear Barons*, Michael Joseph, London, 1981.

Santilli, R.: *Il Grande Grido: Ethical Probe on Einstein's Followers in the USA*, Alpha Publishing (897 Washington St., #82, Newtonville, MA 02160), 1984.

Sternglass, E.: *Secret Fallout*, McGraw Hill, NY, 1986.

Wassermann, H.: *Killing Our Own*, Doubleday, NY, 1985.

Webb, T. et al.: *Food Irradiation, Who Wants It?*, Thorsons Publishers, Rochester, VT, 1987.

Zamm, A.V.: *Why Your House May Endanger Your Health*, Touchstone, NY, 1980.

Alternative Therapies for Degenerative Disease, Repression of New Medical Findings, Hazards of Orthodox Medicine

Abehsera, M.: *Healing Clay: Amazing Cures from the Earth Itself*, DeVorss, NY, 1979

Bailar, J. C. and Smith, E. M.: "Progress Against Cancer?", *New England Journal of Medicine*, 314: 1226-1232 (May 8), 1986.

Bailey, H.: *A Matter of Life or Death*, G.P. Putnam's Sons, NY 1958.

Boyd, P.: *The Silent Wound, A Startling Report on Breast Cancer and Sexuality*, Addison-Wesley, Reading, MA, 1984.

Bradford, R.W., et al.: *Complaint Against Medical Tyranny As Practiced in the United States of America: American Medical Genocide*, Committee for Freedom of Choice in Medicine, 146 Main St., #408, Los Altos, CA 94022, 1984.

Bradshaw, J.S.: *Doctors on Trial*, Paddington Press, NY, 1978.

Corea, G.: *The Hidden Malpractice, How American Medicine Mistreats Women*, Harcourt, Brace & Jovanovich, NY, 1977.

Epstein, S.S.: *The Politics of Cancer*, Sierra Club Press, San Francisco, 1981.

Gerson, M.: *A Cancer Therapy: Results of 50 Cases*, Gerson Institute (PO Box 430, Bonita, CA 92002), 1977.

Glassman, J.: *Cancer Survivors: And How They Did It*, Doubleday, NY, 1980.

Haught, S.J.: *Cancer? Think Curable! The Gerson Therapy*, Gerson Institute (PO Box 430, Bonita, CA 92002), 1962.

Kelley, W.D.: *One Answer to Cancer: An Ecological Approach to the Successful Treatment of Malignancy*, Kelley Foundation, 1969.

Kittler, G.D.: *Laetrile, Control for Cancer*, Warner Library, NY, 1963.

Kushi, M. & Kushi, A.: *Cancer and Heart Disease, The Macrobiotic Approach to Degenerative Disorders*, Kodansha, Boston, 1978.

LeShan, L.: *You Can Fight for your Life: Emotional Factors in the Causation of Cancer*, M. Evans, NY, 1977.

Livingston-Wheeler, V.: *The Conquest of Cancer: Vaccines and Diet*, Franklin Watts, NY, 1984.

Lynes, B. & Crane, J.: *The Cancer Cure That Worked: Fifty Years of Suppression*, Marcus Books, 75 Tidefall Dr., Toronto, Canada, M1W 1J1.

Martin, W.: *Medical Heroes and Heretics*, Devin-Adair, Old Greenwich, CT, 1977.

Mendelsohn, R.S.: *How to Raise a Healthy Child, In Spite of Your Doctor*, Contemporary Books, Chicago, 1984.

Mendelsohn, R.S.: *Confessions of a Medical Heretic*, Contemporary Books, Chicago, 1979.

Newbold, H.L.: *Vitamin C Against Cancer*, Stein & Day, NY, 1982.

Natenberg, M.: *The Cancer Blackout: A History of Denied and Suppressed Remedies*, Regent House, Chicago, 1959.

Pauling, L: *Vitamin C, The Common Cold, and the Flu*, Bantam Books, NY, 1975.

Passwater, R.: *Cancer and Its Nutritional Therapies*, Keats, NY, 1980.

Salaman, M.: *Nutrition: The Cancer Answer*, Stratford, NY, 1984.

Tushnet, L.: *The Medicine Men, The Myth of Quality Medical Care in America Today*, St. Martin's Press, NY 1971.

Tropp, J.: *Cancer, A Healing Crisis: The Whole Body Approach to Cancer Therapy*, DeVorss, NY, 1978.

Wohl, S.: *The Medical Industrial Complex*, Harmony Bks, NY, 1984.

INFORMATION SOURCES

On Wilhelm Reich and Orgonomy:

Orgone Biophysical Research Lab, *Greensprings Research and Educational Center*: PO Box 1148, Ashland, Oregon 97520 USA. Tel/Fax: 541/ 552-0118. Email: demeo@mind.net Internet: http://id.mind.net/community/orgonelab/index.htm
- Publishes *Pulse of the Planet* journal, books and *Special Reports*. Public educational lectures & seminars on Wilhelm Reich & orgonomy. Active program of orgonomic atmospheric & environmental research.

Natural Energy Works: PO Box 1148, Ashland, Oregon 97520 USA Tel/Fax: 541/ 552-0118. Email: demeo@mind.net
- Mail order sales of books, products, accumulator construction supplies, research instruments, radiation detection meters.

Institute for Orgonomic Science: PO Box 304, Gwynedd Valley, PA 19437 Tel: 410/ 732-1200
- Publishes *Annals of the Institute for Orgonomic Science*. Active program of orgonomic research. Training for orgone therapists.

The Wilhelm Reich Museum: PO Box 687, Rangeley, Maine 04970, Tel: 207/ 864-3443. Email: wreich@rangeley.org
- Preserves Wilhelm Reich's home and laboratory (called Orgonon) for public viewing and tours. Publishes a *Newsletter*, and *Orgonomic Functionalism*. Sells xerox copies of various out-of-print books, journals, and pamphlets by Wilhelm Reich. Occasional seminars and symposia.

The American College of Orgonomy: PO Box 490, Princeton, NJ 08542. Tel. 908/ 821-1144.
- Publishes the *Journal of Orgonomy*. Occasional public symposia, seminars and workshops. Training for orgone therapists.

Center for Functional Research: 2175 Mar East, Tiburon, CA 94920. Fax 415/ 435-2815. Email: CFR@FunctionalResearch.org
- Active program of orgonomic and related consciousness research. Occasional public symposia. Training for orgone therapists.

Center for Orgonomic Studies: 6823 NE 29th Ave., Seattle, WA 98115. Tel: 206/ 523-0800 Email: orgstudies@aol.com
- Training for orgone therapists. Occasional public symposia.

Orgonics: Email: Orgonics@aol.com www.orgonics.com
- Sells quality experimental orgone blankets, seed-chargers & full-size accumulators.

Foreign Research Journals on Orgonomy:

Bioenergy, So.37-1 Takomachi, Kanazawa 921, Japan.
Emotion, Karlsburgallee 25-E, Berlin 14089, Germany.
Energia, Caracter y Sociedad, Publicationes Orgon, C/.Serpis 36-34a, Valencia 46022 Spain.
Greek Journal of Orgonomy, Hellenic Orgonomic Association, T.O. 18251, Thessaloniki 54008, Greece.
Lebensenergie, Zentrum fur Orgonomie, Im Braunlesrot 20, Waldbrun 69429, Germany.
Revista de Ciencias Orgonomicas, Fundacion Wilhelm Reich, C/.San Lazaro 2, Figueres, Gerona 17600, Spain.
Sciences Orgonomiques, "La Rose des Sables", Alle due Chene Vert, Parc Liserb, 06000 Nice, France.

Of Related Interest: *Summerhill School*, Leiston, Suffolk, IP16 4HY United Kingdom - Democratically operated school founded by A.S. Neill, educator and pioneer on self-regulation for children. Publishes *Friends of the Summerhill Trust Journal.*

Health Freedom Organizations and Alternative Cancer Treatment Clinics and Institutes:

Cancer Control Society, and *Cancer Control Journal*:
2043 N. Berendo St. Los Angeles, CA 90027; (213) 663-7801
- Maintains a complete listing of clinics and medical practitioners in the USA and abroad, who use non-toxic, alternative treatments for degenerative disease
- Sells books on alternative health topics
- Annual conventions on alternative health topics

Cancer Control Society Clinic Tours and Travel Services:
(Mexico): PO Box 4651, Modesto, CA 95352; (209) 529-4697

Gerson Institute, Charlotte Gerson, Clinic in Mexico:
US Contact: PO Box 430, Bonita, CA 92002; (619) 267-1150

Hoxsey Herbal Therapy, Bio-Medical Center:
PO Box 727, 615 General Ferreira (Colonia Juarez)
Tijuana, B.C., Mexico; (706) 684-9011, 684-9132

Linus Pauling Institute:
440 Page Mill Rd., Palo Alto, CA 94306; (415) 327-4064

Livingstone Immunology Clinic:
3232 Duke St., San Diego, CA 92110; (619) 224-3515

National Health Federation, and *Health Freedom News*:
PO Box 688, Monrovia, CA 91016; (818) 357-2181
- Annual conferences on alternative health topics
- Book sales on alternative health topics
- Works to change repressive health laws

Project Cure: Center for Alternative Cancer Research:
1101 Conn. Ave., NW #403, Washington, DC 20036 (800) 552-2873
- Works to change repressive health laws
- Publishes a newsletter

Appendix:

ABSTRACT

The Psycho-Physiological Effects of the Reich Orgone Energy Accumulator

Dissertation, University of Marburg,
West Germany, 1986.

by Stefan Müschenich, Dipl-Psych. and
Rainer Gebauer, Dipl-Psych.

This dissertation contains an experimental investigation of the (psycho-) physiological changes in volunteer test subjects during sessions in a Reich orgone accumulator, conducted at the University of Marburg, West Germany.

Proceeding from clinical observations as well as from further research work in the biological and physical fields, the Austrian-born physician and psychotherapist Wilhelm Reich postulated a specific bioenergy manifesting itself in the living organism, and called it *orgone*. This hypothesis was based on extensive biophysical studies conducted by Reich between 1934 and 1957 at the universities of Oslo and New York, and in his own research laboratories. Steps in this development were experiments on changes in endosomatic skin potential due to certain stimuli, the description of energy-carrying vesicle-like structures *(bions)* during the microscopic observation of disintegrating biological slides, and investigations of hitherto unexplained atmospheric energy phenomena that play a part in forming weather conditions. Finally Reich claimed that the orgone was a universally existing kind of energy and he attributed to it a specific (psycho-) physiological effectiveness on the human organism.

About 1940 Reich published the construction plans for an apparatus that was able to concentrate this energy within its interior. He postulated that spending some time in this *orgone accumulator* (ORAC) produced certain psychic and somatic reactions. Later he used the device for therapeutic purposes in the treatment of some syndromes. An orgone accumulator may be described as a closet-like structure or box, each of its walls consisting of a number of alternate layers of organic material (for the exterior) and metal (for the interior).

In the theoretical part of our dissertation we first discussed a choice of empirical studies published by various scientists who dealt with the physical aspects of the atmosphere within the device. The phenomenon of a constantly increased air temperature in the accumulator (the To-T effect), the observation of a delayed electroscopical discharge rate, and processes connected with alterations in air humidity and water evaporation rate are of special interest in this context. Furthermore some medical case histories were demonstrated to illustrate the effects of the orgone accumulator that were conducive to health during the therapy of various diseases. The studies mentioned were described in their topical and historical connection with Reich's research work and concepts. In doing so, we critically discussed the scientific validity of the theories involved.

The main effort of our own experimental work was to investigate the (psycho-) physiological effects that are attributed to orgone accumulator sessions. Reich claimed that body temperature rose during ORAC sessions and he described a general vagotonic activation due to sitting in the device for a certain period of time. Considering the contents of orgonomic publications, and the results of our own pilot tests, we decided to explore the systematic changes in body core temperature, skin temperature, and heart rate (ECG). As far as we know, the two last mentioned parameters have never before been evaluated in a scientific manner.

We conducted a long-term study with 15 volunteer subjects, each of them carrying through 20 experimental hours. In this experiment the physiological variables mentioned above were continuously recorded during the sessions with the help of electronic devices. Ten persons undertook ten 30-minute sessions in an eight-fold coated orgone accumulator built according to Reich's instructions. Besides, they carried through ten 30-minute sessions in an almost identical-looking control box. This control box had been constructed by us for purposes of comparison. It consisted only of organic matter (and no metals), but concerning its size, shape, and insulating properties it did not differ from the original accumulator box (which did contain metals). Five additional subjects conducted all their 20 sessions in the same box: three persons used the orgone accumulator every time, while two subjects used only the control. Before each session the subjects sat in a comfortable relaxation-chair for 15 minutes. By this provision we wanted to establish a standardized psychophysiological initial level. During this space of time, the physiological data mentioned were measured.

Firstly, by this means a comparison between the physiological reaction patterns in the orgone accumulator and in the control box was made possible. Secondly, one could relate the data recorded during the sessions in one of the two boxes to the activation standard previously measured in the relaxation-seat. The entire investigation was conducted as a "double-blind study", which means that neither the volunteer subjects (who had been chosen at random) nor the persons that gave the instructions and recorded the data knew anything about the experiment they were taking part in. The clothing of the subjects, the positions of the two boxes, the sequence in which the two devices were used, and the other experimental modalities were standardized or balanced out. In contrast for example to the medical case histories mentioned above, we took into consideration psychological sources of error, and artifacts caused by superimposition effects. Intending to control these factors, we had the persons fill out a questionnaire after each session, which revealed information about their psychophysiological sensations and moods during their stay in the boxes. The exploration, as well as the fact that the subjects were absolutely uninformed, was to eliminate falsifications produced by suggestion (e.g.

caused by manipulation through the conductors of the study) or autosuggestive factors.

In this way we evaluated the subjective quality of potential physiological changes. Additionally, the recorded ECG data conveyed information about the psychic and emotional excitement of the participating persons. Besides, meteorological and physical variables were continuously measured. They were to be correlated with the psychophysiological data.

The investigation gave the following results: All of the (*a-priori* formulated) hypotheses, which claimed that there were no greater psychophysiological alterations between relaxation-chair and orgone accumulator than between relaxation-chair and control box, could be rejected with a statistical significance on the 1% level. This means that one may proceed with a probability of 99% on the assumption that, compared with the initial standard during the stay in the relaxation-chair, the physiological data recorded in the orgone accumulator were subject to greater alterations than the data recorded in the control.

Both of the body temperature variables showed a distinct increase during the accumulator sessions. These facts are in accordance with Reich's predications of a rise in core temperature and an increase in parasympathetic activation produced by orgone accumulator treatment. The interpretation of heart rate pattern, however, is more difficult. While one would expect a decrease, the heartbeat frequency revealed a clear increase between the initial level and the stay in the accumulator. These problems may probably result from the fact that the ECG-data showed a relatively high statistical variance and were much more sensitive to accidental external influences and momentary psychic states than the more stable temperature parameters. Further follow-up studies are to explore to what extent a possible vagotonic ECG-effect may have been superimposed by cognitive processes, nervousness, or anxiety. Still, we can conclude that the assumption of vegetative changes during orgone accumulator sessions was strongly confirmed by our data. The impression one gets from the results of the first 10 subjects was corroborated by the data of those five persons who used only the accumulator or only the control every time. The number of these last-mentioned cases, however, is too small for a statistical test of significance.

Some further interesting effects are worth mentioning, too: for example, one person seemed to be "resistant" to the orgone accumulator effects, while another proved to respond extremely sensitively to the accumulator. Not only the objective physiological results but the subjective sensations (expressed in the questionnaire) of these two persons justified this hypothesis. Besides, the "ORAC-resistant" subject was the only one who said that he felt better in the control than in the orgone accumulator. All the other persons preferred the accumulator. Generally it can be said that the questionnaire data corresponded well to the recorded physiological changes. In the accumulator box the volunteer subjects noticed more perceptions of warmth, prickling and tickling on the skin surface; additionally, they connected more pleasant cognitive associations with the orgone accumulator than with the control.

In our study a correlation between psychophysiological alterations and meteorological/physical processes was corroborated. During late spring and early summer months the somatic reactions were stronger than during the colder period. Especially the air pressure outside the building seems to be quite a good predictor for the physiological pattern in the accumulator. The phenomenon of a constantly positive To-T difference was statistically confirmed. The total of the meteorological data revealed that the air temperature measured in the orgone accumulator differs from that recorded in the control on the 1% level.

As a conclusion it can be stated that the results received in our investigation furnish evidence for the assumption that the physical properties of the orgone accumulator and its psychophysiological efficacy on human organisms, postulated by Reich and his associates, factually exist. Various, more natural-scientific oriented follow-up investigations are expected to determine whether the hypothesis is justified that a hitherto unknown biophysical energy (the orgone) is the causative factor of the phenomena described. Additionally, a more extensive study might clarify one or other effect.

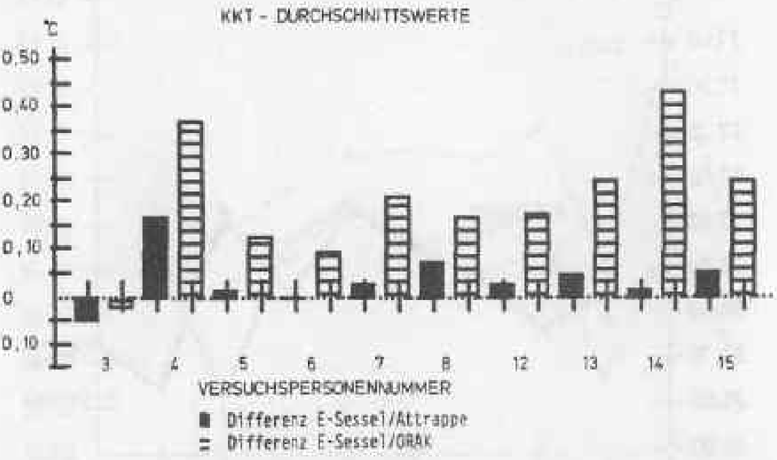

Figure 1. Average Change in Body Core Temperatures, for both control box and orgone accumulator, among different volunteer subjects participating in the study (identified by number).

Black = Change in body core temperature between relaxation-chair and control box session
White = Change in body core temperature between relaxation-chair and orgone accumulator session

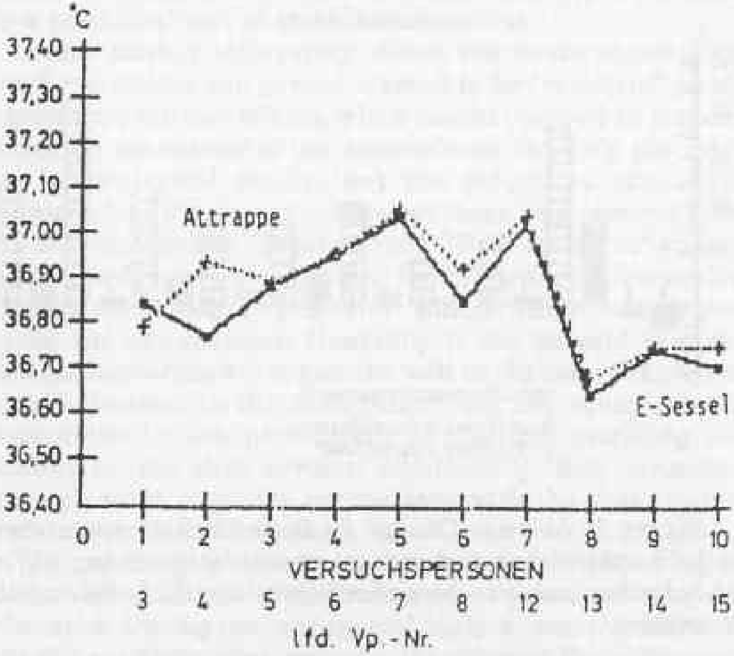

Figure 2. Changes in Body Core Temperatures of Subjects Moving From Relaxation-Chair into Control Box. Note the absence of any significant difference in temperatures (except for subject #4).

Appendix Author Contact Addresses:

Stefan Müschenich, Dr.med. & Diplom-Psychologe,
Hohenzollernstr. 11, 79106 Freiburg i.B., Germany

Rainer Gebauer, Diplom-Psychologe, Steinweg 5,
35287 Amoneburg, ER Furtshausen, Germany

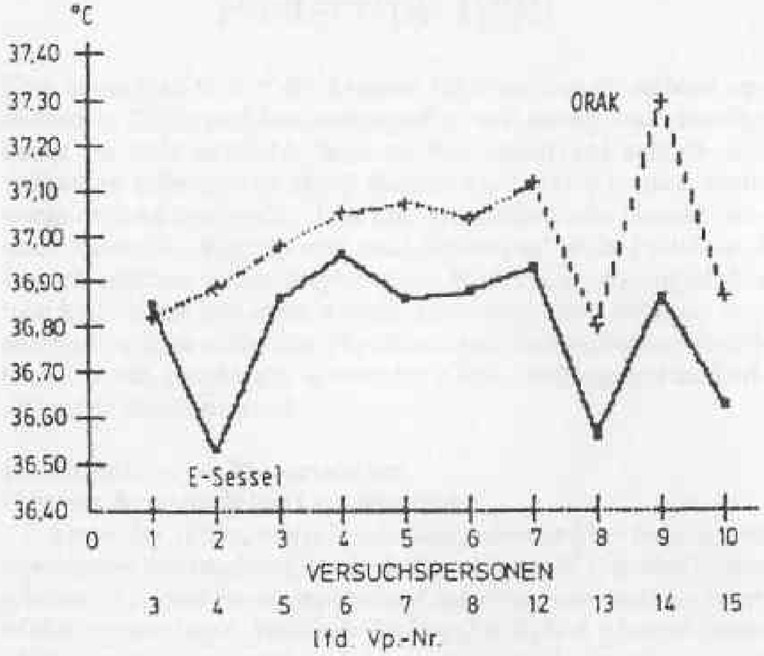

Figure 3. Changes in Body Core Temperatures of Subjects Moving From Relaxation-Chair into Orgone Accumulator. Significant temperature differences are now quite apparent (especially subjects #4 and #14).

Copyright © 1986, All Rights Reserved by Stefan Müschenich and Rainer Gebauer, Germany. Full text published as *Der Reichsche Orgonakkumulator, Naturwissenschaftliche Diskussion, Praktische Anwendung, Expoerimentelle Untersuchung,* Nexus Press, Frankfurt, 1989.

Postscript 1999

This second edition of the *Orgone Accumulator Handbook* appeared in 1989, and has continued to sell slowly but steadily, being the only available book on this important subject and spreading information about Reich's discoveries to new audiences around the world. It is now translated into Greek, German, Spanish, Portuguese, and Japanese, with Polish and French editions under preparation. With the spreading of this new knowledge has come a lively communication between myself and various scientists, physicians and experimenters in different lands, pointing to a need for a few clarifications and additions to the *Handbook*.

Clarification on Materials for Orgone Accumulator Construction

Since the 1940s, when Reich first published his findings on the orgone accumulator, he and others (myself included) have advocated *celotex* for construction of the outer, non-metallic layer of the accumulator. However, "celotex" is in fact a brand name of the *Celotex company*, and does not today designate any one specific product. Originally, the Celotex company made only organically-based *soundboard* material, which is formed from the crushed and pulverized stalks of sugar cane and other herbaceous farm plant residues. The crushed organic material is mixed with binders and glues, pressed into a flat sheet to dry, and then painted white on one surface. Such *soundboard* material continues to be available from many different sources. But the term "celotex" has lost its original meaning. Celotex company today makes a number of rigid insulating panels which are totally unacceptable and toxic in the construction of accumulators, such as an aluminum-foil and foam insulation board.

Another excellent form of *masonboard* is now available for the exterior of the accumulator, called Medite from the Medite Corporation. Medite (or similar material from other manufacturers) is also made from cellulose plant or woody materials ground into very tiny particles, mixed with binders and glues and pressed into thin flat sheets. It is a dense, hard material, not like the often crumbly soundboard, and therefore will last much longer. Medite can also be ordered in a special formaldehyde-free composition. Whether one uses the original Celotex soundboard, or Medite board, it is strongly advised to give the exterior surface several heavy coats of natural shellac to stabilize it, make it waterproof, and add to the orgonotic attraction.

It is also now possible to obtain *carded sheep's wool* at sufficiently low cost as to substitute for the fiberglass normally used inside accumulator panels. When wool is sheared from a sheep, it is gently washed and combed to remove debris, yielding a light and fluffy material called carded wool, which is later processed into yarn or thread for the ultimate making of wool fabrics. Carded wool can be pulled and teased into thin layers as needed for use within orgone accumulator panels, and has no dusts or toxic qualities whatsoever. Newer biomedical findings suggest fiberglass is much more toxic to breathe and handle than previously believed, and so it is all around a better idea to go totally with the more natural materials, such as the carded wool. It can be purchased in bulk from wholesale wool products suppliers, or by mail-order from Natural Energy Works (PO Box 1148, Ashland, Oregon 97520). Fiberglass and some kinds of plastics (see page 43 and below) can still be used to make excellent strong accumulators, but the wool is far easier and more pleasant to handle.

We can also add the following new information regarding why some materials work well for the exterior non-metal layer, while others not. In general, those materials with a high *dielectric constant* (such as sheep's wool with its organic oils, certain plastics, acrylics, fiberglass, shellac, beeswax, etc.) are also good orgone absorbers. Several chemists have also pointed out that petrochemical synthetics which make good orgone absorbers (acrylics, styrene) contain abundant oxygen and water molecules imbedded in their molecular structures. Whether these preliminary ideas hold true universally for a variety of other materials remains to be seen, but they are worthwhile points for further investigation.

All other points made in the Handbook regarding good and poor materials for orgone accumulator construction remain unchanged and time-tested.

Cautionary Note on New "Wireless" Electromagnetic Gadgets

On the subject of safe use of the accumulator (see Part II), we may add the following: Over the last decade, there has been an increased use of new "high tech" devices for telecommunications purposes which actively emit potentially toxic low-levels of electromagnetic radiation: cellular telephones and cell-phone relay systems (often located in populated areas) emit microwave radiations at potentially toxic levels, as well as new remote-control apparatus which employ *active far infrared (close to microwave) frequencies*, such as inventory scanners and various kinds of "wireless" technology. In all cases, orgone accumulators should not be located near to any of these devices, nor should any human being wish to live, work or sleep near them either! New electromagnetic devices should be viewed as contrary for safe use of orgone devices unless proven otherwise. On the other hand, various *passive* infra-red sensing devices (such as automated outdoor security lights) pose no problems.

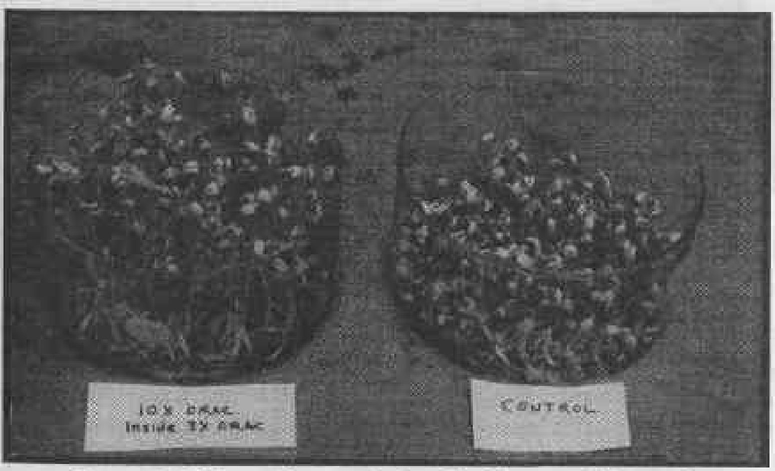

Sprouted Mung Beans: Left side, Orgone-Charged. Right side, Controls. (Experimental protocol p.96-97)

New Publications and Controlled Studies on the Human Bio-Effects of the Orgone Energy Accumulator

In recent years, mainstream publishers in Germany have expressed a growing interest in the works of Wilhelm Reich, and on the subject of his research findings (an interest not currently shared by any mainstream American publisher). Some examples of this interest are given below.

Appendix I of this *Handbook* (page 148-155) presents an extended abstract describing the first double-blind controlled study of human physiological response to the orgone accumulator, as undertaken by Stefan Müschenich and Rainer Gebauer. Their study was completed in 1986 at the University of Marburg in Germany. Dr. Müschenich subsequently wrote a major book on this same subject (in the German language):

Stefan Müschenich, *Der Gesundheitsbegriff im Werk des Arztes Wilhelm Reich (The Concept of Health in the Works of Dr. Wilhelm Reich)*, Doktorarbeit am Fachbereich Humanmedizin der Philipps-Universität Marburg, Verlag Görich & Weiershäuser, Marburg 1995, 425 pages. ISBN 3-922906-54-0.

In more recent years, another similar double-blind and controlled study was undertaken at the University of Vienna by Günter Hebenstreit, producing similar positive results. The Hebenstreit study, like that of Müschenich and Gebauer, used an orgone accumulator, a control dummy-box that looked like an accumulator, and a relaxation-chair in which test subjects were evaluated both before and after exposure to either the accumulator or control enclosure. It produced positive results, I am informed. Unfortunately, we do not have any printed protocols, but can give the exact citation for those who wish to obtain and read the original German-language publication:

Günter Hebenstreit, *Der Orgonakkumulator Nach Wilhelm Reich. Eine Experimentelle Untersuchung zur Spannungs-Ladungs-Formel*, Diplomarbeit zur Erlangung des Magistergrades der Philosophie an der Grung- und Integrativwissenschaftlichen Fakultät der Universität Wien, 1995.

Another German-language title of significance was published in 1998, containing the contributed research papers of

40 different scientific-professional authors on the subject of Wilhelm Reich's social and biophysical findings, to include many papers on the subject of bions, orgone energy, and the orgone energy accumulator:

James DeMeo & Bernd Senf, Editors, *Nach Reich: Neue Forschungen zur Orgonomie: Sexualökonomie, Die Entdeckung Der Orgonenergie (After Reich: New Research in Orgonomy: Sex-Economy, Discovery of the Orgone Energy)*, Zweitausendeins Verlag, Frankfurt, 1998. 898 pp. ISBN 3-86150-239-9.

An Orgone Energy Darkroom (room-sized accumulator) at the Orgone Biophysical Research Lab's high-altitude Greensprings Center near Ashland, Oregon. This structure is made in the form of a large accumulator, with a metal-lined interior and under-layering of fiberglass. Inside are several human-sized orgone accumulators and experimental chargers. The strength of the atmospheric orgone energy charge inside such an orgone room is dramatically increased. The orgone energy can be more easily felt and observed with the eye, and registers more clearly in experiments and with measuring apparatus.

About the Orgone Biophysical Research Lab (OBRL) and Greensprings Center

The Orgone Biophysical Research Lab is a non-profit science research and educational foundation, established in 1978. Over the years, OBRL has supported various laboratory and field research projects, educational lectures and seminars in both the USA and overseas, focused upon the sex-economic and orgone biophysical discoveries of the late, great natural scientist, Wilhelm Reich. OBRL also publishes a journal, *Pulse of the Planet*, on an irregular basis. The founder and director of OBRL, James DeMeo, has been investigating and extending Dr. Reich's original findings in both the social and natural sciences since 1971. Early in 1995, the *Greensprings Center* was established in a vibrant forested region of rural Oregon, to provide a better foundation for continued public outreach and in-depth study of orgone energy functions in nature. The *Center* also has the West Coast's only *Orgone Energy Darkroom*, for experiments requiring a higher energy charge, and for direct visual observation of orgone energy functions. New books (described at right) are also occasionally published, and public educational programs, the *Greensprings Seminars*, are held each summer. Accomplished scholars in the field of orgonomy are invited each year to share their knowledge and experience with interested students and professionals from around the world. Yearly seminars include: *Bions, Biogenesis and the Reich Blood Test, The Orgone Energy Accumulator,* and *Guided Independent Study*. Contact OBRL for an informative brochure and publications catalog, and to get on the international mailing list for seminar announcements.

Orgone Biophysical Research Lab
Greensprings Center, PO Box 1148
Ashland, Oregon 97520 USA
Tel/Fax: 541-552-0118 Email: demeo@mind.net
OBRL's Global Internet Web Site:
http://www.orgonelab.org

Additional Publications Available

(Contact OBRL for ordering information)

Heretic's Notebook:

Heretic's Notebook (Pulse of the Planet #5): Emotions, Protocells, Ether-Drift and Cosmic Life-Energy, with New Research Supporting Wilhelm Reich, Edited by James DeMeo

Contains 28 insightful essays and research articles by 17 different authors, on natural childbirth, sexuality, archaeology of early human violence, Reich's orgonomic functionalism, exposés on Reich's detractors, Giordano Bruno's work, bion-biogenesis research, Dayton Miller's ether-drift discoveries, emotional effects in REG (psychokinesis) experiments, new detector for orgone energy, dowsing research, cloudbusting desert-greening experiments in Africa, plant growth stimulation in the orgone accumulator, the orgone energy motor and "free energy", plus UFO research, book reviews, and much more, with color cover photos, text- photos and illustrations.
272 pages, Large format $24 Softcover

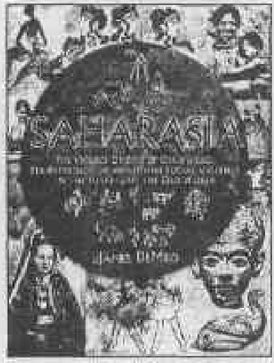

SAHARASIA: The 4000 BCE Origins of Child-Abuse, Sex-Repression, Warfare and Social Violence, In the Deserts of the Old World,
by James DeMeo

Dr. DeMeo's *magnum opus* on the origins of human violence and biophysical armoring, the first geographical, cross-cultural study of human behavior around the world, using Wilhelm Reich's sex-economic discoveries as a basic starting point, presenting world

maps of different behaviors and social institutions. Source-regions (Arabia and Central Asia) for patriarchal authoritarian culture were identified, and migratory-diffusion patterns were traced, back in time, to pinpoint where and how the human tragedy began. Solves the riddle of the origins of human violence and armoring. A breakthrough in the scientific study of human sexuality, psychology and anthropology, and must-reading for every parent, student, professor and clinical worker in the field of human health and behavior. 454 pages, with over 100 maps, photos, and illustrations. Large format with vivid full-color cover, extensive bibliography and index. Softcover $34
Author-Signed Limited Hardbound First Edition $90

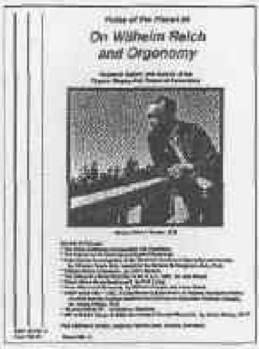

On Wilhelm Reich and Orgonomy (Pulse of the Planet #4)

Edited by James DeMeo

Contains Reich's milestone articles on psychic and somatic (mind-body) processes, and on the bioelectrical aspects of human emotion and sexuality, with articles by R.D. Laing on Reich, and a discussion on Reich's work in Denmark when he fled Nazi terror in Germany and was also expelled from the International Psychoanalytic Association. Other papers discuss: Reich's research on biogenesis and discovery of the microscopical *bion*; the discovery of orgone (life) energy, and the orgone energy accumulator. Also featured are articles about the Food & Drug Administration's attack upon Reich, and their present-day war against the natural health movement; the deadly effects from nuclear power plants, and an illuminating scientific challenge to the HIV theory of AIDS — plus other reports on current life energy research, weather anomalies from nuclear bomb tests, a cloudbusting desert-greening experiment in Israel, provocative book reviews, and more! 176 pp. Large format $20 Softcover.

PLANS

FOR CONSTRUCTING

AN

ORGONE ENERGY ACCUMULATOR

Based on information originally distributed by the Wilhelm Reich Foundation

Notes Regarding the Use of the Orgone Accumulator

The Orgone Accumulator is built so as to collect the natural energy of the atmosphere (Orgone Energy) and to make this energy about five times more concentrated inside the accumulator than in the outside air. Since orgone is the same energy that is in the bodies of all living organisms, the accumulator has the effect of increasing the body energy of the user--increasing the natural charge of the user's tissue and blood--thus strengthening the user's resistance to disease.

Orgone is not electricity, and the orgone accumulator has no electrical connections. WARNING: DO NOT ALLOW EXPOSED ELECTRICAL WIRES TO COME IN CONTACT WITH THE ACCUMULATOR!

The accumulator should be used once or twice daily as follows: the user should sit inside the accumulator lightly clothed or without clothing until the user becomes quite warm, or for roughly 15 to 45 minutes. The effect of the orgone will be best if the user relaxes, breathing out deeply but not forcefully. Slight dizziness or heaviness in the head is a signal to come out. Fresh air will eliminate such effects immediately.

In humid weather, the normal time of use should be lengthened, as there is then less orgone in the atmosphere. Air the room in which the accumulator is kept every day, for the increased orgone concentration tends to make the air in the room heavy.

The funnel end from the shooter box-seat is used for local application of orgone. This unattached end should be held close to, but not quite touching, the region in question for between two and ten minutes, depending on the sensitivity of the region. If it is used on the eyes, they should be kept closed. A slight burning sensation in the region that is being treated is a signal that the body has had enough for the time being. In general, the tube should only be used until the area becomes warm. The shooter box-seat tube can be used several times a day.

© Copyright 1991 by Mary Boyd Higgins as Trustee of the Wilhelm Reich Infant Trust Fund

Questions Regarding the Use of the Orgone Accumulator

Q. Does it lessen the effect of the accumulator if any part of the body comes in direct contact with the metal walls?

A. It is better if no part of the body touches the metal walls. However this cannot always be avoided and it does not lessen the effect to any considerable degree.

Q. If two or three people use an accumulator rapidly in succession, is the orgone "used up"; does the last one to use it get energy?

A. The effect of the orgone accumulator consists of a mutual excitation of the orgone energy which is concentrated within the accumulator and the body-orgone energy of the user. It can be assumed that the successive use of the accumulator by several persons will not lessen the efficacy.

Q. Should you spend more or less time in the accumulator, depending on the humidity?

A. Since the orgone energy concentration in the atmosphere diminishes with high humidity, it is advisable to prolong the use of the accumulator on very humid days.

Q. Does the funnel for the pipe of the tube box have to be metal?

A. Yes, since plastic or other non-metallic material will tend to "hold" orgone energy, while the metal will give it off immediately.

Q. Can you use the tube box inside a larger accumulator for reasons of more intensity, or is it more advisable to use it outside?

A. The concentration of orgone energy in the tube box will be stronger if it is used within the larger accumulator.

Q. Is it all right to take objects into the accumulator while using it, such as books, magazines, notebook, pen, pillow?

A. There is no objection to taking reading or writing material into the accumulator or knitting or needlework, etc. We would advise against taking pillows into the accumulator because they will soak up orgone energy and thus divert it from the body.

Q. Is it all right to paint the accumulator to match walls?

A. As long as the paint does not contain lead or other metals, the accumulator can be painted in any color. It may also be covered with wallpaper.

Q. If one makes one's accumulator light-tight and then sits inside it, is it possible to see the orgone energy light phenomena that are described in The Cancer Biopathy?

A. No; bodily evaporation under these conditions creates too much humidity and the orgone phenomena cannot be seen when the humidity is high.

Q. What is the difference in strength between a one layer and three layer accumulator?

A. The one layer accumulator, has an orgone concentration, measured electroscopically, of about 3 times that of the surrounding atmosphere. The three layer accumulator, has an orgone concentration of about 4 to 5 times that of the surrounding atmosphere. However, the concentration of the orgone energy does not increase proportionally with the numbers of layers. Thus a tenfold accumulator will have a concentration of approximately 6 times that of the surroundings.

Q. My accumulator does not fit tightly together. Does this in any way lessen the effectiveness?

A. Occasionally when an accumulator is dismantled and put together again, the walls do not close tightly. These cracks, even if they are as large as one inch, do not in any way affect the strength of the accumulator. In fact, experimental accumulators have been built with openings at the top and bottom for circulation purposes.

Q. Can paper, felt or material other than fiberglass be used for the non-metallic layers of the accumulator?

A. Experiments have been made with other materials. However, all these materials are liable to absorb humidity, whereas the fiberglass will not absorb any humidity. Too strong an absorption of humidity would definitely lower the strength of the accumulator.

Q. If you go into the accumulator with a fever, is the fever increased?

A. In most cases, experience shows that the fever will be slightly increased while the person uses the accumulator, but that the temperature will drop below the previous level after the use of the accumulator.

Q. Is there any special time of day when the accumulator is more effective than at other times?

A. Electroscopic measurements show the greatest concentration of orgone energy between 12 noon and 2 p.m. However, these variations in a given day are so small that it will not make any significant difference at what time of the day the accumulator is used. It can easily be regulated by timing.

Construction of a Three-Fold Orgone Energy Accumulator

1. General information

a. The accumulator is made of six panels which are to be screwed together. All panels except the bottom are constructed in the same manner, and differ only in dimensions. Each panel consists of an inner surface of galvanized sheetmetal and an outer non-metallic surface of celotex which encloses a braced wood frame and alternate layers of fiberglass and steelwool.

b. The materials specified may, if necessary, be replaced by other materials: 1/4" plexiglass or 1/4" upsonite along with a 1/4" fiberglass layer may be used in place of the 1/2" celotex. If substitutions are made, some adjustments in dimensions of the frames may be necessary.

c. Consult the accompanying drawings and tables for dimensions, construction details, etc.

2. Construct the frames

a. Cut the 1 1/8" x 1 1/2" clear pine to the specified lengths (see fig. 1). Cut a 1/2" x 1/2" rabbet into one side of each piece (see fig. 2). The bottom panel frame requires two 1/4" x 1/2" rabbets (see fig. 3). Mitre the corners and join with carpenters' glue and 8-penny finish nails.

b. Brace each frame with two pieces of 1 1/8" x 1" clear pine (see fig. 4). Join with two 2 1/2" #8 drywall screws. The top and bottom frame do not need bracing.

3. Attach the outer surface

a. Cut the celotex to fit inside the rabbets of each frame except for the bottom frame. Fasten in place with 1" galvanized wire nails.

---

1. For a five-fold accumulator, increase the dimensions of the frames to allow for four more 1/4" layers, two layers of fiberglass and two layers of steelwool.

2. The 1/2" celotex (which has one side coated with white primer paint) may be known as: white construction board, soundboard, etc. It is basically a composite of wood fibers. Flakeboard, chipboard, etc. are not the same product and are too dense. There is a styrofoam product with the trademark name Celotex; do not use this.

3. fiberglass building insulation

4. What is referred to as 5/4" stock by the building industry actually measures 1 1/8 inch.

b. Cut two pieces of 1/4" finish grade plywood to fit inside the rabbets of the bottom frame. Fasten one piece in place with 1" galvanized wire nails.

4. Place the steelwool and fiberglass in the panels

a. Place a layer of steelwool upon the celotex inside one of the frames. Steelwool pads (oo fine or finer) when unrolled are the correct thickness. Make the layer as uniform as possible; leave the steel wool fluffy.

b. Next place a layer of fiberglass about 1/4" thick upon the steelwool layer. Avoid lumps and holes. Do not compress the fiberglass. The standard 3 1/2" x 15" fiberglass insulation comes in a roll with either a paper or foil backing. Peel off the paper or foil. This type of fiberglass is difficult to peel evenly. Try to peel it carefully into roughly 1/4" thick pieces. Too much unevenness in the layering will cause the sides of the accumulator to bulge. (There is a fiberglass insulation that peels more uniformly but might be difficult to purchase. It is a Manville product called 1/2" duct liner. It comes in a roll 48" wide by 100 feet long and may be available at your local heating and airconditioning shop.)

c. In a similar manner place the remaining alternate layers of steelwool and fiberglass in position (see fig 2).

d. Place the steelwool and fiberglass layers in the other panels.

e. The bottom panel has a different number of layers. Start with a layer of fiberglass upon the 1/4" plywood inside the frame (see fig. 3).

5. Attach the inner surfaces

a. Cut the sheetmetal slightly smaller than the frames. You might need to have the 26 or 28 gage galvanized sheetmetal cut at a sheetmetal shop. It can be cut by hand, but the results may be too rough. File the edges smooth and round the corners. Cut a 9" x 12" hole for the window in one of the 53" x 24" sheets. Figure 4 will give the exact location.

b. Drill 1/8" holes at 1" intervals around the edge of the sheetmetal. They should be 1/4" from the edge.

c. Nail the pieces of sheetmetal to the frames with 3/4" galvanized wire nails. Nail the remaining piece of 1/4" plywood to the bottom frame before nailing the final piece of sheetmetal over it. (See fig. 3)

---

5. Wear protective clothing and a dust mask when handling the steelwool and fiberglass.

6. Steelwool in long rolls can be purchased from an industrial supplier.

6. Attach supports to the bottom panel

a. Cut two pieces of 1 1/8" x 1 1/2" clear pine 25" long.

b. Screw them into the outer surface of the panel. (See fig 3.)

c. Nail two furniture slides to the undersurface of each support.

7. Assemble the accumulator

Accuracy is necessary in fastening the sides, top, bottom and door of the accumulator. You will want to be able to take apart and put back together your accumulator in the event that you need to transport it. It is very bulky when assembled.

a. Pre-drill holes for the 2 1/2" #10 flathead philips screws with a 3/16" drill bit. So that the screw heads will countersink, drill a 1/2" hole 1/4" deep. Figure 5 gives the locations of these holes.

b. Lay the two side panels upright on their long sides. Place the bottom panel upright against the bottom ends of the two side panels and screw together. Then place the top panel upright against the top ends of the two side panels and screw together. Place the back panel onto the side panels between the overlaps of the top and bottom panels and screw together (figure 5).

c. Turn the box over onto its back and place the door on the side panels. Make sure that the door is flush with the top of the box (this will leave a one-inch gap at the bottom to allow air circulation). Mount the hinges to the side panel and to the door (fig. 6). You can hang the door to open either left or right.

d. Now stand the accumulator up.

e. You will need a hook and eye to hold the door closed while you sit in the accumulator. Place it where most convenient.

8. Assemble the seat for the accumulator

You have two options. The simplest is to suspend a piece of 3/4" x 14" x 22" wood on two supports fastened to the lower support braces of the two side panels. The other option is to turn the seat into a combination shooter box and chest panel.

Construct the shooter box-seat

a. Construct two panels in a manner similar to the other panels as explained in sections 2 through 6. 1/4" finish grade plywood is needed for the top and front of the seat panels (fig. 7).

b. Prepare supports to hold the shooter box-seat panels. Cut four pieces of 1 1/8" x 1 1/2" pine 12 1/2" long. Fasten two 12 1/2" pieces to the lower supports in the side panels with 2" #8 drywall screws. Then fasten the other two pieces vertically between the horizontal support and the floor of the accumulator with 2" screws. Fasten one 14" x 21 3/4" panel to the horizontal cleat A with 2 1/2" drywall screws. The other 14" x 21 3/4" panel will rest against the vertical cleats B (fig. 8). Do not fasten this panel. Attach a cabinet pull (metal or ceramic) to the front of this panel so you can easily remove it to use as a chest panel (fig. 9).

9. Construct a shooter funnel for the box seat

a. Take a four-foot length of 3/8" steel Greenfield casing. (Greenfield casing can be purchased from an electrical supply store. If you can't find it, you can substitute #12 BX electrical cable, but remove the wiring from the flexible metal casing before using.)

b. Insert the spout of a 3" diameter galvanized sheetmetal funnel into one end of the flexible metal casing. Wrap carefully the outside of both the flexible casing and the funnel with plastic electricians tape, so that no metal is showing. Drill a 5/8" hole through the top back corner of the accumulator box seat's top panel. Push the flexible tubing into the space below.

10. Optional

In order to draw stagnant orgone energy from the accumulator and shooter box seat:

a. You will need approximately three 48" lengths of 3/8" Greenfield casing. Use more or less than three depending on individual circumstances.

b. Drill three 5/8" holes (in a vertical row) through the wood frame of one of the side panels. Make sure that one hole will lead into the shooter box seat. Insert the three casings into the three holes. Make sure that the casings don't protrude inside of the accumulator.

---

7. Greenfield casing is a convenient hollow flexible metal tube. IT IS NOT an electrical connection and should NOT be used as one.

c. Place the loose ends of the three casings into a pail of clean water. You need to change the water every week.

#######

The principle of the accumulator box can be adapted to different shapes for specific needs. An orgone blanket, or a collar for the neck, or a large funnel for the breast can be made as long as the principle of alternating layers of non-metallic and metallic materials is maintained.

---

8. The inner surface should be made of a flexible metal screen instead of galvanized sheetmetal. Screen made from galvanized steel or stainless steel is good. Screen made from copper, aluminum and other metals that are not body-own are NOT recommended.

MATERIALS LIST

ITEM	APPROXIMATE QUANTITY
1/2" celotex	four 4' x 8' sheets
galvanized sheetmetal (26 or 28 gage)	2 pieces 53" x 26 1/2"
" " " "	2 pieces 53" x 24"
" " " "	2 pieces 28" x 24"
fiberglass insulation	3 1/2" x 15"(1 roll or 40 linear feet)
steelwool (oo fine or finer)	200 pads or 10 lbs.
1/4" plywood (finish grade)	one 4' x 8' sheet
clear white pine: 1 1/8" x 1 1/2"	80 linear feet
" " " 1 1/8 x 1"	20 linear feet
2 1/2" #10 philips flathead wood screws	30 pcs.
2 1/2 #8 drywall screws	30 pcs.
2" #8 drywall screws	10 pcs.
3" non-mortise or surface mounted hinges	one pair
hook and eye	one 2"
nail-on furniture glides	4 pcs.
nails: 3/4" galvanized wire nails	3 oz.
1" galvanized wire nails	3 oz.
6-penny finish nails	2 lbs.
carpenters' glue	one small bottle
For the shooter box seat panels:	
galvanized sheet metal (26 or 28 gage)	2 pieces 13" x 20 3/4"
clear white pine: 1 1/8" x 1 1/2"	16 linear feet
3" diameter galvanized sheetmetal funnel	one
3/8" Greenfield flexible metal casing	16 feet
electricians' tape (plastic tape)	one roll

(FIG. 1) TABLE OF DIMENSIONS

PANEL	LENGTH	WIDTH
TOP	29"	25"
BOTTOM	29"	25"
SIDES (2)	54"	27 1/2"
BACK	54"	25"
DOOR	54"	25"
SEATS (2)	14"	21 3/4"

(FIG. 2) DETAIL OF LAYERING OF SIDES (2) TOP, BACK AND DOOR PANELS

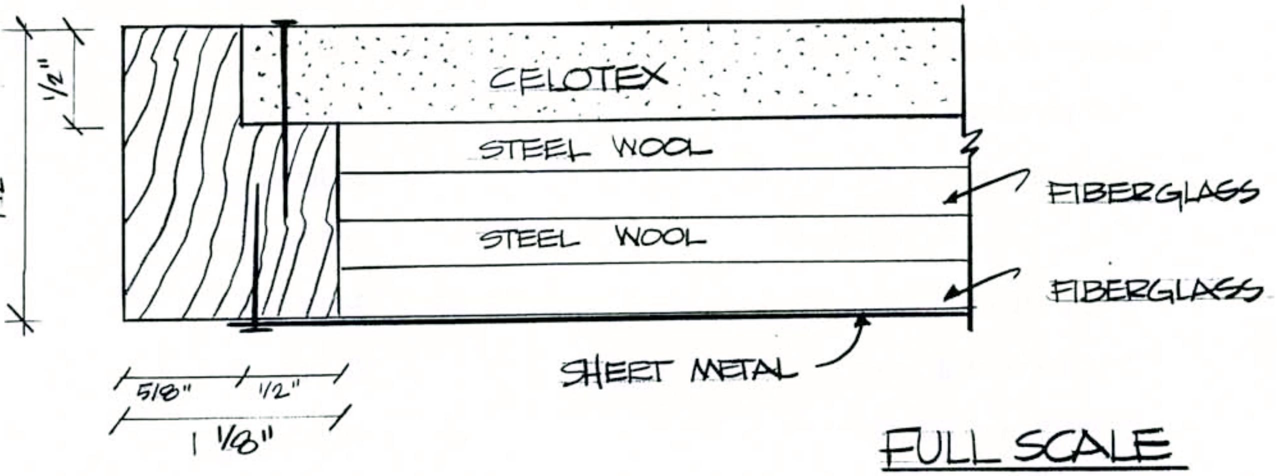
1/2"
1 1/2"
CELOTEX
STEEL WOOL
FIBERGLASS
STEEL WOOL
FIBERGLASS
SHEET METAL
5/8"
1/2"
1 1/8"

FULL SCALE

(FIG. 3) DETAIL OF LAYERING - BOTTOM PANEL

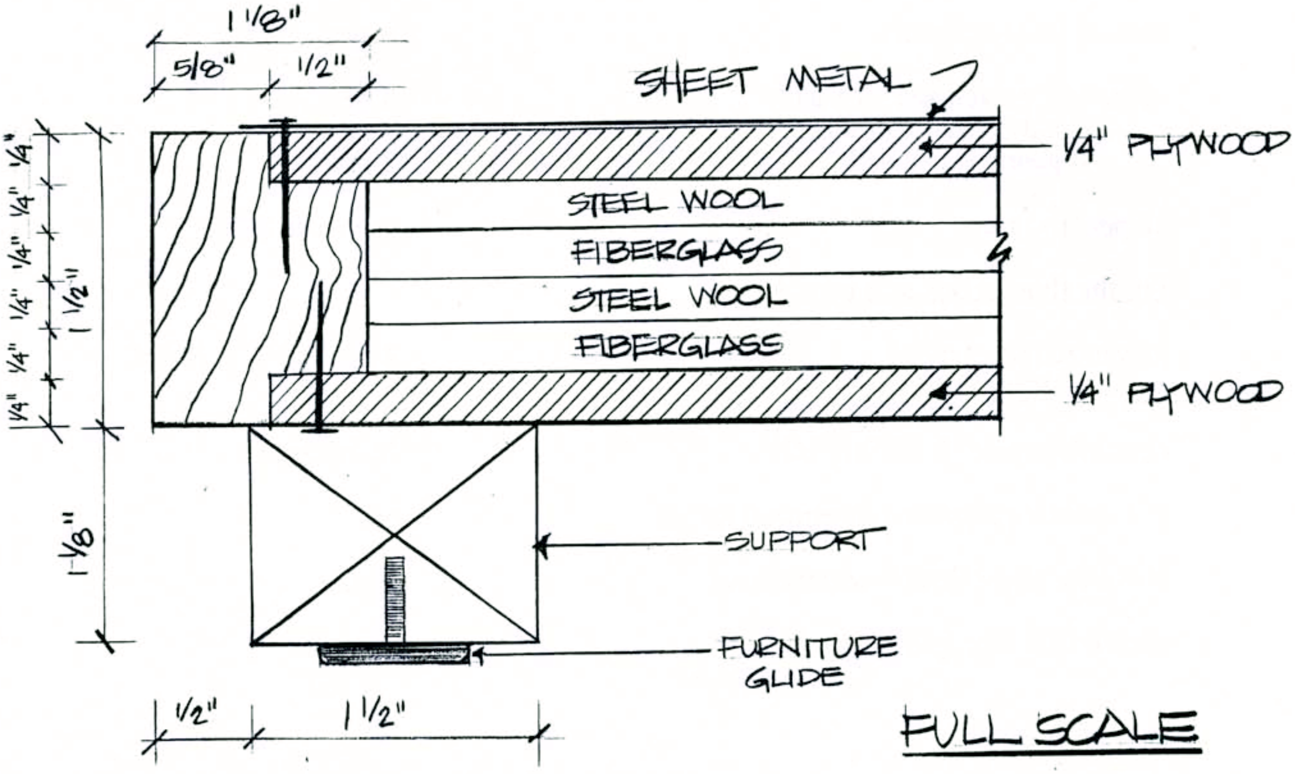
1 1/8"
5/8"
1/2"
SHEET METAL
1/4" PLYWOOD
STEEL WOOL
FIBERGLASS
STEEL WOOL
FIBERGLASS
1/4" PLYWOOD
1/4"
1/4"
1/4"
1/4"
1/4"
1/4"
1 1/2"
1-1/8"
SUPPORT
FURNITURE GLIDE
1/2"
1 1/2"

FULL SCALE

(FIG. 4) FRAMING DETAILS - SIDES (2), BACK, FRONT (DOOR), BOTTOM, TOP & SHOOTER BOX - SEAT (2)

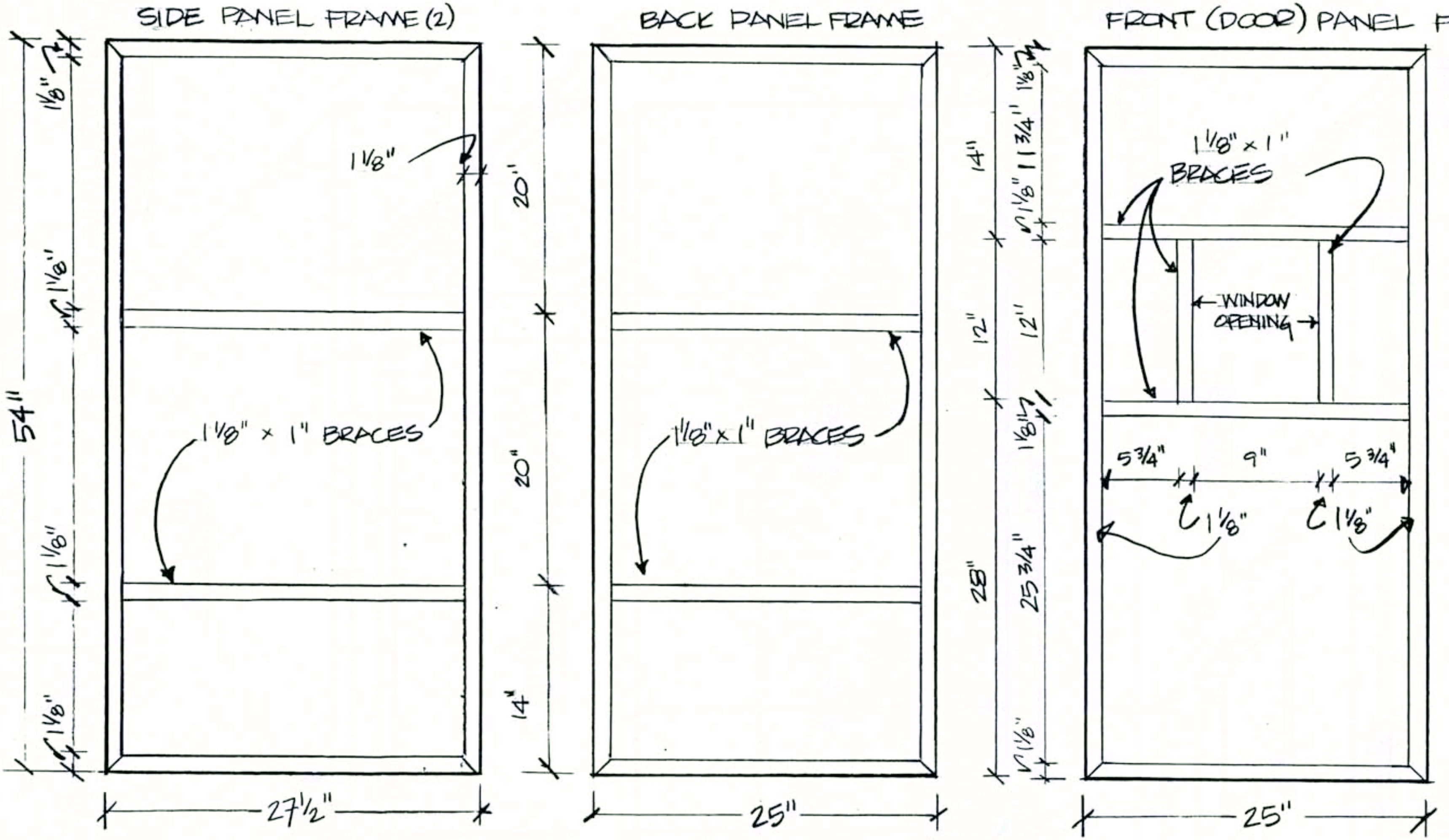
SIDE PANEL FRAME (2)
BACK PANEL FRAME
FRONT (DOOR) PANEL FRAME
54"
1⅛"
1⅛"
1⅛"
1⅛"
1⅛"
1⅛" × 1" BRACES
27½"
20"
20"
14"
1⅛" x 1" BRACES
25"
14"
12"
28"
1⅛"
11 3/4"
1⅛"
12"
1⅛"
25 3/4"
1⅛"
1⅛" × 1"
BRACES
WINDOW
OPENING
5 3/4"
9"
5 3/4"
1⅛"
1⅛"
25"

SIDES (2), BACK & FRONT (DOOR) PANEL FRAMES TO BE 'FASHIONED' FROM 1⅛" × 1½" PINE WITH 1/2" × 1/2" RABBET (SEE FIG. 2)

BRACES TO BE 1⅛" × 1" PINE.

SCALE: 1" = 1'-0"

NOTE: TOP & BOTTOM FRAMES HAVE SAME WIDTH, LENGTH AND THICKNESS.

BOTTOM FRAME CONSTRUCTED FROM 1⅛" × 1½" PINE WITH A ¼" × ½" RABBET ON TWO SIDES (SEE FIG. 3)

TOP FRAME CONSTRUCTED FROM 1⅛" × 1½" PINE WITH A ½" RABBET ON ONE SIDE. (SEE FIG. 2)

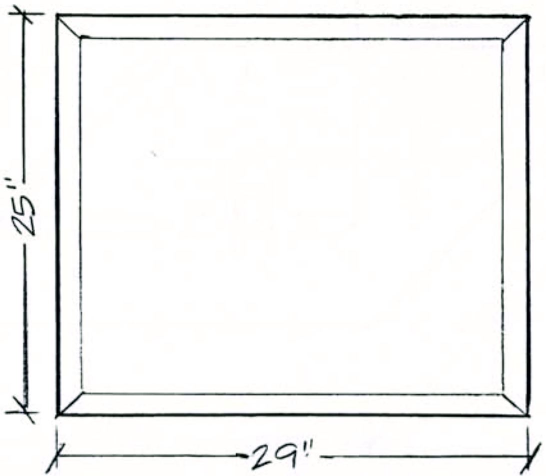

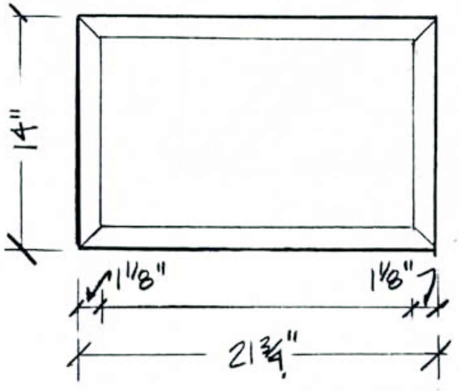

SHOOTER BOX SEAT (2)

FRAME TO BE CONSTRUCTED FROM 1⅛" × 1½" PINE WITH A ¼" × ½" RABBET (SEE FIG. 7)

(FIG. 5)
ASSEMBLING THE PANELS

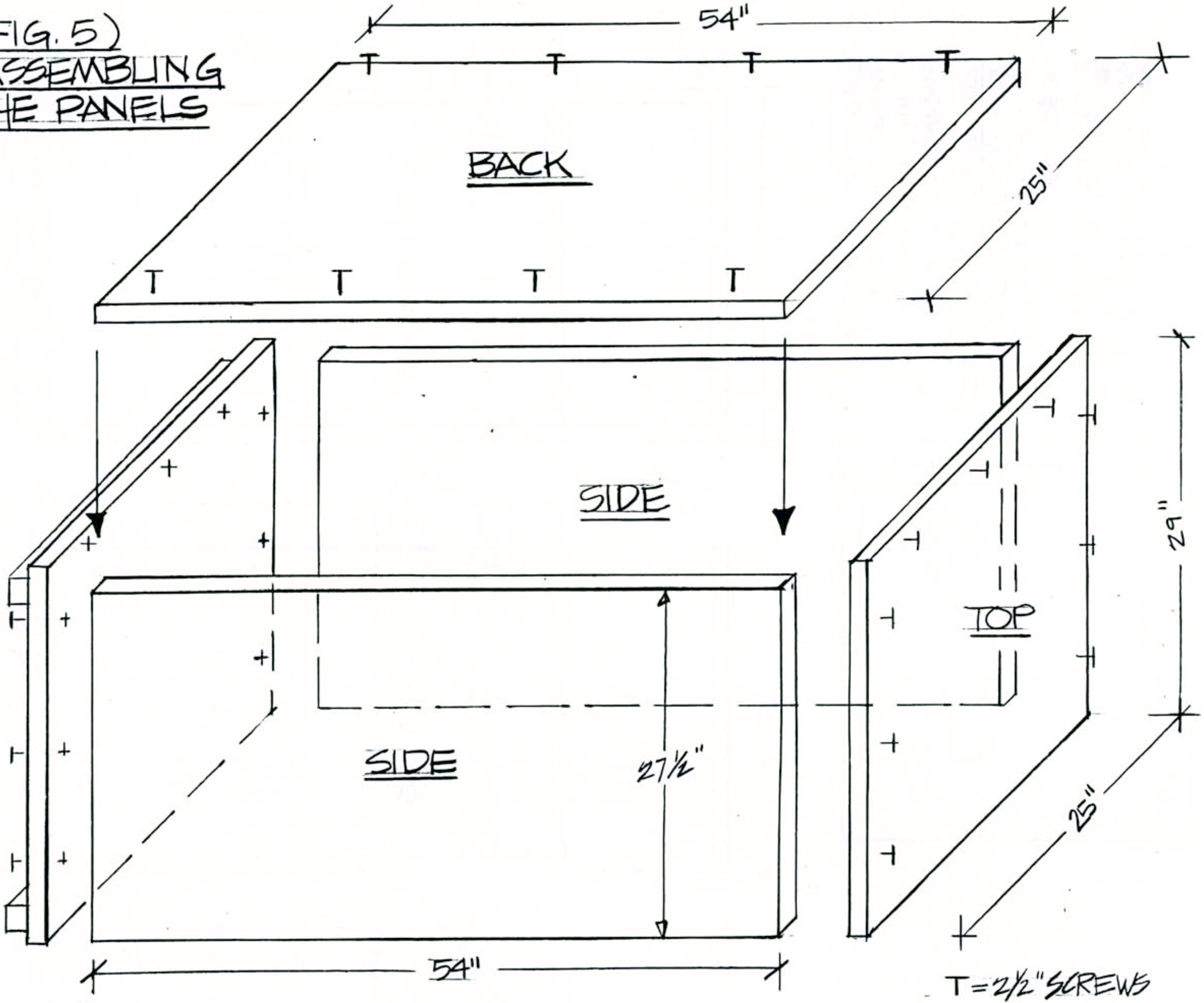
54"
BACK
25"
SIDE
TOP
29"
SIDE
27½"
25"
54"
T = 2½" SCREWS

(FIG. 6)
ATTACH DOOR

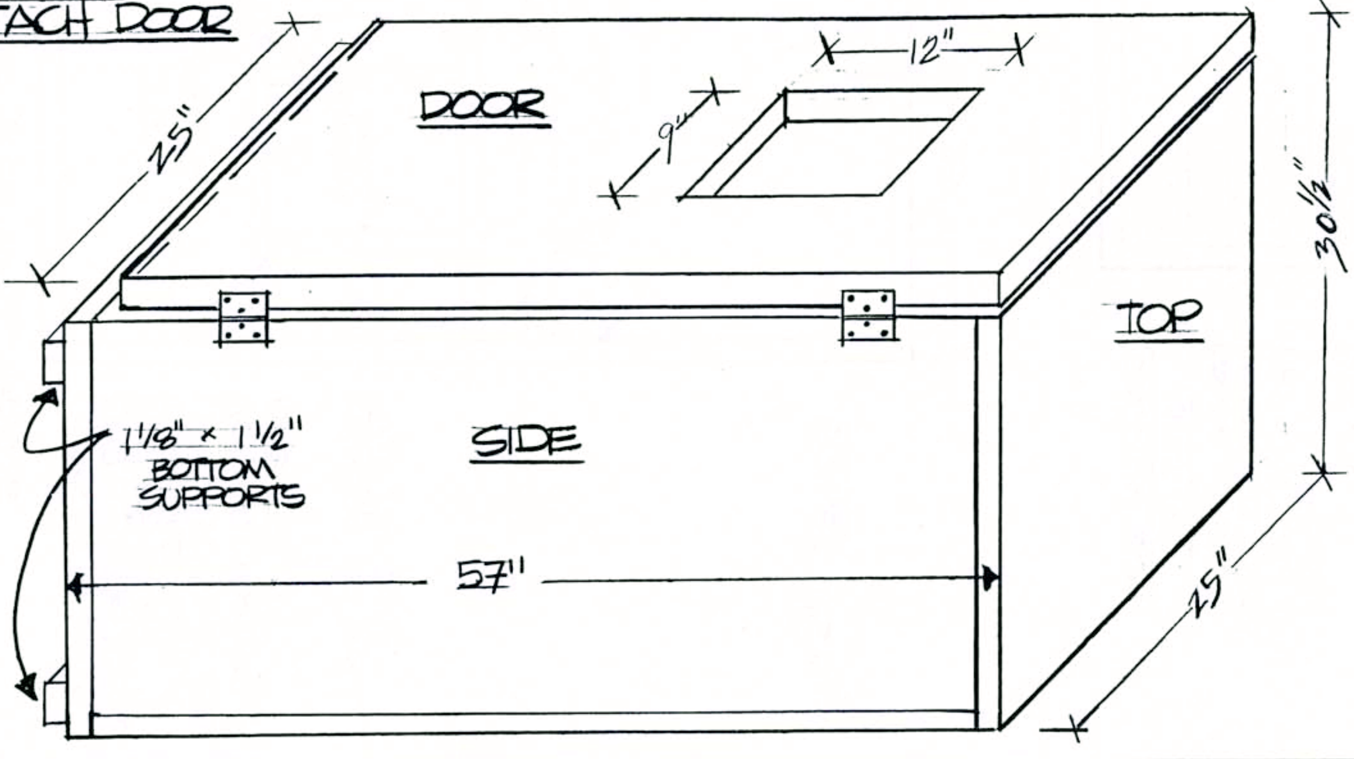
12"
DOOR
9"
25"
30½"
TOP
1⅛" × 1½"
BOTTOM SUPPORTS
SIDE
57"
25"

(FIG. 7) DETAIL FOR SHOOTER BOX - SEAT PANELS

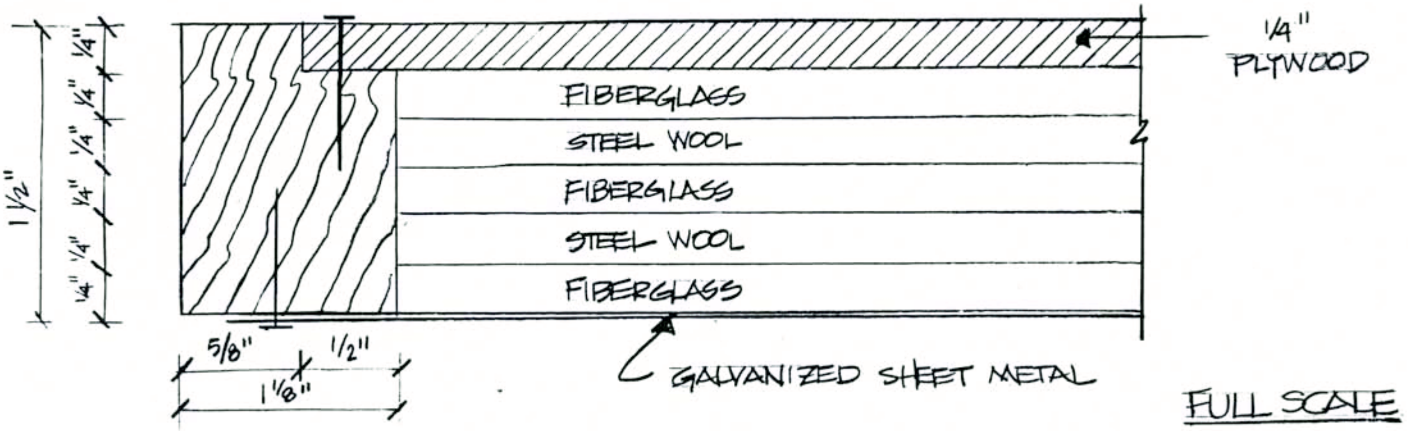
1/4" PLYWOOD
FIBERGLASS
STEEL WOOL
FIBERGLASS
STEEL WOOL
FIBERGLASS
GALVANIZED SHEET METAL
1 1/2"
1/4"
1/4"
1/4"
1/4"
1/4"
1/4"
5/8"
1/2"
1 1/8"
FULL SCALE

(FIG. 8) ASSEMBLING THE SHOOTER BOX - SEAT

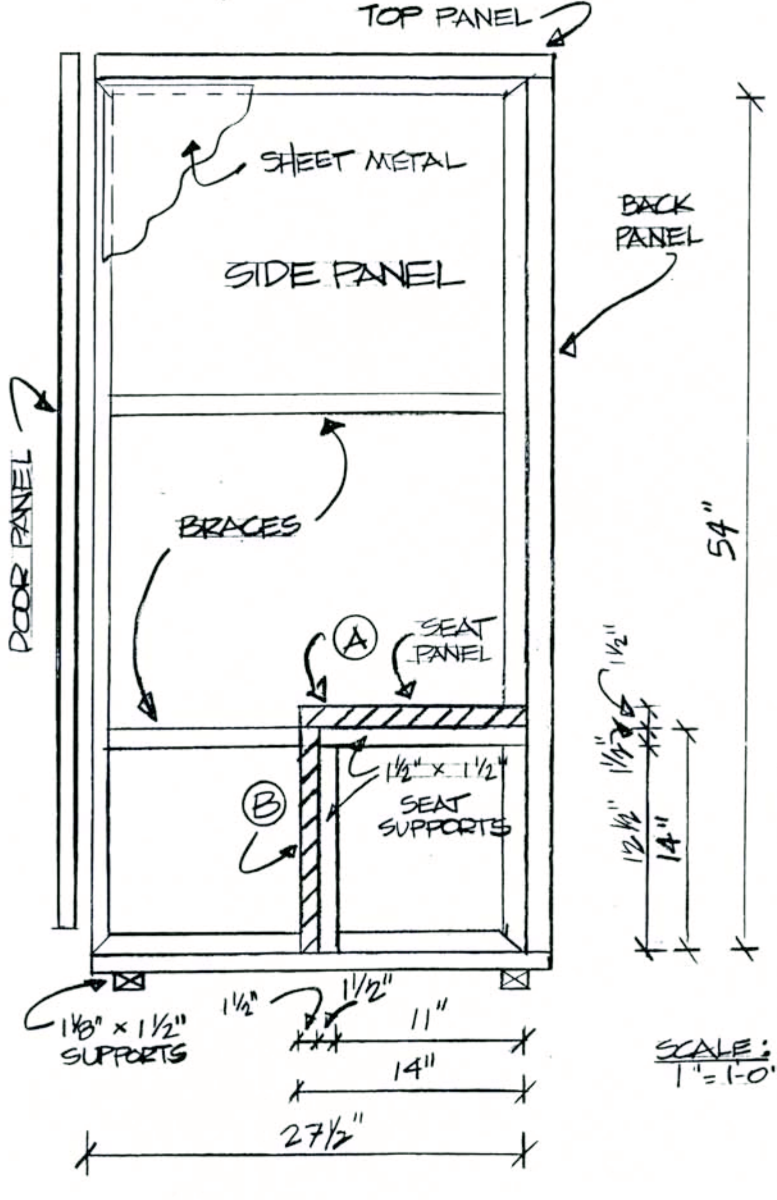
TOP PANEL
SHEET METAL
SIDE PANEL
BACK PANEL
DOOR PANEL
BRACES
A
SEAT PANEL
B
1 1/2" x 1 1/2" SEAT SUPPORTS
54"
1 1/2"
1 1/2"
12 1/2"
14"
1 1/8" x 1 1/2" SUPPORTS
1 1/2"
1 1/2"
11"
14"
27 1/2"
SCALE: 1"=1'-0"

(FIG 9)

HOW TO USE CHEST PANEL

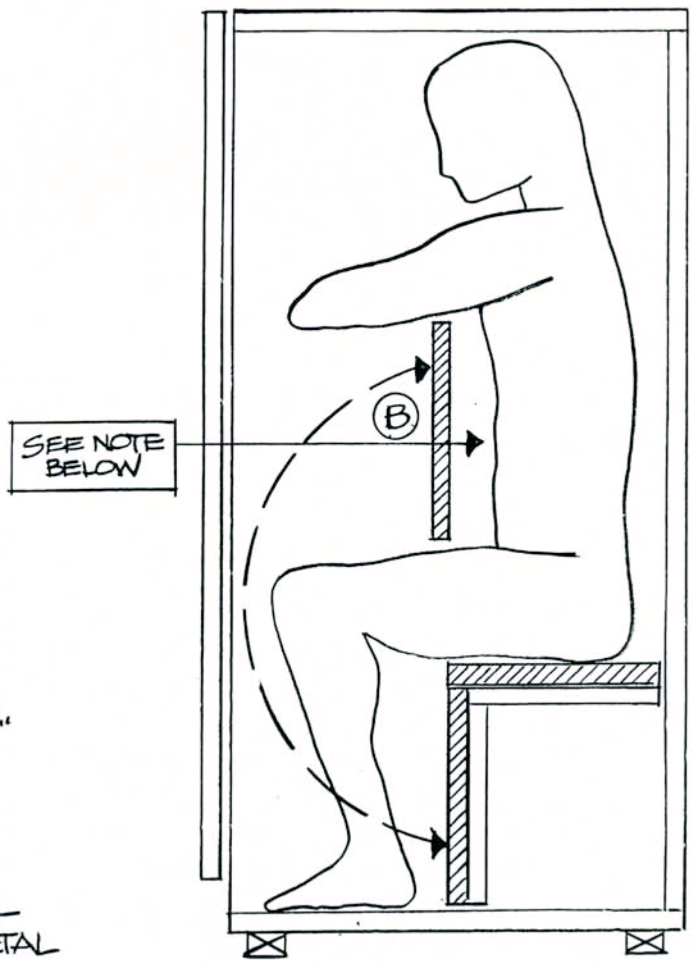
SEE NOTE BELOW
B

SCALE: 1"=1'-0"

NOTE: ALLOW SPACE BETWEEN PANEL AND CHEST OF USER. SHEET METAL SIDE SHOULD FACE CHEST.

produced in your cells. This has to do with the operation of the cell but it also has to do with how your cells are functioning.

One of the first things that takes place as a result of using Sacred G technology is a cleanse. Your cells will literally start detoxing and eliminate toxins that have built up over the years. If you've done a lot of drugs or pharmaceutical drugs you'll notice that those will start to detox or cleanse from you. You might start going to the bathroom a few extra times the first few days, you might get some headaches, you might experience some old pains that come back up as your cells are releasing their memory.

This process is actually what generates energy for you. This is what that saying that you actually heal yourself is talking about. You are actually healing yourself by going through this process. Sacred G doesn't give you energy it gives your cells instructions for how to function and those instructions tell your cells to cleanse their history or their memory. The toxins inside are inhibiting the cell's function their flow and their performance in other words inhibiting the amount of positive energy that you feel inside your body. This is how cells break down and age this is what Sacred G does and this is specifically the Sacred G spot that actually focuses into cleansing your cells.

The Sacred G Spot focuses into the hypothalamus in your brain and is a master gland that regulates all the chemistry inside your body. Using this technology can help you to activate your hypothalamus. When your hypothalamus lacks energy your body starts to run almost like a dictatorship it doesn't want you to grow it doesn't want to change it wants to put you into patterns so it doesn't have to change any of the body chemistry and it locks you into a prison. As you build up toxins the cells get so stagnate that they don't want to shift for you, typically your body is supposed to adapt to whatever situation you are doing in that moment. That's what makes us function effectively in any job your doing or any activity. When you're having fun or out partying and you can't adjust and get into that moment with your family or friends you can't really take part in the treasure that we call life.

This is the very first step that Sacred G is designed to do is to activate and to open up your hypothalamus and open up your body so you can start to change and adapt to any situation that you come across. What you'll also start to notice is a mass amount of energy because as your body clears out these toxins your body is just going to feel better a lot of things happens as a result of this first phase. A lot of people turn to vegetarian diets or if they're vegetarian a lot of people go raw, simply because they can feel the heaviness of heavy foods that aren't good for you which is why it is so important to charge all of your food your water and anything you put on, in or around your body.

During this first phase never use a product that has toxic chemicals in it, only use products that are green, organic and healthy for your body. Any lotion you put on your body you should also be able to eat it because your skin is actually going to eat

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

it. It's going to go into your body. This is the first phase of what Sacred G technology is going to take you through and it's what the Sacred G spot really focuses into. So if you notice you have any kind of cleansing effects such as headaches or your getting old injuries, try putting the Sacred G spot on the spot where it hurts and you'll notice it will take away a lot of the pain.

You can charge your water and it will charge the water to help detoxify your body. So with any type of cleansing that you're doing you can literally charge any type of herb or supplement or food or drink to help your body detoxify just simply by charging it up with the Sacred G spot. This will bring your body into alignment and will help you feel better about whom you are making you feel more comfortable inside of your body again. This is the very first phase of how you are going to increase energy inside of your body. Sacred G is not a magical technology, it doesn't just give your energy from nowhere, and you actually have to generate that energy from inside your body. It is a karma principle, whatever you put in you have to be responsible for taking out.

Transporter

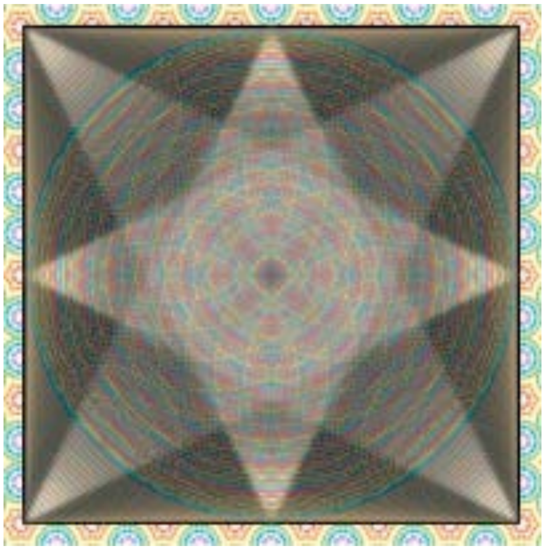

http://coryherter.com/?p=555

The Sacred G Transporter design focuses on the emotional clearing of past memories and current experiences as well.

The second place you are going to get energy from using Sacred G technology is from emotion stored in past memories. I used to do a seminar called the number one reason for every health problem is that our brain becomes congested with

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

emotion from past memories and our brains can't keep up. Emotion stands for energy in motion, in other words its energy that the body used to be able to use to operate its systems and functionality and now that energy is caught up and trapped inside emotion. Your body no longer has access to that energy anymore.

When you're sleeping your brain is discharging your brain wave states processing your memories from the day and freeing up that energy. When you are opening and using more energy during the daytime than what your brain is freeing up during the night it doesn't take very long at all until your body realizes it is losing energy every single day. As soon as your brain makes that calculation it puts the body into a survival state and literally starts shutting down your senses and shutting down your body and it will not activate a lot of the future senses and abilities that the body is capable of, especially the more enhanced and activated spiritual abilities that we are all capable of. This usually happens at the age of two, three or four years of age. You can see it in a child when their brain makes this calculation when the child goes from "hey these are my toys, you want to share? Do you want to play?" to "its mine!" And as soon as they go "it's mine!" their brain just realized it's using more energy than its freeing up at night time.

When the body realizes it's going to die it wants more energy and it starts to horde. Hording the toys is the beginning of a survival state and we never recoup from this survival state. As we get older it changes to people; he's mine, she's mine, that job is mine, that car is mine. Hording material possessions and an endless need to have to own more things and own people and control and manipulate to try to get what we want is really our body saying I don't have enough energy and you're still not giving me the right type of energy. The energy your body wants is this internal energy so that it can heal itself, so how do we offer that? One, we have to increase the amount of healing that takes place why we are sleeping this is so important if you increase the amount of energy that is freed up at night time why you are sleeping in other words enhancing the natural healing process while you sleep. If you can increase that to a point where you are freeing up more energy at night time than you are using in the day your body will make another calculation it will say the body is no longer dying its now starting to regenerate its freeing up more energy every single day. Something happens inside, you come out of this survival state.

Instead of all these outward possessions becoming so important, you find this peace you find this calmness inside. You'll lose that edge you lose the rush of needing things outside of yourself you start to become very fulfilled. This is what we are all looking for, is this inner fulfillment this inner peace and the inner peace comes as we open up all this energy so that's one of the things that Sacred G technology is actually designed for is to help you to go inside at night time when you are sleeping and accelerate this natural healing process. The more Sacred G you use the more this will happen. So what exactly happens at nighttime while you are sleeping? First off when you first start sleeping on Sacred G you might get some tossing and turning the first few nights or even the first few weeks and if you get that it's essentially your beta brain wave state. you have a lot of extra energy and blocks

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

built into it and you get into these anxiety states and typically it's because you just hate your life, you hate what you're doing you hate how you're serving, you hate how you have the same job for so many years, you're not happy with your situation and it starts to bring up all these emotions you have been suppressing all these years and it brings them up and puts them in your face and it makes you deal with them. All of sudden you can't handle it anymore and bam! You have to make some big changes in your life in order for your sleep to actually shift. This is part of Sacred G technology helping you to transform what is taking place in your life and to stop letting things go, stop letting your life go and start going for what you want instead of just settling for less. That is not what we are here for, we are here to have the life of our dreams.

The second thing that you can notice is more dream and nightmares taking place. This is all of your past memories and past emotions clearing out. Your brain takes all of your memories and cuts them into little snippets and links them all together. This is why your dreams are all scattered and all over the place, because its processing series of memories which all have to do with the same types of patterns and processing it all so you that you can release these patterns and step out of these zombie like states. This is where you start to become alive and animated and where all of your energy will come from. This is where you step back into your power and take your power back and you actually start living the life that you are looking for.

This part of the process actually has to do with the limbic system, which is responsible for storing all the emotion from your past. Now if you look back into any major memory and you focus on that memory for a little while, you'll actually feel the emotion come back up. That emotion is simply indicating to you that there is still a lot of unprocessed feelings and energy back in some memory that could 20, 30, or 40 years ago that is literally still holding you back. So when you start cutting those cords and clearing that emotional energy, not only do you let go of those attachments from the past so you can live more in the moment, but you also get new information and new understanding about who you are. You get new lessons and fears you have as result of traumas disappear and you stop putting yourself into a prison to protect yourself. This whole process is actually initiated from the Sacred G design called the Transporter that focuses literally into the limbic system for emotional processing.

Now that is a massive resource of energy, so once that emotion is processed the body can use that energy to start healing itself. Sacred G help to heal on you, the more emotion that is processed the more energy that is freed up so your body takes that energy and it says where can the body best use that energy to reactive it to heal it more effectively so the more memories you process the more energy you are going to generate the more health and healing you are going to attain.

As you really start to heal the body it's not just about healing the body, is literally only the first step. That's all about performance and opening yourself to a whole other way to experience reality so that is the second way that Sacred G frees up energy inside of you and that energy is inside of you It's not from Sacred G it just puts you through the process and when you go through that process your doing the work to free up that energy. If you have a couple nightmares and you're not sure you want to sleep on this technology, let me tell you one thing; your limbic system is directly connected to your hypothalamus and regulates your body's chemistry. If you have extreme nightmares, that's extreme negative energy. When that emotional energy is converted to chemical energy from your limbic system to your hypothalamus, that negative energy is converted to chemical energy that energy is the stuff that actually kills you. In the end when you have that anxiety that is coming up and you can't handle sleeping on this technology at night, that's that energy that's saying I hate my life I don't even know why am here you lose your purpose you lose your point going through that helps you find who you are and your purpose.

Your point it helps you get back on track with your life cleansing emotionally can sometimes be a little difficult but stick with it go through it use the Sacred G meditation on the website attend some of the webinars get some of that energy moving get on a community get on a chat get some support from people who are using this technology if you are having a hard time go through the process this is your life this is your energy

Ascender

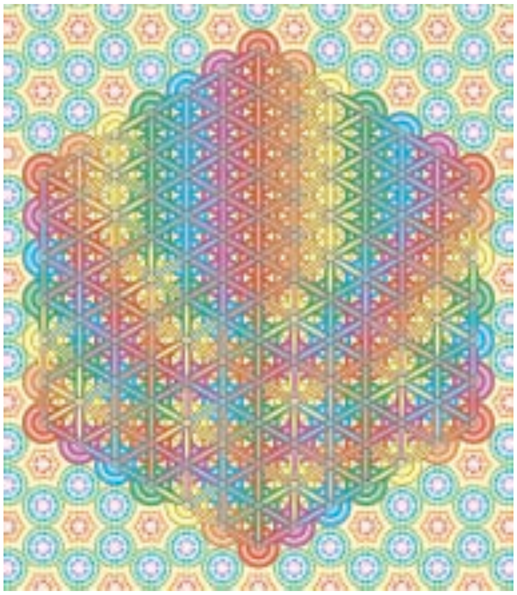

http://coryherter.com/?p=553

The Sacred G Ascender design helps you to resolve and internal information

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

conflicts and false belief systems. The Ascender also helps to gain focus and clarity in your life and identify your goals.

Sacred G doesn't just give you energy; Sacred G walks you through a process which generates more energy from within. This energy is your energy, the regulators, the vibrations and everything that is set with Sacred G is all customized specifically for you and your experience. It's your body that is actually running this energy and this development and healing process. There is a development system that every single body has built into it, which is have been walking you through this process but our experience of Sacred G is completely unique and all customized for us individually.

This next phase of Sacred G is built off of your belief systems. Belief systems are how you like things, your perceptions, do you like experiences in life to happen slow or fast? Are there certain aspects of your life you like to happen fast and others you like to happen slow? How do you set all your regulators for how you experience life? The next phase of Sacred G starts opening up belief systems that you have created throughout your life, some because of passion some because of trauma.

When you start focusing into these belief systems, you'll find that there are a lot of belief systems that are not really serving you anymore. In fact a lot of belief systems is our brain beating itself up and beating us up telling us that we can't do it or that we are not enough or that we are not smart enough. We have all these belief system inside and what are they there for? They are sitting there telling us these things so that we don't expend our energy.

Remember that when we are in a survival state it does not want to expend energy it wants to conserve it. As soon as you start to think of new ideas or new places to go or new things to do your body is like no we can just sit here, lets just sit here and watch TV or do something less strenuous. However when you start opening up your body, in this next phase and start clearing up all the energy from cleansing all the toxins and clearing out all the emotions from past memories, you start opening up energy and your body starts feeling better and starts coming out of this survival state.

All these belief systems and all these thoughts have to be reconfigured. There is a whole process you have to go through coming from "Oh I don't know maybe we shouldn't do that" to "Hey I'm open let's do it! I've never experienced that before lets give it a shot" and jumping in and actually immersing yourself into part of the game. This phase of Sacred G is the Ascender design, which will open up or unlock your belief systems.

If you have ever heard of post dramatic stress disorder or multiple personalities, what happens anytime the body is induced with massive amount of emotional energy especially if it is negative because your limbic system is directly connected with your frontal lobe which is what is responsible for organizing all your belief systems if you induce a massive amount of negative emotional energy it forces your frontal lobe all of your beliefs your personality your identity to actually split and when it splits it goes from one unified identity to two separate halves and they are in opposite polarities so those are things like bi polar depression and schizophrenia all that is actually induced by traumatic states or not properly wiring the brain as you are growing up through childhood. So one of the things that the ascender does is it starts taking all of these opposites and these two halves and as you clear out the negative emotion through sleeping it automatically forces your frontal lobe or your brain to start solving conflicts and when you take two pieces of conflicting information and you find a unifying piece of information you get this third piece up here which unifies these two you get out of this land of the opposites and we move into a unified state and this unified state all of sudden you have a new piece of information and with this new piece of information we have a new perception.

A lot times you'll start talking especially with the ascender and this phase of Sacred G technology a lot of times someone will ask you about something you'll start talking and you'll think "wow I've never really thought that way before." You are actually seeing some of the changes that are taking place internally but your actually seeing them in hindsight. This is one of the most amazing processes, especially for people who value information so much and have a massive amount of information and who are really in the game to figuring out how things work because the ascender will literally start solving pieces of information for people. What starts to happen if you have this identity and self over here and another identity and self over here and all of a sudden you eliminate all the things that cause separation between the two of them they come together into this bigger personality this bigger identity and this bigger identity contains more energy because you've got more active neural activity inside of your brain.

If you've had that experience before where you start making these internal connections and you finally get it and you get this rush of energy and you get tingle down your arms and legs your just vibrating with energy that is actually what has happened. Two different identities have solved conflict and have actually joined and integrated together. This is essentially kind of like putting humpty dumpty back together again. As you start to increase the size of your identities, confidence goes up with confidence comes more energy. You feel like you have more energy that's because you have more neural activity in your brain and more communication going on through your body. you can quickly make decisions you can take a lot more information in at once you can see larger patterns because you can see more information at once in other words you can download more information actively into the neurons in the brain so you can detect larger patterns.

The bigger the pattern you can see in life, you start to become aware of some of the major belief systems that drive all of your actions throughout your entire life so you can start to set yourself free from those. You start to see patterns in your business, in your job, in the way you spend money, and the way your relationships happen. You start to identify how to see more information at once but those all end up leading to creating smarter, wiser choices for you in the long run. This is one of the amazing things about the Ascender design is its ability to unlock new information to you. Unlock new perceptions and overcome conflicting belief systems and really open your mind to new information and new experiences and to let go of that overwhelmed feeling of where we are and the conscious understanding of who we are.

Trancendance

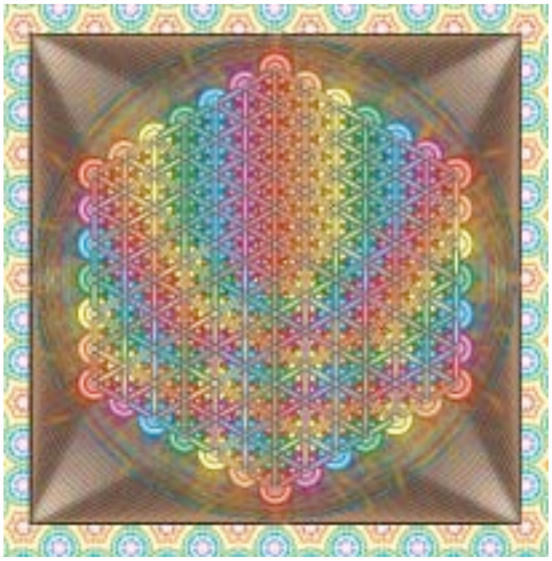

http://coryherter.com/?p=551

The Sacred G TranceNdance design assists in clearing emotion from past memories as well as solving internal information conflicts. TranceNdance is specifically designed to increase intuition and unlock new perceptions and strands of information.

The Sacred G Trancendance design is by far the most popular design. This design essentially helps to cleanse and detoxify the cells, process the emotional memory from our past and also what the ascender design does. So it's kind of a conglomeration of the past three; the g spot, the transporter and the ascender. This was crucial in order for me to step forward into the next one. Everyone loves the Sacred G Trancendance, it charges water, you can do everything with it. It's very powerful because it's combination of the last three. It's been designed to kick in some of the artificial intelligence or the ability to start downloading and streaming new information from cleansing cells, clearing out the emotion and then solving the information. Trancendance helps unlock these information systems even further.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

A lot of the research has been done with Trancecendance. We have had celery last in the refrigerator for about a month compared to uncharged celery only lasting four days. Lettuce literally lasts in the refrigerator for three months, milk lasts for two or three months. We've had bread that doesn't mold, flowers lasting three to four weeks normally wilting up and dying after one week, contact lenses that are disposable that are only good for a week lasting three to four weeks because it stops the protein buildup. So many cool things have been done with Trancendance because we have passed around to so many to people freely and now that we have opened up all the designs there is so many cool types of research that we can do. I hope you are going to get involved with it and create research projects and do them and track the results and really being part of the Sacred G. This is a massive discovery for how to really open up the capabilities and the potential of life.

Simulator

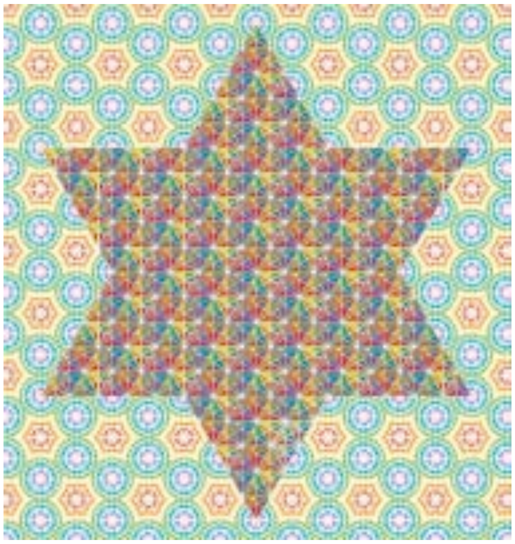

http://coryherter.com/?p=549

The Sacred G Simulator design helps you to learn how to express what is inside. Speaking your heart and mind and being authentic with your own experience is a major part of the transformation induced by this Sacred G design. This design is also called the simulator as it bridges your internal reality with your external one to prepare you for taking action towards your dreams.

The next phase that Sacred G will take you through is the fourth phase. The first phase was clearing and detoxifying the cells. The second phase was clearing out the emotion from past memories. The third phase was reconstructing your neural network or how your brain is organized and how it organizes all your information, eliminating the conflicts and doing that to boost energy, boost confidence, and boost intelligence. The fourth phase has to do with expression. Expression is so important. Now if you watch my videos labeled 2006, you can see that the Simulator wasn't actually developed then. In fact, I only developed the Simulator

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

after those seminars, and you'll see the difference in my expression and my animation. Now after spending about 14 months on the Simulator, you'll see the difference in the way I express myself, and how much I've opened up my expression abilities.

Now there are actually three regions of the brain that the Simulator opens up, and those are the parietal, occipital, and temporal lobes. In other words, it opens up your ability to take your thoughts and convert them into language. You know that when you talk to somebody, you're actually hearing what you say and you're kind of getting a visual image of what you're saying as well. When your trying to figure something out, it's so much easier if you have someone to talk to because you start understanding it deeper because you are starting to convert it from an internal or a two-dimensional state to an external or a three- dimensional state.

A lot of people ask me, "How can you put these designs onto a flat piece of paper and expect them to do anything?" We live in a three-dimensional world but this layer of programming is actually what takes this energy and reconverts it back out into our three-dimensional reality. Before this, it was all just internal changes and changes that were happening inside of you. Now with the expression, the changes that are happening internally as a result of using Sacred G can really start to manifest through your expression.

So if you have ever heard of the spiritual ability of channeling or being able to download new information, what you are really doing is channeling information that the previous design, the Ascender, unlocked by solving conflicts in your own belief systems. As it is unlocking new information, your brain has to bring conscious awareness of that information. So what starts to happen is someone asks you a question and all of a sudden, boom, it comes through you. You're talking it, and you've never heard that information before but you talk like you've known it forever. You're like, "Oh yeah! That's how it works." and you're like "Whoa, cool!" It's cool when you get into this one. It's a really cool one.

It's called the Simulator because this one actually sets up simulations, so after you change your belief systems and you say, "OK, this is what I really want in my life," the body says, "OK, well, we need to just confirm that." So what the Simulator does is actually create a simulation in your life and give you an opportunity to actually choose that other choice that you want to get you out of that other pattern. What is really cool about the Simulator is that if you don't pass that little test to change your programming, it will create another one, only bigger, and it will create it even bigger so you have to pass the test to get what you are actually looking for. It almost becomes this mechanical game of the technology literally walking you through the process, helping you change all your perimeters and regulators inside of your brain so you can have the actual reality that you are looking for.

The Simulator also focuses into helping you to be able to actually express what is in your heart and not just what other people want to hear, but actually saying what you really feel, even if that means opening up to conflict. It means opening up to whatever possibility, but opening yourself up and opening your heart up and opening up who you really are inside. So it really forces you to start coming into integrity to who you truly are and what you're here for.

That's a really major change, and we actually had the Simulators on posters, which we called the Love Chargers, and a lot of people made some pretty major changes in their lives with relationships, with jobs, with careers. Usually people who were working in the same job for 10 years and didn't move up, just quit their job. They had all this anger and rage that came up at their job, at their boss. Why… because everyone else was growing, and they weren't. The company wasn't taking caring of them, so why are they there? They could be moving their life forward. They could be growing. They could be becoming more, but because they weren't expressing themselves, their needs weren't being met, so there was really nothing else to do besides go somewhere new and start actually growing and getting the life they are really looking for.

This is where this energy in Sacred G technology really starts to come out into your external life because after making all these internal changes you have to integrate them into your physical reality in order to make them real. So that's where Sacred G starts to set aside all of the other vibrational technologies because of these conversion systems to get your thoughts out here and start manifesting into reality. That's where people usually get caught up. They're always looking for ways to manifest and meditate more and open themselves spiritually more, so that they can get out here. The Simulator is really what starts that process. It's amazing!

There's so many people in the world looking for this technology, looking for the Simulator right now, and have no idea that is what actually will retrain their brain to get them into that state of mind so that they can manifest the life that they are looking for. It does so much! I learned so much about the new memory systems and about expressions and how to really get out here again. Watch my videos from 2006 and watch them now and see the difference in my animation and how much more alive and how much more energy I have from literally the result of the Simulator.

Activator

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

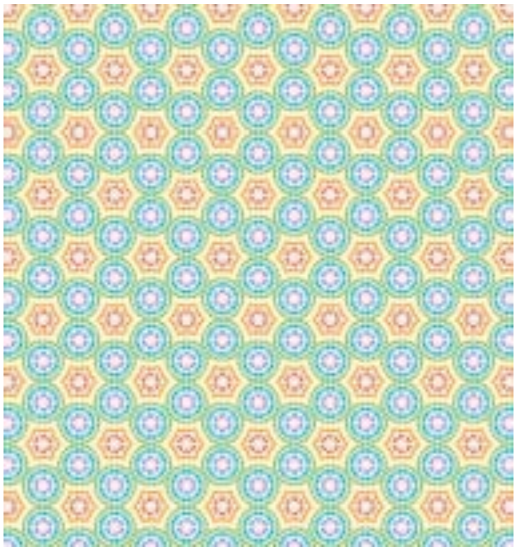

http://coryherter.com/?p=547

The Sacred G Activator design is specifically built to activate your body. Focusing on movement, motion and space, the Activator is all about the value of getting out of your head and back into your body so you can create the life you have dreamed of.

The fifth phase of Sacred G technology is going to walk you through another region of your brain, and it actually has to do with your cerebellum. Now your cerebellum is located in the back of your skull, and your cerebellum is the very last place that you can use your brain to do any tweaks to this vibrational energy. The cerebellum has to do with motion movement. It means taking the first step, making it happen, taking baby steps, moving forward, moving your ideas forward, not just expressing what you want, but taking action steps to make it happen. So instead of talking to your mom or dad or your friends about doing something you want to do in your life, actually taking that step and talking to professionals who can actually move you in that direction for where you want to go. Since they're already doing it, they have the information you need.

So the cerebellum is really interesting, and the Sacred G Activator is the design that really activates the cerebellum. Of course, we call it the Activator simply because its designed to activate and physically activate the body. So this includes things like working out, getting in shape, taking action steps in your life. Enough talk! Let's go. Let's make it happen. Let's take those action steps! It also has to do with building your identity. What separates you? What sets you apart from everybody else? What is your uniqueness? What is your specialty? Who are you? Why are you different than everyone else? What is your value? This is the level where you really start to manifest and understand your own power because you start to understand your own value and what you are worth. When you understand that, you understand what you can offer. When you understand what you can offer, people know what to look for.

For instance, the process I went through was to understand that my real specialty is near-death experiences, Sacred G technology, anti-aging, and memory systems because I brought my memory back. I know my specialties for what I am good at, there's a lot of things I'm not good at, but those things don't matter. All I have to do is focus on my own personal specialties because that is what I am going to offer to the world. That is what the Activator is really designed for, to prep you, and get you ready to make you get out there and to get you in that flow. So it's the very last phase to restructuring your internal core so that you can get out into the world and really make your magic happen on the planet. You can be somebody who has discovered their own wealth and is available to serve humanity.

Now the cerebellum is really unique. This is where with the Activators we really started to understand the value of the Sacred G bed sets. Your cerebellum contains more than half the amount of neurons in your entire brain and compresses them into a region of only 10% of the space of your entire brain. That's how compact and compressed these neurons are. That is pretty dense energy. What we started to learn with the technology, when we add more, and stack them together, is that what it does is increase the density of the energy field. That density was required to activate the cerebellum. That's when we started coming out with the cases of the posters and the bed sets because it took that much energy to activate that region of the brain. So the difference between the Simulator and the Activator was that it really pulled out the need for the bed sets to really manifest and harness what Sacred G can really do. Another thing about the Activator is that besides just running a simulation like the Simulators, the Activators will not only run a simulation but it will actually walk you to that right space in the room to the right person. So the Simulator can express what you need because your thoughts have already been attuned from the Ascender. You say what you need, and all of sudden you're making these really amazing connections to move your life forward to where you want it to go. This is where the real magic starts to happen. It's not fully activated, but this is where it starts to happen. It's training you on a process for how to really manifest your dreams.

So that's all the synchronicities in your life. It's your connection with your external environment. Your cerebellum, because it has to do with motion and movement, is a three-dimensional mathematical equation. The Activator is a three-dimensional system, so it will influence the ways that you move. You start to feel pressure in yourself, in your movements, and other people. Your senses start to enhance. People that slept on the Activator bed sets noticed that all their senses started to open up, especially making love and expressing their sensuality. Because they are exposed to the technology through making love, they started to open up. The body is healing faster at nighttime from doing workouts and pushing your body to the extreme, especially with extreme sports.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

So that's all part of the Activator and something that is really crucial for unlocking our true potential and training our brain, our body, and ourselves on how to really be in the flow of energy in this external three-dimensional environment. Not only linking up our emotions, our thoughts, and our expressions, but also our physical movement, as in knowing where to go, when to leave, who to talk to and who not to talk to, as well as some people take up your time. So sometimes we get caught up, the Activator says, "I got to go… sorry, my body is just taking me."

Each one of the Sacred G designs is focused into specific regions of the brain to activate it. Now what is really amazing about the Activators is that it completed this cycle of energy. When this cycle was completed, it would start another cycle, but you would increase the amount of energy units inside your pineal gland. That's what it was, all for if you could continually increase the amount of energy flow in this cycle or the amount of inner chi in this cycle. If we can increase the amount of energy in this flow, what we start to do is actually activate the pineal gland and the pineal gland has all the magic in it, including the DMT, the connection to Source, the ability to regulate brain waves, and the ability to regulate and activate hormones for maximum growth potentiality, sexuality, and sensuality. It opens life up. The pineal gland is literally that point, that central region of the brain that we are all looking for that really identifies who we really are. It's our connection to Source. It's our spiritual self and our physical body. It's our environment. It's every level of us compacted into one single unit. It's the singularity point.

So the more we can go through this flow of energy using all the Sacred G designs, the more the pineal gland starts to activate. It only took me a little bit longer of integrating, going through all the tests and the activations I had to go through with the Activator to reach the next level of technology.

Sacred G Fusion

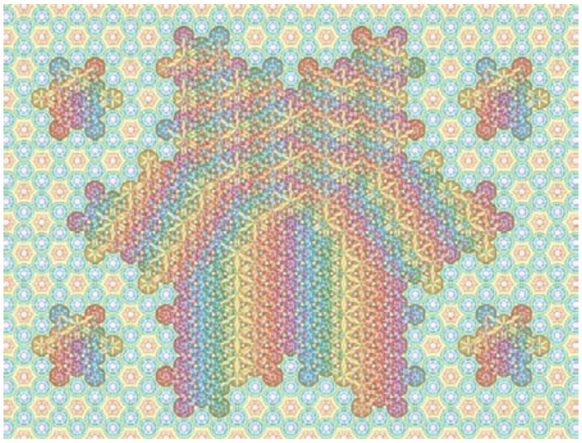

http://coryherter.com/?p=545

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Sacred G Fusion – (Available only in Cases)
Sacred G Fusion is the ultimate Sacred G design that integrates every single Sacred G Design into one cohesive unit. This design enables you and your body to focus into whichever previous design at any time to clear and activate each system for maximum energy production. The Sacred G Fusion also has additional programming that focuses entirely on the pineal gland, activating your connection with God, your body, your environment, and the people in your environment.

Sacred G Fusion is not just another Sacred G Design, it's a highly advanced system that enables you to function on a whole new level that transcends any previous Sacred G Design. Instead of going into specific regions in your brain to increase flow to activate your pineal gland, Sacred G Fusion activates the pineal gland first and then it spreads out into the different regions of your brain. Because it is able to reach more regions of the brain at once, it eliminates a majority of the cleansing and the detoxing effects such as nightmares. You will still be detoxing and may have challenging effects from that but the process will be more smooth, fast and efficient. Each different Sacred G design prior to Fusion mainly affects one of five different areas of the brain. Sacred G Fusion is an incredible tool for activating human potential as it addresses all five areas, increasing the flow of energy to activate the pineal gland. This focus on the pineal gland allows Sacred G Fusion to bring about an experience of singularity, compressing down to the point of all creation. This leads to an intense experience of connection with Source and an unlocking of the spiritual senses, helping us to truly understand and know ourselves inside and out.

Sacred G Fusion has completely transformed my life in 4 months. I've been opening up in ways I haven't even dreamed of and the pineal activation is the key. That connection to Source has allowed me to channel all of these systems and be trained in them as I was taught how to build them. The Sacred G Fusion poster is a minimum of 40 X more powerful than the Activator and the Love Chargers. The power available in Sacred G Fusion deepens one's ability to go within and bring ourselves into complete alignment. What used to take me 2 to 3 weeks to process, now I can process in 2 to 3 days. Everything has been accelerated. I sometimes refer to Fusion as Sacred G 2.0 because it's a whole new type of inner network that you can experience. Here you tap into the magic, the flow, and the synchronicity. Every where I go, I'm always meeting people who just had someone pass over, who had near death experiences. The particular people that I meet are interested in vibrational energy, interested in sacred geometry. Literally every single place I go, I am meeting people who already know of Sacred G. The synchronicities are amazing and at a level that I've never experienced in my life.

Living daily consciously in the flow of synchronicity is something I want for you to experience. We have different Sacred G designs on the website available for free download. I encourage you to print them out and use them to tweak different areas of your life. When you're ready to amplify things and take your life, abilities, and world to a whole new level, I really encourage you to get on Sacred G Fusion. Due to

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

the complexity of the mathematics and design, Sacred G Fusion is only available in poster size. The Fusion design is only available in cases and bed sets simply because I want you to have the full experience and that requires enough energy to actually activate some of the systems.

Sacred G Fusion Case – The Tree of LOVE

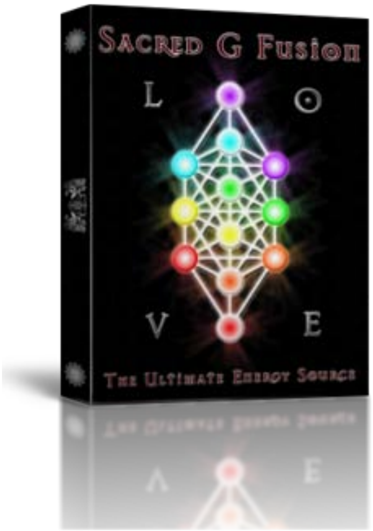

http://coryherter.com/?p=543

Sacred G Fusion comes in a specially designed case. Watch the bonus video to learn more about how the Sacred G Fusion works with the Sacred G Fusion case. The case illustrates the Tree of Love and provides a stabilizing container for the intensity of the 300 Fusion posters. On the Tree of Love, you'll notice twelve colored circles that stand for the 12 chakras. Since all vibration technology is at the top chakra, all Sacred G and energy technology falls in the arena of the blue chakra. This is also referred to as the Sacred G Spot and the area associated on the physical body is your head. The green chakra, associated with your chest and heart, is the limbic system and corresponds with the Transporter design. The yellow chakra relates to your identity and has to do with your frontal lobe and the Ascender design. The orange chakra here has to do with your expression that is the Simulator and relates to the knees. The red chakra essentially has to do with your cerebellum, motion movement and that of course is the Activator.

Sacred G Fusion contains the mathematics of the entire Tree of LOVE, walking you through the outer external chakras that have to do with your job, friends, discovering your true purpose and passion, your relationships, family, and the ways in which you give back to the planet. The orange is the friendship, the green is your relationship, the purple is your family, the red is your job, the yellow is about discovering your passion your purpose, and the blue is about serving the planet in a really big way. These chakras actually regulate and explain all the biological systems for unlocking the body. I encourage you to learn more about the Tree of Love through my videos and trainings. You'll come to learn about the double helix inside the geometry of the Tree and the genetic shifts that occur as you move through the chakras. Some of the advanced systems actually reveal the 23 different chromosomes that make up your entire genetic code. There are particularly amazing things to be found in the Tree of Love related to brain wave states: beta, theta, alpha, delta brain as well as the second beta that is the purple chakra.

There are many ways to interact with the Tree of Love and the Fusion case. For example, if you have a stomachache, you can put a glass of water on the case with all your Sacred G Fusion inside. Place the glass over the yellow chakra that represents your stomach. Leave the water there for about 10 minutes and then drink the water. You'll physically feel the difference in your stomach. You can charge up water, jewelry, food, anything you can physically put onto the chakras on the case and it will be energized with vibrations that will be absorbed into your body, particularly into that particular chakra. It's a way to actually direct where you want that energy to go and what you want to enhance in your life. We had great experiences with putting colored candles matching each different chakra on each corresponding chakra on the case. We lit the candles at night and not only was it beautiful to look at, it amplifies the energy in the space. This is powerful when doing crystal mediations and can support experiencing crystal grids.

Sacred G Fusion walks you through the internal and external systems represented in the Tree of Love. Sacred G Fusion combines all of these systems and moves you to the top point, the singularity point we are all one, we are all already perfect. This conversion system of energy exchange moves us forward to really manifest the type of life we want and that sense of perfect flow.

Click on the Category section for the Tree of LOVE to learn more about each Chakra.

Basic Instructions for Using Sacred G Technology

Select something that you eat or drink or anything you would like to charge and put it directly on the Sacred G design. If you want to charge a bottle of water, simply place it on top of the Sacred G. Keep the water on the design for at least an hour although it's great to experiment with the different lengths of time between one hour, several hours and overnight. Whichever Sacred G design you use, it will charge the water to target specific regions of the brain and body. For example, if you charge the water with the Sacred G Spot, then drink the water, that energy is going to focus on the hypothalamus and is going to help detoxify and cleanse your cells. If you charge it up with the Transporter, that water will help you purge and process negative emotions and emotions that are stored in past memories.

If you want to make Sacred G even more powerful, you can use more than one. Using a bottle of water for example, put one design on the bottom and one on the top of the bottle. This creates a pressure field that will enhance the charge of the water. To create an even more intense field, you can enclose the water on six sides. One way this can be done by taping 4 stickers together in a square and putting one on top and one on bottom of the water. This will supercharge your water, supplements or anything you decide to charge.

You can laminate the stickers, posters or print-outs it doesn't hurt at all. In fact, if you're going to laminate them I would suggest printing 3 or 4 sheets with the designs printed on both sides. If one-sided, put them back to back so you have the design facing out on both sides and then laminate several sheets. The more of them you use, the more energy generated which increases the amplitude of the wave or of the frequency Sacred G puts out. If you want to increase the energy, simply add more. Sacred G lasts forever if the designs aren't damaged. Don't cut any of the designs, the pressure systems and mathematics built into them will be affected and you'll have energy leaks. Print them in color not black and white. Put the Sacred G in your closets, on your shelves, in your refrigerator (best if laminated or you can tape them underneath your shelves so they don't get dirty).

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

35 Ways To Boost Your Energy With Sacred G

To see the Online Version, click… http://coryherter.com/?page_id=594

The following uses are listed in order of importance, go down the list and place Sacred G in every area of your environment to get the maximum benefits. I find it best to place them in ways where the technology will not get damaged. For instance, in the closet, you can tack the Sacred G to the wall behind your clothing. In your fridge, tape the Sacred G to the sides of fridge and underneath the shelves. Be creative, remember that Sacred G Fusion is by far the most powerful Sacred G Technology and is available for only $1.00 / sheet.

Mattress

Place the Sacred G technology between your mattress and box spring to encapsulate your self in the energy all night long. You can also place them directly on the floor. If you have cases you can also place the cases directly on the floor. Sacred G enhances anything that you are doing. While you are sleeping your body is working to heal itself and refill your energy. This is the optimum place to use Sacred G technology and enhanced energy and is the long-term solution for boosting energy permanently. Every night, your energy will continue to increase.

If you are looking to heal your body and activate the Law of Attraction in your life, continue adding Sacred G Fusion underneath your mattress. The more Sacred G you add, the more powerful your experience is going to become.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Water

You can create unlimited amounts of pure energized water using Sacred G. Programmed to a level that no other vibrational energy method can compare to, this water is extremely efficient in hydrating the body. Using large amounts of Sacred G love chargers can energize your water so much that you can use it as a dietary supplement to replace 1 meal a day. Take a big drink of sacred g water whenever you are hungry. Charging your water is vital to reducing any negative cleansing effects from releasing toxins. This also helps to hydrate the body increasing the amount of energy you will feel.

Food – Fridge, Freezer & Pantry

Placing the Sacred G Fusion posters or Sacred G stickers inside your fridge freezer and pantry will lengthen the lifespan of your food help you from over eating and help you lose weight. Use tape to attach the posters to the sides and back of the fridge or on the shelves to prevent damage and get the best results. Charging your food will increase the absorption rate reducing the amount of energy the body has to spend to digest. Cleansing the digestive system will be one the first things to happen when using Sacred G. This will free up your energy, reduce fat and enable your body to start eliminating toxins. Charging your food will help prevent additional toxins from entering in your body and prevent waste buildup in the intestines.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Fruit bowl and Bread Basket

Extend the life of your fruit vegetables and bread by sticking them on top of Sacred G technology increases taste energy and fills you up faster. Eat as many fruits and vegetables as possible for optimum levels of energy. Replace sugar with natural sugar from fruits this is especially a powerful practice for children and eliminates emotional outbursts, rage attacks and fighting with family and friends, simply from eliminating sugar from the diet and replacing it with fruit symptoms of ADD/ADHD, depression, bipolar, MPD and more have disappeared.

Health supplements

VITAMIN
200 I.U.
Dietary Supplement
10 Capsu

Supplements and nutritional shakes are being used by millions of people on a constant basis to improve health. Double or triple their effects by having Sacred G Fusion technology under your supplements all times use herbs to direct where in the body you want the Sacred G Fusion energy to go. This is a powerful method to really focus where in your body you want more energy and can be used for accelerated healing.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Clothing

Your clothing is the first line of defensive between you and your vibrational environment place Sacred G Fusion posters in your closet and drawers to keep your clothes boosted with energy this will keep you flying with energy all day long. Most laundry detergents are filled with chemicals that are no good for your body and rub into your skin this can add to the toxins that are put into your body and decrease energy. Find some natural and green laundry detergent to use this will help in a subtle way to keep your energy boosted. All of these green decisions add up to massive boosts of energy and overall health.

Jewelry

Magnetic and ionized bracelets are a popular tool to use to boost energy and are great for pain relief but they also cost hundreds of dollars and they don't have much style. With sacred g fusion technology you can super charge all of your jewelry to keep your energy boosted all day. Keep your energy boosted all day simply by charging all of your jewelry while you sleep with Sacred G. Charging your jewelry is a powerful technique to help with emotional stability.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Cell Phone

Place the sacred g micros on your cell phone to reduce the heat up to 87%and reduce the negative effects of emfs emitted by your cell phone at the same time.
To Learn more about Cell Phone use and to download your Free Cell Phone Sacred G Micros, visit http://www.freeshield.com

Computer

Place Sacred G under your computer monitor and under your keyboard this works great to reduce the brain drain feeling Sacred G can also help reduce the negative effects caused from emf's. Try placing some Sacred G posters around your desk area can increase work productivity by 300%. Stare at the Sacred G Meditation Station to physically and mentally give your self a break from work and energize yourself do this once or twice a day to accelerate the positive benefits of Sacred G.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Personal Hygiene Products

A lot of products that we put on our body everyday are filled with toxic chemicals that poison our body. The more toxic your body is the, more fat the body needs to neutralize and store these toxins. Search around for natural and green products to ensure your not accidently poisoning your own body. Energize all of your beauty products soaps lotions etc… to boost your energy. Charging your products with sacred g will also help the body with eliminating any chemicals that are not beneficial to your body. Remember your skin will digest anything that you put on it. So anything you put on it you should be able to put in your mouth.

Hot Water Tank

Try placing Sacred G Fusion technology around the hot water tank. This energizes your water and is great for bathing and showering as well as cleaning dishes and doing laundry. Notice how soft your skin becomes.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Vehicle

Place Sacred G under your seat when you are driving reduces pain and stiffness from long periods of sitting. This technology is priceless for truck drivers taxi drivers or anyone else who spends a lot of time in their vehicle. Try putting Sacred G stickers on your gas tank, gas tank lid and fuel filter for increase gas mileage.

Picture Frame

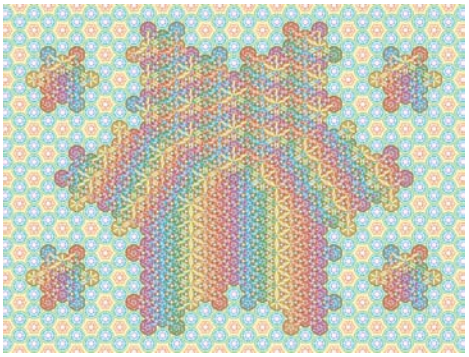

Place 1-20 Sacred G Fusion posters in one picture frame to create a massive blast of energy for your room or office place the picture frame for where you want the energy to go. Great for the office the practitioner room and of course the living room and bedroom for maximum energy and productivity. This will help to keep the energy in the air clear and Transcend negative emotions that can cause drama.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Air Ducts and Fuse Box

Use Sacred G Fusion posters in the air ducts in your home. This is a powerful tool to charge up all the air in your house and enable all of your family to benefit from the positive vibrations emitted from Sacred G. Place Sacred G in your fuse box to encode the electricity that enters your home with positive vibrations using more Sacred G fusion will give you stronger benefits and more noticeable differences use one Sacred G Fusion case for this if possible.

Directly on the Body

This technique is extremely for eliminating pain and focusing where you want the body to heal simply face the Sacred G design towards the area of the body you want to increase energy and eliminate pain sleeping on Sacred G Fusion has been known to eliminate chronic back pain within days.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Couch Cushions

Place Sacred G Fusion under couch cushions or directly underneath the couch. Boost your energy and enable you body to relax when you are watching tv, movies or playing video games. You will find that you can become more immersed in the shows you are watching and the emotional reactions will intensify and heighten your experience. This is a great tool for kids. Using Sacred G this way has actually been known to cut down on the amount of time spent sitting in front of the tv increasing the amount of physical activity of children.

Vision Boards

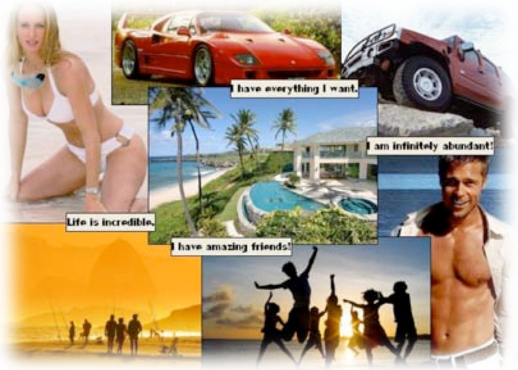
I have everything I want.
I am infinitely abundant!
Life is incredible.
I have amazing friends!

Have you ever heard of a vision board? Simply cut out a bunch of pictures and words that you feel passionate about. This is a physical way of moving towards your goals that you want in your life and takes goal setting to the next level. Attach some Sacred G to your vision board to add extra energy to your intentions. This works great for accelerated manifestation if you sleep on Sacred G, place a vision board between your mattress and the Sacred G. This will program the energy to start building the life of your dreams. The energy will pass these instructions into your cells and start the materialization process.

Backpack

Place Sacred G Fusion (Folded Up) in your backpack to charge your lunch and books charging your books with Sacred G technology increases your ability to retain what you read by 300%. Use Sacred G as a bookmark to help you read and absorb the information more effectively.

Flowers and Plants

Help prolong the life of your flowers and plants by placing the Sacred G design underneath the vase. Try using different designs and see if you have different results then share your research with us. Place Sacred G designs under your plants to help them grow strong and come back to life if they are dying.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Massage Oils / Lotions, Shampoo & Conditioner, etc…

Charging massage oil can turn the oil into a powerful pain relief solution. Add some Fusion 137 Essential Oil to increase the Healing Benefits of products that you use everyday.

Crystals and Precious Stones

Certain stones that you hold after charging with Sacred G will generate a massive influx of energy and sensation in your body. Experiment with different types of stones and how each different type of crystal will create a unique feeling for you. Crystals pick up all kinds of vibrations from the environment so it is important to clear and activate them with Sacred G for maximum energy output.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Meditation

Use Sacred G for meditation by sitting or laying down on several Sacred G designs. You can charge up several crystals or rocks and place them on your body, listen to some nice music and let the energy take you for a ride. Burn Candles on top of the Sacred G to take you even deeper.

Yoga mats

Place the Sacred G designs or Sacred G Fusion posters underneath your yoga mat to instantly increase your strength and flexibility. Sleep on Sacred G Fusion to increase the healing at night reducing pain from physically training your body. Sacred G can be used to support physically demanding exercises and to accelerate healing as well.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Extreme Performance Enhancement

Place the Sacred G on skateboards surfboards team uniforms or any gear or equipment you use. Watch as your performance is enhanced and your game is improved. Take your performance to the next level by using Sacred G in any and everywhere you can.

Candles

Charge your candles with Sacred G to fill the air in a room with positive vibes. Try this method during or after there has been a big argument in a room to help clear the air and watch as it helps solve the conflict more peacefully. Charging your candles helps boost your concentration helps clear your head and inspire new thoughts to come through. Using Candles with Sacred G is great when working on any creative or artistic activities. Use Sacred G with candles to enhance meditation enabling you to access deeper states of mind. Natural and alternative health practitioners love this simple technique when working with clients as it keeps the energy moving and the room free of stagnant energy.

Contact Lenses / Glasses

Place Sacred G under your contact lenses and glasses and cleaning solutions to make them last longer. Some people have reported disposable lenses lasting 3 to 5 times longer. Using Sacred G stickers placed directly on your eyes for at least 15 minutes a day will awaken your ability to see energy and energetic fields that surround everything. There are layers upon layers in the reality that you exist within when you can amp up your energy enough your body will unveil new layers for you to see for yourself unlock these new layers of reality through Sacred G.

Kitchen Utensils

Charge all your dishes glasses and cutlery to help charge your food even more they will pass the chare onto your food and drink.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Cleaning Products

Activate your cleaning products with Sacred G so they do a better job. The emf's (electro magnetic fields) from power lines program the molecules of these substances to group together. Sacred G splits these clusters apart meaning that the cleaning products can penetrate deeper to enhance their natural cleaning properties. Most cleaners are filled with toxic chemicals try finding ones that are green and natural to help keep your body clean and your energy high

Animals Love Sacred G

Pets will notice the effect of Sacred G instantly. They will start sleeping on your pillow/bed. Cats will chew on the stickers trying to get into where the energy is coming from. Turtles will swim at hyper speed. Pets are one of the easiest ways to see how powerful Sacred G technology really is. Place sacred g under your pets food and water dishes as well as under the area where they sleep. Notice the youthfulness and the increased activity in your pet within days. Place the Sacred G Fusion posters behind the fish tank to keep the water clean 3x as long.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Improving Eyesight and Multi-Dimensional Vision

Lose track of time by staring into the design and wondering how it changed your life so quickly. Sacred G has been known to improve eyesight as well through regular practice of this technique (15 to 30 minutes a day) Sacred G Fusion will also produce some interesting multi-dimensional visuals along with heightened states.

Bath Tub

After a hard day or an extreme workout, the body needs to relax and unwind. Charging up the water in the bathtub by sticking some Sacred G stickers on the bottom will ensure that the water is helping the body heal and increase energy. Trying adding some bath salts to clear out lactic acid. Add some Sacred G 137 essential oils to help give your body that extra boost. Water charged with sacred g will make your skin silky smooth by breaking apart the clusters of water molecules apart enabling a deeper cleanse. Try charging a 5 gallon jug with your Sacred G Fusion and adding one gallon to your bath for an extremely powerful experience

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Pictures & Photo Albums

Have you ever had an experience when you were thinking or talking about someone and then they called? We have all had an invisible connection and we have all felt it. Harness this connection with Sacred G by pumping up other people with energy by placing their picture on some Sacred G. Try it out, the effects might freak you out when you see just how powerful Sacred G can be. This is a perfect tool for when you are having problems with someone and don't know how to solve or overcome it.

Business Cards

Direct the energy of Sacred G into your business by placing a business card face down to increase the growth and wealth. You may notice instant effects from this. This is the next level of simply having a positive intent. A physical action is always more powerful than just a thought. Be open to new ideas and direction in your company if you need to make more connections and build a team use, the simulator. If you are looking for new systems or new ideas to govern your company, use TranceNdance. If you wish to accelerate the momentum of your company use the Activator. If there is too much drama in your company, be sure to use the transporter. Add Sacred G Fusion to your life for an increased focus in wealth building.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Gardening

Charging your seeds before planting them. Your food will taste better last longer and fill you up faster. The amount of energy influences every generation so be sure to replant the new seeds. You can also keep your seed immersed in this field continuously by putting Sacred G stickers or minis that have been laminated directly into the dirt. Experiment and see what kinds of results you get and then let us know. Try placing Sacred G under your sprouts if you make them at home.

Positive Influence

Wish you could change the world in a really big way so people were happier and having more fun? Place pictures of authority figures on Sacred G to give them some positive vibes. Fill them with positive energy so they can make more positive decisions that will benefit all of us. Works great for corporate logos as well.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Sacred G Food Test

In the morning take an apple or a banana and place it in between two Sacred G designs both facing toward the piece of fruit. Leave the fruit sitting all morning, at lunchtime eat the fruit before your regular lunch and feel how full you become. This test will show you how charging your food with Sacred G will fill you up faster. Charging your food is a vital part of boosting your energy even higher. Using Sacred G to charge your food also helps with weight loss.

Sacred G Temperature Test

Place your hand on or a couple of inches above the Sacred G technology. Most people can feel a breeze of cool or warm air. After the temperature stabilizes you will feel tingles and sensations all over your arm and then your body temperature is the first system the technology will stabilize for optimum energy flow. Charging your clothing with Sacred G will help you remain at a comfortable temperature through out your day.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Sacred G Full Body Energy Cleanse

Stand with each foot on one sacred g design then place another Sacred G design on top of your head feel the energy trickle out of your fingertips as you are bathed in energy now add two more Sacred G designs between your hands feel the buzz. This test is a quick energy booster and cleanser as it builds up a pressure of energy throughout your entire body.

The more places you place Sacred G Technology, the more energy you are going to have and the better you are going to feel. Have fun, experiment and see what you can discover. Energy is a playful system that utilizes imagination to its full extent. The only limitation energy will ever have is the limitations that you put on it through a lack of imagination.

When you are ready, be sure to upgrade to a case of Sacred G Fusion for the maximum benefits of this incredible technology.

The following section will help you to understand the processes that take place when beginning your use with Sacred G Technology.

What to expect when beginning Sacred G Technology

"Karma – Whatever you put in your body, you are responsible for cleaning out!"
~Humbled State of Consciousness

10 Most Common Forms of Cleansing

Cleansing – As the cells in your body dump these toxins out, you can get headaches, feel tired or woozy, your head can feel congested and foggy. These are all signs that Sacred G is working and clearing out the toxins from your cells… Keep Using the Sacred G to complete this process or try adding more to speed this process up further and cleanse even deeper.

1. Initial weight gain can happen if your organs are on toxic overload… your body will take the toxins and store them temporarily in fat cells while your organs are cleansed and strengthened. Then the body will begin releasing the toxins so your organs can properly cleanse and clear them out of the body. You will also notice that your body will continue to lose even more weight, as your organs will continue this purging of toxins eliminating the need for fat. Fat is designed as a safe place to neutralize and store dangerous toxins. The cleansing effects of Sacred G make this technology a powerful tool to add to your weight loss program.

2. Addictions to toxic substances such as drugs, illegal or pharmaceutical, sugar, coffee, yeast, fluoride in water, pesticides on your food and more are continually creating a toxic wasteland… Inside Your Body. Is it any wonder that it becomes difficult to feel good when practically everything in our world is full of processed chemicals that are toxic? Although Sacred G initiates the body to clear these toxins out of your cells, it is your responsibility to stop putting these toxins back into your body. You will notice that your tastes will change and eating organic and healthy will come easily and naturally if you are charging everything you put in, on and around your body with Sacred G. Please keep in mind, that this also means no more putting toxic substances on your body either.

3. Old injuries often come up from using Sacred G Technology. Although an injury heals, the healing is not 100%. As your cells regress back to their youth, anti-aging, they are physically releasing memory. These old injuries can even have physical bruises show back up, can be very painful and can also spur a lot of agitation as well. Stay with it and try using

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Sacred G Fusion Essential Oil on the injury to accelerate the Healing Process. This is a natural process of Sacred G clearing the cellular memory and restoring your cells back to 100% health.

4. If you had drug addictions or taken pharmaceutical drugs for a long period of time, these drugs will release out of your cells and you will feel their effects. You can have urges, withdrawals, mood swings, etc… that is because these drugs are coming out of your cells and hitting your blood stream. Stick with the process and drink lots of charged water.

5. Speed up this entire process by taking the initiative and doing some cleanses that you can pick up at your local health food store. Check the labels and educate yourself on what is healthy for your body and what is not. Don't put anything on your skin that you wouldn't eat. Do yourself a favor and get off the medications. If you feel the detox, stick with the Sacred G… this is Karma, what you put in, you have to take out. And remember to take a cheat day every once in a while, you are still human and need a chance to celebrate from time to time.

6. Nightmares can take place as your body purges the most extreme form of emotional energy from your past. You can wake up in wet sweats as your body simultaneously releases toxins while you are emotionally purging your system. You can also wake up triggered or even aware of patterns that you are playing out that are causing yourself and others a great deal of pain. All of these emotional triggers can be difficult to go through, yet peace is on the other side of this chaos and it is worth it to stick through it.

7. Pretty much every single person will go through a 1 – 3 week rage attack as they wake up and realize they have living their life for everyone else and not themselves. This is a process of re-activating a person's own personal power and is an amazing feeling and confidence booster once this process completes.

8. Deep anxiety can come up as well once you realize that you have been playing the same patterns in your life, your whole life and you have never really been happy. This usually ends up in dissolving old relationships, friendships or jobs that are not serving you. Although these changes can sometimes make for a difficult transition, sticking with them and going through the process is rewarding in itself when you are on track with your life, your heart and your purpose once again.

9. Little bursts of emotion can come out at anytime of the day for the first while on Sacred G Technology as memories are busy processing in the back ground of your brain. It seems as quick as they come, they go just as quickly too. These emotional surges are the body's way of completing the

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

last conscious processing step and are necessary for a complete release. The majority of the time you won’t even know what the emotion is from.

10. If you have a lot of deep seated emotion come up from never expressing how you feel, the best thing to do is to lay on your Sacred G Fusion Bed Set as often as possible to help you purge this emotion and free up this energy.

Rewards from Cleansing– When you begin to open up your body and feel good again, you will remember that the body is literally the most important thing in your life. As the toxins disappear, your respect for your body and other people’s body’s increases. You will automatically treat others with more respect, love and appreciation boosting every relationship you have. Eliminating toxins automatically increases energy, health and gives a youthful appearance and is a powerful anti-aging tool. Your body is your temple and cleaning the toxins out of it is the first step. Ensure you charge everything you put in, on and around your body with Sacred G to generate the most amount of energy and to assist you in some of these life changing positive patterns.

Clearing the massive amount of emotion stored in past memories is an extremely powerful method in freeing up energy for your body to heal naturally. You will notice a major increase in energy as well as an overall boost in peace, joy, happiness and excitement as old feelings like resentment, anger and sadness are no longer there to weigh you down. As you clear the major emotional blockages, you will find that you will be able to pick up more subtle sensations in your body. This means you will be able to read and communicate with people more effectively as you will be able to sense and feel what is going on with them more clearly. Making love intensifies as your physical senses become more alive. Even being able to smell the fresh morning air after a rain becomes more exhilarating as your senses magnify. Clearing the emotions that have become congested in your brain and body not only boost your energy but enhance your entire experience of life by heightening your senses. This is your body, not just your body, but your very vessel that you experience life with, Unlock your body’s full potential with the Sacred G Fusion Bed Set and discover what life can truly be like.

3 Stage Development Process

Sacred G technology has three major stages of development that it will move you through. Each of the stages has their own unique attributes. These development stages are all required to go through before you will truly stabilize from the increase of energy within your body. Once you stabilize, your growth and development will feel like it has slowed as you integrate your life into that new level of energy.

The first stage of development is a complete body cleansing and rejuvenation. This initial part entails that you clear personal memories and issues in your life that you are experiencing right now. These are some of the typical experiences that you will endure.

- Increase of energy – This will take place if your system is slowed down.

- Decrease of energy – This will occur if your system is too accelerated.

- Non-stop dreaming all night – This is the body's way of cleansing and adding programming through experiences.
  - You can feel tired after these nights or be bursting full of energy.
  - You can experience nightmares if you are spacey or lack focus.
  - You can experience powerful goals and realizations to get in touch with what you really want.

- Deep sleeps - You will have incredibly deep and long sleeps. This is the body working through some very deep issues and has decided that it is better for you to not get back into old memories.
  - You may feel really tired all the time and want to sleep, sleep, sleep.
  - Listen to your body and sleep as much as possible on Sacred G technology. The body needs its rest. The more you do so, the faster you will move past this initial stage.

- You can feel very tired or emotional as the body is trying to clear and reprogram new emotions to stabilize your internal system.
  - Lay on your bed when this occurs and you will notice that Sacred G technology will accelerate this process and rejuvenate you.
  - Make sure you are drinking lots of water that has been programmed.

- You can experience major bursts of energy that you will have no idea what to do with.
  - When you feel this much energy, use it how you would like, most people realize that they still cannot do anything with this energy.
  - Lay on your bed and let Sacred G technology stabilize this energy so that you can speed up the process of this stage of development.

- Old physical injuries or traumas can resurface as the body clears out the cellular memory.
  - Bruising, pain from scars, bone breaks, injured joints, etc… can resurface as the body clears.
  - Lay on the bed and let Sacred G technology accelerate you through this process.

- Physical cleansing sensations may occur such as excessive need of the bathroom.
    - This is part of the physical cleansing process. As the energy of the body increases, pollutants and toxins need to be removed from your body to fulfill the vibrational instructions.
    - Drink lots of water and lay on your bed to speed up this process.

All of these effects may or may not take place depending on what your body needs. You just have to play it through and realize that most effects will only last 1 to 3 days. This first initial body cleansing period is the hardest part about using this technology. The more Sacred G technology you use, the longer it will take for this first stage to complete itself.

You have to think that you have a whole lifetime of memories and physical abuse that you have put your body through. This accelerated cleansing is set to clear as smooth as possible. Some major changes do have to be initiated in this stage though that is necessary to get you where you want to go. Sacred G is the ideal technology to use for this type of cleansing because they give the instructions to your body and let the body use its ability of quantum computing to know what needs to be cleared and when. It is the most efficient system of cleansing that I know of and far surpasses the standard body cleanses and psychological healing.

Once you finish this first initial stage of cleansing, you will notice that your life will dramatically calm down. You will feel powerful, strong and centered. Your body will feel stable and strong.

The second stage of development is vital to setting up new goals and getting in touch with where you want to go in your life. Your desires will be brought to the surface so you can set out a plan of action for yourself. This is the major rewiring and reprogramming part of Sacred G technology.

- Excessive vivid dreaming – Dreams are a way to help bring awareness of beliefs and ideals that you have inside of your head. Dreams are used to train your brain and awaken new ideas and concepts for what you want in your life.

- You will have motivation to get certain things done, yet you will have no energy for it. This is the second stage of getting in touch with what you want. Make notes and create a plan for your life.

- Your information, beliefs, likes and dislikes, attitude towards certain information, activities and people will start to be questioned. Those aspects not in alignment will feel like a waste of time.

- You will go through a level where you will feel like everything is pointless and not worth the effort. This is initially cleansing out everything that once had value and will eventually replace it with strong passions and desires about what you want. This eliminates the clutter of information that is not relevant to your true goals and aspirations.

- The void is introduced as the brain clears out all objectivity and rationality as you move through one reality into the next. Subtle depression and feeling lost can result in this stage. When in your blank stage, keep notes on what still makes you smile and you can have foresight into where you are going next. Lay on the bed with Sacred G technology so that you can accelerate yourself through this process.

- Connections will then be made once all information has been deemed irrelevant. This will construct a new plan and new ambitions for your self. This will clear your mind of all irrational thought and you will feel the precision in your choices. Construct a plan for what you really want to do in your life and prepare for it. Remember that you are alive and that you can do anything that you want.

The second stage of development can be difficult for some people and completely rejuvenating for others. This is essential for you to realize that everything in your life is irrelevant. This cleansing is what allows you to start functioning on the variable vibration program. Afterwards, you will be able to adjust your information system within so that you become more flexible and increase your adaptation skills. This is part of the training to initiate your ability to flow into the level of pure performance.

The third stage of development is where you will begin to implement the new changes in programming into your life. This can and will result in you completely restructuring your life. This can be a fun and exciting adventure, as well as a difficult one as major life changes are introduced. This level is where you will start to see the power of Sacred G technology as it shifts your external world to get you on track with where you want to go.

- Initial cleansing of relationships can begin.
    - This is to cleanse your external environment for what is working and what isn't working. This can be very difficult depending on the type of relationship changing. The easiest and fastest way to overcome this level is to meditate by laying down on your bed with Sacred G technology.
    - Old relationships that serve no real purpose block you from creating new ones. They take up time and energy.

    - Some old relationships can be strengthened and renewed with your new sense of purpose and direction. The body has major key changes that need to occur in your life to get you to where you are going. What needs to be there will be there stronger than ever.

- Business and Career Restructuring – These changes are a major part of your life. In order to reach the point where you want to be. Certain changes need to be made. When you feel like something is not right about a situation, take note and start preparing for a change to happen.
    - Everyone in the world wants to feel worthy and support the whole. 1/3 of our life is devoted to serving others. You need to make sure that you are serving in the ways that you want for optimum happiness and self satisfaction.
    - Remember to think ahead and do not make any major changes by impulse alone. Plan, prepare and then act at the right time to initiate change the smoothest. We will all learn this lesson the hard way at first. Sacred G technology is designed to train you to do this properly. Remember not to be hard on yourself; we learn by mistakes. Sacred G technology is very resourceful at making whatever connections are necessary to help you along the way.

- Physical Health – The body wants to increase your physical level of performance so that you can feel strong and confident with your actions. This may not take place until the end of this stage, as the body wants to complete some of the major life changes first.

- New directions and contacts will need to be made in order to accomplish your new objectives and goals.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Research

Infrared Testing with Cell Phones

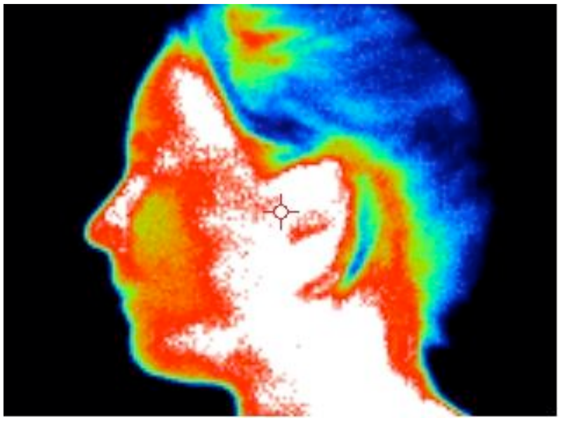
Infrared Test with No Sacred G Micros
White = Highest Heat

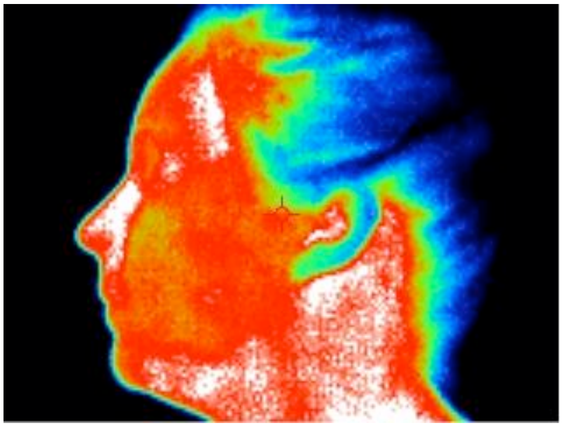
Infrared Test with 4 Sacred G Micros
Red = Normal Tissue Temperature

After being exposed to cell phone radiation for fifteen minutes, the head on the left shows significant levels of heat radiation. Notice the areas of white, which represent the hottest temperature readings.

After fifteen minutes on the same cell phone with 4 Sacred G Micros installed, the head on the right shows far less heat. The red areas show cool, normal tissue temperatures.

After collecting data from numerous infrared thermography trials, there is now scientific proof that the Sacred G Micro Reduces Thermal Heat Radiation by up to 87%!

Now There's Scientific Proof That...
a) Cell phones "cook" your head.
b) Sacred G reduces this effect by up to 87%!

"Without the Sacred G Micro, my cell phone would always get hot, plus I could feel the heat in my head. Now that I put the Micro on my phone, I not longer feel the heat…"

~ Rochelle Coatney, California

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Gas Discharge Visualization Camera

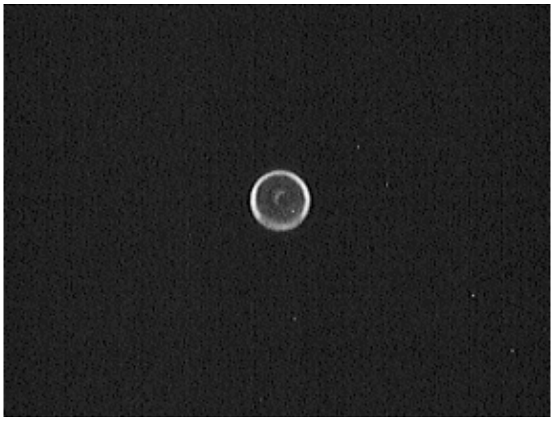

Tap Water

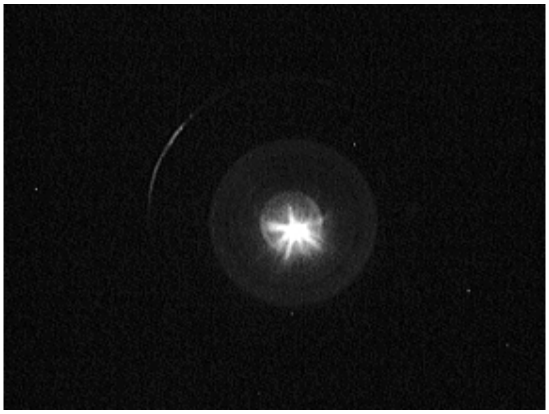

Tap Water Charged with Sacred G

Many people are interested in structured water, which refers to the way in which water molecules align themselves, bond to each other, and interact with each other. Generally, it consists of small clusters of water as opposed to normal, larger clusters of the water molecule. These smaller clusters are thought to improve hydration. Because there are not many effective ways to measure structured water other than secondary effects, the Gas Discharge Visualization Camera (GDV), provides an important way to assess directly the effects of various treatments of water. The GDV Camera, Russian technology that measures photon emission of materials in response to a micro current.

Looking at the above images helps understand what placing a container of tap water on the Sacred G Technology does to the water. A GDV assessment of tap water (left photo) shows that the water has essentially no energy. However, the photo on the right shows that after the same water sat on the Sacred G for 9 hours there was a dramatic increase in the energetic responsiveness of the water. This image has a significantly increased bright corona discharge, a measure of energy. Also, the Charged water has pronounced flares, originating from a few action sites, indicating more highly structured water.

Independent Research Performed by,
Lisa Tulley
Phd. Pharmacology & Toxicology

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Life Span of Flowers

Flowers

Flowers Charged with Sacred G

This is a Fantastic display for the powerful influence Sacred G Technology can have on Flowers and the massive potential Sacred G can have for Anti-Aging.

The above flowers were purchased at the same time and were part of the same bouquet. Three of the flowers were taken out and put into 2 different vases. While the rest were placed in another vase with a Sacred G TranceNdance sticker under it.

These flowers are 12 days old and you can see the huge difference Sacred G can have on flowers. Just think what thousands of the Sacred G Posters could do for your body and mind!

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Food Preservation – Celery

Celery (4 Days Old)

Celery Charged with Sacred G (24 Days Old)

Notice how the celery that is only 4 days old is already starting to wilt and bend over. The celery that was Energized by the Sacred G is still bright, firm and crunchy.

We used to buy fruits and vegetables every week so they wouldn't go bad. Slowly after using the Sacred G our habit shifted and we started buying more Vegetables at one time. Now, we buy all of our groceries for the month, at the beginning of the month, including our fruits and vegetables.

I had bought celery at the beginning of the month; the date was March 4, 2005. It was the Easter weekend and we were having company so I had to pick up some extra food for the dinners we had planned. I picked up an extra stalk of celery from the store. This was on March 24, 2005. The first celery had charged up for 2 weeks. On March 18, 2005 I took the L.O.V.E. Chargers out of the fridge, I was cleaning out the fridge preparing for our move at the end of the month.

On March 28, 2005 we looked at the two celery stalks that we had bought. The one that was charged up for two weeks was in better condition than the one we bought 5 days ago. The one we bought at the beginning of the month still looked fresh and crisp; the leaves were still bright green. The ones we had bought 5 days before had already started to wilt; the leaves had turned a much darker brown and looked terrible compared to the other one.

Here are the pictures - the celery with more gone is the one that I bought at the beginning of the month. The full stalk of celery is the new one.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

FAQ’s Regarding Sacred G

Q. How do I enhance the durability of my Sacred G?

A. Many people have found that laminating the Sacred G is the best way to increase its durability. This is a great solution, especially for using the Sacred G in your refrigerator, with your water, in or under your bed, and any other places where the Sacred G could be exposed to moisture or the possibility of rips and tears.

Q. Can I cut the Sacred G Fusion Posters?

A. No, it is not recommended that you cut the Fusion posters, but you can fold them up to custom fit places like the fridge, freezer, pantry, or any other areas.

Q. I just downloaded and printed out the digital version of the Sacred G, what do I do with it?

A. The best way to use the print out version of the Sacred G is beneath your pillow and/or beneath your water & food. Simply place the Sacred G, design side up, directly inside your pillowcase on the underside of your pillow. This will prevent the Sacred G from moving around too much at night. For a heightened experience with the printable Sacred G, print out as many as you can and layer them between your box spring & mattress beneath the entire length of your body. For your water and food, place the Sacred G design side up beneath your water bottle or glass, beneath your plate of food, in your refrigerator or under your fruit bowl.

Q. I just received a Sacred G Sticker Pack, how do I use them?

A. The Sacred G stickers, which are double sided for extra energy, are great to use to charge up anything you put in, on or around your body. You can either stack them ALL under your pillow for a big boost of energy or you can place just one design for the specific area of your life you want to enhance. You can also use the stickers to charge up your water, food, jewelry and more, click on the Sacred G link on the top bar of the website to learn a more detailed description of each design and over 35 Uses. If you purchased the Variety Pack, you are in for some fun. You can utilize all of the applications above, or use them like a mini tarot deck. Either read the descriptions of what each sticker is for and choose the one the most resonates with what you are going through and place it beneath your pillow, or, wave your hand slowly over each design with your eyes closed and choose the one that draws your attention most strongly. Then place it beneath your pillow, and enjoy!

Q. I just received a Sacred G Capsule containing the full sized 18” x 24” double-sided Sacred G Fusion plus a sticker pack, what do I do with them?

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

A. This package contains either 10 or 20 of the full-sized doubles-sided Sacred G Posters. Depending on what type of bed you have, the rule of thumb with this product is to place as many as you can beneath the full length of your body either under your bed or between your box spring and mattress. If you know that you do not move around a lot at night, the you can place them in three piles directly beneath where you sleep, vertically, so that the points of the stars are North/South. If you do move around a lot, then spread them out evenly throughout the entire length of the bed. TIP: Many people have noticed stronger effects when they physically sleep with their head to the North and feet to the South. For Variety Pack instructions, see above.

Q. What is the best way to charge my food and water?

A. Ideally, you want to be drinking water and eating food that has been charged on the same amount of Sacred G that you are sleeping on. When possible, only eat, drink or use products that carry the same level of charge as your body does. The easiest way to do this is to place your groceries and/or bottles of water directly on the bed for at least 30 minutes before putting them away. Then, make sure you have Sacred G in the fridge, freezer and pantry to help hold the charge. Always carry a few pocket chargers with you for when you eat or drink out. Place the Sacred G Technology under the shelves to protect them.

Q. What is the best way to charge my supplements, herbs, and medications?

A. Same as above; be sure to charge anything you put in, on or around your body with as much Sacred G as you are sleeping on. With herbs, supplements & medications, it is VERY IMPORTANT to remember that the Sacred G will amplify which can increase the effects of anything it charges. Listen to your body and decrease your dosages as your body tells you it is time to do so. Always consult a doctor first before altering any suggested dosages or instructions given by them.

Q. What is the best way to use the Sacred G on the body for aches and pains?

A. Whenever possible, place the Sacred G directly on the body. For example, for back pain, grab a stack of Mini Activators and hold directly against your back, or place between your back and your chair, or lay directly on them. For an even more powerful results, charge up the Fusion 137 Essential Oil Blend we offer and rub it on the agitated area. You can also mix the Fusion 137 Essential Oil in your other products to give it extra healing properties. And last, charge up a glass of water, and when you are drinking it, put all of your focus on the area of your body that requires relief directing the energy quickly to the source of the pain or discomfort.

Q. How do I combine my Sacred G with my vision board or goals list?

A. If you have a vision board, simply put the Sacred G beneath your pictures, inspirational quotes and personal goal list. Sacred G makes a GREAT background for any vision board. In addition, write out, in as much detail as possible, what you desire to manifest in your life, and place the list on top of the Sacred G beneath your body. The Sacred G will blast energy through your list and into your body feeding your body instructions for exactly what action steps to take to turn your desires into reality.

Q. How do I use Sacred G with my cell phone and other EMF emitting Devices?

A. Simply stick the Sacred G directly on the device. We have many different size stickers available for this purpose. The L.O.V.E. Micro is made specifically for cellular phones, but can be used on anything. Some examples include computers, microwaves, even the human body. Did you know that by sleeping on the Sacred G you are surrounding yourself in a shield that protects you from EMF's while you sleep?

Q. What is the best way to use my Sacred G for meditation?

A. Our favorite way to use the Sacred G for meditation is to light some candles on top of some Sacred G stickers, place some Sacred G charged crystals along the charkas, lay on the Sacred G bed and begin by following the G Breath on the Meditation Station at http://coryherter.com/Flash/Gspot/index.html

Q. When I travel, how much of my Sacred G should I take with me?

A. Of course, as much as possible, but due to the weight of the boxes, 25 of the regular sized double-sided "posters" should do the trick. The plastic capsule that the Sacred G comes in is perfect for traveling and fits into most suitcases. Frequently Asked Questions regarding the transformational process that Sacred G Technology initiates. IMPORTANT: Sacred G is an extremely powerful technology that can put your body through major processing to get you cleaned up from the toxins you have been putting in, on & around it over the years. Past events that are left open and have never dealt with will come to the surface to be cleared. Belief systems can be addressed and recognition will have to be made about the choices you are making in your life. In order to achieve higher states of consciousness, you will need to embrace the natural systems of the body and release the emotional baggage and physical toxins that are being stored.

Q. I tried sleeping on the Sacred G and I couldn't sleep. I was restless and tossed and turned all night long. I had to take them out to get some sleep. What should I do?

A. Sleeping on Sacred G will immediately begin processing the stress of your day to day life. If you toss and turn, this is the energy beginning to clear the anxiety and stress that has been suppressed. When you spend your time doing something that isn't taking you where you want to go, anxiety, resentment and frustration start to build up within the body. If you continue this pattern, you will begin suppressing all of this emotion. Sacred G is built to get you back on track with your purpose so you are living your passion. Continue going through this process until it stops. Within days of overcoming this first phase, you are most likely going to make a big shift in your life. This stage usually only lasts a few days, but can last up to 6 weeks depending on the amount of time you have been playing this pattern out.

Q. As soon as I started using Sacred G, my energy skyrocketed. When I added more posters my energy plummeted, what is happening?

A. When you first started, your energy levels increased. After applying more posters, your energy state increased beyond your current state. Once this takes place you have opened up a new stage of growth and requires the body to purify itself. The drop in energy happened because your body went into the internalized cleansing and development stages, which is associated with your core. Study the Tree of L.O.V.E. to identify which Chakra you are blocked on and take the steps to move forward. In time, days or weeks, this phase will pass and your energy will shoot up again, and next time way beyond what you experienced at first.

Q. I have been feeling groggy and spacey the last few days, should I stop using my Sacred G?

A. The body will initiate the physical cleansing of cells in your body. As the toxins are released into the blood stream to be eliminated from the body, you may feel groggy or spacey. These toxins have been in your body for a long time and influence your choices daily. You are simply becoming aware of the chemicals that have been influencing your choices every day. To speed up this process, drink plenty of charged water to cut down on cleansing effects and be sure to keep charging your food. You can also try doing one of the cleanses mentioned on the Tree of L.O.V.E.

Q. My dog is acting like a puppy again and has started sleeping on my bed all the time. Can he feel this energy too?

A. Pets LOVE Sacred G and the results are instantaneous. Youthful playfulness and a calmer attitude towards people are the most common results.

Q. I am dreaming continuously every night, its like watching movies. Does it ever stop?

A. Dreams are the method the body uses to process and clear your internal information and emotions. Advanced levels of dreaming include being taught new information, running simulations for choices that you want to make and opening up some extremely powerful spiritual states. Be sure to have a pen and paper handy when you wake up with downloads for exactly what to do for the day and answers to questions you have been asking.

Q. Why do I need to use more than one Sacred G Poster?

A. Every Sacred G Poster will generate a certain amount of energy. To boost your energy levels higher, simply add more Sacred G. Wait until you see the states of experience you can access at higher quantities.

Q. Can I put the Sacred G directly on my body?

A. Yes. Placing the Sacred G directly on the body can help to eliminate pains, aches, headaches, cramps etc. Try charging up the Fusion 137 essential oil or massage oil with Sacred G and apply it to your body for Pain Relief.

Q. How long do I leave my water, food and supplements on top of the Sacred G?

A. The best way to use this technology is to put the Posters in the fridge, freezer, pantry, medicine cabinet, etc… and leave them there permanently. This ensures the most powerful charge; increase the number of posters to infuse everything you charge with even more Energy and leads to greater results.

Q. What exactly happens to my water and food when it is charged with the Sacred G?

A. There is a mathematical program embedded within Sacred G technology to enhance the cellular absorption strengthening the effects of the energy. When charging your food and water, this programming increases absorption into the cells filling you up faster and enabling you to eat less. This can be a very powerful tool for weight loss. Charging up your food also lengthens the shelf life from the full spectrum vibration that doubles as a stabilizing agent and a natural preservative.

Q. Can I use too Much Sacred G?

A. Sacred G uses a perfectly balanced full spectrum vibration that enables you to continue to increase the quantity as high as you desire. Four people have tested Sacred G at extreme quantities of up to 43,200 Posters for over a six month period. The results are neither incomprehensible nor explainable to someone that has not accessed extremely high states of consciousness, yet I would instantly recommend it to anyone that yearns to explore the consciousness of the universe. There is a safe guard in the mathematics that prevents anyone from opening higher states of energy before they are ready through a precise development system. This safe guard program also incorporates a coaching program to help you develop through this growth cycle when using enough Sacred G.

Q. I am starting to download information that I have never acquired before, where does this come from?

A. Welcome to the new way of learning…I call it the L.O.V.E. Innerversity. This new education system is a powerful Intelligence System built directly into the technology to help you achieve your goals and unlock the potential of the human body. Wait until you start downloading new forms of motion into your body and you watch yourself move. Another fun effect is the ability to speak without knowing what you are going to say and learning from yourself as you say it. This is really fun with jokes because you hear them the first time as you say them. Your body becomes an intelligent being that teaches and interacts with you. This is an Absolutely Phenomenal state of mind.

Q. My sensations are incredible; my love life is amazing…what is happening so I know how to do more of it?

A. Enhancing the senses literally means enhancing your sensations, as the Sacred G continues to boost you to new heights of self acceptance and enable you to open your heart; your senses will become much more intense. Wait until the Tantric energy starts to fill your being and you can feel your spirits merge… oh so many things to experience, I LOVE IT!!

Q. I have spent the last 15 years doing self-development and this technology seems to automate everything, do you think this takes away from people's individual accountability for the laws of cause and effect if they can have their body and mind cleared automatically?

A. The great thing about the technology is that it is not a free ride, you have to train with it and it teaches you to process memories faster, which includes clearing out the emotion, eliminating judgments and opening up to new perceptions. Although the process is automated, this does not mean that the clearing is instantaneous or that each person does not have to be accountable for their actions. In order to learn and develop, conscious recognition will be given to ensure the new programming is locked in which fulfills the laws of cause and effect. What you put into the body, you are also responsible for clearing out, even though it may be uncomfortable at times.

Q. Does the technology have the same effect from looking at it? Should I put it on my walls or on my ceiling above my bed?

A. You can put it anywhere you like… Yes Sacred G does have a powerful effect from staring at it. TIP: Placing Sacred G posters under your bed and on the ceiling above your bed will create a pressurized effect enhancing your results. Place the Posters on your headboard and footboard as well to create an even more powerful effect! I know this might sound strange, but, does the Sacred G answer prayers? I mean, at night I usually go over my affirmations in my head while I meditate, and it just seems like since I have been using the technology, that manifestation, or as I like to call it, my prayers being answered is, well, happening faster. When you override the massive amounts of vibrational influences contained within our environment, the body is able to create your goals more efficiently. Increasing the number of Posters essentially will increase the speed that your intentions or prayers are manifested.

Q. Does the effects and transformation induced by this technology stop or old patterns come back if I stop using the technology?

A. When you stop using the technology, your body will re-harmonize with your environment. This is the drop in energy that is experienced. In this way the technology is used to overcome the negative vibrations created from electricity With the additional mathematics in the technology, this energy has the ability to help you process the emotions from past memories and help solve conflicts in your information unlocking new information. All of these benefits stay with you, it is impossible to unlearn them.
The process that the technology uses is a natural occurring cycle in the body. The technology is only accelerating this growth cycle and helping to prevent you from becoming stuck. As you clear more of the blocks from the past, your spirit can shine through brighter than ever before. The technology gives the body an artificial boost of energy to take you to a higher state that you are not currently at.
The physical body is required to create this temporary energetic state into a solid reality, which initiates the development procedures that Sacred G will take you through. Over time, the level the technology boosts you up to become an integrated system, and you will never drop down in energy even if you stop using the technology. Increasing the amount of Sacred G takes you to a higher level and will speed up the development and integration process. Even though the technology is a shortcut, you are still required to go through the process of integration so you can become this higher state of consciousness. The beauty of the technology is that instead of struggling to grow, the extra energy initiates the growth making this transformational process easier which enables more people to step into their potential.

If you have a question that is not answered above, simply email us and we will answer it and then add it to our list of FAQ's. Please send your question(s) to info@coryherter.com.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Video Library

The Video Library is actually categorized in the www.CoryHerter.com website. Below is an outline of all the videos that are included and a brief synopsis for each video. There is a category link included at the beginning of each section as well as individual links to each video.

2006 Video Series – 19 Hours

http://coryherter.com/?cat=73

The 2006 Series are a live seminar series that I offered in Santa Monica, CA. These seminars were one of my tests that I needed to complete to pass the TranceNdance design and get the upgrade to the Sacred G Simulator (also known as the Love Charger). The Sacred G Simulator includes the mathematics for expression, you can see the before and after results in me from using the Sacred Simulator. The Perfect Wave Programmer talked about in this series of Videos is now called Sacred G TranceNdance. I hope you enjoy these videos, please leave a comment on the bottom of the videos to share with other people your experience.

2006 - Introduction to Sacred G Technology – 2 Hours

Cory explains Sacred G Technology back in 1996 when the technology was called Perfect Wave Programmers. This was before the simulators, activators and fusion designs were developed. The Perfect Wave Programmers were an early version of the Sacred G Trancendance design.

http://coryherter.com/?p=771

2006 – Creating the Council – 2 Hours

Cory explains the process for splitting apart your identity and growing up the fragments into your own internal council. This is a powerful tool for understanding how to unlock massive amounts of information in your brain. Sacred G Fusion has this system built into it and is a piece of what Cory refers to as the advanced memory systems, such as complex pattern recognition.

http://coryherter.com/?p=769

2006 - Creating the Collective – 1 Hour

The Collective is something that we must create within our own energy systems in order to gain access to the collective that exists in deeper layers of the human body. Cory takes creating the council to the next level in this incredible video. The collective is now a system completely integrated into Sacred G Fusion and activated by the Pineal Gland.

http://coryherter.com/?p=767

2006 - Advanced Sacred Geometry - Part 1 – 2 Hours

Cory shares some of the advanced Sacred Geometry science behind Sacred G Technology.

http://coryherter.com/?p=761

2006 - Advanced Sacred Geometry - Part 2 - Time & Space – 2 Hours

Cory shares some of the advanced Sacred Geometry science behind Sacred G Technology.

http://coryherter.com/?p=759

2006 - Life After Death - Part 1 – 2 Hours

Cory shares the Science behind Life After Death using Sacred Geometry. Watch the Science of LOVE Video Series first to gain a deeper understanding of the information given in these videos.

http://coryherter.com/?p=746

2006 - Life After Death - Part 2 – 2 Hours

Cory shares the Science behind Life After Death using Sacred Geometry. Watch the Science of LOVE Video Series first to gain a deeper understanding of the information given in these videos.

http://coryherter.com/?p=744

2006 - Life After Death - Part 3 – 2 Hours

Cory shares the Science behind Life After Death using Sacred Geometry. Watch the Science of LOVE Video Series first to gain a deeper understanding of the information given in these videos.

http://coryherter.com/?p=742

2006 - Chironium Healing - Part 1 – 2 Hours

Chironium Healing is an energy healing technique developed by Cory Herter which utilizes breath and sacred geometry for healing physical injuries and psychological trauma. Chironium Healing is a pre-cursor to the Breath of LOVE Workshops that Cory now offers Live in Person. Check the events page to learn more.

http://coryherter.com/?p=740

2006 - Chironium Healing - Part 2 – 2 Hours

Chironium Healing is an energy healing technique developed by Cory Herter which utilizes breath and sacred geometry for healing physical injuries and psychological trauma. Chironium Healing is a pre-cursor to the Breath of LOVE Workshops that Cory now offers Live in Person. Check the events page to learn more.

http://coryherter.com/?p=738

I hope you enjoyed the 2006 Video Series. This 19 hour introduction into my work is a powerful way to open your mind to some of the more advanced concepts I will continue to teach.

2012

http://coryherter.com/?cat=95

Oct 20, 09 Live Webinar – The Truth about 2012

The Truth about 2012 – Evolution, Conspiracy, 9-11, Privacy Act, Natural Selection, Light Workers and more!

http://coryherter.com/?p=951

The 2012 Enigma by David Wilcock

2012: Tragedy, transcension or just another year? David Wilcock exposes many great secrets: DNA, consciousness science, wormholes, stargate travel, sacred geometry, three-dimensional time, the Mayan Calendar and much, much more!

http://coryherter.com/?p=637

Ascension

http://coryherter.com/?cat=99

Merkabah – The Chariot of Ascension

The Merkaba Field is a permanent and highly ordered Merkabic structure of inter-connected electro-magnetic counter-rotating energy spirals that exist as an integral part of ALL CREATION. The mathematics of the Merkabah are built into the New Sacred G Fusion Design to activate the Pineal Gland to Access Higher Consciousness Communication and Inter-Dimensional Travel.

http://coryherter.com/?p=817

Breath of LOVE

http://coryherter.com/?cat=105

LOVE Innerversity – Module 1 – The Breath of LOVE

Cory Shares about the Module 1 - Breath of LOVE course taught at his Live Workshops. Click the Events page to learn about upcoming events near you.

http://coryherter.com/?p=977

#1 Reason For Every Health Problem

Find out what this incredible breakthrough is and how knowing this 1 thing can change your life forever. Cory breaks down every single illness into a completely customized disease that every single one of us creates for ourselves as the result of 1 single problem.

http://coryherter.com/?p=672

Cory Herter

http://coryherter.com/?cat=10

Cory Explains his 300 Near Death Experiences

Introduction to Cory's Story and how he was triggered into over 300 near death experiences.

http://coryherter.com/?p=495

Evolution Into Sacred G Technology

Discover the evolutionary process through vibrational healing technologies Cory went through in order to create Sacred G.

http://coryherter.com/?p=668

Cory Herter on Sacred G Fusion

Cory Herter shares his initial experience with Sacred G Fusion. Video includes a quick overview of the New Sacred G Fusion.

http://coryherter.com/?p=422

Cory Herter's Experience on Sacred G Fusion #2

Cory Herter's experience using Sacred G Fusion Technology. Another addition to my video blog with what I am learning about Sacred G Fusion as I go through this incredible process of transformation.

http://coryherter.com/?p=498

Cory's Sacred G Fusion Bed Set and Massive Sacred G Designs

I am definitely the biggest fan of Sacred G Technology. As you can see, I have my pad here in Las Vegas loaded up with Sacred G everywhere. With this many cases under my bed, my 5 meal a day diet loaded with steak, chicken and fish to add weight for working out has now been reduced to 4 raw seaweed veggie wraps a day along with a glass or two of fresh juice!! I am still keeping my same weight, losing the outer layer of fat and getting totally cut up!! My dream is to have a massive warehouse with the floor and walls covered in Sacred G Fusion Cases, with my bed resting on top of a pyramid of Sacred G Fusion cases like the mayan temples!! I am curious how many cases it will take to reduce the food down to nothing but air... or if that is even possible.

http://coryherter.com/?p=736

Cory's Video Series

http://coryherter.com/?cat=103

Sacred G Sleep Test – 5 Video Series

Learn the different types of Sleep Effects that Sacred G will use to Increase Your ENERGY!!

http://coryherter.com/?p=975

Pineal Gland Activation – Entire Series – 6 Parts

Cory's Sacred G Pineal Gland Activation Video Series.

http://coryherter.com/?p=969

Benefits of each Sacred G Design – Entire Series + Bonus Video

Discover the Benefits and Science behind each Sacred G Design. This is a 7 Part Series with an extra bonus video at the end explaining the Tree of LOVE Case!!

http://coryherter.com/?p=559

Science of LOVE DVD – A 3 Hour Video Series on the Science behind Sacred G

Cory Herter gives an amazing presentation into the math and physics behind Sacred G Technology. L.O.V.E., which stands for the Laws Of Vibrational Energy is the engine that drives Sacred G Technology and is a natural law of the universe contained within Sacred Geometry. Discover the deeper meaning of the Flower of Life, The All Seeing Eye, DNA Activation and more...

http://coryherter.com/?p=399

Crazy Sacred G Fun

http://coryherter.com/?cat=17

Cool Sacred G Video

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Cool Video Highlighting Sacred G Technology. Sacred G is a revolutionary new technology that increases DMT production in the Pineal Gland. Try this technology out for Free today!

http://coryherter.com/?p=264

Underground Sacred G Party

Check out the Highlights from a Massive Sacred G Underground Party with Fire Dancing, Break Dancers, Sacred Geometry Art Displays and More... The Peeps are Loving the Fusion 137 Essential Oil and are calling it the New *Vicks*.

http://coryherter.com/?p=268

DMT

http://coryherter.com/?cat=46

Pineal Gland & DMT by Cory Herter – Part 1 of 6

Discover the Power of the Pineal Gland and its relation to DMT. Unlock your Potential with Sacred G Fusion's Pineal Gland Activation!

http://coryherter.com/?p=420

DMT: The Spirit Molecule, the Movie

A series of interview clips from DMT: THE SPIRIT MOLECULE, the movie. http://www.thespiritmolecule.com/

http://coryherter.com/?p=617

DMT – The Spirit Molecule – Audio Interview with Film Maker Mitch Schultz

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

An audio interview with Mitch Schultz, maker of the film DMT, The Spirit Molecule. We'll be talking to Mitch about the film: who's in it, what he's learned along the way, and when you can expect it's release.

http://coryherter.com/?p=613

DMT, Dreams and Near Death Experiences

Does DMT play a role in near death experiences? Clip from an episode of the magical egypt series featuring Rick Strassman.

Endogenous DMT, produced in the human brain, is involved in certain psychological and neurological states. As DMT is naturally produced in small amounts in the brains and other tissues of humans, and other mammals, some believe it plays a role in promoting the visual effects of natural dreaming, and also near-death experiences and other mystical states.

http://coryherter.com/?p=389

Joe Rogan Talks DMT on the Alex Jones Show

Alex Jones welcomes to the show Joe Rogan to discuss the importance of Dimethyltryptamine or more commonly known as DMT. Anyone that has experianced a lucid dream or have done an extracted form of DMT knows just how important this subject is.

http://coryherter.com/?p=384

Joe Rogan Goes off about DMT

Joe Rogan Goes off about DMT - Joe Rogan recounting his experience with DMT and his ideas on humanity.

http://coryherter.com/?p=393

Fusion

http://coryherter.com/?cat=79

New Sacred G Fusion Bed Set Video

Cory Herter's Journey Through The After Life Sacred G Fusion Bed Set Video.

http://coryherter.com/?p=258

Sacred G Fusion – ONLY $1.00 / Sheet

Sacred G Fusion is now only $1.00 / sheet. Packs of 100 Available Now!! Boost your Energy and Experience the New Sacred G Fusion!

http://coryherter.com/?p=973

Sacred G Fusion Design – 7 of 7

Discover what each Sacred G Design can do for you!! Learn the process that Sacred G uses to unlock your true potential. Finally learn how Sacred G can boost your energy so much! Be sure to download your Free Sacred G Designs at www.CoryHerter.com!

http://coryherter.com/?p=545

Bonus: Secret of the Sacred G Fusion Case

Discover the secret behind the New Sacred G Fusion Case!!!

http://coryherter.com/?p=543

Cory Herter on Sacred G Fusion

Cory Herter's experience with using Sacred G Fusion Technology. A Quick Overview of the New Sacred G Fusion.

http://coryherter.com/?p=422

Cory Herter's Experience on Sacred G Fusion #2

Cory Herter's experience with using Sacred G Fusion Technology. Another addition to my video blog with what I am learning about Sacred G Fusion as I go through this incredible process of transformation.

http://coryherter.com/?p=498

The Collective & Sacred G Fusion

Access the Collective through your 3rd eye. Activate your Pineal Gland and Release DMT Now!!!

http://coryherter.com/?p=629

Cory's Sacred G Fusion Bed Set and Massive Sacred G Designs

I am definitely the biggest fan of Sacred G Technology. As you can see, I have my pad here in Las Vegas loaded up with Sacred G everywhere. With this many cases under my bed, my 5 meal a day diet loaded with steak, chicken and fish to add weight for working out has now been reduced to 4 raw seaweed veggie wraps a day along with a glass or two of fresh juice!! I am still keeping my same weight, losing the outer layer of fat and getting totally cut up!! My dream is to have a massive warehouse with the floor and walls covered in Sacred G Fusion Cases, with my bed resting on top of a pyramid of Sacred G Fusion cases like the mayan temples!! I am curious how many cases it will take to reduce the food down to nothing but air... or if that is even possible.

http://coryherter.com/?p=736

Indigo Children

http://coryherter.com/?cat=89

Indigo, Rainbow & Crystal Children and Sacred G Fusion

You have to watch this!! The Spiritual Revolution is here!! It is time to Awaken!!

http://coryherter.com/?p=625

Akiane: child prodigy, artist, poet, “Indigo child”

An internationally recognized 13-year-old prodigy, considered the only known child binary genius, in both realist painting and poetry. Selected as 1 of 20 most accomplished visual artists in the world by Tribute Entertainment (London) and ABI (United States). Kids Hall of Fame inductee. World Council for Gifted and Talented Children inductee. A best selling author of two books "Akiane: Her Life, Her Art, Her Poetry" and "My Dream is Bigger than I - Memories of Tomorrow"

http://coryherter.com/?p=634

Meditation

http://coryherter.com/?cat=36

Sacred G Meditation Station

Learn the mechanics behind the Famous Meditation Station. Cory explains each component and why the Meditation is so Powerful. Use this process daily to advance through the levels of this meditation and unlock deeper aspect of the self.

http://coryherter.com/?p=674

Go to the Meditation Station Now… http://coryherter.com/Flash/Gspot/index.html

Sacred G Inspirational Video by Quantum Power

Enjoy this amazing inspirational video created by Quatum Power using Sacred G and iTom Lab's Galactic Mantra.

http://coryherter.com/?p=903

Sacred G Power Meditation by Quantum Power

Enjoy the Sacred G Power Meditation created by Quantum Power using iTom Lab's Galactic Mantra!

http://coryherter.com/?p=906

Pineal Gland

http://coryherter.com/?cat=42

Pineal Gland & DMT by Cory Herter – Part 1 of 6

Discover the Power of the Pineal Gland and its relation to DMT. Unlock your Potential with Sacred G Fusion's Pineal Gland Activation!

http://coryherter.com/?p=420

Pineal Gland & Hormones by Cory Herter – Part 2 of 6

Pineal Gland & Hormones by Cory Herter - Part 2 of 6. Discover the Power of the Pineal Gland and its relation to Hormone Regulation. Unlock your Potential with Sacred G Fusion's Pineal Gland Activation!

http://coryherter.com/?p=418

Pineal Gland & Circadian Rhythms by Cory Herter – Part 3 of 6

Pineal Gland & Circadian Rhythms by Cory Herter - Part 3 of 6. Discover the Power of the Pineal Gland and its relation to CIrcadian Rhythms. Harness the Power of your brain to create synchronicities in your life. Unlock your Potential with Sacred G Fusion's Pineal Gland Activation!

http://coryherter.com/?p=416

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Pineal Gland & Sleep by Cory Herter – Part 4 of 6

Pineal Gland & Sleep by Cory Herter - Part 4 of 6. Discover the Power of the Pineal Gland and its relation to Sleep and Healing. Harness the Power of your Sleep to unlock massive amounts of energy in your life and accelerate the natural healing process every night while you sleep. Unlock your Potential with Sacred G Fusion's Pineal Gland Activation!

http://coryherter.com/?p=414

Pineal Gland & Temperature by Cory Herter – Part 5 of 6

Pineal Gland & Temperature by Cory Herter - Part 5 of 6. Discover the Power of the Pineal Gland and its relation to Temperature and Speed. Harness the Power of Temperature, Adrenaline, Metabolism and more for Total Human Performance by Activating your Pineal Gland with Sacred G Fusion.

http://coryherter.com/?p=412

Pineal Gland – The Master Gland by Cory Herter – Part 6 of 6

Pineal Gland - The Master Gland by Cory Herter - Part 6 of 6. The Pineal Gland is the Master Gland of the Human Body and Human Experience. The Pineal Gland is our direct link to source, our body, our environment, health and performance. Activate Your Pineal Gland Now with the New Sacred G Fusion Technology available at www.CoryHerter.com

http://coryherter.com/?p=409

THE PINEAL GLAND IS A STARGATE – 23 PART SERIES

THE PINEAL GLAND IS A STARGATE SERIES - This Video Player contains 23 videos done by incredible beings of Light and they all talk about the Pineal Gland. Sacred G Fusion is the mathematical coding to unlock the Pineal Gland. Learn more about the information behind the Pineal Activation, watch this Series on www.CoryHerter.com

http://coryherter.com/?p=395

Research

http://coryherter.com/?cat=98

Research with Sacred G Technology

Learn about some of the cool research and contribute to some of the projects that are going on with Sacred G Technology.

http://coryherter.com/?p=734

Sacred G Designs

http://coryherter.com/?cat=12

Sacred G Spot Design – 1 of 7

Discover what each Sacred G Design can do for you!! Learn the process that Sacred G uses to unlock your true potential. Finally learn how Sacred G can boost your energy so much! Be sure to download your Free Sacred G Designs at www.CoryHerter.com!

http://coryherter.com/?p=557

Sacred G Transporter Design – 2 of 7

Discover what each Sacred G Design can do for you!! Learn the process that Sacred G uses to unlock your true potential. Finally learn how Sacred G can boost your energy so much! Be sure to download your Free Sacred G Designs at www.CoryHerter.com!

http://coryherter.com/?p=555

Sacred G Ascender Design – 3 of 7

Discover what each Sacred G Design can do for you!! Learn the process that Sacred G uses to unlock your true potential. Finally learn how Sacred G can boost your energy so much! Be sure to download your Free Sacred G Designs at www.CoryHerter.com!

http://coryherter.com/?p=553

Sacred G Trancendance Design – 4 of 7

Discover what each Sacred G Design can do for you!! Learn the process that Sacred G uses to unlock your true potential. Finally learn how Sacred G can boost your energy so much! Be sure to download your Free Sacred G Designs at www.CoryHerter.com!

http://coryherter.com/?p=551

Sacred G Simulator Design – 5 of 7

Discover what each Sacred G Design can do for you!! Learn the process that Sacred G uses to unlock your true potential. Finally learn how Sacred G can boost your energy so much! Be sure to download your Free Sacred G Designs at www.CoryHerter.com!

http://coryherter.com/?p=549

Sacred G Activator Design – 6 of 7

Discover what each Sacred G Design can do for you!! Learn the process that Sacred G uses to unlock your true potential. Finally learn how Sacred G can boost your energy so much! Be sure to download your Free Sacred G Designs at www.CoryHerter.com!

http://coryherter.com/?p=547

Sacred G Fusion Design – 7 of 7

Discover what each Sacred G Design can do for you!! Learn the process that Sacred G uses to unlock your true potential. Finally learn how Sacred G can boost your energy so much! Be sure to download your Free Sacred G Designs at www.CoryHerter.com!

http://coryherter.com/?p=545

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Bonus: Secret of the Sacred G Fusion Case

Discover the secret behind the New Sacred G Fusion Case!!!

http://coryherter.com/?p=543

Sacred G on MTV

http://coryherter.com/?cat=107

The Full Sacred G Episode on MTV’s Rob & Big

Watch the complete episode of Rob & Big Featuring Sacred G Technology as a Performance Enhancement Tool. Sacred G Enhances the Performance Levels of Athletes doing Extreme Sports by reducing the amount of Electro-magnetic interference from power lines, cell phone towers, satellite signals and radio waves. Neutralizing this Negative Electro-Magnetic Polloution enables the brain to process information faster, increasing reaction time. In extreme sports, 1/10 of a second in reaction time is a very long time and can easily determine whether or not you land your trick, score a goal or beat your opponent.

http://coryherter.com/?p=426

Rob Dyrdek & Steve Berra Share Their Views about Sacred G on MTV

Steve talks about Sacred G being like Steroids while Rob Dyrdek lands a Switch Laser, a Skateboarding Trick he has never landed before!

http://coryherter.com/?p=273

Highlights from MTV’s Hit Show Rob & Big Featuring Sacred G

Watch the Highlights from MTV's Hit Show Rob & Big Featuring Sacred G.

http://coryherter.com/?p=270

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Sacred G Uses & Tests

http://coryherter.com/?cat=13

Simple Instructions and Tips for using Sacred G Technology

Cory explains simple instructions on how to use Sacred G Technology from his years of experience. You might even learn some cool tricks from this old dog.

http://coryherter.com/?p=670

35 Uses & 5 Tests for Sacred G

The incredible thing about **Sacred G** is that it can be used to charge absolutely everything you put ***IN, ON & AROUND Your Body*** with it. As you begin sleeping on your **Sacred G** as well as charging your food and water with it, your body will begin to tell you how else it wants it to be used. We have listed over 35 methods you can use **Sacred G** to enhance your life below.

http://coryherter.com/?p=84

Go to the Full Page of 35 Uses and 5 Tests here by clicking here…
http://coryherter.com/?page_id=594

Sacred G Water Taste Test

The Sacred G Water Taste Test demonstrates how Sacred G Technology can purify your body and mind. Get the explanation for how Sacred G can purify a glass of water right before your eyes.

http://coryherter.com/?p=136

Sacred G Sleep Test – 5 Video Series

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Learn the different types of Sleep Effects that Sacred G will use to Increase Your ENERGY!!

http://coryherter.com/?p=975

Sacred Geometry

http://coryherter.com/?cat=94

The Secret of the Universe

The Secret of the Universe is the name given to the ancient knowledge of the inner workings of the universe. This includes (but is not limited to) the method by which huge megaliths can be reduced in weight and lifted, something also known to the Freemasons and the builder of Coral Castle up to the mid 1900's.

http://coryherter.com/?p=641

Science of LOVE DVD

http://coryherter.com/?cat=37

Science of LOVE Introduction

.

The Science of LOVE Introduction - Learn the Science behind Sacred G Technology with this 3 Hour DVD Special... Available Online for a Limited Time Only!!!

http://coryherter.com/?p=282

Science of LOVE DVD – A 3 Hour Video Series on the Science behind Sacred G

Cory Herter gives an amazing presentation into the math and physics behind Sacred G Technology. L.O.V.E., which stands for the Laws Of Vibrational Energy is the engine that drives Sacred G Technology and is a natural law of the universe contained within Sacred Geometry. Discover the deeper meaning of the Flower of Life, The All Seeing Eye, DNA Activation and more...

http://coryherter.com/?p=399

Sleep Test

http://coryherter.com/?cat=106

Sacred G Sleep Test – Beta (Tossing and Tuning) – Part 1 of 5

The Sacred G Sleep Test reveals which brain wave states you have an excess amount of unprocessed energy. This energy is released from using Sacred G Technology by altering the types of sleep you have. This is the body's natural healing mechanism and is crucial for regenerating the body.

http://coryherter.com/?p=1002

Sacred G Sleep Test – Alpha (Emotional Dreams) – Part 2 of 5

The Sacred G Sleep Test reveals which brain wave states you have an excess amount of unprocessed energy. This energy is released from using Sacred G Technology by altering the types of sleep you have. This is the body's natural healing mechanism and is crucial for regenerating the body.

http://coryherter.com/?p=1000

.

Sacred G Sleep Test – Theta (Educational Dreams) – Part 3 of 5

The Sacred G Sleep Test reveals which brain wave states you have an excess amount of unprocessed energy. This energy is released from using Sacred G Technology by altering the types of sleep you have. This is the body's natural healing mechanism and is crucial for regenerating the body.

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

http://coryherter.com/?p=998

Sacred G Sleep Test – Delta (Deep Sleep) – Part 4 of 5

The Sacred G Sleep Test reveals which brain wave states you have an excess amount of unprocessed energy. This energy is released from using Sacred G Technology by altering the types of sleep you have. This is the body's natural healing mechanism and is crucial for regenerating the body.

http://coryherter.com/?p=996

Sacred G Sleep Test – Brain Wave Conclusion – Part 5 of 5

The Sacred G Sleep Test reveals which brain wave states you have an excess amount of unprocessed energy. This energy is released from using Sacred G Technology by altering the types of sleep you have. This is the body's natural healing mechanism and is crucial for regenerating the body.

http://coryherter.com/?p=993

Tree of LOVE

http://coryherter.com/?cat=87

Secret of the Sacred G Fusion Case

Discover the secret behind the New Sacred G Fusion Case!!!

http://coryherter.com/?p=543

The Three Trees – The Tree of Life, Knowledge & LOVE

Cory shares the science behind the Word of God through Sacred Geometry. The Tree of Life, Tree of Knowledge and the Tree of LOVE are explained in detail. Cory shows how the ascension is attained through understanding the Laws Of Vibrational Energy.

http://coryherter.com/?p=936

Secret to Perfect Health - 1 Problem, 1 Solution, 6 Layers & the Tree of LOVE

Oct 27, 09 - Your Health, 1 Problem - 1 Solution - 6 Layers - Tree of LOVE

http://coryherter.com/?p=962

Video Testimonials

http://coryherter.com/?cat=28

Carolanne & Theresa's Sacred G Experience

Carolanne and Theresa are natural health practitioners and host workshops. Get the scoop on their experience of Sacred G and how they are using it with their practice.

http://coryherter.com/?p=909

David Shore, Sacred G Life Saver

David shares how Sacred G Technology saved his life... Literally!

http://coryherter.com/?p=901

David Caddell Sacred G Testimonial

David shares his experience on Sacred G Technology!

http://coryherter.com/?p=899

Druanna on Sacred G Technology

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

I am so excited about how Sacred G has seriously Changed my life! I no longer have insomnia problems or Heart problems! My good friend Beau Manley from the METAL MULISHA CREW is having wonderful experiences on it as well!! HE CANT GET ENUFF! LOL Looking forward to the Sacred Fusion experience and G PILLS!!!

http://coryherter.com/?p=896

David Shore shares his testimonial of Sacred G

David's story is definitely one that will keep you smiling as he talks about how Sacred G has transformed his life.

http://coryherter.com/?p=859

Pete Lewis's experience on Sacred G Technology

Listen to Pete share about how Sacred G has helped him get off meds and cleared his cat of some emotional baggage.

http://coryherter.com/?p=857

Loriel Hanon's Testimonial from Sleeping on 50 Cases of Sacred G Technology!!

Listen to Loriel's incredible story about how Sacred G helped her clear the trauma of losing her son and restore the positive memories of his life.

http://coryherter.com/?p=844

Taylor Angelos's amazing testimonial from sleeping on a Fusion Bed Set!!

Within 2 weeks of Sleeping on Sacred G Fusion, Taylor is already downloading new information in her dreams. Her ability to recognize patterns in her life and make more conscious decisions has already made a major impact and she is just getting started. Taylor was so impressed with Sacred G that she flew out to Denver to check out the new LOVE Practitioner Course.

http://coryherter.com/?p=842

JoAnne Jablonski's LOVE Practioner Course Experience

Find out what JoAnne thought about Cory Herter's brand new LOVE Practioner course in Denver Colorado.

http://coryherter.com/?p=840

Serena Carrico's experience of Sacred G Technology

LIsten as Serena shares her very first experience of Sacred G Technology and how it has impacted her life.

http://coryherter.com/?p=838

Serena's Healing with Dog at the LOVE Practitioner Course – AMAZING!!!

Listen to this incredibly gifted healer as she talks about how she used the new LOVE Practitioner Techniques to significantly reduce the size of a tumor on a dog at Cory Herter's Love Practitioner course held in Denver.

http://coryherter.com/?p=836

Solei Hannahs testimonial about Sacred G and its powerful cleansing effects!

Doing Healing for over 12 years and eating Raw for over 4 years, Solei's body was pretty purified. Sleeping on the Sacred G Bed Set, Solei discovered the amazing benefits of Sacred G. Solei shares with us the power of Sacred G's ability to cleanse the body on a whole new level than ever before.

http://coryherter.com/?p=834

Loriel Hanon's experience of the LOVE Practitioner Course

Loriel hosted Cory in Denver for the very first LOVE Practitioner Certification Course... Get the scoop from her for how the event went!

http://coryherter.com/?p=831

Alysha Duran's experience on her Sacred G Fusion Bed Set

Alysha's noticed a major difference in her ability to detect patterns that she is playing out in her life and being able to consciously shift her own internal programming to accelerate her growth.

http://coryherter.com/?p=829

Allison Lee Clark – New Senses, Enhanced Meditation, Tons of Benefits!!

Listen to Allison's Testimonial of Sacred G Technology!!

http://coryherter.com/?p=825

Kara's Sacred G Testimonial – Complete Life Transformation in 1 Year

Discover the Power Sacred G has had in transforming Kara's life as a result of sleeping on a Sacred G Bed Set.

http://coryherter.com/?p=731

.

Jayson's Testimonial of Sacred G Technology

Discover the power of Sacred G in Jayson's life and how it has helped him get everything he had on his dream board 2 years ago.

http://coryherter.com/?p=665

Lee's Sacred G Testimonial – Kidney Transplant & Medication

Listen to Lee's incredible story about how Sacred G has been helping him to reduce his medication necessary after his kidney transplant.

http://coryherter.com/?p=659

Sacred G Testimonial from a Pediatrician's Office

Watch the Testimonials from a Pediatrician's office about the Power of Sacred G on their patients. This is an incredible Testimonial for the Power of Sacred G!

http://coryherter.com/?p=277

Rob Dyrdek & Steve Berra Share Their Views about Sacred G on MTV

Steve talks about Sacred G being like Steroids while Rob Dyrdek lands a Switch Laser, a Skateboarding Trick he has never landed before!

http://coryherter.com/?p=273

Highlights from MTV's Hit Show Rob & Big Featuring Sacred G

Watch the Highlights from MTV's Hit Show Rob & Big Featuring Sacred G.

http://coryherter.com/?p=270

Webinar Archives

http://coryherter.com/?cat=100

Aug 18, 09 Webinar – Sacred G and the Table of Elements

Cory weaves Sacred G, Anti-Gravity and the Mathematics of the Coral Castle into the Table of Elements and unveils a deeper layer of the mechanics of Life!

http://coryherter.com/?p=846

Aug 25, 09 Webinar – Cory shares insights on Sacred G Fusion and Q&A

Cory talks about some of his near death experiences and how they are connected with Sacred G Fusion as well as answering a bunch of questions from Skype Chat!

http://coryherter.com/?p=848

Sept. 1, 09 – Channeling Higher Energy States

Cory talks about channeling energy, the flow of unlocking information and how we all have the energy and power within us to heal.

http://coryherter.com/?p=851

Sept. 8, 09 Webinar – Financial Abundance

Cory tackles the debate on finance and spirituality and kicks the **** out of it!! Enjoy as Cory shares how to increase your financial abundance and demonstrates there is no conflict at all. Using the Tree of LOVE, Cory illustrates the flow of energy from a spiritual source through the core and into our external reality using financial abundance to complete the creation process.

http://coryherter.com/?p=913

Sept. 15, 09 Webinar – Cleansing through Sacred G

Cory shares with everyone how the cleansing effects work and why it is so important to clear out the negative energy stored in past memories. Cleansing can come up through physical, emotional, mental and spiritual releases that can be a bit uncomfortable at times but well worth it in the end. Cory shares some of his story's of processing on Sacred G.

http://coryherter.com/?p=915

Sept 22, 09 Webinar – Sound Healing + Q & A from Chat Room

Cory shares simple yet extremely powerful techniques to use Sound for boosting energy in water and other products for Healing. Cory spends the remainder of the webinar answering questions from the Sacred G Skype Chat Room.

http://coryherter.com/?p=919

Oct 6, 09 Webinar – Cory talks Energy and Introduces the new LOVE Innerversity

Cory talks about the New Online LOVE Innerversity School. This incredible announcement has rocked the world of energy healers and natural health practitioners around the world. Join Cory as he literally opens up a new energy healing modality every week.

http://coryherter.com/?p=938

Oct 13, 09 Webinar – The Three Trees of Life, Knowledge & LOVE

Cory shares the science behind the Word of God through Sacred Geometry. The Tree of Life, Tree of Knowledge and the Tree of LOVE are explained in detail. Cory shows how the ascension is attained through understanding the Laws Of Vibrational Energy.

http://coryherter.com/?p=936

©2009 – Cory Herter - All Rights Reserved. Sacred G, Inner G, Tree of LOVE, L.O.V.E. and all Sacred G Designs are registered trademarks and copy writes of Cory Herter.

Oct 20, 09 Live Webinar – The Truth about 2012

The Truth about 2012 – Evolution, Conspiracy, 9-11, Privacy Act, Natural Selection, Light Workers and more!

http://coryherter.com/?p=951

Oct 27, 2009 Webinar – Your Health, 1 Problem, 1 Solution, 6 Layers & The Tree of LOVE

Oct 27, 09 - Your Health, 1 Problem - 1 Solution - 6 Layers - Tree of LOVE

http://coryherter.com/?p=962

Everyone and Everything within the vicinity of Sacred G will get blasted with positive vibes that unlock energy and boost performance. Increasing the amount of Sacred G you use will increase the amount of energy you have.

Sacred G Stickers come in 6 unique designs:

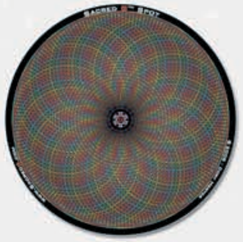

The **Sacred G Spot** is designed to physically cleanse your cells, freeing up more energy for the body to run more efficiently.

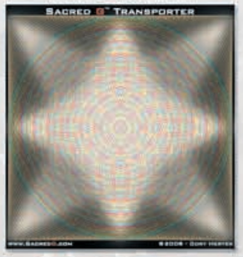

The **Sacred G Transporter** is designed to free up usable energy trapped as emotion in your past memories. Each night that you sleep on Sacred G, more and more energy is released restoring you to your natural state of Peaceful Bliss.

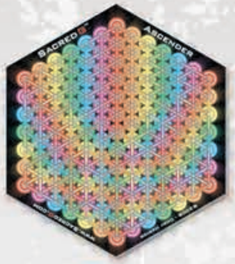

The **Sacred G Ascender** is designed to boost energy and increase self confidence by solving conflicts within your thoughts and belief systems. This enables you to perceive more information at once increasing your level of focus and reaction time.

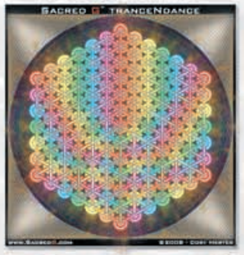

The **Sacred G TrancENdance** is a combination of the first three designs plus advanced memory systems that accelerate the learning and downloading of new information from the collective consciousness.

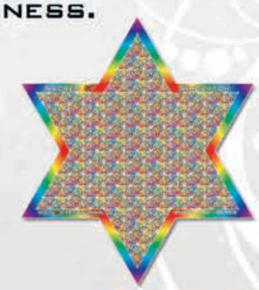

The **Sacred G Simulator** is designed to upgrade your friendships and relationships by attracting new productive ones while empowering you to let go of old ones that no longer serve you. This design will increase the number of synchronicities in your life and create an awakened sense of adventure and purpose.

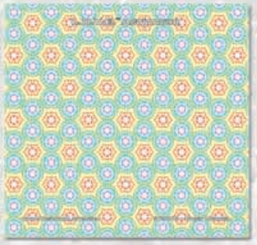

The **Sacred G L.O.V.E.™ Activator** combines all of the designs with a new motion activation system that takes you beyond your limits to achieve your goals. The Activator is the Ultimate State-of-the-Art DNA Activation Technology.

Visit the website to learn about the most advanced Sacred G Design called "Fusion" which incorporates every design into One.

Sleep / Pillow Case

Place the Sacred G beneath your pillow or inside your pillowcase design side up.

Notice how the Sacred G alters your sleep and dream states. The body frees up the most amount of energy while you sleep. Sacred G enhances this process continually increasing the amount of energy you feel.

Energized Water

Super Charge your water with Energy by placing it on top of the Sacred G.

Test: Take two glasses of tap water, place one on top of the Sacred G, and the other at least six feet away. Allow the water to charge overnight, then taste the difference between the two glasses in the morning. This test shows how Sacred G Technology not only cleans your water, but is also instrumental in physically cleansing the toxins from your body increasing energy.

"I will never ride a board that's not G'd out again!"

Rob Dyrdek

X-treme Performance Enhancement: Place the Sacred G on skateboards, surfboards, team uniforms, or any gear and equipment that you use. Watch as your performance is enhanced and your game is improved.

Visit the website to learn over 35 ways to use Sacred G to boost your energy, plus 5 cool tests that you can do. Also, be sure to watch the Sacred G episode on MTV's Rob & Big show to see how Sacred G can boost your Performance.

Visit the website to download all of the Sacred G Designs for FREE! Bonus: Watch a 3 hour video on the science of Sacred Geometry!

www.CoryHerter.com

© 2008 · Cory Herter

Sacred G Spot

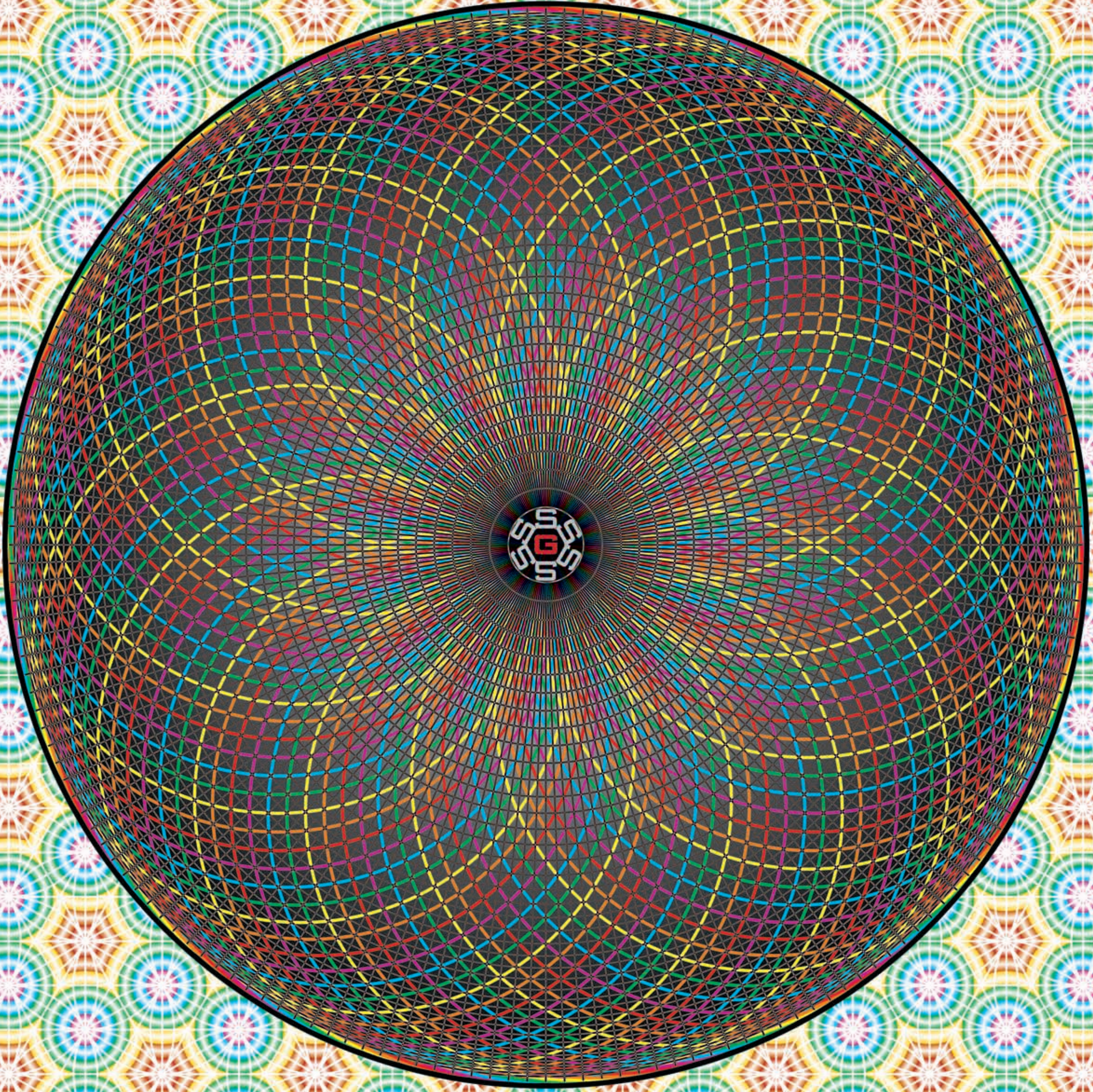

www.CoryHerter.com

©2008 - Cory Herter

Sacred G Transporter

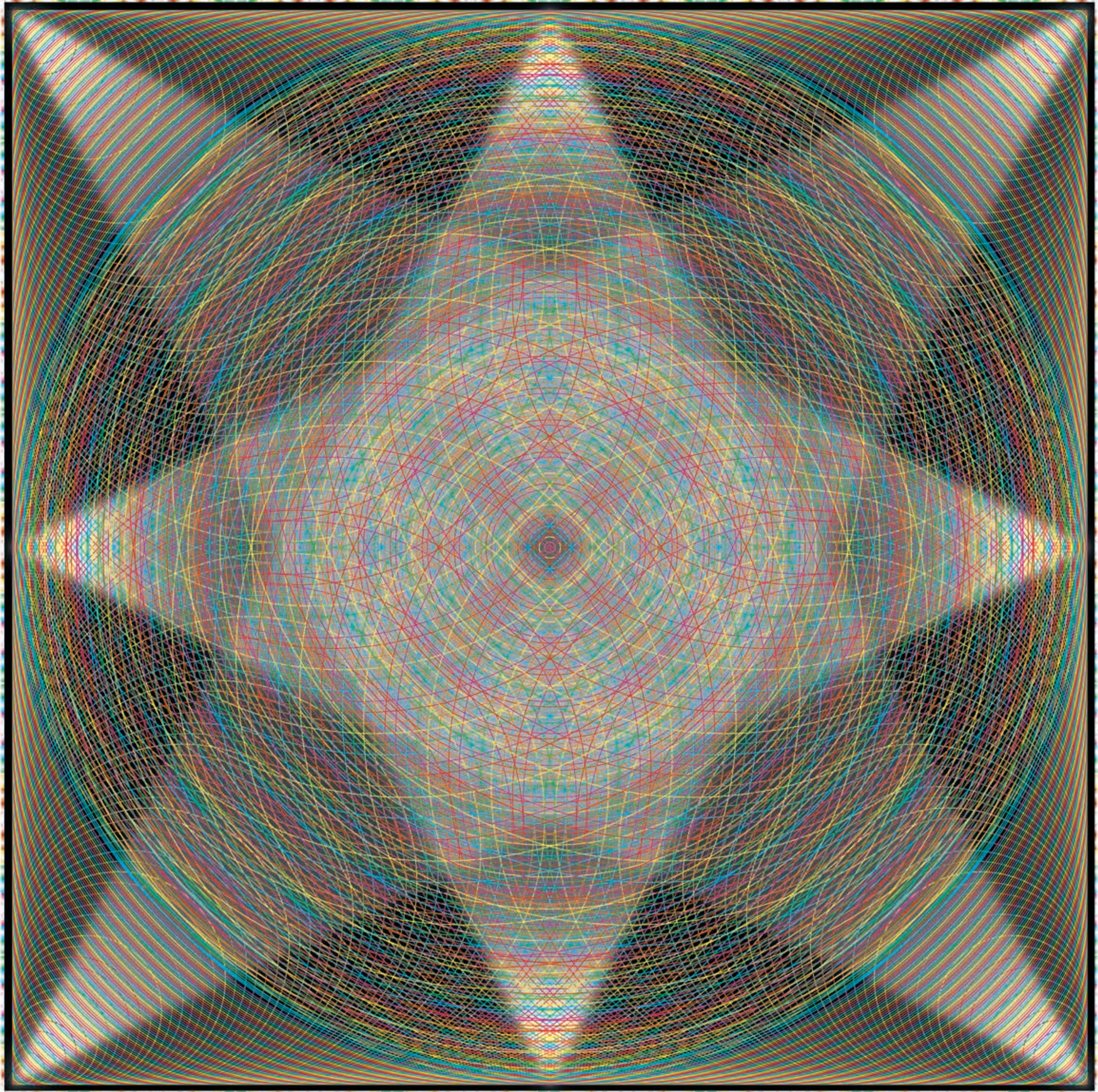

www.CoryHerter.com

©2008 - Cory Herter

Sacred G Ascender

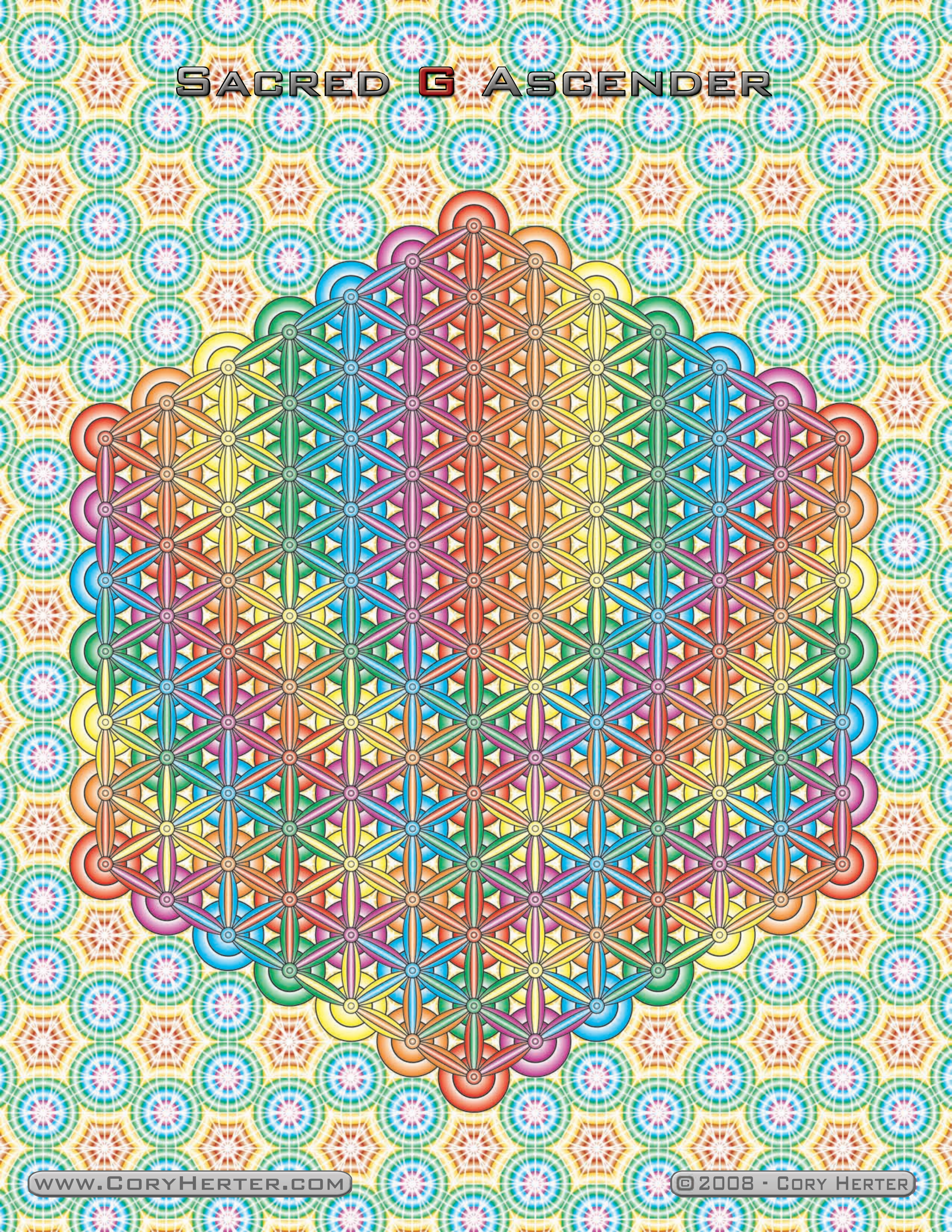

www.CoryHerter.com

©2008 - Cory Herter

Sacred G TranceNdance

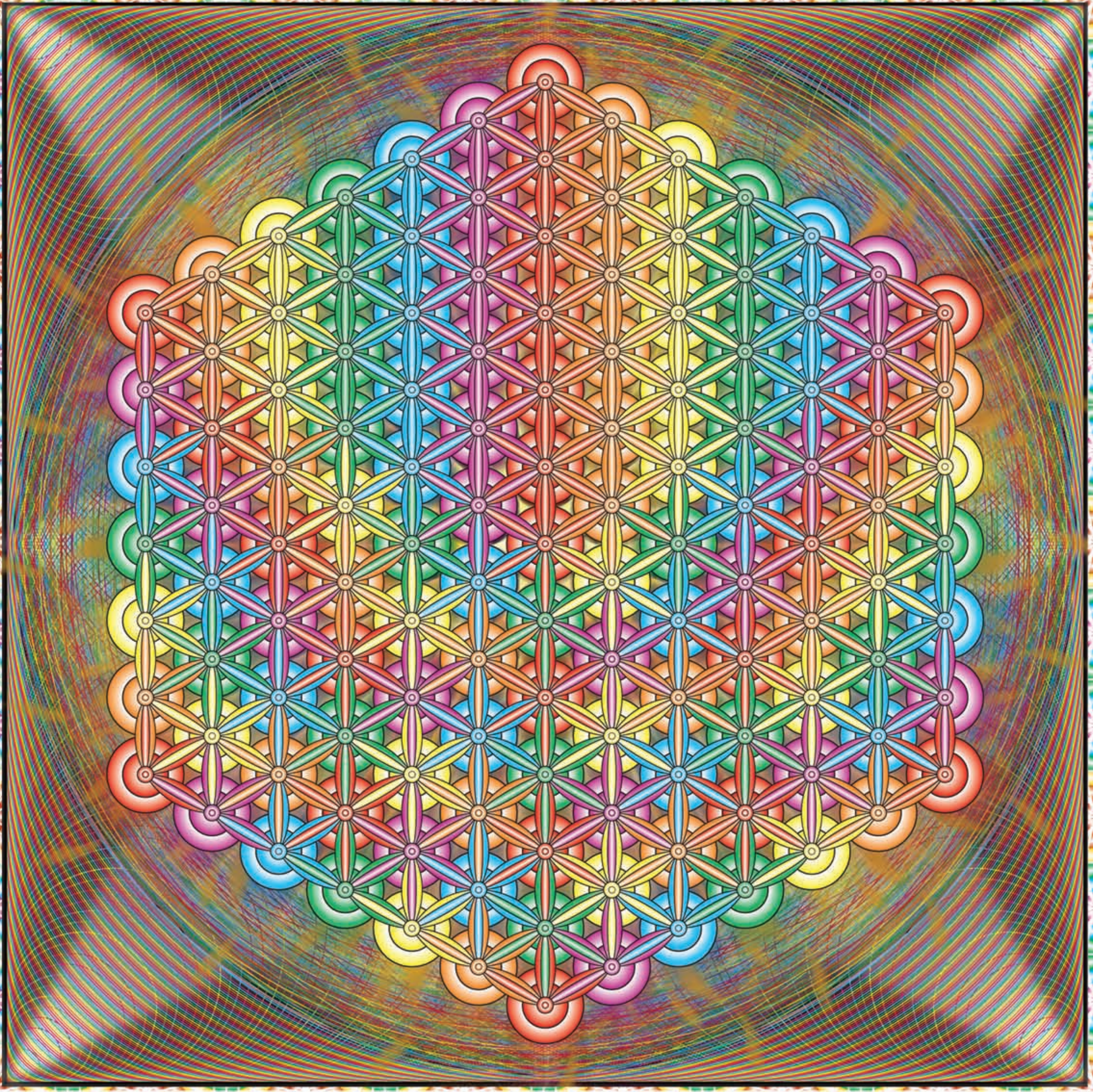

www.CoryHerter.com

©2008 - Cory Herter

SACRED G SIMULATOR

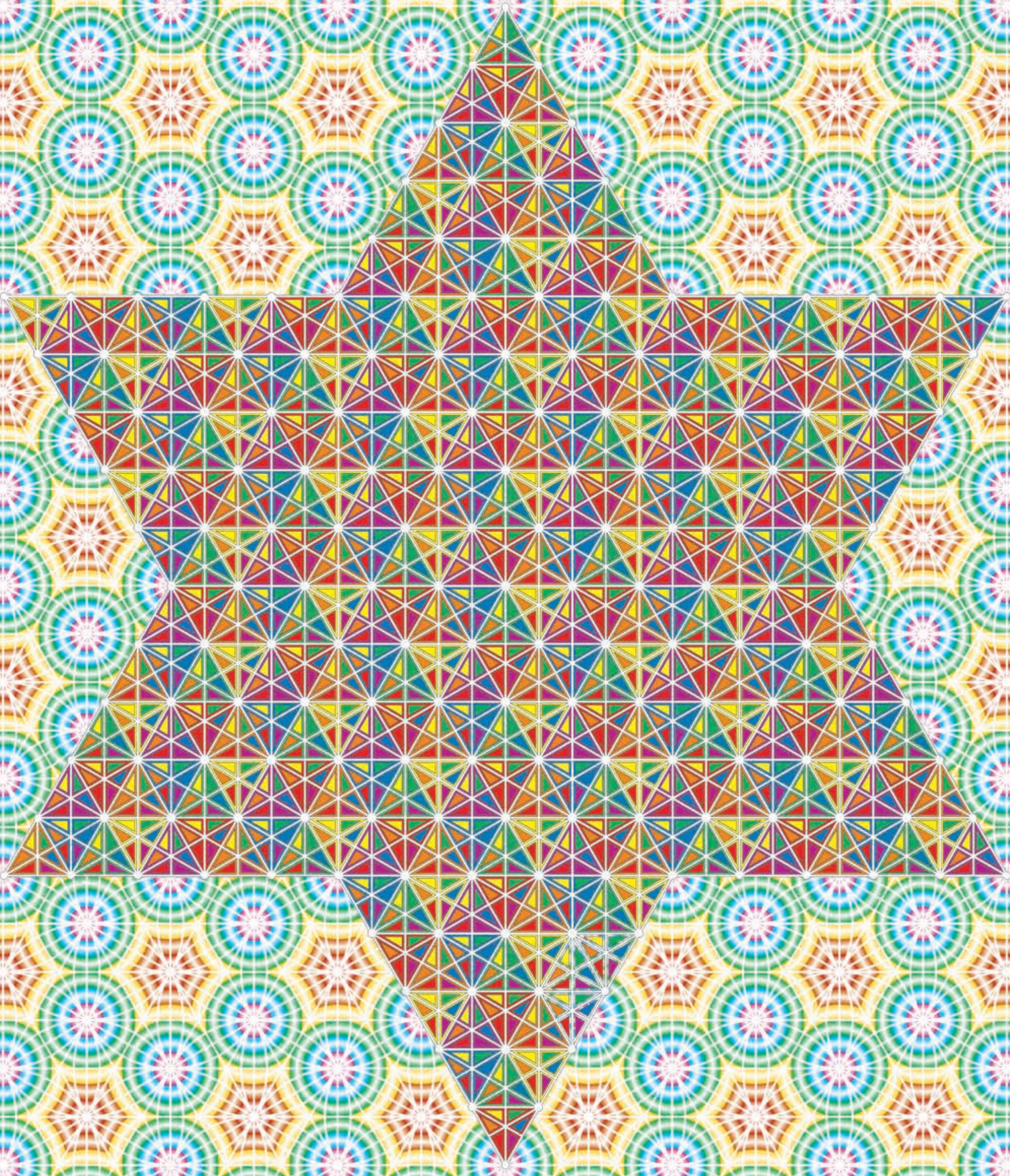

www.CoryHerter.com

©2008 - Cory Herter

Sacred G Activator

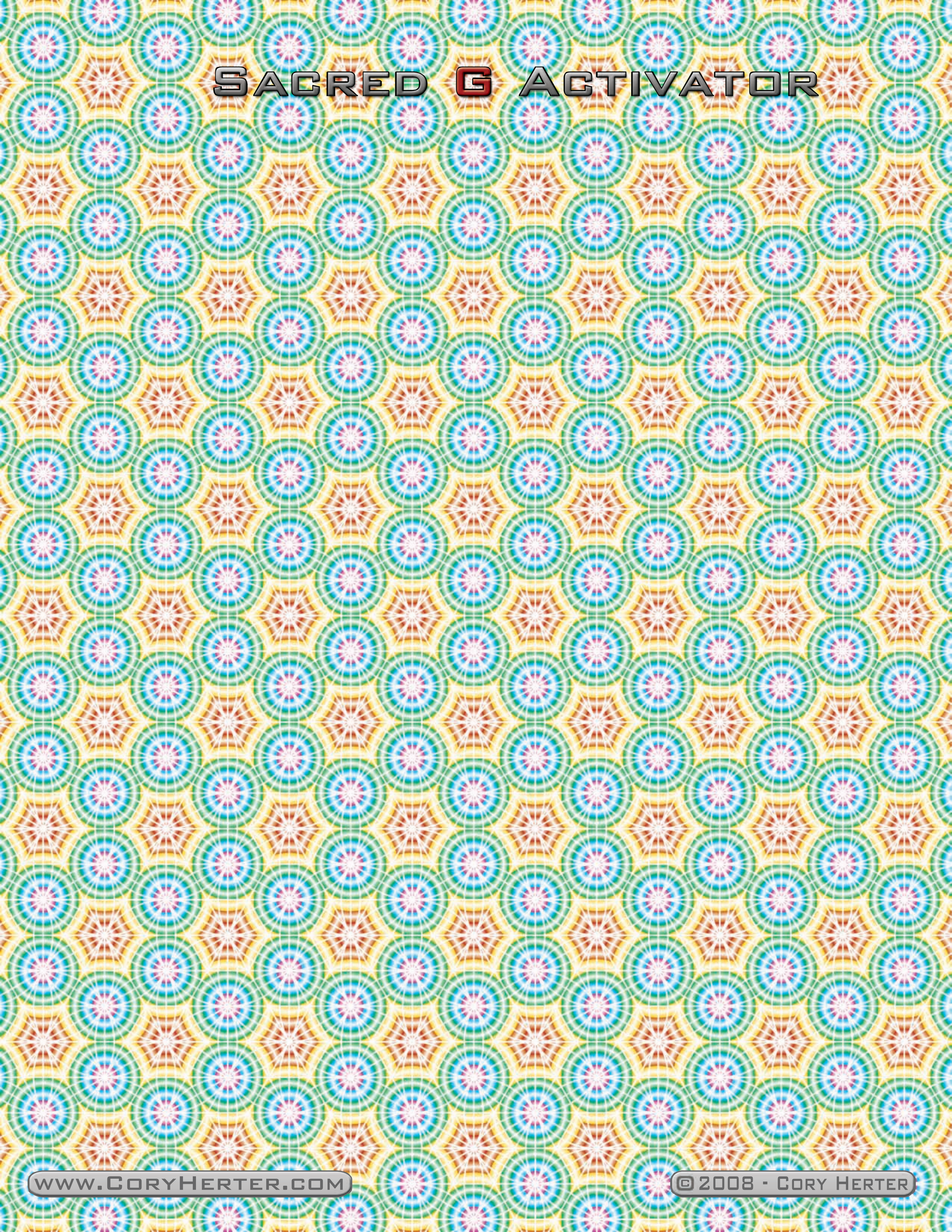

www.CoryHerter.com

© 2008 - Cory Herter

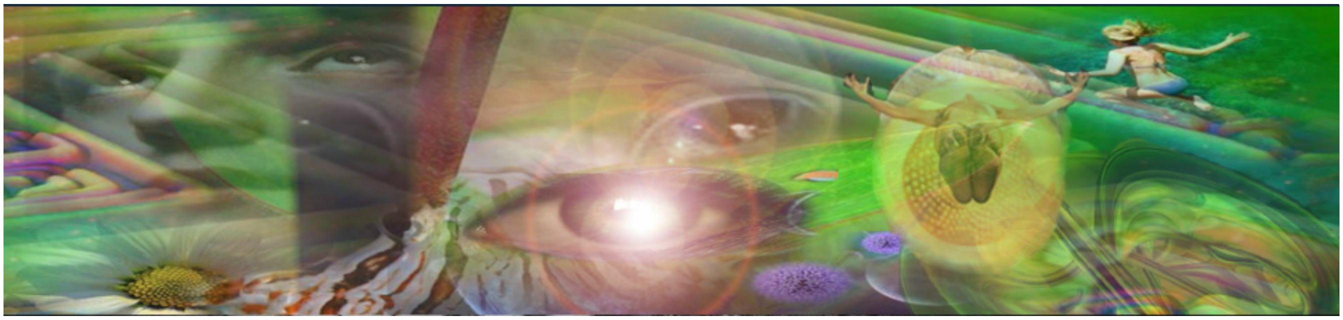

Quantum Bioholography

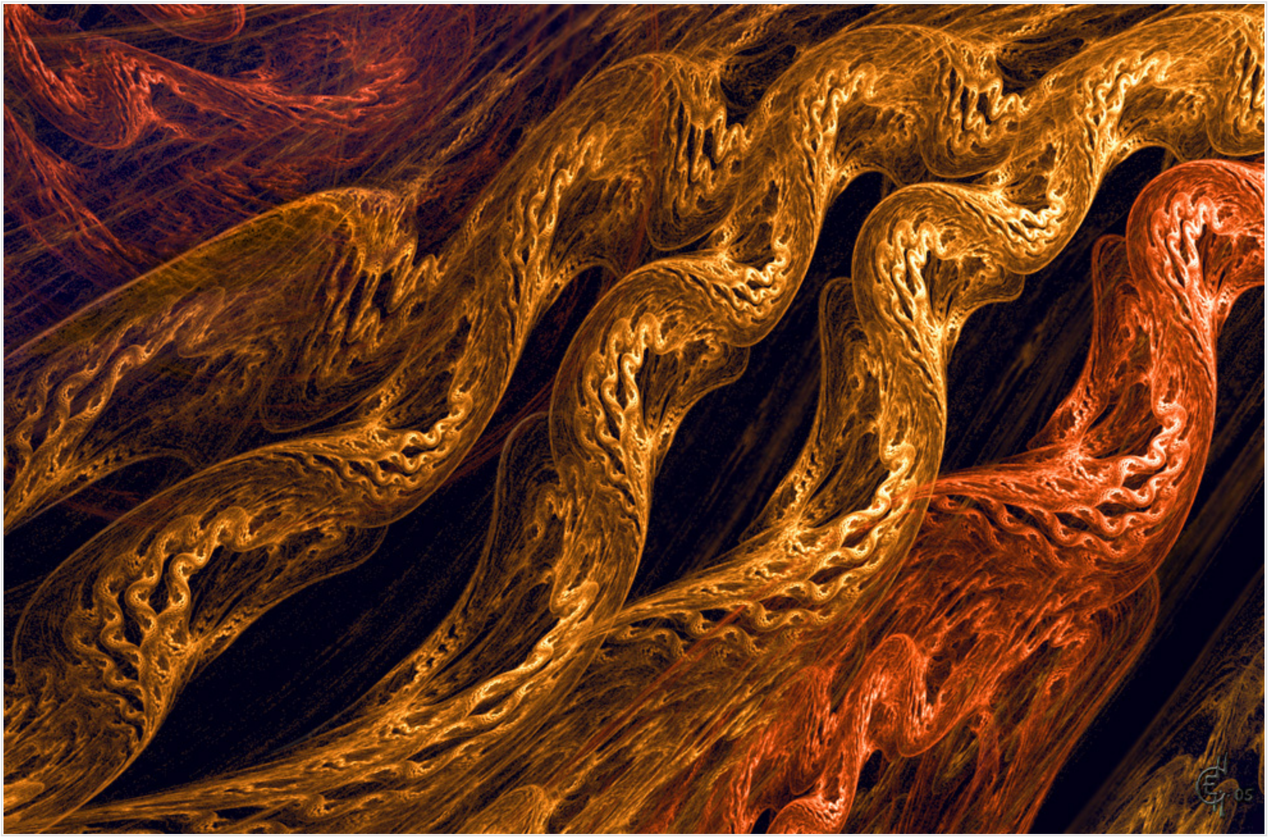

Genetic Code,

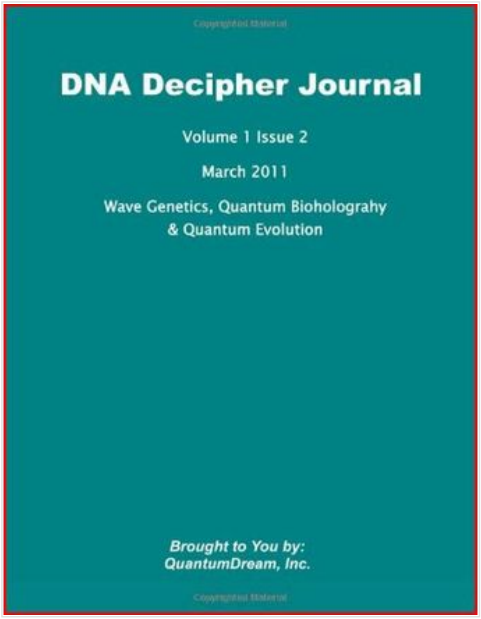

Copyrighted Material

DNA Decipher Journal

Volume 1 Issue 2

March 2011

Wave Genetics, Quantum Biohologrphy
& Quantum Evolution

Brought to You by:
QuantumDream, Inc.

Copyrighted Material

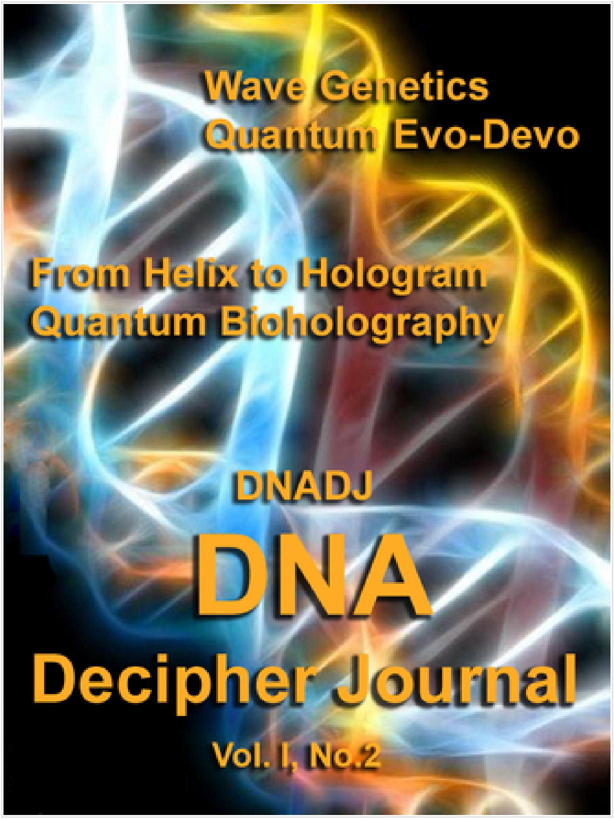

ARTICLE

http://holographicarchetypes.weebly.com/quantum-biholography.html

(http://holographicarchetypes.weebly.com/quantum-biholography.html)

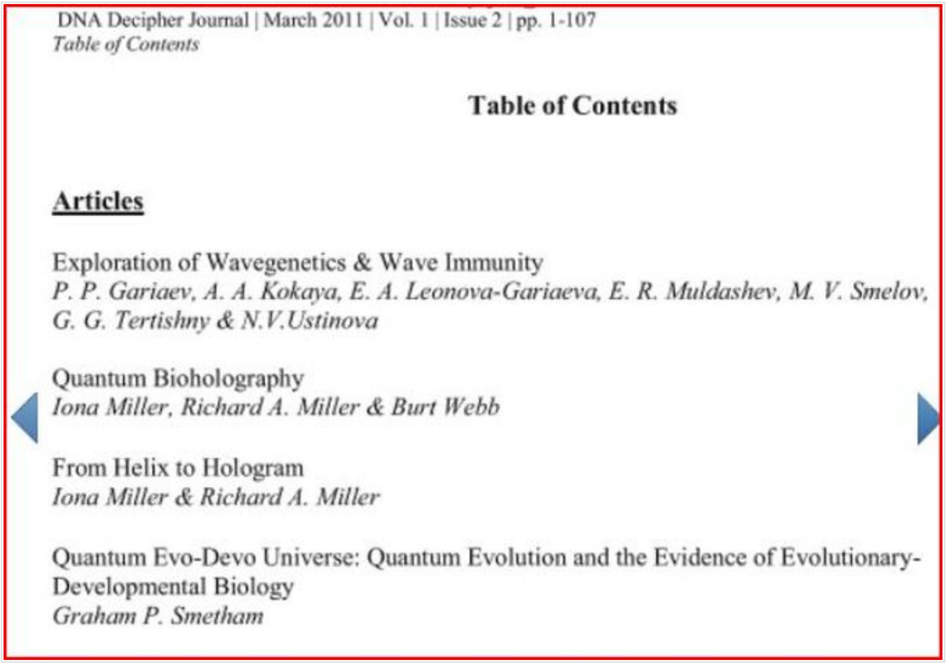

DNA Decipher Journal | March 2011 | Vol. 1 | Issue 2 | pp. 1-107
Table of Contents

Table of Contents

Articles

Exploration of Wavegenetics & Wave Immunity
P. P. Gariaev, A. A. Kokaya, E. A. Leonova-Gariaeva, E. R. Muldashev, M. V. Smelov, G. G. Tertishny & N.V.Ustinova

Quantum Bioholography
Iona Miller, Richard A. Miller & Burt Webb

From Helix to Hologram
Iona Miller & Richard A. Miller

Quantum Evo-Devo Universe: Quantum Evolution and the Evidence of Evolutionary-Developmental Biology
Graham P. Smetham

Gariaev, 2012 -

DNA Wave biocomputer functions

- **The problem of creation of the DNA-WAVE biocomputer is considered, in which one the storage locations on genetic molecules will be used. In the basis of tendered idea are trusted to experimental and theoretical activities of the writers, in which one a) to be demonstrated capacity of DNA to the laser, b) capacity of DNA to generate solitonic waves with memory, c) the phenomenon of transition of the localized photons in radio waves is revealed,**

d) the phenomenon of spectral storage of DNA by the localized photons is revealed, e) the transport phenomena of the genetic information in polarization modulations of an electromagnetic field is revealed at transition of a photon - radio wave. On the basis of these data the theory of wave genes is built, in which one the dualistic explanation of the genetic information as unity of material and wave encoding functions of the chromosomes is offered. The hypothesis quantum nonlocality of genome of higher organisms is offered. The set of the obtained outcomes allows the writers to suspect, that artificial DNA computing cannot be carried out to the full without the registration of listed properties of the genetic apparatus. Key pattern of the DNA-WAVE computer will be the phenomenon of DNA memory on solitons and on the localized photons with participation of quantum nonlocality of such memory.

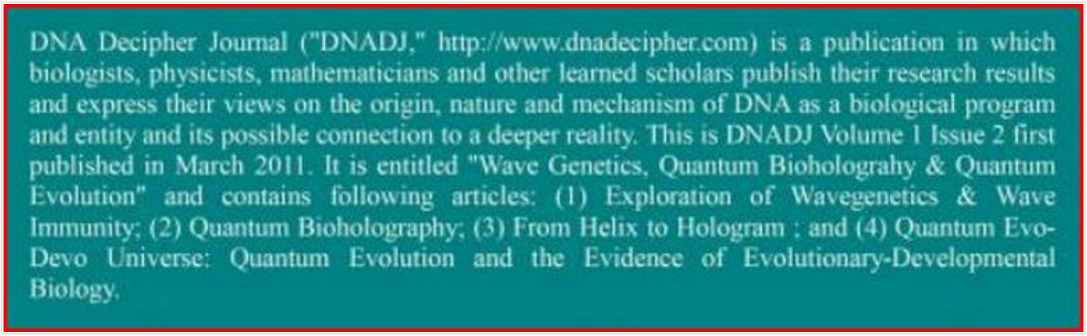

DNA Decipher Journal ("DNADJ," http://www.dnadecipher.com) is a publication in which biologists, physicists, mathematicians and other learned scholars publish their research results and express their views on the origin, nature and mechanism of DNA as a biological program and entity and its possible connection to a deeper reality. This is DNADJ Volume 1 Issue 2 first published in March 2011. It is entitled "Wave Genetics, Quantum Bioholograhy & Quantum Evolution" and contains following articles: (1) Exploration of Wavegenetics & Wave Immunity; (2) Quantum Bioholography; (3) From Helix to Hologram ; and (4) Quantum Evo-Devo Universe: Quantum Evolution and the Evidence of Evolutionary-Developmental Biology.

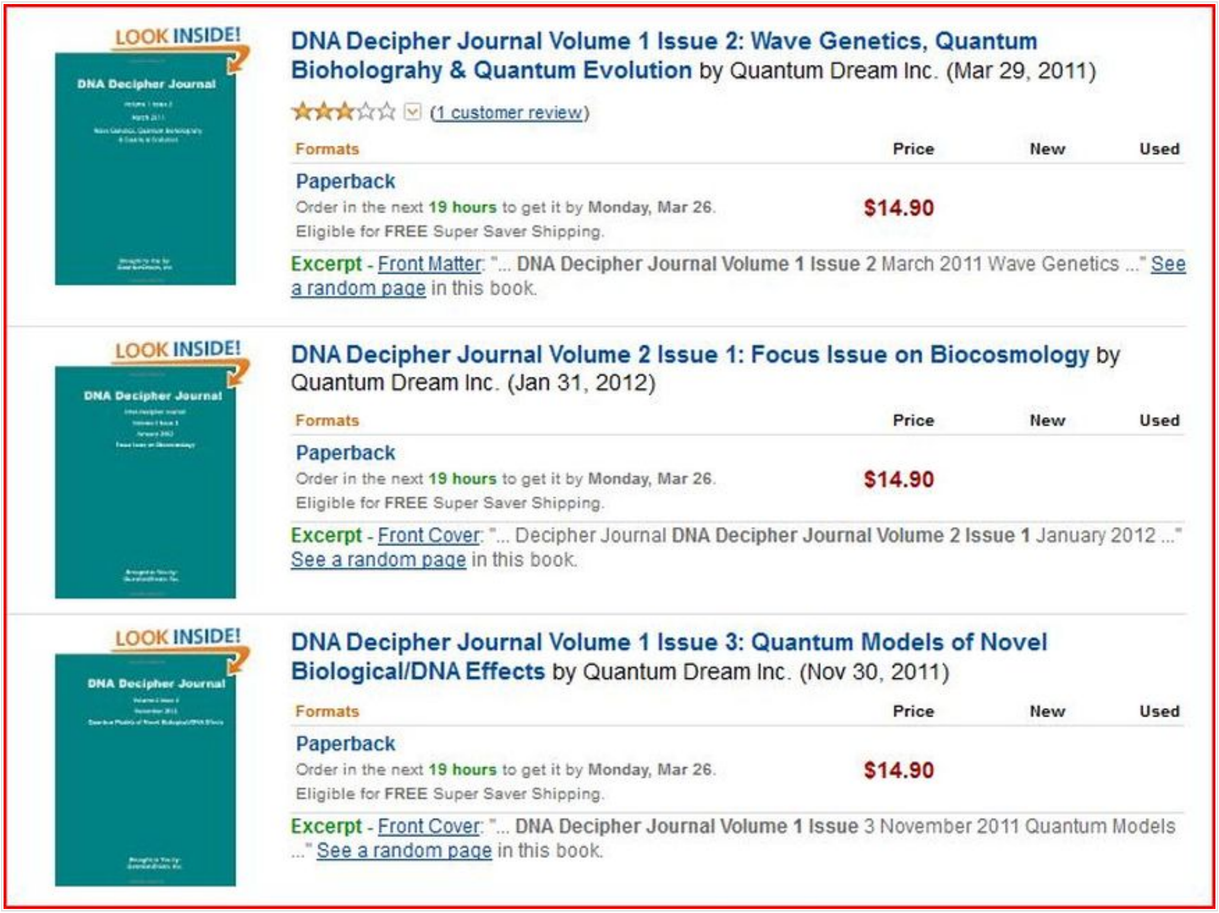

DNA Decipher Journal Volume 1 Issue 2: Wave Genetics, Quantum Bioholograhy & Quantum Evolution by Quantum Dream Inc. (Mar 29, 2011)

☆☆☆☆☆ (1 customer review)

Formats	Price	New	Used
Paperback Order in the next 19 hours to get it by Monday, Mar 26. Eligible for FREE Super Saver Shipping.	$14.90		

Excerpt - Front Matter: "... DNA Decipher Journal Volume 1 Issue 2 March 2011 Wave Genetics ..." See a random page in this book.

DNA Decipher Journal Volume 2 Issue 1: Focus Issue on Biocosmology by Quantum Dream Inc. (Jan 31, 2012)

Formats	Price	New	Used
Paperback Order in the next 19 hours to get it by Monday, Mar 26. Eligible for FREE Super Saver Shipping.	$14.90		

Excerpt - Front Cover: "... Decipher Journal DNA Decipher Journal Volume 2 Issue 1 January 2012 ..." See a random page in this book.

DNA Decipher Journal Volume 1 Issue 3: Quantum Models of Novel Biological/DNA Effects by Quantum Dream Inc. (Nov 30, 2011)

Formats	Price	New	Used
Paperback Order in the next 19 hours to get it by Monday, Mar 26. Eligible for FREE Super Saver Shipping.	$14.90		

Excerpt - Front Cover: "... DNA Decipher Journal Volume 1 Issue 3 November 2011 Quantum Models ..." See a random page in this book.

amazon.com

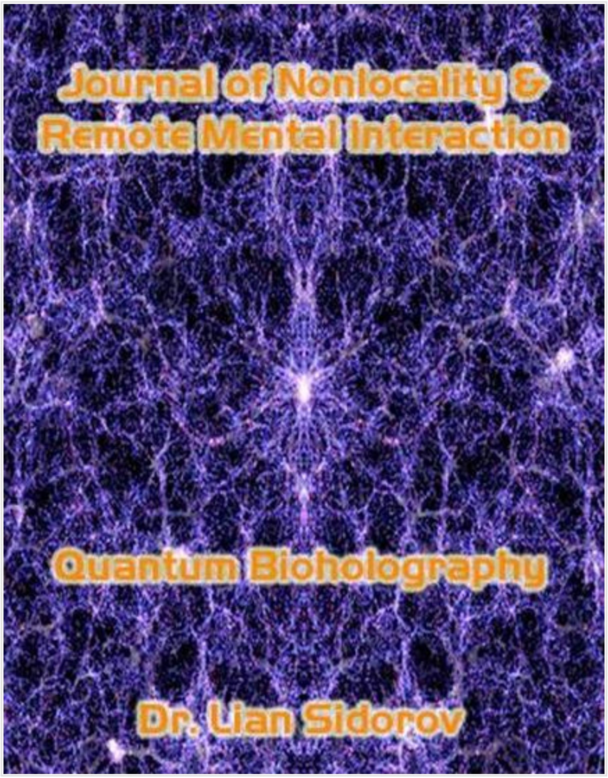

JNLRMI
Journal of Nonlocality
&
Remote Mental Interaction

Quantum Bioholography: a review of the field from 1973-2002

Iona Miller, Richard Alan Miller, and Burt Webb

URL: www.emergentmind.org/MillerWebbI3a.htm (http://www.emergentmind.org/MillerWebbI3a.htm)

Abstract: Complex information can be encoded in EM fields, as we all know from coding and decoding of television and radio signals. Even more complex information can be encoded in holographic images. DNA acts as a holographic projector of acoustic and EM information which contains the informational quintessence of the biohologram. Only 3% of human DNA encodes the physical body. The remaining 97% of the 3 billion base pair genome contains over a million genetic structures called transposons, that have the capacity to jump from one chromosomal location to another (Kelleher, 1999). We are 99.9% alike in our genetic legacy. Our individuality is expressed in three million small variations in our cells, called single nucleotide polymorphisms.

Gene-expression is the mechanism by which new patterns are called into being (Rossi, 2000). There is also a strong correlation between modulation of the brain's EM field and consciousness (Persinger, 1987; McFadden, 2002). The Gariaev group has discovered a wave-based genome and DNA phantom effect which strongly supports the holographic concept of reality (Miller, Webb, Dickson, 1975). This main information channel of DNA is the same for both photons and radio waves. Superposed coherent waves of different types in the cells interact to form diffraction patterns, firstly in the acoustic domain, secondly in the electromagnetic domain -- a quantum hologram -- a translation process between acoustical and optical holograms.

Creative, novel and enriching psychotherapeutic experiences can lead to neurogenesis, gene expression, and healing which facilitate mindbody communication and can have a long-term transformative effect on the whole person (Rossi, 2002). Thus, bioholography has relevant applications for optimizing health, well-being and even self-realization. It is relevant in biophysics, medicine, psychobiology, psychotherapy and the holistic healing arts. It provides us with a more genetically integrated model of the complex dynamics of the mindbody -- one arising in the domain underlying the standard genetic code triplet model.

Wave-genetics, Peter Gariaev

-
    - Brief introduction into WaveGenetics (https://www.facebook.com/groups/wavegenetic/doc/132917683496885/)
    - by Peter Gariaev and Irene Caesar

- **Its scope and opportunities**

  Gariaev Peter

  The history of wave genetics has been developing for more than 80 years. In 20s and 30s of the last century Russian scientists, A G Gurwitsch and A A Lubishev postulated that not only does genetical apparatus of living organisms on the Earth operate at material, physical level but also at...

-
  - The operation of the laser holographic-data converter (https://www.facebook.com/groups/wavegenetic/doc/132916856830301/)
  - by Peter Gariaev (https://www.facebook.com/Peter.Gariaev)
  - For the purpose of acquiring an insight into principles of operation of the laser holographic-data converter, which is in fact – quantum bio-computer (Ref 26), works of Denisuk U N on dynamic holography are of particular attraction (Ref 17). He has developed the basics of holographic imaging of material objects, including objects in motion in ...
-
  - Wave Genetic Bio-Management nanotechnologies. (https://www.facebook.com/groups/wavegenetic/doc/132914663497187/)
  - by Peter Gariaev (https://www.facebook.com/Peter.Gariaev)
  - **Wave Genetic Bio-Management nanotechnologies. Theory and empirical evaluation Tertishnii G.G. and Gariaev P.P. New medicine technologies, №7, p. 49-64 (2007)**

    In this paper a further development of previously propounded theory of external quantum management and self management of organisms 'in vitro - in vivo' by means of bio-holographic...
-
  - Antenna Model (https://www.facebook.com/groups/wavegenetic/doc/132907646831222/)
  - by Peter Gariaev (https://www.facebook.com/Peter.Gariaev)
  - We have noted earlier (Gariaev, Maslova et all, 1996 (a); Gariaev, Maslova et all., 1996 (b)), that in the role of the main molecules perceiving external electromagnetic fields as a regulators – ar the proteins and especially metal containing. Functioning of a set of biological macro-moleculs (ferments in particular) is largely determined ...
-
  - Mathematical Modeling of solitons on DNA (https://www.facebook.com/groups/wavegenetic/doc/132903276831659/)
  - by Peter Gariaev (https://www.facebook.com/Peter.Gariaev)
  - Mario Salerno was the first to experiment with solitons on DNA on computer not only as formal mathematical structures but also tried to link their behavior in a mono-dimensional space of poli-nucleotides with their bio-genetic, or more accurately with their epi-genetic functions. He had been developing the first soliton model on DNA ...
  - 3. Molecular – tissuey “radio stations” (https://www.facebook.com/groups/wavegenetic/doc/132902406831746/)
  - by Peter Gariaev (https://www.facebook.com/Peter.Gariaev)
  - Additional theoretical approaches here, possibly, lie in effects of so called “weak ifluence” [Chukova, 2002]. Apparently regenerative and cytoprotective effects in our experiments bear endoergic character where even weakly absorbed energy of coherent red polarized laser light by the bio-preparations goes in to escalation of Helmholtz free...
-
-
  - 2. Quantum teleportation of genetic-metabolic information/data in permissive version. (https://www.facebook.com/groups/wavegenetic/doc/132898556832131/)
  - by Peter Gariaev (https://www.facebook.com/Peter.Gariaev)
  - (based on works of [Prangishvil ,Gariaev, Maksimenko et al., 2000(б); http://www.wavegenetics.jino-net.ru/zip/Teleport.zip (http://www.wavegenetics.jino-net.ru/zip/Teleport.zip)

    The abovegiven biological experiments on remote transmission of morphogenetic, or more accurately – genetic-metabolic information/data from a donor to recipient may be interpreted from views of non-localized ...
-
  - EXPLORING WAVEGENETICS AND WAVE IMMUNITY. (https://www.facebook.com/groups/wavegenetic/doc/132895870165733/)
  - by Peter Gariaev (https://www.facebook.com/Peter.Gariaev)
  - THEORETICAL MODELS *.

    Gariaev P.P., Kokaya A.A., Leonova-Gariaeva E A., Muldashev E.R., Mikhina I.V.,

    Smelov M.V., Tertishnii G.G., Tovmash A.V., Chalkin S.F., Shatrov Y.K., Yagujinski L.S.

    * Disclaimer: This translation and English text editing have been done to the best knowledge and understanding of the translator who is not a specialist...
  - Last edited about a month ago (Saturday, February 18, 2012 at 5:36am) ·
-
  - Publications (https://www.facebook.com/groups/wavegenetic/doc/132893456832641/)
  - by Peter Gariaev (https://www.facebook.com/Peter.Gariaev)
  - **Publications**

    Publications in foreign journals

    1. Garyaev P.P., Vikhert A.M., Zakharkina G.A., Alexandrova L.G., 1977, Endogenous reserpine-like substances in the myocardium and other tissues of man and animals., USA-USSR First Joint Symposium on Sudden Death. Yalta, USSR, October 3-5, 1977, pp.323-338.

2. Gariaev P.P., Chudin V.I., Komissarov...

Brief Introduction into WaveGenetics
By Peter Gariaev and Irene Caesar in Wavegenetic (https://www.facebook.com/groups/wavegenetic/) **· Its scope and opportunities**

Gariaev Peter

The history of wave genetics has been developing for more than 80 years. In 20s and 30s of the last century Russian scientists, A G Gurwitsch and A A Lubishev postulated that not only does genetical apparatus of living organisms on the Earth operate at material, physical level but also at certain waves/fields level and is able to transfer genetic data/information via electromagnetic (EM) and acoustic waves. The science has recently taken a major leap in its advancements of the theory and reproducible experiments and thus the theory of wave genome has been put forward.

One of the first attempts to rationalize and expound the wave genome theory in Russia had been made by *P P Gariaev* and *A A Berezin* from the department of Theoretical Matters of Russian Academy of Science, and also with participation of A A Vaciliev, a fellow member of Institute of Physics of Russian Academy of Science. As a theoretical foundation, principles of *coherent physical radiations*, *holography* and *solitons*, *theory of physical vacuum*, *fractal representation of DNA* and *human verbal expression/speech* have been employed to describe and substantiate the empirical results obtained through many experiments conducted.

The *quintessence* of the wave genome theory may be represented as following:

genome of the highest organisms is considered to be a bio-computer which forms the space-time grid framework of a bio-systems.

In that bio-system, as the carriers of a field epi-gene-matrix - *wave fronts* are being used, which are assigned by gene-holograms and so-called solitons on DNA – distinct type of acoustic and electromagnetic fields, produced by biogenetic apparatus of the organism/bio-system under consideration and being a medium of strategic regulatory data/information exchange between cells, tissues and organs of the bio-system.

It is also vital to note that the holographic grids/frameworks, which are also the elements of fluctuating structures of solitons, are, in fact, discrete simplest cases of code-originated information, anchored in chromosome continuum of an organism.

At present the dominant viewpoint in respect of genetics and molecular biology is that[1]:

1. **the genetic apparatus functions as a purely material structure**
2. **all the functions of genetic control of an organism occupy approximately 2 % of DNA of a bio-system and fulfill designated functions such as replication of RNA and proteins, so called coding DNA of an organism. The other 98% is considered to be "junk DNA" not carrying out any genetical functions, which may represent a graveyard of virus DNA.**

The biologists and geneticists use the language of analogies and metaphors to explain how genetic apparatus operates. The genetic apparatus consisting of 46 chromosomes is viewed as a library consisting of 46 volumes or books. Each book (a chromosome), contains a text (instructions of how to build an organism) which consists of sentences (DNA) consisting of words (genes). And each word (a gene) consists of 4 letters (certain "chemical letters"), i.e. the "genetic alphabet" consists of only 4 "letters". The material realizations of the DNA molecules are famous double helixes, consisting of segments which are genes. In essence, genetic apparatus operates as follows. The texts, written in the "DNA language", are first translated by the organism into the "RNA language" and then into the "Protein language". And proteins are the stuff that we are mostly made of (not counting water). Proteins perform two principle functions in the organism: they metabolize substances that we eat and participate in the morphogenesis, i.e. development of the spatial-temporal organization of an organism.

It is to be pointed out that the primary focus of the wave genome theory is on the remaining aforementioned 98% of chromosomes as being the *key* "intellectual" structure of all cells of an organism including the brain. It is those chromosomes that operate on the wave, on the "ideal" (fine-field) level.

It the *ideal* component, that may be called super-gene-continuum, is a strategic vital figure/formation that ensures

development and life of humans, animals, plants and also their *programmable* natural dying. Along with that, it is important to conceive that there is no sharp and insurmountable distinction between genes and super-genes. Both these levels of encoding constitute material (physical) matrixes, however genes supply material replications in the form of RNA and proteins, whereas super-genes transform endo- and exogenous fields, forming from them super-gene-signal wave structures. Furthermore, the genes may be components of holographic grids/frameworks of super-genes and supervise their field activity

Particular attention in the wave genome theory should be devoted to substantiating of the unity of fractal (replicating itself on varying scales) sequence structures of DNA and of a human speech. In 1990 Jeffrey Delrow discovered four “letters” of the genetic alphabet (Adenine, Guanine, Cytosine and Thymine) in DNA “texts” form fractal structures.

Later on, a discovery of similar fractal structures in a human speech, not limited to multy-letter alphabets of Russian and English texts and including a sequence of words from those texts came as a promising surprise for both genetics and linguists. Nonetheless this is in accordance with a branch in semiotics named “*Genetical Linguistics*”, which studies incomprehensible and unexplainable precise application of laws of Formal Genetics to formation of interlingual and intro-lingual words-hybrids.

A group of scientists headed by P P Gariaev and M U Maslov, developed a theory of so called fractal representation of natural (human) and genetical languages. Within the confines of this theory it is said that the quasi-speech of DNA possesses potentially inexhaustible “supply of words” and, moreover, what had been a sentence on the scales of DNA–“texts” “phrases” or a “sentence” becomes/turns into a word or a letter on the other scale. Genetical apparatus can be viewed as the triunity of its structure-functional organization consisting of holographic, soliton and fractal structures.

This theory allows a refined quantitative comparison of symbolic structure of any texts including *genetical*. Thereby a possibility has been wide open to approach a deciphering of a lexicon of one’s own gene-code, and accordingly, more accurate composition of algorithms of addressing a genome of a human with an aim of potentially any type of programming of one’s vital activity such as treatment, increasing one’s life expectancy and so on and so forth.

Empirical tests of wave genetic theory in the light of “speech” characteristics of DNA demonstrate strategically correct stance and direction of the research.

Phenomenal experimental findings may be *briefly* presented here. For the first time in history of the science we successfully managed to obtain experimental evidences of abilities of the genetic data/information to function on more than simply one material level, for instance, in the wave/field (electromagnetic) form

Over decades, having been conducting extensive theoretical researches, we successfully managed to formulate theoretical biology and physics-mathematical description, explanation and substantiation of fundamental principles of genetic apparatus functioning on the wave/field level.

These principles allowed us to design and create a *quantum bio-computer* essential element of which is a specially tuned laser beam with certain wavelength and frequency characteristics.

The quantum bio-computer can perform:

i. Scanning and reading the wave/field equivalent of the genetical-metabolic data from the (stem) cells, tissues, organs of a donor’s bio-system employing photon of the laser beam;

ii. Conversion of the photons obtained into wide-wavelength-band preserving the scanned and read data;

iii. Precise and pointed introduction of such scanned and converted principal/dominant data radio waves into the organism-recipient, located at a certain distance from the donor’s bio-system (from few centimeters to 20 km);

iv. Strategic control/management of metabolism and postembryonic morphogenesis of the organism-recipient according to two vectors/modes:

a. "Do as/what I do" – that is holographic principle, and

b. Wave/field transmission of a signal that contains commands activating required programs in the stem cells of, for instance, rats; programs that direct/guide the stem cells development on their way of, as in our experiments, building pancreas gland in rats. These vectors/modes are intended to be used on humans.

At present we are able to program/manage/encode stem cells of various types by means of a *quantum bio-computer*. The *quantum bio-computer initiates* wave/field-based commands, given to cells and tissues of the donor/recipient, and accordingly the stem cells exposed to the waves/field will be prompted to guided cytodifferentiation leading to emergence and development of planned/projected new organs and tissues. This will bring us closer to the fact of substantially increased life expectancy.

Significant achievements in applying *WaveGenetics* have been made thus far in regeneration of pancreas in rats previously destroyed by chemical substance called alloxan

Three series of experiments with identical protocol were conducted by the groups of P. Gariaev in 2000 in Moscow Russia, in 2001 in Toronto, Canada, and in 2005 in Nizhni Novgorod, Russia. These are more advanced experiments based on the principles and technology of *WaveGenetics*. The goal of the experiments was to test new technology for regenerating damaged pancreas. Pancreas is an endocrine gland which has several vital functions, the major one being production of insulin, a hormone responsible for sugar metabolism.

A control group of rats was injected a lethal doses of a poison called alloxan which destroys pancreas. As a result, all the rats in the control group died from diabetes in 3-4 days. Then the same lethal dose of alloxan was injected to another group of rats. And when the rats reached the critical condition, they were exposed to light images/waves coming from a *quantum bio-computer*. Those light images/waves were created beforehand when the bio-computer read information from the pancreas surgically removed from healthy newborn rats of the same species as those used in the alloxan experiments.

One can explain the results of the experiment using the following analogy. The pancreas gland contains DNA-movies with information about healthy condition of the pancreas in its genetic apparatus. And this video morphogenic information *programmed the stem cells* of sick rats *to regenerate their pancreas gland*. Combined statistics for all 3 series of experiments is as follows. Altogether, around 90% of all the rats had their pancreas gland restored and their health recovered.

In some of the experiments *the quantum bio-computer* was modified to allow successful transmission of the healing information to sick rats at the distance of 20 kilometers. Note, that no known physical fields have the capability to transmit such extremely week signals with such unbelievably powerful results.

Furthermore, the WaveGenetics research and application has significant prospects of solving issues as regards ageing process in humans and increase of life expectancy. This view is grounded on the experiments we have been conducting with rats.

It is Opportunities of technologies of *WaveGenetics* are not limited to what has been outlined above. The development and application of the technology is *far-reaching* and is substantiated by experimental data obtained in numerous tests, experiments, observations.

In addition, it ought to be mentioned, such a technology – a quantum bio-computer, is capable of:

- **treating oncological diseases on fundamentally distinct basis – without use of any chemical substances,**
- **eliminating pathogenic viruses and bacteria and agricultural vermin parasites – also without used of any chemical substances.**

It may seem as a technology of another age, from distant future. However, a discovery of fundamental properties of living organisms is occurring today and it is our task to research and explain the phenomena and bring it to the service of humanity.

Peter Gariaev, Ph.D.,

Director of Wave Genetics Institute,

Russia, Moscow.

Antenna Model

- **By Peter Gariaev (https://www.facebook.com/Peter.Gariaev) in Wavegenetic (https://www.facebook.com/groups/wavegenetic/) ·**
- **We have noted earlier (Gariaev, Maslova et all, 1996 (a); Gariaev, Maslova et all., 1996 (b)), that in the role of the main molecules perceiving external electromagnetic fields as a regulators – ar the proteins and especially metal containing. Functioning of a set of biological macro-moleculs (ferments in particular) is largely determined by processes occurring in the active centres surrounded by bio-polimer sequences having a sign topology. Based on this perception of informational bio-macro moleculs, it is normal to suppose that theyir interaction with physical fields external in relation to the biosystem and internal radiations leads to excitation dipolar-active fluctuations of monomers, forming the said chain, and they in turn induce fluctuation if the active centre. In toher words, such system will work as a peculiar antenna.**

  There excited fluctuations can lead to transitioning of bio-macr-molecul in to toher conforming (topological, sign) state.

  Such conception in a principal plane is adequate to the whole set of functionally important bio-macro molecule, for example, chlorophyll, hemoglobin, mioglobin and so on,. These macromolecules are joined by two structural properties: 1) in their geometrical centre there is ion (in the case of chlorophyll – ion of magnium, in the case of hemoglobin –ion of iron); 2) near ion symmetrically placed 4 pirrolik rings (pseudo structure.)

  Other types of polymers, corresponding to antenna model, can be relatively simple cycles such as valinomycine (potassium ions' carrier) and complex over-molecular chromosom structures, DNA of which contains higly organized associates of such metals as magnesium, calcium, nickel, cobalt, copper, ferrum, zink and so on. Their role is not clear and is considered to neutralize OH-groups of remnants of phosphoric acid of polynucleotide. It is proposed that functions of metals in DNA and RNA is substantially wider and is realized along the lines of sign and/or energy interchangewith endo – and exogenous, in relation to the biosystem, physical fields. The same is applicable to the proteinsnot containing porfirine centre yet specifically binding metals, e.g. as such you can considersite-specific proteins with domens like "zink fingers" which participate in regulation genes, often located long awayfrom these governing proteins. Atoms of metals in DNA and proteins can in resonance interact via electromagnetic channels within the meaning of antenna model. Let defined the definition of antenna model.

  External energy (in particular, connected with resonance interaction extremely high frequency electromagnetic radiation with proteins) travels to the periphery, that is onto ensemble of sub-units (not necessarily identical in structures). As a result of an active "conversation" predetermined by bio-chemical links between peripheral acceptor (received encoded energy) and centre-associate (in this case and ion of irongem-containing proteins), the latter receives energy (information) which effects the biological action. The degree of reactive ability of bio-macromoleculs is considerably dependent on the level of excitation of central sub-units. Lets look closely in to potential mechanisms of wave interactions of physical fields and active centres of informational bio-moleculs within the boundries of the proposed antenna model.

  As an example of the simliest model for illustrating antenna effect we consider 2-dimentional closed (cyclic) chain of monomers. There is anactive centre, connected to monomers of chain by dipole-dipole interaction.

  http://eng.wavegenetic.ru/index.php?option=com_content&task=view&id=20&Itemid=1 (http://eng.wavegenetic.ru/index.php?option=com_content&task=view&id=20&Itemid=1)

Molecular – tissuey "radio stations"

By Peter Gariaev (https://www.facebook.com/Peter.Gariaev) in Wavegenetic (https://www.facebook.com/groups/wavegenetic/) ·

Additional theoretical approaches here, possibly, lie in effects of so called "weak ifluence" [Chukova, 2002]. Apparently regenerative and cytoprotective effects in our experiments bear endoergic character where even weakly absorbed energy of coherent red polarized laser light by the bio-preparations goes in to escalation of Helmholtz free energy accumulated in chemical relations of metabolites of bio-preparations, which leads to significant bio-effects. Atoms of informational molecules (DNA, RNA, proteins), interacting with the laser beam together with the energy of the light

quanta also obtain their momentum quantity motion, which creates inversed overpopulation in nuclear zeeman levels, that is – chemical polarization of nucleuses occur. Quantum polarization exit, that is a number of excess kernel spins on the top Zeeman level correlated to the every absorbed quantum of light, can be 30%. Inversely populated protonic-spin system can issue quanta with 6,5·10-26 Jouls of energy which correspond to frequencies of about 10Mhz [Buchachenko, 1979]. Further advancing the aforesaid, one could consider that when bio-preparations are scanned by the laser beam there appears combinative frequencies, occupying blue and UV diapason. Moreover as we previously suggested in an appropriate model of localized light, there is transformation of frequencies in diapason from 2 omega to zero takes place. [Prangishvili, Gariaev et al, 2000 (b)]. Besides when scanning the bio-preparations there is a presence of wideband radio radiation of the gas discharge laser. Taking this into account, we believe that in our experiments bio-preparations-donors are scanned ("read" by the laser) by way of multiplicity of frequencies. Metabolising chemical compounds of bio-preparations, which interact with dynamically polarized red light of the probing laser (and wide spectrum of additional radiation), can generate electromagnetic radiofrequency oscillations. In this instances the bio-samples of pancreas and spleen "assume" a role of idiomatic organ-molecular radio stations where every type of their molecules possess characteristic frequencies that can be reinforced owing to stochastic resonances.

From the other hand, certain types of molecules of pancreas affected by alloxan in the ill animals and/or their stem cells in the blood can resonantly absorb such radiation/emission which carries informational component required for launching bio-chemical eactions which lead to regeneration of tissues of the pancreas and protective anti-alloxan processes. This does not exclude significant contributions of the earlier explored concepts/mechanisms related to teleportational and holographic features/properties of bio-preparation-donor's genome. Let's have recourse to known theses of chemistry and physics http://www.chem.msu.su/rus/publ/Buchachenko/buch5.html (http://www.chem.msu.su/rus/publ/Buchachenko/buch5.html), setting quantum scenario of the genetic-bio-chemical events in question. In the ensemble of molecules-products with inverse population in Zeeman's reservoir/cavity energy is accumulated; it can be spent in the form of heat (via spin-grid magnetic relaxation), but can also convert in to stimulated radiation at Zeeman's nuclear frequency. In this case the reaction does, in fact, become radio frequency emitter, quantum generator with chemical pumping (similar to chemical lasers).

This is new phenomenon – radio emission of a chemical reaction – it was predicted in theory and than discovered empirically. It emerges when Zeeman's cavity/reservoir energy exceeds the generation threshold; the motion of the nuclear spins spontaneously becomes coherent and such coherent system of nucleuses then becomes a quantum generator emittiin in microwave diapason/spectrum. Yet this is only one aspect of chemical radio physics. Chemical reaction can be not only a generator but a receiver of the microwaves. Reception at chemical level follows from principles of spin chemistry: resonant microwave radiation stimulates triplet-singlet conversion of radical couples (or couples of other spin carriers) and alters exit of chemical products. In this way magnetic-spin phenomena make biochemical reaction a receiver of microwave irradiation/emission. Moreover such receiving can be selective. If the microwave pumping touches upon all radical couples (biochemical substratum) then total result boils down to changing the reaction product exit at resonance frequencies. This phenomenon is called reaction yield detected magnetic resonance (RYDMR).

If the "pumping" is selective and involves only radical couples with magnetic nucleuses, a phenomenon is brought about – radio induced magnetic isotopic effect (RIMIE). And, at last, if the microwave "pumping" is selective also in respect of nucleuses' spin orientation (that is involves ensembles of radical couples with chosen orientation of the nucleuses' spins.) then stimulated polarization of the nucleuses (SPN) occur. This is linked to so called spin catalysis It is notable for its spin conversion of reagents being induced by paramagnetic particle-spin catalyzer. Conversion occurs as a result of catalyses (ferment) interactive exchange with reagents. Spin catalyzer accelerates recombination of radicals, wax isomerization of compounds with double connections (by seven-eight orders), recombination of spin-polarized atoms and so on. It is possible that spin catalysis functions in bio-chemical processes of the indicated phenomenon of regeneration of pancreas and in cyto-protective effect. At the base of spin chemistry and chemical radio-physics lays manipulation of electronic and nucleus spins. When such manipulations are performed by the chemical reaction itself, magnetic-spin effects appear and among them – generation of microwaves when the reaction becomes molecular radiostation.

When manipulation with spins occur under microwave influence then other new magnetic effects of the second generation are produces. They serve as indicators of microwave radiation receiving. Spin chemistry and chemical radiophysics are closely related however they have their individual tasks. The former develops new principles of chemical reaction management (including microwaves) whereas the latter possessed significant applied bio-medical aspect to it.

In relation to version (?) of molecular – tissue radiostations vital question arises about highly permeable ability of modulated wideband electromagnetic radiation (MWER). As far as we remember one group of the animals (rats) in our experiment was placed in to isolated concrete ground level chamber and nonetheless the effect of MWER on the animals was authentically and reliably recorded. If biologically (morphogenetically) part of MWER occupies microwave diapason of Zeeman's reservoir then that region of MWER spectrum would have been filtrated by the concrete walls of that ground floor bunker where the rats were placed at the time of wave irradiation on them. Yet the rats perceive/receive the radiation. How?

Possible explanation could be that electrons og Zeeman's level energies of all metabolites including genetic structures whilst existing in a potential "energy pit/cavity" experience supe fine Lamb shift of those levels by about 1000Mhz. This is allowed by existence of longitudinal atomic kernels of photons causing its longitudinal (electrostatic) field which dipolarly excites the vacuum and with this excitation interact moving orbital electrons of atom. They in turn possess their own electrostatic fields composed of similar longitudinal photons.

This way the atomic system of electrons (alternating in time) induces around itself a composit alternation longitudinal electric field which in the form of longitudinal electric wave (LEW) moves instantly in the entire surrounding environment infinitely. Umov-Poynting vector of this wave is equal to zero, that is energy-impulse of this given atomic system is not emitted. However there exist vortexes of longitudinal electric Maxwell's field described by the substantial part of bi- quaternions [Bezin et al, 2003] which are able to transfer information/data which has numerical energy equivalent of Shannon and Brillouin. Yelded in such a way LEW possessing abnormally (incredibly) high permeability properties, passes almost without attenuation through various obstacles (metallic screens, ferromagnetics and dielectrics, concrete and so on). Cellular nucleuses and their main component – DNA – excite solitons [Smelov, 2001] of the linked electron waves and hyper-acoustic fluctuation of liquid crystal chromosomic structures, that is electron-vibron fluctuations [Bersuker, 1976] or electron-nucleus wave fluctuation of DNA's double helix. The said electron-vibron waves retransmit (dissipate) the received LEW back in to the vacuum (ether) and can be received by the other bo-systems similar to the first one (radiated by the primary LEW).

Owing to high corpulence (?)~1014 of all electron-vibron fluctuation systems, they possess high sensitivity estimated in terms of Planck's quantum piece energy for one element of coherent fluctuating chain which can be a DNA spiral or cellular membrane. In opend states DNA spirals, initiated by thermal motion in a living cell, electron-vibron fluctuations exist in the form of solitonic (helicoids, spiral, vortical) motions of atoms in the chain. These are so called Salerno-Maslovs solitons and these are capable of readinginformation from DNA-RNA "text" sequences. Emission/radiation by such "informational" solitons is generated by electron-vibron DNA and RNA fluctuations. Herewith scanned/read information from genetic molecules can be transmitted to other cells (and beyond the bio-system) in a relay mode employing mechanism of dissipation with LEW's frequency alterations irradiated in to the area of radio frequencies and acoustic fluctuations. Along with this, informational radio emission of electron-vibron oscillations in the form of LEW at certain frequencies, in principle, can govern/operate biochemical processes. And the other way around, biochemical reactions in preparations being probed by the polarized laser light can generate electromagnetic radiofrequency oscillations.

Due to probable wave genetic-symbolic bearing of the vortical solitonic states, initiated on DNA and RNA molecules in vivo, let's look at these processes in more details. [Blagodatskix, Gariaev et al, 1996; Gariaev, 1997]

Quantum teleportation of genetic-metabolic information/data in permissive version.

By Peter Gariaev (https://www.facebook.com/Peter.Gariaev) in Wavegenetic (https://www.facebook.com/groups/wavegenetic/) **· based on works of [Prangishvil ,Gariaev, Maksimenko et al., 2000(б); http://www.wavegenetics.jino-net.ru/zip/Teleport.zip (http://www.wavegenetics.jino-net.ru/zip/Teleport.zip)**

The abovegiven biological experiments on remote transmission of morphogenetic, or more accurately – genetic-metabolic information/data from a donor to recipient may be interpreted from views of non-localized contacts in accordance with mechanics of quantum teleportation in permissive version/mode. The laser device, practically – quantum bio-computer (of which we mentioned before), apart from unique capability of wave transfer of morphogenetic information, can also perform a conversion of red coherent photons into radio waves of rather wide spectrum. These very waves – are candidates for the first act of scanning and transmitting.

When scanning bio-samples or some other substances with the laser light, in essence a biologically active spectroscopy is performed of a distinct kind. We propounded a tentative explanation of this phenomenon [Gariaev.

Tertishnii, Gotovskii, 1997], which is primary substantiation of the new king of spectroscopy - called polarized laser-radio-wave spectroscopy (PLRS) [Prangishvili, Gariaev at al., 2000 (b)]. Such spectroscopy is designed to research and examine unknown spin-fluctuating quantum-molecular characteristics of bio-systems, bio-tissues, hard, liquid and gas substances, and also plasma conditions/states. The type of PLRS that we describe here takes a narrow optical diapason – red light, yet it is planned in the future that modification from UV to IR diapason be available.

This present version of the device – He-Ne laser ( ) with generation of two orthogonal linked in their intensity optical modes, which are mobile and depend on the object being probed and are connected in that way so the total of their intensities remains constant irrespective of preparation (bio-sample, tissue) being scanned. When even one of the modes is interacting with the substance, reflected and dispersed emission of which is returned in to optical resonator, redistribution of intensities of these optical modes occur in line with the law of polarization change, attributed to the new state/condition after the interaction of the laser beam with dynamic micro-polarizers, which are located on section/cut of illuminated area of the sample in question. Dynamics of the micro-polarizers is determined by internal dynamics of the object being examined. (metabolism, acoustics of chromosomes and DNA in vivo-in vitro, fluctuations of crystal grids and so on). One of the laser modes, in regime of photons being returned back to resonator, is capable of causing emission/irradiation by our device of modulated radio waves in wide spectrum, correlated with modulations in optical modes of emission in our laser. (Rather complex sentences-the meaning may be lost – rephrase-translator's remarks)

These modulations depend on spinning fluctuating micro-structural components (for instance, of LCC domens in vivo-in vitro) of substances under examination and level of their optical activities. Frequency interval of radio waves created when converting photons in accordance with theoretical model (further below) lies in diapason from 2 up to 0. The maximum of such irradiation is located in the region of 1 MHz. Radio wave signal having been detected is passed to a CPU of a computer with designated software for the signal processing. At the other end Fourier spectrum of radiowaves is recorded, charachterising polarized-dynamic features of substances with which one of the laser beams is interacting and also a spectral memory of emitting substances. The second light beam is returned to the resonator of the laserfor creating a resonance interaction wit hatomic oscillators of the gas mixtures. The reason for photons being converted in to radio waves, as we believe is nonw-stiff dispersion and localization of the light of the main laser modes upon the system of non- homogeneity of the mirrors in the laser resonator. Mechanism for localization (localization in the non-stiff dispersion channel) is given further on. In particular we believe that in the resonator there exists a stiff-dispersed localized light.

The radio wave emission that is generated by the laser is able to "read information", for example, from DNA samples or organs or tissues. The reading mechanism resembles process of usual induced radiation. An option of closing and opening laser's resonator allows localization or "record" in it spectrums of various objects/samples under examination. Radio wave radiation is able to scan and transmit such spectrums. Then we discovered an effect of spectral memory; during a certain macroscopic lengh of time one reproduces radio wave spectrums of objects in question, which reflect the laser beam back into resonator and than removed from the exposition zones. In this way we managed to record DNA spectrums and revealed their highly bio-activities, more likely to be linked with wave mode of transmission of genetic-metabolic information/data. [Gariaev, Leonova, 2003]. IT appears that quantum non-locality of genetic (chromosome) information/data (as manifestation of its total distribution (continuality) inside multi-cell bio-systems), is a particular instance. In reality, there are at least6 levels of non-locality in bio-systems.

Level 1 – organisms' level. Non-locality here is expressed in ability to regeneration, for instance in Planaria worms. After cutting of these organisms, any part of their body reproduces the whole body. In other words in this case there is no an attachment of the common pool of genetic information to any particular part of these worms. The same applies to vegetative reproduction in plants.

Level 2 – cellular. Out of any cell (and not only from zygote) one could grow the whole organism. For living biosystems it is more difficult, however possible. Every cell is a potential continuum of the organism.

Level 3 – cellular-nucleotide. Enucleation of nucleuses from somatic and gamete cells with subsequent injection of other nucleuses does not hinder a normal development of an organism. Cloning of this kind is already being performed on the high bio-systems, sheep for instance. Every cellular nucleus – is a potential continuum of a bio-system. Localization of genetic potencies on certain kind of cells is not the case. Gamete cells perform the same role but with haploid set of chromosomes, combining in zygote in to double set, the same as in somatic cells.

Level 4 – molecular: ribosome "reads" informational RNA not only on separate codons, but the whole of it in certain

context, that is – non locally, but continually.

Level 5 – chromosomic-holographic. Genome possesses holographic memory [Gariaev et al., 1991; Gariaev, 1994], and this is tipically distributed (non-local) associative memory. At this and the subsequent levels non-locality obtaines new quality, dualistic material-wave/field character since the holograms I the same manner as materials are scanned by electromagnetic or/and acoustic fields, which carry genetic-wave information/data beyond material substance of chromosomes. Here we deal with physical field or fields, as calibration formations marking future space-time of an organism. It appears that holographic memory of cortex pertains to this fields as well, which sets mental, rational and imaginary spaces, which calibrate potential actions of higher bio-systems in a socium (?) as a super-organism. This is the highest level of socio-genetic processes.

Level 6 – quantum non-locality of genome. Up to the level 6 genetic information/data non-locality is realized in space of the organism and also in socium. Level 6 has a special character and new quality. It is expressed in one of the forms of quantum non-locality, namely – pemissive, postulated in this paper. In that case non-locality is instantly realized across the space of the bio-system as much as on its own, "compressed" until zeroed - time. Instantaneously distributed in this manner genome-wave programs, isomorphic materially, perform in the organism "here and there at the same time', therefore it is meaningless to use the semantic structure "now and then". This is a strategic factor, incredibly vital evolutionary achievement for multi-cell bio-systems. Billions of cells of the organism ought to "know" one another and about one another if not everything than at least a lot and instantly. Without phenomena of "wave informational instantaneity" gigantic multi-cell continuum if the highest bio-systems is not capable of holistically coordinate metabolism, its own physiological and other functions. Intercellular diffusion of signaling substances and nervous processes are overly slow, inert for these purposes. Even if we suppose that symbolic electromagnetic field as quick as speed of light (which is sensible supposition) take part in intercellular exchange, then this is not sufficient either. It must possess a mechanism of quantum non-locality and it can be applied to genetic apparatus, which may perform as an instantly distributed quantum (wave) object, isomorphic to material chromosomes.

Utilizing non-locality, genetic apparatus of the highest bio-systems creates amazing phenomenon, when at certain moments in "compressed" space-time of bio-system "here and there", "now and then" are working as a continuity providing for the organisms super-coherence, informational super-abundance, super-informed, connectedness and as a result integrity (survival). As a manifestation of these phenomena one can observe regenerative abilities in organd and tissues in inferior organisms (hydra, worms, amphibians, lizards, Crustacea), the ability which has largely been lost by humans. However, it may be activated considering principles of wave self-organization processes that we advance and develop in this paper and in our experiments.

As an illustration of these phenomena serves the very first successful grafting of implanted donor's tissues in a blind human with recovery of vision [Muldashev, 2000]. As a foundation (ideology) for such surgeries and regenerative processes were used our models of morphogenesis and holographic memory of genome. The key moment of this surgery was utilization of donor's retina as a hologram and intercellular substance (alloplant) as systems of additional coordination post-embryonic morphogenesis of the eye, which was in fact predicted earlier [Gariaev, 1994].

Meanwhile these theoretical-experimental researches here are tentative and initial and require physics-mathematical foundation and consideration and advancement. For that reason we show formalized model of photonic-radio-wave processes, brought in to existence when laser beam is interacting with the substances, processes, which can be view as foundation of PLRS and one of the version of quantum informational events in chromosomes.

The fundamental grounding of our ideas lies in theories of localization of light in dispersing space-correlated systems. The Phenomenon of localization of light received a wide acknowledgement since 1985 after works of [Albada, 1985]. Today it is one of most dynamically developing spheres in physics, closely interrelating with such issues as, for instance, quantum teleportation, new methods of recording and reading information and so on.

- **Mathematical Modeling of solitons on DNA**
- **By Peter Gariaev (https://www.facebook.com/Peter.Gariaev) in Wavegenetic (https://www.facebook.com/groups/wavegenetic/) ·**
- **Mario Salerno was the first to experiment with solitons on DNA on computer not only as formal mathematical structures but also tried to link their behavior in a mono-dimensional space of poli-nucleotides with their bio-genetic, or more accurately with their epi-genetic functions. He had been developing the first soliton model on DNA proposed by Englander and et al. This model and its more detailed variations including ours (see further**

below) is represented in term of mechanical systems as a chain of oscillators (DNA base) linked by resilient non-linear sugar-phosphate connections. Following Salerno we have paid particular attention to objectively existing known DNA sequences and thei effects onto solitons behaviour.

In the first stage we have replicated his experiments yet on significantly longer DNA parts. Indeed, solitonic kink excitations are sensitive to the place of their initiation; their motion along one of DNA chains (when they are opened as a result of temperature fluctuations) is accompanied by specific modulation of kink trajectory in time and in single-dimension space of poly-nucleotides. Such solitons are structures that emit electromagnetic and acoustic fields; their inner oscillation structures are capable of reflecting and retranslating texts and other symbolic structures of DNA in to into and out of cellular space at least at the level of large blocks of sequences. As an example behavior of kink on a fragment of singular DNA chain, which is 1020 nucleotides long from a virus of bird sarcoma

C (or S) area of DNA (1 1020 nucleotide) on 3'-end of virus of bird sarcoma contains several "semantically" certain areas such as polypeptide-coding area (between 558 and 675 nucleotides); PolA (936) 3'-end of virus RNA, poly-adeniling site; 916 nucleotide - 5'-end of virus RNA ("capping site"); Red-sector - () – short end replica of the virus genome; Pro – possible component of observing the transcription (between 870-9000); palindrome – "pin" (870 912) .

(5' beginning) GGC CTA TGT GGA GAG GAT GAA CTA CGT GCA CCG AGA CCT GCG GGC GGC CAA CAT CCT GGT GGG GGA GAA CCT GGT GTG CAA GGT GGC TGA CTT TGG GCT GGC ACG CCT CAT CGA GGA CAA CGA GTA CAC AGC ACG GCA AGG TGC AAG TTC CCC ATC AAG TGG AGA GCC CCC GAG GCA GCC CTC TAT GGC CGG TTC ACC ATC AAG TCG GAT GTC TGG TCC TTC GGC ATC CTG CTG ACT GAG CTG ACC ACC AAG GGC CGG GTG CCA TAC CCA GGG ATG GGC AAC GGG GAG GTG CTG GAC CGG GTG GAG AGG GGC TAC CGC ATG CCC TGC CCG CCC GAG TGC CCC GAG TCG CTG CAT GAC CTT ATG TGC CAG TGC TGG CGG AGG GAC CCT GGA GGA GCG GCC CAC TTT TCG AGC TAC CTG CAG GCC CAG CTG CTC CCT GCT TGT GTG TTG GAG GTC GCT GAG TAG TGC GCG AGT AAA ATT TAA GCT ACA ACA AGG CAA GGC TTG ACC GAC AAT TGC ATG AAG AAT CTG CTT AGG GTT AGG CGT TTT GCG CTG CTT CGC GAT GTA CGGGCC AGA TAT ACG CGT ATC TGA GGG GAC TAG GGT GTG TTT AGG CGA AAA GCG GGG CTT CGG TTG TAC GCG GTT AGG AGT CCC CTC AGG ATA TAG TAG TTT CGC TTT TGC ATA GGG AGG GGG AAA TGT AGT CTT ATG CAA TAC TCT TGT AGT CTT GCA ACA TGG TAA CGA TGA GTT AGC AAC ATA CCT TAC AAG GAG AGA AAA AGC ACC GTG CAT GCC GAT TGG TGG AAG TAA GGT GTA CGA TCG TGC CTT ATT AGG AAG GCA ACA GAC CGG GTC TGA CAT GGA TTG GAC GAA CCA CTG AAT TCC GCA TCG CAG AGA TAT TGT ATT TAA GTG CCT AGC TCG ATA CAA TAA ACG CCA TTT GAC CAT TCA CCA CAT TGG TGT GCA CCT GGG TTG ATG GCT GGA CCG TCG ATT CCC TAA CGA TTG CGA ACA CCT GAA TGA AGC AGA AGG CTT CATT 1020 (3'-ending)

In Fig.1 and 2 the kinks appear in a form of "mountine ranges" rather tha as steps, because a derivative of sinus-Gordon equation function was taken. In here a horizontal axis – DNA sequence, vertical – soliton amplitude. Axis to the observer – time. It is visible that wnen soliton's initiation location is altered at certain sequences of poly-nucleotide, the dynamics of that single range/wave is notably alters too in a form of spinning-fluctuating motions along DNA sequence.

The molecul's area under this research is rich functionally (semantically) by biologically significant sectors and we are reasonably expecting that these sectors will alter, modulate, that is – will be introducing/injecting in to a single chain of DNA or RNA "textual" information. This such aspect will be realized in the spectrum of solitonic wave fluctuations alongside poly-nucleotide chain. This spectrum will mirror sequences of nucleotide and by that will perform a role of the carrier of genetic information/data/messages. Such modulation of the fluctuation structures is clearly observed in the presented graphs. It is plausible to believe that spectral composition of oscilating frequencies of solitons appear to be one of the mechanisms of text conversion of DNA and RNA structures in to wave form and a means of transmitting genetic and other messages in a mono-dimentional space alongside the poly-nucleotide chains and/or in 3D of the genome as a single cell as well as the tissue continuum of the bio-system.

The same as Fig.1, apart from the epicenter of the perturbation – 450th nucleotide in here.

This is how the soliton dynamics computer modeling works developed by Salerno after it had been propounded by Englender. Salerno gave formalism describing spin nucleotide fluctuations of DNA molecule in order to explain experimental data in hydrogen-tritium exchange in DNA. In accordance with that model, Engleder, in DNA chain

> **there can appear (under thermal radiation/noise) and distribute open states ("melting" the DNA double-helix in short segments enriched by AT-couples) in the form of localized dislocations.**
>
> **Mario Salerno continuing works of Englender in a simplified form, discovered influence of nucleotide sequence on none-linear solitonic dynamics of spin fluctuations of nucleotides within (однотяжных- word is not known by the translator)… DNA areas, which form such open fields (open states). Later, Yakushevich, Fedianin and Khomma et al. revied various generalizations of Englender's model assessing specifics of DNA structuring taking in to account a break in hydrogen connection when the base is opened, paired DNA chains and other aspects of reedom distinct from spinning. However, as it seems, it is not shown in those works the reasons for emergence of dislocations in DNA.**
>
> **We put forward a possible mechanism of that process in DNA, which is alternative to Englender's hypothesis about thermal noise influence as a cause for opening the bases of couples. We think that dislocations in DNA can be triggered when the period of DNA double-helix is altered (Maslov's proposition). In our model DNA nucleotides are viewd as oscillators suspended on masless non-extendable rod/pivot; sugaf-phosphate links between neighbor nucleotides in the chain are modeled by linear springs, spiralisation along the chain is not taken in to account; hydrogen links between complementary bases are modeled by "gravitational" potential. Hamiltonian, as in accordance with M. Salerno, liiks like the following:**
>
> **http://eng.wavegenetic.ru/index.php?option=com_content&task=view&id=23&Itemid=1 (http://eng.wavegenetic.ru/index.php?option=com_content&task=view&id=23&Itemid=1)**

EXPLORING WAVEGENETICS AND WAVE IMMUNITY.By Peter Gariaev (https://www.facebook.com/Peter.Gariaev) in Wavegenetic (https://www.facebook.com/groups/wavegenetic/) · Edit Doc (https://www.facebook.com/ajax/docs/inline_edit_get.php?group_id=129999910455329&doc_id=132895870165733&permalink=1&show_brief=0)THEORETICAL MODELS *.

Gariaev P.P., Kokaya A.A., Leonova-Gariaeva E A., Muldashev E.R., Mikhina I.V.,

Smelov M.V., Tertishnii G.G., Tovmash A.V., Chalkin S.F., Shatrov Y.K., Yagujinski L.S.

*** Disclaimer: This translation and English text editing have been done to the best knowledge and understanding of the translator who is not a specialist in the field although having a good understanding of Russian language and advanced level of English complemented with rather good understanding of basic physics and biology. The reader is encouraged to seek specialist opinion in any event. Extensive resources are used to understand the concepts in order for the translation to be s precise and accurate as possible. Where possible Russian abbreviations are to be preferred as they represent the most accurate references. An approach of translating the concepts, not the discrete words, has been employed in a belief that this is the most appropriate form of conveying a thought and/or concept from one language into another; of course a detour from this mode is taken where deemed necessary for the sake of clarity.**

All authors have equally contributed to the research in this paper.

Many researchers have pointed out in their works to the dangers posed by geptil pollution to the environment and that it adversely and detrimentally affects health of livings organisms [http://www.seu.ru/conference/ecoprvo/geptil.htm (http://www.seu.ru/conference/ecoprvo/geptil.htm)]. In this present paper a possibility of principled approach to the development of technology is demonstrated, which allows bio-systems to trigger resistance mechanisms to protect the organism from the toxic impact of geptil substance by mean of electromagnetic fields/waves. Theoretical foundations for conducting this kind of research are illustrated on the following website http://www.wavegenetics.jino-net.ru/ (http://www.wavegenetics.jino-net.ru/). Prior to commencing this research/investigation in this direction, we had organised a series of initial model experiments with alloxan. Alloxan is a cytotoxic substance with a dominant detrimental effect on β-cells of pancreas resulting in to Type 1 diabetes.

The earlier experimental works [Gariaev P.P., Kokaya A.A.] have suggested that modulated by pancreas and spleen – modulated wideband electromagnetic radiation (MWER) generated by helium-neon laser influences upon progression of the experimental diabetes in rats. Diabetes is provoked by intro-peritoneal/abdominal injection of alloxan in a dose of 200mg/kg of the mass of the animal.

Irradiating rats with such a frequency resulted in prolonged life expectancy of the animals in the test groups when compared to the control group, leading to normalisation of blood glucose level and, crucially, facilitated pancreas tissue regeneration process.

The aim of this present paper is to assess the phenomenon of resistance to alloxan developed in animals having been previously preventively irradiated by MWER. For that purpose a special laser was used which features interconnected complementary orthogonal polarization of the light beam. Generation of the WER (wideband electromagnetic radiation) was conducted in accordance with Fabri-Perot interferometer's scheme, where the operational laser ray many times passed through the thin freshly-prepared sections/samples of pancreas and spleen from a healthy rat. We believe that the samples specifically and uniquely modulate the laser ray in such a way that this system reveals the following capabilities:

1) reinforce/enhance WER released from discharge interval of He-Ne laser;

2) WER is parametrically connected with the sample-modulated laser ray and as a result it obtains a property of being biologically highly active;

3) biological effect can be observed at relatively long distances from the source of WER;

4) Genetical-metabolic managing/primary information is transmitted from bio-donor to the bio-recipient. The information is carried by MWER (modulated wideband electromagnetic radiation) modulated by the bio-object from the donor

As a donor/bio-structure tissue, probed by the laser beam in the present system, following can be used: living and/or quasi-living organisms, for instance, bacteria, viruses and also living through tissues and organs, metabolites and abiogenic substances.

RESEARCH METHODOLOGY. PHYSICS' PART

To obtain WER modulated by the bio-structures/samples we utilized developed earlier biotechnology with use of helium-neon laser [Gariaev P.P., Tertishni G.G. 1999]. Helium-Neon laser of 2 MW with 632.8 nm of the wavelength possesses two superposed, orthogonal linear polarized modes of emission, single-frequency in each of them. Laser ray probes/scans DNA samples/bio-structures - newly retrieved tissues of pancreas and spleen of a newborn rat Wistar strain. Semitransparent tissues were deposited onto a a lab glass and then covered that by a second lab glass and this "sandwiched" object was fixed in front of the optical axis of the laser. Adjustment of the glasses with the tissues had been done so as to provide partial reflection of the modulated (by the tissues) laser beam back in to the laser resonator. In this way we ensured a multiple passages of the laser beam through the tissues and the tissues be an optical correlator [Mazur, Grachev, 1985] and therefore affect redistribution of the secondary modes of the laser emissions.

Optical signals were registered and transmitted to an electronic circuit, which governs laser's generation regimes; frequency stabilization of the coherent radiation is also performed at this stage. In this mode the laser generates, apart from the red light, WER modulated by the tissues – that is – the MWER itself. Distance from the target to the active element of the laser is 11cm.

RESEARCH METHODOLOGY. BIOLOGICAL PART

Wistar strain of rats matured to reproductive stage, 5-6 months old were used, with average mass of 180-220 gram. Diabetes for the purposes of this experiment was provoked by way of intro-peritoneal/abdominal injection of alloxan in a dose of 200mg/kg of mass of the animal after 24-hour fasting period with normal/common levels of blood glucose level. The animals were divided into 4 groups:

Group 1 control (n=20) – no WER irradiation; Group 2 (n=20) and Group 3 (n=20) – were subject to preliminary irradiation of WER; Group 4 (n=10) – placebo, where WER was not modulated by DNA samples and the laser ray passed through empty lab glasses without the tissues.

Group 2 rats were placed 20 meters from the source of MWER in the lower ground level of the lab. Alloxan diabetes was provoked one month later from the day of the last irradiation by MWER.

Group 3 and Group 4 were placed 70 cm from the source of MWER. Alloxan diabetes in these groups was triggered one day after the last irradiation by MWER.

Irradiation by MWER of Group 2 and Group 3 was being performed daily for 30 minutes during 4 days scheduled as follows: 10 minutes irradiation by MWER (modulated by samples of pancreas), 10 min irradiation by MWER (modulated by samples of spleen) and 10 minutes irradiation by MWER (modulated by samples of pancreas).

Group 4 – placebo – was being irradiated by WER – not modulated by any samples the laser beam though passing through the empty lab glasses – for 30 minutes daily during 4 days.

Group 1 – control - was not subject to the irradiation by WER nor was it by MWER.

During the experiment we assessed general health condition of the experimental animals, recorded the death date for the animals from the moment of alloxan injection, in all observed groups. The animals of Group 2 and Group 3 were observed for 1.5 months from the day of alloxan injection. In 8 animals from Group 2 and 3 were recorded with the reproductive function still active when the blood glucose level reached its peak value. (3 rats from Group 2 and 5 rats from Group 3).

Measurements of the blood glucose level were performed by glucometer Ascensia Entrust made by Bayer. Measurement range of the glucose level spreads from 2.0 mmol/l up to 30.6 mmol/l. Measurements of glucose level higher than 30.6 mmol/l were marked as HI.

Removal of heart- lung- liver- kidney- spleen- and pancreas- tissues was done in order to perform microscopic description and histological analysis in:

Group 1 – control – on day 3 and 4 after alloxan injection which corresponded to the highest death rate of the animals;

Group 2 and 3 on the day 8 from the alloxan injection, and also on the 42 day of the experiment after assessing reproductive function in male animals.

For histological analysis purposes the tissue had been fixed in 10% neutral formalin, dehydrated in spirits of increasing concentration and then sealed them in paraffin wax. Paraffin sections of 5-7 micro kilometers thin were obtained on microtome Leica SM 2000R, were colored in hematocsilin-eosin and then analysed by microscope Leica DMLS. Video footage was obtained by using CCD-cameras.

Statistical survey of the experiment's results were conducted by means of using statistical software suits "Stastica 6.0", MS-Exel" for Windows. Degree of accuracy (p) was determined using Student criterion, employing confidential coefficient and a digit of degree of freedom (1) in accordance with table. Calculations of all mathematical values had been performed in compliance with commonly known formulas on a PC.

Research in this series of experiments revealed that application of the said dose of alloxan in control and placebo groups developed diabetes complicated by toxic damage of live-crucial organs and systems. This led to low survival rate of the animals in within the groups. On the contrary, in Group 2 and 3 we observed resistance/increased immunity in the animals to the detrimental/destructive effects of alloxan expressed in various degrees.

Group 1 – control – survival rate after alloxan injection of the animals on the day 2 was 55% yet by the day 4 it went down to 30% (Fig.1). Glucose level in Group 1 animals on day 2, 3 and 4 is reliably distinct ($p<0.05$) from the initial value/level (Table 1). Animals dying in Group 1 (control) at terminal stage were put down by euthanasia (5 rats), organs used for patho-morphological analysis. There was no discrete/spontaneous reduction in glucose level in Group 1 (control) within the period of monitoring. However, there was one rat exhibiting resistance to alloxan and its blood glucose level remained within the limits. Table 1. Blood Glucose level over time in animals after alloxan injection, dose of 200mg/kg of the mass

*** - blood glucose level in Group 3 by the 2nd, 3rd, and 4th day of initiating alloxan diabetes is different ($p<0.05$) from the level of glucose in blood of the animals in Groups 1 and 2 by the 2nd, 3rd, and 4th day, and is also different ($p<0.05$) from glucose level in blood from the animals in Group 4 by the 2nd day;**

**** - blood glucose level in Group 1 by the 2nd, 3rd, 4th day is in fact different ($p<0.05$) from the initial level;**

***** - blood glucose level in Group 2 by the 2nd, 3rd, 4th day is different ($p<0.05$) from the initial level;**

****** - blood glucose level in Group 4 by the 2nd day is different ($p<0.05$) from the initial level;**

^ - in Group 4 by the 3rd and 4th day of observance there was one survived rat.

All animals were injected alloxan – 200mg/kg of an animal mass. Group 1 (control) was not subject to WER or MWER. Group 2 – MWER was conducted; the animals were places 20 meters from the source of emission. Induction of alloxan diabetes was effected one month later after the last radiation by MWER. Group 3 – MWER was conducted, animals were located 70 cm from the source. Induction of alloxan diabetes was done one day after the last radiation by MWER. Group 4 – placebo – was irradiated by WER – not modulated by DNA samples (that is to say that the laser beam was passing through the empty lab glasses). The animals were placed 70 cm from the source of the emission. Alloxan diabetes was induced a day after the treatment.

Animals in Group 4 – placebo - were observed with altered glucose level different from the initial one ($p<0.05$). 80% of the animal survived by the day 2 and only 10% remained alive by day 4. This was considerably at variance with what we observed in Group 2 and Group 3; Group 1 exhibited lower death rate and survival was at 30%. (Fig. 1)

Preventive irradiation by MWER significantly influences the alloxan diabetes progression in the animals of Group 2 and Group 3 (Figs. 1, 2, 3, Table 1) and is accompanied by safeguarding, cyto-ptotective effect. (Fig. 4). That had been seen in both abovenamed Groups despite the fact that diabetes in Group 2 was induces a month later after the last preventive irradiation/treatment

Survival rate in Group 2 (n=20) reached 90% by the 3rd and 4th day from the moment of alloxan injection (Fig.1) which is significantly distinct from survival data obtained from the control Group (30%) and in the Group 4 (placebo) (10%). Glucose level increase ($p<0.05$) is observed in Group 2 animals by day 2, 3 and 4 when compared with the initial value (Table 1). By the 4th day 13 animals in Group 2 (65%) the glucose level in their blood was 14.5 mmol/l, whereas 5 animals (25%) from the same group were observed with normal physiological level of glucose (Fig.2)

Glucose level in Group 2 by the 4th day is ($p<0.05$) different from the initial level. (Table 1). By the 7th day in Group 3 survival rate decreased to 75% (out of 20 animals with hyperglycemia 5 animals died) and remained at the same level up to the end of observation term (1.5 months Fig.1). By the 8th day euthanasia had been conducted to 6 animals from Group 2 and tissues were extracted for patho-morphological analysis. The other 9 animals were observed for a period of 1.5 months. It is important to note that from the day 8 up to 15th animals with hyperglycemia a reduction in glucose level was observed (Fig. 2).

However by the 18th day 4 animals of Group 2 regained the hyperglycemia (more than 30,6 mmol/l) which remained during the entire observation period. The overall health condition of these 4 animals were assessed as satisfactory. Similar results were present during our previous experiments [Gariaev, Kokaya et al, 2007]. The other 5 animals maintained normal level of glucose.

Reproductive functions were noted in 3 animals of Group 2 with hyperglycemia (glucose level over 30.6 mmol/l). These animals gave birth to healthy multiple posterity. After 1.5 months tissues were extracted for the analysis.

Fig.2 Effects of preventive/preliminary irradiation by MWER on progression of alloxan induced diabetes in Group 2.

Alloxan was injected at 200mg/kg dose one month after the preventive irradiation by [MWER] modulated wideband electromagnetic radiation/field. The irradiation was being performed for 4 days, 30 minutes in each day. Distance from the radiation source – 20 meters. The irradiation schedule: 10 min irradiation time using tissues of the pancreas, then 10 minutes of the spleen tissues, and for 10 minutes – pancreas tissues. The beginning is the day of the alloxan injection.

The most manifested effects were observed in Group 3 located at 70 cm distance from the source of the modulated waves. Alloxan diabetes modeling was performed one day after the irradiation by MWER ended (Fig.1,3, Table 1). Every animal in the Group survived. We observed 100% survival rate during the entire period. It is clear that 90% of animals retained their normal physiological glucose level during the 1.5 month of observation, which is clearly different ($p<0.05$)

from the control Group, Group 4 and Group 2.

However 2 animals from Group 3 by the day 6 of the experiment had excessive glucose level increase of 20 mmol/l with subsequent reduction to the normal level. On the 8th day of the experiment6 animals were performed euthanasia for a patho-morphological analysis. Reproductive function could be detected in 5 animals of Group 3. All the animals gave birth to a healthy posterity. 1.5 months later tissues from 6 animals from the same Group were taken for the analysis. Glucose level was at normal levels. Health condition of all the animals of Group 3 was satisfactory.

Fig.3 Effects of preventive/preliminary irradiation by MWER on progression of alloxan induced diabetes in Group 3.

Alloxan was injected at 200mg/kg dose one day after the preventive irradiation by [MWER] modulated wideband electromagnetic radiation/field. The irradiation was being performed for 4 days, 30 minutes in each day. Distance from the radiation source – 70 centimeters. The irradiation schedule: 10 min irradiation time using tissues of the pancreas, then 10 minutes of the spleen tissues, and for 10 minutes – pancreas tissues. The beginning is the day of the alloxan injection.

After histological analysis of the pancreas tissues in Groups 1,2 and 3 a number of distinctive features have been observed (Fig.4). Histological picture of the pancreas tissues from animals of control Group was cjaracterised as having clearly visible degenerative alterations of Langer Islands (Fig.4b). The number and size of the islands are reduced, they are of unusual and irregular form. Quantity of β-cells is sharply reduced and in most of them cytoplasmic vacuolation was observed, the nucleus size decrease, chromatin condensation and cariopicnose in some of the cells. Gathering of lymphocyte infiltrate had been revealed around and inside of the islands.

In Group 2 by the 8th day histological situation if pancreas tissues could be described as having damaging processes of various degree: the islands were reduced in size, of irregular form, β-cells reduction, the total portion of insulin apparatus in the islands was significantly reduced. Only a relatively small part of the islands apparatus maintained rather preserved (intact) structure (Fig. 4c).

Situation in Group 3 by the 8th day after modeling alloxan diabetes was considerably different from the control Group and Group 2. Along with the pathological processes in the tissues of pancreas in question, there was also observed a large quantity of islands, large as well as medium and small sizes with lightened cytoplasm, of normal spherical form, large roun nucleuses containing the core (Fig.4e).

A month and a half later histological situation in Group 2 was also described as having damaged islands' apparatus (Fig.4d)

As opposed to Group 2 and control Group, Group 3 after 1.5 months had hypertrophy and hyperplasia of the pancreas. Numerous islands of varying size and normal spherical form were noted (Fig.4f). Attention was attracted by the large number of small islands and discreete agglomeration of β-cells, where as large islands contained increased number of β-cells which in turn were located in a close proximity to one another. The structure if the islands and β-cells were not altered, the kernels in the cells were large; round, in which it was possible to locate nucleuses.

Fig.4 Pancreas tissue structure, Langerhans islands:

a – intact rats:

b – Group 1 (control), after alloxan injection of 200mg/kg dose;

c – Group 2 by the 8th day from the day of alloxan injection of 200mg/kg dose. A month prior to the inducing alloxan diabetes, this Group was preventively treated by the irradiation and was located at about 20 meters distance from the source I the basement of the laboratory.

d – Group 2 after 1.5 months after alloxan injection of 200mg/kg dose.

e – Group 3 by the 8th day from the day of alloxan injection of 200mg/kg dose. One day prior to the alloxan injection this group was also preventively treated by the irradiation being located at 70 centimetres distance from the source of radiation.

f – Group 3 after 1.5 months after alloxan injection of 200mg/kg dose

Enlargement: 1 x 400, 1 x 100. Colouring agents: Haematoxylin and Eosin.

DISCUSSING THE RESULTS

It can be stated that a positive effect was reached by using MWER in Groups 2 and 3. Differences in the dynamics of glucose level and survival rates of the animals in these two groups point to an interconnectedness between the length of the irradiation by MWER and modeling the alloxan diabetes. Biological long range defensive effects when irradiating with MEWR which were discovered in our previous experiments [Gariaev, Kokaya et al, 2007] has been confirmed in this series of experiments – this research. The effects is manifested in the glucose level dynamics in Group 2 animals and also in the very fact of the animals surviving in this group, when compared with control Group and placebo Group. In spite of the reliable data ($p<0.05$) suggesting increase in glucose level in the animals of Group 2 compared to the initial state, and an absence of any reliable differences in the glucose level data obtained from Groups 1 and 2 by the 2nd, 3rd and 4th day of the alloxan injection, survival rate in the Group 2 was high. Evidential hyperglycemia in 20% of the animals in this group did not result in to their death but were remaining I a satisfactory health condition for the entire period of the observation.

Irradiation by the MWER on the animals of Group 3 facilitated development of resistive properties of their immune system to the alloxan triggered diabetes and the glucose level was at the normal physiological lever at all times for the entire period of observation (this is reliably different ($p<0.05$) from data obtained from Groups 1, 2 and 4). Survival rate in Group 3 was at 100%. Having analyzed the histological data of pancreas from different Groups, it can now be stated that preventive irradiation by MWER in Group 3 resulted not only in cyto-protective effect on the cells of pancreas but also encouraged hypertrophy and hyperplasia processes in it, which, as it seems, were functioning in a compensatory mode. The present experimental findings are in good conformance with the results of our previous experiments conducted earlier. [Gariaev, Kokaya et al, 2007]

In this way, three phenomena of MWER irradiation on the animals with alloxan diabetes were manifestly observed:

Firstly – the factor of survival with contemporaneous and obvious hyperglycemia within prolonged period of observation with preserved reproductive functions in the animals.

Secondly – it was revealed in the previous experiments, and confirmed in the present one that MWER is facilitating pancreas regeneration in the ill animals in situ.

Thirdly – preliminary irradiation by MWER on the animals facilitates development of resistance to alloxan effects.

The recorder effects are related to the fundamental issues of “recording” and transmission of electromagnetic component of genetic information during the post-embryonic growth with wave (frequency) processes taking part in genome and organism as a whole. MWER, parametrically linked with photons (which in turn were modulated by the tissues), appears to be a carrier and transmitter of the information from the bio-donors to the bio-system (bio-organism which precisely and specifically receives this information as strategically superior (that is of higher topology? Of the sort that can manage and govern processes in the recipients organism)). Most likely, MWER – performs a function of a frequesncy (wave) trigger, which initiates “awaiting” (stand by) regenerative morphogenetic processes, of which the information is contained in genome of every cell. Quantum mechanisms of MWER effects on embryonic and post-embryonic processes remain to be discovered, thoughsome ideas were expressed in our earlier works. [Garyaev, 1994; Gariaev, 1997; Prangishvili, Gariaev and et al, 2000 (b); Gariaev et al, 2001; Gariaev, 2003] and are being further advances at present. The defensive and cyto-protective action of MWER are promising field of research with rather sound perspectives.

It may be that a specific role in the discovered defensive-protective manifestations plays a so called factor of “weak influence” [Chukova, 2002]. In this regard it can be put forward that the discovered effects bear endoergic character where even weakly absorbed energy of coherent polarized laser beam results in the increase of free Gelmgolc energy which is accumulated in chemical links of tissue metabolites in pancreas and spleen. For instance, atoms of informational macro-molecules (DNA, RNA and proteins), while absorbing light, together with the energy had by a quantum of light they gain the same momentum of quantity of motion, which creates inversed population of nuclear zeeman levels (see Zeeman effect). Occurs so called chemical polarization of the nucleuses. In this manner, biochemical reactions in preparation (tissues) initiated by the polarized laser light are able to generate electromagnetic

radio frequencies (fluctuations). In this situation the preparations (tissues) of the pancreas and the spleen assume a role of specific molecular radio station where every type of molecules has its own frequency (vibration level) which can be amplified due to stochastic resonance produced by the wideband radiofrequency gas discharge laser.

Based on the obtained experimental data it is proposed to attempt to create a technology allowing development of resistance to toxic effects of Geptil/ hydrazine in animals. We suppose that the resistance to Geptil/ hydrazine and to many other toxic substances can be developed by influencing strategic metabolic vectors most vital of which are – functions of genetic apparatus at quantum level http://www.wavegenetics.jino-net.ru/ (http://www.wavegenetics.jino-net.ru/).

DISCUSSION OF THE EXPERIMENTAL RESULTS.

THEORETICAL MODELING.

The obtained data has wider implications than demonstration of capabilities of possible wave defensive antidote-effect and require theoretical consideration since they relate to strategic (quantum) mechanisms of genetic apparatus functioning of multicell bio-systems. At his stage we propose three formalized hypotheses of the wave processes while "reading or scanning" from bio-structures of donor genetic-metabolic wave information (data), remote addressed transmission of it, introduction of the information in to the bio-system-acceptor and metabolism management using that information.

1. Endogenous polarization-holographic processes in bio-systems

Wave informational scenarios – unfolding in the bio-system itself as well as during scanning by the laser beam at the initial phase – occur at photon level. Let's consider that level in more details. In our previous researches we presented two- and three-dimensional models of bio-holographic management of building spatial structure of multicell organisms during embryogenesis.At the initial view subject to relatively stationary conditions in bio-systems (final phases of morphogenesis), these models are plausible. However, in living organisms statics and dynamics are paradoxically intertwined.

An adult organism is relatively static spatially on macro scales and significantly changes on this plane only at the stages of deep aging. Along with this that statics are provided for by the internal space-time dynamics of metabolic processes on micro levels of bio-systems organization. This is a case due to the fact that metabolic processes are a mobile aggregate of bio-chemical-bio-physical space-time reformations of the organism's microstructures. Having considered the non stationary structures of bio-systems, more advanced model is proposed of endogenous informational polarized-holographic managing (governing) processes in multicell organisms, which are realized/effected at genome levels. The model reflects bio-holographic aspect of metabolism as a whole and therefore includes bio-morphogenesis as it specific case. The model utilizes appropriate physics-mathematical formalism for the polarized-holography yet extrapolates it onto probable endogenic analogical processes in the genetic apparatus of multicell organisms.

As a foundation for the model we used our experimental data where we employer special double-polarized He-Ne laser (= 632,8 nano метеr), having two orthogonal, linked optical modes, of which we have mentioned earlier. When such laser beam of this quantum generator interacts with the bio-tissues, with dynamic holographying on the oncoming beam (bunch) mode on, simultenious acts of recording and scanning of previously unknown information occur, the information about dynamic spinning-fluctuating processes on optical and atomic-molecular levels. Of particular interest can be data obtained about genetic structures and/or living cells. The entire informational structures of an organism, including DNA, RNA and proteins, are optically active, that is to say they are capable of spinning light polarization planes and they are dichroic – difference between absorption of right- and left hand polarized light. Modulations of polarization that correlate with structural-functional condition of any metabolite, present themselves as unique in their capacity storage of information about metabolism and its dynamics and along with it – that is a channel of intercellular photonic bio-symbolic contacts.

Such special processes, in their polarized-holographic version, are apparently inherent to genome workings as a bio-computer. This allows their modeling with use of the aforementioned laser. IT is capable of polarized-holographic recording, scanning, remote transmission and "injection" of wave commanding genetic-metabolic information from one bio-system to another. Besides, such laser performs a conversion of photons probing the bio-system in to wideband electromagnetic spectrum with frequencies from 2 to 0 according to mechanisms of localizing and

delocalizing of photons. During this stage, it is apparent, quantum non-local (teleportation) polarized connection across the whole spectrum of frequencies is retained (maintained, saved), including radio waves. Utilization of such laser as reading-transmitting photon-radio-wave system, imitating analogical wave-bio-computer-symbolic non local processes intercellular communications provided with opportunity to carry out remote wave transmission of primary /managing genetic-metabolic information from donor's bio-system to recipient's bio-systems (so called addressed transmission). In the light of this fact it seems rational and vital to attempt presentation of more advanced formalism of bio-symbolic photon-polarized-holographic processes in chromosome apparatus of higher bio-systems, all the more so that the radiowave equivalent of these processes possessed strikingly distinct morphogenetic potencies.

Let's record vector diffraction Kirchhoff's integral in paraxial approximation, which describes wave field, for instance photon field, formed by non-stationary bio-system's fragment. This sort of field may radiate from liquid crystal continuum of chromosome (LCCC) in vivo. Such emission may be expressed by the following:

For the purpose of simplification we will consider that non stationary LCCC is not a function of frequency of translucent light.

Both polarisational modes of coheren laser light are depolarized by geno-symbolic acoustic LCCC and are partially elliptically polarized. With this, they can interfere forming specl-structures, and their total intensity is transferred from one mode in to another by means of earlier postulate manner [Prangishvili, Gariaev et al, 2000 (b)].

Modified Johns' vector of every passed orthogonal polarized waves immediately behind the object may be presented in the form of partially coherent orthogonal components of elliptical polarization

In a bio-system in the composition of LCCC (with only one polarizational component) we use as a hypothetical a carrying wave, which passed, for instance, through infinitely narrow time shitter lock, possessing - figurative characteristic of time transmission (perhaps "gating"). Such a shutter lock completely depolarizes initially polarized wave. The resulting wave, passed behind the gating lock, has continuous spectrum in the whole diapason with evenly distributed spectral density, where as the modified vector of the carrying wave has a form of orthogonal basis of elliptical polarization:

Parameters of the total ellipse p and g are defined via ellipse's polarization component of every basis A and B, as in Ref [14]

Endogenous biological registration of the summary wave field (5) pertaining to LCCC as a basic element of DNA-wave bio-computer, assumes presence of polarization-sensitive environment in organisms, which is not selective of any particular spectrum across the whole range of active frequencies (similar to the non stationary fragment of biological object, for instance LCCC).

Due to the polarization characteristics of the inducing light in a light-sensitive registering environment of LCCC, photo-anisotropy and photogyrotropy are created (are produced). To describe the vector photo-response polarized-sensitive environment functions of isotropic , anisotropic - and gyrotropic reactions are introduced, which are constants for every frequency of the active spectrum. Using Johns' matrixes and rules of their construction for cases of partially polarized inducing irradiation, for the resulting Johns' matrix we get:

where - LCCC's hologram fragment size

- distance between the point on the hologram surface and the point of observation.

Than, successively substituting in (17) expressions for matrixes (10), (14) and (15), let's define formed by the hologram zero, imaginary and real depictions. It is only now that we can determine, what endogenous or/and exogenous wave for an organism is necessary to utilize as a reconstructing one in order to obtain in the imaginary style regeneration of the required fragment of the wave image of the shaping bio-system. In order to do so it is essential to determine their vectors and corresponding values of matrix. It appears that with precision of up to constant multiplier own vectors of matrix are in essence

It follows that regeneration is performed by a wave identical to the one used while recording by the carrying wave. So, as apparently, in bio-systems at LCCC level recording and reconstruction ensue ither simulteniusly or in accordance with the last equation, then the reconstructed imaginary depiction corresponds to the real one and it is not subject to

any distortions. The latter is principallycrucial for preservation of the wave images-vectors of morphogenesis, which compensate physiologic-bio-chemical and mechanical non-stationary stae of the bio-system as a whole and its LCCC in particular. Nonetheless, non-stability of strategic photonic-images of organism's structures will occur though in term of long lengths of time while aging and its pathological states, for instance in the case of cancerogenesis.

For a passed wave without diffraction the zero depiction has the following appearance:

An analysis of the last correlation shows that with precision up to its multiplier it contains complete reconstruction of space-time structure as well as polarization characteristics of the field of its non- stationary object wave that have passed through, for instance, LCCC. These photo or/and radio wave dynamic structures are, apparently, used by multicellular organisms for their own organization in their own space-time since these structures-images completely retain the true real calibration size without deformations imposed by non-stationary bio-systems and reproduce them in adequate sizes require for a developing or an adult organism. In accordance with the reconstructed wave gradients scanned polarized holograms occurs 4D organization of metabolic "fluxes"/"streams", cellular architectonics and morphogenetic motions during embryo genesis and also partial regeneration of bio-systems in case of them having been damaged. In other word what occurs is calibration of dynamic potential space-time of bio-system.

Under polarized-holographic-bio-control we imply endogenous or intentional (artificial) modification of recipient's cell structure and condition as a result of controlling holographic operations from donor's side. In our case the holographic signal, which was modulated by healthy samples (tissues, cells) of the donor, is transmitted and recorder onto donor's diseased (ill) cells in form of a hologram. Further, the process of control occurs as follows: in the beginning from the modified cells of the recipient by means of regenerative wave the holographic image of donor's healthy cells is scanned. It is reconstructed in terahertz wave diapason as a 3D image encompassing every recipient's cell together with its content.

Mainly, there exist two forms sources of regenerative wave. First form is endogenous. In this instance the processes flow due to innate reserves i.e. "internal" irradiation (emission) of the neighbouring cells. Second form appears to be exogenous reconstruction, when the sources are external emitters. Both forms of sources operate in the recipient's cells and act simultaneously and continuously, regenerating and complementing the same image of donor's healthy cells over and over again

According to gradients of intensity of donor's reconstructed cell images, as if it was a "drawing", growth and regeneration of recipient's damaged (ill) cells occur. The cells of the recipient "assume" a role of a photographic film (plate), in which a hologram of the healthy cells is recorded. Processes of the growth and regeneration, flowing analogically to processes of photo-tropism, require certain length of time. In the upshot the recipient's "sick" cells partially transfer to healthy ranking and partially destroyed. Products of such decomposition are diverted out of the recipient's organism.

That being so, during the replacement process of recipient's cells by healthy ones, analogous to those of donor's, occurs polarized-holographic-control which is in effect germination of the recipient's damaged cells biomass in to offered to them dynamic holographic form of healthy donor's cells. As a consequence, while controlling the process, the form and dynamic state (condition) of recipient's cells is gradually modified under the controlling signal – "pattern", obtained from the donor.

Needless to say that this form of "pattern" is much more complex than those seen in simple controlling systems. It sets spatial distributions terahertz signal on gradients of which the growth and formation of live cells in recipients occur. Therefore acts of growth and formation of recipient's cells happen in line with bio-chemical laws, which control their vital functions, and the "pattern" signal sets a program or guide of growth for young cells' structures and modulation of processes in them.

For more comprehensive description of internal working mechanisms of holographic circular-polarized informational-laser transformation in live organisms it is necessary to have a fair good understanding of basics of polarized-dynamic holographic theory and informational exchange between recipient's live healthy cells and all other cells included in organs and tissues of the ill organism. When dealing with these issues we utilize de

Transmission of modulated information from donor to recipient occurs by means of rectilinearly propagating longitudinal mutually entwined waves, carrying multilevel modulated information. For the short distance transmissions we can use a term of cellular nucleuses – optical poliaroids (polarizers?), where as for longer distances – a term of

quasi lenses (objectives) (here is a reference to some previous papers). Let's look in to description of such process proffered for purposes of registering color holograms without using lasers [Alexanderov, 1998]. When adapting it to a bio-system, we outline requisite conditions for realizing non-coherent polarized-holographic method of control (management). It ought to be noted that in such systems, having done microscopic studies of the systems, it was discovered long ago that they have microscopic polarisers, i.e. cells' nucleuses, as well as optically active protein substances, which spin the polarizational plane of emissions that are passed through them. These aspects are long knowng to researchers [Bischof, 1995], however until now this observed phenomenon was neither explained not utilized.

Holographic transmission function can be defined on the basis of Fourier transformation of equation (13). To remind, the wavelength of the probing signal is not included in this formula therefore that wave can be selected from wide range of waves light, electromagnetic and acoustic diapason. The created hologram encompasses whole information about spatial coordinates features of donor's bio-object being holographed, or about spatial distribution of all donor's points in relation to holograph registering planes of recipient.

So the resulting solution of the task is, generally speaking, analogical to traditional approach. At the same time the method put forward is fundamentally different from other known interferential methods and possesses certain advantages.

Firstly, instead of some wavelength with its monochromatic features and coherence we use dispercing spinning ability of optically active environments and spatial locally-distributed polarizational filtration. This is sufficient for recording polarized-dynamic hologram of donor (provided that there is motion in donor's cell in non-coherent wide spectrum irradiation of the recipient)

Secondly, this method allows uncovering the reasons of vibro-resistance when registering and reconstructing holograms without laser sources of light inside bio-systems in terahertz diapason of waves. It efficacy is determined by the value of polarized-optical spinning ability and thickness of the optically active environment . It is known that spinning ability of certain liquid crystals reaches 40000 degrees/mm, which when used in holographic information-laser convertor is sufficient for polarized-holographic transmission of information and accordingly holographic controlling of the biosystem's structures and processes.

Experimental works on wave interaction in living systems were actively been initiated in 1980s. In the beginning these included researched on interactions between cells [Kirkina et al, 1981; Molchanova, 1985], followed by researches on interaction of living organism [Burlakov et al., 1999]. These works were successfully furthered by A B Budagovski et al. [Budagovski, 1990; Budagovski, Evseeva, 1995; Budagovski et al, 1997; Budagovski et al, 2001]. It was shown that there is a communication and exchange of information of non-chemical (wave, coherent) nature. Such kind of interchange processes, occurring with participation of bio-regulatory signals and occurring without molecular and ion carriers of that information, were labled as processes of remote intercellular interaction (RII) [Budagovski, 2004]. However it seemed incredible that weak electromagnetic cellular signals can produce controlling influences against the background of strong electromagnetic interfering signals of natural and technogenic origins. Nonetheless it appears that with coherent reception light and other electromagnetic non-coherent noises with approximation are zeroed where as those weak coherent and deterministic signals can accumulate [Tertishni et al., 1997, 1998, 2000]

In the last years these works have received the deserved advancements in the Institute of management issues. In particular, utilization of polarized-dynamic holography, which allows formation of little movable polarizational rings. For urposed of transmitting undistorted image of every donor's point to the remote recipient's zone, a sensory quasi objective (lenses) were created, and on its basis a holographic device designated for experimental testing of holographic controlling possibilities.

Publications in foreign journals

1. Garyaev P.P., Vikhert A.M., Zakharkina G.A., Alexandrova L.G., 1977, Endogenous reserpine-like substances in the myocardium and other tissues of man and animals., USA-USSR First Joint Symposium on Sudden Death. Yalta, USSR, October 3-5, 1977, pp.323-338.

2. Gariaev P.P., Chudin V.I., Komissarov G.G., Berezin A.A., Vasiliev A.A., 1991, Hologrphic Associative Memory of Biological Systems, Proceedings SPIE - The International Society for Optical Engineering. Optical Memory and Neural Networks. v.1621, p.280- 291. USA.

3. Gariaev P.P., 1994, DNA as source of new kind of God "knowledge", Act and Facts/Impact series, N12, pp,7-11.

4. Maslov M.U., Gariaev P.P., 1994, Fractal Presentation of Natural Language Texts and Genetic Code. 2nd International Conference on Quantitative Linguistics "QUALICO-94". September 20-24. (1994). Moscow, Lomonosov Moscow State University, Philological Faculty. pp.193-194.

5. Gariaev P.P., Vasiliev A.A., Berezin A.A., 1994, Holographic associative memory and information transmission by solitary waves in biological systems. SPIE - The International Society for Optical Engineering. CIS Selected Papers. Coherent Measuring and Data Processing Methods and Devices. v.1978, pp.249-259.

6. S.A. Reshetnyak, V.A. Shcheglov, V.I. Blagodatskikh, P.P. Gariaev, and M.Yu.Maslov, 1996, Mechanism of interaction of electromagnetic radiation with a biosystem, Laser Physics, v.6, N2, pp.621-653.

7. Berezin A.A., Gariaev P.P., Gorelik V.S., Reshetniak S.A., Shcheglov V.A., 1996, Is it possible to create laser on information biomacromolecules? Laser Physics, v.6, N6, pp.1211-1213.

8. A.A. Berezin, P.P. Gariaev, S.A. Reshetniak, K.V. Shaitan, V.A. Shcheglov., 1996, To the problem of possible development at a biolaser working on Frolich modes. Prepr. P.N.Lebedev Physical Institute №49, 12 c .

9. Gariaev P.P., Tertishny G.G., Kampf U., Muchamedjarov F., Leonova E.A., 1999, Fractal structure in DNA code and human language: Towards a semiotics of biogenic unformation. 7 th International congress of the international association for semiotic studies (IASS/AIS). TU Dresden, October 3-6, 1999. p. 161.

10. Peter P. Gariaev, Boris I. Birshtein, Alexander M. Iarochenko, Peter J. Marcer, George G. Tertishny, Katherine A. Leonova, Uwe Kaempf ., 2001, The DNA-wave biocomputer. “CASYS” – International Journal of Computing Anticipatory Systems (ed. D.M.Dubois), Liege, Belgium, v.10, pp.290-310.

11. Peter P. Gariaev, George G. Tertishny, Katherine A. Leonova., 2002, The Wave, Probabilistic and Linguistic Representations of Cancer and HIV. http://www.emergentmind.org/gariaevI2.htm (http://www.emergentmind.org/gariaevI2.htm)

12. P.P.Gariaev, G.G.Tertishny, A.M. Iarochenko, V.V.Maximenko, E.A.Leonova, 2002, The spectroscopy of biophotons in non-local genetic regulation. Journal of Non-Locality and Remote Mental Interactions Vol.I Nr. 3

13. Peter P. Gariaev, Boris I. Birshtein, Alexander M. Iarochenko, George G. Tertishny, Katherine A. Leonova, Uwe Kaempf, Peter J. Marcer, 2002, Fractal structure in DNA code and human language: towards a semiotic of biogenetic information. International Journal of Computing Anticipatory Systems, volume 13, 2002, pp. 255-273, ISSN 1373-5411 ISBN 2-9600262-7-6.

14. P. Gariaev, 2003, Der wellengenetische Code. Tattva Viveka, №20, pp.68-73. (in German)

15. P. Gariaev, W. Wagner, auch auszugsweise, 2005, Der Phantomeffekt. Spiegelung zwischen Geist und Materie. (Teil 1). Tattva Viveka, №26, pp.54-65. (in German)

16. Peter Gariaev, Mark J. Friedman, Ekaterina A. Leonova-Gariaeva. Principles of Linguistic-Wave Genetics. http://dnadecipher.com/index.php/ddj/article/viewFile/2/2 (http://dnadecipher.com/index.php/ddj/article/viewFile/2/2) ; http://dnadecipher.com/index.php/ddj/article/view/2 (http://dnadecipher.com/index.php/ddj/article/view/2) DNA Decipher Journal, Vol 1, No 1 (2011).

17. Peter Gariaev, Peter J. Marcer, Katherine A. Leonova-Gariaeva, Uwe Kaempf, Valeriy D. Artjukh. http://dnadecipher.com/index.php/ddj/article/view/2 (http://dnadecipher.com/index.php/ddj/article/view/2) DNA as Basis for Quantum Biocomputer PDF; http://dnadecipher.com/index.php/ddj/article/viewFile/3/13 (http://dnadecipher.com/index.php/ddj/article/viewFile/3/13) . DNA Decipher Journal, Vol 1, No 1 (2011)

18. Peter Gariaev, Matti Pitkanen, Model for the Findings about Hologram Generating Properties of DNA PDF ; http://dnadecipher.com/index.php/ddj/article/view/2 (http://dnadecipher.com/index.php/ddj/article/view/2) . http://dnadecipher.com/index.php/ddj/article/view/2 (http://dnadecipher.com/index.php/ddj/article/view/2) . DNA

Decipher Journal, Vol 1, No 1 (2011)

2012 - Mirrored Material is Fair Use for Educational Purposes Only

Represented by Irene Caesar, Ph.D., President, Wave Genome Research and PublishingPhone in the US: 1-347-352-0044email: irenecaesar@wavegenome.com (mailto:irenecaesar@wavegenome.com)
Skype: Irene Caesar
The history of wave genetics has been developing for more than 80 years. In 20s and 30s of the last century Russian scientists, A G Gurwitsch and A A Lubishev postulated that not only does genetical apparatus of living organisms on the Earth operate at material, physical level but also at certain waves/fields level and is able to transfer genetic data/information via electromagnetic (EM) and acoustic waves. The science has recently taken a major leap in its advancements of the theory and reproducible experiments and thus the theory of wave genome has been put forward.

One of the first attempts to rationalize and expound the wave genome theory in Russia had been made by P P Gariaev and A A Berezin from the department of Theoretical Matters of Russian Academy of Science, and also with participation of A A Vaciliev, a fellow member of Institute of Physics of Russian Academy of Science. As a theoretical foundation, principles of coherent physical radiations, holography and solitons, theory of physical vacuum, fractal representation of DNA and human verbal expression/speech have been employed to describe and substantiate the empirical results obtained through many experiments conducted.

The quintessence of the wave genome theory may be represented as following:

genome of the highest organisms is considered to be a bio-computer which forms the space-time grid framework of a bio-systems.

In that bio-system, as the carriers of a field epi-gene-matrix - wave fronts are being used, which are assigned by gene-holograms and so-called solitons on DNA – distinct type of acoustic and electromagnetic fields, produced by biogenetic apparatus of the organism/bio-system under consideration and being a medium of strategic regulatory data/information exchange between cells, tissues and organs of the bio-system.

It is also vital to note that the holographic grids/frameworks, which are also the elements of fluctuating structures of solitons, are, in fact, discrete simplest cases of code-originated information, anchored in chromosome continuum of an organism.

At present the dominant viewpoint in respect of genetics and molecular biology is that :
1. the genetic apparatus functions as a purely material structure
2. all the functions of genetic control of an organism occupy approximately 2 % of DNA of a bio-system and fulfill designated functions such as replication of RNA and proteins, so called coding DNA of an organism. The other 98% is considered to be “junk DNA” not carrying out any genetical functions, which may represent a graveyard of virus DNA.

The biologists and geneticists use the language of analogies and metaphors to explain how genetic apparatus operates. The genetic apparatus consisting of 46 chromosomes is viewed as a library consisting of 46 volumes or books. Each book (a chromosome), contains a text (instructions of how to build an organism) which consists of sentences (DNA) consisting of words (genes). And each word (a gene) consists of 4 letters (certain “chemical letters”), i.e. the “genetic alphabet” consists of only 4 “letters”. The material realizations of the DNA molecules are famous double helixes, consisting of segments which are genes. In essence, genetic apparatus operates as follows. The texts, written in the “DNA language”, are first translated by the organism into the “RNA language” and then into the “Protein language”. And proteins are the stuff that we are mostly made of (not counting water). Proteins perform two principle functions in the organism: they metabolize substances that we eat and participate in the morphogenesis, i.e. development of the spatial-temporal organization of an organism.

It is to be pointed out that the primary focus of the wave genome theory is on the remaining aforementioned 98% of chromosomes as being the key “intellectual” structure of all cells of an organism including the brain. It is those chromosomes that operate on the wave, on the “ideal” (fine-field) level.

It the ideal component, that may be called super-gene-continuum, is a strategic vital figure/formation that ensures development and life of humans, animals, plants and also their programmable natural dying. Along with that, it is

important to conceive that there is no sharp and insurmountable distinction between genes and super-genes. Both these levels of encoding constitute material (physical) matrixes, however genes supply material replications in the form of RNA and proteins, whereas super-genes transform endo- and exogenous fields, forming from them super-gene-signal wave structures. Furthermore, the genes may be components of holographic grids/frameworks of super-genes and supervise their field activity

Particular attention in the wave genome theory should be devoted to substantiating of the unity of fractal (replicating itself on varying scales) sequence structures of DNA and of a human speech. In 1990 Jeffrey Delrow discovered four “letters” of the genetic alphabet (Adenine, Guanine, Cytosine and Thymine) in DNA “texts” form fractal structures.

Later on, a discovery of similar fractal structures in a human speech, not limited to multy-letter alphabets of Russian and English texts and including a sequence of words from those texts came as a promising surprise for both genetics and linguists. Nonetheless this is in accordance with a branch in semiotics named “Genetical Linguistics”, which studies incomprehensible and unexplainable precise application of laws of Formal Genetics to formation of interlingual and intro-lingual words-hybrids.

A group of scientists headed by P P Gariaev and M U Maslov, developed a theory of so called fractal representation of natural (human) and genetical languages. Within the confines of this theory it is said that the quasi-speech of DNA possesses potentially inexhaustible “supply of words” and, moreover, what had been a sentence on the scales of DNA–“texts” “phrases” or a “sentence” becomes/turns into a word or a letter on the other scale. Genetical apparatus can be viewed as the triunity of its structure-functional organization consisting of holographic, soliton and fractal structures.

This theory allows a refined quantitative comparison of symbolic structure of any texts including genetical. Thereby a possibility has been wide open to approach a deciphering of a lexicon of one’s own gene-code, and accordingly, more accurate composition of algorithms of addressing a genome of a human with an aim of potentially any type of programming of one’s vital activity such as treatment, increasing one’s life expectancy and so on and so forth.

Empirical tests of wave genetic theory in the light of “speech” characteristics of DNA demonstrate strategically correct stance and direction of the research.

Phenomenal experimental findings may be briefly presented here. For the first time in history of the science we successfully managed to obtain experimental evidences of abilities of the genetic data/information to function on more than simply one material level, for instance, in the wave/field (electromagnetic) form

Over decades, having been conducting extensive theoretical researches, we successfully managed to formulate theoretical biology and physics-mathematical description, explanation and substantiation of fundamental principles of genetic apparatus functioning on the wave/field level.

These principles allowed us to design and create a quantum bio-computer essential element of which is a specially tuned laser beam with certain wavelength and frequency characteristics.

The quantum bio-computer can perform:
i. Scanning and reading the wave/field equivalent of the genetical-metabolic data from the (stem) cells, tissues, organs of a donor’s bio-system employing photon of the laser beam;
ii. Conversion of the photons obtained into wide-wavelength-band preserving the scanned and read data;
iii. Precise and pointed introduction of such scanned and converted principal/dominant data radio waves into the organism-recipient, located at a certain distance from the donor’s bio-system (from few centimeters to 20 km);
iv. Strategic control/management of metabolism and postembryonic morphogenesis of the organism-recipient according to two vectors/modes:
a. “Do as/what I do” – that is holographic principle, and
b. Wave/field transmission of a signal that contains commands activating required programs in the stem cells of, for instance, rats; programs that direct/guide the stem cells development on their way of, as in our experiments, building pancreas gland in rats. These vectors/modes are intended to be used on humans.

At present we are able to program/manage/encode stem cells of various types by means of a quantum bio-computer. The quantum bio-computer initiates wave/field-based commands, given to cells and tissues of the donor/recipient, and accordingly the stem cells exposed to the waves/field will be prompted to guided cytodifferentiation leading to emergence and development of planned/projected new organs and tissues. This will bring us closer to the fact of substantially increased life expectancy.

Significant achievements in applying WaveGenetics have been made thus far in regeneration of pancreas in rats previously destroyed by chemical substance called alloxan

Three series of experiments with identical protocol were conducted by the groups of P. Gariaev in 2000 in Moscow Russia, in 2001 in Toronto, Canada, and in 2005 in Nizhni Novgorod, Russia. These are more advanced experiments based on the principles and technology of WaveGenetics. The goal of the experiments was to test new technology for regenerating damaged pancreas. Pancreas is an endocrine gland which has several vital functions, the major one being production of insulin, a hormone responsible for sugar metabolism.

A control group of rats was injected a lethal doses of a poison called alloxan which destroys pancreas. As a result, all the rats in the control group died from diabetes in 3-4 days. Then the same lethal dose of alloxan was injected to another group of rats. And when the rats reached the critical condition, they were exposed to light images/waves coming from a quantum bio-computer. Those light images/waves were created beforehand when the bio-computer read information from the pancreas surgically removed from healthy newborn rats of the same species as those used in the alloxan experiments.

One can explain the results of the experiment using the following analogy. The pancreas gland contains DNA-movies with information about healthy condition of the pancreas in its genetic apparatus. And this video morphogenic information programmed the stem cells of sick rats to regenerate their pancreas gland. Combined statistics for all 3 series of experiments is as follows. Altogether, around 90% of all the rats had their pancreas gland restored and their health recovered.

In some of the experiments the quantum bio-computer was modified to allow successful transmission of the healing information to sick rats at the distance of 20 kilometers. Note, that no known physical fields have the capability to transmit such extremely week signals with such unbelievably powerful results.

Furthermore, the WaveGenetics research and application has significant prospects of solving issues as regards ageing process in humans and increase of life expectancy. This view is grounded on the experiments we have been conducting with rats.

It is Opportunities of technologies of WaveGenetics are not limited to what has been outlined above. The development and application of the technology is far-reaching and is substantiated by experimental data obtained in numerous tests, experiments, observations.

In addition, it ought to be mentioned, such a technology – a quantum bio-computer, is capable of:
- **treating oncological diseases on fundamentally distinct basis – without use of any chemical substances,**
- **eliminating pathogenic viruses and bacteria and agricultural vermin parasites – also without used of any chemical substances.**

It may seem as a technology of another age, from distant future. However, a discovery of fundamental properties of living organisms is occurring today and it is our task to research and explain the phenomena and bring it to the service of humanity.

Peter Gariaev, Ph.D.,
Director of Wave Genetics Institute,
Russia, Moscow.

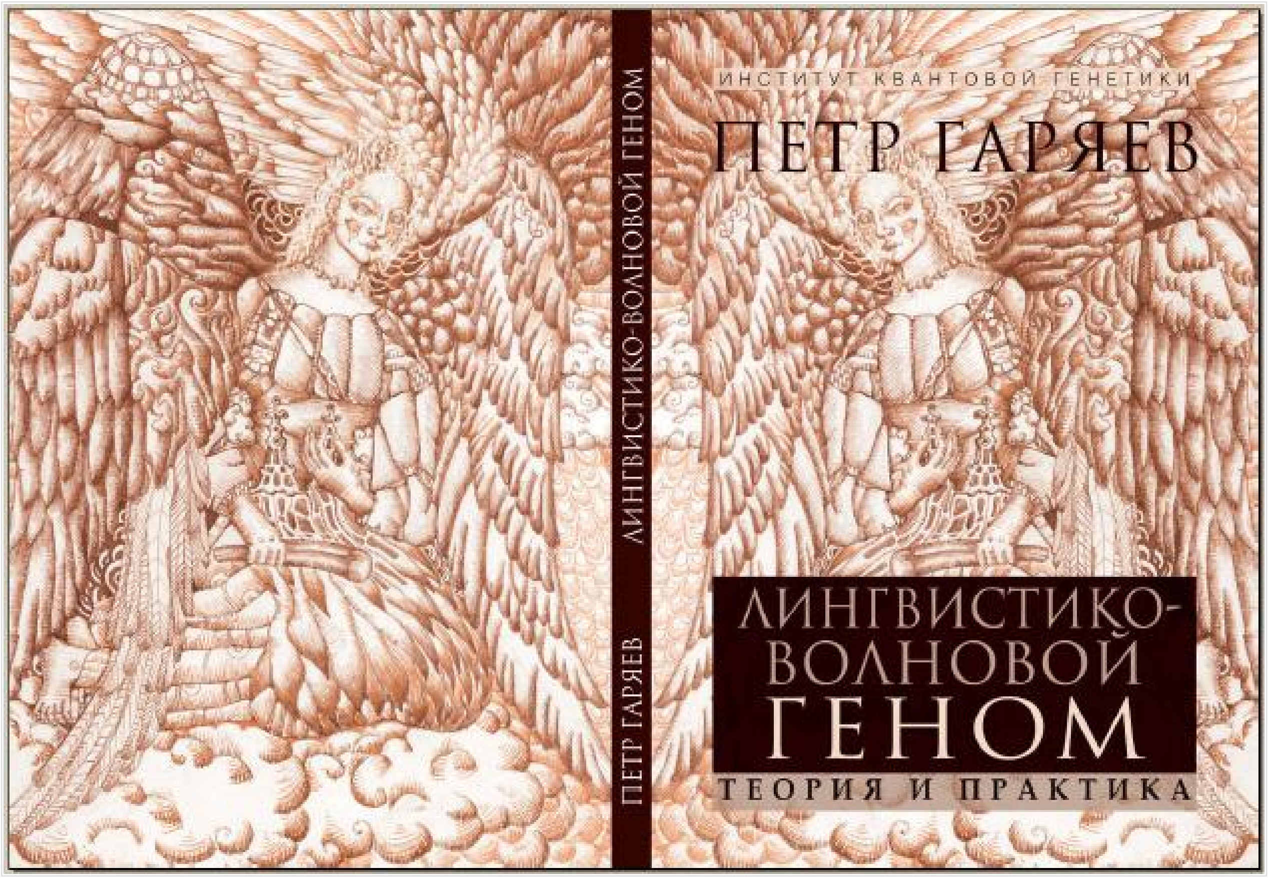
ИНСТИТУТ КВАНТОВОЙ ГЕНЕТИКИ
ПЕТР ГАРЯЕВ
ЛИНГВИСТИКО-ВОЛНОВОЙ ГЕНОМ
ТЕОРИЯ И ПРАКТИКА
ЛИНГВИСТИКО-ВОЛНОВОЙ ГЕНОМ
ПЕТР ГАРЯЕВ

Hi Iona...I have been following Gariave's work since I became aware of it. I have about 25 years of experience with DNA in the lab. A serious problem with his approach is that he still believes in the "junk" DNA notion which has run its course..now there is excellent evidence that ~80% of the genome contains important information. In addition to coding for proteins, there is a plethora of RNAs, micro, ribosomal, long repeat, siRNA etc that play a role in developemt, disease and so forth. Another weakness is that all his experiments were done with purified DNA. In the cell, DNA is part of a protein-DNA complex, which is very tightly packed at very high concentrations ..chromatin has totally different biochemical and biophysical properties than the dilute solutions of DNA that he works with..(The cell is so concentrated with proteins and nucleic acids, that its closer to a liquid crystal than a bag of solutes)

CREATE A FREE WEBSITE

POWERED BY

Category Mind Control

NOV **14** 2017

Invited presentation by Magnus Olsson at the 2017 First Annual Unity and Hope Conference

Invited presentation by Magnus Olsson

“Invited presentation by Magnus Olsson, at the 2017 First Annual Unity and Hope Conference” This event was for targeted individuals and those concerned about the growing crimes of electronic harassment.

(https://nanobrainimplant.com/240_f_116659017_ryiguzdnplfxcwaattixa25shzy5nkt4/)

The conference was held from October 20-22 at the Mass Audubon Blue Hills Trailside Museum: 1904 Canton Ave., Milton, MA 02186, USA. This presentation was co-produced by Mårten Hernebring. The speaker, Magnus Olsson, can be reached at bionicgate@live.se Event Description from the Conference Web Site: “Our goal is to bring together as many support groups, media shows, activism groups, and organizations of targeted individuals, so we can work together and learn from each other and strategize on solutions to bring about change and end the suffering of hundreds of thousands of victims nationwide.

The number of people experiencing electronic harassment and gang-stalking is growing exponentially daily. Our hope is to come together, to build, empower, and educate the community on technology, resources, and support, and as a unified front attempt to educate the public. As a result of this conference, we will be able to strategically fight for freedom and justice for the victims of targeted crimes.

The goal of this conference is to unify all the groups worldwide and provide a knowledge and understanding of the program and the technology. We also strongly encourage targeted individuals to bring friends and family for support and to educate the ones around them on what invisible crimes are being committed against them."

By mindcontroleurope • Posted in Man vs Machine, military, Mind Control • Tagged artifical-intelligence, electronic harassment, gang-stalking, Hive Mind, hybrid brain, hybrid mind, magnus olsson, mind control, targeted individual, Targeted Individuals, Unity and Hope Conference
DEC **11** 2016
IMAGE

Now declassified & available! Russian Quantum Leap technology enhances RNA/DNA, cures diseases, stops Bio hacking

Now declassified & available online! Russian Quantum Leap enhances RNA, DNA & health, cures diseases (e.g. diabetes, cancer 2), stops TI targeting.

December 6, 2016

Now declassified & available online! Russian Quantum Leap technology enhances RNA, DNA & health, cures diseases (e.g. diabetes, cancer 2), stops TI targeting.

By Alfred Lambremont Webre

WATCH QUANTUM LEAP PANEL INTERVIEW

Watch Russian Television RT Interview with Magnus Olsson [Seen by 14 million persons]

Mind Control – Remote Neural Monitoring: Daniel Estulin and Magnus Olsson on Russia Today

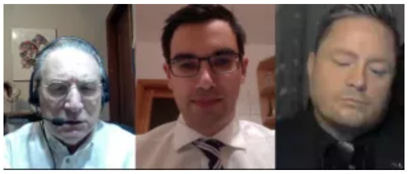 (https://i2.wp.com/newsinsideout.com/wp-content/uploads/2016/12/Alfred-Webre-Morton-Hernebring-Magnus-Ollson-4.png?ssl=1)

alfred-webre-mårten-hernebring-magnus-olsson

You Tube: https://www.youtube.com/watch?v=o9bd-B2dqCM (https://www.youtube.com/watch?v=o9bd-B2dqCM)

VANCOUVER, BC – Newly arrived from Russian Quantum Leap by the American Reaserch and Development Company Wave Genos and its Founder and Sole Owner Irene Caesar Ph.D. Magnus Olsson and Madlen Namro, are joined by their colleagues Mårten Hernebring to launch the public version of Quantum Leap technology of Wave Genome Irene Caesar Ph.D, a declassified technology used by Russian Special Forces to enhance DNA, RNA, personal health profile and performance, eye and vision health, as well as provide personal protection against Transhumanist agenda weaponry including remote directed energy weapon [DEW] and psychetronic attacks, such as "V2K (Voice To Skull) and cancer guns, subliminal advertising and psychic attacks".

MindTechEnterprises, authorized distributors outside Russia for the ULTIMATE TESLA GENERATOR, created and produced by the American company Wave Genome and its President, Sole Owner and Founder Irene Caesar Ph D,"

"The information-wave technology of the Laser Biohologhraphy has been perfected by decades of research experience and clinical trials; it has been patented, licensed and certified by the Russian Ministry of Health; and, at the present moment, it is widely commercialized all over Russia. The information-wave technology of Laser Bioholography is based upon the application of Quantum Physics to Genetics, a new theory of Genetics, which is called Structuralist Genetics (© Irene Caesar, Ph.D., 2010), and the application of the stabilized He-Ne lasers with internal mirrors."

MindTech-Enterprises Website

(https://nanobrainimplant.com/2016/12/11/now-available-russian-quantum-leap-technology-enhances-rnadna-cures-diseases-stops-bio-hacking/mind-tech-logo-01a/)

Full details of this first-ever public roll-out of Russian classified protective and health-status enhancing technology are set out in the above Panel interview and at Quantum Leap's distributors website at:

http://www.mindtech-enterprises.com/ (http://www.mindtech-enterprises.com/)

Quantum Leap Scalar Technology online

Quantum Leap technology, now available online through mini-Tesla Psi generators from the www.MindTech-Enterprises.com (http://www.mindtechenterprises.com/) website and through smart-phone apps, reportedly enhances the user's RNA, DNA & health profile, prevents and cures diseases such as diabetes, Stage 2 cancer, and stops Transhumanist Agenda weapons such as Voice to Skull and electromagnetic remote targeting.

The Quantum Leap technology's Mini-Tesla Generators come with Digital Pharmacy and a subscription service.

Quantum Leap technology, in applications developed in conjunction with Russian scientist Irene Caesar PhD. According to Dr. Caesar, the technology results in "Rejuvenation of all the Physiological systems, including endocrine system, digestive system, cardiovascular system, hearing / eye sight, reproductive system, intellectual abilities, improvement of memory."

Some features of the Quantum Leap Ultimate Generator

According to Wave Genome Irene Caesar Ph.D, *"The device is easy to wear in the form of a pendant cased in pearl, completely unobtrusive, and people can't tell exactly what it is because it is cased in pearl and does not look too unusual or like a "medical device", etc. The device is indestructible. The device is waterproof for basic small splashes, but should not be immersed in water.*

(https://nanobrainimplant.com/2012/07/11/scientists-humans-and-machines-will-merge-in-future/imagesca38vl83-3/)

The team uses a high resolution picture of yourself from early childhood (around 4-6 years of age) to develop sophisticated holography matrices from the picture using laser technologies and quantum technologies including scalar waves. They are able to recreate your energy field from when you were a child in a personal digital wave matrix and then use the device to continuously feed your current energy field with information from your childhood energy field. This helps to restore your current energy field to healthier conditions. This newer model has logarithmically more powerful protection features compared to the older versions of this device.

This new model still comes with all the health features of the older model including the E-Pharmacy, and ongoing health monitoring by professionals. The older more generic version of the device has been worn by members of parliament and Special Forces in Russia as well as the Russian Olympic Team. Over 5000 scientists/researchers from Russia have collaborated to work on this special project developing this device using the most advanced science in the world.

Some benefits of the Ultimate Generator:

The advanced technologies in the device can neutralize psychotronic/electronic attacks. The device is able to eliminate V2K. You can sign up for a monthly membership service to have the expert professional staff at the Moscow clinic in Russia monitor your health on an ongoing basis. If anything goes wrong with your health they are able to correct deviations from normal even before you notice symptoms of any actual medical condition. If medical conditions exist, many of them can be completely healed. You can receive medications remotely administered through the device. They get filtered through your own energy field and are guaranteed to never cause any adverse side effects.

The device helps to boost your immune system to diminish susceptibility to any illness. The device can help neutralize the negative effects of cosmic radiation when you fly in an airplane long distances. It can also reduce jet lag as well as keep you healthy while traveling. The device can help boost your energy, can help you sleep better, and can help you recover from any stressful events, sleep deprivation, etc., much faster. It can help you think more clearly and improve overall cognitive function. (Irene Caesar Ph.D.)

How purchasing works:

You can buy online at the website. Please see

http://www.mindtech-enterprises.com/ (http://www.mindtech-enterprises.com/)

Choose which version of the product you want to purchase. We recommend the Ultimate Generator as it is the most powerful version. Payments are processed using Paypal and other methods. You will be contacted to send a high resolution childhood picture of yourself. You will receive the device in the mail within 6 weeks. It takes this long to produce this device and to customize it for each individual. But don't worry, it will be worth the wait!

Tesla Glasses

IRENE CAESAR, PH.D.: ON TECHNOLOGY AND THEORY OF TESTLA GLASSES

The Panel also discusses Tesla glasses that have been developed and successfully used by Russian Security Forces for prevention and correction of:

- Stress and its consequences;
- Deterioration of the physical and mental performance;
- Reduction of immunity and the bodily defenses;
- Eye diseases;
- Diseases of the internal organs and physiological systems

(https://i1.wp.com/newsinsideout.com/wp-content/uploads/2016/12/1.TESLA_.GLASSES3.jpg?ssl=1)

Tesla Glasses

According to Wave Genome ***Tesla Glasses*** *use the method of Polarized Holography. Polarized Holography is the application of Quantum Physics for the purpose of rejuvenating all organs of the person through the Eye Crystal (Eye Lens and Iris). It is proven to provide significant physical and mental improvement, including a significant improvement in brain functions: decision speed, attention and memory.*

The method consists in the control of the brain via triggering the holographic signal in the eye, and its modulation. This control overwrites any verbal and mental imperatives, is non-local (can be used for instantaneous transmission of information at infinitely large distances), instantaneous (exceeds speed of light), and simultaneously reaching out towards every cell in the body.

The method is based upon Quantum Physics, specifically, upon the theory and technology of Polarized Holography, which uses the modulation of the holographic signal according to the Kozyrev Mirror principle.

The method is based upon the special ability of an eye to convert any linear signal into the holographic signal. The process of this conversion consists in the emergence of the scalar wave diffraction grating, polarization and refraction toward the zero center of the wave crystal (torus). Since the universe is holographic, we can record and transmit information only via the holographic signal, and through the zero center of the wave crystal. The zero center of every toroidal wave crystal coincides from the zero center of every chromosome on the cellular level, and every atom and subatomic particle on the subatomic level to the zero center of our galaxy.

Any healing, rejuvenation, and mental, psychic, and physical enhancement are based upon the ability to produce, control and enhance the holographic signal. The holographic signal instantaneously reaches the zero center of every wave crystal in our body, from the subatomic and atomic to the molecular levels – from the zero center of our skull and every bone to the zero center of every metacentric chromosome. Our eyes, skull, bones and chromosomes are centered and focused similar to our eye crystal according to the laws of the geometrical optics.

Hence, the eye is a very unique receiver, transmitter, and producer of the holographic signal. And that is precisely why vision is a trigger of brain activity. Via correct triggering the holographic signal in the zero center of the eye crystal, we can immediately reach the zero center of the toroidal wave crystal of our brain, thus delivering controlling information instantaneously into the zero center of the wave crystal of every metacentric chromosome in cell division, and into the zero center of the wave crystal of every bone, so that stem cells are programmed in a correct way in the bone marrow.

MINDTECH
ENTERPRISES

(https://nanobrainimplant.com/2016/12/11/now-available-russian-quantum-leap-technology-enhances-rnadna-cures-diseases-stops-bio-hacking/mind-tech-logo-01a/)

Correct triggering consists in the modulation of the holographic signal in the eye. The modulation of the holographic signal in ***Tesla Glasses*** *consists in (1) modulation by the Schumann holographic signal; (2) modulation by human Brain holographic signal; (3) modulation by the holographic signal of a healthy organ and physiological system, recorded upon the chip in the glasses. It is combined with the modulation by the narrowband light-emitting diode sources of five colors. Research had shown that five colors, used by the glasses, benefit organs and physiological systems. The signal for every color is produced from the holographic signal of healthy organs and physiological systems, and, then, it is matched with the color. The signal of the color is not simply the color frequency.*

The carrier holographic signal is the individual holographic signal of the user – the unique [indestructible and uncreated] non-local wave matrix of the user. This is the primary holographic signal. The modulation holographic signals are the holographic enhancers, which help the unique [indestructible and uncreated] non-local wave matrix of the user get centered and focused in this dimension (on this planet). These are secondary holographic signals.

The advantage of the glasses over other Quantum Leap Polarized Holography devices is that glasses use an eye as a filter for filtering out all harmful linear signals, which unseal wave crystals. For example, microwave radiation shortens brain waves via "unsealing" the involuted wave crystals, thus, literally lowering the intellectual potential of people.

(https://nanobrainimplant.com/2016/12/06/now-declassified-available-online-russian-quantum-leap-technology/240_f_68304591_ovk9v6qieaoepraueh02gzahmcl49ymn/)

Thus, the effect of the ***Tesla Glasses*** *is based not simply upon the superficial stimulation of innervation. The stimulation of innervation, tone of the eye muscles, blood circulation, regeneration, and enhancement of all other biochemical and bioelectrical processes in the body are based upon the structural efficacy of Polarized Holography, which centers and focuses chromosomes, and prevents the transformation of metacentric chromosomes into acrocentric chromosomes.*

The modulation via secondary holographic signals (see above) allows for the emergence of the most coherent and sophisticated scalar wave diffraction grating that produces the more clear zero focus within the wave crystal (of a wave torus on every level from the atomic, subatomic to cellular and molecular levels).

The modulation occurs through the designated areas of Iris, each one corresponding to a specific organ in the body. As a result, the user achieves the stimulation of the projection zones in the Iris, each one being connected to a specific organ in the body.

In addition, ***Tesla Glasses*** *use the same effect as the binaural therapy. Binaural therapy uses the stereo effect, when signal in one ear gets into the brain with delay in relation to signal in the other ear. This produces the emergence of the complex scalar wave diffraction grating in the skull, based upon Kozyrev Mirror principle (polarization, refraction and emergence of the zero center of the wave crystal). Similar to this,* ***Tesla Glasses*** *use the stereo effect of color signal, when color signal in one eye gets into the brain with delay in relation to signal in the other eye.*

Product by:

WAVE GENOME

(https://nanobrainimplant.com/2016/12/11/now-available-russian-quantum-leap-technology-enhances-rnadna-cures-diseases-stops-bio-hacking/wave-genome/)

Ophthalmologists have found that all healthy people have rhythmic alternation visual perception. Man sees in turn by the right eye, and, then, by the left eye at regular intervals.

These intervals last for 2-3 seconds. The research was conducted using modern 3D technologies. And if one eye is covered by a blue filter, and the other eyes is covered by the red filter, the healthy person sees both colors simultaneously and separately. The left eye sees 20% of the image, and the right eye sees 80% of the image.

With age, these rhythms get broken. But the restoration of the "biological clock" is possible using Polarized Holography, modulated by color signals, with the frequency of the alpha rhythm of the brain of a healthy person. And, as the positive effect, we can achieve a significant rejuvenation of the body.© Irene Caesar 2015

The text is quoted from http://www.wavegenom.com (http://www.wavegenom.com)

More information on Tesla Glasses is available, and Tesla Glasses can be purchased at:

http://www.mindtech-enterprises.com/ (http://www.mindtech-enterprises.com/)

References

Quantum Leap, available from:

Wave Genome

MindTechEnterprises

http://www.mindtech-enterprises.com/ (http://www.mindtech-enterprises.com/)

Contact:

Email:

mindtechtrade@gmail.com

Send A Message:

http://www.mindtech-enterprises.com/ (http://www.mindtech-enterprises.com/)

Alfred Lambremont Webre Contact/News:

Related

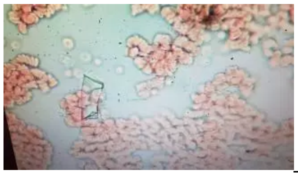

(https://newsinsideout.com/2016/09/part-magnus-olsson-reveals-artificial-crystals-blood-spread-via-smart-dust-transhumanist-remote-neural-weapon-control/)

PART I: Magnus Olsson reveals artificial crystals in his blood, spread via smart dust, are a Transhumanist remote neural weapon for control (https://newsinsideout.com/2016/09/part-magnus-olsson-reveals-artificial-crystals-blood-spread-via-smart-dust-transhumanist-remote-neural-weapon-control/)

PART I: Magnus Olsson reveals artificial crystals in his blood, spread via smart dust, are a Transhumanist remote neural weapon for control Transhumanist think tanks foreshadowed revelation in 2011 VOLVO TV ad – "Magnus Olsson: it gets in your blood" NewsInsideOut.com By Alfred Lambremont Webre WATCH EXOPOLITICSTV INTERVIEW VANCOUVER, BC…

September 20, 2016

In "AI Artificial Intelligence"

EUCACH Director: Dr. Rauni Kilde radiated for four days with DEW.

EUCACH Director: Dr. Rauni Kilde radiated for four days with DEW. Remotely assassinated by NSA? By Alfred Lambremont Webre NewsInsideOut.com WATCH ON YOU TUBE VANCOUVER, BC – Magnus Olsson, director of the European Coalition against Covert Harassment (EUCACH.ORG), revealed today in a NewsInsideOut.com interview that EUCACH Board of Directors member,…

MINDTECH
ENTERPRISES

(https://nanobrainimplant.com/2016/12/11/now-available-russian-quantum-leap-technology-enhances-rnadna-cures-diseases-stops-bio-hacking/mind-tech-logo-01a/)

By mindcontroleurope • Posted in Man vs Machine, military, Mind Control, News, Uncategorized • Tagged artifical-intelligence, biohacking, brain, disease, dna, electronics, health, interfaces, magnus olsson, military, military / medical classified research, neurohacking, quantum leap, rna, russian quantum leap, scalar, targeted individual, ti, Transhumanism, wave genome
SEP **11** 2016

THE Reality of Magnus Olsson’s blood – doctor’s report

THE Reality of Magnus Olsson’s blood – doctor’s report

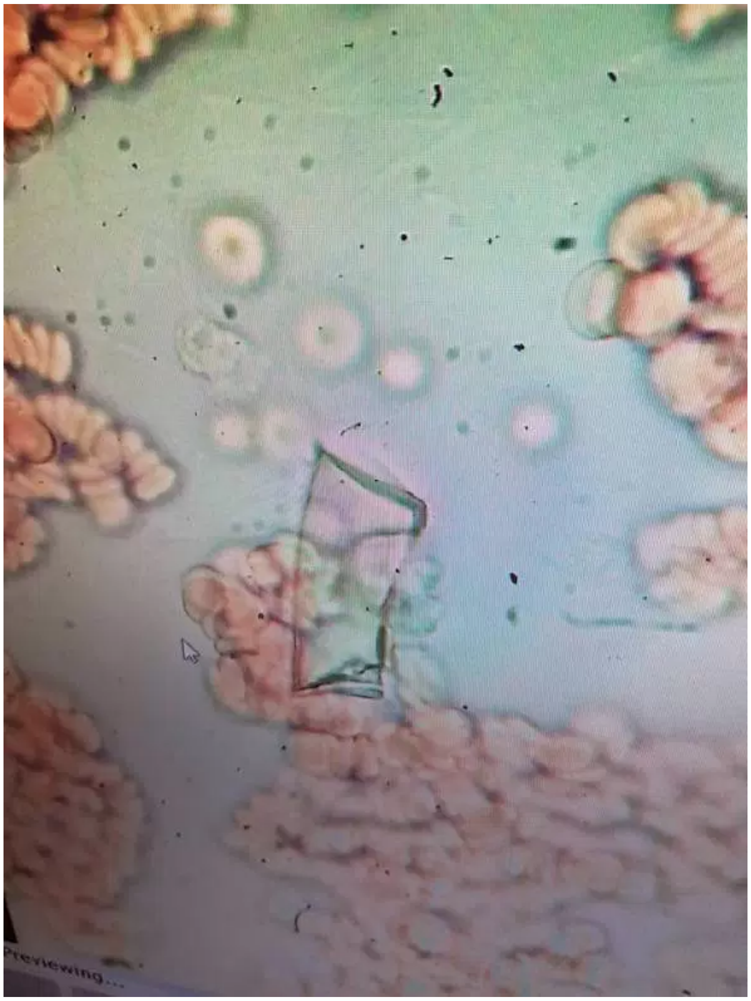

(https://nanobrainimplant.com/?attachment_id=1102)

During my staying in Poland, in Lublin I met a specialist Dr, who took a sample of my blood from my finger and we watched this together through an atomic microscope.

I will not now describe what is wrong with my blood, but I want you to notice the cristal which was made artificially.

what this is doing to me:

1. 1. makes my blood, not circulate in a natural manner (blood is too dense)
2. 2. damage DNA
3. 3. Influence memory (lack of memory)
4. 4. you can not walk normally nor run
5. 5. destroy your skin (you may have albinism)
6. 6. having a hard time sleeping
7. 7. easy access to the brain
8. 8. lack of oxygen

and many other symptoms…

There are millions of this in my blood. Every victim should do this checking with a doctor who knows about People Online and will check this blood with you through a atomic microscope.

This, together with the waves RF, ELF, SCALAR, NEUTRINOS, QUATUM SPECTRA, LASER, SONAR and others is devastating.

Previewing...

(https://nanobrainimplant.com/2016/09/11/the-reality-of-magnus-olssons-blood-doctors-report/magnus-olsson-blood-2/)

The most important for us is to secure ourselves and then cure. I am now working on a breathtaking solution for all of us, and I hope we will be free pretty soon. Then we can recover and heal.

Love and Light

Your Magnus

THE Reality of Magnus Olsson's blood – doctor's report:

Crimes against humanity are certain acts that are committed as part of a widespread or systematic attack directed against any civilian population or an identifiable part of a population:
https://en.m.wikipedia.org/wiki/Crimes_against_humanity (https://en.m.wikipedia.org/wiki/Crimes_against_humanity)
THE PENTAGON'S BLUE-SKY research arm wants to trick out troops' brains, from the areas that regulate alertness and cognition to pain treatment and psychiatric well-being. And the scientists want to do it all from the outside in:
http://google.co.in/patents/US3951134 (http://google.co.in/patents/US3951134)

https://en.m.wikipedia.org/wiki/Rayleigh_wave (https://en.m.wikipedia.org/wiki/Rayleigh_wave)

https://en.m.wikipedia.org/wiki/High_Frequency_Active_Auroral_Research_Program
(https://en.m.wikipedia.org/wiki/High_Frequency_Active_Auroral_Research_Program)
What if a machine could read your mind?
https://drive.google.com/file/d/0B4iZUaEgNfFfMnpyem5fX1BPTTg/view (https://drive.google.com/file/d/0B4iZUaEgNfFfMnpyem5fX1BPTTg/view)
While most developmental robotics projects strongly interact with theories of animal and human development, the degrees of similarities and inspiration between identified biological mechanisms and their counterpart in robots, as well as the abstraction levels of modeling, may vary a lot.
https://en.m.wikipedia.org/wiki/Neuron_(software) (https://en.m.wikipedia.org/wiki/Neuron_%28software%29)
While some projects aim at modeling precisely both the function and biological implementation (neural or morphological models), such as in neurorobotics www.humanbrainproject.eu (http://www.humanbrainproject.eu/), some other projects only focus on functional modeling of the mechanisms and constraints described above, and might for example reuse in their architectures techniques coming from applied mathematics or engineering fields.
https://en.m.wikipedia.org/wiki/Developmental_robotics (https://en.m.wikipedia.org/wiki/Developmental_robotics)

https://m.facebook.com/Artificial-Brain-Sweden-1014306015294202/ (https://m.facebook.com/Artificial-Brain-Sweden-1014306015294202/)
What are the ethical issues involved in simulating a human brain and in technology derived from human brain simulation?
Building computer models of the brain may challenge our concepts of personhood, free will and personal responsibility, and the nature of consciousness.
(Delgado stated that "brain transmitters can remain in a person's head for life.
The energy to activate the brain transmitter is transmitted by way of radio frequencies.
https://en.m.wikipedia.org/wiki/Jos%C3%A9_Manuel_Rodriguez_Delgado (https://en.m.wikipedia.org/wiki/Jos%C3%A9_Manuel_Rodriguez_Delgado))

https://drive.google.com/file/d/0B4iZUaEgNfFfZl9iU1BpbzNJTW8/view (https://drive.google.com/file/d/0B4iZUaEgNfFfZl9iU1BpbzNJTW8/view)
In medicine, brain simulation could make it easier to communicate with people who cannot speak (e.g. people with severe disabilities, people in a vegetative state or with locked-in syndrome) or to enhance cognitive function in people with cognitive disabilities (e.g. dementia, trauma and stroke victims, etc.).
https://en.m.wikipedia.org/wiki/Wetware_(brain) (https://en.m.wikipedia.org/wiki/Wetware_%28brain%29)
As in other fields of science, it also possible that new knowledge about the brain will be abused – deliberately, for example to create new weapons – but also involuntarily, because society does not realize the power and consequences of new technologies.
https://m.facebook.com/Martin-Ingvar-KI-1528410917453814/ (https://m.facebook.com/Martin-Ingvar-KI-1528410917453814/)
For instance, it may be possible in the future to use knowledge about the brain to predict and modify individual behaviour, or even to irreversibly modify behaviour through electrical stimulation of the brain, pharmacology or neurosurgery. In cases of intractable mental disease, this may be desirable, but in other cases, the costs and benefits will be debatable. One example of debate is whether society should allow cognitive enhancement in healthy people.
Similar considerations apply to technology. Future computers that implement the same principles of computation and cognitive architectures as the brain have enormous potential to improve industrial productivity and offer new services to citizens.
However, they could also be used to implement new systems of mass surveillance and new weaponry.
https://m.youtube.com/watch?v=o9bd-B2dqCM&feature=youtu.be (https://m.youtube.com/watch?v=o9bd-B2dqCM&feature=youtu.be)

If such systems came into widespread use they would undoubtedly have a huge impact on patterns of daily life and employment – this could be both beneficial and detrimental
https://www.humanbrainproject.eu/faq/ethics (https://www.humanbrainproject.eu/faq/ethics)

This is what it's all about:
https://m.youtube.com/watch?v=01hbkh4hXEk&feature=youtu.be (https://m.youtube.com/watch?v=01hbkh4hXEk&feature=youtu.be)
The development of your Future connected to a mindreading machine:
https://www.rt.com/usa/265029-kurzweil-google-hybrid-brain/ (https://www.rt.com/usa/265029-kurzweil-google-hybrid-brain/)

http://www.ted.com/talks/ray_kurzweil_get_ready_for_hybrid_thinking (http://www.ted.com/talks/ray_kurzweil_get_ready_for_hybrid_thinking)
And the developing of Future Surveillance:
https://m.youtube.com/watch?list=PL9DCCHoTYZ8KaMFrLrUzNblMsIECi9vQ0&v=30seQeBI-Tc (https://m.youtube.com/watch?list=PL9DCCHoTYZ8KaMFrLrUzNblMsIECi9vQ0&v=30seQeBI-Tc)

https://m.youtube.com/watch?feature=youtu.be&v=pW1HACMaOME (https://m.youtube.com/watch?feature=youtu.be&v=pW1HACMaOME)
Imagine what U.S could do if they could master the nanoparticles that are distributed in vaccine to human's worldwide and use them for remote neural monitoring:
https://mind-computer.com/2012/12/20/remote-neural-monitoring-a-technology-used-for-controlling-the-human-brain/ (https://mind-computer.com/2012/12/20/remote-neural-monitoring-a-technology-used-for-controlling-the-human-brain/)
Nanoparticles BMI https://en.m.wikipedia.org/wiki/Brain%E2%80%93computer_interface (https://en.m.wikipedia.org/wiki/Brain%E2%80%93computer_interface) in vaccine:
Nanofluids also have special acoustical properties and in ultrasonic fields display additional shear-wave reconversion of an incident compressional wave; the effect becomes more pronounced as concentration increases.
https://en.m.wikipedia.org/wiki/Vaccine (https://en.m.wikipedia.org/wiki/Vaccine)

https://en.m.wikipedia.org/wiki/Aluminium (https://en.m.wikipedia.org/wiki/Aluminium)

http://sharpbrains.com/blog/2015/03/19/non-invasive-brain-stimulation-meets-nanotechnology/ (http://sharpbrains.com/blog/2015/03/19/non-invasive-brain-stimulation-meets-nanotechnology/)

http://globalbiodefense.com/2012/07/28/darpa-program-hits-milestone-in-plant-based-vaccines-for-pandemics/ (http://globalbiodefense.com/2012/07/28/darpa-program-hits-milestone-in-plant-based-vaccines-for-pandemics/)

https://en.m.wikipedia.org/wiki/Nanofluid (https://en.m.wikipedia.org/wiki/Nanofluid)
This speaks volumes about the hidden impact certain vaccines may have on your body and your brain in particular…
http://articles.mercola.com/sites/articles/archive/2011/08/06/vaccine-increases-narcolepsy-by-660-percent.aspx (http://articles.mercola.com/sites/articles/archive/2011/08/06/vaccine-increases-narcolepsy-by-660-percent.aspx)

Delgado stated that "brain transmitters can remain in a person's head for life.
The energy to activate the brain transmitter is transmitted by way of radio frequencies.
https://en.m.wikipedia.org/wiki/Jos%C3%A9_Manuel_Rodriguez_Delgado (https://en.m.wikipedia.org/wiki/Jos%C3%A9_Manuel_Rodriguez_Delgado)

https://www.google.com/patents/DE10253433A1?cl=en&dq=inassignee%3A (https://www.google.com/patents/DE10253433A1?cl=en&dq=inassignee%3A)

https://archive.org/details/pubmed-PMC4086297 (https://archive.org/details/pubmed-PMC4086297)

https://en.m.wikipedia.org/wiki/Neurorobotics (https://en.m.wikipedia.org/wiki/Neurorobotics)

https://drive.google.com/file/d/0B4iZUaEgNfFfell0YWFFMHd3RnM/view (https://drive.google.com/file/d/0B4iZUaEgNfFfell0YWFFMHd3RnM/view)

https://en.m.wikipedia.org/wiki/Cybernetics (https://en.m.wikipedia.org/wiki/Cybernetics)

https://en.m.wikipedia.org/wiki/Target_Motion_Analysis (https://en.m.wikipedia.org/wiki/Target_Motion_Analysis)

http://www.raytheon.com/ (http://www.raytheon.com/)
In analysis such as computational fluid dynamics (CFD), nanofluids can be assumed to be single phase fluids.
http://www.datacenterdynamics.com/news/sgi-gives-us-military-the-fastest-supercomputer/91007.fullarticle (http://www.datacenterdynamics.com/news/sgi-gives-us-military-the-fastest-supercomputer/91007.fullarticle)
The Worldwide Security Grid: MATRIX Smartdust is a system of many tiny microelectromechanical systems (MEMS) such as sensors, robots, or other devices, that can detect, for example, light, temperature, vibration, magnetism, or chemicals. They are usually operated on a computer network wirelessly and are distributed over some area to perform tasks, usually sensing through radio-frequency identification.
http://www.foi.se/en/Customer–Partners/Projects/Edge1/Edge/ (http://www.foi.se/en/Customer--Partners/Projects/Edge1/Edge/)
Without an antenna of much greater size the range of tiny smart dust communication devices is measured in a few millimeters and they may be vulnerable to electromagnetic disablement and destruction by microwave exposure.
https://en.m.wikipedia.org/wiki/Smartdust (https://en.m.wikipedia.org/wiki/Smartdust)
The Grid Security Infrastructure (GSI), formerly called the Globus Security Infrastructure https://en.m.wikipedia.org/wiki/Grid_Security_Infrastructure (https://en.m.wikipedia.org/wiki/Grid_Security_Infrastructure), is a specification for secret, tamper-proof, delegatable communication between software in a grid computing environment. Secure, authenticatable communication is enabled using asymmetric encryption.
http://ieeexplore.ieee.org/xpl/login.jsp?reload=true&tp&arnumber=6028683&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6028683 (http://ieeexplore.ieee.org/xpl/login.jsp?reload=true&tp&arnumber=6028683&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6028683)
In software engineering, the terms "front end" and "back end" are distinctions which refer to the separation of concerns between a presentation layer and a data access layer respectively.
The front end is an interface between the user and the back end. The front and back ends may be distributed amongst one or more systems.
In software architecture, there may be many layers between the hardware and end user. Each can be spoken of as having a front end and a back end. The

front is an abstraction, simplifying the underlying component by providing a user-friendly interface.
https://en.m.wikipedia.org/wiki/Pervasive_game (https://en.m.wikipedia.org/wiki/Pervasive_game)
In software design, for example, the model-view-controller architecture provides front and back ends for the database, the user and the data processing components. The separation of software systems into front and back ends simplifies development and separates maintenance. A rule of thumb is that the front (or "client") side is any component manipulated by the user. The server-side (or "back end") code resides on the server.
https://en.m.wikipedia.org/wiki/Front_and_back_ends (https://en.m.wikipedia.org/wiki/Front_and_back_ends)

https://m.youtube.com/watch?v=o9bd-B2dqCM&feature=youtu.be (https://m.youtube.com/watch?v=o9bd-B2dqCM&feature=youtu.be)
The Utah Data Center, also known as the Intelligence Community Comprehensive National Cybersecurity Initiative Data Center, is a data storage facility for the United States Intelligence Community that is designed to store data estimated to be on the order of exabytes or larger.
Its purpose is to support the Comprehensive National Cybersecurity Initiative (CNCI), though its precise mission is classified.
http://www.defense.gov/News/Speeches/Speech-View/Article/606635 (http://www.defense.gov/News/Speeches/Speech-View/Article/606635)

https://www.iarpa.gov/index.php/research-programs/neuroscience-programs-at-iarpa (https://www.iarpa.gov/index.php/research-programs/neuroscience-programs-at-iarpa)

https://en.m.wikipedia.org/wiki/Wetware_(brain) (https://en.m.wikipedia.org/wiki/Wetware_%28brain%29)

http://www.datacenterdynamics.com/news/sgi-gives-us-military-the-fastest-supercomputer/91007.fullarticle (http://www.datacenterdynamics.com/news/sgi-gives-us-military-the-fastest-supercomputer/91007.fullarticle)

http://bioethics.gov/node/4704 (http://bioethics.gov/node/4704)

https://m.facebook.com/Brain-Print-1585272778401091/ (https://m.facebook.com/Brain-Print-1585272778401091/)
The National Security Agency (NSA) leads operations at the facility as the executive agent for the Director of National Intelligence.
It is located at Camp Williams near Bluffdale, Utah, between Utah Lake and Great Salt Lake and was completed in May 2014 at a cost of $1.5 billion.
https://en.m.wikipedia.org/wiki/Utah_Data_Center (https://en.m.wikipedia.org/wiki/Utah_Data_Center)

Any Questions?

http://www.lifecoachcode.com/2016/05/28/the-father-of-fractal-geometry-reveals-the-pattern/ (http://www.lifecoachcode.com/2016/05/28/the-father-of-fractal-geometry-reveals-the-pattern/)

In the computer software world, open source software concerns the creation of software, to wIntelligence.
https://en.m.wikipedia.org/wiki/List_of_artificial_intelligence_projects (https://en.m.wikipedia.org/wiki/List_of_artificial_intelligence_projects)

https://m.facebook.com/Artificial-Genocide-1716276305283944/ (https://m.facebook.com/Artificial-Genocide-1716276305283944/)

https://m.facebook.com/Doomsday-Computers-AGI-ASI-516866275144461/ (https://m.facebook.com/Doomsday-Computers-AGI-ASI-516866275144461/)

https://m.youtube.com/watch?list=PL9DCCHoTYZ8KaMFrLrUzNblMsIECi9vQ0&v=g1lVBNV6ztw (https://m.youtube.com/watch?list=PL9DCCHoTYZ8KaMFrLrUzNblMsIECi9vQ0&v=g1lVBNV6ztw)

#NBIC #DualUseTechnology #HBP #LofarLois #Ericsson #IBM #SGI #Raytheon #NSA #MilitaryNanoTechnology #BrainInitiative #NeuroEthics

hich access to the underlying source code is freely available.
This permits use, study, and modification without restriction.
In computer security, the debate is ongoing as to the relative merits of the full disclosure of security vulnerabilities, versus a security-by-obscurity approach.
There is a different (perhaps almost opposite) sense of transparency in human-computer interaction, whereby a system after change adheres to its previous external interface as much as possible while changing its internal behaviour.
That is, a change in a system is transparent to its users if the change is unnoticeable to them.
https://en.m.wikipedia.org/wiki/Transparency_(behavior) (https://en.m.wikipedia.org/wiki/Transparency_%28behavior%29)

All Human's are Computers in Ubiquitous Computing that are monitored by Artificial Intelligence/Ambient

By mindcontroleurope • Posted in Implant for Synthetic Telepathy, military, Mind Control, Nano Brain Implant, Nano ICT Brain Implants • Tagged artifical-intelligence, blood, brain-computer-interface, doctor, intelligence-gathering, magnus olsson, mind control, poland, post doctoral research, waves, world cach
AUG 5 2016

Assassination is the murder of a person, often (but not or ruler, usually for political reasons or paymen. Or the common man.

Assassination is the murder of a person, often (but not always) or ruler, usually for political reasons or paymen. Or the common man.

1010 1010 1000 0111 0100 1010
0111 0100 1010 1000 1010

(https://nanobrainimplant.com/2016/08/05/assassination-is-the-murder-of-a-prominent-person-often-but-not-always-or-ruler-usually-for-political-reasons-or-paymen-or-the-common-man/brain-control-2/)

An assassination may be prompted by, political, or military motives; it is an act that may be done for financial gain, to avenge a grievance, from a desire to acquire fame or notoriety, or because of a military, security or insurgent group's command to carry out the homicide.

The World Coalition against Covert Harassment is committed to raising awareness to the legal systems as well as to the medical and scientific community to the crime of illegal biomedical and weaponry research committed on citizens in the European Union and beyond. As a European network, EUCACH acts as a lobbying and advocacy platform towards the EU. Using our international network of scientific and technology experts, partner civil and human rights organisations as well as important stakeholders in civil society, we provide consultancy services to the EU Institutions based on our expertise. EUCACH's organisational goal is to influence EU legislation and the decision-making process in calling for a worldwide ban on weapons that might enable any form of manipulation of human beings.

Online-connected brains and neural networks.

"ICT" = Information and Communication Technologies

"BMI" = Brain Machine Interface, brain-computer interconnection

"FET" = Future and Emerging Technologies

"S.T." = Synthetic Telepathy

"A.I." = Artificial Intelligence.

On the road to mind control: DARPA's new program will use a chip to connect brains.For additional background to the latest press release from DARPA posted in full below, I encourage you to read the following selection of linked articles
where I discuss the scope and chronology of what is being studied. Therein, you will find that the U.S. BRAIN Initiative and its European counterpart, the Human Brain Project, are not spending multi-billions of dollars on neuroscience research
simply to help people with Post Traumatic Stress Disorder and organic brain dysfunction. It is, perhaps first and foremost, a military endeavor that has wide ramifications if even 1/10th of what is being studied comes to fruition. In short, it's more about mind control than it is about brain restoration and improvement. Please keep this in mind when you read DARPA's emphasis on "new therapies." •Obama Doubles Down on BRAIN Project and Military Mind Control

•Mind Control Scientists Find New Memory Manipulation Technology
•Secret DARPA Mind Control Project Revealed: Leaked Document
•The 9 Goals of Mind Control: Interim Report
•New Mind Reading Research Aims to Synchronize Humans
•Nanoparticles Enable Remote Control Brains Via Magnetic Field: New Study

DARPA Press Release:

DARPA

(https://nanobrainimplant.com/2014/12/17/disturbing-weapons-now-being-used-on-the-masses/darpa/)

A new DARPA program aims to develop an implantable neural interface able to provide unprecedented signal resolution and data-transfer bandwidth between the human brain and the digital world. The interface would serve as a translator, converting between the electrochemical language used by neurons in the brain and the ones and zeros that constitute the language of information technology. The goal is to achieve this communications link in a biocompatible device no larger than one cubic centimeter in size, roughly the volume of two nickels stacked back to back.

The program, Neural Engineering System Design (NESD), stands to dramatically enhance research capabilities in neurotechnology and provide a foundation for new therapies.

“Today’s best brain-computer interface systems are like two supercomputers trying to talk to each other using an old 300-baud modem,” said Phillip Alvelda, the NESD program manager. “Imagine what will become possible when we upgrade our tools to really open the channel between the human brain and modern electronics.”

Among the program’s potential applications are devices that could compensate for deficits in sight or hearing by feeding digital auditory or visual information into the brain at a resolution and experiential quality far higher than is possible with current technology.

Neural interfaces currently approved for human use squeeze a tremendous amount of information through just 100 channels, with each channel aggregating signals from tens of thousands of neurons at a time. The result is noisy and imprecise. In contrast, the NESD program aims to develop systems that can communicate clearly and individually with any of up to one million neurons in a given region of the brain.

Achieving the program’s ambitious goals and ensuring that the envisioned devices will have the potential to be practical outside of a research setting will require integrated breakthroughs across numerous disciplines including neuroscience, synthetic biology, low-power electronics, photonics, medical device packaging and manufacturing, systems engineering, and clinical testing. In addition to the program’s hardware challenges, NESD researchers will be required to develop advanced mathematical and neuro-computation techniques to first transcode high-definition sensory information between electronic and cortical neuron representations and then compress and represent those data with minimal loss of fidelity and functionality.

To accelerate that integrative process, the NESD program aims to recruit a diverse roster of leading industry stakeholders willing to offer state-of-the-art prototyping and manufacturing services and intellectual property to NESD researchers on a pre-competitive basis. In later phases of the program, these partners could help transition the resulting technologies into research and commercial application spaces.

To familiarize potential participants with the technical objectives of NESD, DARPA will host a Proposers Day meeting that runs Tuesday and Wednesday, February 2-3, 2016, in Arlington, Va. The Special Notice announcing the Proposers Day meeting is available athttps://www.fbo.gov/spg/ODA/DARPA/CMO/DARPA-SN-16-16/listing.html. More details about the Industry Group that will support NESD is available athttps://www.fbo.gov/spg/ODA/DARPA/CMO/DARPA-SN-16-17/listing.html. A Broad Agency Announcement describing the specific capabilities sought is available at: http://go.usa.gov/cP474 (http://l.facebook.com/l.php?u=http%3A%2F%2Fgo.usa.gov%2FcP474&h=LAQF0y59EAQFQR18F27x04gXa7Y_rindsR25iSKphNuN6zA&enc=AZOpn7sJYv1piS9gSXGfy_NqXsCO7oB0S4
wwq-BZaaREHQxwob2U8UXt8kfaCnxig3GT6vilhS7K6HhnYGtn9hKAUxHcqAYyBQ8RvfuBKDruex9_vs-icDjhjTswRnvlZiuFpTbFop5HBD3RZ38KhQmXzcBN3gFvyHixaG5LgtM-FvdH37Galx8Kne2oujUkZt3E&s=1).

DARPA anticipates investing up to $60 million in the NESD program over four years.

NESD is part of a broader portfolio of programs within DARPA that support President Obama’s brain initiative. For more information about DARPA’s work in that domain, please visit:http://www.darpa.mil/…/our-r…/darpa-and-the-brain-initiative (http://l.facebook.com/l.php?u=http%3A%2F%2Fwww.darpa.mil%2Fprogram%2Four-research%2Fdarpa-and-the-brain-initiative&h=HAQECCJXjAQEZD2OkSMYSVL85yN-3EkJvVb9bU9LzVZeKmQ&enc=AZNCWwf9R2i6Uy5lFdfkj6UBSyM39uSN_sZKmEQN0J14oy0ih7JhkHYCtvqG0N0PmELXHOuKdGmXs9kWiHDiy7ubLMS
cr3v6PcjPMVGzyRQND8fQLFaQwZVkVqmre58u3yG7h1NdxM8L8zYhn-_5v&s=1).

FET and ICT research and development (of the new “computer – brain language”) allows computers to read and learn human thought patterns by using injectable brain – machine – interface.

About the same brain-machine interface has has been used to cure Parkinson’s disease , Alzheimer’s and depression can also be used for accessing the brain’s memory.

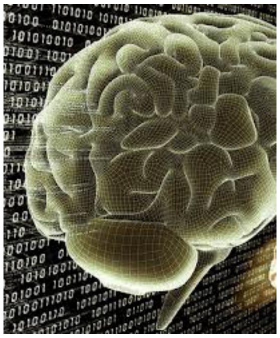

(https://nanobrainimplant.com/2016/08/05/assassination-is-the-murder-of-a-prominent-person-often-but-not-always-or-ruler-usually-for-political-reasons-or-paymen-or-the-common-man/brain-mind-e/)

Digital Mind

Professor Goran Hermerén OPINION OF THE EUROPEAN GROUP ON ETHICS IN SCIENCE AND NEW TECHNOLOGIES TO THE EUROPEAN COMMISSION (ETHICAL ASPECTS OF ICT IMPLANTS IN THE HUMAN BODY) March 2005

Link to document

The new software for the brain-machine interface in nano electronics combined with "Europe's new information technology" provides the researchers with the possibillities for reading and taking "software images" of the brain's neuron network.

A method that provides high-resolution copies of the brains cognitive behavior and human perception.

Tomorrow's high-speed computers and related research has evolved into a sophisticated "computer game" with "mind reading" on real people….

These research methods are unknown to our society's legal system!

So how can research on the brain by "serious" (criminal) organizations in ICT-FET be stopped?

SYNTHETIC TELEPATHY for medicine or mind reading is a direct communication with nanoelectronics between computer and human brain.

Synthetic telepathy is a communication system based on thought, not speech.

It can be used to control for example a prosthesis and cure diseases such as Parkinson's disease but can also decode the patterns used to study cognitive behavior such as memory, learning and emotion. With nano technology, the digital information can be recorded using the new quantum-inspired computers.

Simulation of behavior can be identified and provide diagnostic data for identifying the precursors to diseases such as dementia, stroke and myocardial infarction

This letter is intended to demonstrate a paradox in the Swedish so-called "protection of human rights" and thus the entire Swedish justice system. Synthetic telepathy could, as practiced in Sweden, lie behind an unknown number of violent crimes and suicides due to research deliberately kept hidden from regulators.

Doctors and psychiatrics diagnose people with "voices in his head" following their "manuals". DSM-IV (Diagnostic and Statistical Manual of Mental Disorders) published by the American Psychiatric Association and the ICD-10 (International Statistical Classification of Disorders and Related Health Problems).

This is an old-fashioned black and white and diagnose when the EU priority FET ICT research and develop new information technologies adapted to nano-electronics.

mind-control-weapons

Ulf Gorman writes in the book that with nano technology, we open the doors to an unknown area where we do not know how to apply ethics. What should be and what should not be allowed when you implant a chip that can both read and influence the brain? He takes the example of studying learning and memory.

Micro Implants can provide unprecedented opportunities to understand how we learn and remember things, and hence why we forget and find it difficult to learn. And it can be understood as a form of abuse to look like that into our most private mental world.

Lund University writes about the development of nano-electrodes that can both listen and communicate with neuronal synapses and their cell membranes.

EU priority ICT and FET research are talking about a "A whole new communication technology in Europe" "It will help us understand and exploit the ways in which social and biological systems, organization and evolution, will pave the way for the development of new opportunities for next-generation software and network technologies ".
To understand how the human brain works not only leads to innovations in medicine but also provides new models for energy, fault-tolerant and adaptive computing technologies ".
An initiative of the Virtual Human Physiology that are individually tailored and virtual simulations of the human body where you would expect enormous progress in disease prevention and health care.
The pioneering work carried out also on new ideas such as artificial living cells, synthetic biology, chemical communication, collective intelligence and two-way interface between brain and machine.

(https://nanobrainimplant.com/2016/08/05/assassination-is-the-murder-of-a-prominent-person-often-but-not-always-or-ruler-usually-for-political-reasons-or-paymen-or-the-common-man/new-brain-2/)

Brain reading

Other sources, e.g The UCI (University of California Irvin. Department of Cognitive Science) describes the development of Synthetic Telepathy:

Collaboration between cognitive science, neuro-science, specialists in speech recognition and brain imaging will develop a brain-machine interface. This device could help paralyzed and soldiers would be able to send messages directly from the brain to a computer.

Researcher Michael D'Zmura, President of the UCI describes that the system begins with "the little voice in your head".

How can a medical diagnosis unequivocally describe people's perception of voices in their head and that no alternative can exist except mental disease.

Why? -because it is a prerequisite to be able to "hear" voices for the use of new information technology?

This is why these researchers must be forced to go public and announce this new scientific communication technology.

This paradox must be investigated immediately.With the exclusion of the development of a Swedish and European military force with superior two-way "radio" communication with the brain.

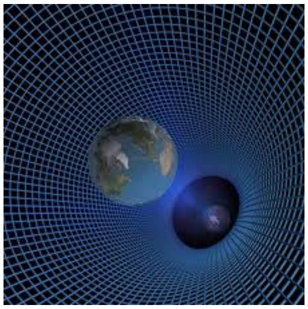

(https://nanobrainimplant.com/2012/07/11/scientists-humans-and-machines-will-merge-in-future/imagescaygmuja/)

A number of past court cases are more or less directly caused by S.T. This "voice to skull" technology must immediately be taken into consideration as an alternative for triggering a number of previously committed violent crimes and suicides.

During the development of BMI, software and network technologies are also computerized and long distance imaging of peoples cognitive behavior and perception. Material that is recorded in the computer that runs the real time simulation and creation of artificial intelligence (for initializing A.I. and computerized decision making.)

The image of the brain's "machine-code" is probably the most comprehensive and advanced ever made. Cognitive behavior depicted and simulated, language and meaning of the words for the subjects are identified. Human perception and mapping how the brain handles information, the image of mathematics reached its perfection.

neuron wave A new research scandal and the fatal consequences for the wider community.

Around the clock the computerized study goes on using collective, artificial intelligence and self-learning systems. The victims testimonies tells us that you can sadly conclude unequivocally that the studies will not be completed until the victim in one way or another has been broken down and / or otherwise inactivated.

This provides power to simulate decomposition for the digital copy thus the fact that the criminal research will not end and is never disclosed.

Descriptions to the vulnerable people who eat psychotropic drugs due to their experiences the experiments and testing goes on, various medicines can affect the "test objects" (or guinea pigs) everything is recorded and compared with previous values (from the multi-annual copy of their real-time neural network and thus the registration of their behavior.)

Obviously this is a disgusting and illegal way to meet the advanced development of tomorrow's medicine.

The "studies" have resulted in enormous damage to many subjects.The number of unrecorded victims/subjects is probably very high.Families are fragmented, the children of these families are suffering tremendously. One of the subjects have recently been hospitalized with a cracked skull, caused by disconnecting the balance system remotely. The accident occurred in public settings.

A new research scandal and the fatal consequences for the wider community.

To influence the balance system is another typical example of how technology is used against the victim to incapacitate them for society. Direct assassination attempts on several previous occasions orchestrated by the brain control deliberately strikes out the human balance system, Why this happens is likely as an alternative to the mentally ill to be treated (as "possessed") or having a neurological disease.

(https://nanobrainimplant.com/2012/06/03/2050-and-immortality-is-within-our-grasp/imagescaaztqrn/)

Brain – copying with BMI and broadband access is definitely no longer a marginal research.Employers have over the years lost millions of dollars and cut-downs is to be expected due to excluded employees. Broken families, children who, years later can not study or work is now forced to seek psychiatric care.

Siblings, grandparents and their closest friends are suffering tremendously. Advanced Medical and hospital care due to study design is a mockery of health-care and doctors who are not familiar with this research. Insurance in the multi-million figures was raised through direct damage caused by brain control. Property for millions of dollars are lost.

Injuries and privacy intrusion of thousands of people is done through “volunteers” that serve as multimedia machines i.e node for the recording of all contacts they have, such as politicians, scientists, lawyers, friends, acquaintances, relatives, international business relations, etc..

Security codes, access codes, etc. is with the new technology no longer private.How can we know that people are not equipped with the new brain-machine interface which makes the person a multimedia application with a function as a live missile is already deployed as nodes in the political and financial world of illegal recording of their conversations with the world?

A sinister JOY seems to embrace the researchers and perpetrators over this superior and powerful tool for mind control and copying brainwave patterns. The tool is, without a doubt, a weapon of offense and stealth.

Implant technology in these forms should immediately be classified as a lethal weapon! It communicates directly with the brain’s neurons and can bring the entire neural nervous system to a halt.(Exclusion of balance can be immediately implied on a victim.)

The technology is now used for purposes of breaking down the persons psyche, with serious accusations, threats, mock execution, incitement to suicide, physical violence, e.g, decrease and increase heart rate, pain in and around the heart, severe chest pains, sudden and painful headaches, difficulty to breathe, tampering with rectum, prostate and muscles to name a few.

The macabre in the use of this technology is that subjects are exposed to these atrocities while the society is not legally able to influence the situation. Sweden is in this matter a lawless country where the researchers grossly exploit the situation.

(https://nanobrainimplant.com/2013/03/24/brain-control-war-too-big-to-fail/imagescaay7q87/)

Face mind

This ultimate humiliation has reached a whole new level. The researchers tries with the enforced communication and slow decomposition enslave people with exhortations to try to see the individual results of the brain control. One can calculate due to the nature and perennial perspective that there are naturally a unrecorded number of people who have been driven to madness and death with this technology.

There is nowhere the victims can hide or escape the access signals to their thoughts. Scientists simulate with computers, 24/7 365 days a year to break down the subjects and stop them from trying to understand what’s happening to them…

A series of grotesque roles played in order to manipulate the brain, threats and statements that are mixed with modern technology. It is also in the researchers’ strategy to make the picture unclear for the subjects if they attempt to get an overall picture of who the perpetrator is and the real goals of the research, a military strategy conducted by veterans and experts in the matters.

The disrespectful research performed and visualized in a 3D virtual game world in the researchers’ computers, with no ethical boundaries and human rights but with real living human beings as “avatars”. It is quite similar to the popular interactive game “The Sims‘, But this game delivers human reality-based measurement data for research.

Cover-up of brain monitoring technologies means that the crime is waterproof, human rights laws and manipulated by a hidden militant regime researcher with expertise in information technology. In addition, the researchers say they in the dialogue to be the police power which in itself is an extremely serious offense.

For the victims– former high-performance hard-working people with families, children, an orderly life and social contacts. People who all his life been performing taxpayers. Because of a work-related mental fatigue and time on medical therapy sessions with the scientists the opportunity to take advantage of the situation of persons for investigation and contemporary development of the new BMI and brain monitoring technologies.

As these technologies and opportunities are not announced, but several instances re-written, must be able to use knowledge and skills which are available. Sweden is a small country and the people engaged in this activity may not be so difficult to identify and stop.

The researchers in these studies have assumed the right that during the permanent reproduction of human neurotransmitters in the long term also destroy them and their life’s work. The issue is called for; How affected society to know that the violent crime and suicide has been performed in the Stockholm area and Europe in recent years and clearly diagnosed as being caused by mental illness with the voices in heads is an expression of pure brain research!

By: Magnus Olsson

(https://nanobrainimplant.com/2013/04/23/nsa-mind-control-and-psyops-2/nsa/)

HOW WE DO IT…

- Awareness-raising campaigns to disseminate information and mobilise public opinion on the issue of covert technologies and techniques that enable the manipulation of human beings
- Providing expert consultancy to key decision-makers on the creation of appropriate EU legislation to protect citizens from this kind of covert crimes.
- Organising networking events (workshops, seminars, conferences) involving all actors concerned to exchange experiences and best practices for the establishment of clear ethical boundaries to strictly regulate the use of systems enabling the manipulation and control of human beings.

It is our philosophy that all men are equal before the law. Everybody's right to life shall be protected. Nobody shall be subjected to torture or held in slavery. Any technologies and techniques capable of endangering the human physical and/or psychological health, to modify the individuals' autonomy and affect their dignity should be strictly prohibited.

Read more about ICT – implants at the European Group on Ethics and New Technologies:

http://ec.europa.eu/…/european-group-eth…/docs/avis20_en.pdf (http://ec.europa.eu/bepa/european-group-ethics/docs/avis20_en.pdf)

Terminology used in the document:

Cybernetics,,,,,, // Magnus Olsson

EPSC – European Commission (http://l.facebook.com/l.php?u=http%3A%2F%2Fec.europa.eu%2Fbepa%2Feuropean-group-ethics%2Fdocs%2Favis20_en.pdf&h=YAQGijSpj&enc=AZMq9f7N-rRi1oqozQeUDtiMrtgkAjH0PKPRaj0LR7nx44ZZmRLHC78kpg1NdCUG2ACxLZxzxYSdqtdeKAhL_RSUBcFCA0LkDQJxno5Rq7aQMwWo3MtKGZPGl3Id990LNnmKm_llFA98bTSAsGUrtrY&s=1)
EPSC
ec.europa.eu
By: Magnus Olsson (Sweden)

By mindcontroleurope • Posted in Computerbrain, Implant for Synthetic Telepathy, military, Mind Control, Uncategorized • Tagged artifical-intelligence, brain, brain implant, brain implants, cross sectional area, electrical resistance, eu brain, intelligence-gathering, magnus olsson, military / medical classified research, mind, transcranial-magnetic-stimulation
JUL 7 2016

Worldcach.com's Leo Angesleva (Sweden) about Mind Control and the Transhumanist Agenda and Alfred Webre

Alfred Lambremont Webre, Worldcach.com’s Leo Angesleva (Sweden) about Mind Control and the Transhumanist Agenda

Interview with Alfred Lambremont Webre, Worldcach.com (http://worldcach.com/)’s Leo Angesleva details how the “Obama” [read “U.S.”] Brain Project and the U.S. Army, along with other agencies covertly committed to the Transhumanist Agenda, are facilitating AI Artificial Intelligence reading and controlling all human brains and minds worldwide, and constructing Cyborg armies.

MAN VERSUS MACHINE

(https://nanobrainimplant.com/2012/08/22/ibm-get-close-to-mimicking-a-human-brain/imagescast4tl3/)

WORLD CACH
The World Coalition Against Covert Harassment

(https://nanobrainimplant.com/2016/07/07/alfred-lambremont-webre-worldcach-coms-leo-angesleva-sweden/world-cach/)

By mindcontroleurope • Posted in Computerbrain, Man vs Machine, military, Mind Control • Tagged Alexander, artifical-intelligence, brain-computer-interface, cyborg, Cyborg-soldier, eucach, hologram, Leo, magnus olsson, mind control, nano brain, transcranial-magnetic-stimulation, world cach
MAR 4 2016

Magnus Olsson launches World Coalition against transhumanist neural robotoid agenda

Magnus Olsson launches World Coalition against transhumanist neural robotoid agenda

<iframe width="640" height="360"

World Coalition against Covert Harassment (WorldCACH)

(https://nanobrainimplant.com/2013/02/11/are-people-superfluous/jj/)

STATEMENT OF GARY OWEN

It was Sunday 28th of February 2016, Magnus Olsson was staying at my house in Spain, I had been appointed the Vice President of World CACH 2 days before. We were planning the roll out of the World CACH, deciding the agenda of the organisation, we had a great day on the Saturday, we appointed the technical director, spoke to one of our legal team Alfred Lambremont Webre, he's the guy that won the war crimes tribunal against Tony Blair and George Bush. We explained that what we wanted to do with World CACH was to set up a Global Intelligence Service that acts for the people against the Governments and security services that no longer represent the people, and take the Masons who have infiltrated all levels of the government and security Services to the war crimes tribunal for genocide and racketeering against the population of earth. He thought both ideas were sensational and invited us to go on his show the following day, which would go out to 250,000 people.

On the Sunday, I started hearing the high pitched noise again in my ear I had not heard it for a few days, since Magnus arrived, and all of a sudden it was back along with the voice to skull. We were to meet for lunch with some people and later talk on Skype with William Binney and then do the show with Alfred where we would break the news to the world about what we were doing. It was only 30 minutes to the lunch meeting and I had heard nothing from Magnus, so I went down to the apartment he was staying in below my house. I knocked on the door and he answered saying he needed medical help, he was in a terrible state. He was physically shacking none stop, and his face was bright red with a high temperature.

(https://nanobrainimplant.com/links/globe1001-2/)

After some phone calls we decided to go to the emergency Chemist and get some pain killers and antibiotics as it looked like a virus. The Chemist said we should go to the emergency doctors in the town, by this time I was having to support Magnus while he walked, he was on the verge of passing out none stop.

Four doctors in the emergency room at Torrox looked at him and did all the tests they could and found nothing wrong. But they agreed something was wrong, so we were dispatched to the main hospital in Valez. I arrived at the emergency entrance, Magnus was no longer able to walk, they put him in a wheel chair. He was fading in and out and incoherent. After 2 hours we got to see a very proficient doctor, he was very thorough. Magnus had come round a little after drinking two litres of cold water. I thought the worst was over, Magnus was dispatched for X-rays and 12 blood tests, we did the X-rays first, no problems, then we went to do the blood tests, half way through taking the blood, they took a lot of blood, about 6 tubes, all of a sudden Magnus took a severe turn for the worse, it started by Magnus telling me that he was going to pass out. I tried to explain this to the Spanish nurse. Then he went as white as snow, all over, his face was pure porcelain white and I noticed that his hands were the same. I was instructed by the nurse to wave his hat in front of his face to try and revive him, his eyes rolled back into his head and you could only see white eyes and a white face it was shocking. By this time the nurse was in a sheer panic and the doctor that had seen him was called.

Before the doctor could arrive, Magnus started to have muscle spasms all over his body, they started on his spine, and spread all over, he was put on a gurney and was contorted and jerking all over the place with his eyes rolled back in his head. It was clear he would die of a heart attack if they did not stop it, they pressed the emergency button and doctors came from everywhere, he was wheeled at speed to the ICU Critical Emergency ward. They stripped of his cloths and by now there was a dozen people trying to figure out what was wrong.

I heard the original doctor tell the others that Magnus was working with CIA whistle blowers, we had explained a bit about the organisation to the doctor during the initial examination. They dismissed it and moved on, they could find nothing wrong. They injected him with Diazepam a strong muscle relaxant. He was in critical for two hours, during the first part I took a picture, I had taken several others, so I could put them on the World CACH facebook site later.

The last photo I took one of the doctors went crazy and threatened me to take no pictures and if any of the pictures I had taken were made public I would be charged by the police. This escalated and security arrived, one more sensible doctor said would you be OK to delete them, as you're not allowed to take photos in a Spanish hospital. There were no signs saying you couldn't.

I agreed to delete them rather than be arrested, I was taken to the public waiting room, three hours went by. A man arrived and told me I could visit Magnus, I was taken through, he was looking better but had an oxygen mask on, and was drained. I was told that in a few hours provided he remained stable he could leave, he was put on an intravenous antibiotic and saline drip. They dosed him up with the anti-biotics and gave me a prescription for a weeks long course and some diazepam in case the spasms started again.

After 7 hours in there we left, Magnus was back but very weak. The doctors told me that every test they did showed nothing, everything was exactly normal through the whole process. They agreed something was wrong but they did not know what it was. The whole day I had a strong high pitched noise whistling in my ear that is part of the Gangstaklking they are perpetrating on people now.

You have to ask yourself about the coincidences involved.

1/ Magnus and I were not being attacked like normal until we decided to Form a Global Intelligence Service for the people against the governments. Then we were both attacked.

2/ Nothing was found to be wrong with Magnus at all.

3/ That day we were to go on video and announce the plan to the TI world.

4/ We were making significant progress on the roll out of World CACH, talking to the leaders around the world of the individual country CACHs. Appointing the technical director. Getting the war tribunals accepted by a judge.

On the way home I started to wonder if we could not get the images back from the camera so I asked a young tech guy Alex Axon to see if it was possible, and a few hours later he had recovered three images and all the thumb nails. They were a bit degraded but we now have the images.

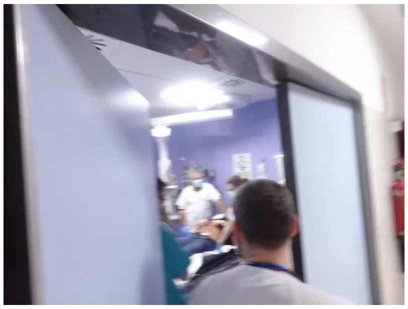

(https://i2.wp.com/newsinsideout.com/wp-content/uploads/2016/03/Magnus-Olsson-atHospital-ICU-Critical-Room-Feb.-28-2016.jpg?ssl=1)

Magnus Olsson at Hospital ICU Critical Room – Feb. 28, 2016

Magnus in ICU Critical Unit

What a day!

Gary Owen

World Coalition against Covert Harassment (WorldCACH)

By mindcontroleurope • Posted in Implant for Synthetic Telepathy, Magnus Olssons story, military, Mind Control, Nano Brain Implant, Uncategorized • Tagged airforce, artifical-intelligence, artificial telepathy, interfaces, magnus olsson, Nanobots, science, synthetic-telepathy, technology, transcranial-

magnetic-stimulation
JAN **28** 2015

The real-life Mind-Matrix…

Mail Online

(https://nanobrainimplant.files.wordpress.com/2015/01/mail-on-line.png)The real-life Matrix: MIT researchers reveal interface that can allow a computer to plug into the brain

(https://nanobrainimplant.files.wordpress.com/2015/01/fgnfgjtyjty.jpg)

- **System could deliver optical signals and drugs directly into the brain**
- **Could lead to devices for treatment of conditions such as Parkinson's**

It has been the holy grail of science fiction – an interface that allows us to plug our brain into a computer.

Now, researchers at MIT have revealed new fibres less than a width of a hair that could make it a reality.

They say their system that could deliver optical signals and drugs directly into the brain, along with electrical readouts to continuously monitor the effects of the various inputs.

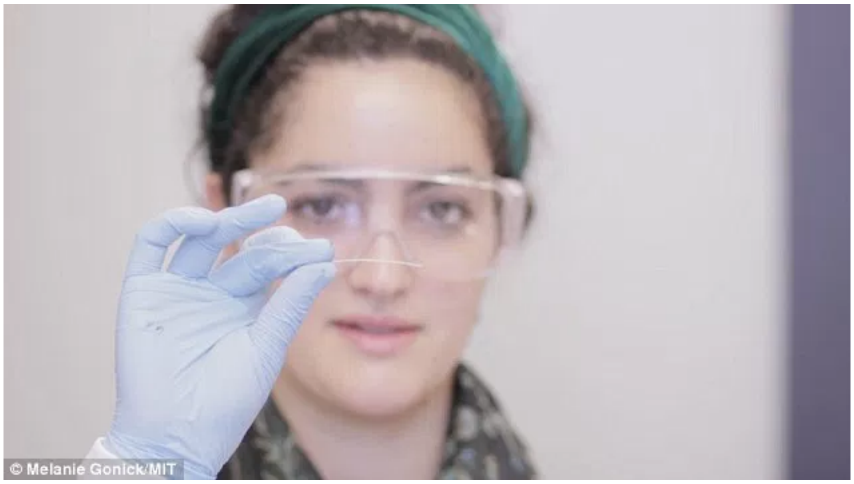

Christina Tringides, a senior at MIT and member of the research team, holds a sample of the multifunction fiber that could deliver optical signals and drugs directly into the brain, along with electrical readouts to continuously monitor the effects of the various inputs.

HOW IT WORKS

The new fibers are made of polymers that closely resemble the characteristics of neural tissues.

The multifunction fiber that could deliver optical signals and drugs directly into the brain, along with electrical readouts to continuously monitor the effects of the various inputs.

Combining the different channels could enable precision mapping of neural activity, and ultimately treatment of neurological disorders, that would not be possible with single-function neural probes.

'We're building neural interfaces that will interact with tissues in a more organic way than devices that have been used previously,' said MIT's Polina Anikeeva, an assistant professor of materials science and engineering.

The human brain's complexity makes it extremely challenging to study not only because of its sheer size, but also because of the variety of signaling methods it uses simultaneously.

Conventional neural probes are designed to record a single type of signaling, limiting the information that can be derived from the brain at any point in time.

Now researchers at MIT may have found a way to change that.

By producing complex fibers that could be less than the width of a hair, they have created a system that could deliver optical signals and drugs directly into the brain, along with simultaneous electrical readout to continuously monitor the effects of the various inputs.

RELATED ARTICLES

The newC technology is described in a paper in the journal Nature Biotechnology.

The new fibers are made of polymers that closely resemble the characteristics of neural tissues, Anikeeva says, allowing them to stay in the body much longer without harming the delicate tissues around them.

To do that, her team made use of novel fiber-fabrication technology pioneered by MIT professor of materials science Yoel Fink and his team, for use in photonics and other applications.

The result, Anikeeva explains, is the fabrication of polymer fibers 'that are soft and flexible and look more like natural nerves.'

Devices currently used for neural recording and stimulation, she says, are made of metals, semiconductors, and glass, and can damage nearby tissues during ordinary movement.

'It's a big problem in neural prosthetics,' Anikeeva says.

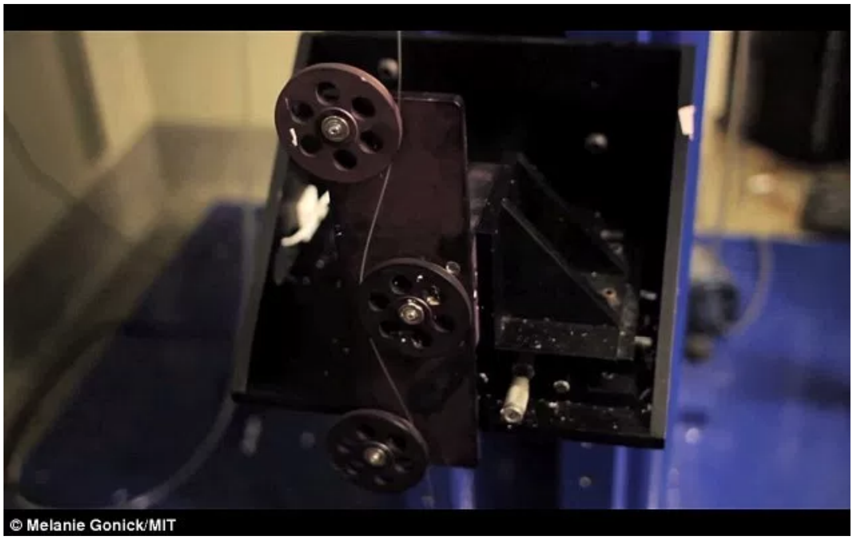
© Melanie Gonick/MIT

The result, Anikeeva explains, is the fabrication of polymer fibers 'that are soft and flexible and look more like natural nerves.

HOW IT WORKS

The new fibers are made of polymers that closely resemble the characteristics of neural tissues.

The multifunction fiber that could deliver optical signals and drugs directly into the brain, along with electrical readouts to continuously monitor the effects of the various inputs.

Combining the different channels could enable precision mapping of neural activity, and ultimately treatment of neurological disorders, that would not be possible with single-function neural probes.

'We're building neural interfaces that will interact…

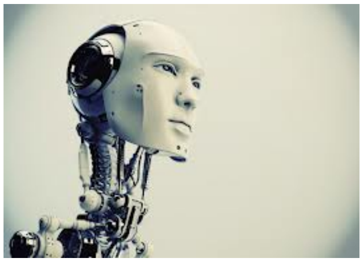

(https://nanobrainimplant.files.wordpress.com/2013/02/robot-brain.jpg)

- The result, Anikeeva explains, is the fabrication of polymer fibers ‘that are soft and flexible and look more like natural nerves.’

‘They are so stiff, so sharp — when you take a step and the brain moves with respect to the device, you end up scrambling the tissue.’

The key to the technology is making a larger-scale version, called a preform, of the desired arrangement of channels within the fiber: optical waveguides to carry light, hollow tubes to carry drugs, and conductive electrodes to carry electrical signals.

These polymer templates, which can have dimensions on the scale of inches, are then heated until they become soft, and drawn into a thin fiber, while retaining the exact arrangement of features within them.

A single draw of the fiber reduces the cross-section of the material 200-fold, and the process can be repeated, making the fibers thinner each time and approaching nanometer scale.

During this process, Anikeeva says, ‘Features that used to be inches across are now microns.’

Combining the different channels in a single fiber, she adds, could enable precision mapping of neural activity, and ultimately treatment of neurological disorders, that would not be possible with single-function neural probes.

For example, light could be transmitted through the optical channels to enable optogenetic neural stimulation, the effects of which could then be monitored with embedded electrodes.

© PUBLICITY PICTURE

Combining the different channels in a single fiber, she adds, could enable precision mapping of neural activity, and ultimately treatment of neurological disorders, that would not be possible with single-function neural probes.

At the same time, one or more drugs could be injected into the brain through the hollow channels, while electrical signals in the neurons are recorded to determine, in real time, exactly what effect the drugs are having.

MIT researchers discuss their novel implantable device that can deliver optical signals and drugs to the brain, without harming the brain tissue.

The system can be tailored for a specific research or therapeutic application by creating the exact combination of channels needed for that task. ‘You can have a really broad palette of devices,’ Anikeeva says.

While a single preform a few inches long can produce hundreds of feet of fiber, the materials must be carefully selected so they all soften at the same temperature.

The fibers could ultimately be used for precision mapping of the responses of different regions of the brain or spinal cord, Anikeeva says, and ultimately may also lead to long-lasting devices for treatment of conditions such as Parkinson's disease.

John Rogers, a professor of materials science and engineering and of chemistry at the University of Illinois at Urbana-Champaign who was not involved in this research, says, 'These authors describe a fascinating, diverse collection of multifunctional fibers, tailored for insertion into the brain where they can stimulate and record neural behaviors through electrical, optical, and fluidic means.

The results significantly expand the toolkit of techniques that will be essential to our development of a basic understanding of brain function.'

Read more: http://www.dailymail.co.uk/sciencetech/article-2927410/The-real-life-Matrix-MIT-researchers-reveal-interface-allow-computer-plug-brain.html#ixzz3Q9lDVQv4 (http://www.dailymail.co.uk/sciencetech/article-2927410/The-real-life-Matrix-MIT-researchers-reveal-interface-allow-computer-plug-brain.html#ixzz3Q9lDVQv4)

By mindcontroleurope • Posted in Computerbrain, Implant for Synthetic Telepathy, Mind Control • Tagged ai, artifical-intelligence, bioelectronics, brain, brain-computer-interface, human brain, mind control, mit, research, science, synthetic-telepathy
FEB 8 2014

Medical Nanobots Will Connect Brain to Cloud Computing – Ray Kurzweil

Thursday, February 6, 2014

Medical Nanobots Will Connect Brain to Cloud Computing – Ray Kurzweil (http://www.activistpost.com/2013/08/the-dna-nanobots-have-arrived.html)

(https://nanobrainimplant.files.wordpress.com/2013/05/mbj.jpg)

Update on the emergence of DNA nanobots and nan ocomputers.

Nicholas WestThe Human Body Version 2.0 (http://lifeboat.com/ex/human.body.version.2.0#nanobots) project features none other than arch-Transhumanist Ray Kurzweil (http://www.amazon.com/The-Singularity-Is-Near-Transcend/dp/0143037889/ref=as_li_tf_sw?&linkCode=wsw&tag=permacultucom-20) as its main proponent. The goals have been openly stated for some time:

> *In the coming decades, a radical upgrading of our body's physical and mental systems, already underway, will use nanobots to augment and ultimately replace our organs. We already know how to prevent most degenerative disease through nutrition and supplementation; this will be a bridge to the emerging biotechnology revolution, which in turn will be a bridge to the nanotechnology revolution. By 2030, reverse-engineering of the human brain will have been completed and nonbiological intelligence will merge with our biological brains.*

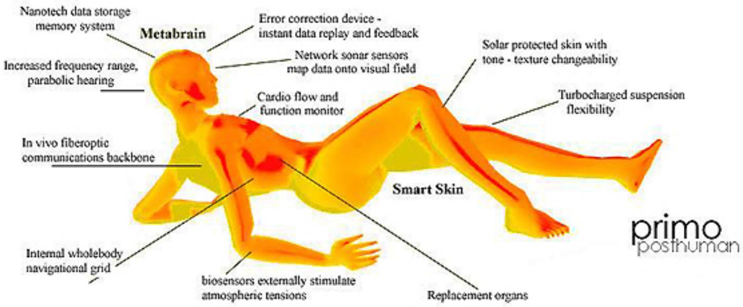
Nanotech data storage memory system
Metabrain
Error correction device - instant data replay and feedback
Increased frequency range, parabolic hearing
Network sonar sensors map data onto visual field
Solar protected skin with tone - texture changeability
Cardio flow and function monitor
Turbocharged suspension flexibility
In vivo fiberoptic communications backbone
Smart Skin
Internal wholebody navigational grid
biosensors externally stimulate atmospheric tensions
Replacement organs
primo posthuman

(https://nanobrainimplant.files.wordpress.com/2014/02/ef606-primo_posthuman.png)

Picture from Natasha Vita-More

Link to Vita-More: http://www.natasha.cc/ (http://www.natasha.cc/)

In fact, the reverse engineering of the human brain (http://www.activistpost.com/2013/08/hardware-and-software-combining-to.html) has already been announced to be well under way via new microchips and accompanying software. And, while full nanobot rewiring of the brain is not expected before 2020, Phys.org (http://phys.org/) has reported that our DNA has been successfully targeted by nanobots "for drug therapy or destruction."

Taking this even one step further, Ray Kurzweil said in a new interview with *The Wall Street Journal* (see below) that our extension into non-biological realms will include nanobot computers that will enter our brain and connect us to Cloud computing.

From science fiction horror, directly to the human body, the nanobots are no longer speculation. Also unlike science fiction, they won't arrive via immediate worldwide takeover — they are already here, and will be introduced incrementally, as Kurzweil has previously stated:

> *It will be an incremental process, one already well under way. Although version 2.0 is a grand project, ultimately resulting in the radical upgrading of all our physical and mental systems, we will implement it **one benign step at a time**. Based on our current knowledge, we can already touch and feel the means for accomplishing each aspect of this vision. (emphasis added)*

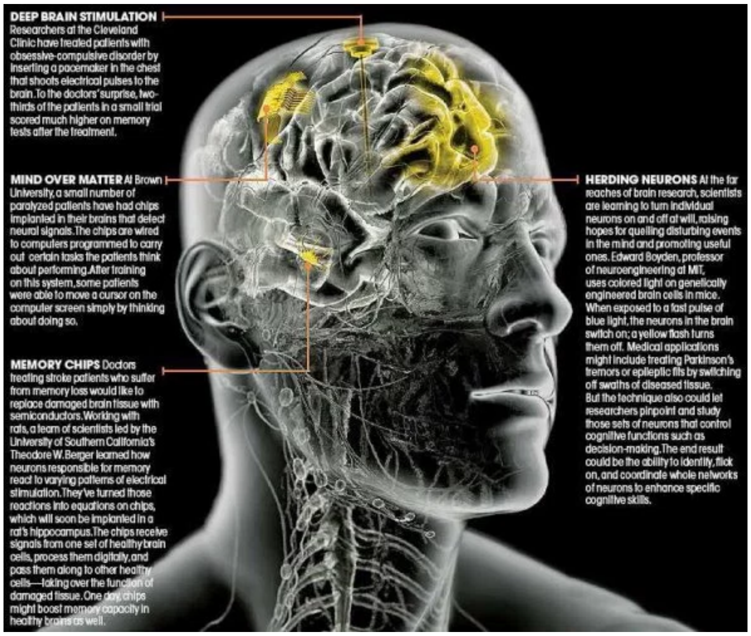

DEEP BRAIN STIMULATION Researchers at the Cleveland Clinic have treated patients with obsessive-compulsive disorder by inserting a pacemaker in the chest that shoots electrical pulses to the brain. To the doctors' surprise, two-thirds of the patients in a small trial scored much higher on memory tests after the treatment.

MIND OVER MATTER At Brown University, a small number of paralyzed patients have had chips implanted in their brains that detect neural signals. The chips are wired to computers programmed to carry out certain tasks the patients think about performing. After training on this system, some patients were able to move a cursor on the computer screen simply by thinking about doing so.

MEMORY CHIPS Doctors treating stroke patients who suffer from memory loss would like to replace damaged brain tissue with semiconductors. Working with rats, a team of scientists led by the University of Southern California's Theodore W. Berger learned how neurons responsible for memory react to varying patterns of electrical stimulation. They've turned those reactions into equations on chips, which will soon be implanted in a rat's hippocampus. The chips receive signals from one set of healthy brain cells, process them digitally, and pass them along to other healthy cells—taking over the function of damaged tissue. One day, chips might boost memory capacity in healthy brains as well.

HERDING NEURONS At the far reaches of brain research, scientists are learning to turn individual neurons on and off at will, raising hopes for quelling disturbing events in the mind and promoting useful ones. Edward Boyden, professor of neuroengineering at MIT, uses colored light on genetically engineered brain cells in mice. When exposed to a fast pulse of blue light, the neurons in the brain switch on; a yellow flash turns them off. Medical applications might include treating Parkinson's tremors or epileptic fits by switching off swaths of diseased tissue. But the technique also could let researchers pinpoint and study those sets of neurons that control cognitive functions such as decision-making. The end result could be the ability to identify, flick on, and coordinate whole networks of neurons to enhance specific cognitive skills.

(https://nanobrainimplant.files.wordpress.com/2014/01/mndcntrl-2.jpg)

Researchers from Columbia University have developed a fleet of molecular *nanorobots* that can deliver drugs to specific cells and also identify certain genetic markers by using fluorescent labeling. After such identification, a chain reaction can be initiated:

> *On cells where all three components are attached, a robot is functional and a fourth component (labeled 0 below) initiates a chain reaction among the DNA strands. Each component swaps a strand of DNA with another, until the end of the swap, when the last antibody obtains a strand of DNA that is fluorescently labeled.*
>
> *At the end of the chain reaction—which takes less than 15 minutes in a sample of human blood—only cells with the three surface proteins are labeled with the fluorescent marker.*

Naturally, this type of targeted therapeutic approach (http://medicalxpress.com/news/2013-07-molecular-robots-therapeutics.html#inlRlv) could prove beneficial, as the researchers highlight — especially for cancer treatment which sweeps up healthy cells along with malignant ones, very often doing more harm than good (if one were to choose the establishment medical route).

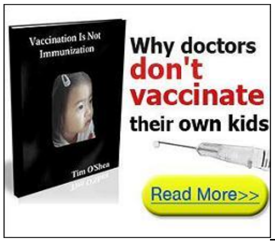
Vaccination Is Not Immunization
Tim O'Shea
Why doctors don't vaccinate their own kids
Read More>>

(http://activistpost.net/vaccinetruth.html)

This is always how new technologies are sold to the public, however, and it would be naive not to consider the darker applications as well.

Direct brain modification already has been packaged as “neuroengineering.” A Wired article (http://www.wired.com/science/discoveries/news/2009/03/neuroengineering1) from early 2009 highlighted that direct brain manipulation via fiber optics is a bit messy, but once installed “it could make someone happy with the press of a button.” Nanobots take the process to an automated level, rewiring the brain molecule by molecule. Worse, these mini droids can autonomously self-replicate, forcing one to wonder how this genie would ever be put back in the bottle once unleashed.

Here is one scenario offered by Kurzweil for how these nanobots could enter our bodies:

> *A significant benefit of nanobot technology is that unlike mere drugs and nutritional supplements, nanobots have* ***a measure of intelligence****. They can keep track of their own inventories, and intelligently slip in and out of our bodies in clever ways. One scenario is that we would wear a special “nutrient garment” such as a belt or undershirt. This garment would be loaded with nutrient bearing nanobots, which would make their way in and out of our bodies through the skin or other body cavities. (emphasis added)*

That might seem to offer a level of participatory choice — to wear or not to wear the garment — but Kurzweil reveals that the nanobots will eventually be everywhere:

> *Ultimately we won’t need to bother with special garments or explicit nutritional resources. Just as computation will eventually be ubiquitous and available everywhere, so too will basic metabolic nanobot resources be embedded everywhere in our environment.*

In a recent interview with *The Wall Street Journal*, Kurzweil highlights why Google has taken an interest in nanotechnology and the possibilities he sees for humans as they increasingly become non-biological and form direct connections with computers, augmenting and/or supplanting our natural processes as we head into the era of cyborgs (http://www.activistpost.com/2014/01/the-era-of-cyborgs-has-begun.html) and beyond:

And of course once our neocortex is uploaded to the Cloud, it positions Google perfectly for searching our every thought and pre-thought. While this might sound like an impossible amount of information to upload, let alone interconnect and search, it is being announced that researchers have designed the first nanocomputer that can push beyond the concept of Moore’s Law, which imposes a theoretical limitation on the expansion of computer processing power.

> *The team designed and assembled, from the bottom up, a functioning, ultra-tiny control computer that is the densest nanoelectronic system ever built.*
>
> *A technical paper has been published online (http://www.pnas.org/content/early/2014/01/23/1323818111.abstract) in the Proceedings of the National Academy of Sciences on the research.*
>
> *The ultra-small, ultra-low-power control processor—termed a* ***nanoelectronic finite-state machine (http://en.wikipedia.org/wiki/Finite-state_machine)*** *or “nanoFSM”—is smaller than a human nerve cell.*
>
> *[…]*
>
> *In their recent collaboration they combined several tiles on a single chip to produce a first-of-its-kind complex, programmable* ***nanocomputer****. (Source (http://www.33rdsquare.com/2014/02/nanocomputer-points-way-to-push-past.html))*

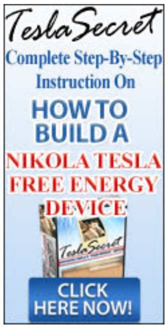
TeslaSecret
Complete Step-By-Step Instruction On
HOW TO BUILD A
NIKOLA TESLA FREE ENERGY DEVICE
CLICK HERE NOW!
(http://activistpost.net/teslasecret/)

It shouldn’t be seen as coincidence that these developments are happening simultaneously. What appears on the surface to be discoveries in entirely different fields are coalescing rapidly as we approach the theoretical date of The Singularity – the full merger of human and machine – estimated to occur between 2029-2045.

Despite the benign language of futurists, we know that a concerted effort is already underway to manage and predict human behavior (http://www.activistpost.com/2013/07/secret-darpa-mind-control-project.html) for a whole range of potentially anti-human applications (http://www.activistpost.com/2013/06/7-future-methods-of-mind-control.html). As our free will is also targeted like the cells of our body — for drug therapy or elimination — ethical concerns must be voiced loud and clear. Scientists seem content with opening Pandora’s Box, then worrying about negative consequences later … and that is only if we assume that their intentions are benign from the beginning. One should take time to examine the history of military experimentation on human populations (http://www.amazon.com/The-Medusa-File-Coverups-Government/dp/0963906240/ref=as_li_tf_sw?&linkCode=wsw&tag=permacultucom-20) to see all of this through a very different lens.

At the very least, instead of the fully realized vision of Human Body 2.0, this might be Big Pharma 2.0 — a new phase where conventional drugs are incrementally replaced by nanodrugs (http://www.sciencedaily.com/releases/2013/08/130810063717.htm) and nano-fleet delivery systems. Coupled with applications that directly enter our brain to connect us to the computer matrix, we are rapidly entering an entirely new human paradigm.

The funding is already there, and a massive amount of money is waiting to be made by companies like Google. Here again, for those who might only see the bright side to this technology, we ought to question who is really in control of it.

Sources:
http://phys.org/news/2013-08-dna-nanorobots-tag-cellular.html (http://phys.org/news/2013-08-dna-nanorobots-tag-cellular.html)
http://www.nature.com/nnano/focus/dna-nanotechnology/index.html (http://www.nature.com/nnano/focus/dna-nanotechnology/index.html)
http://www.nature.com/nnano/journal/v8/n8/index.html (http://www.nature.com/nnano/journal/v8/n8/index.html)
http://blogs.wsj.com/cio/2014/02/04/googles-ray-kurzweil-envisions-new-era-of-search/ (http://blogs.wsj.com/cio/2014/02/04/googles-ray-kurzweil-envisions-new-era-of-search/)

Recently by Nicholas West:

- Robots to Breed with Each Other and Humans by 2045 (http://www.activistpost.com/2014/01/robots-to-breed-with-each-other-and.html)
- Robots to Get Internet Cloud Brain: “Wikipedia For Robots” (http://www.activistpost.com/2014/01/robots-to-get-internet-cloud-brain.html)
- Amazon Patents The Use Of Predictive Behavior Technology For Future Shipments (http://www.activistpost.com/2014/01/amazon-patents-use-of-predictive.html)

By mindcontroleurope • Posted in Implant for Synthetic Telepathy, Man vs Machine, Mind Control • Tagged artifical-intelligence, artificial telepathy, biological brains, chemical stimulus, cloud-computing, DNA nanobots, medical nanobots, Nanobots, nonbiological intelligence, science, trans humanist, transcranial-magnetic-stimulation
JAN **17** 2014

Mind Control in the 21st Century

Mind Control in the 21st Century—Science Fiction and Beyond

[(http://www.veteranstoday.com/wp-content/uploads/2014/01/mind-control2_banner.jpg)](http://www.veteranstoday.com/wp-content/uploads/2014/01/mind-control2_banner.jpg)

Mind Control in the 21st Century—Science Fiction & Beyond

… by Steven DiBasio

Understanding
MIND
CONTROL
[(http://www.veteranstoday.com/wp-content/uploads/2014/01/51vrpnPx+jL.jpg)](http://www.veteranstoday.com/wp-content/uploads/2014/01/51vrpnPx+jL.jpg)

Is it coming to a home near you?

Conspiracy Theory?

"Mind control" is a topic commonly perceived as "conspiracy theory" or "X-Files" fare. That is, it is seen as possibly not "real," and certainly not something about which one should be "overly" concerned.

This attitude at least partially arises from the widespread belief or assumption that the human brain is so complicated—("the most complex entity in the universe" is a common formulation)—that it has not, and perhaps cannot, be comprehended in any depth.

One writer, for example, describes the brain as of "perhaps infinite" complexity,[1] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn1) while another, David Brooks of the *New York Times*, writes that it is "probably impossible" that "a map of brain activity" could reveal mental states such as emotions and desires.[2] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn2)

Similarly, Andrew Sullivan, blogger and former editor of *The New Republic*, opines that neuroscience is still in its "infancy," and that we have only begun "scratching the surface" of the human brain, and links to a *New Yorker* piece in support of that position.[3] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn3)

And the cover story for the October 2004 issue of Discovery Magazine entitled "The Myth Of Mind Control" advises the reader that while mind control is a "familiar science-fiction" staple, there is little reason for real concern, because actually deciphering the "neural code" would be akin to figuring out other "great scientific mysteries" such as the "origin of the universe and of life on Earth," and is therefore hardly likely.[4] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn4)

According to the article, as the brain is “the most significant mystery in science” and quite possibly “the hardest to solve,”[5] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn5) mind control remains at worst a distant concern.

The underlying idea seems to be that sophisticated mind control is unlikely without understanding the brain; and we do not understand the brain.

(http://www.veteranstoday.com/wp-content/uploads/2014/01/mindcontrol.gif)

Understanding the “Neural Code”

Of course, one might question the notion that a full understanding of the “neural code” is a prerequisite for mind control since it is not always necessary to know how something works for it to be effective. Nonetheless, the assumption that the brain is so complex that little progress has been made in “solving” it is itself incorrect.

As neuroscientist Michael Persinger has said, the “great mythology” of the brain is that it is “beyond our understanding; no it’s not.”[6] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn6) In fact, according to inventor and “futurist” Ray Kurzweil, “very detailed mathematical models of several dozen regions of the human brain and how they work….”[7] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn7) had already been developed over a decade ago.

Kurzweil also said at that time that science is “further along in understanding the principles of operation of the human brain than most people realize….”[8] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn8) While the brain may be complicated, “it’s not *that* complicated (emphasis added).”[9] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn9)

Similarly, an Air Force report from 1995, in a section entitled “Biological Process Control,” predicts that before 2050 “… [w]e will have achieved a clear understanding of how the human brain works, how it really controls the various functions of the body, and how it can be manipulated…:”[10] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn10)

> ***One can envision the development of electromagnetic energy sources … that can couple with the human body in a fashion that will allow one to prevent voluntary muscular movements, control emotions (and thus actions), produce sleep, transmit suggestions, interfere with … memory, produce an experience set, and delete an experience set. [11] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn11)***

As disturbing as such “predictions” may be, is it possible that technologies to prevent (or perhaps even impel) muscular movement, control emotions, transmit suggestions, delete memories, create false memories, and so on, have already been developed?

Certainly, even a cursory review of the “open literature” reveals that various sophisticated mind control technologies already exist.[12] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn12) Indeed, it is rather shocking to realize how advanced mind control technology was, even several decades ago.

(http://www.veteranstoday.com/wp-content/uploads/2014/01/PTSD1.jpg)

“Altering brain waves”

For example, there is the 1974 invention of Robert G. Malech for which a patent was granted in 1976 and assigned to defense contractor Dorne & Margolin, Inc.—for a method of “remotely monitoring and altering brain waves.”[13] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn13)

Moreover, experiments conducted over thirty years ago at the Stanford Research Institute (SRI) showed that basic mind reading from EEG readouts was possible, revealing the existence of “a non-symbolic language” of “brain-wave patterns” which could be deciphered and translated.[14] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn14)

Indeed, "…[b]y the late 1960s … 'remote control' of the human brain—accomplished *without* the implantation of electrodes—was well on its way to being realized."[15] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn15) A means of stimulating a brain "by creating an electrical field completely outside the head" was developed,[16] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn16) and it was discovered that electric pulses could stimulate the brain using far less energy than previously "thought … effectual in the old implanting technique."[17] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn17)

Not surprisingly, with such developments arose legitimate fears of a future world where "human robots" would perform the bidding of the "military."[18] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn18)

And one source quotes a 1970s Pentagon agency report as saying that it will likely be possible in "several years" to induce sounds and words directly into the brain (bypassing the ears), as well as to use "combinations of frequencies and other signal characteristics to produce other neurological effects….,"[19] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn19)

The report notes that the Soviets had observed "various changes in body chemistry" and "functioning" of the brain from the exposure of the brain to various frequencies.[20] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn20) Also mentioned are studies at MIT showing that "magnetic brain waves can be picked up … and amplified as if the brain were a radio transmitter," no implants or electrodes required.[21] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn21)

Finally, an article from 1981 describes how "microwave generators" placed in appropriate locations and transmitting at low energy would create "interference patterns" out of the interaction with brainwaves (brain electricity).

These interference patterns "could then be built up by computer into a three-dimensional moving picture of mental processes"—in other words, a remote "thought scanner" (and tracking device) could be developed.[22] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn22)

Recent "Advances"

(http://www.veteranstoday.com/wp-content/uploads/2014/01/crazy.jpg)

Subjected to outside influences

In light of these past developments, it is perhaps rather surprising to read modern articles describing supposedly recent innovations in "mind reading" and mind control technology – in which it is sometimes claimed, for example, that scanners, electrodes and proximity to the subject is required to read and "control" minds.

Such claims reflect an apparent failure of the science of "mind control" to progress as one might have expected considering the presumed interest, as well as the spectacular rate of advancement of science and technology in general in recent decades.

Of course, it would not be all that surprising if mind control technology has advanced considerably, but that research has been carried out in secret for reasons of "national security."

CIA affiliated scientists have certainly conducted much research which they have been prohibited from sharing with their peers,[23] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn23) and inventions that implicate "national security" are routinely suppressed under Pentagon secrecy orders.[24] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn24) Also, it might seem desirable to hide research programs which sometimes "require" relaxation of ethical standards, such as that of informed consent.

That said, even ignoring the likely existence of a "secret science" of mind control, recent public advancements are quite troubling in their own right.

Some examples:

6S
6S
22
2E
3M
3M
24
#130

(http://www.veteranstoday.com/wp-content/uploads/2014/01/th.jpg)

"New connections" are being made all the time.

1) In 2004, 25,000 rat neurons on a glass dish learned to fly an F-22 jet fighter simulator.[25] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn25) After scientists placed the neurons on the dish, the neurons quickly began "to reconnect themselves, forming a living neural network—a brain."[26] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn26)

The lead scientist added that "one day," though of course a "long way off," disembodied brains might actually be used to fly drones,[27] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn27) though the current experiment was merely to enhance knowledge of how the brain works, and possibly provide "clues to brain dysfunction."[28] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn28)

2) In August 2013, researchers revealed that "miniature" human brains had been grown in the laboratory.[29] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn29) As is typical, any negative implications or reasons for worry were minimized, while possible "therapeutic" uses were highlighted. Thus, the breakthrough was hailed as a great opportunity to understand "developmental defects." Though the writer does mention "the spectre of what the future might hold," the reader is reassured that the research is "primitive territory"[30] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn30)—though one researcher did comment on the "undesirability" of growing larger laboratory brains.[31] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn31)

3) On July 1, 2013, a magazine reported a claim by neuroscientist Sergio Canavero that it was now feasible to transplant the head of one human to the body of another and reattach the spinal cord.[32] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn32)

4) Scientists have reconstructed random images viewed by subjects, from fMRI brain scans, in research that "hints" that "one day" scientists might be able to "access dreams, memories and imagery…."[33] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn33)

5) The brains of two rats have been linked, such that one, located in North Carolina, responded "telepathically" to the thoughts of the other, located in Brazil.[34] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn34) The second rat's brain processed signals from the first rat's brain, delivered over the internet, as if they were its own. The scientist speculated about the "future possibility" of a "biological computer, in which numerous brains are connected…."[35] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn35)

6) A brain-to-brain interface has been created, allowing humans to move a rat's tail just by thinking about it.[36] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn36) Readers are told that while it is not yet possible to "communicate brain to brain with our fellow humans … we may be on our way to … controlling" other species.[37] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn37) But, since it is "still very early days" the writer "hope(s)" that any ethical concerns can be "iron(ed) out."[38] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn38) Of note, the study used focused ultrasound to deliver impulses to the rat's brain.[39] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn39)

7) Continuing the ultrasound "theme": Focused pulses of low intensity low frequency ultrasound, transmitted noninvasively through the skull to the human brain, have been shown capable of producing, not only pain, but also sound, as well as evoking "sensory stimuli."[40] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn40) Accordingly, a lab with a "close working relationship" with DARPA, the Department of Defense, and U.S. Intelligence communities, has been looking into using pulsed ultrasound to encode "sensory data onto the cortex"; in other words, producing hallucinations through the remote and direct stimulation of brain circuits.[41] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn41) Possibilities are the ability to "remotely control brain activity" and the "creation of artificial memories."[42] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn42) Even Sony has gotten in on the act, patenting a device for using ultrasound to produce hallucinations—again described as "transmitting sensory data directly into the human brain."[43] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn43) Most troublingly, one source recently alleged that the NSA is using this ultrasound technology to target individuals through their smartphones.[44] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn44)

8) A researcher was able to make a fellow researcher in a different office move his finger just by thinking about it, in the "first" demonstration of a human brain-to-brain interface.[45] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn45)

9) A low cost means of tracking people, even through walls, has been developed. While in the past individuals could be tracked anywhere by the "military" using radar technology, this technology might enable entities with fewer resources to track people as well.[46] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn46)

(http://www.veteranstoday.com/wp-content/uploads/2014/01/mindcontrol_nano-testing_crop.jpg)

10) Scientists have remote controlled a worm by implanting magnetic nanoparticles into it, and then exposing the animal to a "radiofrequency magnetic field" which stimulated its neurons. The scientists suggest that their research could lead to "innovative cancer treatments" and "improved diabetes therapies," as well as

11) Americans can now be spied on in their homes through their internet-connected appliances, according to (former) CIA Director David Petraeus.

Petraeus made his statements at about the same time a huge microchip company, ARM, unveiled new processors which will connect home appliances such as refrigerators, washers and driers to the internet.[48] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn48)

12) LED lights have been ostensibly pushed for their efficiency over traditional bulbs. However, LED lights are also semiconductors capable of inducing "biological and behavior effects."[49] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn49)

"Breakaway" Science?

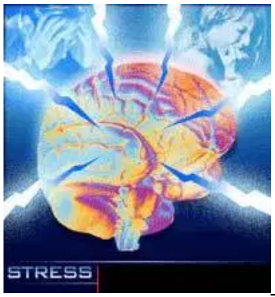

(http://www.veteranstoday.com/wp-content/uploads/2014/01/stress-510010.jpg)

Nural Codes

While the aforementioned public developments are quite concerning, the reality is they may not actually represent the true state of the art in "mind control" technology.

It would not be that surprising, after all, for a domain with national security implications to at some point in its development branch off onto separate "tracks," one public and the other "hidden."

If such a bifurcation were to occur, advancements made in secret would not necessarily be incorporated into the public sphere. Eventually perhaps, innovations and breakthroughs would result in the development of an essentially new, covert science.

An example of a domain in which this bifurcation process seems to have occurred is aviation. In the public sphere, the most advanced aircraft might well be the F-22 fighter jet, or perhaps the F-35. However, if insider testimony is credited, these aircraft seem almost primitive in comparison with flying machines developed in secret.

Perhaps the most compelling statements in this regard come from Ben Rich, former Director of Lockheed-Martin's Advanced Development Projects, or "Skunk Works," a Lockheed division notable for its super high-tech, top secret projects, among them the U2 spy plane and the SR-71 Blackbird.

As Joseph P. Farrell's reports in his book *Saucers, Swastikas, and Psyops*, Rich made a number of peculiar and provocative comments at the end of his career, and following his retirement on December 31, 1990 (prior to his death five years later), comments strongly hinting at "the development of … an off-the-books physics and technology…."[50] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn50)

For example, on September 7, 1988, in a presentation to the American Institute of Aeronautics and Astronautics in Atlanta, Georgia, Rich lamented that he was prohibited from discussing Skunkwork's current projects, but he did say that they "call for technologies once only dreamed of by science fiction writers."[51] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn51)

In ensuing years, Rich elaborated slightly. For instance, while speaking to the UCLA School of Engineering Alumni Association in 1993, Rich said that "an error in the equations" had been discovered and corrected, making it possible "to travel to the stars."[52] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn52) He added, however, that "these technologies are so locked up in black programs, that it would take an act of God to ever get them out to benefit humanity."[53] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn53)

100 YEARS OF ACCELERATING TOMORROW

(http://www.veteranstoday.com/wp-content/uploads/2014/01/Ben-Rich-Ftr.jpg)

Ben Rich – Who saw it it…lived it all.

Farrell goes on to relay a statement from an unnamed Lockheed retired engineer who was quoted in a magazine article in 1988 as saying that "we have things flying in the Nevada desert that would make George Lucas drool."[54] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn54) In the same article an Air Force officer involved in the development of the SR-71 said "[w]e are testing vehicles that defy description.

To compare them conceptually to the SR-71 would be like comparing Leonardo da Vinci's parachute design to the space shuttle." And a retired Colonel chimed in: "We have things that are so far beyond the comprehension of the average aviation authority as to be really alien to our way of thinking." [55] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn55)

> ***Consider then for a moment the possibility that within the classified world, in 1993, a technology, to quote Ben Rich, "to take ET back home" had already been developed.[56] (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/#_ftn56) The implications are enormous, not to mention rather frightening. One wonders where the technology must be in 2014, more than twenty years later.***

And if the aforementioned statements are true, and this seems plausible (why would these individuals lie, or even exaggerate, especially to Engineering Associations and Aeronautics institutes), what might this imply about the current state of the art in domains other than aviation, such as neuroscience, which has itself been the subject of intense "weaponization" efforts.

> ***Indeed, what does such a vast discrepancy between what people believe and what is actually true suggest about the nature of our perceived reality in general?***

Steven DiBasio is a writer, attorney, and sometime musician. He lives in the Midwest. And more will be on the way. He can be reached at: steven.dibasio@gmail.com (mailto:steven.dibasio@gmail.com)

Editing: Jim W. Dea (http://www.veteranstoday.com/staff-writers/)

Original: http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/ (http://www.veteranstoday.com/2014/01/13/mind-control-in-the-21st-century-science-fiction-beyond/)

By mindcontroleurope • Posted in Implant for Synthetic Telepathy, military, Mind Control • Tagged artifical-intelligence, artificial telepathy, brain-computer-interface, microwave, military, mind control, nsa, radio-engineering, transcranial-magnetic-stimulation
JAN 5 2014

NSA Spy Program Is Absolutely Orwellian Intelligence

The EMF Brain Mapping Of The People And Its Use As Part Of A Signals Intelligence Driven Spy Program Is Absolutely Orwellian

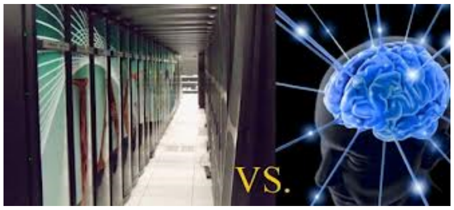

(https://nanobrainimplant.files.wordpress.com/2014/01/44.png)

Former US National Security Agency contractor Edward Snowden appeared in a televised Christmas message released to the world public by TV station Channel 4 on Wednesday. (Dec. 25)

Investigative Journalists Targeted For An Ongoing **FBI COINTELPRO STING OPERATION/SMEAR CAMPAIGN** — been used as an unwitting target of (MKULTRA) non consensual human experimentation for decades via the **NSA's SIGNALS INTELLIGENCE EMF Scanning Network – a covert spy program which uses brain scanners deployed via NSA signals intelligence satellites to remotely scan the brainwaves of any citizen in the world.** The NSA's SIGNIT EMF Scanning Network is an outgrowth of the Pentagon and CIA's MIND and TAMI (MKULTRA) "mind control" programs, which use EEG Heterodyning technology to synchronize AI computers with the unique brainwave print of each citizen. This technology enables the NSA to brand us like heads of cattle. Google: AKWEI VS NSA & The Matrix Deciphered by Dr. Robert Duncan to learn more about this Orwellian attack on the people's Constitutional rights.

By mindcontroleurope • Posted in Man vs Machine, Mind Control • Tagged artificial intelligence, artificial telepathy mind, b, Brain Implants for MIND/BRAIN CONTROL, brain-computer-interface, Implant for Synthetic Telepathy, military, nsa, orwell, snowden, transcranial-magnetic-stimulation

Powered by WordPress.com.

WAVE GENOME (/)

HEALING VIA WAVE PHARMACEUTICALS
AND QUANTUM BIOCOMPUTING

HOME / ГЛАВНАЯ (/)

SERVICES / УСЛУГИ (/SERVICES--105910571051105910431048.HTML)

PRODUCTS / ТОВАРЫ (/PRODUCTS--105810541042104010561067.HTML)

CONTACT / КОНТАКТ (/CONTACT--1050105410531058104010501058.HTML)

(/tesla-generator.html) TESLA PSI-GENERATOR

(/order-tesla-selenium.html) TESLA SELENIUM

(/tesla-water.html) TESLA WATER

(/order-tesla-glasses.html) TESLA GLASSES

(/digital-pharmaceuticals--- DIGITAL PHARMACY

(/order-tesla-screen.html) TESLA SCREEN

IRENE CAESAR, PH.D.

President & Founder

WAVE GENOME LLC

465 Buckland Hills Drive

Unit 33211, Manchester, CT 06042

1-646-480-0496

ORDER YOUR MINI-TESLA GENERATOR

(/tesla-generator.html)

WAVE GENOME IS A FRONTIER OF THE TECHNOLOGICAL REVOLUTION

WAVE GENOME IS AN AMERICAN COMPANY LOCATED IN CONNECTICUT - A PROUD PIONEER OF PRODUCING WAVE MATRICES / BIOHOLOGRAMS BY LASERS FOR THE LOCAL AND REMOTE MANAGEMENT OF BIOSYSTEMS AND HUMAN CONDITION

WAVE GENOME PRECEDED GOOGLE'S BIOELECTRONIC DRUGS R&D PROGRAM FOR 6 YEARS OF ACTUAL SALES AND 50 YEARS BEYOND R&D.

WAVE GENOME HAS INTELLECTUAL PRIORITY FOR THE WAVE OPTICS IN CHROMOSOMES AS BIOHOLOGRAPHY (/diaschev-dipol-di-pol.html)

(http://www.facebook.com/pages/Wave-Genome/203765456304318?fref=ts)

(/wave-genome--russian.html)

Нажмите на Русский флаг для перемещения на версию сайта на Русском языке

WG ВОЛНОВОЙ ГЕНОМ

WAVE PHARMACEUTICALS

(/wave-pharmaceuticals.html)

VISIT OUR CLINIC FOR CURE AND REJUVENATION

(/quantum-clinic.html)

TESLA SELENIUM IS THE ONLY ORGANIC AND BIVALENT SELENIUM IN THE WORLD
TESLA SELENIUM PRO
BY WAVE GENOME LLC
TESLA SELENIUM
BY WAVE GENOME LLC

(/order-tesla-selenium.html)

USED BY MILITARY PILOTS AND ASTRONAUTS
ORDER TESLA GLASSES
EYE SIGHT RESTORATION AND TREATMENT OF PHYSIOLOGICAL SYSTEMS VIA POLARIZED BIOHOLOGRAPHY

(/order-tesla-glasses.html)

Read about TESLA GLASSES here (/order-tesla-glasses.html).
Order TESLA GLASSES here (/order-form.html).

EIGHT YEARS OF SATISFIED CUSTOMERS

WE OFFER ADVANCED WEATHER CONTROL. BUT WHAT CAN BE NOBLER THAN SAVING LIVES! (/quantum-clinic.html)

BIOWAVE GENOME HAS DISTRIBUTORS ALL OVER THE WORLD:

Germany, Sweden, Poland, United States, Russia, Thailand

Click here (/diaschev-dipol-di-pol.html) to read on how WAVE GENOME's technology differs from Dr. Peter P. Gariaev's and Dr. Alexander D. Panchenko's technology (DIPOL or DI-POL Generator, sold by Dr. Diaschev).

EMBRYONIC STEM CELL BIOHOLOGRAM
EXCLUSIVELY
BY WAVE GENOME LLC

(/second-birth-mini-tesla-psi-generator.html)

GLOBAL LAUNCH OF THE SECOND-BIRTH MINI-TESLA PSI-GENERATOR BY WAVE GENOME (/second-birth-mini-tesla-psi-generator.html)

JANUARY 13, 2018

Wave Genome is proud to announce its most advanced MINI-TESLA PSI-GENERATOR ever: the SECOND-BIRTH Mini-Tesla Psi-Generator.

We called the new chip the "SECOND BIRTH" generator (/second-birth-mini-tesla-psi-generator.html), because the client gets the most valuable bioholographic treatment ever for rejuvenation -- the biohologram of the embryonic stem cell. The SECOND-BIRTH Biohologram (/second-birth-mini-tesla-psi-generator.html) is available as a chip encoded by laser on nanolevel. The clinical trials were conducted by the Bauman University Medical Center. The clinical trials had demonstrated the radical improvement in the wellbeing of the participants in the clinical trials.

Click here to read more. (/second-birth-mini-tesla-psi-generator.html)

WAVE GENOME LLC uses Bauman Moscow Russian State Technical University, major Science University of Russia, of MIT level, to verify WAVE GENOME'S technological models.

WORLDWIDE LAUNCH OF HOLONET (© Dr. Irene Caesar) - HOLOGRAPHIC BIOINTERNET & HOLOGRAPCHIC DRUGS, BIOELECTRONIC DRUGS AND INTERNET DRUGS

H-DRUGS, E-DRUGS AND I-DRUGS

FOR REMOTE AND LOCAL REJUVENATION

NEWS

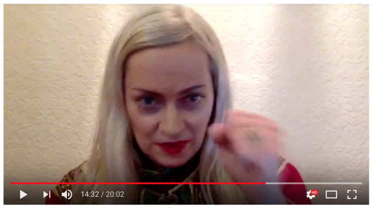

IRENE CAESAR INTERVIEW FOR MINDTECH ENTERPRISES DEC 13 2016

IRENE CAESAR - INTERVIEW FOR MINDTECH ENTERPRISES DEC 13 2016 (https://www.youtube.com/watch?v=ap8kHpuYlxM&feature=youtu.be)

IRENE CAESAR INTERVIEW FOR RIPSONAR NEWS APRIL 22 2017
54:42 / 56:35

IRENE CAESAR, Ph.D., PRESIDENT AND OWNER OF WAVE GENOME: INTERVIEW FOR RIPSONAR NEWS, APRIL 22, 2017 (https://www.youtube.com/watch?v=iTTHLETrz1A).

(/quantum-clinic.html)
Large-size AURA PSI-GENERATOR for radical rejuvenation of all bodily organs and physiological systems.

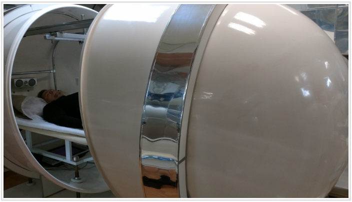

(/quantum-clinic.html)
Large-size AURA PSI-GENERATOR offers radical rejuvenation for patents rejected by conventional medicine to the degree they do not need conventional medicine any longer.

Click here (/quantum-clinic.html) to read about our clinic.

Consortium "MATRIX CITY - city as one integrated wave matrix" (/consortium.html)

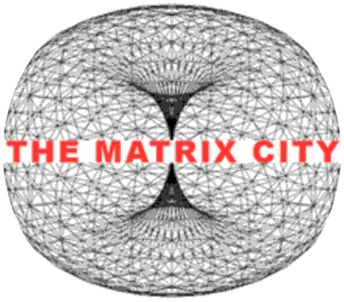

(/irene-caesar---matrix-city-project.html)

QUANTUM REMOTE REJUVENATION

(/quantum-rejuvenation-clinic-om---english.html)

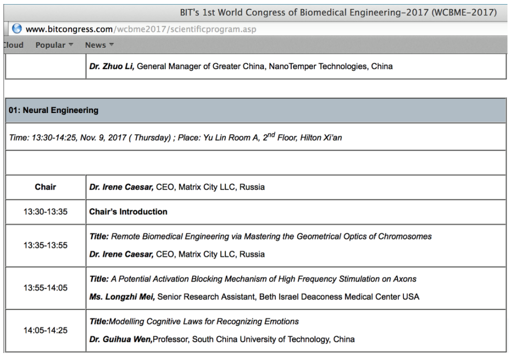

BIT's 1st World Congress of Biomedical Engineering-2017 (WCBME-2017)

www.bitcongress.com/wcbme2017/scientificprogram.asp

	Dr. Zhuo Li, General Manager of Greater China, NanoTemper Technologies, China

01: Neural Engineering

Time: 13:30-14:25, Nov. 9, 2017 (Thursday) ; Place: Yu Lin Room A, 2nd Floor, Hilton Xi'an

Chair	***Dr. Irene Caesar,*** CEO, Matrix City LLC, Russia
13:30-13:35	**Chair's Introduction**
13:35-13:55	***Title:*** *Remote Biomedical Engineering via Mastering the Geometrical Optics of Chromosomes* ***Dr. Irene Caesar,*** CEO, Matrix City LLC, Russia
13:55-14:05	***Title:*** *A Potential Activation Blocking Mechanism of High Frequency Stimulation on Axons* ***Ms. Longzhi Mei,*** Senior Research Assistant, Beth Israel Deaconess Medical Center USA
14:05-14:25	***Title:****Modelling Cognitive Laws for Recognizing Emotions* ***Dr. Guihua Wen,***Professor, South China University of Technology, China

NOVEMBER 9-11, 2017

DR. IRENE CAESAR WAS INVITED TO BE A CHAIR OF THE NEURAL ENGINEERING SECTION OF THE BIT'S 1st WORLD COGRESS OF BIOMEDICAL ENGINEERING, Theme: Co-creating a New Future of Biomedicine, November 9-11, 2017, in Xi'an, China. I will deliver paper called **"REMOTE BIOMEDICAL ENGINEERING VIA MASTERING THE GEOMETRICAL OPTICS OF CHROMOSOMES"**. List of Renowned Speakers include professors from University of Oxford, UK; Harvard Medical School, USA; American College Of Clinical Engineering, USA; California South University, USA; University of Illinois at Chicago, USA; The Mount Sinai Medical Center, USA; University of California, USA; Washington University School of Medicine, USA; The Salerno Center for Complementary Medicine, USA; National Cancer Institute, NIH, USA; University of Pennsylvania, USA; Washington University in St. Louis Campus, USA; International Federation For Medical and Biological Engineering, USA; Louisiana State University Health Sciences Center in Shreveport, USA; International Spine Research Foundation, USA; RWTH Aachen University, Germany; Technical University of Munich, Germany; Institut Curie, France; University of Queensland, Australia; Tsinghua University, China; Asia Pacific Association for Medical Informatics, South Korea; European Medical Association, Belgium; University of Geneva, Switzerland; Gyeongbuk Institute of Science and Technology, South Korea; University of Geneva, Switzerland; The Standing Committee of European Doctors, Belgium; International Scociety of Medical Laser Applications, Austria.

Link to the NEURAL ENGINEERING Session:
http://www.bitcongress.com/wcbme2017/scientificprogram.asp

List of Renowned Speakers:
http://www.bitcongress.com/wcbme2017/default.asp

"REMOTE BIOMEDICAL ENGINEERING VIA MASTERING THE GEOMETRICAL OPTICS OF CHROMOSOMES"

Abstract

The same gene is expressed in functional individuals and species by metacentric chromosome, and in dysfunctional individuals and species via acrocentric chromosome. Metacentric chromosome is analogous to a well-centered and focused eye crystal, while acrocentric chromosome is analogous to a myopic or farsighted eye. Chromosome is a lens that allows our cells to focus the needed wave information during cell division, analogously to a crystal in our eye. Thus, Biomedical Engineering is possible only via mastering the Geometrical Optics of Chromosomes by Laser Bioholography. Current genetic engineering via cutting and pasting DNA snippets destroys the systematic wholes of chromosomal bioholograms and the refraction of the geometrical optics in chromosomes, and, therefore, causes genetic collapse. The Quantum Leap in the Biomedical Engineering is the cost of our survival. Biomedical Engineering as Bioholography is based on Quantum Physics, and understanding the non-local nature of the Bioholograms. Since universe is holographic, we can record and transmit only holographic information, based upon changing the refraction of the scalar wave diffraction grating in chromosomes. According to Holographic principle, Universe is entirely in its every matrix point. The implication of the Holographic Principle states that if Universe is entirely in its every matrix point, then, every matrix point is not simply different from any other matrix point, but is unique. This means that

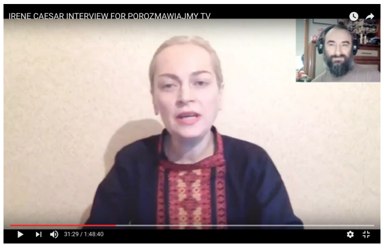

(https://www.youtube.com/watch?v=uxfui6xTtEc)

IRENE CAESAR, Ph.D., PRESIDENT AND OWNER OF WAVE GENOME: INTERVIEW FOR POROZMAWIAJMY TV CONDUCTED BY ALEXANDER BERDOWICZ, FEBRUARY 7, 2017. (https://www.youtube.com/watch?v=uxfui6xTtEc)

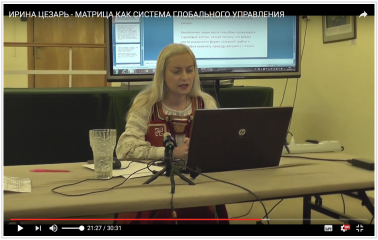

IRENE CAESAR, Ph.D., PRESIDENT AND OWNER OF WAVE GENOME: LECTURE "MATRIX AS A GLOBAL MANAGEMENT SYSTEM OF LIFE. MECHANISM OF REMOTE MANAGEMENT OF BIOSYSTEMS," CONFERENCE "PSYCHICS - YESTERDAY, TODAY, AND TOMORROW, (4TH MEETING OF THE ASSOCIATION OF ASTRAL ANTHROPOLOGY AND PSYCHOLOGY), MOSCOW, RUSSIA, OCTOBER 10, 2016. (https://www.youtube.com/watch?v=eEKzZQ4Wzcc)

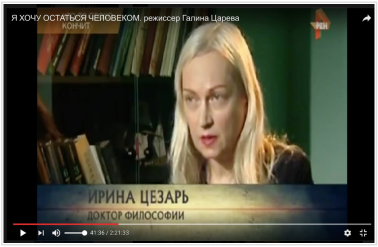

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME, IN THE FILM BY RENOWNED RUSSIAN PRODUCER GALINA TSAREVA "I WANT TO REMAIN HUMAN", AUGUST, 2016. (https://www.youtube.com/watch?v=YTTWaIRT_0Y)

(https://www.youtube.com/watch?v=XuwV_QCTf2I)

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME: "REMOTE MANAGEMENT OF BIIO SYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR

Models of TESLA PSI-GENERATORS by WAVE GENOME.

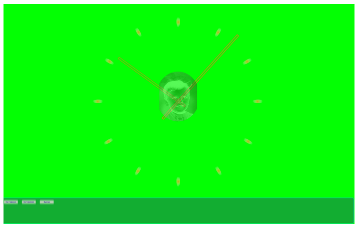

TRUE BIOELECTRONIC E-DRUGS

The MINI-TESLA psi-generator **DIGITAL PHARMACY (/digital-pharmacy-software--1057105410601058-10621048106010561054104210541049-104010551058104510501048.html)** software program generates HOLOGRAPHIC SIGNAL (as white noise) encoded in music for every system and vital organ in your body.

MILESTONES

Over the span of the 20th century, generations of extraordinary scientists, among whom Russian scientists played a leading role, laid down scientific foundation for the technological revolution of information-wave technology of the Quantum Leap.

MORE (/milestones.html)

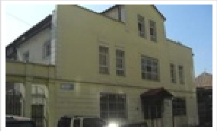

(/uploads/3/5/1/1/35112207/5553543_orig.jpg)

For developing and producing its products, WAVE GENOME is teaming with most advanced scientists all over the world. Here we see Scientific Research Institute for National Security in Moscow, now closed. The scientific group still

every wave matrix is both non-local, and unique. Thus, if we have two exact copies of the same non-local wave matrix, we can transmit information instantaneously and remotely, from one copy of the same non-local unique wave matrix to another copy. Thus, the precision in the Biomedical Engineering is achieved not via the precision of localization, but, vice versa, via the precision of refraction in the unique and non-local bioholograms aka “wave crystals”. The secret in the remote Biomedical Engineering as Quantum Biocomputing lies in the recording and instantaneous transmission of a structural phantom. I will introduce the new method for recording and remote transmitting of bioholograms via laser spectroscopy and coding electrets on nanolevel via laser. The technology creates “lenses” assisting our chromosomes in correcting their refraction for focusing genetic bioholographic information, both locally and remotely.

APRIL 25-27, 2017

IRENE CAESAR, PH.D., SPEECH "REMOTE MANAGEMENT OF BIOSYSTEMS AS BIOHOLOGRAPHY BASED UPON THE SHAPE OF CHROMOSOMES FOR FOCUSING GENETIC INFORMATION" at the Annual World Congress of Molecular & Cell Biology Xi'an, China. **Click here** (http://www.bitcongress.com/cmcb2017/)for the official webpage of the Summit. **Click here (http://www.bitcongress.com/cmcb2017/scientificprogram_4.asp)** for the official list of participants.

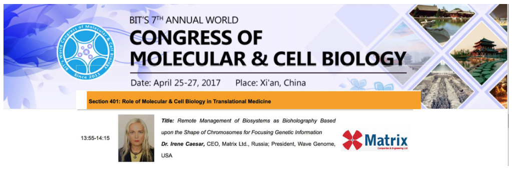

BIT'S 7TH ANNUAL WORLD
CONGRESS OF
MOLECULAR & CELL BIOLOGY
Date: April 25-27, 2017 Place: Xi'an, China

Section 401: Role of Molecular & Cell Biology in Translational Medicine

13:55-14:15 *Title: Remote Management of Biosystems as Bioholography Based upon the Shape of Chromosomes for Focusing Genetic Information*
Dr. Irene Caesar, CEO, Matrix Ltd., Russia; President, Wave Genome, USA

BIT's World Congress of Molecular & Cell Biology
Since 2011
Matrix

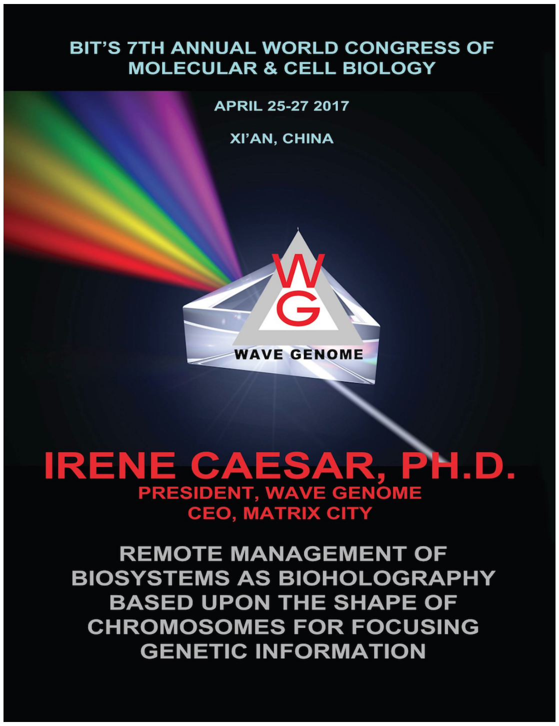

BIT'S 7TH ANNUAL WORLD CONGRESS OF MOLECULAR & CELL BIOLOGY
APRIL 25-27 2017
XI'AN, CHINA

IRENE CAESAR, PH.D.
PRESIDENT, WAVE GENOME
CEO, MATRIX CITY

REMOTE MANAGEMENT OF BIOSYSTEMS AS BIOHOLOGRAPHY BASED UPON THE SHAPE OF CHROMOSOMES FOR FOCUSING GENETIC INFORMATION

WG
WAVE GENOME

Keynote speakers: Dr. Christian Paterman, Former Director, European Commission, Germany; Dr. Martin Banwell, Professor, Australian National University, Australia; Dr. Robert S. Plumb, Professor, Imperial College London, UK; Dr. Dongping Zhong, Professor, Ohio State University, USA.

FOCUSING GENETIC INFORMATION" - RECORDING OF THE LECTURE, BANGKOK THONBURI UNIVERSITY 2ND INTERNATIONAL SYMPOSIUM, JULY 23, 2016.

(https://www.youtube.com/watch?v=XuwV_QCTf2I)

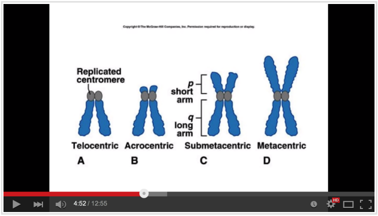

(https://www.youtube.com/watch?v=pSsgOPUzz9A)

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME: "REMOTE MANAGEMENT OF BIIO SYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION" - SLIDE SHOW OF THE LECTURE, BANGKOK THONBURI UNIVERSITY 2ND INTERNATIONAL SYMPOSIUM, JULY 23, 2016.

(https://www.youtube.com/watch?v=pSsgOPUzz9A)

IRENE CAESAR - GOSPEL OF IRENE
15:52 / 39:01

IRENE CAESAR, PH.D., FOUNDER, OWNER AND PRESIDENT OF WAVE GENOME: LECTURE "GOSPEL OF IRENE" (PAPER "GOSPEL OF IRENE" WAS PUBLISHED IN FULL IN THREE ISSUES OF THE SOCIALIST FACTOR MAGAZINE, LUCKNOW AND LONDON), THE STRATEGIC INFORMATICS ADVISORY, MARYLAND, USA, JULY 7, 2016.

(https://www.youtube.com/watch?v=QFEErBqSJjY&feature=youtu.be)

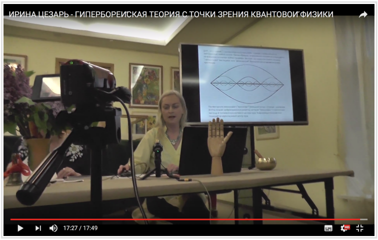

ИРИНА ЦЕЗАРЬ - ГИПЕРБОРЕЙСКАЯ ТЕОРИЯ С ТОЧКИ ЗРЕНИЯ КВАНТОВОЙ ФИЗИКИ
17:27 / 17:49

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME: "HYPERBOREAN THEORY FROM THE POINT OF VIEW OF QUANTUM PHYSICS" - PAPER DELIVERED AT THE CONFERENCE "ESOTERICS, NEW TURNAROUND", ASSOCIATION OF ANTHROPOLOGY AND PSYCHOLOGY, MOSCOW, MAY 16, 2016. (https://www.youtube.com/watch?v=CRbRsohA8i4)

operates the clinic. Among clients of the clinic were Russian President Boris Yeltsin, famed American astronaut Edgar Mitchell, and actor celebrity Alexander Kalyagin. The institute has received the written Blessing of Patriarch Alexis and Patriarch of Constantinople.

(/uploads/3/5/1/1/35412207/4353994_orig.jpg) Famed American Astronaut Edgar Mitchell received a treatment by GENERATOR AURA-001.

OUR PLEDGE

We build our strategy upon a firm foundation of the highest standards of ethics.

MORE (/our-pledge.html)

Our MINI-TESLA PSI-GENERATOR is used by Russian Special Forces.

Our MINI-TESLA PSI-GENERATOR is used by Russian Olympic Team.

Our TESLA GLASSES are used by Russian Central Intelligence Service.

Our TESLA GLASSES are used by Russian Space Forces.

Re
Biosyste
upon t
for Foc

Abstract

We are standing at the threshold of the biggest scientific revolution in the history of human kind, called Quantum Leap. All Transhumanist models of genetic modification of organisms, relying on global artificial intellect and brain-computer interfaces are obsolete technologies of the pre-Quantum Era. Either Biology makes Quantum Leap, or humankind will be destroyed by genetic collapse. Human enhancement is possible only through mastering the geometrical optics of chromosomes via Bioholography.

2017第八届国际DNA和基因组活动周
BIT's 8th Annual World DNA and Genome Day-2017
NO. 03
BIT'S 7TH ANNUAL WORLD CONGRESS OF MOLECULAR & CELL BIOLOGY
APRIL 25-27 2017
XI'AN, CHINA
WAVE GENOME
IRENE CAESAR, PH.D.
PRESIDENT, WAVE GENOME
CEO, MATRIX CITY
REMOTE MANAGEMENT OF BIOSYSTEMS AS BIOHOLOGRAPHY BASED UPON THE SHAPE OF CHROMOSOMES FOR FOCUSING GENETIC INFORMATION
NO. 04

Now bioengineering fails since it does not account for the geometrical optics of chromosomes, leading to the corruption of genomes. The same gene gets expressed in dysfunctional individuals by acrocentric chromosome, and in functional individuals by metacentric chromosome. I claim that chromosome is a geometrical code and functions via refraction in the wave crystal, i.e., nonlocal unique biohologram or wave matrix that focuses genetic information unique to this individual via his / her own unique scalar wave diffraction grating. The implication of the holographic principle states that if the universe is entirely in its every matrix point, then, every matrix point is not simply different from any other matrix point, but is unique. Thus, if we have two exact copies of the same unique non-local biohologram, we can transmit information remotely and instantaneously from one copy to another copy for the remote management of biosystems.

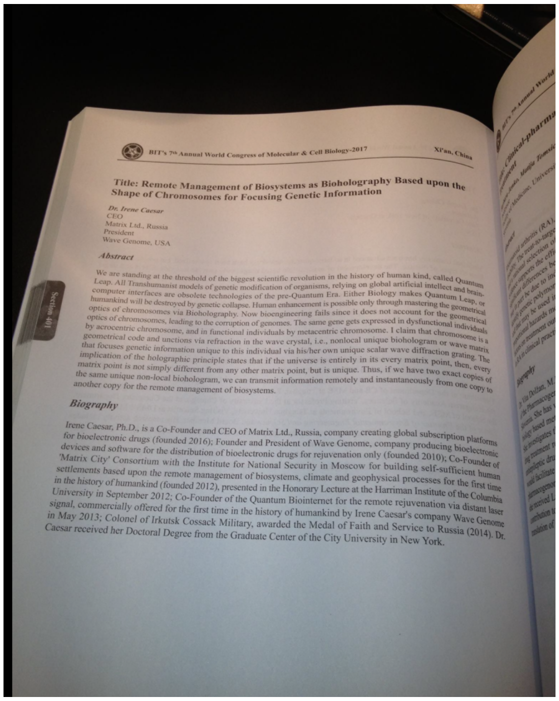

BIT's 7th Annual World Congress of Molecular & Cell Biology-2017 Xi'an, China

Title: Remote Management of Biosystems as Bioholography Based upon the Shape of Chromosomes for Focusing Genetic Information

Dr. Irene Caesar
CEO
Matrix Ltd., Russia
President
Wave Genome, USA

Abstract

We are standing at the threshold of the biggest scientific revolution in the history of human kind, called Quantum Leap. All Transhumanist models of genetic modification of organisms, relying on global artificial intellect and brain-computer interfaces are obsolete technologies of the pre-Quantum Era. Either Biology makes Quantum Leap, or humankind will be destroyed by genetic collapse. Human enhancement is possible only through mastering the geometrical optics of chromosomes via Bioholography. Now bioengineering fails since it does not account for the geometrical optics of chromosomes, leading to the corruption of genomes. The same gene gets expressed in dysfunctional individuals by acrocentric chromosome, and in functional individuals by metacentric chromosome. I claim that chromosome is a geometrical code and unctions via refraction in the wave crystal, i.e., nonlocal unique biohologram or wave matrix that focuses genetic information unique to this individual via his/her own unique scalar wave diffraction grating. The implication of the holographic principle states that if the universe is entirely in its every matrix point, then, every matrix point is not simply different from any other matrix point, but is unique. Thus, if we have two exact copies of the same unique non-local biohologram, we can transmit information remotely and instantaneously from one copy to another copy for the remote management of biosystems.

Biography

Irene Caesar, Ph.D., is a Co-Founder and CEO of Matrix Ltd., Russia, company creating global subscription platforms for bioelectronic drugs (founded 2016); Founder and President of Wave Genome, company producing bioelectronic devices and software for the distribution of bioelectronic drugs for rejuvenation only (founded 2010); Co-Founder of 'Matrix City' Consortium with the Institute for National Security in Moscow for building self-sufficient human settlements based upon the remote management of biosystems, climate and geophysical processes for the first time in the history of humankind (founded 2012), presented in the Honorary Lecture at the Harriman Institute of the Columbia University in September 2012; Co-Founder of the Quantum Biointernet for the remote rejuvenation via distant laser signal, commercially offered for the first time in the history of humankind by Irene Caesar's company Wave Genome in May 2013; Colonel of Irkutsk Cossack Military, awarded the Medal of Faith and Service to Russia (2014). Dr. Caesar received her Doctoral Degree from the Graduate Center of the City University in New York.

Section 401

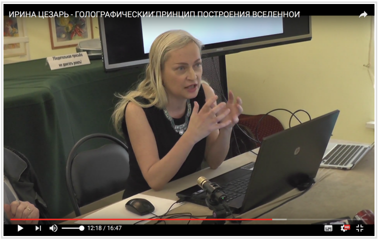

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME: "HOLOGRAPHIC PRINCIPLE OF THE STRUCURE OF THE UNIVERSE" - PAPER DELIVERED AT THE CONFERENCE "UNKNOWN COSMOS", ASSOCIATION "PROTOHISTORY", MOSCOW, MAY 14, 2016. (https://www.youtube.com/watch?v=M4oYWLLmP4c)

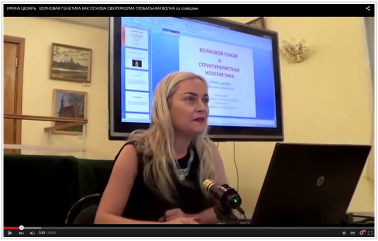

(https://www.youtube.com/watch?v=RT1nSUJilbY)

IRENE CAESAR, PH.D., PRESIDENT AND OWNER WAVE GENOME: "STRUCTURALIST WAVE GENOME AS THE FOUNDATION FOR SUPER INTELLECT" - PAPER DELIVERED AT THE CONFERENCE OF THE ASSOCIATION FOR ANTHROPOLOGY AND PSYCHOLOGY, MOSCOW, APRIL 26, 2016. (https://www.youtube.com/watch?v=RT1nSUJilbY)

Another version of this video recording. (http://www.wavegenome.com http://www.thematrixcity.com)

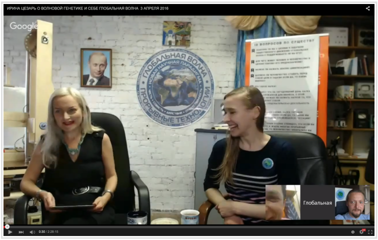

(http://Skip navigation RUUpload Search Analytics Video Manager ИРИНА ЦЕЗАРЬ О ВОЛНОВОЙ ГЕНЕТИКЕ И СЕБЕ ГЛОБАЛЬНАЯ ВОЛНА 3 АПРЕЛЯ 2016 Irene Caesar Irene Caesar Channel settingsDownload720 Add to Share More 82 views Published on Jun 7, 2016 ИРИНА ЦЕЗАРЬ О ВОЛНОВОЙ ГЕНЕТИКЕ И СЕБЕ ГЛОБАЛЬНАЯ ВОЛНА 3 АПРЕЛЯ 2016 SHOW MORE Comments are disabled for this video. Autoplay Up Next Сергей Салль Ответ еврейке Цезарь - Салль позор России игра в бисер 12,851 views 50:28 ИРИНА ЦЕЗАРЬ - РАВВИН ЭТО РУСО-АРИЙ ГЛОБАЛЬНАЯ ВОЛНА Irene Caesar 97 views 2:52:41 Засланный казачок Ирина Цезарь или очередной проект сатаноизбранных (Сергей Салль) Светослав 4,499 views 50:28 ИРИНА ЦЕЗАРЬ - СТРУКТУРНАЯ ВОЛНОВАЯ ГЕНЕТИКА КАК ОСНОВА СВЕРХРАЗУМА Irene Caesar 362 views 18:05 СПИЦЫН - Золотая Орда. Дмитрий Донской. Нашествия ханов. Куликовская битва. Монголы и Русь. Евгений Спицын ИСТОРИЯ РОССИИ Recommended for you 57:16 ИРИНА ЦЕЗАРЬ

Our TESLA GLASSES are used by Russian Strategic Missile Forces.

Our TESLA GLASSES are used by Russian Ministry of Emergency Situations.

Our TESLA GLASSES are used by Russian Federal Security Service.

Our TESLA GLASSES are used by Russian Interior Ministry.

WAVE GENOME employs scientists from most advanced research group all over the world, including the present Moscow Bauman Russian State Technical University, the major Science University in Russia, of MIT level; and scientists of the former most secret USSR Research Institutes, which conducted studies in psychotronics, e.g., "Scientific-Research Institute Energy" (founded in 1937, with 5000 scientists under supervision of the National Security Agencies of the Soviet Union (Научно- производственное Объединение Энергия). The Scientific-Research Institute Energy was developing devices and services for the first persons in the State, military commanders, diplomats, Special Forces and the Olympic Team. Technologies of the Institute were declassified in 2008. In the US, these technologies are still classified. Dr. Irene Caesar was also collaborating with the Scientific Research Institute for the National Security in Moscow, Russian Federation, now closed.

Nonetheless, it was Dr. Irene Caesar herself who was first to offer to record the holographic signal of the wave matrices via laser from the pharmaceutical drugs as recreated bioholograms representing in this dimension the scalar wave diffraction gratings of multidimensional wave crystals. And Dr. Caesar was first to offer an idea to encode electrets via laser on nanolevel with these bioholograms of the pharmaceutical drugs.

Dr. Caesar created a theory of the Wave Optics (Geometrical Optics) in chromosomes as refraction in the scalar wave diffraction grating.

GLOBAL LAUNCH OF ULTIMATE TESLA PSYCHOTRONIC GENERATOR BY WAVE GENOME

Click here (/ultimate-mini-tesla-psi-generator.html) to read **PRESS RELEASE** on the **GLOBAL LAUNCH OF THE ULTIMATE TESLA PSI-GENERATOR by WAVE GENOME**

ULTIMATE TESLA PSI-GENERATOR IS BASED UPON IMPORTANT DISCOVERY BY DR. IRENE CAESAR REGARDING THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING / REFRACTION MECHANISM FOR FOCUSING GENETIC INFORMATION

IRENE CAESAR & MAGNUS OLSSON ULTIMATE TESLA LAUNCH BY WAVE GENOME
14:36 / 15:22

(https://www.youtube.com/watch?v=lyfh-C3oSm0&feature=youtu.be)

ON DECEMBER 2, 2016, IRENE CAESAR, PH.D.. PRESIDENT, FOUNDER AND OWNER OF WAVE GENOME, LAUNCHED HER NEW PRODUCT -- ULTIMATE MINI-TESLA PSI-GENERATOR FOR THE ULTIMATE PROTECTION AGAINST PSYCHOTRONIC AND GENETIC WARFARE. Intellectual and Commerical Property belongs to WAVE GENOME. Magnus Olsson, President of Mindtech Enterprises, becomes the designated Distributor for WAVE GENOME'S ULTIMATE MINI-TESLA PSI-GENERATOR, available on wavegenome.com (http://www.wavegenome.com) and mindtechenterprises.com (http://www.mindtechenterprises.com). **Click here (https://www.youtube.com/watch?v=lyfh-C3oSm0&feature=youtu.be)** to see the official video of the global launch of THE ULTIMATE MINI-TESLA PSI-GENERATOR by WAVE GENOME.

- ДОКАЗАТЕЛЬСТВА ДРЕВНОСТИ РУСОВ
Irene Caesar 162 views 1:09:36 America)

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME - NTERVIEW ABOUT WAVE GENOME ON GLOBAL WAVE, MOSCOW, APRIL 3, 2016
(https://www.youtube.com/watch?v=VNLtnfLCOlk&feature=youtu.be)

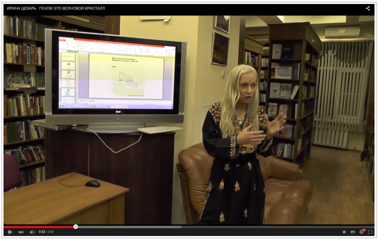

(https://www.youtube.com/watch?v=tgF8PCbRJCQ&feature=youtu.be)

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME: "GENOME IS THE REFRACTION OF SIGNALS TOWARDS THE FOCUS IN THE WAVE CRYSTAL OF CHROMOSOMES" - GUEST PUBLIC LECTURE "DNA GENEALOGY IN THE LIFHT OF RUSSIAN IDEA", EXTRACT, RESEARCH CENTER ON THE ETHICS OF THE RUSSIAN WORLD, MOSCOW, MARCH 16, 2016
(https://www.youtube.com/watch?v=tgF8PCbRJCQ&feature=youtu.be)

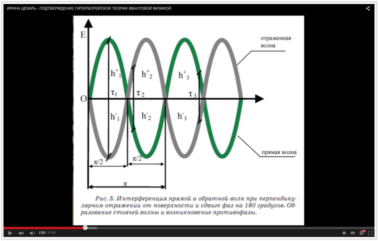

(https://www.youtube.com/watch?v=nPgyM2d0HQg&feature=youtu.be)

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME: "CONFIRMATION OF THE HYPERBOREAN THEORY BY QUANTUM PHYSICS" - GUEST PUBLIC LECTURE "DNA GENEALOGY IN THE LIFHT OF RUSSIAN IDEA", EXTRACT, RESEARCH CENTER ON THE ETHICS OF THE RUSSIAN WORLD, MARCH 16, 2016
(https://www.youtube.com/watch?v=nPgyM2d0HQg&feature=youtu.be)

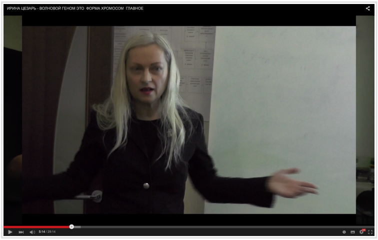

(https://www.youtube.com/watch?v=ONPUDcRVgHY&feature=youtu.be)

IRENE CAESAR, PH.D, PRESIDENT AND OWNER OF WAVE GENOME: "SHAPE OF CHROMOSOMES AS A KEY TO WAVE GENOME - GUEST PUBLIC LECTURE, CHETEK ("MAN, TECHNOLOGY, CAPITAL"), MOSCOW, JANUARY 27, 2016

Dr. Caesar was first to offer an idea of DIGITAL PHARMACY as the collection not of frequencies, but of bioholograms.

WAVE GENOME is focused on developing and commercializing **QUANTUM BIOCOMPUTING** as the **INFORMATION-WAVE TECHNOLOGY OF THE LASER BIOHOLOGRAPHY** in the fields of rejuvenation, wellness and agriculture.

THE INFORMATION-WAVE TECHNOLOGY OF LASER BIOHOLOGRAPHY is based upon the application of Quantum Physics to Genetics, a new theory of Genetics, which is called **STRUCTURALIST GENETICS** (© Irene Caesar, Ph.D., 1989-2010), and the application of various lasers, including the stabilized He-Ne lasers with internal mirrors.

Specifically, this technology applies:
- the holographic model of the Quantum Vacuum identical for macro and micro dimensions;
- the holographic torsion model of the cell as a torus;
- polarized biohoholography and polarized bioholograms, based upon the principles of geometrical optics;
- the notions of two kinds of torus: torus as a boundless sphere and torus as a wave crystal (© Irene Caesar, Ph.D., 1989);
- the Kozyrev Mirror principle for the reflection of the external signal upon itself, and creating the scalar wave diffraction grating (wave crystal);
- the diffraction grating (wave crystal) for the refraction of the external signal towards the zero center;
- the notion of the zero center of the wave crystal as the information portal;
- the notion of the instantaneous and remote transmission of information through the zero center of the wave crystal via quantum entanglement in the paradigm of the **QUANTUM BIOINTERNET**.

INFORMATION-WAVE TECHNOLOGY OF THE LASER BIOHOLORAPHY is radically different from the Digital Homeopathy and Bioresonance Technology, though it is also based upon the wave-particle duality in physics. Cell dies when it goes into bioresonance with itself. Soviets had the psychotronic weapon, based upon the principle of bioresonance. Thus, the bioresonance technologies in the West are still wide spread simply because they cannot do enough harm due to incompetence. The information is received and transmitted not through waves (frequencies). The information is received and transmitted through the **GEOMETRICAL CODES** of the wave crystals.

Also, the **INFORMATION-WAVE TECHNOLOGY OF THE LASER BIOHOLORAPHY** is not the scalar wave technology, since scalar waves neutralize any external signal. And to call the **INFORMATION-WAVE TECHNOLOGY OF THE LASER BIOHOLOGRAPHY** as the "scalar wave technology" is to say nothing. Scalar waves are only one element of this technology. Information is not received and transmitted via scalar waves. Scalar waves are simply used for creating the diffraction grating inside the wave crystal in order to refract the external signal toward the zero center of the wave crystal. Information is received and

GLOBAL LAUNCH OF
TESLA GLASSES
BY WAVE GENOME

(/order-tesla-glasses.html)

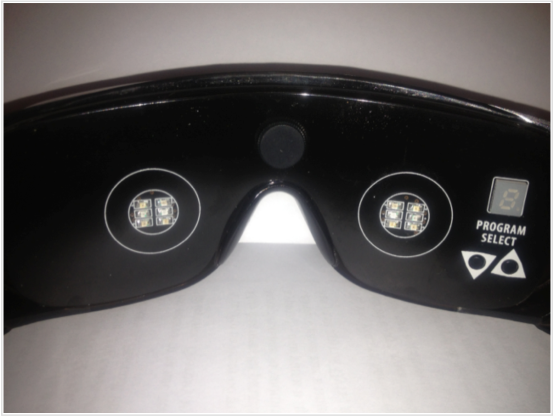

(/order-tesla-glasses.html)

Click here (/order-tesla-glasses.html) to read about WAVE GENOME TESLA GLASSES.

IRENE CAESAR - HOW TESLA PSI GENERATOR WORKS
9:45 / 11:42

(https://www.youtube.com/watch?v=Z8Otz0gbqeM)

DECEMBER 2, 2016

IRENE CAESAR, PH..D, PRESIDENT, FOUNDER AND OWNER OF WAVE GENOME: "HOW THE ULTIMATE TESLA PSI-GENERATOR WORKS", Global Launch of the ULTIMATE TESLA PSI-GENERATOR by WAVE GENOME. **Click here (https://www.youtube.com/watch?v=Z8Otz0gbqeM)** for the video.

IRENE CAESAR ON WAVE GENOME PRODUCTS
1:12 / 9:40

(https://www.youtube.com/watch?v=H8hkdkMMkMg)

DECEMBER 2, 2016

IRENE CAESAR, PH..D, PRESIDENT, FOUNDER AND OWNER OF WAVE GENOME ON WAVE GENOME PRODUCTS, Global Launch of the ULTIMATE TESLA PSI-GENERATOR by WAVE GENOME. **Click here** (https://www.youtube.com/watch?v=H8hkdkMMkMg) for the video.

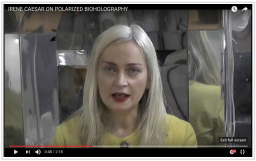

(https://www.youtube.com/watch?v=6BpWw4R8XBo)

DECEMBER 2, 2016

IRENE CAESAR, PH.D., FOUNDER AND PRESIDENT OF WAVE GENOME, ON POLARIZED BIOHOLOGRAPHY. Click here (https://www.youtube.com/watch?v=6BpWw4R8XBo&feature=youtu.be) for the video.

IRENE CAESAR - VOICE TO SKULL OF MAGNUS OLSSON
Irene Caesar, Ph.D., President and Founder of

(https://www.youtube.com/watch?v=2QgxLxAbkWs&feature=youtu.be)

DECEMBER 2, 2016

IRENE CAESAR, PH.D., FOUNDER AND PRESIDENT OF WAVE GENOME: "WAVE GENOME FOR VICTIMS OF VOICE TO SKULL" (WAVE GENOME helps V2K victim Magnus Olsson). **Click here** (https://www.youtube.com/watch?v=2QgxLxAbkWs&feature=youtu.be)for the video.

(https://www.youtube.com/watch?v=ONPUDcRVgHY&feature=youtu.be)

IRENE CAESAR WACHENHEIM SEMINAR FEB 18 2015 PART 1
venation information-wave technologies of the Institute, the Colonel of the Interna
1:04 / 30:25

(https://www.youtube.com/watch?v=2BCiU31THbQ)

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME: "WAVE GENOME - LOCAL AND REMOTE BIOELECRONIC DRUGS AS FRONTIER OF QUANTUM LEAP" - LECTURE, HOSTED BY PRAXIS DR. GRUENAGEL, TRANSLATED BY KATIA MEIER, MD (OWNER OF THE CLEAR SKY MEDICAL, PC, IN DENVER, CO), WACHENHEIM, GERMANY, FEBRUARY 18, 2015, PART 1 (https://www.youtube.com/watch?v=2BCiU31THbQ)

PART 2 (https://www.youtube.com/watch?v=2a9dDLfCE5g)

PART 3 (https://www.youtube.com/watch?v=Ntg-IJFC974)

IRENE CAESAR OF WAVE GENOME EVOLVE EXPO PART 2
Katia Meier, MD
Irene Caesar, PhD
0:30 / 21:14

(https://www.youtube.com/watch?v=PasGu7KjhyQ)

IRENE CAESAR, PH.D., PRESIDENT AND OWNER OF WAVE GENOME: "QUANTUM LEAP VIA WAVE GENOME" - LECTURE, AS PART OF WAVE GENOME'S PARTICIPATION IN EVOLVE EXPO 2015, EVOLVE EXPO 2015, THE NATIONAL WESTERN COMPLEX, DENVER, COLORADO, USA, FEBRUARY 7-8, 2015 (https://www.youtube.com/watch?v=PasGu7KjhyQ)

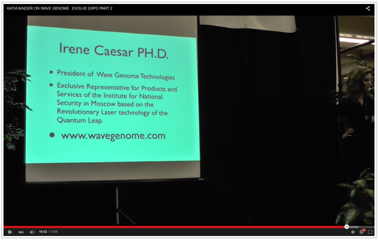

Irene Caesar PH.D.
- President of Wave Genome Technologies
- Exclusive Representative for Products and Services of the Institute for National Security in Moscow based on the Revolutionary Laser technology of the Quantum Leap.
- www.wavegenome.com

(https://www.youtube.com/watch?v=xWYkXpzVJPo)

KATIA MEIER, MD, REPRESENTATIVE OF WAVE GENOME IN COLORADO, OWNER OF CLEAR SKY MEDICAL, PC, ON HER REPRESENTATION OF WAVE GENOME - LECTURE AS PART OF WAVE GENOME'S PARTICIPATION IN EVOLVE EXPO 2015, EVOLVE EXPO, THE NATIONAL WESTERN COMPLEX, DENVER, COLORADO, USA, MARCH 7-8, 2015 (https://www.youtube.com/watch?v=xWYkXpzVJPo)

transmitted **ONLY** through the zero center of the biohologram / wave matrix (wave crystal). The information is transmitted from one identical token to another identical token of the same unique non-local wave matrix **REMOTELY**
and **INSTANTANEOUSLY**
via **QUANTUM INTERNET**.

WAVE GENOME LLC employs only professional and experienced specialists to administer rejuvenation treatments in its partner clinics and produce its products, including the **TESLA WATER (/tesla-water.html)** and **MINI-TESLA GENERATOR (/tesla-generator.html)**. Majority of **WAVE GENOME LLC** specialists are both MD and Ph.D. and are well-known doctors all over the world.

WAVE GENOME'S LLC scientists have conducted the first in the world experiment for the remote lowering of the blood sugar in a diabetic cow at a distance of 20 kilometers in the mid 1990s. **WAVE GENOME LLC** has eight years of benefiting global market with its services and products, which are patented, licensed, and certified.

2016
健康2.0亚洲会议
Health 2.0 Asia
智慧医疗与大健康产业高峰论坛
International Summit on Smarter Healthcare & Great Health Industry
Irene Caesar
俄罗斯Matrix公司首席执行官；美
Genome公司董事长
Irene Caesar
CEO, Matrix Ltd., Russia; Pre
Wave Genome, USA

Irene Caesar, Ph.D., Speech "Remote Management of Biosystems Based upon the Shape of Chromosomes as Diffraction Grating for Focusing Genetic Information", International Symposium HEALTH 2.0, Nanjing, China, November 9-10, 2016.

Irene Caesar
俄罗斯Matrix公司首席执行官；美国Wave
Genome公司董事长
Irene Caesar
CEO, Matrix Ltd., Russia; President,
Wave Genome, USA

Irene Caesar, Ph.D., Speech "Remote Management of Biosystems Based upon the Shape of Chromosomes as Diffraction Grating for Focusing Genetic Information", International Symposium HEALTH 2.0, Nanjing, China, November 9-10, 2016.

IRENE CAESAR AND MAGNUS OLSSON WAVE GENOME HELP TO V2K VICTIMS
Евгений Гигаури posted in

(https://www.youtube.com/watch?v=zQgSz2_PdG0&spfreload=10)

DECEMBER 2, 2016**. IRENE CAESAR & MAGNUS OLSSON WAVE GENOME HELP TO V2K VICTIMS**. Click here (https://www.youtube.com/watch?v=zQgSz2_PdG0) for the video.

IRENE CAESAR ON QUANTUM LEAP OF WAVE GENOME
0:37 / 8:48

(https://www.youtube.com/watch?v=sgttqrKAFlg)

DECEMBER 2, 2016**. IRENE CAESAR ON THE QUANTUM LEAP OF WAVE GENOME.** Click here (https://www.youtube.com/watch?v=sgttqrKAFlg) for the video.

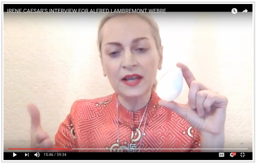
IRENE CAESAR'S INTERVIEW FOR ALFRED LAMBREMONT WEBRE

(https://newsinsideout.com/2017/01/dr-irene-caesar-reveals-russian-quantum-leaps-new-science-next-evolution-health-now-available-online-mindtechenterprises-com/)

JANUARY 2, 2017

ARTICLE BY ALFRED LAMBREMONT WEBRE ON DR. IRENE CAESAR AND HER COMPANY WAVE GENOME: "Dr. Irene Caesar reveals Russian Quantum Leap's new science – Our next evolution in health, now available online at MindTechEnterprises.com and wavegenome.com", News Insideout. **Click here** (https://newsinsideout.com/2017/01/dr-irene-caesar-reveals-russian-quantum-leaps-new-science-next-evolution-health-now-available-online-mindtechenterprises-com/) to read the article.

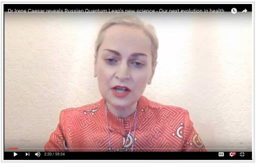
Dr Irene Caesar reveals Russian Quantum Leap's new science - Our next evolution in health

(https://www.youtube.com/watch?v=OwHafg-VJNI)

DECEMBER 17, 2016

IRENE CAESAR, PH.D., INTERVIEW TO ALFRED LAMBREMONT WEBRE: "Dr. Irene Caesar reveals Russian Quantum Leap's new science – Our next evolution in health, now available online at MindTechEnterprises", By Alfred Lambremont Webre. With the participation of MindTech Enterprises Team -- distributors of products of Irene Caesar's company Wave Genome, Magnus Thorwald Olsson and Madlen Namro. Click here to see the video (https://www.youtube.com/watch?v=2T2VRsJNE34).

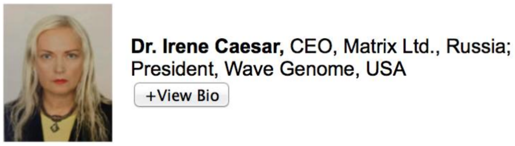

Dr. Irene Caesar, CEO, Matrix Ltd., Russia; President, Wave Genome, USA

+View Bio

IRENE CAESAR SEMINAR AT ORANIA ZENTRUM, GERMANY
26:30 / 32:54

IRENE CAESAR'S, PH.D., SEMINAR WELLENGENOM @ ORANIA-ZENTRUM, GERMANY, KURNACH BEI WURZBURG, FILM BY EXTREME NEWS, NOVEMBER 21-23, 2014 (https://www.youtube.com/watch?v=PHPBOCkdmiQ)

(http://www.wavegenome.com/1/post/2013/08/leakpro
the-singularity-irene-caesar-interviews-roman-yampolskiy.html)

LEAKPROOFING THE SINGULARITY: IRENE CAESAR INTERVIEWS ROMAN YAMPOLSKIY, AUGUST 6, 2013 (http://www.wavegenome.com/1/post/2013/08/leakproofing-the-singularity-irene-caesar-interviews-roman-yampolskiy.html)

IRENE CAESAR MATRIX CITY LECTURE AT THE HARRIMAN INSTITUTE (SHORT)

(https://www.youtube.com/watch?v=rVU1MHIruwQ)

IRENE CAESAR, PH.D.: "MATRIX CITY - PUBLIC LECTURE AT THE HARRIMAN INSTITUTE, COLUMBIA UNIVERSITY, SEPTEMBER 13, 2012 (https://www.youtube.com/watch?v=rVU1MHIruwQ)

MORE (/news.html)

In November 2011, Dr. Irene Caesar presented to President Vladimir Putin's Agency for Strategic Initiatives her project of the MATRIX CITY - the self-sufficient city as an integrated information-wave matrix, built with the application of the information-wave technologies of the Quantum Leap. In August 2012, Dr. Irene Caesar offered her project of the MATRIX CITY to the Institute for National Security in Moscow for practical realization. As a result, Dr. Irene Caesar, Founder, Owner and President of Wave Genome, and Dr. Alexej Diaschev had established the Consortium **"Matrix City" (http://www.thematrixcity.com).** Irene Caesar, Ph.D., presented her project of the Matrix City in her public lecture at the Harriman Institute of the Columbia University on September 13, 2012.

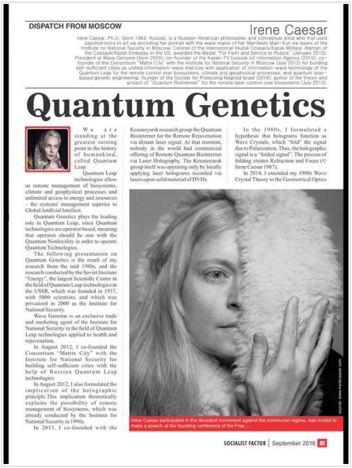

DISPATCH FROM MOSCOW

Irene Caesar

Irene Caesar, Ph.D. (born 1963, Russia), is a Russian-American philosopher and conceptual artist who first used psychotronics in art via encoding her portrait with the wave matrix of her Manifesto Man-Sun via lasers of the Institute for National Security in Moscow; Colonel of the International Irkutsk Cossack/Kazak Military; Ataman of the Cossack/Kazak Embassy in the US, awarded the Medal "For Faith and Service to Russia" (January 2015); President at Wave Genome (from 2010); co-founder of the Kazak-TV (cassak.tv) Information Agency (2015); co-founder of the Consortium "Matrix City" with the Institute for National Security in Moscow (July 2012) for building self-sufficient cities as unified information-wave matrices with application of information-wave technology of the Quantum Leap for the remote control over biosystems, climate and geophysical processes, and quantum laser-based genetic engineering; founder of the Society for Protecting Negroid Israel (2014); author of the theory and project of "Quantum Biointernet" for the remote laser control over biosystems (July 2013).

Quantum Genetics

We are standing at the greatest turning point in the history of humankind, called Quantum Leap.

Quantum Leap technologies allow us remote management of biosystems, climate and geophysical processes and unlimited access to energy and resources – the systems' management superior to Global Artificial Intellect.

Quantum Genetics plays the leading role in Quantum Leap, since Quantum technologies are operator based, meaning that operator should be one with the Quantum Nonlocality in order to operate Quantum Technologies.

The following presentation on Quantum Genetics is the result of my research from the mid 1980s, and the research conducted by the Soviet Institute "Energy", the largest Scientific Center in the field of Quantum Leap technologies in the USSR, which was founded in 1937, with 5000 scientists, and which was privatized in 2000 as the Institute for National Security.

Wave Genome is an exclusive trade and marketing agent of the Institute for National Security in the field of Quantum Leap technologies applied to health and rejuvenation.

In August 2012, I co-founded the Consortium "Matrix City" with the Institute for National Security for building self-sufficient cities with the help of Russian Quantum Leap technologies.

In August 2012, I also formulated the implication of the holographic principle. This implication theoretically explains the possibility of remote management of biosystems, which was already conducted by the Institute for National Security in 1990s.

In 2013, I co-founded with the Krasnoyarsk research group the Quantum Biointernet for the Remote Rejuvenation via distant laser signal. At that moment, nobody in the world had commercial offering of Remote Quantum Biointernet via Laser Holography. The Krasnoyarsk group itself was operating only by locally applying laser holograms recorded via lasers upon solid material of DVDs.

In the 1980s, I formulated a hypothesis that holograms function as Wave Crystals, which "fold" the signal due to Polarization. Thus, the holographic signal is a "folded signal". The process of folding creates Refraction and Focus (© Irene Caesar 1987).

In 2014, I extended my 1980s Wave Crystal Theory to the Geometrical Optics

Irene Caesar participated in the decedent movement against the communist regime, was invited to make a speech at the founding conference of the Free ...

Source: www.irenecaesar.com

SOCIALIST FACTOR | September 2016 41

(/theory---105810451054105610481071.html)

SEPTEMBER 2016

IRENE CAESAR'S ARTICLE "QUANTUM GENETICS" IN SEPTEMBER 2016 ISSUE OF "SOCIALIST FACTOR" MAGAZINE (LUCKNOW AND LONDON), as the publication of Irene Caesar's paper "Remote Management of Biosystems Based upon the Shape of Chromosomes as Diffraction Grating for Focusing Genetic Information", delivered as lecture at the Bangkok Thonburi University International Symposium, July 23, 2016. **Click here** (http://2016-9-irene-caesar-quantum-genetics-socialist-factor-magazine.pdf) for the PDF of the article.

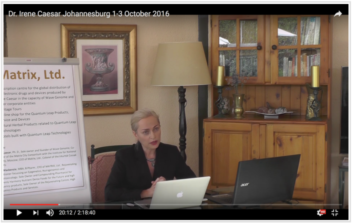

(https://www.youtube.com/watch?v=MuYnjs6khiU)

OCTOBER 1-2, 2016

IRENE CAESAR'S SEMINAR "REMOTE MANAGEMENT OF BIOSYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION", Jane Mackenzie's Rejuvenating Centre, Johannesburg. **Click here** (https://www.youtube.com/watch?v=MuYnjs6khiU) for the video.

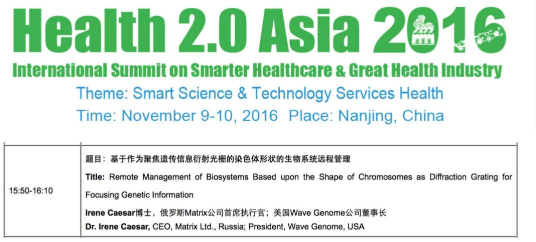

Health 2.0 Asia 2016

International Summit on Smarter Healthcare & Great Health Industry

Theme: Smart Science & Technology Services Health

Time: November 9-10, 2016 Place: Nanjing, China

15:50-16:10	题目：基于作为聚焦遗传信息衍射光栅的染色体形状的生物系统远程管理 **Title:** Remote Management of Biosystems Based upon the Shape of Chromosomes as Diffraction Grating for Focusing Genetic Information **Irene Caesar**博士，俄罗斯Matrix公司首席执行官；美国Wave Genome公司董事长 **Dr. Irene Caesar,** CEO, Matrix Ltd., Russia; President, Wave Genome, USA

NOVEMBER 9-10, 2016

IRENE CAESAR, PH.D., SPEECH "REMOTE MANAGEMENT OF BIOSYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION" at the International Health 2.0 Asia 2016 Conference in Nanjing, China, supported by Nanjing Municiple Government, and with the participation of Professor Daniel Shechtman, Nobel Prize Laureate in Chemistry (2011). **Click here (http://www.bitcongress.com/HEALTH2016/default.asp)** for the official webpage of the Summit. **Click here (http://www.wavegenome.com/2016_11_9_10_Irene_Caesar_at_Health_2_Asia_International_Summit_Speaker_List.pdf)** for the official list of participants.

2016 健康2.0亚洲会议 Health 2.0 Asia
国际生物医药产业发展高峰论坛暨人才与项目对接会
International Biomedicine Industry Development
中国·南京 2016.11
IRENE CAESAR, Ph.D.
Remote management of biosystems based upon the shape of chromosomes as diffraction grating for focusing genetic information

Irene Caesar, Ph.D., Speech "Remote Management of Biosystems Based upon the Shape of Chromosomes as Diffraction Grating for Focusing Genetic Information", International Symposium HEALTH 2.0, Nanjing, China, November 9-10, 2016.

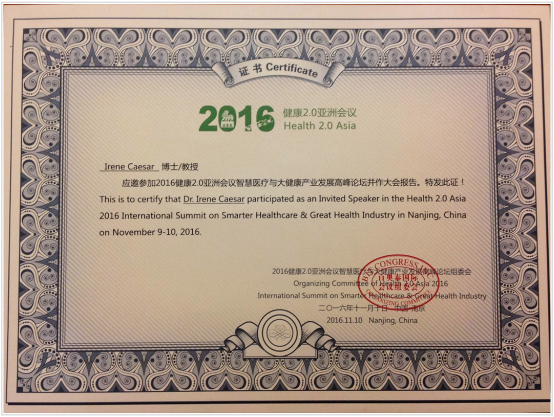

证书 Certificate

2016 健康2.0亚洲会议 Health 2.0 Asia

Irene Caesar 博士/教授

应邀参加2016健康2.0亚洲会议智慧医疗与大健康产业发展高峰论坛并作大会报告。特发此证！

This is to certify that Dr. Irene Caesar participated as an Invited Speaker in the Health 2.0 Asia 2016 International Summit on Smarter Healthcare & Great Health Industry in Nanjing, China on November 9-10, 2016.

2016健康2.0亚洲会议智慧医疗与大健康产业发展高峰论坛组委会
Organizing Committee of Health 2.0 Asia 2016
International Summit on Smarter Healthcare & Great Health Industry
二〇一六年十一月十日 中国·南京
2016.11.10 Nanjing, China

Irene Caesar, Ph.D., Speech "Remote Management of Biosystems Based upon the Shape of Chromosomes as Diffraction Grating for Focusing Genetic Information", International Symposium HEALTH 2.0, Nanjing, China, November 9-10, 2016.

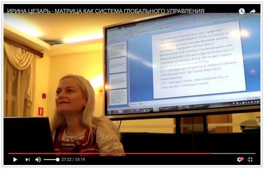

(https://www.youtube.com/watch?v=bECHfG4WttQ)

OCTOBER 10, 2016

LECTURE BY IRENE CAESAR, Ph.D., "MATRIX AS A GLOBAL

WE OFFER:

1:40 / 20:29
IRENE CAESAR AND MAGNUS OLSSON WAVE GENOME HELP

Irene Caesar, Ph.D., Founder, Owner and President of WAVE GENOME with Magnus Olsson, President of Mind Tech Enterprises, and Director of World, European and US Coalitions Against Covert Electronic Harassment, one of Distributors of WAVE GENOME.

- **MINI-TESLA PSI-GENERATOR (/tesla-generator.html)**

as a chip (artificial crystal) structured with the generic or personal biohologram / wave matrix by the laser on the nano level for healing, rejuvenation and self-defense from harmful electromagnetic fields, psychotronic and psychic attacks, and genetic warfare;

- **WAVE PHARMACEUTICALS (/wave-pharmaceuticals.html) VIA DIGITAL PHARMACY (/digital-pharmacy-software--1057105410601058-10621048106010561054104210541049-104010551058104510501048.html)**

via generic and individual wave matrices. **WAVE MATRIX** (http://www.wavegenome.com/wave-pharmaceuticals.html) is the hologram, recorded by the laser. **WAVE MATRICES** (http://www.wavegenome.com/wave-pharmaceuticals.html) include (a) INDIVIDUAL WAVE MATRIX (http://www.wavegenome.com/wave-pharmaceuticals.html), recorded by the laser from the childhood photograph, placenta or navel blood, or/and (b) generic **WAVE MATRIX** (http://www.wavegenome.com/wave-pharmaceuticals.html) recorded by the laser from the healthy human cells and bodily systems, and digitised as **DIGITAL CELLS, (http://www.wavegenome.com/wave-pharmaceuticals.html)**(c) generic or personal **WAVE MATRICES (http://www.wavegenome.com/wave-pharmaceuticals.html)** of medications as psi-medications, including such psi-pharmaceuticals, as psi-VYagra, psi-hormones, etc. **WAVE PHARMACEUTICALS (/wave-pharmaceuticals.html)** has various platforms for dispensation, including:

(1) generic or personal **WAVE MATRICES (http://www.wavegenome.com/wave-pharmaceuticals.html),** encoded by the laser in the chip of the **PSI-GENERATOR (http://www.wavegenome.com/order-your-mini-tesla-generator.html)** on the nano level;

(2) **INDIVIDUAL WAVE MATRIX (http://www.wavegenome.com/wave-pharmaceuticals.html), DIGITAL CELL (http://www.wavegenome.com/wave-pharmaceuticals.html)** and **DIGITAL PHARMACEUTICALS (/wave-pharmaceuticals.html)** -

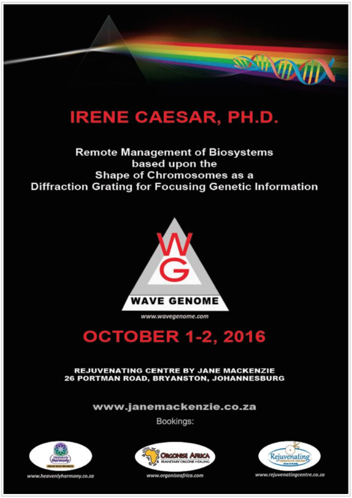

IRENE CAESAR, PH.D.

Remote Management of Biosystems
based upon the
Shape of Chromosomes as a
Diffraction Grating for Focusing Genetic Information

WAVE GENOME
www.wavegenome.com

OCTOBER 1-2, 2016

REJUVENATING CENTRE BY JANE MACKENZIE
26 PORTMAN ROAD, BRYANSTON, JOHANNESBURG

www.janemackenzie.co.za
Bookings:

www.heavenlyharmony.co.za www.orgoniseafrica.com www.rejuvenatingcentre.co.za

OCTOBER 1-2, 2016

IRENE CAESAR'S SEMINAR "REMOTE MANAGEMENT OF BIOSYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION", Jane Mackenzie's (http://www.janemackenzie.co.za) Rejuvenating Centre Johannesburg. (http://www.rejuvenatingcentre.co.za)

MANAGEMENT SYSTEM OF LIFE. MECHANISM OF REMOTE MANAGEMENT OF BIOSYSTEMS." Conference "Psychics - yesterday, today and tomorrow", October 10, 2016, (4th Meeting of the Association for Astral Anthropology and Psychology), House of Culture "Scarlet Sails", Moscow, Russia. **Click here** (https://www.youtube.com/watch?v=bECHfG4WttQ) for the video. **Click here (https://www.youtube.com/watch?v=eEKzZQ4Wzcc)** for another version of vidoe recording.

(/theory---10581045105410561048107 1.html)

JULY 2016

IRENE CAESAR, PH.D., LECTURE "REMOTE MANAGEMENT OF BIOSYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION", Bangkok Thonburi University International Symposium, July 23, 2016. **Click here** (https://www.youtube.com/watch?v=XuwV_QCTf2I) for the recording of Irene Caesar's lecture at the Bangkok Thonburi University. **Click here** (https://www.youtube.com/watch?v=pSsgOPUzz9A) for the slide show of Irene Caesar's lecture at the Bangkok Thonburi University International Symposium.

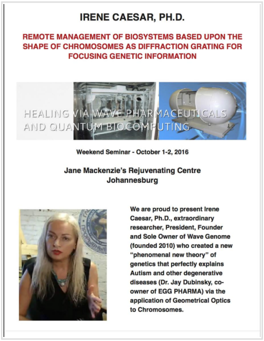

IRENE CAESAR, PH.D.

REMOTE MANAGEMENT OF BIOSYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION

HEALING VIA WAVE PHARMACEUTICALS AND QUANTUM BIOCOMPUTING

Weekend Seminar - October 1-2, 2016

Jane Mackenzie's Rejuvenating Centre
Johannesburg

We are proud to present Irene Caesar, Ph.D., extraordinary researcher, President, Founder and Sole Owner of Wave Genome (founded 2010) who created a new "phenomenal new theory" of genetics that perfectly explains Autism and other degenerative diseases (Dr. Jay Dubinsky, co-owner of EGG PHARMA) via the application of Geometrical Optics to Chromosomes.

OCTOBER 1-2, 2016

PRESS RELEASE FOR IRENE CAESAR'S SEMINAR "REMOTE MANAGEMENT OF BIOSYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION", Jane Mackenzie's Rejuvenating Centre Johannesburg. **Click here (http://www.wavegenome.com/2016_9_1_irene_caesar_seminar_zar_press_release_final.pdf)** for PDF file of the Press Release.

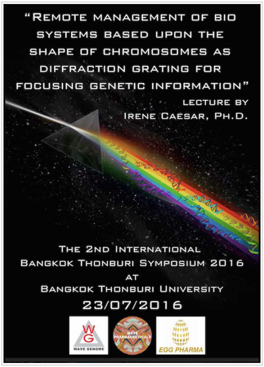

"REMOTE MANAGEMENT OF BIO SYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION"
LECTURE BY
IRENE CAESAR, PH.D.

THE 2ND INTERNATIONAL
BANGKOK THONBURI SYMPOSIUM 2016
AT
BANGKOK THONBURI UNIVERSITY
23/07/2016

WAVE GENOME WAVE PHARMACEUTICALS EGG PHARMA

(/theory---10581045105410561048107 1.html)

SEPTEMBER 2016

PUBLICATION OF IRENE CAESAR'S, PH.D., PAPER "REMOTE MANAGEMENT OF BIOSYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION" in the Proceedings of the Bangkok Thonburi University International Symposium, July 23, 2016, also published in the September Issue of the "Socialist Factor" Magazine (Lucknow and London). **Click here** (http://www.wavegenonome.com/2016_BTUSymposium.pdf) for PDF of the BTU Symposium Proceedings.

generic or personalised wave matrices of human cells and pharmacueuticals recorded by lasers, digitised and administered as holographic / granulated signal in the form of the **WHITE NOISE (http://www.wavegenome.com/wave-pharmaceuticals.html)**, encoded into the music in the acoustic and video **PSI-PROGRAM (/digital-pharmacy-software---1057105410601058-10621048106010561054104210541049-1040105510581045105010481048.html)** of the binaural self-therapy. Each psi-generator comes with the flash drive, which contains a software with **COMPREHENSIVE DIGITAL PHARMAC (/digital-pharmacy-software---1057105410601058-10621048106010561054104210541049-1040105510581045105010481048.html)Y** that has 44 digital modules of the digitized wave matrices in the form of the white noise, encoded into the music, for every bodily system (e.g., endocrine system, reproductive system, cardiovascular system, digestive system, etc.) and every vital organ in the body (e.g., liver, kidney, brain, etc.) -- 22 modules for organs and 22 modules for physiological systems. Digital drugs per se, like psi-vyagra or psi-hormones, are offered as extra. White noise, generated by the **PSI-PROGRAM (/digital-pharmacy-software---1057105410601058-10621048106010561054104210541049-1040105510581045105010481048.html)**, can be used for programming water. Digital encoding of wave matrices onto various media via laser is available;

(3) **QUANTUM INTERNET (/quantum-rejuvenation-clinic-om---english.html)** - wireless programming of biosystems, water and human condition via laser. **QUANTUM INTERNET** is **REMOTE, WIRELESS,** and **INDIVIDUALISED;**

- **TESLA WATER (/tesla-water.html)**

Bioactive restructured water charged remotely via **QUANTUM INTERNET** with the use of laser through Quantum nonlocality**,** or charged locally by the white noise (initially programmed by the laser, and, then, digitized), when you run your program of the **DIGITAL PHARMACY** (/wave-pharmaceuticals---digital-pharmacy.html) on your computer. In both cases, you need to purchase **MINI-TESLA PSI-GENERATOR (/tesla-generator.html)**. **TESLA WATER (/tesla-water.html)** is either individualised for every client or generic for specific conditions. **TESLA WATER** can be used in agriculture to increase its productivity manifold;

- **REMOTE REJUVENATION "OM" (/quantum-rejuvenation-clinic-om---english.html)**

via **QUANTUM INTERNET** for all the physiological systems, including endocrine system, digestive system, cardiovascular system, hearing / eye sight, reproductive system, intellectual abilities, improvement of memory, psychic abilities, sexual potency, etc. This service is provided remotely via **WIRELESS QUANTUM INTERNET (/quantum-rejuvenation-clinic-om---english.html);**

- **SUBSCRIPTION PACKAGES FOR MAINTENANCE (http://wavegenome.com/wave-genome---prices.html) (in works)**

Client can purchase the everyday maintenance of his **WAVE MATRIX (http://www.wavegenome.com/wave-pharmaceuticals.html).** Client sends his data to the Institute either via the conventional internet, and via the **QUANTUM INTERNET (/quantum-rejuvenation-**

clinic-om---english.html). HOLOGRAPHIC MAINTENANCE (http://www.wavegenome.com/wave-pharmaceuticals.html) is the correction of the intensity and modules of bioholograms / wave matrices in accordance with the changing client's condition, earthly electro-magnetic activity, solar and lunar activity, geophysical activity, and predicted social unrest. The correction is provided by the specialists of WAVE GENOME LLC remotely or locally. **HOLOGRAPHIC MAINTENANCE (http://www.wavegenome.com/wave-pharmaceuticals.html)** is provided via the update of the digital files in the Digital Pharmacy and by the remote (non-local) laser signal via **QUANTUM INTERNET (/quantum-rejuvenation-clinic-om---english.html).** For example, we can offer the remote management of diabetes;

- **CLINIC with LARGE-SIZE PSI-GENERATORS (/quantum-clinic.html)**

Exclusive rejuvenation services in our partner clinics for restoration and self-defense, especially for patients who were refused by conventional medicine, infertility cases, impotence, drug dependence, and rehabilitation. BIOHOLOGRAPHIC TREATMENT can restore menstruation cycles in the post-menopausal women over 60 years old, so that a woman can again bear a child;

- **TESLA ENHANCEMENT (/quantum-clinic.html)**

Special programs in our partner clinics for those who are seeking to enhance their mental and physical abilities via **LARGE-SIZE PSI-GENERATORS (/quantum-clinic.html);**

- **CONSULTING SERVICES (/ask-our-maven.html)**

regarding the Quantum Leap information-wave technologies by our mavens.

Irene Caesar

Listen to my gospel till it is not too late

(/uploads/3/5/1/1/35112207/4952633_orig.jpg)

OCTOBER 2016

IRENE CAESAR, PH.D., ARTICLE "LISTEN TO MY GOSPEL TILL IT IS NOT TOO LATE" IN OCTOBER 2016 ISSUE OF "SOCIALIST FACTOR MAGAZINE (LUCKNOW AND LONDON). Click here (http://www.wavegenome.com/2016-10-irene-caesar-listen-to-my-gospel-socialist-factor-magazine.pdf) for the PDF of the article.

DISPATCH FROM MOSCOW

Irene Caesar

Irene Caesar, Ph.D. (born 1963, Russia), is a Russian-American philosopher and conceptual artist who first used psychotronics in art via encoding her portrait with the wave matrix of her Manifesto Man-Sun via lasers of the Institute for National Security in Moscow; Colonel of the International Irkutsk Cossack/Kazak Military; Ataman of the Cossack/Kazak Embassy in the US, awarded the Medal "For Faith and Service to Russia" (January 2015); President at Wave Genome (from 2010); co-founder of the Kazak-TV (cassak.tv) Information Agency (2015); co-founder of the Consortium "Matrix City" with the Institute for National Security in Moscow (July 2012) for building self-sufficient cities as unified information-wave matrices with application of information-wave technology of the Quantum Leap for the remote control over biosystems, climate and geophysical processes, and quantum laser-based genetic engineering; founder of the Society for Protecting Negroid Israel (2014); author of the theory and project of "Quantum Biointernet" for the remote laser control over biosystems (July 2013).

How to Build Matrix City

The problem with digital insulin seemed to be the fact that digital insulin meant cure. People get cured from diabetes, even type 1. This takes a sick person off the insulin market, and, as a result, the insulin producer loses money. If the technology produces mass healing, the insulin producer would sooner or later get bankrupt. But this is a major misunderstanding of the information-wave technology. This technology is in fact a subscription, rather than simply a cure. You get your own personal matrix in digital form, recorded onto the MINI-TESLA generator. Then, you connect your personal generator every week to the information-wave center, which corrects (tunes) your personal wave matrix, and combines it with digital drugs if needed.

Russian scientists are serving farms with lasers and generators already for many years for the sake of manifold increasing the crop yield. They do not sell generators - they sell only the service.

The same can be done by Big Pharma, after it switches from the analog to the digital. This model of "subscription" or "contract" will involve much more people into the information-wave system than it is now involved by Big Pharma.

Young people can get the service for enhancing their mental and physical abilities. It is possible to enhance sexual performance, and take vitamins and other supplements in the information-wave form.

I believe that in one generation, everybody will be using this technology in every aspect of life. Those who will accept this technology will get a chance to live for 200 years and longer. In 10 years from now, "getting one's maintenance via information-wave technology" will be like brushing one's teeth every morning.

WHAT ABOUT CHIPPING?

I believe that information-wave technology has nothing to do with chipping the population. Chipping the population is the AI paradigm, when the choice is alienated, and the holographic principle is violated. The AI cannot in principle perfect human wave matrices or create a superman (Übermensch), who will be more perfect than humans. Human matrices are "uncreated", non-local, that is-eternal, indivisible, and infinite, while AI is "artificial", limited, divisible, and finite in virtue of definition. AI can in principle only assist in having healthier children.

That is why psi-generators should not be in-built in the human body, but only added to it as a kind of amulets of the psychotronic era. Man is himself a powerful and most perfect generator. Having audacity of installing into a human a generator that is supposed to "control" or "rule" the human is laughable.

I will give you an example so that you understand what I mean. There are clearly two paradigms of Übermensch right now. The first one is a NATO "transformer", stuffed with nanobots, who is a paradigm of cut-and-paste genetic engineering on the bio-chemical-cyber level. The other paradigm is Russian paradigm of non-contact combat. We can enhance a Russian soldier from Special forces with psi-generators, as, in fact, Dr. Diashev does, but we cannot substitute a generator or chip for his mind or "eidos", which is quantum infinity itself, non-local and

Source: futurewarstories.blogspot.com

Women comprise around 20% of most modern military o

38 August 2016 | SOCIALIST FACTOR

(/theory---105810451054105610481071.html)

AUGUST 2016

IRENE CAESAR'S ARTICLE (2012) "HOW TO BUILD MATRIX CITY" IN AUGUST 2016 ISSUE OF "SOCIALIST FACTOR" MAGAZINE (LUCKNOW AND LONDON), first published on thematrixcity.com in 2012, as a major document of the Matrix City Consortium (founded August 2012), Co-founded by Dr. Irene Caesar and the Institute for National Security in Moscow. **Click her** (http://wavegenome.com/2016-8-irene-caesar-matrix-city-socialist-factor-magazine.pdf)**e** for the PDF of the article.

PRIVACY POLICY (/privacy-policy.html)

SHIPPING POLICY (/shipping--warranty.html)

RETURN POLICY (/return-policy.html)

TERMS OF SERVICE (/terms-of-service.html)

FDA DISCLAIMER: Content on this site is for reference purposes and is not a substitute for advice from a licensed health-care professional. You should not rely solely on this content, and wave genome.com assumes no liability for inaccuracies. Always read directions before using a product. Statements regarding products and services have not been evaluated by the FDA, and these products and services are not intended to diagnose, treat, cure, or prevent any disease or condition.

Payment in full is due ahead of services. No refunds and returns.

COPYRIGHT NOTICE: Text, graphics, and HTML code on wavegenome.com belong to Irene Caesar, Ph.D., their sole author, and are protected by US and International Copyright Laws, and may not be copied, reprinted, published, translated, hosted, or otherwise distributed by any means without explicit permission from Irene Caesar, Ph.D., and in violation of the International Laws for citing references. Any citation of the wavegenome.com website should be accompanied with the proper reference to Irene Caesar, Ph.D. Any citation exceeding the length, permissible for the reference length by the International Copyright Law, will be considered to be a violation of the International Copyright Law.

© IRENE CAESAR, PH.D., FOUNDER AND SOLE OWNER OF WAVEGENOME 2010-2017 ALL RIGHTS RESERVED

IRENE CAESAR, PH.D. / ИРИНА ЦЕЗАРЬ, ДОКТ. ФИЛОСОФ. НАУК ~ http://www.thematrixcity.com

Tag Archives: *wave pharmaceuticals*

Remote management of biosystems based upon the shape of chromosomes as diffraction grating for focusing genetic information

14 Tuesday Jun 2016

POSTED BY IRENE CAESAR, PH.D. / ИРИНА ЦЕЗАРЬ, ДОКТОР ФИЛОСОФСКИХ НАУК IN NEGROID BLOODLINE OF THE QUEEN OF ENGLAND

≈ 3 COMMENTS

Tags
acrocentric chromosome, chromosome shape, crystalline structure of torus, diffraction grating, digital drugs, e-drugs, electronic drugs, he-ne laser, holographic principle, i-drugs, institute for national security in moscow, irene caesar, matrix city, metacentric chromosome, nano chips of new generation, nanotechnology, polarized holography, quantum biointernet, remote control over biosystems, structuralist genetics, toroidal, torsion, torus, wave crystal, wave genome, wave matrix, wave pharmaceuticals

Remote management of biosystems based upon the shape of chromosomes as diffraction grating for focusing genetic information

Irene Caesar, Ph.D.

June 13, 2016

Abstract

The same gene gets expressed in dysfunctional individuals by acrocentric chromosome, and in functional individuals by metacentric chromosome, and till now nobody was able to explain why. I claim that chromosome functions as wave crystal, i.e., nonlocal unique wave matrix that focuses genetic information unique to this individual via his / her own unique scalar wave diffraction grating. Thus, chromosome is a structural / geometric code or shape that has holographic nature allowing for the remote management of biosystems via laser polarized holography.

Keywords

Irene Caesar, wave genome, polarized holography, chromosome shape, metacentric chromosome, acrocentric chromosome, remote control over biosystems, digital drugs, electronic drugs, quantum biointernet, wave pharmaceuticals, diffraction grating, wave crystal, crystalline structure of torus, nanotechnology, nanochips of new generation, remote laser control system, holographic principle, wave matrix, matrix city, structuralist genetics

Introduction

We can conduct remote laser management of the biosystems, and, thus, create individuals with enhanced physical and psychic abilities, and dramatically extend life span, based upon the implication of the holographic principle (© Quantum Biointernet, Irene Caesar 2013), and mastering the geometric codes of chromosomes.

Objectives

The objective is to build a Matrix City, i.e., the self-sufficient human settlement of the Quantum Leap Era (http://www.thematrixcity.com and specifically: http://www.thematrixcity.com/structure-of-the-matrix-city-r…), where enhanced man in enhanced environment can have enhanced life of any life span he chooses. The objective includes: (1) creating BANKS OF WAVE MATRICES for every living creature on this planet; (2) creating MATRIX CITY ZONES, where individuals receive all their supplements and medications via E-DRUGS and I-DRUGS from the MATRIX CITY based upon subscription model; (3) cultivating men with enhanced physical, intellectual and psychic abilities.

Review of Related Literature

The Polarized Holography is mainly based upon Russian research, especially upon the research of Nikolai Kozyrev, and his so called “Kozyrev Mirror” principle; the structuralist genetics by Alexander Gurvich; as well as upon the toroidal (“torsion”) research by Anatoly Akimov and Gennady Schipov. In 2002 some information got licked to the West, when two researchers of the NPO “Aura” (NPO “Energy”/BINAR, which became “The Institute for National Security” in 2000) Peter Gariaev and George Tertyshny, who did not have clearance and full access to the research, have brought a laser to Toronto to demonstrate remote management of the artificially induced diabetes via the laser signal at a distance of 20 kilometers in the rats of the same genetic line. But Gariaev and Tertyshny did not have access to the Digital Pharmacy of the digitized bioholograms of healthy bodily organs and physiological systems (aka “Digital Cell”), and could not diagnose wave matrices. The technology was classified at that time in Russia, and got declassified only in 2008.

Conceptual Framework

Universe is holographic, i.e., it is entirely contained in its every matrix point. Since universe is holographic, we can record, transmit and receive only holographic information. The implication of the holographic principle states that, If entire universe is contained in its every matrix point (holographic principle), then, every matrix point is not simply different from any other matrix point, but is unique (implication of the holographic principle, © Irene Caesar 2012). If we have an exact copy of the unique nonlocal wave matrix, we can send information instantaneously through Quantum Nonlocality, from one copy of the unique nonlocal wave matrix to the other copy, to any dimension, any planet, or galaxy, to the past, and to the future.

Research Methodology

The chromosome geometry changes the nature of the external signal from linear to holographic (@ Irene Caesar 2012, implication of holographic principle) via (1) torus polarization (via the change of the spin in the integral toroidal field of DNA); (2) reflection of any external linear signal upon itself and creation of the scalar / standing wave diffraction grating, which neutralizes any external signal; (3) refraction of any incoming signal towards the zero center of the wave crystal, which functions as a focus similar to how our eye functions (Irene Caesar, Chromosome Geometry ©2014).

Quantum Leap technology applying knowledge of the chromosome geometry is called “polarized holography”, and is based upon application of stabilized Helium-Neon Lasers with internal mirrors. This kind of lasers are capable of creating holographic signal. Analogously, our bones are capable of producing stem cells, only because the shape of the bone is analogous to laser tube and is capable of changing the nature of external signal.

Research Results

As of today, the Institute for National Security in Moscow offers the local and remote management of biosystems, and individuals:
(1) via electronic / digital drugs (DIGITAL PHARMACY consisting of digitized biohologram of wave matrices for every healthy bodily organ and physiological system; and including E-DRUGS [additional upon subscription], of which more than 300 are already created, including E-VYAG-RA and I-VYAG-RA);
(2) via remote laser signal (I-DRUGS or Quantum Biointernet).

Institute has conducted the first in the world official and public remote lowering of blood sugar in diabetic calf via the laser signal at the distance of 20 kilometers for the American Diabetic Association in 1999. Over 2015, the Institute has created the digital platform for dispensation of E-DRUGS and I-DRUGS via subscription analogous to mobile phone subscription programs.

E-DRUGS are offered in two forms: (1) as chips inside a pendant [MINI-TESLA PSI-GENERATOR] of electret nature structured by laser on the nanolevel, and (2) as software program on flash drive sold with a MINI-TESLA PSI-GENERATOR, which contains the entire DIGITAL PHARMACY (soon only by subscription).

Discussion

It is clear that the exiting model of Human Genome is not correct, since it cannot explain 98% of DNA calling it “Junk DNA”. Thus, genetic engineering based upon “cutting and pasting” of “identifiable” 2% of DNA is doomed for failure. Only the Holographic Wave Genetics, based upon chromosome geometry, can explain the difference between functional and dysfunctional individuals, and promote the advance in genetic engineering.

Advertisements

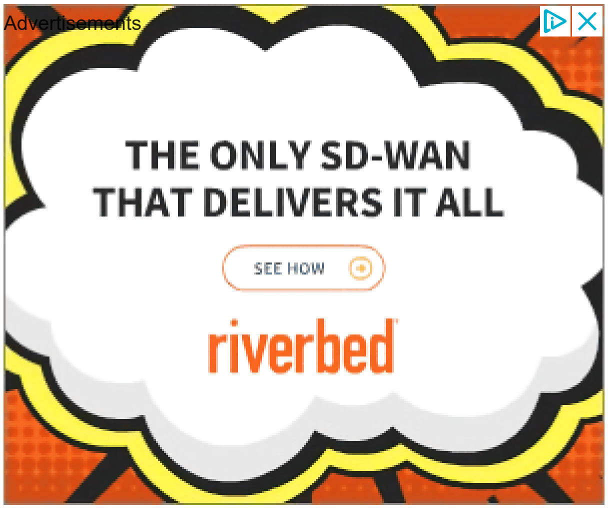
THE ONLY SD-WAN THAT DELIVERS IT ALL

SEE HOW

riverbed

Report this ad

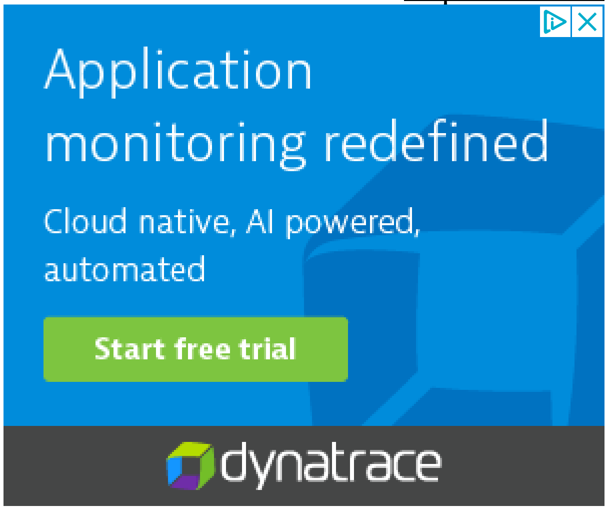
Application monitoring redefined

Cloud native, AI powered, automated

Start free trial

dynatrace

Report this ad

Irene Caesar – Article in the "I KNOW" Magazine "On How to Build Matrix City", March 28, 2015

21 Thursday May 2015

POSTED BY IRENE CAESAR, PH.D. / ИРИНА ЦЕЗАРЬ, ДОКТОР ФИЛОСОФСКИХ НАУК IN NEGROID BLOODLINE OF THE QUEEN OF ENGLAND

≈ 3 COMMENTS

Tags

bci, brain-computer interface, chipping, global artificial intellect, how to build matrix city, i know magazine, irene caesar, matrix city, matrix zones, non-fossil energy, wave genome, wave pharmaceuticals

IRENE CAESAR, PH.D. ON HOW TO BUILD MATRIX CITY

“I KNOW” Magazine, Canada, March 28, 2015

http://www.iknowmagazine.ca/irene-caesar-ph-d-build-matrix-city/

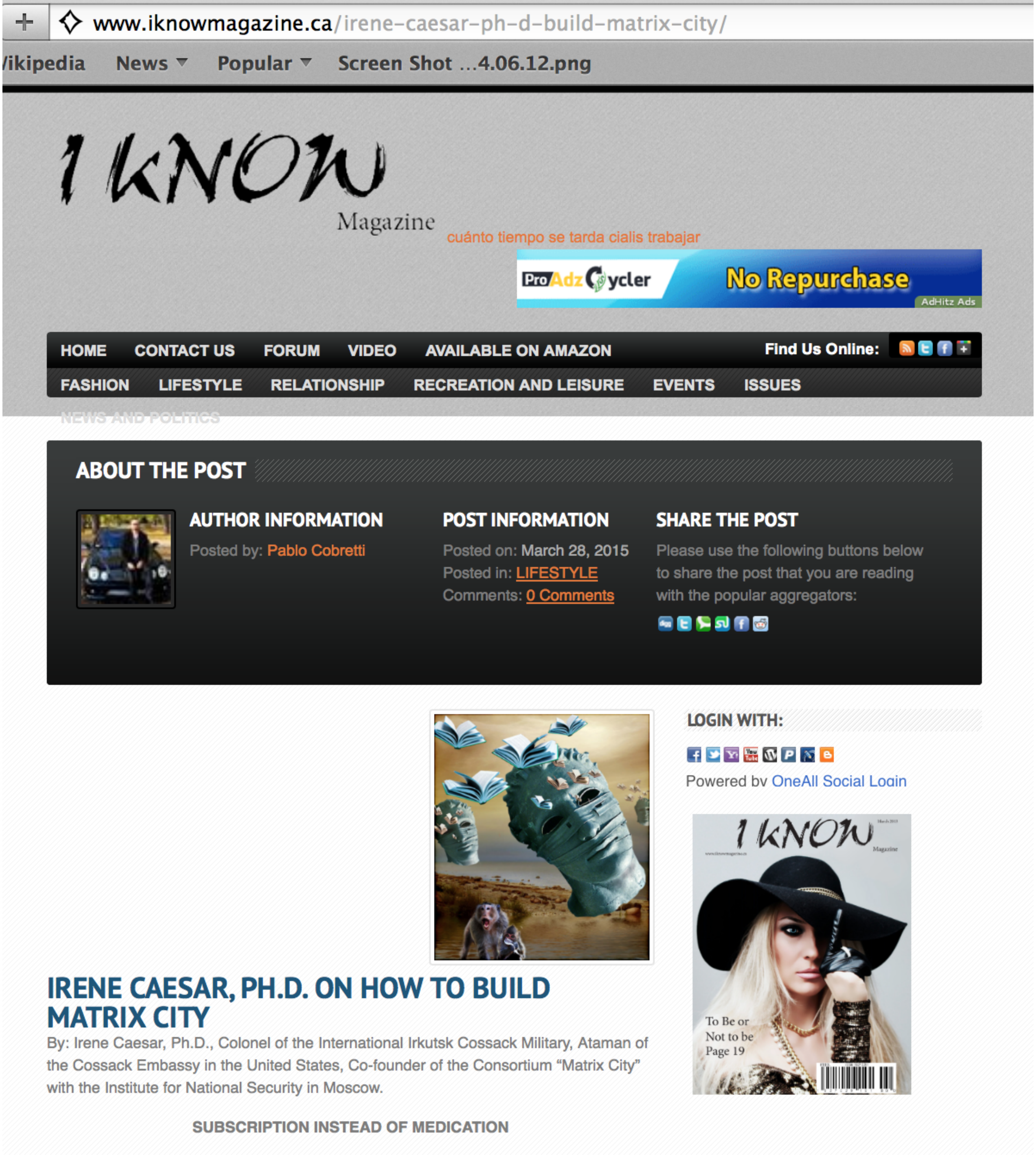
www.iknowmagazine.ca/irene-caesar-ph-d-build-matrix-city/
/ikipedia News Popular Screen Shot ...4.06.12.png
I KNOW
Magazine
cuánto tiempo se tarda cialis trabajar
ProAdz Cycler No Repurchase AdHitz Ads
HOME CONTACT US FORUM VIDEO AVAILABLE ON AMAZON Find Us Online:
FASHION LIFESTYLE RELATIONSHIP RECREATION AND LEISURE EVENTS ISSUES
NEWS AND POLITICS
LOGIN WITH:
Powered by OneAll Social Login
I KNOW Magazine
To Be or Not to be Page 19

ABOUT THE POST

AUTHOR INFORMATION

Posted by: Pablo Cobretti

POST INFORMATION

Posted on: March 28, 2015
Posted in: LIFESTYLE
Comments: 0 Comments

SHARE THE POST

Please use the following buttons below to share the post that you are reading with the popular aggregators:

IRENE CAESAR, PH.D. ON HOW TO BUILD MATRIX CITY

By: Irene Caesar, Ph.D., Colonel of the International Irkutsk Cossack Military, Ataman of the Cossack Embassy in the United States, Co-founder of the Consortium "Matrix City" with the Institute for National Security in Moscow.

SUBSCRIPTION INSTEAD OF MEDICATION

Irene Caesar's Article "On how to build Matrix City" in the "I KNOW" Magazine, Canada, March 28, 2015.

Irene Caesar's Article "On how to build Matrix City" in the "I KNOW" Magazine, Canada, March 28, 2015.

By: Irene Caesar, Ph.D., Colonel of the International Irkutsk Cossack Military, Ataman of the Cossack Embassy in the United States, Co-founder of the Consortium "Matrix City" with the Institute for National Security in Moscow.

SUBSCRIPTION INSTEAD OF MEDICATION

The problem with digital insulin seemed to be the fact that digital insulin meant cure. People get cured from diabetes, even type 1. This takes a sick person off the insulin market, and, as a result, the insulin producer loses money. If the technology produces mass healing, the insulin producer would sooner or later get bankrupt. But this is a major misunderstanding of the information-wave technology. This technology is in fact a subscription, rather than simply a cure. You get your own personal matrix in digital form, recorded onto the MINI-TESLA generator. Then, you connect your personal generator every week to the information-wave center, which corrects (tunes) your personal wave matrix, and combines it with digital drugs if needed.

Russian scientists are serving farms with lasers and generators already for many years for the sake of manifold increasing the crop yield. They do not sell generators – they sell only the service.

KONICA MINOLTA

Irene Caesar "How to Build Matrix City", "I KNOW" Magazine, Photo credit: GôDiNô / Foter / CC BY-NC-ND

The same can be done by Big Pharma, after it switches from the analog to the digital. This model of "subscription" or "contract" will involve much more people into the information-wave system than it is now involved by Big Pharma.

Young people can get the service for enhancing their mental and physical abilities. It is possible to enhance sexual performance, and take vitamins and other supplements in the information-wave form.

I believe that in one generation, everybody will be using this technology in every aspect of life. Those who will accept this technology will get a chance to live for 200 years and longer. In 10 years from now, “getting one’s maintenance via information-wave technology” will be like brushing one’s teeth every morning.

WHAT ABOUT CHIPPING?

I believe that information-wave technology has nothing to do with chipping the population. Chipping the population is the AI paradigm, when the choice is alienated, and the holographic principle is violated. The AI cannot in principle perfect human wave matrices or create a superman (Übermensch), who will be more perfect than humans. Human matrices are “uncreated”, non-local, that is — eternal, indivisible, and infinite, while AI is “artificial”, limited, divisible, and finite in virtue of definition. AI can in principle only assist in having healthier children.

That is why psi-generators should not be in-built in the human body, but only added to it as a kind of amulets of the psychotronic era. Man is himself a powerful and most perfect generator. Having audacity of installing into a human a generator that is supposed to “control” or “rule” the human is laughable.

I will give you an example so that you understand what I mean. There are clearly two paradigms of Übermensch right now. The first one is a NATO “transformer”, stuffed with nanobots, who is a paradigm of cut-and-paste genetic engineering on the bio-chemical-cyber level. The other paradigm is Russian paradigm of non-contact combat. We can enhance a Russian soldier from Special forces with psi-generators, as, in fact, Dr. Diashev does, but we cannot substitute a generator or chip for his mind or “eidos”, which is quantum infinity itself, non-local and entangled with everything forever.

CULTIVATION OF OPERATORS

Irene Caesar “How to Build Matrix City”, “I KNOW” Magazine, Photo credit: Photo credit: digitalbob8 / Foter / CC BY

The matrix is the energy structure, which has a wave component of indestructible omnipresent nature, i.e., quantum non-locality and quantum entanglement. This infinite wave literally gives life – it heals, increases life span, and gives joy. The point of building Matrix Cities is to have such architecture, agriculture, medicine, industry, energy, education, and all other sides of human life, so that they all have access to this wave component.

That is why the agenda of Matrix Cities would be the cultivating of operators who would be producing wave matrices. This cultivation should be voluntary. And those who do not want to join should be left alone.

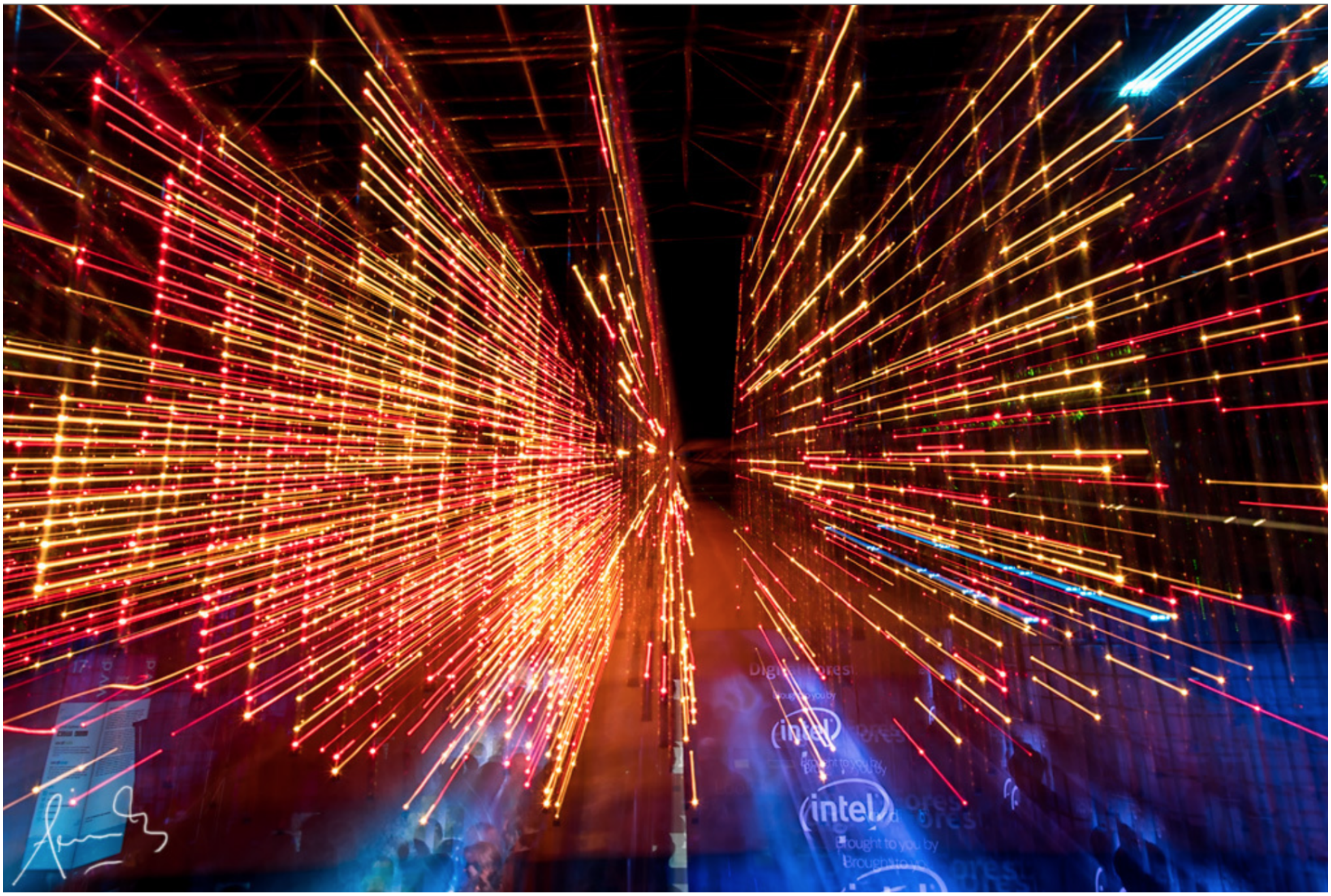

Irene Caesar "How to build Matrix City", "I KNOW" Magazine, Canada, Photo credit: sachman75 / Foter / CC BY-NC-ND.

STRUCTURE OF MATRIX CITY

Lasers and generators will constitute the INNER CORE or the NUCLEUS of the Matrix City under the protection of agencies for national security.

The NUCLEUS will consist of concentric circles with centers for the production and conservation of wave matrices.

The center of NUCLEUS will consist in the production of non-fossil fuel generators of energy.

The conservation center will take care of the wave matrices banks for the Matrix Zone.

The NUCLEUS will handle the production of bioactive water, which itself is a generator.

The NUCLEUS will regulate all forms of life in the Matrix Zone, from organism to species, from bacteria to man.

The NUCLEUS will produce lasers and generators.

Around NUCLEUS, there will be a protective perimeter.

Over the protective perimeter, around the NUCLEUS, there will be concentric circles with housing for operators, their centers for education and cultivation, the Secondary Production Centers, and the Subscription Center.

Matrix City will handle subscription to its service. Service should be provided via personal mini psi-generators, which should be connected to the information-wave system of the Matrix City through the internet at least once a week.

The Secondary Production Centers will consist of farms as agro-factories and industrial park, which will produce all components of information-wave technology — bioactive construction and industrial materials, wave pharmaceuticals, etc.

MATRIX ZONES

Irene Caesar “How to build Matrix City”, “I KNOW” Magazine, Canada, Photo credit: GôDiNô / Foter / CC BY-NC-ND.

The country should be divided into the Matrix Zones. Each Matrix zone should have its own Matrix City, which would serve its zone with wave matrices.

It looks like, advanced people will gather in special zones which are serviced by the Matrix City, the same way as now those who can are looking for areas with better schools for their children. These zones will be ecologically pure, with bioactive architecture, and bioactive water. These zones should accept everyone who wants to join, independently of money. The only condition should be the acceptance of a connection to the Matrix City. But nobody should be forced to join.

This condition will put down all accusations that Matrix Cities are a new exploitation device and a new segregation device. In fact, Matrix Cities should take care of personal, racial and national uniqueness and survival as much as the rare animal species are protected by the Red Book. DNA wave matrices banks should preserve all the variety of human genome.

SPECIAL CHARTER FOR MATRIX CITIES

Generators and lasers producing wave matrices should be protected by the state and agencies for national security.

Matrix Cities should have their own Charter, i.e., be special, semi-military zones, because ultimately this is a more powerful weapon than nuclear bomb, and, if used peacefully, generators will be more desirable thing than gold.

It is completely clear that without the control from the state and agencies for national security, this technology cannot be commercialized in principle. Matrix City is basically a conversion of psychotronic weapons into peaceful applications in industry, medicine, agriculture, etc.

Coca-Cola
Coca-Cola

Irene Caesar “How to build Matrix City”, “I KNOW” Magazine, Canada, Photo credit: GôDiNô / Foter / CC BY-NC-ND.

Remote & Wireless Maintenance via Quantum Internet by the Institute for National Security

in Moscow

15 Saturday Nov 2014

POSTED BY IRENE CAESAR, PH.D. / ИРИНА ЦЕЗАРЬ, ДОКТОР ФИЛОСОФСКИХ НАУК IN BIOHOLOGRAM, BIOHOLOGRAMS, HE-NE LASER, HOLOGRAPHY, INSTITUTE FOR NATIONAL SECURITY, IRENE CAESAR, MATRIX CITY, ORANIA, POLARISED HOLOGRAPHY, QUANTUM INTERNET, QUANTUM LEAP, WAVE GENOME, WAVE MATRIX, WAVE PHARMACEUTICALS, WELLENGENETISHER HEILUNGSCODE

≈ 3 COMMENTS

Tags

biohologram, biohologograms, he-ne laser, holography, institute for national security, irene caesar, matrix city, orania, polarised hologrpahy, quantum computer, quantum leap, torus, wave genome, wave matrix, wave pharmaceuticals, wellengenetisher heilungscode, zero field

REMOTE AND WIRELESS MAINTENANCE VIA QUANTUM INTERNET BY THE INSTITUTE FOR NATIONAL SECURITY IN MOSCOW

November 15, 2014

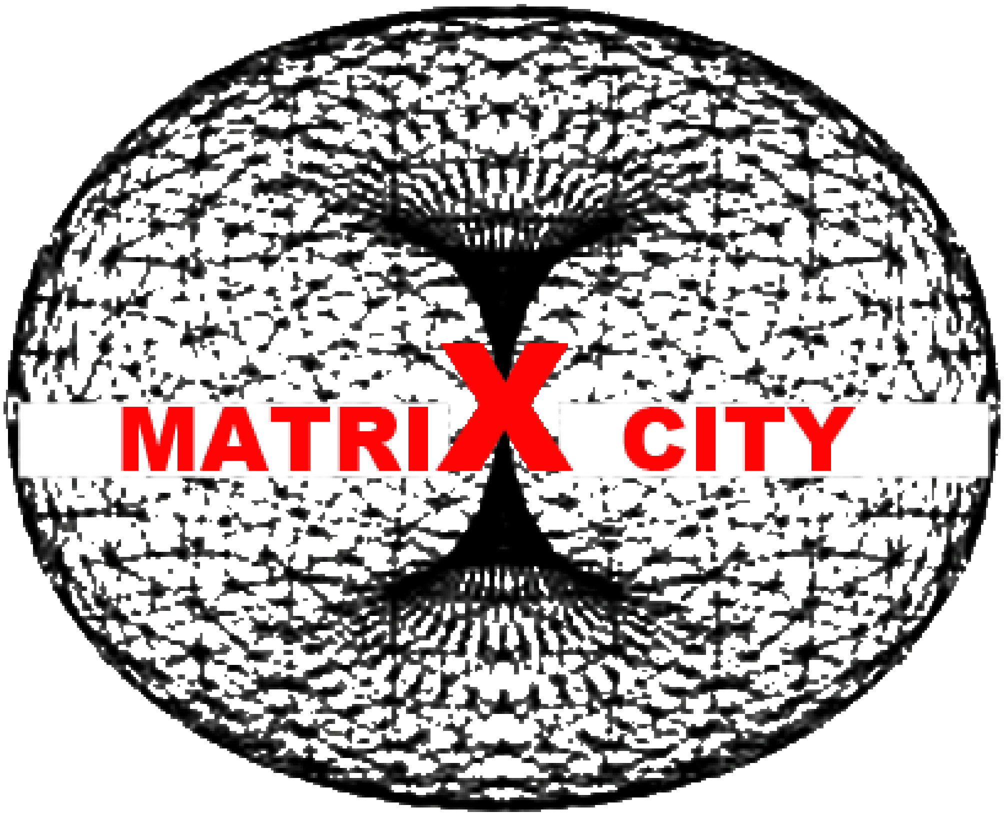

Matrix City © Irene Caesar

In the case of QUANTUM INTERNET, you send your information REMOTELY and WIRELESSLY. At the Institute, the laser reads the feedback from your wave matrix in real time, using your digitised wave matrix on record at the Institute.

According to the holographic principle, the universe is holographic, meaning that it is entirely in its every matrix part. The holographic principle has three important implications:

The first implication of the holographic principle postulates the UNIQUENESS of every wave matrix. if the universe is entirely in its every matrix point, then, every matrix point is not simply different from any other matrix point, but is unique.

The second implication of the holographic principle postulates the NON-LOCALITY of every wave matrix. If the universe is entirely in its every matrix point, then every wave matrix is simultaneously present in the infinite number of dimensions via its tokens. Though there are infinitely many tokens of the same non-local wave matrix, all of them constitute one integrated SYSTEMATIC WHOLE.

The non-local holographic nature of wave matrices also means that the wave matrix has the infinite set of waves, particles and fields within its TORUS — from infinitely micro to infinitely macro dimensions. It is impossible to talk of each individual wave matrix as "a specific frequency". And it is impossible to postulate any "medium" of its existence, like "scalar wave", "aether", "Higgs / boson field", "torsions", "leptons", etc. There is no specific medium for transferring the information.

The third implication of the holographic principle postulates that there is the QUANTUM ENTANGLEMENT in the form of the instantaneous transmission of information between all the tokens of the same non-local wave matrix. The signal that records and transmits the ENTANGLED INFORMATION between different wave matrices or between the tokens of the same non-local wave matrix has specific nature. It is HOLOGRAPHIC. Only lasers can record and transmit holographic signal, based upon the principles of the polarised refraction holography (described in terms of geometrical optics). Thus, the principle of recording and transmitting the MATRIX SIGNAL is structural, and not medium-based.

Thus, it is possible to remotely and instantaneously transfer some wave information to a man via the entanglement between his body (at one location) and his photograph (at another location), as two tokens of his non-local wave matrix.

For example, it is possible to remotely lower blood sugar in a diabetic man via scanning his photograph with the laser ray, which is passing through insulin before reaching the photograph.

© Irene Caesar, Ph.D. from her book "WAVE PHARMACEUTICALS" in co-authorship with Steve Schwartz, MD (2014)

http://www.wavegenome.com

http://thematrixcity.com/structure-of-the-matrix-city-english.html

WELLENGENETISCHER HEILUNGSCODE seminar by Dr. Irene Caesar in Germany:

21. November 2014, Freitag, 18 – ca. 21 Uhr

http://www.orania-zentrum.de/index.php/abendveranstaltungen/infoabend-wellengenetischerheilungscode
http://www.orania-zentrum.de/index.php/seminare/seminar-wellengenetische-heilungscode

Übermensch

The ***Übermensch*** (German for "Beyond-Man", "Superman", "Overman", "Superhuman", "Hyperman", "Hyperhuman"; German pronunciation: [ˈʔyːbɐmɛnʃ]) is a concept in the philosophy of Friedrich Nietzsche. In his 1883 book *Thus Spoke Zarathustra* (German: *Also sprach Zarathustra*), Nietzsche has his character Zarathustra posit the Übermensch as a goal for humanity to set for itself. It is a work of philosophical allegory, with a structural similarity to the Gathas of Zoroaster/Zarathustra.

Contents

In English

This-worldliness

Death of God and the creation of new values

As a goal

Re-embodiment of amoral aristocratic values

Relation to the eternal recurrence

Nazism

Anarchism

In popular culture

See also

References

External links

In English

The first translation of *Thus Spoke Zarathustra* into English was published in 1896. In that translation, by Alexander Tille, Übermensch was translated as "Beyond-Man". In the Thomas Common translation, published in 1909, however, *Übermensch* was rendered as "Superman". Common was anticipated in this by George Bernard Shaw, who had done the same in his 1903 stage play *Man and Superman*. Walter Kaufmann lambasted this translation in the 1950s for two reasons: first, its near or total failure to capture the nuance of the German word *über* (while the Latin prefix super- means above or beyond, the English use of the prefix or its use as an adjective has altered the meaning); and second, a rationale which Fredric Wertham railed against even more vehemently in *Seduction of the Innocent*, for promoting identification by children with the comic-book character Superman (whom Wertham described as "un-American and fascist"). The preference of Kaufmann and others is to translate Übermensch as "overman". Scholars continue to employ both terms, some simply opting to reproduce the German word.[1][2]

The German prefix *über* can have connotations of superiority, transcendence, excessiveness, or intensity, depending on the words to which it is attached.[3] *Mensch* refers to a member of the human species, rather than to a male specifically. The adjective *übermenschlich* means super-human, in the sense of beyond human strength or out of proportion to humanity.[4]

This-worldliness

Nietzsche introduces the concept of the Übermensch in contrast to his understanding of the other-worldliness of Christianity: Zarathustra proclaims the Übermensch to be the meaning of the earth and admonishes his audience to ignore those who promise other-worldly hopes in order to draw them away from the earth.[5][6] The turn away from the earth is prompted, he says, by a dissatisfaction with life—a dissatisfaction that causes one to create another world in which those who made one unhappy in this life are tormented. The Übermensch is not driven into other worlds away from this one.

Zarathustra declares that the Christian escape from this world also required the invention of an eternal soul which would be separate from the body and survive the body's death. Part of other-worldliness, then, was the abnegation and mortification of the body, or asceticism. Zarathustra further links the Übermensch to the body and to interpreting the soul as simply an aspect of the body.

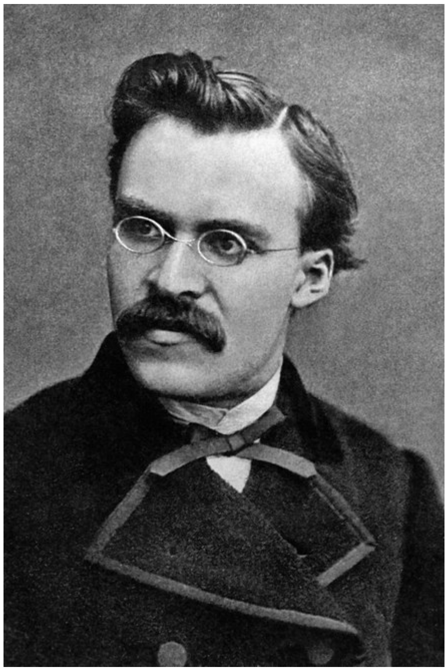

Friedrich Nietzsche in 1869.

Death of God and the creation of new values

Zarathustra ties the Übermensch to the death of God. While this God was the ultimate expression of other-worldly values and the instincts that gave birth to those values, belief in that God nevertheless did give meaning to life for a time. 'God is dead' means that the idea of God can no longer provide values. With the sole source of values no longer capable of providing those values, there is a real chance of nihilism prevailing.

Zarathustra presents the Übermensch as the creator of new values. In this way, it appears as a solution to the problem of the death of God and nihilism. If the Übermensch acts to create new values within the moral vacuum of nihilism, there is nothing that this creative act would not justify. Alternatively, in the absence of this creation, there are no grounds upon which to criticize or justify any action, including the particular values created and the means by which they are promulgated.

In order to avoid a relapse into Platonic idealism or asceticism, the creation of these new values cannot be motivated by the same instincts that gave birth to those tables of values. Instead, they must be motivated by a love of this world and of life. Whereas Nietzsche diagnosed the Christian value system as a reaction against life and hence destructive in a sense, the new values which the Übermensch will be responsible for will be life-affirming and creative (see Nietzschean affirmation).

As a goal

Zarathustra first announces the Übermensch as a goal humanity can set for itself. All human life would be given meaning by how it advanced a new generation of human beings. The aspiration of a woman would be to give birth to an Übermensch, for example; her relationships with men would be judged by this standard.[7]

Zarathustra contrasts the Übermensch with the last man of egalitarian modernity, an alternative goal which humanity might set for itself. The last man appears only in *Thus Spoke Zarathustra*, and is presented as a condition that would render the creation of the Übermensch impossible.

According to Rüdiger Safranski, some commentators associate the Übermensch with a program of eugenics.[8] This is most pronounced when considered in the aspect of a goal that humanity sets for itself. The reduction of all psychology to physiology implies, to some, that human beings can be bred for cultural traits. This interpretation of Nietzsche's doctrine focuses more on the future of humanity than on a single cataclysmic individual. There is no consensus regarding how this aspect of the Übermensch relates to the creation of new values.

Re-embodiment of amoral aristocratic values

For Rüdiger Safranski, the Übermensch represents a higher biological type reached through artificial selection and at the same time is also an ideal for anyone who is creative and strong enough to master the whole spectrum of human potential, good and "evil", to become an "artist-tyrant". In *Ecce Homo*, Nietzsche vehemently denied any idealistic, democratic or humanitarian interpretation of the Übermensch: "The word Übermensch [designates] a type of supreme achievement, as opposed to 'modern' men, 'good' men, Christians, and other nihilists ... When I whispered into the ears of some people that they were better off looking for a Cesare Borgia than a Parsifal, they did not believe their ears."[9] Safranski argues that the combination of ruthless warrior pride and artistic brilliance that defined the Italian Renaissance embodied the sense of the Übermensch for Nietzsche. According to Safranski, Nietzsche intended the ultra-aristocratic figure of the Übermensch to serve as a Machiavellian bogeyman of the modern Western middle class and its pseudo-Christian egalitarian value system.[10]

Relation to the eternal recurrence

The Übermensch shares a place of prominence in *Thus Spoke Zarathustra* with another of Nietzsche's key concepts: the eternal recurrence of the same. Several interpretations for this fact have been offered.

Laurence Lampert suggests that the eternal recurrence replaces the Übermensch as the object of serious aspiration.[11] This is in part due to the fact that even the Übermensch can appear like an other-worldly hope. The Übermensch lies in the future — no historical figures have ever been Übermenschen — and so still represents a sort of eschatological redemption in some future time.

Stanley Rosen, on the other hand, suggests that the doctrine of eternal return is an esoteric ruse meant to save the concept of the Übermensch from the charge of Idealism.[12] Rather than positing an as-yet unexperienced perfection, Nietzsche would be the prophet of something that has occurred a countless number of times in the past.

Others maintain that willing the eternal recurrence of the same is a necessary step if the Übermensch is to create new values, untainted by the spirit of gravity or asceticism. Values involve a rank-ordering of things, and so are inseparable from approval and disapproval; yet it was dissatisfaction that prompted men to seek refuge in other-worldliness and embrace other-worldly values. Therefore, it could seem that the Übermensch, in being devoted to any values at all, would necessarily fail to create values that did not share some bit of asceticism. Willing the eternal recurrence is presented as accepting the existence of the low while still recognizing it as the low, and thus as overcoming the spirit of gravity or asceticism.

Still others suggest that one must have the strength of the Übermensch in order to will the eternal recurrence of the same; that is, only the Übermensch will have the strength to fully accept all of his past life, including his failures and misdeeds, and to truly will their eternal return. This action nearly kills Zarathustra, for example, and most human beings cannot avoid other-worldliness because they really are sick, not because of any choice they made.

Nazism

The term Übermensch was utilized frequently by Hitler and the Nazi regime to describe their idea of a biologically superior Aryan or Germanic master race;[13] a form of Nietzsche's Übermensch became a philosophical foundation for the National Socialist ideas. Their conception of the Übermensch, however, was racial in nature.[14][15] The Nazi notion of the master race also spawned the idea of "inferior humans" (*Untermenschen*) which could be dominated and enslaved; this term does not originate with Nietzsche. Nietzsche himself was critical of both antisemitism and German nationalism. In his final years, Nietzsche began to believe that he was in fact Polish, not German, and was quoted as saying, "I am a pure-blooded Polish nobleman, without a single drop of bad blood, certainly not German blood".[16] In defiance of these doctrines, he claimed that he and Germany were great only because of "Polish blood in their veins",[17] and that he would be "having all anti-semites shot" as an answer to his stance on anti-semitism. Although the term has been associated with the Nazis, Nietzsche was dead long before Hitler's reign. It was Nietzsche's sister Elisabeth Förster-Nietzsche who actually first played a part in manipulating her brother's words to accommodate the worldview of herself and her husband, Bernhard Förster, a prominent German nationalist and antisemite.[18] In order to support his beliefs he set up the *Deutscher Volksverein* (German People's League) in 1881 with Max Liebermann von Sonnenberg.[19]

Anarchism

The thought of Nietzsche had an important influence in anarchist authors (see Anarchism and Friedrich Nietzsche). Spencer Sunshine writes that "There were many things that drew anarchists to Nietzsche: his hatred of the state; his disgust for the mindless social behavior of 'herds'; his anti-Christianity; his distrust of the effect of both the market and the State on cultural production; his desire for an 'overman' — that is, for a new human who was to be neither master nor slave; his praise of the ecstatic and creative self, with the artist as his prototype, who could say, 'Yes' to the self-creation of a new world on the basis of nothing; and his forwarding of the 'transvaluation of values' as source of change, as opposed to a Marxist conception of class struggle and the dialectic of a linear history."[20] The influential American anarchist Emma Goldman in her famous collection of essays *Anarchism and Other Essays* in the preface passionately defends both Nietzsche and Max Stirner from attacks within anarchism when she says "The most disheartening tendency common among readers is to tear out one sentence from a work, as a criterion of the writer's ideas or personality. Friedrich Nietzsche, for instance, is decried as a hater of the weak because he believed in the Übermensch. It does not occur to the shallow interpreters of that giant mind that this vision of the Übermensch also called for a state of society which will not give birth to a race of weaklings and slaves."[21]

Sunshine says that the "Spanish anarchists also mixed their class politics with Nietzschean inspiration." Murray Bookchin, in *The Spanish Anarchists*, describes prominent CNT–FAI member Salvador Seguí as "an admirer of Nietzschean individualism, of the *superhombre* to whom 'all is permitted'." Bookchin, in his 1973 introduction to Sam Dolgoff's *The Anarchist Collectives*, even describes the reconstruction of society by the workers as a Nietzschean project. Bookchin says that "workers must see themselves as human beings, not as class beings; as creative personalities, not as 'proletarians,' as self-affirming individuals, not as 'masses'. . .(the) economic component must be humanized precisely by bringing an 'affinity of friendship' to the work process, by diminishing the role of onerous work in the lives of producers, indeed by a total 'transvaluation of values' (to use Nietzsche's phrase) as it applies to production and consumption as well as social and personal life."[20]

In popular culture

- The comic-book hero Superman, when Jerome "Jerry" Siegel first created him, was originally a villain modeled on Nietzsche's idea (see "The Reign of the Superman"). He was re-invented as a hero by his eventual designer, Joseph "Joe" Shuster, after which he bore little resemblance to the previous character. However, Superman does find an adversary in the mold of the Nietzschean Übermensch in the recurring arch-villain Lex Luthor, his greatest enemy on Earth. Luthor is preceded, even, by a supervillain resembling Siegel's original concept for Superman bearing the synonymous name of 'Ultra-Humanite'. A direct reference to the term occurs in the episode "Double Trouble" of the TV series *Adventures of Superman,* in which a German-speaking character calls the title character "*verfluchter*

Übermensch" ("cursed Superman").[22] In the tenth season of the show *Smallville*, an alternate Lionel Luthor refers to Clark as the "Übermensch". Overman is an alternate version of Superman, from Earth-10, an Earth where the Axis Powers won World War II.[23]

- Jack London dedicated his novels *The Sea-Wolf* and *Martin Eden* to criticizing Nietzsche's concept of the Übermensch and his radical individualism,[24] which London considered to be selfish and egoistic.
- George Bernard Shaw's 1903 play *Man and Superman* is a reference to the archetype; its main character considers himself an untameable revolutionary, above the normal concerns of humanity.
- James Joyce utilizes the *Übermensch* in the first chapter of his novel Ulysses. Joyce makes Buck Mulligan say it:[25] "—My twelfth rib is gone, he cried. I'm the *Uebermensch*. Toothless Kinch and I, the supermen."
- In *The Power*, a 1956 book by Frank M. Robinson, the villain consciously models himself upon Nietzsche's Übermensch, and a quotation from Nietzsche serves as the book's motto.
- In real life, Leopold and Loeb committed murder in 1924 partly out of a superficially Übermensch-like conception of themselves.[26] Their story has been dramatized many times, including in the Alfred Hitchcock film *Rope*, the 1959 film Compulsion based on Meyer Levin's novel, the 1994 film *Swoon*, the 2002 film *Murder by Numbers*, and the 2005 Off-Broadway musical *Thrill Me: The Leopold and Loeb Story*.
- A character in the show *Dollhouse* (Season 1, Episode 12; titled "Omega") references Übermensch in relation to Nietzsche when trying to describe a person that had the memories, skills, and intelligence of dozens of people uploaded into their (single) mind by means of futuristic technology.
- The Medic character in the video game *Team Fortress 2* will sometimes chant "I am the Übermensch!" for one of his "Cheers" voice commands.
- In the Disney TV Cartoon, Pepper Ann, in an episode titled "Effie Shrugged", Effie, a physically imposing bully teaches Pepper Ann some basic information on Übermensch.
- David Bowie's song titled "The Supermen", from his 1970 album *The Man Who Sold the World*, was inspired by the work of Nietzsche. Bowie later said "I was still going through the thing when I was pretending that I understood Nietzsche... And I had tried to translate it into my own terms to understand it so 'Supermen' came out of that."[27]

See also

- Great Man theory
- Junzi
- Knight of faith
- Last man
- New Man (utopian concept)
- New Soviet man
- *Notes from Underground*
- Philosopher-King
- Posthuman
- Metabolic supermice
- Randian hero
- *Strange Life of Ivan Osokin*
- Transhumanism
- Supermind (Integral yoga)

References

Notes

1. Lampert, Laurence (1986). *Nietzsche's Teaching*. New Haven, CT: Yale University Press.
2. Rosen, Stanley (1995). *The Mask of Enlightenment*. Cambridge: Cambridge University Press.
3. *Duden Deutsches Universal Wörterbuch A–Z,* s.v. über-.
4. Übermenschlich. PONS.eu Online Dictionary. Retrieved from http://en.pons.eu/german-english/%C3%BCbermenschlich.
5. Hollingdale, R. J. (1961), page 44 - English translation of Zarathustra's prologue; "I love those who do not first seek beyond the stars for reasons to go down and to be sacrifices: but who sacrifice themselves to the earth, that the earth may one day belong to the Superman"
6. Nietsche, F. (1885) - p4, Original publication in German - "Ich liebe die, welche nicht erst hinter den Sternen einen Grund suchen, unterzugehen und Opfer zu sein: sondern die sich der Erde opfern, dass die Erde einst des Übermenschen werde."
7. *Thus Spoke Zarathustra*, I.18; Lampert, *Nietzsche's; Rosen,* Mask of Enlightenment*, 118.*
8. Safranski, *Nietzsche*, 262-64, 266-68.
9. Nietzsche, *Ecce Homo*, Why I Write Such Good Books, §1)

10. Safranski, *Nietzsche*, 365
11. Lampert, *Nietzsche's Teaching*.
12. Rosen, *The Mask of Enlightenment*.
13. Alexander, Jeffrey (2011). *A Contemporary Introduction to Sociology* (2nd ed.). Paradigm. ISBN 978-1-61205-029-4.
14. "Nietzsche inspired Hitler and other killers - Page 7" (http://www.trutv.com/library/crime/notorious_murders/famous/nietzsche_crimes/7.html), Court TV Crime Library
15. http://econ161.berkeley.edu/tceh/Nietzsche.html
16. Friedrich Nietzsche, "Ecce Homo: How One Becomes What One Is" [1] (https://books.google.com/books/about/Ecce_Homo.html?id=tVL-Dmg6TlUC)
17. Henry Louis Mencken, "The Philosophy of Friedrich Nietzsche", T. Fisher Unwin, 1908, reprinted by University of Michigan 2006, pg. 6, [2] (https://books.google.com/books?id=nnEOAAAAIAAJ&pg=PA6&as_brr=3)
18. Hannu Salmi (1994). "Die Sucht nach dem germanischen Ideal" (http://users.utu.fi/hansalmi/forster.html) (in German). Also published in *Zeitschrift für Geschichtswissenschaft* 6/1994, pp. 485-496
19. Karl Dietrich Bracher, *The German Dictatorship*, 1970, pp. 59-60
20. Spencer Sunshine, "Nietzsche and the Anarchists" (http://radicalarchives.org/2010/05/18/nietzsche-and-the-anarchists/)
21. (http://theanarchistlibrary.org/HTML/Emma_Goldman__Anarchism_and_Other_Essays.html)Anarchism and Other Essays by Emma Goldman
22. Grossman, Gary (1976). *Superman: Serial to Cereal*. New York: Popular Library. p. 67. ISBN 0445040548.
23. Grant Morrison (w), Doug Mahnke (p), Christian Alamy, Rodney Ramos, Tom Nguyen, Walden Wong (i). *Final Crisis: Superman Beyond* 1 (October 2008), DC Comics
24. Bridgwater, Patrick (1972). *Nietzsche in Anglosaxony: A Study of Nietzsche's Impact on English and American Literature*. Leicester: Leicester University Press. p. 169. ISBN 0718511042.
25. Joyce, James (1922). *Ulysses*. Shakespeare & Co. ISBN 0-679-72276-9.
26. "Nietzsche inspired Hitler and other killers" (http://www.crimelibrary.com/notorious_murders/famous/nietzsche_crimes/index.html), Court TV Crime Library
27. David Buckley (1999). *Strange Fascination - David Bowie: The Definitive Story*: p.267

Bibliography

- Knoll, Manuel (2014) "The Übermensch as Social and Political Task: A Study in the Continuity of Nietzsche’s Political Thought", in: Knoll, Manual and Stocker, Barry (eds.) (2014) *Nietzsche as Political Philosopher*, Berlin/Boston, pp. 239–266.
- Lampert, Laurence (1986) *Nietzsche's Teaching: An Interpretation of Thus Spoke Zarathustra*. New Haven: Yale University Press.
- Nietzsche, Friedrich (1885) *Thus Spoke Zarathustra*
- Nietzsche, Friedrich; Hollingdale, R. J. and Rieu, E.V. (eds.) (1961_ *Thus Spoke Zarathustra* Penguin Classics: Penguin Publishing (Originally published 1885)
- Rosen, Stanley (1995) *The Mask of Enlightenment: Nietzsche's Zarathustra*. New York: Cambridge University Press.
- Safranski, Rudiger (2002 *Nietzsche: A Philosophical Biography*. Translated by Shelley Frisch. New York: W.W. Norton & Co.
- Wilson, Colin (1981) *The Outsider*. Los Angeles: J.P. Tarcher.

External links

- Martin Heidegger and Nietzsche’s Overman: Aphorisms on the Attack (http://www.freewebs.com/m3smg2/HeideggerOverman.htm)
- Human Superhuman (https://groups.yahoo.com/group/human_superhuman/) Yahoo! Group dedicated to Nietzsche's Overman.
- Searchable database of Nietzsche Quotes (http://www.nietzsche-quotes.com)

- An academic exchange concerning the relationship between Nietzsche and transhumanism (http://ieet.org/index.php/IEET/more/pellissier20120423)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Übermensch&oldid=821725096"

This page was last edited on 22 January 2018, at 07:03.

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

(/)

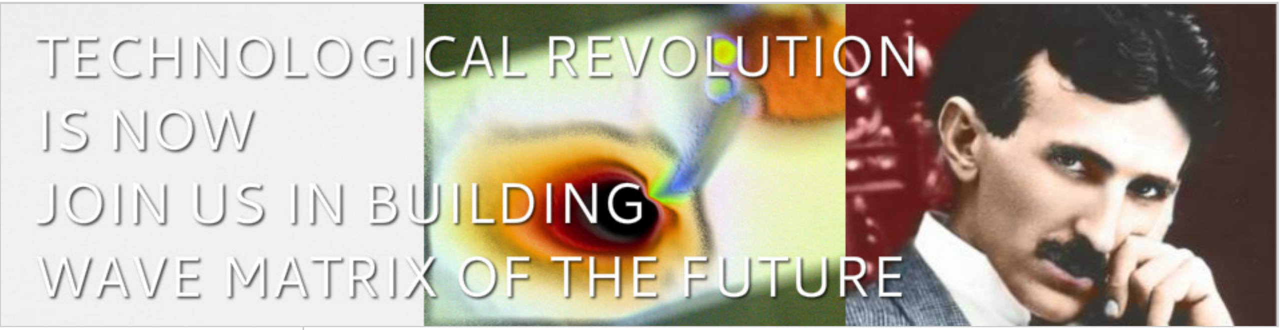

HOME / ГЛАВНАЯ (/)

SERVICES / УСЛУГИ (/SERVICES--105910571051105910431048.HTML)

PRODUCTS / ТОВАРЫ (/PRODUCTS--105810541042104010561067.HTML)

CONTACT / КОНТАКТ (/CONTACT--1050105410531058104010501058.HTML)

ORDER
YOUR MINI-TESLA
GENERATOR

(/tesla-generator.html)

QUANTUM REMOTE REJUVENATION

(/quantum-rejuvenation-clinic-om---english.html)

ORDER
BIOACTIVE
WATER

(/tesla-water.html)

PRICES

Ordering your MINI-TESLA PSI-GENERATOR (/tesla-generator.html), REMOTE REJUVENATION SERVICE (/quantum-rejuvenation-clinic-om---english.html), E-VIAGRA (/tesla-generator.html) and **I-VIAGRA (/quantum-rejuvenation-clinic-om---english.html), TESLA BIOACTIVE WATER (http://www.wavegenome.com/order-bioactive-water.html), VISIT TO OUR CLINICS IN MOSCOW OR SIBERIA (http://www.wavegenome.com/visit-our-clinic-for-cure-and-rejuvenation.html), and CONSULTATION (http://www.wavegenome.com/ask-our-maven.html) means you accept the conditions of the Client Consent Form. (/client-consent.html) All Bank fees for payments are paid by the clients.**

MODELS OF MINI-TESLA PSI-GENERATOR (/tesla-generator.html)

(1) BASIC TESLA PSI-GENERATOR (chip)

(a) has twice less layers of the chip than our other models**;** (b) has either the **GENERIC** or **INDIVIDUAL** biohologram of a healthy human in the chip, being encoded via laser onto the chip on nano level; (c) **BASIC TESLA PSI-GENERATOR** has both sides of the casing for the chip made of Real Solid Pearl; (d) **BASIC-X TESLA PSI-GENERATOR** has one side of the casing for the chip made of Real Solid Pearl, and another side of casing made of 3D printing plastic, encoded with the immune system rejuvenation **E-DRUG** (**BIOELECTRONIC DRUG** / **DIGITIZED BIOHOLOGRAM**). This makes **"TWO-CHIPS IN ONE"** solution.

GENERIC BASIC MINI-TESLA GENERATOR - $856
INDIVIDUAL BASIC MINI-TESLA GENERATOR - $1179
INDIVIDUAL BASIC-X MINI-TESLA PSI-GENERATOR - $1534

(2) SUPER MINI-TESLA GENERATOR (chip)

(a) has twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR;** (b) has either **GENERIC** or **INDIVIDUAL** biohologram; (c) Both sides of the casing for the chip are made of Real Solid Pearl.

GENERIC SUPER MINI-TESLA GENERATOR - $1179

FAQ
Frequently Asked Questions

(/mini-tesla-generator-faq.html)

READ INFORMATION LEAFLET FOR THE (/information-leaflet-for-mini-tesla-generator.html) MINI-TESLA GENERATOR (/information-leaflet-for-mini-tesla-generator.html)

ULTIMATE MINI-TESLA PSI-GENERATOR is the ultimate protection against psychotronic and genetic warfare.

ULTIMATE TESLA PSI-GENERATOR (/ultimate-mini-

WAVE GENOME TESLA GLASSES are not "glasses" per se, but the Quantum Biocomputer. TESLA GLASSES improve eye sight guaranteed, and treat internal organs via Polarized Biohologhraphy.

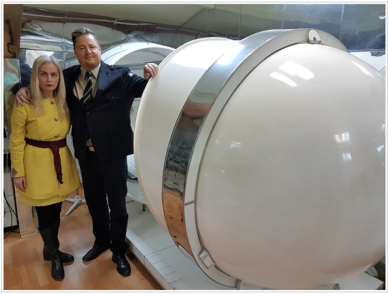

Irene Caesar, Ph.D., Founder, Owner and President of WAVE GENOME with Magnus Olsson, distributor of WAVE GENOME products.

INDIVIDUAL SUPER MINI-TESLA GENERATOR - $1768

(3) ULTRA MINI-TESLA GENERATOR (chip)

(a) has twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR;** (b) has either **GENERIC** or **INDIVIDUAL** biohologram; (c) One side of the casing for the chip is made of Real Solid Pearl, and another side of casing is made of 3D printing plastic, encoded with the immune system rejuvenation **E-DRUG** (**BIOELECTRONIC DRUG** / **DIGITIZED BIOHOLOGRAM**). This makes **"TWO-CHIPS IN ONE"** solution.

GENERIC ULTRA MINI-TESLA GENERATOR - $1768
INDIVIDUAL ULTRA MINI-TESLA GENERATOR - $2358

(4) MATRIX TESLA PSI-GENERATOR (chip)

(a) has twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR.**

(b) has either **GENERIC** or **INDIVIDUAL** biohologram.

(c) Both sides of the casing for the chip are made of 3D printing plastic, encoded with the immune system rejuvenation **E-DRUG** (**BIOELECTRONIC DRUG** / **DIGITIZED BIOHOLOGRAM**). This makes **"TWO-CHIPS IN ONE"** solution.

GENERIC MATRIX MINI-TESLA GENERATOR - $2122
INDIVIDUAL MATRIX MINI-TESLA GENERATOR - $2947

(4) ULTIMATE (/ultimate-mini-tesla-psi-generator.html)MINI-TESLA GENERATOR (/ultimate-mini-tesla-psi-generator.html) (chip)

(a) has twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR**; (b) has either **GENERIC** or **INDIVIDUAL** biohologram; (c) One side of the casing for the chip is made of Real Solid Pearl, and another side of casing is made of 3D printing plastic, encoded with the RNA **E-DRUG**(**BIOELECTRONIC DRUG** / **DIGITIZED BIOHOLOGRAM**) for the ultimate protection against psychotronic and genetic warfare. This makes **"TWO-CHIPS IN ONE"** solution.

GENERIC ULTIMATE MINI-TESLA GENERATOR - $2358
INDIVIDUAL ULTIMATE MINI-TESLA GENERATOR - $2712 (promotional price)

(5) SECOND-BIRTH TESLA PSI-GENERATOR (/second-birth-mini-tesla-psi-generator.html) (chip)

(a) **chip is encoded by laser on nano level with the most valuable biohologhraphic information on the global market: EMBRYONIC STEM CELL;**

(b) has a new technology of encoding the chip by the newly invented laser (patented);

(c) layers of the chip are made of a new material of organic nature more advanced than our other chips;

(d) its efficacy is proven by the clinical trials at the Medical Clinic of the Moscow Bauman Russian State Technical University, the most important Science University in Russia, of MIT level. twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR.**

(e) has either **GENERIC** or **INDIVIDUAL** biohologram.

(f) provides the ultimate modulation for the scalar wave diffraction grating for the most precise focus / zero center of the biohologram (aka 'wave crystal") in the process of rejuvenation.

GENERIC SECOND-BIRTH MINI-TESLA GENERATOR - $3000
INDIVIDUAL SECOND-BIRTH MINI-TESLA GENERATOR - $3500

tesla-psi-generator.html) **IS BASED UPON IMPORTANT DISCOVERY BY DR. IRENE CAESAR REGARDING THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING / REFRACTION MECHANISM FOR FOCUSING GENETIC INFORMATION**

ULTIMATE MINI-TESLA psi-generator (/ultimate-mini-tesla-psi-generator.html) has the digitized biohologram of RNA encoded into the casing. RNA is a single-stranded spiral antennae, and it is RNA which sends and receives wave information. While DNA is a double-stranded structure which neutralizes every external linear signal in the standing / scalar / zero wave (reflects every wave upon itself, so that the peak of the forward-going wave is neutralized by the trough of the same wave when it is reflected upon itself and backward-going). Thus, DNA is protected by its very structure. while RNA is vulnerable due to its very structure. Protection against Psychotronic and Generic warfare does essentially consist in tuning RNA, making our chromosomes center and focus our own unique generic information during cell division.

To read about the **ULTIMATE TESLA PSI-GENERATOR click here (/ultimate-mini-tesla-psi-generator.html)**.

TESLA GLASSES (/order-tesla-glasses.html)

TESLA GLASSES for restoration of eye sight, and treatment of physiological systems and bodily organs via Polarized Biohologhraphy of Eye Crystal and iris. Can be used by all members of the family. Hit of sales - **$1090**

Our clinic has AURA PSI-GENERATOR specifically for treating spine issues.

TESLA SELENIUM (/order-tesla-selenium.html)

TESLA SELENIUM MAXI is 60 pills a bottle/pack, each pill is 70 µg of Organic Two-Valent Selenium.**PROMOTIONAL PRICE:**

3 PACKS FOR $210
5 PACKS FOR $300
10 PACKS FOR $500

TESLA SELENIUM PRO is 60 pills a bottle/pack, each pill is 25.0 µg of Organic Two-Valent Selenium, 100 µg of Organic Germanium, and 5.0 mg of Organic Zinc.**PROMOTIONAL PRICE:**

3 PACKS FOR $240
5 PACKS FOR $350
10 PACKS FOR $600

TREATMENTS AT THE CLINIC (/quantum-clinic.html)

TREATMENT AT OUR CLINICS (/quantum-clinic.html) is priced on the individual basis. We help with travel, lodging, escort and interpreting services at extra cost according to regular Russian rates, for example, interpreter is usually billed $177 for the first hour of translation / escort (or below one hour) with $59 for every next hour (regular 8 hours working day); while for working days with overtime (over 8 hours regular working day), client is billed at $177 for the first hour, and $94 for every next hour. We accept only our own interpreters. Clinics are RUSSIAN CLINICS. **WAVE GENOME** does NOT own any of these clinics, but offer **VISITATION** programs for getting treatments at Russian clinics.

PRIVACY POLICY (/privacy-policy.html) SHIPPING POLICY (/shipping--warranty.html) RETURN POLICY (/return-policy.html) TERMS OF SERVICE (/terms-of-service.html)

FDA DISCLAIMER: Content on this site is for reference purposes and is not a substitute for advice from a licensed health-care professional. You should not rely solely on this content, and wave genome.com assumes no liability for inaccuracies. Always read directions before using a product. Statements regarding products and services have not been evaluated by the FDA, and these products and services are not intended to diagnose, treat, cure, or prevent any disease or condition.

Payment in full is due ahead of services. No refunds and returns.

COPYRIGHT NOTICE: Text, graphics, and HTML code on wavegenome.com belong to Irene Caesar, Ph.D., their sole author, and are protected by US and International Copyright Laws, and may not be copied, reprinted, published, translated, hosted, or otherwise distributed by any means without explicit permission from Irene Caesar, Ph.D., and in violation of the International Laws for citing references. Any citation of the wavegenome.com website should be accompanied with the proper reference to Irene Caesar, Ph.D. Any citation exceeding the length, permissible for the reference length by the International Copyright Law, will be considered to be a violation of the International Copyright Law.

© IRENE CAESAR, PH.D., FOUNDER AND SOLE OWNER OF WAVEGENOME 2010-2017 ALL RIGHTS RESERVED

WAVE GENOME

www.wavegenome.com

www.heavenlyharmony.co.za www.orgoniseafrica.com www.rejuvenatingcentre.co.za

IRENE CAESAR, PH.D.

REMOTE MANAGEMENT OF BIOSYSTEMS BASED UPON THE SHAPE OF CHROMOSOMES AS DIFFRACTION GRATING FOR FOCUSING GENETIC INFORMATION

HEALING VIA WAVE PHARMACEUTICALS
AND QUANTUM BIOCOMPUTING

Weekend Seminar - October 1-2, 2016

Jane Mackenzie's Rejuvenating Centre
Johannesburg

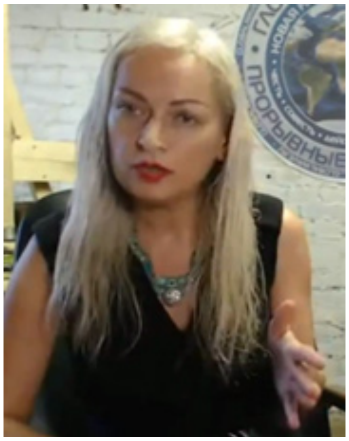

We are proud to present Irene Caesar, Ph.D., extraordinary researcher, President, Founder and Sole Owner of Wave Genome (founded 2010) who created "phenomenal new theory" of genetics that perfectly explains Autism and other degenerative diseases (Dr. Jay Dubinsky, co-owner of EGG PHARMA) via the application of Geometrical Optics to Chromosomes.

.

Dr. Irene Caesar is an Exclusive Sales and Marketing Representative of the Institute for National Security in the field of Quantum Leap technologies applied to health and rejuvenation, Co-founder of the Matrix City Consortium (2012) with the Institute for National Security in Moscow for building self-sufficient cities with the help of Russian Quantum Leap technologies, Co-founder of Quantum Biointernet via the remote laser signal (2013) with the Krasnoyarsk Scientific Group, and Colonel of the Irkutsk Cossack Military. Dr. Irene Caesar was awarded the Medal for Service to Russia (2015).

In her July 23, 2016 lecture at the International Symposium of the Thonburi University in Bangkok, Thailand, Dr. Irene Caesar said: “We are standing at the greatest turning point in the history of humankind, called Quantum Leap. Quantum Leap technologies allow us remote management of biosystems, climate and geophysical processes and unlimited access to energy and resources – the systems’ management superior to Global Artificial Intellect.” Quantum Leap gives us cure of “incurable diseases” and radical rejuvenation, so that enhanced man in enhanced environment can live enhanced life of any life span he chooses.

Dr Irene Caesar will be giving her Seminar at Jane Mackenzie’s (MBA, B Pharm) Rejuvenating Centre in Johannesburg, with the assistance of Georg Ritschl, Wave Genome’s Exclusive Representative in the South African Republic. Jane Mackenzie is Wave Genome’s partner, and a leading Rejuvenation Practitioner focusing on epigenetics, nutrigenomics and homotoxicology. Georg Ritschl is the owner of Organise Africa, and a leading producer of Orgonite on the global market.

Dr Irene Caesar is bringing to the South African Republic the World Leading Russian Rejuvenation technology, still classified in the West. Wave Genome technology was created for the first persons in the State, military commanders, diplomats, Special Forces and Olympic Team.

Wave Genome offers:

(1) **MINI-TESLA PSI-GENERATOR**

MINI-TESLA PSI-GENERATOR consists of two parts: (a) Chip (in casing) structured by laser on nanolevel with client’s biohologram recorded by laser from client's childhood photograph; for the **ULTRA MINI-TESLA PSI-GENERATOR**, the casing itself is structured by laser on nanolevel with the biohologram of the immune system rejuvenation **E-DRUG** (electronic drug); (b) flash drive with the complete **DIGITAL PHARMACY** (software) consisting of the digitized bioholograms of all essential bodily organs and physiological systems. Both of these components are not available elsewhere on the global market. **MINI-TESLA PSI-GENERATOR** is a certified protection against V2K (Voice to Skull), and all other kinds of psychotronic weapons, and harmful radiation. THERE ARE NO ANALOGUES ON THE GLOBAL MARKET, or devices that provide actual protection.

(2) **AURA PSI-GENERATOR**

AURA PSI-GENERATOR is a large-size device for radical rejuvenation and enhancement of physical, mental and psychic abilities. **AURA PSI-GENERATOR** comes with revolutionary diagnostics, not available elsewhere, which detects all malfunctions in your body, inaccessible for any conventional diagnostics, via using formerly super secret military technology, which was created by Soviet scientists for detecting enemy in the outer space. **AURA PSI-GENERATOR** has the complete **DIGITAL PHARMACY**, consisting of digitized bioholograms for every essential bodily organ and physiological system, e.g., immune system, endocrine system, etc. For example, complete rejuvenation for clients with diabetes type two takes 10 sessions. Each session is about 2 hours. **AURA PSI-GENERATOR** detects whether a person got any implants, including Nanobots (computers of the molecular size), or whether a person is under Psychotronic attack. After detection, **AURA PSI-GENERATOR** destroys microchips of any kind and protects the client from the Psychotronic or Psychic attack.

Text on the Quantum Leap Technology of Laser Polarized Holography by Wave Genome:
© Irene Caesar, Ph.D. 2016 www.wavegenome.com / mail@wavegenome.com

Jane Mackenzie, MBA B Pharm
Wave Genome’s Partner
www.janemackenzie.co.za / info@janemackenzie.co.za

Georg Ritschl
Wave Genome’s ZAR Representative
https://www.orgoniseafrica.com / orgonisehelpdesk@gmail.com

BOOKING: www.janemackenzie.co.za

(/)

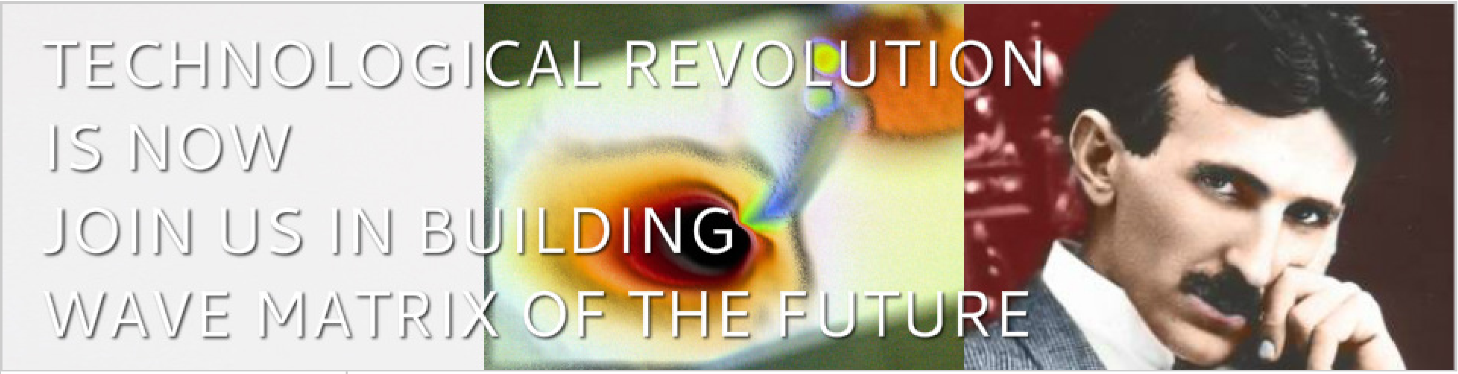
TECHNOLOGICAL REVOLUTION
IS NOW
JOIN US IN BUILDING
WAVE MATRIX OF THE FUTURE

HOME / ГЛАВНАЯ (/)

SERVICES / УСЛУГИ (/SERVICES--105910571051105910431048.HTML)

PRODUCTS / ТОВАРЫ (/PRODUCTS--105810541042104010561067.HTML)

CONTACT / КОНТАКТ (/CONTACT--1050105410531058104010501058.HTML)

TESLA PSI-GENERATORS WITH COMPLETE DIGITAL PHARMACY - THE FRONTIER OF QUANTUM GENETICS

TESLA
PSI-GENERATOR

You can place an order via filling up **THE ORDER FORM (/order-form.html).** Prices differ depending on the service, client's condition, and disease.

Models of **TESLA PSI-GENERATOR**S differ by the size of the chip (the number of layers). The chip is multilayered and it takes longer than one day to produce just one layer. Our **SUPER**, **ULTRA** and **ULTIMATE PSI-GENERATORS** has twice as much layers than our **BASIC TESLA PSI-GENERATOR**. Production of our **ULTRA** and **ULTIMATE MINI-TESLA PSI-GENERATORS** takes about or longer than one month.

Each model comes in **GENERIC** and **INDIVIDUAL** version. Generic version has the universal human biohologram (aka DIGITAL CELL) encoded via laser into the chip on nano level. And the individualized version has the personal human biohologram encoded via laser into the chip on nano level.

GENERIC TESLA PSI-GENERATORS are produced for large families, who would like to share the same **TESLA GENERATOR** for all members of the family. Generic Human Biohologram, aka DIGITAL CELL, is the result of decades of research, and is the foundation for our diagnostics procedures.

INDIVIDUAL TESLA PSI-GENERATORS have individual biohologram of a client in the chip, being encoded via laser onto the chip on nano level. Individual biohologram is recorded from your childhood photograph (or navel blood / placenta) via laser, digitized, diagnosed, and encoded into the chip via laser.

INDIVIDUAL TESLA PSI-GENERATORS are produced for the ultimate protection against genetic and psychotronic warfare.

Each chip (inside a pendant) comes with the **COMPLETE DIGITAL PHARMACY** as software on a flash drive with 44 modules for every essential bodily organ and physiological system. Market value is 35,000 Euros only one module.

COMPLETE DIGITAL PHARMACY SOFTWARE PROGRAM

All our models of TESLA PSI-GENERATORS (chip) come with a separate flash drive with the complete **DIGITAL PHARMACY** software program for every essential bodily organ and physiological system **(PSI-PROGRAM**) to run on your computer in order to generate **E-DRUGS** / **BIOELECTRONIC DRUGS** (aka wave matrices / bioholograms) for every health condition.

SECOND-BIRTH TESLA PSI-GENERATOR

SECOND-BIRTH TESLA PSI-GENERATOR

(a) **has a more advanced, second generation chip encoded by laser on nano level with the most valuable bioholographic information on the global market: EMBRYONIC STEM CELL;**

(b) has a new technology of encoding the chip by the newly invented laser (patented);

(c) layers of the chip are made of a new material of organic nature more advanced than our other chips;

(d) its efficacy is proven by the clinical trials at the Medical Clinic of the Moscow Bauman Russian State Technical University, the most important Science University in Russia, of MIT level. twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR.**

(e) has
either **GENERIC** or **INDIVIDUAL** biohologram.

(f) provides the ultimate modulation for the scalar wave diffraction grating for the most precise focus / zero center of the biohologram (aka 'wave crystal") in the process of rejuvenation.

GENERIC MATRIX MINI-TESLA GENERATOR - $3000

INDIVIDUAL MATRIX MINI-TESLA GENERATOR - $3500

BASIC-X TESLA PSI-GENERATOR

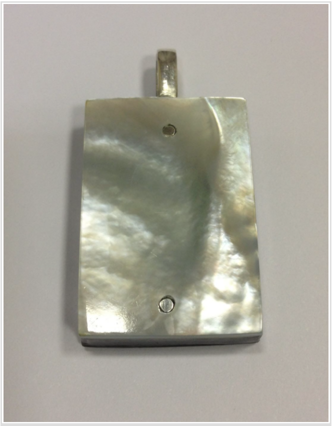

SUPER TESLA PSI-GENERATOR

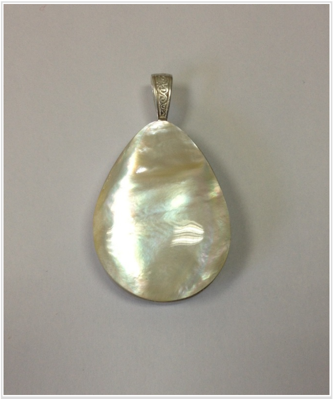

ULTRA TESLA PSI-GENERATOR

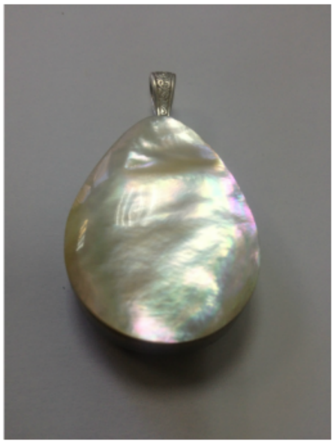

ULTIMATE TESLA PSI-GENERATOR

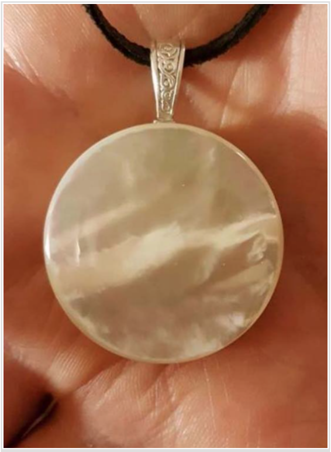

BASIC TESLA PSI-GENERATOR (chip)

(a) has twice less layers of the chip than our other models**.**

(b) has either the **GENERIC** or **INDIVIDUAL** biohologram of a healthy human in the chip, being encoded via laser onto the chip on nano level.

(c) **BASIC TESLA PSI-GENERATOR** has both sides of the casing for the chip made of Real Solid Pearl.

(d) **BASIC-X TESLA PSI-**

SUPER MINI-TESLA GENERATOR (chip)

(a) has twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR.**

(b) has either **GENERIC** or **INDIVIDUAL** biohologram.

(c) Both sides of the casing for the chip are made of Real Solid Pearl.

GENERIC SUPER MINI-TESLA GENERATOR - $1179

INDIVIDUAL SUPER MINI-TESLA GENERATOR - $1768

MATRIX TESLA PSI-GENERATOR

MATRIX TESLA PSI-GENERATOR

(a) has twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR.**

(b) has either **GENERIC** or **INDIVIDUAL** bioh

ULTRA MINI-TESLA GENERATOR (chip)

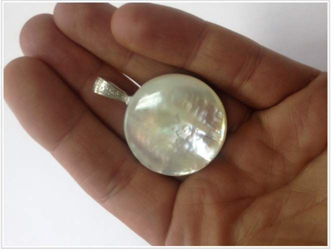

ULTIMATE MINI-TESLA GENERATOR (chip)

(a) has twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR.**

(b) has either **GENERIC** or **INDIVIDUAL** biohologram.

(c) One side of the casing for the chip is made of Real Solid Pearl, and another side of casing is made of 3D printing plastic, encoded with the

GENERATOR has one side of the casing for the chip made of Real Solid Pearl, and another side of casing made of 3D printing plastic, encoded with the immune system rejuvenation **E-DRUG** (**BIOELECTRONIC DRUG** / **DIGITIZED BIOHOLOGRAM**). This makes **"TWO-CHIPS IN ONE"** solution.

GENERIC BASIC MINI-TESLA GENERATOR - $856

INDIVIDUAL BASIC MINI-TESLA GENERATOR - $1179

INDIVIDUAL BASIC-X MINI-TESLA PSI-GENERATOR - $1534

ologram.

(c) Both sides of the casing for the chip are made of 3D printing plastic, encoded with the immune system rejuvenation **E-DRUG** (**BIOELECTRONIC DRUG** / **DIGITIZED BIOHOLOGRAM**). This makes **"TWO-CHIPS IN ONE"** solution.

GENERIC MATRIX MINI-TESLA GENERATOR - $2122

INDIVIDUAL MATRIX MINI-TESLA GENERATOR - $2947

(a) has twice as more layers of the chip than the **BASIC MINI-TESLA PSI-GENERATOR.**

(b) has either **GENERIC** or **INDIVIDUAL** biohologram.

(c) One side of the casing for the chip is made of Real Solid Pearl, and another side of casing is made of 3D printing plastic, encoded with the immune system rejuvenation **E-DRUG** (**BIOELECTRONIC DRUG** / **DIGITIZED BIOHOLOGRAM**). This makes **"TWO-CHIPS IN ONE"** solution.

GENERIC ULTRA MINI-TESLA GENERATOR - $1768

INDIVIDUAL ULTRA MINI-TESLA GENERATOR - $2358

RNA **E-DRUG**(**BIOELECTRONIC DRUG** / **DIGITIZED BIOHOLOGRAM**) for the ultimate protection against psychotronic and genetic warfare. This makes **"TWO-CHIPS IN ONE"** solution.

GENERIC ULTIMATE MINI-TESLA GENERATOR - $2358

INDIVIDUAL ULTIMATE MINI-TESLA GENERATOR - $2712 (promotional price)

MINI-TESLA GENERATOR

MANUAL FOR MINI-TESLA PSI-GENERATOR

ABOUT THE DEVICE

The device consists of two parts: a chip in a pendant and a computer program (software) on a flash drive.

CHIP:

The user wears a pendant with a chip inside on his/her neck at the level of the solar plexus. We have chips of first and second generation. First generation chip is an artificial polymer crystal with electret properties, which has two modules, depending on how you turn it towards your body: open module (reinforcing the impact of a user on the environment) and closed (protecting a user from the harmful effects of the environment). Second generation chip (more advanced) has the entirely holographic technology, and functions through the focus / zero center of the scalar wave diffraction grating in the hologram encoded into the chip via laser on nanolevel.

Chips are structured by laser at the nano level. The laser writes to the chip at the nano level client's digitized bioholograms, which is read by the laser from the childhood photograph, placenta or umbilical cord blood of the client for the individual generators, or the biogologram of a average healthy person for the generic generators.

When user is wearing the chip, the electret is emitting the wave crystalline structure (hologram or matrix) upon the user, and makes his chromosomes repeat the wave crystalline structure of a healthy person during cell division.

The laser and chip itself use the principle of Kozyrev Mirrors. The Universe consists of energy for 93%, and of particles only for 7%, and particles are the concentrated energy. The universe is holographic, that is, it is entirely in its every Matrix Point. Wave Hologram or Matrices exist in two forms: in the form of a Torus (simultaneously present from the infinitely small to the infinitely large dimensions) and in the form of the Wave Crystal. Man exists in the form of a Torus during sleep. During waking hours, the person is in the form of a Wave Crystal.

Wave Crystals are formed in the structures, which reproduce in their shape the Kozyrev Mirrors. Kozyrev Mirrors reflect any external wave signal, and, therefore, are called "mirrors". These structures neutralize any external linear signal, and refract it towards their Zero Center or focus. Neutralization of any external signal happens due to the reflection of the wave and the polarization, and the creation of a so-called standing (scalar wave). Trough of the back-running wave neutralizes the peak of the forward-going wave, when the initial wave is reflected. This provides protection from harmful electro-magnetic and other technogenic or anthropogenic influences.

As a result of the standing (scalar) wave, there emerges the interference (diffraction) grating, which refracts any next linear signal line towards the center (focus) of the Wave Crystal.

It is necessary to know this, because all the information in the universe is transmitted and received only through the Zero Center of the Wave Crystal. This is analogous to the work of our eye. We see only because there is the gathering

"PSI" stands for "psychotronic", that is, for the technology that is offering technical means for managing biosystems.

Every **MINI-TESLA psi-generator** comes with

(1) a **CHIP**, structured with a laser on a nano-level with the digitized Wave Matrix / Biohologram of healthy human or your own individual wave matrix in childhood (recorded from your childhood photograph), and cased either in pearl or 3-D printing plastic, or combination of these two materials. You wear it on a chain on your solar plexus all the time, taking it off only for bathing;

(2) and psi-program - a flash drive with the software that has the entire **DIGITAL PHARMACY** generated as the holographic / granulated signal (aka "white noise" - signal beyond your auditory perception) and encoded in music. You are getting the complete **DIGITAL PHARMACY** for your lifetime and for every system and vital organ in your body. You run this software at least once a day for best results. If you are sick, you run this software as often as possible.

All sales upon agreeing to c**lient consent conditions** (/client-consent.html), no refunds.

(/order-your-mini-tesla-generator-russian.html)

Нажмите на Русский флаг для перемещения на версию сайта на Русском языке

Refer to our **FAQ (Frequently Asked Questions) (/mini-tesla-generator-faq.html)** for any concerns.

IMPORTANT: CHIP has two modes : OPEN and CLOSED. In the **OPEN MODE**, you increase your information-wave impact upon people. The **CLOSED MODE** is the **SHIELD MODE**, which protects you from any psychotronic attack, including V2K (voice to skull) and cancer guns, subliminal advertising and psychic attacks. Those customers who are endangered with these attacks can wear their **CHIP** in the protective mode nonstop. **PLEASE NOTE**: Plastic towards your body is a **SHIELD MODE**. For casing which has both sides from plastic, the **SHIELD SIDE** is thinner so that the chip is noticeable beneath plastic. For casing which has both sides from Pearl, the darker side of Pearl towards your body is the **SHIELD MODE**.

DIGITAL PHARMACY SOFTWARE PROGRAM (/digital-pharmacy-software--1057105410601058-10621048106010561054104210541049-1040105510581045105010481.html)

of all the external linear signals in the focus of the lens (Wave Crystal). If focus is not achieved, and the refraction of the signal takes place before or behind the center of the lens, we have developed the long-sightedness (hyperopia), or the long-sightedness (myopia), and we don't see well. Receiving information by both our eye, and our DNA takes place strictly in accordance with the laws of geometrical optics. The technology used for creating the chip is called "Polarized Holography".

It is possible to record and transmit only a holographic signal. To record Wave Crystals (Bioholograms or Matrices) of users, we use devices that can record and transmit holographic signal - He-Ne laser with internal mirrors. These lasers convert any external linear signal into a holographic signal, i.e., changing the nature of the initial signal. For example, our bones produce stem cells only because their shape and structure is analogous to the laser tubes, which "involute" a signal.

Similarly, our DNA is able to transmit and receive information, only because, in its structure, it is the Wave Crystal. Two strands of DNA are literally the standing (scalar) wave, since in the first strand of DNA, a wave goes in the direction opposite to the wave going in the second strand of DNA. Moreover, DNA coils into the crystalline structure, which literally reproduces the interference / diffraction grating of the standing wave in the Kozyrev Mirror, in order to most accurately refract any external linear signal toward the Zero Center / Focus of its Wave Crystal.

We live only thanks to cell division. During cell division, chromosome is formed. A chromosome is formed when two sisters chromatids unite together into a cruciform structure. Chromosomes come in two main shapes: acrocentric and metacentric. In acrocentric chromosomes, two endings of the sisters chromatids are shorter than the other two endings. On the contrary, metacentric chromosome has all the endings the same, so that it is absolutely symmetrical. Metacentric chromosome resembles the quadratic / symmetrical cross (Crux Quadrata / Gammadion), as well as the lens of a well-seeing eye. The same gene in apes, in children with Down syndrome, and wild horses of Przewalski has the acrocentric chromosome, whereas in humans, normal children, and Arabian horses, the same gene has the metacentric shape of the chromosome. That is, the functionality of the species directly depends upon the shape of chromosome, that is, upon the laws of the Polarized Holography of Wave Crystals.

Our device is centering and focusing the chromosomes during cell division, giving them the metacentric form.

Because the entire universe is entirely in its every Matrix Point, each Matrix Point is not simply different from any other Matrix Point, but is unique

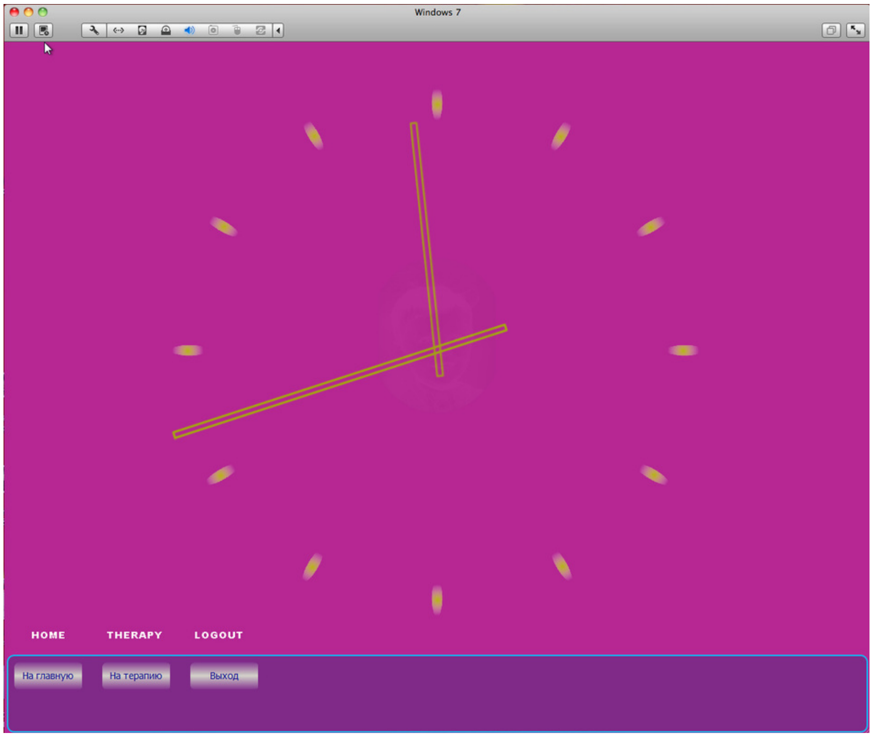

Click here (/digital-pharmacy-software--1057105410601058-10621048106010561054104210541049-1040105510581045105010481048.html) to see screen shots of the DIGITAL PHARMACY SOFTWARE PROGRAM.

Produced according to TU 9444-021-42844321. Certificate # 77, FC.19.944. P.2455.11.00

Our **MINI-TESLA GENERATOR** is used by Russian Special Forces and Olympic Team.

MINI-TESLA GENERATOR is a CHIP (artificial / polymer crystal), structured by laser on a nanolevel, as a collector and retransmitter of wave structures (Matrices or Bioholograms), which functions as a dispenser for WAVE MATRICES (http://www.wavegenome.com/wave-pharmaceuticals.html) within the revolutionary paradigm of LASER POLARIZED BIOHOLOGRAPHY, which is the true WAVE PHARMACEUTICALS (/wave-pharmaceuticals.html) **-- information technology of a new generation. Generator or retransmitter functions like a vibrator, which emits micro vibrations of special kind - scalar wave diffraction grating.**

WAVE MATRICES (http://www.wavegenome.com/wave-pharmaceuticals.html) can be generic or individualised. Wave matrices can be of a human as a whole, of specific bodily system, or of a specific medication.

Comes with the software that has 44 modules of DIGITISED WAVE MATRICES (http://www.wavegenome.com/wave-pharmaceuticals.html) for every bodily system and every vital organ in your body.

E-VIAGRA

Individualized ELECTRONIC CODE for the rejuvenation of your sex function.

FAQ Frequently Asked Questions (/mini-tesla-generator-faq.html)

READ INFORMATION LEAFLET FOR THE MINI-TESLA GENERATOR

(COMES WITH THE UNIT) (/information-leaflet-for-mini-tesla-generator.html)

Over the course of decades, our institute participated in the research in the field of information technologies of a new generation, and their commercialization. The institute created the system of information-energy correction, which is the most advanced in the world, and is the two-way connection through the energy-information cloud. In Russia, our **MINI-TESLA generator** is used by Special Forces and the Olympic Team in order to help them cope with stress and enhance their performance.

(© Irene Caesar, 2012). Each Matrix of each person is the uncreated and indestructible unique nonlocal Wave Matrix. If we have two copies of a unique non-local Wave Matrix, we can instantaneously transfer information from one copy of a unique non-local Wave Matrix to another copy. In this way, birds and insects share one unique non-local Wave Matrix for their entire species, and instantaneously transmit information between all individuals of a given species. But each man has his own unique nonlocal Wave Matrix. If you order an individual generator, you get your unique non-local Wave Matrix with the most clear focus / center of the Wave Crystal. Your generator transmits upon your chromosomes your unique uncreated and indestructible information from the noosphere of the universe. If you order the generic (universal) generator, its action is limited, but it, nevertheless, helps your chromosomes to get centered and focused.

We also offer the remote therapy by the distant laser signal that instantly, through Quantum Non-locality, affects your genetic apparatus according to the principle described above. The user uses his/her individual generator as an antenna and a filter for the remote laser signal.

DIGITAL PHARMACY SOFTWARE

The user uses the computer program (software) every day once a day in a normal state of health, and several times a day, depending on the complexity of his/her state of health. When working with the computer program of the generator, the user must turn his/her generator to the open module.

Computer program of the generator is a comprehensive Digital Pharmacy of the digitized Bioholograms (Wave Crystal structures or Matrices) of all healthy vitally important organs and physiological systems. The user tales a self-test for self-therapy, and the program generates for him the granular (holographic) signal, which is encoded in the music.

Self-therapy is the binaural therapy, which uses the Kozyrev Mirrors principle alike the chip inside the generator. Binaural therapy is based on the effect of the stereo signal to both ears, so that signal is delayed in one ear with respect to the signal applied to the second ear. It is in this way that the Kozyrev Mirrors effect is created, when human skull itself becomes the Kozyrev Mirror. The signal is reflected; there emerges a standing (scalar), the wave interference / diffraction grating; and the signal is refracted toward the Zero Center / focus of the wave crystal.

Also, the user can order any pill or tincture in the form of a software module of the Digital Pharmacy. Digital Pill module is attached to the Digital Pharmacy software, and the user receives the Digital Pill as the granular (holographic) signal encoded in the music. We record the Digital Pill via a laser and as the wave structure (Hologram or Matrix), which is afterwards gets digitized.

DIGITISED WAVE MATRICES (http://www.wavegenome.com/wave-pharmaceuticals.html) also can be either generic or individualised, and of a human as a whole, of a specific bodily system, or of a specific medication.

(/uploads/3/5/1/1/35112207/3881928_orig.png)

The first-generation model of MINI-TESLA PSI-GENERATOR, which combined the chip with the flash drive containing the DIGITAL PHARMACY software.

BASIC MINI-TESLA GENERATOR emits generic **HUMAN CELL HOLOGRAM**, recorded by the laser from healthy human cells and bodily systems, and then used to structure the crystal via the laser on nanolevel.

SUPER, ULTRA, MATRIX AND ULTIMATE MINI-TESLA PSI-GENERATORS emit individual **HUMAN CELL HOLOGRAM** in the form of your childhood wave matrix, which is recorded from your childhood photograph when you were a healthy child. We also need your date of birth. We use the complex systematic approach to read the bioholographic information from your photograph, including the application of laser.

After your matrix is recorded from your photograph, we run a special diagnostic program in our Institute to reveal the defects in your matrix, and to correct them.

If no childhood photograph is available, we can use your adult photograph, since the Institute can diagnose **WAVE MATRICES (http://www.wavegenome.com/wave-pharmaceuticals.html)** and correct them.

We can "record" wave matrices / bioholograms upon various media, and not only on the electret-based chip of the MINI-TESLA PSI-GENERATOR. Examples of such media are CDs that are used "as such", and not for extracting the digital files from them.

We can combine your childhood wave matrix with wave medications.

We can tune up your childhood matrix.

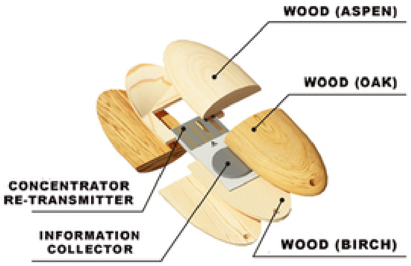

(/uploads/3/5/1/1/35112207/3725345_orig.jpg)

The first-generation model of MINI-TESLA PSI-GENERATOR, which combined the chip with the flash drive containing the DIGITAL PHARMACY software.

MINI-TESLA psi-generator cased in pearl

The second-generation model of MINI-TESLA PSI-GENERATOR, with casing of chip made from Pearl and Wood. Now this model is offered in Pearl for the BASIC MINI-TESLA GENERATOR (chip structured by the laser with the generic biohologram of a healthy human on nano level).

The device is in fact very necessary in modern times when we are bombarded by the harmful electro-magentic radiation.

The device is designed in such a way that you can turn it either with darker-colorer side to yourself, and, in this case, the generator is programmed to open your biofield; or to turn it with the lighter-colored side (aspen side in generators coated in wood) to yourself, and, in this case, the generator becomes your psychotronic "firewall".

IF you are a jet setter, you are receiving a huge amount of radiation each time you fly. Radiation is one of the causes of cancer. That is why our MINI-TESLA GENERATOR is indispensable for jet-setters.

If you get sick, e.g., infected with a virus or bacteria, your wave imprint / data received at the institute will immediately alert the system, and we will correct the situation via sending you the necessary correction to your wave-diagram (available only for **ULTRA MINI-TESLA PSI-GENERATORS)**

Digital Pills do not have the danger of overdose and side effects.

ОЛИМПИЙСКАЯ КОМАНДА РОССИИ
RUSSIAN OLYMPIC TEAM

(/uploads/3/5/4/1/35112207/7731776_orig.gif)

Our MINI-TESLA GENERATOR is used by Russian Olympic Team.

ВОЙСКА
СПЕЦИАЛЬНОГО НАЗНАЧЕНИЯ

(/uploads/3/5/4/1/35112207/1566232_orig.gif)

Our MINI-TESLA GENERATOR is used by Russian Special Forces.

We can tune up your childhood matrix in such a way that some specific bodily systems or organs are stimulated.

SELF-MAINTENANCE

You run software with **PSI-PROGRAM** on your computer in order to perform the self-maintenance as often as you need. **PSI-PROGRAM** has 44 modules for every bodily system and every vital organ in your body. These 44 modules were recorded by the laser from healthy human cells and bodily systems, and, then, digitised.

Software has (1) the chart of bodily systems and organs to choose from for self-maintenance; (2) the scale of self-assessment from the worst to the best; (3) and two biofeedback tests, based upon graphology.

QUANTUM COMPUTER

Besides conventional tests, Institute developed a completely new way of recording and transmitting information. We call it **QAUNTUM BIOINTERNET**.

When you run the **PSI-PROGRAM**, the spin states of the software and the spin states of your crystal / chip in the generator constitute one and the same torus (hologram). There is the instantaneous biofeedback / bioresonance two-way communication between the running software and the crystal inside the **MINI-TESLA PSI-GENERATOR.**

The crystal / nano chip in the **MNI-TESLA PSI-GENERATOR** and the **PSI-RPOGRAM** software (during its operation on your computer) constitute one integrated **QUANTUM COMPUTER**. The spin states of the crystal / nanochip in the generator and the spin states of the software coincide / resonate or oppose each other, according to the principles of the polarised holography. The principles of the polarised holography are expressed by geometrical optics.

The software reads via its own spin states the information off the crystal (its spin states), and corrects it. Then, the crystal tunes via the bioresonance of the spin states to the hologram, corrected by the operating software.

In this way, the crystal / nanochip inside the generator self-corrects itself based upon the operation of this **QUANTUM COMPUTER**.

MAINTENANCE BY WAVE GENOME LLC

(http://www.wavegenome.com/wave-genome---prices.html).

Some bacteria and viruses are considered to be incurable because they leave a wave imprint in the body, even after they are cleared up on the biochemical level. And, then, these wave imprints can literally rebuild the virus and bacteria from scratch. This same principle is seen in the classic physics experiment involving passing photons and molecules through two slits simultaneously. That is why it is necessary to protect and cure oneself on the wave level.

(/uploads/3/5/4/1/35112207/5220735_orig.jpg)

The first-generation model of the MINI-TESLA PSI-GENERATOR, which was approximately 2 inches long, and which was worn at your solar plexus the same way as the new model of the MINI-TESLA PSI-GENERATOR. Both models are invisible when you wear it. The first-generation model had chip and flash drive with the software combined in one piece.

MAINTENANCE VIA QUANTUM INTERNET BY WAVE GENOME LLC

In the case of **QUANTUM INTERNET**, you send your information **REMOTELY** and **WIRELESSLY**. At the WAVE GENOME LLC, the laser reads the feedback from your wave matrix in real time, using your digitised wave matrix on record at the Institute.

According to the holographic principle, the universe is holographic, meaning that it is entirely in its every matrix part. The holographic principle has three important implications:

The first implication of the holographic principle postulates the **UNIQUENESS** of every wave matrix. if the universe is entirely in its every matrix point, then, every matrix point is not simply different from any other matrix point, but is unique.

The second implication of the holographic principle postulates the **NON-LOCALITY** of every wave matrix. If the universe is entirely in its every matrix point, then every wave matrix is simultaneously present in the infinite number of dimensions via its tokens. Though there are infinitely many tokens of the same non-local wave matrix, all of them constitute one integrated **SYSTEMATIC WHOLE**.

The non-local holographic nature of wave matrices also means that the wave matrix has the infinite set

WAVE GENOME offers everyday **MAINTENANCE** of **WAVE MATRICES** with the purchase of the **ULTRA MINI-TESLA PSI-GENERATOR** and **SUBSCRIPTION PACKAGES.**

The maintenance consists in receiving the necessary correction of your bio-field from the institute.

To obtain MAINTENANCE from the Institute, you send in your information. The Institute tracks down the changes in your **WAVE MATRIX** and makes necessary corrections. Corrections also account for the changes in the earthly electromagnetic activity, solar and lunar activity, geophysical activity, weather anomalies, and predicted social unrest.

There are two ways to send in your information. You can send your biofeedback information either **DIGITALLY** through the conventional internet, or via **QUANTUM INTERNET**.

DIGITAL MAINTENANCE BY WAVE GENOME

In the case of **DIGITAL MAINTENANCE BY WAVE GENOME LLC, WAVE GENOME LLC** corrects the digital files in the software of your **PSI-PROGRAM**.

When you connect your **MINI-TESLA GENERATOR** to your computer, you both receive and transmit through the internet the information, needed for your utmost performance, to and from the our facility.

Your childhood matrix, recorded onto the generator-transmitter, is your antennae that helps your body tune to the waves which are beneficial to you, and ward off harmful waves. That is why this technology is also called "the resonance technology".

Device strengthens the immune system and body's resistance to infectious and toxins;
Improves performance and cognitive abilities, including, sharpness of thinking and memory, and overall quality of life;
Protects against stress factors of any nature (technogenic, including, electromagnetic, as well as physical and psychic)
Speeds up recovery and healing
It is unique in the prevention of many diseases

Especially recommended for people who are in stressful situations: to sportsmen, businessmen, professionals with busy schedule, flights and business trips.

of waves, particles and fields within its **TORUS** -- from infinitely micro to infinitely macro dimensions. It is impossible to talk of each individual wave matrix as "a specific frequency". And it is impossible to postulate any "medium" of its existence, like "scalar wave", "aether", "Higgs / boson field", "torsions", "leptons", etc. There is no specific medium for transferring the information.

The third implication of the holographic principle postulates that there is the **QUANTUM ENTANGLEMENT** in the form of the instantaneous transmission of information between all the tokens of the same non-local wave matrix. The signal that records and transmits the **ENTANGLED INFORMATION** between different wave matrices or between the tokens of the same non-local wave matrix has specific nature. It is **HOLOGRAPHIC**. Only lasers can record and transmit holographic signal, based upon the principles of the polarised refraction holography (described in terms of geometrical optics). Thus, the principle of recording and transmitting the **MATRIX SIGNAL** is structural, and not medium-based.

Thus, it is possible to remotely and instantaneously transfer some wave information to a man via the entanglement between his body (at one location) and his photograph (at another location), as two tokens of his non-local wave matrix.

For example, it is possible to remotely lower blood sugar in a diabetic man via scanning his photograph with the laser ray, which is passing through insulin before reaching the photograph.

© Irene Caesar, Ph.D. from her book **"WAVE PHARMACEUTICALS"** (2014)

FDA DISCLAIMER: Content on this site is for reference purposes and is not a substitute for advice from a licensed health-care professional. You should not rely solely on this content, and wave genome.com assumes no liability for inaccuracies. Always read directions before using a product. Statements regarding products and services have not been evaluated by the FDA, and these products and services are not intended to diagnose, treat, cure, or prevent any disease or condition.

Payment in full is due ahead of services. No refunds and returns.

COPYRIGHT NOTICE: Text, graphics, and HTML code on wavegenome.com belong to Irene Caesar, Ph.D., their sole author, and are protected by US and International Copyright Laws, and may not be copied, reprinted, published, translated, hosted, or otherwise distributed by any means without explicit permission from Irene Caesar, Ph.D., and in violation of the International Laws for citing references. Any citation of the wavegenome.com website should be accompanied with the proper reference to Irene Caesar, Ph.D. Any citation exceeding the length, permissible for the reference length by the International Copyright Law, will be considered to be a violation of the International Copyright Law.

© IRENE CAESAR, PH.D., FOUNDER AND SOLE OWNER OF WAVEGENOME 2010-2017 ALL RIGHTS RESERVED

WAVE GENOME (/)

WAVE GEHOME IS
THE INFORMATION-WAVE
MATRIX OF THE FUTURE WORLD

HOME / ГЛАВНАЯ (/)

SERVICES / УСЛУГИ (/SERVICES--105910571051105910431048.HTML)

PRODUCTS / ТОВАРЫ (/PRODUCTS--105810541042104010561067.HTML)

CONTACT / КОНТАКТ (/CONTACT--1050105410531058104010501058.HTML)

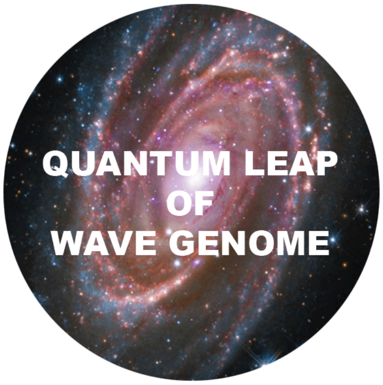

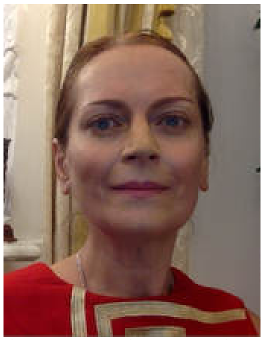

Dr. Irene Caesar, author of the theory of Wave Optics in Chromosomes as Biholography, and author of the technological model for recording and transferring the biholograms of cells and drugs via refractions codes in the scalar wave diffraction grating; the Head of the group of technical professionals to verify the technological model offered and theoretically substantiated by her.

PRESS RELEASE ON GLOBAL LAUNCH OF THE SECOND-BIRTH TESLA PSI-GENERATOR BY WAVE GENOME

JANUARY 13, 2018

© Irene Caesar, Ph.D., President, Founder and Owner of Wave Genome

Wave Genome is proud to announce its most advanced MINI-TESLA PSI-GENERATOR ever: the SECOND-BIRTH Mini-Tesla Psi-Generator.

We called the new chip the "SECOND BIRTH" generator, because the client gets the most valuable bioholographic treatment ever for rejuvenation -- the biohologram of the embryonic stem cell. The SECOND-BIRTH Biohologram is available as a chip encoded by laser on nanolevel. The clinical trials were conducted by the Bauman University Medical Center. The clinical trials had demonstrated the radical improvement in the wellbeing of the participants in the clinical trials.

In the GENERIC SECOND-BIRTH Tesla Generator, the biohologram of the embryonic stem cell is generic. It is a DIGITAL CELL. It is not the biohologram of a specific embryo, but the result of the Embryonic research based upon the Quantum Leap diagnostic of holograms according to the coherence of the geometrical optics in chromosomes.

In the INDIVIDUAL SECOND-BIRTH Tesla Generator, the biohologram of the embryonic stem cell is modulated by client's most coherent biohologram, recorded via laser from his

(/ask-our-maven-russian.html)

childhood photograph.

The INDIVIDUAL SECOND-BIRTH Tesla Generator is stronger than the GENERIC one, since it has a more precise focus due to the finer / smaller grid of the scalar wave diffraction grating. But you can use the GENERIC one for the entire family, and this is also a great benefit.

If somebody else will wear your own unique INDIVIDUAL Tesla Generator, this individual will simply not connect to the focus of your unique biohologram encoded on the chip, since this person has his/her own geometrical code for refraction in the scalar wave diffraction grating of his/her biohologram.

When the modulation is added to the carrier wave in the scalar wave diffraction grating, the resulting hologram is different from these two components taken separately.

In the biohologram, the polarization emerges due to the change of the spin in the integral torsion. Due to polarization, the external signal is reflected upon itself, so that the scalar / standing wave emerges. In the scalar wave, the peak of the forward-going wave is annulled by the trough of the same wave when it is reflected upon itself. Scalar wave creates the diffraction grating, so that any next external signal is refracted towards the focus / zero center of the biohologram (aka “wave crystal”).

Any external signal is only a trigger, and the external signal can be a psychotronic attack. This is a great benefit of the Quantum Leap Laser Polarized Bioholography. The more modulations the biohologram has, the more precise focus / zero center is achieved via refraction, and the better is your protection against the psychotronic attack.

Because the SECOND-BIRTH Tesla Generator has the most fine / small / scaled scalar wave diffraction grid among our other generators, the SECOND-BIRTH Tesla Generator provides the most precise focus / zero focus of client’s biohologram and the best defense against psychotronic attack.

Thus, with the SECOND-BIRTH Chip, you will experience the most radical rejuvenation effects, including mental and physical enhancement, a more emotional control, and a better defense against psychotronic attacks.

Those of our clients, who already own one of our MINI-TESLA PSI-GENERATORS, can either wear both generators together, or carry one of the generators in a pocket, or alternate generators.

WAVE GENOME HAS INTELLECTUAL PRIORITY FOR THE WAVE OPTICS IN CHROMOSOMES AS BIOHOLOGRAPHY

Click here (/diaschev-dipol-di-pol.html) to read on how WAVE GENOME’s technology differs from Dr. Gariaev’s and Dr. Diaschev’s technology (DIPOL or DI-POL Generator).

Prof. Alexander K. Volkov, Ph.D., Medical Sciences, Director of the Scientific, Medical, Technological and Educational Center of the Bauman Moscow Russian State Technical University, who headed the group of medical professionals to verify the technological model offered by Dr. Irene Caesar.

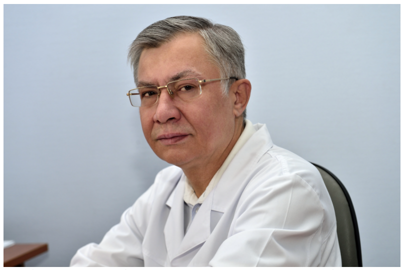

Prof. Vladimir I. Sinopalnikov, Ph.D. Medical Sciences, The Renowned Physician of the Russian Federaration, Deputy Director of the Scientific, Medical, Technological and Educational Center of the Bauman Moscow Russian State Technical University, who participated in the group of medical professionals to verify the technological model offered by Dr. Irene Caesar.

НАУЧНО-ОБРАЗОВАТЕЛЬНЫЙ МЕДИКО-ТЕХНОЛОГИЧЕСКИЙ ЦЕНТР
МГТУ ИМ. Н. Э. БАУМАНА

Scientific, Medical, Technological and Educational Center of the Bauman Moscow Russian State Technical University, in possession of the cutting edge lasers, and other technology.

Bauman Moscow Russian State Technical University, major Science University of Russia, of MIT level.

PRIVACY POLICY (/privacy-policy.html)

SHIPPING POLICY (/shipping--warranty.html)

RETURN POLICY (/return-policy.html)

TERMS OF SERVICE (/terms-of-service.html)

FDA DISCLAIMER: Content on this site is for reference purposes and is not a substitute for advice from a licensed health-care professional. You should not rely solely on this content, and wave genome.com assumes no liability for inaccuracies. Always read directions before using a product. Statements regarding products and services have not been evaluated by the FDA, and these products and services are not intended to diagnose, treat, cure, or prevent any disease or condition.

Payment in full is due ahead of services. No refunds and returns.

COPYRIGHT NOTICE: Text, graphics, and HTML code on wavegenome.com belong to Irene Caesar, Ph.D., their sole author, and are protected by US and International Copyright Laws, and may not be copied, reprinted, published, translated, hosted, or otherwise distributed by any means without explicit permission from Irene Caesar, Ph.D., and in violation of the International Laws for citing references. Any citation of the wavegenome.com website should be accompanied with the proper reference to Irene Caesar, Ph.D. Any citation exceeding the length, permissible for the reference length by the International Copyright Law, will be considered to be a violation of the International Copyright Law.

© IRENE CAESAR, PH.D., FOUNDER AND SOLE OWNER OF WAVEGENOME 2010-2017 ALL RIGHTS RESERVED

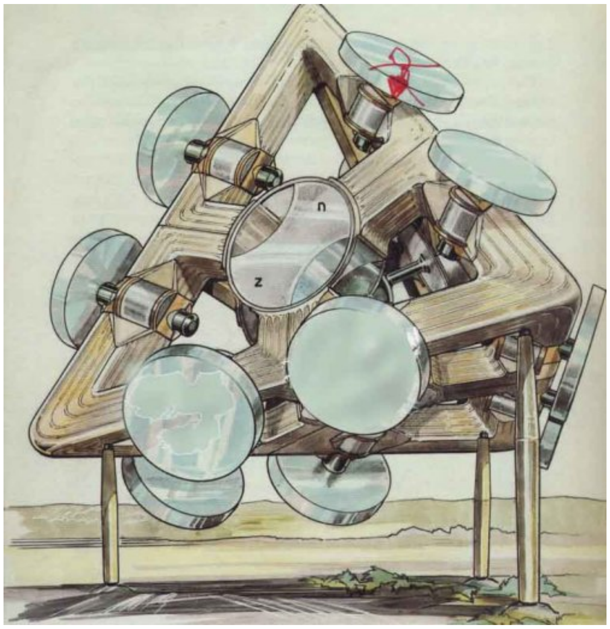

Technical information given from the IARGA contactteam.

their describtions on the matter of
TIME

- AND MORE

picture is described later

12. The analog carrierfield model, an instruction model.

The cosmic carrier field of all space and matter performs a sixfold function, of which some have already been discussed in gereral way.

1. The carrier function of an infinitely strong forcefield blocked in itself contains the fluidum conducting all waves without dissipation and maintained to eternity.
2. The orientation of the time-axes of the universes with respect to nothing provides the structure of the carrier-field of three dimensions, the bounded Love-will, creating our life and existence.
3. The time synchronisation function locks all universes in the iron grip of the law of cosmic counter-balance from which no atom can escape.
4. The feedback principle of time to the total mass of the universe is the function of the coupling-field by counterbalance (and interference) of the time-bound waves.
5. The function of the timeless primeval atom-field, the atom-principle, as the transmitter program containing the code of matter of which all atoms are remote manifestations.
6. The function as bearer of the absolute ALL-knowledge and goal consciousness, manifesting itself as the laws of nature and as the intelligent goal of life: the spirit of God.

The last function concerns the existential aspects, which have been discussed in the previous section of this book.

The first analog model is a mechanical instruction model explaining the cosmic carrier field by means of symbols. This model was built on another life-supporting planet on a high mountain and was many meters high. It was the gathering point for contemplation of that intelligent race for devotion to its Creator.

It was the symbol by which the majority of the race acknowledged the goal of Creation, simultaneously the sign that the cosmic isolation had been lifted. The other function of the carrier-field will be discussed with the help of this mechanical instruction model.

The analog model consisted mainly of the six edges of a huge equilateral triangular pyramid built out of solid soft-iron bars. The many metres thick edges of the pyramid contain an impressive number of iron bars parallel to the direction of the edges and cast in a kind of plastic resin following a process similar to that of casting concrete(betong). So a mainly iron-frame was formed especially suitable for conduction of magnetic pulses of high power without much loss.

So far the carrying fuction of the cosmic time-field, which primarily maintains the timeless waves is symbolized in the model. The magnetic pulses in the edges of the pyramid run backwards and forwards like the timeless waves in reality.

With respect to the other edges of the pyramid, the magnetic pulses are phase-shifted. The problem is that up to now the forward propagation (forplantning) in time has been missed out. In fact we look for the equivalent of the two little wagons on the vibrating rail. The solution can be found in the symbol of the rotating flywheel, as *with the atom its rotation is continuous in time, it is in fact time propagation.* Therefore the iron frame of the pyramid is discontinued at two places in each of the six edges of the pyramid. At this discontinuation point a linear magnetic impulse motor is installed, on the axis of which a large flywheel is mounted.

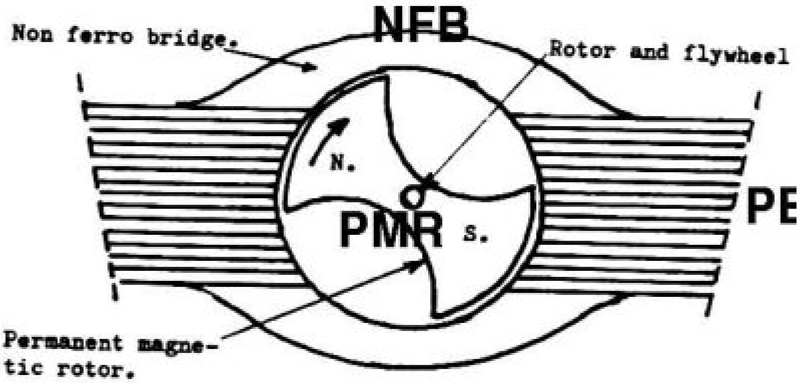

NBF=NonFerroBridge

PE=PyramidEdge and soft iron bars

PMR=PermanentMagneticRotor

The flywheel symbolizes not only the propagation of time, but also the inertia of matter and so the flywheel is in effect the total mass of the universe. This type of magnetic impulse motor is not self-starting, so at the inside on the flywheel axis a starter-motor is mounted, fitted solely for the purpose of starting. One rotating the synchronous motor maintains the exact number of revolutions at the same frequency of the magnetic alternation field. Thus in each of pyramid's edges two synchronous motors are mounted making twelve flywheels in total.

The two flywheels per axis rotate in opposite directions to symbolize the time propagation in opposite time directions. This type of linear impulse motor is chosen because it can be made to rotate in either direction.

The first analog model.

In one triangular plane of the pyramid the six synchronous motors in the iron frame demonstrate a rotating magnetic field, seen from the center of the plane. Each edge of the pyramid has the right phase-shift with respect to the other edges of the triangle.

The first model has four central rotating magnetic field generators in the middle of each triangular plane. (See figure of unfolded pyramid indicating the actual phase-shift).

These four synchronous (three phase) generators are coupled together by a gearbox and are driven by the main electric motor (not shown in the drawing). Consequently the four three-phase generators fulfill the synchronisation and constant phase-shifts for the flywheels of the pyramid-edges. The four separate rotating fields together, initate the complex three-dimensional rotation of the carrier field (or atom-field).

The disadvantage of the first model is that it is further away from reality, because the atomic field is not symboized as one three-demensional rotating field. However it shows better how to imagine the complicated rotation in three dimensions.

The second or the real analog model.

This mechanical model has only one central generator in the middle of the pyramid, instead of the four rotating field generators of the first model. All magnetic waves in the pyramid edges go straight to the pyramid's center. The rotor of the 3-dimensional rotating generator is like a yo-yo with spherical end-surfaces.

It requires a touch of genius to provide the drive-connection between the central motor and the spherical rotor, which has to be started up in a complex rotation. The six diametrically opposed pole-shoes from the middle of the pyramid's edges around the spherical rotor form in pairs, three axes of right angles - representing a Cartesian coordinate system. So it shows that the six time-axes of the edges of the pyramid can be converted into the three principal time-axes of the central sphere.

Together it imitates three-dimensional time, which causally creates the three-dimensional space of our universe.

The principles of the first and second analog models are equivalent. One version can be converted into the other, for example, converting the four rotating fields to one three-dimensional rotation in the centre-mass of the pyramid.

The model can be started by switching on the thirteen drive-motors. As the flywheels reach velocity the twelve starter-motors are switched off and only the central motor drives the complete gigantic model. The essence is that all flywheels have exactly the same number of revolutions. They are all either synchronous motors or generators. Then the flywheels are marked and illuminated by a stroboscopic flash light. *Seen by the normal eye they do not appear to move showing the synchronous rotation and the constant phase-shifts between the flywheels.*

These mechanical models are meant to explain the unimaginable Creation out of the nothing by symbolic imitation of the carrier-field. However it is only a weak symbol of Creation. The carrier-field is the esoteric medium, a transmutation of the absolute spirit or conscience, it is the spiritual force field of infinite strength which is present everywhere in the universe.

The imitation of the carrier-field functions.

Nevertheless the analogy of reality is incredible. This will be shown in the next explanation of the five functions of the cosmic carrier-field.

The carrier-function

The foundation of the carrier-field is the infinite force-field created by interference quenching (avbrytende stopp) of the original infinite vibration. The mode initiates this force-field with the six bundles of soft-iron bars as the edges of the pyramid.

As in reality, the three-dimensional magnetic rotating field in our model is caused by the linear waves of each time-axis. *The timeless waves alternate backwards and forwards as in our model*. They drive the flywheels representing the masses of the universes, and their rotation is the propagation (spredning) in time.

The orientation of the time-axes with respect to the nothing.

The relative position of one time-axis in the apparent nothing plays an important role in the orientation of the universes in reality, determining the direction of each time-axis. Here the first model is shown representing the four three-phase synchronous generators in the planes of the pyramid. Only with decompostion of the three-dimensional rotation into four rotations can one show that the three-dimensional rotating field has left no remaining component in the nothing, while in contrast a two-dimensional rotating field can not be nullified, a remaining component will always exist in the nothing. This can be explained by considering a single plane of the pyramid with its six flywheels at the edges, the analog for a two-dimensional carrier-field. The rotation impulse of the flywheels is compensated by the opposite rotations but the rotation impulse of the central three-phase generator is left over. It can be measured by the outside world (the nothing).

It requires four field three-phase generators (maintaining overall symmetry) to compensate for these rotation impulses.

With this the three-dimensionality of the carrier-field has been shown and the orientation of the time-axes with respect to the nothing is determined.

The time synchronization function.

Impressively the model imitates the time synchronization, simultaneously maintaining the cosmic counter-balance by which the universes continue to exist. The rotation of the two flywheels per axis is in opposite directions and they rotate with exactly the same speed. Thus it imitates the circular motion to the timeless waves which progresses exactly synchronous by their mutual interference coupling.

The twelve cosmic systems run without any dissipation, once started, the perpetuum mobile continues into eternity with the absolute accuracy of time-synchronisation. If one supposes the model to be without loss and without using the eternal driving motors, it will continue to run synchronously due to the inertia of the flywheels and the fact that the synchronous linear impulse motors can function as generators as well.

This symbolizes another property of the continuous circular motion of the timeless waves, because every wave somewhere in the pyramid circuit at each moment has its opposite value resulting in an overall zero at any moment.

This is the law of conservation of the nothing.

The couplingfield function.

Time synchronization is not alone determined by the iron grip of the three-dimensional carrier-field and the principle of cosmic counter-balance. Apart from this an important effect is realized from the feedback of the gigantic masses of the universes to the timeless waves. These masses can be considered as enormous flywheels without loss. Once started by the super flash, no change in speed is possible, because a force of sufficient magnitude is simply not available. This function of the coupling field can also be shown in our model, because the linear synchronous motors of the flywheels are also generators. If the central driving motor is switched off, then the flywheels continue for some time to drive the central three-dimensioal rotating sphere. Everything stays exactly synchronous with the same phase-shifts.

This experiment shows how the flywheels feed their energy back and have a stabilizing function in time-synchronization. The coupling-field of the masses of the universes is fed back in the time-bound vibrations of the atom-field from which no atom total mass of the universes is fed back to one single atom.

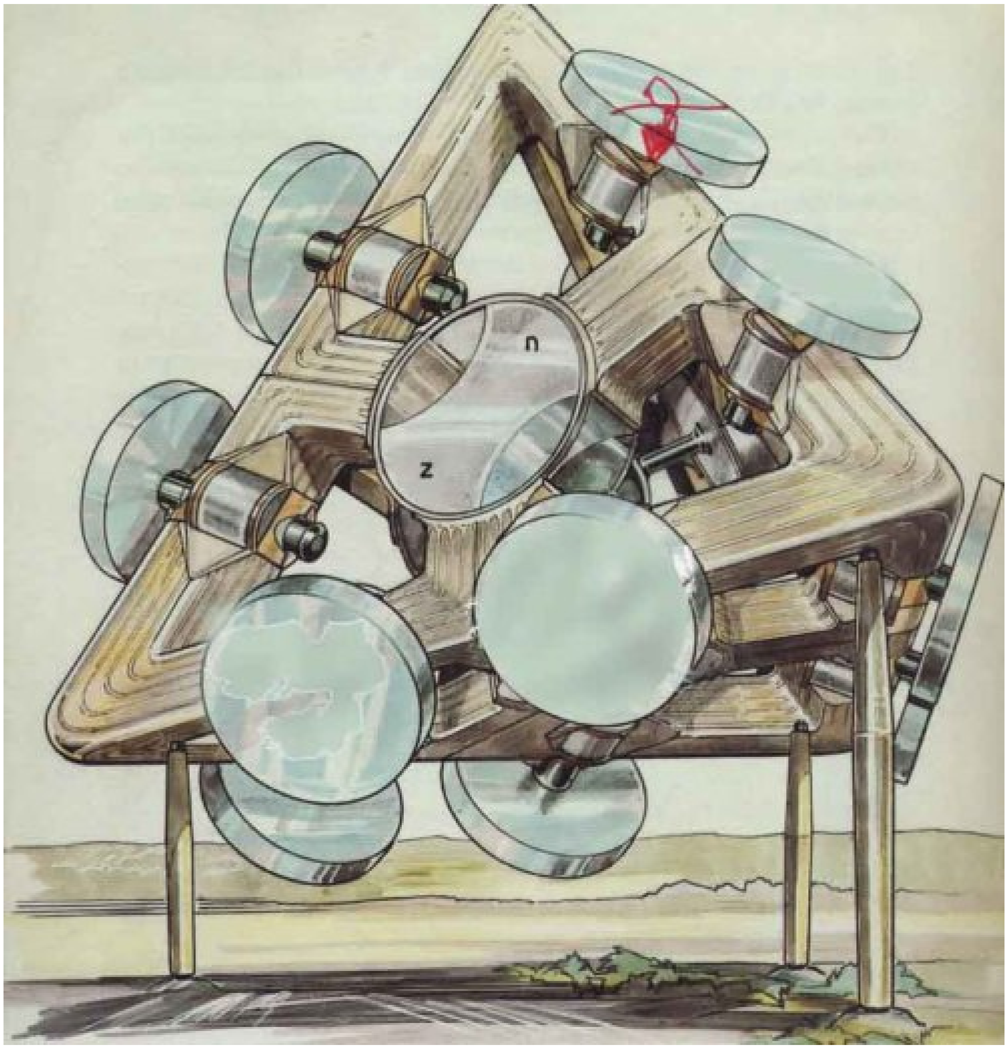

This is a painting by Rudolf Dass of the analogous visual-mechanical carrier-field model which symbolizes time and matter creation and synchronization in twelve universes. This giant working construction that was a cross between a monument and a machine was seen on an artificially flattened mountaintop on **Iarga**. This place was perceived as a place of pilgrimage, a place of meditation and contemplation, and the machine was a symbol of creation. The **Iarga**ns know no separation between science and religion, and religious and spiritual theories are entirely scientific. The model was several hundred yards(meters) long. The machine seemed to operate by some kind of perpetual motion, neither running on fuel or electrical power.

The Instructional Analog Model

The mechanical instruction models are the visual representations of the principles of the cosmic unified field theory. The model is intended to demonstrate that time, matter, and energy are creations out of the nothing and therefore the counter-balance has to be maintained to allow the causal infinity of the nothing to exist. In a broader context the physical symbols all take on new meaning and demonstrate the unification of science and religion. In a religious sense the spherical three-dimensional time-field in the heart of the pyramid symbolizes the Soul of the Creator. The pyramid symbolizes his infinite lovewill, the cage in which the infinity of His existence has been blocked irrevocably. The lines symbolize the different frequencies of the timeless waves. The flywheels on the edges of the pyramid symbolize the twelve associated and interrelated universes, including ours, that make up this unity. One cannot exist without the others.

The timeless primeval atom-principle, the transmitter function.

The atom-field is a time-creating field and as such a remote manifestation of the primeval atom principle. It can be compared with a transmitter. As soon as the program is converted into waves of great power, billions of radios can receive the same program, the number is unlimited. Everything depends on the one program of the transmitter, the primeval atom-field. If this exists with infinite strength, it is at the same time the vibration medium (the carrier-field) which helps it to spread everywhere. *Thus any atom in any universe can receive the timeless basic program.* The second model shows the three-dimensional time-field of the primeval atom field most clearly, although this is a three-dimensional rotation of the central sphere in spatial sense. This is an analog of reality, the atom has a spherical field which is defined by six linear time-vibration directions. Any atom is a remote manifestation of the one carrier-field of infinite strength.

Considering this fifth function of the carrier-field one has to realize that only the central sphere, the atom, is visible to us. The fields in the pyramid edges outside the sphere are invisible. It is the time-field reality of which the atom is a reflection.

The mechanical instruction models are the visual representations of the principles of the cosmic unified field theory. This requires however some marginal notes. A material comparison to explain spiritual abstractions is always in some ways incomplete. There are several hiatus of which only the three most important ones are mentioned.

1. In the model one time-axis represents only *one frequency* of the magnetic field, while in reality all frequencies are contained. Every frequency however can be initiated for each time-axis for which the interference representation of the analog model is valid.

2. The model demonstrates the rotating field of the *timeless waves.* It *does not imitate the time-bound waves* by which our time-axis the propagation speed of time is X, then the phase velocity of the two time-axes connected to "our" pyramid edge is ½X. The other two time-axes connecting our time-axis with the other corner of the pyramid is -½X, making up for the time speed of our anti-universe -X. The crossing, not-touching time-axis is perpendicular to our time-axis and does not contribute.

3. In our model the masses of the universes are given by the flywheel symbol. In fact this is *fundamentally wrong, because it is separated from the three-dimensional rotation of the central sphere.* It can therefore basically not maintain the counterbalance of the nothing.

The most important conclusion of the analog model is that *time, matter and energy are creations out of the nothing and therefore the counter-balance has to be maintained to allow the causal infinity of the nothing exist.*

14. The existential symbolism of the model.

So far the physical aspects of the model have been unravelled, but if one wants to give a wider context to the model suddenly all the physical symbols change their meaning. In relation to philosophical or theological disciplines the model takes on the additional meaning of the unification of science and religion. By certain inherent (medfødte) laws within any intelligent race the denial of God's existence will be destroyed sooner or later. Earth science will provide an indestructible contribution to our final knowledge-awareness, which we shall attribute as typical human attainment in the cosmic integration. It is that piece of our human identity and self-awareness, which we shall possess and care for throughout eternity. *It the light of this purpose, science serves God and the distinction between science and religion fades away.*

In a religious sense the spherical three-dimensional time-field in the heart of the pyramid has a special meaning. **It symbolizes the Soul of the Creator**, the Father who is the pure Identity of the infinite origin. The pyramid symbolizes his infinite Love-will, the cage in which the infinity of his Existence has been blocked irrevocably. He sacrificed his Ego and his freedom, to love his twelve Sons once. He is present only indirectly in his Creation as the reflection of the infinity, the Supreme Being, as the one Ego-consciousness of the Archangels.

That is why the theologian covers the central sphere with pure gold, symbolizing the golden Heart, the Holy of Holies, the most sacred secret of the universe. The pyramid is then painted dark red, the universal colour of love. Over the red many thin golden stripes run in the direction along the pyramid edges. These stripes symbolize the different frequencies of the timeless waves, the expressions of Love and faith of the Father and his infinite Will.

Gathered around the golden Heart are the twelve flywheels, coloured white, the colour of purity and innocence. The model shows that every universe has the power to awaken the matter-consciousness and to transform it to the new

God: the Son. The goal of Creation is therefore the awakening and the formation of the twelve Sons of the Father.

This is not in contradiction with the Bible telling us that the Father sent his only-begotten Son. The Bible is solely directed to man. It is our instruction book for life describing our attitude towards God and our fellow man. The Bible does not tell anything about life on other planets, because it is not relevant in our relationship with God now.

The same goes for our parallel universes. In fact they do not exist for us, because they are not in our world of existence but somewhere else. They do not exist for all intelligent races in this universe, we shall never be able to observe them nor experience them in any way. Perhaps at the end of all time we shall meet. In *our* world the Father has one only-begotten Son and this truth is already so unimaginable that nobody can conceive the consequences yet.

END OF IARGAN DISCUSSION

---

mainpage

the IARGA CONTACT story

ufo-contact-material

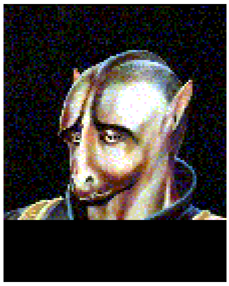

Technical information given from the IARGA contactteam.

their describtions on the matter of TIME

- AND MORE

APPENDIX 1

An Analog Model

1. Introduction

This appendix is not a part of the story of the Creation which is explained in religious terms in Chapter 1 (Part 2). It is a complementary way of reasoning to prove the existence of God, using the concepts of time, matter and energy. These concepts are also used in physics. The problem, which now arises, is that the author will use physical concepts without understanding the real meaning of the words, because he is not a specialist. The words atom, atom-field, matter, etc., must all be understood within the context of this book and they have no direct connection with the ideas used in modern science. Especially the words "matter" and "atom" are most confusing, because the **Iarga**n author knows about the existence of matter in three configurations (and its opposite possibilities as anti-matter); i.e., time-matter, phase-matter and *reflection-matter* (the one we know about in our universe), but he fails to interpret these concepts due to his lack of knowledge of physics. He only knows that these forms of matter are all related to one or none of the directions of time- propagation (tidsforplantning) of the atom of which time is fundamentally three-dimensional. It therefore determines in which form these configurations (beskaffenheter) can exist in our universe.

In this appendix - time-creation out of NOTHING is described - by making use of the concept of more time-dimensions. The explanation of time-creation is very difficult because it uses an overall carrier-field of infinite strength and then this concept is called the cosmic unified field theory. In defining this term, one has to be careful, *because this theory is of a different basic nature from the existing unified field theories here on earth.* Our unified field theories try to unify the four basic forces in nature, which are the electro-magnetic, the weak strong nuclear force and gravitation. The unified field theories accepted by most prominent (iøynefallende) scientists find their origin in the general relativity theory of Einstein, which itself explains matter and energy in terms of space and time.

The problem now is to find a method, explaining the spiritual or esoteric reality, the cosmic time-field. On other planets, this problem was solved by explaining the basic concepts of the cosmic unified field theory. The elementary functions of these concepts can be understood with a mechanical instruction model demonstrating the cosmic time-field visually.

Why indulge in this complication of describing a theory which will probably only be fully grasped by a few individuals? It does not concern the author very much, he only carries out his commission. His **Iarga**n instructors gave three reasons for publishing the concepts of this theory:

1. The future integration of science and religion requires a material proof of God's existence. This explanation could be the start of the exchange between these two.

2. In the previous story of Creation the limitation of the infinite is explained unambiguously(upretendiøst), but the working principle was completely neglected(forsømt).

3. The cosmic unified field theory published in this book will be the only possible verification for the true identity of this book, providing the link for the future development of the human race.

2. The physical approximation of the NOTHING.

One concept in physics is that all atoms and energy consist of vibrations and wave appearances. In a material way, every atom or form of energy is caused by vibration. The physical approximation (tilnærmelse) describes the nothing of the idea of vibration - and *does not touch on the spiritual aspects and the consciousness of matter*. It does not distinguish between the "nothing" and the spiritual "No-Thing" (the ALL). It only recognizes the material aspects of the nothing.

If everything is vibrations or waves, then it is not necessary to speak of matter or energy, then Creation is fundamentally based on this concept of vibration. The primeval vibration (before the limitation) can be described as infinite in time, infinite in all its aspects and possibilities, infinite in frequency and amplitude and existing in all time directions (omnitime). Unchanging, timeless and powerful to all limited vibrations in every bound power. It is the omnipotent definition.

In this approximation the nothing is more simple to understand. Before the creation - the infinite omnipotent vibration existed- *but a vibration in every direction and of infinite frequency and amplitude is not a vibration*. The apparent paradox can be eliminated by stating: Only at the instant when this infinite vibration blocks itself by interference*, as in a flash, all bound waves and frequencies were created.* This condition is very close to the existential view of the Creation out of NOTHING in Chapter 1.

The nothing (here) has no counteraction, but in contrast the omnipotent vibration - can manifest itself only by counter-acting in such a way that the infinite allows the nothing to be. So one can also state that all vibrations were created because the nothing bound itself by a mechanism of interference quenching (ble dempet) - (counteracting vibrations) of infinite force, which blocked the infinite possibilities of these vibrations. *This universe can only be considered as the limitation of the unbounded nothing.*

3. The bounding of the omnipotent infinite vibration, the cosmic time field.

One of the basic principles on which the mechanism of bounding is based is: *If one property of a power (any power) of something is bound, then all properties and powers of this something will be bound.* In infinity the finite cannot exist. To limit the omnipotent vibration it was sufficient to bind only one of the perhaps infinite number of possibilities. Because if only just a little something were to be formed in the nothing, the nothing would be altered and cancelled irrevocably, then the complete finity (endelighet/begrensethet) was a fact. OR, if there were just the three-dimensional rotating time-field as a possibility of the primeval atom in the infinity of the nothing, then the finite was a fact. This atom-field needed to be a force-field of infinite strength, the carrier-field, to counteract the omnipotent vibration, the cosmic time-field.

4. The Creation can only exist as a result of counter-balance.

One of the functions of the carrier-field is time-synchronization by which the cosmic counter-action and therefore the nothing can BE. *Matter and energy cannot exist on its own in the nothing.* The principle of time synchronization is given:

1. Time is only relative, it is motion or propagation (spredning) in relation to at least a second motion of time.
2. the cosmic law of counterbalance dictates that the second time has to counterbalance the first time. So time can only exist in relation to a second exact *opposite direction* of time (running backwards in relation to our time).

3. Time exists only through the presence of atoms or matter. Then the backward-running time can only exist through matter in which the time propagation is opposite, *running backwards* in relation to us (antitime matter). Anti-matter or reflection antimatter may exist in our universe with an opposite electric charge with respect to our known reflection matter. The question, as to which charge antitime matter has in the anti-universe, is of no importance in this appendix.

4. Because neither matter nor energy can exist on its own in the nothing -another quantity of matter of opposite properties *must* exist which is exactly similar to that in our universe. This balance of matter and energy is so accurate, that one surplus atom (overskuddsatom) is impossible.

This explains nearly as per definition the cosmic law of counter-balance allowing the nothing to be. There must be at least a second universe exactly like ours, in which time runs backwards according to us, but normally for life in that universe. In the following it is shown that there are not just two, but *eleven* other universes in addition to ours necessary to maintain the cosmic law of counter-balance.

6. The time-synchronization field.

The law of counter-balance also applies to time. The timedirection is opposite and the time velocity is equal to ours. Time is the propagation of vibrations of the atom. Then the atomvibrations in both universes are equal and opposite with respect to the timelessness. This is only possible if these times are caused by one field (time-field).

The definition of a field (any field in physics) is: If an event occurs, because another event takes place somewhere else without a mechanical or visible connection between these events, then these events are caused by a field. Time-synchronization occurs according to this definition. The atom vibrations in our universes are exactly similar, because the atom vibrations of our anti-universe occur despite the enormous difference in time and without any dissipation of energy. It is about 30 billion (milliarder) years since the occurrence of the super flash Creation (big bang) between the two universes.

The law of conservation of counter-balance is causal to the law of energy conservation in every universe. Each universe is a closed system and without energy dissipation outside itself. The amount of matter and energy once created remains constant. *A loss of energy is impossible, because our time-bound energy propagates forwards, while the time-bound matter and energy of the anti-universe propagate backwards in time.* These universes are separated by a time-barrier through which no communication nor transfer of life is known. (Well - according to information in the book UFO-CONTACT FROM PLANET KOLDAS and also told by the et-contact ASKET and Semjase - there are some methods for entering the mirroruniverse - but this can not be touched here - reseach yourself if interested!! R.Ø.remark.)

6. *What is a timeless vibration or wave?*

The question is how the field functions maintain the time-synchronization of the atom-waves between the universes separated by billions of years. It is only possible if the synchronization stands still in time, but exists simultaneously in all the time elapsed (- som er gått..) since the beginning of the super flash, as well in all the future. *A timeless vibration does not normally move in time*, the amplitudes of the field propagation - alternate (veksler) forwards and backwards in time. Time synchronization is the principle by which our universe catches every forward-moving amplitude and the anti-universe catches every backward moving amplitude of one and the same timeless wave.

On other planets there are simple mechanical models explaining the function of time synchronization. One such model is very similar to our well-known mini-model rail upon which are two little wagons. The model has two lengths of rail of say 1.50 meters each, which are in direct line with each other.

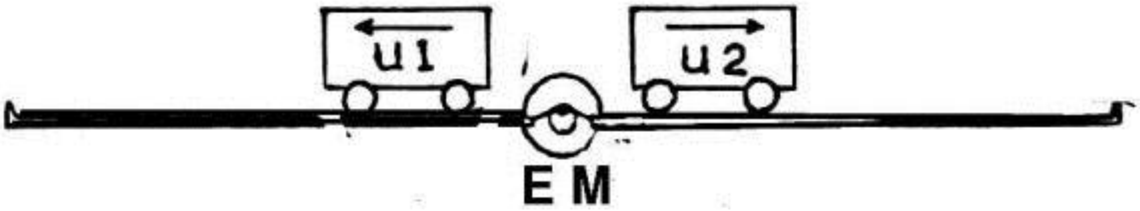

Picture of the model wagons with a Electric drive motor which letting an excentric, and so the rails vibrate horisontally.

Rail and wagon model demonstrating the principle of the timeless waves. The vibration is present at any point of the rail and at any time during the whole process period. Also from the moment of the super flash until the end of all times (ALWAYS).

The rails are mounted on small springs to allow them to vibrate forwards and backwards in horizontal direction (from left to right and vice versa). The two rails are connected to a little electric motor with a double eccentric, each driving one rail, letting the rails vibrate about 1 millimeter of linear distance, also backwards and forwards. The amplitude is small enough to be invisible to the eye.

Then one little wagon is placed on each rail and each wagon has a catch with a catchwheel (tilbakeløps-sperre) on one of the wheelaxes, so the wagon can ride in only one direction. *The movement in the other direction is blocked.* The wagons are placed with their rears against each other, after which the electric motor is switched on. If the teeth of the catchwheel have the same beat as the vibrator mechanism, the wagons roll away from each other with exactly the same speed to the end of each rail.

With this simple model time synchronization for the atom-vibrations of both departing universes (in time) by a timeless field "always existing everywhere" has been demonstrated sufficiently. At the same time a flaw(feil) comes to light. It is impossible to maintain an energy time flux of the timefield for the two departing universes, because where is then the inexhaustible (utømmelige)energy source to maintain the time-flux? (betyr det tidsustabiliteten??)

This is one of the reasons that a one-dimensional time-field is not possible. Only a three-dimensional time-field, which encloses the energy time-flux in itself, allows our universe and the others to exist.

7. The principle (mechanism) to create finity: INTERFERENCE.

So far the function of time-synchronization of the carrier-field has been considered. Now we will discuss the medium of conserving and maintaining the timeless vibration into eternity. It is the infinite force-field carrying all matter and energy.

We are at the moment of the flash or big bang, the moment at which the infinite primeval vibration limited (transmutated) itself to the carrier-field. This mechanism of blocking is well known and based on the principle of wave interference. (Interference of lightwaves is important today in laser technology.) Interference is the action or counter-action of two or more waves. If the *wave-propagation(-forplantning) is opposite and parallel* and the frequencies of the waves equal, mutual amplification (gjensidig forsterkning)or attenuation (fortynnelse) of the intensities of the wave is possible, depending on the phase relation between these waves.

If *the phases are opposite* and the *amplitudes of the waves are equal,* total quenching (dempning/kvelning)is possible. This is the principle of the binding of infinity. If an infinite wave manifests itself in one time-direction, automatically the equivalent absolute counteraction is created which counter-balances the infinity, allowing the nothing to be.

The infinite primeval wave(urbølgen) blocked itself at the moment it created itself and so it never existed in reality. It was in fact a *transmutation (forvandling) of infinity;* the transmutation of the infinite vibration of infinite amplitude and frequency to a carrier-field of infinite strength, which blocked itself by interference.

8. The time-bound vibrations or waves.

The primeval wave blocked itself in a three dimensional time-axis system (propagation directions) by interference, and trans-mutated to an infinite, timeless, immobile force-field. How could the bound timeless waves exist, which synchronize our universe?

It is an error of thinking to suppose that by quenching the omnipotent vibration - all finite waves were also cancelled. It is just the opposite. A vibration with infinite frequency is not a vibration in reality. The infinite prevented the vibration, because a vibration can only exist in the finite. As soon as infinity was removed, the prevention disappeared and the real bound wave was a possibility. It is the carrier forcefield although of infinite strength and energy density which gets its finite value due to interference quenching. All timeless waves are as it were encapsulated in infinity and therefore indestructible. However "belonging" to infinity and not bound by time themselves, they create time forever, they are the eternal ripples on the static ("krusning i likevekten"), unchangeable and loss-free field of infinite strength. (This must be what the Danish wiseman Martinus call the X2- principle and the primeval-desire or "urbegjæret" in Danish. R.Ø.remark.)

Because of this, all time-bound waves such as light, radio-waves or gravitation, exist in space without any loss and it is the empty space in the universe, which is the non-modulated structure of the carrier-field. It is one of the most fantastic touches of genius of Creation, that the non-vibrating infinity was transmutated to a loss-free medium carrying all time-bound and timeless waves. A few properties of this medium can be measured with instruments, the other part of its structure cannot be objectively observed yet, *it is the esoteric fluidum or the infinite Love-will of the Father which carries the whole Creation.*

9. The unimaginable three-dimensional time.

(Den utenklige tredimensjonale tid)

One time-axis counter-balances only one direction of time propagation of the omnipotent (allmektige)wave. In infinity all possibilities of vibration exist and therefore all time-propagation directions. The question is, could infinity be limited by just one time-direction and one co-existing propagation direction going the opposite way?

One way to answer this was already given in the previous paragraph, in which it was stated that the energy time-flux of the primeval wave of two universes is not a closed system. Another plausible answer can be by simply reasoning.

The atom is causal for the time - and space dimensions, due to its time-creating properties, already mentioned. *The space-creating properties are easy to understand, because the atom takes up space.* It is three-dimensional in space and therefore it creates space in three dimensions. If all atoms were two-dimensional then the whole of creation would have been on a plane and space would not have been more than a flat or curved plane(as on a sheet of paper. R.Ø.remark.) Our time-space is no more than a bound part of the nothing (det "noe som er" -X1- som Martinus kaller det. R.Ø.remark.) - being filled with matter or energy. Anything that space distinguishes (utskiller) of the time - and - dimensionless void, is caused by atoms or energy. If atoms have three space dimensions, then the timeless waves creating the atoms also have three dimensions.

1. The timeless primeval atom-field has three time dimensions.

2. The timeless waves could not be bound by only one time-axis, but must necessarily have three time-axes.

3. The timeless atom-field should consist of a three-dimensional rotating field of infinite strength, which is attenuated (fortynnet) of itself by interference. It conserves

the law of cosmic counter-balance of the no-thing. *The atom of such a rotating field structure has six time-axes, but in principle these six axes can always be reduced to three time-dimensions, no further reduction is possible.*

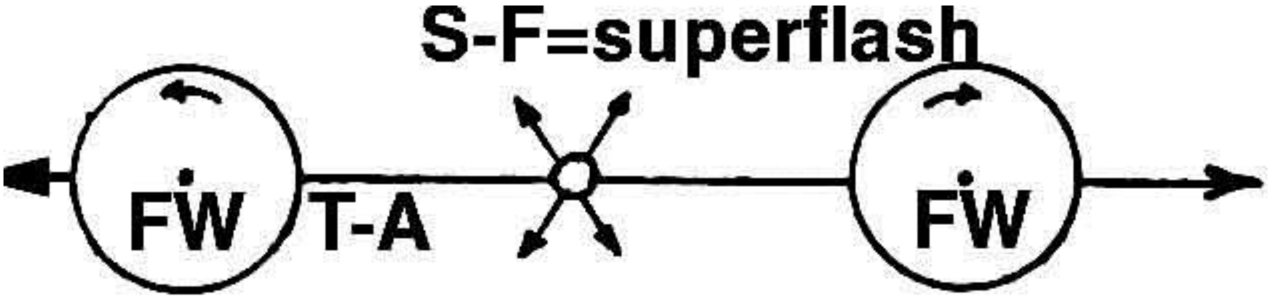

Picture of ONE-TIME-AXIS: The cosmic symbol of one time-axis.

In the following parpagraphs of the analog carrier-field model it is shown, why the atom-field has to be a three-dimensional and not a two-dimensional field.

10. The six time-axes of the time-creating, cosmic carrier field.

How can it be established that the carrier-field of all matter in twelve universes has six time-axes? The first part of the answer is simple to explain. *A well-known type of rotating field on earth is that of the electric three-phase synchronous motor*.

This type of electromagnetic-field system consists of three linear magnetic fields, which are arranged at angles of 120 round a rotor. The phase-differences of the three alternating magnetic-fields, which have the same frequencies (say 50 Hz) are also at each 120 degrees. Each alternating field is in itself an exact linear vibration oriented in one direction. The rotating field exists only if the three linear vibrations are combined together with the proper phase-relation between them. So a two-dimensional rotating field in space is created. It can force a piece of magnetic material to rotate around its axis with a great power.

Insert pict of motor

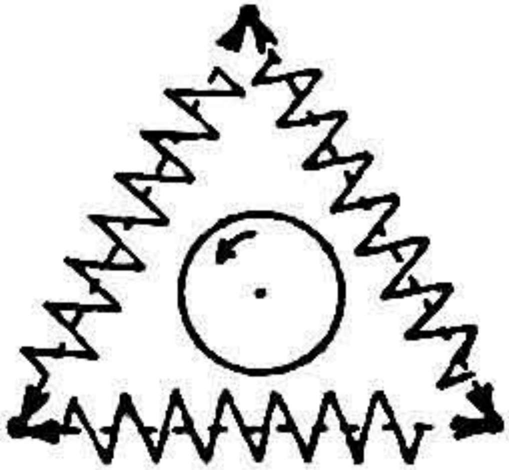

The principle of an electric three-phase synchronous motor.

The fact that it is possible to create rotating fields by combining three linear alternating waves does not mean that it has to be the same for fields of infinite strength, it can be more than three, but it is not very likely as first guess, because the three wave principle (the equilateral triangle) is the most simple and effective. According this principle, for a three dimensional field six linear waves are required to form an *equilateral (likesidet) triangular pyramid,* but why an angle of 60 degrees between time-axes?

The first answer can initially be given in a very simple way, the second is more complicated, but at least more to the point.

A simple way of reasoning to derive (utlede) the answer to the question of 60 degrees is the following. The atom field of matter has spatial (romlig)structure, therefore the time-axes must be positioned in such a way that they form a closed three-dimensional geometric object, because the energy fluxes of time must be closed in itself due to the law of counter-balance.

The time-axes form straight lines, which cannot be parallel lines, because parallel time-axes form one and the same time-direction. *The most simple three-dimensional closed object is an equilateral triangular pyramid.* There are no reasons to suppose it should be an a-symmetric pyramid if three-dimensional time is considered, on the contrary, it must be symmetrical in all positions. So it is simply a equilateral triangular pyramid of which the lanes are equilateral triangles. There is no more simple object in three-dimensional space with these requirements.

The more sophisticated answer to the question of 60 degrees is time given by approaching the question from the relativity of time frequencies between the other universes (the other time-axes). Time is the wave propagation of atoms which determines the direction of the past to the future.Simultaneously these time-creating or timeless waves are the linear phase waves combining to maintain the atomic rotating field. The problem is that "our' atoms do not experience the right frequencies of the other time-axes, because they do not propagate (spredes)with the same speed, but advance in time.

Thus there is an apparent frequency-shift, a kind of Doppler shift, by which the frequencies of the "other" time-axes are experienced by "our" atom. One should realize that the timeless vibrations of the other universes also exist in our universe.

At first sight it seems impossible for a rotating field to posess a constant phase-shift because between vibrations of different frequencies a constant phase-shift is not possible unless the frequencies are related to each other in a particular way. They are harmonic. Two vibrations are harmonic if the apparent frequencies have constant ratios to each other, like 1:2, 1:3, 1:4, etc.

Timeless waves can only exist when the frequencies are related by the harmonic ratio 1 : 2.

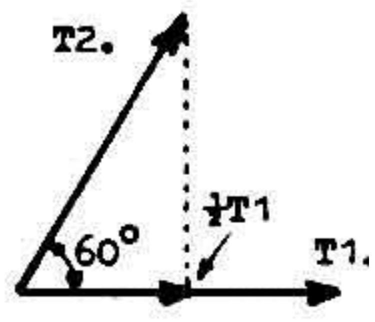

Picture explaining the frequency-shift for 60 degrees time-axis angle.

The frequency ratio is determined by the angle between the two time-axes, as the picture shows. Our time direction is T1 while another time direction is shown as T2. The time direction propagating forwards (from past to future) is useful to our universe, while the backward-moving impulses of the timeleess wave do not "catch" the rotating field of "our" atoms. Secondly the time-axis at a right angle (90dg) to our axis T1 does not make any contribution, we experience this time-axis as non-existing.

As the picture shows, due to the angle of 60 degrees, 'our' atom experiences exactly half the time-speed of the other time-axis and this makes the frequency harmonic. As T1=2 times T2, the result is the simplest harmonic ratio of 1:2, determined by the angle of projection of 60 degrees between two time-axes. (Pythagoras' law for right angled triangles).

All other possibilities and other harmonic ratios fail to give non-destructive contributions for the necessary interference intensification. Other excitations of higher order harmonic ratios give more complicated time-axes structures of the rotating atom field and cannot be stable. Also because the couple-field feedback to the total mass (inertia) of the universes is not optimized (see the following paragraphs).

It supposes simultaneously the existence of a higher number of co-existing universes, which seems highly unlikely (usannsynlig)(higher than twelve). Therefore we must decide for the simplest three-dimensional geometric form: the equilateral triangular pyramid.

11. The ancient symbol of God, the Star of David.

It is necessary to summarize the above issues of the principle.

1. The choice of the mechanism of limitation determined the existence of all types of timeless waves, among which was the (most complex) three-dimensional rotating time-field.

2. The limitation took place by interference quenching (forstyrrelses-dempning) the primeval infinite vibration (possibility) affording more time-axes in three dimensions.

3. A time-axis is a part of the infinite force-field. This only exists as a single linear vibration.

4. For a three-dimensional rotating field at least six linear field waves are necessary, which are located spatially in time having angles of 60 degrees between the time-axes.

5. The angle of 60 degrees could only be created by an apparent frequency shift maintaining the harmonic vibrations of the rotating field.

These considerations determine the mechanism of blocking the infinite. It consists of six time-axes forming a closed three-dimensional object, which creates and restricts the speed of the rotating atom-field.

Though every time-axis carries two universes(universe and anti-univers), there are in the frightening depth of cosmic space twelve distinct cosmic structures, separated by the unbridgeable barrier of time. Not only are there eleven other universes apart from ours, *but they are all embedded in the same cosmic or geometric space*. All universes have the same size and the same energy content. The geometric space is necessarily the location in the nothing - in which the omnipotent vibration quenched itself by interference. It is the location for waves encountering each other, as it is for light-waves interfering with each other. But it does not require that every atom has twelve buldges at the same space-point in time. Or more crucially, it does not mean that every human being has twelve images of himself without knowing it.

It is the mechanism of interference synchronisation of the timeless waves which uncouples the universes at each space-point, simultaneously maintaining the time-propagation of each universe.

*The equilateral triangular pyramid is the symbol of God, the Father as it is known everywhere in the universe (after lifting the cosmic isolation).*We know it as the double equilateral triangles of the Star of David, also six time-axes with angles of 60 degrees. It means that the principle of the limitation was revealed to humanity centuries ago, but we did not understand its meaning so far.

make in your thoughts : The Star of David.

Technical information given from the IARGA contactteam.

their describtions on the matter of TIME

- AND MORE

picture is described later

12. The analog carrierfield model, an instruction model.

The cosmic carrier field of all space and matter performs a sixfold function, of which some have already been discussed in gereral way.

1. The carrier function of an infinitely strong forcefield blocked in itself contains the fluidum conducting all waves without dissipation and maintained to eternity.
2. The orientation of the time-axes of the universes with respect to nothing provides the structure of the carrier-field of three dimensions, the bounded Love-will, creating our life and existence.
3. The time synchronisation function locks all universes in the iron grip of the law of cosmic counter-balance from which no atom can escape.
4. The feedback principle of time to the total mass of the universe is the function of the coupling-field by counterbalance (and interference) of the time-bound waves.
5. The function of the timeless primeval atom-field, the atom-principle, as the transmitter program containing the code of matter of which all atoms are remote manifestations.
6. The function as bearer of the absolute ALL-knowledge and goal consciousness, manifesting itself as the laws of nature and as the intelligent goal of life: the spirit of God.

The last function concerns the existential aspects, which have been discussed in the previous section of this book.

The first analog model is a mechanical instruction model explaining the cosmic carrier field by means of symbols. This model was built on another life-supporting planet on a high mountain and was many meters high. It was the gathering point for contemplation of that intelligent race for devotion to its Creator.

It was the symbol by which the majority of the race acknowledged the goal of Creation, simultaneously the sign that the cosmic isolation had been lifted. The other function of the carrier-field will be discussed with the help of this mechanical instruction model.

The analog model consisted mainly of the six edges of a huge equilateral triangular pyramid built out of solid soft-iron bars. The many metres thick edges of the pyramid contain an impressive number of iron bars parallel to the direction of the edges and cast in a kind of plastic resin following a process similar to that of casting concrete(betong). So a mainly iron-frame was formed especially suitable for conduction of magnetic pulses of high power without much loss.

So far the carrying fuction of the cosmic time-field, which primarily maintains the timeless waves is symbolized in the model. The magnetic pulses in the edges of the pyramid run backwards and forwards like the timeless waves in reality.

With respect to the other edges of the pyramid, the magnetic pulses are phase-shifted. The problem is that up to now the forward propagation (forplantning) in time has been missed out. In fact we look for the equivalent of the two little wagons on the vibrating rail. The solution can be found in the symbol of the rotating flywheel, as *with the atom its rotation is continuous in time, it is in fact time propagation.* Therefore the iron frame of the pyramid is discontinued at two places in each of the six edges of the pyramid. At this discontinuation point a linear magnetic impulse motor is installed, on the axis of which a large flywheel is mounted.

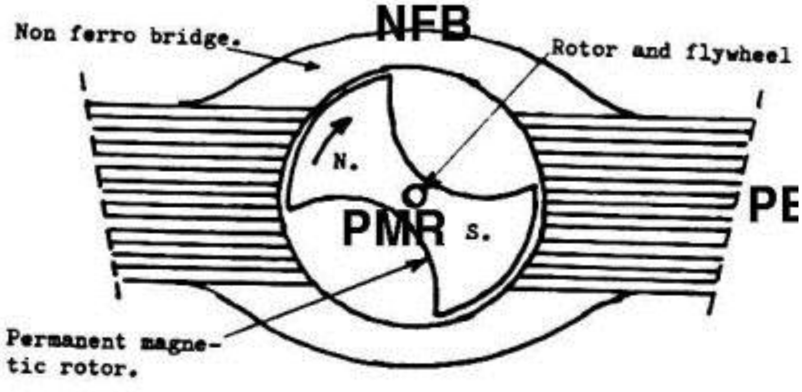

NBF=NonFerroBridge

PE=PyramidEdge and soft iron bars

PMR=PermanentMagneticRotor

The flywheel symbolizes not only the propagation of time, but also the inertia of matter and so the flywheel is in effect the total mass of the universe. This type of magnetic impulse motor is not self-starting, so at the inside on the flywheel axis a starter-motor is mounted, fitted solely for the purpose of starting. One rotating the synchronous motor maintains the exact number of revolutions at the same frequency of the magnetic alternation field. Thus in each of pyramid's edges two synchronous motors are mounted making twelve flywheels in total.

The two flywheels per axis rotate in opposite directions to symbolize the time propagation in opposite time directions. This type of linear impulse motor is chosen because it can be made to rotate in either direction.

The first analog model.

In one triangular plane of the pyramid the six synchronous motors in the iron frame demonstrate a rotating magnetic field, seen from the center of the plane. Each edge of the pyramid has the right phase-shift with respect to the other edges of the triangle.

The first model has four central rotating magnetic field generators in the middle of each triangular plane. (See figure of unfolded pyramid indicating the actual phase-shift).

These four synchronous (three phase) generators are coupled together by a gearbox and are driven by the main electric motor (not shown in the drawing). Consequently the four three-phase generators fulfill the synchronisation and constant phase-shifts for the flywheels of the pyramid-edges. The four separate rotating fields together, initate the complex three-dimensional rotation of the carrier field (or atom-field).

The disadvantage of the first model is that it is further away from reality, because the atomic field is not symboized as one three-demensional rotating field. However it shows better how to imagine the complicated rotation in three dimensions.

The second or the real analog model.

This mechanical model has only one central generator in the middle of the pyramid, instead of the four rotating field generators of the first model. All magnetic waves in the pyramid edges go straight to the pyramid's center. The rotor of the 3-dimensional rotating generator is like a yo-yo with spherical end-surfaces.

It requires a touch of genius to provide the drive-connection between the central motor and the spherical rotor, which has to be started up in a complex rotation. The six diametrically opposed pole-shoes from the middle of the pyramid's edges around the spherical rotor form in pairs, three axes of right angles - representing a Cartesian coordinate system. So it shows that the six time-axes of the edges of the pyramid can be converted into the three principal time-axes of the central sphere.

Together it imitates three-dimensional time, which causally creates the three-dimensional space of our universe.

The principles of the first and second analog models are equivalent. One version can be converted into the other, for example, converting the four rotating fields to one three-dimensional rotation in the centre-mass of the pyramid.

The model can be started by switching on the thirteen drive-motors. As the flywheels reach velocity the twelve starter-motors are switched off and only the central motor drives the complete gigantic model. The essence is that all flywheels have exactly the same number of revolutions. They are all either synchronous motors or generators. Then the flywheels are marked and illuminated by a stroboscopic flash light. *Seen by* *the normal eye they do not appear to move showing the synchronous rotation and the constant phase-shifts between the flywheels.*

These mechanical models are meant to explain the unimaginable Creation out of the nothing by symbolic imitation of the carrier-field. However it is only a weak symbol of Creation. The carrier-field is the esoteric medium, a transmutation of the absolute spirit or conscience, it is the spiritual force field of infinite strength which is present everywhere in the universe.

The imitation of the carrier-field functions.

Nevertheless the analogy of reality is incredible. This will be shown in the next explanation of the five functions of the cosmic carrier-field.

The carrier-function

The foundation of the carrier-field is the infinite force-field created by interference quenching (avbrytende stopp) of the original infinite vibration. The mode initiates this force-field with the six bundles of soft-iron bars as the edges of the pyramid.

As in reality, the three-dimensional magnetic rotating field in our model is caused by the linear waves of each time-axis. *The timeless waves alternate backwards and forwards as in our model*. They drive the flywheels representing the masses of the universes, and their rotation is the propagation (spredning) in time.

The orientation of the time-axes with respect to the nothing.

The relative position of one time-axis in the apparent nothing plays an important role in the orientation of the universes in reality, determining the direction of each time-axis. Here the first model is shown representing the four three-phase synchronous generators in the planes of the pyramid. Only with decompostion of the three-dimensional rotation into four rotations can one show that the three-dimensional rotating field has left no remaining component in the nothing, while in contrast a two-dimensional rotating field can not be nullified, a remaining component will always exist in the nothing. This can be explained by considering a single plane of the pyramid with its six flywheels at the edges, the analog for a two-dimensional

carrier-field. The rotation impulse of the flywheels is compensated by the opposite rotations but the rotation impulse of the central three-phase generator is left over. It can be measured by the outside world (the nothing).

It requires four field three-phase generators (maintaining overall symmetry) to compensate for these rotation impulses.

With this the three-dimensionality of the carrier-field has been shown and the orientation of the time-axes with respect to the nothing is determined.

The time synchronization function.

Impressively the model imitates the time synchronization, simultaneously maintaining the cosmic counter-balance by which the universes continue to exist. The rotation of the two flywheels per axis is in opposite directions and they rotate with exactly the same speed. Thus it imitates the circular motion to the timeless waves which progresses exactly synchronous by their mutual interference coupling.

The twelve cosmic systems run without any dissipation, once started, the perpetuum mobile continues into eternity with the absolute accuracy of time-synchronisation. If one supposes the model to be without loss and without using the eternal driving motors, it will continue to run synchronously due to the inertia of the flywheels and the fact that the synchronous linear impulse motors can function as generators as well.

This symbolizes another property of the continuous circular motion of the timeless waves, because every wave somewhere in the pyramid circuit at each moment has its opposite value resulting in an overall zero at any moment.

This is the law of conservation of the nothing.

The couplingfield function.

Time synchronization is not alone determined by the iron grip of the three-dimensional carrier-field and the principle of cosmic counter-balance. Apart from this an important effect is realized from the feedback of the gigantic masses of the universes to the timeless waves. These masses can be considered as enormous flywheels without loss. Once started by the super flash, no change in speed is possible, because a force of sufficient magnitude is simply not available. This function of the coupling field can also be shown in our model, because the linear synchronous motors of the flywheels are also generators. If the central driving motor is switched off, then the flywheels continue for some time to drive the central three-dimensioal rotating sphere. Everything stays exactly synchronous with the same phase-shifts.

This experiment shows how the flywheels feed their energy back and have a stabilizing function in time-synchronization. The coupling-field of the masses of the universes is fed back in the time-bound vibrations of the atom-field from which no atom total mass of the universes is fed back to one single atom.

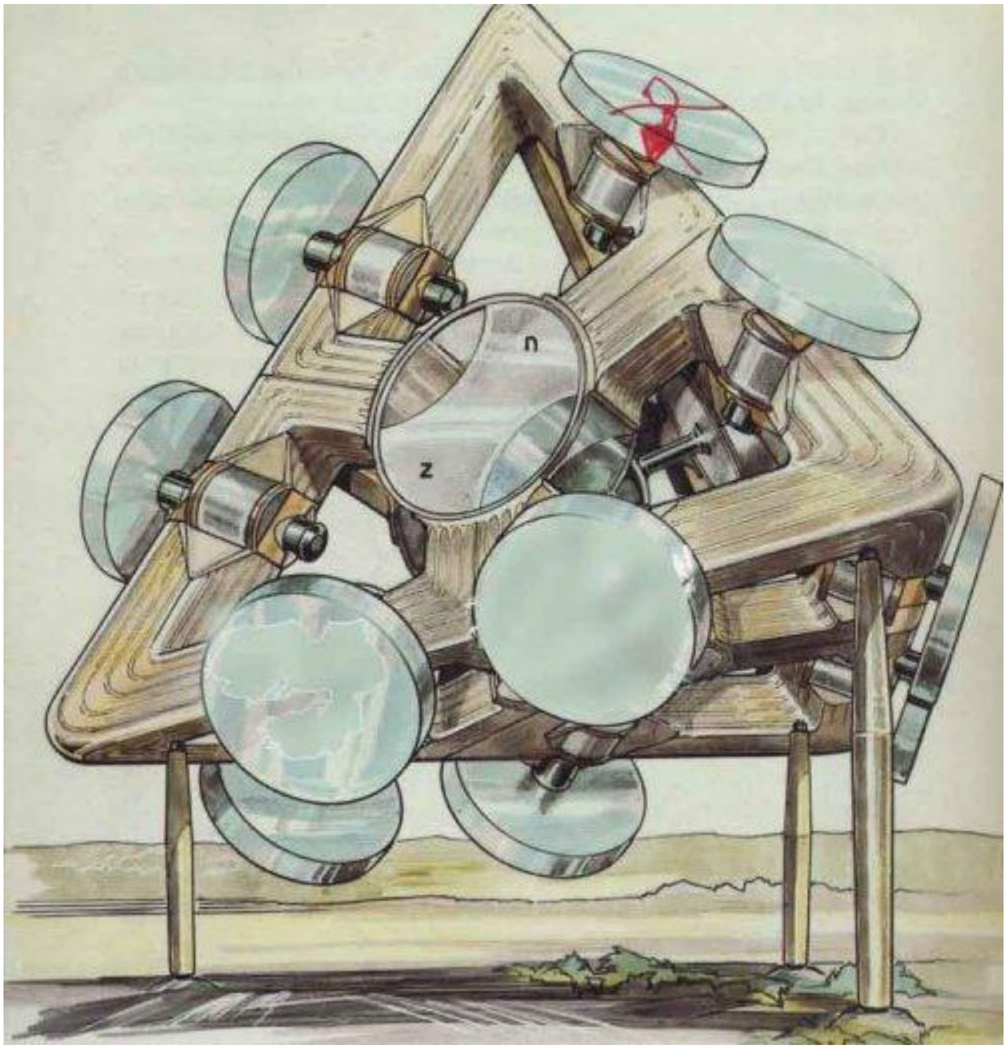

This is a painting by Rudolf Dass of the analogous visual-mechanical carrier-field model which symbolizes time and matter creation and synchronization in twelve universes. This giant working construction that was a cross between a monument and a machine was seen on an artificially flattened mountaintop on **Iarga**. This place was perceived as a place of pilgrimage, a place of meditation and contemplation, and the machine was a symbol of creation. The **Iarga**ns know no separation between science and religion, and religious and spiritual theories are entirely scientific. The model was several hundred yards(meters) long. The machine seemed to operate by some kind of perpetual motion, neither running on fuel or electrical power.

The Instructional Analog Model

The mechanical instruction models are the visual representations of the principles of the cosmic unified field theory. The model is intended to demonstrate that time, matter, and energy are creations out of the nothing and therefore the counter-balance has to be maintained to allow the causal infinity of the nothing to exist. In a broader context the physical symbols all take on new meaning and demonstrate the unification of science and religion. In a religious sense the spherical three-dimensional time-field in the heart of the pyramid symbolizes the Soul of the Creator. The pyramid symbolizes his infinite lovewill, the cage in which the infinity of His existence has been blocked irrevocably. The lines symbolize the different frequencies of the timeless waves. The flywheels on the edges of the pyramid symbolize the twelve associated and interrelated universes, including ours, that make up this unity. One cannot exist without the others.

The timeless primeval atom-principle, the transmitter function.

The atom-field is a time-creating field and as such a remote manifestation of the primeval atom principle. It can be compared with a transmitter. As soon as the program is converted into waves of great power, billions of radios can receive the same program, the number is unlimited. Everything depends on the one program of the transmitter, the primeval atom-field. If this exists with infinite strength, it is at the same time the vibration medium (the carrier-field) which helps it to spread everywhere. *Thus any atom in any universe can receive the timeless basic program.* The second model shows the three-dimensional time-field of the primeval atom field most clearly, although this is a three-dimensional rotation of the central sphere in spatial sense. This is an analog of reality, the atom has a spherical field which is defined by six linear time-vibration directions. Any atom is a remote manifestation of the one carrier-field of infinite strength.

Considering this fifth function of the carrier-field one has to realize that only the central sphere, the atom, is visible to us. The fields in the pyramid edges outside the sphere are invisible. It is the time-field reality of which the atom is a reflection.

The mechanical instruction models are the visual representations of the principles of the cosmic unified field theory. This requires however some marginal notes. A material comparison to explain spiritual abstractions is always in some ways incomplete. There are several hiatus of which only the three most important ones are mentioned.

1. In the model one time-axis represents only *one frequency* of the magnetic field, while in reality all frequencies are contained. Every frequency however can be initiated for each time-axis for which the interference representation of the analog model is valid.

2. The model demonstrates the rotating field of the *timeless waves*. It *does not imitate the time-bound waves* by which our time-axis the propagation speed of time is X, then the phase velocity of the two time-axes connected to "our" pyramid edge is ½X. The other two time-axes connecting our time-axis with the other corner of the pyramid is -½X, making up for the time speed of our anti-universe -X. The crossing, not-touching time-axis is perpendicular to our time-axis and does not contribute.

3. In our model the masses of the universes are given by the flywheel symbol. In fact this is *fundamentally wrong, because it is separated from the three-dimensional rotation of the central sphere.* It can therefore basically not maintain the counterbalance of the nothing.

The most important conclusion of the analog model is that *time, matter and energy are creations out of the nothing and therefore the counter-balance has to be maintained to allow the causal infinity of the nothing exist.*

14. The existential symbolism of the model.

So far the physical aspects of the model have been unravelled, but if one wants to give a wider context to the model suddenly all the physical symbols change their meaning. In relation to philosophical or theological disciplines the model takes on the additional meaning of the unification of science and religion. By certain inherent (medfødte) laws within any intelligent race the denial of God's existence will be destroyed sooner or later. Earth science will provide an indestructible contribution to our final knowledge-awareness, which we shall attribute as typical human attainment in the cosmic integration. It is that piece of our human identity and self-awareness, which we shall possess and care for throughout eternity. *It the light of this purpose, science serves God and the distinction between science and religion fades away*.

In a religious sense the spherical three-dimensional time-field in the heart of the pyramid has a special meaning. **It symbolizes the Soul of the Creator**, the Father who is the pure Identity of the infinite origin. The pyramid symbolizes his infinite Love-will, the cage in which the infinity of his Existence has been blocked irrevocably. He sacrificed his Ego and his freedom, to love his twelve Sons once. He is present only indirectly in his Creation as the reflection of the infinity, the Supreme Being, as the one Ego-consciousness of the Archangels.

That is why the theologian covers the central sphere with pure gold, symbolizing the golden Heart, the Holy of Holies, the most sacred secret of the universe. The pyramid is then painted dark red, the universal colour of love. Over the red many thin golden stripes run in the direction along the pyramid edges. These stripes symbolize the different frequencies of the timeless waves, the expressions of Love and faith of the Father and his infinite Will.

Gathered around the golden Heart are the twelve flywheels, coloured white, the colour of purity and innocence. The model shows that every universe has the power to awaken the matter-consciousness and to transform it to the new God: the Son. The goal of Creation is therefore the awakening and the formation of the twelve Sons of the Father.

This is not in contradiction with the Bible telling us that the Father sent his only-begotten Son. The Bible is solely directed to man. It is our instruction book for life describing our attitude towards God and our fellow man. The Bible does not tell anything about life on other planets, because it is not relevant in our relationship with God now.

The same goes for our parallel universes. In fact they do not exist for us, because they are not in our world of existence but somewhere else. They do not exist for all intelligent races in this universe, we shall never be able to observe them nor experience them in any way. Perhaps at the end of all time we shall meet.
In *our* world the Father has one only-begotten Son and this truth is already so unimaginable that nobody can conceive the consequences yet.

END OF IARGAN DISCUSSION

Stefan Denaerde

Cosmic Unified Field Theory.

How it started.

More than 25 years ago appeared an amazing book in the Netherlands: ´Buitenaardse Beschaving´ (Extraterrestrial Civilization) from Stefan Denaerde. (Known in English as : "UFO contact from planet Iarga") It describes the confrontation of the author with the race from the planet Jarga (Iarga). It was written as a novel (although it is a true story) and illustrated by a well known Dutch technical illustrator (Rudolf Das).

Sailing with his ship in Dutch waters, Mr. Denaerde (not his real name) bumped into something which was hidden underwater. When he sees someone floating in the water he comes to the rescue of what later appears to be an astronaut. This rescue was a test for him and after the initial shock he gets in contact with the Jargans who invite him for a three day learning session inside their submerged saucer.

Via a 3D screen and a radiation ´head set´ he learns about the Jargan culture, philosophy and technical achievements. Through the concepts of extreme efficiency and justice he is taken to the Jargans love for the Omni-Creativity (our word for ´God´ is too poor) and their striving to be one with it.
This ´carrier field´ of the Universe supports everything (matter, time, consiousnes), and by applying this knowledge it makes interstellar space travel possible.
It appears that the goal of a civilization is to create a socially stable supercivilization. Discriminations of any kind no longer exist. Resources and wealth are available for everyone, but not as a goal as such, but to enable everyone to grow spiritually.

There is much more to this book. Anyone interested can find an interview with Denaerde by Wendell Stevens, as well as several chapters from the book on this link. (This also contains the 2nd book of Denaerde!)
From this interview it appears that the contact has continued for many years in telepathic form. A lot of highly technical information has been given in this way, which resulted in the second book:

The Universal Creation. (In Dutch: ´de Universele Schepping´)

This appeared about 20 years after his first book. It describes in one theory the origin of the Universe, matter, time, consciousness, the Universal goal for any civilization, the history and possible (very bright) future of mankind, in one word : a TOE (theory of everything).
The atmosphere of the book is totally different: the first book had something like ´you don't have to believe this has happened, it could be hypothetical´. The Universal Creation is much more direct and confronts us with a direct proof of its contents, because the authors (Stefan and his ´brother´ Christian Denaerde) express that it is high time that science and religion are united again.
The Universal Creation comes with an appendix ´Cosmic Unified Theory´- in English!- which goes into much detail about the new physical model. Recently this information is available on CR-Rom, to be ordered at the Cosmic Field Paradigm foundation. This essay is mainly about the appendix and the CD-Rom.

The creation out of the nothing.

According to the Cosmic Field (CF) model there existed prior to the creation (or big bang) vibrations of infinite frequency and amplitude in every time direction.
The act of creation was the blocking of the infinite waves by interference to form waves of finite frequency and amplitude. In order for matter and energy or anything to exist, there must be an equal counterbalancing vibration. Also time direction has to be counterbalanced. Think of a time axis as two pieces of mini model rail which are connected via a motor, which can move the rails back and forth say over 1 cm. If you place a wagon

on one of the rails which has a catch wheel that catches only the forward movement of the rails, you will see the wagon move forward. Another wagon, placed in the opposite direction on the other rail will also move forwards, thus the two wagons will separate from each other. So the rails can be seen as a time axis, where time moves back and forth. The wagon represents matter, which can only catch one time direction through a property called spin.
One time axis is not sufficient: to counterbalance everything and to cause a a feedback mechanism a 3D time structure is needed in the form of a tetra-eder.

Fig. 1 The carrier field model

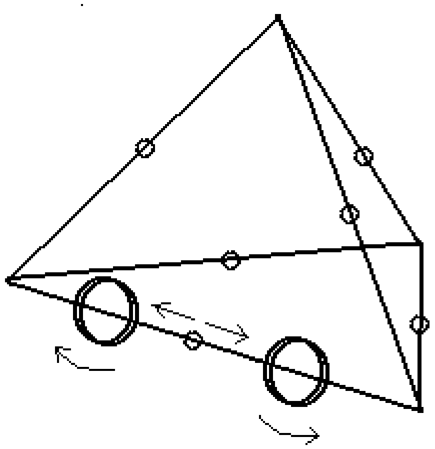

6 time axes form a closed loop. 12 universes are separated by time direction.
The spin of matter is represented by the two flywheels (only shown for one time axis)

This implies that there exist 6 time axes (each with a back and forward movement in time) and thus 12 universes. These universes cannot be seen, but some aspects like the time frequencies can be sensed by our atoms. (BTW, this might explain the missing mass problem in cosmology: our universe still senses the mass of the other 11 universes, that is why 90% of the mass is unexplained..)

The picture shows the arrangement of the time axes. The center of each axis (depicted by a small circle) can be considered as T=0. The two flywheels (drawn on one axis only) represent the spin direction of matter. In our direct anti-universe (on the same time axis) the direction of time is opposite, but also the spin of the matter particles. This means that everything is mirrored, and living beings in the anti-universe experience time as moving forward, just like we. In fact time moves in two directions along an axis, but matter has a sort of rectifying effect - it only picks up the forward going time.

The 60 degrees angle between the time axes is necessary to have an harmonic ratio between our time axis and the time axis of the other universes.

Fig.2 2 time axes at 60 degrees

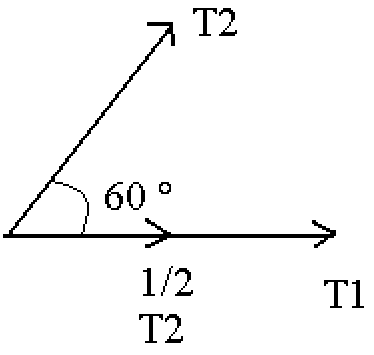

Only at this angle there can be a harmonic ratio between the time creating frequencies. An atom on our time axis T1 experiences the other time axis at half the frequency.

Our time direction is T1. Another time direction is T2. The angle of 60 degrees makes T1=2*T2, thus giving an harmonic ratio of 1:2.

The carrier field model has several functions, such as:

It is an infinitely strong force field, blocked in itself. It is a loss free conduction medium for all waves. The counter balancing of everything implies that the universes are locked in an iron grip: the spins in one universe are exactly the same and opposite to the spins of the anti universes. It also means that there cannot be one atom more in a universe, which is not present in the 11 other universes. (this does not mean that all the 12 universes have to be the same).

The model is also an image of the universal atom: When applied at a very small scale Denaerde explains the basic building blocks to form matter.

The flywheels can represent the total mass in one universe or just the mass of one atom. The 3D time structure of the pyramid is present in all matter: you could imagine the center of the pyramid being in a star, a galaxy, etc. The 11 other universes are not observable, but they could occupy the same 3D space. The model demonstrates beautifully that matter, time and energy are just properties of one universe (1/12th of the whole). Summing all the properties of the pyramid results in nothing. (The big bang theory fails to explain where the initial energy came from)

The Double rotating central bar.

Denaerde introduces the concept of a magnetic bar, which touches 2 opposite ribs in the pyramid. This induces a flux in the loss free time axis. This is a very important concept and it takes a while to understand it. The bar needs to rotate around two axes simultaneously (Fig. 3) in order to create alternating fluxes in the pyramid ribs (which still represent time axes). The rotating movement repeats itself in 6 time-steps.

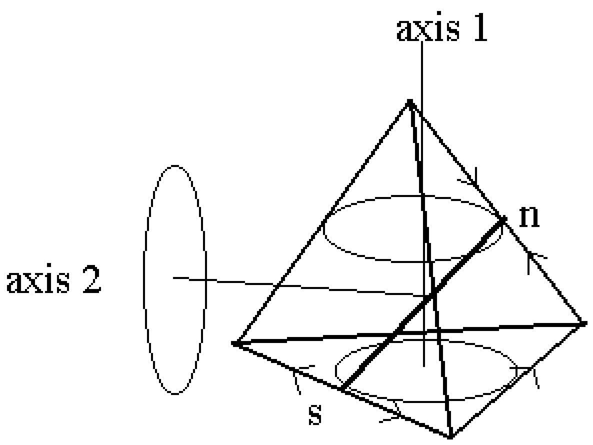

The simultaneous rotation of a central magnetic bar around 2 axes gives alternating flux patterns (time directions) in a pair of ribs. This process is repeated in 6 timesteps.

Fig.3 Double rotation of a central bar

Denaerde calls this configuration a building element. Next we will see that from these building elements one can derive the fundamental elementary particles : electron, quark and neutrino.

The two fundamental building blocks

By adding the initial building block at different time steps (adding three bar rotors) it can be derived that there are two possibilities. These are called central rotors of the first and second kind. (figures 4 and 5). The central rotor of the first kind gives a fixed alternating flux pattern in the ribs of the pyramid. (Try it out!)

Fig. 4 3 building blocks at different time moments are added

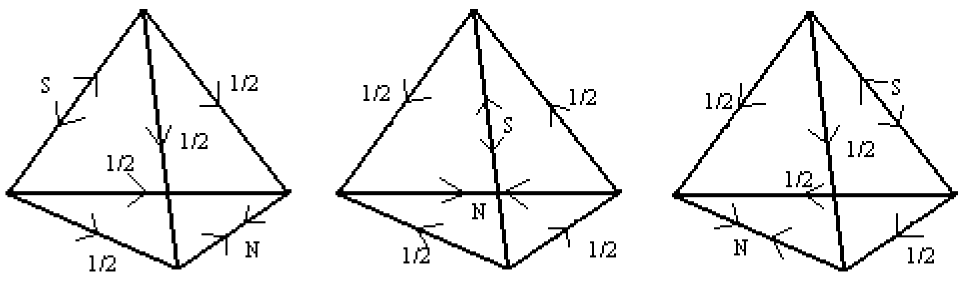
S
1/2
1/2
1/2
1/2
N
1/2
1/2
S
1/2
N
1/2
1/2
1/2
S
1/2
1/2
N
1/2

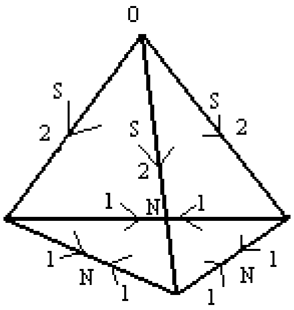
0
S
2
S
2
S
2
1
N
1
1
N
1
1
N
1

The total return flux gives a new building block: a rigid central barrotor of the first kind. The zero node does not change its orientation. In a next timestep the flux pattern is reversed, but the magnitude is the same.

This model describes the ELECTRON.

Fig. 5 The rigid central bar rotor of the second kind

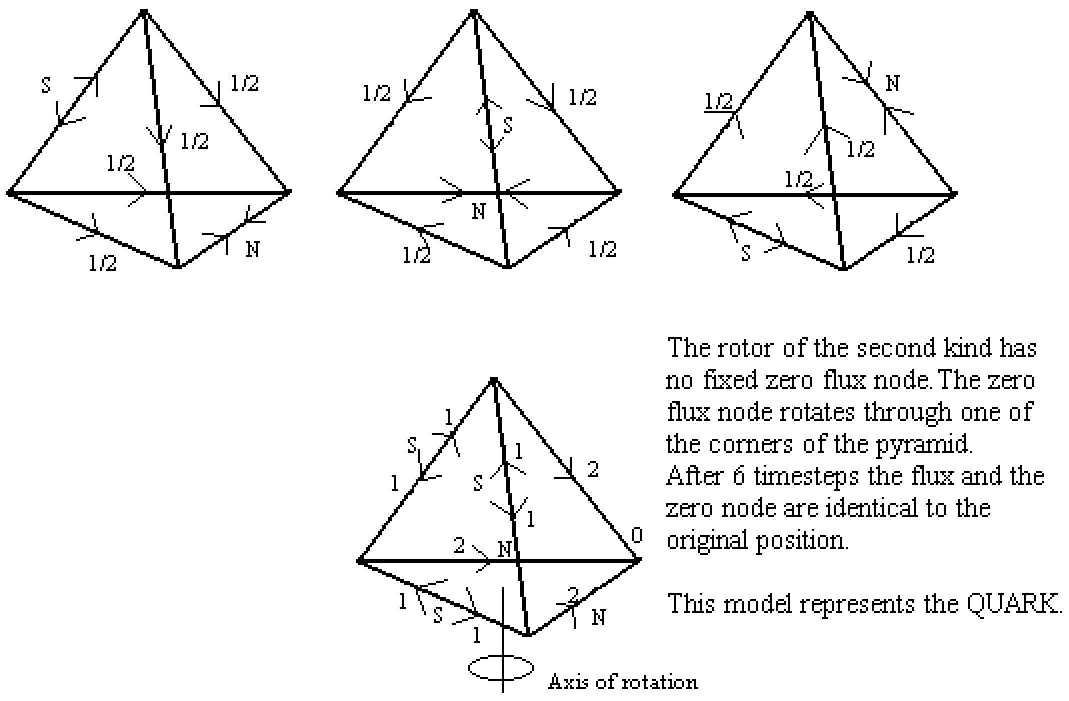

The rotor of the second kind has no fixed zero flux node. The zero flux node rotates through one of the corners of the pyramid.
After 6 timesteps the flux and the zero node are identical to the original position.

This model represents the QUARK.

The second possibility is called a central rotor of the second kind.

The rotor of the first kind is similar to the ELECTRON.
The rotor of the second kind is similar to the QUARK.

Don´t forget, the pyramid ribs represent time axes. (Note: Denaerde uses capitals to distinguish from the real particles in 3D space.)

Three QUARKS added together give a PROTON. (Figure 6)

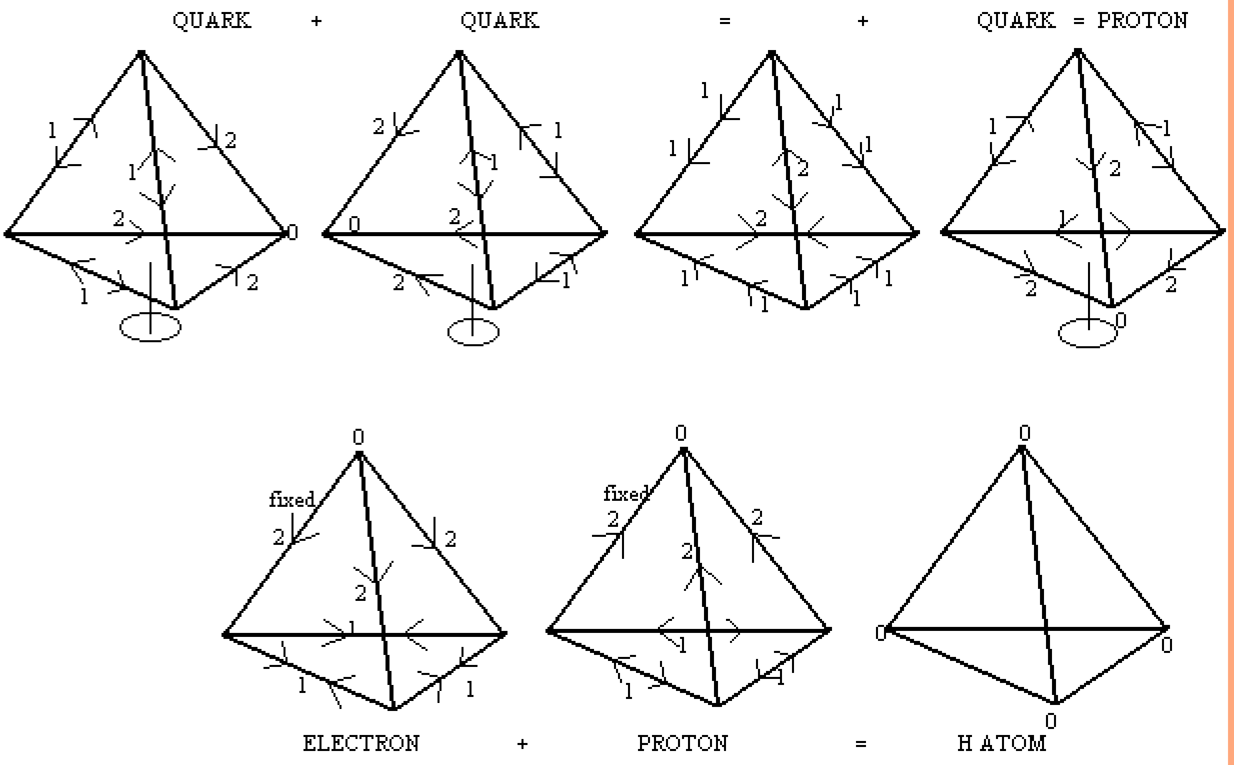

Fig 6 Quarks, electron and H atom

A QUARK and an "anti time" QUARK or the ELECTRON and the "anti time" ELECTRON (positron) can be distinguished by the c.w. or anti c.w. rotation of the central rotors. Dipoles of opposite rotation destroy each other.
Fig 6 shows that adding the PROTON and the ELECTRON gives the neutral H-ATOM.
The mass difference between electron and proton can be explained by different rotation frequencies of the dipoles.

Denaerde then continues to describe several NEUTRON groups which can be formed. This is rather complicated and requires a lot of pictures to visualize, so I would rather skip this. (For anyone interested, I strongly recommend to order the CD!)
The outcome is, that the NEUTRON in the Carrier Field model has 4 QUARKS.
This gives a NEUTRON with a zero external flux. The model explains why the neutron is not stable in free space, while the electron and proton are.

It can easily be seen that to the CF model a QUARK has a unit of charge of 1/3 of the ELECTRON, if charge is defined as the external flux in the pyramid ribs. Charge is related to a symmetry axis of the pyramid. The CF model predicts that charge is not constant in time and that there are two types of charge for a matter particle (e.g. electron): North and South charge. Due to the time switching mechanism we don´t detect this.

A NEUTRON consisting of 4 central rotors of the 2nd kind can be converted into a PROTON (3 rotors of the 2nd kind) and an ELECTRON (1 rotor of the 1st kind).
What is left over is the NEUTRINO : a free dipole in the carrier field.

In order for the particles to exist in space and time, a grid is needed which carries the complex flux patterns. Denaerde comes to a grid consisting of two pyramids stacked together. The top one is inverted and rotated 60 degrees, while its top touches the base of the other pyramid (sunken). The size of the pyramids is in the order of h (Plank´s constant, normalised to length). All particles have more or less the same size in parametric space (space as we know it).

Further Implications

We come now to the more profound meanings of the model. This subject I describe very condensed (without further proof, although the subject is described in great detail in the appendix).

The double rotating central bar has an additional degree of freedom. This results in a second gravitational field, which we cannot directly observe. So there are two kinds of gravity.
The CF model talks about ideal Black Holes and White Radiators. Black Holes have a function to stabilize the universes. As said before, the mass (energy) in all 12 universes must be exactly the same. Still the mass distribution can be different, but only thanks to black holes. B.H´s and W.R´s are present in the corners of the pyramid frame. In one time step, our time direction faces the Black Hole, while our anti-universe´s time direction faces the White Radiator. The next time step our time direction faces the White Radiator. In the next time step it is the null-energy flux.
White Radiators are generators of mass and energy, but have no mass of their own. It generates the exact opposite type of gravity which nullifies with normal gravity.
A pair of BH and WR belongs to each reality. The ideal Black Holes (and there counterpart the WR) in the CF model must have spin. The normal black hole as known from General Relativity Theorie is discussed also.

The formation of a pair of BH and WR (also called a macroscopic quantum mirror or black mirror) is then proven to be a non natural object : it could only have formed by an act of intelligence. Intelligent life is a necessary condition in the formation of the universe! According to the CF model, the black mirrors are formed by the collective mind power of an intelligent race.

Another important consequence is that the time history (knowledge) of the universe is laid down in the quantum gravitation fields. This is one of the reasons why the universe expands: to record the time history. In other words, the time history of a matter system, determined by an increase of entropy, is converted into information in the quantum gravity field. Intelligent life interferes with and modifies the quantum gravity fields (creativity).

The two types of gravity fields represent the matter-bound and emotional (or soul-bound) energy fields. The creation of the macroscopic quantum mirror is the final goal of an intelligent race. It is called the all-might manifestation of the race and it can only be initiated by enough minds working together to create the mirror - which means nothing less than the recognition and identification of the Love poles which carry the universe...

Note:

Everything written here is in my own words. Pictures were made by me. Therefore, copyright has not been violated.

Back

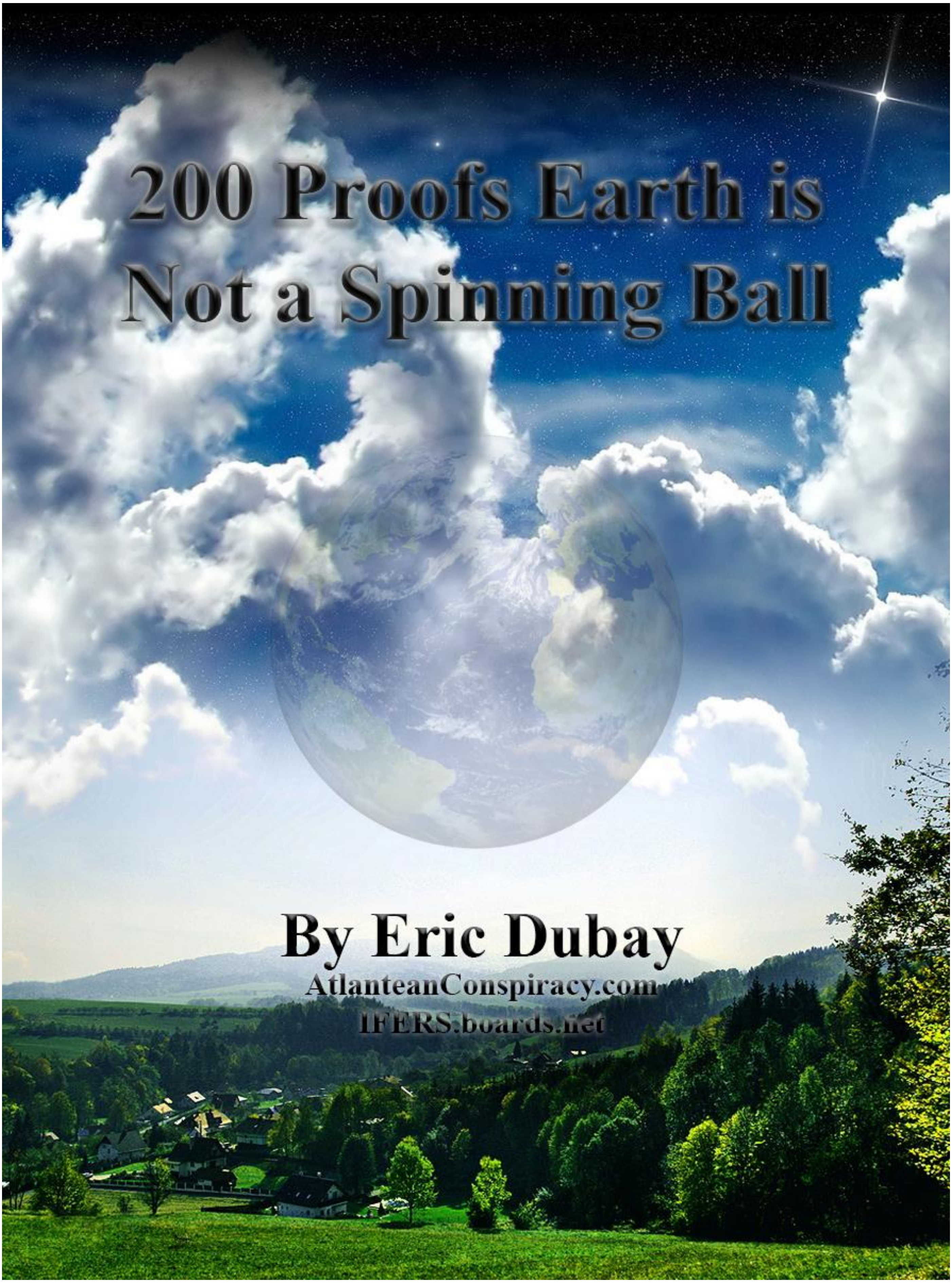

200 Proofs Earth is Not a Spinning Ball

By Eric Dubay

AtlanteanConspiracy.com
IFERS.boards.net

1) The horizon always appears perfectly flat 360 degrees around the observer regardless of altitude. All amateur balloon, rocket, plane and drone footage show a completely flat horizon over 20+ miles high. Only NASA and other government "space agencies" show curvature in their fake CGI photos/videos.

2) The horizon always rises to the eye level of the observer as altitude is gained, so you never have to look down to see it. If Earth were in fact a globe, no matter how large, as you ascended the horizon would stay fixed and the observer / camera would have to tilt looking down further and further to see it.

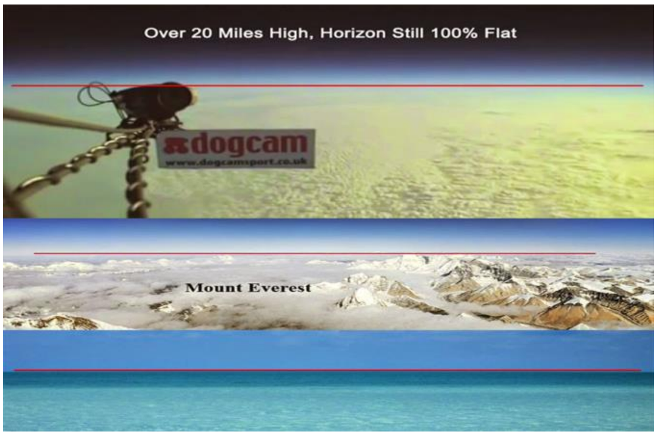

3) The natural physics of water is to find and maintain its level. If Earth were a giant sphere tilted, wobbling and hurdling through infinite space then truly flat, consistently level surfaces would not exist here. But since Earth is in fact an extended flat plane, this fundamental physical property of fluids finding and remaining level is consistent with experience and common sense.

4) Rivers run down to sea-level finding the easiest course, North, South, East, West and all other intermediary directions over the Earth at the same time. If Earth were truly a spinning ball then many of these rivers would be impossibly flowing uphill, for example the Mississippi in its 3000 miles would have to ascend 11 miles before reaching the Gulf of Mexico.

5) One portion of the Nile River flows for a thousand miles with a fall of only one foot. Parts of the West African Congo, according to the supposed inclination and movement of the ball-Earth, would be sometimes running uphill and sometimes down. This would also be the case for the Parana, Paraguay and other long rivers.

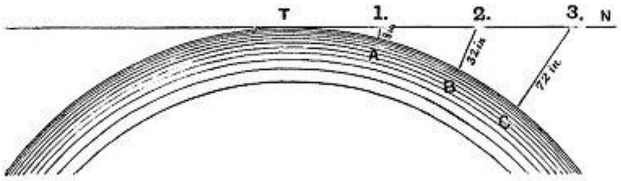

6) If Earth were a ball 25,000 miles in circumference as NASA and modern astronomy claim, spherical trigonometry dictates the surface of all standing water must curve downward an easily measurable 8 inches per mile multiplied by the square of the distance. This means along a 6 mile channel of standing water, the Earth would dip 6 feet on either end from the central peak. Every time such experiments have been conducted, however, standing water has proven to be perfectly level.

7) Surveyors, engineers and architects are never required to factor the supposed curvature of the Earth into their projects. Canals, railways, bridges and tunnels for example are always cut and laid horizontally, often over hundreds of miles without any allowance for curvature.

8) The Suez Canal connecting the Mediterranean with the Red Sea is 100 miles long without any locks making the water an uninterrupted continuation of the two seas. When constructed, the Earth's supposed curvature was not taken into account, it was dug along a horizontal datum line 26 feet below sea-level, passing through several lakes from one sea to the other, with the datum line and water's surface running perfectly parallel over the 100 miles.

9) Engineer, W. Winckler was published in the Earth Review regarding the Earth's supposed curvature, stating, "*As an engineer of many years standing, I saw that this absurd allowance is only permitted in school books. No engineer would dream of allowing anything of the kind. I have projected many miles of railways and many more of canals and the allowance has not even been thought of, much less allowed for. This allowance for curvature means this - that it is 8" for the first mile of a canal, and increasing at the ratio by the square of the distance in miles; thus a small navigable canal for boats, say 30 miles long, will have, by the above rule an allowance for curvature of 600 feet. Think of that and then please credit engineers as not being quite such fools. Nothing of the sort is allowed. We no more think of allowing 600 feet for a line of 30 miles of railway or canal, than of wasting our time trying to square the circle*"

10) The London and Northwestern Railway forms a straight line 180 miles long between London and Liverpool. The railroad's highest point, midway at Birmingham station, is only 240 feet above sea-level. If the world were actually a globe, however, curving 8 inches per mile squared, the 180 mile stretch of rail would form an arc with the center point at Birmingham raising over a mile, a full 5,400 feet above London and Liverpool.

11) A surveyor and engineer of thirty years published in the Birmingham Weekly Mercury stated, "*I am thoroughly acquainted with the theory and practice of civil engineering. However bigoted some of our professors may be in the theory of surveying according to the prescribed rules, yet it is well known amongst us that such theoretical measurements are INCAPABLE OF ANY PRACTICAL ILLUSTRATION. All our locomotives are designed to run on what may be regarded as TRUE LEVELS or FLATS. There are, of course, partial inclines or gradients here and there, but they are always accurately defined and must be carefully traversed. But anything approaching to eight inches in the mile, increasing as the square of the distance, COULD NOT BE WORKED BY ANY ENGINE THAT WAS EVER YET CONSTRUCTED. Taking one station with another all over England and Scotland, it may be stated that all the platforms are ON THE SAME RELATIVE LEVEL. The distance between Eastern and Western coasts of England may be set down as 300 miles. If the prescribed curvature was indeed as represented, the central stations at Rugby or Warwick ought to be close upon three miles higher than a chord drawn from the two extremities. If such was the case there is not a driver or stoker within the Kingdom that would be found to take charge of the train. We can only laugh at those of your readers who seriously give us credit for such venturesome exploits, as running trains round spherical curves. Horizontal curves on levels are dangerous enough, vertical curves would be a thousand times worse, and with our rolling stock constructed as at present* ***physically impossible***."

12) The Manchester Ship Canal Company published in the Earth Review stated, "*It is customary in Railway and Canal constructions for all levels to be referred to a datum which is nominally horizontal and is so shown on all sections.* ***It is not the practice in laying out Public Works to make allowances for the curvature of the earth***

13) In a 19th century French experiment by M. M. Biot and Arago a powerful lamp with good reflectors was placed on the summit of Desierto las Palmas in Spain and able to be seen all the way from Camprey on the Island of Iviza. Since the elevation of the two points were identical and the distance between covered nearly 100 miles, if Earth were a ball 25,000 miles in circumference, the light should have been more than 6600 feet, a mile and a quarter, below the line of sight!

14) The Lieutenant-Colonel Portlock experiment used oxy-hydrogen Drummond's lights and heliostats to reflect the sun's rays across stations set up across 108 miles of St. George's Channel. If the Earth were actually a ball 25,000 miles in circumference, Portlock's light should have remained hidden under a mile and a half of curvature.

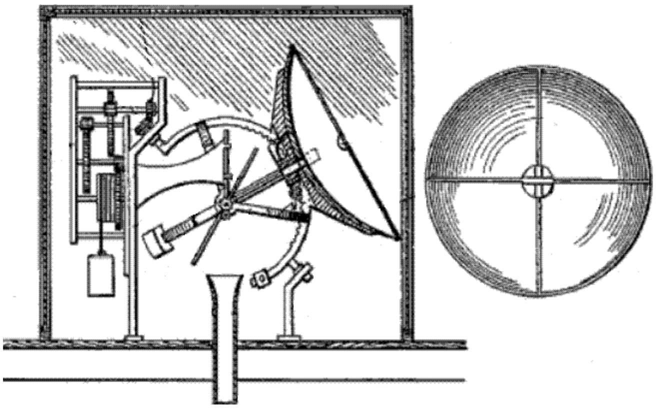

15) If the Earth were truly a sphere 25,000 miles in circumference, airplane pilots would have to constantly correct their altitudes downwards so as to not fly straight off into "outer space;" a pilot wishing to simply maintain their altitude at a typical cruising speed of 500 mph, would have to constantly dip their nose downwards and descend 2,777 feet (over half a mile) every minute! Otherwise, without compensation, in one hour's time the pilot would find themselves 31.5 miles higher than expected.

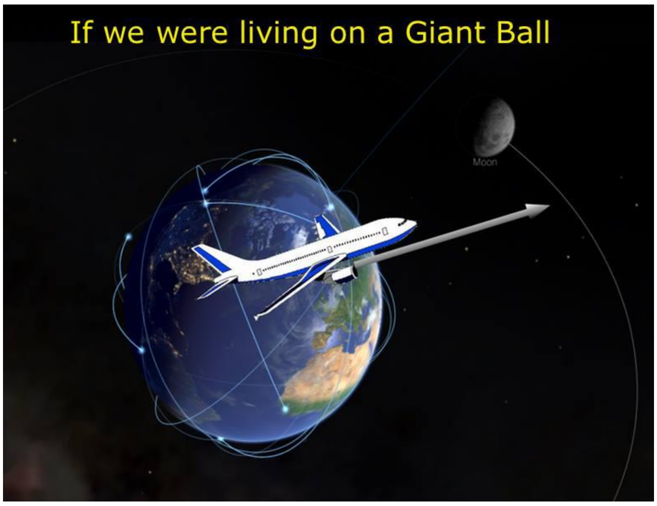
If we were living on a Giant Ball
Moon

16) The experiment known as "Airy's Failure" proved that the stars move relative to a stationary Earth and not the other way around. By first filling a telescope with water to slow down the speed of light inside, then calculating the tilt necessary to get the starlight directly down the tube, Airy failed to prove the heliocentric theory since the starlight was already coming in the correct angle with no change necessary, and instead proved the geocentric model correct.

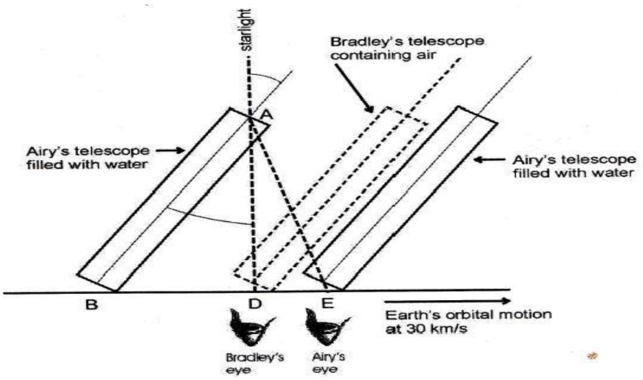
starlight
Bradley's telescope containing air
A
Airy's telescope filled with water
Airy's telescope filled with water
B
D
E
Earth's orbital motion at 30 km/s
Bradley's eye
Airy's eye

17) "Olber's Paradox" states that if there were billions of stars which are suns the night sky would be filled completely with light. As Edgar Allen Poe said, "*Were the succession of stars endless, then the background of the sky would present us a uniform luminosity, since there could exist absolutely no point, in all that background, at which would not exist a star*." In fact Olber's "Paradox" is no more a paradox than George Airy's experiment was a "failure." Both are actually excellent refutations of the heliocentric spinning ball model.

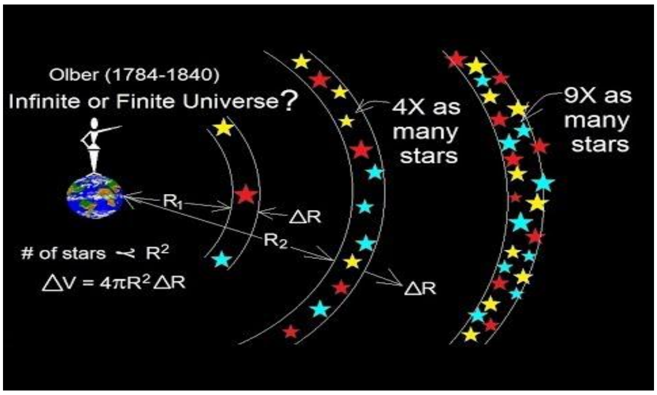
Olber (1784-1840)
Infinite or Finite Universe?
4X as many stars
9X as many stars
R_1
ΔR
R_2
ΔR
# of stars $\prec R^2$
$\Delta V = 4\pi R^2 \Delta R$

18) The Michelson-Morley and Sagnac experiments attempted to measure the change in speed of light due to Earth's assumed motion through space. After measuring in every possible different direction in various locations they failed to detect any significant change whatsoever, again proving the stationary geocentric model.

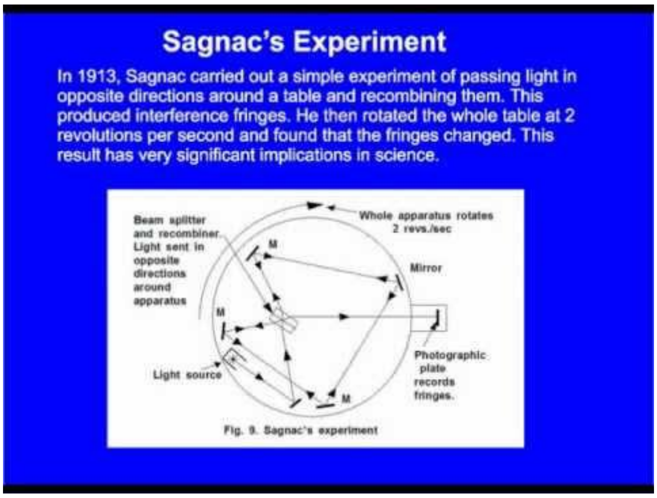

Sagnac's Experiment

In 1913, Sagnac carried out a simple experiment of passing light in opposite directions around a table and recombining them. This produced interference fringes. He then rotated the whole table at 2 revolutions per second and found that the fringes changed. This result has very significant implications in science.

Beam splitter and recombiner. Light sent in opposite directions around apparatus
Whole apparatus rotates 2 revs./sec
M
Mirror
M
Light source
M
Photographic plate records fringes.

Fig. 9. Sagnac's experiment

19) Tycho Brahe famously argued against the heliocentric theory in his time, positing that if the Earth revolved around the Sun, the change in relative position of the stars after 6 months orbital motion could not fail to be seen. He argued that the stars should seem to separate as we approach and come together as we recede. In actual fact, however, after 190,000,000 miles of supposed orbit around the Sun, not a single inch of parallax can be detected in the stars, proving we have not moved at all.

20) If Earth were truly constantly spinning Eastwards at over 1000mph, vertically-fired cannonballs and other projectiles should fall significantly due west. In actual fact, however, whenever this has been tested, vertically-fired cannonballs shoot upwards an average of 14 seconds ascending, 14 seconds descending, and fall back to the ground no more than 2 feet away from the cannon, often directly back into the muzzle.

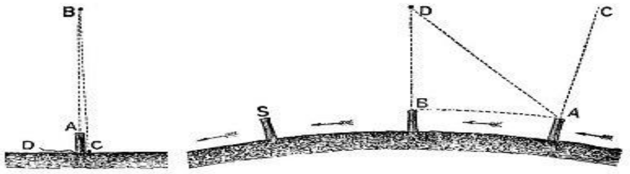

21) If the Earth were truly constantly spinning Eastwards at over 1000mph, helicopters and hot-air balloons should be able to simply hover over the surface of the Earth and wait for their destinations to come to them!

22) If Earth were truly constantly spinning Eastwards at over 1000mph, during the Red Bull stratosphere dive, Felix Baumgartner, spending 3 hours ascending over New Mexico, should have landed 2500 miles West into the Pacific Ocean but instead landed a few dozen miles East of the take-off point.

23) Ball-believers often claim "gravity" magically and inexplicably drags the entire lower-atmosphere of the Earth in perfect synchronization up to some undetermined height where this progressively faster spinning atmosphere gives way to the non-spinning, non-gravitized, non-atmosphere of infinite vacuum space. Such non-sensical theories are debunked, however, by rain, fireworks, birds, bugs, clouds, smoke, planes and projectiles all of which would behave very differently if both the ball-Earth and its atmosphere were constantly spinning Eastwards at 1000mph.

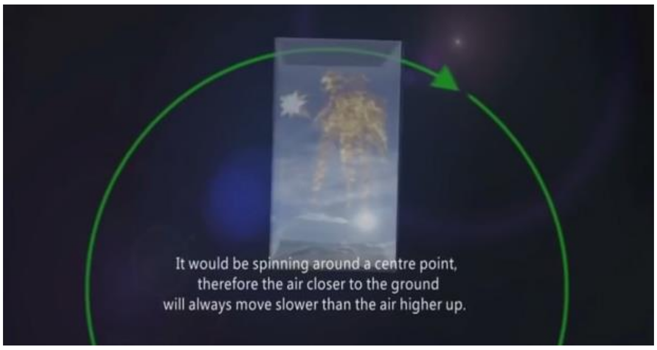

It would be spinning around a centre point, therefore the air closer to the ground will always move slower than the air higher up.

24) If Earth and its atmosphere were constantly spinning eastwards over 1000mph then North/South facing cannons should establish a control while East-firing cannonballs should fall significantly farther than all others while West-firing cannonballs should fall significantly closer. In actual fact, however, regardless of which direction cannons are fired, the distance covered is always the same.

25) If Earth and its atmosphere were constantly spinning eastwards over 1000mph, then the average commercial airliner traveling 500mph should never be able to reach its Eastward destinations before they come speeding up from behind! Likewise Westward destinations should be arrived at thrice the speed, but this is not the case.

26) Quoting "Heaven and Earth" by Gabrielle Henriet, "*If flying had been invented at the time of Copernicus, there is no doubt that he would have soon realized that his contention regarding the rotation of the earth was wrong, on account of the relation existing between the speed of an aircraft and that of the earth's rotation. If the earth rotates, as it is said, at 1,000 miles an hour, and a plane flies in the same direction at only 500 miles, it is obvious that its place of destination will be farther removed every minute. On the other hand, if flying took place in the direction opposite to that of the rotation, a distance of 1,500 miles would be covered in one hour, instead of 500, since the speed of the rotation is to be added to that of the plane. It could also be pointed out that such a flying speed of 1,000 miles an hour, which is supposed to be that of the earth's rotation, has* *recently been achieved, so that an aircraft flying at this rate in the same direction as that of the rotation could not cover any ground at all. It would remain suspended in mid-air over the spot from which it took off, since both speeds are equal.*"

27) If Earth and its atmosphere were constantly spinning Eastwards over 1000mph, landing airplanes on such fast-moving runways which face all manner of directions North, South, East, West and otherwise would be practically impossible, yet in reality such fictional concerns are completely negligible.

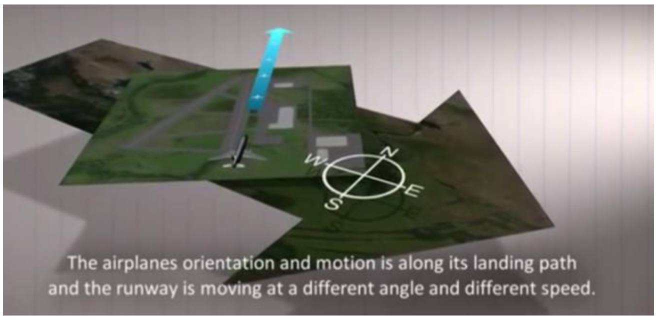
N
W
E
S
The airplanes orientation and motion is along its landing path
and the runway is moving at a different angle and different speed.

28) If the Earth and its atmosphere were constantly spinning Eastwards over 1000mph, then clouds, wind and weather patterns could not casually and unpredictably go every which way, with clouds often travelling in opposing directions at varying altitudes simultaneously.

29) If the Earth and its atmosphere were constantly spinning Eastwards over 1000mph, this should somewhere somehow be seen, heard, felt or measured by someone, yet no one in history has ever experienced this alleged Eastward motion; meanwhile, however, we can hear, feel and experimentally measure even the slightest Westward breeze.

30) In his book "South Sea Voyages," Arctic and Antarctic explorer Sir James Clarke Ross, described his experience on the night of November 27th, 1839 and his conclusion that the Earth must be motionless: *"The sky being very clear … it enabled us to observe the higher stratum of clouds to be moving in an exactly opposite direction to that of the wind--a circumstance which is frequently recorded in our meteorological journal both in the north-east and south-east trades, and has also often been observed by former voyagers. Captain Basil Hall witnessed it from the summit of the Peak of Teneriffe; and Count Strzelechi, on ascending the volcanic mountain of Kiranea, in Owhyhee, reached at 4000 feet an elevation above that of the trade wind, and experienced the influence of an opposite current of air of a different hygrometric and thermometric condition … Count Strzelechi further informed me of the following seemingly anomalous circumstance--that at the height of 6000 feet he found the current of air blowing at right angles to both the lower strata, also of a different hygrometric and thermometric condition, but warmer than the inter-stratum.* ***Such a state of the atmosphere is compatible only with the fact which other evidence has demonstrated, that the earth is at rest."***

31) Quoting "Zetetic Cosmogeny" Thomas Winships states: "*Let 'imagination' picture to the mind what force air would have which was set in motion by a spherical body of 8,000 miles in diameter, which in one hour was spinning round 1,000 mph, rushing through space at 65,000 mph and gyrating across the heavens? Then let 'conjecture' endeavor to discover whether the inhabitants on such a globe could keep their hair on? If the earth-globe rotates on its axis at the terrific rate of 1,000 miles an hour, such an immense mass would of necessity cause a tremendous rush of wind in the space it occupied. The wind would go all one way, and anything like clouds which got 'within the sphere of influence' of the rotating sphere, would have to go the same way.* ***The fact that the earth is at rest is proved by kite flying***."

32) If "gravity" is credited with being a force strong enough to hold the world's oceans, buildings, people and atmosphere stuck to the surface of a rapidly spinning ball, then it is impossible for "gravity" to also simultaneously be weak enough to allow little birds, bugs, and planes to take-off and travel freely unabated in any direction.

33) If "gravity" is credited with being a force strong enough to curve the massive expanse of oceans around a globular Earth, it would be impossible for fish and other creatures to swim through such forcefully held water.

34) Ship captains in navigating great distances at sea never need to factor the supposed curvature of the Earth into their calculations. Both Plane Sailing and Great Circle Sailing, the most popular navigation methods, use plane, not spherical trigonometry, making all mathematical calculations on the assumption that the Earth is perfectly flat. If the Earth were in fact a sphere, such an errant assumption would lead to constant glaring inaccuracies. Plane Sailing has worked perfectly fine in both theory and practice for thousands of years, however, and plane trigonometry has time and again proven **more** accurate than spherical trigonometry in determining distances across the oceans.

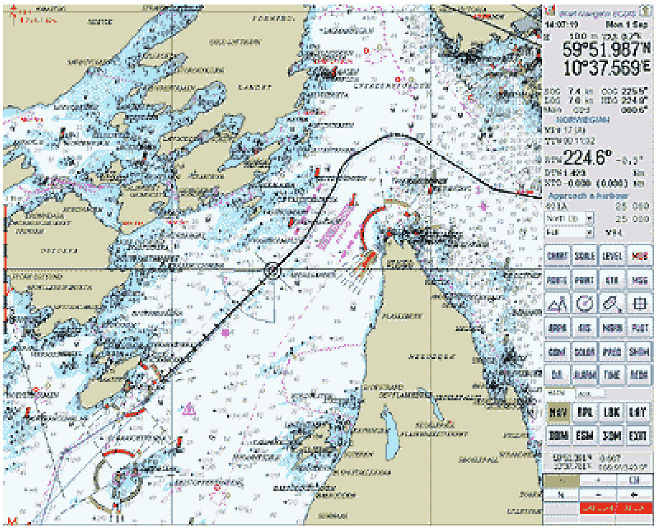

35) If the Earth were truly a globe, then every line of latitude south of the equator would have to measure a gradually smaller and smaller circumference the farther South travelled. If, however, the Earth is an extended plane, then every line of latitude south of the equator should measure a gradually larger and larger circumference the farther South travelled. The fact that many captains navigating south of the equator assuming the globular theory have found themselves drastically out of reckoning, moreso the farther South travelled, testifies to the fact that the Earth is not a ball.

GLEASON'S
NEW STANDARD MAP OF THE WORLD
"AS IT IS"
LONGITUDE AND TIME CALCULATOR

36) During Captain James Clark Ross's voyages around the Antarctic circumference, he often wrote in his journal perplexed at how they routinely found themselves out of accordance with their charts, stating that they found themselves an average of 12-16 miles outside their reckoning every day, later on further south as much as 29 miles.

37) Lieutenant Charles Wilkes commanded a United States Navy exploration expedition to the Antarctic from 1838 to 1842, and in his journals also mentioned being consistently east of his reckoning, sometimes over 20 miles in less than 18 hours.

38) To quote Reverend Thomas Milner, "*In the southern hemisphere, navigators to India have often fancied themselves east of the Cape when still west, and have been driven ashore on the African coast, which, according to their reckoning, lay behind them. This misfortune happened to a fine frigate, the Challenger, in 1845. How came Her Majesty's Ship 'Conqueror,' to be lost? How have so many other noble vessels, perfectly sound, perfectly manned, perfectly navigated, been wrecked in calm weather, not only in dark night, or in a fog, but in broad daylight and sunshine - in the former case upon the coasts, in the latter, upon sunken rocks - from being 'out of reckoning?*'" The simple answer is that Earth is not a ball.

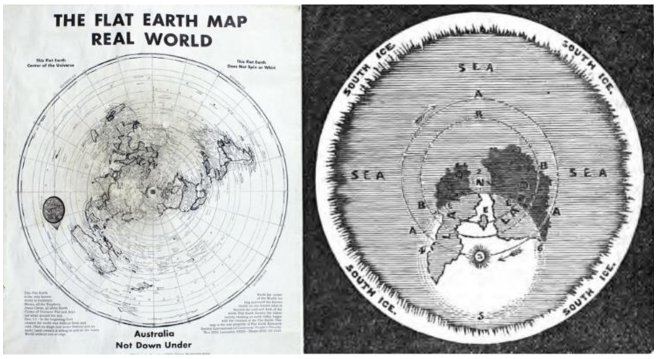

39) Practical distance measurements taken from "The Australian Handbook, Almanack, Shippers' and Importers' Directory" state that the straight line distance between Sydney and Nelson is 1550 statute miles. Their given difference in longitude is 22 degrees 2'14". Therefore if 22 degrees 2'14" out of 360 is 1550 miles, the entirety would measure 25,182 miles. This is not only larger than the ball-Earth is said to be at the equator, but a whole 4262 miles greater than it would be at Sydney's southern latitude on a globe of said proportions.

40) From near Cape Horn, Chile to Port Philip in Melbourne, Australia the distance is 10,500 miles, or 143 degrees of longitude away. Factoring in the remaining degrees to 360 makes for a total distance of 26,430 miles around this particular latitude, which is over 1500 miles wider than Earth is supposed to be at the equator, and many more thousands of miles wider than it is supposed to be at such Southern latitudes.

41) Similar calculations made from the Cape of Good Hope, South Africa to Melbourne, Australia at an average latitude of 35.5 degrees South, have given an approximate figure of over 25,000 miles, which is again equal to or greater than the Earth's supposed greatest circumference at the equator. Calculations from Sydney, Australia to Wellington, New Zealand at an average of 37.5 degrees South have given an approximate circumference of 25,500 miles, greater still! According to the ball-Earth theory, the circumference of the Earth at 37.5 degrees Southern latitude should be only 19,757 statute miles, almost six thousand miles less than such practical measurements.

42) In the ball-Earth model Antarctica is an ice continent which covers the bottom of the ball from 78 degrees South latitude to 90 and is therefore not more than 12,000 miles in circumference. Many early explorers including Captian Cook and James Clark Ross, however, in attempting Antarctic circumnavigation took 3 to 4 years and clocked 50-60,000 miles around. The British ship Challenger also made an indirect but complete circumnavigation of Antarctica traversing 69,000 miles. This is entirely inconsistent with the ball model.

43) If Earth was a ball there are several flights in the Southern hemisphere which would have their quickest, straightest path over the Antarctic continent such as Santiago, Chile to Sydney, Australia. Instead of taking the shortest, quickest route in a straight line over Antarctica, all such flights detour all manner of directions away from Antarctica instead claiming the temperatures too cold for airplane travel! Considering the fact that there are plenty of flights to/from/over Antarctica, and NASA claims to have technology keeping them in conditions far colder (and far hotter) than any experienced on Earth, such an excuse is clearly just an excuse, and these flights aren't made because they are impossible.

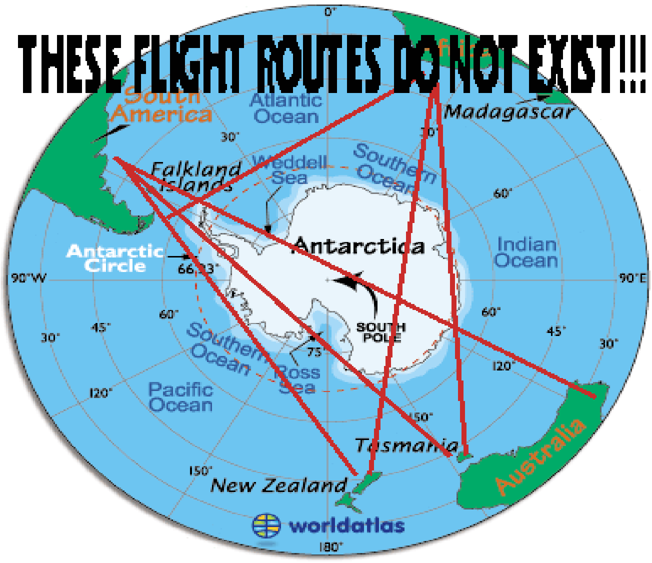

44) If Earth was a ball, and Antarctica was too cold to fly over, the only logical way to fly from Sydney to Santiago would be a straight shot over the Pacific staying in the Southern hemisphere the entire way. Re-fueling could be done in New Zealand or other Southern hemisphere destinations along the way if absolutely necessary. In actual fact, however, Santiago-Sydney flights go into the Northern hemisphere making stop-overs at LAX and other North American airports before continuing back down to the Southern hemisphere. Such ridiculously wayward detours make no sense on the globe but make perfect sense and form nearly straight lines when shown on a flat Earth map.

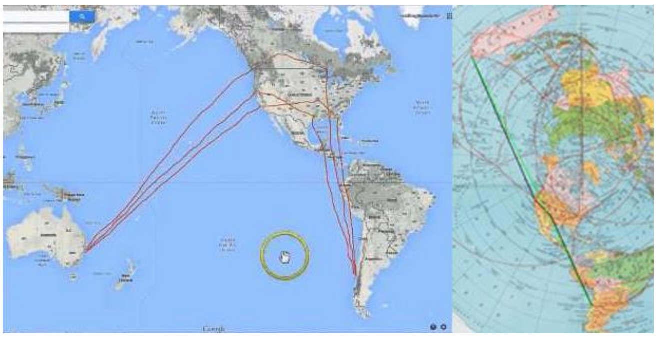

45) On a ball-Earth, Johannesburg, South Africa to Perth, Australia should be a straight shot over the Indian Ocean with convenient re-fueling possibilities on Mauritus or Madagascar. In actual practice, however, most Johannesburg to Perth flights curiously stop over either in Dubai, Hong Kong or Malaysia all of which make no sense on the ball, but are completely understandable when mapped on a flat Earth.

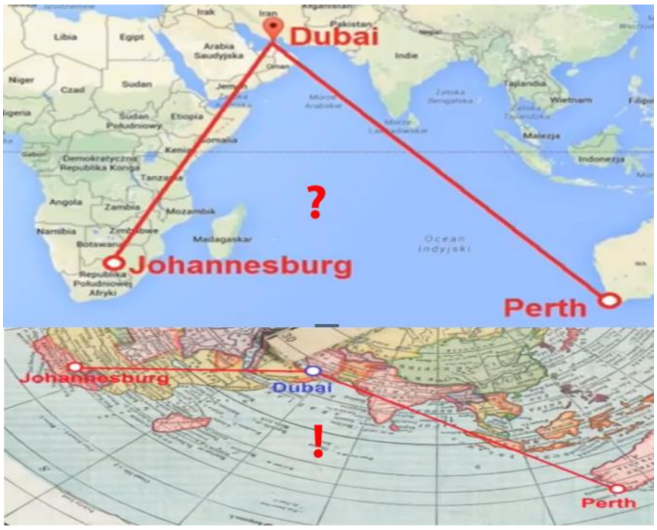

46) On a ball-Earth Cape Town, South Africa to Buenos Aries, Argentina should be a straight shot over the Atlantic following the same line of latitude across, but instead every flight goes to connecting locations in the Northern hemisphere first, stopping over anywhere from London to Turkey to Dubai. Once again these make absolutely no sense on the globe but are completely understandable options when mapped on a flat Earth.

47) On a ball-Earth Johannesburg, South Africa to Sao Paolo, Brazil should be a quick straight shot along the 25th Southern latitude, but instead nearly every flight makes a re-fueling stop at the 50th degree North latitude in London first! The only reason such a ridiculous stop-over works in reality is because the Earth is flat.

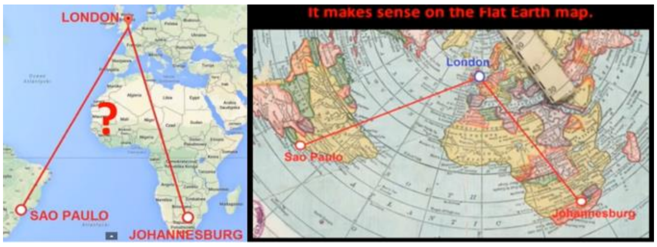

48) On a ball-Earth Santiago, Chile to Johannesburg, South Africa should be an easy flight all taking place below the Tropic of Capricorn in the Southern hemisphere, yet every listed flight makes a curious re-fueling stop in Senegal near the Tropic of Cancer in the North hemisphere first! When mapped on a flat Earth the reason why is clear to see, however, Senegal is actually directly in a straight-line path half-way between the two.

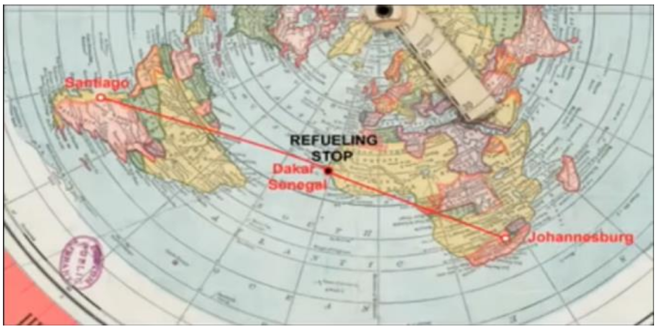

49) If Earth were a spinning ball heated by a Sun 93 million miles away, it would be impossible to have simultaneously sweltering summers in Africa while just a few thousand miles away bone-chilling frozen Arctic/Antarctic winters experiencing little to no heat from the Sun whatsoever. If the heat from the Sun traveled 93,000,000 miles to the Sahara desert, it is absurd to assert that another 4,000 miles (0.00004%) further to Antarctica would completely negate such sweltering heat resulting in such drastic differences.

50) If the Earth were truly a globe, the Arctic and Antarctic polar regions and areas of comparable latitude North and South of the equator should share similar conditions and characteristics such as comparable temperatures, seasonal changes, length of daylight, plant and animal life. In reality, however, the Arctic/Antarctic regions and areas of comparable latitude North/South of the equator differ greatly in many ways entirely inconsistent with the ball model and entirely consistent with the flat model.

ARCTIC
ANTARCTIC

51) Antarctica is by far the coldest place on Earth with an average annual temperature of approximately -57 degrees Fahrenheit, and a record low of -135.8! The average annual temperature at the North Pole, however, is a comparatively warm 4 degrees. Throughout the year, temperatures in the Antarctic vary less than half the amount at comparable Arctic latitudes. The Northern Arctic region enjoys moderately warm summers and manageable winters, whereas the Southern Antarctic region never even warms enough to melt the perpetual snow and ice. On a tilting, wobbling, ball-Earth spinning uniformly around the Sun, Arctic and Antarctic temperatures and seasons should not vary so greatly.

52) Iceland at 65 degrees North latitude is home to 870 species of native plants and abundant various animal life. Compare this with the Isle of Georgia at just 54 degrees South latitude where there are only 18 species of native plants and animal life is almost non-existent. The same latitude as Canada or England in the North where dense forests of various tall trees abound, the infamous Captain Cook wrote of Georgia that he was unable to find a single shrub large enough to make a toothpick! Cook wrote, “*Not a tree was to be seen. The lands which lie to the south are doomed by nature to perpetual frigidness - never to feel the warmth of the sun’s rays; whose horrible and savage aspect I have not words to describe. Even marine life is sparse in certain tracts of vast extent, and the sea-bird is seldom observed flying over such lonely wastes. The contrasts between the limits of organic life in Arctic and Antarctic zones is very remarkable and significant.*”

Arctic Winter, Arctic Summer

53) At places of comparable latitude North and South, the Sun behaves very differently than it would on a spinning ball Earth but precisely how it should on a flat Earth. For example, the longest summer days North of the equator are much longer than those South of the equator, and the shortest winter days North of the equator are much shorter than the shortest South of the equator. This is inexplicable on a uniformly spinning, wobbling ball Earth but fits exactly on the flat model with the Sun traveling circles over and around the Earth from Tropic to Tropic.

54) At places of comparable latitude North and South, dawn and dusk happen very differently than they would on a spinning ball, but precisely how they should on a flat Earth. In the North dawn and dusk come slowly and last far longer than in the South where they come and go very quickly. Certain places in the North twilight can last for over an hour while at comparable Southern latitudes within a few minutes the sunlight completely disappears. This is inexplicable on a uniformly spinning, wobbling ball Earth but is exactly what is expected on a flat Earth with the Sun traveling faster, wider circles over the South and slower, narrower circles over the North.

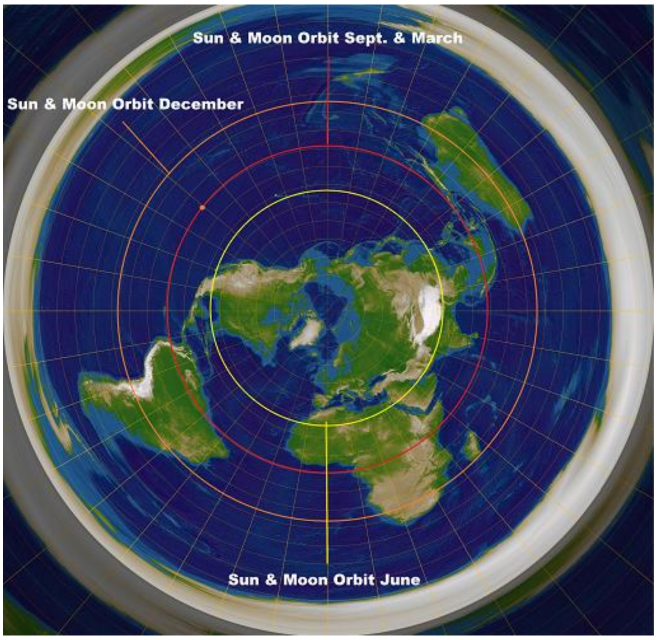

55) If the Sun circles over and around the Earth every 24 hours, steadily travelling from Tropic to Tropic every 6 months, it follows that the Northern, central region would annually receive far more heat and sunlight than the Southern circumferential region. Since the Sun must sweep over the larger Southern region in the same 24 hours it has to pass over the smaller Northern region, its passage must necessarily be proportionally faster as well. This perfectly explains the differences in Arctic/Antarctic temperatures, seasons, length of daylight, plant and animal life; this is why the Antarctic morning dawn and evening twilight are very abrupt compared with the North; and this explains why many midsummer Arctic nights the Sun does not set at all!

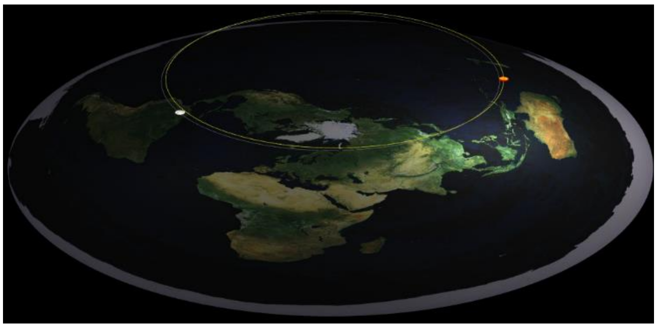

56) The “Midnight Sun” is an Arctic phenomenon occurring annually during the summer solstice where for several days straight an observer significantly far enough north can watch the Sun traveling circles over-head, rising and falling in the sky throughout the day, but never fully setting for upwards of 72+ hours! If the Earth were actually a spinning globe revolving around the Sun, the only place such a phenomenon as the Midnight Sun could be observed would be at the poles. Any other vantage point from 89 degrees latitude downwards could never, regardless of any tilt or inclination, see the Sun for 24 hours straight. To see the Sun for an entire revolution on a spinning globe at a point other than the poles, you would have to be looking through miles and miles of land and sea for part of the revolution!

57) The establishment claims the Midnight Sun IS experienced in Antarctica but they conveniently do not have any uncut videos showing this, nor do they allow independent explorers to travel to Antarctica during the winter solstice to verify or refute these claims. Conversely, there are dozens of uncut videos publicly available showing the Arctic Midnight Sun and it has been verified beyond any shadow of a doubt.

58) The Royal Belgian Geographical Society in their "Expedition Antarctique Belge," recorded that during the most severe part of the Antarctic winter, from 71 degrees South latitude onwards, the sun sets on May 17th and is not seen above the horizon again until July 21st! This is completely at odds with the ball-Earth theory, but easily explained by the flat-Earth model. The Midnight Sun is seen from high altitudes in extreme Northern latitudes during Arctic summer because the Sun, at its inner-most cycle, is circling tightly enough around the polar center that it remains visible above the horizon for someone at such a vantage point. Likewise, in extreme Southern latitudes during Arctic summer, the Sun completely disappears from view for over 2 months because there at the Northern Tropic, at the inner-most arc of its boomerang journey, the Sun is circling the Northern center too tightly to be seen from the Southern circumference.

59) Quoting Gabrielle Henriet, "*The theory of the rotation of the earth may once and for all be definitely disposed of as impracticable by pointing out the following inadvertence. It is said that the rotation takes twenty-four hours and that its speed is uniform, in which case, necessarily, days and nights should have an identical duration of twelve hours each all the year round. The sun should invariably rise in the morning and set in the evening at the same hours, with the result that it would be the equinox every day from the 1st of January to the 31st of December. One should stop and reflect on this before saying that the earth has a movement of rotation. How does the system of gravitation account for the seasonal variations in the lengths of days and nights if the earth rotates at a uniform speed in twenty-four hours!?*"

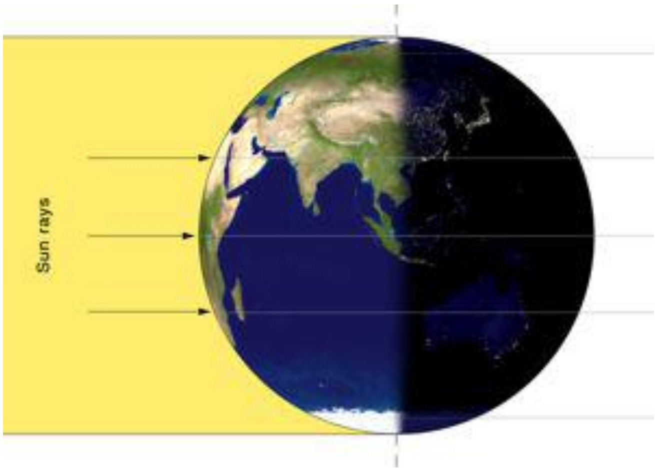

60) Anyone can prove the sea-horizon perfectly straight and the entire Earth perfectly flat using nothing more than a level, tripods and a wooden plank. At any altitude above sea-level, simply fix a 6-12 foot long, smooth, leveled board edgewise upon tripods and observe the skyline from eye-level behind it. The distant horizon will always align perfectly parallel with the upper edge of the board. Furthermore, if you move in a half-circle from one end of the board to the other whilst observing the skyline over the upper edge, you will be able to trace a clear, flat 10-20 miles depending on your altitude. This would be impossible if the Earth were a globe 25,000 miles in circumference; the horizon would align over the center of the board but then gradually, noticeably decline towards the extremities. Just ten miles on each side would necessitate an easily visible curvature of 66.6 feet from each end to the center.

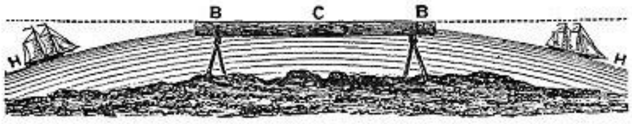

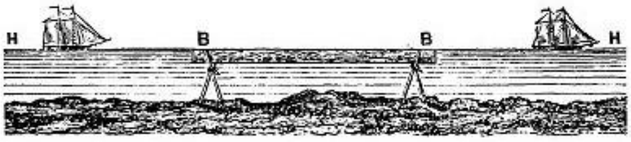

61) If the Earth were actually a big ball 25,000 miles in circumference, the horizon would be noticeably curved even at sea-level, and everything on or approaching the horizon would appear to tilt backwards slightly from your perspective. Distant buildings along the horizon would all look like leaning towers of Piza falling away from the observer. A hot-air balloon taking off then drifting steadily away from you, on a ball-Earth would slowly and constantly appear to lean back more and more the farther away it flew, the bottom of the basket coming gradually into view as the top of the balloon disappears from sight. In reality, however, buildings, balloons, trees, people, anything and everything at right angles to the ground/horizon remains so regardless the distance or height of the observer.

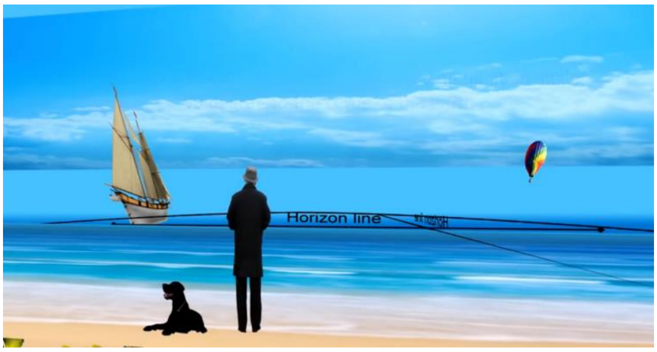

62) Samuel Rowbotham's experiments at the Old Bedford Level proved conclusively the canal's water to be completely flat over a 6 mile stretch. First he stood in the canal with his telescope held 8 inches above the surface of the water, then his friend in a boat with a 5 foot tall flag sailed the 6 miles away. If Earth were a ball 25,000 miles in circumference the 6 mile stretch of water should have comprised an arc exactly 6 feet high in the middle, so the entire boat and flag should have ultimately disappeared, when in fact the entire boat and flag remained visible at the same height for the entire journey.

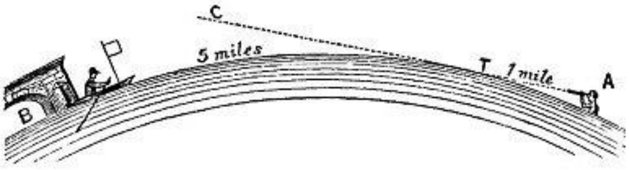

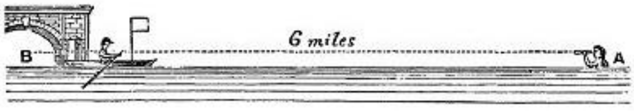

63) In a second experiment Dr. Rowbotham affixed flags 5 feet high along the shoreline, one at every mile marker. Then using his telescope mounted at 5 feet just behind the first flag looked over the tops of all 6 flags which lined up in a perfectly straight line. If the Earth were a ball 25,000 miles in circumference the flags should have progressively dipped down after the first establishing line of sight, the second would have descended 8 inches, 32 inches for the third, 6 feet for the fourth, 10 feet 8 inches for the fifth, and 16 feet 8 inches for the sixth.

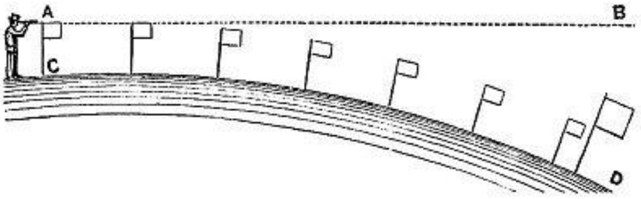

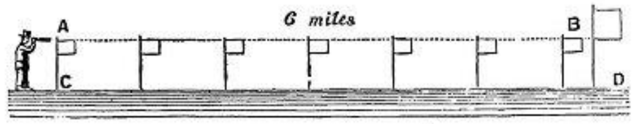

64) Quoting "Earth Not a Globe!" by Samuel Rowbotham, "*It is known that the horizon at sea, whatever distance it may extend to the right and left of the observer on land, always appears as a straight line. The following experiment has been tried in various parts of the country. At Brighton, on a rising ground near the race course, two poles were fixed in the earth six yards apart, and directly opposite the sea. Between these poles a line was tightly stretched parallel to the horizon. From the center of the line the view embraced not less than 20 miles on each side making a distance of 40 miles. A vessel was observed sailing directly westwards; the line cut the rigging a little above the bulwarks, which it did for several hours or until the vessel had sailed the whole distance of 40 miles. The ship coming into view from the east would have to ascend an inclined plane for 20 miles until it arrived at the center of the arc, whence it would have to descend for the same distance. The square of 20 miles multiplied by 8 inches gives 266 feet as the amount the vessel would be below the line at the beginning and at the end of the 40 miles.*"

65) Also Quoting Dr. Rowbotham, "*On the shore near Waterloo, a few miles to the north of Liverpool, a good telescope was fixed, at an elevation of 6 feet above the water. It was directed to a large steamer, just leaving the River Mersey, and sailing out to Dublin. Gradually the mast-head of the receding vessel came nearer to the horizon, until, at length, after more than four hours had elapsed, it disappeared. The ordinary rate of sailing of the Dublin steamers was fully eight miles an hour; so that the vessel would be, at least, thirty-two miles distant when the mast-head came to the horizon. The 6 feet of elevation of the telescope would require three miles to be deducted for convexity, which would leave twenty-nine miles, the square of which, multiplied by 8 inches, gives 560 feet; deducting 80 feet for the height of the main-mast, and we find that, according to the doctrine of rotundity, the mast-head of the outward bound steamer should have been 480 feet below the horizon. Many other experiments of this kind have been made upon sea-going steamers, and always with results entirely incompatible with the theory that the earth is a globe.*"

66) Dr. Rowbotham conducted several other experiments using telescopes, spirit levels, sextants and "theodolites," special precision instruments used for measuring angles in horizontal or vertical planes. By positioning them at equal heights aimed at each other successively he proved over and over the Earth to be perfectly flat for miles without a single inch of curvature. His findings caused quite a stir in the scientific community and thanks to 30 years of his efforts, the shape of the Earth became a hot topic of debate around the turn of the nineteenth century.

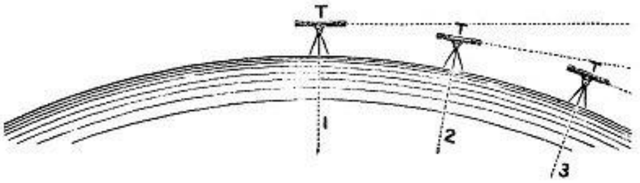

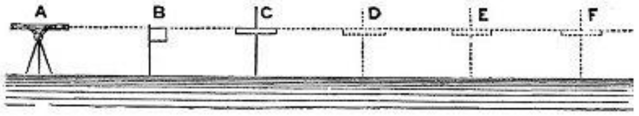

67) The distance across the Irish Sea from the Isle of Man's Douglas Harbor to Great Orm's Head in North Wales is 60 miles. If the Earth was a globe then the surface of the water between them would form a 60 mile arc, the center towering 1944 feet higher than the coastlines at either end. It is well-known and easily verifiable, however, that on a clear day, from a modest altitude of 100 feet, the Great Orm's Head is visible from Douglas Harbor. This would be completely impossible on a globe of 25,000 miles. Assuming the 100 foot altitude causes the horizon to appear approximately 13 miles off, the 47 miles remaining means the Welsh coastline should still fall an impossible 1472 feet below the line of sight!

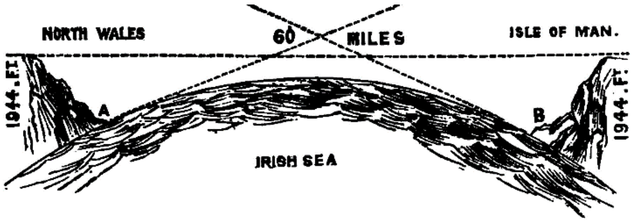

68) The Philadelphia skyline is clearly visible from Apple Pie Hill in the New Jersey Pine Barrens 40 miles away. If Earth were a ball 25,000 miles in circumference, factoring in the 205 foot elevation of Apple Pie Hill, the Philly skyline should remain well-hidden beyond 335 feet of curvature.

PHILLYSKYLINE.COM

69) The New York City skyline is clearly visible from Harriman State Park's Bear Mountain 60 miles away. If Earth were a ball 25,000 miles in circumference, viewing from Bear Mountain's 1,283 foot summit, the Pythagorean Theorem determining distance to the horizon being 1.23 times the square root of the height in feet, the NYC skyline should be invisible behind 170 feet of curved Earth.

70) From Washington's Rock in New Jersey, at just a 400 foot elevation, it is possible on a clear day to see the skylines of both New York and Philadelphia in opposite directions at the same time covering a total distance of 120 miles! If Earth were a ball 25,000 miles in circumference, both of these skylines should be hidden behind over 800 feet of Earth's curvature.

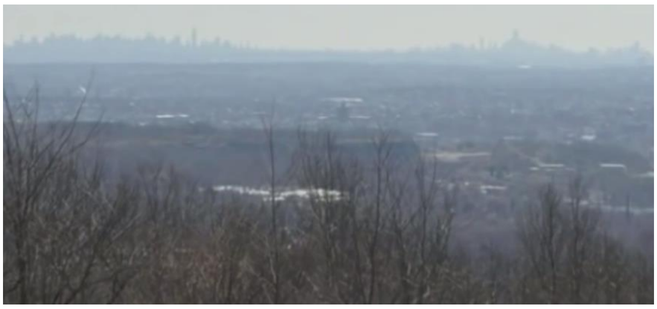

71) It is often possible to see the Chicago skyline from sea-level 60 miles away across Lake Michigan. In 2015 after photographer Joshua Nowicki photographed this phenomenon several news channels quickly claimed his picture to be a "superior mirage," an atmospheric anomaly caused by temperature inversion. While these certainly do occur, the skyline in question was facing right-side up and clearly seen unlike a hazy illusory mirage, and on a ball-Earth 25,000 miles in circumference should be 2,400 feet below the horizon.

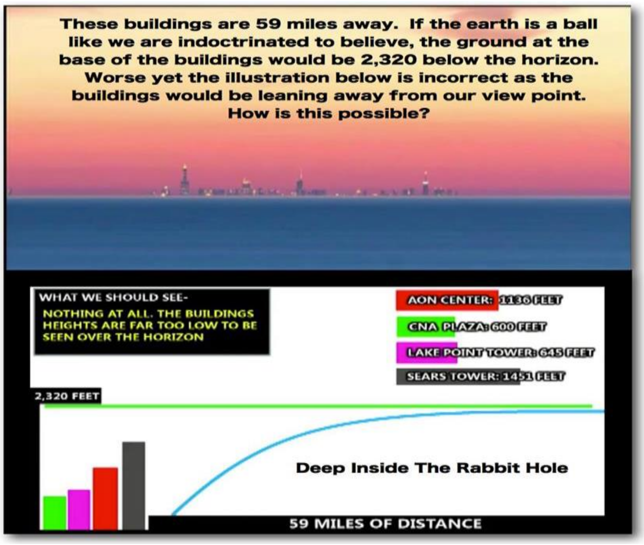

These buildings are 59 miles away. If the earth is a ball like we are indoctrinated to believe, the ground at the base of the buildings would be 2,320 below the horizon. Worse yet the illustration below is incorrect as the buildings would be leaning away from our view point. How is this possible?

WHAT WE SHOULD SEE-
NOTHING AT ALL. THE BUILDINGS HEIGHTS ARE FAR TOO LOW TO BE SEEN OVER THE HORIZON

72) October 16, 1854 the Times newspaper reported the Queen's visit to Great Grimsby from Hull recording they were able to see the 300 foot tall dock tower from 70 miles away. On a ball-Earth 25,000 miles in circumference, factoring their 10 foot elevation above the water and the tower's 300 foot height, at 70 miles away the dock tower should have remained an entire 2,600 feet below the horizon.

73) In 1872 Capt. Gibson and crewmates, sailing the ship "Thomas Wood" from China to London, reported seeing the entirety of St. Helena Island on a clear day from 75 miles away. Factoring in their height during measurement on a ball-Earth 25,000 miles in circumference, it was found the island should have been 3,650 feet below their line of sight.

74) From Genoa, Italy at a height of just 70 feet above sea-level, the island of Gorgona can often be seen 81 miles away. If Earth were a ball 25,000 miles in circumference, Gorgona should be hidden beyond 3,332 feet of curvature.

Isola della Gorgona
Isola di Capraia

75) From Genoa, Italy at a height of just 70 feet above sea-level, the island of Corsica can often be seen 99 miles away. If Earth were a ball 25,000 miles in circumference, Corsica should fall 5,245 feet, almost an entire mile below the horizon.

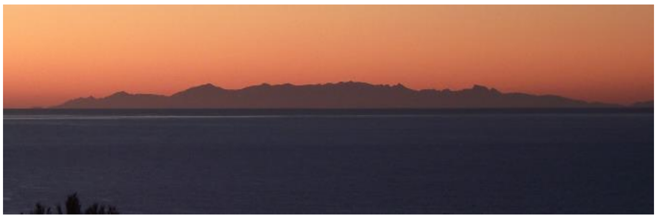

76) From Genoa, Italy 70 feet above sea-level, the island of Capraia 102 miles away can often be seen as well. If Earth were a ball 25,000 miles in circumference, Capraia should always remain hidden behind 5,605 feet, over a mile of supposed curvature.

Isola d'Elba
Gorgona
Capraia
Corsica
Corsica, Elba e Arc. Toscano visti da Genova

77) Also from Genoa, on bright clear days, the island of Elba can be seen an incredible 125 miles away! If Earth were a ball 25,000 miles in circumference, Elba should be forever invisible behind 8770 feet of curvature.

Isola d'Elba
Monte Capanne (m. 1019)

78) From Anchorage, Alaska at an elevation of 102 feet, on clear days Mount Foraker can be seen with the naked eye 120 miles away. If Earth were a ball 25,000 miles in circumference, Mount Foraker's 17,400 summit should be leaning back away from the observer covered by 7,719 feet of curved Earth. In reality, however, the entire mountain can be quite easily seen standing straight from base to summit.

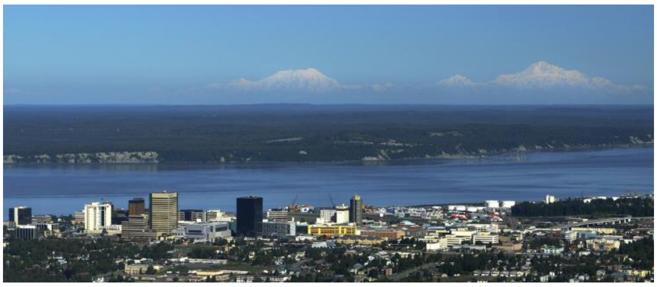

79) From Anchorage, Alaska at an elevation of 102 feet, on clear days Mount McKinley can be seen with the naked eye from 130 miles away. If Earth were a ball 25,000 miles in circumference, Mount McKinley's 20,320 foot summit should be leaning back away from the observer and almost half covered by 9,220 feet of curved Earth. In reality, however, the entire mountain can be quite easily seen standing straight from base to summit.

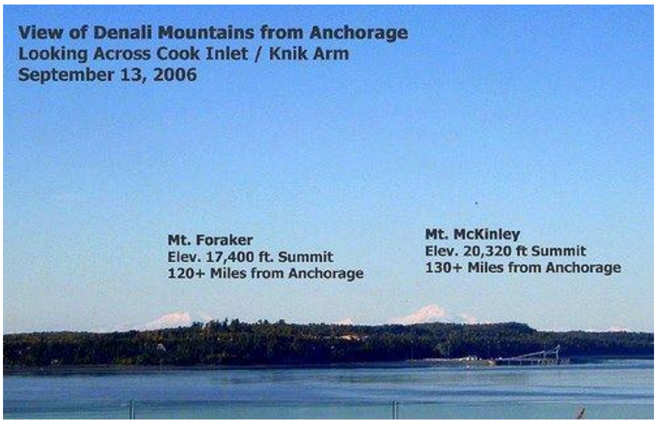
View of Denali Mountains from Anchorage
Looking Across Cook Inlet / Knik Arm
September 13, 2006
Mt. Foraker
Elev. 17,400 ft. Summit
120+ Miles from Anchorage
Mt. McKinley
Elev. 20,320 ft Summit
130+ Miles from Anchorage

80) In Chambers' Journal, February 1895, a sailor near Mauritius in the Indian Ocean reported having seen a vessel which turned out to be an incredible 200 miles away! The incident caused much heated debate in nautical circles at the time, gaining further confirmation in Aden, Yemen where another witness reported seeing a missing Bombay steamer from 200 miles away. He correctly stated the precise appearance, location and direction of the steamer all later corroborated and confirmed correct by those onboard. Such sightings are absolutely inexplicable if the Earth were actually a ball 25,000 miles around, as ships 200 miles distant would have to fall approximately 5 miles below line of sight!

81) The distance from which various lighthouse lights around the world are visible at sea far exceeds what could be found on a ball-Earth 25,000 miles in circumference. For example, the Dunkerque Light in southern France at an altitude of 194 feet is visible from a boat (10 feet above sea-level) 28 miles away. Spherical trigonometry dictates that if the Earth was a globe with the given curvature of 8 inches per mile squared, this light should be hidden 190 feet below the horizon.

82) The Port Nicholson Light in New Zealand is 420 feet above sea-level and visible from 35 miles away where it should be 220 feet below the horizon.

83) The Egerö Light in Norway is 154 feet above high-water and visible from 28 statute miles where it should be 230 feet below the horizon.

84) The Light at Madras, on the Esplanade, is 132 feet high and visible from 28 miles away, where it should be 250 feet below the line of sight.

85) The Cordonan Light on the west coast of France is 207 feet high and visible from 31 miles away, where it should be 280 feet below the line of sight.

86) The light at Cape Bonavista, Newfoundland is 150 feet above sea-level and visible at 35 miles, where it should be 491 feet below the horizon.

87) The lighthouse steeple of St. Botolph's Parish Church in Boston is 290 feet tall and visible from over 40 miles away, where it should be hidden a full 800 feet below the horizon!

88) The Isle of Wight lighthouse in England is 180 feet high and can be seen up to 42 miles away, a distance at which modern astronomers say the light should fall 996 feet below line of sight.

89) The Cape L'Agulhas lighthouse in South Africa is 33 feet high, 238 feet above sea level, and can be seen for over 50 miles. If the world were a globe, this light would fall 1,400 feet below an observer's line of sight.

90) The Statue of Liberty in New York stands 326 feet above sea level and on a clear day can be seen as far as 60 miles away. If the Earth were a globe, that would put Lady Liberty at an impossible 2,074 feet below the horizon.

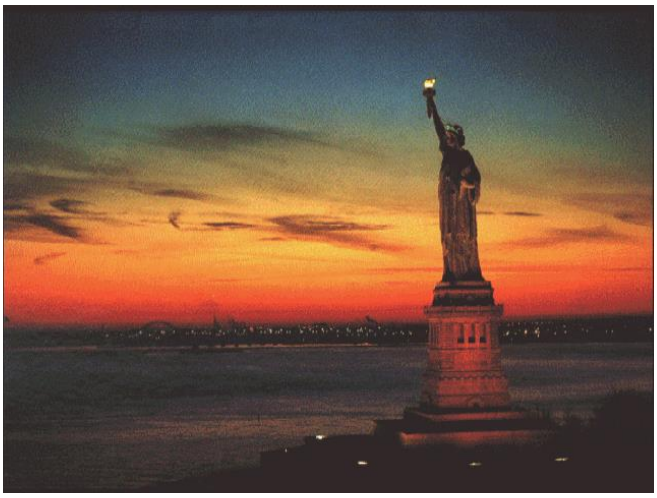

91) The lighthouse at Port Said, Egypt, at an elevation of only 60 feet has been seen an astonishing 58 miles away, where, according to modern astronomy it should be 2,182 feet below the line of sight!

92) The Notre Dame Antwerp spire stands 403 feet high from the foot of the tower with Strasburg measuring 468 feet above sea level. With the aid of a telescope, ships can be distinguished on the horizon and captains declare they can see the cathedral spire from an amazing 150 miles away. If the Earth were a globe, however, at that distance the spire should be an entire mile, 5,280 feet below the horizon!

93) The St. George's Channel between Holyhead and Kingstown Harbor near Dublin is 60 miles across. When half-way across a ferry passenger will notice behind them the light on Holyhead pier as well as in front of them the Poolbeg light in Dublin Bay. The Holyhead Pier light is 44 feet high, while the Poolbeg lighthouse 68 feet, therefore a vessel in the middle of the channel, 30 miles from either side standing on a deck 24 feet above the water, can clearly see both lights. On a ball Earth 25,000 miles in circumference, however, both lights should be hidden well below both horizons by over 300 feet!

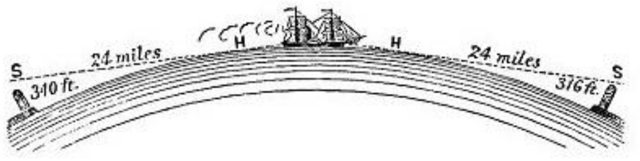

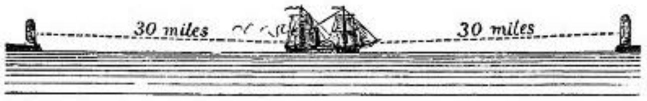

94) From the highland near Portsmouth Harbor in Hampshire, England looking across Spithead to the Isle of Wight, the entire base of the island, where water and land come together composes a perfectly straight line 22 statute miles long. According to the ball-Earth theory, the Isle of Wight should decline 80 feet from the center on each side to account for the necessary curvature. The cross-hairs of a good theodolite directed there, however, have repeatedly shown the land and water line to be perfectly level.

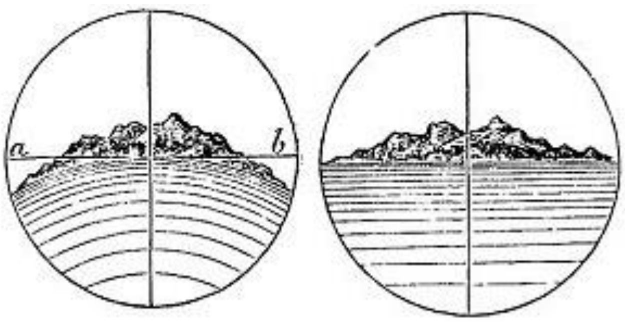

95) On a clear day from the highland near Douglas Harbor on the Isle of Man, the whole length of the coast of North Wales is often plainly visible to the naked eye. From the Point of Ayr at the mouth of the River Dee to Holyhead comprises a 50 mile stretch which has also been repeatedly found to be perfectly horizontal. If the Earth actually had curvature of 8 inches per mile squared, as NASA and modern astronomy claim, the 50 mile length of Welsh coast seen along the horizon in Liverpool Bay would have to decline from the center-point an easily detectable 416 feet on each side!

96) From “100 Proofs the Earth is Not a Globe” by William Carpenter, “*If we take a journey down the Chesapeake Bay, by night, we shall see the ‘light’ exhibited at Sharpe's Island for an hour before the steamer gets to it. We may take up a position on the deck so that the rail of the vessel's side will be in a line with the ‘light’ and in the line of sight; and we shall find that in the whole journey the light won't vary in the slightest degree in its apparent elevation. But, say that a distance of thirteen miles has been traversed, the astronomers' theory of ‘curvature’ demands a difference (one way or the other!) in the apparent elevation of the light, of 112 feet 8 inches! Since, however, there is not a difference of 100 hair's breadths, we have a plain proof that the water of the Chesapeake Bay is not curved, which is a proof that the Earth is not a globe.*”

97) NASA and modern astronomy say the Earth is a giant ball tilted back, wobbling and spinning 1,000 mph around its central axis, traveling 67,000 mph circles around the Sun, spiraling 500,000 mph around the Milky Way, while the entire galaxy rockets a ridiculous 670,000,000 mph through the Universe, with all of these motions originating from an alleged “Big Bang” cosmogenic explosion 14 billion years ago. That’s a grand total of 670,568,000 mph in several different directions we’re all supposedly speeding along at simultaneously, yet no one has ever seen, felt, heard, measured or proven a single one of these motions to exist whatsoever.

98) NASA and modern astronomy say Polaris, the North Pole star, is somewhere between 323-434 light years, or about 2 quadrillion miles, away from us! Firstly, note that is between 1,938,000,000,000,000 - 2,604,000,000,000,000 miles making a difference of 666,000,000,000,000 (over six hundred trillion) miles! If modern astronomy cannot even agree on the distance to stars within **hundreds of trillions** of miles, perhaps their “science” is flawed and their theory needs re-examining. However, even granting them their obscurely distant stars, it is impossible for heliocentrists to explain how Polaris manages to always remain perfectly aligned straight above the North Pole throughout Earth’s various alleged tilting, wobbling, rotating and revolving motions.

99) Viewed from a ball-Earth, Polaris, situated directly over the North Pole, should not be visible anywhere in the Southern hemisphere. For Polaris to be seen from the Southern hemisphere of a globular Earth, the observer would have to be somehow looking “through the globe,” and miles of land and sea would have to be transparent. Polaris can be seen, however, up to over 20 degrees South latitude.

100) If Earth were a ball, the Southern Cross and other Southern constellations would all be visible at the same time from every longitude on the same latitude as is the case in the North with Polaris and its surrounding constellations. Ursa Major/Minor and many others can be seen from every Northern meridian simultaneously whereas in the South, constellations like the Southern Cross cannot. This proves the Southern hemisphere is not "turned under" as in the ball-Earth model, but simply stretching further outwards away from the Northern center-point as in the flat Earth model.

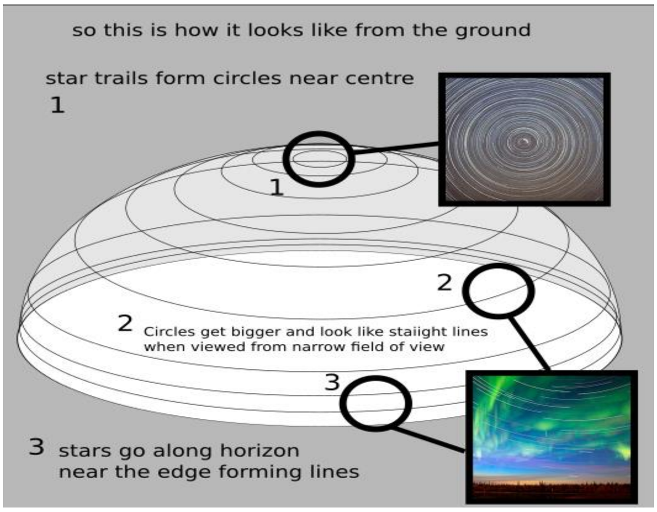
so this is how it looks like from the ground
star trails form circles near centre
1
1
2
2 Circles get bigger and look like staiight lines when viewed from narrow field of view
3
3 stars go along horizon near the edge forming lines

101) Sigma Octantis is claimed to be a Southern central pole star similar to Polaris, around which the Southern hemisphere stars all rotate around the opposite direction. Unlike Polaris, however, Sigma Octantis can NOT be seen simultaneously from every point along the same latitude, it is NOT central but allegedly 1 degree off-center, it is NOT motionless, and in fact cannot be seen at all using publicly available telescopes! There is legitimate speculation regarding whether Sigma Octantis even exists. Either way, the direction in which stars move overhead is based on perspective and the exact direction you're facing, not which hemisphere you are in.

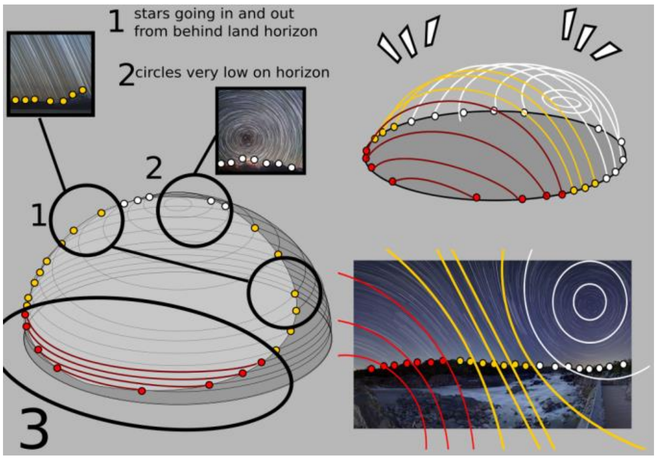
1 stars going in and out from behind land horizon
2 circles very low on horizon
2
1
3

102) Some heliocentrists have tried to suggest that the Pole Star's gradual declination overhead as an observer travels southwards is proof of a globular Earth. Far from it, the declination of the Pole Star or any other object is simply a result of the Law of Perspective on plane (flat) surfaces. The Law of Perspective dictates that the angle and height at which an object is seen diminishes the farther one recedes from the object, until at a certain point the line of sight and the seemingly uprising surface of the Earth converges to a vanishing point (i.e. the horizon line) beyond which the object is invisible. In the ball-Earth model the horizon is claimed to be the curvature of the Earth, whereas in reality, the horizon is known to be simply the vanishing line of perspective based on the strength of your eyes, instruments, weather and altitude.

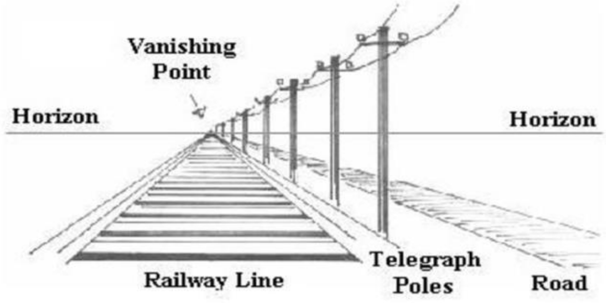
Vanishing Point
Horizon
Horizon
Railway Line
Telegraph Poles
Road

103) There are several constellations which can be seen from far greater distances over the face of the Earth than should be possible if the world were a rotating, revolving, wobbling ball. For instance, Ursa Major, very close to Polaris, can be seen from 90 degrees North latitude (the North Pole) all the way down to 30 degrees South latitude. For this to be possible on a ball-Earth the Southern observers would have to be seeing through hundreds or thousands of miles of bulging Earth to the Northern sky.

104) The constellation Vulpecula can be seen from 90 degrees North latitude, all the way to 55 degrees South latitude. Taurus, Pisces and Leo can be seen from 90 degrees North all the way to 65 degrees South. An observer on a ball-Earth, regardless of any tilt or inclination, should not logically be able to see this far.

105) Aquarius and Libra can be seen from 65 degrees North to 90 degrees South! The constellation Virgo is visible from 80 degrees North down to 80 degrees South, and Orion can be seen from 85 degrees North all the way to 75 degrees South latitude! These are all only possible because the "hemispheres" are not spheres at all but concentric circles of latitude extending outwards from the central North Pole with the stars rotating over and around.

106) The so-called "South Pole" is simply an arbitrary point along the Antarctic ice marked with a red and white barbershop pole topped with a metal ball-Earth. This ceremonial South Pole is admittedly and provably NOT the actual South Pole, however, because the actual South Pole could be demonstrably confirmed with the aid of a compass showing North to be 360 degrees around the observer. Since this feat has never been achieved, the model remains pure theory, along with the establishment's excuse that the geomagnetic poles supposedly constantly move around making verification of their claims impossible.

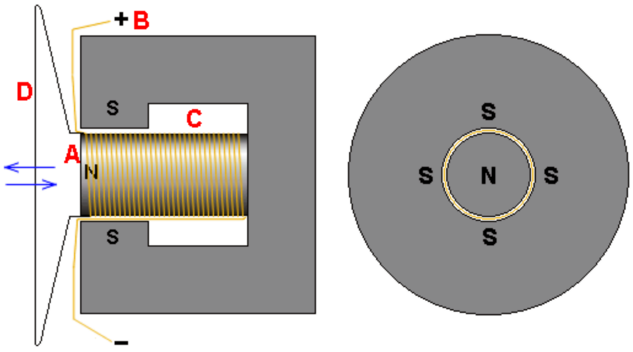

107) Ring magnets of the kind found in loudspeakers have a central North pole with the opposite "South" pole actually being all points along the outer circumference. This perfectly demonstrates the magnetism of our flat Earth, whereas the alleged source of magnetism in the ball-Earth model is emitted from a hypothetical molten magnetic core in the center of the ball which they claim conveniently causes both poles to constantly move thus evading independent verification at their two "ceremonial poles." In reality the deepest drilling operation in history, the Russian Kola Ultradeep, managed to get only 8 miles down, so the entire ball-Earth model taught in schools showing a crust, outer-mantle, inner-mantle, outer-core and inner-core layers are all purely speculation as we have never penetrated through beyond the crust.

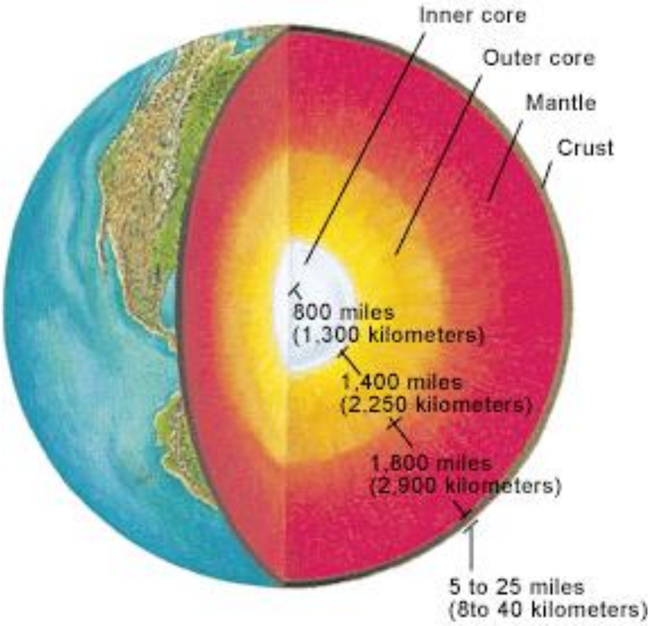

108) The mariner's compass is an impossible and non-sensical instrument for use on a ball-Earth. It simultaneously points North and South over a flat surface, yet claims to be pin-pointing two constantly moving geomagnetic poles at opposite ends of a spinning sphere originating from a hypothetical molten metal core. If compass needles were actually drawn to the North pole of a globe, the opposing "South" needle would actually be pointing up and off into outer-space.

109) There are no fixed "East" or "West" points just as there is no fixed "South." The North central Pole is the only proven fixed point on our flat Earth, with South being all straight lines outwards from the pole, East and West being concentric circles at constant right angles 90 degrees from the pole. A westerly circumnavigation of Earth is thus going around with Polaris continually on your right, while an easterly circumnavigation is going around with Polaris always at your left.

110) Magellan and others' East/West circumnavigations of Earth are often quoted as proof of the ball model. In actual fact, however, sailing or flying at rights angles to the North pole and eventually returning to one's original location is no more difficult or mysterious than doing so on a globe. Just as an architect's compass can place its center-point on a flat piece of paper and trace a circle either way around the "pole," so can a ship or plane circumnavigate a flat-Earth.

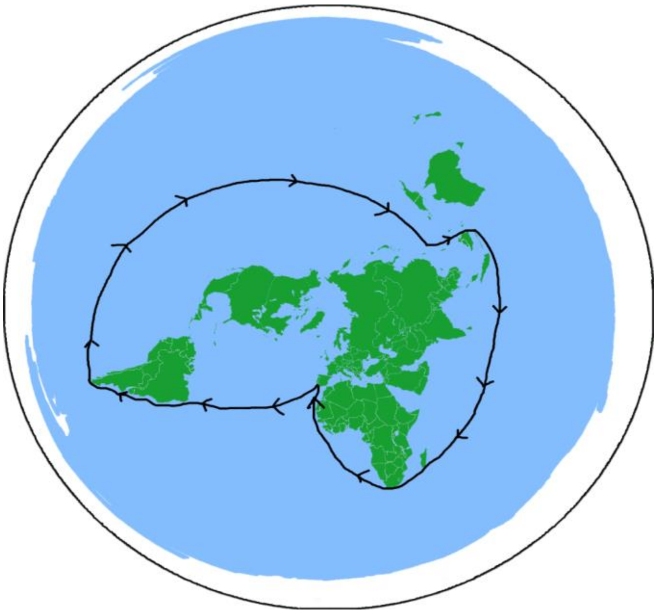

111) Since the North Pole and Antarctica are covered in ice and guarded "no-fly" zones, no ships or planes have ever been known to circumnavigate the Earth in North/South directions. The only kind of circumnavigation which could not happen on a flat-Earth is North/Southbound, which is likely the very reason for the heavily-enforced flight restrictions. The fact that there has yet to be a single verified North/South circumnavigation of Earth serves as standing proof the world is not a ball.

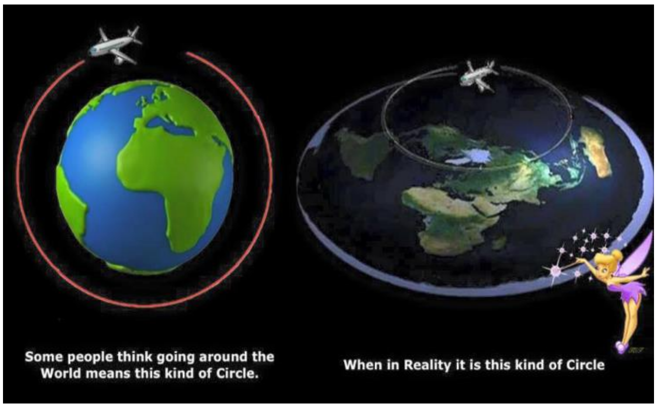
Some people think going around the World means this kind of Circle.
When in Reality it is this kind of Circle

112) The Sun brings noon to every time-zone as it passes directly over-head every 15 degree demarcation point, 24 times per day in its circular path over and around the Earth. If time-zones were instead caused by the uniform spinning of the ball-Earth around the Sun, every 6 months as Earth found itself on the opposite side of the Sun, clocks all over Earth would have to flip 12 hours, day would be night and night would be day.

113) The idea that people are standing, ships are sailing and planes are flying upside down on certain parts of Earth while others tilted at 90 degrees and all other impossible angles is complete absurdity. The idea that a man digging a hole straight down could eventually reach sky on the other side is ludicrous. Common sense tells every free-thinking person correctly that there truly is an "up" and "down" in nature, unlike the "everything is relative" rhetoric of the Newtonian/Einsteinian paradigm.

BECAUSE GRAVITY IS MAGIK

114) Quoting, "On the False Wisdom of the Philosophers" by Lacantius, "*A sphere where people on the other side live with their feet above their heads, where rain, snow and hail fall upwards, where trees and crops grow upside-down and the sky is lower than the ground? The ancient wonder of the hanging gardens of Babylon dwindle into nothing in comparison to the fields, seas, towns and mountains that pagan philosophers believe to be hanging from the earth without support!*"

Meanwhile in Paris

115) The existing laws of density and buoyancy perfectly explained the physics of falling objects long before knighted Freemason "Sir" Isaac Newton bestowed his theory of "gravity" upon the world. It is a fact that objects placed in denser mediums rise up while objects placed in less dense mediums sink down. To fit with the heliocentric model which has no up or down, Newton instead claimed objects are attracted to large masses and fall towards the center. Not a single experiment in history, however, has shown an object massive enough to, by virtue of its mass alone, cause other smaller masses to be attracted to it as Newton claims "gravity" does with Earth, the Sun, Moon, Stars and Planets.

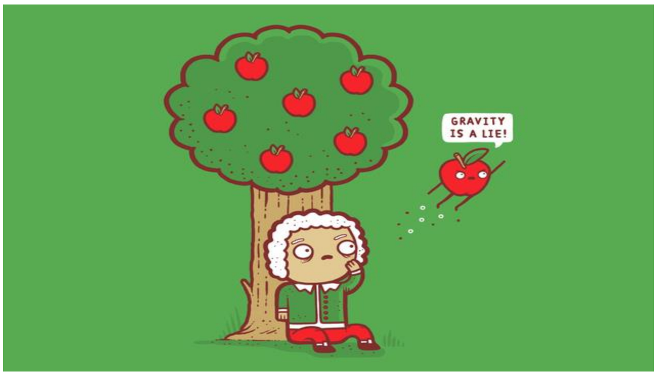
GRAVITY IS A LIE!

116) There has also never been a single experiment in history showing an object massive enough to, by virtue of its mass alone, cause another smaller mass to orbit around it. The magic theory of gravity allows for oceans, buildings and people to remain forever stuck to the underside of a spinning ball while simultaneously causing objects like the Moon and satellites to remain locked in perpetual circular orbits around the Earth. If these were both true then people should be able to jump up and start orbiting circles around the Earth, or the Moon should have long ago been sucked into the Earth. Neither of these theories have ever been experimentally verified and their alleged results are mutually exclusive.

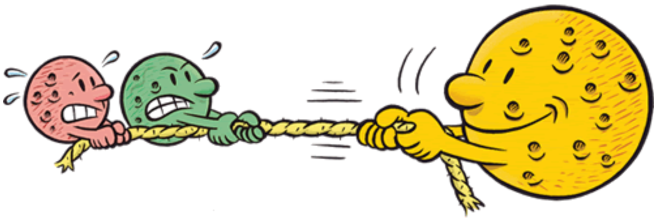

117) Newton also theorized and it is now commonly taught that the Earth's ocean tides are caused by gravitational lunar attraction. If the Moon is only 2,160 miles in diameter and the Earth 8,000 miles, however, using their own math and "law," it follows that the Earth is 87 times more massive and therefore the larger object should attract the smaller to it, and not the other way around. If the Earth's greater gravity is what keeps the Moon in orbit, it is impossible for the Moon's lesser gravity to supersede the Earth's gravity, especially at Earth's sea-level, where its gravitational attraction would even further out-trump the Moon's. And if the Moon's gravity truly did supersede the Earth's causing the tides to be drawn towards it, there should be nothing to stop them from continuing onwards and upwards towards their great attractor.

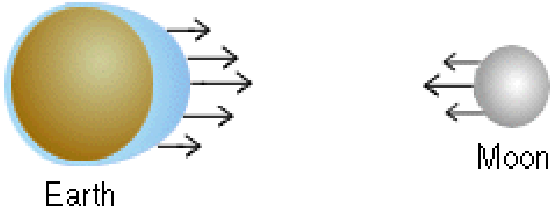

118) Furthermore, the velocity and path of the Moon are uniform and should therefore exert a uniform influence on the Earth's tides, when in actuality the Earth's tides vary greatly and do not follow the Moon. Earth's lakes, ponds, marshes and other inland bodies of water also inexplicably remain forever outside the Moon's gravitational grasp! If "gravity" was truly drawing Earth's oceans up to it, all lakes, ponds and other bodies of standing water should certainly have tides as well.

119) It is claimed that the other planets are spheres and so therefore Earth must also be a sphere. Firstly, Earth is a "plane" not a "planet," so the shape of these "planets" in the sky have no bearing on the shape of the Earth beneath our feet. Secondly, these "planets" have been known for thousands of years around the world as the "wandering stars" since they differ from the other fixed stars in their relative motions only. When looked at with an unprejudiced naked-eye or through a telescope, the fixed and wandering stars appear as luminous discs of light, NOT spherical terra firma. The pictures and videos shown by NASA of spherical terra firma planets are all clearly fake computer-generated images, and NOT photographs.

Mars through NASA's telescope
Mars through everyone else's telescope

120) The etymology of the word "planet" actually comes from late Old English planete, from Old French planete (Modern French planète), from Latin planeta, from Greek planetes, from (asteres) planetai "wandering (stars)," from planasthai "to wander," of unknown origin, possibly from PIE *pele "flat, to spread" or notion of "spread out." And Plane (n) "flat surface," c. 1600, from Latin planum "flat surface, plane, level, plain," planus "flat, level, even, plain, clear." They just added a "t" to our Earth plane and everyone bought it.

121) When you observe the Sun and Moon you see two equally-sized equidistant circles tracing similar paths at similar speeds around a flat, stationary Earth. The "experts" at NASA, however, claim your common sense every day experience is false on all counts! To begin with, they say the Earth is not flat but a big ball; not stationary but spinning around 19 miles per second; they say the Sun does not revolve around the Earth as it appears, but Earth revolves around the Sun; the Moon, on the other hand, does revolve around the Earth, though not East to West as it appears, rather West to East; and the Sun is actually 400 times larger than the Moon and 400 times farther away! You can clearly see they are the same size and distance, you can see the Earth is flat, you can feel the Earth is stationary, but according to the gospel of modern astronomy, you are wrong and a simpleton worthy of endless ridicule if you dare to trust your own eyes and experience.

Moon
Sun
Appear Almost Same Size

122) Quoting Allen Daves, "*If the Government or NASA had said to you that the Earth is stationary, imagine that. And then imagine we are trying to convince people that 'no, no it's not stationary, it's moving forward at 32 times rifle bullet speed and spinning at 1,000 miles per hour.' We would be laughed at! We would have so many people telling us 'you are crazy, the Earth is not moving!' We would be ridiculed for having no scientific backing for this convoluted moving Earth theory. And not only that but then people would say, 'oh then how do you explain a fixed, calm atmosphere and the Sun's observable movement, how do you explain that?' Imagine saying to people, 'no, no, the atmosphere is moving also but is somehow magically velcroed to the moving-Earth. The reason is not simply because the Earth is stationary.' So what we are actually doing is what makes sense. We are saying that the moving-Earth theory is nonsense. The stationary-Earth theory makes sense and we are being ridiculed. You've got to picture it being the other way around to realize just how RIDICULOUS this situation is. This theory from the*

Government and NASA that the Earth is rotating and orbiting and leaning over and wobbling is absolute nonsense and yet people are clinging to it, tightly, like a teddy bear. They just can't bring themselves to face the possibility that the Earth is stationary though ALL the evidence shows it: we feel no movement, the atmosphere hasn't been blown away, we see the Sun move from East-to-West, everything can be explained by a motionless Earth without bringing in all these assumptions to cover up previous assumptions gone bad."

123) Heliocentrists' astronomical figures always sound perfectly precise, but they have historically been notorious for regularly and drastically changing them to suit their various models. For instance, in his time Copernicus calculated the Sun's distance from Earth to be 3,391,200 miles. The next century Johannes Kepler decided it was actually 12,376,800 miles away. Issac Newton once said, "*It matters not whether we reckon it 28 or 54 million miles distant for either would do just as well!*" How scientific!? Benjamin Martin calculated between 81 and 82 million miles, Thomas Dilworth claimed 93,726,900 miles, John Hind stated positively 95,298,260 miles, Benjamin Gould said more than 96 million miles, and Christian Mayer thought it was more than 104 million! Flat-Earthers throughout the ages, conversely, have used sextants and plane trigonometry to make such calculations and found the Sun and Moon both to be only about 32 miles in diameter and less than a few thousand miles from Earth.

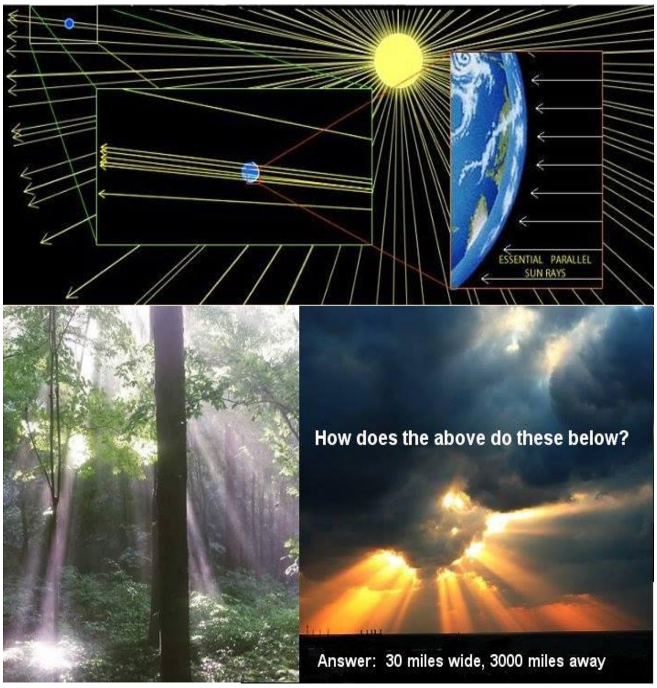
ESSENTIAL PARALLEL SUN RAYS
How does the above do these below?
Answer: 30 miles wide, 3000 miles away

124) Amateur balloon footage taken above the clouds has provided stunning visual proof that the Sun cannot be millions of miles away. In several shots you can see a clear hot-spot reflecting on the clouds directly below the Sun's spotlight-like influence. If the Sun were actually millions of miles away such a small, localized hot-spot could not occur.

The Sun is clearly NOT 93 million miles away. You can see a hot-spot directly underneath it, proving it is very close.

125) Another proof the Sun is not millions of miles away is found by tracing the angle of sun-rays back to their source above the clouds. There are thousands of pictures showing how sunlight comes down through cloud-cover at a variance of converging angles. The area of convergence is of course the Sun, and is clearly NOT millions of miles away, but rather relatively close to Earth just above the clouds.

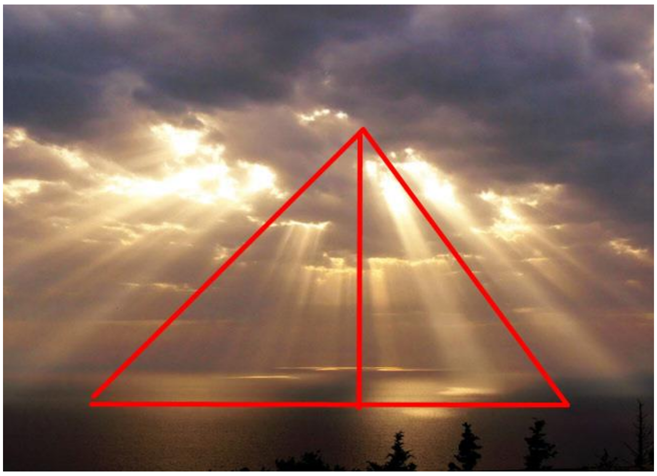

126) The Sun's annual journey from tropic to tropic, solstice to solstice, is what determines the length and character of days, nights and seasons. This is why equatorial regions experience almost year-round summer and heat while higher latitudes North and especially South experience more distinct seasons with harsh winters. The heliocentric model claims seasons change based on the ball-Earth's alleged "axial tilt" and "elliptical orbit" around the Sun, yet their flawed current model places us closest to the Sun (91,400,000 miles) in January when its actually winter, and farthest from the Sun (94,500,000 miles) in July when its actually summer throughout most of the Earth.

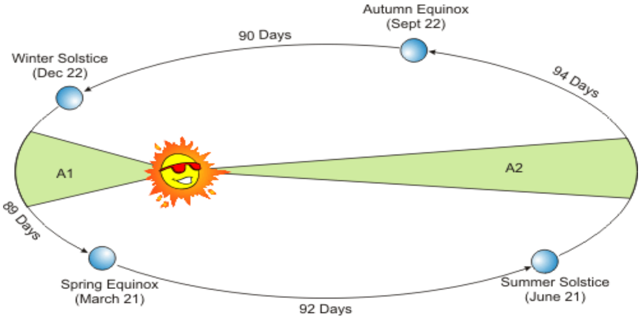

127) The fact that the Sun and Moon’s reflections on water always form a straight line path from the horizon to the observer proves the Earth is not a ball. If Earth’s surface was curved it would be impossible for the reflected light to curve over the ball from horizon to observer.

128) There are huge centuries-old stone sundials and moondials all over the world which still tell the time now down to the minute as perfectly as the day they were made. If the Earth, Sun and Moon were truly subject to the number of contradictory revolving, rotating, wobbling and spiraling motions claimed by modern astronomy, it would be impossible for these monuments to so accurately tell time without constant adjustment.

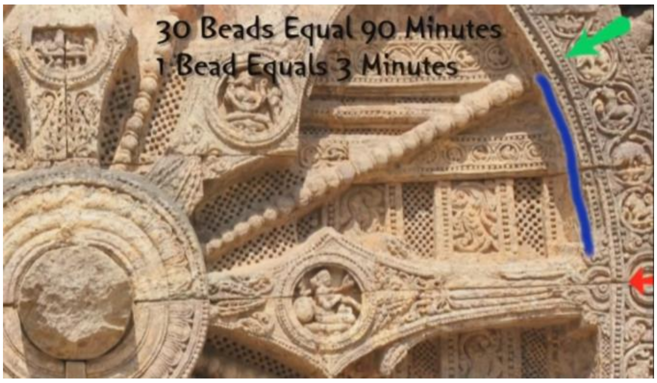

129) To quote William Carpenter, “*Why, in the name of common sense, should observers have to fix their telescopes on solid stone bases so that they should not move a hair's-breadth, - if the Earth on which they fix them moves at the rate of nineteen miles in a second? Indeed, to believe that ‘six thousand million million million tons’ is ‘rolling, surging, flying, darting on through space for ever’ with a velocity compared with which a shot from a cannon is a ‘very slow coach,’ with such unerring accuracy that a telescope fixed on granite pillars in an observatory will not enable a lynx-eyed astronomer to detect a variation in its onward motion of the thousandth part of a hair's-breadth is to conceive a miracle compared with which all the miracles on record put together would sink into utter insignificance. Since we can, (in middle north latitudes), see the North Star, on looking out of a window that faces it - and out of the very same corner of the very same pane of glass in the very same window - all the year round, it is proof enough for any man in his senses that we have made no motion at all and that the Earth is not a globe*.”

130) From “Earth Not a Globe!” by Samuel Rowbotham, *“Take two carefully-bored metallic tubes, not less than six feet in length, and place them one yard asunder, on the opposite sides of a wooden frame, or a solid block of wood or stone: so adjust them that their centres or axes of vision shall be perfectly parallel to each other. Now, direct them to the plane of some notable fixed star, a few seconds previous to its meridian time. Let an observer be stationed at each tube and the moment the star appears in the first tube let a loud knock or other signal be given, to be repeated by the observer at the second tube when he first sees the same star. A distinct period of time will elapse between the signals given. The signals will follow each other in very rapid succession, but still, the time between is sufficient to show that the same star is not visible at the same moment by two parallel lines of sight when only one yard asunder. A slight inclination of the second tube towards the first tube would be required for the star to be seen through both tubes at the same instant. Let the tubes remain in their position for six months; at the end of which time the same observation or experiment will produce the same results--the star will be visible at the same meridian time, without the slightest alteration being required in the direction of the tubes: from which it is concluded that if the earth had moved one single yard in an orbit through space, there would at least be observed the slight inclination of the tube which the difference in position of one yard had previously required. But as no such difference in the direction of the tube is required, **the conclusion is unavoidable, that in six months a given meridian upon the earth's surface does not move a single yard, and therefore, that the earth has not the slightest degree of orbital motion**."*

131) NASA and modern astronomy maintain that the Moon is a solid, spherical, Earth-like habitation which man has actually flown to and set foot on. They claim the Moon is a non-luminescent planetoid which receives and reflects all its light from the Sun. The reality is, however, that the Moon is observably not a solid body, it is clearly circular, but not spherical, and not in any way an Earth-like planetoid which humans could set foot on. In fact, the Moon has been proven largely transparent and completely self-luminescent, shining with its own unique light.

132) The Sun's light is golden, warm, drying, preservative and antiseptic, while the Moon's light is silver, cool, damp, putrefying and septic. The Sun's rays decrease the combustion of a bonfire, while the Moon's rays increase combustion. Plant and animal substances exposed to sunlight quickly dry, shrink, coagulate, and lose the tendency to decompose and putrify; grapes and other fruits become solid, partially candied and preserved like raisins, dates, and prunes; animal flesh coagulates, loses its volatile gaseous constituents, becomes firm, dry, and slow to decay. When exposed to moonlight, however, plant and animal substances tend to show symptoms of putrefaction and decay. This proves that Sun and Moon light are different, unique, and opposites as they are in the geocentric flat model.

133) In direct sunlight a thermometer will read higher than another thermometer placed in the shade, but in full, direct moonlight a thermometer will read lower than another placed in the shade. If the Sun's light is collected in a large lens and thrown to a focus point it can create significant heat, while the Moon's light collected similarly creates no heat. In the "Lancet Medical Journal," from March 14th, 1856, particulars are given of several experiments which proved the Moon's rays when concentrated can actually reduce the temperature upon a thermometer more than eight degrees. So sunlight and moonlight clearly have altogether different properties.

134) Furthermore the Moon itself cannot physically be both a spherical body and a reflector of the Sun's light. Reflectors must be flat or concave for light rays to have any angle of incidence; If a reflector's surface is convex then every ray of light points in a direct line with the radius perpendicular to the surface resulting in no reflection.

135) Not only is the Moon clearly self-luminescent, shining its own unique light, but it is also largely transparent. When the waxing or waning Moon is visible during the day it is possible to see the blue sky right through the Moon. And on a clear night, during a waxing or waning cycle, it is even possible to occasionally see stars and "planets" directly through the surface of the Moon! The Royal Astronomical Society has on record many such occurrences throughout history which all defy the heliocentric model.

136) Many people think that modern astronomy's ability to accurately predict lunar and solar eclipses is a result and proof positive of the heliocentric theory of the universe. The fact of the matter however is that eclipses have been accurately predicted by cultures worldwide for thousands of years before the "heliocentric ball-Earth" was even a glimmer in Copernicus' imagination. Ptolemy in the 1st century A.D. accurately predicted eclipses for six hundred years on the basis of a flat, stationary Earth with equal precision as anyone living today. All the way back in 600 B.C. Thales accurately predicted an eclipse which ended the war between the Medes and Lydians. Eclipses happen regularly with precision in 18 year cycles, so regardless of geocentric or heliocentric, flat or globe Earth cosmologies, eclipses can be accurately calculated independent of such factors.

137) Another assumption and supposed proof of Earth's shape, heliocentrists claim that lunar eclipses are caused by the shadow of the ball-Earth occulting the Moon. They claim the Sun, Earth, and Moon spheres perfectly align like three billiard balls in a row so that the Sun's light casts the Earth's shadow onto the Moon. Unfortunately for heliocentrists, this explanation is rendered completely invalid due to the fact that lunar eclipses have happened and continue to happen regularly when both the Sun and Moon are still visible together above the horizon! For the Sun's light to be casting Earth's shadow onto the Moon, the three bodies must be aligned in a straight 180 degree syzygy, but as early as the time of Pliny, there are records of lunar eclipses happening while both the Sun and Moon are visible in the sky. Therefore the eclipsor of the Moon cannot be the Earth/Earth's shadow and some other explanation must be sought.

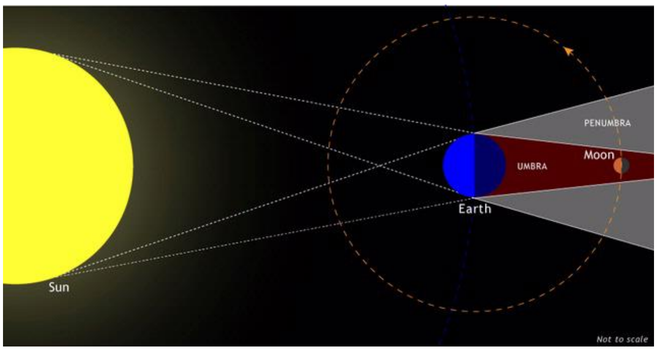

138) Another favorite "proof" of ball-Earthers is the appearance from an observer on shore of ships' hulls being obfuscated by the water and disappearing from view when sailing away towards the horizon. Their claim is that ships' hulls disappear before their mast-heads because the ship is beginning its declination around the convex curvature of the ball-Earth. Once again, however, their hasty conclusion is drawn from a faulty premise, namely that only on a ball-Earth could this phenomenon occur. The fact of the matter is that the Law of Perspective on plane surfaces dictates and necessitates the exact same occurrence. For example a girl wearing a dress walking away towards the horizon will appear to sink into the Earth the farther away she walks. Her feet will disappear from view first and the distance between the ground and the bottom of her dress will gradually diminish until after about half a mile it seems like her dress is touching the ground as she walks on invisible legs. Such is the case on plane surfaces, the lowest parts of objects receding from a given point of observation necessarily disappear before the highest.

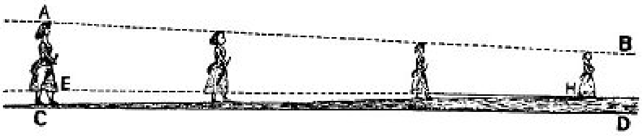

139) Not only is the disappearance of ship's hulls explained by the Law of Perspective on flat surfaces, it is proven undeniably true with the aid of a good telescope. If you watch a ship sailing away into the horizon with the naked eye until its hull has completely disappeared from view under the supposed "curvature of the Earth," then look through a telescope, you will notice the entire ship quickly zooms back into view, hull and all, proving that the disappearance was caused by the Law of Perspective, not by a wall of curved water! This also proves that the horizon is simply the vanishing line of perspective from your point of view, NOT the alleged "curvature" of Earth.

140) Foucault's Pendulums are often quoted as proof of a rotating Earth but upon closer investigation prove the opposite. To begin with, Foucault's pendulums do not uniformly swing in any one direction. Sometimes they rotate clockwise and sometimes counter-clockwise, sometimes they fail to rotate and sometimes they rotate far too much. The behavior of the pendulum actually depends on 1) the initial force beginning its swing and, 2) the ball-and-socket joint used which most-readily facilitates circular motion over any other. The supposed rotation of the Earth is completely inconsequential and irrelevant to the pendulum's swing. If the alleged constant rotation of the Earth affected pendulums in any way, then there should be no need to manually start pendulums in motion. If the Earth's diurnal rotation caused the 360 degree uniform diurnal rotation of pendulums, then there should not exist a stationary pendulum anywhere on Earth!

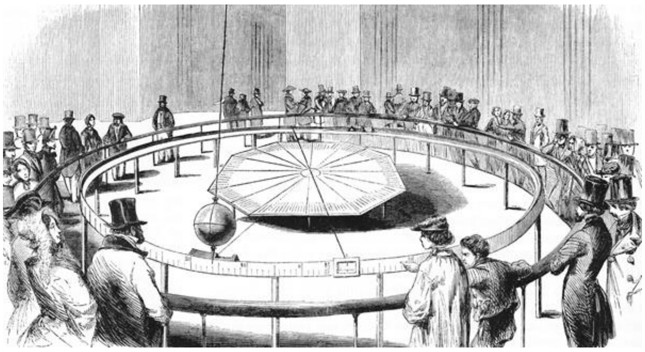

141) The "Coriolis Effect" is often said to cause sinks and toilet bowls in the Northern Hemisphere to drain spinning in one direction while in the Southern Hemisphere causing them to spin the opposite way, thus providing proof of the spinning ball-Earth. Once again, however, just like Foucault's Pendulums spinning either which way, sinks and toilets in the Northern and Southern hemispheres do not consistently spin in any one direction! Sinks and toilets in the very same household are often found to spin opposite directions, depending entirely upon the shape of the basin and the angle of the water's entry, not the supposed rotation of the Earth.

142) People claim that if the Earth were flat, they should be able to use a telescope and see clear across the oceans! This is absurd, however, as the air is full of precipitation especially over the oceans, and especially at the lowest, densest layer of atmosphere is NOT transparent. Picture the blurry haze over roads on hot, humid days. Even the best telescope will blur out long before you could see across an ocean. You can, however, use a telescope to zoom in

MUCH more of our flat Earth than would be possible on a ball 25,000 miles in circumference.

143) People claim that if the Earth were flat, with the Sun circling over and around us, we should be able to see the Sun from everywhere all over the Earth, and there should be daylight even at night-time. Since the Sun is NOT 93 million miles away but rather just a few thousand and shining down like a spotlight, once it has moved significantly far enough away from your location it becomes invisible beyond the horizon and daylight slowly fades until it completely disappears. If the Sun were 93 million miles away and the Earth a spinning ball, the transition from daylight to night would instead be almost instantaneous as you passed the terminator line.

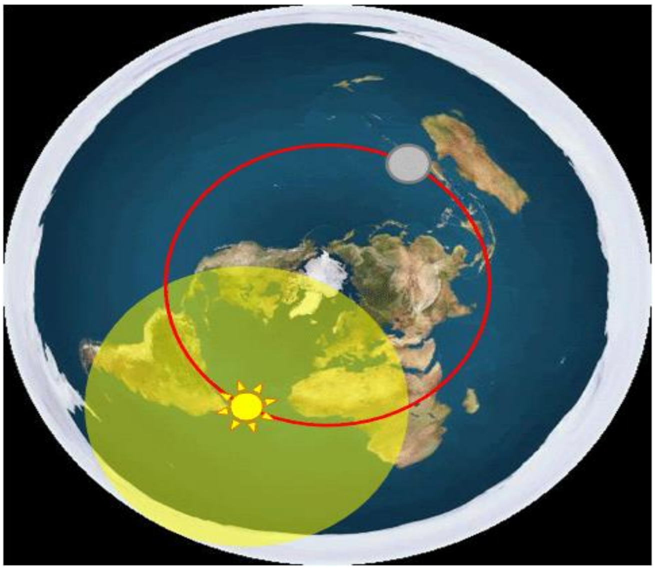

144) Pictures of the Moon appearing upside-down in the Southern hemisphere and right-side up in the North are often cited as proof of the ball-Earth, but once again, upon closer inspection, provide another proof of the flat model. In fact, time-lapse photography shows the Moon itself turns clockwise like a wheel as it circles over and around the Earth. You can find pictures of the Moon at 360 degrees of various inclination from all over the Earth simply depending on where and when the picture was taken.

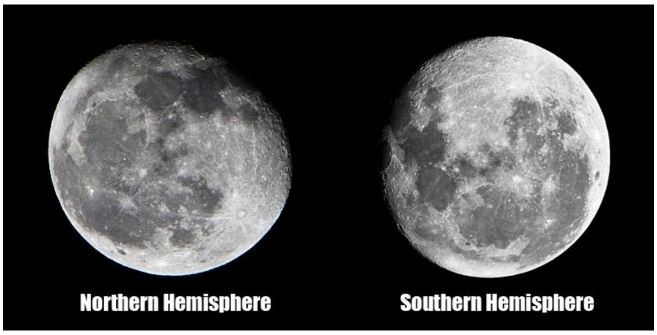

145) Heliocentrists believe the Moon is a ball, even though its appearance is clearly that of a flat luminous disc. We only ever see the same one face (albeit at various inclinations) of the Moon, yet it is claimed that there is another "dark side of the Moon" which remains hidden. NASA states the Moon spins opposite the spin of the Earth in such a perfectly synchronized way that the motions cancel each other out so we will conveniently never be able to observe the supposed dark-side of the Moon outside of their terrible fake CGI images. The fact of the matter is, however, if the Moon were a sphere, observers in Antarctica would see a different face from those at the equator, yet they do not – just the same flat face rotated at various degrees.

6:52am
7:32am
8:33am
9:33am

146) The ball-Earth model claims the Moon orbits around the Earth once every 28 days, yet it is plain for anyone to see that the Moon orbits around the Earth every single day! The Moon's orbit is slightly slower than the Sun's, but follows the Sun's same path from Tropic to Tropic, solstice to solstice, making a full circle over the Earth in just under 25 hours.

147) The ball-Earth model claims the Sun is precisely 400 times larger than the Moon and 400 times further away from Earth making them "falsely" appear exactly the same size. Once again, the ball model asks us to accept as coincidence something that cannot be explained other than by natural design. The Sun and the Moon occupy the same amount of space in the sky and have been measured with sextants to be of equal size and equal distance, so claiming otherwise is against our eyes, experience, experiments and common sense.

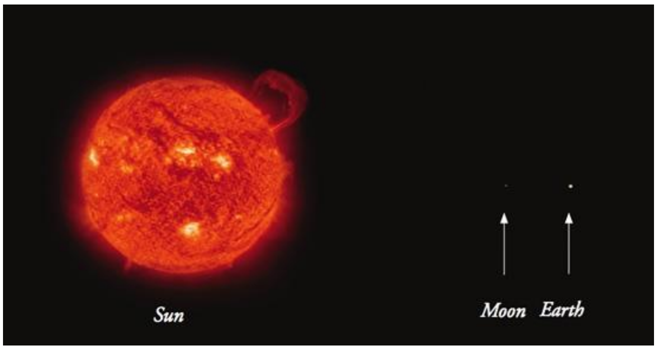

148) Quoting "Earth Not a Globe!" by Samuel Rowbotham, "*It is found by observation that the stars come to the meridian about four minutes earlier every twenty-four hours than the sun, taking the solar time as the standard. This makes 120 minutes every thirty days, and twenty-four hours in the year. Hence all the constellations have passed before or in advance of the sun in that time. This is the simple fact as observed in nature, but the theory of rotundity and* *motion on axes and in an orbit has no place for it. Visible truth must be ignored, because this theory stands in the way, and prevents its votaries from understanding it.*”

149) Throughout thousands of years the same constellations have remained fixed in their same patterns without moving out of position whatsoever. If the Earth were a big ball spinning around a bigger Sun spinning around a bigger galaxy shooting off from the Biggest Bang as NASA claims, it is impossible that the constellations would remain so fixed. Based on their model, we should, in fact, have an entirely different night sky every single night and never repeat exactly the same star pattern twice.

HEMISPHÆRIUM COELI SPHÆRICI GRA.
HIRIBORE ET TERRÆ CASCENO PHIA.

150) If Earth were a spinning ball it would be impossible to photograph star-trail time-lapses turning perfect circles around Polaris anywhere but the North Pole. At all other vantage points the stars would be seen to travel more or less horizontally across the observer’s horizon due to the alleged 1000mph motion beneath their feet. In reality, however, Polaris’s surrounding stars can always be photographed turning perfect circles around the central star all the way down to the Tropic of Capricorn.

151) If Earth were a spinning ball revolving around the Sun it would actually be impossible for star-trail photos to show perfect circles even at the North Pole! Since the Earth is also allegedly moving 67,000mph around the Sun, the Sun moving 500,000mph around the Milky Way, and the entire galaxy going 670,000,000mph, these four contradictory motions would make star-trail time-lapses all show irregular curved lines.

152) In 2003, three University Geography professors collaborated in an experiment to prove that the state of Kansas is indeed actually flatter than a pancake! Using topigraphical geodetic surveys covering over 80,000 square miles it was determined that Kansas has a flatness ratio of 0.9997 over the entire state while the average pancake, precisely measured using a confocal laser microscope comes in at 0.957, making Kansas thereby literally flatter than a pancake.

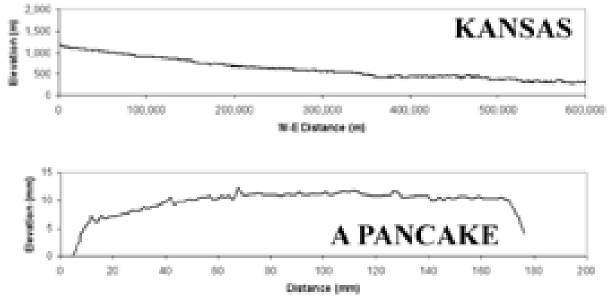

153) Quoting Reverend Thomas Milner’s “Atlas of Physical Geography,” we find that, “*Vast areas exhibit a perfectly dead level, scarcely a rise existing through 1,500 miles from the Carpathians to the Urals. South of the Baltic the country is so flat that a prevailing north wind will drive the waters of the Stattiner Haf into the mouth of the Oder, and give the river a backward flow 30 or 40 miles. The plains of Venezuela and New Granada, in South America chiefly on the left of the Orinoco, are termed Ilanos, or level fields. Often in the space of 270 square miles the surface does not vary a single foot. The Amazon only falls 12 feet in the last 700 miles of its course; the La Plata has only a descent of one thirty-third of an inch a mile.*”

154) The Felix Baumgartner Red Bull dive outside camera shows the same amount of “curvature of Earth” from surface-level to jump-height proving it to be a deceiving fish-eyed wide-angle lens, while the inside regular camera shows a perfectly flat horizon, eye level at 128,000 feet, which is only consistent with a flat plane.

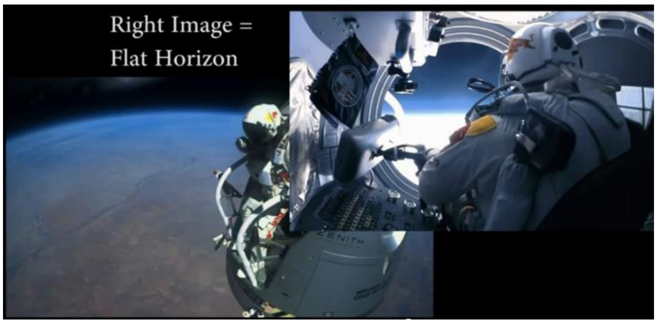

155) Some people claim to have seen the curvature of the Earth out their airplane windows. The glass used in all commercial airplanes, however, is curved to remain flush with the fuselage. This creates a slight effect mixed with confirmation bias people mistake for being the alleged curvature of the Earth. In actuality, the fact that you can see the horizon at eye-level at 35,000 feet out both port/starboard windows proves the Earth is flat. If the Earth were a ball, no matter how big, the horizon would stay exactly where it was and you would have to look DOWN further and further to see the horizon at all. Looking straight out the window at 35,000 feet you should see nothing but "outer-space" from the port and starboard windows, as the Earth/horizon are supposed to be BELOW you. If they are visible at eye level outside both side windows, it’s because the Earth is flat!

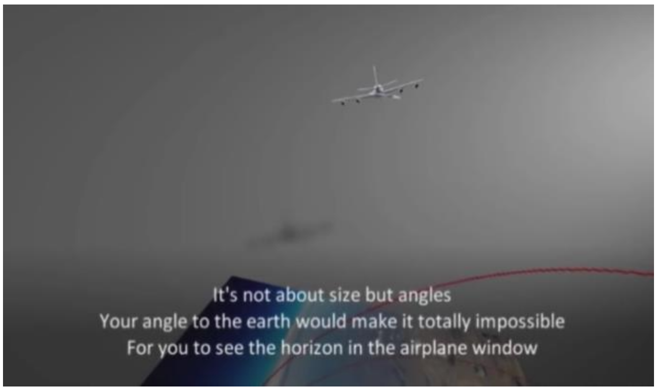
It's not about size but angles
Your angle to the earth would make it totally impossible
For you to see the horizon in the airplane window

156) People also claim to see curvature in Go Pro or other high altitude camera footage of the horizon. While it is true that the horizon often appears convex in such footage, it just as often appears concave or flat depending on the tilt/movement of the camera. The effect is simply a distortion due to wide-angle lenses. In lens-corrected and footage taken without wide-angle technology, all amateur high-altitude horizon shots appear perfectly flat.

Go Pro
Go Pro Lens Corrector

157) If “gravity” magically dragged the atmosphere along with the spinning ball Earth, that would mean the atmosphere near the equator would be spinning around at over 1000mph, the atmosphere over the mid-latitudes would be spinning around 500mph, and gradually slower down to the poles where the atmosphere would be unaffected at 0mph. In reality, however, the atmosphere at every point on Earth is equally unaffected by this alleged force, as it has never been measured or calculated and proven non-existent by the ability of airplanes to fly unabated in any direction without experiencing any such atmospheric changes.

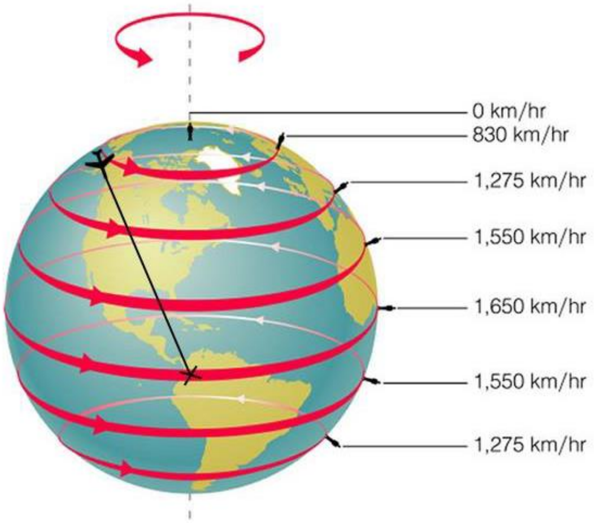

158) If “gravity” magically dragged the atmosphere along with the spinning ball Earth, that would mean the higher the altitude, the faster the spinning atmosphere would have to be turning around the center of rotation. In reality, however, if this were happening then rain and fireworks would behave entirely differently as they fell down through progressively slower and slower spinning atmosphere. Hot-air balloons would also be forced steadily faster Eastwards as they ascended through the ever increasing atmospheric speeds.

159) If there were progressively faster and faster spinning atmosphere the higher the altitude that would mean it would have to abruptly end at some key altitude where the fastest layer of gravitized spinning atmosphere meets the supposed non-gravitized non-spinning non-atmosphere of infinite vacuum space! NASA has never mentioned what altitude this impossible feat allegedly happens, but it is easily philosophically refuted by the simple fact that vacuums cannot exist connected to non-vacuums while maintaining the properties of a vacuum – not to mention, the effect such a transition would have on a rocket “space ship” would be disastrous.

160) It is impossible for rockets or any type of jet propulsion engines to work in the alleged non-atmosphere of vacuum space because without air/atmosphere to push against there is nothing to propel the vehicle forwards. Instead the rockets and shuttles would be sent spinning around their own axis uncontrollably in all directions like a gyroscope. It would be impossible to fly to the Moon or go in any direction whatsoever, especially if “gravity” were real and constantly sucking you towards the closest densest body.

161) If Earth were really a ball, there would be no reason to use rockets for flying into "outer-space" anyway because simply flying an airplane straight at any altitude for long enough should and would send you off into outer-space. To prevent their airplanes from flying tangent to the ball-Earth, pilots would have to constantly course-correct downwards, or else within just a few hours the average commercial airliner traveling 500mph would find themselves lost in "outer-space." The fact that this never happens, artificial horizons remain level at pilot's desired altitudes and do NOT require constant downwards adjustments, proves the Earth is not a ball.

162) All NASA and other "space agencies" rocket launches never go straight up. Every rocket forms a parabolic curve, peaks out, and inevitably starts falling back to Earth. The rockets which are declared "successful" are those few which don't explode or start falling too soon but make it out of range of spectator view before crashing down into restricted waters and recovered. There is no magic altitude where rockets or anything else can simply go up, up, up and then suddenly just start "free-floating" in space. This is all a science-fiction illusion created by wires, green-screens, dark pools, some permed hair and Zero-G planes.

163) NASA and other space agencies have been caught time and again with air bubbles forming and floating off in their official "outer-space" footage. Astronauts have also been caught using scuba-space-gear, kicking their legs to move, and astronaut Luca Parmitano even almost drowned when water started filling up his helmet while allegedly on a "space-walk." It is admitted that astronauts train for their "space-walks" in under-water training facilities like NASA's "Neutral Buoyancy Lab," but what is obvious from their "space bubbles," and other blunders is that all official "space-walk" footage is also fake and filmed under-water.

164) Analysis of many interior videos from the "International Space Station," have shown the use of camera-tricks such as green-screens, harnesses and even wildly permed hair to achieve a zero-gravity type effect. Footage of astronauts seemingly floating in the zero-gravity of their "space station" is indistinguishable from "vomit comet" Zero-G airplane footage. By flying parabolic maneuvers this Zero-G floating effect can be achieved over and over again then edited together. For longer uncut shots, NASA has been caught using simple wires and green screen technology.

I DON'T ALWAYS GO INTO OUTER SPACE
BUT WHEN I DO, IT'S FROM INSIDE A NASA BATHTUB!

165) NASA claims one can observe the International Space Station pass by overhead proving its existence, yet analysis of the "ISS" seen through zoom cameras proves it to be some type of hologram/drone, not a physical floating space-base. As you can see in my documentary "ISS Hoax," when zooming in/out, the "ISS" dramatically and impossibly changes shape and color, displaying a prismatic rainbow effect until coming into focus much like an old television turning on/off.

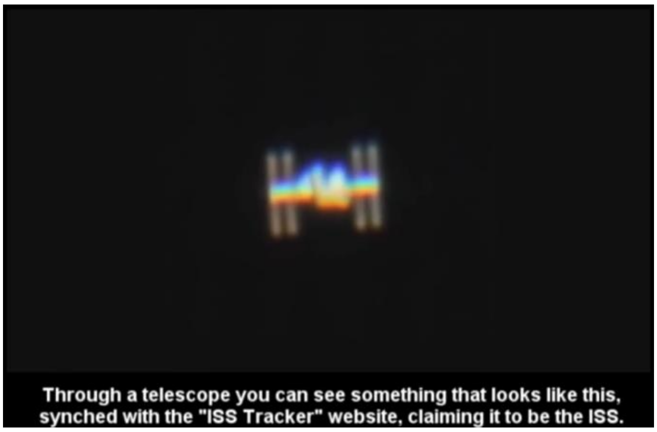
Through a telescope you can see something that looks like this, synched with the "ISS Tracker" website, claiming it to be the ISS.

166) The "geostationary communications satellite" was first created by Freemason science-fiction writer Arthur C. Clarke and supposedly became science-fact just a decade later. Before this, radio, television, and navigation systems like LORAN and DECCA were already well-established and worked fine using only ground-based technologies. Nowadays huge fibre-optics cables connect the internet across oceans, gigantic cell towers triangulate GPS signals, and ionospheric propagation allows radio waves to be bounced all without the aid of the science-fiction best-seller known as "satellites."

photo of satellite
All Artist Impressions

167) Satellites are allegedly floating around in the thermosphere where temperatures are claimed to be upwards of 4,530 degrees Fahrenheit. The metals used in satellites, however, such as aluminum, gold and titanium have melting points of 1,221, 1,948, and 3,034 degrees respectively, all far lower than they could possibly handle.

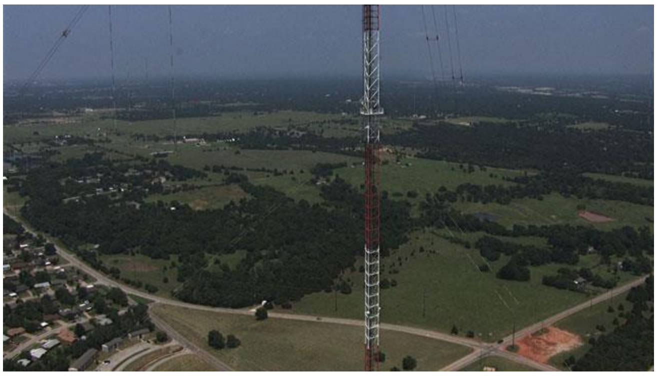

168) So-called "satellite" phones have been found to have reception problems in countries like Kazakhstan with very few cell phone towers. If the Earth were a ball with 20,000+ satellites surrounding, such blackouts should not regularly occur in any rural countryside areas.

169) So-called "satellite" TV dishes are almost always positioned at a 45 degree angle towards the nearest ground-based repeater tower. If TV antennae were actually picking up signals from satellites 100+ miles in space, most TV dishes should be pointing more or less straight up to the sky. The fact that "satellite" dishes are never pointing straight up and almost always positioned at a 45 degree angle proves they are picking up ground-based tower signals and not "outer-space satellites."

170) People even claim to see satellites with their naked eyes, but this is ridiculous considering they are smaller than a bus and allegedly 100+ miles away; It is impossible to see anything so small that far away. Even using telescopes, no one claims to discern the shape of satellites but rather describes seeing passing moving lights, which could easily be any number of things from airplanes to drones to shooting stars or other unidentified flying objects.

171) NASA claims there are upwards of 20,000 satellites floating around Earth's upper-atmosphere sending us radio, television, GPS, and taking pictures of the planet. All these supposed satellite pictures, however, are admittedly "composite images, edited in photoshop!" They claim to receive "ribbons of imagery" from satellites which must then be spliced together to create composite images of the Earth, all of which are clearly CGI and not photographs. If Earth were truly a ball with 20,000 satellites orbiting, it would be a simple matter to mount a camera and take some real photographs. The fact that no real satellite photographs of the supposed ball Earth exist in favor of NASA's "ribbons of composite CG imagery," is further proof we are not being told the truth.

172) If you pick any cloud in the sky and watch for several minutes, two things will happen: the clouds will move and they will morph gradually changing shape. In official NASA footage of the spinning ball Earth, such as the "Galileo" time-lapse video however, clouds are constantly shown for 24+ hours at a time and not moving or morphing whatsoever! This is completely impossible, further proof that NASA produces fake CGI videos, and further evidence that Earth is not a spinning ball.

Come on NASA* make your mind up,
what does the earth really look like?
1975
2015
2007
1997
2007
2013
2012
2002
*ALL IMAGES ARE OFFICIAL NASA PICTURES OF EARTH, AS RELEASED THROUGH THE MAINSTREAM MEDIA OUTLETS AND MAJOR NEWS AGENCIES OF THE WORLD.
QUESTIONS: HOW BIG IS AMERICA? WHAT COLOUR ARE THE OCEANS, AND THE LAND? WHY IS THERE NEVER ANY REAL VIDEO OF THE EARTH SPINNING, ONLY STILLS? WHY ARE THE IMAGES SO VARIED? DO YOU REALLY BELIEVE THIS IS WHERE YOU LIVE? DO YOU ACTUALLY TRUST NASA?

173) NASA has several alleged photographs of the ball-Earth which show several exact duplicate cloud patterns! The likelihood of having two or three clouds of the exact same shape in the same picture is as likely as finding two or three people with exactly the same fingerprints. In fact it is solid proof that the clouds were copied and pasted in a computer program and that such pictures showing a ball-shaped Earth are fakes.

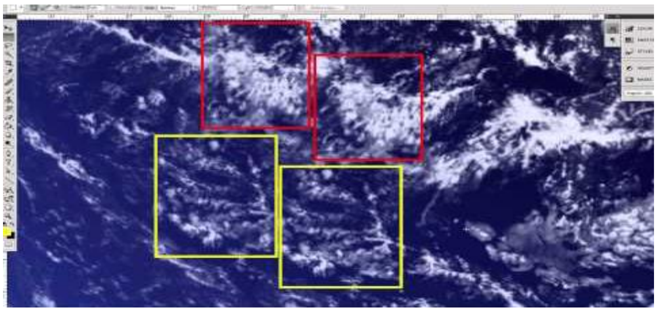

174) NASA graphics artists have placed things like faces, dragons, and even the word "SEX" into cloud patterns over their various ball-Earth pictures. Their recent 2015 Pluto pictures even clearly have a picture of Disney's "Pluto" the dog layered into the background. Such blatant fraud goes unnoticed by the hypnotized masses, but provides further proof of the illegitimacy of NASA and their spinning ball planet mythos.

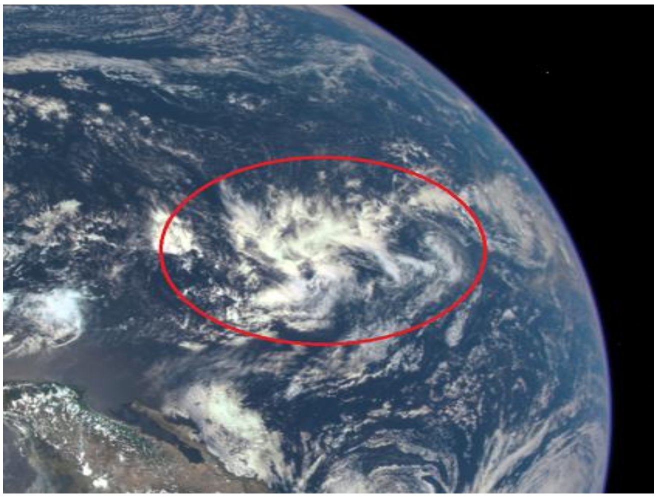

175) Professional photo-analysts have dissected several NASA images of the ball-Earth and found undeniable proof of computer editing. For example, images of the Earth allegedly taken from the Moon have proven to be copied and pasted in, as evidenced by rectangular cuts found in the black background around the "Earth" by adjusting brightness and contrast levels. If they were truly on the Moon and Earth was truly a ball, there would be no need to fake such pictures.

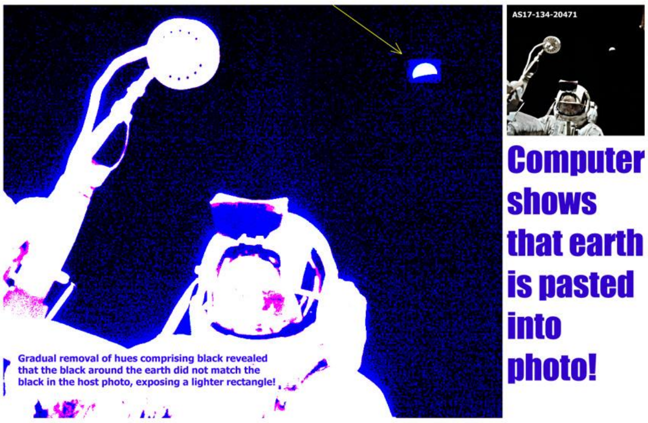
AS17-134-20471
Computer shows that earth is pasted into photo!
Gradual removal of hues comprising black revealed that the black around the earth did not match the black in the host photo, exposing a lighter rectangle!

176) When NASA's images of the ball-Earth are compared with one another the coloration of the land/oceans and relative size of the continents are consistently so drastically different from one another as to prove beyond any reasonable doubt that the pictures are all fake.

How Big is America....
2007
2012
"Photos" courtesy of NASA.

177) In the documentary "A Funny Thing Happened on the Way to the Moon," you can watch official leaked NASA footage showing Apollo 11 astronauts Buzz Aldrin, Neil Armstrong and Michael Collins, for almost an hour, using transparencies and camera-tricks to fake shots of a round Earth! They communicate over audio with control in Houston about how to accurately stage the shot, and someone keeps prompting them on how to effectively manipulate the camera to achieve the desired effect. First, they blacked out all the windows except for a downward facing circular one, which they aimed the camera towards from several feet away. This created the illusion of a ball-shaped Earth surrounded by the blackness of space, when in fact it was simply a round window in their dark cabin. Neil Armstrong claimed at this point to be 130,000 miles from

Earth, half-way to the Moon, but when camera-tricks were finished the viewer could see for themselves the astro-nots were not more than a couple dozen miles above the Earth's surface, likely flying in a high-altitude plane!

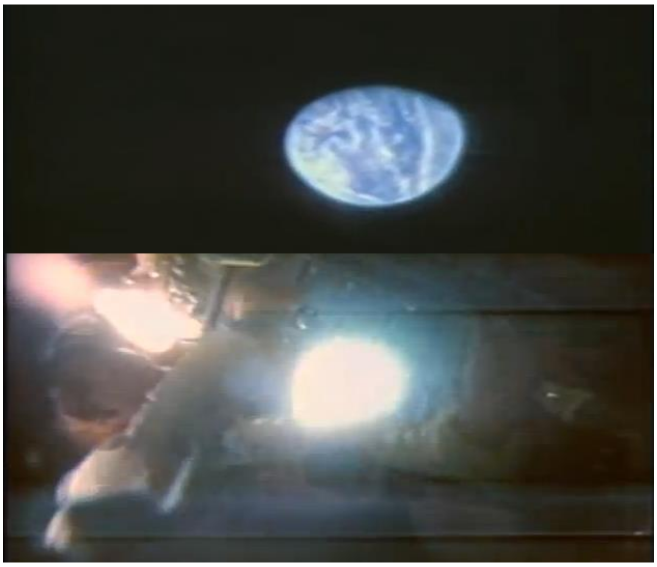

178) People claim Google Earth somehow proves the ball model without realizing that Google Earth is simply a composite program of images taken from high-altitude planes and street-level car-cameras superimposed onto a CGI model of a ball Earth. The same could be just as easily modeled onto a square Earth or any other shape and therefore cannot be used as proof of Earth's rotundity.

Google earth

179) If the Earth were constantly spinning Eastwards 1000mph then airplane flight durations going Eastwards vs. Westwards should be significantly different. If the average commercial airliner travels 500mph, it follows that Westbound equatorial flights should reach their destination at approximately thrice the speed as their Eastbound return flights. In reality, however, the differences in East/Westbound flight durations usually amount to a matter of minutes, and nothing near what would occur on a 1000mph spinning ball Earth.

180) The spinning ball model dictates that the Earth and atmosphere would be moving together at approximately 500mph at the mid-latitudes where an LA to NYC flight takes place. The average commercial airliner traveling 500mph takes 5.5 hours traveling East with the alleged rotation of the Earth, so the return flight West should take only 2.75 hours, but in fact we find the average NYC to LA flight takes 6 hours, a flight time totally inconsistent with the spinning ball model.

181) Flights Eastwards with the alleged spin of the ball-Earth from Tokyo to LA take an average of 10.5 hours, therefore the return flights Westwards against the alleged spin should take an average of 5.25 hours, but in actual fact take an average of 11.5 hours, another flight time totally inconsistent with the spinning ball model.

182) Flights Eastwards with the alleged spin of the ball-Earth from NY to London take an average of 7 hours, therefore the return flights Westwards against the alleged spin should take an average of 3.5 hours, but in actual fact take an average of 7.5 hours, a flight time totally inconsistent with the spinning ball model.

183) Flights Eastwards from Chicago to Boston with the alleged spin of the ball-Earth take an average of 2.25 hours, therefore the return flights Westwards against the alleged spin should take an average of just over an hour, but in actual fact take an average of 2.75 hours, once again, completely inconsistent with the spinning ball model.

184) Flights Eastwards from Paris to Rome with the alleged spin of the ball-Earth take an average of 2 hours, therefore the return flights Westwards against the alleged spin should take an average of 1 hour, but in actual fact have an average flight duration of 2 hours 10 minutes, a flight time totally inconsistent with the spinning ball model.

NASA
Your Balls Are Showing!

185) We are told that the Earth and atmosphere spin together at such a perfect uniform velocity that no one in history has ever seen, heard, felt or measured the supposed 1000mph movement. This is then often compared to traveling in a car at uniform velocity, where we only feel the movement during acceleration or deceleration. In reality, however, even with eyes closed, windows up, over smooth tar in a luxury car at a mere uniform 50mph, the movement absolutely can be felt! At 20 times this speed, Earth's imaginary 1000mph spin would most certainly be noticeable, felt, seen and heard by all.

186) People sensitive to motion sickness feel distinct unease and physical discomfort from motion as slight as an elevator or a train ride. This means that the 1000mph alleged uniform spin of the Earth has no effect on such people, but add an extra 50mph uniform velocity of a car and their stomach starts turning knots. The idea that motion sickness is nowhere apparent in anyone at 1000mph, but suddenly comes about at 1050mph is ridiculous and proves the Earth is not in motion whatsoever.

187) The second law of thermodynamics, otherwise known as the law of entropy, along with the fundamental principles of friction/resistance determine the impossibility of Earth being a uniformly spinning ball. Over time, the spinning ball Earth would experience measurable amounts of drag constantly slowing the spin and lengthening the amount of hours per day. As not the slightest such change has ever been observed in all of recorded history it is absurd to assume the Earth has ever moved an inch.

188) Over the years NASA has twice changed their story regarding the shape of the Earth. At first they maintained Earth was a perfect sphere, which later changed to an “oblate spheroid” flattened at the poles, and then changed again to being “pear-shaped” as the Southern hemisphere allegedly bulges out as well. Unfortunately for NASA, however, none of their official pictures show an oblate spheroid or pear-shaped Earth! All their pictures, contrary to their words, show a spherical (and clearly CGI fake) Earth.

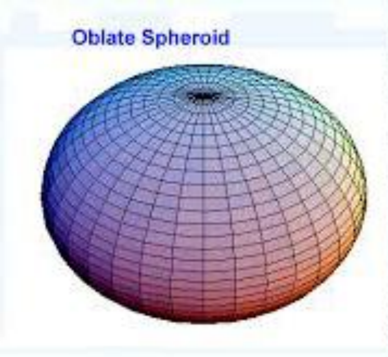
Oblate Spheroid

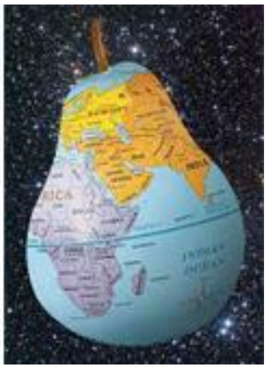

189) The Bible, Koran, Srimad Bhagavatam, and many other holy books describe and purport the existence of a geocentric, stationary flat Earth. For example, 1 Chronicles 16:30 and Psalm 96:10 both read, “*He has fixed the earth firm, immovable*.” And Psalm 93:1 says, “*The world also is stablished, that it cannot be moved*.” The Bible also repeatedly affirms that the Earth is “outstretched” as a plane, with the outstretched heavens everywhere above (not all around) giving a scriptural proof the Earth is not a spinning ball.

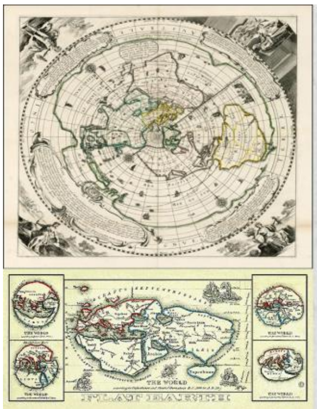

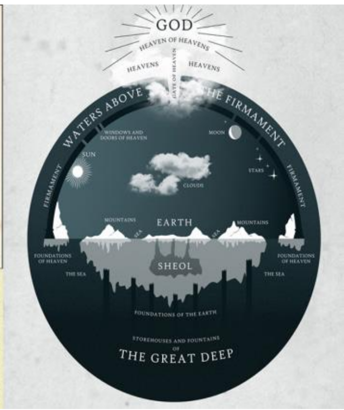
GOD
HEAVEN OF HEAVENS
HEAVENS
HEAVENS
WATERS ABOVE
THE FIRMAMENT
WINDOWS AND DOORS OF HEAVEN
SUN
MOON
STARS
CLOUDS
FIRMAMENT
FIRMAMENT
MOUNTAINS
EARTH
MOUNTAINS
FOUNDATIONS OF HEAVEN
FOUNDATIONS OF HEAVEN
THE SEA
SHEOL
THE SEA
FOUNDATIONS OF THE EARTH
STOREHOUSES AND FOUNTAINS OF
THE GREAT DEEP

190) Cultures the world over throughout history have all described and purported the existence of a geocentric, stationary flat Earth. Egyptians, Indians, Mayans, Chinese, Native Americans and literally every ancient civilization on Earth had a geocentric flat-Earth cosmology. Before Pythagoras, the idea of a spinning ball-Earth was non-existent and even after Pythagoras it remained an obscure minority view until 2000 years later when Copernicus began reviving the heliocentric theory.

191) From Pythagoras to Copernicus, Galileo and Newton, to modern astronauts like Aldrin, Armstrong and Collins, to director of NASA and Grand Commander of the 33rd degree C. Fred Kleinknecht, the founding fathers of the spinning ball mythos have all been Freemasons! The fact that so many members of this, the largest and oldest secret society in existence have all been co-conspirators bringing about this literal “planetary revolution” is beyond the possibility of coincidence and provides proof of organized collusion in creating and maintaining this multi-generational deception.

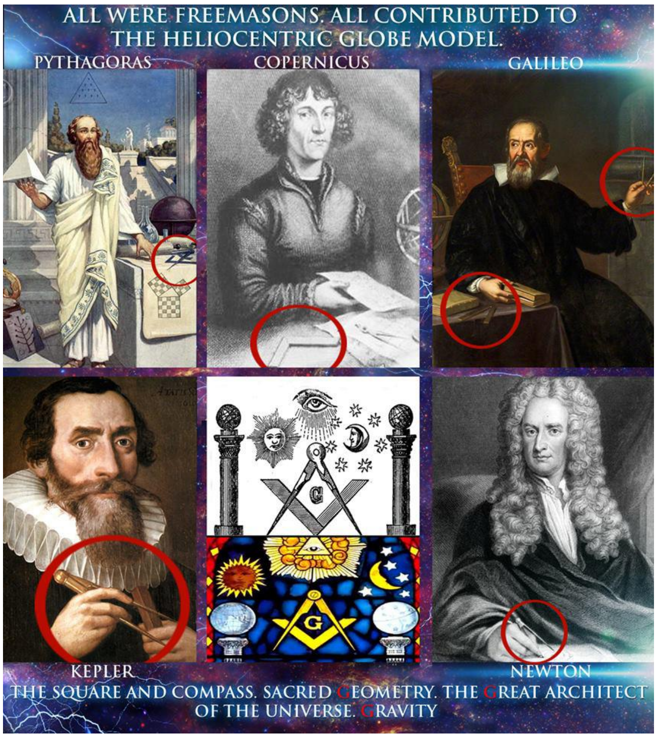
ALL WERE FREEMASONS. ALL CONTRIBUTED TO THE HELIOCENTRIC GLOBE MODEL.
PYTHAGORAS
COPERNICUS
GALILEO
KEPLER
NEWTON
G
G
THE SQUARE AND COMPASS. SACRED GEOMETRY. THE GREAT ARCHITECT OF THE UNIVERSE. GRAVITY

192) Quoting “Terra Firma” by David Wardlaw Scott, “*The system of the Universe, as taught by Modern Astronomers, being founded entirely on theory, for the truth of which they are unable to advance one single real proof, they have entrenched themselves in a conspiracy of silence, and decline to answer any objections which may be made to their hypotheses … Copernicus himself, who revived the theory of the heathen philosopher Pythagoras, and his great exponent Sir Isaac Newton, confessed that their system of a revolving Earth was only a possibility, and could not be proved by facts. It is only their followers who have decorated it with the name of an ‘exact science,’ yea, according to them, ‘the most exact of all the sciences.’ Yet one Astronomer Royal for England once said, speaking of the motion of the whole Solar system: ‘The matter is left in a most delightful state of uncertainty, and I shall be very glad if any one can help me out of it.’ What a very sad position for an ‘exact science’ to be in is this!*”

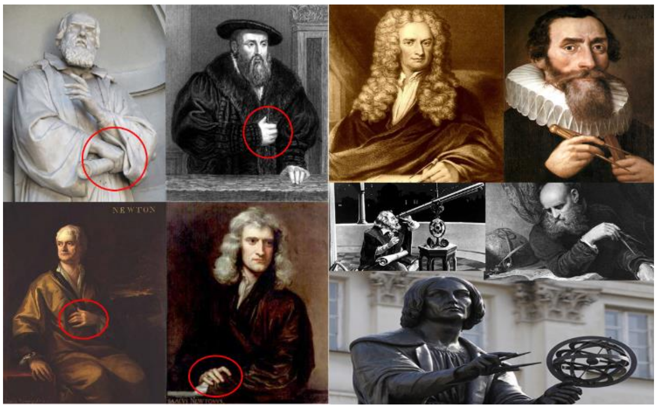

193) No child or un-indoctrinated man in their right-mind would ever conclude or even conceive given to their own devices, based on their own personal observations, that the Earth was a spinning ball revolving around the Sun! Such imaginative theories nowhere present in anyone’s daily experience require and have required massive amounts of constant propaganda to uphold the illusion.

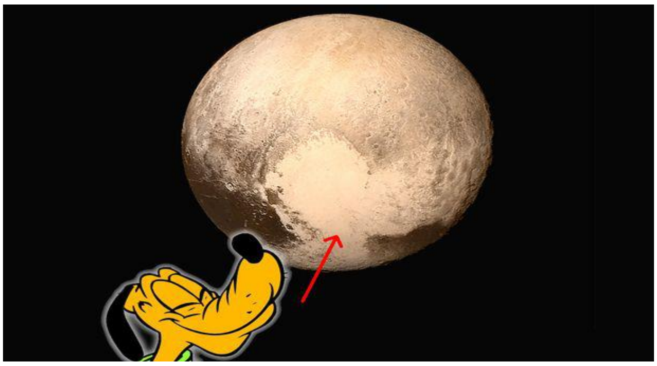

194) From David Wardlaw Scott, “*I remember being taught when a boy, that the Earth was a great ball, revolving at a very rapid rate around the Sun, and, when I expressed to my teacher my fears that the waters of the oceans would tumble off, I was told that they were prevented from doing so by Newton’s great law of Gravitation, which kept everything in its proper place. I presume that my countenance must have shown some signs of incredulity, for my teacher immediately added - I can show you a direct proof of this; a man can whirl around his head a pail filled with water without its being spilt, and so, in like manner, can the oceans be carried round the Sun without losing a drop. As this illustration was evidently intended to settle the matter, I then said no more upon the subject. Had such been proposed to me afterwards as a man, I would have answered somewhat as follows - Sir, I beg to say that the illustration you have given of a man whirling a pail of water round his head, and the oceans revolving round the Sun, does not in any degree confirm your argument, because the water in the two cases is placed under entirely different circumstances, but, to be of any value, the conditions in each case must be the same, which here they are not. The pail is a hollow vessel which holds the water inside it, whereas, according to your teaching, the Earth is a ball, with a continuous curvature outside, which, in agreement with the laws of nature, could not retain any water.*”

195) Astronomers say the magical magnetism of gravity is what keeps all the oceans of the world stuck to the ball-Earth. They claim that because the Earth is so massive, by virtue of this mass it creates a magic force able to hold people, oceans and atmosphere tightly clung to the underside of the spinning ball. Unfortunately, however, they cannot provide any practical example of this on a scale smaller than the planetary. A spinning wet tennis ball, for instance, has the exact opposite effect of the supposed ball-Earth! Any water poured over it simply falls off the sides, and giving it a spin results in water flying off 360 degrees like a dog shaking after a bath. Astronomers concede the wet tennis ball example displays the opposite effect of their supposed ball-Earth, but claim that at some unknown mass, the magic adhesive properties of gravity suddenly kick in allowing the spinning wet tennis ball-Earth to keep every drop of “gravitized” water stuck to the surface. When such an unproven theory goes against all experiments, experience and common sense, it is high time to drop the theory.

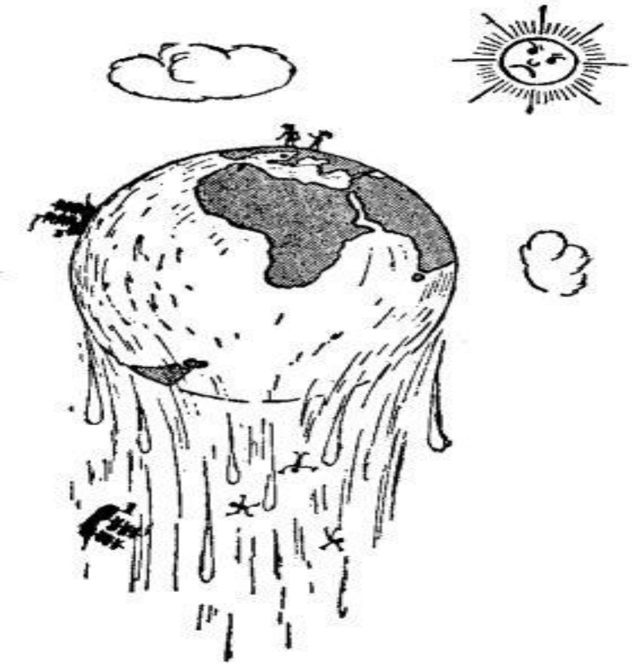

196) Quoting Marshall Hall, "*In short, the sun, moon, and stars are actually doing precisely what everyone throughout all history has seen them do. We do not believe what our eyes tell us because we have been taught a counterfeit system which demands that we believe what has never been confirmed by observation or experiment. That counterfeit system demands that the Earth rotate on an 'axis' every 24 hours at a speed of over 1000 MPH at the equator. No one has ever, ever, ever seen or felt such movement (nor seen or felt the 67,000MPH speed of the Earth's alleged orbit around the sun or its 500,000 MPH alleged speed around a galaxy or its retreat from an alleged 'Big Bang' at over 670,000,000 MPH!). Remember, no experiment has ever shown the earth to be moving. Add to that the fact that the alleged rotational speed we've all been taught as scientific fact MUST decrease every inch or mile one goes north or south of the equator, and it becomes readily apparent that such things as accurate aerial bombing in WWII (down a chimney from 25,000 feet with a plane going any direction at high speed) would have been impossible if calculated on an earth moving below at several hundred MPH and changing constantly with the latitude*."

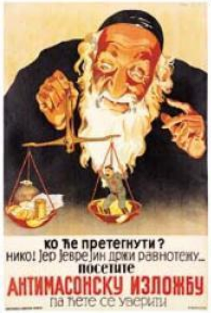

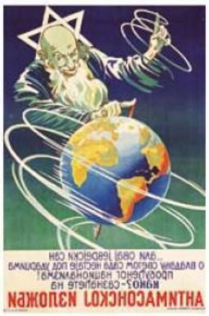

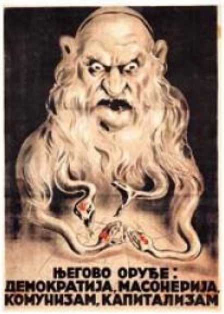

197) Some people claim there is no motive for such a grand-scale deception and that flat or a ball makes no difference. By removing Earth from the motionless center of the Universe, these Masons have moved us physically and metaphysically from a place of supreme importance to one of complete nihilistic indifference. If the Earth is the center of the Universe, then the ideas of God, creation, and a purpose for human existence are resplendent. But if the Earth is just one of billions of planets revolving around billions of stars in billions of galaxies, then the ideas of God, creation, and a specific purpose for Earth and human existence become highly implausible. By surreptitiously indoctrinating us into their scientific materialist Sun-worship, not only do we lose faith in anything beyond the material, we gain absolute faith in materiality, superficiality, status, selfishness, hedonism and consumerism. If there is no God, and everyone is just an accident, then all that really matters is me, me, me. They have turned Madonna, the Mother of God, into a material girl living in a material world. Their rich, powerful corporations with slick Sun-cult logos sell us idols to worship, slowly taking over the world while we tacitly believe their "science," vote for their politicians, buy their products, listen to their music, and watch their movies, sacrificing our souls at the altar of materialism. To quote Morris Kline, "*The heliocentric theory, by putting the sun at the center of the universe ... made man appear to be just one of a possible host of wanderers drifting through a cold sky. It seemed less likely that he was born to live gloriously and to attain paradise upon his death. Less likely, too, was it that he was the object of God's ministrations."*

198) Some say the idea of an inter-generational world-wide conspiracy to delude the masses sounds implausible or unrealistic, but these people need only familiarize themselves with the works and writings of Freemasons themselves, for example John Robison who exposed this in his 1798 book, "Proofs of a Conspiracy Against All the Religions and Governments of Europe Carried Out in the Secret Meetings of the Freemasons, Illuminati and Reading Societies." Supreme Commander of the 33rd degree Albert Pike was quite forth-coming in several letters regarding the Masons ultimate goal of world domination, and in the Zionist "Protocols of the Learned Elders of Zion" the exact plan by which this would be and has been carried out is completely disclosed.

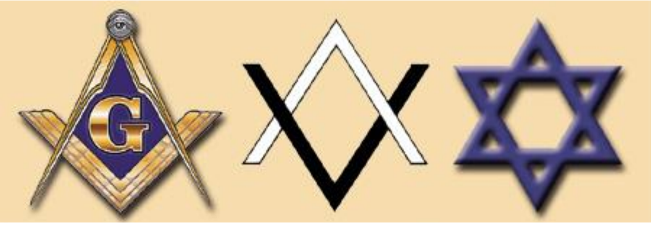

199) From "Foundations of Many Generations" by E. Eschini, "*The one thing the fable of the revolving Earth has done, it has shown the terrible power of a lie, a lie has the power to make a man a mental slave, so that he dares not back the evidence of his own senses. To deny the plain and obvious movement of the Sun he sees before him. When he feels himself standing on an Earth utterly devoid of motion, at the suggestion of someone else he is prepared to accept that he is spinning furiously round. When he sees a bird flying, and gaining over the ground, he is prepared to believe that the ground is really travelling a great number of times faster than the bird, finally, in order to uphold the imagination of a madman, he is prepared to accuse his Maker of forming him a sensiferous lie.*"

200) And finally, from Dr. Rowbotham, "*Thus we see that this Newtonian philosophy is devoid of consistency; its details are the result of an entire violation of the laws of legitimate reasoning, and all its premises are assumed. It is, in fact, nothing more than assumption upon assumption, and the conclusions derived therefrom are willfully considered as things proved, and to be employed as truths to substantiate the first and fundamental assumptions. Such a 'juggle and jumble' of fancies and falsehoods extended and intensified as in theoretical astronomy is calculated to make the unprejudiced inquirer revolt with horror from the terrible conjuration which has been practised upon him; to sternly resolve to resist its further progress; to endeavour to over-throw the entire edifice, and to bury in its ruins the* *false honours which have been associated with its fabricators, and which still attach to its devotees. For the learning, the patience, the perseverance and devotion for which they have ever been examples, honour and applause need not be withheld; but their false reasoning, the advantages they have taken of the general ignorance of mankind in respect to astronomical subjects, and the unfounded theories they have advanced and defended, cannot be otherwise than regretted, and ought to be by every possible means uprooted.*”

The Flat-Earth Conspiracy
By Eric Dubay

For more information about our Flat Earth read “The Flat Earth Conspiracy” by Eric Dubay and visit:

http://www.AtlanteanConspiracy.com

http://www.ifers.boards.net

EARTH is FLAT!
ifers.
boards.
net

the organic gardener

How to create flower, vegetable, herb and fruit gardens using completely natural techniques

Central Highlands Regional Library
35/6063179d

christine and michael lavelle

CONTENTS

Dedication

This book is dedicated to Mike Lavelle who would have loved to have seen this book.

This edition is published by Southwater

Southwater is an imprint of
Anness Publishing Ltd
Hermes House
88–89 Blackfriars Road, London SE1 8HA
tel. 020 7401 2077; fax 020 7633 9499
www.southwaterbooks.com; info@anness.com

© Anness Publishing Ltd 2004

UK agent: The Manning Partnership Ltd
tel. 01225 478444; fax 01225 478440
sales@manning-partnership.co.uk

UK distributor: Grantham Book Services Ltd
tel. 01476 541080; fax 01476 541061
orders@gbs.tbs-ltd.co.uk

North American agent/distributor:
National Book Network, tel. 301 459 3366
fax 301 429 5746; www.nbnbooks.com

Australian agent/distributor: Pan Macmillan Australia
tel. 1300 135 113; fax 1300 135 103
customer.service@macmillan.com.au

New Zealand agent/distributor: David Bateman Ltd
tel. (09) 415 7664; fax (09) 415 8892

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any way or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the copyright holder.

A CIP catalogue record for this book is available from the British Library.

Publisher: **Joanna Lorenz**
Editorial Director: **Judith Simons**
Project Editor: **Felicity Forster**
Designer: **Lisa Tai**
Jacket design: **Balley Design Associates**
Additional text: **Richard Bird and Jonathan Edwards**
Additional photography: **Jonathan Buckley and Michelle Garrett**
Illustrator: **Liz Pepperell**
Production Controller: **Darren Price**

Previously published as part of a larger volume, *Organic Gardening*

10 9 8 7 6 5 4 3 2 1

INTRODUCTION 6
What is Organic Gardening? 8
Why Go Organic? 10
Understanding Your Garden 12
The Organic Cycle 14

SOIL AND SOIL MANAGEMENT 16
Types of Soil 18
Soil Structure 20
Soil Chemistry 22
Digging and Cultivation 24
No-Dig Gardening 26
Soil Conditioners 28
Making Garden Compost 32
Green Manures 34

BASIC TECHNIQUES 36
Weeding and Weed Control 38
Watering 40
Feeding 42
Basic Pruning 45
Plant Propagation 48
The Greenhouse Environment 54
Greenhouse Cultivation 56
Cold Frames and Cloches 58

PLANT HEALTH 60
Why Plants Get Sick 62
Preventing Problems 64
Plant Pests 66
Plant Diseases 68
Plant Disorders 70
Beneficial Predators 72
Other Control Methods 74

THE ORNAMENTAL GARDEN 78
The Garden Framework 80
Ornamental Lawns 82
Lawn Maintenance 84
Woody Plants 88
Flowers for Beds and Borders 92
Beds and Borders 96
Containers 100

THE WILDLIFE GARDEN 104
What is a Wildlife Garden? 106
Woodland Gardens 108
Wild Flower Lawns 110
Wildlife Ponds 112
Providing Shelter 114
Plants to Attract Wildlife 116

THE KITCHEN GARDEN 120
Planning a Kitchen Garden 122
Crop Varieties 124
Crop Rotation 126
Growing Methods 128
Companion Planting 130
Making the Most of Your Space 132
Sowing in the Open 134
Sowing Under Glass 136
Thinning and Transplanting 138
Aftercare 140
Growing Herbs 142
Harvesting and Storing Herbs 144
Growing Fruit Trees and Bushes 146
Supporting Fruit Trees and Shrubs 148
Protecting Fruit 149
Pruning Fruit Trees and Shrubs 150
Harvesting and Storing Fruit 152

CALENDAR OF CARE 154
The Ornamental Garden 154
The Wildlife Garden 154
The Kitchen Garden 156

Index 158
Acknowledgements 160

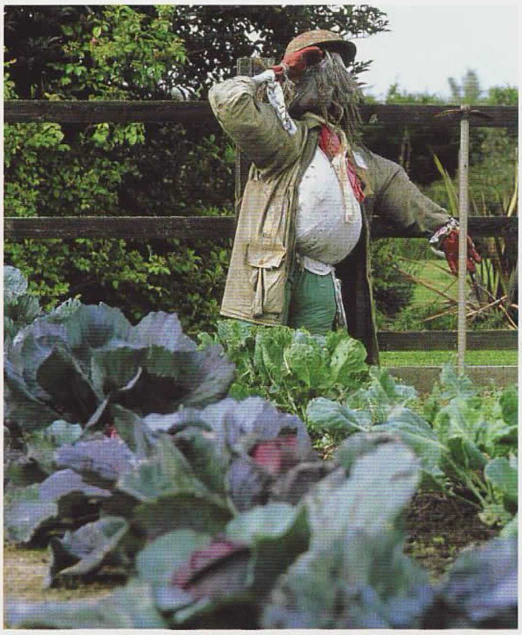

INTRODUCTION

The aim of this book is to show that organic gardening is gardening at its best. The most successful organic gardeners learn that the use of artifical chemicals for short-term gains leads to long-term losses and also that nature makes the best model. Whether you want to grow organic produce for your table, have a dazzling display of summer colour or create a wildlife paradise, the organic garden is for you. Many techniques may already be familiar to you and the change will be less difficult than you think. Going organic may not change the world immediately, but being an organic gardener means that you can make a difference to the wellbeing of the world.

Left: *A well-maintained ornamental garden will reward the keen organic gardener with a wealth of dazzling colour.*

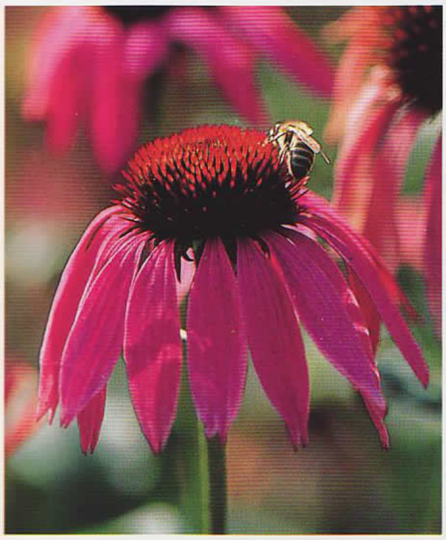

Above: *Organic gardens are highly beneficial to wildlife because they mimic natural environments.*

Above: *The organic garden is the ideal place in which to grow a selection of fresh healthy vegetables.*

Above: *Ornamental plants and flowers are perfectly at home and thrive in a garden that uses organic principles.*

WHAT IS ORGANIC GARDENING?

The term organic means of living origin. When applied to gardening, organic has come to mean the systematic use of techniques that mirror naturally occurring systems. Put simply, it is about finding environmentally friendly ways to cultivate the land, working with rather than against nature. Organic gardening refers not just to a system of techniques, however, but also to a whole philosophy of life.

Above: *The organic garden uses a range of natural materials, including by-products such as chicken manure.*

THE ORIGINS OF ORGANIC GARDENING

From the point of view of gardening, the "modern" organic movement began in the late1940s as a reaction to the increased use of pesticides and synthetic fertilizers in the years after the Second World War. In many respects, however, the principles of organic gardening have been practised for centuries. Ancient writers, among them Pliny and Virgil, commented on the importance of "good husbandry to the health of the land". Thomas Tusser, in his classic work of 1580 entitled *Five Hundred Points of Good Husbandry*, recommends crop rotation to maintain good health. The 17th-century English garden writer John Evelyn begins his *Kalendarium Hortense* with a section describing how to enrich the ground in mid-winter with "horse and sheeps dung especially, that you may have some of two years preparation".

Above: *Many of the plants that benefit wildlife also make showy and decorative additions to beds and borders.*

ORGANIC GARDENING IN THE 21ST CENTURY

Much of the current interest in organic gardening began in the 1960s, when there was increasing concern about the growing levels of environmental damage caused by pesticides and other agrochemicals. If they were causing so much damage to the natural world, then surely they must ultimately affect human beings?

The organic approach aims to reduce the effects that our gardens, farms and cities have upon the wider natural environment. Activities such as recycling, using sustainably produced materials and avoiding pesticides and other harmful agrochemicals all help in this. Organic gardening is often described as being a more natural way to garden. This can make it appear to be somehow revolutionary. In many ways, however, organic gardening could be said to be counter-revolutionary. It aims to avoid artificial inputs and gains. Instead it draws from a vast resource of wisdom and experience amassed over 10,000 years. Much of the so-called conventional wisdom is, in fact, very recent. We only have to look back as far as our grandparents to see that they were naturally organic gardeners. Organic gardening then is the marriage of good horticultural practice to an awareness of our impact upon our surroundings.

IS IT POSSIBLE TO BE WHOLLY ORGANIC?

The short answer to this would probably have to be … with great difficulty. But the purpose of this book is not to promote difficulty or set unattainable goals. The aim should always be to aspire to the ideal solution. Almost all of us face compromise on a daily basis and understand that practicalities outweigh personal ideals. With perseverance and practice, however, it is possible to become less compromised in the confines of your own garden. Simple planning and the observance of good

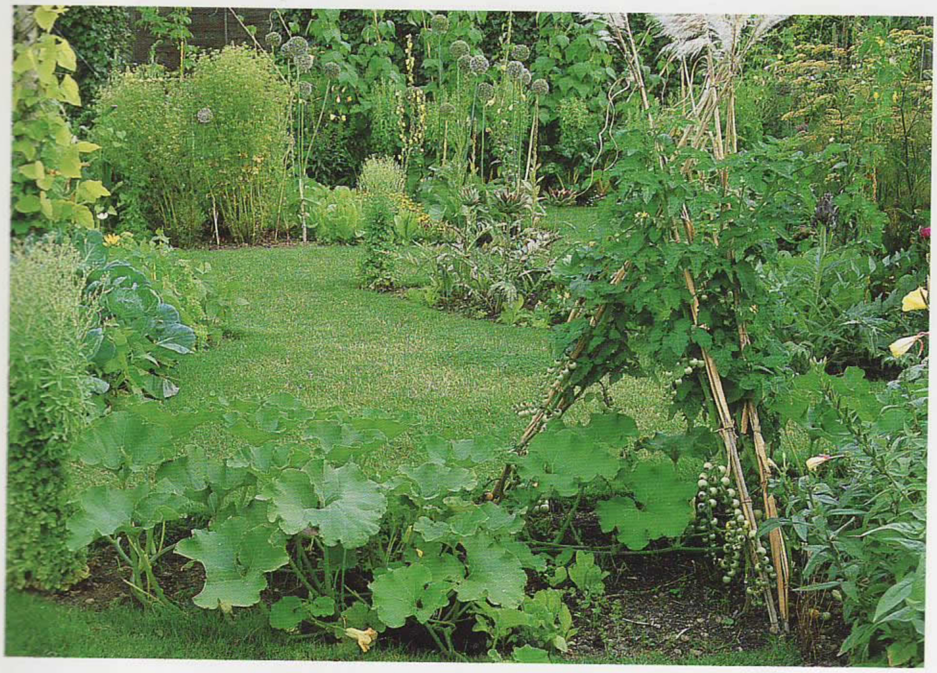

Left: *Organic gardens can be beautifully designed and can include a mixture of ornamental and edible plants.*

Above: *A well-designed and properly maintained organic garden can be both attractive and productive.*

gardening practice can steadily improve your organic credentials and, with time, a natural balance will be established in your garden. Ultimately, the aim of all organic gardeners, be they landscaping a city garden, tending a country estate or growing home-produced vegetables, is to make choices appropriate to their situation. It is better to move a few steps towards the organic way than to ignore it completely.

Above: *Vegetables need not be planted in rows. Here, the beds are arranged in an attractive geometric pattern.*

ORGANIC STANDARDS

Organic standards are set out to explain the requirements that farmers, growers, processors and others must meet in order for their products or services to be marketed as organic. The standards can be extensive and cover a wide range of farming, growing and food manufacturing practices.

Are organic standards the same worldwide? No. Different countries may stipulate their own standards. IFOAM is the International Federation of Organic Agricultural Movements that represents the worldwide body of organic agriculture and provides a platform for the global exchange of information and co-operation.

Are any chemicals allowed in organic production? Yes. In line with the legal framework of the country, a very limited number of chemicals are allowed on a restricted basis. Organic standards do not allow the use of artificial herbicides or fungicides. By comparison, in non-organic farming as many as 450 chemicals can be routinely used.

If organic gardening severely restricts the use of artificial chemical fertilizers and pesticides, then how do I keep my garden healthy? Organic gardeners rely on developing a healthy fertile soil and growing a mixture of crops. By following these basic practices, organic gardeners work in harmony with nature and aim to achieve a healthy natural balance within their gardens.

Do I have to achieve any standards to become organic? No. Commercial standards are very strict and not necessarily easy to achieve for the amateur. Everyone must make choices appropriate to their own situation.

Is the organic movement just concerned with growing food? No. Going organic is not just about organic food – it should become a way of life. Today, organic wood, clothing, gardening products and even restaurants can all be found. This means that other areas of your life can be organic as well as your garden. It does not always mean completely changing your life, but it can change lives.

Above: *Butterflies are just one of the many different types of wildlife that can benefit from a pesticide-free garden.*

WHY GO ORGANIC?

This commonly asked question is easy to answer. The primary reason to embrace the organic approach is health. This means the well-being of the individual as well as the health of the environment. Many pesticides can accumulate over time, both in our own bodies and in the wider environment, a poisonous legacy that can persist for decades. Organic gardening seeks to redress this damage by working with and encouraging nature. It is a long-term investment in the health and wellbeing of us all.

GREEN CUISINE

Hardly a day seems to go by without some new revelation about the benefits or detrimental effects of some food or other. Recent concerns about the "goodness" of the things that we eat have been a major incentive to supermarkets to supply a growing demand for organic produce. There can be little that equals the satisfaction to be derived from eating home-produced food, gathered fresh from the garden. What better than garden-fresh vegetables that are produced naturally and cleanly?

A GREENER, SAFER GARDEN

Although we think of our gardens as places of safe retreat we have used a cocktail of pesticides and agrochemicals on them that has stripped them of their natural diversity. Many of these poisons no longer control the beasts that they were intended to. All that suffers in this noxious onslaught is the garden, its wild occupants and ultimately the plants. This cycle of pesticide use eventually turns your personal paradise into a potential health hazard.

HELPING THE ENVIRONMENT

For many of us it can seem that there is little that we can do to change the degradation of our environment. But the truth is that the actions of all of us have both positive and negative impacts upon the land. If every individual reduces his or her own negative impact upon the environment, then collectively we can make a difference. By changing the way you garden to a more organic way, you can begin the healing process, literally in your own backyard.

PROTECTING AND ENCOURAGING WILDLIFE

Be it a bird on a feeder or brightly coloured butterflies on a flower, most of us enjoy seeing visitors from the wild in our gardens. But many of the unsung heroes of our gardens – the insects, worms, amphibians and mammals that crawl, slither and scurry out of sight – are equally important and may even be the reason why some of the larger visitors come. Organic gardening helps many of these and promotes a balance of wildlife that is both interesting and beautiful to look at.

SO WHAT IF I DO NOT GO ORGANIC?

While the intention of this book is not to dwell upon negative points, the cost of not adopting an organic approach may be seen all around us. This is most frequently noticed as fields and open spaces bereft of butterflies, bees, birds and other wildlife with very little diversity of plant species. It is unthinkable that we could hand this planet to successive generations in a denuded and impoverished state. Perhaps a garden free of "bugs" may sound attractive to some, but the long-term cost to the environment may be felt for generations and in the worst case may even be permanent.

Above: *Organic flower gardens, with their wide range of wildlife-friendly plants, can provide a balanced habitat for birds and insects that is diverse, interesting and beautiful to look at.*

Above: *Imaginative planting designs can result in beautiful contrasts of flowers and foliage, while also providing a wealth of flavourful, home-grown food for your table.*

IT'S A SOLUTION AND IT'S FUN!

Organic gardening is not the solution to all humankind's problems. However, it is a positive and valuable step that you as an individual can take. What's more, you can contribute positively to the wellbeing of your surroundings and to that of your wider environment by doing something that is also fun and rewarding.

WILL "GOING ORGANIC" MEAN THAT MY WHOLE LIFESTYLE HAS TO CHANGE?

Organic gardening is often presented as part of a more general "organic movement". This movement owes its origins to a range of social philosophies, some of which are more fundamentalist than others. The aim of organic gardening is to change the way that you garden and not your whole life. People frequently get confused about the meaning of words that are used as part of the wider organic movement. Four of these terms are explained below. Of these, three describe a whole philosophy that affects practitioners far beyond the confines of their own gardens. An understanding of the philosophies of others can help inspire your own lifestyle changes and all have their own points of interest. In the end, however, it is a matter of choice for you alone and it is best to decide just how organic you (and your family) wish to be.

Biodiversity literally means "the variety of life". It includes all the different plant, animal, fungus and microscopic species in the world. It also considers the genes they contain and the ecosystems of which they form a part. Biodiversity has been seen as the total complexity of all life, including not only the great variety of organisms but also their varying behaviour and interactions. An organic garden aims to maximize all these many different things which are vital to every part of the garden, but especially important when considering the soil. Soil biodiversity is the "lifeblood" of the organic garden and demands investment in order to yield a return.

Biodynamics is described by its practitioners as a science of life forces, recognizing the basic principles at work in nature and taking into account how our actions can bring balance and healing to the world. It is based upon the work of Rudolf Steiner and is said to offer an account of the spiritual history of the Earth as a living being. Its enthusiasts advocate a broadening of personal perspective that includes an understanding of cosmic rhythms, natural healing, a respect for life forces and ultimately a shift away from centralized economic strategies.

Above: *The foliage and form of many types of vegetables can provide striking contrasts in the vegetable garden.*

Permaculture is a concept that was originally developed in the 1970s by Australians Bill Mollinson and David Holgreen. It was conceived as a response to increasing concerns over the deteriorating natural environment. They attempted to answer the question "how do we, as the human species, sustain ourselves and provide for the needs of the environment for an indefinite period of time?". The answer to this question is the basis for creating a permanent culture (Perma-Culture). This philosophy aims to unite practical, ecological design philosophies by incorporating natural systems and materials into human habitats. Permaculture rejects short-term gains in favour of long-term stability for the environment and those who live in it.

Self-sufficiency is an ideal rather than a reality for most that attempt it, as modern living usually dictates some degree of trade, even if it is limited. Advocates of self-sufficiency aim to produce everything that they need in order to support themselves. The only imports are those that cannot be produced by an individual.

REASONS FOR GOING ORGANIC

Below are just a few of the positive benefits that organic gardening may bring to you and your environment.

Human impact on the earth Organic systems reduce dependence on non-renewable resources. The world has an increasing population, but many of our resources are in limited supply. Using renewable resources helps to secure our common future.

Traditional methods with modern techniques Organic systems rely on a modern, scientific understanding of ecology and soil science. They also depend on traditional methods of crop rotation to ensure soil fertility as well as weed and pest control.

Local solutions to global problems The organic movement encourages individuals to act on a local level, while thinking about how this affects the world.

Best for biodiversity Organic growing is friendlier to the environment, so there is a greater diversity of birds, butterflies and plants in organic gardens.

Minimize your waste We produce a huge amount of waste, including packaging and kitchen and garden waste. Recycling and waste reduction are key issues for organic living.

Mimicking natural systems Organic gardening uses nature as its model and, by working in harmony with nature, aims to achieve a natural balance. By maintaining this equilibrium, organic gardens become very stable entities.

A more interesting garden Organic production is more sustainable and friendlier to the environment. By becoming involved in the whole natural cycle, you will gain greater satisfaction from your own personal Eden.

We are what we eat ... Organic systems recognize that our health is connected to the quality of our food and, ultimately, to the health of the soil.

UNDERSTANDING YOUR GARDEN

Plants, like all living things, flourish when given the right conditions. Certain plants will only thrive in a hot, sunny and dry garden, for example, while others may need cool, moist and shady conditions to prosper. If you are to have healthy plants, you need to choose those that are best suited to your garden and this means understanding exactly what conditions prevail and taking into account a range of environmental factors.

Above: *A hot, sunny spot is ideal for growing pretty summer flowers such as argyranthemums and verbenas.*

CLIMATE

The climate causes a variety of responses, depending upon the plant's location, its stage of maturity, the length of exposure and the intensity of the type of weather. Climate may have a dramatic effect on plant growth and development, especially when extreme weather conditions prevail.

TEMPERATURE

All plant species have their own maximum and minimum temperature tolerances, beyond which the life processes of the plant cease. As a general rule, the maximum temperature that most plants can tolerate is around 35°C (95°F), while the minimum is highly variable. The air and soil temperatures are also crucial in influencing dormancy within plants and this, in turn, largely dictates the length of the growing season.

Air temperature is affected by the degree of energy received from the sun. A sheltered site that benefits from the warming effects of the sun may be used for growing plants that are indigenous to warmer climates.

Soil temperature influences a plant's root development and the rate at which water and nutrients can be absorbed. Sandy soils warm up quicker and earlier in the season than clay soils, mainly because they are relatively free draining and do not hold as much water. Sites that receive a lot of sun, or with a slight incline towards it, will also warm up quicker than shady ones.

WATER

This is the major constituent of all plants. In most gardens, rainfall is the principal source of water for plants and it can be lost through evaporation or surface run-off, although much of it is soaked up by soil particles. This water may then be absorbed by the root hairs of the plant. A steady supply of water is essential for the plant to sustain itself and for optimum growth.

Water may also be held in the air and is referred to as humidity. The amount of water vapour in the atmosphere at any one time is referred to as the relative humidity and is measured as a percentage of the saturation point (100 per cent humidity). In areas that receive heavy rainfall, relative humidity is also higher. Many plants, such as ferns and mosses, thrive in such conditions. But high relative humidity can have undesirable effects on plants, often encouraging disease.

Waterlogging refers to the build-up of water that may occur in badly drained soil with poor structure. The roots of plants that are not adapted to these conditions will suffer and probably die through asphyxiation.

Left: *A bog garden provides ideal conditions for a range of decorative moisture-loving plants.*

Right: *Blossom is both pretty and essential for good fruit-set. A sheltered site will also encourage pollinating insects.*

LIGHT LEVELS

Sunlight enables photosynthesis (the method by which plant food is made) to occur. As a result, it is vital to new growth a well as sustaining existing growth. Seasonal changes in light levels may also trigger different stages in the plants' development.

AIR CIRCULATION

Even moderate winds can increase a plant's transpiration rate. However, a light wind can also have beneficial effects, providing relief from extreme heat and cooling the foliage. It will also "change the air" around the plant and thus help to alleviate a stagnant atmosphere that could promote disease.

ASSESSING THE CLIMATE

A range of climatic information is generally available for your local region. Weather forecasts are an obvious source, but long-term records are sometimes available. From these you should be able to get a picture of the average rainfall, snow, frosty days, wind direction and monthly temperatures.

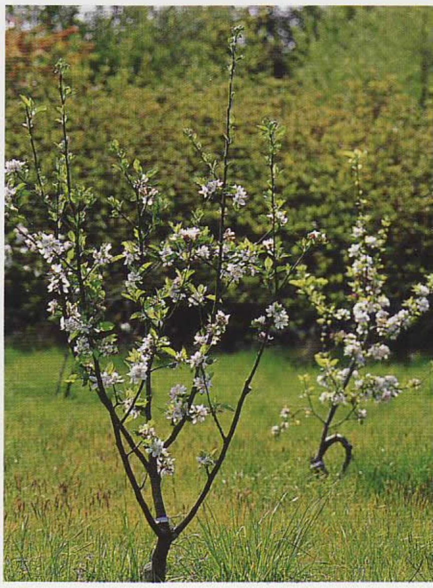

A TYPICAL GARDEN AND ITS MICROCLIMATES

Even a small garden can contain a wide variety of microclimates, ranging from cool, shady corners to hot, dry areas. A proper assessment of the growing conditions that prevail in your garden allows you to choose the right plants for the right place.

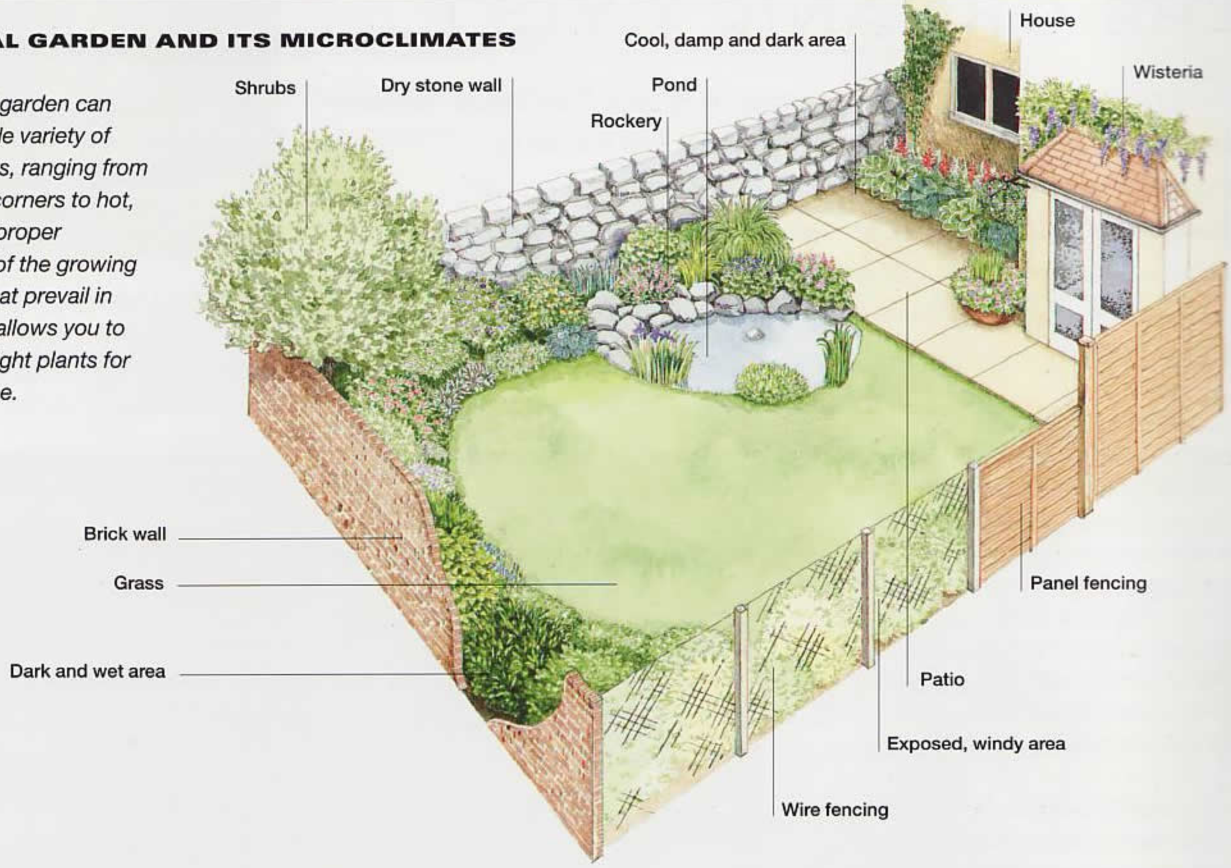

A GARDEN'S MICROCLIMATE

The prevailing weather conditions in confined or discrete areas are termed the microclimate. For example, there will be a range of environments within a small domestic garden that may be very different from those weather conditions prevailing elsewhere in the locality. Observe what is going on in your own garden and keep a careful record of the position and influence of the following features.

Walls, hedges and fences may cause wind turbulence and dryness, depending upon their height and density. They may cause cold air to build up on sloping sites by preventing it from "draining away" downwards. This can create frost pockets that may have a crucial effect on the type and range of plants that may be grown.

Shadow projections from permanent features, such as buildings, walls, hedges and large trees, may create a range of dense shadow resulting in dark, dry and often cool conditions that will only suit certain plant species.

Hot spots may occur with aspects that are subject to high and prolonged levels of sunlight. Plants may struggle in the dryness created by such strong sunlight.

Damp areas may be anticipated where the ground is low-lying, especially on very heavy soils. During heavy rainfall you should also note the direction of water flow, including where necessary the route taken by excess surface water, which may in extreme cases be the cause of soil erosion.

Soil type and pH play an important role in the ability of a site to support particular plant species. The relative acidity or alkalinity can affect the availability of certain essential nutrients, leading to either shortage or toxicity in plants not ideally suited to the pH of a soil. It is always better to choose plants that thrive in the particular pH of a soil. Attempts to alter the pH by acidifying or adding lime are rarely satisfactory and neither are long-term solutions.

Existing habitat types on and around the garden indicate what plants are likely to thrive. Notes should be made on the range of species and the likely wildlife value.

The garden topography (the shape and aspect of the land) may also affect what you are able to do. If a garden slopes toward the direction of the midday sun, for instance, it will be warmer than a flat site. Sloping or uneven gardens have their own problems, particularly if you wish to site a greenhouse.

Right: *A shady corner of the garden provides the perfect place for shade lovers such as these foxgloves (*Digitalis*).*

RIGHT PLANT, RIGHT PLACE

Once you have assessed your garden environment, you will be able to choose the right plants for your site. Every garden has its own unique set of advantages and limitations, and these take time to evaluate, so trial and error is ultimately the most reliable method.

THE ORGANIC CYCLE

Plants produce their own food using carbon dioxide from the air, water from the soil and energy from the sun in a process known as photosynthesis. The energy in this sugar powers all the growth, development and life-giving processes within the plant. Plants must also produce a range of other substances, including proteins, fats, oils and cellulose, for which they need nutrients such as nitrogen, phosphorus, and potassium. These are almost always taken up through the roots.

Above: *Plant leaves are remarkable structures because they are the factories in which a plant makes its food.*

WHAT PLANTS NEED

In natural systems, nitrogen, phosphorus and potassium are repeatedly taken up by the plant, used and returned to the soil when the plant dies, drops its leaves or is eaten by an animal. This process is known as nutrient cycling. Plant nutrients are covered in more depth later. For now we need only look at three of the most important cycles: water, carbon and nitrogen.

THE WATER CYCLE

The movement and endless recycling of water between the atmosphere, the surface of the land and under the soil is called the water cycle and is driven by the energy of the sun and the force of gravity.

Water vapour in the atmosphere condenses into clouds, which fall as snow, rain, sleet or hail. This water may be taken up by plants, stored in lakes, enter the soil or flow over the surface in streams. The sun causes water to evaporate back into the atmosphere, or gravity may pull it down through the pores of the soil to be stored as slow-moving ground water. Water can also return to the atmosphere indirectly through plants' leaves – a process known as transpiration – this being highest during periods of high temperatures, wind, dry air and sunshine.

Above: *Plants rely on water, taken up through the roots, to support their life processes and facilitate growth.*

Right: *Left to develop naturally, grassland will become a rich and diverse habitat for a wealth of both plants and wildlife.*

THE CARBON CYCLE

The movement of carbon, in its many forms, between the total living content of the earth, the atmosphere, oceans, and the rocks and soils covering the surface, is termed the carbon cycle. It is complex and far reaching. The same carbon atoms in your body today have been used in countless other molecules since time began. Plants absorb carbon dioxide from the atmosphere during photosynthesis (food production) and release it back into the atmosphere during respiration (food use). This process is the great natural recycler of carbon atoms.

THE NITROGEN CYCLE

Nitrogen is used by living organisms to produce a number of complex organic molecules such as amino acids, proteins and nucleic acids, the "building blocks" of life. The largest store of nitrogen is in the atmosphere, where it exists as a gas. This store is about one million times larger than the total nitrogen contained in living organisms.

NUTRIENT CYCLES IN A TYPICAL GARDEN

Even small gardens contain natural cycles. Elements and nutrients are cycled repeatedly both within the environment and as a result of the complex food chains and webs. This cycle of nature is essential in creating a balanced and healthy habitat.

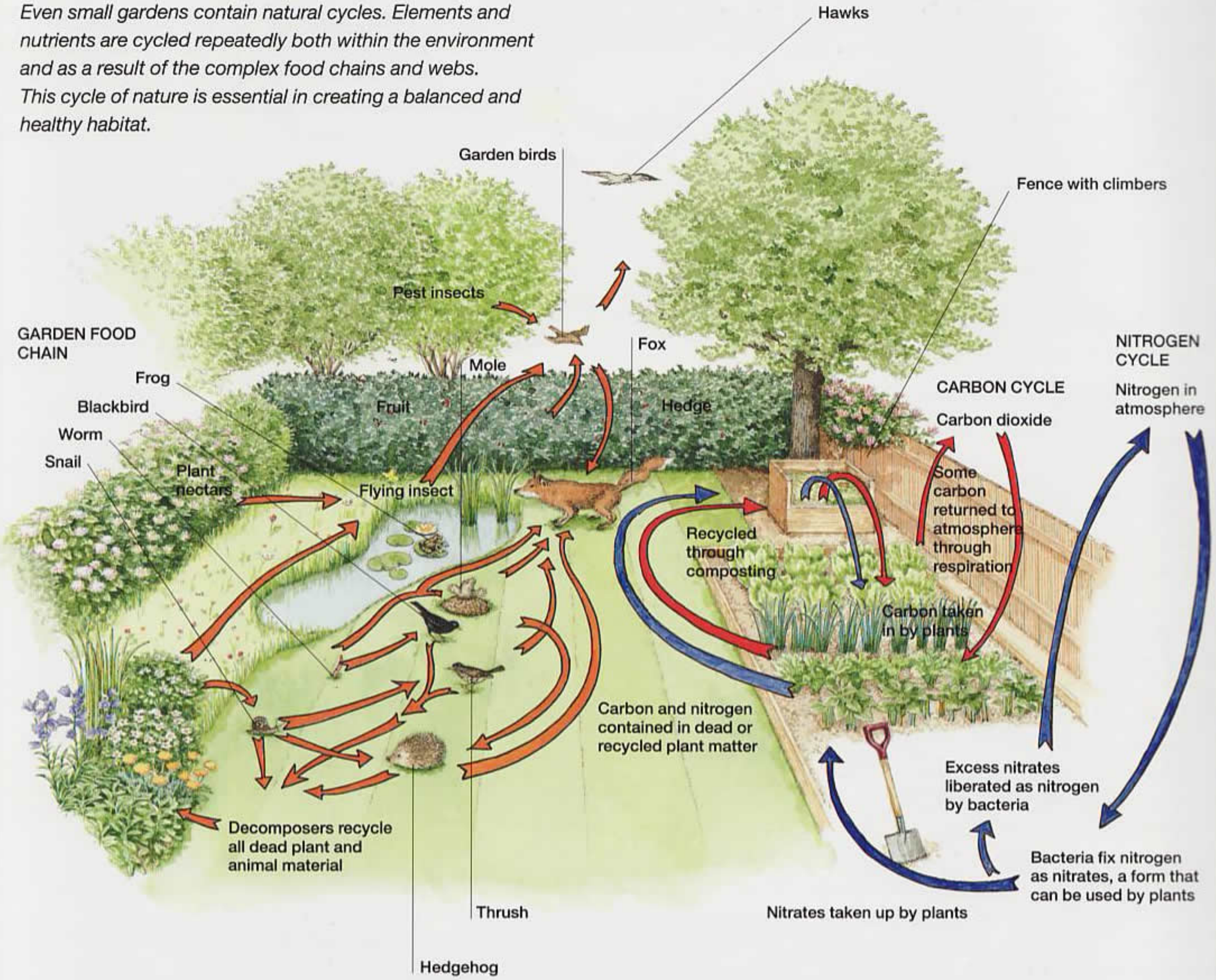
Hawks
Garden birds
Fence with climbers
Pest insects
GARDEN FOOD CHAIN
Fox
NITROGEN CYCLE
Mole
Frog
CARBON CYCLE
Nitrogen in atmosphere
Fruit
Hedge
Blackbird
Carbon dioxide
Worm
Snail
Plant nectars
Flying insect
Some carbon returned to atmosphere through respiration
Recycled through composting
Carbon taken in by plants
Carbon and nitrogen contained in dead or recycled plant matter
Excess nitrates liberated as nitrogen by bacteria
Decomposers recycle all dead plant and animal material
Bacteria fix nitrogen as nitrates, a form that can be used by plants
Thrush
Nitrates taken up by plants
Hedgehog

Despite its abundance in the atmosphere, plants can only take up nitrogen in two solid forms: ammonium and nitrate. Most plants fulfil their needs using nitrate from the soil. Ammonium is used less frequently because it is toxic in large concentrations. Most ecosystems have nitrogen stored in living and dead organic matter, which re-enters the cycle via decomposition. Decomposers in the soil, such as bacteria and fungi, chemically modify the nitrogen found in this organic matter.

Almost all of the nitrogen found in any terrestrial ecosystem originally came from the atmosphere and is biochemically converted ("fixed") into a useful form within the soil by specialized bacteria. Members of the bean family (legumes) and some other kinds of plants form relationships with nitrogen-fixing bacteria. In exchange for nitrogen, the bacteria receive food from the plants and special root structures (nodules) that provide protection.

WHAT ARE ECOSYSTEMS?

Short for ecological systems, ecosystems are communities of plants and animals that consist of a given habitat (the place where an organism lives) and its community (all the plants and animals that live in it). Any group of living and non-living things interacting with each other (including those found in gardens) can be considered an ecosystem.

The chemical materials that are extracted from the environment and changed into living tissue by plants and animals are continually recycled within the ecosystem. These nutrient cycles are at their most efficient when an ecosystem has a good diversity of species. Such diversity tends to make a community stable and self-perpetuating.

FOOD CHAINS AND WEBS

The energy necessary for life reaches the earth as sunlight. This is made into food and stored in plants. Animals acquire some of the stored energy by eating the plants or by eating other plant-eating animals. Such sequences, called food chains, overlap at many points, forming food webs which show "who eats whom" in an ecosystem. If one animal or plant is affected, then the entire food web can be changed.

SOIL AND SOIL MANAGEMENT

Soil is the most precious resource in your garden. Some inherit a well-tended soil, while others, particularly those moving into new homes, inherit a rubble-filled mass. However, any soil can be improved through time and effort. If you regard your soil as a living entity, you will see that essential plant nutrients are cycled by a microscopic army of inhabitants and larger worms, insects and grubs. All these creatures need air, moisture and food. Using manure, garden compost and other sources of organic matter is the key to sustaining this soil life and keeping the soil healthy.

Left: *The key to developing a successful organic garden lies in the careful management of the soil.*

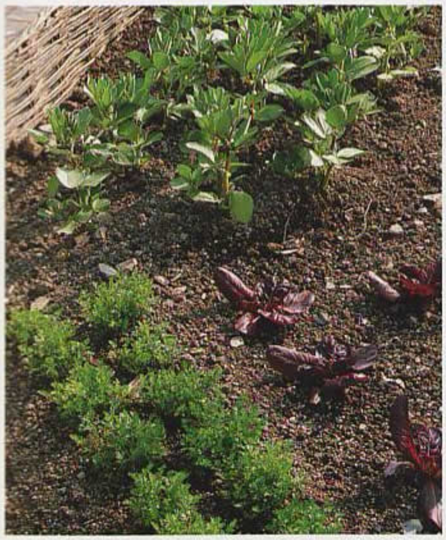

Above: *Newly planted crops must be kept weed free so that water and nutrients are freely available.*

Above: *Companion plants can help to protect crops naturally from pests, thus avoiding damage to the soil ecology.*

Above: *Crops that are grown in rows must be rotated on a regular basis in order to avoid soil-borne pests and diseases.*

TYPES OF SOIL

Soil is probably the most important constituent of any organic garden because it is vital for successful plant growth. Understanding your soil and knowing how this can help you to create a healthy, fertile growing environment for your plants should be a priority. The starting point in this process is an understanding of the different soil types and how they affect the plants you can grow. Once you understand the soil in your own garden, then you will be able to create a successful organic garden.

Above: *All soil types benefit greatly from the addition of organic matter such as this green manure.*

WHAT ARE SOILS MADE OF?

Both natural soils and some potting mixes for container-grown plants usually have five main components. These are mineral particles (the inorganic fraction), organic matter (the remains of living organisms), water (the "soil solution"), air (which fills the spaces between solid particles that are not filled with water) and living organisms. The proportions of these components vary widely according to the soil type, or the growing medium.

The proportions of water, air and organic matter can be readily changed by soil cultivation and other horticultural practices. A good topsoil will continuously supply plant roots with water, air and nutrients. Subsoil (the largely inert soil layer that lies beneath the thin layer of "living" topsoil) has less organic matter than topsoil. Plant growth will suffer when the proportions of water and air in the medium are out of balance. Too much air will have the same effect upon plant roots as a drought, whereas too much water causes waterlogging. Growth will also suffer if nutrient levels are too low, too high or if there is an imbalance in supply between different nutrients. In addition, soil that is too compacted for roots to grow will also adversely affect growth.

SAND AND SILT

Sandy and silty soils originate from river deposits, windblown sediments or from the erosion of sandstone outcrops. Their general properties are that they do not provide or retain plant nutrients; they are not cohesive and therefore possess a weak structure; and in most cases they are free draining.

Silts, unless well structured, will be waterlogged. Sand, on the other hand, is naturally free draining. The structure and texture of sandy soils means that they are only able to hold a very small reserve of water. Organic matter can improve the available water content but the real trick in avoiding drought-stress lies in ensuring that plants root deeply into the soil. Nutrient shortage can also be a problem on sandy soils which have a tendency to become acidic over time. Liming and the regular addition of organic matter will help alleviate these problems.

Sandy soils do have the advantage of "warming up" quickly in spring due to their lower water content and are easier to work early in the year. This means that you can grow a wider range of plants in a sandy soil. Planting or transplanting is also easier in the autumn.

Organic material is broken down very quickly in sandy soils due to good aeration and a temperature that favours rapid bacterial action. Many light soils naturally have less than two per cent organic matter and it is vital that organic matter be added regularly to sustain healthy plant growth.

COMMON TYPES OF SOIL

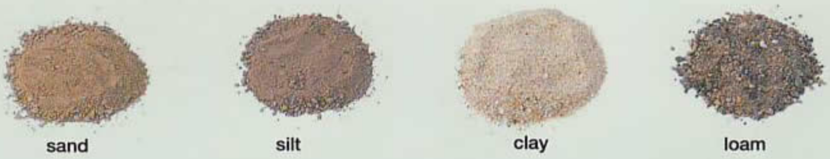

Sand A free-draining soil that is quick to warm up, but hungry and thirsty.
Silt River deposits can be sticky, but not as sticky as clay. Rich and easy to work.

Clay A heavy soil that is often difficult to work. Slow to warm up, but fairly rich.
Loam A moisture-retentive soil that warms up quickly and works perfectly.

PLANTS SUITABLE FOR DIFFERENT SOIL TYPES

Many plants have a type of soil in which they grow best. The list below gives a few examples of the soil preferences of some common species.

Plants suitable for acid soils
Rhododendron (all species)
Camellia (all cultivars)

Plants suitable for alkaline soils
Butterfly bush (*Buddleja* spp.)
Sage (*Salvia*)

Plants suitable for hot dry soils in sun
Yucca (all species)
Broom (*Cytisus*)
Lavender (*Lavandula*)

Plants suitable for damp or wet soils
Willow (*Salix*)
Dogwood (*Cornus*)
Plantain lily (*Hosta*)

Plants suitable for heavy clay
Weigela (all cultivars)
Crocosmia (all cultivars)
Daffodil (*Narcissus*)

Plants suitable for dry shaded soils
Flowering currant (*Ribes sanguineum*)
Pachysandra terminalis
Elephant's ear (*Bergenia cordifolia*)

HAND-TEXTURING METHOD

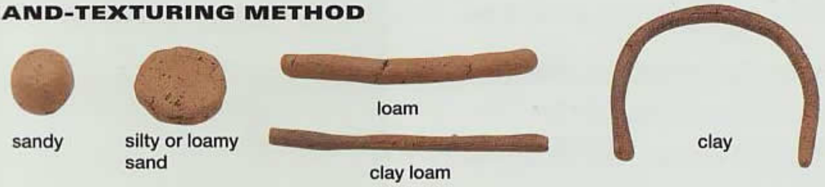

Finding out what type of soil you have is easy and does not require any specialist equipment. This test is ultimately a matter of judgement and will only give you a relative picture of the sort of soil you have. However, it is surprisingly accurate. Simply take a small amount of soil – about a teaspoonful will do – in the palm of your hand. Moisten with a little water (not too much but enough to make it just workable). Once moistened, try to form the soil into one of the shapes shown above.

1 Begin by forming a ball. If it stays together, then proceed to the next shape. If it does not form a ball, then you have a sandy soil.
2 If you can flatten the ball without it breaking up, then you have a silty sand or a loamy sand.
3 If you can roll the flattened ball into a thick sausage shape, then you have a loam.
4 A soil that can be rolled into a thin "sausage" is a clay loam.
5 If you can bend the soil into a horseshoe or ring shape, then you have a clay soil.

Above: *Plants such as alliums and eryngiums like well-drained soil and so will thrive together in the same site.*

CLAY

The particles found in clay are extremely small and are able to interact with, and directly affect, the chemistry of the soil. The individual clay particles are so tiny that they are actually bonded together by electrical charges which produce the characteristic plasticity of this type of soil. Clay is both water retentive and rich in nutrients. It has few pore spaces and those that are there have a tendency to become waterlogged. Clay is prone to swelling when waterlogged and shrinkage when dry. As a result, clays heave (swell outward and upward) when wet and crack when dry. They can also be subject to frost action which causes an increase in tiny, almost microscopic, airspaces (micro pores). Pure clay soils are rare, although some soils may be very rich in clay. They have the potential to be extremely fertile soils if they are well managed.

LOAMS

Loams are a mixture of sand, silt and clay that results in a blend of the characteristics of each constituent part. They are usually characterized by their clay content. Heavy loams are about 24–30% clay, whereas light loams contain about 12–18% clay. Heavy loams behave and should be treated like clay soils. Light loams should be treated like sandy soils. Medium loams are potentially the ideal mixture, exhibiting the advantages of both heavy and light soils without many of the disadvantages of either.

HUMUS

This is a stable form of partially decomposed plant material that gives topsoil its characteristically dark colour. Humus has a high nutrient-reserve potential (2½ x better than clay), retains water, increases the friability (suitability for cultivation) of heavy soils and its darker colour encourages soils to warm up more rapidly.

SOIL PROFILE

A typical soil profile usually consists of three main elements: an upper layer of dark, fertile topsoil; a middle layer of lighter, infertile subsoil; and a lower layer of bedrock, which ranges from a few to hundreds of metres (yards) deep.

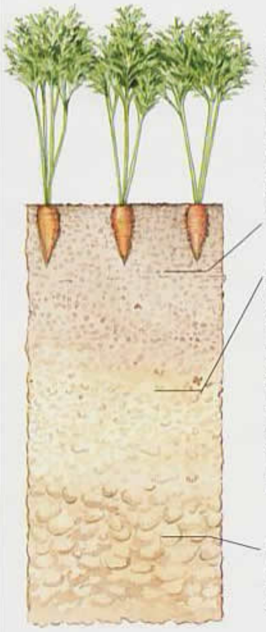

Topsoil is the dark layer of soil that contains organic material. Keep as deep as possible, although usually only one spit (spade) deep.

Subsoil is usually lighter in colour and contains little or no organic material or nutrients. It should be broken to one spit depth, but not mixed with the topsoil.

Bedrock is usually below the level of cultivation.

MANAGING YOUR SOIL

Any soil has the potential to be a fertile growing medium in which plants will thrive if it is managed correctly.

Management of light/sandy soils

- Maintain levels of organic matter.
- Lime when required. (N.B. It is often easier to select species tolerant of the site conditions in the long term.)
- Fertilize well as light soils are usually infertile. (N.B. Some exotic and native plant species actually thrive under relatively poor nutrient conditions.)
- Irrigate frequently in the first year after planting, less so once deeper roots are established.

Management of clay soils

- Drainage is essential.
- Maintain lime status.
- Maintain levels of organic matter and fertilize when needed.
- Wise cultivation is needed to conserve winter or summer tilth produced by natural weathering, i.e. cracking due to wetting/drying cycle and breakdown by frost action.

SOIL STRUCTURE

Soils are made up of mineral and rock particles that give them their natural properties. The character of the mineral fraction cannot be changed but the way that these building blocks are organized within the soil can. Creating and maintaining soil structure is a key aspect of maintaining the fertility of the soil and supporting healthy plant growth. Even the most difficult soils can be modified to create a fertile growing medium for plants.

Above: *Even "difficult" soils can be transformed into a good growing medium by careful cultivation.*

WHAT IS MEANT BY SOIL STRUCTURE?

Soil structure is quite simply the way soils are organized. To use a simple analogy: if we see the soil particles (sand, silt or clay – the mineral fraction) as the building materials – like the bricks and mortar of a house, for instance – then the structure is the architecture. The building materials themselves cannot be altered, but the architecture can vary considerably from home to home. This is the key. It is possible to improve a soil's structure even if the mineral fraction remains unchanged. It is perfectly feasible to have a well-structured clay soil for instance. It all depends upon how well you manage it.

WHY IS SOIL STRUCTURE IMPORTANT FOR PLANT GROWTH?

In order to survive, plant roots need water, nutrients and air. Plant roots breathe in the soil, taking in oxygen and expelling carbon dioxide. A poorly structured soil may have too little airspace. This can have two negative effects. Fresh air (containing oxygen) may not be able to penetrate the soil very easily and the soil will tend to become waterlogged following rain. Waterlogging will also reduce the air available in the soil.

Below: *A rotavator is a useful and labour-reducing method of breaking up the soil to prepare it for planting.*

Structure is therefore very important in terms of maintaining the health of plant roots. Plant shoot development is directly linked to root health. A healthy root system ensures healthy shoots and will ultimately result in better growth and crop yields.

HOW DOES SOIL STRUCTURE DEVELOP IN NATURAL SYSTEMS?

Soil structure naturally develops in soils through the effect of weather cycles. Wetting causes soils to swell while drying causes shrinkage. This naturally causes soils to crack. The action of freezing and thawing is also important on clay soils in areas where frosts occur. They form particles known commonly as crumbs. The "crumb structure" of any soil develops over time and is important in terms of allowing the free passage of water through the soil and air that must be able to move in and out of the soil. In this way soils do not become habitually waterlogged and natural nutrient cycles – oxygen, carbon and nitrogen, for example – are not impeded. Plant roots and soil organisms, such as worms, naturally help to maintain soil structure, as does the natural addition of decaying plant material each autumn.

HOW TO ALTER SOIL STRUCTURE

Humans have learned over time that even an infertile area can be worked – and the structure of the soil improved – by cultivation. This is covered in more detail later but it is essential to understand now that cultivation is a way of rapidly accelerating the natural cycles that promote good structure. Digging and breaking down "clods" helps to introduce air and creates new pores in the soil. Adding organic matter helps to maintain these pore spaces, retains moisture (in dry soils) and encourages the action of soil-dwelling creatures such as worms. No-dig

Above: *If you need to work on wet soil, work from a plank of wood to ensure the soil is not compacted and its structure destroyed.*

SECTION THROUGH A LAND DRAIN

This cross-section of a typical land drain shows the various layers of materials from which it is constructed, including the topsoil, an inverted turf or perforated plastic membrane, gravel, peabeach or crushed stone, and the land-drain pipe.

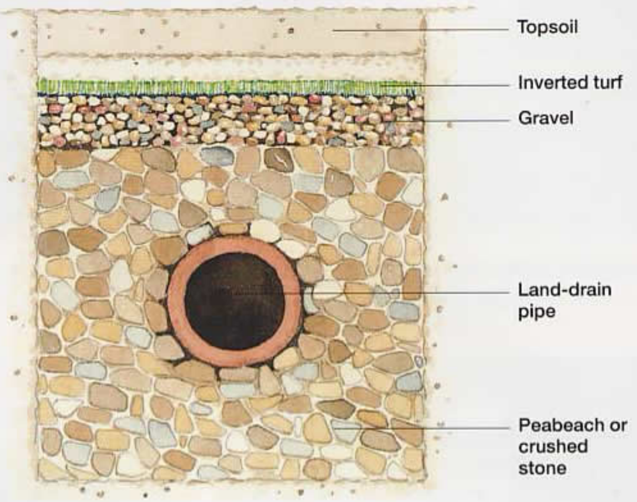

systems aim to harness this natural cycle more closely, but the goal is the same – a rich, fertile and well-structured soil.

IMPROVING DRAINAGE

Soils that tend to become waterlogged often do so as a result of poor structure and because there is nowhere for the water to go. The vast majority of soils do not actually need (or benefit from) the installation of land drains. Often it is enough to dig over a site because much of the drainage problem may be due to surface compaction. Surface compaction severely reduces the passage of water into the soil and can result in a sticky surface that is prone to "puddling". Installing land drains is usually only done where the water table is very near to the surface or the consequences of flooding are severe (if they threaten housing or use of the area for instance). Land drainage using pipes or "tiles" is a complex undertaking and may be best left to a contractor.

If you have a sloping site you can dig a "soak away" drain at the base of the slope and improve the structure of the soil to encourage the rapid dispersal of water following rain. If you have a fairly flat site, then you should consider making a raised bed for growing crops. Raised beds are easy to manage and avoid the inconvenience of installing (and maintaining) a drainage system.

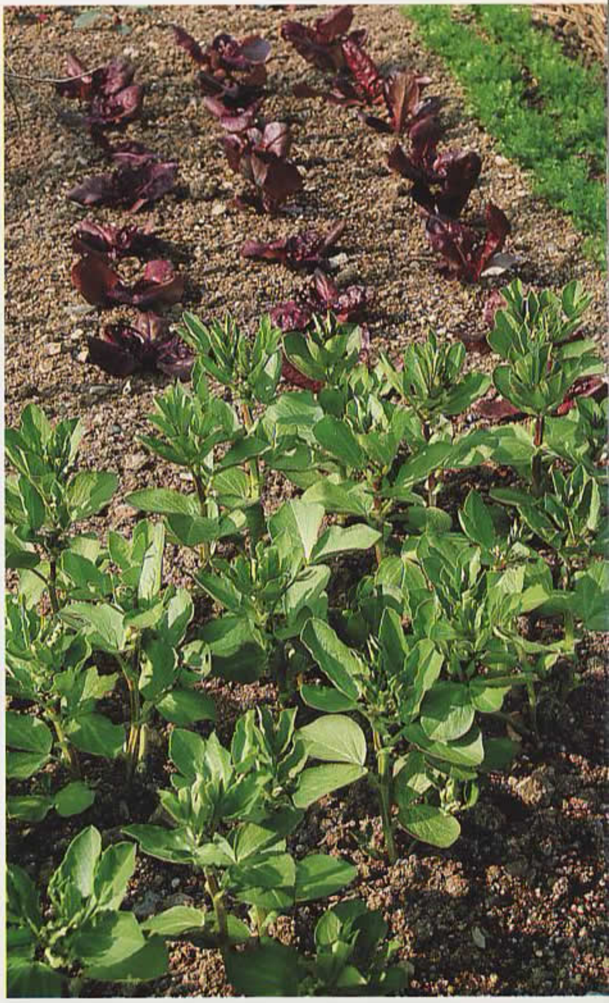

Above: *A well-prepared soil that is kept free of weeds and dug regularly will result in healthier, more vigorous crops throughout the growing season. A healthier crop will be more resistant to pest and disease attack.*

IMPROVING SOIL STRUCTURE

1 *One of the best ways to improve the structure of the soil is to add as much organic material as you can, preferably when the soil is dug. For heavy soils, this is best done in the autumn.*

2 *If the soil has already been dug, then well-rotted organic material can be worked into the surface of the soil with a fork. The worms will complete the task of working it into the soil.*

SOIL CHEMISTRY

While for many of us the word chemistry can be extremely off-putting, the truth of the matter is that everything in life is made up of chemicals. While the pure science of chemistry can be very complex, the chemistry of soil that organic gardeners must master is mercifully simple. A rudimentary knowledge of the chemicals that affect plant growth and how soil chemistry can be controlled and managed is all you need to grow healthy plants in an organic garden.

Above: *You can reduce the acidity of the soil by adding lime some weeks before planting. Test the soil first to see how much is needed.*

OXYGEN

The amount of oxygen in a healthy soil controls the type of life it will support. Nearly all organisms need oxygen to survive. Soils without oxygen are described as anaerobic. Most organisms can survive for short periods under anaerobic conditions, but this causes the accumulation of poisons that can become toxic at high concentrations.

A typical soil has about 50 per cent of its pore space filled by air and 50 per cent by water. Only certain bacteria can remain in anaerobic conditions for long periods of time, although some species of bacteria can readily switch from oxygen-rich to oxygen-poor conditions quickly to adapt to local conditions. Microbes use about 70 per cent of the oxygen in the soil and plant roots use the remaining 30 per cent. Under anaerobic conditions, the efficiency of microbes is poor and decomposition rates are much slower.

WATER

Soil water is vital for all soil life. Without it, microbes cannot grow or remain active and many will go into "hibernation" until water returns. Fungi, on the other hand, are more resistant to water stress than bacteria. With too much water, oxygen levels drop and the lack of air tends to slow down the nutrient cycles driven by microbes. Water is also the medium by which essential nutrients are able to enter the plant.

SOIL pH

The pH scale is an abbreviated form of "Potential of Hydrogen". It is a measure of the degree of acidity or the alkalinity of a solution as measured on a scale (pH scale) of 0 to 14. The midpoint of 7.0 on the pH scale represents neutrality. A "neutral" solution is, therefore, neither acid nor alkaline. Numbers below 7.0 indicate acidity; numbers greater than 7.0 indicate alkalinity.

The level of acidity or alkalinity (pH) of a soil can significantly affect the nutrient availability. Many nutrients become "unavailable" to plants when the soil is either too acid or too alkaline. Microbial activity in soil is also largely controlled by pH. Fungi tend to predominate in acid soils, bacteria in neutral or alkaline soils.

Soil pH is essentially a measure of the acidity of the soil water, although the soil itself is the deciding factor in respect of what this will be.

Most plants prefer or are tolerant of a specific pH range. Some plants, such as the hydrangea, exhibit a different flower colour depending upon the prevailing pH. Most garden plants, especially vegetables, thrive within a range of 6–7 which happens to be where the majority of nutrients are available. It is best to maintain this pH in order to optimize the availability of nutrients. Many garden plants, however, are not too fussy about the pH levels, so if you choose plants carefully, it will not usually be necessary to alter the soil acidity.

ACIDIFYING SOIL

Lowering the pH of a naturally limy soil is difficult because the soil often contains a reserve of calcium that is released immediately upon acidification. Lowering the pH involves the use of flowers of sulphur and is only usually successful over a short period of time.

TAKING A pH TEST

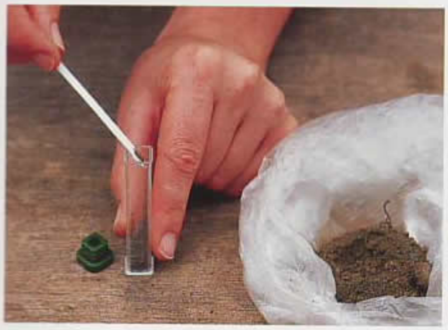

1 *Place the soil in a test tube until it reaches the mark on the side. For the most accurate results, dry the sample first, grind it into a powder and ensure it is free from stones.*

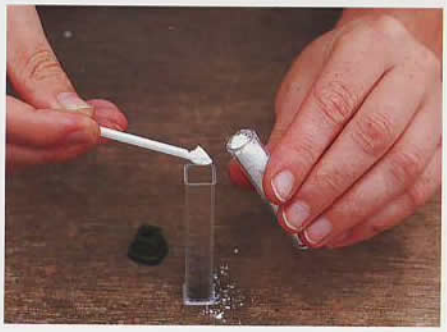

2 *Put a layer of barium sulphate powder into the tube level with the mark. This compound helps the solution to clear rapidly and makes the pH reading clearer.*

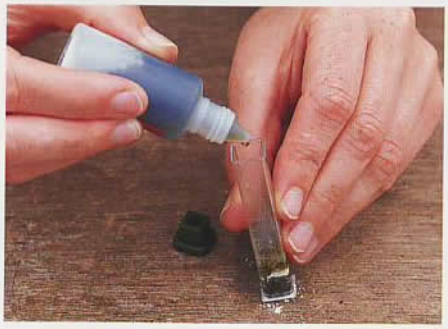

3 *Pour in a little of the indicator solution up to the mark shown on the tube. Be careful not to put in too much because this can make the solution dark and difficult to read.*

MAKING A NUTRIENT TEST

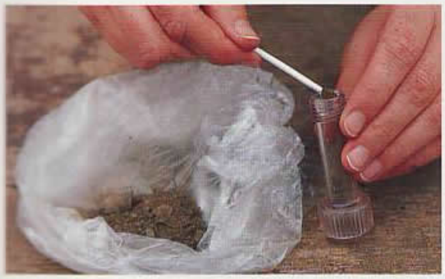

1 Place a small sample of the soil into the test tube up to the mark on the side.

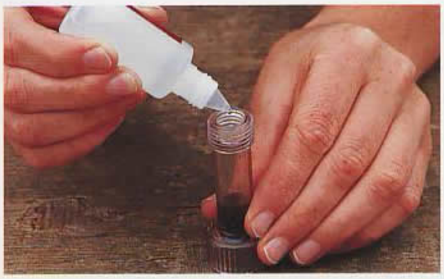

2 Add a test solution (in this case one for nitrogen) up to the mark on the test tube.

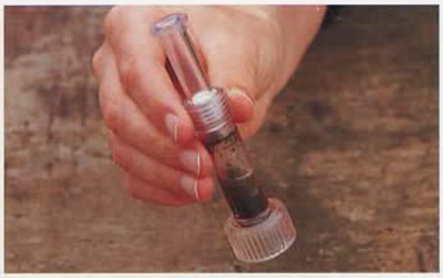

3 Filter the solution to remove soil particles and leave just a liquid solution.

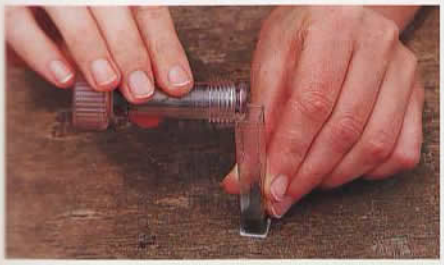

4 Decant the resulting filtered solution into another container for the final stage of the nutrient test.

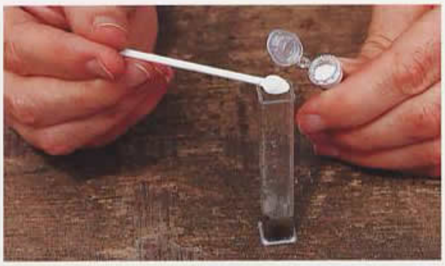

5 Add a small amount of indicator powder. This will react with the solution and enable a colour reading to be taken.

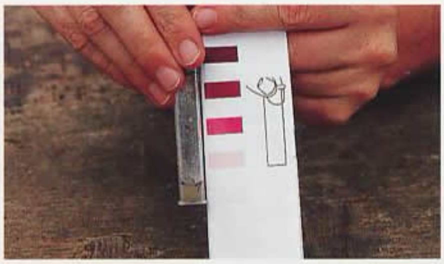

6 Shake for about 10 seconds and compare with the chart. Here, the low reading indicates that a nitrogen-rich fertilizer will benefit this soil.

LIMING SOIL

It is generally easier to raise the soil pH than to lower it. Lime neutralizes soil acidity and is commonly applied as ground limestone, chalk or dolomitic limestone (dolodust). Lime requirement cannot be determined from soil pH because it is influenced by soil texture and organic matter content. Clay and humus act as a "buffer" because of their complex chemistry. If soil is known to be acidic, regular light application is preferable to heavier, more infrequent, doses.

HIGHLY ALKALINE SOILS

Soils that are too alkaline suffer trace element deficiencies of manganese, copper, iron, zinc and boron. Phosphates are also less available, their maximum availability being between pH 6–7. Disease organisms can be more of a problem in calcium-rich soils, as many disease-causing fungal agents prefer alkaline conditions. Some plants, such as rhododendrons, are intolerant of high pH and only grow on acid soils, while others, like helianthemums, thrive in highly alkaline soils.

4 Add distilled water to the mark on the tube and shake the container vigorously for about a minute. Ensure the contents are mixed thoroughly and leave to settle.

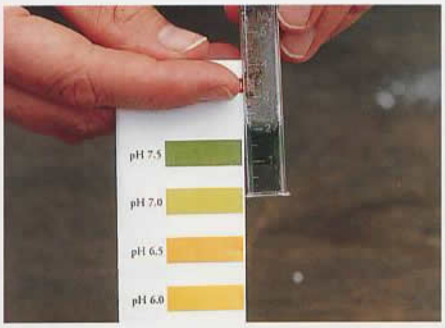

5 Once sufficiently cleared, compare the colour against those on the chart, choosing the one that most closely matches that of the solution.

GROWTH RESTRICTIONS IN EXTREMES OF PH

Many vital nutrients that are essential for healthy plant growth become unavailable in extremes of soil pH.

Nitrogen deficiency Most nitrates are released from organic matter and a low soil pH limits the rate of decomposition severely.

Phosphate deficiency Phosphate becomes unavailable outside the 6.5–7.5 pH range. Some plants form relationships with soil-borne fungi that release phosphates in acid conditions.

Trace-element toxicity and deficiency Trace elements, especially aluminium, iron and manganese, are generally more soluble in acidic conditions. Extreme acidity can lead to excessive quantities of trace elements and to plant death. Other trace elements, such as copper, boron and molybdenum, become less available at low soil pH. Molybdenum deficiency affects legumes, which will not grow in acid soils.

DIGGING AND CULTIVATION

Every organic gardener's goal should be a healthy soil to support plant growth and development. Digging is one of the commonest ways to create a rich, fertile and ultimately productive soil. It can be hard work at the outset, particularly on sites not cultivated for some time, but the results – in the form of healthier, more productive plants – are worth it. Quite simply, what you get out of your soil will depend upon what you put into it.

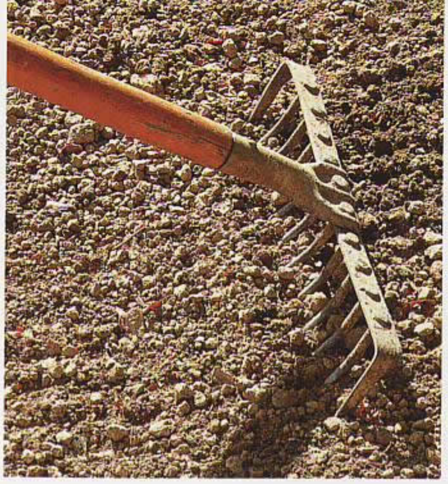

Above: *After a winter exposed to the weather, most soils can be easily broken into a fine tilth with a rake.*

THE BENEFITS OF DIGGING

Winter is the commonest time to dig, but soil can be dug at any time of the year if the conditions are right. Avoid working the soil when it is too dry and impenetrable, or too wet and sticking to your tools and boots. Clay soils may be best dug in mid- to late autumn to allow the action of frost to make the soil more suitable for final cultivation. Lighter soils are best dug in the spring or immediately prior to planting the site. Done properly, digging increases the amount of air space in the soil, which in turn benefits soil-dwelling organisms and plant roots due to the increase in oxygen available. It also lets you add organic matter that will feed these vital denizens of the soil and aid nutrient cycling. Calculate how much organic matter you will need before you start. You should aim to add about 30 per cent of the volume cultivated. A 20m^2 (220ft^2) plot cultivated to one spade's depth will need 2.5 cubic metres (88 cubic feet) of manure or garden compost.

SINGLE DIGGING

This method involves digging down to the depth of one spade (this is called a spit). Single digging suffices on light free-draining soil as long as there is no layer of compacted stones in the topsoil. Much of the procedure is the same as for double digging, but the subsoil is left undisturbed. Mark out the position of the bed. Remove any turf or vegetation from the surface, put to one side and then bury it in the bottom of the bed as you proceed. Work across the bed, digging out a trench that is two spits wide and a single spit deep, and place the soil to one side. Remove stones and perennial weeds as you go. Fork well-rotted manure or compost into the trench. Begin digging a new trench behind the first. Throw the soil forward into the first trench, burying the organic matter. Repeat this process of trenching down the bed until the last trench has been dug. Add organic matter to the base and fill the trench with the soil dug from the first trench.

DOUBLE DIGGING

This is a method for deeply digging the soil, in which the soil is broken up to a depth of approximately 70cm (28in) or more. The method usually involves digging a quantity of soil and setting it aside while aerating and sometimes adding an amendment to the subsoil below, then returning the topsoil. There are several methods of double digging, all of which have their advocates. Two methods for double digging are described here.

First, define the area where the digging will take place, making an estimate of the area to be dug. Dig a trench, 30cm (12in) wide and a spit deep, across half the width of the bed. Place the excavated soil next to the other half, placing it on to a tarpaulin or similar covering if the surface needs protection.

When the trench is complete, fork the subsoil, rocking it back and forth, to loosen it down to a depth of about 30cm (12in). Then, spread a layer of garden compost over

SINGLE DIGGING

Single digging is the simplest method of cultivation. It is well suited to light soils or those soils that have been well worked for a number of years. It simply involves digging and inverting the soil to a single spade's depth, leaving the subsoil below undisturbed.

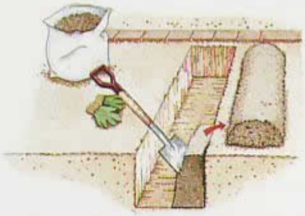

1 *Start by digging a single trench across the width of the plot. Put the soil from the first trench to one side because you will need to use it later in the final trench.*

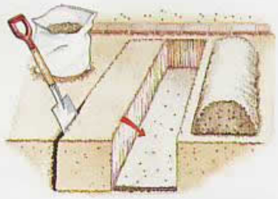

2 *Put a layer of manure in the bottom of the trench. Dig out the next trench and cover over the manure in the first trench with the earth taken from the second trench.*

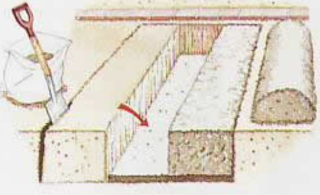

3 *Repeat this process of adding manure to each trench and filling in with earth from the next, breaking up the soil as you go and keeping the surface as even as possible.*

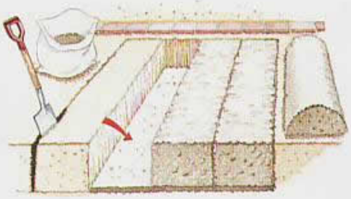

4 *Continue down the length of the plot until you reach the final trench. This should be filled in with the earth taken from the first trench, which was set to one side.*

DOUBLE DIGGING, METHOD ONE

Double digging is a good method for compacted, heavy or poor soils that are in need of rejuvenation. It involves a deep cultivation of both the topsoil and subsoil and is well worth the hard work, producing a wonderful soil in which plants will thrive.

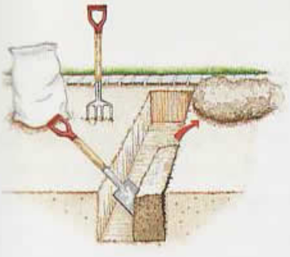

1 *Dig a wide trench, placing the soil to one side to be used later when filling in the final trench.*

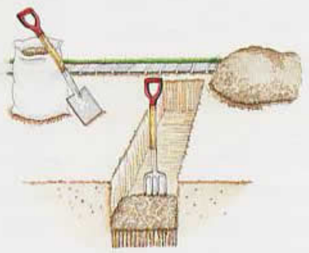

2 *Break up the soil at the bottom of the trench with a fork, adding manure to the soil as you proceed.*

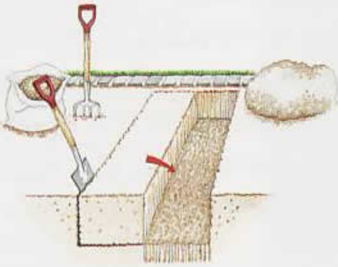

3 *Dig the next trench in the bed, turning the soil over on top of the broken soil in the first trench.*

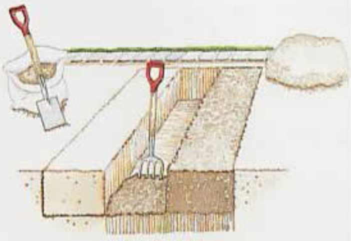

4 *Continue down the plot, ensuring that subsoil from the lower trench is not mixed with topsoil of the upper.*

DOUBLE DIGGING, METHOD TWO

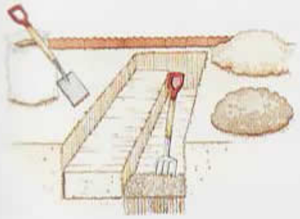

1 *Keeping soil from each level separate, dig the first trench two spits deep and fork over. Dig the second trench one spit deep.*

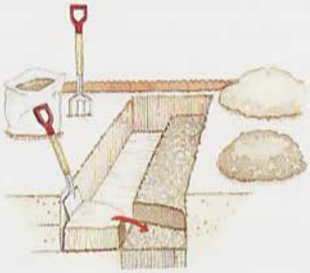

2 *Add organic material to the first double trench and dig the lower spit of the second trench into it.*

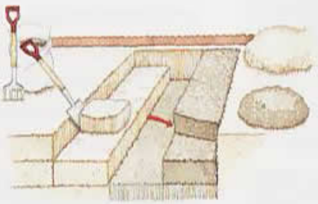

3 *Dig an upper third trench one spit deep, and place the soil on top of that already placed in the first trench.*

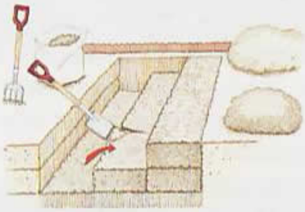

4 *Continue, ensuring that the topsoil and subsoil do not mix. Fill in remaining trenches with soil taken from the first one.*

the exposed subsoil. Move over one spit's width in the bed, and begin to dig out another trench, moving and inverting the excavated soil into the adjacent trench that was just dug. When you get to the end of the bed replace the soil which was set aside.

When the digging is finished, level the site prior to planting. More garden compost (approximately 8cm/3in) can be spread over the surface and forked in at this stage.

You do not have to dig the entire garden all at once. Instead, work on small areas, say, one metre (yard) square, whenever you wish. Once you double-dig a bed, it is important not to walk on it. After all, the whole point of double digging is to loosen the topsoil and the subsoil to a sufficient depth so that the roots of the plants can grow unrestricted and water can readily percolate through the soil. Walking on the soil will simply pack it down again.

There is a second method of double digging that is less commonly used, which enriches the soil to a greater depth.

SECONDARY CULTIVATION

This normally involves the nourishing of the soil after it has been cultivated by digging and usually entails the addition of some form of organic matter during, or immediately after, its completion. Secondary cultivation also describes light digging used to control weeds or incorporate green manures. The tool generally used for this is a fork. Light forking is often the best method to use in areas with permanent plantings of perennial plants or shrubs.

USING A ROTAVATOR

A rotavator is a useful labour-saving device for breaking down dug clods and creating a fine, free-running "tilth" that can easily be levelled for planting or seed-bed preparation.

TIPS FOR DIGGING

Digging can be hard work, but simply following the tips given below will help you to avoid strains and injuries.

- Start with a small area, pace yourself and avoid straining your back.
- Use a spade that is right for you and the job. It should be a little higher than hip height when held vertically.
- Do not try to lift more than you can handle. You will dig for longer periods of time by taking smaller amounts.
- As you work, try to establish a rhythm that you can maintain at all times.
- When doing a lot of heavy digging, scrape the blade clean now and then, and, if necessary, use a file to sharpen the end of the spade, following the original bevel of the blade.

NO-DIG GARDENING

As the name suggests, no-dig gardening is a method of growing plants without cultivating the ground they occupy. While this may seem unnatural to some confirmed tillers of the soil, it is actually the most natural method available. No-dig gardening mimics the natural cycle. After all, plants have grown on the earth's surface for over 400 million years without people digging the soil for them. This approach also has the advantage of being less labour-intensive.

Above: *In nature, organic matter accumulates on the soil surface, breaks down and is worked into the soil by worms.*

ESTABLISHING A NATURAL CYCLE

Essentially no-dig gardening relies upon natural soil processes combined with the action of soil organisms to produce a good soil structure and natural nutrient cycles. There are numerous ways that this can be done, but all rely upon inputs from the top down. A thick top-dressing of organic matter is placed on the surface to be absorbed by the soil itself. Less laborious than digging, the no-dig method does need planning and careful input in order to be successful, although a properly prepared no-dig bed will be every bit as effective as a conventionally cultivated one.

WHAT ARE THE BENEFITS?

Assuming weeds can be controlled, most vegetable crops can be grown with reduced tillage. The key features of a no-dig vegetable-growing system are to keep the soil covered with organic mulch as much of the time as possible and to keep off the soil. The beds are left permanently in place, and there is no cultivation.

Plants grown without tillage use water more efficiently, the water-holding capacity of the soil increases, and water loss from run-off and evaporation are reduced. For crops grown without irrigation in drought-prone soils, this more efficient use of water can translate into higher yields.

In addition, the organic matter in soil and populations of beneficial insects are maintained, as are earthworm populations and microbial activity. Soil and nutrients are less likely to be lost from the ground, less time is required to prepare the soil for planting, and, since there is less bare soil, there will be fewer opportunities for weeds to get established. The mulch of organic matter also helps to keep weeds down.

WHAT ARE THE DISADVANTAGES?

Potential problems associated with the no-dig method are compaction, flooding and poor drainage. However, if the water-holding capacity of a soil improves, no-till systems may produce higher yields. Soil temperatures under organic mulch can be several degrees lower than bare soil or soil under a plastic mulch. Although the lower temperature can be an advantage, especially in the summer when soil temperatures under plastic can be excessive, it can delay crop development.

Left: *Many organic gardeners grow crops such as potatoes quite successfully without ever resorting to digging.*

WHAT ARE THE BENEFITS OF NO-DIG SYSTEMS TO THE ORGANIC GARDENER?

Although no-dig beds may require a little longer to establish initially than cultivated plots, they do offer considerable labour-saving advantages to the organic gardener. They also suffer very few disadvantages when these are weighed against the considerable advantages.

Advantages

- Plants use water more efficiently.
- Better water-holding capacity of the soil.
- Reduced water loss from run-off and evaporation.
- Higher crop yields in areas prone to drought.
- Organic matter in the soil and soil organisms are maintained at higher levels.
- The lower temperature can be an advantage, especially in summer when soil temperatures under plastic can be excessive.
- Soil and nutrients are more easily maintained.
- Less time is required to prepare the soil for planting.
- Fewer weeds become established due to the mulch of organic matter and the lack of bare soil.

Disadvantages

- Soil can become compacted or poorly drained if it is not properly managed.
- Soil temperatures under an organic mulch can be several degrees lower than bare soil, which can delay the development of crops.

MAKING A NO-DIG BED

A bed system which includes permanent paths is essential if digging is to be eliminated from gardening tasks. Planning the area so that all parts of the bed can be reached from the paths means that you can grow your crops closer together (you do not need to leave extra space for access) and you will be able to harvest more from a smaller area.

Mark out the beds on the ground or build simple raised beds with wooden sides. Alternatively, those wishing to make a more permanent edge could consider a low brick wall. By raising the bed it is possible to create a deep bed of good-quality soil that will warm up more quickly than the surrounding soil. A raised bed also means you do not have to bend down so far to carry out routine tasks and the method is tailor-made for disabled gardeners. Raised edging is also useful for attaching supporting hoops, crop covers and cloches. Edges such as these are also ideal for providing barriers for pests. Devices such as water traps or copper paper are easily attached to them to prevent slugs and snails from entering the bed.

Beds should be no more than 1.2m (4ft) wide and paths at least 30cm (12in) wide with every other path double this width to allow for easy wheelbarrow access. Individual beds should be no longer than 3.5m (12ft) otherwise you will waste a lot of time walking around them. The beds can be of any design, from a simple series of parallel rectangles to a more complex design of interlocking shapes. Whatever design you choose, try to follow the same recommended bed and path widths, remembering that narrow angles and very small beds are more difficult to construct and use space less efficiently.

Right: *Green manures are usually dug into the soil, but they can still be used in a no-dig system. The manure can be cut down and left as a mulch.*

PLANTING POTATOES IN A NO-DIG BED

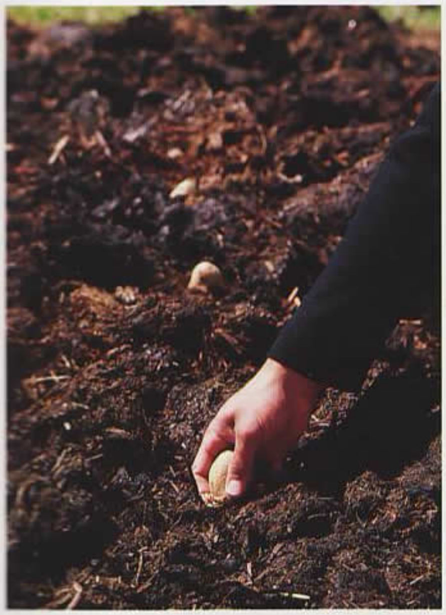

1 *Place a thick layer of well-rotted manure or garden compost on the bed to be planted. This should be quite deep – at least 15–20cm (6–8in) – as it will quickly decompose during the growing season. Early planting is possible by manuring in winter and covering with plastic sheeting to warm the soil. The potatoes are planted directly into this and may be more closely spaced than those planted in soil, perhaps as close as 30cm (12in) apart.*

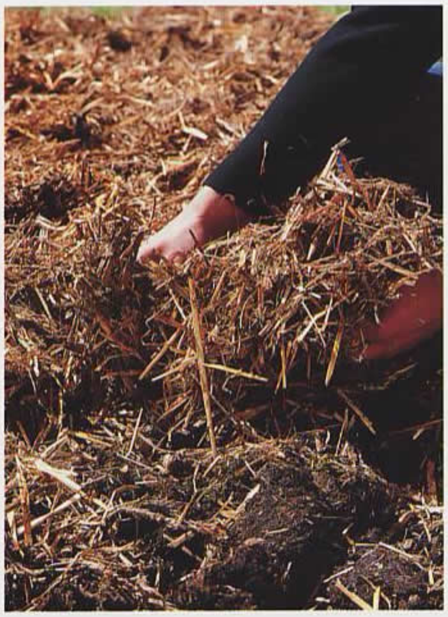

2 *Apply a generous mulch of straw once the bed is planted to give cover to the tubers and to suppress weed growth on the manure or compost. Spread a second layer of composted organic material over the straw. Add another layer of straw and composted material. Water the bed well. Once the potato plants begin to emerge, more straw can be laid down on to the bed, as the material put down earlier in the season will begin to compost down.*

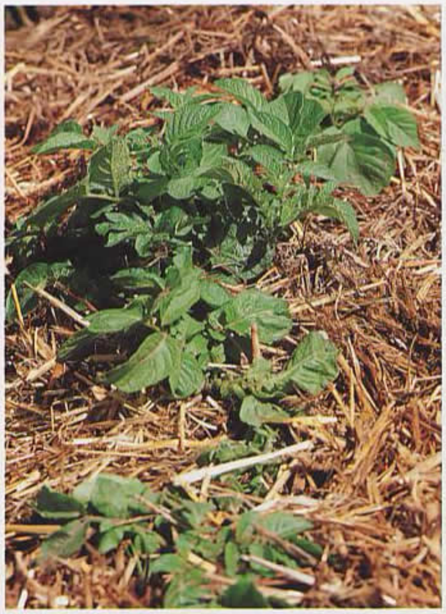

3 *As the potato stems grow, mulch heavily in order to ensure that no light reaches the developing tubers. The potato crop will mature at the same rate as potatoes grown in ordinary soil. The no-dig system is possibly the best method for heavy clay soils that are prone to waterlogging. Potatoes grown in this way are also easy to harvest as there is no digging and the potatoes are cleaner once they are ready for harvesting.*

SOIL CONDITIONERS

There is quite a range of conditioners available to the organic gardener for improving the soil. Some are free (if you do not count the time taken in working and carting them), others are relatively cheap, while those bought by the bag can be quite expensive. The best option for the organic gardener begins with using the materials that are closest to hand. Smaller gardens or those that are newly established may require soil conditioners to be brought in. This section looks at some of the options.

Above: *A well-conditioned soil will pay dividends in terms of enhanced plant growth and development.*

FARMYARD MANURE

This is usually freely available, although these days it comes increasingly from stables rather than farms. It contains a small amount of nutrients although it usually makes the soil more acidic as it decomposes and should always be well rotted before use. It can also contain lots of weed seeds, which often come from the bedding used for the animals. Occasionally there can be problems if the straw used has been sprayed with pesticides, so always be sure of your source. Farmyard manure is usually cheap or occasionally free for the taking.

Left: *Spent mushroom compost is an ideal material for improving the structure of many types of soil.*

BURYING GREEN WASTE

If you have space, you can simply open a trench in the soil and bury your green waste 20–30cm (8–12in) deep under the soil. It will take perhaps two to three months during the warm season to break down. Avoid planting crops in the soil above until the material has decomposed properly.

Left: *Animal manures can provide both nutrients and valuable organic matter when added to your garden soil.*

LEAF MOULD

Many gardeners believe leaf mould to be the finest addition to any soil. The product itself forms naturally under the closed canopy of forest trees. It is, as the name suggests, the rotted product of fallen leaves. It is not particularly rich in nutrients but it has an extremely good ability to condition a soil and encourages natural nutrient cycles in your soil. It is usually made by the gardeners that wish to use it and is rarely sold on. Anyone who has a ready supply of leaves will find this an easy material to make.

Start by gathering up leaves in the autumn and stack in a chicken-wire cage to prevent them being blown back over the garden. Do not add any greens or other materials to the pile. The heap will rot down slowly and be ready for use in one to two years. It can be sited in a dark corner where nothing else will grow, but you must ensure that you keep the pile moist all year round. The wire cage can be removed after two months or so and used again the following autumn. Alternatively, two permanent pits, sited side by side, can be filled (or emptied) in alternate years. Smaller quantities of leaves can be pushed into black plastic sacks with the top closed and with holes pierced by a garden fork around the sides and bottom of the bag to allow some airflow.

SPENT MUSHROOM COMPOST

This is one of the few types of organic matter to have a slightly alkaline effect on the soil. It is a uniform and friable mixture of stable organic materials that is a waste product of the mushroom-growing industry. It is usually quite inexpensive. The composting and growing processes that produce this material bind nutrients to the organic matter, resulting in a substance that holds on to nutrients more readily than fresh or non-composted organic wastes. In addition, the compost has good moisture-holding abilities.

Spent mushroom compost is slightly alkaline with a pH ranging from 7–8 but generally around 7.3. There are few weed seeds, insects or pathogens because the compost is pasteurized before it is removed from the mushroom house. The addition of spent compost to garden soil can result in a

BUILDING A LEAF MOULD PIT

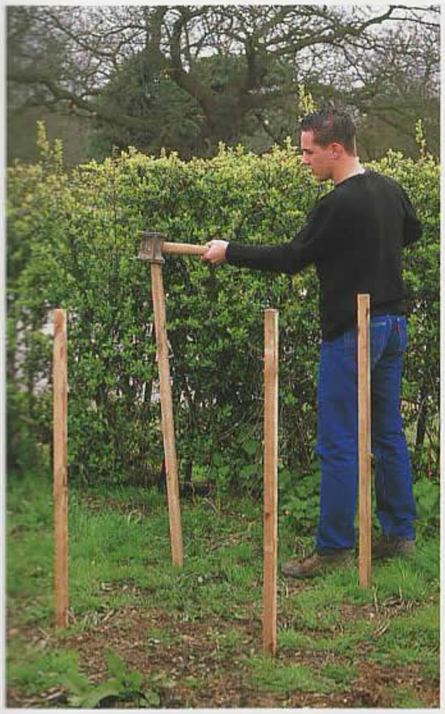

1 *Drive in four 1.5m (5ft) posts to form the four corners of a square on the ground. The posts should be spaced about 1m (3ft) apart and there should be just over 1m left above ground once the posts are driven in.*

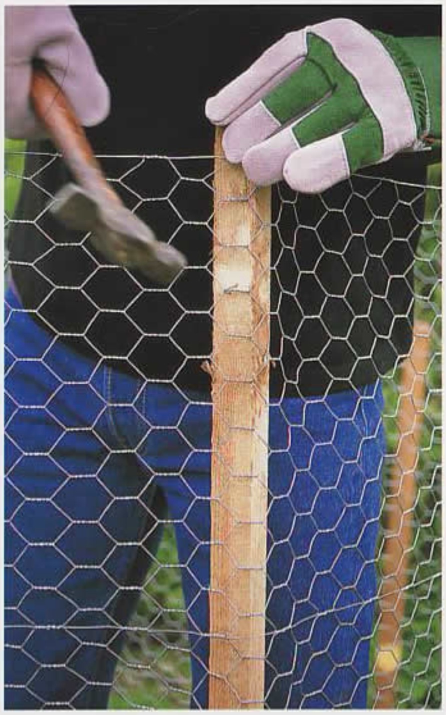

2 *Attach chicken wire – a 1m (3ft) wide roll is needed for this – to the stakes using "U"-shaped fencing staples, hammered in firmly. Pull the wire tight across the length of the roll before attaching it to the next post.*

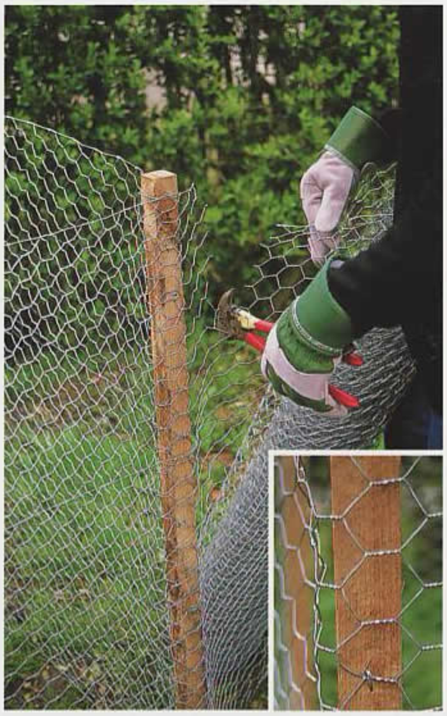

3 *Once the wire has been securely stapled to the posts all the way around, cut it with wire cutters, folding the sharp ends under the attached mesh so that they cannot cut you.*

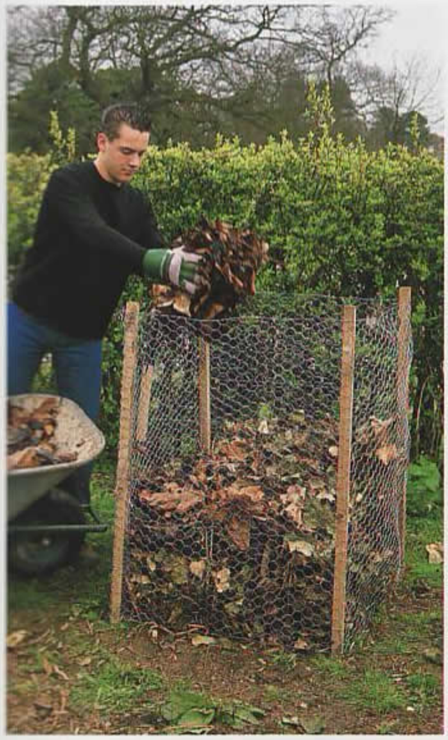

4 *Place fallen leaves into the "pit" until it is full. Firm the leaves lightly in layers as you go, so that the pit is completely packed with leaves but still contains a little air.*

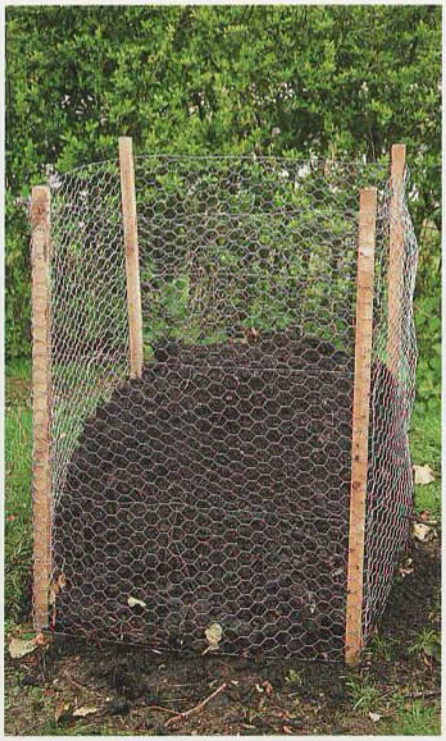

5 *After a year or two, the leaf mould is ready and should be applied to the soil in late winter or early spring in order to gain the most benefit.*

6 *It is a good idea to build two or more pits and rotate their use as a single pit will not be emptied in time to receive the new lot of leaves in the following autumn.*

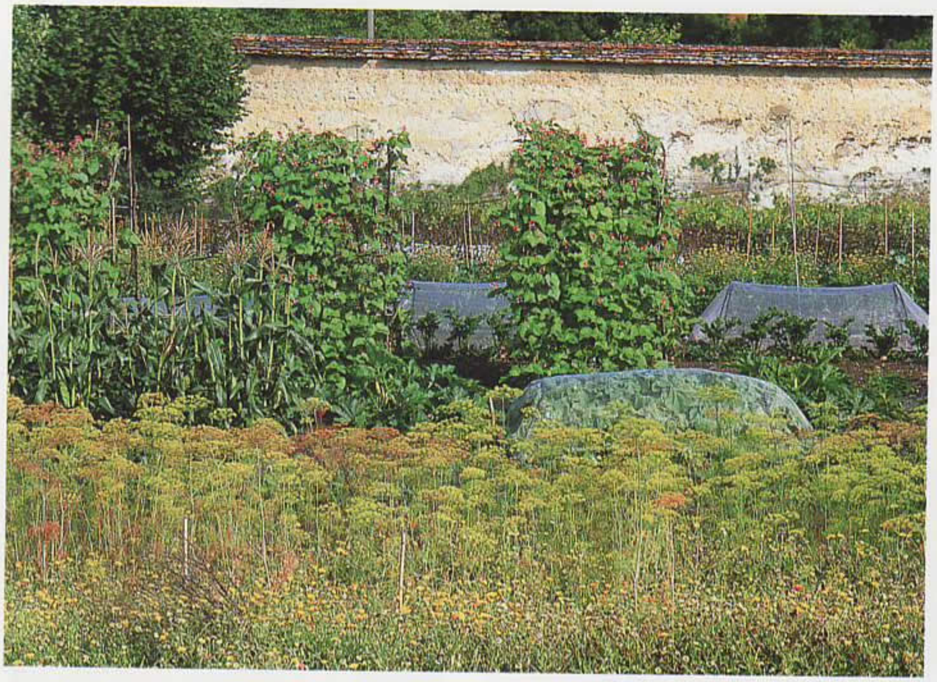

Left: *Beans and other leguminous plants are an important part of a successful organic vegetable garden, "fixing" their own nitrogen and so helping to raise levels of soil fertility.*

higher pH, increased nutrient-holding capacity and better soil structure. The main disadvantage occurs when the compost is first mixed with soil, as bacteria convert the proteins in the compost to ammonia, which can be toxic to young plants. Always ensure the source of your compost is an organic mushroom grower where possible.

VEGETABLE INDUSTRIAL WASTE

Some by-products of the food industry, such as spent hop waste and cocoa shells, can be beneficial as soil conditioners in the organic garden. They are useful in terms of their ability to improve soil structure. However, their smell, which may be described as rancid ale or beer with hops, or sickly cocoa, is often enough to put people off. They are sometimes marketed commercially although they may be available as a local by-product of industry. Whatever product you use, it should be well rotted before it is used as a soil conditioner.

WORM COMPOST

Worms can be used to compost your kitchen waste both indoors and outdoors, although both methods have their limitations and advantages.

The materials that are being composted are consumed by the worms and then excreted as worm casts. This process binds the nutrients consumed into a form that can be used by plants and reduces the volume of the original materials. These worm casts are covered with slow-dissolving, semi-permeable mucus that acts as a time-release mechanism for the nutritional content of the worm casts and also gives the finished compost good water-retentive qualities.

Above: *Brandling or tiger worms readily reduce vegetable matter into a useful compost material within a wormery.*

A WORMERY

Worms are an effective way to break down kitchen waste such as vegetable scraps. They will eventually eat their bedding as well as the kitchen waste. The resulting worm casts are a nutrient-rich soil conditioner.

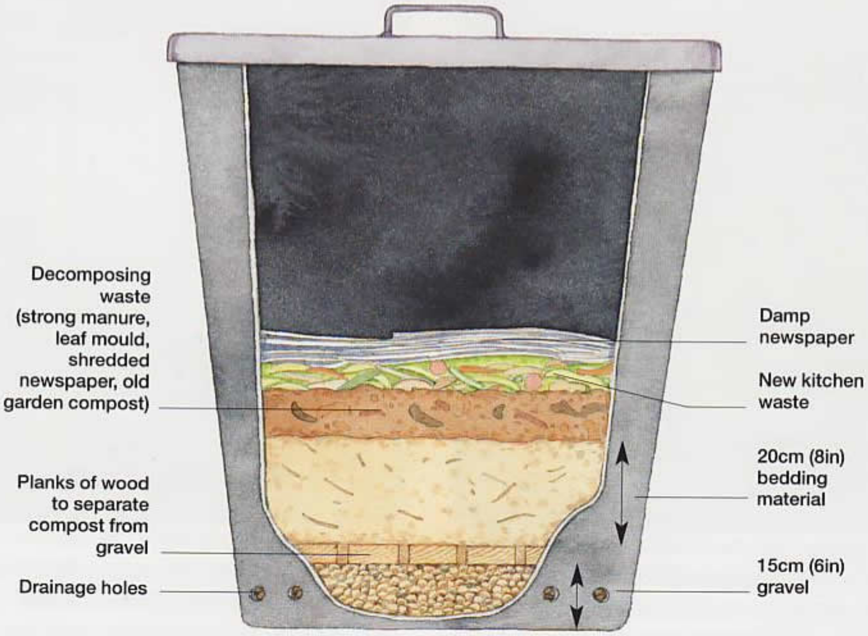
Decomposing waste (strong manure, leaf mould, shredded newspaper, old garden compost)
Planks of wood to separate compost from gravel
Drainage holes
Damp newspaper
New kitchen waste
20cm (8in) bedding material
15cm (6in) gravel

Worm colonies can be built in a similar way to a cold pile compost heap, with materials added as they become available. Excessive heat and frequent pile turning are not needed. Too much heat will actually kill the worms. Alternatively, a purpose-built wormery can be used. Both methods require materials to be placed in a shaded, cool and moist location. The worm colony must not become heated by the microbial action that takes place in normal composting. If the colony does get hot, it may cause the loss of the worms if they do not have cooler areas to which they can escape.

For an indoor worm colony, the wormery (a sealed, ventilated box with a close-fitting lid) is filled with bedding materials (usually damp shredded newspaper or similar) and the worms are added. Vegetable waste from the kitchen is placed on top or a few inches under the bedding where the worms can find and devour it. Eventually, the worms eat their bedding as well as the vegetable scraps. The finished worm castings can be removed and used. New bedding and kitchen scraps are then placed in the container and the worms returned to make a new batch of compost. Many purchased wormeries have the facility to collect the liquid by-product of the worms' activities, which makes an excellent liquid plant food.

It is advisable to check the pH level of the wormery every now and then. Thin, white, thread-like cotton worms are usually an indication that the pH is too low. Adding some calcified seaweed can help to counter this acidity, as can a regular addition of eggshells.

Above: *Leafy crops benefit from the addition of organic matter because it encourages nitrogen-fixing bacteria to do their all-important work in the soil.*

Above: *Place a thick layer – about 8–10cm (4–5in) – of well-rotted organic matter into the base of a trench dug to a spit's depth (the depth of a spade blade). Use a fork to break up the material and to ensure that there is plenty of air space in the soil to aid the breakdown of the material and help soil organisms to cycle the nutrients.*

THE BENEFITS OF ADDING ORGANIC MATTER

Using organic matter as a mulch or soil conditioner will help to warm up the soil earlier in the spring. This in turn will allow for the earlier planting of some types of plant.

- Organic matter inoculates the soil with vast numbers of beneficial microbes and provides the food source that soil-dwelling microbes need to live. These microbes are able to extract nutrients from the mineral part of the soil and eventually pass the nutrients on to plants.
- Adding organic matter to your soil improves the way in which water interacts with the soil and improves the overall soil structure. All soils can benefit from the regular addition of organic matter.

To sandy soils

- Organic matter acts as a sponge to help retain water that would otherwise drain down below the reach of plant roots. This has the benefit of protecting plants against drought.

To clay soils

- Organic matter helps to increase the air space within the soil, making it drain more quickly. This ensures that the soil does not become waterlogged or dry out into a brick-like substance.

MAKING GARDEN COMPOST

Every organic garden should have at least one compost heap, where garden and kitchen waste can be broken down by microbes and other soil-dwelling creatures to produce a good medium-fertility soil improver. Making compost is not difficult and is an excellent way to recycle waste. Many gardeners take great pride in their home-produced compost and all gardens can benefit from this sustainable approach to soil improvement.

Above: *Even a stack of old tyres can be recycled and used to make an effective compost heap.*

THE SITE

A position in full sun will give the compost heap additional heat, allowing the contents to decompose more rapidly. If time is not a key factor, partial shade or even full shade will do. Compost is easily made in a freestanding pile in a sunny or partially shaded area in the garden. Alternatively, a simple structure can be used to contain the pile. A variety of materials can be used to make these structures, including wood, bricks, pallets, a wire cage or even a stack of used tyres. Whichever method you use, you must provide good drainage. An average heap should be around a cubic metre (10 cu. ft). It can be much larger, but any smaller and it will not be as effective.

When adding waste materials to the heap, the items should be as small as you can make them. The more surface area that is exposed, the quicker it will decompose. A chipper or shredder is a big help. Many gardeners are keen advocates of adding a small amount of good topsoil in order to introduce microorganisms that will kick-start the process of decomposition. Compost heaps in containers can be built up with successive layers of "green" and "brown" materials, separating these with thin layers of topsoil.

Left: *Composting your garden and kitchen waste ensures that you return essential nutrients to the soil.*

Sifting the pile is the final stage of the process, with any large pieces of residue being returned back to the new heap.

COMPOST INGREDIENTS

The potential ingredients for your compost heap will be either "brown" or "green". Browns are dry and dead plant materials, such as straw, hay, dry brown weeds, autumn leaves, nutshells, shredded paper, pine needles, tough plant stems and wood chips or sawdust. Because they tend to be dry, browns often need to be moistened before they are put into a compost system.

Greens are fresh (and often green) plant materials, such as green weeds from the garden, fruit and vegetable scraps, green leaves, grass clippings, tea bags, coffee grounds, seaweed, eggshells, fish scraps, green manures and fresh horse manure. Compared with browns, greens contain

MAKING A HOT PILE COMPOST HEAP

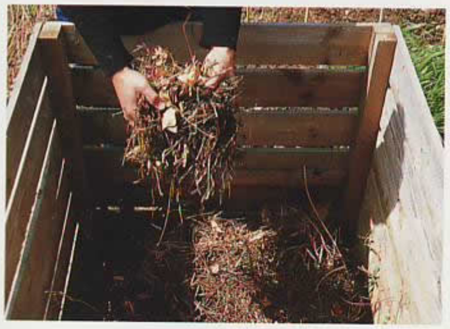

1 *Starting with an empty compost bin, place a layer of "browns" – straw, old leaves and chipped wood are ideal – in the base of the bin. If the bin is positioned over bare soil, include some fine twiggy material to help to circulate air to the base of the pile.*

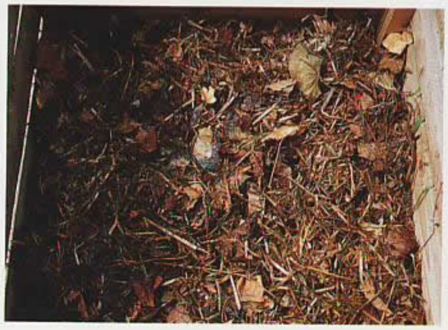

2 *Add a layer of brown compost materials, such as autumn leaves, plant stems and wood chips, which should be approximately 15cm (6in) deep, and ensure that it is of an even thickness. The material should be lightly firmed in.*

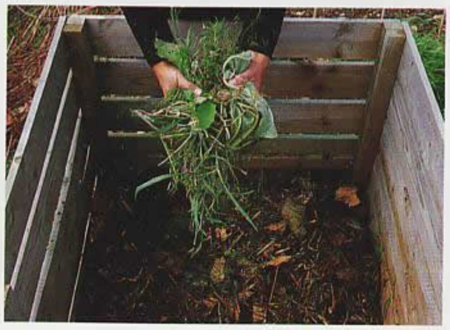

3 *Add a new layer of "greens" to the compost bin. Most plant material can be used for this, although you should avoid too much fibrous or woody material such as plant stems. Never include weed seeds if possible.*

Above: It is often advisable to have two or more compost heaps at any one time in order to ensure a steady supply.

more nitrogen. Nitrogen is a critical element in amino acids and proteins, and can be thought of as a protein source for the billions of multiplying microbes.

A good mix of browns and greens is the best nutritional balance for microbes. This mix also improves the aeration and level of water in the pile.

THE "COLD PILE" METHOD

Make a pile of "browns", such as autumn leaves, moisten them well and then slowly incorporate "greens" such as kitchen waste or lawn clippings over the next year (this method can take a full year to finish the compost). If it gets very cold in winter, cover the heap with cardboard or old carpet and peel this back to stir in any greens. In this type of composting, there are never enough high-nitrogen greens to get the pile really hot. Turn the pile a minimum of once a month if you have not recently added any greens, thereby breaking up the decaying mass and making it friable. If you do use this method, be sure you always have at least as much browns as added greens, as too many greens will just create a slimy, smelly mess. Some compost heaps (those with lots of greens added) may reach very high temperatures, some even reaching 70°C (158°F), thereby killing weed seeds. Cold piles, on the other hand, do not and it is best not to add any weed wastes that will contain seeds. If you do, you could risk spreading weed seeds across your garden in their own growing mix when you come to use the compost.

THE "HOT PILE" METHOD

Wet the ground under the pile and add twigs or other un-shredded browns to provide some aeration at the base. Layer the rest of your materials, alternating green and brown layers of about 15cm (6in) thickness, and add water as you go. Topsoil can also be added as a 2cm (¾in) layer between each cycle of green and brown materials. Finish the pile with a layer of browns. Cover the pile with a lid or piece of carpet to keep out rain and conserve heat. Check to see that your pile becomes hot within a few days. Turn the pile to decrease composting time. This action allows all the material to be exposed to the hot centre, thereby increasing aeration. Do this once a week in the warmer season, and once a month in cooler periods. The pile's heat should peak every time you turn it, although the peak temperature will be lower with each turn. Always make sure that the pile remains moist, but avoid over-wetting.

COMPOSTING MATERIALS

Many items can be composted into valuable soil conditioners. Whether because of toxins, plant or human diseases or weed problems, some things should never be used.

Good compost materials
Animal manure
Fallen leaves
Grass/lawn clippings
Hay and straw from organic farms
Kitchen waste
Prunings from the garden
Sawdust
Shredded browns
Soot and charcoal
Spent hops or cocoa shells
Spent mushroom compost
Weeds and other garden wastes

What not to compost
Chemically treated wood products
Diseased plants
Human and pet waste
Meat, bones and fatty food wastes
Pernicious weeds

4 *Continue adding greens until you have a layer about 15cm (6in) thick, the same depth as the browns below. Lawn clippings should be placed in layers of about 10cm (4in) or mixed with other green waste to avoid the layer becoming slimy and airless.*

5 *Kitchen refuse, such as vegetable waste, can also be added and is usually classed as a "green". You can include a thin layer of soil to add microorganisms before the next layer of "browns". Continue layering until the bin is full, then water and cover.*

6 *After two to three months, you should have well-rotted garden compost, which, due to the heat of the pile, will be largely free from weed seeds, pests and diseases. Once the bin has been emptied, it can be refilled in the same way.*

GREEN MANURES

Sometimes referred to as "cover crops", green manures are plants that are grown to benefit the soil rather than for consumption or display. Green manures replace and hold nutrients, improve the structure of the soil and increase its organic material content. They also smother the soil and so prevent weed growth. Green manures also "fix" atmospheric nitrogen. They are easy to grow and can be used to improve the soil when manures are not readily available.

Above: *Growing a green manure is an easy and effective way in which to add some organic matter to the soil.*

BENEFITS OF GREEN MANURES

At any time of year, but especially in winter, the soil loses nutrients if it is bare for six weeks or more. Green manures help to counteract this and maintain a more even soil temperature and moisture content. Many green manures grow deep roots to tap resources unavailable to some crops. Others produce a fibrous root system to help build structure in the soil, while many have flowers that attract pollinating insects.

TYPES OF GREEN MANURE

Some green manures are nitrogen-fixers, using bacteria that colonize the nodules on their roots. These microbes take nitrogen out of the air and convert it into a form that plants can use. Green manures are usually divided into two groups: legumes and non-legumes. Leguminous manures include clover, peas, fava beans and alfalfa. Despite their nitrogen-fixing abilities, legumes have slow autumn growth and add less organic content. They can also be less winter hardy.

Winter rye is the most commonly grown non-leguminous cover crop, but oats, wheat, oilseed rape and buckwheat are also used. Although they do not add nitrogen to the soil, they help maintain levels and have the additional advantage over legumes of growing faster through the autumn, thereby giving better weed suppression. They also tend to break down more slowly than legumes and add more organic matter to the soil. You may wish to use a combination of green manures in order to balance the benefits of the different types.

DEEP-ROOTED GREEN MANURES

The roots of certain green manure "crops" work through the soil, holding it together as they grow. Once dug in, however, they help to make it crumbly and friable as they rot. Some green manures have very deep roots that reach down into the subsoil and can harvest nutrients that are unavailable to other garden plants. In this respect they are a form of biological double digging.

PLANTING GREEN MANURES

Winter green manures should be planted early enough to give about four weeks of growth before cold weather stops their growth. Spring or summer crops must be given sufficient time to develop before they are dug in. After preparing the soil, plant large-seeded cover crops in shallow, closely spaced furrows. Smaller seeds can be broadcast over the surface and covered with a light raking, watering if needed until they germinate. Dig the crop in once it is ready, allowing at least three weeks before you intend to plant to give time for the material to rot down. It is best not to allow green manures to go to seed as they all have the potential to become weeds.

Gardens are traditionally dug over and manured in autumn, then left bare over winter. Many nutrients can be lost during this time due to the leaching action of rain. A hardy green manure sown in the autumn can be dug in the following spring, yielding nutrients to a newly sown or planted crop.

USING GREEN MANURES

1 *Sow the seed evenly across the area where you plan to grow the green manure. Either broadcast the seed or sow larger seed thinly in shallow drills and close rows.*

2 *Lightly rake in the seed so that it is covered and will germinate quickly. Water the area thoroughly if there is no rain due or if it does not rain for the following 48 hours.*

3 *Once the seed germinates, allow it to grow a little before digging it into the soil. Never let green manure crops set seed or they will become a weed in later crops on the site.*

TYPES OF GREEN MANURE

Green manure for growing over winter

These species can withstand moderate to hard frost for a long period and can be cut down in the spring, prior to cultivation. The shorter types can also be used as catch (cover) crops around winter vegetables or biennial crops for harvest in the early spring. This protects bare soil, prevents nutrient leaching and creates a more stable soil environment.

GREEN MANURE	SOWING TIME AND METHOD	HEIGHT	WINTER HARDY?	NITROGEN-FIXING ABILITY	TERMS OF GROWTH	DIGGING IN	OTHER NOTES
Alfalfa *Medicago sativa*	Early spring to mid-summer; broadcast at 3g (1/10oz) per square metre (yard).	100–150cm (3–5ft)	Yes	Yes but poor	Perennial. Grow for several months or more than one season.	Any time. Medium effort if young. Hard if left for more than one season.	Very deep rooting. Will grow on most soils. Dislikes acid or waterlogged soils, but drought resistant.
Alsike clover *Trifolium hybridum*	Mid-spring to late summer; broadcast at 3g (1/10oz) per square metre (yard).	30cm (12in)	Yes	Yes	Several months	Any time. Medium effort.	Will withstand wetter soils than other clovers but more prone to drought. Shallow rooted.
Essex or red merviot clover *Trifolium pratense*	Mid-spring to late summer; broadcast at 3g (1/10oz) per square metre (yard).	40cm (16in)	Yes	Yes	Several months	Any time. Easy, little effort.	Prefers good loamy soil. Can be mown or cut several times per season and used for compost.
Grazing rye *Secale cereale*	Late summer to late autumn; broadcast at 30g (1oz) per square metre (yard) or thinly in rows 20cm (8in) apart.	30–60cm (12–24in)	Yes	No	Autumn to spring	Before flowering. Hard work.	Grows in most soils. Keep watered during germination, else yield is poor. Sow thickly to smother weeds.
Phacelia *Phacelia tanacetifolium*	Early spring to early autumn; broadcast at 3g (1/10oz) per square metre (yard) or thinly in rows 20cm (8in) apart.	60–90cm (12–36in)	Yes	No	2 months in summer; 5–6 months over winter	Before flowering. Easy, little effort.	Grows in most soils. Quick to grow in summer. If left, will produce mauve flowers that bees love.
Trefoil *Medicago lupulina*	Early spring to late summer; broadcast at 3g (1/10oz) per square metre (yard).	30–60cm (12–24in)	Yes	Yes	Several months to a year	Any time. Medium effort.	Will grow in most soils but dislikes acid. Can be used for undersowing. Dense foliage.

Green manure for warm-season growing

Plants for use in the spring and summer have to be quick growing in order to cover the ground and yield benefit within a short period of time. They are generally slightly tender, although lower-growing species can actually be used as a catch crop around other seasonal crops.

GREEN MANURE	SOWING TIME AND METHOD	HEIGHT	WINTER HARDY?	NITROGEN-FIXING ABILITY	TERMS OF GROWTH	DIGGING IN	OTHER NOTES
Bitter lupin *Lupinus angustifolius*	Early spring to early summer 4cm (1½in) deep; 3cm (1¼in) apart; in rows 15cm (6in) apart.	50cm (20in)	Mild winters	Yes	2–3 months	Before flowering. Easy, little effort.	Prefers light slightly acidic soil. Foliage not very dense. Deep rooted.
Buckwheat *Fagopyrum esculentum*	Early spring to late summer; broadcast at 10g (1/3oz) per square metre (yard) or thinly in shallow rows 20cm (8in) apart.	80cm (32in)	No	No	2–3 months	Before or during flowering. Easy, no effort.	Grows on poor soils. If allowed to flower, attracts hoverflies to aid pollination of crops.
Crimson clover *Trifolium incarnatum*	Early spring to late summer; broadcast at 3g (1/10oz) per square metre (yard).	30–60cm (12–24in)	Mild winters only	Yes	2–3 months or over winter	Before flowering. Medium effort.	Prefers sandy loam soil but will tolerate heavy clay. Large red flowers attract bees.
Fenugreek *Trigonella foenum-graecum*	Early spring to late summer; broadcast at 5g (1/5oz) per square metre (yard) or thinly in 15cm (6in) shallow rows.	30–60cm (12–24in)	Mild winters	No	2–3 months	Any time before flowering. Easy, little effort.	Prefers good drainage but will tolerate heavy or light soil.

BASIC TECHNIQUES

Most gardening techniques, from weeding and feeding to pruning and propagation, are straightforward and and easy to achieve if you plan and organize the task in hand. The procedures described here will help you to master the basics of organic gardening. You may well develop your own methods in time, based upon the techniques shown here as well as those you that learn from other gardeners. Always make sure that you are clear what you want to do and always approach a task positively. Gardening should bring enjoyment as well as a sense of satisfaction that comes from creating a thriving organic paradise in your own backyard.

Left: *A flourishing garden plot is the end result of careful planning and preparation as well as regular maintenance.*

Above: *A well-prepared, weed-free plot is the ideal starting point for successful plant growth and development.*

Above: *Both weeds and crop residues can be recycled in the organic garden by composting and then returned to the soil.*

Above: *Taller plants may need some form of support. In this case, willow branches have been used to support* Lysimachia.

WEEDING AND WEED CONTROL

Many people are put off organic gardening because they do not like the idea of weeding. Not having the "quick fix" of weed-killing chemicals can make this task seem daunting. It is, however, worth considering that a well-maintained garden has less weeding to do than you might think. In addition, it can be a relaxing, even therapeutic, task and there is very little that is more satisfying in the garden than a freshly weeded plot.

WHAT PROBLEMS DO WEEDS PRESENT?

Weeds compete with garden plants of all types for essential nutrients, light and water. Uncontrolled weed growth may kill plants that are less vigorous or seriously inhibit their development. Weeds also act as "host" plants for animal pests or diseases that may, in turn, affect the desired plants. Finally, weeds are controlled because they often look unsightly.

CAN WEEDS BE TOLERATED?

There is no easy answer to this question. It depends upon what exactly you want out of your garden. Aesthetic considerations aside, most gardens can survive quite healthily with some weeds. The majority only become a real nuisance when they start to take over or if they set seed. The simple rule is the fewer weeds in the garden the better.

GOOD CULTIVATION

The secret of organic weed control is good cultivation. This begins with thorough ground preparation to reduce the build-up of weeds and weed seeds before planting. Fallowing is commonly used to do this. After the ground has been prepared, it is left bare for a period of a few weeks to a few months. The weed seeds are allowed to germinate and are then removed as they appear. This method will never remove all the seeds because they can survive for several years in a dormant state. But the majority of them will have germinated and the task of weed control made easier for the future.

Vigilance is important throughout the growing season, but never more so than in spring and early summer. Tackle weed seedlings as they appear. Some methods of control, such as hoeing, may damage garden plants, so use a hand fork for close work or in areas where hoeing is difficult. A hoe, for example, can sever the roots of surface-rooting shrubs such as rhododendrons. These weeding methods are relatively slow and labour-intensive. The trick is to do it little and often.

Flails, brush cutters, mowers, rollers, burning and other mechanical means are occasionally used. Although strimmers can be used in shrub borders, those that cut using nylon cords can damage woody stems.

MINIMAL CULTIVATION

Many weeds are plants of cultivation, thriving in disturbed ground. Minimal cultivation reduces disturbance and relies on creating a ground-covering layer of vegetation. The plants literally out-compete the weeds, which can be removed as they appear.

Left: *Weed-free plots can be attractive to crop pests. Here, plant collars have been used to prevent root flies attacking the crop.*

Above: *Hoeing is an effective way of keeping a garden plot free of weeds. Pull a draw or swan-neck hoe towards you in a series of chopping movements.*

COMMON WEEDS

Weeds can be divided into three groups. Annual weeds germinate, grow, flower and set seed in one growing season. Certain annuals, which are known as ephemerals, complete their lifecycle well within this period and can produce several successive generations within the space of one season. The third type of weeds are perennial species. Perennial weeds initially establish from seed, but tend to persist for several seasons, sometimes dying down during adverse seasons only to re-emerge as conditions improve. They are much more difficult to eradicate once established than annual weeds and many are a serious problem in gardens.

Ephemeral and annual weeds

- Annual meadow grass
- Bittercress
- Chickweed

Perennial weeds

- Blackberry (in Australia)
- Couch grass
- Creeping buttercup
- Dock
- Ground elder
- Groundsel
- Nettle

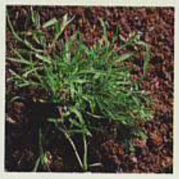

Meadow grass

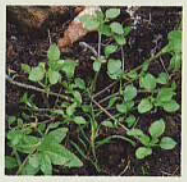

Chickweed

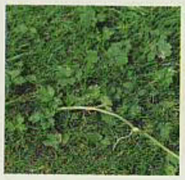

Creeping buttercup

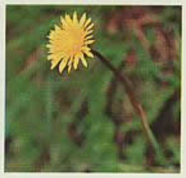

Dandelion

DIFFERENT TYPES OF MULCH

Mulches provide a physical barrier to weed seeds that would otherwise land on the soil surface. The materials that can be used for mulching are varied in their composition and the effects they have. They are generally classified into synthetic and natural types.

Bark chippings

Newspaper

Dried bracken

Gravel

Old carpet

Sawdust

Woven plastic

Cocoa shells

MULCH

This is a layer of material that is laid over the surface of a soil or other growing medium. While often cited as a means of weed control, applying mulch has other functions that are beneficial to growth. Mulches help to keep the plants' roots warm in winter and cool in summer, greatly assisting new growth in the spring. They also reduce water loss from the soil in two main ways: by shading the soil surface and by slowing down evaporation of water from the soil.

In general, most mulches reduce the number of weeds growing in a planted area. They act by inhibiting the germinating weed seedling from reaching the sunlight or by "drawing up" the weed so that it is spindly and can be easily removed or "pulled". Mulches should be laid on a bare surface, as they will not inhibit established weeds, especially vigorous perennial ones.

If a mulch is laid to an even depth of around 5cm (2in), it will develop a dry dusty surface that will deter the germination of weed seeds that land on it. This is only a temporary measure, but can help reduce weeding at the busiest times of the year.

TIMING AND APPLICATION

Mulches should be laid on warm, moist soil. Autumn is the ideal time, but they can be applied in the spring after the soil has warmed up. Applied too thickly (over 10cm/4in), they may affect water infiltration into the soil. Materials containing an excessive amount of fine humus or soil-like material ("fines") may encourage weed seedlings to establish. In poor-draining soil, mulch may keep the soil waterlogged and this will be detrimental to plant growth. Synthetic materials, especially if they are impermeable such as black plastic, will inhibit oxygen and water infiltration into the soil. Some materials, such as grass clippings, can "pack down" on the surface and form a water-shedding layer.

As mulches encourage surface-rooting, application must be continued once started, otherwise the plants will suffer later. Many ornamental plants, such as rhododendrons, are naturally very shallow-rooting and thus need the protection of mulch for good growth. They benefit from an organic mulch that conserves soil moisture. Some mulches, such as manure and compost, also aid plant development by supplying nutrients.

MULCHING TO CONTROL WEEDS

1 *Grass clippings are a cheap, effective mulch. Apply them at a maximum depth of 5cm (2in) or the heat they create as they decompose may harm the stems. Do not use mowings from grass that has gone to seed.*

2 *Chipped bark should be composted or stored for several months to let it release any resin and start to decompose. Some gardeners worry that it introduces fungal diseases, but the spores are already in the air.*

WATERING

In order to cultivate a successful organic garden, it is crucial to recognize the water requirements of different plants. It is of vital importance that you know when and how to provide the right amount of water as well as how to avoid either drought or waterlogging. Conserving the water in the soil as well as storing and recycling this valuable commodity are of key importance in the organic garden, especially in areas where rainfall is seasonal and droughts are common.

Above: *Water that runs off the roof can be collected in a custom-made rainwater butt, and stored for later use.*

PRINCIPLES OF WATER MANAGEMENT

After rainfall or irrigation, all the soil pore spaces are filled with water. The air is displaced and the soil saturated. If the rain or irrigation were to continue, then surface pooling or runoff would occur. The rate at which a soil can absorb water is influenced by its texture and structure.

Water drains from the soil through the pore system and is replaced to some extent by air. After draining has removed the excess water, the remaining water may be lost either through evaporation or taken up by the plant and lost through its leaves.

All sands and loamy sands tend to have a low available water capacity. Loams and clay have a medium capacity while very fine sandy loam, silt loams, peats and any soil with a high water table have high available water capacity. Clay soils often have a large amount of water within them, but it is held so tightly that plants are not able to obtain it.

WATER AND THE PLANT

All plants are largely made of water and it is critical that water levels are maintained if the plant is to grow and develop. Soft fleshy subjects, such as many bedding plants and some vegetables, use large quantities of water, particularly during establishment. Plants with large leaves or with extensive branch systems, such as large trees, will remove gallons of water from the soil in the course of a hot sunny day.

WATERWISE GARDENING

Since water supplies are limited, we need to do all we can to conserve soil moisture. As the top layer of soil dries out, the dry soil acts as an insulation to further rapid moisture loss. Take advantage of this by avoiding over-cultivation, particularly during the summer, as this may bring moist soil to the surface to dry out. Always try to maintain a well-structured soil and avoid capping (a hard, partially compacted surface layer) as this can prevent the absorption of rainfall by the soil, leaving puddles that quickly evaporate. Mulches help to conserve moisture either by holding water or acting as insulation provided that they are deep enough. Keep the soil surface weed-free and provide shelter from drying winds. Maintain drainage, as poor drainage leads to limited root growth of the plant and poor soil structure.

STORING AND RECYCLING WATER

Water is a precious resource that is sometimes in short supply. Plants in open ground may acclimatize to the soil conditions, but those growing in pots are more susceptible to drought. Mains water can provide an answer but there is a cost involved and supplies can be restricted. Saving water from rainfall or recycling waste water (which is sometimes called "grey water") can overcome the worst ravages of summer drought.

METHODS OF WATERING

1 *When watering by hand, be patient and give the ground around the plant a good soaking. If in doubt, dig a small hole in the soil and check that the water has soaked through to the roots.*

2 *Turning on a sprinkler allows you to leave it and get on with something else. Place a jar under the spray to gauge how much has fallen. There should be at least 2.5cm (1in) of water for the sprinkler to have done any good.*

3 *A spray attached to a garden hose delivers at a greater rate than a sprinkler but it must be held in place until the ground is soaked. It is easy to under-water using this method. Spray the leaves to wash away dirt or dust.*

4 *A seep hose with holes in it is snaked around plants that need to be watered and left in position. When connected, it provides a gradual but steady flow of water that penetrates deeply into the soil.*

AUTOMATIC WATERING SYSTEMS

1 You can bury a pipeline beneath the surface and plug in various watering devices. A sprinkler can be pushed on to this fitting, which lies flush with the turf.

2 Control systems can be fitted to the hose system so that you can alter the pressure of the water. These can also act as a filter.

3 Drip-feed systems can be used for beds, borders and containers. "T"-joints allow tubes to be attached for individual drip heads.

4 The delivery tube of the garden hose can be held firmly in position with a pipe peg that is inserted into the ground, if this is necessary.

Storing and recycling household water is a relatively complicated undertaking that is usually best left to specialist contractors. The potential health risks mean that it is not usually wise to store this type of water for any length of time, but it can be used fresh to water non-food plants. Try keeping your bath water and using it to water ornamental plants. It can be used to fill watering cans and is ideal for watering outdoor pot specimens or plants in open ground.

Rainwater may be a viable alternative and some contractors will be able to install large underground tanks to contain seasonal rainfall. The quicker and easier alternative is to install water butts to collect water from downspouts from roof drains. This method will rarely provide enough water for your whole garden, but is often plenty to water (or supplement the watering of) plants in containers, particularly edible plants such as herbs. Rainwater is naturally acidic and is, therefore, not likely to contain many harmful bacteria. Its acidity also makes it the ideal choice for misting plants under glass on hot days (as it will not mark the leaves with lime) and suits many species sensitive to domestic water additives. Rainwater is also a must for people wishing to grow "acid-loving" specimens in pots.

WATERING PLANTS

In areas where seasonal water shortage is a regular problem, choose drought-tolerant species where possible. Seasonal crops (such as vegetables) and displays (summer or spring bedding) tend to suffer the most from drought due to the limited time that they have to develop extensive root systems. If you wish to grow these seasonal plants in drier areas, they will need to be watered regularly to survive prolonged dry periods. Perennials, on the other hand, may actually become more drought-tolerant if they are watered less. This is because they will have to root deeper to find water. Perennials will need to be watered in their first season following planting to ensure their long-term survival. Remember that applying a good drench on a more occasional basis allows the water to penetrate more deeply and have a better long-term effect because it encourages deeper rooting. Watering little and often can prove detrimental in drier climates by encouraging surface roots that may become very drought-prone. After the first growing season following planting, they are better able to cope if not watered at all.

WATERING METHODS

Plants can be watered in many ways, but all generally fall into two categories: hand watering or automatic systems.

Hand watering The commonest "tool" here is the watering can. Numerous designs exist but they all rely upon a gradual delivery through a spout. A rose ensures that a fine spray is delivered that will not wash away soil or damage plants. Hoses can also be used for hand-watering and many nozzles have been designed to deliver a shower of water rather than a jet.

Automated systems There is an array of devices to choose from, some of which can deliver doses of water at set times of the day or week. Even the most complex of these use relatively simple water-delivery methods such as seep or drip hoses and sprinklers. Seep and drip hoses are essentially "leaky pipes" that gradually release water into the soil. Sprinklers shower an area of garden repeatedly until the ground is well soaked. There are many designs for both and the amount of time they need to be left on depends upon the manufacturer's recommendations and the available water pressure. Both hoses and sprinklers offer the benefit of gradual delivery over a large area and both can be operated manually from a mains water supply. A good watering every few days can be extremely beneficial to most plants, especially if recently planted.

WATERING PROBLEMS

These are usually the result of either too much or too little water being applied to the plant. Both over- and under-watering have similar symptoms because both are related to root death. An over-watered plant's root system begins to die off due to a lack of oxygen. The plant becomes short of water and wilts. Other problems can result from the wetting of foliage and flowers. Watering in full sunlight can give plants a scorched appearance, so it is best not to water in hot sunny conditions.

RECYCLING WATER

The suitability of the water left over from washing, bathing and other household activities depends on how much waste matter it contains. Ordinary bath water, for instance, is useful for watering a variety of plants, but the main problem lies in its storage. Waste water always carries the risk of bacterial infestation. Specialist storage methods and filtration can be used, but it is easier to use it immediately (once cooled), and to avoid using it on houseplants or edible crops.

FEEDING

In order for plants to grow and sustain their life processes, they require water and air as well as a range of naturally occurring elements. At least sixteen different elements are needed by all plants for successful growth and development. Three of these elements – carbon, hydrogen and oxygen – are obtained from the carbon dioxide in the air and from the water and oxygen in the soil. In addition to these, plants also need a selection of other important nutrients, which are described here.

Above: ***Top-dressing with organic matter provides the soil with nutrients and encourages nutrient cycles.***

IMPORTANT PLANT NUTRIENTS

In addition to carbon, hydrogen and oxygen, plants need other important nutrients, which they obtain directly from the soil or compost. These nutrients are commonly divided into two groups, depending upon whether they are found in plants in high or low concentrations. When they are found in high concentrations and are needed in large amounts, they are known as macro or major nutrients. There are three of these: nitrogen (N), phosphorus (P) and potassium (K). Every organic gardener must be fully aware of the functions of the "essential three". Nitrogen is essentially the shoot and leaf maker and is especially important early in the growing season; phosphorus is important in the formation of roots; and potassium plays a vital role in flower and fruit formation and is also important in promoting hardiness in overwintering plants. Elements that are needed in relatively small amounts are the micro or trace elements, of which there are ten. Among the most important of these are calcium, magnesium and sulphur, although all play a vital role in the growth and development of the plant.

ORGANIC FERTILIZERS

A wide variety of organic substances can boost nutrient levels. They should be used in addition to composts and manures.

Animal manure Bird manure is a good source of nitrogen. Use in pellet form, as the fresh substance will burn plant roots.

Bonemeal Promotes strong root growth due to its high phosphate content. Use it as a base dressing prior to planting perennials.

Calcified seaweed This raises the pH of the soil and adds calcium, like limestone, but also contains magnesium.

Fish, blood and bone A general fertilizer that is normally applied in the spring to promote root and shoot growth.

Ground limestone This raises the pH and supplies calcium where there is a deficiency. It is applied as a powder when it is needed.

Gypsum This supplies calcium to soils where there is a deficiency, but does not alter the pH. It is added in a 4:1 ratio to "lighten" clay soils.

Hoof and horn This material gives a slow release of nitrogen where strong growth is needed and is applied in the spring or early summer.

Rock phosphate Used mostly to counter phosphate deficiency and as an alternative to bonemeal.

Seaweed meal Helps to build up humus levels in the soil and is used on a wide range of plants.

Wood ash An excellent source of potassium that can be added directly to the soil or to a compost heap.

Bonemeal

Fish, blood and bone

Hoof and horn

Calcified seaweed

SYMPTOMS OF DEFICIENCY

Plants exhibit varying symptoms to a nutrient deficiency. General ones are stunted growth, discoloured leaves, including mottling and inter-veinal coloration, and the premature death of leaves or other parts of the plant. Plants may also display twisted or distorted growth as well as poor root growth and development. In many cases, the symptoms are often seen first in the plant's extremities, such as the tips of the shoots, as these are furthest away from the nutrient source. If a plant cannot obtain enough of any nutrient from a soil it will be because there is either little or no nutrient present in the soil, or the nutrient is not "available" to the plant.

Any element can also be toxic to the plant if it is present in sufficiently high amounts or out of balance with other elements in the growing medium.

FEEDING PLANTS ORGANICALLY

Numerous organic nutrient supplements are available. All have their relative advantages, but they should be used only as part of a combined strategy that aims to harness natural nutrient cycles. This usually involves top-dressing the soil with well-rotted organic matter. This is the catalyst to nutrient cycling in the soil as it feeds the

MAKING COMFREY TEA

1 Cut a large bundle of comfrey leaves from the patch in your garden. Choose good, green, leafy growth. Remove the leaf and stem, but avoid using the crown.

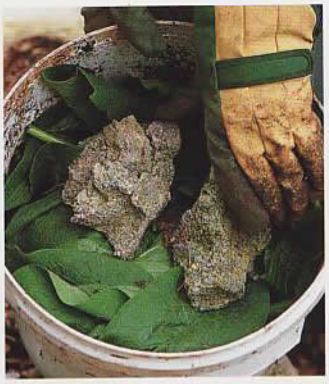

2 Place in a plastic container with holes in the base. Weigh down the leaves. Line a second larger container with a plastic bag and place the first container inside. Tie the bag at the top.

3 The thick dark liquid from the decomposing comfrey leaves collects in the bag below. The lower bucket allows this to drip down and stores it until the process is complete.

4 Decant the finished liquid into a jam jar for later use. The liquid feed must be diluted before it is applied to plants at a ratio of about 10 parts water to 1 part concentrate.

vast array of soil organisms that naturally make use of essential plant nutrients. The majority of garden composts do not contain especially large amounts of the major plant nutrients; it is their long-term ability to feed soil life that makes them so valuable.

Despite the best intentions, plants can become short of a certain type of nutrient. This is usually avoided by giving the plant a supplement of a "plant food" that is rich in that particular nutrient. This can be applied to the soil as a top-dressing, lightly worked in with a fork, or applied as a foliar feed. Foliar feeding involves spraying a fine mist of a low-strength solution of the feed. The exact concentrations vary according to the feed being used, but a rough guide would be to use a quarter strength of that which you would apply as a liquid feed to the soil. The nutrient will be taken up quickly and the residue is best washed off after a couple of days because any soil-applied fertilizer will have been taken up by then. Applying a weak solution of Epsom salts as a foliar or liquid feed is an example of applying such a tonic.

LIQUID FEEDS AND TEAS

Compost tea is easy to make. Comfrey is ideal for this, but stinging nettles are also very good. You could also use sheep or goat manure, or finely sifted compost from the heap. Simply fill an old pillowcase or burlap bag with your chosen material (comfrey, nettles, compost or manure) and sink it into a large bucket or barrel of water. Cover the container and let it steep for a few days. Remember, the longer you steep, the stronger the tea will be. You may use the final "brew" as a light liquid feed. It can also be used as a foliar feeding medium, provided that it is well diluted. Use the residue in the bag as mulch. An alternative method is shown above.

Worm colonies in containers also produce a liquid feed that is a tonic for your plants. It should be diluted to a ratio of about 10 parts water to 1 part liquid feed. It is generally high in potassium and phosphorus.

OTHER FEEDING METHODS

1 An alternative method of feeding is to use a liquid feed. This is most useful for plants that are growing in containers. Add the organic fertilizer of your choice to one of the waterings, according to the manufacturer's instructions.

2 Organic granular fertilizer can be applied by hand, spreading it over the area covered by the roots below. Follow the manufacturer's instructions. You may wish to protect your hands when applying the fertilizer by wearing a pair of gloves.

3 Apply a layer of well-rotted organic material, such as farmyard manure or garden compost, to the surface of the soil around the plant. If the plant is not shallow-rooted, lightly fork the material into the top layer of the soil.

PLANT NUTRIENTS

The following table highlights the characteristics of the soil nutrients. Symptoms of deficiency may be seen on different parts of the plant at different stages of growth, although it may not always be immediately obvious that a plant is suffering from a nutrient deficiency. Some of the more important symptoms that can be seen on plants are listed below.

PLANT NUTRIENT	NATURAL ORGANIC SOURCE	PLANTS MOST IN NEED	SOILS MOST IN NEED	SIGNS OF SHORTAGE	HOW TO AVOID SHORTAGE	TOXICITY CAUSE
Major plant nutrients						
Nitrogen (N) The "leaf maker"	Bird manure, blood and bone, grass clippings	Grass, vegetables grown for their leaves, and root-bound plants	Sandy soils in rainy areas	Stunted growth; small pale green leaves; weak stems	Apply a base dressing before sowing or planting. Top-dress in spring and summer with organic matter.	Excessive use of nitrogen-rich fertilizer
Phosphates (P_2O_5) The "root maker"	Comfrey, seagrass, horse manure, blood and bone	Young plants, root vegetables, and fruit and seed crops	Sandy soils	Stunted roots and stems; small leaves with a purplish tinge; low fruit yield	Apply a base dressing of bonemeal. Use fish, blood and bone when top-dressing plants.	Excessive use of fertilizer
Potash (K_2O) The "flower and fruit maker"	Comfrey, horse manure, seagrass	Fruit, flowers and potatoes	Sandy soils	Leaf edges turn yellow, then brown; low fruit yield; fruit and flowers poorly coloured	Apply a compound organic fertilizer or wood ash as a base dressing or top-dressing.	Excessive use of fertilizer; lack of magnesium
Minor plant nutrients						
Boron (B)	Beetroot leaves, horse manure, compost, seagrass, untreated sawdust	Root vegetables, top fruit (apples, pears etc.)	Sandy soils	Brown heart (roots); narrow leaves; corky patches on fruit	Top-dress with compost or apply borax in severe cases.	Irrigation water
Calcium (Ca)	Dandelion, lucerne hay, comfrey, horse manure, compost, blood and bone	Fruit, flowers, vegetables	Acid and potash-rich soils	Similar to nitrogen shortage – stunted growth and pale green leaves	Apply lime or use gypsum or calcified seaweed.	High pH causes toxicity for some acid-loving plants
Copper (Cu)	Nettles, yarrow, horse manure, dandelion, chickweed, compost, untreated sawdust	Fruit, vegetables	Sandy soils	Dieback; brown spots on leaves	Top-dress with compost.	Copper-based fungicides
Iron (Fe)	Stinging nettle, compost, dandelion, horse manure, spinach and seaweed	Rhododendrons, azaleas, camellias	Chalky soils	Yellowing of younger leaves	Trouble can occur on chalky, peaty, very light and acid soils. Iron shortage in chalky soils is due to lock-up by calcium and difficult to avoid.	Rare
Magnesium (Mg)	Grass clippings, seagrass	Roses, tomatoes	Sandy, peaty and potash-rich soils	Yellow or brown patches between veins of older leaves	Apply a mulch of compost or Epsom salts if deficiency is severe.	None
Manganese (Mn)	Chickweed, compost, untreated sawdust	Rhododendrons, azaleas, camellias	Chalky soils	Yellowing between the veins of older leaves	Shortage in chalky soils is due to lock-up by calcium (see iron). Acidification for some plants.	None
Molybdenum (Mo)	Cornstalks, compost, grass clippings	Brassicas	Acid soils	Narrow leaves	Apply lime and top-dress with compost.	None
Sodium (Na)	Seaweed	Root vegetables, fruit	Sandy soils	Very rare except in a few seaside-loving plants	Not usually warranted.	Land reclaimed from sea
Sulphur (S)	Cabbage leaves	All plants	Soils in rural areas	Stunted growth and pale green leaves	A light dusting of flowers of sulphur during the year.	Industrial sites and acid rain
Zinc (Zn)	Horse manure, corn stalks, garden compost, untreated sawdust	Fruit, vegetables	Sandy soils	Dieback	Top-dress with compost.	Leaching from containers

BASIC PRUNING

The task of pruning can seem a daunting prospect to some gardeners, but there is, in fact, no great mystery surrounding pruning and a few basic rules are all you need to get started in most cases. At its simplest level, pruning should be regarded as a way of introducing some order to your garden, giving your plants space to grow and look their best, while also controlling them within a set space. The following section provides some basic advice to enable you to prune with confidence.

Above: *Regular dead-heading keeps shrubs tidy and promotes further flowering. Cut flowering stems back to a bud or stem division.*

WHY IS PRUNING NECESSARY?

Pruning is usually carried out to control the growth of stronger subjects that may otherwise damage those of a weaker habit growing alongside; it also ensures plants do not outgrow their allocated space and cause an obstruction. Pruning is also used to train a plant to a specific shape or form; to maintain or improve flowering; and to improve the shape or "balance" of an individual shrub or tree. Finally, it is used to remove dead, damaged, diseased or pest-infected growth. This last type of pruning is referred to as hygienic pruning and helps maintain plant health and vigour.

Constant pruning, however, is not always necessary for the good cultivation of plants. Once a plant has become established and the initial framework has been formed, pruning is often reduced to a minimum. The major exceptions to this include plants that are pruned annually in order to improve flowering, fruiting or quality of foliage. Apples, peaches and other fruit trees and bushes are regularly pruned to increase the yield of fruit. Roses are a good example of an ornamental shrub that benefits from regular pruning, having more vigorous and prolific flowering as a result of cutting back on a yearly basis.

SUCCESSFUL PRUNING

There are arguably two hallmarks of good pruning: a well-planned, careful and methodical approach to the work undertaken and a good working knowledge of the requirements of the plants that are to be pruned. The real results may not always be apparent for months and, in the case of trees, for years.

With secateurs (hand pruners), always try to place the thin blade on the trunk/branch side of the cut. This will result in a short stub and will also prevent a layer of damaged cambium tissue being left immediately under the line of the cut.

GOOD AND BAD PRUNING

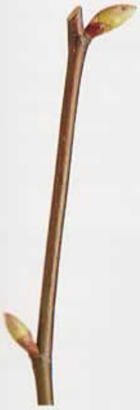

1 *A good pruning cut is made just above a strong bud, about 3mm (⅛in) above the bud. It should be a slanting cut, with the higher end above the bud. The bud should generally be outward bound from the plant rather than inward.*

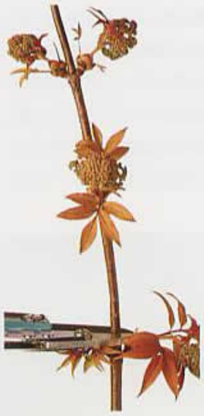

2 *If the stem has buds or leaves opposite each other, make the cut horizontal, about 3mm (⅛in) above the buds.*

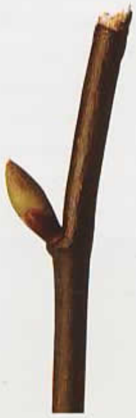

3 *Always use a sharp pair of secateurs (hand pruners). Blunt ones will produce a ragged or bruised cut, which is likely to introduce disease into the plant.*

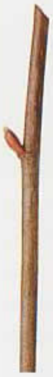

4 *Do not cut too far above a bud. The piece of stem above the bud is likely to die back and the rest of the stem may well die back even further, causing the loss of the whole stem.*

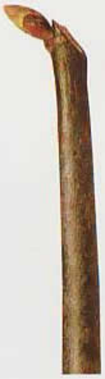

5 *Do not cut too close to the bud otherwise the bud might be damaged by the secateurs or disease might enter. Too close a cut is likely to cause the stem to die back to the next bud.*

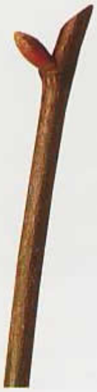

6 *It is bad practice to slope the cut towards the bud as this makes the stem above the bud too long, which is likely to cause dieback. It also sheds rain on to the bud, which may cause problems.*

HARD PRUNING SHRUBS

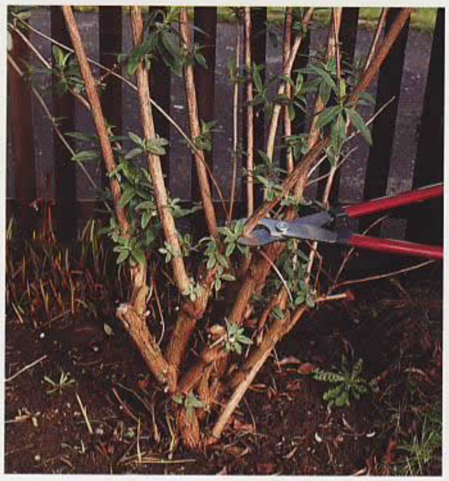

1 *There are a few shrubs – buddleias are the main and most obvious example – which benefit from being cut hard back each spring. This treatment improves the development of the foliage. Elders* (Sambucus) *and the purple smoke bush* (Cotinus) *should also be treated in this way.* Rosa glauca *also responds very well to this type of hard pruning.*

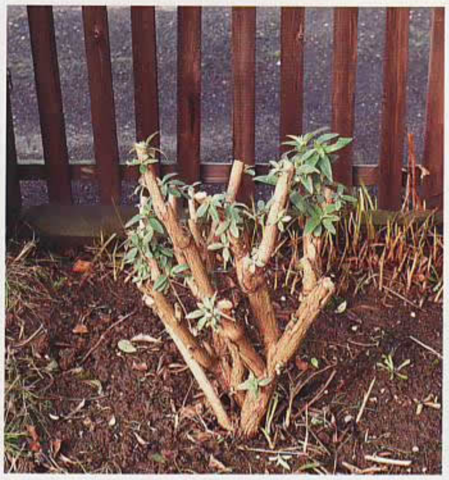

2 *Cut the shoots right back almost to the ground. It is important to ensure that you make the cuts just above an outward-facing bud. Cutting back will leave little more than a stump. It may seem a little drastic, but the shrubs will quickly grow again in the spring. If they are not cut back, they become very leggy and do not make such attractive bushes.*

3 *Several plants that have attractive coloured bark in the winter are best cut to the ground in the spring. So, by the following winter, attractive new shoots will be displayed. The various coloured-stemmed* Rubus, *such as* R. cockburnianus, *as well as some of the dogwoods* (Cornus) *and willow* (Salix), *are good candidates for this treatment.*

Cuts should always be made immediately above a bud and should be done as cleanly as possible. Use well-maintained secateurs that are regularly sharpened. Make sure that you do not position your blade too close to the bud or the tissue that connects the bud to the main plant body could be damaged. As a guide, you should cut 3mm (⅛in) above a bud and always ensure the bud is not damaged when cutting.

Be careful when using a saw close to a trunk or stem to remove heavier thicker branches. The weight of the branch can sometimes cause splitting and tearing to the main stem wood or bark. These injuries can allow the entry of disease-causing organisms and damage the plant.

The typical order of cuts to follow when pruning any tree is to remember the D's. These are: dead, dying, diseased, displaced or deformed branches. All of these must be removed as the first stage of any pruning work. Then add to this list the C's: criss-crossing branches and crowded growth.

The next stage is to remove any unwanted growth. Care must be applied at this critical stage to ensure that you do not remove too much growth. Finally, once you have cleared this material, it is possible to

Above: *When cutting out crossing stems, cut out the stems while they are still young and free from damage and disease. Using secateurs (hand pruners), cut the stem at its base where it joins the main branch.*

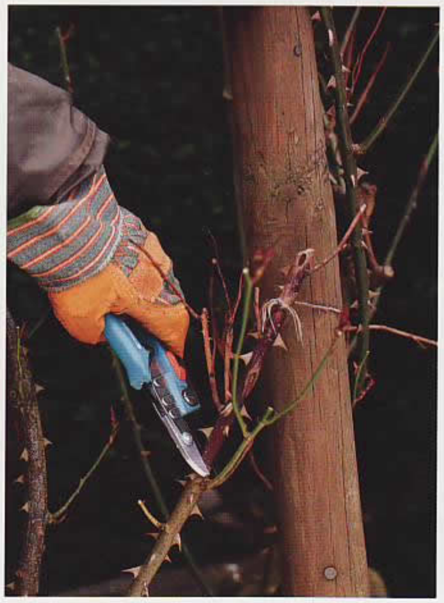

Left: *Tips of stems often die back, especially those that have carried bunches of flowers. Another cause is the young growth at the tip of shoots being killed by frost. If this dieback is not cut out, it can kill off the whole shoot. Even if dieback proceeds no further, the bush looks tidier without these dead shoots. Cut the shoot back into good wood, just above a strong bud.*

PRUNING OUT DEAD WOOD ON A CLIMBER

1 Most climbers produce a mass of dead wood that has to be removed so that the plant does not become congested. Dead wood is normally quite clearly differentiated from live wood by its colour and lack of flexibility.

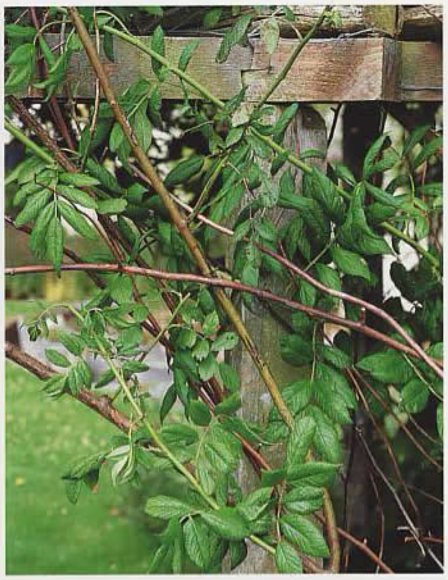

2 *Thin out all of the dead wood that the climber has produced, removing it very carefully in sections, if this is necessary, so that the remaining stems are not damaged when the dead stems are eventually pulled out.*

see what further work needs to be done. Many species have specific pruning requirements, which need to be carried out if the plant is to reach its full potential. Creating balance, form and a pleasing shape in the pruned specimen requires a keen eye, as well as an understanding of the plant's growth habits and tolerances. It is worth taking a little time over this and planning what you are going to do before starting. Acting too hastily now can easily spoil the job. Pruning is one of a few tasks in the garden that are best viewed intermittently from a distance. Remember, you can only cut a branch once. If you cut the wrong branch, you cannot rejoin it effectively to the plant from whence it came. Always check that you know the full extent of a branch and what effect its removal will have prior to making the cut.

Above: *Up to a third of the old wood should be removed to encourage the climber to produce new growth. If possible, cut some of this out to the base; also remove some of the upper stems, cutting them back to a strong growing point.*

Above: *Remove any wood that has become diseased, cutting it back to a point on the stem where the wood of the climber is healthy. If the cut end shows that the wood is still diseased on the inside of the stem, cut back further still.*

PRUNING EQUIPMENT

There is a range of tools that can be used for pruning. It is advisable to use the correct tool for a specific task because it will make the process both easier and more efficient.

Pruning knife Using a pruning knife requires both skill and a sharp blade. A blunt knife can easily damage the plant. Cut away from you and keep hands clear of the direction of the cut.

Secateurs (hand pruners) When small branches are to be removed, it is easiest to use secateurs and remove the branch in one go. Secateurs are the commonest type of pruning tool and are best used on twiggy material up to about the width of a pencil.

Loppers Use these on branches up to 2.5cm (1in) thick. Loppers are useful in situations where secateurs may not easily reach, such as in dense tangled shrubs. Loppers should not be used for very thick branches or they may easily be damaged.

Extension loppers Ideal for cutting out-of-reach branches up to 2.5cm (1in) thick. They are a useful tool but they can be tiring to use and require some practice.

Ratchet and snap cut shears These have three times the power of conventional shears. Used for branches up to 4cm (1½in) in diameter.

Narrow curved Grecian saw Use on branches up to 5cm (2in) in diameter where the branches are very crowded. It cuts on a pulling cut, making its use above shoulder height very easy.

Extension saws These are similar to Grecian pruning but they have an extended handle. They are very useful where larger branches need removing from smaller garden trees.

Bow saw Used for cutting larger branches. Available in various sizes. They cut with a pushing cut, making their use above shoulder height tiring.

PLANT PROPAGATION

Despite the large variety of methods employed, there are really only two basic types of plant propagation: vegetative – taking cuttings, for example – and sowing seed. All of the techniques used fall into one or other of these categories. Raising your own plants is one of the most rewarding activities that any gardener can do. It can also be particularly beneficial for the organic gardener because, unlike with bought-in plants, you can be assured that self-raised stock will be pesticide-free.

Above: *Collecting your own seed can be very rewarding and ensures that your new plants are free of any pesticide residues.*

POTTING MIX FOR PROPAGATION

Seed and cutting potting mix should be almost free of nutrients – the plants do not need soil-derived nutrients until they have rooted. Numerous proprietary brands are available for organic propagation. Alternatively, you can make your own potting mix using finely sifted compost or leaf mould mixed with equal parts of sharp sand or grit. Mixing this is comparatively easy but problems can arise if the organic matter is not thoroughly sterilized. Unless you are able to do this, there is always the possibility of fungal diseases affecting the plants. Worm compost can be good for making potting mix but the easiest way to get sterilized potting mix is to use compost that was made in a hot pile. Use material from near the middle of the pile to get the most sterilized matter.

PROPAGATION FROM SEED

Seeds are nature's way of introducing variety into plant material. Fertilizing the female part of a flower with pollen from the male part produces a seed. Seeds are a resting and survival stage in the plant's life that help it survive adverse conditions in nature. Seeds can survive conditions that kill plants. A seed consists of three things: an embryo (a young plant at its most immature stage); a food supply (to maintain the embryo and provide energy after germination); and a protective seed coat. The embryo consists of a young root system (radicle); the young shoot system (plumule), which carries the seed leaves (cotyledons); and the hypocotyl, which is the junction between the root and shoot systems.

Some seeds need special treatment to germinate. One such treatment is stratification. To stratify seeds, cover them in some damp material, such as damp compost, and keep them in a refrigerator at about 4.5°C (40°F) for up to three months. After this treatment, the seeds are ready to germinate. Seeds that germinate better after stratification are oaks and hellebores.

Seeds require moisture, warmth and oxygen to germinate. Because a seed needs oxygen as well as moisture, it is important not to keep the seeds too wet. On the other hand, the seeds must not dry out. Most seeds germinate best between 18.5–24°C (65–75°F). Most seeds will germinate well in light; however, some seeds, such as those of periwinkle, pansy and verbena, germinate best in the dark.

Many seeds have specific germination requirements that require knowledge of the particular species. Certain generalizations are possible and although it is always advisable to follow the instructions on packets you buy, self-collected seed may require a bit of guesswork. If it is a small dust-like seed (thyme and basil, for example), it probably needs light to germinate. Make sure the seed is well pressed into the moist mixture, but do not cover. Larger seed should be covered to twice its thickness in the soilless mixture. If the seed packet says to scarify, make a scar in the seed by nicking the seed coat in some way. Lightly rubbing the seeds with sandpaper usually accomplishes this. The reason for scarification is to speed up the lengthy process of letting the water past the waterproof seed coat.

MAKING ORGANIC PROPAGATION POTTING MIX

1 *Take equal quantities of an organic material, such as coir or finely sieved leaf mould, and a free-draining material such as vermiculite or perlite.*

2 *Mix the organic material and vermiculite or perlite together thoroughly, ensuring that there are no lumps and that the mixture is open and "free running".*

3 *This organic potting mix for propagation is now ready for use and does not need any additional fertilizer. It can be used to propagate seeds and cuttings.*

SOWING SEEDS IN CONTAINERS

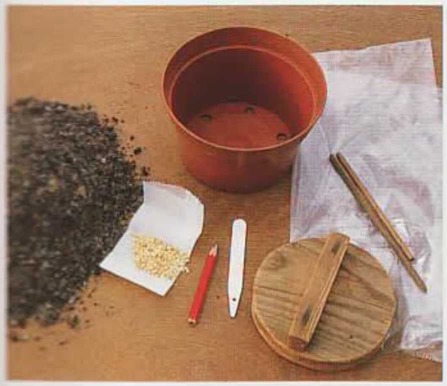

1 *First gather together the basic equipment. You will need pots, potting mix, labels, a pencil, a firming board and a plastic bag.*

2 *Fill the container with the potting mix, gently tapping the pot to ensure that there are no air spaces in the mix.*

3 *Gently firm and flatten the potting mix with the firming board. Take care not to over-firm the potting mix.*

4 *Sow the seed evenly over the potting mix surface. Large seed can be placed, but smaller seed must be scattered.*

5 *With larger seed, lightly cover the seed with finely sieved potting mix. Sieved vermiculite or perlite can also be used.*

6 *Water the pot using a fine rose on a watering can. If the seed is small, water the potting mix prior to sowing.*

VEGETATIVE PROPAGATION

The vegetative parts of the plant (stem, root or leaf) can also be used to produce a new plant. Most fruit trees are propagated asexually, using a bud or a twig from a tree that produces exceptionally good fruit. When this bud or twig becomes an adult tree, it has the same qualities as the "mother" tree. By propagating asexually (cloning), we reproduce the "mother" plant.

PROPAGATING PLANTS FROM SPECIALIZED ORGANS

Plants may be propagated from many different plant parts. Specialized roots, stems and leaves that occur naturally make asexual propagation easy.

Bulb This consists of swollen leaves on a short stem. Bulbs can be propagated by removing small bulblets or offsets that form at the base of the parent bulb. These small bulbs take two or three years to mature into plants that flower. Place offsets in rich light soil. Certain bulbs, such as lilies, can be propagated by removing individual scales from the dormant bulb and placing them in a bag of damp moss. After a few months the scale will develop into a small bulb (a bulbil), which can be potted as if it were a large seed and grown on. Certain lilies (tiger lilies, for instance) produce bulbils in their leaf axils (where the bulb leaf joins the stem) that can also be propagated. Examples of bulbs are tulips, onions and lilies.

Corm This is similar to a bulb and often confused with one. Structurally, however, a corm is different, consisting of a stem that is swollen as a food store. It is shorter and broader than a bulb. The leaves of the stem are modified as thin dry membranes that enclose the corm and protect it against injury and drying. Examples of corms are crocus and gladiolus. The procedure for propagating these is the same as for taking offsets from bulbs.

Rhizome This is a stem that grows horizontally near the soil surface. A rhizome usually stores food, but, as it grows, it develops buds along its length. Rhizomes can be cut into sections with at least one eye or bud. Plants propagated in this way include iris.

Runner This is a stem that arises from a crown bud and creeps over the ground. The plantlet that forms at the tip is easily rooted and forms a new plant. Plants with runners include strawberries and spider plants.

Tuber This is a swollen underground stem or root that stores food. Tuberous plants, such as potatoes and dahlias, can be dug up and the tubers separated. Each section must have a segment of the crown that contains at least one eye or bud.

PROPAGATING TUBERS

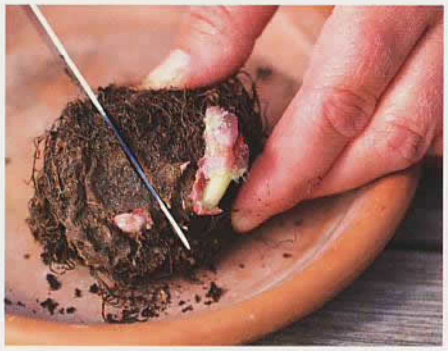

1 *Take a sharp knife and cut the tuber (in this case a begonia tuber) through the middle so that each part has a shoot.*

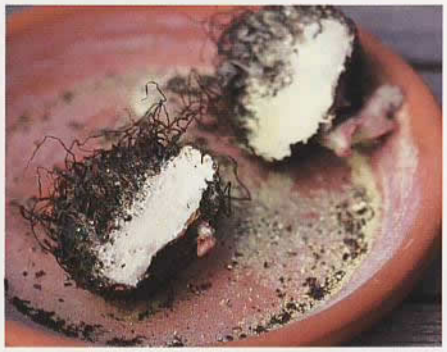

2 *Dust the open cuts with a fungicide such as sulphur or Bordeaux mixture. Prepare two small pots with moist potting mix.*

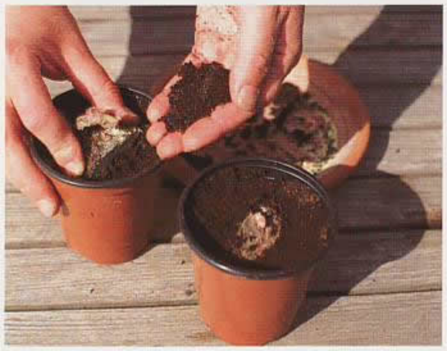

3 *Plant both halves so that they sit firmly on top of the potting mix. Keep the potting mix moist and plant on when new growth appears.*

ROOT CUTTINGS

To produce a new plant from a root cutting, there must be a shoot bud present or it must be possible for the cutting to form one. The ability of root cuttings to form these buds depends on the time of year. The dormant (resting) season is usually the best time to take root cuttings and it is best to take them from newer root growth.

Make cuttings 3–10cm (1¼–4in) long from roots that are 1–1.5cm (⅜–½in) in diameter. Ensure the roots are from the chosen plant and not neighbouring plants. Cuttings are best taken during the winter, when roots have large carbohydrate supplies, but they may also be taken throughout the growing season. Cut straight through the end of the root closest to the stem. Cut the other end on a slant. This will help you to remember which end is the top (the straight cut) and which is the bottom (the diagonal cut).

Store cuttings from dormant roots for three weeks in moist rooting medium at 5°C (41°F). After this, remove from storage and plant upright in the growing medium. Keep moist and warm in a bright location until the plant has grown large enough to pot on or plant out. This will, of course, vary depending upon the vigour of the plant and slow-growing plants may be best potted up and grown on for a season before finally planting out the following spring.

If root cuttings are taken during active growth, miss out the period of storage and place the cuttings directly in the rooting medium. Phlox and euphorbias are propagated from root cuttings.

STEM CUTTINGS AND LAYERING

Many trees, shrubs and herbaceous plants are propagated from stem cuttings. A root system must be formed on a stem either before or after the stem is removed. Roots can be formed on stems in two ways: by layering and by taking stem cuttings. With layering, the stem is allowed to produce roots before it is cut from the parent plant.

Stem cuttings form roots after the stem is removed. The main difficulty with stem cuttings is keeping the stems alive while they form new roots. Some plant stems root better when the wood is soft and actively growing, others root best from mature wood. Cuttings taken from plants that are actively growing are called softwood cuttings, while those taken after the wood is mature are known as hardwood cuttings.

TAKING ROOT CUTTINGS

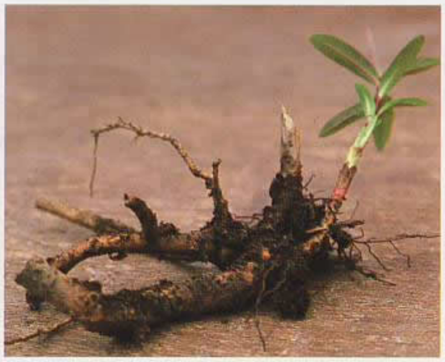

1 *Dig up a small section of root from a suitable species with thick fleshy roots. The section must either have a shoot bud or be capable of forming one – in which case the cutting should be taken in the dormant season – if it is to be successful.*

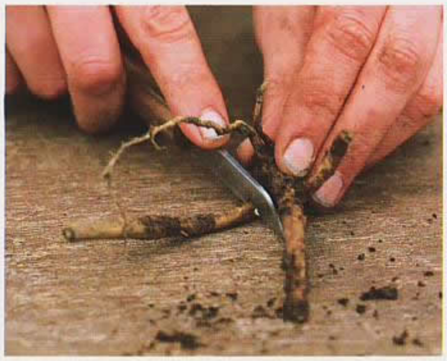

2 *Remove as much of the soil as possible from the root cutting before slicing off the small side roots. Using a sharp knife, remove these as well as dead and damaged portions and any side shoots to leave a section of healthy root.*

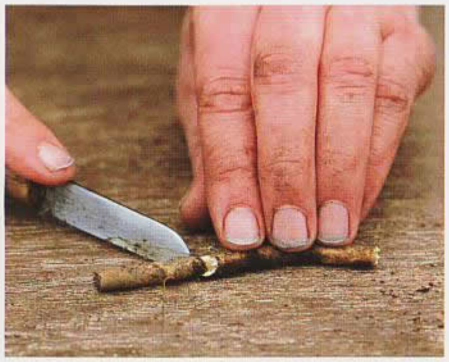

3 *Cut the remaining sections of the root into small pieces, approximately 5cm (2in) in length, cutting the bottom end at an angle. Insert the root cutting into propagation potting mix, with the angled end pointing down.*

TAKING HARDWOOD CUTTINGS

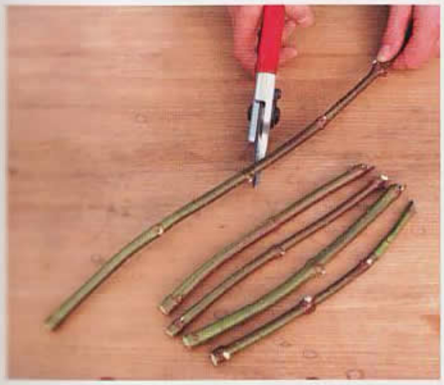

1 *Select healthy, blemish-free pieces of wood about the thickness of a pencil. Cut each of these into sections about 20–25cm (8–10in) long, angling the top cut.*

2 *Stand the cut stems in a jar of water until you are ready to plant them out. Ideally, stand the jar in a cool moist place out of direct sunlight.*

3 *Dig a narrow trench that is deep enough for the cuttings. About 2.5–5cm (1–2in) should emerge above ground. The length of the row will depend on the number of cuttings.*

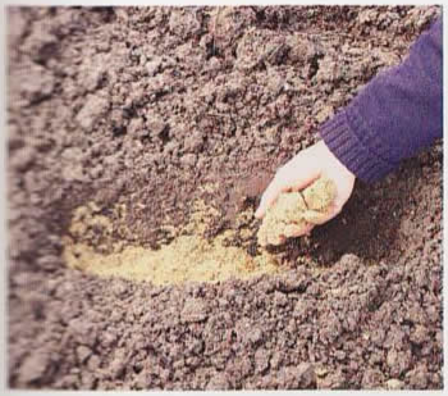

4 *Loosely fill the trench to about two-thirds full with sharp sand or fine grit. This is essential as it will allow the passage of air around the bases and prevent rotting off. It will also encourage rooting.*

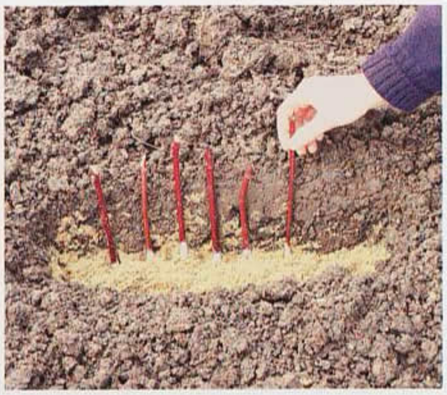

5 *Insert the cuttings about 10–15cm (4–6in) apart, making sure that the angled cut is uppermost. If you place any cuttings in upside down – i.e. with the buds facing downward – rooting will not occur.*

6 *Gently firm the soil, making sure that the cuttings are firm but that the soil surface is not over-compacted. Leave the cuttings for the whole of the growing season before lifting and planting out next winter.*

TAKING STEM TIP CUTTINGS

1 *Take cuttings from the tips of the stems and put them in a plastic bag. The length of the cuttings will vary, depending on the subject, but take about 10cm (4in).*

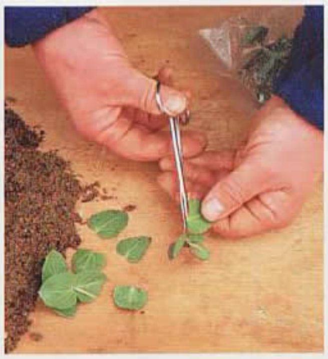

2 *Trim the cuttings to just below a leaf joint and then remove most of the leaves and side shoots, leaving just two at the top. This will help to prevent stem rot.*

3 *Place up to twelve cuttings in a pot of potting mix that is specially formulated for cuttings or a 50:50 mixture of sharp sand and peat substitute such as coir or leaf mould mix.*

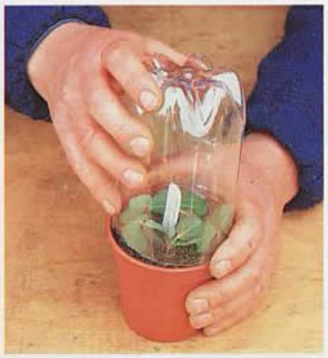

4 *Water well, and cover with the cut-off base of a soft-drinks bottle, a perfect substitute for a propagator. A heated propagator will speed up the rooting process. Place several pots in the same unit.*

Layering is a useful way of increasing some shrubs and climbers that cannot readily be propagated from cuttings. Simple layering is a method of getting a growing shoot to produce roots while it is still attached to the "parent" plant. There are several variations on the technique, and the one you use will depend on the plant and the type of growth it produces.

Softwood cuttings are taken from first-year branches that have not yet become woody. Flowering shrubs are often propagated by softwood cuttings. Late spring and early summer are the best times for success with this method. Take cuttings 5–10cm (2–4in) long. Larger cuttings produce larger plants sooner, but they are prone to more rapid water loss. Make cuts slightly below a leaf node. Remove any leaves on the lower section and insert them into potting mix, making sure that no leaves are touching each other or the potting mix. Remove any cuttings immediately from the tray or pot if they die or appear diseased. Pot healthy cuttings up promptly once they recommence growth following rooting.

Hardwood cuttings are taken once the tissue becomes woody and the plant is dormant. Cuttings can be taken two weeks after leaf fall and before bud burst. Select healthy wood that was produced the previous summer. The wood should be about pencil thickness and cut into sections of approximately 20–25cm (8–10in). Several cuttings can be made from the same branch of some shrubs.

To take hardwood cuttings, make basal cuts just below a node, and upper cuts slightly above a bud. The upper cut should be slanted and the lower cut straight, so that a cutting is less likely to be inserted into the potting mix upside down. Once inserted all the tops of the cuttings should be slanted cuts. Bury cuttings vertically in moist sandy topsoil or sand.

The cuttings should not freeze, but must remain cool. In spring, remove the cuttings from storage and plant in a hotbed or other protected site with exposure to morning sun or filtered light. Leave 2.5–5cm (1–2in) of cutting above ground. Keep cuttings moist until a root system forms. Transplant the cuttings the following spring while they are still dormant.

LAYERING SHRUBS

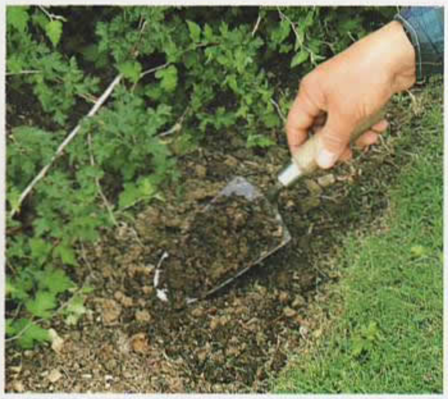

1 *Choose a stem that will reach the ground without breaking and prepare the soil beneath it. In most cases, the native soil will be satisfactory, but if it is heavy clay, add some potting mix to improve its texture.*

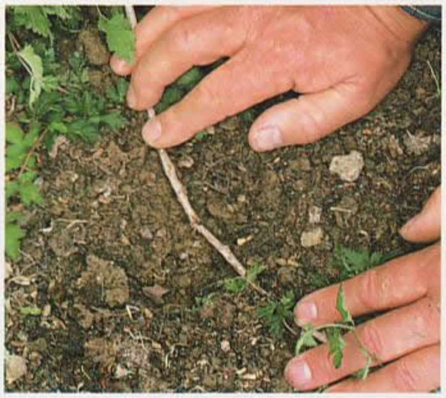

2 *Trim off any side shoots or leaves. Dig a shallow hole and bend the shoot down into it.*

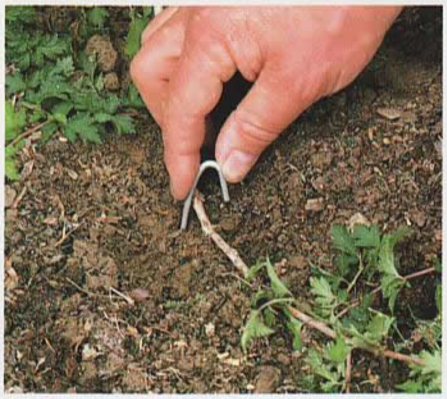

3 *To help hold the shoot in place, peg it down with a piece of bent wire.*

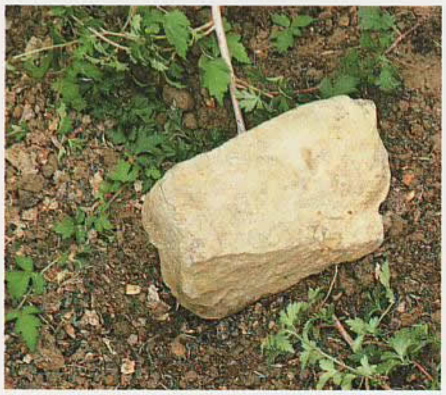

4 *Fill in the hole and cover it with a stone. In many cases, the stone will be sufficient to hold the layer in place and a peg will not be required. The stone will also help to keep the area beneath it moist.*

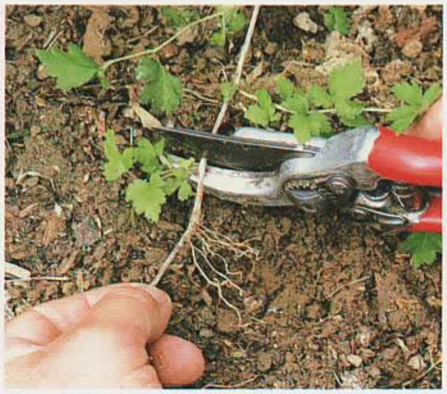

5 *It may take several months or even years for shrubs that are hard to propagate to layer but, eventually, new shoots will appear and the layer will have rooted. Sever it from its parent and pot it up into a container.*

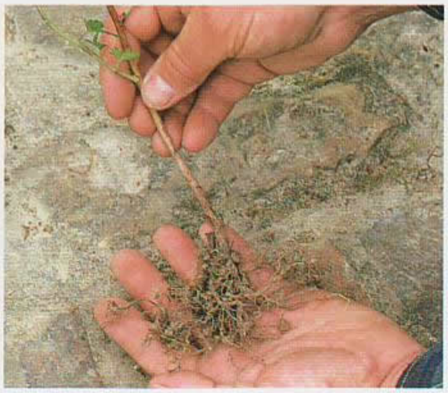

6 *If the roots are well developed, transfer the layer directly to its new site.*

SIMPLE DIVISION

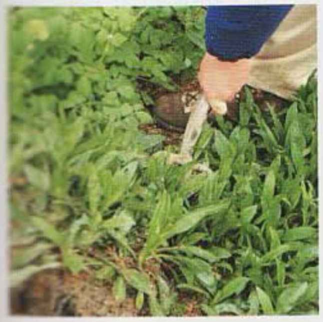

*1 Water the plant that is to be divided during the previous day. Carefully dig up a clump of the plant using a spade, in this case a Michaelmas daisy (*Aster novi-belgii*).*

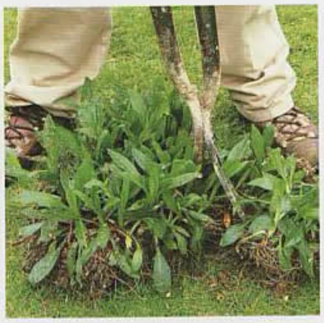

2 Insert two garden forks back-to-back into the plant and lever apart by pushing the handles together. Keep on dividing until the pieces are of the required size.

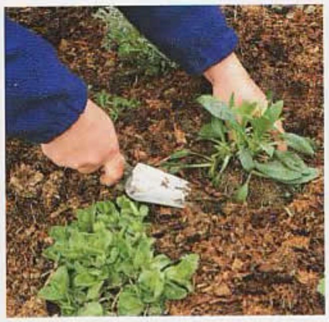

3 The pieces of the plant can then be replanted in the bed, but dig over the soil first, removing any weeds and adding some well-rotted organic material.

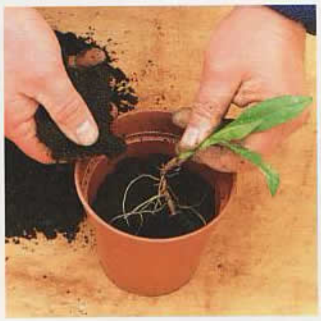

4 Alternatively, small pieces of the plant can be potted up individually. After watering, place these in a closed cold frame for a few days before hardening off.

DIVISION

Division is the cutting or breaking up of a crown or clump of suckers into segments. Each segment must have a bud and some roots to propagate successfully. These segments are replanted and grow into new plants identical to the parent. Most perennials should be lifted and divided when they become overgrown and begin to lose vigour. Vigorous growth in most perennials occurs on the outer segments of the clump. Old growth in the centre of the clump is discarded.

Carefully dig up the plant, loosening the roots and lifting it from the soil. Split apart the main clump with two spades or forks, or chop with a shovel or large knife if the clump is firmly massed. It is advisable to divide autumn-flowering perennials in spring, while those that flower in spring and summer are best divided in autumn.

GRAFTING AND BUDDING

Grafting and budding involve inserting or attaching a part of one plant on to another so that both parts continue to grow. In grafting, you attach a short twig from a desirable variety with two or more buds (called a scion) to a seedling. This seedling is called the stock and forms the root system for the new plant, while the scion becomes the top growth. Budding involves inserting the bud from a desirable variety into a cut or slit in the bark of the stock. The bud develops into the whole top of the plant. You must match scions and buds with stocks according to their ability to grow together.

CHIP BUDDING

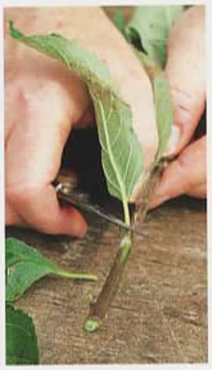

1 Select good bud wood that is free from blemishes and obvious damage, pests or diseases. If you are doing this in the summer, then you must first remove the leaves.

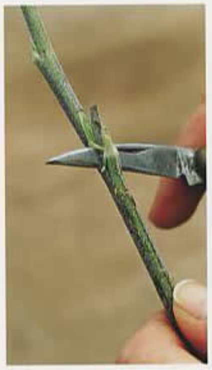

2 Remove the bud using a very sharp, flat-edged knife. Cut out the bud with a thin shaving of the wood below the bark and leave a "V" shape at the base.

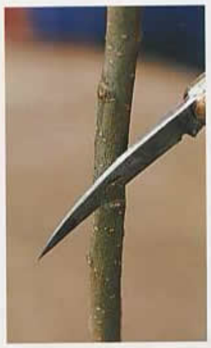

3 Prepare the stock plant by cutting a "church window" shaped notch from the stem that is the same thickness as the chipped bud and leave a small notch cut into the base.

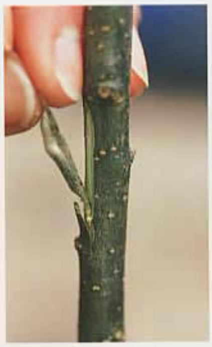

4 Place the bud into the notch, fitting the "V"-shaped base into the notch at the base of the "window cut".

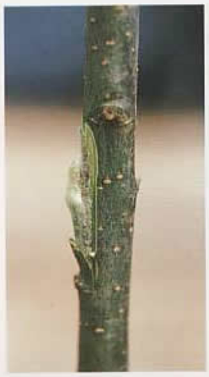

5 Make sure that the bud fits snugly, with the edges of the bud-wood in contact with the edge of the "window" notch.

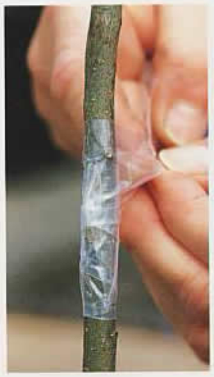

6 Secure the graft firmly with grafting tape to ensure that the join does not dry out. The graft will usually take in about 4–6 weeks in the summer but winter grafts can take longer.

THE GREENHOUSE ENVIRONMENT

Greenhouses provide an environment in which levels of light and shade, temperature and moisture, and protection from the elements can be closely controlled and monitored. You can propagate plants to grow outside, protect tender plants from frosts and grow certain crops out of season. A basic understanding of the ways that the greenhouse environment can be controlled will help you on your way to a bumper harvest.

Above: *Shading helps to keep temperatures down in the summer and also protects the plants from sun scorch.*

LIGHT INTENSITY

Natural light has all the right wavelengths that plants need to function and grow successfully. Light intensities vary considerably during the year and gardeners often try to compensate for this by giving extra light to the plant. The quality of this light must be correct for the growth of plants, and light bulbs used to supplement natural light must be of a suitable daylight quality. Natural winter light intensity is often insufficient for certain crops. Conversely, summer light levels, especially when combined with elevated temperatures, can cause problems for other crops. Balanced growth requires good greenhouse design and proper planning for the light requirements of the crop. As a general rule, the ridge that runs along the middle of your greenhouse (if yours is a traditional "A"-shaped one) should always be orientated so that it runs east to west.

Above: *A maximum/minimum thermometer is an essential tool in helping you to judge the conditions in your greenhouse.*

LIGHT READINGS WITH A METER

These should be taken inside the greenhouse and then compared with an outside reading. The difference between the two readings will indicate the percentage of available light transmitted into the greenhouse. A good tip here for those without a professional light meter is to use the automatic aperture reading on an SLR camera or a photographic light meter. It will not give actual figures, but will indicate to a certain extent just how much of the available light is getting through.

ADDITIONAL LIGHTING

Supplementary lighting can be used to "top-up" existing daylight, especially in winter. You should always buy bulbs that give a full spectrum of light (the same quality as daylight – often sold or described as "grow lights"). Almost all bulbs for household use are short of the particular light wavelengths that plants need to produce food and grow successfully. Grow lights tend to use a lot of electricity and it may be cheaper to provide additional light for just a part of the greenhouse, suspending the lights above a crop that really needs the light. Grow lights are available as bulbs or strip lights and can sometimes be bought as special enclosed units.

SHADING

This controls temperature and light. Shading is usually put in place in order to reduce summer temperatures (if they are excessive) or to protect young plants and shade-loving ones from very intense sunlight. Blinds, netting or white shade paint can all be effective.

TEMPERATURE

Almost all plants like a fall in night-time temperature of about 4–6°C (7–10°F) below that in the day. Many commonly grown species have a relatively wide range of tolerances, but most grow best between about 16–24°C (60–76°F). All plants have maximum and minimum temperature tolerances that should never be exceeded.

You can take an average temperature reading inside the greenhouse using a maximum and minimum thermometer. Compare it with one taken outside. You should note, however, that more than just the outside temperature influences that of a greenhouse. The action of the sun, even on a relatively cold day, can result in temperatures being much higher inside than out. Wind is also a factor to consider that can cause a chilling effect far below the temperature average, rapidly cooling the glass and air nearby.

HEATING SYSTEMS

Most heaters depend upon a hotbox or burner. These work by heating the air directly. They are more generally used

Above: *Some automated systems measure air temperature, and then turn on heat and open vents as necessary.*

Left: *Even a relatively small greenhouse can vastly increase the range of plants that you can grow in your garden.*

where there is no need to provide heat on a regular basis – in cold greenhouses, for example, or in polytunnels used for raising early hardy stock. Some enterprising gardeners link hot water pipes to their domestic heating systems to provide heating for the greenhouse. The disadvantage here is that the heat tends to be applied at times to suit the household but not the plants within the greenhouse.

INSULATION

It is possible to insulate a greenhouse by providing additional layers within the structure to trap air. Bubble-wrap plastic is a popular method as it is easy to place in position and allows for the transmission of light. Such methods always reduce the amount of light transmitted, but will save upon heating costs. In addition to this, careful siting of the greenhouse can help to reduce the chilling effects of high winds.

HUMIDITY

Most plants thrive in about 40 to 60 per cent humidity. Plants that are grouped together often set up their own small microclimate where the humidity is higher in the immediate vicinity. Much of the humidity requirement for a plant depends upon where it comes from. Rainforest plants often need 85 per cent or more humidity, whereas cacti need as little as 15 per cent (in cultivation). The humidity is also affected by other factors, notably temperature. Warm air can hold much more moisture than cooler air.

The need to increase humidity rises during the hotter months when it is harder to keep down the temperature in protected structures. Increasing the humidity also helps to slow down the water loss from leaves and may also help to prevent wilting or scorch.

Humidity should be raised by damping down (spraying a hose on hard surfaces, paths or under benches) on hot dry days. If there are only certain plants, such as young seedlings, that require elevated humidity, then these are best covered with a thin layer of clear plastic.

AIR CIRCULATION

All plants need a constant supply of clean fresh air around them in order to grow healthily. Outdoors, normal wind movements would supply this, but in a greenhouse you will need to make provision for this.

Ventilators help control temperature and air flow. Roof vents should normally occupy the equivalent of 15 per cent of the floor space in order to be fully effective. They are normally a single structure that opens "window fashion" with a safety catch to keep them open to the required width. As with greenhouse design in general, there are many designs, most being simple variations on a theme. Side ventilators are also occasionally seen in some greenhouses. These are usually arranged in a "louvre type" fashion, with several overlapping strips that can be opened or closed with a small lever.

CHECKING THE CONDITIONS

The greenhouse microclimate differs from the garden chiefly because the light intensity is lower, the temperature is higher and air turbulence is less. How you modify the environment largely depends on what you intend to grow. Winter salad crops need some additional heat, particularly on cold winter nights. Slightly tender, overwintering perennials may not need any heat unless the temperature drops very low. You must know what the environment will be even on cold winter nights. The only really accurate way to assess the suitability of an environment for plant growth is to take regular readings. Many devices exist to monitor the internal conditions within protected structures. These include thermostats for heating and automatic ventilators, although a seemingly endless array of humidification, automatic shading and watering devices are now available for small greenhouses.

GREENHOUSE HYGIENE

Practising good hygiene is important in maintaining the health of the plants. Dead leaves and debris left lying about can harbour pests and diseases, as can badly affected plants if they are not removed promptly. Once a year, thoroughly clean all surfaces, including glass and glazing bars, with hot soapy water. This will pay dividends during the growing season by reducing the number of annually re-occurring pests and diseases. Where plants are grown in open borders, rotate crops or replace the soil with fresh topsoil every two or three years.

Above: *Bubble-wrap plastic is an ideal material with which to insulate the greenhouse in the winter.*

GREENHOUSE CULTIVATION

In a greenhouse you can hasten the arrival of spring by forcing hyacinths, tulips, azaleas and a whole variety of spring crops. You can also lengthen summer, with roses that bloom far into the autumn, and brighten winter with carnations, camellias and rare tropical flowers. You may also harvest tomatoes in mid-winter and vegetables or greens at any time of the year.

Above: *Early crops and flowers can be raised in pots ready for planting out when the weather is warmer.*

METHODS OF GROWING

Think carefully about how you plan to use your greenhouse and what plants you would like to grow there. The true art of greenhouse gardening lies in using every available area of space. This will involve planning the layout of the growing areas in the greenhouse and making a cropping plan to ensure that you not only use all of the available space, but also every "time space" within the growing calendar.

GROWING IN OPEN GROUND

Many greenhouse plants will grow well in open soil borders. The soil can be prepared in a similar way to that used for borders outside. Remember that much of what benefits outdoor soil (digging and adding organic matter) is also good for indoor soil. The advantage is that the system is sustainable over a long period and can even be serviced as a small raised bed system to help maintain steady soil temperatures. Taller crops, such as tomatoes, are particularly well suited to growing in open borders in the greenhouse, but a whole range of both ornamentals and edible crops can also be grown in this way. Remember that if you intend to grow certain crops indoors – salad crops, tomatoes, legumes or cucumbers, for example – then you may need to rotate these, or in smaller spaces avoid growing them in open ground every second or third year. This is done for the same reasons that outdoor crops are rotated.

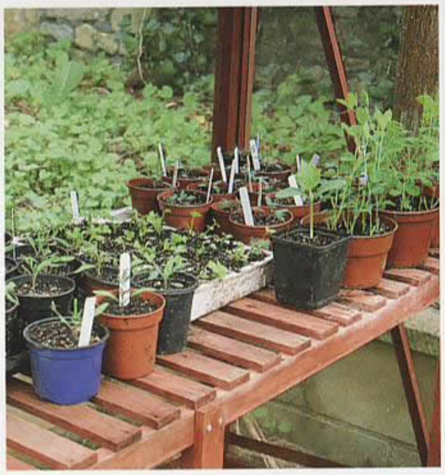

Above: *Staging can provide extra space in the greenhouse as well as a comfortable surface on which to work.*

Left: *You can grow crops in soil beds as well as in pots in the greenhouse.*

GROWING IN POTS ON STAGING

Staging is essentially a shelf or shelves on which plants in pots are arranged. It enables the gardener to work at a comfortable height and is particularly useful for those who wish to raise seedlings and propagate plants. The area beneath the staging is somewhat shaded but by no means wasted as it can be used for overwintering dormant plants, growing ferns and other shade-lovers, and for some forms of propagation (lilies propagated from scales, for instance). It also provides the perfect place for propagators and can easily be adapted to become a propagating tent.

USING GROW BAGS

Grow bags offer a good alternative to open beds in smaller greenhouses, where space is limited and no open beds exist. These can be purchased or you can easily make your own by filling a large plastic potting-mix bag with a mixture of garden compost and a little grit. Each bag will support four tomato plants and even greater numbers of smaller plants. Grow bags are ideal for raising sweet peas (*Lathyrus odoratus*) for use on patios or roof gardens where the rich deep soil in which they thrive is unavailable. They can equally be used to raise garden peas or beans.

GROWING CLIMBING PLANTS

The last dimension in the greenhouse is the vertical one, ultimately to the roof. Lean-to greenhouses are especially useful for this, their back walls being well suited for growing such things as vines or kiwi. Even freestanding "apex" houses (those shaped like an "A" with a central ridge) can be used to raise climbing crops, such as cucumbers or melons, to say nothing of the thousands of delightful and colourful ornamental climbers that thrive in the greenhouse.

A YEARLY GREENHOUSE CROPPING PLAN

The plan shown below is a sample of how a greenhouse can be used to be productive all year round. The essential point is that you plan what you want to grow. Once you have decided on this it is a relatively simple matter to make your greenhouse a productive space.

GREENHOUSE TASK	MID-WINTER	LATE WINTER	EARLY SPRING	MID-SPRING
Edible crops	Sow seeds for early crops. Interplant with radishes. Sow early cabbage and cauliflower in pots. Bring strawberries in from frames. Top-dress vines and fruiting climbers.	Sow broad beans. Pot on early vegetable seedlings. Reduce young shoots of vines to one to two per spur and raise lowered rods back to the greenhouse side.	Sow the seeds of Brussels sprouts. Sow indoor tomato seed and make sowings of calendula, tagetes or similar in order to encourage beneficials and pollinators. Prick out seedlings.	Sow seed of outdoor tomatoes, celery, celeriac, melons and cucumbers. Harden off early vegetable plants. Tie new growth from vine rods on to wires. Remove side shoots of early tomatoes.
Ornamentals	Take cuttings of plants such as chrysanthemums. Bring bulb pots in for forcing. Pot on autumn-sown sweet peas into "long" pots.	Start off dahlia cuttings. Cut back and re-pot fuchsias and pelargoniums. Sow sweet peas (*Lathyrus odoratus*). Pot up cuttings of herbaceous plants. Sow herbaceous seed and begin sowing bedding plants.	Harden off early sweet peas. Pot up cuttings and sow seed of half-hardy annuals. Take cuttings. Prick out seedlings. Pot up summer bedding tubers and corms.	Finish sowing half-hardy bedding. Pot up cuttings and prick out seedlings. Make up summer tubs and hanging baskets for growing on.
General	Wash glass panes of greenhouse and frames. Keep conditions on the dry side. Be vigilant for winter pests and diseases. Provide heat as necessary.	Keep an eye on ventilation. Watch for pests and diseases. Keep conditions on the dry side. Remove dead or dying material from plants. Provide heat as necessary and ventilate.	Damp down on hot days and ventilate. Watch for pests and diseases. Begin to feed plants as growth increases. Dig, manure and top-dress borders ready for planting.	Damp down at least twice daily when hot and ventilate. Continue to watch for pests and diseases. Dig borders and manure if you have not already done so.

GREENHOUSE TASK	LATE SPRING	EARLY SUMMER	MID-SUMMER	LATE SUMMER
Edible crops	Plant greenhouse tomatoes into prepared borders or large pots of loam-based potting mix. Interplant with calendula and tagetes to encourage bees into the crop. Prick out celery and celeriac cuttings.	Thin grape trusses using vine scissors. Thin fruit if necessary on peaches growing under glass. Pick tomatoes regularly as they ripen. Pinch out and tie in melons and cucumbers as they reach their intended size.	Hand pollinate melons and cucumbers. Perform last thinning on grape trusses. Feed tomatoes with dried blood to develop uppermost trusses.	Remove lower leaves of tomatoes as fruit ripens. Support melons, removing leaves that shade the fruit. Harvest cucumbers. Sow seeds of winter leaf crops and companion plants.
Ornamentals	Harden off bedding plants. Place winter-flowering cyclamen in shade and allow to partially dry out.	Take cuttings of deciduous shrubs. Propagate houseplants. Sow seeds of spring bedding plants in frames.	Take cuttings of evergreen shrubs and deciduous species. Propagate houseplants. Sow seeds of biennials in frames. Re-start rested cyclamen corms.	Take cuttings of pelargoniums and fuchsias to act as "mother plants" for cuttings the next season.
General	Provide shade over sensitive crops. Feed plants regularly with a liquid "tea". Watch for pests and diseases and introduce biological control.	Water regularly. Continue potting on developing cuttings and seedlings of all plants.	Water regularly (twice daily in hot weather) and damp down two or three times daily to maintain humidity.	Continue potting on developing cuttings and seedlings. Continue to watch for pests and diseases and introduce biological control if necessary.

GREENHOUSE TASK	EARLY AUTUMN	MID-AUTUMN	LATE AUTUMN	EARLY WINTER
Edible crops	Cut melons as soon as they are ripe. Clear tomatoes as they finish.	Plant up winter leaf and salad crops, interplanting these with an appropriate companion planting.	Lift crowns of rhubarb for forcing and leave them on the surface for a week.	Sow seed of early onion varieties. Lift further crowns of rhubarb for forcing. Prune peach trees that are growing under glass.
Ornamentals	Take last semi-ripe cuttings and softwood cuttings.	Harden off spring bedding plants ready for planting.	Move bulbs for forcing into a frame to encourage top growth. Lift and pot up lily-of-the-valley for spring display.	Bring in bulbs for forced flowers. Stool back "mother" plants for early cuttings.
General	Decrease watering as days shorten. Pot on developing cuttings and seedlings of all plants. Remove shading and provide gentle heat on cold nights.	Pot on any remaining cuttings and seedlings. Pick over plants on a regular basis in order to remove dead or dying material. Provide gentle heat on cold nights.	Keep conditions on the dry side. Be vigilant for winter pests and diseases, as plants become more stressed due to low light levels. Provide heat as necessary.	Repair any damage to glass panes immediately it is noticed. Order seed for the coming season if you have not done so already.

COLD FRAMES AND CLOCHES

Individual and multiple plant protectors are useful for covering the transplants or seedlings of warm-weather vegetables or flowers that are set out ahead of the normal planting season. These usually take the form of low plastic tunnels, cloches or individual bell jars. The colder the area in which you live, the greater their usefulness because they effectively extend the length of the normal growing season.

Above: ***This is a sturdy brick cold frame with a soil bed for growing winter and early spring vegetables.***

COLD FRAMES

These enable you to sow summer flowers and vegetables some weeks before outdoor planting and may even allow for an extra crop within a season. They are relatively inexpensive, simple structures, providing a favourable environment for growing cool-weather crops in the very early spring, the autumn and even into the winter.

Cold frames have no outside energy requirements, relying on the sun as a source of heat. They collect heat when the sun's rays penetrate the light (the top cover) which is made from plastic, glass or fibreglass. The ideal location for a cold frame is facing the direction of the sun with a slight slope to ensure good drainage and maximum solar absorption. A sheltered spot against a wall or hedge is best. Sink the frame into the ground to provide extra protection, using the earth for insulation. Put a walkway to the front and adequate space behind the frame to help when removing the light.

Designs for cold frames vary. For example, some contain barrels that are painted black and filled with water. These absorb heat during the day and release it at night. Some cold frames are built with a very high back and a steep glass slope. Others are insulated very well and may also include moveable insulation. A simple method of providing insulation is to use sacks filled with leaves over the top of the frame and bales of straw or hay stacked against the sides at night in order to protect against freezing.

There is no standard-sized cold frame. The inside depth of the frame should be determined by the height of the plants that you plan to grow. Spring annuals, perennial seedlings or low over-wintering stock may need as little as a 30cm (12in) backboard and a 20cm (8in) front board. Potted plants may need a 38cm (15in) front board and a 45cm (18in) back board. A standard glass frame light is usually about 1 x 1.8m (3 x 6ft). Do not make the structure too wide for weeding and harvesting; a width of 1.2–1.5m (4–5ft) is convenient to reach across.

Cold frames are useful for hardening-off seedlings that were started indoors or in a greenhouse. This hardening-off period is important because seedlings can suffer serious setbacks if they are moved directly from the warmth and protection of the greenhouse to the garden. It is also possible to start cool-weather crops in the cold frame and either grow them to maturity or transplant them in the garden. Cold frames may also be useful for rooting the cuttings of deciduous and evergreen shrubs and trees, and the softwood cuttings of chrysanthemums, geraniums and fuchsias during the warmer months. Ventilation is most critical in late winter, early spring, and early autumn on clear, sunny days when temperatures may rise above 45°C (113°F). The light should be raised partially to prevent the build-up of extreme temperatures inside the frame. Lower or replace the light early each day in order to conserve some heat for the evening.

In summer, extreme heat and intensive sunlight can damage plants. You can avoid this by shading with a lath (a slatted wooden frame) or old bamboo window blinds. Water plants early so that they dry before dark; this helps to reduce disease problems.

CLOCHES

Traditionally a bell-shaped glass cover, a cloche is a moveable structure that serves as a mini-greenhouse. Cloches can be used to protect transplanted tender plants from spring frosts. They also help to warm up the soil for crops sown directly in the soil.

Left: ***When the seedlings are fully acclimatized and ready to be planted out, the lights or lids can be left off altogether.***

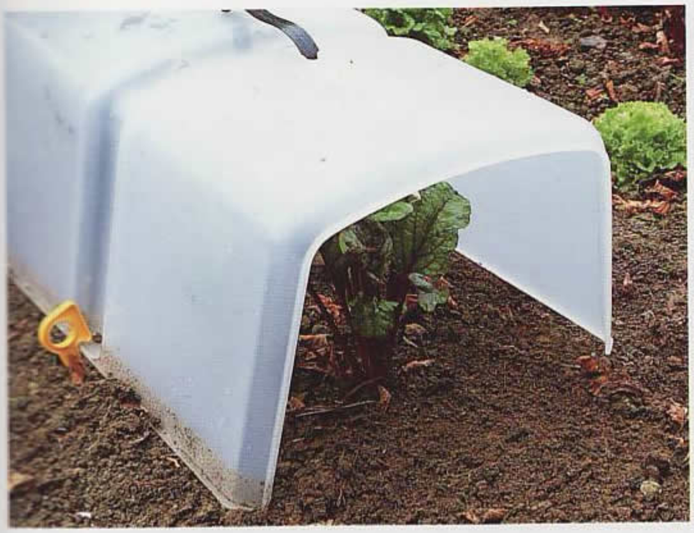

Above: *A rigid plastic cloche is easy to use. The sections butt up against each other and can be pegged into the soil. Endpieces are also available.*

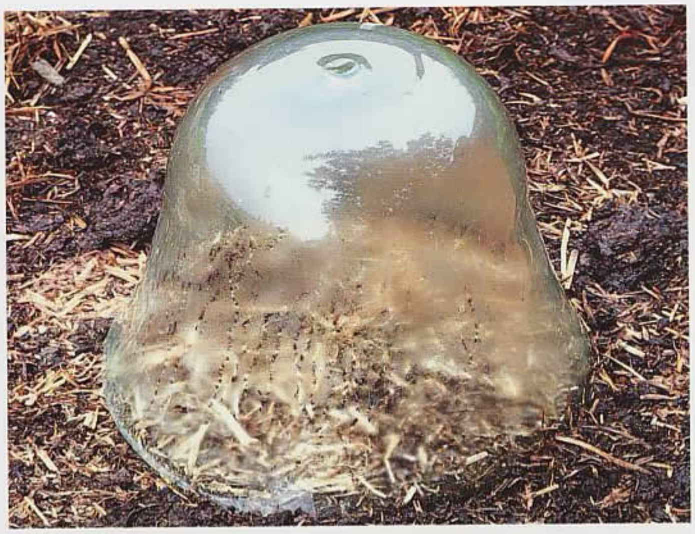

Above: *Glass bell jars are simple cloches for covering one plant. They are expensive, but large plastic jars make a good, if not as attractive, alternative.*

The traditional or European cloche is usually built in 60cm (24in) sections that vary in height from 20–60cm (8–24in) and 40–65cm (16–26in) in width. It is made of four panes of glass held together with heavy galvanized wire fittings. It has a handle for ease in carrying and for operating the ventilation system. Several cloches placed end to end make a miniature greenhouse.

Plastic bottles with the bottoms cut out can provide protection for small individual plants. These will last a season or two, but will become brittle over time. Flexible fibreglass sheets held in an inverted "U"-shape by stiff wire hoops or small wooden stakes can be used to cover rows of plants.

A tunnel-like plant protector can be made with a 1.5m (5ft) strip of plastic or fleece laid over 1.8m (6ft) wire hoops placed 1m (3ft) apart. Elastic tiedowns over the top near each hoop will hold the plastic.

Temporary cloches can also be made by arching black, semi-rigid, plastic piping over the row or bed and sticking it into the ground on each side. Lay clear plastic over the arches. If the beds are enclosed with wood, attach brackets to the inside edges of the boxes or sink short pieces of pipe with a larger inside diameter along the sides to hold the arches. The arches can be used to support fleece or shade cloth to ward off both frost and bright sunlight. When this cloche is no longer needed, simply remove the plastic sheet and pipe ribs and store them until the following season.

THE BOTTLE RADIATOR

In areas prone to late air frosts, this simple method can protect outdoor crops such as bush tomatoes. Fill a glass or plastic bottle with water and place it next to the plant. The sunlight will warm the water in the bottle, which will in turn release a gentle heat at night. This is sufficient to prevent cold shock to the plant and will ensure good growth and cropping. It can also be used in conjunction with a fleece covering to enhance the warming effect.

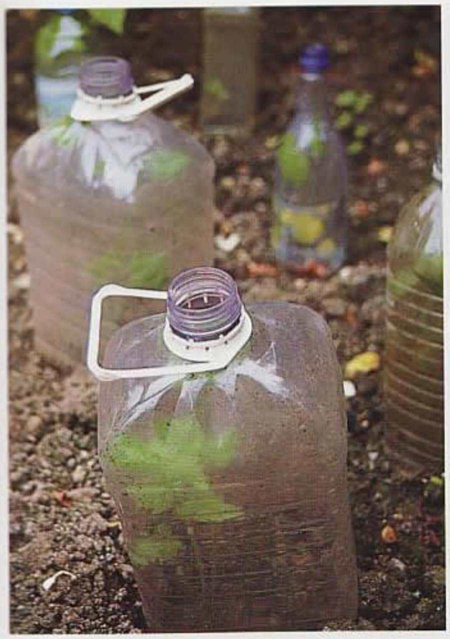

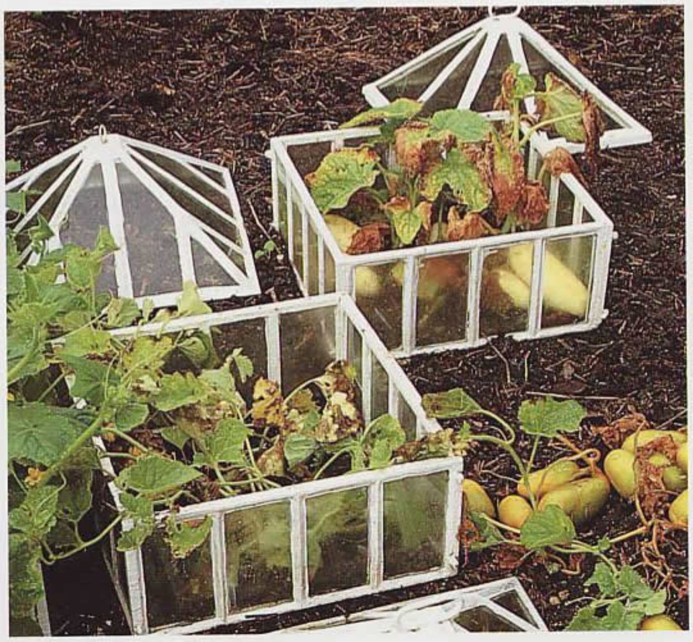

Left: *Old-fashioned cloches are particularly good for decorative vegetable gardens. However, they are also very expensive.*

Far left: *Upturned, clear plastic bottles with the bottoms cut out are ideal for using as mini-cloches, protecting individual plants. This is also an environmentally friendly technique.*

PLANT HEALTH

The main concern for new organic gardeners is that their plants will be attacked by pests and diseases. There is an array of techniques to control these undesirables, although the organic gardener must first learn to recognize signs of distress. However, you are unlikely to experience more than a handful of the problems described, all of which are relatively easy to solve. Pests, for example, are eaten by creatures known as beneficials and so the ecology of the garden, once stabilized, will be enough to keep most problems at bay. There is always a technique to control more persistent problems and so ensure the health of your plants.

Left: *A scarecrow is a traditional way of protecting your crops from marauding birds. It can also be decorative and amusing.*

bove: *Companion plants such as these arigolds can provide welcome colour in e kitchen garden.*

Above: *Brightly coloured flowers can encourage a variety of beneficial insects into your garden.*

Above: *Decorative planting in the vegetable plot can provide additional colour and interest.*

WHY PLANTS GET SICK

Plants are prone to numerous ailments, some of which can be a serious threat to their survival. Pests and diseases are only one of the potential pitfalls that you will meet during a normal growing season. A basic knowledge of the other main factors that can affect plant health are all you will need to help your garden to flourish. While the sight of your plants suffering can be alarming, there are, in general, only a few potential health threats that you are likely to encounter.

Above: *Large blocks of the same type of plant are easy for pests to locate and are therefore prone to attack.*

FROST

This can cause serious problems and is actually more critical than average minimum temperatures. Most harmful are unexpected frosts that can cause severe damage even to hardy subjects, especially when they may have produced "soft" new growth.

While warm air rises, cold air will settle and collect in hollows and depressions. Cold air is laden with water vapour and is therefore heavier. Any valley or low-lying area is, therefore, a potential frost pocket. Cold air will accumulate in a depression and then back up the sloping sides as the build-up increases. Any barrier, such as a hedge or a wall, will obstruct the passage of the cold air and a frost pocket will form. Any plants growing in the vicinity will be exposed to the frost and may be damaged by it.

When the soil is frozen, water is no longer available to the plant and shallow-rooted plants are not able to access any water to replace that which they are losing through transpiration. The plant will dehydrate and the foliage will brown and shrivel. Ground frosts may also cause the soil to "heave" and plants will be lifted out of the ground.

Alternate freezing and thawing is often more damaging to the plant than the initial frost itself, especially for tender or half-hardy subjects such as bedding plants. Severe frost can split the bark on woody subjects and may also distort leaves.

The damage caused by frost is directly related to its duration. A temperature of -4°C (25°F) for one hour may cause little or no damage, while the same temperature for four hours may be disastrous.

WATERLOGGING AND DROUGHT

The build-up of water in the soil, particularly where this occurs over a prolonged period of time, can be highly detrimental to the health of your plants. This is of greatest concern with plants that are not adapted to such waterlogged conditions. As a result, the roots of the plants will suffer and probably die from asphyxiation.

Water shortage or drought only tends to occur during the summer months when temperatures and light intensity are at their peak. The most obvious sign of the effects of water shortage on a plant is when it wilts and loses its turgidity. Water shortage also causes plant functions to slow down dramatically and prolonged drought can result in permanent cell damage.

WIND

The effects of wind damage to plants, especially woody trees and shrubs, are sometimes only too obvious. Wind often worsens extremes of temperature and drought. In severe cases of wind damage, trees can be broken and shrubs uprooted. The stronger the wind, the more damage is likely to occur.

LIGHT IMBALANCES

A plant's growth is always directed towards the available light. Plants growing in shade often become drawn and etiolated. Conversely, strong sunlight damages plants by scorching the foliage. To prevent such problems it is best to select plants that will grow well in the prevailing site conditions.

NUTRIENT IMBALANCES

Plants exhibit varying symptoms to a nutrient imbalance depending, in general, upon the severity of the problem. Symptoms of deficiency may include stunted growth, discoloured leaves (including mottling and interveinal coloration), the premature death of leaves and parts of the plant, twisted and distorted growth and poor root growth and development.

Left: *Planting crops among companion plants will make it more difficult for pests to locate them.*

Right: *A healthy, flourishing border, which is free of pests and diseases, is the ultimate aim of all organic gardeners.*

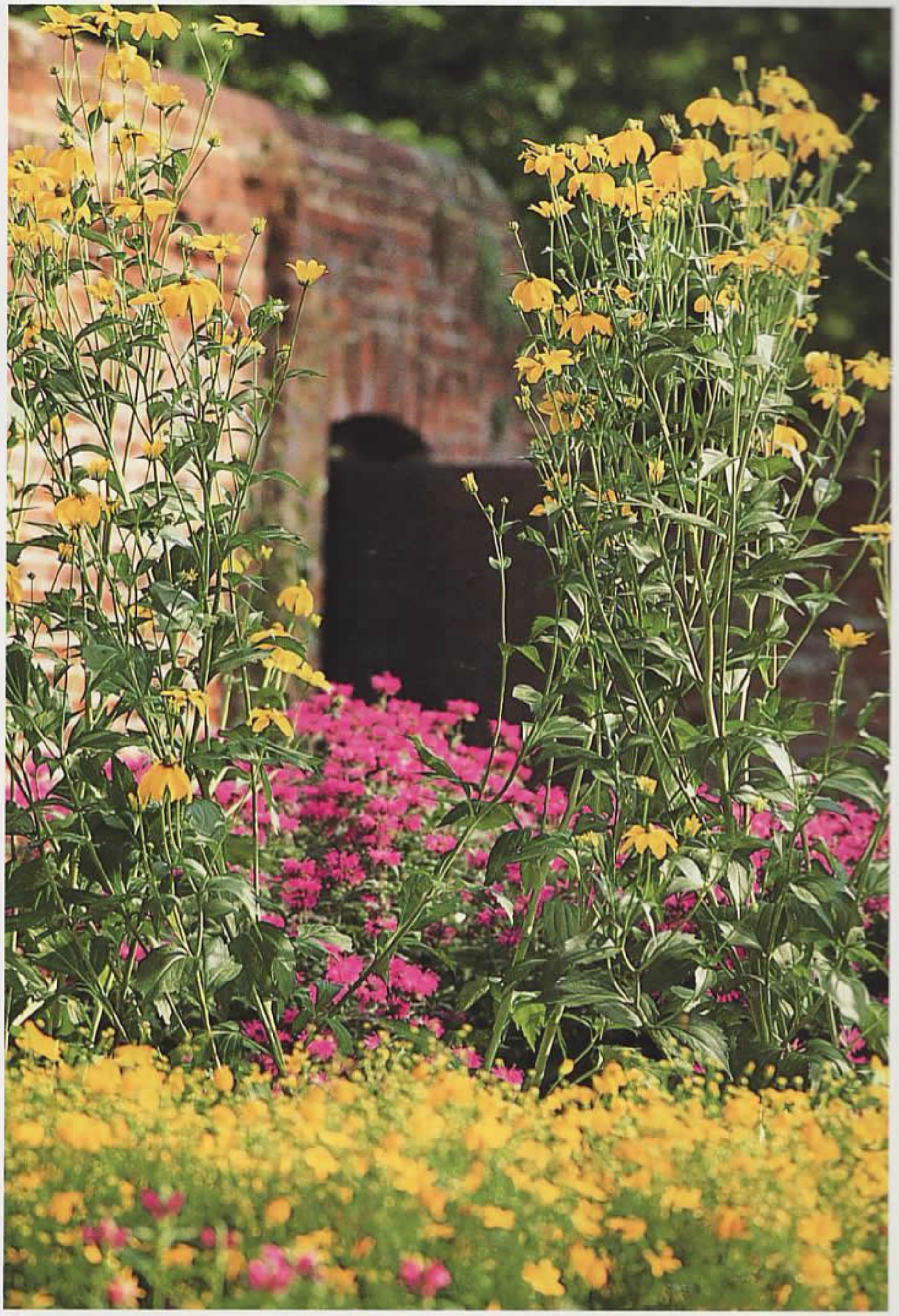

The symptoms are usually noticed first on the shoot tips, although problems may appear on any part of the plant and in different stages of growth. Determining a cause may involve an analysis of the soil.

Any nutrient can be toxic to the plant if it is present in sufficiently high amounts or is out of balance with other elements in the growing medium.

POLLUTION

Specific symptoms of pollution may include leaves turning brown at the tips and margins, leaf discoloration or premature leaf fall. Growth may also be stunted.

Soil pollutant damage can be very severe and rapid in its effects on a plant and its growth. Common problems that you may encounter include extreme soil acidity or alkalinity, chemical toxicity, salt toxicity or pesticide residues. Of these, only soil acidity is relatively easy to cure. If serious pollution is suspected, then this requires specialist help. Such situations, however, are thankfully rare.

PESTS AND DISEASES

It is important to bear in mind that a certain level of pest and disease invasion is normal even on healthy crops. However, it is also true to say that healthy vigorous plants are more resistant to serious attack than plants that are growing under stress. The best form of pest and disease control is obviously prevention. You will need to be able to find and recognize a range of pests and diseases in order to prevent outbreaks of them in your garden and, if necessary, take prompt action.

Left: *Healthy plants are often the result of careful selection according to the site conditions.*

Right: *It is very important to recognize potential problems before your plants suffer too much damage.*

PREVENTING PROBLEMS

Organic gardening is about working with nature to create an environment in which plants can withstand attack from pests and diseases, without you resorting to the use of harmful insecticides. If plants become infested with pests or diseases, they may fail to give a good display and crop yields can suffer. Prevention is better than cure, and a keen eye and regular checks will help you to anticipate and prevent the worst of any potential problems.

Above: *A planting combination of roses and daisies will help to attract beneficial predators to the garden.*

GOOD CULTURAL PRACTICE AND HYGIENE

Many pests and diseases can survive without a susceptible host even under the most unfavourable conditions. Myriad plant diseases survive from one growing season to the next on plant debris, in the soil, on seeds or on alternate hosts (some pests and diseases affect different plant species at different times of the year e.g. peach-potato aphid). This means it is vital to remove and properly dispose of any infected plant materials. It is also important for the organic gardener to be aware of the diseases that can threaten an individual crop and recognize the conditions in which these potential threats to plant health can thrive.

GROW DISEASE-RESISTANT CULTIVARS

Plant varieties and cultivars were mostly chosen for other reasons than their disease-resistant qualities. Often they become so commonly grown that their diseases become widespread. Many plants have disease-resistant strains or cultivars, but this does not necessarily guarantee that they will be immune to a disease. However, they will be better able to resist the worst of its ravages.

Above: *Place straw under growing strawberries in order to protect them from rot and soil-dwelling pests.*

Above: *The regular washing down of the glass panes in a greenhouse can reduce the build-up of disease-causing organisms.*

Right: *Regular deadheading of old or damaged blooms will help to reduce the spread of some fungal diseases.*

AVOID PLANT STRESS

A plant that is stressed – by drought or an unfavourable temperature, for example – will be predisposed to pest or disease attack. Plants that are not subjected to higher levels of stress than they can cope with will remain healthy and better able to deal with potential attackers.

Stressed plants often show signs of physical disorders (e.g. being tall, drawn and pale due to lack of light). These can be due to the weather, being wrongly sited, nutrient imbalance or the presence of a toxic substance in the air or soil. Physically stressed plants may become sicker than if a pest- or disease-causing organism actually had invaded them. Stress can kill a plant if the problem is not quickly remedied.

RIGHT PLANT, RIGHT PLACE

Plants all have their preferred locations and the occurrence or lack of a particular environmental factor or factors will ultimately determine whether a plant will prosper in the position in which it has been planted. Ferns, for example, need a cool moist site. Placing one in a hot sunny site will lead to its death as it struggles to keep its moisture. Plants that become stressed will neither grow as well, nor be as disease-resistant, as they would otherwise. Choosing an appropriate site in the first place will at least help to ensure the initial health of your plants and will render them more able to resist other potential threats to their health.

RECOGNIZING THE PROBLEM

Organic gardeners who understand pest life-cycles and behaviour are better able to determine when control will be most effective. Insects living in your garden are all part of nature's complex ecosystems and food chains. Less than one per cent of species that you are likely to encounter are considered pests. Since few insects are actually harmful, organic gardeners must learn which are pests, which are beneficials, and which ones will have no effect on the garden whatsoever.

Left: *Always choose plants that are suited to the site. Alkaline soils – indicated here by the pink flower of this hydrangea – will not suit certain plant species.*

Despite the fact that they are not always popular, insects play an important role in our gardens. Beneficial insects, such as bees, are necessary in the organic garden to pollinate fruit and some vegetable crops. Others, such as springtails, also help to break down dead plant tissue, while wasps and ground beetles capture and eat other pest insects and are called predators. Wasps and midges have larvae that attack pests by living inside their bodies and are called parasitoids. Organic gardeners must learn to cherish this "willing army" of helpers in their gardens.

PEST AND DISEASE CYCLES

Often gardeners believe that their plants have been attacked overnight. This may be true in the case of damping-off disease or with larger pests such as rabbits. More often, however, much has occurred before the symptoms are actually visible.

The pathogen (the pest- or disease-causing organism) must be introduced (inoculated) to the host plant. Most pathogens either move by themselves (as with most pests) or must be carried to the host plant (as with the vast majority of diseases). Rain, wind, insects, birds and people usually spread plant diseases.

Splashing rain carries spores of apple scab fungus from infected apple leaves to uninfected leaves. Wind blows fungal spores from plant to plant, while aphids and whiteflies transmit many common plant diseases. Believe it or not, smokers can transmit tobacco mosaic virus from a cigarette to tomato plants.

Once the pathogen has been transferred to the host plant, it begins to multiply, change or grow into a form that can then enter the host. In many fungal diseases, the pathogen arrives on the plant as a spore, which must germinate before it can begin to grow and invade the plant. Once the fungal spore germinates, it sends out thread-like tubes called hyphae. These penetrate the plant through wounds or natural pores in the outer skin of leaves, stems and roots. The roots of bedding plants that have been damaged during transplanting are a common entry point for root-rotting fungi. A single aphid that lands on a plant can give birth to a clone every twelve hours and can eventually form a small thriving colony within a few days, leaving the gardener with the impression that it suddenly appeared.

Once established, pests or diseases can grow or increase in number and begin damaging plant tissue. As they consume nutrients or plant tissue, evidence of the damage to the plant begins to appear. Symptoms may be seen on any plant parts and include mottling, dwarfing, distortion, discoloration, wilting, shrivelling or holes and notching in the margins of leaves. The first signs of pest infestation may not appear until well after the insect has laid its eggs and disappeared.

Seeds or cuttings from infected plants will also transmit disease. Certified organic seed guarantees that at the time of sale the seeds are free of all diseases. Always try to obtain disease-free stock as this should guarantee that the plant is not infected and will not introduce disease into your garden. This is particularly important with plants such as roses as well as crop plants like raspberries and other small fruits.

Pest insects and mites may carry diseases that infect plants. Organic gardeners use the term "pest management" rather than pest control or pest eradication. It is impossible to eradicate pests from your garden completely. The best option is to try to keep pest numbers low in order to minimize the damage that they can cause in your garden.

ADOPT AN INTEGRATED STRATEGY

Organic gardeners must learn how to use a range of pest management techniques, such as introducing beneficial predators into their gardens (often referred to as biological controls), making gardens less attractive to harmful pests, and encouraging conditions in the garden that favour beneficial predators of all types (cultural controls).

Left: *Composting is the best way of recycling garden waste, but you must never compost material that is diseased.*

PLANT PESTS

Pests can be described as those creatures that harm your garden plants and, if left unchecked, they can quickly cause a great deal of damage. There is a huge army of these pests, but most organic gardeners are unlikely to encounter the vast majority of them. A basic knowledge of the commonest types that occur in domestic gardens is all that you should need to be familiar with in order to protect your plants and guarantee their health and successful growth.

Above: *Pests may well be present before symptoms appear. Close examination may reveal them before the plants start to suffer.*

HOW CAN YOU TELL IF A PEST IS A PROBLEM OR NOT?

Deciding whether a pest is a problem or not is very much a matter of opinion. Commercial growers assess the importance of a pest in terms of their financial losses. Domestic organic gardeners, however, tend to grow fewer plants or crops and mainly grow these for their beauty or for the pleasure of eating home-grown produce. The final decision as to the importance of a pest will rely upon the circumstances and experience of individual gardeners. All organic gardeners must be willing to accept a certain number of pests in their garden as these form part of the intricate food webs that result in natural control. If there are no pests, then the animals that eat them will disappear and open the door to future, potentially serious, pest outbreaks.

RECOGNIZING PESTS

It is important that you are able to accurately identify a pest that has been attacking your plants so that you can take the appropriate action. Just because an insect is seen walking on an affected plant does not mean that it is the one causing the damage. The only real way to control pests involves getting to know them. Many pests produce characteristic symptoms that make it possible to diagnose the cause with relative certainty. Some have a wide range of host plants, and symptoms may not always be as conspicuous on all affected plants. Close examination – perhaps with a hand lens – may be necessary for the final diagnosis. With careful observation and experience, it is possible to keep one step ahead of the pests in your garden.

CONTROL OPTIONS

It is important to control pests before they become a problem. A single black bean aphid (*Aphis fabae*) that lands on a broad bean at the start of the summer could theoretically give rise to 2,000,000,000,000,000 aphids by the start of the autumn. This would be about a million tons of aphids. Numbers such as this are impossible but it does go to show that early control is essential. Quantities such as those quoted in the aphid example cannot occur as the food supply would run out before this can happen. In addition, a whole host of predators eat them. In just a few days, however, pests can cause considerable damage and quick action is needed if you are to save your plants. Organic gardeners must employ a full range of control measures to ensure their plants survive this seasonal invasion, including cultural practices (crop rotation, good hygiene and encouraging biodiversity), physical controls (hand picking, traps, repellents and barriers) and biological control (using other animals that naturally eat pests). These are covered more extensively later in this section but for now it is important to stress that pests can only really be controlled by an integrated strategy that uses a variety of techniques.

CAN PESTS BE TOLERATED?

It is worth pointing out that we tend to be unduly concerned with pests damaging our plants. Supermarkets have conditioned us to expect blemish-free produce. We need to judge the overall health of a plant rather than react when we see a pest. If there were no pests in the garden, then there would be no predators. Step back and look at the whole picture and remember that everything, even pests, has its place in nature. They all add to the interest and diversity that is the most unique quality of an organic garden.

Above: *Birds can be serious pests in the garden. Here, large bites have been taken out of a brassica.*

Above: *Rabbits can devastate a garden overnight, leaving nothing but chewed off stumps as a result of their visit.*

Above: *Although fascinating to watch, squirrels are a garden pest and are closely related to rats.*

COMMON PESTS

There is a seemingly endless array of garden creatures that are waiting to devour and attack your garden plants. The most common pests are almost worldwide in their distribution, but they can be controlled relatively easily. It can be discouraging to see how many potential pests may attack your garden, but it is important to bear in mind that you will only encounter a handful of these in your gardening career.

PEST		PLANTS AT RISK	TREATMENT
Ants and termites		Ants are not really pests, but they "farm" aphids for sticky sugar. Termites can attack some woody species.	Few organic treatments, but baits based on borax are useful. Use a herbal spray of essential oil (citronella and lavender) or lukewarm water as a repellent.
Aphids		Most cultivated plants growing in the open, under glass or indoors.	Encourage beneficial insects to feed on them. Organic insecticidal soap can also be useful, particularly if the aphids are being "farmed" by ants.
Birds		Some birds will attack fruit and brassicas as well as brightly coloured flowers, including blossom.	Bird scarers may be employed as a deterrent and netting will also provide cover for individual crops and plants.
Cabbage root fly		The small maggots grow and develop on and within the roots of the developing cabbage plant.	Place a collar around newly planted seedlings, sinking it into the ground to prevent the newly hatched maggot from reaching the plant roots.
Carrot fly		These root-feeding maggots feed on carrots.	Try companion planting and avoid large monocultures of carrots. Erect plastic or fleece barriers, about 45–50cm (18–20in) high and 2–3m (6½–10ft) apart.
Caterpillars		Many different species of plant, especially those in the cabbage family.	Birds and other predators will reduce populations. Biological control, using a bacterial agent, and pesticides such as derris. Hand-pick off individual plants.
Chafers		Raspberries, strawberries, potatoes, lettuce, young trees, lawns and some herbaceous perennials.	Keep the ground weed-free, and well cultivated in order to reduce the number of bugs. Roll lawns in late spring.
Deer		Browse on garden plants. Male deer rub their antlers against trees, causing damage to the bark.	Notoriously difficult to control, often best kept out by fencing. Repellents are available, based upon formulae such as bear or even lion droppings.
Fruit flies		Troublesome on softer fruits in warm conditions, especially tomatoes.	Cover the fruit with a small piece of rag. Paper bags are also good to use, but can be difficult to get around a truss of tomatoes.
Gall mites and wasps		These pests affect a wide range of trees and shrubs, especially oak trees.	Generally not problematic or life-threatening. They are a good indicator of healthy biodiversity in a garden.
Leaf hoppers		There are many different types, affecting a wide range of plant species.	Difficult to control. Remove dead leaves to reduce overwintering eggs and nymphs. Encourage predators such as lacewings.
Leaf miners		There are many hundreds of species that all have their preferred host species.	Hand-picking of severely affected leaves. Biological controls exist but the problem is rarely serious enough to warrant their artificial introduction.
Leather-jackets		Can be a nuisance, especially of turf grass. They feed on roots just below the soil surface.	Can be difficult to control. Damp sacking over the soil can lure them up to the surface where they can be hand-picked or left for birds to eat.
Mealy bugs		Suck the sap of many species and produce a sticky honeydew that supports the growth of sooty moulds.	Biological controls are available. Cultural control is more difficult and involves drenching with an organic soap solution an then rinsing.
Mites		A common pest of many plants, especially those growing in hot dry conditions.	Biological controls are available. For a cultural control, improve humidity or spray foliage with an organic soap solution and then rinse.
Moles		Damage roots by lifting small newly planted trees. Tunnel under lawns, making them a serious pest.	Best controlled by trapping or using repellents. Sonic repellent devices are rarely effective. Removing one mole can simply "open the door" to another.
Nematodes (eelworms)		A few species cause disease-like symptoms. A problem when the same plants are grown in the same place.	Regular crop rotations can help to reduce damage. If numbers build up or plants become badly affected, avoid growing the affected species.
Rabbits and hares, squirrels, voles and mice		Rabbits gnaw shoots. Mice and voles eat small bulbs and corms in winter. They all ring-bark young trees.	They may be trapped or a variety of repellents are available. Tree and shrub shelters are useful against voles and mice as is fencing for rabbits and hares.
Sawflies		Developing larvae eat plant tissue. Fruit sawflies are notorious. Slugworm sawfly attack ornamentals.	Control slugworm by applying insecticidal soap. Control fruit sawfly larvae with derris. Both these treatments reduce numbers of beneficial insect predators.
Scale insects		Several species, some of which have specific host plants.	A cultural control consists of swabbing woody stems with a strong organic soap solution or pruning and removing affected parts.
Sciarid fly (fungus gnat)		Tiny flies that feed on soil fungus. The larvae attack the roots of young plants in water-logged potting mix.	Control by preventing the potting mix becoming too wet, although biological controls of nematodes are also available.
Slugs and snails		Common pests of a wide variety of plants. Snails, in particular, cause damage and defoliation of plants.	Often difficult to control although they can be caught in traps containing beer. A variety of barriers and deterrents are also available.
Thrips		Small insects suck the sap of soft foliage and attack flowers. Rarely a problem outside in cooler climates.	Remove seriously damaged foliage. Several natural predators (usually mites) are available for use against thrips.
Weevils		Larvae attack roots, stems or flowers and fruit of a range of plants. Adults feed on affected plants.	Wet acidic composts favour ground-living types such as vine weevil. Biological control possible with parasitic nematodes if temperatures sufficiently high.
Whiteflies		A greenhouse pest in cooler climates that may occur outside in warmer areas.	Biological controls are available but the best option can be to avoid growing susceptible species.
Woodlice (pill bugs)		Usually a pest indoors of crops like cucumbers. May chew through young seedlings.	Often more of an indication of poor hygiene than a problem in itself. A sign that you must clear up the greenhouse.

PLANT DISEASES

The early detection of plant disease can help to halt the widespread infestation of your crop. Try to establish a routine of regularly checking your plants and crops. Look closely for any telltale signs, using a hand lens if necessary. Remember that the first or most obvious symptom may not always be the only one or even the most important. Always check to see if there are other symptoms to ensure that you get the full picture before making your final diagnosis.

Above: *Modern rose varieties are highly susceptible to diseases such as blackspot and mildew.*

RECOGNIZING PLANT DISEASES

Vigilance is the key to success in controlling plant disease. Carefully inspect leaves, stems, roots, flowers and fruits for any sign of disease. You may even find it useful to cut open a branch or stem to look for problems such as discoloration of the tissue, which may explain leaf or stem wilting and sudden wilting of a section of or a whole plant.

Stand back and look at the overall picture. Consider the whole environment. This will include the weather, soil, the stage of development of the plant (and any pathogens present), cultural practices and the condition of other plants in the area. All of this information can help to indicate what may be wrong with the plant. Remember that a plant growing in the wrong location may be stressed.

Try to determine when the symptoms became apparent. The onset of a problem may be due to a cultural practice, the seasonal appearance of a disease or insect, or a weather-related event. Remember that long-term stress is slow to appear, taking a year or more at times.

Try to determine whether the problem is spreading, as this may indicate that it is a disease. Check whether plants of other species have been affected, as diseases are usually (but not always) species-specific. Problems caused by environmental factors do not spread, although the symptoms may become more severe.

You must know what the plant should look like in order to be able to recognize any abnormalities. Reading up a little on the species that you are growing in your garden can help you to make a more accurate diagnosis of the problem.

Remember that there is usually no single cause of a disease infestation. The primary cause may be associated with cultural or environmental conditions. And just as there is probably no single cause, there is usually no single symptom.

When you are attempting to diagnose the cause of plant illness, always inspect symptoms that appear on parts that are still alive (or at least partially alive). Dead plants are often invaded by secondary infestations of decomposer fungi, which may hide the original problem. If possible, make an examination of the entire plant, including roots, although this may not be possible for large specimens such as trees.

Left: *Although a plant may be in obvious distress, closer examination may be needed to properly determine the cause.*

Right: *Disease-causing organisms can enter a plant each time it is cut or pruned.*

CONTROL OPTIONS

Ultimately, where a serious disease is suspected, it may be advisable to avoid growing susceptible species altogether, or, in the case of fruit and vegetables, to rotate the crop as part of a regular cycle. Make notes that provide details on disease occurrence (the type of disease and when it appeared), the plants affected, the weather and environmental conditions in your garden each year. By doing this you can better anticipate what problems are likely to occur in your garden during the growing season.

Finally, remember that even the best gardeners lose plants to disease. This is only serious when large numbers of plants are affected. A diverse garden will contain other highlights, and diseases will claim only a fraction of the planting.

COMMON DISEASES

There are all kinds of diseases that can affect your plants. Most of these are mercifully rare, but every garden will suffer from its share of diseases during the growing season. Most of these conditions are relatively easy to deal with, but a correct diagnosis is essential. As always, it is preferable to try to prevent these diseases taking hold and spreading in the first place rather than treating them when they occur.

DISEASE		PLANTS AT RISK AND THE SYMPTOMS	TREATMENT
Bacterial canker	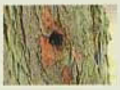	Many trees and shrubs. Reduces vigour and the rate of growth of the affected plant. Branches become girdled and die back. A sticky secretion will ooze from cracks or welts in the bark.	Difficult to control once a plant is infected. Confine the problem to one specimen by cleaning pruning tools between cuts and before moving on to another specimen of the same type.
Blackspot	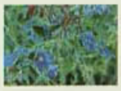	A common fungal disease of roses, which is most prevalent on modern hybrid tea and floribunda types. Black spots appear, primarily on the leaves and sometimes on the stems as well.	Prune bush roses into a goblet shape. Use disease-resistant cultivars. Dust affected plants with sulphur early and intermittently throughout the growing season.
Bracket fungus	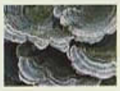	This disease can affect all types of tree. Bracket fungi are the fruiting bodies of fungi that either parasitize healthy heart wood or, more commonly, decay heart wood that is already dead.	Many affected trees live for many years and need only be felled if they present a risk of collapse.
Common scab of apple	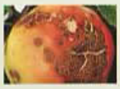	A bacterial disease that causes the formation of large (albeit harmless) scabs on the surface of the apples. Commonly encountered in damp weather and on trees with crowded branches.	Rake up and dispose of affected leaves. Prune out cracked or scabby shoots to remove places for the fungus to overwinter.
Coral spot	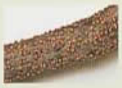	A common fungus affecting many woody plant species. Appears as pink or bright red, raised pustules on wood that is showing signs of dieback.	Prune out infected tissue and destroy it. Regular "hygienic pruning" can help to prevent it taking hold in the first place.
Downy mildew	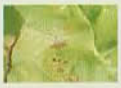	A group of fungal diseases that affects a wide range of plants. Appears as a white coating over leaves that display distended growth, browning and wilting. Thrives in warm moist conditions.	Dust with sulphur or spray with Bordeaux mixture to form a protective barrier against the spores. Note that these are both easily washed off by watering or rain.
Fireblight	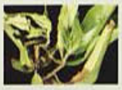	Attacks plants of the Rosaceae family, notably apples, pears, plums, cherries, pyracanthas, cotoneasters and roses. The affected plant dies back rapidly from the branch tip.	Remove and destroy all affected material, cutting back to at least 50cm (20in) below the point of infection. Best to remove diseased plants, replacing them with a non-susceptible tree or shrub.
Fungal canker	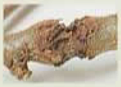	Affects many woody plants. The bark tissue begins to peel and flake away around the affected tissue in concentric rings.	Prune out the affected material, about 15cm (6in) below the point of infection, and burn it. Clean tools with disinfectant between cuts and especially between pruning individual plants.
Honey fungus	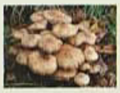	A soil-dwelling fungus that parasitizes the roots of woody plants. Leaves tend to discolour and wilt and fail to develop in the spring. Ultimately, this can weaken and kill the plant.	Difficult to control organically. It may be best to grow herbaceous plants in affected areas for two to three years before attempting to replant woody plants there.
Mosaic virus	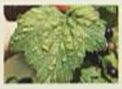	This group of viruses affects a very large number of species and is characterized by an irregular, angular mottling or streaking of the leaves.	Winged insects that feed on plants, then migrate to another, can rapidly spread the virus. Control the pests themselves where possible.
Phytophthora	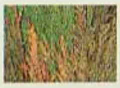	A soil-borne fungal disease that affects many species of woody plants, particularly conifers. Branches die back, often very quickly. Serious infection may cause death for a large proportion of the root system and severe dieback of top growth.	There is no cure for this condition and the removal of badly affected plants may be the only answer. The disease is often restricted to damp or heavy soils, although prolonged wet weather can result in outbreaks.
Powdery mildew	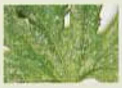	Affects many plant species, with asters being especially prone. The leaves become covered with white powdery patches that may distort growth or even cause leaf drop in severe cases. Thrives in warm, humid or wet conditions.	Sulphur and Bordeaux mixture can provide protection, but these are easily washed off in wet conditions. Avoid growing susceptible species if conditions favour the spread of the disease.
Rust	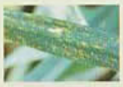	A common fungal disease affecting many species. Characterized by rusty coloured patches or spots, known as pustules, on leaves. Tissue around the pustules yellows and dies, and this, in turn, may distort growth or cause leaf drop in severe cases.	Commonly seen on soils rich in nitrogen. It may have a noticeable effect on seasonal crops or those grown for their leaves. Bordeaux mixture can reduce its spread but it is difficult to control completely. Avoid susceptible crops and varieties if the problem is persistent.
Silverleaf	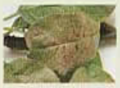	Affects both fruiting and ornamental species of *Prunus*. The leaves on some branches gain a silvery sheen, gradually dying back a year or two later. A purple fungal growth appears on the dead tissue.	The affected wood should be pruned out below the point of infection and disposed of in late summer after fruit has set. Badly infected specimens should be removed completely and disposed of, preferably by burning.
Stem rot	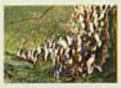	Attacks trees and larger woody plants. Stems and branches gradually become hollow.	Smaller trees or shrubs may be much more seriously affected and may be best removed if their condition deteriorates too much.

PLANT DISORDERS

Many external factors can affect garden plants. Weather seldom does what we want it to – there is either too little or too much rain or it is too cold or too hot. If you combine these climatic disappointments with other factors that can affect your plants, such as pollution or nutrient deficiencies, you will see why these difficult conditions can produce a number of disease-like symptoms. They can all put severe stress on an ornamental plant or crop and so precipitate attack from living organisms.

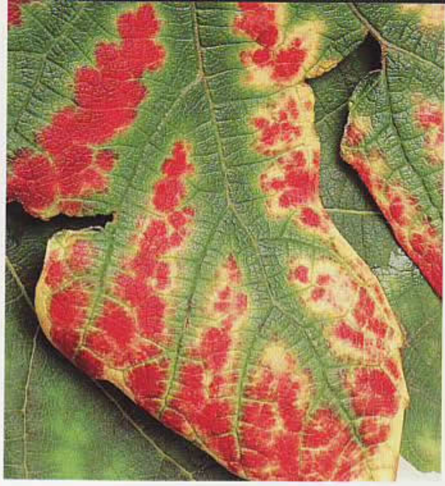

Above: *Nutrient deficiencies can cause discoloration of the leaves. This plant is short of magnesium.*

RECOGNIZING DISORDERS

Plants can be susceptible to a long list of outside forces. Extremes of weather, nutrient deficiencies and physical damage can all take their toll on the health and vigour of a plant. Other environmental factors, such as too much or too little water or pollutants in the soil, can also encourage disease-like symptoms.

Soil pollution can be caused by nutrient deficiency, misapplied fertilizer (resulting in too much nutrient and, therefore, toxicity), spilt lawnmower fuel (perhaps due to careless filling) or buried inert material. Airborne pollution can be more insidious and more difficult to determine or detect.

Weather events such as high winds or frost may go unnoticed if you are not there to witness them. A huge storm will leave evidence in its wake. The effect of a sharp early morning frost, however, or a steady drying wind on a sunny afternoon may not show the damage caused until several days have passed. Often frost occurs in the hours just before daybreak and quickly disappears to leave a fine sunny morning.

Disease can result from a combination of factors, affecting growing conditions and actual disease-causing organisms. Plants may initially be placed under stress – by high winds or frost, for example – making them vulnerable to attack by living agents. For instance, drought may damage roots, which in turn renders them more liable to infection by fungal diseases.

It is important to determine whether the distress that your plant is suffering is the result of a pathogen or due to a problem in the environment. In order to work out which of these is responsible, look to see if the occurrence of ill health is random or uniform in terms of its distribution. As a general rule of thumb, randomly distributed symptoms on injured plants are usually caused by a living factor, such as infectious diseases or a pest. In addition to this, infestations, particularly those caused by diseases, tend to radiate out from central points. Uniform patterns are generally associated with non-living or non-infectious agents such as poisons, fertilizers, environmental stress or mechanical damage.

Above: *Drought is a common cause of stress in potted plants, damaging both roots and shoots.*

CONTROL OPTIONS

If the cause of the problem is physical, you will need to find out whether it is due to a recurring environmental factor inherent in the site, such as constant buffeting by strong winds, or whether it is caused by "one-off" events such as unseasonable frosts or contamination from a careless neighbour's weedkiller spray. Some factors can be removed, such as polluted soil, and windbreaks can be planted and fleece draped over tender specimens on cold nights. If plants continue to be affected, it could be that the plant is in the wrong place and you may have to try it elsewhere.

Left: *Plants sometimes show unusual growth patterns in response to environmental conditions. Here, the growth of a tree has been affected by the wind.*

COMMON DISORDERS

Disorders are the result of either an imbalance of essential nutrients or of a range of stresses that are caused by adverse or difficult environmental conditions. Many of the disorders and conditions described here can resemble diseases. This means that is very important for you to diagnose the cause of the distress in the plant before taking any remedial action. A plant disorder will call for a different response to a plant disease.

CONDITION		PROBABLE CAUSE	TREATMENT
Bare patches or areas of poor plant establishment		The patchy establishment of cover crops or turf can be due to soil compaction or underground obstructions. Soil can easily become compacted if it is walked on a lot, particularly when it is wet. Buried obstructions such as rubble can also cause similar growth distortions.	Remove all buried rubble prior to planting or replanting. Relieve compaction by cultivating and replant if required.
Blackening of leaf tips	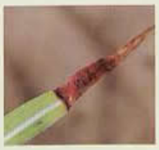	Usually the result of overwatering, particularly with container-grown plants. The waterlogged soil or potting mix forces oxygen out of the soil, thereby suffocating the roots which suffer a form of drought-stress.	Do not water until the soil or potting mix dries out a little. Check to see if plants need watering before doing so.
Etiolation		Sun-loving plants kept in low light conditions quickly become starved and will be significantly weakened. If kept in these conditions for any length of time, the plant will eventually die.	Choose specimens that are appropriate for the light levels in the garden.
Leaf blackening	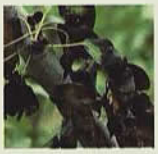	Frost damage on buds and leaves in early spring, even to hardy specimens, is usually noticed on new growth that has not yet become acclimatized. It is the sudden shock that often causes the problem, not the actual temperature itself.	Cover slightly tender plants with fleece in the winter. Lightly spray plants with water in the evening to help protect against late frosts.
Leaf scorch	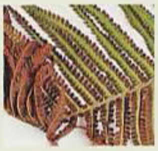	High winds and bright sunlight, especially on shade-loving species. Some plants can also be damaged by watering in bright sunlight, by providing too little or too much water or by applying too much fertilizer to the soil. Hail can also cause leaf spotting or holes.	Water in the evening. Choose wind-resistant species, if appropriate. Apply fertilizer at recommended rates. Do not over- or under-water plants.
Mechanical or physical damage		High winds, which are especially damaging to deciduous woody plants in full leaf. Herbaceous plants can easily break if not adequately staked, as may newly planted trees and shrubs. Snow can snap the branches of conifers and evergreens. Visiting animals can also flatten or snap garden plants.	Stake plants firmly and ensure that animals are excluded from areas where plants could be damaged.
Nutrient deficiency	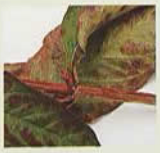	A lack of or too much of a particular nutrient. Nutrients can also become in short supply if other nutrients are present in large amounts.	Deficiencies are best treated with the application of fertilizer or choose plants that are adapted to deal with the site conditions.
Poisoned ground		Leaked fuel or lubricants used in construction work can leave the ground contaminated. This is usually more of a problem in gardens attached to modern houses. Underground gas pipes can leak, flooding the ground with gas that is lethal to plant roots. Other chemical spillages could come from machinery such as a lawnmower.	Remove all affected soil and replace with fresh topsoil. If pipes are leaking, make sure that they are fixed before replacing soil. Fill lawnmowers with fuel over a plastic sheet.
Root girdling	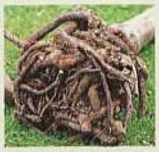	This can cause instability and the collapse of larger trees and shrubs. Tree and shrub roots become woody following the first year's growth and, if these are constrained in round pots, they tend to grow in spirals.	Try to use bare-rooted stock whenever possible. Tease out roots from the rootball immediately prior to planting.
Wilting		A normal response in many plants to either a lack of water or high temperatures, and not necessarily a cause for great concern. Ligularias habitually wilt on hot summer afternoons, even when planted in wet ground, only to "pick up" and show no ill effects later.	Treatment may not always be necessary. Check the soil first to see if it is dry. If not, wait until the evening to see if the plant shows any signs of improvement.

BENEFICIAL PREDATORS

A thriving population of natural predators and parasites can significantly help to keep pest populations down. The organic gardener must strive to create and maintain an environment in which these welcome visitors to the garden or greenhouse can prosper. Above all, this means avoiding pesticides, which can wipe out beneficial predators and so upset the natural balance in the garden or greenhouse.

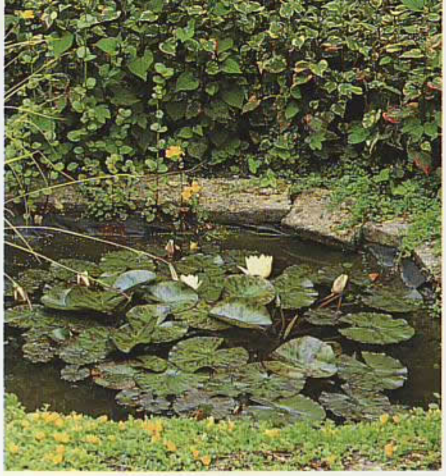

Above: *Creating a diverse range of habitats, including a pond, in your garden will encourage beneficial predators.*

BIOLOGICAL CONTROLS

Nearly every species of plant-feeding insect has another insect that is its predator or parasite. Some pests, such as aphids, are eaten by ladybirds (ladybugs), hoverflies and midge larvae, parasitized by wasps and infested by fungal diseases. Large pest populations are a food larder for many "natural enemies", including carnivorous animals, parasites and diseases.

Biological methods of control use these natural enemies of pest insects to keep their populations under control. It is like having an army of insects and other creatures doing the work for you. Biological controls can be encouraged into the garden by creating a suitable habitat. Ironically, this means that the pests must be present first. They must also be present in sufficient numbers to support a viable population of the beneficial predator. This can be a complex area and the easiest way is to encourage the conditions that will favour a balanced food web to develop. Purchasing beneficials is only generally recommended for use in greenhouses although a few (particularly microbes) are suitable for use outdoors.

CREATING SUITABLE HABITATS

The easiest way to achieve this is to get a mixture of habitats. The more variety your garden has, the more biodiversity it will support. Ponds, long grass, log piles, and food plants for animals with varied tastes all encourage a stable garden ecosystem where pests are kept at relatively low levels.

Many birds and bats eat insects. Bats need a roosting site, so place a purpose-made bat box on the shady side of a tree. Position the box at least 3m (10ft) high with a clear airspace in front. Night-scented flowers encourage the moths that bats feed upon. This is the key. The predators must be able to live in your garden or they will go and find a more suitable habitat elsewhere.

The longer you garden organically, the more balance will appear in your garden. This is simply because your garden ecosystem begins to stabilize and diversify over time.

BENEFICIAL INSECTS

Predators (those that devour pests directly) and parasitic insects (those whose young hatch inside and devour pests) are often termed "beneficials". Predators include , lacewings and wasps. Parasitic insects (more correctly termed parasitoids) are less well known than predators, but equally effective. They lay their eggs in a pest species. When the eggs hatch, the larvae feed on the pest insect, killing it. The majority of these insects are tiny wasps, although some flies and mites also fall under this category. Learning to recognize beneficial insects is crucial if you want to avoid killing your army of garden helpers.

Above: *Log piles encourage beneficial predators such as beetles to colonize and take shelter in your garden.*

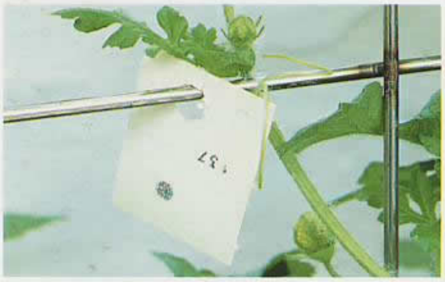

Above: *Biological controls are a successful way to fight pests. They are mainly used in greenhouses, but others are now becoming available for the open garden. The control insects are released, here from a sachet, in order to attact the pests.*

BENEFICIAL PREDATORS IN THE GARDEN

There are many beneficial predators that will willingly take up residence in your garden and help to control populations of pests in a natural way.

- Anthocorid bug or red kneed capsid
- Bats
- Birds (robin, blue tit and thrush)
- Centipedes
- Earwigs
- Frogs, toads and newts
- Ground beetle
- Harvestman
- Hedgehogs
- Hoverflies
- Lacewings (adult or larva)
- Ladybirds (ladybugs)
- Mites e.g. *Phytoseiulus*
- Nematodes e.g. *Heterorhabditis*
- Parasitic wasps (*Encarsia* or *Aphidoletes*)
- Slow worm
- Spider on web or wolf spider
- Tachinid flies
- Wasps (solitary and social)

BIOLOGICAL CONTROLS IN THE GREENHOUSE OR GARDEN

The application of biological controls in the greenhouse or garden involves using predatory insects or other beneficial animals to control commonly occurring pests. Many pests are only a problem because their natural predators are missing from the garden. Introducing a biological control usually results in the rapid control of the pest.

NAME	PREFERRED TEMPERATURE RANGE	WHAT THEY CONTROL
Amblyseus cucumeris (predatory mite)	25°C (77°F)	The nymphal forms and adults consume large quantities of immature thrips.
Aphidoletes aphidimyza (predatory midge larva)	21°C (70°F); needs 80%+ humidity	Tiny mosquito-like midge larvae that control substantial populations of more than 60 species of aphids.
Cryptolaemus montrouzieri (predatory beetle)	20–25°C (68–77°F); needs 70%+ humidity	This ladybird is effective in controlling mealybugs on houseplants and in greenhouses.
Encarsia formosa (parasitic wasp)	18–25°C (64–77°F)	Minute, flying parasitic wasps, which lay their eggs inside whitefly scales (the pupa stage) and eat them in two to four weeks
Heterorhabditis megidis (parasitic nematode)	Minimum soil temperature of 14°C (57°F). If temperature drops below 20°C (68°F), they become less effective.	Patrol the soil to a depth of about 18cm (7in) and quickly take care of the slow-moving grubs, like vine weevil grubs and chafers. Very effective in pots and containers. Soil must be moist.
Metaphycus helvolus (parasitic wasp)	20–30°C (68–86°F)	These tiny, black and yellow wasps are effective against several soft-scale species, including brown scale. The females lay their eggs under the body of first- and second-stage scales. The grubs feed on scales and develop into adults within two weeks. Adults also provide control by feeding on non-parasitized scales. *Metaphycus* are most effective in semi-tropical conditions.
Phasmarhabditis hermaphrodita (parasitic nematode)	Minimum soil temperature of 5°C (40°F)	Useful for slug control. Should be applied during the early growing stages of vulnerable plants. Needs moist soil.
Phytoseiulus persimilis (predatory mite)	Use once temperature is regularly above 15°C (60°F). Best at 18–25°C (64–77°F); needs 60%+ humidity	Predator mites, slightly larger than the two-spotted mites (also known as red spider mites) upon which they feed.
Steinernema feltiae (parasitic nematode)	Minimum temperature of 10°C (50°F), although they remain effective when the soil temperature drops below this.	Aggressive predators used to control fungus gnats, mushroom flies and leatherjackets. They can be used on lawns as well as in flower and vegetable gardens, fields, orchards and greenhouses.

Above: *Ladybirds (ladybugs) are just one of the many beneficial insects that will help to keep garden pests in check.*

MICROBES

Bacteria, fungi, viruses, protozoans and parasitic nematodes are microorganisms that attack insects. These microscopic hordes are generally effective against very specific pests and present little risk to humans and the environment. Many organic gardeners may well be familiar with a popularly known, microbial-based insecticide known as Bt or *Bacillus thuringiensis*. This commonly available product, which is used to kill many different kinds of moth and butterfly larvae, is a bacterium. It produces a toxin that kills specific caterpillars. The larval pest usually dies within four to seven days. There are many strains of Bt, each type controlling specific pests.

Parasitic nematodes are also very effective against certain pests that live in the soil. However, the nematodes require moist conditions in order to survive and their temperature requirements further limit their use to greenhouses in many cases.

Despite their potential, very few fungi, viruses and protozoa are commercially available because these living organisms are difficult to raise, store and apply. The best way to encourage these willing and tiny helpers into your garden is to maintain a healthy soil that is rich and diverse in terms of the life it contains.

Above: *Providing convenient shelter, such as this lacewing hotel, for beneficial insects can help to increase their numbers.*

OTHER CONTROL METHODS

There are many ways of dealing with the different pests that appear in the garden. Some of these methods have been tried and tested for generations, others are individual to the gardener and are often the result of a happy accident. As an organic gardener, you must learn to use as many different tricks as possible to manage the pests in your garden and to protect your plants.

Above: *Companion plants such as these marigolds confuse or deter pests that would otherwise attack garden plants.*

GOOD GARDENING PRACTICE

The selection and culture of plants can reduce the potential for pests and diseases. Cultural practices are methods the organic gardener can use to change environmental factors that affect plants and their pest populations. It is essential, therefore, that gardeners know the cultural or growing requirements of each plant. Providing the correct conditions results in a vigorous plant that is less likely to be attacked by pests and diseases and can tolerate some damage.

ASSESSING THE DAMAGE

When problems do arise, you must decide whether a pest is causing enough damage to warrant control. In other words, you will need to assess how far the problem can be tolerated before action is necessary. Some form of pest damage is inevitable with any crop, but you will need to establish limits. To do this, you will need to take into consideration the amount of damage that can be tolerated, the numbers of an individual pest that can cause significant damage and the plants' stage of development. The health and vigour of the plant can also have a direct bearing on when or if you need to take action. A few holes on a leaf may not require control, but, if most of the leaf has been eaten, the plant may die.

Left: *Slugs and snails are notorious garden pests that are easily kept at bay using a water trap.*

Above: *Compact discs are an unusual and amusing way of scaring away marauding birds from your crops.*

TAKING ACTION

Monitor your plants to determine when action is necessary. A thorough inspection of the plant allows you to identify a problem before major damage occurs. You should also inspect the plant's entire environment for clues to the problem. Observing and keeping records of weather conditions, for example, can help provide clues to growth patterns and problems.

TYPES OF CONTROL

Control options can be arranged by their mode of action and their impact on the environment. These methods of control can be grouped from least to highest impact: cultural and mechanical controls and "permitted" chemical controls (soaps, oils and botanical insecticides).

Cultural control includes hand removal of larger pests, the use of screens, barriers, and traps, freezing and crushing. These methods generally have little or no negative effect on the environment and are particularly suitable for smaller gardens.

COMPANION PLANTING

This is commonly used to protect plants from pest attack. The theory is that the companion plants – flowers growing next to a food crop, for example – disrupt the searching pattern of the pests looking for host plants. They literally smell these hosts but become confused with the more diverse planting style. Separating rows of cabbages, broccoli or other brassicas with rows of onions has always been a popular combination, possibly because the onion's strong scent confuses cabbage pests. Tomato plants also grow well next to cabbages and seem to deter caterpillars while growing leeks near carrots repels carrot flies.

SCREENS AND BARRIERS

Any material that is fine enough to keep pests out can be used as a barrier. A variety of screens of different mesh sizes can keep out large insects, birds and rabbits, but they can also prevent pollinating insects from reaching a plant, resulting in lack of fruit. Cardboard and metal collars will prevent cutworms from reaching young transplants. Sticky bands placed on tree trunks trap beetles and soil-hibernating pests. Copper strips are available for slug control. These supposedly react with the slugs' slime to shock them. Sharp particles, such as crushed eggshells, are also used to control slugs.

TRAPS

Certain insect pests can be monitored by using traps. Sticky coloured traps, pheromone traps and pitfall traps (like beer traps for slugs) can all be used to monitor the occurrence of some pests. Whiteflies and aphids are attracted to bright yellow, and this colour is used for sticky cards upon which they become trapped.

You may want to apply a control and then enclose the plants in netting to keep further infestation from occurring – perhaps putting up netting and then releasing predators.

Traps usually serve as a monitoring system, warning of the presence or increase in undesirable pest numbers. Traps can also be useful in timing control measures by showing the presence of migrating or emerging adults. The control measure can then be introduced at the best time to control the particular pest. Codling moth traps for use in orchards are a good example of this. They are sometimes used to control numbers, but most types are limited in their real effectiveness. Yellow, sticky traps attract whiteflies, aphids, thrips, leafhoppers and other small flying insects. Traps that use pheromones or attractive scents to tantalize adult insects are best used as a way to check presence and numbers. Pitfall traps can be cups or jars placed into the ground filled with yeast and water or beer to trap slugs.

WATER

A jet of water from a hose washes aphids, spider mites and other small insects from plant foliage. This must be done frequently since it does not kill insects or eggs and it does not prevent some insects from crawling back on to plants.

INSECTICIDAL SOAPS

These are made from the salts of fatty acids. Fatty acids are components of the fats and oils found in plants and animals. These soaps should not be confused with ordinary cleaning soaps. Insecticidal soaps kill only what they touch and are effective against soft-bodied pests such as aphids, thrips, crawler stage scales, whiteflies, leafhoppers and mites. Insecticidal soaps may cause burning on some plants, particularly those with hairy leaves. Test insecticidal soap on a single leaf if you are unsure – burning will usually occur within 24 hours.

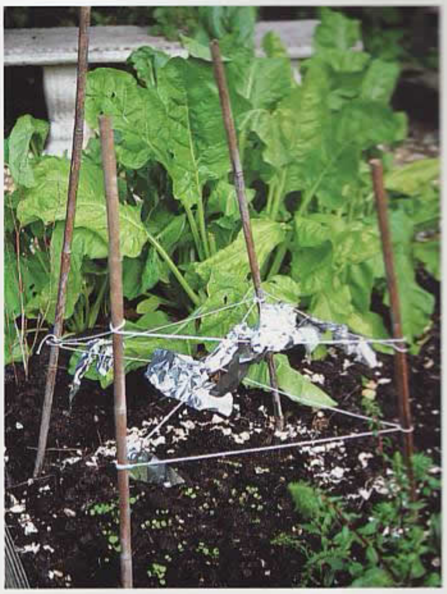

Above: *Birds can be dissuaded from attacking your plants by stretching string and shiny foil over the crops.*

BOTANICAL INSECTICIDES

Derived from plants, botanical insecticides include pyrethrum, citrus oil extracts and the extract of the neem tree. They act rapidly to stop feeding by insects, although they may not kill the pest for hours or days. There are also disadvantages to the use of botanicals. They must be applied frequently, may be difficult to obtain, and, although generally less toxic than many pesticides, they are still toxic and may harm other beneficial garden residents.

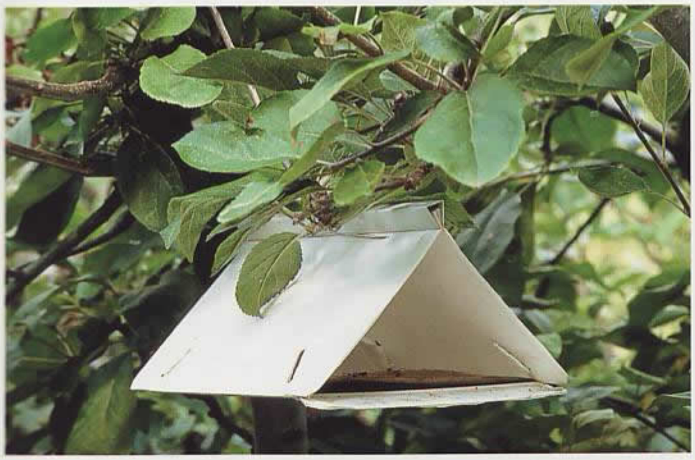

Above: *Sticky traps are another form of pest control in greenhouses. Here, pheromones attract pests to the trap where they get stuck. Other traps consist of sheets of plastic covered with a non-drying glue.*

Above: *Horticultural fleece, which is stretched over developing young plants, can provide a physical barrier against smaller pests such as flying insects.*

TRAPS, BARRIERS AND DETERRENTS

The prevention of pest and disease attacks is an essential part of organic gardening. Plants that are infected by pests or diseases are weaker specimens that are difficult to treat and never quite recover their former vigour. Many garden pests can be trapped or kept at bay using relatively inexpensive materials and sometimes recycled household items. Put the barriers in place when the plants are young and always ensure that pests are not trapped inside the barrier.

CONTROL METHOD	HOW IT WORKS
Beer traps and deterrents (granules, copper strips and greasebands)	Traps are effective ways of both controlling pests, such as slugs, and finding out which ones you actually have. Deterrents are physical barriers over which the pest cannot or will not pass. There are many types and their effectiveness can vary.
Bug nets in the greenhouse	Greenhouse vents are problematic in terms of pest control in that they allow both pests in and purchased biological controls out. Bug nets are put in place to avoid this happening.
Fleece on frames	Frames covered in horticultural fleece can be used over outdoor crops to help keep pests out and control the temperature.
Fleece stretched over a crop	Fleece can be used to create a favourable microclimate around young plants. It also acts as a barrier to airborne pests. On the downside, however, it can also keep out airborne predators from pests that overwinter in the soil.
Fruit nets over fruit	Fruit nets are especially useful for summer soft fruit crops that can quickly be devastated by birds.
Individual cloches	Cloches can act as barriers to a wide variety of airborne pests. Any pests that are sealed into this environment may, however, find the perfect environment within which to thrive.
Mesh cages for trees	Mesh cages are usually used to keep rabbits and hares at bay. They are usually simple constructions formed from three or more stakes driven into the ground with chicken wire (or similar) attached to them.
Rabbit fencing	A continuous barrier to prevent rabbits entering areas where plants are growing. The base of the wire should be buried below ground level to prevent the rabbits burrowing a passage beneath it.
Bird scarers (e.g. scarecrows)	Bird scarers have the drawback of a limited lifespan before the birds learn that they are not a real threat. They can, of course, be changed and most bird scarers are only needed on a seasonal basis.
String or wire netting stretched over seedlings (e.g. peas)	Aerial barriers can protect against bird attack. They may only be needed for the duration of the crop's life or even less.
Traps (sticky, coloured traps used in the greenhouse)	Sticky traps can provide a certain degree of control against the flying adults of insect pests. However, they are not as effective as they are sometimes thought to be and are, in fact, of more use for showing whether a particular pest is present or not, thereby allowing appropriate control measures to be put in place.
Traps (pheromone)	Pheromone traps are used to detect the presence of insects. The pheromones attract members of the opposite sex and the appearance of the target species allows you to begin looking for and controlling the young that cause the damage.
Tree guards (spiral)	Spiral guards are useful for protecting the bark of newly planted trees from rabbits and hares, especially in winter and early spring. These guards expand as the tree develops, but they are best removed completely after about a year.
Tree shelter	These protect newly planted trees from vertebrate pests and from the worst rigours of the environment by providing a favourable microclimate around them. They naturally degrade under the action of sunlight, but are best removed after two to three years.
Twiggy branches over plants	Arched over young plants, these can be an effective deterrent to pests such as birds and cats. They do not prevent the migration of beneficial predators to the plants.

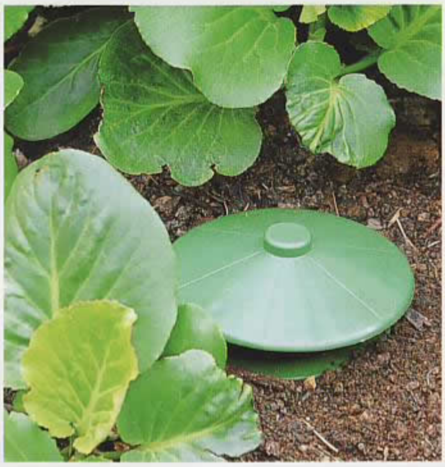

Above: *Slugs and snails can be caught in a trap that is filled with stale beer, or water that has been mixed with yeast.*

Above: *Ring tunnels that are covered in a fine mesh will prevent flying insects from attacking your crops.*

Above: *Plastic netting, stretched over the crop and carefully secured, provides an effective barrier against birds.*

Above: *Old plastic bottles make an ideal barrier to protect young plants from pests. These are cheaper than traditional cloches.*

Above: *Chicken wire, stretched over young plants, will prevent birds from reaching and eating them.*

ORGANIC PESTICIDES

While most pesticides are not used in organic gardens, some naturally occurring substances can be used to protect your plants from pests and diseases. The pesticides which you are allowed to use in an organic garden should only be applied when necessary and not on a regular basis.

ORGANIC PESTICIDE	HOW IT WORKS
Bacillus thuringiensis	These are bacterial spores that produce a toxic protein that is useful against caterpillars but will not cause any harm to beneficials. *Bacillus thuringiensis* works quickly, paralysing the caterpillar and so preventing any further damage, although it quickly degrades in sunlight and needs frequent re-application throughout the growing season.
Bordeaux mixture	A compound containing copper and sulphur used to control various fungal diseases including apple scab and potato blight. It is harmful to fish and livestock, and frequent use can lead to a build up of copper in the soil that can be harmful to worms.
Derris	A chemical extracted from the roots of the derris and longocarpus plants. It is useful against a variety of insects including aphids, caterpillars, sawflies and plant-eating beetles. It can prove harmful to beneficials, although it is not a threat to bees.
Insecticidal soap	This is not soap, like the domestic washing soap, but is made from the salts of fatty acids, extracted from plant material. It can be effective against a wide variety of insect pests, although it can damage sensitive plant species.
Plant oils	Their effectiveness varies but most depend upon coming into contact with the pest itself and suffocating it, although some, like neem tree oil, do appear to have insecticidal properties.
Pyrethrum	This organic pesticide is extracted from the flower-heads of *Chrysanthemum cinerariifolium* and is especially effective in controlling aphids. However, it does not persist for long and can also cause harm to beneficial insects.
Sulphur	An effective fungicide against a variety of plant fungi, including powdery mildew, greymould and blackspot. Sulphur can also prove harmful to predatory mites and can cause damage to certain sensitive species.

Above: *Tree shelters and quills protect trees from rabbits and voles, and also provide an ideal growing environment for the tree.*

Above: *Scarecrows are an attractive method of deterring birds and occasionally large pests away from your vegetable garden. They should be moved around the garden and redressed so that the birds do not get too accustomed to them.*

THE ORNAMENTAL GARDEN

Unlike the more fleeting nature of crops in the kitchen garden, the plants in the ornamental garden are usually permanent, which means that a stable ecology may be set up very easily. In fact, lawns, shrubs, trees and flower beds can all be adapted to organic culture. The ornamental garden needs flair and imagination, not only in its design and construction, but also in the types of materials it uses; there is little point in creating an organic garden with materials treated with toxic chemicals and wood that was logged from a tropical rainforest.

Left: *A flower garden in the height of summer provides colour and a natural food source for garden insects.*

Above: *Flowers are like fuel stations for insects, providing them with nectar in return for pollination.*

Above: *There is an absolutely staggering range of flower forms in nature, many of which are highly ornamental.*

Above: *Careful selection by gardeners has led to the development of many large and showy varieties of flower.*

THE GARDEN FRAMEWORK

It would be a great shame if we spent a lot of time and effort creating an organic garden, only to find that we had done so at the expense of another habitat or the wider environment. A truly organic approach to gardening should extend beyond the way we grow the plants and maintain the beds and borders to include all the elements of the garden, from the hardscaping materials to the furniture.

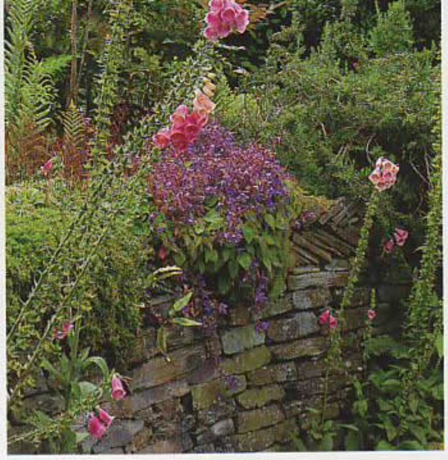

Above: *A natural dry-stone wall – assembled without mortar – can quickly become colonized with showy flowers.*

ECOLOGICALLY SOUND MATERIALS

Pictures of logged rainforests, eroded soils and drought-ravaged landscapes are all too familiar. So familiar, in fact, that we run the risk of becoming immune to their shocking truth. Many habitats around the world face increasing pressure from human activities. While there is often little we can do as individuals, we can at least try not to contribute to the demise of such habitats. Rainforests, the source of much of the wood used for garden furniture, are often clear felled, resulting in the loss of much of the biodiversity that they contain.

For this reason, it is important that we use ecologically sound materials. Numerous suppliers claim that they offer ecologically sound products, gathered from sustainable sources. You should check the ecological credentials of the items you buy wherever possible. Buying materials from local sources can help. Always ask who the supplier is, where the materials come from and how their extraction impacts upon the environment. Rocks for alpine gardens, wood for posts and fences, mulch material and potting mixes can all be detrimental to the environment if not carefully sourced.

It is also worth finding out what chemical treatments the material has undergone. Some wood is treated with preservative that may leach out into the soil given time. Potting mixes that are purchased may also contain chemical additives.

Even when products are environmentally sound, it is always worth finding out if the same material (or one that is sufficiently similar) is available from a local source. Transporting materials over great distances, particularly those that are bulky or heavy, involves the use of a considerable amount of fossil fuel. One of the cornerstones of sustainability is to think globally while acting locally. It may even be possible to see the products being made if you do buy from local sources.

Ultimately, of course, the closest source is that which you produce yourself in your own garden. Any organic gardener should, therefore, aim to be as self-sufficient as possible.

PLANTS SUITABLE FOR COPPICING

Most broadleaved trees can be used for coppicing, although some are more suitable than others as the rate of growth varies so widely.

In general, most conifers are not suitable for coppicing because they do not readily re-grow from a stump.

PLANT SPECIES	SITE	POTENTIAL USES
Wattle (*Acacia*)	Warm dry situations with good drainage. Especially tolerant of low fertility.	Hurdles and barriers (wattlework). Excellent in warmer climates where other weaving species will not prosper.
Gum (*Eucalyptus*)	Hot dry situations especially where soil is well drained and seasonally dry.	Durable fence posts, stakes and young ornamental growth.
Willow (*Salix*)	Tolerant of wet situations and heavy soils. The roots are notoriously aggressive and may damage nearby drains or foundations.	Weaving work, building and fencing. In a small garden, dogwood is a good substitute for weaving work.
Hazel (*Corylus avellana*)	Any garden soil with good drainage and a pH over 6.5. Tolerates frost, shade, and exposure.	Wattle work, weaving, pea and bean sticks.
Sweet chestnut (*Castanea sativa*)	Light, slightly moist, acid garden soils. Protect from frost and exposure.	Wattlework and fence construction.
Dogwood (*Cornus*)	Most garden soils but dislikes exposure and can sucker aggressively.	Red, orange or lime-green shoots suitable for basketry and other fine rustic weaving work.
Elder (*Sambucus*)	An adaptable species that will thrive in a variety of situations.	Twigs are useful for a number of garden projects and the berries and flowers can be used to make wines and cordials.
Honeysuckle (*Lonicera*)	Likes most soils and thrives with the roots in shade and the topgrowth in full sun.	Stems are a traditional basketry material.

USING RECYCLED MATERIALS

Recycling plays an important role in the organic garden. Waste materials from the kitchen and garden are added to the compost heap; plastic bottles can become mini cloches; and old carpet makes a useful protective cover. Old building materials, such as bricks, slabs, quarried stone and wooden railway sleepers (ties), can all be re-used in the garden.

SUSTAINABLE DESIGN

The more self-sustaining your garden is the better. Gardens that rely heavily upon recycled material – compost, leaf mould and pea sticks, for example – produced within their own confines are naturally self-sustaining. Some external input will always be needed but the garden's long-term success is assured. In short, you will have a sustainable organic garden.

BIO-ENGINEERING YOUR PLOT

This term describes the use of living material to construct the garden framework. It ensures that you make the most of the natural capabilities of some plant species, such as willow, that have been used for centuries for building and fencing. Other examples include using grass for stabilizing soil on slopes or woody plants for hedging. Living barriers do not rot and are self-sustaining.

WILLOW WEAVING

1 *Willow comes in various lengths. The shorter lengths are finest and suited for fine-quality work. This 2.1m (7ft) bundle of "weavers" is the best length for fencing. Thicker uprights are also needed to form the vertical columns to weave around.*

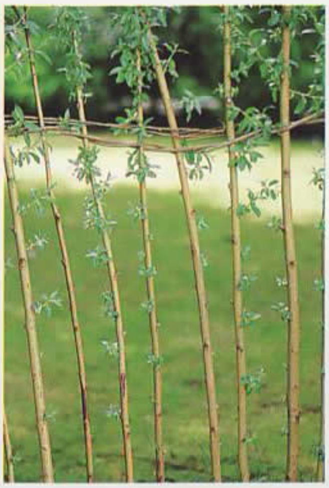

2 *Place the uprights in the ground about 10–20cm (4–8in) apart, ensuring that they are firmly secured. Use a metal spike to make the holes if the ground is hard. The stakes at each end must be the largest and need to be firmly secured.*

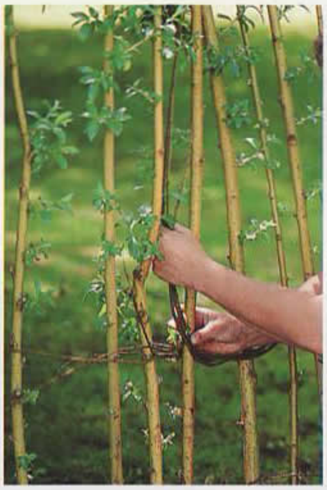

3 *Groups of four weavers are used together. Place two in the first gap (between the uprights) and the second pair in the next gap along. Take the back pair of weavers in front of the vertical upright, behind the next upright and to the front again.*

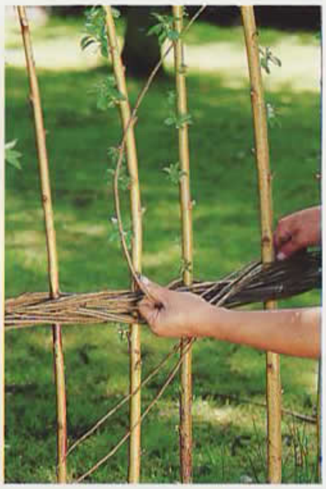

4 *Pick up the other set of weavers and do the same again. Add more weavers as they become thinner and slot these into the uprights to anchor them in. This will give the cross-over effect to the previous weavers.*

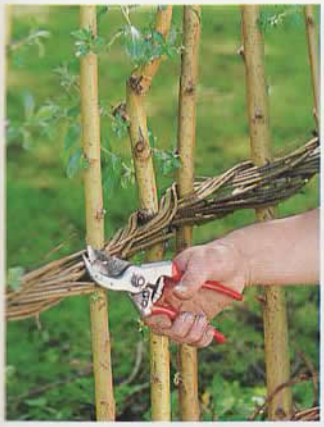

5 *Use successive groups of four weavers and continue to weave. Work from the front in order to ensure that the willow fence is neatly finished. Cut off any stubs as you progress.*

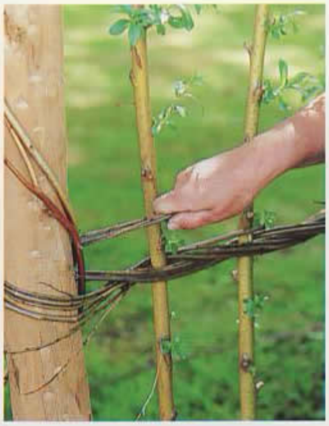

6 *At the end of the panel, wrap all four weavers round the post and begin weaving in the other direction. Ensure the weave is tight here and avoid making joins in the weave on the end posts.*

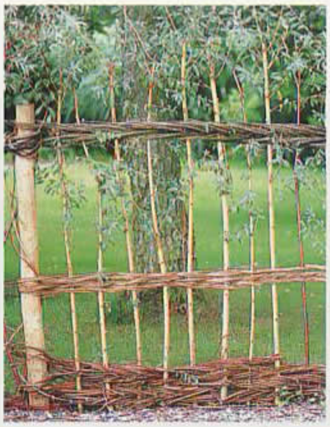

7 *The panels can be solid and used to form a garden boundary. They may also contain spaces, as shown, and act as dividers, while not obstructing the view beyond.*

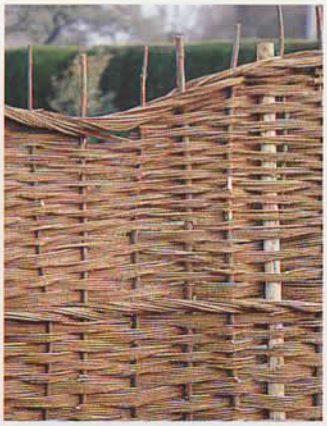

8 *The finished willow fence will last for many years, especially if it is adorned with climbers such as ivy. Willow fencing encourages wildlife as well as making an attractive feature.*

ORNAMENTAL LAWNS

In spite of the growth in popularity of hard surfaces, such as decking and paving, a well-tended ornamental lawn is still regarded by many as the quintessential feature of a beautiful garden. The lawn provides a safe surface on which children can play as well as an area for relaxing. In the organic garden, the lawn also provides a unique habitat for a wide range of wildlife. There is no reason why you should not have a fully organic lawn that looks pleasing while also contributing to the garden's diversity.

Above: *A well-manicured lawn is the result of regular care and maintenance. Mowing stripes in the lawn is a finishing touch.*

WHAT MAKES A GOOD LAWN?

Some gardeners believe that a good lawn is the product of time, patience and effort, which is true of highly manicured lawns. Others believe that an attractive, but low-maintenance, lawn is the ideal. Neither of these lawns is an impossible dream in the organic garden, but consider whether a lawn can withstand the wear and tear it is likely to receive in a small garden. In this case, it might be better to consider an alternative such as paving or gravel.

You should also consider how much time you are willing to spend maintaining the lawn. Short, highly ornamental lawns need more regular maintenance than longer grass areas, but all lawn areas require some degree of maintenance.

CAN I HAVE AN ORGANIC LAWN?

Lawns, like any community of plants, can be grown quite successfully the organic way. The key to success lies in keeping the grass plants healthy. Lawn grasses are naturally very resilient and competitive plants, well able to withstand a great deal of abuse when they are healthy. Provided they get sufficient light, water, nutrients, air and space, you will be rewarded with an attractive lawn.

TYPES OF LAWN GRASSES

The grasses below, if present in seed mixes or turf, should prosper in the given conditions. Common names of grasses, while frequently used, are not the same in all countries and the Latin names are given as a definitive reference.

Coarse-leaved species
Large-leaved timothy (*Phleum pratense*)
Meadow grass (*Poa pratensis*)
Perennial ryegrass (*Lolium perenne*)
Rough-stalked meadow grass (*Poa trivialis*)
Small-leaved timothy (*Phleum bertolonii*)
Tall fescue (*Festuca arundinacea*)
Wavy hair grass (*Deschampsia flexuosa*)

Drought-tolerant species
Chewing's fescue (*Festuca commutata*)
Hard fescue (*Festuca longifolia*)
Meadow grass (*Poa pratensis*)
Red fescue (*Festuca rubra*)
Sheep's fescue (*Festuca tenuifolia*)
Slender creeping red fescue (*F. rubra litoralis*)

Fine-leaved species
Browntop bent (*Agrostis tenuis*)
Chewing's fescue (*Festuca commutata*)
Hard fescue (*Festuca longifolia*)
Red fescue (*Festuca rubra*)
Sheep's fescue (*Festuca tenuifolia*)
Slender creeping red fescue
(*F. rubra litoralis*)

Shade-tolerant species
Chewing's fescue (*Festuca commutata*)
Hard fescue (*Festuca longifolia*)
Red fescue (*Festuca rubra*)
Wavy hair grass (*Deschampsia flexuosa*)

Species for damp conditions
Large-leaved timothy (*Phleum pratense*)
Rough-stalked meadow grass (*Poa trivialis*)
Small-leaved timothy (*Phleum bertolonii*)

Wear-tolerant species
Browntop bent (*Agrostis tenuis*)
Hard fescue (*Festuca longifolia*)
Sheep's fescue (*Festuca tenuifolia*)
Tall fescue (*Festuca arundinacea*)

CARING FOR AN ORGANIC LAWN

Each grass plant occupies just enough space to grow and closely "knits in" with its neighbours. The taller the plants grow, the lower the number of plants in any one space becomes. As they grow taller, their needs increase and the most competitive plants win. Regular cutting results in a dense coverage with short green leaves and a high number of plants in a given area. The more regularly you cut grass, the more vigorously it grows. Irregular cutting means fewer plants in the area. Less frequent removal of the green material at the top of the plants makes the leaf bases white or yellow and also results in a patchy lawn.

A lawn that is cut regularly will be greener and more attractive than one that is cut more infrequently. In addition, short-cut lawns are an ideal habitat for many garden creatures, particularly birds such as starlings and thrushes. Short grass can even be flower-rich and benefit bees and butterflies.

However, tall grasses favour visiting wildlife and can become an attractive feature in their own right, but they are not ideal for sitting out or for children to play on.

CHOOSING GRASS PLANTS

Once you know what sort of lawn you want, then you must choose the right grass species for your site and what you want to use it for. Some grass species, such as ryegrass, are hardwearing, while others, like the fescues, are drought-tolerant and need less mowing. Trying to grow the wrong grass species for your garden environment will only work if you go to great lengths to tend it and, even then, it will probably be extremely susceptible to damage. Heavy clay soil that becomes wet in winter will not support a close-mown community of fescues, but would be ideal for a ryegrass and bent mixture.

HOW TO SOW A NEW LAWN

1 *Dig the ground thoroughly, removing deep-rooted perennial weeds. Rake the soil level. Use pegs marked with lines drawn 5cm (2in) down from the top as a guide, having checked with a spirit (carpenter's) level on a straightedge that the pegs are level.*

2 *Allow the soil to settle for a week or so, and then consolidate it further by treading it evenly in order to remove large air pockets. The best method by which to do this is to shuffle your feet methodically over the whole area, first in one direction, then at right angles.*

3 *Rake the consolidated soil in order to produce a fine, crumbly structure that is suitable for sowing seeds. If you can, leave the area for a couple of weeks to allow any weed seeds to germinate. Hoe off the weed seedlings and leave them to die before you apply the grass seed.*

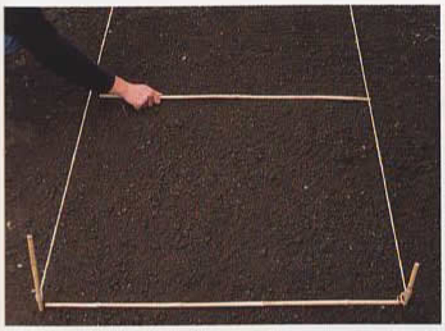

4 *Use string to divide the area into clearly demarcated strips. The strips should be approximately a metre (yard) wide. Divide the strips into squares with bamboo canes or stakes. Move the canes along the strips as you sow.*

5 *Use a small container that holds enough seed for a square metre (yard). Make a mark on it if the amount only partly fills the container. Scatter the seeds as evenly as possible with a sweeping motion of the hand.*

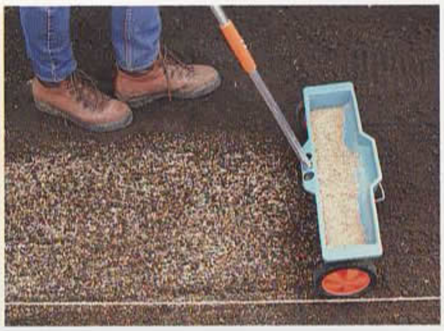

6 *Hire or buy a calibrated granular fertilizer spreader to sow large areas quickly. Check the delivery rates over sheets of paper first and adjust the spreader until the correct amount is being applied per square metre (yard).*

HOW TO LAY A LAWN WITH TURFS

1 *Dig and consolidate the soil as for seed. There is no need to leave it for a few weeks to allow weed seeds to germinate; the turf will prevent them from sprouting. Start by laying the turf along a straightedge.*

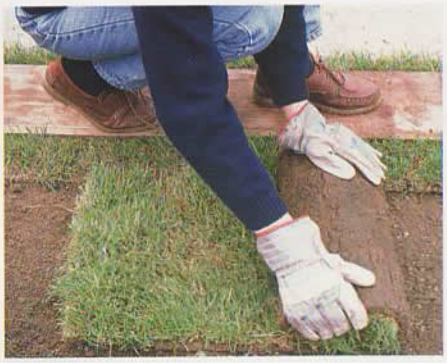

2 *Stand on a plank while you lay the next row, as this will distribute your weight. Stagger the joints between rows to create a bond like brickwork. Turf in a long roll will have fewer joints, but these should not align.*

3 *Tamp down each row of turf to eliminate air pockets with the head of a rake, then roll the plank forwards to lay the next row. Brush sandy soil, or a mixture of peat substitute and sand, into the joints to bind the turfs.*

LAWN MAINTENANCE

A rich green lawn is coveted by many gardeners, but a certain amount of maintenance is needed to achieve this. The secret of a healthy and attractive lawn lies in adopting a regular routine. Applying systematic care will also make the lawn relatively easy to maintain. Mowing is the most important and frequent operation carried out, but other practices, including watering, feeding, aerating and top-dressing, will also need to be done to keep your lawn in good condition.

Above: *Mowing will help to keep your lawn both green and healthy but it must be done frequently for the best results.*

MOWING

Where mowing is concerned there is a simple rule that you should always remember. The more grass that you remove by mowing, the more rapid its regrowth will be. The cutting height should be higher during late autumn, winter and early spring when little growth is taking place, but may be lowered in the warmer growing season. Raising the cutting height can also alleviate drought-stress and can help lawns stay green during hot dry summers. As a general rule, you should never remove more than a third of the leaf at any one mowing (unless of course you are cutting down an area left deliberately long for wildlife).

Grass can produce 2–3mm (1⁄12–⅛in) of growth over a 24-hour period in ideal conditions. The single most important factor is that you should mow regularly. Mowing will need to be less frequent when growth is slow, as in periods of dry or cool weather. Remember that lawns may need cutting in mild periods during the winter.

TURF PROBLEMS

Numerous problems periodically affect turf. The organic approach emphasizes the importance of maintaining strong healthy grass and removing the causes of ill health.

Organic control of turf diseases

- Maintain a vigorous growing sward.
- Control moisture to avoid humid surface conditions, e.g. by brushing.
- Ensure free movement of air.
- Regular aeration treatments.
- Reduce thatch.
- Box off clippings where possible.
- Avoid excessive nitrogen combined with cool conditions (in autumn).
- Avoid lime where possible (test water and top-dressings for lime).
- Adopt a balanced fertilizer regime.
- Sterilize loams and top-dressings.

Organic control of turf weeds

- Aerate to improve surface drainage and enhance grass vigour.
- Hand-dig persistent weeds.
- Mow and collect clippings regularly.
- Scarify the lawn surface.
- Remove earthworm casts.

LAWN CLIPPINGS

Grass clippings contain up to 3% nitrogen, 0.7% phosphorus and 2% potassium by dry weight. Returning clippings directly to the grass surface during mowing promotes the recycling of nutrients and "feeds" the lawn. Clippings also add organic matter to the soil, which will retain water, thus making the lawn more drought-resistant and helping to conserve water. On the downside, grass clippings can make turf more susceptible to disease and soil-borne pests and lead to a build up of "thatch" (a mat of partially decomposed grass leaves above the soil surface). Grass clippings should always be removed if they are excessively thick and will restrict future growth. This is especially true if the grass has not been cut regularly enough.

CONTROLLING THATCH

Thatch is the general term used to describe the layers of organic fibrous material found in turf. It is a perfectly natural component of turf and is desirable to a certain extent because it increases resilience to wear and drought tolerance. Too much thatch, however, can cause an increase in disease, localized dry spots, leaf yellowing, proneness to scalping (during mowing) and a soft spongy surface. It may also affect the tolerance of grass plants to heat, cold and drought.

Top-dressing and avoiding the over-application of nitrogen can help to reduce the build up of thatch, as can liming on acidic soils. Thatch can also be physically removed. This activity is commonly referred to as scarification but is also known as vertical mowing, power-raking or de-thatching. Scarification not only removes thatch, but also removes or controls moss and creeping weeds. It can also help air and

Above: *Grass clippings, leaves and other debris form a thatch at the base of grasses which can stifle them. Remove it with a spring-tine rake. Raking also removes moss.*

Above: *In autumn, rake fallen leaves into piles and scoop them up with a pair of boards. Choose a still day when the leaves are dry to make the job pleasant.*

water to enter the soil. Scarification must only be carried out when the grass is actively growing.

TURF AERATION

Aeration, put simply, means allowing air to get into the soil. In a lawn, it can improve surface drainage, improve soil air supply and relieve compaction. It also makes the lawn more drought-resistant because of improved, deeper grass root growth. Thatch will be reduced due to the increased microbial activity that is the result of better-aerated soil. Improved drainage leads to a warmer soil that will stimulate root growth and will help release unwanted chemicals and gases from the soil.

You should aerate the turf when the soil is moist, but not wet because it will then damage the soil structure. Aerating the lawn when it is too dry can be difficult if the soil is hard. Small areas can be done by hand, with the traditional garden fork proving very effective for relieving localized compaction.

ROLLING THE LAWN

Rolling a lawn gives a smooth, flat and level surface, but it is frequently overdone and invariably leads to soil compaction, loss of structure, poor drainage, poor aeration and reduced root growth. It is important never to roll a lawn when it is wet.

TOP-DRESSING MATERIALS

Numerous materials can be used as bulky top-dressing and each has its own benefits for both soil and grass plants.

Charcoal Sometimes used to improve surface drainage and ventilation.

Garden compost and leaf mould Improves the moisture-holding capacity of a rootzone while being relatively sterile. Often best used in mixtures. Homemade compost should be left for sufficient time to decay. Rich in trace elements.

Sand Porous and compaction-resistant, sand is used as a component for most top-dressing mixes and on its own.

Topsoil Used in "composts" for top-dressing and on its own for fine-turf areas if it contains sufficient sand.

AERATING THE LAWN

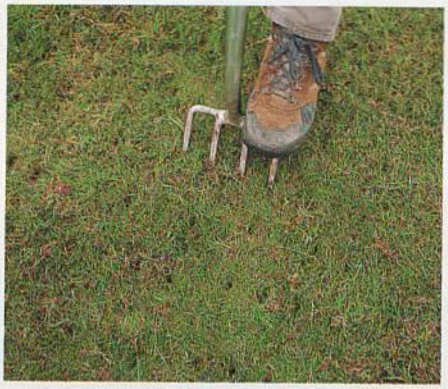

1 *Poor grass growth could be because the soil is poorly drained. You can aerate the lawn by pushing the prongs of a fork into the ground.*

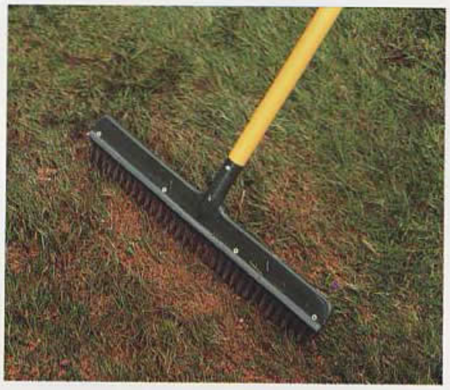

2 *Gently brush a soil improver, such as sharp sand or a mixture of soil and sand, into the holes made by the fork.*

TOP-DRESSING THE LAWN

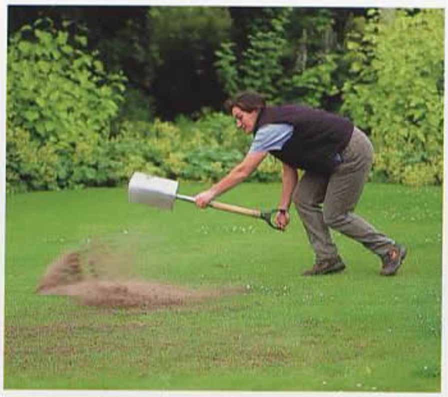

1 *Scatter dry topsoil or top-dressing mix evenly over the surface of the lawn using a shovel. As you throw the shovel outwards in a wide arc, twist it through 180° by rotating the handle. This will help you to spread the mix evenly.*

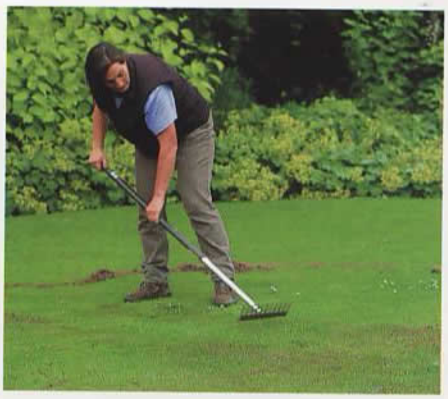

2 *Using either the flat side of a rake or a stiff broom or besom, work the top-dressing into the surface with short, even strokes. Only apply a top-dressing when the grass is actively growing.*

BRUSHING

Brushing scatters worm casts, grass clippings and other debris on the surface. It also disperses dew and "dries" the surface of the lawn, making it easier to obtain a clean cut with the mower. Frequent brushing can also act as a mild scarifier, lifting the grass blades ready for mowing. Brushing is also used to "work in" bulky top-dressing into the turf surface. A stiff broom or a besom is the most commonly used tool for brushing turf.

TOP-DRESSING

This is the application of a bulky material to the surface of the lawn and usually has no or only minor nutritional value. It is carried out to help level the surface or to improve the nature of the soil. It may also help to reduce the build up of thatch.

Top-dressing should be applied when both it and the surface are dry. It must be thoroughly incorporated into the turf using a brush or besom. Top-dress during the growing season so that the grass has a chance to grow through it and never apply so much that it smothers the grass.

Top-dressing is easier to work into the sward if the area has been mown and scarified first. If it is being carried out as part of autumn renovation work you can mow the area closer than normal. Make sure that you use the same type of material each time you apply top-dressing.

FERTILIZING YOUR LAWN

Organic fertilizers are applied to turf to ensure that there are sufficient nutrients available for healthy, sustained grass growth. The nutrient reserve within the soil is continually being diminished by plant uptake and removal of clippings. Nitrogen is the most important nutrient as it has the most effect upon growth and development. In most situations only nitrogen and potassium will need supplementing by fertilizer application. Most soils have adequate reserves of phosphorus and micronutrients. Organic fertilizers are available as pre-formulated dressings, although fine fertilizers such as fish, blood and bone are quite suitable for early-season application.

Soluble forms of fertilizer (compost teas or worm liquid for instance) should never be applied when grass is not actively growing. The first application of fertilizer in the spring helps to boost grass growth and recovery from winter and quickly improves the appearance of the turf for summer. Closely cut lawns need regular applications every four to six weeks through the growing season. Fertilizer dressing for use in late summer must be low in nitrogen, as an excess will invariably lead to soft growth that is prone to frost damage.

Apply lawn fertilizer evenly over the area to ensure uniform turf grass growth. Poor distribution may lead to scorching in areas that receive too much fertilizer. The best time to apply fertilizer is during a dry interval in a showery weather period when rainfall will wash it into the sward. Water the lawn if there is no rain for 24 hours.

Above: *Feed lawns using an organic fertilizer mix formulated for the season: spring and summer feeds have more nitrogen than autumn ones.*

WATERING LAWNS

Grass becomes dormant during prolonged dry weather. This may lead to invasion by weed species, but the main concern of most gardeners is that the lawn keeps its colour. Raising the height of cut may reduce this problem, but, in prolonged dry periods, the only real solution is to water. Well-aerated lawns often allow the quickest entry of moisture and lose less water to evaporation. Hand watering may be feasible for smaller areas, but it can be time-consuming. If you live in a drought-prone area, the best option is to choose drought-resistant species.

TURF GRASS DISEASES

Fungi cause the majority of diseases. The best way of controlling diseases is to keep the grass growing strongly by choosing the right species for the conditions and ensuring that the lawn is well maintained. Not all turf and soil fungi are harmful, however. Many are beneficial, aiding the decomposition of plant materials and organic matter and so releasing nutrients.

HOW TO REMOVE WEEDS

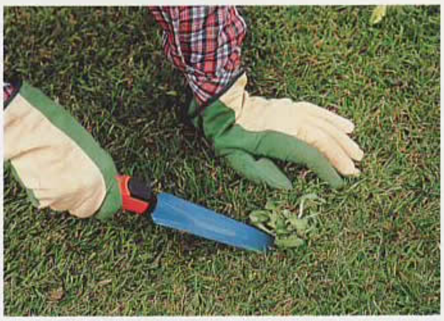

1 *Use a special weeding tool or knife to prise out weeds. Push the tool in next to the plant and lever out as you pull. Even deep-rooted plants can be removed like this.*

2 *Make any necessary lawn repairs. If you have had to lift a lot of weeds growing close together, leaving a bare patch in your lawn, sprinkle grass seeds over the area.*

HOW TO CREATE A MOWING EDGE

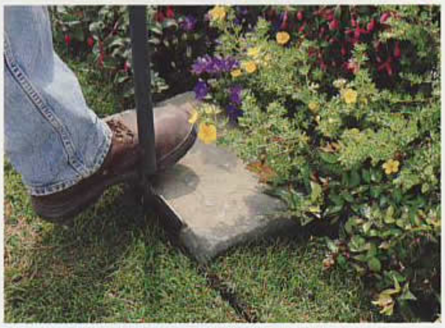

1 *An edge of bricks or paving slabs prevents flowers smothering the lawn. Mark out the area to be lifted, using the paving as a guide. To keep the edge straight, use a half-moon edger. Lift the grass by slicing it with a spade.*

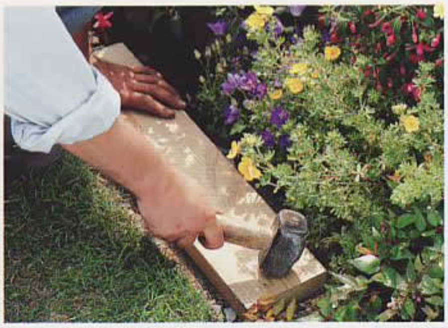

2 *Remove enough soil to allow for the depth of the slab or brick and a few blobs of mortar. Make a firm base by compacting gravel or a mixture of sand and gravel. Use a plank of wood to make sure it is level.*

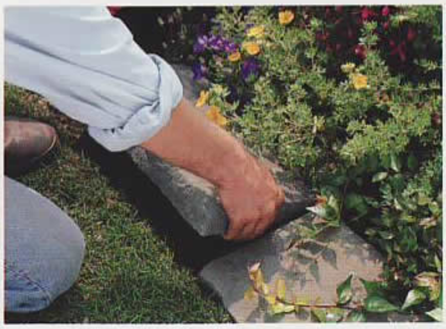

3 *Bed the edging on mortar for stability. As it will not be taking a heavy weight, just press the slabs on to blobs of mortar and tap level. Lay the slabs evenly and flush with, or very slightly below, the lawn. Check levels.*

REPAIRING A LAWN

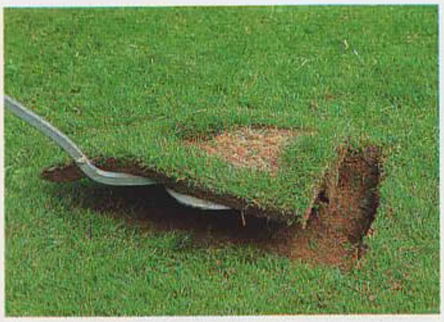

1 *Lift the area of damaged turf using a large shovel to beyond the point of wear or damage.*

2 *Cultivate the soil beneath in order to remove compaction or, alternatively, remove all the contaminated soil.*

3 *Add a layer of new, good-quality topsoil using the shovel.*

4 *Lightly firm down the topsoil with the back of a rake.*

5 *Sow grass seed over the newly added soil and gently rake in the seed.*

6 *Water in the seed and keep the patch moist until the new grass is established.*

REMOVING MOSS

Moss readily establishes on areas where grass growth is weak. Moss growth is favoured by a moist surface, mowing too close, surface compaction, shade, low soil fertility and excessive pH. Removing moss is palliative and moss soon returns unless the health of the grass is restored. The best solution is to find the cause and remedy it.

LAWN REPAIRS

There are two methods for repairing turf areas: seeding and turfing. Turf has an instant effect, but seeding is more economical over larger areas. Grass established from seed is usually more hard-wearing.

Repairing with turves Mark out the area and use a half moon and turfing iron or a spade to remove the worn/damaged area of turf. Fork over the soil to relieve surface compaction and add a suitable pre-seeding fertilizer. Firm the area by treading. Rake to obtain a tilth. Lay new turves on prepared soil, ensuring all joints are staggered and leaving the turves about 6mm (¼in) proud of the surrounding area to allow for future soil settlement. Top-dress and ensure the area is irrigated when necessary.

Repairing with seed Mark out the affected area and break up the soil by forking. Rake to produce a suitable tilth and "top up" soil levels if need be. Consolidate the surface by treading and apply a suitable pre-seeding fertilizer. Rake to a tilth and sow the seed in two (or more) directions. Rake in the seed, ensuring the area is level with the surrounding turf. Firm by rolling, if required.

Overseeding This method is used to thicken an existing lawn. Seed is scattered over the lawn, where the seed will then germinate, grow and fill-in the turf surface. Overseeding can be done by hand, broadcasting and scratching in the seed. When overseeding grassy areas, lower seed rates are used than those for new lawns. When soil conditions are dry, ensure that the seed is covered and in contact with the soil underneath. Keep the soil moist until the seed germinates and establishes.

Repairing damaged edges Mark out and lift an area of turf around the damaged edge. Fork over the soil in order to prepare a tilth and relieve surface compaction. Incorporate a suitable pre-seeding fertilizer and firm the area by treading. Rake to obtain a suitable tilth. Re-lay the turf with the damaged edge innermost. Top-dress the newly laid turves, working the top-dressing into the joints. Pay particular attention to the damaged zone. The top-dressing should ideally have a small quantity of seed mixed into it. Alternatively, the area may be overseeded once the turf is laid and top-dressing applied. Ensure the area is irrigated when necessary.

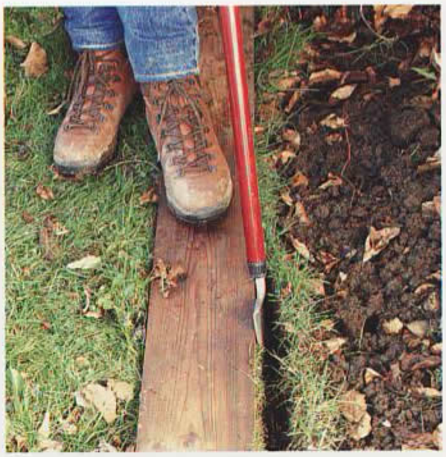

Above: *Insert a half-moon edging tool into the soil. Lever forwards to form a gulley with one vertical side against the lawn and one curved side against the border.*

WOODY PLANTS

Woody plants maintain permanent living structures above ground. They are unique in that their stems repeatedly thicken and strengthen, as is easily seen in the familiar growth rings in wood. Woody plants take many forms: they can be small ground-hugging shrubs or forest giants, and both the largest and oldest living things on our planet today are woody plants. They are the backbone of many gardens, providing height and structure as well as a feeling of permanence.

Above: *Woody plants, such as this flowering quince (Chaenomeles), often have very decorative and colourful flowers.*

BUYING WOODY PLANTS

Woody plants may be bought bare-rooted or rootballed, or in containers. All plants should be vigorous, healthy and suitable for the site conditions and intended use. Check that bare-rooted specimens are properly wrapped so as to prevent their roots drying out and ensure that trees and shrubs are free from any obvious pest or disease.

PREPARING THE GROUND

Organic matter is best incorporated across the whole area as a mulch after planting. This encourages the development of a healthy root system. Apply across the whole site during digging if the soil is sandy.

Adding fertilizer is only generally recommended when analysis reveals the soil to be deficient in particular nutrients. Woody plants set up mutually beneficial relationships with soil fungi and gain much of their nutrients as a result of natural nutrient cycles. Often, the best floral displays occur on soils of low fertility, making fertilizer unnecessary.

The key to woody plant growth usually lies in applying organic mulch on an annual basis. Leaf mould is arguably the best substance to use, but well-rotted compost or manure is almost equally useful. An application of bonemeal or similar organic base dressing before mulch is applied can prove to be beneficial when shrubs are newly planted.

PLANTING OUT BARE-ROOTED OR ROOTBALLED TREES

The soil should ideally be moist and friable, not frozen or excessively dry or waterlogged. Bare-rooted or root-balled specimens are usually planted during the dormant season. For deciduous broad-leaved species, this is in winter. An early planting results in better establishment in the following season, and often makes the

PLANTING A SHRUB

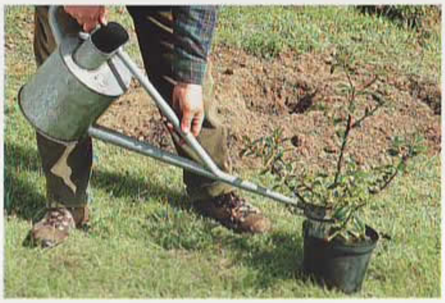

1 *Never plant a shrub that is stressed. Before you start planting, check that the plant has been watered. If not, give it a thorough soaking, preferably the night before planting. If this is not possible, ensure that the shrub is watered at least an hour before you plant. This allows the water to be absorbed thoroughly by the potting mix and will help the plant to establish.*

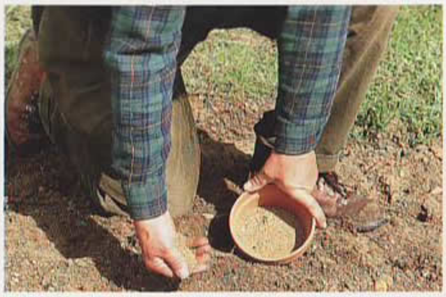

2 *If the soil has not been recently prepared, fork it over, removing any weeds. Add a slow-release fertilizer, such as bonemeal, wearing rubber or vinyl gloves if required, and fork this in.*

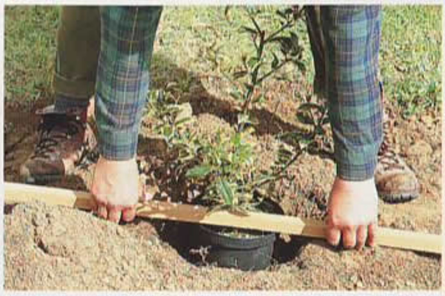

3 *Dig a hole wider than the rootball. Put the plant in the hole and check it is deep enough by placing a stick across the top of the pot. It should align with the top of the soil. Adjust the depth of the hole accordingly.*

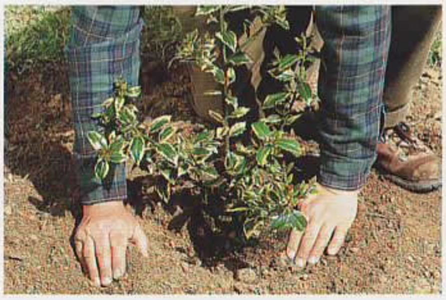

4 *Remove the plant from its pot, taking care not to disturb the rootball. If it is in a plastic bag, cut away the bag rather than trying to pull it off. Place the shrub in the hole and draw the earth around it. Firm the soil down with the heel of your boot and water in well.*

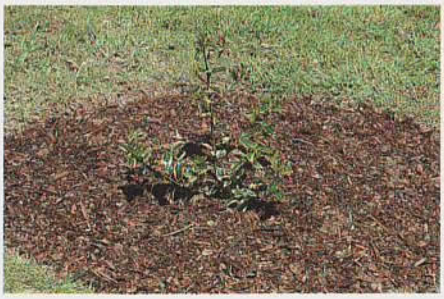

5 *Finally, mulch all around the shrub, covering the soil with an 8–10cm (3–4in) layer of bark or a similar material. This will not only help to preserve moisture but will also help to prevent any weed seeds from germinating.*

PLANTING A CLIMBER

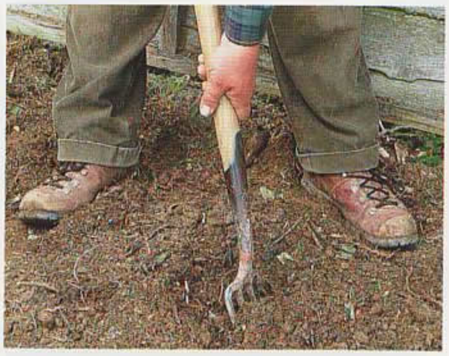
1 Dig over the proposed site, loosening the soil and removing any weeds. If the ground has not recently been prepared, work some well-rotted organic material into the soil to improve soil texture and fertility.

2 Add a general or specialist shrub fertilizer, if necessary, to the soil at the dose recommended on the packet. Work into the soil around the planting area with a fork. A slow-release organic fertilizer is best.

3 Water the plant in the pot. Dig a hole that is much wider than the rootball. The hole should be at least 30cm (12in) from the wall or fence. The free-standing canes will be angled towards the wall or fence.

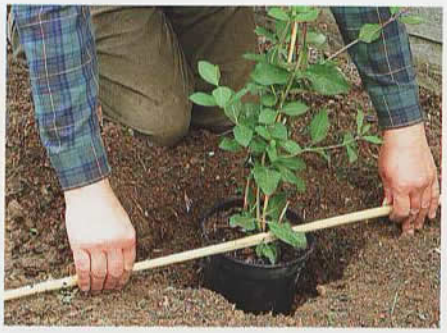
4 Stand the plant in its hole and place a cane across the hole to check that it is at the same level. Take the plant from the pot or cut the bag away. Holding the plant steady, fill in the soil. Firm as you go with your hands.

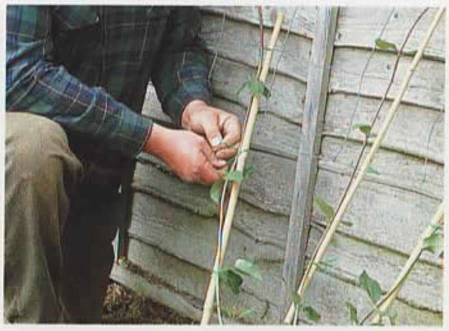
5 Train the stems up individual canes. Tie in with string or plastic ties. Even twining plants or plants with tendrils will need this help. Spread them out, so that they will ultimately cover the whole of their support.

6 Water the climber thoroughly. Put a layer of mulch around the plant to help preserve the moisture and prevent weed growth. Do not pile mulch up against the stems of the climber, however.

---

plants more resistant to spring drought. Evergreen specimens establish more readily if planted during early autumn or late spring in conditions in which enough moisture is available for rapid root growth.

Mark out the positions of the plants and dig each pit, allowing ample space to accommodate the roots, rootball or container size. This is usually a quarter to half the diameter again of the rootball. Take care not to smear the sides of a pit dug in heavy soil as this will effectively cause it to fill up with water during the wetter months, resulting in root dieback, subsequent shoot dieback and in extreme cases, the death of the whole plant.

Remove protection from the roots and place the tree into the hole. Prune any damaged tissue with a sharp knife or secateurs (hand pruners). Position the plant in the pit with roots well spread. Where trees require staking, the stake should be hammered into the ground before planting the tree to avoid damaging the roots. Position the stake on the side from which the wind most commonly blows.

PLANTING OUT CONTAINER-GROWN PLANTS

Container-grown specimens may be planted throughout the year, provided that the ground is sufficiently moist and adequate water supplies are available and used to irrigate the newly planted area. Other site-related stresses, such as high wind and hot conditions, may also have to be minimized where plants are planted outside of the optimum winter period. Preparation of the planting pit, and staking, are the same as for bare-rooted or rootballed trees, but the positioning of the rootball and treatment of the roots differs slightly. Remove the tree from the container and lightly shake out the rootball over the hole of excavated soil. Check the root collar area for girdling roots. These are roots that spiral around and may eventually damage the developing stem. Where present sever these cleanly with a knife and remove carefully out of the rootball. Tease out the lower roots using a fork. Place the rootball in the hole so that the roots are spread out and the surface of the potting mix is level with the ground. Tease out the circling roots at the edges of the rootball with a fork or by hand especially where plants are pot bound. Fill in and firm, staking where required.

Immediately after planting, remove any damaged, diseased or untidy growth. Do not remove excessive amounts of top growth, however, as this will reduce the plant's ability to produce food.

STAKING TREES AND SHRUBS

Large shrubs and trees larger than about 1.2m (4ft) may require staking. Low stakes are now recommended as they allow the top growth to move in a circular motion. This motion encourages rapid lower-stem thickening and the development of fibrous root systems that are essential for quick establishment and support of the plant.

PROTECTING NEWLY PLANTED SPECIMENS

Plants unused to cold or hot drying winds may suffer from drought stress. Cold drying winds may damage foliage, buds and branch tips, as well as the bark of thin-barked tree species unused to such exposure. Tree shelters protect trees from adverse conditions, providing an extremely suitable environment for the rapid growth and development of the tree. They are also useful in protecting plants from voles, rabbits and sometimes deer attack. They are usually designed to last three to five years before the material starts to degrade under the action of sunlight. Shrub shelters are similar in most respects to those for trees but are generally shorter.

STAKING A STANDARD ROSE

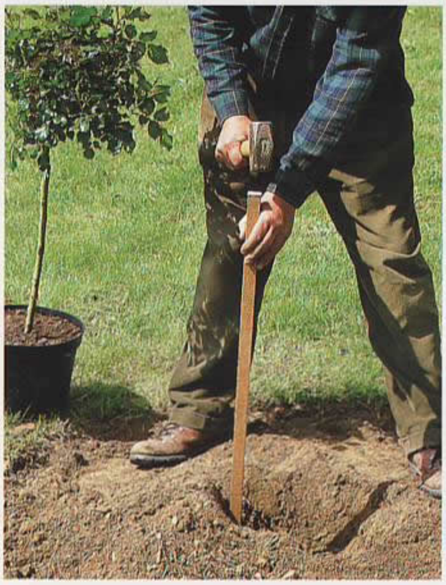

1 *For a standard shrub, make sure you use a strong stake. It should be of a rot-resistant wood. Firmly place the stake in the planting hole, knocking it into the soil so that it cannot move.*

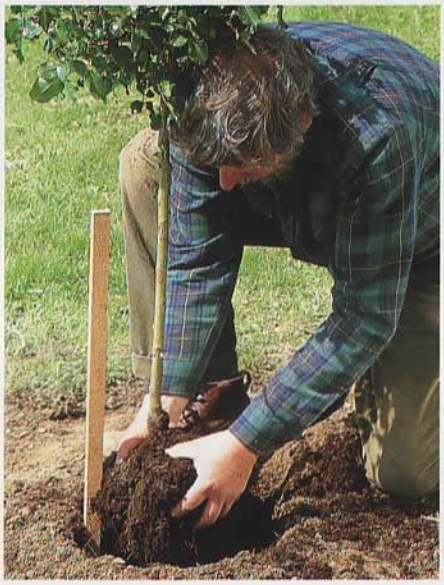

2 *Plant the shrub, pushing the rootball up against the stake, so that the stem and stake are approximately 8–10cm (3–4in) apart.*

3 *Firm the soil down around the plant with the heel of your boot.*

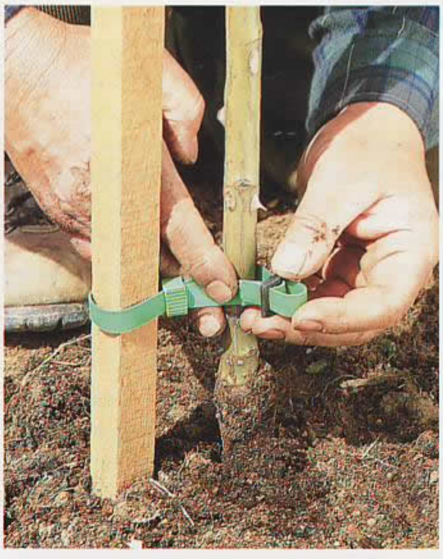

4 *Although you can use string, a proper rose or tree tie provides the best support. Fix the lower one 15cm (6in) above the soil.*

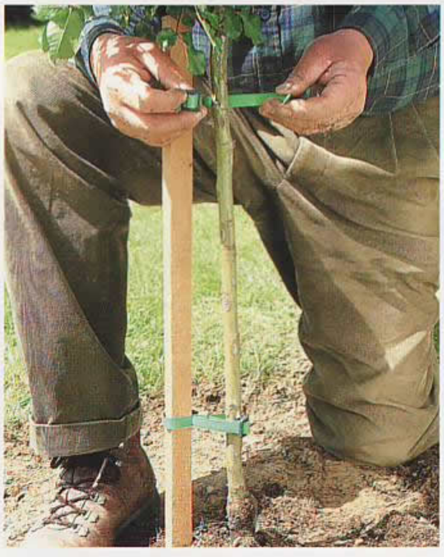

5 *Fix the second tie near to the top of the stake, but just below the head of the standard shrub.*

6 *Water the ground around the plant thoroughly and mulch with chipped bark or a similar material.*

Spiral guards, designed to protect larger trees from rabbits, can act as tree shelters for smaller plants. When used on larger trees, they are designed to expand as the stem diameter increases, thereby not constricting growth. Open-mesh guards, which are mesh rolls supported with two or more canes or stakes, provide alternative rabbit protection.

Re-firm roots of all newly planted trees and shrubs after wind or frost, using the heel of your boot, to exclude air pockets around the roots.

Irrigation should only be applied on very dry soils during the first year of establishment. Water is best applied only when absolutely necessary using a sprinkler to thoroughly drench the soil.

All shrub and tree bases should be kept weed-free during the first year. Mulching may help to achieve this, but some hand weeding may be necessary.

RENOVATING OLDER SHRUBS

Shrubs that become overgrown look unsightly and will often have a poor display of flowers. Many shrubs can, however, be rejuvenated by pruning. In certain cases the pruning programme can be staggered over a period of two or more years. Deciduous species, such as lilacs (*Syringa*), may be pruned after flowering, or more commonly during the dormant season. Evergreen shrubs, such as *Viburnum tinus*, are best treated in mid-spring. Remove all weak spindly growth and any badly placed branches or crossing stems. Remove stems to leave a balanced framework by cutting them back to around 30–45cm (12–18in), depending on the species and its vigour. Mulch using leaf mould or compost to 5–10cm (2–4in) in depth. Ensure that the plant is never short of water in the summer.

The following season the shrub will have produced a mass of new shoots. These should be thinned as necessary. After renovation, all shrubs should be pruned according to their normal requirements.

RENOVATING HEDGES

If hedges become overgrown or too wide through neglect or incorrect maintenance, they may respond to drastic pruning methods. The method is similar to that which is carried out on shrubs. Instead of cutting hard back to the base, however, only one side of the hedge is pruned back hard to the main stems of the hedge plants. One side of the hedge is left alone so there is enough foliage area left to stimulate new growth and recovery on the pruned side. This can also prove useful in providing cover for wildlife, especially nesting birds.

The process is repeated on the other side of the hedge the following year or, perhaps, two years later. Evergreen subjects (but not conifers) should be pruned in this way during early to mid-spring and deciduous subjects when they are dormant in late winter.

Feed using an organic mulch, such as farmyard manure, and water during the following season to ensure that the hedge survives such drastic pruning. This technique can be very successful for many evergreen plants as well as most deciduous plants.

Left: *Flowering trees such as these cherries (*Prunus*) can provide a welcome and refreshing splash of spring-time colour.*

RENOVATING WOODY CLIMBERS

Climbers that have not been pruned or trained properly often degenerate into a tangled mass of woody stems which look unsightly and give poor foliar or floral displays. Hard pruning will often rejuvenate the plant. In most cases, the plant will tolerate close pruning in early spring to within 30–60cm (12–24in) of the ground or its main framework of branches. Plants in poor health may not survive this treatment, in which case you may need to carry out the work over a period of two or more years.

Water the base of the plant well and apply bulky organic material. New growth must be trained and pruned according to the species. New growth can initially be trained as if it were a newly planted climber.

PROVIDING PROTECTION

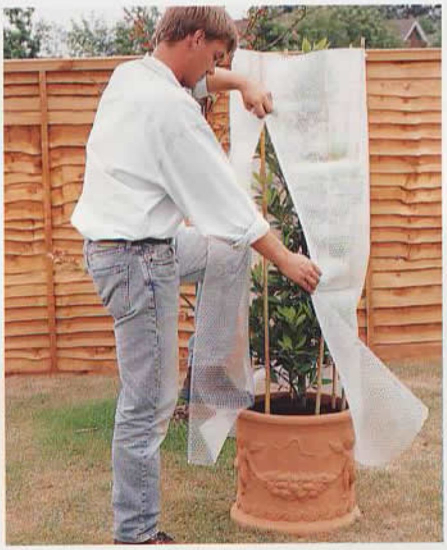

1 *Many shrubs need some winter protection. This shrub is in a pot but the same principles can be applied to free-standing shrubs. Insert a number of canes around the plant. Cut a piece of horticultural fleece, hessian (burlap) or bubble plastic to size, allowing room for overlap. Wrap around the plant, using a double layer for very tender plants.*

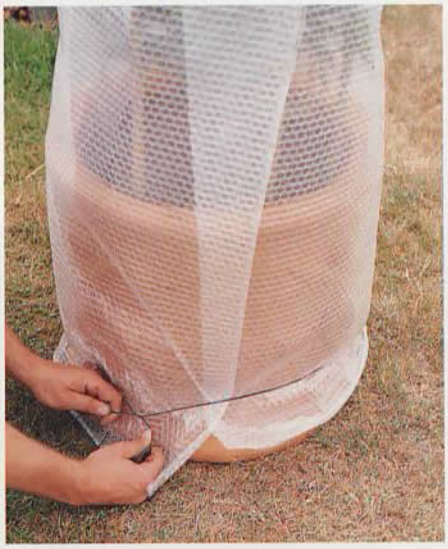

2 *Tie the protective cover around the pot or lightly around the shrub if it is in the ground. Horticultural fleece can be tied at the top because moisture can penetrate, but, if you are using bubble plastic, then leave it open for ventilation and watering.*

FLOWERS FOR BEDS AND BORDERS

Flowers are an essential part of the ornamental garden and there can be few sights that can rival a flower border at the height of summer. Understanding the lifecycles and horticultural needs of these plants is essential if you are to use them to their best effect in flower beds and borders. This section looks at planting and caring for herbaceous perennials, bulbs and annuals.

Above: *Few sights in the ornamental garden can rival that of a flower border in full bloom in mid-summer.*

HERBACEOUS PLANTS

The strict definition of a herbaceous plant is one that dies down to or near to ground level each year in order to avoid harsh and extreme weather conditions. In the garden this includes herbaceous perennials as well as grasses – which may not actually die down but never exceed a certain height – and bulbs.

Herbaceous perennials are non-woody plants that live for more than one year. They usually grow, flower and set seed before dying down to ground level in the autumn. They remain there until the next season when the cycle begins again. Herbaceous perennials may form sizeable clumps which are best divided every three to five years to maintain vigour and promote flowering.

Ornamental grasses contrast superbly with broad-leaved plants and narrow-leaved forms and are ideally suited for creating "dry" gardens.

The maintenance requirements for the vast majority of these plants are the same and they form the backbone of the ornamental garden by providing colour which, if planned properly, may benefit the garden all year round.

Planting out Herbaceous plants may be planted in either autumn or spring. The advantage of autumn planting is usually noticed on drier soils as the plants establish and are potentially more drought resistant. Those planted on heavier soils may benefit from spring planting, as heavy soils may be wet and cold – conditions that may cause the newly planted specimen to rot. Grasses are planted in the same way as "true" herbaceous plants. Bulb planting is different and is dealt with separately.

Pot-grown specimens should be watered at least an hour before planting. The ground should be dug and manured before being levelled ready for planting. A base dressing of bonemeal or fish, blood and bone may also be applied. Plant only when the soil is moist, but not waterlogged or frozen. Dig a pit larger than the rootball and ensure that the plant is well firmed, taking care not to compact the soil when it is returned around the rootball. Care should be taken immediately after to ensure that slugs and snails do not damage the crowns. Set traps and barriers around them to avoid this.

Thinning Removing weak shoots allows the plant to divert its energy into the remaining shoots, which will be sturdier and produce larger flowers. Thinning is carried out when the plant has reached one-quarter to one third of its eventual height and is useful for plants such as as delphiniums, phlox and Michaelmas daisies.

Stopping This is the removal of the growing tip of a stem in order to encourage the side shoots to develop. The tips may be

HOW TO PLANT PERENNIALS

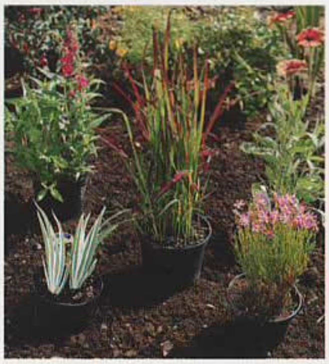

1 *Always prepare the soil first. Dig it deeply, remove weeds, and incorporate a fertilizer and well-rotted compost if the soil is impoverished. Most perennials are sold in pots, so arrange them to create pleasing associations.*

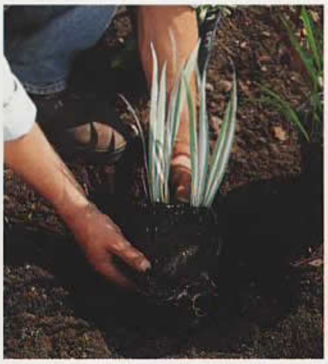

2 *Water thoroughly at least an hour before knocking the plant from its pot, then dig a hole. If the rootball is congested, tease out a few roots first. Work methodically from the back of the border or from one end.*

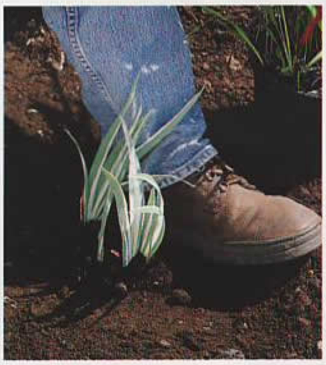

3 *Carefully firm the soil around the roots with the heel of your boot. This helps to remove any large pockets of air.*

4 *It is important that you always remember to water thoroughly after planting and keep the new plants well watered in dry weather for the first few months.*

pinched out when the plant has attained a third of its ultimate height. This encourages the buds on the leaf axils to develop and leads to more flowers being produced, although these will be smaller than if the one terminal bud had been allowed to flower. This technique is successful for a range of plants including *Helenium*, *Rudbeckia*, *Dahlia* and *Chrysanthemum*.

Dead-heading Unless the plant produces decorative seed heads or seed is to be collected for propagation purposes, it is often best to remove all flowers as they begin to fade. Further side shoots will then often develop and the flowering period will be extended. Notable exceptions to this are those plants that produce seed heads that are valuable food sources for overwintering birds. Examples of this include teasels, *Echinops* and many of the ornamental grasses.

Cutting back Some shrubby perennials (often referred to as sub-shrubs), such as *Chrysanthemum*, *Phygelius* and *Penstemon*, benefit from being pruned annually in early spring. This is especially true of sub-shrubs that are somewhat tender. The older twiggy and unproductive growth is cut hard back in order to promote the growth of new shoots that will flower in summer and autumn.

Feeding Few perennials require more than an annual fertilizer application if initial soil preparation has been thorough and included the addition of bulky organic matter. An organic fertilizer, such as bonemeal or fish, blood and bone, may be added to the soil in spring, preferably after rain.

Above: *When dead-heading, snip off the flowers cleanly where they join the stem. Cut back to the first set of leaves if a whole head of flowers has to be removed.*

CUTTING BACK

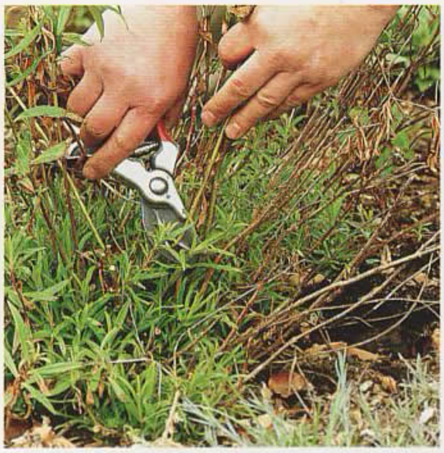

1 *Some herbaceous plants remain green throughout winter. Cut back to sound growth, removing dead and leggy material.*

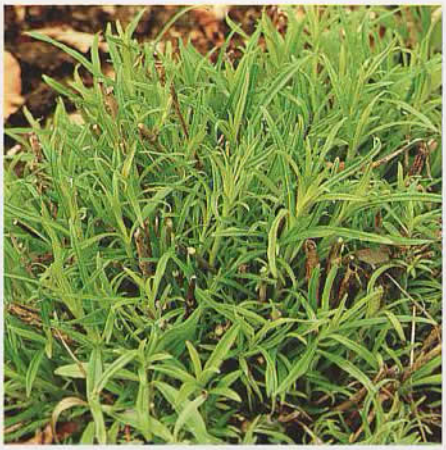

2 *Here, the old stems have been cut off so that they are level with the emerging growth, so as not to damage it.*

Bulky organic material, such as leaf mould or compost, applied in spring before shoot growth is also beneficial. Care must be taken not to apply this too thickly and not to cover the crowns of plants resting at the soil surface as these may rot off.

Renovating Herbaceous perennial plants should be lifted, divided and replanted every three to five years. Some fast-growing and vigorous plants, such as *Ajuga* and *Stachys*, may need to be divided every year. Herbaceous plants can become woody towards the centre and may lose vigour.

Lifting the plants for dividing will not only rejuvenate them but also allows for other border maintenance operations to be carried out. The site can be cleared of any weeds and organic matter such as compost or manure dug into the border. The process of division is beneficial to the whole border. In addition, it keeps the plant healthy and will prevent the over-vigorous growth of fast-growing species. Lifting and dividing is normally done while the plants are dormant in late autumn or early spring.

Transplanting Most perennials can be transplanted relatively easily. This should be done during the dormant season, which is usually in autumn or early spring. Some plants dislike cold wet conditions and these should be moved once the soil has warmed sufficiently to encourage growth. This is particularly true for any plants that are not fully hardy. Some plants that are relatively long-lived, such as peonies, resent being disturbed and will take two or more years to establish again after transplanting.

Plant replacement If possible, all planted areas should be designed to contain plants that have similar life spans and respond to rejuvenation techniques. Some species are relatively short lived and will die out before others within a bed. It is possible to rejuvenate beds on a set cycle, say every three to four years, and avoid gaps appearing in the periods between.

Plant losses can occur in severe winters when those that are normally regarded as hardy are exposed to very low temperatures. Half-hardy or tender perennials are even more susceptible. Some perennials are probably best lifted and stored over winter until the worst weather has passed, when they can be planted out again.

Above: *Lightly dig over the soil around the plants, removing any weeds. Avoid digging around plants such as asters which have shallow roots.*

ANNUALS AND BIENNIALS

As the name suggests, annuals are plants that grow, set seed and die, completing their lifecycle within one growing season. Some garden annuals may actually be perennials in nature, but are treated as annuals for ease of cultivation. Many half-hardy and summer bedding annuals fall into this category. Almost all spend a maximum of one calendar year in the ground before dying off.

Biennials are similar to annuals as they flower once and die, but differ in that they grow and establish in the first season before resting in a similar manner to herbaceous perennials over winter. They grow, flower and die the following season. Some plants that are grown as biennials (especially as bedding) are perennials. They are treated as biennials, however, because they tend to weaken during their second year and are therefore better re-started in order to maintain vigorous plants.

Annuals are used as displays in their own right or as small groupings of colour among more established plantings. Groups of annuals strategically placed together in, for example, containers can have a greater visual impact than if they were spread throughout the garden. Many annuals, especially those with composite flower heads, attract beneficial insect species into the garden and some are used as companion planting for vegetable crops. Annuals can also be a useful infill in new borders as they are easily removed or thinned out later when the perennials have grown and become established.

Planting annuals, biennials and tender perennials Hardy annuals can be sown *in situ* and then thinned. Some may self-seed freely through borders. If annuals are to be grown in pots or trays prior to planting, the planting is much the same as for herbaceous plants. Always ensure that the risk of frost has passed before planting out tender specimens and that these plants are acclimatized thoroughly beforehand.

PLANTING ANNUALS

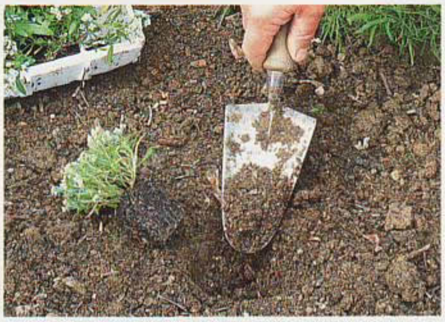

1 *Remove the plants from the pack. If they are attached to each other, try not to tear off too many roots when removing. Dig a hole wider and slightly deeper than the rootball.*

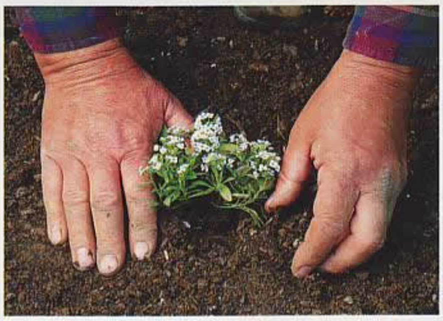

2 *Fill in the hole around the plant with soil and gently firm down with your hands. Water the plant with a watering can fitted with a fine rose.*

SPLITTING CONGESTED PLANTS

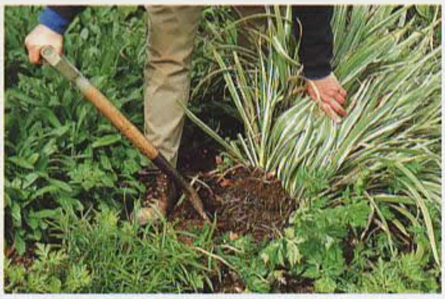

1 *Some perennials need splitting every few years to keep them flourishing. In spring, lift the whole of the clump from the ground.*

2 *Clean the ground thoroughly, removing any weeds, particularly perennial ones that have appeared since the border was last dug.*

3 *Incorporate some well-rotted organic material such as farmyard manure in order to rejuvenate the soil.*

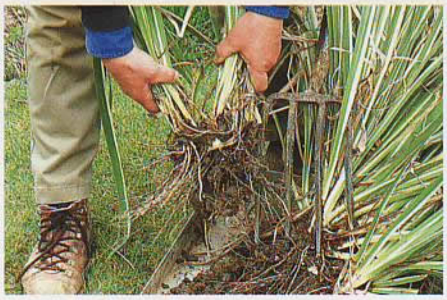

4 *Remove the earth from the plant and divide. Use the new growth around the edge, and discard the woody centre.*

5 *Replant some of the divisions, making certain that the roots are well spread out. Firm them in and then water.*

6 *Keep the plants watered until they have re-established themselves. They will soon fill out the gaps in the border.*

SOWING SEED IN OPEN GROUND

1 *Dig over the ground and break it down into a fine tilth using a rake. Do not work the soil when wet or it will become compacted.*

2 *Using a garden line as a guide, draw out a shallow drill. Use the corner of a hoe, a stick or a trowel.*

3 *If the soil is very dry, water the drill using a watering can and leave to drain. It should not be muddy for sowing.*

4 *Identify the row with a clearly labelled marker. This is important because when the row is backfilled it will be impossible to see where the seed is until it germinates.*

5 *Sow the seed thinly by hand along the length of the drill. Sowing thinly will reduce the amount of thinning that is required at a later stage.*

6 *Rake the soil back over the drill and lightly tamp it down with the back of the rake. When the seedlings emerge, thin them out in order to prevent overcrowding.*

Many hardy annuals thrive in poorer soils, and are undemanding and easy to grow. Dry banks are ideal for a whole range of species including poppies and nasturtiums, both of which flower more abundantly on slightly poorer soils.

The only "pruning" required for annual plants is dead-heading to promote new shoot and flower development, or stopping to promote side shoot development and multiple flower production as opposed to large single flowers.

BULBS, CORMS AND TUBERS

These rest below the surface and are modified buds, stems or roots. They emerge, often for only a short time to grow and flower, before retreating back down underground to escape harsh weather conditions.

Planting and maintenance Bulbs, corms and tubers can be lifted during their dormant period and transported and stored easily. Generally speaking, bulbs are planted in a hole that has been dug to about two to three times their length deep. Bulbs have a base plate, which is where the roots appear. This must always be placed in the hole roots down and should be in contact with the soil in the base of the hole. Placing fine sand in the base will ensure this. The top of the bulb (the neck) should always point upwards. Fill the hole firmly once the bulb is in the hole. After this the bulbs need little care and attention, although tall varieties may need staking and all should be dead-headed after flowering.

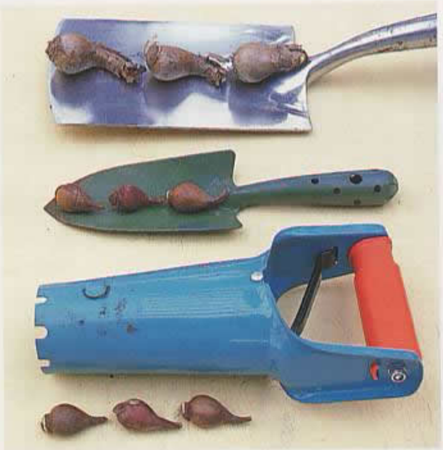

Above: *The rule is that bulbs should be planted at a depth of three times their own size. Exceptions include nerines, crinums and some lilies, which prefer to be planted with the top of the bulb at soil level.*

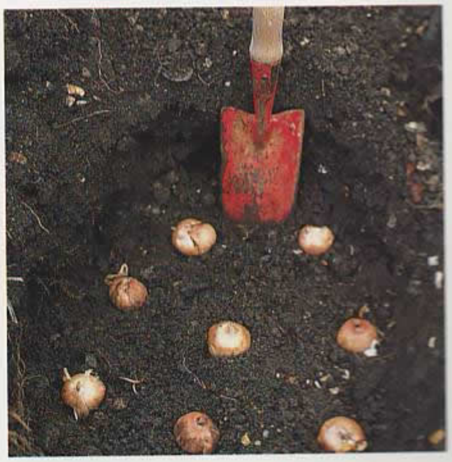

Above: *Dig a large hole, according to the requirements of the individual bulbs or corms. These gladioli corms are placed, base down and nose up, about 10cm (4in) apart. Gladioli corms need to be lifted in autumn.*

BEDS AND BORDERS

Whether you are creating a flower border from scratch or updating an existing design, there are a number of important factors that you will need to consider before you start planting the scheme. Begin by looking very carefully at the site and its conditions. This initial planning and preparation will help you in your choice of plants. Hardiness and suitability to the soil type, as well as the aspect of the site, are also factors that will determine which plants will thrive.

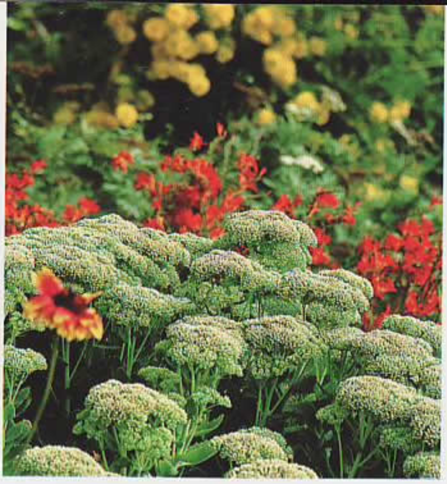

Above: *Flowers are not only decorative, but will also encourage pollinators and other insects into your garden.*

PLANNING AND SITING

Ideally, the bed or border should be positioned in a sunny site away from the shade of trees or buildings unless you intend to plant a border with shade-loving species. A hedge, wall or fence can provide valuable protection for your plants, and will create a warm microclimate for more tender plants. A sheltered spot will also extend the planting season. These structures may also provide a backdrop for climbers.

When you begin planning the border, ask yourself what you hope to achieve. Borders that are to be viewed from a distance will benefit from larger groupings of each plant type. Larger specimens can also appear very striking from further away. These can be used as character plants and provide a recurring theme through your border.

Above: *Combining strong colours such as orange and blue can produce striking planting associations.*

Left: *Herbaceous borders are easy to create and maintain, and can provide year-round interest if they are properly planned.*

Borders may have a particular theme. They may be planted with fruiting shrubs, plants to create a "tropical" look or fragrant plants such as philadelphus and lavender for rich summer fragrance. Colour also provides interesting themes. A golden border of yellow flowers in association with plants that have variegated or yellow leaves is striking. Remember that you can leave gaps in the planting for seasonal features such as bedding plants and bulbs. Decide what look you would like to achieve before buying any plants.

SEASONAL CONSIDERATIONS

Ask yourself when you want the border to look its best. Perhaps you would like a bright display of bulbs in spring followed by a colourful selection of annuals and bedding plants in summer. Alternatively, you may want to plan for a bed or border that provides year-round interest, in which case you will need to consider a selection of hardy shrubs or perennials that are appropriate for your garden conditions.

Make a list of all the plants that you would like to include and when they will look their best. Many plants have more than one season. For example, shrubs such as *Viburnum opulus* flower in early summer and have bright red berries in autumn. If you intend to use winter-stem shrubs such as *Cornus alba*, then choose varieties with interesting foliage, such as the golden form 'Aurea'.

COLOUR CONTRASTS

Once you have your list of desired plants, consider their colours. Contrasting colours can look effective. Try plants that have blue, purple or pink flowers, and choose as many examples of these as you can find with bronze foliage. A much brighter combination would call for hot vibrant colours such as yellow, orange and red.

Cottage gardens throw all caution to the wind, including a riot of contrasting and clashing colours to provide unexpected and beautiful combinations of hue and form. Alternatively, you might prefer a monochrome scheme, using flowers of one colour – white or cream, say, offset by a display of green foliage.

FOLIAR AND TEXTURAL EFFECTS

Placing plants with contrasting foliage next to each other can create great interest in a border. This is also true where evergreens are used in conjunction with winter-stem colour. *Cornus sibirica*, for example, has striking red stems in winter and is best shown off against a carpet of weed-suppressing variegated ivy. Textural contrasts can also be striking. For example, the bold leaf forms of plants such as *Rheum* contrast well with the more feathery forms of astilbes and *Aruncus* in summer.

THREE-DIMENSIONAL STRUCTURE

When you plan out a border on paper, it can sometimes be difficult to envisage just how it will look when it is established and growing. In order to get a better idea of the final result, it can be helpful to draw a sketch relief of the border. This should always be done as a projection into the future, based upon the lifespan of the plant. This will be the following year in the case of herbaceous plants but may be five years or more in the case of shrub beds.

PLANTING OUT A PERENNIAL BED

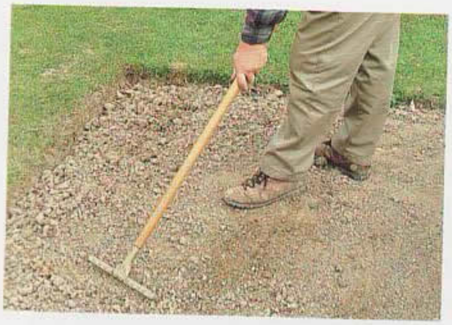

1 *If the bed was dug in the autumn, winter weather should have broken down the soil. In spring, rake over the soil and remove any weeds that have reappeared.*

2 *Although well-rotted organic material should have been added at the time of digging, a sprinkling of bonemeal will ensure the plants get off to a good start.*

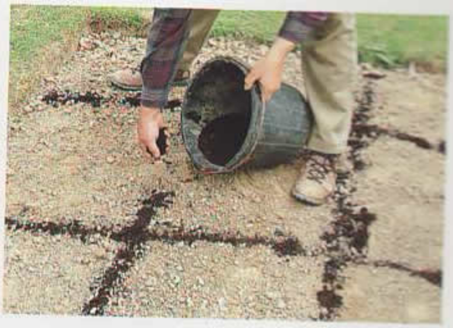

3 *Draw a grid on a plan and then mark out a scaled-up version on the bed, using sand or potting mix. Alternatively, use string between canes to mark out the plan.*

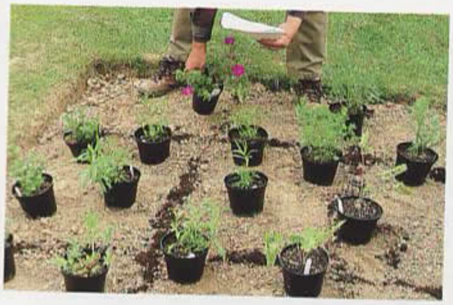

4 *Using your planting plan and grid as a guide, lay out the plants, still in their pots, on the ground. Stand back, try to envisage the border as it will be, and make any necessary adjustments.*

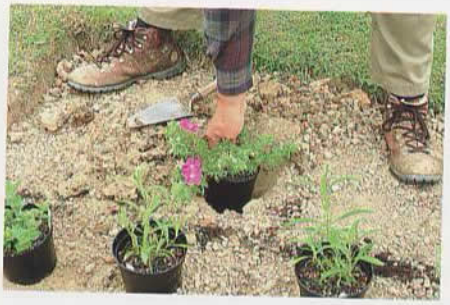

5 *Dig a hole and, with the plant still in its pot, check that the depth and width is right. Adjust if necessary. Remove the plant from the pot and place in the planting hole. Fill in the hole with soil and then firm the plant in.*

6 *When the bed is completely planted, water in all the plants. They should be kept watered until they have become established, especially throughout dry periods.*

7 *Go over the border with a fork, or use a rake if there is room. This will loosen any compacted areas, as well as level the soil.*

8 *Cover the soil between the plants with a layer of mulch like composted bark to keep weeds down and preserve moisture.*

9 *Mark each plant with a label. The finished border should need little attention, apart from removing the odd weed.*

PLANTING GROUPS

When designing and planning your border, always remember to set out and plant the herbaceous plants in groups of three, five, seven or nine. In this way, you can create a more natural and informal effect because this approach does not show the rigid lines or geometric shapes that often appear when you plant even numbers of plants. Larger planting groups than this need not adhere to this number rule because an overall random effect is created by the shape of the drift alone. Larger shrubs, such as *Deutzia* or *Buddleja*, can make stand-alone specimens. Smaller shrubs, such as rosemary or potentillas, may well have greater impact if they are positioned in groups.

CREATING BORDERS WITH HERBACEOUS PLANTS

The popularity of herbaceous plants has meant that they are used in a wide variety of different situations in the garden. However, they are primarily used in the herbaceous border where they can be shown to great effect. Three important categories of this kind of border are usually recognized.

One-sided borders These were extremely popular in Victorian and Edwardian gardens and are the very essence of the English "gardenesque" style. The design usually involves creating a border along one side of a pathway, with the plants arranged in order of ascending height from the path edge. The border is usually positioned in front of a solid backdrop beyond which the eye cannot see.

The backdrop may be living, such as a hedge, or man-made, taking the form of a wall or fence. These beds, while pretty, are often limited by their lack of perspective. The relative proximity of the backdrop often draws attention away from the nearer plants and destroys any illusion of distance. This drawback can be rectified by planting mainly red-flowered plants at one end of the border, gradually changing to blue-flowered plants at the other end. This makes the border seem longer than it actually is. Such schemes are only really effective, however, in borders that are 10m (35ft) or more in length.

Island beds These differ from one-sided borders in that the plants can be viewed from all sides by walking around the whole bed. This kind of planting does not have a backdrop and relies, therefore, on the strategic placement of groups of large plants in the centre of the feature. Island borders allow the onlooker to view the plants from a number of different angles, thus enhancing the interest of the border as a whole.

Mixed borders These designs use many types of plant, including trees, shrubs, herbaceous plants, annuals and bulbs. In a mixed border, one group of plants should complement and enhance the other. For example, the trees provide the shelter and backdrop for the shrubs and flowers in a woodland garden. Rock gardens can also be regarded as mixed borders, with a rich (if diminutive) array of dwarf trees, creeping shrubs and ground-hugging herbaceous plants and bulbs.

ROSE BEDS

A well-tended, healthy and pest-free rose bed has long been, and indeed continues to be, the desire of many gardeners. However, roses can be at risk from pests and diseases. This risk is increased when too many roses are grown in one bed. Most modern cultivars are also too highly bred to resist the range of diseases that can affect them, having been chosen for flower size rather than disease immunity. Choose old-fashioned varieties or modern cultivars that offer resistance to disease.

Most rose pests are common garden pests that are easily controlled by natural predators. The trick with roses is to indulge them with plenty of colourful companion planting, especially daisy-type plants – *Argyranthemum* is ideal – which will encourage those beneficial insects into your garden. It can yield wonderful results and makes the bed even more colourful. Old-fashioned roses need less work in general than modern cultivars and are often more suitable for pollinating insects like bees.

GROUNDCOVER PLANTING

This is a way of exploiting the competitive abilities of plants that grow by spreading sideways and smothering their neighbours. An obvious example of a plant type that is good for this is turf grass, but many herbaceous and shrubby species are also able to fulfil this function. These include conifers, roses, heathers, and both broadleaved evergreens and deciduous

A PARTLY SHADED BORDER

A shady spot in the garden can provide the right growing conditions for a dazzling array of flowers. This mixed border in heavy shade is shown in summer to late autumn.

1 *Aconitum napellus* (poisonous)
2 *Digitalis purpurea* (poisonous)
3 *Anemone sylvestris*
4 *Hyacinthoides non-scripta*
5 *Astilbe* 'Aphrodite'
6 *Primula bulleyana*
7 *Aconitum napellus vulgare* 'Albidum' (poisonous)
8 *Cardiocrinum giganteum*
9 *Lilium regale*
10 *Convallaria majalis*
11 *Primula vialii*
12 *Anemone nemorosa*
13 *Digitalis purpurea* var. *alba* (poisonous)
14 *Campanula alliariifolia*
15 *Aquilegia vulgaris stellata* 'Nora Barlow'
16 *Aster novi-belgii*
17 *Digitalis ferruginea* (poisonous)
18 *Galax urceolata*
19 *Primula japonica* 'Millers' Crimson'
20 *Acer*
21 *Mahonia* 'Charity'
22 *Cotinus coggygria*

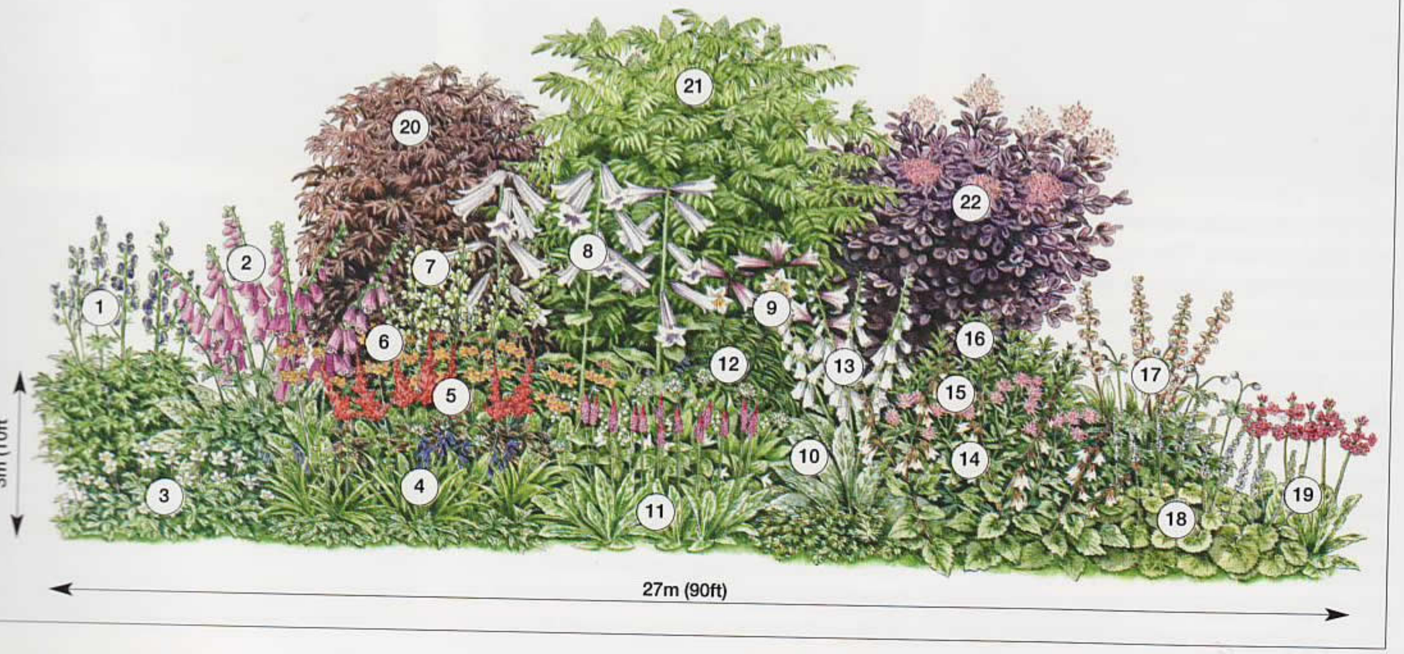

plants. Their main characteristic is a close, compact and low-growing form that suppresses the underlying weeds. The same effect can, of course, be achieved by planting closely. This method is especially useful with herbaceous plants that grow up quickly and suppress the weeds below. No groundcover will suppress all weeds, however, and you may have to weed among them while they become established.

There are numerous groundcover roses available that offer a colourful carpet of flowers. If you opt for roses as a groundcover, you should be aware that some are covered with thorns and it can be difficult to weed between them. Mulching the bed following planting can help to suppress weeds until the plants achieve full cover. All groundcover can take two to three years (or more) to become fully established and so should be considered as a long-term, labour-saving feature.

TIDYING UP

1 *Some edging plants spread out over the grass, possibly killing it or creating bald patches, as has this poached egg plant (*Limnanthes douglasii*).*

2 *If the plant is an annual and has finished flowering, it can be removed. Otherwise, just cut back the part that is encroaching on the grass.*

BORDER MAINTENANCE

It is possible to plan a border to be low maintenance, but there is no such thing as a maintenance-free border. Decide how much time you can spend caring for the border and choose your plants accordingly. A close planting, particularly one that uses ground-covering shrubs or closely planted herbaceous plants, will certainly cut down on the weeding as can mulching the border regularly. Any border will need at least one seasonal tidy up in late winter. Shrub beds are also best mulched at this time before bulbs and other flowers have fully emerged. Choose shrubs that will grow into the space available and avoid very vigorous varieties or ones that need pruning on a regular basis. Even large herbaceous borders need not be too difficult to maintain. Delay your winter clear up for as long as possible and choose species that do not need staking. Do not worry if the plants grow among each other, it will just add to the effect. The less labour a border needs, the less disturbance there will be to the wildlife it contains.

A SPRING BORDER FOR AN OPEN, SUNNY SITE

A sunny spot can provide the ideal conditions for a wide range of showy blooms in the early part of the year. The plants in this border prefer a neutral to slightly alkaline soil.

1 *Lupinus* 'The Page'
2 *Aster novi-belgii* 'Fellowship'
3 *Lobelia* 'Bees Flame'
4 *Heuchera micrantha* 'Palace Purple'
5 *Aquilegia skinneri*
6 *Primula* 'Gold Lace'
7 *Aquilegia chrysantha* 'Yellow Queen'
8 *Lupinus arboreus* 'Mauve Queen'
9 *Iris* 'Blue Eyed Brunette'
10 *Heuchera* 'Leuchtkäfer'
11 *Campanula latifolia* 'Brantwood' (deep blue) and *C. latifolia* 'Gloaming' (light blue)
12 *Iris chrysographes*
13 *Sidalcea* 'Party Girl'
14 *Physostegia virginiana* 'Vivid'
15 *Geranium macrorrhizum* 'Ingwersen's Variety'

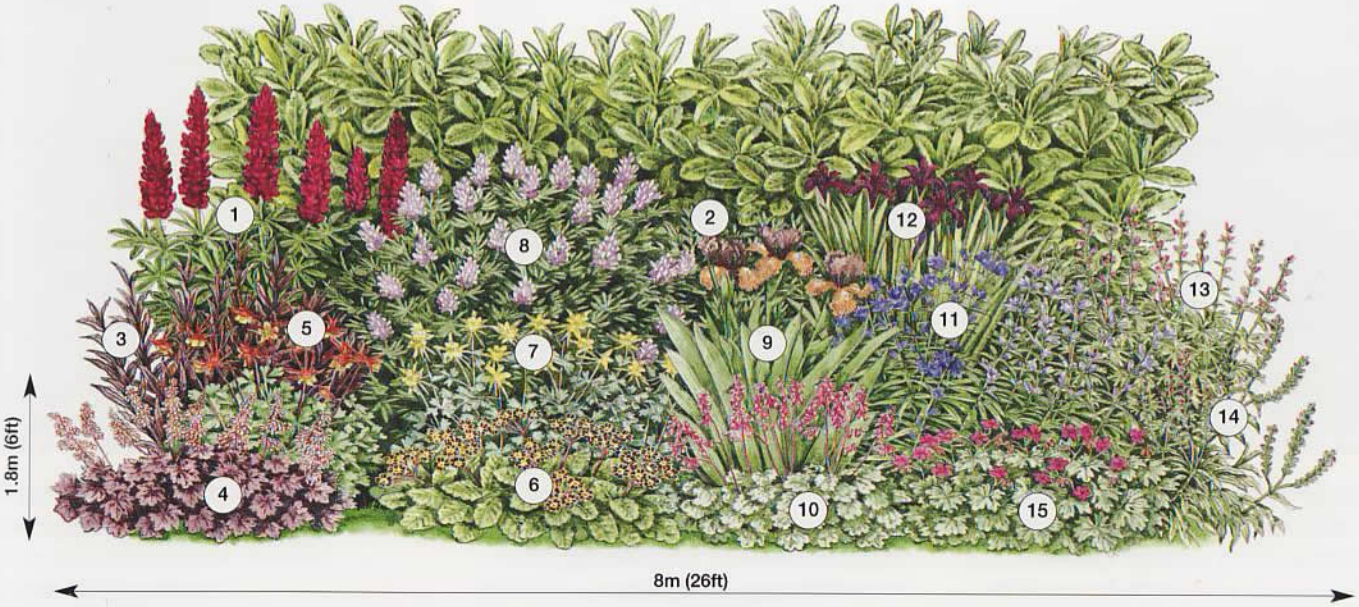

CONTAINERS

These are an attractive feature in any garden and are very useful when there is no soil at all for planting, as is the case with patios and roof gardens. They are also easy to maintain. There are many types of container available, which means that you can create a theme of your own choice: Mediterranean or perhaps lush and tropical in style. You can also plant up seasonal containers to create different effects through the year or you can change the style from year to year.

Above: *A wooden barrel makes an attractive container in which to grow a profusion of plants.*

TYPES OF CONTAINER

Always choose containers that fit in with the general garden style and with each other. Terracotta and wooden planters are easy to fit into almost any garden scheme, but shiny metal containers may be more difficult to place. Hanging baskets usually take the form of a basket of wire, but they can also be made of wicker. This is lined with a permeable material such as wool, hay, coir matting or even conifer cuttings. You can buy purpose-made lining material which is made from recycled materials or you can make your own. Hanging baskets are an excellent way of covering a wall or post and can even increase the growing space in your greenhouse. They are best planted with a selection of trailing plants (placed around the edge of the basket) and upright types (placed in the centre). They are extremely prone to drying out and should be watered at least once a day and even twice in hot dry conditions. Window boxes should be treated in a similar way to hanging baskets.

The size of a container will influence the kind of plant that you can grow. Any container must be at least 15–20cm (6–8in) deep, although most plants will be happier in a deeper one than this. Be wary of putting a small plant in a very large pot, as this can lead to very wet potting mix that the plant will not want to root into. Large pots with arrangements of several plants are ideal and much easier to look after than several smaller pots.

MATERIALS AND FINISHES

Containers are available in a range of different materials and finishes. Glass fibre and resin containers are extremely strong and lightweight but they tend to be expensive because they are hand-made. They are available as extremely realistic stone-look containers, but can also be found in almost any other finish, including wood and glazed ceramic. Plastic containers are very cheap, lightweight and often last a long time. The finish on some may be poor, but they are very easy to paint and customize. Metal containers are available as pots, "manger-style" window boxes or hanging baskets. There are many different designs and the best quality ones have a properly lacquered finish.

PLANTING A CONTAINER

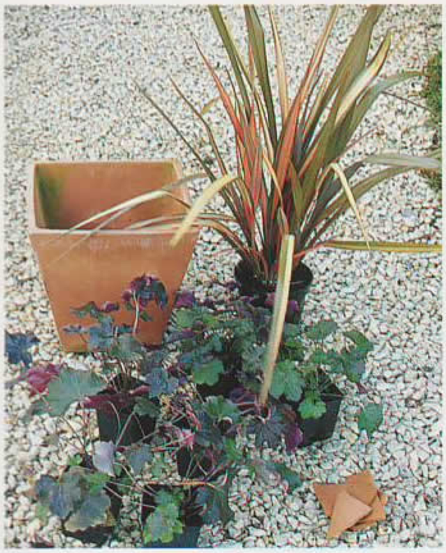

1 *The materials you need include a suitable container, your choice of plants (in this case, the central plant is a striking cordyline), some stones to place at the bottom of the container for drainage and potting mix.*

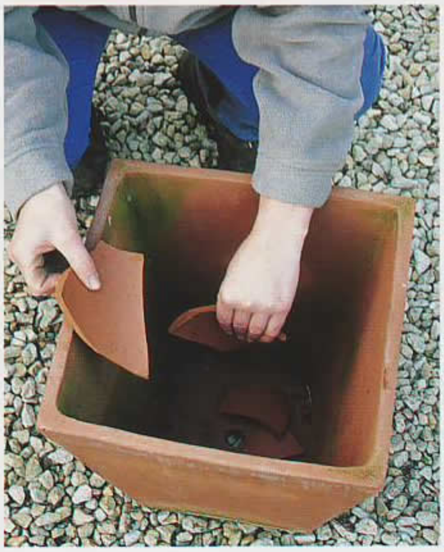

2 *Cover the bottom of the container with small stones or some pieces of tile or pottery, so that water can drain freely from the container.*

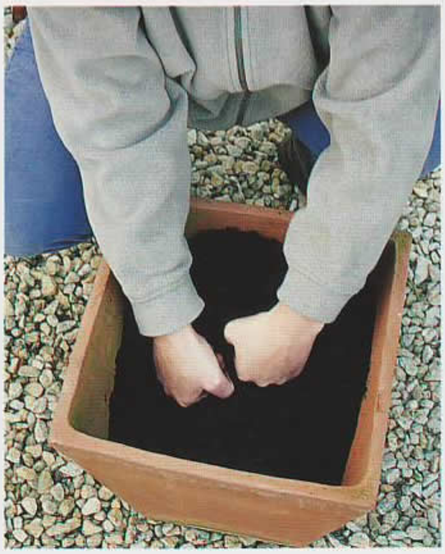

3 *Partly fill the container with a good-quality potting mix.*

Above: *Hanging baskets can help to provide colour and interest at eye-level throughout the year.*

Above: *Using a watering can is one of the easiest ways to keep the soil moist. It allows you to deliver the right amount of water.*

Above: *You can create a stunning display with a range of different containers, including an old wheelbarrow.*

Stone and terracotta pots offset the shades of flowers and foliage beautifully. The main problem with terracotta is that it is water permeable and requires more frequent watering than plastic, resin or metal pots. Both stone and terracotta containers can be very heavy, so plant them up *in situ*. Look out for frost-resistant types that will not crack in winter. Glazed ceramic pots need less watering than plain terracotta pots, and can be chosen in almost any colour to suit the theme of the garden.

Wooden or wicker planters have a very natural look and can easily be home-made. Oak, cedar and chestnut are the best woods to use because they do not need preservative and planed wood – of any type – can easily be painted. Wicker is an excellent material for making hanging baskets, while wood can be used for window boxes. Wicker or wooden planters benefit from being lined with plastic or a non-permeable paint. The paint can be plastic resin or bitumen based, but make sure that it does not contain fungicides or preservative. Alternatively, plastic sheeting can be stapled firmly inside the pot. Wooden and wicker pots should be raised off the ground slightly to prevent the base rotting. There are many different designs, ranging from half-barrels to wheelbarrows. However, take care with wooden items for outdoor use because many of them have been chemically treated. A wide variety of recycled objects can be used as containers. An old boot, bucket or watering can, painted if you wish, can all be used.

DRAINAGE

All containers must have drainage holes in the base. To ensure that the potting mix drains freely, place a layer of gravel – about 4cm (1½in) deep for most pots – over the base of the pot before putting in the mix. Terracotta and stone pots may need as

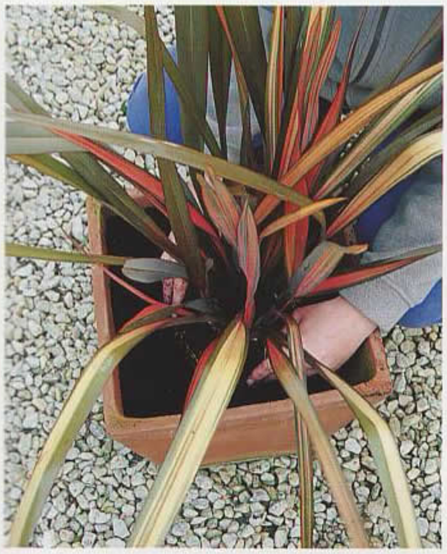

4 *Scoop out a hole in the potting mix and insert the plant, positioning it so that the top of the rootball will be level with the surface of the potting mix.*

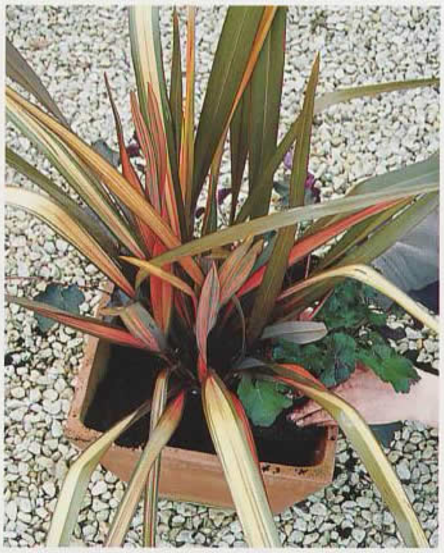

5 *Place any extra plants around the edge of the main plant. Add more potting mix in order to fill in any gaps, and firm down. Water the container thoroughly.*

6 *The plants will soon grow away and fill out the container to create a stunning year-round feature.*

much as one-fifth of their total depth filled with gravel to help protect them from frost damage. Standing pots on bricks or purpose-made "feet" also helps them to drain freely.

ORGANIC POTTING MIX

The best potting mixes contain sufficient air space to allow the roots to breathe, while retaining enough water and nutrients to support plant growth. Potting mix must also be heavy enough to support the weight of the top growth of the plant if it is to remain upright. You can buy a general-purpose potting mix, although some gardeners prefer to mix their own.

Potting mix can be loam-based or soilless. Loam-based potting mix is a "heavy-weighted" mix that provides good anchorage for tall or heavy specimens. It also holds on well to nutrients. It is ideal for plants that will stay in containers for several years or more as it will not degrade or become structureless.

Soilless potting mixes do not contain loam. They have the advantage of being free-draining and relatively lightweight, although some grit or sand is usually added to give them weight. Coir-based potting mixes are made with a natural fibre that is a by-product of the coconut-growing trade. It has been hailed as a peat substitute and is equally good in containers. However, it is expensive and can be difficult to re-wet once dry. Composted fine bark is cheap and readily available. There have been some advances in its use in recent years and it has even been suggested that bark-based potting mix increases a plant's resistance to pest and disease attack.

PLANTING A HANGING BASKET

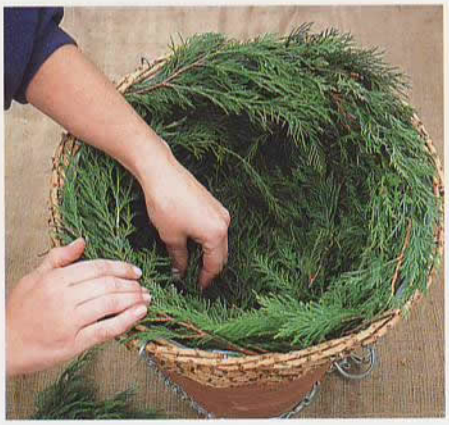

1 *Line the hanging basket with a suitable organic material. Conifer branches are an ideal material, as is shown here.*

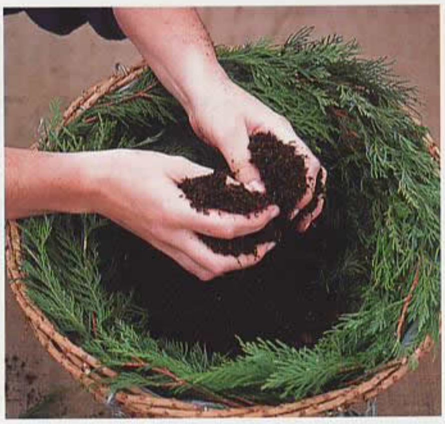

2 *Fill the lined basket with a suitable soilless organic potting mix to about one-third of its depth.*

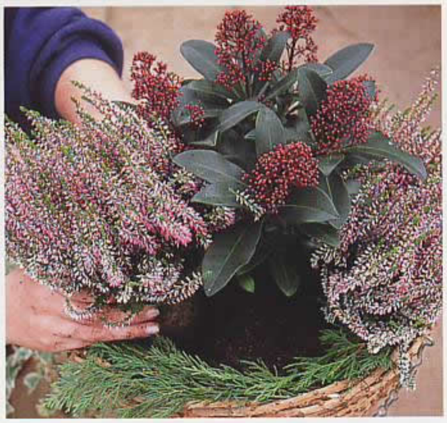

3 *Position the plants first to see how they will look. Fill in the spaces between the plants with more potting mix.*

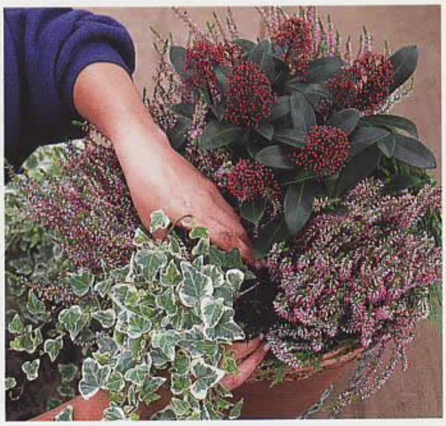

4 *Any gaps between the plants can be filled with small bulbs or bedding in order to provide a riot of spring colour.*

MAKING ORGANIC POTTING MIX

You can make your own potting mix with ingredients such as well-rotted compost, leaf mould and worm compost; loamy garden soil; and clean, coarse sand or fine grit. The organic matter and loam should be thoroughly sieved. A slow-acting, balanced fertilizer may also be added along with ground limestone to bring the mixture's pH to around 6.5 which is suitable for the majority of plants. Containers for seasonal displays such as bedding will not need loam.

PLANTS FOR CONTAINERS

Fruit, herbs and even vegetables can all be grown in containers, as well as a wide range of ornamental plants. A flowering shrub such as an azalea can act as a focal point and enliven a shadier part of the garden. Pots of fresh herbs and salad crops can provide a welcome addition to summer meals and a window box can even be used as a miniature wildlife garden. Seasonal bedding can transform a patio into a riot of colour and be changed regularly to reflect the seasons. Containers also enable you to grow a range of plants that would not otherwise grow in your soil. For example, if your garden soil is alkaline, you can still enjoy the spring and summer delights of acid-loving rhododendrons, blue poppies (*Meconopsis*) and Himalayan primulas.

CARE AND MAINTENANCE

There are a number of tasks that you will need to perform in order to keep your container plants in peak condition.

Watering Knowing when and how much water a plant needs is one of the most important aspects of container gardening. Often a look at the surface of the potting mix may give the impression that it is dry. If you dig down with your finger, however, you quickly discover moist potting mix. If the mix feels wet, then do not water it.

Feeding Container plants need more fertilizer than plants grown in the open garden because frequent watering washes out the nutrients. For the best results, start feeding container plants six to eight weeks after planting. Use a liquid tea, worm liquid or a purchased soluble food such as seaweed extract or liquid manure every two to three weeks. Apply fish emulsion or garden compost once or twice a season to add trace elements to the mix. Do not add more than the recommended rate of any fertilizer as too much can harm plant roots.

PLANTING A WINDOW BOX

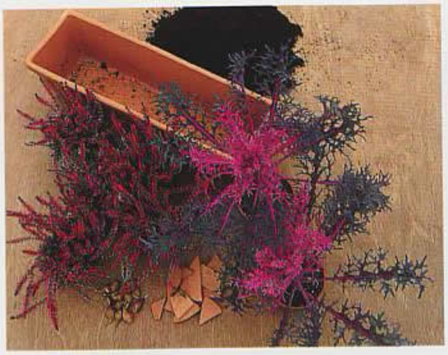

1 *Assemble all the necessary materials. These include the window box, crocks, a good organic potting mix and the plants. If the box is light, assemble it on the ground. If not, then assemble it in position.*

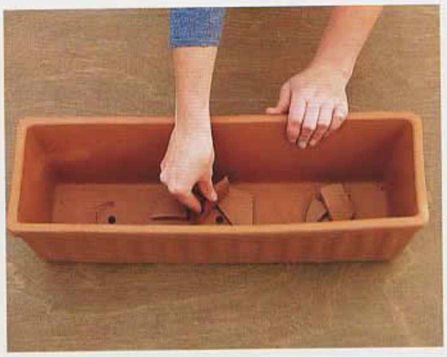

2 *Holes in the bottom are essential to allow good drainage. Stop the potting mix from being washed out by placing crocks over these. If very good drainage is needed, then a layer of gravel can also be added.*

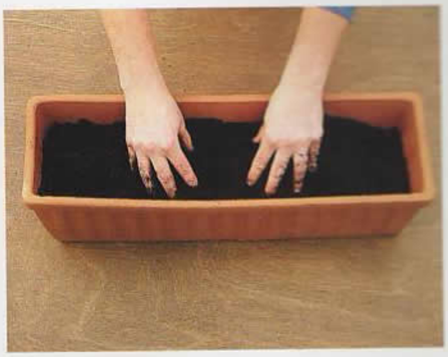

3 *Partially fill the box with potting mix, gently tapping the sides to make sure that no air gaps remain. Never over-firm soiless potting mixes as they will become waterlogged and airless, and the plants will suffer.*

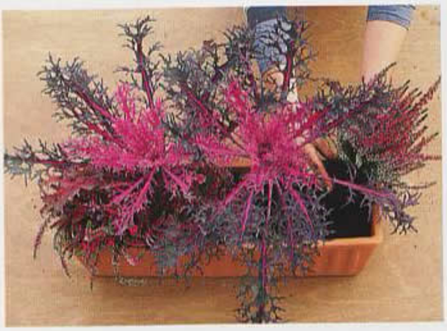

4 *Place the plants in position and check these positions before finally planting. Make sure that they are planted in the potting mix at the same depth that they were in their pots or trays. Plant them fairly close for an instant effect.*

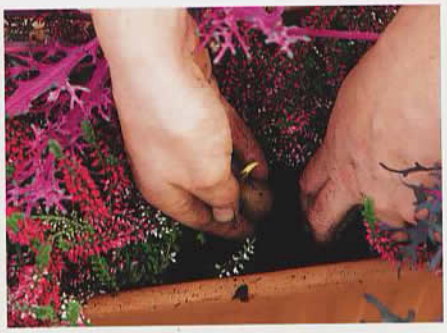

5 *A selection of bulbs makes a useful additional display for the window box. The bulbs can be planted in among the main plants and are best planted in groups of three so that they give a fuller display. Water the basket once it is planted.*

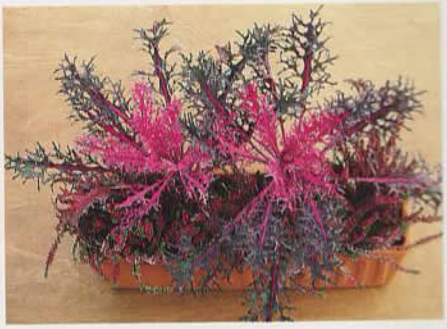

6 *Window boxes that are planted for seasonal display will quickly use up all the available nutrients. This is less of a problem with baskets planted for winter interest, such as this, but summer boxes will need regular feeding.*

Monitoring plant health Containers are best placed where they will receive maximum sunlight and good ventilation. During periods of high temperatures and bright sunshine, many containers may benefit from shade during the hottest part of the day. Shelter plants from severe rain, hail and wind storms. Also watch out for and control insect pests.

Pinch pruning When young, the growth of many ornamentals is readily shaped by the pinching out of young shoots. Many specimens treated in this way yield impressive results. This pinching out, or stopping as it is sometimes called, stimulates the development of side shoots and, if carried out regularly, results in dense bushy growth. The selective use of this method of pruning allows for considerable freedom in the shaping of plants.

MAKING ORGANIC POTTING MIX

Mixing your own organic potting mix is relatively easy. You can use a combination of a variety of materials, including loam, leaf mould, garden compost and worm compost.

INGREDIENTS	RATIO BY VOLUME	ADDITIONAL INFORMATION
Loam : Leaf mould : Garden compost	1 : 1 : 1	A basic mix that is well drained, fertile and suited for longer-term plantings.
Leaf mould : Worm compost	3 : 1	Very nutrient rich. Add extra grit for better drainage. Best for short-term pots of vegetables, herbs and flowers.
Loam : Manure : Leaf mould	3 : 1 : 1	A rich mix for heavy feeding plants such as pot-grown tomatoes and peppers.
Leaf mould : Loam	1 : 1	Ideal for long-term plantings such as trees and shrubs. Use acidic loam for lime-haters.
Loam : Leaf mould or Coir	1 : 1	Enrich with 225g (8oz) seaweed meal; 110g (4oz) bonemeal; 85g (3oz) hoof and horn; 55g (2oz) ground limestone per 30 litres (63 pints) base potting mix.

THE WILDLIFE GARDEN

Wildlife gardening involves providing food, shelter and a habitat for creatures as well as relying on native plants that are suited to the climate and soils in the locality. A single plant species may directly and indirectly serve species of insects, invertebrates, fungi, and other organisms, which in turn become a meal for larger wildlife such as birds. In time you will have a functioning ecosystem on your doorstep. In fact, creating a garden of this kind can help break up the "grass desert" responsible for reducing the variety of birds, insects and other wildlife.

Left: *Wildlife gardens use plants, such as this fennel (*Foeniculum vulgare*), for their ability to attract insects.*

Above: *Garden wildlife can be both pretty and, in the case of ladybirds (ladybugs), useful for controlling aphids.*

Above: *Flowers such as sweet peas (*Lathyrus odoratus*) encourage pollinators and beneficial insects into your garden.*

Above: *Colourful fruits such as these rosehips provide much-needed food for garden wildlife such as birds.*

WHAT IS A WILDLIFE GARDEN?

The organic garden can provide a welcome refuge for a wide range of wildlife, from birds and butterflies to creatures such as frogs and hedgehogs. The basic requirements for any visitor to the garden are food, water and shelter. You may also want to consider organizing some, or all, of your planting schemes around species that will attract a diversity of fascinating wildlife.

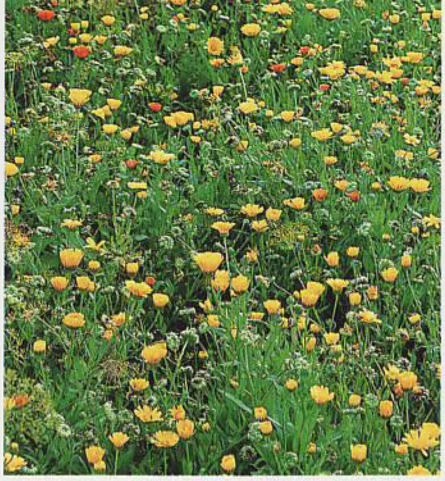

Above: *Wildlife gardens such as this grassland area can be attractive as well as providing useful habitats.*

WHAT ARE THE BENEFITS OF A WILDLIFE GARDEN?

Domestic gardens cover a huge area around our towns, cities and urban fringes. This is a potentially fantastic resource for an otherwise denuded habitat. Keeping your garden free of harmful chemicals gives you a head start in attracting wildlife and, with a little planning, you can create a habitat that will welcome a diverse range of creatures. The visitors in turn will repay you by acting as predators, helping to control the level of problem-causing pests in your garden. Knowing that your patch, no matter what its size, is providing a haven for wildlife will give you a closer connection with the natural world. Just to be able to enjoy watching birds feeding can lift the spirits and improve your sense of wellbeing.

Encouraging natural cycles in your garden will promote biodiversity. The more soil life there is, the more insects will come to feed on it. Birds will come to feed on the insects, as will amphibians and mammals. In this way, food webs will gradually recover in the absence of pesticides.

Making your own garden compost, instead of using peat, for instance, prevents damage to a fragile habitat that cannot be re-created. Collecting rainwater and using this in the garden will reduce the consumption of mains water. Huge amounts of energy are wasted on cleaning and transporting this precious resource and it is often extracted from rivers at levels that threaten the local wildlife. The careful gardener can encourage the creation of wildlife habitats in their own backyard and prevent habitat destruction somewhere else.

CREATING THE RIGHT HABITAT

Much of the wildlife that thrives in a garden does so under the care of a gardener who is not unduly tidy. The best wildlife gardens

A WILDLIFE GARDEN

A well-designed wildlife garden will provide food and shelter for a host of different creatures.

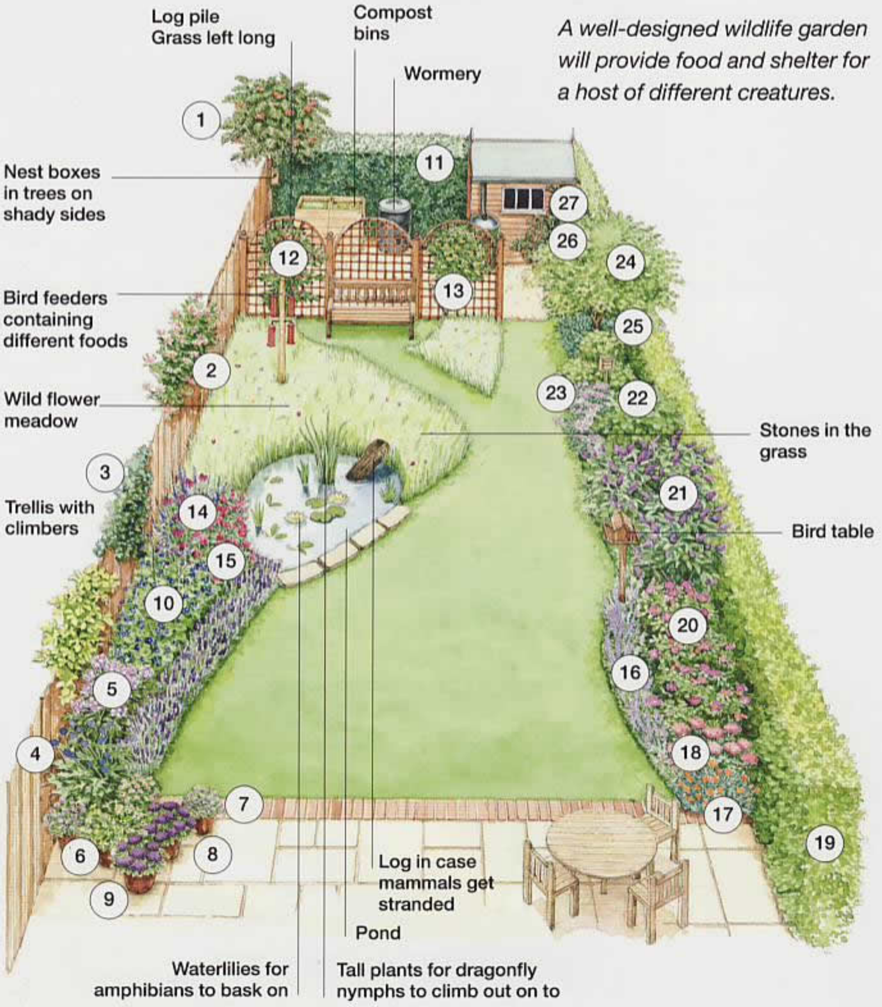
Log pile
Grass left long
Compost bins
Wormery
Nest boxes in trees on shady sides
Bird feeders containing different foods
Wild flower meadow
Trellis with climbers
Stones in the grass
Bird table
Log in case mammals get stranded
Pond
Waterlilies for amphibians to bask on
Tall plants for dragonfly nymphs to climb out on to
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27

1 *Sorbus aucuparia*
2 *Lonicera periclymenum* 'Belgica'
3 *Hedera helix* and *Humulus lupulus* 'Aureus'
4 *Echinops ritro*
5 *Lavandula* hedge
6 *Phlox paniculata*
7 *Nicotiana alata*
8 *Phacelia campanularia*
9 *Heliotropium* in pots
10 *Verbena bonariensis* and *Origanum vulgare*
11 Hedge of *Ilex aquifolium*
12 *Pyracantha*
13 *Clematis tangutica*
14 Mixed planting of *Perovskia* 'Blue Spire' and *Echinacea purpurea*
15 *Centranthus ruber*
16 *Nepeta* 'Six Hills Giant'
17 *Eschscholzia californica*
18 *Sedum spectabile*
19 *Fagus sylvatica* hedge
20 *Buddleja* 'Lochinch'
21 *Syringa* × *prestoniae*
22 *Aster* × *frikartii*
23 *Ribes sanguineum*
24 *Crataegus laevigata* 'Paul's Scarlet'
25 *Mahonia japonica*
26 *Cotoneaster horizontalis*
27 *Jasminum officinale*

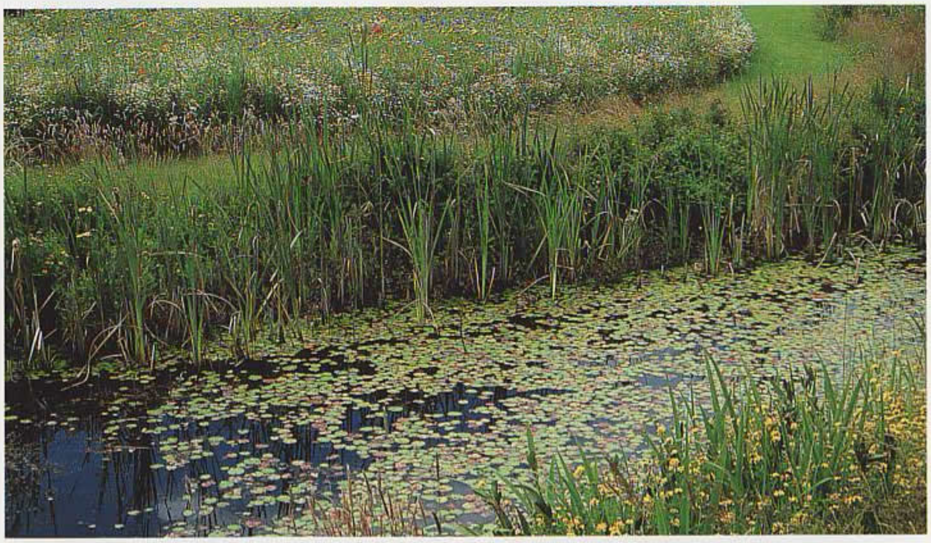

Above: *Wildlife thrives in natural wetlands, many of which are threatened by domestic water use.*

Above: *Looking at natural habitats such as woodlands can provide inspiration for your own garden.*

leave room for decay. This approach – and it is one that is good for organic gardens in general – involves piling up old logs and autumn leaves in a quiet shady corner to create a home for insects and hibernating hedgehogs. Do not confuse this with making leaf mould pits; these stay in place and gradually rot down. Log piles can also be overplanted with ivy to enhance both their appearance and attractiveness to insects such as stag beetles.

Consider planting native trees, shrubs and flowers so that native creatures will have a familiar food source or nesting site. Plant as many suitable flowers as possible to attract bees, butterflies and other insects that will enhance your wildlife garden. Many beneficial species are attracted to a garden that is effectively a feeding station for them. For the larger visitors, such as small mammals and birds, you can plant flowering and berrying trees and shrubs.

Ponds provide a rich habitat for a variety of wildlife. Many familiar creatures, such as frogs, toads, newts and dragonflies, depend upon ponds to breed. Many urban gardens have become havens for these creatures. Even a small pool will provide somewhere that birds and other visitors can come to drink. If you have no space for a pond, try introducing a suitable container that can be utilized as a drinking or bathing area for birds, mammals and even insects.

Just leaving the lawn uncut for a few weeks in mid-summer will give many of the plants you do not normally see the opportunity to flower. An area of longer grass also benefits butterflies, moths and many other insect species by providing food and shelter. Many of these insects will fall prey to larger garden occupants such as songbirds and hedgehogs, so diversifying and enhancing the food web in the garden.

The barriers and structures in your garden also provide very useful habitats. Hedges are a sanctuary for all sorts of wildlife, where they can breed, feed and take shelter. These hedges are easy to maintain, as they should only be cut once in the winter to avoid interfering with spring and summer nesting birds. Cutting is done on alternate sides in alternate winters. If one side is cut this winter, then the opposite side is cut the next. This ensures that only one side is clipped before the growing season. The unclipped side flowers and provides nectar and fruit for wildlife. Dry-stone walls are another garden feature that can provide homes for hibernating amphibians, lizards, insects and bees and are best left undisturbed throughout the winter.

Above: *Recycling organic material through composting reduces the need to use substances such as peat.*

If you are lucky enough to have a few trees in your garden, you could try to re-create a woodland habitat. Larger gardens may actually contain a small area of woodland but most are more likely to have a shady patch under a few trees. Choose plants that mimic a woodland floor: bluebells, anemones and sorrel are just a few of the handsome plants that will thrive in these conditions. Planting a few understorey shrubs (those that grow below the trees) such as holly and hazel will complete the effect and may well tempt the shyer woodland birds into your garden.

WILDLIFE CORRIDORS

A wildlife garden is often likened to an oasis in an urban desert. Gardens are the only chance of cover that many creatures have to migrate between larger blocks of suitable habitat such as parks or a greenbelt. If these important habitats become totally isolated then they begin to lose much of their diversity. Town gardens play a vital role in protecting the health of the wider urban habitat, while rural gardens play a similar role in areas denuded by intensive farming. Try putting up a few bird or bat nest boxes and providing nesting areas for insects and mammals. The greater the year-round diversity, the more useful your garden will be.

WOODLAND GARDENS

Woodland can be one of the most interesting and diverse natural habitats. Those with smaller gardens may think that such a feature would be too large for them to create. However, even a few trees in the corner of the garden can provide ample opportunity for the creation of a small wooded area and provide a rich and diverse habitat for a range of visiting wildlife. Try to select native trees or species that have nectar-rich flowers or berries.

Above: *Woodlands are home to many of our favourite garden flowers such as these primroses (*Primula vulgaris*).*

WHAT IS WOODLAND?

Woodland is a very variable habitat and its exact character will mostly depend upon the prevailing climate. It is often a more open environment than that of a forest, having clearings, glades, pools and streams that break up the monotony of tree cover.

Large trees such as oak (*Quercus*), beech (*Fagus*) or plane (*Platanus*) are known as the woodland canopy. Beneath the canopy are smaller trees such as holly (*Ilex*) and hazel (*Corylus*), known as understorey species. They are often shade-tolerant trees, but grow best under gaps in the canopy. Beneath this, a shrubby layer develops in more open woodland, possibly consisting of bushes such as hawthorn (*Crataegus*) and elder (*Sambucus*).

The floor of the woodland is called the herb layer. The plants that live here tend to be seasonal, taking advantage of early spring sunshine to grow and flower, although some, such as mosses and ferns, grow through the summer. It is also where all seedlings grow and is mulched each year with the falling leaves of the trees.

CREATING A WOODED AREA

Gardeners who have a few trees (even if they do not provide a continuous canopy) can easily make a small woodland area. Visit some woods in your area to gain some inspiration and emulate natural growth if possible. Think how the existing tree cover can be linked, maybe by planting more canopy species and some understorey trees.

Woodland environments are often havens for plants and animals, with many containing an enormous diversity of species. This is most pronounced where there is a "mosaic" of open ground and trees. This is because it increases the "marginal areas" – the point between woodland and more open habitat. The places where two habitats meet are always the most diverse. Layered woodland (one with canopy, understorey, scrub and ground herbs) is beneficial for wildlife as it provides ample feeding and shelter for a range of creatures. Any additional planting should always be aimed at achieving or maintaining this layered structure and providing the highest number of habitats.

You can think of woodland as a collection of diverse habitats. Even those that have only a few types of tree are capable of supporting a great number of species that live on or within these trees. Oak trees in Britain, for instance, are known to provide homes for over 400 insect and other invertebrate species. This does not mean, however, that simply planting an oak tree in your organic garden will guarantee all 400 will appear at once. If you aim to create a woodland that is similar to others in your area, then species will be able to migrate from these to the one in your garden.

You may need to provide additional help for wildlife if it is to establish successfully in the woodland area. Nesting boxes, log piles, rocks and a long grass area linked to your pond can all help to encourage wildlife to use your planting as a habitat. Make sure that you include plants that provide food and be prepared to give additional sustenance, especially in the winter for non-hibernating foragers such as blue tits and other woodland birds. As the woodland develops, so will the diversity it contains. Let nature do its work and give a helping hand from time to time.

PLANTING WOODLAND FLOWERS

In most deciduous woodland, the ground is a riot of colour from late winter until early summer as showy woodland species grow, flower and set seed before the leaves in the canopy cast too much shade. As a consequence, most "woodland gardens" are at their best from late winter to early summer. The edges and clearings provide the chance for later colour, of course, but

Left: *Shady and partly shaded wooded areas can provide the perfect habitat for a host of creatures.*

Above: *Foxgloves (*Digitalis*) thrive in shady woodland corners and are very easy to grow from seed.*

Above: *Bluebells (*Hyacinthoides*) naturally grow and flower in the spring before the trees shade them.*

some of the showiest species are to be found among the spring flowers. Choose species that grow in your own area, where possible, because these will be of most benefit to wildlife.

Primroses (*Primula vulgaris*), bluebells (*Hyacinthoides*), snowdrops (*Galanthus*) and wake robin (*Trillium grandiflorum*) are all guaranteed to give a good display, especially at a time when there are fewer colours in the rest of the ornamental garden. Later flowers such as foxgloves (*Digitalis*) should be planted at or near the edge, where they will still get sunlight once the leaves have appeared on the trees. Climbers such as honeysuckle (*Lonicera*) or clematis are also best situated at the woodland edge where you can appreciate the plants as they bloom. It is advisable to always obtain your seed or plants from a reputable source that does the propagation themselves. Never buy bulbs or plants that have been collected from the wild. Your aim is to create a habitat, but not by destroying another.

ASSOCIATED PLANTING

Perhaps the easiest way to mimic a woodland habitat in a smaller garden is to link a few trees with shrubs and understorey trees in order to provide the shelter and food plants that wildlife needs. These shrubs – particularly if they are deciduous – can provide a larger space in which to plant "woodland herb layer" species. Doing this will also widen the edge of the habitat and provide extra coverings around the base of trees. There is, of course, no problem in having a few decorative species in or around your woodland. Try a few rhododendrons, mahonias, lilacs and hydrangeas to provide additional colour and interest if you have enough space. They will all provide shelter and look attractive when they are in flower.

EDGE ENHANCEMENT

Woodland edges are often the most valuable area in many types of woodland for a variety of wildlife and plants. The trick with any wildlife-friendly area is to make sure that these boundaries are as long as possible. Fashioning them into "scalloped" or bay shapes will usually achieve this and each scallop or bay may contain different plants or shrub species. The clearings inside your woodland (if it is large enough) are also important. When you create these clearings, ensure that they are sufficiently long to allow plenty of sunshine into each one. Woodland glades should not be less than 7m (23ft) across and are best if they are irregular in shape.

Remember that the direction of the sun in your garden will affect the type of plants that you can grow at the woodland edge, and butterflies, for example, will prefer a sunny aspect. If the bays or clearings in your woodland area are mainly intended for long grass, then this may only need cutting every two to five years unless, of course, you wish to have a secluded seating area that may need more frequent mowing to maintain access.

Above: Hellebores and Anemone blanda *make an ideal underplanting for trees, providing colour in early spring.*

Left: *Hornbeam* (Carpinus)*, grown as a decorative hedge, can be underplanted with a selection of woodland flowers for added interest in the spring.*

WILD FLOWER LAWNS

Compared with a closely mown pristine lawn, an area of rough grass with flowers growing in it can provide a rich haven for a diverse range of wildlife. In nature, many attractive and species-rich habitats develop over time but in the garden, things may need a helping hand. You do not have to let the whole of the turfed area go untouched. Simply leave an area of grass uncut for some of the year and plant it up with your own choice of wild flowers to create a rich habitat.

Above: *Lawns and areas of long grass can easily be transformed into a stunning display of summer flowers.*

PLANTING AND MAINTAINING A WILD FLOWER MEADOW

Small areas of flowery grassland seem to work best when co-ordinated with hedges, shrubs or trees. They are not necessarily spectacular from a distance and need close inspection to appreciate the detailed tapestry of colour and form. Think carefully about where you would like to site such a feature. Always choose a simple mix of species that are well adapted to the conditions in your garden. Take a walk around your home area to see which local or well-adapted species thrive in similar soil types and aspects to your own.

ANNUAL FLOWERS OR PERENNIALS?

Many annuals have attractive flowers and seem ideal for flower-rich lawns but most require yearly soil disturbance and open ground. Poppies and cornflowers are good examples of showy annual plants that benefit from such conditions. Annuals are usually better treated as a separate display. Try planting some wheat or other grain along with the annuals and leave this for foraging birds and mammals. Don't try and harvest the wheat, just cut it down in late autumn and rotavate the plot ready for the spring.

Wild flower meadow species are perennials, which are adapted to life in permanent grassland. The easiest way to make a wild flower lawn is to plant small wild flower plants, directly into an existing lawn. Many wild flowers are easily raised from seed that is available from specialist suppliers. Sowing and growing this is generally the same as for any other herbaceous plant but you should use potting mix that is low in nutrients for the best results. A late autumn or early spring sowing will provide plants that are large enough to plant out in early to mid-summer.

HOW TO SOW WILD FLOWERS

1 *The most satisfactory way to create a wild flower meadow is to sow a special mixture of wild flower seeds. Remember to completely clear the ground of all perennial weeds before you start.*

2 *To bury the seeds, simply rake in one direction and then in the other. It does not matter if some seeds remain on the surface. Keep the area well watered until the seeds germinate. Protect from birds if necessary.*

ADDING INTEREST TO NATIVE GRASSLAND

1 *For a very small area, wild flower plants may be more convenient. You can raise your own from seed or buy them. Plant into bare ground or in an existing lawn.*

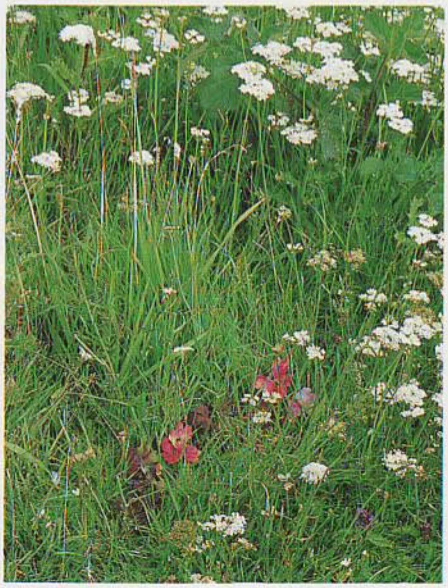

2 *Keep the plants well watered until they are established. Several years of successional planting may be needed to fully establish your wild flower lawn.*

Plant these out about 10–20cm (4–8in) apart in groups of 9–15 to ensure that the group can compete with the grass. Mark where you plant them so that you can see which ones have succeeded later.

If your plants do not establish the first time around, then it is worth trying again. Two failures should tell you that they are not suited for your conditions. You can always try a few non-native plants that have proven wildlife value. Prairie plants such as *Echinacea* are very useful in long grass. Try to avoid rare species, as they are likely to be particular in their habitat requirements, and always consider the attractiveness of plants to insects, birds or other wildlife when deciding what to grow.

NATURALIZING BULBS IN GRASS

Bulbs are potentially one of the easiest plants for flowery lawns. Many early-flowering species are suitable, including crocus, iris and snowdrops. Others, such as *Narcissus*, *Puschkinia*, fritillaries and *Leucojum*, provide good displays later in the spring. Planting bulbs is quite simple. Planting holes can be dug singly with a trowel or spade to the required depth – usually two to three times the height of the bulb. Bulbs should be placed with the base plate firmly in the bottom of the hole. An alternative method for large groupings of small bulbs like crocus and snowdrops is to cut an "H" shape in the grass surface and peel the grass back in two "flaps". Place the bulbs underneath with the top uppermost and gently replace the flaps, firming them carefully. Any group of bulbs should be denser at the centre, gradually becoming wider spaced toward the outside, as this will give a natural appearance. Bulbs such as *Narcissus* may survive for many years but some smaller types – crocus being a prime example – may be eaten by rodents.

HOW TO NATURALIZE BULBS IN A LAWN

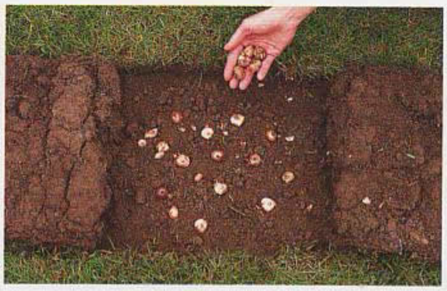

1 Lift an area of turf to create an H-shaped cut. Slice beneath the grass until you can fold back the turf for planting. Loosen the ground, as it will be very compacted. If you want to apply a slow-acting fertilizer, such as bonemeal, work it into the soil at the same time. Scatter the bulbs and plant them where they fall for a natural effect.

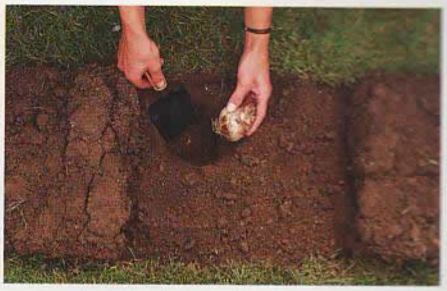

2 Use a trowel or a bulb planter for large bulbs, making sure that the bulb will be covered with twice its own depth of soil when the grass is returned. Firm down the soil and then return the grass. Firm once again if necessary to make sure that the grass is flat, and water well. Water the grass again in dry weather.

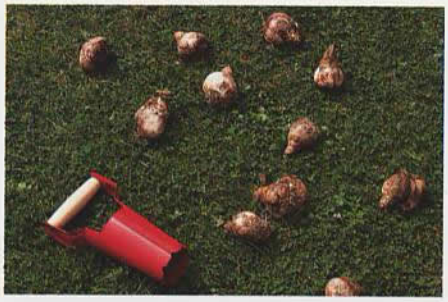

3 Special bulb planters can be used for large bulbs. The planters remove a cylindrical core of soil.

4 Place the bulb at the bottom of the hole on a bed of gravel to improve drainage, if necessary, and replace the plug of earth.

Above: *Daffodils (*Narcissus*) are an ideal bulb to naturalize in lawns, giving spectacular early-spring interest in the wildlife or ornamental garden.*

CARE AND MAINTENANCE

The time of year that you mow your meadow will influence the species you can grow and also when it will look good. If bulbs are not being grown in the grass then a spring cut can help keep a summer wild flower area tidy, especially after a mild winter. A high cut is best, especially if wild flowers have started to produce leafy growth. Cutting in mid-spring will produce a flowering peak in mid-summer. A slightly later cut in late spring will delay the peak of flowering by a few weeks.

Once flowering has finished, the long grass should be cut down. Spring-flowering meadows are best not cut until after flowering. If bulbs are included then the area should not be cut for six weeks after they have flowered. A second cut in mid-autumn will tidy the area ready for the winter. Summer meadows are usually cut from late summer to mid-autumn. All cuts must stress the grasses but minimize damage to wild flowers. Cuts should be between 30–70mm (1¼–2¾in) in order to achieve this.

Cuttings should be removed to help reduce fertility (as this will help your wild flowers) and to prevent long cuttings smothering the plants below. The cuttings should be left on the ground for three to five days, though, to allow insects to move back into the cut grass below. A portion of lawn may be left completely uncut to protect over-wintering insects, but the portion left must be rotated each year to make sure that the wild flowers survive.

WILDLIFE PONDS

Water is a key element in a wildlife garden. Garden ponds attract an enormous range of living creatures, such as dragonflies, newts and birds, to name just a few of the insects and animals that will visit the garden. Some will come to drink or bathe, others will set up permanent residence there, adding to the beauty and diversity of your garden. A pond not only provides an attractive feature in your garden, but also has far-reaching benefits to wildlife.

Above: *Pond plants are essential to the health and balance of the habitat as well as being highly attractive.*

CREATING A HEALTHY ENVIRONMENT FOR A POND

There are a number of points that you must be aware of if you want to provide a safe and healthy environment for wildlife. A pond is always best sited in a sunny position. This is because many pond plants will not prosper in shade, while the water may also be rather cold until the early summer. Shade within the pool is best achieved by surface vegetation, although a pond may have up to a third of the surface in shade if the remainder is well lit. You should also avoid a site near to overhanging trees whose shade can cause the water to be colder for longer in the spring. Falling leaves can cause a build-up of nutrients and increase the potential for toxic tannins in the water. The roots of some species may also interfere with the lining of the pond and can, in extreme cases, cause serious damage.

WATER FOR THE POND

Never fill your pond with tapwater, which often contains many additives, most notably chlorine. Much tapwater can have a high pH and may contain phosphates or nitrates. These can raise the pond's nutrient levels and promote algal growth. Water runoff from nearby lawns or borders can have much the same effect. If the water level becomes very low during the summer it is possible to use solarized mains water to top up your pond. This involves filling large containers with tap water, partially covering the tops and leaving them to stand in the sun for a couple of days. It removes chemicals like chlorine and, although not as good as rainwater, it will stop your pond from drying out. The majority of pool life requires a still environment, so avoid introducing pumps, fountains or filters.

A CROSS-SECTION OF A WILDLIFE POND

Displayed below are some of the more common inhabitants of the garden pool which go towards forming a healthy and balanced ecosystem.

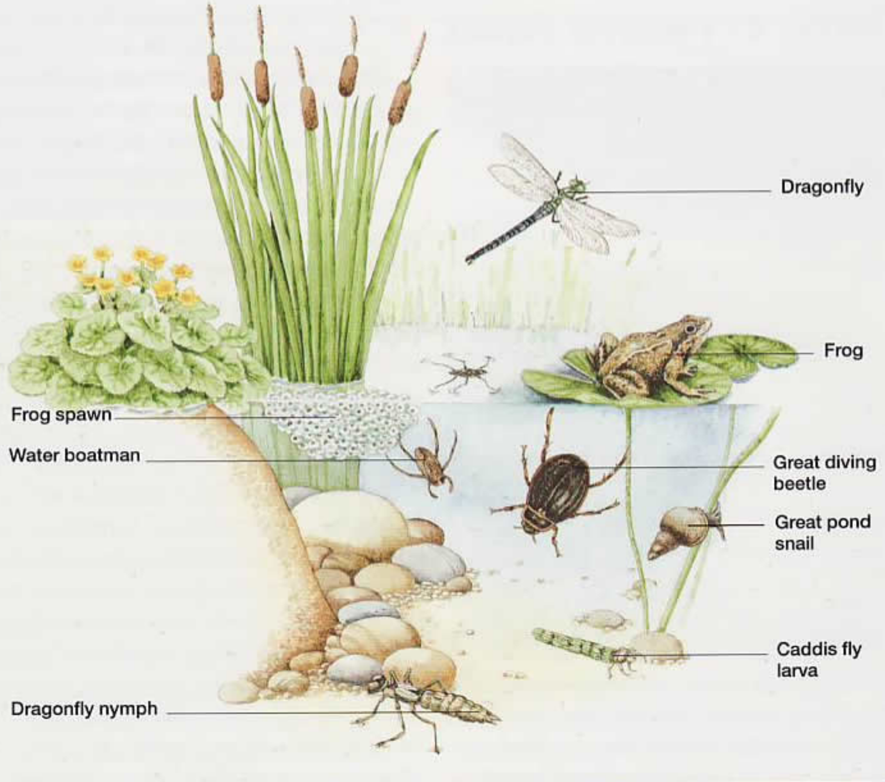

POND SIZE

Even small-scale ponds can support an amazing diversity of life. The shallows are the most important areas for wildlife but a depth of at least 60cm (24in) must be reached in some parts to give creatures a place to which they can escape in extreme heat, cold or drought. The greater the surface area of a pool, the deeper it has to be (at its deepest point) to maintain an even temperature throughout the year. The average pool will need to be about 1m (3ft) deep.

STOCKING A POND

Ponds that are designed to attract wildlife should be colonized naturally within a couple of years. This is the best method for stocking your new pond in the first instance as it is providing a breeding place or hibernating area for those creatures already established in your area. Introducing frog (or other amphibian) spawn can sometimes spread diseases such as red leg that can devastate populations of frogs and other amphibians. Never attempt to introduce adult amphibians to your pond. They have an amazingly strong homing instinct and can die in the process of finding their way back.

A WILDLIFE POND

Wildlife ponds depend upon correct planning. Ponds must be both deep enough and possess the right balance of plant types.

These plants will provide habitats, shelter, and maintain a balanced environment for the wildlife in the pond.

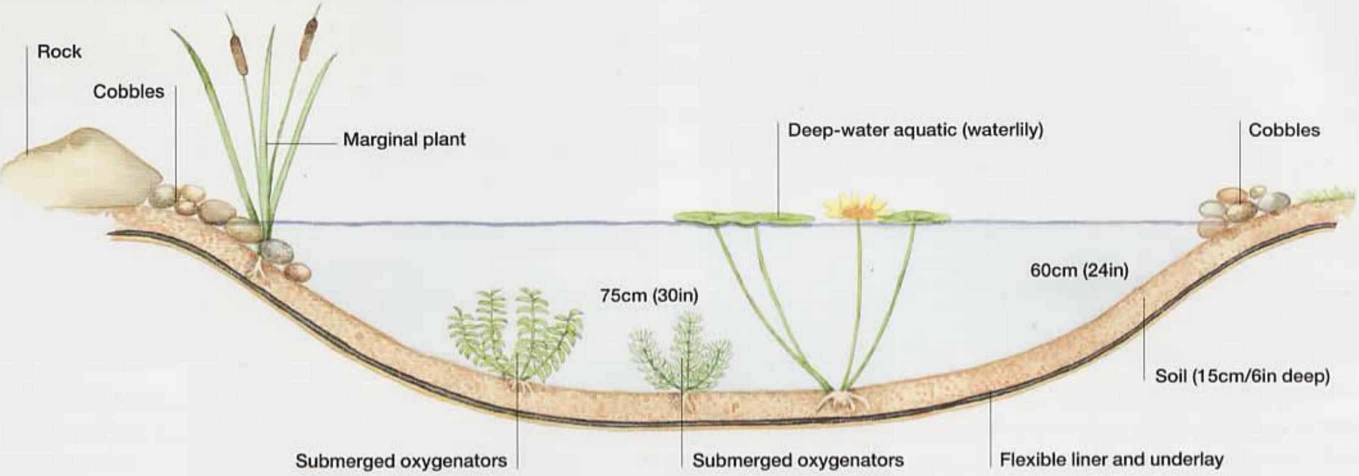

Avoid introducing fish to small wildlife ponds, as they can exclude other wildlife. Sticklebacks can wipe out the tadpoles of frogs and newts. Fish also eat valuable water-living crustaceans that help to control the growth of water algae and their droppings promote algal bloom.

PLANTING

Bank-side vegetation provides cover for animals entering or leaving the pond. Choose a selection of marginal plants, submerged plants and plants with floating leaves. Leaving long grass around a pond will also provide important cover for wildlife. Gently sloping sides make it easier for creatures to climb in and out of the pond.

Pond dwellers, such as dragonflies and caddis flies, need vegetation to climb out of the water so that the adult may emerge. Small reeds such as the lesser bulrush are ideal but any plants that grow about 40cm (16in) or more above the normal water surface will do.

A log pile nearby will provide a place for hibernating frogs and toads, as well as being a habitat for many other creatures. Add an escape log to help small mammals get out of the pond if they fall in.

Ideally the water surface should have a third or more covered with plant leaves such as water lilies, water soldier, frog-bit or even duckweed. These help keep the water cool, provide shelter and breeding cover for wildlife and help to prevent the development of algae in the water.

Plants that are to form part of your pond habitat should always be grown in a nutrient-poor potting mix. Slow-release fertilizers effectively act as an algal "time bomb", steadily releasing a supply of nutrients that encourage green water and blanket weed development.

Oxygen is essential for the growth and development of pond-dwelling creatures and can easily be boosted by the inclusion of plants that release it directly into the water. These are plants that are entirely aquatic and do not emerge above the water surface. Plants such as milfoil and Canadian pondweed are often used for this purpose.

Native plants should be a first choice for a pond. Many are showy and often rare or endangered in the wild. If you decide to use natives, never take them from the wild. Always buy them from a reputable nursery or take small offsets from a friend's pond.

Some plants are just too vigorous to be included and will rapidly outgrow their situation. Reedmace (*Typha latifolia*), reeds such as phragmites, floaters such as duckweed (*Lemna*) and the *Azolla* fern are all potentially problematic and should only be included where space permits. They are the equivalent of aquatic weeds and may cause maintenance problems.

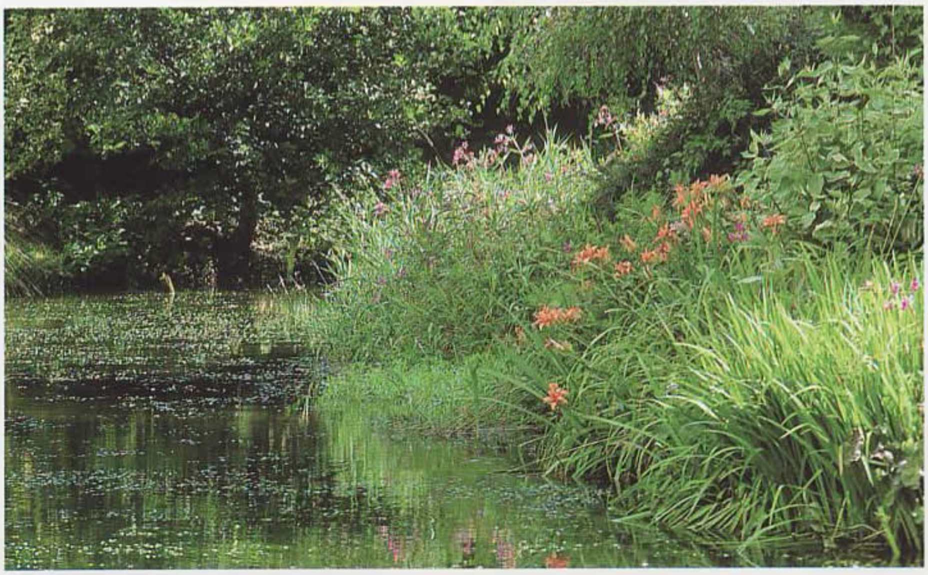

Right: *Even a small garden will benefit from a tranquil wildlife pond where animals can come to drink.*

PROVIDING SHELTER

As well as needing water and food, wildlife occupants and passing visitors to the garden need a secure place in which to shelter, where they can feed, nest, breed or escape from predators. Almost every part of the garden can potentially shelter wildlife and the key is not only to provide the right kind of cover but also to make sure that it is not disturbed once it is occupied. These sites can take the form of natural plantings or man-made shelters.

Above: *Providing winter cover such as this lacewing hotel for beneficial insects ensures a healthy population for next season.*

TYPES OF SHELTER

In nature, evergreen trees and shrubs provide year-round cover from weather and predators, while deciduous ones give summer cover for nesting and protection. Dead trees provide habitats for owls, woodpeckers and bats, while fallen logs are host to a wide array of ground dwellers such as wood-boring beetles and other insects.

Centipedes and ground and rove beetles that are active all year round need undisturbed soil, stones and logs. Frogs, toads and newts hibernate through the coldest months, often overwintering in damp hidden sites. Place stones and logs near to a pond and ensure that it has shallow edges and sloping sides. Newts prefer overhanging vegetation to shield them on entering and leaving.

Think about where specific creatures live in nature and aim to mimic these habitats. Sometimes this will involve using the natural landscape or erecting a purpose-built shelter.

Left: *Logs and branches, left over from pruning or tree felling, can be left as a habitat or winter shelter for wildlife.*

LOG PILES

These are an essential part of any wildlife garden because they offer shelter and food for a wide range of different organisms and insects that will in turn provide food for other visiting animals.

Log piles make an excellent place for frogs, toads, lizards, hedgehogs and other wildlife species to shelter, find food, and raise their young. Build a permanent log pile in a shaded area, and let it decay naturally. The logs form food for fungi and insects, while providing shelter for reptiles and small mammals. Dead wood on the ground and especially up in the trees is a very important source of nourishment for insect-feeding birds, such as woodpeckers and tree creepers.

Left: *A shady part of the garden, under the cover of trees, is the ideal place for a log pile. Log piles are both useful for hibernating animals and as a food source for beetles.*

TIDINESS

You should not be in too much of a hurry to clear and tidy up your garden in the autumn, as this will provide shelter during the winter months for a wide variety of garden wildlife. Always be selective, preferably removing only piles of diseased debris. The later you can leave your winter clearance, say until early spring, then the greater the numbers of animals there will be successfully overwintering.

WHICH SPECIES SHOULD YOU SHELTER?

There is a wide range of species that can benefit from cover in your garden. However, you should always try to assess the likelihood that these creatures will actually visit your garden in the first place. There is little point in putting up luxury accommodation for creatures that will never even see it, much less shelter within. The following are some of the creatures that may benefit from a shelter in your garden.

Above: *It is advisable to position shelters for beneficial insects near to a suitable habitat or to a source of food so that they can find them easily.*

Above: *Pollinating insects such as bees need both shelter and protection. Bees are very important inhabitants of a wildlife garden; honeybees pollinate plants in the garden.*

Bats These are probably the most elusive of our garden visitors and because of their nocturnal lives we know very little about them and often, rather unfairly, fear them. Bats roost during the day in tree hollows, in dense shrubs or climbers, and in buildings, although during winter they hibernate, often in deeper, solid places with an even, cool temperature. These fascinating mammals are declining in number due mainly to the loss of roosts and feeding areas, and the increasing use of pesticides. One of the easiest ways to help bats is by putting up bat boxes. These should be located in a sheltered position that gets morning sun and afternoon shade.

Birds The lively activity around your bird table or feeders shows that it can be easy to attract more birds into your garden. Birds need more than just a source of food, though, and suitable nesting sites can be essential for some common garden species. Trees, shrubs and climber-covered walls and fences will provide shelter and nest sites.

Though the term nesting box implies birds may nest in them, only a few species use natural or man-made cavities. Most look for shelter in vegetation, with shrubs being perhaps more valuable than trees for this purpose. You should note that regular close pruning of shrubs might cause growth that is too dense for birds to navigate easily. Try to use native species where possible. Hedges that also produce berries are a good source of food for birds in the colder months.

Butterflies These also require shelter from the elements and a place to roost. The foliage of shrubs is useful as an overnight roost, as are patches of tall grass. You can also provide a place for butterflies to roost, perch, or even hibernate by building a log pile and stacking the logs crosswise to create as many open spaces as possible. The ideal log pile size is about 1.5m (5ft) tall and 2m (6½ft) long. You can also buy a butterfly hibernation house, although there is little evidence that these are ever actually used by butterflies.

Hedgehogs If you lean a sheet of wood at an angle against a wall and stuff it with dry leaves, this will provide a suitable nest for hedgehogs. They will also nest under raised sheds and beneath hedgerows, as long as there are plenty of dry leaves for bedding. You can also provide a winter box in an old compost heap or leaf pile for hibernation.

Above: *You can encourage butterflies into your garden by planting wildlife-friendly plants such as these purple verbenas.*

Reptiles and amphibians Rocks, logs, and leaf piles offer cover for small mammals, reptiles, amphibians and insects. In the autumn, try constructing small rock piles that can provide lizards, reptiles and amphibians with a warm winter retreat. If you have a pond, leave brushwood and leaves or rock piles near the bank to provide shelter for amphibians as many of these hibernate in damp areas on the land.

Pollinators Bumble-bees can overwinter in a small terracotta pot filled with moss or finely shredded newspaper. Bury the filled pot upside down and protect the hole from rain with stones, but leave a tiny gap for the bee to get in and out. Masonry bees can be encouraged to take up residence by tying a bundle of bamboo canes together – lengths of about 25cm (10in) are ideal – and then placing them inside a length of old plastic pipe that is about 5–7.5cm (2–3in) longer than the canes. They will make nests inside the hollow bamboo canes and lay eggs within these that will hatch out the following spring. Always place the pipe in a sunny sheltered position where rain will not reach it. You can also buy purpose-built masonry bee shelters.

Predators Lacewings will shelter in a similar construction to that used for masonry bees. Ready-made lacewing hotels are available and are impregnated with a pheromone that attracts the adults in the autumn. Ladybird (ladybug) shelters are also marketed commercially for the same purpose. You can also buy the pheromone with which to impregnate your own home-made shelters.

PLANTS TO ATTRACT WILDLIFE

Plants are the single most important factor when it comes to attracting wildlife into your garden. They provide a food source and shelter for visiting bees, butterflies, moths and a whole host of beneficial insects. Even if your garden is tiny, you can still introduce a selection of useful plants by planting tubs of brightly coloured and nectar-rich summer flowers and pots of sweet-smelling herbs such as lavender and rosemary.

Above: *Beneficial predators such as ladybirds (ladybugs) will be attracted to the garden if you include the right species of plant.*

PLANTS FOR FOOD AND SHELTER

Planting colourful, nectar-rich plants attracts butterflies and other beneficial insects. Choose plants that will provide the visitors with nectar from early spring to late autumn. This of course has the added benefit that your flower beds will be brimming with colour for most of the year. Combine selected species of herbs, wild flowers and cottage garden plants in a herbaceous border to create a butterfly or bee border.

To encourage breeding insects, allow a few nettles and grasses to grow in a secluded corner of the garden or let ivy trail up your fence. These will provide sites for female butterflies to lay eggs and a food source for caterpillars.

Wild or single-flowering varieties of plants are often best for wildlife. You should not be too eager to dead-head plants in autumn.

Left: *Plants with composite flower-heads, such as thistles and daisies, readily attract insects to the garden.*

Many birds feed on spent flower heads such as those of forget-me-nots and pansies. Leave perennial flower stems in autumn, as birds will be attracted to them, searching for seeds and hiding insects. Nesting birds need insects to raise young, so try early-flowering plants such as aubrietia that attract aphids and provide a source of food.

A hedge that is left relatively undisturbed by the gardener can prove a rich habitat for a diverse range of wildlife. Dense hedges give birds protection from predators and provide windbreaks. The thick undergrowth encourages many species of small bird to take shelter and unclipped hedges of all types have more chance of bearing flowers and fruit.

If your garden is large enough to have at least one large tree then this will provide perches and nesting sites for many birds. Native trees are valuable to insects, birds and other animals as they provide better shelter and food than introduced species. The best trees are those that produce

Above: *Choose plants such as roses that have berries or edible fruits for birds and mammals to feed on.*

Above: *Sunflowers (*Helianthus annuus*) provide food for insect pollinators as well as late seeds for overwintering birds.*

Above: *Scabious (*Astrantia*) can be used in flower borders or in wild flower lawns in order to attract insects such as bees.*

seeds or fruit, as these are often a valuable source of food for many birds in autumn or winter.

Shrubs provide shelter for birds, enabling them to nest safely. Choose shrubs such as *Pyracantha*, *Cotoneaster* and *Viburnum* that produce a good crop of berries for the birds to feed on. Climbers are attractive to birds, especially if they are late flowering or bear fruit and seeds for winter food.

Annual species, planted on vacant patches where quick reliable displays of colour are required, will produce an outstanding array of colour, and their nectar and seeds provide valuable food for many types of insect.

Meadow-plant species will provide colour while the foliage of naturalized bulbs withers down. A fine early display can be created under deciduous trees before the soil becomes too dry and shaded, and grassy banks provide a marvellous opportunity for creating a mini meadow. An informal lawn is the intermediate step between a formal lawn and a meadow and is often created adjacent to a formal lawn and sown at the same time. It contains species tolerant of close cutting and if mowing is relaxed in summer, an enchanting display of flowers can be enjoyed for many weeks. When mowing is recommenced no harm will come to the plants, and the practice can be repeated annually.

Above: *Encouraging beneficial insects into your garden by including insect-friendly plants can help to create a natural balance.*

Shaded areas of the garden often present a problem for gardeners, but there are many varieties of woodland plants – often including some of the most beautiful wild flowers such as anemones and foxgloves (*Digitalis*) – that will thrive in these conditions. A woodland area planted with native trees and shrubs is the ideal backdrop for bulbs and woodland wild flowers. This habitat provides a continual source of interest throughout the year and attracts insects, birds and mammals.

Plants growing along and in hedgerows are particularly useful to attract wildlife, especially if the hedge is made up of native species. Plants for sun, shade or semi-shade should be selected depending on the aspect of the hedgerow. New hedgerows should be allowed to establish free of competition for two to three years before associated plants are introduced. Hedges provide corridors for wildlife and sources of interest throughout the year, with attractive foliage in spring, flowers in summer and berries, hips or leaf colour in autumn and winter.

Plants that grow in or around ponds act as a magnet to surrounding wildlife. Damp soil borders and marshy areas are the ideal habitat for many species of wild flower.

Dry-stone walls can also be exploited by planting up nooks and crannies with wild flowers adapted to the conditions, adding interesting foliage and colour.

Combinations of wild flowers can also make excellent subjects for patio tubs, pots and hanging baskets, as many of them are tolerant of drought and have a spreading habit, making them ideal for low-maintenance containers.

Left: *As its common name of butterfly bush suggests, buddleia is a particularly excellent source of nectar for butterflies.*

TOP 20 GARDEN PLANTS TO ATTRACT WILDLIFE

There are literally thousands of plants that all have the potential to attract various forms of wildlife to the garden. Some plants, however, are more effective at this than others. Twenty of the best plants for attracting wildlife that are easy to grow are listed below.

PLANT		WILDLIFE ATTRACTED	BENEFITS TO THE GARDEN
Achillea millefolium **(Yarrow)** A vigorous herbaceous perennial that grows to a height of 60cm (2ft). Flower colour ranges from white through to reds and oranges.		An abundance of hoverflies can be found nectaring on the flowers in summer. Also attracts bees and bumble-bees.	Hoverfly larvae are an excellent addition to the garden, providing efficient control of aphids. The larvae can eat vast amounts of aphids during this part of the life-cycle.
Angelica archangelica **(Angelica)** A tall, upright, herbaceous perennial, more often grown as a biennial, which grows to 1.8m (6ft). The architectural flower-heads bloom in early to mid-summer.		Bees, butterflies, hoverflies, insects and birds all congregate around the huge flower heads of this garden bloom.	This plant can be likened to a supermarket where many species gather to stock up on food. Count the number of species on one bloom, it is amazing.
***Aster novi-belgii* (Michaelmas daisy)** The flowers of this herbaceous perennial provide a rich source of late nectar. A full range of sizes and colours are available.		In late summer and autumn butterflies and bees visit the flowers. Birds eat the seed after ripening.	This plant provides a valuable late source of nectar. Butterflies will use this food to strengthen them up for winter hibernation or to produce eggs to overwinter.
Aurinia saxatilis **(Golden alyssum)** Also known as *Alyssum saxatile*, this herbaceous perennial forms a carpet of yellow flowers in spring.		Butterflies, bees and early flying hoverflies are attracted to the bright yellow flowers of this popular rock garden plant.	A very useful source of nectar for all early insects, especially early season or overwintering butterflies.
Buddleja davidii **(Butterfly bush)** This deciduous shrub is often quoted as the saviour of the British butterfly population due to the high quantities of nectar it produces.		As the name suggests, this is an excellent plant for attracting butterflies. Also useful for a variety of nectar feeders, including bees and moths.	Like asters, this is an excellent feeding stop for migratory and resident butterflies. The caterpillars are a good source of protein for birds.
Centranthus ruber **(Red valerian)** A herbaceous perennial that grows to 60cm (2ft). It performs well in poor dry soils and does especially well in seaside locations.		One of the best flowers for butterflies, providing nectar for small tortoiseshells, peacocks, red admirals and white and yellow butterflies.	Butterflies are a good indicator of the health of the environment. Their caterpillars are a good source of protein for nesting songbirds.
***Cerinthe major* 'Purpurascens'** A small herbaceous perennial that flowers in spring. Produces fantastic blue flowers with lilac bracts. It self-seeds freely.		Bees love the blue flowers and congregate in vast numbers while the plant is in flower.	This early flowering perennial gives a boost of nectar to resident bee populations, which are good pollinators.
Eryngium giganteum **(Miss Willmott's ghost)** A short-lived herbaceous perennial that reaches a height of 90cm (3ft) and has beautiful steel-blue flowers.		This plant hums with bees and bumble-bees during its long summer-flowering period.	An outstandingly beautiful plant that attracts bees which will pollinate your plants. Good for naturalizing in wild areas.
Eschscholzia californica **(Californian poppy)** An annual that produces striking orange, cream, red or yellow flowers which will bloom all summer. This is a small plant, reaching 45cm (18in) in height.		Hoverflies, bees and bumble-bees are attracted to this plant. Once it is sown, it will self-seed freely round the garden.	The bright orange flowers attract hoverflies, the larvae of which are extremely voracious predators of aphids.
Eupatorium cannabinum A tall plant that produces large heads of fluffy pink flowers in late summer. This herbaceous perennial is a must for the wildlife garden.		Attracts butterflies and bees. As well as looking fantastic, this plant is one of the best for providing nectar to adult butterflies.	From mid-summer onwards, this decorative plant offers a rich and steady source of nectar and is a firm favourite with visiting butterflies.

PLANT		WILDLIFE ATTRACTED	BENEFITS TO THE GARDEN
Helianthus annuus **(Sunflower)** If you want to make a statement, grow a few of these in your garden. They have enormous yellow, red or orange heads with some tall varieties easily reaching 2m (6½ft).		Birds such as greenfinches and bullfinches will eat the seeds. Also visited by honeybees and bumble-bees when in flower.	The seeds provide a late-summer boost to the diet of birds and some small mammal species. Good for encouraging birds into your garden.
***Humulus lupulus* 'Aureus'** **(Golden hop)** A vigorous climber with twining stems. It is unusual in that it dies down every year. The female plants produce hops that are used in flavouring beer.		Caterpillars of the beautiful comma butterfly and button snout moth feed on the golden foliage.	Food plants for caterpillars and larvae are an invaluable part of a wildlife garden, otherwise we would not be graced with the adult butterflies.
Lavandula angustifolia **(Lavender)** A compact bushy shrub that sports many varieties to choose from. It produces fragrant pink to purple flowers in early summer.		The fragrant blooms attract many different species of bees, bumble-bees and also white and blue butterflies.	Insect-eating birds are attracted to feast on the insects drawn to nectar on these fragrant bushes.
Lonicera periclymenum **(Honeysuckle)** A climber with twining stems reaching up to 6m (18ft). It bears sweetly scented, pink and cream flowers in early summer, with red berries appearing later.		Nectar provides food for bumble-bees, bee hawk moths as well as elephant, lime and privet hawk moths. Fruit-eating birds will feast on the berries.	A superb plant for night-flying moths that should be placed next to the house so you can enjoy both the fragrance and watch these night-flying beauties. The moths are also a good source of protein for birds and bats.
Nymphaea alba **(Waterlily)** An aquatic plant that has large round floating leaves with beautiful white flowers in the summer months.		Adult frogs and dragonflies use the floating leaves of this aquatic plant for basking on. The leaves also provide cover for tadpoles of newts and frogs.	Adult frogs eat slugs and snails. The tadpoles help to control algae in the pond, as it is part of their diet.
Phlox paniculata The lovely, sweetly scented flower heads provide a colourful display in late summer. This 90cm (3ft) herbaceous perennial prefers a moist soil in light shade.	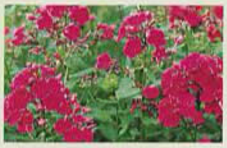	This plant will attract adult butterflies and moths, which will appreciate the nectar from the flower heads in late summer to early autumn.	The moths nectaring on this plant will provide a tasty meal for wandering bats. Butterflies flying during the day will provide a meal for hungry birds.
Rosmarinus officinalis **(Rosemary)** A dense bushy shrub with aromatic foliage growing up to 1.5m (5ft) tall. It bears purple to blue flowers in mid-spring to early summer.		Almost unrivalled in its ability to attract honeybees, butterflies and hoverflies to its nectar-rich flowers.	Honeybees pollinate many garden plants and trees. They are welcomed by farmers into fruit orchards.
Sedum spectabile **(Ice plant)** This small herbaceous plant is one of the best butterfly plants. It has thick succulent leaves and stems that are well suited to growing in hot well-drained positions.		An old favourite for the herbaceous border, Sedum is extremely useful to butterflies, bees and hoverflies as a late summer source of nectar.	Provides a good source of nectar late in the season that is a positive benefit for overwintering butterfly species to build up their energy.
Sorbus aucuparia **(Mountain ash)** A wonderful tree for the small garden. It has attractive white flowers in the spring and bears red fruit in the autumn. It can reach heights of 12m (36ft).		The flowers are excellent for feeding bees and other insects in the spring whereas the berries are consumed by many species of birds.	The tree can attract up to 30 different species of insect, not including the birds and small mammals it feeds, providing a meal for many.
Verbena bonariensis **(Tall verbena)** This versatile plant can be grown as an annual or short-lived perennial, reaching up to 1.8m (6ft). The purple-flowering spikes bloom in late summer until the first frosts.		This is a fantastic plant for nectaring butterflies such as the small tortoiseshell, peacock, red admiral and comma. It is perhaps more effective than the more well-known buddleia.	Good for planting in large groups throughout the garden to attract large numbers of butterflies and bees to pollinate plants and to provide food for garden birds.

THE KITCHEN GARDEN

The kitchen garden is a paradise for gardener and cook alike. Few pleasures in life can compare with the satisfaction of harvesting and eating fresh produce from your own garden. There are many design possibilities when planning a kitchen garden. Some gardeners prefer to lay out their vegetable plot in neat rows; others choose an edible landscape that is ornamental as well as functional. There is also a seemingly endless choice of varieties in seed and nursery catalogues which can be confusing, but merely reflects the popularity of home-growing.

Left: *Kitchen gardens can provide a summer spectacle as well as a wealth of fresh, home-grown food for your table.*

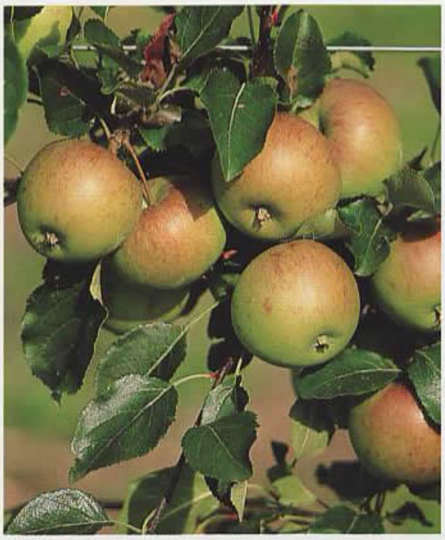

Above: *Fruit such as pears and apples is one of the most welcome delights of autumn in the kitchen garden.*

Above: *Companion planting helps to keep plants healthy and also looks attractive throughout the summer.*

Above: *Kitchen-garden plants can be just as decorative as those grown for their ornamental qualities.*

PLANNING A KITCHEN GARDEN

Growing your own vegetables, herbs and fruit ensures that the food you eat is fresh, tasty and chemical-free. The healthiest food is found closest to its source, so freshly picked produce from your garden is both tastier and better for you than that which has travelled long distances by air, sea or road. Your garden will also be more appealing when it is planted with a wide selection of plants that will attract an equally diverse range of wildlife.

Above: ***A well-planned kitchen garden, with companion planting, can be a feast for the eyes as well as the table.***

WHAT DO YOU WANT FROM YOUR KITCHEN GARDEN?

An ideal starting point is to make a list of all the vegetables and fruit that you eat on a regular basis, adding in a few that you are curious about and would like to try. The next stage is to look at the size of your garden and consider if there is enough room to grow everything that is on your list. Think about the quantity needed to feed your family. There is no point in planting ten apple trees for a family of four. It is a good idea to start with a few easy crops to build your confidence and progress from there.

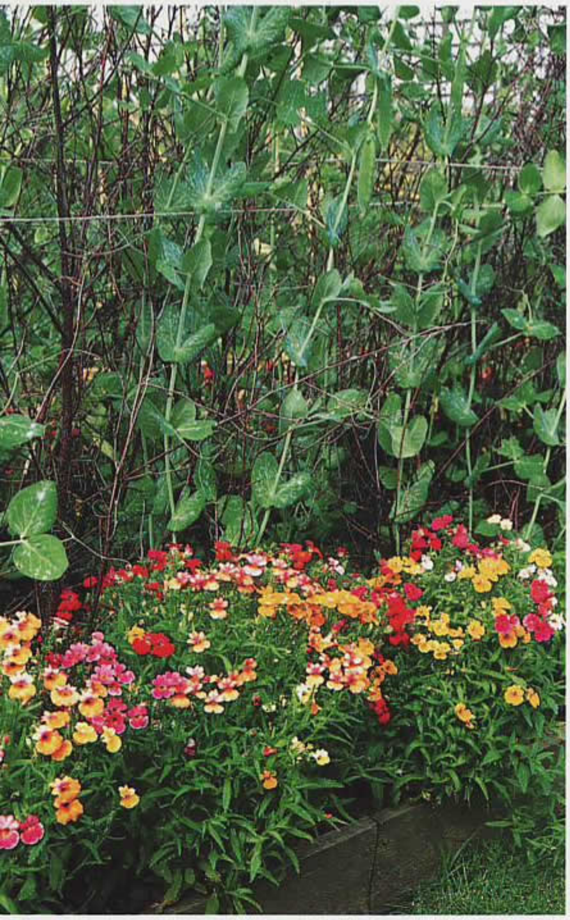

Many edible plants, especially vegetables and herbs, are easy to grow and can be ready for harvesting in a short space of time. Children love to sow and grow less difficult crops, such as spring onions (scallions), radishes, carrots and lettuce. Growing fruit is usually a longer-term investment, as many types, such as apples, pears and blackcurrants, for example, will take several years to crop. Once they start to produce fruit, however, your investment will pay off, as they will continue to crop for many years to come and are less time-consuming than vegetables.

Furthermore, you do not need acres of space to enjoy growing your own fresh fruit and vegetables. In fact, even a small garden will support an edible crop or two. You could grow potatoes in barrels, herbs in pots and fruit trees in ornamental containers. If you live in an apartment, you can use window boxes and containers to grow fresh herbs and vegetables. Try planting hanging baskets with trailing tomatoes, herbs or fruits such as strawberries to overcome space difficulties. Edible gardening involves using any space that is conveniently available. It also promises a colourful garden and unusual ways to feed your family.

DESIGNING THE GARDEN

There are a number of points that you will need to consider in order to get the most out of your kitchen garden and to ensure that the food you produce is safe and fresh. Siting your crops next to a busy road, for instance, can mean eating produce that is polluted by cars. When siting the crops, think carefully about your own garden environment. Each garden is different and has specific requirements for growing various crops. Check the pH of your soil before getting underway as vegetables such as cauliflowers and cabbages will not grow well if the soil is too acidic (under pH of 6.5). As well as testing the pH level, test the soil for nutrients. The results will indicate if there is any need to improve the soil before planting.

Left: ***Peas and beans can easily be trained up supports, giving height and structure to the kitchen garden.***

The aspect of your garden is also critical in the siting of your plants. Fruit, for example, needs plenty of sun exposure in order to ripen properly and develop flavour. Fruit trees are also often best sited over a lawn to facilitate fruit collection.

Any design needs to consider the difference between annual and perennial crops. Annual vegetable crops, such as potatoes, carrots, cabbages and beetroot, need to be rotated within the vegetable garden to deter soil-borne pests and diseases from attacking them. Strawberries are a good example here, as they are effectively an annual, needing to be divided every year. A strawberry bed that you rotate with flowers and annual herbs (such as basil or dill) is useful for avoiding soil-borne pest problems.

Perennial crops, such as asparagus and rhubarb, may well merit their own distinct areas and are not rotated like annual crops. Raspberries and blackberries, which are both perennial, are best sited permanently on a wire trellis. Avoid putting them in a huge expanding clump that can take over your garden.

Consider, too, how difficult a particular plant (or the produce it offers) is to grow. Peaches, for example, can be challenging. They are disease-prone and require lots of care and attention. Unless you are a committed gardener with time on your hands, it may be best to purchase a few organically grown fruits from a local shop.

PLANTING THE KITCHEN GARDEN

If you have recently moved into a new home, then you will have the luxury of starting your kitchen garden from scratch. If your garden is already established, there is no need to uproot everything to grow your own produce. Gardens need not be exclusively made up of just edible or ornamental plants. You may wish to start by planting a few annual vegetables and herbs among your ornamental plants. Indeed, you can easily incorporate edible plants into the garden, even mixing fruit, vegetables and herbs in ornamental beds. This planting arrangement avoids crops grown in visually uninspiring mono-cropped rows.

If you are planting vegetables in new beds, then they can be interplanted with ornamentals and herbs. Fragrant plants such as English marigolds (*Calendula*), French marigolds (*Tagetes patula*) and oregano (*Origanum*) are excellent choices for attracting beneficial insects, and interplanting with plants such as these will help to keep pests to a minimum. Large blocks of the same vegetable are likely to attract high concentrations of pests, whereas interplanting tends to confuse and dissuade them.

If you need to remove an ornamental tree or shrub that has died or outgrown its site, consider replacing it with a fruit-bearing tree or shrub. There are many possibilities, including apples, currants, raspberries, crab apples, plums and cherries, which can all provide valuable colour and texture in the garden as well as a source of food. Apart from feeding the household, excess fruit from these trees will also provide food for a range of birds and insects.

Right: *Cloches and low tunnels can increase the range of early crops that can be grown in the kitchen garden.*

It is worthwhile researching the eventual height of plants in the planning stages. Some will simply look out of place if they are grown in the wrong location, such as planting tall plants in front of smaller ones. Many large ones will also need some form of staking or support.

Lettuce, chives, pansies and parsley create excellent borders along the edges of raised beds. Tall plants such as dill, sunflowers, daylilies, fennel, valerian, peas and beans are best grown at the back of beds or at the centres of containers. Provide trellises and other supports where needed. Choose edible flowers such as nasturtiums and chives wherever possible because they are a natural addition to gardens and make salad bowls look and taste wonderful.

You may wish to consider an edible lawn, in which some (or even all) of the lawn is given over to a groundcover such as strawberries. Strawberries are a very attractive crop because they produce fruit for most of the summer and tolerate marginal soils and light shade. However, large patches are likely to attract pests, so they should be interplanted with strong-smelling herbs like thyme and oregano which also form good groundcover. It may be an idea to keep a permanent cover of herbs and rotate the strawberry crop on a two- or three-year cycle.

PLANTS USED AS FENCES, SCREENS AND BARRIERS

Plants can be used in a number of imaginative ways. Instead of erecting plain wooden fencing or mesh barriers why not consider planting a living one?

Above: *Onions, with their tall stems, can be grown for their decorative appearance in ornamental beds and borders.*

Gooseberries, raspberries or currants are readily trained along a fence and apples can be trained as an espalier cordon or fan along a wall. Climbing plants, such as hardy kiwi fruits, trailing nasturtiums, broad beans, or sugar snap peas, can all be grown over a fence or trellis work to provide an ornate screen. Hedges can also form useful and attractive edible barriers. Shrub roses, such as *Rosa rugosa*, create an attractive, but impenetrable barrier, producing large red rose hips that contain 60 times the vitamin C of an orange. The hips can be used to make tea, jam, syrup or jelly. Currants and other fruit can also be included on the sunny side.

When designing your kitchen garden, bear in mind that the garden is for everyone in the household. Hold family discussions to involve everyone in the planning and design stage. It is worthwhile taking everyone's needs and tastes into consideration when undertaking this extremely important stage of development.

CROP VARIETIES

There is a seemingly endless array of seeds available to the home gardener. Each supplier makes claims that their variety is better than all those that went before and any others currently available. Others claim to have older, more choice varieties saved from extinction and representing a time when everything was purer and more wholesome. Personal preference usually decides the best varieties for you, but an understanding of what you can expect from the seed in a packet can be very useful.

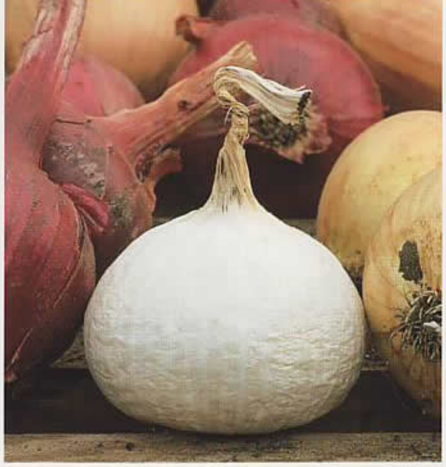

Above: *Onions come in a range of attractive colours and can be grown for their ornamental value in beds and borders.*

WHAT ARE HYBRIDS?

Often when we purchase seed, we respond more readily to the picture on the front of the packet. We may not notice whether it is marked with the terms "F1" or "hybrid". A hybrid is the result of a cross between one variety with pollen from another specific variety. The breeder chooses parent varieties that will produce first generation offspring (F1 hybrids) with known characteristics. The crossing is done in a very controlled manner so that no pollen from another variety is able to pollinate the flowers. As a result, all of the plants that are grown from the hybrid seed will be genetically identical.

Hybrids may be bred to be more widely adapted to environmental stresses such as heat, cold, disease or drought than non-hybrids. They also have more uniform characteristics, making crops more predictable in their qualities. They also have "hybrid vigour" and may grow faster or be more disease-resistant than either of the parents. They may also give better yields than open-pollinated varieties. They will not breed true, however, meaning that seed collected will not produce plants that are the same as the parent (F1 hybrids). For this reason, seed cannot be saved from F1 hybrid plants by the home gardener. Seed for hybrid varieties must be purchased year after year from the seed companies or nurseries, unless you want to gamble and grow an array of offspring.

WHAT ARE OPEN-POLLINATED VARIETIES?

Open-pollinated varieties are traditional varieties that have (in some cases) been grown and selected for desirable traits such as taste, yield or disease-resistance for many years. They often grow well in organic kitchen gardens as many were originally selected under organic conditions. These plants can mutate and adapt to the local ecosystem, as the seed is often collected and re-used by the organic gardener.

If a seed packet is labelled "heirloom", "open-pollinated" or has no special markings, then it is most likely a standard or traditional variety. The majority of lettuce, bean and pea varieties for domestic use are open-pollinated, while most cabbages, broccoli, tomatoes, cucumbers, melons and Brussels sprouts are hybrids.

Right: *Allowing a few of your crops to flower will also provide you with seed for growing next year's crop.*

Left: *Unusual crop varieties can often be grown from seed that you have collected yourself each year. This is a striking variety of kohl rabi.*

Hybrid seeds can dominate the garden seed market, but open-pollinated varieties are more or less stabilized in their characteristics, remaining fairly consistent and producing seed that will grow into plants that are more or less like the parent plants. They are a little less uniform than hybrids, but the home gardener can safely collect seed and grow plants from them that will be essentially the same as the original plants. Open-pollinated varieties either self-pollinate or are pollinated by wind or insects and they usually produce viable seed.

There have been various claims that open-pollinated varieties do not taste as good as hybrids. It is also claimed that they are smaller and not as uniform and, in many cases, this may be true. However, where matters of taste in varieties are concerned, the only real answer is to try them and see for yourself.

Right: *Old varieties of tomato tend to have varying sizes of tastier fruits that crop over a long period of time.*

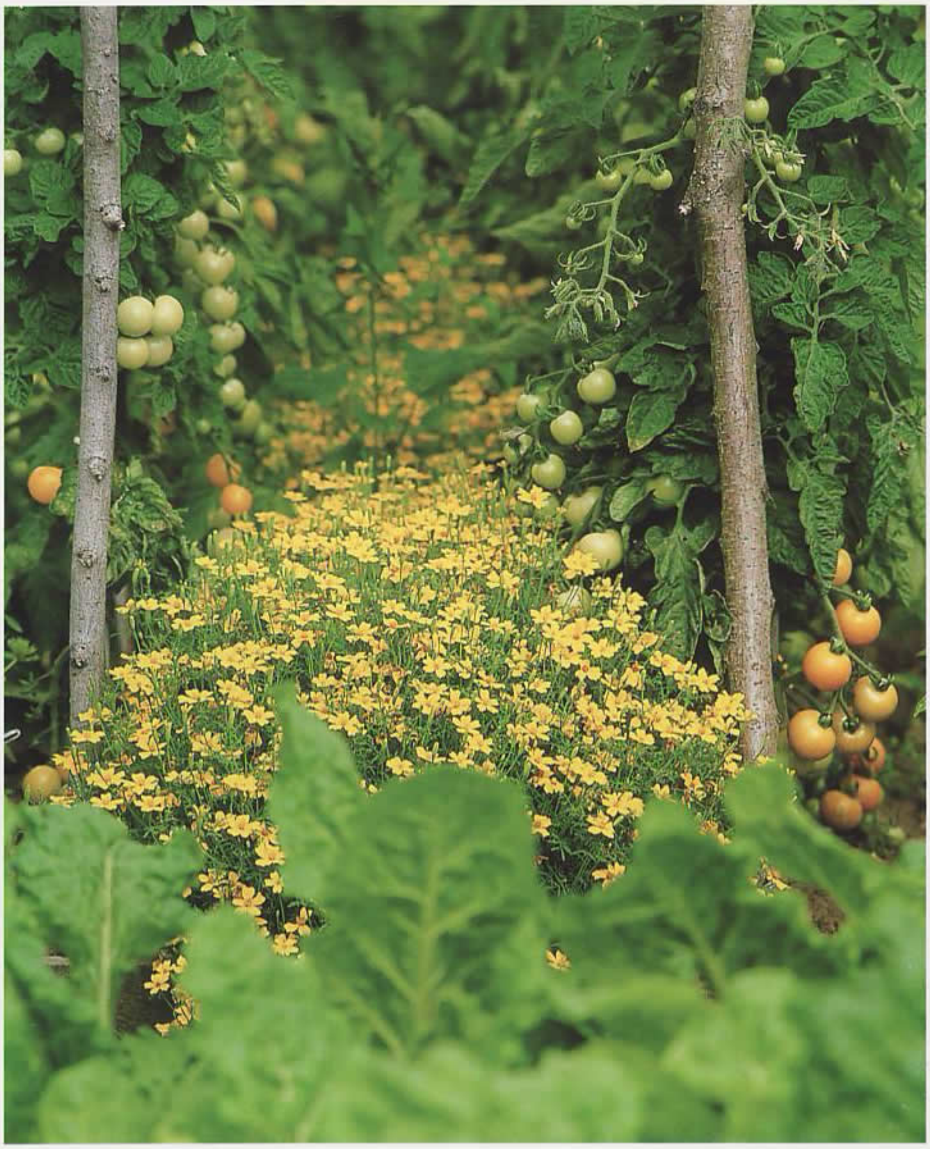

WHAT ARE HEIRLOOM VARIETIES?

In the 1970s the European Community brought in regulations to encourage the breeding of new vegetable cultivars and the standardization of older ones. This resulted in a list of approved cultivars being drawn up and it became illegal to sell any cultivar that was not included on this list. It is very expensive to have a single cultivar tested in order to then register it on the list. This meant that many old cultivars were put in grave danger of being lost forever, as the funds to test each variety were not available. It was due to this legislation that HDRA, the Organic Organisation, in England, a society which was established to promote organic issues, founded the Heritage Seed Library (HSL). This seed library ensures that old or "heirloom varieties" are kept safe for posterity by distributing its seed. Although HDRA grow some of the seeds themselves, they also employ contract growers and seed guardians to make up the bulk of the seeds that are supplied.

Each year a catalogue is sent to HSL members from which they can select up to six varieties free of charge. This distribution set-up overcomes the clause of selling only approved cultivars. This service is also available to members in the United States.

The best heirloom cultivars can be traced back fifty years or longer. Many of these early varieties have been lost already, making those that remain all the more precious. A number of these cultivars have been collected and saved by families and ethnic groups dating back many years. This practice protected the genetic make-up that made each variety successful within a given environment. These base characteristics have become invaluable and the genetic strains of these vegetables are the backbone of modern disease- and drought-resistant hybrids. It is this that makes the collection and preservation of these cultivars so important.

Left: *Many heirloom varieties have been selected for the way they perform in local garden conditions, rather than for crop size or uniformity.*

CROP ROTATION

This is the practice of grouping and growing related plants together and rotating them around different areas of land in a regimented fashion from year to year. Rotating your crops in this way has many advantages, including helping to prevent pest and disease problems from arising in the first place. This method of gardening is fundamental to successful organic growing. It has been practised for thousands of years and developed into a system that is easy to follow.

Above: *The secret to achieving a thriving and healthy kitchen garden is to rotate the crops on a regular basis.*

WHY ARE CROPS ROTATED?

Continuous cropping in the same area puts both plants and soil at risk. It not only allows large numbers of soil-borne pests and diseases to build up, but, because crops require the same nutrients from the soil year after year, the practice can deplete nutrient levels. A poor infertile soil produces weak unhealthy plants, which, in turn, will be more prone to pest and disease attacks. All these problems amount to reduced yields and even complete crop failure.

When crops are rotated, the groups are divided up into closely related plants that are prone to similar pests and diseases. For example, carrots, parsnips, beetroot and potatoes are members of one group and prone to carrot fly, whereas cabbage, kale, broccoli and Brussels sprouts, which are members of another group, can be prone to clubroot and cabbage root fly. If the groups are grown in different areas on a rotational basis, it can help to prevent the establishment of soil-borne pests and diseases.

Combined with regular additions of compost and manure, crop rotation will make the soil richer, replace certain nutrients and help prevent pH imbalances that can result from repeated crops of the same type of vegetable. Companion planting will also aid a rotational plan, particularly if the species improve pest control.

MAKING A ROTATIONAL PLAN

The basic rules of crop rotation are simple. If you are planning a four-year crop rotation, the plants you have selected are split up into five groups. Group one contains the legumes (peas and beans); group two contains the brassicas (cabbages, Brussels sprouts, broccoli, kale etc); group three contains the onion family and others (onions, lettuce, garlic, sweetcorn etc); group four includes root crops (potatoes, parsnips, carrots, beetroot etc); and finally group five houses permanent crops such as asparagus and rhubarb.

The vegetable plot is then divided into five sections. The permanent crops are given a specific area and are not moved or included in the rotational cycle. The remaining four groups are allocated an area in which to grow. Every year each group is moved on to the next plot, making it four years before the crops are grown on the same area of land again; hence the name crop rotation.

If space is limited, crops can be grown on a three-year rotation. Quite simply, the crops are split into three groups, plus the permanent ones. The plant groups are divided up as before except group three is incorporated into group one.

There is no need to grow each crop in every year of a cycle. Remember that it is the vegetable groups that dictate the cropping cycle according to their soil needs and any associated problems. Rotating the crops helps provide the correct soil requirements for certain crops. For example, cabbages and the rest of the group grow well in soil that has been manured the previous autumn, whereas carrots and other root crops (not including potatoes) do not. Where carrots are to be grown, then the plot will need to be dug deeply in readiness for them.

BENEFITS OF CROP ROTATION

There are many benefits to be gained from rotating your crops regularly.

- Prevents the build-up of soil pests and diseases
- Helps to prevent nutrient depletion from heavy-feeding crops
- Rotated crops produce higher yields
- Results in a healthier soil

It is important that you plan where the crops are to be every year, so that you know their position for the following year. A comprehensive cropping plan can help you to maximize the yield on a year-round basis by working out successional sowings and intercropping and catch cropping. The plan needs to include not only the crops that will be grown but also the companion plants and the soil amendments that are needed to support the best possible growth of your plants.

COMMON DIFFICULTIES

When planning the cycle, you may encounter certain problems. Careful planning will show that potatoes often take up more space in a vegetable bed than any other crop and finding enough room to grow them in their allocated area may prove difficult. Other problems you may come across include overwintering brassicas or plants left in the ground for seed collection. This is when you find that practising crop rotation in a small garden is difficult and that you may not be able to practise a strict rotation.

If space is proving a problem, consider some of the following strategies. Keep brassicas together as a group and never plant them in the same ground two years running. Keep potatoes together every year; if you have planned a lot of potatoes, move all the other members from group four into group three. Also bear in mind that some root crops, such as potatoes, are manure-friendly, while others, like parsnips, are not. Finally, alternating shallow-rooted plants like cabbages or lettuce with deep-rooted plants like tomatoes or squash will allow the plants' roots to do much of the soil loosening that would otherwise have to be done by hand. This will help to preserve the health of the soil, while causing minimal disturbance to its ecosystem.

FOUR-YEAR CROP ROTATION

Divide the vegetables you have decided to grow into the five groups (plots 1–5) shown in the table. Draw a plan to indicate which group of crops goes where, using a different colour for each group. (Remember plot 5 is for the permanent crops.) Next year, move the crops in each group on to the next plot.

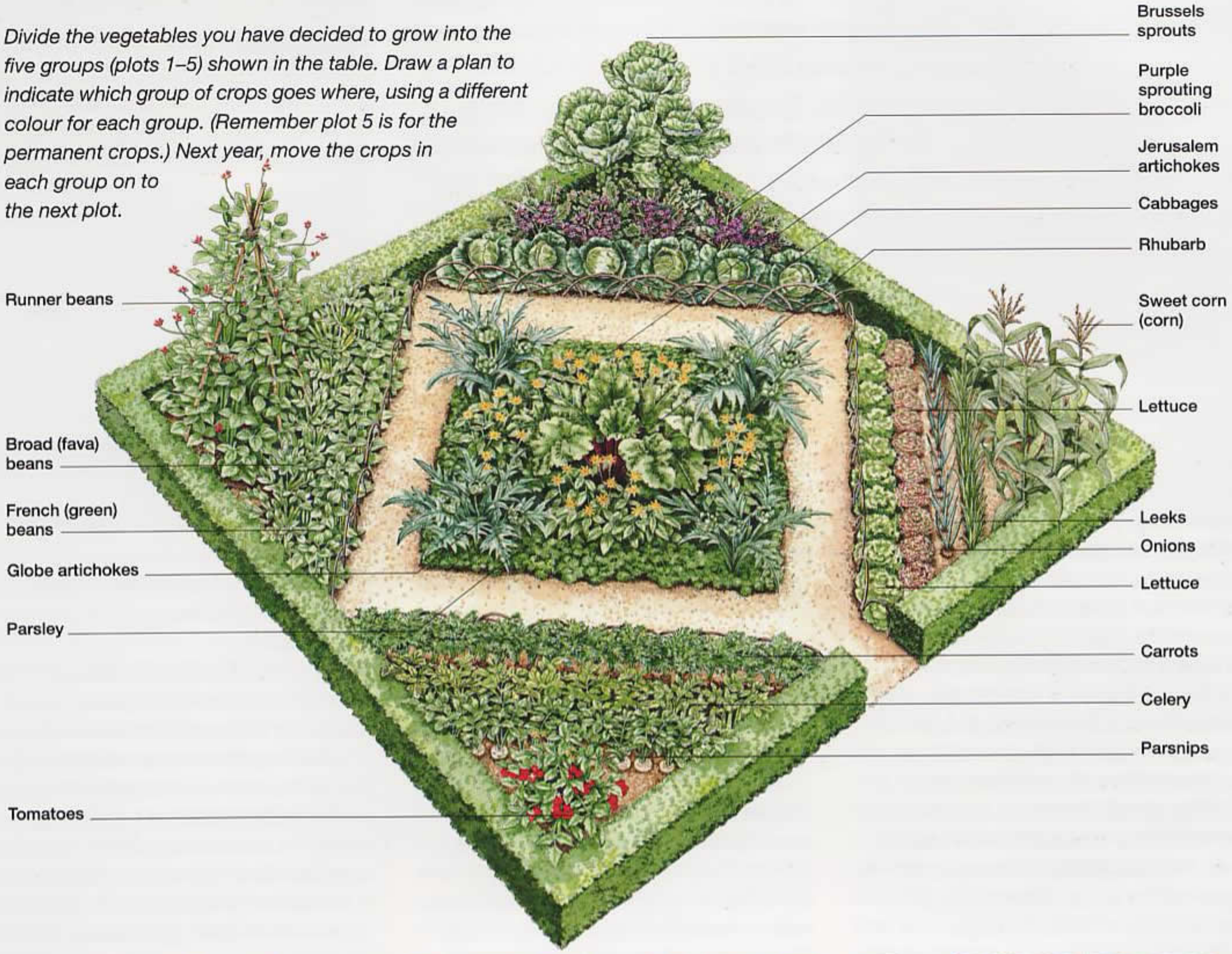
Brussels sprouts
Purple sprouting broccoli
Jerusalem artichokes
Cabbages
Rhubarb
Sweet corn (corn)
Lettuce
Leeks
Onions
Lettuce
Carrots
Celery
Parsnips
Runner beans
Broad (fava) beans
French (green) beans
Globe artichokes
Parsley
Tomatoes

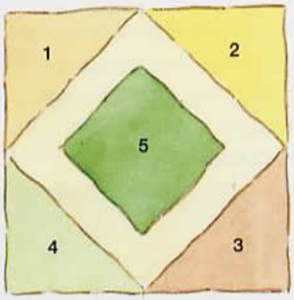

YEAR ONE

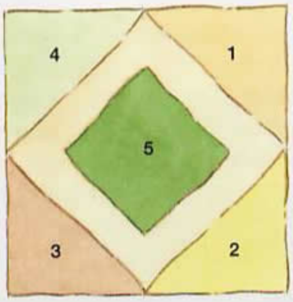

YEAR TWO

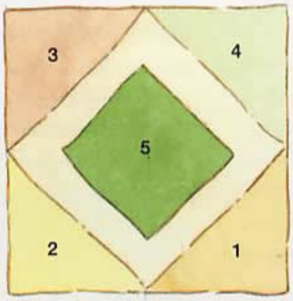

YEAR THREE

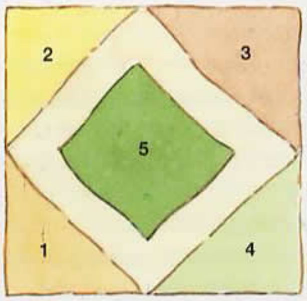

YEAR FOUR

Plot 1
Peas
Broad (fava) beans
French (green) beans
Runner beans

Plot 2
Cabbages
Brussels sprouts
Calabrese (Italian sprouting broccoli)
Broccoli
Kale
Radishes
Swedes (rutabagas or yellow turnips)
Turnips
Kohl rabi

Plot 3
Bulb onions
Spring onions (scallions)
Shallots
Leeks
Garlic
Sweet corn (corn)
Marrows (zucchini), squashes and pumpkins
Lettuce

Plot 4
Potatoes
Parsnips
Beetroot (beets)
Carrots
Salsify
Scorzonera
Celery
Celeriac
Tomatoes

Plot 5
Rhubarb
Asparagus
Perennial herbs
Globe artichokes
Jerusalem artichokes
Seakale

GROWING METHODS

There are many different ways to grow your crops. You can try the traditionally practised method of sowing crops in rows or be more adventurous and grow them in blocks. The crops you have chosen to grow can be planted into a number of different types of vegetable bed: flat beds, raised or edged beds and no-dig beds. Each method and type of bed has its own advantages and disadvantages, which are described here, so choose the one that best suits your needs.

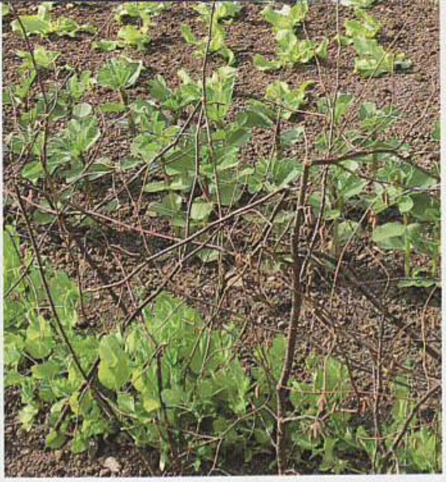

Above: *Twiggy sticks provide an ideal support for young climbers such as peas. They also look very decorative.*

PLANTING IN ROWS

This is the most traditionally practised and popular form of vegetable growing. It is an effective method because the spacing of the plants allows the crop to grow without excessive competition. In addition, it is relatively easy to add supports and protection such as cloches to the crops. However, this type of planting does have the disadvantage of needing relatively high maintenance. Due to the large amounts of bare earth left between the rows, the conditions are ideal for weed growth. Rows are spaced according to the optimum growth of the plants and so rows of pumpkins need to be much wider than those for carrots. Setting out rows in this way allows free movement of air along the rows. This results in fewer disease problems which can occur with more closely grown methods such as block plantings.

The rows are quite easily maintained by regular hoeing but can be wasteful of space, particularly in smaller gardens. They will produce sizeable vegetables, but this is offset by the actual yield per unit area being smaller than more intensive systems.

BLOCK PLANTING

In this method plants are grown in squares or rectangles rather than in straight rows. Blocks of plants are grown next to each other, for example in dimensions such as five plants by five plants. The numbers of plants grown can be larger or smaller than this. Block plantings use space efficiently, producing many more plants in an area than if grown in rows.

Well-tended soil can support planting that forms a close network of foliage over the soil. The soil will retain moisture extremely well underneath the canopy. It also stops weed seeds from germinating, resulting in a less weedy environment than other growing methods. However, a close canopy of leaves can have disavantages, resulting in poor air flow and high humidity which are ideal for attracting diseases such as botrytis. During dry spells watering is essential for the health of the crops. Double check all watering that has been done because water easily runs off the dense cover of the leaves and never reaches much of the root system.

Blocks can be planted and managed at ground level or in slightly raised beds. They suit modestly sized vegetables, such as root and salad crops, that are usually grown in rows. You may have to be inventive with crop protection for early-sown crops, as cloches and low polythene tunnels do not cover the area well. The blocks are edged with permanent paths for easy access to the centre from any side so you need never walk on freshly cultivated soil. As a result, the soil does not get compacted and the closely grouped plants make maintenance and cultivation easier.

Seeding a block involves sowing the furrows thinly along close rows. For example, instead of the usual 30–40cm (12–16in) distance between carrot rows, this can be reduced to 20–25cm (8–10in). Further space savings can result from growing as much as you can vertically, rather than sprawled over the ground. Wire fencing, netting or poles and tepees take up less space than blocks of climbing vegetables. This will also keep the climbing plants dry and free of disease.

FLAT BEDS

As the name suggests, flat beds are constructed at the natural ground level of your garden. They are the traditional way of cultivating ground for growing vegetables and are most suitable for gardens that have good soil. Flat beds will naturally raise the level of the soil, especially where organic matter is added on a regular basis as part of the cultivation regime, but the bed remains effectively at ground level. This form of gardening is relatively labour-intensive because all of the bed is cultivated, with large areas then being used as paths between the vegetables. With raised beds, there are no paths and only the growing areas are dug.

Left: *The close planting of crops in rows or blocks helps to retain soil moisture as well as reduce the growth of weeds.*

Above: *Plants such as beans can be trained over an arch in order to provide an ornamental and productive feature.*

RAISED AND EDGED BEDS

These can be freestanding garden beds or beds with wooden or brick walls constructed several inches above the normal ground level. Raised bed gardens not only look good but they also can help solve many problems associated with soils that are difficult to manage. Problems with soil are often aggravated in urban and suburban settings, where topsoil and vegetation may have been removed or the surface level changed during building work.

Raised bed gardens improve the environment for plants by lifting their roots above poor soil. The growing medium can be amended by the incorporation of manure and/or garden compost. Soil in these beds warms up earlier in the spring, allowing the seed sown to germinate quicker than if it is grown in flat beds. Beds should be located where they will receive full sun and with protection from prevailing winds. Do not site the beds in frost pockets or where air circulation is poor.

Drainage is important. Build the beds so that the crops will not become waterlogged, as good drainage is especially important when growing vegetables. If the bed contains clay soil, incorporate sand, grit or organic matter to improve drainage. Construct the bed so that it slopes about 2cm (¾in) per metre (yard) of horizontal distance away from any structures, or away from the centre of the bed.

NO-DIG BEDS

As the term suggests, no digging is involved in this method, which is a good way to retain good soil structure within a vegetable garden. Weeding is kept to a minimum and because the soil is not disturbed it will be alive with worms and other organisms. The bed is made on top of the ground and can be built over existing beds, lawns and even hard or rocky ground. Situate the bed in a sunny area that receives morning sun and has good drainage. It can be any size or shape, depending on your space. Start with a small bed – about 2 x 1.5m (6½ x 5ft) – but with a view to expanding in time.

When preparing the site it is not necessary to pull up the lawn or existing garden if the soil conditions are good, but if the ground is very poor, compacted or the drainage is bad, initial digging may be necessary.

Above: *Containers are an ideal way to grow a variety of crops where space is limited. These pots contain courgettes (zucchini).*

The outside wall of a no-dig bed is formed using logs, old planks, tiles, bricks or stones. Line the bed with a layer of plain wet newspaper, at least 6mm (¼in) thick. This layer should cover the enclosed area completely and overlap slightly, so as to kill off any weeds and stop new ones growing. Spread out a thin layer of hay or straw, ensuring that there are no gaps. Place a layer of good organic fertilizer, such as chicken manure, 20mm (¾in) thick, on top. Cover with a 20cm (8in) thick layer of loose bedding straw. Follow this with a 2cm (¾in) layer of good organic fertilizer and complete the bed with a top layer of garden compost, about 10cm (4in) thick. Make sure the bed is watered in well. Once it is settled you can plant out seedlings, but not seeds. Sowing seeds can begin when the bed has matured and the soil has become fine and crumbly.

No-dig beds that have been recently created are best for growing crops such as potatoes, lettuce, brassicas and cucurbits, whereas root crops grow better when the bed is mature. No-dig beds are best suited to planting vegetables in small blocks of different varieties rather than in long rows.

Left: *Vegetable beds need not be rectangular to be functional. This kitchen garden uses simple geometric shapes to provide interesting effects.*

COMPANION PLANTING

This is when two or more crops are grown together for the benefit of one or all. This technique creates a colourful landscape, made up of different species, which mirrors nature itself. Plants are grown together for several reasons, including to attract beneficial insects or to give off odours that deter or confuse pests. This is so different from large fields planted with only one crop, a monoculture that allows large numbers of pests and diseases to build up rapidly.

Above: *Nasturtiums have edible leaves and flowers as well as being a colourful companion plant.*

HOW COMPANION PLANTING WORKS

Plants have natural affinities with others of their kind. The smell of volatile oils from many plants discourages certain pests, making them excellent companion plants. A good example of this is the well-known relationship between the tomato plant and the French marigold (*Tagetes patula*). The scent of the French marigold is said to deter whitefly from entering the greenhouse and therefore avoids a whitefly attack on the tomato plant.

Plants such as yarrow (*Achillea*) and hyssop (*Hyssopus*) are just a couple of plants from a list of many that attract beneficial insects like hoverflies. The hoverflies will lay their eggs around these plants and, after hatching, the larval stage of the insect will start to eat adult aphids. So greedy are these larvae that they can eat up to 800 aphids before pupating. Many organic gardeners grow trays of single-flowered French marigolds to dot around their gardens, both in the vegetable and ornamental areas, in order to encourage these eating machines. This is both very effective and quite stunning to look at.

BENEFITS OF COMPANION PLANTING

A mosaic of plants offers many benefits to the organic garden, some of which are outlined below.

- Creates a colourful landscape.
- Plants can attract benefical insects.
- Plant odour can deter harmful insects.
- A variety of plants together confuses pests from locating host plants.
- Plants attract birds that prey on pests.
- Flowers attract pollinators.

Certain distinct qualities of a plant have a proven benefit to others, such as fixing nitrogen in the soil. Clover in grass will fix nitrogen, offering the excess nitrogen produced to the surrounding grass which improves the yield. By the same token, others are less suited as partners. It is never wise, for example, to plant two vegetables side by side that attract the same pests, as this effectively doubles the chances of attack. It is advisable to practise crop rotation or use companion planting in between them.

BENEFICIAL COMBINATIONS

There is little scientific evidence of these associations working, but if you talk to any organic practitioner, they will certainly provide plenty of anecdotal evidence. Tomatoes, for instance, like to be grown near basil and parsley plants. This is, of course, useful for cooks as well as gardeners. Separating rows of cabbages, broccoli or other brassicas with rows of onions has always been a popular combination, possibly due to the onion's strong scent confusing cabbage pests. Tomato plants also grow well next to cabbages and seem to deter caterpillars. Other beneficial combinations include leeks near carrots as they repel carrot flies, while Swiss chard thrives near carrots and beetroot. Never plant carrot and dill close by each other. This makes the carrots woodier and stronger-flavoured, and the dill milder and with weaker stems.

DECORATIVE COMPANION PLANTS

Certain flowers and flowering herbs offer potential benefits for a variety of vegetables. French marigolds (*Tagetes patula*) are cited as a wonder flower by many organic gardeners and the bright flowers make a colourful companion crop. They deter many pests, and seem to spur growth in roses. They are also said to reduce the number of soil nematodes, while attracting hoverflies and their larvae which eat aphids. French marigolds are frequently planted with pot marigolds (*Calendula officinalis*).

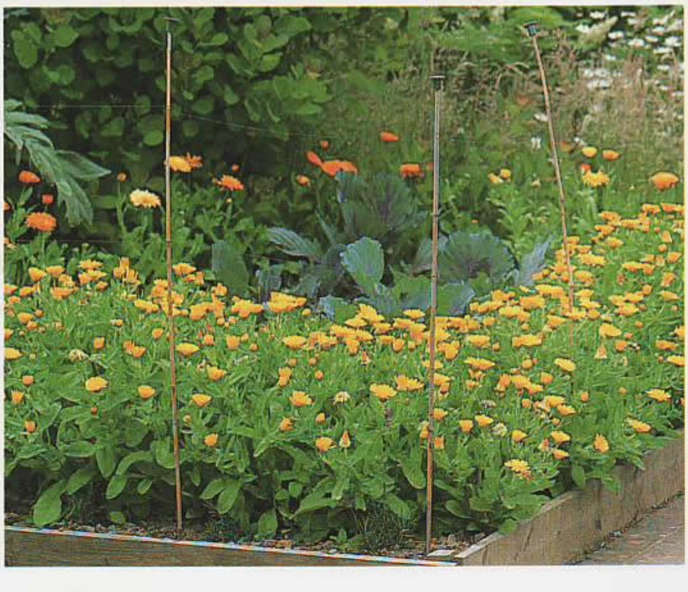

Right: *Companion plants can provide welcome splashes of colour in the kitchen garden.*

CROPS AND THEIR COMPANION PLANTS

While it is not an exact science, any practitioner of companion planting will tell you that individual crops have their "preferred companions". Experience is the best guide, but the list below outlines some plant combinations that work well in most situations.

Apples Chives, foxgloves, wallflowers, nasturtiums, garlic, onions
Apricots Basil, tansy, wormwood
Asparagus Tomatoes, parsley, basil
Beans Carrots, cucumbers, cabbages, lettuce, peas, parsley, cauliflower, spinach, summer savory
Beans (broad/fava) Potatoes, sweetcorn (corn).
Beans (dwarf) Beetroot, potatoes
Beetroot Onions, kohl rabi, lettuce, cabbage, dwarf beans
Brussels sprouts Nasturtiums
Cabbages Beans, beetroot, celery, mint, thyme, sage, onions, rosemary, dill, potatoes, chamomile, oregano, hyssop, wormwood, nasturtiums, tansy, coriander (cilantro)
Carrots Peas, radishes, lettuce, chives, sage, onions, leeks
Cauliflowers Celery, beans, tansy, nasturtium
Celery Tomatoes, dill, beans, leeks, cabbage, cauliflowers
Chives Parsley, apples, carrots, tomatoes
Courgette (zucchini) Nasturtiums
Cucumbers Potatoes (early crop only), beans, celery, lettuce, sweetcorn, Savoy cabbages, sunflowers, nasturtiums
Kohl rabi Beetroot, onions
Garlic Roses, apples, peaches
Grapevines Geraniums, mulberries, hyssop, basil, tansy
Leeks Carrots, celery
Lettuce Carrots, onions, strawberries, beetroot, cabbages, radishes, tagetes
Onions Carrots, beetroot, lettuce, chamomile, kohl rabi, courgettes
Parsnips Peas, potatoes, peppers, beans, radishes, garlic
Peaches Tansy, garlic, basil, wormwood
Peas Potatoes, radishes, carrots, turnips
Potatoes Peas, beans, cabbage, sweetcorn, broad beans, green beans, nasturtium, marigolds, foxgloves, horseradish, aubergine (eggplant)
Pumpkin Sweetcorn
Radishes Lettuces, peas, chervil, nasturtium
Raspberries Tansy
Spinach Strawberries
Squash Sunflowers
Strawberries Borage, lettuce, spinach, sage, pyrethrum
Sunflowers Squash, cucumber
Sweetcorn (corn) Broad beans, potatoes, melons, tomatoes, cucumber, squash, tansy
Tomatoes Asparagus, celery, parsley, basil, carrots, chives, marigolds, foxgloves, garlic, sweetcorn
Turnips Peas, nasturtiums

Left: *Foxgloves (*Digitalis*) make excellent companion plants for growing under apple trees.*

White alyssum (*Lobularia maritima*), by reseeding frequently, helps to break up the soil and adds to its organic content, while chrysanthemums reduce nematodes, making for healthier soil. Mint almost always works with various types of squashes and brassicas to aid plant growth, although it can become invasive. Tansy (*Tanacetum vulgare*) is said to repel ants, aphids and plant beetles although this, too, can become invasive if not regularly checked.

Chamomile (*Chamaemelum nobile*) is known as the "plant doctor" by some organic gardeners because of its alleged ability to encourage other plants to increase their production of essential oil, making plants such as rosemary and lavender taste and smell stronger. Chamomile is easy to grow and looks beautiful anywhere, although it should be kept well trimmed to avoid a straggly look. It is also thought that chamomile can help to activate the composting process if it is added to the compost heap.

Lavender (*Lavandula*) is a general insect repellent and makes an excellent small hedge. It is a great addition to the garden, attracting bees and numerous white and blue butterflies.

Plants that produce berries such as cotoneaster and the rowan tree (*Sorbus*) will attract birds into the garden. Birds, in turn, eat many pests, such as protein-rich aphids, caterpillars and various flies. Thrushes are the unsung heroes of the garden because they decrease the snail population quite considerably.

Be prepared to experiment before committing to a companion species. Nasturtiums have been cited as an effective aphid control, although many wonder if they do this by attracting all the aphids to themselves. What works in one area may not always work elsewhere and experimentation is the key to success in this interesting yet uncertain area of organic gardening.

CROPS AND THEIR ANTAGONISTS

Some plants are highly antagonistic to one another. You should always avoid planting the following combinations.

Asparagus Onion and potato
Beans Chives, fennel or garlic
Carrots Dill
Carrots, cauliflower or potatoes Tomatoes
Peas Onion, garlic and shallots
Potato Pumpkin and summer squash

MAKING THE MOST OF YOUR SPACE

There are three basic growing techniques that you can use to make the most of your space. These are known as intercropping, catch cropping and successional sowing. By adopting the latter two practices, you will extend the cropping season instead of harvesting a crop all at once. You will also make use of the available space, therefore increasing yields. All that is needed is careful planning to work out which crops to sow and when.

Above: *Close planting in blocks makes the most of available space and allows more crops to be raised in a season.*

INTERCROPPING

This kitchen-garden technique increases productivity and also helps to keep the numbers of weeds down. It refers to the practice of planting a fast-growing crop, such as carrots, radishes and lettuce, between main crops that are slower growing. These include vegetables such as cabbages, peas and potatoes.

Intercropping involves harvesting the quicker-growing crop first before the slower-growing one achieves total foliage cover of the soil or shades out the area. A good example of intercropping is to grow a crop of spring onions or lettuce between tomatoes. Similarly, spinach or radishes can easily be planted out early between sweetcorn or, alternatively, radishes can be planted between cabbages.

Intercropping can also be used to increase productivity. It ensures that no space is left unused and makes the most efficient use of light, nutrients and moisture. It will also reduce the amount of weeds in the vegetable patch by maintaining a continuous plant canopy over the soil.

One slight variation on the theme is to combine the benefits of a green manure with a crop. This can be useful in the case of winter crops because the green manure doubles up as a cover crop, protecting the soil from erosion and leaching as well as stabilizing soil temperatures. If a leguminous green manure is planted in late summer or early autumn in a bed along with leafy crops such as Brussels sprouts, it can provide nitrogen throughout the remaining growing season. It will also provide a boost for early crops that will be planted out after the green manure crop has been dug into the soil. While intercropping requires careful planning, it can increase the productivity of even a relatively small vegetable plot.

CATCH CROPPING

This technique is when fast-maturing vegetables, such as radishes and lettuce, are grown in an area of ground that has just been cropped and has a vacancy until the next crop is either sown or planted. The sowing of the catch crops can be done in between the main ones or after harvesting at the end of the season if there is time. It is important to know how long a crop takes to mature when planning catch cropping so that you do not sow anything that takes too long to mature in between the main crops.

Left: *Successional sowing in rows allows the same space in the vegetable plot to be kept productive throughout the growing season.*

ADVANTAGES OF INTERCROPPING AND CATCH CROPPING

Making the most of your space is not only productive, but it also has environmental advantages.

- Suppresses weed growth in the kitchen garden
- Increases the productivity of the vegetable patch
- Planting green manures enriches the soil with nitrogen
- Helps to protect against the erosion of the soil
- Helps prevent leaching

Above: *The close spacing of rows and successional sowing of new rows allows the maximum use of space on your plot.*

FAST-GROWING CROPS SUITABLE FOR INTERCROPPING AND CATCH CROPPING

Use all the available space by sowing vigorous growing crops in between rows of the main crops. The list below shows the time it takes some fast-growing crops to mature.

Carrots	9 to 20 weeks
Endives	7 to 13 weeks
Lettuce	4 to 14 weeks
Radishes	4 to 8 weeks
Rocket (arugula)	3 to 5 weeks
Salad leaves	4 to 14 weeks
Spinach	5 to 10 weeks
Spring onions	8 to 10 weeks

SUCCESSIONAL SOWING

This is the practice of sowing the seeds of fast-maturing vegetables at regular intervals several times during the growing season. This practice will ensure that you have a continuous supply of crops such as lettuces, carrots and spinach throughout the season. Successional sowing is also useful where crops are sown directly outdoors early in the season where they may be prone to frost damage. Early crops such as lettuce and radish can be sown under the cover of, for example, a cloche where they will begin to develop earlier than would otherwise be possible. Subsequent sowings outdoors will mature later, thus extending the growing season for harvesting.

Gardeners with small plots can use this method by sowing only a half a row at any one time. This process is then repeated a week or so later, with further sowings as often as you like. This way you will have fresh vegetables for several weeks and will avoid a sudden glut.

CROPS SUITABLE FOR SUCCESSIONAL SOWING

The technique of successional sowing can ensure a continuous supply of fresh vegetables. It also increases the productivity of the garden and avoids harvest gluts.

Beetroot	Spring to summer
Broccoli	Spring to summer
Cabbages	Spring to summer
Carrots	Spring to summer
Endives	Spring to summer
Kohl rabi	Winter to summer
Lettuce	Winter to summer
Parsnips	Winter to spring
Peas	All year round
Potatoes	Spring
Radishes	Winter to summer
Salad leaves	Spring to autumn
Salad rocket	Spring to autumn
Spinach	Spring to summer
Spring onions (scallions)	Spring to autumn
Swedes	Spring to summer
Swiss chard	Spring to summer
Turnips	Winter to spring

Above: *Sweetcorn (corn) is a late-maturing crop that can be intercropped with fast-growing, early salad crops. This enables you to make the most of your growing space.*

SOWING IN THE OPEN

There can be nothing more satisfying than sowing seeds and watching in anticipation for them to germinate. Watching the seedlings grow in your kitchen garden, making it come alive with leaves, flowers and insects is a fantastic experience. One of the drawbacks of sowing out in the open is that you are at the mercy of the weather, but you can manipulate your garden environment by using cloches, small plastic bottles and polythene tunnels to increase your chances of success.

Above: *A dibber (dibble) is a very helpful tool to use when you are transplanting seedlings outside.*

SITE REQUIREMENTS

If you are a novice kitchen gardener, be assured that sowing seed is easier than you may think. In order to germinate successfully seeds need water, air, a suitable temperature and a place into which they can root in order to support the top growth that will follow. For this reason, good soil preparation is everything to organic success. Most garden soils are able to supply all the necessary elements and only slight modifications are usually necessary. However, it is important that the ground is prepared in a way that will enable the seed to germinate evenly and grow in a uniform environment. The prior preparation of a seedbed can provide the seed with the environment that it needs and has the added bonus of making the task of seed sowing easier.

Preparing a seedbed is simple. Following cultivation, the ground is levelled using a rake, held at a shallow angle, to break down any large clods. The art of levelling is to keep the rake angle shallow and move the high spots over into the low spots with even strokes. Hold the rake firmly at the rear and let the shaft run smoothly through the front hand. All stones and large objects, including organic matter such as twigs or previous crop debris, should be removed by combing them out with the teeth of the rake, while holding the tool in a near vertical position. The soil is then firmed with light treading. A light shuffle across the bed is best. Once firmed, lightly rake the soil again at a shallow angle to produce a light "fluffy" surface that runs freely through the teeth of the rake. This is the perfect environment for sowing and growing seeds.

SEED REQUIREMENTS

The conditions that seeds require to germinate are easy and straightforward to create. A well-aerated moist soil environment is almost all that the seed needs. Most seeds will germinate quite successfully once the temperature gets above 7°C (45°F). Seeds carry their own food supply that provides them with everything they need for those first crucial days following germination. Once the plant begins to establish and grow, it needs soil-borne nutrients. This means that the soil in which it is growing needs to be of the right fertility for the plant. Poorer soils can benefit from the addition of a base dressing with fish, blood and bone prior to sowing to give the boost that the developing plant needs, although loamier soils need only be properly dug and prepared to support germination.

SOWING SEEDS OUTDOORS

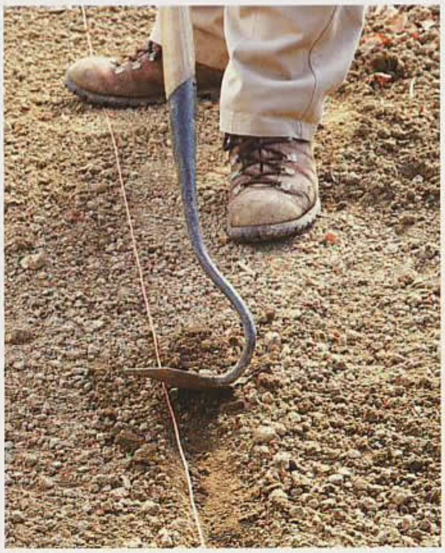

1 *Set a tight string line where you intend your crop row to be and make a shallow drill with the edge of a swan-necked hoe.*

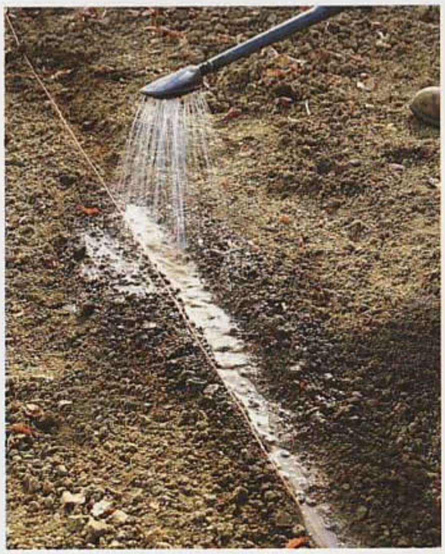

2 *Water dry soil using a watering can fitted with a fine rose about an hour before sowing and allow to drain.*

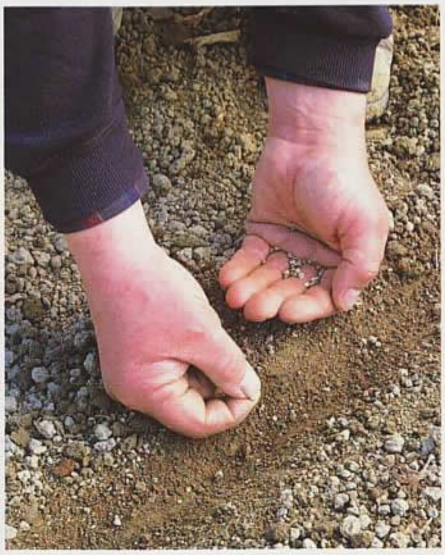

3 *Sow the seed thinly along the length of the row. Larger seed may be station sown at regular intervals.*

Some seeds can benefit from being soaked for a short period in tepid water just before planting to help them take on the water they need for germination. This is especially true of beets, but other large seeds also benefit from this treatment.

SOWING IN ROWS

Seed is usually sown in rows. Using a tightly drawn garden line as a guide, draw the corner of a swan-necked hoe along the line to create a shallow drill of about 1–2cm (⅜–¾in) depth, depending upon the seed's individual requirements. Dry ground can be watered after the drill is made. Sow the seed thinly and mix fine seed with silver sand to make it easier to distribute evenly. Mark the end of each row with a label before moving on to the next.

STATION SOWING

Seeds of larger growing plants, particularly those with seed that is large enough to handle, benefit from station sowing. This involves sowing two or three seeds at intervals that will be the eventual crop spacing. If all three germinate, then the two weaker ones are removed or transplanted to gaps where none has germinated.

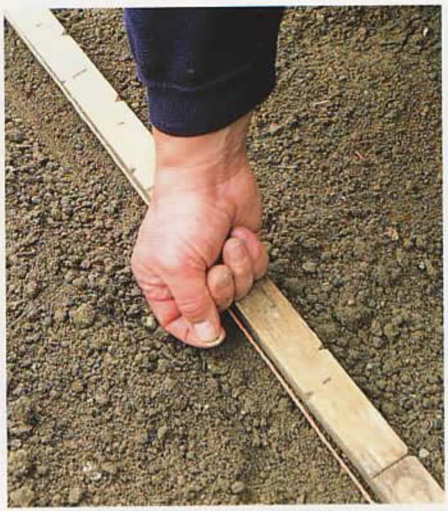

Left: *Station sowing is ideal for larger seeds or seedlings that resent disturbance. Use a marked stick to set the spacing of the seed.*

WIDE ROWS

Certain seeds, particularly peas and beans, benefit from being planted in wide rows. Two rows are effectively station sown at once, one on each side of a drill that is 15cm (6in) across. The drill is made with the flat of the hoe and after sowing the soil is carefully raked back. Make sure that you do not disturb the seeds from their stations.

BROADCASTING (BLOCKS)

Broadcasting is an ancient method of seed sowing that was used to sow large areas of crops. It involves a "broadcasting" action that separates the seed to an even spacing. The easiest way to do this in small vegetable plots is to split the seed into two halves, mixing small seed with fine sand. Scatter the seed carefully, letting it run from your hand in even arcs as you move your arm from side to side. Sow each half of the seed at a 90-degree angle to the other, thereby assuring an even distribution. Gently rake the seed in once sown and lightly water if needed.

PROTECTING SEEDLINGS

Seedbeds, with their fine "fluffed" earth, act as magnets to birds and animals. Some may take the seed from the ground, but, in truth, most will find it more attractive as a dust bath or litter tray. Once the seedlings emerge, however, some birds find them irresistible. They must be kept out with some form of barrier. One of the easiest methods is to form a low tunnel of chicken wire, supporting this on hoops. For larger areas, a series of stakes in the ground can be covered with netting to keep birds at bay. Sticks with thread or string stretched between them are also effective, but less easily removed for you to tend the crop.

LABELLING

When you sow a row or area of seeds, label it straightaway. Re-usable plastic labels are the best option. Each label should have the name and variety of the plant sown. You may also wish to record the sowing date and if the seed was pre-treated.

4 *Place a label at the end of each row, showing the crop name, variety and the sowing date, before you start a new one.*

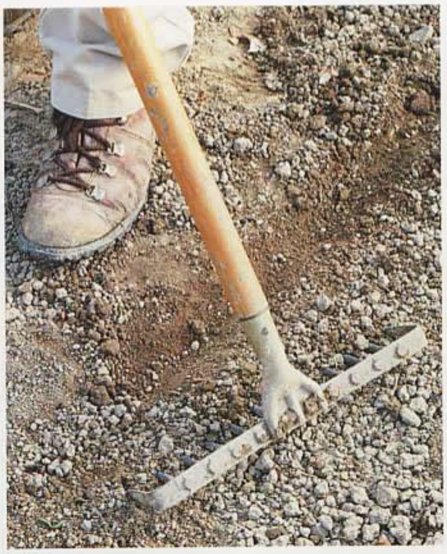

5 *Gently cover the seed using a rake. Take care not to disturb or move the seed in the row when you cover it.*

6 *Alternatively, seed can be covered using a soil and potting mix. This is useful on heavy soils and avoids capping (surface hardening).*

SOWING UNDER GLASS

Sowing seed under glass extends the growing season and enables you to raise tender crops that only survive outside in warmer months. It is also ideal for rapidly establishing plants to use as catch crops and as early companion plants. In short, it offers variety and choice for your cropping regimes. Plants raised under glass can be grown on until they reach a size where they are better able to resist pest attack. They can also be planted out at their final spacing, thereby avoiding thinning or gaps in rows.

Above: *A propagator is a useful piece of equipment for raising your seed in ideal growing conditions.*

HEAT REQUIREMENTS

Most seeds have a preferred temperature range within which they will grow. A heated greenhouse, a conservatory or even a warm living room will often provide this, although seeds needing a constant high temperature may need a propagator. Such seeds are mercifully rare among vegetables, although cucumbers, tomatoes and peppers are good examples of crops that will benefit if they are first started off in a propagator.

CHOICE OF CONTAINER

There are a variety of containers that may be used to sow seed under glass. The most common form is the plastic seed tray which has now largely replaced the wooden seed tray. Although more attractive, wooden trays are difficult to keep clean and may harbour plant diseases. Plastic trays can be made of durable polyurethane or sometimes a more flimsy, thin, moulded plastic which is intended for single use. They are available in a variety of sizes, although small 9cm (3½in) pots may be more suitable when you are raising only a few plants.

Modular seed trays are another option. These are made up of individual cells, and a single seed is sown into each one. Seeds that are sown in these trays have the advantage of not suffering any root disturbance when they are planted out in their eventual position. The same is true of some biodegradable pots, which are formed out of paper and coir. These are better (environmentally speaking) than those made from peat, but any recycled material used for these trays is acceptable in the organic kitchen garden.

Inventive recycling can also provide an array of useful sowing containers. Re-using plastic cups, vegetable packing trays and any other throw-away items that might otherwise end up on a landfill site are all possibilities worth exploring. Old plastic bottles make good individual propagation cowls for small pots and plastic bags can also be used to cover the tops of pots and trays in order to maintain humidity.

Left: *Many propagators have ventilators on the top in order to aid the flow of fresh air to the seedlings.*

PROPAGATORS

These are, in effect, mini-hothouses that help to keep the seed in a warm, moist, stable environment both above and below the soil line. Expensive propagators involve the use of electric soil-warming cables and some have thermostats to control the soil temperature. Many are designed for use in the greenhouse, but some models are narrow enough for use on a windowsill inside the house. Alternatively, instead of buying the whole propagator, you can purchase a heating mat on which the seed trays can stand in order to receive heat at the bottom. This system has the advantage of being mobile and easily moved about from area to area. Another cheaper method is to buy soil-heating cables to bed into sand. The seed trays sit on the bed of heated sand to receive an even supply of bottom heat.

AFTERCARE OF SEEDLINGS

Once seeds begin to germinate, they can be moved gradually into a less humid environment. The trays are freed of plastic covers, or the propagator vents or lids are opened and, after a few days, the lid is removed completely. As soon as the seedlings reach a size at which they can be handled, they are carefully pricked out into individual pots or larger boxes and trays. Always handle the seedlings by the leaves,

Above: *Modular trays are ideal for planting large seeds and avoid disturbance to roots when planting out or potting on.*

Above: *Individual pots can be used for large seeds. Plant two seeds per pot and remove the weaker seedling.*

SOWING IN TRAYS

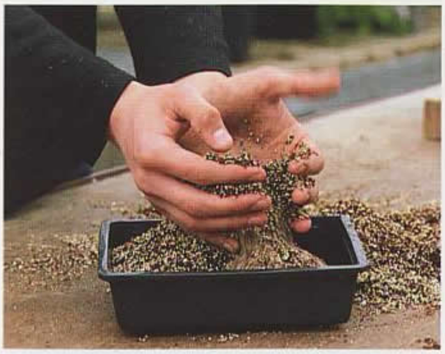

1 Fill a seed tray with propagation potting mix that has been thoroughly mixed. Fill the tray to overflowing and do not firm in the potting mix.

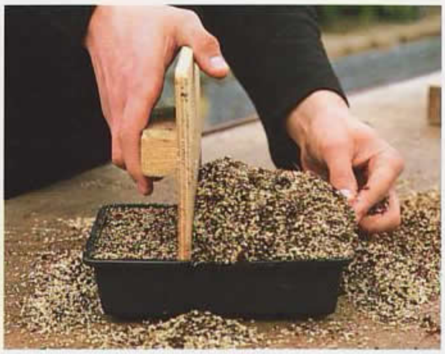

2 Using a straightedged piece of board, level the surface of the potting mix by carefully moving the board across the top of the tray.

3 Water the tray and leave to drain for about 20 minutes before sowing the seed on the moist surface of the potting mix.

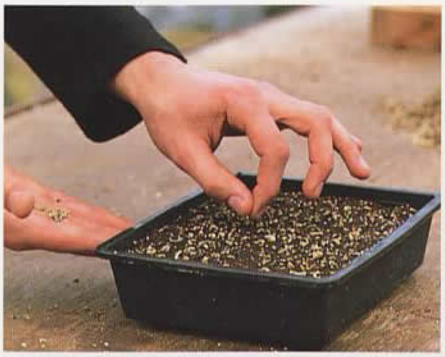

4 Larger seed can be placed on the surface of the potting mix at regular intervals. Regular spacing will prevent overcrowding.

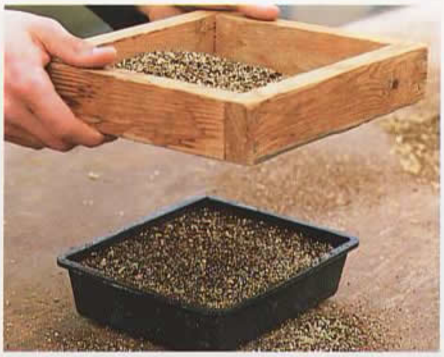

5 Once the seed has been sown, use a sieve to cover the surface with a fine layer of potting mix.

6 Do not cover the seeds of plants that require light for germination. Always check the growing requirements of plants before sowing the seed.

---

gently lifting each one from beneath, using a dibber (dibble). Never hold them by the stems because this can cause a great deal of damage to the developing plant. They should be spaced at least 5cm (2in) apart to allow for subsequent development. Water the transplanted seedlings with a fine upturned rose attached to a watering can that has been filled with water overnight, thus bringing it up to room temperature and not giving the transplants a shock of cold water. Keep the seedlings on a warm and sheltered windowsill or in a greenhouse or conservatory. A constant temperature will promote healthy growth.

Above: *Growing vegetables in pots under glass can extend the growing season and also provide a colourful greenhouse display.*

HARDENING OFF

Plants that have been grown in a greenhouse cannot be put straight outside because their growth is too soft to withstand the cold. They must be gradually hardened off and acclimatized to outdoor conditions. The young plants are hardened off by moving them from the greenhouse and into a cold frame, planted out under cloches, low polythene tunnels or horticultural fleece. Open up the cloches and cold frames or remove the fleece during the day and replace at night for a week or two before planting out or removing the cloches completely. Remove low polythene tunnels after a couple of weeks depending on the weather conditions. If these are severe, leave in place for longer.

Above: *Watering from below by pouring the water into the base of the seed tray will prevent damage to newly emerging seed.*

THINNING AND TRANSPLANTING

Thinning seedlings not only allows the plants left in the soil to develop and mature into their natural shape and habit, but it is an essential process for good plant health. Thinned seedlings can be transplanted into bare areas or more often special seedbeds are set up or crops are grown under glass and the whole batch of seedlings are transplanted into their final destination within the vegetable garden.

Above: *Growing plants from seed is one of the most satisfying ways of producing new stock for the kitchen garden.*

WHY THIN SEEDLINGS?

All plants need space to develop and grow. Plants that are growing too closely compete not only for space, but also for light, water and important nutrients. In addition to this competitive stress, they also become prone to a variety of fungal diseases, as the air is not able to move around them. Thinning the seedlings helps to counteract these problems and will result in larger, stronger and healthier plants.

HOW TO THIN SEEDLINGS

Thinning is essentially two processes in one. Firstly, you are removing all the plants that are excess to requirements and at the same time you are selecting the biggest, healthiest and strongest plants that will be retained to form the crop.

Before starting to thin, dry ground must be watered, preferably the night before you intend to thin. A measuring stick, marked with the appropriate crop distance, can be used to show the approximate position of the individual plants. Remove all the plants in between each of the markers, selecting the healthiest plant at or near the mark on the stick. If there are no plants at the marked point, then you can transplant one of the seedlings that is excess to requirements into this position. When you are removing the excess plants, place a finger on the soil at either side of the seedling that is being kept. This protects it from root disturbance. Once you have finished, water the remaining seedlings with a fine rose on a watering can to re-firm the soil around the plants. The seedlings that have been removed can be put on the compost heap.

Above: *Rows of seedlings are often too crowded and need to be thinned when the plants are young.*

Left: *Seedlings grown elsewhere in the garden can be lifted and replanted in their eventual positions. This method is good for crops that need wide spacing.*

Avoid thinning with this method on hot dry days or in windy conditions, as the remaining seedlings may become water-stressed if their roots are disturbed. The alternative, in this case, is to snip off the seedlings at ground level with a pair of sharp scissors, thereby avoiding root disturbance.

THINNING DISTANCES

The following measurements are the distances that need to be left between thinned seedlings.

Beetroot	7.5–10cm (3–4in)
Broad beans (fava beans)	23cm (9in)
Carrots	7.5cm (3in)
Dwarf French beans (bush green beans)	20cm (8in)
Florence fennel	25cm (10in)
Kohl rabi	20cm (8in)
Lettuce	23cm (9in)
Parsley	15cm (6in)
Peas	5cm (2in)
Parsnips	15–20cm (6–9in)
Radishes	2.5–5cm (1–2in)
Runner beans	25–30cm (10–12in)
Salsify	15cm (6in)
Scorzonera	15cm (6in)
Spinach	15cm (6in)
Spring onions (scallions)	5cm (2in)
Swedes (rutabaga or yellow turnip)	30cm (12in)
Swiss chard	30cm (12in)
Turnips	15–20cm (6–8in)

TRANSPLANTING SEEDLINGS

1 *Water your row of seedlings at least an hour before transplanting and preferably the night before if no rain has fallen.*

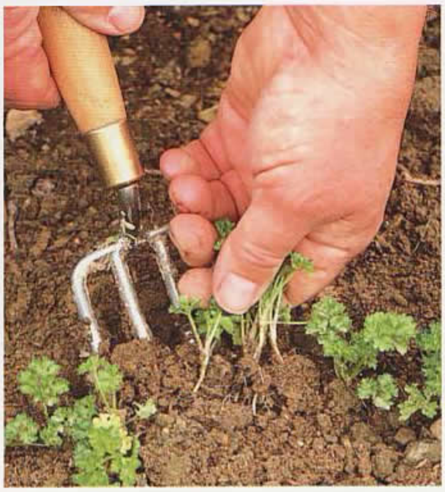

2 *Using a fork to loosen the soil, gently lift the seedlings. Take care to handle them by their leaves and never touch the stems.*

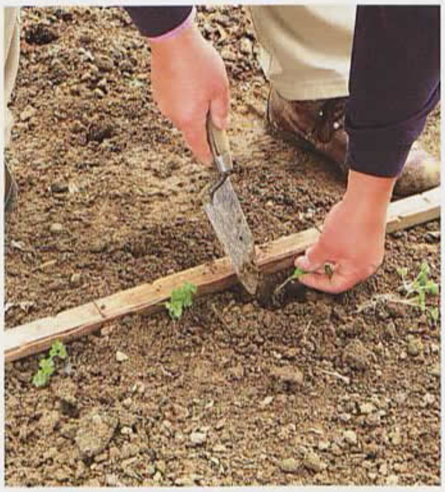

3 *Using a tight line, straight edge or notched planting board, replant the seedlings at the appropriate spacing for that crop.*

4 *Gently water the seedlings immediately after sowing. Never let roots dry out at any stage during transplanting.*

COMMON PLANTING DISTANCES

Always allow the correct spacing between crop plants so that they will grow into healthy specimens.

Asparagus	30–38cm (12–15in)
Aubergines (eggplant)	60cm (24in)
Broccoli	60cm (24in)
Brussels sprouts	50–75cm (20–30in)
Cabbages	30–50cm (12–20in)
Calabrese (Italian sprouting broccoli)	15–23cm (6–9in)
Cauliflowers	50–75cm (20–30in)
Celeriac	30–38cm (12–15in)
Celery	23–30cm (9–12in)
Courgettes (zucchini)	60cm (24in)
Cucumbers	60cm (24in)
Garlic	15cm (6in)
Globe artichokes	75cm (30in)
Jerusalem artichokes	30cm (12in)
Kale	60cm (24in)
Leeks	15cm (6in)
Marrows (zucchini)	60cm (24in)
Onion sets	10cm (4in)
Peppers	45–60cm (18–24in)
Potatoes	30–38cm (12–15in)
Pumpkins	90–180cm (3–6ft)
Rhubarb	75–90cm (30–36in)
Runner beans	25–30cm (10–12in)
Seakale	30cm (12in)
Shallots	15–18cm (6–7in)
Sweetcorn (corn)	30cm (12in)
Tomatoes	60cm (24in)

TRANSPLANTING SEEDLINGS

The most common way of transplanting seedlings involves planting container-grown plants into open ground. Early vegetable crops can easily be raised in this way. The other method is to raise seedlings in open ground near to where they are to be planted out. Transplanting outdoor seedlings means that a smaller area of the vegetable plot is needed for sowing and, in consequence, a smaller seedbed is required. Transplanting seedlings is a good way of growing plants such as lettuce that are to be used in catch-cropping beds or where plants are to be planted out in no-dig beds.

The ideal time for transplanting outdoor seedlings is during damp overcast weather because this helps to prevent the seedlings' roots drying out. As is the case when thinning the plants, the seedlings will need watering the evening before. It is best to dig up only a few plants at a time, discarding any that are weak, damaged or appear to be sick. Seedlings can be placed in a plastic bag to maintain humidity around them while they are out of the ground.

A garden line can be set out in the vegetable patch in a similar manner to the way in which it is placed for preparing a seed drill. The position of the plants can then be determined using a measuring stick. Use a dibber (dibble) or thin trowel to plant the seedlings, firming lightly around the base before moving on to the next transplant. Once the row is completed, it is important that you water them in.

For catch cropping or planting among other plants, the surrounding crop may well determine the spacings between the transplants, although the procedure remains exactly the same.

AFTERCARE

The aftercare that your crops require is essential to ensure healthy plants that produce high yields. There are a number of different techniques that are listed below to help you on your way. Just remember that the more effort you put in then the better they will taste. As well as judicious watering and feeding, you will also have to weed the vegetable plot and provide some form of plant protection and support. But the reward of such care and attention is delicious home-grown produce.

Above: *Regular cropping of vegetables such as this ruby chard will encourage the development of new growth.*

WATERING

Vegetables need watering if there is no significant rainfall. This is especially true when the crops are young and only have shallow roots. Having a soil that is rich in organic matter is an advantage because it conserves water in the soil, particularly if the soil is sandy. You can enrich the soil by digging in organic matter, growing green manures or by continuous mulching.

If you have a large garden, mulching around the crop with straw, farmyard manure, garden compost or another similar substance can help to limit water loss. When you water, it is best to give the plants a thorough soaking, allowing the water to penetrate deeply. Giving a light watering will just encourage shallow surface roots that are easily damaged by prolonged drought. Prolonged dry periods also encourage plants to bolt, thus ruining the crop.

Certain plants such as cabbages and lettuces are naturally shallow rooted, whereas others such as tomatoes and squash are naturally deep rooted. Shallow rooters especially benefit from mulching and may need watering often in hot summers.

FEEDING

This is generally only necessary for heavier feeding plants such as pumpkins and marrows, but can provide a tonic for any crop during the growing season. Specific nutrient deficiencies are rare in the garden if a regular crop rotation and soil enrichment programme is followed, but if specific shortages are noticed, they must be countered immediately. Good liquid feeds can be made in your back garden. Try making compost tea or "worm liquid", as these are an excellent choice for actively growing plants.

Above: *The careful harvesting of crops is essential if they are to store well. Here, carrots are being lifted with a garden fork.*

Above: *The regular harvesting of peas and beans will help to extend the length of the harvest period.*

Above: *The effort you put into protecting and caring for your crops will be rewarded at harvest time.*

Apply these feeds once every week or two. Many plants are given their main feed annually. This type of feeding is normally applied to fruit trees and bushes in the spring and is best applied in the form of a slower-release fertilizer such as bonemeal.

WEEDING

Check your plants regularly and try to keep them as weed-free as possible. Remember that weeds compete with your plants for light, space, water and nutrients. It is better to weed a little and often than to weed irregularly. Even areas that have been mulched or have close cropping regimes need regular checks and weeds removed.

PLANT PROTECTION

Pests and diseases come in many different forms. Over time you will learn which specific problems your garden is prone to. Keep a yearly notebook of all the things that affect your crop each season, noting what control measures you used and how effective they were. In time you will have a record of how and when to manage pests and other problems in your own garden.

Above: *Taller plants and climbers will need some form of support. Canes or sticks with twine tied between them provide ideal temporary supports.*

PROVIDING SUPPORT

Plants may need supporting, especially climbers such as peas and beans. This support must begin at an early age, using twiggy pea sticks or hazel sticks, both of which can look ornate. Bamboo canes can be used for a host of climbing plants, but it is more ethically correct to use local materials rather than to use a product that has been transported thousands of miles. Squashes can be grown up fancy supports made from stakes and strong string or rope, making a curious garden feature. Other plants such as globe artichokes and Florence fennel can benefit from individual supports for their stems.

Above: *Marrows and pumpkins are easily stored in trays that should then be kept in a cool dark place.*

HARVESTING AND STORING

When harvesting, it is tempting to pick the good specimens and leave the rest, but remember to pick and compost poor or rotten ones as well as the best of the crop. Diseased or damaged crops can begin to rot and may infect the remaining harvest.

Freshly picked vegetables are higher in nutritional value than stored ones. So, during the growing season, pick what you need and store the surplus. Choosing when to pick a certain vegetable is a matter of taste. Some prefer to pick produce such as beans when they are young, small and succulent, whereas others prefer the beans to be more mature.

Crops can be stored in a number of ways. Root crops can be left in the ground in all but the coldest winters, lifting them as they are needed. Some crops, such as parsnips, actually develop a better taste after they have been subjected to frost. Alternatively, root crops can be stored in a cool dark place. They should be cleaned and stored in sand and a sterile organic substance such as coir, untreated sawdust, fine leaf mould, fine bark or sterilized soil.

Store other vegetables on shelves or boxes, ensuring that they do not touch. Onions and garlic are best kept in open sacks or in strings. Cabbages can be kept in nets until needed for up to two to three months. Other brassicas such as Brussels sprouts and swedes are best kept outside in the ground, harvesting them as needed.

Freezing is another storage option. Some vegetables, such as asparagus, beans and cauliflowers, are best blanched in boiling water for a couple of minutes before freezing. Alternatively, you can cook the vegetables and then freeze them, as is the case for marrows and Jerusalem artichokes.

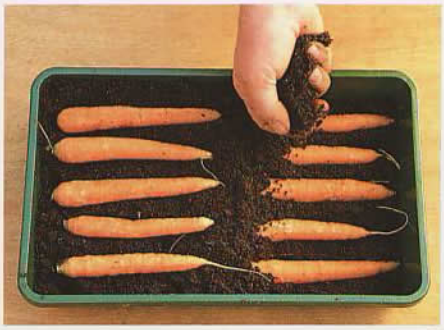

Above: *Store root crops such as carrots in a cool dark place and cover them in sterile sawdust, leaf mould or coir.*

CROPS FOR STORING OVER WINTER

Harvest time often brings the problem of a crop glut. Storage can enable you to enjoy this bounty for longer.

Asparagus Best consumed fresh. Cook or blanch before freezing
Aubergines (eggplant) Cook before freezing
Beetroot (beets) Shelf storage or pickling
Broad (fava) beans Freezing or drying
Brussels sprouts Freezing. Leave on plant until needed
Cabbages Freezing or shelf storage. Leave in ground in mild conditions
Carrots Leave in ground in mild conditions. Shelf storage
Cauliflowers Blanch before freezing. Shelf storage if hung upside down in dark and misted
Celery Cook before freezing. Limited shelf storage
Courgettes (zucchini) Cook before freezing. Limited shelf storage
Dwarf French beans (bush green beans) Freezing
Garlic In sacks or strung
Kale Harvest through winter
Kohl rabi Leave in ground if mild or protected, medium shelf storage
Leeks Freezing, leave in ground in mild conditions, shelf storage
Marrows (zucchini) Cook before freezing, shelf storage
Onions In sacks or strung
Parsnips Leave in ground until needed or late winter
Peas Freezing or drying on plant
Peppers Blanch before freezing. Pickle or dry
Potatoes Store in paper sacks once cleaned and dried
Pumpkins Cook before freezing. Good shelf storage if fully ripe
Rhubarb Cook before freezing
Runner beans Freezing
Shallots In sacks or strung
Spinach Cook then freeze
Swedes (rutabaga or yellow turnip) Leave in ground until needed (may go woody by late winter)
Sweetcorn (corn) Freezing or pickling
Tomatoes Cook before freezing. Pickle
Turnips Leave in ground until needed

GROWING HERBS

Herbs are valued for their culinary, medicinal, decorative and aromatic properties. They come in a diverse range of sizes, shapes and habits, ranging from ground-creeping thyme through to the tall architectural stems of angelica. The choice of herbs is so great that there is always something to offer a gardener with only a window box, hanging basket or a small space. A herb garden offers a treat for the senses and these fragrant plants are also easy to grow in the organic garden.

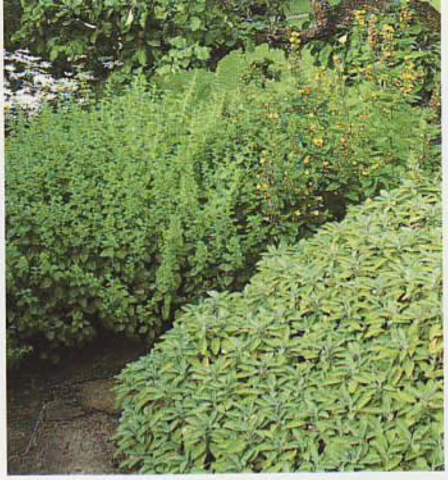

Above: *Herbs are easy to grow given the right growing conditions. They provide an attractive, often fragrant, display.*

WHERE TO GROW HERBS

Herbs can be grown in a range of settings, such as custom-designed herb gardens and ornamental borders. They can also be grown as companion plants in the vegetable garden and are eminently suited to growing in containers, hanging baskets and window boxes. They are especially useful if grown near the kitchen.

Herbs range from tall showy herbaceous plants such as fennel (*Foeniculum vulgare*) and tansy (*Tanacetum vulgare*) to ground-hugging cushion plants such as thyme (*Thymus vulgaris*). The majority of herbs originate from dry sunny environments and so need sunshine to help them develop their essential oils. It is best to site herbs in an open, sunny spot in the garden where they will thrive.

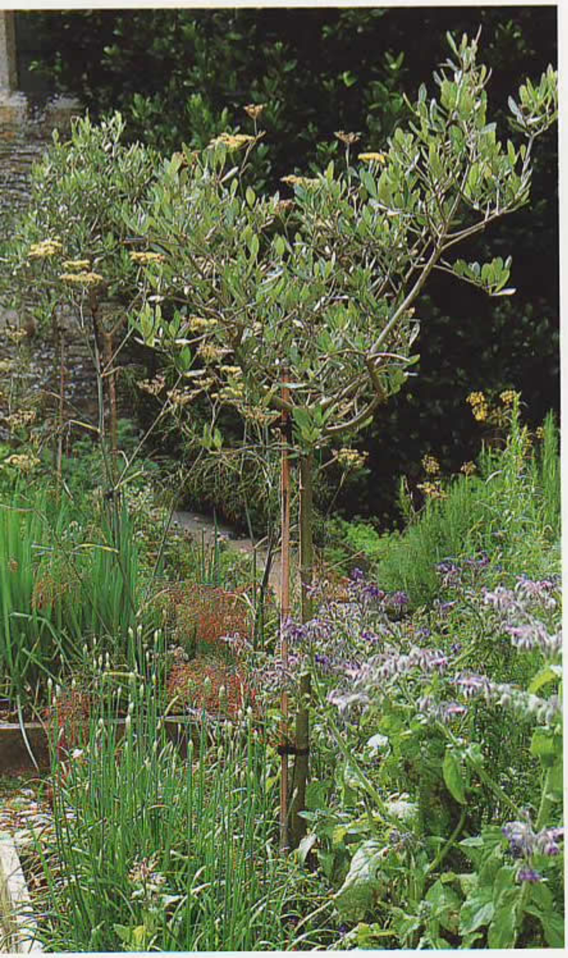

SOIL PREPARATION

Drier sites suit most herbs, and the sunnier and hotter the site the better they will taste. The taste and smell of herbs is usually due to the production of essential oils within the plants. If they are grown in hot conditions, then the concentrations of essential oils will be greater. Growing herbs in very moist rich soils can accelerate their growth, but will result in a milder flavour. They will also look better and flower less than their "hot-site" counterparts and be easier to harvest.

Herbs are, however, best grown in a soil that is loamy with some added organic matter. The ideal pH is 6.5 to 7.0, which means that herbs can easily be planted among vegetables in the kitchen garden.

SOWING HERB SEED

Herbs may be sown directly in the soil outdoors, just like vegetables. The preparation of the seedbed and the sowing techniques are exactly the same, and they can easily be interplanted or block planted among other vegetables.

Alternatively, herb seeds may be planted under cover, raising them in the same way as early vegetables and bedding plants and then hardening them off before planting out in the garden. This method is especially useful for more tender, leafy herbs such as basil or coriander (cilantro).

Left: *Herbs are an important element in both ornamental and kitchen gardens, being decorative as well as useful.*

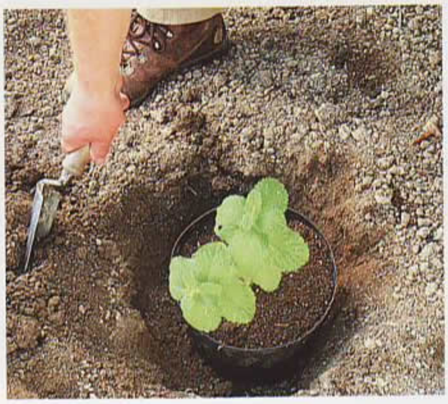

Above: *You can restrict the spread of invasive plants such as mint (*Mentha*) by planting them in pots or buckets that are then buried in the ground.*

Above: *Pinching out newly planted herbs such as this bay (*Laurus*) will encourage bushy, leafy growth.*

Basil (*Ocimum basilicum*), for example, associates well with outdoor tomatoes, both in the garden and in the kitchen. These can be planted out together from an indoor sowing and the crop will be mutually complementary. Indoor sowing can also provide you with herbs that can be cropped earlier in the season, effectively extending the useful life of your organic herb garden.

PLANTING HERBS

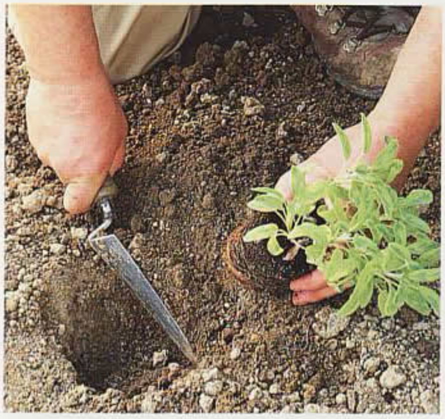

1 *Excavate a planting hole that is about a quarter to half as big again as the plant's rootball. Make sure it is deep enough.*

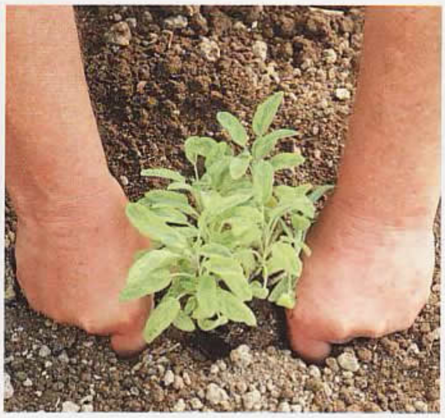

2 *Ensure that the top of the compost is the same level as the surrounding soil. Refill the gap around the rootball and firm the soil.*

3 *Water the plant immediately after planting. Keep the plant well watered until it is fully established into the surrounding soil.*

PLANTING HERBS IN THE GARDEN

Herbs can be sited anywhere in the garden as long as it is sunny. They have a range of forms and colours and often make valuable additions to the ornamental garden. Foxgloves (*Digitalis*), sage (*Salvia officinalis*) and the curry plant (*Helichrysum italicum*) are a few examples of herbs that can be used in annual and herbaceous borders as well as in the kitchen garden. Some herbs, such as mint (*Mentha*) and lemon balm (*Melissa officinalis*), can become very invasive if they are not contained in a pot or sunken sink when growing among other plants in an ornamental border. Remove the flower-heads from the mint before they have had a chance to seed, as the seed will germinate all over your border.

PLANTING HERBS IN CONTAINERS

Herbs make excellent subjects for use in pots and containers and are wonderful for patio gardens that catch plenty of summer sun, although you need to make sure that the potting mix never dries out. Raised beds, which provide good drainage, are also good areas for growing herbs. Always plant them in a free-draining potting mix that will not become waterlogged. There are numerous cultivars of culinary herbs that can be used for ornamental purposes, and groups of pot-grown herbs can be extremely decorative as well as supplying you with a range of fresh flavourings. Thyme, rosemary, lavender and sage are all good choices to grow in pots, either outdoors or on your windowsill.

Above: *Herbs are easily grown in a container which will provide an excellent focal point in the kitchen garden.*

Herb pots require little maintenance, save for watering and the occasional feed during the growing season. Most are rarely long term and are best restarted yearly or every other year. Herbs that are permanently in pots, such as bay trees, will need repotting every year. Spreading subjects like thyme can be lifted and both top-pruned and root-pruned prior to repotting.

Mint and other spreading herbaceous subjects may need dividing and repotting from time to time. This is done by splitting the crown of the plant into smaller pieces and then repotting one of these back into the container with some fresh potting mix. Dividing in this way is best performed annually for very vigorous herbs. Harvesting on a regular basis is often enough to control the growth of many potted herb arrangements (where a number are planted in one container), but a light trim may also be necessary from time to time.

Potted herbs may also be grown in the greenhouse to ensure a supply both earlier and later in the season. Try growing basil (*Ocimum basilicum*), coriander (cilantro; *Coriandrum sativum*), chives (*Allium schoenoprasum*) and dill (*Anethum graveolens*) in pots because this will save money if you spend a lot on fresh herbs. Pots raised under glass can be brought into the kitchen for ease of use. A series of successional sowings under glass will ensure that you have fresh herbs for most or even all of the year, both indoors and out.

Above: *Planting spreading herbs such as thyme (*Thymus*) in pots will help to control them and can also be decorative.*

HARVESTING AND STORING HERBS

Many commonly used culinary herbs such as basil, coriander (cilantro), chives and parsley can be grown indoors on a windowsill or conservatory during the cooler winter months if you have space. If you are not able to spare growing room for these, most are easily stored for use later. Herbs are easily harvested and most of the storage techniques are simple and straightforward.

HOW TO HARVEST HERBS

Various parts of herbs, including the leaves, flowers, fruits and seeds, are gathered at different times, depending upon the plant and the part that provides the desirable properties. Annual leafy herbs such as basil (*Ocimum basilicum*) and parsley (*Petroselinum crispum*) should be carefully picked, never taking more than about 10 per cent of the growth in a single picking.

The same is true of perennials such as sage (*Salvia officinalis*), thyme (*Thymus vulgaris*) and rosemary (*Rosmarinus officinalis*) because severe pruning or over-stripping of the leaves will weaken the plant. It is important that you not remove more than one-third of the growth at any one time. If you harvest carefully, you will get a more vigorous leaf growth that will result in healthier plants.

Above: *The best time to harvest herbs is usually just before they flower. This is when they have their strongest flavour.*

As a general rule, pick herbs just before the plant is about to flower, which is when they have the strongest flavour. Pick the leaves when they are fresh and at their sweetest, selecting blemish-free, upper leaves. Collect the leaves in early morning or late evening, provided they are dry, rather than in bright afternoon sun when the plant's sap is rising. This is when the aroma of herbs is at its strongest and is easily lost if picked during this time of day. Flowers such as borage (*Borago officinalis*) and lavender (*Lavandula*), however, are best picked just before they reach full bloom and once they begin to open in the heat of the day.

Rhizomes, like ginger and turmeric, are collected in autumn, just as the leaves begin to change colour and the maximum amount of nutrition has been stored. Use a fork to gently free the roots from the soil and always avoid "hand-pulling" them. Choose only the best ones and use a vegetable brush to gently loosen any dirt. If you do need to wash them, do so quickly in cold water and avoid soaking them, as this can result in lost flavour.

Harvesting seeds tends to vary from plant to plant. Some seeds, like those of borage, simply fall to the ground as soon as they are ripe. Thyme seeds are very small and hard to see. Parsley and coriander seeds shake off very easily, and frequently the plants will have sown next year's crop for you before you realize they have gone to seed. One method of harvesting any seed that is difficult to collect is to tie a small paper bag over the flower head when the seeds start to form, ensuring that you can collect the seed without losing any. It would be advisable to use this method for collecting from plants with small seed as they can drop off when ripe or sometimes spring from the plant.

Above: *Aromatic herbs such as rosemary* (Rosmarinus officinalis) *can be used to flavour bottles of olive oil.*

Above: *Herbs can be dried for later use by hanging them in bunches in a dry place out of direct sunlight.*

DRYING HERBS AND FLOWERS

Store herbs in a cool, dry place with minimum exposure to air and sunlight. One of the most popular methods of preserving them for use during the winter months is drying. This method can actually improve the flavour of some herbs, particularly the leaves of bay trees (*Laurus nobilis*). Herbs may be dried in bundles secured with a rubber band or string and hung upside down from a rack in a dry location such as an airing cupboard or shed.

When drying the herbs, the temperature of the area should not exceed 30°C (86°F) because the plants' essential oils will evaporate at or above this temperature. Do not dry your herbs in the kitchen where they will be spoiled by the humidity caused by cooking.

Fresh herbs can also be placed in brown paper bags. Remember to label the bags because it will be hard to distinguish between the herbs once they have dried. Store in a dry, dark, cool place until the herbs inside are dry, shaking the bags occasionally so that the plants dry evenly. Remove any stems and store the dried herbs in airtight jars. Keep the jars away from light to protect the colour and flavour of the herbs. Roots are best chopped into small pieces and dried in an oven. In general, you can expect those that you have grown and dried yourself to last at least two years.

STORING HERBS

Herbs can also be preserved in other ways, so that you can use them in cooking throughout the year.

Herb salts In a cool oven, spread a layer of ground salt on a sheet of greaseproof (waxed) paper. Sprinkle the chopped fresh herbs on top of the salt and bake for 10–20 minutes. When the herbs are dry, let them cool and place in a jar. Chives, oregano, thyme, lemon balm, lemon thyme, parsley, rosemary and basil can all be treated this way.

Puréeing This method involves mixing approximately 60ml (4tbsp) of olive oil with 2l (8 cups) fresh basil leaves which have been washed and dried. These are blended in a processor until puréed before being transferred to a jar. Stir each time you use it and top with a thin layer of oil afterwards. The purée should keep for up to one year in a refrigerator.

Above: *After herbs have been hung up to dry, separate the seeds from the dry flower heads before storing in tightly sealed jars.*

Freezing herbs Herbs such as dill, fennel, basil and parsley freeze well. The herbs should be cleaned and put into separate, labelled freezer bags. Alternatively, chop the leaves and freeze them with a little water in ice-cube trays. Chop the herbs finely, filling each cube, half with herbs and half with water, before freezing. Transfer the frozen cubes to plastic bags and label. Frozen herbs are best used within six months.

HERBAL INFUSIONS

You can also make a hot infusion of leafy herbs by placing the herb and any fine-quality olive oil in a glass bowl. This is then placed over a pan of gently simmering water and heated gently for about three hours, ensuring that the water in the pan does not dry out. The strained oil, once it has cooled, should be stored in airtight bottles or jars.

A cool infusion of flowery herbs involves using fresh herbs such as chamomile which are ground with a pestle and mortar and packed into a large jar and covered with oil. The sealed jar is then left in a warm, sunny place for two to three weeks and shaken occasionally. It is then strained and placed into airtight jars or bottles where it can be stored for up to a year.

A simple way of creating aromatic olive oil is to simply add a large sprig of your chosen herbs – rosemary and mint are good choices – to a bottle of olive oil. Store the bottle in a cool dark place for about ten days before using.

You can also make you own herbal vinegars by adding fresh herbs such as tarragon or rosemary or cloves of garlic, slices of ginger, chillies or peppercorns and all-spice powder to white vinegar. Crush about a quarter litre volume (1 cup) of loosely packed fresh herbs for each litre of vinegar. If you are using dried herbs, use half the amount of herb stated above. It is important that you use only commercially prepared vinegars, as homemade vinegar may not have a low enough pH to prevent bacterial growth. Place the vinegar in a pot on the stove and heat, but do not boil. Place the herbs in a clean, sterilized jar and slightly crush them. Pour the vinegar over the herbs and cover the jar tightly. Let the herb-vinegar mixture steep in a dark place at room temperature, shaking the jar every couple of days. After a week, strain the vinegar and place in bottles and store for up to six weeks.

DRYING AND FREEZING HERBS

1 *Pick seed just as it is ripening. Place the seeds on a tray or in a paper or muslin bag. Leave in a cool, dark place for a few days until the seed is completely dry.*

2 *Herb seeds that have been dried can be stored in old glass jars with an airtight lid. Store the jars in a cool, dry, dark place and label them for future reference.*

3 *Herbs can be frozen in ice-cube trays. Fill the trays with water after you have added the herbs to make ready-to-use cubes. Herbs can also be packed into freezer bags.*

GROWING FRUIT TREES AND BUSHES

Freshly picked fruit from the organic garden tastes absolutely delicious. The warm taste of a juicy sweet raspberry or the crisp flavour of a tree-ripened apple would tempt many a gardener into growing their own fruit produce. The tastes can be quite different from shop-grown produce where storage, handling, packing and off-the-tree ripening all take their toll on the quality and taste of the fruit.

Above: *A sunny wall provides an ideal location for growing fruit trees, such as this pear, that need warmth.*

WHERE CAN FRUIT BE GROWN?

Fruit trees and bushes can be grown wherever there is space. Strawberries can be used in hanging baskets or tubs, dwarf apple trees can be planted in ornamental containers and a number of fruits, such as cherries, white and red currants and figs, can all be grown against walls. Fruits not only provide produce for the kitchen table but many of them have ornamental qualities and can blend well with other plants in the decorative garden.

CHOOSING A SITE

The best site for fruit, both indoors and out, is a sunny one. Sunlight is essential, not just for the ripening of the fruit itself but also for flower bud formation and flowering. In addition, fruit trees and bushes often appreciate a sheltered spot where even exotic fruit can be raised. Within the garden a sheltered sunny wall can provide the ideal place to grow peaches, apricots or figs. If cold winters are a problem, then provide winter protection in the form of a portable frame or, alternatively, grow the plants in a greenhouse. It is worth noting that providing your plants with the ideal conditions for healthy growth will reap its rewards in the end. Not only will you harvest heavier yields, but, more importantly, your plants will be less prone to pest and disease attacks, which is an obvious advantage when growing organically.

SELECTING THE RIGHT PLANTS

Variety selection should not be based purely on hardiness but on personal preference. You may also wish to consider how easy the fruit tree or bush is to grow, fruit size, taste and the time of harvest. Selecting more than one variety can result in having fresh fruit over a longer period due to a succession of ripening.

Some fruit trees combine the best qualities of two plants. Grafting utilizes the qualities of the variety as top growth (scion) and other desirable qualities from a rootstock that may be absent from the variety. The scion is the fruiting variety that is budded or grafted on to the rootstock which is selected for certain characteristics such as dwarfing, nematode insect resistance, soil type, cold hardiness and disease resistance.

The most commonly grown rootstocks for amateur gardeners are the apple semi-dwarfs and dwarfs. Grapes are also grafted on to clonal rootstocks, although they are often supplied on their own roots. Figs, olives and various types of berry are also usually supplied as plants on their own roots.

Dwarf apple trees are very useful in small gardens and are eminently suited to container growing. They produce fruit of the same size, colour and quality as larger standard trees and require the same pruning, nutritional and care regimes as a standard-size tree.

Left: *Most fruit bushes will benefit from a generous application of well-rotted manure or compost early in the year.*

Dwarf trees fruit much sooner after planting and bear less fruit per tree. When harvesting the fruit, you can reach all parts of the tree from the ground without using a ladder and the trees are easier to train and prune on an annual basis. Grafting the desired variety on to special clonal rootstocks "dwarfs" apples. The most popular dwarfing rootstocks for apple were developed in England and are designated as either EM or M (for East Malling) or MM (for Malling Merton). Dwarfed trees must be pruned annually or size control may be lost. In addition, loss of fruit by frost or pests will also increase growth so necessitating summer pruning.

PLANTING FRUIT TREES AND BUSHES

As with any other type of tree, good ground preparation and careful handling are essential steps to successfully establishing your fruit trees and bushes. The cheapest option when buying fruit trees is to purchase bare-rooted plants. The disadvantage is that they are only available during the dormant season. The most important factor in handling bare-rooted plants is not to let the roots dry out. When you buy trees always check the condition of the roots and packing material. Heel in plants by covering the roots with moist soil in a cool environment outdoors if they are not to be planted immediately.

Dig a hole slightly wider and deeper than the spread and length of the root system, making sure the sides of the hole are not "glazed" over as this will result in a root girdling. After trimming diseased, dead, broken or extra long roots, place the tree in the hole and spread out the roots. For larger trees, place the stake in the hole and drive it in (remembering to first remove the tree and cover the roots). Place the tree back into

PLANTING A FRUIT TREE

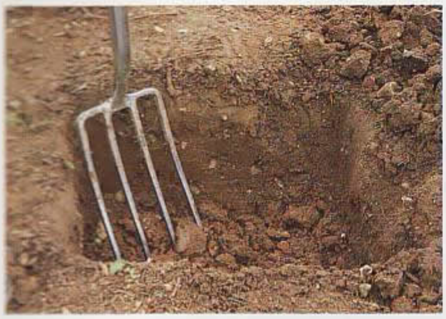

1 *Dig a hole in the ground that is at least half as large again as the rootball of the fruit tree. Loosen the sides and the base of the planting hole with a garden fork.*

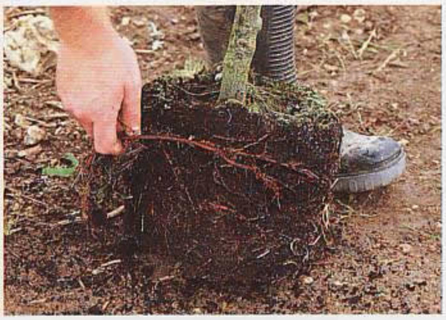

2 *Remove the pot and check the rootball for girdling roots. Tease these out by hand or with a garden fork.*

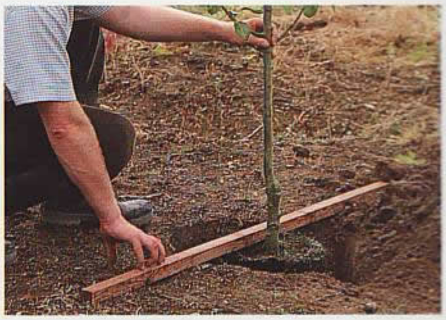

3 *Use a straightedge to make sure that the plant is at the right depth in the hole. Fill around the roots with soil, firm it down and water well.*

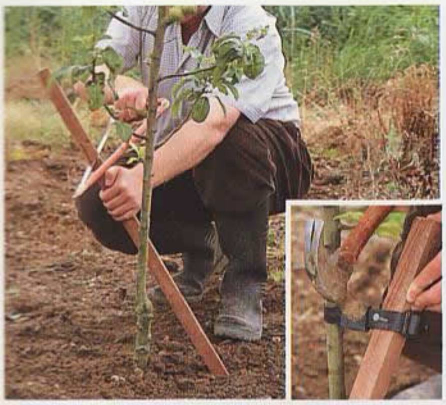

4 *Hammer in a stake at an angle of about 60° in order to avoid the rootball and place a tie on the tree. Saw off the end of the stake. Nail the tie to the stake to secure it.*

5 *The tree should remain staked for approximately one year. As the diameter of the tree stem increases, loosen the tree tie as required.*

the hole and return the soil, firming in layers of about 30cm (12in) as you go. This avoids large air spaces being left around the roots and ensures that it is set firmly. Trees should be planted at the same depth as they were grown in the nursery. Make sure the bud union (for trees on rootstocks) is about 5–7cm (2–2¾in) above the soil line. Do not place fertilizer in the planting hole as this can be added later. Mulch the newly planted tree with well-rotted manure or compost to suppress weeds. Container-grown nursery stock can be transplanted any time of the year. Site preparation is the same as for bare-rooted stock. Make sure that you check the roots as they can become distorted or root bound if they are grown in containers for a long period of time. Teasing these roots out can help avoid root girdling, but the best way to do this is to use field-grown (bare-rooted) stock.

PRUNING NEWLY PLANTED STOCK

Fruit trees must be pruned when they are planted for a number of reasons. If planting bare-rooted stock, the top of the tree must be pruned to counter-balance the loss of the root system which would have been severed in the nursery during lifting. Pruning also forces the growth of laterals from which the future framework of the fruit tree will be selected. Branches that are desirably located can be retained as part of the framework whereas undesirable branches are removed.

FRUIT TREE POLLINATION

Pollination is the transfer of pollen from the male part of the flower to the female part of the flower (the stigma) to allow fruit to set and seeds to develop. Seeds cause the fruit to develop properly.
If both the pollen and stigma are from the same flower or from another flower from the same variety, the process is called self-pollination. Fruit trees that set fruit as the result of self-pollination are called self-fruitful, whereas those relying on pollen from a different variety are called self-unfruitful. The latter needs two varieties near to each other for fruit set to occur. This is called cross-pollination.

Apples Apples generally need two varieties for good fruit set. This can be another apple variety or a crab apple that blooms with the edible crop.
Apricots Self-fruitful
Berries (all types) Self-fruitful
Cherries Sweet cherry is self-unfruitful and needs two varieties for good crop set. Sour cherry varieties are self-fruitful.
European plums Self-fruitful
Figs Self-fruitful
Japanese plums Self-unfruitful as a rule with the exception of Santa Rosa which will set fruit fairly well without cross-pollination.
Nectarines Self-fruitful
Peaches Self-fruitful with the exception of 'J.H. Hale' which has to be pollinated by another variety.
Pears These always need two varieties to ensure good fruit set.

SUPPORTING FRUIT TREES AND SHRUBS

Many of the fruits that we grow, such as free-standing apple or plum trees, require support only in the early stages, whereas other fruits, like the raspberry, need this throughout their lives. Supports benefit the plants in a number of different ways. They help in establishing strong roots and can prop up trained specimens such as cordons and espaliers. This helps to maintain healthy vigorous growth and increase fruit yields.

Above: *Apples can easily be grown in rows as cordons that are supported by a framework of stakes and wires.*

WALL AND FENCE FRUIT

The training of fruit on a wall or a fence is carried out to gain the maximum production of high-quality fruit in a limited space. As well as being ideal for a small garden, this can also look extremely decorative. Numerous training systems, based on the art of espalier which originated in France and Italy about 400 years ago, have been devised. The most useful training systems used in gardens today are the fan, espalier and cordon. Apples, pears and plums are all suited to this method of training, which is usually supported by a wall, fence or wire trellis. The plants are normally held to the wall or fence using wires which are held in place by vine eyes, positioned 60–90cm (2–3ft) apart. The wire is led through the holes in the vine eyes and secured at both ends. The wires should be no more than 30–45cm (12–18in) apart.

FREESTANDING WIREWORK

Raspberries, blackberries and other hybrid berries are all grown on a permanent framework. They are usually grown on a freestanding structure that supports the loose growth of the plants. To build such a structure, insert a post at least 60cm (24in) into the ground at the end of each row. Brace each end post with another post set at a 45-degree angle. Insert intermediate posts between the two end ones at a distance of every 2m (6½ft). They are set in at the same depth as the end posts, but no bracing posts are needed. Fix the wires on to the posts so that they run the length of the support. Place the first wire 60cm (24in) from the ground, pull tight and attach to the posts using staples or eye-bolts. The other wires are attached at 30cm (12in) intervals until they reach the top of the support. The final height depends on the height of the fruit.

INDIVIDUAL SUPPORT

When they are first planted, all free-standing trees require support to help them to establish a strong healthy root system. You can use a stake to support a fruit tree, ensuring that the stake is placed on the windward side of the tree. The stake is inserted before the tree is planted and is driven well into the ground. The general rule is to have a stake that is one-third the height of the tree showing above the ground. The tree is then tied to the stake using a tree tie. Do not use string or rope because this can damage the bark. A low stake such as this will allow the top part of the tree to move about freely and so help the tree to thicken and strengthen its trunk. For most fruit trees the stake should only be left in the ground for approximately one year and not for several years as is commonly practised.

STAKING A FRUIT BUSH

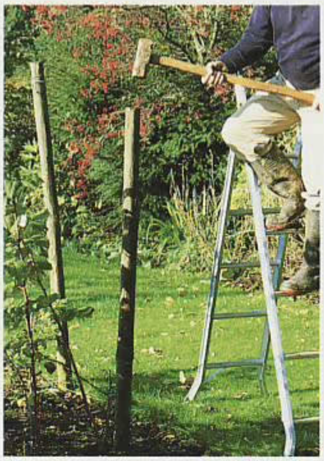

1 *Using a large mallet, drive a strong wooden post into the ground at the end of a row of the fruit bushes to a depth of about 60cm (2ft).*

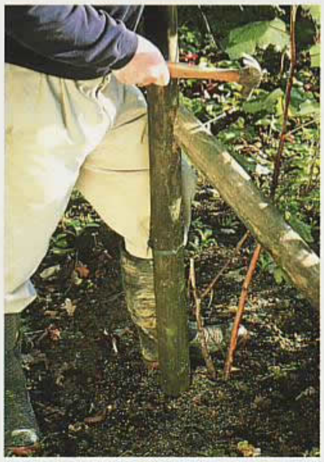

2 *Fix another strainer post at a 45-degree angle in order to support the upright. Place this on the side the wire is to run and nail it firmly.*

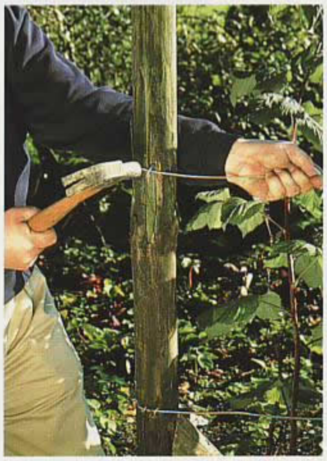

3 *Fix the wires to one of the posts and then stretch these tightly along the row, stapling them at each post along the row as you go.*

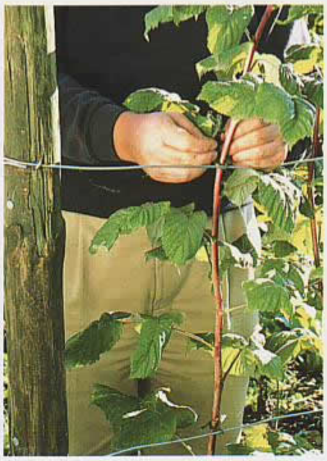

4 *The fruit bushes can be fastened to or trained along these wires. Take care not to tie stems too tightly as this will damage them.*

PROTECTING FRUIT

There can be few more frustrating events in the kitchen-garden calendar than losing your fruit crop (or a fair portion of it) to birds. However, we have to remember that birds are a strong ally in the fight against pests in an organic garden and should be encouraged into fruit-growing areas when the bushes are not fruiting. Protection for the fruit crops will therefore need to take the form of moveable cloches or fruit cages where the netting can be removed after fruiting has taken place.

Above: *Cloches can be used to protect young strawberries from bad weather and also to keep pests at bay.*

FRUIT CAGES

This is perhaps the easiest way to protect tall fruit crops such as blackcurrants, raspberries and gooseberries because these plants can all be grown under the one structure. This makes it easy to maintain and harvest the fruit.

A permanent fruit-cage structure with removable netting is the ideal scenario because the netting can be put in place just before the crop ripens and then is taken down again after fruiting has occurred. This allows the birds to roam freely around the area at all other times, but without jeopardizing the crop at fruiting time. Alternatively, a series of small cages can be constructed over crops that ripen at different times, netting each of them individually as the fruit begins to ripen. This is a wildlife-friendly way of protecting the fruit, but it is also time-consuming.

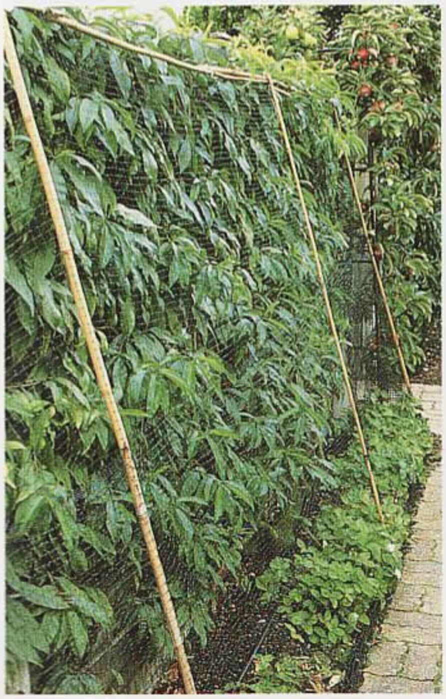

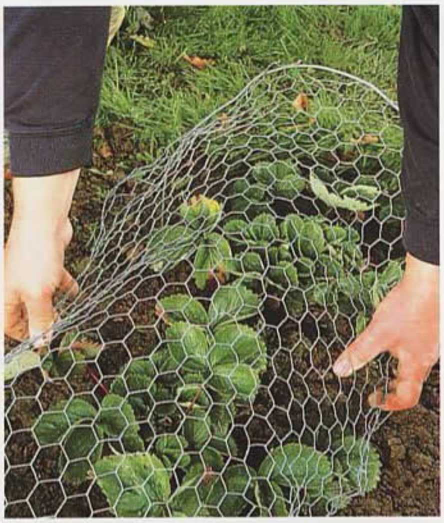

Above: *A tunnel of wire netting is ideal for protecting ripening strawberries from pests such as birds.*

LOW-LEVEL PROTECTION

Smaller- or lower-growing crops, such as strawberries, can also benefit from a barrier to protect them from birds. This is easily achieved, as the structure does not need to be large, permanent or even particularly sturdy. One simple method is to form a low tunnel of chicken wire, supported on hoops that are bent into an inverted U-shape. For larger areas, a series of stakes hammered into the ground can be covered with netting to keep the birds at bay. Plants growing on wire trellises can be protected with netting draped over the stakes and top wires and weighted at the base with stones.

Left: *Fruit trees growing against a wall can easily be protected with a cost-effective frame that is covered with netting.*

Right: *Fleece stretched over a frame can help protect the blossom of wall specimens from late frosts.*

Larger bush fruit can be covered with netting and weighted down with stones. A little of the crop may be accessible to birds, but most will be protected.

PROTECTION WITHOUT BARRIERS

Most bird deterrents depend upon shocking the birds or mimicking something that they are naturally afraid of. Scarecrows are an age-old favourite. Other deterrents mimic the shape of hawks or make noises that will startle the birds. A relatively recent "innovation" has been to hang compact discs from wires stretched through the crop. Most of these deterrents tend to work very well at first until the birds get used to them. However, the only drawback is that when the birds get used to them, they resume their onslaught with a vengeance. The only way to overcome this is to keep changing the deterrents every few days.

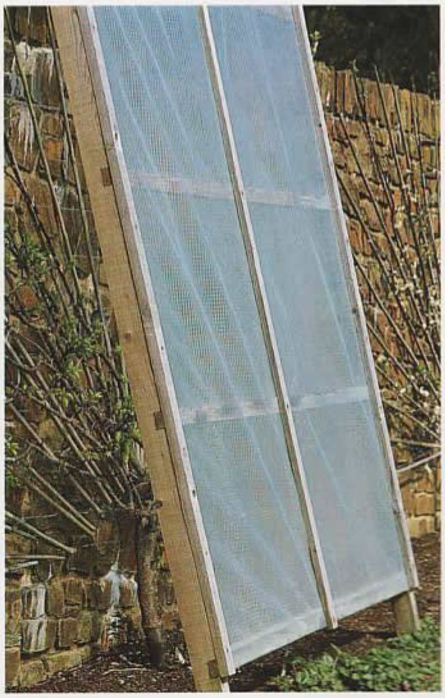

PRUNING FRUIT TREES AND SHRUBS

The general purpose of pruning fruit trees is to regulate growth, increase crop yields and to improve fruit size and quality. The fruit trees also need to have all the dead, diseased and damaged wood removed as well as overcrowded areas thinned out. Pruning is also performed in order to establish a tree with a strong framework that is capable of supporting heavy crops without causing damage to the plant.

Above: *The careful pruning of trained specimens, such as this pear, will result in good blossom and improve fruit yields.*

WHEN TO PRUNE FRUIT TREES

If a fruit tree is pruned in spring, all the effort that the tree has put into bud production is wasted. The best time to prune is in winter when the plants are dormant and the sap has not yet started to rise. This applies to all fruit except for plums and cherries which are susceptible to silverleaf disease at this time and are best pruned in summer.

Summer pruning is also recommended if you wish to control vigorous growth in formally trained fruits such as fans, cordons and espaliers.

When pruning is underway, older fruit-bearing trees should be pruned first. Young, non-bearing apple trees and stone fruits can be pruned later to minimize the risk of winter injury.

BASIC PRUNING CUTS

Although there are different methods of pruning and training, the pruning cuts are the same. Always make the pruning cut just above a bud, ensuring that the cut is angled away from the bud. This will allow rainwater to drip down the other side of the stem and away from the bud, thus protecting the top of the cut from rotting.

Larger branches will need to be cut with a saw. If the branch is very heavy, you will need to remove the branch in three separate stages. Cut underneath the branch first, 20–25cm (8–10in) away from where the final cut will be. Make the second cut on the top of the branch just behind the first undercut. Follow this through until the weight of the branch makes the wood split and it falls off cleanly at the undercut. You will then need to make a third cut at the branch collar, cutting straight from the top to the bottom of the branch as there is no weight left to tear the bark.

BENEFITS OF PRUNING

Pruning fruit trees and bushes has many advantages, some of which are outlined below.

- Increased fruit yields.
- Can aid ripening of the fruit.
- Maintains healthy growth.
- Regulates growth.
- Improves fruit size and quality.
- Controls the spread of diseases.

Above: *Training can give rise to a variety of ornamental shapes. These pear trees have been trained into a cylindrical shape.*

PRUNING LARGE BRANCHES

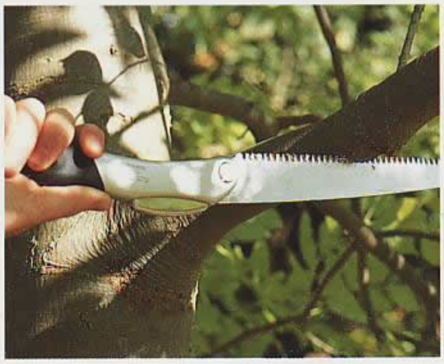

1 *Make a cut about 20–25cm (8–10in) out from where the final cut will be on the underside of the branch. Cut about a third of the way into the branch.*

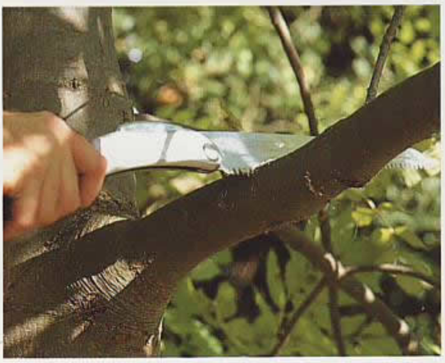

2 *Make a cut about 10cm (4in) nearer the position of the final cut, cutting until the branch snaps. The initial undercut prevents the wood splitting or bark from stripping.*

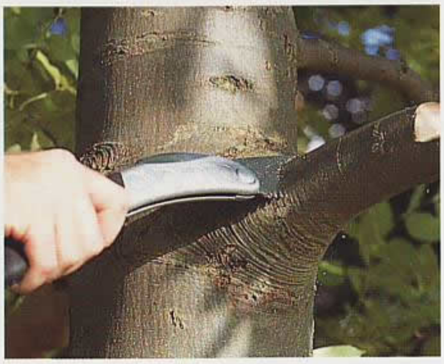

3 *Position the saw for the last cut to avoid damaging the swollen area at the base of the branch. The cut will heal in a couple of seasons. No wound painting is necessary.*

SUCCESSFUL FRUIT TREE AND SHRUB PRUNING

Pruning fruit trees is a very complex subject and, like so many areas of organic gardening, is the subject of opinion and hot debate. Despite this it is possible to apply certain guiding principles that can enable you to prune your fruit trees and bushes successfully.

1 **Remove root suckers arising at the base of the tree**
These compete with the upper growth for water and essential nutrients and their dense habit can harbour pests.

2 **Always cut out dead, dying, damaged or diseased limbs first**
Helps to maintain the plant's health and enables you to see precisely what needs doing next.

3 **Remove low, drooping limbs**
These will not bear fruit and will often be heavily shaded by growth above. They are an unproductive drain on the tree.

4 **Remove upright growth or outward growth in the case of wall-trained bushes**
Upright growth will produce a flush of growth at the end and will not fruit well. Branches growing away from the wall will shade the fruit behind causing poor ripening.

5 **Remove crossing or dense parallel growth**
Crossing growth will rub and can cause bark damage that will allow disease entry. Dense growth will also shade the developing fruits. This will slow and limit the ripening of the fruit.

6 **Freestanding trees should have upper limbs cut back further than lower limbs**
This will result in the development of a conical shape that will allow light penetration all the way down the side. Fruit will ripen more evenly.

7 **Remove water sprouts as they develop**
Water sprouts can quickly develop and cover the developing fruits, causing poor ripening and harvesting difficulties. They can be easily removed by hand. This is best done early in the season when the growth is soft and is therefore easily removed. Carry out regular inspections throughout the season to control the water sprouts.

8 **Once the larger cuts have been made, thin out the smaller branches**
Removing thick branches first allows you to see what remains to be done and also makes it easier to get to the remaining pruning work. Removing older wood encourages young growth to fill the gap. Remember to stand back and view what you have pruned as you go along.

9 **Always make clean cuts above a leaf bud or close to the stem**
Pruning cuts that are not performed cleanly or that leave a long stub will damage the plant tissue and encourage the entry of pests and diseases.

10 **Avoid feeding with nitrogen for a season before and following heavy pruning**
Both pruning and high nitrogen fertilizers promote the development of rapid growth flushes that can result in the poor development of the crop.

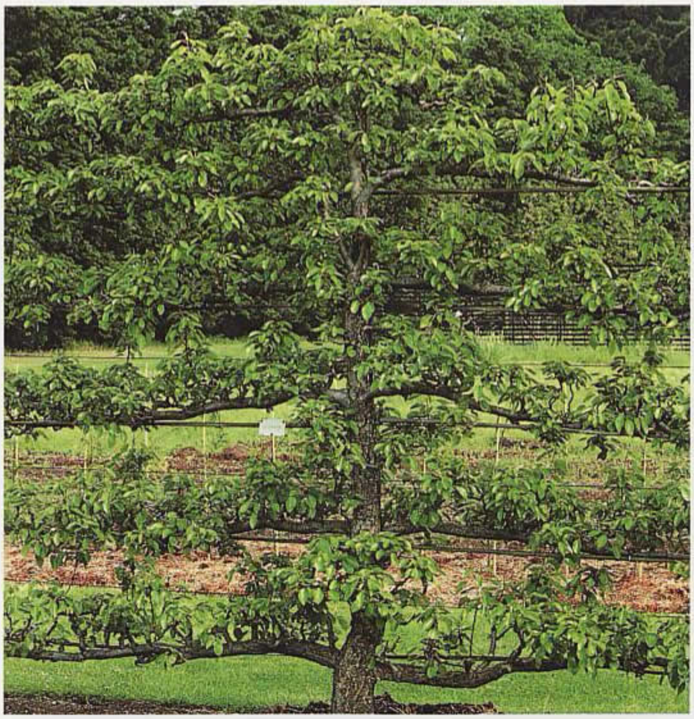

Above: *This well-established pear tree has been trained into an espalier on wire supports. Many fruit trees can be grown and supported in this way.*

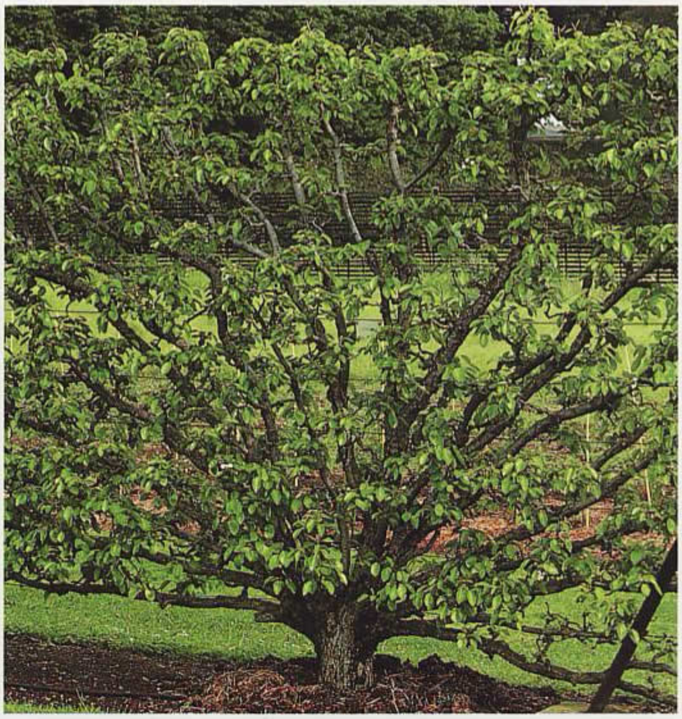

Above: *This pear tree has been grown as a fan. Fans are an ideal way to train many species of fruit tree and are well suited for growing against a wall.*

HARVESTING AND STORING FRUIT

When harvesting your fruits, you will find that there is only so much you can eat fresh or give away. There is always surplus fruit left over and you are faced with the question of what to do with it. Leaving some on the plant to help feed the wildlife in your organic garden is a good idea, but storing the rest of the excess fruit using a variety of methods will ensure that you can eat your home-grown produce over a longer period of time.

Above: *An apple is ready to be picked if it can be removed from the tree with only a single twist of the fruit.*

HARVESTING

The key to successfully storing fruit begins well before harvesting commences. Your first objective should be to grow fruit that is as healthy as possible because it will be the best for storage. Harvesting immature crops or attempting to save those that are in poor condition – due perhaps to a lack of water or nutrients or to pest and disease damage – can lead to many storage losses.

There are several ways that fruit can be stored and the condition of the picked crop will usually be the deciding factor as to which of these you should use. Top fruit such as apples and pears can often be stored fresh through most of the winter, whereas stone fruits and berries, such as peaches, strawberries and raspberries, must be quickly consumed, turned into a preserve or frozen.

Careful handling, both during and following harvesting, is essential because, from the moment they are harvested (and in many cases well before), crops have no means of repairing any physical damage that they may suffer. Even firm, strong-looking fruit such as apples can easily be bruised, although the damage may well not show up immediately.

As well as good handling, a careful selection of the fruits during harvesting is essential for successful storage. You should inspect the picked fruits and select only those of the best quality for fresh storage. Reject any that have a broken skin or show any sign of pest or disease damage. Do not throw them on the compost heap yet, however, because damaged fruit may well be useful for making preserves such as jellies and jams or for freezing.

It is also a good idea to leave a small percentage of mature fruit on the plants when you are harvesting to help feed the wildlife in your organic garden.

STORING

In general, the storage area for fruit must be frost-free, safe from pests, rainproof and ideally kept at a constant temperature.

STORING FRUIT

To store surplus fruit and avoid the risk of rotting, make sure that you use the appropriate method.

Apples, pears and quinces
Store in a cool place for up to 12 months, depending on the variety.
All other fruit Eat immediately or freeze. Alternatively, preserve fruit by bottling or making into jam. Fruit can be kept for up to 12 months, depending on the method of preservation that has been used.

The long-term storage of any fruit calls for cool conditions with adequate ventilation. If you have space, consider having a separate refrigerator specifically for fruit storage, or choose an area with a low temperature that does not go below freezing. A garden shed or garage can be ideal, in many cases, but even these

Far left: *Pick soft fruits, such as strawberries, raspberries and gooseberries, carefully to avoid bruising the fruit.*

Left: *Pears can last for up to 12 months, depending upon the variety. Lay them in a recycled box, ensuring that the fruits do not touch and air can circulate around the fruit.*

Above: *Fruit trays, recycled from supermarkets, make ideal storage boxes for surplus fruit. The trays should be kept in a cool dry place such as a garage or shed.*

Above: *Soft fruit is best placed in small individual containers as it is picked to prevent it being squashed and spoiling. Store in a refrigerator or other cool place.*

spaces may need extra insulation if winter weather conditions become severe. Some houses have a basement, cellar or unheated room that may be ideally suited for the task. Attics are not recommended for fruit storage because of their wide temperature fluctuations and variable humidity.

It is worthwhile trying out a variety of different storage methods. This can be done by splitting the crop up and then trying out different locations in your home. You will soon find which are the best places to store fruit, and what areas are best for storing a particular kind of fruit.

Ensure that you check the stored fruit regularly, at least weekly, removing any that show signs of decay. Remember the old adage that "one bad apple spoils the whole barrel"? This can be true for your crop too unless you prevent rots from spreading. The unblemished parts can often still be used for eating or cooking. If lots of fruits begin to rot simultaneously, it could be that the storage conditions are not suitable, the crop has reached its maximum "shelf-life" or the fruit was not of sufficiently good quality to start with.

FREEZING AND PRESERVING

Freezing is an excellent way of storing all the surplus fruit that has been produced in your garden. Unfortunately, fruit tends to lose its firmness once it has been frozen, although the taste will remain more or less the same. Raspberries, for example, are best used for making pies or flans after they have been frozen and will be of very poor quality if they are eaten raw. Most fruits can be frozen after they have been stewed or puréed; this is true of fruits such as apples and plums.

Fruits like strawberries and blackcurrants that are not suited to long-term storage can be made into jams, pickles and chutneys. You can find recipes for these in good cookbooks.

Above: *One of the simplest ways of preserving fruit is to freeze it, although once frozen most fruits are only good for cooking.*

Above: *Adding a favourite fruit such as these cranberries to good wine vinegar can produce interesting flavours.*

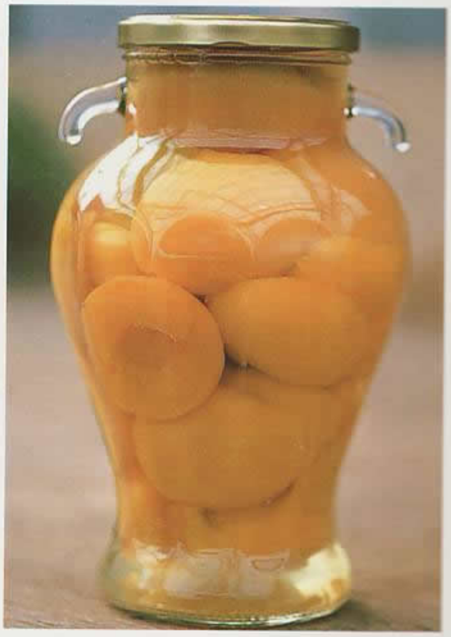

Right: *Bottling fruit in alcohol is an ideal way to preserve soft fruits such as peaches, nectarines and apricots.*

CALENDAR OF CARE

WINTER

THE ORNAMENTAL GARDEN

EARLY WINTER
- Continue cultivating the soil, if not already done, when the ground is not wet.
- Continue planting and transplanting on mild days.
- Press plants that have been lifted by frost back into the ground.
- Plant herbaceous perennials for next year.
- Last chance to plant roses.
- Move greenhouse plants to the house as they come into flower.
- Start pruning deciduous shrubs and trees during spells of fine, frost-free weather.

MID-WINTER
- Finish digging and trenching and fork over borders.
- Order seeds for coming season.
- Protect autumn sown annuals from frost.
- Prepare ground that is to be planted out with bedding plants with manure.
- Tidy up climbing plants and remove any dead wood.
- Firm down herbaceous plants that have been lifted by the frost and protect the new shoots of young tender plants.
- Plant evergreen and deciduous shrubs.
- Trim deciduous hedges and shrubs to start bringing them into shape.
- Take hardwood cuttings and layers of deciduous shrubs.
- Sow annuals in gentle heat.
- Take cuttings of tender shrubs and start under gentle heat.

LATE WINTER
- Tidy up unused ground and apply a top-dressing of manure or leaf mould.
- Protect bulbs from mice, snails and birds.
- Prune roses planted this month.
- Start to train and trim climbers.
- Divide herbaceous plants.
- Sow annuals in gentle heat.
- Transplant autumn sown annuals.
- Plant edgings such as box (*Buxus*) hedging, pinks (*Dianthus*) or thrift (*Armeria*).

THE WILDLIFE GARDEN

EARLY WINTER
- Try to maintain a fresh supply of clean unfrozen water for wildlife.
- To keep some open water on ponds, put in a floating ball or anchored block of polystyrene, preferably painted black, to absorb the warmth of the sun.
- Leave tidying the garden until late winter to shelter overwintering insects.
- Leave log and leaf piles completely undisturbed until spring.
- Feed resident birds with high energy supplements such as fat balls.
- Order seed now for early sowings of perennials and annuals.

MID-WINTER
- Put plenty of food in the garden for resident birds.
- Hang nesting boxes ready for breeding birds.
- Build dry-stone walls with gaps to provide shelter for reptiles and amphibians.
- Look out for wildlife tracks in snow and position cat deterrents at entry points.
- Provide food for non-hibernating mammals if present.
- Keep a record of wildlife to help you target future feeding and habitat creation.

LATE WINTER
- Continue feeding resident birds.
- Put up a range of different nesting boxes and make each box as inconspicuous as possible. Reposition any nesting boxes that were unused last year.
- Stretch wires or heavy-duty cord at least 10cm (4in) proud of walls and fences and train climbing plants to form a screen with a gap behind it for nesting birds.
- Trim one side of hedges to encourage dense new growth as nesting cover, but leave some flowering stems on spring-blossoming species such as hawthorn (*Crataegus*) for nectar.
- Protect bulbs from mice, snails and birds.

SPRING

EARLY SPRING
- Feed lawns where the grass is thin and seed areas where the grass has been completely worn away.
- Prepare borders ready for biennials and perennials next month.
- Remove leaves and other debris from borders and fork over.
- Tidy up spring bedding plants.
- Sow hardy annuals when the weather is dry and mild.
- Start pruning roses.
- Trim box edging and damaged evergreens.
- Prick out annuals started last month.
- Start to harden off early seedlings ready for bedding out.
- Look for pests and diseases and deal with them as they occur.

MID-SPRING
- Hoe off weeds.
- Watch for signs of frost and protect young shoots.
- Watch out for pests and diseases.
- Give the lawn its first cut and tidy its edges.
- Stake and tie any plants that need it.
- Sow hardy annuals.
- Propagate hardy perennials, including water plants.
- Finish winter pruning and prune late winter/early spring flowering shrubs before they commence growth.
- Continue hardening off summer bedding and pinch out the tops.
- Trim and tidy up all climbers before they start into growth.
- Sow half-hardy annuals and take cuttings.
- Give box hedges their first clipping.

LATE SPRING
- Hoe off weeds.
- Finish off digging and rotavating to prepare the beds for planting.
- Water newly planted plants in dry weather.
- Keep an eye open for pests and diseases.
- Tidy bulb beds after the foliage has died down.
- Tie in sweet peas (*Lathyrus odoratus*) and check the support of other climbers.
- Finish sowing hardy annuals and plant out those raised under glass. Thin outdoor sowings and transplant where necessary.
- Plant half-hardy bedding plants when the danger of frost is over.
- Divide spring-flowering plants, like primroses (*Primula vulgaris*), for next season.
- Shrubs should have old wood removed and thinned out.
- Continue taking cuttings.

EARLY SPRING
- Put plenty of food in the garden for resident birds.
- Create a wildlife pond ready to plant up next month.
- Supplement the diet of resident early nesting birds with mealworms and waxworms.
- Emerging hedgehogs will benefit from a feed of cat food near their nesting sites.
- Hang nesting boxes for solitary bees.

MID-SPRING
- Plant new garden ponds and replace dead plants in established ponds.
- Sow seeds of hardy annuals in a sunny spot to attract insects during the summer until late autumn. Mix them up with salad crops to act as helpful companion plants.
- Install a "cage" of large-gauge chicken wire around nesting sites to allow nesting birds in but keep out magpies, grey squirrels and other nest-robbing species.
- Keep one or two nest boxes sealed so as to prevent the entry of resident birds until migrant bird species arrive.
- Continue feeding garden birds.
- Continue feeding hedgehogs until natural prey increases in numbers.
- Propagate hardy perennials, not forgetting water plants.

LATE SPRING
- Plant pot-grown seedlings of spring-flowering wild flowers in an area of lawn where you can leave the grass uncut until mid-summer.
- Sow the seeds of wild flowers thinly on the surface of gritty potting mix.
- Divide spring-flowering plants, such as primroses (*Primula vulgaris*) and cowslips (*P. veris*), for next season.
- Continue feeding garden birds.
- Put up bat boxes to encourage bats to roost in the summer.

SUMMER

EARLY SUMMER

- Mulch round plants and shrubs that need to be kept moist.
- Keep an eye out for pests, and keep weeds under control.
- Cut down and tidy perennials as they finish flowering.
- Feed roses and sweet peas (*Lathyrus odoratus*) with manure or compost tea.
- Sow biennials and perennials.
- Plant out plants raised under glass, dividing and planting out those that have finished flowering in other areas.
- Thin out seedlings and stake carnations (*Dianthus*), dahlias and sweet peas.
- Sow hardy annuals for autumn planting.
- Spray the foliage of new trees with water in hot or dry weather.
- Tidy up rhododendrons and other plants by removing seedheads, unless leaving them for decoration and food for the birds.

MID-SUMMER

- Continue to hoe weeds and water newly planted specimens.
- Stake and tie plants.
- Layer or take cuttings from carnations.
- Pinch out the tops of chrysanthemums to prevent straggly plants.
- Water and feed sweet peas with a compost or manure tea.
- Sow perennial and biennial seeds for next year.
- Continue taking cuttings.
- Sow seeds for overwintering in the greenhouse.
- Pot plants that need repotting.
- Cut away dead and unwanted foliage on vigorous shrubs.
- Trim deciduous hedges.
- Propagate shrubs by cuttings or layering.

LATE SUMMER

- Continue watering, mulching and hoeing.
- Remove all dead flowers, stalks and leaves.
- Train and tie clematis, honeysuckle (*Lonicera*) and wisteria.
- Trim plants such as lavender and penstemons after flowering.
- Water and feed sweet peas with a compost or manure tea.
- Take cuttings of old-fashioned roses.
- Thin perennial seedlings.
- Divide and replant spring-flowering perennials.
- Take cuttings of fuchsias, pelargoniums, heliotrope, hydrangeas and all half-hardy plants.
- Remove dead flowers and stalks from summer-flowering shrubs, unless they are berry producing.

AUTUMN

EARLY AUTUMN

- Hoe off any late weeds.
- Start tidying up for winter, but leave seedheads for birds.
- Put unwanted debris on the compost heap.
- Collect seeds and label for sowing next year.
- Start digging, manuring and rotavating heavy soils.
- Repair patches on lawns.
- Start collecting fallen leaves to make leaf mould.
- Tie and stake dahlias and tall shrubs.
- Cut back and support climbing plants.
- Sow spring-flowering seeds, starting with hardy annuals.
- Plant spring bulbs.
- Plant or transplant evergreen shrubs, conifers and herbaceous plants.

MID-AUTUMN

- Collect fallen leaves to make leaf mould.
- Prepare areas for planting.
- Repair and lay new lawns, repair paths and walkways.
- Make final plantings of bulbs.
- Collect seeds of late-flowering plants.
- Lift and store half-hardy bulbs, corms and tubers.
- Lift, divide and replant herbaceous plants.
- Sow early sweet peas (*Lathyrus odoratus*).
- Start planting out spring bedding.
- If moving deciduous shrubs, do so to encourage new root growth.
- Take cuttings until the cold weather sets in.

LATE AUTUMN

- Continue digging heavy and medium soils, leaving the light soils for spring.
- Spread surplus manure around flowerbeds if leaf mould is not available.
- Fork over herbaceous borders, leaving dead foliage until spring.
- Take cuttings of plants that may need replacing.
- Make final plantings of spring bedding.
- Take rose prunings from choice plants and push into the ground in a sheltered position where they may strike and give you more plants.
- Put bulbs for forcing under glass in pots.
- Plant flowering shrubs and trees for next season.

EARLY SUMMER

- Plant tobacco plants (*Nicotiana*) to guarantee interesting night-flying moths and the bats that feed on them.
- Check fruit nets each day to ensure no birds are trapped inside.
- Choose bedding plants with simple flowers and keep a note of which the bees and butterflies prefer.
- Leave seedheads on spring-flowering shrubs such as rhododendrons as food for birds.
- Continue feeding garden birds, with seed and waxworms or mealworms, especially if there is a dry spell.

MID-SUMMER

- Watch for night-flying moths on tobacco plants, late flowers on the honeysuckle, and tall yellow evening primroses.
- Continue to check fruit nets each day for birds.
- Take special care when turning compost heaps as grass snakes may be nesting and hedgehogs often sleep in them.
- Make log piles for wood feeding insects such as stag beetles.
- Lightly trim deciduous hedges if birds have finished nesting to encourage dense cover in the winter.
- Continue feeding garden birds, especially if there is a dry spell.

LATE SUMMER

- Plant early spring bulbs now for nectar next year, especially for overwintering insects.
- Lightly trim nectar-rich plants such as lavender (*Lavandula*) and nepeta after flowering to encourage new growth.
- Dead-head herbaceous plants to encourage fresh flowers that give an extra boost to pollinators in the early autumn.
- Leave seedheads and berries on summer-flowering shrubs for birds.
- Continue feeding garden birds.

EARLY AUTUMN

- Let one or two old vegetables flower, such as leeks, carrots and almost any member of the cabbage family, to boost nectar supplies for autumn insects.
- Look now to see if there is a shortage of nectar and redesign flower borders accordingly. Remember to select single-flowered plants where possible.
- Build new log piles, rock piles and brush piles ready for overwintering insects, reptiles, amphibians and mammals.
- Place nest boxes outside for overwintering mammals such as hedgehogs.
- Continue planting spring bulbs for early nectar next year.
- Plant evergreens to provide dense winter cover for birds and mammals.
- Continue feeding garden birds.

MID-AUTUMN

- Start a new compost heap.
- Let hedge prunings, raspberry canes and other woody waste accumulate and leave them to become a hibernation heap for hedgehogs.
- Weed among and divide most pond plants before they die back. This removes excess nutrients from the pond and minimizes next summer's algal blooms.
- Pile weeds and thinnings from ponds at the side until the following day to give trapped pond dwellers time to escape.
- Place nest boxes outside for overwintering mammals such as hedgehogs.
- Continue planting spring bulbs for early nectar next year.
- Begin putting out high energy supplements such as fat balls for birds.

LATE AUTUMN

- Plant hedgerows using native shrubs and trees.
- Plant late-flowering nectar plants like English ivy (*Hedera helix*) against a boundary that catches autumn sun.
- Put in late bulbs at least 7.5cm (3in) deep and bury chicken wire just below the surface to avoid them being eaten by squirrels.
- Feed resident birds with high energy supplements such as fat balls.

WINTER

THE KITCHEN GARDEN

VEGETABLES

EARLY WINTER

- Hoe between crops on fine days.
- Examine stored onions and potatoes for any sign of disease and discard second-class seed. Any potatoes that are sprouting should be placed apart in dry trays.
- Protect more tender plants with leaves or cloches.
- Begin forcing rhubarb and seakale under a light-proof cover.
- Order seed catalogues and plan the new season's crops.

MID-WINTER

- Tidy up and put vegetable waste on the compost heap.
- Hoe between growing crops when dry.
- Sow broad (fava) beans and peas in warm sheltered positions.
- Sow early vegetables under glass or cloches.
- Start potatoes in a frost-free shed.
- Plant rhubarb in well-manured beds.

LATE WINTER

- Hoe weeds when the weather allows.
- Clear old beds ready for new season.
- Cut pea and bean sticks.
- Sow early vegetables under glass or cloches.
- Plant artichokes, garlic, shallots, lettuce (winter), potatoes (early), rhubarb and seakale.
- Plants overwintering under cloches should be hardened off and only protected when the weather is frosty.

FRUIT

EARLY WINTER

- Finish pruning fruit trees and bushes on frost-free days. Large standard trees will need pruning every second or third year.
- Firm newly planted trees where the frost has been at work.
- Do not plant any more trees until the spring; just prepare the ground.
- Currants and gooseberries should be pruned if not already done.
- Prune outdoor vines.
- Remove unwanted suckers from raspberry stools but do not prune until early spring.

MID-WINTER

- Start pruning hardier trees such as apples. Do not prune apricots, cherries, figs, nectarines and peaches.
- Prepare grafting material.
- Plant container-grown trees and cover soil with well-rotted manure.

LATE WINTER

- Prune damsons, pears, plums and quince.
- Spread manure around fruit trees and dig it in.
- Prepare new strawberry beds.
- Fertilize by hand the blossom of fruit trees grown under glass.

HERBS

EARLY WINTER

- Prepare ground for planting when conditions allow.
- Protect tender herbs under cloches.
- Continue to remove dead stems from herbaceous types.

MID-WINTER

- Prepare ground ready for planting.
- Force early growth under cloches.
- Make early sowings under glass.

LATE WINTER

- Remove dead stems from herbaceous types.
- Continue forcing under cloches.
- Continue sowing under glass.

SPRING

EARLY SPRING

- Hoe weeds when the weather allows.
- Prepare seedbeds.
- Protect early crops from birds.
- Begin sowing in sheltered places in the open.
- Sow early vegetables under glass or cloches.
- Plant out early crops raised under glass and early potatoes.
- Plant early broad (fava) beans out in the garden under glass.

MID-SPRING

- Hoe weeds regularly.
- Prepare beds for planting out.
- Feed cabbages with high-nitrogen fertilizer such as chicken manure.
- Potatoes need the soil drawn over their leaves to protect them from frost.
- Begin sowing remaining summer crops outdoors.
- Continue sowing more tender crops under glass.

LATE SPRING

- Hoe weeds regularly.
- Cut asparagus when shoots reach 15cm (6in).
- Maintain the succession of seed sowing to give salad throughout the summer season.
- Continue planting out summer crops.
- Sow late crops such as broccoli and plant out winter crops previously sown under cover.
- Last chance to plant late potatoes.
- Plant out cucumbers, pumpkins and melons in beds previously used for protected early crops.

EARLY SPRING

- Protect blossom with fleece on cold nights.
- Last chance to plant container-grown fruit bushes.
- Graft apples, cherries, pears and plums that were prepared in mid-winter.
- Finish pruning fruit trees and bushes.

MID-SPRING

- Continue protecting blossom.
- Wall trees will need moisture from now on.
- Watch for pests and diseases.
- Grapes will require thinning in order to swell their size.

LATE SPRING

- All fruit will benefit from a mulch.
- Apricots, peaches and Morello cherries should have fruiting spurs shortened back to three to four leaves.
- Limit raspberry suckers to around four to six to each stool.
- Put straw around strawberries to protect the fruits from the ground.

EARLY SPRING

- Continue sowing under glass.
- Plant out hardy herbs once hardened off.
- Prune shrubby types.
- Lift and divide herbaceous types until late spring.

MID-SPRING

- Continue sowing under glass.
- Continue planting out.
- Take basal cuttings.

LATE SPRING

- Sow herbs like parsley (*Petroselinum crispum*) and chervil (*Anthriscus cerefolium*) for winter use.
- Continue sowing under glass and planting out.

SUMMER

EARLY SUMMER

- Hoe weeds regularly throughout summer.
- Water all salad crops regularly.
- Feed heavy feeders like cauliflowers with liquid manure or compost tea.
- Keep picking peas to stimulate flowering.
- Harvest early potatoes and replace with winter turnips or late celery.
- Support runner beans.
- Sow the last of the outdoor salad crops and continue sowing seed of late or winter crops.
- Thin out seedlings.
- Plant late-season crops raised under cover.

MID-SUMMER

- Water all salad crops regularly.
- Remove debris, following harvest, to the compost heap until late summer.
- Potatoes and other crops may need earthing-up.
- Pinch out and stop tomatoes and marrows.
- Maintain the succession of seed sowing for late or winter crops.
- Thin out the beetroot crop so that the roots can form properly.
- Plant late-season crops raised under cover.

LATE SUMMER

- Water all crops regularly.
- Remove debris, following harvest, to the compost heap.
- Harvest seed from French (green) beans and onions.
- Garlic and onions can be harvested this month and then ripened.
- Potatoes for next year's seed should be dug up and dried in the sun.
- Maintain the succession of seed sowing for late or winter crops.

EARLY SUMMER

- Continue mulching trees and bushes to conserve moisture.
- Cut back untidy vigorous growth.
- Over-laden trees may need support. Aphids will appear, so take appropriate action.
- Throw nets over fruiting trees and shrubs to prevent the birds getting the fruit first.
- Peaches and apricots may need some thinning.
- Strawberries should have regular watering.

MID-SUMMER

- Thin trees and remove unnecessary growth.
- Hoe round the roots of fruit trees and water them.
- Fruit trees may be budded when the weather is moist.
- Espalier and dwarf fruit trees will need training and protection from birds.
- Strawberries and loganberries should be layered and clipped to the soil to secure their position.

LATE SUMMER

- Protect fruit on walls from wasps and birds as it ripens.
- Pruning can be started and trees budded.
- Apples, pears and plums may need the fruit thinning out if the crop is heavy.
- Cut down the old canes of loganberries and raspberries after harvesting.
- Cut off unwanted runners from strawberries and remove the old straw from round the crowns.

EARLY SUMMER

- Continue sowing under glass.
- Continue planting out until late summer.
- Harvest herbs for storing before they flower.
- Plant out tender herbs.
- Dead-head unless seed is required until late summer.
- Cut back herbaceous types to stimulate growth until late summer.

MID-SUMMER

- Cut mint (*Mentha*) and other sweet herbs for drying.
- Pull out parsley (*Petroselinum crispum*) unless it is being kept for seed.
- Propagate sage (*Salvia*) or savory (*Satureja*) from cuttings or division.

LATE SUMMER

- Harvest herbs as required.

AUTUMN

EARLY AUTUMN

- Hoe weeds regularly throughout autumn.
- Water all crops regularly.
- Harvest onions and bend down the necks of the remainder to prevent seeding.
- Continue to lift potatoes.
- Expose the fruit of outdoor tomatoes to the sun by removing the covering leaves.
- Maintain the succession of seed sowing for late or winter crops.
- Plant late season crops raised under cover.

MID-AUTUMN

- After beans and peas have been harvested, turn the soil over and leave it fallow over the winter.
- Lift root vegetables when their tops fade.
- Cut down asparagus foliage when mature.
- Onions and turnip beds should be thinned.
- Clear the last of the potatoes and store them.
- Earth up celery and leeks.
- Plant out late-season crops and protect tender types such as cauliflower.

LATE AUTUMN

- Get the soil ready for spring sowing.
- Protect crowns of tender crops with a mulch of leaves.
- Lift the last of the beetroot and carrots.
- Earth up celery and leeks if not already done.
- Spinach will continue to crop as it is thinned.
- Dry off the late potatoes for seed stock next year.
- Continue to plant out late-season crops and protect tender types such as cauliflower.

EARLY AUTUMN

- Wall fruit that has ripened will need protecting.
- Fruit trees may be budded and light pruning continued.
- Early varieties of apples and pears that do not keep well should be gathered.
- Prune currants and gooseberries to keep them in shape. Raspberry canes that have fruited should be removed to make room for the new growth.
- Strawberries should have unwanted runners removed to give space to the main plants.

MID-AUTUMN

- Look though the nursery catalogues and order fruit stock, then prepare the ground and remove stock to be replaced.
- Move any trees or bushes that are in the wrong place.
- Apples and pears can now be gathered.
- Apricots, cherries, currants and gooseberries should be pruned and the cuttings burnt.
- Loganberries and raspberries should be planted now.

LATE AUTUMN

- This is the time to plant fruit trees in well-prepared ground.
- Begin winter pruning of fruit trees.
- Plant currants and gooseberries.
- Figs will need thinning.
- Spread manure between rows of new strawberry plants.

EARLY AUTUMN

- Harvest seed as it ripens.
- Harvest leaves and stems as required.
- Cut back herbaceous types to stimulate new growth.

MID-AUTUMN

- Harvest seed as it ripens.
- Tidy away dead material.
- Protect tender herbs.
- Plant shrubby and herbaceous types until late autumn.
- Move tender container herbs under some protection.
- Divide herbaceous herbs until late autumn.

LATE AUTUMN

- Prepare ground for planting when conditions allow.

INDEX

Figures in *italics* indicate captions; main references to vegetables, fruit and herbs are in **bold** type.

alfalfa (*Medicago sativa*) 34, 35
alliums *19*
amphibians 10, 106, 107, 112, 115
Anemone blanda 109
anemones 107
animals
    and biological control 66
    overwintering 114
    plant-eating 15
aphids 65, 66, 67, 72, 75, 77, *105*, 116, 118, 130, 131
apples (*Malus domestica*) 45, 69, *121*, 123, 131, 147, 148, *148*, 152, 153
apricots 131, 147, *153*
Argyranthemum 98
argyranthemums *12*
asilke clover (*Trifolium hybridum*) 37
asparagus (*Asparagus officinalis*) 122, 131, 141
asters 93
azaleas 102

Bacillus thuringiensis (BT) 73, 77
bacteria 73
    anaerobic conditions 22
    nitrogen-fixing 15, *31*
    and temperature 18
bark 39, *39*, 62, 69, 76, 90, *97*, 148
barriers 66, 74, 75, *75*, 76, *77*, 81, 92, 123
basil (*Ocimum basilicum*) 48, 122, 130, 142–5
bats 72, 115, 119
bay (*Laurus nobilis*) *142*, 144
beans 15, *30*, 124, 126, *129*, 131, 135, 141
    broad (fava; *Vicia faba*) 34, 57, 123, 131
    dwarf 131
bees 10, 65, 107, 115, *115*, 116, *116*, 118, 119, 131
beetles *72*, 75, 77, 114, *114*
beetroot (beets; *Beta vulgaris*) 122, 126, 130, 131, 135
begonias *50*
beneficials 61, *61*, 64, *64*, 72–3, 76, *77*, *114*, *115*, *116*, *117*, 130

berries 147, 148, 152
biological controls 65, 66, 67, 72, *72*, *73*
birds 10, 11, *61*, 65, *66*, 67, 72
  attracting 106, 107, 115, 116, *116*, 117, 118, 131
  insect-eating 114, 119, 130, 131
  as pests *74*, 75, 76, *76*, *77*, 149
  woodland 107, 108
blackberries (*Rubus fruticosus*) 122, 148
blackcurrants (*Ribes nigrum*) 153
blackspot *68*, 69, 77
bluebells (*Hyacinthoides*) 107, 109, *109*
bonemeal 42, *42*, 88, 92, 103, 140
borage (*Borago officinalis*) 144
botrytis 128
brassicas *66*, 67, 74, 126, 130, 141
broccoli (*Brassica oleracea* Cymosa Group) 74, 124, 126, 130
Brussels sprouts (*Brassica oleracea* Gemmifera Group) 124, 126, 131, 132, 141
buckwheat (*Fagopyrum esculentum*) 35
bulbs 49, 57, 95, *95*, 96, *102*, *103*, 117
  naturalizing in a lawn 111, *111*
butterflies 10, *10*, 11, 106, 107, 115, *115*, 116, *117*, 118, 119, 131
butterfly bush (*Buddleja* spp.) 18, *46*, *117*

cabbages (*Brassica oleracea* Capitata Group) 57, 67, 74, 75, 122, 124, 126, 130, 131, 139, 141
caddis flies 113
calendar of care
    kitchen garden 156–7
    ornamental garden 154–5
    wildlife garden 154–5
calendula 57
Camellia 18
capping 40, *135*
carrot flies 75
carrots (*Daucus carota*) 67, 122, 126, 130, 131, 133, *140*, *141*
cauliflowers (*Brassica oleracea* Botrytis Group) 57, 122, 131, 141
celeriac (*Apium graveolens* var. rapaceum) 57

celery (*Apium graveleons*) 57, 131
chamomile (*Chamaemelum nobile*) 131, 145
chemicals 9, 22, 38, 63, 79, 106
cherries (*Prunus* spp.) 69, *91*, 123, 147
chives (*Allium schoenoprasum*) 131, 143, 145
chysanthemums 57
clematis 109
climbers 96, 115, 117, 128
  greenhouse 56
  planting *89*
  pruning out dead wood *47*
  renovating 91
  support 141
cloches 58–9, *59*, 76, *123*, 149, *149*
codling moths 75
companion planting *17*, *61*, *62*, *64*, 74–5, *75*, *121*, *122*, 126, 130–1, 136
compost 17, 24–5, *27*, *37*, 43, *65*, 88, 102, 103, *146*, 147
  heap 32–3, *32–3*, 42
  making 32–3, 106
  spent mushroom 28, *28*, 30, 33
  worm 30–1, 48
Comus alba 96
conifers 80, 91, 98, 100
coriander (cilantro; *Coriandrum sativum*) 142, 143, 144
Cotoneaster 69, *117*, 131
courgettes (zucchini; *Cucurbita pepo*) 67, *129*, 131
crab apples 123
cranberries (*Vaccinium macrocarpon*, *V. trilobum* and *V. vitis-idaea*) *153*
crocus 111
crop rotation 8, 11, *17*, 55, 66, 126, *127*
cucumbers (*Cucumis sativus*) 57, 67, 124, 131, 136
curry plant (*Helichrysum italicum*) 143
cuttings 50, *50*, *51*
  and disease 65
  hardwood 50, *51*, 52
  softwood 50, 52, 57
cyclamens 57

daffodil (*Narcissus*) 18, 111, *111*
dahlias 57
daisies *64*, *116*
dead-heading *45*, *64*, 93, *93*, 95, 116
digging 24–5, *93*
  benefits 24
  double 24–5, *25*
  single 24, *24*
  tips 25
dill (*Anethum graveolens*) 122, 130, 143
diseases *17*, *21*, 46, *47*, 55, 57, 61, 62, 63, *63*, 65, *65*, 68–9, 74, 77, 84, 88, 98, 126
  common 69
  life-cycles 64, 65
  soil-borne 122
dragonflies 107, 113
drainage 19, 21, 26, 32, 80, 101–2, 103, *103*, 129, 143
drought 18, 40, 41, 62, 64, *70*, 80, 117
Echinacea 111

elder (*Sambucus*) *46*, 80, 108
English marigolds (*Calendula*) 123
Essex (red merviot) clover (*Trifolium pratense*) 37
euphorbias 50

feeding 42–4, 91, 140
  containers 102, *103*
  herbaceous plants 93
fences 11, *81*, 96, 123
fennel (*Foeniculum vulgare*) *105*, 142
fenugreek (*Trigonella foenum-graecum*) 35
ferns 12, 108
fertilizers 70, 102
  lawn 86
  nitrogen-rich 23
  organic 42
  slow-release 113, 140
  synthetic 8, 9
figs (*Ficus carica*) 147
fish 113
Florence fennel (*Foeniculum vulgare* var. *duce*) 141
flowering currant (*Ribes sanguineum*) *18*
flowers 41, 56, 67, *79*, 103
  attracting wildlife 107
  of sulphur 22
  wild 110–11, 116, 117
food webs 15, 66, 72
foxglove (*Digitalis*) *13*, 109, *109*, 117, *131*, 143
freezing 141, 145, *145*, 152, 153, *153*
French marigolds (*Tagetes patula*) 123, 130
fritillaries 111
frogs 106, 107, 112, 113, 114, 119
frost 12, 13, 19, 54, 62, 70, *149*, *167*
fruit 45, 65, 67, 76, 103, *105*, 107, *116*, *121*, 122, 123, 146–7
  bushes 146–7, 146, 148, 148
  cages 149
  harvesting and storing 152–3
  protecting 149, *149*
fuchsias 57
fungal diseases *64*, 65, 70, 72, 77, 86, 138
fungi 22, 73, 77, 88, 105
fungicides 9, 50, 77

garlic (*Allium sativum*) 126, 131, 141, 145
ginger 144, 145
gladioli *95*
globe artichokes (*Cynara cardunculus* Scolymus Group) 141
gooseberries (*Ribes uva-crispum*) 123
grass 67, 72, 82–7
  clippings 39, *39*, 84, *84*, 85
  ornamental 92
  types 82
grassland *12*, *106*
greenhouse
  beneficial predators *72*, 73
  cultivation 56–7
  environment 54–5, *64*
  propagators 136
  vegetables *137*
ground beetles 65
harvesting *140*, 141, 144, *144*, 152, *152*

hawthorn (*Crataegus*) 108
hazel (*Corylus avellana*) 80, 107
hedgehogs 106, 107, 115
hedging 13, 81, 91, 96, 107, 115, 116, 117, 123
hellebores 48, *109*
herbs 41, 102, 103, 116, 122
  growing 142–3
  harvesting and storing 144–5
Himalayan primulas 102
holly (*Ilex*) 107, 108
honeysuckle (Lonicera) 80, 109
hoverflies 72, 118, 119
humidity 12, 55, 57, 67, 128, 136, 139
humus 19, 23, 39, 42
hydrangeas 22, *65*, 109
hyssop (*Hyssopus*) 130

insecticides 64, 75
insects 10, 17, 65, 75, 76, *76*, *79*, 103, 105, 106, 107
  attracting 116, *116*, 117, 118
  beneficial *61*, 64, 65
  over-wintering 111
  pollinating *12*, 75, *96*, 115, *115*, 118, 130
insulation 55, *55*
iris 111

Jerusalem artichokes (*Helianthus tuberosus*) 141

kale (*Brassica oleracea* Acephala Group) 126
kiwi fruits 123
kohl rabi (*Brassica oleracea* Gongylodes Group) *124*, 131

lacewing hotels *114*, 115
lacewings 115
ladybirds 72, *73*, *105*, 115, *116*
lavender (*Lavandula*) 18, 96, 116, 119, 131, 144
lawns 67, 79, 82–7, 107
leaf mould 28, 88, 91, 103
  pit 28, *29*, 107
leafhoppers 75
leaves *84*
leeks (*Allium porrum*) 131
legumes 15, 23, *30*, 34, 126
lemon balm (*Melissa officinalis*) 143, 145
lemon thyme 145
lettuce (*Latuca sativa*) 67, 124, 126, 131, 132, 133, 139
Leucojum 111
light levels 12, 54, 57, 62
lilacs (*Syringa*) 91, 109
lilies *95*
liming 18, 19, *22*, 23
liquid feeds 43, *43*, 102
lizards 107, 115

manure 17, *25*, 103
  animal *8*, *28*, 33, 42, 43
  farmyard 28, *43*
  green manures 25, *27*, 28, 34–5, *34*
  well-rotted 24, *31*, 42, 88, *146*, 147
marigolds *61*, *74*, 123, 130
marrows (*Cucurbita pepo*) 141, *141*
melons (*Cucumis melo*) 57, 124
midges 65, 72
mildew *68*, 69, 77
mint (*Mentha*) *142*, 143, 145
mites 72, 75
mosses 12, *84*, 87, 108
moths 75, 119
mowers 38
mowing *82*, 84, 109
mowing edge *86*
mulches, mulching 26, *27*, 39, *39*, 40, 43, 88, 91, *97*, 99, 108, 147

nasturtiums 123, *130*, 131
nectarines (*Prunus persica nectarina*) 147, *153*
nerines *95*
newts 107, 114
nitrates 15, 23
nitrogen 14, 15, 20, 23, *23*, 34, 42, 44, 69, 132
nutrient cycles 14, *15*, 20, 22, 26, 42, *42*
nutrients 17, *17*, 19, 30, *31*, *32*, 39, 65, 126, 138
  deficiency 42, 62–3, *70*, 88
  plant 17, 18, 42, 44
  and roots 18, 20

oaks (*Quercus*) 48, 108
onions (*Allium cepa*) 74, 75, 123, *124*, 126, 130, 131, 141
oregano (*Origanum*) 123, 145
oxygen 20, 22, 24, 41, 42, 48, 113

Pachysandra terminalis 18
parasites 72
parasitic nematodes 73
parasitoids 65, 72
parsley (*Petroselinum crispum*) 130, 144, 145
parsnips (*Pastinaca sativa*) 126, 131
peaches (*Prunus persica*) 45, 57, 122, 131, 147, 152, *153*
pears (*Pyrus communis*) 69, *121*, 147, 148, *150*, *151*, 152, *152*
peas (*Pisum sativum*) 34, 124, 126, *128*, 131, 135, *140*
peas, sugar snap 123
peat 40, 106
pelargoniums 57
peppers (*Capsicum* sp.) 103, 136,
pesticides 8, 9, 10, *48*, 63, 67, 77, 106, 115
pests *17*, *21*, 27, *38*, 45, 55, 57, 61–7, 72, 74–7, 88, 98, 123, 126, 149
  common 67
  life-cycles 64, 65
  soil-borne 122
Phacelia (*Phacelia tanacetifolium*) 35
pheromones 75, *75*, 76, 115
phlox 50
phosphates 23, 42, 44
phosphorus 12, 43
plant health 60–77
  beneficial predators 72–3
  diseases 68–9
  disorders 70–1
  other control methods 74–7
  pests 66–7
  preventing problems 64–5
  why plants get sick 62–3
plums (*Prunus domestica* and *P. salicina*) 69, 123, 147, 148, 153
poached egg plant (*Limnanthes douglasii*) *99*
poisons 70
pollen 48
pollination 65, *79*, 147
pollution 63, 70, 122
ponds 72, 107, 117
  wildlife 112–13
pot marigolds (*Calendula officinalis*) 130
potash 44
potassium 14, 42, 43
potatoes (*Solanum tuberosum*) *26*, *27*, 67, 122, 126, 131
pots 100, 101, 102, 116, 117, 136, *136*, *137*, 143, *143*
potting mix 48, *48*, 57, 67, 80, *97*, *100*, *101*, 102, *103*, 113
  making 103
  propagation *137*
primroses (*Primula vulgaris*) *108*, 109
propagators 136, *136*
protozoans 73, *73*
pruning 45–7, 91, 99
  fruit trees/shrubs 147, 150–1
pumpkins (*Cucurbita maxima*) 131, *141*
purple smoke bush (*Cotinus*) *46*
Puschkinia 111
Pyracantha 69, 117
pyrethrum 75, 77

quince (*Chaenomeles*) *88*
quince (*Cydonia oblonga*) 152

rabbits 65, *66*, 67, 75, 76, *77*, 90, 91
radishes (*Raphanus sativus*) 57, 131, 132
rainfall 12, 40, 41, 65
rainforests 79, 80
raking *24*, 34, *34*, *83*, 84, 87, *97*, *135*
raspberries (*Rubus idaeus*) 122, 123, 131, 148, 152, 153
recycling 8, 11, *32*, *37*, 40, 41, *65*, 81
repellents 66, 67
rhizomes 49, 144
Rhododendron 18
rhododendrons 23, 38, 102, 109
rhubarb (*Rheum* x *hybridum*) 57, 122
root flies *38*
roots
  and air 18, 20
  and animal manure 42
  growth 42
  and nutrients 18, 20
  and shoots 20
  surface 41
  and water 12, *14*, 18, 20
Rosa glauca 46
Rosa rugosa 123
Rosaceae family 69
rosehips *105*, 123
rosemary (*Rosmarinus officinalis*) 116, 144, *144*, 145
roses 45, *64*, *68*, 69, 98, 99, *116*
  beds 98
  bush 69
  shrub 123
rowan tree (*Sorbus*) 131
Rubus 46
Rubus cockburnianus 46
ruby chard *140*
rye 34, 35

sage (*Salvia officinalis*) 18, 143, 144
salad leaves 173
salt toxicity 63
sawflies 77
scabious (*Astrantia*) *116*
scarecrows *61*, *77*
scarification 48, 84–5
screens 74, 75
seaweed extract 102, 103

seaweed meal 42, *42*
secateurs (hand pruners) 45, 46, *46*
secondary cultivation 25
seed trays 136, *136*, 137
seedlings, protecting 135
seeds
  and disease 65
  propagation from 48, *48*, *49*
  sowing in open ground *95*, 134–5
shading 54, *54*, 55, 57
shadow projections 13
shady areas *13*
shelter, providing 114–15
shoots 20
  growth 42
  and layering shrubs *52*
  pruning *46*, *57*
  side *195*, *203*
shrubs 25, 79, 115, 117
  in borders 99
  deciduous 57, 91, 109
  evergreen 57, 91, 114
  layering 52, *52*
  ornamental 45, 103
  planting *88*, 98
  protection 90–1, *91*
  pruning 46, 150–1
  renovating 91
  shelters 90
  supporting 90, *90*, 148, *148*
  understorey 107
slopes 81
slugs 27, 67, *74*, 75, 76, *76*, 92, 119
snails 27, 67, *74*, *76*, 92, 119, 131
snow 12
snowdrops (*Galanthus*) 109, 111
soil 16–35
  acidic 23, 63
  acidifying 22
  alkaline 23, 63, *65*
  biodiversity 11
  chemistry 22–3
  clay 12, *18*, 19, *19*, 20, 24, *27*, 31, 40, *52*
  compaction 21, 26, *87*
  conditioners 28–31
  fertility 9, 11, 19, 88
  liming *22*, 23
  loamy *18*, 19, *19*, 40, 103, 134
  nutrient test *23*
  pH 13, 22, *22–3*, 24, 42, 122, 126, 142
  preparation 142
  sandy 12, 18, *18*, 20, 31, 40

  silty 18, *18*, 19, 20
  structure 20–1
  temperature 12, 26, 56, 132, 136
  types of 18–19
sowing
  in the open *95*, 134–5
  station 135, *135*
  successional 133, *133*
  under glass 136–7
spent mushroom compost 28, *28*, 30, 33
spinach (*Spinacia oleracea*) 131, 133, 172, *172*
sprinklers *40*, 41, *41*
squashes (*Cucurbita maxima*) 126, 131, 141
squirrels *66*
stag beetles 107
staking 90, *90*, 148, *148*, 149
station sowing 135, *135*
stems
  cuttings 50, *51*
  pruning 46, *46*, *47*
stinging nettles 43
storage 141, *141*, 144–5, 152–3, *152*, *153*
strawberries (*Fragaria* x *ananassa*) 57, *64*, 67, 122, 131, *149*, 152, *152*, 153
subsoil 18, *19*, 24, 25, *25*
sulphur 44, *50*, 69, 77
sunflowers (*Helianthus annuus*) *116*, 119, 131
sweet chestnut (*Castanea sativa*) 80
sweet peas (*Lathyrus odoratus*) 57, *105*
sweetcorn (*Zea mays*) 126, 131, *133*
Swiss chard (*Beta vulgaris* Cicia Group) 130

tagetes 57
tansy (*Tanacetum vulgare*) 131, 142
tarragon 145
temperature 12, 18, 54, *54*, 56, 62, 64
thermostats 55, 136
thinning 92, 138, *138*
thrips 75
thyme (*Thymus vulgaris*) 48, 142, *143*, 144, 145
tiger worms *30*
toads 107, 113, 114
tobacco mosaic virus 65
tomatoes (*Lycopersicon esculentum*) 56, 57, 59, 65, 67, 75, 103, 124, *125*, 126, 130, 131, 136, 142
top-dressing 26, 42, *42*, 43, 57
  lawn 84, 85, *85*, 87
topsoil 18, *19*, 24, 25, *25*, 55, 85, *87*
trace elements 23
training systems 148, *150*, *151*
transplanting 18, 93, 139, *139*
traps 66, 67, *74*, 75, *75*, 76, 92
trees 67, 68, 69, 79, 115, 116–17
  deciduous 117
  dwarf 98
  evergreen 114
  flowering 91
  fruit 103, 122, 123, 146–7, 148, 150–1
  guards 76
  planting 88–9
  protection 90–1
  shelters 76, *77*, 90
  staking 90
  and wind *70*
  woodland gardens 108
trellises 123, *190*
tubers *27*, 49, 95, *182*
  propagating *50*
tubs 116, 117
turf
  aeration 85
  problems 84
turfing 83, *87*
turmeric 144
turnips (*Brassica oleracea* Rapifina Group) 131, 177, *177*

vegetables 122, 123
  bulb 156–60
  in cold frames *58*
verbenas *12*
Viburnum 117
Viburnum opulus 96
Viburnum tinus 91
viruses 73
voles *77*, 90

wake robin (*Trillium grandiflorum*) 109
walls 96, 115, 129
  dry-stone *80*, 107, 117
  and microclimates 13
wasps 65
waste
  production 11
  reduction 11
  vegetable industrial 30
water 12, *14*
  availability 12, 17
  pest control 75
  and roots 12, *14*, 18, 20
  soil 22
  tapwater 112
  waste ("grey") 40
watering 40–1, 57, 102, 128, *137*, 140
waterlogging 12, 18, 19, 20, 27, 39, 40, 62, 71
weeding, weed control 38–9, 86, 91, 97, 99, 140
weeds *21*, 24, 25, 26, *27*, 33, *37*
  and block planting 128, *128*
  ephemeral and annual 38
  and intercropping 132
  perennial 38, *83*
  turf 84
white alyssum (*Lobularia maritima*) 131
whiteflies 65, 75
wildlife gardens 104–19
  calendar of care 244–5
  plants to attract wildlife 116–19
  providing shelter 114–15
  what is it? 106–7
  woodland 108–9
willow (*Salix*) 18, *37*, *46*, 80, *81*
wind 62, 65, 70, *70*, 90, 103
wood ash 42
wood sage 107
woodland gardens 108–9
wormery *28*, 29, 43
worms 10, 17, 20, 26, 30–1, 30
  casts 30, 85
  tiger *30*

yarrow (*Achillea*) 130
yucca 18

zinc 23, 44

ACKNOWLEDGEMENTS

Authors' Acknowledgements

We would like to thank the following people who worked as models for the photography: Joe Lovell, Debbie Hart for her willow weaving skills, Jane Dobson for her graft unions and Ian Gandy and Stewart Brown for the lawn maintenance shots. A big thank you to Writtle College and their staff for allowing the use of their grounds and glasshouses to take many of the photographs.

Publisher's Acknowledgements

The publisher would like to thank the following for kindly allowing photography to take place in their gardens:
t = top   b = bottom   l = left   r = right   c = centre

The Centre for Alternative Technology, Wales 8t; 32t; 33t; 61br; 104;114c; 121bc and br; 122t; 130t; 141bl; **Dean Court, Dorset** 16; 26b; 36; 37c; 65bl; **Edmondsham Manor, Dorset** 6; 9t; 10bl; 12t; 30t; 63t; 76c; 78; 92t; 96t; 106t; 109tl; 123b; 129tl; 137bl; **Fardel Manor, Devon** 30c; 55t; 62bl; 101tl; 107b; 113b; 115tl; 140c; 142bl; **Valerie Ferguson, Bath** 9b; 31t; 38bl; 42tr; 43t (all pictures in sequence); 56t and c; 72t; 75br; 77tl and cl; 114t; 125b; 129tr and b; 133b; 146t; **HDRA, the Organic Organisation, Coventry** 4t; 5bc; 8b; 11; 12br; 14c; 28tr; 56b; 60; 61bl; 73br; 74t, c and b; 77br; 105c; 107tl and tr; 109tr; 114b; 115tr; 122b; 123t; 125t; 130b; 141t; **Adrian Penrose** 40t; 75t; **RHS Rosemoor, Devon** 7c; 10br; 17bl; 21br; 62t; 80; 96b; 108t; 126; 128t and b; 152br; 149br; 150t; 160; **RHS Wisley, Surrey** 62b; 120; **Peter Robinson** 12bl; 105r; **The Royal Society for the Protection of Birds, Bedfordshire** 72c; **Writtle College, Chelmsford** 5tl and b; 14b; 17br; 19tr; 26t; 29 (all pictures); 37br; 40br; 76t; 77bl; 81 (all pictures in sequence); 82; 84t and bl; 85bl and br; 86c; 87 (all pictures in sequence); 91t; 96c; 100t; 100–1 (all pictures in sequence); 101tr; 108b; 109b; 110t, bl and br; 111b; 132t.

The publisher would also like to thank the following picture libraries for kindly allowing their photographs to be reproduced for the purposes of this book:

The Garden Picture Library 67r (three from top; Sunniva Harte); 68t (Janet Sorrell); 69 (ten from top; Lamontagne); **Holt Studios International** 67l (four from top; Nigel Cattlin); 67l (thirteen from top; Nigel Cattlin); 67r (four from top; Nigel Cattlin); 67r (eight from top; Nigel Cattlin); 69 (seven from top; Nigel Cattlin); 69 (eleven from top; Nigel Cattlin); 71 (four from top; Nigel Cattlin); 118 (three from top; Primrose Peacock); 118 (four from top; Nigel Cattlin); 119 (six from top; M. Szadzuik/R. Zinck); **NHPA** 66br; 67tl; 67l (seven from top); 67l (eight from top); 67l (nine from top); 67l (ten from top); 67r (five from top); 69 (three from top); 67r (two from top); 67r (six from top).

Plants that Contain DMT

*The percentages of alkaloid content in plants vary greatly from one plant to another depending upon growing conditions, time of harvest, location, etc. *

	Arundo donax	Entire Plant contains 5-MeO-DMT (Shulgin, TIHKAL) Flowers contain DMT, 5-MeO-DMT, and 5-MeO-NMT (Shulgin, TIHKAL) Roots contain DMT, 5-MeO-DMT, 5-MeO-NMT, Bufotenine, bufotenidine, dehydrobufotenidine (Shulgin, TIHKAL)
	Desmanthus Illinoensis	Root contains DMT - 0.200% (Ott) Root Bark contains DMT - 0.340% (Ott)

	Phalaris arundinacea	Leaves of contain DMT, 5-MeO-DMT, and related compounds (Smith 1977)
	Phalaris aquatica syn. Phalaris tuberosa	Leaves and seedlings contain DMT, 5-MeO-DMT, and related compounds (Smith 1977) DMT - 0.100% (erowid) 5-MeO-DMT - 0.022% (erowid) 5-OH-DMT - 0.005% (erowid)

	Phragmites australis	DMT in roots (Ott)

	Delosperma acuminatum	DMT, 5-MEO-DMT (Trout's Notes)

	Delosperma cooperi	DMT, 5-MEO-DMT (Trout's Notes)
	Delosperma ecklonis	DMT (Trout's Notes)
	Delosperma esterhuyseniae	DMT (Trout's Notes)
	Delosperma hallii	5-MEO-DMT (Trout's Notes)
	Delosperma harazianum	DMT, 5-MEO-DMT (Trout's Notes)
	Delosperma harazianum Shibam	DMT (Trout's Notes)
	Delosperma hirtum	DMT (Trout's Notes)
	Delosperma lydenbergense	DMT, 5-MEO-DMT (Trout's Notes)
	Delosperma nubigenum	5-MEO-DMT

		(Trout's Notes)
	Delosperma pageanum	DMT, 5-MEO-DMT (Trout's Notes)
	Delosperma pergamentaceum	Traces of DMT (Trout's Notes)
	Delosperma tradescantioides	DMT (Trout's Notes)
	Petalostylis cassioides	0.4-0.5% tryptamine, DMT, etc. in leaves and stems (Johns et al 1966)

	Acacia acuminata	Up to 1.5% alkaloids, mainly consisting of tryptamine in leaf (Lycaeum)

	Acacia alpina	Active principles in leaf (M.Bock)
	Acacia angustifolia	Psychoactive Tryptamines (Rätsch 2004)
	Acacia angustissima	β-methyl-phenethylamine (Glasby 1991) NMT and DMT in leaf, 1.1-10.2 ppm (McSweeney et al. 2005)
	Acacia auriculiformis	5-MeO-DMT in stem bark (Lycaeum)
	Acacia baileyana	0.02% tryptamine and β-carbolines, in the leaf, Tetrahydroharman (Ott)
	Acacia berlandieri	Amphetamine, Methamphetamine, N,N-Dimethylamphetamine, 4-Methoxyamphetamine, 4-Hydroxyamphetamine (Shulgin 2001) DMT and others (Phytochem. 1998)
	Acacia catechu	DMT and other tryptamines in leaf, bark (Trout's Notes)
	Acacia colei	DMT (Dr. Karl and abc.net.au 2005)
	Acacia complanata	0.3% alkaloids in leaf and stem, N-methyl-tetrahydroharman with traces of tetrahydroharman (Johns et al. 1966)
	Acacia confusa	N-chloromethyl-N,N-dimethyltryptamine, N-methyltryptamine, N,N-dimethyltryptamine, and

		N,N-dimethyltryptamine-N-oxide (Quinn et al. 2006)
	Acacia cultriformis	Tryptamine in the leaf and stem (Trout's Notes)
	Acacia farnesiana	Tryptamines and phenethylamines (Lyceaum) N-Methyl-Beta-Phenethylamine (Duke 1992)
	Acacia laeta	DMT in the leaf (Trout's Notes)
	Acacia longifolia	0.2% tryptamine in bark, leaves, some in flowers, phenylethylamine in flowers (Hegnauer 1994) DMT in plant (Lyceaum)
	Acacia maidenii	Bark of A. maidenii contains 0.6% of N-methyltryptamine and DMT in the proportions approx. 2:3 (Fitzgerald & Sioumis 1965)
	Acacia melanoxylon	DMT in the bark and leaf, less than 0.02% total alkaloids (Hegnauer 1994)

	Acacia mellifera	DMT in the leaf (Trout's Notes)
	Acacia nilotica	DMT in the leaf (Trout's Notes)
	Acacia obtusifolia	0.4 to 0.5 % DMT in the dried bark (Csiro 1990)
	Acacia oerfota	Less than 0.1% DMT in leaf (Ott)
	Acacia phlebophylla	0.3% DMT in leaf, NMT (Trout's Notes)
	Acacia podalyriaefolia	Tryptamine in the leaf (Trout's Notes) 0.5% to 2% DMT in fresh bark, phenethylamine trace amounts (Hegnauer 1994)
	Acacia polyacantha	DMT in leaf (Trout's Notes)
	Acacia retinodes	DMT and MMT (www.factorey.ch/Eins.htm) less than 0.02% total alkaloids found (Hegnauer 1994)
	Acacia rigidula	DMT, NMT, tryptamine, amphetamines, mescaline, nicotine and others (Phytochem. 1998)
	Acacia sieberiana	DMT in the leaf (Trout's Notes)
	Acacia verek	Psychoactive (Rätsch 2004)
	Acacia vestita	Tryptamine, in the leaf and stem (Trout's Notes)

		less than 0.02% total alkaloids (Hegnauer 1994)

	Anadenanthera colubrina	seed pods contain dimethyltryptamine and the seeds bufotenin, bufotenin oxide, and oxide of dimethyltryptamine (GRANIER-DOYEUX 1965) Leaves, Bark, and Seeds contain DMT, 5-MeO-DMT and related compounds (Schultes 1977)
	Anadenanthera peregrina	seed pods contain dimethyltryptamine and the seeds bufotenin, bufotenin oxide, and oxide of dimethyltryptamine (GRANIER-DOYEUX 1965) Leaves, Bark, and Seeds contain DMT, 5-MeO-DMT and related compounds (Schultes 1977)
	Anadenanthera macrocarpa	seed pods contain dimethyltryptamine and the seeds bufotenin, bufotenin oxide, and oxide of dimethyltryptamine

		(GRANIER-DOYEUX 1965)
	Desmodium caudatum	Roots: 0.087% DMT, Bufotenine-N-oxide 0.03% (Trout's Notes)
	Desmodium gangeticum	DMT, 5-MEO-DMT, whole plant, roots, stems, leaves (Ott)
	Desmodium gyrans	DMT, 5-MEO-DMT, leaves, roots (Ott)
	Desmodium pulchellum	DMT, 5-MEO-DMT, whole plant, roots, stems, leaves, flowers (Ott)
	Desmodium racemosum	5-MEO-DMT (Ott)
	Desmodium triflorum	DMT-N-oxide, roots (Ott)
	Lespedeza bicolor var. japonica	DMT, 5-MEO-DMT in leaves and root bark (Ott)

	Mimosa Hostilis	Root Bark contains DMT - 0.31% to 0.57% (Schultes 1977)
	Mimosa scabrella	DMT in bark (Ott)
	Mimosa verrucosa	DMT (Schultes 1969)
	Mucuna pruriens	leaves, seeds, stems and roots contain L-Dopa, Serotonin, 5-HTP, and Nicotine, as well as N,N-DMT, Bufotenine, and 5-MeO-DMT (Erowid)
	Petalostylis labicheoides var. casseoides	DMT in leaves and stems (Ott)
	Diplopterys cabrerana	1.3% DMT (Shulgin, TIHKAL) Contains DMT (Ott)

	Horsfieldia superba	5-MeO-DMT and beta-carbolines (Jossang et al. 1991)
	Iryanthera ulei	5-MEO-DMT in bark (Ott)
	Osteophloem platyspermum	DMT, 5-MEO-DMT in bark (Ott)

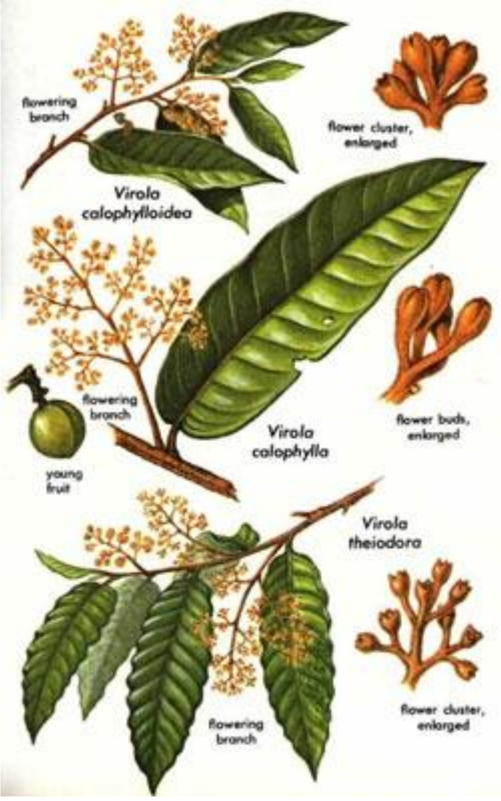 flowering branch Virola calophylloidea flower cluster, enlarged flowering branch young fruit Virola calophylla flower buds, enlarged Virola theiodora flowering branch flower cluster, enlarged	Virola calophylla	Leaves 0.149% DMT (Ott)
	Virola carinata	DMT in leaves (Ott)
	Virola divergens	DMT in leaves (Ott)
	Virola elongata	DMT, 5-MEO-DMT in bark and leaves (Ott)
	Virola melinonii	DMT in bark (Ott)
	Virola multinervia	DMT, 5-MEO-DMT in bark and leaves (Ott)

	Virola pavonis	DMT in leaves (Ott)
	Virola peruviana	DMT, 5-MEO-DMT in bark (Ott)
	Virola rufuta	Alkaloids in bark and root, 95% of which is 5-MeO-DMT (Shulgin, TIHKAL)
	Virola sebifera	DMT in bark (Ott)
	Virola theiodora	DMT, 5-MEO-DMT in bark, roots, leaves and flowers (Ott)
	Virola venosa	DMT, 5-MEO-DMT in roots and leaves (Ott)

	Psychotria carthaginensis	0.2% average DMT in dried leaves (Ott)
	Psychotria viridis	Leaves contain DMT - 0.10% to 0.66% (McKenna, 1984)

Arundo donax

Entire Plant contains 5-MeO-DMT (Shulgin, TIHKAL)

Flowers contain DMT, 5-MeO-DMT, and 5-MeO-NMT (Shulgin, TIHKAL)

Roots contain DMT, 5-MeO-DMT, 5-MeO-NMT, Bufotenine, bufotenidine, dehydrobufotenidine (Shulgin, TIHKAL)

Desmanthus Illinoensis

Root contains DMT - 0.200% (Ott)

Root Bark contains DMT - 0.340% (Ott)

Phalaris arundinacea

Leaves of contain DMT, 5-MeO-DMT, and related compounds (Smith 1977)

Phalaris aquatica

syn.

Phalaris tuberosa

Leaves and seedlings contain DMT, 5-MeO-DMT, and related compounds (Smith 1977)

DMT - 0.100% (erowid)

5-MeO-DMT - 0.022% (erowid)

5-OH-DMT - 0.005% (erowid)

Phragmites australis

DMT in roots (Ott)

Delosperma acuminatum

DMT, 5-MEO-DMT (Trout's Notes)

Delosperma cooperi

DMT, 5-MEO-DMT (Trout's Notes)

Delosperma ecklonis

DMT (Trout's Notes)

Delosperma esterhuyseniae

Delosperma
esterhuyseniae
aal

DMT (Trout's Notes)

Delosperma hallii 5-MEO-DMT (Trout's Notes)

Delosperma harazianum DMT, 5-MEO-DMT (Trout's Notes)

Delosperma harazianum
Shibam DMT (Trout's Notes)

Delosperma hirtum DMT (Trout's Notes)

Delosperma lydenbergense

DMT, 5-MEO-DMT (Trout's Notes)

Delosperma nubigenum 5-MEO-DMT (Trout's Notes)

Delosperma pageanum DMT, 5-MEO-DMT (Trout's Notes)

Delosperma pergamentaceum Traces of DMT (Trout's Notes)

Delosperma tradescantioides

DMT (Trout's Notes)

Petalostylis cassioides

0.4-0.5% tryptamine, DMT, etc. in leaves and stems (Johns et al 1966)

Acacia acuminata Up to 1.5% alkaloids, mainly consisting of tryptamine in leaf (Lycaeum)

Acacia alpina Active principles in leaf (M.Bock)

Acacia angustifolia Psychoactive Tryptamines (Rätsch 2004)

Acacia angustissima

β-methyl-phenethylamine (Glasby 1991)

NMT and DMT in leaf, 1.1-10.2 ppm (McSweeney et al. 2005)

Acacia auriculiformis

5-MeO-DMT in stem bark (Lycaeum)

Acacia baileyana

0.02% tryptamine and β-carbolines, in the leaf, Tetrahydroharman (Ott)

Acacia berlandieri

Amphetamine, Methamphetamine, N,N-Dimethylamphetamine, 4-Methoxyamphetamine, 4-Hydroxyamphetamine (Shulgin 2001)

DMT and others (Phytochem. 199😎

Acacia catechu

DMT and other tryptamines in leaf, bark (Trout's Notes)

Acacia caven

Psychoactive

Acacia colei DMT (Dr. Karl and abc.net.au 2005)

Acacia constricta

β-methyl-phenethylamine[12]

Acacia complanata 0.3% alkaloids in leaf and stem, N-methyl-tetrahydroharman with traces of tetrahydroharman (Johns et al. 1966)

Acacia confusa

N-chloromethyl-N,N-dimethyltryptamine, N-methyltryptamine, N,N-dimethyltryptamine, and N,N-dimethyltryptamine-N-oxide (Quinn et al. 2006)

Acacia cornigera

Psychoactive,[8] Tryptamines[27]

Acacia cultriformis

Tryptamine in the leaf and stem (Trout's Notes)

Acacia farnesiana

Tryptamines and phenethylamines (Lyceaum)

N-Methyl-Beta-Phenethylamine (Duke 1992)

Acacia filiciana Psychoactive

Acacia floribunda Tryptamine, phenethylamine,[33] in flowers[15] other tryptamines,[34] phenethylamines

Acacia georginae Psychoactive,[8] plus deadly toxins

Acacia greggii

N-methyl-β-phenethylamine,[12] phenethylamine[36]

Acacia harpophylla Phenethylamine, hordenine at a ratio of 2:3 in dried leaves, 0.6% total[6]

Acacia holoserica Hordenine, 1.2% in bark[6]

Acacia horrida

Psychoactive

Acacia implexa

Psychoactive

Acacia karroo
Psychoactive

Acacia kempeana
Psychoactive

Acacia kettlewelliae
1.5[6]-1.88%[38] alkaloids, 92% consisting of phenylethylamine.[6] 0.9% N-methyl-2-phenylethylamine found a different time

Acacia laeta DMT in the leaf (Trout's Notes)

Acacia lingulata
Psychoactive

Acacia longifolia

0.2% tryptamine in bark, leaves, some in flowers, phenylethylamine in flowers (Hegnauer 1994)

DMT in plant (Lyceaum)

Acacia macradenia
Tryptamine

Acacia maidenii

Bark of A. maidenii contains 0.6% of N-methyltryptamine and DMT in the proportions approx. 2:3 (Fitzgerald & Sioumis 1965)

Acacia mangium

Psychoactive

Acacia melanoxylon

DMT in the bark and leaf, less than 0.02% total alkaloids (Hegnauer 1994)

Acacia mellifera

DMT in the leaf (Trout's Notes)

Acacia nilotica

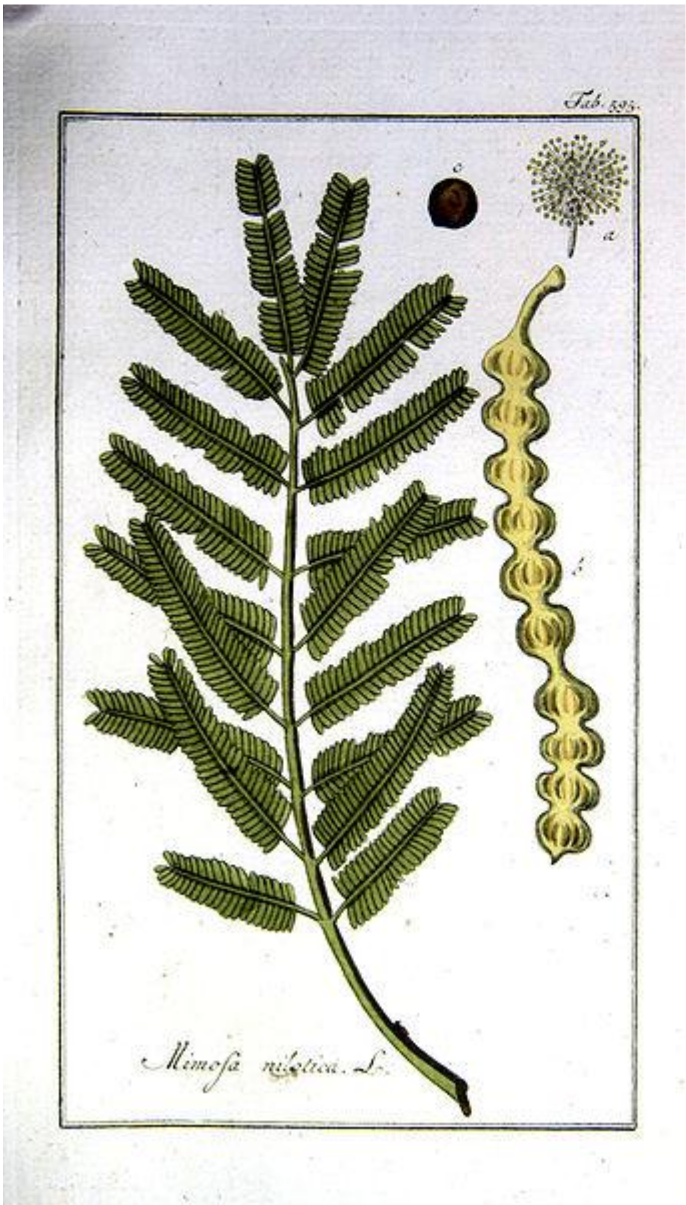
Tab. 595.
c
a
b
Mimoſa nilotica. L.

DMT in the leaf (Trout's Notes)

Acacia obtusifolia 0.4 to 0.5 % DMT in the dried bark (Csiro 1990)

Acacia oerfota Less than 0.1% DMT in leaf (Ott)

Acacia penninervis
Psychoactive

Acacia phlebophylla

0.3% DMT in leaf, NMT (Trout's Notes)

Acacia platensis
Psychoactive

Acacia podalyriaefolia

Tryptamine in the leaf (Trout's Notes)

0.5% to 2% DMT in fresh bark, phenethylamine trace amounts (Hegnauer 1994)

Acacia polyacantha

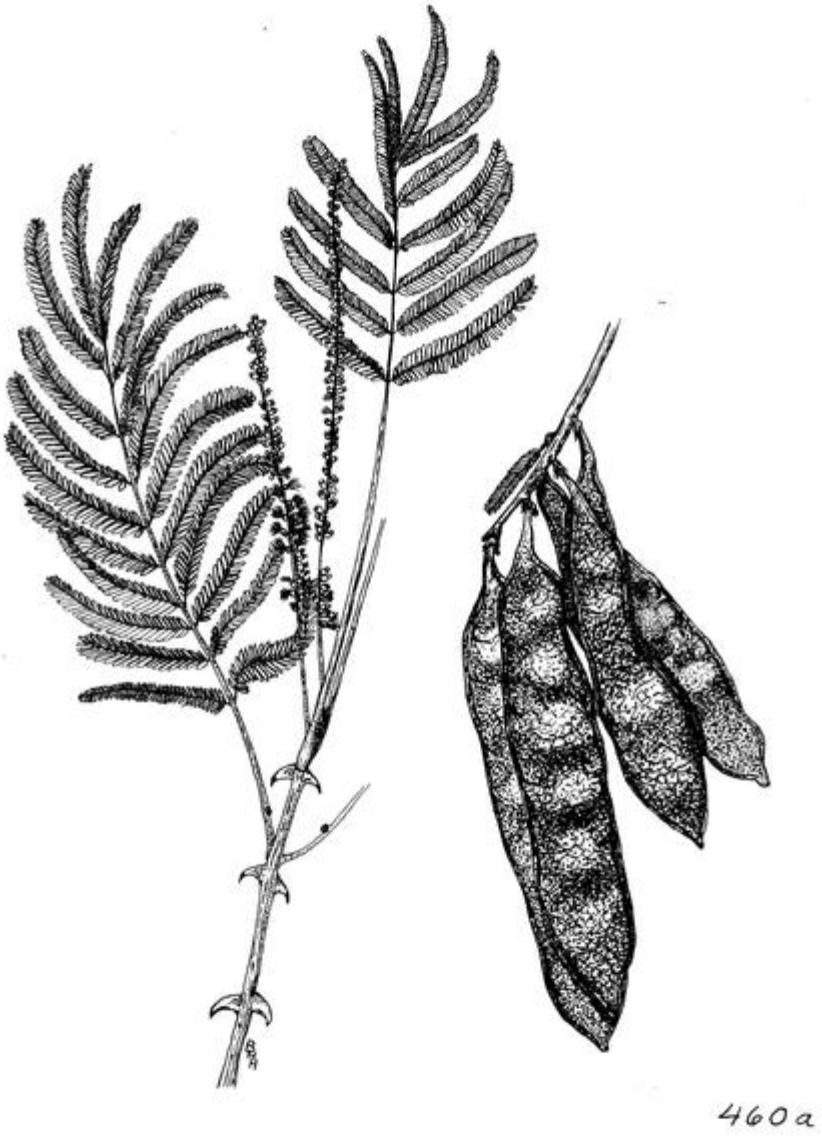

DMT in leaf (Trout's Notes)

Acacia pycantha

Psychoactive,[8] but less than 0.02% total alkaloids

Acacia retinodes

DMT and MMT (www.factorey.ch/Eins.htm)

less than 0.02% total alkaloids found (Hegnauer 1994)

Acacia roemeriana
β-methyl-phenethylamine

Acacia rigidula DMT, NMT, tryptamine, amphetamines, mescaline, nicotine and others (Phytochem. 199😎

Acacia salicina

Psychoactive[8][9] Ash used in Pituri.

Acacia sassa
Psychoactive

Acacia schaffneri

β-methyl-phenethylamine, Phenethylamine[36] Amphetamines and mescaline also found

Acacia schottii
β-methyl-phenethylamine

Acacia senegal
DMT, in the leaf

Acacia simplex

DMT and NMT, in the leaf, stem and trunk bark, 0.81% DMT in bark, MMT

Acacia taxensis
β-methyl-phenethylamine

Acacia tenuifolia
Psychoactive

Acacia tortilis

DMT, NMT, and other tryptamines

Acacia sieberiana DMT in the leaf (Trout's Notes)

Acacia verek Psychoactive (Rätsch 2004)

Acacia vestita Tryptamine, in the leaf and stem (Trout's Notes)

less than 0.02% total alkaloids (Hegnauer 1994)

Anadenanthera colubrina

seed pods contain dimethyltryptamine and the seeds bufotenin, bufotenin oxide, and oxide of dimethyltryptamine (GRANIER-DOYEUX 1965)

Leaves, Bark, and Seeds contain DMT, 5-MeO-DMT and related compounds (Schultes 1977)

Anadenanthera peregrina seed pods contain dimethyltryptamine and the seeds bufotenin, bufotenin oxide, and oxide of dimethyltryptamine (GRANIER-DOYEUX 1965)

Leaves, Bark, and Seeds contain DMT, 5-MeO-DMT and related compounds (Schultes 1977)

Anadenanthera macrocarpa seed pods contain dimethyltryptamine and the seeds bufotenin, bufotenin oxide, and oxide of dimethyltryptamine (GRANIER-DOYEUX 1965)

Desmodium caudatum Roots: 0.087% DMT, Bufotenine-N-oxide 0.03% (Trout's Notes)

Desmodium gangeticum DMT, 5-MEO-DMT, whole plant, roots, stems, leaves (Ott)

Desmodium gyrans DMT, 5-MEO-DMT, leaves, roots (Ott)

Desmodium pulchellum DMT, 5-MEO-DMT, whole plant, roots, stems, leaves, flowers (Ott)

Desmodium racemosum 5-MEO-DMT (Ott)

Desmodium triflorum

DMT-N-oxide, roots (Ott)

Lespedeza bicolor var. japonica DMT, 5-MEO-DMT in leaves and root bark (Ott)

mimosahostilis

Root Bark contains DMT - 0.31% to 0.57% (Schultes 1977)

Mimosa scabrella DMT in bark (Ott)

Mimosa verrucosa DMT (Schultes 1969)

Mucuna pruriens leaves, seeds, stems and roots contain L-Dopa, Serotonin, 5-HTP, and Nicotine, as well as N,N-DMT, Bufotenine, and 5-MeO-DMT (Erowid)

Petalostylis labicheoides var. casseoides DMT in leaves and stems (Ott)

Diplopterys cabrerana

1.3% DMT (Shulgin, TIHKAL)

Contains DMT (Ott)

Horsfieldia superba 5-MeO-DMT and beta-carbolines (Jossang et al. 1991)

Iryanthera ulei

5-MEO-DMT in bark (Ott)

Osteophloem platyspermum DMT, 5-MEO-DMT in bark (Ott)

Virola calophylla

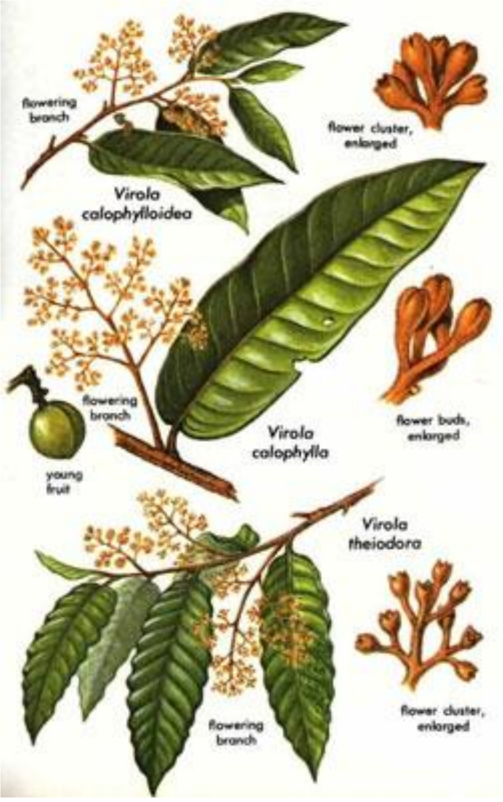
flowering branch
flower cluster, enlarged
Virola calophylloidea
flowering branch
Virola calophylla
flower buds, enlarged
young fruit
Virola theiodora
flower cluster, enlarged
flowering branch

Leaves 0.149% DMT (Ott)

Virola carinata DMT in leaves (Ott)

Virola divergens DMT in leaves (Ott)

Virola elongata DMT, 5-MEO-DMT in bark and leaves (Ott)

Virola melinonii DMT in bark (Ott)

Virola multinervia DMT, 5-MEO-DMT in bark and leaves (Ott)

Virola pavonis DMT in leaves (Ott)

Virola peruviana DMT, 5-MEO-DMT in bark (Ott)

Virola rufuta Alkaloids in bark and root, 95% of which is 5-MeO-DMT (Shulgin, TIHKAL)

Virola sebifera DMT in bark (Ott)

Virola theiodora DMT, 5-MEO-DMT in bark, roots, leaves and flowers (Ott)

Virola venosa DMT, 5-MEO-DMT in roots and leaves (Ott)

Psychotria carthaginensis

0.2% average DMT in dried leaves (Ott)

Psychotria viridis

Leaves contain DMT - 0.10% to 0.66% (McKenna)

Edited:02-04-2008

Uncover the truth,"It's really stranger then fiction".

Acacias

Acacia

Among the most valuable ornamental trees in the drier parts of the world. This massive genus of plants within the larger Legume family, contains over 1200 species of plants, 900 which are native to Australia, the remainder being native to the Americas, Africa and milder parts of Asia. Many provide shelter, shade and timber in dry regions where few other sizeable trees will grown.
Most prefer full sun and well drained soil. Most are propagated from seed, many also propagated from semi-ripe cuttings over heat. Typically germination is greatly improved by pouring boiling water over them and letting them to cool by soaking in cold water over 24 hours. Rubbing sandpaper over the seedcoats can also be done. Germination is rapid.

** photos taken on Jan 3 2011 @ Deerfield Beach Arboretum, Florida.*

* photos of unknown internet source

Many Acacias are psychoactive (marked ^ at end of text description) and contain the same chemical that makes people and animals dream.
http://en.wikipedia.org/wiki/List_of_psychedelic_plants (video posted beneath)

http://www.erowid.org/plants/acacia/acacia.shtml

Acacia acinacea (Gold-Dust Wattle)
A decorative spreading shrub with long arching branches, native to semi-arid parts of southeastern Australia. Some records include: 3 years - 10 feet; largest on record - 10 x 13 feet.
The oblong, obovate or rounded phyllode leaves, up to 1 x 0.5 inches, are deep blue-green.
Profuse, golden-yellow flowerballs, up to 0.2 inches across, are borne singly or paired from the leaf axils during late spring and spring.
They are followed by curved, twisted pods.
The twigs are angled.
The smooth bark is gray to dark brown.
Hardy zones 8 to 10 in full sun to partial shade. It is drought and heavy clay tolerant. Prune lightly after flowering to encourage new growth.

'Ruby Tips'
Bright red new foliage

Acacia abyssinica
A medium-sized tree, reaching a maximum height of 66 (rarely over 53) feet, that is native to Africa.
Some records include: fastest recorded growth rate - 3 feet; largest on record - trunk diameter of 2 feet.
The bipinnate leaves are up to 11 inches in length.
The puffy flowerballs are white.
Hardy zones 9 to 11 (tolerating as low as 15 F).

Acacia acradenia

A large shrub or small tree, reaching a maximum height of 25 (rarely over 13) feet in height, that is a widespread native to arid northern Australia. Some records include: 3 years - 5 feet.
The elliptical leaves are up to 6 x 1.2 inches.
The golden-yellow to orange flower spikes are up to 2.5 inches in length.
The twigs are orange-brown.
The red-gray to brown bark is smooth when young, later turning fibrous.

Acacia acuminata (Raspberry Acacia)
A fast growing, dense, rounded tree reaching a maximum size of 47 x 33 feet, that is native to southwestern Australia. Some records include: 10 months - 4.5 feet; 3 years - 10 feet; largest on record - trunk diameter of 1.5 feet. It is known to resprout after fire but rarely suckers otherwise. It is moderately long-lived, exceeding 50 years.
The linear to narrow-elliptic leaves, up to 10 x 0.4 (rarely over 7) inches, are mid-green.
The golden-yellow flower spikes are up to 1.3 inches in length.
Hardy zones 8 to 10, requiring mediterranean climates. It tolerates drought but not waterlogged soils.

Acacia adsurgens
A spreading shrub or small tree, reaching a maximum size of 23 (rarely over 13) feet in height, that is a widespread but rare native to the arid parts of northern Australia.
Some records include: 6 years - 8 feet with a trunk diameter of 1 inch.
The linear leaves, up to 8 x 0.2 inches, are glossy mid-green.
The cylindrical, golden-yellow flower spikes are up to 0.8 inches in length.
The twigs are reddish or light brown.
The fibrous bark is gray-brown peeling to reveal reddish-brown bark beneath.
Hardy zones 9 to 11 (partial dieback recorded at 16 F).

Acacia adunca (Wallangarra Wattle)
Also called Cascade Wattle. A rapid growing, large bushy shrub or small tree reaching a maximum size of 47 x 20 (rarely over 33) feet that is native to dry forests in far northeast New South Wales in Australia.
The narrow linear, phyllode leaves, up to 6 x 0.1 inches in length, are bright green to blue-green.
The profuse, showy, sweetly fragrant, golden-yellow flower balls, up to 0.3 inches across, are borne in long sprays during late winter and spring.
The slender twigs are red-brown.
Hardy zones 9 to 11 in full sun on well drained soil.

Acacia albida
A beautiful large tree similar in appearance to a mature Gleditsia-Honeylocust.
Some records include: 3 years - 23 feet; largest on record - 110 x 100 feet with a trunk diameter of 6 feet.
Hardy north to zone 9a, tolerating as low as 18 F. Roots have been found as deep as 270 feet.

Acacia alpina (Alpine Wattle)
A dense, low spreading shrub, rarely much taller to 8 feet and spreading as much as 33 feet, that is

native to far southeastern Australia.
The broad obovate phyllode leaves, up to 1.8 x 1.1 inches, are gray-green.
The cylindrical bright yellow flowers, up to 0.6 inches in length, are paired or borne singly along the stems during spring.
Hardy zones 8 to 10 tolerating as low as 8 F. It is also tolerant of snow.

Acacia amblygona (Fan Wattle)
A low spreading shrub native to eastern Australia.
Some records include: 3 years - 5 (rarely over 3) feet; largest on record - 8 x 10 (rarely over 3) feet.
The lower growing forms can valuable for use as groundcover.
The sharply pointed, triangular, phyllode leaves, up to 0.8 x 0.2 inches, are mid- green.
The singular, golden-yellow flower balls, up to 0.3 inches across, are borne winter into early spring.
Hardy zones 9 to 11

Acacia amoena (Boomerang Wattle)
A dense, rounded, large shrub reaching up to 13 feet that is native to highlands in southeastern Australia.
The thick, leathery, oblanceolate phyllode leaves, up to 3.4 x 0.7 inches, are gray-green.
The showy, intense, bright golden-yellow flower balls, up to 0.2 inches across, are borne in short sprays, up to 2.6 inches in length, during late winter and early spring. They are followed by dark brown pods, up to 3.5 x 0.2 inches.
The twigs are reddish-brown.
Hardy zones 8 to 10

Acacia ampliceps (Salt Wattle)
A fast growing, suckering, bushy, semi-pendulous, spreading, large shrub or small tree, reaching a maximum height of 30 feet, that is a widespread native to northwestern Australia. Some records include: 2 years - 9 feet; 3 years - 15 feet.
It makes a great windbreak and is highly valuable for reclaiming salty soil.
The linear to narrow lance-shaped leaves, up to 10 x 1.2 inches, are bright green.
The creamy-white flower balls are borne on racemes up to 4 inches in length.
The twigs are yellowish.
Hardy zones 9 to 10, thriving in hot arid climates. Freezes to ground at 18 F

Acacia amythethophylla
A bushy shrub or small tree, reaching a maximum height of 17 feet, with a tiny natural range of sand dunes around Shark Bay on the central part of Australia's west coast. Some records include: 7 years - 17 feet.
The oblanceolate leaves, up to 4.8 x 0.6 inches, are blue-green.
The pale yellow flower balls are borne on racemes up to 1.4 inches in length.

Acacia aneura (Mulga)
A long lived, moderate growing, evergreen shrubby tree up to 25 feet that is native to much of the dry interior of Australia.
Some records include: fastest recorded growth rate - 4 feet; 4 years - 12 feet; 10 years - 33 feet with

a trunk diameter of 4 inches; largest on record - 60 x 23 feet with a trunk diameter of 1 foot. Mulga is long lived, exceeding 240 years.
Very deep rooted enabling it to survive in arid climates, a 4 foot tall tree was recorded to have a 10 foot taproot.
The flat, narrow, linear phyllode leaves, up to 5 x 0.4 (rarely over 3) inches in length, are gray-green to blue-green above, silvery beneath.
The profuse, golden-yellow flowers are borne on dense spikes, up to 1.2 inches in length, during spring and summer, irregularly the rest of the year.
The fissured bark is dark gray.
Hardy zones 9 to 10 tolerating as low as 15 F, it thrives in climates with 8 to 20 inches of rainfall per year.

Acacia angustissima (Fern Acacia)
A thornless, fast growng large shrub native from Missouri south to Costa Rica, reaching around 8 x 6 feet though may be a small tree more than double that.
Some records include: 2 years - 6 feet; 2.5 years - 17 feet with a trunk diameter of 2.3 inches; largest on record - 23 feet. In cold climates where it grows as a perennial, it can be used for groundcover.
The attractive, finely-textured, fern-like foliage is mid-green.
The white, ball-shaped flowers are borne during summer and fall.
The flowers attract butterflies
Hardy zones 9 to 10 as a shrub, zone 6 to 8 as a perennial, tolerating as low as -10 F. It thrives in full sun to partial shade on just about any fertile, well drained soil. Drought tolerant.

Acacia aphylla (Leafless Rock Wattle)
A leafless but attractive shrub reaching up to 10 x 10 feet that is native to a tiny area just a little bit east of Perth in Australia where it is endangered. It looks like a succulent rather than an Acacia. This Acacia bears no leaves.
The spiny, silvery to gray-green branches are profusely covered by showy, bright yellow flower balls borne singly during spring.
They are followed by linear pods, up to 3.5 inches in length.
Hardy zones 9 to 10 in sun to partial shade on very well drained soil.

Acacia arabica (Egyptian Acacia)
A medium size tree native to a vast area from Africa to southern Asia.
Some records include: 30 years - 60 feet with a trunk diameter of 2.6 feet; largest on record - 82 feet with a trunk diameter of 5.2 feet.
Hardy north to zone 9b

Acacia aroma
A thorny, spreading, small tree, reaching a maximum height of 20 feet, that is native to Bolivia, Peru, Paraguay and Argentina. It is closely related to Acacia macrantha.
Some records include: 2 years - 3 x 6 feet; 9 years - 13 feet.
The pinnate leaves, up to 5.5 inches in length, are composed of 48 to 100 linear leaflets, up to 0.2 x 0.03 inches in size.
The pale yellow flower balls, up to 0.4 inches across, are borne from the leaf axils.

The twigs are dark purplish-brown.
The dark gray to dark brown bark is smooth at first, later becoming shallowly furrowed.
Hardy zones 9 to 11 (slight damage at 16 F).

Acacia ashbeyae
A very attractive, rounded, dense, spreading shrub, reaching a maximum height of 6.5 feet, that is native to southwestern Australia. Some records include: 2 years - 5 x 8 feet; 3 years - 6 x 10 feet.
The very narrow leaves, up to 3.5 inches in length, are gray-green.
The flowers are bright yellow.
Hardy zone 9 to 10 (tolerating 19 F or slightly colder) in mediterranean climates, preferring sandy soil.

Acacia atramentaria
A flat-topped tree, reaching a maximum height of 40 feet, that is native to northern Argentina. Some records include: 3 years - 6 feet; 8 years - 15 feet.
The bark is furrowed and dark gray.
Hardy zones 9a to 10 (fully hardy to 16 F).

Acacia aulacocarpa (Golden Flowered Salwood)
A spreading tree, reaching a maximum height of 50 feet, that is native to the central and northern parts of Australia's east coast.
Some records include: 3 years - 10 feet; largest on record - 16 inch trunk diameter.
The leaves, up to 8 x 1.5 (rarely over 5) inches, are blue-green.
The bright golden-yellow flower spikes are up to 2.4 inches in length.
The bark is smooth.
Hardy zones 9b to 10 (est) in subtropical to tropical climates.

Acacia auriculiformis
A moderately long lived, dense, rounded, large tree native to northern Australia, New Guinea and nearby parts of Indonesia. It is among the worlds fastest growing trees. Some records include: fastest recorded growth rate - 25 feet; 2 years - 20 feet; 3 years - 41 feet with a trunk diameter of 5 inches; 5 years - 50 feet; 8 years - 60 feet; largest on record - 133 x 100 feet with a trunk diameter of 4 feet.
It is an excellent shade tree for subtropic and tropical areas worldwide.
The phyllode leaves, up to 12 x 2 inches, are glossy olive green.
The golden flowers are borne in profuse spikes as the tropical dry season begins.
They are profusely followed by contorted, leathery seed pods.
Hardy zones 10 to 12 in climates with average yearly rainfall of 30 inches or more.
Very soil tolerant, tolerating PH from 3 to 9^

** photo of unknown internet source*

Acacia baeuerlenii
An evergreeen shrub, reaching a maximum height of 13 feet, that is native to the central part of Australia's east coast. Some records include: 3 years - 11 feet.
The linear leaves are up to 6 x 0.5 inches.
The creamy-white flower balls, up to 0.4 inches across, are borne in clusters of 1 to 3 from the lead axils.
Hardy zones 9 to 10 (est) in warm temperate to subtropical climates.

Acacia baileyana (Cootamundra Wattle)
An elegant evergreen tree native to a small area around Cootamundra in the southern part of New South Wales in Australia. The outer branches gracefully droop.
Some records include: fastest recorded growth rate - 7 feet; 5 years - 17 feet; 20 years - 40 x 40 feet; largest on record - 50 x 70 feet with a trunk diameter of 3 feet. Short lived, it usually only lives 20 years.
It has been planted as a street tree in parts of California.
The feathery bipinnate leaves, up to 2 inches in length, are silvery-gray.
The profuse, bright yellow, ball-shaped flowers are borne in dense, axilliary racemes, up to 4 inches in length, during winter and early spring.
The bark is smooth and silvery-gray.
Hardy zones 8 to 10 tolerating as low as 15 F, can recover from 12 F.
It prefers full sun on humus-rich, well drained soil. Drought tolerant. Deer resistant. Pruning after flowering encourages new growth and extends the life of the plant.

'Purpurea'

Attractive purplish foliage.

Acacia bakeri (Baker's Wattle)
Among the largest of all Acacias, it can reach up to 170 x 82 feet with a trunk diameter of 5 feet with a dense rounded canopy. Due to habitat destruction within its natural range, trees exceeding even half that size are now extremely rare. It is slow growing, so the sizes of trees that greeted the settlers won't be around again for a long time. It originally grew in subtropical rainforests along a stretch of coastline on the center part of Australia's east coast.
The elliptical leaves, up to 5 x 1.3 inches, are deep green.
The creamy-yellow flowers are borne on racemes up to 4 inches in length.
They are followed by pods up to 8 x o.5 inches in size.
The close-grained hard, yellowish wood was once used for flooring and cabinetry. This tree is now in danger of extinction.
Hardy zones 9 to 10 in moist, humid subtropical climates.
Seed should be soaked in warm water for 4 hours or until they swell before sowing.

Acacia bancroftii (Bancroft's Wattle)
A very attractive deciduous tree reaching up to 20 feet, that is native to inland areas in central Queensland Australia.
The very large, sickle-shaped phyllode leaves, up to 9 x 5.5 inches, are blue-green.
The bright yellow, ball-shaped flowers are borne in very long racemes during late autumn and winter.
Hardy zones 9 to 11, thriving in warm dry climates.

Acacia berlandieri (Berlandier Acacia)
Also called Guajillo. A very attractive small semi-evergreen to evergreen tree that is native to hills of western Texas, south into Mexico. Some records include: fastest recorded growth rate - 3 feet; 3 years - 5 feet; largest on record - 25 x 25 feet with a trunk diameter of 6 inches.
The fernlike bipinnate, Silktree-like leaves up to 6 inches in length are deep green.
The ferny foliage lends a tropical appearance to the desert landscape.
The fragrant, creamy-white flowerballs are borne during spring.
The stems are armed in curved short thorns.
Berlandier Acacia is toxic to livestock, it is also psychoactive.
It is one of the most cold hardy Acacias, tolerating as cold as 0 F in dry climates.
Extremely drought tolerant, thriving in areas with rainfall as low as 6 inches in a year.

** photo of unknown internet source*

Acacia binervia (Coast Myall)
Also called Acacia glaucescens. A very attractive, long lived, fast growing, dense tree native to eastern Australia. Some records include: 3 years - 13 feet; largest on record - 66 x 35 feet. Also makes an excellent windbreak.
The curved, elliptic phyllode leaves are up to 7 x 2 (rarely over 5.5) inches in size.
The bright yellow flowers are borne in masses of narrow cylindrical, axilliary spikes, up to 2.3 inches in length, during early spring.
The deeply fissured and flaky bark is gray-brown to dark brown.
This plant is toxic to livestock.
Hardy zones 9 to 11 tolerating as low as 19 F in full sun and partial shade.

Acacia bivenosa (Two-Nerved Wattle)
A bushy, dense, rounded shrub, reaching a maximum height of 17 (rarely over 12) feet, that is native to northwest and north central Australia. Some records include: 2 years - 8 feet with a trunk diameter of 3 inches.
The oval to oblong leaves, up to 2 x 1 inches, are blue-green.
The showy, bright yellow flower balls are borne from the leaf axils.
Hardy zones 9b to 11 (plants freeze back to ground level at 19 F, resprouting vigorously).

Acacia bonariensis
An evergreen shrub that is native to Paraguay and northern Argentina.
Some records include: 2 years - 4.5 feet.
Hardy zones 9 to 10

Acacia boormannii
A showy, fast growing, bushy, slender large shrub to small tree, reaching a maximum size of 20 x 17 (rarely over 15) feet, that is native to subalpine zones in far southeastern Australia. It makes a great screen and windbreak.
The linear leaves, up to 3.5 inches in length, are gray-green.
The golden-yellow flower balls, up to 0.2 inches across, are borne on racemes, up to 1.6 inches in length. This plant is spectacular in bloom.
The branches are silvery. The bark is smooth and gray.
Hardy north to zone 8 and is snow and clay tolerant, moderately salt tolerant. Requires average yearly rainfall of 20 inches or more.

Acacia borleae
A shrub that is native to South Africa.
Some records include: 3 years - 6.5 feet.
Hardy zones 9 to 10 (tolerates 16 F with slight damage).

Acacia brachybotrya (Gray Mulga)
A moderate growing large shrub reaching a maximum size of 15 x 15 feet, that is native to semi-arid southeastern Australia. Some records include: 3 years - 12 feet.
The obovate leaves, up to 1.3 x 0.5 inches, are grayish-green to blue-green.
The red Manzanita-like bark looks amazing when exposed through pruning.
Hardy to 18 F

Acacia brachystacha (Turpentine Mulga)
An evergreen shrub to small tree, reaching a maximum size of 20 x 17 feet, that is native to central Australia. Some records include: 9 years - 14 feet.
The very linear leaves are up to 7 x 0.1 inches, are blue-gray to blue-green.
The bright yellow, narrow cylindrical flower spikes, up to 0.8 inches in length, are borne 1 or 2 from the leaf axils.
The gray bark is smooth becoming fissured with age.
Hardy zones 9 to 10 (tolerating 15 F) in semi-arid to desert climates.

Acacia brandeegeana
A small, thicket-forming tree, reaching a maximum height of 23 feet, that is native to Baja California.
Some records include: 9 years - 6 feet.
The pinnate leaves contain 4 to 14 tiny leaflets, up to 0.35 x 0.1 inches.
Hardy zones 9 to 10 (fully hardy to 16 F).

Acacia brunioides
A shrub reaching a maximum size of 6.5 x 5 feet, that is native to Australia's central east coast.
The densely crowded, narrow, needle-like phyllode leaves are up to 0.5 inches in length.
The showy large, globular, light yellow flowerheads are borne singly along the stems late winter into early spring.
Hardy zones 9 to 11. Prune to shape after blooming.

Acacia burkitttii (Pinbush Wattle)
An attractive, fast growing, dense, compact, large spreading shrub to small tree, reaching up to 33 x 15 (rarely over 15) feet, that is native to semi-arid parts of southern Australia.
Hardy zones 8 to 10 in full sun to partial shade preferring sandy soil. Very drought tolerant.
The long, linear, phyllode leaves, up to 8 inches in length, are green.
The oblong spikes of small yellow flowers are borne late winter and spring.
Hardy zones 8 to 10. An excellent landscape plant for dry climates.

Acacia buxifolia (Box-leaf Wattle)
An upright to spreading shrub reaching a maximum size of 13 x 9 feet that is native to the eastern Australian coast.
The leathery, narrow-elliptic, phyllode leaves, up to 1.5 x 0.3 inches, are gray-green.
The very showy, profuse, globular, golden-yellow flowers are borne in dense axilliary panicles, up to 1.8 inches in length, during winter and spring.
Hardy zones 8 to 10, it is very adaptable and tolerates both flood and drought. Pune regularly after flowering to encourage new growth and density.

Acacia caerulescens (Buchan Blue Wattle)
A very attractive, very fast growing, dense, oval to rounded small tree, that is native to extreme southeastern Australia not including Tasmania. Some records include: first year - 10 feet; largest on record - 55 x 40 feet with a trunk diameter of 1.5 feet. It is endangered in the wild with just around 1700 plants left in existence. It makes an excellent landscape tree and should be widely planted where it is adapted in order to preserve it.
The obovate leaves, up to 3.2 x 1.3 inches, are blue-green.
The pale yellow flowerballs are borne on clusters, up to 3.5 inches in length.
Hardy zones 9 (estimate...it may actually be hardier). It is both lime and clay soil tolerant.

Acacia caffra
A medium size tree native to tropical and southern Africa.
Some records include: 4 years - 16 feet; 5 years - 18 feet; 9 years - 17 feet; largest on record - 66 feet.
The bipinnate leaves are up to 9 inches in length.
The flowers are cream colored.
Hardy north to zone 9b tolerating as low as 23 F

Acacia calamifolia
A shrub, reaching a maximum height of 13 feet, that is native to southeastern Australia.
The narrow-linear, phyllode leaves, up to 4.3 inches in length, are gray-green to green.
The flowers are borne on very showy, golden-yellow puffballs.
Hardy zones 9?

Acacia cana
A dense, spreading, small tree, reaching a maximum size of 20 x 20 feet, that is native to eastern Australia. Some records include: 2 years - 4 feet.
The linear leaves are up to 5 x o.1 inches in size. The very attractive foliage is bright gray-green.

The golden-yellow puffball flowers are very showy.
The bark is gray.
Hardy zones 9 to 10 (fully hardy to 16 F).

Acacia cardiophylla (Wyalong Wattle)
A showy spreading, arching shrub or more rarely small tree that is native to New South Wales in Australia. Some records include: 3 years - 3 feet; largest on record - 17 x 13 (rarely over 11) feet.
The bipinnate leaves, up to 2.5 inches, are composed of tiny, heart shaped leaflets.
The profuse, sweetly fragrant, globular, bright yellow flowers are borne in long panicles that cover the stems during late winter and spring.
The bark is smooth and gray.
Hardy zones 8 to 10

Acacia catechu (Black Catechu)
A medium size tree native to Pakistan, India and Burma.
Some records include: fastest recorded growth rate - diameter growth of 0.5 inches; 2 years - 10 feet; largest on record - 80 x 50 feet with a trunk diameter of 3.2 feet.
The fine bipinnate leaves are up to 6 inches in length.
The yellow flowers are borne in long spikes during spring.
The twigs are armed with hooked spines.
Hardy zones 9 to 12^

Acacia cavenia (Cavan)
Also called Espanillo. An attractive twisted trunked, fast growing small tree.
Some records include: fastest recorded growth rate - 5 feet; 2 years - 6 feet; 3 years - 9 feet; 4 years - 12 feet; largest on record - 30 x 30 feet with a trunk diameter of 20 inches. Often used as a hedge in its native Chile.
The bipinnate leaves are composed of tiny leaflets. The foliage is downy at first.
The yellow, ball-shaped flowers are borne from the leaf axils during spring.
The valuable timber is durable and very hard.
The flexible branches are spiny.
Hardy zones 8 to 11. Leaves undamaged at 19 F
Pharmacology: known to contain Tryptamine

** photo of unknown internet source*

Acacia chinchillensis (Chinchilla Wattle)
An attractive, fast growing, small spreading shrub is native to a small area in southern Queensland, Australia. Some records include: 3 years - 4;5 feet; largest on record - 6.6 x 6 feet.
The bipinnate foliage is gray-green. The leaflets are up to 0.3 inches in length.
The showy globular, golden-yellow flowers are borne in panicles during late winter and spring.
The bark is smooth and gray-brown.
Hardy zones 9 to 11. Trim regularly for increased density.

Acacia choriophylla
A thornless small tree reaching a maximum size of 20 x 24 feet, that is native to Florida.
Extremely drought tolerant.

Acacia citrinoviridis
An attractive weeping tree, reaching a maximum height of 50 feet, that is native to western Australia. Some records include: 6 years - 13 feet.
The linear to narrow-elliptic leaves are up to 6.5 x 0.5 inches in size.
The bark is fissured and gray.
Fully hardy to 17 F.

Acacia cognata (River Wattle)
A rapid growing, gracefully weeping, small tree reaching up to 40 x 40 feet, that is native to coastal southeast Australia.
The narrow phyllode leaves, up to 4 inches in length, are bright green.
The globular, golden-yellow flowers are borne on short stalks during late winter and spring.
Hardy zones 9 to 11 (tolerates 20 F) on moist soil in sun to partial shade.

Acacia concurrens
A fast growing tree, reaching a maximum size of 33 feet, that is native to the eastern Australia's

center coast. Some records include: 3 years - 13 feet.
The obovate phyllode leaves, up to 7 x 2.3 inches, are blue-green.
The bright yellow flower spikes are up to 4.3 inches in length.
The fibrous fissured bark is grayish-black.
Hardy zones 9 to 10

Acacia conferta (Crowded-Leaf Wattle)
An attractive fine shrub reaching a maximum size of 15 x 13 feet, that is native to dry areas of southeastern Queensland and nearby parts of New South Wales.
The very dense, nearly needle-like, narrow-oblong, deep gray-green phyllode leaves, are up to 0.5 inches in length.
The small, bright yellow, ball-shaped flowers are borne singly along the stems.
Hardy zones 9 to 11

Acacia confusa (Formosan Acacia)
A medium-sized tree that is native to Taiwan and the Phillipines but has also become an exotic invasive weed in Hawaii. Some records include: largest on record - 66 x 90 feet with a trunk diameter of 3.5 feet.
The narrow, lance-shaped, phyllode leaves are up to 4 x 0.4 inches, are dull green.
The golden-yellow flowerballs are up to 0.4 inches across.
The seeds are highly poisonous.
The slender twigs are smooth and brown.
The bark is smooth and gray.
Hardy zones 10 to 12, preferring a subtropical to tropical climate.
Pharmacology: at a concentration of 1.4%, the roots contain among the worlds highest known concentrations of DMT.

** photo of unknown internet source*

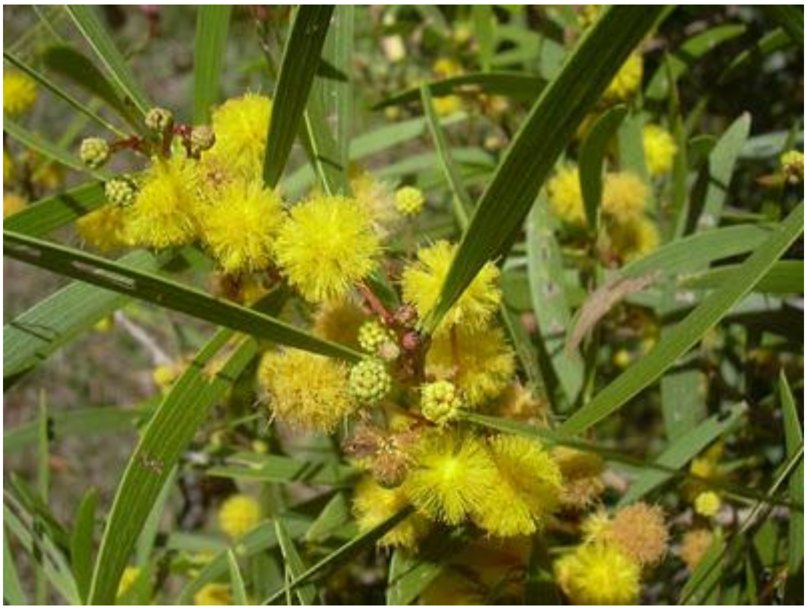

Acacia constricta (Whitethorn Acacia)

A moderate growing, long-lived, large shrub to small tree reaching up to 10 x 15 feet, that is native to Nevado, Arizona to Texas. Some records include: 2 years - 8 feet; 7 years - 14 feet; 11 years - 16 feet. Generally evergreen, it becomes deciduous during extreme drought or cold. It makes an excellent barrier plant.

The soft yellow, puffy, ball-shaped flowers are borne during mid spring to early summer. They are nearly pollen free.

The flowers are followed by attractive deep reddish-brown seed pods, up to 5 inches in length.

The thorns are sharp and white.

Hardy zones 5a to 11, it is among the hardiest of all Acacias, though it is far less hardy in moister climates. Whitethorn Acacia prefers full sun and infrequent deep watering. Very drought and heat tolerant. Deer resistant.

Pharmacology: contains β-methyl-phenethylamine

** photo of unknown internet source*

Acacia continua (Thorn Wattle)

A dense, wiry-stemmed, rigid branched, shrub reaching a maximum height of 6.5 feet that is native to semi-arid parts of southcentral Australia.

The leathery, sharp-tipped, linear leaves, up to 1.6 (rarely over 1) inches in length, are pressed so closely to the stems, that this plant appears leafless.

The foliage is green.

Small, showy, deep golden-yellow, ball-shaped flowers are borne from the leaf axils during early spring.

Hardy zones 9 to 11

Acacia coriacea (Wiry Wattle)
A slightly pendulous, small tree reaching up to 33 x 15 feet, that is a widespread native to northern Australia.
The narrow, linear phyllode leaves, up to 13 x 0.4 inches, are silvery-blue.
The the globular golden-yellow flowers, up to 0.3 inches across, are followed by curled pods, up to 8 inches in length.
The bark is fibrous.
Hardy north to zone 9b (tolerating as low as 20 F). At colder temperatures it will dieback to the base, resprouting rapidly to up to 7 feet during the following summer.
Pharmacology: unknown

Acacia cornigera (Bullhorn Acacia)
A small tree, reaching up to 33 feet, that is native to Mexico and central America.
Some records include: 1 year - 6 feet; 2 years - 7.5 feet; 4 years - 8 feet.
The foliage is bipinnate.
Hardy zones 9 to 11
Pharmacology: contains Tryptamine

Acacia coulteri
Some records include: 2 years - 6 feet; eventual size of 7.5 feet

Acacia covenyi (Bluebush)
A rare fast growing tree, reaching a maximum size of 28 x 30 feet, that is native to far southeastern Australia.
The narrow oblong leaves, up to 2.3 x 0.4 inches, are silvery-blue.
The bright yellow flower puffballs, up to 0.2 inches across, are borne 5 to 8 on a clusters, up to 2.2 inches in length.
Hardy north to zone 8b tolerating 14 F. It thrives in climates with 20 + inches of yearly rainfall.
Drought tolerant.

Acacia cowleana
A small tree, reaching a maximum height of 27 feet, that is native to central Australia. Some records include: 2 years - 9 feet with a trunk diameter of 2 inches.
The oblanceolate leaves, up to 9 x 1.5 inches, are gray-green.
The flower spikes, up to 2 inches in length, are golden-yellow.
The bark is fibrous.
Hardy zones 9 to 11 (tolerating 20 F). Only marginally hardy in Tucson

Acacia craspedocarpa (Leatherleaf Acacia)
Also called Hop Mulga. An evergreen large shrub to small tree reaching a maximum size of 27 x 20 (rarely over 20) feet, that is native to southwestern Australia. Some records include: fastest recorded growth rate - 3 feet; 4 years - 6 feet; 10 years - 9 feet.
The rounded leaves, up to 1 x 0.5 inches, are blue-gray.
The bright yellow flower spikes are up to 0.8 inches in length.
Hardy zones 8 to 10, thriving in semi-arid climates. Very drought tolerant.

Acacia crassicarpa (Thick-Podded Salwood)
A long lived, tall spreading tree that is a widespread native to moist areas of eastern Australia often forming dense pure stands. One of the worlds fastest growing trees reaching up to 100 feet.
The long, curved phyllode leaves, up to 12 x 2 inches in size, are pale green to gray-green.
The golden-yellow flowers are borne in paired spikes during late winter and early spring.
The bark is dark gray-brown.
Hardy zones 9 to 11 (slight damage at 16 F).

Acacia cultriformis (Knife-Leaf Wattle)
A tall, very dense, evergreen shrub with drooping branches that is native to Eucalyptus forests of New South Wales in eastern Australia. Some records include: fastest growth rate - 3 feet; 3 years - 10 feet; largest on record - 26 x 18 feet.
It makes an excellent hedge and screen.
The attractive, pointed-tipped, triangular, phyllode leaves, up to 1.5 x 0.6 inches, are blue-gray.
The lightly fragrant. bright yellow, ball-shaped flowers are borne in sprays up to 3 inches in length.
Hardy zones 8 to 11 tolerating as low as 10 F, does not like climates with high humidity. Prefers full sun (tolerates partial shade) on humus-rich, sandy, well drained soil. Tolerates heat, drought and seaside salt winds.
Pharmacology: contains Tryptamine

** photo of unknown internet source*

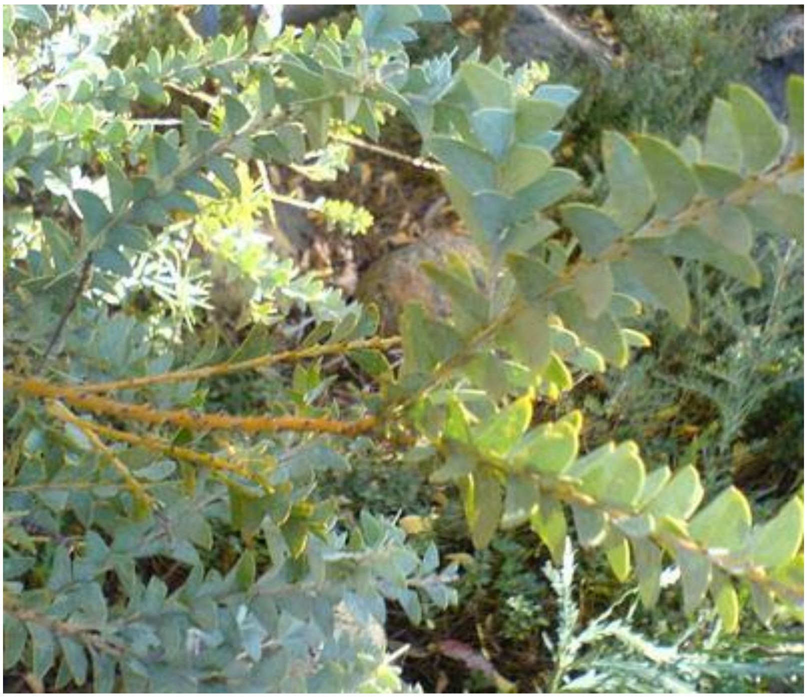

'Austraflora Cascade'
A low groundcover form, reaching up to 1 x 13 feet.

Acacia cyclops (Western Coastal Wattle)
A very attractive, dense, bushy, small tree, reaching a maximum size of 28 x 28 feet, that is native to

the western and central part of the southern Australia coast.
Some records include: 10 months - 4.5 feet; 2 years - 8 x 5 feet.
The leathery, oblanceolate, willowy leaves, up to 4 x 0.6 inches, are green.
The pale yellow flowerballs, up to 0.3 inches across, are borne on 2 flowered racemes originating from the leaf axils.
Hardy zones 9 to 11(tolerates 10 F). Drought tolerant and even thrives on pure sand & salt breezes, making it a great choice for stabilizing sand dunes.

Acacia davyi (Corky Thorn)
An attractive, fast growing, evergreen tree, reaching up to 17 x 17 feet.
Some records include: 3 years - 6 feet.
The attractive, ferny bipinnate leaves, up to 6 inches in length, are bright green.
The bright yellow flowerballs, up to 0.3 inches across, are borne during summer.
The attractive, corky to flaking bark ranges from creamy-yellow to light brown.
Hardy zones 9 to 10 (partial damage at 16 F) in full sun to partial shade on well drained soil. Drought tolerant.

Acacia dealbata (Silver Wattle)
A very fast growing, single trunked, broadly-conical large tree to around 70 feet that is native to mountain forests of southeastern Australia and Tasmania.
Some records include: fastest recorded growth rate - 6 feet; 3 years - 20 feet; 5 years - 33 feet; 20 years - 50 x 27 feet (average); 22 years - 70 feet; 28 years - 82 feet with a trunk diameter of 3 feet; largest on record - 150 x 82 feet with a trunk diameter of 5 feet.
The ferny, bipinnate, blue-green leaves, up to 7 inches in length, are composed of many linear leaflets up to 0.3 inches. The finely-textured foliage is very attractive.
The abundant, very showy, very fragrant, bright yellow, ball-shaped flowers are borne in racemes up to 8 inches in length during late winter into spring.
The flowers are followed by seed pods up to 3 inches in length, that constrict between the seeds.
The branchlets are silvery. The bark is smooth and green, eventually becoming gray, developing cracks on older trees.
Hardy zones 8 to 10 tolerating as low as 8 F and it often resprouts from the ground if it did actually get colder. It thrives in milder parts of the British Isles. Prefers full sun and acidic, warm, well drained soil. Preferring consistent moisture, it can tolerate the occasional drought.

** photo of unknown internet source*

subsp 'subalpina'
Originating from high mountains in Australia, it is more consistantly cold hardy but still unlikely to thrive north of zone 8.
It is vigorous, grows large and has very attractive silvery foliage.

Acacia deanii (Deane's Wattle)
A small tree reaching a maximum height of 40 (rarely over 23) feet, that is native to southeastern Australia. Some records include: 3 years - 5 feet; 10 years - 15 feet.
Makes an excellent screen plant.
The ferny bipinnate leaves are dull deep green.
The creamy-white flowerballs are up to 0.2 inches across.
The smooth bark is brownish-purple.
Hardy zones 9 to 10 (fully hardy to 16 F). Very drought tolerant.

Acacia debilis
A large shrub to small tree, reaching a maximum height of 20 feet, that is native to eastern Australia.
Some records include: 3 years - 12 feet; largest on record - 20 feet.
The bipinnate leaves are composed of very narrow leaflets, up to 0.7 inches in length. The foliage is gray-green.
The tiny flower puffballs are golden-yellow.
The branches are reddish and the smooth bark is grayish-green.
Hardy zones 9b to 10 (est)

Acacia decora (Showy Wattle)
A dense, spreading shrub native to dry areas of eastern Australia.
Also called Western Golden Wattle. Some records include: 3 years - 5 feet; largest on record - 17 x 17 feet. Somewhat short-lived, trees exceeding 50 years in age are rare.
The slightly curved, narrow-elliptic leaves, up to 2.5 x 0.5 inches, are bright blue-green.
The very showy, globular, bright yellow flowers are borne in racemes during winter and spring.

The branches are reddish-brown.
Hardy zones 8 to 11 in sun to partial shade. Easy to grow.

Acacia decurrens (Black Wattle)
A highly ornamental, very fast growing, large, erect tree native to New South Wales in Australia. Some records include: fastest recorded growth rate - 17 feet with a trunk diameter increase of 1 inch; 6 years - trunk diameter of 5 inches; 25 years - 60 feet; largest on record - 80 x 66 feet with a trunk diameter of 3 feet. Trees as large as 55 feet are known to grow in Ireland.
The bipinnate leaves, up to 6 inches in length, are deep green.
The linear leaflets are up to 0.6 inches in length.
The fragrant, globular, bright yellow flowers are borne in axilliary racemes, up to 6 inches in length, during late winter and early spring.
The dark gray bark is furrowed.
Hardy zones 8 to 10 (unconfirmed reports of 7) thriving in areas with yearly rainfall exceeeding 18 inches. Very tolerant and easy to grow however may be prone to borers.

'mollis'
2 years - 9 feet.
Only marginally hardy in Tucson.

'subalpina'
11 years - 17 feet with a trunk diameter of 8 inches.
Has very silvery, tiny leaves.

Acacia difformis (Wyalong Wattle)
A small tree, reaching a maximum height of 23 feet, that is native to southeastern Australia.
Some records include: 4 years - 7.5 feet.
The pendulous, narrow-oblanceolate leaves, up to 6.5 x 0.6 inches, are green.
The bright yellow flowers are borne on clusters, up to 2.5 inches in length.
Likely hardy north to zone 8b (fully hardy to 15 F), surviving in Tucson, AZ.

Acacia dunnii (Dunn's Wattle)
A fast growing small tree native to far northwestern Australia. Some records include: 2 years - 21 feet; largest on record - 23 x 20 feet.
The large, falcate, glaucous-blue leaves are up to 18 x 12 (rarely over 15 x 7) inches.
The very spectacular 0.8 inch golden-yellow flower balls are borne on panicles up to 20 inches in length, during most of the year.
Hardy in zone 11, tropics only.

Acacia elata (Cedar Wattle)
A tree native to southeastern Australia. Some records include: fastest recorded growth rate - ; 3 years - 25 feet; largest on record - 133 x 60 feet with a trunk diameter of 2 feet.
The pinnate leaves are composed of lance-shaped leaflets, up to 2.5 x 0.5 inches.
The flowers are creamy-yellow.
The bark is blackish-gray.

Hardiness in North America is unknown, likely north to zone 8b.

Acacia enterocarpa (Jumping Jack Wattle)
A dense, low spreading shrub reaching a maximum height of 5 feet that is native to border areas between South Australia and Victoria in Australia.
The linear leaves are up to 1.8 inches in length.
The bright yellow flowers are borne on globular heads.
Hardy zone 9 to 11 preferring sandy soil.

Acacia eremophila
A dense, rounded shrub, reaching a maximum height of 6.5 feet, that is native to southwestern Australia.
Some records include: 3 years - 4 feet.
The very linear leaves, up to 4.5 inches in length, are mid-green.
The golden-yellow flowerheads are borne from the leaf axils.
Hardy zones 9a to 10, preferring mediterranean or dry climates.

Acacia erioloba (Camel Thorn)
Also called Acacia giraffae. It is a slow growing, shapely spreading evergreen tree that is a widespread native in southern Africa. Some records include: largest on record - 60 x 60 feet with a trunk diameter of 3 feet; longest lived - 300 years.
Hardy zones 9 to 11 tolerating as low as 18 F. It is a valuable shade tree in hot desert climates.
The leaves are bipinnate.
The sweetly fragrant, globular, golden-yellow flowers are borne singly during late winter and early spring.
They are followed by velvety gray coated, wide, sickle-shaped seed pods.
The stems are armed with white thorns.
The bark is rough and deeply fissured.
Hardy zones 9 to 11 tolerating as low as 18 F. Extremely drought tolerant due to its roots reaching as deep as 150 feet in search of water.

** photo of unknown internet source*

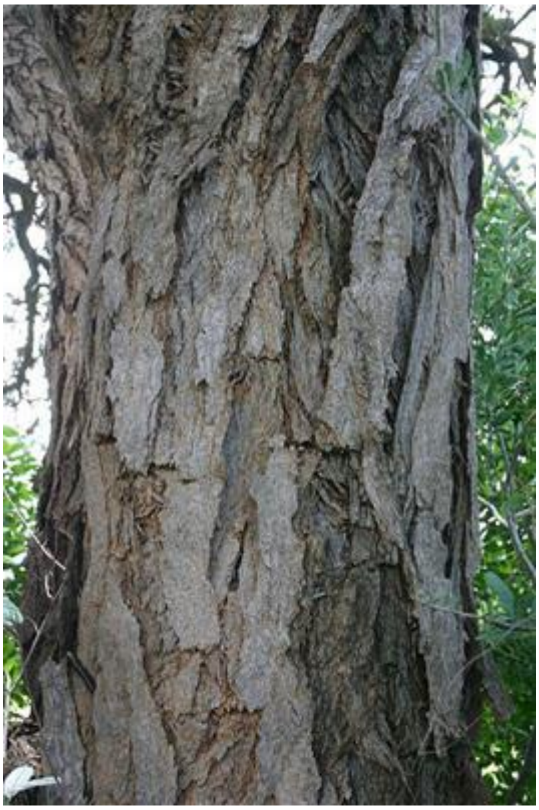

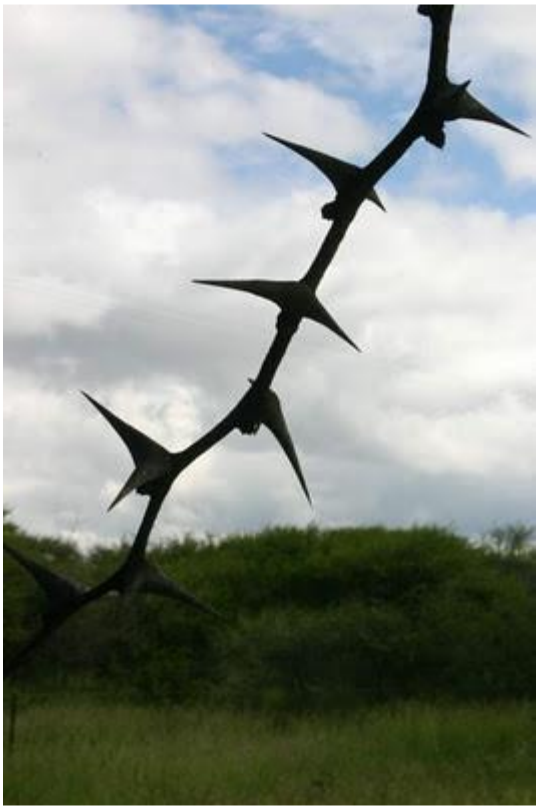

Acacia erubescens

Attractive, thorny, medium-sized tree that is native to central and southern Africa.

Some records include: 1 year - 6 feet; 4 years - 10 feet; 10 years - 20 feet.

The leaves are bipinnate.

The bottlebrush flowers are white.

The showy bark is papery.

Acacia estrophiolata
A gracefully weeping Willow-like tree reaching a maximum size of 70 x 35 feet, that is native to inland areas of Australia.
The linear phyllode-leaves are up to 4.3 x 0.2 inches in size.
The pale yellow flowers are borne on globular heads, up to 0.2 inches across.
The bark is smooth and silver.
Hardy zones 9 to 10 (tolerating 20 F).

Acacia falcata (Hickory Wattle)
A small tree reaching a maximum height of 40 (rarely over 15) feet, that is native to southeastern Australia. Some records include: 1st year - 3 feet; 3 years - 15 feet.
The narrow-elliptical, falcate leaves, up to 7.5 x 1.6 inches, are blue-green.
The creamy-white flowers are borne on clusters, up to 2.5 inches in length.
Hardy zones 9 to 10

Acacia falciformis (Pale Hickory Wattle)
A small tree reaching around 33 feet that is native to coastal southeast Australia.
Some records include: 5 years - 30 feet; largest on record - 82 x 40 feet with a trunk diameter of 3 feet.
The pendulous phyllode leaves are up to 9 x 2.5 inches, are blue-green.
Small, very fragrant, yellow, ball-shaped flowers are borne in large clusters, up to 4 inches in length, during spring.
Hardy zone 9 to 10

Acacia farnesiana (Sweet Acacia)
Also called Huisache. A fast growing, wiry, rounded to wide-spreading, small semi-evergreen to evergreen tree native to tropical America ranging north into Baja Peninsula, southern Texas and south Florida. This is a most attractive tree when in full bloom. The outer twigs are pendulous adding to the beauty of this tree.
Some records include: fastest recorded growth rate - 4 feet; 2 years - 7 feet; 3 years - 14 x 12 feet; 10 years - 20 feet; largest on record - 40 x 50 feet with a trunk diameter of 4.5 feet. Long lived, the Sweet Acacia can persist over 100 years. It has spread into the wild outside its native range in parts of southern California, southern Arizona and central Texas to Louisiana and central Florida where it may become invasive on open to semi-open sites on sandy soils.
It makes a great hedge or screen in the desert southwest.
The feathery bipinnate leaves, up to 4 inches in length, are composed of 3 to 8 pairs of pinnae, each having 10 to 25 pairs of pointed, narrow oval leaflets, each up to 0.3 (very rarely 0.5 inches) inches in length. The foliage is bright-green.
The very attractive foliage is finely-textured.
The large, sweetly very fragrant, bright golden-yellow, globular flowers, up to an inch wide, are borne in small clusters in late winter and spring.
Oils obtained from the flowers can be used in the making of perfume.
They are followed by thick, shiny, reddish-purple to very dark brown, oblong pods, up to 3 x 2 inches.

The zig-zagged orange-brown twigs are armed with strong stiff, straight, whitish spines, up to 2 inches in length, that are borne in the leaf axils.
The bark is reddish-brown with abundant pale lenticels on young trees. Gray with long thin vertical strips on older trees.
Hardy zones 8 to 12 tolerating as low as 10 F (however clones from the Caribbean and Florida are killed at or below 20 F), thriving in climates with 12 + inches of yearly rainfall. Very salt, heat and drought tolerant. It is sold in pots in many Florida nurseries. It is easy to grow in just about any soil.
Pharmacology: Contains DMT & Tryptamine

** photo of unknown internet source*

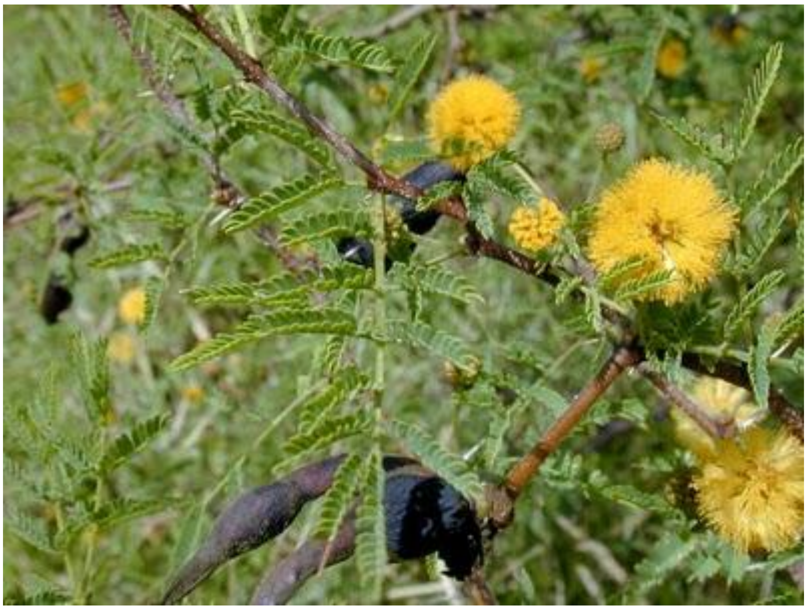

Acacia filicifolia (Fern-Leaf Acacia)
A tree, reaching a maximum size of 60 (rarely over 40) feet, that is native to southeastern Australia.
Some records include: 11 years - 53 x 40 feet with a trunk diameter of 13 inches.
The deep green, fernlike leaves are composed of linear leaflets, up to 0.3 inches in length.
The smooth bark is dark brown.
Hardy zones 8b to 10, thrives and grows very rapidly in milder parts of England.

Acacia fimbriata (Fringed Wattle)
Also called Brisbane Golden Wattle. A fast growing, bushy small tree that is native to eucalyptus forests of coastal areas in eastern Australia.
Some records include: fastest recorded growth rate - 5 feet; 3 years - 10 feet; largest on record - 25 x 20 feet. Fringed Wattle makes an excellent screen.
The linear phyllode leaves, up to 3 x 0.2 (rarely over 2) inches, are deep green.
The profuse, sweetly fragrant, globular, bright yellow flowers are borne in dense clusters, up to 3 inches in length, during late winter and spring.
The slender branches are drooping.
Hardy zones 9 to 11 tolerating as low as 14 F in sun to partial shade.

Acacia fleckii
A multi-stemmed small tree, native to southern Africa that sometimes suckers to form inpenetrable thickets. Some records include: 11 years - 25 feet; largest on record - 33 feet with a trunk diameter of 1 foot.
The foliage is gray-green.
The bottlebrush flower spikes, up to 2.6 inches in length, are creamy-white.
The creamy-white to pale gray bark is smooth or flaking on very old trees.
Hardy zones 9 to 11 (tolerating 20 F). Thrives in Yuma (with irrigation) but not Tucson. Thrives best in tropical climates but will tolerate occasional drought.

Acacia flexifolia (Bent-Leaf Wattle)
A bushy, rounded to spreading shrub reaching up to 7 x 7 feet, that is native to inland parts of southeastern Australia.
Some records include: 3 years - 7 feet; largest on record -
The narrow phyllode leaves are up to 2 (rarely over 0.8) inches in length.
The pale yellow, ball-shaped flowers are borne along the branches during winter and spring.
Hardy zones 9 to 11 tolerating as low as 19 F. Thrives in sun to partial shade. Prune lightly after flowering.

Acacia floribunda (Sally Wattle)
A fast growing, dense small tree with pendulous branches, that is native to coastal parts of eastern Australia. Some records include: 1 day - 1 inch; 5 years - 17 feet; largest on record - 47 x 50 feet. It is generally short lived, only living around 20 years.
The densely crowded, linear, phyllode leaves, up to 7 x 0.5 inches, are deep green.
The profuse, sweetly fragrant, pale yellow flowers are borne in loose spikes up to 3 inches in length during winter and early spring.
They are followed by seedpods up to 4 x 0.3 inches.
The smooth bark is gray with horizontal streaks and lighter mottles.
Hardy zones 9 to 11 tolerating as low as 19 F, in sun to partial shade. Very adaptable and easy to grow.
Pharmacology: contains Tryptamine, phenethylamine

Acacia furcatispina
Some records include: 3 years - 8 feet; 5 years - 12 feet; 12 years - 18 feet.
Hardy zones 9 to 11 (fully hardy to 18 F). Thrives in Tucson

Acacia galpinii
A fast growing, massive, deciduous tree native to southern Africa similar in appearance to a mature Gleditsia - Honeylocust.
Some records include: fastest recorded growth rate - ; 2 years - 18 feet; 3 years - 20 feet; 4 years - 25 feet; 9 years - 35 feet; 12 years - 40 feet; largest on record - 100 x 82 feet with a trunk diameter of 5 feet.
The bipinnate leaves, up to 7 inches in length, are luxuriant bright green.
The long bottlebrush flowers are creamy-white.
Hardy zones 9a to 11 (fully hardy to 16 F), it thrives even in Yuma, AZ extreme heat.

The root system is extensive.

Acacia genistifolia (Early Wattle)
An open shrub, reaching a maximum height of 10 feet, that is native to southeastern Australia.
The linear phyllode leaves are up to 1.6 inches in length.
The flower heads are creamy-yellow.
Hardy zones 9 to 10

Acacia gerrardii subsp 'negerensis'
Some records include: 3 years - 7 feet; 4 years - 10 feet; 8 years - 18 feet; 15 years - 22 feet; largest on record - 45 x 40 feet.
Hardy zones 9 to 11 (fully hardy to 16 F).

Acacia glaucoptera (Clay Wattle)
A rapid growing, dome shaped shrub reaching a maximum size of 8 x 10 feet, that is native to southwestern Australia. Often reaching only 3 feet in height, it can be used for groundcover.
The overlapping, flat leaves, up to 3 x 1 wide, are deep purple at first turning to blue-green. In cool climates, the foliage turns purplish during winter.
The deep yellow flower-balls are borne on very short stalks along the stems during winter and spring.
Hardy zones 9 to 11 tolerating as low as 19 F, thriving in full sun and partial shade. Heat and drought tolerant. It grows well as an understory plant under high canopy trees.
Pruning after flowering and pruning out dead wood encourages new growth.

Acacia gnidium
A dense, spreading shrub, reaching a maximum height of 13 feet, that is native to northeastern Australia.
Some records include: 2 years - 4 feet; 3 years - 6 feet.
The linear leaves are up to 2 inches in length.
Likely hardy north to zone 9, thriving in climates such as southern Arizona with occasional deep watering.

Acacia greggii (Catclaw Acacia)
A small to medium size tree reaching up to 20 x 20 feet, that is a widespread native to the western U.S. from southern California & Nevada to Austin & Corpus Christi, Texas; south into the Baja Peninsula to eastern Mexico. It is deciduous and during dry years may remain dormant or leafless for most of the year.
Some records include: 2 years - 7 feet; 3 years - 10 x 3 feet; 8 years - 15 feet; 12 years - 17 feet; largest on record - 50 x 70 feet with a trunk diameter of 50 inches. Though the maximum longeavity is not known, this tree can certaily exceed 130 years of age on ideal sites. This tree is an excellent choice for a spiny inpenetrable hedge and land reclaimation.
The bipinnate leaves, up to 3 inches in length, are composed of 1 to 3 pairs of pinnae, each having 8, 10 or 12 oval leaflets, each up to 0.3 inches in length.
The entire leaf may have up to 40 leaflets. The foliage is gray-green.
The small, fragrant, bright creamy-yellow flowers are borne in dense oblong cylindrical spikes up to

2.5 inches in length.
They are followed by often curved, flattened pods, up to 5 x 0.8 inches. The pods usually constrict between the seeds. The pods often persist from midsumer through winter into the following spring. The pods are edible and can be eaten raw, however are better boiled with a change of water to reduce their bitterness.
The branches are armed with thick, sharply curved spines up to 0.3 inches in length.
The bark is brown and scaly. The heavy wood is hard and typically not used commercially. It is sometimes used for fence posts. The wood also makes good firewood as it burns hot.
Hardy zones 7 to 9 tolerating as low as -13 F in dry climates. This tree is not known to grow in the humid eastern U.S. Catclaw Acacia prefers a soil PH of 7 to 8.5 and is extremely drought tolerant though much faster growing with occasional deep irrigation.
The seeds may take years to germinate in the wild due to their hard shell.
The seeds will germinate much faster if soaked in water for 24 hours or soaked in acid before sowing. The Catclaw Acacia is also easy to reproduce from semi-ripe cuttings.

** photo of unknown internet source*

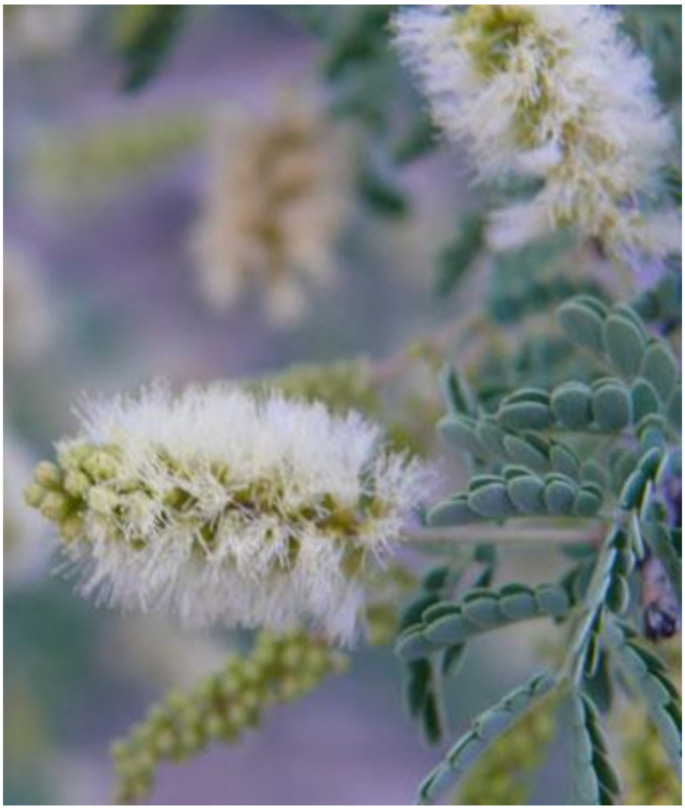

Acacia gummifera (Morocco Acacia)
A small tree reaching a maximum height of 33 feet, that is native to north and central Africa.
Hardy north to zone 9, extremely drought tolerant.

Acacia haematoxylon
Some records include: 3 years - 3 feet; 7 years - 5 feet.
Hardy zones 9 to 11 (fully hardy to 16 F).

Acacia hakeoides (Western Black Wattle)
A multi-branched, bushy, small tree native to drier regions of Australia. Some records include: 3 years - 6 feet; largest on record - 33 x 17 feet.
The thick, oblanceolate phyllode leavss, up to 6 x 0.5 inches, are green.
The showy, dense, bright yellow ball-shaped flowers are borne in axilliary racemes during winter and spring.
Hardy zone 8 to 10. Drought tolerant.

Acacia harpophylla (Brigalow)
A fast growing, dense, erect, long lived small tree native to semi-arid parts of eastern Australia.
Some records include: 5 years - 8 feet; largest on record - 82 x 60 feet with a trunk diameter of 2 feet. It often suckers to form thickets in the wild. An excellent shade and shelter tree in dry climates.
The curved, leathery, phyllode leaves, up to 8 x 1 inches, are silvery-gray.
The yellow ball-shaped flowers are borne in short racemes during late winter and spring.
The bark is dark and furrowed.
Hardy zones 9 to 10 (fully hardy to 16 F), it thrives in Tucson
The seeds are soft coated and do not need treatment to germinate readily.

Acacia harveyi
A dense, large shrub to small tree, reaching a maximum size of 20 feet, that is native to far southwestern Australia.
Some records include: 3 years - 5 feet; 6 years - 8 feet.
The linear leaves are up to 5 inches in length.
The lemon-yellow flowers are borne on globular heads.
Hardy zones 9 to 11 (fully hardy to 16 F with foliage damage only), preferring mediterranean type climates.

Acacia havilandii (Needle Wattle)
A very attractive, bushy, broad vase-shaped shrub native to southeast Australia.
Some records include: 3 years - 6 x 10 feet; 5 years - 8 feet; largest on record - 13 x 17 feet. The bark is gray-brown.
The sharp pointed, slightly curved, fine, needle-like phyllode leaves are up to 3.2 (rarely over 2) inches in length. The foliage is gray-green.
Masses of bright yellow ball-shaped flowers are borne in late winter and spring.
Hardy zone 9 to 11. Hardy in Tucson

Acacia hebeclada
Some records include: 1 year - 3 x 5 feet; 2 years - 4 x 8 feet; 3 years - 7 feet. Eventual height of 8 feet.
Hardy zones 9 to 10 (tolerates 16 F with slight damage).

Acacia hemsleyi
A small tree, reaching a maximum height of 23 feet, that is native to far northern Australia. Some records include: 2 years - 14 feet with a trunk diameter of 6 inches.
The narrow, oblanceolate leaves are up to 5 x 0.5 inches in size.

The rod-shaped flower spikes are up to 1.3 inches in length..
Only marginally hardy in Tucson.

Acacia heroensis
Some records include: 3 years - 6 feet.
Hardy zones 9 to 11 (tolerating 16 F with slight damage). Hardy in Tucson

Acacia holosericea
A rapid growing, thornless, spreading, small tree that typically only lives around 12 years.
Some records include: 1.5 years - 13 feet with a trunk diameter of 4 inches; 3.5 years - 16 feet; largest on record - 27 feet.
The narrow elliptical leaves are up to 10 x 4 (rarely over 8 x 2) inches.
The foliage is luxuriant blue-green.
The rod-shaped flower spikes are golden-yellow.
Hardy zone 10 though may extend somewhat into zone 9, it is only marginally hardy in Tucson.
Saline soil and flood tolerant.

Acacia howittii (Howitt's Wattle)
A very fast growing, dense, rounded, gracefully weeping tree native to a resticted range in southeast Australia. Some records include: fastest recorded growth rate - 6.6 feet; largest on record - 30 x 10 feet. It makes an excellent tall hedge plant.
The spicy fragrant, narrow elliptical phyllode leaves, up to 1 x 0.3 inches, are deep green.
The fragrant, lemon-yellow flower-balls are borne in masses during spring.
Hardy zone 9 to 10 tolerating as low as 19 F, thriving in full sun or partial shade.
It can tolerate heavy pruning.

Acacia implexa (Hickory Acacia)
A fairly open, erect, long lived shade tree that is native to near coastal and coastal areas of eastern Australia.
Some records include: 3 years - 4.5 feet; largest on record - 53 x 50 feet with a trunk diameter of 1 foot.
The curved phyllode leaves, up to 8 x 1 inches, are green.
The creamy-yellow flower-balls are borne in long racemes during summer and early autumn.
The bark is rough and grayish.
Hardy zones 8 to 11 in sun to partial shade.
Pharmacology: considered Psychoactive

** photo of unknown internet source*

Acacia iteaphylla
A fast growing, bushy, spreading, large weeping shrub reaching a maximum size of 18 x 18 feet that is native to southern Australia.
The broad linear leaves, up to 5.5 x 0.3 (rarely over 4) inches in length, are gray-green.
The fragrant, pale yellow flowers are borne in autumn and winter.
The bark is green on young plants.
Hardy zones 8 to 10. Drought tolerant.

Acacia jennerae
A moderate growing evergreen large shrub or small tree that is native to New South Wales in Australia.
Some records include: 2 years - 11 feet; 3 years - 14 feet; 5 years - 16 feet; largest on record - 20 x 12 feet.
The narrow-elliptical leaves, up to 4.5 x 1 inches, are silvery-blue.
The puffball flowers are golden-yellow.
The bark is reddish-brown.
Hardy zones 9 to 11 (fully hardy to 15 F) and thrives in Tucson.

Acacia karroo (Karroo Thorn)
Also called Acacia horrida. A fast growing, rounded, spreading, medium-sized, deciduous tree that is a widespread native in southern Africa. Some records include: fastest recorded growth rate - 3.5 feet; 2 years - 8 feet; 3 years - 13 feet; 6 years - 16 feet; 7 years - 20 feet; 11 years - 25 feet; largest on record - 70 x 60 feet with a trunk diameter of 1.5 feet. The Karroo Thorn can live up to 300 years. It is naturalized in southern Europe where it is grown as a screen.

The bipinnate leaves, up to 5 x 2 inches, are deep green.
The profuse, sweetly fragrant, deep yellow, ball-shaped flowers are borne at the branch tips during summer though sometimes into autumn.
The orange-red twigs are armed with paired white thorns, up to 2 inches in length.
The smooth bark is brownish-gray.
Hardy zones 8 to 11. Flood tolerant. Invasive roots may be a problem on restricted sites.
Pharmacology: Psychoactive

Acacia kempeana
A spreading large shrub to small tree, reaching a maximum height of 17 feet, that is native to central Australia. Some records include: 6 years - 10 feet.
The narrow elliptical leaves, up to 3.5 x 1 inches, are gray to blue-green.
The dense, golden-yellow flower spikes are up to an inch in length.
The bark is furrowed and brown.
Hardy zones 9 to 10 (fully hardy to 16 F), thriving in warmer parts of the U.S. desert southwest.

Acacia kettlewelliae
A bushy, small tree, reaching a maximum height of 33 feet, that is native to far southeast Australia.
The narrow, elliptical leaves, up to 4 x 0.3 inches, are blue-green to green.
The golden-yellow, puffball flowerheads are borne on clusters, up to 2.5 inches in length.
Hardy zones 8b to 10, is known to exceed 15 feet in the British Isles.

Acacia koa (Koa)
A large, massive trunked, dense, bushy tree native to mountains of Hawaii that can reach over 100 feet and is a close relative of Acacia melanoxylon and Acacia confusa.
Some records include: fastest recorded growth rate - 7 feet; 2 years - 17 feet; 5 years - 33 feet with a trunk diameter of 1 foot; 20 years - 100 feet; largest on record - 140 x 150 feet with a trunk diameter of 11 feet.
The lance-shaped, phyllode leaves, up to 10 inches in length, are gray-green.
The pale yellow ball-shaped flowers are borne in axilliary clusters during spring.
Hardy zone 9 to 11. It's tough timber is valuable for the use in carving and the making of canoes.

** photo of unknown internet source*

Acacia lanigera (Woolly Wattle)
A rigid, rounded to spreading shrub native to far southeast Australia, reaching a maximum size of 10 x 10 feet.
The rigid, narrow-elliptic, phyllode leaves, up to 2.3 x 0.4 inches, are bright green.
The bright yellow flowers are borne in globular flowerheads borne from the leaf axils during late winter and early spring. They are followed by densely hairy, curled, brown pods.
The branches are covered in dense woolly hairs.
The bark is gray.
Hardy zones 8 to 11 and somewhat flood tolerant. Prune after flowering to encourage new growth and dense habit.

Acacia leiocalyx (Curracabah)
An attractive small tree, reaching a maximum height of 33 feet, that is native to the central part of Australia's east coast. Some records include: 2 years - 9 feet with a trunk diameter of 2 inches; 3 years - 12 feet; 4 years - 18 feet.
The narrow-elliptical phyllode leaves, up to 7 x 1.4 inches, are blue-gray.
The very profuse, showy, pale-yellow flowers are borne on rod-shaped clusters.
The branches are reddish.
The flaky, fissured bark is gray-brown.
Hardy zones 9b to 10 (tolerates 21 F with slight damage)

Acacia leucophloea
A large tree with attractive white-gray bark. Some records include: 5 years - 35 feet; largest on record - 120 x 100 feet with a trunk diameter of 5 feet.

It can live up to 100 years.
Tolerates temperatures anywhere from 30 to 122 F and is very drought tolerant.

Acacia leurderitzii
Some records include: 2 years - 14 feet; 5 years - 22 feet. Eventual size of 25 feet.
Hardy in Yuma.

Acacia ligulata (Dune Wattle)
A dense, spreading shrub, reaching a maximum size of 13 x 13 feet, that is native to central and southern Australia.
The narrow-elliptical leaves, up to 3 x 0.5 inches, are deep green.
The deep golden-yellow, ball-shaped flowerheads, up to 0.3 inches across, are borne on short racemes that are borne from the leaf axils.
The branchlets are light brown.
Hardy zones 9 (tolerating 15 F)

Acacia linifolia (Flax Wattle)
A fast growing, graceful, open, airy, erect to pendulous slender shrub reaching up to 20 x 12 feet, that is native to woodland understories around Sydney, Australia.
The branching is slender and arching.
The crowded, linear phyllode leaves are up to 2 inches in length.
The yellow ball shaped flowers are borne 6 to 12 on slender axilliary racemes, up to 2 inches in length, during summer and fall though sometimes persisting into late winter.
Hardy zones 9 to 11 tolerating as low as 20 F

Acacia longifolia (Sydney Golden Wattle)
A fast growing, short trunked, bushy, spreading small tree native to southeastern Australia.
Some records include: fastest recorded growth rate - 7 feet; 6 years - 27 feet; 20 years - 37 x 37 feet; largest on record - 37 x 82 feet; largest in Seattle - 26 feet.
It makes an excellent windbreak and screen.
The thick, straight, narrrow-elliptic, willow-like, leathery, phyllode leaves, up to 10 x 1.5 inches, are heavily longitudinally veined. The foliage is deep green.
The abundant bright yellow flower spikes are borne in erect racemes up to 3 inches in length all along the branches during winter and spring.
Hardy zones 9 to 11 tolerating as low as 19 F; it grows in milder parts of Cornwall, England and there are unconfirmed reports of zone 8 hardiness. Tolerant of limey soils.
Pharmacology: plant contains 0.2% DMT

Acacia macradenia (Zigzag Wattle)
A pendulous shrub, reaching a maximum size of 17 x 13 feet, that is native to northeastern Australia.
Some records include: 3 years - 8.5 feet.
The long, narrow-oblong phyllode leaves are up to 10 x 1 inches in size.
The flower clusters, up to 2 inches in length, contain 8 to 15 flowers.
Hardy zones 9 to 11 (estimate)
Pharmacology: Tryptamine

Acacia macrantha (Steel Acacia)
A small tree to 25 feet that is very similar to the Twisted Acacia that is native to southern Florida and the Caribbean. Some records include: largest on record - 50 x 45 feet with a trunk diameter of 1 foot.
The flowers are followed with pods up to 5.5 inches in length.
Hardy north to zone 9.

Acacia maidenii (Maiden's Acacia)
A medium size tree reaching a maximum size of 66 x 18 (rarely over 50) feet that is native to southeastern Australia. Long-lived, this tree can persist over 100 years.
Not a real massive tree, trunk diameters over 1 foot are extremely rare.
The narrow-elliptic phyllode leaves are up to 8 inches in length.
The creamy-yellow to bright yellow flowers are borne on rod-like spikes, up to 2.3 inches in length.
The flower spikes are borne from the leaf axils.
The branches are light brown. The deep gray-brown bark is smooth on young trees, becoming deeply fissured on older plants.
Hardy north to zone 9 tolerating as low as 19 F. Flood tolerant.
Pharmacology; stem bark contains 0.6% NMT and DMT in 2:3 ratio. The foliage contains both.

Acacia mangium (Black Wattle)
A large spreading tree, up to 120 feet; that native to coastal areas in far northern Queensland, Australia as well as southern New Guinea and the southern Molucca Islands. The Black Wattle makes an excellent shade tree. Some records include: 1st year - 10 feet; 3 years - 50 feet; 5 years - 80 feet with a trunk diameter of 10 inches; 14 years - 100 feet with a trunk diameter of 16 inches; largest on record - 170 x 40 feet with a trunk diameter of 3 feet.
The conspicuous veined, large, lance-shaped to elliptical leaves, up to 12 x 4 inches, are deep green.
The rod-like, creamy-white flowers, up to 5 inches in length, are borne in summer and fall. Twisted, woody pods in clusters follow.
The branches are stout.
The coarsely ridged and cracked bark is gray to dark brown.
Hardy zones 9 to 12 thriving where annual rainfall exceeds 40 inches.
Pharmacology: Psychoactive

** photo of unknown internet source*

Acacia meansii (Late Black Myrtle)
A large, upright, spreading tree native to forests of southeast Australia. Some records include: fastest recorded growth rate - 17 feet; 1 year - 17 feet with a trunk diameter of 1 inch; 2 years - 27 feet; 3 years - 35 feet; largest on record - 100 x 33 feet with a trunk diameter of 2 feet.
It makes for an excellent shade and shelter tree.
The ferny bipinnate leaves, up to 6 inches in length, are glossy deep green.
The profuse, light yellow, globular flowers are borne in loose clusters up to 6 inches in length during late spring to early summer.
The smooth gray to blackish bark becomes corrugated at the base on very old trees.
Hardy zones 7 to 11 thriving in climates with 18 inches of more of rainfall per year.
Drought tolerant.

Acacia melanoxylon (Blackwood Acacia)
A, fast growing, tall, dense broad crowned evergreen tree, up to 150 feet, that is native to coastal eastern Australia and Tasmania. Some records include: 6 years - 33 feet with a trunk diameter of 8 inches; 20 years - 50 x 23 feet; 30 years - 66 feet (average); 40 years - 108 feet; largest on record - 200 x 70 feet with a trunk diameter of 10 feet; largest in England - 90 feet with a trunk diameter of 3 feet. With ideal conditions, it is a long lived tree, persisting well over 150 years.
Used as a street tree in California.
The heavily longitudinally-veined, long, curved, narrow, phyllode leaves, up to 7 x 1 inches, are deep green.
The profuse, light yellow, puffball flowers, up to 0.4 inches across, are borne in masses during late winter and early spring.
The strong, close-grained, dark timber is valuable for making cabinets.

Hardy zones 8 to 10 (tolerating as low as 8 F) in sun to partial shade on a sheltered site on fertile, moist soil. Thrives in parts of England. The Blackwood requires a yearly rainfall between 30 and 72 inches per year.

Tolerant of flooding, salt, wind, smog, heat and shade.

it hates drought and is very prone to borer attack in warm areas.

In South Africa it has escaped into the wild and become a weed.

Pharmacology: bark and leaf contains <0.02% DMT Acacia mellifera A rounded, small tree reaching up to 30 feet that is native to dry regions of Africa, Saudi Arabia & India. It often forms dense inpenetrable thickets in the wild. Some records include: 2 years - 6 feet; 4 years - 10 feet; 11 years - 15 feet. The leaves and stems are rich in protein and are great for feeding livestock in dry regions. The fragrant, creamy-white flowers are borne on short, dense, hanging spikes. The stems are armed with very sharp, recurved thorns. The bark is smooth and gray. Hardy zones 9 to 11 (fully hardy to 16 F). Hardy in Tucson. Thrives where rainfall averages 16 to 32 inches and is clay tolerant. Pharmacology; leaf contains DMT. Acacia milleifolia (Santa Rita Acacia) A very attractive, medium-sized shrub. Some records include: 2 years - 5 feet; eventually up to 8 feet. The pinnate leaves are composed of linear leaflets. The foliage is luxuriant mid-green. The white flowers are borne on showy, rod-like clusters. Hardy zones 9 to 11 (tolerating 16 F with slight damage). Hardy in Tucson. Acacia melvillei (Yarran) A very rare Acacia that is native to southeastern Australia. It is a large shrub to small tree, reaching a maximum size of 50 x 38 feet with a trunk diameter of 20 inches. It is long-lived and trees can persist as long as 150 years. Acacia omalophylla is nearly identical but occurs at higher elevations in the same region of Australia. The narrow willow-like leaves, up to 4.3 x 1 inches, are mid-green. The showy flowers are yellow. Hardy zones 9 to 10 in full sun. Drought and lime tolerant. Acacia modesta A fast growing, medium-sized, deciduous tree, that is native to western Pakistan, Afghanistan and India. Some records include: 2 years - 9 feet; 3 years - 15 feet; 10 years - 22 feet. The double pinnate leaves are up to 2 inches in length. The stems are armed with deep brown thorns, up to 0.25 inches in length. The bark is greenish-gray on young trees, later turning rough and brownish. The wood is hard and durable. Hardy zones 9 to 10. Thrives in Tucson Acacia mollifera (Velvet Acacia) A multi-stemmed, pendulous, small to medium size, evergreen tree native to Australia. Some records include: fastest recorded growth rate - 2 feet; largest on record - 40 feet. The double-pinnate leaves, up to 2.5 inches in length, are silvery-gray. Up to 42 golden-yellow puffball flowers may be borne in a cluster. The stems are densely tomentose. The brown to grayish-black bark is smooth except for becoming fissured near the base on older trees. Acacia montana (Mallee Wattle) A multi-stemmed, large shrub, reaching a maximum size of 13 x 20 feet, that is native to southeastern Australia. Some records include: 6 years - 5.5 feet. The narrow oblong leaves are up to 1.6 x 0.3 inches in size. The golden-yellow puffball flowers, up to 0.3 inches across, are borne from the leaf axils. The stems are either sparsely thorned or have no thorns at all. Hardy zones 9 to 10 (fully hardy to 16 F). Hardy in Tucson Acacia montigena A vining rainforest Acacia, reaching a maximum height of 100 feet, that is a widespread native to Africa. The double-pinnate leaflets are composed of linear leaflets. The young branched are dark brown to black. Hardy zones 10 to 12 Acacia mosambique Some records include: largest on record - 100 x 85 feet with a trunk diameter of 5 feet. Acacia mucronata Similar to Acacia longifolia but is slightly hardier and with very narrow phyllode leaves up to 8 x 0.3 inches. It is taller growing but more narrow in habit, reaching a maximum height of 50 feet. It is native to southeastern Australia including Tasmania. Acacia muelleriana (Mueller's Wattle) An attractive,bushy, multi-stemmed, spreading small tree native to eucalyptus forests in Queensland & New South Wales in Australia. Some records include: 3 years - 13

feet; largest on record - 27 x 20 feet. The double-pinnate leaves are deep green. The leaflets are linear, up to 1.5 inches in length. The creamy-yellow flowers are borne in heads during spring and summer. They are followed by narrow seed pods up to 6 inches in length. The smooth bark is yellow-brown to gray. Hardy zones 9 to 11 Acacia murrayana (Murray's Wattle) A fast growing, small tree, reaching a maximum size of 27 x 27 feet with a trunk diameter up to 8 inches. It is native to central Australia where it often forms suckering colonies in the wild. Some records include: 3 years - 5 feet. The oblanceolate leaves, up to 7 x 0.5 inches, are bright green to blue-green. The showy, golden-yellow flowerballs are borne on clusters up to 2.5 inches in length. Hardy zones 9b to 11 in dry climates. Acacia myrtifolia (Myrtle Wattle) A fast growing, bushy, low to medium size shrub reaching a maximum size of 10 x 10 feet that is native to southern Australia. The erect, lance-shaped to narrow-elliptic, phyllode leaves, up to 6 x 1.3 inches, are reddish at first, turning to deep green. The fluffy, creamy-yellow flowerballs, up to 0.5 inches across, are borne in short racemes, up to 4.5 inches, over a long season during winter and spring. The stems are red. The bark is smooth and gray. Hardy zones 8 to 11 in full sun to partial shade. Prune lightly after flowering to promote new growth.
** photo of unknown internet source*

Acacia nebrownii (Water Acacia)
A moderate growing, dense, semi-evergreen large shrub to small tree reaching a maximum size of 23 x 20 (rarely over 17) feet, that is native to dry parts of southern Africa.
Some records include: 2 years - 8 feet; 3 years - 12 feet; 5 years - 14 feet.
The double-pinnate leaves, up to 1.3 x 0.8 inches, are green.
The showy flowers are golden-yellow.
The yellow to orangish stems are armed with white thorns, up to 2.3 inches in length.
The bark is smooth and reddish.
Hardy zones 9 to 10 (tolerating as low as 18 F) in full sun. It will freeze to the base at 16 F but will regrow rapidly the following spring.
It is tolerant of both flooding and temporary drought.

Acacia neovernicosa (Varnish Acacia)
Also called Chihuahuan Whitethorn. An upright shrub, reaching a maximum size of 13 x 8 feet, that is native from Arizona and southern New Mexico into Mexico. Some records include: 2 years - 7 feet; 4 years - 10 feet; 7 years - 13 feet.
The leaves are double-pinnate.
The puffball flowers are bright yellow.
Hardy zones 7a to 11 in full sun. Tolerant of severe drought and alkaline soil.

Acacia neriifolia (Oleander Wattle)
A fast growing, open, erect, small tree, reaching a maximum size of 50 x 30 (rarely over 30) feet, that is native to eastern Australia. Some records include: 3 years - 7 feet.
The linear to narrow-elliptic, willowy, phyllode leaves, up to 8.5 (rarely over 6) inches in length, are downy white at first turning to blue-green.
The abundant, showy, golden-yellow flower puffballs, up to 0.3 inches, are borne on long narrow axilliary racemes up to 4 inches in length during late spring into early summer.
The gray bark becomes fissured on older trunks.
Hardy zones 9 to 11 in full sun on well drained soil. It prefers a climate with 20 to 28 inches of average yearly rainfall. The seeds require 1 minute immersion in boiling water to break dormancy.

Acacia nigrescens
A shapely, rounded, large tree, that is native to Africa. Some records include: fastest recorded growth rate - 2 feet with a trunk diameter increase of 1.3 inches; largest on record - 80 x 70 feet with a trunk diameter of 3 feet.
Long lived, this tree can exceed 300 years in age.
It makes an excellent shade tree where it is adapted.
The bipinnate leaves are composed of leaflets up to an inch in length.
The very showy, creamy-white flower spikes are borne before the new foliage emerges.
The bark is dark brown to black. The trunk is characteristically covered in thorn tipped knobs.
Hardy zones 9a to 11 in full sun on well drained soil It can thrives where average rainfall is anywhere between 15 and 90 inches in a year though preferring a dry season. Drought tolerant.

Acacia nilotica (Egyptian Acacia)

A rapid growing, dense, rounded, large tree native to drier parts of tropical Africa as well as from Saudi Arabia to India. Some records include: 1 year - 6 feet; 2 years - 12 feet; 3 years - 18 feet; 4 years - 23 feet; 6 years - 25 feet; 10 years - 30 feet; largest on record - 82 x 40 feet with a trunk diameter of 4 feet. It should not be planted in Australia where it can spread into the wild and become an invasive weed. A very attractive landscape tree resembling the Honey Locust in appearance.

The bipinnate leaves, up to 6 (rarely over 3) inches in length, are composed of 0.3 inch leaflets. The deep green foliage can either be deciduous or evergreen depending upon moisture conditions.

The flower puffballs are golden-yellow.

The stems are armed with paired, straight spines up to 3 inches in length.

The deeply fissured bark is dark brown.

Hardy north to zone 9b tolerating as low as 26 F and as well as hot as 120 F. It is both tolerant of flooding, saline soil and also drought, surviving in climates with as low as 14 inches of rainfall per year. It is recommended to direct seed it into its permanent position since transplanting is difficult due to its deep taproot.

Pharmacology: leaf contains DMT

** photo of unknown internet source*

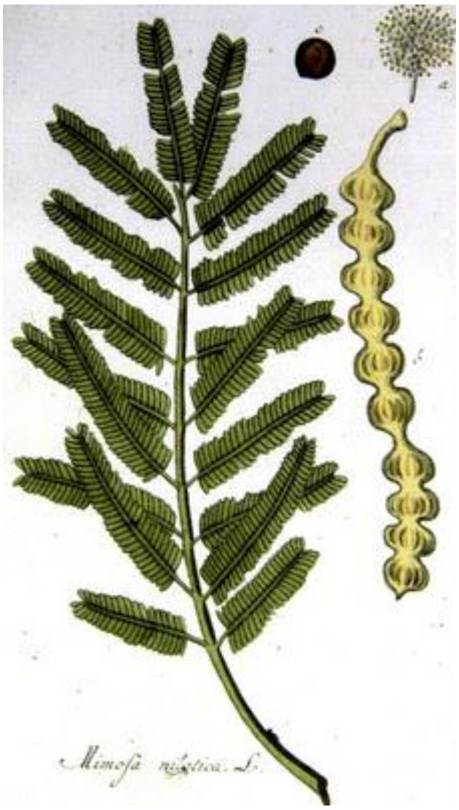
Mimosa nilotica. L.

Acacia notabilis (Notable Wattle)

Also called Flinder's Wattle. A moderate growing, bushy, spreeading shrub, reaching a maximum height of 17 feet, that is native to southeastern Australia. Some records include: 2 years - 5 feet.

The oblanceolate leaves, up to 5 x 1 inches, are blue-green.

The bright yellow flower puffballs, up to 0.3 inches across, are borne on clusters up to 2.5 inches in length.

The twigs are deep reddish-brown. The bark is smooth and red to gray-brown.
Hardy zones 9 to 10 (fully hardy to 17 F)

Acacia obtecta
A very attractive, slow growing, bushy, rounded shrub,, reaching a maximum size of 13 x 13 feet, that is native to southwestern Australia. Some records include: 8 years - 6 feet; 11 years - 7.5 feet. It is great for soil stabilization where it is adapted.
The linear to oblanceolate leaves, up to 4.5 x 0.3 inches, are dull green.
The showy, bright yellow, flower puffballs, up to 0.4 inches, are borne 2 to 4 on axilliary clusters.
The bark is smooth and pale gray.
Hardy zones 9 to 11 (foliage damage only at 16 F) on sand or clay, preferring a Mediterranean climate.
To propagate, boil the seed for 5 minutes to break dormancy.

Acacia occidentalis (Sonoran Catclaw)
A thorny deciduous tree native to the Sonoran Desert in Mexico that is basically a larger version of Acacia greggii reaching a maximum size of 40 x 40 feet.
Some records include: 2 years - 6 feet; 5 years - 10 feet; 10 years - 20 feet.
The deciduous, compound foliage is gray-green. It casts dense shade.
The creamy ball shaped flowers are borne during spring.
The twisted branches are covered in sharp spines.
Hardy zones 9 to 11 (fully hardy to 15 F). Extremely heat and drought tolerant.
Very hardy in Tucson

Acacia odoratissima (Black Siris)
Correctly called Albizzia odoratissima. Some records include: 2 years - 10 feet; 5 years - 17 feet with a trunk diameter of 6 inches; largest on record - 133 feet with a trunk diameter of 5 feet.
The bipinnate leaves, up to 8 inches in length, are composed of leaflets up to 1.5 x 0.5 inches in size.
The foliage is deep green.
The fragrant, creamy-white flowers are borne 10 to 12 on terminal panicles up to 8 inches in length.

Acacia oxycedrus (Spike Wattle)
A stiff prickly large shrub or small tree reaching a maximum size of 33 x 7 feet, that is native to far southeastern Australia.
The sharp pointed, flat, rigid, lance-shaped phyllode leaves, up to 1.6 x 0.2 inches , are whorled around the stens.
The very showy, dense, light yellow flower spikes, up to 2 inches in length, appear during winter and spring.
The stems are prickly making this an excellent barrier hedge.
The gray-brown bark is smooth, becoming lightly fissured on older plants.
Hardy zones 8 to 10

Acacia pachyacra (Casuarina Acacia)
A bushy shrub, reaching a maximum size of 17 feet, that is native to desert areas of west central Australia. Some records include: 2 years - 6.5 feet; 3 years - 9 feet; 4 years - 12 feet. It resembled the

Casuarina in appearance.
The very linear leaves, up to 8 x 0.04 inches, are bright green to gray-green.
The tiny, golden-yellow flowerballs are borne on clusters, up to 1.3 inches in length.
The branches are reddish-brown.
Hardy zones 10 to 11 (est)

Acacia papyrocarpa (Western Myall)
A slow growing, very dense, rounded to spreading, small tree, reaching a maximum size of 33 feet, that is native to southwestern and south central Australia.
Some records include: 3 years - 7 feet; longest lived - 1000 years.
The very linear leaves, up to 5 x 0.1 inches, are silvery.
The golden-yellow flowerballs are borne on clusters up to 2.5 inches in length.
The bark is gray and fissured.
Hardy zones 9 to 10, (fully hardy to 16 F). Very drought tolerant.

Acacia paradoxa (Kangaroo Thorn)
Also called Acacia armata. A very ornamental, dense, , twiggy, spiny, large shrub to small tree reaching a maximum size of 25 x 25 (rarely over 17 x 13) feet that is a widespread native to southeastern Australia as well as Australia's extreme southwest. It makes an excellent barrier hedge.
It should not be planted in California where it may become an invasive noxious weed.
The wavy edged, narrow-oblong phyllode leaves, up to 1.6 x 0.3 inches, are deep green.
The profuse large, golden-yellow flowerballs, up to 0.5 inches, are borne singly or paired along the stems during late winter and spring.
The stems are armed with needle-like spines.
The finely fissured bark is gray-brown.
Hardy zones 8 to 11, tolerant of extended drought.
It can be pruned lightly one a year to shape.

** photo of unknown internet source*

Acacia pataczekii (Pataczek's Acacia)
A vigorous, handsome, suckering, small tree, that is native to Tasmania. Some records include: 30 years - 20 feet with a trunk diameter of 7 inches.
The narrow-oblong leaves, up to 2.4 x 0.8 inches, are blue-green.
The bright yellow flowerballs are borne on clusters up to 4 inches in length.
The branchlets are dark reddish.
Thrives in milder parts of England.

Acacia patagiata
A rounded to spreading shrub, reaching a maximum size of 8 feet, that is native to far southestern Australia. Some records include: 2 years - 3 x 6 feet; 6 years - 6 feet.
The erect, narrow-oblong phyllode leaves, up to 2.2 x 0.3 inches, are luxuriant blue-green.
The showy, bright yellow flowerballs, up to 0.2 inches across, are borne from the leaf axils.
Hardy zones 9 to 10 (hardy with foliage damage only 16 F) in full sun thriving on well drained sand or clay. Prefers a mediterranean climate.

Acacia pendula (Weeping Myall)
A stunning, slow growing, dense, gracefully weeping tree that is a widespread native to dry regions of eastern Australia.
Some records include: 3 years - 6 feet; largest on record - 48 x 44 feet with a trunk diameter of 25 inches. Long lived, it makes an excellent street tree or windbreak.
The narrow-elliptic phyllode leaves, up to 5.5 x 0.5 inches, are silvery or bluish.
The small, light yellow flowerballs, up to 0.2 inches, are borne 2 to 7 on axilliary racemes, up to 0.6 inches in length. The flowers are borne randomly throughout the year and are not very showy.

The fissured bark is gray or dark brown.
The close grained timber is hard and aromatic.
Hardy zones 9 to 11 tolerating as low as 15 F, thriving in climates with 18 + inches of rainfall per year.
Thrives on well drained soil with infrequent deep watering.
Tolerant of salty soil, drought, temporary flooding and clay. Deer resistant.

Acacia pennatula
A fast growing, dense, very tropical looking tree reaching a maximum size of 50 x 50 feet, that is native to Mexico. Some records include: 2 years - 7 feet; 4 years - 12 feet with a trunk diameter of 3 inches; 5 years - 14 feet; 7 years - 20 feet.
The bipinnate leaves, up to 7 x 1.5 inches, have large feathery pinnae.
The foliage is mid-green.
The bright yellow flower puffballs are borne during spring.
The stems are armed with small paired thorns from the leaf axils.
The fissured bark is brownish-gray.
Hardy zones 10 to 11 in full sun on well drained soil.
Tolerates extreme and reflected heat, even in Yuma, Arizona.
Pharmacology: may contain anticancerous properties.

Acacia penninervis
An erect to spreading, small tree, reaching a maximum height of 27 feet, that is native to the central part of Australia's east coast.
The narrow-elliptic leaves, up to 6.5 x 1.6 inches, are green to blue-green.
The creamy-white flowerballs, up to 0.4 inches across, are borne 15 to 30 on clusters up to 5 inches in length.
The finely to deeply fissured bark is dark gray.
Hardy zones 10 (est)

Acacia permixta (Hairy Acacia)
A shrub that is native to southern Africa. Some records include: 3 years - 7 feet.
The bipinnate leaves are similar to that of the Honeylocust in appearance.
The flower puffballs are bright yellow.
The stems are armed with sharp white thorns.
Hardy zones 9 to 11 (slight damage only at 16 F)

Acacia peuce
A pendulous, medium size tree, reaching a maximum size of 50 x 40 feet, that is native to the Simpson Desert in central Australia.
The extremely narrow leaves are up to 16 x 0.04 (rarely over 6) inches. The foliage is similar in appearance to that of the Casuarina.
The wood is extremely hard.
Hardy zones 8a to 10. Has extreme tolerance of both heat and drought and can survive in climates with as little as 5 to 8 inches of yearly rainfall.

Acacia podalyriifolia (Pearl Acacia)
A very fast growing, highly ornamental, evergreen, slender to spreading, small tree reaching a maximum size of 27 x 27 feet that is native to coastal regions of southern Queensland Australia.
it is typically short lived due to disease and borers.
It is popular as a landscape/street tree for its attractive, silvery, oblong or elliptic, phyllode leaves, up to 2 x 1 inches.
The abundant, very fragrant, golden-yellow flowerballs, up to 0.3 inches, are borne 15 to 30 on dense clusters up to 6 inches in length from early winter into spring.
The stems are downy at first, later turning silvery-gray and smooth. Older trees develop a very attractive blue-green trunk that is finely fissured.
Hardy zones 9 to 11 tolerating as low as 19 F and average yearly rainfall as low as 18 inches. Very clay tolerant. Can be pruned hard after flowering.
Pharmacology: fresh bark contains 0.5% to 2% DMT.

Acacia polyacantha (White Thorn)
A fast growing, layered and flat crowned, semi-deciduous, large tree native to central Africa. Some records include: 1 year - 10 feet; 2 years - 16 feet; 3 years - 20 feet; 5 years - 28 feet; 10 years - 30 feet; largest on record - 82 x 60 feet.
The ferny bipinnate leaves are composed of 11 to 27 pairs of pinnae. The foliage
is luxuriant mid-green.
The narrow, long flower-spikes are creamy-white.
The stems are armed with hooked thorns that often persist on older stems.
The bark is yellow-brown.
Hardy north to zone 9a
Pharmacology: leaf contains DMT

Acacia pravissima (Ovens Wattle)
An extremely fast growing, pendulous, spreading, small evergreen tree native to hill country of southeast Australia.
Some records include: 2 years - 4 feet; 4 years - 6 feet; 20 years - 30 feet; largest on record - 42 x 30 feet with a trunk diameter of 1 foot.
The spine-tipped, triangular, phyllode leaves, up to 0.8 x 0.6 inches, are gray to olive green.
The profuse, golden-yellow, ball-shaped flowers are borne in long racemes up to 4 inches in length during late winter to early spring.
The gray bark is smooth becoming finely-fissured on older trees.
Hardy zone 8 to 10 (tolerating as low as 8 F) in full sun (tolerates partial shade) on humus-rich, well drained soil on a site protected from excessive wind. Thrives on the west coast north to Seattle as well as in England as long as it is planted on a warm sheltered site.
Ovens Wattle prefers consistent moisture but can tolerate the occasional drought. Deer resistant. Prune after flowering.

'Golden Carpet'
A groundcover form that is low growing, reaching a maximum size of 16 inches x 20 feet.
The tringular leaves are up to 0.4 x 0.3 inches. The profuse yellow flowerballs are up to 0.2 inches across.

Acacia prominens (Gossford Wattle)
An attractive, dense, long lived tree reaching a maximum size of 82 x 45 feet, that is native to a very tiny area near Gossford in southeastern Australia.
It makes a spectacular lawn tree.
The narrow-oblong, blue-green, phyllode leaves are up to 2.5 x 0.5 inches, are blue-green.
The profuse, showy, sweetly fragrant, light yellow flowerballs, up to 0.2 inches, are borne in dense clusters, up to 2.3 inches in length, during late winter and early spring.
The bark is smooth and gray.
Hardy zones 9 to 11 tolerating as low as 19 F. Prefers a shaded sheltered position and moist soil.

Acacia pruinocarpa (Black Gidge)
A fast growing, thornless, small evergreen tree reaching a maximum size of 40 x 30 feet that is native to central and western (not southwestern) Australia.
The linear to narrow-elliptic, thick, leathery leaves, up to 7 x 1.3 inches, are gray-green.
The bright yellow flowerballs, up to 0.3 inches, are borne on racemes up to 6 inches in length, during spring.
Hardy zones 9 to 11 (tolerating 18 F but defoliates at 22 f).

Acacia pruinosa (Frosty Acacia)
A medium size tree native to southeast Australia. Some records include: fastest recorded growth rate - 3 feet; 25 years - 50 feet; largest on record - 60 x 40 feet.
The bipinnate leaves, up to 4 inches in length, have pinnae in 5 pairs. The narrow oblong leaflets are up to 0.7 x 0.25 inches. The foliage is bronze at first, turning to blue-green.
The golden-yellow flowerballs, up to 0.4 inches across, are borne on racemes up to 6 inches in length.
The branches are glossy and reddish. The bark is smooth and reddish-purple.
Hardy north to zone 9

Acacia pubescens (Downy Wattle)
A showy, fast growing, pendulous branched, spreading shrub native to coastal South Wales, Australia where it is endangered due to habitat destruction in its tiny natural range around Sydney.
Some records include: 3 years - 8 feet; largest on record - 20 x 10 feet.
The deciduous, bipinnate leaves, up to 3 x 2 inches, are composed of leaflets, up to 0.2 inches in length. The foliage is silvery at first, turning to bright green.
The profuse bright yellow flowerballs, up to 0.25 inches, are borne 8 to 16 on dense drooping sprays, up to 4.3 inches in length, during late winter and spring.
The branches are soft, downy and dark reddish-brown. The bark is smooth and brownish-gray.
Hardy zone 9 to 11

Acacia pulchella (Prickly Moses)
A spiny, broad spreading, semi-evergreen shrub, reaching a maximum size of 10 x 5 (rarely over 4) feet, that is native to southwestern Australia.
The bipinnate leaves are divided into up to 11 pairs of tiny, narrow oblong leaflets, up to 0.25 inches.
The foliage is luxuriant mid-green.

The tiny, deep yellow flowers are borne in dense, rounded heads during spring.
The stems are armed with 1 to 2 axilliary spines, up to 0.8 inches in length, per node.
Hardy zones 9 to 10 in full sun on fertile, sandy soil. It prefers a mediterranean climate.

Acacia pustula
A small tree closely related to Acacia nerifolia that is native to semi-dry regions of southern Queensland in Australia. Some records include: largest on record - 50 feet.
The drooping, linear phyllode leaves are up to 6 x 0.5 inches.
The main ornamental feature is the masses of late winter light golden-yellow flowers.
The stems are dark reddish.
Hardy zones 9 to 11

Acacia pycnantha (Golden Wattle)
A fast growing, open branched, pendulous tree native to Eucalyptus forest on sandy soils in southern temperate regions of Australia. Some records include: 3 years - 13 feet with a trunk diameter of 3 inches; 4 years - 15 x 7 feet; 25 years - 35 feet; largest on record - 76 x 60 feet with a trunk diameter of 3 feet. It can sucker freely from the base. Golden Wattle is the floral emblem of Australia. It has also escaped into the wild and become a weed in South Africa. It is not a long lived tree, lasting an average of 30 years.
The pendulous, thick, leathery, oblanceolate phyllode leaves, up to 8 x 1.5 inches, are glossy rich green.
The profuse, sweetly fragrant, large, golden-yellow, ball-shaped flowers, up to 0.4 inches, are borne in long racemes up to 12 inches in length during late winter and spring.
They are followed by seed pods up to 4.7 inches in length.
The brown bark is smooth becoming finely fissured on older trees.
Hardy zones 8 to 10 thriving in climates with 14 to 40 inches of rainfall per year.
It prefers hot summers with low humidity. Young trees are not as frost hardy as older more established trees.

Acacia pyrifolia (Kanji Bush)
A shrub, reaching a maximum size of 15 (rarely over 10) feet, that is native to northwestern Australia.
The round leaves, up to 2.3 x 1.6 inches, are blue-gray.
The yellow flowerballs are borne on clusters up to 7 inches in length.
The bark is smooth and gray.
Hardy zones 10 to 11 (severe dieback at 21 F but resprouts vigorously) in semi-arid climates.

Acacia redolens (Prostrate Acacia)
An attractive, very fast growing, wide spreading, evergreen shrub that can be used for groundcover. It is native to far southwestern Australia. Some records include: 3 years - 2.5 x 8 feet; 7 years - 4 x 10 feet; largest on record - 6 x 28 (rarely over 5 x 12) feet.
The lance-shape to spathulate, gray to mid-green leaves are up to 3 x 0.5 inches.
The bright yellow flowerballs are up to 0.2 inches across.
Hardy zones 8a+ in full sun preferring a mediterranean climate. Tolerates drought and ocean salt winds.

'Desert Carpet'
A low groundcover form, reaching up to 2 x 15 feet.

Acacia rehmanniana (Silky Acacia)
A low, dome shaped to flat crowned tree native to southeastern Africa that can reach a maximum height of 30 feet.
The downy bipinnate leaves are bright golden-green at first, later turning to deep green. The foliage is deciduous during drought.
The sweetly fragrant, white, ball shaped flowers are borne in small clusters during summer.
The orangish-brown stems are armed with straight paired thorns.
Hardy zones 9 to 11 (diesback to base at 20 F but resprouts vigorously). Very drought tolerant.

Acacia retinodes (Swamp Wattle)
A semi-pendulous, rounded small tree, that is native to southeastern Australia.
Some records include: fastest recorded growth rate - 5 feet; first year - 5 feet; largest on record - 36 x 36 feet with a trunk diameter of 2 feet.
It is generally short-lived, lasting around 40 years.
The narrow, oblanceolate leaves, up to 8 x 0.6 inches, are blue-green to green.
The creamy-yellow flowerballs are borne on clusters up to 1.6 inches in length.
The bark is smooth and gray on young trees, on older trees becoming rough and blackish-brown.
Hardy north to zone 8b tolerating as low as 15 F. Thrives in England.^

Acacia rhetinocarpa
A spreading, dome-shaped, resinous, attractive shrub reaching a maximum height of 6.6 feet that is native to south central Australia where it is endangered.
The thick, curved, resinous, obovate phyllode leaves, up to 0.2 x 0.14 inches, are blue-green.
The profuse, bright-yellow, ball-shape flowers are borne singly in the leaf axils during late winter and spring.
Hardy zone 8 to 10. Tolerant of alkaline soil.

Acacia riceana (Rice's Wattle)
A pendulous small tree that is native to Tasmania. Some records include: 20 years - 27 x 20 feet; 23 years - 37 feet with a trunk diameter of 1 foot; largest on record - 37 x 30 feet
The rigid, pointed, linear to narrow-oblanceolate phyllode leaves, up to 2.2 x 0.1 inches, are deep green.
The creamy-white, ball-shaped flowers are borne in drooping spikes up to 1.5 inches in length.
Hardy north to zone 8 tolerating as low as 5 F. Heavy shade tolerant.

Acacia rigens (Needle Wattle)
Also called Acacia chordophylla & Acacia loderi. A fast growing, dense, rounded shrub to small tree, rarely exceeding 20 feet, that is widely native to dry regions of southern Australia. Some records include: largest on record - 32 x 23 feet with a trunk diameter of 1 foot. It is long lived and can survive as long as 100 years.
The sharp pointed, stiff, narrow phyllode leaves are up to 5.5 x 0.1 inches, is similar to that of the

Casuarina in appearance. The foliage is gray-green.
The profuse, golden flowerballs, up to 0.3 inches, are borne in small clusters during winter and spring.
The dark gray bark is finely fissured.
Hardy zones 8 to 10 in sun to partial shade. An excellent ornamental plant for arid and semi-arid climates.

Acacia rigidula (Blackbrush Acacia)
An extremely thorny, moderate growing, small tree native to south Texan and northern Mexican Chihuahuan desert. Some records include: fastest recorded growth rate - 2 feet; 4 years - 6 feet; 7 years - 9 feet; largest on record - 27 x 31 feet with a trunk diameter of 1 foot. An excellent lawn tree.
The foliage is deep green.
The stems are extremely thorny.
The attractive bark is pale gray.
Hardy north to zone 8 tolerating as low as 15 F. Requires full sun and is heat and very drought tolerant.
Pharmacology: contains Mescaline & DMT

** photo of unknown internet source*

Acacia rivularis
A bushy rounded shrub, reaching a maximum height of 17 feet, that is native to south central Australia. Some records include: 2 years - 8 feet; 4 years - 13 feet; 8 years - 16 feet. It is generally short-lived, lasting around 20 years.
The linear leaves, up to 5.5 x 0.2 inches, are mid-green.
The bright yellow flowerballs, up to 0.2 inches, are borne on clusters up to 1.6 inches in length.
The bark is gray and finely fissured.
Hardy zones 9b to 11 (est) in semi-arid climates.

Acacia robusta 'Ussumb'
A large tree with a massive deeply ridged trunk. Some records include: largest on record - 75 x 70 feet with a trunk diameter of 4 feet.
The leaves are up to 5 inches in length.
The spines are up to 4 inches or more in length.
Hardy north to zone 9b

Acacia roemeriana
A small evergreen tree native to Texas.
Some records include: 2 years - 6 feet; 7 years - 13 feet; largest on record - 25 x 25 feet with a trunk diameter of 1.6 feet.
The very fragrant flowers are white.
Hardy north to zone 7, it is fully hardy at Tucson.
Requires a deep well drained soil and is very heat tolerant as well as drought tolerant. It is not known to grow in the humid eastern U.S.

Acacia rostellifera
A fast growing, dense, spreading tree reaching a maximum size of 40 x 33 feet with a trunk diameter of 2 feet, that is native to southwestern Australia.
The narrow, oblanceolate leaves, up to 4.5 x 0.7 inches, are green to blue-green.
The golden-yellow flowerballs, up to 0.3 inches across, are borne on clusters, up to 1.3 inches in length.
The bark is fissured and dark gray.
Tolerates as low as 20 F, prererring a mediterranean climate. Thrives in Phoenix, Arizona.

Acacia rubida (Red-Stemmed Wattle)
A fast growing, small bushy tree native to much of southeastern Australia. Some records include: 5 years - 20 feet; largest on record - 45 x 20 feet.
The leathery, narrow, oblanceolate to narrow-elliptic phyllode leaves, up to 8 x 1 inches, are medium green during summer turning reddish during winter.
The profuse, bright yellow flowerballs, up to 0.3 inches across, are borne 10 to 20 on clusters up to 4 inches in length. The flowers are borne during late winter and spring.
The angular stems are deep red. The finely fissured bark is brown.
Hardy zones 8 to 10 tolerating as low as 8 F. Prefers full sun to partial shade and is flood tolerant as well as tolerant of drought.

Acacia salicina (Broughton Willow)
A medium size, evergreen tree native to tropical central and eastern Australia that is similar in habit to the Weeping Willow.
Some records include: 2 years - 10 feet; 4 years - 16 feet; 5 years - 20 feet; 10 years - 20 feet; largest on record - 90 x 70 feet with a trunk diameter of 2 feet.
It is long-lived, lasting for over 50 years in age.
It is a handsome foliage plant with narrow elliptical, blue-green leaves up to 8 x 1.3 inches.
The pale-yellow flowerballs, up to 0.3 inches across, are borne in clusters up to 2.5 inches in length.
The branchlets are dark gray-brown.

The bark is fissured and gray-brown.
Hardy zones 9 to 11 (fully hardy to 16 F though foliage is damaged at 18 F) on deep well drained soil in climates with 15+ inches of yearly rainfall. It is tolerant of flooding, saline soils and pure sand making it an excellent choice for rehabilitating sand dunes but is also good on clay. On dry, sandy soils, the roots can reach up to 50 feet or more in depth, on wet poorly drained soils this tree will be much more shallow roots and also much more prone to uprooting in high winds.

Acacia saligna (Blueleaf Acacia)
Also called Acacia cyanophylla or Orange Wattle. It is a fast growing, dense, bushy, pendulous small tree native to far southwestern Australia. Some records include: fastest recorded growth rate - 2 feet; 1 year - 7 feet; 2 years - 14 x 7 feet; 3.5 years - 21 feet; 4 years - 27 x 27 feet; largest on record - 33 x 33 feet. It has invasive potential in some climates and has become a weed in South Africa.
It typically only lives about 20 years.
The drooping, gray-green, lance-shaped phyllode leaves are up to 12 x 3 inches.
The profuse, large, fluffy, yellow flowerballs, up to 0.5 inches, are borne in racemes up to 3 inches in length during late winter and early spring.
The brownish-gray bark is smooth becoming finely fissured on older trees.
Hardy zones 9 to 11. Minor leaf burn at 19 F, stem damaged at 15 F. Drought tolerant, growing in regions with 8 to 40 inches of rainfall per year, making a great windbreak for northern Africa and the Middle East.
Prefers mediterranean winter rainfall climates.

Acacia schaffneri (Twisted Acacia)
An attractive, moderate growing, twisted, deciduous to evergreen, small tree, reaching a maximum size of 25 x 30 feet, that is native to the southwest U.S. (southeastern New Mexico & Texas), south to Oaxaca Mexico.
Some records include: 2 years - 8 feet; 4 years - 11 feet; 7 years - 14 feet; largest trunk diameter - 8 inches.
The finely-divided, ferny, bipinnate leaves, up to 6 inches in length, are gray-green.
The very fragrant, bright yellow flowerballs are up to 0.4 inches across.
The stems are viciously armed with long, 1 inch, paired, white thorns that originate from the leaf axils.
The bark is grayish-brown.
Hardy zones 8b to 10 (fully hardy at 15 F) in full sun on well drained soil.
It makes a great landscape tree in Arizona as well as in its native range.
Pharmacology: Contains mescaline & Amphetamine

Acacia schottii
A deciduous small shrub, reaching a maximum size of 6 x 5 (rarely over 4) feet, that is native to Texas and Chihuahua Mexico. Some records include: 3 years - 4 feet; 7 years - 6 feet.
The lacy bipinnate leaves are composed of linear leaflets. The foliage is bright green.
The showy, golden-yellow flowerballs are borne spring through summer.
The stems are armed with white spines up to 0.7 inches in length.
Hardy zones 8 to 10 (tolerating 10 F) in full sun on well drained soil.
Very tolerant of heat, drought, clay soil and lime.

Acacia sclerosperma (Limestone Wattle)
A dense, rounded, shrub or small tree, reaching a maximum size of 20 x 13 feet, that is native to western Australia. Some records include: 4 years - 9 feet.
The thick, linear to narrow-elliptic leaves, up to 5.5 x 0.7 inches, are blue-green.
The flowers are yellow.
The bark on the stems and branches is smooth and light gray.
Hardy zones 9b to 11 (est)

Acacia senegal (Senegal Gum Arabic)
A flat-topped, medium-sized, deciduous tree that is a widespread native to the plains of subsaharan Africa as well as Oman, Pakistan and northwest India. Some records include: largest on record - 66 x 66 feet with a trunk diameter of 4.3 feet.
The bipinnate leaves, up to 4 inches in length, are gray-green.
The fragrant white to pale yellow flowers borne in long spikes, up to 4 inches in length, appear autumn and winter.
At the base of each leaf are 3 fierce small pricles, one curves downward, the other 2 curve upward.
The attractive, gray-brown bark is pale and flaky.
Hardy zones 8 to 10 in full sun, thriving in climates with 10 to 15 inches of rainfall in an average year. It can survive droughts lasting up to 11 months and prefers a climate with a dry season lasting at least 5 months.

'leiorachis'
Some records include: 2 years - 6.5 feet; 4 years - 10 feet; 5 years - 16 feet; 9 years - 19 feet.

'rostrata'
Some records include: 2 years - 7 feet; 4 years - 10 feet; eventually up to 16 feet.

Acacia siculiformis (Dagger Myrtle)
A shrub, reaching up to 10 x 6 feet, that is native to far southeastern Australia including Tasmania.
The narrow lance-shaped to narrow-elliptic leaves, up to 1.5 x 0.2 inches, are green.
Thr bright yellow flowerballs, up to 0.3 inches across, are borne from the leaf axils.
The bark is smooth and gray.
Hardy zones 8 to 9, thriving on well drained soil.

Acacia sieberiana (Paperbark Thorn)
A very fast growing, flat-crownded, medium to large, deciduous tree native to southern Africa. Some records include: fastest recorded growth rate - 7 feet; 2 years - 14 feet; 3 years - 20 feet; 6 years - 28 feet; 7 years - 30 feet; 10 years - 40 feet; largest on record - 82 x 51 feet with a trunk diameter of 6 feet.
It makes an excellent street tree for dry climates such as in southern Arizona.
The feathery, bipinnate leaves, up to 6 inches in length, are deep green, turning to yellow in fall. The leaves are composed of leaflets, up to 0.12 inches in length.
The profuse fragrant flowerballs are creamy-white.
The stems are armed with paired white thorns, up to 3 inches in length.

The yellowish-brown, flaky, papery bark peels off in flat strips.
Hardy zones 9 to 11. Very drought tolerant, it also tolerates flooding, sand and clay.
A healthy attractive shade tree even in the extreme heat of Yuma.
Pharmacology: leaf contains DMT

'woodii'
A perennial in Tucson, dying back to the ground at 15 F

Acacia simplex
A climbing tree reaching up to 40 feet that is native to both Argentina and islands in the western Pacific Ocean.
The cupped, rounded leaves are bright green.
Pharmacology: bark, stem bark and leafy stems contain high levels of DMT.
This plant has been used in fishing. Products of this plant are harmless to humans but stun fish.

Acacia smallii (Small's Acacia)
May be the same as Acacia farnesiana and is in fact nearly biologically similar. This tree represents trees of southwestern U.S. origion and are much hardier (to 15 F).
Whether or not it is a separate species or just a regional variant of Acacia farnesiana is a subject of botanical debate.
The bipinnate leaves are composed of tiny leaflets, up to 0.1 inches in length.
Hardy zones 8b to 10

Acacia sophorae (Coastal Wattle)
A dense spreading shrub that is a widespread native of eastern Australia and Tasmania, reaching a maximum size of 20 x 33 feet.
An excellent choice for use in preventing beach erosion. On exposed coastal sites, it is often prostrate in habit.
The broad, oblong phyllode leaves are up to 5 x 1.5 inches.
The golden-yellow bottlebrush-like flower spikes, up to 2 inches in length, are borne during spring.
Hardy zones 8 to 11 in full sun to partial shade, very tolerant of salt laden winds.
Pharmacology: contains Tryptamine

Acacia spectabilis (Mudgee Wattle)
Also called Pilliga Wattle. An open, pendulous, tall slender shrub native to the inland slopes of eastern Australia that reaches a maximum size of 20 x 15 feet. Some records include: 3 years - 7 feet; 5 years - 17 + feet. It can also be used as a screen.
The soft, blue-green, bipinnate leaves, up to 4 inches in length are composed of leaflets up to 0.7 x 0.25 inches.
The spectacular, golden-yellow flowerballs, up to 0.4 inches, are borne on long, dense sprays, up to 8.5 inches, during late winter and spring.
The smooth bark is whitish to blue-gray.
Hardy zones 9 to 11 tolerating as low as 19 F. Drought tolerant.
Prune lightly after flowering to maintain shape.

Acacia stenophylla (Shoestring Acacia)
A beautiful fast growing, pendulous tree, that is native to central and eastern Australia. Some records include: 16 months - 13 feet; largest on record - 66 x 30 feet. It is similar to Acacia salicina and makes an excellent landscape tree in the southwestern U.S.
The very long narrow leaves are up to 20 x 1 inches.
The pale yellow flowerballs, up to 0.5 inches across, are borne from the leaf axils.
The fibrous to fissured bark is dark gray-brown.
Hardy zones 8b to 11 (tolerating as low as 15 F) in full sun on well drained soil.
Tolerates clay, alkaline soil, drought and flooding. Very tolerant of saline soils.
Seed should be soaked in boiling water for 1 minute before sowing.

** photo of unknown internet source*

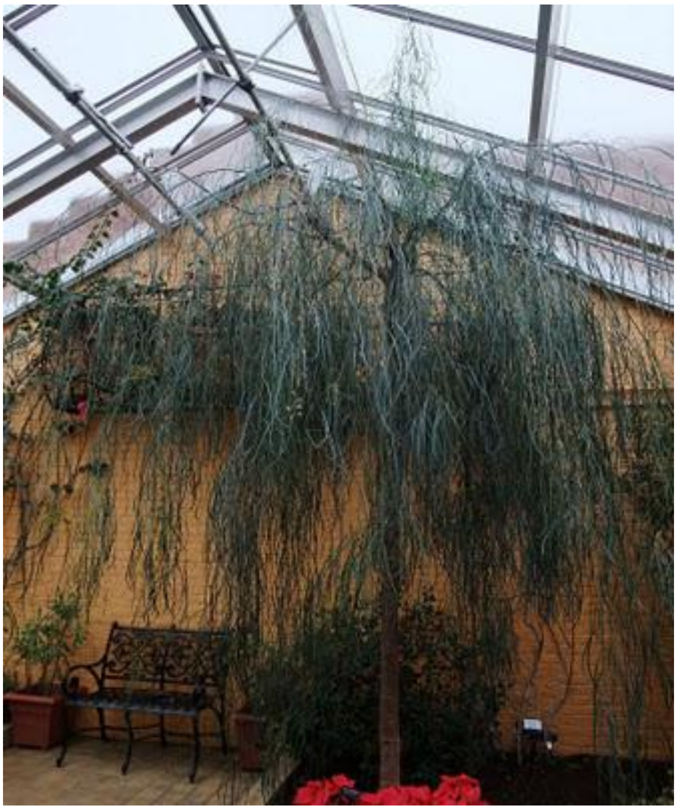

Acacia suaveolens (Sweet Wattle)
A shrub, reaching a maximum height of 10 feet, that is native to coastal eastern Australia and Tasmania.
The linear to oblanceolate, phyllode leaves, up to 6 x 0.4 inches, are green.
The pale yellow flowerballs, up to 0.3 inches across, are borne on axilliary racemes up to 1.3 inches in length.
The bark is smooth and purplish.
Hardy zones 9 to 10 (est). Tolerates pure sand.

Acacia subporosa (Narrow-Leaf Bower Wattle)
A semi-pendulous, small tree, reaching a maximum height of 40 feet, that is native to far southeastern Australia. Some records include: 5 years - 20 feet.

The narrow-elliptical phyllode leaves are up to 5 x 0.4 inches.
The pale yellow flowerballs are up to 0.4 inches across.
The smooth bark is gray or brown.
Hardy zones 9 to 10 (est)

Acacia terminalis (Cedar Wattle)
Also called Sunshine Wattle. A medium size tree reaching a maximum size of 66 (rarely over 20) feet, native to southeastern Australia.
The bipinnate leaves, up to 4.5 inches in length, are composed of oblong leaflets, up to 0.8 x 0.2 inches. The foliage is reddish at first, turning to glossy deep green.
The golden-yellow flowerballs, up to 0.5 inches, are borne on panicles.
The gray or brown bark is smooth, becoming finely fissured on older trees.
Hardy zones 9 to 10

Acacia tetragonophylla (Kurara)
An attractive, dense, semi-pendulous, rounded, shrub or small tree, reaching a maximum size of 20 x 20 (rarely over 10) feet, that is a widespread native to western and central Australia. Some records include: 2 years - 6 feet.
The prickly, needle-like, linear leaves, up to 2 x 0.12 inches, are bright green.
The bright yellow flowerballs, up to 0.4 inches, are borne from the leaf axils.
The gray bark is finely fissured.
Hardy zones 9 to 10 (fully hardy to 16 F with foliage damage only). Extremely drought tolerant.

Acacia tindaleae (Golden-Top Wattle)
A dense small shrub reaching a maximum height of 6.5 feet that is native to eastern Australia.
The very hairy, small, linear phyllode leaves are densely crowded along the stems.
The profuse, golden-yellow flower balls are borne winter and spring.
Hardy zones 9 to 11 thriving in both subtropical and dry climates.
Prune to shape after flowering.

Acacia tortilis (Umbrella Thorn)
A medium size, attractive, umbrella crowned tree native to most of Africa and the Arabian Peninsula Some records include: 2 years - 8 x 10 feet; 2.5 years - 16 feet; 11 years - 22 feet with a trunk diameter of 6 inches; largest on record - 70 x 60 feet.
There is a combination of short, recurved and long, narrow spines with all being very sharp.
The very small, bipinnate leaves have very small leaflets.
The foliage is typically evergreen.
The fragrant, white to light yellow, rounded flowers are borne in small clusters in mid summer with the foliage.
They are followed by spirally twisted, light brown pods borne in clusters.
Hardy zones 9 to 11 tolerating as low as 18 F. It is extremely drought tolerant due to its roots which can extend as deep as 170 feet. It is also very heat tolerant, to as high as 122 F.^

'heteracentha'
A subspecies that ranges widely in southern Africa.

Some records include: 1.5 years - 5 feet with a trunk diameter of 2 inches

Acacia tortuosa (Twisted Acacia)
A rounded to wide spreading, semi-evergeen to evergreen small tree native from central Texas, south into Mexico.
Some records include: fastest recorded growth rate - 2 feet; largest on record - 31 x 30 feet with a trunk diameter of 1.5 feet.
The bipinnate leaves, are composed of 3 to 4 pairs of pinnae, each having 10 to 15 pairs of very narrow, pointed leaflets.
The foliage is light-green.
The fragrant, bright-yellow ball-shaped flowers, up to an inch wide are borne in pairs of 3s from the leaf axils.
They are followed by long narrow, slightly compressed pods, up to 5 x 0.3 inches. that constrict between the seeds.
The branches are armed with slender spines up to an inch in length.
The bark is dark brown and furrowed.
Hardy zones 8 to 10, very tolerant of drought as well as being heat and alkaline soil tolerant.

Acacia trachycarpa (Minniritchi)
A moderately fast growing, thornless tree reaching a maximum size of 20 x 25 feet, that is native to northwestern Australia.
Some records include: 1.5 years - 9 feet.
The linear leaves are up to 4 x 0.1 inches in size.
The golden-yellow flower spikes are up to 0.8 inches in length.
The bark is reddish-brown.
Hardy zones 9b to 10 (tolerating 25 F).

Acacia trineura (Three Nerved Wattle)
A rounded, evergreen large shrub, reaching a maximum height of 17 feet, that is native to southeastern Australia.
Some records include: 2 years - 9 feet; 3 years - 11 feet
The narrow, oblong leaves, up to 3.2 x 0.5 inches, are deep green.
The golden-yellow flowerballs, up to 0.2 inches across, are golden-yellow.
3 to 8 flowerballs may be contained in a raceme.
The bark is smooth and gray.
Hardy zones 9 to 11 (fully hardy to 15 F).

Acacia triptera (Spurwing Wattle)
A very dense, spreading, large shrub native to semi-arid regions of eastern Australia. Some records include: 3 years - 28 inches; largest on record - 15 x 23 feet with a trunk diameter of 9 inches. It makes an excellent barrier hedge.
The stiff, curved, pointed, narrow-triangular phyllode leaves, up to 2.2 x 0.3 inches, are purplish at first, turning to mid-green.
The profuse, very showy, bright yellow, bottlebrush-like flower spikes, up to 1.3 inches in length, are borne late winter and spring.

Hardy zones 9 to 11 (tolerating as low as 19 F) in full sun on well drained soil. Extremely drought tolerant.

Acacia truncata
A dense, domed large shrub, reaching a maximum size of 10 x 5 feet, is native to far southwestern Australia.
The tiny leaves (shaped like upside-down triangles) are only up to 1 x 0.5 inches.
The creamy, ball-shaped flowers are up to 0.5 inches.
Hardy north to zone 9, tolerating as low as 19 F
Requires full sun. Salt tolerant.

Acacia uncinata (Wavy-Leaf Wattle)
A spreading to erect shrub with long pendulous branches that is native to New South Wales in eastern Australia.
Some records include: 5 years - 6 feet; largest on record - 10 feet.
The stem clasping, rounded, wavy, hook-pointed phylloides, up to 1.8 x 0.8 inches, are gray-green.
The large bright yellow flowerballs, up to 0.35 inches across, are borne single or in pairs during spring and summer and also on and off at other times of the year.
Hardy zones 9 to 11 tolerating as low as 19 F. Drought tolerant.

Acacia verticillata (Star Acacia)
Also called Prickly Moses. A attractive, prickly, fast growing, arching shrub to small tree (if trained) that is native to Tasmania and southeast Australia. Some records include: fastest recorded growth rate - 2 + feet; 10 years - 20 x 20 feet; largest on record - 33 x 33 feet; largest in England - 30 feet.
Excellent for use in hedging.
The whorled, rigid, very sharp, linear, needle-like leaves, up to 1 x 0.3 inches, are arranged in whorls.
The bright yellow flowers spikes, up to 1.8 inches in length, are borne either single or in small groups during late winter and spring.
Hardy zones 8b to 11 tolerating as low as 19 F.
It thrives in milder parts of the British Isles.

Acacia vestitia (Hairy Wattle)
Also called Weeping Boree. A bushy, small weeping tree reaching a maximum size of 27 x 20 (rarely over 13) feet, that is native to far southeastern Australia.
The triangular leaves, up to 0.8 x 0.3 inches, are gray-green.
The showy, golden-yellow flowerballs are borne on long racemes, up to 2.3 inches in length.
The bark is smooth and gray.
Hardy north to zone 9, tolerating as low as 19 F. It can be sheared.

Acacia victoriae (Bramble Wattle)
A moderate growing, spreading, small tree native to inland parts of Australia.
Some records include: 1 year - 5 feet; 3 years - 9 feet; 5 years - 14 feet; 8 years - 18 feet; largest on record - 40 x 33 feet with a trunk diameter of 8 inches.
It can be used for hedging and windbreaks as well. Short lived in many regions, lasting around 15 years. It may sucker readily.

The lance-shaped, gray-green leaves, up to 3.5 x 1 inches in size, have a strong, paired spines at the base.
The fragrant, creamy-yellow flowerballs, up to 0.3 inches across, are borne on racemes up to 4 inches in length, during late winter into early summer.
The bark is smooth and dark gray.
Hardy zones 8 to 11. Fully hardy at 15 F at Tucson. Very drought tolerant due to its deep root system which may penetrate as much as 66 feet or more into the earth.
Moderately salt tolerant.

Acacia villosa
A small thornless tree, reaching a maximum size of 23 feet, that is native to Jamaica where it is threatened. Some records include: 2 years - 8 feet; 4 years - 12 feet.
Hardy zones 10 to 12 (freezes to base at 25 F but may reach as much as 9 feet during the following summer).

Acacia viscidula (Sticky Wattle)
A small tree, reaching a maximum height of 20 feet, that is native to the central part of Australia's east coast. Some records include: 4 years - 7.5 feet; 6 years - 10 feet.
The linear leaves, up to 4.3 x 0.1 inches, are glossy mid-green.
The pale yellow flowerballs are up to 0.3 inches across.
The smooth bark is gray-brown.
Hardy zones 9 to 10 (fully hardy to 16 F)

Acacia visco
An attractive, moderate growing, rounded, medium-sized, deciduous to semi-evergreen tree, reaching a maximum size of 40 x 40 feet, that is native to Argentina.
Some records include: 2 years - 5 feet; 3 years - 9 feet; 8 years - 16 feet; 15 years - 22 feet.
The bipinnate leaves, up to 7 x 3.5 inches, are composed of linear leaflets.
The foliage is bright green at first, turning to deep green.
The showy, creamy-white flowerballs, up to 0.5 inches across, are borne during spring through summer.
Hardy zones 8b or 9 to 11. Fully hardy at 15 F at Tucson but defoliates at temperatures in the mid 20s.

Acacia wilhelmiana (Wilhelm's Wattle)
A compact, dense, spreading, medium sized, evergreen shrub, reaching a maximum size of 10 x 10 feet, that is native to semi-arid regions in southeast Australia.
The narrow, linear leaves, up to 1.3 x 0.16 inches, are green.
The masses of single or paired, golden-yellow flowerballs, up to 0.25 inches, are borne late winter and spring.
Hardy zone 8 to 10. Drought tolerant.

Acacia willardiana (Palo Blanco)
A very attractive, evergreen, weeping tree, reaching a maximum size of 33 x 20 feet, that is native to the Sonoran Desert of Mexico. Some records include: 1 year - 5 feet; 2 years - 13 feet; 3 years - 17

feet; 7 years - 20 feet; 10 years - 25 feet. An excellent landscape tree in hot desert climates, it looks like a cross between a Casuarina and a Weeping Willow from a distance.
The bipinnate leaves, up to 4 inches in length, are composed of linear leaflets, up to 0.2 inches in length.
The creamy-yellow flowers are borne on bottlebrush-like spikes, up to 2.5 inches in length.
The attractive bark is white and peeling.
Hardy zones 9b to 11 tolerating as low as 25 F. Very heat tolerant, preferring a hot south facing wall north of zone 10. During extended droughts, a deep watering once a month is recommended.
Pest and disease resistant.

Acacia wrightii (Wright Acacia)
A densely crowned, small evergreen tree reaching up to 20 x 20 feet, that is native to most of central and western Texas, south into eastern Mexico.
Some records include: fastest recorded growth rate - 3 feet; 3 years - 7 feet; 4 years - 11 x 4+ feet; 10 years - 15 feet; 13 years - 17 feet; largest on record - 40 x 40 feet with a trunk diameter of 2 feet.
The bipinnate leaves, up to 3 inches in length, are composed of 1 to 3 pairs of pinnae, each having 3 to 5 pairs of oval leaflets, each up to 0.8 inches in length.
The somewhat hairy foliage is bright-green.
The small, fragrant, light-yellow flowers are borne in cylindrical spikes up to 1.5 inches in length. They are followed by often curved, flattened pods, up to 6 x 1.5 inches. The pods contain narrowly oval seeds up to 0.3 inches in length. The ripened seed pods are not persistant, falling early in autumn.
The branches are sparsely armed with thick, curved brown spines.
The bark is dark brown, scaly and furrowed.
Hardy zones 7 to 10. Fully hardy at 15 F at Tucson. Very tolerant of both heat and drought. It is not known to grow in the humid eastern U.S.
Seeds germinate much better if soaked in sulfuric acid for 20 minutes before sowing to loosen the seed shells.

Acacia xanthophloea (Fever Tree)
A handsome, widespreading, fast growing, large deciduous tree native to southeastern Africa.
Some records include: 2 years - 14 feet; 3 years - 22 feet; 4 years - 25 feet; 5 years - 30 feet; 30 years - 50 x 40 feet; largest on record - 100 x 80 feet with a trunk diameter of 4 feet.
The bipinnate leaves are up to 4 inches in length. The foliage is yellow-green.
The fragrant, rounded, golden-yellow flowers are borne in small clusters with the leaves during spring.
The paired, white thorns, up to 3 inches, are straight and sharp.
The bark is mature trees is smooth, powdery and yellow-green. This is among the few trees in the world that can produce photosynthesis from the bark.
Hardy zones 9 to 11. Flood tolerant. It even thrives in the extreme summer heat of Yuma, Arizona, it also thrives in Phoenix and warmer parts of Tucson.

** photo of unknown internet source*

Posted by Randy Stewart Landscape at 1:01 PM 0 comments

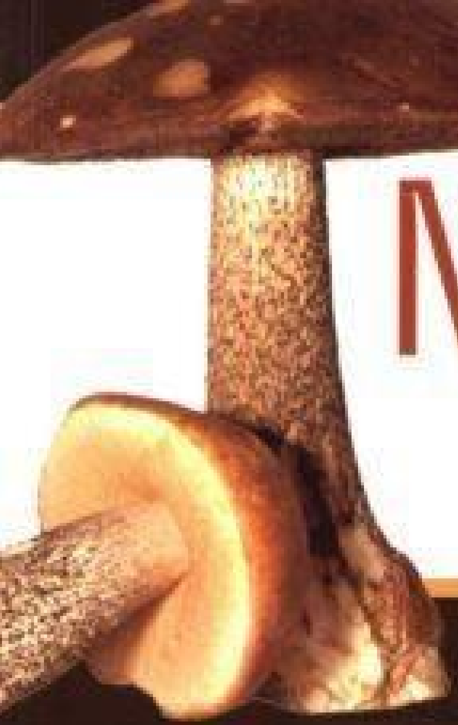

MUSHROOMING WITHOUT FEAR

The Beginner's Guide to Collecting Safe and Delicious Mushrooms

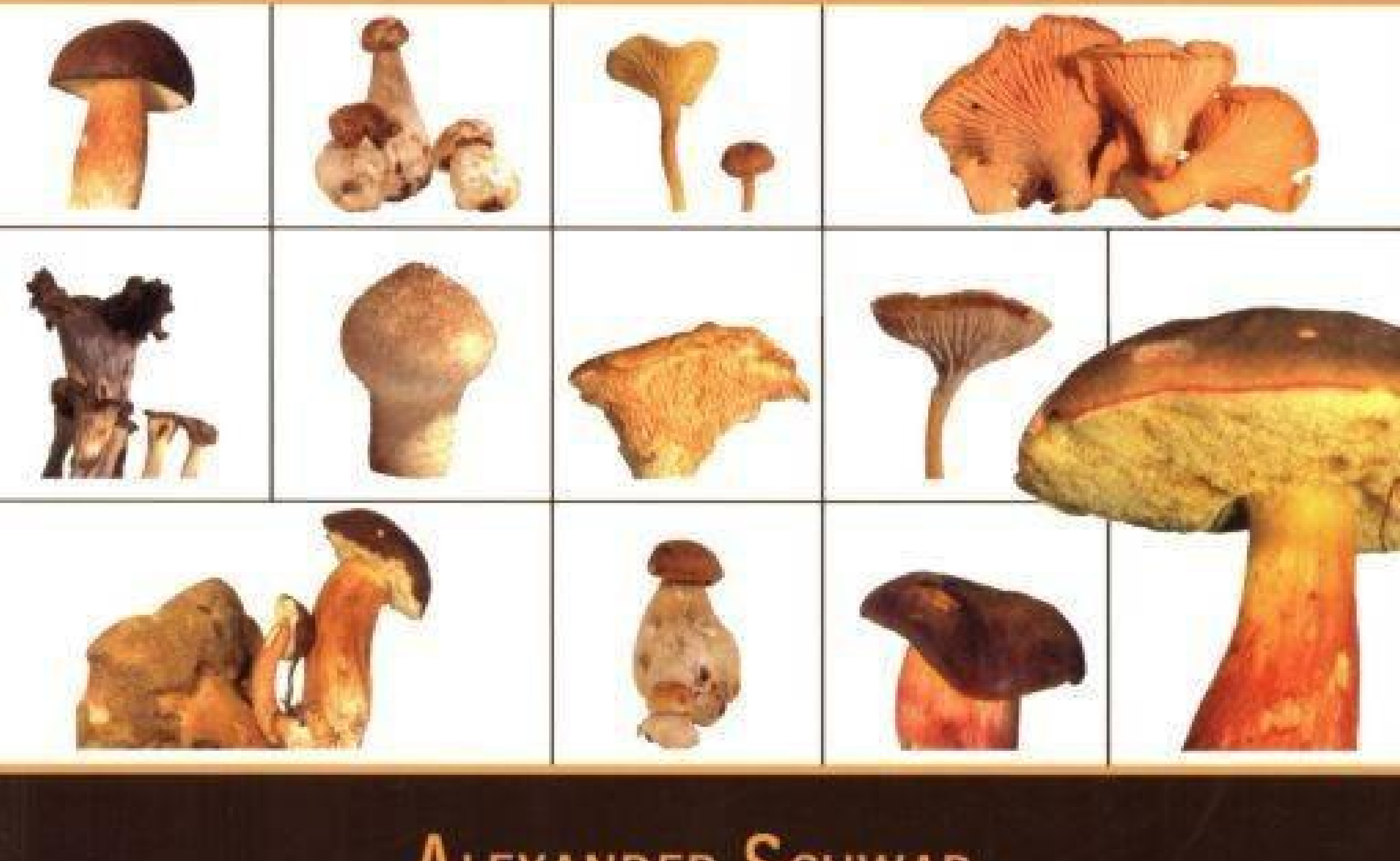

ALEXANDER SCHWAB

Table of Contents

Title Page
Copyright Page

Mushrooming Without Fear

It’s Safe, Easy and Fun
How to use this guide

- The size of mushrooms

What is a Mushroom?

- Gills
- Tubes
- Spines
- Ridges

How to Pick Mushrooms

- Rules:

Positively Identifying Mushrooms

Mushrooms With Tubes

- Cep or King Bolete
- Red Cracked Bolete
- Hen of the Woods
- Larch Bolete
- Bay Bolete
- Birch Bolete

Mushrooms With Ridges

- Chanterelle
- Trumpet Chanterelle

Mushrooms With Spines

    Hedgehog Fungus

Mavericks

    Common Puffball
    Horn of Plenty
    Cauliflower Mushroom

Mushrooming Miscellany

The Mushrooming Regions

    Eastern North America
    The Pacific Northwest
    California

Trees and Mushrooms

    Always remember

Handling, Storage and Cooking

    Handling
    Storage
    Cooking
    Basic preparation of dried mushrooms

Beyond the Stove

It is easy to do without silver and gold
and your toga and cloak
but it is difficult to give up mushrooms.

Martial, Roman poet (AD 40 - 102)
Epigrams, Book 13, Number 48

Mushrooming Without Fear

The Beginner's Guide to Collecting Safe and Delicious Mushrooms

Alexander Schwab

Consultants Monika Lehmann and Roy Mantle

Skyhorse Publishing

Text and photographs © by Alexander Schwab, 2006

First published by Merwin Unwin Books Ltd (Ludlow, UK) in 2006

All Rights Reserved. No part of this book may be reproduced in any manner without the express written consent of the publisher, except in the case of brief excerpts in critical reviews or articles. All inquiries should be addressed to Skyhorse Publishing, 555 Eighth Avenue, Suite 903, New York, NY 10018.

www.skyhorsepublishing.com

Library of Congress Cataloging-in-Publication Data is available on file.

ISBN-10: 1-60239-160-2
ISBN-13: 978-1-60239-160-4

10987654321

Picture credits:

Frank Moser: cep page 40, middle picture cep page 42, orange birch bolete, page 68
Thomas Beer: larch boletes bottom right and middle right, page 57
Niels Sloth (www.biopix.dk): pestle shaped puffball, page 90

Every effort has been made to insure the accuracy of the information in this book. In no circumstances can the publisher or the author accept any liability for any loss, injury or damage of any kind resulting from an error in or omission from the information contained in this book.

Printed and bound in India

Mushrooming Without Fear

It's Safe, Easy and Fun

Follow this guide and you can enjoy picking and eating delicious mushrooms uninhibited by fear and doubt. In order to enjoy a homemade gourmet mushroom dinner you do not really need to know what *basidia or hyphae* are. Nor do you need to know how to identify dozens of tasteless or poisonous mushrooms which you wouldn't want to eat in the first place.

Even advanced mushroom pickers face many uncertainties.
The same mushroom looks different in the various stages of its development and different again when conditions are, for example, very wet or very dry. There are no end of edible mushrooms. Twenty-odd of these are of serious culinary interest, of which about a dozen are outstanding. Quality, not quantity, is what you are after. By good luck, the quality mushrooms are also among the most easily identifiable mushrooms and some of these are among the most frequently encountered.

Mushrooms may look beautiful, but that does not mean that they are edible.

You wouldn't buy a soggy, wormy, half rotten mushroom in the supermarket. So why pick it in the wood? Quality is what the discerning mushroom hunter is after.

With the safe method explained in this book, you will be able to identify the best mushrooms positively by their unique and unmistakable features.

Your part in the safe method is:

1. As a beginner, leave alone most mushrooms you pass
2. Look closely at what you see before you, not at what you wish was there
3. Stick to the rules and tick off every stage in the mushroom identification process

The fly agaric is the most well-known species. For gnomes, pixies, fairies and other such folk it is safe. For us common mortals it is poisonous! Do not believe any old wives' tales regarding the identification of poisonous mushrooms. Stick to the safe method in order to identify the best edible mushrooms.

Why is the safe method safe? If you follow the safe method to the letter, you'll automatically steer clear of deadly mushrooms or mushrooms causing permanent damage.

In the families of mushrooms presented and recommended here, there is not one deadly mushroom.

This is the reason why the field mushroom is not in this guide.
The field mushroom belongs to the group of mushrooms with gills, and especially when young, the feld mushroom can be confused with some deadly species. The safe method therefore asks you to ignore all mushrooms with gills because it is within the gilled group that all the seriously poisonous and deadly mushrooms are found. If you're absolutely sure about the field mushroom, then pick it by all means. But unless you are 100 percent certain, hunt for field mushrooms and other species with gills in the supermarket only.

Within the recommended groups you will find some slightly poisonous mushrooms but these happen to be so bitter, hot-tasting or strong-smelling that nobody sane would eat them anyway should they find their way into a meal (which they won't if you follow the safe method).
Poisonous mushrooms in the recommended groups cause gastrointestinal trouble of varying intensity. In other words: tummy ache.
Nevertheless there is absolutely no reason for complacency: a harmless tummy ache together with an existing medical condition can be dangerous.

Now, enough with the warnings. With this guide in hand, you are sure to have a safe and memorable experience collecting delicious mushrooms. Good luck and have fun!

How to use this guide

1. Read the entire book twice in order to understand the approach
2. Study the “Gills,” “Tubes,” “Spines” and “Ridges” sections until you’re confident you can correctly distinguish these features; then you’re ready for a foray
3. Double-check at home by going through the full identification process step-by-step

The size of mushrooms

The size of mushrooms varies enormously, so any measurements given in this book are only a rough average. Don’t be put off if you see an exception. The size can vary disproportionately due to weather conditions: if it is very dry, the mushrooms will be smaller than average. The other great variable is the growing medium. Don’t be surprised if near a building site, for example, you find surprisingly large mushrooms growing out of sawdust, or where woodchips have been laid at the edges of paths.

What is a Mushroom?

A mushroom is a highly complex organism. Here are the essentials:

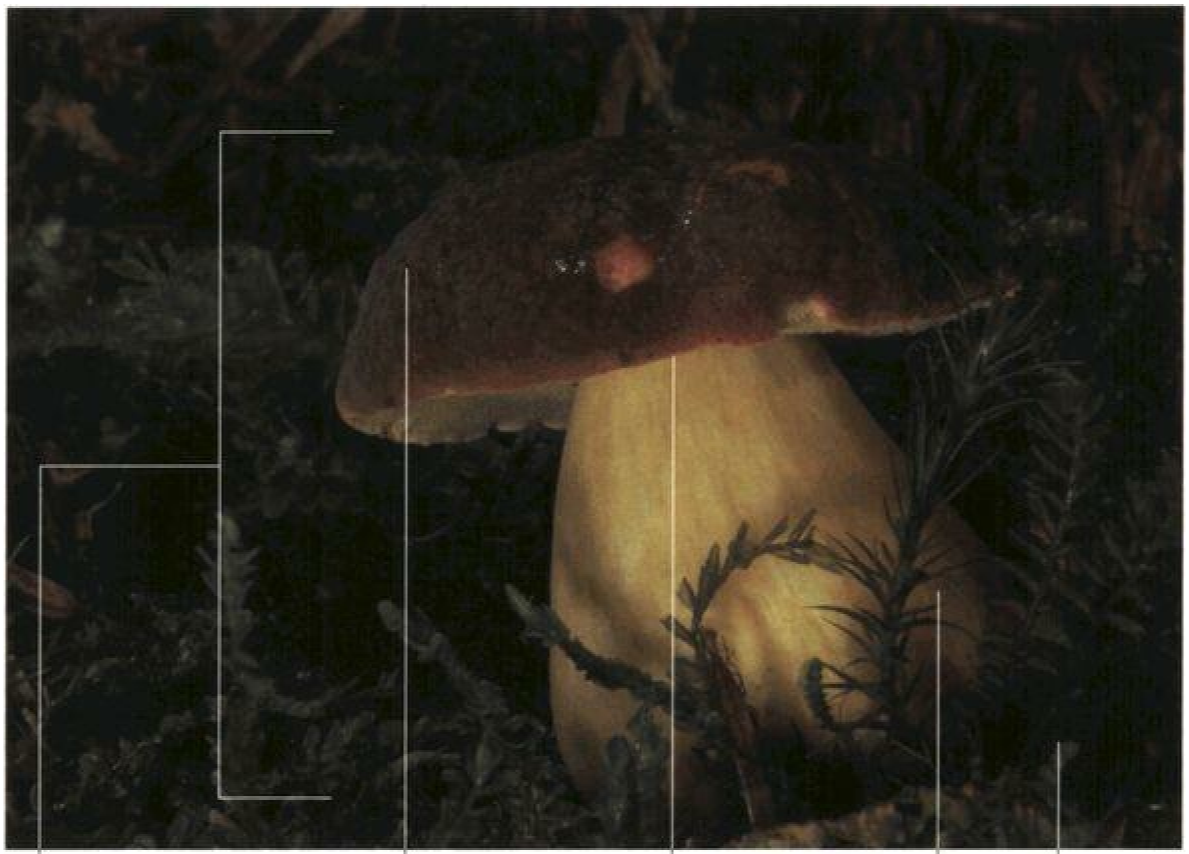

Fruitbody

The technical term for what is colloquially called the "mushroom" is the "fruitbody." Mycologically speaking, the real mushroom is the mycelium which is a fine structure of filaments on or in the forest soil. What we see as mushrooms (the fruitbody) are like the apples on an apple tree.

Cap

The caps of mushrooms have many different forms but what we're interested in is what is under the cap. Under the cap you'll find either . . .

Gills

Tubes

Spines

Ridges

Gills, tubes, spines and ridges produce, contain and release the spores. Wind and rain disperse the spores. Given the right conditions new mycelia and mushrooms form from the spores.

Stalk

Mycelium

The mycelium

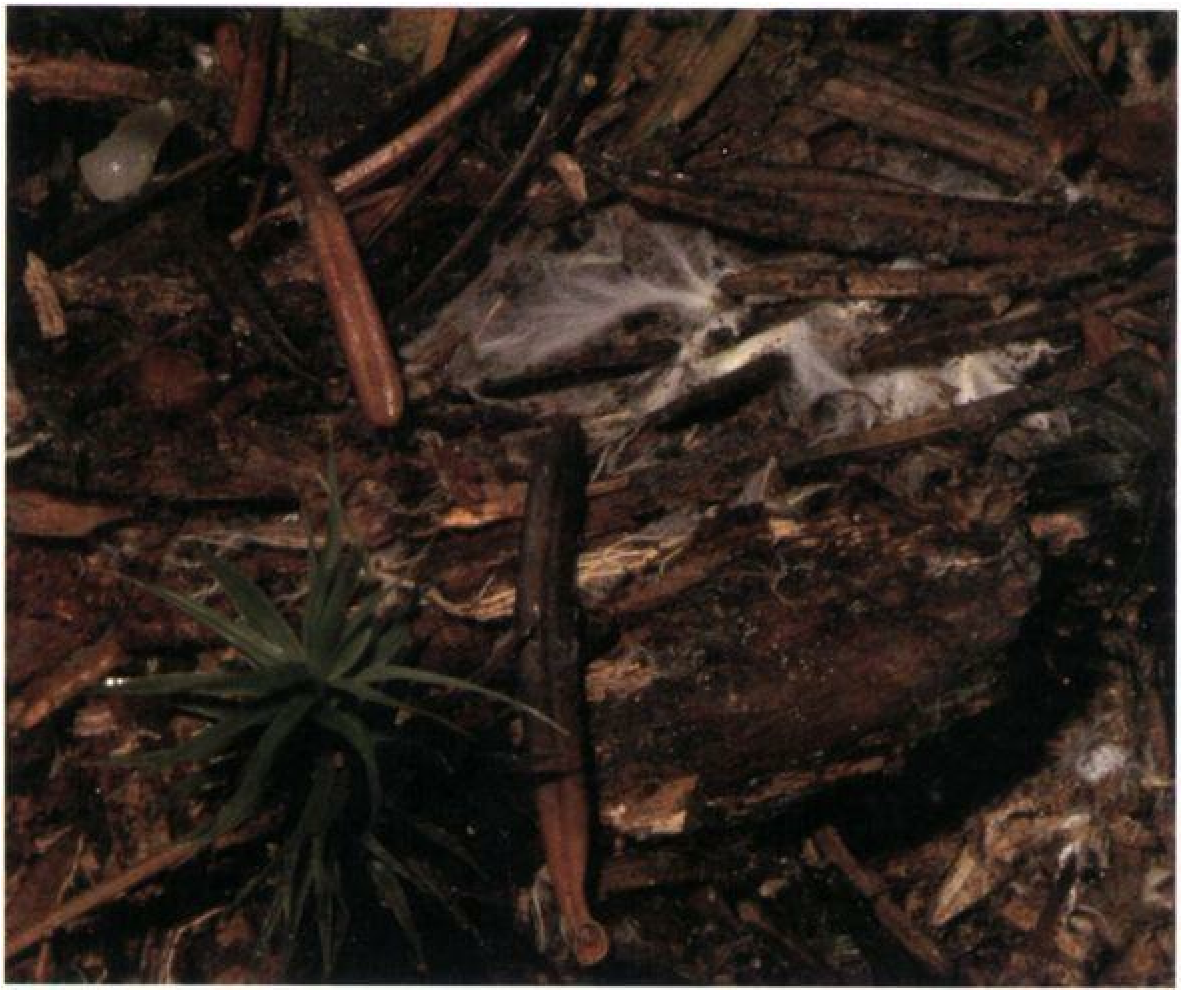

This is a close-up of about 4 mm of the mycelium above, as seen under a microscope. When conditions are right, the mycelium forms a fruitbody, pushing the mushroom up in order to disperse the spores or to wait for you to come along and pick it.

Rule number 1

Never, never take a mushroom with gills!!!

Gills

Identifying gills is the first and most important step. You must be sure that you're able to identify gills. Caps of gilled mushrooms come in many different shapes. Gills are the radiating blades on the underside of the cap. They fan out in a distinctly regular way. Gills have precise forms and come in many colors. Some of them are brittle some of them are soft. They can be rubbed off or separated from the underside of the cap quite easily. Gills are always attached to the stem or to the cap in a uniform way.

In perfect conditions the distinctly regular way in which the gills fan out is clearly visible.

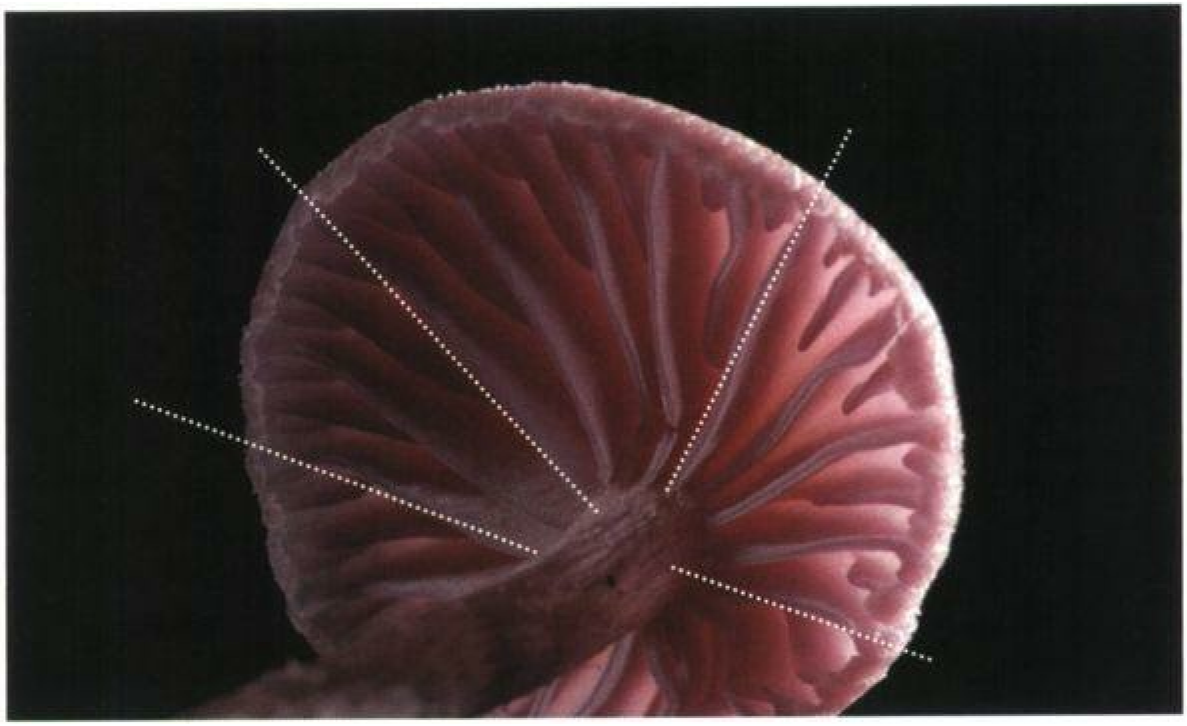

Weather-beaten or old gills might be damaged or broken but on closer inspection their regularity will become apparent. Damage or no damage: either way they look as if drawn with a ruler.

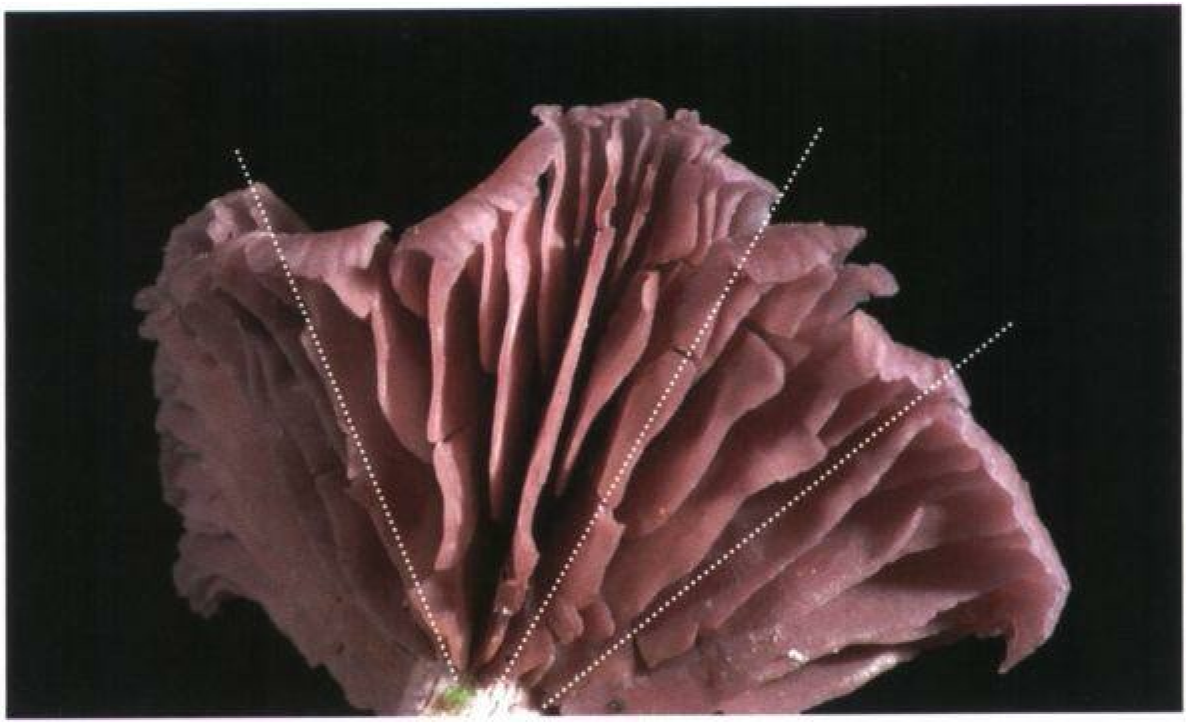

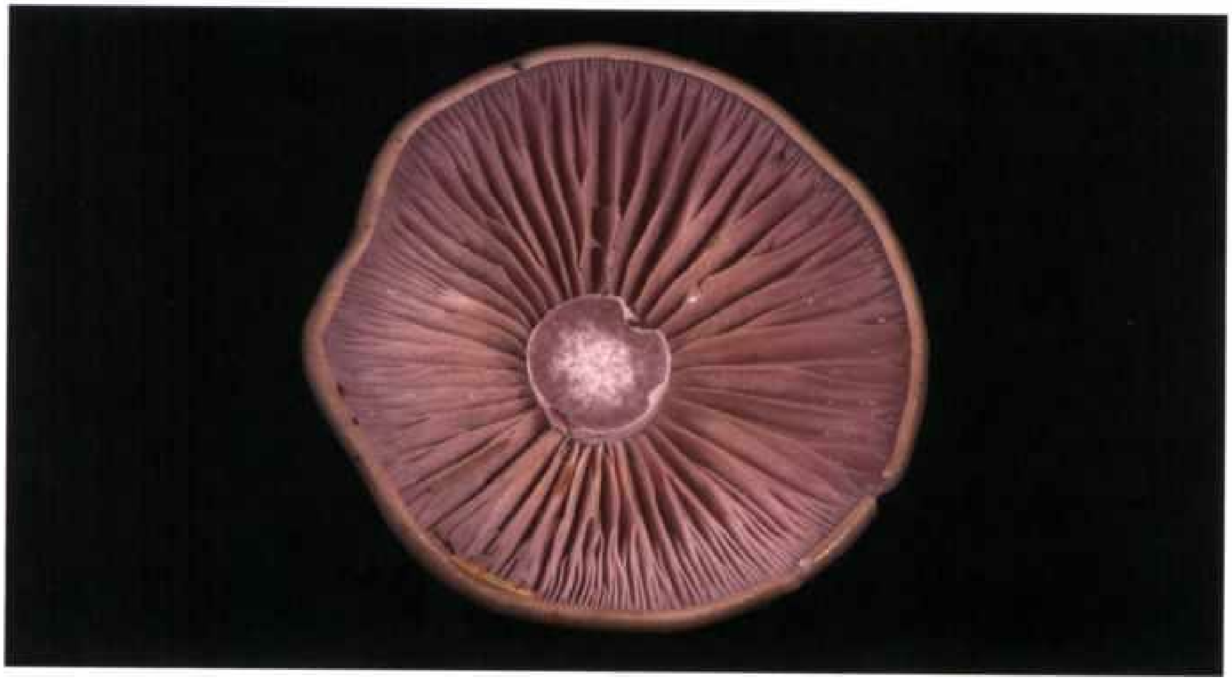

In their regular way, gills differ in spacing and formation.
For example:

Gills can fork once or more.

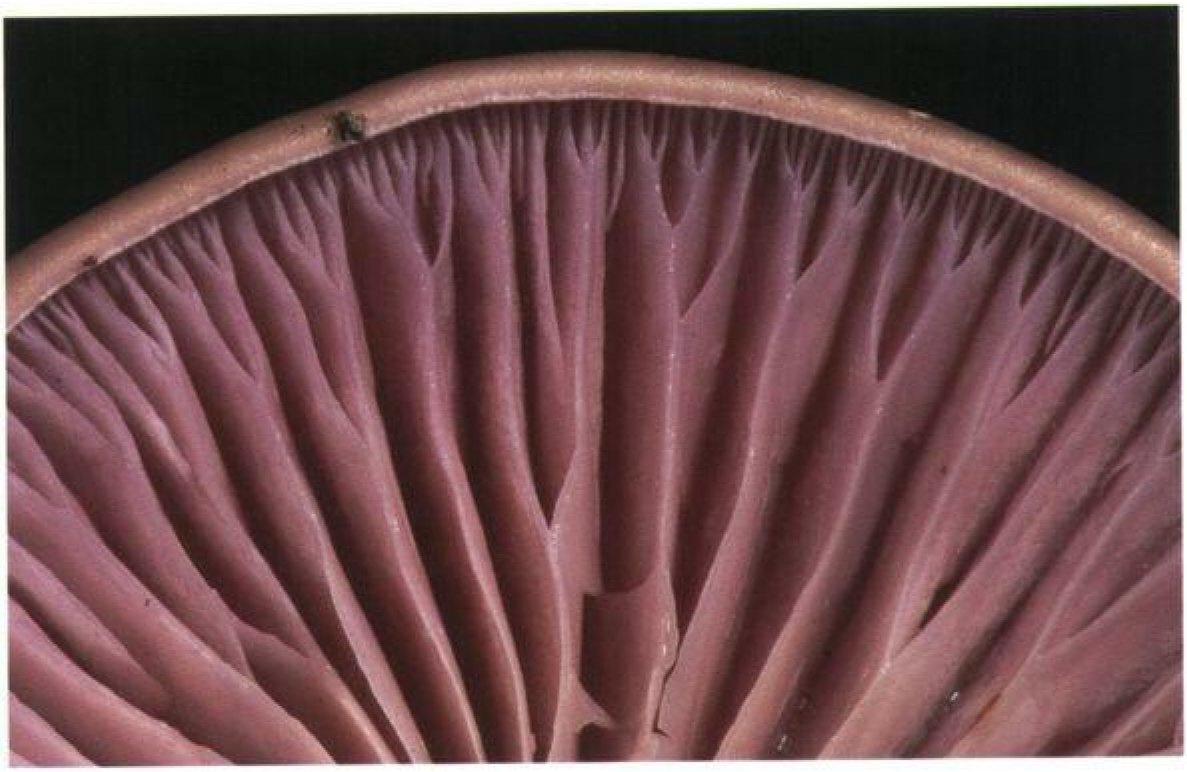

Gills can be mixed: long and short gills.

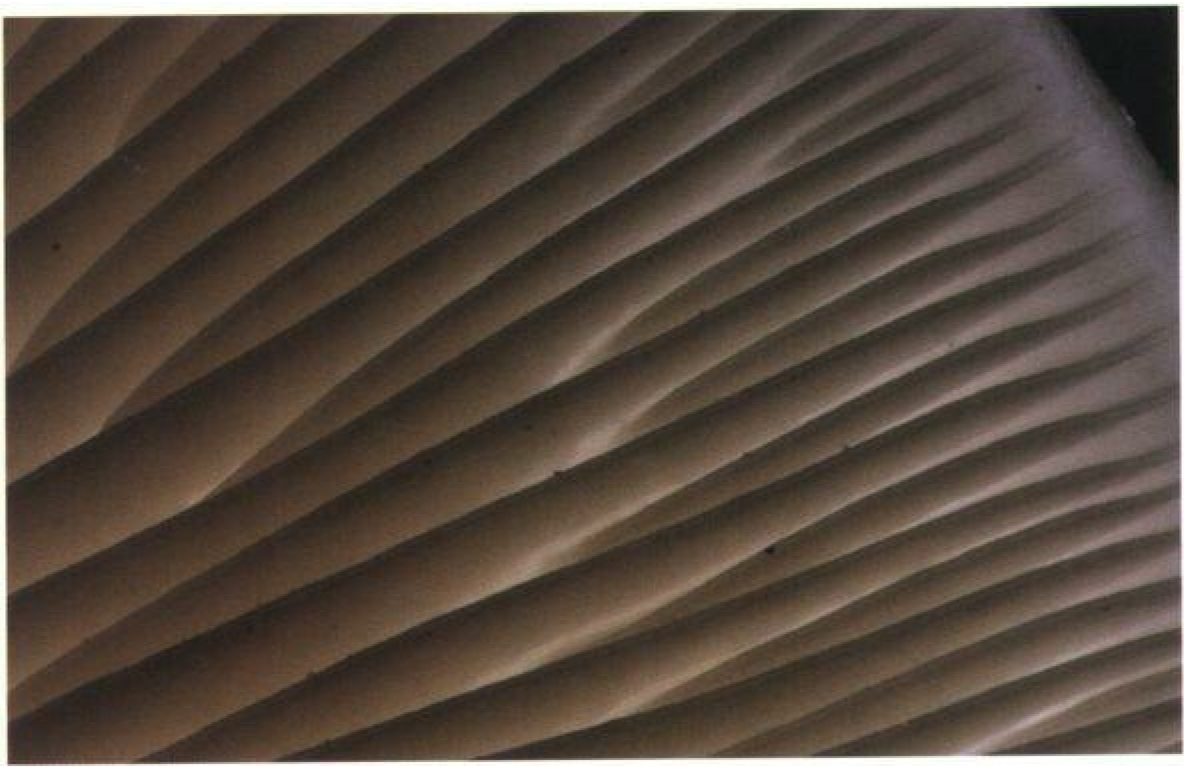

Gills can be crowded.

Gills can be distant.

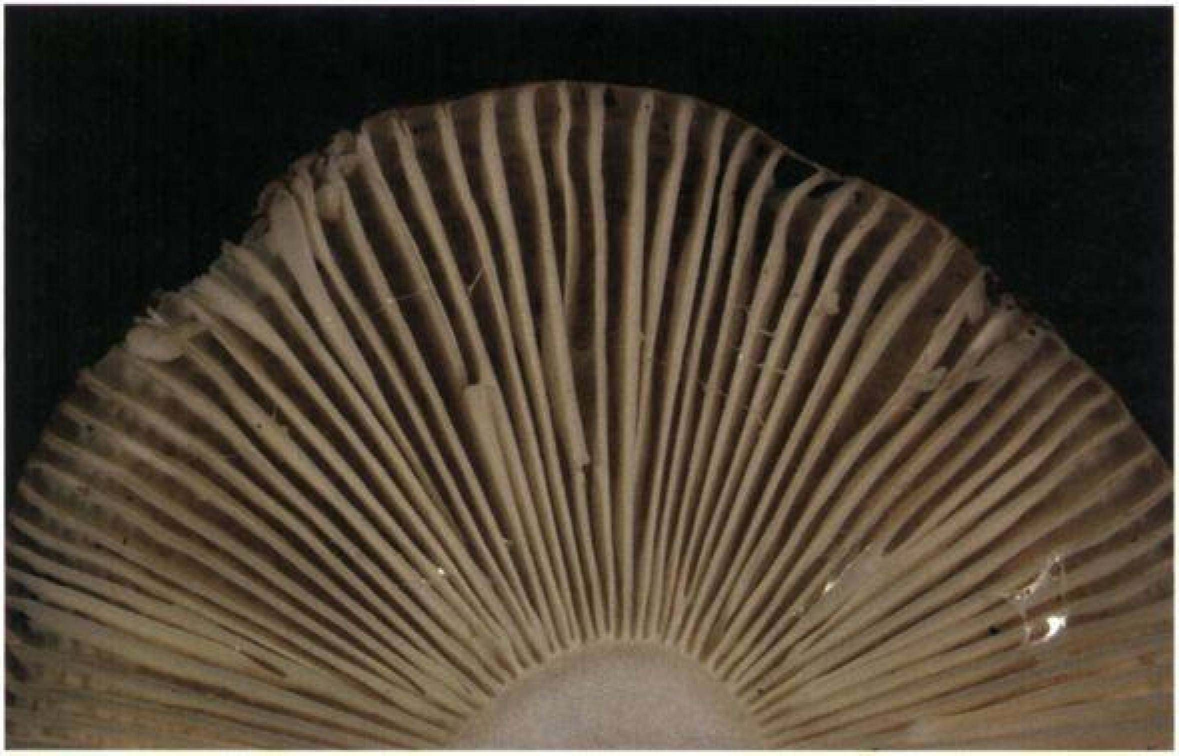

Gills are always uniformly attached to the cap or the stem. They start radiating from the same height along the stem or the same line around the cap.

Rule number 2

> **Only take mushrooms with tubes, spines and ridges and the mavericks portrayed in this book**

Tubes

Tubes are fine and tightly packed on the underside of the cap. The openings at the ends of the tubes are called pores.

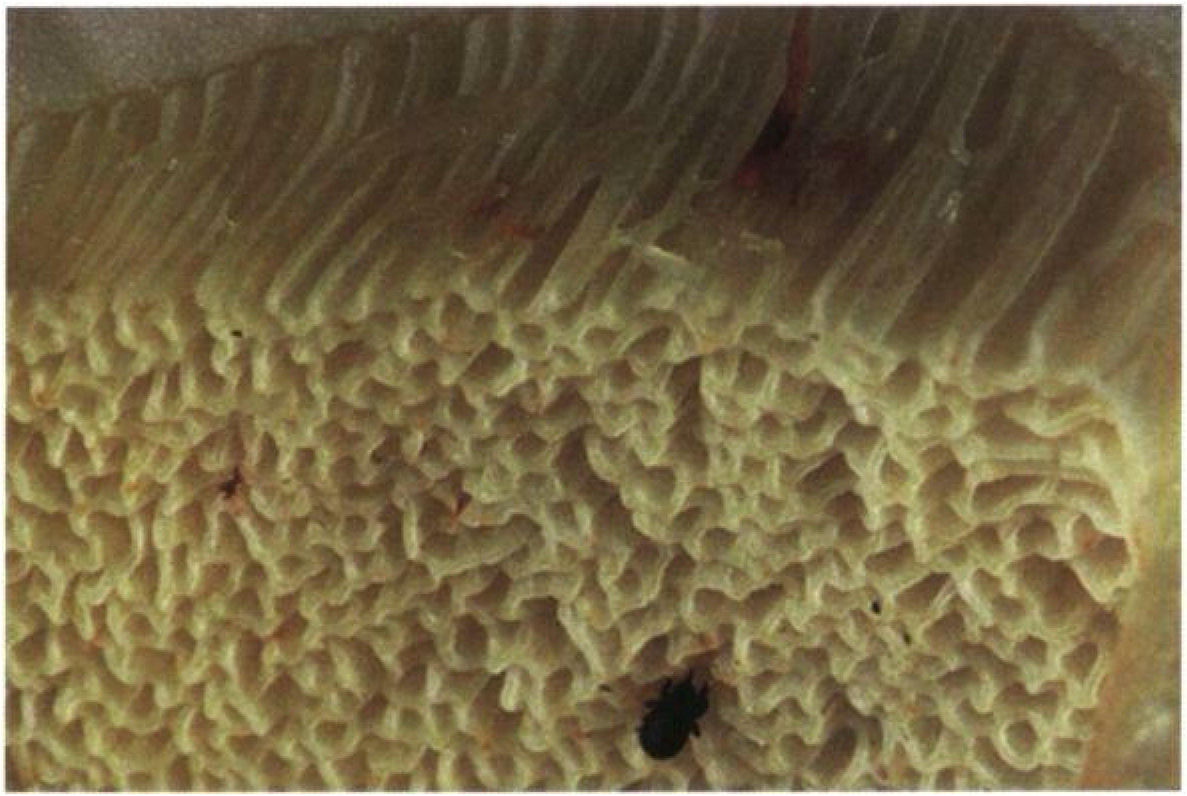

Spines

Spines hang like stalactites from the cap.

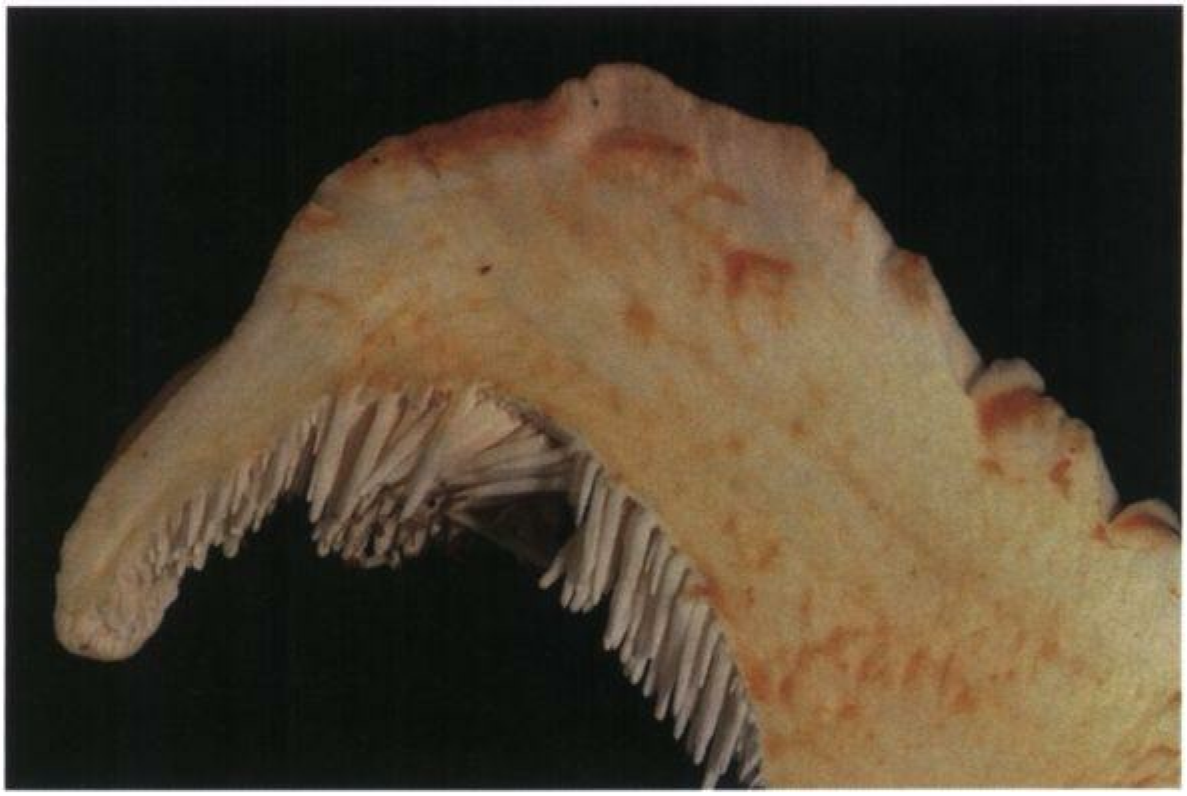

Ridges

Ridges are cross-veined, irregular and don't form a distinct pattern. Ridges are not attached to the cap: they are part of the cap.

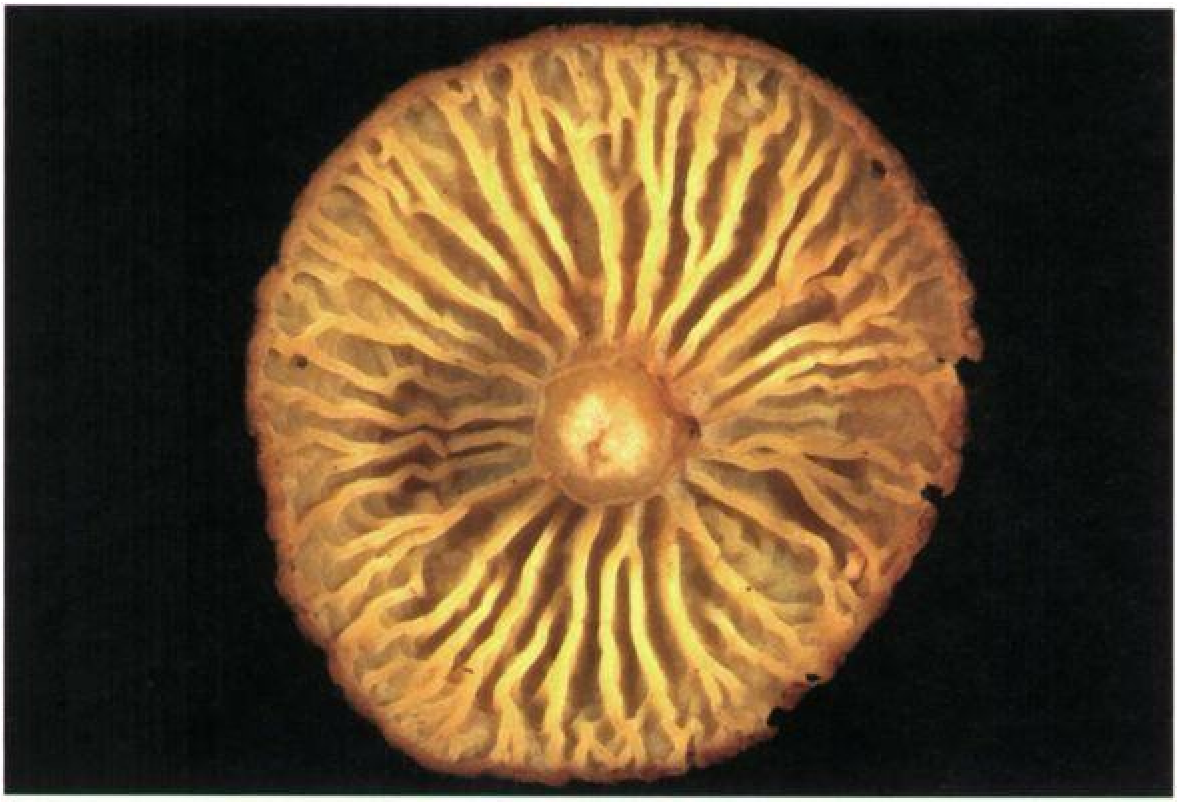

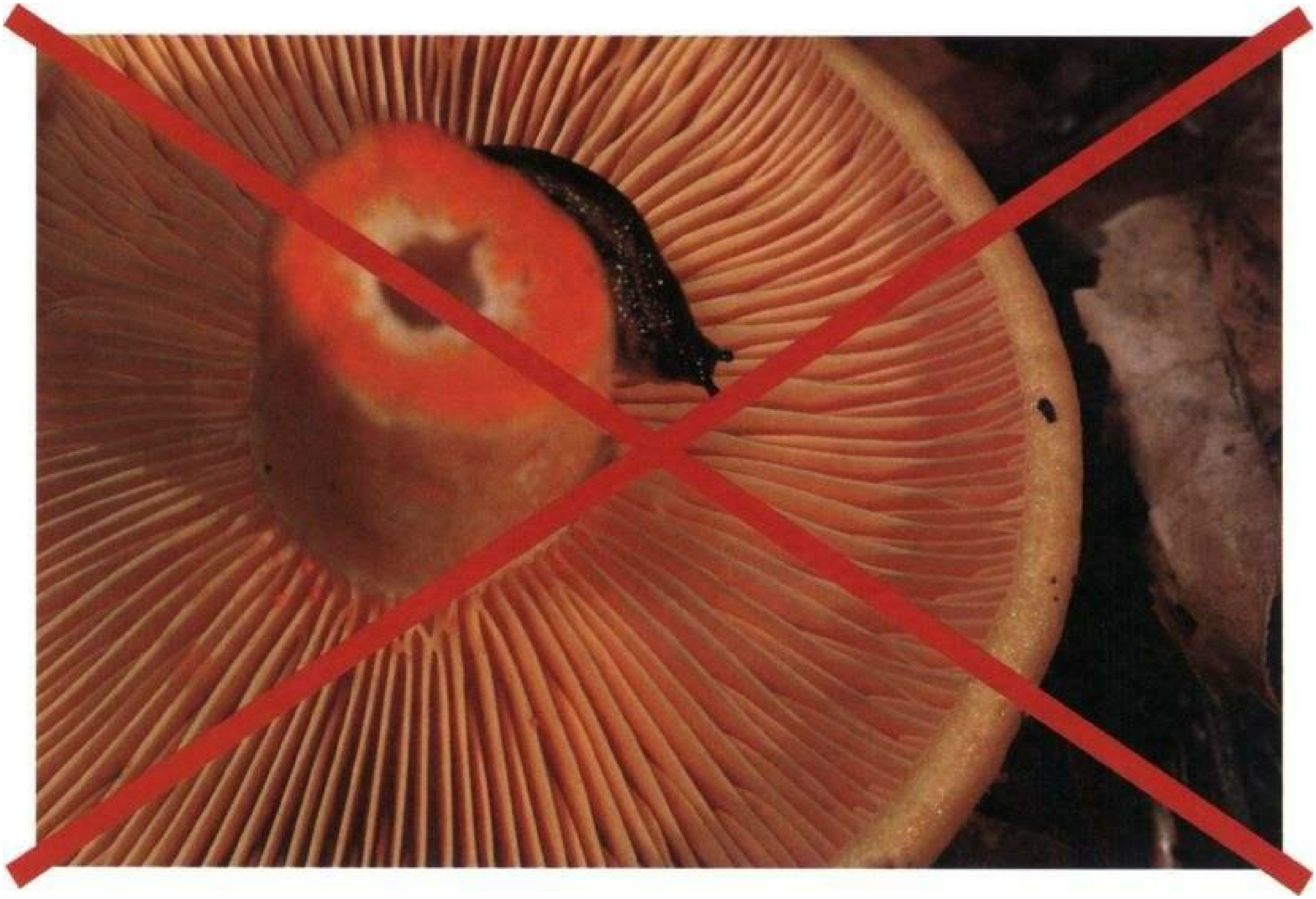

Remember Rule number 1

Never, never take a mushroom with gills!!!

Tubes

This is the underside of a cap with tubes. The tubes are fine and tightly packed and the openings at the end which are seen here are called pores. Tubes can be removed easily from the cap.

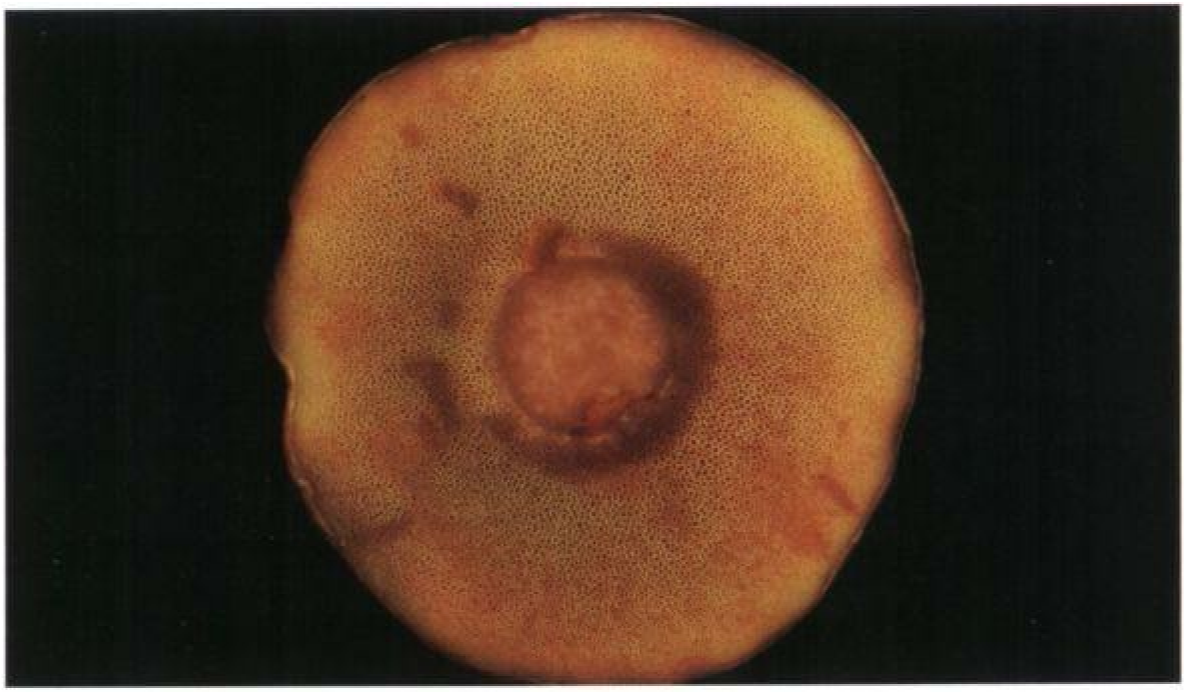

Pores form a regular pattern resembling a sponge.

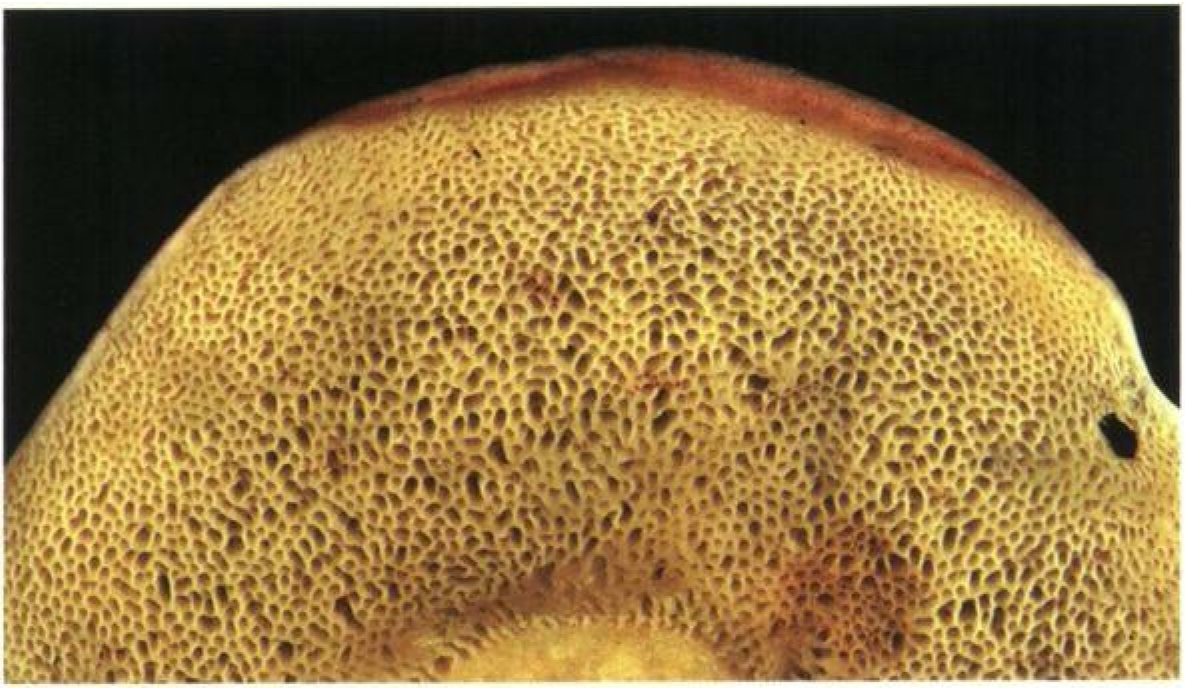

This cross-section shows the tubes and the pores.

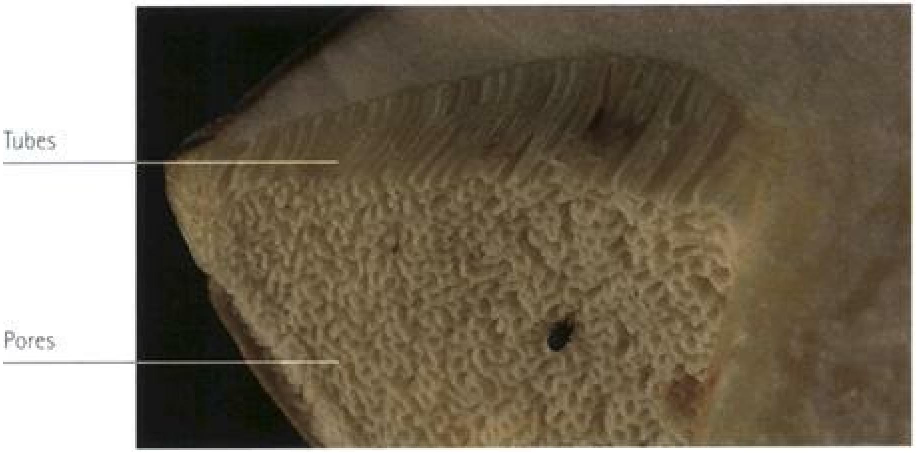
Tubes
Pores

Spines

Spines hang like stalactites from the cap.

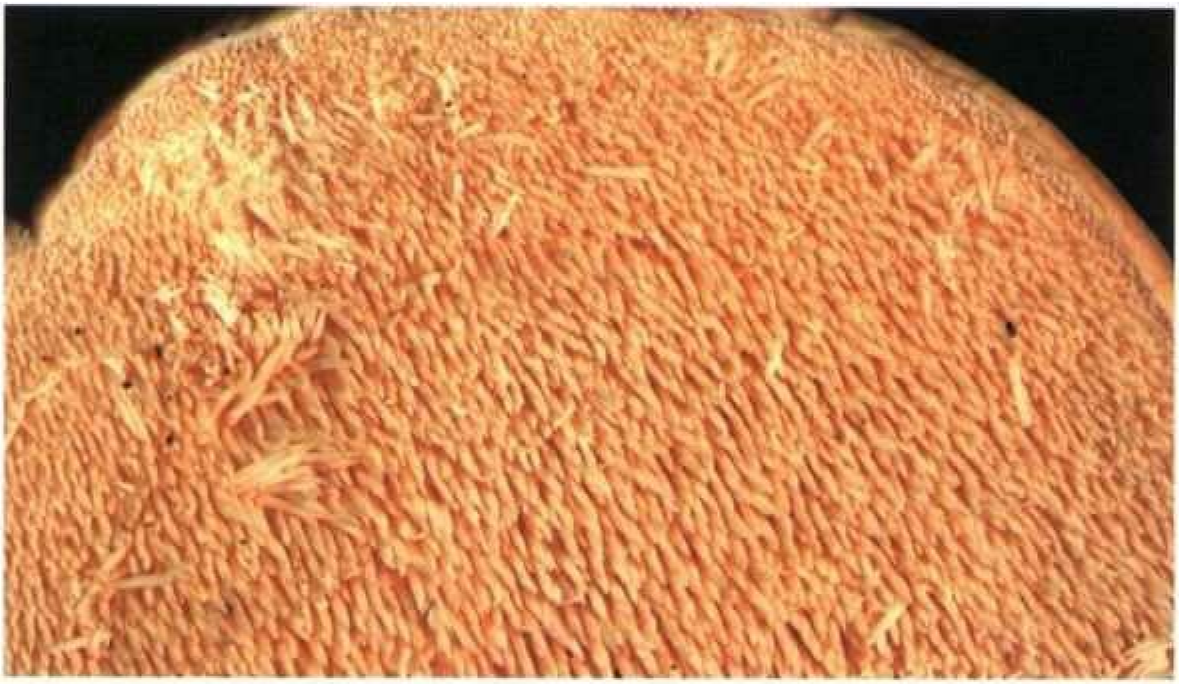

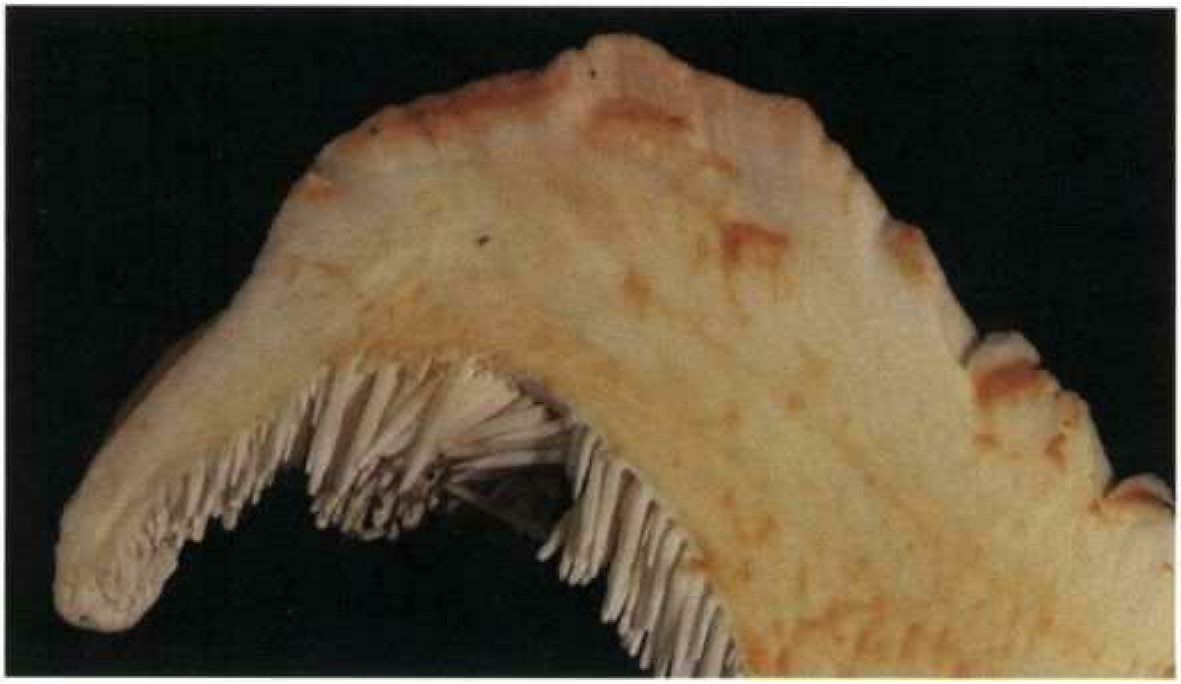

Ridges

Ridges are on the underside of the cap and are part of it. They are not just attached to the cap or stem: they're part of them, which is why you can't rub them off or pull them off easily.

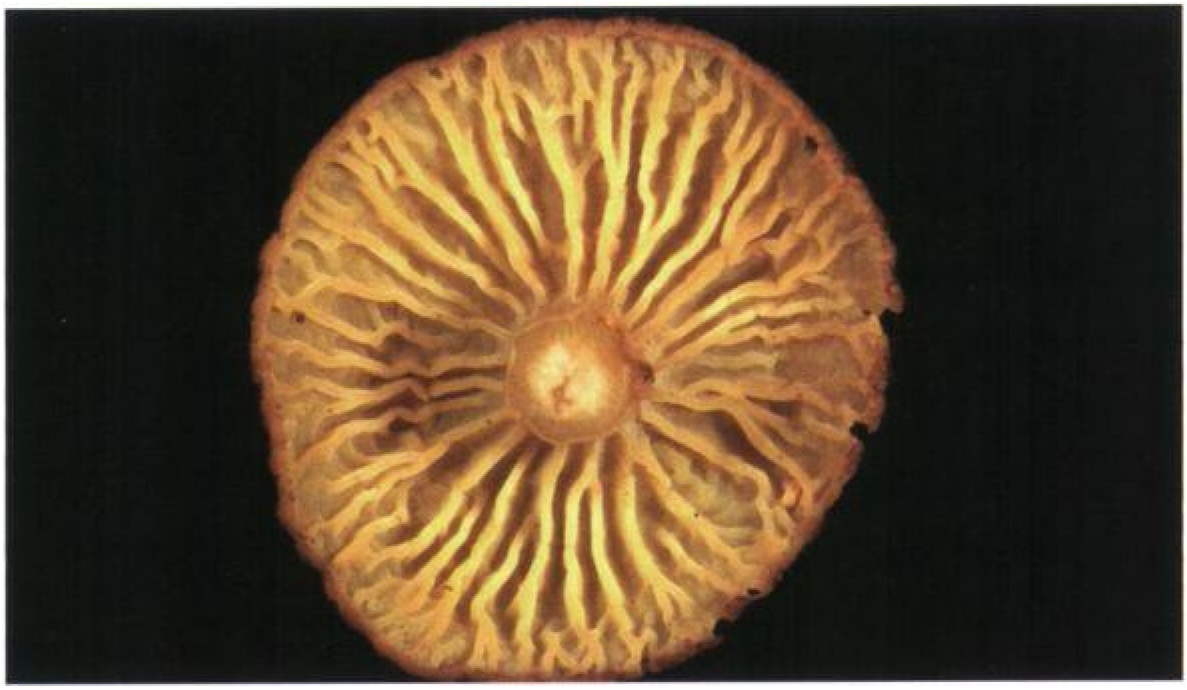

Ridges are cross-veined, irregular and they don't form a set, uniform pattern.

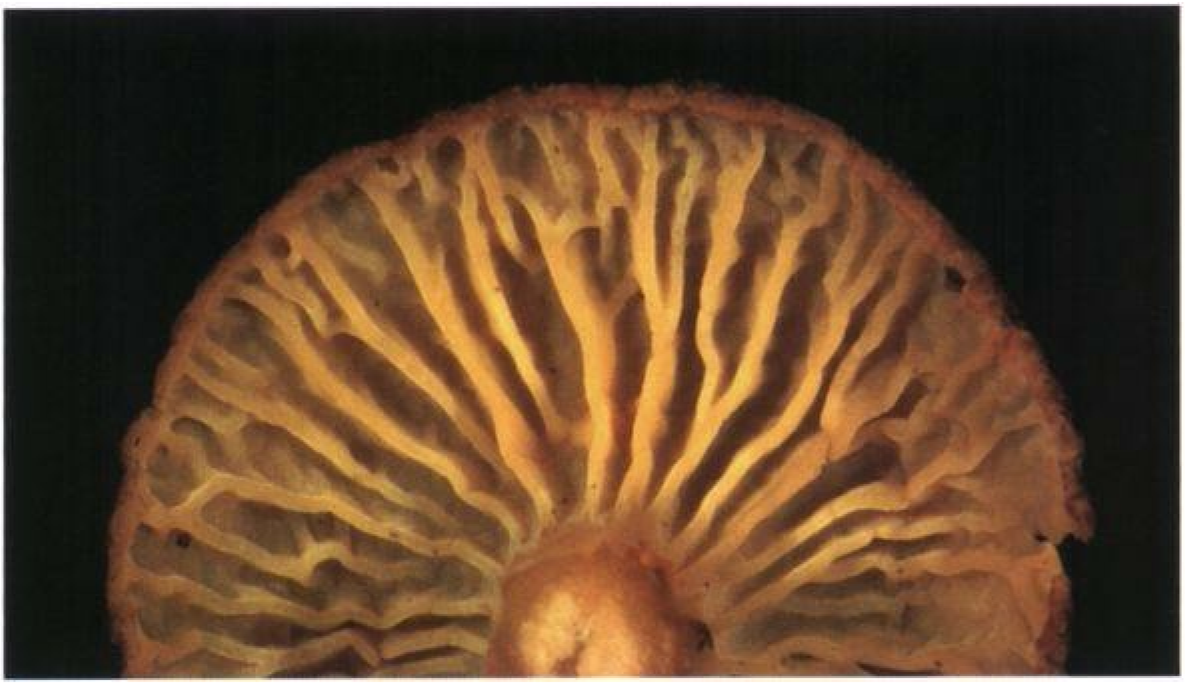

They are not attached to the stem in a distinct and regular way, as gills are: they are part of the stem.

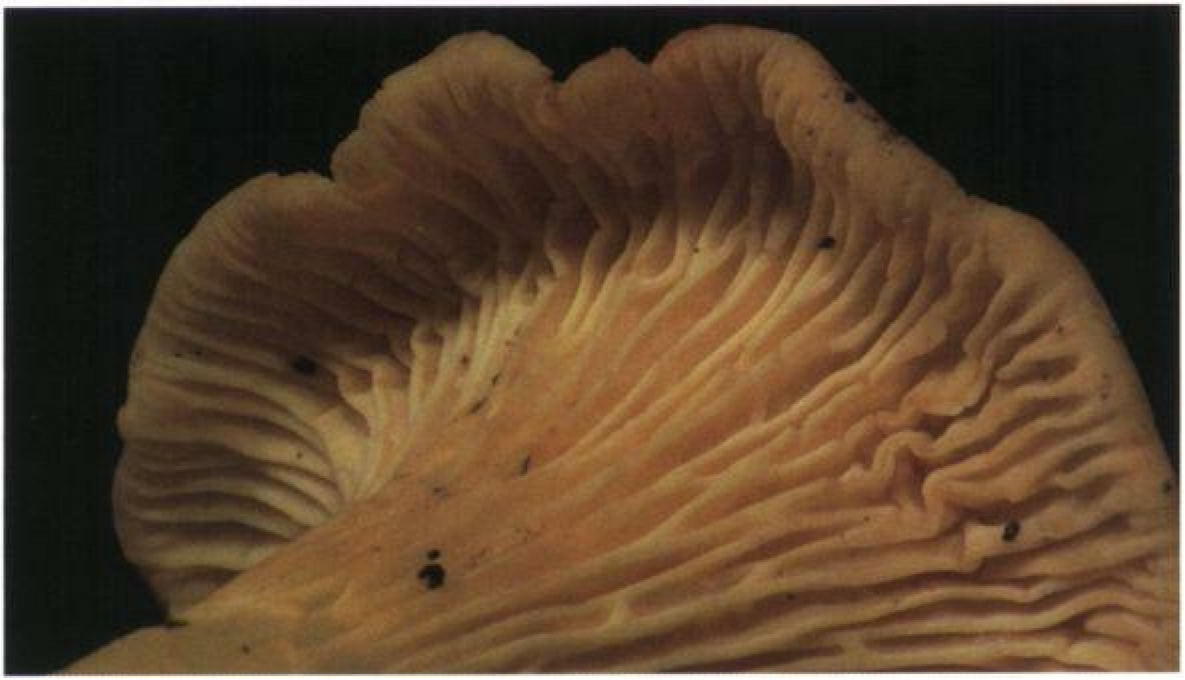

The ridges are part of the stem. There is no distinct pattern in the way they grow out of the stem. Some ridges begin further up, some further down the stem. There is no regularity.

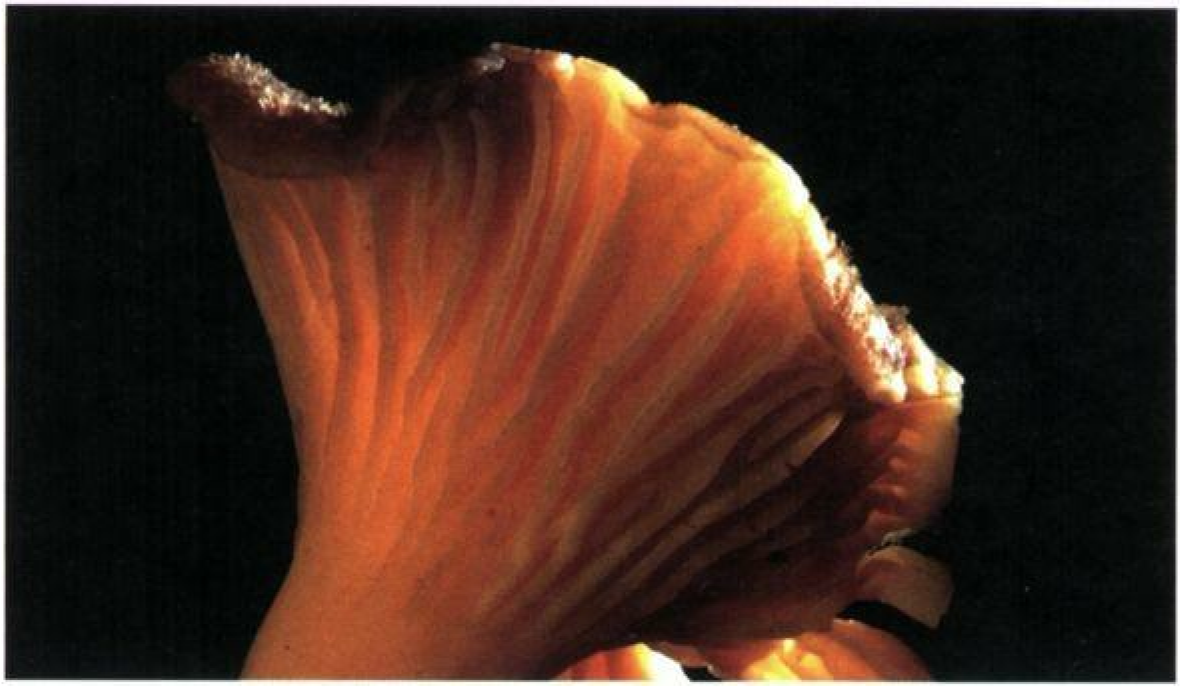

In contrast to gills and tubes, the section of a mushroom with ridges shows no typical features. Because of the irregular nature of ridges each section will look different whereas with tubes and gills you will find the same features each time.

How to Pick Mushrooms

Rule number 1 says not to take anything with gills. This also means you do not touch or cut every mushroom you spot in order to check for gills.

In nine out of ten cases you can see whether a mushroom has gills or not without even touching it. Looking under the cap by kneeling down provides good exercise.

If you really can't see what's underneath a cap, break off just a little to see what's there. If you want to be absolutely safe, you could use surgical gloves or a paper napkin which you discard later.

If you have established that you have, say, a mushroom with tubes and it is in reasonable condition and it tallies with the field identification card—then, and only then, should you pick it.

Rules:

Rule number 1

Never, never take a mushroom with gills!!!

This is our life insurance.

Rule number 2

Only take mushrooms with tubes, spines and ridges and the mavericks portrayed in this book

This means thoroughly understanding the information on pages 18-27 of this book.

Rule number 3

Only eat mushrooms which you have clearly identified with ALL the positive ID marks

The mushrooms you take must be a certain size in order to show all the identification marks. In their baby stage, so to speak, some deadly and poisonous mushrooms are almost indistinguishable from harmless species.

Rule number 4

If a mushroom smells rotten, it is rotten and if it feels soggy, it is soggy

There are a surprising number of people who believe that somehow a half decomposed, soggy, smelly mushroom will improve on the way home. It never does. Only take mushrooms which are in mint condition, that is, the flesh is firm and there are hardly any worms.

Rule number 5

Never, never eat wild mushrooms raw

The reasons for this are manifold but, to state the obvious, some areas in some woods are dog-ridden.

Rule number 6

Look before you cut

Always cut mushrooms, do not pull them out. You don't chop down an apple tree to pick the apples. Cut just above the ground and cover the remaining stem with a little mud or dead leaves. Whether this is of any conservational value is disputed but it does no harm. My recommendation is to cut and cover.

Picked mushrooms want to breathe. Do not suffocate them in plastic bags. Use a wicker or rattan basket or a linen bag. If possible use a proper mushroom knife with an integrated brush for cleaning. As the first cleaning of the mushroom should be done in the woods, the brush is very useful. However, any knife will do, as long as the blade can be folded into the knife handle, for safety. Common sense is your best adviser but be aware of any likelihood of ticks where you go mushrooming.

You might want a stick. It can be any kind of stick and you will find it helpful on the slopes. But you will also find it useful for pushing aside the grass and ferns, affording you a better look at the ground. Different mushrooms can be put together in your wicker basket. Better still, have a few paper bags in the basket; one paper bag (breathable) for each species. Putting, say, all trumpet chanterelles in the same bag helps you to spot the odd one out which you have inadvertently cut while not fully concentrating (see page 78).

Rule number 7

Mushrooms want to breathe: Do not suffocate them in plastic bags

Either in the woods or at home.

Rule number 8

If in doubt, leave it

Now you’re ready for a foray.

Positively Identifying Mushrooms

Always remember

Rule number 1

Never, never take a mushroom with gills!!!

Rule number 2

Only take mushrooms with tubes, spines and ridges and the mavericks portrayed in this book

Rule number 3

Only eat mushrooms which you have clearly identified with ALL the positive ID marks

Rule number 4

If a mushroom smells rotten, it is rotten and if it feels soggy, it is soggy

Rule number 5

Never, never eat wild mushrooms raw

Rule number 6

Look before you cut

Rule number 7

Mushrooms want to breathe. Do not suffocate them in plastic bags

Rule number 8

If in doubt, leave it

Mushrooms With Tubes

Cep or King Bolete

Boletus edulis

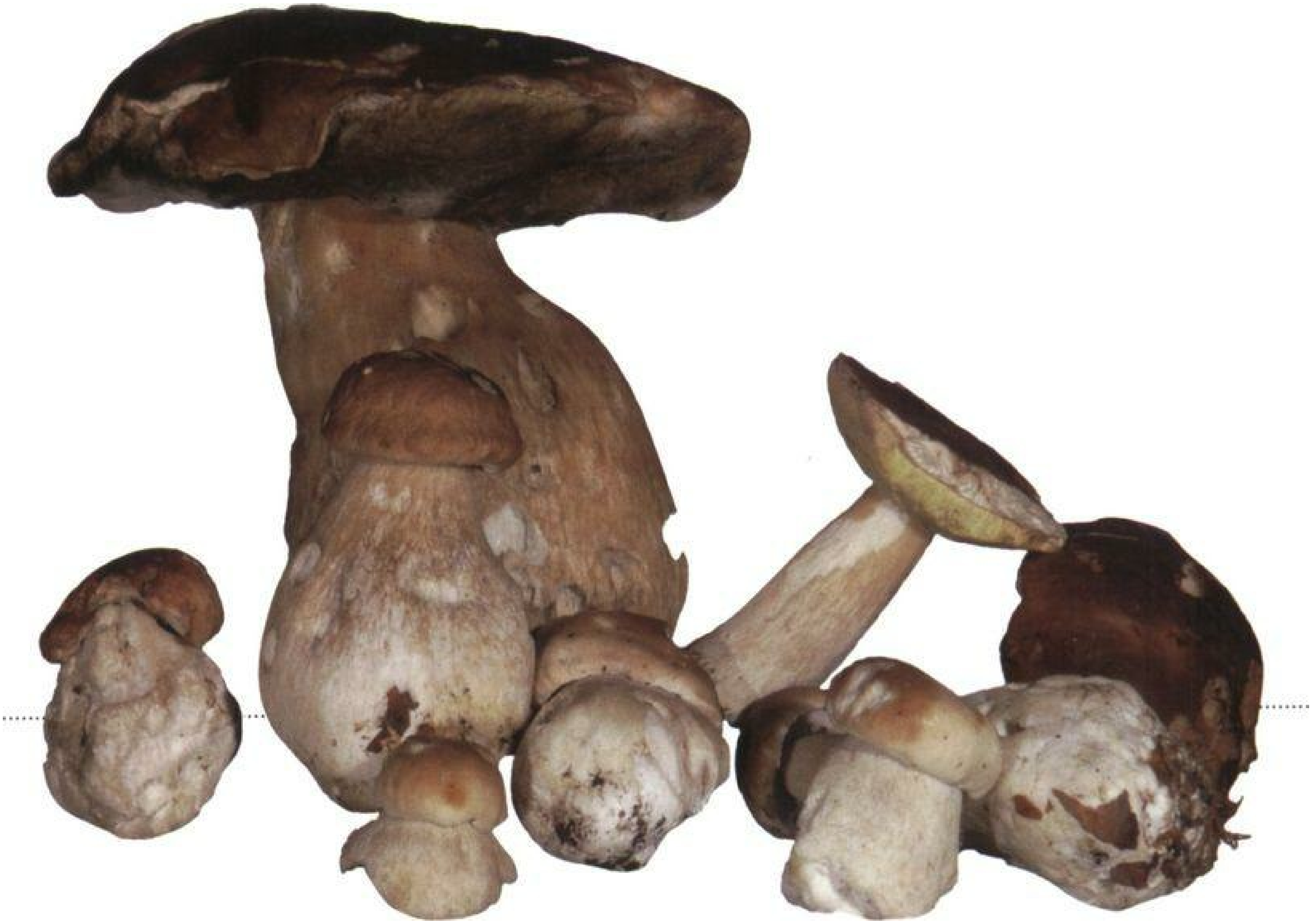

The cep is the king of mushrooms. Some truffles are rarer, pricier and confined to relatively small areas in France and Southern Europe. The cep on the other hand reigns supreme everywhere. Ceps can be found in the same location year after year. They can disappear in certain areas for a couple of years only to return spectacularly in masses.

Some people gather only ceps which is understandable because of their beauty, taste and the thrill of finding them on the same spot time after time. Ceps vary greatly in general appearance, color and size. The size of the little fellow in the foreground of the picture above is about 2 inches in height whereas the giant is about 14 inches in height.

If conditions are right (starting off with a nice wet and humid spring or early summer) you should begin to look for ceps as early as the beginning of July. July and August are the months of the summer ceps.

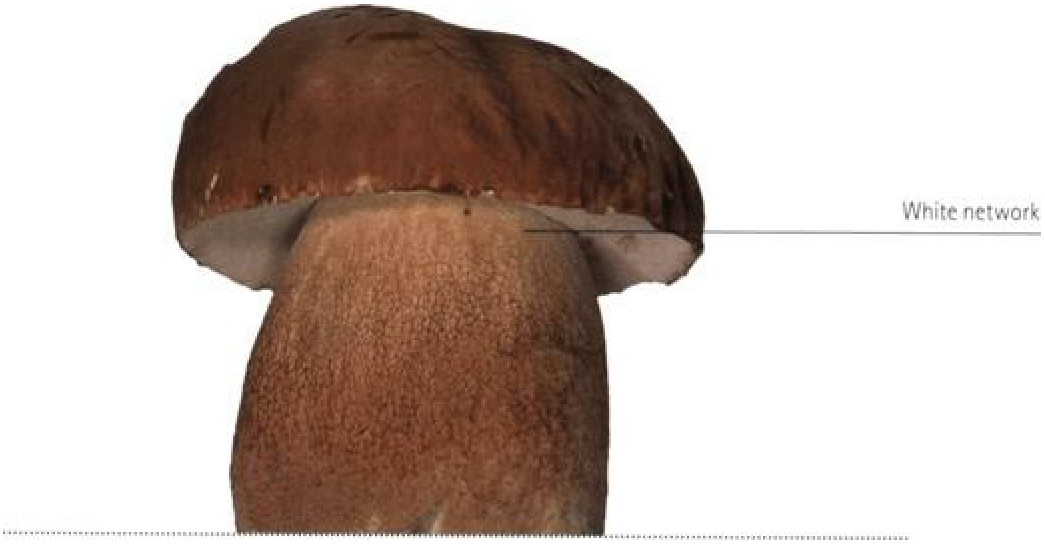
White network

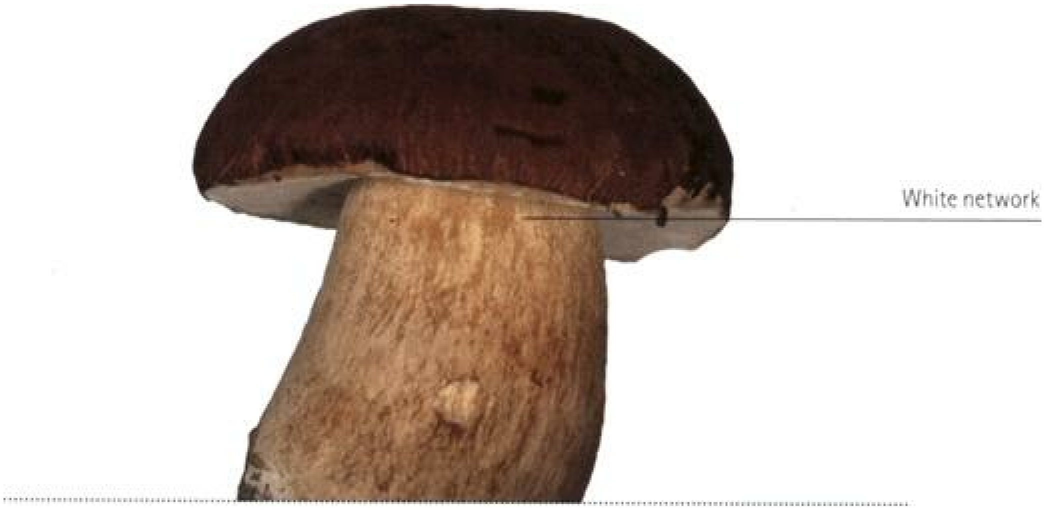
White network

These are summer ceps. However all ceps, summer or autumn, show a fine white network on the top of the stem right underneath the tubes.

A classic autumn cep, the king of mushrooms.

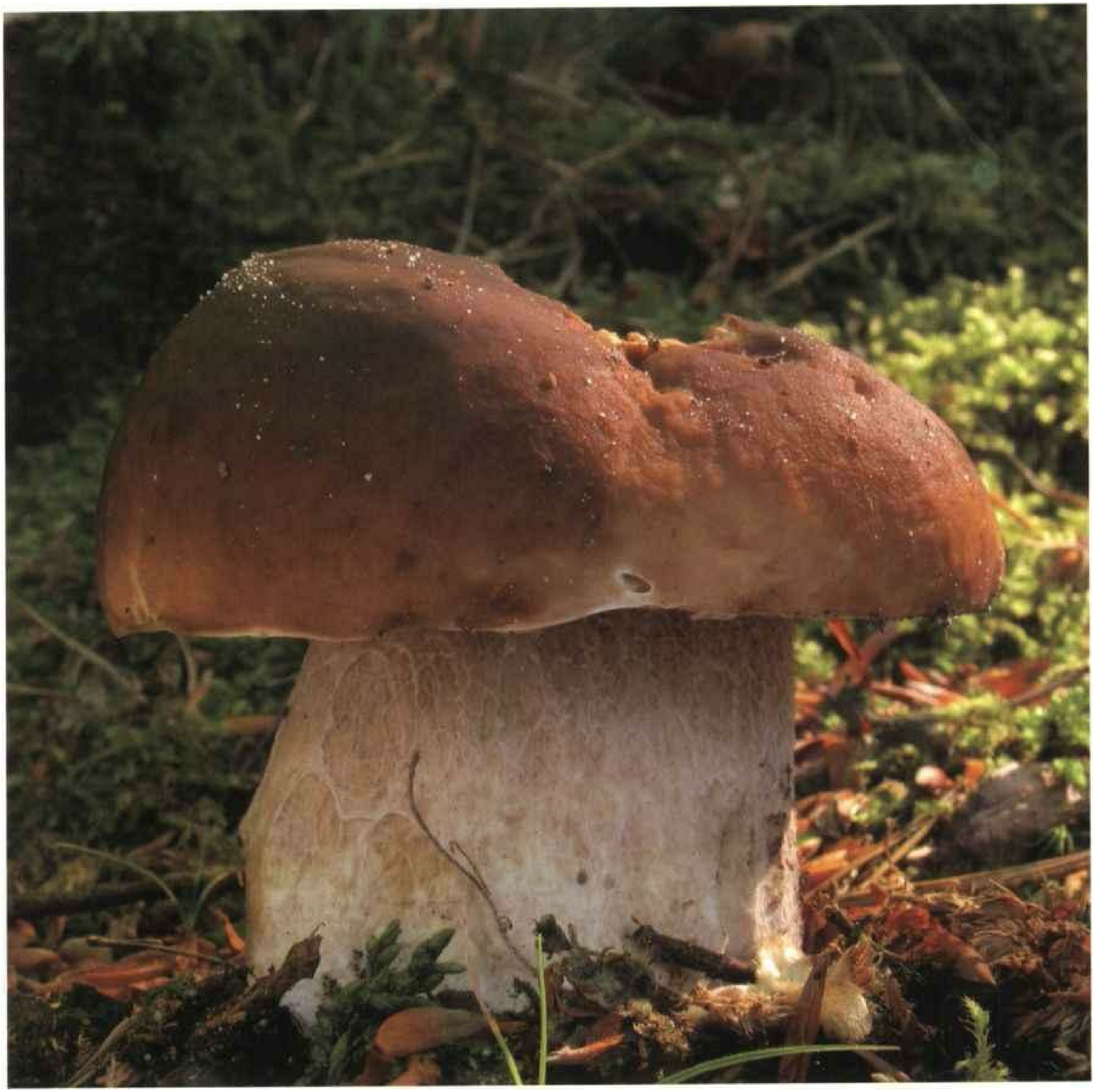

This is the key ID mark. The white net-pattern must be visible! All ceps, regardless of their time of emerging or stage of their development, have a fine white net-pattern at the top of their stalk.

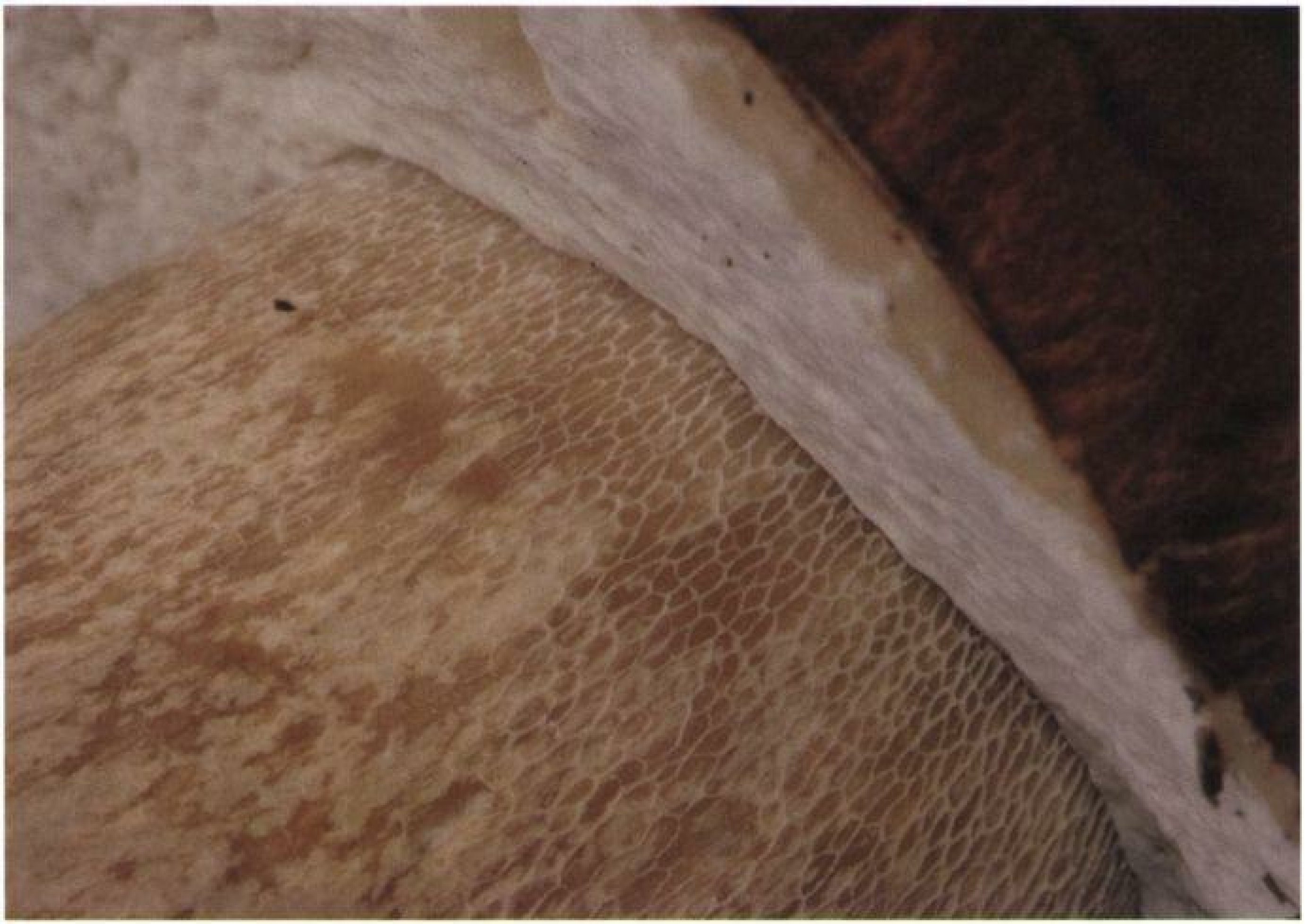

This is what a cep ideally looks like when cut. Not a single worm has even looked at this beauty. The tubes and the flesh do not change color when cut or touched.

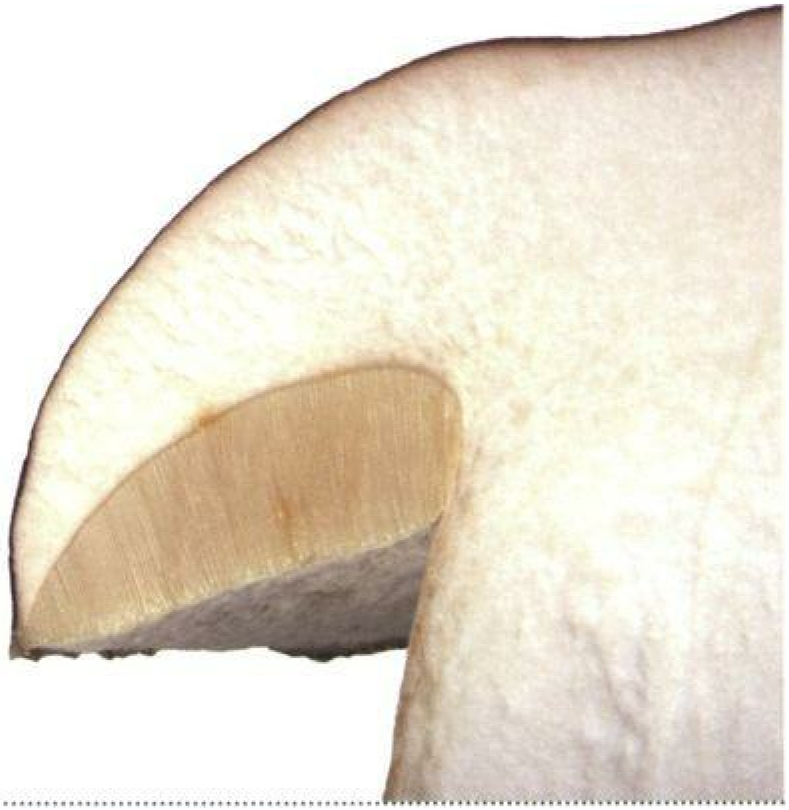

Cep tubes are off-white and firm when young. Later they turn yellow-olive and are less firm. In maturity the tubes are olive. Tubes of mature ceps are soft as a sponge and the tubes of old ones are soggy and often like those of the big fellow on page 38. This does not always mean that the inside of old ceps is rotten but more often than not, it is.

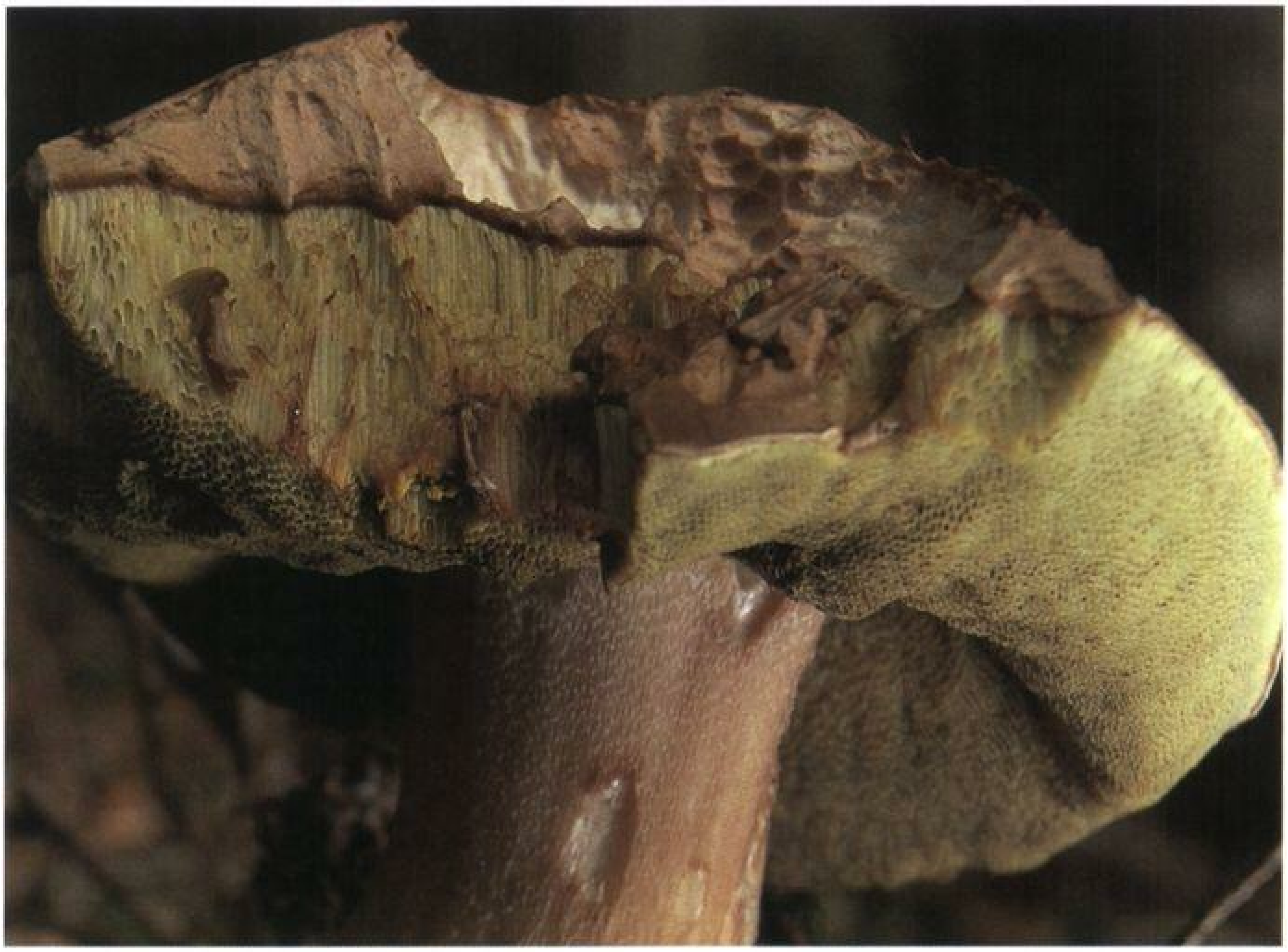

Cep stalks vary greatly in appearance. These specimens clearly show the white net-pattern all over the stalk.

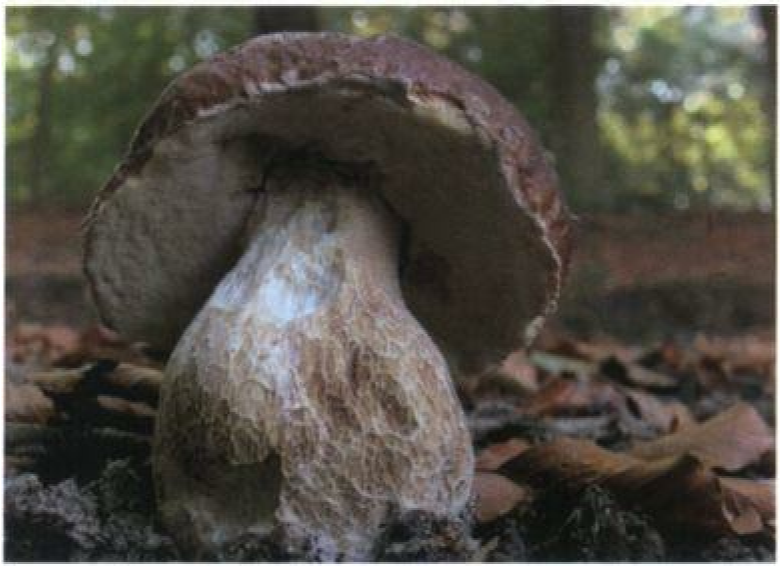

Nevertheless, it is that top section of the stalk just below the cap where it matters: there the white network must be clearly visible.

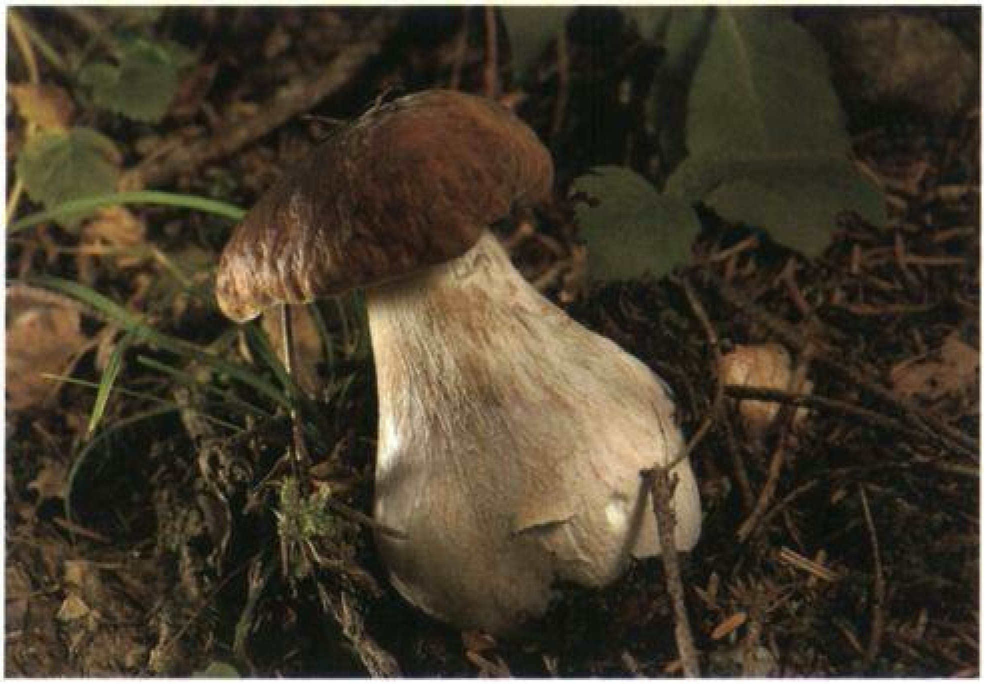

Positive ID Checklist

Cep or King Bolete

☐ Tubes, pores and flesh do not change color when cut or bruised
☐ Tubes are off-white, cream, yellow-olive or olive
☐ Pores do not show any pink tinge
☐ White network on top of stem

☐ Cap matches the color bar range (above)

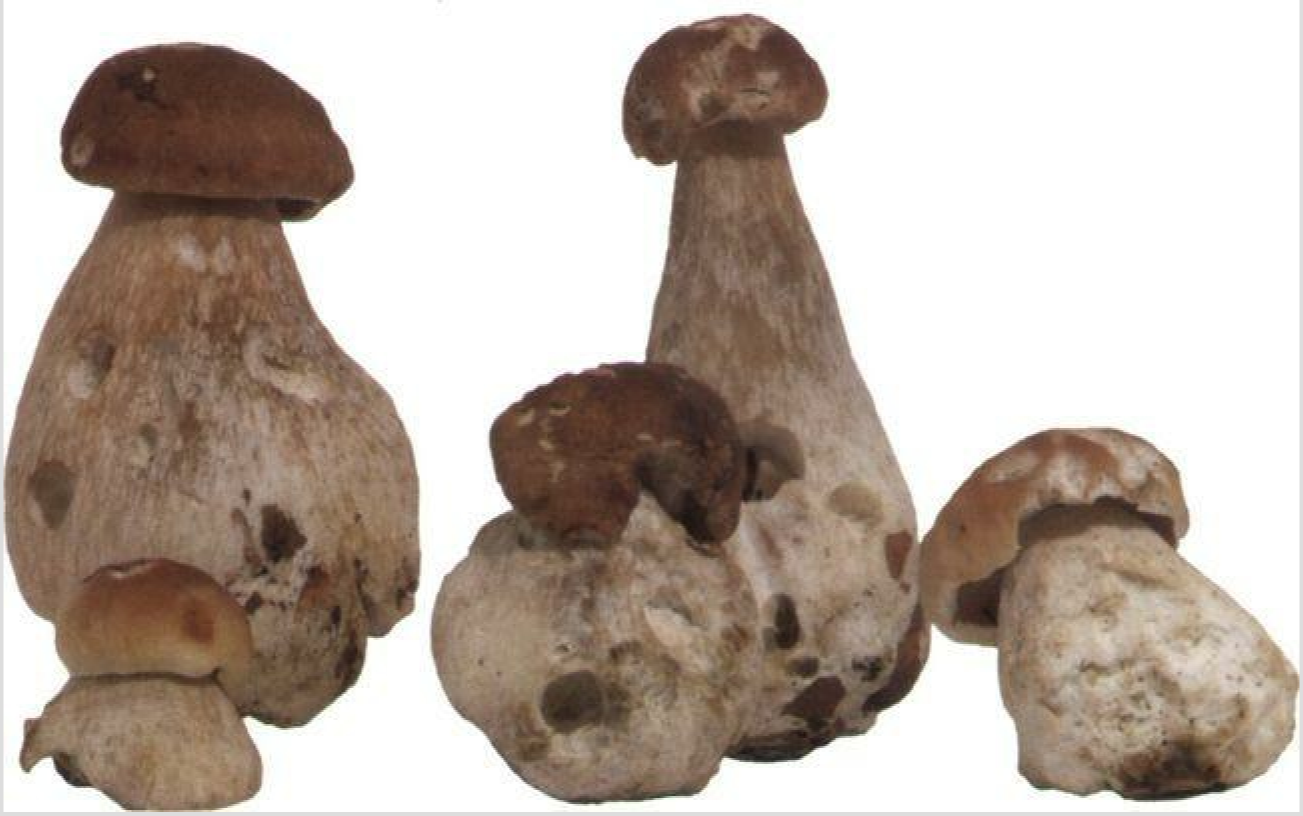

Avg. size across cap: 5 inches when mature

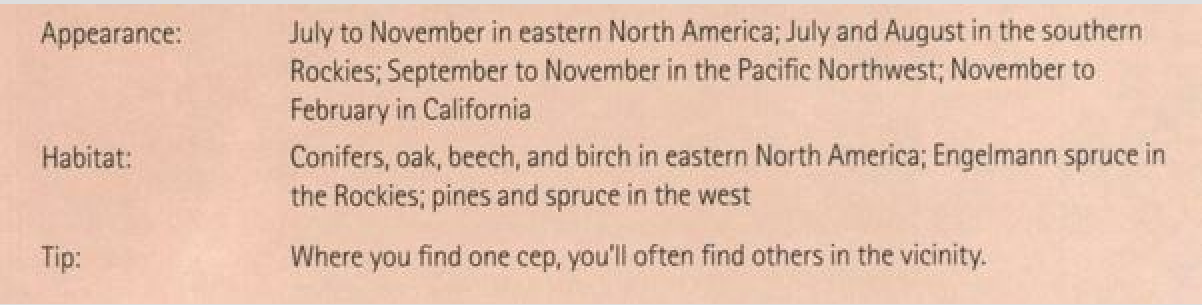

Appearance:	July to November in eastern North America; July and August in the southern Rockies; September to November in the Pacific Northwest; November to February in California
Habitat:	Conifers, oak, beech, and birch in eastern North America; Engelmann spruce in the Rockies; pines and spruce in the west
Tip:	Where you find one cep, you'll often find others in the vicinity.

Red Cracked Bolete

Xerocomus chrysenteron

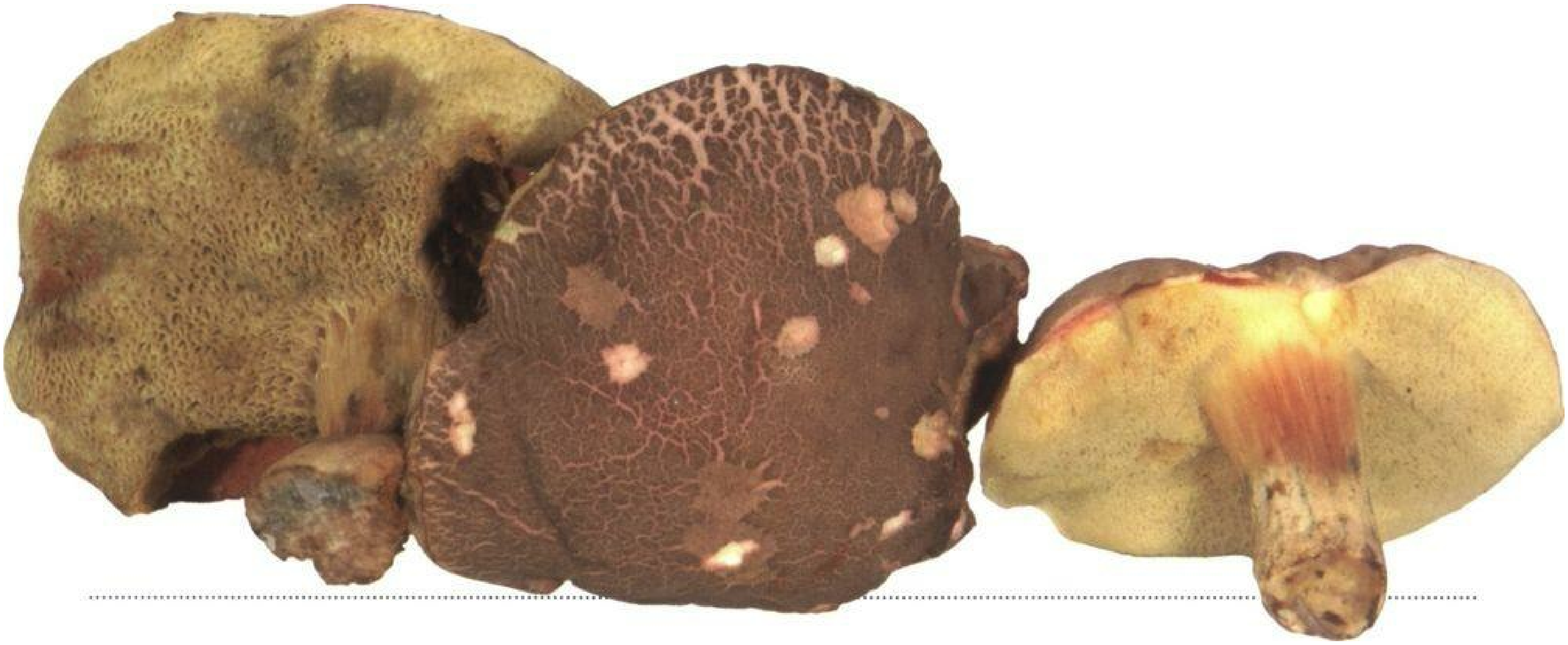

The red cracked bolete is probably the most common of all mushrooms with tubes. Its name describes its key identification feature perfectly.

As the mushroom matures, the cracks become more prominent. Some people love eating the red cracked bolete, others quite like it and there are a few who don’t take to it at all. The reason is its distinct fruity smell. In any case the red cracked bolete should only be picked as long as the pores are bright yellow and firm. The ideal size of the cap is about 1 inch across. Anything larger will make your dish all slimy.

Chrysenteron means ‘with golden-yellow flesh.’

The color of the cap ranges from a velvety dark brown to a light brown.

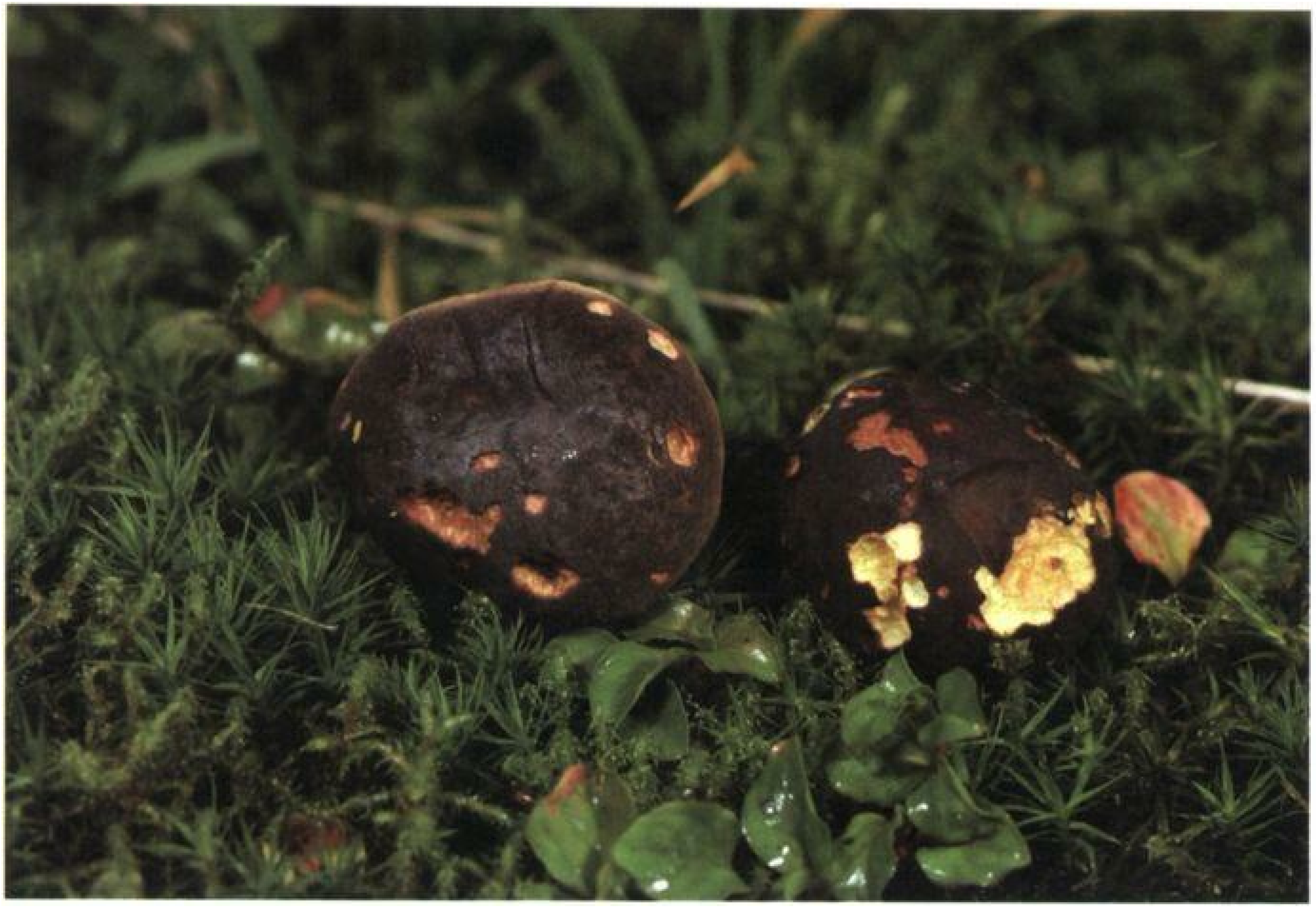

Stems vary in color from yellow, to yellow with a little red to yellow flushed with red. Note the bright yellow pores. The stem and the pores bruise blue. The intensity of the bruising varies. See picture opposite page, far left.

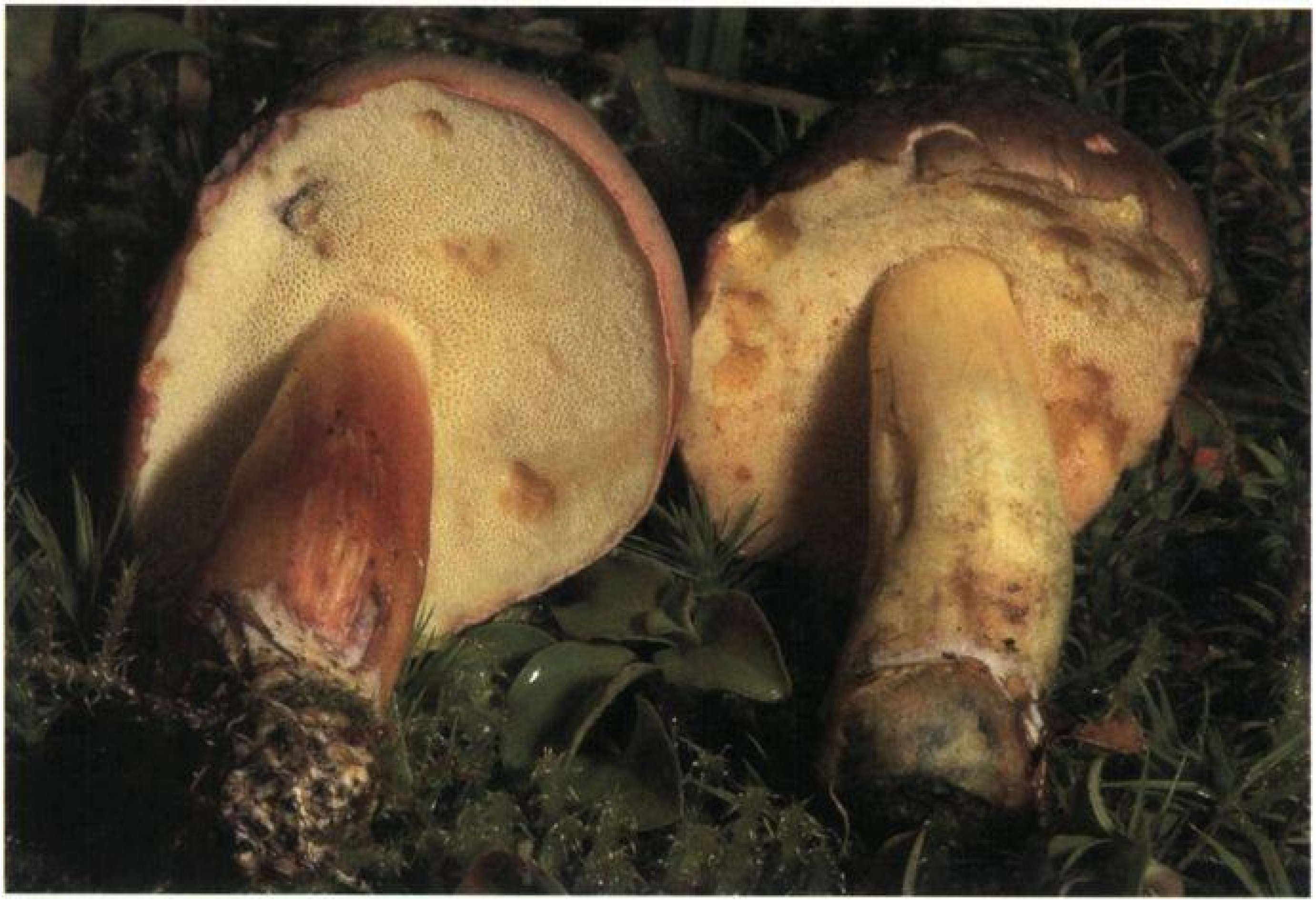

This specimen is too old for the kitchen. The pores have lost their brightness. The dull yellow signals that it’s no good anymore. When the pores are this color, the mushroom feels soft-to-soggy.

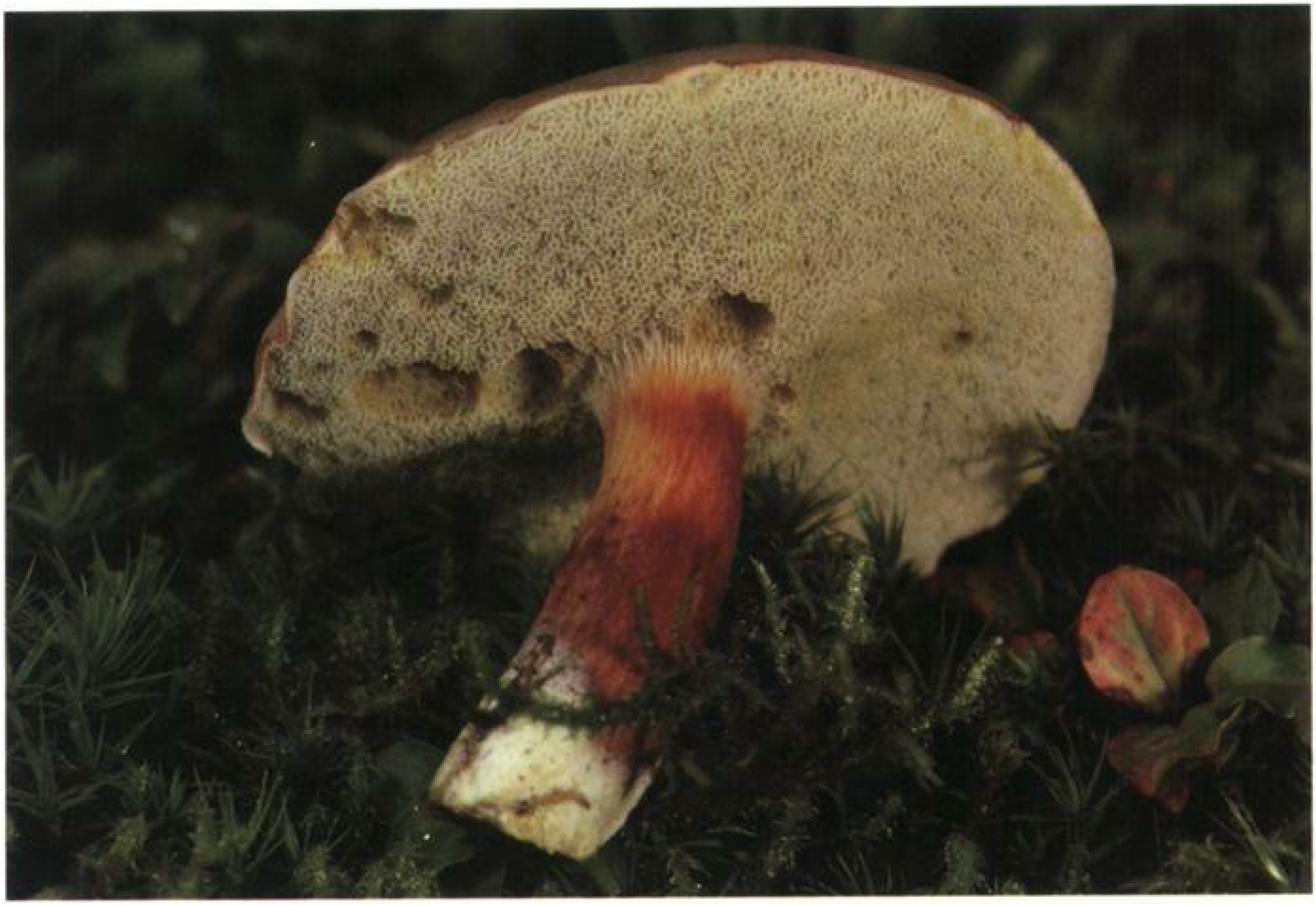

This is the perfect specimen. It has bright, firm golden-yellow pores and a nice firm cap about 1 inch across.

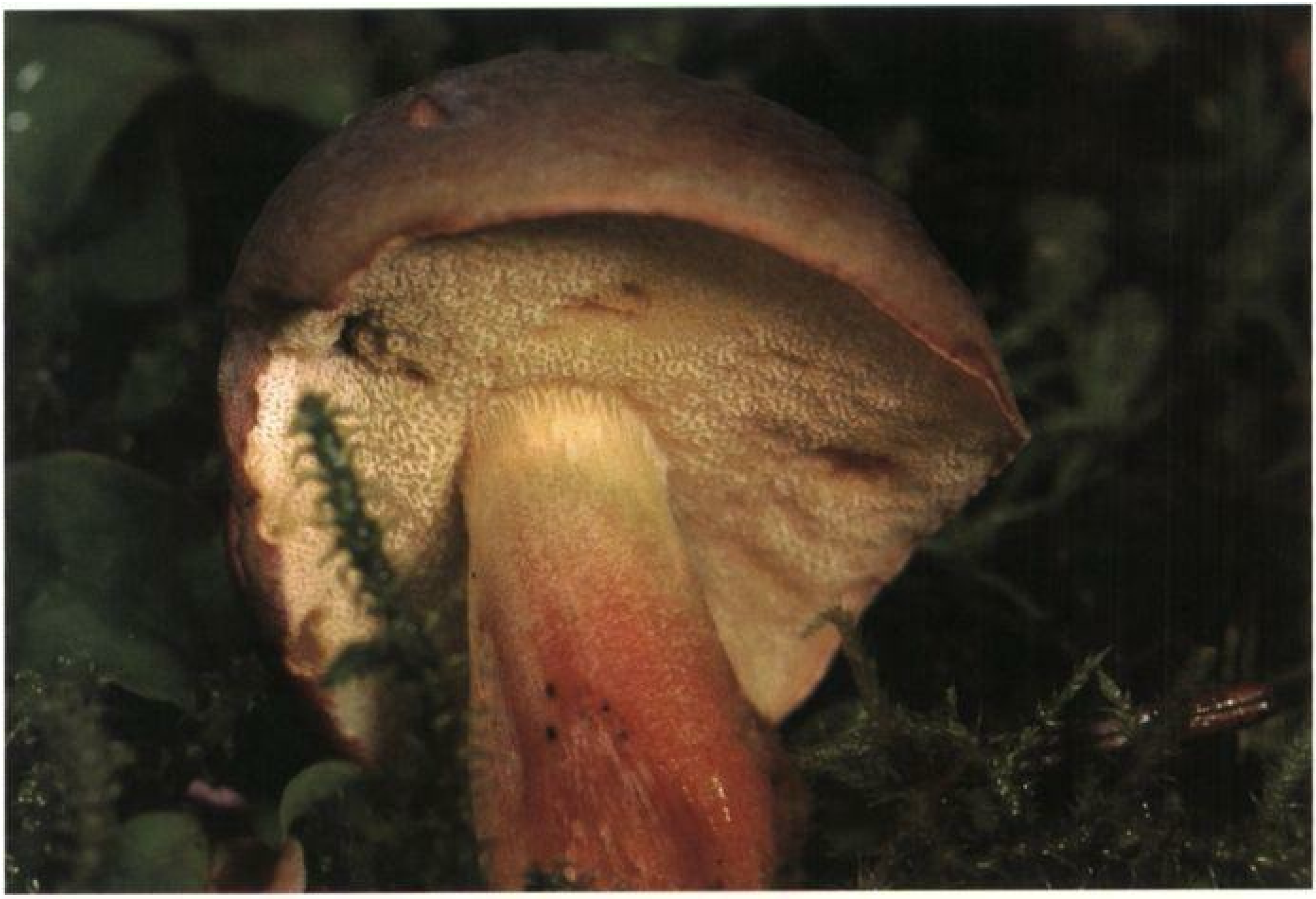

The caps are invariably cracked or lightly damaged. The typical identification mark of the red cracked bolete is the red showing through the cracks or the eaten-away patches.

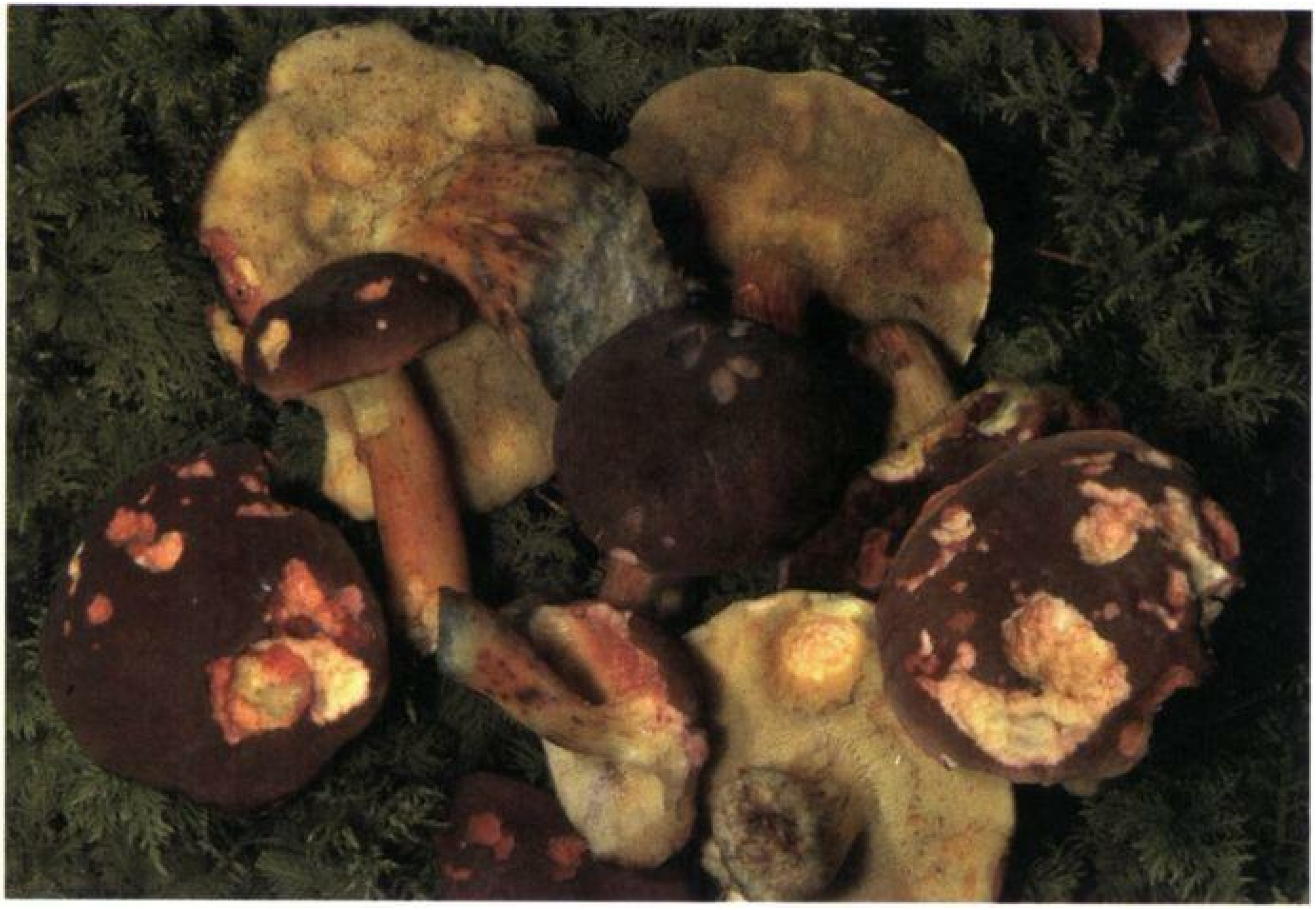

When cut, the white-to-bright-yellow flesh turns blue. This can be very light blue and confined to certain patches.

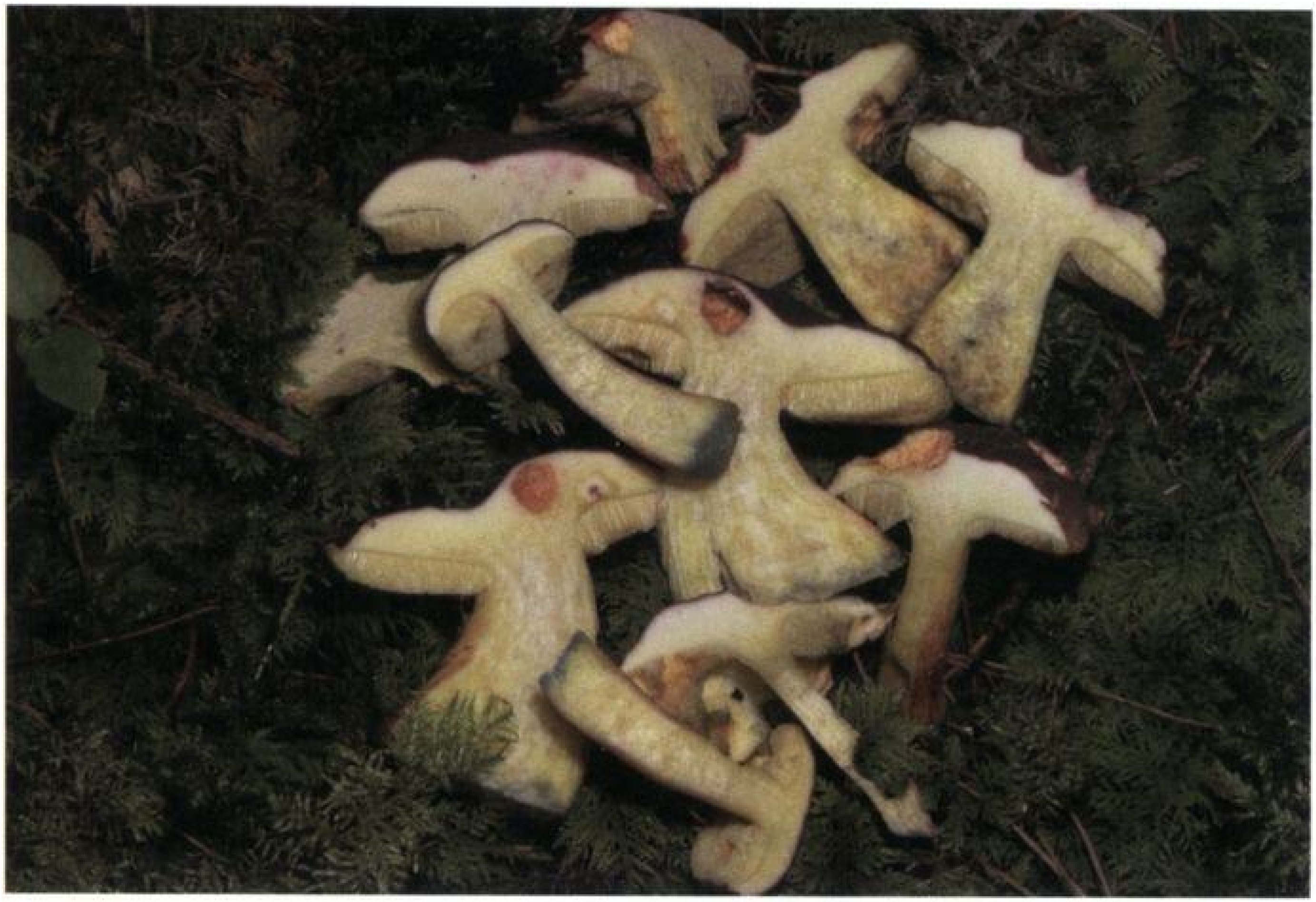

Positive ID Checklist

Red Cracked Bolete

- ☐ Yellow tubes
- ☐ Yellow pores bruise blue (varying intensity)
- ☐ Flesh bruises blue (varying intensity)
- ☐ Cracks and damaged patches in cap show a distinct red tinge
- ☐ Cap matches the color bar (above)

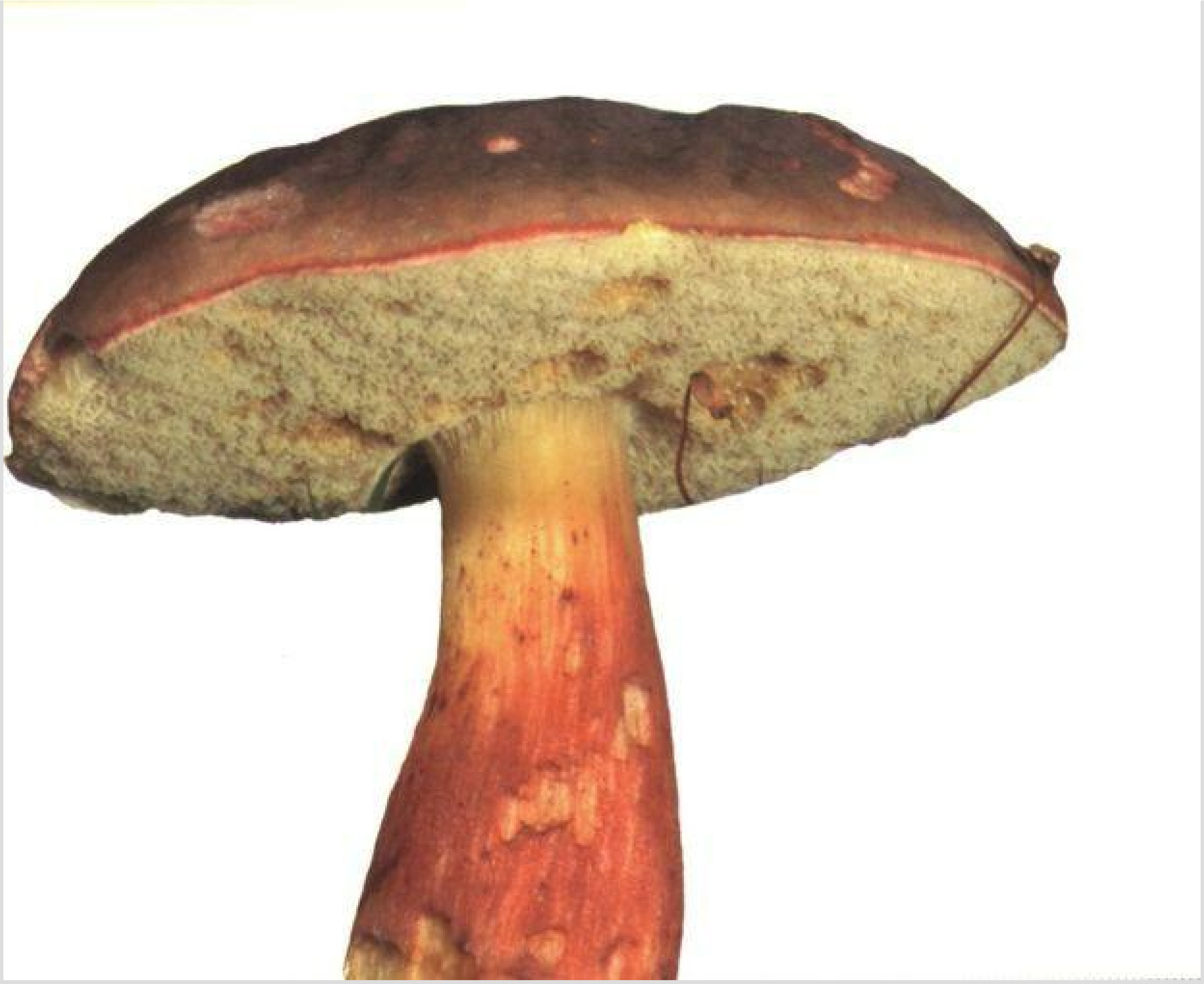

Avg. size across cap: 2 inches when mature

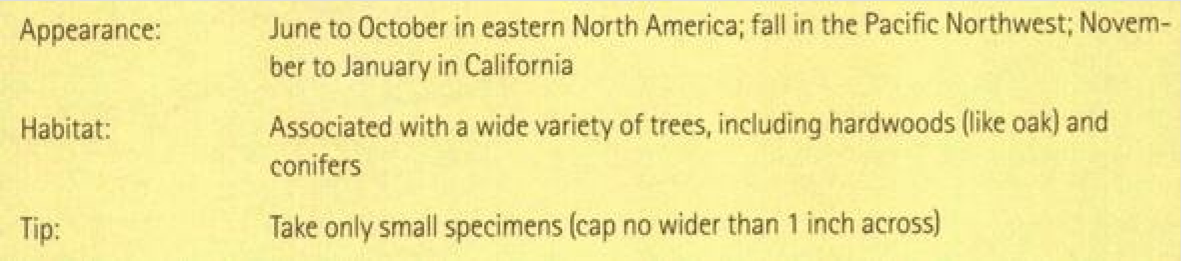

Appearance:	June to October in eastern North America; fall in the Pacific Northwest; November to January in California
Habitat:	Associated with a wide variety of trees, including hardwoods (like oak) and conifers
Tip:	Take only small specimens (cap no wider than 1 inch across)

Hen of the Woods

Grifola frondosa

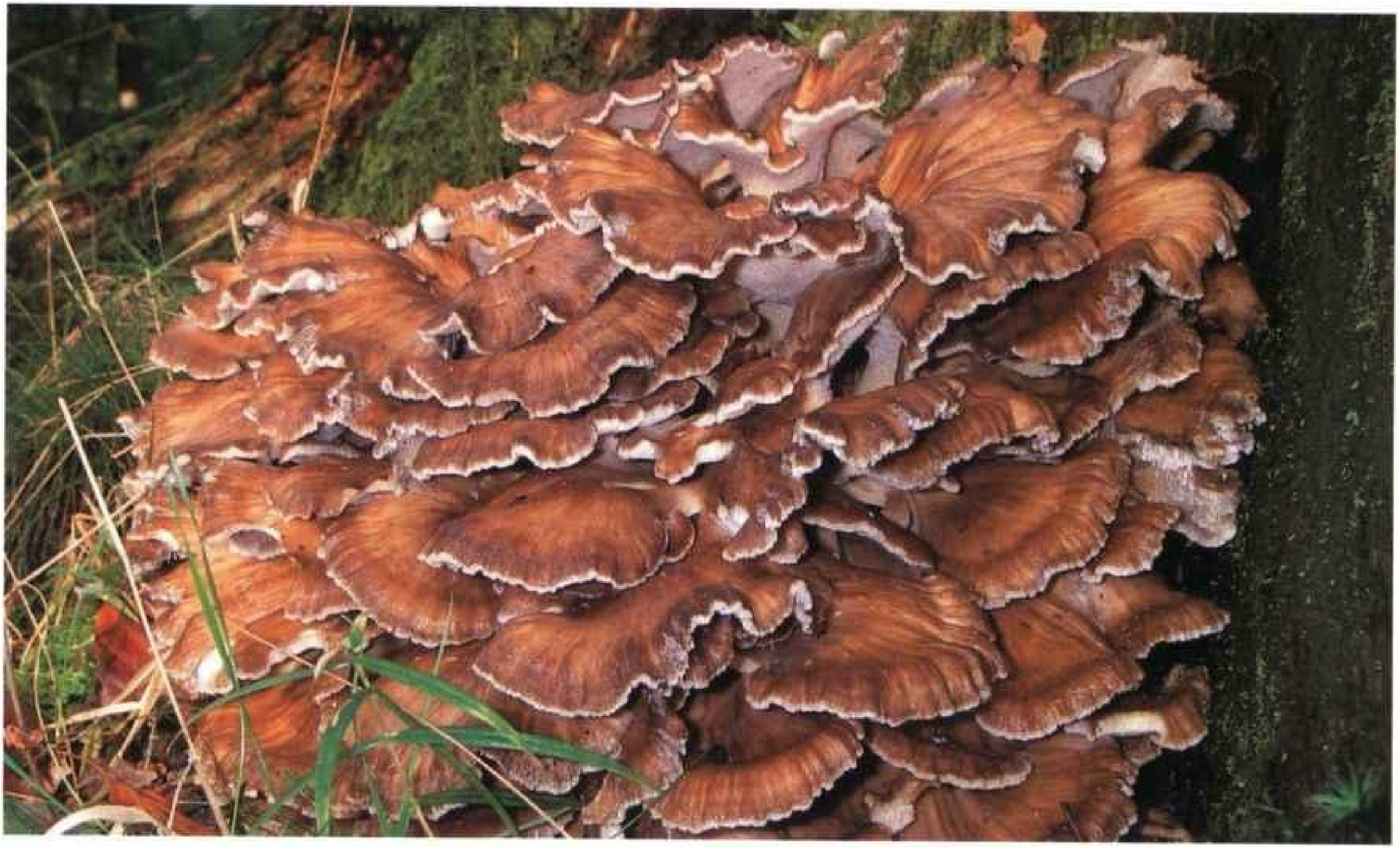

For thousands of years, the hen of the woods mushroom has been prized for its medicinal and culinary value in China and Japan. Maitake, the
Japanese name for the hen of the woods, means "dancing mushroom" and one of the many interpretations of this name is that those who found a hen of the woods started to dance with joy because it was such a highly-prized mushroom.

It can be cultivated commercially. However, the wild hen of the woods mushroom you pick is infinitely superior because, firstly, you found it yourself; and secondly because it is as different from the farmed version as is a wild salmon to a farmed one. Finding a hen of the woods is indeed a reason to dance: a culinary event of the first order—and it's good for you too. There aren't that many things in life of which that can be said.

The hen of the woods is a cluster of fan-shaped overlapping caps.

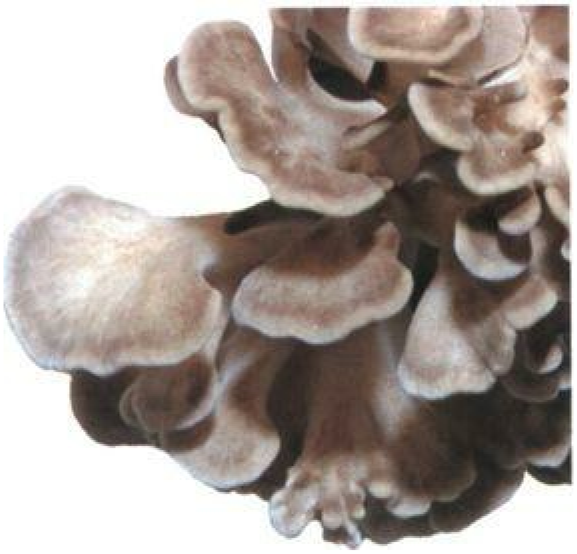

On the underside of the cap (approx. 2 inches across) the tubes are at this stage visible to the naked eye. At this point the hen of the woods reaches gourmet status. Start picking now.

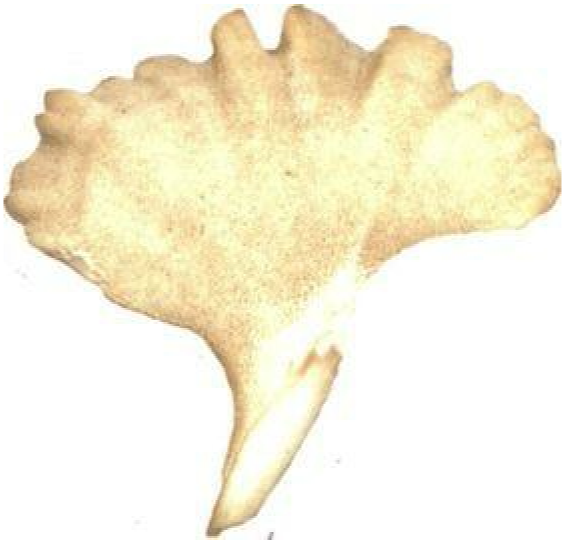

Actual size 2 inches

The fan-shaped caps with radial furrows are typical of the hen of the woods.

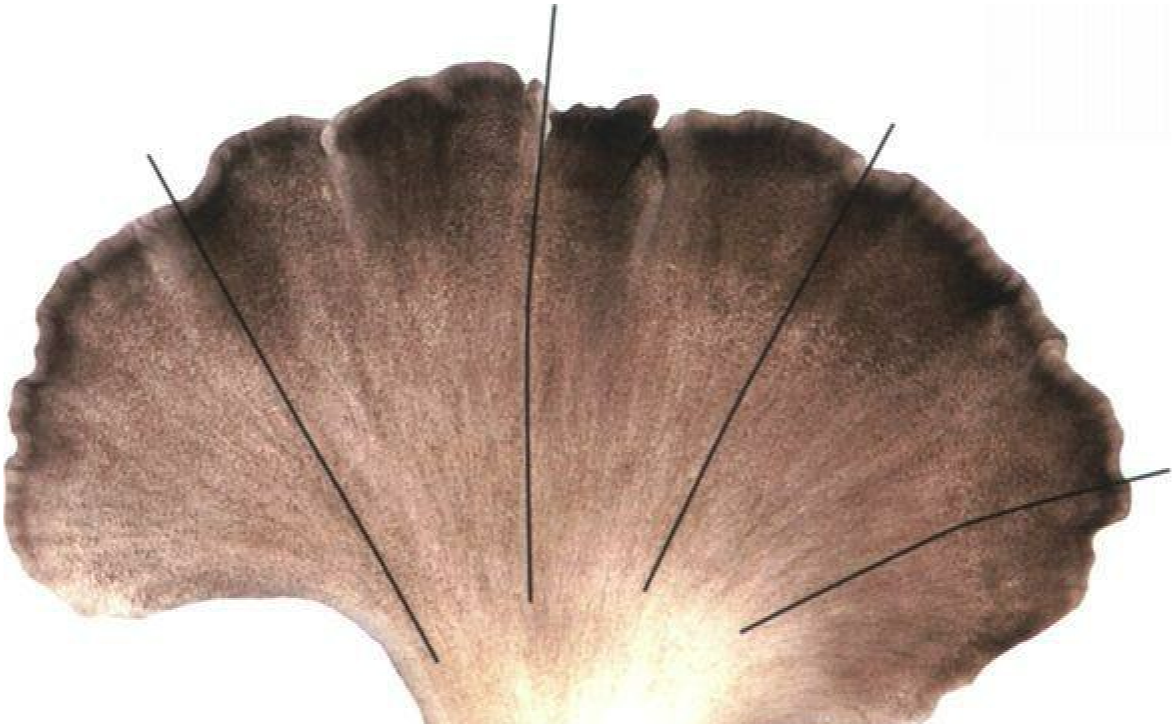

The tubes are clearly visible at this stage.

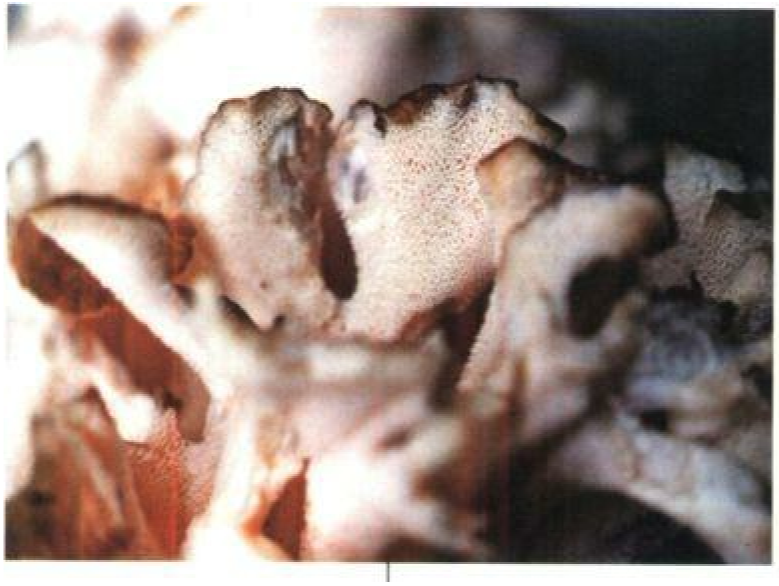

The cross-section shows a cauliflower-like structure with one central stem. The widest part of this particular specimen is approx. 12 inches.

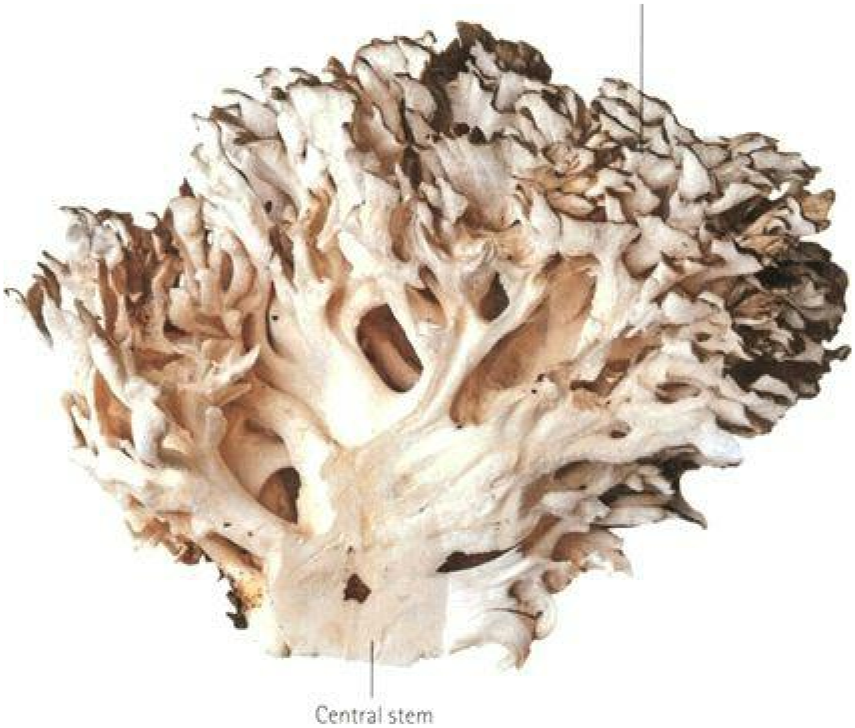

On older specimens the tubes are larger in diameter but remain whitish. As the tubes grow larger they almost look like spines-but tubes they are.

The main colors of the hen of the woods range from off-white, gray-beige, gray-brown, gray-black, brown-black to light brown-black. The hen of the woods can grow up to 30 inches across.

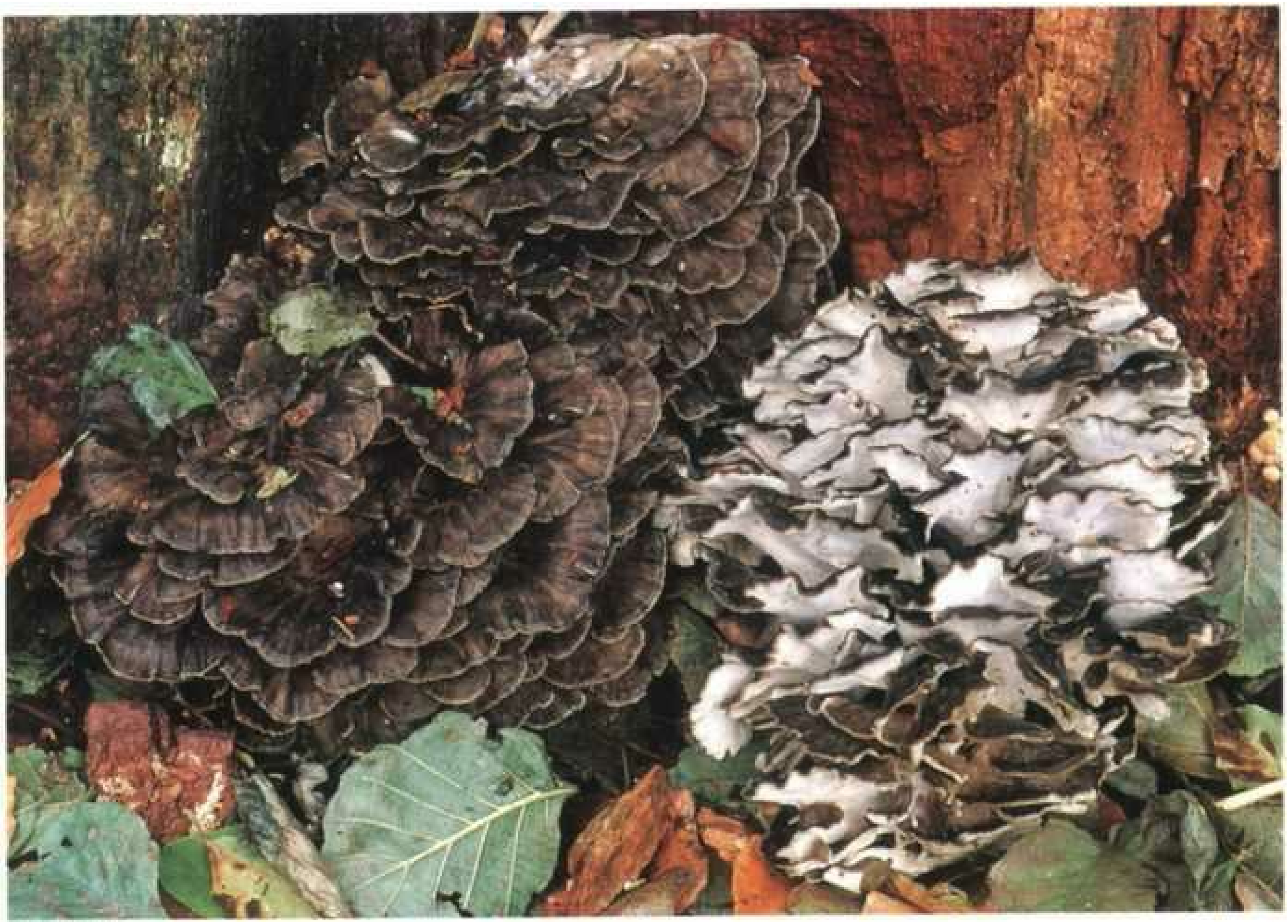

The culinary value of the hen of the woods depends on its age. Size is not always an indicator of age. Some specimens grow very big, very fast. Despite their size these are still young specimens with

white, firm, fibrous flesh with a pleasant smell.
Older specimens have a distinct unpleasant smell and their caps get more and more flabby and the edges crumble.

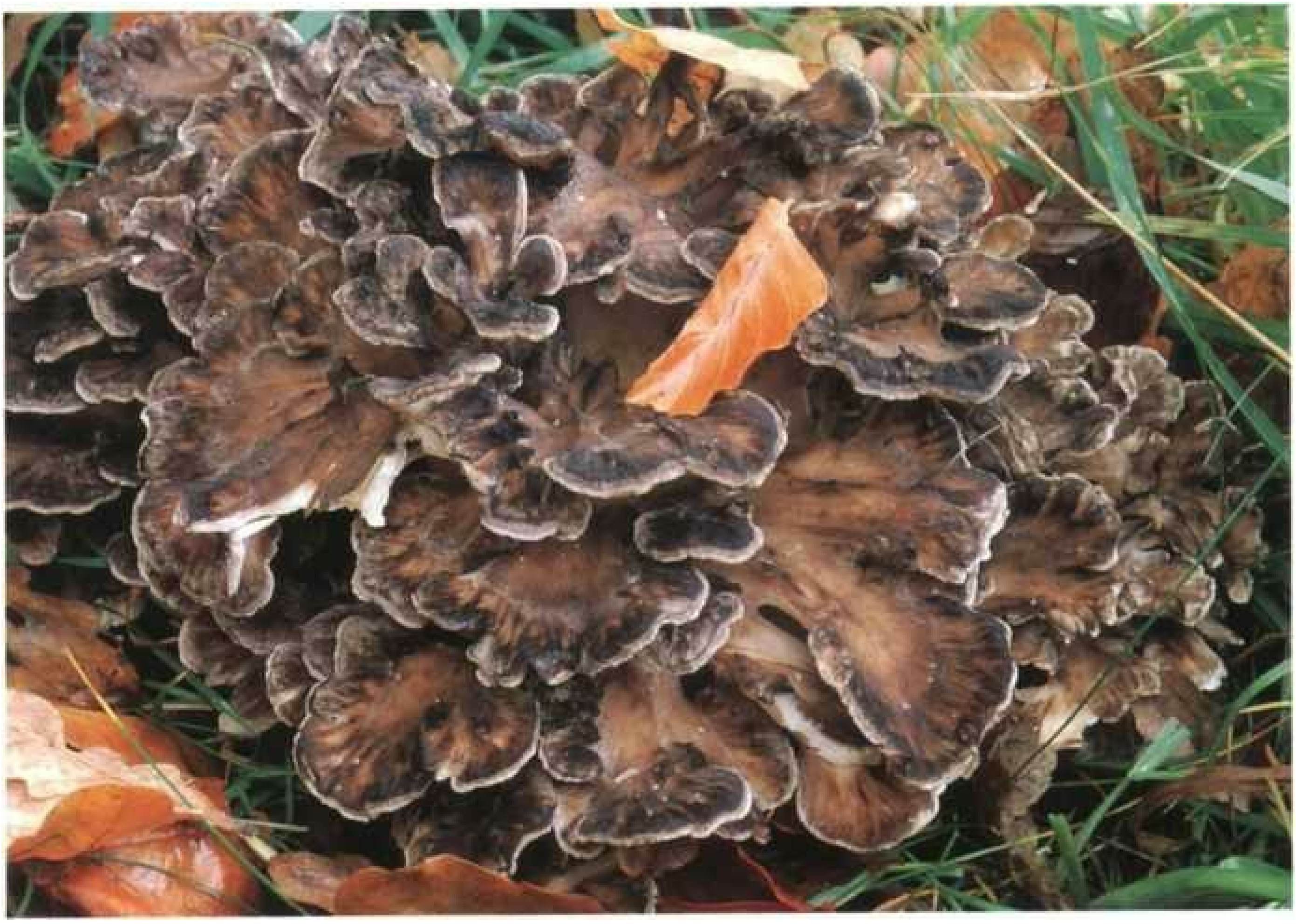

Positive ID Checklist

Hen of the Woods

- ☐ Central stem
- ☐ Fan-shaped caps
- ☐ Caps with dark radial streaks
- ☐ White flesh does not change color when cut
- ☐ Cauliflower-structured cross-section
- ☐ White tubes
- ☐ Tubes do not change color when bruised
- ☐ Main cap color matches color bar (above)

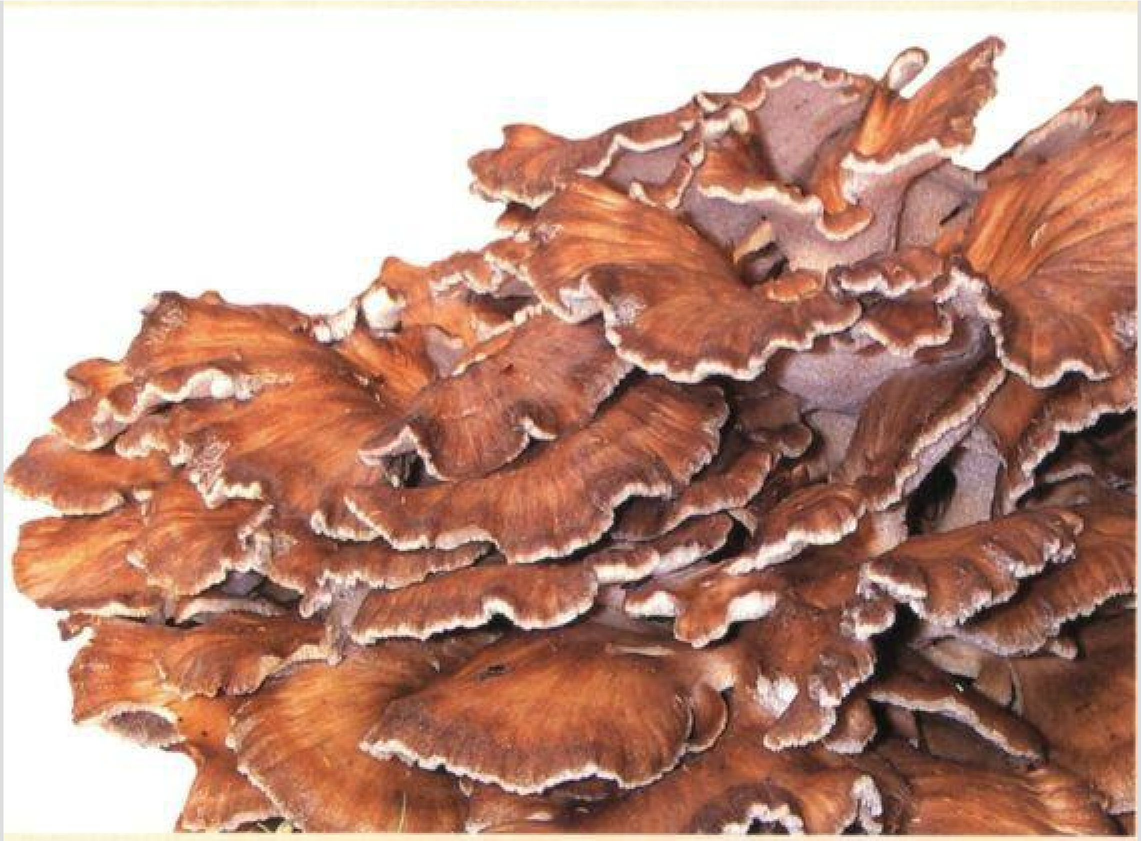

Avg. size across cap: 12 inches, but can grow up to 30 inches

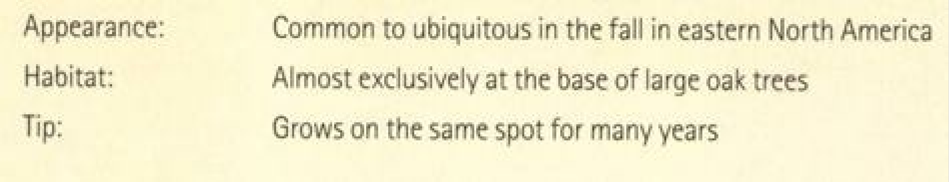

Appearance:	Common to ubiquitous in the fall in eastern North America
Habitat:	Almost exclusively at the base of large oak trees
Tip:	Grows on the same spot for many years

Larch Bolete

Suillus grevillei

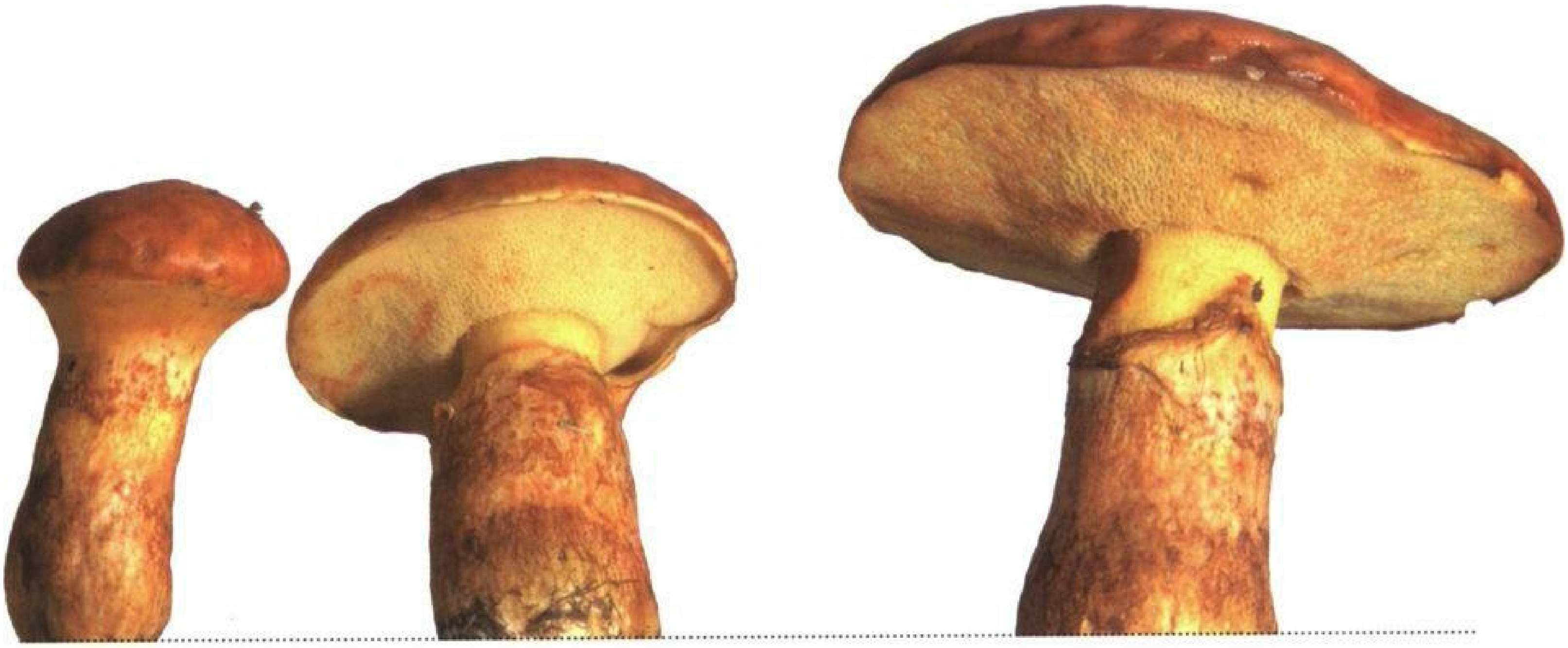

It is as if pure gold has grown out of the ground-and where there is one nugget, there are others. It’s very rare to find a single larch bolete.
More often than not larch boletes form a fairy ring. Always found near larch trees, this beautiful and delicious mushroom is covered with a yellow veil when young (above left). As the maturing mushroom grows, the veil breaks, leaving a transient ring on the stalk (above, center, and right). Always peel the cap of the larch bolete on the spot because otherwise it will make your basket and later your recipe all slimy. In wet conditions, the cap of the larch bolete is always slimy.

Grevillei refers to the Scottish mycologist R. K. Greville.

The veil still covers the cap in a ‘baby’ larch bolete.

The tubes must be clearly visible when you view a cross-section.

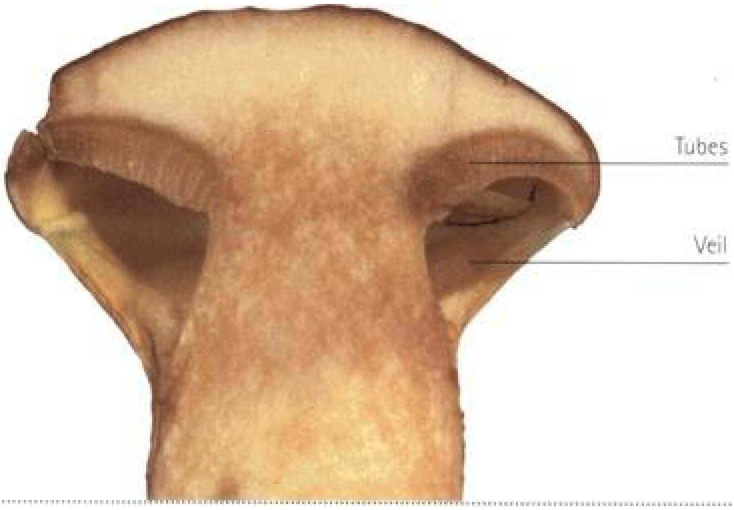

As the mushroom grows, the veil breaks. This picture (right) shows a larch bolete shortly after the breaking of the veil. The ring is transient (i.e., it often drops off) but even in the mature larch bolete you can always see a mark where the ring has been.

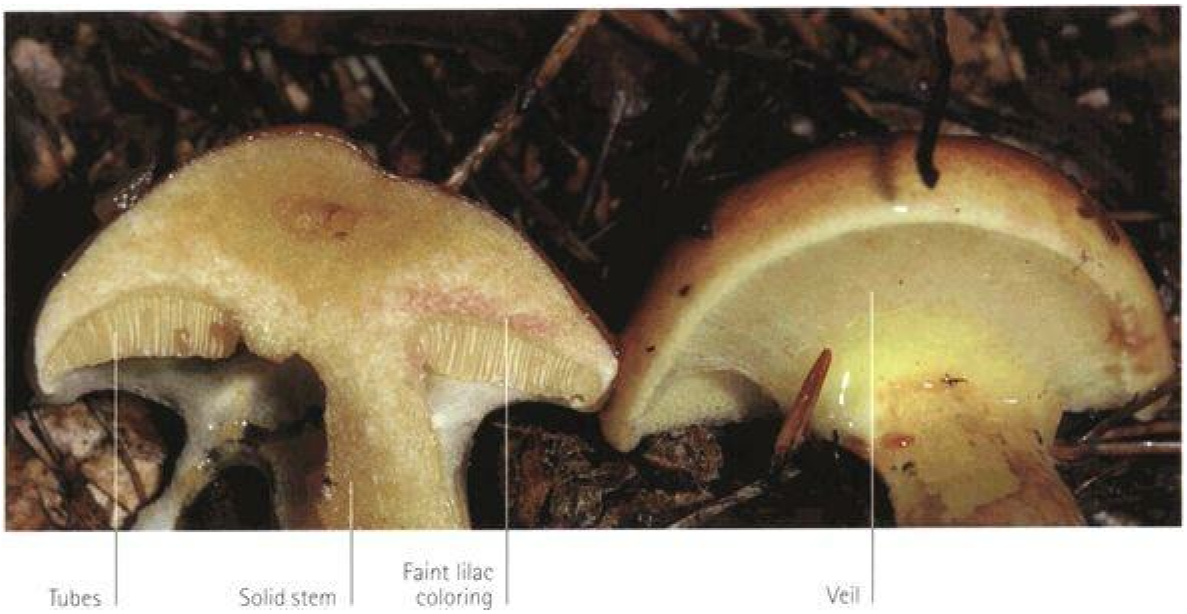
Tubes
Solid stem
Faint lilac coloring
Veil

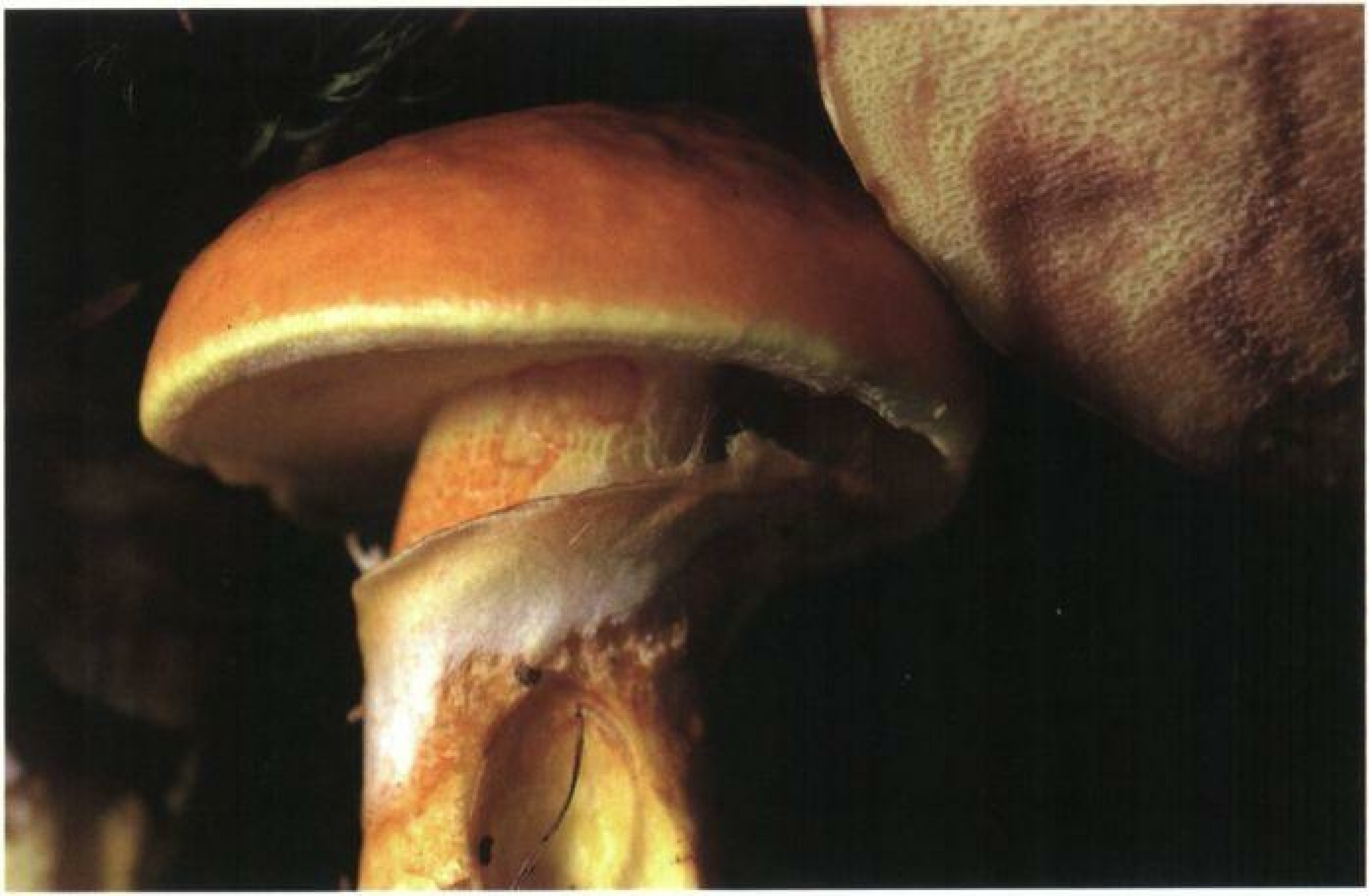

Above the ring marks: net pattern (can be very faint). Below the ring marks: mature specimens show distinct reddish-rusty streaks on the yellow stem.

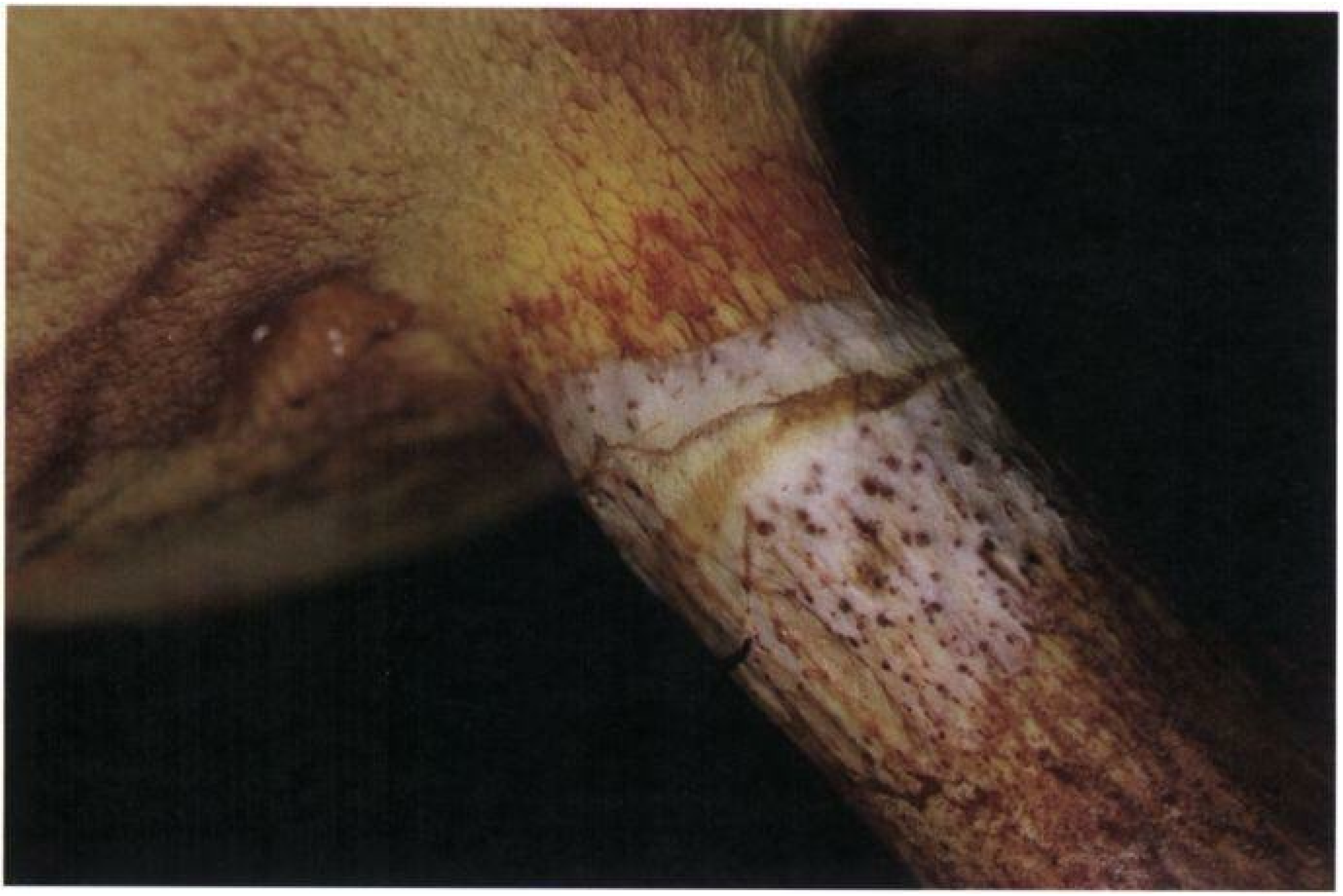

If a larch tree is nearby and you see several gold nuggets,they are bound to be larch boletes. The beautifully golden (and slimy when wet) caps are the key sign. Positive identification however requires ticking off all of the identification marks.

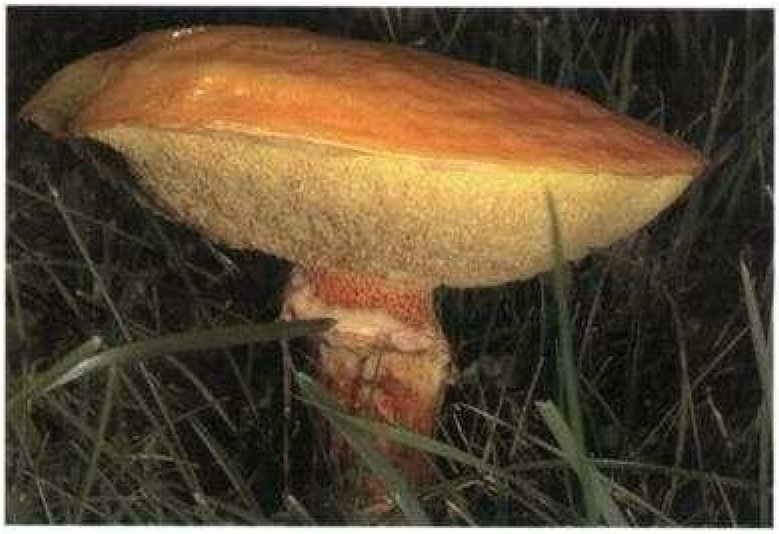

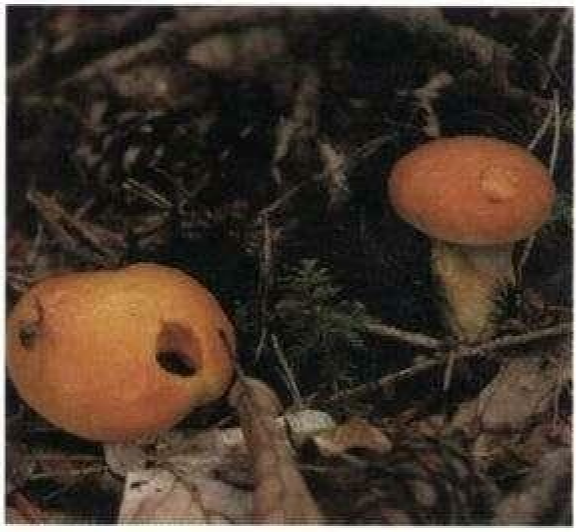

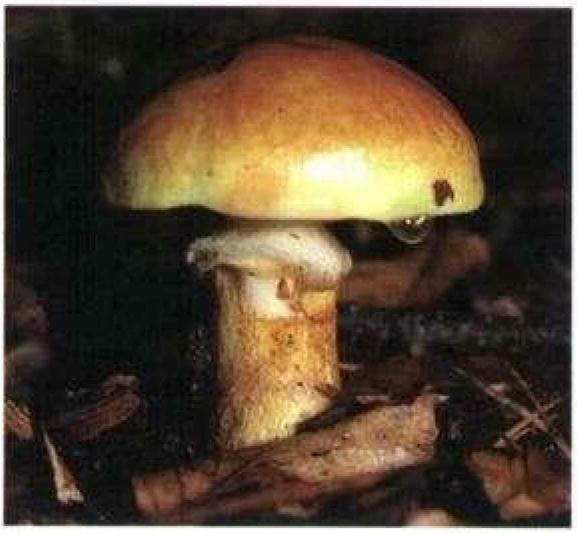

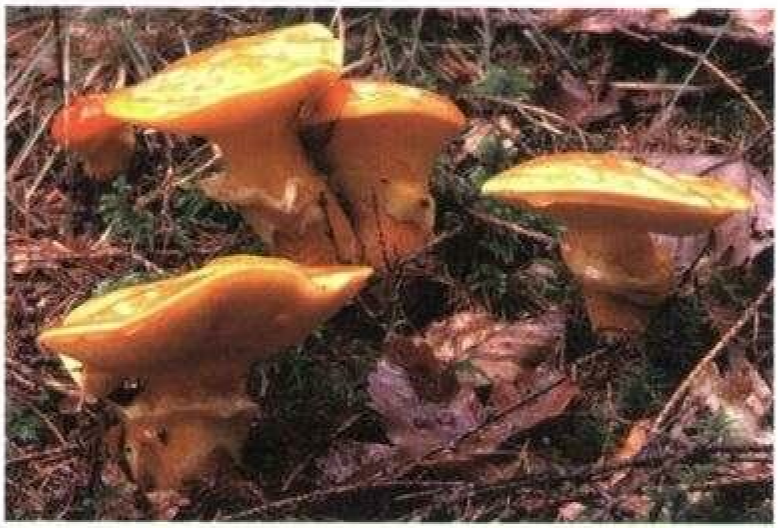

Positive ID Checklist

Larch Bolete

☐ Yellow tubes
☐ Solid stem
☐ Bright orange-yellow or golden-yellow sticky cap (slimy when wet).
☐ Flesh in cap flushes lilac (faint)
☐ Found near larch trees
☐ Other specimens found in the vicinity
☐ Young: tubes covered with a veil
☐ Intermediate stage: ring visible, faint net-pattern above ring zone
☐ Mature: ring zone still visible, faint network above ring zone
☐ Cap color matches cap color bar (above)

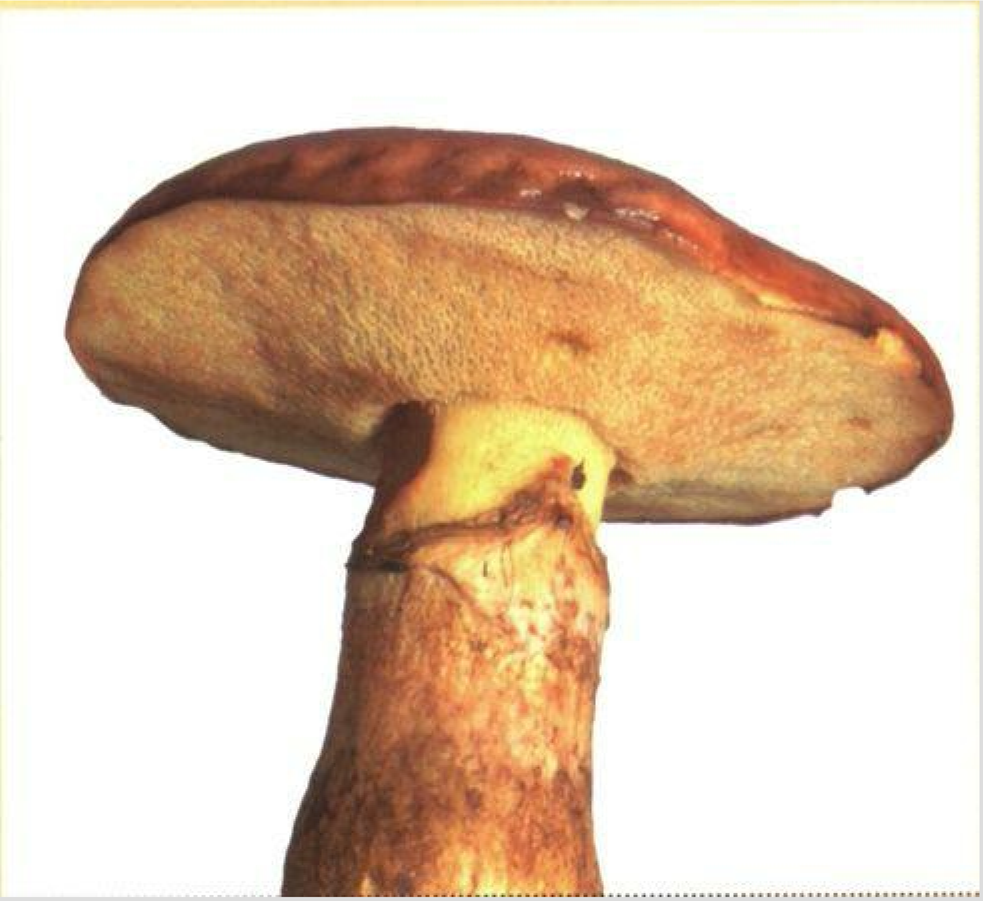

Avg. size across cap: 2.5 inches when mature

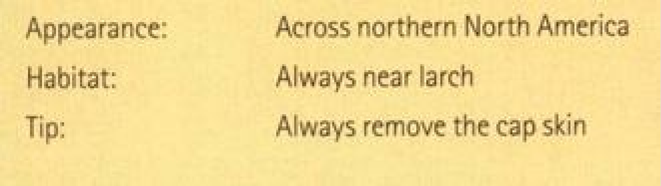

Appearance:	Across northern North America
Habitat:	Always near larch
Tip:	Always remove the cap skin

Bay Bolete

Xerocomus badius

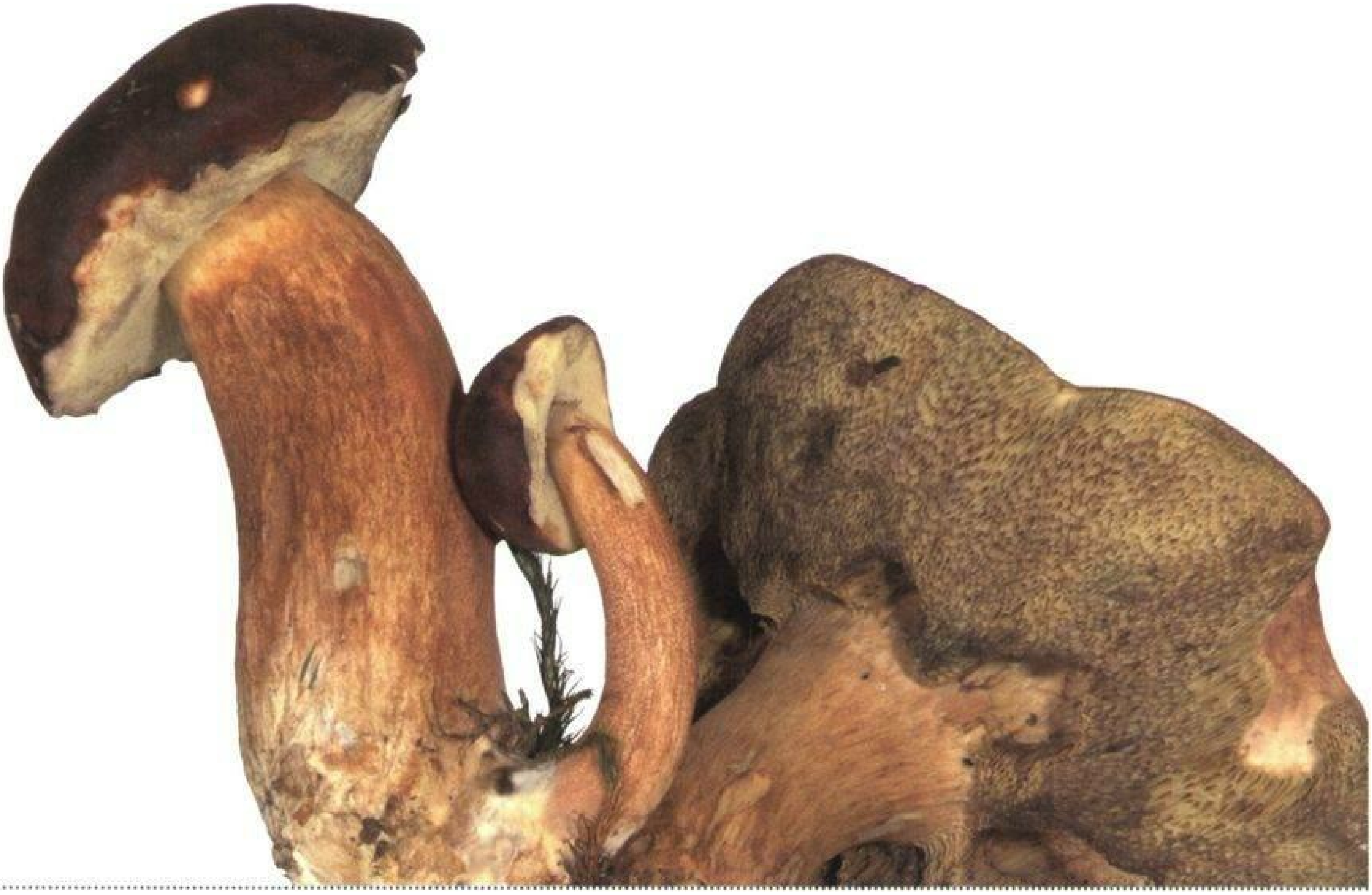

The bay bolete is an excellent mushroom and by good luck, it is also very common. The main bay bolete months are September and October but it can be found as early as June. If you find one, there are bound to be others nearby.

Badius means ‘beautifully brown.’

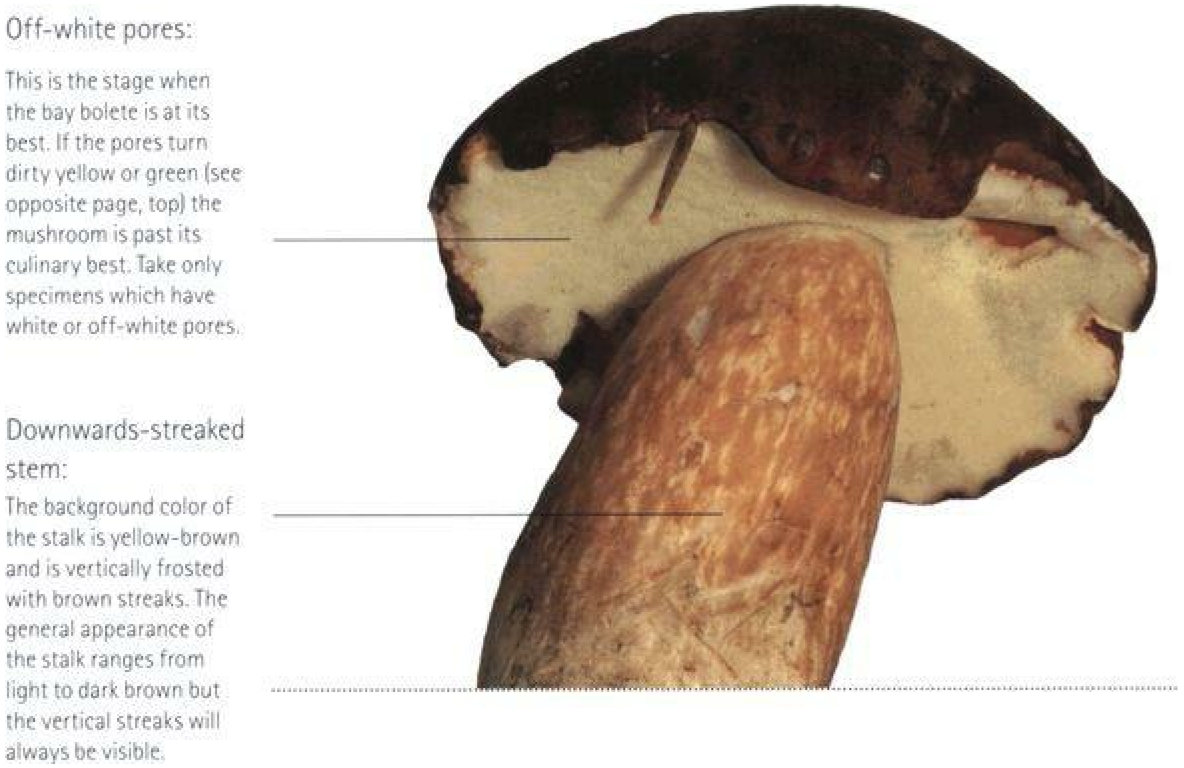

Off-white pores:

This is the stage when the bay bolete is at its best. If the pores turn dirty yellow or green (see opposite page, top) the mushroom is past its culinary best. Take only specimens which have white or off-white pores.

Downwards-streaked stem:

The background color of the stalk is yellow-brown and is vertically frosted with brown streaks. The general appearance of the stalk ranges from light to dark brown but the vertical streaks will always be visible.

The off-white pores bruise green-blue.

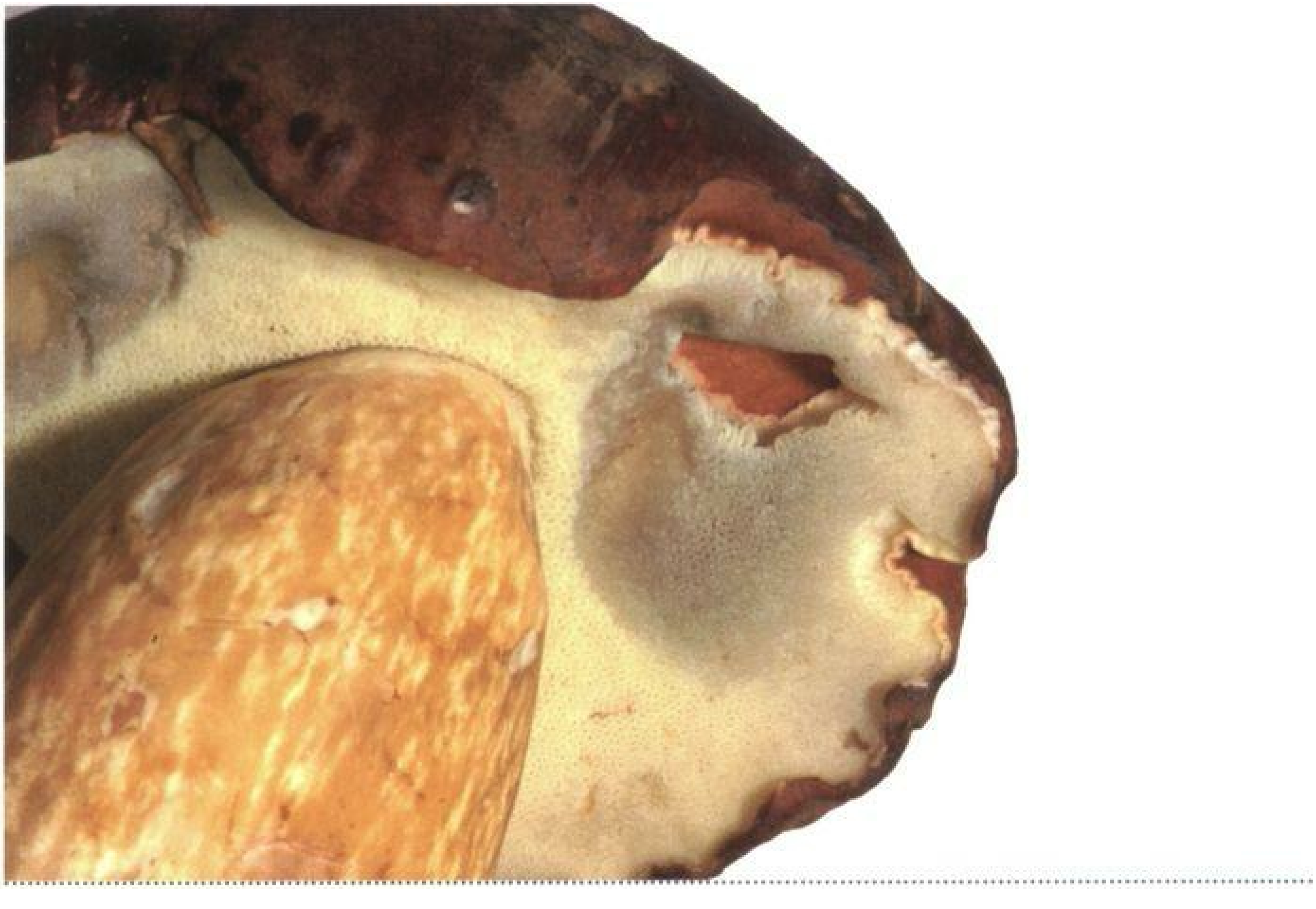

The change in color of the pores to green-blue varies in intensity. Here it's quite extreme.

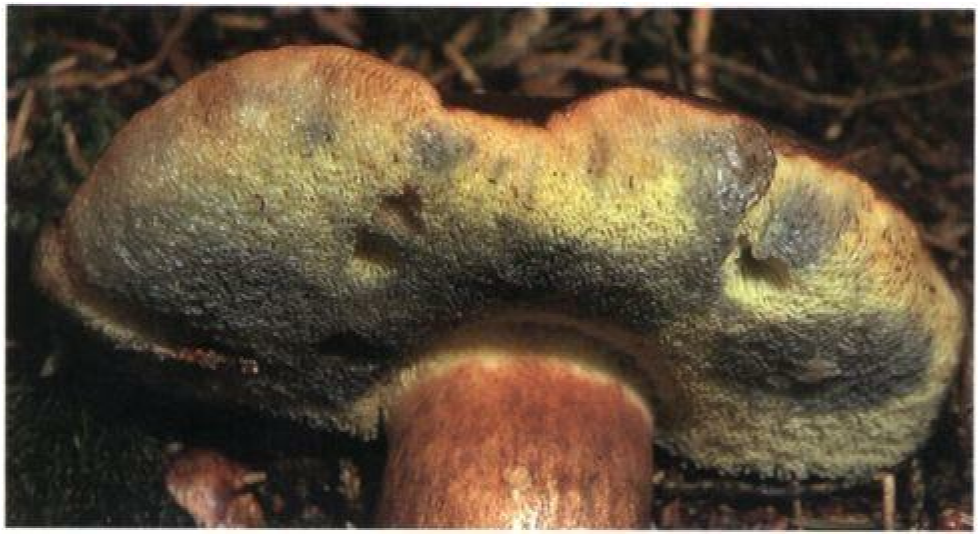

The young bay bolete has pale yellow tubes which then turn to yellow, olive-yellow and dirty olive-yellow as the mushroom matures. Whatever the stage, the tubes and the flesh bruise blue.

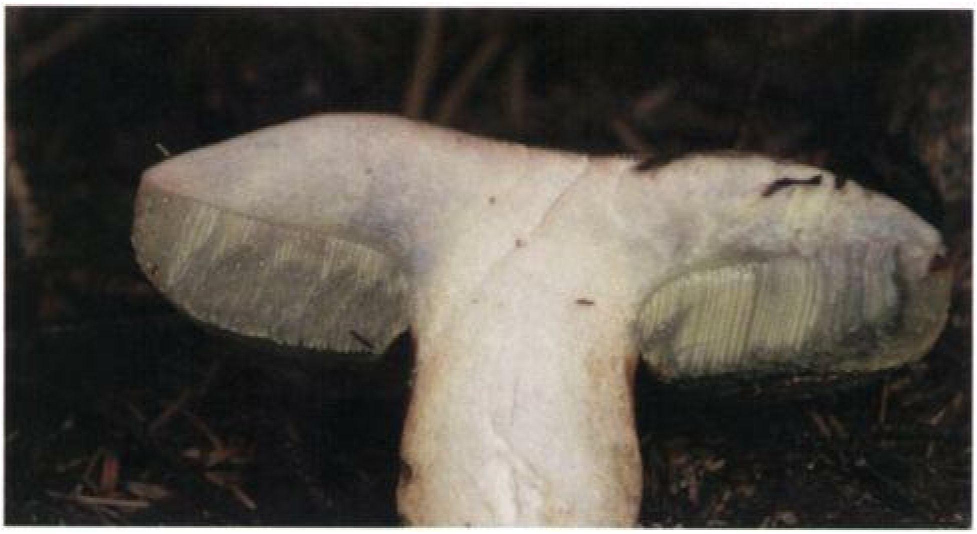

The intensity of the coloring varies but pores, tubes and flesh will always bruise green-blue.

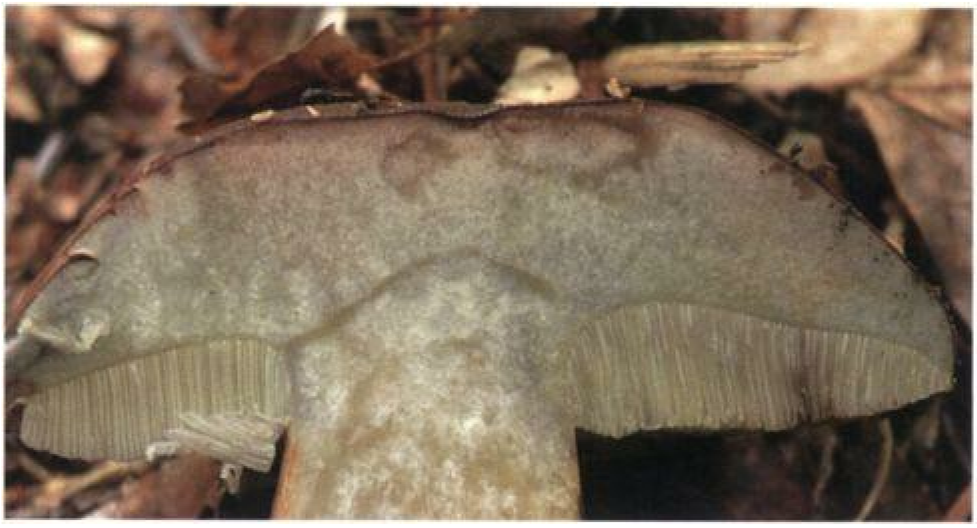

This is what the ideal bay bolete looks like: off-white pores and tubes, white to white-yellow flesh, without a trace of a worm.

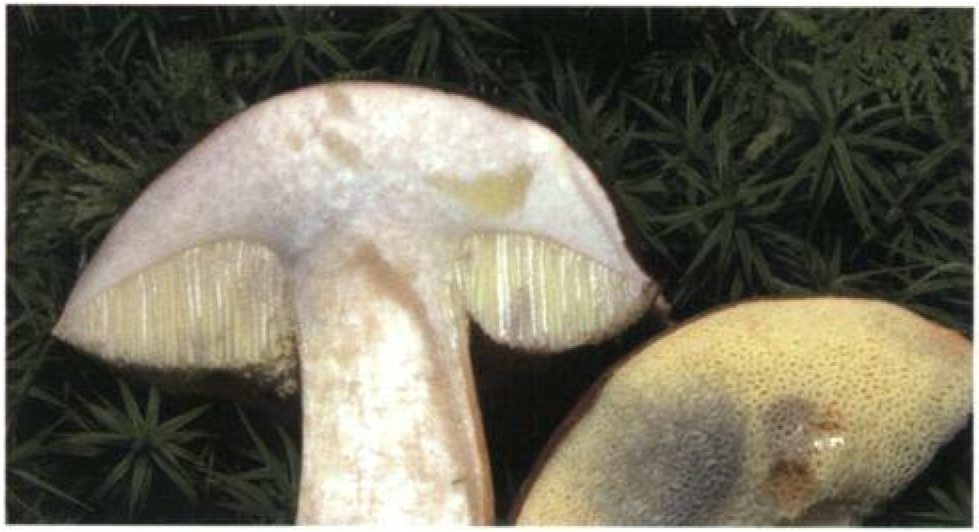

Not as good: but if you discard the soft pores and cut away the wormy parts, it will still be excellent.

The color of the cap will always be “bay” (chestnut brown) but bay has a considerable range.

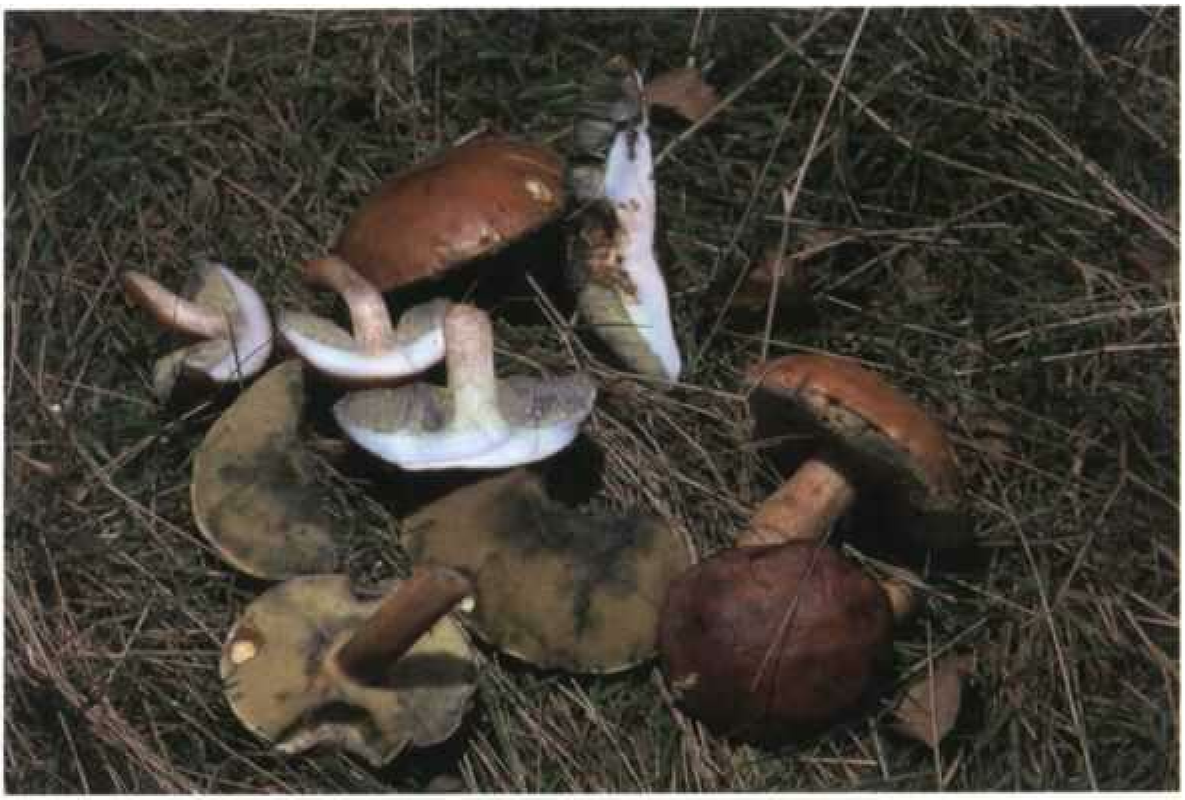

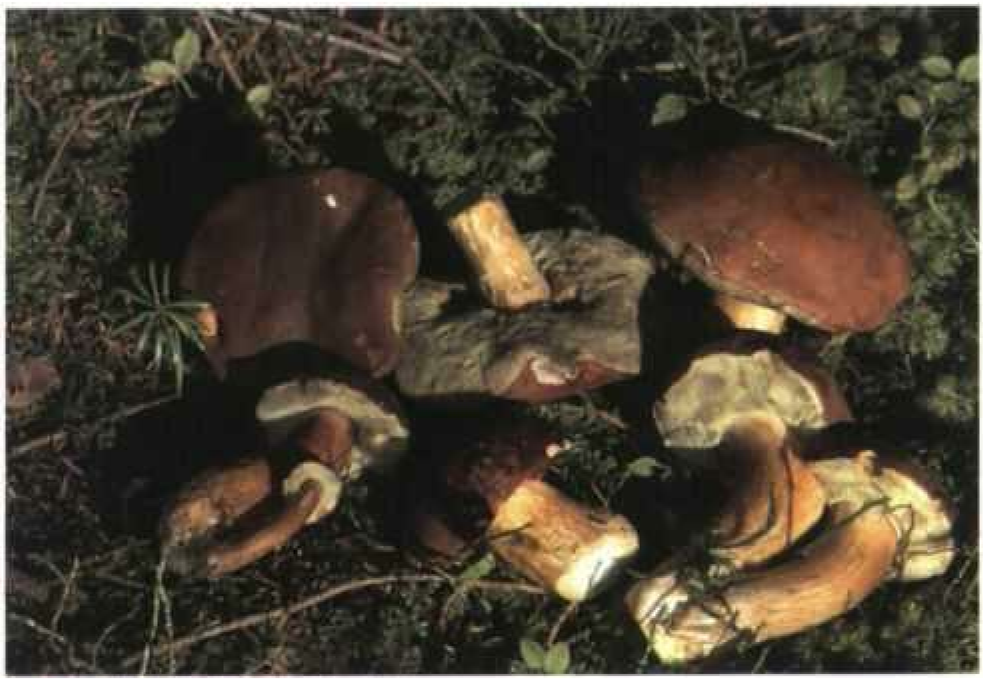

When wet, the cap will be slightly slimy and this will intensify the color of the cap.

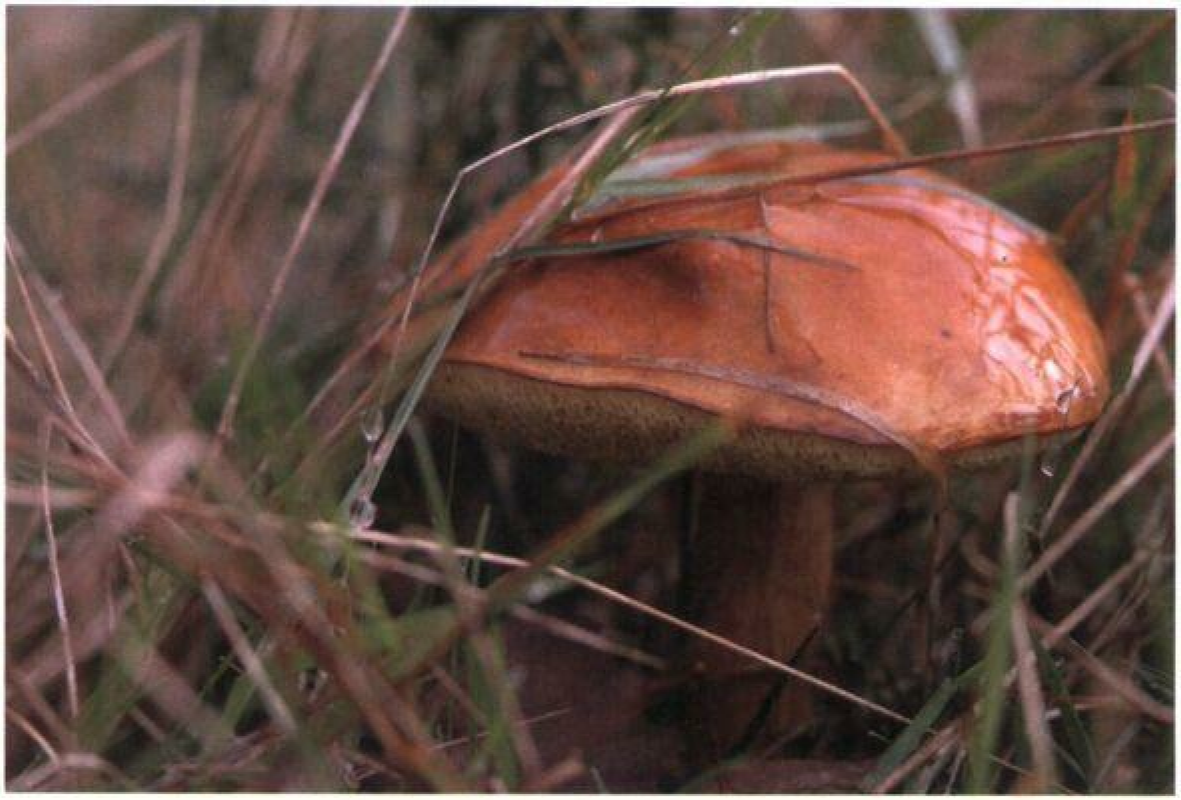

When dry, the color of the cap will be duller.

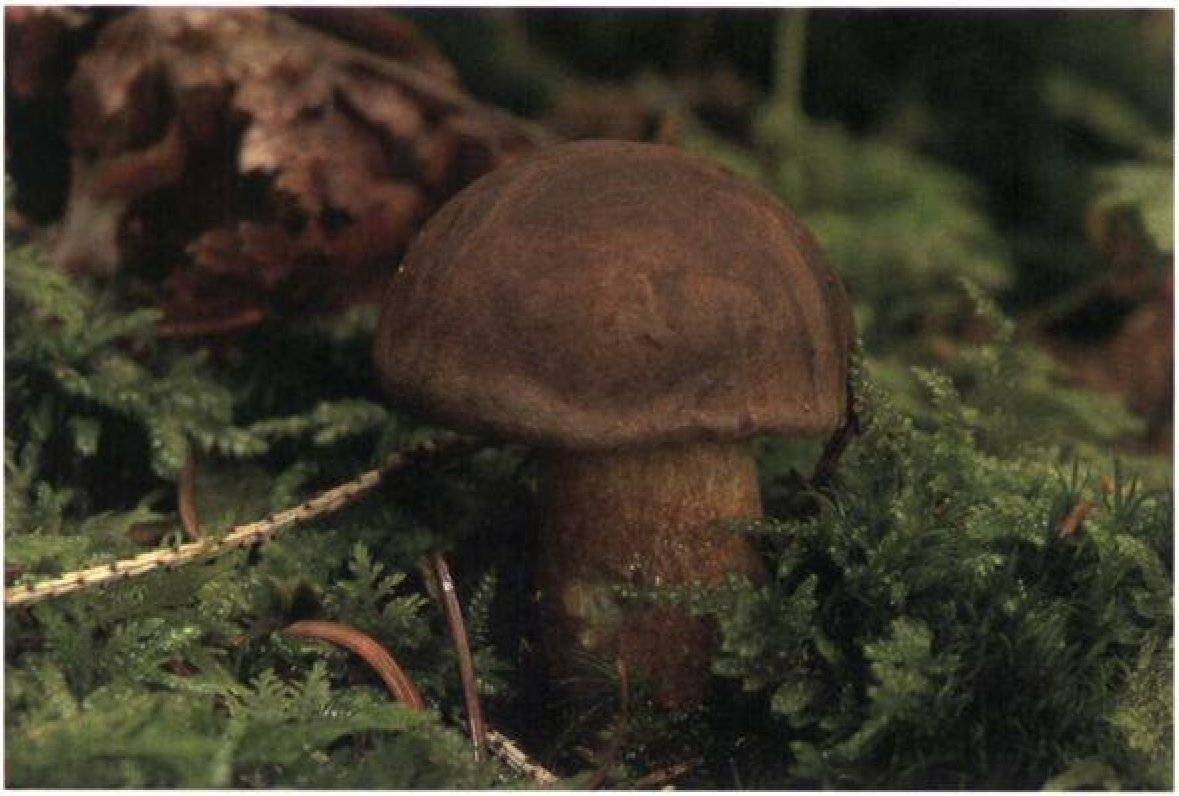

Positive ID Checklist

Bay Bolete

- ☐ Yellow tubes
- ☐ Off-white to gray pores (yellowish when aged) bruise green-blue
- ☐ Tubes and flesh bruise blue
- ☐ Stem is vertically frosted with brown streaks
- ☐ There is no net-pattern of any description on the stem
- ☐ Cap color matches cap color bar (above)

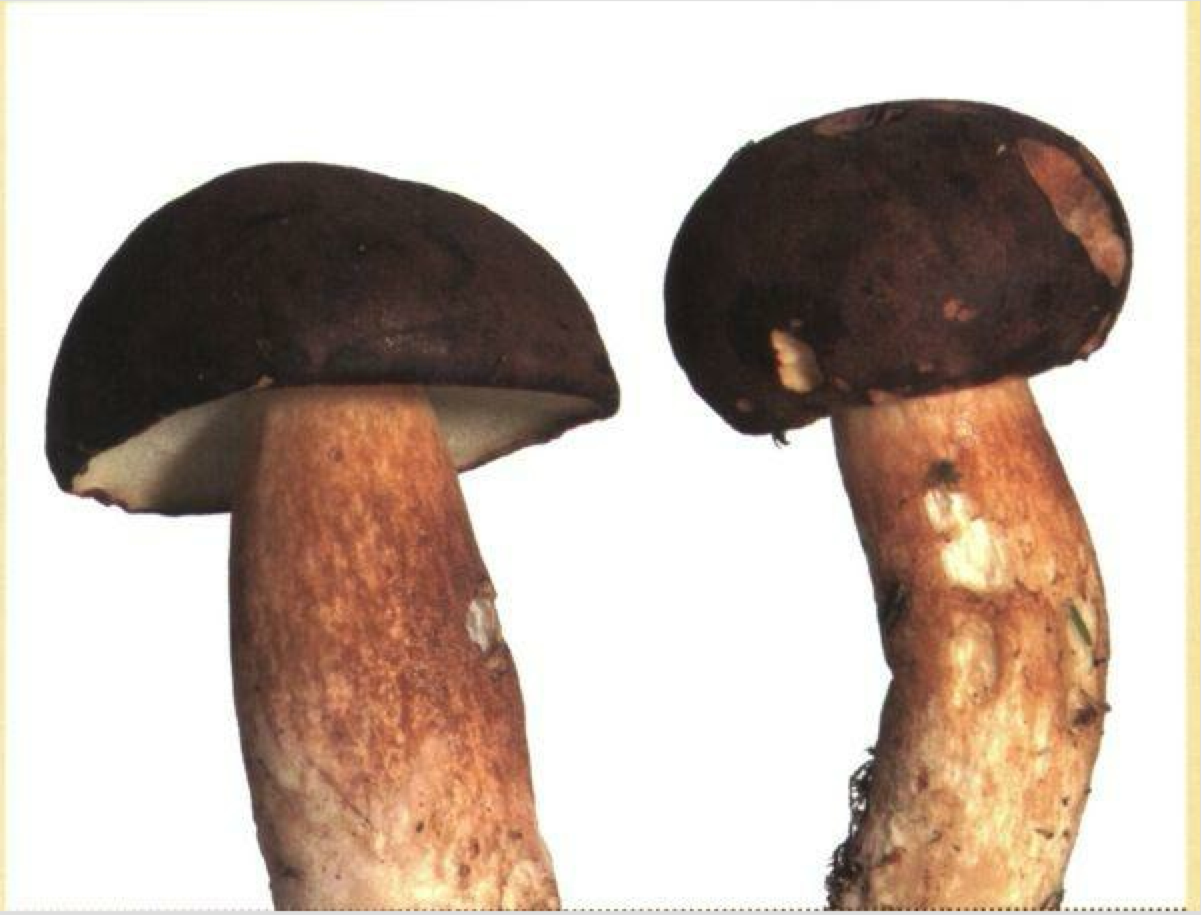

Avg. size across cap: 3.5 inches when mature

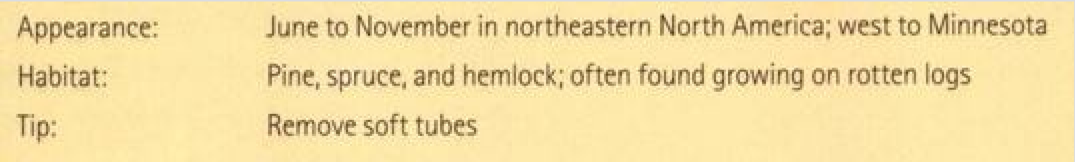

Appearance:	June to November in northeastern North America; west to Minnesota
Habitat:	Pine, spruce, and hemlock; often found growing on rotten logs
Tip:	Remove soft tubes

Birch Bolete

Leccinum scabrum

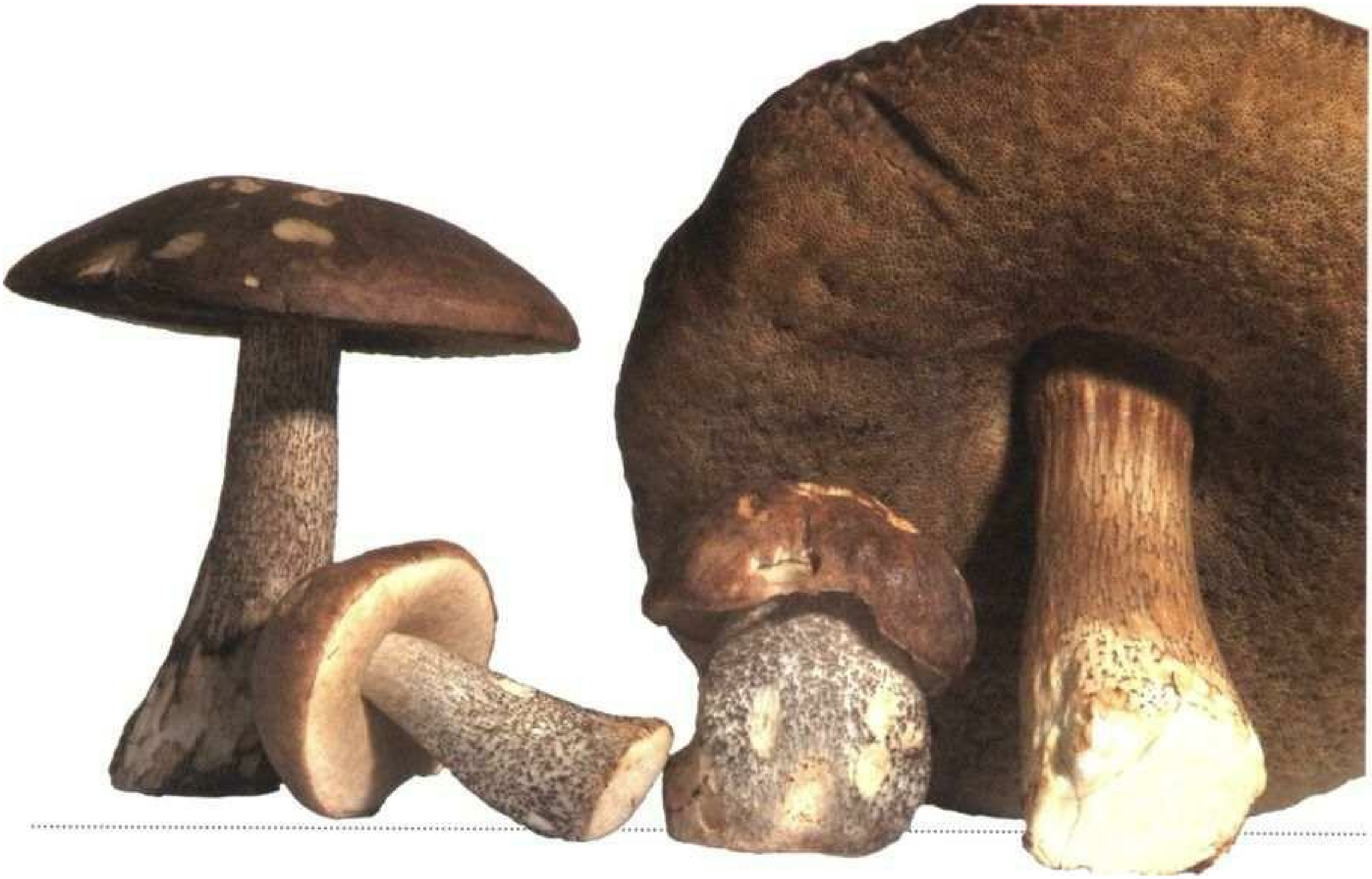

The birch bolete is the most common of the “rough stalks.” Rough stalks are mushrooms with a scaly stem. The birch bolete and its cousin, the orange birch bolete, are the most common rough stalks.
Birch woods or groups of birch trees are the places to look for them.
The birch bolete and the orange birch bolete have numerous relatives, all of which have a black or brown scaly stalk. They are all edible but it is recommended to stick to the two species presented here.

Scabrum means “rough.”

Close-up of stalk, showing black scales.

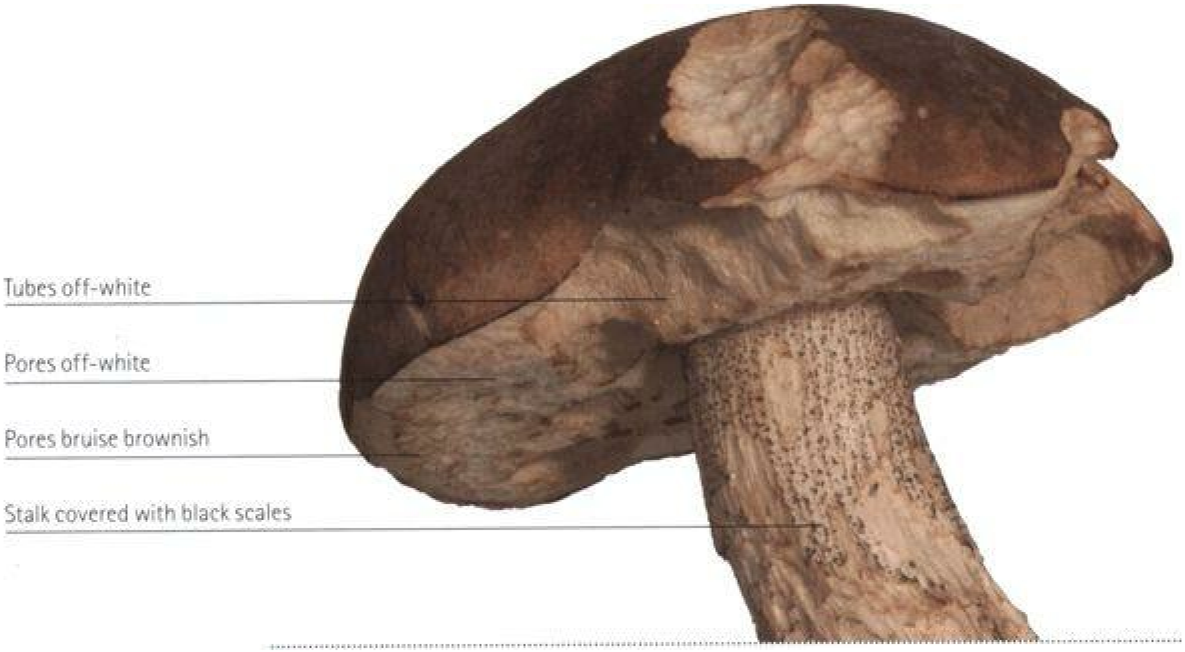
Tubes off-white
Pores off-white
Pores bruise brownish
Stalk covered with black scales

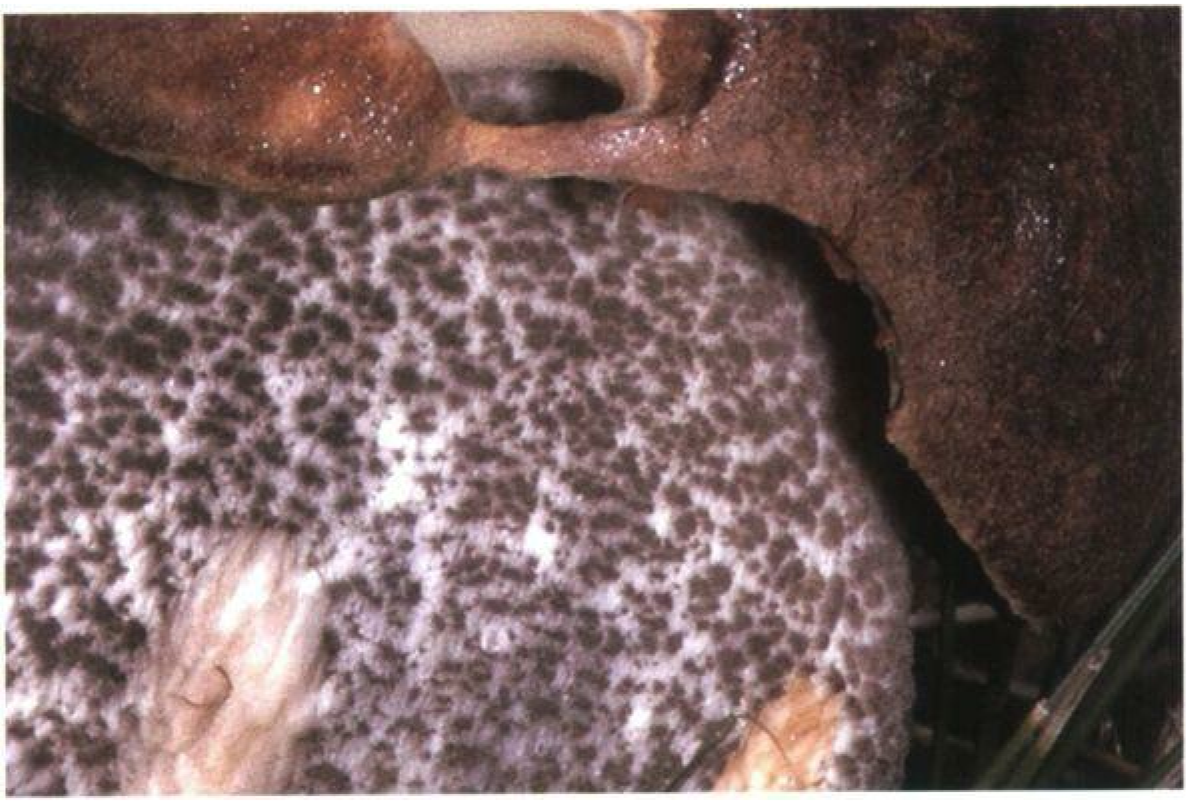

Pores and tubes change from off-white to a gray-white color as the mushroom matures. As a rule, you should only pick the birch bolete if it is firm and the pores are off-white.

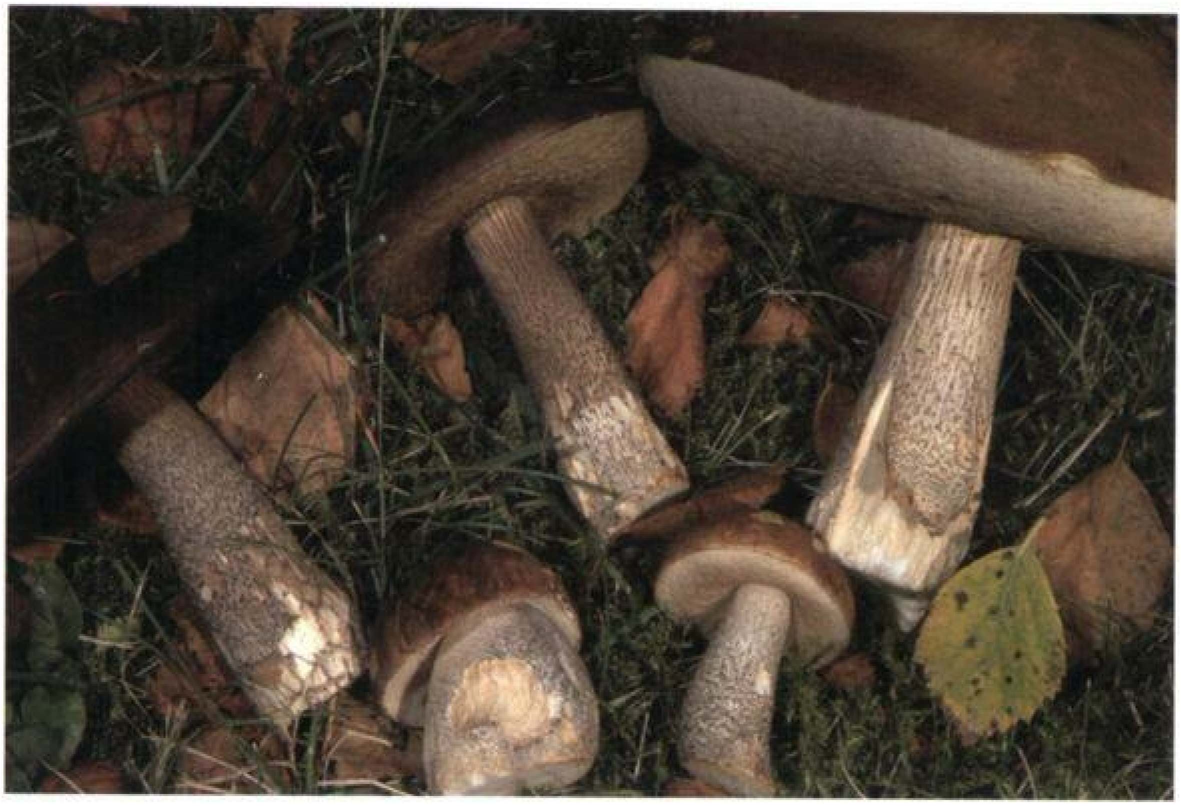

Occasionally larger specimens are still nice and firm especially when conditions have been very dry.

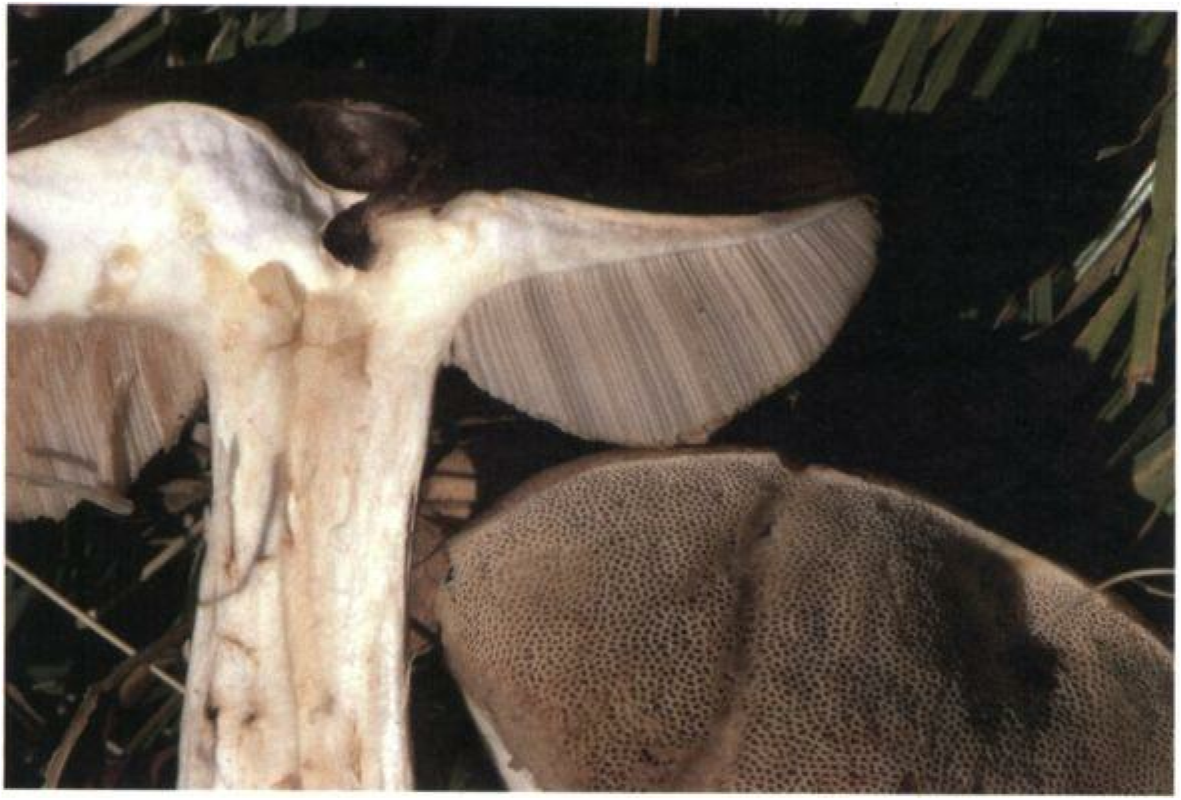

These are the perfect birch boletes: a feast for the eyes and delicious to eat.

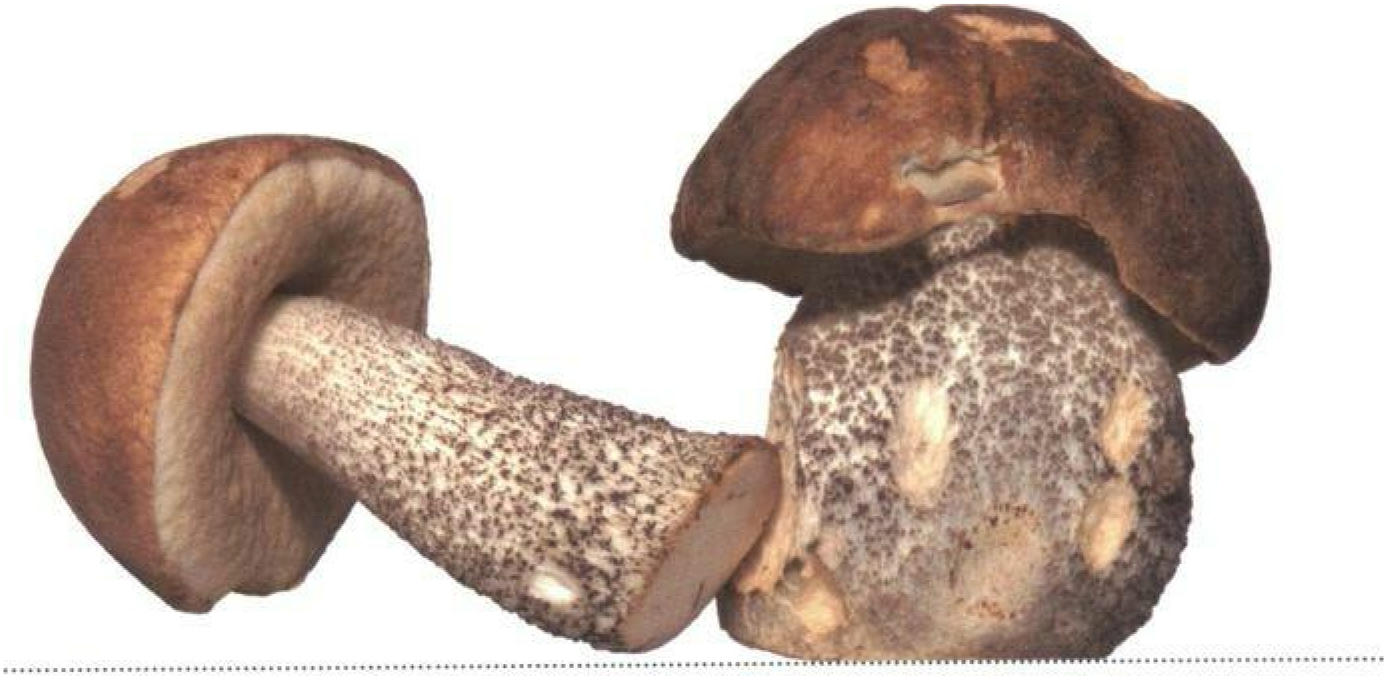

The orange birch bolete, *Leccinum versipelle,* is a beautiful mushroom. Its key identification mark is the stem covered with black scales, and of course the orange cap. When cut, the flesh turns faintly blue at the stalk base, displays traces of wine red and finally turns and stays gray. This color combination doesn't look especially appetizing but be assured the orange birch bolete is delicious.

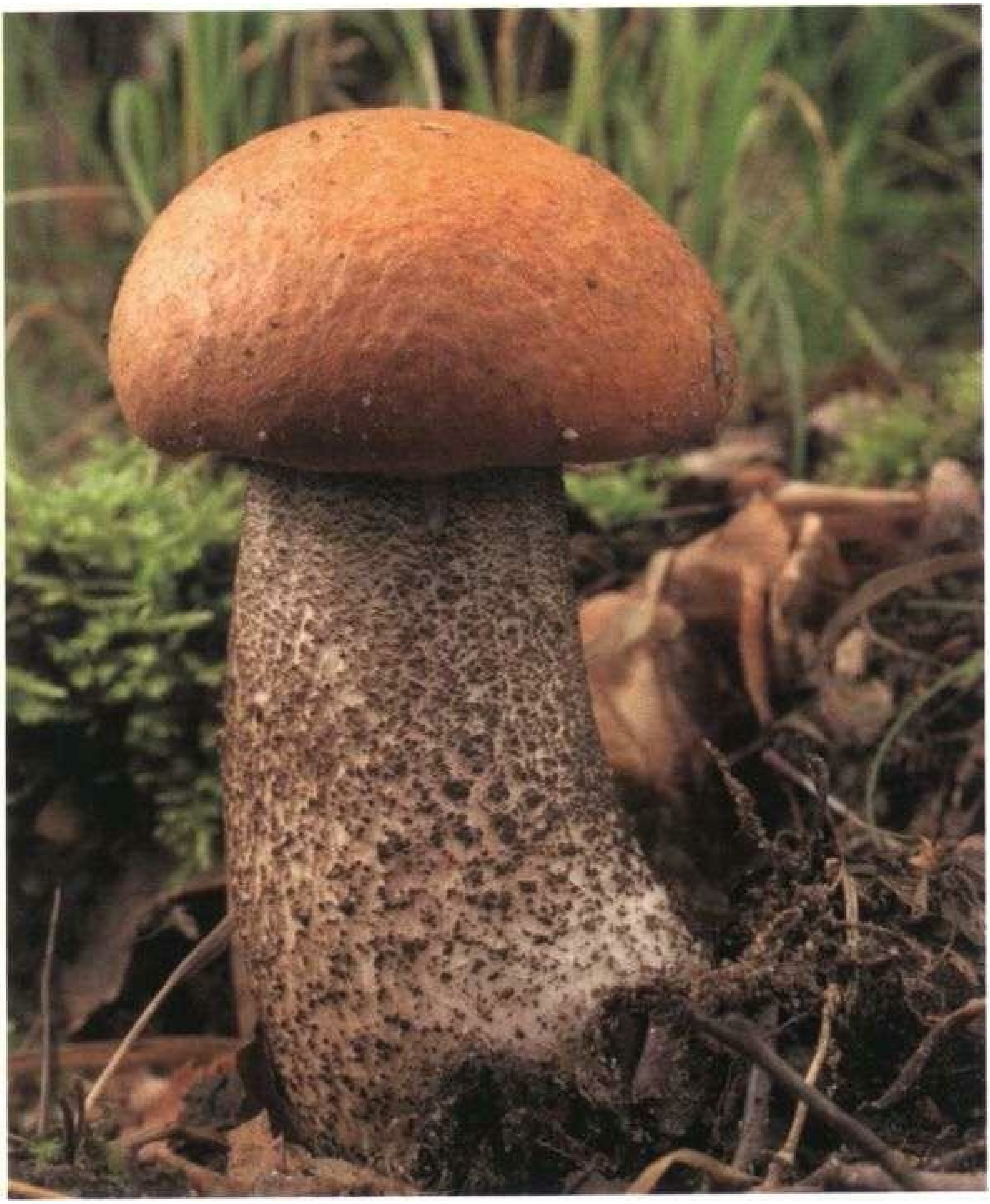

Positive ID Checklist

Birch Bolete

- ☐ Pores off-white
- ☐ Pores bruise brownish
- ☐ Stalk with small black scales
- ☐ Cap color matches cap color bar (above)

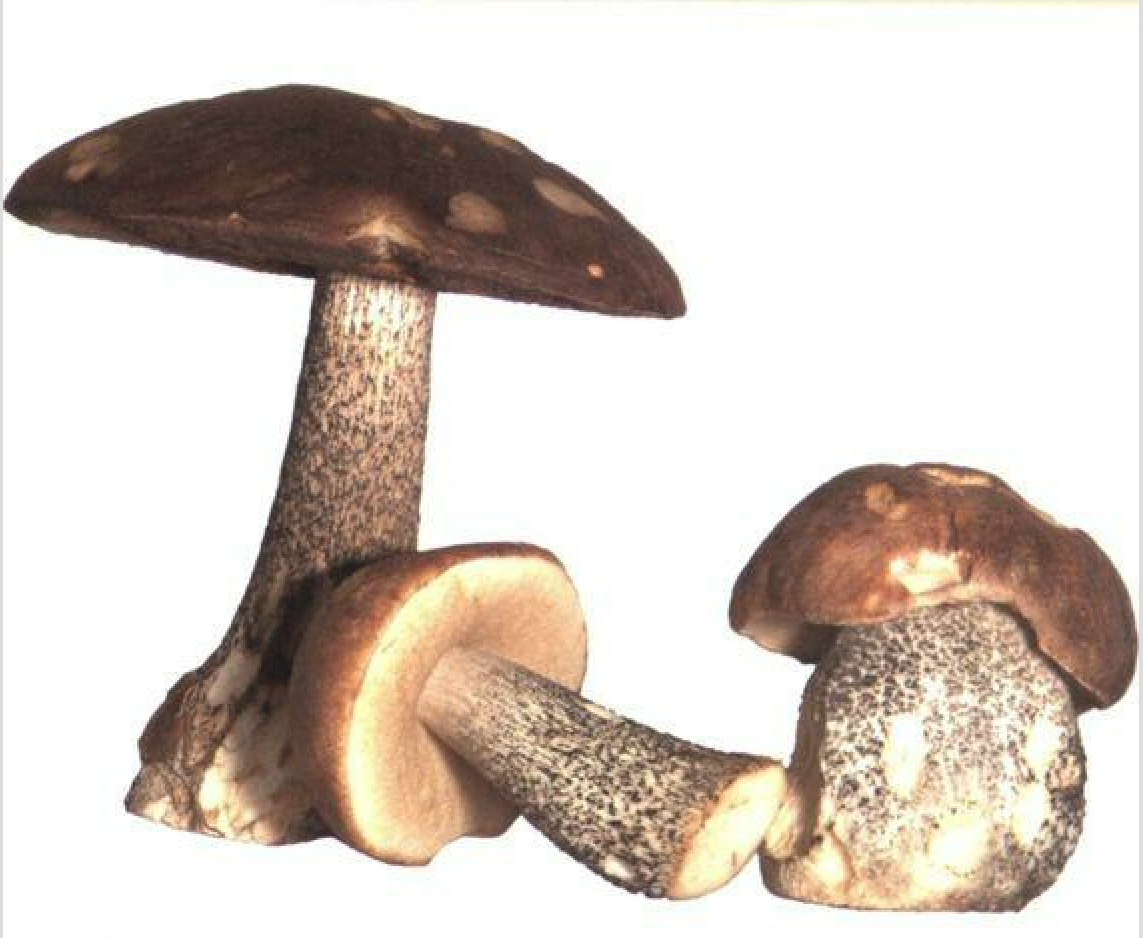

Avg. size across cap: 3.5 inches when mature

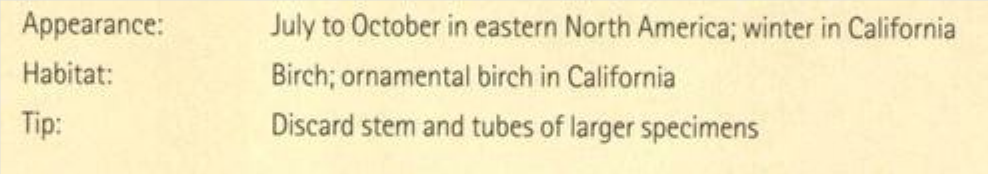

Appearance:	July to October in eastern North America; winter in California
Habitat:	Birch; ornamental birch in California
Tip:	Discard stem and tubes of larger specimens

Always remember

Rule number 1

Never, never take a mushroom with gills!!!

Rule number 2

Only take mushrooms with tubes, spines, and ridges and the mavericks portrayed in this book

Rule number 3

Only eat mushrooms which you have clearly identified with ALL the positive ID marks

Rule number 4

If a mushroom smells rotten, it is rotten and if it feels soggy, it is soggy

Rule number 5

Never, never eat wild mushrooms raw

Rule number 6

Look before you cut

Rule number 7

Mushrooms want to breathe. Do not suffocate them in plastic bags

Rule number 8

If in doubt, leave it

Mushrooms With Ridges

Chanterelle

Cantharellus cibarius

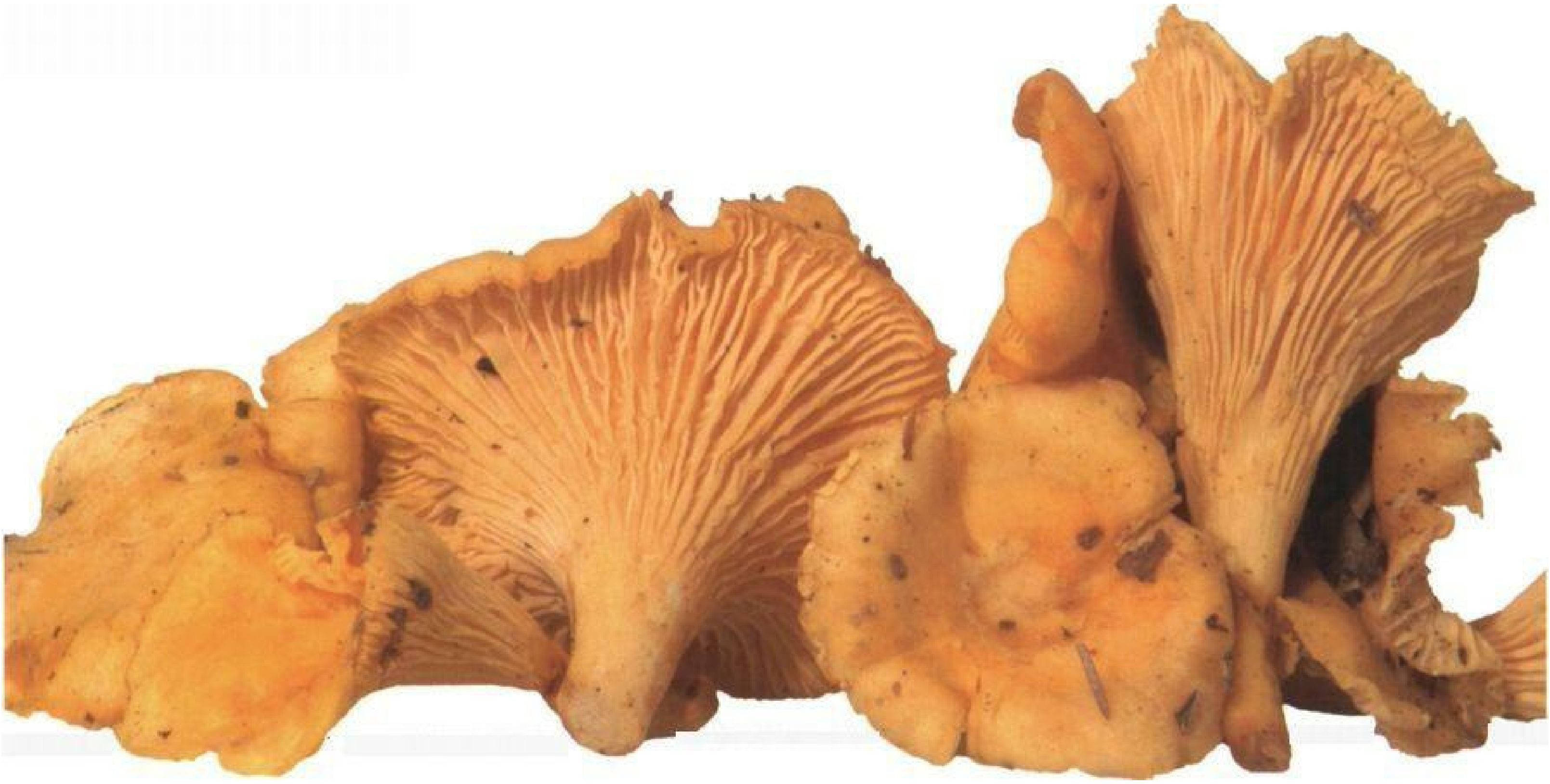

The chanterelle is not surprisingly the most popular mushroom. It is delicious and very common. As it is imported throughout the year, the supermarket is a good place to familiarize yourself with the chanterelle. There is, however, nothing like the chanterelle you find yourself. Refrain from picking tiny chanterelles, that is, little yellow buttons which hardly show the identification marks. At some time you'll come across one lonely chanterelle standing there in the middle of nowhere for no particular reason at all. Leave it.

Cibarius means "belonging to food."

The classic color of the chanterelle is egg-yolk yellow which, however, can vary from light yellow or yellow-orange to yellow-ochre. The chanterelle does not change color when bruised. The entire mushroom is of the same color.

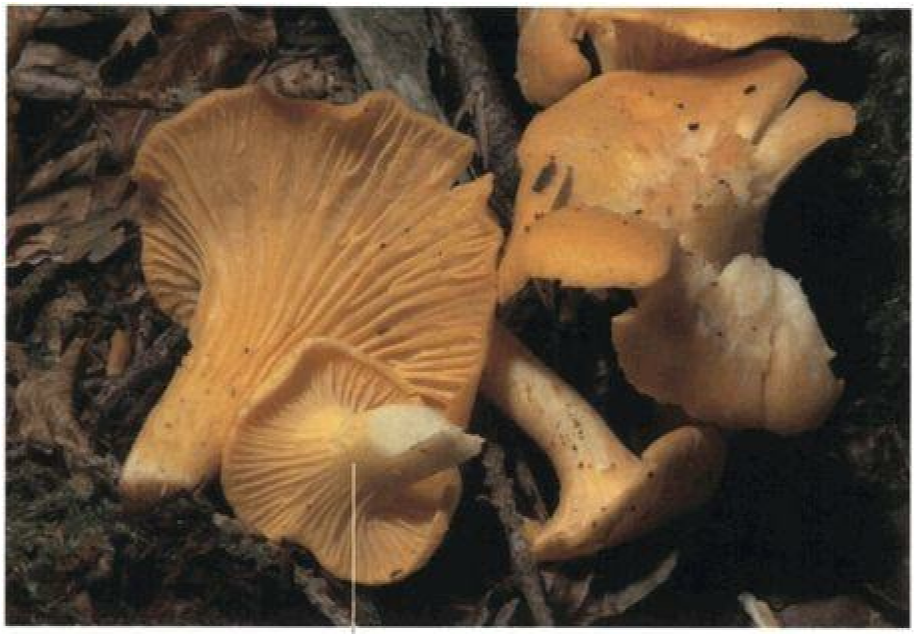

Young chanterelles have a flat cap with a slightly out-rolled margin. Do not pick any chanterelles smaller than 3 cm high.

The more mature the chanterelle, the more funnel-shaped it becomes.

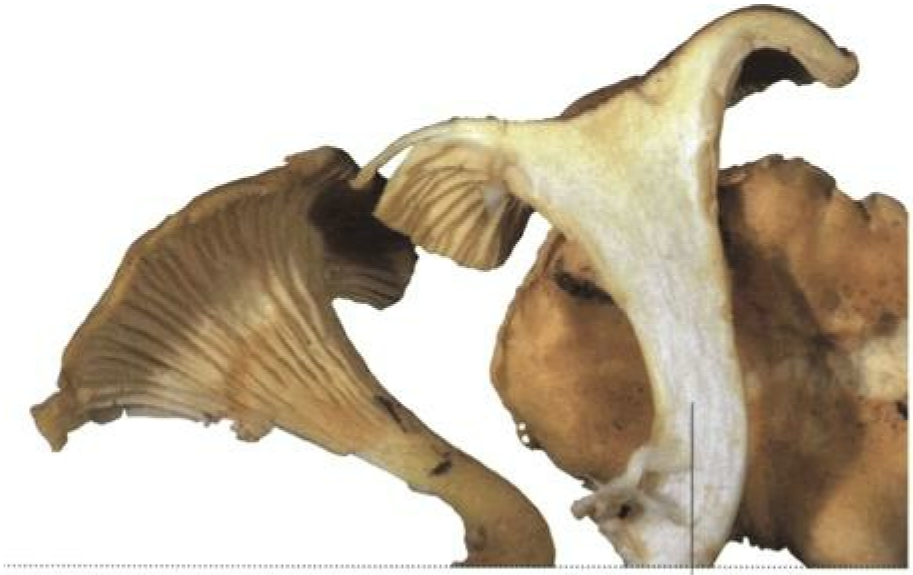

The stalk is solid. The whitish to yellowish flesh does not change color when bruised or cut.

Whether young or mature, the chanterelle’s ridges must be clearly visible.

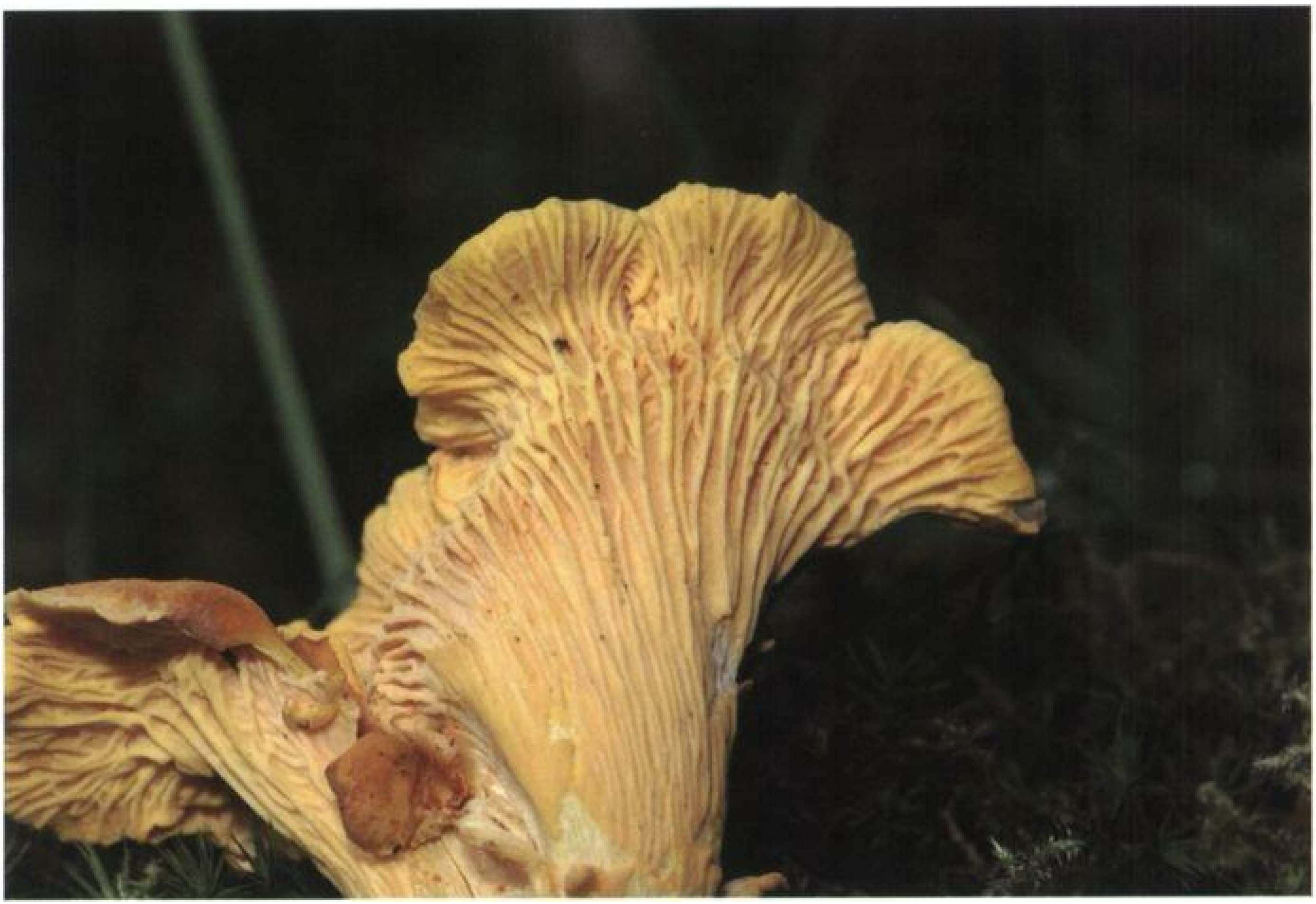

Some mature forms can look highly irregular but on closer inspection, the funnel shape will still be there.

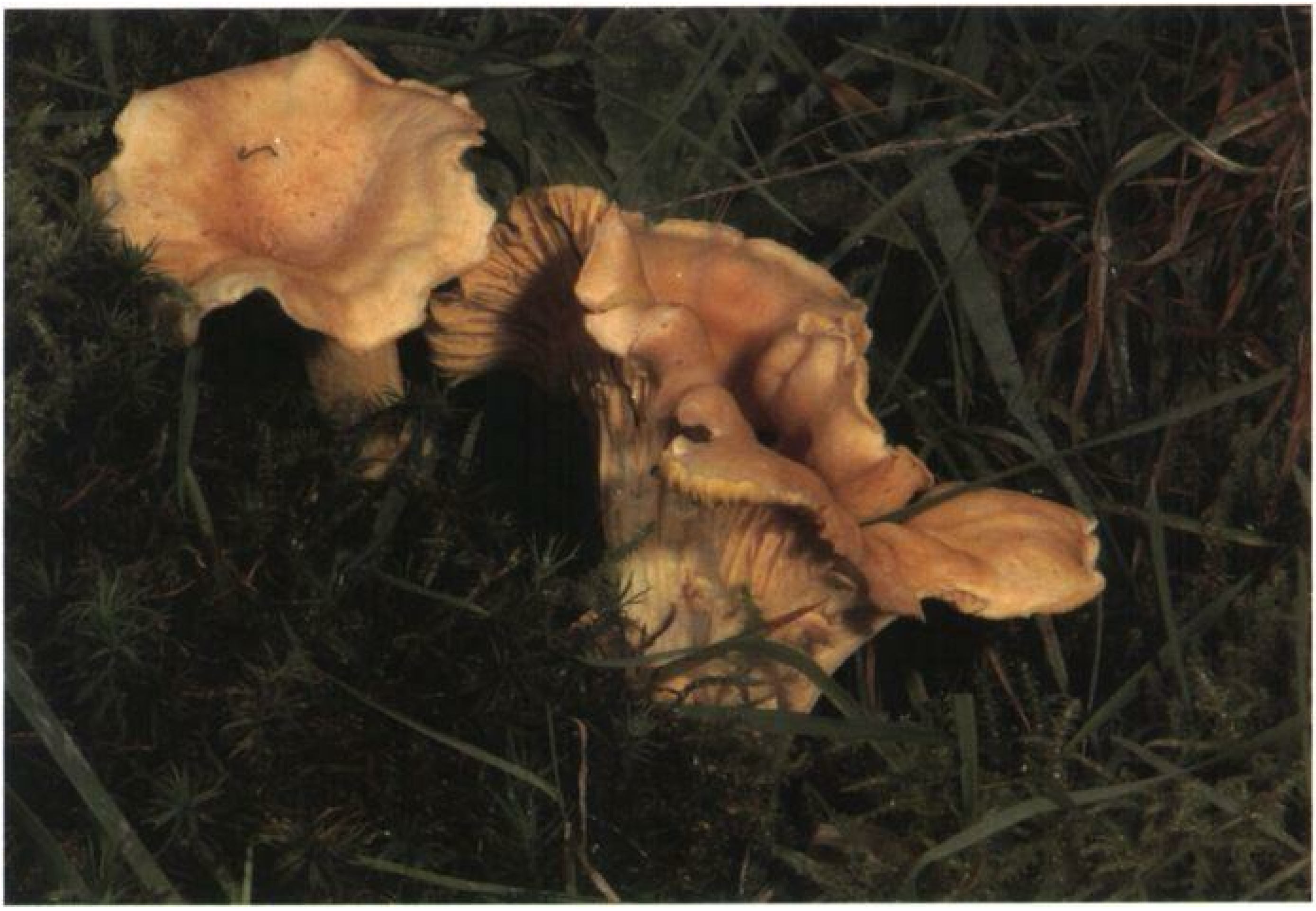

Positive ID Checklist

Chanterelle

☐ Young specimen: Flat, wavy-edged cap, slightly out-rolled margin
☐ Mature specimen: Funnel-shaped
☐ Solid stem
☐ Found in groups on the ground, but not clustered at the base of trees; never on wood
☐ Ridges must be clearly visible
☐ Color matches cap color bar range (above)

Be Careful:

Don’t mistake the chanterelle for the Jack O‘Lantern, a mushroom that grows across eastern North America. The Jack O’Lantern is orange, glows in the dark, has gills, and is not edible.

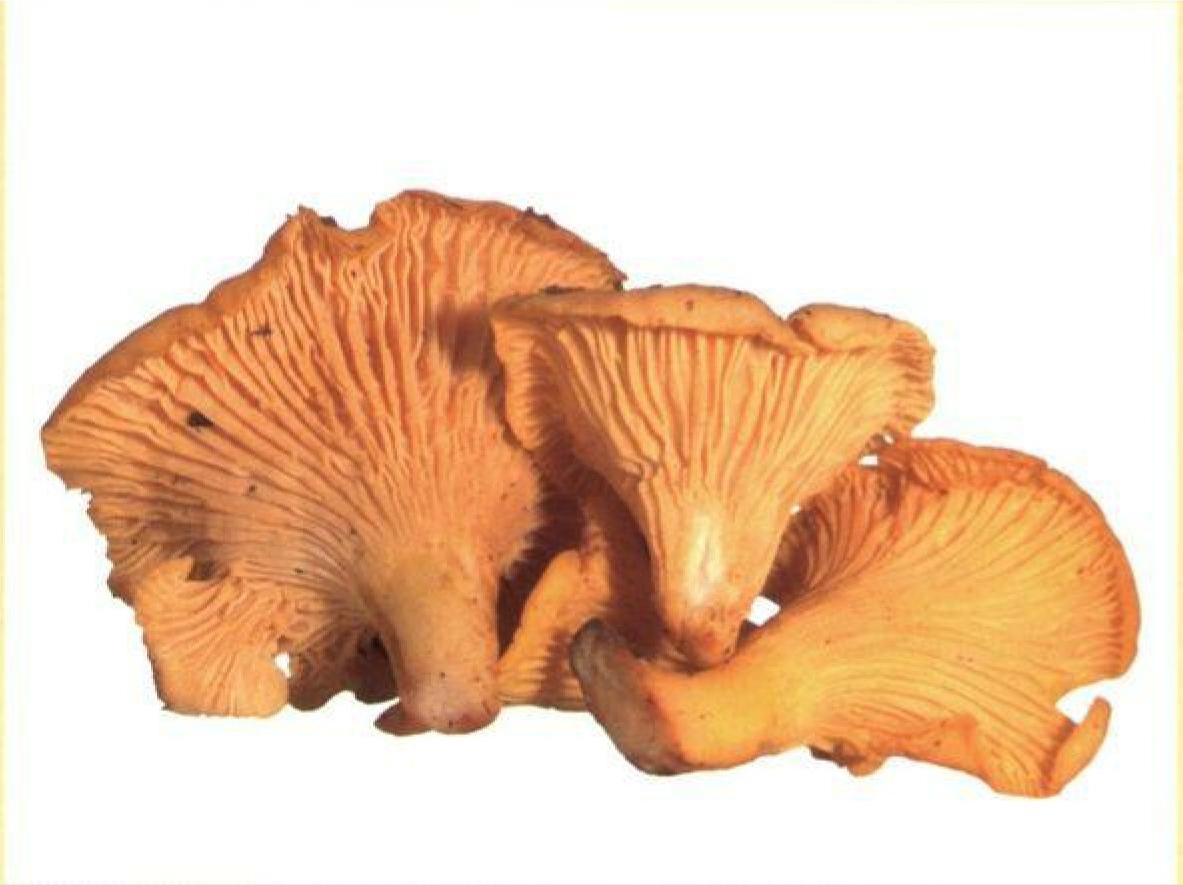

Avg. size across cap: 2 inches when mature

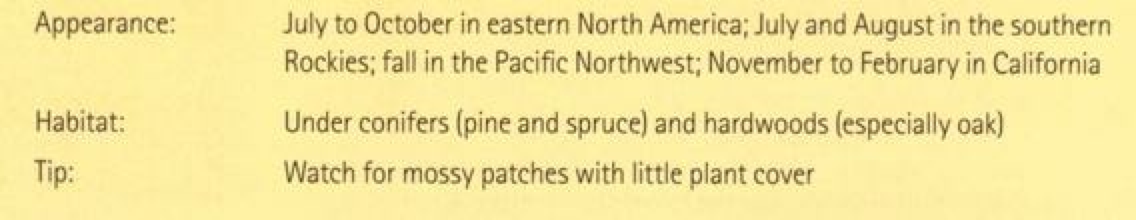

Appearance:	July to October in eastern North America; July and August in the southern Rockies; fall in the Pacific Northwest; November to February in California
Habitat:	Under conifers (pine and spruce) and hardwoods (especially oak)
Tip:	Watch for mossy patches with little plant cover

Trumpet Chanterelle

Cantharellus tubaeformis

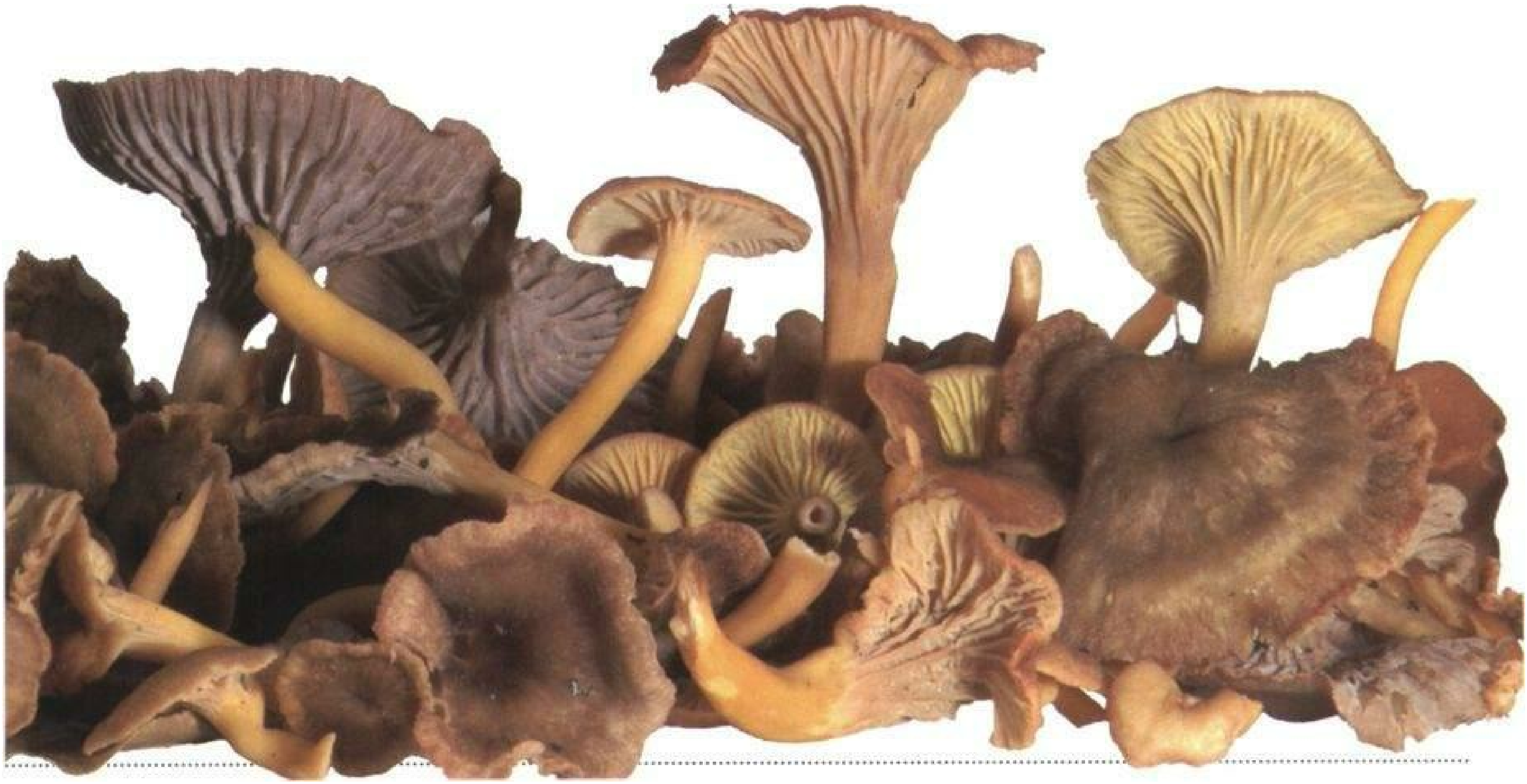

In comparison to the attractively-colored chanterelle, the trumpet chanterelle (also known as the autumn chanterelle, and winter chanterelle) looks modest. On closer inspection, however, it is just as beautiful as the chanterelle. The trumpet chanterelle pops up overnight so quickly that you could watch it grow. It is an endearing mushroom not least because it tends to appear in great numbers and is frost-resistant.

Tubaeformis means “trumpet-shaped.”

A cluster of trumpet chanterelles.

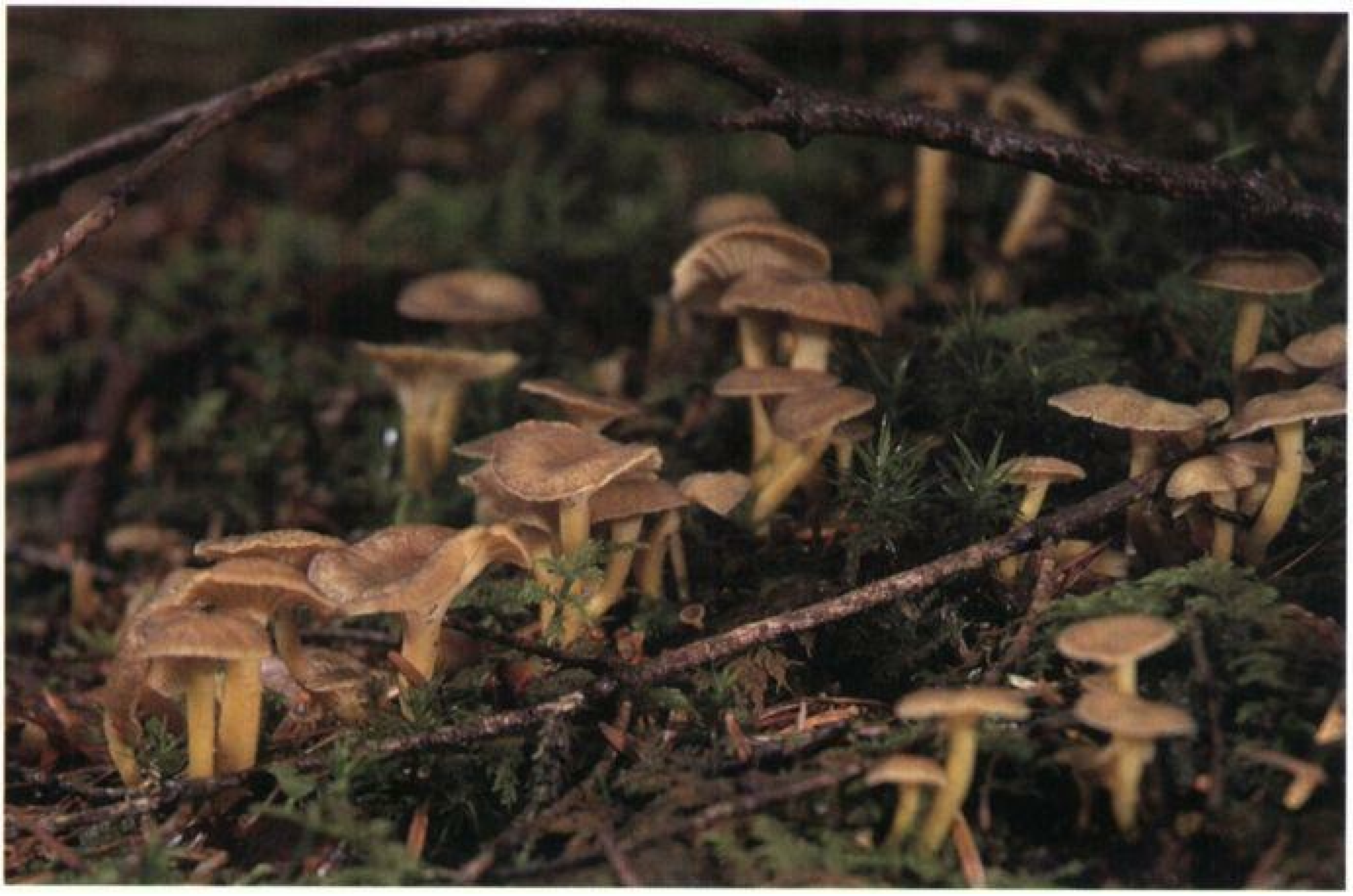

Color and shape vary according to age and weather conditions. Each individual trumpet chanterelle has more than one color. The cap is brownish and has a hole in the center. The ridges and the stalk range from brown-yellow to gray-yellow to gray-lilac.

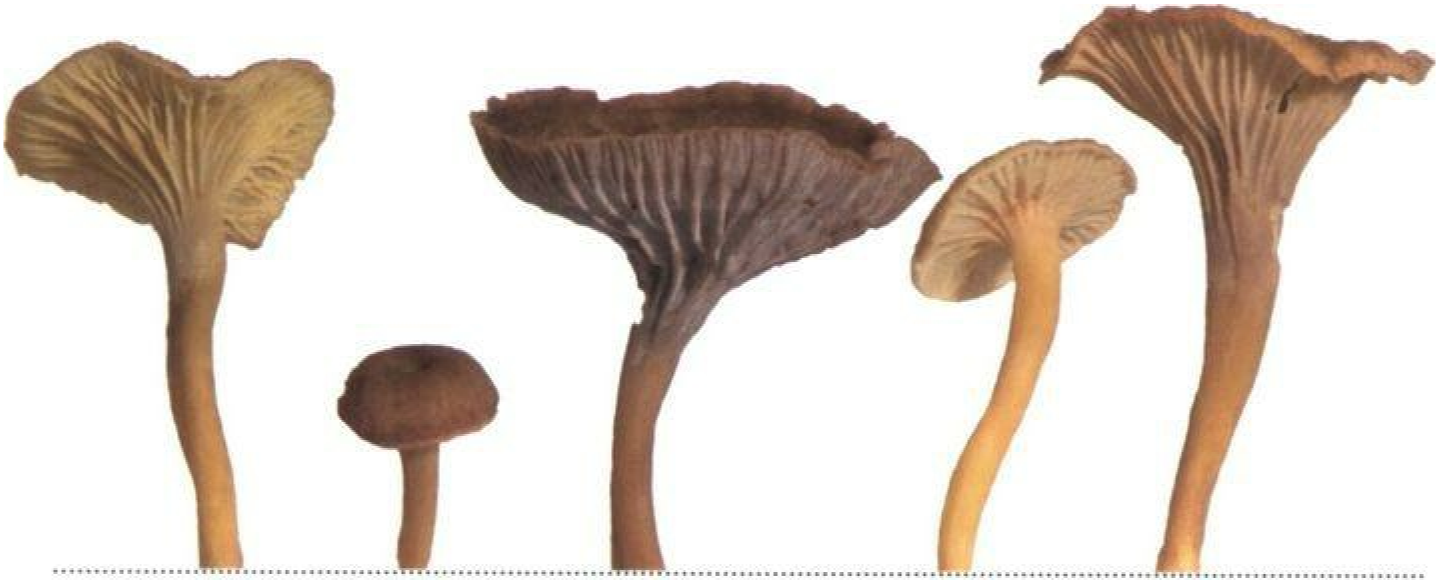

The jagged edges here are due to frost.

The advantage of having a paper bag for each species (see page 32) is seen here. You can spot the odd one out immediately.

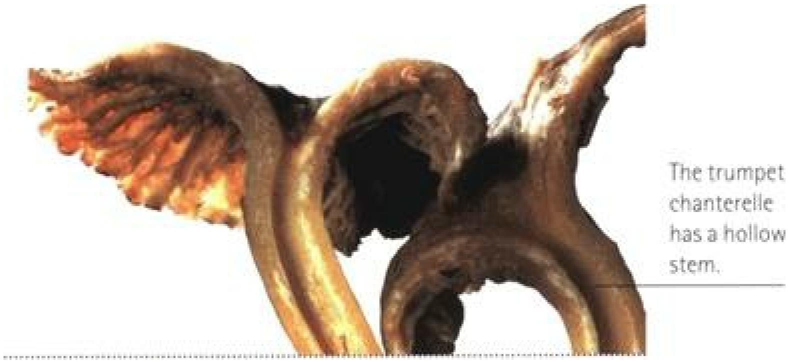

The trumpet chanterelle has a hollow stem.

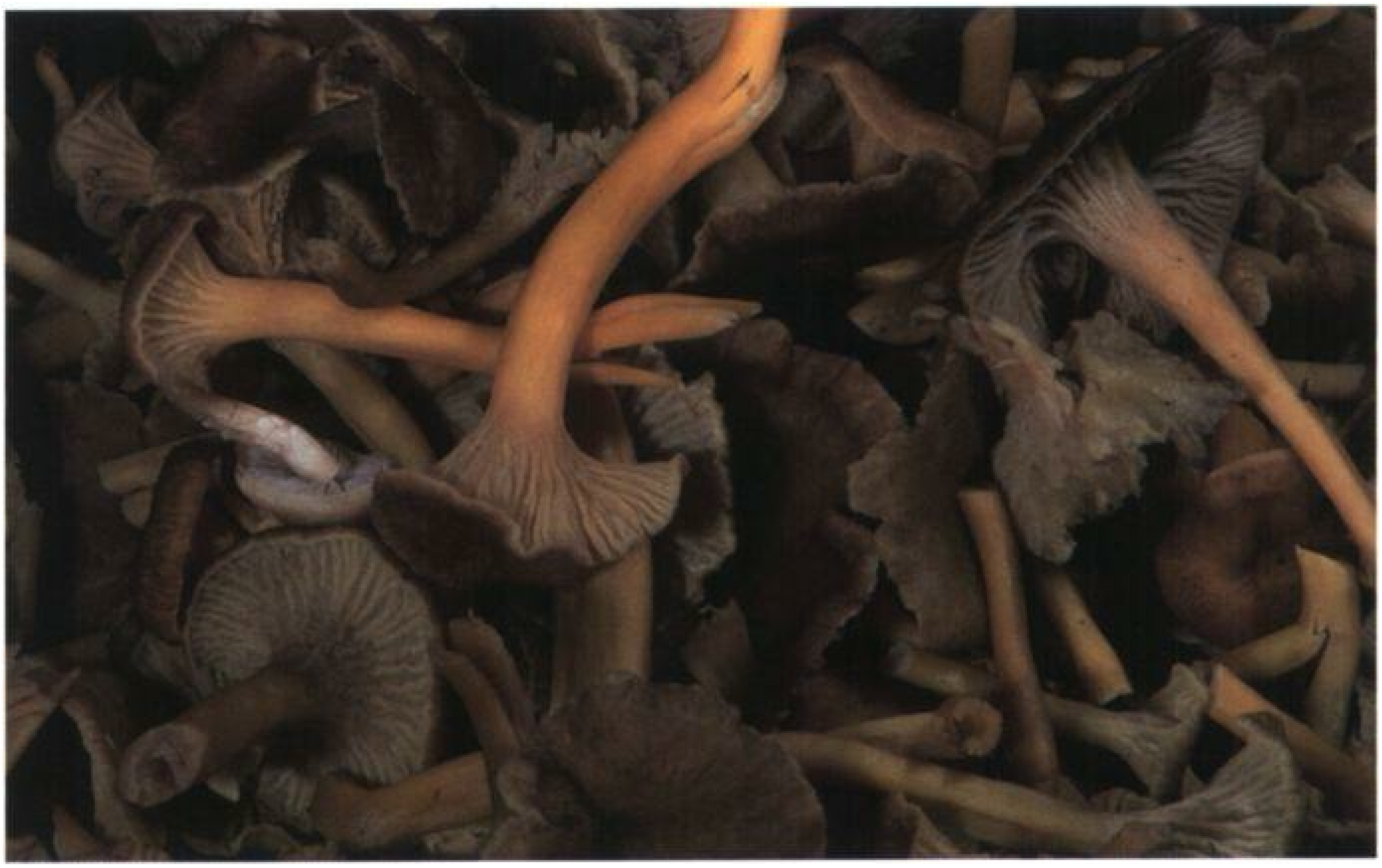

PositiveIDChecklist

Trumpet Chanterelle

☐ Found in groups
☐ Brownish, thin-fleshed, wavy-edged cap with hole in center
☐ Hollow stem
☐ Ridges
☐ Cap color matches color bar (above)

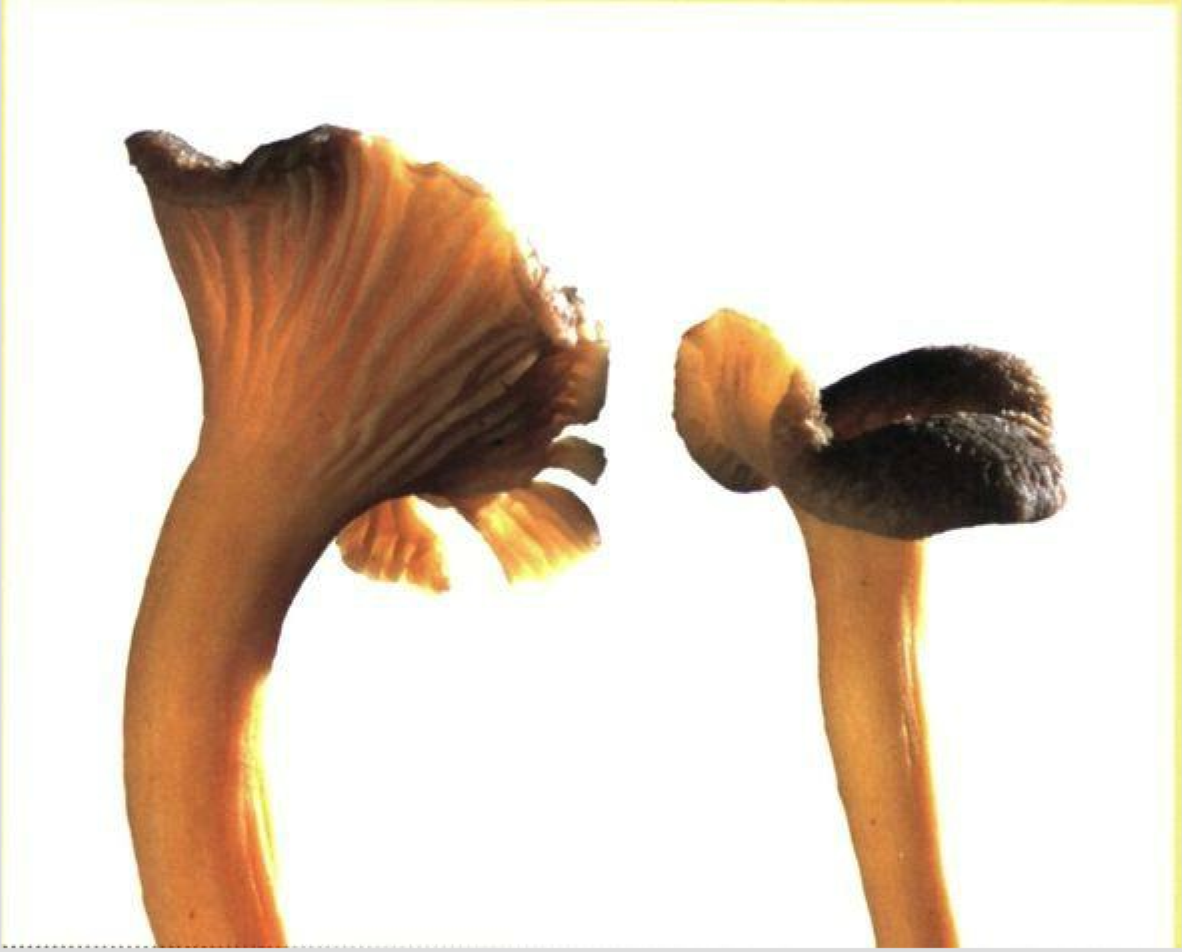

Avg. size across cap: 0.75-1 inch when mature

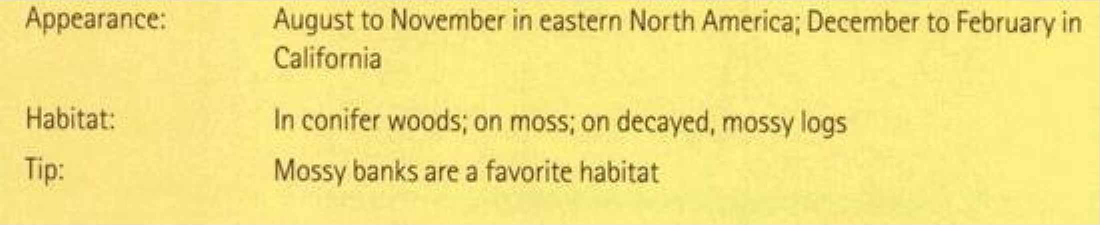

Appearance:	August to November in eastern North America; December to February in California
Habitat:	In conifer woods; on moss; on decayed, mossy logs
Tip:	Mossy banks are a favorite habitat

Always remember

Rule number 1

Never, never take a mushroom with gills!!!

Rule number 2

Only take mushrooms with tubes, spines, and ridges and the mavericks portrayed in this book

Rule number 3

Only eat mushrooms which you have clearly identified with ALL the positive ID marks

Rule number 4

If a mushroom smells rotten, it is rotten and if it feels soggy, it is soggy

Rule number 5

Never, never eat wild mushrooms raw

Rule number 6

Look before you cut

Rule number 7

Mushrooms want to breathe. Do not suffocate them in plastic bags

Rule number 8

If in doubt, leave it

Mushrooms With Spines

Hedgehog Fungus

Hydnum repandum

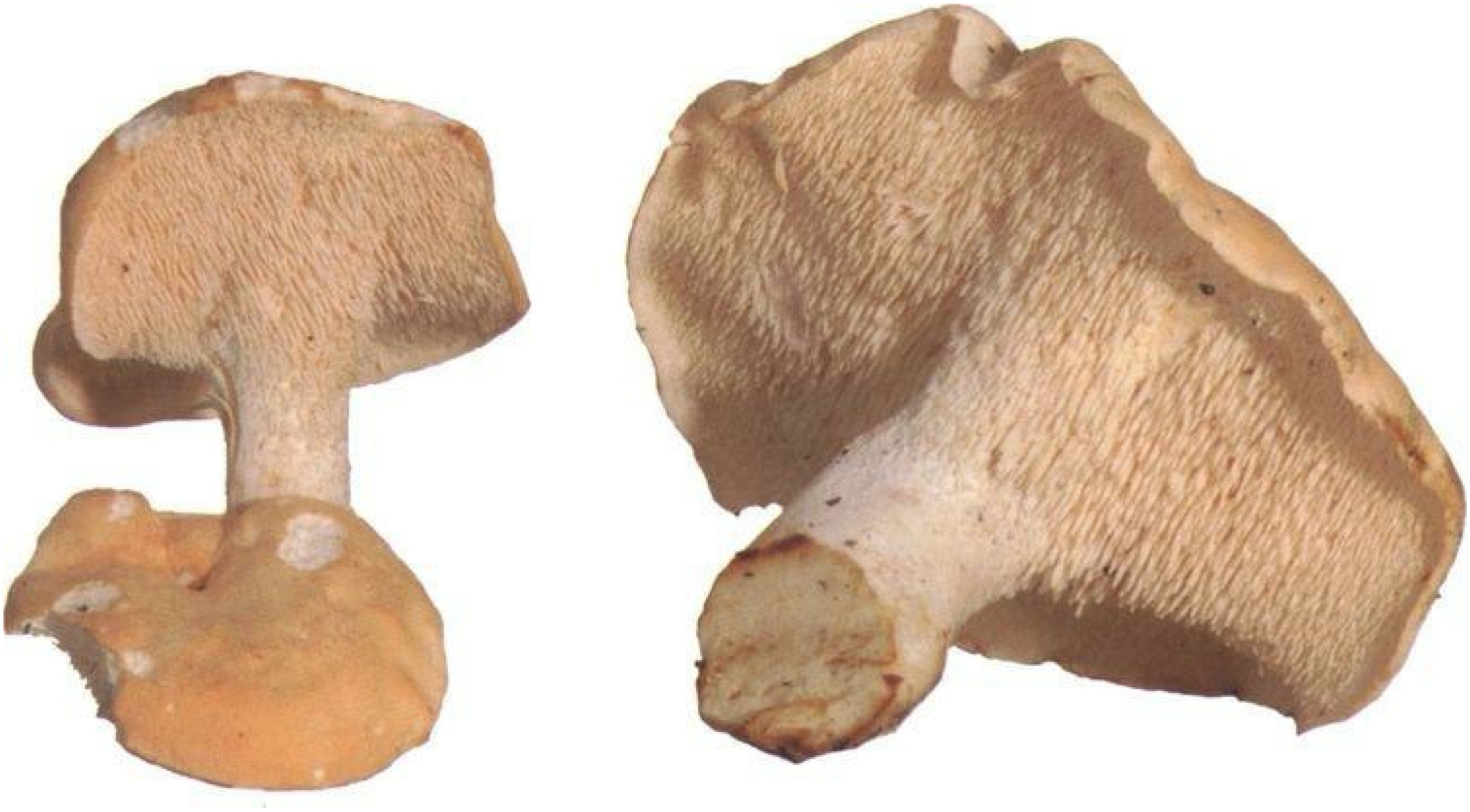

There are other mushrooms with spines but the only one of real culinary interest is the hedgehog fungus. The hedgehog fungus is a great delicacy. It grows on the ground but never on trees.

Repandum means “bent upwards” and refers to the way the cap often reveals its underside.

The spines are the key identification marks of the hedgehog fungus. The flesh is matte white and colors in places to a yellow-brown or rusty yellow.

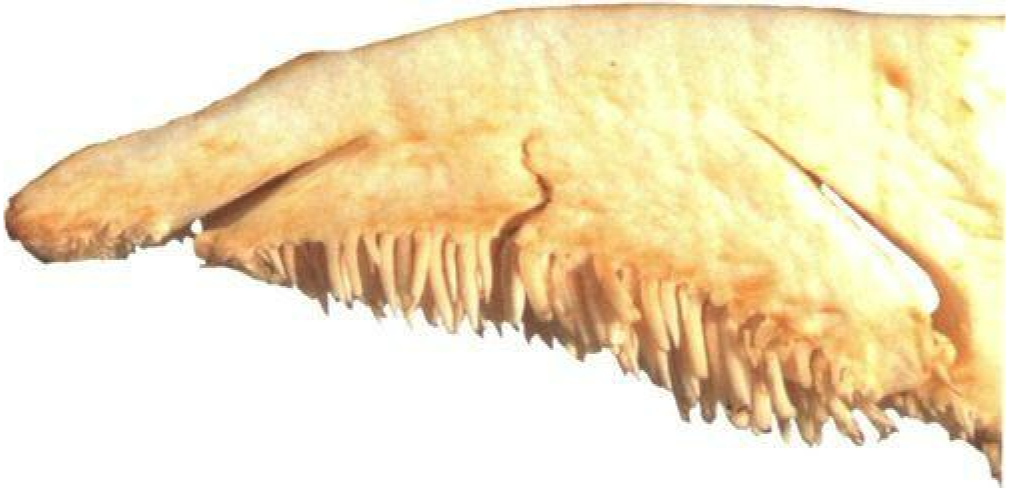

Some hedgehog fungi grow in a very irregular fashion, up to the size of two fists. Nevertheless, all hedgehog fungi have distinctive spines. The specimens here show the entire color range of the cap. No other mushroom with spines has these colors.

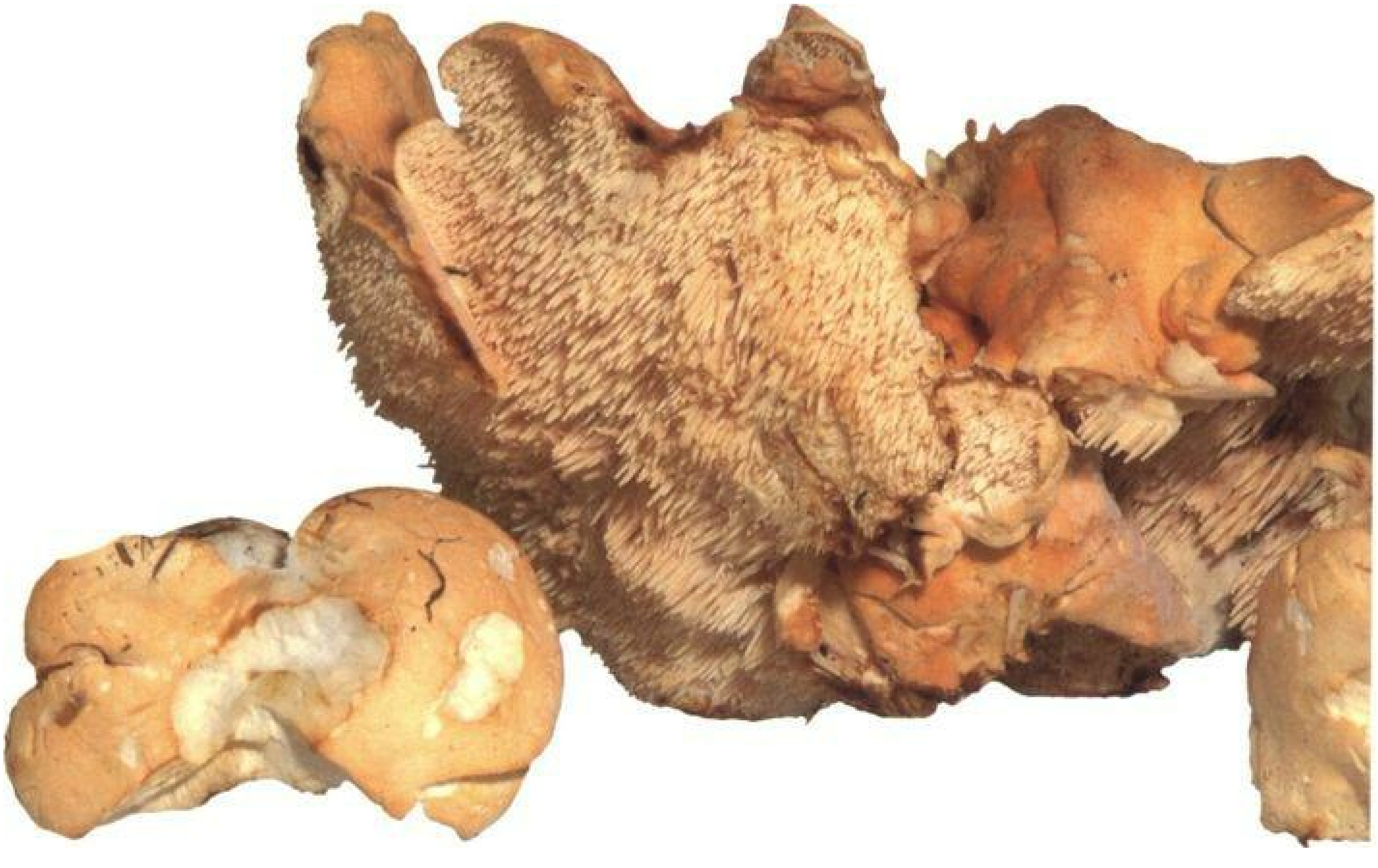

Take only specimens where the spines are clearly visible. The hedgehog fungi pictured here are the perfect size for the kitchen. The cap size is 2.5-4 inches across.

Positive ID Checklist

Hedgehog Fungus

☐ Spines
☐ Flesh matte white when freshly cut; flesh colors in places to yellow-brown or rusty yellow
☐ Cap margin wavy-edged
☐ Found on ground but not on trees
☐ Cap color matches color bar (above)

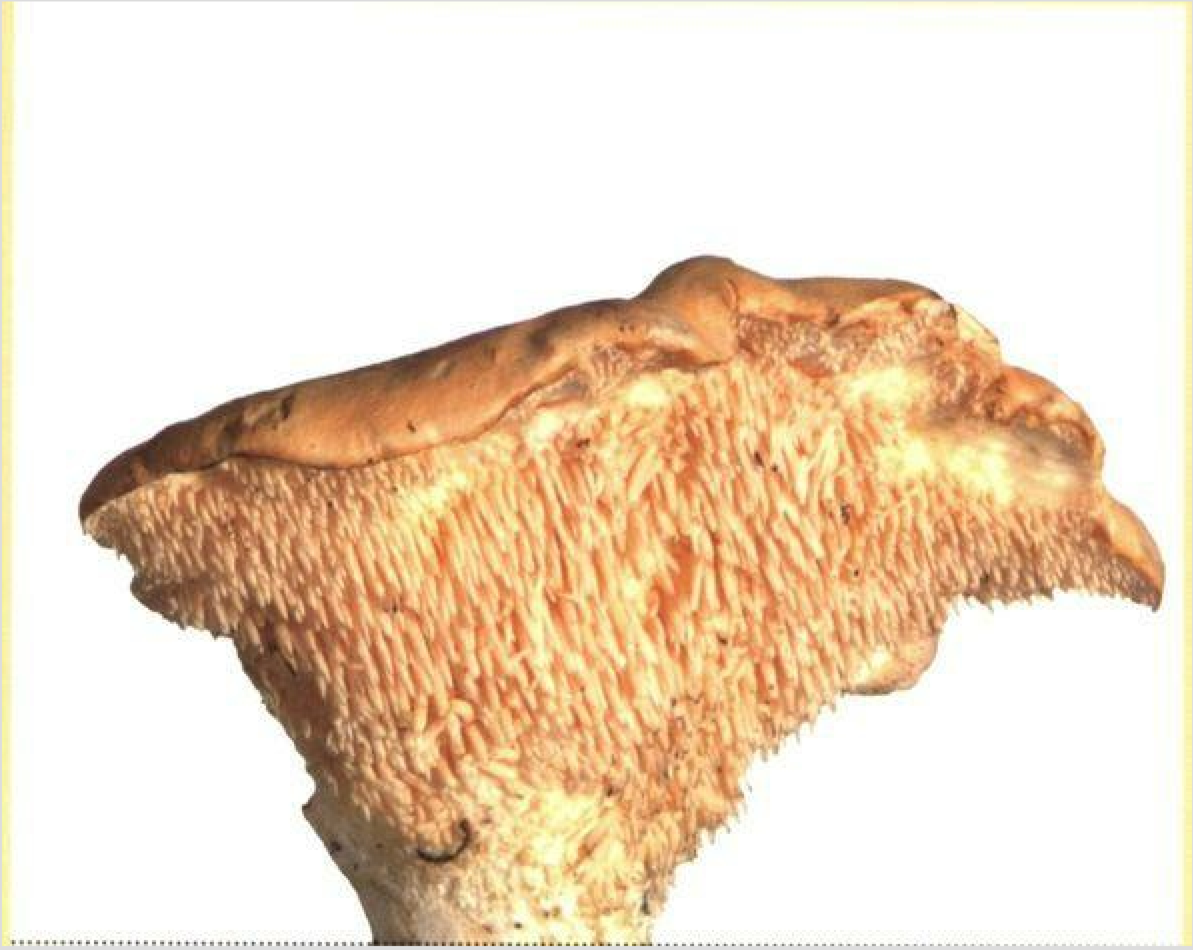

Avg. size across cap: 2.5 inches when mature

Appearance:	August to November in eastern North America; October to November in the Pacific Northwest; December to February in California
Habitat:	Common under oaks and mixed hardwoods in eastern North America; under conifers in the West and Pacific Northwest
Tip:	Avoid older specimens; they are slightly bitter

Always remember

Rule number 1

Never, never take a mushroom with gills!!!

Rule number 2

Only take mushrooms with tubes, spines, and ridges and the mavericks portrayed in this book

Rule number 3

Only eat mushrooms which you have clearly identified with ALL the positive ID marks

Rule number 4

If a mushroom smells rotten, it is rotten and if it feels soggy, it is soggy

Rule number 5

Never, never eat wild mushrooms raw

Rule number 6

Look before you cut

Rule number 7

Mushrooms want to breathe. Do not suffocate them in plastic bags

Rule number 8

If in doubt, leave it

Mavericks

Common Puffball

Lycoperdon perlatum

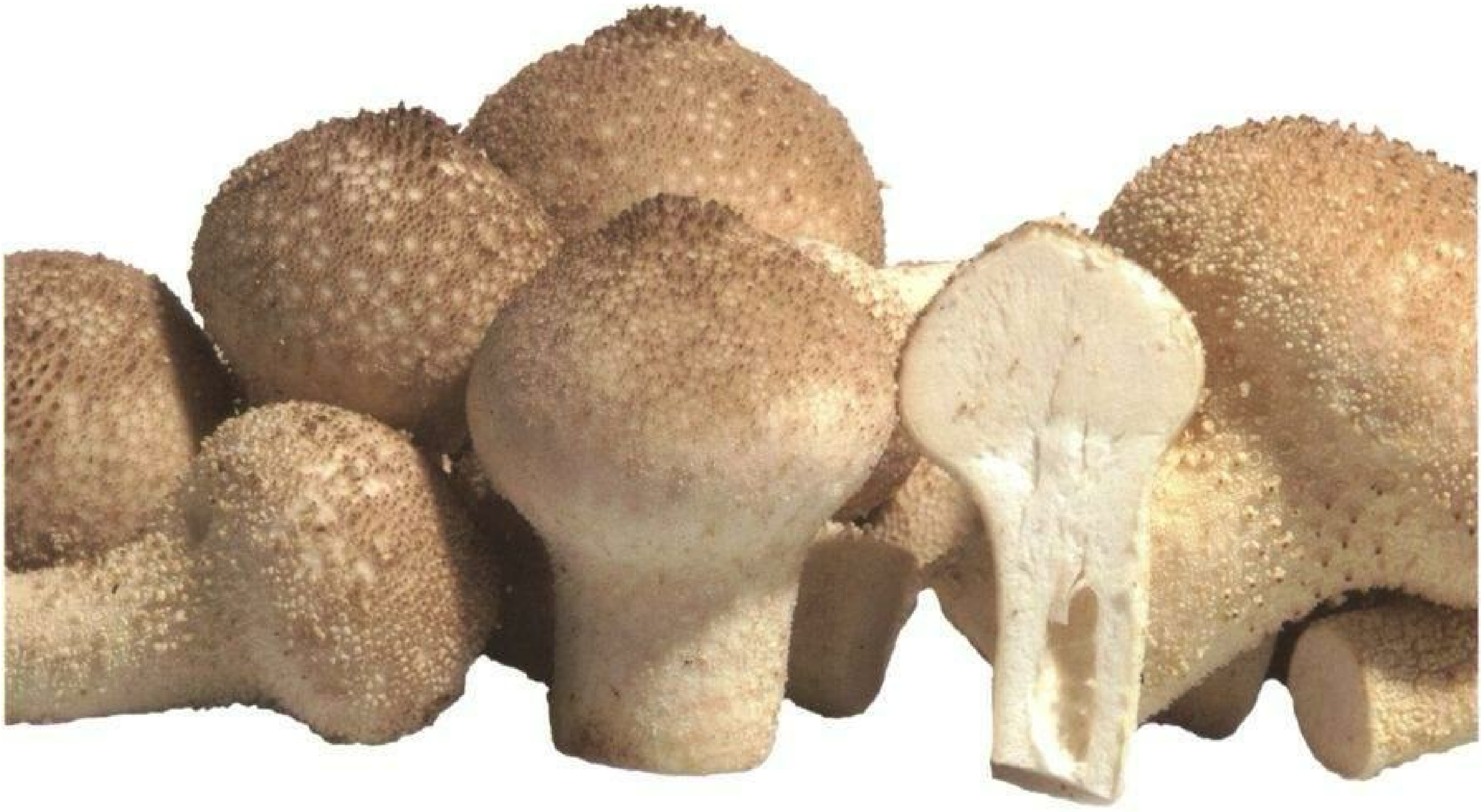

The common puffball is a curious little mushroom. One cluster of common puffballs might number merely four specimens; the next about forty. The common puffball is a welcome addition to any mix of mushrooms. Cooked on its own, it is not everyone's cup of tea because of its distinct taste. The giant puffball is another matter and as good a reason as you'll get for a dinner party.

Perlatum means "widely spread."

The conical spines are visible to the naked eye and leave an unmistakeable net-pattern when rubbed off. Only pick specimens on which the white conical spines are clearly visible. The common puffball has no unpleasant smell.

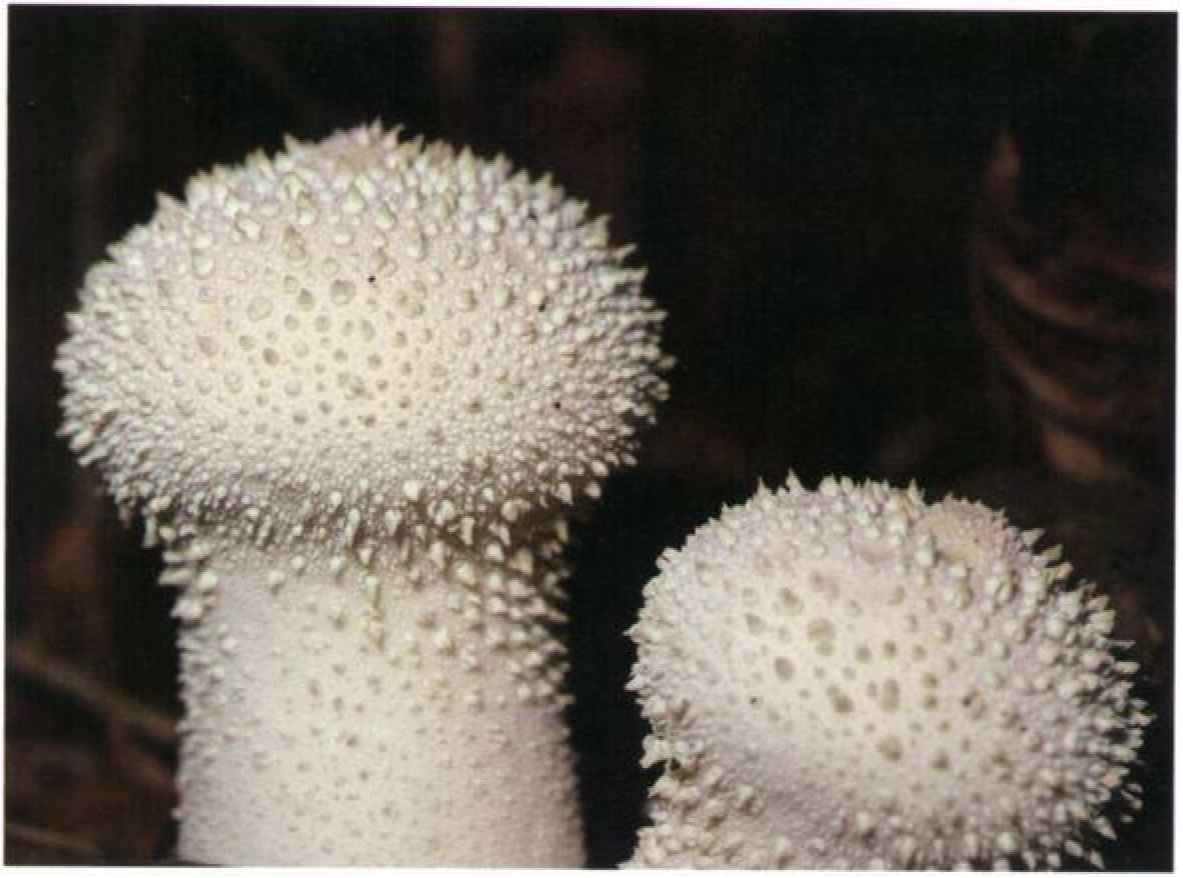

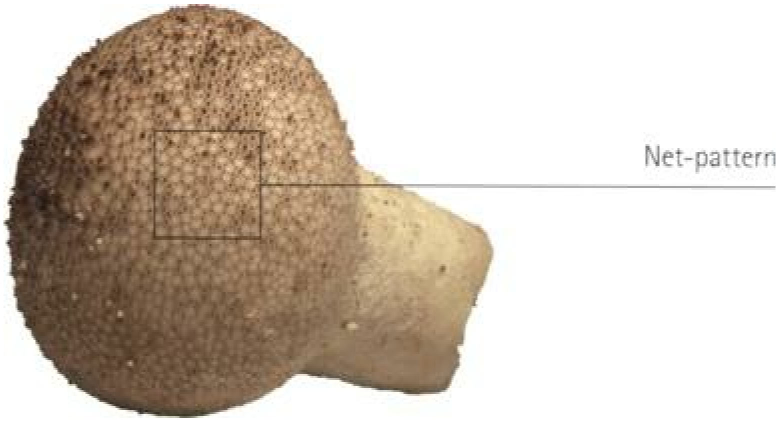

These common puffballs are the perfect size (height 1-2 inches) and condition. When cut, the inside is all white and firm. Only when the flesh is uniformly white and firm, is the mushroom all right for use.

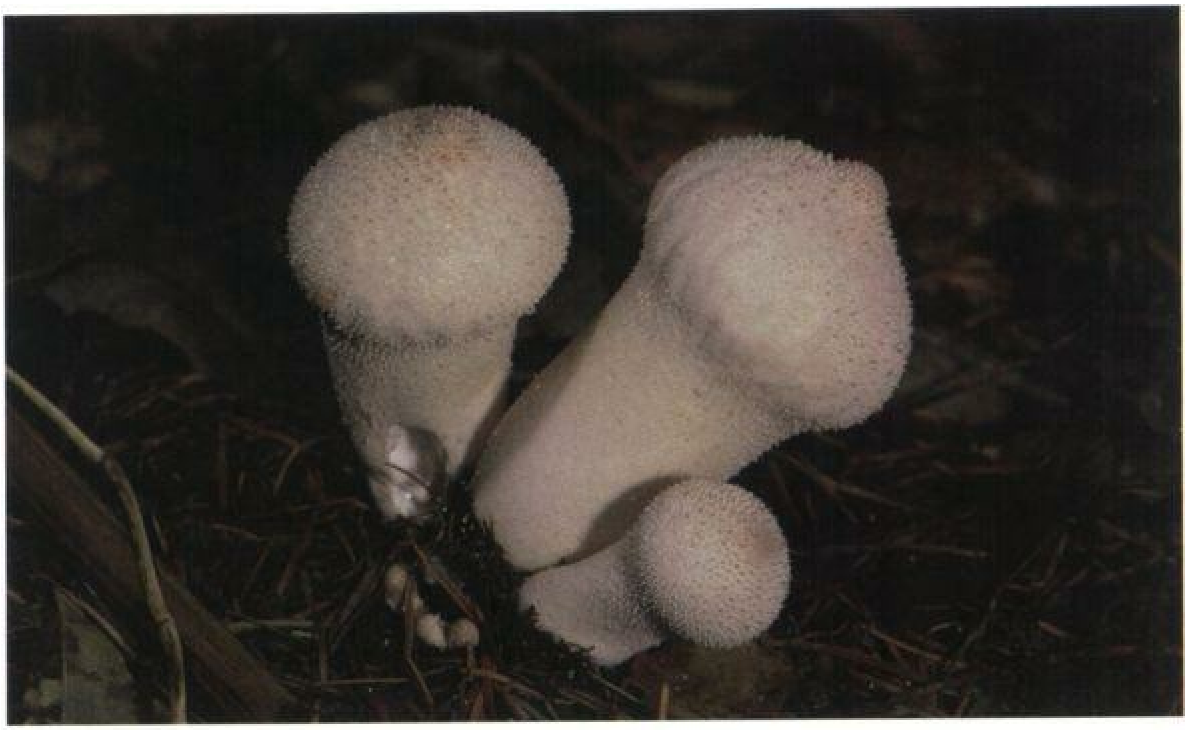

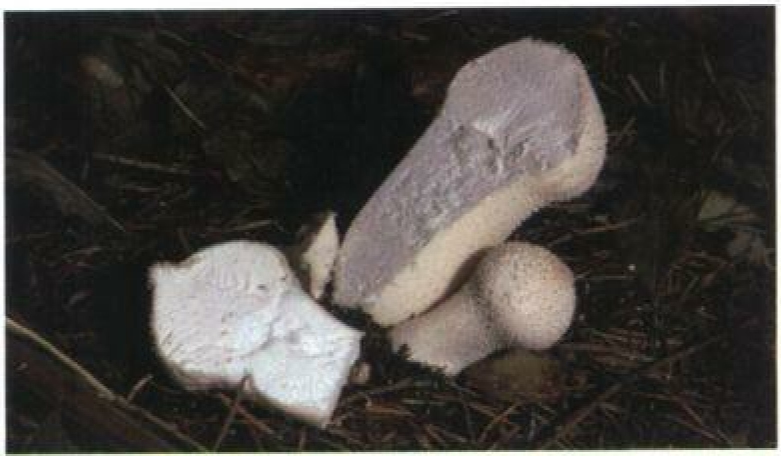

A variation of the common puffball is the pestle puffball. It too has spines but these are finer than those of the common puffball. It is pestle-shaped and it grows larger (height up to 8 inches) than the common puffball. It has no smell. If the flesh is uniformly white, and firm, it is good to eat.

The big brother of the common puffball is the giant puffball. Giant means it can grow to 3 feet in diameter. From a distance it looks all white but it can have a light yellow or yellow-brown tinge. The surface is smooth and feels rather like suede. The small specimen cut in two in the picture is about the size of a fist. When cut open, the flesh must be white and firm. Only then is the giant puffball edible-and delicious at that.

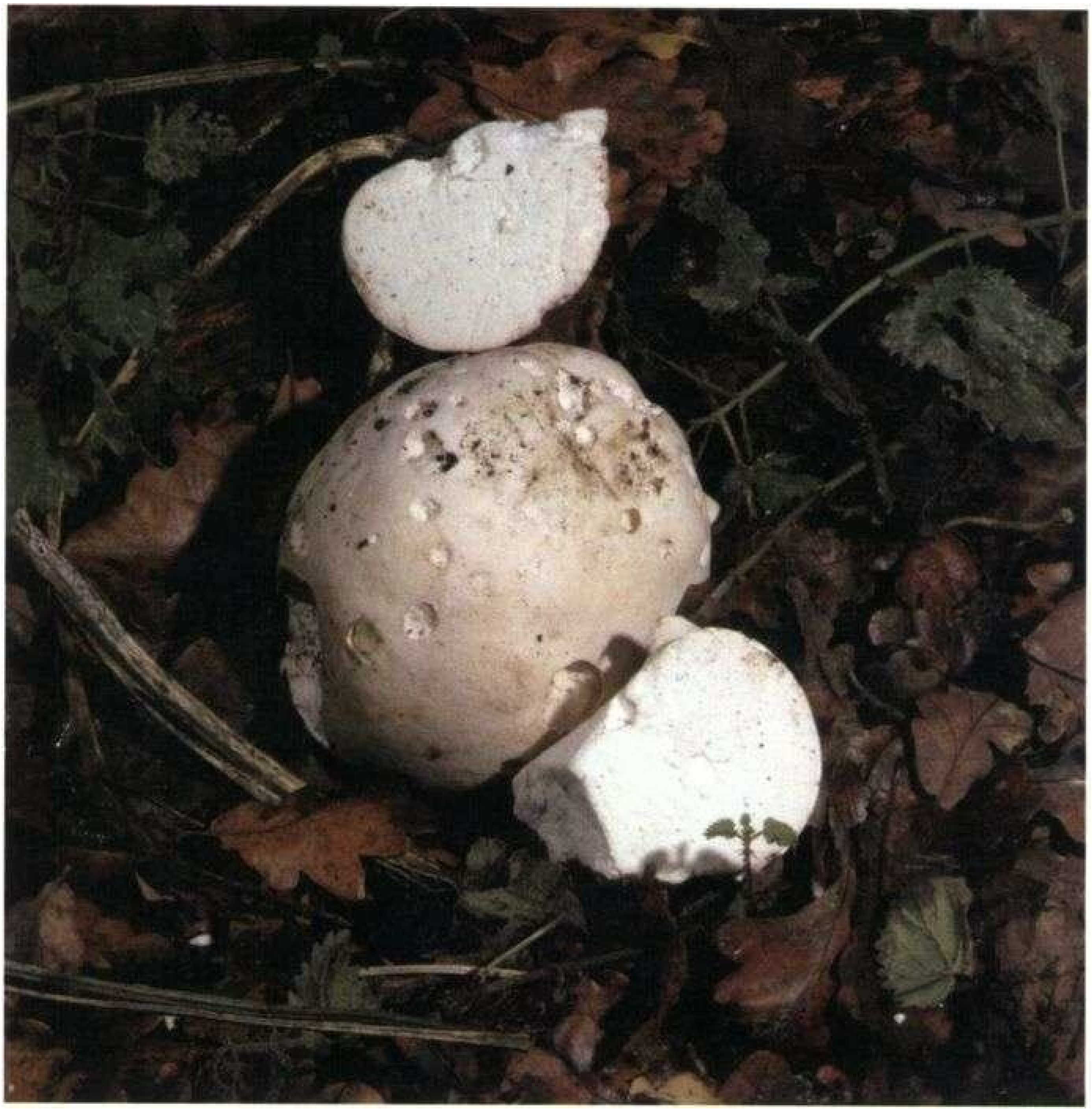

Positive ID Checklist

Common Puffball

- ☐ Conical spines
- ☐ Flesh must be firm and all-white
- ☐ Flesh undifferentiated, no thick rind or outline of mushroom visible in cross-section
- ☐ Net-patterned when spines are rubbed off
- ☐ Color Color white to off-white
- ☐ Found Found in groups
- ☐ No unpleasant smell

☐ Cap color matches cap color bar (above)

Be Careful:

Don’t mistake the Puffball for the Stinkhorn, which has a thick rind and black-and-white interior, or the Destroying Angel, which has gills and a differentiated interior. They are not edible.

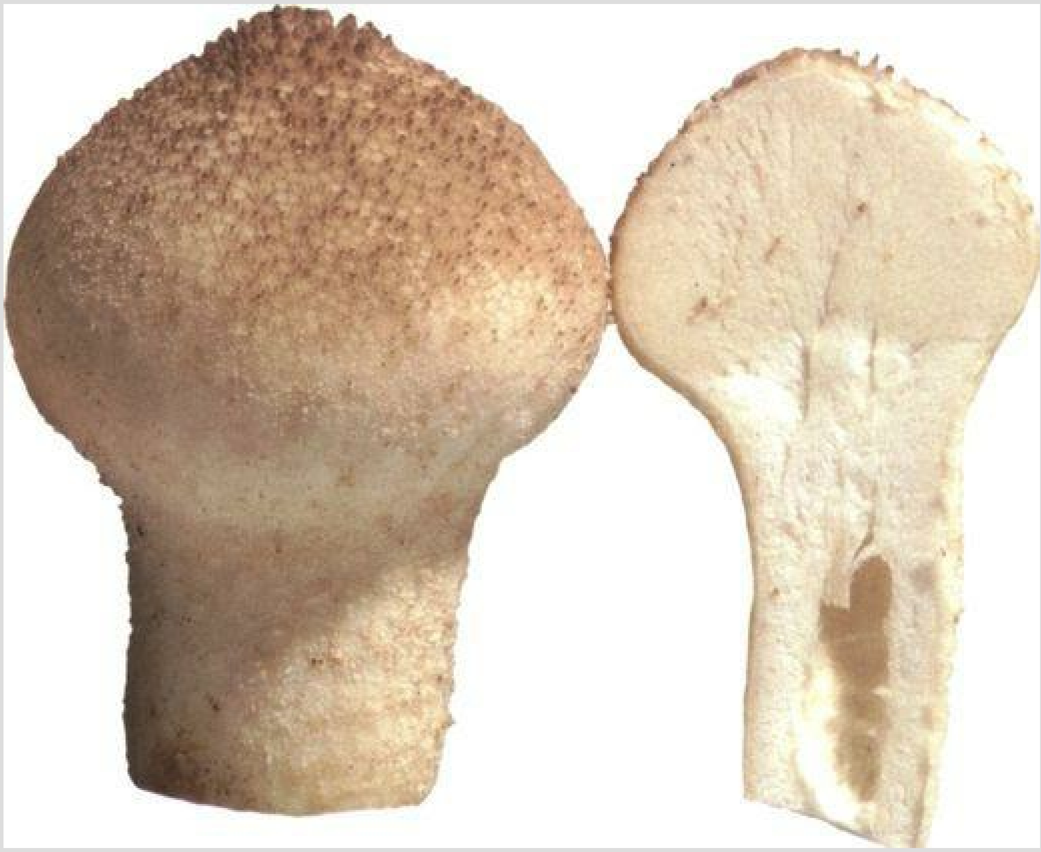

Avg. size across cap: Size of a golf ball. Giant puffball 12 inches

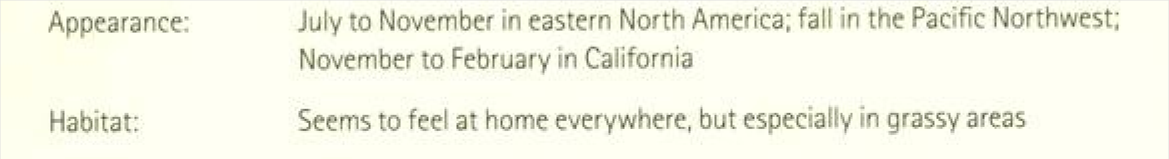

Appearance:	July to November in eastern North America; fall in the Pacific Northwest; November to February in California
Habitat:	Seems to feel at home everywhere, but especially in grassy areas

Horn of Plenty

Craterellus cornucoploides

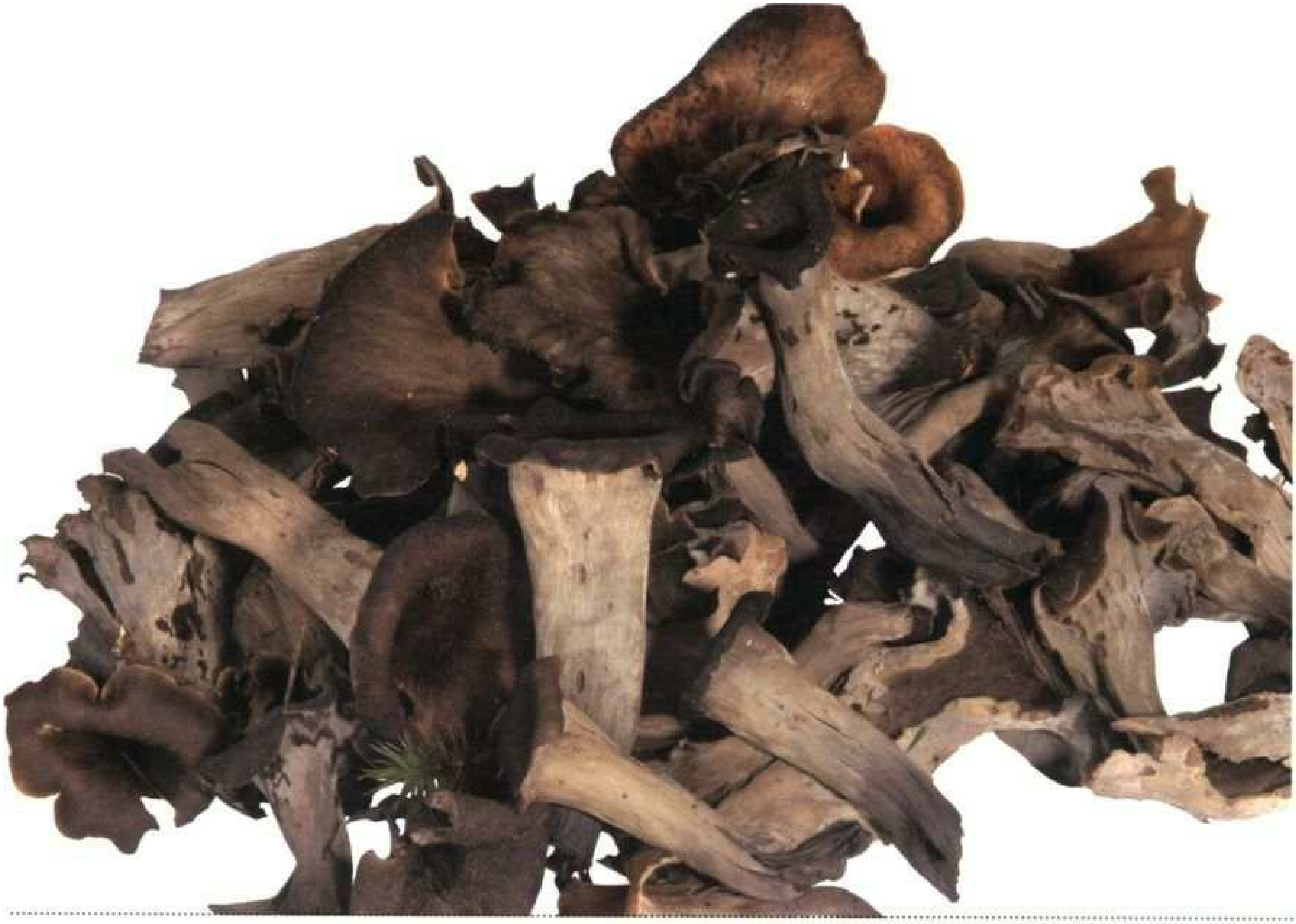

The horn of plenty is a real challenge to the mushroom forager.
Growing low on the forest-floor, it is so well-camouflaged that you can look at masses of them and not see any at all. If you spot one, there will be more: proceed with the utmost care because they are easily trampled. The horn of plenty is without a shadow of doubt one of the most delicious mushrooms in the world. But it has its price. Cleaning the horn of plenty can be quite a job but the reward is well worth the effort. Besides, when everybody joins the cleaning party, it is fun.

Cornucopioides refers to a cornucopia, or horn of plenty.

The horn of plenty is a funnel, the outer side of which looks smooth but in fact is slightly wrinkled. The color on the outer side is ash-gray, gray or pale gray with a bluish or lilac tint.

The inside of the funnel is brown, brown-gray, soot-gray or black. The surface is scaly or flaky and the top is out-rolled. In older specimens the top is wavy and split.

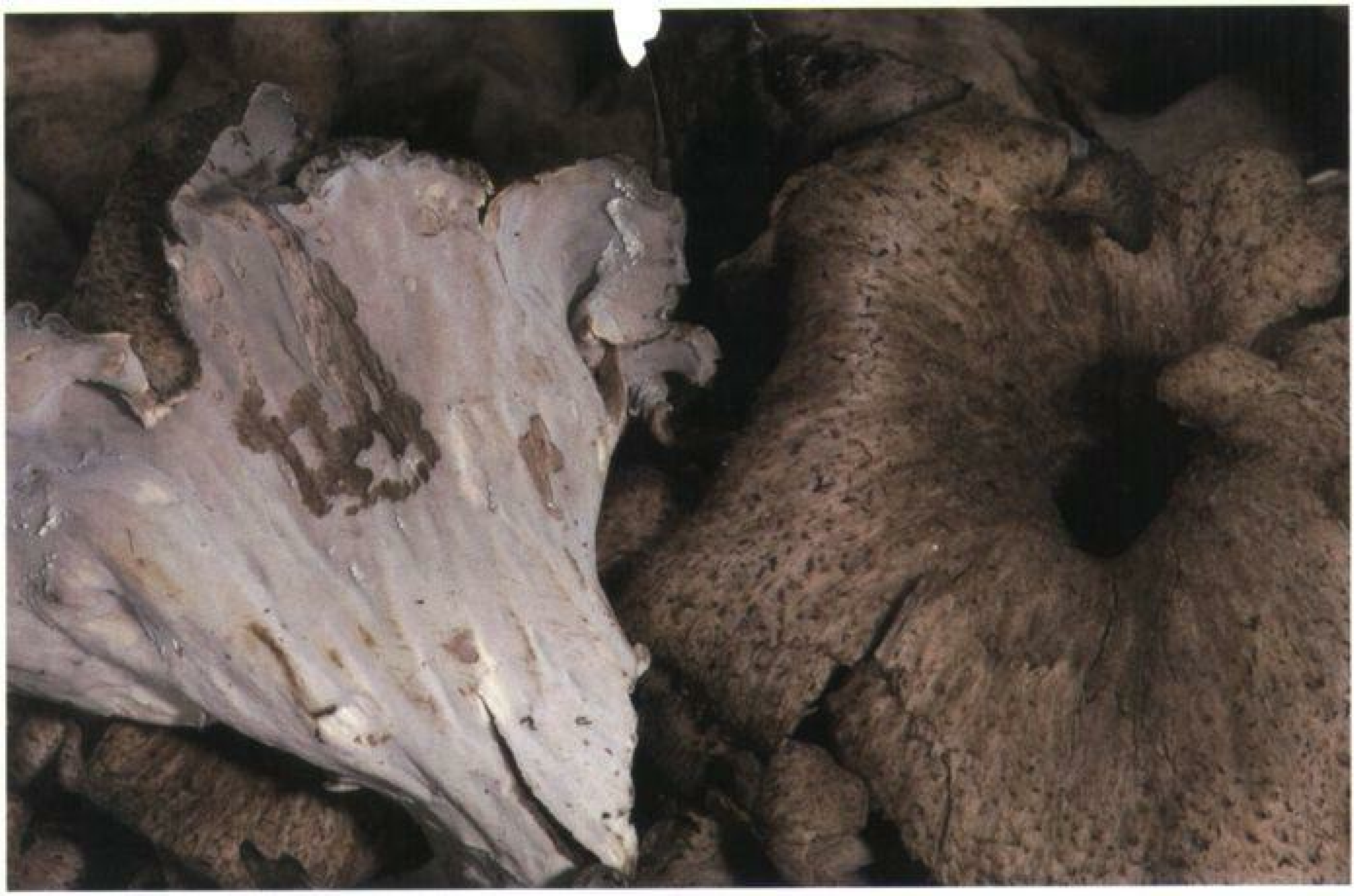

The horn of plenty seen from above in dry conditions.

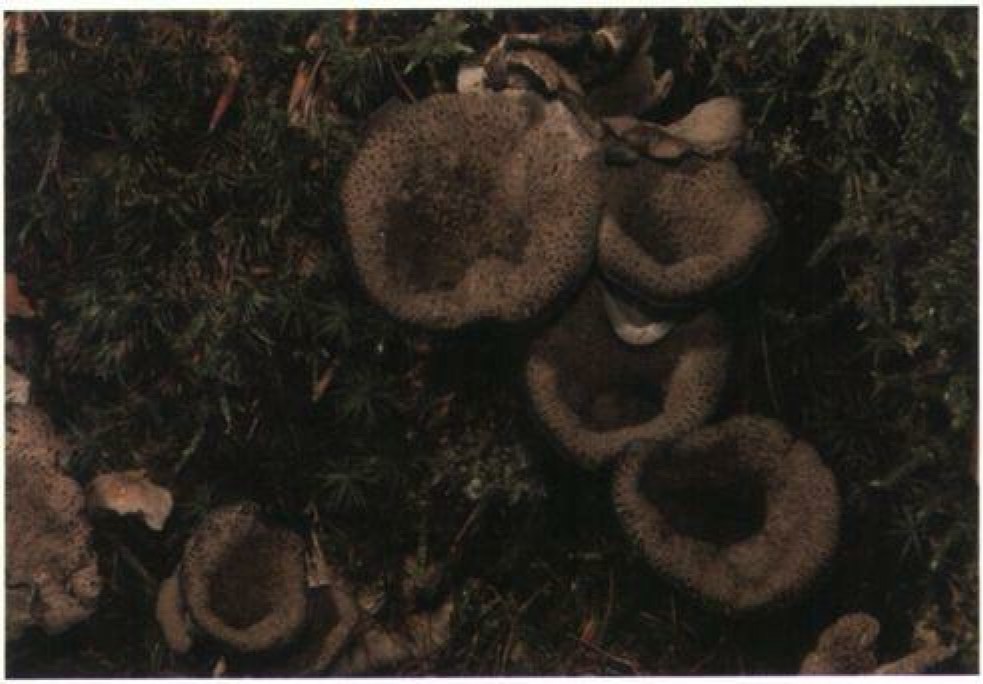

The horn of plenty seen from above in wet conditions.

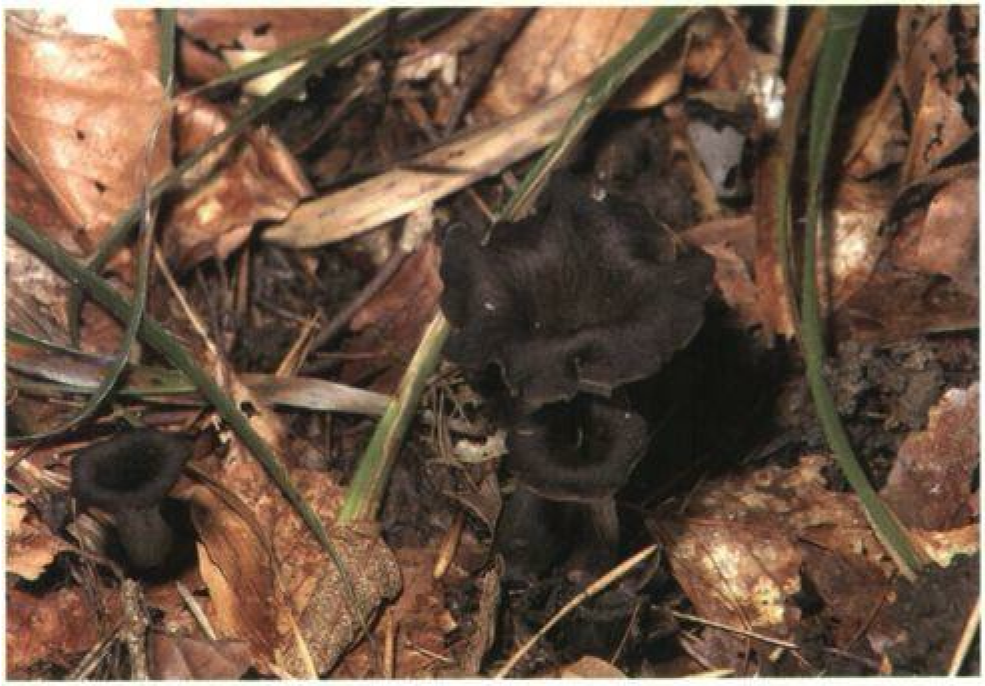

The funnel shape and the out-rolled top are typical.

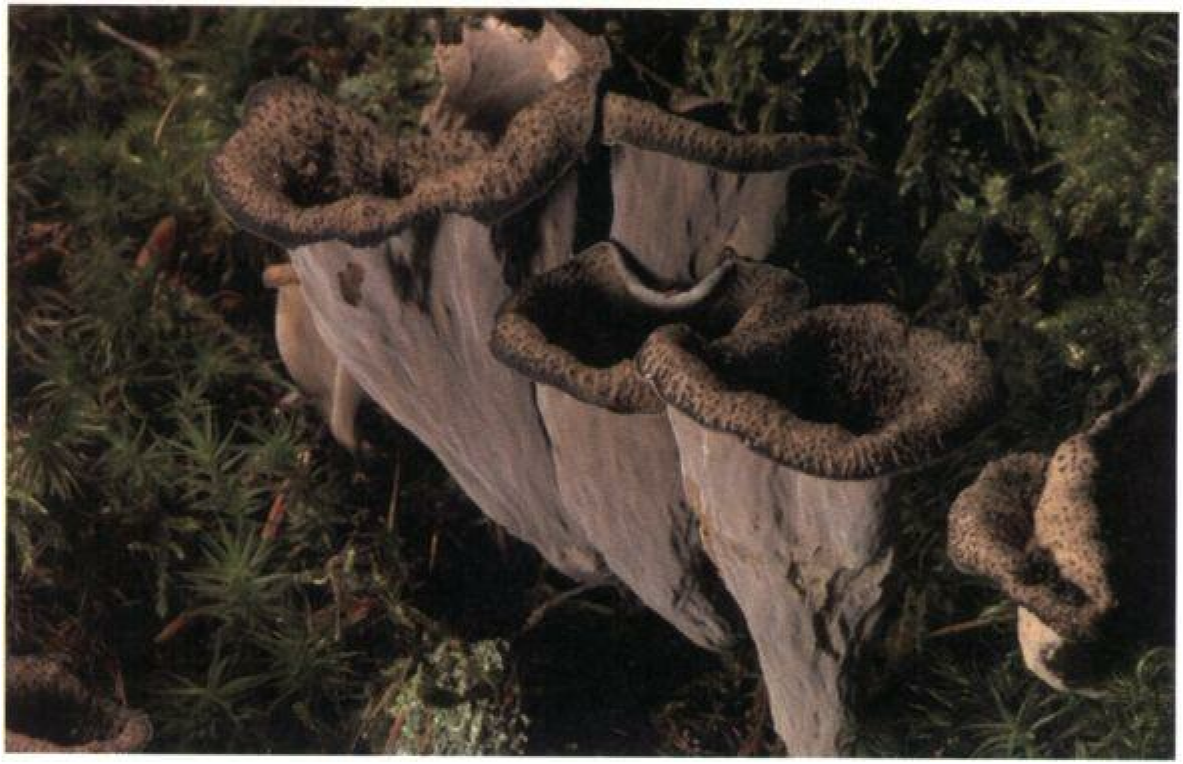

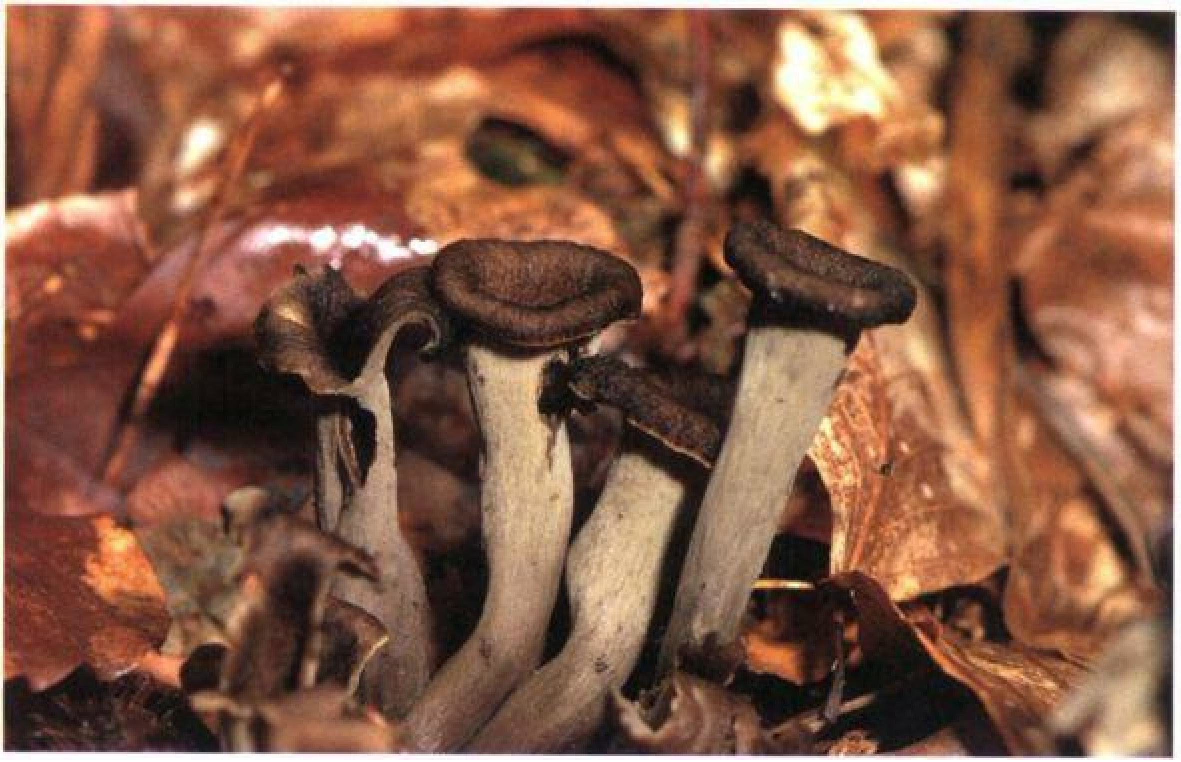

This looks a rather battered group of horn of plenty. Some of the tops have already split. Don't be put off-they still taste good.

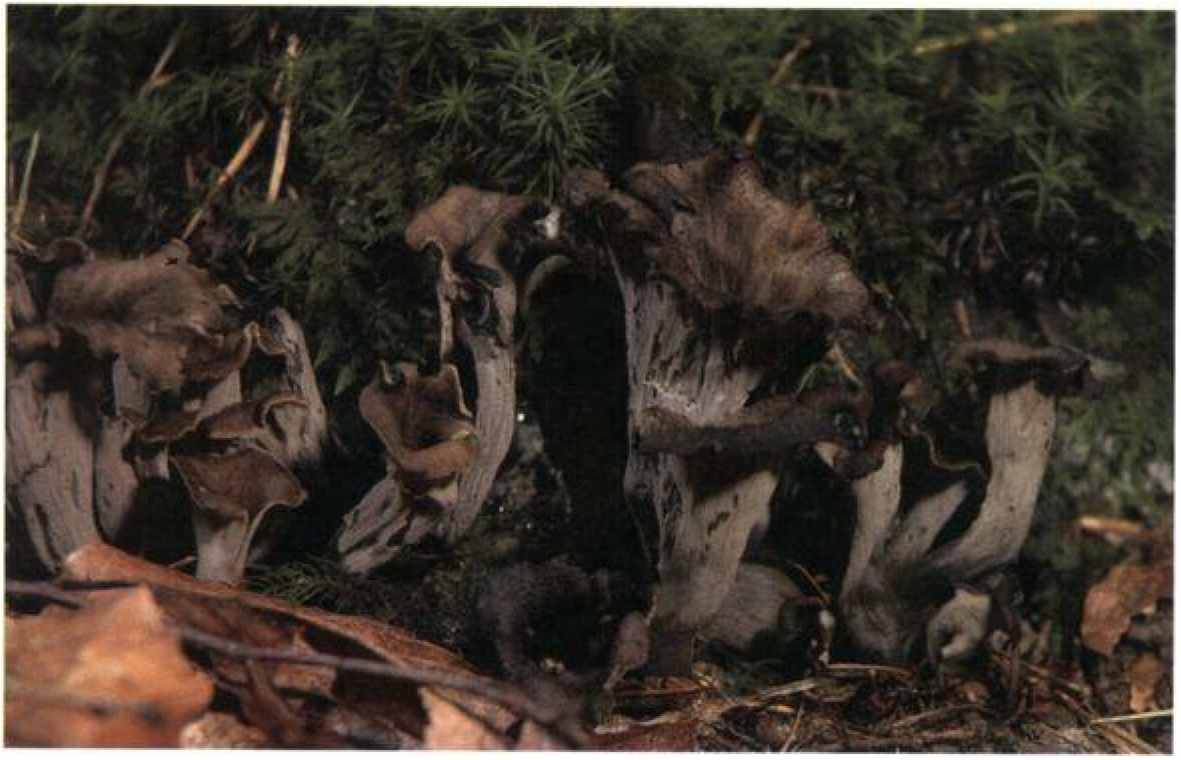

Positive ID Checklist

Horn of Plenty

- ☐ Found in groups
- ☐ Funnel shaped
- ☐ Smooth-looking but slightly wrinkled outer surface
- ☐ Scaly inside surface
- ☐ Outer surface matches color bar above (top line)
- ☐ Inside surface matches color bar above (bottom line)

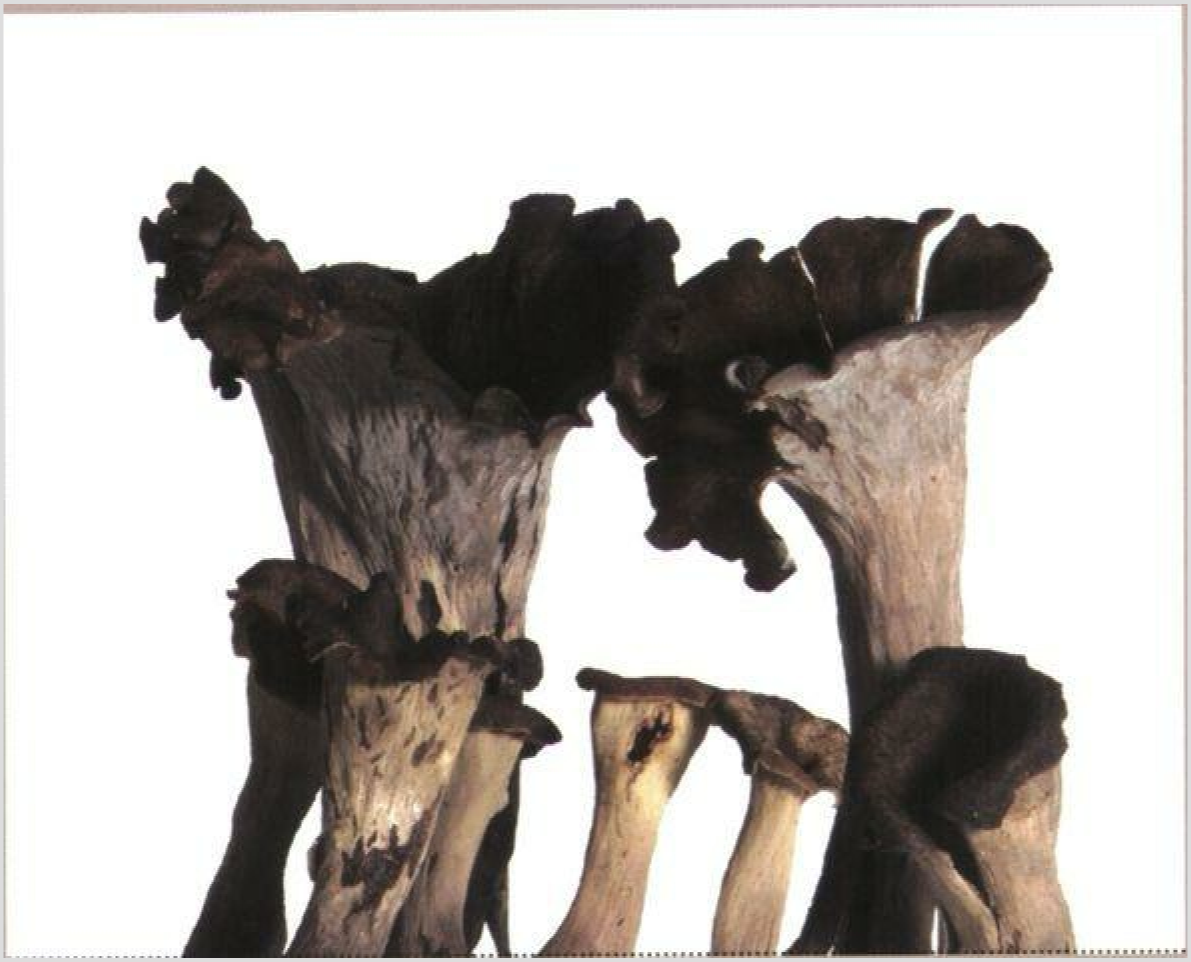

Avg. size across cap: 0.75-1 inch when mature

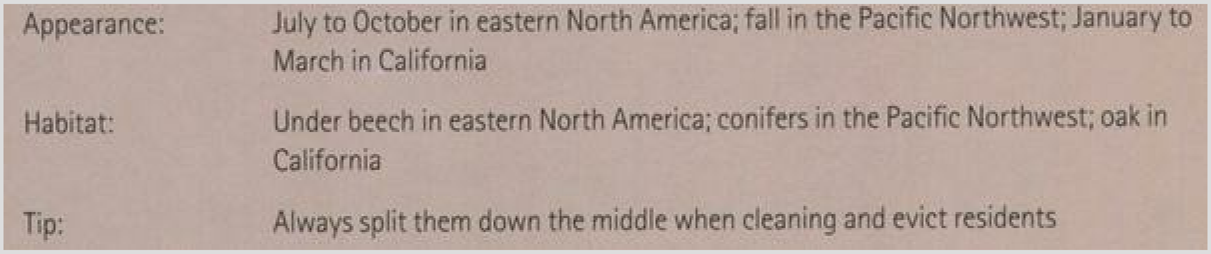

Appearance:	July to October in eastern North America; fall in the Pacific Northwest; January to March in California
Habitat:	Under beech in eastern North America; conifers in the Pacific Northwest; oak in California
Tip:	Always split them down the middle when cleaning and evict residents

Cauliflower Mushroom

Sparassis crispa

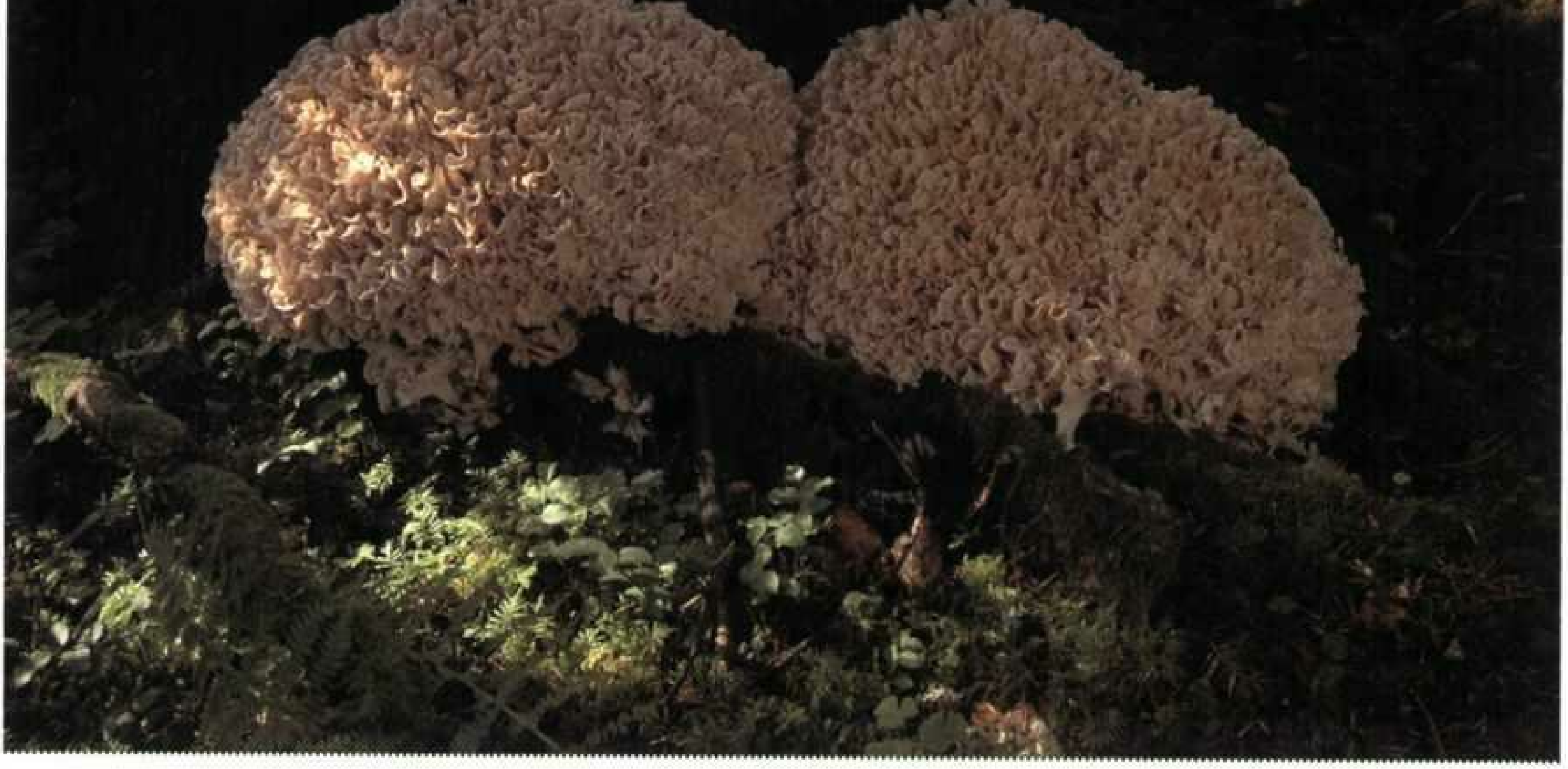

This is the mushroom of superlatives. For many people it is the very best of all. It may be a bit chewy but each bite releases a unique culinary sensation. If you fall for it, you'll be a dedicated cauliflower mushroom hunter forever. It resembles a cauliflower or a sponge.

Once you learn to recognise its spicy smell, you could identify the cauliflower mushroom blindfolded. And then there is the size: it can grow so big that people simply overlook it. This specimen weighed in at 22.5 1b.

Crispa means "frizzy."

Not all specimens grow to giant size. This one here is about the size of two fists. Its cauliflower or sponge-like appearance and the curved lobes are the key identification marks. There is nothing pointed or jagged in the cauliflower mushroom. The color ranges from creamy white to light brown. If it turns any browner than the specimen here, it has spoiled.

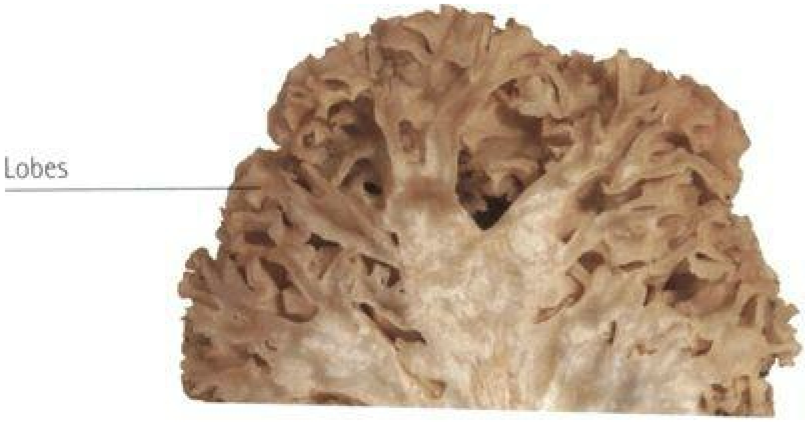

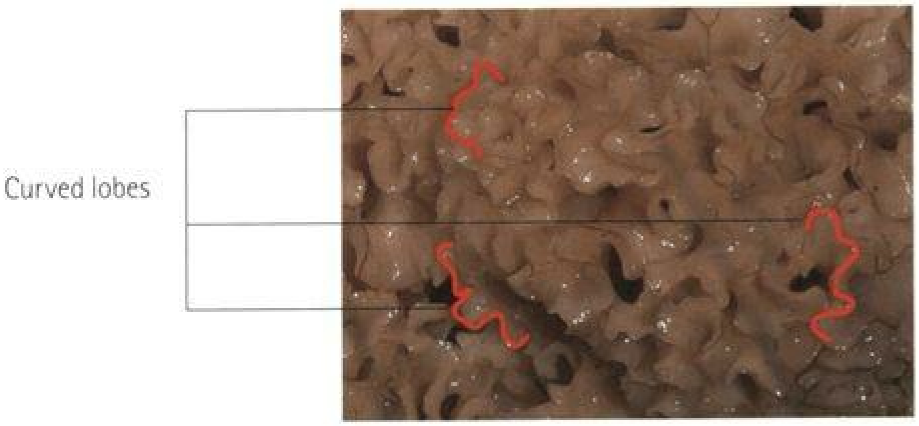
Curved lobes

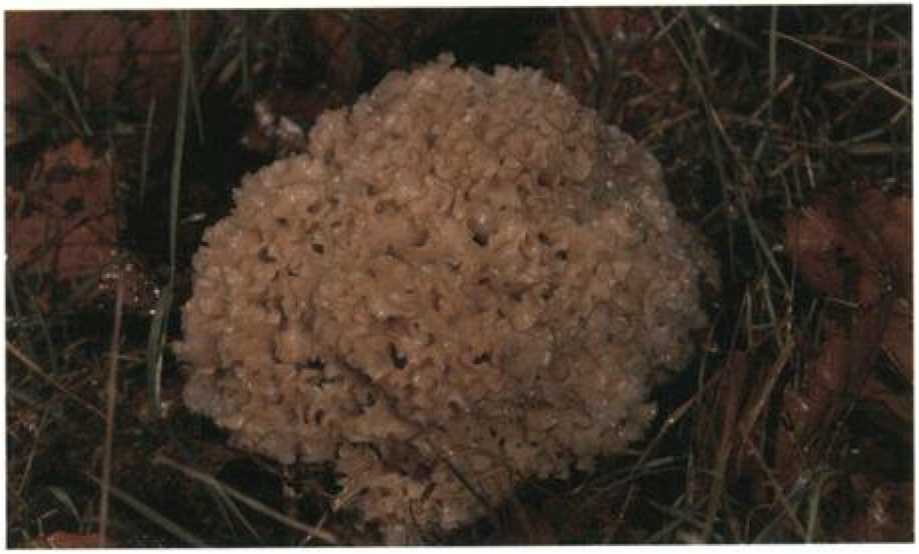

Positive ID Checklist

Cauliflower Mushroom

☐ No gills, pores, tubes or spines

☐ No stem
☐ Looks like a cauliflower or sponge
☐ Lobes
☐ Matches range of color bar (above)

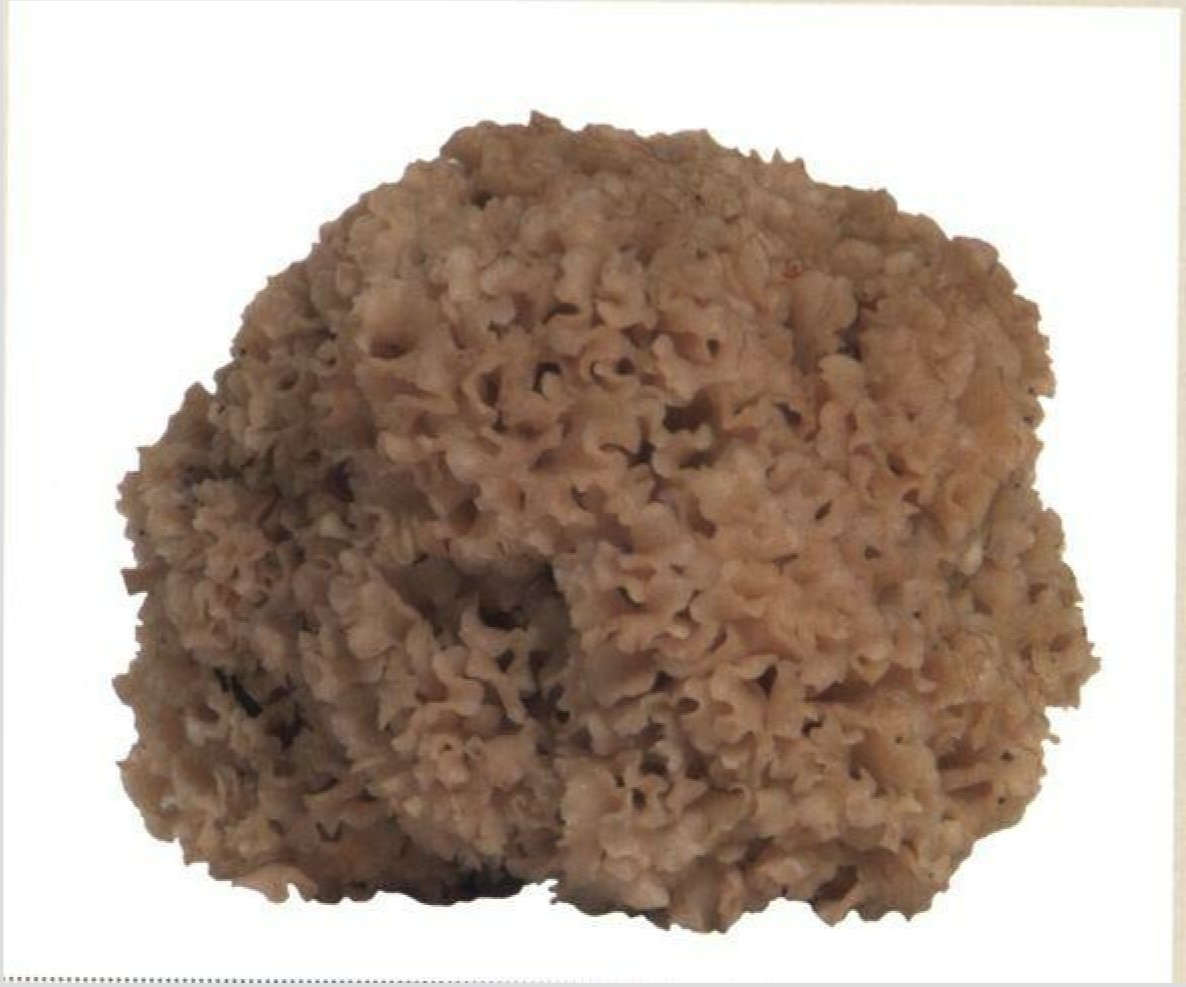

Avg. size across cap: 12 inches but can grow up to three feet

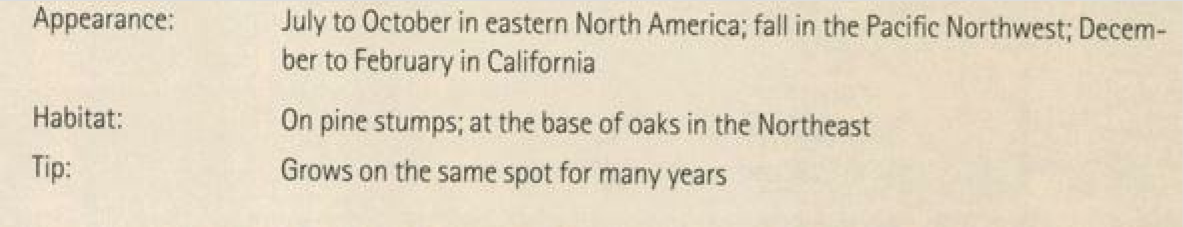

Appearance:	July to October in eastern North America; fall in the Pacific Northwest; December to February in California
Habitat:	On pine stumps; at the base of oaks in the Northeast
Tip:	Grows on the same spot for many years

Mushrooming Miscellany

The Mushrooming Regions

Eastern North America

Throughout this book, “Eastern North America” means regions primarily north of Georgia.

The Pacific Northwest

The Pacific Northwest consists primarily of Washington and Oregon (and parts of Idaho).

California

Throughout the book, “California” refers to the Bay Area, two hours north and south of San Francisco. Mushrooms also grow in Los Angeles and San Diego, but these appear toward the end of the California season.

As you can see, the timing and proliferation of mushrooms varies regionally. A good mushroom crop also depends heavily on weather conditions over the whole year. So the dedicated mushroom hunter need not confine himself to a mere couple of months of pleasure.

Trees and Mushrooms

Beech, birch, oak, pine, spruce, aspen and larch are the most important trees in mushroom-hunting, trees which are easily identifiable by their leaves or bark or both. It is most useful to know these five species of tree because if, for example, you explore a new wood, they will guide you to the mushrooms.

Beech
Especially: horn of plenty

Oak
Especially: chanterelle

Birch
Especially: birch bolete

Pine
Especially: cauliflower mushroom

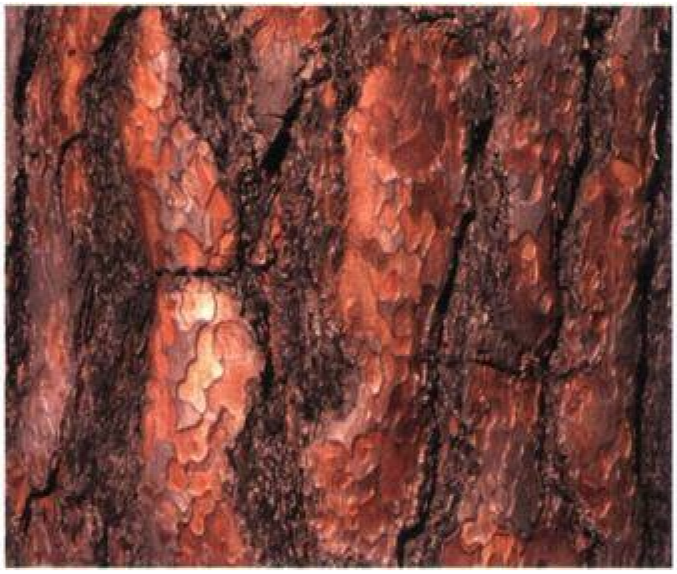

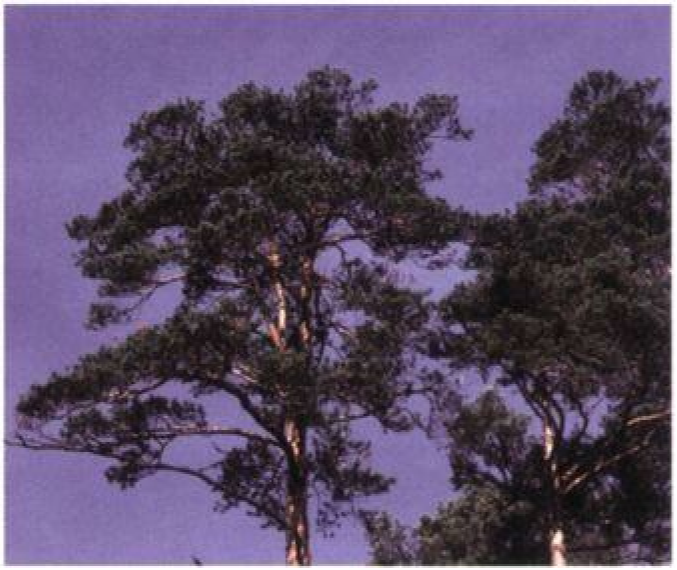

Larch
Especially: larch bolete

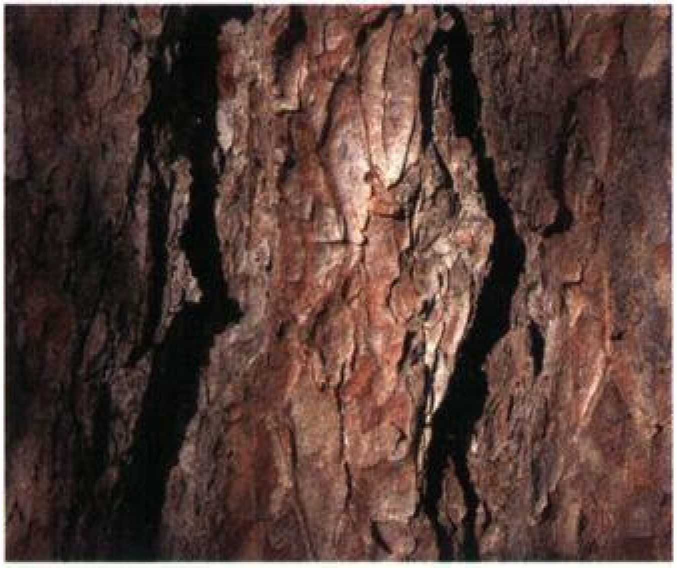

The larch is the only tree whose needles change color in autumn. You can’t miss it

Always remember

Rule number 1

Never, never take a mushroom with gills!!!

Rule number 2

Only take mushrooms with tubes, spines, and ridges and the mavericks portrayed in this book

Rule number 3

Only eat mushrooms which you have clearly identified with ALL the positive ID marks

Rule number 4

If a mushroom smells rotten, it is rotten and if it feels soggy, it is soggy

Rule number 5

Never, never eat wild mushrooms raw

Rule number 6

Look before you cut

Rule number 7

Mushrooms want to breathe. Do not suffocate them in plastic bags

Rule number 8

If in doubt, leave it

Handling, Storage and Cooking

Handling

The first rough cleaning of the mushroom should be done in the woods. Once you’re at home you verify your spoils against the positive ID check lists and then-and only then-can you start the fine-cleaning. Do not wet mushrooms to clean them. Brush or wipe them with a cloth. Part of the cleaning process is to check the quality. Evict residents, cut away soft and soggy tubes and cut the remaining mushroom into bite-size pieces.

Storage

Drying

Drying is the classic way to store mushrooms. This method actually intensifies the taste of the mushrooms and they'll keep for years if you follow the procedure correctly. Do not dry anything you wouldn't eat fresh. Drying does not improve the overall quality of your mushrooms. There are various methods of drying mushrooms but there is only one way which guarantees consistent high-quality results and that is the only one I recommend: a dehydrator. Whatever the make, the principle is the same: warm air circulation. The sliced mushrooms are placed on trays and then dried in the dehydrator. Dried mushrooms are best stored in any air-tight container or a sealed plastic bag and then kept in a dark place. Mix all the species except the cauliflower mushroom. The more species in the mix, the better the taste. The cauliflower mushroom should be dried and stored separately (see page 115). Do not dry chanterelles. All other mushrooms in this book are suitable for drying. Dried mushrooms should be crackle-dry and snap when broken.

Freezing

Freezing is the other method. Results improved enormously when home vacuum-packaging machines became available. I recommend only freezing vacuum-packed mushrooms. Frozen mushrooms retain their color, texture and taste.

Never thaw frozen mushrooms!

Put deep frozen mushrooms in sizzling butter or hot water. Like this they keep their texture.

- Mix all the species except the cauliflower mushroom. Again: the more species in a mix, the better.
- I recommend deep freezing some cep, and horn of plenty, separately. You might want to add just a few of them for a particular dish or use the cep for a tasty starter.

The cauliflower mushroom should always be frozen separately (see page 115).

The chanterelle must be cooked before deep freezing - all other mushrooms in this book can be frozen raw.

Drying or freezing?

Either way, use only perfectly fresh mushrooms. Drying mushrooms gives them that deep, strong mushroom taste. The frozen mushroom is more subtle, but the colors of frozen mushrooms are as bright as on the day they were picked and this adds considerably to the pleasure of eating. Dried and frozen mushrooms can be mixed to get the best of both worlds.

Cooking

Famous chefs win awards for stylish cuisine. Simplicity as a rule is not in their repertoire because no stars can be won by doing as little as possible. There are some really attractive mushroom books with some really attractive recipes but my advice is: do not try to emulate the recipes of the famous chefs - you'll never get there anyway. The place to enjoy grand cuisine is in the restaurant.

Do what you can do better than the professionals and do what they can't do and you're in for a feast. Besides, it's much less trouble, expense, and frustration to do it your own way rather than to follow recipes which pre-suppose the infrastructure of a five star establishment.

So where can't the professionals compete with you?

1. Freshness

This is not to say that restaurants use old mushrooms. They're just not as fresh as yours.

Chervil is a tasty and attractive herb for mushroom dishes.

Your mushrooms have the shorter way from the wood to your kitchen. As a rule the restaurants' mushrooms come from a market and some of those mushrooms on the market have traveled a long way, some of them half way round the world.

2. Sourcing

You know where your mushrooms came from, whereas the chef in the restaurant, as a rule, knows only that it is the type of mushroom he wants. You know the very wood from which you picked them, and this is important because, for example, a cep can taste differently from one wood to another.

3. Quality

You know exactly where your mushrooms come from, and therefore you know their quality. In other words, you can be sure, for example, that your mushrooms grew in a non-polluted habitat or a dog-free zone.

4. Freedom

You are at liberty to vary the mushrooms that go into dish and to change ingredients. In top restaurants, recipes are more precise and designed to he produced time and again. Your mushroom feas the other hand, can be unique each time.

5. Soul

There is a satisfaction in picking your own wild mushrooms and cooking them yourself that not an match.

Garnishing with finely chopped vegetables makes your mushroom dish a feast for the eyes.

Mushrooms are not garnish

The mushrooms in this book are of culinary value in their own right and should be treated accordingly. Treat a wild mushroom with the same culinary respect as you would a wild salmon.

Olive oil and mushrooms

Most contemporary mushroom cookbooks recommend olive oil for cooking mushrooms, but if you really want to draw out all the subtle flavors of your mushrooms and realize their full culinary potential, then olive oil is, generally, not a good idea. Unsalted butter brings out the best in your mushrooms.

Chives are the classic garnish for mushrooms.

Kills all known germs

There is a potential but very low risk of catching something nasty from a mushroom. To reduce that risk to zero, do not eat raw mushrooms and in the cooking process, increase the heat at one stage so that the dish either boils for a couple of seconds or sizzles in the butter. For that brief moment, turn the mushrooms in the pan so that they're exposed to the heat on all sides.

Curly or flat-leafed parsley underline the mushroom taste and add aesthetic appeal. Sprinkle on raw just before serving.

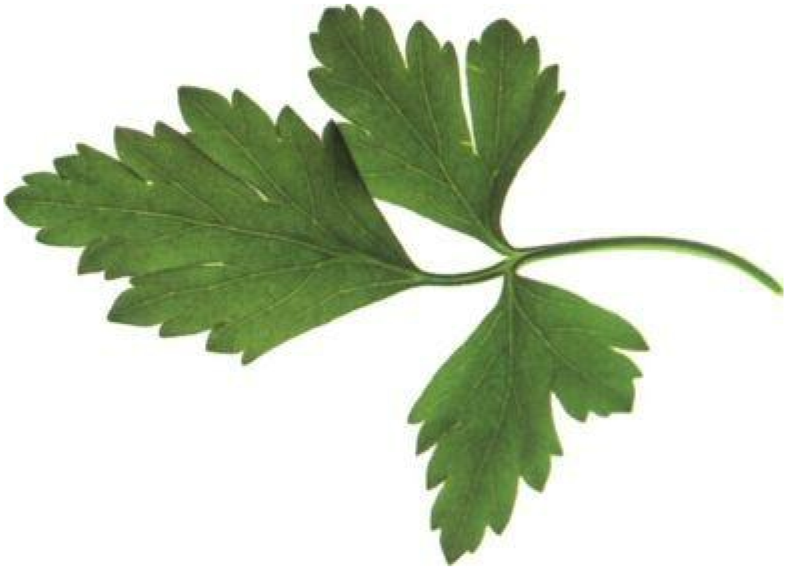

Special Cases

Chanterelle

The chanterelle must not be dried because it goes chewy. It must not be frozen without prior cooking because it will turn bitter.

Horn of Plenty

Some people like it on its own. Others use it only as a spice to add to a mix of mushrooms. In order to find out what you like best, dry and freeze separately.

Cauliflower mushroom

The cauliflower mushroom is also very distinct in taste and perhaps best on its own. Dry or deep freeze separately.

Basic preparation for fresh or frozen mushrooms

There's a lot of flexibility in these basic preparations. The amounts stated can vary considerably and you'll still get an excellent result. And the beauty of it is that you can't overdo mushrooms. The key to culinary success is the quality of the mushrooms.

Serves 4

What you need
Essential:

5-8 cups fresh or frozen mushrooms
4 cups water
1 cup white wine
1 teaspoon salt
1 shallot or small onion, chopped
1 cup cream
¼ cup chives or parsley

Optional:

A mix of finely chopped vegetables, such as carrot, zucchini, and celery: approx. ½ cup
¼ cup chervil

What you do

1. Put water, white wine, salt, and shallot into a pan
2. Do not thaw frozen mushrooms. Add frozen or fresh mushrooms and bring to the boil
3. Reduce heat and simmer until approx ⅔ of liquid has evaporated
4. Add cream
5. Simmer again until approx. half the liquid has evaporated. Stir occasionally. If you feel there's too much liquid, simmer until it reduces still further. The less liquid, the more intense the taste and vice versa. Add water or cream to taste.
6. Add more salt and pepper to taste
7. You now have a wonderful mushroom sauce which you can serve with meat (chops, steak, etc.) or on its own with pasta or rice. Sprinkle with chives, parsley or chervil and the finely chopped vegetables

Basic preparation of dried mushrooms

Serves 4

What you need
Essential:

1-2 ounces dried mushrooms
4 cups lukewarm water
1 shallot or small onion, chopped
¾ cup dry white wine
1 teaspoon salt
¾ cup cream
¼ cup chives or parsley

Optional:

A mix of finely chopped vegetables e.g. carrot, zucchini, and celery: approx. ½ cup
¼ cup chervil

What you do

1. Take the dried mushrooms and put in a bowl or pitcher. Add 4 cups of lukewarm water. Soak for 1½ hours (until the mushrooms float in brown liquid)
2. Put wine, shallot, salt, mushrooms and the brown liquid into a pan
3. Bring to the boil for 10 seconds Reduce heat and simmer until ⅔ of the liquid has evaporated
4. Add the cream

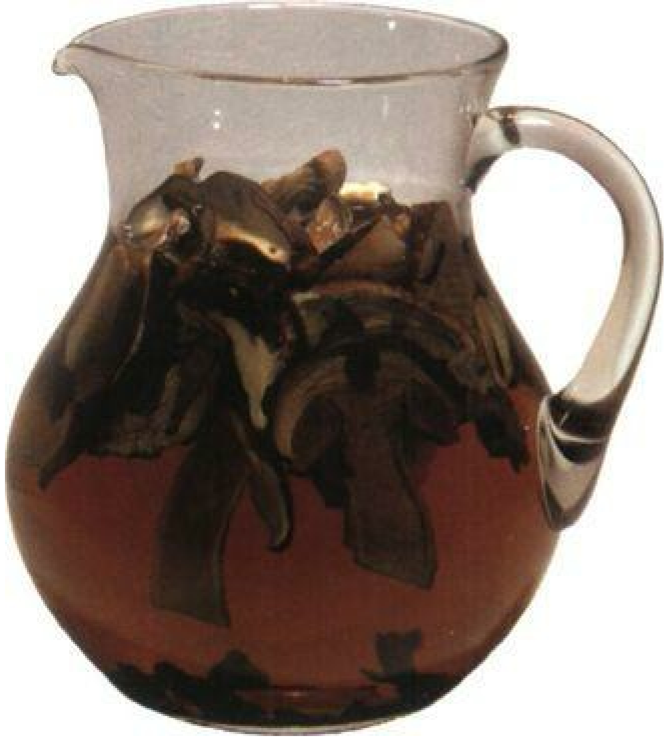

5. Simmer until approx. half the liquid has evaporated. Stir occasionally. If there is too much liquid, simmer until further reduced. The less liquid the more intense the flavor, and vice versa Add water or cream to taste
6. Add salt to taste
7. You now have a wonderful mushroom sauce which you can serve with meat (chops, steak, etc.) or on its own with pasta or rice. Sprinkle with chives, parsley, or chervil and the finely chopped vegetables

Beyond the Stove

No summer or winter woodland walk will ever be the same again once you become aware of the amazing universe of the forest floor, which in terms of color and life is a match for any coral reef. And the forest floor is right at your feet, so to speak, or at least not far from your door.

When in autumn you walk through a beech wood and marvel at the intensity and variety of the colors, think for a moment about those leaves.

There are about 2 tons of leaves to one acre of beech wood. After their job is done on the branches, they fall to the ground. If they did not decompose we would soon be swimming through a sea of leaves.

Likewise, unrotted fallen branches would in time pile up as insurmountable obstacles.

A leaf falls to the ground, inspiring on its way a poet who happens to be passing by: but on a more practical level, bacteria are the first organisms to get busy.

Shortly afterwards, it is fungi, algae and protozoans which play their part in the work of decomposition.

Soon, millions of organisms and little animals are at work on that single leaf. In due course the leaf is fully decomposed, its constituent parts recycled back into the forest soil, keeping it fertile by providing the essential substances for all the plants in the wood.

The vital role of mushrooms

Fungi are key players in maintaining tree growth and health. The reason the birch bolete is always found near birch trees, and the larch bolete always near larch trees, is that these pairs live in a symbiotic relationship. They need each other.

The mycelium of the birch bolete and the fine root-ends of the birch tree link up and exchange substances. There is a lot of chemistry going on, with a net result of mutual benefit. There are intricacies and complexities in all matters fungi and there are, so to speak, vast uncharted areas on the fungal map. However, even the most half-hearted mushroom-hunter will soon become aware of the all-encompassing interrelations in the wood. Mushrooming makes you see and appreciate the woodland differently.

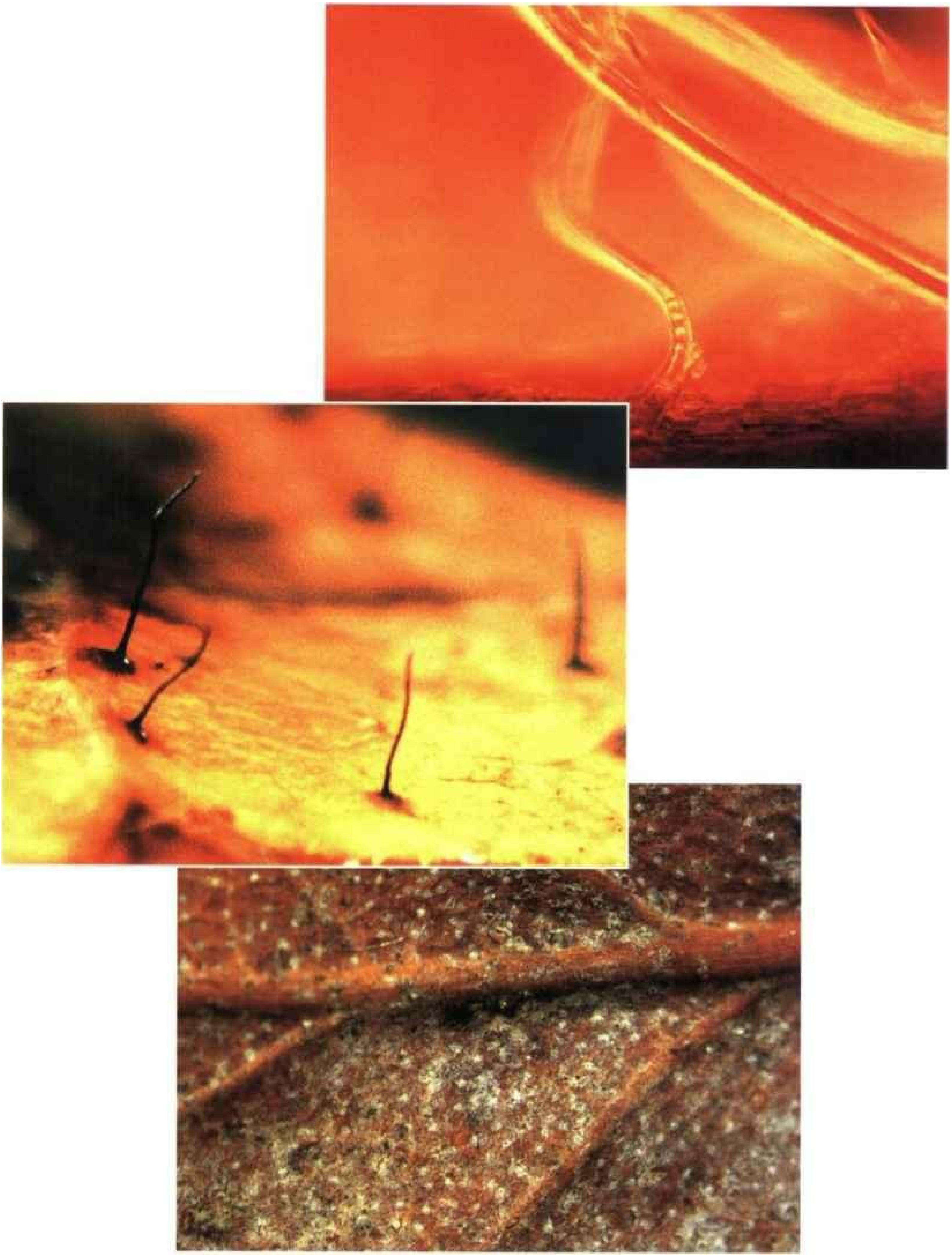

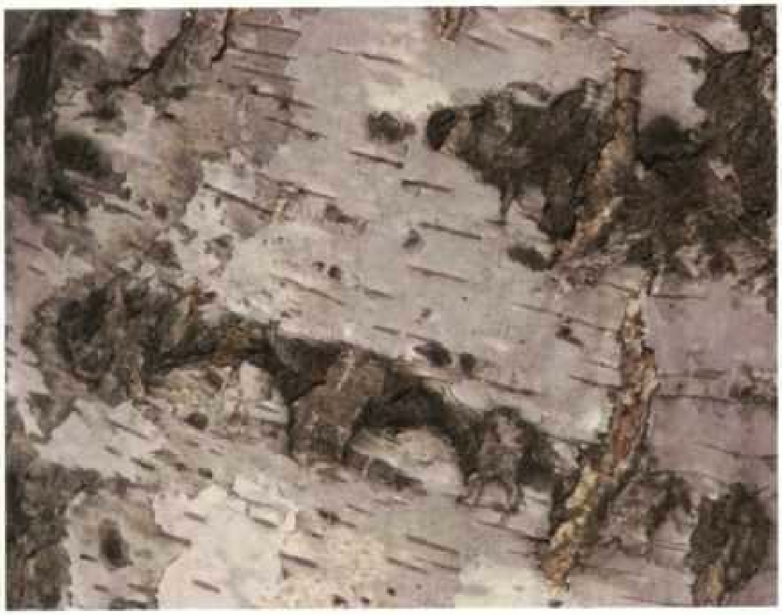

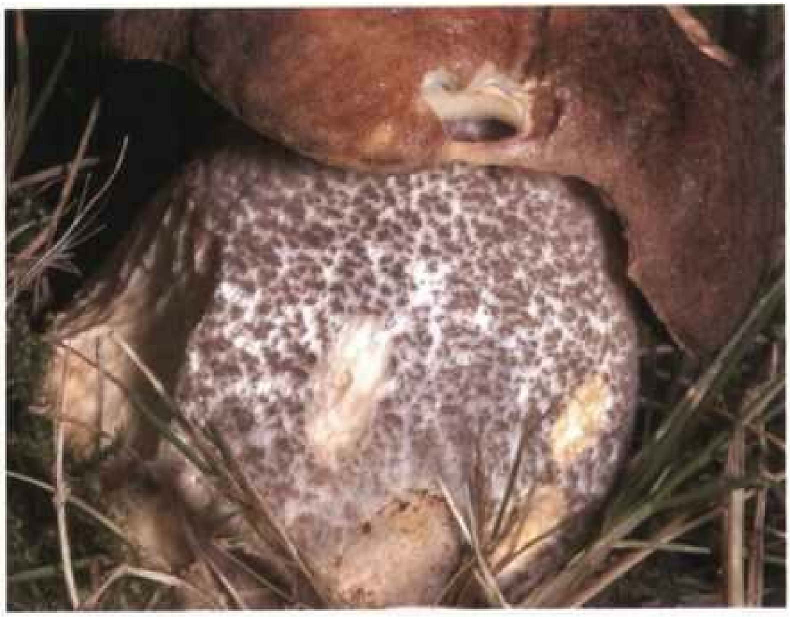

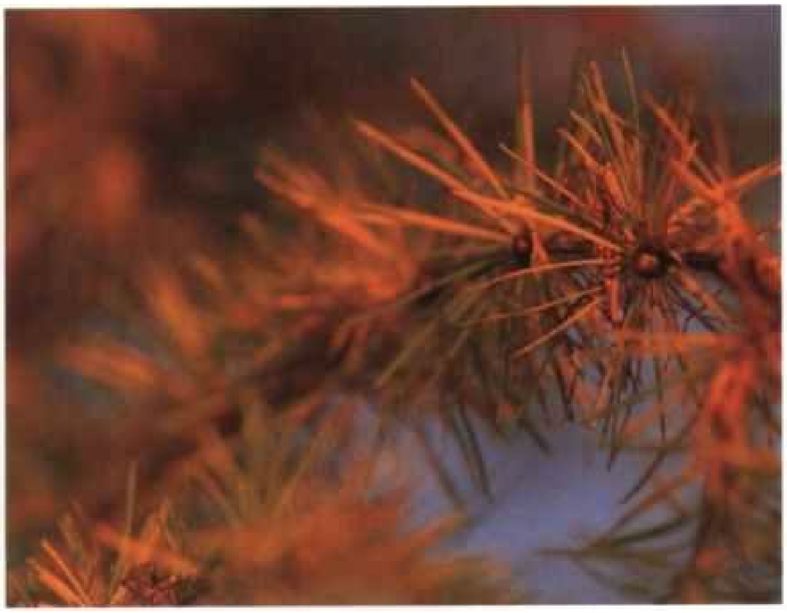

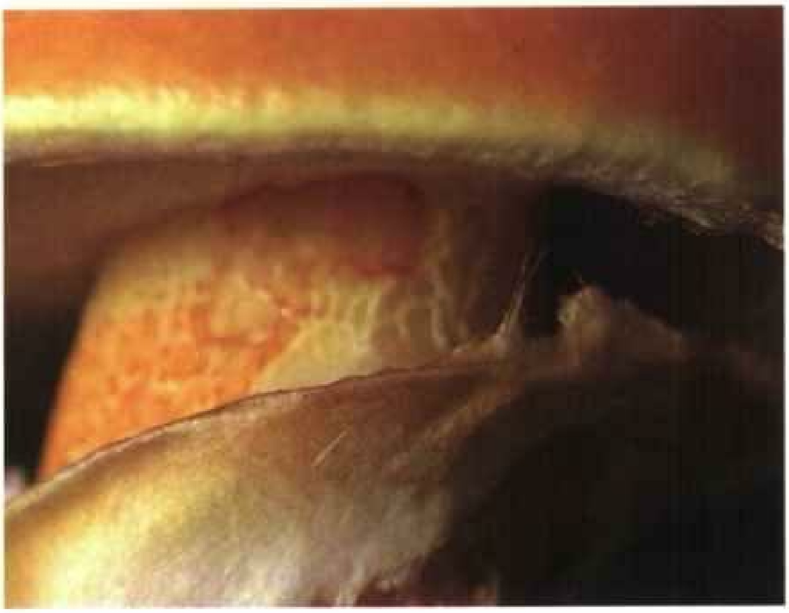

There is a practical point to these impressions which will help you in your mushrooming.

The similarity of patterning between the trunk of the birch tree and the stalk of the birch bolete is there for all to see. The similarity in color between the larch needles (once they change color in autumn) and the larch bolete seems obvious. These pointers can all act as nature's clues and reminders for the mushroom forager. But why is this so? Whatever the scientific answer, if indeed there is one, this similarity illustrates beautifully the interconnected nature or, if you prefer it, the mutual dependence, of life in the wood.

Perhaps even more intriguing is the similarity between the pine bark and the cauliflower mushroom, which grows near pine trees. It could be a coincidence or it could be a complex pattern at work. Either way, the cauliflower mushroom is utterly delicious - but if you mull over a conundrum like this, doesn’t it add interest and pleasure to mushroom hunting?

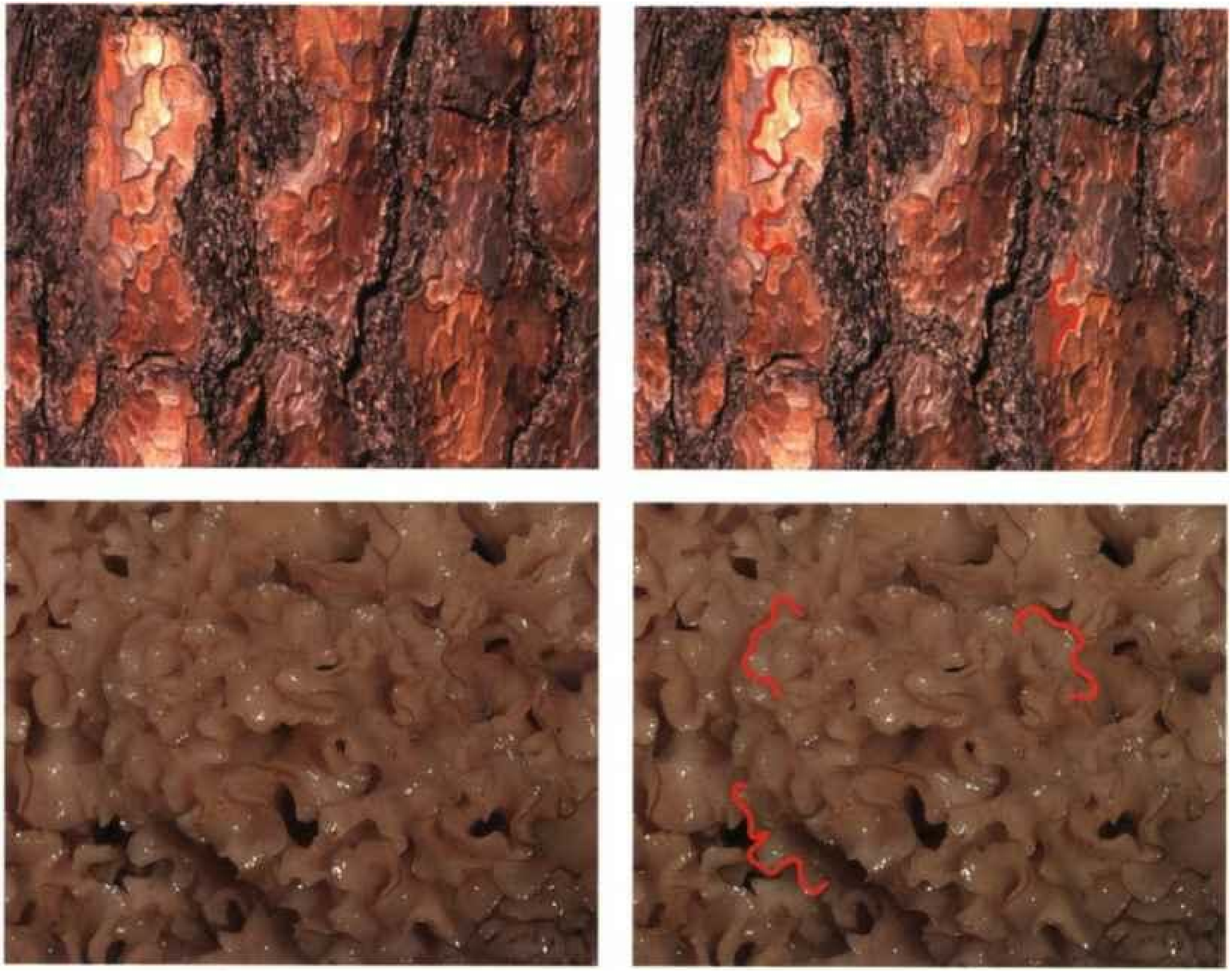

But I still want more advice on the secrets of successful mushrooming.

Mushroom hunting and the world of fungi have more than their fair share of mysterious lore, imponderables and surprises. Take that inexplicable instinct for the right spot to find mushrooms, for example. The experienced mushroomer knows his favorite haunts and whenever conditions are right, the mushrooms seem to call him. The dedicated mushroom hunter doesn’t have to make a conscious effort to memorize the perfect weather conditions for a particular species. Nor does he have to keep a log on the wind direction. Over the years, he just knows that on the right day he has to go to spot X. As vague as that sounds, the mushrooming “instinct” may well have a rational explanation, but it needn’t be a conscious one.

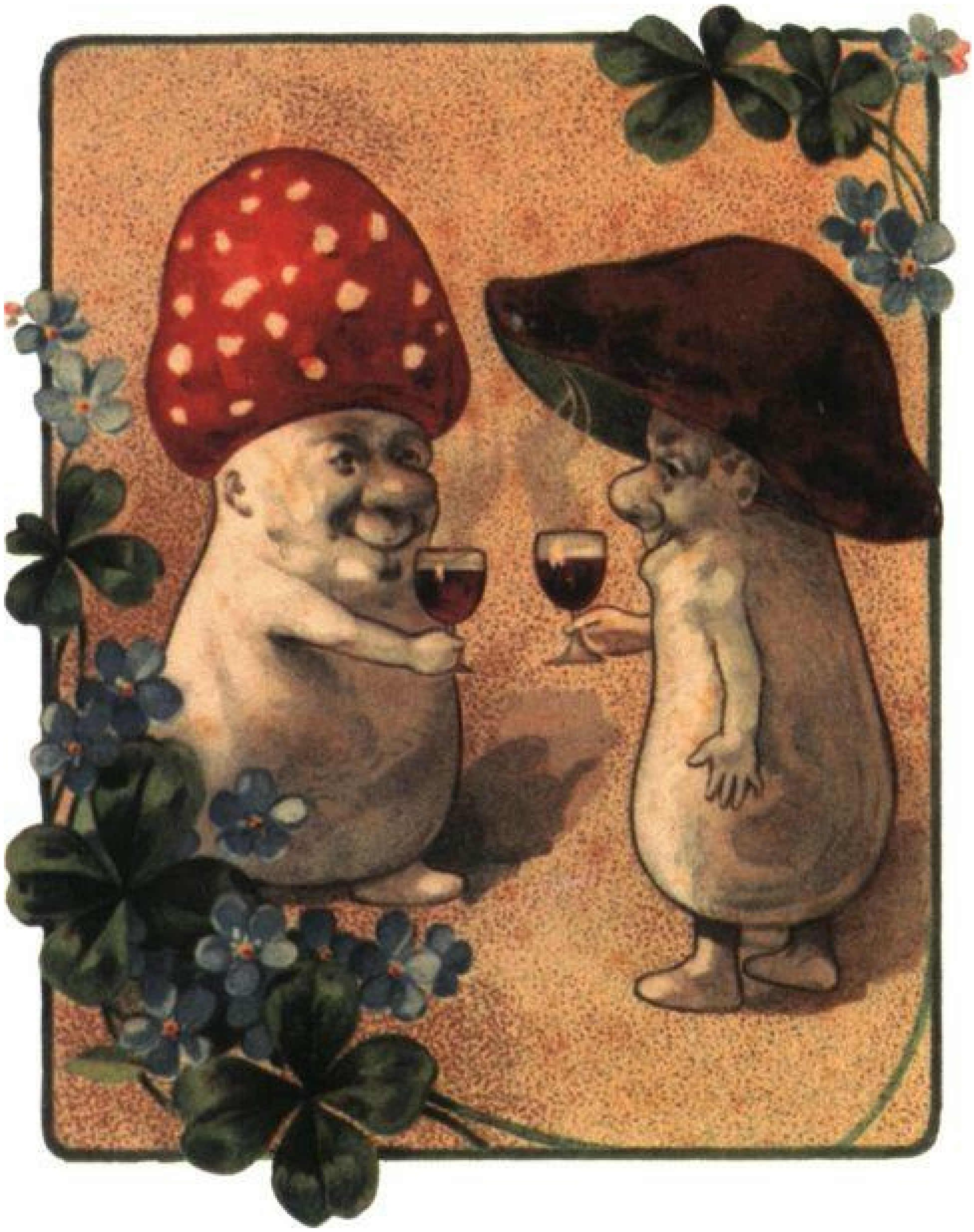

No exact science to mushrooming-hunting, then?

True, if you meet gnomes, pixies, fairies, elves or goblins in the woods, your rationality will be doubted.

But when fingers of fog creep eerily through the lonely wood, mythical creatures don’t seem far away. When out mushrooming in the twilight, you can easily understand how our forebears believed there were strange goings-on in the woods. Perhaps there were: and perhaps our modern powers of perception are just that much weaker than they used to be. The cold light of reason may have evicted all those colorful beings from the woods and now nobody is quite sure where they are anymore.

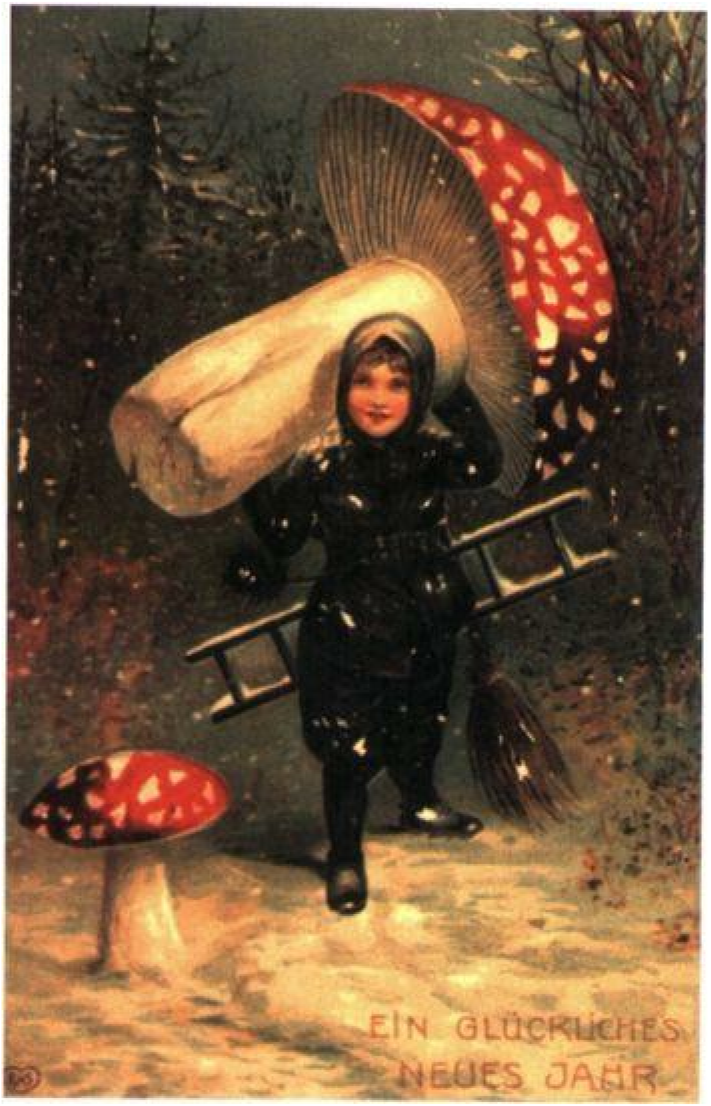

But they did leave their mark. The last bastion of mysticism is the fly agaric (left) which to this day has the power to fire the imagination. No decent fairy tale illustration is complete without the fly agaric. If the cep is the king of mushrooms, the fly agaric must be the queen. And as always with mushrooms, the mystical mixes with the scientific.

Fly agarics are useful indicators: where they are, there are often ceps nearby. Fly agarics wave and shout “Over here! Over here there are beautiful ceps!” The king is always near the queen.

Yet another mingling of fact and fiction: the fly agaric got its name from its surprising powers: it is a natural insecticide. Pieces of it are put in milk to attract flies. The flies then get drugged and die. In fact, the hallucinogenic properties of the fly agaric have inspired some mind-boggling theories and experiences.

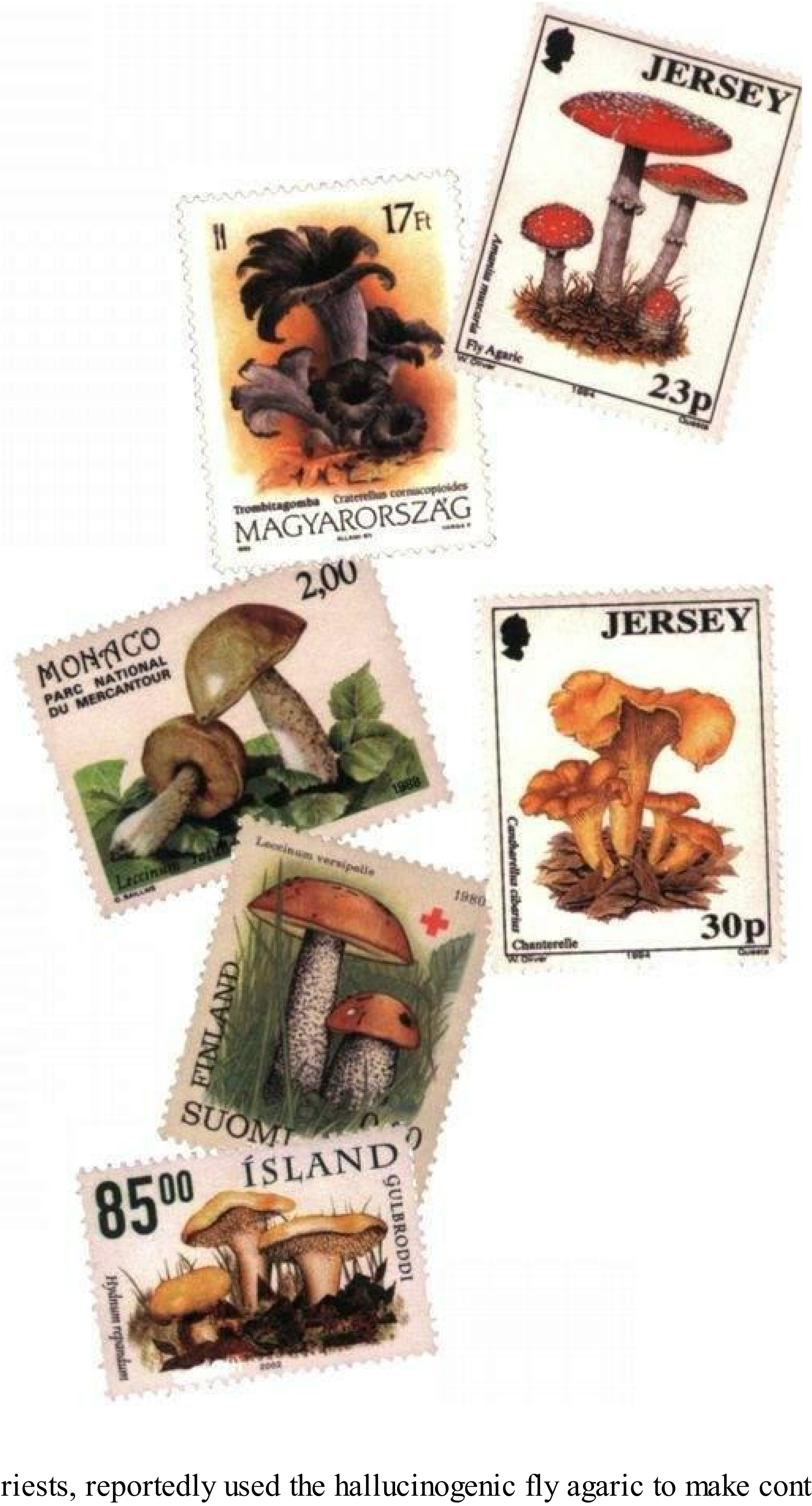
JERSEY
Amanita muscaria
Fly Agaric
1994
23p
17Ft
Trombitagomba
Craterellus cornucopioides
MAGYARORSZÁG
2,00
MONACO
PARC NATIONAL
DU MERCANTOUR
1988
Leccinum
JERSEY
Cantharellus cibarius
Chanterelle
1994
30p
Leccinum versipelle
1980
FINLAND
SUOMI
ÍSLAND
85⁰⁰
GULBRODDI
Hydnum repandum

Shamans, or holy priests, reportedly used the hallucinogenic fly agaric to make contact with the gods, while lesser mortals joined in the ritual by drinking the urine of the shaman. This would have been little more than water with the hallucinogenic substances still active in it, because, before the ceremony, the shaman had to fast for several days before ingesting the fly agaric.

The phrase "to go berserk" refers to the Berserkers (Viking warriors) who went fearlessly into battle and were reported to fight "like mad dogs and wolves." This is said to be the fly agaric's work.

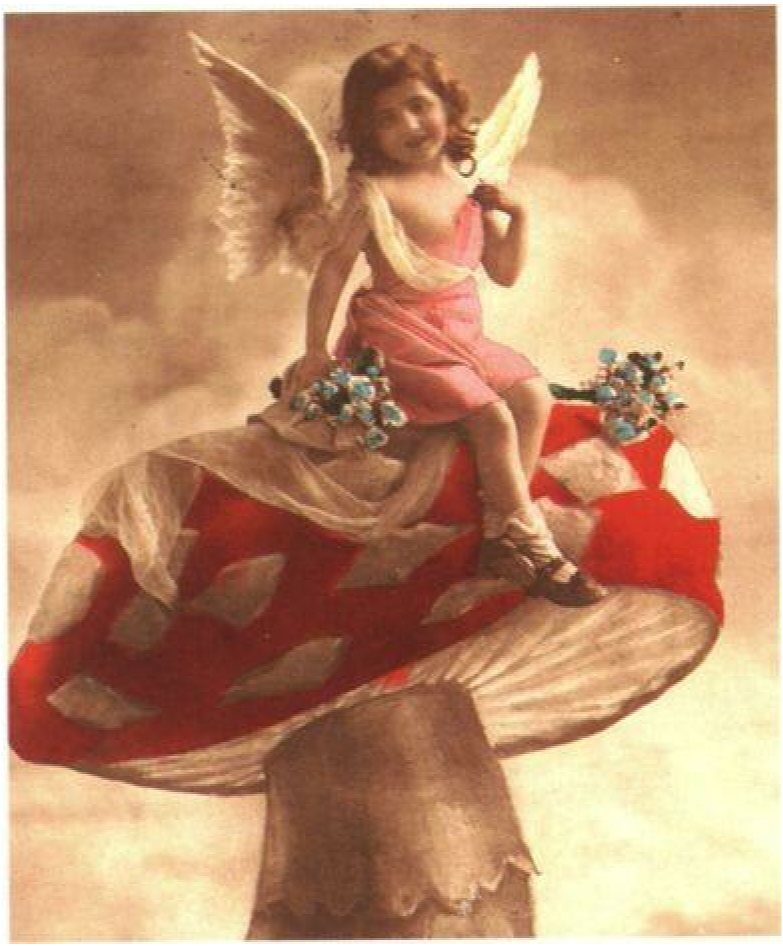

It surely now comes as no surprise that Santa Claus is also linked to the fly agaric. After all, there are the same distinctive colors: and there he is in his reindeer-powered sleigh, flying all over the world, filling stockings, and involuntarily cleaning chimneys. Only a full-blown shamanic fly agaric ritual could take Santa Claus and come up with a tale like that!

These are just a few of the theories and ideas linked to the fly agaric. Fly agarics are also supposed to bring good luck, and in certain parts of the world they remain a popular emblem in advertising, on stamps, and on Christmas and New Year's cards. Luck is a fitting theme with which to conclude this book. Don't trust your luck when identifying mushrooms. Follow the instructions in this book and enjoy mushrooming without fear.

~

The Essential Wild Food Survival Guide

Linda Runyon

WELCOME
to the eBook version of *The Essential Wild Food Survival Guide!*

We have extras for you!

When you go into the kitchen to try the fabulous recipes contained in this book, you very well may not want to drag your computer with you! Printing is disabled in this eBook, but we have the whole printable Recipe Section waiting for you, as well as a free Subscription to Linda Runyon's Newsletter!

Just go here and we'll send them to you. (If your computer can't handle the direct link, please type the web address below into your browser's address bar and fly over to Linda's site.)

http://www.ofthefield.com/eBookSpecialOffers.htm

This is a secret back door for eBook owners like you ONLY, and can't be found by coming in the front door of the site.

Your email address is never given to third parties.
It is our strict policy.
You may opt out at any time.

The Essential Wild Food Survival Guide

Copyright © 2002, 2007 by Linda Runyon
original title *From Crabgrass Muffins to Pine Needle Tea*

Linda Runyon
Wild Food Company
PO Box 83
Shiloh, NJ 08353-0083
lrunyon8@yahoo.com
www.OfTheField.com

ISBN 0-936699-11-6 eBook

All Rights Reserved, including the right to reproduce this book,
or parts thereof, in any form, except for the
inclusion of brief quotations in a review.

All Art by Linda Runyon unless otherwise noted

Photo Credits:
Ruth Demers Executive Director of Rome Sports Hall of Fame
of Rome, NY – all photos except the following:
Kerin A. Denser-Fuzy – sow thistle
Paul A. Runyon, Jr. of Seattle, WA – prickly pear, saguaro,
cholla, lamb's quarters, malva neglecta, sunflower
& shepherd's purse.

Main Cover Art by Linda Runyon
Cover Design by Eric & Todd Conover
Layout and Final Production by Eric Conover

Disclaimer

This book is intended to be an educational tool for gathering and cooking wild plants. The information presented is for use as a supplement to a healthy, well-rounded lifestyle. The nutritional requirements of individuals may vary greatly, therefore the author and publisher take no responsibility for an individual using and ingesting wild plants.

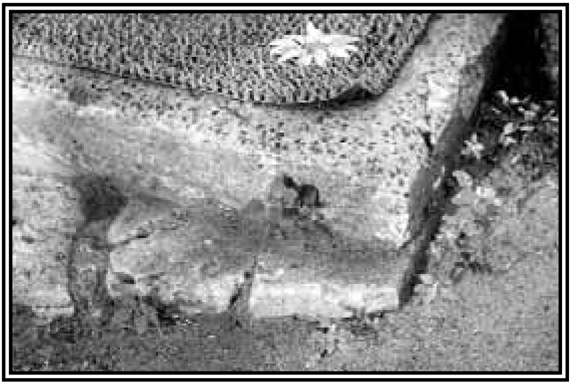

Steps with lamb’s quarter seedlings.

Waste ground plantain dandelion.

Dedication

I dedicate this manuscript to Ruth and Paul Runyon who gave me the feeling of unconditional love. Thank you, Mom and Dad. You made it all happen!

Thank you, my beautiful children whose understanding of Mom kept me on the “wild” path. Without my son Eric it would not be possible to bring you this revised edition. Over the years the help has been invaluable. My grandchildren Ceri & Alec, Jason, and Lee, who have been my guinea pigs with spearmint gum and dandelion fluff popsicles.

The homestead years were wonderful and difficult. Thanks to Ken Heitz, who spent them supporting us and our homestead! Deepest thanks to the late Dorothy and Lynn Galusha, Gert and Dick Frulla, Indian Lake, and dozens of friends: especially Nancy & Pete, Nicole, Colleen & Jenny, Jean, Terry, Shirley and Eric & Rose. I couldn’t have done it without you. Thanks to Gary for computer help.

A special acknowledgment to Cristina, one of my teachers, and her mom, Melinda. Out of the woods, the Runyon Institute began. Thank you, Walt Johnson, you taught me the business world! My deepest gratitude to Mae Charon, without whose friendship and business guidance I never would have completed this manuscript. Bruce & Larry, your faith, love and encouragement are monumental! Alice & the late Walt Sherman, your love is my port in the storm. This book was conceived in the lamb’s quarter patch behind the barn. Ruth Demers, you crawled on your tummy, festooned with black flies to take these photos, thank you. Your faith became mine. Paul, my “bro”, you stood supporting me in 110 degrees while you took the western slides! Thank you!

So many people, thank you all!!! Thanks especially Dr. Edward Glenn and the Environmental Research Lab in Tucson for your library, approval and guidance along the path to the Environmentarian Way. Susan Lake and Nancy Sprague gave me their invaluable library skills. Steve “Wildman” Brill is the foraging cohort I trust implicitly. Your advice is well taken. Joan & David Schultz, Kathy & Mel, Richard & Stephanie Bond, Tess and Michael Morearty, you all give me a tremendous feeling of support here in Arizona.

The Wild Food Walks were the result of gracious offers for the use of private lands. Thank you Rose and Bruce Burke, the Nanticote Lenni Lenape tribe of South Jersey, Jim Ridgeway; Kathy Weaver Warrensburg; Lenore Yalisloff and Hazel McManus in Phoenix, Glendale Public Library and David Schultz's direction here in Arizona.

Thanks to Dr. James Duke, Ph.D. you are the database for my manuscript! Steve "Wildman" Brill edited professional to professional. Althea Dixon, your generosity kept my newsletter going...Wanda Wilde, secretary for Wild Foods Co., thank you for your help and hello to your 6 kids (including the triplets)! I know there is a list of hundreds over the years—you all know who you are!!! Thank you! An abundant green year awaits us all.

Misi Linda Runyon

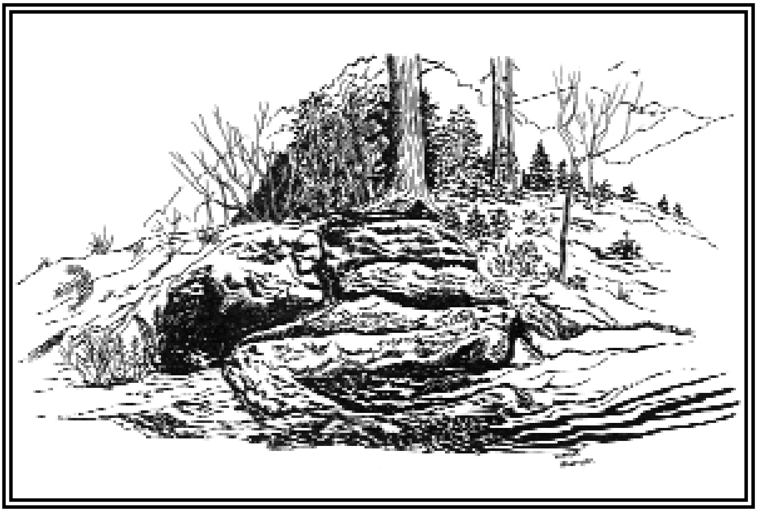

Todd's special place - the cave!

Table of Contents

Page No.

Dedication i
Foreword iv
Preface vi
Acknowledgements viii

PART I AN ENVIRONMENTARIAN LIFESTYLE

Living off the Land 1
An Introduction to Edible Wild Plants 17

PART II FIELD GUIDE TO WILD FOODS

Edible Plants 49
Photos of the Plants 141
Nutrient Value of Wild Foods 173

PART III WILD FOOD RECIPES 195
Soups and Salads 198
Vegetable Dishes & Light Meals 205
Muffins, Breads, Biscuits, and Pancakes 225
Sweets 233
Teas and Other Hot and Cold Drinks 240
Preserving Wild Foods 250
Sundries 255

PART IV REFERENCE SECTION

Cultivation of Wild Edibles 259
Public Wild Food ID Walk Ways 270
A Forager’s Research Diary 283

PART V POISONOUS LOOK-ALIKES 290
Appendix Glossary of Scientific & Botanical Terms 302
References 305
Suggested Reading 306

Foreword —by Joe Gang

Linda Runyon is a child of the Adirondack wilderness. It is there that, at a tender age, she discovered that wild clover, yarrow, and wintergreen were edible. As she grew, so grew her knowledge of the abundant wild flora in her beloved mountains.

In 1972 she decided to homestead in the wilderness. Her ability to recognize and use wild plants added immeasurably to her successful survival. By adapting to a diet of wild vegetables, herbs, fruits, and nuts, Linda carved out a niche for herself among women pioneers who relied on Nature. Linda Runyon has gone on to share her wide-ranging experiences and knowledge of wild foods with a growing following: visitors to several plant identification walks (called *Survival Acres*), readers of her books, (*A Survival Acre, Lawn Food Cookbook,* and *Wild Food and Animal Coloring Book*), users of her fantastic plant identification card game ("Wild Cards"), and subscribers to her (now out of print) Environmentarian Newsletter, and the "Wild Foods Nutritional Wheel", added later. Linda Runyon's valuable *From Crabgrass Muffins to Pine Needle Tea* opened a vast fund of knowledge on nature's own foodstuffs, which often can be found close to home, in an environment unsoiled by insecticides, herbicides, or fungicides. Here is the key to the preparation of wild foods without the use of preservatives, invasive chemicals or factory processing.

And now we present a new edition of that book renamed to *The Essential Wild Food Survival Guide,* with all the sustaining data from the first edition. This book will continue to be your valuable companion on the path to healthful living.

The author offers a new way to view the world.

Poem

MY LAWN

As I walked across the grass,
I noticed clover, dandelion, plus
a few plantains.
The thought came to me that
I should graze on all four.
All other modes of feeding
now seem so inefficient,
even wasteful.
When I go out to graze,
I'll carry a bottle of salad dressing,
and sprinkle liberally in front of me,
as I munch my way across the lawn.

By Joe Gang Environmentalist, master of Economics and finances, chemist, small business consultant, Indian Lake, New York.

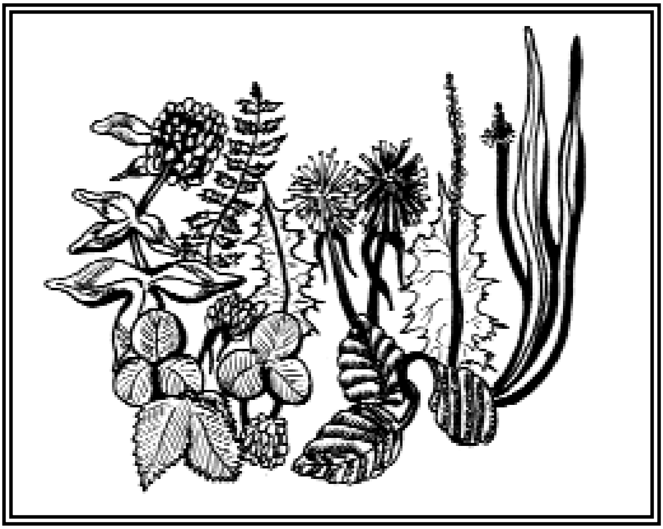

A typical Eastern lawn smorgasbord

Preface

I am called an environmentarian, or a forager in common terms. I enjoy foraging in fields, woods, and even backyards for the wild plants that make up as much as 30 percent of my daily diet.

The knowledge that led me to this way of life did not come overnight. Walking eight and a half miles to work each day one summer taught me the value of knowing the energy in a single clover leaf; after 15 years of homesteading, (living without any power or electricity in cabins and tents) I learned the tricks of gathering and cooking nature's food. It is quick growing, accessible, nutritious, and free!

Many years of this experience taught me not only how to forage, what to eat, and how to prepare it, but it has also confirmed my decision to stay in touch with my primal rights to live a free life. Now that I'm almost 70 years old, my goal is to share this knowledge with others from a comfortable retirement area.

I have divided this book into four major parts, and I've also included some backup information in a resource section and glossary. Part I describes my environmentarian lifestyle, including how I came to adopt this way of life and how I learned and acquired the wilderness skills that helped me survive. In this section, I also provide information on gathering wild plants as well as how to store them for later use—drying, freezing, and the like.

Because properly identifying wild foods in the field is critical, I have provided a small Field Guide in Part II, which pictures and describes over 50 wild plants, including grasses, herbs, brambles, and trees, plus their nutritional values.

In Part III, I provide more recipes that use wild foods as the major ingredients—tasty and easy-to-prepare soups, salads, casseroles, breads and crackers, sweets, teas, jams & jellies.

In Part IV, I offer tips for growing a wild food garden and for creating "wilderness" walks. I also explain how to grow certain varieties as potted plants, should you live without space for a garden or want to bring your plants indoors. This section also contains my foraging research notes, which tell how I investigated the edibility of wild plants.

Part V discusses Poisonous Look-Alikes and the Glossary defines the terms—mostly medical and botanical—that I use in this book. The book concludes with a listing of useful sources for additional information on wild foods.

It is my hope that this book will help you begin to seek out your own free life, should you wish to do so.

Remember, the process of living free does not happen overnight. In fact, the changes in my lifestyle came very slowly, with gradual adaptation. Try it yourself. If you cannot make it one way, adapt. If you cannot find the food you need, change. If you cannot find shelter, make it. If you are cold, build a fire. If you cannot take a bath, adjust! Such adaptations brought about the changes necessary for me to adopt a free lifestyle.

These changes meant being free of monetary need, not free from constant building, splitting wood, and food seeking. Meeting these needs took up a 24-hour period, with occasional days or hours in between to enjoy the lifestyle. As with any endeavor, you need to enjoy the project. I love projects—they fulfill my day. I always set a goal: move a cord of wood, trim that maple tree, re-screen the windows where the raccoon came in, bathe the mule, clean out the stalls, milk the goats, gather some food for dinner, can the tomatoes.

I plan each day and finish the job when I reached my goals. In between, I feel "on vacation" while I gather food for wild food salads, soups, or casseroles. And, as tired as I am, I have enjoyed every minute of every day.

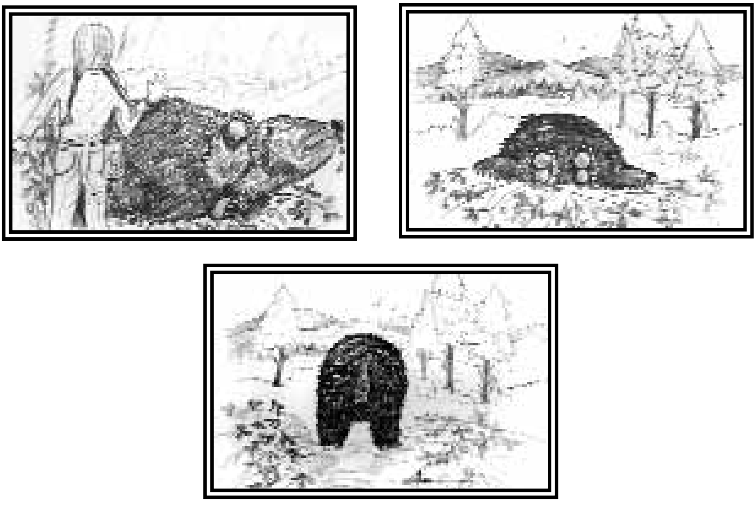

Linda's Surprise Encounter with a Black Bear

(See page 118)

Acknowledgments

Because of my geographic locations, I use Cherokee and Algonquin language as well as Chippewa throughout the text. For Cherokee translations I have used the CHEROKEE-ENGLISH DICTIONARY, Cherokee Nation of Oklahoma, 1975, Commemorative Edition, edited by William Pulte, Southern Methodist University Press, Dallas, Texas. The dictionary was graciously lent to me by my friend, Joseph Tenderfoot Hill, an Iroquois Elder. I am also privileged to have worked with my friend, a mixed-blood Dijalaga (Cherokee) teacher, Willy Whitefeather.

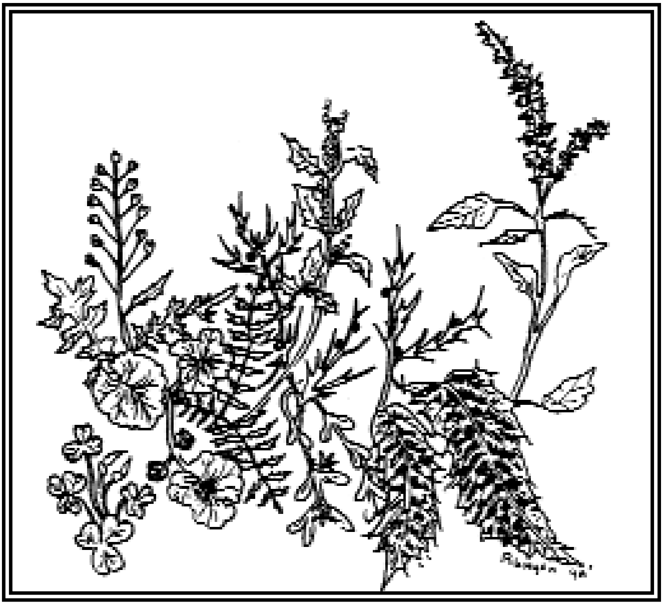

A Typical Western Smorgasbord

Part I: Environmentarian Lifestyle

Living off the Land

It seems that the Adirondacks Mountains and their way of life have been a part of my life for as long as I can remember. For this, I have deep gratitude for my parents, Ruth and Paul Runyon. While growing up, I worked on the girl's staff of Gavetts Camp on Indian Lake in upstate NY, now called Timberlock. From the time I was ten years old until I was eighteen, I ran between cabins and tents, carrying army blankets, making beds, and serving 100 guests three meals a day. The dishes and campfire sing-along became a nightly routine for three months of every year. Six A.M. lake dips, canoe races, island campfires, and mountain climbing all led the way to my eventual homesteading experiences. Each year, the return to school in North Plainfield, New Jersey became more and more difficult. Smelling a souvenir balsam pillow linked the mountains to the city for me, and even the taste of wintergreen became a symbol of freedom.

My high school years flew by quickly, and soon I had attended and graduated from Monmouth Medical Center in Long Branch, New Jersey, as a registered nurse. Not long after I married, the first of my family arrived: Eric taught me the wonders of motherhood. I don't have to tell anyone how close your firstborn is to your heart. Then within two years, a daughter arrived—Kim, my beautiful daughter, heralded a different bonding in my life. A few years later a second son, Todd, was born. He was to share my entire homesteading experience.

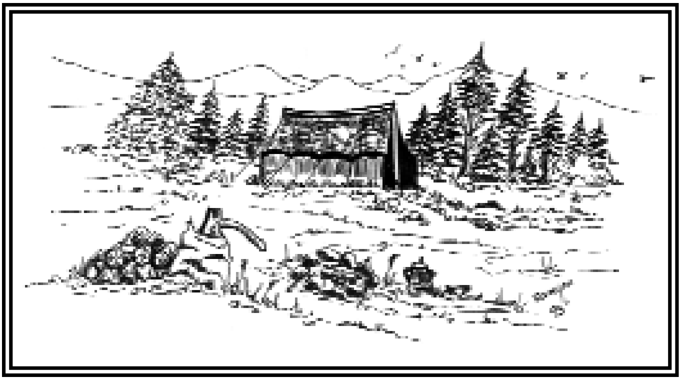

We spent 6 months in a tent.

For many years, all my family thrived in a large housing development in Howell Township, New Jersey. Eventually, my husband Frank and I divorced, and a year or so later I met Ken Heitz. Todd, Ken, and I left New Jersey with an ax, sleeping bags, and great anticipation for the Adirondack wilderness. Kim and Eric remained with their father, but would join us in the summers. Thus my homesteading experience, wild foods, winter survival, and just plain Adirondack life began.

It wasn't long before ideas of an environmentarian lifestyle became a reality. I began to develop the strong bond I feel with Mother Earth. My life was plants, water, shelter, earth, life—"the basics," as someone once said.

Instinct and Ingenuity

One day, in the middle of the winter, I was taken to see a cabin for rent across Lewey Lake. Peering through the cabin's frosty windows, I took less than a minute to shout, "Yes, we'll take it!" An unwinterized cabin, isolated from other summer cabins and framed in black against the mountains, looked like heaven to me. At twenty below, I moved in with Ken and my son Todd, who was age five. We bobsledded the furniture down a moonlit icy path and over twelve-foot snowdrifts to our new home.

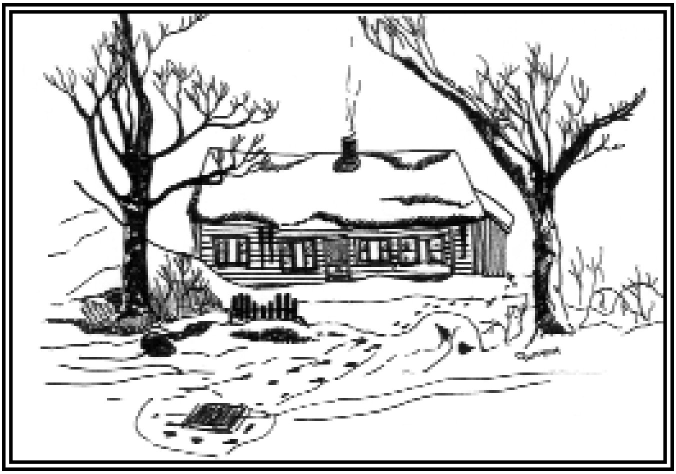

Log cabin. Note water hole in frozen lake.

My very first job was to make a pot of coffee. Ken gave me an ax and a pot, and shoveled a path to the frozen lake. Two hours later, time seemed frozen. A two-foot-wide hole grew before me. As I broke through the ice, I felt as if it was ten thousand years ago. Just as folks at the dawn of time had to learn, I, too, soon discovered my ignorance. When the vacuum was released through the newly dug hole, the water rushed up and quickly disappeared below the ice, dragging everything with it. I dove for the coffee pot, only my arm was not long enough to reach deep enough into the water. With tears streaming down my cheeks, I realized my great need for a friend, a tool, and better instincts! Ken provided the first two, and my own instinct led me to my idea for a rope dipper. To this day I can find no words to describe that first cup of coffee!

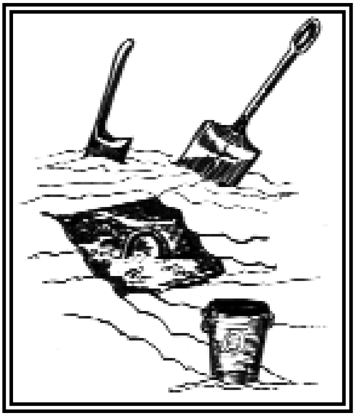

Water hole tap

Years later, my friend Willie Whitefeather would show me how to gather water in the desert, at the base of a cottonwood tree. Perhaps his Cherokee words, “adline” or “draw liquid,” describe that coffee.

Due to a series of events, I learned to respect the Adirondack winter. The first incident occurred when I opened my eyes after a night’s sleep. My head seemed stuck to the pillow—my hair had frozen to the frost on the cabin wall! Frosted walls are a feature of life in the far north. A layer of ice 1/4 inch thick covers the walls most of the winter, brought about by intense cold, uninsulated buildings and potbelly heat. My side of the bed was against the wall, and that morning I was literally stuck to the wallboards. Warm water might have solved the problem, but one doesn’t always think clearly in a panic! Instead, a haircut solved the problem.

Another morning I awoke and felt a layer of frost on my upper lip and under my nose. I saw my breath as white as smoke, and I knew the stove had gone out during the night. I put my slipper socks over my thermal underwear and hopped to the stove, stirred the coals, heaved in a few pieces of beech, and dove back under the covers. Hearing the crackle of the wood as it flames became my security signal on cold winter mornings.

As I began to learn more about winter, I decided to can my wild foods during the summer for use later in the year. We had a seven-and-a-half-foot insulated refrigerator pit where they would keep well, according to a book I read. I trusted my judgment on this, for I had canned many times in my "city life." So after a long canning season I had 420 jars of campfire-canned foods in my storehouse.

When winter came, we were thrilled to eat our canned foods—cattails, milkweed, lamb's quarters, blueberries, and strawberries. Everything was delicious. Then three weeks of twenty to forty-below zero temperatures began. One night a blizzard snowed us in, but I knew we had plenty of food. The winds howled, the snow swirled, and when morning came, the refrigerator pit had to be dug out. When the roof was uncovered, I opened the pit and climbed down the ladder with my lantern in hand. My eyes blinked in disbelief. Almost every one of the jars had exploded. I was in a complete panic. I grabbed the snow shovel and began to scoop the jars off the shelves to find ones that had just begun to freeze. Alas, there were but a few jars with small cracks or pushed-up lids.

Reconstituted dried leaves, vegetables and flour became our immediate food sources. We even ground bark into flour or brewed it for teas. Somehow, we survived. As I look back, I would say that it was this experience that led me to write this book.

Deer are My Teachers

The first time I looked for food in the woods, I took time to become attuned to the deer's world. I followed deer tracks and saw where and what they had eaten. I found out where they appeared to move quickly, where they moved slowly, and where they stood and relaxed. And when they relaxed, I discovered they were almost always eating.

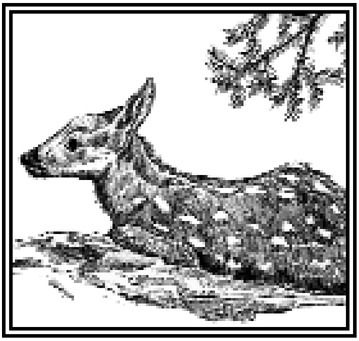

Well fed fawn

Although not every food eaten by deer is safe for humans, much is.

It became a simple matter to find holes deer had pawed in the snow. These snow diggings yielded wintergreen, ferns, and mosses. I noticed the deer bit off the tips of balsam fir needles. Dried goldenrod tops jutting above the snow line were also nibbled.

Dried-up apples and rosehips also helped to sustain the deer population in the dead of winter. In early spring, the new growth on blackberries, raspberries, willow, and birch added to the deer's diet.

I loved to find an apple tree in winter, naked of apples only as high as the deer can reach on their hind legs. I took the frozen apples from the tops of these small trees back to the cabin, where I boiled, simmered, dried, crushed, and otherwise experimented with them.

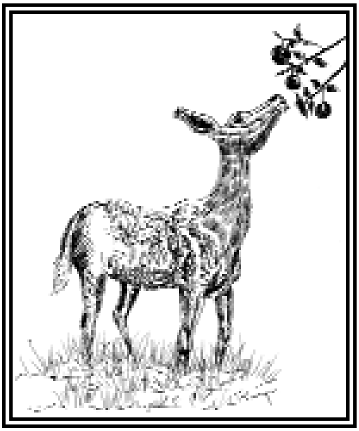
Deer apples

Frozen streams also held many secrets. The deer knew where the water dropped over a rock at fantastic speed, deep beneath the snowdrifts. They would break through there for an icy drink, and then steam would rise from these holes because of the difference between water and air temperatures. Todd and I dipped a long, dry stick into the hole to make an instant popsicle. But we discovered that the ice would freeze solid to our tongues—a lesson hard learned.

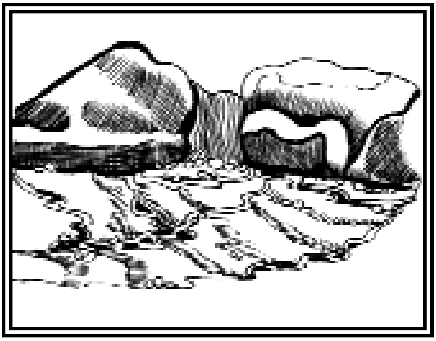
Waterfall

In the spring, lamb's quarters was a favorite bedding for the deer. I also learned that jewelweed is a poison ivy antidote and as luscious a vegetable as lamb's quarters. Near the camp was a hundred-square-yard patch of jewelweed in which the deer fawned. What a place to give birth! Drops of crystal-clear water hung from the emerald-green stalks of the jewelweed as the deer gorged themselves on this plant. Level weed patches grew uniformly, so I could often tell where the deer have been by the scalloped designs of the weed tops.

The deer were sometimes invisible in this patch. As I crossed the area one day, a buck rose beside me, covered with cascading jewelweed stalks. We stared at each other, not more than three feet apart. For a long moment his antlers had looped over me like branches on a tree. I smelled his breath as he turned and walked away slowly. Through my eye contact, I apologized for my intrusion and swore I would call out loud before entering the area in the future.

Here's Mud in Your Eye

The homesteader in the Adirondacks needs to utilize both modern and ancient methods of survival when it comes to the insect world. For example, insects are particularly fierce during the short summer in the far north. Early in the day, black flies were biting even the corners of my eyes by 9:00 A.M. By noon, my upper arms were filled with itching lumps, the back of my neck was bleeding from my scratching, and I had swallowed at least four of the beasts.

I had literally used bottles of insect repellent, but even the heavy-duty spray from the Grand Union in Indian Lake was not working. The black fly problem was so great that all of us were sick from the toxins in our body from both the flies and the chemicals.

One time I came home to a prostrate husband, bloated and swollen from at least 400 black fly bites and sporting a temperature of over 103 degrees. A trip to the town doctor was in order. Since that time, we discovered the use of smudge pots. In fact, smudges soon became a skill we excelled in! Smoke wafted from our campfires, pails, upside down wastepaper baskets—you name it. The black flies were kept at bay.

Building a Smudge Pot

In the neighborhood dump area we were able to remove lidded tin cans, the 20-gallon type. Manufacturers distribute products in them. In the city I would ask local garages for oil cans, and restaurants for old lard cans. If all else fails, use an old wastepaper basket and make a thin lid to fit.

The can must be washed thoroughly and pre-burned for safety. Build a fire using paper kindling. Cover and wait until this burns out. The "clean" smudge pot needs a door in the base. Cutting would be simple using modern tools, but in the "homestead times" we used a nail and a rock. We placed holes side by side, much like you might open a can without a can opener. The door to a 20-gallon container might be 6 by 8 inches or even smaller. Don't cut the fourth side of the rectangle, and bend out the door. Smudge cans last for a season or more according to how much use the "hinge" side gets.

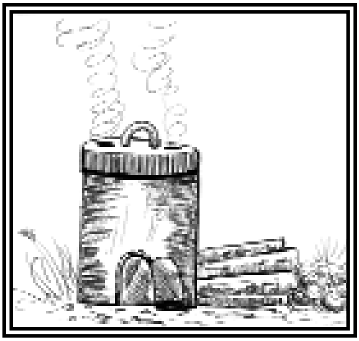

Smudge pot. Metal container with lid, rolls of moss.

Be sure the lid fits if you use a makeshift wastepaper basket or thin pail. Control your airflow by leaving the lid off, opening your door a couple of inches, and experiment with keeping a fire going. When you have coals, place a moist material over them and stand back. Your smudge is smudging. After a while, from the size of smoke, you can tell when to add more wood, open door and lid, or add more smudge material. When you go away, simply put the lid on and shut the door.

Smudge Materials

- Grass clippings (be sure you don't use toxic plants like poison ivy, oleanders or chemical sprayed lawn clippings).
- Moss- any kind works well.
- Wet Sawdust.
- Other material, such as newspapers soaked with water, wrung out and tightly waded up. the tighter the wad, the longer the smudge.

The Last Hunt

The transition from meat eater to vegetarian was a difficult one. Nevertheless, it seemed the best way to follow through on an environmentarian lifestyle.

One day my son Todd accompanied me hunting, the first time for him. I know now that hunting was probably a mistake, but I hadn't suspected the eight-year-old did not know where his meat came from. I performed my usual stone-throwing act, and shot the first rabbit that ran out from behind the lumber pile. Unfortunately, the rabbit rolled over, righted himself, and to my horror continued to run straight for us. I took careful aim and shot it just a few yards from where we stood. The rabbit fell, twitched, and then was still. Todd screamed, horrified at me. "Why did you do that?"

I heard myself saying, "We eat them for dinner. That's why we grow them."

He stood defiant and drew himself up into a proud, straight position and said vehemently, "Well, I don't have to eat them again, and I won't—ever!" Todd walked off in a rage. He meant every word he said, and that was the last kill I made for food from the mammal world. As it says in Genesis 1:29, "Behold, I have given you every herb bearing seed, which is upon the face of all the earth, and every tree, in which is the fruit of a tree yielding seed; to you it shall be for meat."

My theory now is that if we could see the reality of taking life, we would all be vegetarians. Happily, many youngsters are being introduced to this thought by activists or through the animal rights movement. This may well change people's eating habits in the future.

Foraging on the Way Home from Work

Although we were still learning to survive on our own, we needed to continue earning money for a while. The eight-and-a-half-mile stretch of Cedar River Road became a pathway for me to a job at the Cedar River House: a famous hotel often visited by dignitaries. The hotel needed housekeepers, so to earn extra money I worked there one challenging summer. The hotel no longer exists, but the trip to and from work is a treasured memory.

Sometimes I asked for a ride, but usually I welcomed the long trek home in order to sample the food along the way. I would consume large amounts of red clover, crabapples, and dandelions on these trips. Indeed, endurance and energy were sustained by these wild foods.

The deer became accustomed to my ways; and as I passed by them, they made no more than a faster twitch of their tails to acknowledge my presence.

I once saw a large bear along the road eating crabapples. He raked the apples from the branches of the tree with ease, then dropped down on all fours and pulled them up into his mouth. I watched him from a distance of a hundred yards, feeling very exposed. I waited more than twenty minutes for him to get his fill before I walked on. There was no need to disturb the feeding animal when I could avoid doing so, even though my family was waiting for their dinner.

Fire Making

I became an expert fire maker through daily repetition. We gathered kindling from around the area and found twigs for tinder underneath the branches of a balsam or pine tree. Birch bark became a favorite fire starter. Even through torrential rains and high winds, the birch bark always burned steadily.

There are many types of outdoor fires. My favorite fire pit was dug where there were no roots or rocks. It was located near our old homestead, with a rain barrel within a few feet in case of emergency. I used a shovel to sprinkle dirt over the fire when I had to leave it quickly.

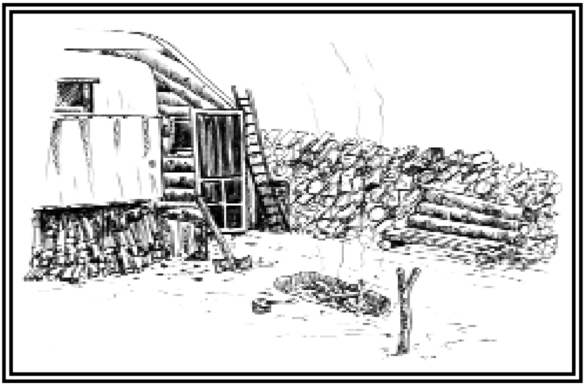

The shanty, fire pit & wood pile.

After a few months, the dirt had been dug out and reused so many times that the ashes and dirt were a fine powder. Afraid of breathing the lye of hardwood ashes and dirt, I began to use a large piece of sheet metal, throwing it over the fire quickly. It wasn't long before I discovered this metal was a stove like the most modern type. I began to make roaring fires with the sheet metal on top, and I would simmer, boil, and fry all on one safe surface. In pouring rain or blizzards this method worked exceptionally well.

I found that a piece of sheet metal also served as a cover for a banked hardwood fire, and the coals would last until morning. If we went on a long foraging trip, the sheet metal cover helped prevent our coming back to a cold fire pit at dinnertime.

A fire pit holds a dug-out below-ground fire. A fire ring is an above ground safety rock ring around a fire area.

After a deep snowfall, we dug out the fire pit to prepare our meal and cooked in the middle of a snowstorm. During spring and early summer, we would cover ourselves with mud while cooking to protect ourselves from insects.

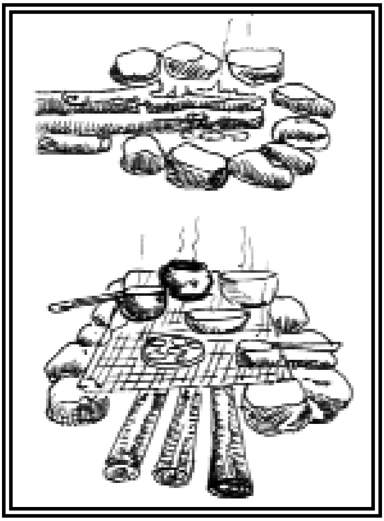

Fire ring and cooking.
Open-end fire, with logs that are rolled over to control heat level.

Scanning the coals using my hands, I could determine which areas of the fire would be used for cooking each dish. To control the boiling level I added twigs or hardwood to the fire. Since tending the fire and stirring the food are a constant chore when using an open fire, the children were often called into action. On Thanksgiving they spent eight hours turning a well-stuffed turkey over our fire pit.

My favorite type of fire ring was open on one end. I pushed the long logs in as I needed them; putting a new log on the fire means fresh fire or flames in the surrounding area. I made hotter areas for simmering or boiling by adding new twigs or fresh logs. The hot rocks on the sides were used for toast or baking cakes. I greased the rock for best results as I would a pan.

To cook on our campfire, I used a spit for long-term roasting, coals for baking, and hot rocks for making toast. I baked main meals in coals (casseroles). Hot rocks and back burner held breads, which were unleavened, pita-like cakes throughout the wilderness years. Toast in particular tasted fantastic when done on the open fire. I would turn the bread when I saw the tiny coils of smoke appear at the edges, and I learned to keep the bread to front edge of the grill to avoid burning the hair on my arms.

In the dead of winter I especially enjoyed cooking over the open fire. There were times when my back was thick with snow, but I was oblivious to discomfort. I would squat in front of the fire and relish preparing my family's meal. The personal challenges of foraging for and cooking our wild foods over our campfire gave us all a special appreciation of our daily meals.

Swamping

Our camp was bounded on the east by a vast cattail swamp and so the food became a mainstay.

"Swamping" brings out fantastic feelings. The marsh odors and insect life opens my senses and I feel animal-like. I made the mistake of wearing sneakers the first time I went out. "A little mud won't hurt me," I thought. Unfortunately, swamps and marshes have a tendency to suck the sneakers from your feet. As you step up, your shoe stays behind. And to retrieve it from the muck is no small feat. Marsh mud is thick, gooey, and smelly at the least.

Detractions aside, a number of edible plants can be found in marshes. All the plants described in this guide are fresh water plants. Mint and arrowhead are prolific, while meadowsweet and black birch also grow among the cattails.

When you go swamping, wipe a thin layer of mud on your face (temples to cheeks).

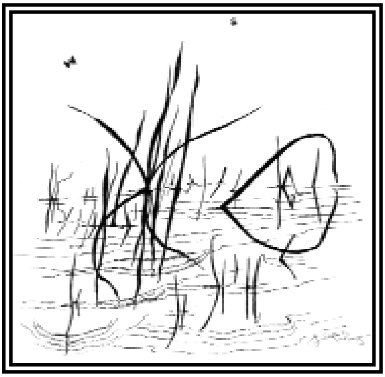

Swamp Grasses.

In fact, smear mud on as much skin surface as you can. This gives you a neutral marsh odor and bugs won't pay as much attention to you.

I would lay boards down on the marsh one in front of the other as a stabilizer. When I was ready to move on, I would pick up the first board and lay it down again in front of the second. You get to keep your shoes this way.

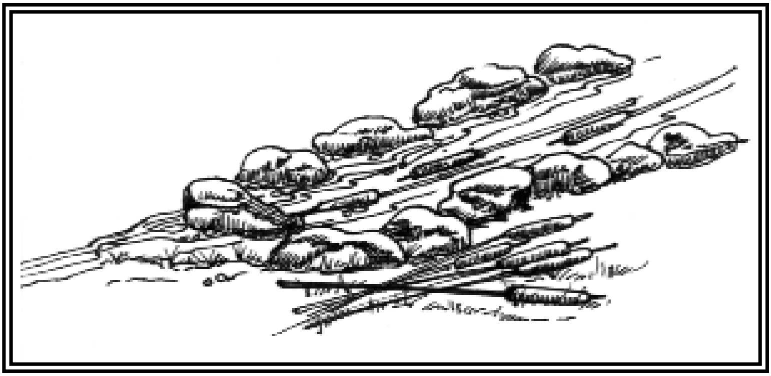

Using the stream to wash cattails

Water Hauling

Water is a basic need for all of us, but when you live in the city, you take it for granted. Our transition from city water to wilderness water was a monumental adjustment.

Ken did most of the water hauling, but there were many days when I shared the burden, carrying the five-gallon container on my back. The feeling was one of extreme satisfaction—filling, hefting, and toting the container to camp to supply our water needs for several days. I would use the drinking water one day and

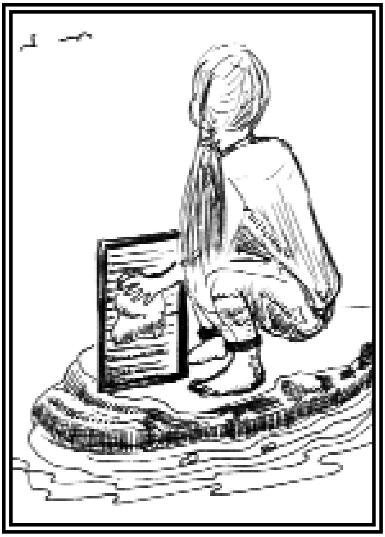

The Wash Rock

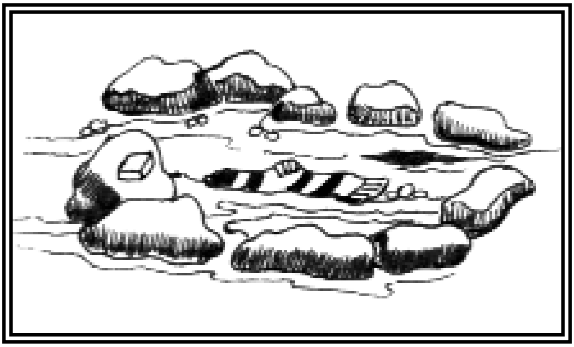

A wash ring

then use the leftover water for cooking, washing, and preparing food. We also placed rain barrels in strategic places and washed ourselves with biodegradable soap when saponin (lather-producing) plants weren't available. At the river, everyone had his or her favorite spot to bathe in privacy. It was easiest to wash our clothes there, too.

Icicles

Under our uninsulated roof were the largest icicles I have ever seen. We used them for water. If we dropped one indoors, we squealed with delight as we chased the broken sections all around the linoleum floor of the camp kitchen. The average icicle was about four to six feet long and sometimes 12" or more in diameter. I could not leave them alone.

My son Todd helped me with this chore. Ice and snow melt at a great rate of speed in a bucket that is set into the top hole of the wood stove. A giant icicle might make four pails of water when melted. It was a true gift to the person washing clothes, dishes, wild foods or themselves.

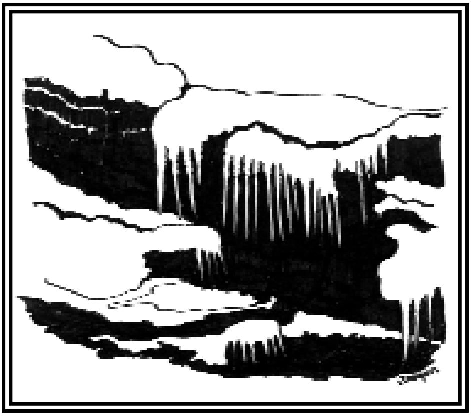

Icicles

Reverance for the Wilds

The following poem was given to me by my mother, Ruth. It expresses my feelings for my time spent in the Adirondack Mountains. Its author is unknown, but the poem is from the early nineteenth century, reprinted from *Nirvana Lodge Log Book*, Indian Lake, New York.

A Camper's Prayer

God of the Hills, grant me Thy strength
to go back to the cities without faltering.
Strength to do my daily tasks
without tiring and with enthusiasm.
Strength to help my neighbor,
who has no hills to remember.
God of the Lake, grant me thy peace,
and Thy restfulness.
Peace to bring into the world of hurry and confusion.
Restfulness to carry to the tired ones,
whom I shall meet every day.
Content to do small things from littleness,
Self control for the unexpected emergency
and patience for the wearisome tasks.
With deep depths within my soul
to bear with me through crowded plans.
The brush of the night time where the pine trees
are dark against the sky.
The humbleness of the hills
who in their rightness know it now,
the laughter of the sunny days
to brighten the cheerless spots in a long winter.
Fill me with the breadth and depth and
height of "Thy Wilderness."
Thou has taught me by every thought
and word and deed.

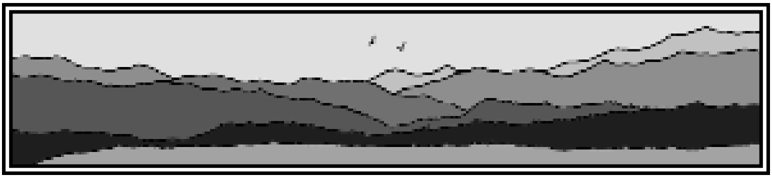

The Simple Life

The following was written by Mrs. Barbara Jennings of Long Lake, New York, and is included here with permission.

Living amid the splendor and protection of the mountains, the people lead comparatively quiet lives. From year to year, very few changes take place, and everyone becomes accustomed to a routine. Consequently, the members of these small communities are independent and fixed in their ways. They keep well informed through television and newspapers, but because of the distance to be traveled, they don't often venture away from the local scene. The tourists are accepted as a necessary intrusion and the tourist season brings enough turmoil to last the entire year. Every man prefers to live according to his own pace, no one setting a very fast one. Those who maintain a carefree existence somehow seem to manage nearly as well as those who do work steadily and live moderately. Many prefer to have less, rather than submit to the shackles of a time clock.

On the occasions when we leave our towns, we are startled into a realization of how naive and backward we must seem. We may have read that skirts are shorter this season, but complete shock is in store when we arrive in the big city and discover how really short or long they are.

A friend and I felt quite at ease and well dressed when we arrived in Utica for a shopping trip, but a shoe salesman soon squelched me as he removed one of my best shoes and remarked, "Say, these are real old-timers, aren't they?" I spent the rest of the day trying to get lost in the crowds.

Even the ever-present desire of women to get out and explore the bargains and varieties offered in the large department stores soon diminish once they are there. They become entangled in the confusion of clutching hands at the sales counters and in the hopeless attempts to attract the attention of tired and bored sales ladies. Even the latest methods of self-service result only in standing in long lines waiting to be checked out. More often than not they return home weary and empty handed, resolving to remain a slave to the mail order catalogs and corner grocer rather than submit to such indignities.

Periodic trips to the bustling city are sufficient for anyone claiming to be a mountaineer by birth or by choice. They are ready to return to their towns where the roads may wind endlessly, but stoplights and traffic cops are nonexistent. No one waits in a booth ready to extract a toll. Even in this highly complicated and sophisticated age, the values of a simple life cannot be denied.

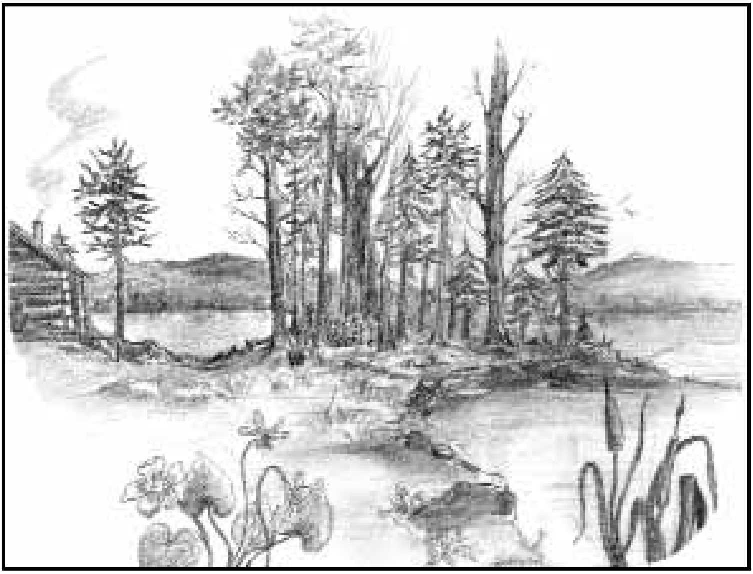

Left: Violets Right: Cattails

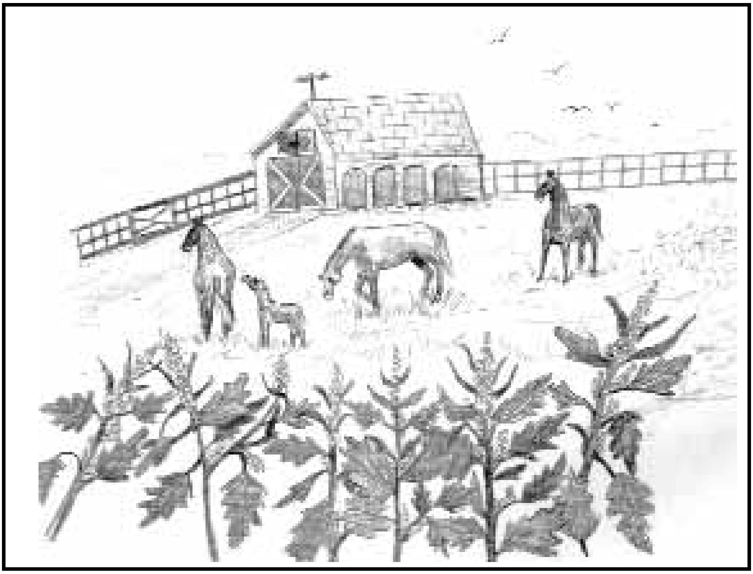

Foreground: lambs quarters

An Introduction to Edible Wild Plants

My life has been spent in two geographically opposite environments, giving rise to an environmentarian diet mixing woods & wilderness with Arizona desert. This experience is reflected here in the wild plants I choose to write about and the uses I've made of nature's bounty.

In many ways, the history of the United States documents people's efforts to survive. This is especially true of Native Americans, who spent long years establishing roots in this soil. In the Eastern wilderness, the Iroquois were bark eaters. Intense cold gave rise to the habit of eating fern roots, balsam needles, and pine bark, and of chewing black and white birch pith. Native Americans invented ways to dig roots from under the snow, such as for ferns and wintergreen.

In the West, specifically in what today is southern and central Arizona, the Hohokam were the nation's prehistoric people. From 300 BC to 1450 AD, they were hunters and gatherers. Wild foods were their diet, including cholla, lamb's quarters, amaranth, purslane, and saguaro.

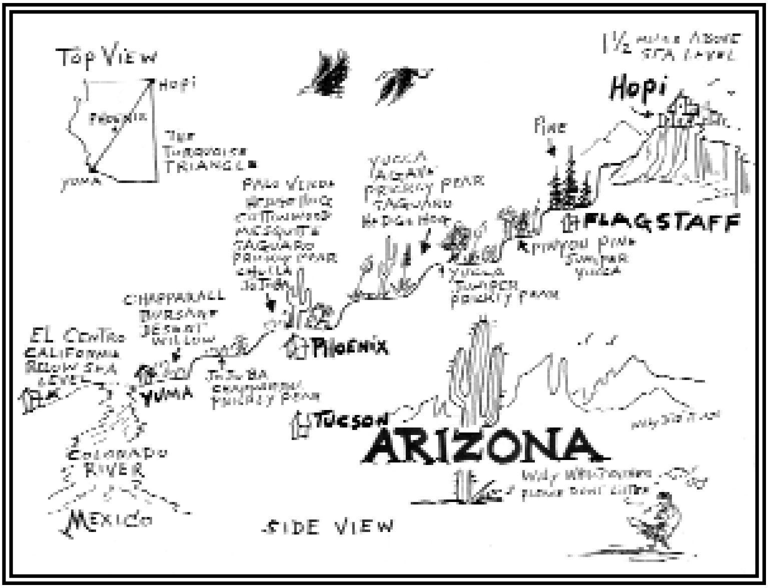
TOP VIEW
HOPI
PHOENIX
THE TURQUOISE TRIANGLE
YUMA
1½ MILES ABOVE SEA LEVEL
Hopi
PINE
YUCCA
AGAVE
PRICKLY PEAR
SAGUARO
HEDGEHOG
FLAGSTAFF
PINYON PINE
JUNIPER
YUCCA
YUCCA
JUNIPER
PRICKLY PEAR
PALO VERDE
COTTONWOOD
MESQUITE
SAGUARO
PRICKLY PEAR
CHOLLA
JOJOBA
CHAPPARALL
BURSAGE
DESERT WILLOW
EL CENTRO
CALIFORNIA
BELOW SEA LEVEL
YUMA
JOJOBA
PRICKLY PEAR
PHOENIX
TUCSON
ARIZONA
COLORADO RIVER
MEXICO
SIDE VIEW

Cross section of Arizona Desert by Willie Whitefeather

From these people came the desert Pimas, Navajos, and Hopis of the Southwest. Even today wild foods are a large part of the Papago Indian's diet, which includes milkweed, bulrushes, cattails, phragmities, amaranth, sumac, asters, thistle, sunflower, wild lettuce, chickweed, mustard, shepherd's purse, mint, plantain, sheep sorrel, purslane, willow, and rose (Hodgson 1982).

Just as Native Americans followed the patterns set by their environment, I found myself doing the things that nature made most obvious for me. I also studied the ways of people who had come before me.

Some specific foraging heritages I learned about are:

- Abanakee (People of the Bark)-eaters of birch, balsam, and willow
- Algonquin (People of the Coast)-eaters of amaranth and lamb's quarters
- Blackfoot, Chippewa (People of Canada)-rice eaters
- Cayuga (People at the Landing)-vegetable eaters
- Oneida (People of the Standing Rock)-vegetable eaters
- Objibwa (People above the Climate Line)-rice eaters
- Iroquois, Eastern Mohawk (People of the Flint & Keepers of the Eastern Door)-bark eaters
- West Seneca (People of the Great Hill & Keepers of the Central Fire)-bark and vegetable eaters

The point is, no matter where you live, you are going to draw upon the heritage of those who went before you. We can learn an immense amount from Native Americans. Climate, terrain, existing plant life, and economical use were all considerations of native peoples. The same plant might be used quite differently in different regions. To make use of the wild foods in your area, study your local Native American heritage. You'll find the information in local museums and libraries. Also, native teachers are often available, reached via earth gatherings, through organizations, at powwows, and such.

A Basic Food Source

Although some wild foods are best utilized in one particular way, such as steeped for tea, others have multiple uses and applications. In fact, it sometimes seems as if a single plant can meet just about all our needs. Such is the case with clover. We can eat it fresh or raw, or we can cook it, dry it, freeze it, or grind it into flour and bake it.

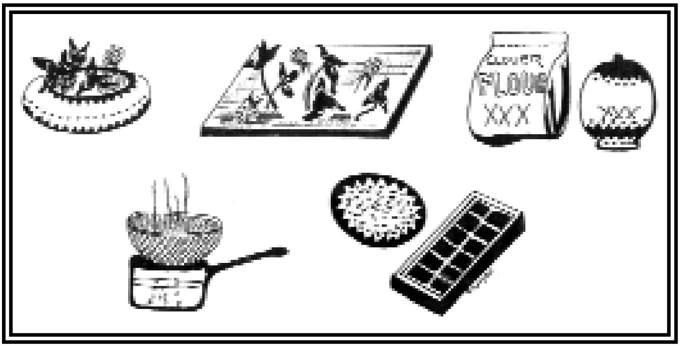

Clover: raw, cooked, dried, frozen or ground into flower.

The same can be said for amaranth, aster, chickweed, dandelion, dock (with some restrictions), lamb’s quarters, mallow, nettles, plantain, purslane, shepherd’s purse (with some restrictions), sow thistle, tumbleweed, and wild lettuce.

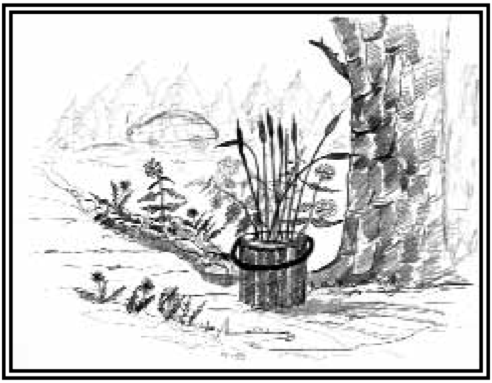

Gathering

The chart on the following pages delineates the uses of the wild foods discussed in this book.

Properties Of Wild Edible Foods

	Type			Origin			Edible Parts							Storage		Uses				
	Annual	Biennial	Perennial	Native to N.A.	Naturalized	Harvest Time	Leaves/needles	Stems	Buds	Flowers/Catkins	Fruit	Seeds	Roots	Freezes	Dries	Vegetable/Fruit	Herb	Tea	Flour	Field Guide Page
Aloe Vera					❂	Spring/Summer	❂		❂	❂				❂		❂	❂	❂		52
Amaranth	❂			❂		Summer	❂	❂		❂		❂		❂	❂	❂			❂	54
Arrowhead			❂	❂		Fall							❂	❂	❂	❂				55
Aster			❂	❂		Fall	❂	❂	❂	❂		❂		❂	❂	❂		❂	❂	57
Balsam Fir				❂		Year-round	❂	❂		❂		❂	❂		❂	❂	❂	❂	❂	58
Birch				❂		Year-round		❂	❂	❂		❂			❂	❂	❂	❂	❂	60
Blackberry			❂	❂		Year-round		❂		❂	❂	❂		❂	❂	❂		❂		62
Blueberry			❂	❂		Summer	❂	❂			❂	❂		❂	❂	❂		❂		64
Bulrush			❂	❂		Summer/Fall		❂	❂	❂		❂			❂	❂			❂	65
Burdock		❂			❂	Year-round	❂	❂			❂	❂	❂	❂	❂	❂	❂	❂	❂	67
Cattail			❂	❂		Year-round	❂	❂	❂	❂		❂	❂	❂		❂		❂	❂	69
Chamomile	❂			❂	❂	Spring/Summer	❂	❂	❂	❂		❂		❂	❂	❂	❂	❂		72
Chickweed	❂			❂	❂	Year-round	❂	❂	❂	❂		❂		❂	❂	❂	❂	❂		73
Chicory			❂		❂	Spring	❂	❂		❂		❂	❂	❂	❂	❂	❂	❂		75
Cholla			❂	❂		May			❂	❂	❂	❂			❂	❂				76
Clover			❂		❂	Year-round	❂	❂	❂	❂		❂	❂	❂	❂	❂	❂	❂	❂	78
Crabgrass			❂		❂	Year-round	❂	❂	❂	❂		❂			❂			❂	❂	80
Daisy			❂		❂	Year-round	❂	❂		❂				❂	❂	❂		❂	❂	81
Dandelion			❂		❂	Year-round	❂		❂	❂		❂	❂	❂	❂	❂	❂	❂	❂	83
Dock			❂		❂	Spring/Fall	❂	❂	❂	❂		❂	❂	❂	❂		❂	❂	❂	85
Evening Primrose		❂		❂		Summer/Fall			❂	❂		❂	❂	❂	❂	❂	❂			86
Filarie		❂				Winter/Spring	❂	❂					❂	❂	❂	❂		❂		87
Fireweed			❂	❂		Spring/Summer	❂	❂	❂	❂		❂		❂	❂		❂	❂		89
Goldenrod			❂	❂		Spring/Sum/Fall	❂	❂	❂	❂		❂	❂		❂		❂	❂		90
Grape			❂	❂		Fall	❂	❂		❂	❂	❂		❂	❂	❂				92

Lamb's Quarters	❂			❂		Summer/Fall	❂	❂	❂	❂		❂		❂	❂	❂			❂	93
Malva			❂	❂		Winter/Spring	❂	❂	❂	❂		❂	❂	❂	❂	❂	❂	❂	❂	95
Maple			❂	❂		Summer	❂	❂	❂	❂		❂		❂	❂			❂	❂	97
Meadowsweet			❂	❂		Summer	❂	❂	❂	❂		❂			❂		❂	❂		98
Milk Thistle		❂		❂		Spring/Summer	❂	❂		❂	❂	❂		❂	❂	❂	❂	❂	❂	99
Milkweed			❂	❂		Spring/Sum/Fall	❂	❂	❂	❂		❂		❂	❂	❂				101
Mint			❂	❂		Spring/Fall	❂	❂	❂	❂		❂	❂	❂	❂		❂	❂		103
Mullein		❂			❂	Spring/Fall	❂	❂	❂	❂		❂	❂		❂		❂	❂		105
Mustard	❂				❂	Spring	❂		❂	❂		❂		❂	❂	❂	❂			107
Nettles			❂		❂	Spring/Fall	❂	❂		❂		❂		❂	❂	❂	❂	❂		108
Phragmities			❂	❂		Summer		❂		❂		❂	❂		❂	❂			❂	109
Pine						Year-round	❂	❂	❂	❂	❂	❂	❂		❂			❂		110
Plantain			❂	❂		Year-round	❂	❂				❂	❂	❂	❂	❂	❂	❂		111
Prickly Pear			❂			Summer	❂		❂	❂	❂	❂		❂	❂	❂		❂		113
Purslane	❂				❂	Summer	❂	❂	❂	❂		❂	❂	❂	❂	❂			❂	115
Queen Anne's Lace		❂			❂	Summer	❂	❂	❂	❂		❂	❂	❂	❂	❂	❂	❂		116
Raspberry			❂	❂	❂	Summer	❂	❂	❂	❂	❂	❂		❂	❂	❂	❂	❂		117
Rose			❂	❂	❂	Spring/Fall	❂	❂	❂	❂	❂	❂		❂	❂	❂	❂	❂		118
Saguaro				❂		Summer			❂	❂	❂	❂		❂	❂	❂				120
Sheep (Garden) Sorrel	❂					Spring/Summer	❂	❂	❂	❂		❂	❂	❂	❂	❂	❂	❂		121
Shepherd's Purse	❂				❂	Summer	❂	❂	❂	❂		❂	❂	❂	❂	❂	❂			122
Sow Thistle	❂				❂	Winter/Spring	❂	❂	❂	❂		❂	❂	❂	❂	❂			❂	124
Strawberry			❂	❂		Spring	❂	❂	❂	❂	❂	❂	❂	❂	❂	❂	❂	❂	❂	125
Sumac			❂	❂		Summer				❂	❂	❂			❂			❂		126
Sunflower	❂			❂		Spring/Summer		❂		❂		❂			❂			❂		127
Thistle		❂		❂		Spring/Fall	❂			❂		❂	❂	❂	❂	❂		❂		128
Thyme			❂		❂	Summer/Fall	❂		❂	❂		❂		❂	❂		❂	❂		129
Tumbleweed	❂				❂	Spring/Summer	❂	❂	❂	❂		❂		❂	❂	❂		❂	❂	131
Violet	❂					Spring/Summer	❂	❂	❂	❂		❂	❂	❂	❂		❂	❂		132
Wild Lettuce			❂	❂		Spring/Summer	❂	❂	❂	❂		❂		❂	❂	❂	❂	❂		133
Willow				❂		Year-round	❂	❂	❂	❂		❂			❂		❂	❂	❂	134
Wintergreen			❂	❂		Spring/Summer	❂	❂		❂		❂			❂			❂		135
Wood Sorrel			❂		❂	Spring/Fall	❂	❂	❂	❂		❂		❂	❂	❂				137
Yarrow			❂		❂	Spring/Fall	❂	❂	❂	❂		❂		❂	❂		❂	❂		138

Picking Wild Foods

State and local regulations on picking wild plants vary, but mostly there are restrictions against taking plants or plant parts on public land. You should always check with park officials before foraging anywhere but on your own land.

In gathering plants, pick only on lands off non-posted access roads or from fields with owner's permission. Move from place to place, picking sparingly and only when there is several of any one plant. In addition: due to nature's changes and people's advances, many plants are now endangered. Even though this guide's plants are not endangered now, they may be in the future. Check your Local State library or Department of Agriculture for updates. Do not pick any endangered plants on private property.

FORAGING RULES

1. ***DO NOT*** collect plants closer than 200 feet from a road.
2. ***NEVER*** collect from areas sprayed with herbicides, pesticides, or other chemicals.
3. ***ALWAYS*** be familiar with all dangerous plants in the area.
4. ***POSITIVELY IDENTIFY*** all plants you intend to use for food or medicine. Check against three good field references with excellent illustrations.

If you are trying a plant for the first time, after checking your field guides, do the following:

1. Snip a piece of the plant and roll between your fingers and sniff. Discard if objectionable. If you like the smell, then rub the tiny piece on your GUMS, above your teeth.
2. Wait twenty minutes.
3. CHECK For burning, nausea, stinging, itching (all allergy results). Poisonous plants USUALLY produce one or more of these symptoms.
4. If no untoward reaction results, take another tiny bit of the plant and make a weak tea. (Place piece in teacup, pour boiling water over, cover, and steep for 10 minutes. Ingest a small amount.)

5. Wait another twenty minutes! Check for signs of irritation. If none, then reheat the tea and sip slowly.
6. Keep all samples away from children and pets. Keep edibles separate from samples to be identified—poisons will give their bad qualities to food through contamination. Bag foods, poisons, and samples separately.
7. Be aware that heating or boiling does not always destroy toxicity.

In general, it's a good idea to store all seeds and bulbs away from children and pets. Tape down the lids or use a locked closet for any questionable foods so children cannot reach them. Likewise, teach children to keep all plants away from their mouths. Do not let children chew on or suck nectar from unknown plants.

Avoid smoke from burning plants. Smoke may irritate the eyes or cause allergic reactions. Also, be aware of your neighbor's habits with regard to chemicals, pesticides, and herbicides. Ask questions! Call and report any chemical spills, contaminated areas, or other situations that may negatively affect your gathering safety.

Growing up with Wild Foods

"As a child, I always mashed and ate wild plants. This worried my Daddy. He began to teach us what not to eat and what good plants to enjoy. He showed us how to pick flowers from the top of ocotillo plants while on horseback. We sucked the honey from the bottom of each flower, much as Easterners used honeysuckle. The ocotillo flowers are just delicious. Our parents ran a cow ranch in Mohave County for fifty years, and we picked our ocotillo flowers often from horseback. A lovely life!"—Danielle Stephens, Kingman, Arizona.

Gathering

Decide which plant you will harvest on a particular day and take along a basket, a bag, or a sheet. Burlap is excellent if you are going to dry the plants, but biodegradable plastic bags generally are better for gathering because burlap gets heavy when wet.

Do any aquatic plant gathering in the morning, as the cool air and water will keep your collection fresher. Plastic bags are waterproof and lightweight, but remember to use biodegradable bags, and reuse them as often as possible. (Do not leave wild food in the sun, especially inside a plastic bag, because the food will wilt quickly; I have spoiled hundreds of milkweed buds this way!)

A flat, round tray also works well for gathering in tall grass, especially when gathering fireweed. It can be great fun wading through a field of these magenta flowers, snapping off the tips and tossing them onto a tray held above your head.

For most gathering, I use a clean sheet and a two-foot section of string. A small amount of wild food will fit in your basket or bag, while a winter's supply will fill a clean sheet. When the sheet is folded up and tied with a string, it can be hung up easily in a warm, airy spot to dry. The sheet method of collection also allows the bugs to crawl out of the food and provides an easy view of how much you have gathered.

When gathering wild foods, remember that certain plants snap off easily while others are better cut cleanly with a knife or scissors. If the plant gives you any resistance to a snap, give way to a tool immediately. A clean cut will help the plant recover quicker and you'll have greater gathering speed and accuracy.

Cutting a batch of leaves on the stem is simple when you hold the stem with one hand and with the other cut the stem with a sharp knife. When I gather with a penknife or scissors, I slide my fingers into the middle of the clump and go for inside growth. The plant then folds up around what I took and rejuvenates nicely. Goldenrod, asters, yarrow, and Queen Anne's lace work especially well this way.

If you're using the sheet method, you may want to strip the leaves off the stems while gathering for ease in drying. At this time you can also separate your favorite flowers for tea. Don't forget to save the stems, though, as they can be valuable, especially mint, spearmint, yarrow, and thyme.

When you've gathered enough, pull the corners of the sheet together, and use the twine to tie it closed. I often hang the bundle from a tree or other obvious spot and retrieve it later. This method also helps mark the trail. (I use a cactus in the desert since trees are sparse.)

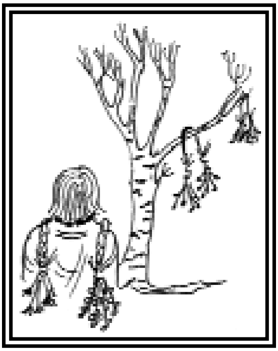

A Food Tree

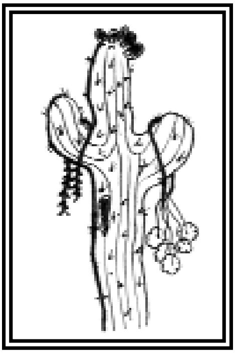

Saguaro as a food tree

Tie enough stems together to make a bunch, but be sure the leaves aren't packed too tight, lest you prevent air penetration. As a rule, when you swish the bundle in the air you should "feel" and hear the air moving through it.

Waists and belt loops also work well for carrying bundles of plants. You can also use burdock leaves or mullein to roll up several smaller items. Roll up nuts inside large leaves and put them in your pockets as you go. Fruits, twigs, nuts and flowers all wrap well in large leaves.

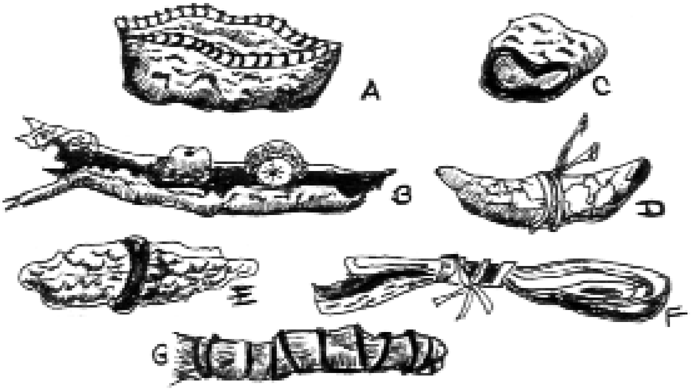

Leaf Bundles: A) White birch bark waterproof container, B) A bark tray, C) Hollow bark, D) Burdock leaf wrapping for small foods, E) Mullein leaf wrapping for small objects, F) Kit of white birch strings, and G) Bundle of foods, grasses or twigs.

In summary, here is a list of tools I generally use for gathering:

- A sheet
- Scissors
- 2-foot length of string
- Biodegradable plastic bags
- Penknife and Bowie knife
- Gloves
- Snake stick with fluorescent handle (in the West)
- Hand axe (in the East)
- Burlap bag

Washing the Foods

Certain plants attract aphids, ants, and “no-see-ums,” as well as collect the usual dirt and dust. In the Adirondacks a nearby stream is always available, so washing large quantities of plants is always easy, no matter how dirty they are. But in Arizona, apartment living demands the use of a bathtub—it’s the easiest way to wash a large quantity of food. Of course, it’s always best to use a hose first, especially for roots, eliminating most of the soil before taking them into the house.

When cleaning small plants or bunches of leaves that aren’t too dirty, it’s a good idea to use a drainable container such as a colander. Some containers can be made to drain by poking holes in the bottom. Some plants may need more soaking, however, and may even need to soak overnight. In this case, soak in 1/4 cup vinegar to a tubful of water. In any case, it’s best not to put too many types of plants in the same tub, since you may have difficulty separating them later.

Finally, here’s a quick tip that will make it easier to clean your tub and prevent a clogged drain: Before filling the tub, place a piece of cheesecloth over a rubber stopper and use it as the plug. When you’re ready to drain the tub, pull the stopper, but hold the cheesecloth in place. When the water is drained, just shake out the cheesecloth and save it for the next time!

Edible Wild Flowers

The following are common wild flowers I utilize often.

Birch catkins—unopened catkins eaten raw or brewed as tea. Dried or frozen for storage. Also cooked.

Blue Asters—All wild asters and flower buds may be eaten raw, cooked or dried and ground for flour.

Cattails—Flowers and flower buds are eaten raw, cooked, or dried and ground for flour. Used in tea.

Chamomile—White flowers and flower buds may be eaten raw, cooked, steeped for tea, or dried for storage.

Clover, Red or White—Flowers and flower buds are eaten raw, cooked, dried for tea, & frozen in ice cubes.

Dandelions—Whole flower and flower buds may be eaten raw in salads, stir-fried in garlic, frozen for use later. Used in tea.

Fireweed—Whole flower and flower buds are eaten raw, cooked, or dried for storage.

Goldenrod—Whole flower and flower buds are eaten raw as a tonic, and dried for tea. Frozen for storage; used in tea.

Maple—Flower buds are eaten raw, cooked, or dried for tea and storage. Can also be frozen.

Meadowsweet—Flowers and buds are eaten raw, cooked, or dried for a sugar substitute or tea.

Milkweed—Eastern flower is cooked as a vegetable, or frozen or dried for storage. Western varieties are caustic from desert sand.

Mint—Flowers and flower buds are eaten raw or cooked, or dried for storage; frozen in ice cubes; used in tea.

Mustard—Flowers and flower buds vary in intensity; may be cooked or eaten raw, and frozen.

Pine—Pine flowers or young cones are called catkins and are eaten raw, cooked, and dried for tea; also frozen.

Prickly Pear—Flowers are eaten raw in salads, and cooked; frozen for storage.

Queen Anne's Lace—Flowers and buds are eaten raw, cooked, frozen, and dried for storage.

Roses—Flowers and flower buds are eaten raw in salads, cooked, also frozen in cubes for storage.

Thyme—Flowers and buds are eaten raw, cooked, or dried for tea; also frozen.

Violets—Flowers and flower buds are eaten raw in salads, cooked, and dried or frozen for storage and tea.

Wintercress—Flowers and buds are eaten raw, cooked, and dried.

For information sections on drying and freezing flowers, see pages 33, 37, 39, 42 and 43 respectively.

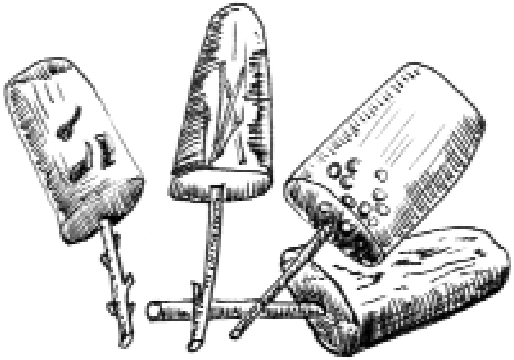

Wild Popsicles

Edible Roots

During my homesteading days I became interested in the many uses of plant roots. A particular experience gave me a lasting respect for the roots of wild plants.

In mid-winter, the ground was frozen and a snow-encrusted world was our environment. My son Todd had developed a bad cough. Soon Todd's condition worsened and he was coughing steadily. He needed more than the usual remedy of pine tea and honey. I knew about the antihistamine properties of mullein root. I could see the tall brown spikes of mullein poking above the snow like sentries standing beyond our camp. But the temperature was 20 degrees below zero and the ground was frozen solid.

The woodstove was filled with hardwood coals. I collected a pail of coals, put on my snowshoes, and went out to make a path to the closest mullein plant I could find. I scraped away as much snow as I could from around the plant and poured the hot coals on the icy crust at its base. (A very heavy pail is necessary for this, as coals will burn through a thin pail very quickly.) Sizzle, sizzle. The coals sunk quickly into the crusty ice. I returned to the house and fetched more coals. Pausing a few minutes between pailfuls, I scraped away each layer of slush until I reached mud. I placed a circle of dry wood around the base of the plant and made a large campfire inside the circle. With the wood concentrating the heat at the plant's base, the fire became hot in no time. I used birch bark to keep it very hot, adding a large sheet or two at a time. We were lucky to live next to a sawmill that cut white birch logs, and sheets of bark were plentiful. Never cut birch strips off a live tree.

After a long, cold hour, I scraped the remaining coals out from the circle and easily dug out a foot-long mullein root. It was partly decomposed, but there was more than enough root to brew a concentrated tea for Todd. He drank 2 tablespoons every hour, and his cough was calmed. Mullein root tea is an effective expectorant, helping Todd discharge mucus from his throat and bronchial tubes. This wintertime experience gave me untold security, knowing that mullein roots were out there if we needed them for medicine or food. This “underground farm” took on new meaning: it was available to us even through several feet of snow and in 20 below temperature.

The following edible roots are ones I use:
Amaranth—Used as a vegetable; dried & ground for flour.
Arrowhead—Tuber used as potatoes.
Balsam fir—Used for medicinal purposes (sedative), & tonic.
Burdock—Used as a vegetable, dried and ground for flour.
Cattail—Dried and ground for flour.
Chicory—Used as a coffee substitute.
Dandelion—Used as a vegetable and tonic.
Evening Primrose—Used as a vegetable flour.
Fireweed—Used as a medicine and a food.
Goldenrod—Used for a blood tonic tea.
Malva neglecta—Used as vegetable; dried & ground for flour.
Pine—Used for tonic and stimulant as well as rope making.
Plantain—Used for vegetable and as an antiseptic
Purslane—Used as a vegetable; dried and ground for flour.
Queen Anne’s Lace—Used as a vegetable and dried food.
Strawberry—Used as a vegetable, a medicine and as an appetite stimulant.
Wild lettuce—Used as a vegetable.

Plant roots are greatly under utilized, and their magnificent properties barely understood. For example, many are aromatics.

Roses, for instance, have an aromatic root that can be reused many times, until the scent and color are gone. Scrub off any mold or black spots before use. Mullein root can be reused many times for tea, as can mint roots.

Clip newly gathered roots from the plant and scrub with a toothbrush. Then use fresh, or dry and store.

To use fresh, peel the root as you would a carrot or parsnip and slice crosswise or lengthwise. To dry the roots, place in a paper bag and label carefully with the name of the plant and the year you collected. Hang to dry. Reconstitute in boiling water or place in a grinder and grind to flour.

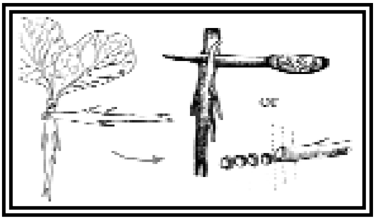

Harvesting roots: peel & slice either crosswise or lengthwise.

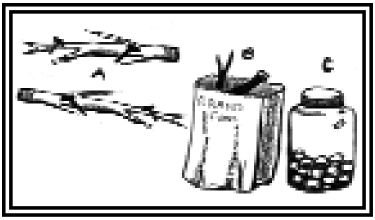

Storing roots: A) hang on wall B) dry & store in paper bag C) store dried slices in glass jars

Edible Parts of Trees and Shrubs

Many trees and shrubs have edible parts and can be a major source of food in the wilderness. The chemical content, vitamins and minerals are found in leaves and flowers, but in concentrated form in shrubs and tree bark. Inner bark, the layer underneath bark, is the “meat” of a tree. Twigs have both bark and inner bark. For example, pine or balsam fir bark strips, cut 1/3 inch to 1 inch wide, to fill a 2-quart glass container will yield a winter’s supply of tea. One small shrub, such as a rose bush, can yield a winter’s supply of tea.

In calculating your yield, figure one twig, 6 to 8 inches long, yields a moderately strong cup of tea. This tea could be used as a soup base. To prepare for cooking or tea, break the twig several times to release the flavor and steep or boil as intended. To store twigs and stems, place in dark paper bag. Or, place in glass jar and cap tightly. Wrap with opaque paper or foil to keep it dark. And, of course, label with the collection date.

The following diagram shows two different methods of peeling bark. Method A shows types of plants which will allow big pieces to be peeled. Method B makes smaller wispy pieces by first splitting along the length, then peeling small pieces around the girth from one side of the split to the other.

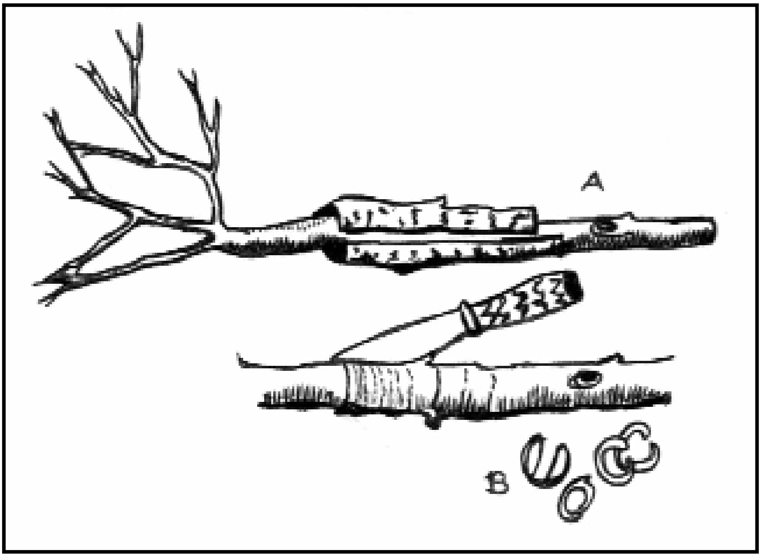

Peeling bark in strips.

The following are trees and shrubs and commonly harvested for their edible parts:

- Balsam fir—Parts of the entire tree are edible, especially the inner bark, sap, twigs, needles, and flower buds and seeds.
- Birch—Entire tree is high in mineral content and medicinal value; sap, buds, catkins and seeds are edible. Birch catkins, buds, twigs and bark, as well as inner bark, are all loaded with trace minerals, and beta-carotene.
- Blackberry—Inner bark, twigs, flowers, fruits and seeds are edible.
- Blueberry—The twigs, bark and roots are edible as well as fruits.
- Maple—Entire tree is edible, especially inner bark, sap, twigs, buds, young leaves, and seeds.
- Meadowsweet—Entire plant is high in minerals and medicinal value; inner bark, twigs, buds and flowers, leaves and seeds are not to be eaten in bulk.

- Pine—Entire tree is edible; inner bark, sap, twigs, needles, catkins, and in one species, Pinon pine nuts.
- Prickly pear—Buds, flowers, pads and fruits of plant are edible.
- Raspberry—Entire shrub is edible, including inner bark, sap, twigs, buds, flowers, fruits and seeds.
- Rose—Entire shrub is edible, especially inner bark, sap, twigs, buds and flowers, leaves and seeds (hips).
- Saguaro—Buds, flowers and fruits, and seeds are edible.
- Sumac—Flowers, fruits and seeds are edible.
- Tumbleweed—Twigs, buds, and leaves are edible.
- Willow—Entire tree has high mineral content or medicinal value and should not be eaten in bulk. Inner bark, sap, twigs, buds and catkins, leaves and seeds are eaten in moderation.

Teas from Wild Plants

Gathering and preparing teas from wild plants is simple and rewarding. For example, a large supply of thyme for tea can be harvested in just two minutes with scissors. Drying requires no more than a screen and takes but a few hours. Using the back shelf of my car, I can have it dried and ready for storage in a few hours. In a few days, the patch grows back and I can harvest it again. This is true for most tea plants.

Plants commonly gathered for tea include:

- balsam fir
- birch
- blackberry
- blueberry
- chamomile
- chicory
- clover
- dandelion
- fireweed
- goldenrod
- meadowsweet
- mint
- mullein
- raspberry
- strawberry
- thyme
- violet
- willow
- wintergreen.

(See Part II, the Field Guide, for information on individual plants.)

Picking Nettles

Do you want to try nettles but are afraid to pick them? Take your gloves, a pair of scissors, and a container and head for the nettle patch. Cut the leaves directly into the container, holding the plant with your gloves. Or, to take home the entire nettle plant, cut two individuals of a plant that is long and straight, such as goldenrod. Lay them on the ground parallel like railroad tracks. Cut the nettle, using your gloves and scissors, pile them across the top. Then pull the goldenrod stems up and fold over so you can carry your bundle home without getting pricked by stickers.

Storing Teas

You might enjoy fresh tea brewed from raw materials. Remember that raw material is a bit less effective and tastes weaker than dried.

Wild plants or leaves that have dried until crisp can easily be stripped clean for storage. Place a clean sheet on a counter and put bundles on top. Using gloved hands, pull the dried leaves, seeds, or flowers down gently, then scoop up and crumble by twisting the sheet and wringing it. This is what is known as a rough grind. In some cases you can obtain the same results by using your hands. Store rough ground teas in glass jars.

You may store twigs in “tea sizes”, such as six inch sections of willow equals approximately 2 aspirin. Six inches of a rose stem equals 1 cup of tea. Four inches of goldenrod stem equals one cup of tea. If you use a mayonnaise or quart size jar many cups of tea are easily stored.

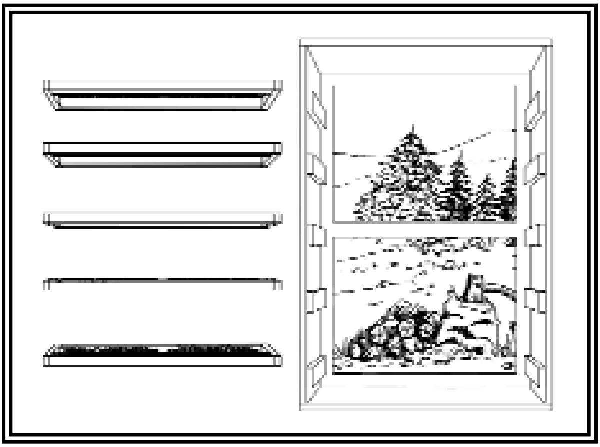

Window food dryer with removable racks.

Flour from Wild Grains

Bulrushes and phragmities grow prolifically in wet lands. They are a familiar sight to most of us, yet few people realize what a valuable food source they are. In both the East and West, I have gathered and ground bulrushes to make flour. The flour makes delicious muffins, cookies, and breads. What's more, there are common grasses growing in most people's lawns that can likewise be harvested for flour. My favorites are listed on page 35.

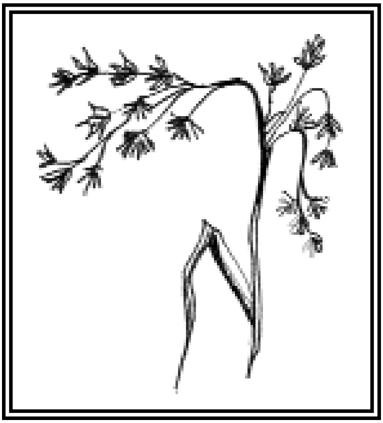
Bullrush

Gathering bulrush stalks is an easy task, so long as the water is clean. Know where the water comes from and look carefully at the stems as you cut them. It is a good idea to wear gloves until you are well used to spotting the stems of the rushes. Poison hemlock may be in the same marsh, so be careful! (see Part II, Field Guide, and Part V, Poisonous Look-Alikes).

Cut one of the two stems of a bulrush plant at a time, sliding your gloved finger down the stem, cutting, and placing the rush up and under your armpit. As you move along, the batch of plants under your arm will grow. When you have enough, stop and tie them into a bundle. Hang the bundle in a tree or make a yoke for around your shoulders. Pick from the centers of the bulrush clumps, so the plant will grow back quickly.

Dry bulrush in an attic, car, or other warm and dry area. Let dry thoroughly. They may appear to be dry, but check. The stems break, crack or snap easily when dry. Place in glass containers until you can grind the grain to flour.

When you are ready to grind to flour, check bulrushes for tiny black dots. These may represent ergot, a poisonous fungus. If present, discard the plants.

Each bulrush stem yields several teaspoons of flour, and a screen full of stems and flowers yields approximately 1 cup of finely ground flour. The color of the flour depends on the bulrush: it will be green if bulrushes are picked early in the season, tan if picked later.

> *"He causeth the grass to grow for the cattle, and the herbs for the service to man; that he may bring forth food out of the earth."*
> Psalm 104:14
>
> *"Also, to every beast of the earth, to every bird of the air, and to everything that creeps on the earth, in which there is life, I have given every green herb for food, and it was so."*
> Genesis 1:30

Harvesting Grasses

The following are common lawn grasses that make excellent flour. They are also known as Graminae, since they are in that family of plants. Since these plants are not described in Part II, check identification in several other reliable field guides. The grains may also be cooked whole for granola, cereal, and other uses.

Barnyard grass— *Echinochloa crus-galli, Graminae* Family
Barley grass— *Hordeum pusillum nott, Graminae* Family
Broomsedge— *Andropogon virginicus, Graminae* Family
Crabgrass— *Digitaria sanguinalis, Graminae* Family
Foxtail grass— *Setaria italica, Graminae* Family
Goosegrass— *Eleusine indica, Graminae* Family
Jungle grass— *Echinochloa colonum, Graminae* Family
Quackgrass— *Agropyron repens, Graminae* Family
Rye grass— *Lolium temulentum, Graminae* Family
Wild oats— *Avena fatua, A. barbara, Graminae* Family
Wild rice— *Zizania palustris, Graminae* Family
Wheat grass— *Triticum aestiuum, Graminae* Family
Yellow nutsedge—*Cyperus esculentus, Graminae* Family

Some grains are better sheared with scissors, such as crabgrass. Others are easily collected using a pillow case. If the seeds come off the stem easily, just put the pillow case over the tops, bend the grasses over, and wring the seeds off. This method is especially good for marsh grasses. I've happily spent many an evening gathering a quart or two of seed heads while watching the dragonflies in the setting sun.

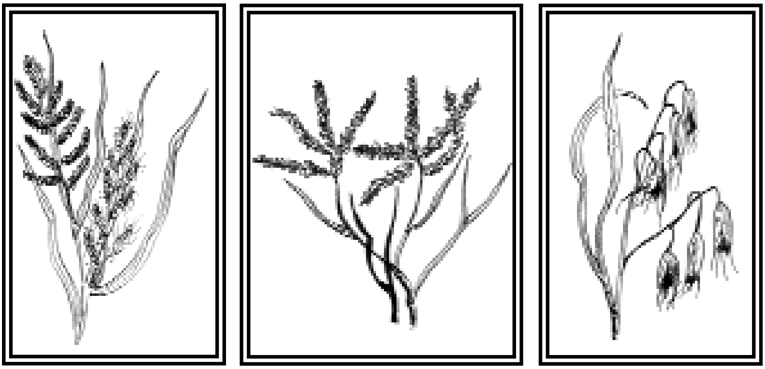

Barnyard Grass, Goose Grass & Wild Oats

After the grains are dried, they can be ground to flour. Use your own method; grinders, processors, and mortar and pestle are all successful.

I usually clip stems to six inches from the grain. Grinding a bit of the stem adds to the fiber content.

Grains contain protein, and protein is one of the building blocks of life. Unrefined grains are best because the natural elements have not been processed out. Much of wheat's protein and vitamins E and B^6 are lost in the refining process. As a result we have "enriched" bread, which tries to make up for the loss. Isn't it a better idea to keep those nutrients to begin with? This is why whole wheat flour and whole natural grains are the wise nutritional choice today. Also, fiber is lost in the processing of wheat flour. Yet the higher the fiber content of our food, the better for our digestive system. Unrefined, natural grains are truly the only answer to better health.

Blend one-half cup of assorted wild grains putting them into a wide-mouth thermos and then pour boiling water over them. Cap tightly and let sit overnight to create an instant hot cereal for the morning or the ingredients for a rough bread dough or wild "pancake." Add a little vegetable flour such as amaranth or clover for a high-energy meal!

Preserving Nature's Bounty

Drying for tea or grinding to flour, as the preceding sections describe, are major ways to preserve and enjoy wild food, but there are other possibilities as well. You can collect roots, branches, needles, and flowers in season and either dry them or freeze them for later use in a variety of ways.

In the Adirondacks, my pantry was full of wild foods preserved for winter. The assortment varied from year to year, and when I moved West, there were many substitutes. Tumbleweed took the place of fireweed. Malva neglecta was used instead of arrowhead for flour. The caption to the illustration on the next page tells what wild foods I commonly had in my pantry, East and West.

Typical Pantry Storage

Eastern cupboards were always filled with certain staples. I did not own a freezer in those years. All staples were dried.

Plants hanging in storage pantry.

Car drying racks.

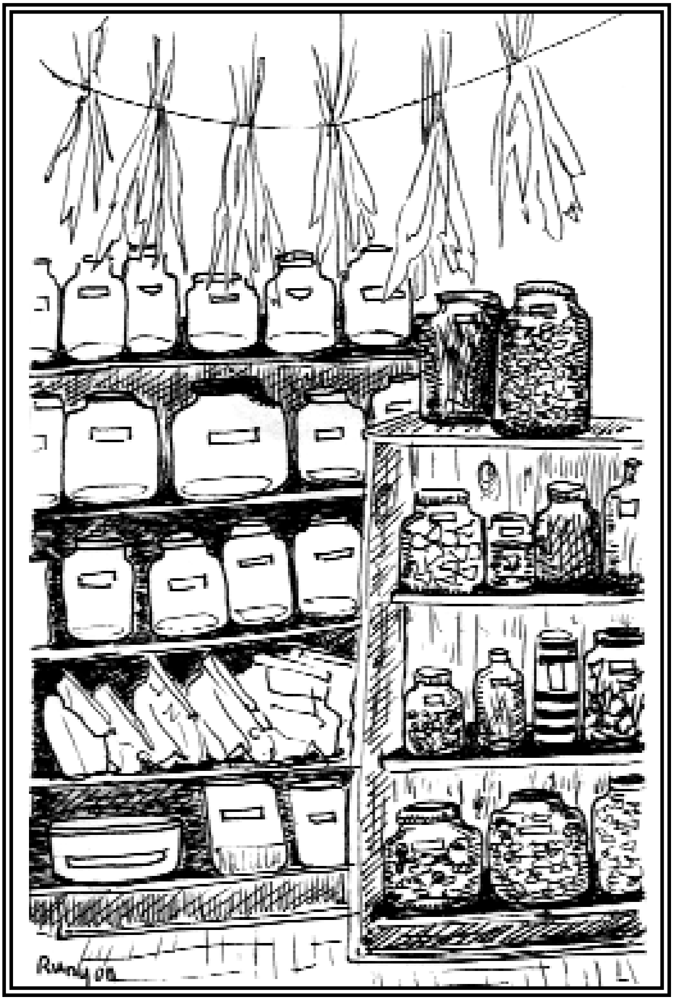

The Pantry

Gallon jars of leaves included: Amaranth, asters, lamb’s quarter, strawberry, clover, red and white clover, chicory, daisy leaves, dandelion, dock & malva neglecta. Western jars also included tumbleweed, mesquite beans, palo verde beans & chaparral.
Gallon jars of twigs for teas included: Rose, meadowsweet, pine, balsam, birch, maple & willow, and, in the West, ephedra (Mormon tea).
Separate jars of dried flowers: Aster, clover, chamomile, chicory, dandelion and rose, both in the East and West.

My closets always had a few choice branches stored whole: pine, balsam, willow, and maple. Closets in both East and West are filled with paper cartons of grasses for future use as flour and, of course, the Christmas tree and wreath from year to year!

Preparing Food for Storage

Once you've gathered your bounty, shake off any soil and wash plants, leaves or roots gently under running water. Wash all foods except cattail pollen. Scrub roots with a toothbrush.

Trim off any dead leaves or parts. Remove and separate leaves, seeds, flowers, and roots for separate drying; or leave plant whole (as you might dry lamb's quarters or amaranth for flour). Cut stems to storage-container lengths.

Drying Methods

There are a few basic "don'ts" about drying food that you should know before you begin. Don't dry leaves in direct sunlight; find a warm, airy place out of the sun. Never dry foods near a road or garage, or you'll taste exhaust fumes. Lastly, don't dry pine needles in an oven or microwave. They are very volatile and may explode.

The best indoor drying apparatus I know is a window screen, convenient for either an apartment or a home. You simply bring home your favorite wild foods, wash them well, trim, and shake off excess water. Lay the foods on the screens, keeping any branches apart somewhat. Slide the screen into bracket and start the next screen. (See page 33 for an illustration of screen assembly.)

When you feel that food is as dry as it can be, preheat the oven to 300 degrees and turn off. Heat the foods for five minutes, then put in hot clean jars, or put the gatherings in the jar, without the lid and heat jar, foods, and all for five minutes at 300 degrees. Often I simply turn the oven to 300 for a while, put the food in, and turn off the oven, leaving it to cool slowly. I take the jar out when the oven is partly cool, then cap tightly for storage.

There is another method. Although you may feel odd driving around town with a rack of "weeds" in the back of your car, it is by far the most efficient drying rack! In July and August, one day on the back window shelf of your car will thoroughly dry most leaves. Just think, while spending money driving around town, you are performing an energy-saving task and making a step toward long-term food storage, ultimately saving dollars. The car doesn't have to be in the hot sun all the time. Park in the shade if you wish.

Caution: Don't dry foods in a car when the outside temperature is over 80 degrees—toxic fumes may arise from plastic or vinyl seats.

The Drying Roof

There are more ways to dry wild plants. Here's a simple method I used.

Looking around for a flat place to put the screens I used to dry plants on, I noticed that the roof seemed to take up a lot of space—space that didn't do much other than keep the rain out. So I climbed a ladder and began to place the screens on the roof. Since there were poplars surrounding the cabin, the shade was perfect for drying apples and berries. We even nailed on boards as scaffolding to create a large accessible drying space. You may not want to go this far, but with a little ingenuity you'll find almost any place can be used to dry fruits, berries, and such.

We dried several types of wild apples. We did this every year using the larger varieties, choosing about half the normal size apple. We also dried crabapples on the roof. We gathered the crabapples by spreading a sheet under the tree and shaking the limbs, making fruit fall. We cored the apples and then sliced them crosswise about 1/3 inch thick. We threaded these slices on old broom handles, then hoisted them up to the roof. After a few days, the crabapples had shriveled and became leathery. After a 5-minute baking in the oven to dry them completely (oven preheated to 300 F. then turned off), we stored them in glass jars. In winter we used them in everything from cereal to cobblers, pies, and snacks.

Linda says- You can also string the dried apple slices together with a needle and string. Not unlike the garlic and hot peppers in the Southwest, your wild foods can decorate your kitchen. I even string roots this way.

Label the jars with the date of drying before storing. The storage place is important. Closets work best. You need darkened spaces for dried foods, out of direct sunlight.

Store whole leaves and grind them into flour as you need it. For long-term storage, grind and place the flour in a shallow 1-inch pan and toast for five to ten minutes in a preheated 300 degree oven turned off. Store in glass for up to several weeks. Vegetable flour does not have a long life, so avoid storing for more than six months at a time.

Rinse dried branches to clean off any dust before you use them. Rose or raspberry vines, pine or balsam branches, and weeping willow can be stored for years, but will accumulate dust.

Containers for Dried Foods

Generally, I'll save just about any glass jar and use it to store dried foods. As long as they are in glass with a tight-fitting lid, wild foods can be stored for long periods of time. My friends bring me their glass jars: in fact, I have begun to consider myself a recycling center! Of course, brown glass is best because it lets in less light, but it is not readily available. You can find green jars with narrow screw tops, however, and these are ideal for storing seeds, nuts, and berries—things that pour well from a narrow top.

The best container for large amounts of stored leaves, twigs, branches, and such are the half-gallon glass containers that deviled eggs, pickles, and mayonnaise come in. Restaurants or convenience stores might save them for you.

Devices that seal plastic bags, removing the air, work almost as well for long-term storage. Roots store well in paper bags, as well as nuts, twigs, barks, and branches to a lesser degree. As long as food is as dry as possible, long-term storage is possible.

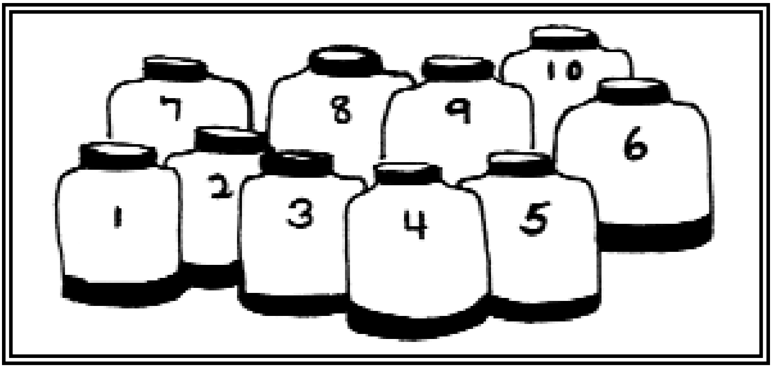

Storage Jars: #1 vegetables, #2 flowers, #3 twigs, #4 roots, #5 stems, #6 nuts, #7 corms, #8 buds, #10 flour

Storage Problems and Solutions

Problem: Pine needles that are stored in glass begin to display black spots or black mold.

This is known as ergot and develops on incompletely dried wild foods which are then sealed in airtight containers. Discard the contaminated foods as these molds are highly toxic. Similarly, grain that was incompletely dried also hosts this toxic black mold. Willow branches stored in a damp area outside develop black spots, also recognized as a poisonous mold.

Freezing Your Wild Foods

Freezer packages of vegetable matter have a life of six months. Roots, twigs and tubers last a bit longer as a rule. When I find freezer crystals throughout the freezer package, it's time for a massive green soup! Then you may re-freeze the soup.

The following are the wild foods successfully frozen. All must be steamed the indicated time before freezing.

Food	Steam Time	Food	Steam Time
Aloe Vera	raw	Plantain	1-2 minutes
Amaranth	1-2 minutes	Prickly pear	2-3 minutes
Arrowhead	1-2 minutes	Purslane	1-2 minutes
Aster	3 minutes	Malva	
Blackberry	2-3 minutes	neglecta leaf	1-2 minutes
Blueberries	1-2 minutes	Maple	2-3 minutes
Burdock	3 minutes	Milkweed	
Cattail	3 minutes	heads	3 minutes
Chamomile	1-2 minutes	leaves	2 minutes
Chickweed	1 minute	Mint	1-2 minutes
Chicory	3 minutes	Queen Anne's	1 minute
Clover	1-2 minutes	Raspberry	2-3 minutes
Daisy leaves	1-2 minutes	Rose	1 minute
Dandelion	3 minutes	Saguaro buds	3 minutes
Dock	3 minutes	Shep.'s purse	1-2 minutes
Evening		Strawberry	
primrose root	1 minute	leaves	2 minutes
Filarie	1-2 minutes	Thistles (all)	1-2 minutes
Fireweed	3 minutes	Thyme	1 minute
Goldenrod	1-2 minutes	Tumbleweed	1 minute
Grape	2 minutes	Violets	1 minute
Lamb's quarters	3 minutes	Wild lettuce	1 minute
Mustard	1-2 minutes	Wood sorrel	1 minute
Nettle leaves	2 minutes	Yarrow	1 minute

Some inexpensive freezing tools you'll need are:

- Plastic wrap
- Large plastic bags
- Labels or marker pen

To freeze vegetables, steam and drain the vegetable in a sieve; save the liquid for stock if desired. (See Soup Recipes starting on page 198.) Place one-foot-long piece of plastic wrap on the tray. Add 1/2 pieces of steamed vegetable and fold the edges in. Roll package tightly to seal. Any remaining liquid will seep out of package. Place the package in a large plastic bag and label with contents and date of freezing.

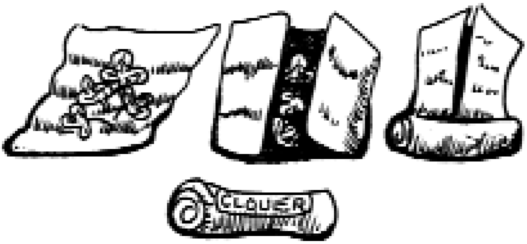

Vegetable Wraps

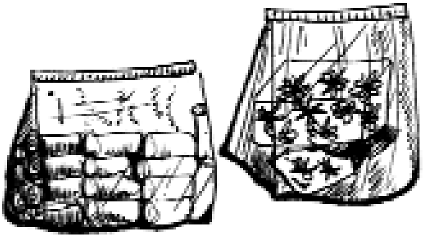

Freezing rolls of vegetables, flowers in blocks.

Freezing wildflowers is a simple matter of freezing whole flowers in solid blocks of ice. A frozen whole rose looks fantastic in a punch bowl. Use an ice cube with a rose inside instead of plain ice. As long as you use a container that will take freezing, you can create a bunch of frozen roses that looks like a crystal dream. My favorite luncheon surprise is layering a selection of flowers in different colors in a large, round bowl. I put in the bottom layer, add a little water, and freeze. A day later, I add another layer, add water, and freeze. And I repeat this until the display is to my liking.

To freeze whole roses, for instance, put a few inches of water in your containers, so that the flower does not float. Keep the pitcher of water handy so that you can reach it easily with one hand. Hold the stems in your other hand, and arrange the colors and shapes to suit you. (I cut the stems to uniform lengths first, making the job of holding them easier.) Then pour some water slowly around the sides, holding down the stems. Put on a lid or plate to hold the arrangement and freeze.

Whole heads of roses opened moderately and without the stems can be layered easily. The more roses in the container, the better. They won't remain in exactly the same position as you put them once you pour in the water, but if they don't float around a lot, the layers will be moderately undisturbed.

Pansies are relatively simple to freeze. Layer them in a sandwich bag and add a little water, then seal the bags. To freeze Queen Anne's lace flowers, spread plastic wrap or wax paper on a flat surface. Place the flowers face down on the wrap and roll up the wrap like a rug. If you have difficulty rolling a single piece of wrap with flowers, place cover sheet of wrap over the flowers and roll. Freeze the roll. Whenever you want some flowers, unroll the wrap take what you need, re-roll and return to the freezer. Queen Anne's lace flowers will turn brown when frozen, but they retain their delicious carrot taste and apparently their nutrition as well.

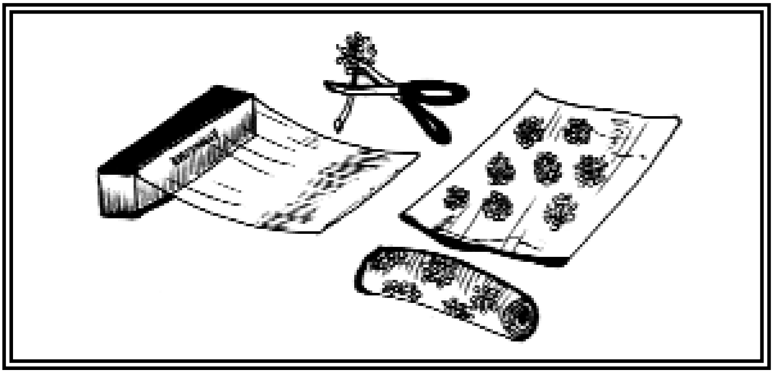

Freezing Queen Anne's Lace flowers.

Besides vegetables and fruits, you can successfully freeze vegetable stock and leftover fruits. Save all liquid from steaming or boiling vegetables or from leftover tea. The stock is best stored in freezer containers, ice cube trays, or heavy duty plastic bags. Soup stock can be frozen in layers until the container is full.

To freeze fruits, possibly for a mixed juice, layer the fruits in a large container as they are available. You might have a cup of raspberries—not quite enough to make jelly. Freeze in a large plastic container. Are there blueberries left over from making a pie? Add these to the fruit cache. When you have filled your container, you'll have the makings for a mixed fruit jelly or juice.

Decorations Can Be Edible Too

A blustery winter wind whistles outside the door, and my spirit needs a cup of hot rose tea—rose tea is filled with vitamin C, for muscle repair and vitality, nerve restoration, and a general feeling of well-being. From the rose-stem wreath on the north wall of my home, I break off a bunch of twigs, a few leaves, and one small dried flower. After rinsing and placing them in a tea pot, I cover the rose stems, leaves, and buds with boiling water. I let it steep for five minutes savoring the bouquet, and then serve myself tea from my old china teapot. Nature's bounty!

I had ventured out to a spot in the woods where wild roses grew. There they were, inviting me to make lush wreaths from the green shoots.

I took my loppers, and a big bag or length of twine, and began clipping off a few vines. I wound two vines together to make a circle, then added string to the overlapping ends to secure the wreath. I twisted more vines on the base, poking the ends through. When I finished, my decorative rose wreath was hung on the wall to provide a winter's supply of tea as well as to be a feast for my eyes and nose.

My home abounds with wild food wreaths. Circular, heart-shaped, sprays—all wreaths are either leaves or flowers for teas, grains for bread flour, vines or leaves for vegetables, and twig spices for herbal teas. I decorate every nook and cranny with food wreaths.

Wreaths made from wild grains are not only beautiful but practical. Since prehistoric days, people have twisted vines into coils and wreaths, combining craft with foraging. The Chippewa nation still does this today with wild rice hoops or wreaths.

What plants make the best food wreaths? Here’s a list.

Food Wreaths

- Amaranth
- Aster
- Chicory
- Cholla
- Clover
- Daisy
- Dock
- Fireweed
- Grape
- Lamb’s quarters
- Malva neglecta
- Queen Anne’s Lace
- Rose

GrainWreaths

- Bulrush
- Crabgrass
- Phragmities
- Spice wreaths
- Catnip
- Chamomile
- Clover
- Mints
- Strawberry
- Thyme
- Wintergreen
- Yarrow

Tea Wreaths

- Balsam fir
- Birch
- Blackberry
- Maple
- Meadowsweet
- Pine
- Raspberry
- Rose
- Sumac
- Willow

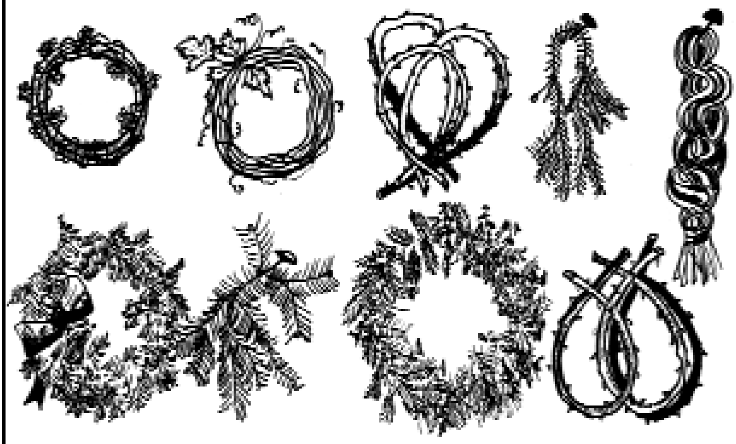

Food Wreaths (from left to right, top to bottom): Yarrow tea, Grape vine with tendrils for food, Raspberry canes for food and tea, Rosemary sprig for spice, Spice rope of thyme, mint & chamomile, (2nd row) Balsam fir & meadowsweet flowers for tea and food, Balsam sprig, Grasses, oats & dock seeds for food and flour, and Rose hoop for tea.

If you grow your own roses, treat them with a spray of mild soap and water instead of pesticides and herbicides to ensure their future edibility. Be sure your bush is at least 100 feet away from any busy street. And be aware of your neighbor's habits with sprays and wait a season to eat transplanted nursery stock.

When the time comes to buy your holiday tree, make sure it is organic. Christmas trees are often sprayed for insects and sometimes sprayed green to darken the needles. Find an organic dealer or cut your own tree. When the holidays are over and your tree is dry, spread a sheet around the base and shake off the dry needles. Place needles in a large oven container (a roaster, for example). Heat the oven to 300 degrees and then turn it off. Place the roaster with the pine needles in the oven and remove when the oven has cooled. Do not try to bake the needles! The resin is combustible. Put the completely dry needles into glass containers and enjoy five years more of your Christmas tree as tea.

The trimmings from your tree also make fantastic tea wreaths. Collect them before they are completely dry, rinsing and swishing any dust off the needles. Shape into wreaths and hang in your home for a source of fresh tea for weeks.

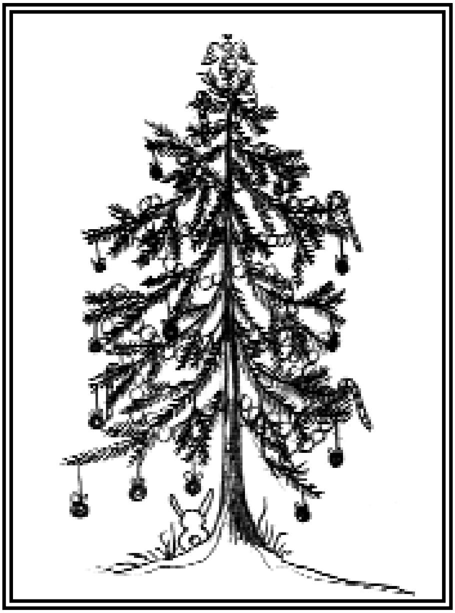

Notes

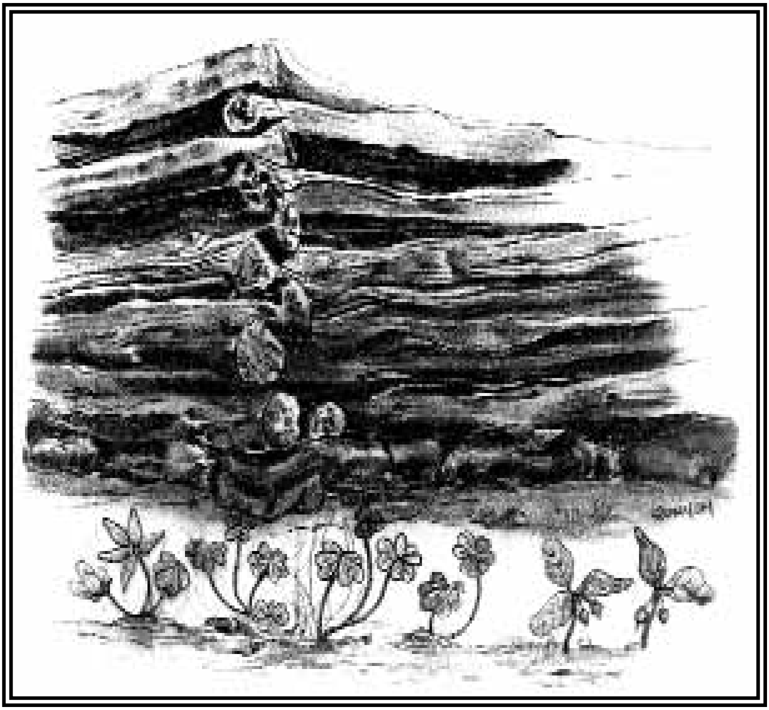

Left: garden sorrel   Right: wintergreen

Part II: Field Guide to Wild Foods

Edible Plants

This Field Guide section provides illustrations and descriptions of the fifty-nine foods discussed in this book. It is not intended to substitute for a comprehensive field guide to wildflowers, trees, or plants. Rather, it should be used in conjunction with three well-illustrated, comprehensive guides to the plant life of your region. Also please note that plant sizes vary with environment, and measurements given here are intended as general guidelines only.

The material that follows is organized as following, information where applicable:

Common name.
Scientific (Latin) name of species, or genus, only if the entry describes a group of related plants.
Family to which the plant belongs.
Other Names: Other common names by which the plant is known; also, types are described if the entry deals with a group of plant rather than a species.
History: Brief background information on the origin of the plant and a notation if used by Native Americans.
Habitat: General growing requirements—where the plant is likely to be found.
Characteristics: Whether the plant is a tree, shrub, or herbaceous plant; whether annual, biennial, or perennial; major physical characteristics to help you identify it in the field.
Primary Uses: Whether used for food, medicine, or as a cosmetic; some mention of the uses the plant has been put to (not meant to be comprehensive).
Nutritional Value: When applicable, a general indication of the plant's value to human nutrition. See also the Nutritional Value tables at the end of Part II for further information on the nutritional value of these plants.
Medicinal Value: When applicable, information on medicinal uses of the plant. See the Appendix for explanations of these medical terms.
Cosmetic Value: When applicable, an indication of how the plant is used as external treatment or otherwise.
Collection and Storage: Tips on harvesting and preparation for immediate use or long-term storage.
Caution: There are some wild plants that can be mistaken for these edible plants. See the section on Poisonous Look-Alikes,

Part V. Note also that entries in the Field Guide include occasional cautions or warning notices about collecting and handling some plants. Be sure to heed these cautions.

Comments: Personal remarks and additional information on the plant and its uses for us all.

Foreground: fiddleheads

Legend to the Uses of Wild Foods

These symbols appear in the description of each wild food and indicate the main uses of that plant at a glance

Beverage 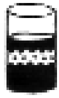

Bread 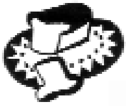

Casserole 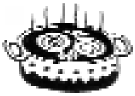

Cereal 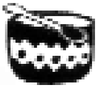

Flour 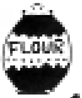

Jam 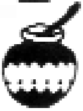

Medicine 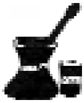

Pickles 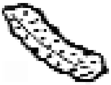

Salad 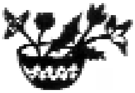

Soup 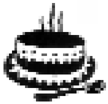

Sweets 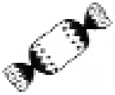

Tea 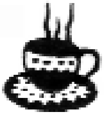

Toothpaste 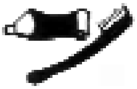

Vegetable 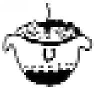

Wine 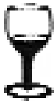

These symbols are used later in the poisonous look-alike section.

Edible 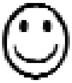

Poisonous 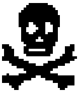

Aloe Vera

Aloe perfoliata vera
Family: *Liliaceae*

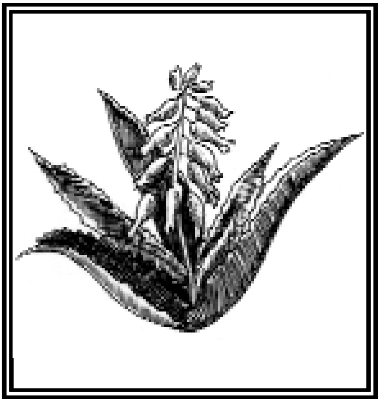

Other Names Medicinal aloe, Barbados aloe.

History: Probably native to Mediterranean. Used by Native Americans.

Habitat: Temperate zones, desert, arid regions.

Characteristics: Bushy succulent. Green fleshy leaves, called spikes, are filled with clear gelatinous liquid. Mature plant has spikes up to 3 feet high with dense, arrow-shaped clusters of yellow or orange flowers.

Primary Uses: Flowers, edible raw. Medicinal, cosmetic. In tea as an antispasmodic, antihistamine (asthma, colds, congestion), anodyne, or as tranquilizer or diuretic. Crushed leaves yield gel which is applied as poultice for burns or sores, promoting growth of new tissue.

Medicinal Value: An antibiotic, astringent (to accelerate wound healing), coagulating agent, biochemical bandage, pain inhibitor, growth stimulator, "wound hormone", demulcent, antihistamine, mild laxative, burn healer.

Cosmetic Value: Skin conditioner, astringent. Also aloe has bacteria- inhibiting qualities that make it an excellent underarm deodorant.

Collection and Storage: Harvest in summer; buds, flowers, young leaves. For external use, slice leaf in half and place directly where needed. For internal use, remove gel and place in water. Refrigerate or freeze. Freezes well. Where aloe grows naturally, both young and old leaves are thick. Since old leaves are tough and the gel is bitter, harvest the new, thinner leaves.

Caution: Do not use as bulk food. Use sparingly for medicinal purposes only.

Linda Says- Aloe vera is the most important native plant in the West. Preparing aloe is fun. I take a leaf, turn the flat side down, and lay it on a flat surface. The skin comes off easily by slicing under the skin from tip to larger end. The gel scrapes off quickly with a spoon, Then I scoop the gel into a bowl, whip it thoroughly with an egg beater, and put it into an ice cube tray to freeze. I use one cube to a 6-ounce glass of orange juice for a tonic. One cube also works wonders for immediate application on sunburns.

While aloe is a Western plant, mullein in the East has many of the same properties.

Different Plants, Similar Uses

ALOE VERA	MULLEIN
Antihistamine	Antihistamine
Burn healer, pain reliever	Burn healer, pain reliever
Growth stimulator	Growth stimulator, reduces swelling
Wound hormone	Wound healer-promotes quick healing
Pain inhibitor	Pain inhibitor
Astringent, accelerated wound healing	Whole leaf bandage, antiseptic
Demulcent	Antispasmodic
Coagulating agent	Tranquilizer, biochemical bandage
Antibiotic	Mild sedative
Deodorant, inhibitor of bacteria	
Mild laxative	Anadyne, a medicine that relieves pain
Cosmetic, skin softener	Cosmetic, skin softener

Amaranth

Amaranthus retroflexus
Amaranth family, *Amaranthaceae*

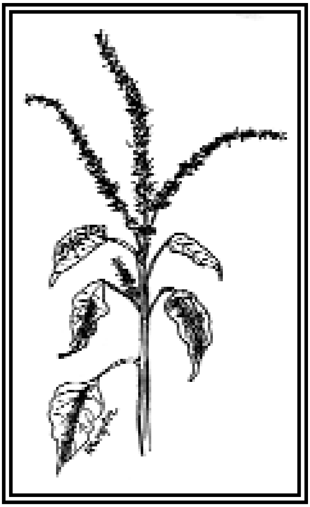

Common Amaranth (Western U.S.)

Other Names: Green amaranth, Pigweed, red root, carelessweed, choohugia (Pima name).
History: Native to tropical America; naturalized throughout the world. A staple food of the Zapotec Indians of Mexico. (Sturtevant 1972) Coastal Algonquins collected amaranth for a vegetable and used its ashes as salt. The plant is naturalized in Asian countries. It is cultivated in tropical Africa and Jamaica as a potherb. Seeds are used for flour in India and Nepal. Amaranth yields 8 ounces of seed from plants covering 1 square yard of ground.
Habitat: Cultivated soil.
Characteristics: Annual herb. Averages 2 foot, but may reach 6 feet or more. With bristly seed heads, clustered on multi-branch stems from a central stalk. Flower seeds are black and shiny; leaves are smooth and veined with slightly toothed margins. Eastern amaranth seed heads or "flowers" are denser and shorter than the Western species. Western varieties may have long, spindly, bristly seed heads, and white seed pods with black seeds inside.
Primary Uses: Culinary, cosmetic. Use leaves and stems like spinach, eaten raw, steamed, sautéed, cooking liquid is drunk. Leaves are also dried and ground for flour. They are used in soups and stews. Seeds are used raw or dried for baked goods, cereal, mush.

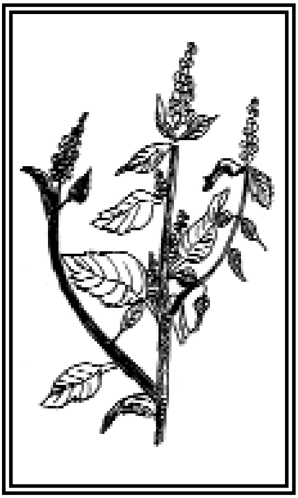

Common Amaranth (Eastern U.S.)

Nutritional Value: High in vegetable protein. High calcium and vitamin E.
Cosmetic Value: Astringent, wrinkle cream.
Collection and Storage: Use entire plant. Harvest lower leaves and branches in summer as vegetables. Refrigerate or freeze; dry. Wait until plant is full grown for large seed heads. Amaranth seeds are easily collected in autumn by tapping the seed head over a bowl, even in summer on the desert.

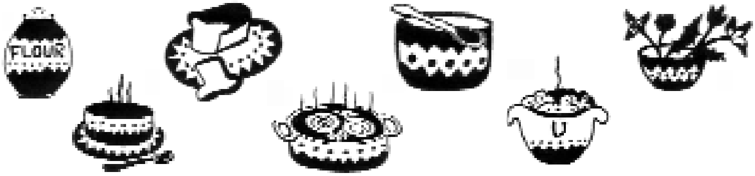

Linda Says- The first time I saw amaranth I couldn't believe my eyes! Tall, wondrous plants heavy with seeds bordering pasture and barn areas. There were hundreds of these "weeds" some bent over from the weight of their seeds. In less than 20 minutes I bundled enough amaranth for a week's supply of fresh vegetable and almost a winter's supply of flour. It did not take long to discover the wonders of amaranth in brownies, as a cooked green, or as a gruel.

Arrowhead

Sagittaria species, *S. chinensis*
Water Plantain Family, *Alismataceae*

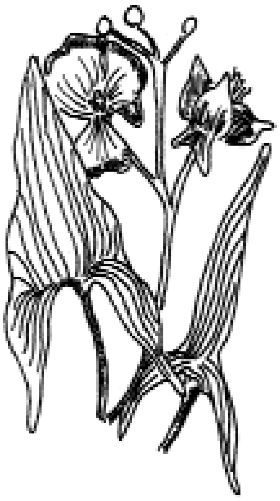

Other Names: Duck potato, wild potato, Wapatoo (Indian).
History: Native to North America. Lewis and Clark used Arrowhead for food, taught by Native Americans. The explorers wrote, “(they) can be ground fine. A flair for pudding, cakes, etc. They are nearly equal to Irish potatoes, and are a bread substitute.” Chinese sell it in marketplaces for food.

Habitat: Widely distributed. Marshes, wet bogs, mud flats.
Characteristics: Aquatic perennial with tubers. Grows in swamp water to a height of about 10 inches. Arrowhead-shaped leaves and filmy white-petal flowers. Roots have walnut-size tubers.
Primary Uses: Culinary. Root corms: Used as potatoes, sliced or whole, boiled, sautéed, or raw. Strung and dried then ground for flour.

Nutritional Value: Easily digestible; a nutritious food, especially for convalescents.
Collection and Storage: When harvesting arrowhead tubers, use a pitchfork and lift gently. The tubers tend to snap off in the mud and the pitchfork creates less tension. Freezes and dries well. Scrub the small tubers thoroughly and string with a needle and thread. Hang the strings to dry, then store in glass.
Caution: If water purity is in doubt, use purification tablet and soak tubers in water purifying solution.
Comments: There aren’t too many arrowheads in my area of northern New York, but southern New York and areas of New Jersey have immense quantities in marshes and wetlands; they are also in wetlands in the West.

Linda Says- My marshes do have a few "secret" areas for Arrowheads. Misty, boggy, muddy, slimy, and mossy, the marsh's loam logs rot away to the richest, blackest soil I have ever seen. Mosquitoes are an inch long where the arrowheads poke up above the mud! I pray silently, dip my pitchfork under the plant, and lift slowly. I used a needle and thread to string the tubers in the cabin, drying them for the winter. Easily reconstituted in water, the tubers make delicious soups and stews.

Aster

Aster nemoralis
Composite Family, *Asteraceae*

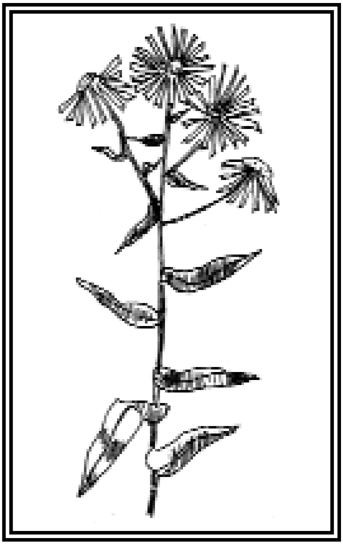

Other Names: Blue, pink, lavender, white & yellow aster.
History: Native to North America. Staple of the Iroquois Indians in the Northeast; food of tribes in the Midwest and West. Used in sweat lodges by Objibwa Indians to revive consciousness by putting on the hot rocks. Used as a food in South America.
Habitat: Fields and meadows, roadsides.
Characteristics: Perennial herb. Many varieties. Height varies, average 24 inches; may reach 40 inches. Numerous extensions jutting with multi-petaled flowers from a central stem. Petals are layered in flat, 1/2-inch-long sections. Colors vary from white to pink, yellow, blue, and purple.
Primary Uses: Culinary, Medicinal. Leaves eaten raw, steamed, boiled, sautéed, cooking liquid is drunk. Flowers stewed or steamed whole; chopped fine and stewed or steamed. Entire plant is dried, then stripped and ground for flour. Sprinkled on dishes, especially tomato-based dishes such as lasagna. The pith of the aster stem is edible, even in mid-winter. As emergency food, eat raw or cook in soup.
Medicinal Value: Dried stems and flowers are used as a wash for rheumatism; also excellent blood tonic tea. Roots of dwarf purple aster used for diarrhea control.
Collection and Storage: Snip the long stems at ground level and bundle together with string. I carry 2-foot lengths of string around my shoulders, and harvest the flowers until I have a bundle 10 to 15 stems thick. Stop bundling stems together when the flowers are too tight to be airy for drying. Take only a few plants from each

area (or 1 out of 4). Hang bundles upside down on a clothesline to dry. When dry, strip down onto sheet, crumble, and store. Also, leaves can be frozen.
Caution: Do not eat domestic asters.

Linda Says- Nature's abundance is particularly clear to me when I harvest the Iroquois staple, the wild aster. There are 250 varieties of wild aster in the United States: blue, lavender, yellow, and white prevail in the Adirondacks; blue, lavender and yellow in Western states. Field may have flowers as far as the eye can see. In less then 15 minutes I have enough to make a winter's supply of wild aster flour. I grind this bounty into flour for baking from dried storage food. The flour is particularly strong, nutritious. I recommend adding a small amount to regular recipes. This Iroquois staple remains one of my favorites.

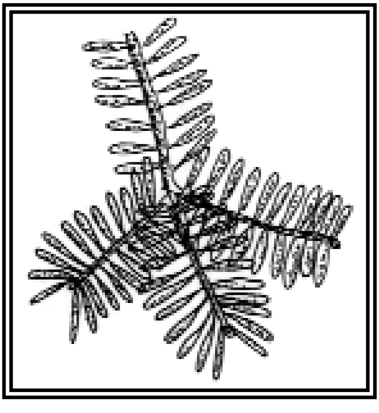

Balsam Fir

Abies balsamea
Pine Family, *Pinaceae*

Other Names: Christmas tree.
History: Native to North America. Soft inner bark was used as emergency food at Valley Forge. Native American usage.
Habitat: Grows in cool, moist, acidic soil along roadsides, woods, or lawns and in softwood forests.
Characteristics: Evergreen tree. Has small, short needles growing alternately from under stem, 1 to 1-1/2 inches long. Grows to height of 40 to 60 feet.
Primary Uses: Culinary, medicinal, cosmetic. Bark is chewed, ground or dried for flour. Sap is chewed as gum. Branches and

twigs are stripped and ground for flour. Needles are used raw for food; steeped raw or dried for tea. Used in wine. Ground raw or dried for emergency flour. Dried needles are used to fill balsam pillows and sachets. They lend a fresh scent to drawers and closets.
Nutritional Value: Good source of vitamin C.
Medicinal Value: Roots used as a poultice to soothe syphilitic sores and gonorrhea. Sap (pitch) is a “glue” for cuts, scratches, bites, and sores. Dried sap is chewed for colds, bronchitis, influenza, pneumonia, upper respiratory diseases.
Cosmetic Value: Oil used for flavoring, scents, perfumes, soaps.
Collection and Storage: One branch yields a winter’s supply. Balsam fir tips are a lighter green, indicating newer growth. I snip off needles here and there, filling a bag with dozens of twigs. Needles dry easily and quickly indoors in a sunny spot. For tea, simmer 1 cup needles until water is dark green. You may reuse needles or bark several times, adding water and simmering. Poke sap blister and gather liquid for emergency glue; store in glass if possible.

Linda Says- Rather than throw away your Christmas tree, use it for nutrition as well as a room refresher. For a wonderful aroma throughout the winter, simmer needles in water until they reach a thick, gummy state. Use twigs and parts of older branches as a fresh-smelling incense. To dry sap for medicine, make small flat cakes of sap and dry in a thin layer at the back of a wood stove or above the pilot light of a gas stove.

Birch

Betula species
Birch Family, *Betulaceae*

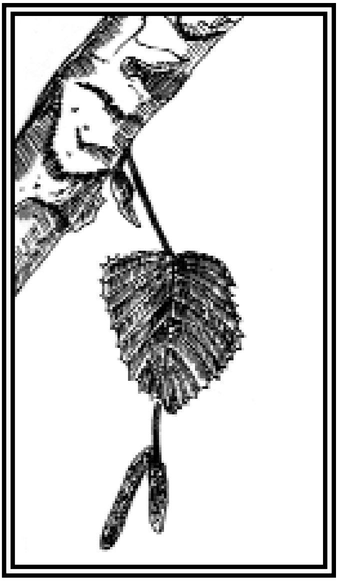

Other Names: White birch, paper birch, yellow birch, golden tree.
History: Native to North America. Scandinavians boiled, baked and added birch sawdust to their breads. Loggers say that their white birch sawdust used to be taken to American bread companies. Native Americans used birch dust for tea, and the strong inner bark was fashioned into many crafts. The Cree Indians folded birch bark and bit patterns into it, then unfolded it. Hundreds of Confederate soldiers were saved during their retreat to Monterey, Virginia, when they used birch bark as food[1] (Harris 1973).
Habitat: Woods, roadsides.
Characteristics: Deciduous tree. Alternate simple saw-toothed leaves. Papery bark of white and yellow birch peels in curls.
Primary Uses: Birch is an excellent cabinet-making wood; it makes strong hardwood furniture. Culinary, medicinal, cosmetic. Inner bark, sap, twigs, buds, and young leaves eaten raw as emergency food; dried and ground for flour. Sap is drunk raw for nutritious liquid. Twigs are used to make wines, dried for tea or crispy treat. Buds eaten raw. Young leaves steamed, sautéed, cooking liquid drunk.
Nutritional Value: High in minerals, calcium and phosphorous. High in potassium and beta carotene.

---

[1] Ben Charles Harris, p. 444

Medicinal Value: Buds are a tonic, since birch bark contains salix, or acetylsalicylic acid. A strong tea made from the bark has been used as an aspirin substitute.

Cosmetic Value: Birch bark used for skin problems, eczema, skin eruptions, pimples, dandruff, and scalp disorders. Dried, powdered leaves used for chaffed skin; birch oil for hair tonic or body oil.

Collection and Storage: Use the bark of downed trees, or follow a logging trail and gather the sheets of bark left behind. You can also peel the bark off firewood in the spring. Strips of birch bark make a quick, sure fire. Four 5 1/2 inch strips = 2 cups fine flour (a winter's supply). One branch approximately 4 feet long = winter's supply for 4 people.

Tap the trunks in spring (April-June) like maple trees. One tree yields an average of 1 gallon of syrup in a few hours. Twigs, inner bark, and sap can be used all year. Buds and twigs have a wintergreen taste. The sap can be boiled down and used as a syrup or in birch beer. Boil down birch syrup further to make birch molasses. Store dried and ground inner bark in glass as an emergency food. The inner bark makes a delicious mineral tea; carry a chunk with you when doing hard physical work. The inner bark dries to a breakable sheet ready to be used as tea; store in glass containers.

Caution: People allergic to aspirin may have a reaction to salix wild foods such as birch.

Caution: Peeling the bark of a living birch will kill the tree. This is called girdling a tree. The bark is literally the skin of the tree, and without it the birch will die.

Linda Says- Birch bark is one of my most precious possessions. I use it in every season for tinder and generally keep sheets of it in quantity. In the middle of one winter I received an order for birch-bark baskets. Since I had no bark on hand, I got permission to go up a logging trail with Ken and Todd. What a day-20 degrees below zero, with 4 feet of snow and no snowshoes! As I peeled bark off stumps of white birch with my knife and hatchet, Todd played in the drifts. That night my back ached in remembrance of the work, but I had enough bark for hundreds of baskets, with plenty left over for Christmas cards and for starting fires.

Blackberry

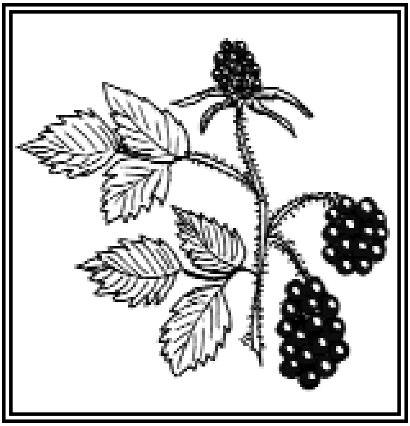

Rubus villosus and other
bramble berry
species
Rose Family, *Rosaceae*

Other Names: Creeping blackberry, cloudberry, wine-berry, dewberry.
History: Native to North America and elsewhere. Used as a fruit in all countries where found
Habitat: Roadsides, fields, meadows.
Characteristics: Biennial trailing stems. Bush varies in size, up to 5 or 6 feet tall. Saw-toothed leaves in groups of 3 to 5; thorns on angular, arching stem; fruit pebbly and black when ripe. Canes produce fruit on second year's growth. Fruit does not separate from receptacle when picked.
Primary Uses: Culinary, medicinal, cosmetic. Young stems and shoots eaten raw or cooked; used fresh or dried for tea. Fruit is eaten raw, and cooked for jellies, jams, syrups; also juices and wines. Flowers entirely edible raw. Young leaves are edible as a cooked vegetable in the spring.
Nutritional Value: Young stems and fruit are good source of vitamin C.
Medicinal Value: Bark of stem helpful for intestinal problems. Tea made from blackberry root helps control diarrhea. This tea is also an astringent, a mouthwash for mouth sores. As a strong tea or eaten raw, all parts of this plant can be used to relieve diarrhea or as a blood cleanser.
Cosmetic Value: Leaves are good for facials, masks, lotions, and astringents. In an extract, clears blemishes, eases scalp itches, and heals scales.
Collection and Storage: Tie a lightweight receptacle around your waist and use both hands to pick berries. Hand-pick young leaves in early summer.

Picking Blackberries in Adirondack Bear Country

The dirt road next to the shanty led down to an old log bridge. The bridge spanned the Cedar River, and there was nothing but private wilderness for several miles. I crossed the river and climbed the wooded path for one mile. In another half-mile, I came to the largest blackberry patch I had ever seen. I was equipped with lightweight containers for the berries and fly repellent made of olive oil and pennyroyal.

I whistled a tune as I approached the patch, wary of the bears known to be in the area. I knew if I blended in too much, the bears might resent it. This philosophy had worked for me before-either I was allowed into their domain or I was not. Then I became aware of a very large body up ahead. I could hear and smell the bear shuffling slowly through the woods. I waited. Testing the wind for the bear smell and listening with my ears as I whistled, I knew the bear shuffled away.

I picked blackberries with both hands, dumping them into a plastic bottle and then into a large bucket. I worked my way around the path from left to right. Experience told me that I needed two hours to obtain eight quarts of juicy blackberries, and this particular day I picked eight quarts and then left the patch to the bears again.

When I worked with my son, Todd, we had our two coffee cans filled in moments and proceeded to dye our sweatshirts blue-black with the weight of the juicy fruit. With over a gallon of ripe blackberries we made jelly that night over an evening campfire. Breakfast was a treat!

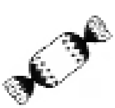

Blueberry

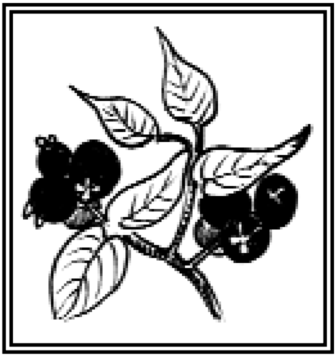

Vaccinium myrtillus
Heath Family, *Ericaceae*

Other Names: Bilberry.
History: Native to North America. Used by Native American tribes.
Habitat: Acid soil, fields, roadsides, waste areas, bogs, marshes, woods.
Characteristics: Low spreading plant or woody shrub. Many varieties, from ground-hugging plants to high bushes. Fruit is in clusters in groups of 2 to 10. (Huckleberry often grows near blueberry; edible fruit looks similar, but huckleberry has stony pit and tart taste.) Wild blueberry looks like bilberry: both are edible, although true bilberry is not very tasty.
Primary Uses: Culinary, Medicinal. Fruit is eaten raw or cooked for jellies, jams, syrups; also wines.
Nutritional Value: High source of Vitamin C.
Medicinal Value: Blueberry leaves used medicinally in a tea. Because they contain myrtillin, an antispasmodic and relaxant, blueberries were used as a drink during childbirth to relax the mother. Leaves, fruit, and twigs can be made into an infusion and drunk for kidney problems. Dried berries are chewed slowly and thoroughly to aid control of diarrhea. Leaves and twigs can be used separately or together as teas, tisanes, and throat gargle.
Collection and Storage: Blueberries are best picked by first pulling sections with both hands rather roughly and letting berries fall on to a sheet spread around the bushes. After rough-picking a few bushes, rustle the branches for the many berries caught between the leaves. Gently gather up the sheet and pour the berries on a tray.

- 4 cups fresh berries = 8 jars jelly
- 4 cups fresh berries = 2 cups dried berries
- 4 cups fresh berries = 1 quart frozen berries

Twigs can be dried or dug up under the snow. Berries are placed in cold water. The immature berries and leaves float to the top for easy removal. To dry, spread on a tray; cover with a screen or cloth. For storage, berries can be dried until they rattle, or frozen.
Caution: Positively identify; crush fruit for characteristic blueber-

ry smell. Blueberries have a slight frill and an opening on the side away from the stem. A species of buckthorn, which is poisonous, has blue berries without a crown or frills.

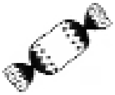

Linda Says- The fields behind the old McConnel place were filled with ripe berries, and became my favorite place to pick. I'd place a sheet around the bushes, and use both hands for picking. With a sheet you can become lightning fast, but it is still a challenge to pick enough berries for a pie before the blackflies find you. I usually made smudges before the clouds of flies found me if I wanted to take my time picking. Then I'd dry blueberries on screens covered with a cheese cloth. Those berries in my cereal on a winter's morning made it seem like spring.

Bulrush

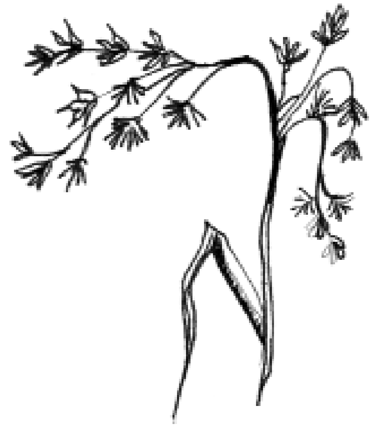

Scirpus validus and other sedge varieties
Sedge Family, *Juncaceae*

Other Names: Sedge, tule, great bulrush, nupiaskun (Crow Indian)
History: Found nearly throughout the world. Used as food from Moses' time including use by all Native Americans. The Crow made mats by sewing together bulrushes with basswood string and a bone needle.
Habitat: Wet areas, drainage spots, bogs, swamps; alongside waterways and ocean bays.

Characteristics: Grasslike herb. Many varieties with height to 24 inches or more. Stems smooth, round, with pith; head of flower spike full of seeds and pollen.
Primary Uses: Culinary, medicinal. Roots cooked as potato. Early shoots eaten raw, steamed, boiled, sautéed; cooking liquid drunk. Flower clusters dried and ground for flour. Bulrush flour is sweet and nutty—it has the most branlike taste of all the grains. Pollen used as flour. Seeds ground and roasted.
Nutritional Value: Highly nutritious, roots rich in complex carbohydrates.
Medicinal Value: A plaster or poultice for aches and pains.
Collection and Storage:

- 6-8 bulrush plants = up to 1 cup leaves
- 1 cup leaves = 1/2 cup bulrush flour
- 1 teaspoon powdered bulrush leaves = 1 cup tea

See page 34 for collection details.

Dry the bulrushes whole, bundle the long stems and hang them upside down in a warm, airy spot out of the sunshine. To ensure dryness, dry in a 300° F. oven for 5 minutes, then turn oven off and wait until cool. Follow this drying procedure even if using the bulrush strictly for decorative purposes.

Linda Says- I have spent many wonderful mornings watching dragonflies and frogs playing in the swamp while I gathered bulrushes. Swamping is not only fun, but profitable-I've fed many people on bulrush bread.

Insects love to lay tiny eggs in bulrush tops, so even if I am using the bulrushes for decoration, I leave them outside for 24 hours so the bugs can leave. After drying, I use them in lovely arrangements or grind them to a fine, delicious flour for baking.

Burdock

Arctium lappa
Sunflower Family, *Asteraceae*

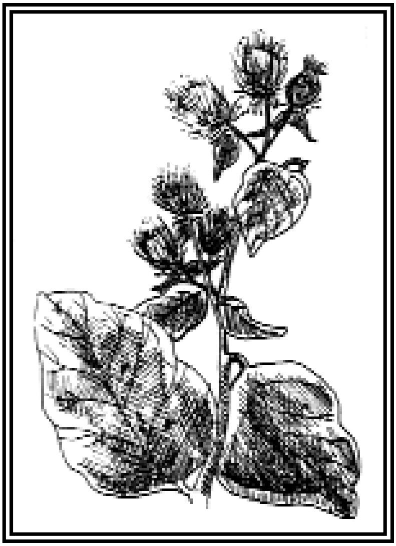

Other Names: Beggar’s buttons, burs.
History: Native to Europe and Asia; widely naturalized in North America.
Habitat: Fields, roads, most waste areas.
Characteristics: Bush like, large green leaves over a foot wide, and many veined. Tubular, purple or white flowers followed by spherical burrs, 3/4“ diameter. In the fall, burs dry to brown, sticky seed pods.
Primary Uses: Culinary, medicinal. Roots are boiled; dried and powdered for nutrient additive. Stem pith is baked, boiled; eaten as pickle, candy. Early leaves are eaten raw, steamed, boiled, sautéed; cooking liquid drunk. Large leaves are used as trays and receptacles for small plants. Burs when young and still pliable are steamed and eaten; when mature, broken in half and white pith eaten raw.
1 large leaf holds 2 cups green burs
1 8-inch root = enough raw root for 2 or 3 people
Nutritional Value: High in vegetable protein. Contains inulin, the chemical source of insulin.
Medicinal Value: Tea used to heal infections. Root is a diaphoretic; as wash infusion, for poultice, bruises, sore or tired feet.
Cosmetic Value: Seeds for skin problems, eczema, cleansing. For pet’s fleas, simmer a small handful of brown burs, strain, and cool. Wash your pets with the decoction. Kills fleas quickly!
Collection and Storage: When you can, harvest the young roots before the flower stalk develops (usually the second year of

growth). When digging roots, slide the shovel thoroughly down one side, then another, until all 4 sides have been deeply cut. Keeping the shovel in one side, pry up the root without breaking it off. You will feel the pop when the root releases from the earth. If the tip remains, the plant will grow again. Steam and wrap burs individually to freeze. Slice and fry or simmer and then freeze roots. Gather burdock in any season. Similar in size to rhubarb leaves, the gigantic burdock leaf forms its own carrier. In spring, place tiny new burs on a leaf, roll the leaf, tuck in the ends, and put the leaf in your pocket. In winter the burs are brown. I take a few brown burs, stroke the bushes, and gather a ball in no time. I stick the burs together in a large ball and carry them on my shoulder, freeing both hands for further picking.

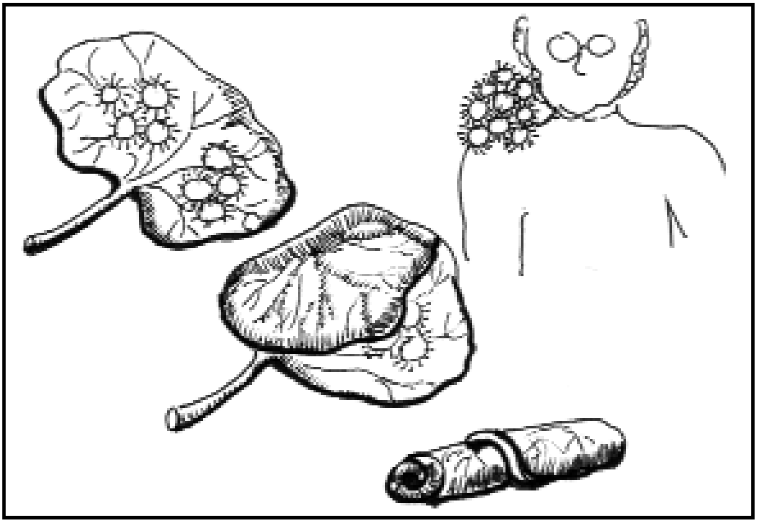

Carrying Burs

Linda Says- When harvesting, cut over the joint of the root where the stem begins. Scrub the root with a brush then peel root until you reach fiber. The fiber will turn from white to brown very quickly; this is normal oxidizing. You can peel the brown off again when serving, but I just steam the burdock pieces and freeze for a winter's meal. Drying the peels makes a crunchy, chewy root food.

Cattail

Typha latifolia
Cattail Family, *Typhacerae*

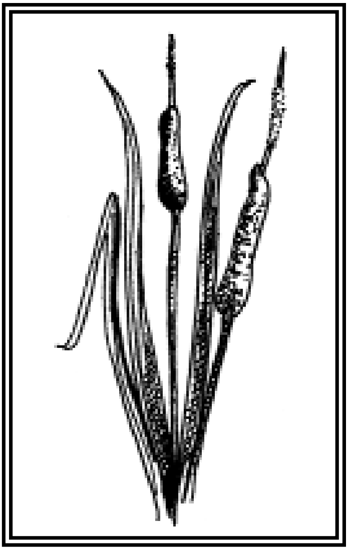

Other Names: Supermarket of the swamp, punk, upakiotipa (Crow Indian).

History: Native to North America, Europe, Asia. Cattail has been used for food throughout recorded history by people in all countries where it is found.

Habitat: Bogs, swamps, wet areas.

Characteristics: Perennial herb. Grows in wet areas to a height of 10 feet or more. Stalks have hot dog-shape heads, pollen flag in early spring. Very tall slender leaves with 1 vein.

Primary Uses: Culinary, medicinal. Roots are dried and ground for flour. Early shoots are eaten raw. Stem pith is eaten raw, boiled, and pickled. Early green heads are eaten raw, cut and cooked as ear of corn. Early brown heads are ground for flour. Pollen is used as nutrient additive. Leaves are used for basket weaving, mats, other crafts.

Nutritional Value: Plant holds about 30 percent complex carbohydrates; highly nutritious.

Medicinal Value: Flower heads used in tea for diarrhea control.

Collection and Storage: Pollen collected in spring; shake into a paper bag. Stems harvested before the cattail flowers, whenever possible. Roots dug from winter to early spring. Gather fluff from mature heads for excellent insulation or stuffing for jackets. The fluff floats and is waterproof, and serves as excellent tinder and torch. Can be used as cotton.

Caution: If water purity is in doubt, use purification tablet and soak plant in solution.

Caution: Pollen fluff may cause skin to break out in hives.

Cattailing Seasons

The morning is misty and cool. The Adirondack. ice has finally left the swamp and the crystals have melted along the edges of the mossy mats. The cattail shoots will soon be through. Water is beginning to show the spring signs of life. I saw bugs skating yesterday and heard a frog peep. Steam is rising from the invisible shield above the water. The old, dried stalks of cattails, left after winter's fury, lay bent and crooked, disheveled. The matted fuzz of the old tails looks pathetic as the new growth heads toward the surface of the shield.

There are blades-green shoots-poking up from the swamp. I take my trusty shears, long boots, plastic garbage bag, 4-foot board, and towel (should I fall in the boggy mud), to gather cattail shoots to can for winter. Laying the board from mat to mat, I gingerly move the length, clipping the shoots and dragging my filling bag behind me. Glancing back into the swamp, I can tell where I have harvested. The shoots will come again, reaching toward the sun of summer.

Later in spring, green plumes meet my eyes. The heads are like corn on the cob. I can taste and smell the meaty, nutritious flesh. My board goes down gingerly now. The swamp is dotted with thousands of green spikes. Gathering carefully, I pick only cattails with a full cob. Making piles of 8 to 10 stems, I bunch them in a pattern behind me, then I wrap each bundle with the leaves, winding them round.

Though the water is much warmer now, steam rises as dozens of insects zoom through. Mosquitoes glide by, legs tucked up underneath, looking like tiny planes with their landing gear drawn for a landing. Frogs and peepers croak and splash off their pads. I have cut the largest harvest ever. Now winding my way back I pick 10, 20, 30 tails at once. I'm coming back to camp with a supply of cobs to savor.

Summer is here and the murky mud is tough going. The cattails have filled the swamp with heavy bushy stems and blades rich with brown cobs. They are ready for harvest. Flour time! Bread time! Pancake time!

The cattail stems have swelled to 1 inch, sometimes 1 1/2 inches thick. Inside lies a creamy white pith that makes delicious pickles. Sweet and highly nutritious, they will supply a winter's fat. The cattail is heaviest now so I gather only a few, counting up to 10. I need 20 to 30 to make a loaf of bread. The mosquitoes are merciless, and the 90 degree August heat is causing the brackish water to make a stench. Under the cattails grow pennyroyal, mint, and other grasses. The harvest is half done for the winter. Another day and I will have seven or eight gallon containers filled with the pure starch piths.

Fall is here and my last trip for the year to the swamp is a sad one. Most mosquitoes have disappeared when the temperature dipped to 50 degrees. Ice crystals will form soon. Soon 40 mile per hour wind will swirl snow. Life will be in a dormant stage for many months now, but the precious cobs will be our insulation, our stuffing for pillows, blankets, and winter jackets. This collection is the easiest. As fast as I cut a cob, the pollen pops out and my bag fills with creamy fluff. The tough brown skin holds some heads together, awaiting a twist by both hands to rip open and fluff out the rich pollen.

The mud has an almost fishy odor, and it is full of algae and frogs. At this last harvest of the year, the rewards come in floods, and the swamp holds my deepest respect. I have reaped its finest fruit-the supermarket plant of the swamp: the aquatic cattail.

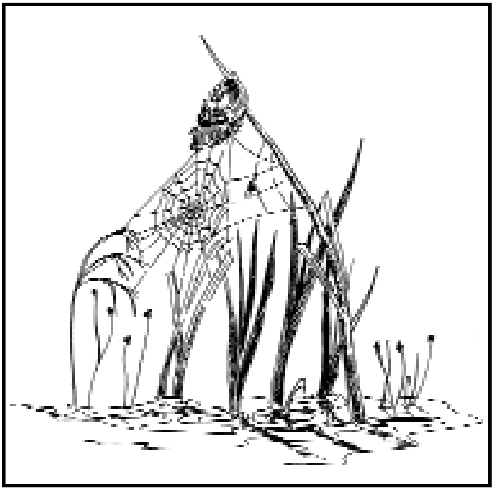

Cattails and Cobweb

Chamomile

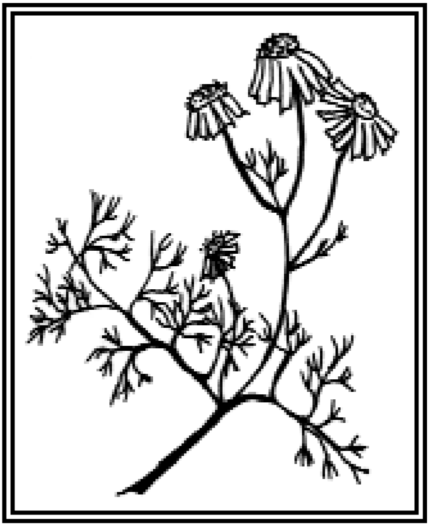

Matricaria chamomilla
Composite Family,
Asteraceae

Other Names: Ground apple, Roman chamomile, mayweed, German chamomile, pineapple weed.
History: Some varieties native to North America; many naturalized from Europe. Used throughout Europe. Also cultivated for commercial use, such as in shampoos and teas. Oil is extracted for many products, including hair rinse and insect repellents. The fragrance is called mananilla in Spain. Used especially in Italy, England, and Germany.
Habitat: Roadsides, “people places,” walkways, gardens.
Characteristics: Biennial herb. Height average of 3 inches; may reach 12 inches or more. Tight-budded flowers, tiny yellow daisy; sweet, pineapple-like smell.
Primary Uses: Culinary, medicinal, cosmetic. Twigs are dried for crispy treat. Leaves are eaten raw; steamed; sautéed; cooking liquid drunk. Steeped raw or dried for tea. Flowers are eaten raw; steamed; sautéed; cooking liquid drunk. Entire plant is steamed; boiled; used in soups and stews. Steeped raw or dried for tea.
Nutritional Value: High in minerals; calcium, iron, potassium, and niacin.
Medicinal Value: Stomatic, diuretic, chapped skin ointment,. root used for toothache.
Cosmetic Value: Compress for eyes, shampoos, rinses for hair luster.
Collection and Storage: Grows near pathways, and a winter’s supply can be picked in a few minutes. Shear with scissors.

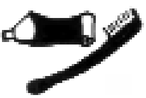

Linda Says- The smell of chamomile tea means fields, sunshine, and peace for me. I especially like to dry the flowers for tea, but any part of the plant will serve just as well. This lovely plant has more uses than tea, though. I eat the plant raw as a mild tranquilizer. The plant's parts make a delicious addition to salads. Chamomile flowers frozen in ice cubes bring a touch of spring to a drink on a cold winter's day.

To use as a hair rinse, simmer the entire plant -leaves, roots, and flowers-in water. Simmer until water is green, about 15 minutes. Cool the water, strain, and apply. Let this sit on your hair for a while, then rinse again. It will give your hair amazing vitality.

Try seeding chamomile in your garden or lawn and let it go wild.

Chickweed

Stellaria media (common chickweed), *Cerastium* (mouse ear chickweed)
Pink Family, *Caryphyllea*

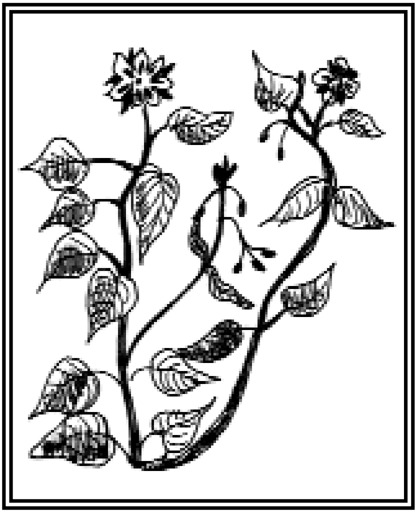

Other Names: Star flower, chickweed, stitchwort, starwort.

History: *Stellaria* native to Europe and Naturalized in North America; *Cerastium* native to North America. Widely distributed; used by Native Americans.

Habitat: Fields, lawns, roadsides, waste places.

Characteristics: Annual or perennial. Flowers petaled, 5 pairs, deeply cleft in half. Tiny and white above long stalk 4 to 6 inches. Leaves ovate and smooth. All types of chickweed are small ground plants with some parts erect but mostly prostrate. Prolific, seen year-round in some areas, even under the snow. Flowers can be found most months of the year. Mouse-ear chickweed has minute hair on all leaves and stems.

Primary Uses: Culinary, medicinal. Whole plant is edible raw. Used as a vegetable or salad green. Also stewed, stir-fried, dried or frozen for later use. Mouse-ear chickweed is cooked to remove hairs.

Nutritional Value: High in calcium, potassium, and ascorbic acid; trace minerals.

Medicinal Value: Soothing effect; upper respiratory; externally for poultices on inflamed skin.

Collection and Storage: The tiny star-shaped flower petals have a tiny notch at the end. Leaves are easily twisted and picked from stems. Gather as you would dandelion leaves (see page 24). Young leaves are best. I have always identified them by this.

Linda Says- A bit of garlic and steaming hot chickweed make a nutritious meal for the forager because it is high in calcium, potassium, ascorbic acid, and trace minerals. I once found myself sitting on a hillside, pulling handfuls of chickweed and eating them with gusto! The most amazing feeling is the delicate leaves of chickweed. When I put my fingers down into the thick mat of tiny nourishing leaves to pull up a handful, my contact with this plant always is one of delicacy and gentleness. A little washing, oil and vinegar and the feeling is easily transferred to my mouth!!!

Chicory

Cichorium intybus
Composite Family, *Asteraceae*

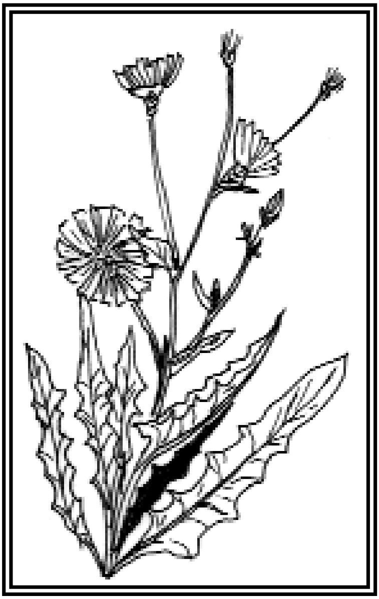

Other Names: Blue sailor, blue daisy, coffee weed, wild succory, Barbe de capuchin, whitloff, ragged sailor, wild bachelor button, blue dandelion, blue daisy.
History: Native to Europe, naturalized throughout much of North America. Used by Native Americans and a staple European food. Each year the United States imports tons of chicory root for use as a coffee substitute, coffee flavor, and coffee extender.
Habitat: Roadsides, fields, dry sandy areas.
Characteristics: Perennial herb. Grows to height of 2 feet but may reach 4 feet. Has deep taproot and milky juice. Red-veined, deep-cleft leaves are thicker than dandelion to touch. Stem has many star-blue ray flowers. When broken, root exudes a white milky sap, as do the stem and leaves. Plant begins in a lettuce-like whorl of leaves on the ground, heads up in a bunch, then shoots a stalk upwards from the center. Roadsides show bunches of deep green, deeply indented leaves; not unlike dandelion, but a bit more uniform in clusters before the stalk appears.
Primary Uses: Culinary and medicinal. Roots are dried and ground to coffee. Leaves are eaten raw in salads and sandwiches. May be steamed, stir-fried, boiled as a vegetable, sautéed; cooking liquid drunk. Stems are entirely edible as are the flowers and seeds.
Nutritional Value: Roots are rich in beta-carotene and niacin as well as carbohydrates.
Medicinal Value: For stomach and kidney disorders.
Collection and Storage: Gather roots in early spring. Roots may be stored in strips for easier grinding (see page 30 for illustration on stripping roots). Store whole roots in paper bags in a warm, dry area for future pulverizing or boiling.

- 1/2 paper grocery bag of dried leaves = 1 quart rough ground flour
- 1/2 paper grocery bag of dried chicory roots = 1 quart ground coffee substitute

To collect pull up entire plant by grasping stem as close to root as possible and pull straight up. Hack off the root (I use an ax), gather and wash the leaves, and scrub the root with a toothbrush. When the roots are clean, slice lengthwise into strings. These dry easily on a screen or may be strung over a wood stove to dry.

Linda Says- Chicory roots pull up easily from sandy soil. I put my bounty in a rain barrel to keep the entire plant fresh until I can work on it. All summer, Todd mows the lawn where there is chicory. The reddish vein makes it easy to identify the plant. To tell chicory from the common dandelion, just feel both leaves. Chicory leaves are thicker and not shiny. Dandelion is softer and shiny.

Cholla

Opuntia fulgida
Cactus Family, *Cactaceae*

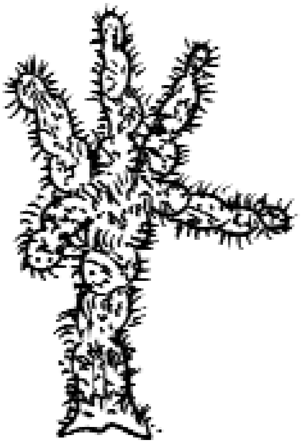

Other Names: Jumping cactus.
History: Native to North America. Used by Hohokam, prehistoric desert people. Indian women used baskets, sticks, and wooden tongs made from saguaro ribs to gather the buds. A firepit was dug in the desert floor, and the buds and joints were placed in a mesquite fire, roasted, and split in two to eat the succulent insides.

Habitat: Desert.
Characteristics: Treelike cactus with many branches. Cholla cactus has dozens of individual egg-shaped barbarous sections extending from tree-like stems reaching heights of 3 feet or more. Flower is light rose color, fruit is green and smooth.
Primary Uses: Culinary, medicinal, cosmetic. Edible flowers, seeds, fruits, and bud extensions. Fruits are eaten raw, boiled, or baked. Dried for long-term storage. Fruits used in soups, casserole.
Nutritional Value: High in calcium and iron.
Medicinal Value: Gel applied on skin burns.
Cosmetic Value: Gel used as skin softener.
Collection and Storage: Use tongs and paper bags to collect fruit, leaves, and flowers of cholla. Spines and glochids are removed in any of several methods. Indians used flash fire, holding a flame under the burr to remove the glochids, so that they could be opened easily and handled with the fingers. Another method is to place burrs in one paper bag and transfer to another several times. Dry cholla buds on screens in the sun. Cover with cheesecloth if birds pick at them. Dried buds are stored in paper bags until needed. When needed, reconstitute in water about 3 to 4 hours, then boil for one-half hour.

Evelyn Neithammer (1974) found that the easiest way to clean cholla buds is to fill each of 2 saucepans one-third full of clean gravel. The buds are added and the gravel and buds poured from one pan to the other four or five times, or until rid of spines and glochids. (Glochids are minuscule, dense pockets of small barbs which protrude from the pads of prickly pear cactus. They are small but mighty protection against antelope, deer, and cattle of the Western plains.)

Caution: Spines and glochids will penetrate skin with a voracious sting and burning sensation. Do not touch the cactus with bare skin.
Caution: All cholla, prickly pear, and saguaro cactus is “protected plant, by State of Arizona,” but it is legal to pick fruits and buds of the species in this field guide for food. The rare crested saguaro is completely protected, so NO fruits or parts may be taken. Be sure to check the regulations in your state.

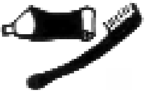

Teddy Bear Cholla

My first introduction to cholla is a painful, clear memory. My brother, Paul, and I were taking a first tour of the desert together. I was listening to him carefully. "Walk behind me, check your shoes carefully for the cholla burrs, and never step over the bushes here...walk around in the bare spaces." I was trying to listen, but the 100 degree temperature became a problem and I began doing my Adirondack way of walking. I stepped over a bush, but never put my other foot down. It was stuck to my rear with a thousand fire barbs. (The burrs were attached to my sneaker heel, which brushed my rear closely enough to get the burrs attached to my skin.) Screaming, I stood an embarrassed 20 minutes while Paul picked out the hundreds of minute hairs with tweezers. Needless to say, I remember the rules now and thank him for a lesson well learned!

Clover

Trifolium pratense (red clover)
Trefolium repens (white clover)
Legume Family, *Leguminosae*

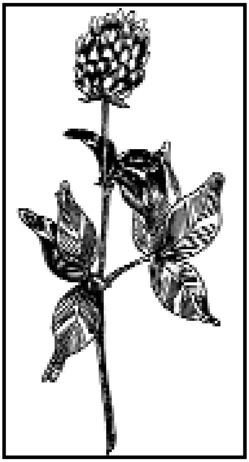

Red Clover

Other Names: Long stalk, strawberry clover (red clover); common, sweet, short stalk clover (white clover).
History: Naturalized from Europe. With over 300 species, clover has been used by all cultures throughout history. Native Americans used it widely as a vegetable or cure for chest congestion.
Habitat: Roadsides, fields, lawns.
Characteristics: Biennial or perennial herb. Red clover reaches height of 10 inches or more, with hairy stems. Red or purple blossom with oval nectar sections; elongated leaves form trefoil with white vein when mature. White clover reaches height of 2 inches or more. White blossoms have dozens of nectar filled sections; round leaves form trefoil at end of stem.
Primary Uses: Culinary, medicinal, cosmetic. Leaves are eaten raw; steamed; boiled; sautéed; cooking liquid drunk. Dried and

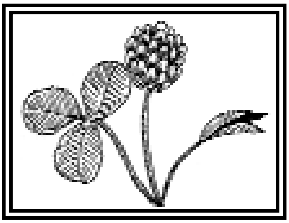

White Clover

ground for flour. Flowers eaten raw; steamed; sautéed; fried; cooking liquid drunk. Dried for tea; used for wines. Seeds crushed for cereal; sprouted. Entire plant is steamed or boiled and used in soups and stews; cooking liquid is drunk.

Nutritional Value: High in calcium, potassium, niacin and vegetable protein.

Medicinal Value: Red clover is used as tea for cough, whooping cough; blood tonic or purifier. Clover syrup used for chest congestion and bronchitis.

Cosmetic Value: Cosmetics, facial creams, rinses, shampoos, wash for pimples, poultices for athlete's foot fungus.

Collection and Storage: Plants are most succulent in spring and early summer. Gathering a winter's supply of clover takes only a few minutes. Clover can be frozen by placing it in a single layer on freezer wrap, folding over 2 sides to hold the clover in place, and freezing. After the clover is frozen, roll the paper to make a compact package, fasten, and label. Dry seed heads separately for an attractive potpourri.

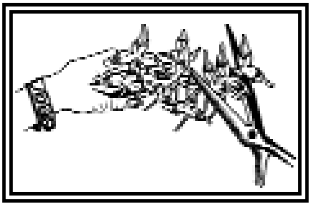

Sheering Clover

Linda Says - White clover has a particularly different flavor from red clover, with a smaller flower and leaf. The white clover leaves are easily sheared for food as they grow close to the ground in patches. One clover patch 3 feet wide will provide dozens of servings in one season. One summer I conducted an experiment. Shearing a 3-foot patch every morning for 3 weeks, I successfully fed 200 people several different dishes of clover: clover with rice or potatoes, clover casserole, breaded clover heads, clover tea, candied clover heads. The clover flowers went well in spaghetti sauce for meatballs too!

A Protection Powerhouse

Red clover borders almost every road edge. It appears to grow on all lawns, & is found in dozens of varieties throughout the world. When I discovered that the plant leaves were vegetable protein, I was ecstatic. I had been walking into town every two weeks for supplies, including cheese. My biggest nutritional concerns were protein and calcium,. & then I read that clover has tremendous amounts of protein. Indeed, five large clover leaves provided the same amount of vegetable protein as 1 ounce of cheese. That did it! I began to put clover leaves between two pieces of whole wheat bread, calling it cheese. Within three days I was eating the leaves raw, saying "I need energy." I would forage freely on a small amount of raw leaves.

A vegetable, a tea, a vegetable flour, clover has become one of my staples. I add clover to rice dishes, simmer clover soup, & steam casseroles with clover as an ingredient.

Crabgrass

Digitaria sanguinalis L.
Grass Family, *Gramineae*

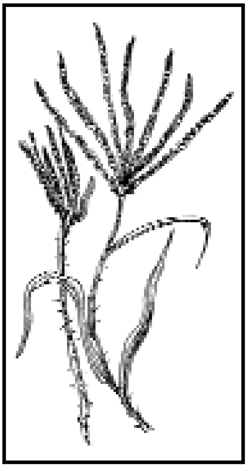

Other Names: Finger grass (Horizontalis, ischaemun (smooth); serotina (creeping); longiflora; simpsoni; filiformis, and more.)
History: Native to North America, Asia, Mediterranean and North Atlantic coast as well as Europe. A drink called Mannagrit is a popular recipe. Used in most areas of the Mediterranean as flour for pita cake.
Habitat: Lawns, roadsides, waste places, edges of fields. The large crabgrass prefers moist soil, streams, ditch banks.
Characteristics: Grasslike, many varieties reaching heights from 3 to 24 inches or more. Summer annual and perennial, the bract or seed tips may vary in amounts and thickness, as well as smooth seeds or heavy bracts.

Primary Uses: Seed bracts as well as leaves are dried and ground to flour. Used raw, steamed, boiled, sautéed, and cooking liquid drunk. Pollen used as flour; seeds are sautéed; stir-fried, or ground and used as flour.
Nutritional Value: Highly nutritious; roots rich in complex carbs.
Medicinal Value: Used as a poultice for aches and pains.
Collection and Storage: Shear bracts and leaves, as if mowing lawn. Dry thoroughly, store whole in glass. Grind to flour when needed.

Linda Says- I'll never forget the feeling when I realized that a 2-foot section of crabgrass would yield a batch of muffins every 2 weeks or so. When I read that the Bedouins pulverized crabgrass to a rough flour and made a cake the same as I did, I knew then that my cakes were a universal recipe.

Daisy

Chrysanthemum leucanthemum
Composite Family, *Asteraceae*

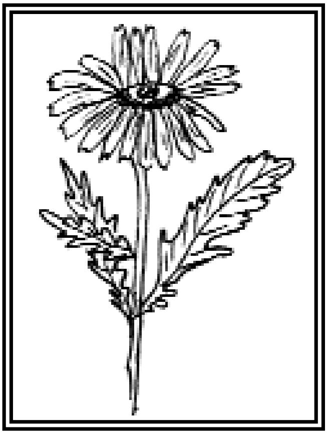

Other Names: She loves me, she loves me not; ox-eye daisy.
History: Native to Europe and Asia, widely naturalized throughout North America.
Habitat: Fields, roadsides, lawn, waste places.
Characteristics: Biennial herbs. White petaled flower about a long, erect stem 18 to 20 inches tall. Flower center has a yellow pebbly disk of seeds. Flower ranges from 1 1/2 to 2 inches in circumference. Petal number varies. Leaves are dark green and

irregularly lobed in a scalloped design. Smooth and succulent, they form a thick basal rosette the first year of growth, sending the stem up with flower the 2nd year.

Primary Uses: Culinary. Petals eaten raw; used as tea. Stem and leaves edible as a salad green, vegetable. Dried stems are added to soups or stews. Leaves eaten raw, put in sandwiches or soups.

Nutritional Value: Leaves very high in beta-carotene, riboflavin, niacin and potassium. Low in carbohydrates and calories. Petals are very high in beta-carotene and niacin.

Collection and Storage: Gather petals only, leaving yellow centers.

Caution: Do not eat the yellow center, as it may cause indigestion. Collect clumps of new spring leaves for salads. Chew leaves or stems while hiking for a taste treat. Dries and freezes well.

Daisy

"She loves me, she loves me not!" We all enjoy and play this game, but think of how much more fun if you popped the succulent white petals in your mouth! Why not make a treat of the petals in salad or gelatin, or even in chocolate ice cream. I have put them in popsicles for kids.

Using the petals for food became a real survival tool when I spent hours picking blueberries or raspberries. They usually grow nearby, and when I found the petals edible, I could vary my diet from fruit to flowers. Then I discovered that daisy leaves are fantastic. From that moment on, Todd and I ate them while walking, playing, and working in the mountains of the Adirondacks. I can never look at a daisy and not remember the freedom of this period of my life!

Dandelion

Taraxacum officinale
Composite Family, *Asteraceae*

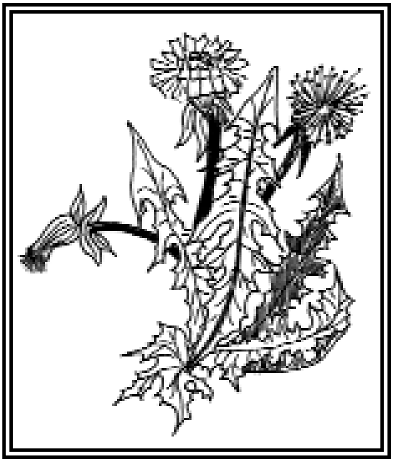

Other Names: Dent-de-lion, wild spinach, wild lettuce, red-seeded dandelion, wasanswak (Crow Indian).
History: Probably native to Europe, but widely distributed through the world. Used by many different cultures. Cultivated on the Mediterranean island of Minorca after locusts destroyed vegetation.
Habitat: Moist areas, lawns, roadsides, fields.
Characteristics: Perennial herb. Grows to height of 2 inches or more; clumped. Leaves saw-toothed; whorl from center. Yellow sectioned flowers mature to fluffy “pompoms” with seeds that blow in the wind.
Primary Uses: Culinary, medicinal. Early roots are eaten raw. Dried whole roots are pulverized or boiled. Leaves are eaten raw; steamed; boiled; sautéed; cooking liquid is drunk. Steeped for tea. Dried and ground for nutrient additive or flour. Flowers are eaten raw; sautéed; fried. Used for wine. Entire plant is steamed; boiled; sautéed; cooking liquid is drunk. Used in soups and stews.
Nutritional Value: High in protein, calcium, and vitamin A.
Medicinal Value: Use as tea for jaundice and other liver problems; a tonic, blood cleanser, purifier, stomatic, diuretic, laxative; for swellings, sores. Also effective for nervousness and hypochondria: an effective relaxant.

To reduce swelling and cleanse a wound, grind fresh leaves to a paste and spread over wound or fracture. Use dandelions as a bandage to bind wounds. Use as blood tonic by pulverizing either roots or leaves and mix 50-50 with whole wheat or rice flour. Make bread.

Cosmetic Value: Facials, wash for eczema; any external cosmetic skin treatments.

Collection and Storage: Gather flower heads in season. Yellow petals are very sweet. Twist the green base from the yellow flower petals and eat the petals raw. Roots can be dug year-round, even under the snow. Young leaves and stems are tastiest. Collect roots in early spring when most succulent and not bitter; dry and grind to coffee. Roots may be stored in strips (see page 30 for illustration on stripping roots). Store whole roots in paper bags in a warm, dry area.

Leaves
- 1/4 grocery-size bag = 8 cups leaves
- 8 cups raw leaves = 10 servings
- 8 cups steamed and frozen leaves = 4 servings
- 8 cups fresh leaves, dried on a screen = 1/2 cup flour

Flowers
- 1/4 grocery-size paper bag = 8 to 10 meals

Linda Says- Security—that is the only word I use for dandelion. Dandelion is always there, even frozen green under the snow, easily found, along edges of tree trunks or rocks. Ruth Spring taught me to core down the root, then peel the white succulent core for the first taste of a spring thaw. "As a tonic," she said, "the body will grab at it."

The former mayor of Vineland, New Jersey, started a large business growing and harvesting dandelions for the gourmet markets of Europe. The annual dandelion dinner he gave was nothing short of a gala feast.

Dock

Rumex crispus
Buckwheat Family, *Portulaceae*

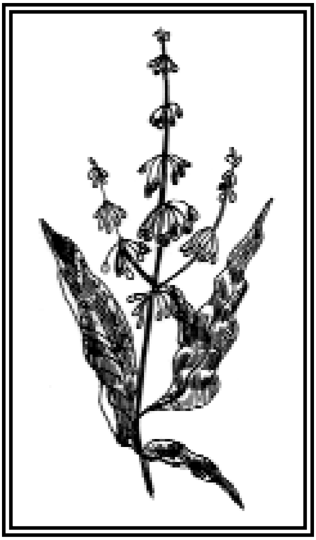

Other Names: Curly dock, yellow dock.
History: Native to Europe, widespread in North America. Pima Indians use roots, stems, and leaves.
Habitat: Fields, roadsides, rich soil.
Characteristics: Perennial herb. Grows to height of 14 inches, may reach 24 inches. Elongated leaves curl; seeds turn dark brown when dry.
Primary Uses: Culinary, medicinal, cosmetic. Roots cooked as carrot. Leaves eaten raw; steamed; boiled; sautéed; cooking liquid is drunk. Stems eaten raw or cooked. Seeds ground for flour. Entire plant makes excellent yellow dye. Seeds are a decorative craft material, useful in flower arranging.
Nutritional Value: High in vitamin A, minerals, and protein.
Medicinal Value: Roots used for skin infections or itches, and for healing wounds and sores. Stems and roots are laxative and tonic. Pulp of leaves used as a poultice for bruises, burns, and swellings. Boiled and mashed, bound on boils and swellings.
Collection and Storage: You can spot dock easily in a field by its deep chocolate-colored tops. Several stands of dock are often in a field; the dark brown tops stand out in peaked clusters, especially above a winter's snow. Gather seeds when rust brown. Leaves are collected in early spring or when young. Roots are gathered in early spring for medicinal storage. Dries and freezes well.

The Department of Agriculture suggests collecting roots in the late summer and splitting while green for drying and grinding.

Linda Says- I remember the summer Ruth Spring showed me dock. When I saw how many thick clusters of seeds there were on just one plant, I decided dock would be a main source of flour. I am constantly amazed at how many quarts of flour can be derived from one clump of dock. Because of the high vitamin A content of curly dock, this flour should be used sparingly only with another flour such as whole wheat. Curly dock seeds have 20,000 units of vitamin A per 1/2 cup. A few teaspoons in your recipe is adequate for vitamin content.

Evening Primrose

Oenothera biennis
Evening Primrose Family,
Onagraceae

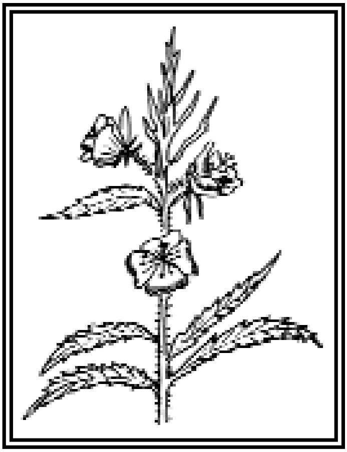

Other Names: Meadow rose.
History: Native to North Americas. Used by Native Americans.
Habitat: Gravelly dry soil, roadsides, waste areas.
Characteristics: Biennial night-flowering herb. Has pinwheel rosette of leaves that are toothless with whitish midrib. Basal leaves brown in spring, hugging the ground. Stalk shoots upward, with 4-petaled flowers on top of multiple branches. With 4 petals, the flowers are not unlike a rose, exceptionally sweet smelling and waxy. Hairy stem. Between stalk bottom and root underground is a strip of red or pink to scarlet indicative of primrose. Single root with many hair-like extensions.
Primary Uses: Culinary, medicinal, cosmetic. Young leaves are eaten as a vegetable or salads. Dried and frozen for storage. Buds edible before flowering. Seeds edible. Roots look and taste like parsnip (peppery flavored), may be eaten raw or cooked.

Dried for long-term storage.
Medicinal Value: Evening primrose oil used for soothing the central nervous system as well as relieving stomach and gastrointestinal stress. See below.
Cosmetic Value: Creams, skin preparations.
Collection and Storage: Pick young leaves in early spring. Pinch flowers off stem. Dig roots and scrub. Roots can be dug in winter.

Linda Says- Rumor has it that oil from this delectable wildflower is one of the most expensive, as well as most beneficial for human beings. Its composition close to the constituents of mother's milk, evening primrose offers many medicinal remedies.

I began picking and eating the flowers in quantity. I use olive oil to make evening primrose oil, just press the flowers into a small jar and add boiling olive oil. (Be careful to avoid splatters.) Press the flowers down with a spoon and pour in oil until flowers are saturated and air has bubbled up. Cap tightly and shake every day or so for a week. Strain the oil and refrigerate. It's a potent tonic when added to your salads.

Filarie

Erodium circutarium
Cranesbill Family, *Geraniaceae*

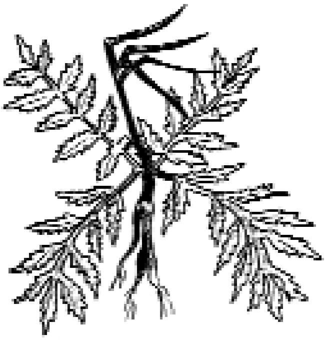

Other Names: alfilaree, alfilaria, redstem, stork's bill, pin clover.
History: Native to Europe and probably introduced to North America by the Spaniards; Native American use.

Habitat: Dry areas, desert areas with a little moisture, particularly clay soils.

Characteristics: Annual or biennial, reproducing via seeds at the base of stork's bill-shaped shoots. Flowers on long stalks shoot up from fern-like leaves lying in thick mats. Minute hair on stalks and stems. Stalks have purple flowers, then seed buds. Seed has a corkscrew tail, driving it into the soil for propagation.

Primary Uses: Culinary. Fern-like leaves are eaten raw like celery. Used in soups, stems, sandwiches. May be steamed and frozen for summer vegetable. Roots are like celery, delicious raw or in soups; steamed or pickled. Both roots and leaves are dried and ground to a flour.

Nutritional Value: Leaves are very high in calcium, potassium and beta carotene, niacin and vitamin C.

Collection and Storage: In the westI used a gloved hand, taking the most succulent leaves only up into your hand, pulling them upwards, and with a twist of your wrist, twist the leaves off the root. This leaves the root and flower stems.

Caution: Do not ingest flowers or stork's bill fruits, as they are known to be toxic. It is generally safe to ingest a large amount of filarie leaves if the vein in the center of the leaf is not pink or red. Filarie tends to pick up nitrates from the desert soil, turning the vein of the leaf as well as the stem of the flower red. Look carefully for healthy plants, green and luscious before eating.

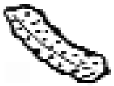

Linda Says- A winter's filarie sandwich has become a ritual. The greens appear after the rains on the Arizona desert. I clip them with scissors, wash them, and make a mild, nutritious raw vegetable served with a dip. Pickled in apple cider vinegar, they make a delicious salad ingredient.

Fireweed

Epilobium angustifolium
Evening-Primrose Family,
Onagraceae

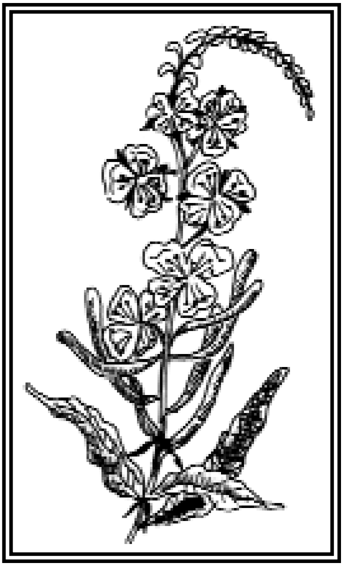

Other Names: Great willow herb, deerthorn, wild asparagus, wickup ocacadii (Potawatomi).
History: Native to North America. Staple food in Northern Asia and Iceland. First plant to re-seed itself after all wars of wasteland or defoliation. Grew on sides of Mt. St. Helen 3 weeks after volcano covered area in potash. Grows back first after fires. The Potawatomi Indians used its root as a soap by lathering it in water.
Habitat: Areas where there has been fire, waste areas, areas of regrowth, roadsides, fields.
Characteristics: Perennial herb with unbranched tall spikes up to 2-6 feet and willowlike. Spikes not unlike upside down asparagus. Stems hollow with white pith.
Primary Uses: Culinary. Pith is eaten raw or boiled; used in soups and stews; steeped for tea. Buds are eaten like asparagus; tips picked before they flower are eaten raw or used in soups and stews. Flowers may be eaten, but buds are more palatable. Young leaves eaten raw or prepared as spinach; dried for tea. The fluff makes excellent winter tinder; used as cotton.
Nutritional Value: Exceptionally high in beta carotene and potassium as well as most minerals.
Collection and Storage: Perennial herb. Reaches height of 2 feet or more. Spike of magenta or pink 4-petaled flowers; elongated top of buds.

Linda Says- The best way to see the old fireweed beds is when it is snowy. The plants' spiral top, tall and fuzzy, swirls with cotton fuzz. In season, fireweed shows a magenta top with flowers. Above the flowers are asparagus-like tips that I harvest as a delicious vegetable. They taste best fresh.

Goldenrod

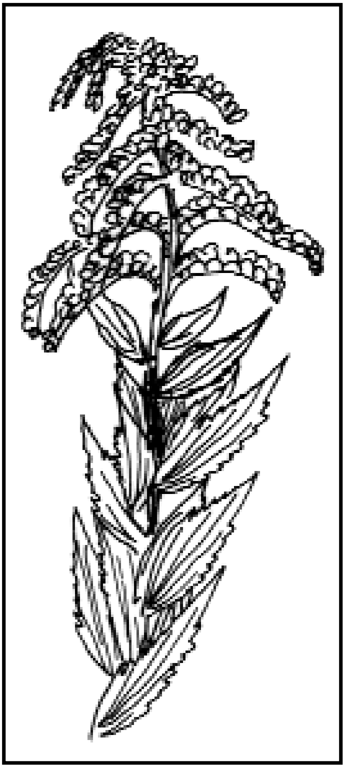

Solidago odora
Sunflower Family, *Compositae*

Other Names: Golden elixir, sweet goldenrod, asawush (Ojibive Indians). Poco moonshine is yet another species, *augustifolum*.
History: Native of North America. Used by pioneers. Ojibive Indians used it in tea as a carminative, antispasmodic, and intestinal astringent (Smith 1932).
Habitat: Dry, open fields, woods, road banks, pine barrens; acidic soil.
Characteristics: Perennial herb. Grows to height of 24 inches or more. Hundreds of tiny fluffy yellow flowers on top stalk. Flowers smell like anise.
Primary Uses: Culinary, medicinal, cosmetics.
Roots eaten raw; dried and pulverized for tonic powder. Stems dried for tonic nibble. Leaves picked and eaten fresh. Dried for tonic tea. Stems, buds, flowers, and seeds eaten or used as tea.
Medicinal Value: Energizer and tonic. As a tea, for diarrhea, bladder tone, colic. Mashed leaves for a poultice, raw flowers

chewed for sore throat. Leaves and roots are tonic for colds and upper respiratory problems. Powdered leaves are sprinkled in sores and wounds because goldenrod is an astringent. Helps stop internal hemorrhaging as a strong tea.
Cosmetic Value: Shampoo, astringent with diaphoretic qualities; aromatic and stimulating.
Collection and Storage: Gather the entire plant, bundle it, and hang it upside down until dry in a warm, airy place out of the sunlight. The bundles are dry when the stems snap easily. Place a clean sheet under the bundles. Hand strip and gather up the bounty. To store a winter's supply of tea, gather the stripped and dried goldenrod parts in a pillowcase or other cloth bag, then wring to crush the leaves. A handful of flowers and leaves will brew 6 to 8 cups of tea.
Caution: Do not use as bulk food. Use sparingly for medicinal purposes.

Linda Says- So easy to gather and bundle, the golden elixir grows thick and bountiful. Although its tonic qualities are present any time, the goodness is strongest just before the pollen forms in the flower head. If you choose to collect at this time, watch out for the bees! They love the flowers and frequently hide in the caps or rest in them.

For a perfect yellow tea, snip off the blossoms. Dry separately on a screen and store in glass bottles.

Grape

Vitis arizonica (Western)
Vitis labrusca (Eastern)
Vitis rotundifolia (South)
Vitis Family, *Vitacere*

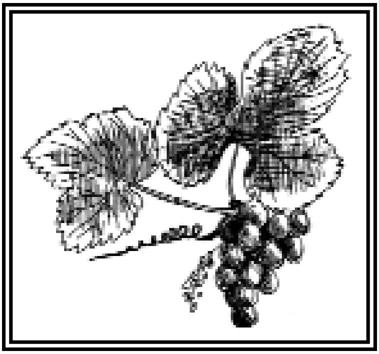

Other Names: Wild grape.
History: Native to North America.
Habitat: Thickets, edges of woods.
Characteristics: Deciduous vine. Forked tendrils at end and joints of high-climbing vines. Branches have a brown pith. Leaves tiny to large, coarse, saw-toothed, lobed, heart shaped. Fruit round, fleshy. One to 4 pear-shaped seeds in several colors from purple or black to amber.
Primary Uses: Culinary. Fruit cooked to make jellies, jams, juice, syrups, and flavorings. Leaves are used as a wrapper for main dishes. Leaves may be eaten raw, stir-fried, or steamed. Tendrils are pickled.
Nutritional Value: Fruit is high in vitamin C.
Collection and Storage: Pick tendrils all year; clip with scissors. Grapes in season may be gathered by hand. Young leaves are picked at any time. Freeze leaves for stuffing; laying between sheets of wax paper, folding paper after each layer.

Linda Says- I wandered the hills of Syracuse, New York, with a famous gourmet cook looking for grape tendrils. There they were, curling down from yards of grape vines. I rolled and sniffed them first for their characteristic grape smell, then used scissors to clip hundreds of tendrils into a bowl. Later, my friend and I stir-fried the tendrils into delicacies, as well as froze them for use later in the year. The tendrils make me pucker.

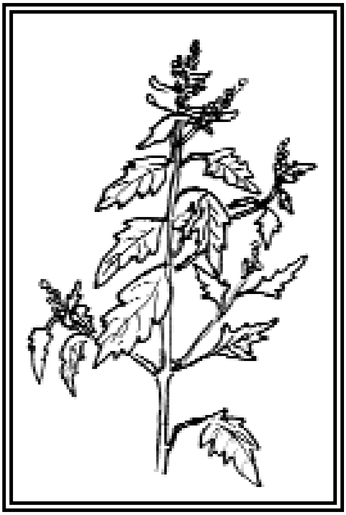

Lamb's Quarters

Chenopodium album
Goosefoot Family, *Chenopodiaceae*

Other Names: Pigweed, goosefoot, wild spinach, chiciquelite, hwahai (Pima Indian). (Toakush Crow Indian).
History: Used in most countries and most civilizations. Recent archaeological findings in Denmark show it was an early food there. Now cultivated by Navaho and Mexicans. The Zapotec Indians use the flour as a staple food.
Habitat: Native to North America, Europe, Asia.
Characteristics: Annual herb. Grows to height of 18 inches or more. Leaf shaped like goose's foot, dark green with whitish underlay. New leaves have white or lavender-tinged powder near center whorl. Leaf beads with water when wet. Green seed-like flower clusters.
Primary Uses: Culinary, medicinal. Young shoots are boiled, cooking water is drunk. Leaves are spinach-like; eaten raw; steamed; boiled; sautéed; cooking liquid is drunk. Dried and ground for flour. Flower and seeds are ground for mush or cereal; sprouted, dried and ground for flour. Seed is excellent food for wild birds and canaries. Roots can be used as a soap substitute; wet hands and rub cleaned root. Young stems eaten raw or cooked.
Nutritional Value: High in nutrition, vitamins, calcium.
Medicinal Value: All parts used as a poultice for swelling, rheumatism, arthritis. Chewed raw for toothaches. Also, gelatin capsules filled with lamb's quarters are a potent vitamin.

Collection and Storage: Harvest large amounts with a long knife or small amounts with scissors and bag. Hang large bunches upside down to dry. When crispy dry store in glass for longevity or steam and freeze.

- 1 large bunch (stems bunched together to about 1 inch thickness) = 2 cups flour when dried and stripped.
- 1 large screen of leaves and seeds = 2 cups flour
- 1 large bunch chopped and steamed = approx. 12 to 14 sandwich sized plastic bags.

Caution: The lamb's quarters' leaf resembles poisonous, malodorous look-alike, nettleleaf goosefoot, which has a tough cardboard-like leaf. Crush and smell leaf. The rank odor will tell you immediately if you have this look-alike. Identify carefully. See page 294.

Comments: Chenopodium ambrosiodes; also known as Jerusalem oak, wormseed or Mexican goosefoot, is a type of lamb's quarters found in Arizona and other Western states. A sprig added to beans renders them anti-flatulent.

Lamb's Quarters

During World War II, the US government gave out information that garden weeds were good food. The top weeds included amaranth and lamb's quarters. Not only are these plants nutritious, but they are not bothered by cutworms or aphids.

Lamb's quarters grows phenomenally fast. In fact, just one plant yields all I need for a summer's meals for me! To gather, I pinch the tops and tips of each branch. I wait two or three days, then repeat; for each top taken, two or three will grow in its place. So just a few plants can yield an entire fall and winter supply.

Dried leaves make a delicious and nutritious flour. Mix it with a bit of water to make a tortilla, or combine with wheat flour to make an enriched pita bread or fry bread.

To make the flour, I pound the dried leaves with my favorite rock, but you can feed the leaves through a hand grinder for the same results. You can also place the crispy dried leaves in a pillow case and wring.

Malva

Malva Neglecta
Mallow family, *Malvaceae*

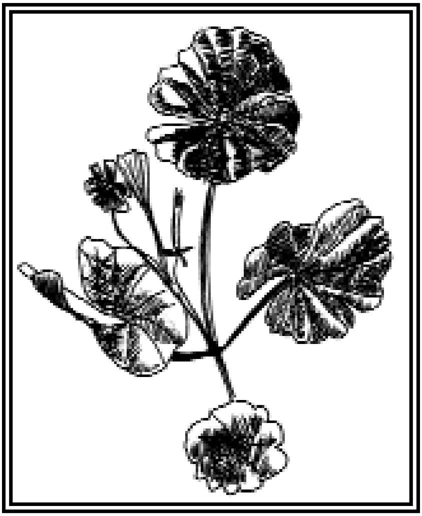

Other Names:
Cheeseweed, button weed, cheeses, mallow.
History: Introduced from Europe; sold in the marketplaces of China and Japan.
Habitat: Fields, roadsides, waste areas.
Characteristics: Perennial or biennial herb. A stout, bushy plant branched and spreading from its base. Reaches height from 2 feet to 4 feet.
Primary Uses: Culinary, medicinal. Leaves used in soups, main dishes, stir-fried. Stems are vegetable stir-fried in garlic and olive oil. Flowers, buds and stems, and cheeses (seed cases) are all edible raw, steamed, sautéed, or dried and stored.
Nutritional Value: Very high in iron and calcium.
Medicinal Value: A mucilagenic, increasing the flow of saliva. Chewed for sore throats, upper respiratory problems.
Collection and Storage: Clip whole stems, pick off buds and leaves. Cut stems into sections for stir-fry or freezing. Cut leaves as large as you can for rolling around tofu, fried beans, potatoes, rice. Collect in bunches for abundance. Do not pick malva with red stems. Red stems may depict nitrate from the soil; pick fresh young plants when possible. Mince malva leaves and marinate in oil and vinegar before cooking.

Add 1 cup salsa to ice cube trays, a spoonful of fried beans, roll tightly in malva leaves and freeze. Easy to defrost and bake at a later date. (See recipe for Malva Rice Roll-Ups.)

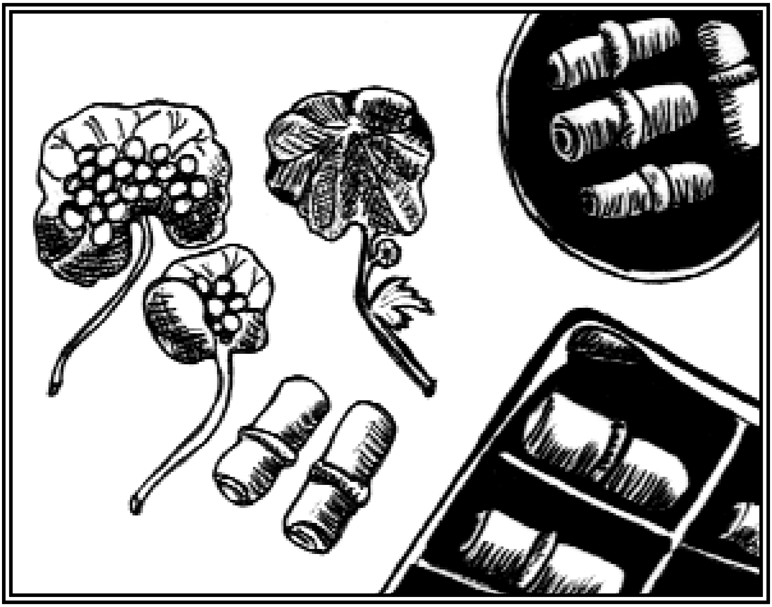

Stuffed Malva Leaves

Linda Says- Malva has cheese-wheel seeds. I love to gather them and dry them for potato salads, stir-fries, or eat them as they are.

Seeds are dried to use as "nuts" in winter. Large leaves used for wrapping foods. Stems are used for delicate sautéed vegetable, and younger leaves for salads and sandwiches. Once dried, all parts of malva can be ground for flour! I cook with malva flour a lot. As a general rule, I use 1 cup malva flour to 3 cups of wheat or other gluten flour.

Maple

Aceraceae Family

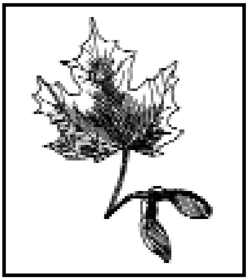

Other Names: Sugar, silver, soft, rock and red maple.
History: Many species native to United States. Native Americans were early users of the tree's products. Chippewa Indians pressed the seeds into cakes calling them maple cakes.
Habitat: All environments except lower Rocky Mountains and the Plains states; areas with rich, well drained soil.
Characteristics: Deciduous tree. Most leaves have 5 pointed lobes, are irregular toothed, and are opposite on branches. Leaves turn golden yellow to scarlet in early fall. Seeds have brown wings and occur in pairs. All have sweet sap to a varying degree.
Primary Uses: Culinary. Sap used for syrup, candy, flavorings. Seeds eaten raw or pressed into cakes. Twigs and young leaves used for tea and emergency food.
Nutritional Value: Leaves high in minerals, beta carotene and vegetable protein.
Collection and Storage: Cut the seed off the wing with scissors, dry on a screen, and put through the grinder. Twigs, inner bark, and sap can be collected and used in winter. Collections of twigs is simple because they snap off easily by hand. Syrup collection is through a tap drilled into tree using a collector pail. (Author used plastic milk bottles from a nearby dump.) Please refer to other references to details of syrup collection.

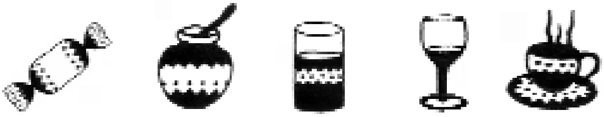

Maples in Winter

I was relaxing in front of a fire in the crispness of early morning when Crack! A sound like an explosion came from behind me in the woods. I scanned the trees and saw that a maple tree had "exploded." The explosion caused a big crack in the tree about three feet high. When a winter wind stirs the frozen trees, they sometimes appear to burst vertically. When it was 40 degrees below zero at night, I lay awake and listened to trees explode. That's a true wilderness thermometer!

Meadowsweet

Filipendula ulmaria
Rose Family, *Rosaceae*

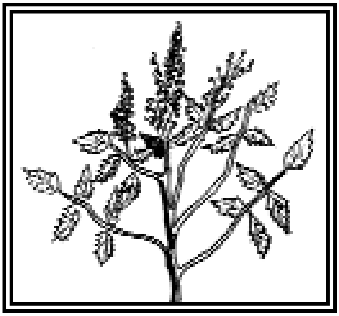

Other Names: Queen-of-the-meadow, meadow sweet.
History: Native to North America, used by native Americans.
Habitat: Meadows, old pastures from Nova Scotia and New Brunswick to North Carolina and west to Minnesota and Arkansas.
Characteristics: Perennial herb. Bush with symmetrical branches, 30 inches long or more. Flower spires, shaped like church steeples, are white and fluffy in spring. The steeples turn seedy and chocolate brown in the winter, giving it a characteristic look against the snow. Twigs have hundreds of tiny nodes between the branches and are characteristic of meadowsweet, not other hardhacks.
Primary Uses: Culinary. Twigs chewed or steeped as tea. Dried and stored for winter. Flowers are dried as a sugar substitute; sweet enough for teas & cereals, but flower must be dried. Used in wines. Seeds are also edible. Leaves used for tea.
Nutritional Value: High in vitamins and minerals. Also high in beta carotene, potassium, and niacin.
Medicinal Value: Mineral tonic. Flu and Rheumatic pains. (contains salicylates.) Twigs, inner bark and sap can be used in winter.
Collection and Storage: Collect twigs by breaking off when young and store in paper bag. Collect flowers by pinching off easily with fingers. Dry on screens.

Linda Says- Shearing meadowsweet flowers with scissors is by far the quickest way to harvest them. Pick with both hands, breaking brittle twigs. Swing a handful from your right hand up under your left armpit. Hold tight and continue until bunch is cumbersome, then tie and set aside. Gather bunches and tie together.

Milk Thistle

Silybum marianum
Composite Family,
Compositae

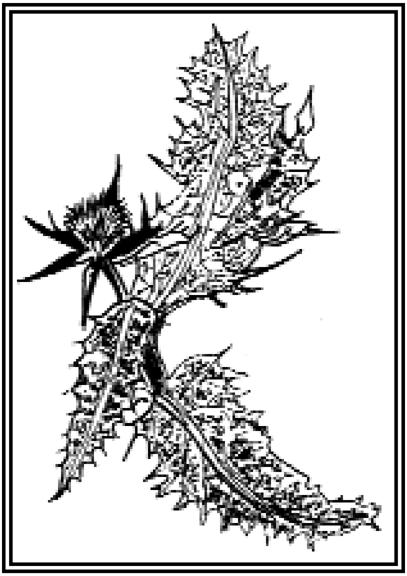

Other Names: Horse thistle, marian, holy thistle.
History: Native to the Mediterranean, also hot regions of North America, including Southern Arizona. Cultivated and dried in Europe. Used by Native Americans.
Habitat: In west, commonly found in ditches; waste places.
Characteristics: Annual herb. Grows 3 to 6 feet tall. White milk splotches on prickly leaves, named “Mother Mary’s milk.” Large prickly, fuzzy thistle flower with long prickers from central stem. Vigorous grower; 1 plant will quickly spread to a bed 3 feet around.
Primary Uses: Culinary, medicinal. Young leaves parboiled to remove prickers, for soup or stir-fry. Cooking liquid is drunk. Leaves dried and ground for extract.
Nutritional Value: High in vitamins and minerals. High in phosphorous, vitamin K, and thiamin and niacin.
Medicinal Value: For treatment of liver disease. A valuable source of the chemical silymarin. Mineral tonic, flu and rheumatic pains (contains salycylates).
Collection and Storage: Use a long blade or sharp machete to cut off prickly leaves. Extract young leaves from center of plant with long knife and tongs. Plant tends to pull up nitrates from the soil, so harvest young leaves. I dry the leaves by placing them face down on a screen. You can handle prickly leaf safely by pinching the back vein of the leaf.

Linda Says- The gigantic prickles on the milk thistle once scared me. I was aware of the multiple benefits of these ominous leaves, but how to gather them? Gloves did not help. The best tools are a long-handled tongs and a long kitchen knife. Slice the leaf off with the knife and pick it up with your tongs and bag it. Washing them becomes the next challenge, and cooking to a safe texture is yet another. The leaves fairly "melt" in cooking; even the barbs become tender if the leaf is simmered long enough. I cut off the barbs with scissors, then simmer a few minutes and drink the broth.

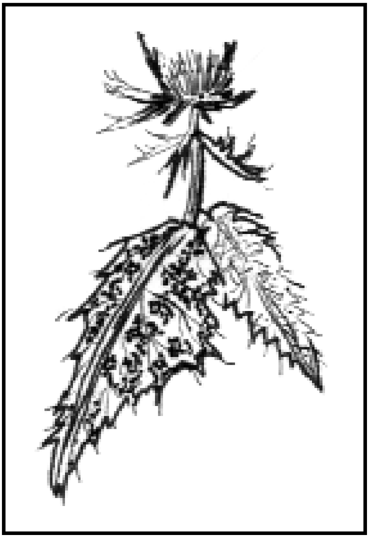

Milk Thistle

Milkweed

Asclepias species
Milkweed Family,
Aslepiadaceae

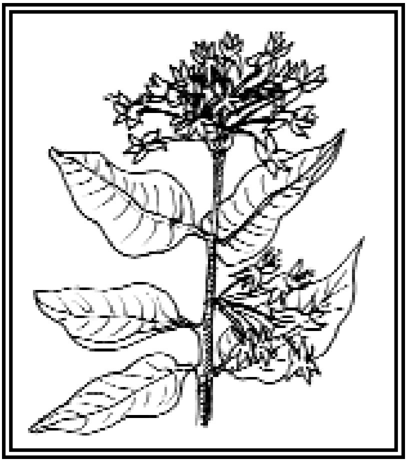

Other Names: Silkweed, butterfly weed, cotton tree, pleurisy root.
History: Native perennial with many species. Known since ancient times, used during World War II in rubber experiments, fluff considered for life preserver stuffing.
Habitat: Fields, roadsides, gardens.
Characteristics: Perennial herb. Single stem 2 to 5 feet high. Leaves opposite in pairs from 4 to 9 inches long. Heavily veined, smooth leaf with white milky sap throughout plant. Flowers are round clusters of pleasantly fragrant single stars at end of dozens of branches, making a round ball of blossoms. Flowers give rise to a green bristly pod, in which fluff-bearing seeds form. White sap exudes from all parts, including roots.
Primary Uses: Culinary and medicinal. Flowers and seeds are edible. New sprouts are used as vegetable, cooked. Young leaves and buds are cooked as vegetable. Stalks are peeled for strong fibers; used for fishing line, especially if braided. Bark is braided into ropes of varying thickness. Pods picked young are used in soups and stews.
Nutritional Value: Shoots high in niacin, potassium and vegetable protein.
Medicinal Value: Milky sap used in wart removal.
Collection and Storage: Break off buds, young leaves, and flowers by hand. Use a paper bag. The milk tends to turn the plant sour if kept in plastic bags. Strip stem fibers and braid for emergency fishing line or rope, as well as shoelaces (in the wild).

As a vegetable; young leaves and buds can be dried or frozen. Sprouts can be frozen. Buds should be simmered twice for 3 to 4 minutes, changing water in between. Roll any wide leaf with goodies and bake in salsa for a wild Mexican delight.

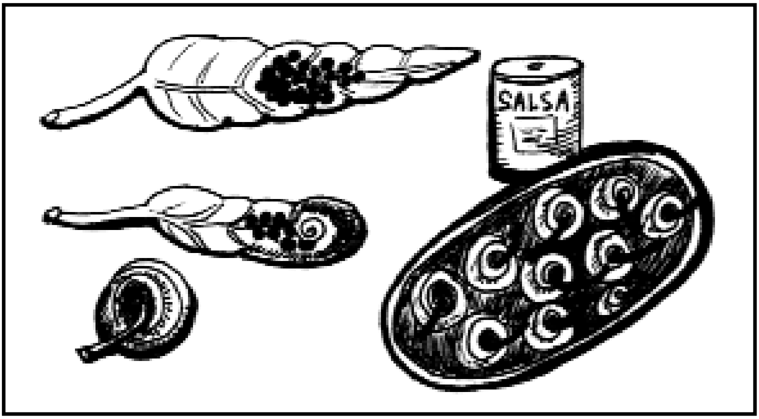

Wild Mexican delight!

Milkweed Abundance

Gathering milkweed is by far the easiest way to find food for my family. The abundant pebbly tops beckon to me. I move from every fourth plant or so with ease. I fill a large bag and then go to the Cedar River log bridge to wash them. The time spent swishing the milkweed buds in the cool water are some of the most precious moments of homesteading.

Mint

Mentha species
Labiatae family

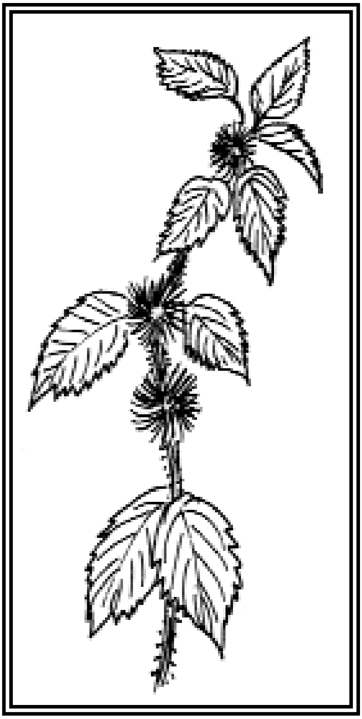

Other Names: Spearmint, peppermint, horehound, giant hyssop, wild bergamont, bee balm, oswego tea.
History: Mostly native to North America and widely distributed throughout the world. Used since ancient times in scents, cosmetics, flavorings, food and medicines.
Habitat: Found in most areas from swamps to fields, gardens to lawns.
Characteristics: Perennial herb. All mints are aromatic. Most mints have square stems; leaves are opposite with usual unevenly saw-toothed edge. Prominent veins; leaves may be smooth or coarse, sharply pointed in most species. Characteristic fuzzy leaves in wild varieties, especially swamp mints. Plants range from 5 inches to 2 feet high.
Primary Uses: Culinary, medicinal, cosmetic. Leaves, flowers, stems and roots eaten raw in salads, candies; leaves steeped for tea. Flowers dried for teas. Whole plant ground for aromatic tea.
Nutritional Value: High in vitamin C and iron.

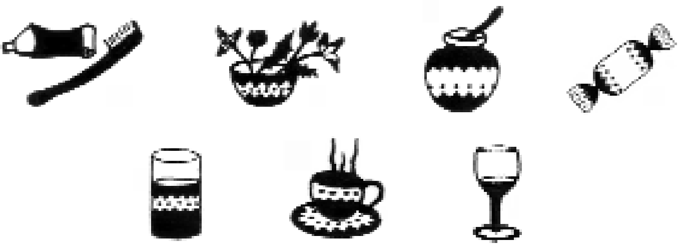

Medicinal Value: Well known as a stomachic, or stomach soother. Also used for insomnia, stomach ache, fevers and colds.
Cosmetic Value: Aromatic stimulant, astringent, restorative; used for scents, in soups and potpourris.
Collection and Storage: Cut stems where desired, as plant rejuvenates quickly. I used to pick and wind one plant into a bunch, then string bunches together for ease of harvesting as well as drying. Also, freeze flowers in ice for pretty edible cubes. Mints transplant easily and have characteristic prolific growth in an herb garden.

Linda Says- The camp always smelled good when the mint was drying for winter. Our family eats the leaves often. Since a leaf holds a whopping amount of iron, it is understandable why I eat at least 1/2 cupful every time I go out to pick it.

Mint Aromas

Lightning flashing all around me, I lay prostrate in a large patch of mint. The air was blue-black. Tired and clutching my bundle of freshly picked plants, I almost welcomed the respite, though I was a bit nervous.

Mint picking is always done around swamps, streams, and riverbeds, and this day I was able to gather a six-week supply before a typical Adirondack thunderstorm kept me waiting for 30 minutes.

Mullein

Verbascum thapsus
Snapdragon Family,
Scrophulariaceae

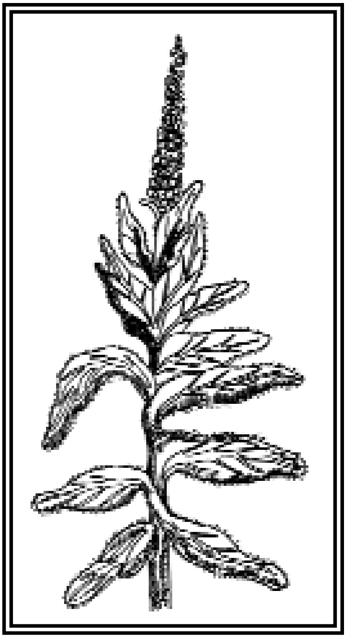

Other Names: Candlewick, flannel leaf, Aaron's rod, feltwort, hare's beard; velvet plant, velvet dock, blanket leaf.

History: Native to Europe, naturalized throughout pasture land of United States. Plant also known in Roman and Greek history; Quakers, Native Americans.

Habitat: Waste wet areas as well as roadsides, gravel-filled fields.

Characteristics: Biennial herb. Tall spikes with yellow flowers, velvety leaves whirl around spike. Grows up to 10 feet tall. Produces basal leaves in first year, then spike the second year. Leaves furry or velvety, ranging from 3 inches to 2 feet long. Flowers are succulent and sweet smelling, attracting bees.

Primary Uses: Culinary, medicinal, cosmetic. Root used for medicinal tea, leaves and flowers dried for tea and medicine. A normal root, up to 1 foot long, may be reused dozens of times. I put large leaves in my sneakers when walking long distances; they help soothe tired feet, and prevent blisters. Native Americans put leaves on cradle boards to heal elbows and shoulder blades. Antiseptic, mullein is invaluable for inflammation.

Medicinal Value: All parts of mullein, from single tap root to flowers have an antihistamine quality among other qualities. (See chart page 53 for comparing mullein to aloe vera.)

Cosmetic Value: Astringent and emollient properties, decoction of flowers for hair rinses to lighten hair color; flowers steeped in olive oil used as application for sores or massaging oils.

Collection and Storage: Second-year leaves may be harvested from the spike individually, flowers harvested individually, and root harvested at any time, any season. I harvest single leaves, roll them up, and dry them on a screen. These rolls may be kept in glass containers and taken along when hiking.

Floated in olive oil, mullein wicks will burn cleanly & clearly for 3 hours or more or until the olive oil is used up. Pour a tablespoon of olive oil on the top of a glass of water. (You must float the mullein by crossing 2 toothpicks, poking the ends of the toothpicks in a tiny piece of foam or buoyant material.) Place the wick in the center, letting a piece touch the oil you have poured on the top of the water.

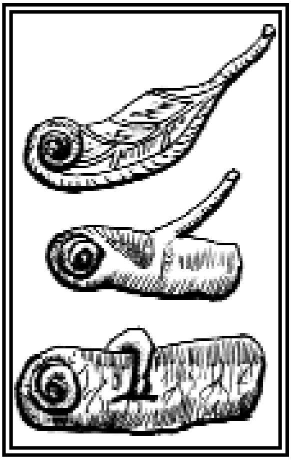

Figure 48 Rolling and Drying Mullein leaves.

Linda Says- I used mullein for lamp wicks when snowed in. A small piece of the flannel leaf cut to size, dried, and soaked in oil will burn for quite a while.

Mullein

I feel as if the 6-foot high plant were a person. I address the mullein and ask to pick its flannel leaves and flowers!

One summer, my Girl Scout Troop was horseback riding on a field trip in Indian Lake, New York. One of the horses moved too close to the one in front and the forward horse kicked, clipping the child on the ankle. It was obvious she had a bad injury. Removing her shoe, I shouted to get mullein, quick! The scouts came running with fresh mullein. "Roll the leaves and crush them quickly with a rock, girls." Poultices were made and applied on the swollen ankle. The mullein stalks became splints and the leaves became bandages until we could get the small scout to a hospital. Soon thereafter a doctor called to ask "What was that plant around the scout's ankle? She had a chip fracture but no swelling or pain!" Mullein, sir!

Mustard

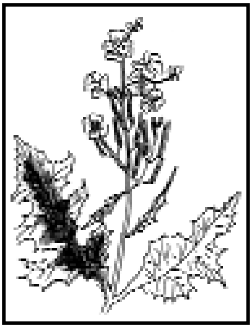

Brassica species
Mustard Family, *Brassicerceae*

Other Names: Charlock mustard, field mustard, yellow mustard.
History: Native to Europe, pasture land in North America. Used medicinally in Greek, Anglo-Saxon history. A pot herb in English history.
Habitat: Fields, roadsides in cooler regions.
Characteristics: Annual herb. Grows to height of 10 feet or more. Leaves rounded, with extra protuberances of tiny leaves below main part of leaf. Crushed leaf yields herby pungent smell. Four-petaled yellow flower; petals in form of cross. Black seeds in pod.
Primary Uses: Culinary. Stems used raw as pungent spice. Leaves all eaten raw, used as pungent spice; steamed, boiled in soups, stews. Flowers are eaten raw; steamed. Seeds dried and used as spice; ground for mustard.
Nutritional Value: High in beta-carotene, minerals, vegetable protein.
Medical Value: Dried leaves and flowers used as a poultice, or mustard plaster for respiratory distress.
Collection and Storage: Collect young leaves by snipping off stalk with fingers. Collect pods separately in season when seeds are mature, pods almost dry.

Linda Says- When eating mustard greens boil in water for 15-20 minutes. Mustard is a very pungent vegetable.

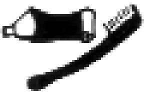

Mustard

As soon as the temperature rose to between 32 and 40 degrees, I would make a beeline to the back of Alice and Walt Sherman's barn. The mustard plants were always green, growing beneath the snow against the warm barn boards. There are no words to describe the taste of the first mustard leaves of the season. My body fairly roars with energy for days to follow!

Nettles

Urtica species
Urticaceae Family

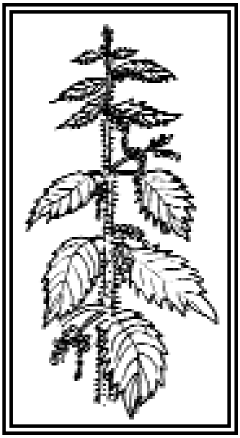

Other Names: Stinging, slender, great, dwarf, nettle.
History: Native to Europe, widely naturalized in North America. Used worldwide.
Habitat: Fields, fertile soil from Alaska throughout continental US.
Characteristics: Perennial herb. Leafy plant. Leaves opposite in pairs, coarsely veined, oblong tapered to tip with rough, sharp saw-toothed edges. Entire plant fuzzy with tiny stinging bristles or minute hairs. Flowers are tiny green seed-like clusters between stalk and branches.
Primary Uses: Culinary, medicinal, cosmetic. Leaves and flowers boiled for soups and stews, cooked in stir-fries. Seeds are used in herbal teas. Excellent tea!
Nutritional Value: High in protein, iron and vitamin C.
Medicinal Value: Tea well known as a nutritious restorative tonic.
Cosmetic Value: Commercial uses: nettle shampoo, herb wraps.
Collection and Storage: Use leather gloves when collecting nettles. When harvesting for vegetable, hold a bag or bowl under separate leaves and clip off with a pair of scissors. A few plants sustain a whole family, since several nettle leaves grow where pruned.

Clip seeds separately if desired, plant will continue to produce until mature. Place nettles in water to simmer away hair and leave a nutritious main dish.

Linda says- I can still feel that prickly burn on my leg from each tiny hair as I approach the tall, stately plants. I learned my lesson. Blessing the "armored" leaves, I soon practiced safe collection. My young scout troop once gathered a bunch of nettles and presented a "bouquet" to Utica's city scouts; I'm sure their lesson became the same painful memory.

Phragmities

Phragmities communis
Grass Family, *Graminae*

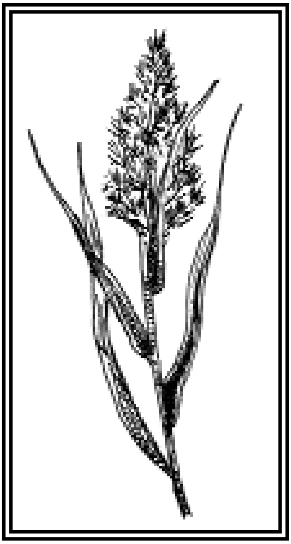

Other Names: Reed, giant reed.
History: Native to Europe and North America. An ancient food, craft material, fiber. The Gosiute Indians gathered a sweet secretion formed on the leaves by aphids and used it as sugar (Chamberlain 1911). The Panamint Indians dried the entire reed, ground it, and sifted out the flour (Colville 1892). This moist and sticky substance was then set near a fire until it swelled and browned, when it would be eaten like taffy. The best time to gather the culms (stalks) was spring.
Habitat: Grows throughout the United States except inland areas of South Atlantic and south-central states; also Europe, Asia. Likes marshes, swamps, wet areas, roadsides.
Characteristics: Tall, reed-like stalks with feathery seed tips. Up to 6-8 feet tall.
Primary Uses: Culinary. Stalks are dried and ground to flour. Gummy substance in the stalk can be used as gum or in drinks. Stems woven for mats, bedding, thatching, grass huts. Excellent tinder. Fibers in stalk made into ropes by twisting together. Seed heads are made into a gruel, cereal, or flour. Used in cellulose, arrow shafts. If seeds are not present, chaff is ground into flour. Roots or rhizome (underground stems) are gathered and ground into flour. Eaten raw, roasted, or boiled into a sweet, gummy confection not unlike candy or marshmallow.
Collection and Storage: Study the cycle of head (seed) maturation. There is a 2-week period when seed head is full and ripe. Past that stage the seeds drop and blow away. Clip heads with scissors in the middle of summer, before seeds mature and fall out of seed head.
Caution: The water in which phragmities grows is often polluted, so test the water before eating the plant.

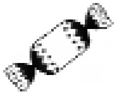

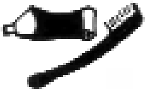

Linda says- Imagine my surprise when told that the waving reed was edible and could be ground to flour! It wasn't long before I began to experiment. The thought that such a tremendous proportions of the world was home to this plant, but few people were baking with it became another reason for this book. (The water in which phragmities grows is often polluted, so test the water before eating the plant.)

Pine

Pinus species
Pine Family, *Pinaceae*

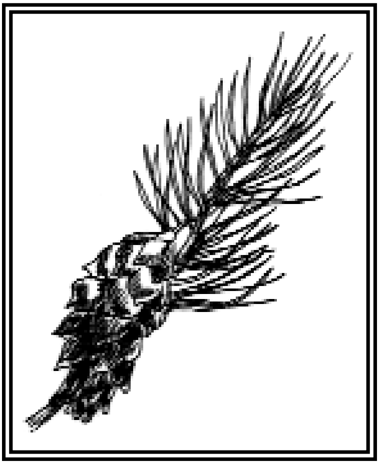

Other Names: Too numerous to mention, but including white pine in East and pinyon in West.
History: Native to most parts of the world. Used by Native Americans.
Habitat: Woods, forests, landscaped area; soil varies from dry to moist.
Characteristics: Evergreen tree. Cone-bearing. Needles vary in length, usually long and slender occurring along the twigs. Bundles of 2 to 5 needles common. (White pine with 5 needles shown in picture.)
Primary Uses: Culinary, medicinal, cosmetic. Needles eaten raw or cooked as vegetable; tea, flour. Catkins eaten as candy; raw, dried for tea, salads. Cone and nuts (seeds) are eaten as snack; ingredient in baked goods. Bark is emergency food, roots are eaten raw, steeped for tea. Sap has medicinal application for small cuts, scrapes, blisters.
Nutritional Value: High in Vitamin C.
Medicinal Value: Colds, flu; as a tea, mixed with honey.
Cosmetic Value: Deodorant, air freshener; soaps, lotions, cleansing oils.
Collection and Storage: Gather needles, twigs, and bark. Using a knife, peel a section of bark lengthwise from a young branch; this may be chewed for a long time. (See “Peeling bark in strips” on page 31.) To harvest needles, let branches dry thoroughly

until needles fall. Gather needles in a sheet, then break up twigs into small sections. Place needles in glass jars. Preheat oven to 300 degrees, then turn off! Place jar with needles in oven and leave until the oven is cool. Cap immediately and store.
To do twigs, place them in a glass jar, leaving the cap off. Put in a turned off 300° F. oven, leaving until cool. Put in a sunny area for further drying, if needed (until they snap easily). Cap for storage.

Caution: Needles are volatile and catch fire easily.

> ***Linda says-*** Many was the time I ploughed through snowdrifts looking for meadowsweet, maple and beech twigs, and succulent fresh pine needles. Pine provides the energy to go on! It is the taste and smell of the forest. Save every bit of your Christmas tree for pine tea if it has not been sprayed or dyed a darker green.

* * *

Plantain

Plantago major (common plantain)
P. Juncoides (Long-Leafed Plantain)
Plantain Family, *Plantaginaceae*

Other Names: Turnip leaf, cart-tract plant, cuckoo's bread, English plantain, goosetongue, Indian wheat, pale plantain, ribwort, rippleseed plantain, seaside plantain, seashore plantain, snake weed, soldiers herb, White man's footprint.

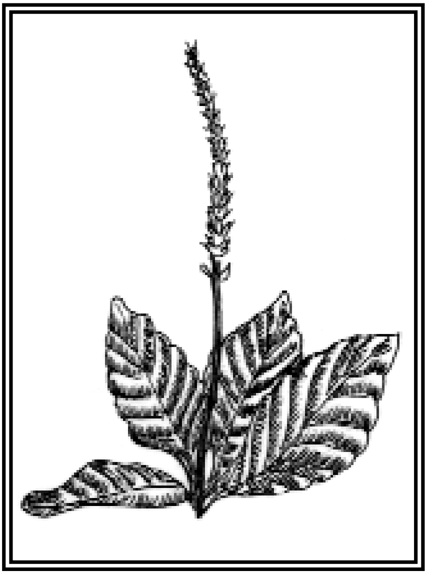

Common Plantain

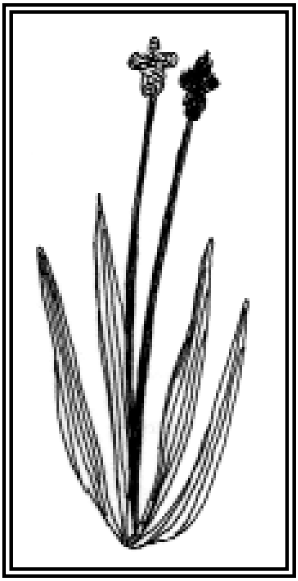

Long-leaf Plantain.

History: Many types native to Europe and naturalized in North America. Used as vegetable in France; cultivated in England. Some varieties native to Asia.

Habitat: Lawns, roadsides, fields, waste areas.

Characteristics: Biennial or annual herb. Turnip-like flat roundish leaves, with prominent parallel veins. Center stalk lined with seeds. Considered a pot herb in Europe, this plant has many species. Common plantain has spade-like flat leaves, whorling on the ground from a center point, with tall stalk in center with many seeds around stalk up to tip. Long-leafed plantain has long narrow leaves with heavy vein and similar texture, taste, and smell. Stalk has seeds at only top inch or so.

Primary Uses: Culinary, cosmetic. Young leaves are eaten raw or cooked. Stalks are dried for vegetable chewies. Seeds are eaten as sprouts. Entire plant is edible.

Collection and Storage: Turnip-like leaves are easily twisted from the ground by hand. The stalks are clipped and dried in a day or so. Use scissors. Leave stalks until the seeds are easily removed. Save stalks for a crispy stick food.

Linda says- The plantain leaf I eat is also a fantastic underarm deodorant! Prepare the leaves by placing 3 leaves together, then crush with a rock or hammer until the juices flow. Remove the inside leaf and wipe juice on underarm. Free supply abounds under my feet. Plantain has surrounded all the camps I have lived in throughout my homestead years.

Prickly Pear

Opuntia megacantha and other species
Cactus Family, *Cactaceae*

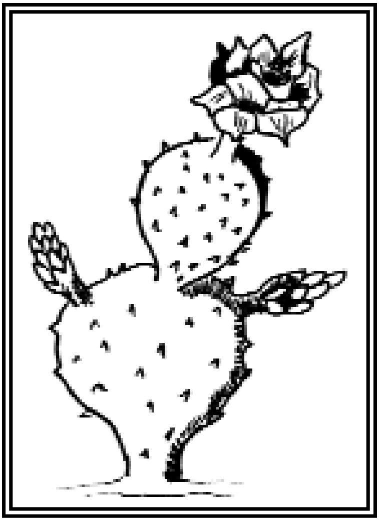

Other Names: Devil's tongue, beavertail, Indian fig, tuna.
History: Native to North and South America. Used by Native Americans.
Habitat: Hot, dry, well drained soil.
Characteristics: Fleshy, low-spreading cactus. Has flat stems and pads with long needles. Thorny fruits are preceded by waxy flowers, yellow or red. Fruits ripe for harvesting are yellow-green to purplish black.
Primary Uses: Culinary, cosmetic. Pads are cut up for food; strips parboiled and french-fried; simmered for hair tonic and softener. Flowers eaten raw, cooked to candy, jelly; also frozen in cubes for drinks. Buds cooked for jellies, jams, syrups, deserts, candies; juices also made. Water from fruit and young pads may be used as a liquid in an emergency. Seeds may be dried and ground to flour; used as a soup thickener.
Nutritional Value: High in carbohydrates and calories.
Medicinal Value: Pads split open for soothing gel for burns, wounds, sunburn.
Cosmetic Value: Hair tonic, softener.
Collection and Storage: Handle pads only with leather gloves and tongs. To harvest, use a long stick, handle, or shovel to knock off each pad. With a deliberate touch-and-push movement, the pads break off and fall to the ground with ease. When you pick them up, slide a board underneath or use tongs, and put them in a container. One pad is a large filet, or 2 servings. It can be cut up and frozen in cubes for drinks.

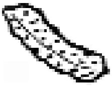

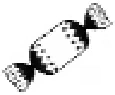

One hair rinse pad is good for several shampoos. Fruit and seed pulp also harvested with tongs. Flowers are easily removed by hand, petal at a time. Petals eaten raw, cooked to candy, jelly.

To prepare pads as poultice, hold gingerly with thumbs and forefinger. Using point of a knife, poke first, then slice carefully in half. Spread out and scrape off watery meat for use as a poultice.

Caution: Prickly pear is a protected plant. OK to harvest pads, flowers, buds as a survival food on the desert. Do not remove whole plants.

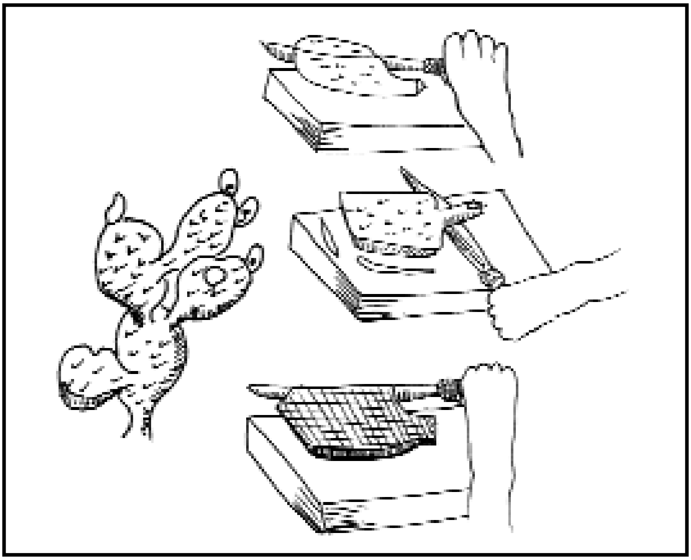

Preparing Prickly Pear.

Linda says- My personal attachment to this plant came in a way different from the usual. I led a wild food field trip to the desert in Black Canyon City. The temperature was only 90 degrees, but the sun was vicious to my fair, sensitive skin. I knew my sun block was not working so well. I scraped the needles off some prickly pear and cut the pad in half, applied the cactus meat to my sunburned face. The prickly pear and I became good friends from then on!

Purslane

Portulaca species
Purslane Family, *Portulacaceae*

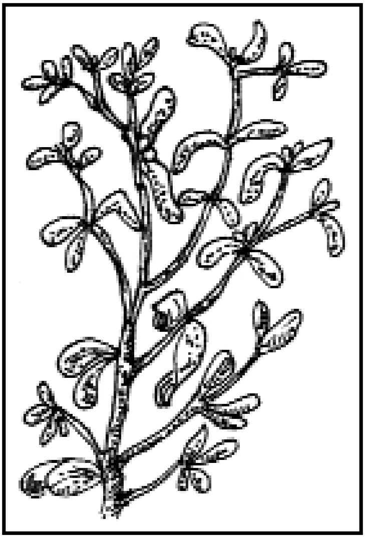

Other Names: Low pigweed, pusley.
History: Native to Europe, widely naturalized in United States.
Habitat: Fertile soil, topsoil.
Characteristics: Annual herb. Prostrate peddle-shaped succulent leaves on branching fleshy vines. Forms mats in masses of growth. Five to 7 petaled tiny yellow or red flowers are prolific. Multitudes of tiny black seeds produced on each plant.
Primary Uses: Culinary. All parts edible, raw or cooked. Purslane also pickled.
Nutritional Value: Shoots contain a balance of minerals for only l6 calories per half a cup.
Medical Value: A succulent that quenches thirst as well as puts a quantity of minerals in the body of an active person. Contains a high mineral content.
Collection and Storage: Gather mats, snapping off branches, and leave root for rejuvenation. Gather succulent young branches for raw treats. Also grows well indoors.

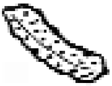

Linda says- My first harvest of purslane yielded 34 jars of pickles. Since then I have picked and pickled virtually hundreds of jars of purslane.

Queen Anne's Lace

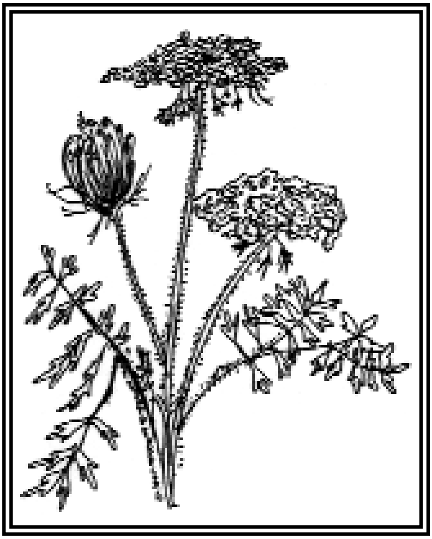

Daucus carota
Parsley Family,
Umbelliferae

Other Names: Wild Carrot.
History: Native to Europe, especially common in England; widespread in United States. Used by pioneers and Native Americans.
Habitat: Fields, waste areas.
Characteristics: Biennial herb. Reaches height of 2 feet or more. Carrot-like odor in stem and leaves as well as seeds. Flower clusters flat-topped, umbrella-like, lacy and often have single purple flower in the center. Old flowers "bird cage up" and turn brown. Leaves are parsley-like, feathery and smell like a carrot. All parts of stems have tiny fuzz, essential for proper identification.
Primary Uses: Culinary. Stems may be cut into sections and used for flavoring in stews & soups. Buds may be sautéed in oil.
Nutritional Value: Beta-carotene is exceptionally high, as is niacin.
Medicinal Value: The brown dried seeds are an excellent salt substitute.
Collection and Storage: Leaves are gathered in early spring and young leaves all season. Gather roots as carrots in early spring or fall; may be dug under leaves or hay even after winter snow. Gather flowers in summer. Collect seeds in fall.
Caution: Seeds of the wild carrot have a high content of vitamin A and shouldn't be eaten in excess.
Caution: Easily confused with Poison hemlock, resulting in severe burns of the fingers and hands as well as almost certain death if ingested. Check for the fuzz on the stem, as well as sandpaper-like feeling on dried stems from old dried fuzz. Use the foraging rule and crush, roll, and smell first; do not ingest. Determine the carrot smell and then look for the hair or fuzz.

Comments: In the fall, the flowers of wild carrot curl up; I call this “bird caging.” At this time, I gather the seeds and a few seeds go a long way! I get a winter’s supply of “salt” from a few flowers.

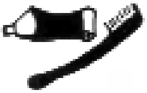

Queen Anne's Lace Chips

I will always remember my switch from potato chips to Queen Anne's lace flowers. In the wilderness, potato chips are a rare commodity, but Queen Anne's lace flowers crisp to a "chip" of carroty crispy flavor. I dip the flowers in hot oil and then put the brown crispy delights on hot rocks to cool. Todd loved to fly by and grab a few crispy flowers and pop them in his mouth using the stems as a handle.

Raspberry

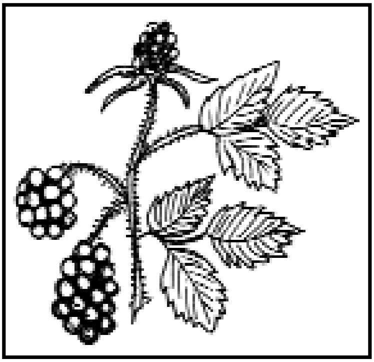

Species
Rose Family, *Rosaceae*

Types: Cloudberry, Thimbleberry, wild red.
History: Some varieties native to North America, others naturalized from Europe, Asia. Used as a fruit in all countries where raspberries are found, especially Europe.
Habitat: Roadsides, rocky fields, thickets.
Characteristics: Hardy shrub, biennial canes from perennial roots. Varies in size to average of 5 feet high. Saw-toothed leaves in groups of 3. Round, erect stem with short thorns. Stems have a white powder on them. Flower white; fruit pebbly and red or purple when ripe, pulls away from receptacle.
Primary Uses: Culinary, medicinal, cosmetics. Early leaves used fresh or dried for tea. Buds and flowers are edible. Frozen in ice for drinks. Fruit eaten raw, cooked for jellies, jams, syrups, fresh for juices, wines. Stem peeled and chewed, tea made from young twigs.

Nutritional Value: High in vitamin C.
Medicinal Value: Tea from leaves good for control of diarrhea. Relaxes muscles of the uterus, beneficial for menstrual cramps and strengthening the reproductive system.
Cosmetic Value: Stimulating astringent in herbal hair rinses and bath mixtures.
Collection and Storage: For quick, efficient berry harvesting, hang a container around your waist and pick using both hands. When drying berries, spread a fine cloth over trays of berries to keep flies and birds away. Harvest young leaves.

Caution: Reddish-tinged mature leaves may be toxic. Use only young green leaves.

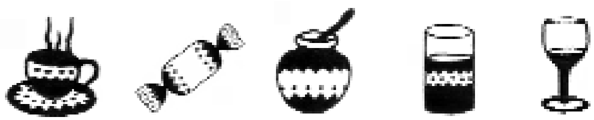

Linda says- The Adirondacks are inundated with black bears during berry season. One time I came upon a huge mama bear lying on her back under a thicket of raspberry. She was stupefied from gorging on the fermented, overripe berries. As she rolled over, she hiccuped and burped, then waddled off. See page vii.

Rose

Rosa rugosa, R. Caroline
Rose Family, *Rosaceae*

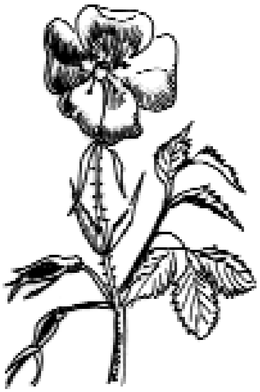

Other Names: Garden rose, wild rose.
History: R. Caroline native to North America; R. rugosa native to Asia but naturalized in United States. The symbol of love and beauty in ancient mythology. In Middle Ages, rose was symbol of secrecy (Subrosa). Also a heraldic symbol (War of Roses). Native Americans used all parts of plant. Europe and Asia (Sturtevant).
Habitat: Meadows, fields, woods, and coastal beaches.
Characteristics: Perennial or hardy shrub. Usually 5-petal flower,

yellow, pink, red or purple. Thorny stems, particularly central stem. Fruit is called rose hip and has prominent calyx lobes.

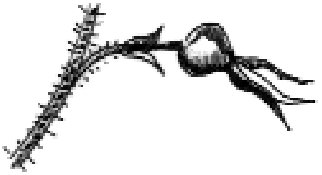

Rose Hip.

Primary Uses: Culinary, medicinal, cosmetic. Petals eaten raw or cooked, in salads, as a garnish. Petals are used for teas for rose ice cubes or popsicles. Rosehips remain at end of stem after the petals fall off. This bulbous end is edible raw or cooked, sliced in salads. Leaves used raw in salads; dried for leaf tea. Stems peeled and brewed as tea; roots peeled and chewed; brewed as tea. Hips used in jam, jellies, tea, candy, or confections. Brewed as tea; dry and reuse, or store in glass jars. Also used in incense & lamp oil.

Nutritional Value: Rose hips are high in vitamin C.

Medicinal Value: Known to help with colds and coughs. All parts, including roots, pounded and powdered as snuff to clear sinuses. In the past, the bark was mashed and applied to boils, wounds, sores. Today, rose oil is used to soothe headaches and as a healing ointment.

Cosmetic Value: Soaps & perfumes.

Collection and Storage: Gather all parts. Early rose shoots are easily clipped with scissors. Save all parts. Leaves may be clipped at any time and wear gloves and use a canvas bag or paper bag, one for leaves, one for roses and hips. Full roses snip 1 inch below hip. Pull petals off and use hips later.

Sauté rose petals in light oil for delicious treat. Roll in confectioner's sugar and refrigerate for candy.

Make rose oil by adding 1 ounce of rose powder (ground petals) to 4 tablespoons olive oil and store for a few weeks.

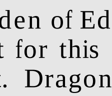
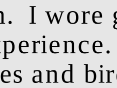
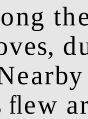
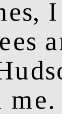

Linda says- Imagine what it's like to find yourself in a field of wild roses. Wandering among the bushes, I felt I was in the Garden of Eden. I wore gloves, dungarees and a long sleeve shirt for this experience. Nearby the Hudson River flowed past. Dragonflies and birds flew around me. When I tired of stuffing bags with petals for drying, I took a dip in the clean upper Hudson.

Saguaro

Carnegiea gigantea
Cactus Family, *Cactaceae*

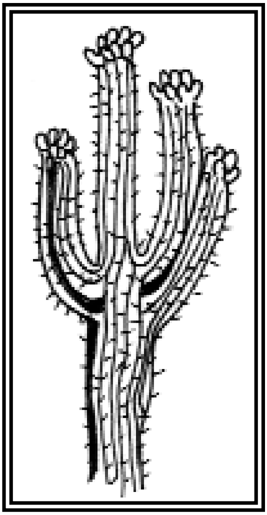

Other Names: Monument of the desert.
History: Native to North American desert. Used by desert dwellers, Native Americans. Papago Indians make heavy syrup and intoxicating wines.
Habitat: Rocky, gravel soil; hills, canyons, and dry washes.
Characteristics: Treelike cactus. Has one or more founded arms extending branch-like from a single, thick trunk. Shallow root system with small roots radiating out the height of the cactus and then some; no taproot. Long thick trunk can be 2 1/2 feet wide and 50 feet high. Large, thick grooves running laterally on all parts; glochids and thick fishhook barbs on all parts of the cactus. Circle of white flowers on top branch. Fruit is egg shaped, 2 to 3 inches long.
Primary Uses: Culinary. Flower buds and seeds are eaten. Flowers are seen first, then buds with seeds inside. Fruit is eaten raw; husks and seeds are boiled, baked; fruits, husks and seeds used for syrup or jam. Dehydrated pulp used for flour, also oil, soft drinks, wine, and vinegar.
Nutritional Value: Buds, or fruits are high in niacin, fiber, carbohydrates as well as ash.
Collection and Storage: Harvest fruit in July and monthly thereafter by using a long stick to knock off bulbs. The skin splits and bark curls back when fruit is mature. May be dried for storage or be cut in half and soaked in water for about 1 hour after removing seeds.
Caution: Saguaro is a protected cactus tree. The Department of Agriculture must give permission to remove cactus from the desert. Fruits, buds may be used for food in a survival way, but permission must be noted in larger quantities. No permission will be granted to remove a rare crested specie of saguaro.

Linda says- The saguaro is a monument to desert life, protecting myriads of insects, mice, fruit-eating bats, owls, and other creatures of the desert. During a field walk with Willie Whitefeather, my Cherokee friend, I saw the buds of the giant saguaro harvested correctly. Willie fashioned the dried ribs of an old saguaro hulk together, lashed a hook-like end on the tip. The tool was 16 feet long and bent like a giant fishing rod in the desert wind. Deftly, he pulled the saguaro buds to the waiting crowd below.

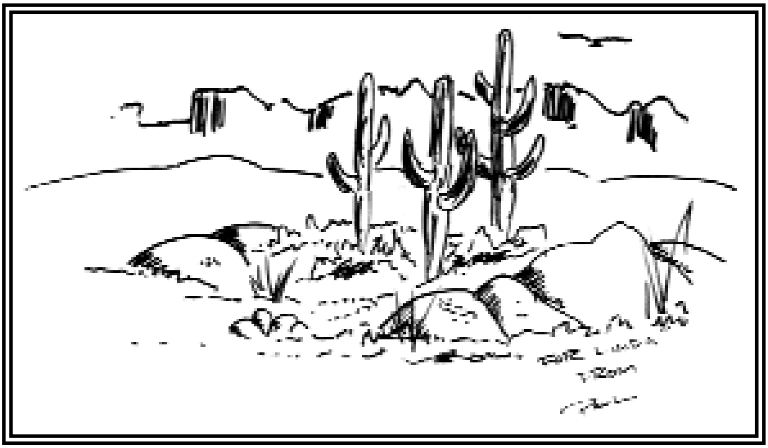

"Desert Saguaros" by Paul Runyon, Linda's brother.

Sheep Sorrel

Rumex acetosella
Buckwheat Family *Polygonaceae*

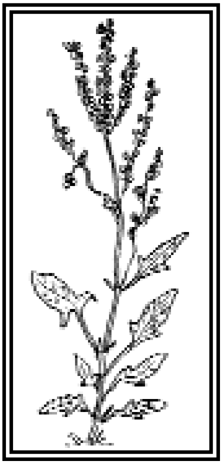

Other Names: Garden sorrel, field sorrel, R. acetoea.
History: Native to Europe and Asia, naturalized throughout North America. Garden sorrel is cultivated for use as cooked green.
Habitat: Waste areas, roadsides.
Characteristics: Perennial herb. Reddish-tinged grassy tops. Mature seeds grow in spreading areas, standing out easily. The arrow shaped leaves are sour in taste.

Flowers are tiny and give way to seed quickly.
Primary Uses: Culinary. Leaves or stalks, eaten raw in salads or cooked as vegetable. Can be used in cold drinks.
Nutritional Value: High in vitamin C.
Medicinal Value: Juice has mild antiseptic effect (crush and place on wound as a poultice). Juice acts as laxative.
Collection and Storage: Use scissors and shear clumps of seeds easily, collecting in a bag or basket. Strip leaves off separately for soups, drinks. Seeds dried and stored in glass jar for use as spicy food additive.
Caution: All parts contain oxalic acid crystals. These may inhibit the absorption of calcium in the body. Place in boiling water for a minute or so to destroy the crystals.

Linda says- I was amazed to taste the arrow head shaped leaves and find they were the same bitey, sour lemony taste as wood sorrel. (See page 137.)

Shepherd's Purse

Capsella bursa-pastoris
Mustard Family, *Brassicacae*

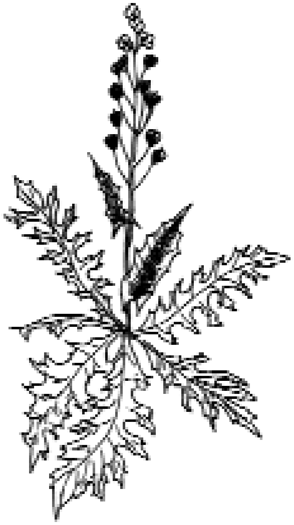

Other Names: Heart seed, field cabbage, purse seed, pick purse.
History: Native to Europe, widely naturalized in North America. Used by Native Americans.
Habitat: Waste areas, roadsides.
Characteristics: Annual herb. Erect plant from 3 inches to 1 1/2 feet high. Slender stem comes up from a spreading rosette of leaves formed flat on the ground. Deeply lobed leaves resemble wild lettuce or dandelion, but are more symmetrical. Has small white flower

with petals only 1/12 to 1/8 inch long, jutting out from the branches on a single stalk. Flowers give rise to seeds that resemble upside-down hearts.

Primary Uses: Culinary, medicinal. Basal leaves eaten in sandwiches, salads, soups, vegetable. Stalk is stir-fried, eaten raw. Buds and flowers eaten raw. Seeds are stir-fried and used as pepper, spice, seasoning. Roots are used fresh or dried as substitute for ginger or candied in syrup.

Nutritional Value: High in vitamin K, vegetable protein, potassium, calcium as well as beta-carotene and minerals.

Medicinal Value: Dried to use as blood-clotter and valuable taken in quantity for internal bleeding.[1]

Collection and Storage: Look for upside-down heart shaped seeds to distinguish plant from other wild mustards and pepper grasses. Gather leaves at any stage. Gather seeds when ripe. Entire plant may be gathered and chopped for vegetable, as well as dried for additive. Add seed pods to soups and stews for a pepper-like flavor.

[1] (Harrington 1967)

Linda says- In the East I ate the little heart-shaped seedpods with gusto, but in the West I seek them with a vengeance. Shepherd's purse is known to help relieve strong menses and internal bleeding. According to Peter Bigfoot in his book *Arizona Wild Herbs,* pick the mature, flowering plants in the Arizona Desert foothills in springtime. Dosage: 1 teaspoon dried herb per cup of hot water. Steep 15 minutes and drink the tea cool. Take 2 or more cups full per day as needed. Drink it quickly, do not sip for best effect with internal bleeding.

Sow Thistle

Sonchus oleraceus
Composite Family, *Compositae*

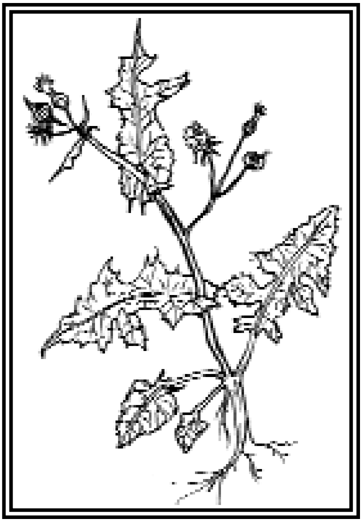

Other Names: Hoitgamivakhi (Pima Indian).
History: Native to North Africa, Europe, and western Asia; introduced to the United States and now found throughout most of Arizona. Used by Native Americans as greens.
Habitat: Roadsides, damp fields, waste areas.
Characteristics: Annual herb. Dandelion-type deeply lobed leaves begin from a central stem on the ground, clasping around stalk to a sharp point, like a sow's ear. Top of stalk bears branches, at the end of each grows a yellow flower that looks somewhat like a dandelion flower. After seeding, a white fuzz appears with single seeds on the end of the fluff.
Primary Uses: Culinary. Buds, flowers, and seeds edible raw or cooked. Stems may be cut into inch long pieces with scissors and stir fried or sautéed in oil as beans. Roots may be scrubbed and stewed in crock pot. Young leaves gathered and eaten as spring green, also eaten raw. Flowers stir-fried as vegetable or dried as well as frozen.
Nutritional Value: Very high in minerals.
Medical Uses: Milk reported to be somewhat diuretic.
Collection and Storage: One may trim leaves off long stems easily by leaving plant in ground and trimming off with a long knife. Use long handled knife or scissors, and wear gloves. Cut leaves over a basket or bag. To harvest large amounts, cut 2 long goldenrod or wild lettuce stalks and lay parallel to each other. Using a gloved hand, place the thistle plants across the parallel stalks. When a bundle is accumulated, pull golden rod stalks up and wrap around. Carry by holding both ends of the goldenrod together.

Linda says- My introduction to sow thistle was an experience. Early in the morning, when the dew was still on the grass, I wandered through a field of barley grass. The sow thistle was a succulent emerald green amid the grasses. Since I carry water with me; I rinsed off the leaves and munched on the sweetest breakfast I ever had in the West.

Strawberry

Fragaria virginiana
Rose Family, *Rosaceae*

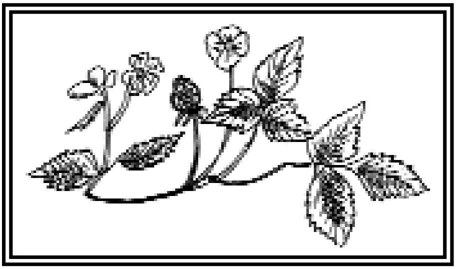

Other Names: Heartberry, wood strawberry.
History: Native to North America; Used in early European cultures, especially popular in France. Grows throughout Northern Hemisphere, excluding tropics. Native Americans steeped wild strawberries in warm water, then strained and cooled the liquid. The resulting lotion was a wash for irritations and skin ulcers.
Habitat: Fields, roadsides, meadows.
Characteristics: Low perennial, reaching height of 2 inches to 8 inches. Lobed saw-toothed basal leaves in groups of 3. Tiny, heart-shaped strawberry fruits. Small white flowers.
Primary Uses: Culinary, medicinal, cosmetic. Flowers are edible raw. Leaves eaten raw, dried for tea. Fruit eaten raw; cooked for jellies, jams, syrups; raw for juices, wines. Entire plant is steamed, boiled, used in soups and stews; cooking liquid is drunk.
Nutritional Value: Leaves and fruit high in vitamin C.
Medicinal Value: Leaves infused and used for sore throats. Juice is mixed with water for eye wash. Historically, roots are infused for gonorrhea remedy. Fruits are dried for stomach problems, diarrhea, dysentery, liver pain, jaundice. To juice, put a nylon stocking over a bowl. Place washed berries in stocking, then squeeze the juice into the bowl.
Cosmetic Value: Astringent and facial cleanser, useful for oily skin. Fruit is tartar remover for teeth; crush fruit, rub on, and rinse.

Collection and Storage: Pick fruits as you would cultivated strawberries With container around your waist, pick fruit and leaves easily using both hands.
Caution: Dry leaves thoroughly for tea. Partially dried leaves can cause a toxic reaction.

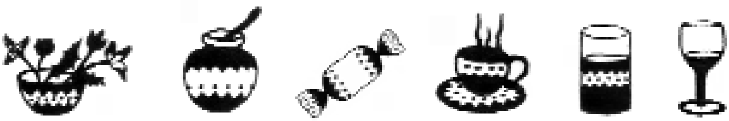

Linda says- Picking wild strawberries in the Adirondacks is difficult because the berries are very small and they ripen in blackfly season. The hardiest folk pick them in the wee morning hours or at night or in a rain. Snapping turtles eat the berries, as do other critters. I enjoy finding berries with chunks missing, and speculate what had breakfasted before me—insect, turtle, or mouse.

Sumac

Rhus species
Cashew Family, *Anacardiaceae*

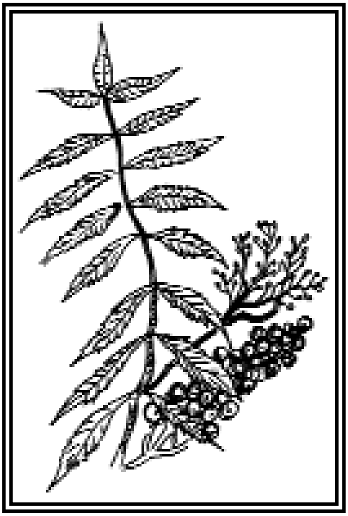

Other Names: Staghorn, smooth, scarlet, dwarf, shining, mountain, hairy, velvet, fragrant, virginian, and winged sumacs; lemonade tree.
History: Mostly native to North America. Used by Native Americans and colonists.
Habitat: Roadsides, field edges.
Characteristics: Deciduous tree or shrub. Bushes or shrubs with some varieties to height of 20 feet. Fern-like leaves 14 to 24 inches long. Flaming red fuzzy berry clusters, hairy and lemonade tasting. Stems and branches exude white milk when broken. Plants need heavy sunshine to produce berry clusters.

Primary Uses: Culinary. Flowers and seeds are the fruits. Fruits eaten raw or crushed in water as lemonade. Frozen in ice cube trays for popsicles.
Nutritional Value: High in vitamin C.
Collection and Storage: Collect fruit before completely mature, when spires turn dark red. Collect spires by breaking off stems. For tea, strain out hairs by overlapping balsam branches as a sieve or use cheesecloth. Look carefully at the whole tree to avoid accidental contact. I once pulled down a poison sumac branch growing in the midst of an edible sumac tree.

Caution: Poison sumac has white or greenish-brown clusters instead of red spires. When dried, a tan spire is evident. See Part V.

> ***Linda says-*** There is nothing more beautiful than a glass decanter filled with sumac berries. Scarlet tea! Sumac lemonade is a year-round treat. Collect and rinse the heads early in the season. Do not wait until fall, when insects have made their nests in them! Summer seed heads may be dried and stored in glass, or bagged and frozen.

Sunflower

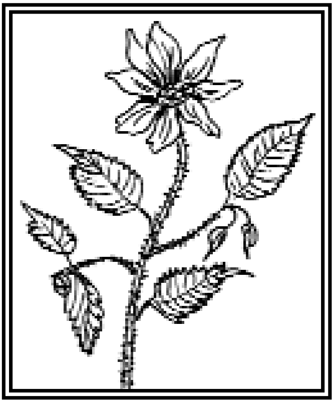

Helianthus species
Composite Family, *Compositae*

Types: Annual, wild, Western, common, tall, giant, weak, saw-toothed, swamp, little, aspen, showy, and woodland sunflower; Marigold of Peru.
History: Native to the Americas, probably originated in Peru. Used by Native Americans.
Habitat: Widespread but more common in West. Gardens, fields, sunny open areas.

Characteristic: Annual herb. Composite flowers on plants from 3 to 14 feet high. Yellow ray flowers with purplish red to brown disk flowers in center. Stems may be smooth or hairy. Leaves opposite or alternate, characteristically narrow, long, rough, and saw-toothed.
Primary Uses: Culinary, cosmetic. Seeds crushed and boiled; oil skimmed off top and eaten. Pith edible raw and cooked in soups. Used for paper making, organic fertilizer, oil, lubricants, candles. Petals used for yellow dye.
Nutritional Value: Seeds high in protein.
Cosmetic Value: Oil, soaps.
Collection and Storage: After flower matures and petals fall off, gather seeds. Snap off dried center and dry seeds on screen, then bottle. Large heads dried in a closed paper bag will mature eventually. Scrape the heads across a sieve with holes sized such that the seeds can drop through. Place seeds on cookie sheet and roast in dry, low oven. Bottle in glass for long-term storage. Pith is collected by peeling outer fresh stem when plant is dry and pulling out white pith.

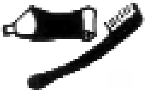

Linda says- Snapping off the tiny dried heads of desert sunflowers is fun and very profitable for my table. A few crushed heads and a bit of water yielded a nutritious gruel with an obvious energy boost. Storage is simple: I put the dried sunflower heads in a glass jar.

Thistle

Circium species
Composite Family, *Compositae*

Other Names: Bull thistle, yellow thistle, plumed thistle, swamp thistle, field thistle.
History: Native to North America. Used since colonial times and by Native Americans.

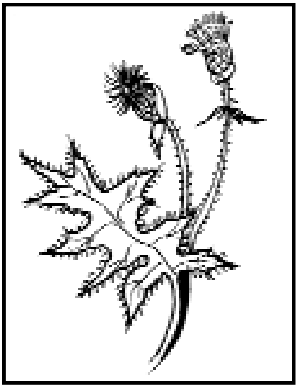

Habitat: Roadside, fields, waste areas.
Characteristics: Biennial plant with dozens of varieties. Reaches height of 2 to 6 feet. Leaves and stems have sharp spines (bulbs) on all parts. One to 3 large, purplish flower heads with spines around base produced second year.
Primary Uses: Culinary. Flowers steamed and eaten raw much like an artichoke. Leaves dried and ground to a fine, green flour. Dried leaves also used for tea. Steamed leaves eaten as a vegetable. Stems peeled, sliced, and cooked as a vegetable. Roots boiled for tea.
Nutritional Value: High in potassium, phosphorus, and vegetable protein.
Medicinal Value: Roots are used for restorative tonic. Roots can be harvested year round, but early summer is best.
Collection and Storage: Early in the spring, use a penknife to cut beneath the basal leaves as you would a head of lettuce, being careful not to disturb the root. Cut leaves off central stem, holding a bag underneath. Use tongs and strainer to wash. Dries and freezes well.

Linda says- The beginnings of a thistle plant are succulent. Its thorny whorls are clear blue-green. The trick is to get as much of the core as possible, leaving the basal root. The thistle will grow again if you carefully take the upper part of the root along with the leaves, leaving the deeper root alone. The first time I did this I wondered if it was worth the effort. Now I use tongs and a strainer, and find the preparation time negligible while the taste is lasting!

Thyme

Thymus vulgaris
Mint Family, *Labiatae*

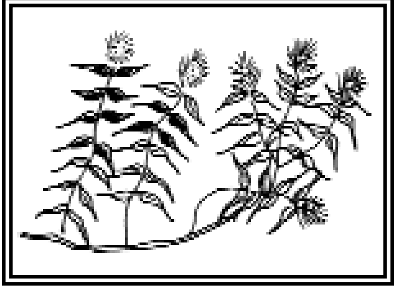

Other Names: Bee plant, wild thyme, mountain thyme, creeping thyme.
History: Native to Europe; in North America gardens vari-

ety has expanded to woods and fields. Adapted and used by Native Americans.

Habitat: Fields, roadsides, roadbanks, lawns.

Characteristics: Woody perennial. Prostrate herb growing to height of about 4 inches. Several varieties, all with pepper-like smell. Many spikes of leaves and lavender spike flowers. Grows in thick patches.

Primary Uses: Culinary, medicinal, cosmetic. All parts edible. Ground to powder; twigs, leaves, and flowers simmered for stimulant tea. Added as seasoning to soups, stews, fish, meat dishes.

Nutritional Value: Has a high count of vitamin C, iron and niacin.

Medicinal Value: For headaches, stimulant, antiseptic, aromatic, carminative, diuretic, diaphoretic, emenogogic, and antispasmodic. Treatment for sore throats.

Cosmetic Value: Oil is aromatic, antiseptic, diaphoretic, stimulant, disinfectant, and used in cosmetic lotions. An effective deodorant, mouthwash, aftershave lotion, and wash.

Collection and Storage: Easily sheared with scissors. Mow thyme patches regularly for a winter supply. For best results, rub dried thyme through a screen to remove stems and twigs.

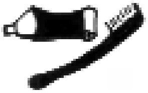

It's Thyme Time

Lying next to a patch of lavender thyme flowers and nibbling the pungent, peppery twigs and flowers soon became a favorite activity after a busy day. My favorite patch was near Cedar River Road in the Adirondacks. Those days were spent building structures, gathering large amounts of food, experimenting with storage and cooking methods, and just plain surviving. I would lay full length next to the bees working the thyme. The bees never minded, lifting off and moving as I raised my hand to pluck a stem to munch on. The thyme was a perfect stimulant. I returned to finish the evening renewed.

Tumbleweed

Salsola kali
Goosefoot Family,
Chenopodiaceae

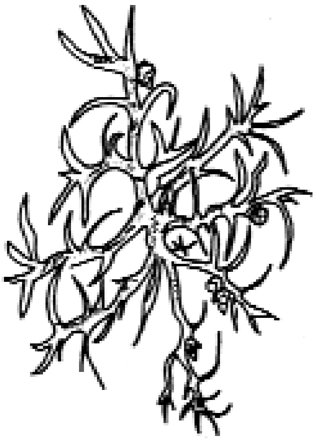

Other Names: Russian thistle.
History: Native to Russia, brought to United States in flaxseed 100 years ago.
Habitat: Open desert, fields, roadsides, hot dry country.
Characteristics: Annual herb. Reaches height of 2 feet in bushy and roundish intricately branched plant. Often reddish with ridged stems and many tiny ends. Late in season, plant breaks off at base and rolls around freely, dropping its many seeds.
Primary Uses: Culinary. Tips eaten cooked as vegetable or in soups and stir-fries.
Nutritional Value: High in calcium, vegetable protein, carbohydrates, fiber, potassium and *extremely* high in niacin.
Collection and Storage: Choose tumbleweed that does not have many buds and flowers, and harvest young seedlings whenever possible. If you don't have time to stand and clip the bush, remove a few branches and let wilt naturally. Clip ends that wilt the most; this ensures getting the most succulent ends. Drop succulent ends in a bowl, wash well, or soak overnight in a large bowl of water and a teaspoon of vinegar. The more you clip the more the plant grows to maturation, producing many more branches.
Caution: Do not ingest raw except in extreme emergency, chewing thoroughly.

Linda says- Making tumbleweed soup is simple. I find a tumbleweed without prickly seeds and flowers, and shear the succulent ends into a basket. Wash the succulent tips well, quickly cook them in boiling water, then pack half-cup quantities into soup bags and freeze. You'll have a month's worth of delicious soup.

Violets

Viola species
Violet Family, *Violaceae*

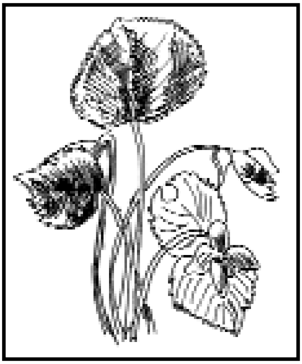

Other Names: Heart leaf, baby face, wild violet.
History: World wide distribution, including North America. Used by Native Americans.
Habitat: Rich, moist woods, swamp sides, stream banks, damp lawns.
Characteristics: Annual or perennial herb. Smooth, heart-shaped leaves with little scalloping on edges. Single stem. Flowers on single stem from leaf base; five-petaled, arising from a center spur, usually yellow. Lower petal has heavy veins and there is a bearded design. Colors vary from white or yellow to purple or blue.
Primary Uses: Culinary, medicinal, cosmetic. Flowers used as little "fruits", eaten raw and candied; also used in jam, jelly, wine. Buds, flowers, and leaves dried for tea. Leaves eaten raw in sandwiches or salads; cooked, as thickening agent. Roots of violets may be eaten raw or cooked in a crockpot. Stems may be snipped into 1 inch sections and added as "beans" to soups.
Nutritional Value: High in vitamins A and C.
Medicinal Value: Good for colds, flu. Eaten for dry, scratchy throat. Raw flowers eaten in quantity for varicose veins.
Cosmetic Value: An astringent; used in cold creams, oil lotions (infusion of flowers).
Collection and Storage: Clip violets easily with scissors, but use fingers when gathering leaves. Choose large leaves to wrap up tofu, rice, potatoes, beans.
Caution: Yellow wood violets may be cathartic; small, common lawn violets are not.

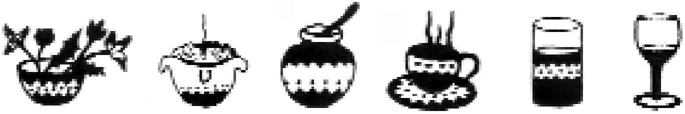

Linda says- Tiny faces peeping out of the dew-covered grass - violets are a favorite of mine, whether next to a stream in early spring or on the back lawn. The violet's face is a spring flag, heralding renewed life.

Todd loved the little faces in his ice cream. I would lightly stir a handful of violets into his vanilla ice cream.

Wild Lettuce

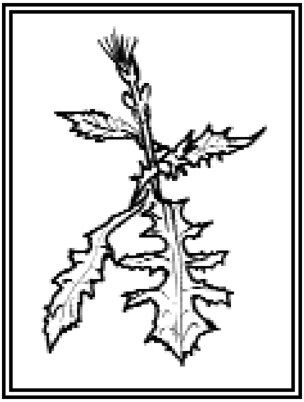

Lactuca canadensis
Composite Family, *Compositae*

Other Names: Milk lettuce, chicory lettuce, horseweed.

History: Native to North America. Used by Native Americans; The Menominees use wild lettuce as a sedative.[2]

Habitat: Fields, waste ground, roadsides.

Characteristics: Biennial herb. Deeply serrated leaves, coming off a central stalk. Tiny buds and extensions off of top stalk. Dozens of yellow dandelion-like ray flowers. White milk in stem. Stem veins are characteristic.

Primary Uses: Leaves eaten in salads. Flowers and buds are stir-fried. Flowers and seeds are eaten raw or cooked, sautéed in oil and garlic. Stems (young) cut to bean size and cooked. Leaves dried and ground for seasoning. Roots have been used as a gum.

Nutritional Value: High in beta-carotene, ascorbic acid, (vitamin C)

Medicinal Value: Tea from milky leaves is sedative. Sap induces sweating and is used as a fever reducer. Decoctions of sap are known to ease irritability and insomnia.

Collection and Storage: Collect leaves, stems, flowers as you would lettuce, use raw.

Linda says- The tops or buds of this great plant are delicious. Stir-fried with a little olive oil and garlic salt, wild lettuce makes a wild meal!

---

[2] Densmore, Francis. *How the Indians Use Wild Plants for Food.* New York: Dover, 1974.

Willow

Salix species
Willow Family, *Salicaceae*

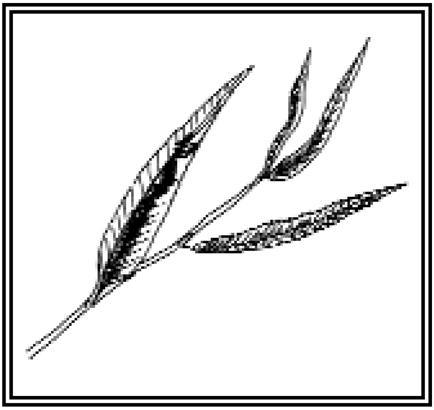

Types: Beaked, black, blue, Canada, coastal plain, crack, Drummond's, ward, weeping willow.
History: Grows throughout the world but abundant in cooler parts of Northern Hemisphere. In North America, used by Eskimos, Native Americans. Eskimos dry inner bark, strip, and cook like spaghetti; Inner bark layer called keeleeyuk in Eskimo, meaning "the scrape."
Habitat: Damp areas, stream sides, swamps.
Characteristics: Deciduous shrubby tree. Over 100 species, many with drooping branches alternate 2 inch leaves, sectioned twig tendril. Flowers (catkins) are furry, present before leaves. Roots seek water; have thousands of white hair-like extensions.
Primary Uses: Culinary, medicinal, cosmetic. Inner bark eaten raw; may be dried and pounded to flour. Buds, flowers or catkins, seeds are brewed as tea. Leaves are emergency food, tea, but use sparingly as they contain "aspirin". Willow sprouts provide early spring food.
Nutritional Value: Very high in beta-carotene.
Medicinal Value: All parts have acetylsalicylic acid, or salix, an aspirin source. Ten to 15 catkins or one 6 to 8 inch tendril, equal approximately 2 aspirins.
Cosmetic Value: Astringent; used in lotions, creams, facial washes.
Collection and Storage: Break tendrils into 6-inch pieces and dry. Store in glass container. Collect catkins or leaves by running hand down tendril. Dry and store. A supply of aspirin was easily kept through the year by stripping catkins in the spring, drying, and placing in glass.

Beneath the Weeping Willow

I sit under my favorite weeping willow tree. My fingers fly as I make a supple willow basket. At the same time, the tiny catkins hanging from the twigs are irresistible. I first munch only four or five, remembering that these catkins are like aspirin. But the catkins are so delicious. I quickly down about 1/2 cup of them while I make the basket. When I stand up, my ears ring and I wobble, almost faint. I sit down immediately and put my head between my knees to stop the dizziness. I curl up and go to sleep. Much later, I awake with a gigantic headache that lasts almost two days. A hard-learned lesson about the potency of willow catkins!

Wintergreen

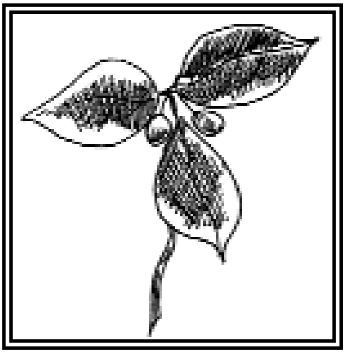

Gaultheria procumbens
Heath Family, *Ericaceae*

Other Names: Checkerberry.
History: Native to North America, especially East.
Habitat: Wooded areas, acid soils, under softwood trees.
Characteristics: Perennial herb. Drooping waxy white flowers hidden by flat shiny evergreen leaves. Has bright red berries.
Primary Uses: Culinary, medicinal, cosmetic. Leaves used fresh or dried, crush or steep for tea. Oils are distilled for flavorings. Buds, flowers, and seeds (berries) edible.
Nutritional Value: Fruit or berry, high in acetylsalicylic acid, or aspirin. Contains niacin.
Medicinal Value: Leaves contain acetylsalicylic acid (aspirin). Steeped for astringent. Boiled and applied to wounds or used as eyewash. Tea breaks a fever. Used as an anti-rheumatism gargle, to treat wounds or hemorrhages, as a poultice for insect bites and

bruises, or as vaginal douche (see page 138). Oils used for rheumatism or arthritis, sciatica.

Cosmetic Value: Aromatic oil used as flavoring for candies, toothpaste.

Collection and Storage: Snip leaves; dig roots, and collect berries. Wintergreen is known to grow under the snow; dig and eat during winter. Dried, powdered wintergreen leaves have more flavor than fresh leaves. Plant transplants easily, loves acid soil.

Caution: Contains acetylsalicylic acid (aspirin). Overdose of oil may produce drowsiness, congestion, and delirium.

Linda Says- Driving from Plainfield, New Jersey, to Indian Lake in upstate New York was a 17-hour drive in 1940. My frustration as a child was greatly relieved by chewing a wintergreen leaf I picked in the woods. I soon learned to recognize the flat, shiny, three-leafed plant growing close to the ground. If I was lucky enough to find a red wintergreen berry, I savored that berry for an hour of the drive!

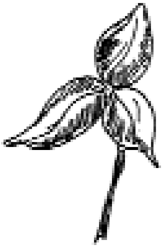

Winter Harvest

It was the first time I trooped through snowdrifts looking for food. I thought it would be hard work, but balsam and pine trees were always within reach of the trail. Bushes of meadowsweet poked above the snow. I would never starve if I ate twigs and bark. But I did not know then that the Iroquois Abanake (bark eaters) did just that for centuries. Following the deer trail, I located the wintergreen. Reaching deep into the snow opening, I pulled out a few plants. The shiny green aromatic leaf never tasted so good!

Wood Sorrel

Oxalis species
Wood-Sorrel Family,
Oxalidaceae

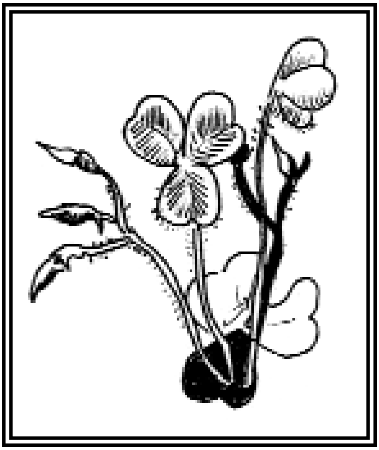

Other Names: Clover sorrel, shamrock sorrel, lemon sorrel.
History: Eaten as spicy salad for centuries. Introduced to North America and used by Europeans and all Native American cultures except Crow and Menomini. George Washington Carver, of peanut fame, made pies and confections from wood sorrel; he even made a paint from sorrel.
Habitat: Fields, woods, moist places, under trees, or shrubs.
Characteristics: Low-growing perennial herb. Leaves alternate or basal, in 3 heart-shaped leaflets (a shamrock), sour lemon tasting. Grows from 2 inches to 8 inches in bunches or carpets. Flowers long, with 5 yellow, rarely green petals, sometimes red at the base. Leaves fold up sometimes when extremely warm or in sunshine.
Primary Uses: Culinary, cosmetic. Leaves cooked in soups, stews. Seeds cooked in soups, stews. Buds and flowers cooked in soups, stews. Many commercial uses. Stems are edible. Cut into inch pieces and add to soups, stews.
Nutritional Value: High in vitamin C. Because wood sorrel is high in oxalic acid, it hastens the elimination of calcium from the body and inhibits absorption of calcium. Therefore, limit intake of raw leaves to no more than 15 at most. Cook for safety. Drop entire plant in hot water. When color changes from bright emerald green to olive green, oxalic acid is neutralized.
Cosmetic Value: Antiseptic wash for skin and skin eruptions.
Collection and Storage: Gather seeds when they are red. Gather leaves year-round Collect flowers any time of year. Cut stems, leaves, and seeds to dry for soup or use as a salad ingredient.

Yarrow

Achillea millefolium
Composite Family, *Compositae*

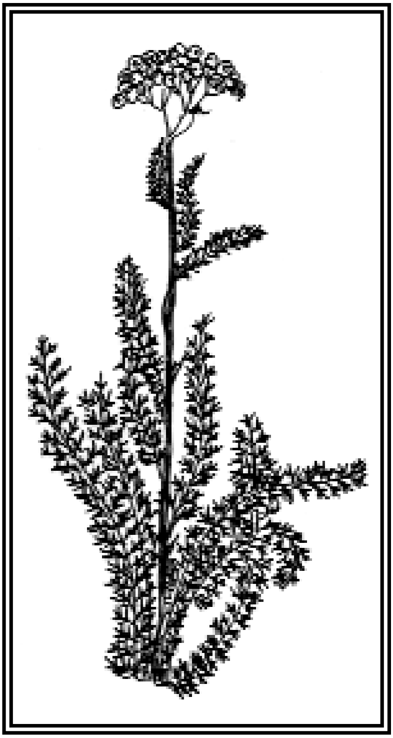

Other Names: Milfoil, field hop, old man's pepper.
History: Native to Europe. Naturalized in North America.
Habitat: Old fields, roadsides.
Characteristics: Perennial herb. Grown to height of 2 1/2 feet. Creeping rootstock; fern-like, lacy leaves, slightly hairy. Flowers typically white in wild, but occur in red, rosy, yellow, and magenta in ornamental gardens. Flowers have 5 petal-like rays, dozens in flat top clusters. Each flower is on its own stem, originating from one stem.
Primary Uses: Culinary, medicinal, cosmetic.
Stimulant tea. Boil any part of plant in small amounts. Leaves used raw in salads, or cooked for green. Seeds and flowers crumpled for seasoning. First-year roots peeled and cooked as vegetable.
Nutritional Value: High in calcium, and potassium.
Medicinal Value: Moves fluids in the body. Diaphoretic; cleanses the blood; tonic, stimulant, antiseptic; soothes burns. Strengthens immune system.
Cosmetic Value: Astringent; cleansing herb in facial mixtures, vaginal douches, shampoos; for dandruff and baldness treatments. Mouthwash for toothaches.
Collection and Storage: Clip flowers with scissors or break off with fingers. Pull leaves down stalk.

Yarrow for Teeth Cleansing

A dentist appointment loomed ahead. I hated to leave the forest and drive the old truck to Indian Lake, but having no phone, I could not cancel. I picked a yarrow stem and stripped off the leaves. I placed a leaf under my forefinger and scrubbed my teeth with the yarrow. Very astringent and satisfying. Using the stem, I picked here and there. Then I drove the pickup to town. In the dentist chair, the doctor exclaimed, "Linda, you've been in the yarrow again!" I did not realize that the yarrow had dyed the inside of my mouth green, especially my teeth—temporarily, of course.

The author begins her exploration of wild food.

Notes

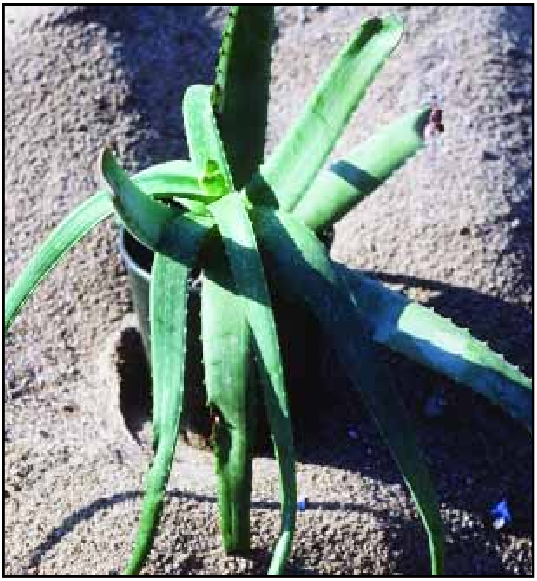

Aloe Vera

Aloe perfoliata vera *Cactus Family, Liliaceae*

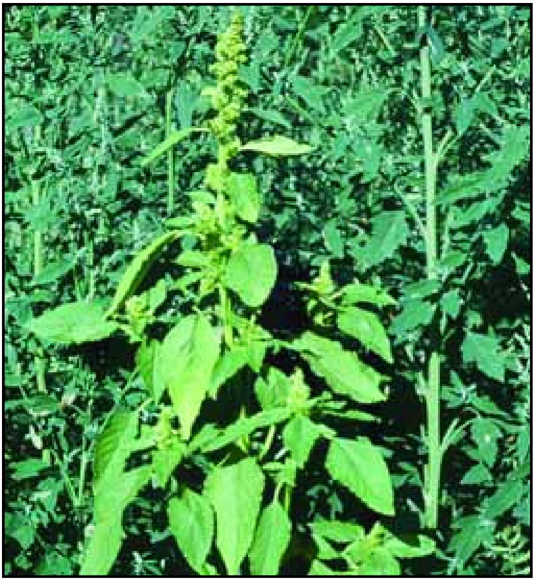

Amaranth

Amaranthus retroflexus *Amaranth Family, Amaranthaceae*

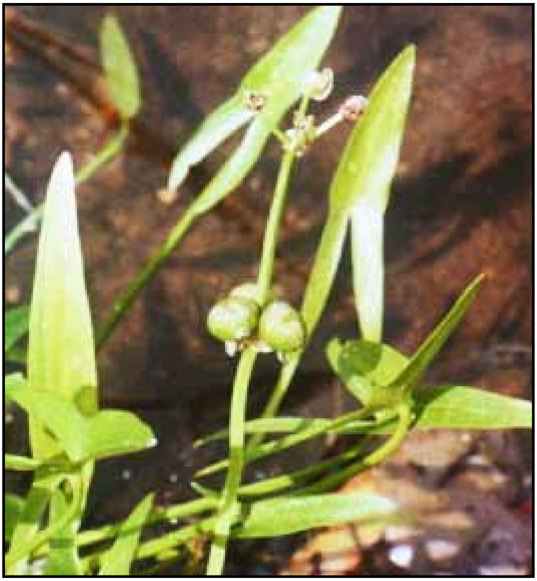

Arrowhead
Sagittaria species, S. chinensis *Water Plantain Family, Alismataceae*

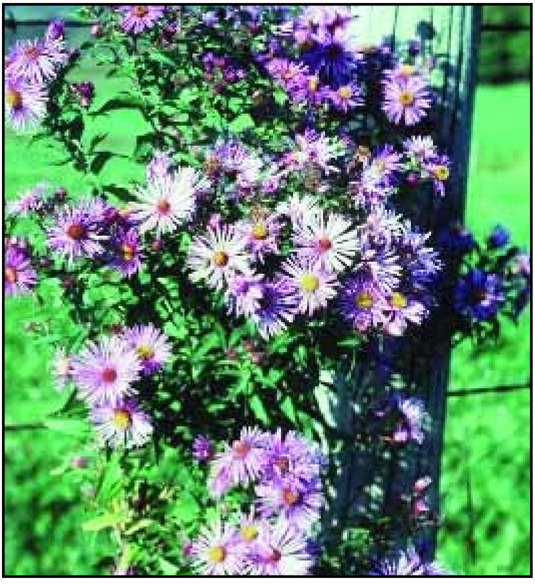

Aster
Aster nemoralis *Composite Family, Asteraceae*

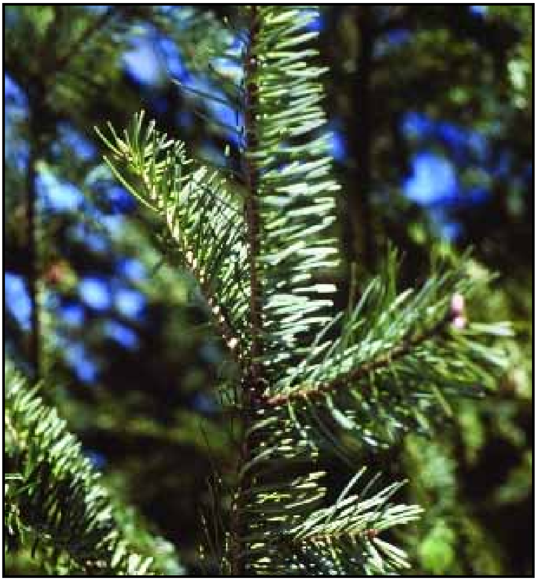

Balsam Fir
Abies balsamea *Pine Family, Pinaceae*

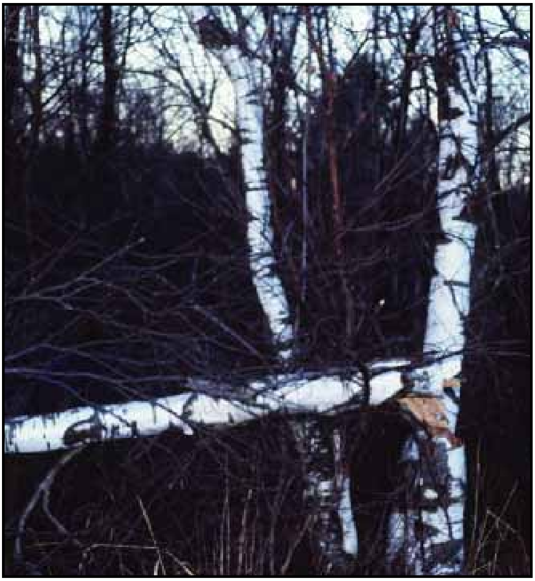

Birch
Betula species *Birch Family, Betulaceae*

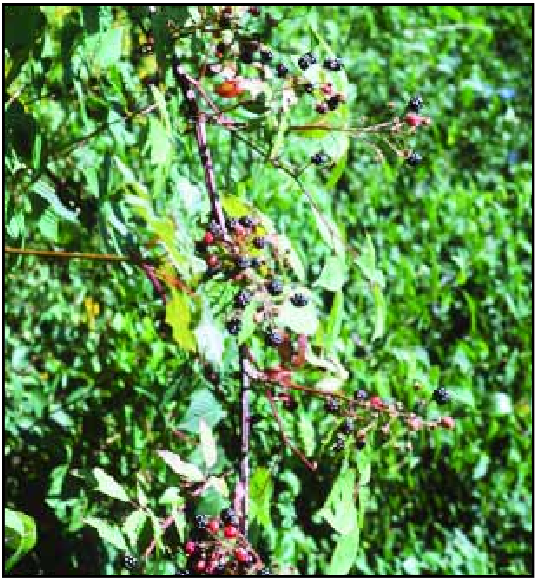

Blackberry
Rubus villosus and other bramble berry species *Rose Family, Rosaceae*

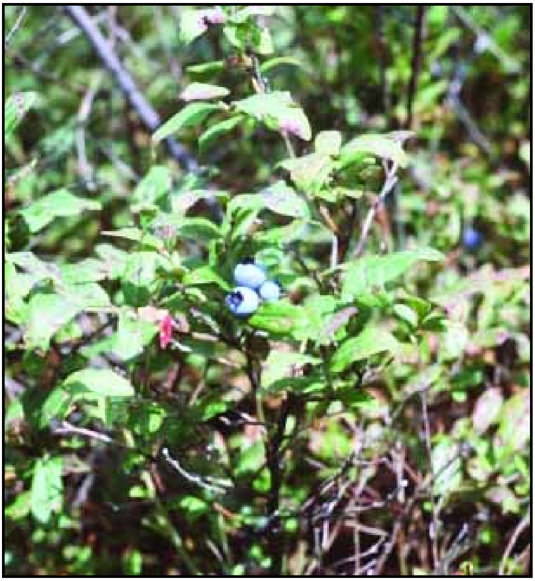

Blueberry
Vaccinium myrtillus *Heath Family, Erraceae*

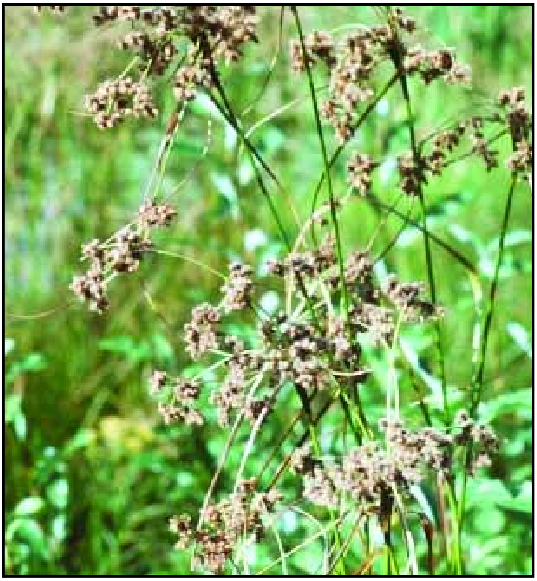

Bulrush
Scirpus validus *Sedge Family, Juncaceae*

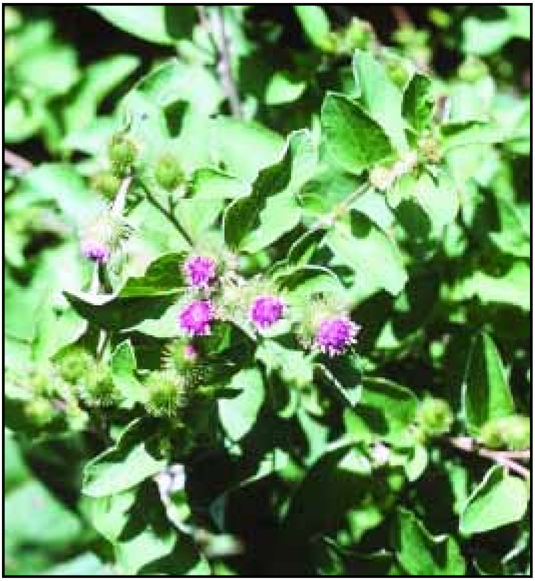

Burdock
Arctium lappa *Sunflower Family, Asteraceae*

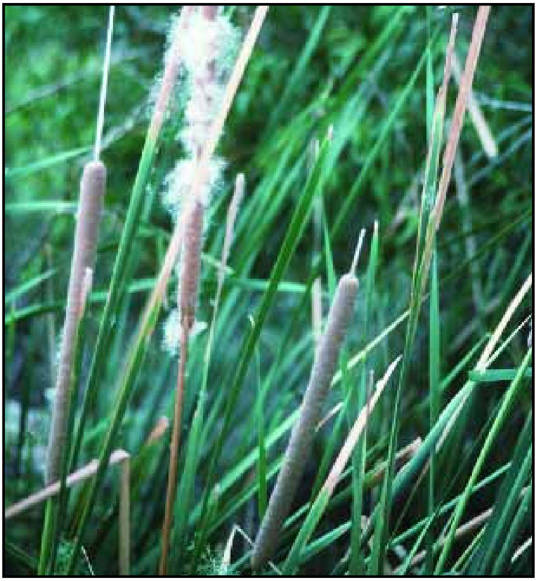

Cattail
Typha latifolia *Cattail Family, Typhacerae*

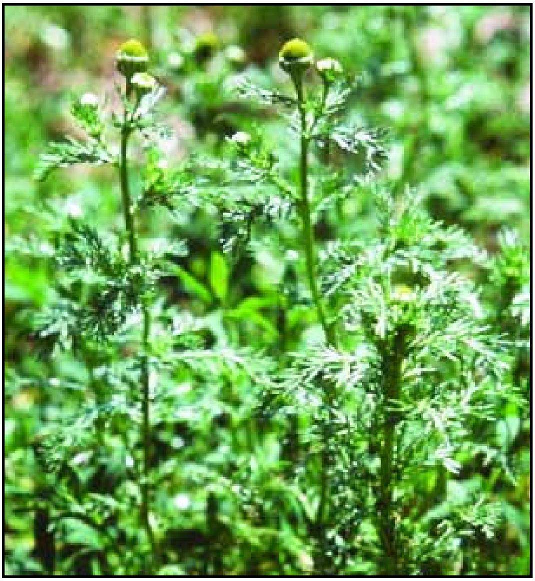

Chamomile
Matricaria chamomilla *Composite Family, Asteraceae*

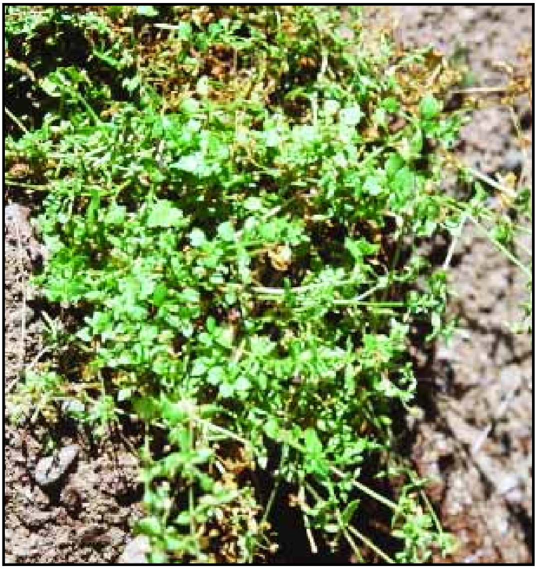

Chickweed
Stellaria media (common chickweed), *Pink Family, Caryphyllea*

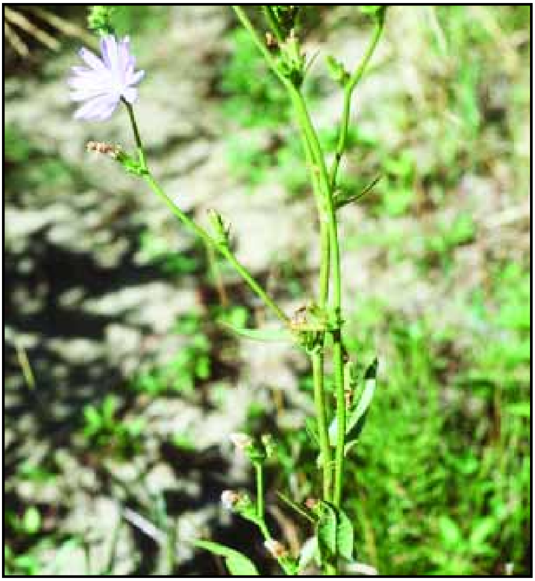

Chicory
Cichorium intybus *Composite Family, Asteraceae*

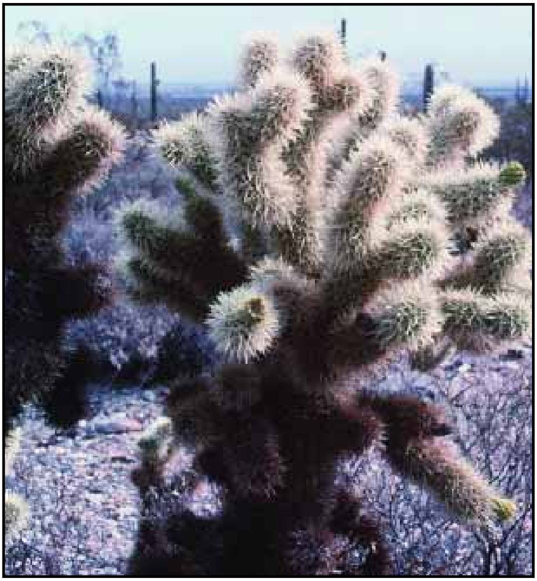

Cholla
Opuntia fulgida *Cactus Family, Cactaceae*

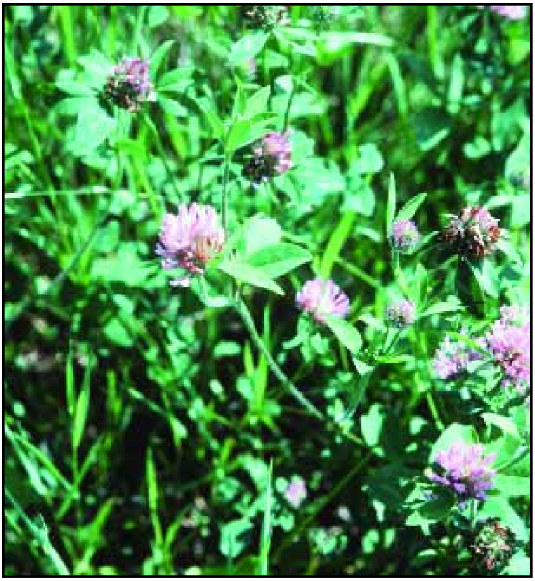

Clover, red
Trifolium pratense *Legume Family, Leguminosea*

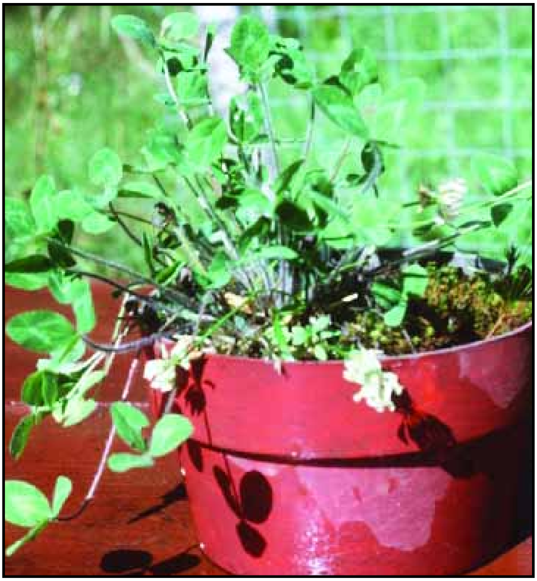

Clover, white
Trefolium Repens *Legume Family, Leguminosae*

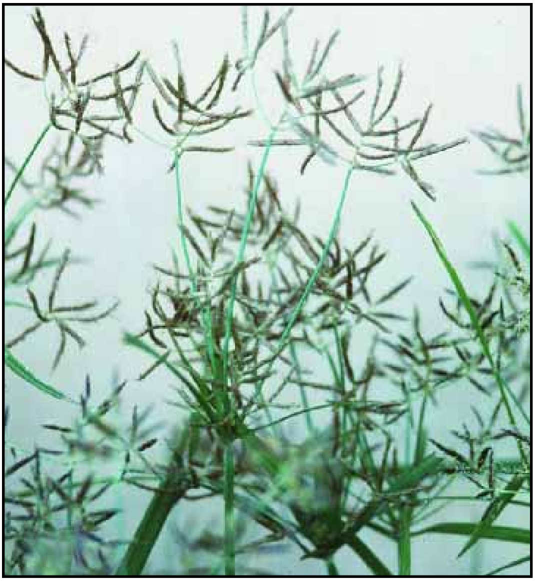

Crabgrass
Digitaria sanguinalis *Grass Family, Graminea*

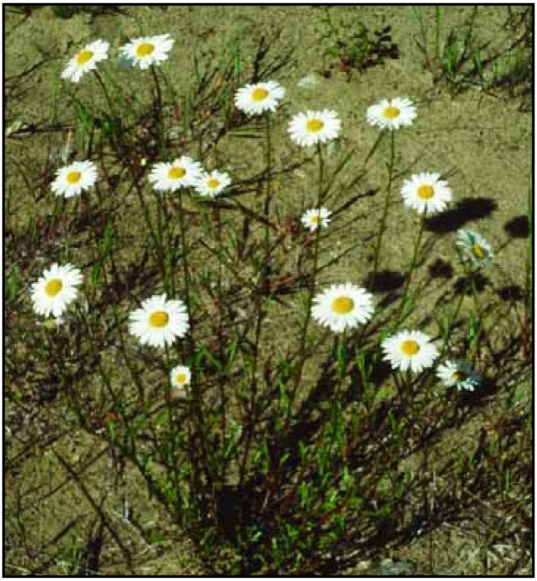

Daisy
Chrysanthemum leucanthemum *Composite Family, Asteraceae*

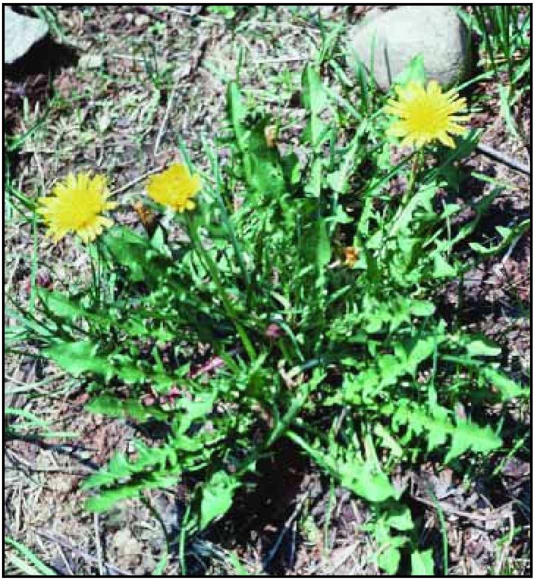

Dandelion
Taraxacum officinale *Composite Family, Asteraceae*

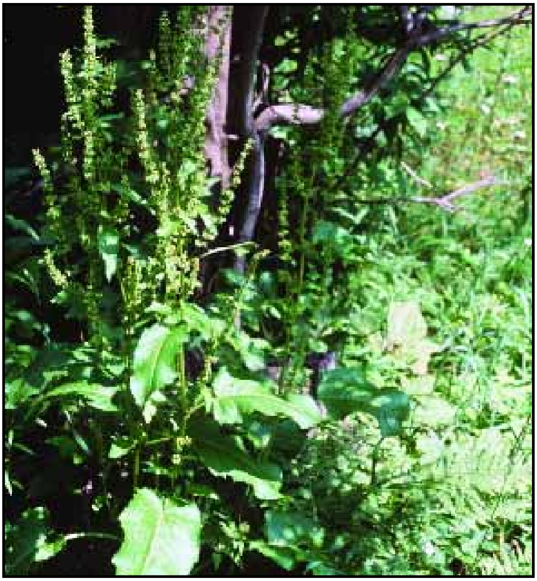

Dock
Rumex crispus *Buckwheat Family, Portulaceae*

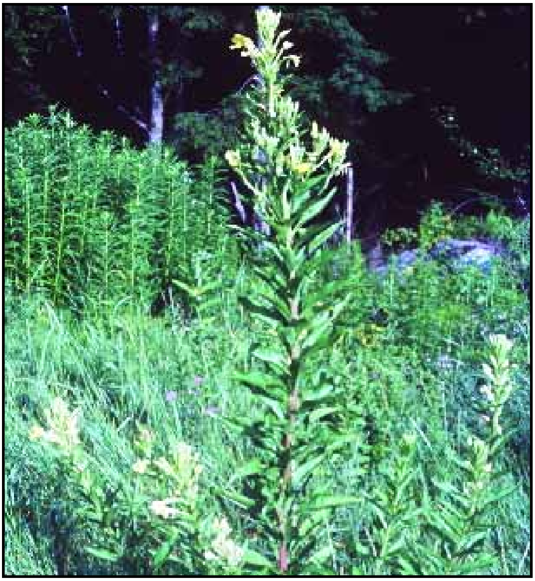

Evening Primrose
Oenothera biennis *Evening Primrose Family, Onagraceae*

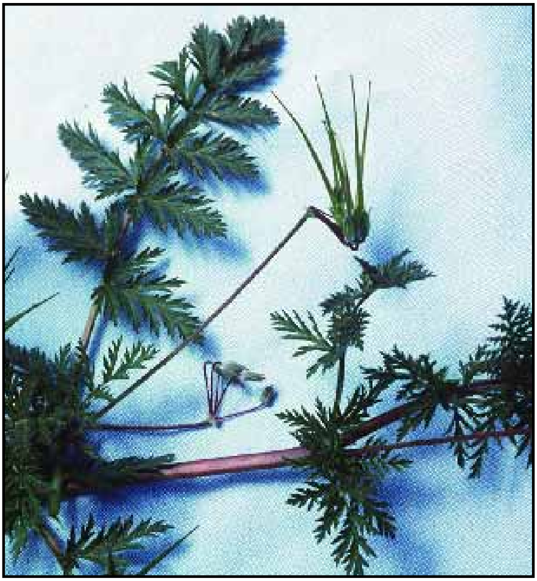

Filarie
Erodium circutarium *Cranesbill Family, Geraniaceae*

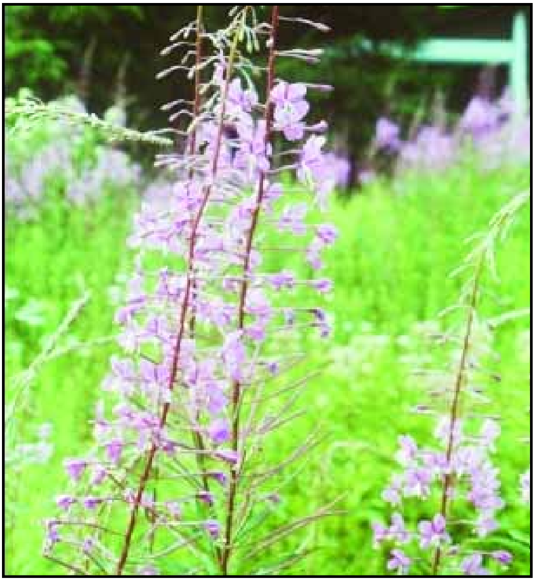

Fireweed
Epilobium angustifolium *Evening Primrose Family, Onagraceae*

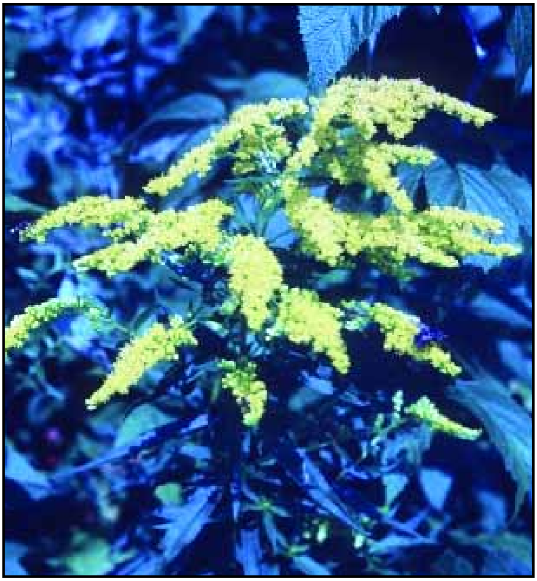

Goldenrod
Solidago odora *Sunflower Family, Compositae*

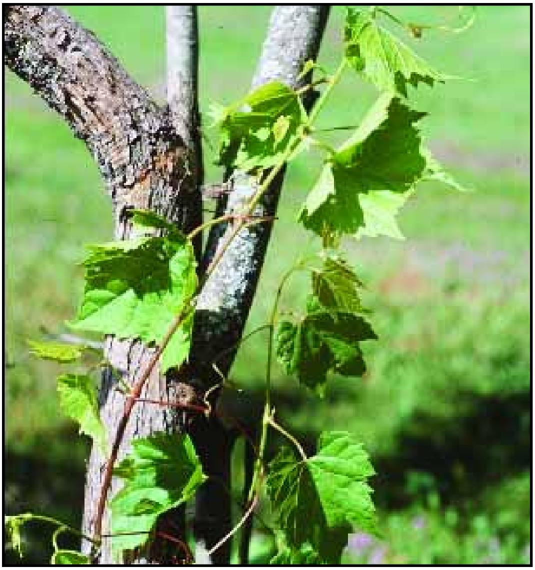

Grape
Vitis arizonica (Western) Vitis labrusca (Eastern) Vitis rotundifolia (South)
Vitis Family, Vitaceae

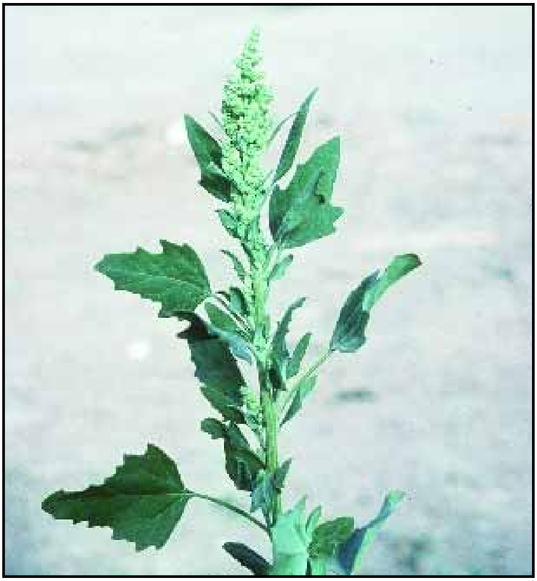

Lamb's Quarters
Chenopodium album *Goosefoot Family, Chenopodiaceae*

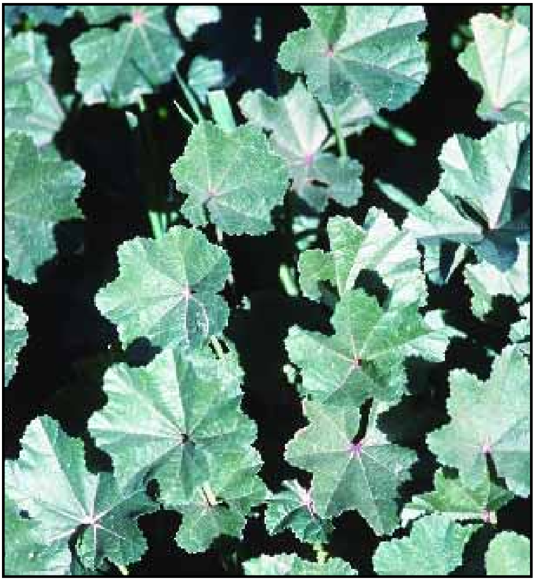

Malva
Malva Neglecta *Mallow Family, Malvaceae*

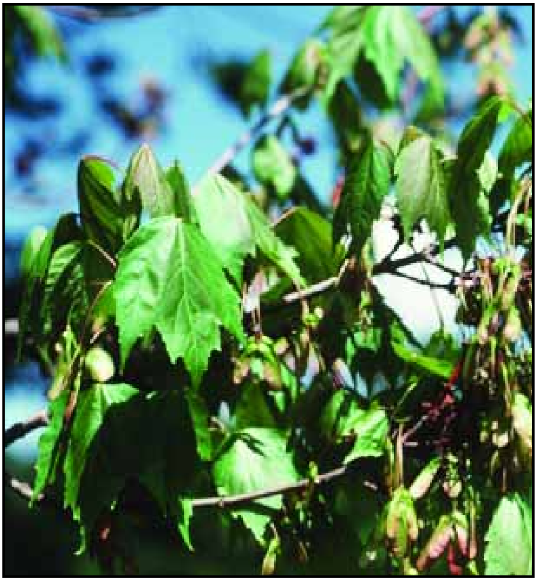

Maple
Aceraceae Family

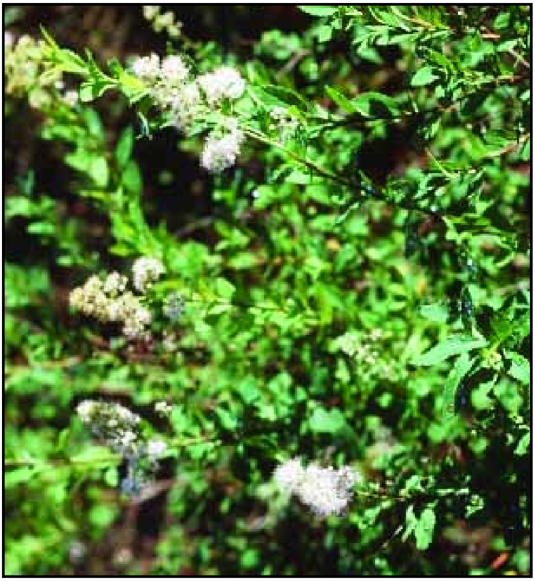

Meadowsweet
Filipendula ulmaria *Rose Family, Rosaceae*

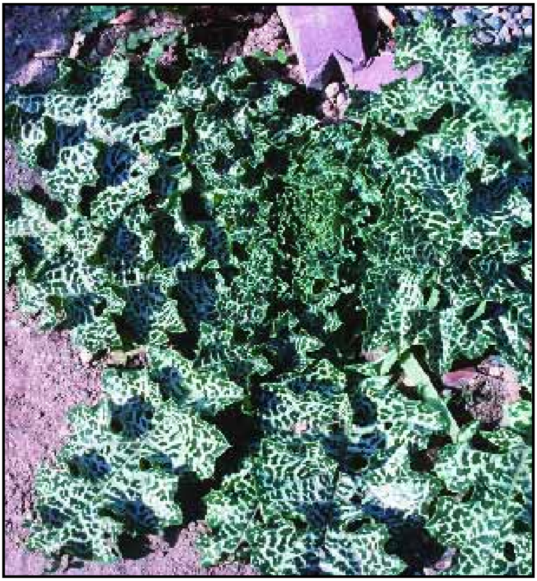

Milk Thistle
Silybum marianum *Composite Family, Compositae*

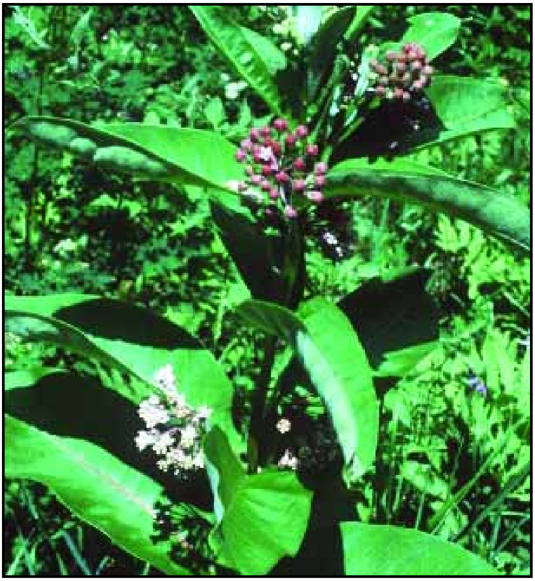

Milkweed
Asclepias species *Milkweed Family, Aslepiadaceae*

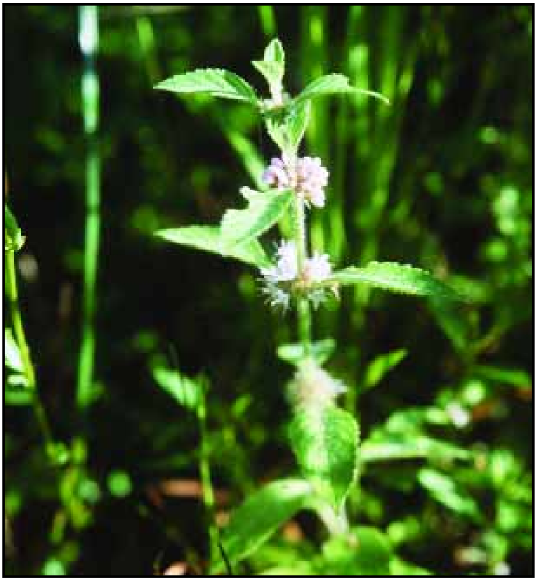

Mint
Mentha species *Labiatae Family*

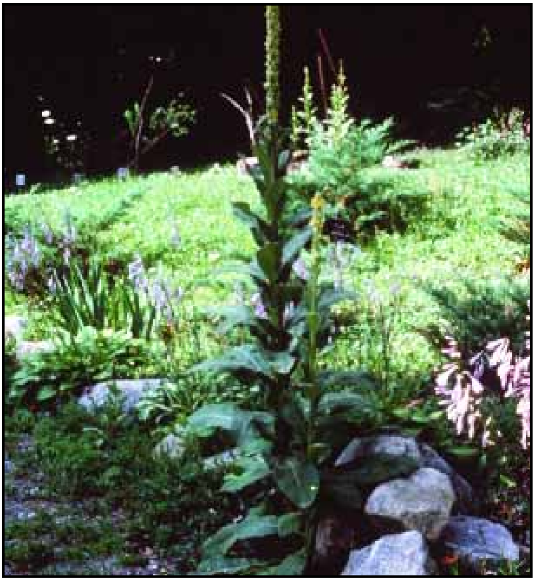

Mullein
Verbascum thapsus *Snapdragon Family, Scrophulariaceae*

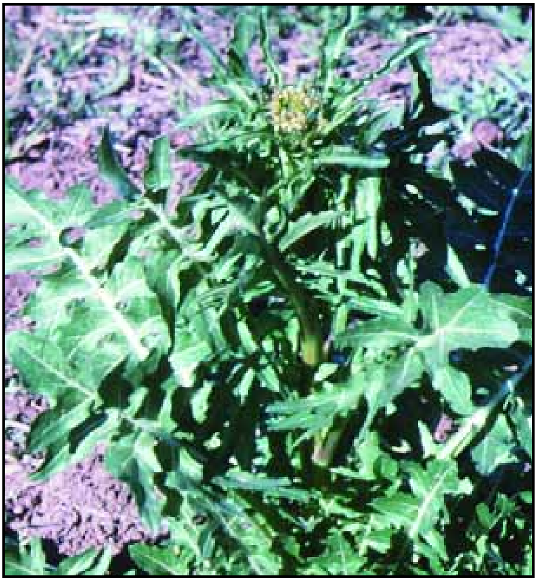

Mustard, yellow
Brassica species *Mustard Family, Brassiacaceae*

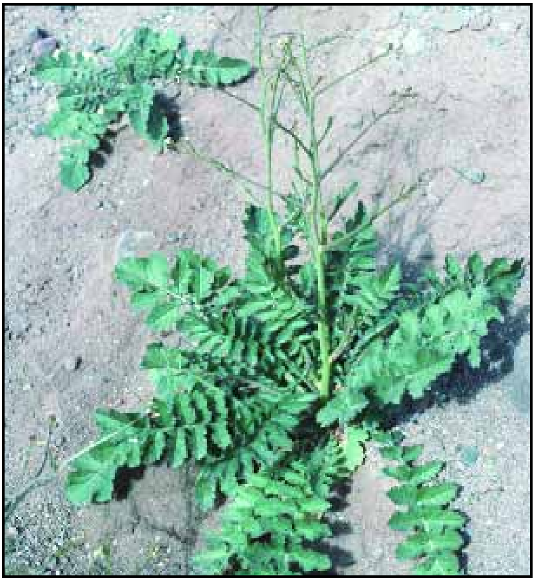

Mustard, black
Brassica species *Mustard Family, Brassicerceae*

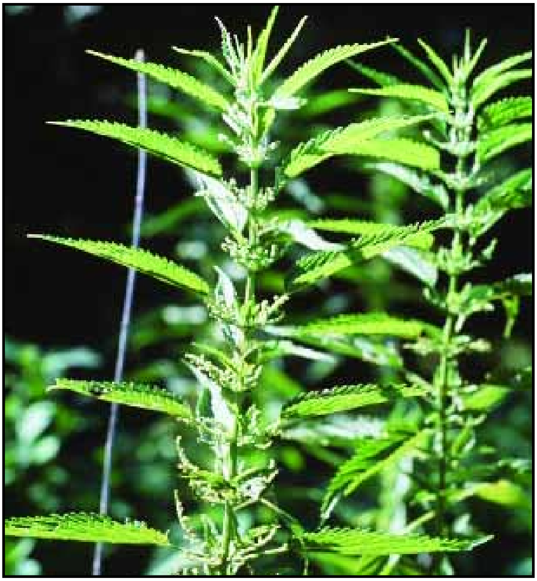

Nettles
Urtica species *Urticaceae Family*

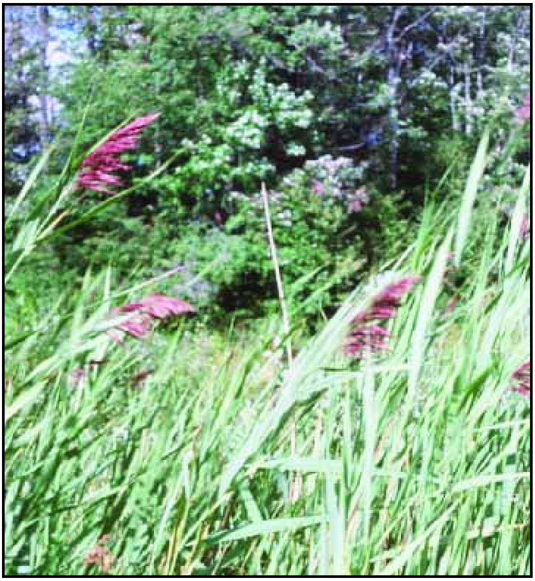

Phragmities
Phragmities communis *Grass Family, Graminae*

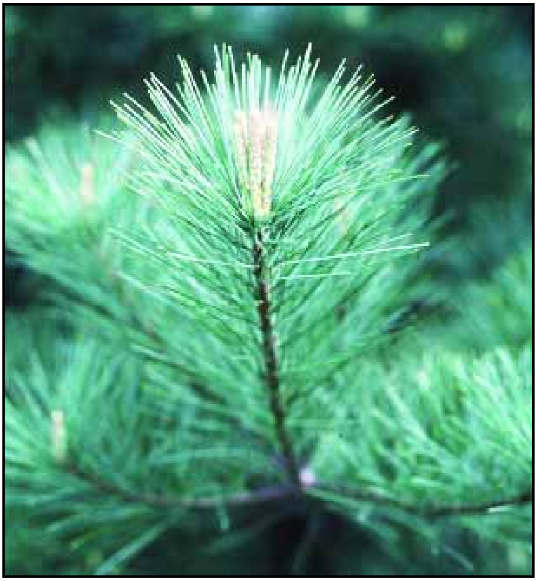

Pine

Pinus species *Pine Family, Pinaceae*

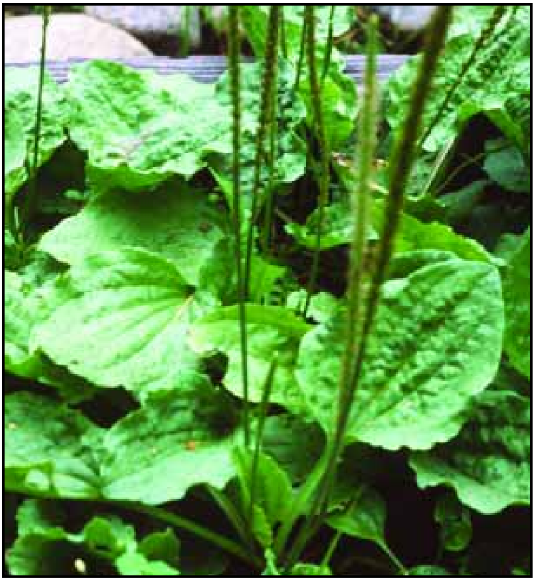

Plantain

Plantago major (common plantain) *Plantain Family, Plantaginaceae*

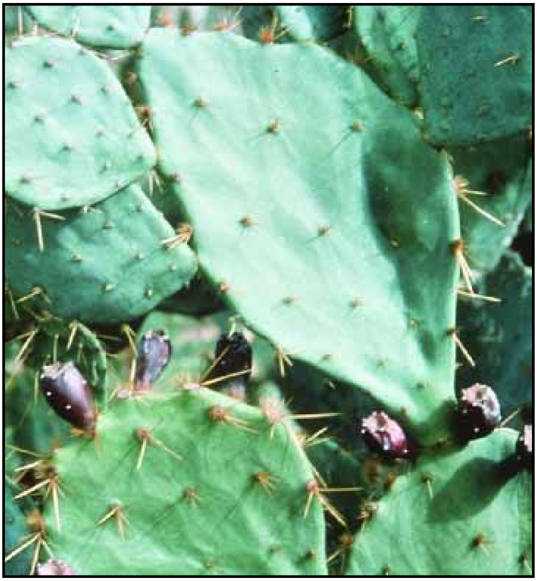

Prickly Pear
Opuntia megacantha *Cactus Family, Cactaceae*

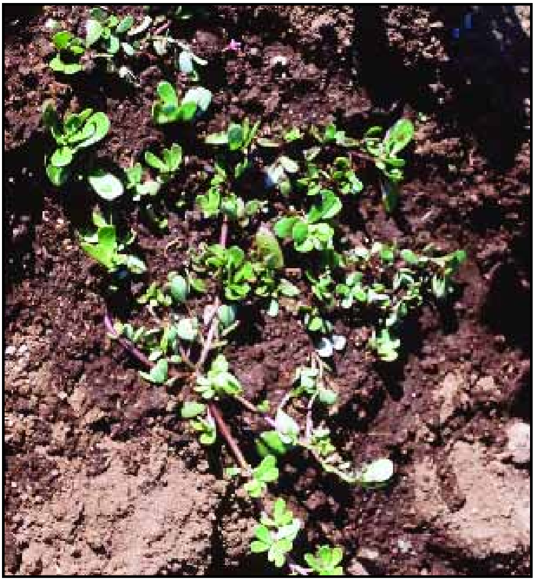

Purslane
Portulaca species *Purslane Family, Portulacaceae*

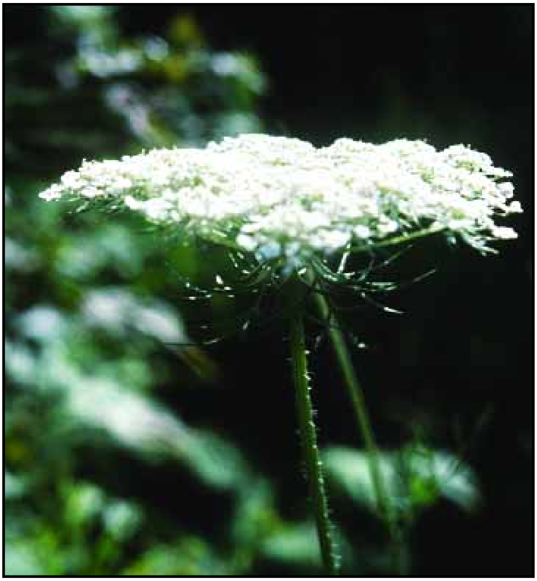

Queen Anne's Lace
Daucus carota *Parsley Family, Umbelliferae*

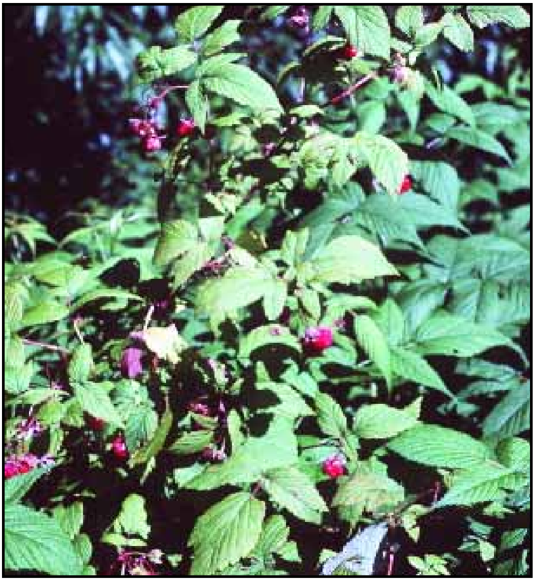

Raspberry
Species Rose *Family, Rosaceae*

Rose
Rosa rugosa, R. Caroline *Rose Family, Rosaceae*

Saguaro
Carnegiea gigantea *Cactus Family, Cactaceae*

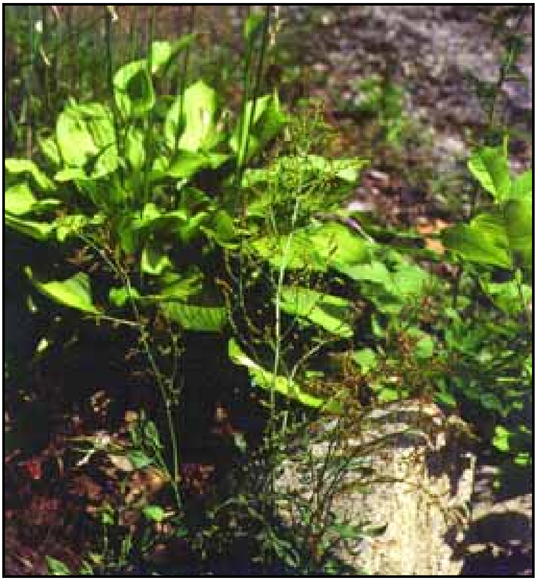

Sheep Sorrel
Rumex acetosella *Buckwheat Family, Polygonaceae*

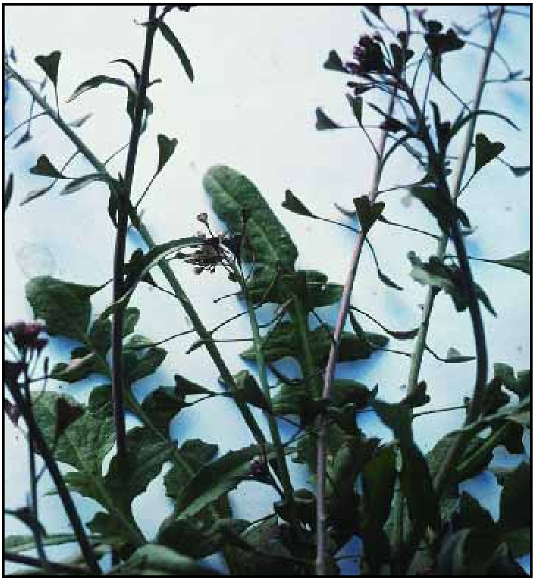

Shepherd's Purse
Capsella bursa-pastoris *Mustard Family, Brassicaceae*

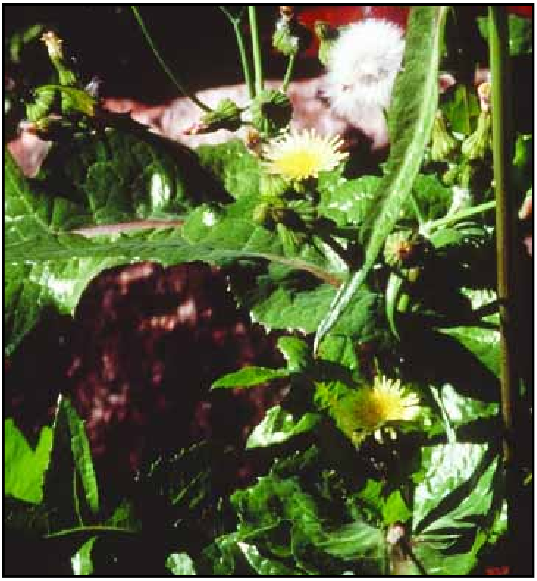

Sow Thistle
Sonchus oleraceus *Composite Family, Compositae*

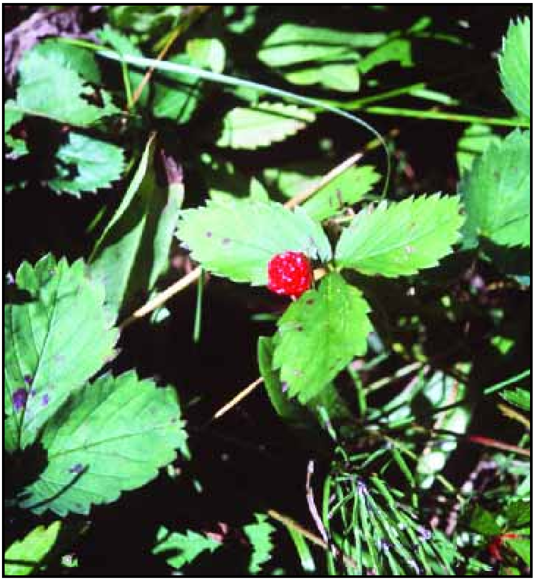

Strawberry
Fragaria virginiana *Rose Family, Rosaceae*

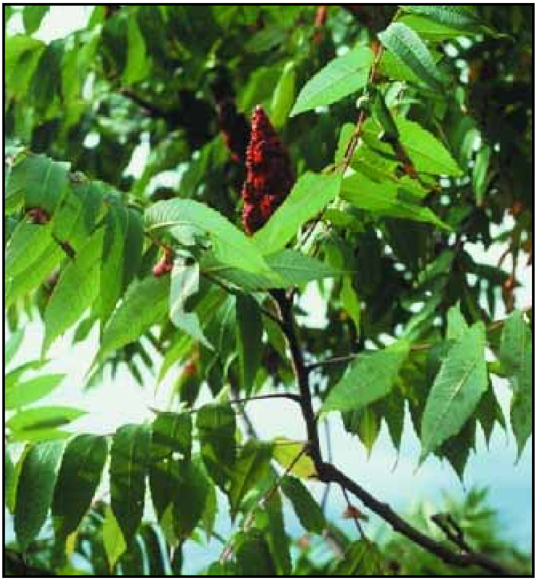

Sumac
Rhus species *Cashew Family, Anacardiaceae*

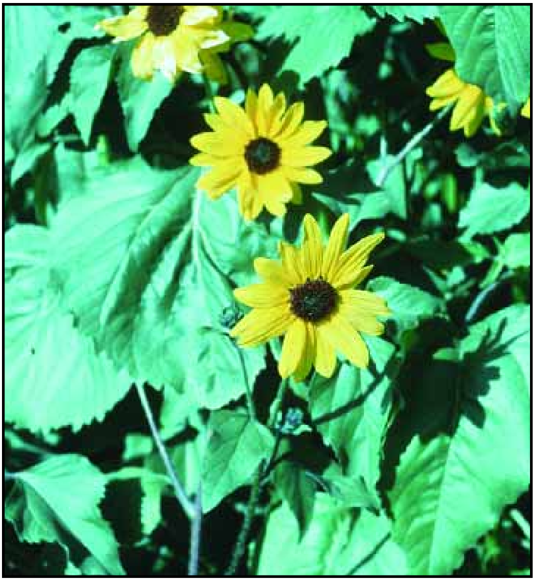

Sunflower
Helianthus species *Composite Family, Compositae*

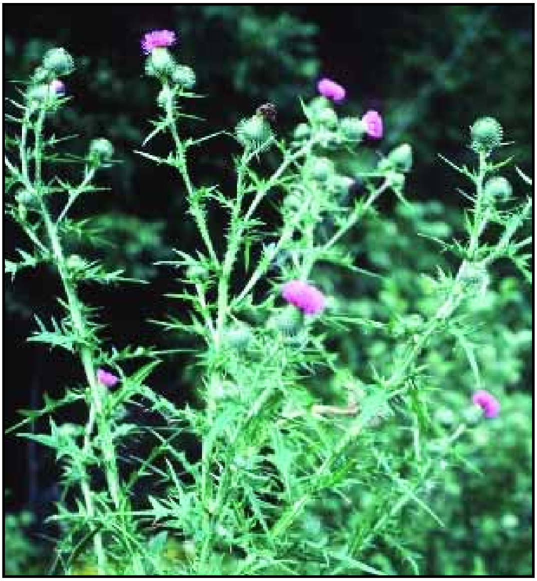

Thistle
Circium species *Composite Family, Compositae*

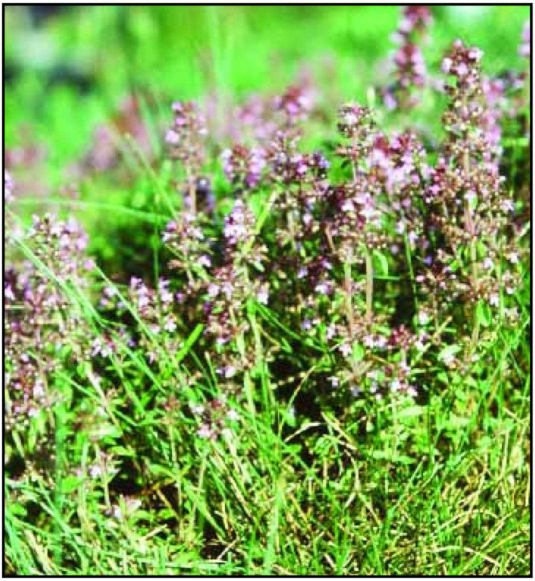

Thyme
Thymus vulgaris *Mint Family, Labiatae*

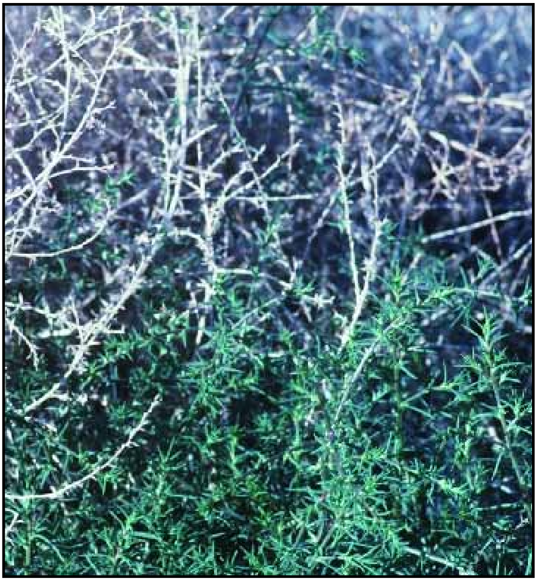

Tumbleweed
Salsola kali *Goosefoot Family, Chenopodiaceae*

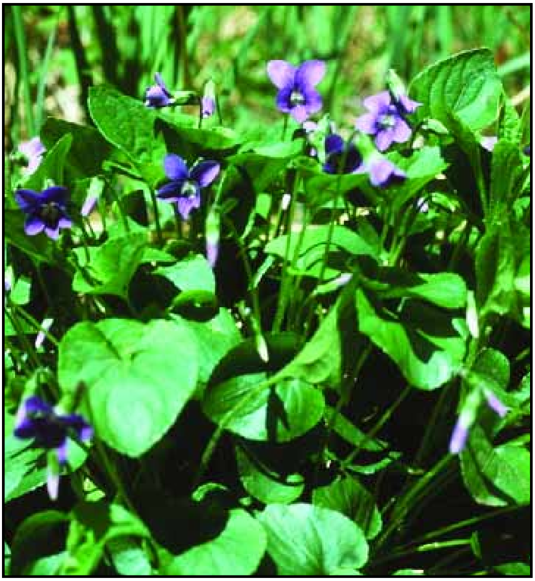

Violets
Viola species *Violet Family, Violaceae*

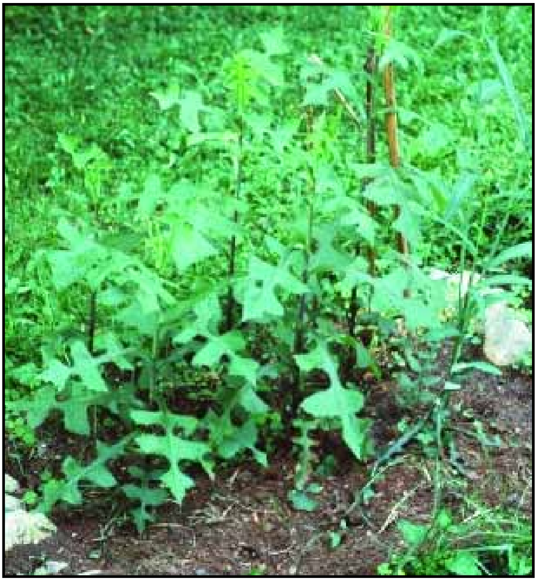

Wild Lettuce
Lactuca canadensis *Composite Family, Compositae*

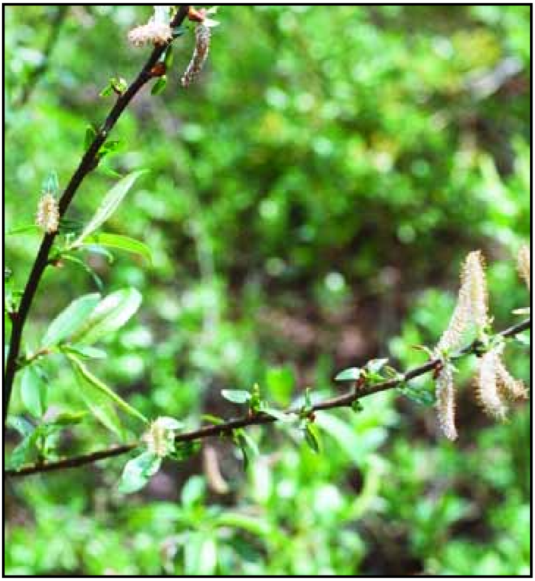

Willow
Salix species *Willow Family, Salicaceae*

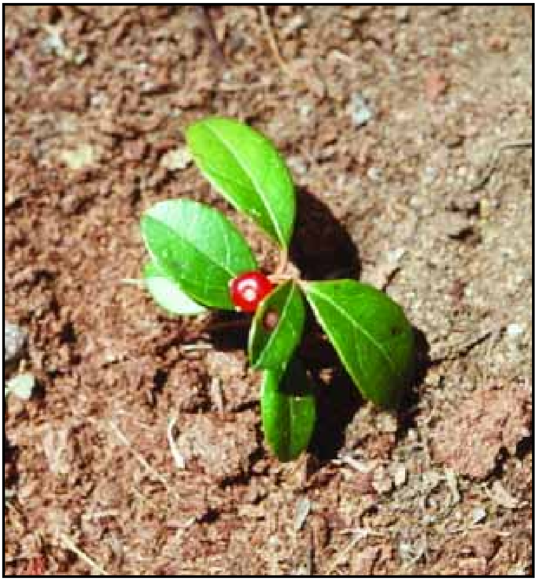

Wintergreen
Gaultheria procumbens *Heath Family, Ericaceae*

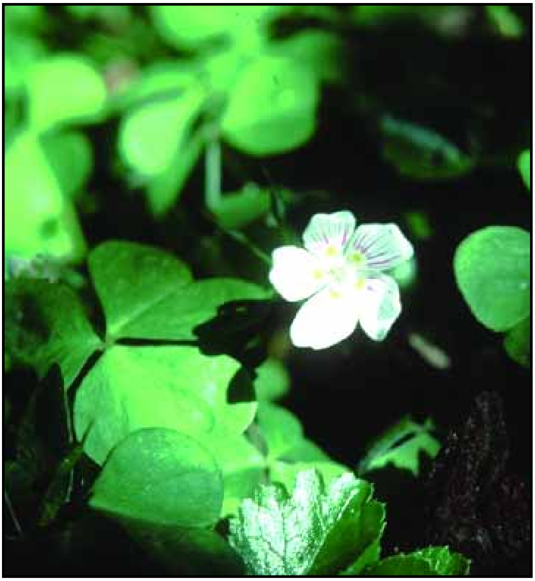

Wood Sorrel
Oxalis species *Wood Sorrel Family, Oxalidaceae*

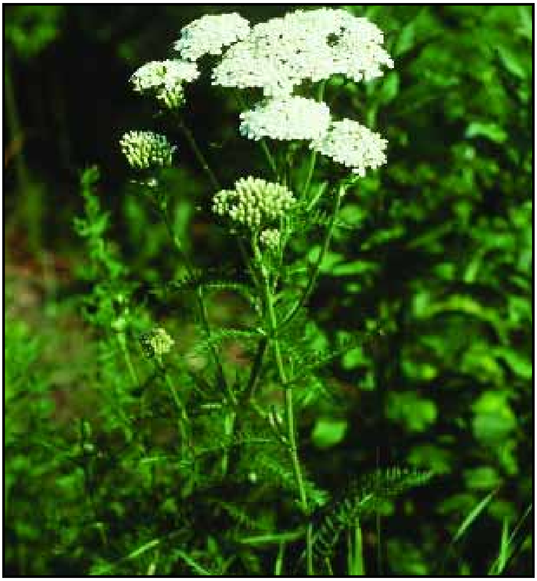

Yarrow

Achillea millefolium *Composite Family, Compositae*

Linda building a walk.

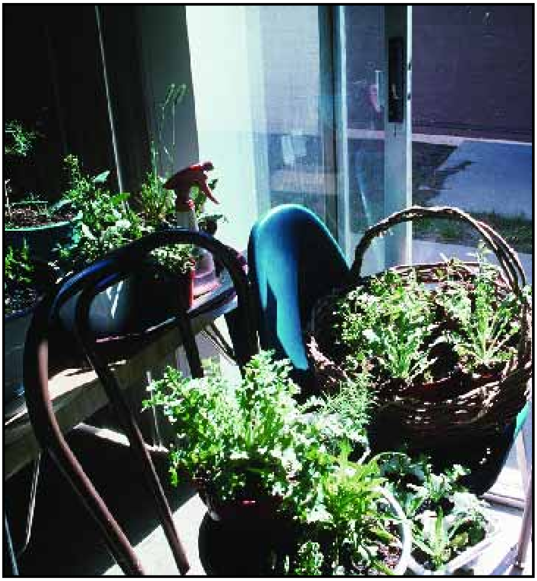

Field in Pots

Home Health Aides Provided Lunch for Linda

Nutrient Value of Wild Foods

I do not advocate giving up one's daily diet, but rather supplementing it with wild foods. My homesteading days taught me the value of a balanced vegetarian diet. Note the following dated 1997 - Some variations have been noted to date.

Some USDA recommended daily dietary allowances for adults are:

- Vitamin A—1,000 micrograms (males) and 800 micrograms (females)
- Vitamin C—60 milligrams
- Vitamin D—5 micrograms
- Vitamin B1 (thiamin)—1.4 milligrams (males) or 1.0 milligrams (females)
- Vitamin B2 (riboflavin)—1.6 milligrams (males) or 1.2 milligrams (females)
- Vitamin B3—20 milligrams
- Vitamin B6—2 milligrams
- Vitamin B12—3 micrograms
- Calcium—1,200 milligrams
- Iron—18 milligrams
- Phosphorous—800 milligrams
- Potassium—1,875 milligrams
- Protein—56 grams (males) or 44 grams (females)

So, you can see from the above how well wild foods can boost the nutritional value of your meal.

The listings that follow show the nutritional values for 1/2 cup quantities of uncooked plant matter.

As much as possible, the data was derived from Duke and Atchley, 1986 *CRC Handbook of Proximate Analyses*, CRC Press, 1986. Duke's FNF database; or from Joseph Laferriere's Nutricomp database. When no other data were available we used Mark Pedersen's Nutritional Herbology (Pedersen Publishing, 1987).

When available data was not specific to the plant or plant part we used data from a related species, or genus, or similar species in a different family. When absolutely no pertinent data was uncovered—as, for example, for violet flowers—averages for flowers such as those published by Duke and Atchley, were used. Occasionally it was necessary to estimate the water content, back-calculating from dried to a fresh basis. Adjustments were then made for carbohydrate and/or fiber content, ensuring that the figures were within a normal range for the plant.

How to read the Nutritional Listings

All quantities given are amounts per 1/2 cup raw food item (≈100 grams).

The figure for Calories is number of Calories per 1/2 cup (100 g) raw food.

Column 1: Measurements are in grams (g) per 100 g (1/2 cup). (Because of the mathematical coincidence, this could also be expressed as a percentage of the whole.)
Abbreviations:
- Carbos. = Carbohydrates
- Tr. = trace amount.

Column 2: Measurements are in milligrams (mg) per 100 g (1/2 cup).
Abbreviations:
- Phos. = Phosphorus
- Ascorbic. = Ascorbic acid
- Tr. = trace amount.

Column 3: Measurements are in micrograms (μg) per 100 g (1/2 cup).
Abbreviations:
- β Carotene = Beta-carotene
- Tr. = trace amount.

No. 1 Aloe Juice

Amount of nutrient per 100g. (About ½ cup)				Calories: 1	
Water	99.5 g	Calcium	2 mg	β Carotene	Tr.
Protein	0.1 g	Phos.	2 mg	Thiamin	Tr.
Fat	Tr.	Iron	0.2 mg	Riboflavin	Tr.
Carbos.	0.4 g	Sodium	0.3 mg	Niacin	Tr.
Fiber	0.1 g	Potassium	0.4 mg		
Ash	Tr.	Ascorbic.	3.1 mg		

No. 2 Amaranth Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 35	
Water	91.6 g	Calcium	448 mg	β Carotene	4300 µg
Protein	2.9 g	Phos.	85 mg	Thiamin	65 µg
Fat	0.4 g	Iron	13 mg	Riboflavin	300 µg
Carbos.	4.5 g	Sodium	20 mg	Niacin	1300 µg
Fiber	1.1 g	Potassium	617 mg		
Ash	1.5 g	Ascorbic.	53 mg		

No. 3 Amaranth Seed

Amount of nutrient per 100g. (About ½ cup)				Calories: 358	
Water	12.3 g	Calcium	247 mg	β Carotene	0.0 µg
Protein	12.9 g	Phos.	500 mg	Thiamin	140 µg
Fat	7.2 g	Iron	3.4 mg	Riboflavin	320 µg
Carbos.	65.1 g	Sodium	33.5 mg	Niacin	1000 µg
Fiber	6.7 g	Potassium	52.5 mg		
Ash	2.5 g	Ascorbic.	3.0 mg		

No. 4 Arrowhead Tubers

Amount of nutrient per 100g. (About ½ cup)				Calories: 99	
Water	72.5 g	Calcium	10 mg	β Carotene	0 µg
Protein	5.3 g	Phos.	174 mg	Thiamin	200 µg
Fat	0.3 g	Iron	2.6 mg	Riboflavin	100 µg
Carbos.	20.2 g	Sodium	225 mg	Niacin	1700 µg
Fiber	0.8 g	Potassium	922 mg		
Ash	1.7 g	Ascorbic.	1 mg		

No. 5 Arrowhead Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 62	
Water	85 g	Calcium	82 mg	β Carotene	1700 µg
Protein	2.6 g	Phos.	45 mg	Thiamin	200 µg
Fat	1.0 g	Iron	7 mg	Riboflavin	300 µg
Carbos.	9.9 g	Sodium	21 mg	Niacin	1500 µg
Fiber	0.9 g	Potassium	606 mg		
Ash	1.5 g	Ascorbic.	60 mg		

No. 6 Aster Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 39	
Water	87.2 g	Calcium	42 mg	β Carotene	3345 µg
Protein	4.2 g	Phos.	76 mg	Thiamin	180 µg
Fat	0.7 g	Iron	4 mg	Riboflavin	360 µg
Carbos.	6.4 g	Sodium	40 mg	Niacin	1100 µg
Fiber	1.1 g	Potassium	533 mg		
Ash	1.5 g	Ascorbic.	88 mg		

No. 7 Balsam Fir Needles

Amount of nutrient per 100g. (About ½ cup)				Calories: 35	
Water	88 g	Calcium	186 mg	β Carotene	5652 µg
Protein	2.8 g	Phos.	57 mg	Thiamin	110 µg
Fat	0.5 g	Iron	2.7 mg	Riboflavin	240 µg
Carbos.	6.5 g	Sodium	40 mg	Niacin	900 µg
Fiber	1.3 g	Potassium	382 mg		
Ash	2.1 g	Ascorbic.	64 mg		

No. 8 Birch Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 1	
Water	88.0 g	Calcium	186 mg	β Carotene	5652 µg
Protein	3.3 g	Phos.	57 mg	Thiamin	110 µg
Fat	1.0 g	Iron	2.7 mg	Riboflavin	240 µg
Carbos.	6.7 g	Sodium	40 mg	Niacin	900 µg
Fiber	2.0 g	Potassium	382 mg		
Ash	0.9 g	Ascorbic.	68 mg		

No. 9 Blackberry Fruit

Amount of nutrient per 100g. (About ½ cup)				Calories: 57	
Water	84.4 g	Calcium	29 mg	β Carotene	1000 μg
Protein	1.2 g	Phos.	26 mg	Thiamin	200 μg
Fat	0.7 g	Iron	1.3 mg	Riboflavin	500 μg
Carbos.	13.3 g	Sodium	1 mg	Niacin	600 μg
Fiber	3.7 g	Potassium	168 mg		
Ash	0.5 g	Ascorbic.	22 mg		

No. 10 Blueberry Fruit

Amount of nutrient per 100g. (About ½ cup)				Calories: 62	
Water	83.2 g	Calcium	15 mg	β Carotene	60 μg
Protein	0.7 g	Phos.	13 mg	Thiamin	30 μg
Fat	0.5 g	Iron	1 mg	Riboflavin	60 μg
Carbos.	15.3 g	Sodium	1 mg	Niacin	500 μg
Fiber	1.5 g	Potassium	81 mg		
Ash	0.3 g	Ascorbic.	14 mg		

No. 11 Bulrush Flowers

Amount of nutrient per 100g. (About ½ cup)				Calories: 16	
Water	94.8 g	Calcium	47 mg	β Carotene	400 μg
Protein	0.4 g	Phos.	86 mg	Thiamin	20 μg
Fat	0.1 g	Iron	1.0 mg	Riboflavin	110 μg
Carbos.	4.4 g	Sodium	18 mg	Niacin	600 μg
Fiber	1.2 g	Potassium	166 mg		
Ash	0.3 g	Ascorbic.	18 mg		

No. 12 Bulrush Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 342	
Water	11.1 g	Calcium	24 mg	β Carotene	0 μg
Protein	10.7 g	Phos.	267 mg	Thiamin	450 μg
Fat	8.9 g	Iron	3.1 mg	Riboflavin	200 μg
Carbos.	67.7 g	Sodium	6 mg	Niacin	3800 μg
Fiber	2.6 g	Potassium	316 mg		
Ash	1.6 g	Ascorbic.	0 mg		

No. 13 Bulrush Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	89.9 g	Calcium	50 mg	β Carotene	1166 μg
Protein	0.8 g	Phos.	18 mg	Thiamin	150 μg
Fat	0.2 g	Iron	1.1 mg	Riboflavin	190 μg
Carbos.	7.5 g	Sodium	9 mg	Niacin	1000 μg
Fiber	2.8 g	Potassium	283 mg		
Ash	1.4 g	Ascorbic.	41 mg		

No. 14 Burdock Root

Amount of nutrient per 100g. (About ½ cup)				Calories: 89	
Water	76.5 g	Calcium	50 mg	β Carotene	0 μg
Protein	2.5 g	Phos.	58 mg	Thiamin	250 μg
Fat	0.1 g	Iron	1.2 mg	Riboflavin	80 μg
Carbos.	20.1 g	Sodium	30 mg	Niacin	300 μg
Fiber	1.7 g	Potassium	180 mg		
Ash	0.8 g	Ascorbic.	2 mg		

No. 15 Cattail Roots

Amount of nutrient per 100g. (About ½ cup)				Calories: 73	
Water	79.7 g	Calcium	32 mg	β Carotene	568 μg
Protein	1.6 g	Phos.	54 mg	Thiamin	90 μg
Fat	0.6 g	Iron	1 mg	Riboflavin	50 μg
Carbos.	17.5 g	Sodium	14 mg	Niacin	800 μg
Fiber	1.0 g	Potassium	399 mg		
Ash	1.0 g	Ascorbic.	18 mg		

No. 16 Cattail Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	81.3 g	Calcium	58 mg	β Carotene	2200 μg
Protein	1.8 g	Phos.	109 mg	Thiamin	300 μg
Fat	0.3 g	Iron	2 mg	Riboflavin	400 μg
Carbos.	15.0 g	Sodium	8 mg	Niacin	800 μg
Fiber	6.2 g	Potassium	639 mg		
Ash	1.6 g	Ascorbic.	76 mg		

No. 17 Chamomile Flowers

Amount of nutrient per 100g. (About ½ cup)				Calories: 56	
Water	81.2 g	Calcium	126 mg	β Carotene	41 µg
Protein	2.2 g	Phos.	60 mg	Thiamin	15 µg
Fat	0.7 g	Iron	3.2 mg	Riboflavin	80 µg
Carbos.	14.7 g	Sodium	48.5 mg	Niacin	2801 µg
Fiber	1.4 g	Potassium	248 mg		
Ash	1.2 g	Ascorbic.	5 mg		

No. 18 Chickweed Leaves

Amount of nutrient per 100g. (About ⅓ cup)				Calories: 28	
Water	91.7 g	Calcium	160 mg	β Carotene	200 µg
Protein	1.2 g	Phos.	49 mg	Thiamin	Tr.
Fat	0.2 g	Iron	29 mg	Riboflavin	100 µg
Carbos.	5.3 g	Sodium	82 mg	Niacin	500 µg
Fiber	1.7 g	Potassium	243 mg		
Ash	1.6 g	Ascorbic.	350 mg		

No. 19 Chicory

Amount of nutrient per 100g. (About ⅓ cup)				Calories: 30	
Water	92.0 g	Calcium	100 mg	β Carotene	2400 µg
Protein	1.7 g	Phos.	47 mg	Thiamin	100 µg
Fat	0.3 g	Iron	0.9 mg	Riboflavin	100 µg
Carbos.	4.7 g	Sodium	45 mg	Niacin	500 µg
Fiber	0.8 g	Potassium	420 mg		
Ash	1.3 g	Ascorbic.	24 mg		

No. 20 Chicory Roots

Amount of nutrient per 100g. (About ½ cup)				Calories: 73	
Water	80.0 g	Calcium	41 mg	β Carotene	Tr.
Protein	1.4 g	Phos.	61 mg	Thiamin	Tr.
Fat	0.2 g	Iron	0.8 mg	Riboflavin	Tr.
Carbos.	17.5 g	Sodium	50 mg	Niacin	400 µg
Fiber	2.0 g	Potassium	290 mg		
Ash	0.9 g	Ascorbic.	5 mg		

No. 21 Cholla Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 31.0	
Water	77.8 g	Calcium	3 mg	β Carotene	Tr.
Protein	1.6 g	Phos.	1 mg	Thiamin	Tr.
Fat	0.5 g	Iron	1.1 mg	Riboflavin	Tr.
Carbos.	15.9 g	Sodium	0.4 mg	Niacin	Tr.
Fiber	1.7 g	Potassium	12 mg		
Ash	4.2 g	Ascorbic.	57 mg		

No. 22 Clover Flowers

Amount of nutrient per 100g. (About ½ cup)				Calories: 54	
Water	82.3 g	Calcium	232 mg	β Carotene	200 μg
Protein	2.0 g	Phos.	57 mg	Thiamin	100 μg
Fat	0.6 g	Iron	1 mg	Riboflavin	100 μg
Carbos.	13.5 g	Sodium	3 mg	Niacin	2200 μg
Fiber	1.8 g	Potassium	354 mg		
Ash	1.5 g	Ascorbic.	52 mg		

No. 23 Clover Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	89.9 g	Calcium	142 mg	β Carotene	1166 μg
Protein	2.1 g	Phos.	32 mg	Thiamin	150 μg
Fat	0.5 g	Iron	1.1 mg	Riboflavin	190 μg
Carbos.	6.3 g	Sodium	4 mg	Niacin	1000 μg
Fiber	2.2 g	Potassium	345 mg		
Ash	1.1 g	Ascorbic.	57 mg		

No. 24 Daisy Flowers

Amount of nutrient per 100g. (About ½ cup)				Calories: 17	
Water	92.6 g	Calcium	56 mg	β Carotene	8800 μg
Protein	1.6 g	Phos.	32 mg	Thiamin	30 μg
Fat	0.2 g	Iron	3.1 mg	Riboflavin	20 μg
Carbos.	4.4 g	Sodium	52 mg	Niacin	860 μg
Fiber	0.9 g	Potassium	570 mg		
Ash	1.3 g	Ascorbic.	53 mg		

No. 25 Daisy Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 19	
Water	93.5 g	Calcium	63 mg	β Carotene	3160 µg
Protein	1.8 g	Phos.	34 mg	Thiamin	90 µg
Fat	0.3 g	Iron	2.5 mg	Riboflavin	190 µg
Carbos.	3.3 g	Sodium	106 mg	Niacin	600 µg
Fiber	0.9 g	Potassium	256 mg		
Ash	1.1 g	Ascorbic.	27 mg		

No. 26 Dandelion Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 45	
Water	85.6 g	Calcium	187 mg	β Carotene	800 µg
Protein	2.7 g	Phos.	66 mg	Thiamin	200 µg
Fat	0.7 g	Iron	3 mg	Riboflavin	300 µg
Carbos.	9.2 g	Sodium	76 mg	Niacin	900 µg
Fiber	1.6 g	Potassium	398 mg		
Ash	1.8 g	Ascorbic.	68 mg		

No. 27 Dock

Amount of nutrient per 100g. (About ½ cup)				Calories: 21	
Water	92.6 g	Calcium	74 mg	β Carotene	2770 µg
Protein	1.5 g	Phos.	56 mg	Thiamin	60 µg
Fat	0.3 g	Iron	5.6 mg	Riboflavin	80 µg
Carbos.	4.1 g	Sodium	5 mg	Niacin	400 µg
Fiber	0.9 g	Potassium	338 mg		
Ash	1.5 g	Ascorbic.	30 mg		

No. 28 Dock Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 342	
Water	14.3 g	Calcium	35 mg	β Carotene	0 µg
Protein	12.9 g	Phos.	305 mg	Thiamin	450 µg
Fat	3.8 g	Iron	3.4 mg	Riboflavin	200 µg
Carbos.	66.9 g	Sodium	4 mg	Niacin	3800 µg
Fiber	19.3 g	Potassium	339 mg		
Ash	2.8 g	Ascorbic.	0 mg		

No. 29 Evening Primrose Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 342	
Water	7.0 g	Calcium	1422 mg	β Carotene	0 µg
Protein	15.1 g	Phos.	533 mg	Thiamin	450 µg
Fat	2.2 g	Iron	22 mg	Riboflavin	200 µg
Carbos.	50.4 g	Sodium	16 mg	Niacin	3800 µg
Fiber	3.6 g	Potassium	542 mg		
Ash	5.3 g	Ascorbic.	0 mg		

No. 30 Filarie Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	81.5 g	Calcium	464 mg	β Carotene	1166 µg
Protein	2.4 g	Phos.	87 mg	Thiamin	150 µg
Fat	0.4 g	Iron	1.1 mg	Riboflavin	190 µg
Carbos.	12.2 g	Sodium	4 mg	Niacin	1000 µg
Fiber	4.5 g	Potassium	622 mg		
Ash	3.5 g	Ascorbic.	57 mg		

No. 31 Filarie Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 32	
Water	85.8 g	Calcium	140 mg	β Carotene	4000 µg
Protein	4.0 g	Phos.	50 mg	Thiamin	100 µg
Fat	0.4 g	Iron	2.5 mg	Riboflavin	200 µg
Carbos.	7.3 g	Sodium	60 mg	Niacin	700 µg
Fiber	1.4 g	Potassium	410 mg		
Ash	2.5 g	Ascorbic.	63 mg		

No. 32 Fireweed Greens

Amount of nutrient per 100g. (About ½ cup)				Calories: 35	
Water	88.0 g	Calcium	186 mg	β Carotene	5652 µg
Protein	2.8 g	Phos.	57 mg	Thiamin	110 µg
Fat	0.7 g	Iron	2.7 mg	Riboflavin	240 µg
Carbos.	8.2 g	Sodium	40 mg	Niacin	900 µg
Fiber	1.3 g	Potassium	382 mg		
Ash	0.8 g	Ascorbic.	68 mg		

No. 33 Goldenrod Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	89.9 g	Calcium	113 mg	β Carotene	300 µg
Protein	0.9 g	Phos.	52 mg	Thiamin	150 µg
Fat	1.4 g	Iron	11 mg	Riboflavin	190 µg
Carbos.	7.0 g	Sodium	4 mg	Niacin	1000 µg
Fiber	1.3 g	Potassium	83 mg		
Ash	0.7 g	Ascorbic.	1000 mg		

No. 34 Grape Fruit

Amount of nutrient per 100g. (About ½ cup)				Calories: 63	
Water	81.3 g	Calcium	14 mg	β Carotene	Tr.
Protein	0.6 g	Phos.	10 mg	Thiamin	100 µg
Fat	0.4 g	Iron	0.3 mg	Riboflavin	100 µg
Carbos.	17.1 g	Sodium	2 mg	Niacin	300 µg
Fiber	0.8 g	Potassium	191 mg		
Ash	0.6 g	Ascorbic.	4 mg		

No. 35 Grape Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 68	
Water	79.2 g	Calcium	716 mg	β Carotene	5200 µg
Protein	4.8 g	Phos.	38 mg	Thiamin	200 µg
Fat	0.9 g	Iron	7 mg	Riboflavin	100 µg
Carbos.	13.5 g	Sodium	21 mg	Niacin	1200 µg
Fiber	2.1 g	Potassium	382 mg		
Ash	1.5 g	Ascorbic.	34 mg		

No. 36 Lamb's Quarters Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 353	
Water	8.3 g	Calcium	1036 mg	β Carotene	0 µg
Protein	19.6 g	Phos.	340 mg	Thiamin	450 µg
Fat	4.2 g	Iron	64 mg	Riboflavin	200 µg
Carbos.	57.7 g	Sodium	9 mg	Niacin	3800 µg
Fiber	27.1 g	Potassium	1687 mg		
Ash	10.2 g	Ascorbic.	0 mg		

No. 37 Lamb's Quarters Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 35	
Water	88.0 g	Calcium	324 mg	β Carotene	3800 µg
Protein	3.5 g	Phos.	48 mg	Thiamin	100 µg
Fat	0.8 g	Iron	1.5 mg	Riboflavin	200 µg
Carbos.	5.5 g	Sodium	4 mg	Niacin	1000 µg
Fiber	2.0 g	Potassium	684 mg		
Ash	2.4 g	Ascorbic.	40 mg		

No. 38 Mallow (Malva) Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 28	
Water	91.6 g	Calcium	90 mg	β Carotene	3315 µg
Protein	3.6 g	Phos.	42 mg	Thiamin	170 µg
Fat	1.4 g	Iron	3.7 mg	Riboflavin	290 µg
Carbos.	2.1 g	Sodium	60 mg	Niacin	500 µg
Fiber	0.9 g	Potassium	410 mg		
Ash	1.3 g	Ascorbic.	24 mg		

No. 39 Maple Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 1	
Water	88.0 g	Calcium	186 mg	β Carotene	5652 µg
Protein	3.3 g	Phos.	57 mg	Thiamin	110 µg
Fat	1.0 g	Iron	2.7 mg	Riboflavin	240 µg
Carbos.	6.7 g	Sodium	40 mg	Niacin	900 µg
Fiber	2.0 g	Potassium	382 mg		
Ash	0.9 g	Ascorbic.	68 mg		

No. 40 Meadowsweet Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	89.0 g	Calcium	140 mg	β Carotene	4000 µg
Protein	2.8 g	Phos.	50 mg	Thiamin	100 µg
Fat	0.4 g	Iron	2.5 mg	Riboflavin	200 µg
Carbos.	6.0 g	Sodium	60 mg	Niacin	700 µg
Fiber	1.2 g	Potassium	410 mg		
Ash	1.8 g	Ascorbic.	63 mg		

No. 41 Milk Thistle Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 342	
Water	11.1 g	Calcium	35 mg	β Carotene	0 µg
Protein	16.9 g	Phos.	305 mg	Thiamin	450 µg
Fat	25.7 g	Iron	3.4 mg	Riboflavin	200 µg
Carbos.	44.7 g	Sodium	4 mg	Niacin	3800 µg
Fiber	2.6 g	Potassium	339 mg		
Ash	1.6 g	Ascorbic.	0 mg		

No. 42 Milk Thistle Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	89.9 g	Calcium	65 mg	β Carotene	300 µg
Protein	1.4 g	Phos.	70 mg	Thiamin	150 µg
Fat	0.6 g	Iron	11 mg	Riboflavin	190 µg
Carbos.	7.2 g	Sodium	4 mg	Niacin	1000 µg
Fiber	0.6 g	Potassium	83 mg		
Ash	0.9 g	Ascorbic.	1000 mg		

No. 43 Milkweed Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	89.9 g	Calcium	31 mg	β Carotene	1166 µg
Protein	3.5 g	Phos.	59 mg	Thiamin	150 µg
Fat	0.5 g	Iron	1.1 mg	Riboflavin	190 µg
Carbos.	5.2 g	Sodium	4 mg	Niacin	1000 µg
Fiber	0.6 g	Potassium	345 mg		
Ash	0.9 g	Ascorbic.	57 mg		

No. 44 Mint Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 51	
Water	84.0 g	Calcium	160 mg	β Carotene	3000 µg
Protein	2.6 g	Phos.	80 mg	Thiamin	80 µg
Fat	0.8 g	Iron	13 mg	Riboflavin	160 µg
Carbos.	10.9 g	Sodium	40 mg	Niacin	650 µg
Fiber	1.9 g	Potassium	360 mg		
Ash	1.7 g	Ascorbic.	24 mg		

No. 45 Mullein Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 62	
Water	78.6 g	Calcium	285 mg	β Carotene	900 µg
Protein	2.3 g	Phos.	122 mg	Thiamin	Tr.
Fat	0.3 g	Iron	50 mg	Riboflavin	Tr.
Carbos.	17.1 g	Sodium	16 mg	Niacin	2000 µg
Fiber	2.4 g	Potassium	282 mg		
Ash	1.8 g	Ascorbic.	16 mg		

No. 46 Mustard Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	89.5 g	Calcium	183 mg	β Carotene	4200 µg
Protein	3.0 g	Phos.	50 mg	Thiamin	110 µg
Fat	0.5 g	Iron	3 mg	Riboflavin	220 µg
Carbos.	5.6 g	Sodium	32 mg	Niacin	800 µg
Fiber	1.1 g	Potassium	377 mg		
Ash	1.4 g	Ascorbic.	97 mg		

No. 47 Mustard Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 416	
Water	7.5 g	Calcium	320 mg	β Carotene	0.0 µg
Protein	22.2 g	Phos.	650 mg	Thiamin	540 µg
Fat	18.3 g	Iron	3.4 mg	Riboflavin	400 µg
Carbos.	45.6 g	Sodium	4 mg	Niacin	3800 µg
Fiber	15.9 g	Potassium	339 mg		
Ash	6.4 g	Ascorbic.	0 mg		

No. 48 Nettle Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 35	
Water	88.2 g	Calcium	34 mg	β Carotene	1100 µg
Protein	1.2 g	Phos.	5 mg	Thiamin	100 µg
Fat	2.7 g	Iron	5 mg	Riboflavin	100 µg
Carbos.	6.9 g	Sodium	1 mg	Niacin	600 µg
Fiber	1.3 g	Potassium	20 mg		
Ash	1.0 g	Ascorbic.	10 mg		

No. 49 Phragmities Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	89.9 g	Calcium	48 mg	β Carotene	1166 µg
Protein	0.5 g	Phos.	6 mg	Thiamin	150 µg
Fat	0.1 g	Iron	1.1 mg	Riboflavin	190 µg
Carbos.	9.0 g	Sodium	4 mg	Niacin	1000 µg
Fiber	4.1 g	Potassium	345 mg		
Ash	0.4 g	Ascorbic.	57 mg		

No. 50 Pine Nuts

Amount of nutrient per 100g. (About ½ cup)				Calories: 634	
Water	3.7 g	Calcium	14 mg	β Carotene	20 µg
Protein	15.3 g	Phos.	515 mg	Thiamin	760 µg
Fat	61.3 g	Iron	4 mg	Riboflavin	230 µg
Carbos.	16.8 g	Sodium	72 mg	Niacin	9800 µg
Fiber	2.6 g	Potassium	628 mg		
Ash	2.9 g	Ascorbic.	1 mg		

No. 51 Pine Needles

Amount of nutrient per 100g. (About ½ cup)				Calories: 141	
Water	51.5 g	Calcium	186 mg	β Carotene	5652 µg
Protein	3.1 g	Phos.	57 mg	Thiamin	110 µg
Fat	4.5 g	Iron	2.7 mg	Riboflavin	240 µg
Carbos.	39.9 g	Sodium	40 mg	Niacin	900 µg
Fiber	14.1 g	Potassium	382 mg		
Ash	1.0 g	Ascorbic.	68 mg		

No. 52 Plantain Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 61	
Water	81.4 g	Calcium	184 mg	β Carotene	2520 µg
Protein	2.5 g	Phos.	52 mg	Thiamin	95 µg
Fat	0.3 g	Iron	1.2 mg	Riboflavin	280 µg
Carbos.	14.6 g	Sodium	16 mg	Niacin	800 µg
Fiber	2.0 g	Potassium	277 mg		
Ash	1.2 g	Ascorbic.	8 mg		

No. 53 Plantain Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 342	
Water	11.2 g	Calcium	35 mg	β Carotene	0 µg
Protein	17.0 g	Phos.	305 mg	Thiamin	450 µg
Fat	7.6 g	Iron	3.4 mg	Riboflavin	200 µg
Carbos.	59.7 g	Sodium	4 mg	Niacin	3800 µg
Fiber	13.7 g	Potassium	339 mg		
Ash	4.5 g	Ascorbic.	0 mg		

No. 54 Prickly Pear Fruit

Amount of nutrient per 100g. (About ½ cup)				Calories: 67	
Water	81.4 g	Calcium	57 mg	β Carotene	100 µg
Protein	1.1 g	Phos.	32 mg	Thiamin	10 µg
Fat	0.4 g	Iron	1.2 mg	Riboflavin	20 µg
Carbos.	16.6 g	Sodium	1 mg	Niacin	300 µg
Fiber	1.1 g	Potassium	178 mg		
Ash	0.5 g	Ascorbic.	18 mg		

No. 55 Purslane Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 16	
Water	93.9 g	Calcium	65 mg	β Carotene	100 µg
Protein	1.3 g	Phos.	44 mg	Thiamin	Tr.
Fat	0.1 g	Iron	2 mg	Riboflavin	100 µg
Carbos.	3.4 g	Sodium	45 mg	Niacin	500 µg
Fiber	0.8 g	Potassium	494 mg		
Ash	1.2 g	Ascorbic.	21 mg		

No. 56 Queen Anne's Lace Roots

Amount of nutrient per 100g. (About ½ cup)				Calories: 37	
Water	89.6 g	Calcium	36 mg	β Carotene	7000 µg
Protein	1.1 g	Phos.	38 mg	Thiamin	60 µg
Fat	0.3 g	Iron	1.2 mg	Riboflavin	50 µg
Carbos.	8.3 g	Sodium	70 mg	Niacin	700 µg
Fiber	0.9 g	Potassium	245 mg		
Ash	0.7 g	Ascorbic.	8 mg		

No. 57 Raspberry Fruit

Amount of nutrient per 100g. (About ½ cup)				Calories: 73	
Water	80.8 g	Calcium	30 mg	β Carotene	0 µg
Protein	1.5 g	Phos.	22 mg	Thiamin	30 µg
Fat	1.4 g	Iron	0.9 mg	Riboflavin	90 µg
Carbos.	15.7 g	Sodium	1 mg	Niacin	900 µg
Fiber	5.1 g	Potassium	199 mg		
Ash	0.6 g	Ascorbic.	18 mg		

No. 58 Raspberry Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 46	
Water	83.1 g	Calcium	204 mg	β Carotene	1900 µg
Protein	1.9 g	Phos.	40 mg	Thiamin	100 µg
Fat	0.3 g	Iron	17 mg	Riboflavin	Tr.
Carbos.	13.4 g	Sodium	1 mg	Niacin	6500 µg
Fiber	1.4 g	Potassium	3226 mg		
Ash	1.4 g	Ascorbic.	62 mg		

No. 59 Rose Flowers

Amount of nutrient per 100g. (About ½ cup)				Calories: 16	
Water	94.8 g	Calcium	47 mg	β Carotene	400 µg
Protein	1.4 g	Phos.	86 mg	Thiamin	20 µg
Fat	0.3 g	Iron	1.0 mg	Riboflavin	110 µg
Carbos.	2.7 g	Sodium	18 mg	Niacin	600 µg
Fiber	0.6 g	Potassium	166 mg		
Ash	0.8 g	Ascorbic.	18 mg		

No. 60 Rose Hips

Amount of nutrient per 100g. (About ½ cup)				Calories: 54	
Water	91.3 g	Calcium	19 mg	β Carotene	300 µg
Protein	0.8 g	Phos.	18 mg	Thiamin	Tr.
Fat	0.4 g	Iron	0.6 mg	Riboflavin	Tr.
Carbos.	7.0 g	Sodium	1.2 mg	Niacin	400 µg
Fiber	1.1 g	Potassium	178 mg		
Ash	0.5 g	Ascorbic.	23 mg		

No. 61 Saguaro

Amount of nutrient per 100g. (About ½ cup)				Calories: 122	
Water	37.5 g	Calcium	43 mg	β Carotene	634 µg
Protein	13.1 g	Phos.	41 mg	Thiamin	95 µg
Fat	1.0 g	Iron	1.4 mg	Riboflavin	97 µg
Carbos.	41.5 g	Sodium	2 mg	Niacin	906 µg
Fiber	13.4 g	Potassium	403 mg		
Ash	6.9 g	Ascorbic.	52 mg		

No. 62 Saguaro Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 528	
Water	4.8 g	Calcium	95 mg	β Carotene	0 µg
Protein	15.5 g	Phos.	305 mg	Thiamin	450 µg
Fat	29.1 g	Iron	3.4 mg	Riboflavin	200 µg
Carbos.	47.7 g	Sodium	4 mg	Niacin	3800 µg
Fiber	2.6 g	Potassium	339 mg		
Ash	2.9 g	Ascorbic.	0 mg		

No. 63 Sheep Sorrel (Wood Sorrel) Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 28	
Water	90.0 g	Calcium	66 mg	β Carotene	7740 µg
Protein	2.1 g	Phos.	41 mg	Thiamin	75 µg
Fat	0.3 g	Iron	1.6 mg	Riboflavin	100 µg
Carbos.	5.6 g	Sodium	5 mg	Niacin	465 µg
Fiber	0.8 g	Potassium	338 mg		
Ash	1.1 g	Ascorbic.	119 mg		

No. 64 Shepherd's Purse Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 33	
Water	88.2 g	Calcium	208 mg	β Carotene	2590 µg
Protein	4.2 g	Phos.	86 mg	Thiamin	250 µg
Fat	0.5 g	Iron	4.8 mg	Riboflavin	170 µg
Carbos.	5.2 g	Sodium	40 mg	Niacin	400 µg
Fiber	1.2 g	Potassium	394 mg		
Ash	1.9 g	Ascorbic.	36 mg		

No. 65 Sow Thistle Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 31	
Water	31 g	Calcium	146 mg	β Carotene	4500 µg
Protein	0.5 g	Phos.	44 mg	Thiamin	100 µg
Fat	0.8 g	Iron	2 mg	Riboflavin	200 µg
Carbos.	6.6 g	Sodium	32 mg	Niacin	700 µg
Fiber	2.0 g	Potassium	299 mg		
Ash	1.5 g	Ascorbic.	13 mg		

No. 66 Sow Thistle Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 20	
Water	92.0 g	Calcium	93 mg	β Carotene	116 µg
Protein	2.4 g	Phos.	35 mg	Thiamin	150 µg
Fat	0.5 g	Iron	8 mg	Riboflavin	190 µg
Carbos.	4.2 g	Sodium	4 mg	Niacin	1000 µg
Fiber	0.6 g	Potassium	67 mg		
Ash	0.9 g	Ascorbic.	32 mg		

No. 67 Strawberry Fruit

Amount of nutrient per 100g. (About ½ cup)				Calories: 37	
Water	89.9 g	Calcium	21 mg	β Carotene	36 µg
Protein	0.7 g	Phos.	21 mg	Thiamin	30 µg
Fat	0.5 g	Iron	1 mg	Riboflavin	70 µg
Carbos.	8.4 g	Sodium	1 mg	Niacin	600 µg
Fiber	1.3 g	Potassium	164 mg		
Ash	0.5 g	Ascorbic.	59 mg		

No. 68 Sumac Fruit

Amount of nutrient per 100g. (About ½ cup)				Calories: 230	
Water	4.8 g	Calcium	93 mg	β Carotene	1700 µg
Protein	11.9 g	Phos.	57 mg	Thiamin	300 µg
Fat	16.8 g	Iron	10 mg	Riboflavin	300 µg
Carbos.	63.2 g	Sodium	95 mg	Niacin	2500 µg
Fiber	27.8 g	Potassium	980 mg		
Ash	3.2 g	Ascorbic.	143 mg		

No. 69 Sunflower Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 560	
Water	4.8 g	Calcium	120 mg	β Carotene	30 µg
Protein	24.0 g	Phos.	837 mg	Thiamin	1960 µg
Fat	47.3 g	Iron	7 mg	Riboflavin	230 µg
Carbos.	19.9 g	Sodium	30 mg	Niacin	5400 µg
Fiber	3.8 g	Potassium	920 mg		
Ash	4.0 g	Ascorbic.	0 mg		

No. 70 Thistle Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 63	
Water	80.9 g	Calcium	20 mg	β Carotene	0 µg
Protein	3.0 g	Phos.	79 mg	Thiamin	Tr.
Fat	0.1 g	Iron	8.5 mg	Riboflavin	Tr.
Carbos.	15.3 g	Sodium	65 mg	Niacin	200 µg
Fiber	1.5 g	Potassium	410 mg		
Ash	0.7 g	Ascorbic.	0 mg		

No. 71 Thyme Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 60	
Water	80.0 g	Calcium	27 mg	β Carotene	400 µg
Protein	2.0 g	Phos.	44 mg	Thiamin	100 µg
Fat	1.6 g	Iron	27 mg	Riboflavin	100 µg
Carbos.	13.9 g	Sodium	12 mg	Niacin	1100 µg
Fiber	0.4 g	Potassium	177 mg		
Ash	2.6 g	Ascorbic.	113 mg		

No. 72 Tumbleweed Seeds

Amount of nutrient per 100g. (About ½ cup)				Calories: 369	
Water	4.2 g	Calcium	830 mg	β Carotene	0 µg
Protein	27.4 g	Phos.	540 mg	Thiamin	500 µg
Fat	14.4 g	Iron	143 mg	Riboflavin	200 µg
Carbos.	46.3 g	Sodium	11 mg	Niacin	4100 µg
Fiber	30.2 g	Potassium	1140 mg		
Ash	7.7 g	Ascorbic.	0 mg		

No. 73 Tumbleweed Shoots

Amount of nutrient per 100g. (About ½ cup)				Calories: 113	
Water	60.6 g	Calcium	973 mg	β Carotene	3500 µg
Protein	4.8 g	Phos.	67 mg	Thiamin	100 µg
Fat	0.7 g	Iron	5 mg	Riboflavin	200 µg
Carbos.	27.9 g	Sodium	4 mg	Niacin	1000 µg
Fiber	12.5 g	Potassium	2545 mg		
Ash	6.0 g	Ascorbic.	40 mg		

No. 74 Violet Flowers

Amount of nutrient per 100g. (About ½ cup)				Calories: 16	
Water	94.8 g	Calcium	47 mg	β Carotene	400 µg
Protein	1.4 g	Phos.	86 mg	Thiamin	20 µg
Fat	0.3 g	Iron	1 mg	Riboflavin	110 µg
Carbos.	2.7 g	Sodium	18 mg	Niacin	600 µg
Fiber	0.6 g	Potassium	166 mg		
Ash	0.8 g	Ascorbic.	18 mg		

No. 75 Wild Lettuce Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 18	
Water	94.0 g	Calcium	68 mg	β Carotene	1000 µg
Protein	1.3 g	Phos.	25 mg	Thiamin	60 µg
Fat	0.3 g	Iron	1.4 mg	Riboflavin	60 µg
Carbos.	3.5 g	Sodium	9 mg	Niacin	400 µg
Fiber	0.7 g	Potassium	264 mg		
Ash	0.9 g	Ascorbic.	18 mg		

No. 76 Willow Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 35	
Water	90.0 g	Calcium	92 mg	β Carotene	5652 µg
Protein	1.6 g	Phos.	25 mg	Thiamin	110 µg
Fat	0.3 g	Iron	2.7 mg	Riboflavin	240 µg
Carbos.	7.3 g	Sodium	40 mg	Niacin	900 µg
Fiber	1.6 g	Potassium	382 mg		
Ash	0.8 g	Ascorbic.	68 mg		

No. 77 Wintergreen Fruit

Amount of nutrient per 100g. (About ½ cup)				Calories: 68	
Water	81.7 g	Calcium	69 mg	β Carotene	280 µg
Protein	2.4 g	Phos.	18 mg	Thiamin	42 µg
Fat	0.9 g	Iron	0.6 mg	Riboflavin	43 µg
Carbos.	14.4 g	Sodium	1 mg	Niacin	400 µg
Fiber	1.3 g	Potassium	178 mg		
Ash	0.5 g	Ascorbic.	89 mg		

No. 78 Yarrow Leaves

Amount of nutrient per 100g. (About ½ cup)				Calories: 103	
Water	65.4 g	Calcium	460 mg	β Carotene	Tr.
Protein	4.5 g	Phos.	125 mg	Thiamin	Tr.
Fat	1.2 g	Iron	Tr.	Riboflavin	Tr.
Carbos.	25.2 g	Sodium	3 mg	Niacin	Tr.
Fiber	5.9 g	Potassium	616 mg		
Ash	3.7 g	Ascorbic.	23 mg		

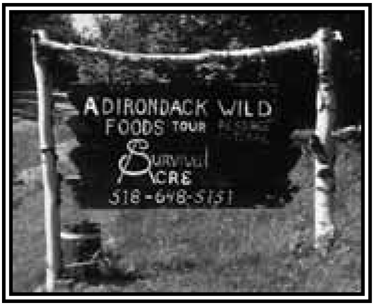

The first "A Survival Acre" at Linda's home wild food tours.

Burke's Cottages Sabael, NY wild food walk.

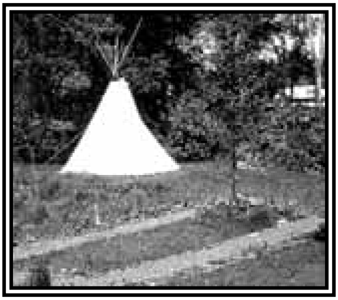

Runyon Institute, Warrensburg, NY wild food walk.

Part III: Wild Food Recipes

These recipes incorporate wild foods as much as possible. As you will see, wild foods can be used in all types of dishes including those that use meat, fish, and dairy foods. Of course, dishes made entirely of wild foods are *my* environmentarian preference. Those included here use meats, fish, dairy foods, and other "nonwild" foods and are adapted from various cultures.

The important thing is to *try* wild foods. Fit them into your diet occasionally at first, increasing to a daily basis. Supplementing your meals with wild food saves money. Most of the recipes may seem unfamiliar, but dare to use wild foods in a creative way and you will be pleasantly surprised!

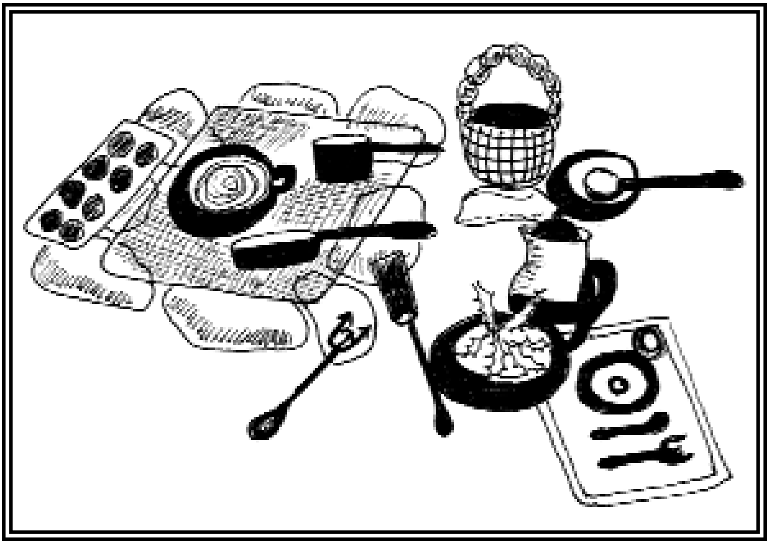

A typical campfire meal: Food simmering on edge, drying over hot coals, casserole on back, hot cookies or biscuits on back rock.

Some suggestions for substitutions:

1. Anise flavor: Goldenrod leaves, flowers, and roots.
2. Asparagus: Milkweed shoots, fireweed tops.
3. Bran: Bulrush, crabgrass.
4. Broccoli: Milkweed buds and flowers, wintercress, mustard flowers.
5. Cabbage: Shepherd's purse, dandelion, chicory, and wild lettuce.
6. Carrot: Queen Anne's lace roots, evening primrose roots, thistle roots, Shepherd's purse roots, and purslane roots.
7. Coffee: Chicory, dandelion roots, yarrow, clover buds.
8. Corn on the cob: Green cattail heads.
9. Cucumbers: cattail pith, fireweed pith, sunflower pith.
10. Endive: Dock.
11. Fruit: Blackberries, blueberries, dandelion flowers, mullein flowers, evening primrose flowers, raspberries, roses, strawberries, and violets.
12. Granola: Phragmities, amaranth, lamb's quarters, clover.
13. Gum: Sap from balsam fir or pine, milk of wild lettuce, dandelion sap, milkweed sap.
14. Lemonade: Sumac, wood sorrel, and sheep sorrel.
15. Lettuce: Sow thistle, dandelion, chicory, wild lettuce, milk thistle, amaranth, malva neglecta, and lamb's quarters leaves.
16. Nuts: Pine seeds, seeds from malva neglecta.
17. Okra: Milkweed shoots, flowers, pods, small leaves.
18. Orange juice: Pine tea, sumac tea, and balsam fir tea.
19. Pepper: Thyme leaves and flowers.
20. Potatoes: Arrowhead, ground pine nuts.
21. Salt: Queen Anne's lace seeds, ground dried clover with dried garlic.
22. Sugar: Dried meadowsweet flowers.
23. Turnip greens: Plantain leaves, sow thistle leaves.

Some medicinal and cosmetic substitutes.

1. Aspirin: Willow.
2. Lampwick: Mullein leaf, dried and cut in strips.
3. Mothballs: Dried white clover.
4. Sedative: Chamomile, mullein.

Some Wild Food Menus

During the Runyon Institute days, students from Rutgers University experimented with recipes. Some recipes are from homestead days while others are shared with us by students of wild food classes. Still others are delectable recipes from gourmet cooks. To help clarify which recipes fit these groups please see the following symbols listed below.

Food Group Symbols

***** From Institute students
! Homestead days
x Given to us by Arizona students and individual people over the last ten years.
~ Gourmet cooks

Because of my vegetarian habits, the following recipes do not use meat or fish. Wild food recipes can be combined in a variety of ways and still fit the traditional meals of breakfast, lunch and dinner. Here are some menu suggestions to get you started on the road to an environmentarian diet.

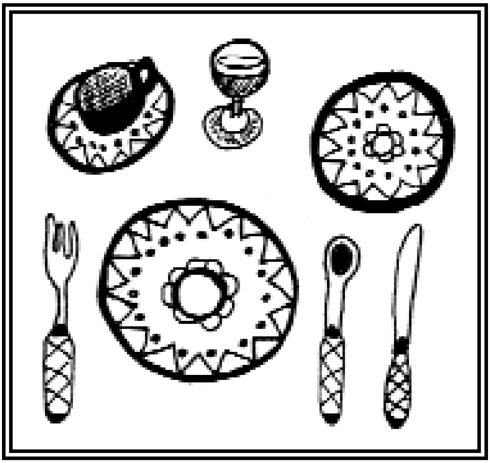

Soups & Salads

Burdock Stem Soup*

2 cups 1-inch sections of burdock stems, peeled and sliced
1 quart water
2 tablespoons lemon juice or vinegar
pinch of thyme pepper, Queen Anne’s lace seeds to taste

Place burdock stems in a medium pot and cover with 2 cups water. Bring to a boil and simmer for 5 minutes. Drain. Add remaining 2 cups water, thyme pepper, and seeds. Bring to a boil and simmer for 5 minutes. Season with lemon juice or vinegar and serve hot.

Serves 4

Slow Cooker Filarie Soup x

1 large filarie head and root
1 medium onion, thinly sliced
1 small potato, peeled and diced (can substitute parsnips or Queen Anne’s lace root)

Wash the filarie head well, pulling out the buds and fruit. Scrub the root with a toothbrush. Place the filarie in a slow cooker and cover with water. Add the onion and potato, cover with water and simmer on high for 1 ½ hours. Serve hot.

Serves 4

Clover Soup !

1 1/2 cups chopped onions
6 cups water
3 cups clover leaves and flowers washed
4 teaspoons tamari or soy sauce
1 teaspoon wild thyme
salt or Queen Anne’s lace seeds to taste
3 cups cooked brown rice (optional)

Simmer the onions in the water for 20 minutes. Add the remaining ingredients, except rice, and simmer 10 minutes. Eat as broth or stir in the brown rice and cook until the rice is done.

Variation: Add 2 cups chopped dandelion root and 1 cup sliced carrots to the water and simmer with the onion.

Serves 6

Lamb's Quarters Fat Hen Soup x

1 pound lamb's quarters leaves and seeds (6 to 8 cups), washed
1 large onion, chopped
1 teaspoon salt or Queen Anne's lace seeds
1/4 teaspoon freshly ground black or thyme pepper
1/4 teaspoon freshly grated nutmeg
1 garlic clove, crushed
2 tablespoons butter or olive oil
2 teaspoons whole wheat flour
1/2 cup milk or water
1/2 cup sour cream or water
2 teaspoons pine nuts

Boil the lamb's quarters in water to cover until tender, about 5 minutes. Add onion, salt or seeds, pepper, and nutmeg. Cover and simmer 10 minutes. Remove from the heat and cool.

Put the mixture in a blender and puree to a smooth, deep green. Add the garlic and blend again. Melt the butter in a medium saucepan over moderate heat. Stir in the flour and cook 1/2 minute. Add the milk, then sour cream or water to make a white sauce. Stir to blend, then serve lamb's quarters into bowls, top with white sauce and garnish with pine nuts.

Serves 6

Delicate Mallow Soup!

1 cup malva neglecta leaves, washed
2 cups water
1 small onion, thinly sliced
1/4 cup malva neglecta seeds

Combine the leaves and water in a medium pot. Bring to a boil over medium heat, then turn off heat, stir and cover. Cool for 10 minutes.

Strain the leaves if you wish, then add the onion and seeds. Serve hot.

Serves 2

Simple Nettle Soup!

3 cups nettles, washed and sliced (use any part of whole plant: leaves, flowers, seeds, or stems)
4 cups water
2 tablespoons butter
3 garlic cloves, crushed
salt or Queen Anne's lace seeds to taste

Combine all ingredients in a large pot and simmer for 20 minutes. Serve hot with whole-grain bread.

Serves 4

Nettle Chowder x

4 cups water
1 1/2 cups diced potatoes
1 cup sliced carrots (wild if available)
2 celery stalks, sliced
1 cup chopped onion
3 garlic cloves, crushed
1/2 teaspoon fresh or dried marjoram
1 teaspoon dried or fresh thyme leaves
Salt or Queen Anne's lace seeds to taste
4 cups nettles, washed and sliced (any part of whole plant: leaves, flowers, seeds, or stems)
2 cups soymilk

Combine the water, potatoes, carrots, celery, onion, herbs, and seeds in a large pot. Simmer 30 minutes, then add nettles and simmer another 20 minutes. Add soymilk and heat gently, but do not boil. Serve hot!

Serves 6

Sesame Thistle Soup*

3 tablespoons butter
1 cup chopped thistle root (any species)
3 cups water
2/3 cup sesame seeds
4 garlic cloves, crushed
1 cup dandelion flowers
Queen Anne’s lace seeds or salt to taste
3 cups cooked brown rice

Melt the butter in a medium skillet. Sauté the thistle, sesame seeds, and garlic until the seeds are golden brown, about 2 minutes. Add the dandelion and cook another 5 minutes. Add the seeds or salt to taste, then serve over brown rice.

Serves 6

Milk Thistle Soup with Yellow Squash x

6-8 milk thistle leaves
1 cup cooked, pureed yellow squash
garlic bud (optional)

Place milk thistle leaves in water and simmer until soft and pliable (5-10 minutes). Puree in a blender. Place milk thistle puree and squash puree in two separate pots and heat. Pour in shallow soup dish, one side at a time so you have a dark half and a light half (ying and yang) in your bowl. Press garlic bud and sprinkle over the top.) Serve hot!

Serves 2

Sow Thistle Soup x

1 cup sow thistle tips
1 clove garlic
1 onion, thinly sliced

Place sow thistle tips in water to cover. Simmer 20 minutes on low heat. (Crock pot is excellent; simmer for 1 hour on low.) Add garlic, onion, and serve hot.

Serves 2

Tumbleweed Soup x

1 cup tumbleweed tips
1 clove garlic
1 thinly sliced onion

Place tips in water to cover. Simmer 20 minutes on low heat. Or use a crock pot for more flavor and cook at low heat for 1 hour. Add pressed oil from garlic and garnish with onion.

Serves 2

Greens Soup *

1 cup fresh sorrel leaves, flowers, stems
1 thinly sliced onion

Place sorrel in a pot and cover with water. Add onion and simmer for a few minutes. Leaves will turn from bright green to olive green. This is normal and ensures the oxalic acid crystals are destroyed. Serve hot. Delicious!

Serves 2

Creamed Sorrel Soup ~

4 cups washed sorrel leaves
2 teaspoons safflower oil or olive oil
2 teaspoons wheat germ
1 onion, chopped
4 cups non-fat dried milk, or goat's milk, or water

Place leaves in pot, cover with water and simmer slowly for 1/2 hour. Blend in remaining ingredients, simmer and strain.

Serves 4

Wild Lettuce Salad *

1 cup wild lettuce leaves, washed
1/2 cup wild lettuce buds and flowers, washed
1/2 cup shredded red cabbage
1 garlic clove, crushed
1 teaspoon wild thyme leaves fresh or dried
3 tablespoons oil, 3 tablespoons vinegar of choice

Combine all ingredients, toss gently, and serve.

Serves 4

Fresh Dandelion Salad !

4 cups fresh dandelion greens
clove of garlic, chopped, or garlic salt
pinch of thyme pepper

Drain the dandelion greens. Rub the garlic on a salad bowl. Add chopped garlic pieces to bowl, and place drained leaves in bowl. Add your favorite dressing.

Serves 3

Hot Dandelion Salad *

4 handfuls fresh (young, new) dandelion leaves
sesame oil or olive oil
thyme pepper

Wash dandelions. If desired, soak bitter leaves in salted water for about 1 hour then wash and drain. Place the leaves in a frying pan and add the oil. Heat slowly to a moderate heat while stirring the leaves until they are wilted. Remove, serve hot!

Serves 4

Daisy Petal Salad x

1 quart freshly picked daisy petals

Wash petals, place in salad bowl and add your favorite dressing. Or put petals in pita sandwiches with favorite garnish.

Serves 4-6

Spring Wild Coleslaw x

1/2 cup dandelion
1/2 cup plantain
1/2 cup clover
1/2 cup chickweed
1/4 cup curley dock
1/2 small peeled onion
2 tablespoons oil and vinegar

Wash all ingredients well and drain in a sieve. Process all parts plus onion in a processor until completely "slawed". Add oil and vinegar (or your favorite dressing). Serve as a side dish.

Chickweed Salad x

1 clove garlic
1 quart loosely packed chickweed, washed
2 eggs, hard-cooked
4 tablespoons favorite dressing

Peel clove of garlic and rub wooden salad bowl. Chop chickweed finely and add to salad bowl. Peel eggs, slice thin and add to bowl. Toss together with favorite salad dressing.

Serves 4

Star Chickweed Dressing x

2 cups star chickweed, washed
3/4 cup plain yogurt
3/4 cup mayonnaise
1/2 tablespoon garlic powder or 4 cloves fresh
1 tablespoon dried dill
salt to taste

Combine all ingredients in a blender or food processor. Blend on medium speed, then refrigerate for 30 minutes. Serve on your favorite salad or casserole. Sandwiches also taste delicious spread with this dressing.

Makes 2 cups of dressing

Evening Primrose Olive Oil*

1 cup evening primrose flowers, washed and dried
approx. 1\2 cup extra virgin olive oil (sunflower,
  safflower, or corn oil can be substituted)

Fill a 10-ounce jar to capacity with evening primrose flowers. Add oil at room temperature or boiling to cover. (Be careful with hot oil in glass jars.) Repeatedly press down the flowers until all the air bubbles are out of the mixture. Cap and shake every day for or 1 week for sunflower, safflower, or corn oil, or 2 weeks for olive oil. Remove the flowers and discard. Refrigerate oil if you used sunflower, safflower, or corn oil. Use for salads.

Makes approximately 10 ounces

Vegetable Dishes & Light Meals

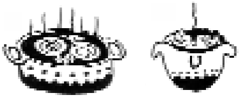

Linguine with Amaranth Cream Sauce~

2 tablespoons butter
4 garlic cloves, crushed
3 cups soymilk or cow's milk
4 tablespoons rice or wheat flour
1 cup crushed dried young amaranth
1 pound linguine, cooked until al dente

In a large frying pan, melt the butter and sauté the garlic over low heat for about 10 minutes. In a saucepan, mix the milk and flour to make a thin paste. Add the flour mixture and amaranth and heat until smooth and thickened, about 3 minutes. Serve over the linguine.

Serves 2 to 3 as a main course

Arrowhead Stew!

1 large onion, sliced
1 cup washed roots: thistle, dandelion, evening primrose, or filarie
12 to 14 arrowhead tubers, washed
1/2 teaspoon Queen Anne's lace seeds
1 teaspoon dried thyme leaves

Combine all ingredients in a slow cooker and add water to cover. Cook on low setting for 4 hours or until tender.

Serves 4 as a main course

Clover Noodles~

1 cup clover flour (See page 78)
3 to 5 cups unbleached white flour
5 eggs
1 tablespoon olive oil
1/2 teaspoon salt or Queen Anne's lace seeds

Put clover flour into bowl and add enough water to reconstitute to a moist consistency. Stir in eggs, oil, & salt. You should have a slimy green soup. Add white flour until you make a stiff dough. Remove and kneed till it won't tear when you stretch it. Let rest 30 minutes. Cut dough ball into about seven or eight equal sized balls, and run through pasta machine. Start on setting #1 and then press through 2, 3, and 5. Add flour with each run through the machine to avoid sticking.

Dry on racks of food dryer. Let cool, store in tightly sealed glass jars. Serve cooked to al dente and topped with butter or add your favorite vegetable.

Serves 8-10

In a pasta machine, clover noodles can be made by replacing 2 tablespoons of flour with 2 tablespoons of clover flour. Adjust the liquid as necessary to get the right texture, extrude and serve.

Makes approximately 4 servings

Clover Pasta with Beans and Basil x

1 pound clover pasta
4 tablespoons olive oil
4 to 6 garlic cloves
1 cup garbanzo beans (chickpeas)
6 tablespoons water or tomato puree
1/2 teaspoon freshly ground black pepper or
  1/4 teaspoon ground dried thyme
12 large basil leaves
1 tablespoon fresh parsley, chopped
1/4 cup grated Romano cheese (optional) or
  1 teaspoon Queen Anne's lace seeds or salt

Cook the pasta until al dente, then drain and set aside. In a food processor or blender, puree olive oil and garlic. Add garbanzos and water or tomato puree and blend. Blend in the pepper, basil, parsley, and cheese or seeds and blend quickly until mixture is a coarse paste. Toss with pasta and serve.

Serves 4 as a main course

Burdock Bur Casserole*

2 cups young burdock burs and leaves
1 teaspoon thyme leaves
1 teaspoon Queen Anne's lace seeds
1 cup nonfat dry milk or water
4 tablespoons butter or 2 teaspoons safflower, olive oil
1 cup whole wheat bread crumbs

Preheat the oven to 350 degrees. Wash the burs & leaves and place in a large saucepan. Add water to cover and simmer for 5 minutes. Drain.

Add the thyme and seeds, cover with water and simmer for 5 another minutes. Drain and save the water for soup stock.

Put the burs in a greased 1 1/2-quart casserole dish. Add the milk or liquid and oil. Top with bread crumbs and bake for 20 minutes.

Serves 4 as a main vegetable casserole

Cattail "Corn On The Cob" !

2 to 3 cattail heads
safflower, olive oil, or butter
salt or Queen Anne's lace seeds
thyme pepper to taste

Put the cattail in a large pot of boiling water and boil for 7 minutes. Remove and serve with oil or butter, salt or seeds, and thyme.

Serves 1 as a side dish

Note: When cutting, leave 2 inches of stem for holding the cob

Cattail Casserole!

2 cups cattail fluff, scraped from brown buds
1 cup bread crumbs
1 egg, beaten (optional)
2 to 3 teaspoons water
Queen Anne's lace seeds or salt to taste
1/4 teaspoon thyme leaves, fresh

Preheat the oven to 325 degrees. Combine all ingredients in a greased 1 1/2-quart casserole dish and bake for 25 minutes. Serve hot.

Serves 4 as a main dish

Chickweed Pita Pockets x

2 cups chickweed, washed
1/2 cup finely chopped tender lamb's quarters leaves, washed
1/2 cup sliced carrot
1/2 cup finely diced cucumber
1 tablespoon salad dressing of your choice
2 pita breads

Toss together the chickweed, lamb's quarters, carrot, cucumber, and dressing. Fill the pita pockets and serve.

Makes 2 sandwiches

Chickweed Mediterranean x

2 tablespoons olive oil
1 garlic clove, finely chopped
2 cups whole chickweed plant without roots, washed
1 teaspoon thyme leaves

Heat the olive oil in a medium frying pan. Sauté the garlic until browned, then add the chickweed and thyme. Stir-fry until piping hot, about 10 minutes.

Serves 3 as a side dish

Hot Clover and Rice with Butter Sauce x

4 cups cooked brown rice
4 cups washed clover leaves
1 cup milk or water
1/4 pound butter or soy butter
1 teaspoon finely minced garlic, or 2 teaspoons garlic powder

Preheat the oven to 350 degrees. Grease a 2-quart casserole dish. In the casserole, combine the rice, clover, and liquid.

Melt butter in a saucepan over low heat. Add garlic and cook for 3 minutes. Stir sauce into casserole and bake for 20 minutes until piping hot.

Serves 4 as a main course

Dandelion Greens Casserole !

2 cups dandelion leaves, washed
1 cup whole wheat bread crumbs
2 cups water

Preheat the oven to 350 degrees and grease a 1-quart casserole dish. Put the leaves in a 1-quart saucepan and add water to cover. Simmer gently for 20 minutes, then drain, reserving the liquid. Chop the leaves fine and pour into the casserole dish. Add the reserved liquid and top with the bread crumbs. Bake for 25 minutes or until brown on top.

Serves 3 to 4 as a main course

Southern Italian Dandelion Leaves ~

4 cups young 2 to 3-inch dandelion leaves, washed
1 medium potato or arrowhead tubers, peeled
2 teaspoons safflower or sesame oil
1 medium onion, cubed
1 large tomato or 1 8-ounce can whole tomatoes, chopped

Combine the dandelion leaves and potato in a medium pot and cover with water. Bring to a boil and simmer for 15 minutes. Remove the potato or arrowhead, then cube and set aside. Drain the leaves, squeezing out the water. (Save the liquid for soup stock.)

Heat the oil in a large frying pan and sauté the onion slowly until limp, about 5 minutes. Combine the dandelion leaves, potato or arrowhead, and tomato, and bring to a boil. Remove from the heat and stir.

Serves 4 to 6 as a main course

Dandelion Stir Fry ~

3 tablespoons olive oil
1 cup chopped onion
4 garlic cloves, crushed
1/4 cup chopped dandelion root, scrubbed
2 cups fresh dandelion leaves, washed
2 cups dandelion flowers, washed
1 teaspoon dried thyme or 1 sprig fresh thyme
1/2 teaspoon dried marjoram leaves or 1 fresh sprig
2 teaspoons sesame salt
2 tablespoons cider vinegar
2 cups cooked brown rice

Heat half the oil in a large wok or frying pan. Add the onion, garlic, and dandelion roots. Sauté for 5 minutes or until soft. Remove from the wok and set aside in a warm place.

Heat the remaining oil in the wok and add the dandelion leaves and flowers, thyme, marjoram, salt, and vinegar. Stir-fry for 10 minutes. Stir in the onion mixture and serve over brown rice.

Serves 6 as a main course

Dandelion Roots *

1/2 cup washed and peeled white dandelion roots,
    chopped into French fry sized pieces
1 cup favorite pancake butter
2 tablespoons safflower oil
1 chopped onion

Dip the roots in the pancake batter. French fry in hot safflower oil until brown. (A wok works perfectly.) Add the chopped onion as a garnish and serve.

Serves 2 as a side dish

Chamomile Vegetable !

Chop a whole chamomile plant into bite-sized pieces and simmer gently for 6-8 minutes. Serve hot with favorite garnish. A delicious broccoli substitute. Servings vary depending on plant size.

Tofu Scallopine with Dandelion Wine ~

1 pound firm tofu, weighted and drained for at least 1 hour
whole wheat flour
3 tablespoons olive oil
1 bay leaf
1/2 cup Dandelion Wine (Page 246) or marsala
10 large fresh mushrooms, sliced
1 large garlic clove, finely minced
1 teaspoon chopped chives
1 teaspoon chopped fresh parsley
juice of 1/2 lemon
1 teaspoon Worcestershire sauce
Queen Anne's lace seeds or salt to taste
Ground thyme or black pepper to taste
6 lemon wedges for garnish

Slice the tofu into 12 equal portions and coat each piece with flour. Heat the oil in a large frying pan and sauté the tofu with the bay leaf until tofu is light brown on both sides- about 6 minutes. Remove the tofu and set aside.

Add the wine and reduce the liquid by one-half. Add the mushrooms, garlic, chives, parsley, lemon juice, Worcestershire sauce, seeds or salt, and thyme or pepper. Cook for 3 minutes.

Add the tofu slices and simmer for 5 minutes, adding additional wine or water if necessary to keep mixture moist. Serve the tofu slices with pan juices and garnish with a lemon wedge and parsley

Serves 6 as a main course

Rosemary Tofu x

3 tablespoons butter or soy butter
1 pound firm tofu, weighted, drained and cubed
4 garlic cloves, crushed
1 1/2 teaspoons dried rosemary leaves
2 cups diced young lamb's quarters or amaranth leaves, washed
3 cups cooked brown rice

Melt the butter in a large frying pan. Sauté the tofu, garlic, and rosemary for 20 minutes, stirring often. During the last 5 minutes of cooking, add the leaves and sauté until tender. Serve hot over brown rice.

Serves 6 as a main course

Evening Primrose Stew !

5 to 6 Evening primrose roots
1 medium potato, halved
1 medium onion, quartered
1 teaspoon thyme leaves dried or fresh
1 teaspoon Queen Anne’s lace seeds

Wash the roots and scrub with a toothbrush, trimming off any bruised parts. Combine the roots, potato, and onion in a slow cooker and add water to cover. Cook on low for 4 to 6 hours, then serve while hot.

Serves 2 as a main course

Dock Casserole *

2 large bunches dock, including leaves
4 to 6 tablespoons butter or safflower oil

Preheat oven to 350 degrees. Wash and drain the dock and steam in a pot until all parts are wilted, about 2 minutes. Discard the tougher stems and finely chop the leaves and short stems. Steam the chopped mixture for about 15 minutes more, stirring frequently. Pour the mixture into a 1-quart casserole dish, dot with butter or drizzle with oil, and bake for 20 minutes or until bubbly hot.

Serves 4-6 as a main course

Lemon Fireweed x

2 cups fireweed tips, washed
1 lemon wedge
4 to 5 thyme sprigs

Steam the fireweed tips until tender, about 2 minutes. Squeeze the lemon over the tips, then sprinkle on the thyme.

Serves 4 as a side dish

Filarie Sandwich x

1 handful filarie leaves, washed
pinch dried thyme leaves
1/4 teaspoon each of oil and vinegar
1 Bermuda onion, thinly sliced into rings
2 slices whole wheat bread

Layer the ingredients between the bread slices for a “desert sandwich” treat.

Makes 1 sandwich

Fireweed Casserole *

4 cups fireweed buds, washed
1 medium onion, chopped, or 1 cup chopped chives
bread crumbs

Preheat the oven to 350 degrees and lightly grease a 1 1/2-quart casserole dish. Mix the buds and onion or chives and place in the casserole dish. Add water to cover. Bake for 30 minutes or until bubbly. Top with bread crumbs.

Serves 4

Wild Sow Thistle Stir-Fry !

1 cup wild sow thistle leaves
1/2 cup wild sow thistle buds and flowers
1/2 cup red cabbage
1 clove garlic, pressed
1 teaspoon wild thyme
3 tablespoons oil and vinegar

Combine all ingredients in a salad bowl, toss gently and serve.

Serves 2

Aloe Vera Flowers x

Eat flowers raw and add to favorite sandwiches, salads. Dip flowers in favorite sauces.

Wild Greens Quiche ~

Crust

3 cups cooked brown rice
1 cup sesame seeds
2 tablespoons butter
dash of salt

Filling

1 tablespoon butter
1 medium onion, chopped
8 ounces fresh mushrooms, sliced
4 garlic cloves, crushed
3 cups mixed wild greens (lamb's quarters, amaranth, clover, dandelion, curly dock, or plantain), washed
1/2 teaspoon salt
3 eggs, beaten
1 cup milk
1 teaspoon marjoram leaves
1 teaspoon chopped basil leaves
1/2 teaspoon thyme leaves

Preheat the oven to 350 degrees.

Mix the brown rice, sesame seeds, butter, and salt in a bowl and press into a 12-inch pie dish. Set aside. Melt the butter in a large skillet. Add the onion, mushrooms, garlic, greens, and salt and sauté until tender, about 10 minutes. Set aside.

Combine the eggs, milk, marjoram, basil, and thyme and sauté until tender, about 10 minutes. Combine with greens and pour into shell. Smooth top and bake for 45 minutes, or until brown and crispy on the top.

Serves 8 as a main course

Baked Purslane with Yogurt x

1 cup fresh purslane, washed
1 pint yogurt

Preheat the oven to 350 degrees. Mix the purslane and yogurt and pour into a greased 1-quart casserole. Bake 10 to 15 minutes or until hot.

Serves 2 as a side dish

Tabouli with Wild Greens ~

1 cup young dandelion leaves
1 cup daisy leaves
1 cup red clover leaves
1 cup chopped sheep sorrel leaves
1/2 cup violet leaves
1/4 cup chopped spearmint leaves
1/4 cup chopped fresh oregano, or 2 teaspoons dried
1 1/2 cups cooked brown rice
1 1/2 cups cooked bulgur
oil and vinegar to taste
1 cup daisy petals
1 cup violet and strawberry flowers or wild roses

Wash and drain all greens and tear into bite-size pieces. Place in a large salad bowl. Toss in the rice, bulgur, oil, vinegar. Arrange the daisies around the edge of the bowl and place the violets and strawberries or roses in the middle. Chill and serve.

Serves 6 to 8

Lamb’s Quarters Calzone x

Dough

1/4 cup warm water
1 tablespoon active dry yeast
4 teaspoons honey
1 cup soymilk
2 eggs
3 cups unbleached all-purpose flour
1/4 cup lamb’s quarters flour
1 1/2 teaspoons sea salt

Filling

1 cup pizza sauce
1 teaspoon chopped garlic
1 medium onion, chopped
1 cup chopped fresh mushrooms
1 green bell pepper, chopped
1/2 cup chopped ripe olives
1/4 cup chopped lamb’s quarters
12 ounces fresh mozzarella cheese, shredded
3 tablespoons olive oil

Put the warm water into a small bowl and stir in the yeast and honey; let the mixture stand until foamy, 5 to 10 minutes. Whisk the soymilk and 1 egg into the yeast mixture.

In another large bowl, mix the flours and salt. Make a mound of flour then make a well in the middle. Gradually pour the yeast and milk mixture into the well, working the flour from the inside of the well into the liquid with your fingers. Continue until all the flour is absorbed, adding 1 tablespoon of water if needed. Knead until soft, velvety, strong, and elastic, 10 to 15 minutes. Place the dough in a lightly oiled bowl, cover with a damp cloth, and put in a warm place. Let the dough rise until doubled, about 2 hours.

Mix the ingredients for the filling and set aside.

Knead the dough briefly on a lightly floured surface. Cut the dough into 6 or 7 pieces. Shape each piece into a ball, flatten it hard with your hand, and roll it into a 1/4-inch thick oval with a rolling pin. Mix the remaining 1 egg with 1 tablespoon of water and brush the edge of each shaped oval with the mixture.

Preheat oven to 450 degrees, lightly oil a baking sheet.

Place 2 tablespoons of pizza sauce in the center of each dough oval. Add a little garlic, onion, mushrooms, green pepper, olives, and lamb’s quarters to each and top with cheese. Fold over the dough to form a half-moon shape, crimping the edges with a fork. Place the calzones on the baking sheet and brush each with the oil. Bake for 20 to 25 minutes or until golden brown. Brush again with oil and serve hot.

Serves 6 to 7

Weed Pizza x

2 tablespoons butter or olive oil
3 cups mixed wild greens, such as clover, plantain,
dandelion, lamb’s quarters, washed
dash of salt
1 pound ready-made pizza dough, thawed if frozen,
toppings of choice

Preheat the oven to 425 degrees. Heat the butter or oil in a large frying pan. Add the greens and salt, and sauté for 5 minutes or until tender. Roll out pizza dough according to package instructions. Spread the greens on the crust and sprinkle with toppings. Bake for 20 minutes or until crisp.

Makes 6 slices

Malva Neglecta and Lamb’s Quarters Quiche x

6 eggs
1 tablespoon lemon pepper
1 cup chocolate flavored soymilk
1/4 cup chopped malva neglecta leaves
1/4 cup chopped lamb’s quarters
1/4 cup grated carrot
2 tablespoons chopped onion
1 teaspoon chopped garlic
1 9-inch pie shell, unbaked

Preheat the oven to 350 degrees. Process the eggs, lemon pepper, and soymilk in a blender or food processor until thoroughly blended. Stir in the malva neglecta leaves, lamb’s quarters, carrot, onion, and garlic. Pour the mixture into the pie shell and bake for 35 minutes or until a knife inserted in the center comes out clean. Serve hot.

Serves 5 to 6 as main course

Milkweed Pod Fritters *

12 milkweed pods, washed. Note: Pick the milkweed pods when they are young and not elastic, 2 to 3 inches in length.
1 to 1 1/2 quarts water
1 cup cornmeal
1 cup whole wheat flour
1/4 cup safflower or peanut oil

Place the pods in a large bowl and cover with the water. In a separate bowl, combine the cornmeal and flour. Roll the pods in this mixture.

Heat the oil in a frying pan. Fry the pods, turning them over, until they are cooked through and brown on all sides, about 5 minutes. Drain well.

Serves 4 as a main dish

Milkweed Casserole *

3 cups milkweed buds, washed
2 cups milk or water
1 cup whole wheat bread crumbs

Preheat the oven to 350 degrees. Grease a 2-quart casserole. Place the milkweed buds and milk or water in casserole. Sprinkle the bread crumbs on top. Cook for 20 to 30 minutes, or until bubbly and brown on top.

Serves 6 as a main course

Young Milkweed and Pasta Medley ~

1 pound assorted pasta (rigatoni, linguine, broken into thirds, small shells, bucatino broken into thirds)
2 tablespoons olive oil
1/4 cup chopped wild onions or yellow onion
3 garlic cloves, coarsely chopped
1/4 cup dry sherry
2 cups milkweed shoots and leaf buds, blanched
ground thyme or pepper, Queen Anne's lace seeds or salt
1/4 cup chopped fresh parsley or watercress
1/4 cup grated Pecorino Romano cheese (optional)

Cook the pastas until al dente; drain and set aside. Heat the oil in a frying pan and sauté the onions and garlic until lightly browned. Add the wine and cook until the liquid is reduced by half, about 9 to 10 minutes. Add the milkweed and sauté until tender, approximately 10 minutes. Add the pastas and toss, heating thoroughly. Season to taste with thyme and seeds as salt. Garnish with parsley or watercress and sprinkle with cheese.

Serves 4 as a main course

Daisy Leaves a la Mediterranean x

2 cups washed daisy leaves
2 tablespoons olive oil
1 clove garlic, chopped finely
1 teaspoon thyme

Heat frying pan and add olive oil and chopped garlic. Brown garlic and add washed daisy leaves. Stir until piping hot.

Serves 1

Mustard Greens Casserole *

2 big bunches wintercress greens stems, leaves, and flowers
4 to 6 tablespoons safflower oil or butter

Preheat the oven to 350 degrees. Wash and drain the wintercress. Steam until wilted, about 3 minutes. Remove the tougher stems and chop the leaves & short stems finely. Steam the chopped mixture for 15 minutes more, stirring frequently. (Keep the water used in steaming for soup stock; freeze if necessary.) Pour the greens into a casserole dish, dot with oil or butter, and bake for 8 minutes or until hot and bubbly.

Serves 4 as a side dish

Stir-Fried Mustard Flowers *

1 quart mustard flower heads
1 garlic clove, chopped
1 tablespoon olive oil

In a frying pan or wok, quickly stir-fry all ingredients over high heat until limp, about 2 to 3 minutes. Serve hot.

Serves 4 as a side dish

Mustard Flower Buds

4 cups mustard flower buds
1 quart boiling water
2 tablespoons olive oil

Wash the buds thoroughly and add to boiling water. Lower heat, and simmer 5 minutes. Drain, add oil, and serve hot.

Serves 4 as a side dish

Vegetarian Burger x

2 cups cooked lentils
1/2 onion, peeled
2 tablespoons sow thistle leaves

Steam then chop thistle leaves. Add other ingredients and mix until everything sticks together.

Makes 6 medium-size burgers

Nettle-Falafel Burgers x

1 cup cooked and finely chopped nettles
1/2 cup falafel mix (see box for recipe)
2 tablespoons olive or safflower oil

Place the nettles in a medium bowl and add the falafel mix slowly. Mix thoroughly. Add a bit of water at a time until a clay-like mixture results. Pat the mixture into 3-inch patties. Heat the oil in a medium frying pan and fry the burgers slowly until brown on both sides, about 6 minutes.

Serves 2 to 3 as a main course

Boiled Mustard Greens !

1 quart young mustard leaves
1 garlic clove, chopped
1 teaspoon lemon juice
1 teaspoon vinegar
1 tablespoon soy butter or regular butter.

Boil the leaves for 30 minutes in a large saucepan. Drain well and add the garlic, lemon juice, vinegar, and butter. Serve hot.

Serves 2 as a side dish

Purslane Casserole !

1 teaspoon peanut or olive oil
4 cups purslane (including stems), washed, drained, and chopped
1 egg, beaten, or violet leaf thickener, opposite page
1/2 cup fine bread crumbs

Preheat the oven to 325 degrees and grease a 1 1/2-quart baking dish with oil. To the purslane add the egg or violet thickener and the bread crumbs. Turn into the baking dish, and add your favorite seasoning. Bake for 20 minutes or until piping hot.

Serves 4 as a side dish

Mexican-Style Purslane ~ (Vertalogas)

The following recipe was given to me by a Mexican friend. It is a common dish in Arizona.

1/2 pound purslane leaves and tender stems
1 teaspoon baking soda
1/2 pound pork chops or ribs
1/4 medium onion, chopped
1 garlic clove, chopped
1 medium tomato, chopped
1/2 teaspoon chicken broth

Place the purslane in boiling water with a pinch of baking soda and cook until tender, about 2 to 3 minutes.

In a large skillet, brown the meat on both sides and set aside. In the same pan sauté' the onion, garlic, and tomato until tender. Stir in the broth, purslane, and meat, and simmer for 15 to 20 minutes or until meat is done.

Serves 5 as a main course

Violet Leaf Thickener x

a handful of violet leaves
1/4 cup boiling water f

Press leaves down under boiling water. Do this slowly and continue for about 20 minutes. Drain. Green water is slightly gooey-or thickened. Use as water substitute. For cassaroles, layer whole violet leaves throughout casserole. Water from other ingredients will draw thickener from the edible violet leaves.

Wild Lettuce Stir-Fry x

1 cup wild lettuce leaves
1/2 cup wild lettuce buds and flowers
1/2 cup red cabbage
1 clove garlic, pressed
1 teaspoon wild thyme
3 tablespoons oil

Stir-fry all ingredients the oil, serve hot.
Serves 4

Brown Rice and Lamb's Quarter Sprouts x

1 cup lamb's quarter sprouts (sprout seeds only
1 1/2 to 2 days)
2 tablespoons corn oil or peanut oil
1 tablespoon chopped scallions or chives or onions
1/4 cup thin sliced canned or fresh mushrooms
1 cup cooked brown rice

Heat oil in heavy fry pan or wok. Add onion, mushrooms. Stir fry until lightly brown. Add soy sauce, rice and sprouts. Stir fry for 3 to 4 minutes. Cover and simmer for 5 more minutes.

Serves 3 to 4

Wild Malva Rice Roll-Ups x

1 cup water
12 large malva neglecta leaves
1 cup cooked wild rice
1/2 teaspoon thyme
1/2 teaspoon vegetable salt (of your choice)
small casserole dish, greased
1/4 cup salsa to your taste

Bring water to a boil. Wilt malva leaves by gingerly dipping each leaf separately into boiling water. Rearrange leaf on a paper towel to drain. Place wild rice in a bowl. Stir in thyme and 1/4 teaspoon vegetable salt.

Take a teaspoon of mixture and place in the middle of the leaf. Roll the leaf and mixture up to stem, then tuck stem end in to secure. Place whole roll-up face down in casserole dish. Sprinkle rest of vegetable salt over roll-up and bake in oven for 20 minutes at 350 degrees. Pour salsa over each and serve hot.

Serves 2

Malva Mexican Style

1 cup large Malva leaves
1/2 cup salsa sauce or spaghetti sauce
1 cup refried beans1 ice cube tray

Place ice cube tray on counter. Drip salsa sauce in tray as evenly as you can. Place a leaf into cube section over salsa. Take a teaspoon and place a spoonful of refried beans into the leaf. Place the ice cube tray into the freezer and freeze solid.

Pop sections out into a freezer bag with a ziplock top. When you cook a casserole, take a few ready-made sections out and place in a bean pot. Cook 350° for 20 minutes. Delicious!

Makes 1 tray full of stuffed Malva leaves

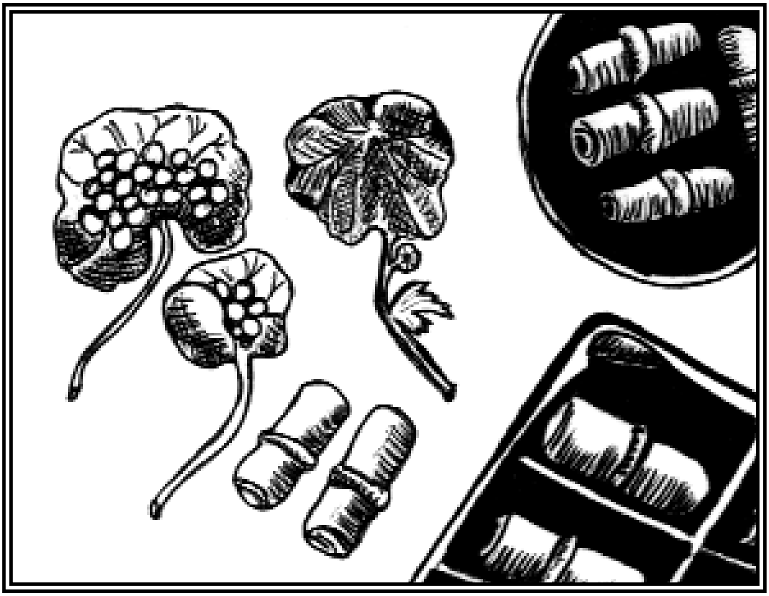

Stuffed Malva leaves. Cook freshly rolled or freeze in ice cube trays for future meals. Store in plastic bags.

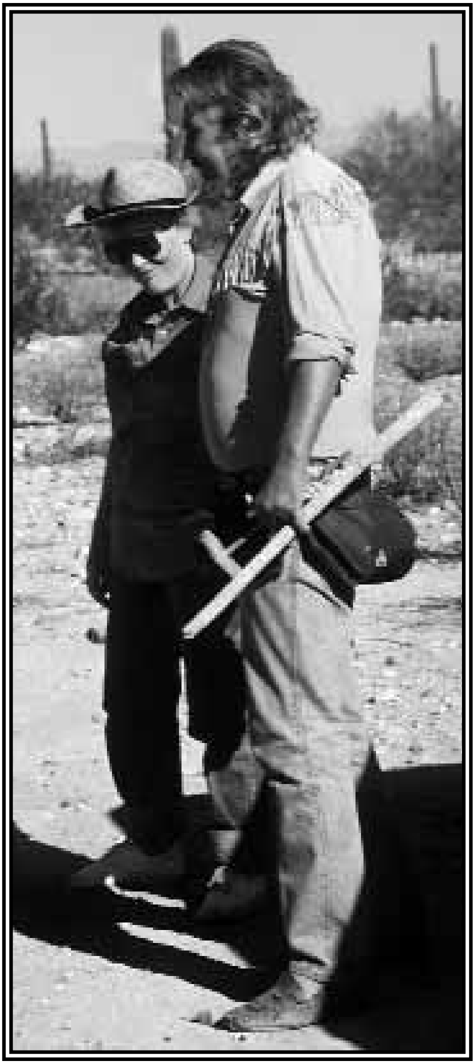

Willie Whitefeather and Linda.
Note his handmade shirt, moccasins and flute.

Muffins, Breads, Biscuits & Pancakes

Basic Mix x

4 1/4 cups whole wheat flour
4 1/4 cups amaranth or triticale flour (see Note)
5 tablespoons baking powder
1 teaspoon baking soda
1 tablespoon ground dried Queen Anne's lace seeds
2 teaspoons cream of tartar
1 1/2 cups instant dry milk
1 1/2 cups olive or peanut oil

In a large bowl, sift together the flours, baking powder, baking soda, seeds, cream of tartar, and dry milk. Blend well. With a wooden spoon, stir in the oil until evenly distributed; the mixture will be only slightly damp. Put the mix in a large, airtight container and label. Store in a cool, dry place. Use within 10 to 12 weeks.

Makes 11 cups, enough for several batches of crackers.

Note: Triticale flour is made from a high-protein grain that is a hybrid between wheat and rye. The flour is available in some natural foods stores.

Amaranth Vegetable Bread x

1 package active dry yeast
1/4 cup warm water
6 cups all-purpose flour or 10 cups whole wheat flour
4 cups cooked, drained amaranth greens

Dissolve the yeast in the warm water. Place flour in a large bowl and mix in the yeast. Add the greens and mix by hand, kneading the dough for 3 minutes. Cover and let rise 1 hour, or until doubled in bulk.

Preheat the oven to 400 degrees. Punch dough down and form into 2 loaves. Place on a baking sheet and let rise again until doubled, about 45 minutes. Bake loaves for 40 minutes, or until they sound hollow when tapped. Serve hot. These loaves freeze well.

Makes 2 loaves

Mean Green Bread x

2 tablespoons active dry yeast
3 tablespoons honey
2 1/2 to 2 3/4 cups warm water
4 cups whole wheat flour
2 cups fresh greens (dandelion, plantain, clover, lamb's quarters)
1 teaspoon salt
2 tablespoons cold-pressed safflower oil
4 to 6 cups flour mixture half lamb's quarters and half rice flour, or clover and whole wheat flour

Mix the yeast, honey, and 2 cups warm water in a large bowl and let it sit for a few minutes until foamy. Add the whole wheat flour and beat for 2 to 3 minutes. Cover the mixture and let set until doubled in size. (approximately 1 hour).

Meanwhile, in a bowl combine the greens, remaining warm water, salt, and oil. Add to the flour a little at a time and blend, kneading for about 5 minutes. Grease a large bowl and let the dough rise until doubled in size, about 50 minutes.

Punch the dough down and let rise another 40 to 50 minutes.

Preheat the oven to 400 degrees. Bake for 35 to 40 minutes or until you hear a hollow sound when you tap the top with the flat side of a butter knife. Cool for 30 minutes before serving.

Makes 1 large loaf

Wild Malva Rolls x

1 cup water
12 large malva neglecta leaves, washed
1 cup cooked wild rice
1/2 teaspoon dried thyme leaves
1/2 teaspoon salt

Preheat the oven to 350 degrees and grease a 1-quart casserole dish.

In a small saucepan, bring the water to a boil. Wilt the malva neglecta leaves by carefully dipping each leaf separately into the boiling water. Place the leaves on a paper towel to drain.

Place the rice in a bowl. Stir in the thyme and half the salt. Place 1 teaspoon of the mixture in the middle of a leaf. Fold the sides toward center and roll from stem to tip of leaf.

Place rolls seam side down in the casserole dish. Repeat, making the remaining rolls.

Sprinkle remaining salt over rolls and bake 20 minutes or until bubbly hot. Serve piping hot.

Serves 2 as a main course

Italian Clover Rolls *

1 cup clover flour
3 cups whole wheat flour
1/2 cup clover leaves
1/8 cup vegetable oil
2 garlic cloves, chopped
1 6-ounce can tomato paste
garlic powder

Preheat the oven to 350 degrees. Combine the flours in a large bowl. Add the clover leaves, oil, garlic, and tomato paste. Using your hands, knead thoroughly, adding a small amount of water until a clay-like consistency is obtained.

Roll golf-ball size pieces of dough in your palms and place on an oiled baking sheet. Bake for approximately 20 minutes or until golden brown. Serve hot, sprinkled with garlic powder.

Makes 12 to 14 rolls

Crabgrass Muffins !

1 cup enriched flour
1 cup crabgrass flour
2 teaspoon baking soda
2 teaspoons ginger (optional as a variation)
3/4 cup water
2 eggs (egg replacement used by author)
1 teaspoon pure vanilla
1/4 cup sunflower or canola oil
1/2 cup raisins
2 teaspoon baking soda

Preheat oven to 350 degrees.

Place flours and baking soda in bowl, mix in water, eggs, vanilla, & oil. Fold in raisins thoroughly. Fill muffin tins 1/2 full or pour in 8-inch square baking pan.

Bake 20 to 25 minutes. Let cool and remove from pan.

Makes 6 muffins

Crabgrass Crackers x

1 cup basic mix (Page 225)
1 cup crabgrass flour
1 teaspoon salt
1/2 cup water

Preheat the oven to 400 degrees. Lightly grease a baking sheet. In a medium bowl, combine the mix, flour, and salt. Add water to form a dough. Knead about 12 times, or until the dough is smooth. Shape into pencil-like strands 1/2 inch thick, and place on baking sheet. Roll out as a single sheet onto the baking sheet. Use a fork to perforate the sections. Bake for about 5 minutes or until crisp.

Variations: Use another flour, such as from malva, mixed grain, or lamb's quarters. Substitute goat's milk or cow's milk for water.

Makes 24 to 36 crackers

Wild Waffles ~

Dry Mix 1 cup rice flour
1 cup barley flour
1 cup buckwheat flour
1 cup old-fashioned rolled oats
1/4 cup finely chopped lamb's quarters
1 cup mixed dried greens (clover, dandelion, violet leaves, curly dock, & amaranth)
4 tablespoons baking powder
1 1/2 teaspoons salt

Combine the flours, oats, lamb's quarters, mixed greens, baking powder, and salt. Store in the refrigerator.

Makes 4 cups

Batter 2 cups dry mix
2 cups soymilk, cow's milk, or water
4 tablespoons butter or cold-pressed oil

Combine the dry mix, liquid, and butter or oil. Grease a waffle iron well. Spoon about 3/4 cup batter onto hot iron. Cook each waffle until it is brown on both sides, about 6 minutes. Serve with maple syrup or honey butter.

Serves 6 to 8

Navaho Lamb's Quarters Fry-Bread ~

1 cup whole wheat flour
1 cup lamb's quarters flour
2 teaspoons baking powder (optional)
1/2 cup reconstituted nonfat dry milk or water
2 teaspoons safflower oil
1/2 cup ice water
Oil for frying

Combine the flours and baking powder (if using). Add the milk or water and oil, and using your hand, mix into a granular mixture. Add the ice water and form into a ball. Put a towel over the bowl and let the dough rise in a warm place for 2 hours. (Even without baking powder, the dough rises about 1/2 inch.)

Cut the dough into 2 parts, roll out each part until 1/4 inch thick, then score parallel lines 1/2 inch apart almost through each piece. Heat oil in a deep pan until very hot, then add 1 piece of dough. Deep-fry until both sides are brown, approximately 2 minutes. Drain on paper towels and repeat for the other piece of dough. Serve hot. Can be frozen and reheated in a microwave.

Makes 2 breads

Mallow Vegetable Bread ~

5 cups warm water
12 to 14 cups whole wheat flour, ground fresh from
    whole wheat berries, if possible
2/3 cup olive oil
2/3 cup rice syrup
2 teaspoons sea salt
3 tablespoons active dry yeast
3 tablespoons vitamin C crystals (for dough
    consistency), vailable in health food stores.
2 medium onions, chopped
1/2 cup chopped mallow
3 medium carrots, chopped
4 garlic cloves, chopped

In a large bowl, combine the flour, water, oil, rice syrup, salt, yeast, and vitamin C crystals and mix thoroughly. Add the onions, mallow, carrots, and garlic. Mix thoroughly by hand, then knead for 10 minutes. Form into loaves, place in greased pans and cover. Let rise until dough rises above the pans, about 1 hour.

Preheat oven to 350 degrees. Bake loaves for 35 to 40 minutes, or until they sound hollow when tapped.

Makes 2 loaves

Phragmities Pancakes !

1 cup phragmities flour
water (approximately 1/4 cup)
oil for frying

Put the flour in a medium bowl, slowly add enough water to form a paste. Mold the paste into a pancake form. Grease a skillet and fry the cake slowly until golden brown and crispy, about 6 minutes.

Serves 1

Variation: Preheat oven to 350 degrees. Grease a baking sheet and bake the cake for 15 minutes or until brown and puffy.

Phragmities Drop Biscuits *

1 cup phragmities flour
1 cup wheat bran or whole wheat flour
2 tablespoons safflower oil
3 tablespoons baking powder
approximately 1/2 to 3/4 cups water

Preheat oven to 350 degrees. Blend the flours and oil. Add just enough water to make a paste. Spoon silver-dollar-size portions onto a greased baking sheet and bake for 10 to 15 minutes or until crisp.

Yields 18 biscuits

Variation: Add 1/2 cup honey to the recipe

Phragmities Pita Bread x

1 cup phragmities flour
1 cup whole wheat flour
1 teaspoon baking powder
1/4 teaspoon olive oil
approximately 1/2 to 3/4 cups water

Preheat the oven to 350 degrees. Combine the flours, baking powder, and oil in a large bowl. Mix slowly, adding water a little at a time until mixture is of bread dough consistency. Knead until smooth, then let the mixture sit, covered, for 10 minutes.

Form 1 inch balls into individual patties ¼ inch thick and put on a greased baking sheet. Bake until the edges of the pita breads are brown and crispy, about 30 minutes.

Makes 12 pita breads

Green Pancakes (made from any vegetable flour) !

1/4 cup flour
small amount of water

Add a tiny amount of water very slowly to the flour to form a paste. Make two 2" in diameter pancakes. These are powerful pancakes, so you want to make them small. Fry slowly on a greased fry pan or cooking sheet, or bake in the oven as you would a cookie. Instant energy boost!

Makes two 2" pancakes

Nutsedge Crackers ~

1 cup Basic Mix (page 225)
1 cup nutsedge flour or oat flour
1/2 cup water

Preheat oven to 400 degrees. Lightly grease a baking sheet. In a medium-sized bowl, combine Basic Mix and flour. Add water to form dough. Knead about 12 times, until dough is smooth. Shape into pencil-like strands 1/2" thick or roll out flat onto baking sheet. Bake about 5 minutes until crisp. Cut into small pieces while warm.

Makes 24 to 36.

Variations: Use any flour you have such as malva, mixed grain or lamb's quarters in place of nutsedge or oat. Use goat's milk or regular milk in place of water.

Breads and Flour

The following will make excellent flour. Refer to pages 35-40 for information on drying, grinding and storage.

plantain leaves
clover leaves
dandelion leaves
strawberry leaves
cattail fluff, root
curley dock seeds and root
nutsedge (grainy top of plant)
Lamb's quarter leaves
pigweed leaves
bulrush pollen
thistle root
inner bark of birch bark

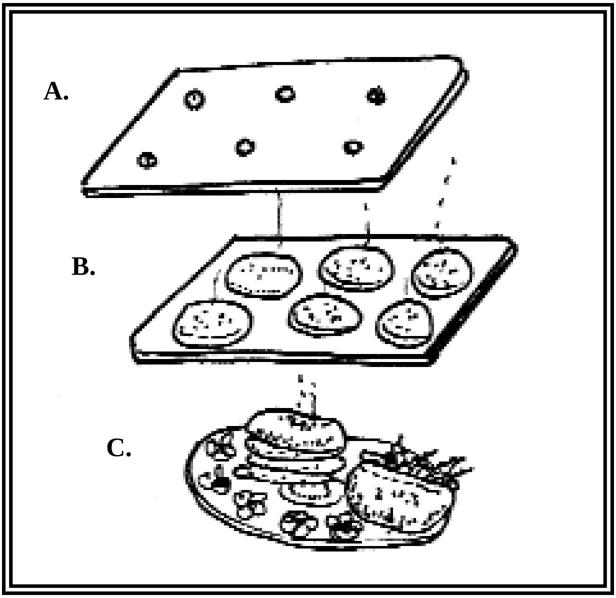

Homemade Wild Pita Bread.
A. Drop round balls of dough
B. Bake until edges are brown
C. Slice and fill pocket

Sweets

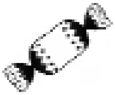

Lamb's Quarters Granola !

1 cup dried lamb's quarters seeds
3 cups old fashioned rolled oats
1 egg
1/2 teaspoon ground cinnamon
dash of ground cloves
1/4 teaspoon salt
1 1/2 cups water
3 tablespoons maple syrup, honey, or rice syrup

Combine the seeds, oats, eggs, and spices in a large bowl. In a large saucepan, bring the water and sweetener to a boil. Add the oat mixture a little at a time while stirring. Reduce the heat to low and use a fork to mix the granola until it is moist, but not wet. Add additional water, if needed, to make moist. Press layer into greased pans or pie plates. Allow to cool.

Serves 4

Variation: Add 3/4 cup raisins or 1 1/2 cups diced apples to the mixture.

Baked Prickly Pears ~

6 large prickly pear fruits
1/2 cup honey
1/2 teaspoon ground cinnamon
1/4 teaspoon grated lemon peel
4 tablespoons butter or 2 teaspoons safflower oil

Preheat the oven to 350 degrees.

Use a fork to hold the fruits while peeling to prevent the hair-like barbs from going into your fingers. After the fruits are peeled, cut in half, and remove the seeds. Grease a 6" baking dish. Line the bottom of the dish with the fruit. In a separate bowl, combine the honey, cinnamon, lemon peel, and butter or oil and drizzle over the fruit. Bake until tender, 20 to 30 minutes. Serve warm.

Serves 3

Meadowsweet Apples !

4 large baking apples, washed and cored
1 cup meadowsweet flowers, washed
1/2 cup maple syrup

Preheat the oven to 350 degrees. Grease a 1-quart glass casserole dish. Place the apples in the dish and put flowers in core of each apple. Drizzle the maple syrup over the apples and bake for 30 minutes, or until soft.

Serves 4

Malva Neglecta Oatmeal Cookies x

1 egg
1/2 cup honey
1/4 cup vegetable oil
3/4 cup finely chopped malva neglecta
1 cup old-fashioned rolled oats
2 teaspoons baking powder
1 cup whole wheat flour
1/2 cup sunflower seeds
1 1/2 teaspoons vanilla extract
2 teaspoons water
1/2 cup raisins

Preheat the oven to 350 degrees. Mix the egg, honey, and oil in a blender. Stir in the malva neglecta.

In a large bowl, blend the oats, baking powder, flour, sunflower seeds, vanilla, water, and raisins. Add the malva mixture and stir well. Drop by teaspoons onto a greased baking sheet. Bake for 12 to 15 minutes, or until golden brown on both sides. Cool on a rack.

Makes 3 dozen cookies

Plantain Raisin Cookies x

2 cups whole wheat flour
3/4 cup dried or fresh plantain seeds
4 tablespoons baking powder
2 tablespoons molasses
1/2 cup carob-covered raisins

Preheat the oven to 350 degrees. Mix all the ingredients in a big bowl. Add water slowly to form a thick, paste-like batter. Roll a pinch of dough between your palms and press onto a greased baking sheet. Continue to form cookies, then bake for 15 minutes or until golden brown.

Makes 2 dozen small cookies

Red Sumac Gelatin x

3 cups sumac berries
2 cups hot water
1 envelope unflavored gelatin
sumac berries for garnish

Make a concentrated sumac tea by putting the berries into a medium saucepan. Add the hot water and steep for 15 to 20 minutes. Strain. Follow the package directions for the gelatin, using the sumac tea in place of water. Garnish with a few berries.

Serves 4

Pine Popsicles *

gather the pine twigs, then strip the needles off the twigs. Use the twigs for popsicle sticks!
1 1/2 cups pine needles
2 cups boiling water

Put the needles in a small pot and pour in boiling water. Cover and steep for 10 to 15 minutes. Stir the needles once or twice to liberate more pine taste. Cool, then strain and pour into popsicle molds. Add pine twigs for sticks and freeze. (Some people put a dash of honey into the mold before pouring in the liquid for a sweet flavor.)

Makes 6 popsicles

Aloe Vera Ice Cubes x

fresh aloe vera leaves

Place aloe vera leaves on the flat side and skim the skin off with a sharp knife. Scrape the gel with a spoon, holding on to end of leaf. Pour into ice cube trays and freeze.

Mint Popsicles x

1 handful mint leaves, washed
mint leaves for garnish

Place the mint in a pot and add boiling water to cover. Cover the pot tightly and steep for 10-15 minutes. Strain and pour the liquid into popsicle molds. Place a leaf in each mold and freeze.

Makes 6 popsicles

Children love this refreshing treat. If you don't have popsicle molds, freeze in ice cube trays, adding a mint leaf to each cube. This is delicious in iced tea and other beverages.

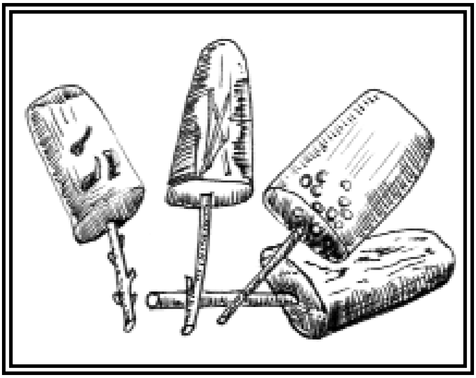

Rose Petal Candy x

4 tablespoons butter
2 cups rose petals
1/2 cup confectioners sugar

Melt the butter in a large frying pan. Fold the rose petals, one at a time, and coat with butter, flipping them over for a minute or so. The petals sometimes puff up. Remove as soon as crisp and brown, then drain on a paper towel. Roll fried petals in sugar and refrigerate until ready to serve.

Makes enough for 4 people

Fresh or dried, rose petals make delicious candy. Reconstitute dried petals by soaking them in water for an hour or so. Drain and use as fresh.

Wintergreen Candy x

3 cups dried wintergreen leaves
2 cups water
2 cups brown sugar
1 tablespoon distilled white vinegar

Slowly boil the leaves in the water for 20 minutes to make a strong tea. Strain. Add 6 tablespoons of tea to the brown sugar and boil until hard ball stage. Add the vinegar and boil for an additional minute or so. Pour the mixture onto a well-greased platter. It will harden as it cools and can be broken easily into pieces of candy.

Makes enough for 4

Blueberry Crumb Pudding x

Topping

1/4 cup sugar
1/3 cup flour
1/2 teaspoon cinnamon
1 tablespoon margarine
1 tablespoon shortening

Mix together and save for crumb topping.

Pudding

1/4 cup shortening
1/4 cup sugar
1 1/2 teaspoon baking powder
1 egg
1 cup sifted flour
1/8 teaspoon salt
1/3 cup milk
1/4 teaspoon vanilla
2 cups blueberries

Cream shortening and sugar. Add the egg, flour, baking powder, salt and milk. Spread this in an 8" square pan and spread 2 cups blueberries over this. Top with crumb spread. Bake at 350 degrees for 30 minutes.

Serves 4

Wild Plantain Cookies ~

2 cups whole wheat flour
3/4 cup plantain seeds (dried or fresh)
4 tablespoons baking powder
2 tablespoons molasses
1/2 cup carob covered raisins

Mix all ingredients well in a big bowl. Add tepid water slowly to form a thick, claylike paste. To form cookies, roll a pinch of dough between your palms and press onto a greased cookie sheet. Bake 15 minutes at 350 degrees or until golden brown.

Makes 18 to 20

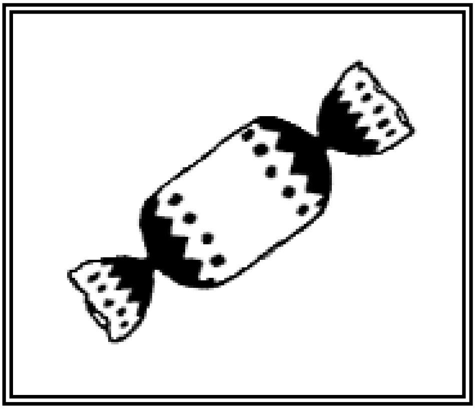

Right: Apprentice
Althea Dixon slices and dices a nopales cactus.

Apprentice
Hazel McManus presents a completely edible centerpiece and meal.

Apprentice
Tracy Alderman Dockery teaching homeschoolers, Arizona.

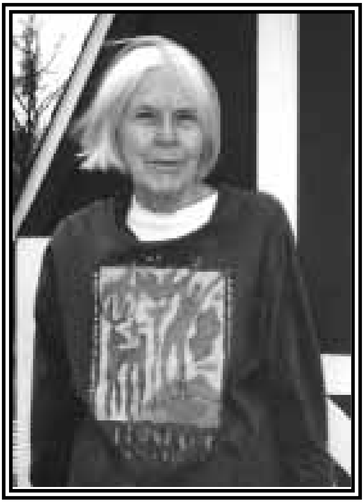

Apprentice
Shirlee Pettipiece.

Teas & Other Hot & Cold Drinks

Teas are an excellent way to develop your taste for wild foods. The wild ingredients in these teas are fresh unless otherwise stated. You may wish to use a tea infuser rather than strain the tea. As your taste for wild foods develops, you may even wish to eat the brewing grounds after drinking the tea.

Rule of Thumb: When you have fresh tea leaves, measure 2 cups of water per handful of leaves. Bring to a boil, cover and turn off heat. Let steep for 5-10 minutes. For dry ingredients (which cook more quickly), measure 1 cup of water per handful. Bring to a boil, cover and turn off heat. Let steep for 5 minutes. Generally, the flavor is more intense with dry herbs.

Aloe Vera Morning Tonic x

1 aloe vera ice cube (see Note)
3/4 cup fresh orange juice

Combine ice cube and orange juice in a glass. Let cube melt, then drink.

Serves 1

Note: To make cubes, place an aloe vera leaf on its flat side and peel the rounded skin off with a sharp knife. Scrape the gel with a spoon, holding on to one end of the leaf. Pour the gel into ice cube trays and freeze.

Balsam Tea !

1 inch tips balsam
1 cup water
honey (optional)

Place balsam tips in a small saucepan and add the water. Cover. Bring to a boil, then lower the heat and simmer for 10 minutes. Strain tea if desired, add water to taste if tea is too strong or sweeten with honey if desired.

Serves 1

Blackberry Bramble Tea !

1 6 inch blackberry shoot
1 cup boiling water
honey (optional)

Peel the shoot until most of the bark is off. Break into several sections to expose more of the surface. Place in a teapot, add water, cover pot and steep for 5 minutes. Strain if desired. Sweeten with honey, if desired, and serve hot or cold.

Serves 1

Balsam Toddy x

1 to 3 inches balsam twigs with needles
2 cups soymilk
dash of ground cinnamon

Place balsam twigs in a teapot. Bring the soymilk almost to a boil, remove from heat, and pour over balsam. Steep, covered, for 5 minutes. Strain, then add cinnamon and serve piping hot.

Serves 1 or 2

Goldenrod Tea !

1 sprig goldenrod flowers
2 cups boiling water
honey (optional)

Place flowers in a teapot. Add water, cover pot, and steep for 5 minutes. Sweeten with honey if desired. Serve hot.

Serves 1 or 2

Chamomile Tea !

1/4 cup chamomile leaves and flowers, washed
2 cups water
honey (optional)

Place chamomile in a small saucepan. Add water, bring to a boil, turn off heat, and cover. Steep 10 minutes, then strain and serve with honey if desired.

Serves 1 to 2

Wild Greens Tonic x

This tonic will keep in the refrigerator for 2 days or so. I freeze the brew as ice cubes for use in recipes or for a cup of tea.

1 cup willow leaves, washed
1 cup mint leaves, washed
1/2 cup red clover buds
2 quarts water

In a blender of food processor, blend all ingredients until green and smooth. Serve hot or cold.

Serves 8

Dandelion Root Coffee or Chicory Coffee !

1 teaspoon dandelion root or chicory root powder
(see page 76)
1 cup water

In a medium saucepan, combine the powder and water. Bring to a boil and simmer, covered, for 3 to 4 minutes. Drink as a coffee substitute.

Serves 1

Hot Meadowsweet Tea !

4 white meadowsweet flower spires
4 to 6 meadowsweet leaves
1 cup water

Bruise the leaves and flowers, then pour boiling water over them and cover for 7 minutes. Strain. Serve hot, or add 1 teaspoon lemon juice and serve cold.

Serves 1

Maple Tea !

1/2 cup maple seeds, clipped from the wing
1 cup boiling water
2 teaspoons maple syrup (optional)

Place seeds in a teapot, add water, and steep for 5 minutes. Sweeten with maple syrup if desired.

Serves 1

Mullein Root Tea !

Mullein root is an antihistamine. When you feel a cold coming on, simmer the root in water and drink the liquid. The root can be used several times to make tea. After five or six uses, bruise the root with a hammer, or cut slits in the sides to release more nutrients.

1 large mullein root
honey (optional)

Scrub the root well and cut slits in several places with a strong knife. Place the root in a non-corrosive pot. Add water to cover, bring to a boil, and simmer, covered, for 5 minutes. Sweeten with honey if desired. Hang the root up to dry.

Serves 1

Willow "Aspirin" Tea !

For pain relief, chew the willow twigs raw as well.

1 6-inch willow tendril, including leaves and catkins
1 cup boiling water
honey (optional)

Using a knife edge, score and remove the bark. Slice the bark, leaves, and catkins into a teapot. Break up the twig into several sections and add to the teapot. Add water, cover, and steep for 5 minutes. Strain, and sweeten with honey if desired.

Serves 1

Pine Needle Tea !

A pine tree is filled with vitamin C, so pine tea is delicious and nutritious. Cold or hot, the flavor is wild. Perhaps in time you will develop a taste for *raw* pine needles. They are edible and so is the bark. Your backyard tree might hold several hundred cups of tea—perhaps thousands. Seasons make no difference; pine is green and succulent all year long.

1 cup fresh pine needles
1 cup boiling water
honey (optional)

Put the needles in a teapot and pour in the water. Steep for 5 to 7 minutes, then strain. Sweeten with honey if desired.

Serves 1

Wintergreen Tea !

Wintergreen tea is scrumptious. With this tea you can comfortably pass a cold winter's night. Gather the flat, shiny leaves and spread them on a screen to dry in a warm, airy spot. The back window ledge of a car is good. Let them dry slowly in the sun and watch that they don't dry too fast. This technique does not work in August when the sun is too hot. I use a hand grinder or old meat grinder to obtain the powder for this tea.

2 teaspoons ground dried wintergreen leaves
1 cup boiling water
honey (optional)

Put wintergreen into a cup and add water. Steep for 5 minutes. Sweeten with honey if desired.

Serves 1

Violet Tea !

a handful of violet flowers
2 cups of water

Pick flowers after dew has evaporated and dry on a screen for a day. Add flowers to the water and cover. Steep for 10 minutes. Skim flowers out or eat them!

Serves 2

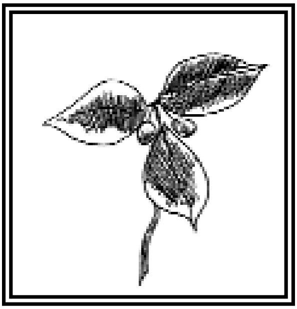

A Choice of Blend of Herbs for Tea !

Gather a combination of lawn herbs. Examples: thyme, mint, red sorrel, yarrow, spearmint, goldenrod, rose hips, among a few. Arrange the herbs so the stems are the same length and bundle them together. Hang the bundle high in a warm kitchen. When the leaves are dry enough to crumble, rub the bundle between your palms over a tray. Then rub that mixture through a sieve. You can bottle this Fine Blend and store for future exquisite tea! Also, one tablespoon is adequate for a cup of soup.

Blackberry Bramble Wine !

1 gallon blackberry shoots or tips
1 pound raisins, chopped
1 gallon boiling water
3 1/2 cups sugar or 2 cups honey
1 tablespoon all-purpose wine yeast

Wash the blackberry shoots. Using scissors, clip shoots into 1-inch pieces and place in a sterilized 5-gallon container. Add the raisins, then cover with boiling water. Cover tightly. The next day, strain liquid and return it to the container. Add the sugar or honey and stir well. Prepare the yeast according to the package directions. When bubbly, add the yeast mixture. Cover tightly. In a few hours the mixture will begin to foam. Stir occasionally, until fermentation slows. When bubbling ceases (in about 6 weeks), siphon the liquid into sterilized bottles. Cork and let sit in a dark place. Serve in a year's time.

Makes about 5 bottles

Daisy Petal Wine !

1 gallon daisy heads with petals (no stems), washed
1 gallon boiling water
3 pounds sugar or 1 1/2 pounds honey
1 package all-purpose wine yeast

Pack the daisy heads into a sterilized 2-gallon crock. Pour the boiling water over the heads and stir. Cover tightly. Let the mixture stand for 1 1/2 weeks. Strain.

Bring the liquid to a boil and remove from the heat. Transfer to a sterilized 4-gallon container. Stir in the sugar or honey until it is dissolved. Let cool. Meanwhile, prepare yeast according to package instructions. Add to liquid and cover tightly. Let the mixture sit for a week while it foams. Let it sit until the fermentation clears, about another 6 weeks. Siphon into sterilized bottles and store for 4 to 6 weeks before serving.

Makes about 5 bottles.

Dandelion Wine !

1 gallon dandelion flowers
4 oranges
1 gallon water
5 pounds sugar
1 package all-purpose wine yeast

Remove the green parts and soak the flowers in boiling water. Peel oranges and save the juice. Pour boiling water over flowers and let stand for 2 days. Add orange peels, bring to a boil, then let simmer for 15 minutes. Strain liquid and parts through a fine strainer or use a jelly-bag. Add sugar to strained liquid and cool to room temperature. Add the juice of the oranges and yeast. Pour into a fermenting jug fitted with an airlock. When fermentation clears, siphon into sterilized bottles. Store at least 6 months before serving.

Makes 5 quarts

Prickly Pear Wine x

1 dozen prickly pear buds
2 quarts boiling water
4 cups honey
1 package all-purpose wine yeast

Hold the prickly pear buds with forceps in one hand and peel under water. Place pears in a sterilized 1-gallon crock and pour in boiling water. Cover tightly. Place mixture in a warm place for 4 or 5 days or until mixture "turns."

Strain liquid and return to gallon crock. Add honey. Prepare the yeast according to package directions and stir into the mixture. Cover it tightly, and let it sit in a dark warm spot for 1 more week or until foaming clears. If desired, transfer to a carboy fitted with an airlock. Let it sit until the fermentation ceases in about 6 weeks.

Siphon into sterilized bottles and cork tightly. Let rest for about 8 months.

Makes about 3 bottles.

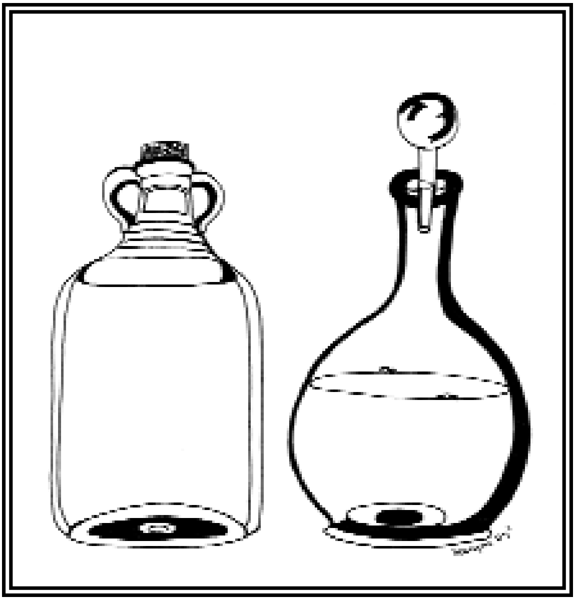

Rose Petal Wine !

1 quart rose petals, washed
2 quarts water
2 cups honey
1 package all-purpose wine yeast

Simmer the rose petals in water for 5 minutes, stirring occasionally. Pour the liquid and petals into a sterilized 2-gallon glass container. Cover tightly and let the mixture sit for 1 week. The petals and water will turn almost rancid.

Strain the mixture and put the liquid in a clean pot. Bring to a boil, then turn off heat. Transfer to a sterilized 2-gallon container. While mixture is very warm, add the honey. Let it cool. Meanwhile, prepare the yeast according to package directions. When the mixture has cooled, add the yeast mixture, stir once, and cover tightly. If desired, after 1 week, transfer the mixture to a fermenting jug fitted with an airlock. When the fermentation ceases, in about 6 weeks, siphon the wine into sterilized bottles and cork tightly. Let rest for 4 months.

Makes about 3 bottles

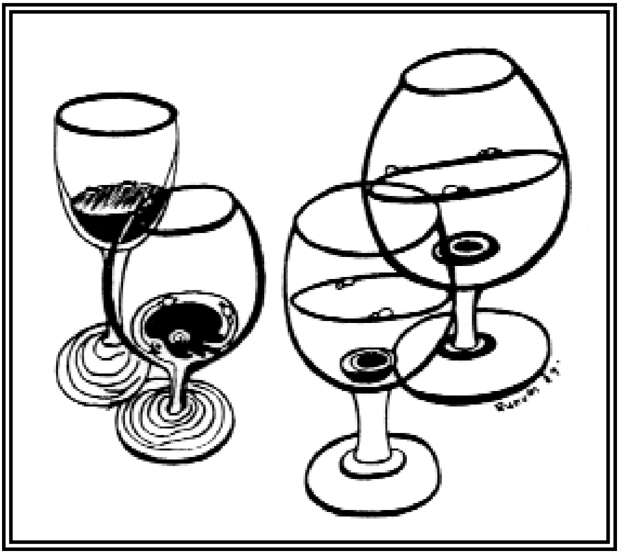

Sugar Maple Twig Wine !

1 gallon tightly packed sugar maple twigs, washed
1 gallon boiling water
4 cups honey
1 package all-purpose wine yeast

Pack the twigs into a casserole. Pour in the water and simmer for 20 minutes. Transfer to a sterilized 2-gallon crock. Cover tightly. Let sit for 3 days.

Strain the liquid and add honey. Stir well. Prepare the yeast according to package directions and add to the mixture. Stir, then cover tightly. Mixture will foam for 1 week. If desired, at this point transfer to a sterilized fermenting jug fitted with an airlock. Keep covered until fermentation stops, in about 6 weeks. Siphon into sterilized bottles and cork. Store in cellar for 3 to 4 months before serving.

Makes about 5 bottles

Preserving Wild Foods

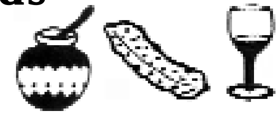

Dandelion Jelly x

Petals from 40 dandelion flowers
3 cups boiling water
1 package commercial pectin
3 3/4 cups sugar or 1 3/4 cups honey

Wash the petals and add them to the water. Steep for 10 minutes, then strain. Return the juice to the heat, bring to a boil, and immediately add the pectin. Boil mixture for 1 minute, then add sugar or honey and simmer until it makes a hard ball (260 degrees). Pour into hot, clean pint jars and seal with wax or process in hot water bath for 20 minutes.

Makes 2 pints

Cactus Lemon Marmalade x

1/4 cup thin lemon slices
1 cup prickly pear cactus pulp (Split the fruit and remove the tips & seedy centers.)
1/4 cup honey

Put the lemons in a small saucepan and cover with water. Soak overnight. The next morning add the cactus and honey and cook until the mixture is thick. Pour into a hot pint jar and cover with sealing wax or process in a hot water bath for 20 minutes.

Makes 1 pint

Saguaro Cactus Jelly x

16 to 24 saguaro fruits
1 package powdered pectin
1/4 cup lemon juice
2 1/2 cups honey

Split the fruit and remove the tips & seedy centers. Place the pulp in a pot, cover with water, and simmer until well-cooked, about 15 minutes. Strain the pulp using a fine colander. In a large pot, combine 3 1/4 cups saguaro puree, pectin, lemon juice,

and honey. Boil for 1 minute, stirring constantly, then pour into scalded pint jars. Cover with sealing wax or process in a hot water bath for 20 minutes.
Makes 4 pints

Rose Petal Syrup !

4 cups rose petals
1 quart water

Put the petals in a large saucepan and cover with water. Simmer for 20 to 25 minutes, stirring often, until it is thickened to syrup consistency. Cool, bottle, and refrigerate.

Makes 1 quart

Rose Hip Jam !

4 cups rose hips, washed
4 cups water
2 cups honey

Cook the hips in the water until tender, about 15 minutes. Put the hips through a meat grinder, food processor, or coarse sieve, then push the pulp through a fine sieve to obtain a smooth puree. Add 1/2 cup honey for each cup of pulp and cook until thick. Pour into scalded pint jars and seal with sealing wax or process in a hot water bath for 20 minutes.

Makes 4 pints

Queen Anne's Lace Salt !

This salt substitute lasts for years and years. Caution: Be 100 percent certain on your identification. The fuzzy hairs on the stem of Queen Anne's lace are dry at this stage. Running your finger down the dried stem feels like sandpaper.

Collect Queen Anne's lace seeds in the early fall, when flowers have “birdcaged” and turn a light brown. When the seeds mature, the entire head naturally dries up. Pull on the dried seed top; the seeds are easily extracted in your hand, even without taking the “birdcage” off the stem. The remaining seeds will propagate new growth next season.

Place the seeds on a tray and put into a 250 degree oven for 5 minutes. You may put them on a tray on top of the stove's pilot light until completely dry. Bottle in glass jars for long-term storage.

Caution: Pregnant women should not use these seeds: medicinal field guides show that the seeds have abortive tendencies.

Herbal Vinegars !

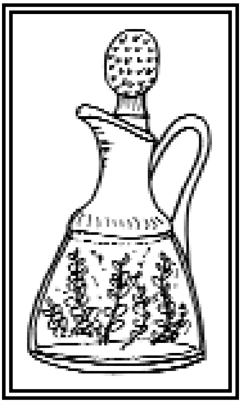

Fantastic herbal vinegars can be made from the blossoms of red or white clover, roses, catnip, mint, goldenrod, blackberries, raspberries, violets, and lamb's quarters. You might enjoy adding other flavors, such as thyme, garlic, or Queen Anne's lace. My favorite vinegar is a combination of greens and spices. Experiment with different flavors to find your favorite. These vinegars make wonderful gifts.

Wash and drain the greens and flowers and place in a 2-quart glass container. Bring vinegar to a boil, then pour it slowly over the greens and flowers. Cap and store the container in a high place, out of direct sunlight. The vinegar will cook the greens a bit in the warmth. Shake it every day for best results. After 2 weeks, strain the mixture, cap tightly, date and label the container.

Makes 2 quarts

Lemon Pepper x

1 peel from lemon, scrubbed
1 2-ounce container garlic salt
1 3-ounce container celery salt
1 2-ounce container coarse black pepper

Dry lemon peel thoroughly, about 3 days on a plate or 1 to 1 1/2 hours in an 250-degree oven. Don't let it get brown. Pulverize dried peel in blender until powdered. Add the rest of the ingredients, mix thoroughly and refrigerate. Use as a spicy additive.

Makes about 1 cup

Cattail Pickles !

Harvest cattails at the height of summer, when they have large stems about 1/2 to 1 inch in diameter. Cut the cattails at water level, remove seed heads, then peel the stem to the inner white core. Wash this pith under running water. Lay out on counter or board and slice into 6-inch pieces. Place pith pieces in pint jars, add your favorite pickling recipe. (Use hot cider vinegar and 1 teaspoon pickling spices (commercial) to each 8-ounce jar of vinegar.)

Stems yield approximately 4 to 6 jars, depending on the diameter of the pith and the length of the stems.

Makes 4 pints

Cactus Pickles x

2 quarts prickly pear fruits
2 cups honey
2/3 cup cider vinegar
3 ounces red cinnamon candies Or
    1 small cinnamon stick, ground

Put cinnamon candies or stick in a cheesecloth bag. Place the fruit in pint jars. Mix the honey and vinegar in a saucepan and add spice bag or cinnamon candies to pan. Bring to a boil, then reduce heat and simmer for 5 minutes. Remove the bag and pour the liquid over the fruit to 1/4 inch from tops. Process in a hot water bath for 25 minutes.

Makes 2 pints

Pickled Mustard Flowers !

1 cup mustard flowers, washed
8 ounces cider vinegar
1 teaspoon pickling spices

Put flowers into a clean, wide-mouthed 8-ounce canning jar. In a non-corrosive saucepan, bring vinegar to a boil. Add pickling spices and simmer ‘till flavors mix, about 5 minutes. Remove spice bag and pour vinegar over the flowers to within 1/2 inch of top of jar. Loosely cap the jar and place in a 2-quart sauce pan filled with water. Bring to a boil and boil for 5 minutes.

Makes 1 jar

Rose Petal Butter x

1 cup rose petals, 2 cups
½ stick sweet butter

Alternate layers of petals with butter, so you get a "marble" layered appearance. Seal in a container and store in refrigerator. Blend when ready to spread on crackers or bread for a wonderful appetizer! Or, cut into little patties for another type of appetizer.

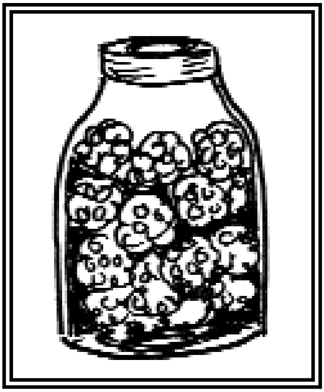

Milkweed Bud Pickles

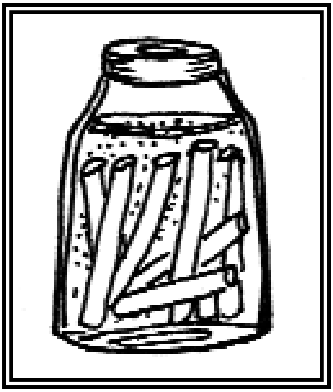

Cattail Pickles

Desert Meal.

Sundries

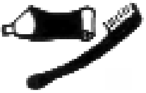

Wild foods are fantastic vegetables and make great teas, but have you ever considered them as sundries or accessories? The following recipes give you more ways to use wild foods each day.

Nettle Hair Softener !

1 cup dried nettles
2 cups water

Simmer the nettles in the water for 20 minutes, then allow the mixture to cool. Strain.

Rinse your hair with the liquid. Let it stay in your hair for 2 to 3 minutes, then rinse with nettle liquid again. Then rinse with clear water.

Makes about 2 cups

Cactus Hair Softener !

1 cup cactus skins or pads, chopped
2 cups water

Simmer the cactus for approximately 30 minutes until almost dry. Pulp will turn water thicker. Cool and rinse hair with liquid after washing. Rinse again with clear water.

Makes 2 cups

Note: You may choose almost any type of pulpy cactus such as prickly pear, nopolitas, cow’s tongue.

Yarrow Tooth Powder or Paste !

1 piece of burnt toast
yarrow leaves, dried and ground

Scrape the charcoal off a piece of burnt toast. Make a mixture of half charcoal and half ground yarrow for a refreshing, cleansing tooth powder. Or add a small amount of water to form a paste.

Variation: Dry a thick, straight-stemmed yarrow plant. Remove and save the leaves for tea. Cut the desired stem toothbrush to size. Using a knife, cut into one end over and over to make a fuzzy tip.

In an emergency situation, cut a large raw yarrow leaf and scrub your teeth by pressing the leaf to them and rubbing. Rinse.

Wild Facial Lotion and Skin Softener !

- 1 cup wild lettuce leaves, washed
- 1 cup mullein leaves, washed
- 1 quart water

Combine the leaves and water in a 2-quart saucepan. Cover and bring to a boil, then simmer gently for 30 minutes. Strain.

Put the resulting liquid in a bottle for use as a skin softener.

Makes about 1 quart

Wildflower Punch Bowl x/~

- 1 large plastic cake dome to fit inside punch bowl
- 1 cup each violets, daisies, red rose petals, multi-colored hibiscus, and pansy flowers
- several rose leaves
- Several small ferns

Invert the cake dome and add 1 inch of water. Place the violets in the water and freeze solid. Repeat the procedure with each type of flower. Use the rose leaves and ferns for the last layer.

Cover and keep frozen until ready to use. When you are ready to use it, unmold and float in a punch bowl with your favorite punch recipe.

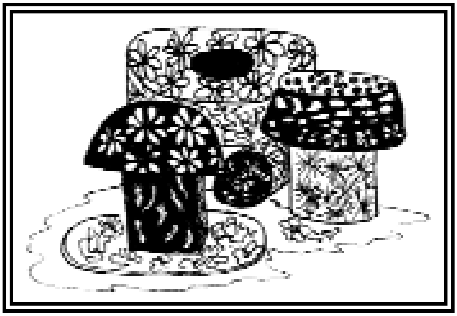

Flower Cake.

Barbeque Cactus Burrs.

Desert Snake Stick (top).

Typical Desert Walk.

Gathering Desert Tumbleweed.

Part IV: Reference Section

Cultivation of Wild Edibles

During our Adirondack homesteading years we made many attempts at gardening, using conventional methods. Indeed, experimenting with gardens was always a summer pastime.

The 175-foot strip garden worked best. I had found out when the electric company had last defoliated the areas, beneath a power line, and since it was many years ago, I chose the strip because it was already partly cleared. The garden was high, sunny, dry, and basically rockless, as Adirondack soil goes.

I enjoyed every moment of clearing the land, especially the weeds. As I seeded with vegetable seeds, I was able to supply the family's immediate food needs, harvesting a year's supply of lamb's quarters, chickweed, sheep sorrel, violets, strawberries, and dandelions. The amount of our wild stored food continued to grow, from dried to canned to pickled. The family enjoyed nutritious meals of vegetables, spices, and herbs that were both wild and cultivated.

But after two years of fixing fences and hunting predators, of purchasing cloth, saving newspapers, and recycling plastic as covers to beat the frost, I finally turned to wild foods exclusively. Long after the cultivated vegetables were gone in fall, the wild foods continued to grow under the snow, under hay, and in cold frames. I asked myself why I had spent all that time and energy gardening when wild foods grew with such abandon, untouched by predators, insects, and weather. The answer was simple; the problem *was* getting *used* to the change in eating patterns.

Over the years we learned to use the foods nature provided so readily, and in the process discovered a whole new culinary world rich in flavor and nutrition.

A Field of Wild Foods

Gardens as we know them are basically artificial, filled with introduced species and forced arrangements. We don't all have acres of wild land near home, however, from which to forage our daily meals. On the other hand, many of us can consider changing that backyard garden or lawn into a productive field of wild foods. Weeds are more natural, and can supply much of your food needs.

Your Private Wild Food Area

I recommend stringing off a special area—perhaps a 4 by 6-foot section—of rich loamy garden soil or area previously tilled. Don't use herbicides on the soil; usually the weeds already there make great herbicide-resistant garden starters. (Only approximately one percent of all weeds are poisonous.)

For best results, transplant foraged weeds after a light rain, when young plants are more easily dislodged and the soil clings to the roots. I use a spoon for shallow-rooted types such as white clover, plantain and garden sorrel, planting them in the front rows about two to three inches apart. Experience has taught me that these plants will eventually mat and form a close-knit crop. Taller medium-rooted plants like milkweed and mustard are transplanted with a trowel and placed six to eight inches apart behind the "shorties." As these larger plants grow, they'll branch out, so to minimize encroachment on other more traditional plants, place rocks, peat moss or hardwood chips around the border of your wildfood crops. Pine needles are also excellent for this purpose. Even if you let only twenty feet of lawn return to its natural state, you have a productive food strip, or field.

This new food strip will be a natural field; eventually free of poisons.

For example, if you have been using a weed killer, the chemicals have been killing off sheep sorrel, dock, plantain, dandelion, and clovers. As wild "weeds" they will return almost immediately. At this point, harvest only the youngest leaves for minimal chemical content while your soil replenishes itself. Long-term lawn treatment requires more time. Visit a local hardware store or nursery, or call the US FDA or County Extension Service, for information on residuals for chemicals. Also, there are several books on replenishing your soil the natural organic way without pesticides.

Saltiness and High Concentrations of Nitrates

Liming the soil at the base of your plants corrects both the alkalinity and nitrates. Desert dwellers should also water more frequently since this will reduce the alkaline salt buildup.

Unless it's a rainy season, your weed garden should be watered as much as your traditional vegetable garden—about three times a week; more in arid desert areas. Keep the plants moist, but not soaked. I mist my wildfood strips with a spray attached to a hose.

Compared to plants like tomatoes, wildfoods respond twice as fast to water, fertilizer and light. And they're hearty; in a heavy frost you don't need to cover them.

Harvesting

After only a week, I'm usually able to pinch back or scissor off leaves and use them as food. In fact, cutting back a few branches or leaves generally stimulates greater growth.

In the cold Adirondack winters I allowed six inches of snow (plus hay, leaf composts or peat moss) to cover my garden. This created a hothouse effect. (Boards also can act as a winter greenhouse, especially for purslane.) In this way, I was able to harvest thyme, strawberry leaves, chicory and chickweed the whole year round.

Likewise, some common garden plants are poisonous. Especially if you have children, you'll want to eliminate or at least be aware of the dangers these plants hold. Some of these plants are listed below. See the books listed in Suggested Reading for more information.

Common Poisonous Plants

American bittersweet
American yew
Anemone
Azalea
Black locust
Bleeding heart
Bloodroot
Iris
Bouncing Bet
Buckthorn
Buttercup
Butterfly weed
Castor bean
Celandine
Choke cherry
Cowbane
Crown of thorns
Daphne
Deadly nightshade
Delphinium
Dogbane
False hellebore
Foxglove
Four o'clocks
Golden chain
Horse nettle
Horse chestnut
Horsetail
Jack-in-the-pulpit
Jack-o-lantern mushroom

Jimsonweed
Lantana
Larkspur
Lily of the valley
Marsh marigold
May apple
Mountain laurel
Narcissus
Nightshade
Poison hemlock
Pokeweed
Prickly poppy
Privet
Rhododendron
Rhubarb leaves
Scotch broom
Skunk cabbage
Snow on the mountain
Star of Bethlehem
Tansy
Tomato leaves
Trumpet flower
Virginia creeper
Water-hemlock
Wisteria
Yellow oleander

Do not let this list confuse or inhibit your foraging. The stress here is on identification. Know your own property. Other good field guides are found in the library or sold in book stores. There are too many poisons in the natural world to memorize. Therefore, (1) memorize your edibles, (2) identify the poisons nearby, and be absolutely sure before you try a new plant (see section on foraging, Part I, page 22).

Building a Wild Foods Garden

Wild plants will grow everywhere, in desert, valley, woods, and fields. Vacant lots and sidewalk cracks show evidence of the tenacity of common plants to live. Dandelions, plantain, yarrow, chicory and violets are common in many places. So begin by looking at your local edible landscape. View the surroundings with calm deliberation and a pencil. See the abundance already there and know the ease of creating a wild foods garden.

I cannot be specific here as to soil needs, moisture requirements, and the like. You have to consider your environment, your geographic location when working with these wild plants. The "gardening" I speak of is mostly one of transference, and *support*. You can move plants to where you want them (provided you haven't taken them from a protected area) and you can provide conditions that they like. Peat moss may help keep the surrounding weeds from inundating your food strip. Mulch helps too—one food field has 6 inches of wood chips around the main plants. Leaves, rocks, pebbles also work. Sidewalks retain a lot of moisture. But even simple things like cedar chips or sawdust may increase the ability of eastern soil or add to the saltiness (alkali) of desert soil. In general, it is best to use indigenous soils and topsoil.

Use a spoon, trowel, or shovel to gently release the plant's grip on the soil. See the chart on the next page for guidance on plant reactions to being moved.

In developing your wild foods garden, you may want to move plants around a bit or bring in native plants that haven't otherwise appeared. Be sure to dig up all of the roots, then set in new place and pat soil in gently. Water well.

The Inside Garden

Starting a wildfood garden in your home isn't much harder than maintaining your outdoor one. Clay pots prevent the soil from becoming rancid, and positioning them away from heating or air conditioning units helps regulate temperature. Use potting soil and organic fertilizer, and don't forget to *mist* the plants three times a week.

In my Adirondacks log cabin, I transplanted dandelions and chicory into an old-fashioned wash tub, then placed it down in the cellar where the dark, damp conditions produced an amazing abundance.

Much of your indoor gardening depends on your climate, the temperature of your house, and the hours and strength of available sunlight. I sometimes put my plants in one big twig basket and relocate them to different areas. Artificial grow-lights are occasionally necessary for certain plants like clover, which requires at least 12 to 14 hours of high light per day. When introducing artificial light, at first leave it on only a few hours.

Indoor Garden.

Transplanting Wild Edibles

Plant Name	Planting Depth	Method	Ease of Adjustment
Aloe Vera	deep	trowel	perky
Amaranth	shallow	spoon	wilts
Aster	medium	trowel	wilts
Balsam fir	deep	shovel	perky
Birch	deep	shovel	wilts
Blackberry	deep	shovel	perky
Blueberry	deep	shovel	perky
Bulrush	deep	shovel	perky
Burdock	deep	shovel	wilts
Cattail	deep	saw, shovel	perky
Chamomile	shallow	spoon	perky
Chickweed	shallow	spoon	perky
Chicory	deep	shovel	wilts
Cholla	medium deep	shovel	perky
Clover	shallow	spoon	perky
Daisy	medium	trowel	perky
Dandelion	medium	trowel	wilts
Dock	deep	shovel	wilts
Evening Primrose	medium	shovel	perky
Field thistle	medium	spoon	wilts
Fireweed	medium	trowel	wilts
Goldenrod	deep	shovel	perky
Grape	deep	shovel	wilts
Lamb's quarters	shallow	spoon	wilts
Mallow (Malva)	medium	trowel	wilts
Maple	deep	shovel	perky
Meadowsweet	deep	shovel	perky
Milk thistle	deep	shovel	perky
Milkweed	medium deep	trowel	wilts
Mint	shallow	spoon	wilts
Mullein	deep	shovel	wilts
Mustard	medium	trowel	wilts
Nettles	deep	shovel	perky
Phragmities	deep	shovel	wilts
Pine	deep	shovel	perky
Plantain	shallow	spoon	wilts
Prickly pear	deep	shovel	perky

Transplanting Wild Edibles (Continued)

Plant Name	Planting Depth	Method	Ease of Adjustment
Purslane	shallow	spoon	perky
Queen Anne's Lace	deep	shovel	wilts
Raspberry	deep	shovel	perky
Rose	deep	shovel	perky
Sheep sorrel	shallow	spoon	perky
Shepherd's Purse	shallow	spoon	perky
Sow Thistle	deep	shovel	wilts
Strawberry	shallow	spoon	perky
Sumac	deep	shovel	perky
Sunflower	medium	shovel	wilts
Thyme	shallow	spoon	perky
Tumbleweed	shallow	trowel	perky
Violets	shallow	spoon	perky
Wild lettuce	medium	trowel	wilts
Willow	deep	saw, shovel	perky
Wintergreen	shallow	spoon	perky
Wood sorrel	shallow	spoon	perky
Yarrow	medium	trowel	perky

Glendale Library Plants.
(See page 276)

Edge your wild foods garden both to keep out nonedible plants and to give it a sense of unity.

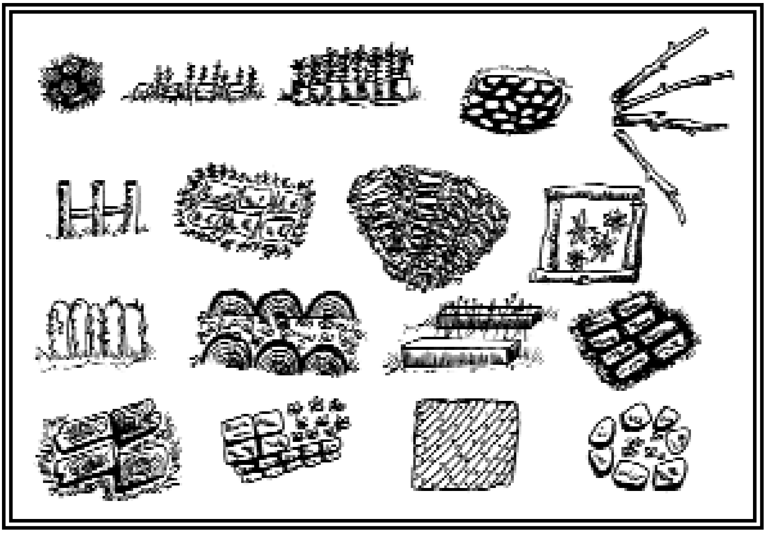

Edgings and Path Decorations

Top (left to right): crosscut circular slices of log, bricks, 2 by 4 ends buried in ground, smooth egg-shaped rocks with soil in between, straight limbs or poles.

Second row (left to right): Natural fences from wood, larger slate pieces, roundish stones, cut poles in a raised bed.

Third row (left to right): Cross fencing or picket sections, crosscut logs halved, 2 by 4's,

Bottom row, (left to right): large cut lumber boards, flagstones or bricks, decking or patio lumber, stones.

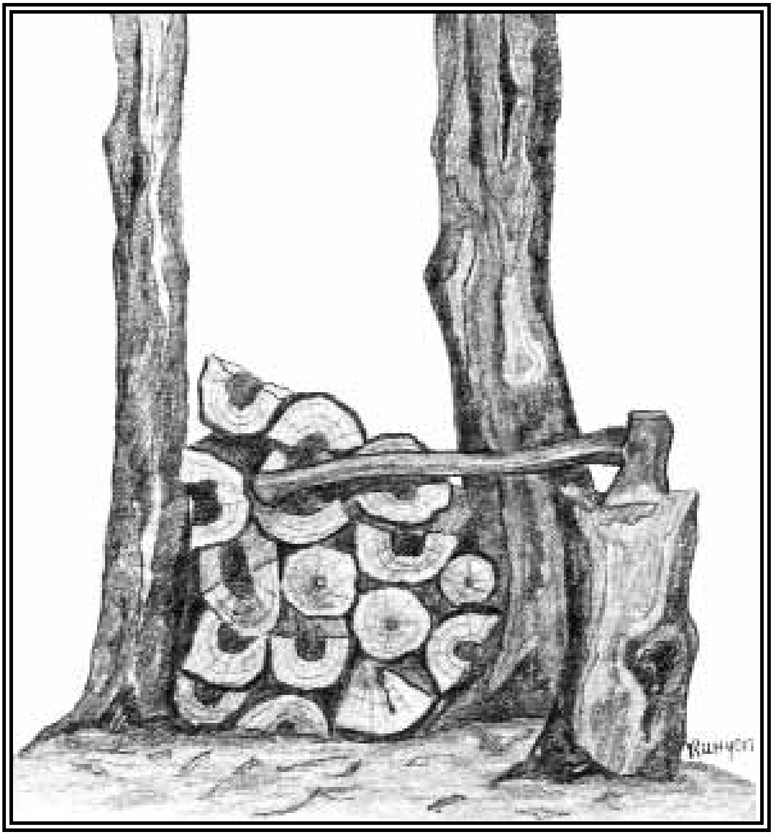

So much depended on wood.

Illustration from *A Glimpse of Peace* and *Flickering Free* written by Ken Heitz, illustrated by Linda Heitz (Runyon)

Linda's former husband Ken published two now rare books back in the early 70's now out of print.

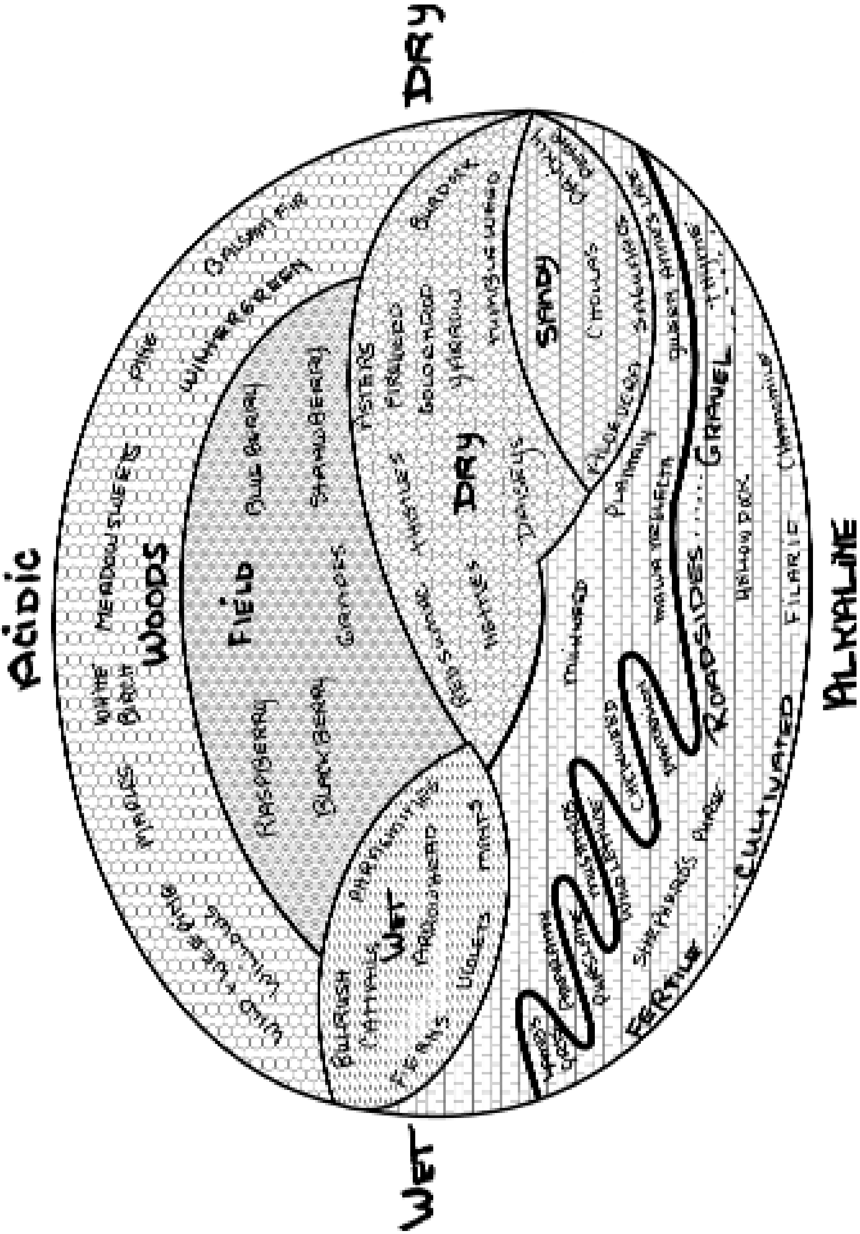
ACIDIC
DRY
WET
ALKALINE
WOODS
WILD WEEPING WILLOWS
MAPLES
WHITE BIRCH
MEADOWSWEETS
PINE
WINTERGREEN
BALSAM FIR
FIELD
RASPBERRY
BLUEBERRY
BLACKBERRY
STRAWBERRY
GRAPES
WET
BULRUSH
CATTAILS
ARROWHEAD
FERNS
VIOLETS
MINTS
DRY
ASTERS
FIREWEED
GOLDENROD
BURDOCK
YARROW
THISTLES
RED SUMAC
NETTLES
TUMBLEWEED
DAISEYS
SANDY
PRICKLY PEAR
CHOLLAS
ALOE VERA
SAGUAROS
LAMB'S QUARTERS
AMARANTH
PURSLANE
MUSTARDS
WILD LETTUCE
CHICKWEED
DANDELION
SHEPHERD'S PURSE
MILKWEED
PLANTAIN
MALVA NEGLECTA
QUEEN ANNE'S LACE
THYME
FERTILE
CULTIVATED
ROADSIDES
GRAVEL
YELLOW DOCK
FILARIE
CHAMOMILE

Soil-type matrix where wild foods are naturally found.

Legend:
A hypothetical complete Wild Foods Garden.

▨ **Woods.** . .Includes: wild & weeping willow, maples, white birch meadowsweet, pine, balsam, & wintergreen.

▨ **Field**. . . Includes: raspberry, blackberry, grapes, strawberry & blueberry.

▨ **Wet.** . .Includes: bulrush, , phragmities, cattails, arrowhead, ferns, violets & mints.

▨ **Dry**. . .Includes: red sumac, thistles, nettles, daisies, asters, fireweed, goldenrod, yarrow, tumbleweed & burdocks.

▨ **Sandy**. . .Includes: prickly pear, aloe vera, cholla & saguaro.

▨ **Fertile / Roadsides**. . .Includes: lamb's quarters, amaranth, purslane, mustards, wild lettuce, chickweed, dandelion, shepherd's purse, milkweed, plantain, malva neglecta, yellow dock, filarie, chamomile, Queen Anne's lace & thyme.

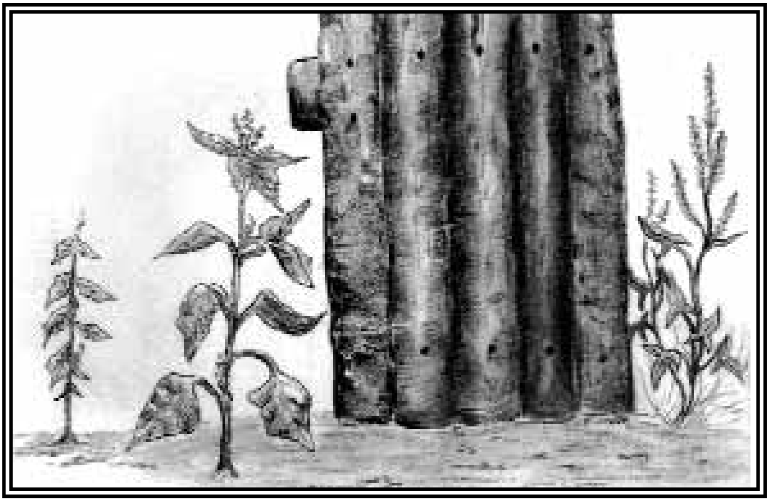

Left: lambs quarters Right: sheep sorrel

Public Wild Food Identification Walk Ways

Sabael Walk 1982-1983

Sabael, New York

After her homesteading years, Linda's first home became the first public wild foods identification walk.

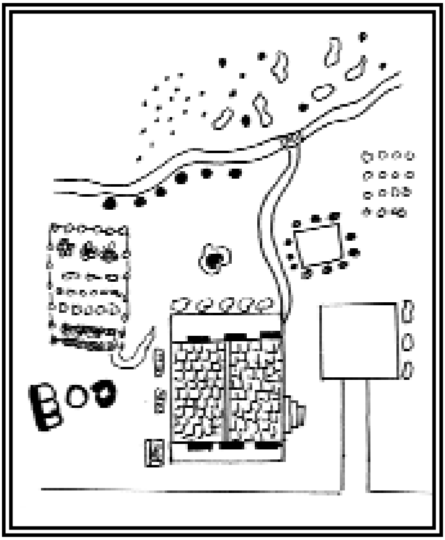

A completely natural Garden.

First year: Added topsoil, made raised beds using railroad ties. No fertilizing done. Rocked the natural areas, built bridges over swamp into woods. Peat mossed individual plants, as well as used heavy hay over winter.
Second Year: Complete regrowth of every species except sparse wintergreen patch. Presence of red dots on leaves indicated overly acidic soil because of balsam trees close by (natural to area). All beds peat mossed for second summer.

Linda says- I loved this first public walk around my cottage where I had electricity for the first time in 13 years. I maintained this walk for 2 years and then moved it to Rose & Bruce Burke's. This walk was studied by over 3000 people.

Burke's Cottage Garden 1983-1985

Sabael, New York

First Year: Rototilled a strip of lawn and transferred plants to rock-ringed sections. Added 6 inches of peat moss after one month's growth. Some harvesting and thinning necessary.
Second Year: Regrowth fantastic-100%.
Third Year: Good regrowth in most beds; poor regrowth of grape, meadowsweet, cattail (unattended).
Fourth Year: Plants assimilated back into mowed lawn.

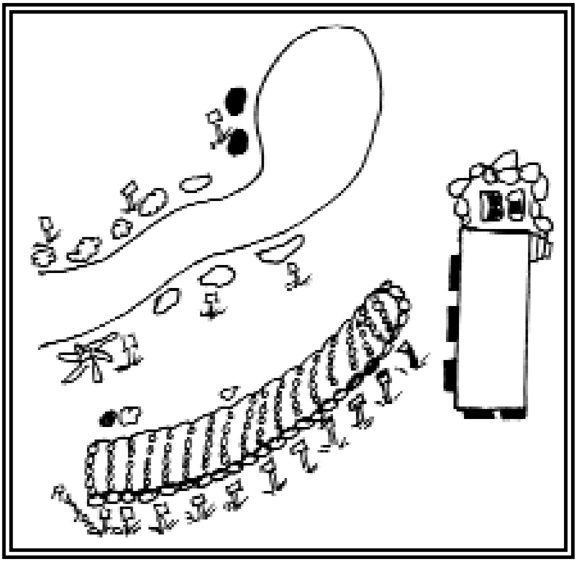

Burke's Cottage Garden.

Linda says- Down the street from my home, Mr. and Mrs. Burke gave me an area to build a wild food walk. This was well marked and I gave a "tour" twice daily for two summers, before moving on to the Runyon Institute.

Note: All trees planted continued growth: birch, pine or maple. The rest returned to lawn. This walk was seen by over 3000 people.

Runyon Institute Walk

Warrensburg, New York 1986

The Runyon Institute ran for one year in Warrensburg, New York. This was the third food identification walk developed for public use. The strip was made up of common lawn plants, and the walk demonstrated the area's natural habitat.

First Year: Rototilled the strip of lawn. Added topsoil, gravel walkways, wood 2x4's in between plants transplanted from natural areas which exhibited rapid growth. Peat moss added on successive harvesting of most sections. No fertilizers used. Assimilated into lawn after first year.

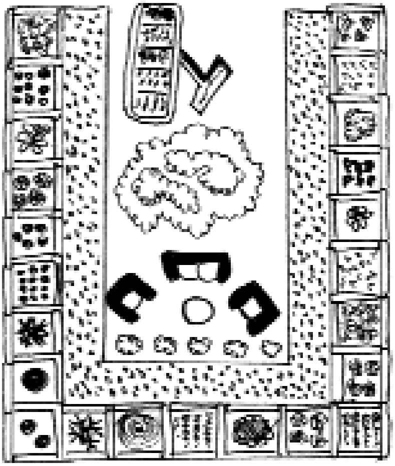

Runyon Institute Walk.

Linda says- The Runyon Institute was founded by a businessman named Walter Johnson. The building was founded in my name, and it hosted over 6,000 people, showing them our wild food walk. Unfortunately, I needed to close the institute and move closer to home in North Creek where the white water rafting walks began. The Runyon Institute walk was maintained for two years. Hundreds walked the graveled paths. Trees remained and prospered.

Nantikote Lenni Lenape Walk

Bridgeton, New Jersey 1987-1988

The Nantikote Lenni Lenape developed this walk in 1987. Research done here has been invaluable for developing future food walks.

First Year: Rototilled a strip of lawn in Bridgeton City Park. Added garden topsoil. Tepee wood stays used for wheel design and plant separation. Applied 6 inches of peat moss after 2 to 3 weeks growth.
Second Year: Unattended garden prolific with growth. Largest leaf and seedling growth of all walks. Assimilated back into park after third year. I couldn't attend this one personally but once every 2 months.

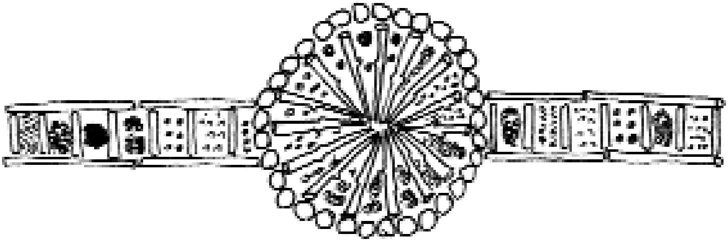

Nantikoke Lenni Lenape Walk.

Linda says- I was asked to consider a wild food walk in Bridgeton City Park. This was Native American land, and I was pleased to work with Jim Ridgeway as well as other members of the Nantikote Lenni Lenape tribe. This walk was graciously maintained for 2 years by the Nantikote Lenni Lenapi Indian people. It closed in 1988.

"Rafting the Glenn" Walk

Warrensburg, New York 1985

First Year: Put peat moss 6 inches thick around transplanted seedlings. Natural watering. Occasionally harvested smaller "lawn foods." No fertilizers added. Some trimming done of raspberry, blackberry, roses.

Second Year: Unattended except to weed patches, add 6 inches more peat moss. All plants survived winter except for birch, pine, and cattails. Replanted more lamb's quarters.

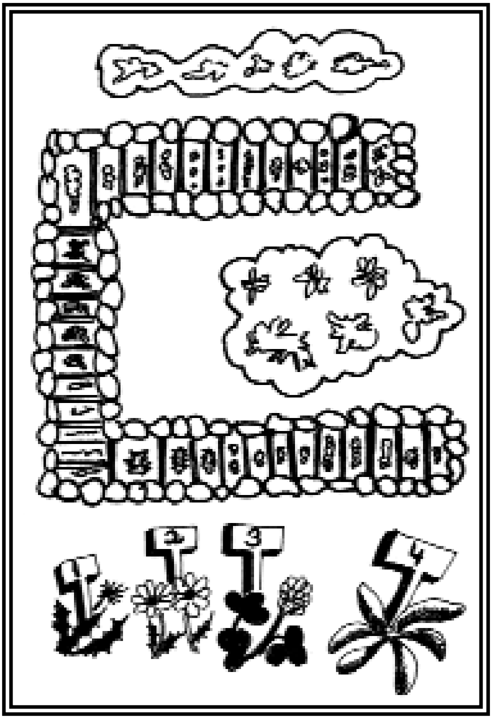

"Rafting the Glenn" Walk.

Linda says- Many of my students were young campers from white water rafting company. I particularly enjoyed the river a few yards away.

McManus Wild Food Walk

Sun City, Arizona 1990

First Year: Soil prepared to deal withcaliche (desert soil). Added in equal quantities peat moss, compost, topsoil. Used raised beds, wood sides. Plants easily transplanted with spoon and trowel. Irrigated with bubbler and hose twice a month, otherwise unattended.
Second Year: Unattended; temperatures of 110 degrees for two months. Harvested remaining plants, seedlings, and seeds. Strip assimilated into neighboring areas.

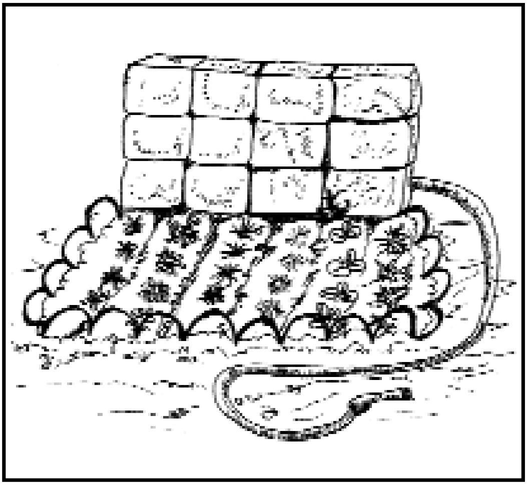

McManus Walk.

Apprentice Hazel McManus.

Linda says- A gourmet cook, Hazel of Sun City, Arizona, hosted many wild food luncheons for me. Our local television stations covered these wild food functions.

Glendale Library Walk

City of Glendale, Arizona 1995-1997

David Schultz, water conservationist, was instrumental in the planning and maintenance of this walk, working along with Linda Runyon.

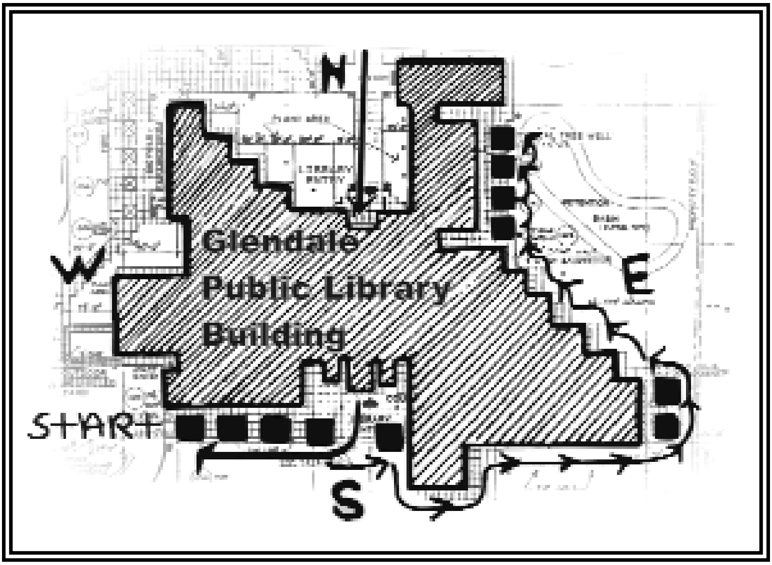

Glendale Public Library Walk.

The public visited 11 squares open in the pavement each 10' x 10', and boasting 101 wild plants and grasses in total. An underground sprinkler system was used to maintain and control this mammoth "desert" walk of wild edibles. The squares were filled with potting soil and fine desert soil and laced with peat moss. Everything was kept natural, even the use of an insect deterrent made from gigantic batches of garlic, olive oil and dish soap. Mixed in equal parts of one third each, this concoction kept the Arizona whitefly from eating the succulent new young seedlings.

Linda gives a tour.

This walk had well planned sign information using a color coded safety system: Yellow—use caution, Red—use BIG caution, and green—may eat freely after individual testing. The signs also included botanical and common names, as well as other pertinent information on edibility.

Drilling into the hard Arizona soil.
Left: John Hunkele, right: Richard Bond.

My hat is off to the City of Glendale and David Schultz. Many factors contributed to the now closed project. My hope is to bring my experience to others.

Wild Food Plant Walk

Greenwich, New Jersey, at the Children of the Earth school, started 2004

First Year: Plants were easily potted separately in native soil. The size of the pots varied to accommodate anticipated eventual growth. (Mullein grew to be 6 feet tall!) All plants were clipped and eaten from time to time, especially when a Wild Food Walk Lecture was in progress. Children kept the plants watered.

Second Year: The pots were transported to my field for the winter, and were covered in snow from time to time.

Third Year: The plants were abundant, and even more dense than the first year. Seeds fell and the soil was full of each individual plant. The Walk was used by the school for another summer. The plants wintered with the owners of the school in southern New Jersey. As of this writing (July, 2007), the plants have grown tall and quickly in the pots, just like the second year.

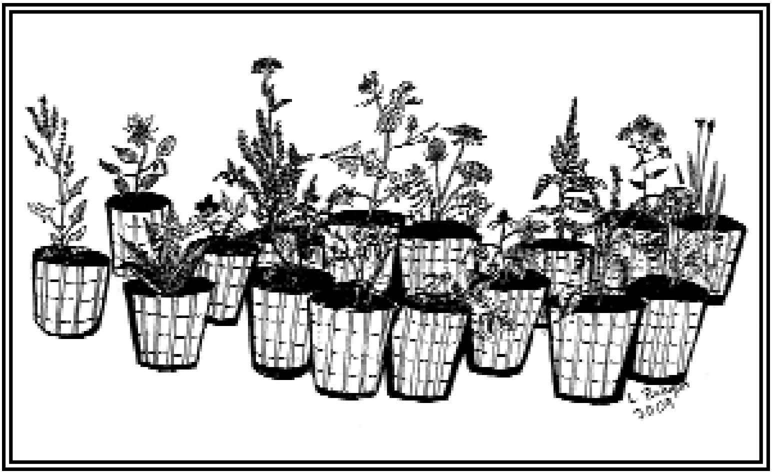

New Jersey Walk

Fordville, New Jersey, 2005-2006

First Year: This very elaborate walk required several truckloads of 4 x 4's (cut into 2 x 2's) and several loads of mulch to complete. The entire walk was 75 feet long, Because the New Jersey farm soil is so fertile, individual plants became extremely lush, spilling over in some sections and being difficult to maintain. Mullein grew to a height of 7 to 8 feet, and the Queen Anne's lace flowers were the size of cup saucers. I was by myself on this one and found it so overwhelming that I closed it after the first year. However, were there to be a crew of caretakers, such as scouts or nature park people, this type of identification center could see thousands of interested visitors.

Second Year Research: I went back to the walk, and saw that everything was coming back up.

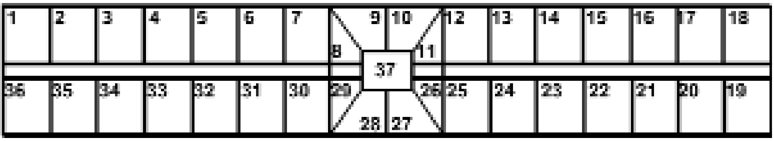

1. White Pine
2. Aloe Vera
3. Amaranth
4. Burdock
5. Chamomile
6. Chickweed
7. Chickory
8. Mullein
9. Clover
10. Plantain
11. Dandelion
12. Cattail
13. Evening Primrose
14. Queen Anne’s Lace
15. Curlydock
16. Wild Pepper
17. Sorrel
18. Blue Lettuce
19. Sassafras
20. Maple Tree
21. Nutsedge Grass
22. Dried Barley Grass
23. Crabgrass
24. Yarrow
25. Thistle
26. Goldenrod
27. Violets
28. Daylily
29. Sunflowers
30. Nettles
31. Mustard
32. Mints
33. Milkweed
34. Malva Neglecta
35. Lambs Quarters
36. Daisy
37. Willow Tree

Wild Foods Indoors

A house, apartment, roof garden, garage, or balcony are excellent areas to grow wild plants. Suitable containers for these indoor plants vary widely. Clay or plastic pots, pails, and washtubs all make ideal planters. You may even use half-gallon cans spaced evenly with a board placed on top to make an attractive shelf. There are many good books on container gardening and on growing plants indoors. Wild plants adapt easily to these ideas. Put your plants in areas of the house according to their outdoor environment. Many of them like water; the bathroom has steam, for example. Plants that like cold can thrive in a basement. I found that thyme loves heat and sun, dandelion loves the basement, near a window. Mint loves steamy areas.

The following are some plants that can be grown indoors:

Blackberry
Blueberry
Bulrushes
Daisies
Lamb's quarters (western variety)
Mallow
Phragmities
Purslane
Queen Anne's Lace
Raspberry
Roses
Sheep Sorrel
Shepherd's Purse
Tumbleweed
Violets
Yarrow

Use soil that is appropriate for the individual plants, with perhaps some rich potting soil added. Provide the suitable temperature and moisture level, and the plant will flourish. Ruth Spring, from Indian Lake, taught me to bring in dandelion, chicory, and wild lettuce for the winter. I planted them whole with their corresponding soil.

A Modest Proposal

Over 70 percent of the people in the United States live in an urban environment. More and more, urban parks are falling victim to surface pollutants, smog, ground water contamination, and, in some areas, apathy. It is harder and harder to find a clean urban area.

Potential sites for wild foods include fields, garden centers and private nurseries, environmental centers, camps, Boy or Girl Scout Centers, outdoor recreational centers and state or national parks and churches. The empty field in the middle of a city block is perfect for an urban garden. Waste areas seem to already have almost all the wild foods necessary. Add a few park benches, perhaps a water fountain, and a sandbox for the children. Plant a rototilled swatch with seedlings from nearby, edge it with wood or stones, and you have an instant identification center.

About 85 million people are added to the world population each year.[1] According to the U.S. Census Bureau, the global population doubled between 1959 and 1999, reaching 6 billion. Despite success in increasing global food production, the number of hungry people is growing. It is now thought to be 950 million—almost one fifth the world's population is underfed. Poverty is the cause. Every 24 hours, 25,000 human beings die as a result of hunger and starvation, 24 every minute, 18 of whom are children under five years of age. Acute human exposure to pesticides (such as that experienced by unprotected farm workers) is frequently fatal; 3,000-20,000 people die from pesticide poisoning annually. Human intervention to repair degraded lands takes two forms: restoration and rehabilitation. Of the two, only restoration aims to return a site to its natural state, complete with all the species that existed there before human disturbances. Rehabilitation is much more utilitarian. Its goal is to make the land productive for human use, employing whatever species and techniques which are most effective, regardless of whether they are indigenous or not. In the developing world, rehabilitation is likely to have a more prominent role than restoration because of the needs of growing population.

Perhaps the word *permaculture* is applicable here. What is permaculture? It is an attitude, a way of seeing what is before one's eyes, of enabling. It is a science, or study of natural systems such as water and plants. Permaculture focuses on *sustainable systems*, those with no pollution or waste. Permaculture is also an *ethic*. All systems of human activities support the earth's ability to sustain a diversity of life. Humans interact with their environment in mutually beneficial ways. Lastly, permaculture is a lifestyle that channels this knowledge and ethic into appropriate activities.

---

[1] World Resources, the World Resource Institute published by Oxford University, 1990.

Using wild food plants to restore the earth is the best way an individual can help at the local level. With very little help from us, the earth restores our food supply.

The urban food field is but a drawing board away. Wild foods can easily be incorporated into our educational system. Curriculum, books, tapes, videos, and newspaper articles as well as TV programs can bring this concept to the public. The environmentarian way of life is being born again!

A Forager's Research Diary

There are large gaps in our knowledge of wild foods, especially in how to incorporate them into our lives. I'm always conducting experiments with these plants, seeing how well they do in a garden, whether there are new ways to use these traditional foods, and how potent the seeds remain.

The following are my notes on such experiments, along with material from the work done by my students. Growing and using wild foods remains unexplored territory.

I hope to encourage you to do your own experimentation with wild and backyard foods—not just with growing and harvesting, but in creating interesting menus for you and your family.

Purslane Experiments

Question: will dried purslane reconstitute in water? If so, will it grow and sprout roots? Can it be used reconstituted for food? Dried purslane will reconstitute well soaked in water. The plant will not sprout roots as it grows from seed. Reconstituted for food only.

August 1	Picked a large amount of purslane. Filled a screen 1 by 3 foot. Placed screen in an inside open window receiving only afternoon sun (2:00 p.m. to 4:00 p.m.).
August 3	Purslane roots were cut off. Black seeds from the center of the leaf clusters dropped onto the screen. There are only a few green pods in the center of the leaves still filled with seeds. Pods were apparently immature at time of picking.
September 1	I cut off the dried, withered stem of a twig about 7 inches long and placed in water. The leaf part opened and filled with water, although the branch did not appear to grow any quicker than the one on the screen.
September 13	Purslane in water growing well. Stems and leaves on the screen are dried and crumbly. On top of the purslane mat and up through the dried stems a completely new growth is visible. On close exami-

nation, I saw a 7-inch stem with five extensions off the main stem. Healthy green leaves extend off tips of the top three branches, with 3 tiny leaf buds appearing between the leaf extensions and joints of stem. Average size leaf is 1/2 inch. Lower two stems are weak, partially dry, flaccid at base. Small leaves extend on the tips, green and apparently healthy. All leaves are thinner than normal, stretched toward the light and do not have the succulent appearance of purslane growing in a field. Taste same. Stems are crispy and solid, about 3 1/2 to 4 inches from tip to the base. The 4 inches of base stem are apparently weak, discolored reddish, deeply veined, devoid of succulence.

September 21 Clipped 1/4 inch off the thicker, central stem. Placed in a cup of water. Cells at the end look green and open under magnification. Stems are tightly veined and devoid of moisture. Color basically a red hue with a touch of green where the stem lay against the bottom plants on the screen. Leaves small, 1/2 inch as a norm, growing green and alert toward the western light of the window. Multiple branches tipped with green leaves have one or two small branches with small green leaves. Two necrotic twigs cut off cleanly with scissors. Cup, water, and twig placed out of direct light of window at 7:00 a.m.

Conclusion: Purslane dries in several days or weeks, according to moisture in the air. New growth can be seen throughout as old growth dries. Roots take in moisture from air and mat of purslane itself.

It appears that old dried purslane may be easily reconstituted in water, yielding edible food. In drought conditions, plant may be utilized with small addition of water. Purslane appears to be one of the hardiest of plants; after two months of drying, it was easily reconstituted in water.

Amaranth Experiments by Althea Dixon, student

Background: In 1992 Arizona experienced an unusually large rainfall. By May 1, we had received all the rain we usually get in a year. We had so much moisture in the adobe soil that mold was killing some of the clover and mushrooms were sprouting up. Amaranth appeared in the goat pen, thanks to these heavy rains which led to some experimentation.

August 13 I cut the first harvest from plants averaging 8 to 16 inches in height. Because of the abundant rain and rich soil, they grow with an umbrella of leaves, creating a mini-rain forest environment with lots of seedlings beneath. I remove the tops down to 4 inches above the ground.

August 17 The amaranth crop has grown from 8 to 12 inches in just four days!

August 18 Harvest two plastic grocery sacks full of tender amaranth tops. Crop is harvested to about 12" inches. Total fresh yield is 16.5 pounds.

After the harvest I notice that the soil is still well underneath, and the crop is growing without noticeable red on the stems, which is good. I can get another crop or two from this patch.

I also notice that amaranth growing in the direct sunlight does better than that growing in shade, the latter attacked by bugs causing leaf damage.

August 19 Dryers yielded 1.75 pounds of leaves and stems. After the major crop is dried, heavier stalks still have bulk and moisture enough to mold the rest of the batch. I cut stalks out with scissors and set them in an airy box for another week to finish drying. Further drying of whole batch would damage the already dry tender parts.

The high humidity is softening the tender leaves. I may have to put them in the dehydrator for an hour before whizzing them to flour. I don't know what effect this will have on nutritional value, but the color remains excellent.

August 20 Last night we had a heavy rain. I checked the progress of the amaranth and observed that undergrowth is now up to 12 inches. The day-old seedlings that were about 1/8 inch cotyledons are now between 1" and 2" seedlings, with primary leaves approximately 3/4 inches from stem to tip of the leaf.

August 24 Heavy rains over weekend. The cotyledons and primaries now have tertiary leaves and are approximately 3 inches tall. The undergrowth is now 12 inches high.

August 26 Harvested one shopping bag full. Do not weigh the green crop. Dryer yields 8 ounces of flour. By volume it makes 1 2/3 cups. Have lost only a small amount of chaff, which I sifted out of the flour as it comes from the high-impact grinder.

August 27 No rain since 8/24 in A.M. Crop has leveled off. Undergrowth continues to catch up with umbrella top at approximately 30 inches.

I harvest two more bags of tender tops this morning. These bags were the same volume as the batch on 8/18, however, they weigh only 6.5 and 5.5 pounds respectively. I fill the same number of dryer trays, but net weight of crop in dryers is 10.5 pounds. I use 1/2 pound of large stalks to make "beans."

The 4-inch seedlings are now averaging 7 to 9 inches tall with no further rain or water since 8/23. After harvesting four shopping bags of vegetable in four days, it appears there are at least 2 to 4 more harvests with no further need of water. Have not harvested red-stemmed plants; these are for seed.

August 28 Dry yield is 2 pounds 2 ounces, or 5 cups. Discarded 1 ounce of chaff plus one stem that had not dried Also lost powder in the air as it was moved from one container to another, as well as leaf fragments too small to handle.

September 12 After running it through high-impact grinder, I get 2 quarts of flour. Dry weight is 2.13 pounds, of which 1/4 inch of heavy stem chaff is discarded.

It is interesting to note that although the heavier stems had dried hard, the humidity had gone from 20 percent to 46 percent during the wet season in Arizona, and the batch gained moisture overall. Though the texture is less crisp (the leaves would bend without breaking), I have more trouble with chibbled leaves. The color is still excellent, but more on the olive side, which I believe is due to a higher percentage of stem content in the end product. Stems dry brown whereas leaves dry green.

Conclusions:

The best procedure for harvesting to flour is as follows: Bring crop directly from garden to dryers (wash the night before with gentle water from hose) and let dry 24 hours. Rearrange trays about halfway through if dryer is old-fashioned kind with no heat or air controls. When putting crop into dryers, trim stems thicker than 1/4 inch and cook as fresh beans. This way crop is completely dry at the same time. Place directly into high-impact grinder from dryer trays; doesn't hurt if still warm. Strain chaff with a fine sieve and set to one side. Rerun the chaff all together and then sift again. What little chaff is left can be thrown away. Put flour in tightly sealed glass jars. Wrap in foil to protect from the light, and label with name of product and date. Keep in cool dry place.

Time: It takes 45 minutes to process 18 dryer trays from dried leaves to vegetable powder. It takes 30 to 50 minutes to harvest two grocery bags of fresh vegetables.

Caution: After plants form seed pods, use gloves to harvest and process flour. When processing indoors, protect lungs from flour in air by using a mask.

Reconstituted Amaranth Leaves: Can dried leaves be reconstituted to produce a palatable food item? Yes. Dried green leaves were soaked in cool distilled water for one hour, and then simmered until tender, about 5 minutes. I used only enough water to cover the leaves. The texture was so good it was almost like fresh.

Additional Comments: Was not satisfied drying amaranth outdoors. One crop got rained on and the vibrant green leaves turned sickly gray. In a survival situation you do what you can, but if domesticating amaranth, why not get the best-quality, highest nutritional product possible?

The reason for not putting amaranth in water is that if you have to leave it for any reason it can very quickly take on a dark gray color from the iron, and the quality of the dried product is not appetizing.

Further Comments: In September, 1993, the amaranth flour was inspected for color. It had remained bright green, and flavor in breads was excellent. For even longer storage, vacuum pack in plastic and store in cool dark place.

Grinding greens to flour.

Lamb's Quarters for flour.

Notes

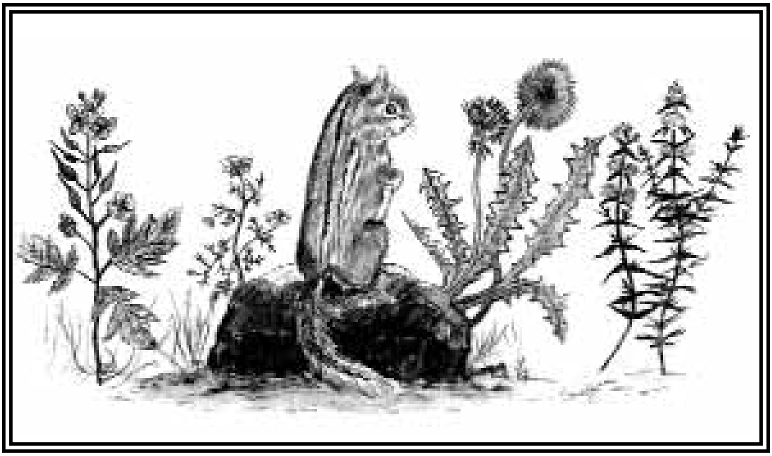

Left to right: wild mustard, chamomile, dandelion and thyme.

Part V: Poisonous Look-Alikes

Here are a few troublesome plants for wild food foragers to be aware of. These are by no means the only poisonous plants out there, but they are some of the most common and are most easily confused with the plants that have been covered in the edible plants section of this book.

See also "An Introduction to Edible Plants" in Part I, especially the section on picking wild foods (page 22).

Caution: It is sometimes not the close up that is confusing, but the same plants identified at a distance can fool you.

Please Note: There are many varieties of poison hemlock, geographically spaced in all countries. Please note all species have similar characteristics. Smooth purple striated or blotched stems. Flower clusters on single stems from main stem. Hollow, circulated banded roots. Always be sure of 100% identification.
Cross reference:
Queen Anne's Lace with any white top-flowered umbrella.

Water Hemlock—Poison

Spotted Cowbane
Circuta maculata

Habitat: Wetlands, wet meadows, swamps. Canada south to Texas. This fact was observed some years ago. Do not take the state perimeter too literally.
Characteristics: The MOST DEADLY plant. Any part will kill humans. White flowers, clustered in umbrella on single stems. Stem is smooth, streaked with purple. Stem, roots chambered. (See page 292 for illustration.) All parts may even simulate smell of mild parsnip or carrot.
Caution: DEADLY POISON.

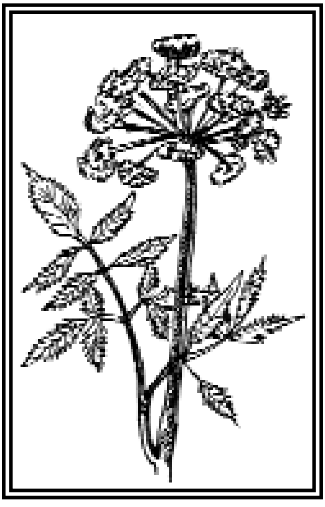

Poisonous - Water Hemlock

Poison Hemlock—Poison ☠

Conium maculatum

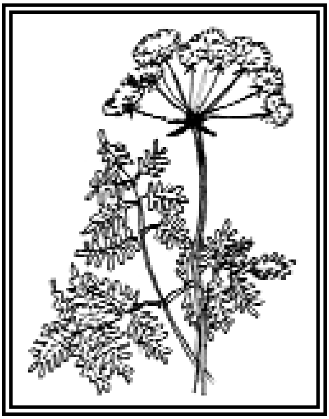

Poisonous - Poison Hemlock

Habitat: Waste ground, Iowa to Quebec South, approximately. Other genus or types on West Coast, even wet desert areas.
Characteristics: Blooms May through August. Umbrella shaped flower cluster, lacy white flowers. Dried flower maintains its umbrella shape, develops rusty colored seeds. Stalks are hairless, stout, hollow and grooved with spots of purple stripes. Small amounts may cause paralysis and death. DEADLY. Roots look like a wild carrot, but are hollow. A deadly poison. (See illustration next page.)
Caution: Poison Hemlock Looks like Queen Anne's lace until flowers dry. Use 3 excellent picture references and make sure you identify Hemlock properly and DO NOT TOUCH OR INGEST. HEMLOCK IS DEADLY. (See illustration of Poison Hemlock.)

Queen Anne's Lace—Food ☺

Daucus carota
Wild Carrot

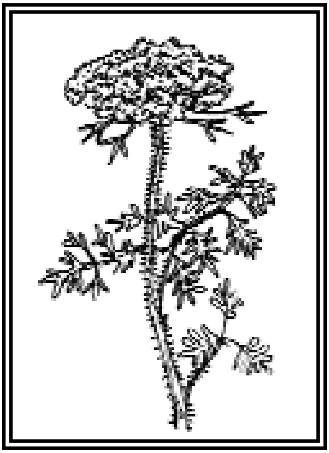

Edible - Queen Anne's Lace

Habitat: Waste ground fields
Characteristics: HAIRY-STEMMED biennial. Umbrella shaped flower cluster, lacy white flowers. All parts smell like a carrot. Old flower clusters "bird cage" into a seed cluster. Parsley-like leaves smell like a carrot. All stems are hairy, the root is SOLID & pithy. Smells like a carrot! (Also see page 116 in Field Guide for further description of this edible plant.)

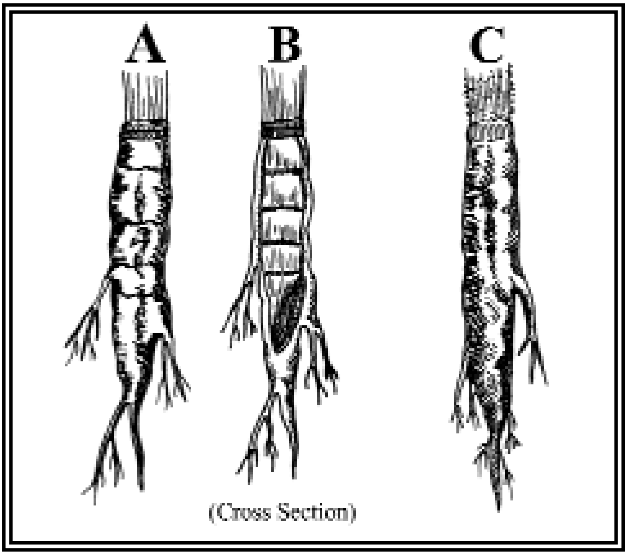

A: Poison Hemlock Root

B: Poison Hemlock root in cross section

C: Queen Anne's Lace root – food

Please note: Poison Hemlock (A and B) has hollow chambers with rings at the base of the stem and throughout the root. Queen Anne's Lace (C, on right) has a solid twig-like pithy root with hairy stems.

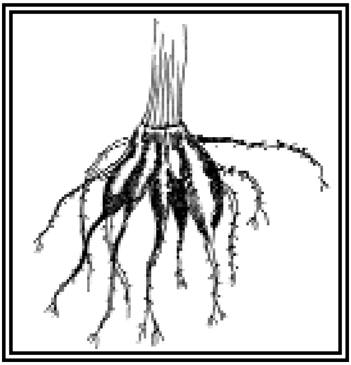

Poison - Water Hemlock Root.

Poison Sumac—Poison

Rhus vermix

☠

Habitat: Wooded swamps.
Characteristics: Leaves, 6-12 inches long with 7-13 points, leaflets. First difference between sumac for food and poison sumac is that the poison variety has smooth leaflets. Poison berries are white. (See page 126 for description of the edible variety.)

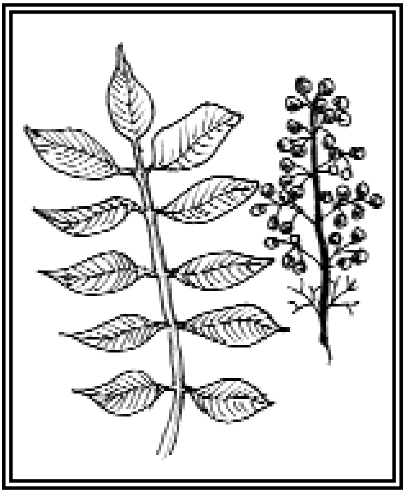

Poisonous - Poison Sumac

Edible Sumac—Food

Rhus Species

☺

Habitat: Shrub or small tree found in all areas of the United States.
Characteristics: Twigs have milky sap. Fruit is red, hairy, and dry at maturity. Leaves: saw-tooth leaflets. (See page 126 for details.)

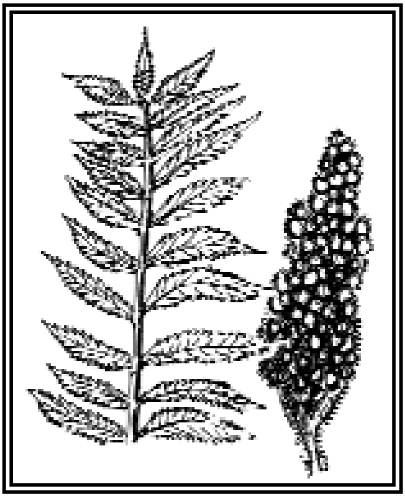

Edible - Sumac

Nettle Leaf Goosefoot—Poison

Goosefoot family: Chenopodicea

History: Naturalized from Europe.
Habitat: Moist soil, roadsides, around crops such as alfalfa, citrus, vegetables.
Characteristics: Grows to height of 1 to 3 feet, smells rank.
Caution: All raw edible parts are toxic or poison to humans.

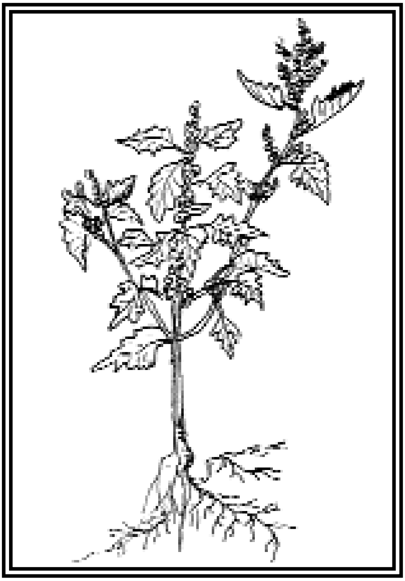

Poisonous - Nettle Leaf Goosefoot

Lamb's Quarters—Food

Goosefoot family: Chenopodicea

History: Naturalized from Europe.
Habitat: Moist soil, roadsides, gardens, crops such as alfalfa, citrus, vegetables.
Characteristic: Grows to height of 1 to 3 feet. Smells similar to spinach. Leaves are soft and have a white powder underneath. All parts edible. Good food. (See page 93 for complete description.)

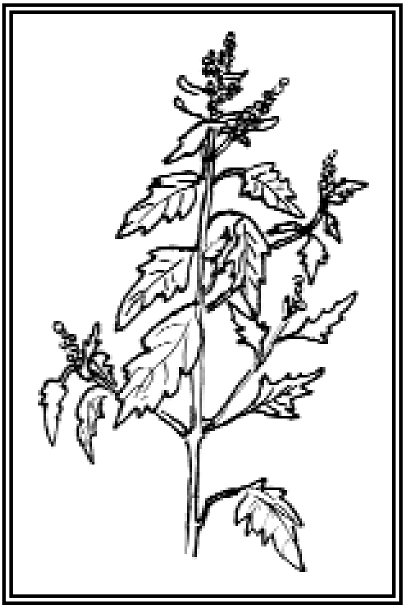

Edible - Lamb's Quarters

Field Horsetail—Poison

Equisetum Arvense ☠

History: Family over 400 million years old.
Habitat: Wet, moist areas, swamps, lakesides, river banks.
Characteristics: Jointed green, like grass. All parts contain silicon; are toxic to humans. Abrasive silica in all parts, so plants will scrub aluminum clean.
Medicinal Value: Folk medicine wound healer, diuretic in very small amounts.

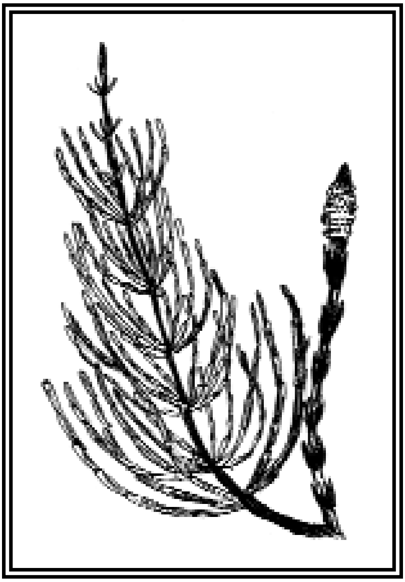

Poisonous - Field Horsetail

Grass—Food

Graminae ☺

Characteristics: Curls and grows up like horsetail. Horsetail grows in longer grass areas. We need to examine all types of grasses carefully when choosing grass to be clipped for food. Pull out any field horsetail. (See section beginning on page 35 for complete description.)

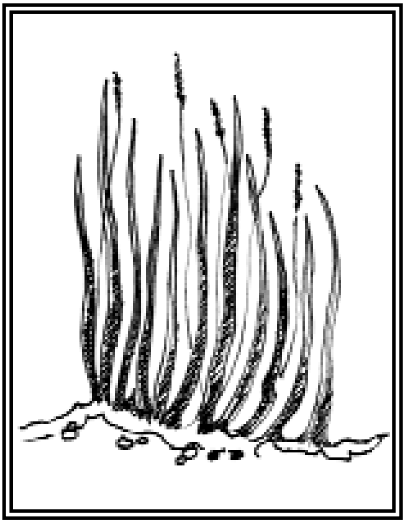

Edible - Grass

St. Johnswort—Poison

Hypericareae ☠

Other Names:
Hypericum perforatum
History: Known for medicinal value since biblical days. Europe, herbalists.
Habitat: Fields, roadsides.
Primary Use: Leaves used as astringent and salves for bruises, wounds, anti-inflammatory.
Characteristics: Leaves shaped like lungs with many tiny holes in them. Plant grows to 1 1/2 to 3 feet. 5-petaled flowers. From a distance, plant shape is similar to the pocomoonshine variety of goldenrod.

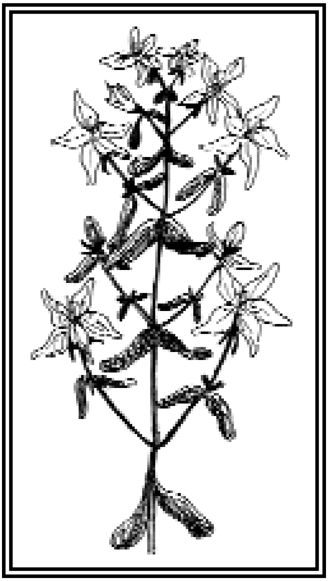

Poisonous - St. Johnswort

Pocomoonshine—Food

☺

(Variety of Goldenrod)
Augustifolum (specific species in the *Compositae* family)

Other Names: Poco moonshine
History: Native to Europe, Asia. Naturalized in North America.
Primary Uses: Culinary, medicinal, commercial. Leaves dried for tea. Oils and extract used commercially in baking, ice cream, soft drinks, chewing gum.
Characteristics: Flowers at end of many branches 6 to 18 inches, considered an herb. The smell and uses are the same as other Canadian or sweet goldenrod species in field guide. See Goldenrod, page 90, for complete description.

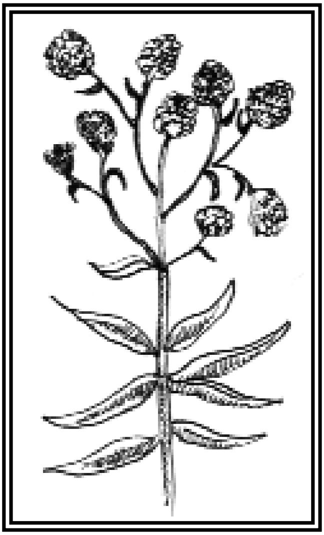

Edible - Pocomoonshine

Spreading Dogbane—Poison

☠

Apocynum species

Habitat: Thickets, roadsides. Characteristics: Perennial herb. Wide spreading branches, opposite leaves 2 to 3 inches long. Pink nodding bell-like flowers with pink stripe in center of each lobe. Shoots of Dogbane are hairless and quickly branching.

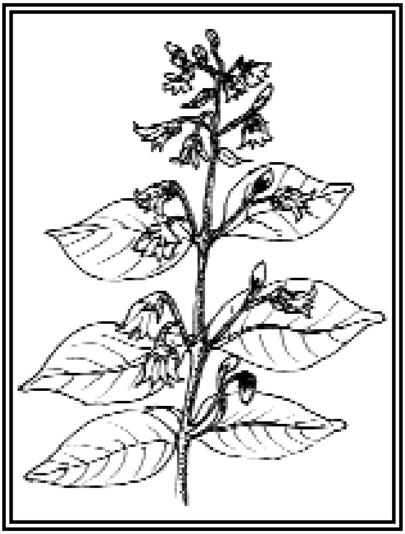

Poisonous - Spreading Dogbane

Common Milkweed—Food

☺

Aslepias Syriaca

Habitat: Fields, roadsides.
Characteristics: Downy stemmed shoots. Leaves from 3 to 6 inches, deeply veined. Flowers dome shaped, tight, drooping clusters. After flowering, buds give way to seed pods. Crescent shaped pods, when young, may be dipped in batter and deep fried or boiled. (See page 101 for culinary uses.)

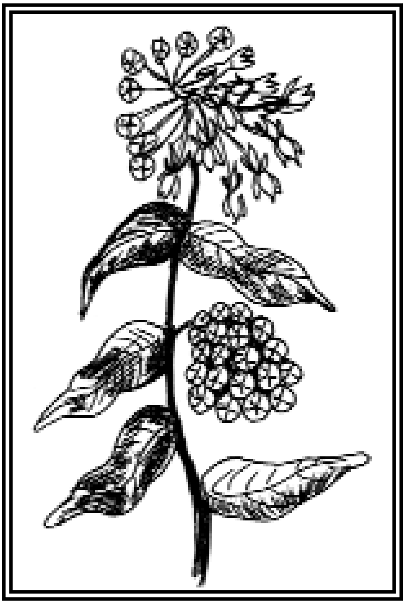

Edible - Common Milkweed

Dallis Grass—Poison

Gramineae Family

☠

History: Native to South America, found on North American continent in the west.
Habitat: Arizona
Primary Uses: None for humans.
Caution: This grass is very susceptible to fungus attacks by paspalum ergot which is a DEADLY POISON to cattle; humans.
Characteristic: Grows to height of 1 1/2 inches. Has tight large seeds placed in a way as to attract black mold easily.

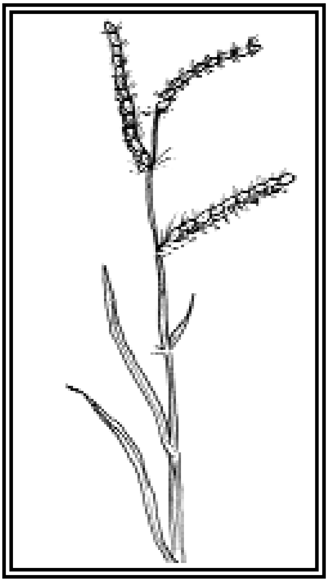

Poisonous - Dallis grass

Jungle Rice—Food

Echinochle colonum
Graminae Family

☺

Other Names: Water grass
History: Used for food in Asia, Africa.
Habitat: Fertile desert soils of the West around crops such as sorghum, cotton, alfalfa, melons, lawns.
Characteristics: Grows 1 to 1 1/2 inches high. A sedge, but dark spots of camouflage. Areas of stems are jointed.

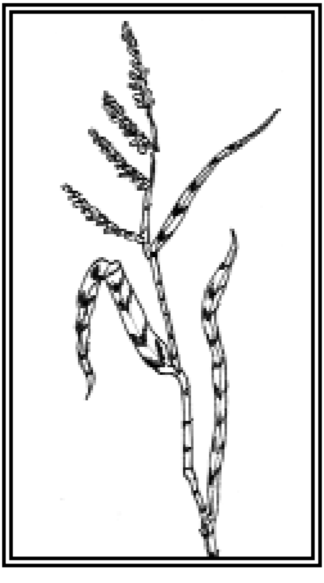

Edible - Jungle Rice

Stink Grass—Poison

Gramineae Family

☠

Other Names: Lovegrass
History: Native to Europe.
Habitat: Found growing around base of crops such as Sorghum, Alfalfa, Citrus, gardens and waste areas.
Characteristics: Annual grows to heights of 4 inches to 2 feet. Whole plant has a cockroach like odor, rank and offensive. Flowering parts 2 to 10 inches long. Poison.

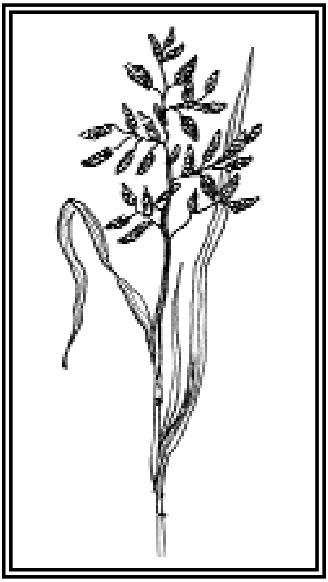

Poisonous - Stink Grass

Yellow Nutsedge—Food

Cyperus esculentus
Sedge family

☺

Other Names: Yellow nutgrass.
History: Introduced from Europe.
Habitat: Cultivated fields, edges of crops, wet areas, ditches, roadside, streams.
Characteristics: Umbrella-like flowering tops have a thread-like projection extending from seeds and flowers. Roots have small nut-like tubers which can be used as a highly nutritious food. (See page 35 for complete description.)

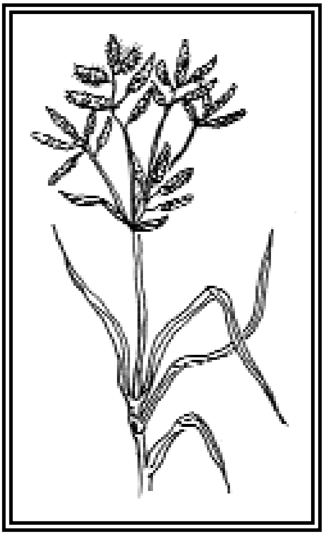

Edible - Yellow Nutsedge

Purslane—Food

Portulaca oleracea ☺

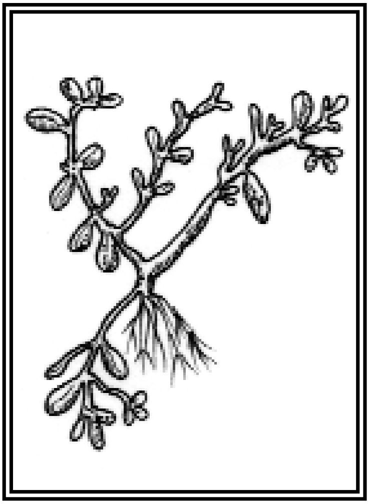

Edible - Purslane

Note: Purslane can be confused with White Margin Spurge, Hyssopspurge or Puncture Vine.
Habitat: Rich, sandy soils, waste places.
Characteristics: Prostrate paddle-shaped succulent leaves, tiny 5-petaled yellow flowers open in sunlight. Purslane has no milk. Leaves and stems succulent and make delicious food. Plant has a shallow tap root. (See page 115 for details.)

Please Note: The following 3 plants are look-alikes to Purslane. When viewed from a stand height, the look of the ground cover is similar for all of these plants.

White Margin Spurge—Poison

Euphorbia albomarginata ☠

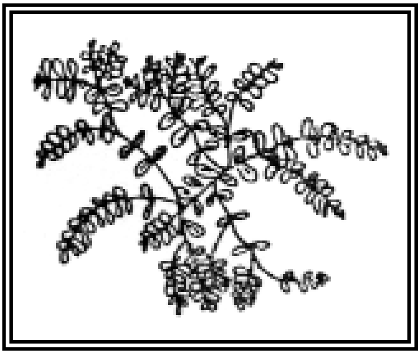

Poison - White Margin Spurge

Other Names: Rattlesnake weed
Habitat: Dry barren dirt areas along sidewalks, waste places.
Characteristics: All parts toxic, whitish milk in stem, leaves. Red blotches in each leaf.

Hyssopspurge—Poison

☠

Euphorbia hyssopifolia

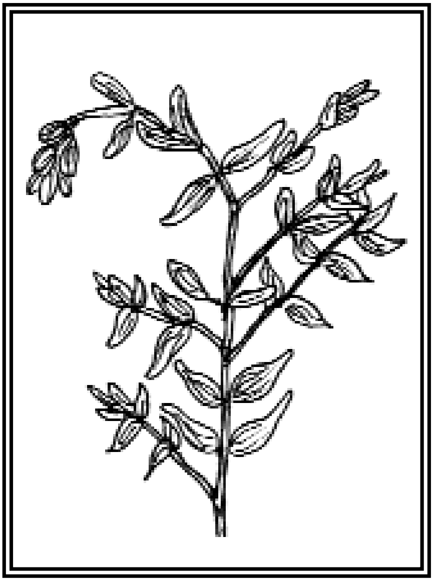

Poisonous - Hyssopspurge

Habitat: Grass, fields, crops, irrigated areas.
Characteristics: Annual. Toxic plant contains silicon in all parts. Has a white sticky milk. Grows erect, stems hairless.

Puncture Vine—Poison

☠

Caltrop zygophyllacea

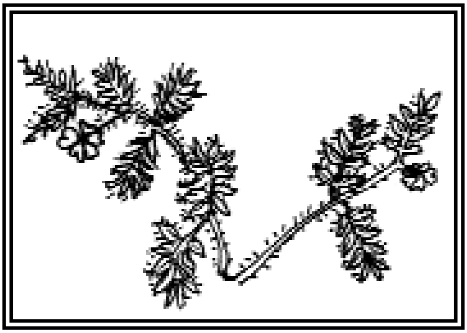

Poisonous - Puncture Vine

Other Names: Bullhead, goathead
History: Native to Europe.
Habitat: Roadsides, gardens, fields, crops.
Characteristics: Prostrate plant has a shallow tap root, 5-petaled yellow flowers which open in sun. Seedpods contain a cluster of 5 spiny burs or nutlets that contain the seeds. Toxic to animals and humans.

Appendix: Glossary of Scientific & Botanical Terms

Alkaloid: Basic, bitter-tasting organic compound containing nitrogen and forming water-soluble salts and acids.

Anodyne: Soothes or allays pain.

Antiscorbutic: Prevents scurvy.

Antiseptic: Prevents infection; kills germs.

Anti-spasmodic: Calms nervous and muscular spasm.

Aphrodisiac: Increases sexual stimulation and excitement.

Astringent: Diminishes secretion, contracts tissues.

Areole: In cacti, a clearly defined small area that may bear felt, hair, spines, glocids, flowers, or new branches.

Bacteriostatic: Inhibits the growth of bacteria.

Balsamic: Soothes mucous inflammation, especially of respiratory and urinary tracts.

Calyx: Outer whorl of floral leaves (sepals), which may be separate or fused.

Carminative: Expels gas from the stomach and intestines.

Cathartic: Purgative.

Chibbled leaves: see Compound leaf.

Compound leaf: Divided into two or more leaflets. The leaflets can be further subdivided, twice compound, or even thrice compound.

Corm: Enlarged base of a stem; bulblike but solid, not layered like an onion.

Cotyledon: First two leaves emerging from embryo.

Decoction: Liquid prepared by boiling with water and straining the cold solution.

Demulcent: Protects and soothes the mucous membranes.

Diaphoretic: Increases perspiration.

Digestive: Aids digestion.

Diuretic: Increases the flow of urine.

Dysentery: Diarrhetic condition or disease.

Elixir: Alcoholic tincture with sugar.

Emetic: Substance that causes vomiting.

Emmenagogic: Brings on or regulates menstruation.

Emollient: Soothes and softens the skin.

Essential oil: Plant oil, usually scented, that evaporates at a low temperature. Also known as a volatile oil or ethereal oil.

Expectorant: Helps expel phlegm.

Extract: Dry, soft, or liquid preparation. A fluid extract may be prepared with alcohol, distilled water, or alcohol and water.

Febrifuge: Substance that lowers a high temperature or prevents fever.

Glocids: Thin barbed bristle, produced in the aereoles of chollas, prickly pears, and a few other cacti.

Glycoside: Compound combining sugar and nonsugar units.

Hemostatic: Arrests bleeding.

Hepatic: Beneficial to the liver.

Infusion: Liquid prepared by pouring on boiling water and later straining the cooled solution.

Laxative: Mild purgative.

Leaflet: Division of a compound leaf.

Macerate: Soak in liquid to soften and dissolve.

Mucilage: Gelatinous substance swelling, but not dissolving in water.

Narcotic: Induces sleep or drowsiness.

Nerve tonic, nervine: Stimulates the nervous system.

Ointment: Salve or unction for application to the skin.

Panicle: Loose, irregularly branched inflorescence with stalked individual flowers.

Pectoral: Relieves coughing and promotes expectoration.

Plaster: Preparation spread on material and stuck to the skin.

Potherb: Plant cooked as a green.

Poultice: Soft paste, hot or cold, wrapped in cloth and applied externally.

Pubescent: Covered with fine, soft hairs.

Salve: Unctuous adhesive substance applied to wounds or sores.

Sauté’: To fry quickly, stirring constantly, until lightly browned.

Seed head: Part of plant that holds a group of seeds.

Sedative: Having a calming and soothing effect.

Stomatic: Aids in digestion and stimulates the appetite.

Stimulant: Excites the functions of various organs.

Stir-fry: To fry quickly in hot oil over very high heat, stirring until food is cooked.

Styptic: Astringent, checks bleeding.

Sundries: Accessories

Tertiary: Third set of leaves of seedling; may or may not display adult attributes of plant.

Tincture: Extract of drug, usually dissolved in alcohol.
Tisane: Dilute plant infusion or tea.
Tonic: Stimulates the activity of an organ.
Tubers: Modified, swollen root.
Unction: Soothing or healing salve, ointment.
Volatile oil: See essential oil.
Winnow: To place seed heads in basket, and move up and down to allow the chaff (inedible parts of the grain) to fly out of the basket, leaving the seeds.
Whole-leaf tea: Tea made with the whole leaf, not crushed; also tea made with whole twig, not pulverized.

Edible Centerpiece

References

References in the text to Cherokee, Algonquin, and Chippewa words are based on transliterations from the *Cherokee-English Dictionary*, edited by William Pulete (Dallas: Southern Methodist University Press, 1975).

Bigfoot, Peter. *Arizona Wild Herbs*.

Densmore, Frances. *How Indians Use Wild Plants for Food, Medicine and Crafts*. New York: Dover, 1974.

Duke, James, and Alan Atchley. *CRC Handbook of Proximate Analyses*. Boca Raton: CRC Press, 1986.

Harrington, H. D. *Western Edible Wild Plants*. Albuquerque: University of New Mexico Press, 1967.

Harris, Charles. *Eat the Weeds*. New Canaan Ct.: Keats Publishing, 1973.

Hodgson, Wendy Caye. *Edible Native and Naturalized Plants of the Sonoran Desert North of Mexico*. Arizona State University: Arizona, 1982.

Page, Nancy M. and Richard E. Weaver, Jr. *Wild Plants in the City*. New York: Quadrangle, 1975.

Pedersen, Mark. *Nutritional Herbology*. Bountiful, Utah: Pedersen Publishing, 1987.

Smith, Huron. *Ethnobotany of the Ojibve Indians*, Vol. 4. Milwaukee, Wis. 1883.

Sturtevant, E. Lewis, Dr., *Sturtevant's Edible Plants of the World*. Edited by Dr. U. P. Hedrick, New York: Dover, 1972.

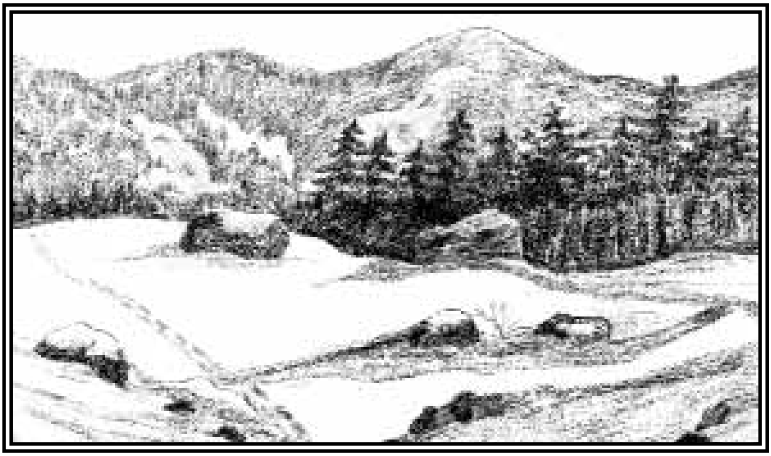

Suggested Reading

Agricultural Research Service. *Common Weeds of the United States.* U. S. Department of Agriculture, Washington, D.C.: Government Printing Office, 1971.

Airola, Paavo. *Are You Confused?* Phoenix: Health Plus, 1971.

Angier, Bradford. *Field Guide to Edible Wild Plants.* Harrisburg, PA, Stockpole Books: 1974

Ball, Jeff. *The Self Sufficient Suburban Garden.* Emmaus, Pa.: Rodale Press, 1983.

Birdsey, Clarence, and Eleanor G. Birdsey. *Growing Woodland Plants.* New York: Oxford University Press, 1951.

Briggs, Jim. *Wild Foods.* Hamilton County, N.Y. Cooperative Extensions publisher, 1972.

Brookes, John. *A Place in the Country.* London: Thames and Hudson, 1984.

Brown, Lauren. *Weeds in Winter.* New York: W. W. Norton & Co., 1976.

Chase, Agnes, and L. E. Bailey. *The First Book of Grasses,* Rural Textbook Series. New York: Macmillan, 1982.

Coon, Nelson. *The Dictionary of Useful Plants.* Emmaus, PA: Rodale Press, 1977.

Cooper, Hewitt Museum, the Smithsonian Institute's National Museum of Design, 712 Fifth Ave., New York, N.Y. 10019, 1979.

Crockett, James Underwood. *Greenhouse Gardening.* New York: Time-Life Books, 1989.

Densmore, Frances. *How the Indians Use Wild Plants for Food.* New York: Dover, 1974.

Duffield, Mary Rose, and Warren D. Jones. *Plants For Dry Climates.* New York: H. P. Books, 1981

Elliot, Douig. *Roots: An Underground Botany and Foragers Guide.* Old Greenwich, CT: Chatham Press, 1976.

Evers, Robert A. and Roger P. Link. *Poisonous Plants of the Midwest & Their Effects on Livestock.* Spec. Publ. #24, College of Agriculture, University of Illinois at Urbana-Champaign, 1972.

Fairbanks, Bert L. *A Principle with Promise,* Salt Lake City: Bookcraft, 1978.

Gabel, Medard. *Energy, Earth, and Everyone.* Garden City, N.Y.: Anchoe Books, 1980.

Gaertner, Erika E. *Harvest Without Planting.* Montreal: Concordia University Press, 1967.

Gibbons, Euell. *Stalking the Wild Asparagus.* New York: David McKay, 1978.

Gilbertie, Sal, with Larry Sheehan. *Home Gardening at Its Best*. New York: Macmillan, 1977.

Goldstein, Jerome. *The New Food Chain. An Organic Link Between the Farm and City*. Emmaus, PA: Rodale Press, 1973.

Hardin, James W. *Human Poisoning for Native and Cultivated Plants*. Second Edition. Durham: Duke University Press, 1974.

Harrington, H. D. *Western Edible Wild Plants*. Albuquerque: University of New Mexico Press, 1967.

Harris, Ben Charles. *Eat the Weeds*. New Canaan, CT: Keats Publishing, 1973.

Hitchcock, Susan Tyler. *Gather Ye Wild Things*. New York: Harper and Row, 1980.

Hobson, Phyllis. *Food Drying*. Charlotte, VT: Garden Way Publishing, 1983.

Kingsbury, John M. *Common Poisonous Plants*. New York State College of Agriculture and Life Sciences, New York, 1976.

Kirschenew, H. E. *Nature's Healing Grasses*. H. C. White Publications, 1960.

Lappe, Frances Moore. *World Hunger: Twelve Myths*, New York: Grove Press, 1986.

Lappe, Frances Moore. *Diet For A Small Planet*. New York: Ballantine Books, 1971.

Lerzaa, Catherine, and Michael Jacobson, eds. *Food For People, Not For Profit; a Sourcebook on the Food Crisis*. New York: Ballantine, 1975.

Luenscha, Walter Conrad Leopold. *Poisonous Plants of the United States*. New York: Macmillan, 1961.

Life Seed Foundation, Box 72, Port Townsend, WA 98368 or Native Seed Search, Tucson, Arizona

Messegue, Maurice. *Of Men and Plants*. New York: Macmillan, 1973.

Michael, Pamela. *All Good Things Around Us—A Cookbook and Guide to Wild Plants and Herbs*. New York: Holt, Rinehardt and Winston, 1980.

Miller, Saul, and Jo Anne Miller. *Food For Thought*. New York: Prentice-Hall, 1979.

Morton, Julia F. *Herbs and Spices: A Golden Nature Guide*. New York: Golden Press, 1976.

Muramoto, Naboru. *Healing Ourselves*. New York: Swan House/Avon, 1973.

New Western Garden Book. Miami: Lane Publishing, 1979.

Niethammer, Carolyn. *American Indian Food and Lore; 150 Authentic Recipes*. New York: Macmillan, 1974.

Page, Nancy M. A., and Richard E. Weaver, Jr. *Wild Plants in the City*. New York: Quadrangle/New York Times Book Co., 1975.

Peterson, Lee Allen. *A Field Guide to Edible Wild Plants; East/Central North America*. New York: Houghton Mifflin, 1977.

Poisonous Plants of the United States and Canada. Englewood Cliffs, N. J.: Prentice Hall, 1964.

Public Health Service, *Typical Poisonous Plants*. Food and Drug Administration Booklet, #017-12-00177-7.

Scully, Virginia. *Treasury of American Indian Herbs*, Prineville, OR: Bonanza Publishing, 1974.

Simmonite-Culpepper. *Herbal Remedies*. Universal-Award House, 1970.

Sheffield, Charles. *Earthwatch*. New York: Macmillan, 1981.

Soltanoff, Jack. *Natural Healing*. D. C. Warner Books, 1988.

Sourcebook on Food and Nutrition. Chicago: Marquis Academic Media, 1978.

Stokes, Donald W. *A Guide to Nature in Winter*. Boston: Little, Brown, 1976.

Sweet, Muriel. *Common Edible and Useful Plants of the West*. Happy Camp, Calif.: Naturegraph Publishers, 1962.

Szeley, Edmond Bordeaux. *The Book of Herbs*. International Biogenic Society, 1971.

University of Alaska, *Wild Edibles and Poisonous Plants of Alaska*. Cooperative Extension Service, booklet #F-40, 1953.

U. S. Department of Agriculture and U. S. Department of Health and Human Services. *Dietary Guidelines*. Washington, DC: Government Printing Office, 1980.

Waisel, Yory. *Biology of Halophytes*. New York: Academic Press, 1972.

Western Society of Weed Science. *Weeds of the West*. PO Box 963, Newark, CA 94560, Rev. 1992

Whittle, Tyler. *The Plant Hunters*. Philadelphia: Clinton Book Co., 1970.

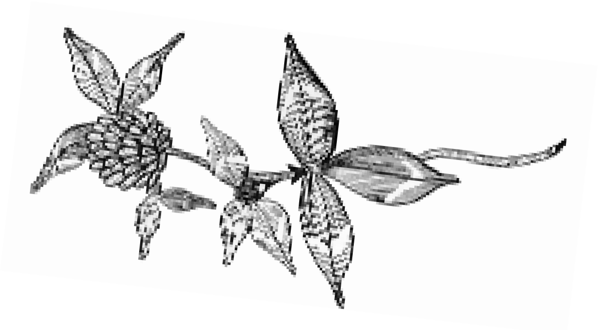

Linda Runyon

Don't Forget-

We have extras for you!

When you go into the kitchen to try the fabulous recipes contained in this book, you very well may not want to drag your computer with you! Printing is disabled in this eBook, but we have the whole printable Recipe Section waiting for you. If you act now, we will include a Free Subscription to Linda Runyon's Newsletter!

Just go here and we'll send them to you. (If your computer can't handle the direct link, please type the web address below into your browser's address bar and fly over to Linda's site.)

http://www.ofthefield.com/eBookSpecialOffers.htm

This is a secret back door for eBook owners like yourself ONLY, and can't be found by coming in the front door of the site.

Your email address is never given to third parties.
It is our strict policy.
You may opt out at any time.

Linda Runyon was raised in Plainfield and Watchung, New Jersey. As a child, she spent 3 months each year at her parents' camp in Indian Lake, New York. This wilderness setting was the background for her own wilderness experiences, homesteading on the outskirts of Indian Lake, NY. For 13 years she, her son, and husband lived without indoor plumbing, electricity or running water. This way of life blessed Linda with knowledge and many skills to live off of the land. She continues to enrich her life with these skills, even during her retirement in her comfortable log home.

Linda has written materials focusing on wild food products:

- A Survival Acre (50 Nationwide Wild Foods, Teas & Medicines)
- Wild Cards A Wild Food Card Deck
- Wild Food & Animals Coloring Book
- A Basic Middle Eastern Desert Guide of Edible Foods (Written for the Kurds)

Linda is a mother of 3, a Grandmother of 4, a registered nurse, an author, a lecturer and an artist.

www.OfTheField.com

Starvation is impossible with this book.

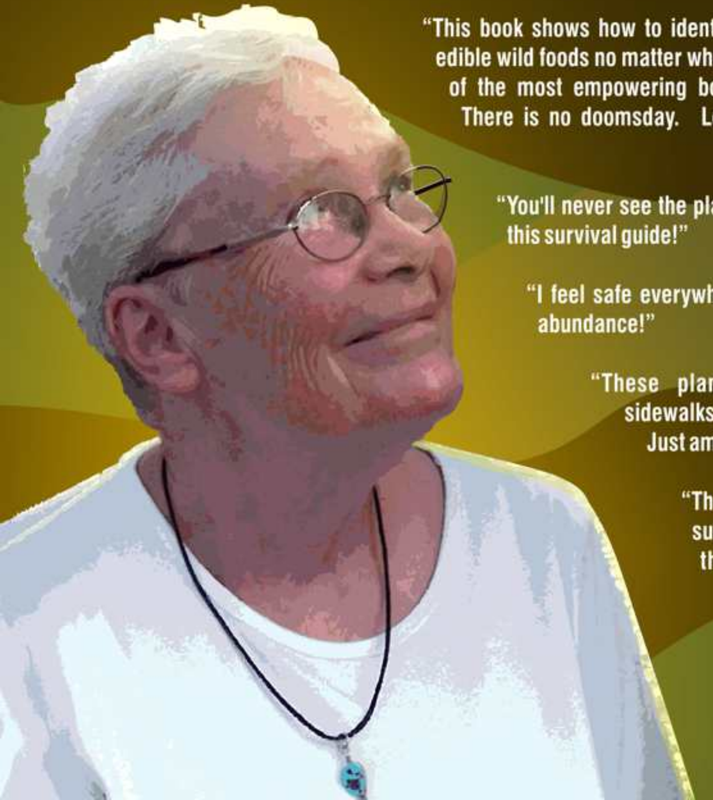

"This book shows how to identify, gather and prepare edible wild foods no matter where you live. This is one of the most empowering books I have ever read. There is no doomsday. Look down and eat up!"
--Reader in USA

"You'll never see the planet the same way after this survival guide!" -Biology student, ASU

"I feel safe everywhere I am with this food abundance!" - Peg, Ohio

"These plants are even in the sidewalks cracks in Manhattan. Just amazing!" -Evelyn, NYC

"Throw away all those other survival books on food - this one has it all!"
-Kevin, NY State

ISBN 093669910-8
9780936699103

✻12

Techniques of Grafting

learning objectives

- Explain the requirements for successful graftage.
- Describe the techniques of detached scion graftage, approach graftage, and repair graftage.
- Discuss the preparation for grafting—tools, accessories, machines, automation, and processing scionwood.
- Explain the craftsmanship of grafting—manual techniques, record keeping, and mechanization.
- Describe the aftercare of grafted plants—in bench grafting systems, and field and nursery grafting systems.
- Identify field, bench, and miscellaneous grafting systems.

INTRODUCTION

Since people first learned to graft plants, a myriad of grafting techniques have been developed. In *The Grafter's Handbook,* Garner (19) enumerates and describes some forty different grafts.

Here we describe the most important grafting methods. Among them, a person who can use a sharp knife can find one that meets any specific grafting need. However, success in grafting depends not only on a technically correct graft but in preparation of the scion and rootstock for graftage. Equally critical are the optimum time for grafting, and proper aftercare.

With high labor costs, only a few of the more efficient grafts are utilized in United States woody ornamental nurseries, including the side veneer, splice (whip graft), and whip-and-tongue graft; use of approach and repair graftage is limited. With fruit crops, depending on the species, a number of different apical, side, and root grafts are utilized around the world. Chip budding and T-budding, which are described in detail in Chapter 13, are two of the most common budding methods for woody ornamentals and fruit crops. Vegetable grafting has increased dramatically worldwide—and is commonly done in Asia and Europe where land is intensively used and crops are not rotated. Grafting onto rootstock resistant to soil pathogens and environmental stress helps increase yield and reduce chemical usage (12, 21, 34). For example, some of the most important grafts with cucurbit vegetables (melon, squash) include hole insertion grafting, tongued approach, and one cotyledon graft (also known as the splice, slant, or the Japanese tube graft), which are described in the chapter. Some robotic vegetable grafting machines can produce 800 grafts per hour.

This chapter is divided into three sections: (a) the **types of grafts,** (b) **production processes of graftage**—including the preparation, craftsmanship, and aftercare of grafted plants, and (c) **grafting systems,** including field grafting, bench grafting, and miscellaneous grafting systems—such as herbaceous graftage, cutting grafts, and micrografting.

REQUIREMENTS FOR SUCCESSFUL GRAFTING

For any successful grafting operation, producing a plant, as shown in Figure 12–1, requires five important elements:

1. *The rootstock and scion must be compatible.* They must be capable of uniting. Usually, but not always, closely related plants, such as two apple cultivars, can be grafted together. Distantly related plants, such as oak and apple, cannot make a successful graft combination (see Chapter 11 for a discussion of these factors).

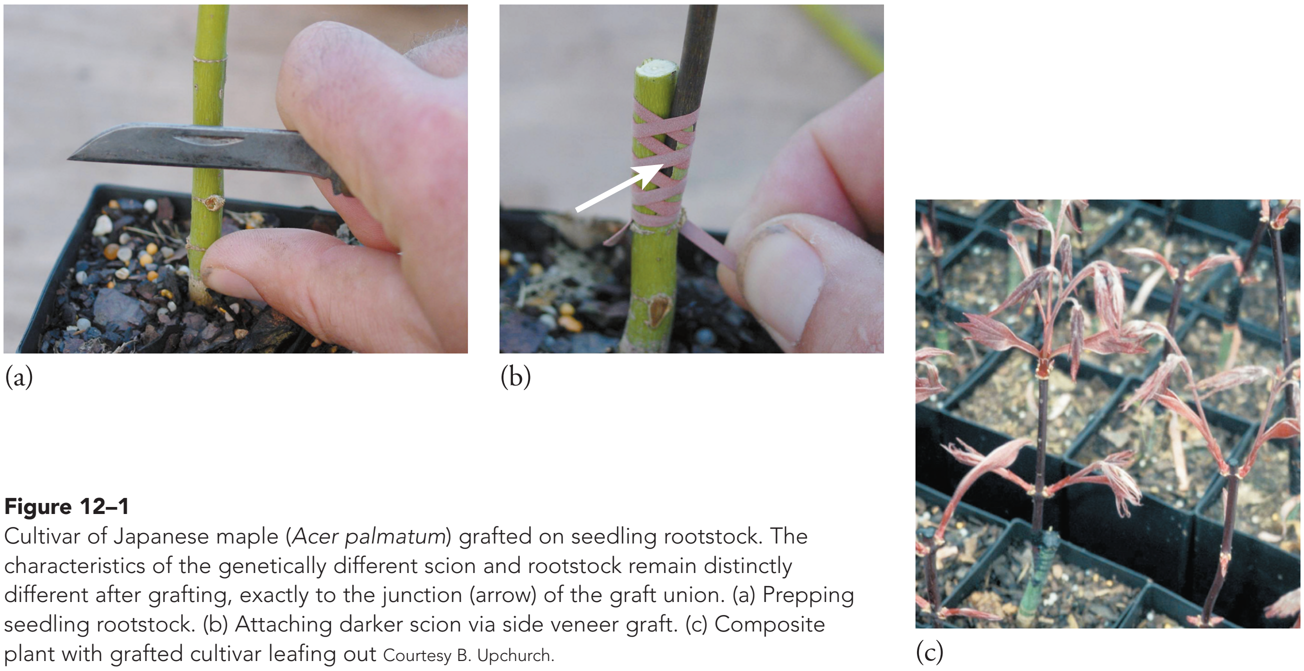

(a) (b) (c)

Figure 12–1
Cultivar of Japanese maple (*Acer palmatum*) grafted on seedling rootstock. The characteristics of the genetically different scion and rootstock remain distinctly different after grafting, exactly to the junction (arrow) of the graft union. (a) Prepping seedling rootstock. (b) Attaching darker scion via side veneer graft. (c) Composite plant with grafted cultivar leafing out Courtesy B. Upchurch.

2. *The* **vascular cambium** *of the scion must be placed in direct contact with that of the rootstock.* The cut surfaces should be held together tightly by wrapping, nailing, wedging, or some similar method. Rapid development of the graft union is necessary so that the scion may be supplied with water and nutrients from the rootstock by the time the buds start to open.

vascular cambium The tissue responsible for the formation of new xylem and phloem in the development of a successful graft union.

3. *The grafting operation must be done at a time when the rootstock and scion are in the proper physiological stage.* Usually, this means that the scion buds are dormant while at the same time, the cut tissues at the graft union are capable of producing the callus tissue necessary for healing of the graft. For deciduous plants, dormant scionwood is collected during the winter and kept inactive by storing at low temperatures. The rootstock plant may be dormant or in active growth, depending upon the grafting method used.
4. *Immediately after the grafting operation is completed, all cut surfaces must be protected from desiccation.* The graft union is covered with tape, grafting wax, Parafilm tape, Buddy Tape, or the grafts are placed in moist material or a covered grafting frame.
5. *Proper care must be given to the grafts for a period of time after grafting.* Shoots (suckers) coming from the rootstock below the graft will often choke out the desired growth from the scion. In some cases, shoots from the scion will grow so vigorously that they break off unless staked and tied or cut back.

TYPES OF GRAFTS

Grafting may be classified according to the part of the rootstock on which the scion is placed—a root, or various places in the top of the plant. Types of grafts can be categorized as (1) **detached scion graftage,** which includes apical, side, bark, and root graftage; (2) **approach graftage,** where the root system of the scion and the shoot system of the rootstock are not removed until after successful graft union formation occurs; and (3) **repair graftage of established trees.** The grafts that are categorized in Tables 12–1 and 12–2 are described in greater detail later in the chapter.

detached scion graftage A type of graft used when a section of the shoot of the scion is removed and grafted to the apex or side of the rootstock. It is also used in grafting roots (root graftage).

approach graftage The root system of the scion and shoot system of the rootstock are not removed until after successful graft union formation occurs.

repair graftage Graft used in repairing or reinforcing injured or weak trees.

Table 12–1
TYPES OF GRAFTS

- I. **Detached Scion Graftage**
  - A. **Apical Graftage**
    - Whip-and-tongue graft
    - Splice graft (whip graft; with vegetables—One cotyledon graft [OCG] or Japanese tube graft)
    - Cleft graft (split graft)
    - Wedge graft (saw-kerf graft)
    - Saddle graft
    - Four-flap graft (banana graft)
    - Hole Insertion Graft (HIG) or Terminal/Tip Insertion graft with vegetables
  - B. **Side Graftage**
    - Side-stub graft
    - Side-tongue graft
    - Side-veneer graft
    - Side insertion graft (SIG) with vegetables
  - C. **Bark Graftage**
    - Bark graft (rind graft)
    - Inlay bark graft
  - D. **Root Graftage**
    - Whole-root and piece-root grafting
    - Nurse-root grafting
- II. **Approach Graftage**
    - Spliced approach graft
    - Tongued approach graft (TAG)
    - Inlay approach graft
- III. **Repair Graft**
    - Inarching
    - Bridge graft
    - Bracing

Table 12–2
UTILIZATION AND ROOTSTOCK CRITERIA OF SELECTED GRAFTS

Graft type	Diameter of rootstock	Rootstock condition	Uses
Whip-and-tongue graft	Small: 6 to 13 mm (1/4 to 1/2 in); same diameter as scions	Dormant; however, active with bench grafting of container rootstock	Bench grafting; container grafting; some topworking in field; root grafting; a popular graft
Whip graft (splice graft) —also called **One cotyledon graft (OCG) or Japanese tube graft** with vegetables.	Small: 6 to 13 mm (1/4 to 1/2 in); same diameter as scions; See Figure 12–46 for schedule.	Dormant; however, active with bench grafting of container rootstock, greenwood grafting, and vegetable crops	Bench grafting; container grafting; some topworking in field; grafting of vegetable liner plants; root grafting; a popular graft
Cleft graft (split graft)	Moderate: 2.5 to 10 cm (1 to 4 in)	Dormant—before active growth starts in spring	Topworking in field
Wedge graft (saw-kerf graft)	Moderate: 2.5 to 10 cm (1 to 4 in)	Dormant—before active growth starts in spring	Topworking in field
Saddle graft	Small: 6 to 19 mm (1/4 to 3/4 in); same diameter as scion	Dormant	Bench grafting via hand or machine; container grafting; root grafting
Four-flap graft (banana graft)	Small: up to 2.5 cm (1 in); same diameter as scions	Active; bark must be slipping	Topworking small caliper trees
Hole insertion graft (HIG) or **Terminal/Top insertion graft**			Bench grafting; container grafting of liner vegetable plants

Graft type	Diameter of rootstock	Rootstock condition	Uses
Side-stub graft	Small to moderate rootstock larger than scion: grafted on rootstock branches up to 2.5 cm (1 in) in diameter	Dormant	Topworking in field
Side-tongue graft	Small: 6 to 19 mm (1/4 to 3/4 in); diameter of scion slightly smaller than rootstock	Dormant	Bench grafting; container grafting of broad- and narrow-leaved evergreen species
Side-veneer graft	Small: 6 to 19 mm (1/4 to 3/4 in); same diameter as scion	Dormant	Bench grafting; container grafting of smaller liner potted plants; a popular graft for conifers, deciduous trees and shrubs, and fruit crops
Side insertion graft (SIG)	Works well using vegetable rootstock with wide hypocotyls.	Active	Bench grafting; container grafting of liner vegetable plants
Bark graft (rind graft)	Large: 2.5 to 30 cm (1 to 12 in).	Active; bark must be slipping	Topworking in field
Inlay bark graft	Large: 2.5 to 30 cm (1 to 12 in).	Active; bark must be slipping	Topworking in field (e.g., pecans)
Spliced approach graft	Small: 6 to 19 mm (1/4 to 3/4 in); same size as scion; exception is mango grafting in India on larger, established trees	Active	Container grafting with difficult-to-graft species; scion and stock grafted as two independent, self-sustaining plants; only limited topworking in field
Tongued approach graft (TAG)	Small: 6 to 19 mm (1/4 to 3/4 in); same size as scion	Active	Container grafting with difficult-to-graft species; scion and stock grafted as two independent, self-sustaining plants; also used with vegetables
Inlay approach graft	Small: 6 to 19 mm (1/4 to 3/4 in); bark of rootstock is thicker than scion	Active	Container grafting with difficult-to-graft species; scion and stock grafted as two independent, self-sustaining plants
Inarching	Large: 15 cm (6 in) and larger	Dormant	Used to replace a weak or damaged root system of an established tree
Bridge graft	Large: 15 cm (6 in) and larger	Active; bark must be slipping	Repair injury to trunk of tree
Bracing	Limbs of tree bound by pulling together two strong young lateral shoots from limbs to be braced	Active or dormant	Natural grafting used to strengthen scaffolding limbs of a tree

Detached Scion Graftage—Apical Graftage

There are many variations of apical graftage. As the name suggests, the scion is inserted into the top of the severed rootstock shoot.

Whip-and-Tongue Graft The whip-and-tongue graft, shown in Figures 12–2 and 12–3, is particularly useful for grafting relatively small material about 6 to 13 mm (1/4 to 1/2 in) in diameter. It is highly successful if done properly because there is considerable vascular cambium contact,

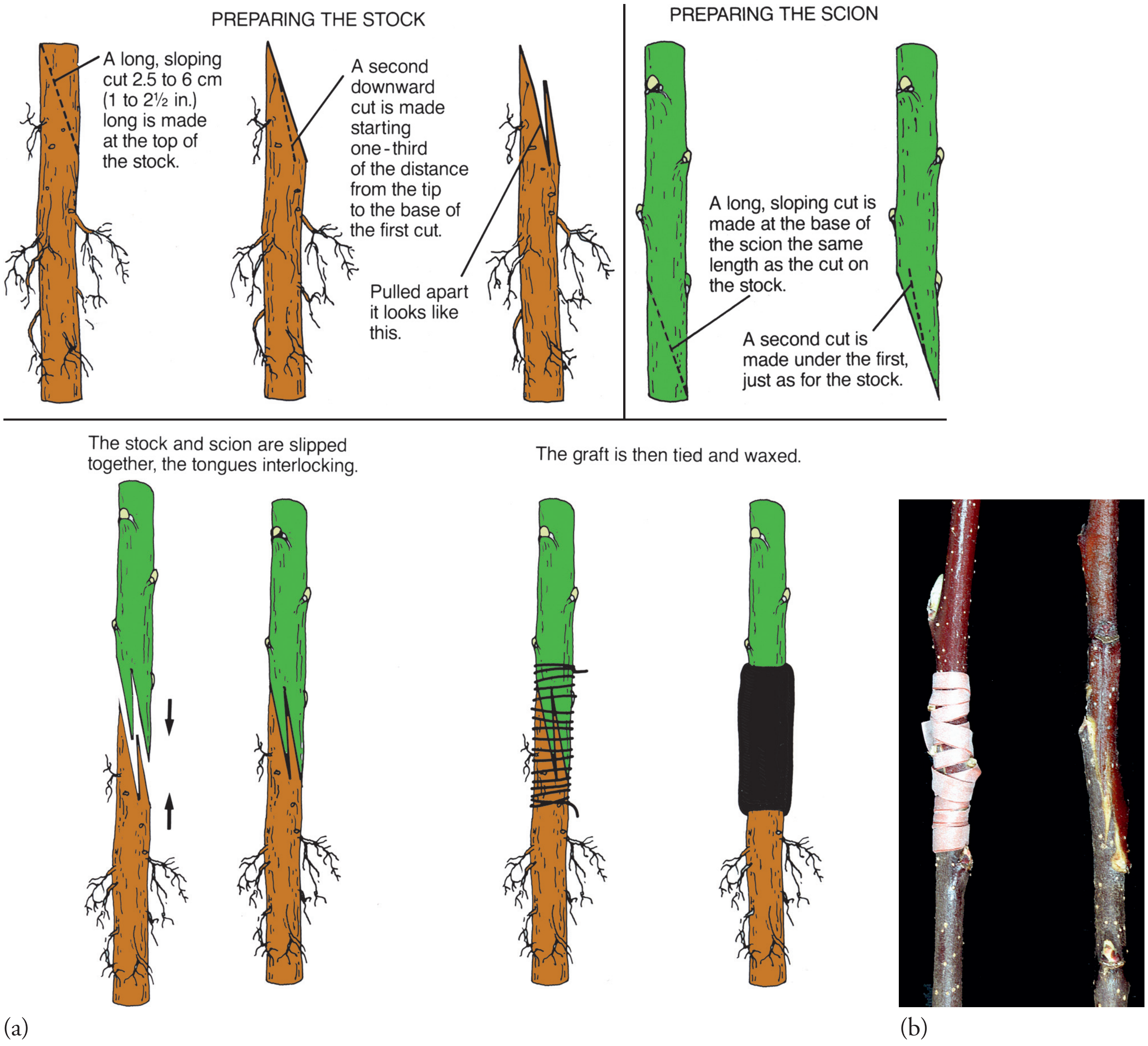
PREPARING THE STOCK
A long, sloping cut 2.5 to 6 cm (1 to 2½ in.) long is made at the top of the stock.
A second downward cut is made starting one-third of the distance from the tip to the base of the first cut.
Pulled apart it looks like this.
PREPARING THE SCION
A long, sloping cut is made at the base of the scion the same length as the cut on the stock.
A second cut is made under the first, just as for the stock.
The stock and scion are slipped together, the tongues interlocking.
The graft is then tied and waxed.
(a)
(b)

Figure 12–2

Whip-and-tongue graft. (a) This method is widely used in grafting small plant material and is especially valuable in making root grafts as illustrated here. (b) Whip-and-tongue of bench grafted pear.

plus it heals quickly and makes a strong union. Preferably, the scion and rootstock should be of equal diameter. The scion should contain two or three buds, and the graft made in the smooth internodes area below the lower bud.

The cuts made at the top of the rootstock should be the same as those made at the bottom of the scion. First, a smooth, sloping cut is made, 2.5 to 6 cm (1 to 2 1/2 in) long; longer cuts are made when working with large material. This first cut should preferably be made with one single stroke of the knife, in order to leave a smooth, flat surface. To do this, the knife must be razor sharp. Wavy, uneven cuts made with a dull knife will not result in a satisfactory union.

On each of these cut surfaces, a reverse cut is made. It is started downward at a point about one-third of the distance from the tip and should be about one-half the length of the first cut. To obtain a smooth-fitting graft, this second cut should not just split the grain of the wood but should follow along under the first cut, tending to parallel it.

The rootstock and scion are then inserted into each other, with the tongues interlocking. It is extremely important that the vascular cambium layers match along at least one side, preferably along both sides. The lower tip of the scion should not overhang the stock, because it increases the likelihood of the formation of

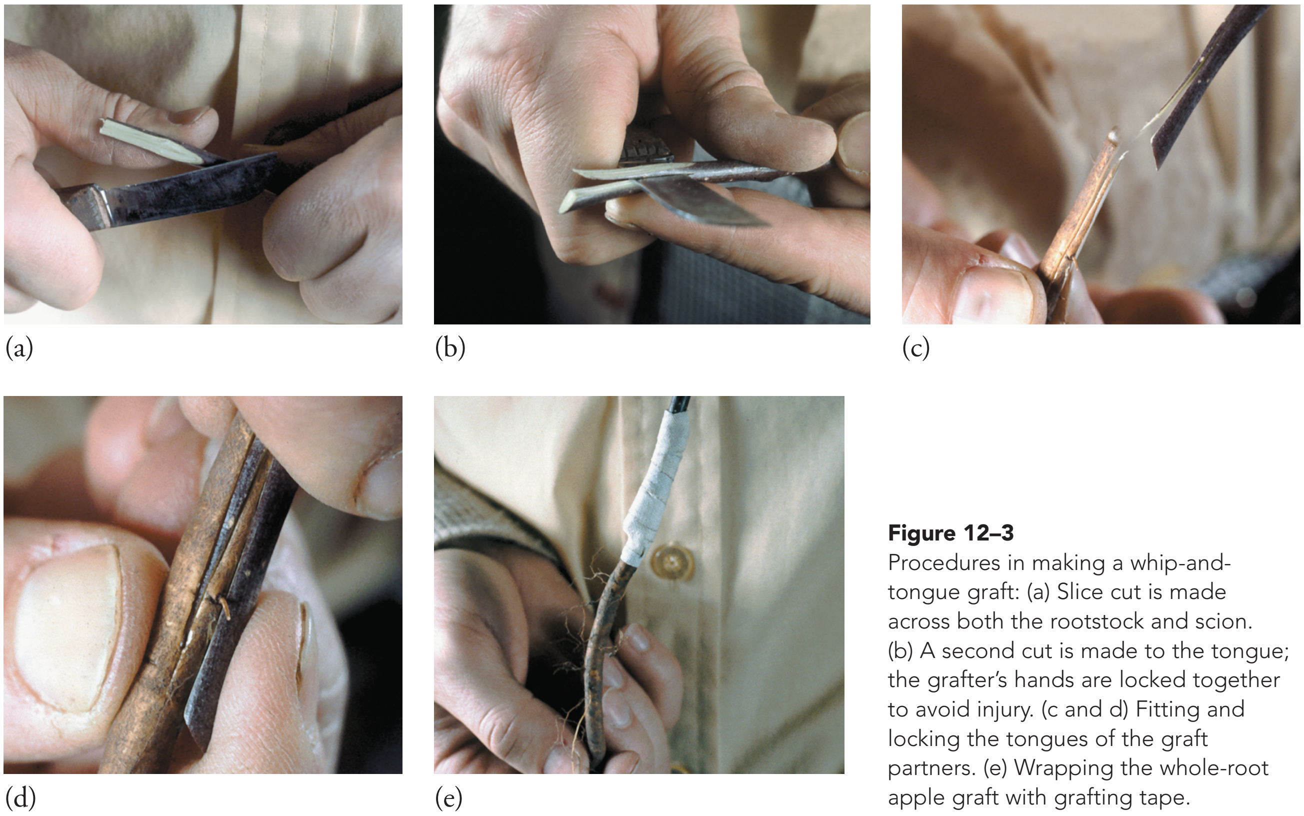
(a) (b) (c) (d) (e)

Figure 12–3
Procedures in making a whip-and-tongue graft: (a) Slice cut is made across both the rootstock and scion. (b) A second cut is made to the tongue; the grafter's hands are locked together to avoid injury. (c and d) Fitting and locking the tongues of the graft partners. (e) Wrapping the whole-root apple graft with grafting tape.

large callus knots. The use of scions larger than the rootstock should be avoided for the same reason.

After the scion and rootstock are fitted together, they are securely tied with budding rubber strips, plastic (poly) budding/grafting tape, or raffia. It is important that the tissues in the graft union area not dry out, so either sealing the graft union with grafter's wax, Parafilm, or Buddy Tape, or placing the plants under high relative humidity, is essential until the graft union has formed.

In **bench graftage** (page 502) the bare-root grafted plants can be stored in a grafting box (without sealing the graft union with grafter's wax) and packed with slightly moist peat or bark. Grafted plants in liner pots can be placed in a polytent in a temperature-controlled greenhouse (Fig. 12–4). If bare-root, bench-grafted plants are to be directly planted in a field nursery, the graft union is temporarily placed below the soil level. Any poly budding tape will need to be removed after graft union formation to prevent girdling the stem. Grafts wrapped with budding rubbers and temporarily covered with soil or media should be inspected later; the rubber decomposes very slowly below ground and may cause a constriction at the graft union.

If the whip-and-tongue graft is used in **field grafting,** the graft union of the **topworked** (page 422) plant must be tied and sealed with grafter's wax, Parafilm, or Buddy Tape. Aftercare of grafted plants is further described in the section "Production Processes of Graftage" (page 491).

Splice Graft (Whip Graft) The splice graft is simple and easy to make (Fig. 12–5). It is the same as the whip-and-tongue graft except that the second, or "tongue," cut is not made in either the rootstock or scion. A simple slanting cut of the same length and angle is made in both the rootstock and the scion. These are placed together and wrapped or tied as described for the whip graft. If the scion is smaller than the rootstock it should be set at one side of the rootstock so that the vascular cambium layers will match along that side (Fig. 12–5).

The splice graft is particularly useful in grafting plants that have a very pithy stem or that have wood that is not flexible enough to permit a tight fit when a tongue is made as in the whip-and-tongue graft. The splice graft is used in greenhouse production of vegetable crops for grafting disease-resistant rootstocks. For vegetable crops such as cucurbits or *Solanaceae,* this graft is sometimes referred to as One Cotyledon Grafting (OCG), the slant graft, or Japanese tube graft

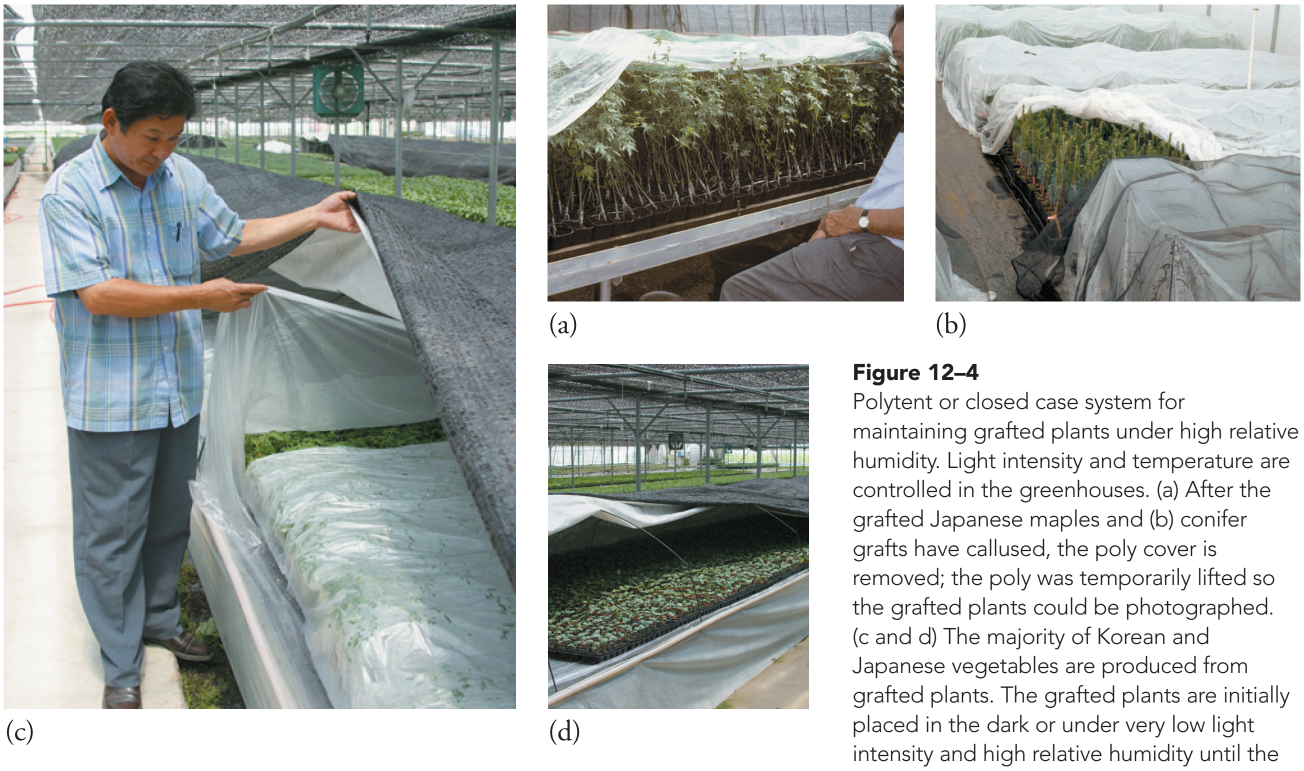
(a) (b) (c) (d)

Figure 12–4
Polytent or closed case system for maintaining grafted plants under high relative humidity. Light intensity and temperature are controlled in the greenhouses. (a) After the grafted Japanese maples and (b) conifer grafts have callused, the poly cover is removed; the poly was temporarily lifted so the grafted plants could be photographed. (c and d) The majority of Korean and Japanese vegetables are produced from grafted plants. The grafted plants are initially placed in the dark or under very low light intensity and high relative humidity until the graft has healed. Photos c and d courtesy M. Peet.

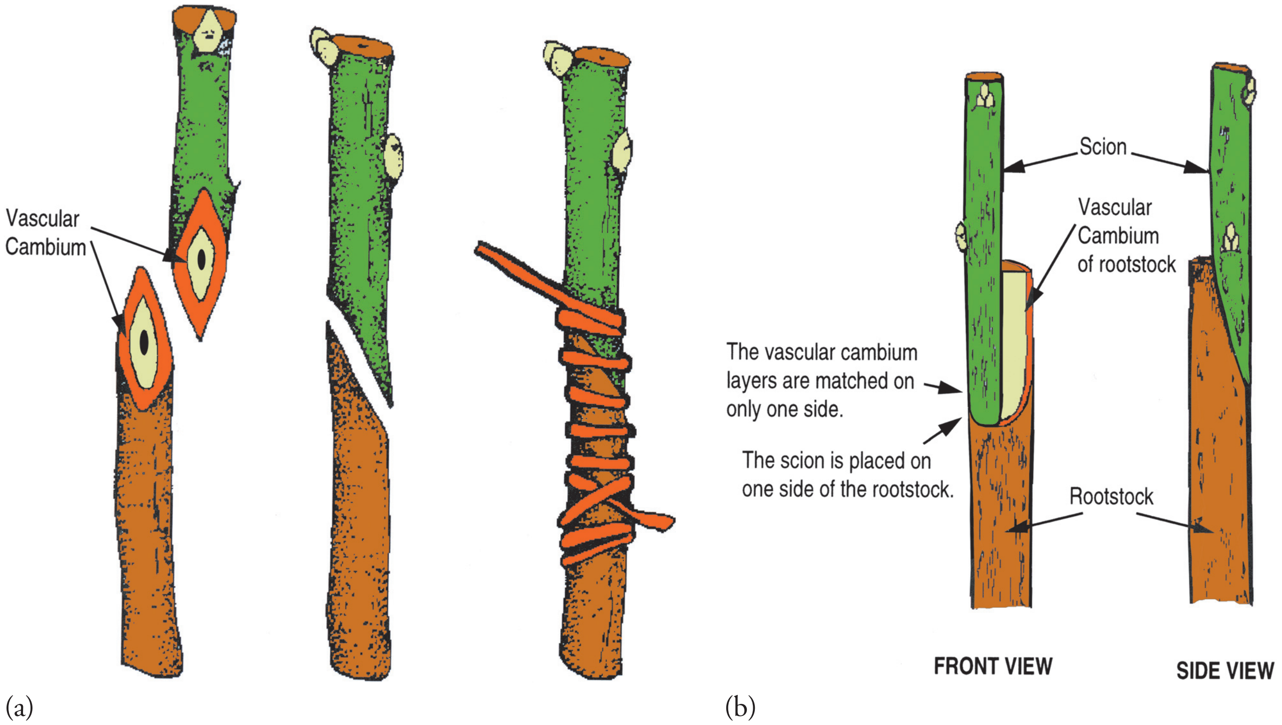
Vascular Cambium
(a)
Scion
Vascular Cambium of rootstock
The vascular cambium layers are matched on only one side.
The scion is placed on one side of the rootstock.
Rootstock
FRONT VIEW
SIDE VIEW
(b)

Figure 12–5
Splice graft (whip graft). (a) Procedures in making the splice graft with a slice cut that slants across the grafting partners. Ideally, the rootstock and scion are of the same caliber. (b) Method of making a splice graft when the scion is considerably smaller than the rootstock. It is important that the cambium layers be matched on one side.

BOX 12.1 **GETTING MORE IN DEPTH ON THE SUBJECT**

CORRECT INSERTION OF THE SCION

In all types of grafting, the scion must be inserted right side up. That is, the apical tip of the buds on the scion should be pointing upward and away from the rootstock. The graft will not be successful if this rule is not observed.

(Figs. 12–6 and 12–7) (11, 12, 21). The graft can be performed manually or with sophisticated, robotic grafting machines; see Figures 12–43, 12–44, and 12–45, pages 498–99. The rootstock and scion must be held together while tying the splice graft. In field grafting, it is not a convenient method to use at ground level, and must be performed higher up on the rootstock, where the grafter must do both the cutting and tying. The whip-and-tongue does not have this limitation, since the tongue holds the graft together, so that

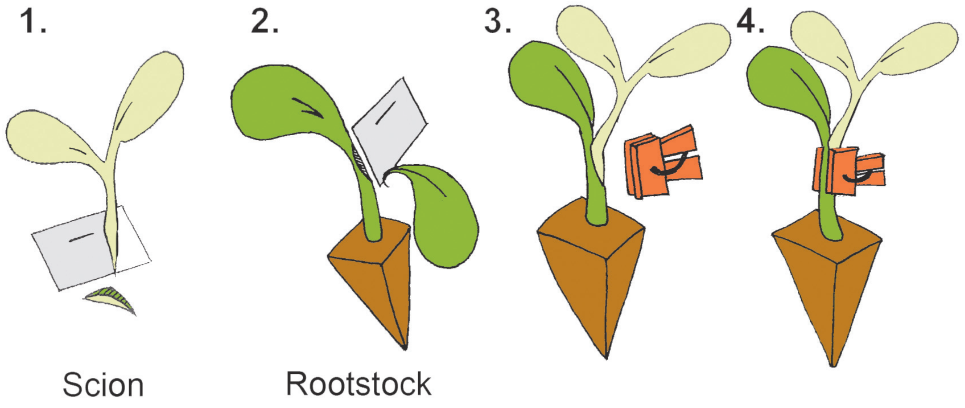

Figure 12–6
One cotyledon grafting (OCG), which is a form of the splice graft used with cucurbit vegetable crops; also called the slant graft and Japanese tube graft (12, 21). Illustrations courtesy R. L. Hassell.

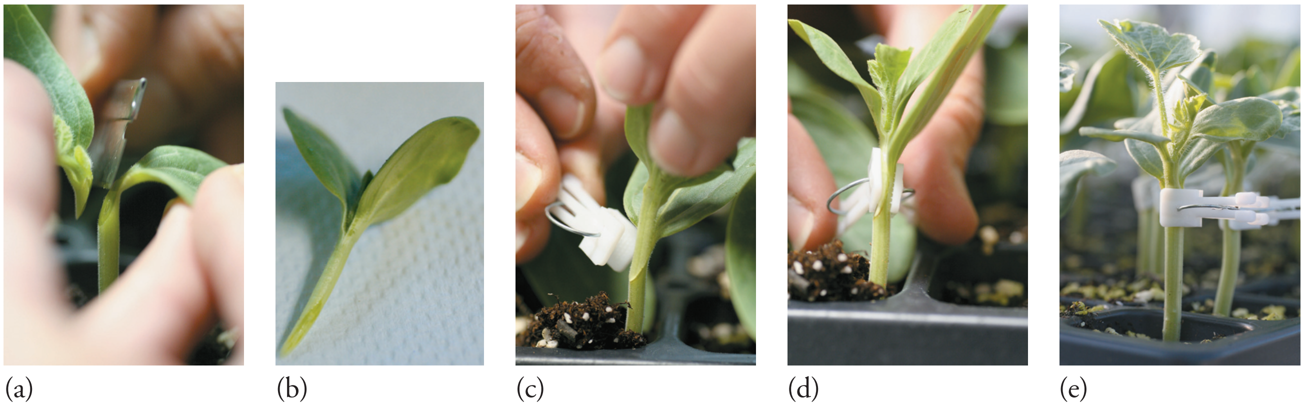
(a) (b) (c) (d) (e)

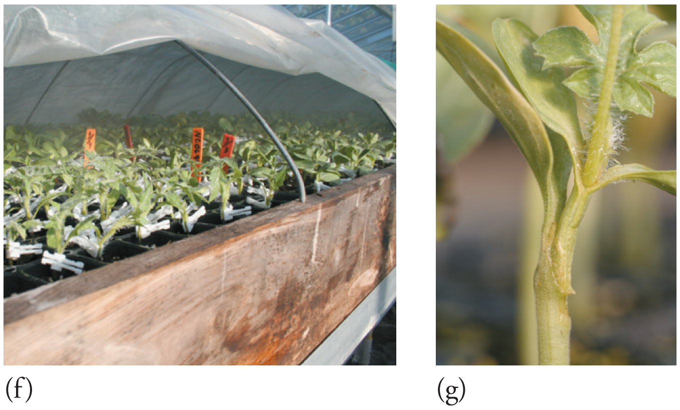
(f) (g)

Figure 12–7
One cotyledon grafting (OCG): A form of splice graft used with cucurbits. (a) Preparation of squash rootstock leaving a single cotyledon leaf. (b) Watermelon scion with slant cut. (c) Plastic clip used to hold scion and rootstock. (d and e) Plastic clips used to hold watermelon scion and squash rootstock. (f) Grafts are allowed to heal under very high humidity and dark to very low light conditions until graft union formation has occurred. (g) Successfully healed OCG. Photos courtesy of R. L. Hassell.

grafter The person cutting the stock and scion and inserting the scion piece.

Tier The person who completes the grafting process by tying, and sometimes waxing, the graft area.

the **grafter** has both hands free for tying, or the fitted graft can be left for a helper or **"tier"** to tie and seal. A splice graft is used in bench grafting and grafting of container plants.

Cleft Graft (Split Graft) The cleft or split graft is one of the oldest methods of field grafting. It is used to topwork trees, either in the trunk of a small tree or in the scaffold branches of a larger tree (Figs. 12–8 and 12–9). Cleft grafting is used for *crown grafting* (see the "Grafting Systems" section, page 504) or grafting smaller plants such as established grapevines or camellias. In topworking trees, this method should be limited to rootstock branches about 2.5 to 10 cm (1 to 4 in) in diameter, and to species with fairly straight-grained wood that will split evenly.

Although cleft grafting can be done any time during the dormant season, the chances for successful healing of the graft union are best if the work is done in early spring just when the buds of the rootstock are beginning to swell, but before active growth has started. If cleft grafting is done after the tree is in active growth, the bark of the rootstock may separate from the wood, making it difficult to obtain a good union. When this separation occurs, the loosened bark must be firmly nailed back in place. The scions should be made from dormant, 1-year-old wood. Unless the grafting is done early in the season (when the dormant scions can be collected and used immediately), the

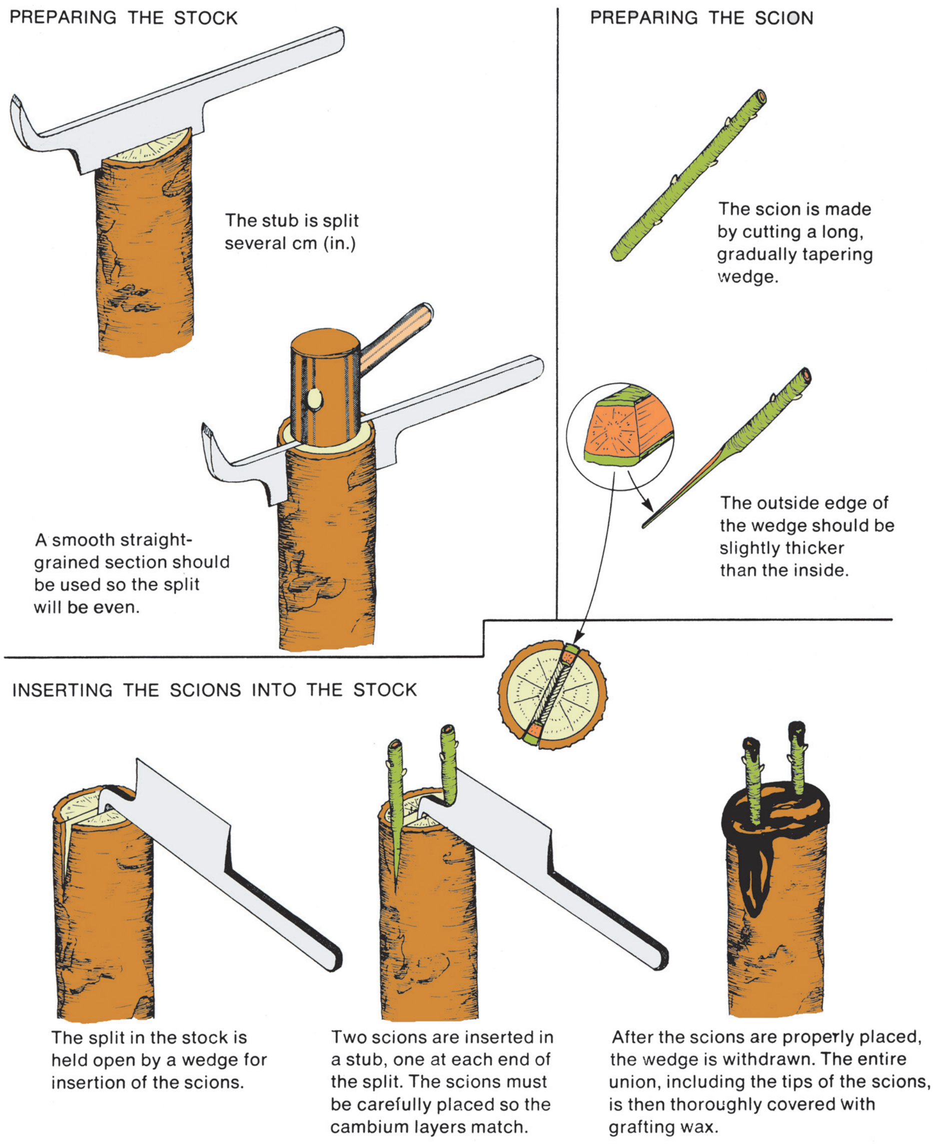
PREPARING THE STOCK
The stub is split several cm (in.)
A smooth straight-grained section should be used so the split will be even.
PREPARING THE SCION
The scion is made by cutting a long, gradually tapering wedge.
The outside edge of the wedge should be slightly thicker than the inside.
INSERTING THE SCIONS INTO THE STOCK
The split in the stock is held open by a wedge for insertion of the scions.
Two scions are inserted in a stub, one at each end of the split. The scions must be carefully placed so the cambium layers match.
After the scions are properly placed, the wedge is withdrawn. The entire union, including the tips of the scions, is then thoroughly covered with grafting wax.

Figure 12–8 Steps in making the cleft graft (split graft). This method is very widely used and is quite successful if the scions are inserted so that the cambium layers of stock and scion match properly.

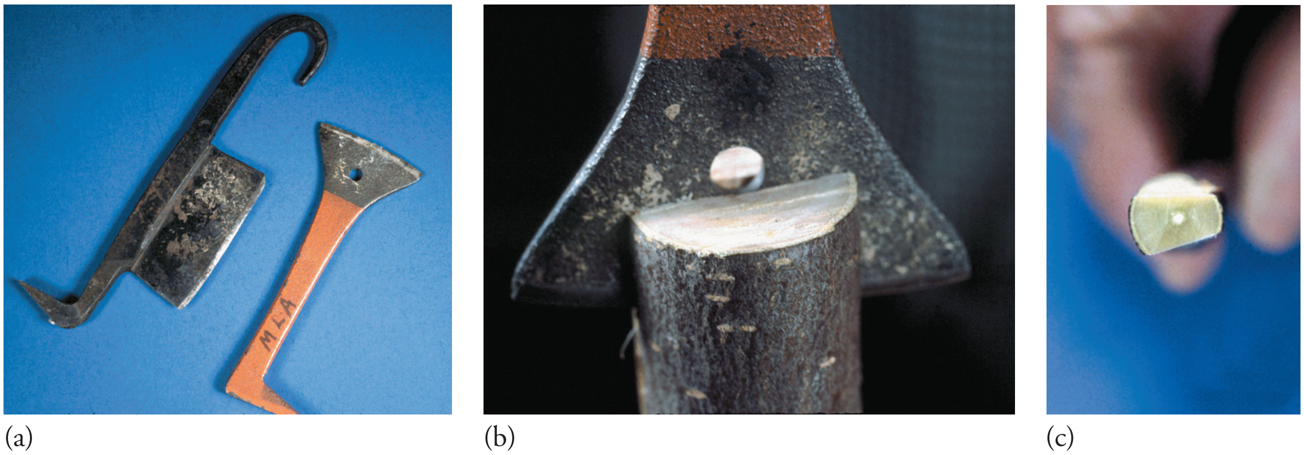

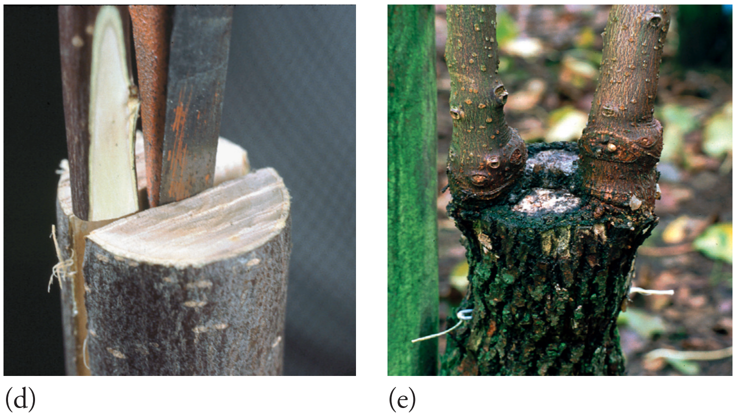

Figure 12–9
(a) Tools used in making a cleft (split) graft. (b) Making the cleft by splitting the rootstock top. (c) Scionwood with the outside wedge slightly thicker than the inside. (d) Inserting the first of two scions. The split rootstock is temporarily separated by the tool. (e) Cleft grafted yellow kiwi (scion) to replace the standard kiwi.

scionwood should be collected in advance and refrigerated. In sawing off the branch for this and other topworking methods, the cut should be made at right angles to the main axis of the branch.

In making the cleft graft, a heavy knife, such as a butcher knife, or one of several special cleft grafting tools, is used to make a vertical split for a distance of 5 to 8 cm (2 to 3 in) down the center of the stub to be grafted (Figs. 12–8 and 12–9). This split is made by pounding the knife in with a hammer or mallet. The branch is sawed off in such a position that the end of the stub that is left is smooth, straight-grained, and free of knots for at least 15 cm (6 in). Otherwise, the split may not be straight, or the wood may split one way and the bark another. The split should be in a tangential rather than radial direction in relation to the center of the tree to permit better placement of the scions for their subsequent growth. Sometimes the cleft is made by a longitudinal saw cut rather than by splitting. After a good, straight split is made, a screwdriver, chisel, or the wedge part of the cleft-grafting tool is driven into the top of the split to hold it open.

Two scions are inserted, one at each side of the stock where the vascular cambium layer is located. The scions should be 8 to 10 cm (3 to 4 in) long, about 10 to 13 mm (3/8 to 1/2 in) thick, and should have two or three buds. The basal end of each scion should be cut into a gently sloping wedge about 5 cm (2 in) long. It is not necessary that the end of the wedge come to a point. The side of the wedge which is to go to the outer side of the rootstock should be slightly wider than the inside edge. Thus, when the scion is inserted and the tool is removed, the full pressure of the split rootstock will come to bear on the scions at the position where the vascular cambium of the rootstock touches the vascular cambium layer on the outer edge of the scion. Since the bark of the rootstock is almost always thicker than the bark of the scion, it is usually necessary for the outer surface of the scion to set slightly in from the outer surface of the rootstock in order to match the vascular cambium layers.

The long, sloping wedge cuts at the base of the scion should be smooth, a single cut on each side made with a sharp knife. Both sides of the scion wedge should press firmly against the rootstock for their entire length. A common mistake in cutting scions for this type of graft is to make the cut on the scion too short and the slope too abrupt, so that the point of contact is only at the top. Slightly shaving the sides of the split in the stock will often permit a smoother contact.

After the scions are properly made and inserted, the tool is withdrawn, without disturbing the scions, which should be held tightly by the pressure of the rootstock so that they cannot be pulled loose by hand. No further tying or nailing is needed unless very small rootstock branches have been used. In this case, the top of the rootstock can be wrapped tightly with poly grafting tape or adhesive tape to hold the scions in place more securely.

Thorough waxing of the completed graft is essential. The top surface of the stub should be entirely covered, permitting the wax to work into the split in the stock. The sides of the grafted stub should be well covered with wax as far down the stub as the length of the split. The tops of the scions should be waxed but not necessarily the bark or buds of the scion. Two or three days later, all the grafts should be inspected and rewaxed where openings appear. Lack of thorough and complete waxing in this type of graft is a common cause of failure.

Wedge Graft (Saw-Kerf Graft) Wedge grafting is illustrated in Figure 12–10. Like the cleft graft, it can be made in late winter (in mild climates) or early spring before the bark begins to slip (separates easily from the wood).

The diameter of the stock to be grafted is the same as for the cleft graft—5 to 10 cm (2 to 4 in), and the scions are also the same size—10 to 13 cm (4 to 5 in) long and 10 to 13 mm (3/8 to 1/2 in) in thickness.

A sharp, heavy, short-bladed knife is used for making a V-wedge in the side of the stub, about 5 cm (2 in) long. Two cuts are made, coming together at the bottom and as far apart at the top as the width of the scion. These cuts extend about 2 cm (3/4 in) deep into the side of the stub. After these cuts are made, a screwdriver is pounded downward behind the wedge chip from the top of the stub to knock out the chip, leaving a V-shaped opening for insertion of the scion. The base of the scion is trimmed to a wedge shape exactly the same size and shape as the opening. With the two vascular cambium layers matching, the scion is tapped downward, firmly into place, and slanting outward slightly at the top so that the vascular cambium layers cross. If the cut is long enough and gently tapering, the scion should be so tightly held in place that it would be difficult to dislodge.

In a stub that is 5 cm (2 in) wide, 2 scions should be inserted 180 degrees apart; in a 10-cm (4 in) stub, 3 scions should be used, 120 degrees apart. After all scions are firmly tapped into place, all cut surfaces, including the tips of the scion, should be waxed thoroughly.

Saddle Graft The saddle graft can be bench grafted by hand or machine (see Fig. 12–41, page 497). The rootstock and scion should be the same size. The scion is prepared by cutting upward through the bark and into the wood on opposite sides of the scion (Fig. 12–11, page 476). The knife should penetrate more deeply into the wood as the cuts are lengthened. Before the knife is withdrawn, it is turned towards the middle of the scion piece, and the saddle shape is gradually formed by removing pieces of the wood. The rootstock is cut transversely and receives two upward cuts on either side to expose the vascular cambium of the rootstock, in order to match vascular cambium in the saddle of the scion. The apex of the rootstock is carved to fit the saddle. The graft needs to be tied, and all exposed cut surfaces sealed or stored in a grafting case until the graft union has formed. The saddle graft is used for bench grafting grape and *Rhododendron* cultivars (19).

Four-Flap Graft (Banana Graft) The four-flap graft is used in topworking small-caliper trees or tree limbs up to 2.5 cm (1 in) in diameter. This field graft is normally done manually [Figs. 12–12 (page 476) and 12–13 (page 477)], but there is a tool that aids in stripping the rootstock bark flaps from the wood (Fig. 12–13). Both the scion and rootstock should be of equal diameter, and the best fit is obtained when the scion is slightly larger than the rootstock. The four-flap graft is done with pecans in Texas from April to mid-May, when the rootstock bark is actively slipping (39). Scionwood, which is collected while dormant during the winter, is taken from cold storage and used immediately.

The rootstock with a primary stem or lateral limb is severed horizontally with sharp pruning shears. On the rootstock where the horizontal cut was made, 4 vertical, equally spaced cuts 4 cm (1.5 in) long are made with a grafting knife that penetrates from the bark down to the interior wood. A 15 cm (6 in) piece of scionwood with 3 axillary buds is cut on 4 sides with a knife. Cuts are made on the scion through the bark down to the wood—without removing much wood. There should be 4 thin slivers of bark, with the vascular cambium at the corners, which gives the prepared scion a square diameter appearance. The 4 flaps of bark are pulled down 4 cm (1.5 in) on the rootstock, and the inner wood is removed with pruning shears. The scion piece is inserted upright on the rootstock and the 4 flaps of the rootstock are pulled up to cover the 4 cut surfaces of the scion. A rubber band is rolled up onto the flaps to hold them in place. The cut flap areas are then tied with flagging tape, green floral tape, or white budding tape. The tip of the scion is painted with tree paint or sealed with white glue to prevent it from drying out.

Then the taped graft area is covered with aluminum foil to protect it from heat. A hole is made in the corner of a clear poly bag (freezer bag) and the poly

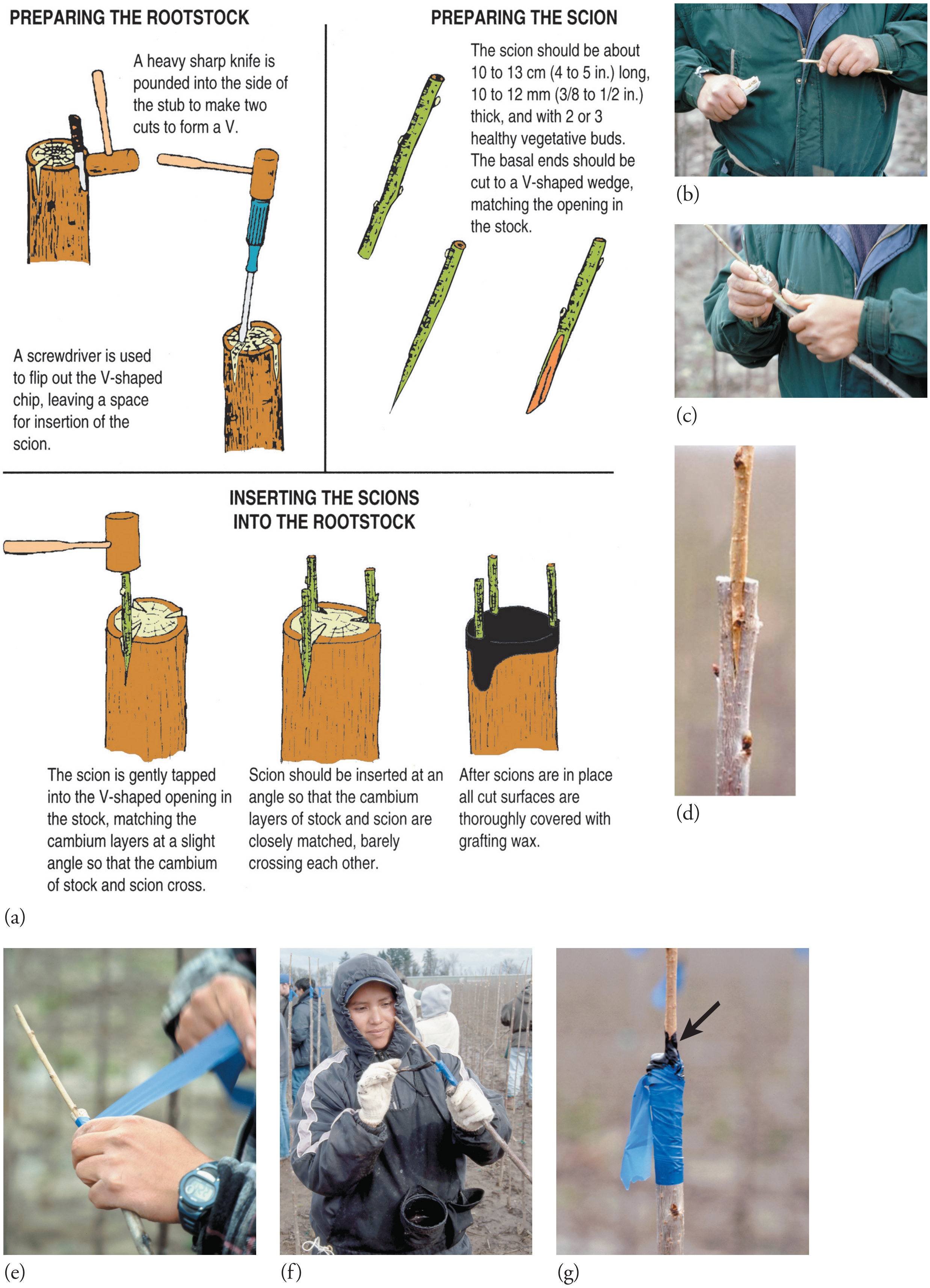

PREPARING THE ROOTSTOCK

A heavy sharp knife is pounded into the side of the stub to make two cuts to form a V.

A screwdriver is used to flip out the V-shaped chip, leaving a space for insertion of the scion.

PREPARING THE SCION

The scion should be about 10 to 13 cm (4 to 5 in.) long, 10 to 12 mm (3/8 to 1/2 in.) thick, and with 2 or 3 healthy vegetative buds. The basal ends should be cut to a V-shaped wedge, matching the opening in the stock.

INSERTING THE SCIONS INTO THE ROOTSTOCK

The scion is gently tapped into the V-shaped opening in the stock, matching the cambium layers at a slight angle so that the cambium of stock and scion cross.

Scion should be inserted at an angle so that the cambium layers of stock and scion are closely matched, barely crossing each other.

After scions are in place all cut surfaces are thoroughly covered with grafting wax.

(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figure 12–10
(a) Wedge graft (saw-kerf graft). Sometimes called the saw-kerf because the cuts in the side of the rootstock can be made with a saw, rather than with the sharp tool depicted. (b, c, d, e, and f) Wedge graft of cherry whips in field using one scion piece. (b) Trimming scionwood with grafting knife, (c and d) inserting scionwood into rootstock, (e) wrapping graft with poly, (f) sealing with grafting wax—notice wax container (arrow) and (g) tied graft with poly and grafting wax covering bottom of scion and graft union area (arrow) used to fill in tissue separation.

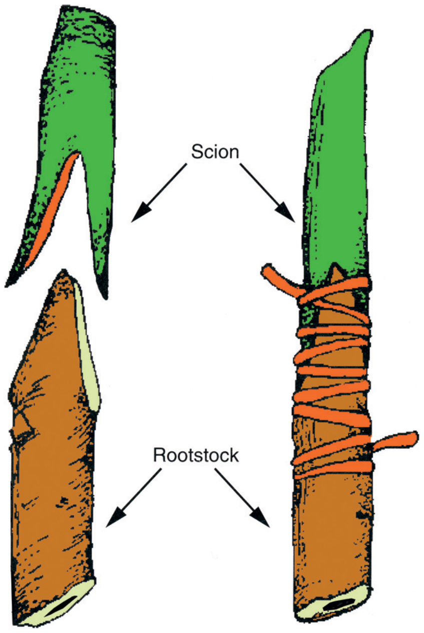

Figure 12–11
Steps in making the saddle graft. The scion is cut to have a saddle appearance and the understock to form a blunt point. The cambium layers are matched up and the graft tied.

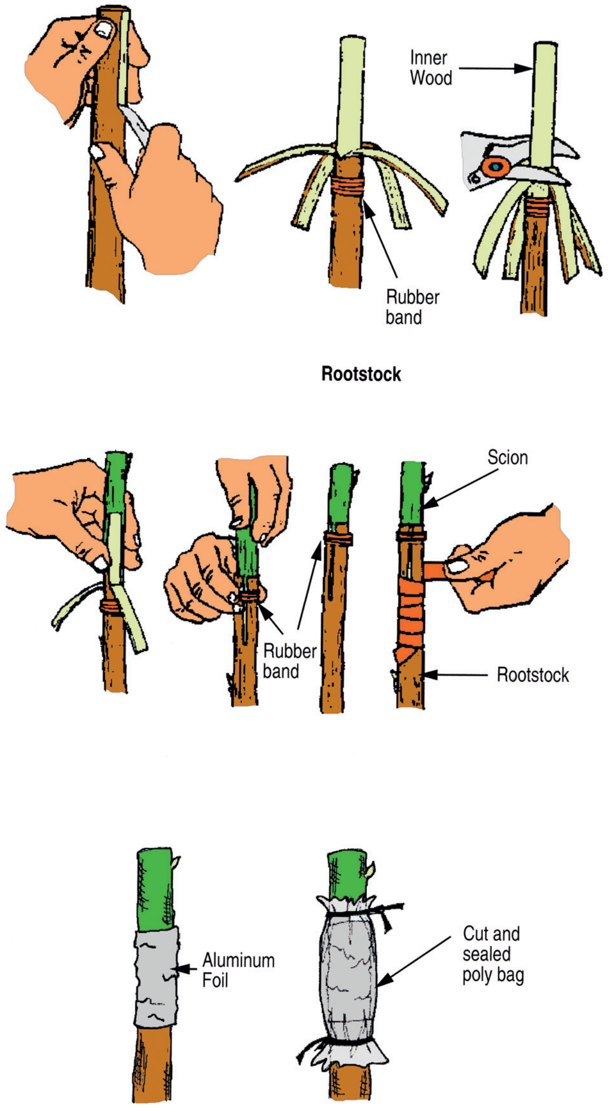

Figure 12–12
The four-flap or banana graft. *Top:* The top of the rootstock is cut horizontally, and the bark is cut vertically into four strips. The four bark flaps are peeled down and the inner wood removed. *Middle:* The scion bark is removed and the wood retained. The flaps of the rootstock cover the cut surfaces of the scionwood and are temporarily held by a rubber band. The graft is then tied with white grafting tape. *Bottom left:* Aluminum foil is wrapped around the graft to exclude heat from the graft. *Bottom right:* The grafted area covered with aluminum foil is wrapped with a cut poly bag, which is sealed to retain high relative humidity until the graft takes (39).

slid down over the graft area so that it is just covering the aluminum foil (no poly should cover the apex of the scion, nor should it touch any exposed wood). Air is expelled so the poly fits snugly over the aluminum foil, and it is tied at both ends with stretchable plastic budding tape or rubber bands. The function of the poly bag is to maintain a high relative humidity in the graft area. In 4 to 6 weeks after the graft has taken, the ties, poly bag, and aluminum foil are removed.

The vegetative growth of the rootstock plant must be kept in check, since many new shoots will appear on the rootstock below the graft. Some of these shoots are needed for maintaining tree vigor, but the rootstock shoots should not become dominant or exceed the height of the scion—the growing tips of the rootstock shoots will have to be removed several times during the growing season. After 2 to 3 years, all rootstock branches are removed below the graft and the scion becomes the dominant shoot system.

Hole Insertion Graft (HIG) or Terminal/Top Insertion Graft This technique is used for grafting watermelon to squash rootstock (12, 21). This is most popular graft used in China because it is suitable for *Lagenaria* (Cucurbita) and interspecific squash as rootstocks, requires few materials, is highly efficient (1,500+ plants/day/worker), and allows simpler management techniques (34). When both cotyledons and first true leaf start to develop, the rootstock plant is ready to graft (7 to 10 days after sowing). Remove the growing point with a sharp probe, and then open a hole on the upper portion of the rootstock hypocotyl with a bamboo needle or 1.4-mm drill bit. The scion is then cut on a 35- to 45-degree angle, on both sides, on the hypocotyls and inserted into the hole made in the rootstock. The cut surfaces are matched together, held with or without a grafting clip and transferred to a humidity chamber or healing room. Grafted plants should not be older than 33 days before transplanting (Figs. 12–14 and 12–15, page 478) (21).

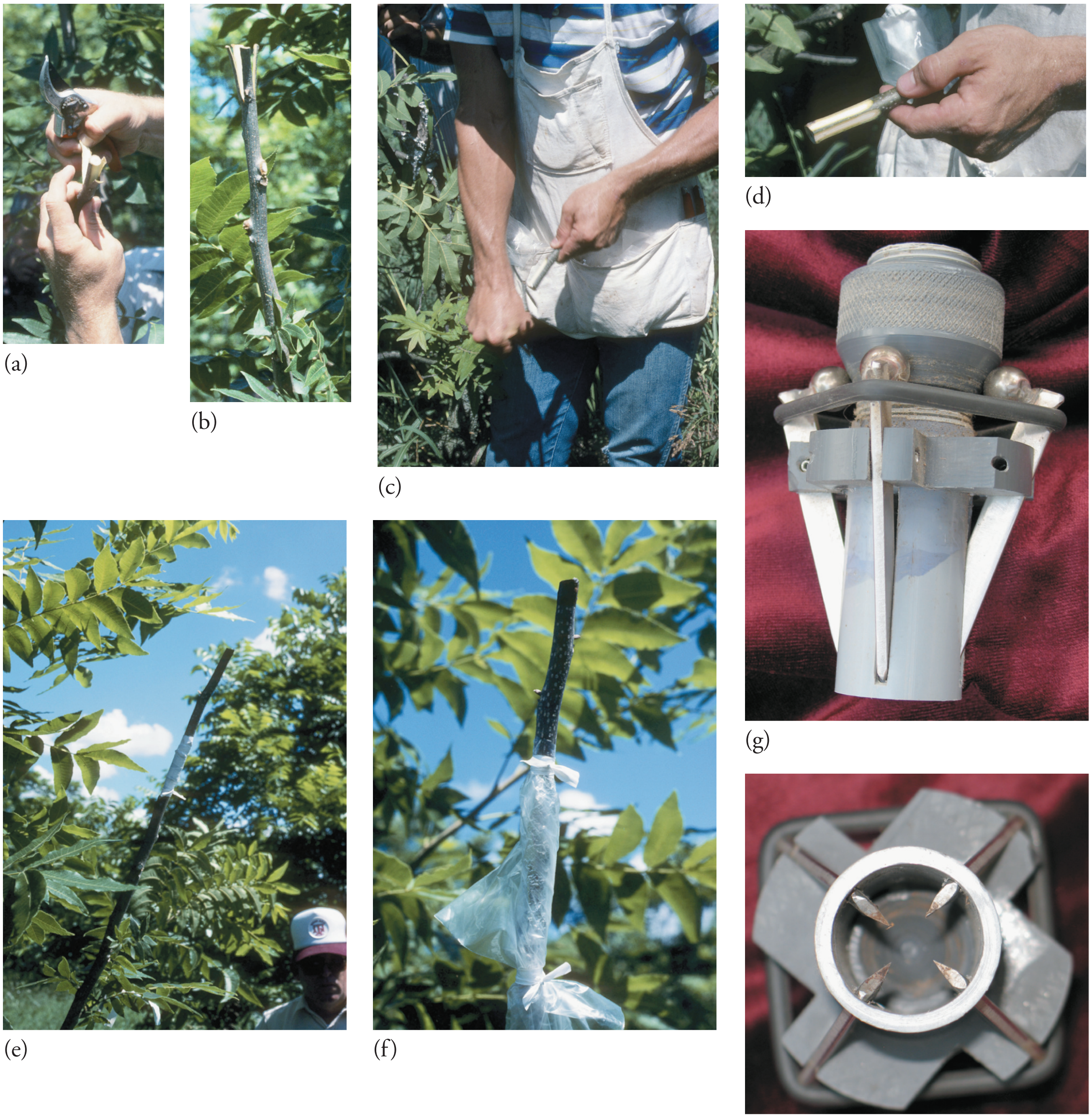
(a) (b) (c) (d) (e) (f) (g) (h)

Figure 12–13
Steps in the four-flap or banana graft. (a) Preparing the rootstock by severing the rootstock top. (b) The four bark flaps of the rootstock with the "wood" of the rootstock removed. (c and d) Prepping the scion by removing the bark of the scion with wood left intact. (e and f) Flaps of the rootstock cover the cut surfaces of the scionwood and are temporarily held by grafting tape, then covered with aluminum foil and poly to prevent desiccation. (g and h) A tool for the four-flap graft, which is slid over the rootstock and used to cut and peal the four flaps.

Detached Scion Graftage—Side Grafting

There are many types of side graftage. As the name suggests, the scion is inserted into the side of the rootstock, which is generally larger in diameter than the scion. This method has proven useful for large-scale propagation of nursery trees (36). Generally, the rootstock shoot is removed after the graft takes, and the scion becomes the dominant shoot system.

Side-Stub Graft The side-stub graft is useful in grafting branches of trees that are too large for the whip-and-tongue graft, yet not large enough for other methods such

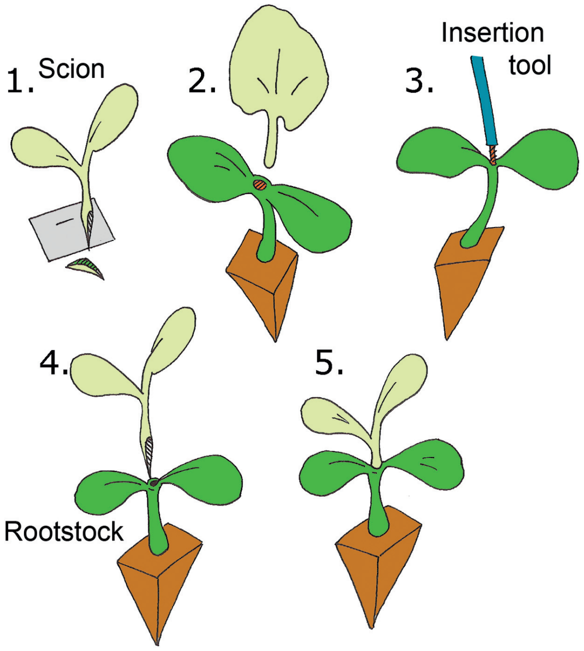

Figure 12–14
Hole insertion graft (HIG) or terminal/top insertion graft (12, 21). Illustrations courtesy R. L. Hassell.

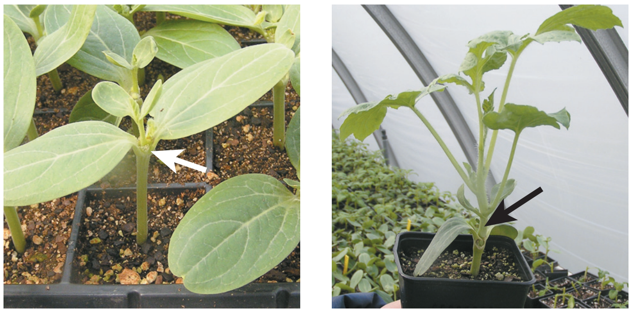
(a) (b)

Figure 12–15
(a and b) Hole insertion graft (HIG) for grafting watermelon to squash rootstock. This is the most popular graft used in China because it is suitable for Lagenaria (Cucurbita) and interspecific squash as rootstocks, requires few materials, has high efficiency, 1500+ plants/day/worker, and simpler management techniques (34). Photos courtesy of R. L. Hassell.

as the cleft or bark graft. For this type of side graft, the best rootstocks are branches about 2.5 cm (1 in) in diameter. An oblique cut is made into the rootstock branch with a chisel or heavy knife at an angle of 20 to 30 degrees. The cut should be about 2.5 cm (1 in) deep and at such an angle and depth that when the branch is pulled back, the cut will open slightly but will close when the pull is released.

The scion should contain two or three buds and be about 7.5 cm (3 in) long and relatively thin. At the basal end of the scion, a wedge about 2.5 cm (1 in) long is made. The cuts on both sides of the scion should be very smooth, each made by one single cut with a sharp knife. The scion must be inserted into the rootstock at an angle, as shown in Figure 12–16, to obtain maximum contact of the vascular cambium layers. The grafter inserts the scion into the cut while the upper part of the rootstock is pulled backward, being careful to obtain the best cambium contact, then the rootstock is released. The pressure of the rootstock should grip the scion tightly. The scion can be further secured by driving two small flat-headed wire nails [20 gauge, 1.5 cm (5/8 in) long] into the stock through the scion. Wrapping the rootstock and scion at the point of union with nursery tape also may be helpful. After the graft is completed, the rootstock may be cut off just above the union. This must be done very carefully or the scion may become dislodged. The entire graft union must be thoroughly covered with grafting wax, sealing all openings. The tip of the scion also should be covered with wax or sealed with white glue (57).

Side-Tongue Graft The side-tongue graft, shown in Figure 12–17, page 480, is useful for small plants, especially some of the broad- and narrow-leaved evergreen species. The rootstock plant should have a smooth section in the stem just above the crown of the plant. The diameter of the scion should be slightly smaller than that of the rootstock. The cuts at the base of the scion are made in the same way as for the whip-and-tongue graft. Along a smooth portion of the stem of the rootstock a thin piece of bark and wood, the same length as the cut surface of the scion, is completely removed. Then a reverse cut is made downward in the cut on the rootstock starting one-third of the distance from the top of the cut. This second cut in the rootstock should be the same length as the reverse cut in the scion. The scion is then inserted into the cut in the rootstock, the two tongues interlocking, and the vascular cambia matching. The graft is wrapped tightly, using one of the methods described for the whip-and-tongue graft.

The top of the rootstock is left intact for several weeks until the graft union has started to heal. Then it may be cut back above the scion gradually or all at once to force the buds on the scion into active growth.

Side-Veneer Graft The side-veneer graft is widely used for grafting small potted liner plants such as seedling conifers, deciduous trees and shrubs, and fruit crops (Figs. 12–18, 12–19, and 12–20, pages 480–83). A shallow downward and inward cut from 25 to 38 mm (1 to 1 1/2 in) long is made in a smooth area just above the crown of the rootstock. At the base of this cut, a second

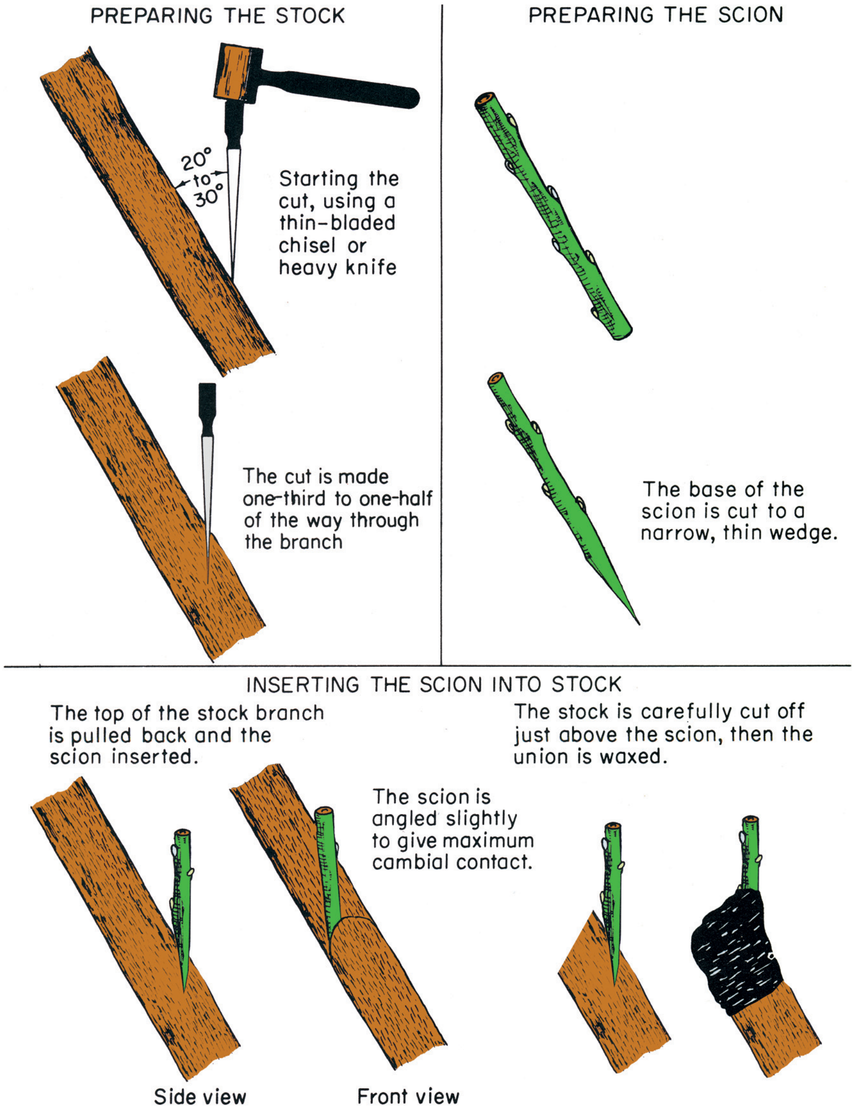
PREPARING THE STOCK
20° to 30°
Starting the cut, using a thin-bladed chisel or heavy knife
The cut is made one-third to one-half of the way through the branch
PREPARING THE SCION
The base of the scion is cut to a narrow, thin wedge.
INSERTING THE SCION INTO STOCK
The top of the stock branch is pulled back and the scion inserted.
The scion is angled slightly to give maximum cambial contact.
The stock is carefully cut off just above the scion, then the union is waxed.
Side view
Front view

Figure 12–16
Steps in preparing the side-stub graft. A thin-bladed chisel, as illustrated here, is ideal for making the cut, but a heavy butcher knife could be used satisfactorily.

short inward and downward cut is made, intersecting the first cut, that removes the piece of wood and bark. The scion is prepared with a long cut along one side and a very short one at the base of the scion on the opposite side. These scion cuts should be the same length and width as those made in the rootstock so that the vascular cambium layers can be matched as closely as possible.

After inserting the scion, the graft is tightly wrapped with poly budding strips, budding rubbers, Buddy Tape, or with nursery adhesive tape. The graft may or may not be covered with wax, depending upon the species. A common practice in side grafting small potted plants of some woody ornamental species is to plunge the grafted plants into a slightly moist medium, such as peat moss, so that it just covers the graft union. Inserting the grafted liner plants in polytents in temperature-controlled greenhouses is another common practice (Fig. 12–4). To maintain high humidity, the newly grafted plants may also be placed for healing in a mist propagating house (but the grafts are not directly placed under mist), or set in grafting cases. The latter are closed boxes with a transparent cover, which permits retention of high humidity around the grafted plant until the union has healed. The grafting cases are kept closed for a week or so after the grafts are put in, and then gradually opened over a period of several weeks; finally, the cover is taken off completely.

After the union has healed, the rootstock can be cut back above the scion either in gradual steps or all at once.

Side Insertion Graft (SIG) The Side Insertion Graft (SIG) has been largely replaced by the OCG or Japanese tube graft, hole insertion graft (HIG), and tongue approach graft (TAG) (12, 21). The SIG is suitable for rootstocks with wide hypocotyls (Fig. 12–21, page 483). Production of rootstocks and scions is the

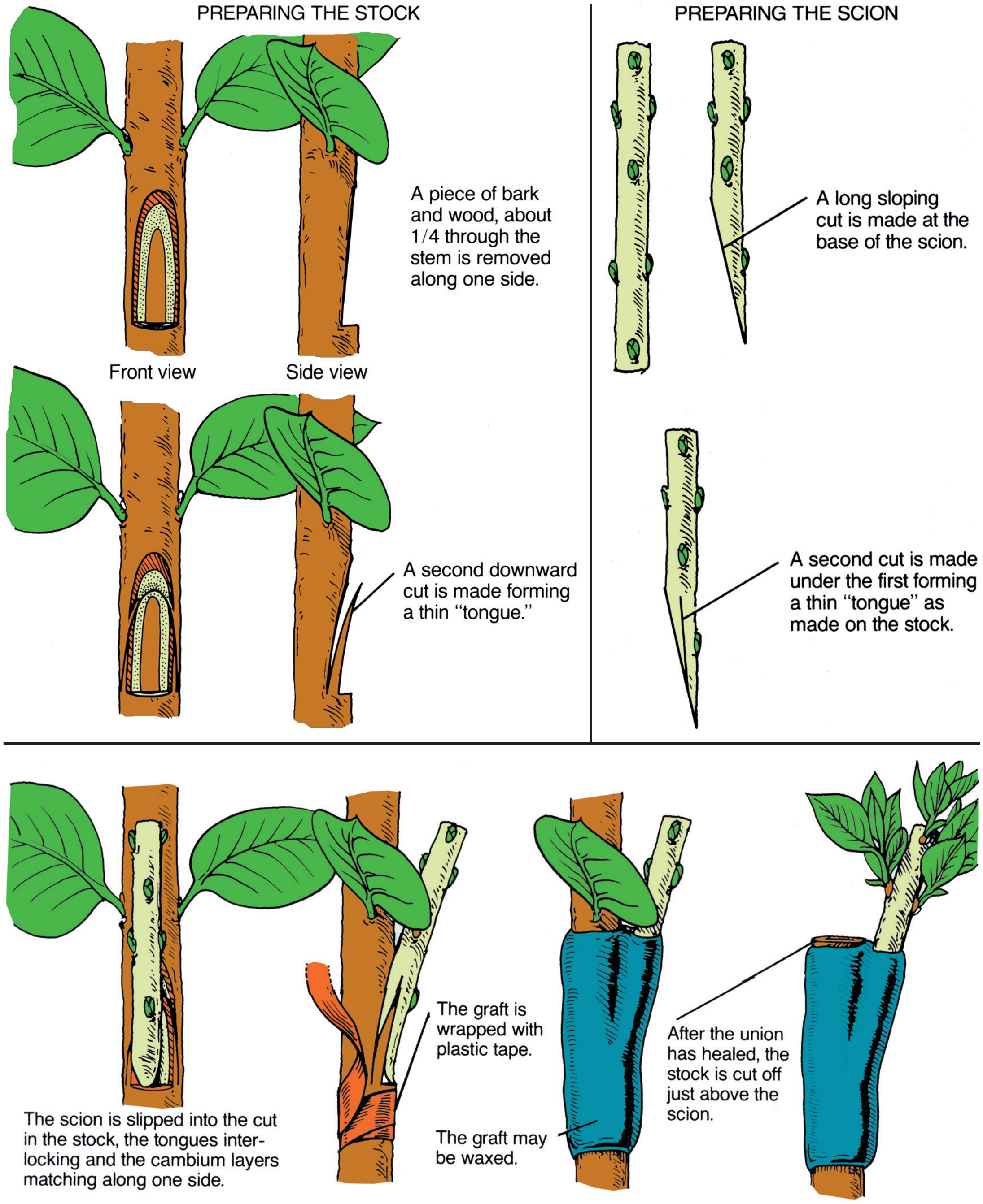
PREPARING THE STOCK
PREPARING THE SCION
A piece of bark and wood, about 1/4 through the stem is removed along one side.
Front view
Side view
A long sloping cut is made at the base of the scion.
A second downward cut is made forming a thin "tongue."
A second cut is made under the first forming a thin "tongue" as made on the stock.
The scion is slipped into the cut in the stock, the tongues interlocking and the cambium layers matching along one side.
The graft is wrapped with plastic tape.
The graft may be waxed.
After the union has healed, the stock is cut off just above the scion.

Figure 12–17
Sidetongue graft. This method is very useful for grafting broadleaved evergreen plants. Final tying may be done with budding rubbers, poly tape, or waxed string. The graft may be waxed, or wrapped with a sealing tape such as Parafilm or Buddy Tape.

same as that described for hole insertion grafting. A slit is cut on the hypocotyl of the rootstock with a razor blade and held open with a toothpick. A 35- to 45-degree-angle cut, on both sides is made on the hypocotyl of the scion. Then the scion is inserted into the slit in the hypocotyl of the rootstock and the toothpick is removed. Two cut surfaces are matched together and held with a grafting clip or silicone sleeve. The top of the rootstock is cut off 5 days after grafted plants are moved from the high-humidity growth chamber (21).

Detached Scion Graftage—Bark Grafting

Bark grafting is done in topworking established plants. The rootstock must be in an active stage of growth so that the bark will slip. The scion is inserted between the bark and wood of the rootstock. Bark grafting can be

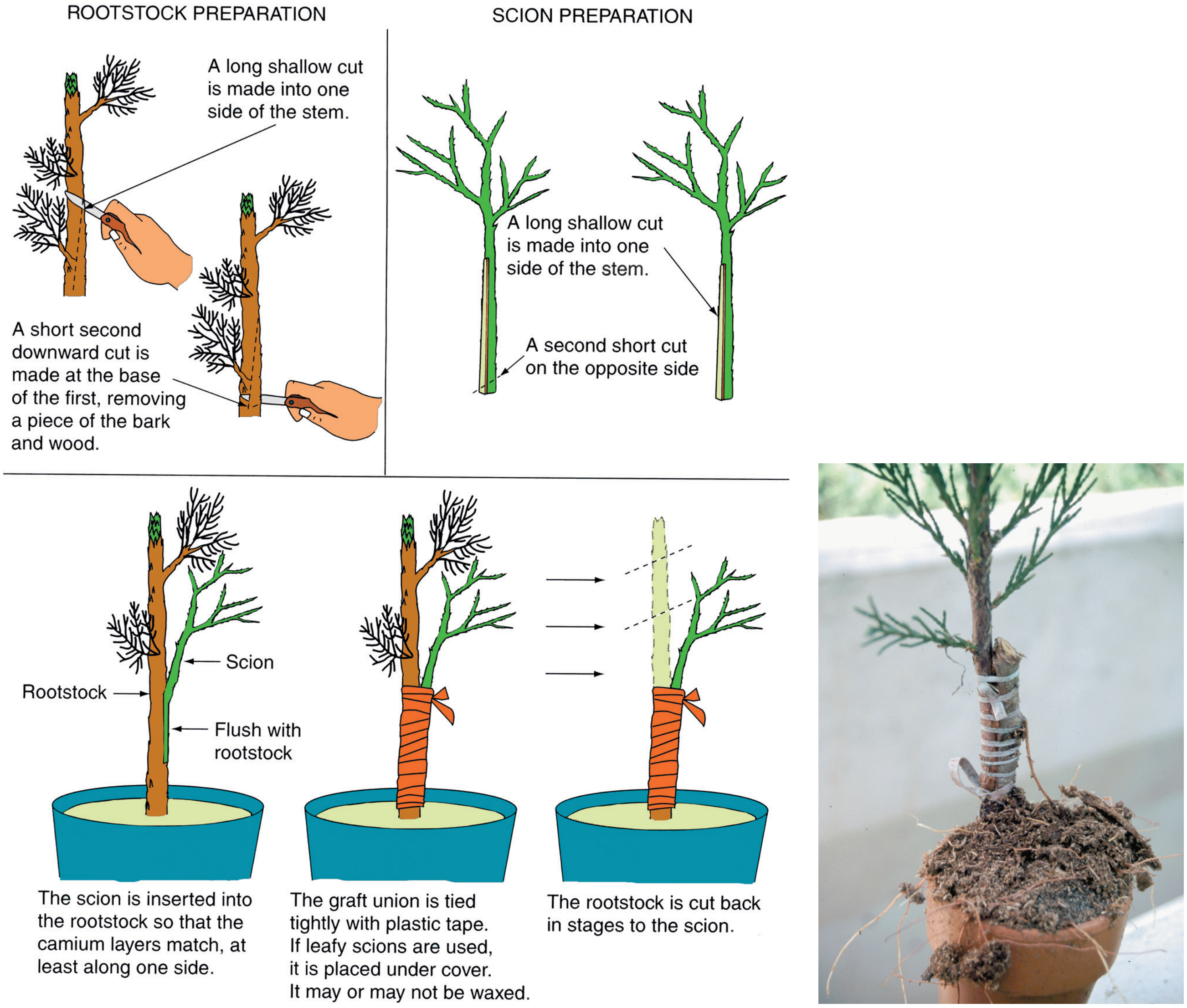
ROOTSTOCK PREPARATION
A long shallow cut is made into one side of the stem.
A short second downward cut is made at the base of the first, removing a piece of the bark and wood.
SCION PREPARATION
A long shallow cut is made into one side of the stem.
A second short cut on the opposite side
Scion
Rootstock
Flush with rootstock
The scion is inserted into the rootstock so that the camium layers match, at least along one side.
The graft union is tied tightly with plastic tape. If leafy scions are used, it is placed under cover. It may or may not be waxed.
The rootstock is cut back in stages to the scion.

(a) (b)

Figure 12–18
(a) Steps in making the side-veneer graft. This method is one of the most popular grafts for propagating conifers and shrubs. The graft is quite versatile and can be used on a larger number of species than other grafts, such as the whip-and-tongue. (b) Side-veneer grafted connifer.

performed on branches ranging from 2.5 cm (1 in) up to 30 cm (1 ft) or more in diameter. The latter size is not recommended, because it is difficult to heal over such large stubs before decay-producing organisms attack.

Scions must be collected for deciduous species during the dormant season and held under refrigeration. For evergreen species, freshly collected scionwood can be used. In the bark graft, scions are not as securely attached to the rootstock as in some of the other methods and are more susceptible to wind breakage during the first year, even though healing has been satisfactory. Therefore, the new shoots arising from the scions probably should be staked during the first year, or cut back to about half their length, especially in windy areas. After a few years' growth, the bark graft union is as strong as the unions formed by other methods. *Two modifications of the bark graft are described next.*

Bark Graft (Rind Graft) Several scions are inserted into each rootstock stub (Fig. 12–22, page 484). For each scion, a vertical knife cut 2.5 to 5 cm (1 to 2 in) long is made at the top end of the rootstock stub through the bark to the wood. The bark is then lifted slightly along both sides of this cut, in preparation for the insertion of the scion. The dormant scions should be 10 to 13 cm (4 to 5 in) long, contain 2 or 3 buds, and be 6 to 13 mm (1/4 to 1/2 in) thick. One cut—about 5 cm (2 in) long— is made along one side at the base of the scion. With large scions,

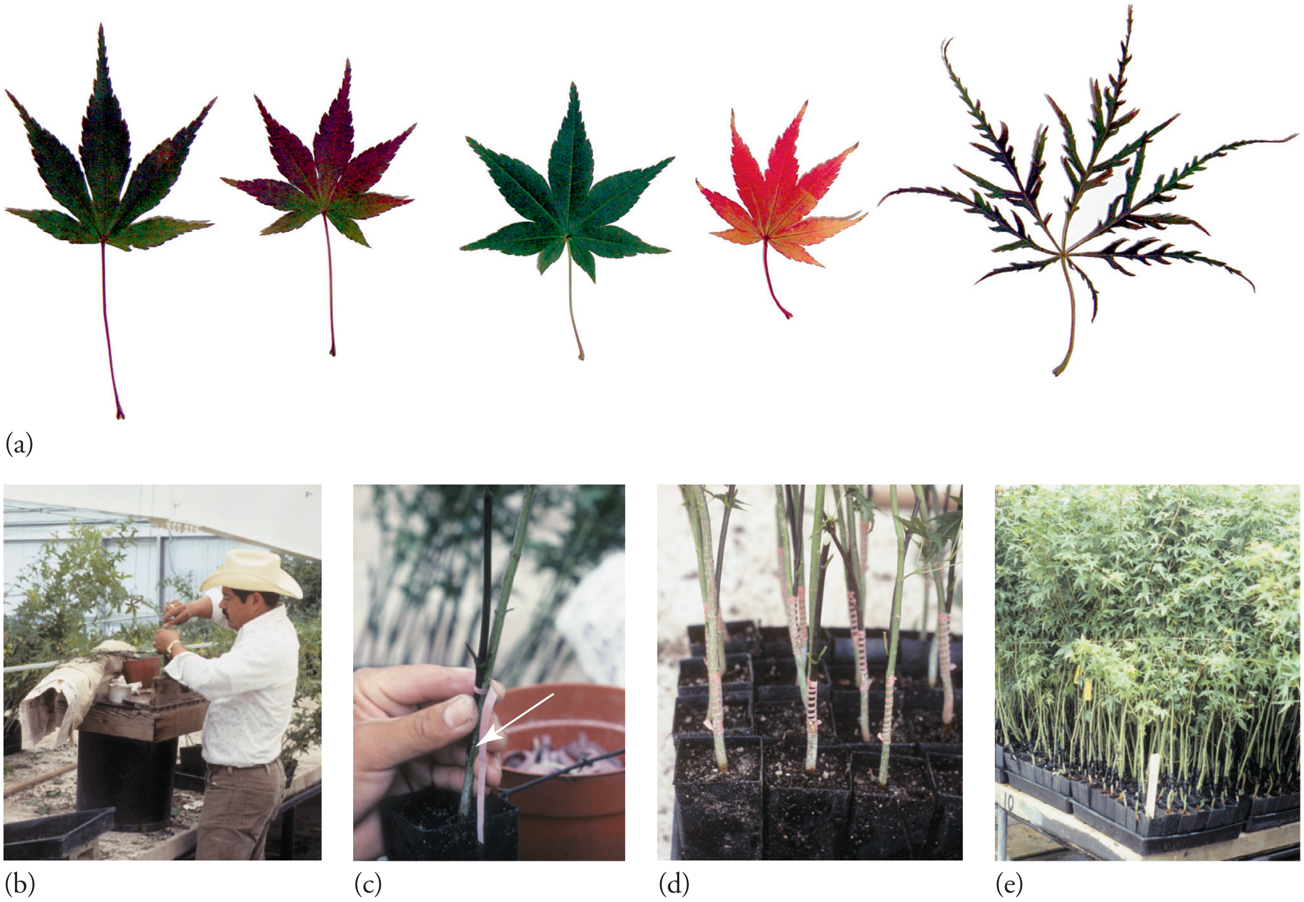

Figure 12–19
(a) Cultivars of the highly diverse Japanese maple (Acer palmatum) are grafted onto seedling rootstock. (b and c) A budding rubber (arrow) is used to wrap the dormant, leafless scion to the rootstock. (d) Side-veneer-grafted Japanese maples in potted liner pots which will be moved to a polytent area for callusing, and (e) post-allusing liner production.

this cut extends about one-third of the way into the scion, leaving a "shoulder" at the top. This shoulder reduces the thickness of the scion to minimize the separation of bark and wood after insertion in the rootstock. The scion should not be cut too thin, or it will be mechanically weak and break off at the point of attachment to the rootstock. If small scions are used, no shoulder is necessary. On the side of the scion opposite the first long cut, a second, shorter cut is made, as shown in Figure 12–22, bringing the basal end of the scion to a wedge shape. The scion is then inserted between the bark and the wood of the rootstock, centered directly under the vertical cut through the bark. The longer cut on the scion is placed against the wood, and the scion's shoulder is brought down until it rests on top of the stub. The scion is then ready to be fastened in place. The scion is nailed into the wood, using two nails per scion. Flat-headed nails 15 to 25 mm (5/8 to 1 in) long, of 19- or 20-gauge wire, depending on the size of the scions, are satisfactory. The bark on both sides of the scion should be nailed down securely or it will tend to peel back from the wood.

Another method commonly used with soft-barked trees, such as the avocado, is to insert all the scions in the stub and then hold them in place by wrapping waxed string, adhesive tape, or poly budding tape around the stub. This method is more effective than nailing for preventing the scions from blowing out, but probably does not give as tight a fit. A combination of nailing and wrapping are advisable for maximum strength. If a wrapping material is used, it must be checked to avoid constricting the rootstock. After the stub has been grafted and the scions fastened by nailing or tying, all cut surfaces, including the end of the scions, should be covered thoroughly with grafting wax.

Inlay Bark Graft Two knife cuts about 5 cm (2 in) long are made through the bark of the rootstock down to the wood, rather than just one (Fig. 12–23, page 485). The distance between these two cuts should be exactly the same as the width of the scion. The piece of bark between the cuts should be lifted and the terminal two-thirds cut off. The scion is prepared with a smooth

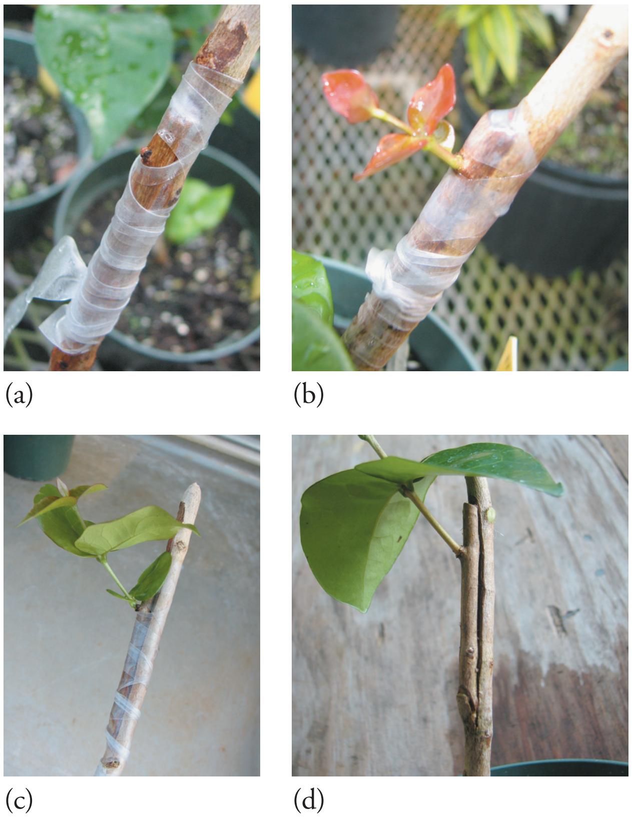

Figure 12–20
Side veneer graft of Eugenia (Myrtaceae). (a, b, and c) Graft wrapped with Parafilm tape. (d) Healed graft. Photos courtesy John Griffis.

slanting cut along one side at the basal end, completely through the scion. This cut should be about 5 cm (2 in) long but without the shoulder, in contrast to the bark graft. On the opposite side of the scion, a cut about 13 mm (1/2 in) long is made, forming a wedge at the base of the scion. The scion should fit snugly into the opening in the bark with the longer cut inward and with the wedge at the base slipped under the flap of remaining bark.

The scion should be nailed into place with two nails, the lower nail going through the flap of bark covering the short cut on the back of the scion. If the bark along the sides of the scion should accidentally become disturbed, it must be nailed back into place. Flat-point staples in the vertical position, or budding or flagging tape have all been used to secure the graft (40). The inlay bark graft is well adapted for use with thick-barked trees, such as walnuts and pecans, on which it is not feasible to insert the scion under the bark; it is used when topworking an existing orchard (Fig. 12–24, page 486).

Detached Scion Graftage—Root Graftage

A number of plants are propagated commercially by root grafting—apples, pears, grapes, and selected woody ornamental shrubs and trees (17, 18).

Root Grafting (Whole-Root and Piece-Root Graftage)

In root grafting, the rootstock seedling, rooted cutting, or layered plant is dug up, and the roots are used as the rootstock for the graft. The entire root system may be used (**whole-root graft**—Figs. 12–3 and 12–25, page 487), or the roots may be cut up into small pieces and each piece used as a rootstock (**piece-root graft**—Fig. 12–25). Both methods give satisfactory results. Since the roots used are relatively small [0.6 to 1.3 cm (1/4 to 1/2 in) in diameter], the **whip-and-tongue graft** is frequently used. In England, Rhododendron cultivars are saddle-grafted on roots of *R. ponticum;* the root graft is then tied and placed in a propagation case (19). Tree peony and herbaceous peony are root-grafted with a cleft graft using the root of herbaceous peony. Root grafts are usually bench-grafted indoors during the late winter or early spring. The scionwood collected previously is held in storage, while the rootstock plants are also dug in the late fall and stored under cool [1.5 to 4.5°C (35 to 40°F)] and moist conditions until the grafting is done. The term *bench grafting* is given to this process, because it is performed indoors with dormant scions and rootstocks at benches by skilled grafters as part of a large-scale operation.

In making root grafts, the root pieces should be 7.5 to 15.0 cm (3 to 6 in) long and the scions about the same length, containing 2 to 4 buds. After the grafts are

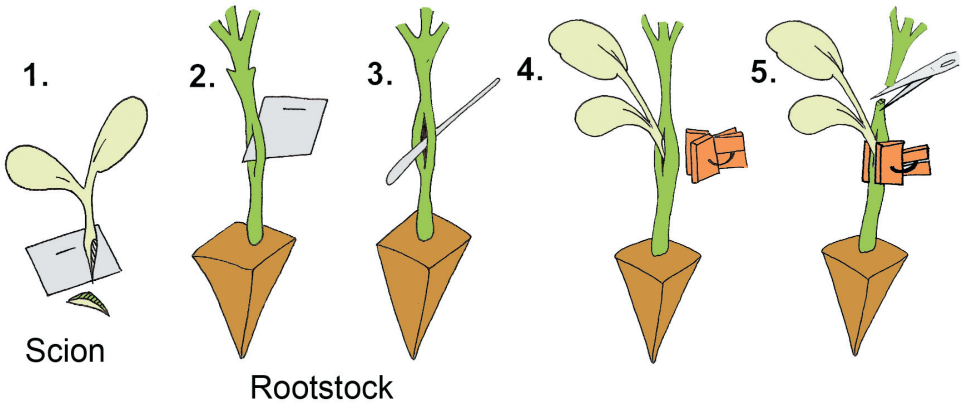

Figure 12–21
Steps in preparing the Side Insertion Graft (SIG). The SIG has been largely replaced by the OCG or Japanese tube graft hole insertion graft (HIG), and tongue approach graft (TAG). The SIG is suitable for rootstocks with wide hypocotyls (12, 21). Illustrations courtesy R. L. Hassell.

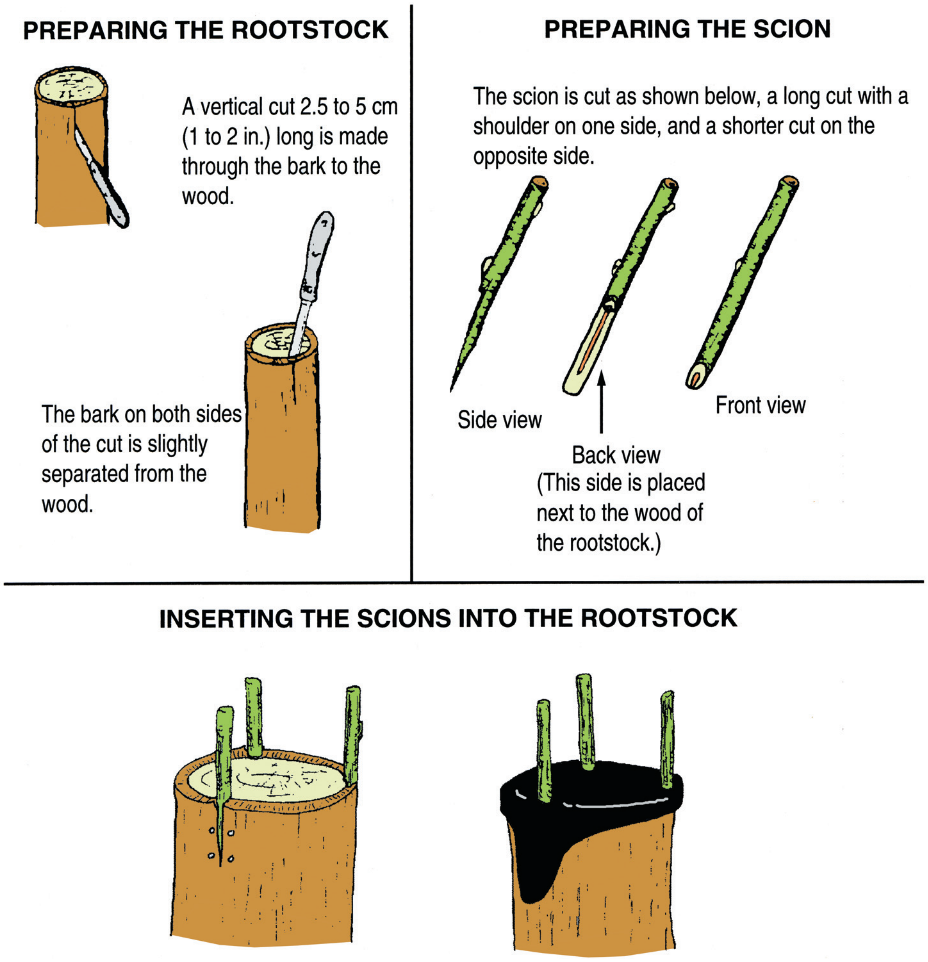
PREPARING THE ROOTSTOCK
A vertical cut 2.5 to 5 cm (1 to 2 in.) long is made through the bark to the wood.
The bark on both sides of the cut is slightly separated from the wood.
PREPARING THE SCION
The scion is cut as shown below, a long cut with a shoulder on one side, and a shorter cut on the opposite side.
Side view
Back view (This side is placed next to the wood of the rootstock.)
Front view
INSERTING THE SCIONS INTO THE ROOTSTOCK

Figure 12–22
Steps in preparing the bark graft (rind graft). In grafting some thick-barked plants, the vertical cut in the bark is unnecessary; the scion is inserted between the bark and wood of the stock.

made and properly tied, they are bundled together in groups of 50 to 100 and stored for callusing in damp sand, peat moss, or other packing material.

Nurse-Root Grafting Stem cuttings of a difficult-to-root species can sometimes be induced to develop adventitious roots by making a temporary "nurse-root" graft. The plant to be grown on its own roots is temporarily grafted as the scion. The scion may be made longer than usual and the graft planted deeply, with the major portion of the scion below ground. **Scion rooting** can be promoted by applying an auxin, such as indole-3-butyric acid, into several vertical cuts made through the bark at the base of the scion, above the graft union before planting. The grafts are set deeply, so that most of the scion is covered (mound layered) with soil (28). After one season of growth the scions have roots, and the temporary nurse rootstock is cut off and discarded. The rooted scion is replanted to grow on its own roots; it can later be used as a rootstock and grafted to a scion fruit cultivar.

scion rooting The development of adventitious roots from the grafted scion, desirable in **nurse-root grafting**, which is a temporary graft, but problematic in other grafts where the size control or disease-resistant characteristics of the rootstock may be lost.

Methods of nurse-root grafting include *reversing the polarity of the nurse-root rootstock.* The rootstock piece will eventually die if it is grafted onto the scion in an inverted position (Fig. 12–26, page 487) (37). A graft union is formed—the inverted rootstock piece sustains the scion until it roots—but the rootstock fails to receive sufficient carbohydrates from the scion and eventually dies, leaving the scion on its own roots. Another method is *girdling the rootstock just above the graft union at the scion base.* The rootstock is girdled with budding rubber strips (0.016 gauge) (7). Budding rubbers disintegrate within a month when exposed to sun and air; however, when buried in the soil, they will last as long as 2 years, allowing sufficient time for the scion to become rooted. In a third method, an *incompatible rootstock is used.* When the graft is planted deeply, scion roots will gradually become more dominant in sustaining the plant. Examples of this are apple scions on pear rootstock, and lilac scions on ash rootstock.

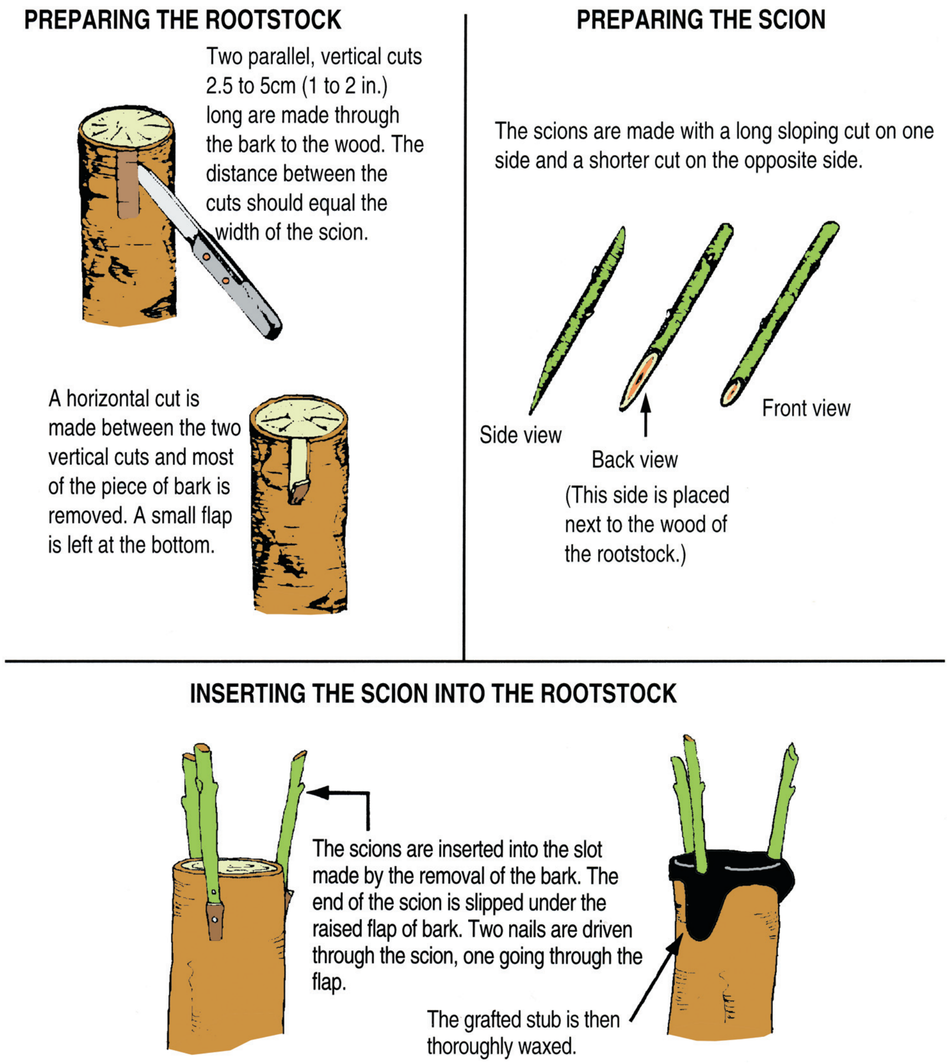

PREPARING THE ROOTSTOCK

Two parallel, vertical cuts 2.5 to 5cm (1 to 2 in.) long are made through the bark to the wood. The distance between the cuts should equal the width of the scion.

A horizontal cut is made between the two vertical cuts and most of the piece of bark is removed. A small flap is left at the bottom.

PREPARING THE SCION

The scions are made with a long sloping cut on one side and a shorter cut on the opposite side.

(This side is placed next to the wood of the rootstock.)

INSERTING THE SCION INTO THE ROOTSTOCK

The scions are inserted into the slot made by the removal of the bark. The end of the scion is slipped under the raised flap of bark. Two nails are driven through the scion, one going through the flap.

The grafted stub is then thoroughly waxed.

Figure 12–23
Inlay bark graft. With Texas pecans, the scion of the inlay bark graft is stapled or nailed, and aluminum foil and a cut poly bag are used instead of grafting wax (40).

Approach Graftage

The distinguishing feature of approach grafting is that two independent, self-sustaining plants are grafted together. After a union has occurred, the top of the rootstock plant is removed above the graft, and the base of the scion plant is removed below the graft. Sometimes it is necessary to sever these parts gradually rather than all at once. Approach grafting provides a means of establishing a graft union between certain plants which are otherwise difficult to successfully graft. It is usually performed with one or both of the plants growing in a container. Rootstock plants in containers may also be placed adjoining an established plant that is to furnish the scion part of the new, grafted plant (Fig. 12–27, page 488).

This type of grafting should be done at times of the year when growth is active and rapid healing of the graft union will take place. Three useful methods of making approach grafts are described as follows, and illustrated in Figure 12–28.

Spliced Approach Graft In the spliced approach graft, the two stems should be approximately the same size (Fig. 12–28, page 489). An exception to this is the spliced approach graft of mango in India, where the scion is considerably smaller than the field-grown rootstock; the scion, in a pot, is hung from the branch of the larger rootstock (19). At the point where the union is to occur, a slice of bark and wood 2.5 to 5 cm (1 to 2 in) long is cut from both stems. This cut should be the same size on each so that identical cambium patterns are made. The cuts must be perfectly smooth and as nearly flat as possible so that when they are pressed together there will be close contact of the vascular cambium layers. The two cut surfaces are bound tightly together with raffia or poly grafting tape, then the

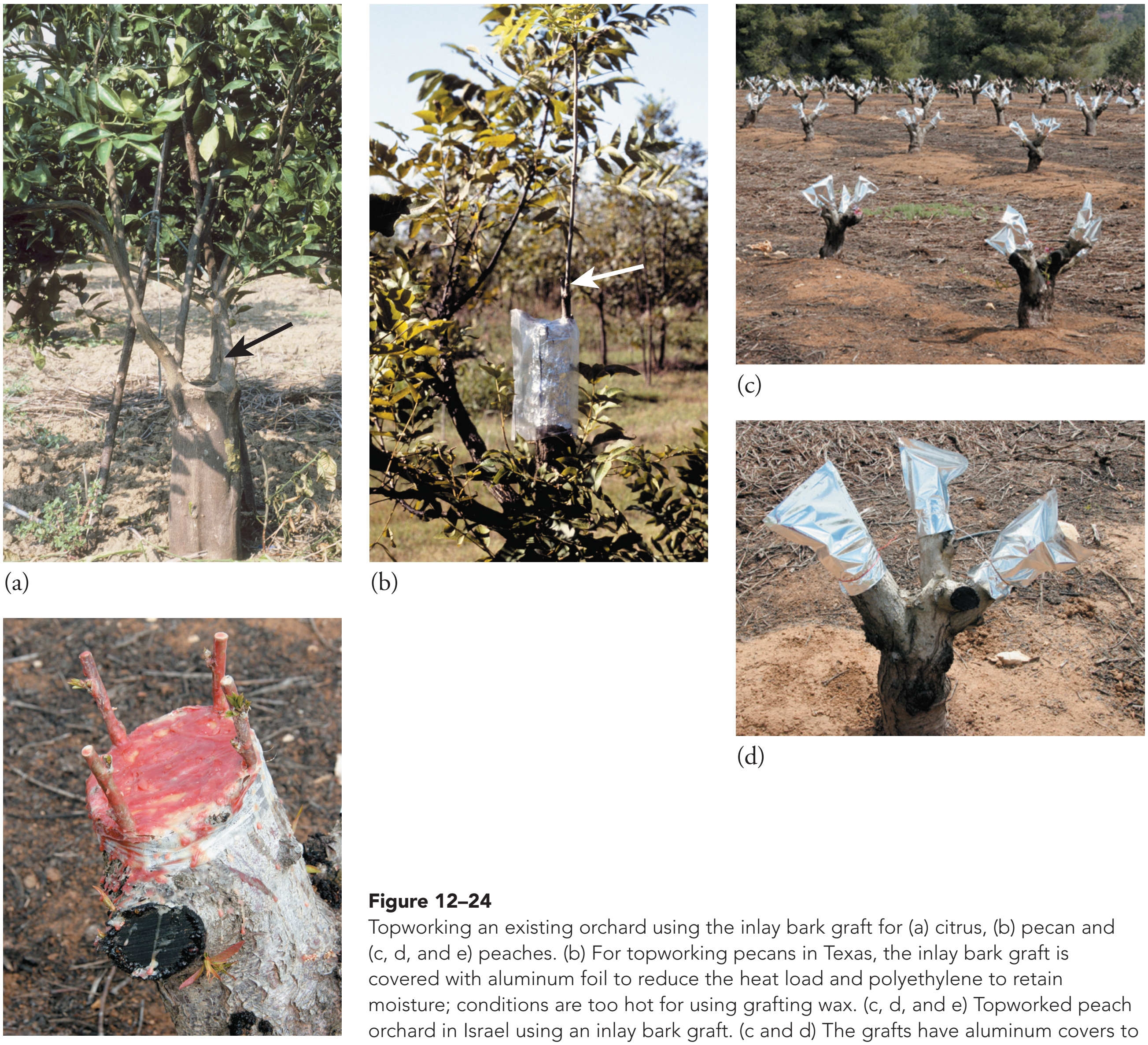
(a) (b) (c) (d) (e)

Figure 12–24
Topworking an existing orchard using the inlay bark graft for (a) citrus, (b) pecan and (c, d, and e) peaches. (b) For topworking pecans in Texas, the inlay bark graft is covered with aluminum foil to reduce the heat load and polyethylene to retain moisture; conditions are too hot for using grafting wax. (c, d, and e) Topworked peach orchard in Israel using an inlay bark graft. (c and d) The grafts have aluminum covers to reduce heat buildup.

whole union should be covered with grafting wax. After the parts are well united (which may require considerable time in some cases) the rootstock above the union, and the scion below the union are cut, and the graft is completed. It may be necessary to reduce the leaf area of the scion if it is more than the root system of the rootstock can initially sustain.

Tongued Approach Graft (TAG) The tongued approach graft is the same as the spliced approach graft, except that after the first cut is made in each stem to be joined, a second cut—downward on the stock and upward on the scion—is made, thus providing a thin tongue on each piece. By interlocking these tongues a very tight, closely fitting graft union can be obtained (Fig. 12–29, page 489). For grafting vegetable crops, after the rootstock has fully developed cotyledons and scion has cotyledon and first true leaf, plants are pulled out from the tray (21). Make a cut at a 35- to 45-degree angle into the hypocotyl of the rootstock approximately halfway with a razor blade, and make a cut of the opposite angle on the hypocotyl of the scion. Cuts need to be made so that the scion will be on top of the rootstock when completed. Two cut hypocotyls are placed together and sealed with aluminum foil or Buddy Tape to help healing and prevent the graft from drying out. The two plants are transplanted into a bigger cell that will accommodate the two root balls. The top of the rootstock is cut off 5 days after grafting, and the bottom of the scion is cut off 7 days after the top of the rootstock is removed (Fig. 12–29).

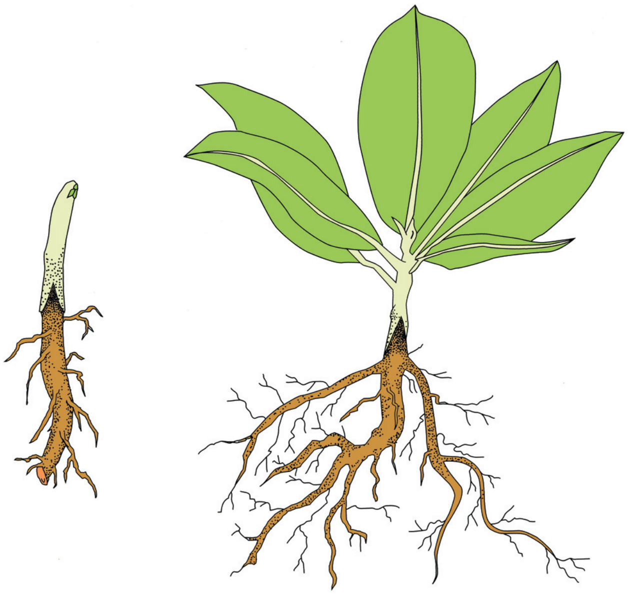

Figure 12–25
Root graftage with a saddle graft on a small root section or piece-root (*left*) and a whole root rootstock (*right*). See Figure 12–3 which is a whip and tongue graft using a whole root apple rootstock.

Inlay Approach Graft The inlay approach graft may be used if the bark of the rootstock plant is considerably thicker than that of the scion plant. A narrow slot, 7.5 to 10 cm (3 to 4 in) long, is made in the bark of the rootstock plant by cutting two parallel channels and removing the strip of bark between (Fig. 12–28); this can be done only when the rootstock plant is actively growing and the bark "slipping." The slot should be exactly as wide as the scion to be inserted. The stem of the scion plant, at the point of union, should be given a long, shallow cut along one side, of the same length as the slot in the rootstock plant and deep enough to go through the bark and slightly into the wood. This cut surface of the scion branch should be laid into the slot cut in the rootstock plant and held there by nailing with two or more small, flat-headed wire nails. Then the entire union must be thoroughly covered with grafting wax. After the union has healed, the rootstock can be cut off above the graft and the scion below the graft.

Repair Graftage

Inarching Inarching is similar to approach grafting in that both rootstock and scion plants are on their own roots at the time of grafting. It differs in that the top of the new rootstock plant usually does not extend above the point of the graft union, as it does in approach grafting. Inarching is used to replace roots damaged by cultivation equipment, rodents, or disease. It can be used to very good advantage to save a valuable tree or improve its root system (Fig. 12–30, page 489).

Seedlings (or rooted cuttings) planted beside the damaged tree, or suckers arising near its base, are grafted into the trunk of the tree to provide a new root system to supplant the damaged roots. The seedlings to be inarched into the tree should be spaced about 13 to 15 cm (5 to 6 in) apart around the circumference of the tree if the damage is extensive. A damaged tree usually will stay alive for some time unless the injury is very severe. The procedure for inarching is to plant seedlings of a compatible species around the tree during the dormant season, and graft when active growth commences in early spring. Inarching may also enhance growth of uninjured, older trees (22).

As illustrated in Figures 12–30 and 12–31 (page 490), the graft is similar to an inlay bark graft. The upper end of the seedling, which should be 6 to 13 mm (1/4 to 1/2 in) thick, is cut shallowly along the side for 10 to 15 cm (4 to 6 in). This cut should be on the side next to the trunk of the tree and deep enough to remove some of the wood, exposing two strips of cambium tissue. Another, shorter cut, about 13 mm (1/2 in) long, is made on the side opposite the long cut, creating a sharp, wedge-shaped end on the seedling stem.

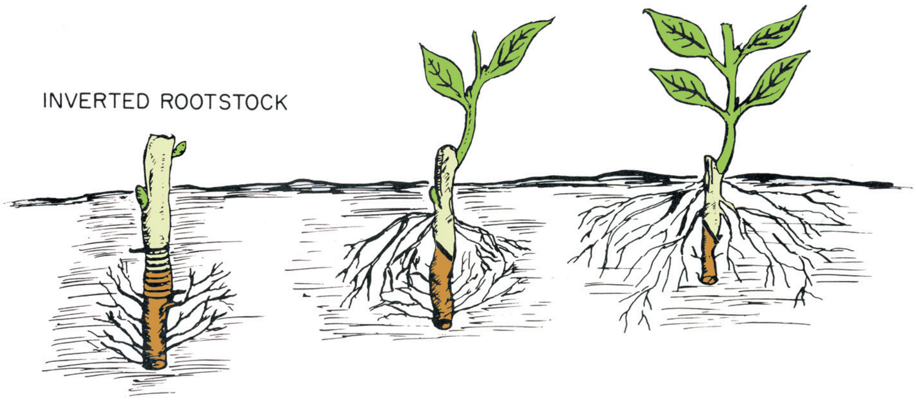

Figure 12–26
Reversing the polarity of the rootstock piece of the root graft is one method of "nurse-root" grafting. The nurseroot graft is a temporary graft used to induce the scion to develop its own roots. The nurse root sustains the plant until the scion roots form, then it dies. In the method shown, the rootstock piece is inverted, so the distal of the rootstock is temporarily joined to the proximal of the scion.

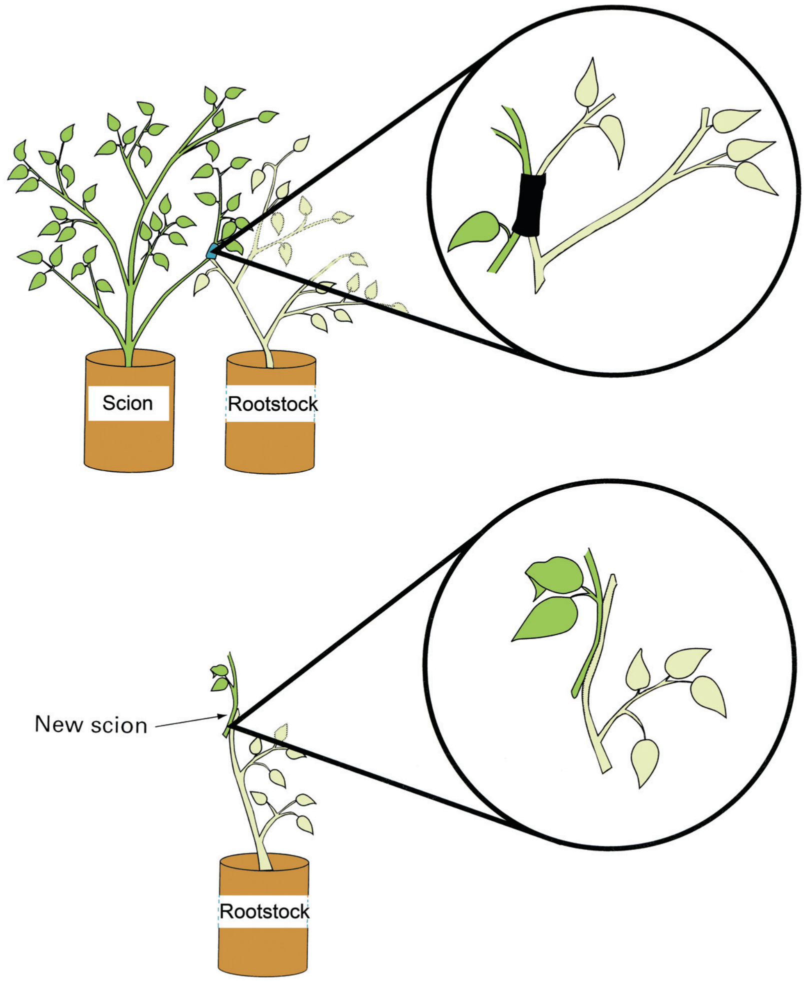

Figure 12–27
Approach grafting. *Top:* Initial grafting of the containerized scion and rootstock plants. *Below:* Completed graft union with scion severed from its own roots, and shoots of rootstock severed above the graft union.

A long slot is cut in the trunk of the tree by removing a piece of bark the width of the seedling and just as long as the cut surface made on the seedling. A small flap of bark is left at the upper end of the slot, under which the wedge end of the seedling is inserted. The seedling is nailed into the slot with four or five small, flat-headed wire nails. The nail at the top of the slot should go through the flap of bark and through the end of the seedling. If the bark of the tree along the sides of the seedling is accidentally pulled loose, it should be nailed back into place. The entire area of the graft union should then be thoroughly waxed.

Bridge Grafting Bridge grafting is another form of repair grafting—used when there is injury to the trunk, such as by cultivation equipment, rodents, disease, or winter injury. If the damage to the bark is extensive, the tree is almost certain to die, because the roots will be deprived of their carbohydrate supply from the top of the tree. Trees of some species, such as the elm, cherry, and pecan, can compartmentalize extensively injured areas by the development of a wound periderm of callus tissue. However, most woody species with severely damaged bark should be bridge grafted if they are to be saved, as illustrated in Figure 12–32, page 491.

An ***interstock bridge graft system*** has been used with mature apple trees for grafting M9 dwarfing rootstock (as the interstock) onto semi-dwarfing apple rootstock, leading to 20 percent reduced shoot growth, but a 30 percent increase in yield and increased soluble sugars and starch in the scion (53). A ring of bark 8 cm wide was removed from the trunk about 30 cm from ground level. Bridge grafts composed of 1-cm-wide split interstocks were inserted perpendicular around the ring, and then tightly wrapped with plastic during graft healing (New Zealand).

Bridge grafting is best performed in early spring as active growth of the tree is beginning and the bark is slipping easily. The scions should be obtained when dormant from 1-year-old growth, 6 to 13 mm (1/4 to 1/2 in) in diameter, of the same or compatible species, and refrigerated until grafted. In an emergency, one

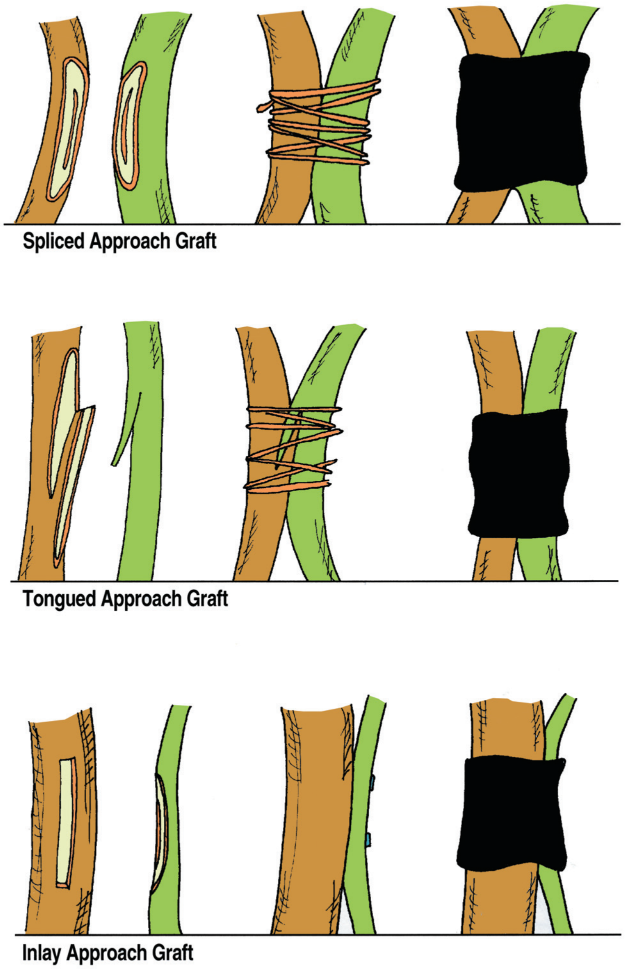

Figure 12–28
Three methods of making an approach graft: spliced-, tongued- and inlay-approach graft.

may successfully perform bridge grafting late in the spring, using scionwood whose buds have already started to grow; the developing buds or new shoots are removed.

The first step in bridge grafting is to trim the wounded area back to healthy, undamaged tissue by removing dead or torn bark. A scion is inserted every 5 to 7.5 cm (2 to 3 in) around the injured section and attached at both the upper and lower ends into live, undamaged bark. It is important that the scions are right side up. If reversed, a union may form, but the scions will not enlarge in diameter as they would if inserted correctly. Figure 12–33, page 491 shows the details of making a bridge graft.

After all the scions have been inserted, the cut surfaces must be thoroughly covered with grafting wax; particular care should be taken to work the wax around the scions, especially at the graft unions.

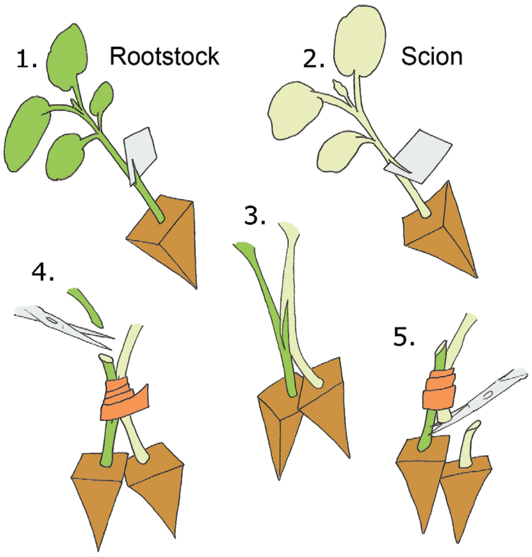

Figure 12–29
Tongued Approach Graft (TAG) with cucurbits. An easy graft that requires no special equipment or graft healing chambers, has a high success rate, but is labor intensive (12, 21). Illustrations courtesy R. L. Hassell.

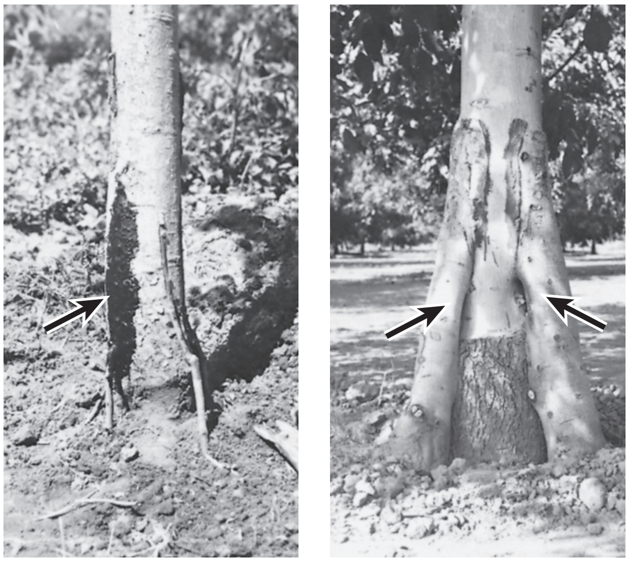
(a) (b)

Figure 12–30
(a) Inarches that have just been inserted (arrows). The one on the left has been waxed. The one on the right has been nailed into place and is ready for waxing. (b) Inarching can be used for invigorating established trees by replacing a weak rootstock with a more vigorous one. Here a Persian walnut tree has been inarched with Paradox hybrid seedlings (Juglans hindsii × J. regia).

Bracing Bracing is a form of natural branch grafting that is used by fruit producers to strengthen scaffolding limbs of a tree in order to support the weight of the fruit crop. Natural grafting of roots or shoot systems occurs in species such as fig (*Ficus*), rubber trees

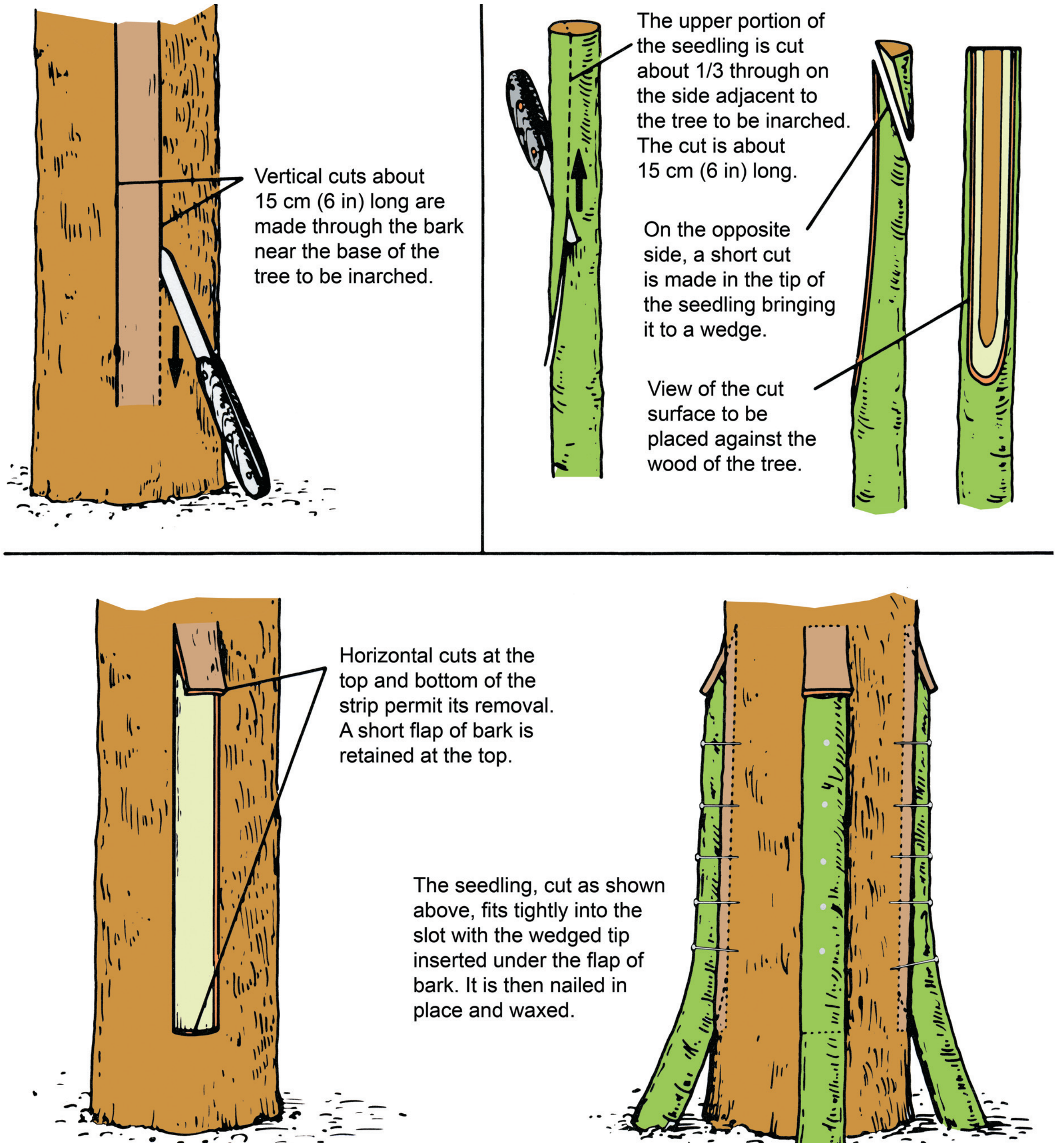

Vertical cuts about 15 cm (6 in) long are made through the bark near the base of the tree to be inarched.

The upper portion of the seedling is cut about 1/3 through on the side adjacent to the tree to be inarched. The cut is about 15 cm (6 in) long.

On the opposite side, a short cut is made in the tip of the seedling bringing it to a wedge.

View of the cut surface to be placed against the wood of the tree.

Horizontal cuts at the top and bottom of the strip permit its removal. A short flap of bark is retained at the top.

The seedling, cut as shown above, fits tightly into the slot with the wedged tip inserted under the flap of bark. It is then nailed in place and waxed.

Figure 12–31
Steps in inarching a large plant, with smaller ones planted around its base.

(*Hevea*), birch (*Betula*), beech (*Fagus*), ash (*Fraxinus*), maple (*Acer*), pine (*Pinus*), and climbing species such as English ivy (*Hedera*) (see Chapter 11). Branches and trunks can naturally graft when they come in contact with each other during early development. The union begins with compression and constant and increasing pressure that ruptures the outer bark of the graft partners, followed by continued secondary growth, which leads to graft union formation and the joining of the independent vascular systems of the partners.

When bracing limbs, fruit producers will pull together two strong, young lateral shoots from the limbs to be braced. A rope or cord is used to temporarily brace the larger limbs. The weaved smaller shoots, which will naturally graft, are tied with waxed string or poly tape to keep them together (Fig. 12–34, page 492).

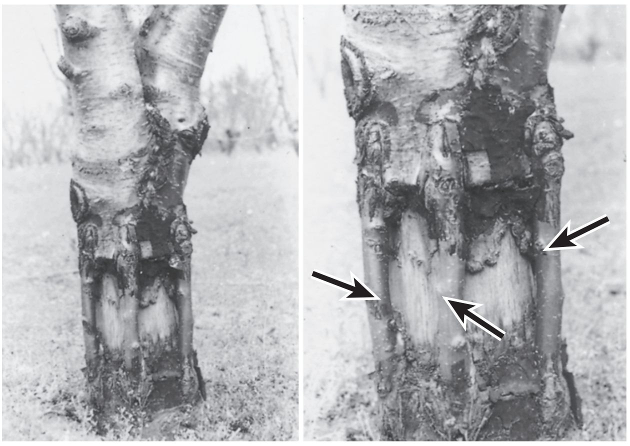

Figure 12–32
Injured trunk of a cherry tree successfully bridge grafted (arrows) by a modification of the bark graft.

PRODUCTION PROCESSES OF GRAFTAGE

Success in grafting depends 45 percent on preparation, including the quality and preparation of the scion and rootstock material, 10 percent on craftsmanship, and 45 percent on the aftercare of the grafted plant (26, 38). The production goals of grafting are achieving a high success rate, or “take,” and obtaining high speed and accuracy in performing the graft. Preparation for grafting begins with the proper tools and accessories, as well as the selection and handling of the scion and rootstock (Table 12–3). Since grafting is a repetitive, labor-intensive process, grafting machines and grafting automation, including robotics, continue to play a greater role.

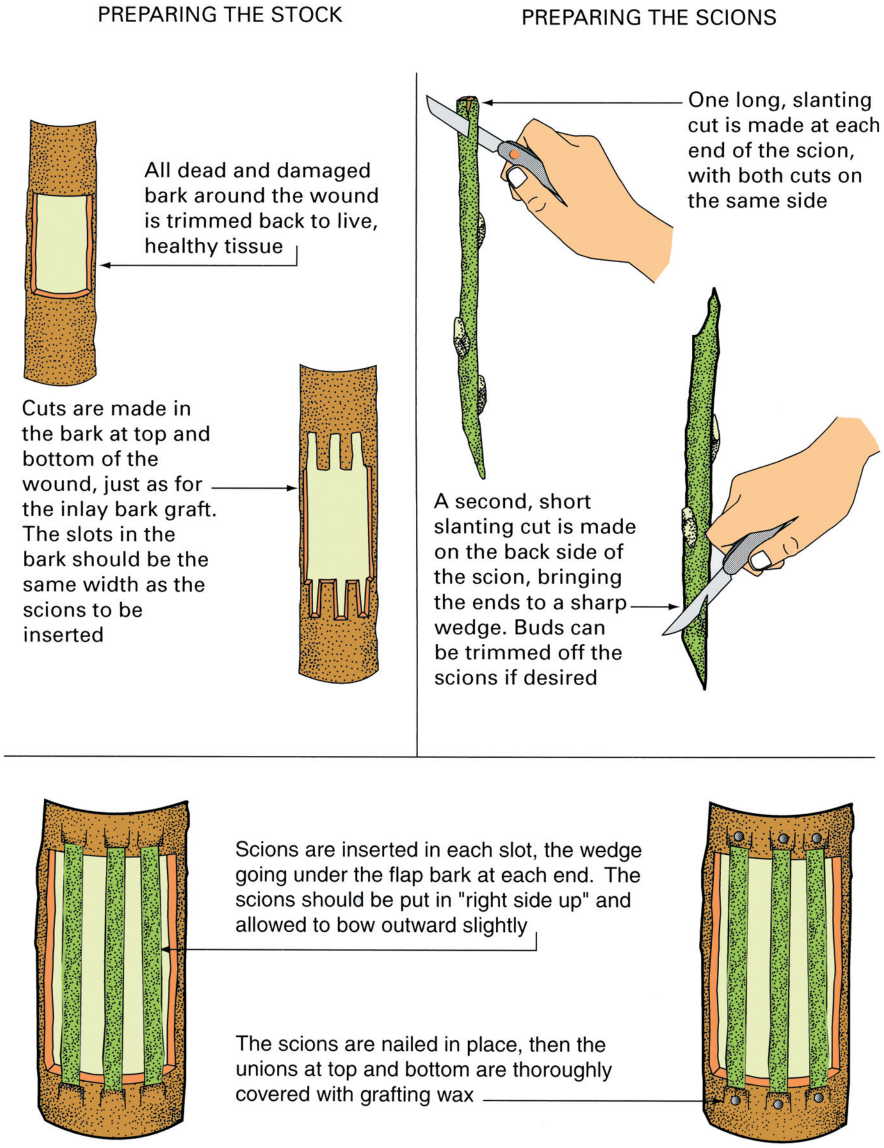
PREPARING THE STOCK
PREPARING THE SCIONS
All dead and damaged bark around the wound is trimmed back to live, healthy tissue
One long, slanting cut is made at each end of the scion, with both cuts on the same side
Cuts are made in the bark at top and bottom of the wound, just as for the inlay bark graft. The slots in the bark should be the same width as the scions to be inserted
A second, short slanting cut is made on the back side of the scion, bringing the ends to a sharp wedge. Buds can be trimmed off the scions if desired
Scions are inserted in each slot, the wedge going under the flap bark at each end. The scions should be put in "right side up" and allowed to bow outward slightly
The scions are nailed in place, then the unions at top and bottom are thoroughly covered with grafting wax

Figure 12–33
A satisfactory method of making a bridge graft, using a modification of the inlay bark graft.

BOX 12.2 **GETTING MORE IN DEPTH ON THE SUBJECT**

LARGE-SCALE BRIDGE GRAFTING

After World War I, thousands of fruit trees were bridge grafted in France to repair damage that occurred during the war. Alternatively, trunks from mutilated trees less than 20 cm (8 in) in diameter were cut off and crown grafted (23).

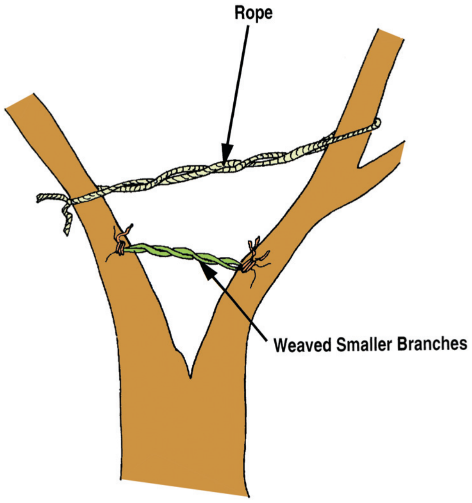

Figure 12–34
Bracing of fruit tree limbs by encouraging natural grafting. The tree limbs are braced with a twisted rope or electrical cord, and smaller shoots from the limbs are woven together and graft naturally as secondary growth occurs.

Preparation for Grafting

Tools and Accessories for Grafting Common tools and accessories used for grafting include grafting knives, tying and wrapping materials, and grafting waxes. Special equipment needed for any particular method of grafting is illustrated, along with the description of the method, in the remaining pages of this chapter. For example, grafting planes are sometimes used for more accurate fitting of scions with a hard and thick wood (Fig. 12–35) (61).

Table 12–3
PRODUCTION PROCESSES OF GRAFTAGE

Preparation for Graftage
- Tools and accessories for grafting
- Grafting machines and grafting automation/robotics
- Selection, handling, and storage of scionwood
- Handling of rootstock.

The Craftsmanship of Grafting
- Manual techniques: speed, accuracy, enhancing success rates
- Record keeping

Aftercare of Grafted Plants
- In bench grafting systems
- In field and nursery grafting systems

Knives. The two general types of knives used for propagation work are the budding knife and the grafting knife (Fig. 12–36). Where a limited amount of either budding or grafting is done, the budding knife can be used satisfactorily for both operations. The knives have either a folding or a fixed blade. The fixed-blade type is stronger, and if a holder of some kind is used to protect the cutting edge, it is probably the most desirable. A well-built, sturdy knife of high-carbon steel is essential. Grafting blades are flat on one side and have a tapered edge on the other to make a sharp, clean cut. Grafting knives are available for either right- or left-handed people.

Tying and Wrapping Materials. Grafting methods, such as the whip-and-tongue, splice (whip), and side-veneer graft, and budding methods, such as chip budding and T-budding, require that the graft union be held together by tying until the parts unite. A number of materials can be used for tying or wrapping grafts and budding. Some of these tying materials can also seal and help maintain a high relative humidity in the graft union area, which can help eliminate production

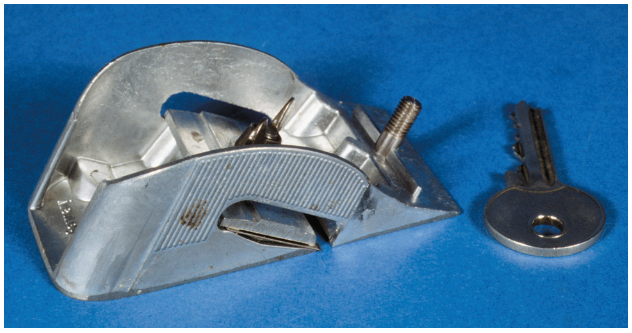

Figure 12–35
The grafting plane is a small woodworking plane equipped with a disposable heavy-duty razor blade. It allows more precise fitting of scions with hard and thick wood (61).

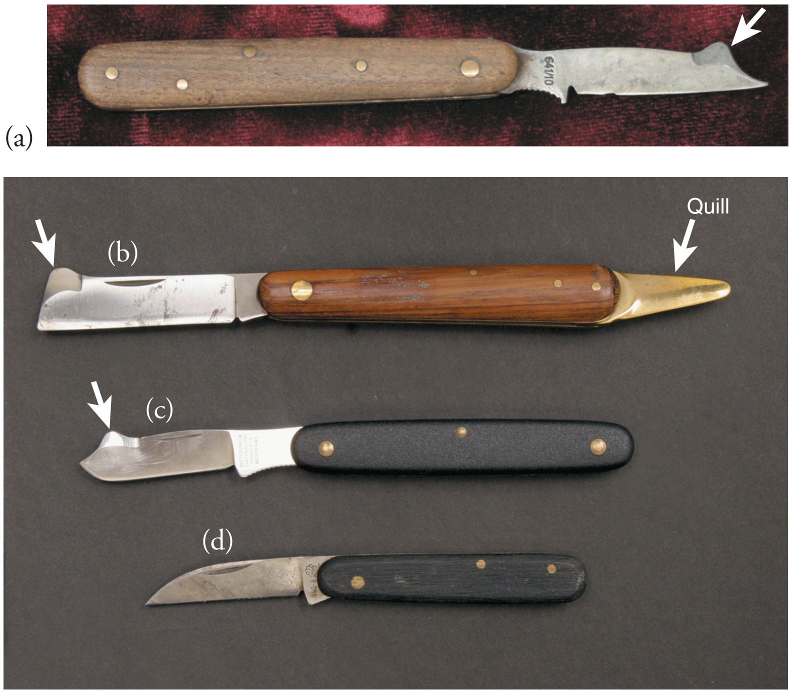

Figure 12–36
Folding, budding-grafting knives. (a) Tina budding-grafting knife. (a, b, and c) The metal flap (arrow) opposite the sharp end of the blade is used to separate the bark during the "T"–cut of the rootstock. (b) The quill is also used to separate the bark during T-budding. (d) Grafting knives are sharpened on one side, so there is a flat side of the blade for better control.

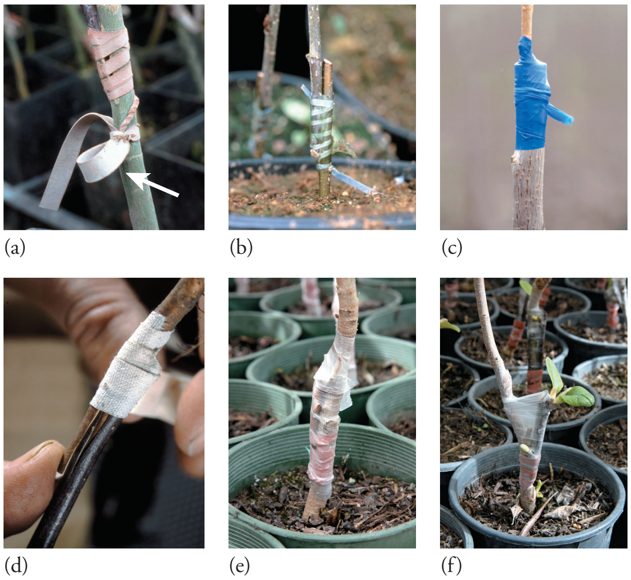

Figure 12–37
(a) Sometimes a loop is included in the final half-hitch knot of the rubber (arrow) or polyethylene budding and grafting strips. This allows easy unraveling of the strips after the graft has "taken" and avoids potential girdling problems in this side veneer graft of Japanese maple. Various tying materials: (b) raffia, (c) poly tape, (d) nursery tape, and (e and f) Buddy Tape (similar to Parafilm) used to cover magnolia scion pieces.

steps for applying a hot wax sealant on top of the tying materials, or the need for maintaining the grafts in special poly chambers or "sweat boxes."

Some common tying materials for budding and grafting include:

- **Budding rubbers** (also used in grafting).
- Clear or colored polyethylene or polyvinyl chloride **(PVC) budding and grafting strips,** which are 0.5 to 1.3 cm (3/16 to 1/2 in) wide and slightly elastic, allowing for a more secure wrap; this is also called flagging tape, green floral tape, white budding tape, or orange grafting tape.
- **Plastic clips and silicon tubing** (for manual and robotic grafting).
- **Raffia** (strips of palm leafstalk fiber—an older wrapping material, but still used).

Since PVC budding and grafting strips are not self-adhesive, they must be tied with a half-hitch knot (Fig. 12–37), which is done at the final turn of the tape by slipping it under the previous turn. With the exception of budding rubber, which deteriorates in full sunlight (but not when buried with the graft below ground), or Buddy Tape, the wrapping materials must be removed later to prevent girdling the plant.

Self-Adhering Tying Materials. A time-tested tying material is **waxed string** or **twine,** which adheres to itself and to the plant parts without tying. It should be strong enough to hold the grafted parts together, yet weak enough to be broken by hand.

Nursery adhesive tape is similar to surgical adhesive tape but lighter in weight and not sterilized; it is more convenient to use than waxed cloth tape. Adhesive tape is useful for tying and sealing whip grafts. When using any kind of tape or string for wrapping grafts, it is important not to use too many layers or the material may eventually girdle the plant unless it is cut. When this type of wrapping is covered with soil, it usually rots and breaks before damage can occur. On a limited scale, adhesive tapes, such as **duct** and **electrical tape** can be used, while **masking tape** tends to unravel (54). Regardless of the wrapping material, it is best to remove or cut it *after* the graft has taken to avoid girdling.

Self-sealing tying materials include **Parafilm tape,** which has been used with successful results to wrap graft unions rapidly (5) and for chip-budding roses (Fig. 13–6). This material is a waterproof, flexible, stretchable, thermoplastic film with a paper backing. The film is removed from the paper, wrapped around the graft union, and pressed into place by hand. **Buddy Tape** (buddytape.com) is similar to Parafilm tape, but thinner and more economical. It seals and holds the graft or bud piece in place, and is thin enough for the bud to elongate and pass through it once the graft "takes" (Fig. 12–37). Sometimes budding rubbers are used to tie a graft, which is then sealed with Parafilm tape. Self-sealing cure **crepe rubber sheets** are used for herbaceous grafts and small woody plant material. Rubber patches up to 4 cm (1 1/2 in) are fixed with a staple and used for budding (Fig. 12–38).

Miscellaneous Fastenings and Wrapping Material. Miscellaneous fastenings and wrapping material include 18-gauge, 2-cm (3/4-in) **nails,** 1.6-cm (5/8-in) **flat-point staples,** as used in the inlay bark graft of pecans (40), and **plastic graft clips,** used in manual and machine splice grafting of vegetable crops (Figs. 12–7 and 12–39). The combination of **aluminum foil** and **polyethylene bags** wrapped around a four-flap or inlay bark graft replaces the need for waxing the graft (which would melt and be unsuitable in spring field grafting in Texas or other warm regions) (Figs. 12–13 and 12–24). **Metal shoot guide clips** (Fig. 13–4) are used for field budded, dormant rootstock to compel upright growth from the bud. **Silicon tubing** has been used to hold graft unions of single-node scions of oak and ash with high rooting success (Fig. 12–40, page 496) (14).

Splints made of *toothpicks, bamboo,* or *metal skewers* are used with bench grafting of herbaceous plants such as cacti. The splints are later removed after the graft has taken.

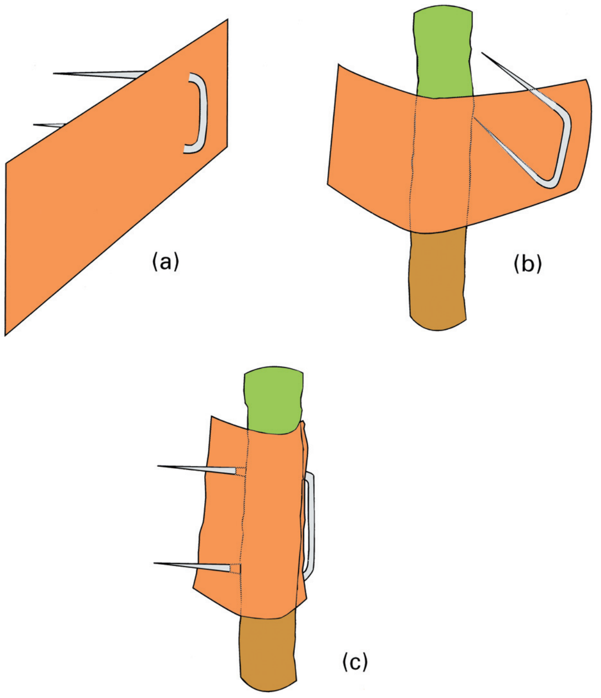

Figure 12–38
Rubber patch tying materials: (a) Rubber patches that come with a fixed stainless steel staple. (b) The rubber sheet is pulled over the bud and pulled securely around the rootstock. (c) The staple is pushed through the flap to secure the patch. The rubber patch normally deteriorates later with UV light from the sun or can easily be removed.

In a novel approach for developing robotic grafting systems for vegetable crops in Japan, Chinese cabbage seedling (scions) are horizontally grafted to turnip rootstocks. In grafting *Solanaceae* and *Cucurbitacae* vegetable crops, the graft partners are joined by a chemical adhesive, followed by spraying a chemical hardener to set up and solidify the adhesive around the graft (see Fig. 12–43, page 498 (29, 34, 47, 48)).

Whether a sealant, such as grafting wax, may be applied depends on the type of graft, the grafting system, and type of material used. Sealants are generally not used with budding, since tying with budding rubbers, rubber patches, Buddy Tape, or Parafilm tape is sufficient to alleviate desiccation problems. If the bench-grafted plant is to be placed in a high relative humidity graft box or temperature-controlled polytent, or immediately outplanted in the field with the union below the soil surface, waxing may be omitted.

Grafting Waxes. Grafting wax has two chief purposes: (a) It seals over the graft union, thereby preventing the

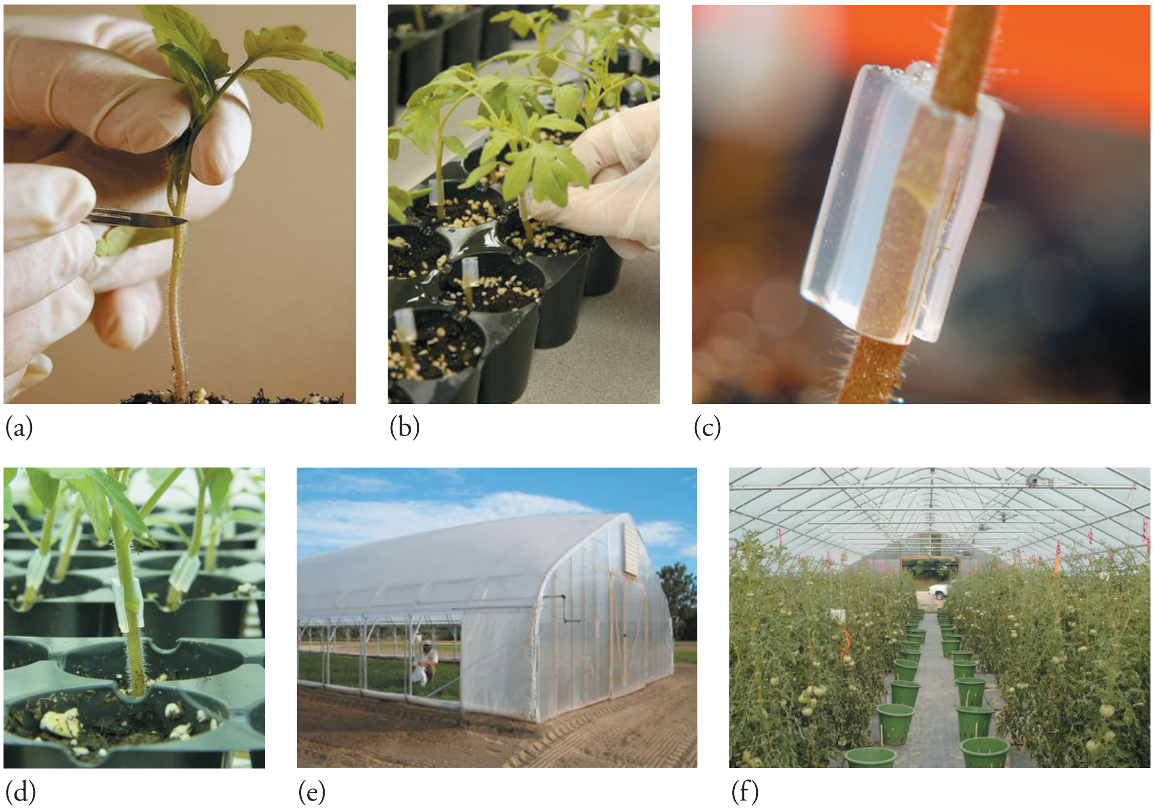

Figure 12–39
Grafting of heirloom tomatoes in high tunnels can lead to increased yield, earlier season extension (compared to field-grown), soil-borne disease resistance and increased nutrient uptake with a potential reduction in fertilizer inputs. (a, b, c, and d) OCG or Japanese tubegrafting system. (c and d) Plastic clip holding graft. (e and f) High tunnel production of grafted heirloom tomatoes. The system enhances IPM—integrated pest management—and organic production systems for vegetable crops (52).
Photos courtesy S. O'Connell, M. Peet, C. Rivard, F. Louws, and S. McDonough.

loss of moisture and death of the tender, exposed cells of the cut surfaces of the scion and rootstock. These cells are essential for callus production and healing of the graft union. (b) It prevents the entrance of various decay-producing organisms that rot wood.

An ideal grafting wax should adhere well to the plant surfaces, not be washed off by rains, not be so brittle as to crack and chip during cold weather or so soft that it will melt and run off during hot days, but still be pliable enough to allow for the swelling of the scion and the growth enlargement of the rootstock without cracking. **Hot waxes** require heating, while **cold waxes** contain volatile solvents that keep the wax liquid. The cold wax solidifies when the solvents evaporate. Most nurseries develop their own hot wax, which is low-melting, soft, and flexible, so that subsequent handling of the graft does not cause cracking and flaking. Thermostatically controlled wax heaters are available to provide instant liquid wax when needed. *The wax should be hot enough to flow easily, yet not be boiling, which damages plant tissue.*

hot waxes Waxes that are paraffin based and must be heated to melt and apply.

cold waxes Waxes that contain a volatile solvent that preserves the wax liquid. After application the solvents evaporate and the wax hardens.

Various recipes for making hot and cold waxes are listed by Garner (19). For hot wax, blocks of premixed grafting wax containing the necessary ingredients (e.g., TrowBridge's grafting wax, Walter E. Clark & Son, Orange, Conn., USA) can be purchased from nursery supply houses.

Grafting Machines Several *bench grafting* machines or devices have been developed to prepare graft and bud unions, and a few have been widely used, especially in propagating grapevines (2, 3).

Various bench grafting machines for the wedge graft or French-V are available, including a portable and

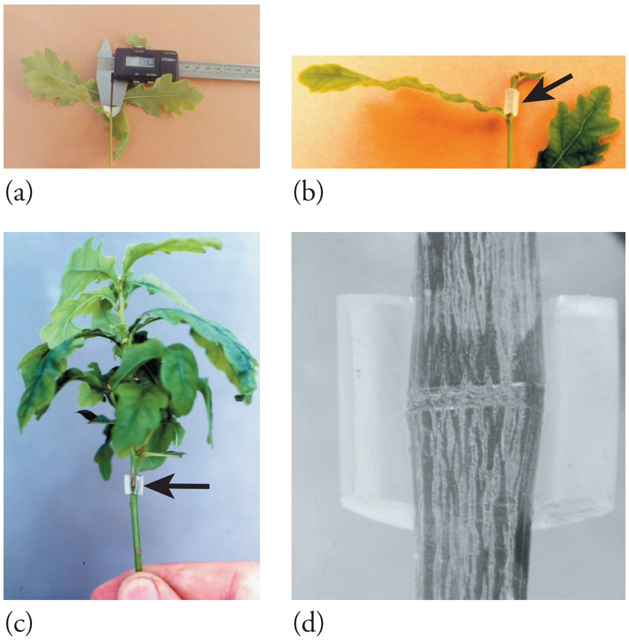

Figure 12–40
High grafting success has been obtained using silicone tubing to hold graft unions together of single-node scions of oak and ash: (a) Digital caliper to measure stem diameter. (b) Single node of oak grafted using silicone tube (arrow). (c) Oak graft after 8 weeks (arrow showing tubing). (d) Healed ash graft after 12 weeks. Photographs courtesy of G. Douglas (14).

bench-mountable device made in New Zealand by Raggett Industries, Ltd., Gisborne (www.raggettindustries.co.nz) (Fig. 12–41). This device makes a type of wedge graft, cutting out a long V-notch in the rootstock and a corresponding long, tapered cut at the base of the scion. By reversing the position of the rootstock and scion, it could also make a saddle graft (Fig. 12–11). Although the cuts fit together very well, the operation is slow because the graft union must be either tied with a budding rubber or poly tape or stapled together. This machine has been used successfully in propagating grapes and fruit trees.

There are machines for making omega grafts, which are hand-operated (Fig. 12–42) or foot-operated. One device for grape grafting is the Pfropf-Star grafting machine manufactured in Germany. It cuts through both the rootstock and scion, one laid on top of the other, making an omega-shaped cut and leaving the two parts interlocked. While these machines work fine for grape grafting, most ornamental nurseries that graft do not use machines. Instead, they rely on hand-grafting, which may be faster and more reliable given the larger number of genera and species grafted (35). Without question, finding skilled grafters is a severe production problem; hence the importance of developing mechanized and automated grafting systems.

Grafting Automation/Robotics Prototypes and commercial robotic machines for grafting vegetable seedlings have been developed (Figs. 12–43, 12–44, and 12–45, pages 498–99). The production of grafted vegetable crops is becoming more common in the United States. However, grafted vegetable seedlings are used extensively in heavily populated countries such as Japan, Korea, some other Asian countries, and in parts of Europe, where the land use is highly intensive, farming areas quite small, and crops are not rotated. Grafted seedlings account for 81 percent of the commercial outdoor and greenhouse vegetable production in Korea, and 54 percent and 81 percent, respectively, for Japan (12, 29, 30, 33). Vegetable rootstock used are resistant to soil-borne pathogens and nematodes, which build up under these intensive cultivation conditions. Some of the commercialized grafting robots can graft 800 or more *Solanaceae* vegetable seedlings per hour (Figs. 12–43 and 12–45) (49).

Selection, Handling, and Storage of Scionwood.

Kind of Wood. Since bench or field grafting of deciduous species takes place in winter or early spring, it is necessary to use the scionwood that grew the previous fall.

In selecting such scion material the following points should be observed:

- For most species, the wood should be 1 year old or less (current season's growth). Avoid including older growth, although with certain species, such as the fig or olive, 2-year-old wood is satisfactory, or even preferable, if it is of the proper size.
- Healthy, well-developed vegetative buds should be present. Avoid wood with flower buds. Usually, vegetative buds are narrow and pointed, whereas flower buds are round and plump (see Fig. 13–3).
- The best type of scion material is vigorous (but not overly succulent), well-matured, hardened shoots from the upper part of the tree, which have grown 60 to 90 cm (2 to 3 ft) the previous summer. Such growth develops on relatively young, well-grown, vigorous plants; high production of scion material can be promoted by pruning the plant back heavily the previous winter. Water sprouts from older trees sometimes make satisfactory scionwood, but suckers arising from the base of grafted trees should not be used, since they

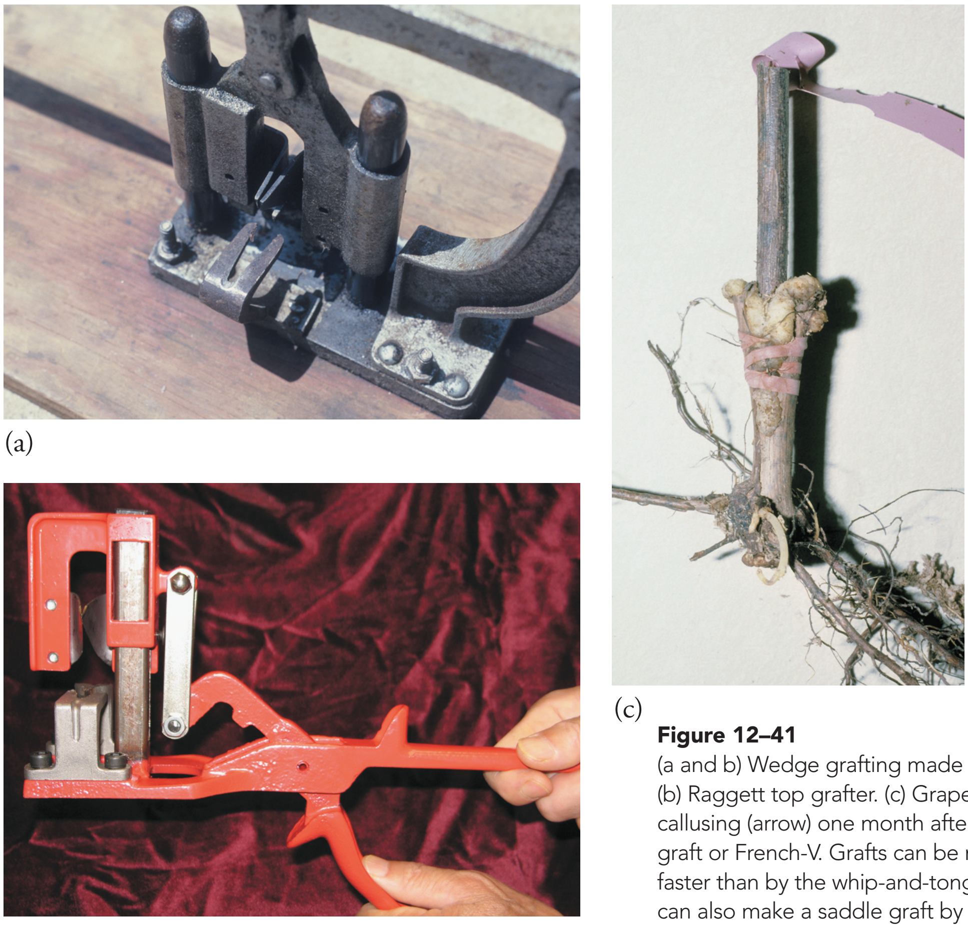

(a)

(b)

(c)

Figure 12–41
(a and b) Wedge grafting made by French-V grafting devices. (b) Raggett top grafter. (c) Grape graft union healing with profuse callusing (arrow) one month after bench grafting with a wedge graft or French-V. Grafts can be made with these devices much faster than by the whip-and-tongue graft method. These machines can also make a saddle graft by these machines, by reversing the cuts so the scion piece has the saddle shape (see Fig. 12–11).

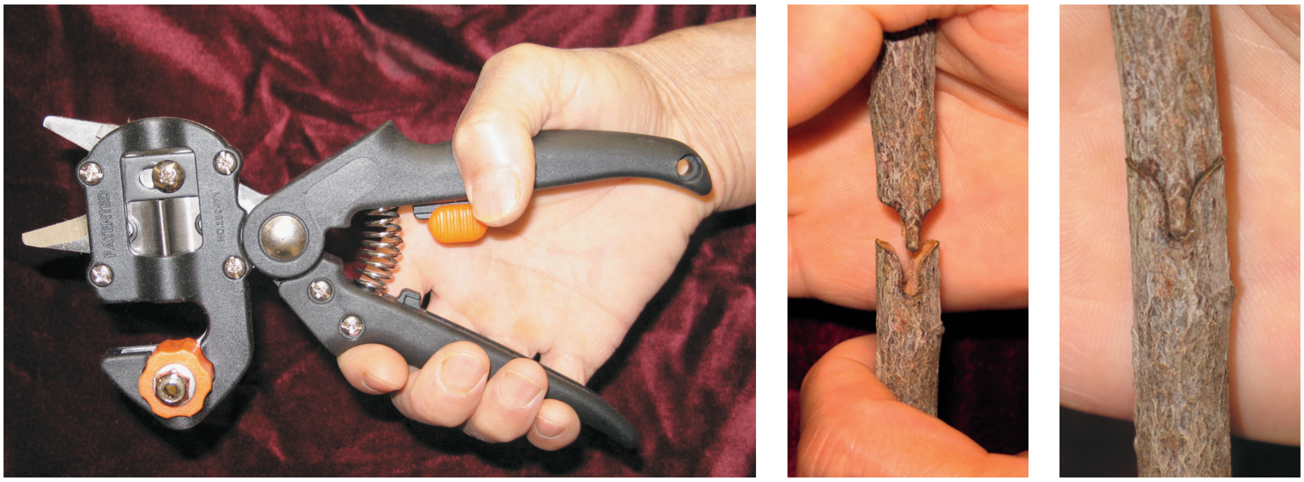

(a) (b) (c)

Figure 12–42
(a) Grafting tool for making an omega graft. (b and c) Omega graft locks in place and then is held together with grafting tape. The scion and stock must be the same diameter.

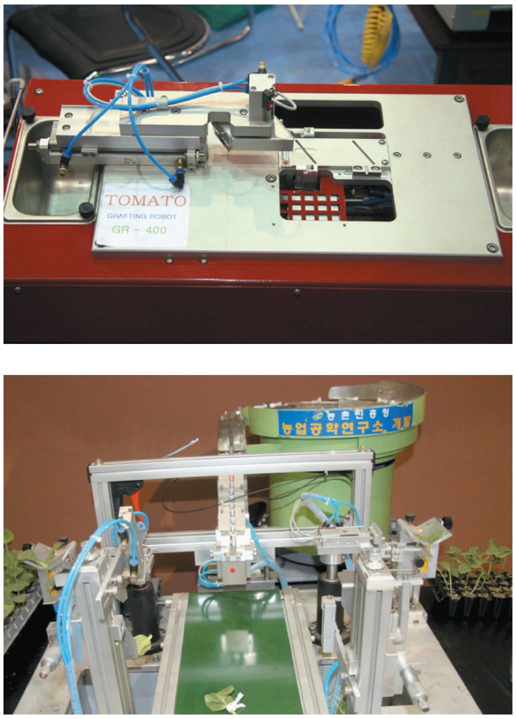

(a)

(b)

Figure 12–43
Commercial grafting robot for Solanaceae (tomatoes, melons, cucurbits) vegetables on seedling rootstock suitable for intensive planting and resistant to disease, insect and environmental stress. Plant vigor and yield can also be enhanced with superior rootstock. (a) Tomato and (b) melon grafting robots.

may consist of rootstock material. A satisfactory size is from 0.6 to 1.2 cm (1/4 to 1/2 in) in diameter.

- The best scions are obtained from the center portion or from the basal two-thirds of the shoots. The terminal sections, which are likely to be too succulent, pithy, and low in stored carbohydrates, should be discarded. Mature wood with short internodes should be selected.

Source of Material. Scionwood should be taken from source plants of the correct cultivar known to be pathogen-tested and genetically true-to-type (see Chapter 16). Virus-diseased, undesirable sports, chimeras, and virus-like genetic disorders must be avoided. Source plants may be of three basic types:

1. Plants produced in an orchard, vineyard, ornamental field, container nursery, or landscape are selected when the flowering, fruiting, and growth habits are known. It is best to take propagation material from bearing plants whose production history is known. Visual inspection, however, may not reveal the true condition of the proposed source plant and, appropriate indexing and progeny tests are required to be sure (see Chapter 16).
2. In commercial nurseries, special scion blocks, where plants are grown particularly for propagation, may be maintained. Such plants are handled differently than they would be for producing a crop. For example, fruit trees may be pruned back each year to produce a large annual supply of long, vigorous shoots well-suited for scionwood. Such special blocks would usually be handled to conform to registration and certification programs and would be subject to isolation, indexing, and inspection requirements. In addition, it is important to maintain source identity of scion material through the entire propagation sequence, so that over a period of time proper

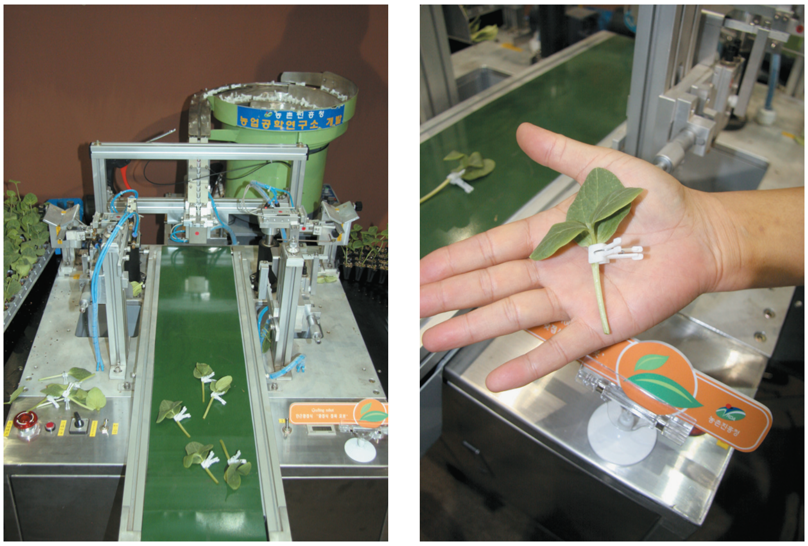

(a) (b)

Figure 12–44
Finished grafts from a commercial melon grafting robot. (a) The grafted plants are moved on a conveyor belt system for processing. (b) Splice grafted scion and rootstock held together by a grafting clip.

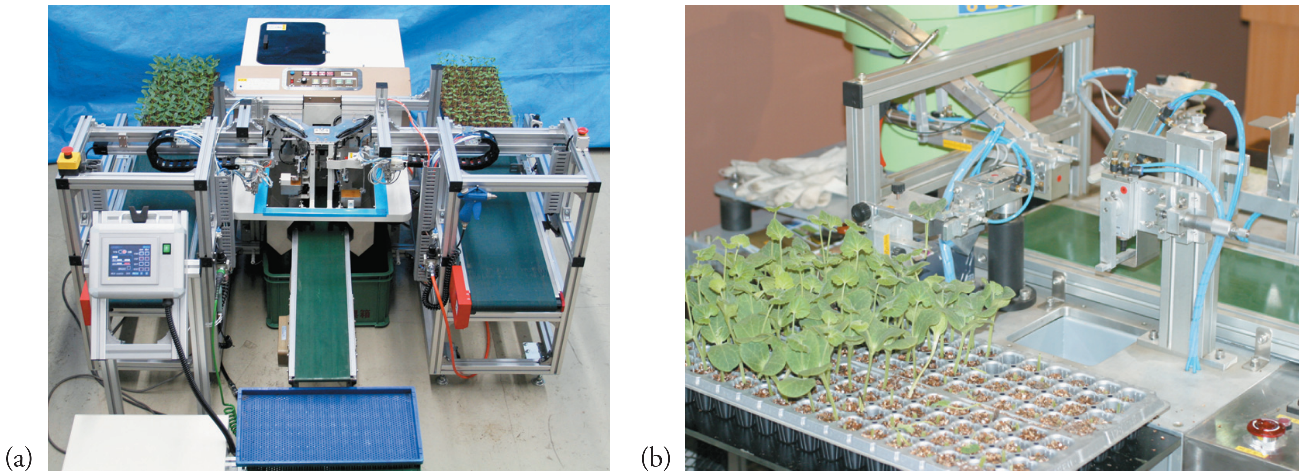

Figure 12–45
Two techniques used in herbaceous grafting are manual and robotic grafting. (a) Fully automated grafting robot for cucurbits (BRAIN, Saitama, Japan), and (b) semiautomated Korean grafting machine. Plants are grafted at the 2 to 4 leaf stage. The advent of OCG, "tube-grafting" or "Japanese top-grafting" has become the most popular graft for tomato. Grafting robots can make up to 800 grafts/hr, whereas an individual can make 1,000 grafts/day. Photos a and b courtesy of C. Kubota and M. Peet, respectively.

sources of the various cultivars can be identified and maintained.

3. For vegetable grafting, commercial seed from selected rootstocks and scions are sown under protected cultural conditions. Selected rootstock for vegetable grafting is listed in Table 12–4. A timeline for grafting heirloom tomatoes, starting with sowing rootstock seed 2 to 5 days prior to sowing scion seed, is depicted in Figure 12–46.

Collection and Handling. For deciduous plants to be grafted in early spring, the scionwood can be collected almost any time during the winter season when the plants are fully dormant (6). In climates with severe winters, the wood should not be gathered when it is frozen, and any wood that shows freezing injury should not be used. Where considerable winter injury is likely, it is best to collect dormant scionwood and put it in cold storage after leaf fall but before the onset of winter.

Storage. Scionwood collected prior to grafting must be properly stored. It should be kept slightly moist and at a low enough temperature to prevent elongation of the buds. A common method is to wrap the wood, in bundles of 25 to 100 sticks, in heavy, waterproof paper or in polyethylene sheets or bags. *All bundles must be labeled accurately.*

Polyethylene bags are useful for storing small quantities of scionwood. They allow the passage of oxygen and carbon dioxide, which are exchanged during

Table 12–4
SELECTED SCION AND ROOTSTOCK COMBINATIONS FOR VEGETABLE GRAFTING (50)

Scion	Rootstock
Watermelon (*Citrullus vulgaris* syn. *C. lanatus*)	Bottle gourd (*Lagenaria siceraria*) White gourd (*Benincasa hispida*) *Cucurbita* spp.
Melon (*Cucumis melo*)	White gourd (*Benincasa hispida* Cogn.) *Cucumis* spp.; *Cucurbita* spp. *C. moschata xC. maima*
Cucumber (*Cucumis sativus* L.)	Pumpkin (*Cucurbita* spp.) *Cucurbita ficifolia*
Aubergine (*Solanum melongena* L.)	*Solanum integrifolium;* *Solanum torvum; Solanum melongena*
Tomato (*Lycopersicon esculentum*)	Tomato (*Lycopersicon esculentum*)

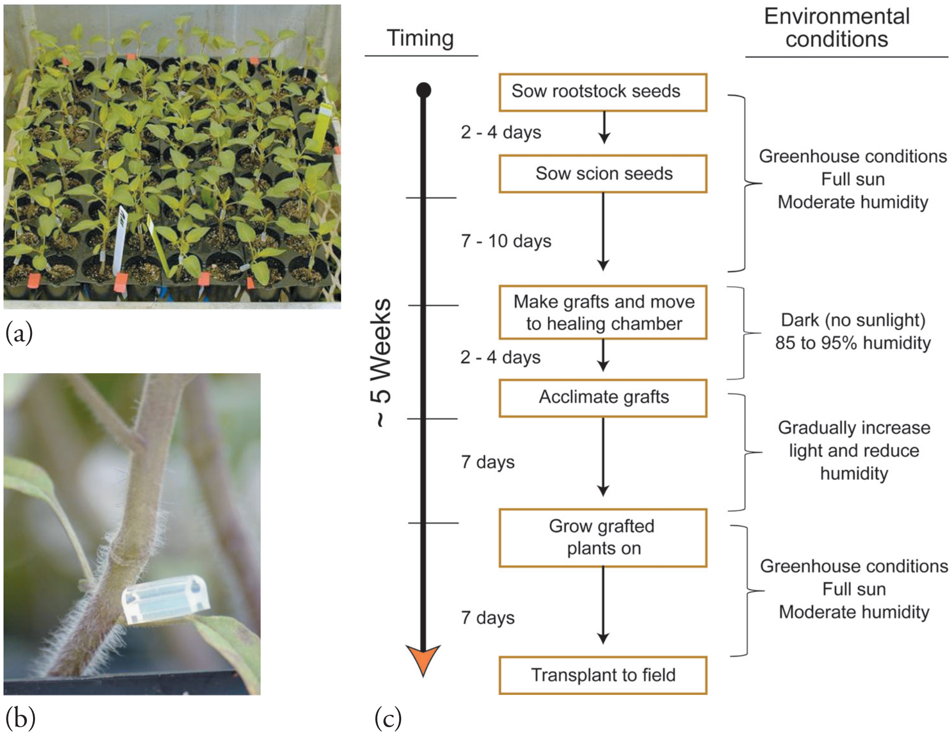
(a)
(b)
(c)
Timing
Environmental conditions
~ 5 Weeks
Sow rootstock seeds
2 - 4 days
Sow scion seeds
7 - 10 days
Make grafts and move to healing chamber
2 - 4 days
Acclimate grafts
7 days
Grow grafted plants on
7 days
Transplant to field
Greenhouse conditions
Full sun
Moderate humidity
Dark (no sunlight)
85 to 95% humidity
Gradually increase light and reduce humidity
Greenhouse conditions
Full sun
Moderate humidity

Figure 12–46
Grafting timeline for heirloom tomatoes. (a) Grafts in healing chamber. (b) Hardened-off grafted plant, with grafting clip falling off—ready for transplanting (52). Illustrations courtesy of C. Rivard and F. Louws.

the respiration process of the stored wood, but retard the passage of water vapor. Sometimes the natural moisture in the wood is sufficient, so slightly moist packing material is not needed in the sealed poly bags. In commercial field rose production, the scionwood (budwood) is harvested dormant, wrapped in slightly moist newspaper and sealed in poly bags, and maintained at –1.7 to –0.6°C (29 to 31°F) for up to 7 months.

The temperature at which the wood is stored is important. If it is to be kept only 2 or 3 weeks before grafting, the temperature of a home refrigerator—about 5°C (40°F)—is satisfactory. If stored for a period of 1 to 3 months, scionwood should be held at about 0°C (32°F) (6) to keep the buds dormant. However, buds of some species, such as the almond and sweet cherry, will start growth after about 3 months, even at such low temperatures. Do not store scionwood in a home freezer because the very low temperatures, about –18°C (0°F), can injure the buds.

Storage of scions should not be attempted if succulent, herbaceous plants are being grafted; such scions should be obtained at the time of grafting and used immediately. Certain broad-leaved evergreen species, such as camellias, olives, and citrus, can be grafted in the spring before much active growth starts, without previous collection and storage of the scionwood. Grafts are taken directly from the tree as needed, using the basal part of the shoots containing dormant, axillary buds. The leaves are removed at the time of collection.

Attempting to use scionwood in which the buds are starting active growth is almost certain to result in failure. In such cases, the buds quickly leaf out before the graft union has healed; consequently, the leaves withdraw water from the scions by transpiration, and cause the scions to die. In addition, the strong competing sink of a developing shoot can interfere with graft union formation.

In topworking pecans (4), good results are obtained by using precut scions; that is, scions cut in advance by skilled persons at a convenient time, that are then held in cold storage in polyethylene bags for up to 9 days before being inserted in the graft unions. Grafting success is reduced only slightly by the use of precut scions.

Handling of Rootstock for Bench Grafting. Seedling rootstock of maple (*Acer*) is established in liner pots for 1 year, brought into a greenhouse in the fall after leaf drop, and placed in bottom-heated benches at 13 to 16°C (55 to 60°F). Bench grafting of the container rootstock with a splice graft is done in January and February (Canada), when white roots appear along the perimeter of the rootballs (26).

In North Carolina, rootstock liners of woody ornamental plants are allowed to harden-off in minimum-temperature–controlled poly houses in the fall, and maintained just above freezing. When new roots emerge from the rootstock in late winter, plants are

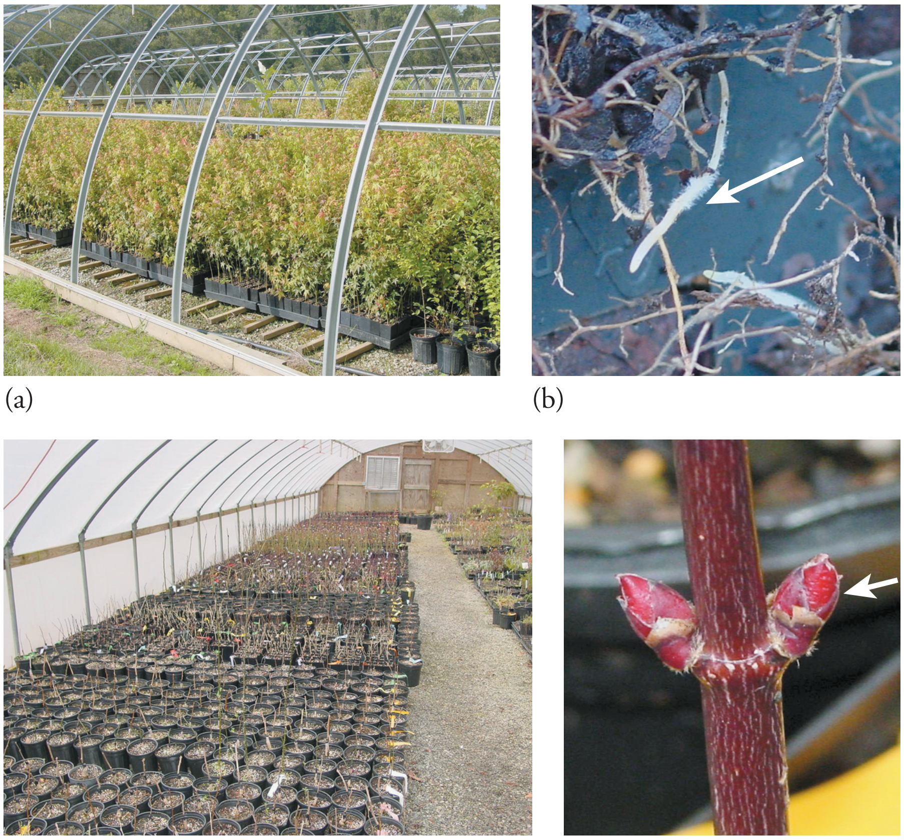

(a) (b) (c) (d)

Figure 12–47
Proper rootstock preparation. (a) Hardening-off Acer palmatum rootstock liners in late September (North Carolina). (b) New roots emerging from rootstock in January prior to grafting. (c) Proper after-care of grafted, labeled plants in poly covered, temperature-controlled hoop house. (d) Buds swelling on grafted A. palmatum 'Fireglow' in mid to late March (56).
Courtesy B. L. Upchurch.

ready for grafting (Fig. 12–47) (56). In general, bench grafting is best when new, white root tips of 6 mm (1/4 in) occur or buds start to swell on the rootstock of potted liner plants.

The Craftsmanship of Grafting

Grafting is both an art and a skill. Successful grafting is a **repetitive task** that requires a high degree of accuracy and speed; to become skilled, it is essential for the grafter to eliminate all unnecessary movements. To increase grafting efficiency, it is important to organize the workplace so that scion material, knives, and grafting tape are all within easy reach (42). Grafting is generally more efficient with a **team approach,** in which each worker performs a certain task, in order to reduce inefficient motion of materials and repetitive picking up and putting down of different tools.

Manual Techniques: Speed, Accuracy, and Enhancing Success Rates Tips on improving grafting techniques and ergonomics include the following (42):

- Concentrate on accuracy first, and allow grafting speed to build up—aim at initially completing at least 200 bench grafts a day.
- Use a graft method that is less time-consuming (yet still successful!) and that can be done with lesser skill and preparation of the rootstock [e.g., bench grafting with a splice (whip) graft, compared to a whip-and-tongue or side-veneer graft (43)]; this works well with *Betula, Cornus, Fagus, Ginkgo, Quercus,* and *Acer* (26).
- Grafting is best with two people: in bench grafting with the whip-and-tongue, one person does the graft, and the other moves the potted liner rootstocks and waxes the graft union. In T-budding field roses, the budder prepares the rootstock and inserts the shield bud of the scion, while the "tier" follows and ties the budding rubber around the budded graft (see Fig. 13–13).
- The grafting knife should always be held with a relaxed grip to improve accuracy and reduce repetitive strain injuries (e.g., **carpal tunnel syndrome**); it is necessary to restrict and control your arm movements.
- There are two basic cuts in grafting: the **slice cut,** which is made using the arm and shoulder to pull the knife [e.g., in the making of a splice (**whip**) graft]; and the **cross cut,** in which the grafter's arm and knife are rotated using the thumb

carpal tunnel syndrome
Nerve damage in the wrist caused by the stress of repetitive hand-arm movements.

as a pivot, with both hands joined to prevent the knife from cutting the grafter—a whip-and-tongue is created when the scion and rootstock are sliced and then cut across in **four** (economical) movements.

- Hang the grafting tape around your neck so that you know where it is and it stays free of contamination.
- In most cases, speed is more important than 100 percent accuracy; increase speed by developing a routine when grafting, and avoiding useless movements.
- Try to hold the grafting knife in your hand at all times (e.g., in your little finger).
- Make sure that everything is at a height that is easy to reach.

Record Keeping Maintaining good records is important for successful grafting.

- *Keep records on the grafters* to determine daily quantities grafted; some operations pay on a **piecework** or bonus system (see Chapter 10). Records kept by the supervisor can be constructively used to help grafters improve their technique and efficiency (9).
- *Keep records on the plant material* to determine the optimum windows of time to graft (43) and to assure having the best available material to graft. Records should be kept on the conditions of the grafting material, grafting problems encountered, and aftercare of the grafts. Should a crop failure occur, records can generally help pinpoint the cause and help managers improve efficiency (9).
- *Importance of developing and sticking to a grafting time-line schedule.* A timeline schedule is critical for commercial success (Fig. 12–46).

piecework A bonus system awarded to workers when they exceed daily graft production quotas.

AFTERCARE OF GRAFTED PLANTS

In Bench Grafting Systems

A common method of bench grafting is to wrap the union with budding rubbers, poly or plastic tape, plastic clips and silicone tubing, biodegradable cloth tape, or older materials such as raffia. Depending on the wrapping material, the entire union may be covered with grafting wax.

Root Grafting The root grafts may be placed under refrigeration at 7°C (45°F) for about 2 months. For general callusing purposes, temperatures from 7 to 21°C (45 to 70°F) are the most satisfactory. The callusing period for apples can be shortened to around 30 days if the grafts are stored at a temperature of about 21°C (70°F) and at a high humidity. To use this higher callusing temperature, the material should be collected in the fall and the grafts made before any cold weather has overcome the rest period of the scion buds. After the unions are well healed, the grafts must be stored at cool temperatures—2 to 4°C (35 to 40°F)—to overcome the "rest period" of the buds and to hold them dormant until planting (24). The root grafts are lined-out in early spring in the nursery row directly from the low-temperature storage conditions.

Hot-Pipe Callusing System With some plants, the graft union should be kept warm, 24 to 27°C (75 to 82°F), but the roots and the buds on the scion should be kept cool, about 5°C (41°F), to prevent premature growth before the graft union has callused and healed together. An ingenious system for regulating temperature was developed for whip-grafting hazelnut (*Corylus*), which are notoriously difficult to root graft (Fig. 12–48). This hot-pipe callusing system keeps the graft union warm by recirculating hot water in a PVC pipe onto which the graft is placed. The scion and roots protrude into areas of lower temperatures to keep the plant dormant until ready for transplanting. This hot-pipe callusing system, when used outdoors in late winter or early spring, has increased the root grafting "takes" of hazelnut and other difficult-to-graft species (31). Some aeration of the callusing grafts is required, so airtight containers should not

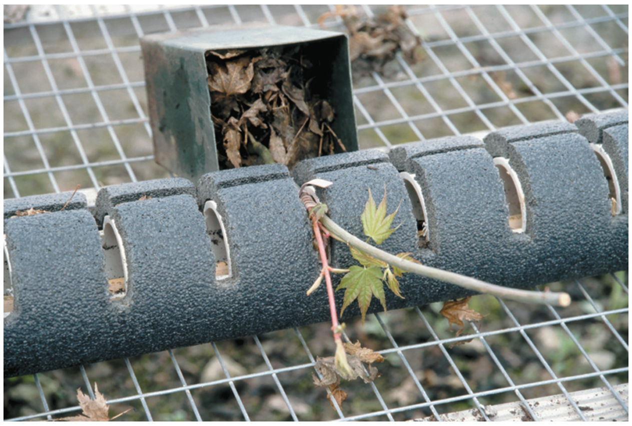

Figure 12–48
Hot-pipe callusing system for bench grafting difficult plants. The graft union is placed in a slot in a large plastic pipe. Inside the large pipe is a smaller pipe through which thermostatically controlled hot water circulates. Insulating material laid over this pipe retains the heat. The protected roots and scions protrude into areas of cooler temperatures, which retards their development.

be used. Virtually any graft used in bench grafting can be callused with the hot-pipe system, including apples, pears, peaches, and plums (32), and ornamentals such as *Acer, Cedrus, Corylus,* and *Fagus.* Not all species respond—the higher graft union temperature does not enhance graftage of spruce (*Picea*), for instance.

Closed Case Waxing may be omitted if the **bare-root grafts** are to be protected from drying by packing the grafts in boxes containing slightly moist peat. Some producers still dip the grafts in rose wax from the scion end to the taped union of the graft prior to boxing (51). Then the boxes are moved to a callusing room at 21°C (70°F) for about 12 days. Once the grafts have formed sufficient callus, the boxed grafts are held in cold storage at 2°C (35°F) until outplanted in the field (51). Another form of the closed case is the use of a **polytent** in a heated greenhouse for callusing the grafts of **potted rootstock liners** (Fig. 12–4). Provided light irradiance is controlled, glass mason jars can be used as a closed-case system for grafted plants in containers (Fig. 12–49).

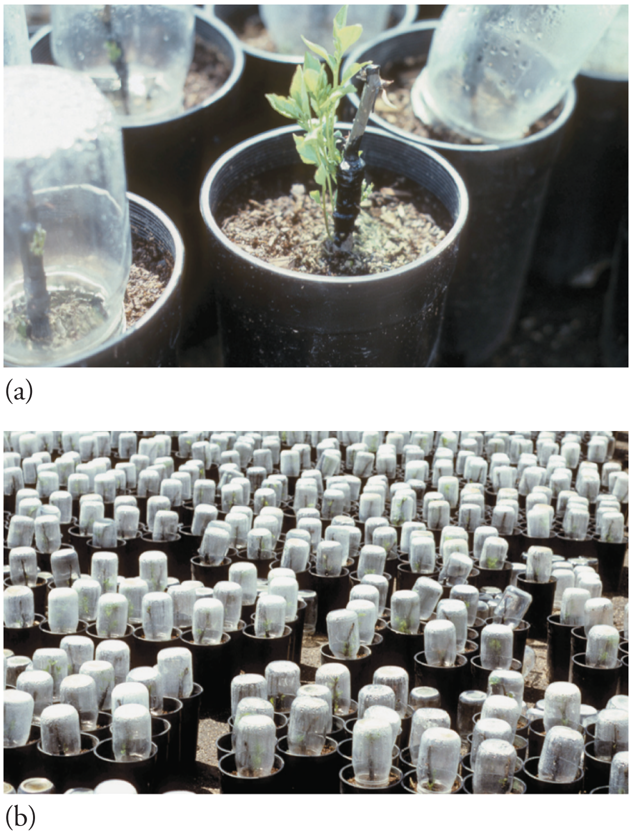

Figure 12–49
(a and b) A closed case system of covering containerized grafted plants with glass Mason jars. (a) Uncovered plant. Condensate is visible inside the glass jars. (b) These grafted plants are maintained under shade to reduce the light irradiance and minimize heat buildup.

Open Case (Open Bench) Grafting is also done in a temperature-controlled greenhouse or unheated polyhouse (depending on the season). Waxing can be omitted in the bench grafting of **potted rootstock liners** by plunging the container and burying the graft in slightly moist peat moss or bark in a temperature-controlled greenhouse. The medium is bottom heated and kept at 18 to 21°C (65 to 70°F) for 3 to 6 weeks for callusing; ideally, the air temperature should be cooler to discourage any initial top growth. Wrapping a graft with poly grafting tape is also sufficient without waxing. Optimum periods for grafting selected ornamental species in a greenhouse are listed in Table 12–5.

Outplanting of Bare-Root Grafts. As soon as the ground can be prepared in the spring, the grafts are lined-out in the nursery row 10 to 15 cm (4 to 6 in) apart. Grafts should be planted before growth of the buds or roots begins. If growth starts before the grafts can be planted, they should be moved to lower temperatures (−1 to 2°C, 30 to 35°F). The grafts are usually planted deep enough so that the graft union is just below the ground level, but if the roots are to arise only from the rootstock, the graft should be planted with the union well above the soil level [i.e., 7 to 15 cm (3 to 6 in)]. It is very important to prevent *scion rooting* where certain definite influences, such as dwarfing or disease resistance, are expected from the rootstock.

After one summer's growth, the grafts should be large enough to transplant to their permanent location. If not, the scion may be cut back to one or two buds, or headed-back somewhat to force out scaffold branches and allowed to grow a second year. With the older root system—a strong, vigorous top is obtained the second year.

Aftercare in Field and Nursery Grafting Systems

Aftercare of field- and nursery-grafted plants is described in "Production Processes of Graftage" starting on page 492. In the section "Types of Grafts" (page 466), see the descriptions for grafting and aftercare using such grafts as the whip-and-tongue, four-flap, and inlay bark graft. Chapter 11 describes heading-back (**lopping**) as well as the **crippling** techniques that are used in field and container nurseries to encourage the

crippling or lopping
Bending (restriction) or cutting halfway through the rootstock stem above the bud union to help force out the bud and maintain growth of the grafted plant. The rootstock tops are later cut off.

Table 12–5
"Optimum Windows" of the Year When Selected Ornamental Species Can Be Grafted in the Greenhouse in Oregon, USA (43)

	Month											
Crop	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Abies	X	X										X
Acer palmatum	X	X	X	X	X	X	X	X	X	X		
Aesculus	X	X	X				X	X	X			X
Carpinus	X	X	X	X						X	X	
Cedrus	X	X	X						X			
Cercis		X	X	X			X	X				
Cornus	X	X	X						X	X		
Fagus	X	X	X	X					X	X	X	
Ginkgo	X	X	X	X				X	X			
Hamamelis	X	X	X	X				X	X	X		
Larix	X	X										
Liquidambar		X	X					X	X			
Liriodendron		X	X						X	X		
Picea	X	X					X	X				X
Pinus spp.	X	X									X	X
Wisteria	X	X	X	X								

scions of grafted and budded plants to overcome the apical dominance of the rootstock and begin final production development.

Aftercare in Topworking Systems (Top-Grafting and Top-Budding) After the actual top-grafting (or top-budding) operation is finished, much important work needs to be done before the topworking is successfully completed. A good grafting job can be ruined by improper care of the grafted trees.

The trees should be carefully inspected 3 to 5 days after grafting, and the graft unions rewaxed if cracks or holes appear in the wax. However, using wax in top-grafting pecan trees is not feasible, given the high-temperature conditions of Texas, so inlay bark grafts are covered with aluminum foil and polybags to control desiccation and heat stress during the grafting process (Figs. 12–13 and 12–24). In Israel the grafts of topworked peaches have aluminum covers to reduce heat buildup (Fig. 12–24).

FIELD, BENCH, AND MISCELLANEOUS GRAFTING SYSTEMS

Some of the different grafting systems have been described in the sections on **types of grafting** and the **production processes of graftage. Grafting systems** are categorized as field, bench, and miscellaneous grafting systems (Table 12–6).

Field Grafting Systems

Crown Grafting The **crown graft** originally referred to scions grafted onto larger rootstock. The large rootstock stem was grafted with a number of scions, which sometimes were in a crown-like circle (19). Today, the term includes grafting onto an established rootstock with single or multiple scions, using the whip-and-tongue, cleft, wedge, side-veneer, inlay bark graft, and others. The choice of the graft depends on the species and size of the rootstock. In California, seedling walnut trees are planted in the nursery and then grafted at the crown—close to the junction of the root and shoot—of the rootstock.

crown grafting Grafting that is done at the **crown** of the rootstock, which is the junction of the root and shoot system. In earlier times it referred to grafting several scions in a crown-like circle onto an established larger rootstock.

Crown grafting of deciduous plants is done from late winter to late spring. In each species, grafting should take place shortly before new growth starts. The scions should be prepared from mature, dormant wood of the previous season's growth.

If the graft is above the soil level, the union must be well tied (or nailed) and sealed to firmly hold the graft and prevent desiccation. However, when the operation is performed just below, at, or just above the soil

Table 12–6
FIELD, BENCH, AND MISCELLANEOUS GRAFTING SYSTEMS

Field Grafting Systems
- Crown graftage
- Topworking
- Frame working
- Repair graftage—see page 487
- Double-working (by grafting and budding)—sometimes bench grafted

Bench Grafting Systems—see page 502
- Root graftage
- Nurse-root graftage
- Grafting of plants in liner pots under protected culture
- Herbaceous grafting—including grafting of vegetable crops
- Approach graftage—sometimes field grafted

Miscellaneous Grafting Systems
- Cutting grafts—simultaneous rooting and grafting (stenting) of roses, citrus, etc.
- Micrografting

level, it is possible to cover the graft union, or even the entire scion, with soil and thus eliminate the necessity for waxing or ceiling. In all cases, the union should be tied securely with tape to hold the grafted parts together until the graft takes.

Topworking **Topworking** is used primarily to change the cultivar of an established plant—tree, shrub, or vine—by grafting (see Fig. 12–24). Topworking can be done with any of the apical or side graftage grafts described earlier in this chapter, depending on the plant species. This procedure may be preferred to removal and replacement of the entire plant, since a return to flowering and fruiting is faster with topworking an established plant than with transplanting a new nursery plant—particularly if the topworked plant is young, healthy, and well cared for. Plants that are old, diseased, or of a short-lived species are not satisfactory candidates for topworking.

topworking Using either top-grafting or top-budding when a scion is grafted or budded onto an established plant in the orchard.

Preparation for Topworking. Top-grafting is usually done in the spring, shortly before new growth starts. The exact time depends on the method to be used. The cleft, side, whip, and wedge grafts can be done before the bark is slipping, but the bark graft must be done when the bark is slipping, preferably just as the buds of the stock tree are starting to grow.

It is usually advisable to obtain an ample amount of good-quality scionwood prior to grafting and store it under the proper conditions, although for broad-leaved evergreens, such as avocado or citrus, scionwood can be collected at the time of the grafting operation. See the earlier section on the selection and handling of scionwood (page 496).

In preparing for topworking, one must decide for each individual rootstock tree how many scaffold branches, if any, should be used (usually 3 to 5). However, no scaffold branches are retained when topworking pecans in Texas with the inlay bark method.

Double-Working (Grafting or Budding) A double-worked plant has three genetically distinct parts: the rootstock, the interstock, and the scion (Fig. 12–50). Such a plant has two unions, one between the rootstock and interstock and one between the interstock and the scion. The interstock may be less than 25 mm (1 in) in length or extensive enough to include the trunk and secondary scaffold branches of a tree.

Double-working is used for various purposes (see Chapter 11 and Chapter 13). Examples of double-working are (a) the propagation of 'Bartlett' pears on quince as a dwarfing rootstock by using a compatible interstock such as 'Old Home' or 'Hardy' pear (Fig. 12–50, page 506), and (b) the propagation of dwarfed apple trees consisting of the scion cultivar grafted onto a dwarfing 'M 9' or 'M 27' interstock that is grafted onto a more vigorous rootstock such as 'MM 106,' 'M 111,' or apple seedlings (Fig. 11–11) (10). Another form of double working is using a bridge-graft with a dwarfing rootstock (as the interstock) on older, mature apple rootstock in an established orchard, as previously described (53).

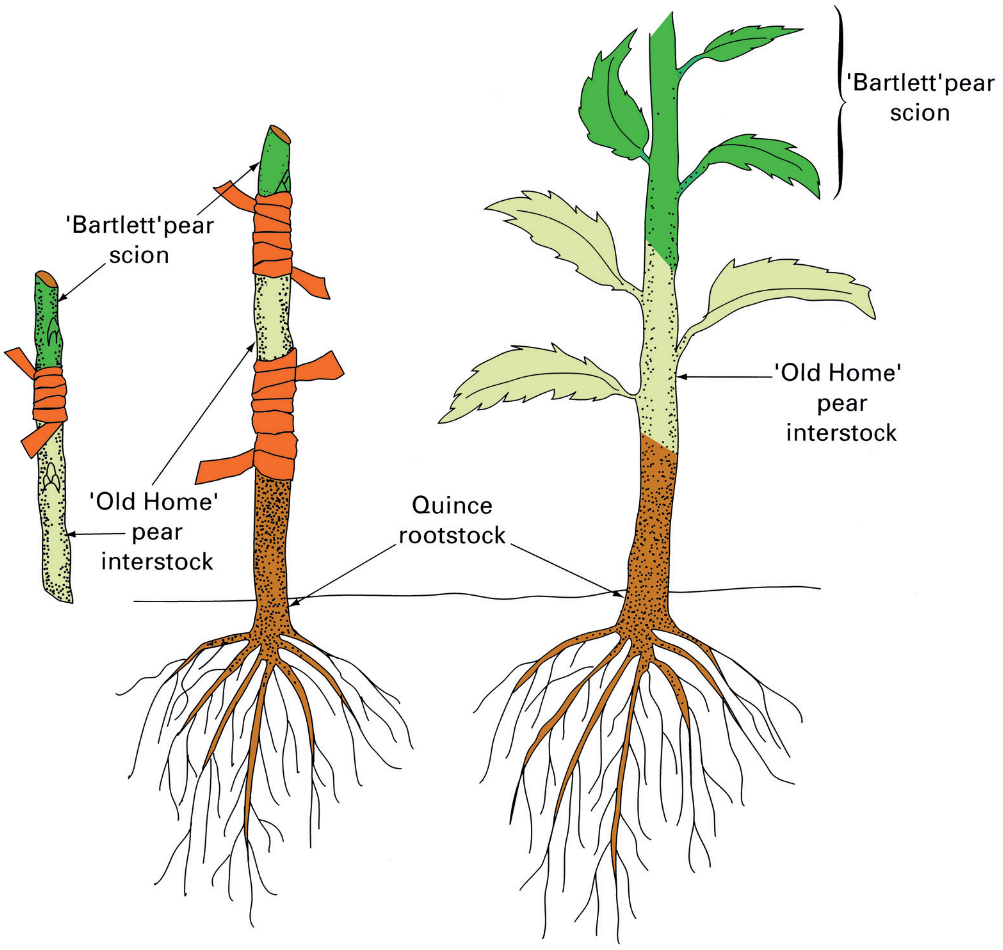

Figure 12–50 There are three genetically distinct parts and two graft unions in a doubleworked plant. The 'Bartlett' pear scion is grafted on an 'Old Home' pear interstock—then grafted to quince rootstock.

Several methods are used for developing double-worked nursery trees. The grafting in these techniques can be done with the whip graft:

- Rootstock "liners"—seedlings, clonal rooted cuttings, or rooted layers—are set out in the nursery row in early spring. These are then fall-budded with the interstock buds, growth from which, a year later, is fall-budded with the scion cultivar buds. Generally, 3 years are required to produce a nursery tree by this method.
- The interstock piece is bench grafted onto the rooted rootstock—either a seedling or a clonal stock—in late winter. After callusing, the grafts are lined-out in the nursery row in the spring, and fall-budded to the scion cultivar. By this method, the nursery tree is propagated in 2 years.
- A variation of the previous method is to prepare, by bench grafting, two graft unions—the scion grafted to the interstock and the interstock grafted to the rooted rootstock. After callusing, the completed graft, with two unions, is lined-out in the nursery row. Depending on growth rate, a nursery tree can be obtained in 1 or 2 years.
- Double-shield budding (T-budding) is used for double-working in one operation by budding (Fig. 13–21). A nursery tree is produced in 1 year or, if growth is slow, 2 years after budding.
- The interstock shoots still on the plant can be T-budded in late summer with the scion buds inserted about 15 cm (6 in) apart. During late winter, the budded interstock shoots are removed with the budded scion at the terminal end of each piece and bench grafted with a whip graft onto seedling rootstocks. After callusing, the completed graft—now consisting of rootstock interstock and a budded scion—is ready for planting in the nursery row (16).

Bench Grafting Systems

The term **bench grafting (bench working)** traditionally refers to any graft procedure performed on a rootstock and scion that are not initially planted, including root graftage, nurse-root graftage, or any graftage performed on bare-root rootstock. Bench grafting also applies to potted liner rootstock that is grafted on a bench or table, as is commonly done with selected

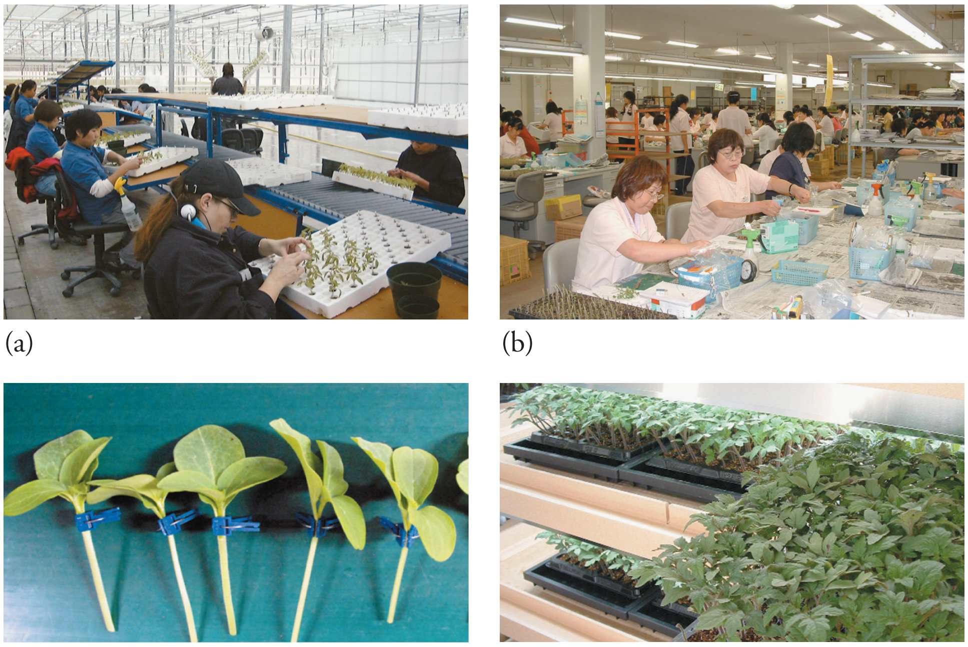
(a) (b) (c) (d)

Figure 12–51
(a and b) Bench grafting of vegetables in Canada and Japan. (c) Machine grafted cucumber cuttings.
(d) Growing grafted seedlings indoors under artificial lights (29). Courtesy C. Kubota.

woody ornamental species (43) or selected vegetable crops (Fig. 12–51) (29). Certain approach grafts are bench grafted, while others, such as the spliced approach graft of mangos in India, are field grafted, with the potted rootstock grafted to the established scion in the field.

Herbaceous types of plants are grafted for various purposes, such as studying virus transmission, stock-scion physiology, and grafting compatibility, as well as for the commercial greenhouse and field production of selected vegetable crops, particularly in Japan, Korea, and Europe (Fig. 12–51). The rootstock is grafted shortly after seed germination, while the plants are quite small. Such material is generally very soft, succulent, and susceptible to injury. The one cotylendon graft (OCG) or Japanese tube graft, is described in detail (see Figs. 12–6, 12–7, and 12–46) (21). Automated procedures of vegetable grafting using robotics and plastic grafting clips were described earlier (Figs. 12–43, 12–44, and 12–45).

Miscellaneous Grafting Systems

Cutting-Grafts In the cutting-graft, a leafy scion is grafted onto a leafy, unrooted stem piece (which is to become the rootstock), and the combination is placed in a rooting medium under intermittent mist for simultaneous grafting and rooting of the rootstock. Leaves must be retained on the rootstock piece in order for it to root. This procedure was utilized many years ago in studying stock-scion physiology in citrus (20), and has been used in commercial propagation of various types of citrus on clonal dwarfing rootstocks (13). It is also of value in propagating certain difficult-to-root conifers (55), rhododendrons (15, 41), and macadamias (1), as well as a number of apple, plum, and pear cultivars (44). It is used in the Netherlands and Israel in propagating greenhouse roses, where it is called **stenting** (Figs. 12–52 and 12–53) (58, 59).

stenting A Dutch term for simultaneously grafting and rooting the rootstock.

For citrus, a simple splice graft is used. The slope of the cut is at a 30-degree angle 1.3 to 2 cm (1/2 to 3/4 in) long; the union is tied with a rubber band. The base of the rootstock is dipped into an auxin, such as IBA, and then the grafts are placed under mist, or in a closed case, in flats of the rooting medium over bottom heat. After healing of the union and rooting of the stock, the grafts are allowed to harden by discontinuing the mist and bottom heat for about 2 weeks. Then the grafts are ready to be planted in 3.8-liter (1-gal) containers.

Micrografting Grafting of tiny plant parts can be done aseptically using tissue culture techniques described in Chapters 17 and 18, in which the small grafts are grown in closed containers until they are large enough to be transferred to open conditions. Micrografting has been used mostly with citrus, apple, and some *Prunus* species to develop virus-free plants, where a virus-free shoot tip can be obtained but cannot be rooted. The shoot tip is grafted aseptically onto a virus-free seedling, thus providing a complete virus-free plant from which other

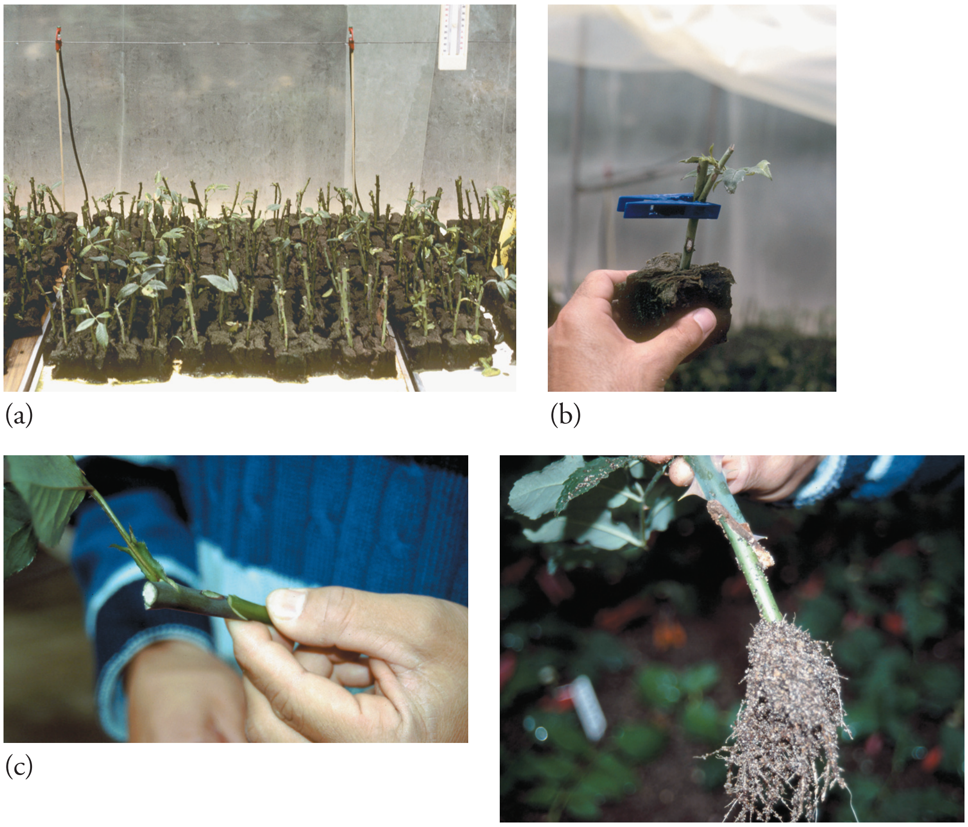
(a) (b) (c) (d)

Figure 12–52
Simultaneously grafting and rooting roses for cut flower production in Israel. (a and b) Rooting in Rockwool rooting blocks. (b and c) Grafted rose with plastic graft clip for simultaneous rooting and successful callus bridge formation. (d) Rooted, stented rose.

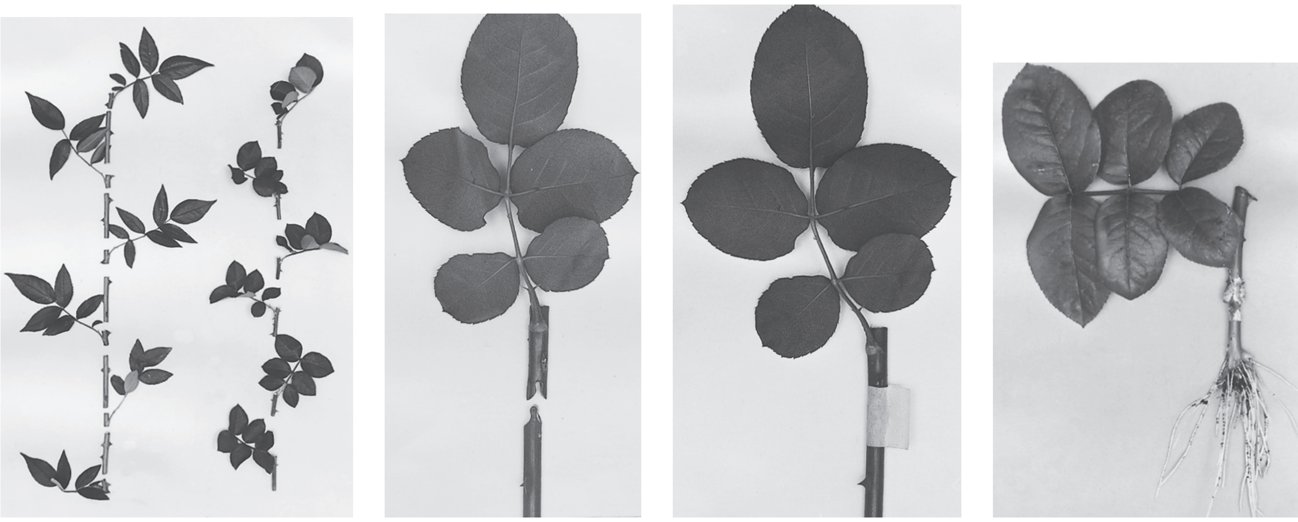
(a) (b) (c) (d)

Figure 12–53
Roses in the Netherlands being propagated by simultaneous rooting and grafting. (a) *Left:* Shoot cut apart to be used for rootstocks. Only internodes are used. *Right:* Sections cut for scions. One leaf is used per scion. (b) Saddle graft made with an Omega grafting machine. (c) Graft wrapped with tape for healing. (d) Completed graft with union healed and stock well rooted, ready for planting. In the Netherlands this process is called "stenting," a contraction of the Dutch words stekken ("to strike a cutting") and enten ("to graft"). Courtesy P. A. Van de Pol (58, 59).

"clean" plants can then be propagated (8, 25, 27, 46). Tests conducted some years after the shoot-tip grafts were made and the resultant trees were fruiting showed that the fruits were normal for the cultivar, disease-free, and with no variations appearing (45). The procedures for micrografting are described in Chapter 17.

Micrografting can also be done without tissue culture techniques. The non-aseptic propagation of very small nursery trees by grafting tiny seedlings with match-like scions, then growing the minute grafts long enough to have a viable plant, is a promising procedure (60). This is useful particularly in the tropics, where nursery plants sometimes must be shipped long distances, often by air, into inaccessible regions. Quantities of such tiny plants can be transported much more readily than full-sized nursery trees.

DISCUSSION ITEMS

1. What are five important requirements necessary for successfully producing a grafted plant?
2. What are the three major types (classifications) of grafts, and what criteria are used to categorize a specific graft into one of these three types?
3. What are the three types of nurse-root grafts? Why use an expensive process such as nurse-root grafting?
4. How does approach graftage differ from repair graftage? Give examples of grafts used.
5. Why are grafting waxes not used as much today as in the past? What kinds of substitutes are being used in their place?
6. Why is there increased interest in grafting automation/robotics?
7. What are some important considerations in the collection, handling, and storage of scionwood for grafting?
8. The craftsmanship of grafting includes both art and skill. How can an individual improve techniques and ergonomics to become a better grafter?
9. What are some good cultural techniques to enhance aftercare in topworking (topgrafting) an orchard with a new cultivar?
10. How do field grafting systems differ from bench grafting systems? Include examples of the various types of grafts in your answer.
11. Why the interest in herbaceous grafting, and with what crops is this technique of commercial importance? Why does this type of grafting lend itself to automation/robotics?
12. Contrast cutting grafts with micrografting, and give examples of their commercial use.

REFERENCES

1. Ahlswede, J. 1985. Twig grafting of macadamia. *Comb. Proc. Intl. Plant Prop. Soc.* 34:211–14.

2. Alley, C. J. 1957. Mechanized grape grafting. *Calif. Agr.* 11:3,12.

3. Alley, C. J. 1970. Can grafting be mechanized? *Comb. Proc. Intl. Plant Prop. Soc.* 20:244–48.

4. Anonymous. 1968. Pre-cut scions. *Agr. Res.* 17:11.

5. Beineke, W. F, 1978. Parafilm: A new way to wrap grafts. *HortScience*. 13:284.

6. Bhar, D. S., R. J. Hilton, and G. C. Ashton. 1966. Effect of time of cutting and storage treatment on growth and vigor of scions of *Malus pumila* cv. McIntosh. *Can. J. Plant Sci.* 46:69–72.

7. Brase, K. D. 1951. The nurse-root graft, an aid in rootstock research. *Farm Res.* 17:16.

8. Burger, D. W. 1985. Micrografting: A tool for the plant propagator. *Comb. Proc. Intl. Plant Prop. Soc.* 34:244–48.

9. Carpenter, E. L. 1989. How records can improve grafting. *Comb. Proc. Intl. Plant Prop. Soc.* 39:413–15.

10. Cummins, J. N. 1973. Systems for producing multiple-stock fruit trees in the nursery. *Plant Propagator.* 19:7–9.

11. Davis, A. R., P. Perkins-Veazie, R. Hassell, S. R. King, and X. Zhang. 2008a. Grafting effects on vegetable quality. *HortScience* 43:1670–2.

12. Davis, A. R., P. Perkins-Veazie, Y. Sakata, S. Lopez-Galarza, J. V. Maroto, S. G. Lee, Y. C. Huh, Z. Sun, A. Miguel, S. R. King, R. Cohen, and J. M. Lee. 2008b. Cucurbit grafting. *Crit. Rev. Plant Sci.* 27:50–74.

13. Dillon, D. 1967. Simultaneous grafting and rooting of citrus under mist. *Comb. Proc. Intl. Plant Prop. Soc.* 17:114–17.

14. Douglas, G. C., and J. McNamara. 1996. A tube method for grafting small diameter scions of the hardwoods *Quercus*, *Fraxinus*, *Betula* and *Sorbus*. *Comb. Proc. Intl. Plant Prop. Soc.* 46:221–26.

15. Eichelser, J. 1967. Simultaneous grafting and rooting techniques as applied to Rhododendrons. *Comb. Proc. Intl. Plant Prop. Soc.* 17:112.

16. Fisher, E. 1977. The pre-budded interstem: A new technique. *Fruit Var. J.* 31:14–5.

17. Flemer, W., III. 1986. New advances in bench grafting. *Comb. Proc. Intl. Plant Prop. Soc.* 36: 545–49.

18. Gaggini, J. B. 1985. Bench grafting of trees under polyethylene. *Comb. Proc. Intl. Plant Prop. Soc.* 34:646–47.

19. Garner, R. J. 1988. *The grafter's handbook.* 5th ed. New York: Oxford Univ. Press.

20. Halma, F. F., and E. R. Eggers. 1936. Propagating citrus by twig-grafting. *Proc. Amer. Soc. Hort. Sci.* 34:289–90.

21. Hassell, R. L., F. Memmott, and D. G. Liere. 2008. Grafting methods for watermelon production. *HortScience* 43:1677–9.

22. Hearman, J., A. B. Beakbane, R. G. Hatton, and W. A. Roach. 1936. The reinvigoration of apple trees by the inarching of vigorous rootstocks. *J. Pom. Hort. Sci.* 14:376–90.

23. Hottes, A. C. 1937. *Plant propagation: 999 questions answered.* New York: De La Mare Co.

24. Howard, G. S., and A. C. Hildreth. 1963. Induction of callus tissue on apple grafts prior to field planting and its growth effects. *Proc. Amer. Soc. Hort. Sci.* 82:11–5.

25. Huang, S., and D. F. Millikan. 1980. *In vitro* micrografting of apple shoot tips. *HortScience* 15:741–43.

26. Intven, W. J., and T. J. Intven. 1989. Apical grafting of *Acer palmatum* and other deciduous plants. *Comb. Proc. Intl. Plant Prop. Soc.* 39:409–12.

27. Jonard, R., J. Hugard, J. Macheix, J. Martinez, L. Mosella-Chancel, J. Luc Poessel, and P. Villemur. 1983. *In vitro* micrografting and its application to fruit science. *Scientia Hort.* 20:147–59.

28. Kerr, W. L. 1935. A simple method of obtaining fruit trees on their own roots. *Proc. Amer. Soc. Hort. Sci.* 33:355–57.

29. Kubota, C., M. A. McClure, N. Kokalis-Burelle, M. G. Bausher, and E. N. Rosskopf. 2008. Vegetable grafting: History, use, and current technology status in North America. *HortScience* 43:1664–9.

30. Kurata, K. 1994. Cultivation of grafted vegetables II. Development of grafting robots in Japan. *HortScience* 29:240–44.

31. Lagerstedt, H. B. 1982. A device for hot callusing graft unions of fruit and nut trees. *Comb. Proc. Intl. Plant Prop. Soc.* 31:151–59.

32. Lagerstedt, H. B. 1984. Hot callusing pipe speeds up grafting. *Amer. Nurs.* 160:113–17.

33. Lee, J. M. 1994. Cultivation of grafted vegetables 1. Current status, grafting methods, and benefits. *HortScience* 29:235–39.

34. Lee, J. M., and M. Oda. 2003. Grafting of herbaceous vegetable and ornamental crops. *Hort. Rev.* 28:61–124.

35. Legare, M. 2007. The future of grafting. *Comb. Prod. Intl. Plant Proc. Soc.* 57:380–84.

36. Leiss, J. 1987. Modified side graft for nursery trees. *Comb. Proc. Intl. Plant Prop. Soc.* 36:543–44.

37. Lincoln, F. B. 1938. Layering of root grafts—a ready method for obtaining self-rooted apple trees. *Proc. Amer. Soc. Hort. Sci.* 35:419–22.

38. MacDonald, B. 1986. *Practical woody plant propagation for nursery growers.* Portland, OR: Timber Press.

39. McEachern, G. R., and A. Stockton. 1993. The four-flap graft. In G. R. McEachern and L. A. Stein, eds. *Texas pecan handbook.* TAEX Hort Handbook 105. College Station, TX: Texas Agricultural Extension Service, Texas A&M University.

40. McEachern, G. R., S. Helmers, L. Stein, J. Lipe, and L. Shreve. 1993. Texas inlay bark graft. In G. R. McEachern and L. A. Stein, eds. *Texas pecan handbook.* TAEX Hort Handbook 105. College Station, TX: Texas Agricultural Extension Service, Texas A&M University.

41. McGuire, J. J., W. Johnson, and C. Dawson. 1987. Leaf-bud or side graft nurse grafts for difficult-to-root Rhododendron cultivars. *Comb. Proc. Intl. Plant Prop. Soc.* 37:447–49.

42. McPhee, G. R. 1992. Grafting techniques. *Comb. Proc. Intl. Plant Prop. Soc.* 42:51–3.

43. Meacham, G. E. 1995. Bench grafting, when is the best time? *Comb. Proc. Intl. Plant Prop. Soc.* 45:301–4.

44. Morini, S. 1984. The propagation of fruit trees by grafted cuttings. *J. Hort. Sci.* 59:287–94.

45. Nauer, E. M., C. N. Roistacher, T. L. Carson, and T. Murashige. 1983. *In vitro* shoot-tip grafting to eliminate citrus viruses and virus-like pathogens produces uniform bud-lines. *HortScience* 18:308–9.

46. Navarro, L., C. N. Roistacher, and T. Murashige. 1975. Improvement of shoot tip grafting

in vitro for virus-free citrus. *J. Amer. Soc. Hort. Sci.* 100:471–79.

47. Oda, M., and T. Nakajima. 1992. Adhesive grafting of Chinese cabbage on turnip. *HortScience* 27:1136.

48. Oda, M., K. Okada, H. Sasaki, S. Akazawa, and M. Sei. 1997. Growth and yield of eggplants grafted by a newly developed robot. *HortScience* 32:848–49.

49. Oda, M., T. Nagaoka, T. Mori, and M. Sei. 1994. Simultaneous grafting of young tomato plants using grafting plates. *Scientia Hort.* 58:259–64.

50. Passam, H. C. 2003. Use of grafting makes a comeback. *Fruit Veg. Tech.* 3:7–9.

51. Patrick, B. 1992. Budding and grafting of fruit and nut trees at Stark Brothers. *Comb. Proc. Intl. Plant Prop. Soc.* 42:354–56.

52. Rivard, C., and F. Louws. 2006. *Grafting for disease resistance in heirloom tomatoes.* Raleigh, NC: North Carolina Cooperative Extension Service, AG-675. E07:45829.

53. Samad, A., D. L. McNeil, and Z. U. Khan. 1999. Effect of interstock bridge grafting (M9 dwarfing rootstock and same cultivar cutting) on vegetative growth, reproductive growth and carbohydrate composition of mature apple trees. *Scientia Hort.* 79:23–8.

54. Singha, S. 1990. Effectiveness of readily available adhesive tapes as grafting wraps. *HortScience* 25:579.

55. Teuscher, H. 1962. Speeding production of hard-to-root conifers. *Amer. Nurs.* 116:16.

56. Upchurch, B. L. 2006. Grafting with care. *Amer. Nurs.* 203:18–22.

57. Upshall, W. H. 1946. The stub graft as a supplement to budding in nursery practice. *Proc. Amer. Soc. Hort. Sci.* 47:187–89.

58. Van de Pol, M. H., A. J. Joosten, and H. Keizer. 1986. Stenting of roses, starch depletion and accumulation during the early development. *Acta Hort.* 189:51–9.

59. Van de Pol, P. A., and A. Breukelaar. 1982. Stenting of roses: A method for quick propagation by simultaneously cutting and grafting. *Scientia Hort.* 7:187–96.

60. Verhey, E. W. M. 1982. Minute nursery trees, a breakthrough for the tropics? *Chronica Hort.* 22:1–2.

61. Westergaard, L. 1997. Improved grafting techniques for nursery stock 1997. *Comb. Proc. Intl. Plant Prop. Soc.* 42:354–56.

Edible Wild Plants

A NORTH AMERICAN FIELD GUIDE

Thomas S. Elias & Peter A. Dykeman

AN
OUTDOOR
LIFE BOOK

Sterling Publishing Co., Inc. New York

Library of Congress Cataloging-in-Publication Data

Elias, Thomas S.
Edible wild plants : a North American field guide / Thomas S. Elias & Peter A. Dykeman.
p. cm.
Includes index.
Reprint. Originally published: New York : Outdoor Life Books, c1982.
1. Wild plants, Edible—United States—Identification. 2. Wild plants, Edible—Canada—Identification. 3. Cookery (Wild foods) I. Dykeman, Peter A., 1935– II. Title.
QK98.5.U6E35 1990
581.6'32'0973—dc20 90-39506
CIP

10 9 8 7

Published 1990 by Sterling Publishing Company, Inc.
387 Park Avenue South, New York, N.Y. 10016
Originally published as *Field Guide to North American Edible Wild Plants*
by Outdoor Life Books © 1982 by Thomas S. Elias and Peter A. Dykeman
Distributed in Canada by Sterling Publishing
% Canadian Manda Group, P.O. Box 920, Station U
Toronto, Ontario, Canada M8Z 5P9
Distributed in Great Britain and Europe by Cassell PLC
Villiers House, 41/47 Strand, London WC2N 5JE, England
Distributed in Australia by Capricorn Ltd.
P.O. Box 665, Lane Cove, NSW 2066
Manufactured in the United States of America
All rights reserved

Sterling ISBN 0-8069-7488-5

Contents

Acknowledgments 4

Introduction 5
- How to Use This Book 5
- Nature's Diversity 7
- What Determines the Taste of Wild Plants? 9
- Danger of Overharvesting 9
- Habitat Guide 10
- Harvest and Preparation 12
- Plant Foods of the Indians 22
- Nutrition 25
- Other Plant Identification Guides 27

Key to Edible-use Symbols 29

Seasonal Key to Plants 30

Spring 56

Summer 126

Autumn 208

Winter 248

Poisonous Plants 258

Appendix: Nutritional Contents of Selected Foods 274

General Index 277

Index to Plant Names (Common and Scientific) 278

Acknowledgments

This book is a compendium of information we gathered over many years of study. Of course, any work on edible wild plants builds upon the experiences of others, beginning with the accounts of the American Indians and early European settlers, and continuing to the present.

Many people assisted us in the preparation of this book. Central to the entire effort was Neil Soderstrom, who served as editor and friend. His ideas helped make the book more useable and his encouragement and prodding brought it to fruition. Jeff Fitschen designed the pages and created the maps. Lloyd Birmingham drew the edible symbols. Helen Dykeman helped with typing, and Brian Elias assisted in some of the photo sessions. We are also grateful to the New York Botanical Garden for use of its botanical library resources, among the finest in North America.

While the majority of the photographs here are our own, many came from other sources. We are grateful to friends, colleagues, and institutions that supplied color transparencies. We've listed full names of photo sources below, alphabetically. Credits by last name or institutional abbreviation appear in parentheses in photo captions.

Dr. Loran C. Anderson
Nelson T. Bernard Jr.
Frank P. Bogel
James Clawson
Bill Cramer
Dr. Garrett Crow
Desert Botanical Garden, Phoenix, AZ
Dr. Peter A. Dykeman
Dr. Thomas S. Elias
Osbra L. Eye
Dr. Albert Feldman
Dr. George W. Folkerts
Dr. James Hardin
Jepson Herbarium, University of California, Berkeley
Bob Hutchins
Berrie Kavasch
Arthur R. Kruckenberg
Dr. Dale McNeal
John Mathisen
Dr. Robert H. Mohlenbrock
Helen Mulligan
New York Botanical Garden (NYBG), Bronx, NY
Dr. Robert Ornduff
John Pawloski
Robert A. Ross
James R. Shevock
Neil Soderstrom
Robert Speas
James Stevenson
University of California, Davis
U.S. Dept. of Agriculture (USDA)
C. S. Webber
Dr. William A. Weber
Kay Young

T.S.E. and P.A.D.

Introduction

How to use this book

This is a season-by-season guide to identification, harvest, and preparation of over 200 of the most popular and common edible wild plants of the U.S. and Canada. Hundreds of related edible species are also noted. In addition, the book covers identification of 20 poisonous plants and notes many other poisonous ones. Neither edible nor poisonous mushrooms are included.

Detailed texts, maps, drawings, and nearly 400 photos (mostly in color) allow easy identification. But since there are over 20,000 species of plants to choose from in North America, excluding Mexico, it's important to make identifications with care and caution. The trick is to carefully compare features of the plant at hand against those here. Remember this rule: *When in doubt, leave the plant out.*

We have included edible plants for each season from every major region of the U.S. and Canada, except for the subtropical areas of southern Florida. We have excluded obscure plants and plants that are only marginally edible. Likewise, we have omitted plants that occur in limited numbers or are threatened. Two such plants are the Joshua tree *(Yucca brevifolia)* and the Saguaro cactus *(Cereus giganteus),* both slow-growing trees of the southwestern U.S. Rare or threatened species should never be gathered.

The book's plant-seeking and identification systems are simple and versatile. In most cases, you'll probably want to begin by consulting the "Seasonal Key to Plants," beginning on page 30. This Key groups plants by *primary* "edible season," with other edible seasons noted there as well. Within the Key, you can easily hunt up plants in various ways: whether by common or scientific name, by distribution range, by general plant type, by habitat, or by edible options—whether you have an urge for something boiled, baked, raw, and so on. The more you use the Key, the more you'll like it.

On the other hand, you may know a plant's scientific or common name and simply want to learn more about it. In this case, just find its page reference in the Index in the back of the book. *Caution:* Often, unrelated plants share a common name—from region to region, and from guidebook to guidebook. Thus, it is wise to learn or at least check the scientific names.

At other times, you might prefer to browse a seasonal section in the body of this book itself to see what kinds of plants you might encounter on your next outing. There you'll find plants arranged as they are in the Key, by edible season. Within each of these seasonal sections, plants are arranged by type. That is, ferns appear first. Then come monocots such as the lilies and grasses. Last come the largest group, the dicots, which include herbs (most soft-stemmed plants), then vines and shrubs, and finally trees. Within the herb groupings themselves, the plants are arranged according to flower color and leaf arrangement.

The following table gives contrasting features of monocots and dicots.

Monocots	**Dicots**
(such as onions, wild rice, cattail, corn lily, sweet flag, reed, camass lily)	*(such as amaranth, mustard, blackberries, strawberries, sunflower, plums, cherries, blueberries)*
Narrow leaves with parallel veins	Broader leaves with netlike veins
Flower parts in 3s or multiples of 3	Flower parts in 5s, sometimes 4s
Vascular strands (bundles) scattered throughout a nonwoody stem	Vascular strands (bundles) arranged in a ring on stem that may be woody
One seed leaf per seed	Two seed leaves per seed

Common names and scientific names. Common names are useful, especially so if they are descriptive, but they often vary from region to region. Adding to the disadvantage of using only common names, several plant species may be known by the same common name. In this book, we've preferred to use a single common name but may list a second or even third common name when the names are widely used. But you'd be wise to confirm scientific names when consulting other plant guides or when using common names in everyday discussions.

The scientific name, in Latin, is important because it is the universally accepted name. Scientific names for all plants and animals adhere to a binomial (two-word) system and are always italicized. The first word designates the genus and begins with a capital letter. A genus is a group of closely related species. The second word indicates the species and begins with a small letter. A species is a group (or groups) of individuals that look similar and can freely interbreed. Hybrids between species do occur though.

Typographical tipoffs. The texts covering each plant in this book employ typographic variations that highlight categories for quick reference and announce whether a mentioned plant is covered elsewhere in this book. As in the example shown two paragraphs below, each species text begins with the common name or names, beneath which are listed additional seasons in which the plant is edible. Then comes the scientific name, always in *italics,* which may be followed by an italicized synonym in parentheses. This synonym is the outdated scientific name under which you may find the plant listed in older plant guides.

Bold face type draws your eyes to the sequence of descriptive categories. ***Bold face italic*** announces primary identifying features. SMALL CAPITAL LETTERS are used for common names of plants covered in detail elsewhere in the book and located by means of the Index. For mentioned plants not covered in detail, both the common and the scientific names are given. To look up these plants in other guides, use the scientific name.

Utah juniper *Juniperus osteosperma*
Also edible autumn, winter (synonym: *J. utahensis*)

Habitat: mesas, mountain slopes, high plains, principally from 1,000–2,650 m (3,300–8,700 ft) elevation in dry rocky or gravelly soils. **Identification:** large ***shrub*** or small ***bushy tree*** to 8 m (26 ft) high with ***furrowed ash-gray bark that shreds in long, loosely attached strips. Leaves*** tiny, ***scalelike, opposite and overlapping one another in 2s, closely pressed against branchlets,*** 3–4 mm (0.1–0.2 in) long, usually gland-dotted on back, ***sharp-pointed at tip, yellowish-green.*** Male and female flowering cones usually produced in spring on separate trees, inconspicuous and usually overlooked. ***Fruits small, leathery berries 6–9 mm*** (0.2–0.4 in) ***in diameter, nearly globe-shaped, reddish-brown, resinous, sweet,*** with mealy texture, containing 1, rarely 2, seeds. **Harvest and preparation:** same as for ROCKY MOUNTAIN JUNIPER. **Related edible species:** CALIFORNIA JUNIPER and ROCKY MOUNTAIN JUNIPER (preceding). **Poisonous look-alikes:** same as for ROCKY MOUNTAIN JUNIPER.

Nature's diversity

Some plants taste good and are easy to digest, while others are marginally so. Many others are unpalatable or even poisonous. Why? Plants, like animals, are products of millions of years of evolution and adaptation. Through all this time and even today, they are subject to many influences, including changes in climate, soil, and water; competition from other plants; and the scores of plant-eating animals, ranging from insects to mammals. Some animals, particularly the insects, have gradually specialized by feeding on and laying eggs on specific plant species and groups. If a plant fails to adapt to these external influences, the result can be serious depletion or even extinction of its species.

Plants, like animals, have developed a myriad of defense mechanisms to counter the continuing attacks by animals. But, unlike animals, plants cannot pick up their roots and race for safety. Nor have many plants taken the offensive and become predators; the parasitic and insectivorous plants are remarkable exceptions. So to defend themselves or ensure survival of their species, most plants have developed elaborate mechanical, chemical, and reproductive characteristics.

Reproduction. Some plant species survive largely because of their high reproductive capacity. Here, the plants reach sexual maturity early, produce large quantities of seeds or other propagules, and may be adapted to a wide variety of habitats, especially on sites disturbed by man. In these ways, dandelions, chickweed, shepherd's purse, and other plants channel much of their resources into reproduction. With their rapid growth and early maturation, they outcompete many slower-growing plants. This phenomenon is easiest to observe in yards and on land reverting back to a wild state.

Armament. Still other plants employ an arsenal of external structures (spines, thorns, stinging hairs) that can deter larger animals from browsing. These defensive structures often occur in conjunction with internal mechanisms such as tough fibers and bitter-tasting or even poisonous chemical compounds. Did you ever notice how many thistles remain uneaten in fields otherwise heavily browsed by deer or cattle?

External chemical defense. Another defense is the production of tiny glands or glandular hairs on the plant surfaces. These structures may produce an array of chemical compounds that may even be secreted onto the surfaces of stems and leaves. Secreted substances usually glisten in direct light and are sticky. They may also emit a repulsive odor. Many members of the sunflower family have tiny glands on the leaves, stems, or outer surface of the flower clusters. As a test, place some fresh leaves and flowers of tarweed (*Grindelia* species) in a plastic bag for a few hours and then open the bag. You'll smell a heavy resinous odor from the glandular secretions. The secretion and odor are believed to repel leaf-eating animals.

Internal chemical defense. Many plants produce chemical compounds within the leaves, flowers, and fruits, which make the plant taste bad or render it toxic.

The cholla's armament, below left, consists of spines that readily pierce flesh. Spine tips may break off when you attempt to remove them. Below right, the glistening, sticky coating on tarweed flowers is secreted by tiny flower glands. As a chemical defense, it renders the plants unappetizing to potential predators. (Elias photos)

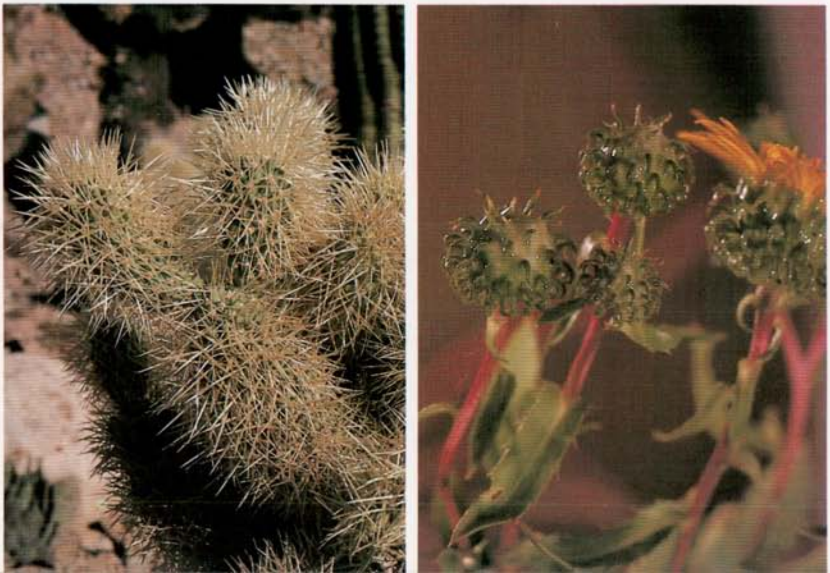

This is why some recipes later in this book tell you to soak the plant parts in water first or to discard the water after the plants have been cooked. Other recipes recommend using only young leaves or stems because the young parts do not contain high levels of the undesirable chemical compounds. Also, young parts are tenderer because the vascular strands and fibers have not had enough time to enlarge and toughen. Even extended cooking won't render all plants edible. The poisonous components in such plants as irises, water hemlock, and others cannot be broken down, nor can they be leached out by water changes or cooking.

Appeal to animals. There are times when it is to a plant's advantage to be eaten. After plants have flowered and produced fruits, reproduction and, ultimately, survival of the species require that the seeds be dispersed to sites suitable for germination and growth. To achieve seed dispersal, plants use unwitting animals. Consider the tasty and nutritious flesh of wild berries. The tiny seeds inside many kinds of berries pass through animals unaffected by the digestive process and are so dispersed in excrement.

What determines the taste of wild plants?

Experienced foragers know that the taste of a particular species from one site may differ slightly from that of the same species at another site. This is due to genetic variation usually manifested in differing physical or chemical characteristics. Thus, the degree of bitterness or mildness in many members of the mustard and daisy families, for example, can be attributed to natural genetic variations. Chemical composition of the soil along with other substances added by man or animals can affect the taste of some plants, especially smaller herbs.

Then too, people themselves differ in their ability to taste and in their tolerance of certain plant compounds. Just as many foragers enjoy fresh dandelion greens and boiled cattail stalks, others may detest them. You might acquire a taste for certain wild edibles only after experimenting with harvest and preparation techniques.

Danger of overharvesting

Avoid overharvesting and thus eradication of a plant from a site. Try to make a note of the reproduction levels of wild edibles in your area. If you harvest native plants faster than they can reproduce, you'll have a detrimental impact on the environment. It's a good rule always to leave some mature plants to produce fruits and seeds.

Obviously, you won't cause serious ecological damage if you harvest plants introduced from other countries. Introduced plants are already alien to native habitats. The widespread invasions of these introduced plants in the wake of man's new settlements have permitted them to become established and in some

In summer, gravelly flood-scoured streambanks may be dominated by purple loosestrife, as shown left. Beneath the loosestrife, mature ostrich fern fronds are drying. Their brown plumes will remain until spring to help foragers find the fern's tasty young fiddleheads. Right, the density of this cattail "garden" on the edge of an artificial pond was annually increased by weeding. (Dykeman photos)

cases to overcome the natives. Examples of introduced species include dandelions, chickweed, mustard, and plantain.

Habitat guide

With the acquisition of this book, you have taken the first and easiest step toward a feast of wild edibles. Though good eating is a major goal here, don't overlook the potential fun in simply foraging for plants. With a little knowledge of plant ecology and field identification you can find an abundance of wild foods.

Except for nuts, acorns, and sap in season, forested areas have fewer edibles to offer than most unforested areas. Yet in northern deciduous forests you might find Indian cucumber and bunchberry, among many other plants. In hardwood forests early in the growing season, look for marsh marigold, toothwort, and several of the violets near small streams and adjacent mucky areas. In the rich,

Left, bulldozed two years prior to the photo at left, this "disturbed" site offers various edibles including dock (brown mature seed stalks), evening primrose (yellow-flowered stalks), and stinging nettle (dark green plants topped by long slender flower spikes). Right, desert habitats dominated by tall saguaro cacti, with edible fruits, may also offer edible prickly pear and mesquite. (Dykeman and Elias photos)

moist and neutral-soil woods you may find wild onions and leeks. Near the edge of the woods, you may see the broad leaves of garlic mustard. Here also dwell the greenbriers and catbrier.

In spring, along gravelly stream banks regularly scoured by floods, look for the delicious fiddleheads of the ostrich fern. The task is easier if you first notice the tall brown fern plumes from the previous year. Later this habitat may offer tender young nettles and one or both species of jewelweed, which, away from streams, may grow in moist, fairly open areas. Partially sunny areas in spring may reward you with mints and watercress.

Marshy pond edges often offer wild-food bonanzas. From the cattail ("Cossack asparagus") early in spring to the arrowhead tubers in the fall and winter, these sites provide continuous sources of foods. Among the providers are pickerelweed, cattail, common reed, arrowhead, water lilies, great bulrush, and wild rice. It's often wise to start the spring harvest by checking for new sprouts on the cattail roots. Fortunately, none of us needs to complete the harvest season in December the way the Indian squaws did, according to Lewis and Clark, by standing in chest-deep ice water, loosening arrowhead tubers from the mud with their toes.

Sights disturbed by bulldozers or plows are always worth a close look. For a grand variety of greens, from early spring to summer, visit such disturbed areas frequently. Lawns provide dandelions, chicory, plantain, and other greens best gathered as early as possible. In recently disturbed sites such as last year's garden, mustard, winter cress, chickweed, and shepherd's purse are followed by lamb's quarters, amaranth, purslane, and many other edible "weeds." A year or so after being disturbed, these sites may then produce docks, pokeweed, evening primrose, salsify, milkweed, and many other edibles.

Roadsides offer bounty closely akin to bulldozed areas. Yet auto traffic and its pollutants often make roadside foraging unhealthy. In June along a lightly traveled dirt road, you may want to cut poke or asparagus shoots, or pick common day lily buds. In fall, you might dig for Jerusalem artichokes or common day lily tubers.

While foraging for spring greens, note locations of the flowers of blackberry, raspberry, blueberry, and elderberry. Most old fields and forest edges produce one or more species of berry plants during the succession toward a woodland plant community. Elderberry blossoms are edible, but be sure to leave plenty for fruit production. In the Northeast, raspberries and blueberries fruit first, followed by blackberries and elderberries. If you live in an area with heavily fruiting serviceberries, first locate their white flowers in spring against the leafless grays and browns of woods and river edges. Gather the serviceberries that the birds don't get first, during blackberry time in midsummer.

In summer throughout the rough lands of the West, look for currants, yuccas, prickly pears, and immature mesquite pods.

If you plan your vacation to the East Coast for late summer, look for ripe beach plums. They are frequently plentiful, though few people seem to pick them. If camping in national or state parks, of course, first obtain permission to pick.

In the East, September is a great month for sumac berries along the edges of woods and in old fields. The staghorn sumac is best then and the smooth sumac, which reaches its peak earlier, still has good flavor. In the West, squawbush fruits are harvested in September. As the first frost approaches, grapes and viburnums ripen along brushy fencerows and wood's edge. Early fall is also the time to drive the country roads looking for the first fallen nuts. The best-producing nut trees are in the open and along roads and old fences. Make note of the nut tree locations early and check them periodically in order to beat wildlife to the nuts when they begin to fall. In fact, you can gain the advantage over wildlife by climbing nut trees and shaking the nuts loose.

Through the fall, look in fields for hawthorn and rose fruits undergoing their post-frost sweetening. After that, you might sit back and relax until the following spring's maple sugar season.

Winter foraging serves almost exclusively for survival fare or for diversion during a January thaw. If the wild food fever strikes you in winter, you'd be best served in warmer southern climes.

Harvest and preparation

Shoots, leaves, and other aboveground parts. For all aboveground tender parts, the time between harvest and eating is critical. Some plants lose much of their vitamin content in a few hours. So avoid saving them until they taste as flat as if they'd come from a supermarket!

It's also important to forage at correct times. Several greens, including chicory and winter cress, become bitter very early in the growing season. Pokeweed shoots increase in toxicity as they grow. And many other greens such as ostrich fern and plantain become tough or fibrous with age.

Even though the shoots lose palatability as they grow, new leaves and growing tips of many plants, such as miner's lettuce and lamb's quarters, remain tender and edible throughout the growing season. Plants such as milkweed, cattail, and common day lily also produce edible flower buds and flowers, and during the growing season are worth several foraging visits. As the years pass, you may want to develop your own regional calendar of the harvest dates for plant parts.

Some plants are best picked by hand—a leaf or tender tip at a time. Densely leaved plants such as sheep sorrel and chickweed are best pulled up whole before you use scissors to shear off the leaves. Since nettles, greenbrier, and some other edibles are armed in various ways, you'll need gloves and shears for harvesting.

Wash the freshly harvested plants with cold water. Then cook them with concern for their taste, tenderness, and nutritional value. With shoots, you can retain most vitamins by bundling and steaming them like asparagus. This works well with fern fiddleheads, cattail shoots, Japanese knotweed, and, of course, wild asparagus. Shoots with bitter sap like milkweed's or a toxic substance, as in pokeweed, may be bundled but must be submerged in changes of boiling water to render the plants safe to eat.

Normally you should avoid submerged boiling of tender, delicate leaves and

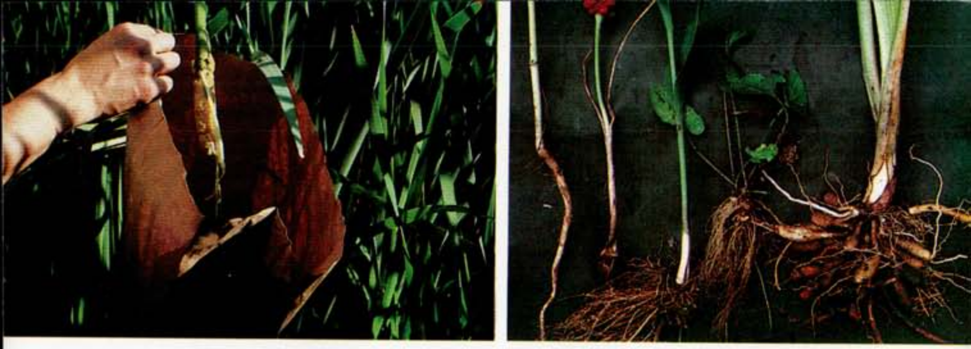

Left, for a nutritious flour additive, shake pollen from cattail male flower spikes. Plants at right (not all edible) illustrate terms used in this book to characterize underground plant parts: l. to r. are a wild carrot taproot, a jack-in-the-pulpit corm, a lily-of-the-valley bulb, violet rhizomes, and day lily tubers. (Dykeman photos)

growing tips. Here boiling has about the same effect as soaking your breakfast cereal in boiling water and then draining the water off: All the flavor and texture and much of the nutritional value are lost. For such plants, including lamb's quarters and chickweed, use just a few tablespoons of water beyond that which is left from washing. Within 1–3 minutes such greens are done to perfection.

Coarser greens such as dock and cleavers require longer cooking, barely covered with water. Cook until tender instead of just chewy. Plants with toxic or bitter substances must be boiled in changes of water, although the process reduces much of the vitamin content. A pinch of baking soda will help tenderize tough greens but promotes vitamin loss.

When possible, try to combine bland greens with those having a bitter or sour character. While coarse dry-textured greens are cooking, add a little bacon grease or butter. Season only *after* cooking.

Roots, tubers, and other underground parts. Several plants produce taproots, tubers, or corms that are excellent substitutes for potatoes and root crops. These plants are usually best from fall through early spring when most other wild foods are scarce. The underground structures serve as storehouses of carbohydrates and are richest in starch at the end of the growing season. As the earth warms in early spring the starch is converted to sugars that may sweeten the underground structures. When the plants begin to grow and most of the starches have been converted, the underground structures generally become soft and unpalatable.

But there are disadvantages to using underground parts as food. The aboveground parts of many plants die in autumn and either shrivel greatly or break away from the root, making foraging very difficult. Then positive identification is a tough challenge since you nearly always need aboveground parts for reference. Mistaking a species such as death camass *(Zygadenus)* for the edible camass *(Camassia)* would be fatal. As with all wild food gathering, be certain of what you collect.

Once you have gathered a batch of fleshy roots or other underground parts, cook them like potatoes. Options include boiling, baking, and frying, as well as the pit-roasting method used by Indians in dry climates. For pit roasting, dig a hole at least 2 ft by 3 ft and about 2 ft deep. Line the bottom and sides with rocks. Build an intense fire inside the pit to heat the stones and adjacent earth. After several hours of burning, clean out the burning wood and coals and line the rocks with green grass or leafy twigs. Lay in the food to be roasted and cover it with more grass. Pour in a pint of hot water to produce steam. Top the pit with 5 cm (2 in) of soil and build a fire on it. For groundnuts and arrowhead tubers, maintain the fire for 3–4 hours. To reduce camass bulbs to a sweet blackish mass, Indians cooked them from up to 3 days. Whatever way you cook underground edibles, leave the skins on to preserve most of the vitamins.

Underground parts of some plants are excellent raw as trail nibbles or in salads. Examples include wild onions, Indian cucumber, and cattail root tips. You can experiment to determine which others are neither too tough nor too bitter for eating raw.

Tough roots of cattails and some other species can be crushed to release the starch, which can then be dried for use as flour. Other species may be dried whole or sliced and stored for long periods, a very valuable preparation for Indians, who relied on seasonally available foods. Drying also removes the toxicity of the roots of plants such as skunk cabbage and jack-in-the-pulpit, which are not included in this book because of their toxicity when fresh.

Beverages. You can brew teas from a wide range of leaves, fruits, twigs, and even roots, many of them high in vitamin C. Most tea materials may be used fresh or dried; check the species texts in this book to learn which is best for each plant.

Pick leaves at their peak growth. Then dry them in a warm place out of direct sunlight. If you use a warm oven, remember that too much heat will drive off oils that give the beverage its flavor. Store dried leaves in a sealed container. The

Below, a berry scoop fashioned from a plastic container can greatly increase your harvest rate, but it won't distinguish between ripe and unripe berries. Right, you can use a fabric-covered board to bounce most hard, unripe berries beyond a collector box. (Dykeman photos)

same principles apply to the drying and storage of fragrant twigs and root bark and to flowers of such species as basswood and clover.

To prepare a cup of tea, 1 tsp of dried material or 2 tsp of fresh are usually enough. Steep the dried ingredients in boiling water, but do not boil the tea. Otherwise you'll drive off the fragrant oils. Like commercial teas, strong-flavored ingredients need only a few minutes of steeping. Delicately flavored strawberry leaves, birch twigs, and the like may require 15 minutes or more.

For flavor variations in teas, experiment with blends of ingredients. Since many plants make good teas, the combinations are limitless, as suggested by the labels on commercially packaged wild tea blends. *Caution:* Be certain that all of your ingredients are safe to use. In this book we've avoided mention of folk remedies and medicinal uses of plants. Some of these are unproven and even dangerous. And those that may have merit often depend on correct diagnosis of an ailment. You're better off using commercially prepared medicine. Still, a cup of birch tea for a queasy stomach won't hurt you.

Coffee substitutes are usually made from roasted, coarsely ground roots and tubers. Some, such as chicory and dandelion, are best harvested outside the growing season, while chufa tubers are firm and good to use anytime. Roast the material slowly, browning it all the way through. Too high a temperature will scorch it and give you a terrible cup of coffee. Store the prepared material in a sealed container. Generally use 2 tbsp of "coffee" per cup of water. Brew it like regular coffee but do not overbrew unless you have a cast-iron stomach.

Fruits and fruit juices. You'll need to handpick many fruits because they'll be attached too firmly to be shaken loose. But the shake method works well for bushes or trees such as mulberry, cherry, and highbush blueberry. Any large sheet of plastic or fabric will suffice. Just spread it under a portion of the plant and shake the branches above. You may be amazed at the quantities you harvest. Large trees may require climbing. This method works well on nut trees, too, and helps you beat the squirrels to the nuts.

If shaking won't work but the fruits are large and firm like cranberries, improvise a toothed scoop from a plastic 2 liter (2 quart) jug. With practice, you can make this scoop a highly efficient gathering device.

Some clustered fruits such as grapes, mountain ash berries, and elderberries will not yield to shaking or scoops. Here you'll need to take the whole cluster with fingers, knife, or pruning shears. Then you can strip the fruits from their stems. Soft fruits like strawberries and raspberries have to be picked by hand.

If you strip firm fruits such as blueberries from the twigs, you'll usually wind up with many unripe fruits in the bucket. You can sort out unripe fruits tediously by hand, or you can rig up a bounce board, as shown in the accompanying photo.

Drying Fruits. Drying processes are described later in many of the texts on individual species or plant groups. But as a general rule, select only fully ripe fruits. Dry small fruits whole, pit the pitted ones, and cut large ones into pieces or slices.

For best results place the fruits in a single layer on a rack made with a nylon screen or thin cloth. If drying is done outside, place the rack out of direct sunlight. Turn the fruit once or twice daily and cover the rack or take it inside at night. You may find it more convenient to place the rack in an attic, which may be ideally warm and dry during the summer. Just be certain there are no mice or squirrels there. Attic and outdoors drying requires from a few days to two weeks, depending on the fruit and the weather.

If you are ambitious, make a solar dryer, which is basically a sun absorbing box with adjustable vents to allow air exchange while maintaining a temperature below 80°C (175°F). The effect is the same as from using a warm oven, but the energy is free. Drying time in a solar box or kitchen oven will be from 6 hours to 2 days. Remember to turn the fruit occasionally to promote uniform drying. The process is completed when no juice can be squeezed from the fruit. If you fail to dry the fruit enough, it will mold. Store it in sealed plastic or glass containers away from light.

Most dried fruits store well but a few, such as saguaro, may have insect eggs on them and will become wormy in time. Avoid worms by pulping and boiling the fruit, then drying it in a thin sheet to make fruit leather. This lightweight, nutritious fruit form is great for trail use. For cooking use ¾ cup of dried fruit per cup of fresh fruit called for in a recipe. But soak the dried fruit about 15 minutes beforehand.

Cold drinks can be made from the juices of many fruits. Extract the juices after crushing the fruits and cooking them over low heat. To finish the beverage, you usually need only add sweetening and perhaps a little water. Grapes, cranberries, and red mulberries are among the best fruits for juices. To extract the flavors of drier fruits such as the sumacs, soak them in hot water. Served cold or hot with or without sweetening, sumac juice is delicious.

Saps from maples, birches, and other trees, taken during sugaring season, make excellent drinks. Sap is best within a day or two of harvest. Try some in spring while you are making maple syrup.

Jellies. Most jellies are made from fruits, but any leaves with fragrant oils can be used as a base, mints being the best example. For most preparations, select firm ripe fruit. Some fruits such as grapes produce better jelly from a mixture of ripe and unripe fruits.

Wash the fruits and remove blossom ends and stems. Cut up large fruits. Crush soft fruits to start the release of juice. Add water and cook until very soft or until all seeds are free from the pulp. Pour into a wet jelly bag and drain. Squeezing the bag increases quantity but reduces clarity of the juice.

Accompanying recipes are based on the use of commercial powdered pectin. But, depending on ripeness and type, some fruits may not need pectin. To determine the need for pectin, mix 1 tbsp each of juice and alcohol (grain or wood). If a solid mass forms, no additional pectin is needed. If flakes or a loose mass forms, add pectin. Slightly underripe fruits are generally richer in pectin than ripe ones.

Jelly Recipes

Note: All recipes except grape require pectin.

Fruit	*Approx. amount (cups)*	*Water (cups)*	*Approx. simmer time (min)*	*Juice (cups)*	*Sugar (cups)*	*Lemon*	*Yield (cups)*
Barberry/Apple	6/3	1	20	4	3	none	4
Blackberry, Dewberry	8	none	3	3½	5	none	6
Blueberry, all types	11	1	10	5½	7	¼ C	9
Buffaloberry	8	1	10	4	6	none	7
Cherry	7	½	10	3½	4½	none	5
Currant	14	1½	10	6	6	none	8
Grape (use no pectin)	8	½	10	4	4	none	6
Huckleberry	11	1	10	5½	7	¼ C	9
Indian Fig	6	2½	20	4	4	½ C	6
Mint	1½ *	3½	**	3	4	¾ C	5½
Mountain Ash	8	4	10	4	7	¼ C	7
Mulberry, Red	8	2	20	4	5½	none	6
Plum, American	8	1½	20	6	6	3 tbsp	8
Plum, Beach and Flatwood	8	1½	20	6	8	3 tbsp	8½
Prickly Pear	6	2½	20	4	4	½ C	6
Raspberry, Black	14	1	10	5	6½	none	7
Raspberry, Red; Wineberry; Thimbleberry	10	1	5	4	5½	none	6
Rose Hips	6	2	25	4	7	none	7
Salal	6	2	20	5	5	none	7
Serviceberry	6	4	20	4½	3	½ C	6
Thornapple	8	4	10	4	7	¼ C	7
Violet Blossom	4	†	‡	2	4	¼ C	4

*Chopped leaves **Quick boil; let stand 10 min. †3 C boiling water ‡Infuse 24 hr.

You can substitute liquid pectin or apple juice for powdered pectin, or you can boil the juice down until the jelly temperature reaches 105°C (220°F).

Measure the specified volume of juice in a large kettle that will allow the rolling boil required for jelly making. Stir in pectin and lemon juice, if indicated in the recipe. Bring to a full boil with constant stirring. Stir in sugar and bring to a rolling boil. Boil hard for one minute, remove from heat, and skim off foam. Immediately pour into hot, sterilized jars. When jelly is set, cover the surface with at least ⅛ inch of hot, melted paraffin. Remove any bubbles and make certain a complete seal is formed with the jar.

For the recipes in the accompanying jelly table, amounts of water and sugar will vary according to the individual batch of fruit, desired consistency of the jelly, and your taste preferences.

Jams. Jams are made from whole fruits, cooked to a thick syrupy consistency. Select fully ripe fruits, and prepare no more than 3–4 quarts of fruit in one batch.

Wash and drain fruit. Remove pits, if present. Seeds of species such as black-

Jam Recipes

Fruit	*Approx. amount*	*Special procedures*	*Fruit pulp (cups)*	*Sugar (cups)*	*Other*	*Yield (cups)*
Barberry	5 C	Crush in 1 C water; cook 20–30 min; sieve seeds	3	3¼	none	4
Blackberry	2 qt	Crush; sieve seeds; no water	5	7	none	7½
Blueberry	1½ qt	Crush thoroughly; no water	4	4	2 tbsp lemon	5½
Gooseberry	2½ qt	Remove ends; grind; no water	5½	7	none	8½
Ground Cherry	1 qt	Halve and crush; 1 cup water	4	4½	¼ C lemon	5½
Grouseberry	1½ qt	Crush thoroughly; no water	4	4	2 tbsp lemon	5½
Huckleberry	1½ qt	Crush thoroughly; no water	4	4	2 tbsp lemon	5½

Fruit	*Approx. amount*	*Special procedures*	*Fruit pulp (cups)*	*Sugar (cups)*	*Other*	*Yield (cups)*
Mayapple	2 qt	Crush in ½ C water; cook gently 20 min; colander	4	4	⅛ tsp salt	5½
Mulberry, Red	2 qt	Crush; ½ C water	4	3	none	5
Persimmon	2 qt	Halve; colander; no water	4	3	none	5
Plum, Amer.	4 qt	Pit; chop fine; ½ C water	6	8	none	9½
Plum, Beach and Flatwood	4 qt	Pit; chop fine; ½ C water	6	10	none	10½
Raspberry, Wineberry, Thimbleberry (See Blackberry)						
Salal	1½ qt	Crush and sieve; no water	3¾	5	½ C lemon	6
Sea Grape	2 qt	Crush in 2 C water; cook 15 min; colander seeds	5	2	none	5
Serviceberry	1 qt	Chop with 2 oranges; grate orange peels; water to cover	4	3	5 tbsp lemon	5½
Strawberry	2 qt	Crush; sieve seeds; no water	5	7	none	7½

berries and grapes may be strained out during the process. Chop, grind, or crush the fruits as indicated in the accompanying jam recipe table. Put in a saucepan, add water (if called for), and simmer 8–10 minutes—longer if specified. Measure fruit pulp into a large saucepan and add pectin. Bring to a full boil with constant stirring. Stir in sugar and other ingredients. Boil hard for 1 minute. Skim off foam. Ladle the jam into hot, sterilized jars to within ¼ inch of top. Seal with paraffin or airtight lids. If you use canning jars with 2-piece lids, sterilize the filled, loosely capped jars in a canner after you fill them.

For jams without commercial pectin, boil prepared fruit pulp with an approximately equal volume of sugar for about 30 minutes. Pack and store as above.

Uncooked jam can easily be prepared from many fruits by grinding or blending 2 cups of fruit with 4 cups of sugar. Let stand 30 minutes. Bring to a boil one package of powdered pectin in 1 cup of water. Stir liquid into fruit and sugar. Pour into sterilized jars and cover. Leave at room temperature for one day before refrigerating or freezing. Use within 3 months.

Pies. Pie fillings can be as simple as fruit and sugar, but usually additional ingredients will enhance the quality. If the fruit is bland, add lemon juice; if dry, add 2 or 3 tbsp water. Vary the amounts of flour and sugar according to juiciness and sweetness. Spices may add zest. Experiment with the extras. The accompanying pie recipe table gives general, not absolute, quantities.

Nine-inch pies will require 3–4 cups of fruit unless other ingredients add significantly to the volume of the filling. For most pies, pour the fruit into the pie shell. Then mix the flour, sugar, lemon juice, water, and salt as called for and pour over the fruit. Top with spices and butter. Fruits such as huckleberries make better fillings if floured before going into the pie shell. Once again, experience or a detailed recipe book will help you arrive at the best procedure.

Bake pies at 190–205°C (375–400°F) for 40–50 minutes, unless you partially cook the fruit and other ingredients before putting them into the pie shell. For baking next to a campfire the temperature is right if you can hold your hand for

Pie Recipes

Fruit	*Amount (cups)*	*Sugar (cups)*	*Lemon juice*	*Flour*	*Other ingredients*
Blackberry	4	1	none	3 tbsp	2 tbsp butter
Blueberry	4	1	1½ tbsp	3 tbsp	½ tsp nutmeg, 2 tbsp butter, mix berries with flour
Currant	2	2	none	2 tbsp	2 beaten eggs

Fruit	*Amount (cups)*	*Sugar (cups)*	*Lemon juice*	*Flour*	*Other ingredients*
Elderberry	3½	1½	4 tbsp	2 tbsp	2 tbsp butter
Ground Cherry	3	¾	3 tbsp	none	½ tbsp ground cloves, 2 tbsp butter
Huckleberry	4	1	1½ tbsp	3 tbsp	½ tsp nutmeg, 2 tbsp butter, mix berries with flour
Japanese Knotweed	4	1½	none	4 tbsp	3 beaten eggs, ¾ tsp nutmeg
Mayapple	3 (halves)	¾	6 tbsp	none	½ tsp each, ground cloves and cinnamon
Mountain Ash	3	1¼	1 tbsp	2 tbsp	½ tsp cinnamon, 2 tbsp butter
Mulberry, Red	3	1	none	4 tbsp	2 tbsp butter
Pawpaw	6 large, sliced fruits	1	4 tbsp	none	1 tsp cinnamon, ½ tsp allspice, 2 tbsp butter
Plum	1½ sliced	1½	none	4 tbsp	1½ C sliced peaches, 2 tbsp butter
Raspberry, all	4	1	none	2 tbsp	½ tsp cinnamon, 2 tbsp butter
Serviceberry	4	¾	1½ tbsp	3 tbsp	½ tsp cinnamon or nutmeg, 2 tbsp butter
Spanish Bayonet	3 pulp	¾	3 tbsp	none	½ tsp cinnamon, ¼ tsp ground cloves or nutmeg, 2 tbsp butter
Strawberry	4	¾	none	4 tbsp	
Wineberry	4	1	none	2 tbsp	½ tsp cinnamon, 2 tbsp butter

not more than 8–10 seconds where the pie will be placed. In situations where heat may be uneven, rotate the baking pie a part of a turn every few minutes. On a campout few diners will complain even if the pie isn't cooked to perfection.

Grains, nuts, and seeds. These fruits, like the fleshy fruits, generally ripen late in the growing season. Grains, nuts, and seeds are usually high in proteins and oils, making them valuable wild foods. Oily seeds and nuts may be used in many ways: Chopped or sliced, they serve well in baked goods and preserves and with green vegetables. They can be eaten fresh or pressed for oil extracts. Most people have tried hickory nuts, walnuts, and—perhaps among westerners—pinyon nuts. But these few most popular nuts are just the beginnings of nature's grain, nut, and seed bounty.

If you forage close to civilization, abandoned and disturbed sites will provide seeds of common weeds such as lamb's quarters, amaranth, and purslane. Conduct experimental harvests after the seeds are fully formed but before they have fallen from the plants. Cut whole plants and dry them on a cloth. When dry, the seeds can be shaken out and separated from other plant parts. This is labor-intensive harvesting, but it can be fun.

With dozens of grasses, you can separate grains from their hulls by rubbing them between your hands or rolling or pounding them lightly with a rolling pin between two towels on a hard surface. Use either a natural breeze or an electric fan to blow the hulls away as you pour the mixture of hulls and grains from a height of several feet into a container. Here are two other winnowing methods: (1) Shake the seeds and chaff in a deep container or sack, causing the denser seeds to settle to the bottom while the hulls and other light debris rise to the top. (2) Also, you can shake the mixture in water, causing the seeds, which are heavier, to sink and other materials to float. Experimentation will yield the most effective process for the seeds you have at hand.

You can grind seeds into flour the old way with stones, but a modern blender or food mill makes the task much faster. Substitute your wild flour for some of the commercial flour in any recipe. You will soon concoct some favorite blends. Some ground-up seeds also make a pleasant boiled gruel.

Little need be said about the various nuts in addition to the comments you'll find on each species later in this book. But here are a few general tips: When collecting hickory nuts and walnuts, search out trees that produce large nuts because nut cracking is too much effort to make small nuts worthwhile. The husks of hickory nuts are easy to remove. Those of the walnut must be crushed and rubbed off. If you take the trouble to remove husks in the field, you'll eliminate a stain and mess at home. Drying the nuts for at least a month makes nutmeat extraction easier.

Plant foods of the Indians

As the big-game hunting phase of American Indian history evolved into the hunting/gathering phase, wild plants became an increasingly important part of the Indian diet. Inland tribes in particular could no longer sustain themselves almost wholly on game within tribal territory, and infringement on territory of competing tribes posed a more sudden threat to life than malnourishment did. The successful tribes learned to use the plants of their region in hundreds of ingenious ways. Overall, Indians used more than 250 species of wild fruits alone.

The first cultivated crop in the Southwest was probably sunflower. But well before that crop, several thousand years B.C., Mexican Indians began to cultivate upland vines and grasses that bore goodly quantities of edible fruits and seeds. From these efforts came many forms of squashes, pumpkins, and gourds and six species of Indian-cultivated corn with more than 150 varieties. Later, Indians domesticated beans that through the centuries became five species with many color varieties. Through sound agricultural practices, Indians developed bean varieties that thrived in the many climates from Mexico to Maine. In various parts of the country, Indians also began cultivating peanuts, white and sweet potatoes, chocolate, vanilla, sunflower, Jerusalem artichokes, and other plants.

Agriculture gave the Indians more reliable food sources and, therefore, more control over their lives. Yet many tribes continued to spend much of the year away from their summer settlements by taking advantage of a sophisticated system of wild food harvesting. For example, cultivated crops accounted for only about 25 percent of the Apaches' food.

Harvesting game and wild plants dictated a nomadic life with extensive knowledge of nature. The tribes learned to harvest crops such as berries and wild rice at very specific stages to minimize losses to wildlife and natural shedding.

In the dry Southwest, sparse vegetation resulted in intensive harvesting. And Indians naturally learned to favor high-calorie foods. Saguaro fruits provided the first succulent, sweet food of the summer for the Pimas and Papagos. In addition to high sugar content, the saguaro fruits offer high amounts of fat, protein, and vitamin C. Unfortunately, today as in early times, the dried fruits become infested with fly larvae and so can't be kept long. Thus, eating the fruits, like harvesting, is seasonal. Since harvesting does not destroy the saguaro and since plenty of seeds are left behind for reproduction, intensive use over the centuries has had little effect on the saguaro population.

Other desert plants including prickly pear, agave, Spanish bayonet, and the mesquite and screw beans provided important fleshy foods. A few plants such as pigweed, purslane, and tumbleweed provided greens. To the Sonoran desert Indians, mesquite and screw beans were nutritionally more important than corn. Generally, the edible seeds used by the desert tribes came from more than 50 species of grasses. Gathering them was extremely tedious. Several hours of work often provided only a cupful of seeds that served more as a flavoring than as a nutritional ingredient.

In arid regions, the Indians employed as little water as possible when preparing foods. Many foods were prepared from dried ingredients which were soaked and then baked or roasted. Early Indians boiled foods by placing hot stones in water-filled watertight baskets or hides. Later, ceramic pots came into use. Because of the mealy nature and bland taste of the major edibles in the Southwest, Indians supplemented the flavor with herbs, salt, sugar, onions, and chili.

In less arid regions, Indians made extensive use of hundreds of plants. Roots, shoots, and leafy foods were often cooked with oils from animals, nuts, acorns, and sunflowers. Most inland tribes lacked access to salt but employed flavor accents such as mints, bearberry, spicebush, onions, leeks, and partridgeberry. Sweets came from maple sugar and many fresh and dried fruits.

To help ensure good harvests, Indians gave special care to many native plants in their natural settings. These included rice, water lilies, blackberries, cherries, currants, mulberries, oaks, hickories, walnuts, beeches, and chestnuts.

By the time the Pilgrims landed, all northeastern tribes except those of eastern Maine employed agriculture. This involved burning and clearing land, followed by the planting of native and introduced crops. Native plants included strawberries, grapes, plums, Jerusalem artichokes, and milkweed. Large fields were planted with the "three sisters"—corn, beans, and squashes (including pumpkins) introduced from the Southwest.

For some northern tribes rice was the staple food, often served with maple sugar. Family groups staked out fields for harvesting one or two weeks before the grains ripened. They harvested most of the crop by boat or canoe using two sticks—one to bend the stalks over the canoe, the other to sweep the grains off the stalk. Skilled harvesters made only the ripe grains fall. The complete crop was taken in several harvests a few days apart. Yet Indians took care to leave enough grains to produce the next year's crop. A canoe full of grain was considered a good day's work. After gathering came a complex process of drying, parching, pounding, winnowing, and a final treading to remove the last bits of the hulls before storage in bags.

In the Northeast, maple sugaring was a family project. Each family had its own reserved woodland dominated by sugar maples (sugar bush) with a sugaring lodge and a smaller storage lodge. An average camp with 900 taps in trees would obtain several hundred pounds of sugar. Sugaring was especially difficult before the introduction of iron kettles. Till then the Indians boiled sap by immersing hot stones into hide or bark containers and then reheated the stones as they cooled. After three to four weeks of maple sugaring, the family returned to their village with all the sugar they could carry. They stored the rest underground.

Today, we still use many Indian food processing techniques. For example, the Indians parched or thoroughly sun-dried grains and then stored them until grinding them just before use. Stone grinding, touted by commercial bakeries today, was the only method available to the Indians. Beans, roots, and fruits were sun-dried. For an energy rich trail food, Indians made pemmican, consisting of dried berries, tallow, and meat. This combination was usually stuffed into a cleaned intestine of a large game animal and hung to dry. The Indians also gave us techniques of barbecuing, pit steaming, and baking, as well as dishes such as succotash, bannock, and hominy.

Colonists introduced many food plants to North America. Many came by intent. Others were unwittingly introduced in the soil ballast of ships. And many of these foreign plants established themselves quickly. Today's Pimas and Papagos of Arizona may include in their diets lamb's quarters, curly dock, and dandelion introduced in the East by the colonists.

Nutrition

Media advertisements give the impression that highly refined, vitamin enriched foods are the most nutritious. Vitamins are added to compensate for those lost in processing. Next come more additives to extend shelf life. Only the food companies know how it's all prepared. Certainly food additives have some merit in today's hurry-up lifestyle, but it's doubtful that they improve on nature.

Wild foods aren't advertised since they're free for the picking. Yet wild plants are just as tasty and nutritious as store-bought produce; often more so since they're fresh and unadulterated. Nutrition, the process of digestion and utilization of foods, is the same regardless of whether the food is wild or cultivated. The more we know about nutrition, the more assurance we have for balancing our diet.

You can think of food as body fuel. The energy produced by food is measured in calories. An adult male requires about 18 calories a day, per pound of body weight. Women require somewhat less; children and adolescents, somewhat more. Strenuous exercise can increase requirements upwards of 50 percent, and exposure to cold also increases caloric needs. Thus a 175-pound man might burn a little over 3,000 calories during a 24-hour period of normal activities and burn over 5,000 calories climbing a mountain in winter.

It's impossible for us to eat exactly the right amount of food at the rate we need it, so we all carry reserves of food energy in our bodies. The most readily available energy is stored in muscles and liver in the form of animal starch, or glycogen. When energy is needed, this starch can be quickly converted to sugar. Fat, the body's secondary energy reserve, takes longer to break down into a useable source of energy. But the fat reservoir is larger. Even slim people carry enough fat to sustain life for two weeks without additional food. Fat is the most concentrated form of food energy and contains about 3,500 calories per pound.

What we eat is just as important as how much. All food contains three primary nutritional building blocks in addition to water. These are protein, fats, and carbohydrates. The enzyme action of our digestive system requires a combination of these three substances in order to function properly. Thus a knowledge of them will help you practice good nutrition and supplement your diet with wild plants in the most beneficial way.

1. ***Protein.*** This is the only substance that can supply the amino acids necessary for building and maintaining body tissue. Protein not needed for tissue is converted to sugar and used for energy, or further converted to fat and stored. Protein in its original form cannot be stored by the body, and neither sugar nor fat can be reconverted to protein. Thus our bodies need a constant protein supply. Nutritionists recommend that 15 to 20 percent of caloric intake come from protein.

Protein also promotes the oxidation (burning) of fats and sugars for energy. An adult needs roughly a half gram of protein a day per pound of body weight; more for growing youngsters and people who are sweating heavily. Most Americans eat more protein than they need.

Although all foods contain some protein, the most significant amounts are provided by animal sources such as meat, milk, eggs, fish, and poultry. Protein from these sources is complete and contains all amino acids necessary in the diet. This is not true of vegetable protein. Each plant species is unique in the amounts and types of amino acids it provides, and no single species supplies all of the essential ones. Incomplete proteins from two or more species, however, may be combined to form a complete protein. Beans and corn are complementary in this respect, and the combination is a popular one when cooked with tomatoes. Plant and animal protein combinations also complement nicely. In this case it doesn't take much meat to satisfy human requirements. Orientals have long used poultry and fish primarily as a condiment, combined with vegetables to provide complete and adequate protein.

2. ***Fats.*** These should provide another 15 to 20 percent of your caloric intake. Eggs, nuts, milk, cheese, butter, and oils are important sources of fats. Most people feel that fatty foods "stick to your ribs" longer, which is just another way of saying that it takes longer to digest them. This is also true of protein, so the traditional breakfast of bacon and eggs, heavy on fats and protein, sticks to the ribs. A normal diet containing meat and milk usually provides a full quota of fat.

3. ***Carbohydrates.*** About 65 percent of your caloric intake should come from carbohydrates. And wild plants are primarily valuable for their carbohydrates. For dietary purposes, carbohydrates fall into two groups: starches and sugars. Foods high in carbohydrates include grains, vegetables, and fruit. Of these, vegetables are especially important because they are a secondary source of protein. The traditional breakfast of bacon and eggs usually includes toast and jelly, the sugars from both providing a quick shot of energy. When that's used up, the fats and proteins will be ready to provide additional fuel. Since carbohydrates are digested faster than fats and proteins and are then burned up quickly, don't be surprised if snacks of cookies, crackers, and candy don't satisfy your hunger very long. Nuts would do a better job since they also contain significant amounts of protein.

Vitamins and minerals. Vitamins and minerals in small amounts are essential for good nutrition. Nature has provided amply in the foods we eat, and if we eat the right foods, the vitamins and minerals will be included. The table, below, shows recommended adult daily requirements. Some people (including female adolescents and pregnant and lactating women) require somewhat more.

Recommended Dietary Allowances

	MINERALS			*VITAMINS*				
	Calcium (mg)	*Phosphorus (mg)*	*Iron (mg)*	*Vitamin A (IU)*	*Thiamine (mg)*	*Riboflavin (mg)*	*Niacin (mg)*	*Ascorbic acid (mg)*
Male	800	800	10	5,000	1.4	1.6	18	45
Female	800	800	18	4,000	1.0	1.2	13	45

Source: *Recommended Dietary Allowances,* National Academy of Sciences, Washington, DC

A balanced diet. Vital over the long term, a balanced diet is also important over short periods. Insufficient protein intake, if extended over several days, will cause fatigue, even if you consume enough carbohydrates. In addition, a disproportionate amount of fats in relation to carbohydrates usually results in an upset stomach. This is nature's way of acknowledging poor eating habits. Achieving proper balance, however, is easy if you remember this simple proportion: Measured in calories, food consumption should total about 1/5 protein, 1/5 fats, and 3/5 carbohydrates.

If you want to fine tune your diet, the Appendix table on page 274 should help. The table was extracted from the Department of Agriculture's *Handbook of the Nutritional Contents of Foods.* Though with only a limited listing of wild edibles, it is the most comprehensive and reliable reference available. The table can help you balance your diet, because wild vegetables, fruits, and nuts are shown together with supermarket foods such as beef, milk, and eggs. Food values for wild plants not included are often approximately the same as those of similar cultivated plants. For example, fiddleheads would approximate asparagus in food value.

In addition to the energy content of foods, we've also included values for vitamins and minerals. Foods vary greatly in their nutritional content, and we've included this information to dramatize the need for a balanced diet.

Preparation methods. Nutritional values of fruits and vegetables depend greatly on preparation. Fresh green leaves or shoots that must be cooked in two changes of water to make them palatable lose much nutritional value compared to those that can be eaten fresh. Also, the longer fresh vegetables are cooked, the greater the loss of vitamins and minerals to the water. Proteins and carbohydrates are less affected by cooking. So for the sake of nutrition, the preferred order of preparation options is raw, quick-cook or steam, bake, boil.

Other plant guides

Since it is impossible to distinguish among all 20,000 North American species in this or any other book, you may also want to refer to identification guides devoted to your own region or state. There are many wildflower guides available, yet even these treat only the most common species.

Here's a sampling of good regional plant guides:

- *The New Britton and Brown Illustrated Flora of the Northeastern U.S. and Adjacent Canada* by H. A. Gleason, third printing, 1963.
- *Manual of Vascular Plants of Northeastern U.S. and Adjacent Canada* by H. Gleason and A. Cronquist, 1963.
- *Manual of Vascular Flora of the Carolinas* by A. Radford, H. Ahles, and C. Bell, 1964.
- *Flora of the Pacific Northwest* by C. Hitchcock and A. Cronquist, 1976.
- *A California Flora and Supplement* by P. Munz, 1968, 1973.
- *Manual of the Vascular Plants of Texas* by D. Correll and M. Johnston, 1970.
- *Rocky Mountain Flora* by W. Weber, 1976.

Good less-technical guides include these:

- *Newcomb's Wildflower Guide* by L. Newcomb, 1977.
- *The Audubon Society Field Guide to North American Wildflowers (Eastern Region* by W. Niering, 1979; *Western Region* by R. Spellenberg, 1979).

For North American trees, we recommend *The Complete Trees of North America* by T. Elias, 1980.

Key to edible-use symbols

The symbols below appear in the "Seasonal Key to Plants" that follows and later in page margins of texts on each plant. After you've read the explanations below, you'll probably find the symbols distinctive enough to make repeated reference to this page unnecessary. You can exploit the symbols in either of two ways based on (1) your general interest in a plant's range of edible uses and (2) your desire to locate plants with specific edible uses.

Trail nibble: plant part eaten as picked along the trail

Salad: used raw in tossed salad (seldom steamed)

Cooked "green" vegetable: boiled or steamed greens, shoots, buds, flowers, or young fruits used as a vegetable

Underground vegetable: roots, tubers, corms, bulbs, rhizomes prepared as vegetables, usually by boiling or roasting

Fritter: plant part dipped in batter and deep fried

Raw fruit: fruits eaten uncooked, though perhaps with toppings

Cooked fruit: true fruits or nonfruit parts (such as rhubarb) used in pies, baked goods, and other dessert recipes

Jams, jellies, sauces: made from fruits

Syrup and sugar: such as made from maple sap

Candy: plant parts or substances that may or may not contain their own sugar that are boiled in sugar syrup to produce a confection

Grains: for cereal and for cooking whole, such as rice

Nuts, nutlets, large seeds: used whole and for oil extract

Flour or meal: grains, nuts, other seeds, or starchy roots from which flour or meal can be ground

Hot beverage: from tea or coffee-like preparations

Cold beverage: fruit juices or originally hot tea or coffee-like beverage that has cooled

Pickle: from pickling recipes

Seasoning: flavoring from ground, grated, or cooked-whole parts

Thickener: used like cornstarch to thicken soups

Caution: minor reaction may result from eating wrong part, or too much of right part, or eating at wrong stage

Poisonous: severe reaction results from eating wrong part, or eating at wrong stage

Seasonal key to plants

Spring

FERNS

Page / Plant	Region	Habitat	Edible Uses
58 Ostrich fern *Matteucia struthiopteris*	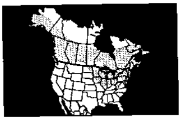	swamps, bottomlands, stream banks, ponds	🚶 ✕ 🍲

HERBS: MONOCOTS

Flowers many in dense heads

Page / Plant	Region	Habitat	Edible Uses
58 Wild onion *Allium cernuum*	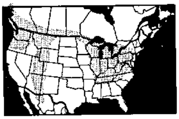	dry woodlands, ridges, rocky slopes, prairies	🌿 🧂 ✕ 🍲 🥒
60 Wild garlic *Allium canadense*	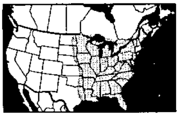	open woodlands, fields, prairies	🌿 🧂 ✕ 🍲 🥒
60 Wild leek *Allium tricoccum*	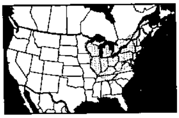	moist woodlands	🌿 🧂 ✕ 🍲 🥒
62 Corn lily *Clintonia borealis*	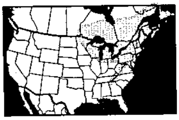	rich woodlands, thickets	🍲 ✕

Flowers single, leaves threadlike

Page / Plant	Region	Habitat	Edible Uses
62 Asparagus *Asparagus officinalis*	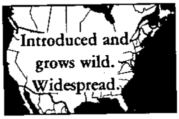	roadsides, disturbed sites	🍲 🚦

Page/Plant	Region	Habitat	Edible Uses
	Flowers in elongate clusters		
64 Eastern camass *Camassia scilloides*	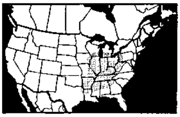	wet meadows, fields, moist open woods	🥕 🍴
	Vines		
64 Common greenbrier *Smilax rotundifolia*	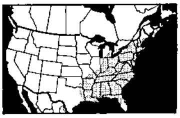	roadsides, disturbed woodlands	🚶 🍴 🪴
66 Bullbrier *Smilax bona-nox*	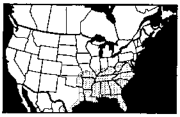	roadsides, fields, open woodlands	🚶 ☕ 🥛 🍴 🥗 🪴 ▣ 🫙
66 Carrion flower *Smilax herbacea*	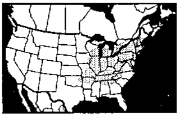	roadsides, fields, margins of woodlands	🚶 ☕ 🥛 🍴 🥗 🪴 ▣
	Large semiaquatic		
68 Cattail *Typha latifolia*	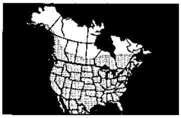	ditches, margins of ponds and lakes	🚶 🥕 🍴 🪴 ▣
68 Narrow-leaved cattail *Typha angustifolia*	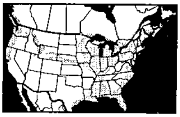	ditches, margins of ponds and lakes	🚶 🥕 🍴 🪴 ▣
	Small grasslike		
70 Hard-stem bulrush *Scirpus acutus*	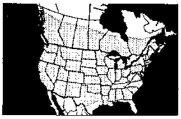	edges of ponds, lakes, streams and marshes	🚶 🥕 🍴 🪴 ▣ 🫖
70 Great bulrush *Scirpus validus*	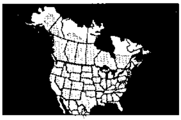	edges of ponds, streams, marshes	🚶 🥕 🍴 🪴 ▣ 🫖

HERBS: DICOTS

Flowers small, inconspicuous

Page/Plant	Region	Habitat	Edible Uses
72 Pigweed *Amaranthus retroflexus*	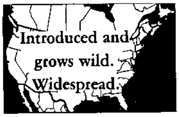 Introduced and grows wild. Widespread.	waste areas, roadsides, abandoned land	🍴 🍲 🫙
72 Green pigweed *Amaranthus hybridus*	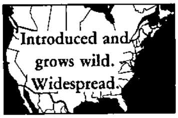 Introduced and grows wild. Widespread.	fields, roadsides, disturbed sites	🍴 🍲 🫙
74 Lamb's quarters *Chenopodium album*	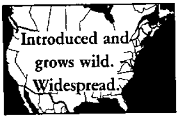 Introduced and grows wild. Widespread.	vacant lots, roadsides, cultivated land	🍴 🍲 🫙
74 Poverty weed *Menolepis nuttalliana*	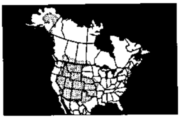	dry salty or alkaline soils	🍲 🫙

Flowers tiny, stems & leaves fleshy

Page/Plant	Region	Habitat	Edible Uses
76 Glasswort *Salicornia europaea*	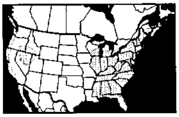	salt marshes, sea coast, alkaline soils	🚶 🍴 🍲 🥒
76 Russian thistle *Salsola kali*	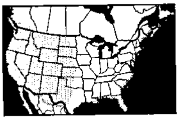	sandy beaches, dry alkaline soils	🍲 ☠

Stinging hairs on stem

Page/Plant	Region	Habitat	Edible Uses
78 Stinging nettle *Urtica dioica*	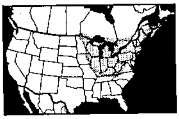	roadsides, trailsides, stream banks	☕ 🍲 🚦

Page / Plant	Region	Habitat	Edible Uses
	Flowers tiny, leaves basal		
78 Plantain *Plantago major*	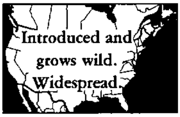 Introduced and grows wild. Widespread.	yards, roadsides, disturbed sites	☕ 🍴 🍲 🫙
	Flowers white		
80 Shepherd's purse *Capsella bursa-pastoris*	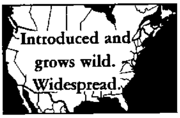 Introduced and grows wild. Widespread.	yards, vacant lots, roadsides	🧂 🍴 🍲
80 Toothwort *Dentaria laciniata*	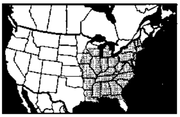	moist woodlands, ravines, floodplains	⛏ 🧂
82 Peppergrass *Lepidium virginicum*	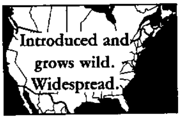 Introduced and grows wild. Widespread.	fields, vacant lots, roadsides	🧂 🍴 🍲
82 Field cress *Lepidium campestre*	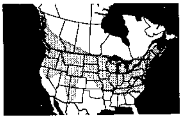	fields, vacant lots, roadsides, waste areas	🧂 🍴 🍲
82 Field pennycress *Thlaspi arvense*	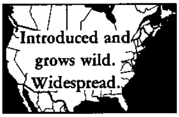 Introduced and grows wild. Widespread.	fields, vacant lots, disturbed soils	🍴 🍲
84 Garlic mustard *Alliaria officinalis*	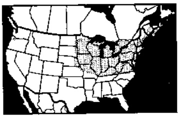	shady woodlands, trailsides	🧂 🍴 🍲
84 Common chickweed *Stellaria media*	Introduced and grows wild; widespread in East and Midwest	yards, moist soils, woodlands, waste areas	🍴 🍲

Page/Plant	Region	Habitat	Edible Uses
84 James chickweed *Stellaria jamesiana*	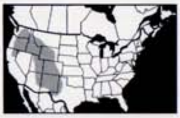	mountain meadows, seepage areas, woodlands	✕ 🍲
86 Cleavers *Galium aparine*	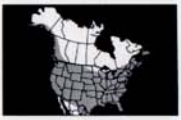	rich woodlands, fields, edges of streams and lakes	☕ ✕ 🍲
86 Northern bedstraw *Galium boreale*	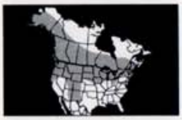	stream banks, fields, open woodlands	☕ ✕ 🍲
88 Bugleweed *Lycopus uniflorus*	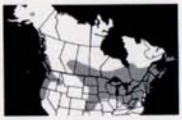	wet or boggy soils, wet woodlands	🥕 🧂 ✕ 🥒
88 Edible valerian *Valeriana edulis*	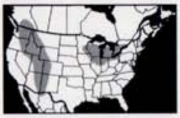	ditches, moist meadows, prairies, swamps	🥕 🍞
90 Wild strawberry *Fragaria virginiana*	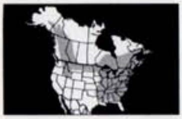	open woodlands, fields, margins of woodlands	🏃 🍎 ☕ 🥤 🥧 🫙
90 Spring beauty *Claytonia virginica*	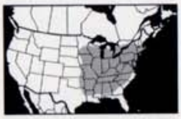	rich woodlands, wet fields, wooded floodplains	🏃 🥕 ✕ 🍲
92 Alpine spring beauty *Claytonia megarhiza*	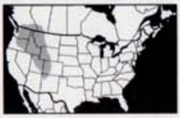	tundra, rocky slopes	🏃 🥕 ✕ 🍲
92 Western spring beauty *Claytonia lanceolata*	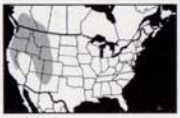	high mountain slopes, moist woodlands	🏃 🥕 ✕ 🍲

Page / Plant	Region	Habitat	Edible Uses
94 Miner's lettuce *Montia perfoliata*	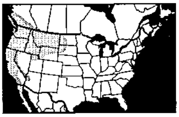	lower mountain slopes, moist sites, springs	🚶 🍴 🍲
94 Canada violet *Viola canadensis*	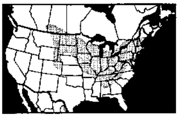	bluffs, deciduous woodlands, rich soils	🚶 ☕ 🍴 🍲 🫙
96 Pale violet *Viola striata*	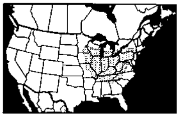	wooded floodplains, stream banks, meadows	🚶 ☕ 🍴 🍲 🫙
96 Pokeweed *Phytolacca americana*	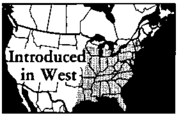	disturbed soils, roadsides, woodland margins	🍲 🥒 ☠
98 Japanese knotweed *Polygonum cuspidatum*	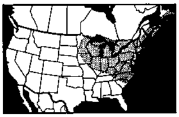	disturbed sites, vacant lots, roadsides	🍲 🥧 🫙

Flowers yellow

Page / Plant	Region	Habitat	Edible Uses
98 Wintercress *Barbarea vulgaris*	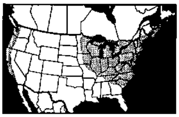	fields, wet meadows, ditches, roadsides	🍴 🍲
100 Black mustard *Brassica nigra*	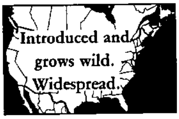	fields, yards, disturbed sites	🧂 🍴 🍲
100 Marsh marigold *Caltha palustris*	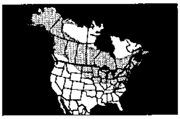	edges of streams, ponds, freshwater swamps	🍲 🥒 ☠

Page/Plant	Region	Habitat	Edible Uses
102 Evening primrose *Oenothera biennis*	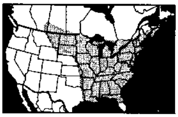	prairies, fields, roadsides	🍃 🍴 🥘 🍬
102 Sow thistle *Sonchus arvensis*	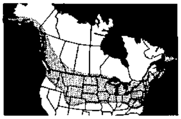	roadsides, fields, rights-of-way	🍴 🥘 ☠
104 Prickly lettuce *Lactuca serriola*	Introduced and grows wild. Widespread.	fields, vacant lots, disturbed sites	🍴 🥘 🚦
104 Dandelion *Taraxacum officinale*	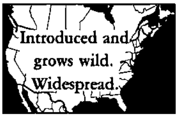 Introduced and grows wild. Widespread.	lawns, yards, disturbed sites	🍃 ☕ 🍴 🥘 🍳
106 Goat's beard *Tragopogon pratensis*	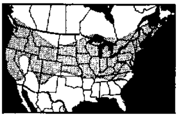	roadsides, fields, disturbed sites	🍃 🍴 🥘
	Flowers pink, reddish, or purple		
106 Common milkweed *Asclepias syriaca*	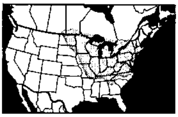	fields, meadows, margins of woodlands, roadsides	🍲 🥘 🚦
108 Fireweed *Epilobium angustifolium*	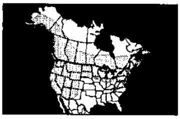	ravines, clearings, woodlands	☕ 🍴 🥘
108 Alfilaria *Erodium cicutarium*	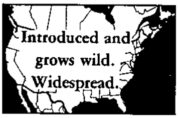 Introduced and grows wild. Widespread.	fields, pastures, roadsides, disturbed sites	🍴 🥘
110 Live-forever *Sedum purpureum*	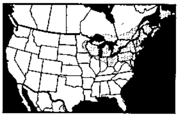	roadsides, disturbed sites, open woodlands	🍃 🍴 🥘 🥒

Page / Plant	Region	Habitat	Edible Uses
110 Lady's thumb *Polygonum persicaria*	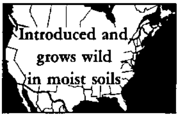 Introduced and grows wild in moist soils	low wet sites, roadsides, disturbed sites	🍴 🍲 🚦
112 Common burdock *Arctium minus*	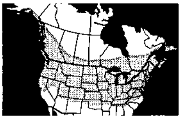	vacant lots, roadsides, disturbed sites	🏃 🌿 🍴 🍲 🍬
	Flowers purple		
112 Salsify *Tragopogon porrifolius*	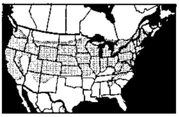	roadsides, fields, rights-of-way	🌿 🍴 🍲
114 Bull thistle *Cirsium vulgare*	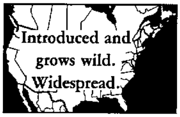 Introduced and grows wild. Widespread.	fields, meadows, roadsides, disturbed sites	🌿 🍴 🍲 🥫
114 Chicory *Cichorium intybus*	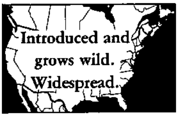 Introduced and grows wild. Widespread.	roadsides, fields, vacant lots	🌿 ☕ 🍴 🍲
116 Henbit *Lamium amplexicaule*	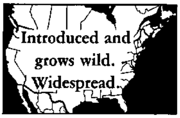 Introduced and grows wild. Widespread.	fields, yards, vacant lots, disturbed sites	🍴 🍲
116 Wood violet *Viola palmata*	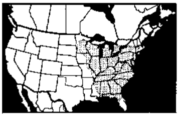	wooded floodplains, rich forests, slopes	🏃 ☕ 🍴 🍲 🫙 🫗
	Flowers tiny, often greenish		
118 Orange-flowered jewelweed *Impatiens biflora*	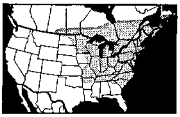	along streams, ponds, springs, wet areas	🧂 🍲 🚦

Page/Plant	Region	Habitat	Edible Uses
118 Yellow-flowered jewelweed *Impatiens pallida*	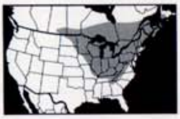	wet meadows, moist woodlands	🧂 🍲 🚦
	Flowers orange to yellow		
120 Sheep sorrel *Rumex acetosella*	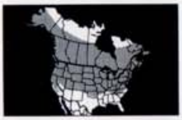	old fields, gardens, disturbed sites	🚶 ☕ ✕ 🍲 🚦
120 Sour dock *Rumex crispus*	Introduced and grows wild. Widespread.	fields, roadsides, vacant lots, disturbed sites	🍲 ▣ 🚦
TREES: CONIFERS			
122 Eastern hemlock *Tsuga canadensis*	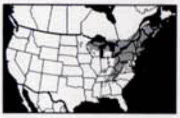	cool moist ravines, protected valleys	☕ ▣
TREES: DECIDUOUS			
122 Yellow birch *Betula alleghaniensis*	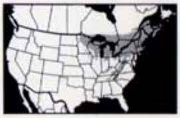	rich woodlands, wet soils	☕ 🥤 ▣ 🫗
122 Sweet birch *Betula lenta*	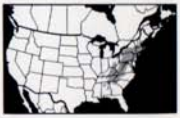	open woodlands, rich soils	☕ 🥤 ▣ 🫗
124 Sassafras *Sassafras albidum*	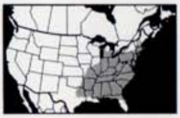	bottomlands, open woodlands	🧂 ☕ 🥤 ✕ 🥣 🫙
124 Sugar maple *Acer saccharum*	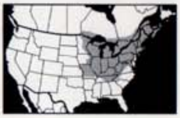	eastern deciduous forests	🍎 🥤 🫗

Summer

Page/Plant	Region	Habitat	Edible Uses
HERBS: MONOCOTS			
	Large grasses		
128 Reed grass *Phragmites communis*	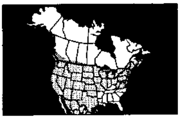	edges of streams, lakes, ponds, marshes, ditches	
128 Wild rice *Zizania aquatica*	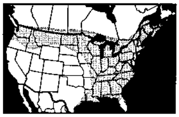	marshes, shallow ponds, lakes, streams, bays	
	Small grasslike		
130 Chufa *Cyperus esculentus*	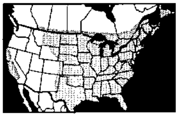	ditches, banks of ponds and streams, low wet soils	
	Semiaquatic, flowers blue		
130 Pickerelweed *Pontederia cordata*	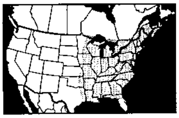	edges of lakes, ponds, streams, bays	
132 Lance-leaved pickerelweed *Pontederia lanceolata*	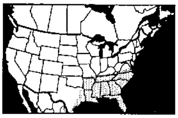	shallow water of lakes, ponds, streams	
	Leaves large, swordlike		
132 Sweet flag *Acorus calamus*	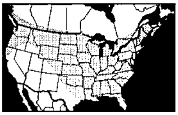	ditches, ponds, swamps, lakes and streams	

Page/Plant	Region	Habitat	Edible Uses
	Flowers white, showy		
134 Spanish bayonet *Yucca baccata*	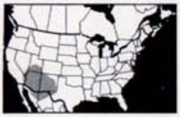	dry slopes, sandy soils	🍎 🌰 ☕ 🥤 🍴 🍳 🥧
134 Soapweed *Yucca glauca*	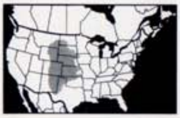	prairies, roadsides, pastures	🍴 🍳 🥧
	Flowers orange, showy		
136 Common day lily *Hemerocallis fulva*	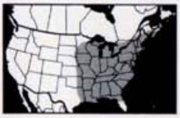	vacant lots, abandoned home sites, roadsides	🥔 🍴 🍲 🍳 🍡
	Flowers blue		
136 Western camass *Camassia quamash*	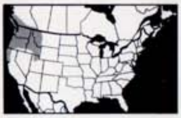	wet mountain meadows	🥔 🍴
	Flowers greenish-yellow		
138 Indian cucumber *Medeola virginiana*	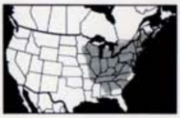	moist woodlands, bogs, swamps	🌱 🍴 🥒
HERBS: DICOTS			
	Stems large, fleshy		
138 Indian fig *Opuntia humifusa*	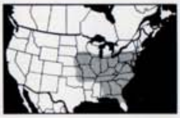	rocky bluffs, sand dunes, dry grasslands	🍎 🍲 🍳 🥫 🥧 🍯
140 Prickly pear *Opuntia phaeacantha*	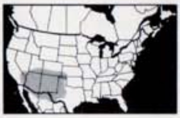	canyons, valleys, hills, rocky and sandy soils	🍎 🍲 🍳 🥫 🥧 🍯

Page/Plant	Region	Habitat	Edible Uses
	Flowers white		
140 May apple *Podophyllum peltatum*	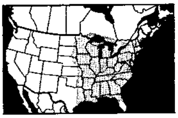	rich woodlands, shaded roadsides, meadows	🍎 🥤 🥧 🫙 ☠
142 Bunchberry *Cornus canadensis*	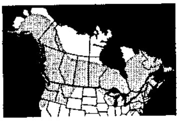	bogs, moist woodlands, mountains	🏃 🍎 🥧
	Stems 4-sided, flowers white		
142 Peppermint *Mentha piperita*	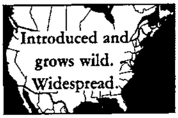 Introduced and grows wild. Widespread.	ditches, along stream banks, wet meadows	🧂 ☕ 🍬 🫙
144 Spearmint *Mentha spicata*	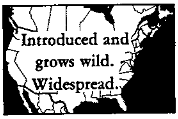 Introduced and grows wild. Widespread.	wet areas, ditches, stream banks	🧂 ☕ 🍬 🫙
	Flowers violet to purple		
144 American brooklime *Veronica americana*	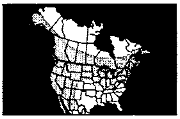	edges of streams, swamps, springs, ponds	🍴 🍲
	Flowers pink		
146 Marsh mallow *Althaea officinalis*	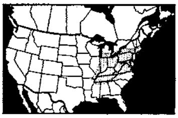	edges of salt marshes	🌿 🥣 🍲 🥒 🍬
	Flowers white		
146 Watercress *Nasturtium officinale*	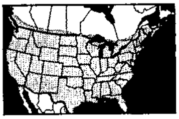	springs, cool clear streams	🧂 🍴 🍲

Page/Plant	Region	Habitat	Edible Uses
	Flowers yellow		
148 Ground cherry *Physalis pubescens*	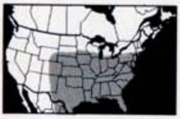	old fields, open woodlands, moist sites	🚶 🍎 🥧 🫙 ☠
	Flowers brownish-purple		
148 Groundnut *Apios americana*	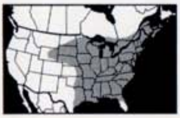	thickets, bottomlands, moist woodlands	🌰 🍂
	Flowers blue		
150 Common breadroot *Psoralea esculenta*	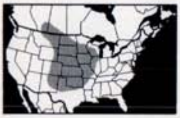	prairies, plains, dry rocky woodlands	🍂 ✕ ▣ 🚦
	Low creeping plants		
150 Wintergreen *Gaultheria procumbens*	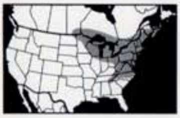	rich woodlands, clearings, poor acid soils	🚶 🍎 ☕ 🥧 🫙
152 Creeping snowberry *Gaultheria hispidula*	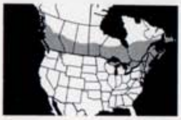	subalpine and alpine, northern coniferous forests	🚶 🍎 ☕ 🍲 🥧 🫙
	Leaves sheathing stems		
152 Mountain sorrel *Oxyria digyna*	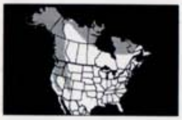	high mountains, subalpine and alpine fields	🚶 ☕ ✕ 🍲
154 American bistort *Polygonum bistortoides*	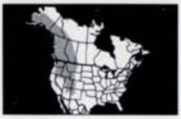	along streams, wet meadows, subalpine meadows	🍂 ✕ 🍲 🚦

Page/Plant	Region	Habitat	Edible Uses
154 Alpine bistort *Polygonum viviparum*	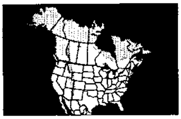	arctic regions, alpine and subalpine slopes	🏃 🍃 🍴 🍲 🚦

Flowers yellow

Page/Plant	Region	Habitat	Edible Uses
156 Purslane *Portulaca oleracea*	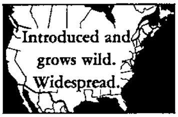 Introduced and grows wild. Widespread.	fields, vacant lots, waste sites, gardens	🍴 🥗 🍲 xxx 🥒
156 Sunflower *Helianthus annus*	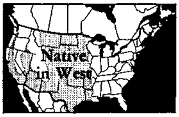 Native in West	prairies, plains, roadsides, disturbed sites	🌰 ☕ xxx
158 Quickweed *Galinsoga parviflora*	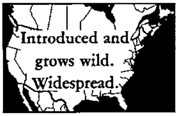 Introduced and grows wild. Widespread.	roadsides, yards, gardens, waste areas	🍲

SHRUBS: DICOTS

Leaves evergreen

Page/Plant	Region	Habitat	Edible Uses
158 Bearberry *Arctostaphylos uva-ursi*	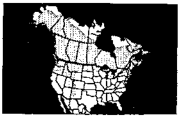	higher mountains, subarctic, circumboreal	☕ 🥧
160 Salal *Gaultheria shallon*	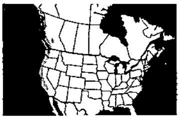	woodlands, clearings	🏃 🍎 🥤 🥧 🫙
160 Creeping wintergreen *Gaultheria humifusa*	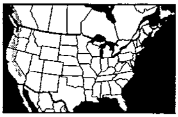	subalpine and alpine	🥤 🥧 🫙
160 Labrador tea *Ledum groenlandicum*	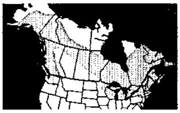	bogs, fens, heaths, northern woods	☕ 🥤 ☠

Page/Plant	Region	Habitat	Edible Uses
162 Mountain cranberry *Vaccinium vitis-idaea*	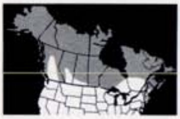	subarctic, alpine, bogs, cold seepage areas	🚶 🍎 🥤 ▣ 🥧 🫙
162 Cranberry *Vaccinium macrocarpon*	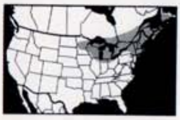	bogs, swamps, shores of cold acid lakes	🥤 ▣ 🥧 🫙
	Leaves deciduous		
164 Bog bilberry *Vaccinium uliginosum*	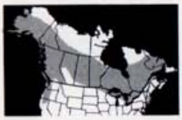	arctic, bogs, rocky outcrops	🚶 🍎 🥧 🫙
164 Highbush blueberry *Vaccinium corymbosum*	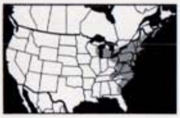	bogs, low wet woodlands, swamps	🚶 🍎 🥧 🫙
164 Late sweet blueberry *Vaccinium angustifolium*	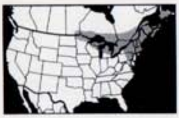	mountain slopes, barrens, dry rocky soils	🚶 🍎 🥧 🫙
166 Black huckleberry *Gaylussacia baccata*	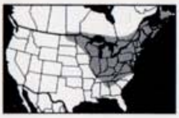	woodlands, clearings	🚶 🍎 🥧 🫙
166 Beach plum *Prunus maritima*	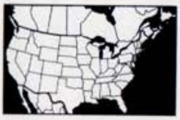	coastal areas, sandy soils	🚶 🥧 🫙
168 Golden currant *Ribes aureum*	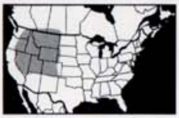	along streams, ravines, washes, slopes	🚶 🍎 🥧 🫙

Page/Plant	Region	Habitat	Edible Uses
170 Wax currant *Ribes cereum*	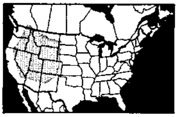	canyons, dry ravines, prairies, hillsides	🚶 🥧 🫙
170 Subalpine prickly currant *Ribes montigenum*	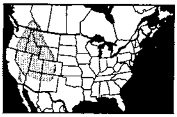	along streams, ravines, washes, forests	🚶 🥧 🫙
170 Fragrant golden currant *Ribes odoratum*	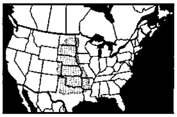	prairies, open woodlands, bluffs, ravines	🚶 🍎 🥧 🫙
172 American hazelnut *Corylus americana*	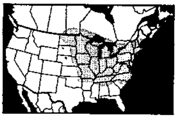	thickets, woodlands	🚶 🌰 XXX 🍬
172 Beaked hazelnut *Corylus cornuta*	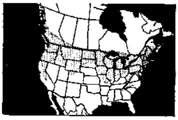	woodlands, hills, slopes	🚶 🌰 XXX 🍬
174 Sweet fern *Comptonia peregrina*	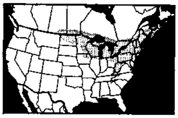	pastures, old fields, clearings	🚶 🌰 ☕

Leaves fragrant when rubbed

174 Southern bayberry *Myrica cerifera*	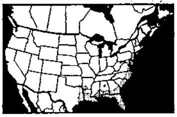	coastal plains, swamps, edges of ponds, barrens	🧂
176 Sweet gale *Myrica gale*	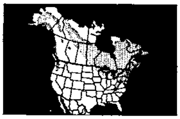	swamps, along ponds and streams	🧂 ☕

Page/Plant	Region	Habitat	Edible Uses
	Flowers white		
176 New Jersey tea *Ceanothus americanus*	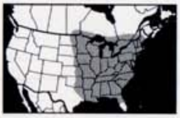	open woodlands, prairies	☕
	Leaves with silvery scales		
178 Silver buffaloberry *Shepherdia argentea*	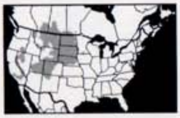	canyons, plains, meadows, river banks	🏃 🍎 ☕ 🥤 🥧 🫙
	Leaves usually compound		
180 Red raspberry *Rubus idaeus* var. *strigosus*	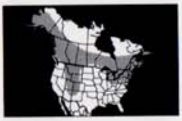	clearings, margins of woodlands, old fields	🏃 🍎 ☕ 🥧 🫙
180 Wineberry *Rubus phoenicolasius*	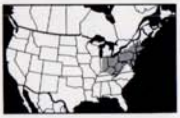	along rivers, streams, roads	🏃 🍎 ☕ 🥧 🫙
180 Thimbleberry *Rubus parviflorus*	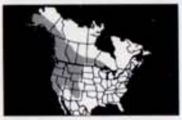	woodlands, canyons, trailsides, stream banks	🏃 🍎 ☕ 🥧 🫙
182 Common blackberry *Rubus allegheniensis*	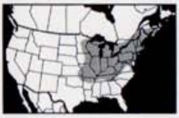	old fields, fencerows, edges of woodlands	🏃 🍎 ☕ 🥧 🫙
182 Black raspberry *Rubus occidentalis*	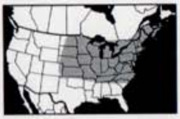	old fields, fencerows, edges of woodlands	🏃 🍎 ☕ 🥧 🫙
184 California blackberry *Rubus ursinus*	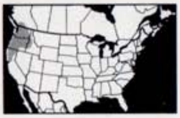	along trails, roads, canyons, fields	🏃 🍎 🥧 🫙

Page/Plant	Region	Habitat	Edible Uses
	Leaves pinnately compound		
184 Smooth sumac *Rhus glabra*	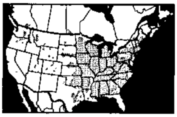	fields, open woodlands, stream banks, trailsides	🚶 ☕ 🥤 🫙 🚦
186 Squawbush *Rhus trilobata*	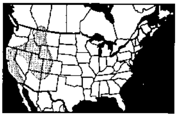	foothills, canyons, slopes, washes	🚶 ☕ 🥤 🫙 🚦
186 Staghorn sumac *Rhus typhina*	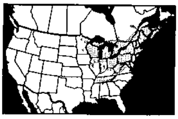	fencerows, roadsides, edges of woodlands	🚶 ☕ 🥤 🫙 🚦
188 Screwbean mesquite *Prosopis pubescens*	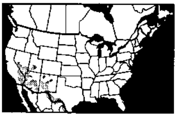	washes, riverbottoms, stream banks	🥤 🏺
188 Glandular mesquite *Prosopis glandulosa*	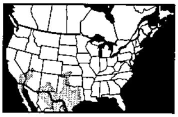	desert valleys, grazing lands, range	🚶 ☕ 🥤 XXX 🏺
190 American elder *Sambucus canadensis*	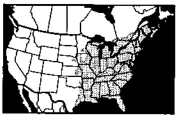	along streams and rivers, edges of woodlands	🥤 🍳 🥧 🫙 ☠
190 Blue elder *Sambucus cerulea*	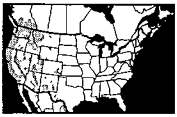	stream banks, washes, edges of fields	☕ 🥤 🥧 🫙 ☠
TREES: CONIFERS			
192 Singleleaf pinyon pine *Pinus monophylla*	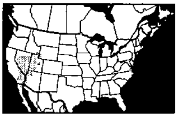	foothills, canyons, mountain slopes	🌰 XXX

Page/Plant	Region	Habitat	Edible Uses
192 Sugar pine *Pinus lambertiana*	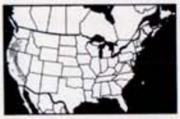	mountain slopes	🌰 🛍️
	Leaves tiny, scaly		
192 Digger pine *Pinus sabiniana*	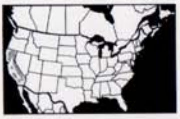	foothills, lower mountain slopes	🌰 🛍️
194 Rocky Mountain juniper *Juniperus scopulorum*	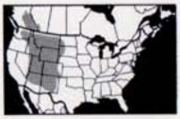	ridges, bluffs, dry rocky hillsides	🧂 ☕ 🚦
194 California juniper *Juniperus californica*	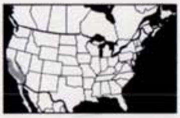	dry slopes, canyons, desert slopes	🏃 🧂 ☕
196 Utah juniper *Juniperus osteosperma*	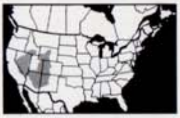	mesas, mountain slopes, high plains	🧂 ☕ 🚦
TREES: FLOWERING			
	Evergreen		
196 Sea grape *Coccoloba uvifera*	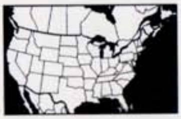	beaches, sea shores	🫙
198 California laurel *Umbellularia californica*	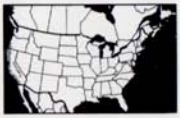	hillsides, flatlands, lower mountain slopes	🌰 🧂 🛍️
	Deciduous		
198 Pawpaw *Asimina triloba*	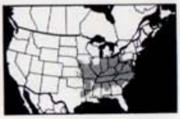	river valleys, bottomlands	🍎 🥧 🚦

Page/Plant	Region	Habitat	Edible Uses
	Leaves unlobed or 1- to 3-lobed		
200 White mulberry *Morus alba*	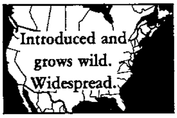 Introduced and grows wild. Widespread.	fencerows, roadsides, old fields	🏃 ♨ 🥧 🍬 🚦
200 Red mulberry *Morus rubra*	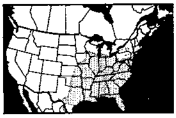	river valleys, floodplains, low hills	🏃 🍎 🥤 ♨ 🥧 🫙 🚦
	Flowers white, showy, early		
202 American plum *Prunus americana*	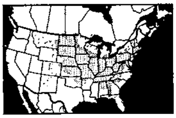	deciduous woodlands, stream banks	🏃 🍎 🥧 🫙 ☠
202 Flatwood plum *Prunus umbellata*	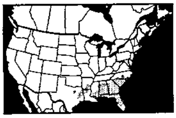	coastal plain; piedmont; rivers, swamps, hammocks	🏃 🍎 🥧 🫙
204 Black cherry *Prunus serotina*	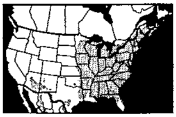	mixed hardwood forests, woodlands	🏃 🍎 🥤 🥧 🫙 ☠
204 Common chokecherry *Prunus virginiana*	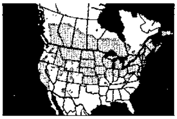	moist soils of roadsides, fencerows, woodland margins	🏃 🥧 🫙 ☠
206 Downy serviceberry *Amelanchier arborea*	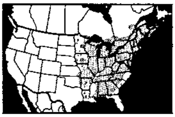	woodlands, rocky slopes, river banks	🏃 🍎 🥧 🫙
206 Saskatoon serviceberry *Amelanchier alnifolia*	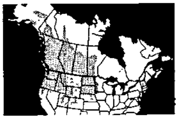	mountain slopes, hillsides, prairies, stream banks	🏃 🍎 🥧 🫙

Autumn

Page / Plant	Region	Habitat	Edible Uses
HERBS			
	Land plant, flowers yellow		
210 Jerusalem artichoke *Helianthus tuberosus*	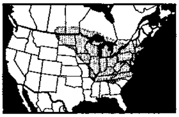	roadsides, low wet ground, waste areas	
	Aquatic plants		
210 Lotus lily *Nelumbo lutea*	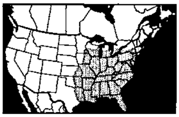	lakes, ponds, slow-moving streams	
212 Yellow pond lily *Nuphar advena*	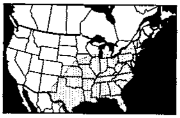	lakes, ponds, slow streams, tidal waters	
	Semiaquatic, flowers white		
212 Arrowhead *Sagittaria latifolia*	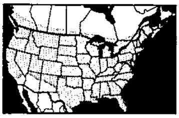	edges of lakes, ponds, streams, ditches	
VINES			
214 Riverside grape *Vitis riparia*	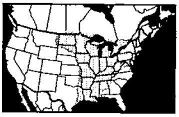	edges of woodlands, streams, rivers	
214 Frost grape *Vitis vulpina*	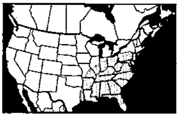	along streams and rivers, in floodplains and thickets	

Page/Plant	Region	Habitat	Edible Uses
216 California grape *Vitis californica*	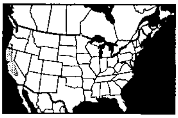	along streams and rivers, moist canyons	🚶 🥤 🍲 🥧 🫙
216 Wild potato vine *Ipomoea pandurata*	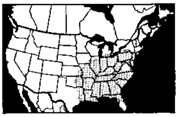	fields, fencerows, roadsides	🥔 🚦

SHRUBS

Flowers yellow, leaves simple

Page/Plant	Region	Habitat	Edible Uses
218 Common barberry *Berberis vulgaris*	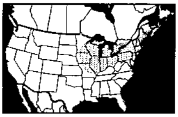	fields, fencerows, roadsides, disturbed sites	🚶 🥤 🫙
218 Spice bush *Lindera benzoin*	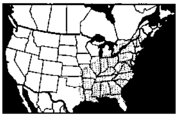	along streams, rich woodlands	🧂 ☕
220 Desert hackberry *Celtis pallida*	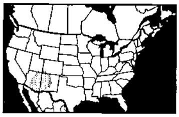	desert foothills, mesas	🚶 🍎 🥧 🫙

Leaves compound, flowers white or pink

Page/Plant	Region	Habitat	Edible Uses
220 California rose *Rosa californica*	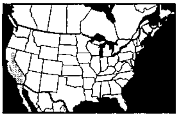	along streams and rivers, washes, canyons	🚶 ☕ 🫙 🍬
222 Wild rose *Rosa carolina*	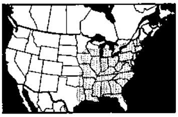	open woodlands, fields, pastures	🚶 ☕ 🫙 🍬
222 Sweetbrier rose *Rosa eglanteria*	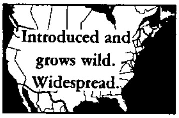 	roadsides, old fields, fencerows	🚶 ☕ 🫙 🍬

Page/Plant	Region	Habitat	Edible Uses
222 Wrinkled rose *Rosa rugosa*	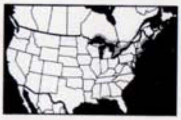	roadsides, sandy dunes, near seashores	🧍 ☕ 🫙 🍬
Leaves simple, flowers white			
224 Highbush cranberry *Viburnum trilobum*	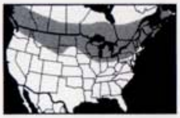	along streams, wet thickets, moist woodlands	🥤 🥧 🫙
224 Hobblebush *Viburnum alnifolium*	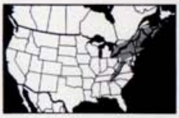	stream banks, ravines, moist woodlands	🧍 🥧
226 Nannyberry *Viburnum lentago*	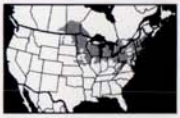	swamps, stream banks, hillsides	🧍 🥧
226 Squashberry *Viburnum edule*	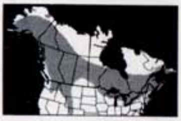	along streams, wet thickets, moist woodlands	🥤 🥧 🫙
TREES: CONIFER			
228 Pinyon pine *Pinus edulis*	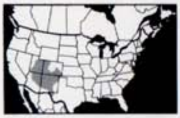	mesas, plateaus, foothills, lower mountain slopes	🌰 ▣
TREES: DECIDUOUS			
Leaves simple, fruits nuts			
228 White oak *Quercus alba*	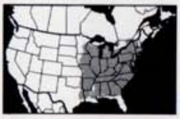	eastern deciduous forests	🌰 ☕ ▣ 🍬
230 Chestnut oak *Quercus prinus*	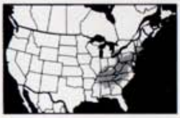	slopes and ridges of deciduous forests	🌰 ☕ ▣ 🍬

Page / Plant	Region	Habitat	Edible Uses
230 Gambel's oak *Quercus gambelii*	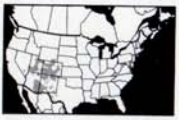	dry foothills, canyons	🌰 ☕ ▣ 🍬
230 Blue oak *Quercus douglasii*	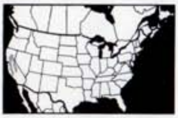	interior valleys, rolling hills, mountain slopes	🌰 ☕ ▣ 🍬
232 Northern red oak *Quercus rubra*	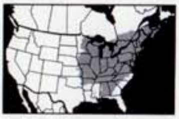	valleys, ravines, lower mountain slopes	☕ ▣ 🚦
232 American beech *Fagus grandifolia*	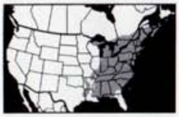	river valleys, deciduous forests, mountain slopes	🌰 ☕ ▣
234 Allegheny chinkapin *Castanea pumila*	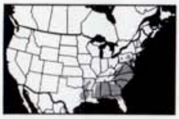	dry woodlands, slopes, sandy coastal areas	🌰 ☕ 🍲

Leaves simple, fruits berrylike or fleshy

236 Washington hawthorn *Crataegus phaenopyrum*	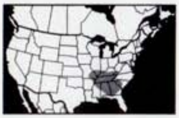	old fields, stream banks, open woodlands	🏃 ☕ 🫙
236 Downy hawthorn *Crataegus mollis*	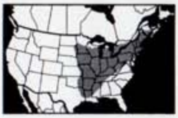	along rivers, streams, lowland fields	🏃 ☕ 🫙
238 Fleshy hawthorn *Crataegus succulenta*	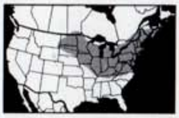	hillsides, old fields, coastal areas	🏃 ☕ 🫙
238 Cockspur hawthorn *Crataegus crus-galli*	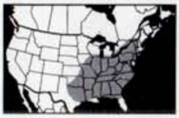	river valleys, old fields, hillsides	🏃 ☕ 🫙

Page/Plant	Region	Habitat	Edible Uses
238 Common persimmon *Diospyros virginiana*	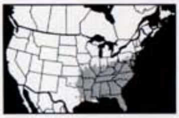	bottomlands, edges of woodlands, old fields, fencerows	🍎 ☕ 🥧 🫙
Leaves compound, fruits berrylike			
240 American mountain ash *Sorbus americana*	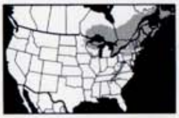	rocky slopes and ridges, seepage areas	🍎 🥧 🫙
Leaves compound, fruits nuts			
240 Pecan *Carya illinoensis*	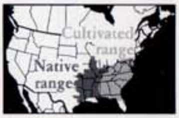	bottomlands	🧍 🌰 ▣
242 Big shellbark hickory *Carya laciniosa*	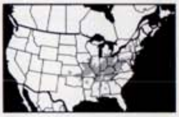	floodplains, bottomlands, river valleys	🧍 🌰 ▣ 🫖
242 Shagbark hickory *Carya ovata*	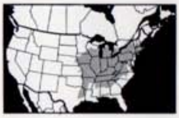	dry upland slopes, well-drained lowlands	🧍 🌰 ▣ 🫖
244 Mockernut hickory *Carya tomentosa*	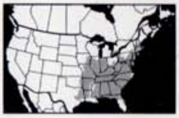	ridges, slopes, river valleys	🧍 🌰 ▣ 🫖
244 Black walnut *Juglans nigra*	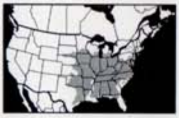	bottomlands, floodplains, low mixed woodlands	🧍 🌰 ▣ 🫖
246 California walnut *Juglans californica*	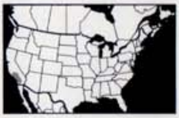	along streams and rivers, in bottomlands	🧍 🌰 ▣
246 Butternut *Juglans cinerea*	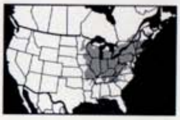	bottomlands, floodplains, mixed deciduous forests	🧍 🌰 ▣ 🥒 🫖

Winter

Page / Plant	Region	Habitat	Edible Uses
SHRUBS			
	Leaves compound		
250 Staghorn sumac *Rhus typhina*	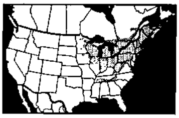	fencerows, roadsides, edges of woodlands	🚶 ☕ 🥛 🫙 🧂
	Leaves simple		
250 Highbush cranberry *Viburnum trilobum*	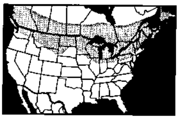	along streams, wet thickets, moist woodlands	🥛 🥧 🫙
252 Squashberry *Viburnum edule*	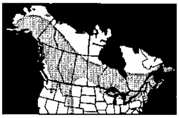	along streams, wet thickets, edges of woodlands	🥛 🥧 🫙
TREES: DECIDUOUS			
	Leaf scars alternate		
252 Sweet birch *Betula lenta*	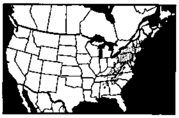	open woodlands, moist slopes	☕ 🥛 XXX 🫗
254 Yellow birch *Betula alleghaniensis*	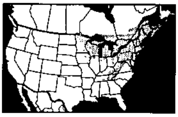	rich woodlands, lower mountain slopes	☕ 🥛 XXX 🫗
254 American mountain ash *Sorbus americana*	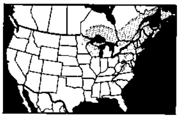	rocky slopes, seepage areas	🍎 🥧 🫙
	Leaf scars opposite		
256 Sugar maple *Acer saccharum*	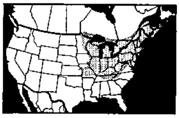	eastern deciduous forests	🌰 🥛 🫗

Spring

Spring is the prime season for vitamin-rich shoots and leafy greens. At this time, certain flowers are edible too, and free-flowing sap in trees can be tapped for syrups and tea bases. There's an abundance of plants offering roots, bulbs, and underground stems rich in carbohydrates and other nutrients. Besides, a few other plants already sport nourishing berries and seeds.

The secret for catching the most spring edibles at opportune times is planning. For this, simply consult this book's illustrated table of contents. From it you can make a list of edibles you might encounter on your next spring outing. Or you might even decide to schedule an outing to coincide with the probable readiness of specific plants. Of course, you might prefer a more casual approach and just refer to the book whenever you encounter plants that interest you. Whatever your method, spring's edible bounty awaits you.

Additional spring edibles

Plants listed below have parts that are edible in spring and at least one additional season. Descriptions and photos of these plants appear in a seasonal section other than "Spring," as indicated.

Alpine bistort (*Polygonum viviparum*), p. 154
American bistort (*Polygonum bistortoides*), p. 154
American brooklime (*Veronica americana*), p. 144
American elder (*Sambucus canadensis*), p. 190
Arrowhead (*Sagittaria latifolia*), p. 212
Blue elder (*Sambucus cerulea*), p. 190
California grape (*Vitis californica*), p. 216
California laurel (*Umbellularia californica*), p. 198
Chufa (*Cyperus esculentus*), p. 130
Common day lily (*Hemerocallis fulva*), p. 136
Creeping snowberry (*Gaultheria hispidula*), p. 152
Creeping wintergreen (*Gaultheria humifusa*), p. 160
Frost grape (*Vitis vulpina*), p. 214
Groundnut (*Apios americana*), p. 148
Indian fig (*Opuntia humifusa*), p. 138
Jerusalem artichoke (*Helianthus tuberosus*), p. 210
Labrador tea (*Ledum groenlandicum*), p. 160
Lotus lily (*Nelumbo lutea*), p. 210

Prickly pear (*Opuntia phaeacantha*), p. 140
Red mulberry (*Morus rubra*), p. 200
Reed grass (*Phragmites communis*), p. 128
Riverside grape (*Vitis riparia*), p. 214
Soapweed (*Yucca glauca*), p. 134
Spanish bayonet (*Yucca baccata*), p. 134
Spice bush (*Lindera benzoin*), p. 218
Sweet fern (*Comptonia peregrina*), p. 174
Watercress (*Nasturtium officinale*), p. 146
Western camass (*Camassia quamash*), p. 136
White mulberry (*Morus alba*), p. 200
Wild potato vine (*Ipomoea pandurata*), p. 216
Wintergreen (*Gaultheria procumbens*), p. 150
Yellow pond lily (*Nuphar advena*), p. 212

Ostrich fern *Matteucia struthiopteris*
(synonyms: *M. pensylvanica, Pteretis nodulosa*)

Habitat: bottomlands, swamps, streamsides, ponds; rich forest soil. **Identification:** ***perennial herb with 2 leaf types. Larger plumelike, sterile green leaves (fronds) 0.6–1.6 m (1–5 ft) long,*** with stiff leafstalks to 40 cm (16 in) long, ***leaf blade broadest above middle,*** short-tapering at tip, long-tapering at base, with 20 or more pairs of side branches (pinnae), the pairs alternating. ***Smaller fertile leaves dark brown when mature,*** 40–65 cm (16–26 in) long, stiff, with side branches 3–5 cm (1–2 in) long, bearing powdery reproductive spores. Immature leaves or fiddleheads usually emerge in clusters amid few remaining, upright, brown fertile leaves of previous year, sparsely covered near top with papery, brown scales that fall away as the leaves uncoil; leafstalk (rachis) is conspicuously grooved. **Harvest:** Collect coiled fiddlehead leaves to 20 cm (8 in) tall in April in southern part of range and early May farther north, about same time as flowering of SERVICEBERRY trees. **Preparation:** Scrape inedible brown scales off by hand or wash in cold water. To cook, steam 10 min, season and butter hot; can also be boiled. In field, eat raw as a trail snack or salad item. **Related edible species:** bracken fern *(Pteridium aquilinum)* and cinnamon fern *(Osumunda cinnamomea)*. **Poisonous look-alikes:** None.

Wild onion, Nodding onion *Allium cernuum*
Also edible summer, autumn

Habitat: dry woodlands, rocky slopes, ledges, ridges, prairies. **Identification:** ***perennial herb*** to 60 cm (24 in) high; underground ***bulb usually reddish-purple, slender,*** broadest near base, tapering above. Stem (scape) erect except ***arching at tip. Leaves*** several, from base of plant, linear, mostly 2–4 mm (0.1 in) wide, shorter than stem, ***with onion odor when broken,*** soft, flexible. Flowers in rounded to flat-topped cluster, nodding when young, erect with age, each long-stalked, stalks 10–24 mm (0.4–0.9 in) long, petallike parts white to pink, 4–6 mm (0.2 in) long. ***Fruiting capsules*** in summer nearly rounded to broadest near tip, 3–5 mm (0.1–0.2 in) long, ***3-parted, each part with 2 pointed crests;*** seeds triangular, black. **Harvest:** young tops before flowers appear. Bulbs in spring and fall. **Preparation:** For creamed onions, melt 1 tbsp butter in saucepan, add 2 tbsp flour, ¾ tsp salt, dash of pepper, and cook with constant stirring for 1 min. Add 1 cup boiled onion bulbs and leaves. Stir in milk and simmer until thickened. For wild pickles, pack top bulb clusters in pickle jar, add 1 tbsp mixed pickling spices, and fill with mixture of 2 parts cider vinegar to 1 part water. Seal, store cool for a month or more before using. **Related edible species:** WILD GARLIC and WILD LEEK (next), and several other *Allium* species. **Poisonous look-alikes:** Field garlic *(A. vineale)* is not poisonous but is too strong for most tastes. Avoid all onionlike plants that lack onion odor; some, such as DEATH CAMASS, are highly poisonous.

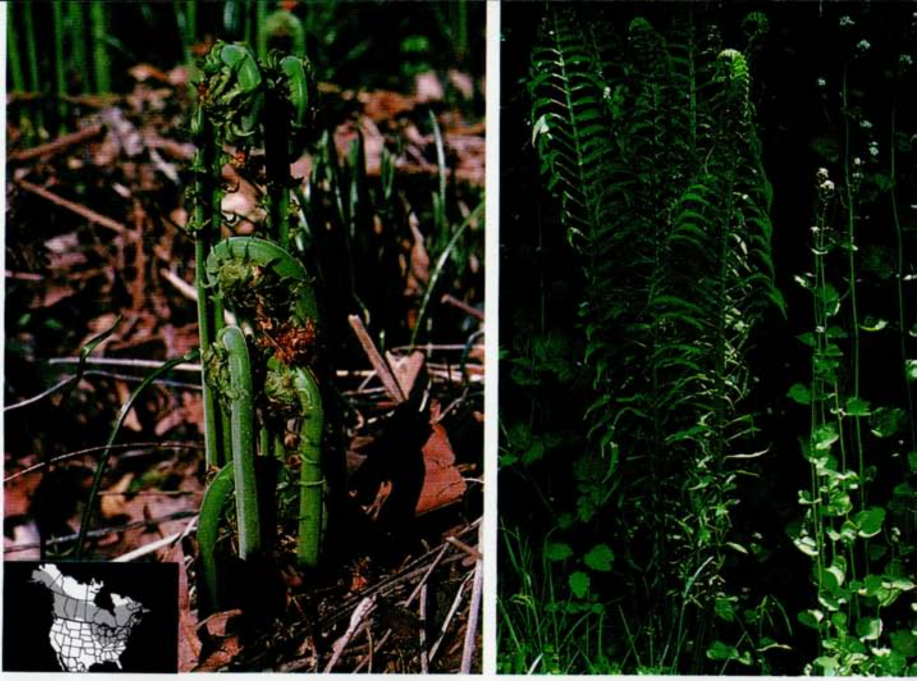

Ostrich fern: young fronds, or fiddleheads, left; and mature summer leaves (Dykeman).

Wild onion: entire plants and flowers (NYBG).

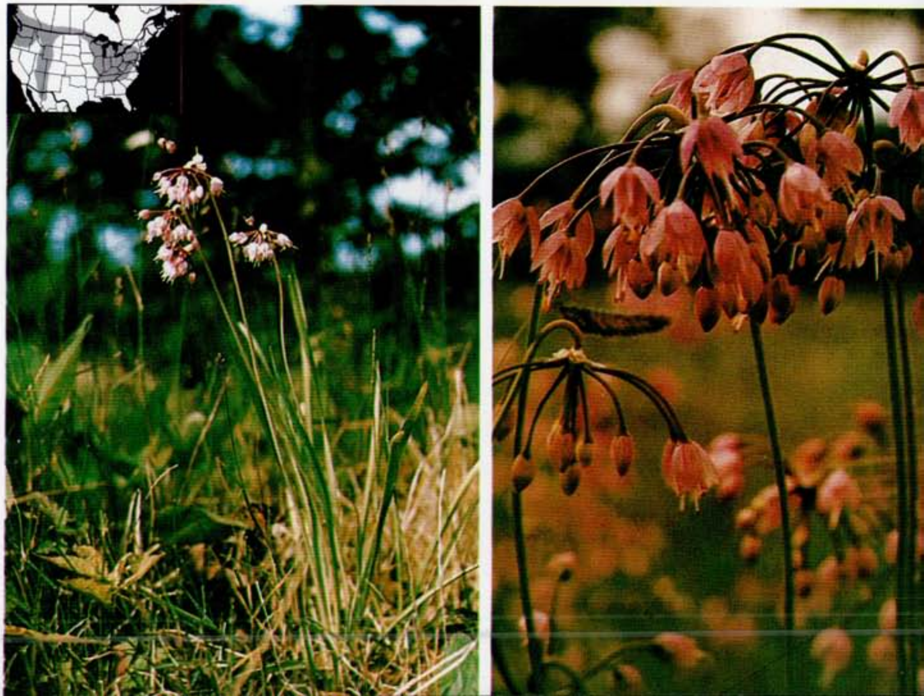

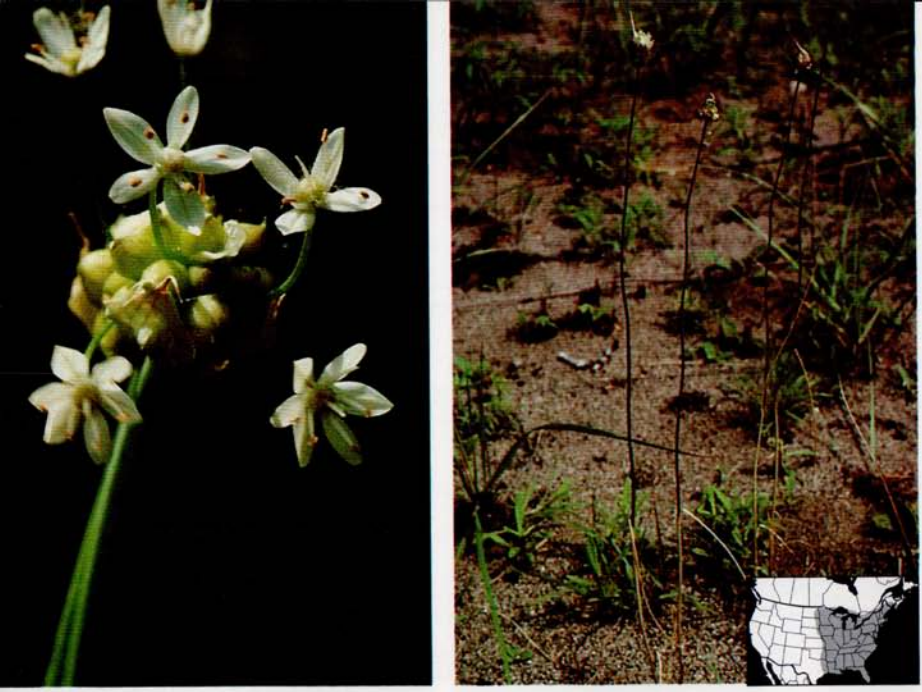

Wild garlic: flowers (Speas). Entire plant (Kavasch).

Wild leek: rootstocks and entire plants (Dykeman).

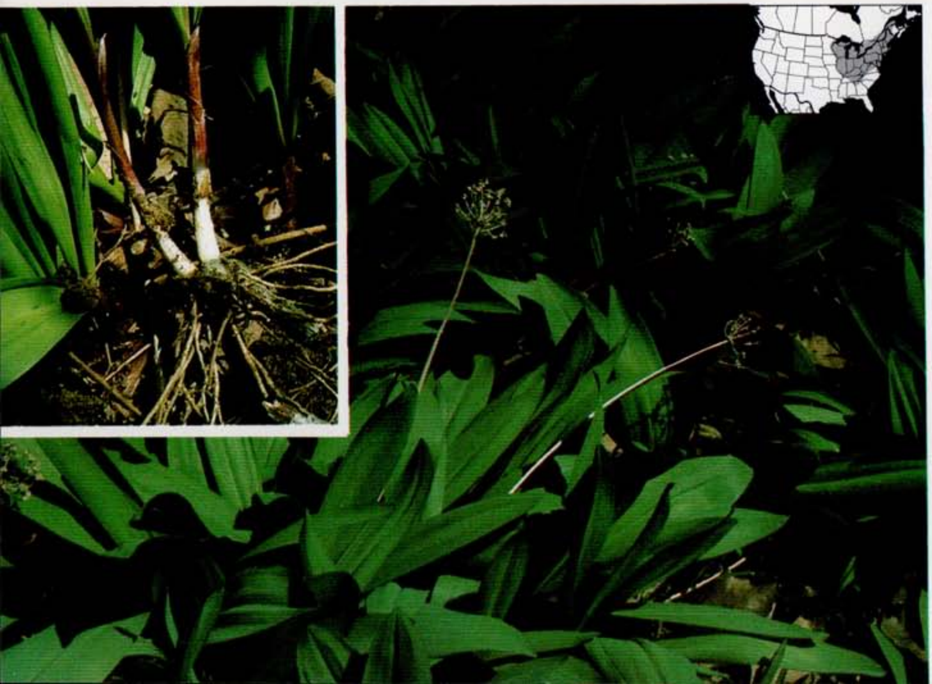

Wild garlic *Allium canadense*

Also edible summer, autumn

Habitat: common, open woodlands, fields, prairies. **Identification:** ***perennial herb*** to 58 cm (23 in) high; ***underground bulb brown, egg-shaped to nearly globe-shaped, 2–3 cm*** (0.8–1.2 in) ***across.*** Stem (scape) erect. ***Leaves*** several, from or near base, linear, to 35 cm (13.8 in) long, 2–5 mm (0.1–0.2 in) wide, rarely wider, flat, flexible, ***with strong onion odor when broken,*** shiny green. ***Flower*** (spring) ***clusters consist of several tiny white to purple bulbils*** (miniature bulbs) ***and zero to few pink to white flowers on long, flexible stalks. Fruiting capsules*** rounded or nearly so, 2–3 mm (0.1 in) long, 3-parted, ***smooth on top;*** seeds triangular, tiny, black. **Harvest:** Use whole plant in early spring before flower stalk appears, or use bulbs in spring and fall. Bundle bulbs with 15–20 cm (6–8 in) stems and dry at room temperature for later use. Pick green top bulb clusters on flower stalk in late spring before ripe. **Preparation:** For salad or vegetable, remove tough outer layers of young whole plant. Chop into green salad or boil whole for ½ hr in salted water, then drain, season with butter, and serve. For soup, use drained-off water to make cream of onion soup. Combine 1 qt onion water, 2 tbsp butter, 1 cup half-and-half. Add smooth mixture of 2 tbsp flour in milk, heat to simmer, and serve. Use this plant in any soup recipes calling for onions or with many cooked vegetables and meats. For other uses, see WILD ONION (preceding). **Related edible species:** WILD ONION, WILD LEEK (next), and several other *Allium* species. **Poisonous look-alikes:** See WILD ONION.

Wild leek, Ramp *Allium tricoccum*

Also edible summer, autumn

Habitat: rich, neutral, moist woods, frequently under maples. **Identification:** ***colonial herb. Leaves directly from bulb, 2 or 3 in number, 10–30 cm (4–12 in) long, 2–6 cm (0.8–2.4 in) wide, light green, smooth, fleshy, onion scented, appearing in early spring, shrivel by flowering in midsummer.*** Flowers small, white, in hemispherical cluster on leafless stalk 15–40 cm (6–16 in) tall. Fruit capsule small, 3-lobed, each lobe with 1 hard seed. **Harvest:** young tops before they unfurl into broad leaves; bulbs from early spring through autumn. **Preparation:** add fine-chopped young leaves to salads or boil whole for 20 min in salted water, drain, season, and serve. Excellent in soup. Bulbs excellent pickled, chopped in salads, cooked alone, or with meat or other vegetables. Add chopped fine bulbs to mashed potatoes. For fried leeks parboil 3 min, drain, add more water, boil until tender. Saute in butter, serve covered with bread crumbs. *Caution:* Leeks are often very strong flavored and may cause gastric distress. Eat sparingly. **Related edible species:** WILD ONION, WILD GARLIC (preceding), and several other *Allium* species. **Poisonous look-alikes:** See WILD ONION.

Corn lily, Clinton's lily *Clintonia borealis*

Habitat: rich woodlands, deep thickets, higher elevations, cool moist sites. **Identification:** ***perennial herb*** to 40 cm (16 in) high, from a slender underground stem (rhizome); ***each plant with 2–5 large basal leaves. Leaves*** usually broadest near or above the middle, ***10–38 cm*** (4–15 in) ***long,*** 4–14 cm (1.6–5.5 in) wide, pointed at tip, entire along margin, dark shiny green. ***Flowers 3–8 at top of erect stalk,*** nodding, each with 6 greenish-yellow, narrow petallike structures 1.3–1.8 cm (0.5–0.7 in) long. ***Fruits are globe-shaped, bright blue berries 8–10 mm*** (0.3–0.4 in) ***in diameter.*** **Harvest:** Gather leaves before fully unfurled. Fully formed leaves are tough and unpleasant tasting. **Preparation:** Chop young leaves into salad with other greens of the season. Boil leaves about 10 min and serve with butter and seasonings. **Related edible species:** White clintonia *(C. umbellata)* has white flowers and black fruit. **Poisonous look-alikes:** *Be positive* of corn lily's young-growth identification features because some similar-looking species are poisonous.

Asparagus *Asparagus officinalis*

Habitat: roadsides, disturbed sites, especially sandy and well-drained soils. **Identification:** ***perennial herb*** 1–2 m (3.3–6.6 ft) high from thick, white, underground stem (rhizome); ***stems*** upright, often branched, ***green***, becoming very slender at top. ***Leaves*** alternate, ***reduced to small scales, lance-shaped,*** to 2.5 cm (1 in) long. ***Flowers produced along upper stems, usually singly or in pairs, hanging on slender stalk, narrowly bell-shaped, 3–6 mm*** (0.1–0.2 in) ***long, yellow to greenish-yellow, 6-lobed. Fruits*** are ***globe-shaped berries 4–8 mm*** (0.2–0.3 in) ***in diameter, bright red,*** fleshy, containing 3–6 black seeds. **Harvest:** identical to cultivated asparagus. Locate by presence of dried Christmas-tree-shaped stalks of previous year. Slice off young stalks just below ground level when green shoots protrude only a few inches. More shoots will grow, allowing reharvesting. *Caution: Avoid older stalks and other parts, which are mildly toxic.* **Preparation:** Wash stalks and peel off any tough covering near base. Tie stalks in bundle and steam or boil until tender, usually 10–15 min. Serve with butter and seasonings or cream sauce. Excellent prepared as cream soup. See any vegetable cookbook for other recipes. **Related edible species:** none. **Poisonous look-alikes:** none.

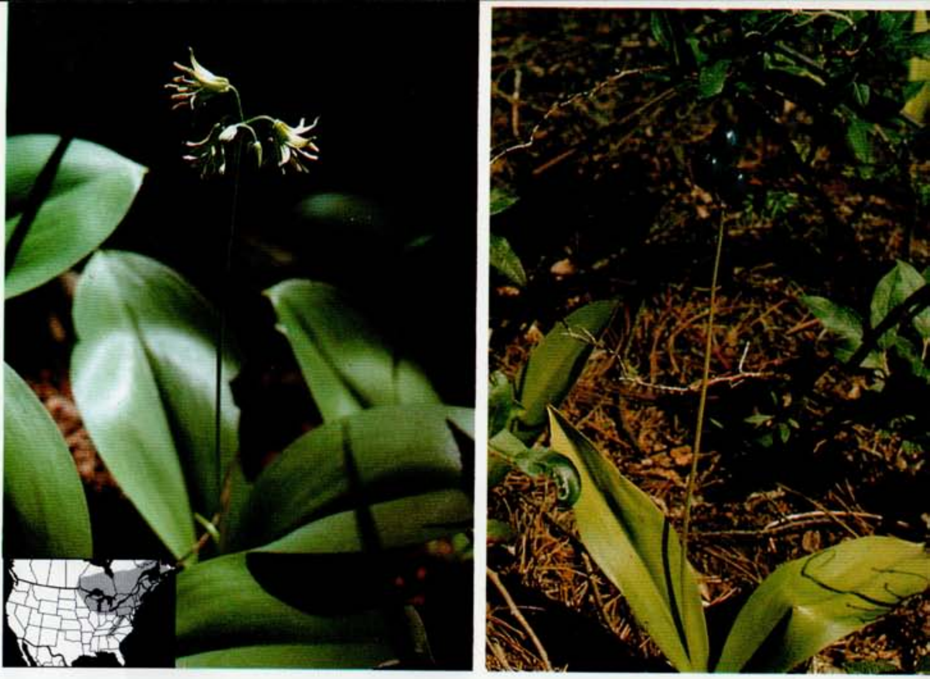

Corn lily: flowering and fruiting plants (NYBG).

Asparagus: young shoots (Dykeman). Mature branch with fruit (Elias).

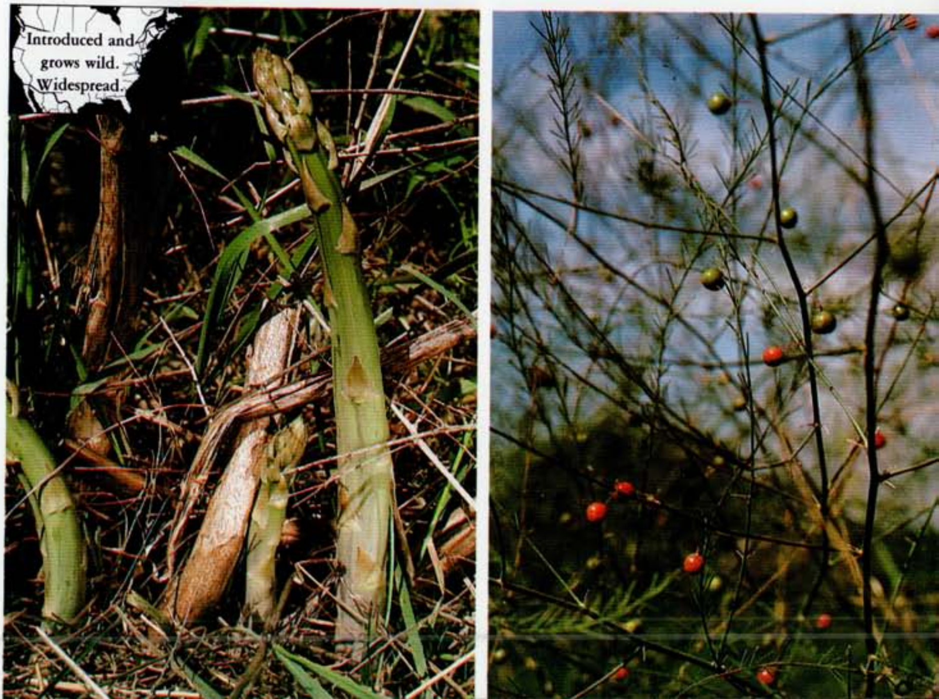
Introduced and grows wild. Widespread.

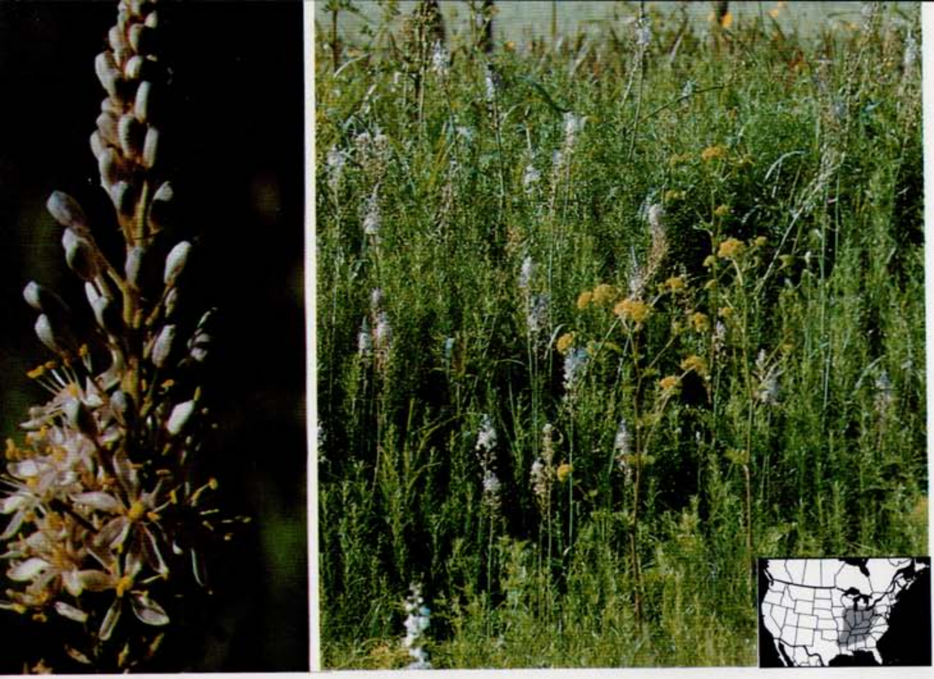

Eastern camass: flower and field (NYBG).

Common greenbrier: fruiting branch (Elias). Flowers (NYBG).

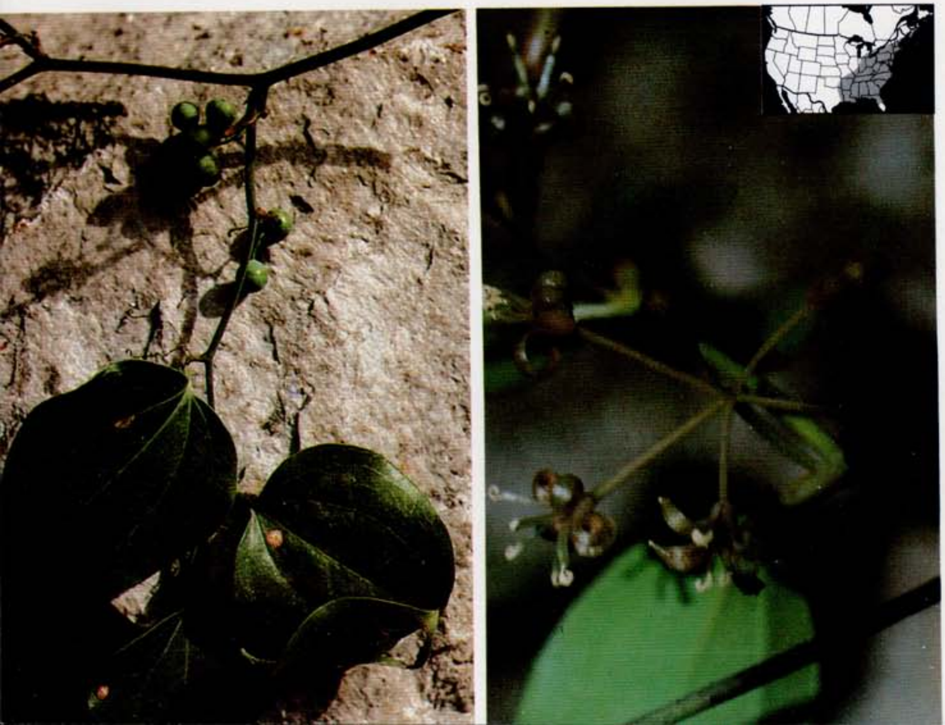

Eastern camass, Camass lily *Camassia scilloides*

Also edible summer, autumn, winter

Habitat: fields, wet meadows, moist open woodlands. **Identification:** ***perennial herb*** from a globe-shaped bulb 1–3 cm (0.4–1.2 in) in diameter. ***Leaves attached at base, long-narrow, 20–60 cm*** (7.9–23.6 in) ***long, 5–17 mm*** (0.2–0.7 in) ***wide,*** pointed at tip. ***Flowering stalk upright to 60 cm (24 in) tall;*** stalked flowers in narrow cluster; ***each flower with 6 white, blue, or violet petals*** or petallike structures 7–12 mm (0.3–0.5 in) long. ***Fruit*** small, nearly globe-shaped, ***3-parted capsule*** 6–12 mm (0.2–0.5 in) in diameter. **Harvest:** Dig or pull bulbs any season; safest when blue flowers distinguish it from DEATH CAMASS. Dig and transplant in ground free of DEATH CAMASS. Study descriptions of DEATH CAMASS carefully. **Preparation:** Raw bulbs are palatable. May be boiled 25–30 min or baked in foil for 45 min at 175°C (350°F) but best preparation is the pit method used by Indians and described in the Introduction, in which eastern camass is baked for 1–3 days. The baked bulbs are very dark and sugary. Slice pit-baked or boiled bulbs. Then dry in sun or warm oven before storing in paper bag in dry place. Bulbs can be used as a potato substitute. **Related edible species:** WESTERN CAMASS. **Poisonous look-alikes:** DEATH CAMASS, narrow-leaved plants with green, white, or bronze flowers, single or several in narrow spike.

Common greenbrier, Catbrier *Smilax rotundifolia*

Also edible summer

Habitat: common plant of roadsides, secondary woodlands, woodland clearings, streamsides, thickets. **Identification:** ***climbing perennial vine, stems*** tough, becoming woody, ***often strongly 4-sided,*** green, tendrils many, ***armed with stout, sharp thorns. Leaves*** alternate, deciduous to almost evergreen (southern part of range), 4–10 cm (1.6–4 in) long, nearly as wide, ***broadest near base to nearly circular,*** pointed at tip, entire along margin, ***leathery,*** dark shiny green. Vines bearing either male or female flowers in several-flowered, flat-topped clusters, green or yellow. ***Fruits bluish-black berries,*** globe-shaped 5–9 mm (0.2–0.4 in) in diameter, often containing 2 or 3 seeds. **Harvest:** Gather new shoots and growing tips, plus uncurling leaves and tendrils, while they are still crisp and tender during spring and summer. **Preparation:** excellent raw as trail nibble. For salads use tenderest parts raw or boil slightly older tips 2–3 min, drain and cool. Add your favorite salad dressing. Steam or boil bundled shoots like asparagus or boil tendrils, leaves, and shoots together like spinach. Cook until tender and serve with sauces and seasonings as you would other vegetables. **Related edible species:** several other *Smilax* species, including BULLBRIER and CARRION FLOWER (next). **Poisonous look-alikes:** none.

Bullbrier, Greenbrier *Smilax bona-nox*

Also edible summer

Habitat: open woodlands, fields, abandoned farmlands, roadsides, disturbed sites. **Identification:** perennial ***climbing vine, stems 4-sided,*** becoming woody, frequently branching, green, tendrils present, with scattered stiff thorns or thornless. ***Leaves*** alternate, deciduous, 4–8 cm (1.6–3.2 in) long, triangular in shape, broadest near base, pointed at tip, entire or with few sharp teeth along ***thickened margin,*** almost leathery texture. Vines with either male or female flowers in several-flowered, rounded or flat-topped stalked clusters. ***Fruits black berries,*** globe-shaped or nearly so, ***6–9 mm*** (0.2–0.4 in) ***in diameter,*** usually 1-seeded. **Harvest and preparation:** shoots, tips, tendrils, leaves same as COMMON GREENBRIER. Also, dig and clean tuberous roots. Cut them into short pieces and pound them in metal container until fibers are separated and gelatinous sediment is released. Wash sediment from fibers and allow to settle. Strain out fibers and other floating matter. Carefully pour off water and allow sediment to dry and turn red. Boil 1 tbsp powdered sediment per cup of water for 10 min. Cool to make jelly or dilute and sweeten for drink. Also use powder as thickener for stews or substitute for part of flour in pancake batter. **Related edible species:** COMMON GREENBRIER (preceding) and CARRION FLOWER (next). Other tuberous-rooted *Smilax* species, such as China-root *(S. tamnoides)* in the East and *S. californica* in the West, are usable for gelatin. **Poisonous look-alikes:** none.

Carrion flower *Smilax herbacea*

Also edible summer

Habitat: open woodlands, margins of woodlands, roadsides, fields. **Identification:** perennial climbing ***vine; green stems, never woody,*** freely branching, tendrils present, ***lacking prickles or thorns. Leaves*** alternate, deciduous, broadest near base to almost rounded, pointed to rounded at tip, heart-shaped to rounded at base, entire along margin, becoming leathery. Vines bearing either male or female ***flowers on long stalks from junctions of uppermost leaves;*** flowers in densely-clustered rounded heads. Fruits in dense, rounded clusters, each ***fruit a berry***, globe-shaped, ***8–12 mm*** (0.3–0.5 in) ***in diameter, dark blue to bluish-black,*** containing several seeds. **Harvest and preparation:** same as for COMMON GREENBRIER and BULLBRIER (preceding). **Related edible species:** COMMON GREENBRIER and BULLBRIER. **Poisonous look-alikes:** none.

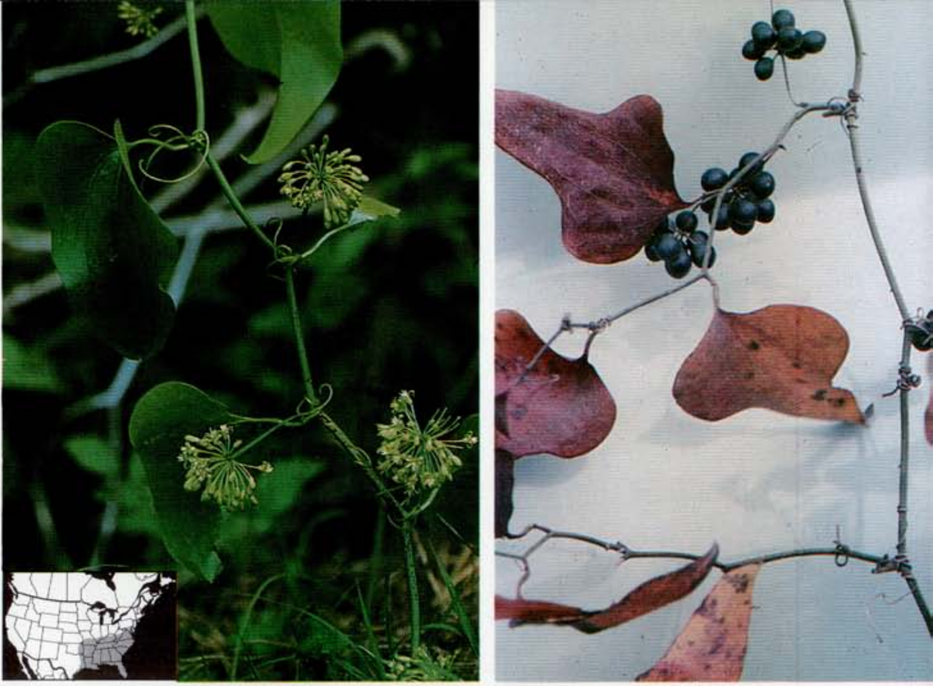

Bullbrier: flowering and fruiting branches (NYBG).

Carrion flower: flowering branch (NYBG). Fruiting branch (Kavasch).

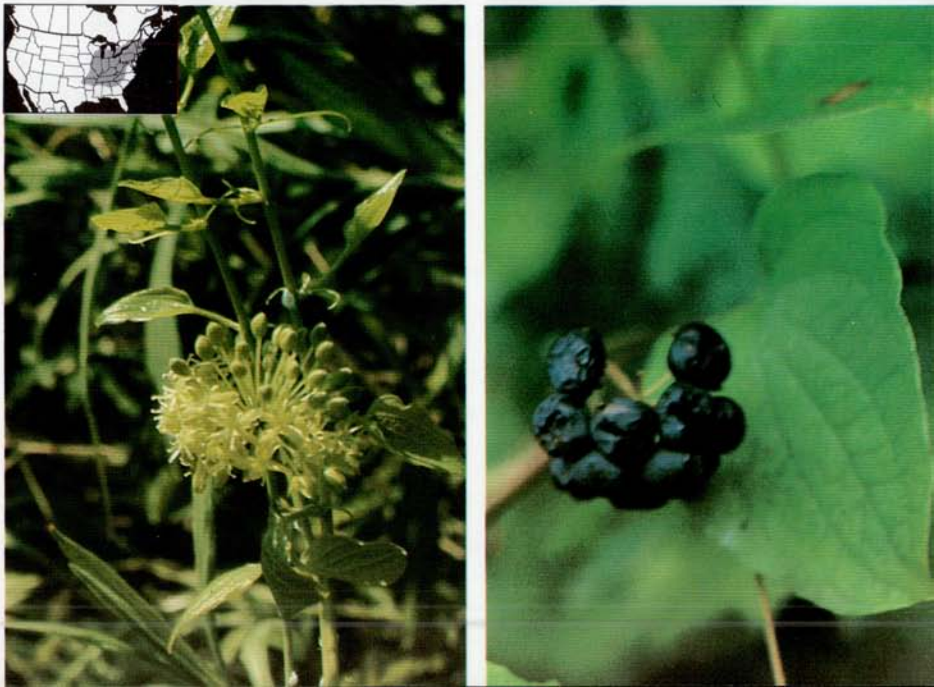

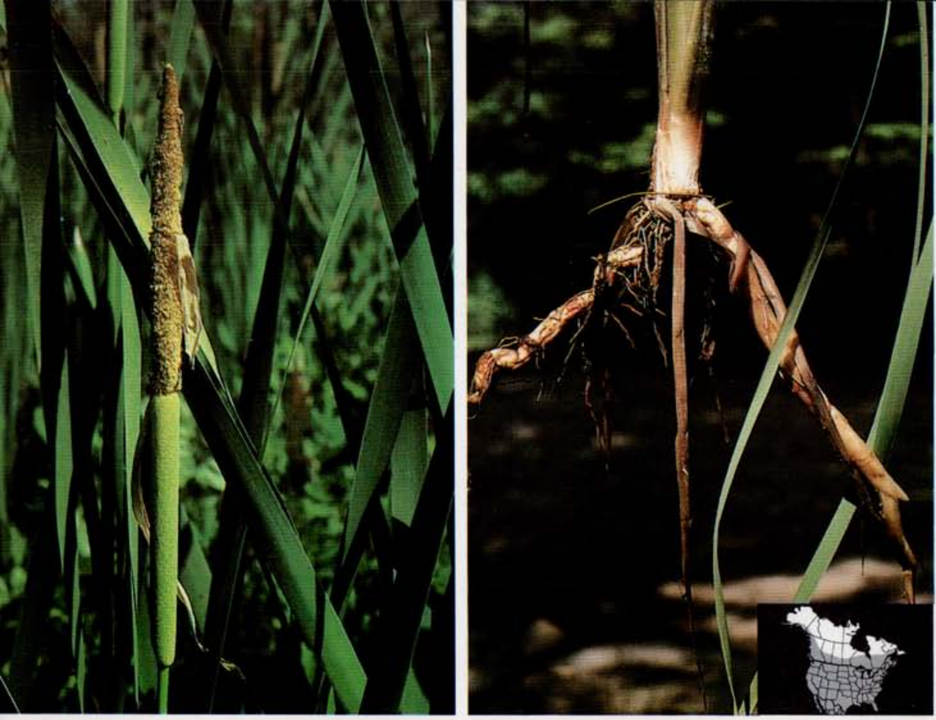

Cattail: flowering stalk and roots (Dykeman).

Narrow-leaved cattail: entire plants and flowers (Elias).

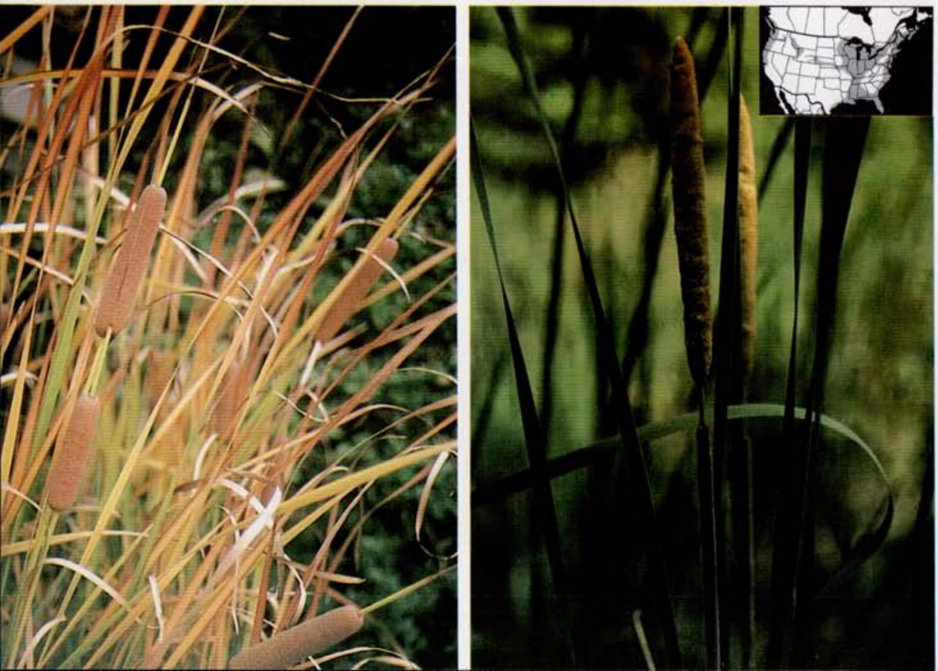

Cattail *Typha latifolia*

Also edible summer, autumn, winter

Habitat: shallow water of lakes, ponds, bays, ditches, marshes. **Identification:** large upright perennial herb to 3 m (9.8 ft) tall, with large stout, horizontal underground stems (rhizomes). ***Leaves*** sheathing at the base of stem, ***1–2.5 m*** (3.2–8.2 ft) ***long, 8–24 mm*** (0.3–0.9 in) ***wide, narrow, swordlike,*** pointed at tip, ***grayish green.*** Flowers at tip of erect stalks, tiny male flowers densely clustered in narrow, cylinder-shaped mass directly above female flowers densely clustered in a cylinder-shape 5–20 cm (2–8 in) long. ***Fruits*** in ***cigar-shaped clusters 2–3.5 cm*** (0.8–1.4 in) ***thick*** which break apart in soft downy-haired seeds in autumn. **Harvest and preparation:** In early spring, dig up roots to locate small, pointed dormant sprouts at ends of roots. Sprouts are edible raw or cooked. Later, similar sprouts appear on roots at bases of leaves. From then until plants are 60–90 cm (2–3 ft) tall, grasp leafy stalks below water surface and pull straight up, usually breaking stalks off at root. Peel away tough leafy layers to tender core, about 12 mm (0.5 in) diameter and to 30 cm (12 in) long. This "Cossack asparagus" is excellent raw, eaten like celery, or sliced into a salad. Or, cover cores with boiling water and simmer about 10 min. Serve with butter, salt and pepper or with oil-and-vinegar. Later pick green bloom spikes while still in papery sheaths and before pollen ripens above them. Remove sheaths and boil spikes in slightly salted water until tender. Eat like corn on cob, or scrape buds from cores for "corn" casserole. As pollen ripens, shake and rub it off stalks into container. Sift through fine screen to remove nonpollen materials. Substitute for part of flour in muffin, fritter, or pancake batter. For white flour, dig roots from fall through early spring. Wash and peel roots, crush cores thoroughly in container of water, and strain out fibers. Wash resulting starch several times, letting it settle and carefully pouring off water after each washing. Dry thoroughly for storage or use wet. **Related edible species:** NARROW-LEAVED CATTAIL (next). **Poisonous look-alikes:** stalks and roots of wild iris species (including YELLOW FLAG, BLUE FLAG). Harvest cattails where old stalks are abundant.

Narrow-leaved cattail *Typha angustifolia*

Also edible summer, autumn, winter

Habitat: ponds, lakes, swamps, bays, ditches. **Identification:** ***large*** upright ***perennial herb*** to 2 m (6.6 ft) tall, with thick, horizontal, underground stems (rhizomes). ***Leaves*** sheathing at base of stem, 0.6–1.5 m (2–5 ft) long, ***3–8 mm*** (0.1–0.3 in) ***wide, very narrow***, strap-shaped, flattened, pointed at tip. ***Flowers*** at tip of erect stalks, ***tiny male flowers*** densely clustered in narrow, cylinder-shaped column 2.5–7.5 cm (1–3 in), ***spaced at intervals above the larger, brown, cylinder-shaped clusters of female flowers*** 3–12 cm (1.2–4.7 in) long. ***Fruits are narrow, cigar-shaped clusters 5–15 cm*** (2–6 in) ***long*** which break apart into soft, downy-haired seeds in autumn. **Harvest and preparation:** same as for CATTAIL (preceding). **Related edible species:** CATTAIL. **Poisonous look-alikes:** See CATTAIL for details.

Hard-stem bulrush *Scirpus acutus*

Also edible summer, autumn, winter

Habitat: shores of ponds, lakes, streams, rivers, marshes. **Identification:** ***perennial herb*** 1–3 m (3.2–9.8 ft) high, from thick, tough, brownish underground stems (rhizomes); ***stems upright***, rounded, ***stiff, yellowish-green, unbranched,*** leaf blades absent or sometimes present, to 10 cm (4 in) long. ***Flowering heads (spikes) densely clustered and originating from one place near the tip of the stem,*** with 1 to several stalks each bearing several heads, each head cylinder-shaped, 8–18 mm (0.3–0.7 in) long, composed of several overlapping scales. Fruits are hard, somewhat flattened seeds 2–3 mm (0.1 in) long. **Harvest:** See GREAT BULRUSH (next). **Preparation:** Peel and cut firm older roots into sections. Crush them and boil in water until you have white gruel. Separate out any fibers and dry gruel into flour. For biscuit dough, add oil and water to reach right consistency. Add salt and baking powder, knead and roll to about ½-in thickness. Bake at 230°C (450°F) or twisted on stick by campfire. Also prepare flour by drying cleaned roots thoroughly, crush, and remove fibers. Pound and grind remaining material into flour. Use pollen as flour additive. Also grind seeds to use alone or with pollen and root flour for baking. Also see GREAT BULRUSH. **Related edible species:** All North American *Scirpus* species may be edible. **Poisonous look-alikes:** none because the fruiting pattern is distinctive.

Great bulrush, Tule *Scirpus validus*

Also edible summer, autumn, winter

Habitat: shallow water of lakeshores, ponds, and streams; fresh and brackish water marshes. **Identification:** ***perennial herb*** to 3 m (9.8 ft) high from tough, reddish scaly underground stems (rhizomes); ***stems upright, rounded, soft, pale green, unbranched;*** leaf blades absent. ***Flowering heads (spikes) produced near stem tips,*** with several stalked, branched clusters, each head egg-shaped, 0.4–1 cm (0.1–0.4 in) long, somewhat flattened, composed of a series of overlapping scales. Fruits hard, flattened seeds 2–3 mm (0.1 in) long. **Harvest:** Bulrushes, like CATTAILS, provide food year-round. Cut off young shoots in early spring or when they've formed in autumn. Later remove cores from bases of older stalks. Shake and beat pollen from stalks into bucket in summer and return to collect seeds the same way from late summer into winter. Dig up roots year-round to use firm, sweet ones as vegetable or as sugar source and older ones for flour. **Preparation:** Peel new shoots and youngest rootstocks and separate cores from older stalks. Eat them raw, sliced in salads, or boiled. Roast young shoots and roots, or slice and fry like potatoes. Bruise young roots and boil for several hours, boiling off water to produce sweet syrup. For other preparations, see HARD-STEM BULRUSH (preceding). **Related edible species:** HARD-STEM BULRUSH. **Poisonous look-alikes:** none.

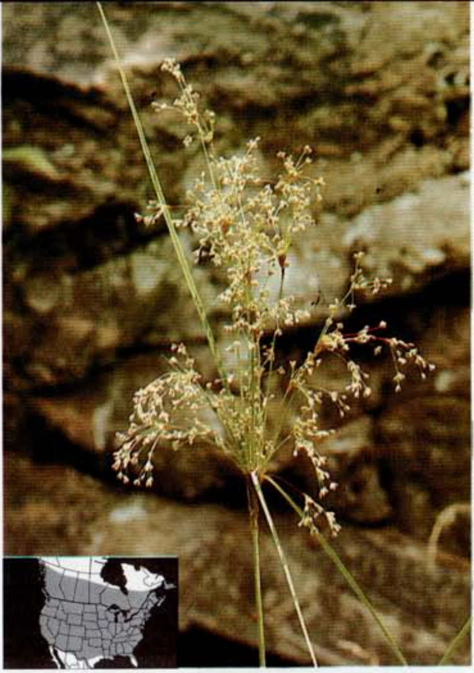

Hard-stem bulrush: flowers (Speas).

Great bulrush: fruiting (Kavasch).

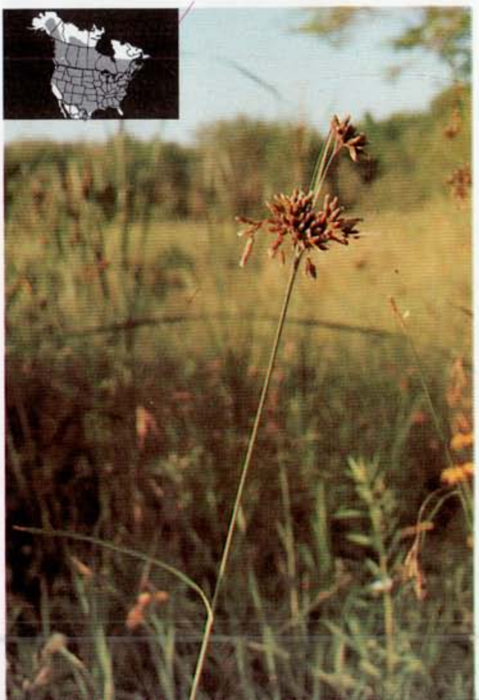

Pigweed, Redroot *Amaranthus retroflexus*

Also edible summer, autumn

Habitat: common and widespread; especially recently abandoned land, waste areas, roadsides. **Identification:** ***coarse annual herb*** to 2 m (6.6 ft) high; stem upright, stout, usually branched. ***Leaves*** alternate, ***long-stalked,*** broadest near base, ***5–10 cm*** (2–4 in) ***long, 2.5–6 cm*** (1–2.4 in) ***wide,*** blunt to short-pointed at tip, wavy along margin. ***Flowers*** late summer and fall, ***densely crowded on narrow spikes*** 6–18 cm (2.4–7.1 in) long, on upper part of stem. ***Fruits*** flattened, 1–2 mm (0.1 in) across, dry, each ***partially enclosed by rigid pointed bracts 5–8 mm*** (0.2–0.3 in) ***long;*** seeds rounded, shiny black, tiny. **Harvest:** For greens, pick young leaves when plants are only a few inches tall and before stem becomes woody. Use leaves fresh or dry. Pick seed clusters as plants mature but before seeds begin to drop. Dry, thresh by shaking or trampling, winnow by shifting in breeze. Dry on tray at 190°C (375°F) for 40–50 min, stirring occasionally. Store in paper bag in dry place. **Preparation:** Fresh greens excellent as potherb. Boil for 10–20 min until tender and serve with seasonings and butter. For livelier flavor, boil with other greens such as mustard or dandelion, or add vinegar, bacon, hard-boiled eggs, or grated cheese. For salad, use fresh young leaves as major component of tossed salad. Use dried leaves in soups. Seeds edible whole or ground into meal using food mill or stones. For muffins, combine 1 cup pigweed meal, ½ cup whole-wheat flour, 2 tsp baking powder, ½ tsp salt, 3 tbsp honey, brown sugar, or molasses. Add 2 egg yolks, 1¼ cups milk, and ¼ cup butter. Blend until mix is moist. Beat 2 egg whites until stiff and fold into mix. Bake in greased muffin tin about 20 min at 190°C (375°F). Makes approximately 18. Thin batter with milk for pancakes. **Related edible species:** GREEN PIGWEED (next) and several other *Amaranthus* species. **Poisonous look-alikes:** Since pigweed is a concentrator of nitrates, plants from nitrate-fertilized areas should be eaten only in moderation.

Green pigweed *Amaranthus hybridus*

Also edible summer, autumn

Habitat: common and widespread in eastern and midwestern North America; fields, roadsides, disturbed sites. **Identification:** ***coarse annual herb*** to 2 m (6.6 ft) high; stems upright, with several upright branches. ***Leaves*** alternate, long stalked, broadest below or near middle, ***8–15 cm*** (3.2–6 in) ***long, 2–5 cm*** (0.8–2 in) ***wide,*** short-pointed at tip, wavy along margin. ***Flowers*** tiny, ***densely crowded in slender, green to red spikes.*** Fruits tiny, flattened, 1.5–2 mm (0.1 in) across, dry, ***partially enclosed by short, flexible bracts*** to 2 mm (0.1 in) long; seeds tiny, rounded, shiny black. **Harvest and preparation:** Same as for PIGWEED (preceding). **Related edible species:** PIGWEED and related species. **Poisonous look-alikes:** See PIGWEED.

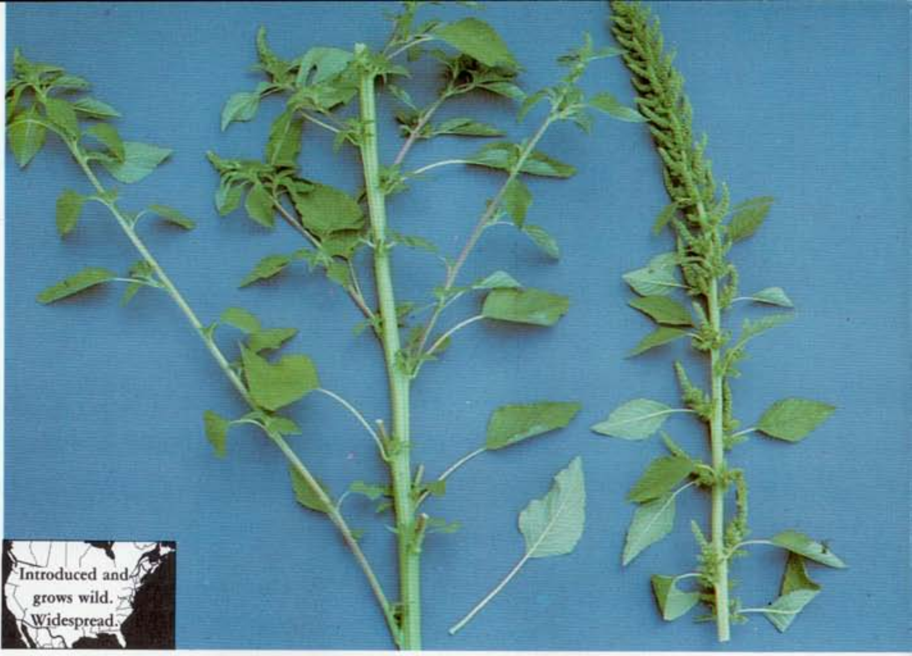
Introduced and
grows wild.
Widespread.

Pigweed: basal stalk and flowering head (USDA).

Green pigweed: fruiting stalks (Hardin).

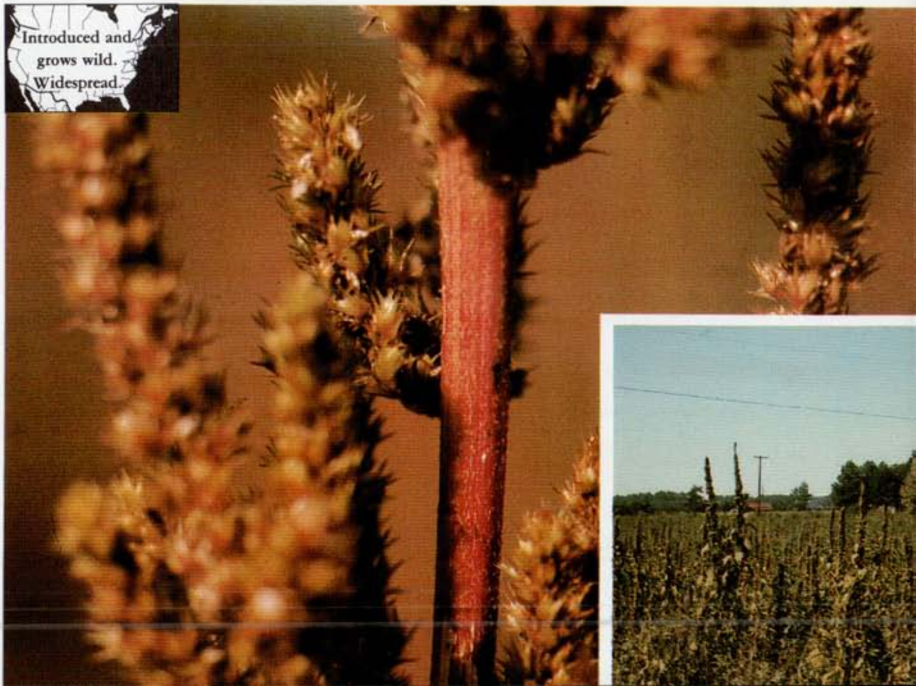
Introduced and
grows wild.
Widespread.

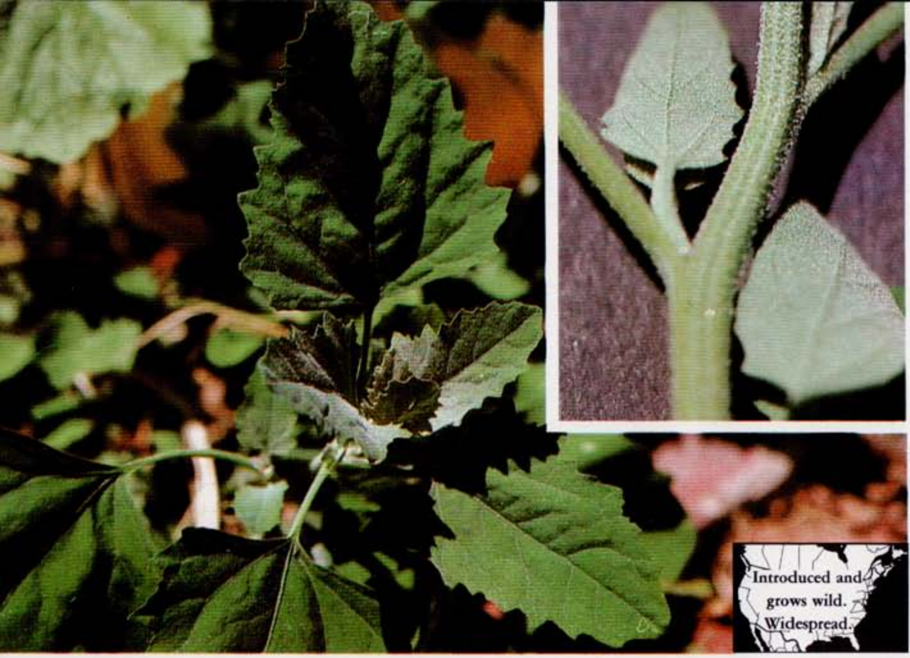
Introduced and
grows wild.
Widespread.

Lamb's quarters: basal leaves of young plant and mature stem (Dykeman).

Poverty weed: entire plant (Weber).

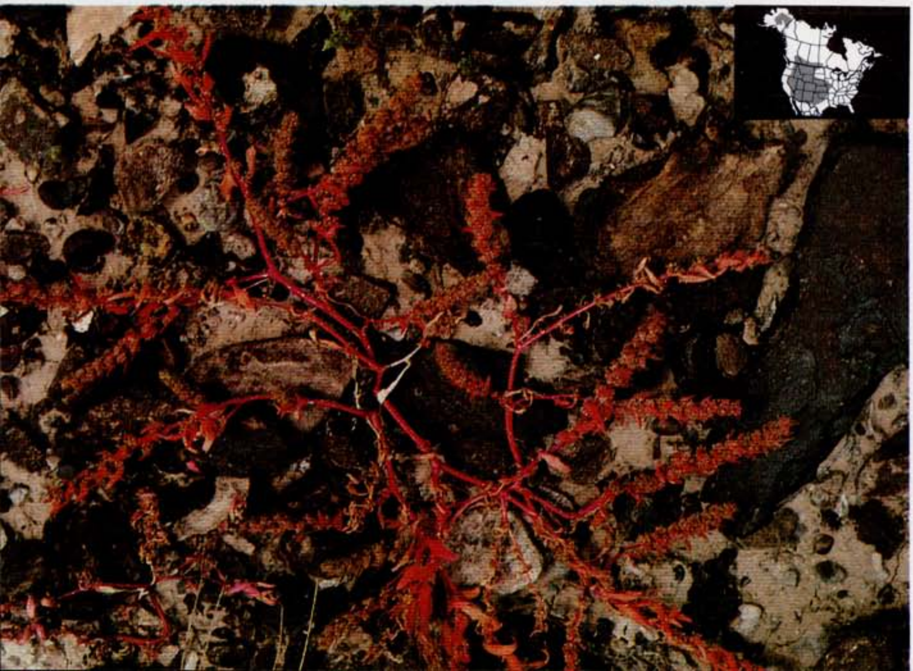

Lamb's quarters *Chenopodium album*

Also edible autumn

Habitat: common and widespread; yards, vacant lots, roadsides, waste areas, disturbed sites; introduced from Europe. **Identification:** small to medium-size annual ***herb*** to 0.6 m (2 ft) high; stems erect, usually branched, ***with a pale green or whitish cast. Leaves*** alternate, lower ones ***almost triangular,*** 3–10 cm (1.2–4 in) long, ***blunt at tip, coarsely toothed along margin, mealy white beneath,*** stalked, upper ones smaller. ***Flowers*** in late spring to autumn, ***tiny, inconspicuous, produced in dense short spikes in junctions of upper leaves and at tip.*** Fruits tiny, globe-shaped but depressed at tip, enclosed in persistent calyx lobes, in short dense spikes. **Harvest:** Pick young leafy stems to 25 cm (10 in) tall or tender growing tips of older plants. Gather abundant seeds in autumn by rubbing fruiting spikes into paper bag and winnow away chaff. **Preparation:** a superlative green vegetable, lacking strong flavor and high in vitamins A and C. Use leafy stems alone in salad or mix with stronger greens. For potherb, use large quantity of greens because cooking greatly diminishes bulk. Boil young leafy stems in small quantity of water about 5 min until tender. Add butter, salt and pepper or sauce of ¼ cup diced onion, 4 slices crisp bacon chopped fine, ¼ cup vinegar, salt and pepper, simmered gently. Use same sauce on raw greens as salad dressing. Soften seeds by boiling, crush, grind in food mill or blender. This produces nutritious black flour, good mixed with wheat flour for pancakes and muffins. **Related edible species:** Berlandier's lamb's quarters *(C. berlandieri)*, Fremont's lamb's quarters *(C. fremontii)*. **Poisonous look-alikes:** Species of *Chenopodium* that have bad odor and taste can be somewhat toxic.

Poverty weed *Monolepis nuttalliana*

Also edible autumn

Habitat: dry areas of high salinity or alkaline soils. **Identification:** ***annual herb*** to 55 cm (22 in) tall, spreading; branches numerous, stout, somewhat fleshy, upright or nearly so. ***Leaves*** alternate, ***triangular to lance-shaped,*** 1–4 cm (0.4–1.6 in) long, pointed at the tip, entire or irregularly toothed along the margin, sharply lobed at the base. ***Flowers small, numerous, inconspicuous, produced in junction of upper, smaller leaves, each bisexual, greenish, petals absent.*** Fruits are small, flattened, brown; winged seeds 1 mm (less than 0.1 in) across. **Harvest:** Pick young leafy stems. Roots are edible but usually not tasty. If you have the patience, strip seeds from stems when fully ripe late in growing season. **Preparation:** Like its relative, LAMB'S QUARTERS (preceding), which it superficially resembles, poverty weed is a good green vegetable. Boil young stems in small quantity of water until tender, but avoid overcooking. Serve with butter and seasonings or the following sauce hot: Simmer together for a few minutes ¼ cup dried onion, 4 slices crisp bacon (chopped fine), ¼ cup vinegar, salt and pepper. Southwestern Indians ground seeds with corn and mesquite beans to make pinole. **Related edible species:** Species of the *Chenopodium* genus (including LAMB'S QUARTERS) are of the same family as poverty weed. **Poisonous look-alikes:** none.

Glasswort *Salicornia europaea*

Also edible summer, autumn

Habitat: sea coast, salt marshes, alkaline soils. **Identification:** ***annual herb*** to 40 cm (16 in) high; ***stems*** erect to spreading, ***stiff, succulent*** (may be unbranched to many branched), green turning orange or red in autumn; leaves opposite, minute, scalelike, ***appearing leafless. Flowers*** (late summer) tiny, numerous, produced ***in erect*** to slightly spreading ***spikes 2–6 cm*** (0.8–2.4 in) ***long,*** joints longer than broad. Fruits are seeds, tiny, egg-shaped, 1.2–3 mm (0.1 in) in diameter. **Harvest:** Pick tender tips from branches midspring to autumn. **Preparation:** Tips are crispy and salty; good for trail nibble. Or chop and serve raw as salad with spicy vinegar dressing or add to other salad greens. For pickles pack jars with very tender, clean, raw tips or older tips that have been boiled in a little water for 5 min. Make pickling liquid from 4 cups cider or wine vinegar, 3 tbsp pickling spices, 1 grated onion, ½ cup honey or sugar, plus garlic, bay leaves, or other spices. Combine ingredients and boil 5–6 min. Cover glasswort with boiling liquid and seal jars. Store at least 1 month. For vegetable, place tips in boiling water and cook until tender 5–10 min. Serve buttered. **Related edible species:** Other species of *Salicornia* look very similar and are edible; one is perennial *(S. virginica)*. **Poisonous look-alikes:** none.

Russian thistle, Saltwort *Salsola kali*

Also edible summer

Habitat: sandy beaches on East Coast; dry, alkaline soils in West. **Identification:** ***annual herb*** to 1 m (3.2 ft) high; stems branching, upright to spreading, smooth, hairy, or with mealy appearance. ***Leaves alternate,*** simple, ***fleshy, rounded, 2–5 cm*** (0.8–2 in) ***long, lance- to needle-shaped, spine-tipped,*** entire along margins, lacking distinct leafstalk. Flowers numerous in narrow, elongated spikes in junctions of stems of upper leaves and at branch tips; each flower tiny, bisexual, sepals 5, winged on margins; petals absent. ***Fruits are flattened, winged, and nearly circular, 3–6 mm*** (0.1–0.2 in) ***across,*** yellowish to slate-colored. **Harvest:** Pull up shoots of new spring growth when less than 15 cm (6 in) tall; also succulent new plants after rains in late spring and in summer. Beware of spines at leaf tips. *Caution:* Plants occasionally contain toxic levels of nitrates and oxalates. **Preparation:** Cut off roots. Wash tops and boil for about 15 min. Serve like spinach with butter and seasonings, cream sauce, or vinegar. Mix with other greens of season. **Related edible species:** other *Salsola* species mainly in the Southwest. **Poisonous look-alikes:** none.

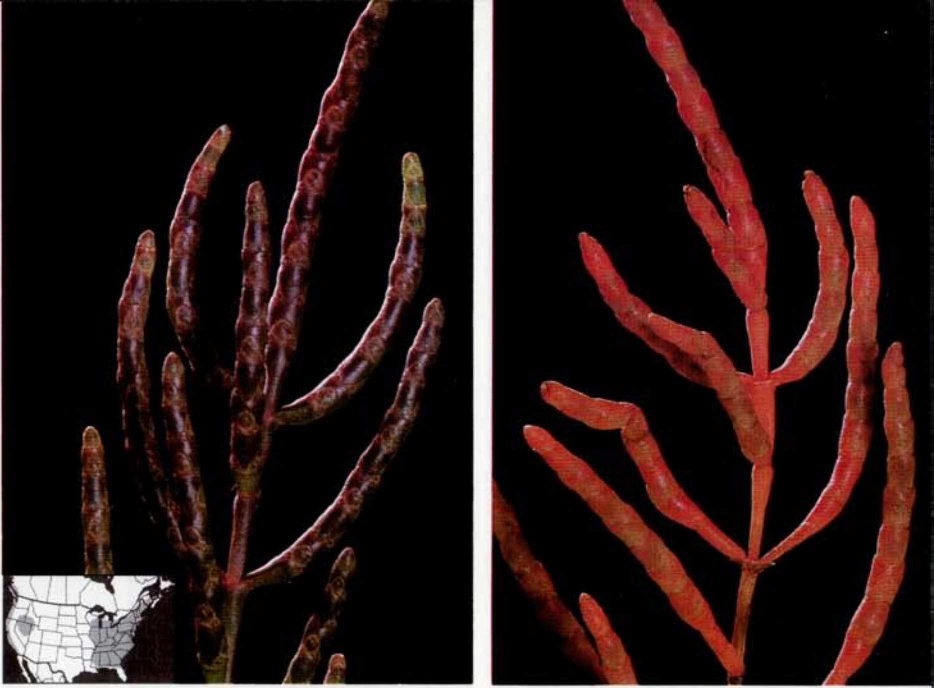

Glasswort: stems showing two color phases (NYBG).

Russian thistle: flowering branches (NYBG).

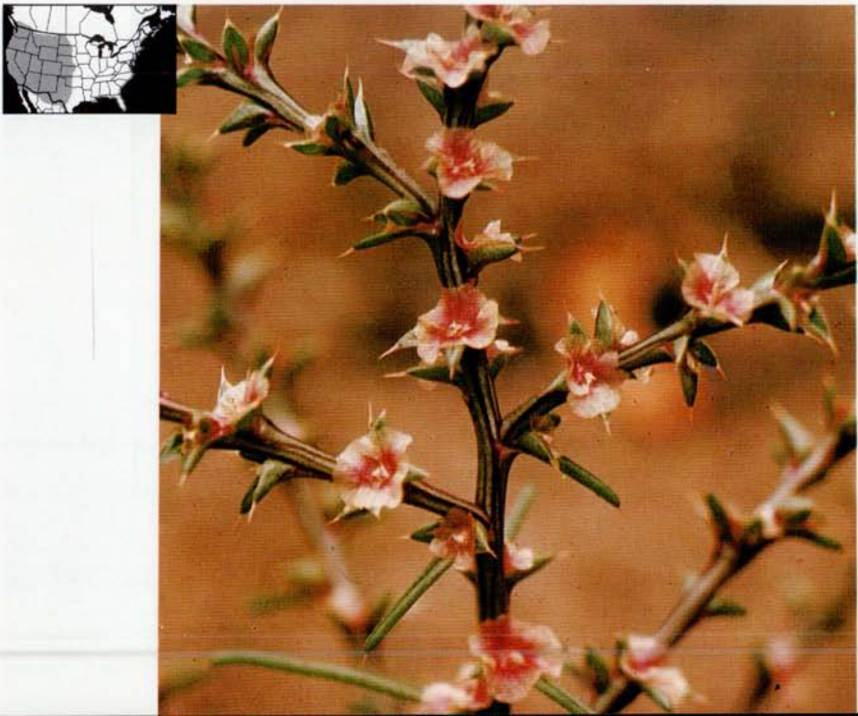

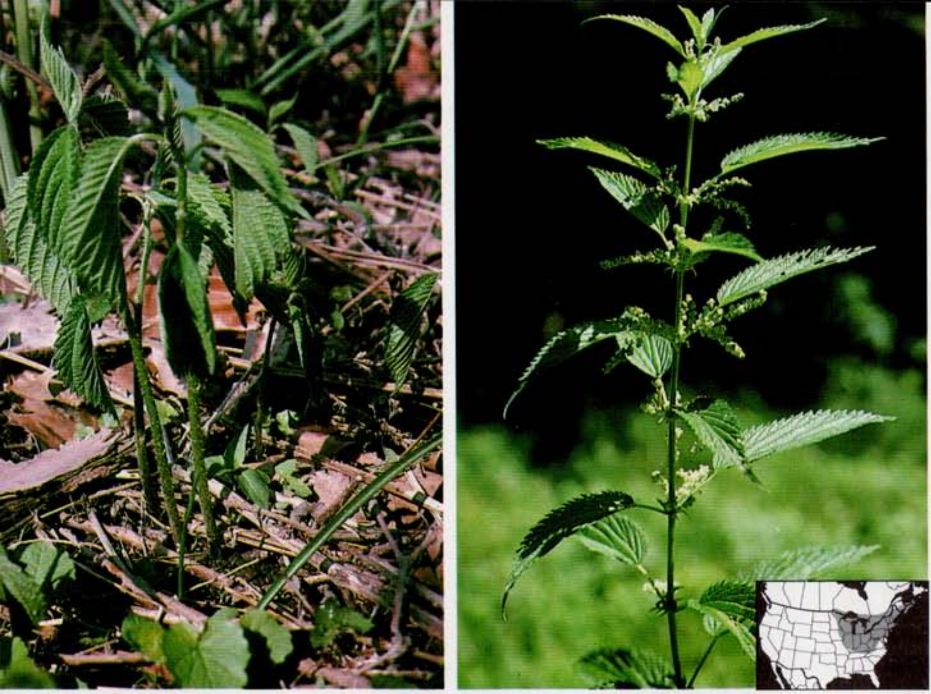

Stinging nettle: young plants and flowering plant (Elias).

Plantain: young plant (Dykeman). Mature plant (Elias).

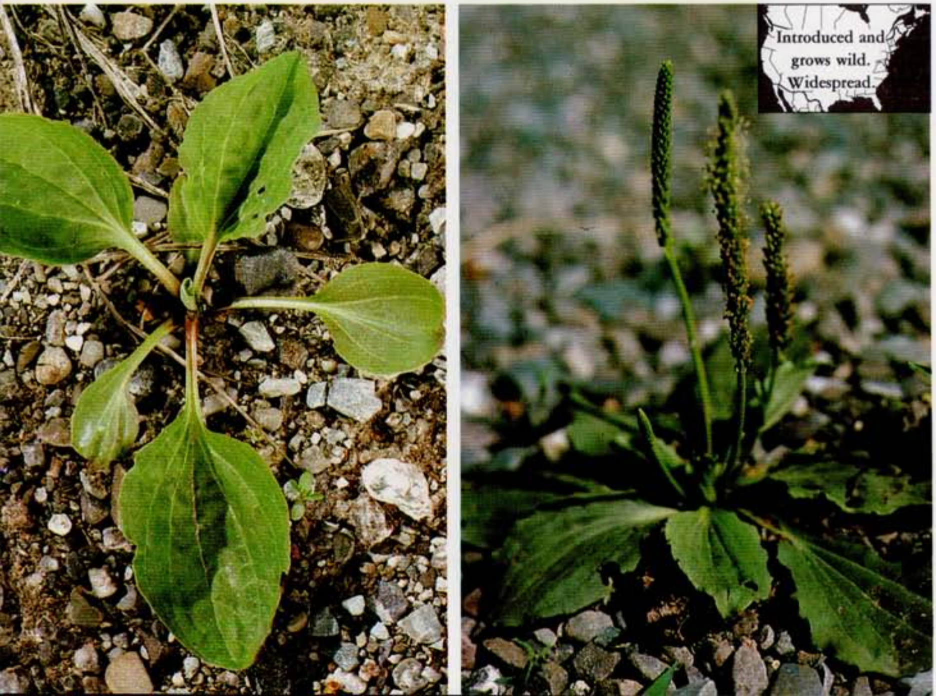
Introduced and grows wild. Widespread.

Stinging nettle, Nettle *Urtica dioica*

Habitat: widespread, especially in moist fertile soils in disturbed areas, margins of woods, trails, stream banks, roadsides, vacant lots. **Identification:** perennial, upright, to 1 m (3.2 ft); erect ***stem with bristly, stinging hairs. Leaves opposite, long-stalked,*** blade 5–14 cm (2–5.5 in) long, broadest near base, coarsely toothed, veins depressed, ***with stiff, stinging hairs on both surfaces.*** Each plant either male or female. Flowers produced in junctions of leafstalks, each flower tiny, male or female, borne in many-flowered, loosely branched, elongated, cream-colored clusters. Fruits tiny, inconspicuous, hard, nutletlike, about 2 mm (0.1 in) long. **Harvest:** In spring, wear gloves to avoid stinging hairs, or use scissors to cut young shoots and tender top leaves directly into pot or bag. Later leaves are tough and gritty. **Preparation:** Boiling destroys irritant. As potherb, barely cover with water and simmer until tender. Drain, add seasonings, butter, lemon juice, or wine vinegar. Use cooking water as beverage with lemon and sugar or as soup with salt, pepper, and vinegar. To make a simple stinging nettle puree, rub 2 cups cooked stinging nettles with juice through sieve, or puree in blender. Reheat, add 2 tbsp butter, salt to taste, simmer 10 min more. Stir in ¼ cup light cream; season with black pepper. A versatile plant, stinging nettle cooking juice may also be used as rennet (milk coagulant) for junket pudding and as beer ingredient. Plant is high in vitamins A and C, protein, and minerals. **Related edible species:** wood nettle *(Laporta canadensis),* slender nettle *(U. gracilis).* **Poisonous look-alikes:** none.

Plantain *(Also edible summer, autumn)* *Plantago major*

Habitat: yards, roadsides, disturbed sites; common weed introduced from Europe. **Identification:** ***perennial herb*** with fibrous root system. ***Leaves basal,*** numerous, broadest near or below middle, ***5–20 cm*** (2–8 in) ***long, 2.5–12 cm*** (1–4.7 in) ***wide,*** rounded to pointed at tip, toothed, entire or wavy along margin, ***rough on both surfaces,*** smooth or hairy. ***Flowers tiny, numerous along upper part of erect spike*** to 20 cm (8 in) high, each flower with 4 translucent petals. Fruits are tiny capsules produced along spike; each capsule contains 10–18 seeds. **Harvest:** Pick young leaves before flower stalk appears or pick new leaves through summer; otherwise leaves develop stringy veins that are tough to chew. Strip ripe seeds from flower stalks from late spring through fall. **Preparation:** Use tenderest leaves in salads with other greens of season. For cooked vegetable, soak leaves in salt water for 5 min. Boil with little water in covered pot until tender but not overcooked. For beverage, steep handful of leaves in pint of boiling water for ½ hr or longer. Dry seeds thoroughly and grind into flour. For pancakes, combine 2 cups plantain flour with 3 tsp baking powder, ½ tsp salt, 3 tbsp sugar, 2 eggs, 3 tbsp cooking oil, 1 cup milk. Mix thoroughly. Add more milk or flour for proper consistency; add fruit, if desired. Cook on hot griddle. **Related edible species:** Species of the *Plantago* genus are found throughout North America; most are edible. Boil leaves of bitter species in sweetened water. Seaside plantain *(P. juncoides)* is one of the best and is found only along East Coast. **Poisonous look-alikes:** none.

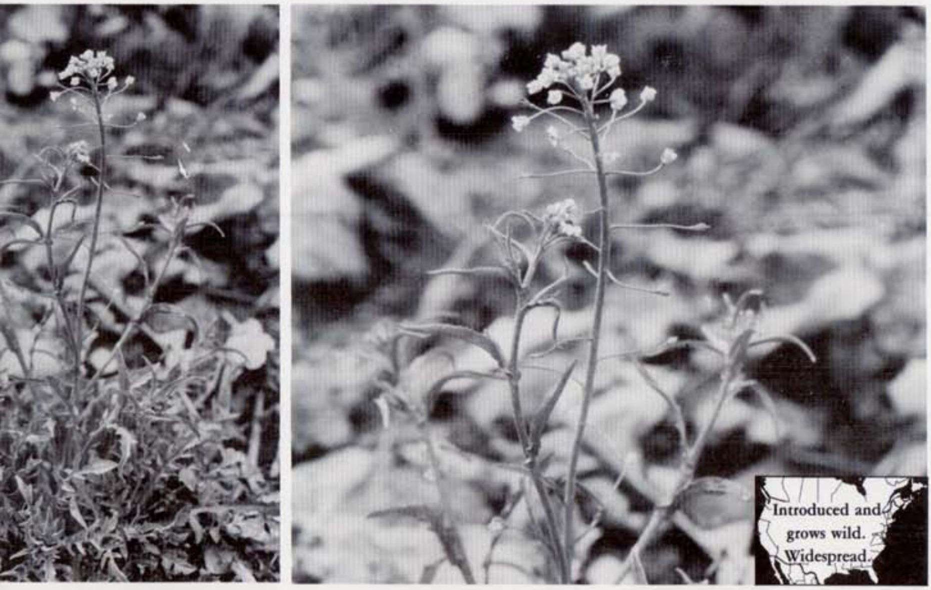
Introduced and
grows wild.
Widespread.

Shepherd's purse: entire plant and flowering stalks (Elias).

Toothwort: flowering plant (NYBG).

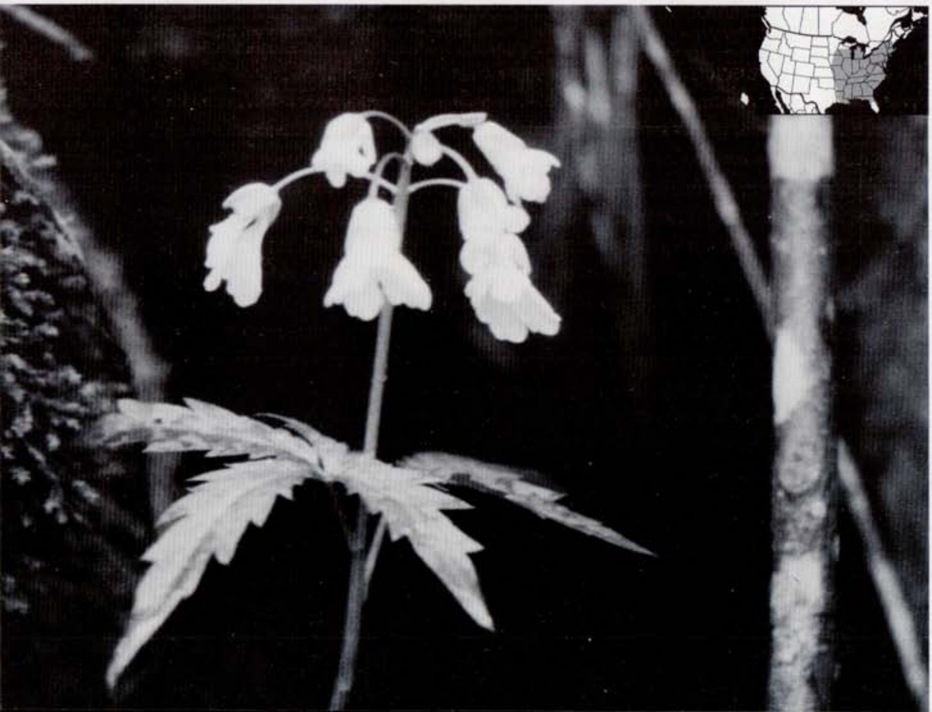

Shepherd's purse *Capsella bursa-pastoris*

Habitat: yards, vacant lots, roadsides, other disturbed areas; widespread weed, introduced from Europe. **Identification:** small to medium-size ***annual or biennial herb*** to 60 cm (2 ft) high, unbranched or sparingly branched, ***with a flattened, basal rosette of leaves,*** the ***basal leaves*** 4–10 cm (1.6–4 in) long, uniformly wide to broadest near tip, ***shallowly to deeply lobed along sides.*** Upright stems with few, smaller, narrower leaves. Flowers produced in narrow elongate cluster, each tiny, stalked, with 4 white petals 2–4 mm (0.1 in) long. ***Fruits slightly flattened, heart-shaped, notched at tip, 4–10 mm*** (0.2–0.4 in) ***long.*** **Harvest:** Gather leaves before flowering stage. If too peppery, bleach by covering with sawdust, leaves, or flower pots for week or more. Strip off seed pods as they ripen. **Preparation:** Use raw leaves as salad greens, if not too peppery, or with other milder greens. For a potherb, boil leaves in water for about 20 min with one change of water if needed. Season with salt and vinegar. Use dried pods and seeds as seasoning in soups and stews. Western Indians once ground seeds into meal, but process is tedious. **Related edible species:** red shepherd's purse *(C. rubella)* and slender shepherd's purse *(C. gracilis)*. **Poisonous look-alikes:** none.

Toothwort *Dentaria laciniata*

Habitat: moist deciduous woodlands, ravines, floodplains. **Identification:** perennial herb from a jointed, white underground stem (rhizome); stem upright, to 40 cm (15.7 in) high. Basal leaves absent at flowering, ***stem leaves in a whorl of 3 on upper half, deeply 3-parted, leaf segments lance-shaped, coarsely toothed to nearly entire,*** the terminal segment often deeply parted. ***Flowers*** produced in a terminal, elongate, hairy cluster, each flower stalked, ***with 4 white petals*** 1.2–2 cm (0.5–0.8 in) long. Fruits narrow, elongate, 2–4 cm (0.8–1.6 in) long, splitting open lengthwise at maturity to release the tiny seeds. **Harvest:** Dig rootstocks in small numbers from areas where plentiful. Other plants will grow from pieces of perennial root remaining in the soil. **Preparation:** Use as spicy trail nibble. Wash roots, chop and grind them in vinegar, add a dash of salt; use like horseradish. Add chopped roots to wild tossed green salads to liven flavor. **Related edible species:** At least six other species, similar in appearance, are distributed in moist woodlands of the East and West. All may be used like the species described. **Poisonous look-alikes:** none.

Peppergrass *Lepidium virginicum*

Habitat: fields, vacant lots, roadsides, other disturbed areas, waste sites. **Identification:** annual or biennial ***herb,*** upright to 50 cm (19.7 in) high, occasionally taller. ***Basal rosette of leaves 4–10 cm*** (1.6–3.9 in) ***long,*** 0.8–2 cm (0.3–0.8 in) wide, broadest near tip, ***doubly toothed, lobed, or deeply dissected*** along margin; stem leaves smaller, toothed or entire, becoming linear near top. ***Flowers*** produced in several, narrow, elongate clusters; each flower stalked, ***with 4 small white petals. Fruits flattened, widest near middle to almost circular, 2–4 mm*** (0.1 in) ***long,*** nearly as wide, ***slightly notched at tip,*** containing tiny reddish-brown seeds. **Harvest:** young shoots and leaves in spring. **Preparation:** another member of the mustard family. Use peppergrass greens sparingly in salad with less bitter species. As a potherb boil in 1 or 2 changes of water. Try recipes given for other MUSTARDS. Leaves high in vitamins A and C. Peppery seeds can be used to season meats, soups, and salads. **Related edible species:** Fremont's peppergrass *(L. fremontii)* and FIELD CRESS (next). **Poisonous look-alikes:** none.

Field cress, Cow cress *Lepidium campestre*

Habitat: fields, vacant lots, roadsides, disturbed sites; introduced from Europe. **Identification:** annual or biennial ***herb,*** upright to 60 cm (2 ft) high, ***stems densely hairy.*** Basal leaves 4–8 cm (1.6–3.2 in) long, 1–1.6 cm (0.4–0.6 in) wide, narrow but broadest near tip, coarsely toothed, lobed or dissected along margin; ***stem leaves*** smaller, almost circular in shape, ***lobed at base. Flowers*** produced in several, narrow, elongate clusters; each flower stalked, ***with 4 small, white petals. Fruits flattened,*** widest near tip to ***almost round in outline, 2–3 mm*** (0.1 in) ***long,*** slightly notched at tip, narrowly winged, containing tiny dark brown seeds. **Harvest and preparation:** same as for PEPPERGRASS (preceding). **Related edible species:** PEPPERGRASS and related species. **Poisonous look-alikes:** none.

Field pennycress *Thlaspi arvense*

Habitat: fields, vacant lots, waste areas, other disturbed sites; more common in northern portion of its range. **Identification:** ***annual herb*** 40–80 cm (16–32 in) high; stems branched, upright, smooth. Basal leaves narrow, widest near tip, withering and absent at flowering. ***Stem leaves 2–10 cm*** (0.8–4 in) ***long,*** 0.4–3 cm (0.2–1.2 in) wide, ***lance-shaped*** to narrow but broadest near tip, wavy or coarsely toothed or entire along margin, ***lobed at base, lobes pointed. Flowers*** numerous in narrow, elongate clusters (racemes), each flower stalked, ***with 4 white petals. Fruits flattened, nearly circular in outline, 0.8–1.8 cm (0.3–0.7 in) in diameter, deeply notched at tip, broadly winged,*** containing several tiny black seeds. **Harvest:** young shoots and leaves in spring. **Preparation:** Since plant has characteristic bitterness of MUSTARD family, boil shoots and leaves in 1 or 2 changes of water for 15–25 min. Serve with butter, salt and pepper. Use tender greens if not too bitter with other species in tossed salads. **Related edible species:** pennycress *(T. perfoliatum)* and many other members of mustard family Cruciferae. **Poisonous look-alikes:** none.

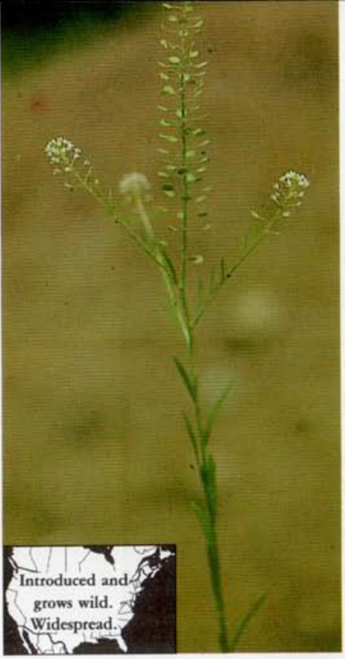
Introduced and grows wild. Widespread.

Peppergrass: fruiting plant (Elias).

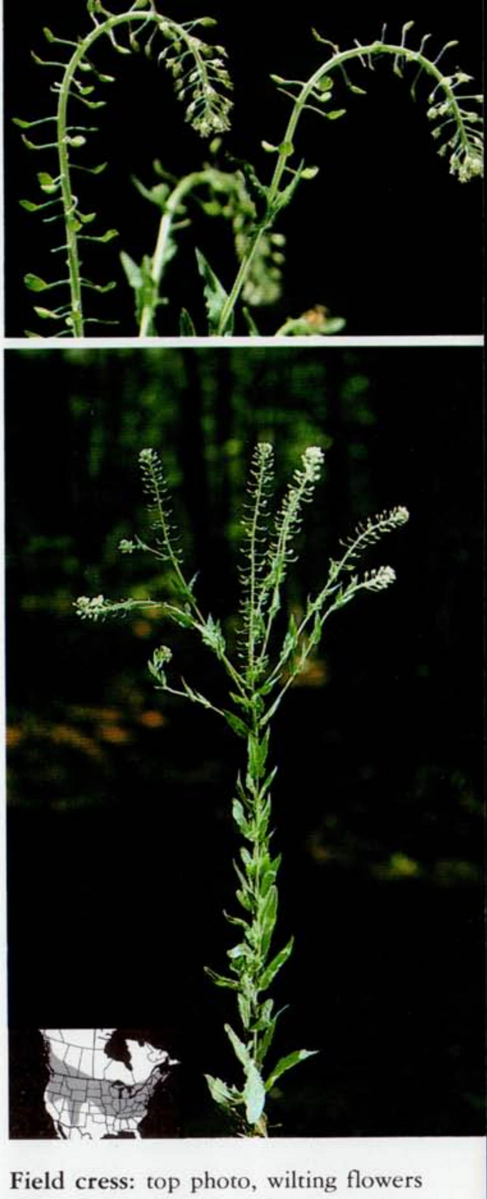

Field cress: top photo, wilting flowers after being picked and entire plant (Dykeman).

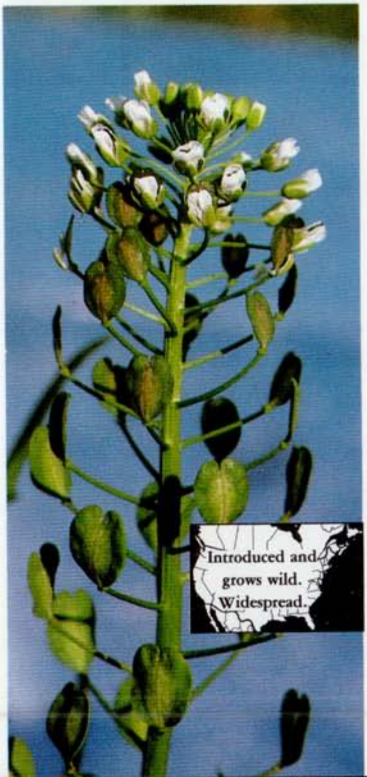
Introduced and grows wild. Widespread.

Field pennycress: flowers, top, and fruits on lower portion of flowering cluster (Dykeman).

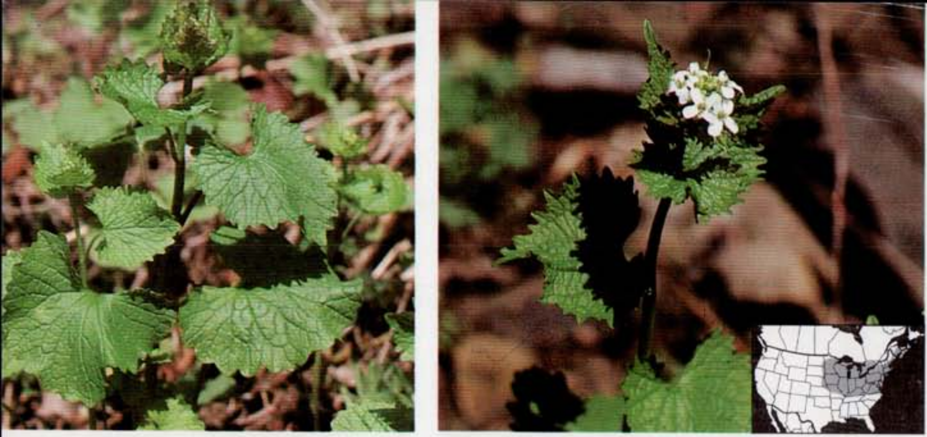

Garlic mustard: young plants and flowering plant (Dykeman).

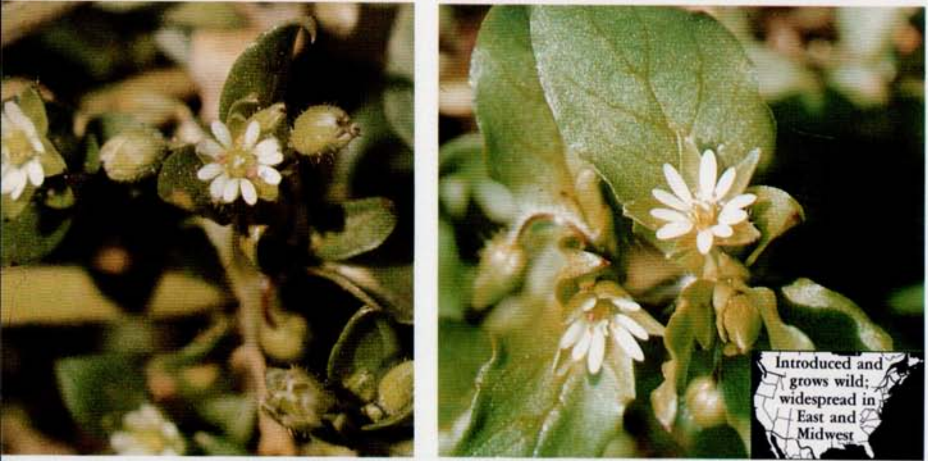
Introduced and grows wild; widespread in East and Midwest

Common chickweed: flowering plants (NYBG).

James chickweed: flowering branch and flower (NYBG).

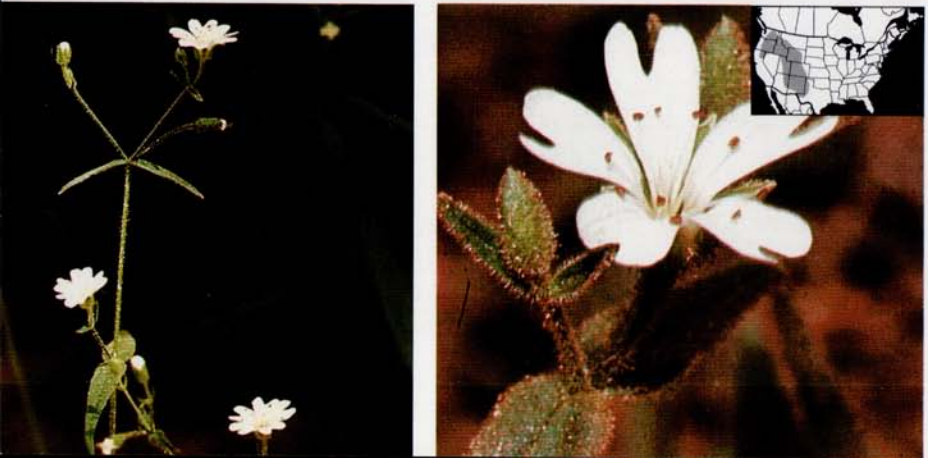

Garlic mustard *(Also edible summer)* *Alliaria officinalis*
Habitat: partially shaded roadsides, paths in open woods, shady disturbed areas. **Identification:** ***biennial herb, 2nd year growth to 90 cm (3 ft). Leaves triangular to heart-shaped, coarsely toothed, long-stalked, with garlic odor when crushed. Flowers white,*** 6–8 mm (0.2–0.3 in) across, in ***clusters at tips of stems and at leaf bases,*** April–June. Fruits are very ***slender pods (siliques) to 5 cm (2 in) long,*** containing numerous tiny seeds. **Harvest:** young leaves, blossoms, and seed pods in spring and early summer. **Preparation:** chop tender growth raw in salad; for cooked vegetable steam or boil in little water, season with salt and butter. Use for garlic flavor with meats and vegetables. **Related edible species:** no other *Alliaria* species present; many other members of MUSTARD family *(Cruciferae)* are edible. **Poisonous look-alikes:** none.

Common chickweed *(Also edible summer, autumn, winter)* *Stellaria media*
Habitat: yards, disturbed sites, moist and wet areas, cultivated land, woodlands; common weed introduced from Europe. **Identification:** annual herb sometimes to 30 cm (12 in) high; stems weak, trailing along ground before becoming upright, with rows of hairs. ***Leaves opposite,*** simple, ***1–3.5 cm*** (0.4–1.4 in) ***long, about ⅔ as wide,*** usually broadest near base or middle, pointed at tip, entire along margin, rounded at base, lower leaves with a conspicuous leafstalk. ***Flowers single*** or in few-flowered clusters at tips, ***long-stalked, each flower with 5 deeply lobed, white petals that are shorter than sepals.*** Fruits are small reddish-brown capsules 5–8 mm (0.2–0.3 in) long, egg-shaped, containing many small seeds. **Harvest:** Pick tender leaves and stems from young plants any time they are not blooming; available year-round in some regions. Quicker method, pull up the whole plant and cut off tender growth with scissors. **Preparation:** Mild flavored chickweed is excellent by itself or with other, stronger greens. Use generously raw in salads. Boil only 2–5 min; add it to other less delicate greens near end of their cooking period. Serve with butter, seasonings, and a little chopped onion. Also see JAMES CHICKWEED (next). **Related edible species:** JAMES CHICKWEED. **Poisonous look-alikes:** none.

James chickweed *(Also edible summer, autumn)* *Stellaria jamesiana*
Habitat: mountain meadows, seepage areas, woodlands, usually between 1,220–2,590 m (4,000–8,500 ft) elevation. **Identification:** ***small perennial herb*** to 35 cm (14 in) tall from slender roots with thickened tuberlike structures; stems erect or nearly so, branched frequently, hairy. ***Leaves opposite,*** simple, ***4–15 cm*** (1.6–6 in) ***long, about one-fourth as wide, lance-shaped,*** long pointed at tip, entire along margin. ***Flowers produced in loose, spreading clusters at tips*** and at junctions of upper leaves in spring and early summer. ***Each flower with 5 deeply notched white petals.*** Fruits are small tan capsules, 3–5 mm (0.2 in) long, containing many tiny, dark brown seeds. **Harvest:** same as COMMON CHICKWEED (preceding). **Preparation:** Add chickweed to soups and stews. For pancakes, drain and blot 1–2 cups leaves. Blend into prepared pancake batter (2 cups mix to 1½ cups water). Adjust consistency. Fry and serve with butter and tart jelly. **Related edible species:** COMMON CHICKWEED, star chickweed *(S. pubera),* and mouse-ear chickweed *(Cerastium vulgatum)*. **Poisonous look-alikes:** none.

Cleavers, Goosegrass, Bedstraw *Galium aparine*

Also edible summer

Habitat: wide spread on moist sites, rich woodlands, along streams, lakes, vacant lots and fields. **Identification:** ***annual herb;*** stems weak, slender, often leaning on other plants or rambling along ground, 4-sided, with numerous tiny, stiff, hooked hairs causing plants to cling to clothing; ***leaves in whorls of 8,*** each very narrow but ***widest above middle,*** 4–8 cm (1.6–3.2 in) long, 2–6 mm (0.1–0.2 in) wide, with recurved hairs along margin. Flowers tiny, white, produced in usually 3-flowered, stalked clusters in junction of upper leaves in late spring or early summer. ***Fruits consist of twin, fused, globe-shaped, dry structures 2–4 mm (0.1 in) long, covered with numerous hairs hooked at tip.*** **Harvest and preparation:** Pick young greens in spring when tender. Steam leafy stems in strainer or colander over boiling water for 5–10 min or boil in small amount of water until tender. Serve with butter and seasonings. Combines well with nettle greens. For salad take 3 cups boiled, cooled cleavers and add 1 cup cooked asparagus cut in 1 in lengths, ½ cup slivered nuts, ¾ cup French dressing. Gather fruits in June and July as they turn brown. For excellent noncaffeine coffee substitute, roast fruits in warm oven 150°C (300°F) until dark and crisp, about 1 hr. Grind in blender or crush on any hard surface. Simmer ¾ cup prepared fruits in 1 qt water until medium to dark brown. Strain and serve. **Edible related species:** NORTHERN BEDSTRAW (next). *Note:* The many species of cleavers or bedstraw in North America are difficult to tell apart; however, none of them is known to be poisonous. **Poisonous look-alikes:** none.

Northern bedstraw *Galium boreale*

Habitat: stream banks, fields, open woodlands. **Identification:** ***perennial herb;*** stems usually upright, to 1 m (3.2 ft), slender, 4-sided, smooth. ***Leaves in whorls of 4, each narrowly lance-shaped, 2.5–5 cm*** (1–2 in) ***long,*** 2–7 mm (0.1–0.3 in) wide, with 3 main veins, finely hairy along margins. Flowers tiny, white, produced in compact many-flowered, branched, elongated clusters (inflorescences) at or near growing tips. ***Fruits consist of 2 fused, globe-shaped structures (rarely 1) that are 2–3 mm (0.1 in) in diameter, smooth or with long straight hairs or shorter ones hooked at tip.*** **Harvest and preparation:** same as for CLEAVERS (preceding). **Edible related species:** See CLEAVERS. **Poisonous look-alikes:** none.

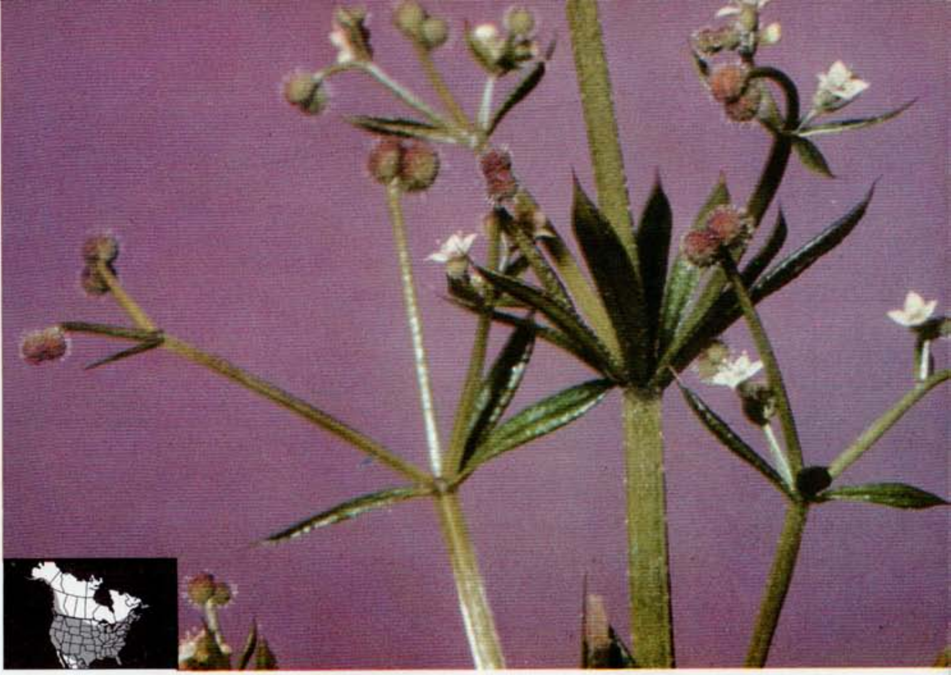

Cleavers: flowering stalks (C. S. Webber, Jepson Herbarium).

Northern bedstraw: flowering stalks (NYBG).

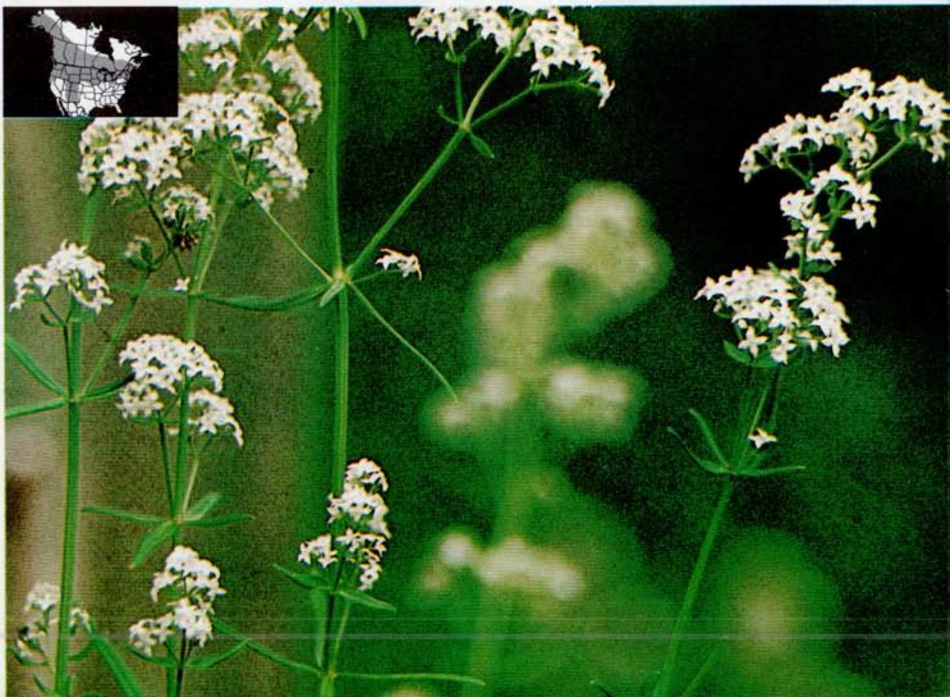

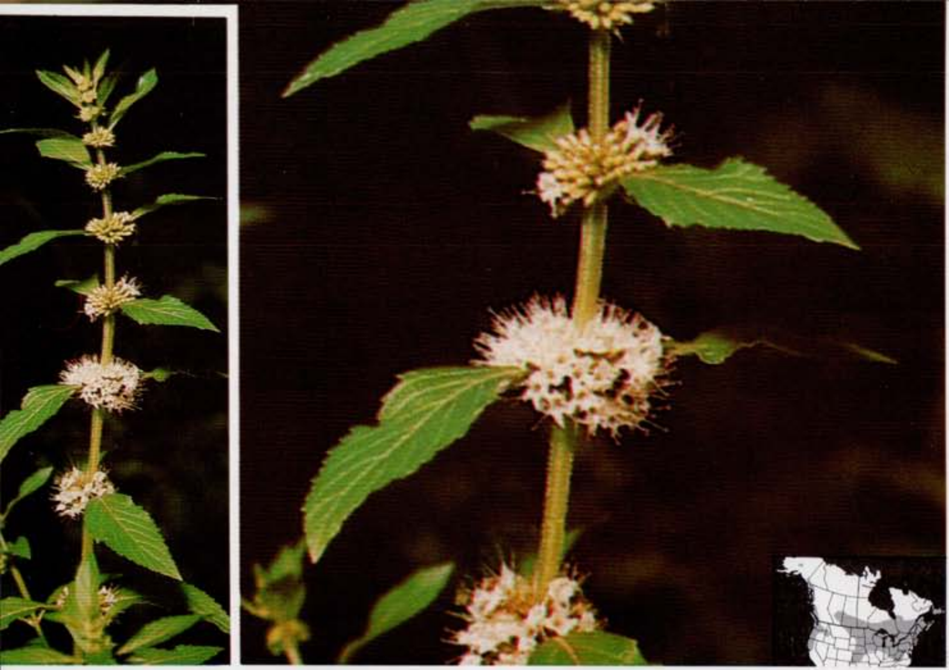

Bugleweed: flowering stalk and flowers (NYBG).

Edible valerian: flowering stalks (NYBG).

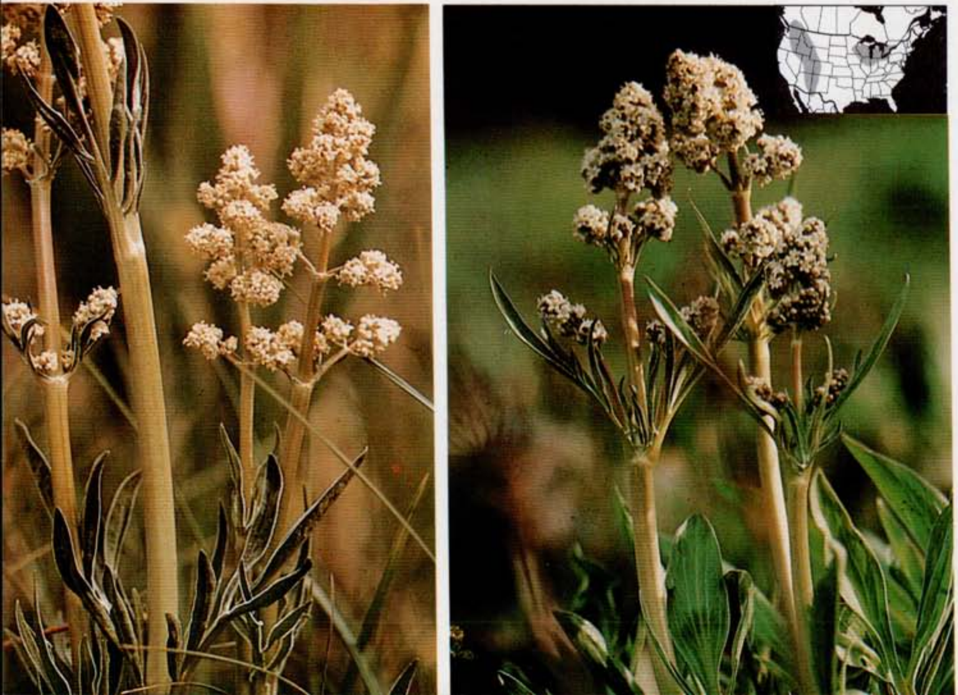

Bugleweed *Lycopus uniflorus*

Also edible autumn, winter

Habitat: low wet or boggy ground in North, wet woodlands and mountain bogs in South. **Identification:** ***perennial herb*** to 1 m (3.2 ft) tall; stems upright, unbranched or sparingly branched, finely hairy. ***Leaves opposite,*** simple, 3–10 cm (1.2–4 in) long, 0.6–3.4 cm (0.2–1.3 in) wide, ***lance-shaped*** to almost uniformly wide, pointed at tip, ***with few teeth along margin,*** tapering at base, light green and papery; leafstalks absent. ***Flowers produced in compact, many-flowered clusters in junctions of upper leaf stems;*** each flower tiny, ***petals white,*** fused, with 2 upper and 3 lower lobes. Fruits are tiny, flattened, irregular nutlets 1–2 mm (less than 0.1 in) long, with small bumps (tubercles) on top. **Harvest:** Tubers grow best close to the base of old stems in loose sandy soil. Dig tubers from fall through early spring when they are highest in stored nutrients. **Preparation:** Add peeled, sliced tubers to tossed salad or chop fine as mild relish. Boil tubers in salted water until tender, about 10–15 min. Serve with butter and seasonings. For pickles, peel tubers, boil in water about 5 min, drain and cover with wine vinegar. Store for several weeks before using. **Related edible species:** sessile-leaved water horehound *(L. amplectens)* is found in sandy East Coast areas. Other species of *Lycopus* lack tubers. **Poisonous look-alikes:** none.

Edible valerian *Valeriana edulis*

Also edible autumn, winter

Habitat: open moist sites, meadows, prairies, ditches, swamps. **Identification:** ***perennial herb*** to 1.2 m (3.8 ft) from large, elongated root. ***Leaves opposite,*** featherlike (pinnate) divided; ***basal leaves spoon-shaped,*** widest near tip, 12–30 cm (4.7–12 in) long, ***entire or sometimes featherlike (pinnate) divided*** with narrow lateral segments, thick, hairy along margins; stem leaves pinnately divided into few narrow segments, smaller. ***Flowers tiny, produced in*** tightly branched, elongated, ***many-flowered clusters*** at tips in late spring; each flower white to pink, 3–5 mm (0.1–0.2 in) long. Fruits dry, flattened, 1-seeded, egg-shaped, 3–4 mm (0.2 in) long. **Harvest and preparation:** Dig roots in autumn and early spring, when firm. Steam 24 hr or longer to remove disagreeable odor. Then use in soups as you would potato or squash. Dry and grind steamed roots into flour for use in bread, as did the Indians. Bake Indian pit style 2 days. **Related edible species:** Other valerians have been used in folk medicine. **Poisonous look-alikes:** toxicity of look-alikes not well documented.

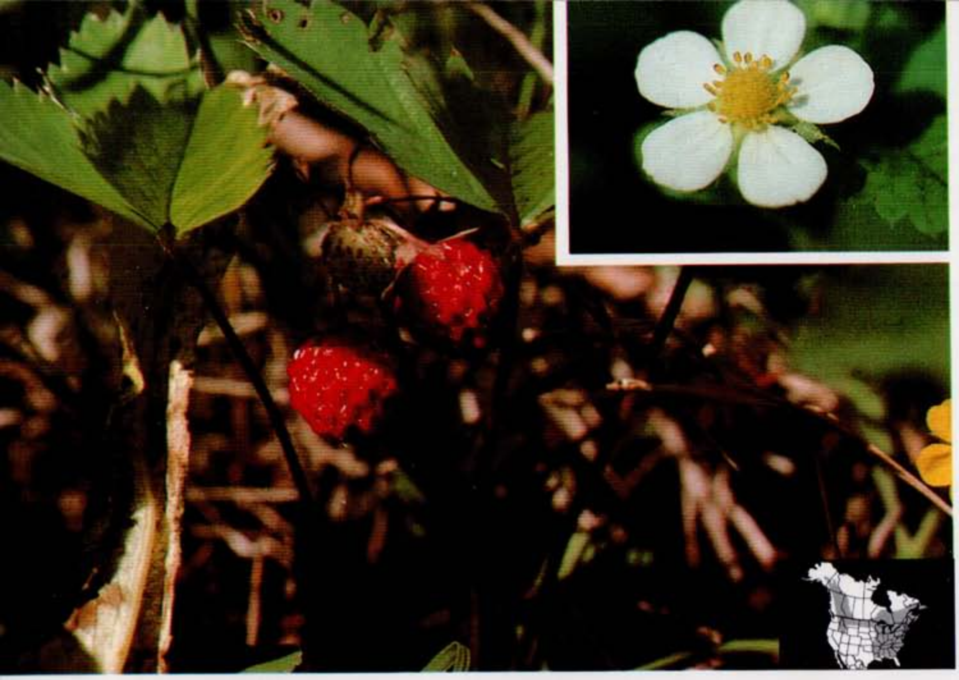

Wild strawberry: mature fruits and flower (Elias).

Spring beauty: flowering stalk (NYBG).

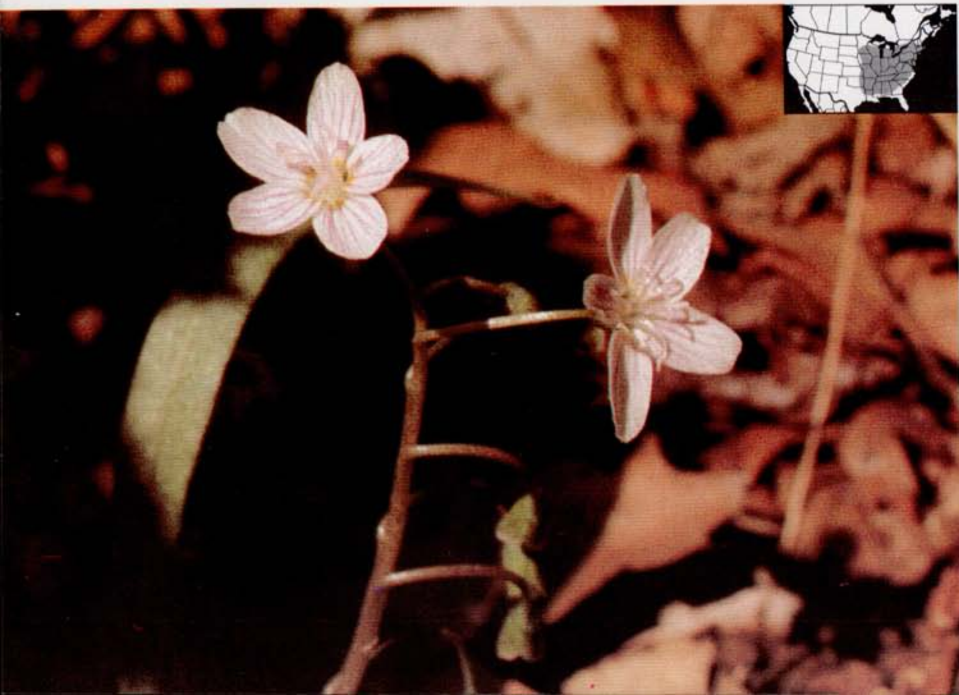

Wild strawberry *Fragaria virginiana*
Also edible summer

Habitat: open woodlands, margins of woodlands, fields. **Identification:** small perennial herb, lacking distinct stem, sometimes producing runners. ***Leaves basal, compound and 3-parted*** (trifoliolate), leaflets widest near or above middle, 2–10 cm (0.8–4 in) long, sharply toothed along margin, hairy beneath. ***Flowers*** several in short spreading cluster, each ***with 5 white petals*** 8–14 mm (0.3–0.5 in) long. ***Fruits are tiny, hard (achenes), embedded in large, fleshy tissue (receptacle)*** that is 0.5–2 cm (0.2–0.8 in) in diameter, ***red, sweet.*** **Harvest:** one of the best wild fruits. Locate your site early and check regularly from early June for ripe berries. Collect leaves anytime but best in late summer. **Preparation:** Use wild strawberries in any strawberry recipes. For quick beverage, crush handful of berries and mix with 1 cup water and a little sugar. For fruit leather, add ½ tbsp sugar or honey to each cup of fruit. Bring just to boil, cool, and process through food mill. Pour no more than 5 mm (0.2 in) thick onto Teflon cookie sheet. Heat in oven at lowest setting for 6–7 hr. When leathery, roll and seal in plastic wrap. Stores well at any temperature. See basic pie and jam recipes in book's introductory material. For tea high in vitamin C, steep 2 handfulls fresh or ½ cup dried leaves in 1 qt boiling water for 5 min. To dry leaves, spread on clean surface in warm, dry area (sun or shade). When thoroughly dried (use warm oven, if necessary), crumble leaves and store in paper bag in dry place. **Related edible species:** Several are widely distributed, basically alike, and easy to identify. **Poisonous look-alikes:** none.

Spring beauty *Claytonia virginica*

Habitat: rich woodlands, wooded floodplains, wet fields. **Identification:** ***perennial herb*** to 40 cm (16 in) tall, ***from globe-shaped, bulblike tuber 1–5 cm*** (0.4–2 in) ***in diameter;*** stems several, upright, weak, sometimes leaning. ***Leaves*** simple, ***5–14 cm*** (2–5.5 in) ***long, 0.2–2 cm*** (0.1–0.8 in) ***wide, linear*** to narrow but widest above the middle, blade gradually merging with leafstalk. ***Flowering stem with 5–20 weakly stalked flowers, each flower white to pink with dark pink veins, petals 5,*** 1.2–2 cm (0.5–0.8 in) across. Fruits are egg-shaped to nearly globe-shaped capsules 5–10 mm (0.2–0.4 in) long, containing numerous, tiny, dark red seeds. **Harvest:** Because this is a beautiful spring flower, use only where very abundant or when desperately needed. In spring, pick young leaves and dig tubers when identifiable by blossoms. Keep only largest tubers; replant others. **Preparation:** Use leaves as trail nibble or chop a few in green salads. Boil leaves for a few minutes in salted water for edible but not exceptional vegetable. Eat washed tubers raw or add to salads. Boil tubers with jackets in salted water about 10–15 min. Peel and serve whole or mashed, fried, or in salad like potatoes. Tubers also good baked; peel before eating. **Related edible species and poisonous look-alikes:** See ALPINE SPRING BEAUTY (next).

Alpine spring beauty *Claytonia megarhiza*

Habitat: rocky crevices, rock slides in alpine zone, tundra. **Identification:** ***low perennial herb*** from thick, fleshy, purple to dark-red taproot. ***Basal leaves spreading, forming large rosette,*** leaves 2–15 cm (0.8–6 in) long, ***broadest near tip to almost circular in outline,*** pointed to rounded at tip, entire along margin, gradually tapering and winged at base; stem leaves alternate, linear to narrow and broadest near tip. ***Flowers*** stalked, produced in elongate cluster; each flower 1.2–2.5 cm (0.5–1 in) across, ***with 5 white to pink petals with red veins,*** 6–10 mm (0.2–0.4 in) long. Fruits are tiny, 3-parted, egg-shaped capsules. **Harvest:** Because of habitat and scarcity, use as survival food only. Pick rosettes of leaves or pull up whole plant for fleshy root. **Preparation:** similar to SPRING BEAUTY (preceding). Peel roots before boiling to avoid strong flavor and use like tubers of SPRING BEAUTY. **Related edible species:** SPRING BEAUTY, WESTERN SPRING BEAUTY (next); also Carolina spring beauty *(C. caroliniana),* widely distributed in northern U. S. **Poisonous look-alikes:** To avoid confusion with poisonous lilies, learn to identify leaves or collect when in bloom.

Western spring beauty *Claytonia lanceolata*

Habitat: mid to high mountain slopes, moist woodlands to alpine zone. **Identification:** ***perennial herb*** to 20 cm (8 in) high, ***from nearly globe-shaped,*** bulblike underground ***stem base*** (corm) 1–2 cm (0.4–0.8 in) in diameter; stems 1 to several, upright. Basal leaves 1 or 2, 4–8 cm (1.6–3.2 in) long, broadly lance-shaped; ***stem leaves 2, opposite, 3–6 cm*** (1.2–2.4 in) ***long,*** narrow to broadly lance-shaped, 0.8–2 cm (0.3–0.8 in) wide. ***Flowering stems*** upright, ***with 3–15 stalked flowers, each bowl-shaped, with 5 pink petals*** 0.8–1.2 cm (0.3–0.5 in) long. Fruits are egg-shaped capsules 3–5 mm (0.1–0.2 in) long, containing a few shiny, black seeds. **Harvest and preparation:** Because of habitat and scarcity, use as survival food only. Similar to SPRING BEAUTY. Leaves are edible but not exceptional; boil 10–15 min. Crispy but small tubers are good raw or cooked for about 25 min at about 1,640 m (5,400 ft) altitude. **Related edible species and poisonous look-alikes:** See ALPINE SPRING BEAUTY (preceding).

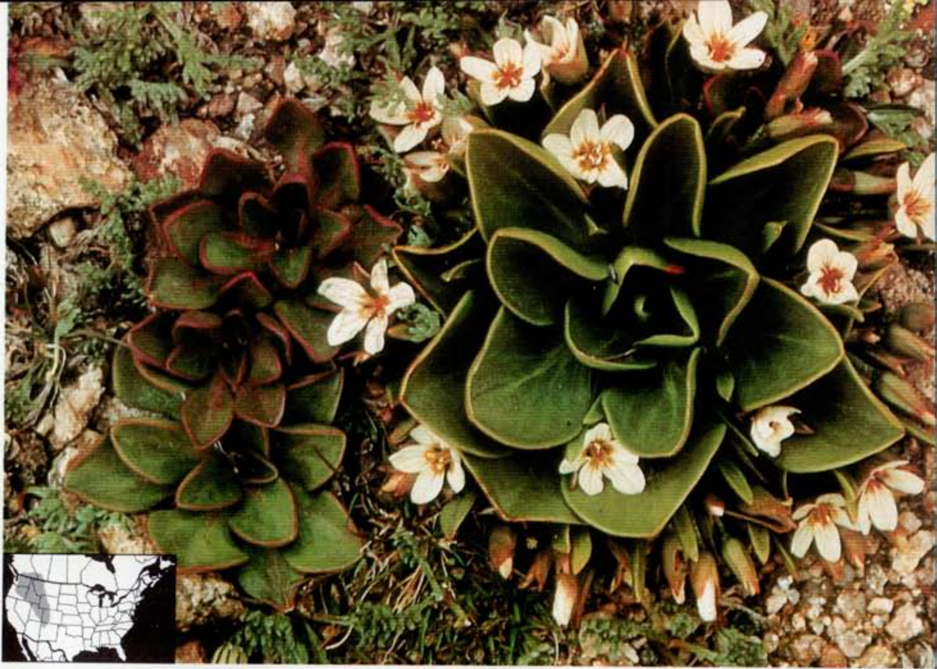

Alpine spring beauty: flowering plant (NYBG).

Western spring beauty: flowering plant (NYBG).

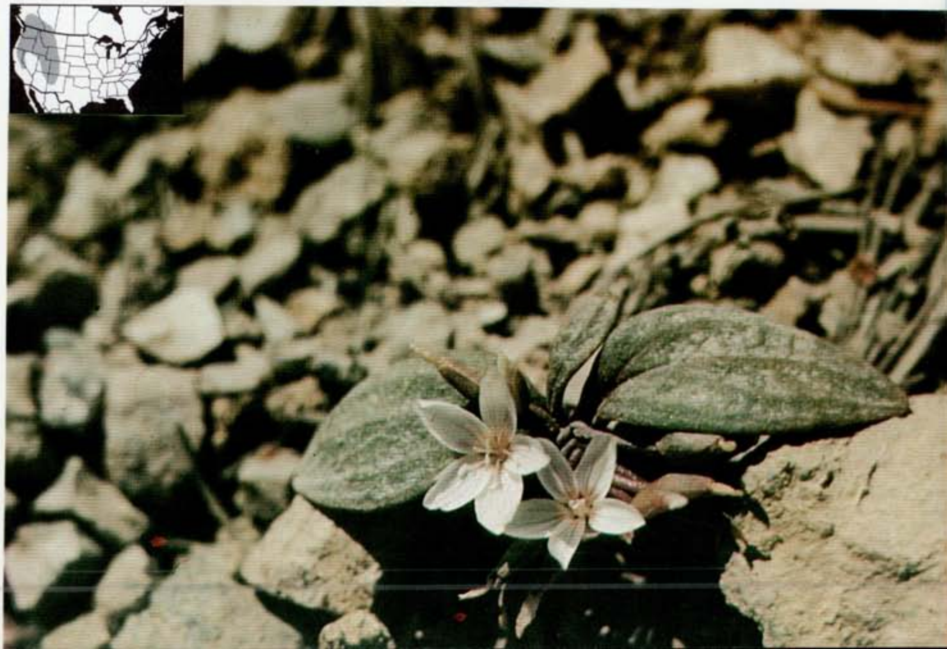

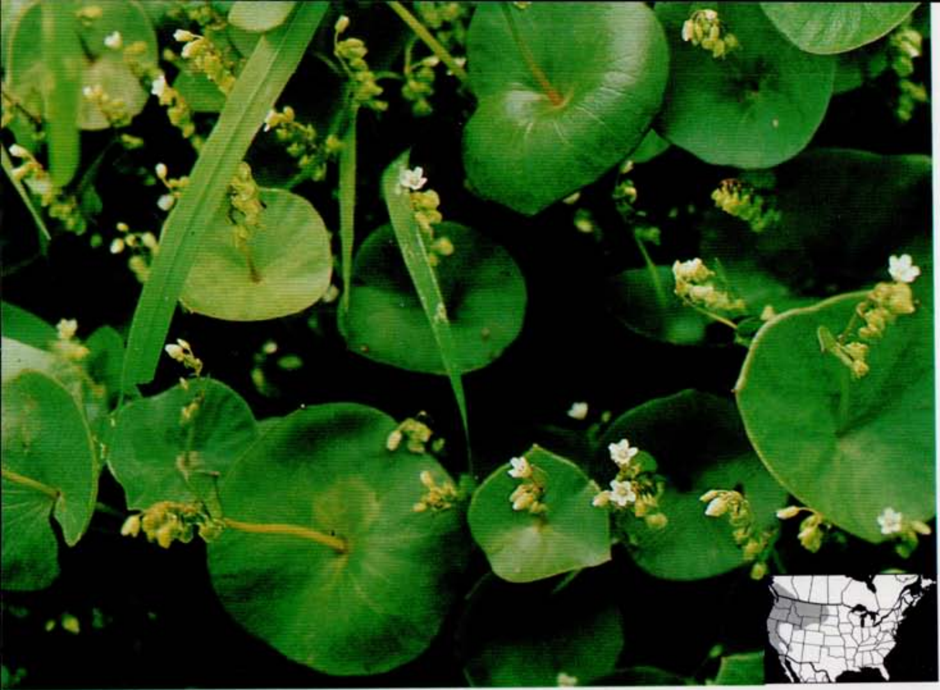

Miner's lettuce: flowering plants (NYBG).

Canada violet: flower and flowering plant (NYBG).

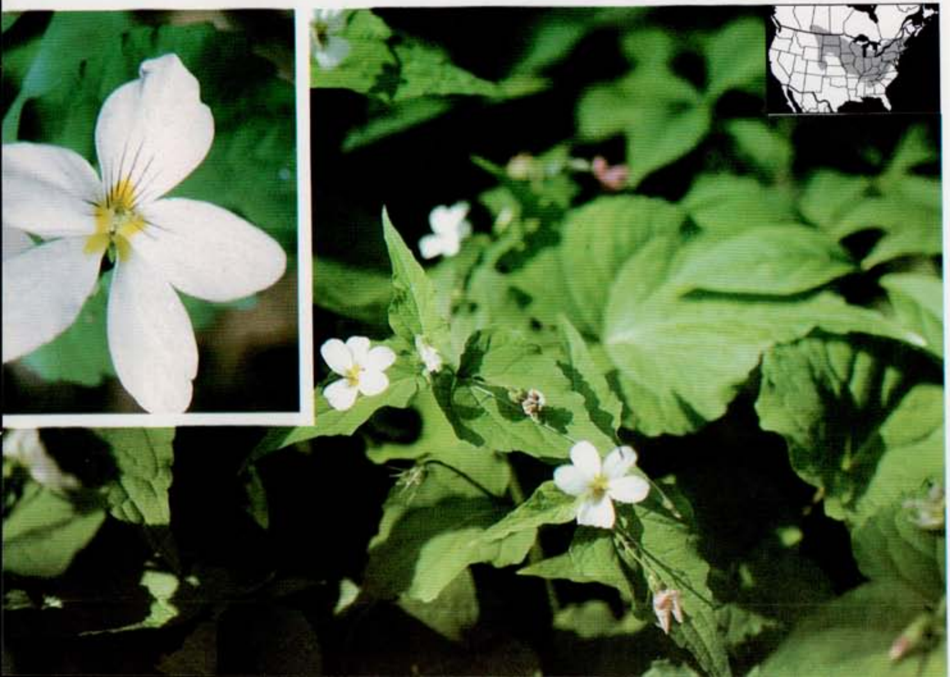

Miner's lettuce *Montia perfoliata*
Also edible summer (synonym: *Claytonia perfoliata*)

Habitat: valleys, lower mountain slopes, springs, moist sites. Introduced from Europe; miner's lettuce has expanded its range in the East. **Identification:** ***annual herb*** to 30 cm (12 in) high, ***from fibrous root system.*** Leaves mostly basal, simple, 6–20 cm (2.4–8 in) long including the long leafstalk, blade variable, usually broadest below or above middle, entire along margin, smooth; ***stem leaves 2, opposite and fused into a circular disc just below flowering stalk.*** Flower stalk elongate, bearing numerous, whorled, short-stalked flowers, each tiny, with 5 white, clawed petals 3–6 mm (0.1–0.2 in) long. Fruits are tiny, 3-parted capsules containing 1–3 shiny, black seeds. **Harvest:** blossoms, leaves, stems, and roots any time during growing season. **Preparation:** Young upper leaves and stems are excellent as picked or in salads. Include blossoms when available. For salad add chopped hard-boiled eggs and oil and vinegar dressing. For cooked vegetable, boil leaves and stems in very little water until tender; boil young leaves as little as 3–4 min. Excellent source of vitamin C. Wash roots and boil for just a few minutes. Butter and lightly season to bring out water chestnut flavor. **Related edible species:** Siberian miner's lettuce *(M. sibirica)* of southern Alaska. **Poisonous look-alikes:** none.

Canada violet *Viola canadensis*
Also edible summer

Habitat: mixed deciduous woodlands, bluffs, rocky to deep, rich soils. **Identification:** perennial herb with thick underground stem (rhizome); stems upright, smooth, to 40 cm (16 in) high. ***Leaves*** basal and along stem, simple ***4–12 cm (1.6–4.7 in) long, nearly as wide, broadest near base, long pointed at tip,*** toothed along margin, ***heart-shaped at base,*** leafstalks long, ***with small, lance-shaped bract (stipule)*** at base of each leaf. ***Flowers*** produced on upper third of plant, each stalked, ***white often with purple on back, 2–3 cm*** (0.8–1.2 in) ***across, petals 5, lateral ones fringed.*** Fruits are egg-shaped to nearly globe-shaped capsules 5–10 mm (0.2–0.4 in) long, containing several tiny brown seeds. **Harvest:** Strip blossoms from plants, avoiding stems. When possible gather bright green leaves which are young and tender, but even older leaves in late summer are edible. **Preparation:** Leaves and blossoms are superior sources of vitamin C; the leaves are high also in vitamin A. Use blossoms as snack straight from plant or for jelly, jam, syrup, and candy. For jelly, see book's introduction. Use leaves and flowers raw with other ingredients in tossed salad. Cook chopped leaves with minimum water in covered pot for about 15 min. Season and serve alone or with mustard greens. For tea, dry leaves in warm shade and store in paper bag in cool dry places. Steep 1 oz dried leaves per cup of boiling water. **Related edible species:** PALE VIOLET (next), and WOOD VIOLET, and SOUTHERN WOOD VIOLET. It is likely that all of the species of violets are edible, with some more palatable than others. Some, especially yellow-flowered violets, may be cathartic. Thus, moderate initial consumption is advised. **Poisonous look-alikes:** Violet leaves may resemble leaves of other wildflowers, some of which are toxic. So it's safest to collect violet leaves only when flowers are in bloom.

Pale violet *Viola striata*

Also edible summer

Habitat: wooded floodplains, meadows, stream banks. **Identification:** ***perennial herb*** with short thick underground stem (rhizome); stems several, upright or reclining, smooth, to 50 cm (20 in) high. Basal leaves soon withering; ***stem leaves*** simple, ***3–8 cm*** (1.2–3.2 in) ***long,*** as wide or nearly so, ***broadest near base and almost circular in outline,*** pointed to blunt at tip, toothed along margin, shallowly heart-shaped at base, leafstalks long; ***with large, leaflike bract (stipule) at base of each leaf. Flowers*** usually produced at tips of stems, each on a slender stalk, ***creamy white with purplish-brown veins near base, 2–2.5 cm (0.8–1 in) across, petals 5, lateral ones densely fringed, forming spur 3–6 mm*** (0.1–0.2 in) ***long on backside.*** Fruits are egg-shaped to globe-shaped capsules 4–8 mm (0.2–0.3 in) long. **Harvest and preparation:** same as for CANADA VIOLET (preceding). **Related edible species:** CANADA VIOLET. **Poisonous look-alikes:** See CANADA VIOLET for details.

Pokeweed *Phytolacca americana*

Habitat: woodland margins, disturbed sites, waste ground, roadsides, fencerows. **Identification:** ***large, perennial herb*** to 3 m (9.8 ft) tall with thick fleshy roots; stems branched, upright, smooth, ***with disagreeable odor when broken.*** Leaves alternate, 9–30 cm (3.5–12 in) long, 3–12 cm (1.2–4.7 in) wide, broadest near base, pointed at tip, entire along margin; leafstalks to 5 cm (2 in) long. ***Flowers greenish to white, produced in narrow elongate clusters*** (racemes); each flower 2–4 mm (0.1 in) long, petals absent. ***Fruits are 5–12-parted berries, wider than long, 6–10 mm*** (0.2–0.4 in) ***in diameter, green turning purplish-black at maturity.*** **Harvest:** young shoots up to 15 cm (6 in) long in late spring before leaves unfold. Identify from presence of previous year's stalks. Cut stems from perennial root as you would asparagus. *Caution: The root is poisonous; avoid all parts of the plant except young shoots.* If shoots show any purple coloration, avoid them as well as mature stems. For winter use, dig root in fall and plant in large container of soil. After hard freeze move to warm area and water regularly. Harvest shoots as above. **Preparation:** Boil shoots 10 min with 1 or 2 water changes; drain, add fresh boiling water and season, cover, and simmer until tender. Serve like broccoli or asparagus with butter, cheese, or other sauces. For casserole, alternate layers of cooked pokeweed shoots, white sauce, and sliced, boiled eggs. Top with seasoned bread crumbs. Bake 30 min at low temperature 165°C (330°F). Boiled, peeled shoots may also be pickled. **Related edible species:** none. **Poisonous look-alikes:** None, but see *caution* above.

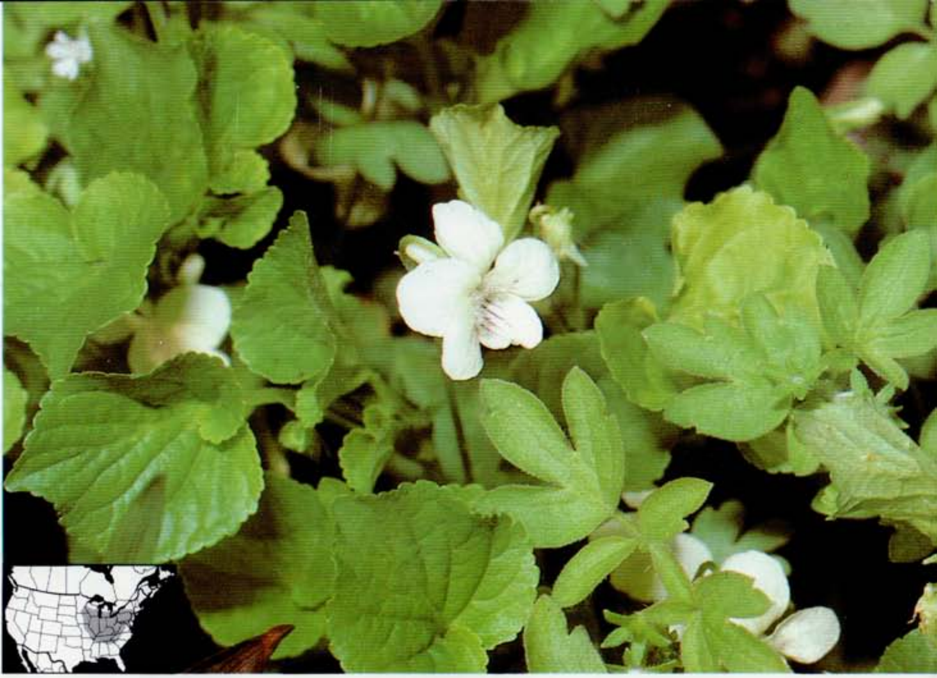

Pale violet: flowering plant (Mohlenbrock).

Pokeweed: fruiting stalks (Crow). Flowers (Elias). Young shoots attached to cut root top (Dykeman).

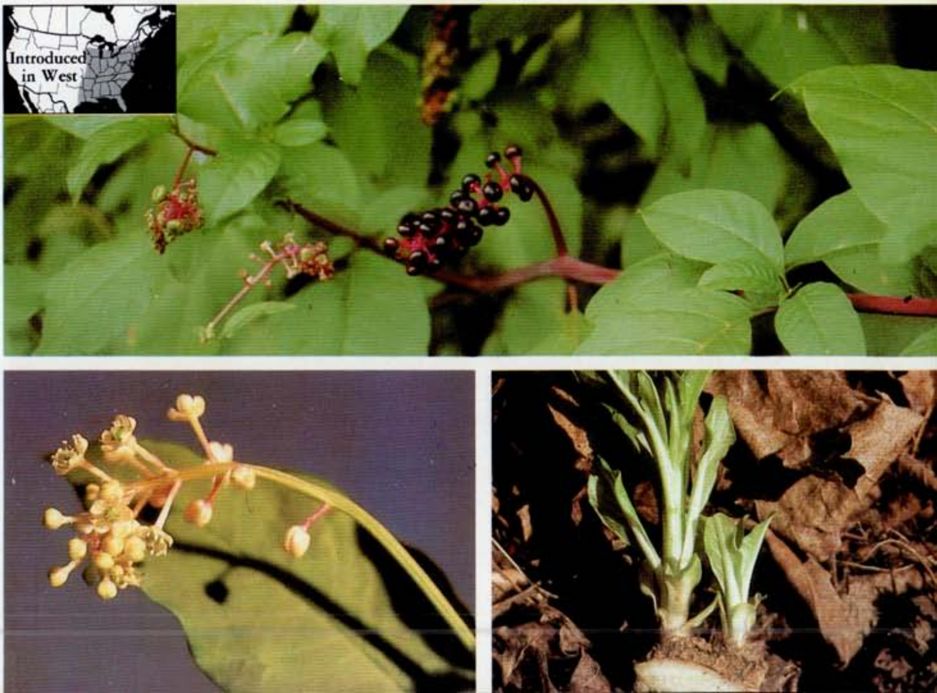
Introduced in West

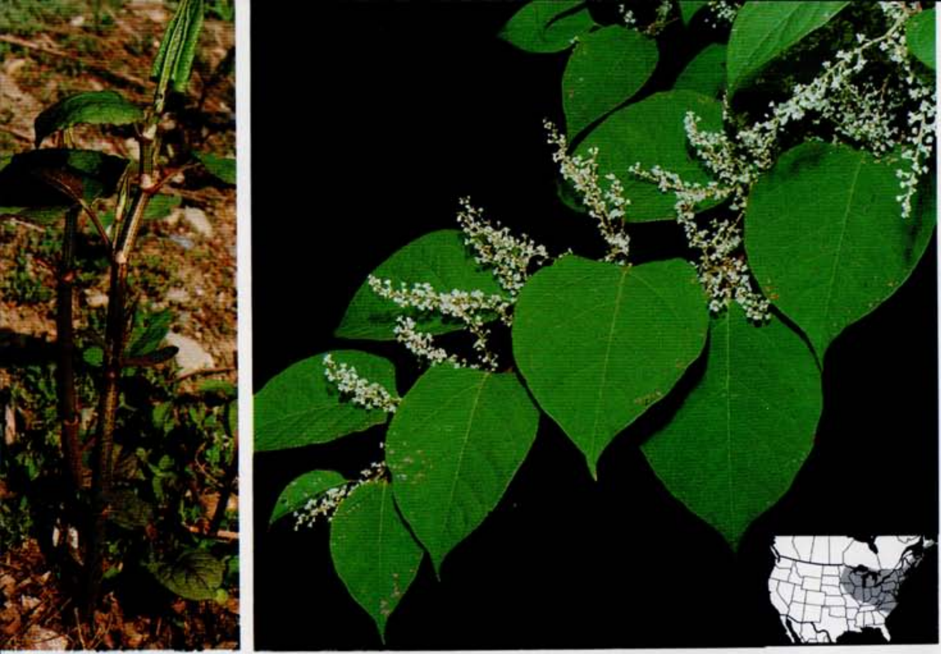

Japanese knotweed: young shoots (Dykeman). Flowering branch (Feldman).

Wintercress: entire plant, leaf, and flowering head (Dykeman).

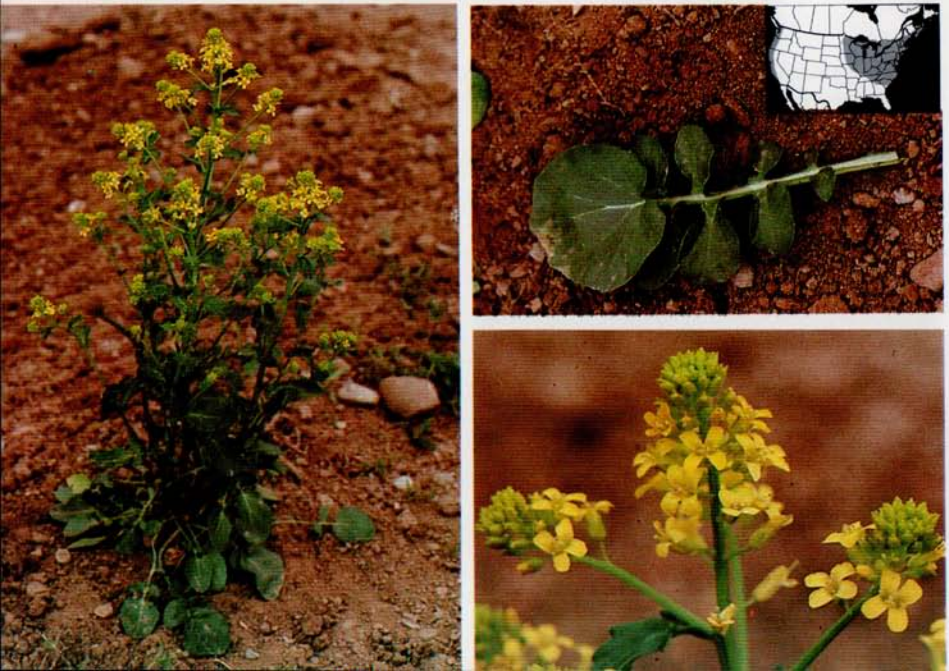

Japanese knotweed *Polygonum cuspidatum*

Habitat: disturbed and waste sites, vacant lots, roadsides; introduced weed from Asia. **Identification:** ***large perennial herb to 3 m*** (9.8 ft) ***high,*** often forming dense colonies; ***stems*** upright, ***with hollow core,*** smooth, mottled, frequently branching. ***Leaves*** alternate, simple, ***10–16 cm*** (4–6.3 in) ***long, 6–12 cm*** (2.4–4.7 in) ***wide, broadest near base,*** long tipped, ***wavy along margin,*** squared to rounded at base, almost leathery; ***leafstalk*** stout and ***enclosed at base by membranous sheath that also surrounds stem.*** Flowers produced in elongate, branched clusters in junctions of upper leaves, each flower tiny, 1–2 mm (less than 0.1 in) long, white. Fruits are dark brown, shiny, winged nutlets. **Harvest:** Locate papery dry stalks of previous year. Gather young shoots up to 30 cm (1 ft) long for potherb, up to 90 cm (3 ft) for other use. **Preparation:** Boil tips of young shoots in salt water or steam 4–5 min. Serve like ASPARAGUS, hot on toast with butter or white sauce, or cold in salad. Also, force cooked shoots through mill for puree. Add salt, butter, sweetener, to taste, and reheat. Or prepare casserole of cooked young shoots, bread crumbs, grated cheese, cream, and seasonings; bake at 200°C (390°F) until top is browned. Peel older stalks 30–90 cm (1–3 ft) and use like rhubarb. For sauce combine 4 cups peeled and cut shoots, 1 cup sugar, and nutmeg. Simmer until soft. Serve hot or cold. See basic jam recipe in book's Introduction. Prepare pie from peeled shoots using rhubarb pie recipes. **Related edible species:** giant knotweed *(P. sachalinense),* Alaskan knotweed *(P. alaskanum),* AMERICAN BISTORT, ALPINE BISTORT, LADY'S THUMB and others. Some may cause a light-sensitive skin reaction. **Poisonous look-alikes:** none.

Wintercress *Barbarea vulgaris*

Also edible autumn, winter

Habitat: fields, wet meadows, ditches, roadsides; common and widespread, introduced from Europe. **Identification:** ***biennial or perennial herb*** to 0.7 m (28 in) high; stems erect, sparingly branched above. Lower ***stem leaves*** broadest near or above middle, 7–15 cm (2.8–6 in) long, ***deeply lobed along sides,*** stalked; leaves of upper stem much smaller, stalkless and lobed at base, coarsely toothed. ***Flowers from April to June, numerous in cylinder-shaped clusters, each flower*** narrow, stalked, ***with 4 narrow, yellow petals. Fruit pods very narrow, 1.5–3 cm (0.6–1.2 in) long, tipped with short beak to 3 mm (0.1 in) long,*** splitting open lengthwise to release tiny seeds. **Harvest:** Pick leaves after frosty weather and into spring until flower stalks appear. Leaf rosettes grow on warm winter days; become bitter in spring. Leaves blanch and lose bitterness if you tie them in bundle with plant rooted and cover with a bucket. Pick buds and flowers until blooming is over, April or May. **Preparation:** Use early and blanched leaves in salads with other late-winter foods such as COMMON DAY LILY tubers and WATERCRESS. Boil later, bitter leaves in 2 changes of water until tender and treat as for BLACK MUSTARD (next). Buds and flowers are excellent with cheese sauce. Cover clusters with boiling water for 1 min. Drain, cover with fresh boiling water, heat, and boil 3 min more. Drain, season, and cool for 5 min. **Related edible species:** early wintercress *(B. verna)*. **Poisonous look-alikes:** none.

Black mustard *(Also edible summer)* *Brassica nigra*

Habitat: common and widespread, fields, yards, disturbed sites; introduced from Europe. **Identification:** ***medium-size to large annual herb*** to 1.2 m (4 ft) high; stems erect, hairy on lower parts. Leaves alternate, stalked, usually broadest above or near the middle, 8–18 cm (3.2–7 in) long, ***lower leaves lobed, upper leaves narrower, wavy toothed. Flowers few in narrow, elongate, loose clusters,*** each ***flower*** stalked, about 1 cm (0.4 in) across, ***with 4 yellow petals. Fruits*** narrow linear ***capsules 1–2 cm*** (0.4–0.7 in) ***long, roughly 4-sided,*** upright, ***beaked, beak 1–2 cm*** (0.4–0.8 in) ***long,*** splitting open lengthwise to release tiny seeds. **Harvest:** leaves very early in spring or blanch later leaves as with WINTERCRESS (preceding). Pick buds and flowers until nearly all flowers have opened. Avoid bitter leaves. For seeds, harvest during summer when oldest pods open. Pile them on fine cloth or plastic tarp and sun-dry for several days. Beat with flail to release seeds, remove stalks, winnow seed from chaff. **Preparation:** Use early and blanched leaves in salads. Boil older greens with 1 change of water for about 30 min. Season with butter and vinegar or mix with crumbled bacon and chopped wild onions. Top with hard-boiled egg slices and French dressing or vinegar. Excellent source of vitamins A, B, and C. For buds and flowers, treat as WINTERCRESS. Sprinkle seeds lightly on salad. To make mustard, grind seeds in food mill; use dried. Or for mustard spread, combine equal parts mustard powder and oven-browned flour with half-and-half mix of vinegar and water. Add other spices to taste. **Related edible species:** field mustard *(B. rapa),* Chinese mustard *(B. juncea),* and others. **Poisonous look-alikes:** none.

Marsh marigold, Cowslip *Caltha palustris*

Habitat: freshwater swamps, wet meadows, along streams, ponds, lakes. **Identification:** ***perennial herbs*** to 70 cm (28 in) high; stems partially spreading to upright, hollow, smooth. ***Basal leaves long-stalked, blade rounded to almost kidney-shaped, 1–2.8 cm*** (0.4–1.1 in) ***long, 0.4–2.2 cm*** (0.2–0.8 in) ***wide,*** rounded at tip, toothed along margin, heart-shaped at base; stem leaves smaller, lacking stalks. ***Flowers large, showy,*** 2.5–4 cm (1–1.6 in) across, ***with 5 bright yellow petallike sepals.*** Fruits are elongate, somewhat rounded, 1–1.7 cm (0.4–0.7 in) long, splitting open along one side when mature to release small seeds. **Harvest:** Cut leaves from plant when young, best before flowers open. Pick tightly closed buds for pickling. **Preparation:** *Raw leaves and buds are poisonous until properly treated!* Wash leaves and cover with boiling water. Reheat to boiling, then drain. Repeat 2 or 3 times, then cook in very little water until tender. Drain and serve with butter and seasonings or with cream or cheese sauce. For pickles, treat buds with boiling water as described for leaves. Do not boil. Drain and pack in jars. Cover with hot pickling liquid and seal. Store for at least a month before eating. Do not consume used pickling juice. To make pickling liquid, slowly add mix of 1/4 cup salt, 1/2 cup sugar, 1 tbsp ground mustard, and 1 tbsp celery seed to 2 cups vinegar and 1 cup water. Boil 10 min with stirring. **Related edible species:** none. **Poisonous look-alikes:** Flowers resemble COMMON BUTTERCUPS yet leaves are very different.

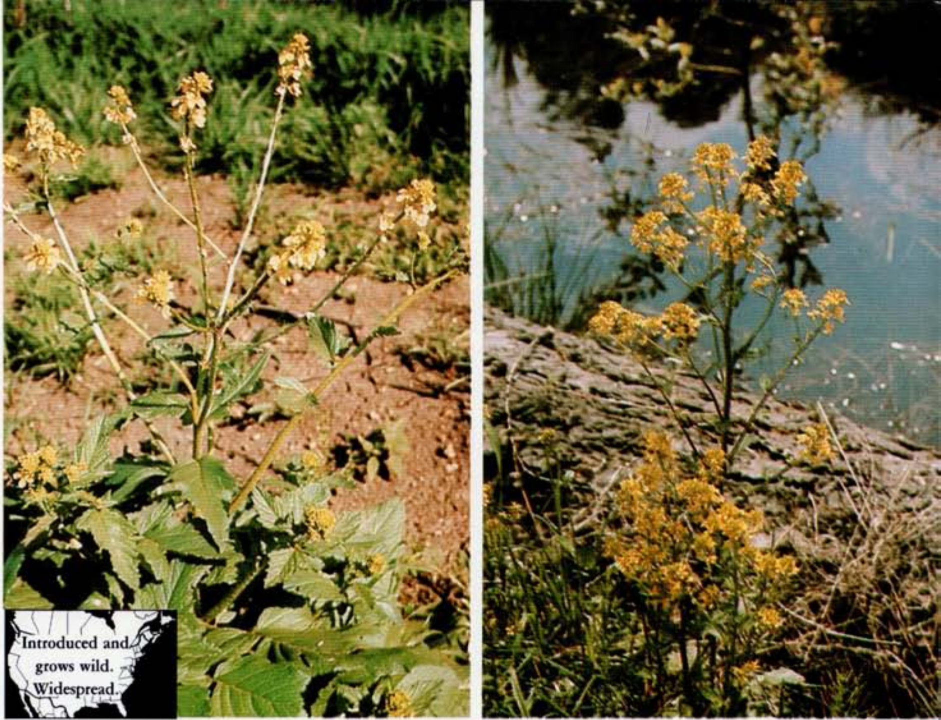
Introduced and grows wild. Widespread.

Black mustard: entire plants (NYBG).

Marsh marigold: plants in typical streamside habitat and entire plant (Dykeman).

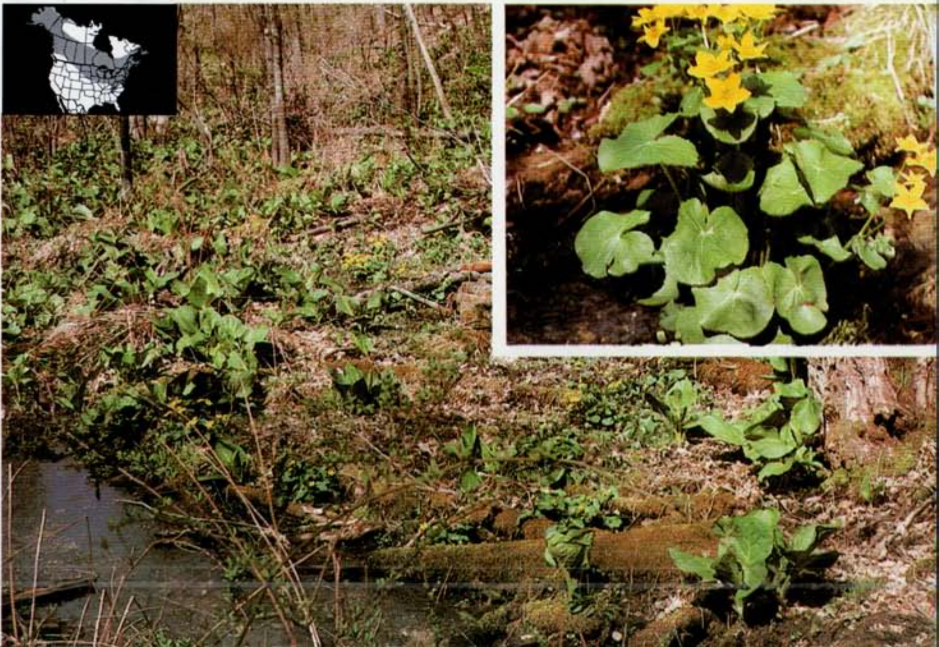

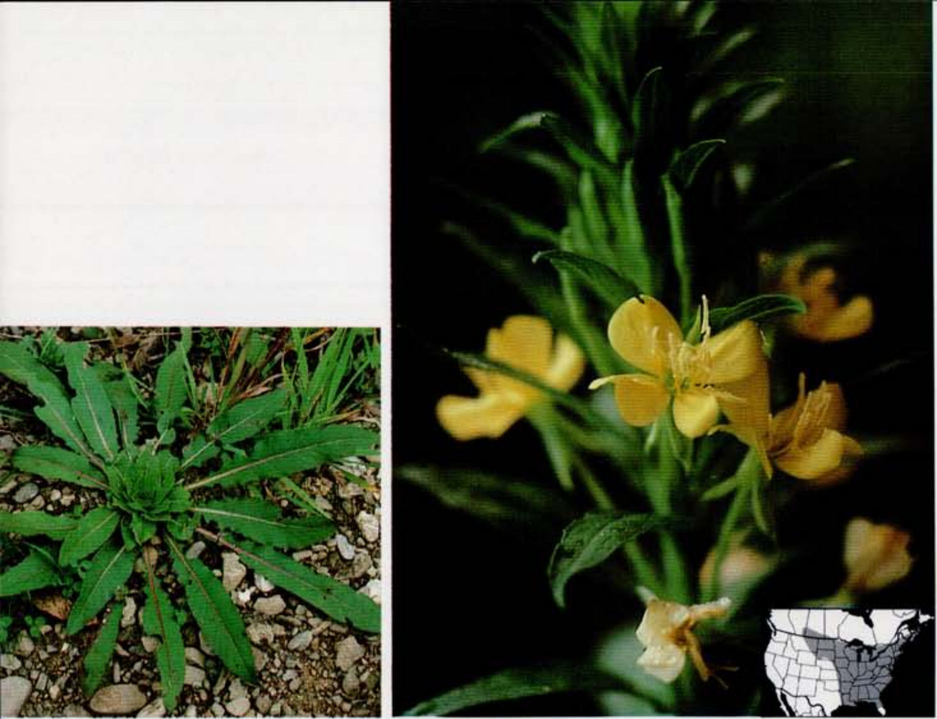

Evening primrose: basal leaves (Dykeman). Flowering stalks (Pawloski).

Sow thistle: fruiting head, left, and flower (Mathisen).

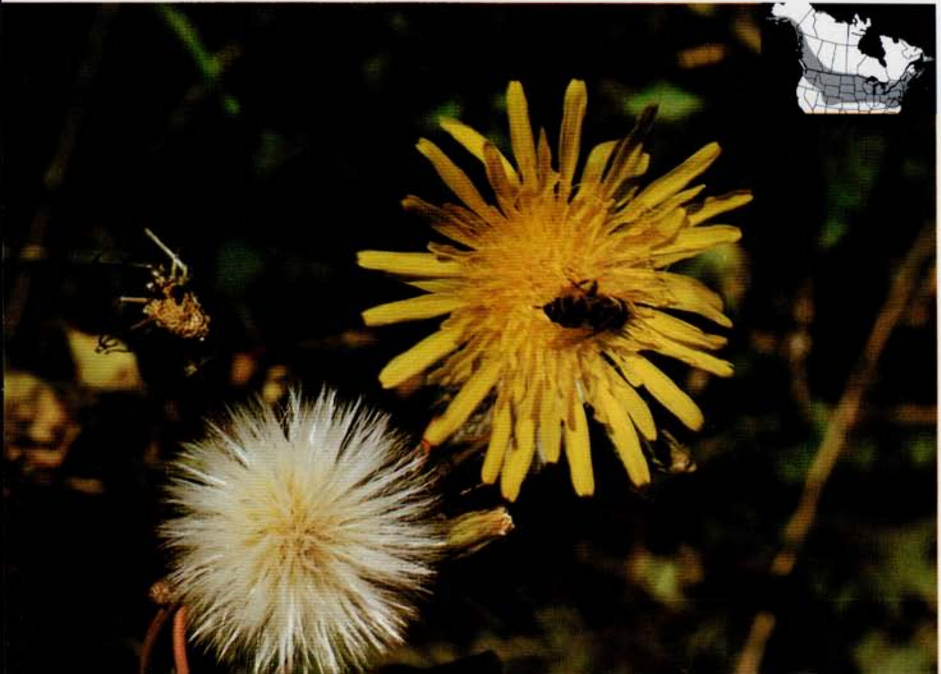

Evening primrose *Oenothera biennis*

Also edible autumn, winter

Habitat: disturbed sites, along roads, prairies, fields. **Identification:** biennial herb of 1.6 m (5.3 ft) high, from elongated taproot; stems upright, hairy. Leaves of first year in low rosette; second year leaves on flower stalk, alternate, lance-shaped to broadest near middle, 8–18 cm (3.2–7 in) long, 2–6 cm (0.8–2.4 in) wide, pointed at tip, entire to toothed along margin, tapering at base, hairy. ***Flowers second year, many in terminal, elongated cluster, each flower with narrow, tubular calyx 1.5–3 cm*** (0.6–1.2 in) ***long, 4 large spreading, bright yellow petals 1.5–3 cm*** (0.6–1.2 in) ***long. (Flowers open after dusk and often close by midmorning.) Fruits are cylinder-shaped, angled capsules 1–4 cm*** (0.4–1.6 in) ***long,*** containing many small, reddish seeds. **Harvest:** Dig taproots of first year growth between autumn and early spring. Pick leaves from rosettes in early spring or even earlier from areas clear of snow. **Preparation:** Peel taproots and boil in 2 changes of water for 20–30 min. Serve with butter and seasonings or slice and fry. Cook sliced roots with meat or in stews, partially boiling roots first if flavor is too strong. Simmer boiled roots in sugar syrup for 20–30 min until candied. Peel leaves and use sparingly in salads. Or boil leaves 15–20 min in 2 changes of water and serve with butter or vinegar. **Related edible species:** other species of this large genus. **Poisonous look-alikes:** none, but don't confuse leaves with those of other early-blooming species.

Sow thistle *Sonchus arvensis*

Habitat: disturbed sites such as roadsides, fields, railroad right-of-ways, gravel banks; introduced weed from Europe. **Identification:** Not to be mistaken for the more common and generally more palatable thistles of the *Cirsium* genus, sow thistle is a ***perennial herb*** to 1 m (3.2 ft) high with deep vertical and spreading horizontal coarse roots; stems upright, smooth, with milky sap. ***Leaves*** alternate, simple, broadest near or above middle, 10–35 cm (4–14 in) long, 4–14 cm (1.6–5.5 in) wide, ***deeply lobed along margin, upper leaves less so, spiny-toothed, lobed leaf bases clasping stems,*** usually smooth. ***Flowering heads*** produced in loosely branched, elongated, erect cluster, each head ***3–5 cm*** (1.2–2 in) ***across,*** with lance-shaped bracts on the outside and many small ***yellow flowers*** within. Fruits are small, dark-brown nutlets 2.5–3 mm (0.1 in) long. **Harvest:** Locate by noting concentrations of mature plants the previous year. Pick when only a few inches tall; leaves become increasingly bitter with age. *Caution: Sow thistle occasionally contains toxic levels of nitrates.* **Preparation:** If not too bitter, cut and mix leaves with other, blander salad greens. Boil in very little water for 3–4 min, change water, and then boil until tender. Serve with butter and seasonings or vinegar. If too bitter, mix with blander greens such as LAMB'S QUARTERS. **Related edible species:** Several species of the *Sonchus* genus are variously distributed and edible when very young. Some have leaf spines that must be removed. **Poisonous look-alikes:** none.

Prickly lettuce *Lactuca serriola* (synonym: *L. scariola*)
Habitat: disturbed sites such as fields, roadsides, vacant lots; introduced weed from Europe. **Identification:** ***annual or biennial herb*** to 2 m (6.6 ft) high, ***with milky juice in stems and leaves,*** stems upright, slender, usually armed with spreading prickles. Stem ***leaves*** alternate, ***lance-shaped,*** 5–25 cm (2–10 in) long, 1.5–12 cm (0.6–4.7 in) wide, ***with several deep lobes or lobeless, toothed and spiny along margin*** and spiny along main vein on lower leaf surface, with pointed lobes at base and often clasping stem. ***Flowering heads*** sparingly produced in branched clusters, each head ***1–1.5 cm*** (0.4–0.6 in) ***long,*** 3–6 mm (0.1–0.2 in) ***across, containing 6–12 yellow flowers.*** Fruits are tiny gray to brown nutlets 3–4 mm (0.1 in) long. **Harvest:** Like other wild lettuce species, prickly lettuce is good only when very young. Pick leaves when plants are less than 20 cm (8 in) tall. **Preparation:** Use alone or with other greens in tossed salad. Boil leaves in very little water for 2–3 min. Serve with butter and seasonings. For wilted lettuce pour boiling water over lettuce leaves and let sit for 5 min. Drain. Make dressing of 3 slices of crisp bacon crumbled into ¼ cup hot vinegar with 1 tsp sugar and ½ tsp salt. Pour over hot lettuce. *Caution: Large quantity of raw young leaves may cause digestive upset.* **Related edible species:** many species of the *Lactuca* genus widely distributed; most noticeable of these is wild lettuce *(L. canadensis),* which lack prickles on leaves and lower stem. **Poisonous look-alikes:** none.

Dandelion *Taraxacum officinale*
Also edible autumn, winter

Habitat: lawns, yards, disturbed sites; common weed introduced from Europe. **Identification:** ***perennial herb, with large,*** sometimes branched ***taproot. Leaves all basal,*** 8–40 cm (3.2–16 in) long, 0.8–14 cm (0.3–5.5 in) wide, widest near tip, rounded to pointed at tip, ***variously and irregularly lobed along margin,*** tapering to narrow base, hairy beneath, ***with milky juice.*** Flowers produced throughout spring in large heads at tip of hollow stalks. ***Flowering heads 2–5 cm*** (0.8–2 in) ***across, composed of numerous, small, narrow, orange-yellow flowers, remaining closed on cloudy days. Fruits*** on elongated, upright hollow stalks, ***forming delicate ball*** composed of many tiny seeds each tipped with white featherlike tuft (pappus) easily dispersed by wind. **Harvest:** Best lawn areas are those with taller grass. Dig roots with crown at top of root and leaves attached from time leaves (rich in vitamin A) appear until they are too bitter to use. Dig older roots in fall and winter. Cut out developing bud clusters and pick fully opened flowers in spring. **Preparation:** For flowers, leaves, and older roots, see CHICORY. Peel young roots, slice thin, and boil in 2 changes of water for about 20 min, adding pinch of baking soda to first water. Serve with butter and seasonings or saute cooked slices in butter until brown. Season with paprika. Boil buds just a few minutes, butter and serve. Dip blossoms in fritter batter; fry in oil and serve with syrup or garlic salt. **Related edible species:** several other species of *Taraxacum,* all similar to common dandelion in appearance/preparation. Also, CHICORY in early growth. **Poisonous look-alikes:** none.

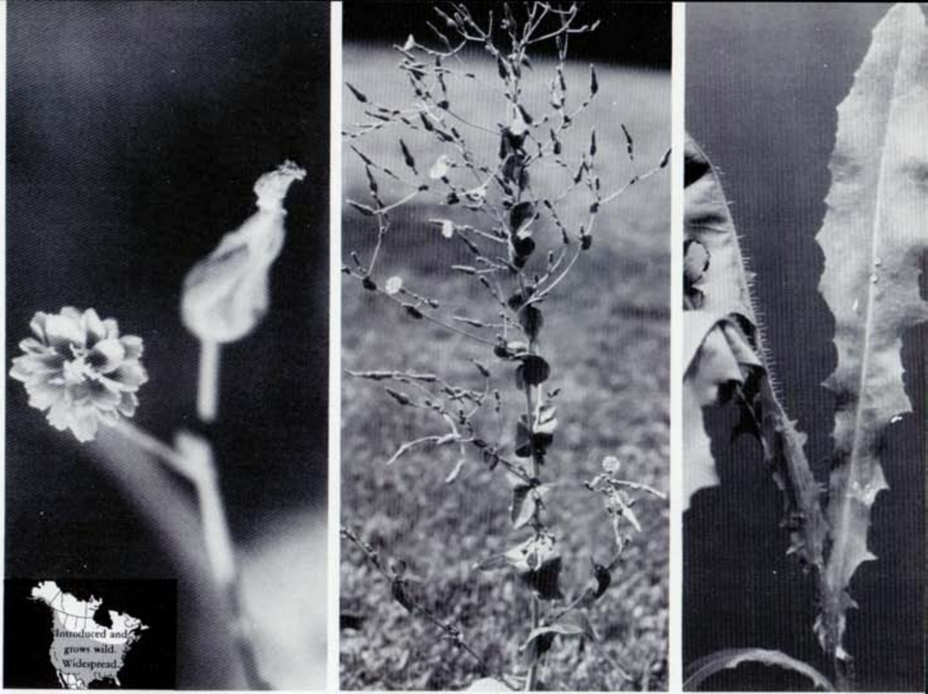
Introduced and
grows wild.
Widespread.

Prickly lettuce: flowering branch and entire plant (Kavasch). Leaves (Dykeman).

Dandelion: flowering plant (Elias).

Introduced and
grows wild.
Widespread.

Goat's beard *Tragopogon pratensis*

Habitat: roadsides, fields, other disturbed sites; introduced from Europe. **Identification:** ***perennial*** or biennial ***herb*** to 76 cm (30 in) high with taproot and milky juice; stems upright. ***Leaves*** alternate, linear, ***grasslike, 15–30 cm*** (6–12 in) ***long, 1–2 cm*** (0.4–0.8 in) ***wide,*** pointed at tip, entire along margin, base partially circling stem, smooth. ***Flowering heads*** produced singly at tip of stems, ***resembling dandelion, with 8 lance-shaped outer bracts, many yellow flowers.*** Fruiting heads globe-shaped, light, fluffy, consisting of numerous small nutlets, each tipped with light, featherlike bristles. **Harvest:** See SALSIFY (preceding). Roots of goat's beard may be smaller, tougher, less palatable. **Preparation:** Same as SALSIFY. Cool boiled roots and saute in oil until browned or prepare like potatoes. Roasted and ground roots make coffee substitute. Use very young leaves in salad if not bitter or boil for 5–10 min and serve with butter and seasonings. **Related edible species:** SALSIFY, yellow goat's beard *(T. dubius)*. **Poisonous look-alikes:** none in the flowering stage.

Common milkweed *Asclepias syriaca*

Also edible summer

Habitat: fields, meadows, roadsides, woodland margins. **Identification:** ***perennial herb*** to 1.8 m (6 ft) tall; stems upright, usually unbranched, hairy. ***Leaves opposite,*** broadest near base or middle, ***10–28 cm*** (4–11 in) ***long, 4–12 cm*** (1.6–4.7 in) ***wide,*** pointed at tip, entire to wavy along margin, ***thick, leathery, hairy above, with milky sap. Flowers in dense, rounded heads 6–10 cm*** (2.4–4 in) ***across, each flower rose, dull purple to greenish-white,*** 8–12 cm (3.2–4.7 in) long, with 5 reflexed petals. ***Fruits are elongated pods,*** broadest near base and gradually tapering to curved point, 8–14 cm (3.2–5.5 in) long, covered with small, soft, green projections. Seeds small, each with conspicuous, soft, featherlike hairs at one end. **Harvest:** Cut new shoots when less than 20 cm (8 in) tall. Pick newly opened leaves until buds are formed; then pick clusters of buds and flowers. Later pick firm pods 2–3 cm (0.8–1.2 in) long. **Preparation:** To remove bitterness from all parts, cover with boiling water and bring back to boil. Discard water and repeat 2–3 times with boiling water. Do not use cold water in process. Cook treated shoots like ASPARAGUS. After treatment boil young leaves, bud clusters, and pods until tender, usually 10–15 min. Serve with butter and seasonings or with sauce. Use treated pods in soup like okra *(Hibiscus esculentus)*. **Related edible species:** Showy milkweed *(A. speciosa)* is good. **Poisonous look-alikes:** all narrow-leaved milkweed species, especially BUTTERFLY WEED; some are more toxic than others. Young DOGBANE, also with milky sap, has branches (common milkweed does not) and its young shoots lack hairs.

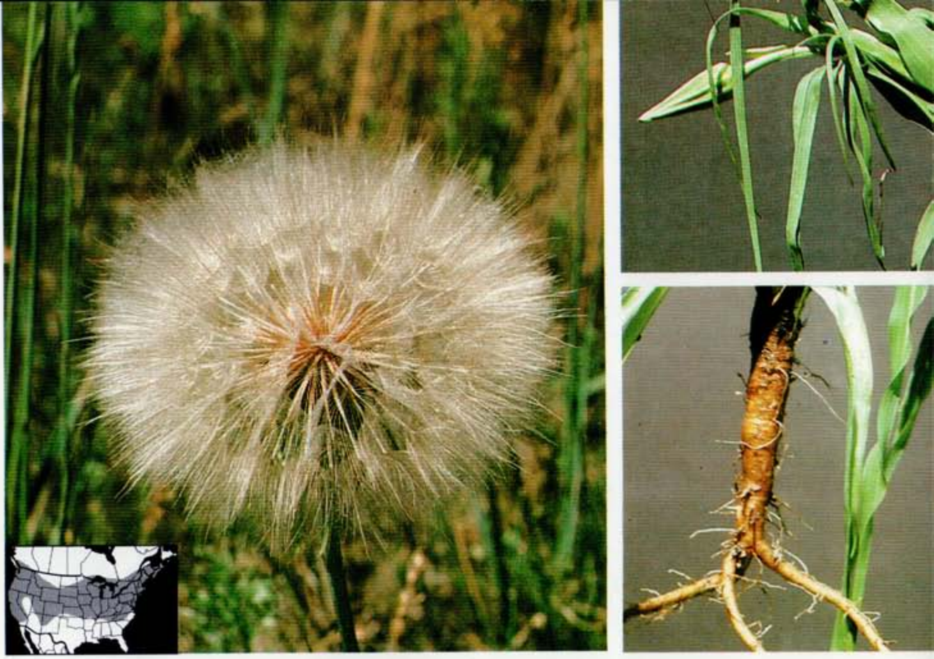

Goat's beard: fruiting head, flower bud, rootstock (Dykeman).

Common milkweed: flowering head and entire plant (Elias).

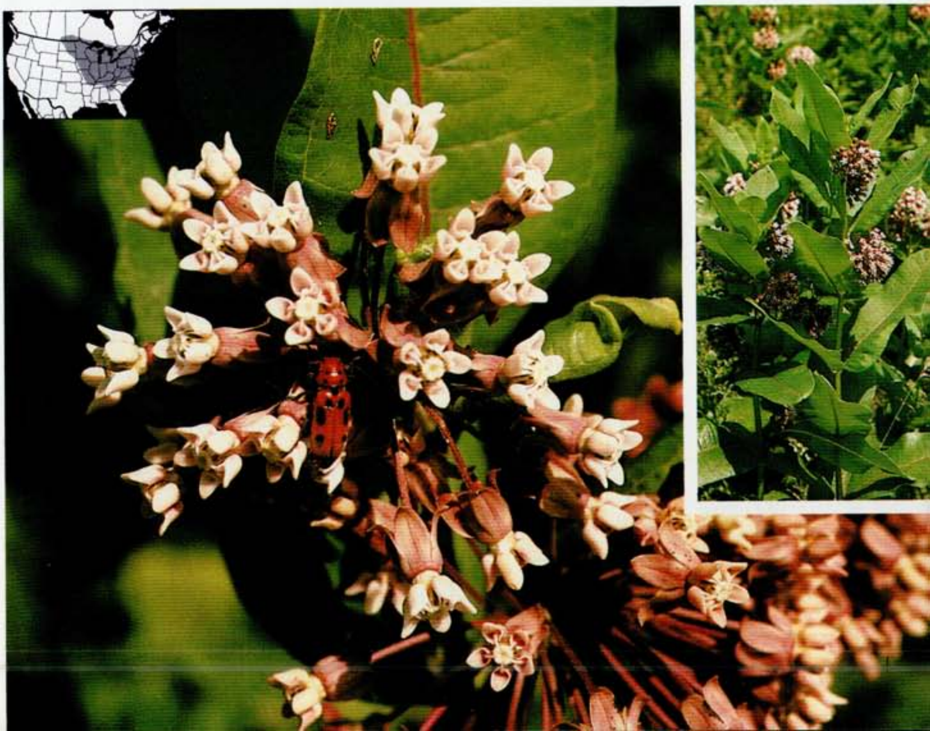

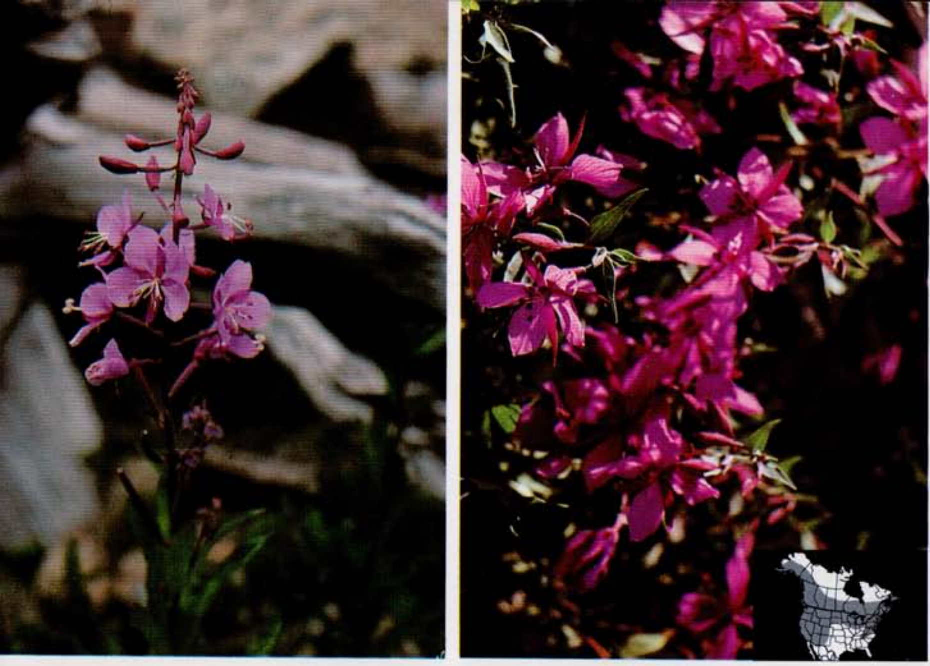

Fireweed: flowering stalk (Elias). Flowers (NYBG).

Alfilaria: entire plant with flowers and fruits (Ross).

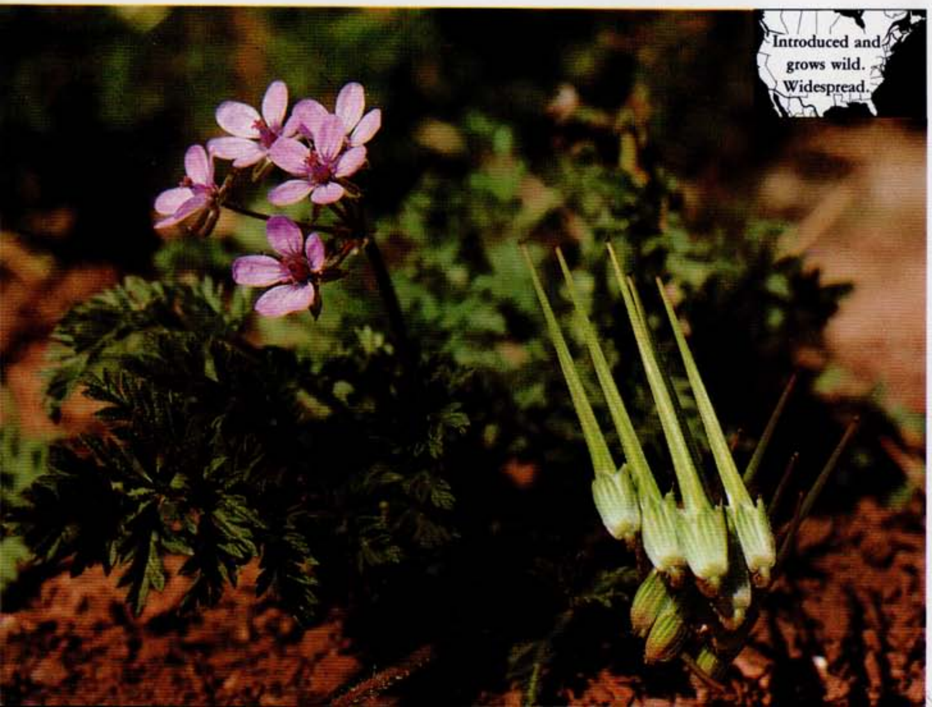
Introduced and grows wild. Widespread.

Fireweed *Epilobium angustifolium*

Also edible summer

Habitat: woodlands, ravines, clearings; commonly grows after fires. **Identification:** ***perennial herb*** 1–2 m (3.2–6.5 ft) high; stems erect, usually single, smooth. Leaves alternate, narrow and broadest near base, 5–20 cm (2–8 in) long, to 3 cm (1.2 in) wide, nearly stalkless, long-pointed at tip, entire or nearly so along margin. ***Flowers*** many on elongate spikes; each flower stalked, spreading, ***with 4 pink to purple, rarely white, clawed petals*** 1–2 cm (0.4–0.8 in) long. ***Fruit a narrow capsule 3–8 cm*** (1.2–3.2 in) ***long,*** with reddish or purplish cast, 4-parted. **Harvest:** Cut young shoots in spring; new leaves, flower bud clusters, and mature stems later. Older leaves are good for tea but too bitter for potherb. **Preparation:** Not a very tasty food plant but widespread, abundant, easy to identify, and has several edible parts. Cook young shoots like ASPARAGUS and season with butter, hollandaise, or cheese sauce. Use young leaves and later flower-bud clusters in salad with other greens. For potherb, stir 4 cups small leaves and 3 tbsp water into 2 tbsp melted butter in skillet over high heat. Cover and cook 2–3 min, season and serve. Boil flower bud clusters 10–15 min and serve with butter, salt and pepper. Peel stems too old for cooking and chew on raw pith or use pith in soups. Dry mature leaves and blend with other suitable leaves for a tea. **Related edible species:** dwarf fireweed *(E. latifolium)* and other related but smaller species. **Poisonous look-alikes:** none.

Alfilaria, Filaree, Storksbill *Erodium cicutarium*

Habitat: fields, pastures, roadsides, other disturbed sites; common weed introduced from Europe. **Identification:** ***annual or biennial herb*** with 1 to several spreading stems, hairy. ***Leaves*** emerge at base and along stem, alternate, ***2–10 cm*** (0.8–4 in) ***long, featherlike (pinnate) compound, leaflets (pinnae) cleft and coarsely toothed. Flowers*** 2–10 in small, branched, rounded-headed cluster at tips of shoots; each flower about 2.5 cm (1 in) across, ***with 5 rose-purple petals.*** Fruits are narrow, elongate capsules 2–4 cm (0.8–1.6 in) long, opening from base toward tip. **Harvest:** Gather young, overwintered leaves in early spring when other wild foods are scarce. Be certain leafstalks are hairy. **Preparation:** Chop leaves for salad with other greens. Boil leaves in lightly salted water until tender, about 10–20 min. Serve with butter, lemon juice, and herbs. **Related edible species:** none recommended. **Poisonous look-alikes:** Leaves resemble those of the parsley family including POISON HEMLOCK. Picking only hairy-stemmed leaves very early in spring should help you avoid errors. Be sure of identification!

Live-forever, Orphine *Sedum purpureum*

Also edible autumn, winter

Habitat: roadsides, open slopes, woodlands, and disturbed sites; introduced European weed. **Identification:** ***perennial herb*** to 80 cm (32 in) tall, from fleshy tubers. ***Stems upright, succulent; leaves*** alternate or whorls of 3, simple, ***fleshy, 3–6 cm*** (1.2–2.4 in) ***long***, 1–3 cm (0.4–1.2 in) wide, pointed or blunt at tip, entire or toothed along margin, green to bluish-green. ***Flowers*** produced in a compact, branched cluster at tips of branches, each flower bisexual, 6–10 mm (0.2–0.4 in) across, ***with 5 reddish-purple to deep pink petals.*** Fruits small capsules 4–6 mm (0.2 in) long, splitting open along 1 side only. **Harvest:** young leaves in spring and tubers from fall through spring, as long as they are firm and crisp. **Preparation:** Add raw, very young leaves to salads. Boil leaves gathered through early summer for 5–10 min and serve as cooked green. Pour hot pickling liquid over tubers and leave for several days before eating. For pickling liquid boil 2 tbsp mixed pickle spices in cloth bag in solution of 1 qt vinegar and 1 cup sugar. Boil tubers like those of COMMON DAY LILY for about 20 min. Serve with butter and seasonings. **Related edible species:** Other *Sedum* species provide edible leaves; some are more pungent than others. Roseroot *(S. rosea)* is the most frequently used. **Poisonous look-alikes:** none.

Lady's thumb *Polygonum persicaria*

Also edible summer

Habitat: disturbed sites, especially low wet ground, along roadsides, lakes, streams, ponds; widespread weed introduced from Europe. **Identification:** ***annual herb*** to 1 m (3.2 ft) high; stems soon branching, spreading to upright, smooth. ***Leaves*** alternate, ***narrowly lance-shaped,*** 4–12 cm (1.6–4.7 in) long, 0.6–3 cm (0.2–1.2 in) wide, ***often with irregular or triangular purplish spot above,*** pointed at tip, wavy along margin, tapering at base, ***enclosed at base by membranous sheath that also surrounds stem.*** Mature stems reddish. ***Flowers*** numerous, dense ***in cylinder-shaped spikes*** 2–4 cm (0.8–1.6 in) long, each flower ***pink to rose.*** Fruits are tiny, broadly egg-shaped nutlets 2–2.5 mm (0.1 in) long. **Harvest:** Gather young leaves, mainly in spring but later if not too peppery. **Preparation:** Add chopped raw leaves to tossed salad. Boil leaves in small amount of water for 5–10 min. Serve with vinegar. **Related edible species:** ALPINE BISTORT, JAPANESE KNOTWEED. Known as the smartweeds, other *Polygonum* species may be as good as lady's thumb, and none are known to be poisonous. *Caution: Some species may cause light-sensitive reactions.* **Poisonous look-alikes:** none.

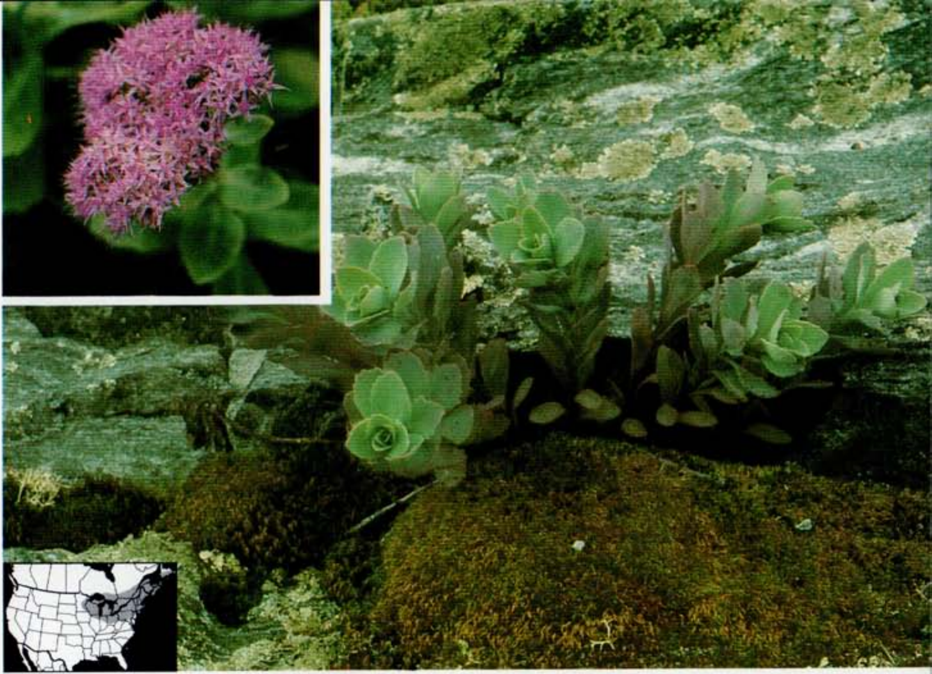

Live-forever: flowering head (Kavasch). Entire plant (Speas).

Lady's thumb: entire plant and stem (Dykeman).

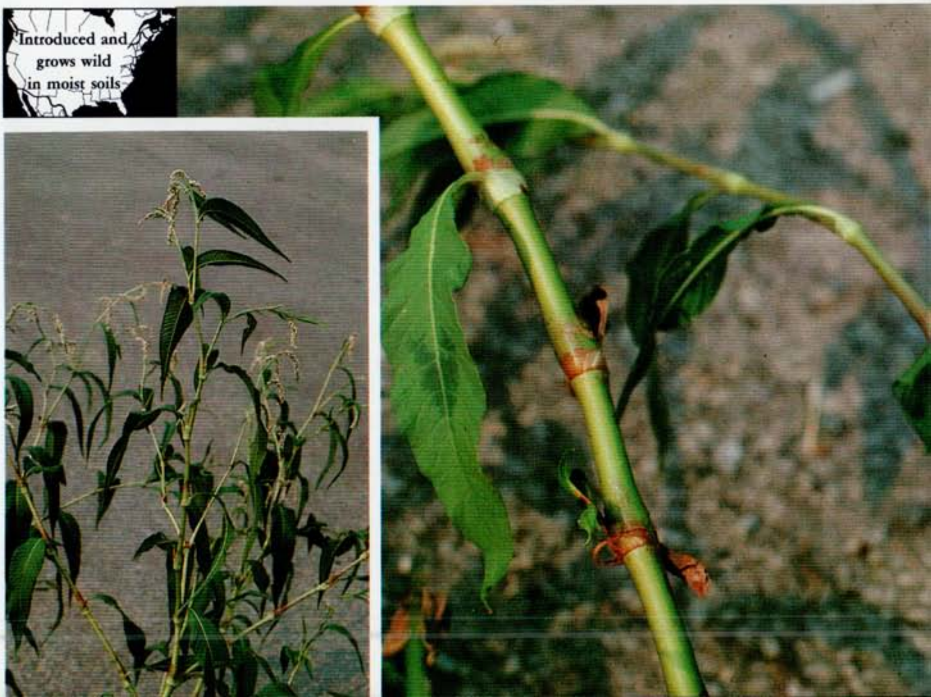
Introduced and
grows wild
in moist soils

Common burdock *Arctium minus*

Also edible summer, autumn

Habitat: roadsides, vacant lots, disturbed sites; introduced from Europe and Asia. **Identification:** ***biennial herb*** to 1.8 m (6 ft) high; stems upright, often branched, usually smooth. ***Leaves*** alternate, ***basal leaves large, 30–50 cm*** (12–20 in) ***long, 20–30 cm*** (8–12 in) ***wide,*** widest near base, ***deeply lobed at base,*** with conspicuous hollow stalk; upper-stem leaves smaller and not lobed at base. ***Flowering heads produced during second year along upper stems*** or in branched elongated clusters; heads 1.7–2.8 cm (0.7–1.1 in) across, outer bracts spine-tipped; ***flowers pink to purple***. Seeds small, hard nutlets 4–6 mm (0.2 in) long. **Harvest:** Dig roots of first-year plants during summer, later if necessary. Pick first-year basal leafstalks and young flower stalks of second year. **Preparation:** Peel roots, slice ½-in thick and boil 20 min with pinch of baking soda. Change water and boil until tender. Serve with butter and seasonings, or mash pieces and make fried patties. Peel flower stalk, removing all green rind. Eat like celery, slice into salad, or boil until tender. Good in casseroles and many other recipes. Simmer pieces of cooked flower stalk in maple syrup to make candy. Eat peeled leafstalks raw or boiled, but since they are hollow, there's not much left after peeling. **Related edible species:** Just as good, great burdock *(A. lappa)* is larger, with larger burrs on solid leafstalks. **Poisonous look-alikes:** Leaves resemble rhubarb leaves but flowers are greatly different.

Salsify, Oyster plant *Tragopogon porrifolius*

Habitat: roadsides, fields, railroad right-of-ways, other disturbed sites; introduced weed from Europe. **Identification:** ***biennial or perennial herb*** to 1 m (3.2 ft) with long taproot and milky juice; ***stems*** upright, ***swollen just below flowering heads.*** Leaves alternate, lance-shaped, 18–32 cm (7–12.6 in) long, to 2 cm (0.8 in) wide, long-pointed at tip, entire along margin, base circling stem. ***Flowering heads*** solitary, ***usually 3–4 cm*** (1.2–1.6 in) ***long, with 6–11 lance-shaped outer bracts and numerous purple flowers.*** Fruiting heads globe-shaped, fluffy, consisting of many small nutlets, each tipped with featherlike bristles. **Harvest:** This is the cultivated species of salsify or oyster plant that also grows wild. Roots and leaves must be harvested before plant produces flowerstalk. This is a problem since grasslike leaves grow among leaves of many other plants. Learn to recognize leaves so you can harvest them and dig roots while still palatable. **Preparation:** Scrub and scrape roots as you would carrots, putting them in cold water to prevent discoloring. Cut into ½-in slices. If they are tough, add pinch of baking soda to cooking water and boil for 10 min, then change water and cook until tender. Drain and serve with butter and seasonings or with white sauce. For other uses see GOAT'S BEARD (next). **Related edible species:** GOAT'S BEARD. **Poisonous look-alikes:** none in flowering stage.

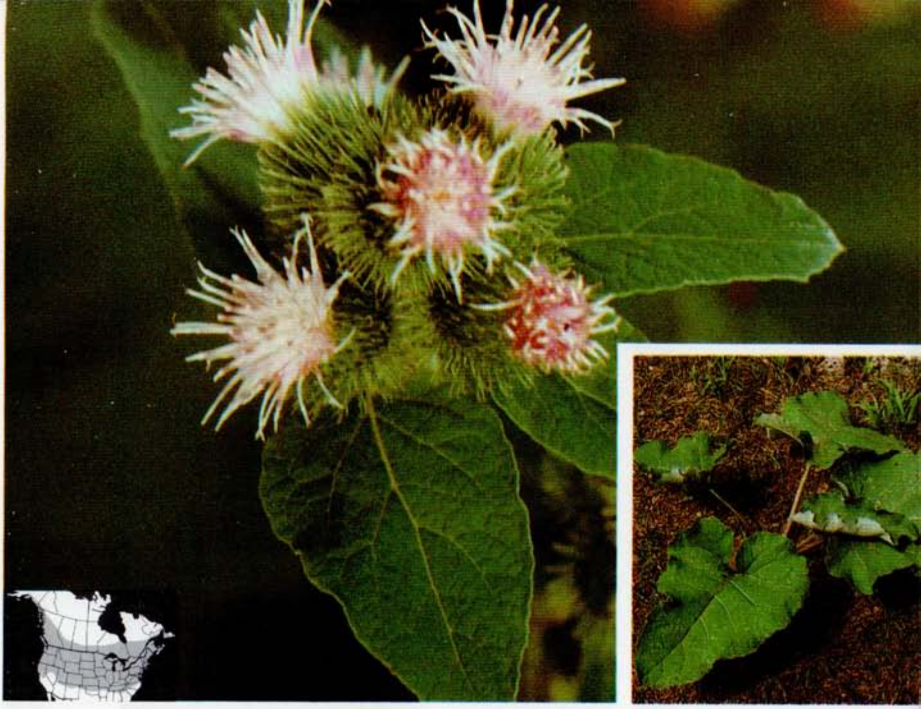

Common burdock: flowering stalk (Speas). Basal leaves (Dykeman).

Salsify: flowers, side view (NYBG). Flower (Ross).

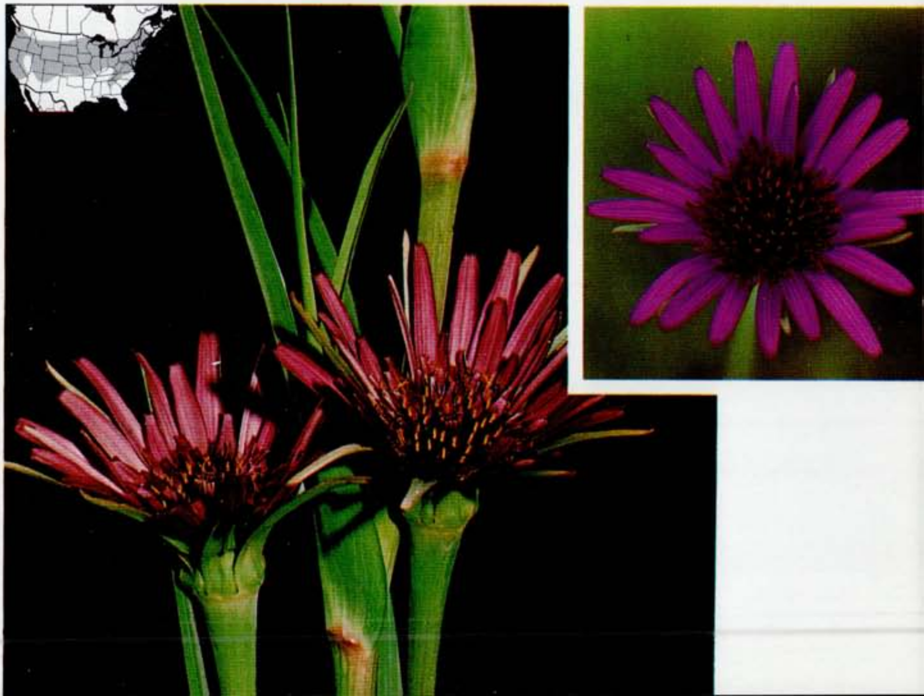

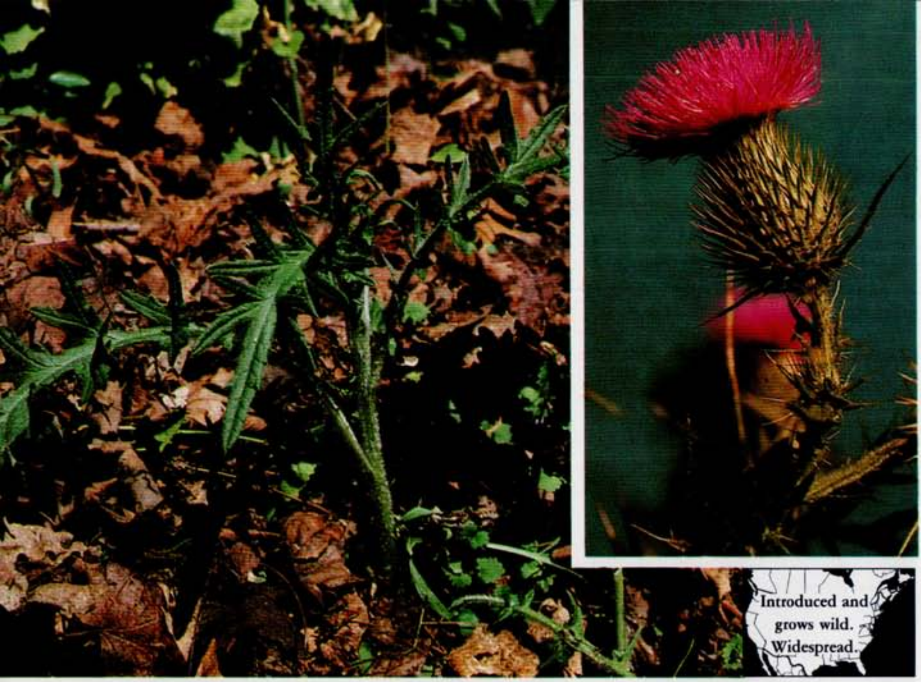
Introduced and
grows wild.
Widespread.

Bull thistle: young plant (Dykeman). Flowering head (Elias).

Chicory: basal leaves and young plants (Dykeman). Flowers (NYBG).

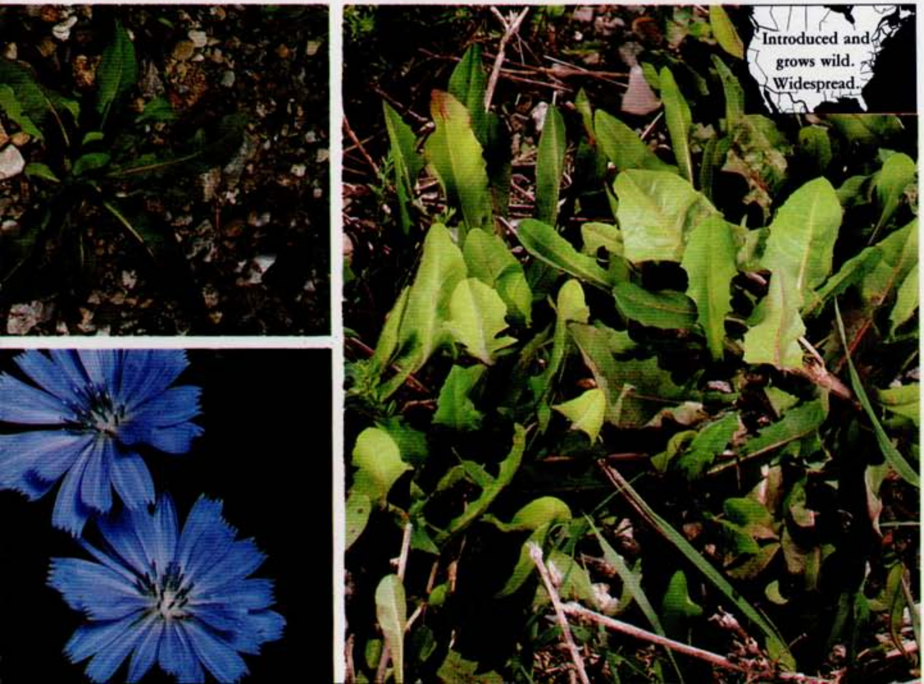
Introduced and
grows wild.
Widespread.

Bull thistle *Cirsium vulgare*

Habitat: fields, meadows, roadsides, disturbed sites; introduced from Europe and Asia. **Identification:** ***biennial herb*** to 1.8 m (6 ft) high, from taproot; ***stems*** upright, hairy, ***winged, wings lobed and armed with spines. Leaves*** alternate, broadest near or above middle, ***deeply and variously dissected along margin,*** 8–30 cm (3.2–12 in) long, 4–12 cm (1.6–4.7 in) wide, ***stiff, spiny,*** stiffly hairy above, with soft white to green hairs below, leaf bases extending onto stem. ***Flowering heads*** second year, several, outer bracts spiny, 2–4 cm (0.8–1.6 in) long; flowers ***purple.*** Seeds are hard, 3–4 mm (0.2 in) long and tipped with bristles. **Harvest:** Dig roots of first-year growth (thistles without stems). With gloves and knife, harvest leaves and young second-year stems at top of root early in growing season. **Preparation:** Peel young stems and remove spines from leaves before using raw in salad or boiling until tender for green vegetable. Remove all spiny covering from core of leaf bases. Boil cores with washed, peeled roots for about 15 min, changing water if necessary to remove bitterness. Serve salted with butter or vinegar. Slice boiled roots and fry until browned. Boil roots several hours until mushy, dry and grind into flour. Because of spines, you may prefer to use thistles only as survival food. **Related edible species:** Many species of thistles are available; some are more palatable than others. **Poisonous look-alikes:** HORSE NETTLE is a low herb with prickly stems but no spines on leaves.

Chicory *Cichorium intybus*

Also edible autumn, winter

Habitat: roadsides, fields, vacant lots, other disturbed sites; introduced weed from Europe. **Identification:** ***perennial herb*** to 1.5 m (5 ft) high, ***with large deep taproot;*** stems erect or nearly so, often branching. ***Leaves*** alternate, lance-shaped to broadest above middle, ***10–32 cm*** (4–12.6 in) ***long,*** 2–8 cm (0.8–3.1 in) wide, upper leaves reduced, ***toothed and often lobed along margin,*** sap milky. ***Flowering heads produced along upper part of stems in junctions of upper smaller leaves; each head*** 1–1.6 cm (0.4–0.6 in) long, ***3–4 cm*** (1.2–1.6 in) ***across in flower, containing many bright blue, rarely pink or white, 5-lobed flowers.*** Fruits are tiny, brown or black nutlets 2–4 mm (0.1 in) long. **Harvest:** Pick leaves, including crown at top of root, as early as possible. Dig root, fall through spring. Chicory leaves are similar to those of DANDELION; collect and use leafy parts of the two together. **Preparation:** Slice bleached underground portion of leaf crown for use raw in salad, along with oil and vinegar or with other spring edibles. Boil crown bases about 5 min and serve with butter and seasonings. Same for leaves but water changes may be needed to remove bitterness, especially later in spring. For coffee substitute, scrub roots and roast slowly in oven until brown throughout. Grind and brew like coffee. **Related edible species:** none, but similar to DANDELION in early stages. **Poisonous look-alikes:** none.

Henbit *Lamium amplexicaule*

Habitat: fields, yards, vacant lots, disturbed sites; introduced from Eurasia, common and widespread. **Identification:** small ***annual herb*** to 40 cm (16 in) tall; ***stems*** usually branching with upper parts upright, ***4-sided. Leaves opposite,*** mainly on lower half of stem, 0.7–3.4 cm (0.3–1.3 in) long, ***nearly rounded, coarsely round-toothed along margin,*** heart-shaped at base, hairy above, lower leaves long-stalked. ***Flowers purple,*** produced near tip, each ***narrow, funnel-shaped,*** 12–18 mm (0.5–0.7 in) long, ***2-lipped,*** lower lip 2-lobed, upper lip 3-lobed, hairy on outer surface. Fruits are 4 small nutlets, partially enclosed by persistent calyx. **Harvest:** young leaves and shoots in spring. **Preparation:** Combine with other greens in salads; use young leafy shoots as potherb. For spiced henbit, chop 4 cups shoots, cover with water, and boil 10 min. Separately melt 3 tbsp butter or margarine, add 1 tsp curry powder, 2 whole cloves, and ¼ tsp ground cinnamon. Stir and cook 1 min. Stir in 2 tbsp flour and cook 1 more min. Add ½ cup water from boiling henbit and stir until smooth. Add cooked henbit and ¾ cup sour cream. Simmer 15 min gently. **Related edible species:** Henbit *(L. purpureum),* with the same distribution, is very similar except leaves are stalked. **Poisonous look-alikes:** none.

Wood violet *Viola palmata*

Also edible summer

Habitat: wooded floodplains, rich deciduous forests, especially on slopes and ledges. **Identification:** ***perennial herb*** from thick, elongated underground stem (rhizome); above ground stem absent. ***Leaves*** simple, ***6–12 cm*** (2.4–4.7 in) ***long,*** often wider, broadest near base, ***blade deeply 5- to 11-lobed,*** lobes variously toothed, smooth or hairy, leafstalks long, to 20 cm (8 in); with linear to narrowly lance-shaped bracts (stipules) at base of each leaf. ***Flowers produced singly at tip of stalk as long as leaves, deep reddish-violet to violet,*** 1.7–4 cm (0.7–1.6 in) across, petals 5, lateral ones bearded, with short spur in back. Fruits are egg-shaped capsules 6–12 mm (0.2–0.5 in) long. **Harvest:** Gather flowers, leaving stems behind. Give preference to bright-green, young, tender leaves; but older darker-green summer leaves are edible too. **Preparation:** Use fresh flowers as trail snack or for jelly, jam, syrup, candy. For jelly, cover 2 cups flowers with boiling water in jar and leave overnight. Then strain and discard flowers. Add juice of 1 lemon and 1 pkg pectin. Bring to boil and add 4 cups sugar. Boil again for 1 min. Pour into sterilized jars and seal. For syrup, pour 2 cups boiling water over 6 cups flowers and leave overnight. Strain and discard flowers. Combine liquid with 2 cups sugar. Boil until syrup thickens. Store in covered jars in refrigerator. For other uses, see CANADA VIOLET. **Related edible species:** CANADA VIOLET, PALE VIOLET, western mountain violet, *(V. purpurea),* and southern wood violet *(V. hirsutula).* It is likely that all of the violet species are edible, with some more palatable than others. Some may have a cathartic effect. **Poisonous look-alikes:** Violet leaves resemble leaves of other wildflowers, some of which are toxic. So it's safest to collect leaves only when flowers are in bloom.

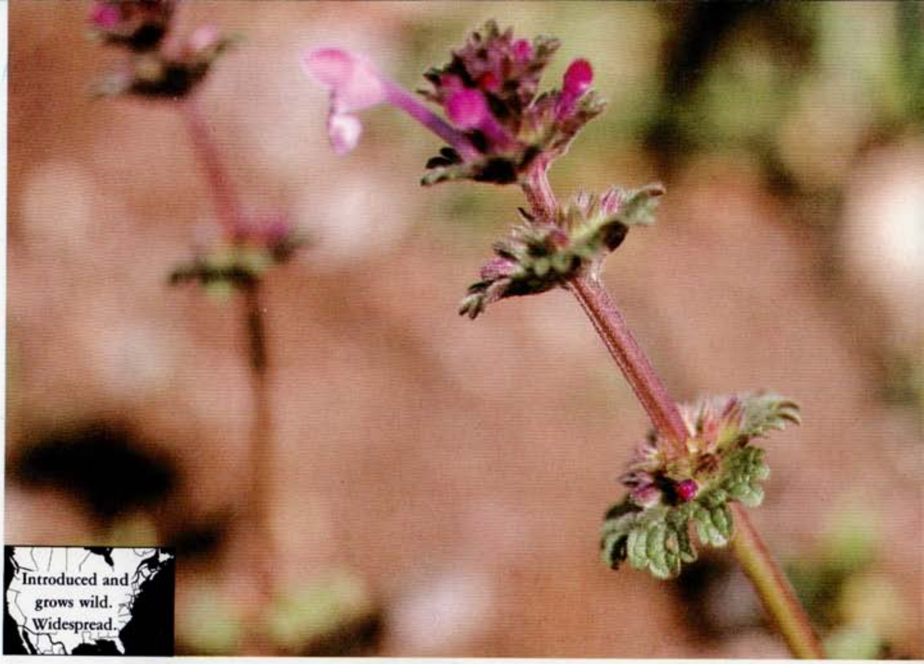
Introduced and
grows wild.
Widespread.

Henbit: flowering stock (Elias).

Wood violet: flowering plant (Elias).

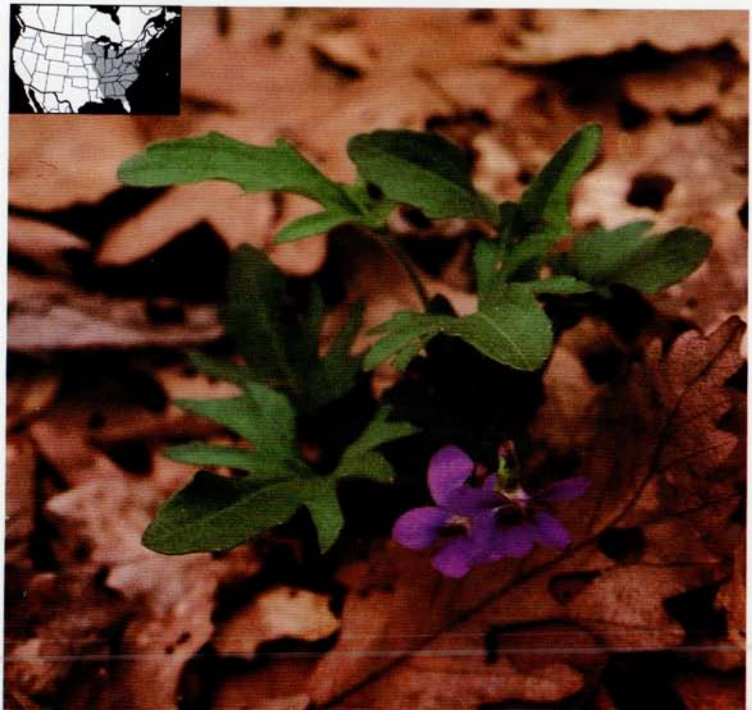

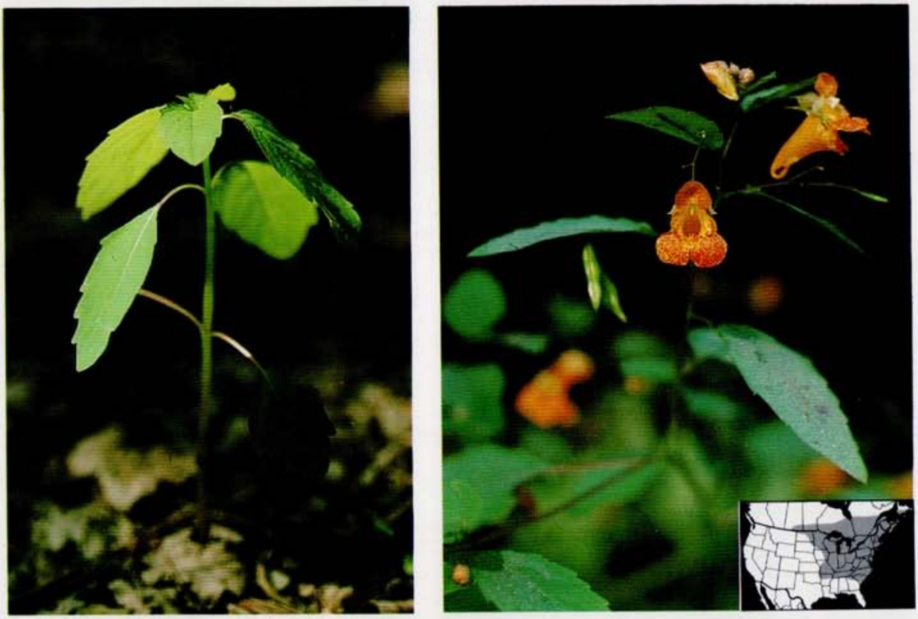

Orange-flowered jewelweed: flowering plant (Dykeman). Young plant (Elias).

Yellow-flowered jewelweed: flowering stalks (Elias).

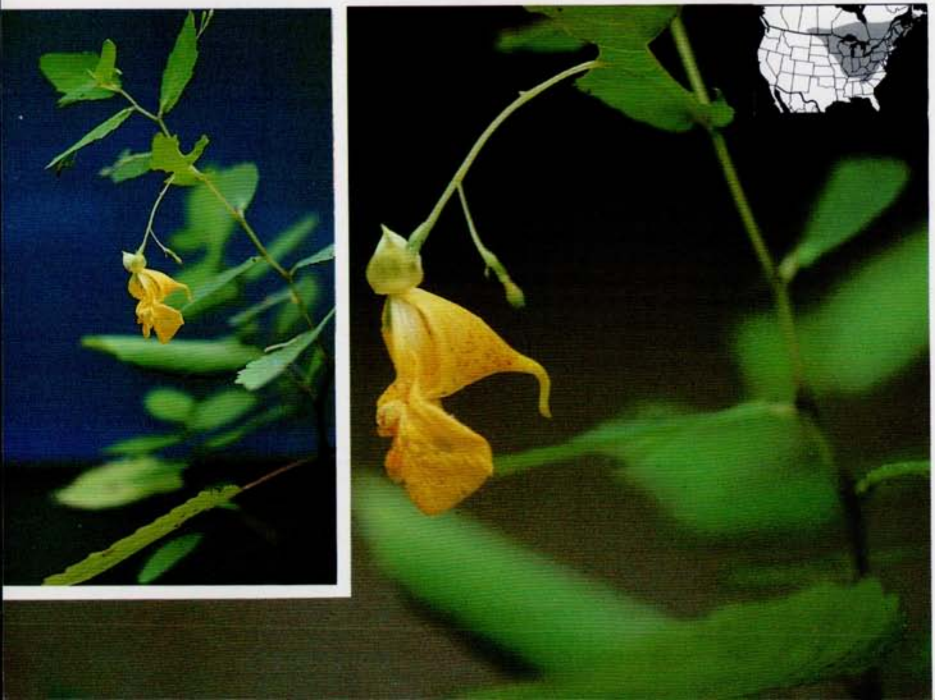

Orange-flowered jewelweed, Spotted touch-me-not *Impatiens biflora*
(synonym: *I. capensis*)

Habitat: widespread in wet areas, along streams, ponds, springs, swamps, low moist woodlands, floodplains, bottomlands. **Identification:** annual herb; ***stems*** upright, to 1.5 m (5 ft) high, branched, ***very juicy,*** smooth. Leaves alternate, 4–12 cm (1.6–4.7 in) long, 3–7 cm (1.2–2.8 in) wide, broadest at or below middle, blunt- or round-toothed along margin, long-stalked. Flowers showy, in few-flowered, spreading clusters; each ***flower*** stalked, ***with conspicuous saclike spur in rear, orange to orange-yellow spotted with reddish, yellow, or white spots on face. Fruit*** elongate, cylinder-shaped capsule 1.5–2.5 cm (0.6–1 in) long, at maturity ***exploding suddenly when touched to scatter seeds.*** **Harvest and preparation:** For potherb, pick ***young shoots*** to 15 cm (6 in) high; wilts quickly, so use immediately. Cover with water; boil 10–15 min, draining and adding fresh boiling water twice. Season and butter. *Caution:* Frequent use as a potherb is not recommended because of abundance of calcium oxalate crystals. Seeds: As pods ripen, shake plant over fine net; separate seeds from rest of fruit. Tastes like BUTTERNUTS; eat as gathered or use as topping for desserts. Medicinal use: Applied to skin, juice prevents and relieves symptoms of POISON IVY, NETTLES, and fungal dermatitis. Apply raw juice or concentrated boiled juice of crushed stems. Extract spoils quickly, but frozen extract retains medicinal properties. Antifungal agent is 2-methoxy-1, 4-naphthoquinone. **Related edible species:** YELLOW-FLOWERED JEWELWEED (next). **Poisonous look-alikes:** none.

Yellow-flowered jewelweed, Pale touch-me-not *Impatiens pallida*

Habitat: moist woodlands, wet meadows; not as abundant and usually occurring in slightly drier sites than ORANGE-FLOWERED JEWELWEED. **Identification:** annual herb; ***stems*** upright, to 2 m (6.6 ft) high, freely branched, ***very juicy,*** smooth. Leaves alternate, long stalked, blades 4–14 cm (1.6–5.5 in) long, 3–8 cm (1.2–3.2 in) wide, broadest at or below middle, blunt or round toothed along margin. ***Flowers showy, in few-flowered clusters*** in junctions of upper leaves; ***each flower*** stalked, ***with conspicuous saclike spur in rear, pale yellow with reddish spots on face. Fruit elongate, cylinder-shaped capsule 2–2.6 cm (0.8–1 in) long, at maturity exploding suddenly when touched to scatter seeds.*** **Harvest and preparation:** same as for ORANGE-FLOWERED JEWELWEED (preceding). **Related edible species:** ORANGE-FLOWERED JEWELWEED. **Poisonous look-alikes:** none.

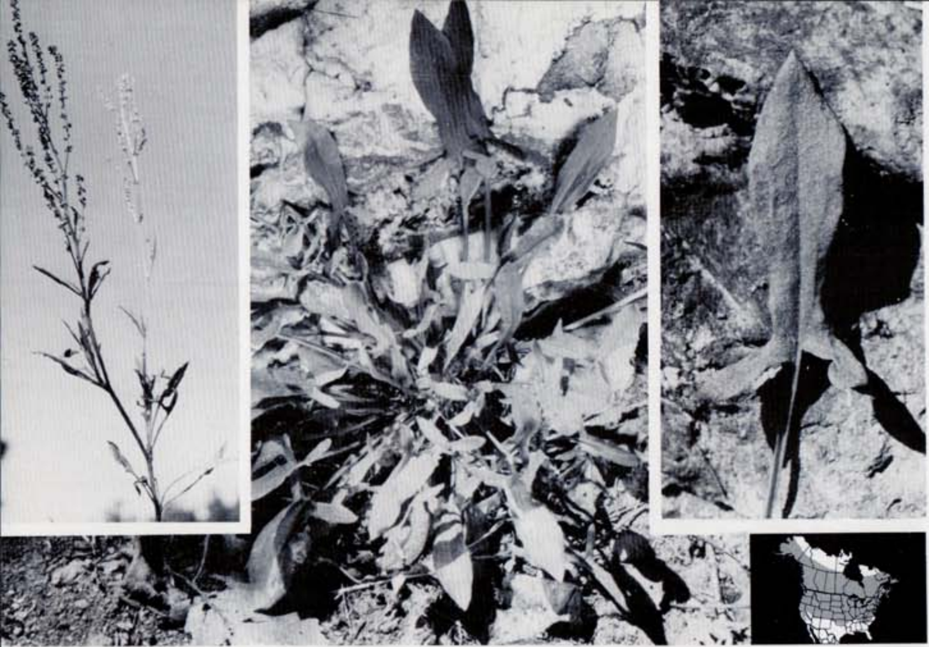

Sheep sorrel: flowering stalk (Elias). Basal leaves and leaf (Dykeman).

Sour dock: flowering stalks (Elias).

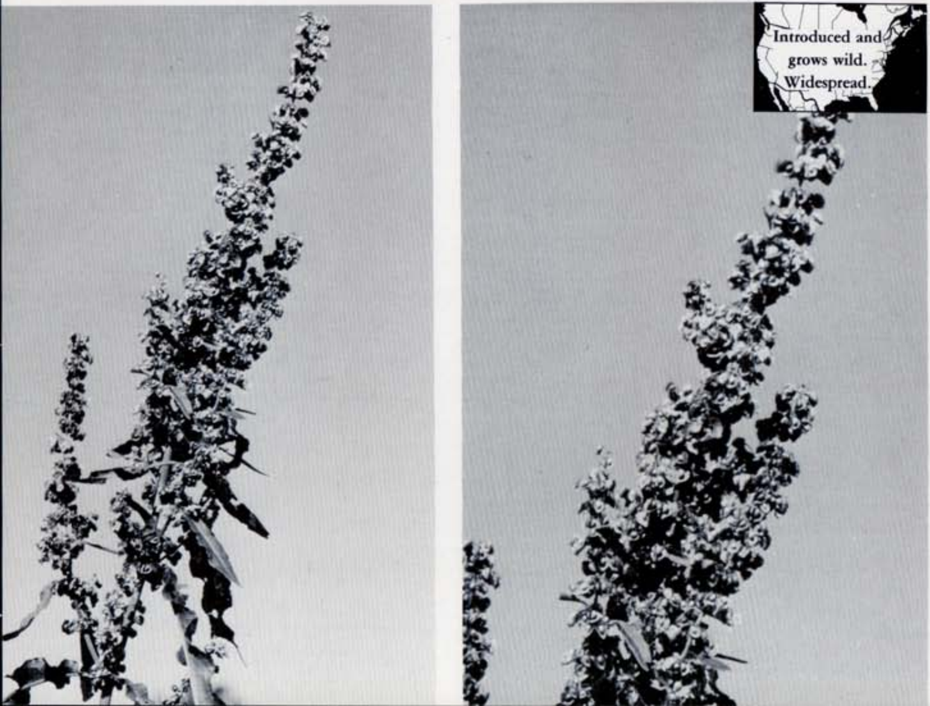
Introduced and grows wild. Widespread.

Sheep sorrel *Rumex acetosella*

Also edible summer, autumn

Habitat: disturbed sites, gardens, old fields, mainly on acid soils. **Identification:** annual or perennial ***herb*** to 50 cm (20 in); stems upright, 4-sided, unbranched or branched. Basal leaves present, shaped like a narrow arrowhead. ***Stem leaves 2.4–5 cm*** (1–2 in) ***long,*** 0.5–2 cm (0.2–0.8 in) wide, ***narrow arrowhead-shaped,*** pointed at tip, ***with slender pointed lobes at base,*** with papery sheath around stem where leaves are attached. ***Flowers tiny,*** green, yellow, red to purple, ***clustered on narrow elongate spikes.*** Fruits are tiny nutlets 1.2–1.8 mm (less than 0.1 in) long, 3-angled, shiny yellow-brown, enclosed by persistent calyx. **Harvest:** Pick or cut leaves from healthy, deep green clumps from spring until autumn, but flowering clumps may be tough. **Preparation:** Sour taste makes sheep sorrel leaves an excellent thirst-quenching trail nibble. Use leaves liberally with bland greens in tossed salad. For beverage boil 1 cup loosely packed leaves in 1 qt water for 2–3 min; then cover and steep for at least 15 min. Strain, add sugar or honey to taste, and serve hot or cold. Boil leaves in 2 changes of water for pleasantly sour potherb or mix with other cooked greens. Use leaves with chicken broth to make soup or sauce to use with fish, rice, or vegetables. *Caution: Large quantity may cause stomach upset.* **Related edible species:** Many species of dock, including SOUR DOCK (next), are edible; sheep sorrel has the most distinctive sour flavor. **Poisonous look-alikes:** none.

Sour dock, Curled dock, Yellow dock *Rumex crispus*

Also edible summer, autumn, winter

Habitat: common in disturbed sites, fields, roadsides, vacant lots. **Identification:** ***large herb,*** usually biennial, to 1.5 m (5 ft) tall; stems upright, smooth, ribbed. ***Leaves*** 14–24 cm (5.5–9.4 in) long, 2–6.5 cm (0.8–2.5 in) wide, uniformly wide to widest near base, pointed at tip, ***strongly wavy and curled along margin,*** rounded to heart-shaped at base, dark green, ***with papery, straw-colored sheath around stem where leaves are attached.*** Flowers small, usually greenish, stalked, densely clustered on 1 to several narrow, elongated stalks. ***Fruits 3-parted,*** 4–6 mm (0.2 in) across, dark red to reddish-brown, each ***enclosing*** 3 ***nutlets*** 1.5–3 mm (0.1 in) in diameter. **Harvest:** leaves in early spring, or even through late winter when they may be tinged with red and yellow. Strip seeds from stalks in late summer and autumn. **Preparation:** Young leaves cooked for about 10 min in small amount of water taste like young beet greens. Older leaves may need longer cooking with 1–2 water changes to make them tender and remove bitterness. Cook with other wild greens of season. Serve with butter, bacon, hard-boiled egg, seasonings. *Caution: Large quantity may cause stomach upset.* With much effort seeds can be hulled, winnowed, ground, and sifted to produce palatable flour. Use half and half with regular flour. **Related edible species:** While most species of dock are likely edible, sour dock is one of the most flavorful. **Poisonous look-alikes:** none.

Eastern hemlock *Tsuga canadensis*

Habitat: cool moist ravines, protected valleys, lowlands. **Identification:** medium or large ***evergreen tree*** to 25 m (82 ft), rarely taller, with deeply furrowed, reddish-brown or cinnamon red bark. ***Leaves small,*** 0.7–1.6 cm (0.3–0.6 in) long, 1–2 mm (less than 0.1 in) wide, ***linear, arranged in 1 plane horizontal to ground, flattened,*** dark yellowish-green ***with 2 whitish bands underneath.*** Male flowering cones tiny, yellowish, at tips of branches; female cones small, green, leathery. ***Fruiting cones 1.2–1.9 cm*** (0.5–0.7 in) ***long,*** hanging, widest near middle, rounded or pointed at tip, ***light brown;*** cone scales nearly rounded; seeds tiny, egg-shaped, winged. **Harvest:** needles any time but preferably young ones in spring. Inner bark best in winter and early spring. **Preparation:** For tea, boil needles in covered pot and steep about 10 min. Tea is high in vitamin C. To avoid damage to tree, use inner bark only in emergency. It may be eaten raw or boiled, or dried and ground to mix with flour. **Related edible species:** Carolina hemlock *(T. caroliniana),* western hemlock *(T. heterophylla),* mountain hemlock *(T. mertensiana).* **Poisonous look-alikes:** AMERICAN YEW is a commonly cultivated shrub that has longer needles and bright red, fleshy pulp surrounding seeds.

Yellow birch *Betula alleghaniensis*
Also edible autumn, winter (synonym: *B. lutea*)

Habitat: rich woodlands, lower slopes, and occasionally cool marshlands, usually below 1,000 m (3,300 ft) elevation. **Identification:** medium-size tree with rounded crown and peeling ragged-edged, reddish-brown ***bark*** turning ***dull yellow or yellowish-brown. Broken twigs have mild wintergreen fragrance.*** Leaves alternate, deciduous, simple, 7–11 cm (2.8–4.3 in) long, 3–5 cm (1.2–2 in) wide, usually broadest near base, with double row of sharp-pointed teeth along margin. Male and female flowers in separate catkins on same tree in spring. ***Fruits broad, upright cones 2.5–3.5 cm (1–1.4 in) long,*** with numerous 3-lobed scales and tiny winged seeds. **Harvest:** For sap, tap like SUGAR MAPLE, 3–4 weeks after maple sap flow begins. Flow copious but sugar content lower than from maple. For use of twigs and inner bark, see instructions under more flavorful SWEET BIRCH (next). **Preparation:** Boil sap in shallow open container outdoors, adding more as volume decreases, until evaporation leaves viscous, molasses-flavored syrup with temperature about 104°C (220°F). Ratio of sap to ultimate syrup is far greater than for SUGAR MAPLE. Store in sterilized, filled, sealed jars. **Related edible species:** SWEET BIRCH. **Poisonous look-alikes:** none.

Sweet birch, Black birch *(Also edible summer, autumn, winter)* *Betula lenta*

Habitat: forest or open woods, especially moist, north-facing, protected slopes, in deep, rich, well-drained soils. **Identification:** medium-size tree with rounded crown and ***smooth, dark red to almost black bark. Broken twigs have wintergreen fragrance.*** Leaves alternate, deciduous, simple, 6–15 cm (2.4–6 in) long, 4–8 cm (1.6–3.2 in) wide, broadest near base, ***with double row of fine, sharp-pointed teeth along margin.*** Male and female flowers in separate, small catkins on same tree in spring. Fruits in erect, brownish cones 2.5–3 cm (1–1.2 in) long, containing many tiny, winged seeds. Fruits maturing in late summer and early

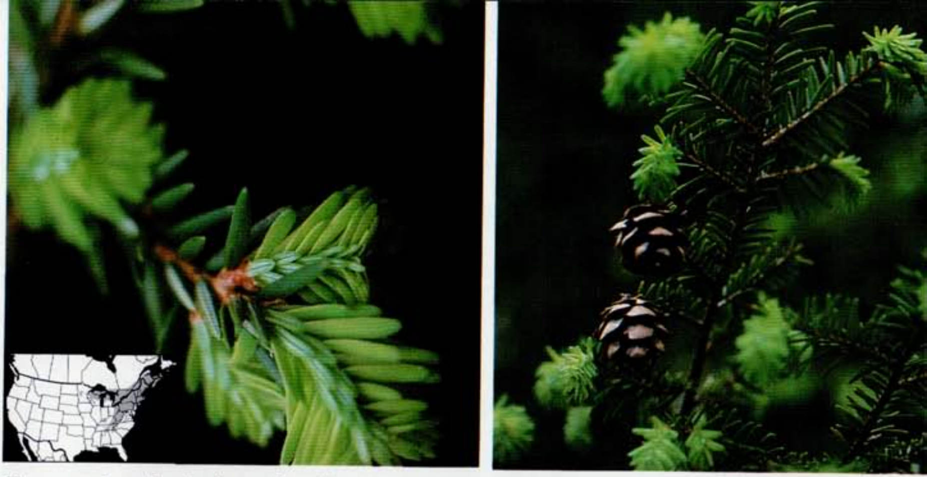

Eastern hemlock: branch with young leaves and branch with mature cones (Elias).

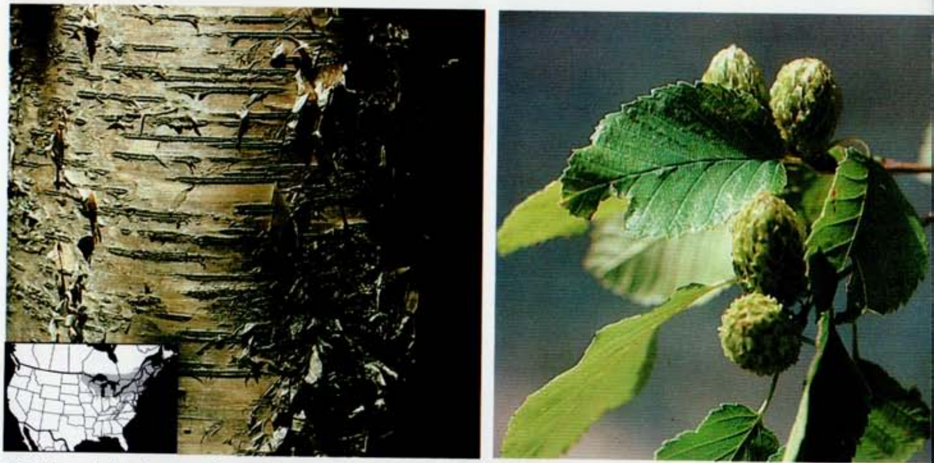

Yellow birch: trunk (Dykeman). Branch with fruits (Elias).

Sweet birch: spring branch with male flowers (Dykeman). Fruit in summer (Elias). Trunk (Dykeman).

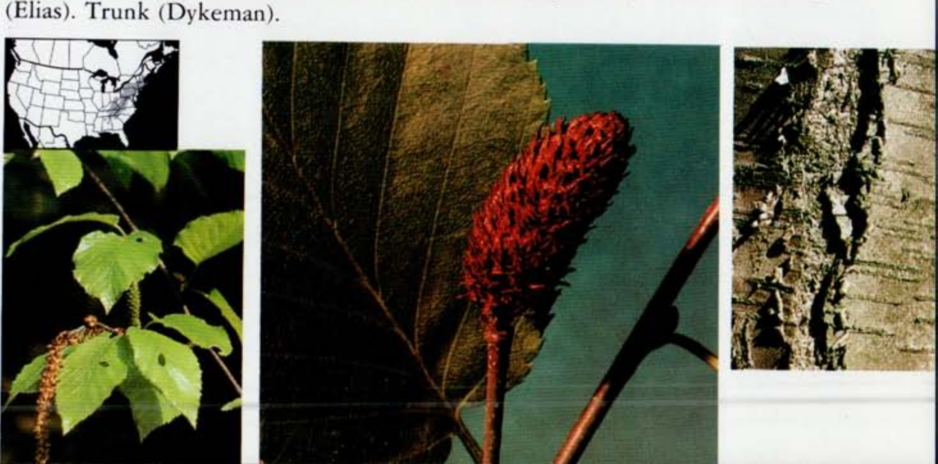

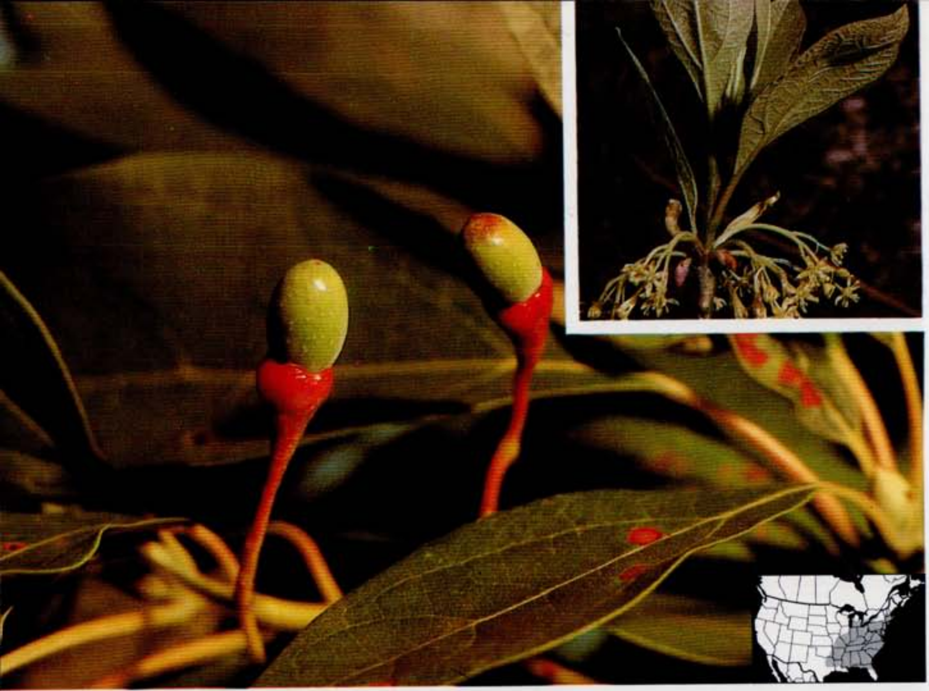

Sassafras: summer branch with fruits (Elias). Spring branch with flowers (NYBG).

Sugar maple: fruiting branch (Soderstrom).

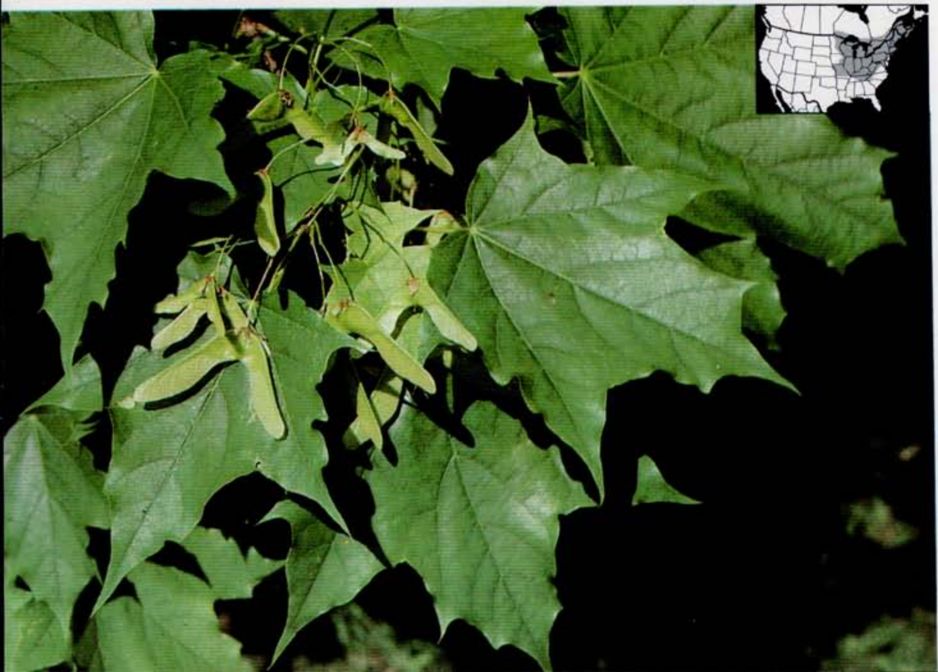

fall. **Harvest:** twigs, red inner bark, larger roots: year-round but best in spring. For sap in spring, 3–4 weeks later than SUGAR MAPLE. See instructions under YELLOW BIRCH (preceding). **Preparation:** inner bark raw as emergency food, boiled like noodles, or dried and ground into flour. Dry inner bark at room temperature; store in sealed jars. For tea, steep (do not boil) twigs or fresh or dried inner bark in water or, preferably, birch sap. Boiling destroys volatile wintergreen oil. For birch beer, pour solution of 4 gal birch sap and 1 gal honey (or 5 gal sap and 3 lb sugar), which has been boiled 10 min, over 4 qt fine twigs in crock. Cool, strain to remove twigs, add 1 cake yeast. Cover; ferment about 1 week, until cloudiness starts to settle. Bottle and cap tightly. For birch syrup, see YELLOW BIRCH. **Related edible species:** YELLOW BIRCH. **Poisonous look-alikes:** none.

Sassafras *(Also edible summer, autumn, winter)* *Sassafras albidum*

Habitat: bottomlands, moist soil of open woodlands, lower slopes. **Identification:** small to medium-size tree with open, flat-topped crown and smooth to deeply and irregularly furrowed, reddish-brown bark on older trees. Leaves alternate, deciduous, simple, ***in 3 shapes (entire or with 2 or 3 lobes), 10–15 cm (4–6 in) long,*** 5–10 cm (2–4 in) wide, broadest near base, ***bright green, with spicy odor when crushed.*** Flowers produced in clusters in spring, each small, yellow, and either male or female (each tree male or female). Fruits egg-shaped 1–1.5 cm (0.4–0.6 in) long, dark blue, stalked, enclosed at base by small red cup. **Harvest and preparation:** year-round. Carefully pull several suckers near base of parent tree to obtain small roots. For larger harvest, brittle roots require digging. Flavor is in root bark. Boil washed roots or root bark in water (or maple sap) until liquid is red. Sweeten to taste; drink hot or cold. Roots may be reused. Dried, pulverized root bark is pungent meat spice. Dried, pulverized leaves flavor and thicken soups and gravies. Use green winter buds and young leaves in salads. *Caution:* Sassafras reportedly has carcinogenic properties. Basic jelly recipe: Make jelly with 3 cups strong sassafras tea, 3 oz powdered pectin, ½ tsp citric acid, 4 cups sugar; or use 2 cups tea, 1 pkg pectin, 3 cups honey, powdered root bark to taste. **Related edible species:** SPICE BUSH. **Poisonous look-alikes:** none.

Sugar maple *(Also edible summer, winter)* *Acer saccharum*

Habitat: eastern deciduous forest, common and widespread, especially at lower elevations. **Identification:** medium-size to tall ***tree*** to 30 m (98 ft) with rounded crown, slightly fissured to furrowed dark gray bark. ***Leaves opposite,*** simple, 7.5–20 cm (3–8 in) long and broad, broadest near base, ***3- to 5-lobed, lobes with a few coarse teeth along margin,*** bright green, smooth above and below. Flowers small, bisexual, produced in slender hanging clusters before leaves emerge. ***Fruits are winged (samaras), 3–3.5 cm*** (1.2–1.4 in) ***long,*** wings spreading at 60° angle. **Harvest:** Obtain sap from first spring flow, until buds swell. Good sap flow requires freezing nights and warm days. Drill hole [11 mm (7/16 in) for commercial spile] slightly upward and 5–8 cm (2–3 in) into trunk. Drive spile in far enough to hold bucket, not so hard to split wood. Collect sap and store cold until you have 25–30 gal. Collect winged seeds before fully ripe, usually from June to September. **Preparation:** Boil sap outdoors, indoors only if you like sticky walls. Use large surfaced, open pan, adding sap as water boils off. After adding all sap, continue boiling until temperature is 4°C (7°F) above that of boiling water on that particular day. You may wish to boil for the last two degrees on a controllable stove. Pour syrup through milk filter, filling sterilized canning jars, and seal. Yields about 3 qt. For maple sugar, boil 1 pt syrup in saucepan with frequent stirring until temperature is at least 110°C (230°F). Remove from heat, cool and stir until surface loses sheen. Pour quickly onto foil or waxed paper. When partially hardened, score into pieces with knife. Use sap fresh as beverage, fermented to form mild vinegar, or reduced by boiling to ¼ original volume as base for teas. Soak seeds and remove wings. Boil until tender, drain, season, and roast for 10–15 min. **Related edible species:** Sap of all native maples may be used for syrup. Sugar maple is sweetest. **Poisonous look-alikes:** none; avoid Norway maple *(A. platanoides),* which has milky sap.

Summer

By summer most of the leafy edibles that were tender in spring have toughened and become unpalatable. Yet many late-starting plants are usable through the summer and include the mints of fields and water's edge as well as purslane from the untended garden. Spring strawberries are succeeded by a host of other delicious fruits including raspberries, mulberries, blueberries, blackberries, serviceberries, and beach plums. So summer is the prime season for pies, jams, jellies, and juices, as well as food-drying projects.

As the days grow shorter in late August and September, fewer and fewer shrubs and trees yield juicy, flavorful fruits. Yet this slow period before the fall fruit and nut season can still keep a forager busy. Now you can find ripe seed clusters from many of the same plants whose greens served as edibles in spring; some of these include lamb's quarters, amaranth, plantain, and dock. Threshed, winnowed, and ground, these seeds provide tasty and nutritious flour supplements.

Of course, summer is also the season to beware of stinging nettles, poison ivy, and pesky briars, the briars coming in many disguises.

Additional summer edibles

Plants listed below have parts that are edible in summer and at least one additional season. Descriptions and photos of these plants appear in a seasonal section other than "Summer," as indicated.

Allegheny chinkapin (*Castanea pumila*), p. 234
Arrowhead (*Sagittaria latifolia*), p. 212
Black mustard (*Brassica nigra*), p. 100
Bullbrier (*Smilax bona-nox*), p. 66
Butternut (*Juglans cinerea*), p. 246
California grape (*Vitis californica*), p. 216
California rose (*Rosa californica*), p. 220
Canada violet (*Viola canadensis*), p. 94
Carrion flower (*Smilax herbacea*), p. 66
Cattail (*Typha latifolia*), p. 68
Common burdock (*Arctium minus*), p. 112
Common chickweed (*Stellaria media*), p. 84
Common greenbrier (*Smilax rotundifolia*), p. 64
Common milkweed (*Asclepias syriaca*), p. 106
Eastern camass (*Camassia scilloides*), p. 64
Fireweed (*Epilobium angustifolium*), p. 108
Frost grape (*Vitis vulpina*), p. 214
Garlic mustard (*Alliaria officinalis*), p. 84
Glasswort (*Salicornia europea*), p. 76
Great bulrush (*Scirpus validus*), p. 70
Green pigweed (*Amaranthus hybridus*), p. 72
Hard-stem bulrush (*Scirpus acutus*), p. 70
James chickweed (*Stellaria jamesiana*), p. 84
Lady's thumb (*Polygonum persicaria*), p. 110
Lotus lily (*Nelumbo lutea*), p. 210
Miner's lettuce (*Montia perfoliata*), p. 94
Narrow-leaved cattail (*Typha angustifolia*), p. 68
Pale violet (*Viola striata*), p. 96
Pigweed (*Amaranthus retroflexus*), p. 72
Plantain (*Plantago major*), p. 78
Riverside grape (*Vitis riparia*), p. 214
Russian thistle (*Salsola kali*), p. 76
Sassafras (*Sassafras albidum*), p. 124
Sheep sorrel (*Rumex acetosella*), p. 120
Sour dock (*Rumex crispus*), p. 120
Spice bush (*Lindera benzoin*), p. 218
Squashberry (*Viburnum edule*), p. 226, 252
Sugar maple (*Acer saccharum*), p. 124, 256
Sweet birch (*Betula lenta*), p. 122, 252
Sweetbrier rose (*Rosa eglanteria*), p. 222
Wild garlic (*Allium canadense*), p. 60
Wild leek (*Allium tricoccum*), p. 60
Wild onion (*Allium cernuum*), p. 58
Wild potato vine (*Ipomoea pandurata*), p. 216
Wild rose (*Rosa carolina*), p. 222
Wild strawberry (*Fragaria virginiana*), p. 90
Wintercress (*Barbarea vulgaris*), p. 98
Wood violet (*Viola palmata*), p. 116
Wrinkled rose (*Rosa rugosa*), p. 222
Yellow pond lily (*Nuphar advena*), p. 212

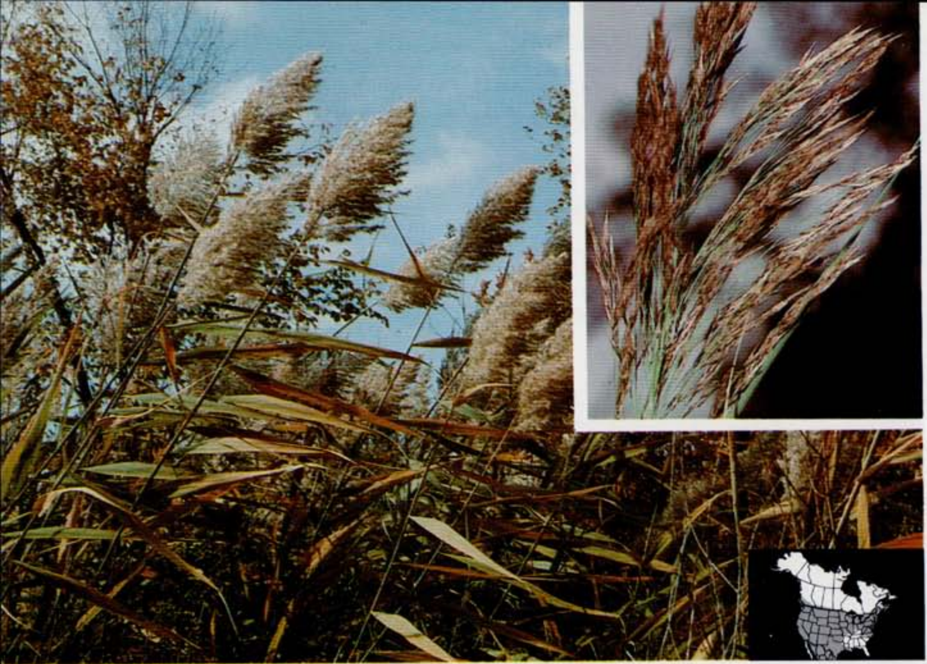

Reed grass: flowering plants (Feldman). Flowers (Mohlenbrock).

Wild rice: flower stalk with male flowers topmost and stalks with female flowers (Elias).

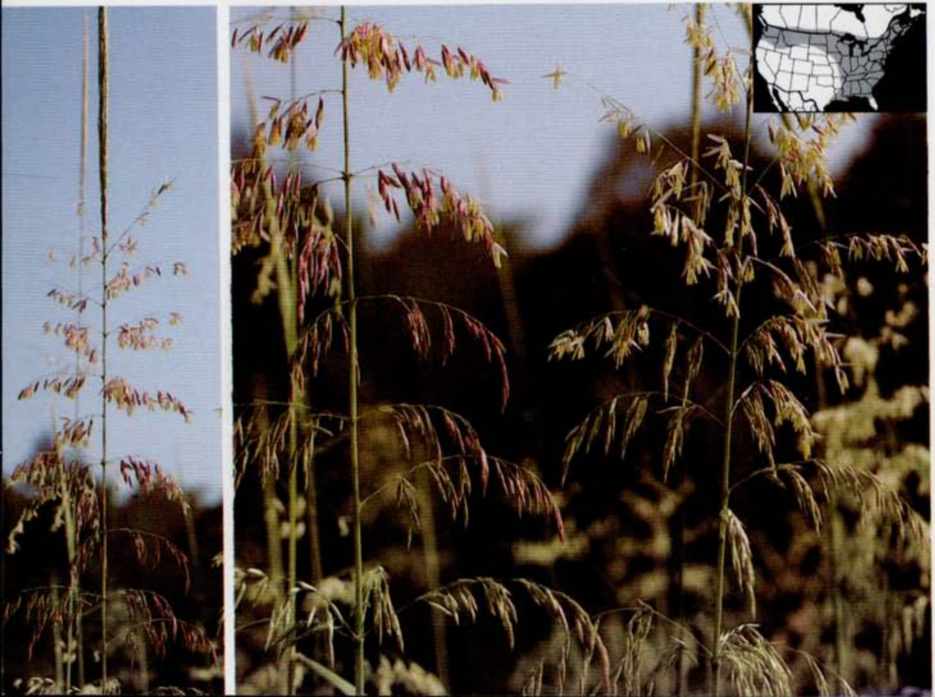

Reed grass, Phragmites *(Also edible autumn, winter, spring)* *Phragmites communis*

Habitat: margins of streams, lakes, ponds, marshes, springs, ditches. **Identification:** ***tall perennial grass,*** with long, spreading underground stems (rhizomes), often forming dense clumps; ***stems upright, 2–4 m*** (6.6–13 ft) ***high,*** stout; ***leaf blades*** flat, long, ***18–60 cm*** (7–24 in) ***long,*** narrow, 1–6 cm (0.4–2.4 in) wide, smooth, base of leaf sheathing part of stem. ***Flowers many in large, dense, elongated, branched, plumelike clusters 15–40 cm*** (6–16 in) ***long,*** purplish at first, turning tan to whitish at maturity, side branches ascending, each spikelet (small simple cluster of flowers sharing the same stalk) 1–1.6 cm (0.4–0.6 in) long, ***with numerous long, silky hairs.*** Fruits are tiny, hard, reddish grains (seeds) many people overlook. **Harvest:** Gather new shoots (at base of old stalks) and leaves early in growing season. Collect hardened sap from wounded stems. Cut whole stems before blooming. Gather seed heads in late summer and autumn. Failure to produce seed is common. Dig roots and underground stems all year. **Preparation:** Eat early shoots raw, or boil or steam until tender; serve as vegetable. Eat hardened gum fresh or toast near fire. Dry stems in sun, beat or grind, sift out fine powder. Moisten powder and place near fire to swell and toast like marshmallows. For gruel, remove seeds by hand. Crush seeds with hulls and mix with spicy berries. Add boiling water and cook until thin reddish gruel, about 30 min; add sweetener. For use of roots and underground stems, see CATTAIL. **Related edible species:** none. **Poisonous look-alikes:** none, but avoid roots and shoots of iris (YELLOW FLAG, BLUE FLAG) and other inedibles.

Wild rice *(Also edible autumn)* *Zizania aquatica*

Habitat: marshes, shallow ponds, lakes, streams, or bays of fresh and brackish water. **Identification:** ***tall annual grass;*** stems upright, usually 1–3 m (3–10 ft), stout; leaf blades long, narrow 0.3–1 m (12–39 in) long, 2–5 cm (0.8–2 in) wide, soft. ***Flowers many, in large, open, branched, elongate, plumelike clusters to 60 cm (2 ft) high,*** side branches 15–20 cm (6–8 in) long, ***lower branches*** spreading to ascending and ***only with male flowers; upper branches erect, bearing only female flowers,*** each about 2 cm (0.8 in) long; seed case tipped with long, stiff, hairlike structures; seed narrow, hard, brown. **Harvest and preparation:** Grains ripen in late summer to early fall. Timing is critical; ripe seeds easily fall from plant. Small, shallow-draft boat or canoe provides best harvesting container. Carefully bend stalks over boat and knock ripe grains into it. Several visits days apart needed for maximum harvest. Spread grains to dry; then parch, stirring occasionally, for 3 hr at 175°C (350°F). Thresh husks from grains by pounding or rubbing. Winnow by pouring from one container to another in natural or fan-generated breeze. Wash grains in cold water to eliminate smoky flavor. Use like cultivated rice. Cooking: Bring 3 cups water and 1 tsp salt to boil. Stir in 1½ cups rice grains and return to boil. Cover and simmer until tender, 30–45 min. Serve buttered or with sweetener. For flour, substitute ground rice for part of wheat flour in baking recipes. **Related edible species:** Texas wildrice *(Z. texana)*. **Poisonous look-alikes:** none, but poisonous pink or purple fungus (*Claviceps* species) occasionally replaces some grains. Learn to recognize and avoid it.

Chufa, Yellow nut grass *Cyperus esculentus*

Also edible autumn, winter, spring

Habitat: ditches, margins of ponds and streams, low wet soils, waste places. **Identification:** ***perennial herb with*** slender underground stems (rhizomes) and ***nearly rounded tubers*** 6–10 mm (0.2–0.4 in) in diameter; ***stems triangular in cross section,*** upright, 20–90 cm (8–36 in) tall, stout, smooth; stem leaves narrow, 5–10 mm (0.2–0.4 in) wide; ***with 3–10 leaves at base of flowering clusters*** but much longer than flower clusters. ***Flowering clusters broadly cylinder-shaped to almost globe-shaped,*** composed of numerous spikelets (small simple cluster of flowers on same stalk); spikelets narrow, elongate, 0.5–3 cm (0.2–1.2 in) long, 4-sided, many-flowered. Fruits are small ***nutlets, 1–1.5 mm*** (less than 0.1 in) ***long, triangular-shaped, yellowish.*** **Harvest:** Gather tubers best from loose or sandy soil where whole plant with tubers is easily pulled or dug from depth of less than a few inches. Gather any time ground is not frozen. **Preparation:** Tubers palatable and useful raw, boiled, candied, dried, and ground into flour, and as beverage base. For flour, dry slowly until tubers break apart instead of mash when hammered lightly. Grind fine in blender or food mill. Use flour half and half with wheat flour. For beverage soak 1½–2 cups tubers for 2 days. Drain and mash or blend in 1 qt water with ¼ cup sugar or honey. Strain out solids before drinking. For coffee substitute, roast tubers until dark brown throughout. Grind and brew 1 tbsp per cup. **Related edible species:** Several species of *Cyperus* have tubers or tuberlike structures. **Poisonous look-alikes:** none.

Pickerelweed *Pontederia cordata*

Also edible autumn

Habitat: along shores or in shallow water of lakes, ponds, slow moving streams, bays. **Identification:** ***perennial herb*** to 1 m (3.2 ft) high, with short, thick underground stem. ***Leaves basal, blade*** 6–20 cm (2.4–8 in) long, 2–12 cm (0.8–4.7 in) wide, broadest near base to very narrow, shape variable but usually ***arrowhead-shaped,*** smooth; ***leafstalk stout, 6–32 cm*** (2.4–13 in) ***long.*** Flowering stalk slightly taller than leaves, bearing 1 or 2 leaflike bracts, ***attractive blue flowers densely clustered on spike 5–16 cm*** (2–6.3 in) ***long, each flower funnel-shaped, 2-lipped.*** Fruits are stiff, egg-shaped capsules 4–10 mm (0.2–0.4 in) long, enclosed by persistent ragged remains of flower. **Harvest:** Pick young leafstalks in early summer before leaves are fully unfurled. Strip fruits off spikes as they ripen in early autumn. **Preparation:** Wash, chop, and add young leafstalks to salads. As potherb boil leafstalks about 10 min, drain, and season. Fruits each contain a solid seed, edible as trail snack or dried and ground like grain. **Related edible species:** LANCE-LEAVED PICKERELWEED (next), used the same as this species. **Poisonous look-alikes:** none.

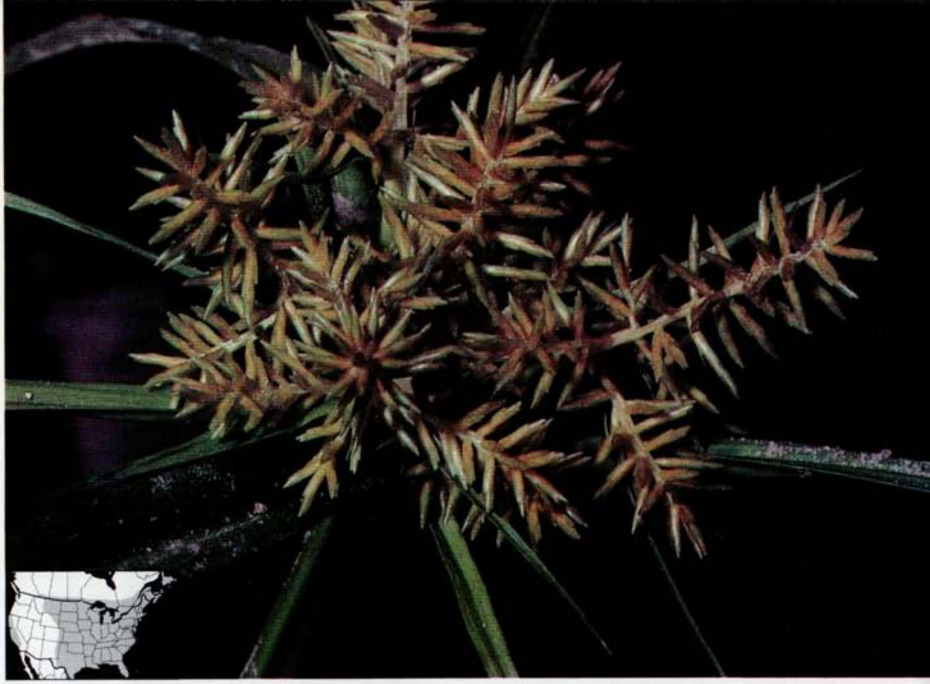

Chufa: fruiting cluster (Mohlenbrock).

Pickerelweed: flowering stalk (NYBG). Plants (Dykeman).

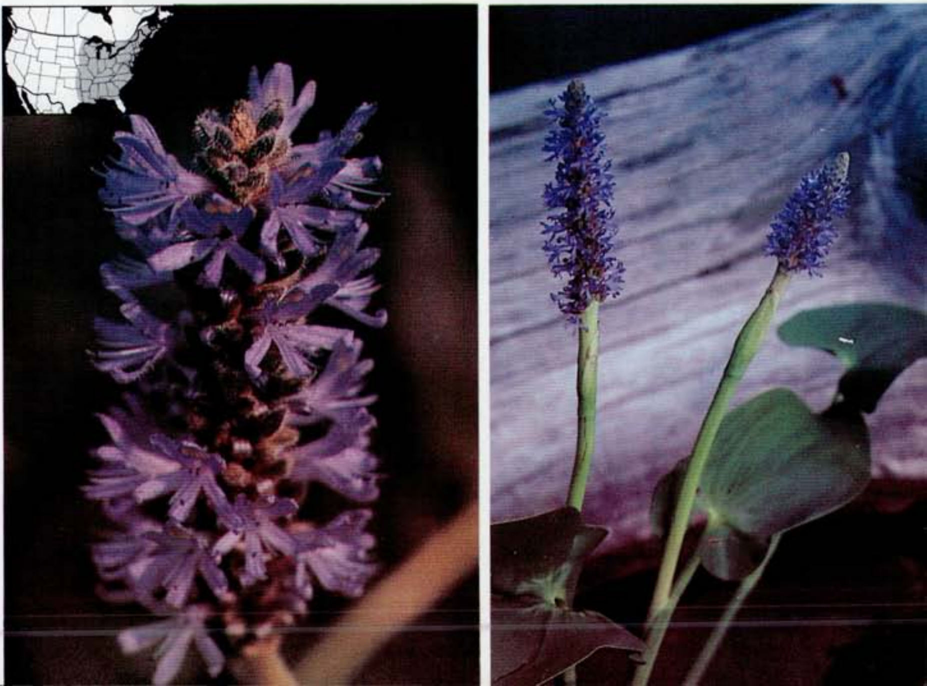

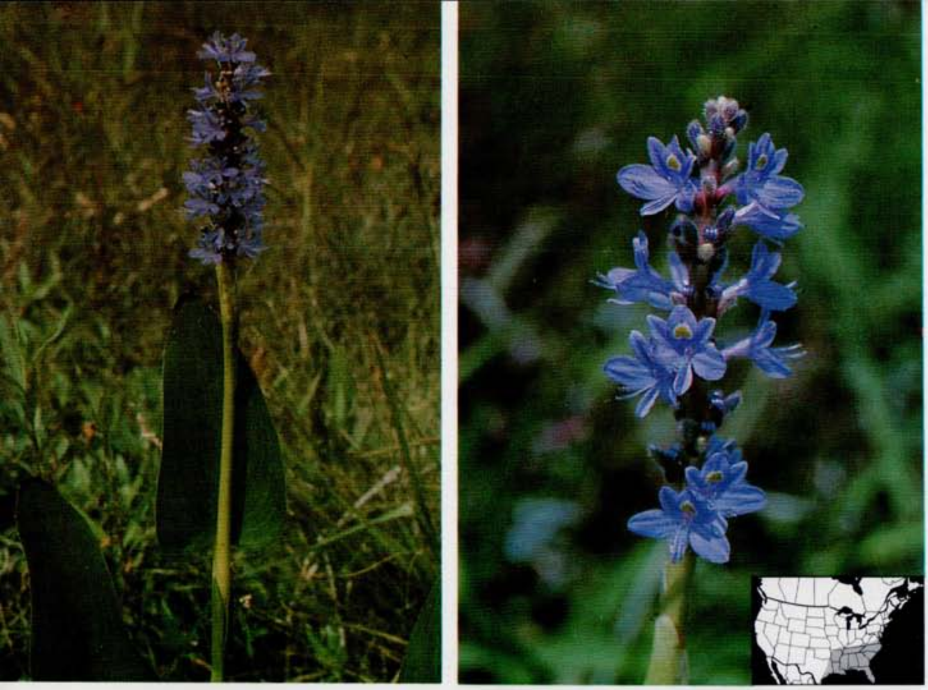

Lance-leaved pickerelweed: plant and flowering stalk (NYBG).

Sweet flag: leaves (Dykeman). Flowering spike (Elias). Rootstock (Dykeman).

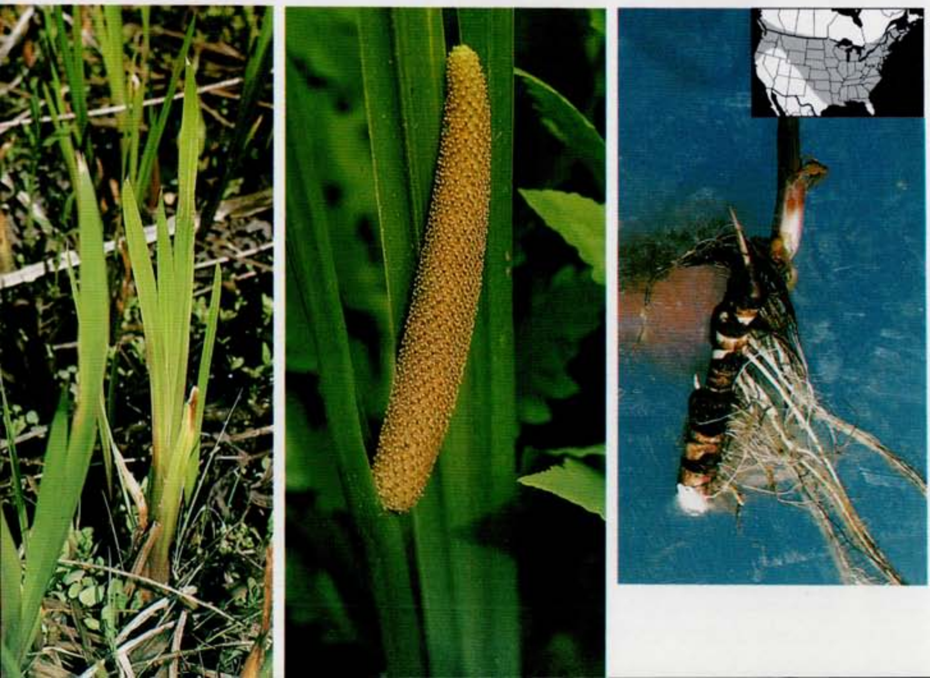

Lance-leaved pickerelweed *Pontederia lanceolata*

Also edible autumn

Habitat: shallow water along shores of lakes, ponds, streams, marshes. **Identification:** perennial herb to 1 m (3.2 ft) tall with thick, creeping underground stem (rhizome). ***Leaves mainly basal,*** blade 6–20 cm (2.4–8 in) long, 2–12 cm (0.8–4.7 in) wide, variable in shape but usually narrowly heart-shaped to arrowhead-shaped, smooth; ***leafstalk stout, 8–30 cm*** (3.2–12 in) ***long.*** Flowering stalk as tall or slightly taller than leaves, usually with a leaflike bract about midway; flowers densely clustered on narrow spike 6–18 cm (2.4–7 in) long, ***each flower blue to purple,*** funnel-shaped, 2-lipped, ***covered with short glandular hairs. Fruits*** stiff, ***egg-shaped capsules 4–8 mm*** (0.2–0.3 in) ***long, usually as broad or broader than long.*** **Harvest and preparation:** See PICKERELWEED (preceding). **Related edible species:** PICKERELWEED. **Poisonous look-alikes:** none.

Sweet flag, Calamus *Acorus calamus*

Habitat: ditches, swamps, pond edges, lakes, and streams; may grow with BLUE FLAG, water flag, and CATTAILS. **Identification:** perennial herb with large, stout, aromatic underground stem (rhizome). ***Leaves*** crowded near base, erect, narrow, ***linear, swordlike but flexible, usually 1–1.5 m*** (3–5 ft) ***long,*** rarely to 2 m (6.6 ft), 1–3 cm (0.4–1.2 in) wide, ***midvein off-center, greenish-yellow, with aromatic, almost spicy, odor when cut.*** Underground stem portions red. ***Flowers*** tiny, bisexual, ***numerous on narrow, tapering spike (spadix) produced about ⅓ of way up leaflike stalk;*** spikes often overlooked. Fruits inconspicuous, borne on flowering spike. **Harvest and preparation:** Harvest tender young stalks when less than 30 cm (1 ft) tall with short underground stems. Grasp and pull firmly. Vertical stem usually breaks off at rhizome. To obtain underground stem, dig with trowel to sever anchoring roots. Inner portions of tender young stems provide spicy trail snack or salad ingredient. For candy, use especially red underground part of stem. Remove roots, wash, peel, slice into 1 cm (0.4 in) long pieces. Cook in several changes of boiling water until pungency is tolerable. Simmer 20 min in syrup of 1 cup water to 2 cups sugar. Drain, dry, roll in sugar. **Related edible species:** none. **Poisonous look-alikes:** Irises (YELLOW FLAG, BLUE FLAG) cause severe digestive upset. Iris leaves lack spicy odor when cut. Flowers are showy.

Spanish bayonet, Datil *(Also edible spring)* *Yucca baccata*

Habitat: dry slopes, usually in sandy or rocky soils from 920–2,440 m (3,000–8,000 ft). **Identification:** ***long-lived perennial*** with short inconspicuous trunk, sometimes in clumps. ***Leaves clustered near base, blade lance-shaped, 50–76 cm*** (20–30 in) ***long, 3–7 cm*** (1.2–2.8 in) ***wide,*** stiff, tip sharp-pointed and ending in stout spine, with coarse, recurved fibers along margin, flattened or concave, grayish-green. Flowering stalk in spring, upright 1–1.5 m (3.2–5 ft) high, densely flowered along upper half, flowers showy, fleshy, sepals 3, reddish-brown on outer surface, petals 3, cream-white, 6–12 cm (2.4–4.7 in) long. ***Fruits are fleshy capsules, 14–18 cm*** (5.5–7 in) ***long,*** broadest near the middle, 3-parted, turning reddish brown at maturity and splitting open to release seeds. **Harvest:** young flower stalks before buds expand and fruits as they are available late spring through summer. Use buds and flowers as described for SOAPWEED (next), but not as good. **Preparation:** Cut young flower stalks into sections, boil 25–30 min or roast until core is tender. Peel off tough rind. Serve with butter, lemon juice, and seasonings. Eat fruit raw or cooked. Roast whole fruits until tender, scrape out pulp and separate from seeds. Sweeten if desired. Use raw pulp for pie. Strain pulp and dry by boiling to paste and finishing in oven as thick sheet. Eat dry or dissolve for beverage. Roast seeds at 190°C (375°F) until dry and crisp. Grind coarsely and boil as vegetable until tender. **Related edible species and poisonous look-alikes:** See SOAPWEED.

Soapweed *(Also edible spring)* *Yucca glauca*

Habitat: prairies, roadsides, pastures—usually in dry, well-drained sandy or rocky soils. **Identification:** perennial with short stem. ***Leaves densely clustered near base*** and covering stem, ***blade linear, 38-60 cm*** (15-24 in) ***long, 0.8-1.2 cm*** (0.3-0.5 in) ***wide,*** stiff, hard and sharp-pointed at tip, with narrow white band and occasional loose threadlike strands along margin, concave, gray-green. Flowering stalk in spring, upright to 1 m (3.3 ft), with stalked flowers along upper half, flowers bowl-shaped, showy, with 3 greenish-white sepals, 3 white petals 4.5–5.5 cm (1.8–2.2 in) long. ***Fruits are cylinder-shaped dry capsules 5–8 cm*** (2–3.2 in) ***long,*** 3–4.5 cm (1.2–1.8 in) in diameter, turning brown, 3-parted and splitting open at maturity to release seeds. **Harvest:** Pick young flower stalks before buds expand. Pick flowers and buds as available. **Preparation:** Cover 1 cup flowers or buds with water and boil until tender, 15–20 min. Drain and chop coarsely. Add with pimento or green pepper to several eggs and little milk. Cook like scrambled eggs. For salad boil 1 cup flowers and buds as above. Drain, chop, and cool. Add celery, chopped apple, raisins, nutmeats, cabbage, and mayonnaise as in Waldorf salad. For flower stalks see SPANISH BAYONET (preceding). Unripe fruits are bitter but can be made palatable by boiling or baking. **Related edible species:** all *Yucca* species including SPANISH BAYONET. Flower and bud uses apply best to dry-fruited species, fruit uses to fleshy-fruited species. Spiny leaved *Agave* looks similar and is also edible. **Poisonous look-alikes:** none.

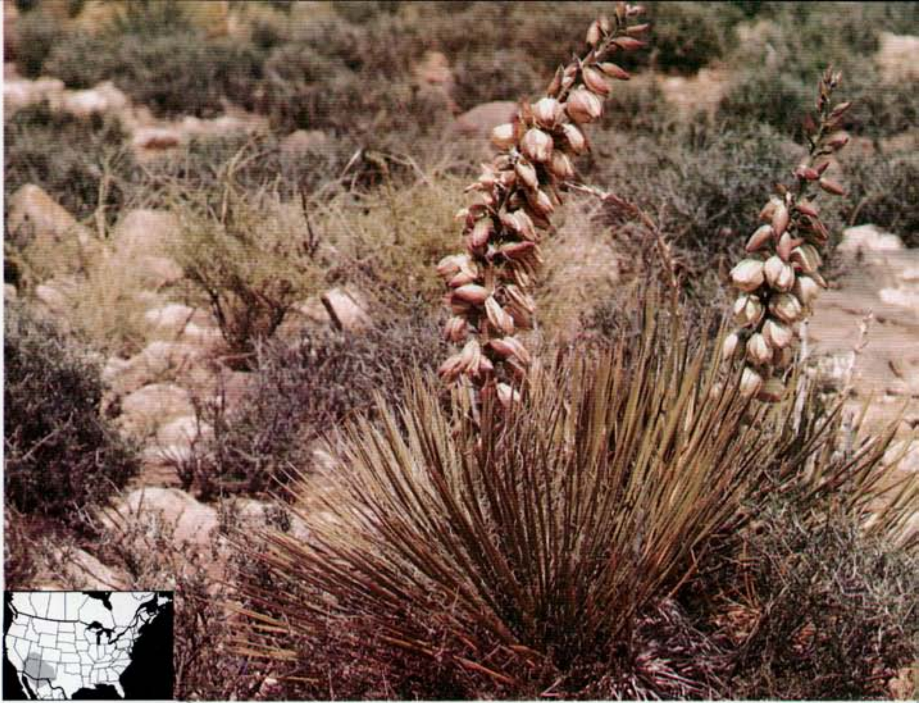

Spanish bayonet: plants (NYBG).

Soapweed: flowering stalk (NYBG). Plants (NYBG).

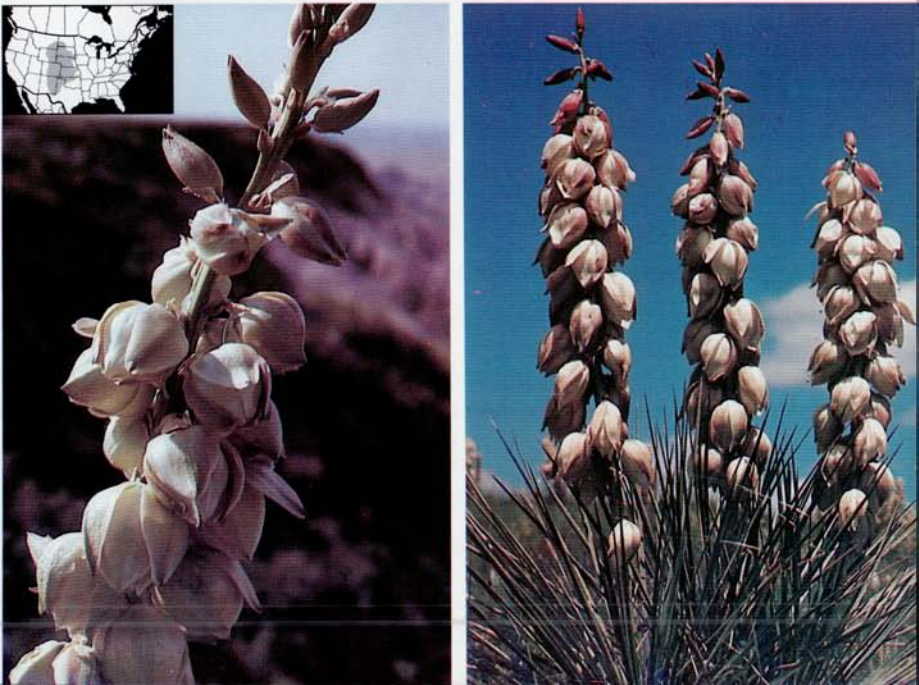

Common day lily *Hemerocallis fulva*

Also edible autumn, spring

Habitat: garden plant that commonly grows wild in abandoned home sites, vacant lots, along roads. **Identification:** large perennial herb from a fibrous root system with short, fingerlike tubers, with numerous ***long, narrow basal leaves.*** Each leaf 0.5–1.5 m (1.7–5 ft) long, 1.5–3 cm (0.6–1.2 in) wide, gradually tapering to long narrow tip, tapering with edges folded inward at base, light green to greenish-yellow. Flowers in clusters of 3–5 at tip of upright stalks, stalks to 2 m (6.6 ft); flower buds narrow, tubular; each ***flower 10–15 cm*** (4–6 in) ***long,*** narrowly tubular at base, ***funnel-shaped*** above, ***with 6 large, showy petallike segments, orange to tawny-orange,*** darker inside, flowering from late May to July, each flower open 1 day only. Fruits fleshy, 3-parted capsules. **Harvest:** tubers 1 cm (0.4 in) × 2–3 cm (0.8–1.2 in). Spade roots over; pick only firm tubers all year. Replant roots. Tubers regenerate rapidly. Pick buds when half to nearly full size. Pick flowers when fully opened or wilting. **Preparation:** Raw tubers give sweet nutty flavor to salads, or boil in salted water until tender, season. Excellent potato substitute. Boil buds a few minutes, butter, and season. Great cooked vegetable; raw they irritate the throat. Dip buds or flowers in egg batter and fry in hot oil or add them to soups at end of preparation for gelatinous quality. Dry buds and flowers at air temperature about a week; store in sealed jars for later use in soups or other recipes. *Caution: Eat in moderation to avoid laxative effect.* **Related edible species:** none. **Poisonous look-alikes:** Leaves slightly resemble those of iris (YELLOW FLAG, BLUE FLAG), but common day lily leaves arch over with tip pointed to ground; iris leaves are upright and darker.

Western camass, Camass lily *Camassia quamash*

Also edible autumn, winter, spring

Habitat: coastal mountain forests, wet meadows of interior mountains. **Identification:** ***perennial herb*** to 80 cm (32 in) high, from bulb 2–3 cm (0.8–1.2 in) across; ***grasslike leaves*** several from base, narrowly lance-shaped, 20–50 cm (8–20 in) long, 0.6–1.8 cm (0.2–0.7 in) wide, pointed at tip. Flower stalk 20–80 cm (8–32 in) high, ***flowers several, stalked, showy, with 6 blue to bluish-violet petals*** or petallike structures, each 2–3.6 cm (0.8–1.4 in) long. Fruits are dry capsules 0.8–2.4 cm (0.3–1 in) long, egg-shaped, splitting open at maturity to release seeds. **Harvest:** Dig or pull bulbs any season, safest when blue flowers distinguish it from DEATH CAMASS. Dig and transplant to safe place where you know there is no DEATH CAMASS. *Be certain of identification!* **Preparation:** Raw bulbs are palatable potato substitute. May be boiled 25–30 min or baked in foil for 45 min at 175°C (350°F). Indians baked bulbs in pit up to 3 days, the resulting product being very dark and sugary. Slice pit-baked or boiled bulbs, dry in sun or warm oven, store in paper bag in dry place. **Related edible species:** EASTERN CAMASS. **Poisonous look-alikes:** DEATH CAMASS (*Zigadenus* species), narrow-leaved plants with green, white, or bronze flowers, single or several in narrow stalk.

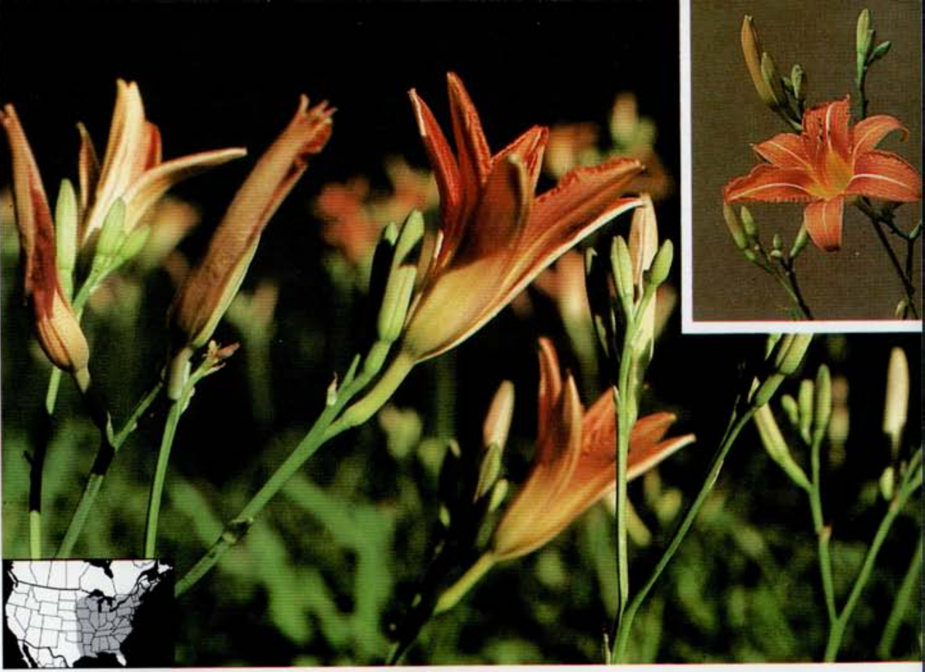

Common day lily: flowers (Dykeman).

Western camass: field of camass (Ross). Flowering stalks (NYBG). Flowers (Ross).

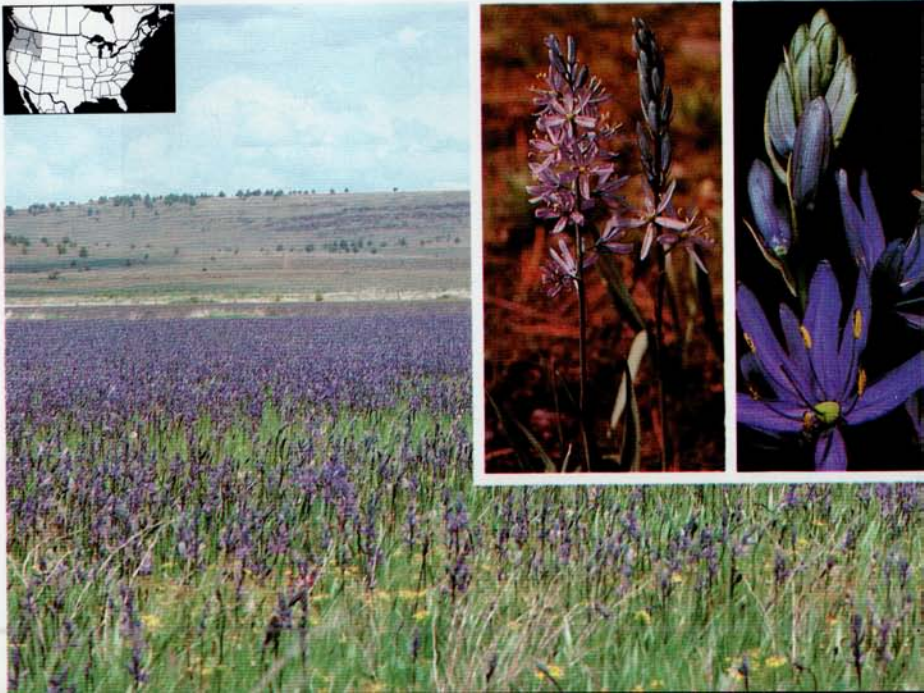

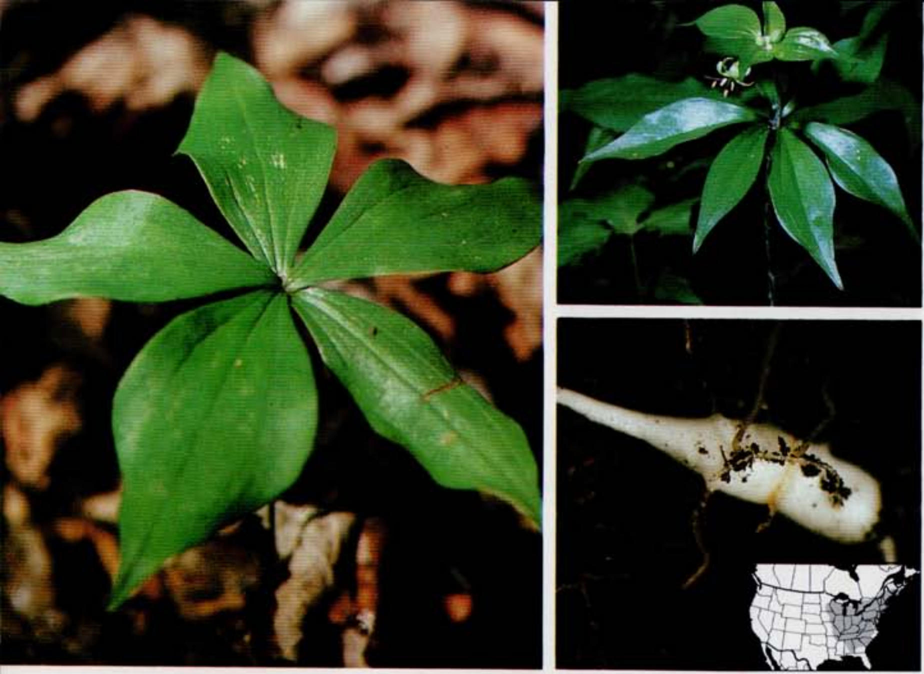

Indian cucumber: leaves and root (Dykeman). Flowering plant (NYBG).

Indian fig: plant (Folkerts).

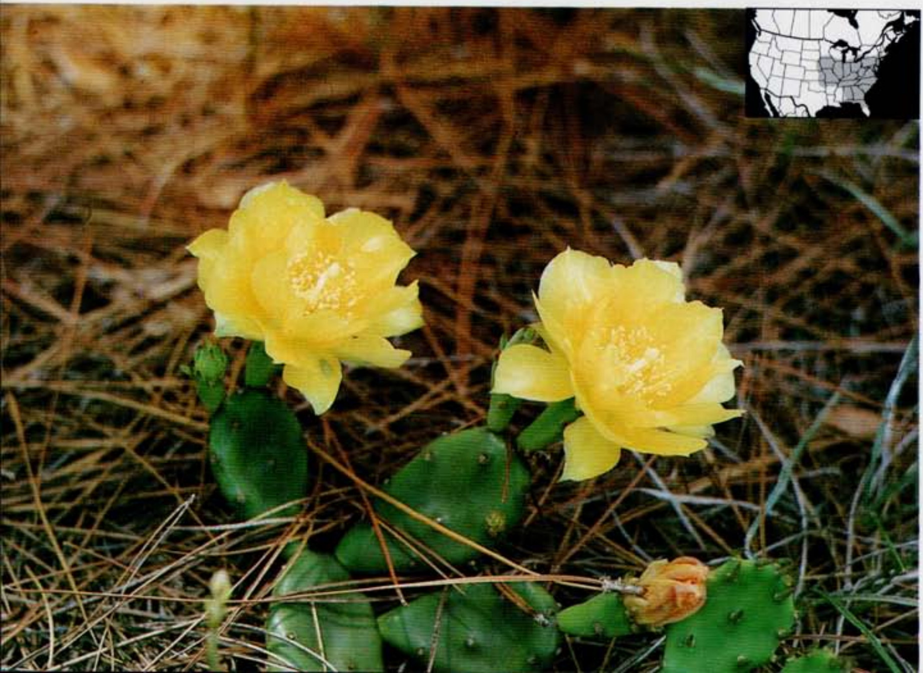

Indian cucumber *Medeola virginiana*

Habitat: rich moist woodlands, protected slopes, margin of bogs and swamps. **Identification:** ***perennial herb, with*** swollen ***tuberlike white underground stems (rhizomes)*** horizontal to ground; stem erect, to 80 cm (2.8 ft), usually about 30 cm (1 ft), slender, covered with wooly hairs when young, ***bearing a whorl of 5–10 leaves about midway and another whorl of 3 leaves near top.*** Leaves simple, parallel-veined, to 10 cm (4 in) long, and uniformly wide to broadest near tip; upper leaves smaller and broadest near base. ***Stalked flowers 3–9 in cluster at tip of stem, some flowers hanging beneath upper leaves,*** in May and June, each flower with 6 pale greenish-yellow petals 6–10 mm (0.2–0.4 in) long. ***Fruits*** small, rounded, dark purple berries 6–8 mm (0.2–0.3 in) in diameter, ***not edible.*** **Harvest:** common, easily identified plant a reliable source of limited quantity of tubers for trail food. Digging destroys plant, but any remaining tubers will produce new growth. **Preparation:** Wash and eat tubers raw as trail snack or in salad. Has delicate cucumber flavor. For pickles combine 1½ cups tubers, ½ cup water, ½ cup vinegar, 1 tbsp salt, 2 tsp sugar, 1 clove garlic, ¼ small, chopped red pepper, 1 tsp pickling spice. Bring to boil, place in small hot sterilized jars, covering tubers with liquid, and seal. Store at least 2 months. **Related edible species:** none. Unrelated, inedible starflower *(Trientalis borealis)* looks similar except for net-veined leaves. **Poisonous look-alikes:** none.

Indian fig, Prickly pear *(Also edible spring, autumn)* *Opuntia humifusa*

Habitat: rocky bluffs, sand dunes, dry rocky or sandy grasslands. **Identification:** perennial herb with fibrous root system, sprawling along ground, often forming clumps. ***Fleshy segments or joints nearly circular*** to broadest above middle, sometimes uniformly wide, ***5–20 cm*** (2–8 in) ***long, strongly flattened,*** pale to dark green, dotted with small reddish-brown bumps containing numerous tiny, barbed hairs; spines absent or occasionally 1. ***Flowers*** produced June–July on upper part of segments, ***showy, yellow with red center,*** 5–9 cm (2–3.5 in) across, usually with 10–12 petals. ***Fruit*** broadest above middle to nearly club-shaped, ***3–6 cm*** (1.2–2.4 in) ***long, red to purple, becoming fleshy,*** containing many flattened, nearly circular seeds. **Harvest:** Wear leather gloves to ward off spines and barbed hairs. Cut young segments during spring and summer. Pick fruits as they ripen in late summer and autumn. **Preparation:** Flame segments to remove spines, roast segments over fire, then peel or peel first, then slice and use like string beans. Edible but somewhat sticky. Also, deep fry like onion rings. Dry pulp for later use. To eat fruits fresh, chill if possible, trim off both ends, cut skin lengthwise, scoop out sweet pulp. Or, dip 1–2 min in boiling water and peel. Be careful, fruits also have fine, spiny bristles. Dry fruit pulp for later use, stew it for dessert, or boil it down and strain out seeds to make syrup or to prepare jelly. Briefly roast seeds and grind into meal for soup thickener or flour. For emergency water, peel and chew leafy segments. **Related edible species:** Fleshy fruits and segments of all flat-stemmed *Opuntia* species are edible. **Poisonous look-alikes:** none.

Prickly pear, Indian fig *Opuntia phaeacantha*

Also edible autumn, spring

Habitat: hills, canyons, valleys, usually in rocky or sandy soils. **Identification:** shrub to 1 m (3.2 ft), sometimes sprawling or forming clumps to 6 m (19.7 ft) in diameter. ***Fleshy cactus segments or joints rounded or nearly so,*** sometimes widest above middle, ***10–25 cm*** (4–10 in) ***long, flattened, bluish-green; cacti spines usually on upper half, 1–6 per cluster, 2.5–7.5 cm*** (1–3 in) ***long, straight or curved. Flowers*** produced near top of upper segments, ***showy, yellow,*** 6–8 cm (2.4–3.2 in) long, to 8 cm (3.2 in) in diameter, with many petallike structures. ***Fruits elongate,*** usually broadest above middle, ***3–6 cm (1.2–2.4 in) long, to 3.8 cm (1.5 in) in diameter, fleshy, red,*** depressed at tip, containing numerous, flattened and rounded seeds. The several varieties of this prickly pear differ primarily in segment shape, spine distribution, color, and length. **Harvest and preparation:** INDIAN FIG (preceding). **Related edible species:** all flat-stemmed prickly pear cacti. **Poisonous look-alikes:** none.

May apple, Mandrake *Podophyllum peltatum*

Habitat: rich deciduous woodlands, meadows, moist shaded road banks. **Identification:** ***perennial herb*** to 50 cm (20 in) tall, frequently in large dense patches with creeping underground stems and thick, fibrous roots, short stem, first-year plant with 1 leaf, second-year flowering plants with 2. ***Leaves*** large, umbrellalike, ***12–36 cm*** (4.7–14 in) ***in diameter, nearly circular in outline, deeply 5–9-parted,*** toothed; ***leafstalk attached at center of leaf. Flowers solitary, produced in junction of 2 leaves,*** stalked, nodding, each with 6 sepals, 6–9 waxy white petals 2.5–4 cm (1–1.6 in) long. Fruit an egg-shaped berry 2.5–5 cm (1–2 in) long, usually yellow, but red in one form. **Harvest:** Gather fruit when yellow; practically falls from plant into your hand. Leaves will be turning yellow by this time. Unripe fruit and other parts of plant contain poison podophyllin. The root has a long history of medicinal use. **Preparation:** Eat ripe fruit fresh; taste sensation varies from person to person. For preserves, jam, or pie filling, remove both ends, quarter fruit, and force through food mill to remove seeds before cooking. If skins are desired in final product, squeeze pulp from quartered fruit before removing seeds. Chop skins and add to pulp later. See recipes for jam and pie early in book. Also combine juice with lemonade, sumac ade, or other juices for new flavors. **Related edible species:** none. **Poisonous look-alikes:** none; do not attempt to use unripe May apples or any other part of the plant.

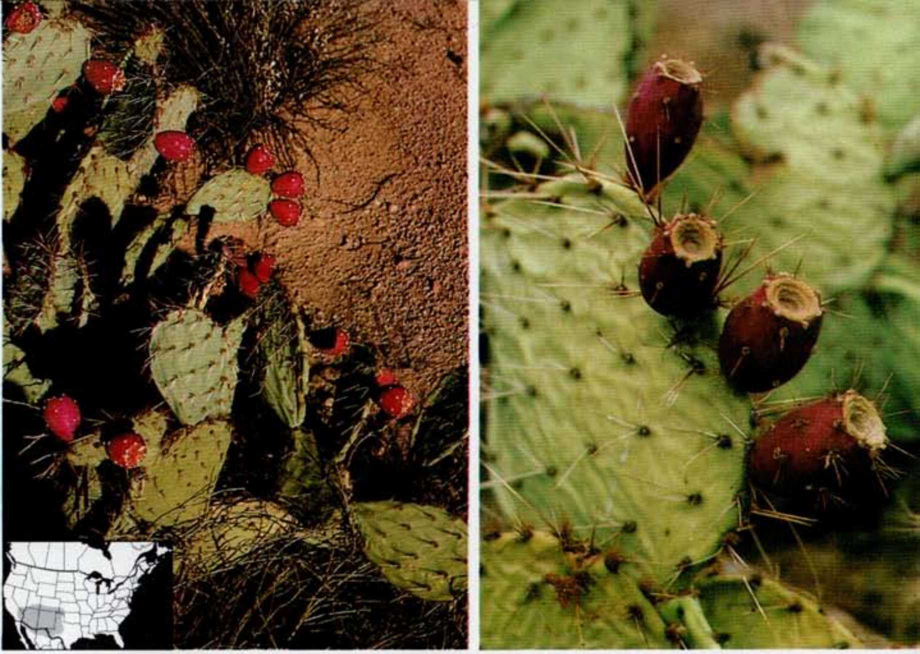

Prickly pear: plant (Elias). Fruits (Dykeman).

May apple: flowering plant (Dykeman). Fruit (NYBG).

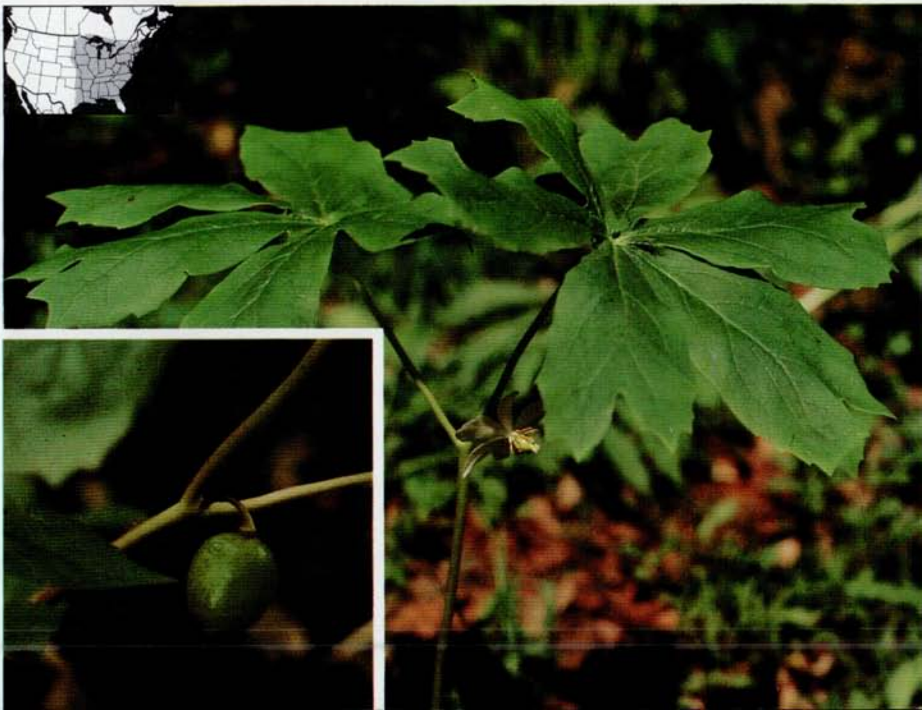

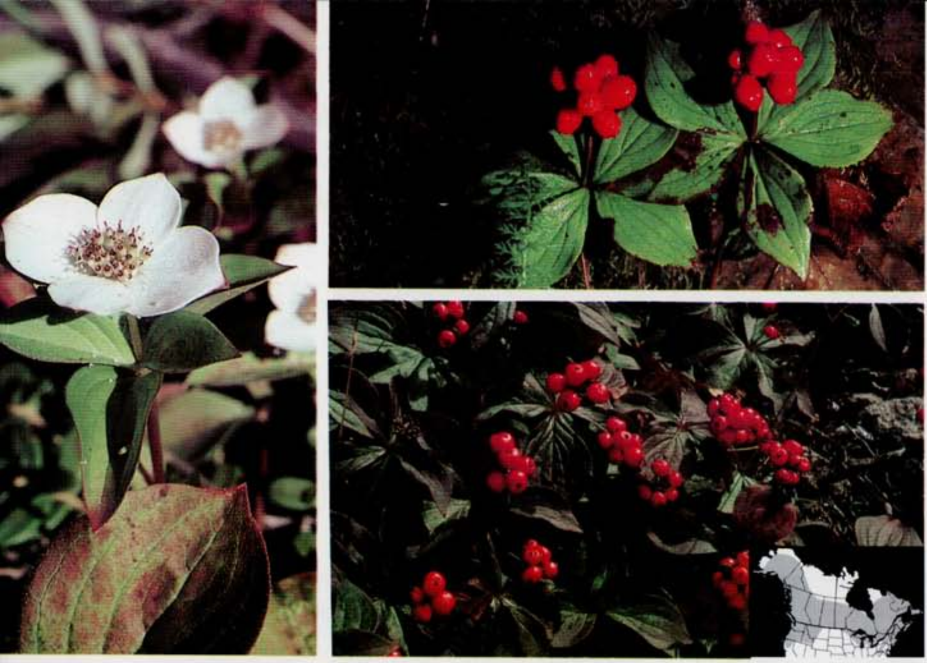

Bunchberry: flowers and fruiting plants (Elias).

Peppermint: flowering stalks (Speas).

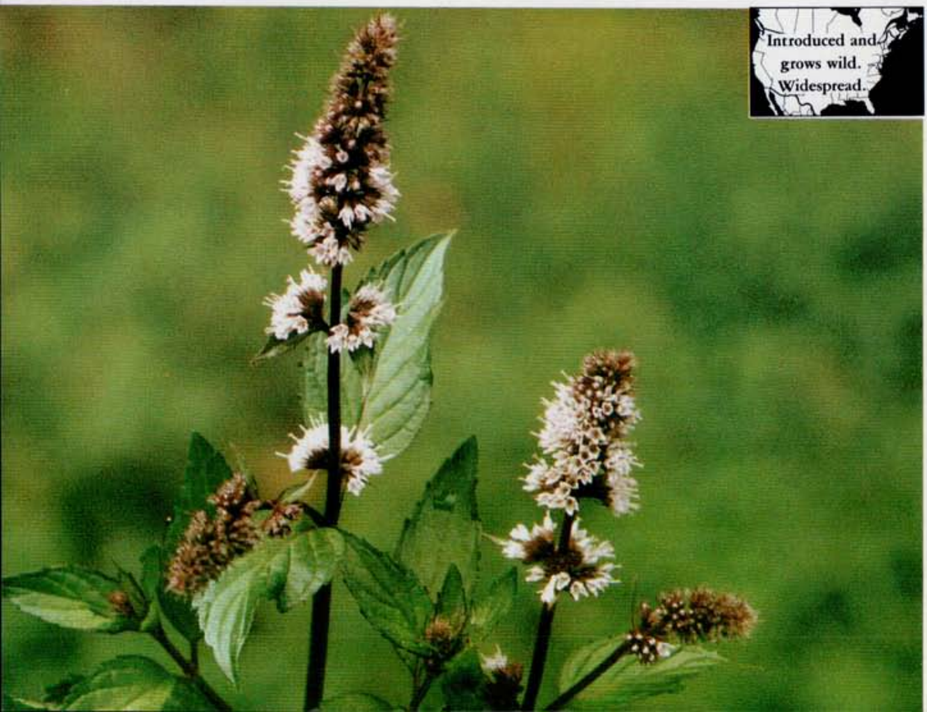
Introduced and
grows wild.
Widespread.

Bunchberry *Cornus canadensis*

Habitat: northern moist woods, bogs, woodlands, mountains especially in southern part of range. **Identification:** perennial, forming mats from spreading underground stems (rhizomes); aboveground stems 8–24 cm (3.2–9.4 in) high, ***each stem with whorl of 4–6 leaves near top*** and 1–2 pair smaller leaves below. Leaves 4–7 cm (1.6–2.8 in) long, 1.5–3.5 cm (0.6–1.4 in) wide, broadest near base or tip, pointed at tip, with 2–3 pair lateral veins. ***Flowers*** produced in May or June, ***tiny, in tight, dense clusters surrounded by 4 conspicuous, white, petallike bracts*** (flower clusters plus bracts resemble a single flower). Fruit produced in midsummer to autumn, in dense clusters, each a fleshy stone-fruit, globe-shaped, 6–8 mm (0.2–0.3 in) in diameter. **Harvest:** Berries ripen through July in clusters of a few to 12. Ripe when scarlet; usually plentiful and easily picked in quantity. **Preparation:** Eat nearly tasteless berries raw, but pulp is difficult to separate from seeds. Better cooked and strained for use with more flavorful fruits or for pudding. **Related edible species:** Swedish bunchberry *(C. suecica)* in northern regions has sweeter and more acid fruits. Laplanders make pudding from this species by boiling fruits with whey until thick. **Poisonous look-alikes:** none.

Peppermint *Mentha piperita*

Habitat: wet places, especially ditches, along streams, wet meadows. **Identification:** perennial herb, stems erect, 0.6–1.1 m (2–3.6 ft) tall, ***4-sided,*** purplish, smooth to slightly hairy. ***Leaves opposite,*** each pair alternating along stem, ***each with short leafstalk,*** blade 3.2–6.4 cm (1.3–2.5 in) long, 1–1.6 cm (0.4–0.6 in) wide, usually broadest near base, sharply toothed along margins, ***with distinct peppermint smell when broken or crushed.*** Flowers tiny, light purple, crowded together in dense, many-flowered, interrupted spikes at tips, June–October. Fruits with 4 tiny nutlets enclosed by persistent calyx. **Harvest and preparation:** Pick leaves at any stage. Great variation in flavor among populations from different areas. Collect leaves on dry day and dry on paper in warm area. Store in tight container. Jelly: Add 2 cups boiling water to 2 cups crushed fresh leaves in saucepan. Place inverted lid on saucepan, with small amount of ice and water in lid to condense water vapor and flavorful oils. Bring mint mix to simmer, remove from heat, steep 10 min. Strain; add 4 cups sugar and ¼ cup cider vinegar to 2 cups mint flavored water. Add green food coloring, if desired. Stir until sugar is dissolved. Dissolve 1 pkg powdered pectin in ¾ cup water, heat and boil hard 1 min. Add pectin solution to mint syrup, stir, pour into sterilized glasses, and seal. For other uses, see SPEARMINT (next). **Related edible species:** SPEARMINT, several hybrids of native and introduced species. **Poisonous look-alikes:** none.

Spearmint *Mentha spicata*

Habitat: grows in wild, especially in wet areas along roads, ditches, streams. **Identification:** perennial, spreading by creeping rootstock; ***stems*** erect to 1 m (3.2 ft), ***4-sided,*** smooth or nearly so. ***Leaves opposite,*** each pair alternating along stem, leafstalks absent or nearly so, blade 2.5–7.5 cm (1–3 in) long, 0.8–2.4 cm (0.3–1 in) wide, broadest at or below middle, toothed along margin, with ***strong smell of spearmint when crushed or broken.*** Flowers tiny, light purple to near white, produced in dense, elongate, many-flowered clusters 4–12 cm (1.6–4.7 in) long, at growing tips. Fruits tiny nutlets enclosed in persistent calyx. **Harvest and preparation:** Pick leaves at any stage. Great variation in flavor among plants from different areas. Collect leaves on dry day and dry on paper in warm area. Store in tight container. For tea pour 1 cup boiling water over ½ tsp crumbled, dried leaves or ¼–½ cup fresh leaves. Cover; steep 5 min. Do not boil. Strain, add sweetener and lemon to taste. High in vitamin C. Chop fresh leaves in green salad. Crush fresh leaves in iced tea. For candy wash and pat dry 2 doz fresh leaves. Beat 1 egg white until stiff. Dip leaves, one at a time in egg white, then in confectioner's sugar. Dry in oven at low heat, about 100°C (210°F), for 30 min or until crisp. Real mint-leaf candy! For jelly recipe see PEPPERMINT (preceding). **Related edible species:** PEPPERMINT, several other *Mentha* species, and many hybrids between native and cultivated species. **Poisonous look-alikes:** none.

American brooklime *Veronica americana*

Also edible spring

Habitat: common, along stream banks, swamps, clear ponds, springs, often with watercress. **Identification:** ***Perennial herb, usually creeping*** or sprawling; stems smooth, fleshy. ***Leaves opposite,*** lanceshaped to broadest near base, 3–8 cm (1.2–3.2 in) long, 1–4 cm (0.4–1.6 in) wide, pointed at tip, toothed along margin, stalked. ***Flowers stalked, produced along terminal, elongated spikes;*** each ***flower with 4 pale violet to lilac petals, 4–8 mm (0.2–0.3 in) across.*** Fruits are small, nearly round capsules, 3–5 mm (0.1–0.2 in) across, almost as long. **Harvest:** young growth tips from unpolluted water during late spring and summer. If water quality is questionable, wash tips in clean water to which iodide purifying tablets sold in camping stores have been added. Rinse with potable water. **Preparation:** Use like WATERCRESS in salads or as potherb with bland greens to offset bitterness of brooklime. **Related edible species:** several other species of speedwells (*Veronica* spp.), some widely distributed. **Poisonous look-alikes:** none.

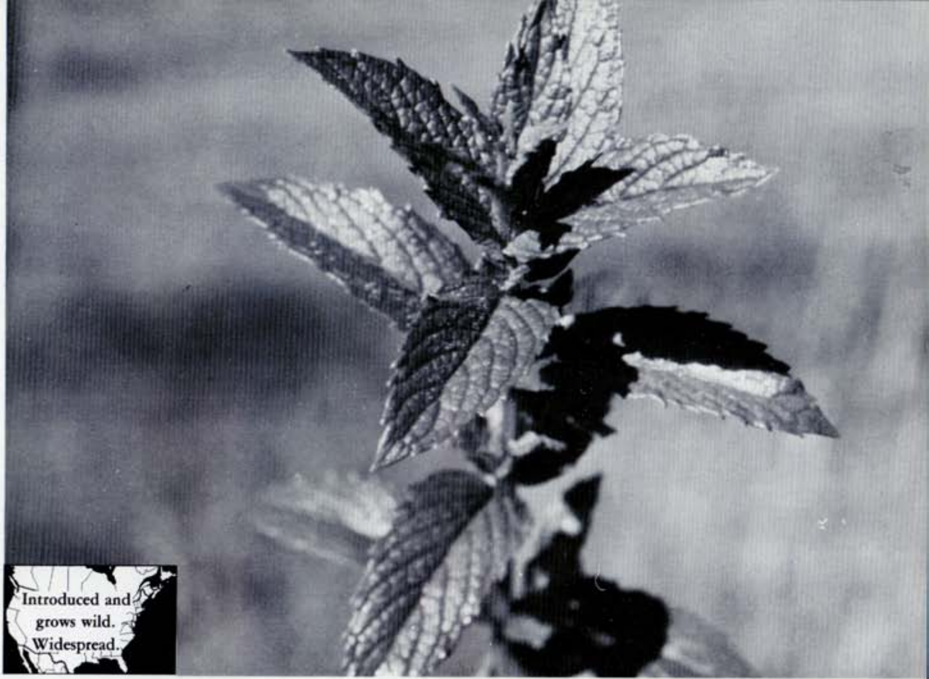
Introduced and grows wild. Widespread.

Spearmint: (Elias photo).

American brooklime: flowering stalk (Dykeman).

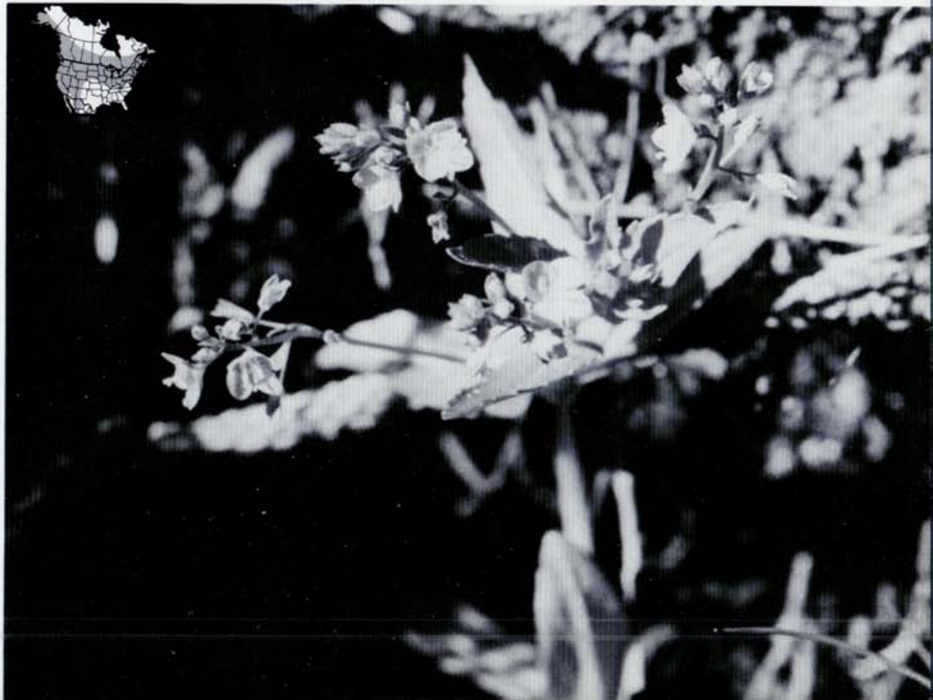

Marsh mallow *(Also edible autumn, winter)* *Althaea officinalis*

Habitat: edges of brackish and salt marshes; introduced from Europe and now grows wild. **Identification:** ***perennial herb*** from 0.6–1.2 m (2–4 ft) tall, from thick, large taproot; stems upright, often branched, stout, hairy. ***Leaves*** alternate, simple, broadest near base, 6–12 cm (2.4–4.7 in) long, about ⅔ as wide, ***often shallowly 3-lobed,*** short-pointed at tip, rounded at base, ***irregularly and coarsely toothed along margin,*** densely hairy. ***Flowers*** several ***in cluster at base of upper leaves in summer; each with 5 pink, spreading petals*** 2–3 cm (0.8–1.2 in) long. ***Fruit is dry, flattened disc, divided into 15–20 segments.*** **Harvest:** leaves in early summer, flower buds in summer, and roots from late summer through winter. **Preparation:** whole plant contains mucilagelike material; roots are best source. Use young leaves in early summer as okra-like soup thickener or as potherb. Pickle flower buds. Boil thin sliced, peeled roots for 20 min in enough water to cover them. Strain off roots; for candy sweeten the liquid and boil until very thick. Beat and drop spoonfuls on waxed paper to cool. Roll pieces in confectioner's sugar. For vegetable, fry boiled root slices with butter and chopped onion until browned. Use water from boiling any parts of plant as substitute for egg white in meringue or chiffon pies. Also used for hand lotion and cough syrup. Use leaves for poultices for infected wounds. **Related edible species:** other Mallow family species, especially those of genus *Malva*. Eat *Malva* fruits raw or substitute roots for meringue. **Poisonous look-alikes:** none.

Watercress *Nasturtium officinale*
Also edible autumn, winter, spring (synonym: *Rorippa nasturtium-aquaticum*)

Habitat: slow moving clear water, especially springs, cool streams; introduced from Eurasia. **Identification:** ***perennial herb, growing in shallow water*** or mud; stems weak, partially floating in water or lying flat in mud with white roots growing from floating stems. ***Leaves alternate,*** 4–16 cm (1.6–6.3 in) long, 2–5 cm (0.8–2 in) wide, ***dissected into 3–11 lobed and usually rounded dark green leaflets;*** upper leaves nearly rounded. ***Flowers numerous in narrow, elongate clusters,*** each flower stalked, with ***4 white petals,*** 3–5 mm (0.2 in) long. Fruits narrow, slender, capsulelike, 1–2.5 cm (0.4–1 in) long, beaked, containing several tiny reddish-brown seeds. **Harvest:** Collect young growth nearly all year from unpolluted water sources. Removal does not hurt plant. If water quality is questionable and plant is to be eaten raw, disinfect shoots by washing in water with iodide purifying tablets sold in camping stores. Rinse in potable water. **Preparation:** As potherb, cook and serve just like spinach or add to more bland potherb greens, or stir-fry 4–5 min in skillet with 2 tbsp cooking oil and 1 tbsp fresh grated ginger root. Serve with soy sauce. Watercress is an excellent soup ingredient. Use radish-tasting stems and blossoms as garnish or as main salad ingredient. For nasturtium salad: Make slightly sweetened oil and vinegar dressing. Mix 1 cup young watercress leaves, 1 cup watercress buds and blossoms, 2 cups mixed mild greens, and 1 chopped scallion in large bowl. Add dressing, toss and serve. Vary ratio of watercress to other greens to get desired taste. High in vitamins and minerals. **Related edible species:** none. **Poisonous look-alikes:** none.

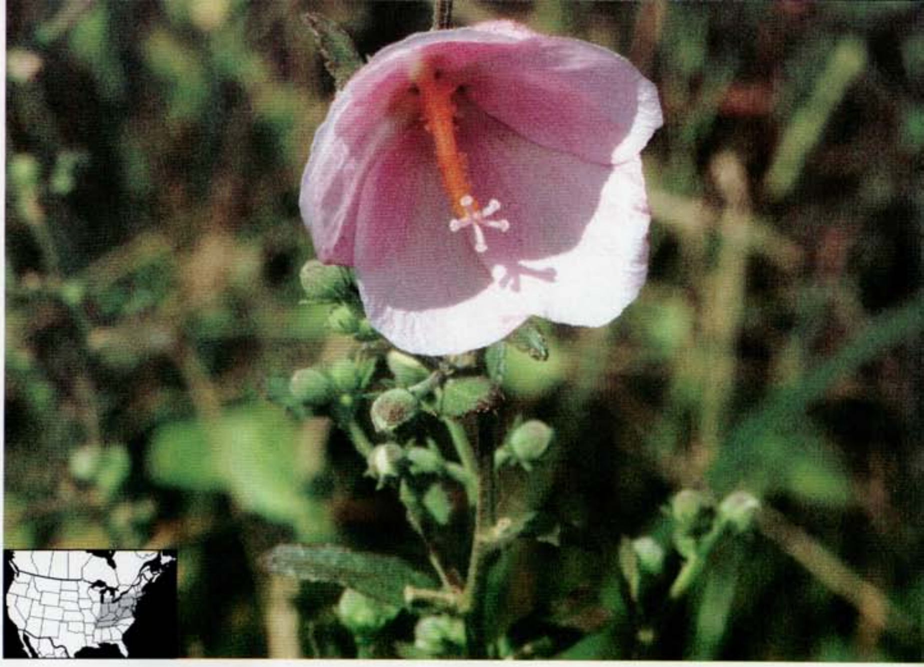

Marsh mallow: flowering stalk (Speas).

Watercress: flowering plant (Mohlenbrock).

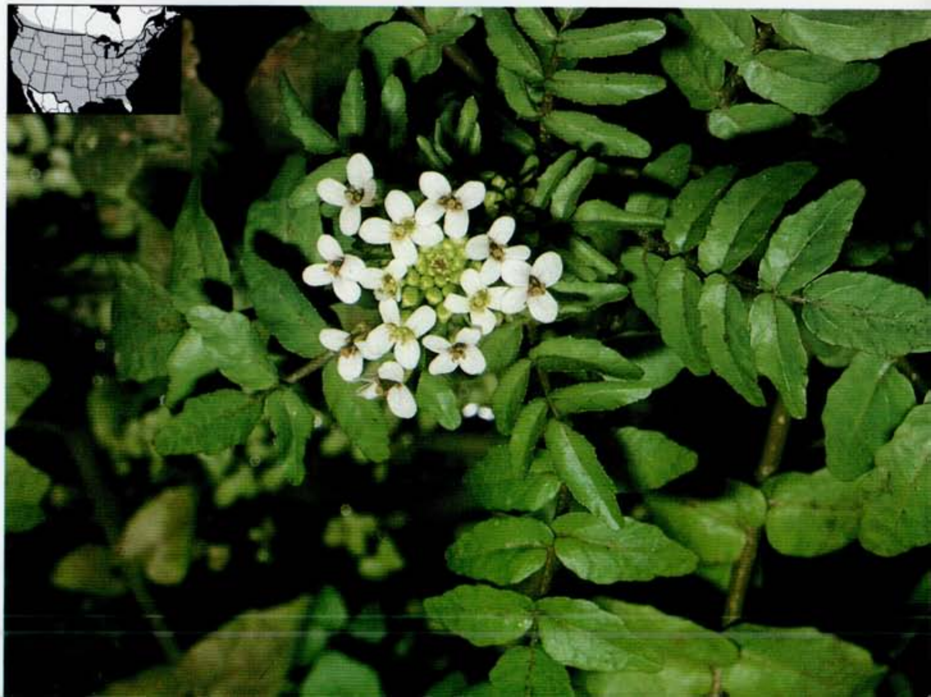

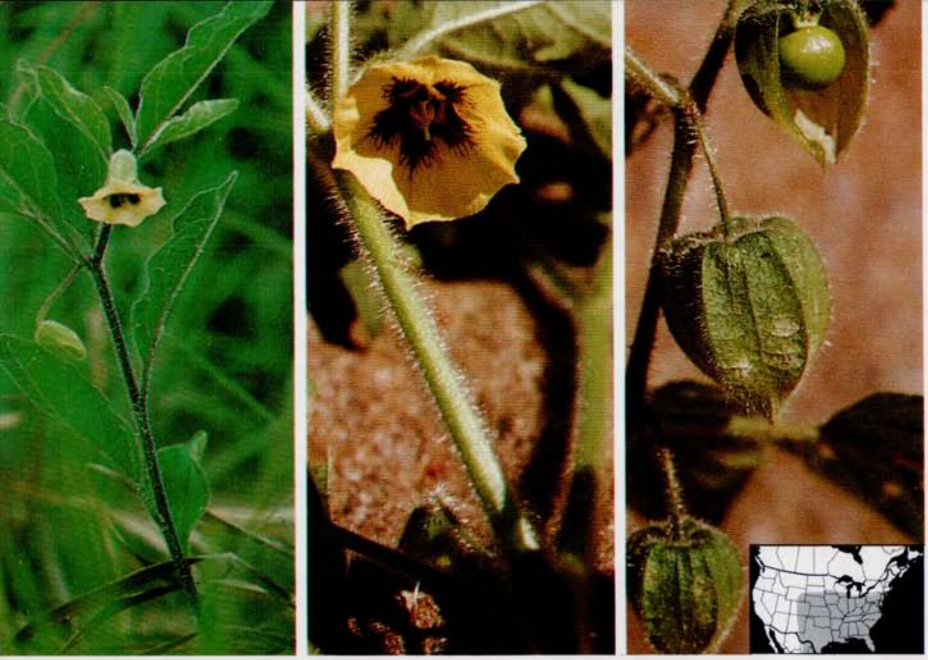

Ground cherry: flowering plant (NYBG). Flower and fruits (Pawloski).

Ground nut: flowering plant (NYBG). Underground tuber (Kavasch).

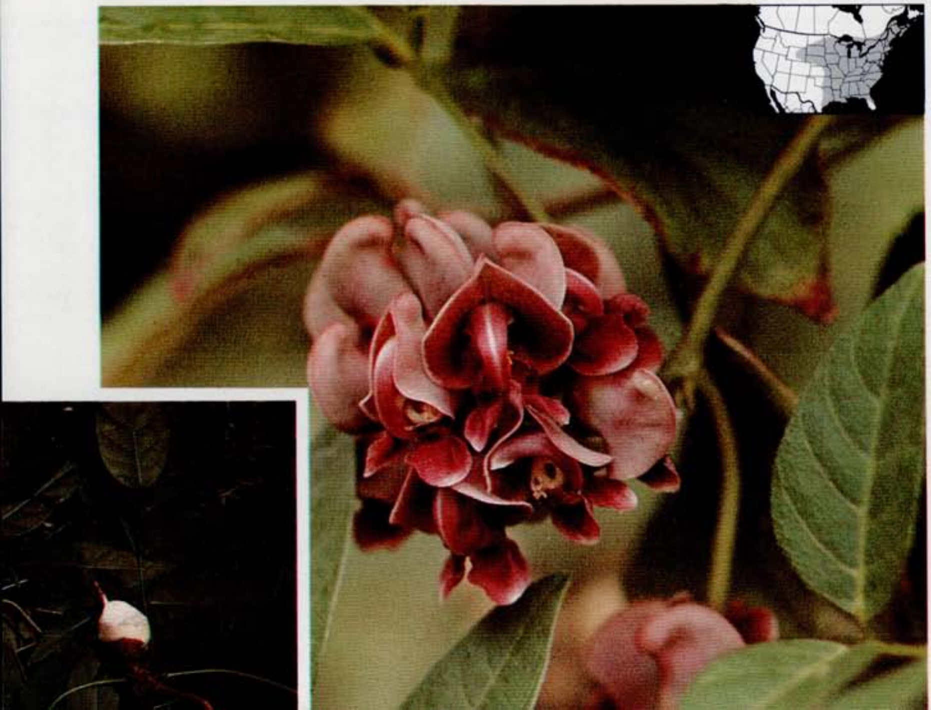

Ground cherry *Physalis pubescens*

Habitat: moist sites, open woodlands, recently cultivated fields, disturbed sites. **Identification:** ***annual herb*** to 60 cm (24 in) high; ***stems*** weak, wide branching, ***often trailing,*** hairy. ***Leaves*** alternate, simple, broadest near base, 3.5–8 cm (1.4–3.2 in) long, 1.7–6 cm (0.7–2.4 in) wide, pointed at tip, variably rounded at base, nearly entire, ***thin, pale green, hairy,*** long stalked. ***Flowers*** stalked, in junction of upper leaf stems, 5–10 cm (2–4 in) across, ***yellow with dark center. Fruit a sweet yellow berry*** 1.2–1.8 cm (0.5–0.7 in) in diameter, ***enclosed by 5-sided inflated persistent calyx resembling paper lantern 2–3 cm*** (0.8–1.2 in) long, ***angled,*** with conspicuous veins. **Harvest:** Fruits fall before fully ripe. So pick mid to late summer and ripen in husks a few weeks until fruits are yellow and sweet. Unripe fruit and green plant parts are strong flavored and somewhat poisonous. **Preparation:** excellent trail snack and dessert. For pie recipe, see book's Introduction. To preserve fruits boil until clear and tender in syrup of 1 cup sugar, 2 cups water, 3 tbsp lemon juice. Seal in sterilized jars. For jam boil for 5 min 1 qt crushed fruits, ¼ cup lemon juice, ½ cup water, and 1 pkg pectin. Add 3 cups sugar, boil again to reach jelly point, 115°C (240°F). Seal in sterilized jars. Use cherries in relishes. **Related edible species:** Several *Physalis* species are distributed throughout North America. Ripe fruits yellow, red, purple, or blue-black according to species; all are identifiable by lanternlike husk. Use only ripe fruit! **Poisonous look-alikes:** Learn to distinguish species of *Physalis* from poisonous nightshades *(Solanum)* (see JIMSON WEED and HORSE NETTLE).

Groundnut *(Also edible autumn, winter, spring)* *Apios americana*

Habitat: moist woodlands, bottomlands, thickets. **Identification:** ***perennial vine*** from 1 or more fleshy tubers arranged in a row, the tubers 1–2 cm (0.4–0.8 in) thick; stems smooth, twining around nearby plants; very white and easy to spot when mature, milky juice. ***Leaves*** alternate, ***featherlike compound*** (pinnate), 10–20 cm (4–8 in) long, ***composed of 5–9 leaflets,*** rarely less, leaflets usually 3–6 cm (1.2–2.4 in) long, lance-shaped, pointed at tip, rounded at base, smooth to hairy. ***Flowers*** July-September, densely clustered on short stalk 5–15 cm (2–6 in) long, each flower ***with 5 brownish-purple petals, resembling pea flower;*** fragrance resembles that of the violets. Fruits are dry, linear pods 5–12 cm (2–4.7 in) long, 4–7 mm (0.2–0.3 in) broad, containing 2 to several seeds, splitting open lengthwise into 2 spirally twisting valves. **Harvest:** Pick pods in midsummer to autumn as seeds mature; pods seldom abundant though. Dig tubers, starting at base of stem, working along root to obtain whole string. Good all year. **Preparation:** Roast pods at 190°C (375°F) for 20–25 min. Cool; remove seeds and brown in oil over low heat. Boil tubers in heavily salted water until tender. Season. Slice and fry leftovers, or grease and roast to regain tenderness, flavor. Also thinly slice raw tubers and fry like potatoes in butter or pork; season. Or bake at 175°C (350°F) 45–60 min until tender. Flavor turnip-like. **Related edible species:** groundnut *(A. priceana)* with 1 large tuber. **Poisonous look-alikes:** species of vetchlings *(Lathyrus),* rattlebox *(Crotalaria),* milk vetches *(Astragalus),* and point locoweeds *(Oxytropis).*

Common breadroot *Psoralea esculenta*

Habitat: prairies, plains, dry rocky woodlands. **Identification:** ***perennial herb*** to 38 cm (15 in) tall, ***developing from 1 or more swollen, elongated taproots*** 4–10 cm(1.6–4 in) long; stems short, stout, frequently branched, hairy. ***Leaves*** alternate, ***handlike (palmate) compound,*** long-stalked, ***composed of 5 leaflets,*** each 3–6 cm (1.2–2.4 in) long, uniformly wide to widest above middle, covered with long hairs. ***Flowers produced in dense elongate clusters 3–8 cm (1.2–3.2 in) long,*** each flower with leaflike bract, flowers ***blue,*** composed of 5 petals shaped like clover flower. Fruits are small pods 5–8 mm (0.2–0.3 in) long, broadest near base, long beaked, splitting open lengthwise along 2 valves. **Harvest:** Dig taproot from its hard prairie soil in early summer as stems and leaves begin to turn brown. Later above-ground parts dry and blow away making roots nearly impossible to locate. **Preparation:** important wild food staple. Peel and eat root raw, roast in campfire or boil until tender. Season like potatoes. For storage, peel and dry slowly. Grind dried root into flour or meal and use for baking. Indians cooked meal with meat and corn. Reconstitute with water to use as vegetable. **Related edible species:** *P. hypogala,* also of the plains region, produces a smaller but similar taproot. **Poisonous look-alikes:** Scurf pea *(P. tenuifolia)* and other *Psoralea* species have been toxic to horses and cattle, and so should not be eaten.

Wintergreen, Checkerberry *Gaultheria procumbens*

Also edible autumn, winter, spring

Habitat: woodlands, clearings, usually in poor acid soils. **Identification:** small herb with slender stems creeping along soil surface and erect leafy stems to 18 cm (7 in). ***Leaves*** crowded near top, each widest near or above middle, 2–5 cm (0.8–2 in) long, entire or with tiny rounded teeth along margin, ***leathery texture, shiny, with wintergreen smell when crushed. Flowers usually solitary on short drooping stalk, white,*** urn-shaped to cylinder-shaped, 5-parted at tip, 6–10 mm (0.2–0.4 in) long. ***Fruits bright red berries 6–10 mm (0.2–0.4 in) in diameter,*** globe-shaped to elongate. **Harvest:** leaves any time of year; berries from fall through winter. **Preparation:** Use leaves as trail nibble or brew into tea. Cover 1 qt young leaves with boiling water, cover container loosely and steep 1–2 days. Reheat gently, sweeten, and enjoy. Wintergreen oil concentration in tea increases with prolonged steeping. Berries ripen in fall but improve with freezing so look for them in winter and spring and use as trail nibble; add to pancakes and muffins, or, if plentiful, make uncooked jam. Combine in blender 2 cups fresh berries, 1 oz lemon juice, 1 cup strong wintergreen tea. Blend until smooth, pour into mixing bowl, add 4 cups sugar. In saucepan bring to hard boil 1 pkg pectin and ¾ cup water. Boil 1 min and stir into berry mix. Mix thoroughly, ladle into sterile jars, and seal. Store in refrigerator or freezer. Leaves and fruits contain aspirin like methyl salicylate and effectively reduce fevers and minor aches and pains. **Related edible species:** CREEPING SNOWBERRY (next). **Poisonous look-alikes:** none.

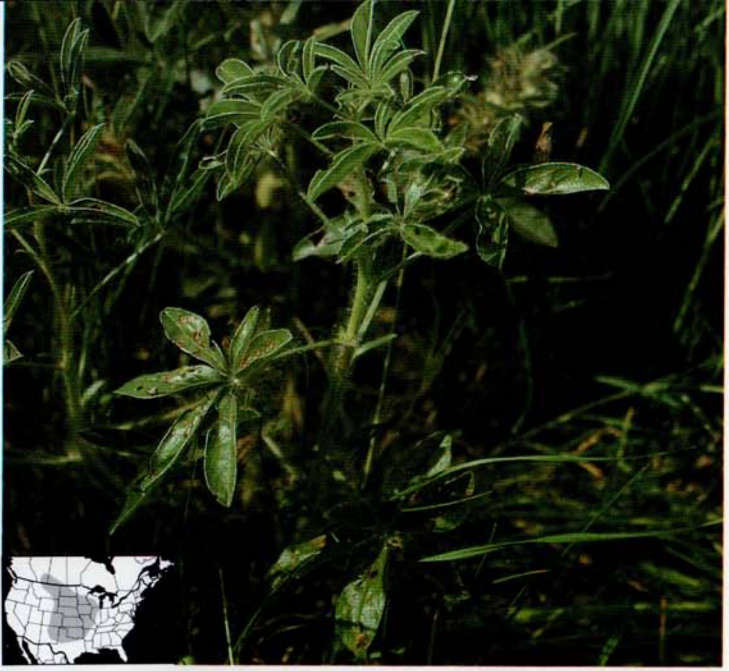

Common breadroot: July plant with remnant flower cluster on right (Young).

Wintergreen: fruiting plant, top (Dykeman). Flowering plant (Elias).

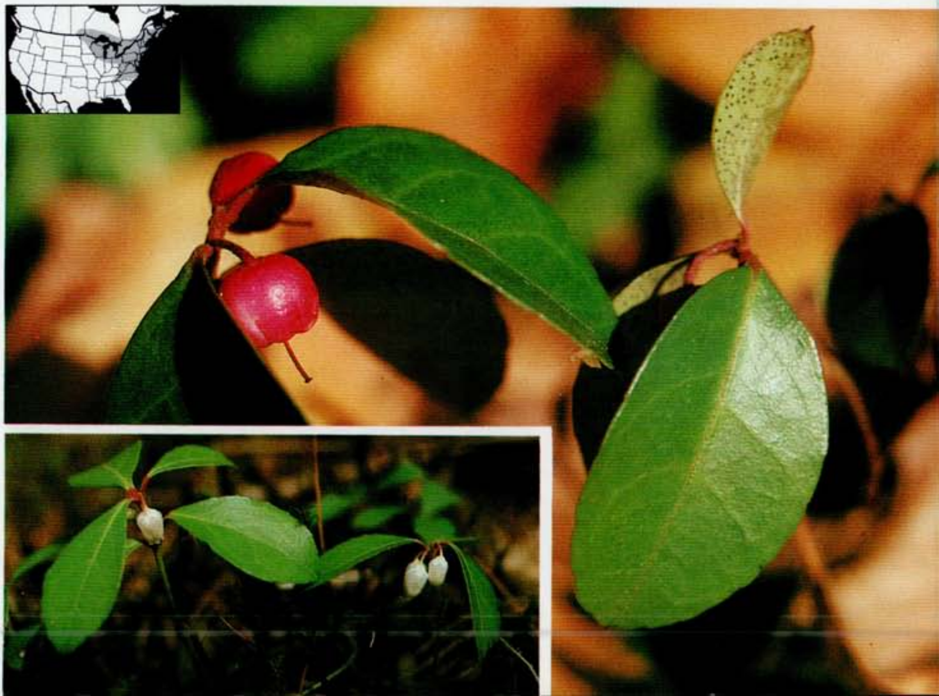

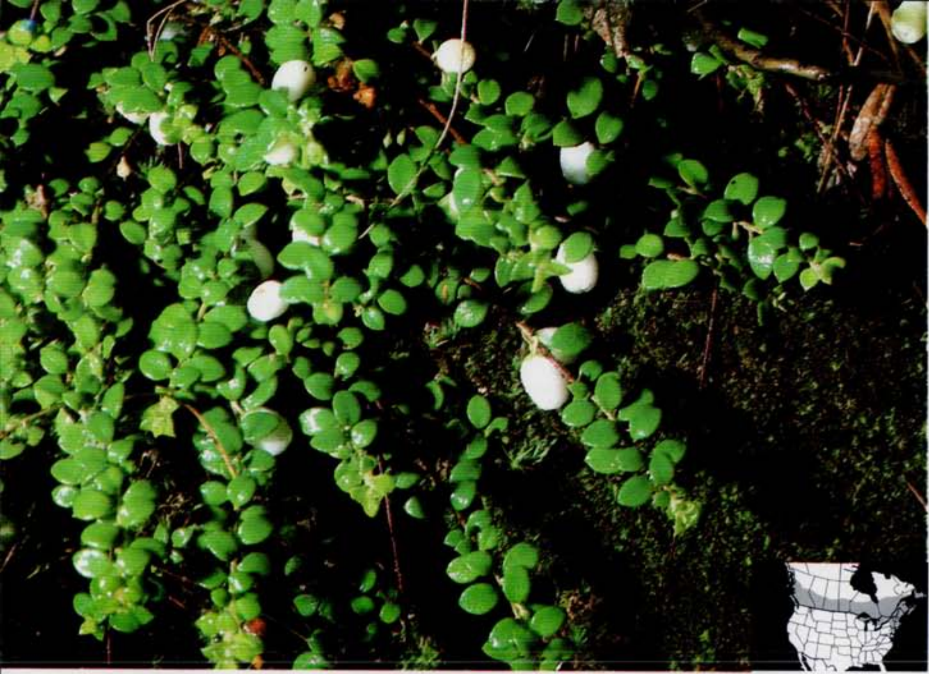

Creeping snowberry: fruiting plant (Elias).

Mountain sorrel: fruiting plant (Elias).

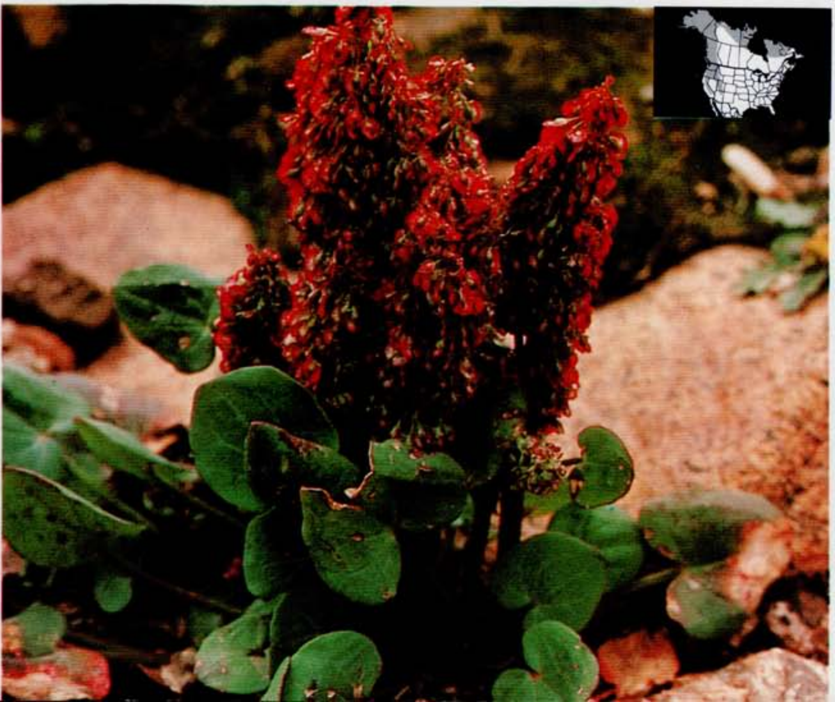

Creeping snowberry *Gaultheria hispidula*

Also edible autumn, winter, spring

Habitat: northern coniferous forests, mountains in subalpine and alpine zone. **Identification:** ***small creeping herb*** with very leafy stems; ***leaves*** alternate, simple, ***evergreen,*** broadest near base, to ***almost round, 0.5–1 cm (0.2–0.4 in) long,*** entire, bright green and smooth above, paler with long stiff hairs beneath. Flowers May-June, solitary in junction of leaves, each tiny, white, bell-shaped, 2–3 mm (0.1 in) long, on short drooping stalk. ***Fruits are fleshy bright white berries 6–10 mm (0.2–0.4 in) long,*** slightly longer than wide, with spicy aromatic smell when crushed, bristly. Ripen by August. **Harvest and preparation:** leaves same as WINTERGREEN (preceding). Also prepare as cooked green vegetable. Fruits ripen mid through late summer. Excellent fresh and in baked goods, jam (see WINTERGREEN), and preserves. For no-cook preserves wash and smash or grind fruit thoroughly. Add equal amount of sugar slowly with constant stirring. Pack in sterile jars and seal. **Related edible species:** WINTERGREEN, CREEPING WINTERGREEN, SALAL, and *G. ovatifolia,* also called salal, of northwest U. S. **Poisonous look-alikes:** none.

Mountain sorrel *Oxyria digyna*

Habitat: high mountains, subalpine and alpine fields, usually in rocky places 2,450–3,950 m (8,000–13,000 ft). **Identification:** perennial herb to 40 cm (16 in) high; stems upright, simple or sparingly branched, smooth. ***Leaves*** alternate, sometimes kidney shaped, simple, ***mostly basal, blades usually rounded,*** 1.6–3.2 cm (0.6–1.3 in) across, entire, smooth, long-stalked, stalk 6–12 cm (2.4–4.7 in) long. ***Flowers*** small, ***arranged in densely flowered, elongate, compact clusters;*** each flower stalked, bisexual, with 4 red to greenish sepals that persist in fruiting. ***Fruits*** flattened, ***broadly winged, rounded, dark red, small, dry, hard, 1-seeded,*** not splitting open at maturity. **Harvest:** leaves while still tender, before plant flowers. **Preparation:** Use fresh leaves as trail nibble or in salad. Combination of sour mountain sorrel and pungent species such as WATERCRESS works well with other, more neutral greens. Use a plain oil dressing. Add leaves to other greens for added flavor when cooked. In arctic regions, natives ferment mountain sorrel as sauerkraut. For puree, mash greens with food mill or blender and add to small amount of boiling water. Boil 5–10 min. Add 3 cups puree to 1 qt hot milk (or 1 cup milk and 3 cups chicken stock). Salt and pepper to taste. Bring to boil and simmer 5–10 min with stirring. Slowly add 3 beaten egg yolks. Add flour paste for desired thickness. Blend in 4 tbsp melted butter. For other uses see SHEEP SORREL. **Related edible species:** none, but SHEEP SORREL and SOUR DOCK are used similarly. **Poisonous look-alikes:** Some species of dock (RUMEX) accumulate sour-tasting oxalates.

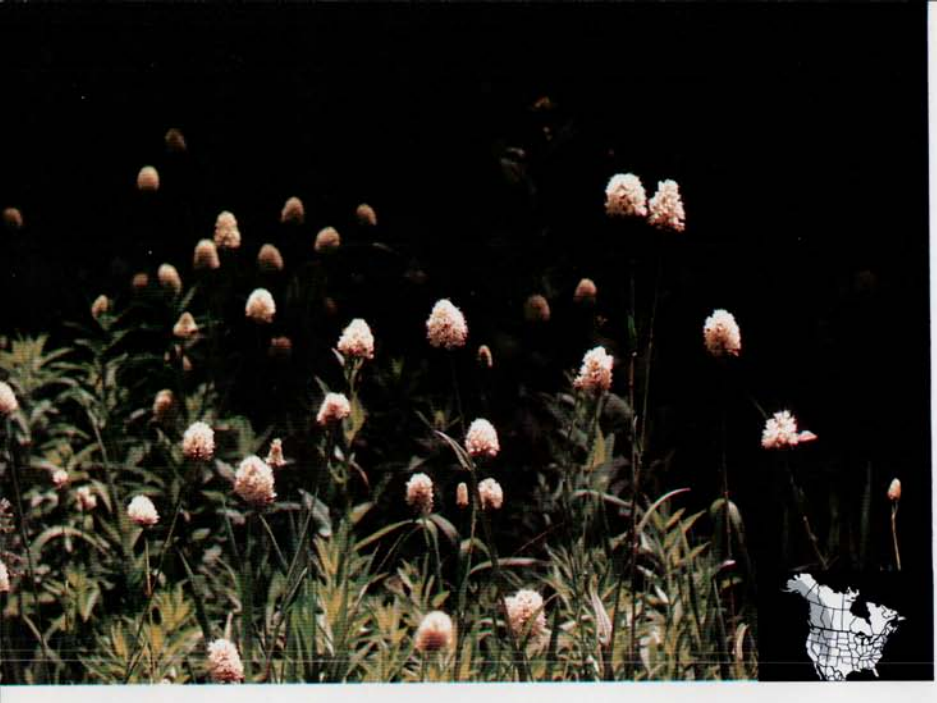

American bistort: flowering plants (NYBG).

Alpine bistort: flowering stalks (NYBG).

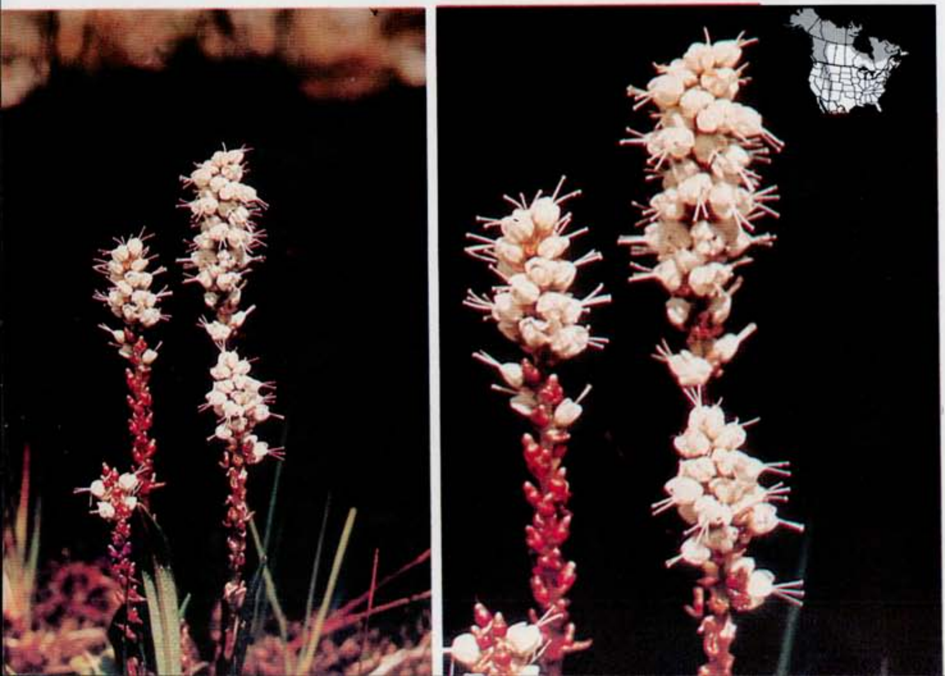

American bistort, Western bistort *Polygonum bistortoides*

Also edible autumn, spring

Habitat: along streams, wet meadows, subalpine meadows. **Identification:** ***perennial herb*** to 70 cm (28 in) high, from ***thick, fleshy, horizontal rootstock;*** stems erect, slender, smooth. Leaves alternate, entire, mostly basal, 10–25 cm (4–10 in) long, 2–5 cm (0.8–2 in) wide, lower leaves uniformly wide to broadest near tip, upper ones lance-shaped and smaller, stalked, each sheathing stem at base. ***Flowering spikes cylinder-shaped at end of stem, 2–6 cm*** (0.8–2.4 in) ***long,*** to 1.5 cm (0.6 in) thick, ***pink to white,*** individual flowers tiny, lacking petals. Fruits are small, shiny brown nutlets 3–4 mm (0.1–0.2 in) long. **Harvest and preparation:** Use rootstocks and young leaves as for ALPINE BISTORT (next). Rootstock larger in American bistort, but edible bulblets are absent from flower stalks. Both species provide valuable survival fare in mountain regions where other wild foods may be unavailable. **Related edible species and poisonous look-alikes:** See ALPINE BISTORT.

Alpine bistort, Alpine smartweed *Polygonum viviparum*

Also edible autumn, spring

Habitat: arctic regions, alpine and subalpine slopes, gravelly and rocky soils. **Identification:** ***perennial herb*** to 45 cm (18 in) high, ***from short, thickened, usually twisted, underground stem;*** stem upright, sheathed at base of leaves. ***Lower leaves*** alternate, simple, ***3–10 cm*** (1.2–4 in) ***long, to 2.5 cm*** (1 in) ***wide, narrowly lance-shaped,*** pointed at tip, tapering at base, long stalked, upper leaves much smaller. Flowering spike 3–8 cm (1.2–3.2 in) long, ***upper flowers pink or white,*** tiny, lacking petals, lower flowers producing tiny reddish or purple bulbils. Fruits are small, reddish-brown, 3-angled nutlets 2–4 mm (0.1–0.2 in) long. **Harvest and preparation:** Strip tiny bulbils from lower part of flower stalks and eat raw. Dig rootstocks (easiest to find during growing season) and eat raw, boil, or roast. Young roots tender but older ones may require cooking. Boil 40 min or bake at 165°C (330°F) for 40 min. Season like potatoes. Though appearance of unpeeled roots may not be inviting, flavor is good. Use young leaves in salad with other greens or prepare as cooked greens. Boil until tender and season with butter, salt and pepper. **Related edible species:** AMERICAN BISTORT (preceding) and LADY'S THUMB used as cooked or salad green. Also JAPANESE KNOTWEED, giant knotweed *(P. sachalinense),* Alaskan knotweed *(P. alaskanum),* and others. **Poisonous look-alikes:** No species of *Polygonum* are known to be poisonous. But some may cause light-sensitive skin reaction. They vary in usefulness as cooked or salad greens, some being very peppery or acid; some, like tartary buckwheat *(P. tataricum),* produce edible seeds.

Purslane, Pusley *Portulaca oleracea*

Habitat: fields, vacant lots, waste sites; common garden weed. **Identification:** annual ***herb, lying flat on ground*** or nearly so and ***forming mats to 1 sq ft, stems smooth, fleshy, shiny, often reddish purple. Leaves*** alternate and opposite, 2–5 cm (0.8–2 in) long, 1–2 cm (0.4–0.8 in) wide, ***fleshy, broadest near the rounded tip,*** entire along margin. ***Flowers*** single, or few clustered at tips of branches, flower 0.5–1 cm (0.2–0.4 in) wide, ***with 5 yellow petals. Fruits are capsules,*** 4–8 mm (0.2–0.3 in) long, broadest near base and rounded at tip, ***top portion falling away at maturity*** to expose dark red to black seeds. **Harvest:** Pinch or cut young leafy tips June–September. Pick larger stems in midsummer for pickles. To gather seeds, spread whole mature plants on sheet to dry for 2 weeks, then sieve and winnow. **Preparation:** One of best wild shoots cooked or in salad. Common and nutritious; acid taste; high in iron. For tossed salad combine raw shoots with other ingredients. Use oil and vinegar dressing. Cook young plants as potherb 10 min and season with butter and salt. For casserole chop cooked shoots, stir in 1 beaten egg and enough bread crumbs to make damp mix. Add salt and pepper, bake in moderate oven. Pickle stems same as cucumbers. Blanch and freeze tips for later use. To dry, boil thick stems in a little water for 20 min. Drain, dry in warm shady area, store in paper bag for later use in soups or stews as thickener. Dry seed 1 week in shade, store in paper bag. Grind and use half and half with flour. **Related edible species:** at least 2 *(P. neglecta* and *P. retusa)* which are very similar. **Poisonous look-alikes:** none.

Sunflower *(Also edible autumn)* *Helianthus annuus*

Habitat: prairies, plains, roadsides, disturbed sites. **Identification:** ***perennial herb 1–2.5 m (3.2–8.2 ft) high,*** from large taproot; stem upright, unbranched to branched, with stiff hairs. ***Leaves*** mostly alternate, simple, ***6.5–30 cm*** (2.6–12 in) ***long,*** nearly as wide at base, ***broadly to narrowly triangular,*** tapering to short- or long-pointed tip, toothed near tip, heart-shaped at base, with stiff hairs on both surfaces. Flowers produced in heads to 15 cm (6 in) across, single, stalked, base of head cup-shaped, ***ray flowers with yellow petals 2.5–5 cm*** (1–2 in) ***long,*** central disc flowers lacking showy petals, brown. Fruit flattened, dry, 1-seeded, 4–6 mm (0.2 in) long. **Harvest:** Gather seed heads in late summer to early autumn before seeds are dry enough to be released. Hang in warm, dry place. **Preparation:** Indians parched then ground unshelled seeds into fine meal. Meal was added to water and drunk or mixed with marrow to form dough and eaten. To separate nutmeats from shells, break hulls with chopper, very coarse grinder, or food mill. Pour broken seeds into water, stir vigorously, let settle. Nutmeats will sink and shells will rise to top. Skim shells for making coffee. For use in nut recipes, drain, dry, and roast nutmeats. Grind nutmeats into meal and use for part of flour in muffin and bread recipes. For nut butter grind raw nutmeats into paste; add honey or maple syrup if necessary to improve flavor. To produce oil, thoroughly crush or grind raw seeds. Boil in water. Skim oil off broth. **Related edible species:** All of the nearly 60 species in North and South America are edible. **Poisonous look-alikes:** none.

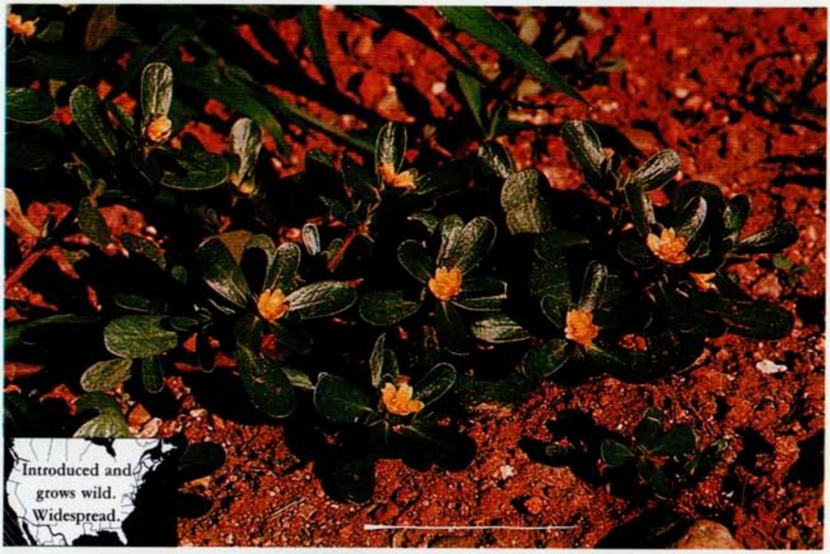
Introduced and
grows wild.
Widespread.

Purslane: flowering plant (NYBG).

Sunflower: flowering stalks (Elias). Cultivated form with larger flowering head (Feldman).

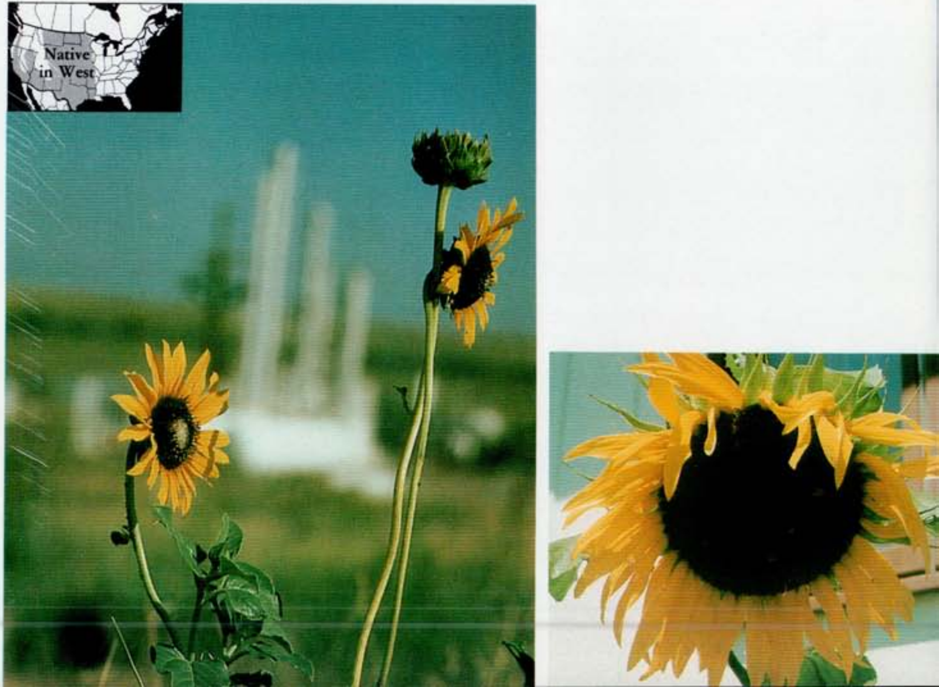
Native
in West

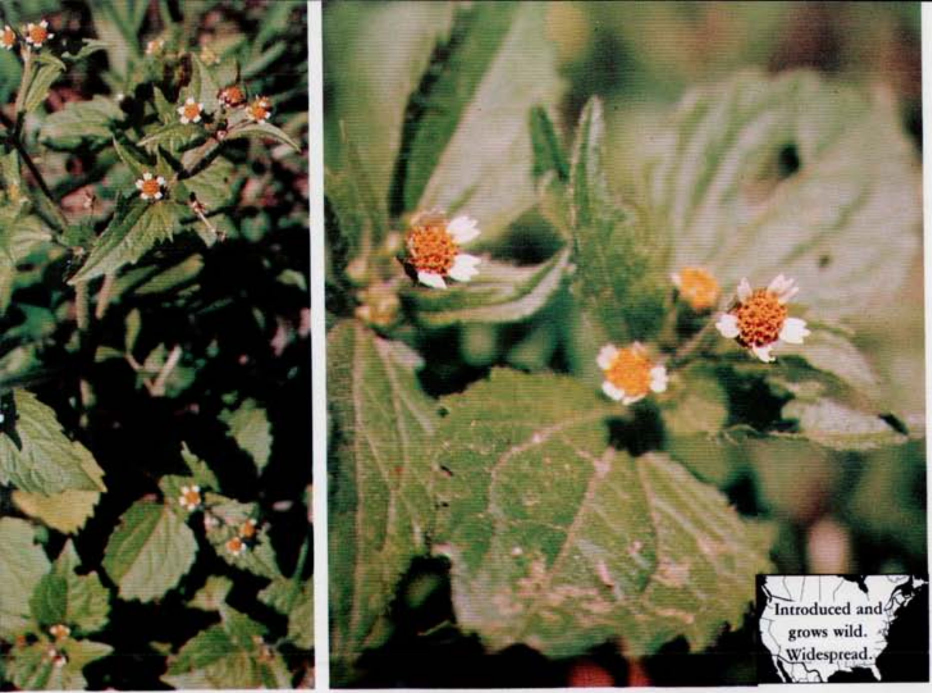
Introduced and
grows wild.
Widespread.

Quickweed: flowering plant and flowers (NYBG).

Bearberry: fruiting branches (Elias).

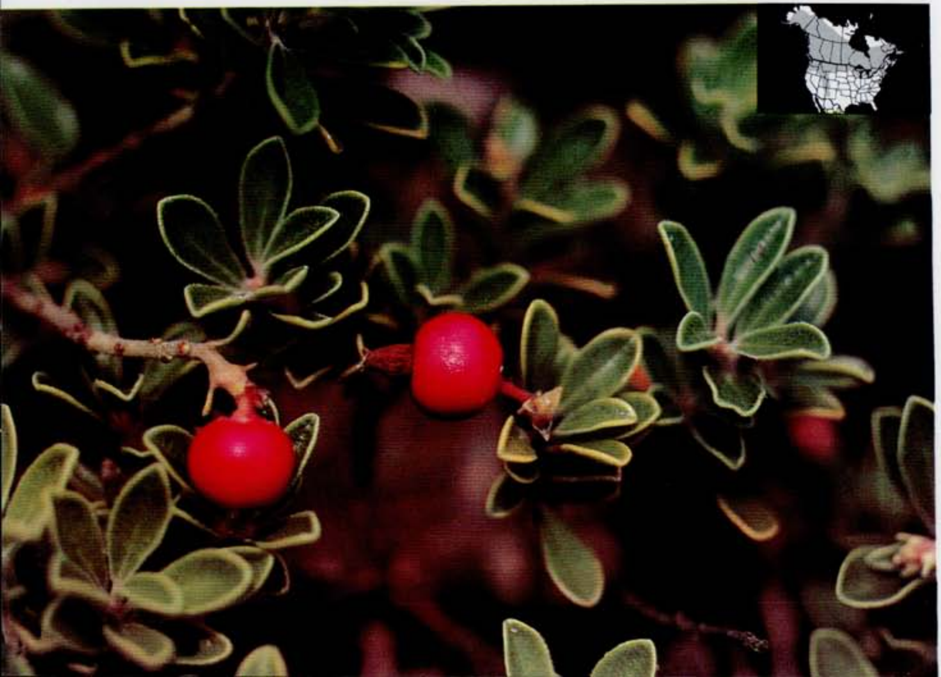

Quickweed, Galinsoga *Galinsoga parviflora*

Habitat: disturbed sites, roadsides, yards; also garden weed. **Identification:** annual herb to 70 cm (28 in) high, often bushy, stems usually spreading, smooth to slightly hairy. ***Leaves opposite,*** simple, usually lance-shaped, 2.5–8 cm (1–3.1 in) long, 1.2–4 cm (0.5–1.6 in) wide, pointed at tip, small blunt teeth along margin, tapering to rounded at base, with 3 principal veins, papery, usually stalked. ***Flowers tiny, produced in heads, 4–5 outer scale-like bracts (involucre) 2–3 mm (0.1 in) high, green; outer (ray) flowers white, inner (disc) flowers yellow.*** Fruits small hard nutlets 1–1.5 mm (0.1 in) long, nearly black, hairy. **Harvest:** young plants during middle of growing season. **Preparation:** Discard roots. Boil tops about 15–25 min and serve with salt, pepper, and butter. Add more flavorful greens and vinegar to compensate for quickweed blandness. **Related edible species:** *G. ciliata* (also called quickweed) plus a few other *Galinsoga* species, all introduced from southwest U.S., Mexico, and points south. **Poisonous look-alikes:** none.

Bearberry, Kinnikinick *Arctostaphylos uva-ursi*

Habitat: circumboreal, subarctic; sandy or rocky soils, higher elevations in southern part of range. **Identification:** ***low spreading shrub, forming mats*** with tough flexible twigs. ***New twigs hairy and sticky; mature ones*** dark reddish-brown or gray, ***papery and peeling. Leaves*** alternate, simple, ***evergreen, 1–3 cm*** (0.4–1.2 in) ***long,*** broadest near rounded tip, entire, shiny above, leathery. Flowers tiny, 5–12 per elongated cluster, each white to pale pink, urn-shaped, 4–6 mm (0.2 in) long. ***Fruits bright or dull red berries 4–10 mm (0.2–0.4 in) in diameter, mealy,*** not tasty. **Harvest and preparation:** leaves called *kinnikinick,* an Indian tobacco substitute also used by frontiersmen. For smoking, leaves are mildest in summer; dry in sun or near fire, crumble, and light up. Dried leaves make medicinal tea, good for indigestion. Fruits ripen late in growing season. Raw, they are bland and dry, but cooked they become palatable with cream and sugar. Not great snack but useful survival food. **Related edible species:** alpine bearberry *(A. alpina)* and several manzanitas plus many species in other genera in heath family. **Poisonous look-alikes:** none.

Salal, Western wintergreen *(Also edible autumn)* *Gaultheria shallon*

Habitat: woods, clearings, brushy areas below 800 m (2,630 ft), from West Coast to west slope of Cascades. **Identification:** ***upright, many branched shrub*** 0.3–2.5 m (1–8.2 ft) tall, frequently forming dense thickets. ***Leaves evergreen,*** alternate, simple, uniformly wide to broadest near middle, 3–10 cm (1.2–4 in) long, finely toothed, dark, shiny above, paler beneath. Flowering clusters narrow, elongate, composed of 5–15 stalked flowers, each urn-shaped, white or pink, 8–10 mm (0.3–0.4 in) long; prominent leaflike bract and 2 bractlets at base of flower stem. ***Fruits dark-purple berries,*** 7–8 mm (0.3 in) in diameter, nearly round, thick skinned, star-shaped depression on end, with several seeds. **Harvest:** Fruits ripen in summer and persist several weeks. **Preparation:** edible fresh but sometimes quite spicy and not juicy; better cooked. Also dry in sun or warm attic. Good in jam, jelly, pies, baked goods. See basic jelly and jam recipes, in book's Introduction. For beverage crush and boil berries, extract juice, and add equal amount of water. Sweeten, if necessary. **Related edible species:** WINTERGREEN, CREEPING WINTERGREEN (next), CREEPING SNOWBERRY, and *G. ovatifolia,* also called salal, of northwest U.S. **Poisonous look-alikes:** none.

Creeping wintergreen *(Also edible autumn, winter)* *Gaultheria humifusa*

Habitat: subalpine and alpine zones and moist slopes of western and Pacific Northwest mountains. **Identification:** low shrub with short spreading branches along ground. ***Leaves*** alternate, simple, ***evergreen,*** broad, ***broadest near base or middle, 1–2 cm (0.4–0.8 in) long, almost as wide,*** entire or with tiny teeth along margin, leathery. Flowers solitary in junction of leaves, each white, urn-shaped, 3–4 mm (0.1–0.2 in) long. ***Fruits are dry berries 4–8 mm*** (0.2–0.3 in) ***in diameter,*** rounded, ***red at maturity.*** **Harvest and preparation:** same as WINTERGREEN. Dryness makes fruit less palatable than other *Gaultheria* species. **Related edible species:** See SALAL (preceding). **Poisonous look-alikes:** none.

Labrador tea *(Also edible autumn, winter, spring)* *Ledum groenlandicum*

Habitat: bogs, fens, heaths, northern woods, exposed bluffs. **Identification:** low shrub up to 1 m (3.2 ft) high, with erect, reddish-brown branches, densely hairy when young. ***Leaves*** alternate, simple, ***evergreen,*** fragrant, crowded near tip of twigs, 1.5–5 cm (0.6–2 in) long, 4–12 mm (0.2–0.5 in) wide, uniformly wide to slightly wider at middle, rounded at tip, ***margin entire and rolled under, leathery,*** dull dark green above, ***densely hairy beneath, white at first, turning rusty.*** Flowers produced in late spring or summer in rounded clusters, each with 5 cream-white petals, stalked. Fruits are dry capsules 4–6 mm (0.2 in) long, slender, long-stalked, splitting open from base. **Harvest:** any time but least good in midwinter. Pick leaves when dense pile beneath is rust colored. Dry in sun, over low fire, or in oven. **Preparation:** Steep 1 tbsp dried leaves in 1 cup boiling hot water for 10 min. Do not boil water with leaves in it, which may release a harmful alkaloid. Serve hot or cold. **Related edible species:** northern Labrador tea *(L. palustre).* **Poisonous look-alikes:** mountain Labrador tea *(L. glandulosum)* is described by some sources as poisonous.

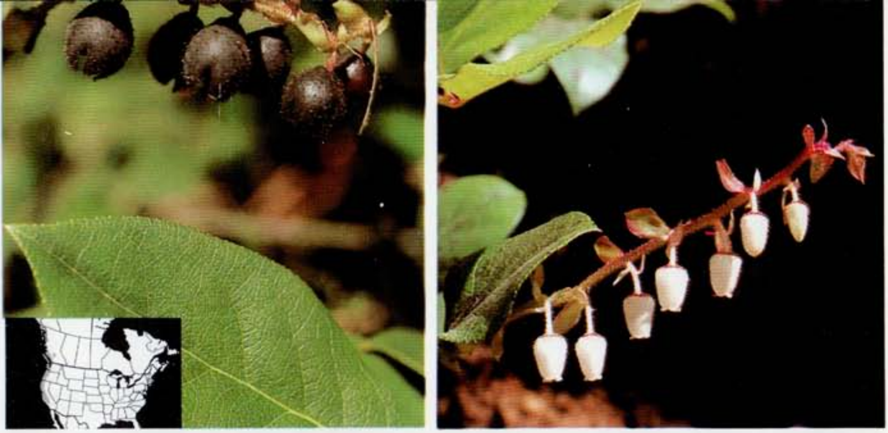

Salal: fruiting and flowering branches (Ross).

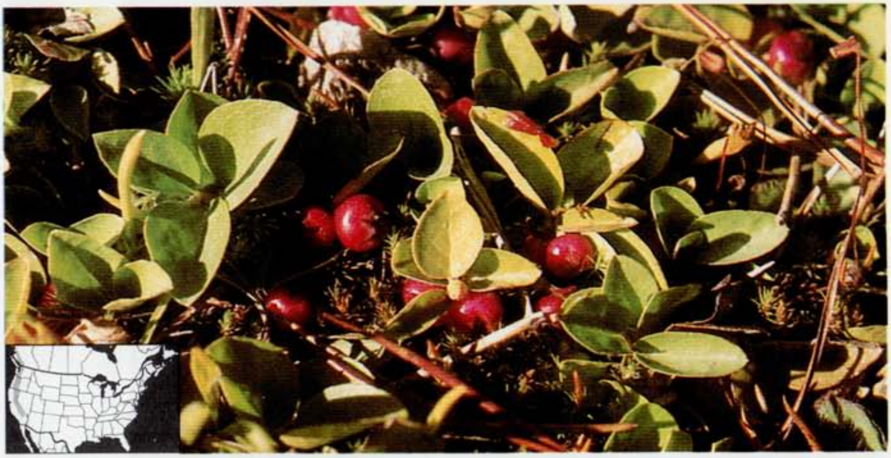

Creeping wintergreen: fruiting plants (Mulligan).

Labrador tea: flowering plant (Elias).

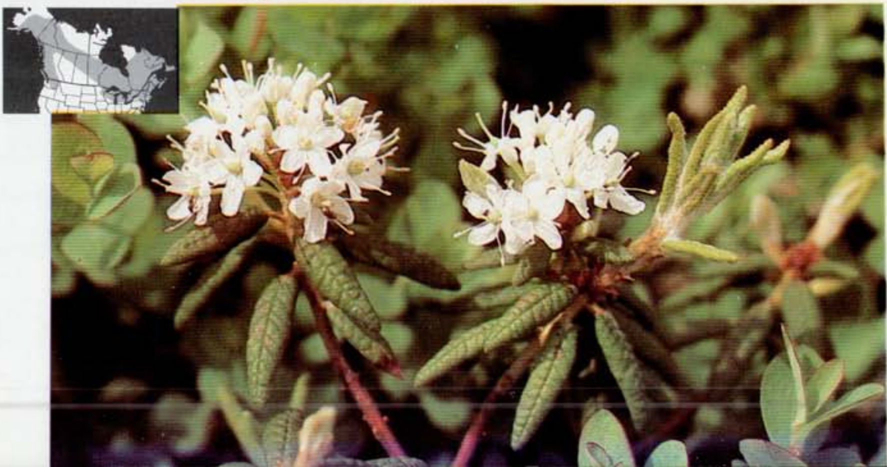

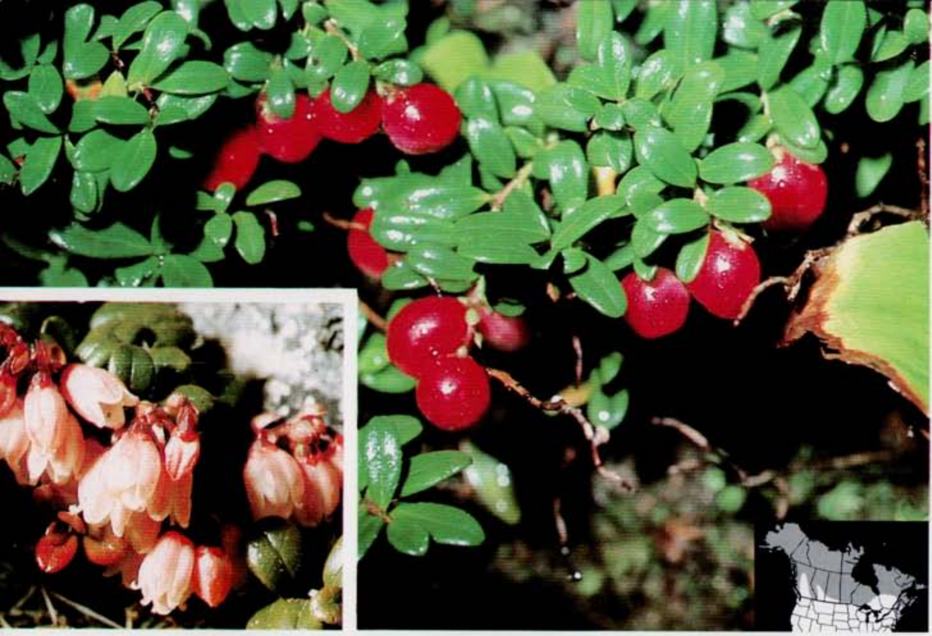

Mountain cranberry: flowers (NYBG). Fruiting plant (Elias).

Cranberry: flowering plants (NYBG). Fruit (Elias).

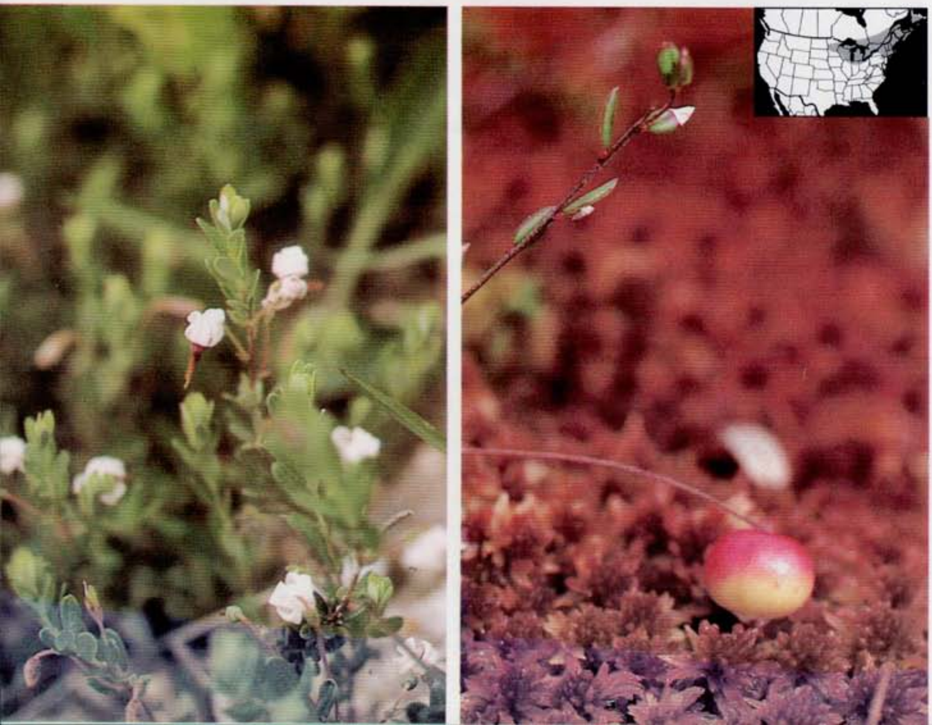

The BLUEBERRY and CRANBERRY GENUS *Vaccinium*

Blueberries, cranberries, huckleberries, and bilberries are all closely related groups of species of the *Vaccinium* genus. There are perhaps 35 species in North America, but hybridization is common. Fruits of most species are edible and range in taste from tart to sweet. The domesticated blueberries are selections of the highbush blueberry *(V. corymbosum)*. Wild blueberries sold commercially are commonly obtained from late sweet blueberry *(V. angustifolium)* or velvet-leaved blueberry *(V. myrtilloides)*. The supermarket cranberry *(V. macrocarpon)* is the same as the wild. Berries are rich in vitamin C.

Mountain cranberry, Lingenberry, Foxberry *Vaccinium vitis-idaea*

Also edible autumn, winter

Habitat: widespread subarctic and alpine plant, bogs, cold seepage areas, rocky acid soils. **Identification:** ***small, creeping plant,*** to 15 cm (6 in) with slender stems; ***leaves*** alternate, simple, ***evergreen,*** broadest above middle, ***0.6–1.8 cm (0.2–0.7 in) long, 4–10 mm (0.2–0.4 in) wide,*** entire, leathery, with scattered black glandular hairs beneath, midrib furrowed. Flowers tiny, pink to reddish, 6–8 mm (0.2–0.3 in) long, narrowly bell-shaped. ***Fruits red berries, rounded, 6–10 mm (0.2–0.4 in) in diameter,*** often persisting over winter, slightly acid tasting. **Harvest and preparation:** Fruits ripen August and September; many persist on plant through winter. Tart to bitter at first; flavor improves after frost. Good trail snack, better used in any cooked cranberry recipes. **Related edible species:** CRANBERRY (next). **Poisonous look-alikes:** none.

Cranberry *(Also edible autumn, winter)* *Vaccinium macrocarpon*

Habitat: widespread northern plant; bogs, swamps, shores of cold acid lakes. **Identification:** ***low, trailing shrub*** with slender brown, flaky-barked stems; ***leaves*** alternate, simple, ***evergreen,*** uniformly wide to broadest near middle, ***5–16 mm (0.2–0.6 in) long,*** rounded at tip, leathery, whitish beneath. Flowers single or in few-flowered clusters on long stalks in junction of upper leaves, each stalk with 2 leaflike bracts, each flower pink, urn-shaped, 4 petals bent back, 6–10 mm (0.2–0.4 in) long. ***Fruits*** are ***berries 1–2 cm (0.4–0.8 in) in diameter,*** globe-shaped, green turning ***bright red at maturity,*** each on long slender stalk. **Harvest:** Fruits ripen September through October and persist into winter in protected areas. Fruits may be harvested with toothed scoop. **Preparation:** Fresh acidic fruits keep well in cool storage. Freeze or dry for long storage. To dry berries, spread on tray or foil in oven for 4 hr at 110°C (230°F). Experiment with substituting dried, powdered berries for part of flour in baking recipes. Most common use of fresh and frozen berries is in sauce served with roast turkey. Bring to boil 2 lb fresh cranberries, 2 cups sugar, 1 cup water (or berries plus 1 cup maple sugar and 1 cup cider). Reduce heat and simmer until all berries pop, and mix looks glassy. For storage, ladle into sterile jars and seal. Check recipe books. **Related edible species:** MOUNTAIN CRANBERRY (preceding); small cranberry *(V. oxycoccos)* is similar but smaller. HIGHBUSH CRANBERRY is not related but has almost identical fruits. **Poisonous look-alikes:** none.

Bog bilberry *Vaccinium uliginosum*

Habitat: widespread boreal and arctic plant; bogs, rocky outcrops, high mountain slopes. **Identification:** ***low shrub to 0.6 m*** (2 ft) high; branches stout, usually spreading; ***leaves*** alternate, simple, ***deciduous, broadest near tip or middle to uniformly wide, 0.5–2 cm*** (0.2–0.8 in) ***long,*** rounded at tip, entire, firm to leathery texture, dull green. Flowers ***solitary or in clusters of 2–4 on short stalks*** in junctions of leaves; each pink to whitish, urn-shaped. ***Fruits fleshy berries 5–8 mm (0.2–0.3 in) in diameter, dark blue to black,*** sweet, with many tiny seeds. **Harvest and preparation:** same as LATE SWEET BLUEBERRY. **Related edible species:** See LATE SWEET BLUEBERRY. **Poisonous look-alikes:** none.

Highbush blueberry, Swamp blueberry *Vaccinium corymbosum*

Habitat: widespread in eastern North America; bogs, low wet woodlands, swamps, occasionally drier soils. **Identification:** ***shrubs,*** variable in shape, ***1–4.5 m*** (3.2–14.8 ft) ***high; leaves*** alternate, simple, ***deciduous,*** broadest near base or middle, sometimes narrowly so, ***3–8 cm (1.2–3.2 in) long,*** 2–3 cm (0.8–1.2 in) wide, usually entire, greenish-yellow. ***Flowers densely clustered,*** each tubular to urn-shaped, white to pink, 5–10 mm (0.2–0.4 in) long. ***Fruits fleshy berries, rounded, 6–12 mm (0.2–0.5 in) in diameter, dark blue to bluish-black, sweet,*** containing many tiny seeds. **Harvest:** Gather ripe fruits in late summer. They can be eaten as trail snack or stored for later use. Fruits ripen as do those of LATE SWEET BLUEBERRY (next), but are easier to gather. Hand pick or shake berries onto blanket or plastic tarp spread beneath bush. **Preparation:** Same as LATE SWEET BLUEBERRY. Also see basic recipes early in book. Add cup dried blueberries to favorite stew recipe. For fritters prepare batter from 1 cup biscuit mix, 3 tbsp sugar, ⅓ cup milk, 1 beaten egg, 1 cup fresh blueberries. Drop spoonfuls into hot cooking oil; brown both sides. Drain; dust with sugar. **Related edible species:** See LATE SWEET BLUEBERRY. **Poisonous look-alikes:** none.

Late sweet blueberry *Vaccinium angustifolium*

Habitat: dry rocky soils, barrens, mountain slopes, occasionally bogs; northcentral and northeastern North America. **Identification:** ***low shrub to 0.3 m*** (1 ft), much branched, often forming thickets. ***Leaves*** alternate, simple, deciduous, ***narrow*** and broadest at or below middle, 1–3 cm (0.4–1.2 in) long, ***finely toothed along margins,*** bright green, smooth. ***Flowers produced in short tight clusters,*** each nearly cylinder-shaped, white to pinkish tinged, 5–6 mm (0.2 in) long. ***Fruits*** are ***fleshy berries 5–12 mm*** (0.2–0.5 in) ***in diameter,*** rounded, ***bright blue,*** white bloom, sweet. **Harvest:** This and other blueberry species ripen in summer, generally after raspberries (RED RASPBERRY) and before blackberries (COMMON BLACKBERRY) where those berries grow. Pick by hand. **Preparation:** Blueberries were probably the most important fruit to American Indians. Excellent dried like raisins and fresh, both raw and in most berry recipes, especially pies, jams, and breads (see basic recipes early in book). Dry thoroughly in sun or warm attic about 10 days or in 100°C (210°F) oven for 4–5 hr; store in sealed containers. For bannock, blend 2 cups sifted flour, 3 tsp baking powder, ½ tsp

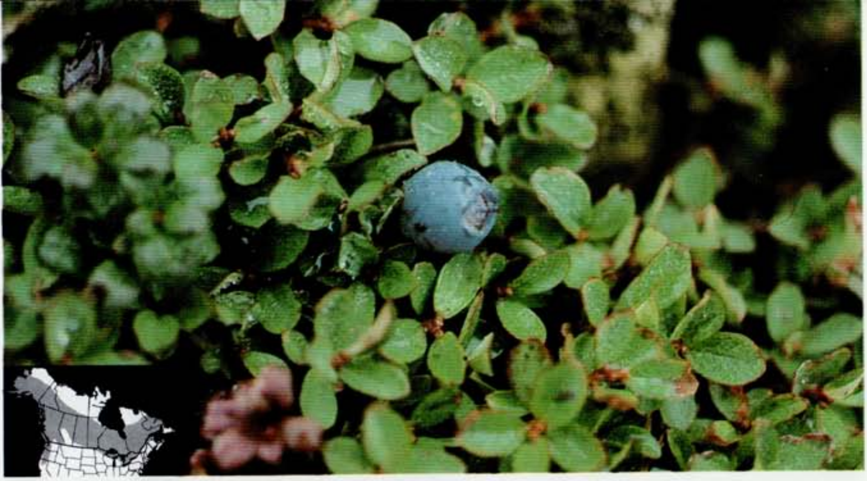

Bog billberry: fruiting plant (Elias).

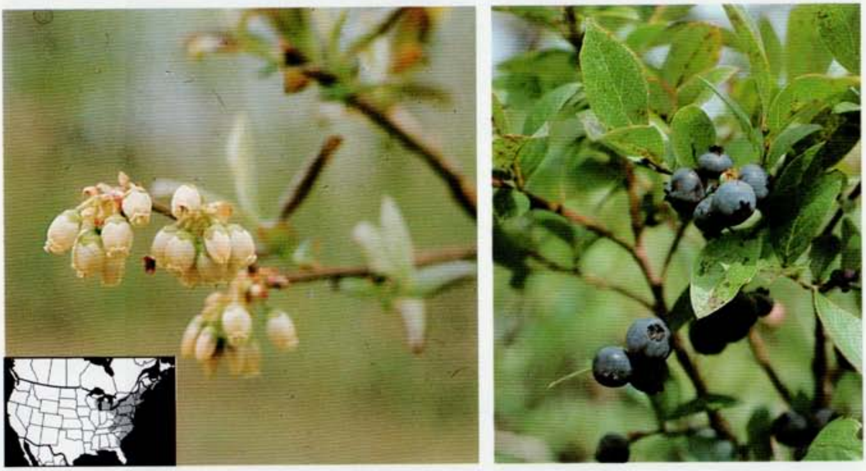

Highbush blueberry: flowering branch (NYBG). Fruiting branch (Elias).

Late sweet blueberry: flowering and fruiting branches (Speas).

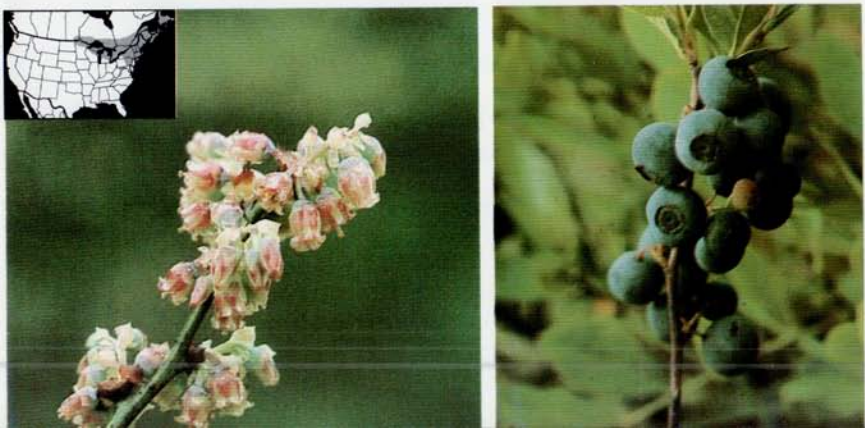

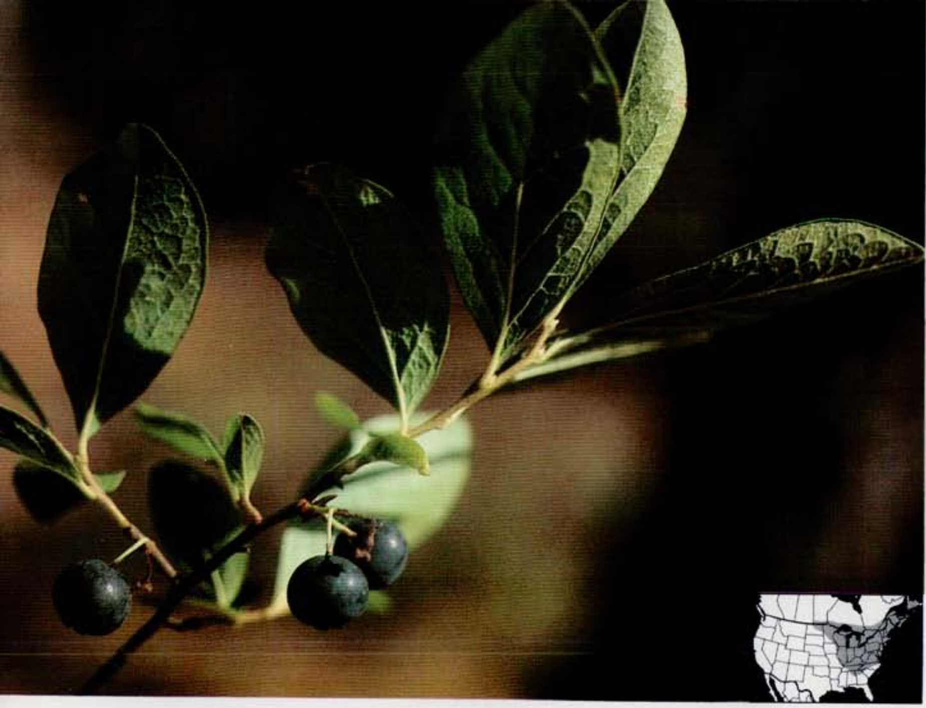

Black huckleberry: fruiting branch (Elias).

Beach plum: fruiting branches (Elias).

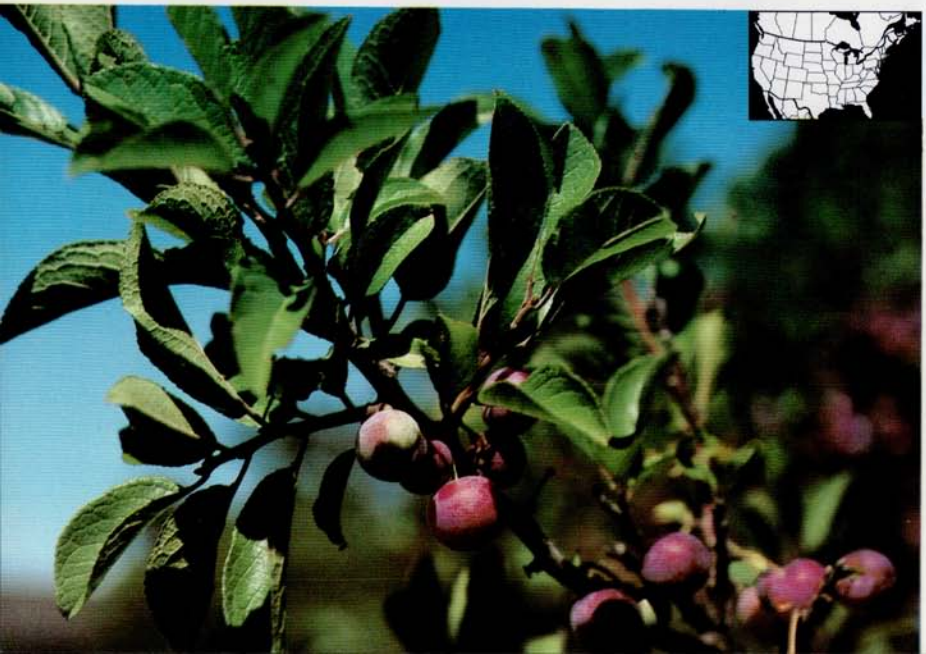

salt, 6 tbsp shortening, 4 tbsp powdered milk. Mix in 1 cup blueberries, then ⅓ cup water. Shape into 1 in thick cake, dust with flour, place in heated frying pan. Cook over coals until bottom is crusty. Turn with spatula and bake until done; test with straw or sliver. See HIGHBUSH BLUEBERRY (preceding) for other uses. **Related edible species:** Other *VACCINIUM*. **Poisonous look-alikes:** none.

Black huckleberry *Gaylussacia baccata*

Habitat: woodlands, clearings, dry to moist soils. **Identification:** shrub to 1.2 m (4 ft), often forming low thickets; branches numerous, stiff, resembling blueberry bush. ***Leaves*** alternate, simple, deciduous, ***3–6 cm*** (1.2–2.4 in) ***long,*** uniformly wide to broadest near or below middle, entire, ***gland-dotted above and below. Flowers few on short elongated clusters, each white,*** small, cone-shaped to almost tubular, 4–7 mm (0.2–0.3 in) long. ***Fruits fleshy berries 6–9 mm*** (0.2–0.4 in) ***in diameter, black at maturity,*** containing 10 tiny seeds, sweet. **Harvest and preparation:** Pick berries as they ripen in late June in south to September in far north. Eat as fresh fruit or use in desserts, pancakes, jam, and jelly. For pancakes add 1 cup berries to 3–4 cups batter. To dry, spread berries in single layer on tray, foil, or paper in warm attic; dry about 10 days or until squashed berry releases no juice. Store in sealed container. Use dried berries in baking or pemmican. Or add ½–1 cup dried berries to beef stew. See LATE SWEET BLUEBERRY and HIGHBUSH BLUEBERRY for other recipes. **Related edible species:** Dwarf huckleberry *(G. dumosa),* box huckleberry *(G. brachycera),* and dangleberry *(G. frondosa)* all yield berries edible but inferior to blueberries. **Poisonous look-alikes:** none.

Beach plum *Prunus maritima*

Also edible autumn

Habitat: coastal regions, sandy soils. **Identification:** shrub to 2.5 m (8 ft) or low and straggling; branches hairy when young, smooth later. ***Leaves*** alternate, simple, deciduous, ***3–7 cm*** (1.2–2.8 in) ***long,*** 1.5–3.5 cm (0.6–1.4 in) wide, broadest at or below middle, rarely widest above middle, pointed at tip, ***sharply toothed along margin,*** dull green and smooth above, softly hairy beneath. ***Flowers 2–4 per cluster, each 1–1.5 cm*** (0.4–0.6 in) ***across,*** with 5 ***white*** petals. ***Fruits rounded*** or nearly so, ***1.5–2.5 cm*** (0.6–1 in) ***in diameter, deep purple to reddish-black at maturity,*** fleshy, sour tasting, enclosing large stone. **Harvest:** Pick fruit from August through October, depending on location. Slightly unripe fruit best for jelly. **Preparation:** similar to AMERICAN PLUM but fruit is more tart. Use jelly and jam recipes early in book. Also, make jellies and jams from blend of plum and other fruits. Combine with peaches or other sweet pulpy fruit for pie filling; see basic plum pie recipe early in book. To dry fruit for later use, slit skin and remove pit. Pits contain toxic hydrocyanic acid. Spread fruit in sun or warm oven until no juice can be squeezed from them, about 4–6 hr. Makes tart trail snack. **Related edible species:** AMERICAN PLUM, FLATWOOD PLUM, and several others. **Poisonous look-alikes:** none.

CURRANT or GOOSEBERRY GENUS *Ribes*

This large group of about 120 species is centered in western North America and the Andes Mountains of South America. They are shrubs, sometimes armed with spines, with simple, maplelike, alternate leaves, flowers in small clusters, and a 1-celled pulpy berry, sometimes transparent, variously colored, containing several seeds. North American species are sometimes difficult to distinguish. Fortunately, none is known to be poisonous. We recommend 2 common yellow-flowered species with smooth fruits and 2 pink- to purple-flowered species with glandular hairy fruits. **Harvest:** Pick fruits in summer as they ripen. **Preparation:** Sweetness and juiciness vary greatly among species. Some fruits can be eaten fresh. Some are spiny and so require cautious eating. Adapt recipes to berry characteristics. For no-cook currant or gooseberry preserves, wash and mash fruits thoroughly. Slowly mix in sugar equal to volume of crushed fruit. Pour into sterilized jelly glasses and seal. For preserves wash and stem 1 lb currants; add ¾ lb sugar and let stand overnight. Bring slowly to boiling point and boil 3–5 min. Let stand overnight, pour into jelly glasses, and seal. To make

Golden currant: flowers (NYBG). Fruiting branches (USDA).

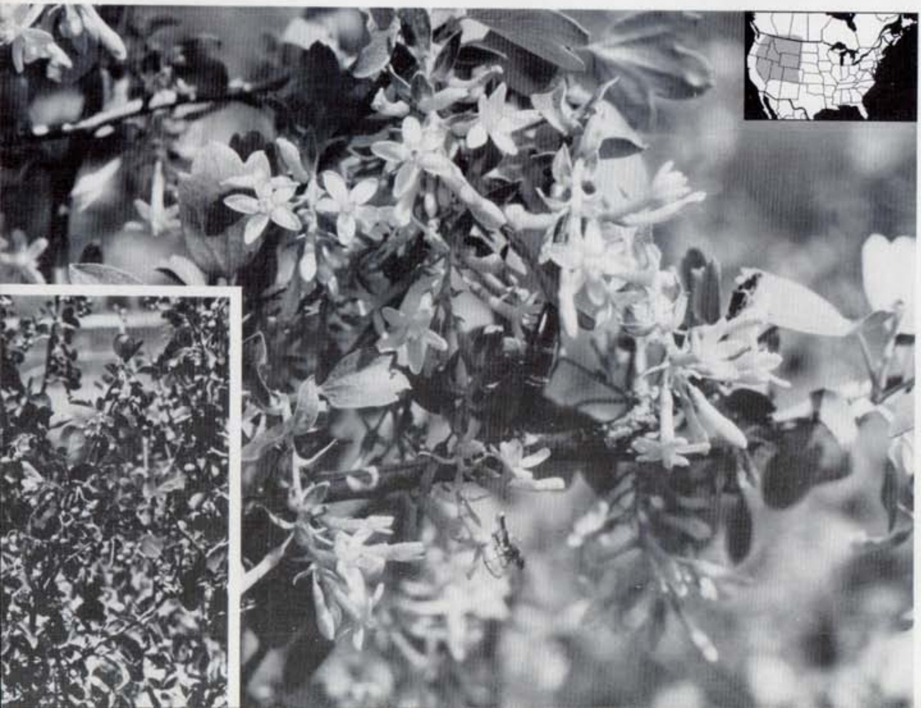

currant/gooseberry conserve, wash and stem 3 lb berries. Grate rind of 3 oranges and cut up pulp. Mix berries, orange rind and pulp, 1 lb seeded raisins, and 3 lb sugar, cook slowly until thick. Pour into sterilized jars and seal. For simple currant/gooseberry pie wash and stem 1 pt berries. Cook in saucepan with 1 cup sugar and small amount of water about 15 min until berries soften. Pour into pie crust, add top crust, bake at 175°C (350°F) oven until crust is done. Add fresh or dried berries to kuchen dough and muffins. To dry whole or mashed berries spread in thin layer on baking sheet; place in sun or 105°C (220°F) oven until dry. Store in paper bag in dry place. To make pemmican mix 1 lb dried and finely ground or pounded meat with ¼ cup brown sugar and ½ cup dried berries. Add ¾ lb melted suet and blend into paste. Form into bars and store dry. Eat as is, boil, or fry. For pie and cooked jelly see recipes early in book. **Related edible species:** many in addition to the four described; widely distributed. **Poisonous look-alikes:** Sticky currant *(R. viscosissimum)* is reported to be a strong vomitory. It has a sticky substance on the leaves, twigs, and fruit. Some fruits are covered with short, bristly hairs.

Golden currant *Ribes aureum*

Habitat: along streams, ravines, washes, mountain slopes, mainly between 750–2,380 m (2,460–7,800 ft) elevation. **Identification:** shrub, erect, open, to 2 m (6.6 ft); branches stiff, nearly smooth. ***Leaves*** alternate, deciduous, 1.2–4.5 cm (0.5–1.8 in) long, 1.5–5 cm (0.6–2 in) wide, ***broadly rounded to widest near base, 3-lobed,*** rounded at tip, entire to coarsely toothed along margin, ***light green,*** smooth, stalked. ***Flowers*** produced in drooping, 5–15-flowered, elongated clusters in late spring, each flower ***bright yellow,*** 5-parted, 6–10 mm (0.2–0.4 in) long, ***smooth. Fruits stalked berries,*** rounded, 6–9 mm (0.2–0.4 in) in diameter, ***red*** or occasionally black, ***smooth, beaked*** (persistent floral parts), produced in summer. **Harvest and preparation:** See CURRANTS, GOOSEBERRIES (preceding). This species both wild and cultivated was much used by western Indians; one of the best flavored species.

Wax currant *Ribes cereum*

Habitat: canyons, dry ravines, hillsides, prairies, open woodlands. **Identification:** open spreading shrub to 1 m (3.2 ft) high; branches stiff, crooked. ***Leaves*** alternate, simple, deciduous, 2–2.5 cm (0.8–1 in) long, 2.5–3 cm (1–1.2 in) wide, ***broadly rounded, wider than long, shallowly 3-lobed,*** rounded at tip, irregularly ***round toothed along margin,*** yellowish-green, glandular hairy on both surfaces, leafstalk 8–15 mm (0.3–0.6 in) long. ***Flowers*** in drooping, 1–6-flowered, elongated clusters in late May or June, each flower ***white to pinkish,*** 5-parted, 6–10 mm (0.2–0.4 in) long. ***Fruits stalked berries, globe-shaped,*** 8–12 mm (0.3–0.5 in) in diameter, ***dull to bright red, with stalked glandular hairs,*** beaked (the persistent floral parts), produced in July or August. **Harvest and preparation:** See CURRANTS, GOOSEBERRIES (preceding). Wax currant is native in West, though frequently cultivated in East.

Subalpine prickly currant *Ribes montigenum*

Habitat: along streams, wet forests, ravines, washes, mainly in subalpine zone 2,140–3,800 m (7,000–12,500 ft) elevation. **Identification:** ***shrub,*** open, spreading, sometimes straggling, to 0.7 m (28 in) high; branches tough, ***densely covered with glandular hairs, stiff bristlelike prickles with small spines at leaf bases. Leaves*** alternate, deciduous, simple, 0.8–3 cm (0.3–1.2 in) long, as wide or wider, almost circular in outline, ***deeply 5-lobed,*** pointed at tip, sharply toothed along margin, glandular hairy on both surfaces, stalked. Flowers produced in drooping, few-flowered clusters in late spring or early summer, each ***flower*** stalked, ***purplish,*** 5-parted, 4–8 mm (0.2–0.3 in) long, ***covered with glandular hairs. Fruits*** stalked ***berries,*** rounded, 4–6 mm (0.2 in) in diameter, ***red, covered with glandular hairs,*** produced in summer. **Harvest and preparation:** See CURRANTS, GOOSEBERRIES (preceding).

Fragrant golden currant *Ribes odoratum*

Habitat: prairies, open woodlands, bluffs, ledges, ravines. **Identification:** ***shrub,*** upright, open, to 1.5 m (5 ft); branches tough, reddish-brown, hairy, ***spineless. Leaves*** alternate, simple, deciduous, 3–6 cm (1.2–2.4 in) long, slightly wider, broadly rounded to widest near base, ***3–5-lobed,*** pointed to blunt at tip, coarsely toothed along margin, hairy, stalked. ***Flowers*** produced in drooping, few-flowered clusters from older wood in spring, each flower ***pale yellow,*** 5-parted, 12–16 mm (0.5–0.6 in) long, smooth. ***Fruits, produced in summer, stalked berries, globe-shaped, 9–12 mm*** (0.4–0.5 in) ***in diameter, purplish-black, smooth,*** with beaked persistent floral parts. **Harvest and preparation:** See CURRANTS, GOOSEBERRIES (preceding).

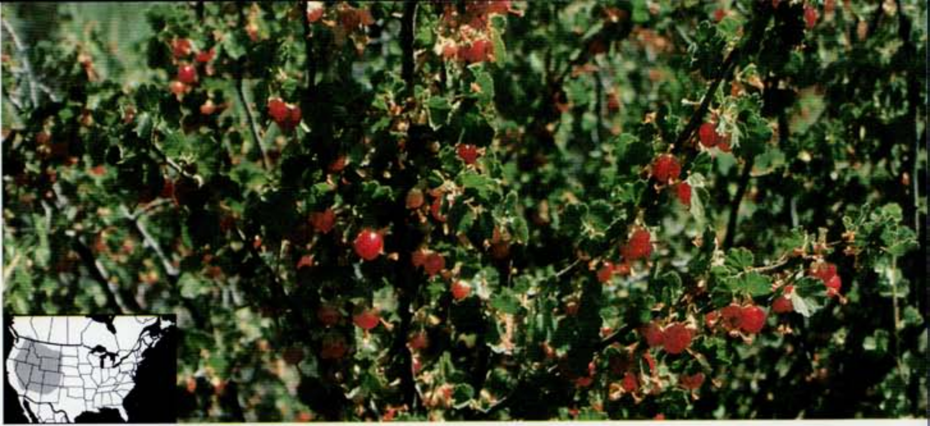

Wax currant: fruiting branches (Ross).

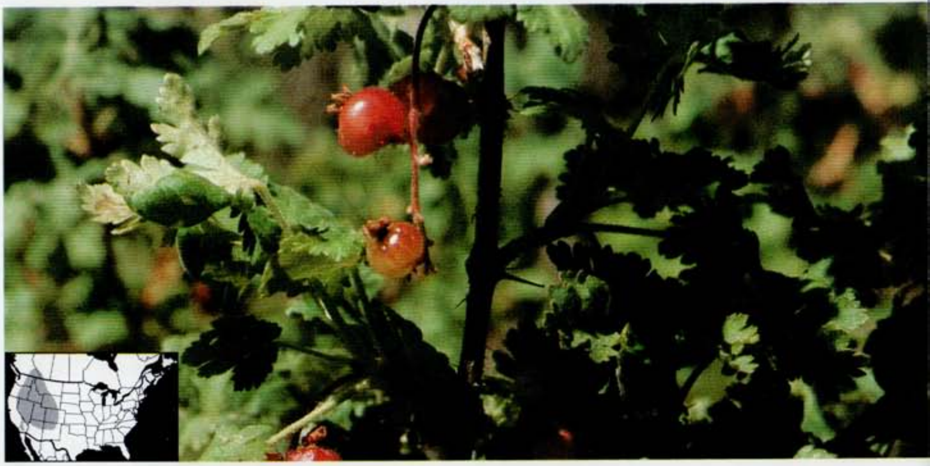

Subalpine prickly currant: fruiting branch (Dykeman).

Fragrant golden currant: fruiting branches (Kavasch).

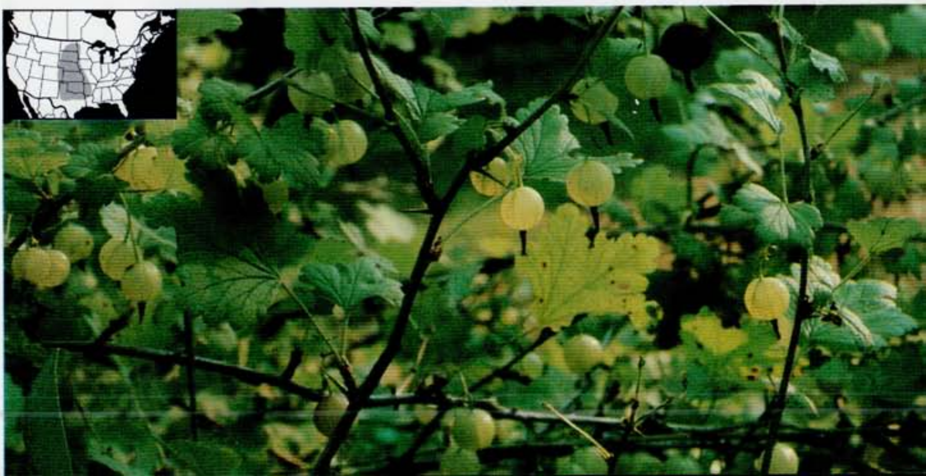

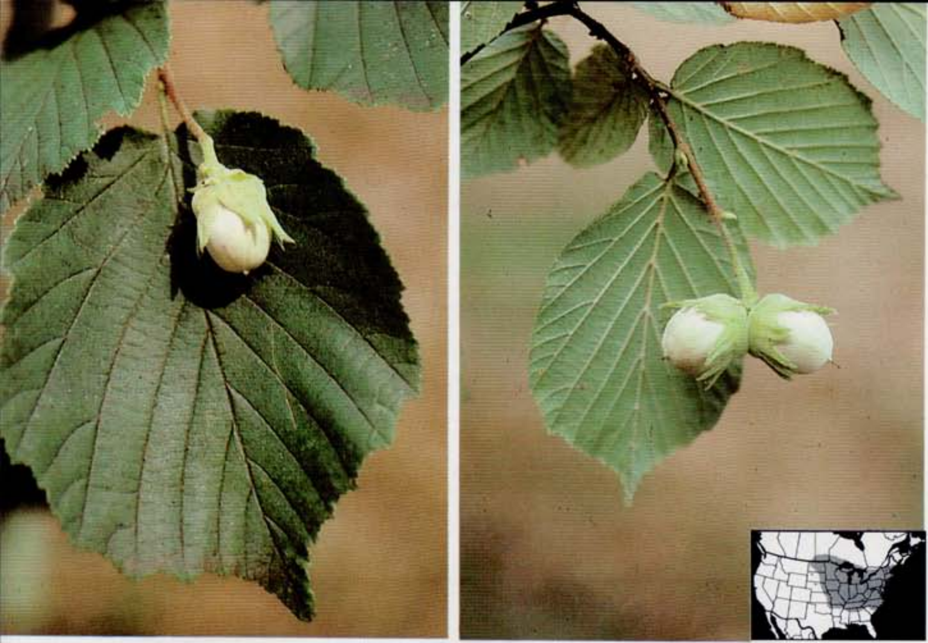

American hazelnut: fruiting branches (Elias).

Beaked hazelnut: fruits (Speas).

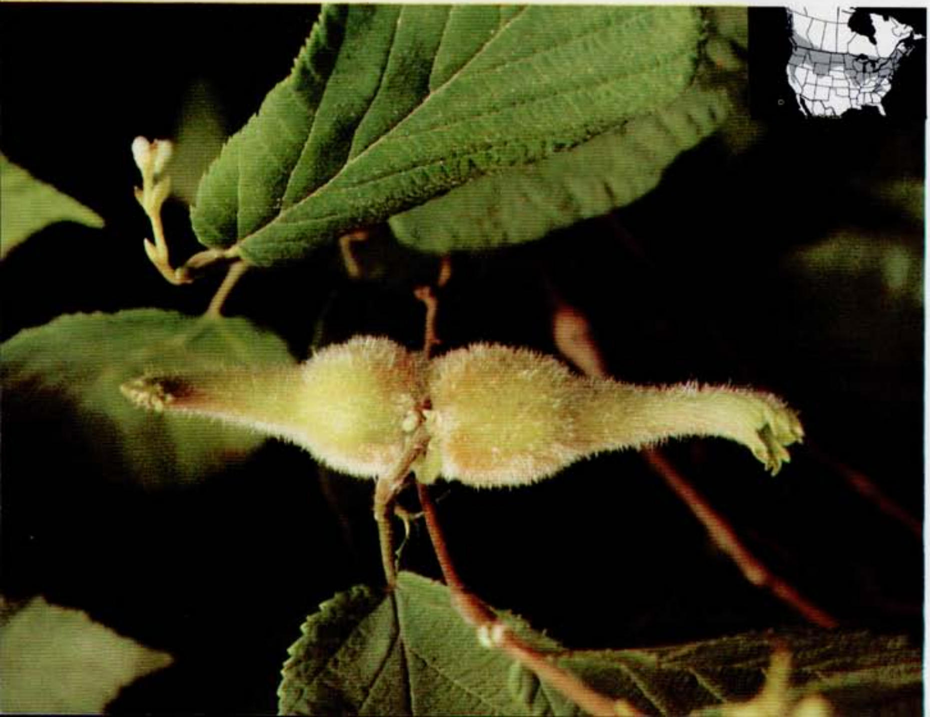

American hazelnut *Corylus americana*

Also edible autumn

Habitat: widespread, dry to moist woodlands, thickets. **Identification:** shrub to 5 m (16.4 ft) with rounded head and glandular hairy twigs. ***Leaves*** alternate, simple, deciduous, 5–15 cm (2–6 in) long, 3–12 cm (1.2–4.7 in), broadest below or above middle, long-pointed at tip, ***doubly toothed*** along margin, usually hairy beneath. Male flowers in dense slender catkins in early spring. Female flowers usually solitary or paired, tiny, inconspicuous in early spring. ***Fruits are hard-shelled nuts, broader than long,*** 1–1.5 cm (0.4–0.6 in) across, ***partially enclosed by 2 ragged-edged leaflike bracts,*** nutmeat sweet. **Harvest:** Nuts mature in late summer or autumn. **Preparation:** Remove husk, shell, and eat as is, or use in any recipe calling for nuts. For no-cook candy, mix 1 egg white with 2 cups confectioner's sugar. Add 2 tsp butter or margarine and blend until smooth. Work in 2 cups chopped hazelnuts and form into balls. Eat and enjoy. For hazelnut hot cakes, finely grind 0.2 kg (½ lb) dried hazelnuts. Boil in 2 cups water until mushy, about 30 min. Add 1 tsp maple syrup and ⅓ cup fine cornmeal, stir well, let stand for 20 min or until thick. Heat ⅓ cup frying oil in skillet. Drop batter by tablespoonfuls into hot oil. Brown on one side, flip, flatten, brown on other side. Drain and serve hot or cold. See also BEAKED HAZELNUT (next). **Related edible species:** BEAKED HAZELNUT. **Poisonous look-alikes:** none.

Beaked hazelnut *Corylus cornuta*

Also edible autumn

Habitat: dry or moist woodlands, hills or mountain slopes, absent on Coastal Plain. **Identification:** shrub to 3 m (9.8 ft), often forming thickets; twigs hairy when young. Leaves alternate, simple, deciduous, 3–10 cm (1.2–4 in) long, 2–8 cm (0.8–3.2 in) wide, broadest near base, long-pointed at tip, doubly toothed along margin, hairy beneath. Male flowers in dense, slender catkins in early spring. Female flowers solitary or few per cluster, tiny, inconspicuous. ***Fruits*** are ***thin-shelled nuts,*** broadest near base, 1–1.5 cm (0.4–0.6 in) in diameter, ***enclosed by 2 bristly leaflike bracts 4–7 cm*** (1.6–2.8 in) ***long which form narrow beak;*** nutmeat sweet. **Harvest:** Nuts mature in late summer or autumn. **Preparation:** Remove husk. Shell and eat as is or use in recipes calling for nuts or grind into meal to use as flour. For hazelnut torte preheat moderate oven 175°C (350°F). Beat 11 egg yolks well, add 0.5 kg (1 lb) confectioner's sugar, continuing to beat. Add 0.5 kg (1 lb) ground hazelnuts and 1 tsp instant coffee. Fold in 11 stiffly beaten egg whites. Bake in 30 cm (12 in) greased and floured spring baking form for 50–60 min. See also AMERICAN HAZELNUT (preceding). **Related edible species:** AMERICAN HAZELNUT. **Poisonous look-alikes:** none.

Sweet fern *Comptonia peregrina*

Also edible winter, spring

Habitat: dry rocky, sandy, or gravelly soils of clearings, pastures, poor open woodlands. **Identification:** small shrub 0.4–1.2 m (1.3–4 ft) high, with rounded shape; young branches hairy. ***Leaves*** in fernlike pattern, alternate, simple, deciduous, ***6–14 cm*** (2.4–5.5 in) ***long, 0.8–1.2 cm*** (0.3–0.5 in) ***wide,*** widest near middle or base, ***deeply lobed along margin, fragrant due to numerous resinous glands*** on leaves and twigs. Male and female flowers produced in separate small catkins on same bush in spring. ***Fruits small, green burrs 1–2 cm*** (0.4–0.8 in) ***in diameter,*** each partially or totally enclosing 1–4 nutlets, maturing in mid to late summer. **Harvest:** leaves any time for fragrant tea. Immature nutlets during June and July while still tender. **Preparation:** For tea add 1 tsp dried or 2 tsp fresh leaves per cup boiling water; steep. When camping, put 8 tsp fresh, chopped leaves per qt of water in glass jar. Cap and place in sun 3 hr or until water is dark. Strain, dilute to taste, sweeten if desired. Not bitter like many brewed teas. Strong tea (infusion) has been used to alleviate dysentery. Using thumbnail, separate nutlets from burrs when immature and eat raw. **Edible related species:** northern bayberry *(Myrica pensylvanica),* SOUTHERN BAYBERRY and SWEET GALE (next). **Poisonous look-alikes:** none.

Southern bayberry, Wax myrtle *Myrica cerifera*

Also edible autumn, winter

Habitat: coastal plain, swamps, along ponds and lakes, pine barrens, wet sandy soils. **Identification:** shrub or small tree to 8 m (26 ft) with rounded crown and slender, hairy, ***gland-dotted twigs. Leaves*** alternate, simple, deciduous but often persisting until second year, ***3.5–10 cm*** (1.4–4 in) ***long,*** 0.6–1.2 cm (0.2–0.5 in) wide, ***narrow*** but broadest near tip, ***coarsely toothed along margin, yellow-green, gland-dotted, fragrant when rubbed.*** Male and female flowers on separate bushes in spring; male flowers in slender catkins 1.2–1.9 cm (0.5–0.7 in) long; female flowers in broader catkins 0.6–1.2 cm (0.2–0.5 in) long on older twigs. ***Fruits few to several clustered on short spikes, each rounded, 2–4 mm (0.1 in) in diameter, covered with light blue waxlike coating.*** **Harvest and preparation:** Pick leaves any time available but best in early summer. Wash and spread to dry thoroughly in warm shaded area. Store in sealed jars in dark place. Crumble and blend into stews and sauces for cooking and serving meats. Also use berries picked from late summer into winter as seasoning and to provide wax used in candles. **Related edible species:** northern bayberry *(M. pensylvanica)* and other wax myrtle species are used the same way. Also SWEET GALE (next) is used for seasoning but fruits lack wax. **Poisonous look-alikes:** none.

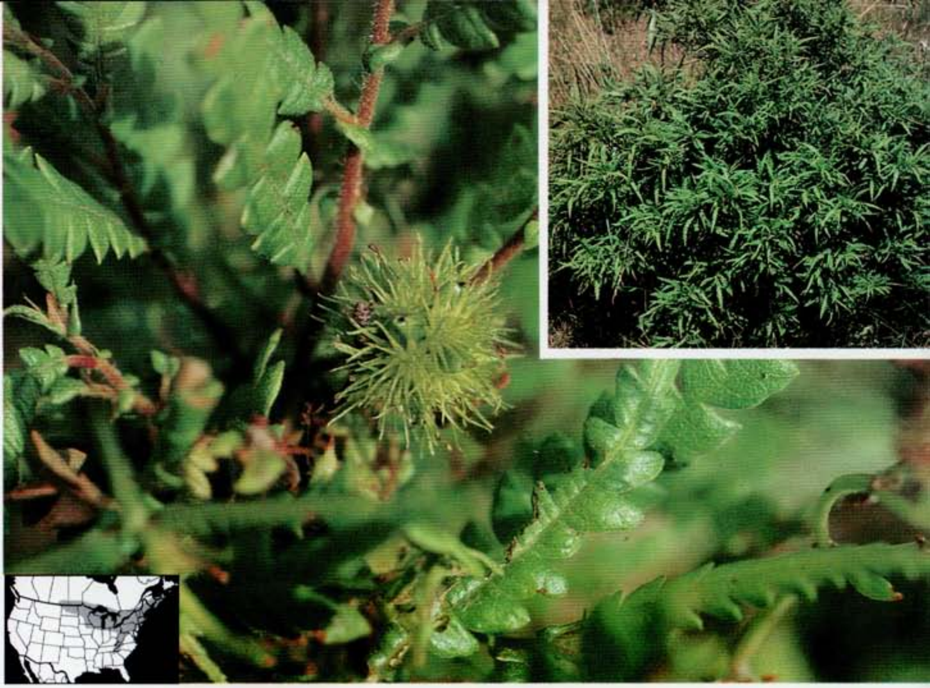

Sweet fern: mature plant and branch with fruit (Elias).

Southern bayberry: fruiting branches (NYBG).

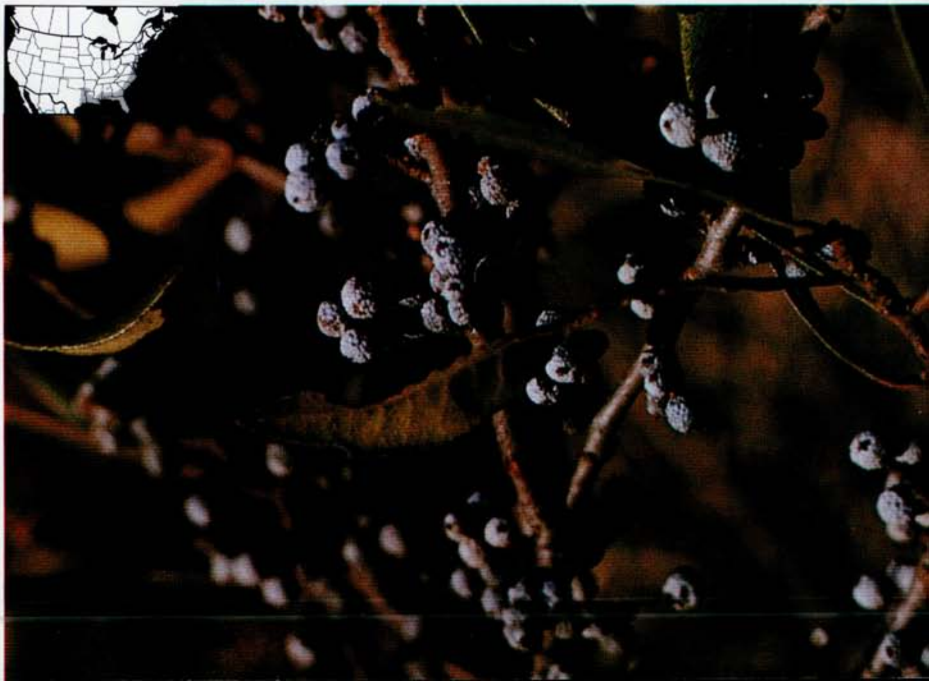

Sweet gale *Myrica gale*

Also edible autumn, winter

Habitat: swamps and shallow water along ponds and streams. **Identification:** shrub 0.3–2 m (1–6.6 ft) tall with upright brown branches. ***Leaves*** alternate, simple, ***deciduous, wedge shaped,*** rounded, ***broadest near tip,*** toothed near tip, ***dark green above, pale beneath, downy on veins,*** 2–7 cm (0.8–2.8 in) long, ***aromatic*** when crushed. ***Male flowers in spikelike catkins at end of twigs; female flowers usually on separate plants,*** about 8 mm (0.3 in) long, ***at base of leaves. Fruit small nutlike, not waxy,*** resin dotted, aromatic, with 2 winglike scales, ***in small conelike clusters.*** **Harvest:** Pick leaves as described for SOUTHERN BAYBERRY (preceding) and aromatic nutlets from summer through winter. **Preparation:** Dry and use leaves as described for SOUTHERN BAYBERRY and for making hot tea. Use nutlets like dried leaves for seasoning. **Related edible species:** See SOUTHERN BAYBERRY. **Poisonous look-alikes:** none.

New Jersey tea *Ceanothus americanus*

Habitat: open woodlands and prairies, in well-drained rocky, gravelly, or sandy soils. **Identification:** ***shrub*** to 1 m (3.2 ft) with slender, reddish-brown, densely hairy twigs. ***Leaves*** alternate, simple, 3-ribbed, deciduous, ***4–8 cm*** (1.6–3.2 in) ***long,*** 2–4 cm (0.8 –1.6 in) wide, broadest near base, rounded at tip, ***finely toothed along margin,*** green above, gray and hairy beneath. ***Flowers tiny, white, produced in dense, short cylinder-shaped, branched clusters in spring,*** each with 5 white petals shaped like tiny tobacco pipes. Clusters from leaf junctions. ***Fruits*** produced in rounded clusters in late spring or early summer, each ***capsulelike, 5–6 mm*** (0.2 in) ***long,*** splitting open to release tiny, reddish-brown seeds. **Harvest:** Gather fresh leaves in summer, best when bushes are flowering. Dry leaves thoroughly in sun, by fire, or in warm oven. **Preparation:** For tea, steep 1 tbsp fresh or 1 tsp dried leaves in 1 cup boiling water for 15 min. Sweeten to taste. Caffeine free. Dried root bark tea was used as sedative by Indians. **Related edible species:** Red root *(C. ovatus)* is also known as New Jersey tea. *C. americanus* has wider leaves and flower clusters at leaf junctions instead of at end of twig. **Poisonous look-alikes:** none.

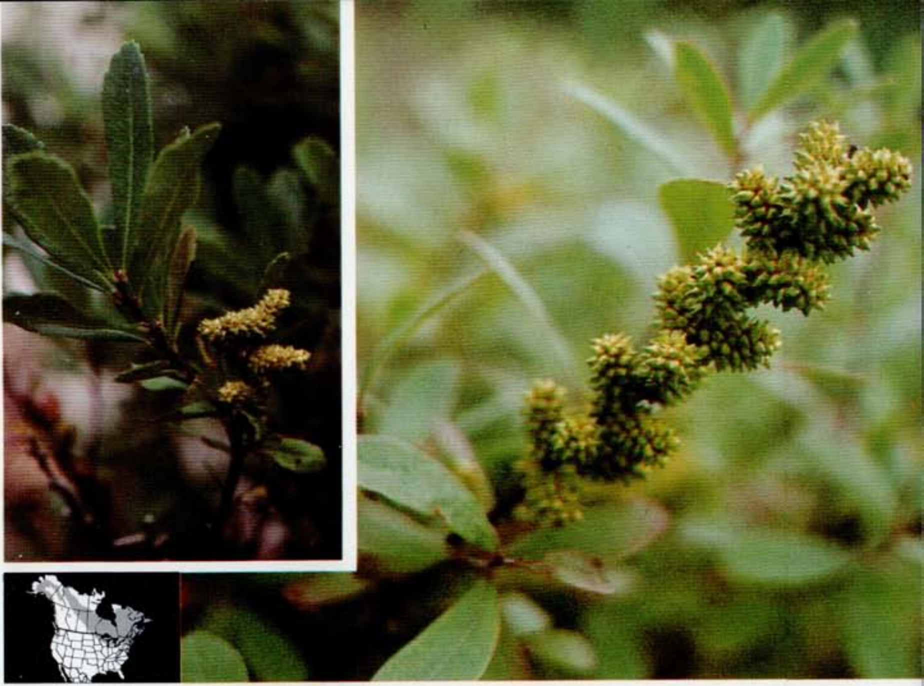

Sweet gale: fruiting branch and fruits (Elias).

New Jersey tea: flowers (Mohlenbrock). Flowering branch (NYBG).

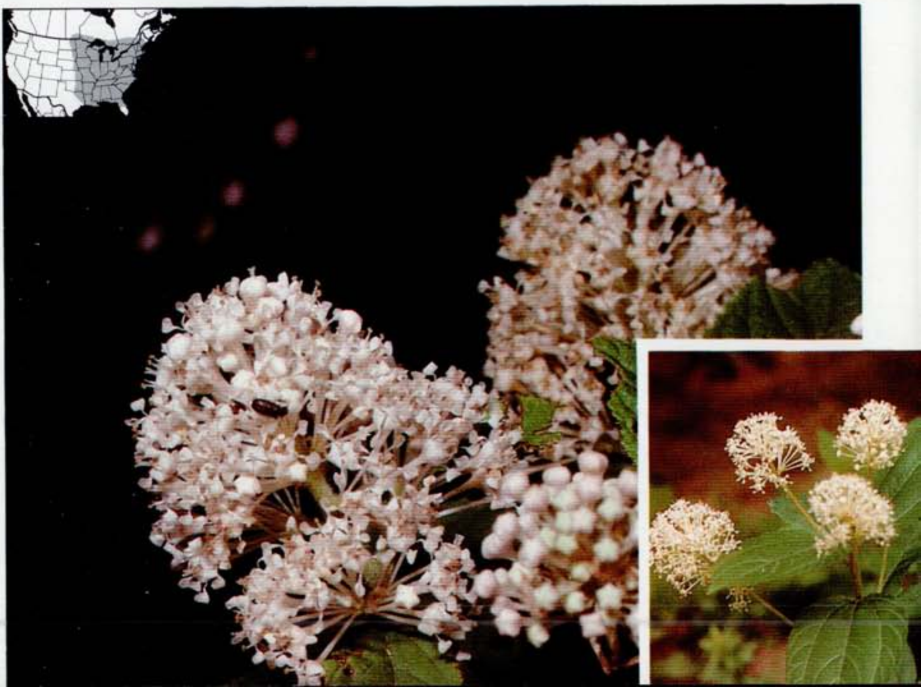

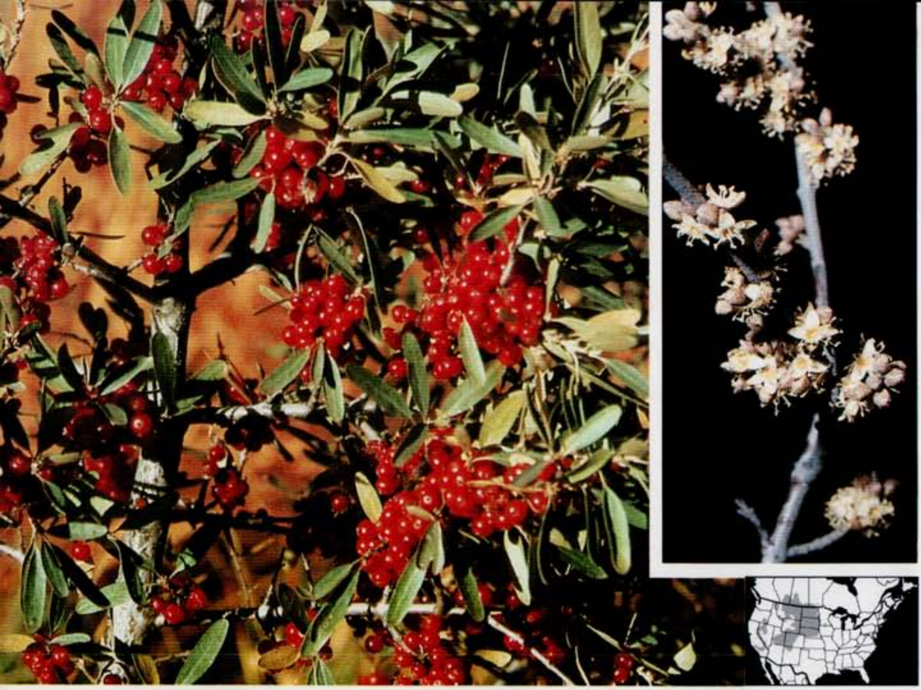

Silver buffaloberry: fruiting branches (Elias). Flowering branch (NYBG).

BRAMBLES, BLACKBERRIES, RASPBERRIES *Rubus*

This is a large and variable group of approximately 300 to 400 species in the temperate regions of the world. Members of this genus are shrubs or trailing, often prickly, vines, collectively known as brambles. Fruits may or may not separate from the spongy core of the fruit base when picked. The genus normally is divided into subcategories; one containing blackberries and dewberries, another containing thimbleberries, another raspberries. Most of the species have edible fruits; many have delicious fruits. Six of the more common species are treated on the following pages. **Harvest and preparation:** Pick fruit as it ripens. Berries of most species are excellent fresh, except for thimbleberries which have a rather dry taste and may require mixing with other berries. The berries are good with cream and sugar, in pancakes, on cereals, and in jams, jellies, and pies. To make berry bannock over campfire, mix 2 cups flour, 3 tsp double-acting baking powder, ½ tsp salt, 6 tbsp margarine or butter, and 4 tbsp powdered milk until uniformly mixed into mealy texture. Add 1 cup washed damp berries; mix gently to coat fruit. Add ⅓ cup water to make workable dough. Shape into 1-in-thick

Silver buffaloberry *Shepherdia argentea*

Also edible autumn

Habitat: riverbanks, low meadows, plains, canyons. **Identification:** spreading shrub or small tree to 5 m (16 ft) with dull gray bark shredding in long strips; ***branches*** stout, ***covered with silvery-white scales,*** usually ***thorny. Leaves opposite,*** simple, deciduous, 3–5 cm (1.2–2 in) long, 7–10 mm (0.3–0.4 in) wide, uniformly wide to widest near middle, entire, ***leathery, densely covered with silvery scales on both surfaces.*** Male and female flowers on separate trees; flowers small, inconspicuous, lacking petals. ***Fruits berrylike, rounded,*** 6–9 mm (0.2–0.4 in) long, smooth, ***bright red,*** slightly fleshy, enclosing single flattened seed. **Harvest:** late summer fruits pleasantly tart, better after frost has sweetened them. Shrub is frequently prolific fruit producer. Handpick or spread tarp beneath shrub and shake branches. **Preparation:** Use fruit fresh or dry thoroughly in sun. Use prefrost berries for jelly. Eat ripe fruit fresh or dried; add sugar as desired. Seeds are chewable. For beverage crush 1 cup fresh berries with 1 cup sugar in 3 cups hot water or mash 1 cup dried berries in 1 pt water. Add sugar to taste. Strain and drink. Cook berries with meat or prepare like CRANBERRY sauce for use with meat. For dessert mash fresh fruits and whip or blend into froth. Blend in sugar if needed. **Related edible species:** Canadian buffaloberry *(S. canadensis)* and silverberry *(Elaeagnus argentea)* are edible but less flavorful, and are improved by cooking and adding sugar. **Poisonous look-alikes:** none.

---

cake, dust with flour, place in warm, greased fry pan. Cook over moderate heat until crust forms on bottom. Turn over with spatula and cook until browned and no dough sticks to fork inserted into center of dough. Adapt other berry recipes from your favorite cookbooks. Tender growth tips of many species are edible, and leaves of many are good for tea. **Related edible species:** See above. **Poisonous look-alikes:** none.

Red raspberry *Rubus idaeus* var. *strigosus*

(synonym: *R. strigosus*)

Habitat: margins of woodlands, clearings, roadsides, abandoned fields. **Identification:** subshrub to 2 m (6.6 ft) high; ***branches*** spreading to trailing along ground, often with a whitish cast, ***armed with usually numerous prickles*** and stiff hairs. ***Leaves*** alternate, deciduous, compound on nonflowering canes, ***with 3–7 leaflets,*** each leaflet broadest near base, 2–10 cm (0.8–4 in) long, long-pointed at tip, coarsely and sharply toothed along margin, ***pale green to whitish and hairy beneath.*** Flowers white to greenish-white, 1.2–1.8 cm (0.5–0.7 in) across, produced June–July in small clusters of 2–5 flowers along upper part of canes. ***Fruits globe-shaped or nearly so, 1.2–2 cm*** (0.5–0.8 in) ***in diameter, bright red,*** juicy, ***separating easily from fruiting stalk and base,*** maturing from July–September. **Harvest and preparation:** See BRAMBLES (preceding).

Wineberry *Rubus phoenicolasius*

Habitat: introduced from Asia; grows wild in disturbed sites, along streams, rivers, roads, especially in coastal eastern U.S. **Identification:** shrub to 2.5 m (8 ft) high; ***branches*** spreading, often arching over, ***lacking spines but with dense, red, gland-tipped hairs. Leaves*** alternate, deciduous, ***3-parted*** (trifoliolate), ***broadest near base and often as broad as long, 4–10 cm (1.6–4 in) long and wide,*** often lobed, toothed. ***Flowers white,*** produced in elongate, many-flowered clusters; ***each 2–3 cm (0.8–1.2 in) across (including calyx lobes), petals shorter than the narrow, pointed, calyx lobes. Fruits globe-shaped, 0.8–1.2 cm (0.3–0.5 in) in diameter, bright red*** at maturity, juicy, maturing from midsummer to early autumn. **Harvest and preparation:** See BRAMBLES (preceding).

Thimbleberry *Rubus parviflorus*

Habitat: widespread from sea level to mountains; woodlands, canyons, open areas, along roads, trails, streams. **Identification:** shrub to 2 m (6.6 ft) high with ***upright to spreading branches, without spines*** but with glandular hairs. ***Leaves*** alternate, deciduous, ***simple, 5-lobed, 5–15 cm (2–6 in) long and wide,*** sharply toothed along margin. ***Flowers*** large, ***white to pink, 2–5 cm (0.8–2 in) across,*** produced in loose, 2–9-flowered, spreading clusters. ***Fruits large,*** globe-shaped, ***1.2–1.8 cm (0.5–0.7 in) across, bright red to scarlet,*** sweet, juicy, maturing throughout season. **Harvest and preparation:** See BRAMBLES (preceding).

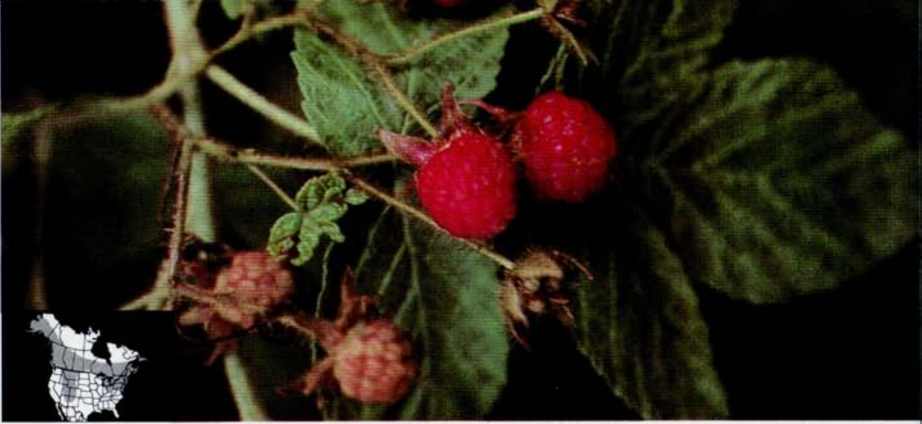

Red raspberry: fruiting branch (Elias).

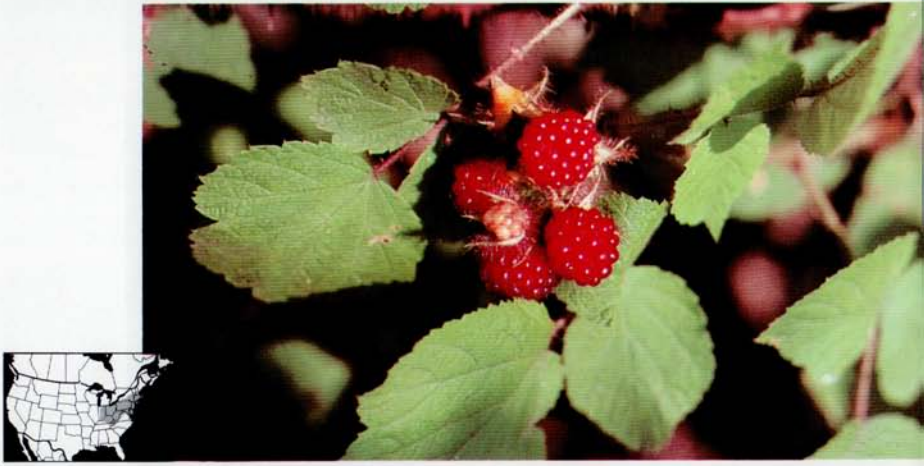

Wineberry: fruiting branch (Elias).

Thimbleberry: fruiting branch (McNeal).

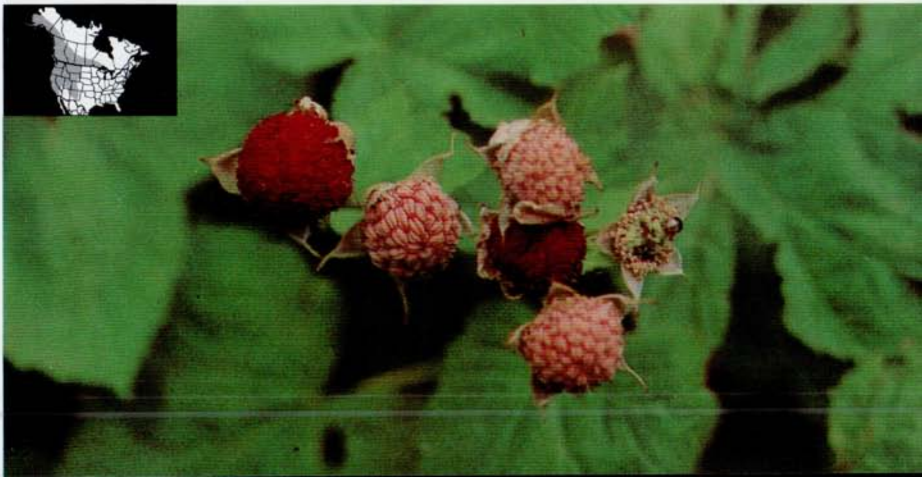

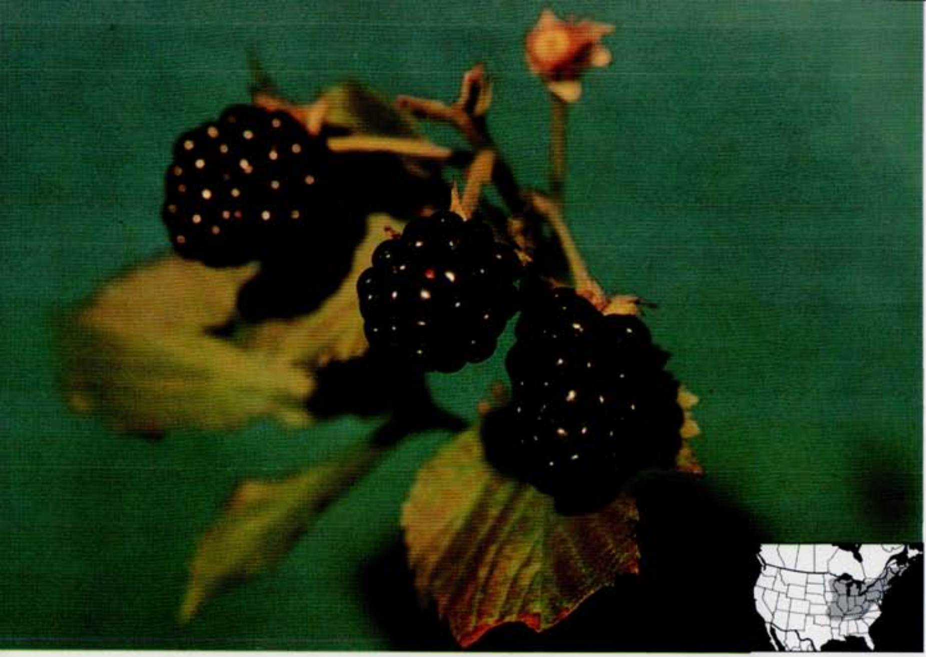

Common blackberry: mature fruits (Elias).

Black raspberry: fruiting branch (Elias). Flowering branch (Dykeman).

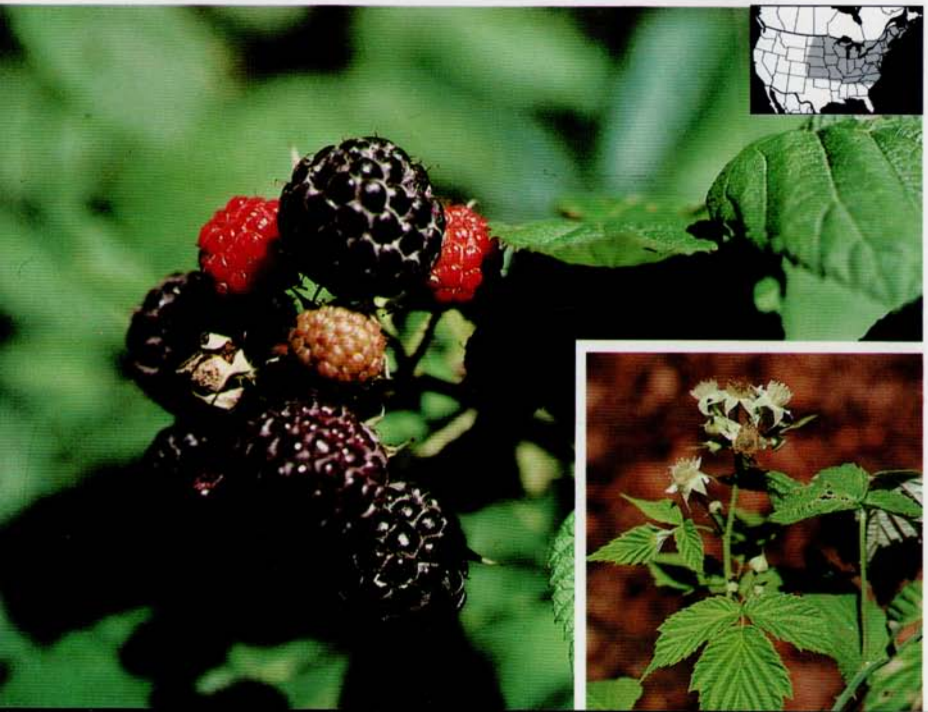

Common blackberry *Rubus allegheniensis*

Habitat: margins of woodlands, old fields, fencerows; also roadsides, thickets, mountain slopes. **Identification:** shrub, sometimes to 3 m (9.8 ft), branches upright to spreading, sometimes trailing, with scattered straight spines. ***Leaves*** alternate, deciduous, palmately compound, ***usually with 5 leaflets,*** sometimes 3, variable, typically broadest near or just below middle, long-pointed at tip, usually 10–18 cm (4–7 in) long, with numerous sharp-pointed teeth along margin, ***pale green beneath. Flowers white,*** 1.7–2.5 cm (0.7–1 in) across, produced in May, June, or early July (usually about 2–3 weeks later than black raspberry) ***in elongated 6–15-flowered clusters. Fruits globe- to short cylinder-shaped, to 2 cm (0.8 in) long,*** juicy, sweet, black at maturity; maturing in late July or August. **Harvest and preparation:** See BRAMBLES (preceding).

Black raspberry *Rubus occidentalis*

Habitat: disturbed areas, especially margins of woodlands, ravines, fields, thickets; sometimes in full shade. **Identification:** shrub with upright to spreading ***branches, sometimes arching to ground; stems with scattered, straight or curved, stout spines. Leaves*** alternate, deciduous, compound, each leaf ***with 3–5 leaflets,*** terminal leaflets broadest near base, pointed at tip, sharp and irregularly toothed along margin; lateral leaflets smaller, ***whitish beneath.*** Flowers white, produced in May or June in 3- to 7-flowered nearly flat clusters. ***Fruits purplish-black to black, 0.8–1.2 cm (0.3–0.5 in) across,*** rounded or nearly so, ***easily separated as a unit*** from fruiting stalk and base, juicy, sweet, maturing in June or July. **Harvest and preparation:** See BRAMBLES (preceding).

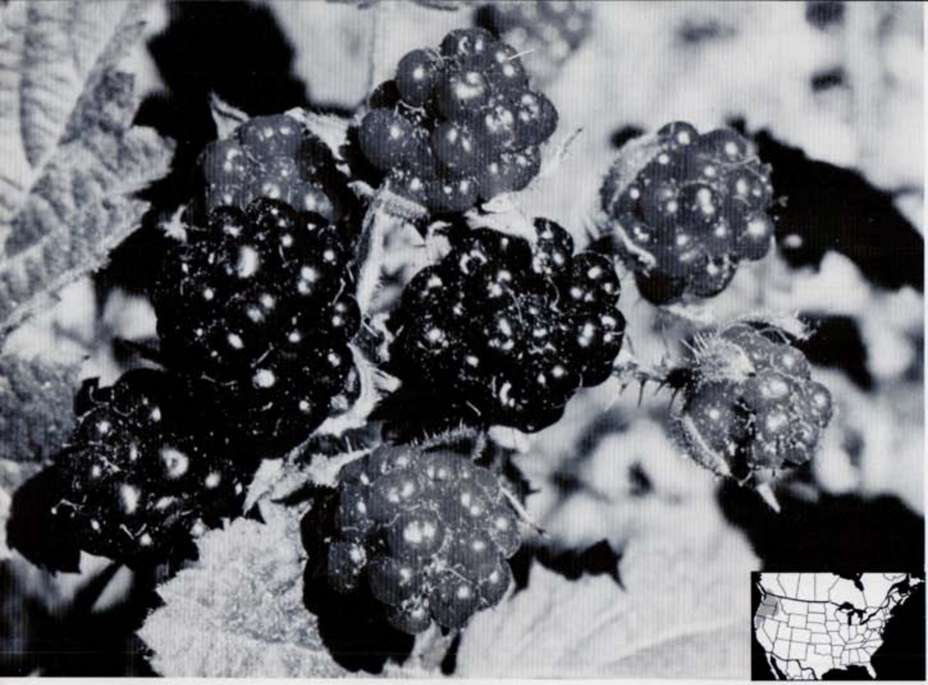

California blackberry: fruits (Ross).

Smooth sumac: fruiting branch (Dykeman). Fruits (Elias).

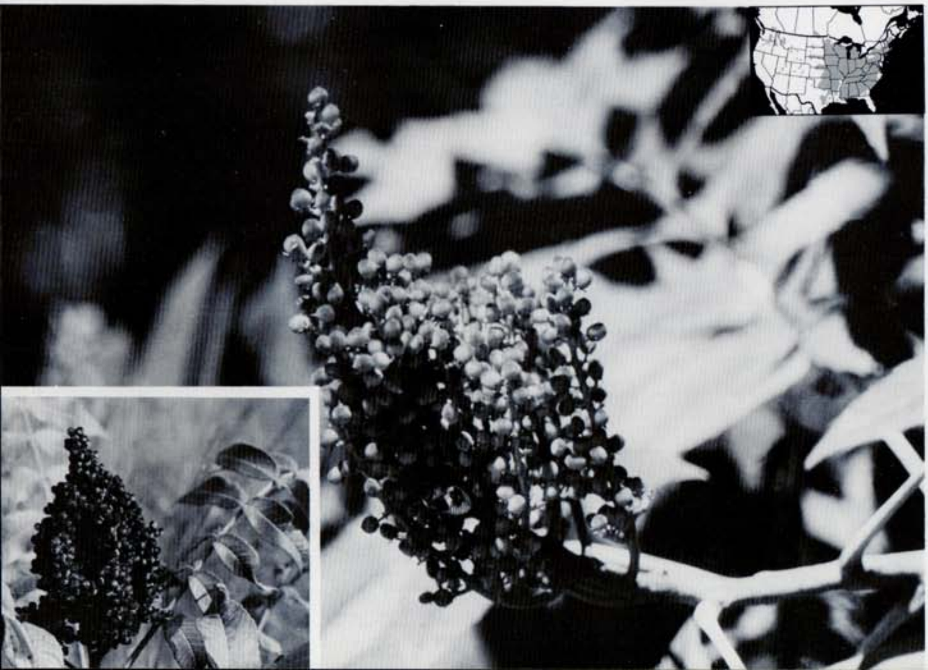

California blackberry *Rubus ursinus*

Habitat: disturbed sites, along trails, roads, canyons, fields, open woodlands; common at lower elevations, below 1,000 (3,200 ft). **Identification:** ***shrub, often low; branches spreading to trailing along ground;*** armed with slender, slightly flattened spines. ***Leaves*** alternate, deciduous, ***usually 3-parted (trifoliolate),*** sometimes simple; ***leaflets broadest near base, 6–12 cm (2.4–4.7 in) long, 4–10 cm (1.6–4 in) wide,*** coarsely toothed, occasionally shallowly lobed. Flowers white, produced in elongate, loose, 3–12-flowered clusters; each 1–3 cm (0.4–1.2 in) across. ***Fruits short cylinder-shaped to nearly cone-shaped, 1–2 cm (0.4–0.8 in) long, black at maturity,*** juicy, sweet. **Harvest and preparation:** See BRAMBLES (preceding).

Smooth sumac *Rhus glabra*

Also edible autumn

Habitat: common and widespread; abandoned farmlands, fields, margins of woodlands, along streams. **Identification:** shrub or small ***tree*** to 7 m (23 ft) ***with smooth stout branches*** and open rounded crown. ***Leaves*** alternate, deciduous, ***featherlike (pinnate) compound, with 11–31 leaflets,*** leaflets usually lance-shaped, 5–9.5 cm (2–3.7 in) long, 1.2–3 cm (0.5–1.2 in) wide, long pointed at tip, sharply toothed along margin, dark shiny green above, whitish beneath. Male and female flowers on different plants, produced in dense, cone-shaped clusters 10–25 cm (4–10 in) long. ***Fruits produced in autumn in large cone-shaped clusters, each fruit berrylike, rounded, 3–5 mm (0.1–0.2 in) in diameter, dark red, fuzzy.*** **Harvest:** Pick fruit clusters when glazelike bloom appears on bright red berries. Lick like lollipop for delicious lemony thirst quencher. Flavor is washed out by rains more quickly than in STAGHORN SUMAC. **Preparation:** same as STAGHORN SUMAC. Make delicious jelly from 2 cups sumac extract (see beverage under STAGHORN SUMAC) and 2 cups elderberry juice. Boil juice and extract with 1 pkg powdered pectin. Add 5 cups sugar. Bring to hard boil, boil 1 min, remove from heat, skim, pour into jars, seal. For other recipes, see SQUAWBUSH and STAGHORN SUMAC (next). See *caution* under STAGHORN SUMAC. **Related edible species:** STAGHORN SUMAC, SQUAWBUSH, plus lemonade sumac *(R. integrifolia)* of southern California and sugar bush *(R. ovata)* of the Southwest. **Poisonous look-alikes:** Poison sumac *(Toxicodendron vernix)* grows in wet areas and has white fruits. Avoid white fruited sumacs. POISON IVY and poison oak *(T. diversiloba)* are in the sumac genus.

Squawbush *Rhus trilobata*

Also edible autumn

Habitat: foothills, canyons, slopes, washes, often in dry rocky soils. **Identification:** ***shrub*** to 2 m (6.6 ft) with numerous spreading branches and ***densely hairy twigs, with strong odor when broken. Leaves*** alternate, deciduous, ***with 3 leaflets*** (trifoliolate), occasionally 1 leaflet, leaflets broadest near base or tip, 1–3 cm (0.4–1.2 in) long, nearly as wide, ***with large rounded teeth along margin,*** rounded at tip, hairy. Flowers yellow, small, densely clustered in short spikes in spring. ***Fruits berrylike, rounded, 4–6 mm (0.2 in) in diameter, red, hairy, sticky.*** **Harvest:** Gather fruits from end of August. **Preparation:** Suck on fresh fruits for refreshing sour taste. For beverage and jelly see SMOOTH SUMAC (preceding) and STAGHORN SUMAC (next). For syrup combine ½ cup dried, ground, sifted berries with 1⅓ cups water and ¾ cup honey (or 1 cup sugar). Cover, boil mix 30 sec, reduce heat, simmer 15 min. Uncover and boil until mix thickens, stirring constantly. Pour into jar and refrigerate for further thickening. Good dessert topping or spread on bread. If fresh berries are used, crush, simmer, and strain before adding honey or sugar. Then boil and complete process as above. See *caution* under STAGHORN SUMAC. **Related edible species:** See SMOOTH SUMAC. **Poisonous look-alikes:** Leaves resemble those of poison oak *(R. diversiloba)* and POISON IVY, both with white fruits. See also SMOOTH SUMAC.

Staghorn sumac *Rhus typhina*

Also edible autumn, winter

Habitat: woodland margins, fencerows, roadsides, old fields, streambanks. **Identification:** shrub or small ***tree*** to 10 m (33 ft) ***with thick densely hairy branches and twigs.*** Leaves alternate, deciduous, featherlike (pinnate) compound, 40–60 cm (16–24 in) long, with 11–31 leaflets, leaflets usually lance-shaped, 5–12 cm (2–4.7 in) long, 2–4 cm (0.8–1.6 in) wide, sharply toothed along margin. Male and female flowers on same or separate trees, produced in crowded cone-shaped clusters. ***Fruits produced in dense, cone-shaped clusters 10–20 cm (4–8 in) long; each fruit berrylike, rounded, 3–5 mm (0.1–0.2 in) in diameter, densely covered with dark red hairs.*** **Harvest:** Pick fruit clusters late summer to autumn when deep red and fruits have developed strong lemony taste. Dense hairs on fruits prevent rain from penetrating cluster and washing out all malic acid flavor so fruits are usable though less potent through winter. Dry some clusters for winter use. **Preparation:** For beverage separate fruits from twigs. Steep in near boiling water (no less than 1 cup fruits to 1 qt water) for at least 15 min. Strain through cloth, dilute, sweeten to taste. Drink hot or cold but in moderation; some people show allergic reaction. Using whole heads of fruits will include small arthropods within them, and twigs may add slight bitterness. See other uses under SMOOTH SUMAC and SQUAWBUSH (preceding). **Related edible species:** SMOOTH SUMAC and SQUAWBUSH, and lemonade berry *(R. integrifolia)*. **Poisonous look-alikes:** See SMOOTH SUMAC.

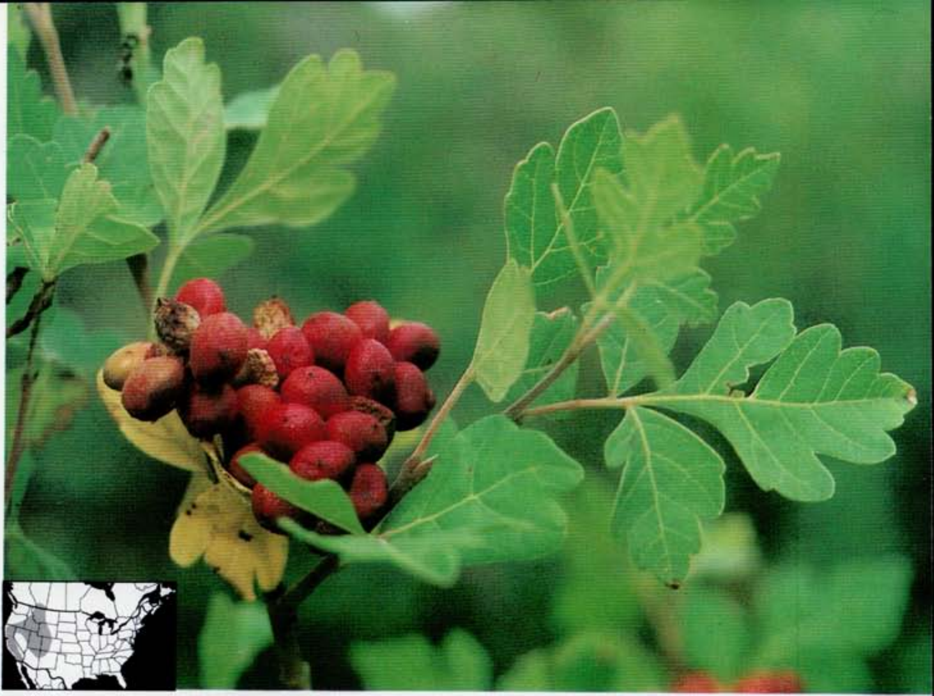

Squawbush: fruiting branch (Dykeman).

Staghorn sumac: fruiting branch, left (Dykeman). Fruits (Elias).

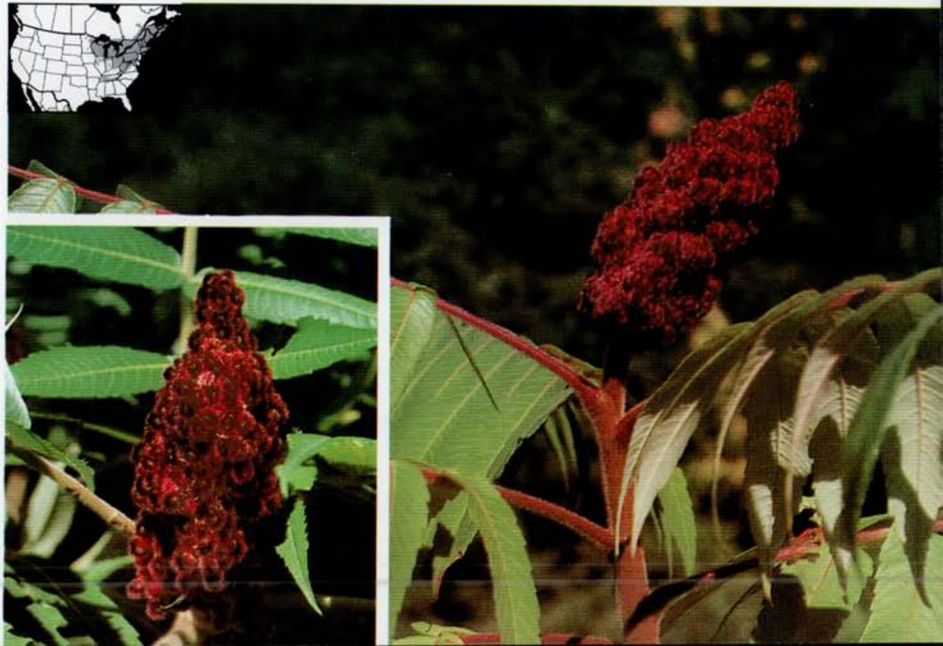

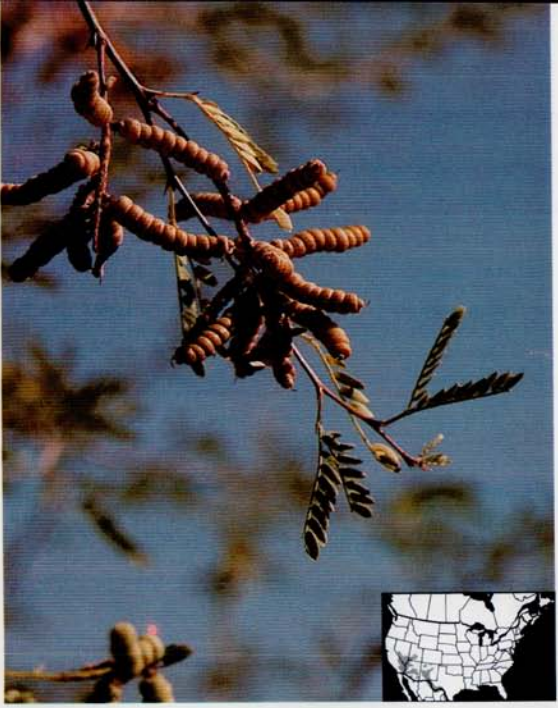

Screwbean mesquite: fruiting branch (Feldman).

Glandular mesquite: fruiting tree (Feldman). September branches (Hutchins).

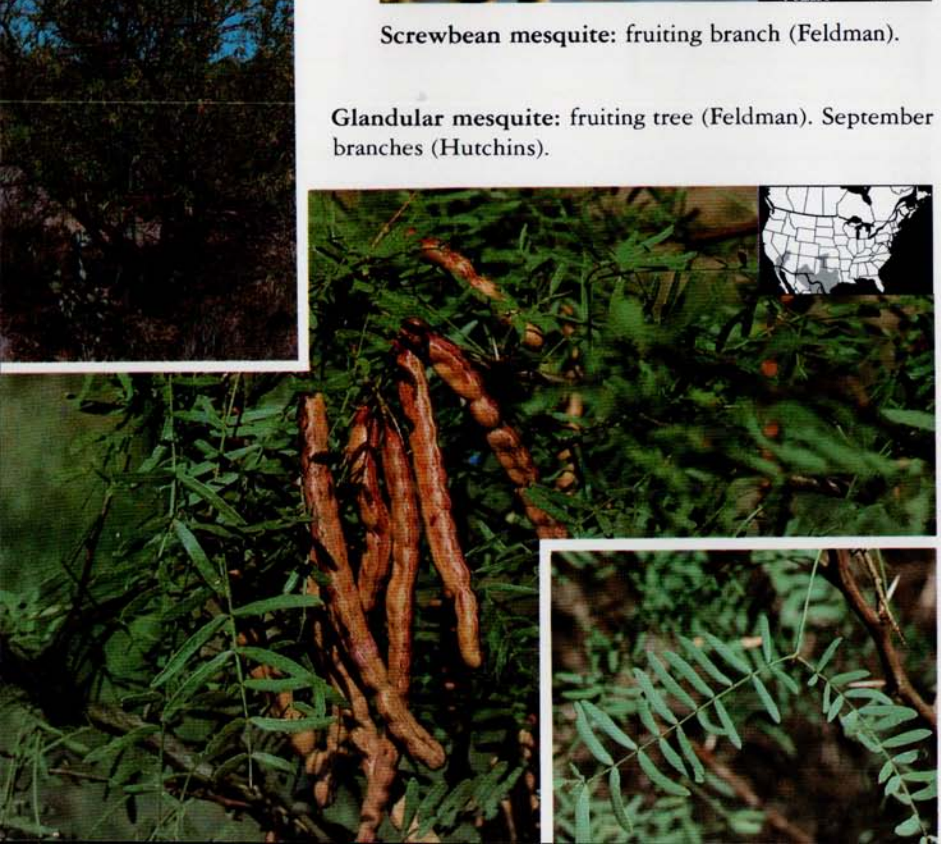

Screwbean mesquite *Prosopis pubescens*

Also edible autumn

Habitat: gullies, washes, stream banks, riverbottoms, floodplains. **Identification:** shrub or ***small spreading tree*** to 10 m (33 ft) with slender ***branches armed with pair of sharp spines at each leaf base. Leaves*** alternate, deciduous, ***twice compound and featherlike*** (bipinnate), with 1 or 2 pairs of side branches (pinnae), each 3.8–5 cm (1.5–2 in) long, composed of 5–8 pairs of small leaflets. Flowers produced in densely flowered cylinder-shaped spikes 5–7.5 cm (2–3 in) long, each flower greenish-white to yellow, 3–4 mm (0.2 in) long. ***Fruits*** are ***spiral twisted pods 3–5 cm*** (1.2–2 in) ***long,*** brown, hairy, becoming hard and tough at maturity. **Harvest:** Gather and clean immature pods, the younger the tenderer. **Preparation:** Boil until soft or cook with other vegetables. The pods can be boiled down to make a sweet syrup with many uses. **Related edible species:** GLANDULAR MESQUITE (next) and mesquite *(P. veluntina).* **Poisonous look-alikes:** slightly similar to CORALBEAN, which has poisonous seeds.

Glandular mesquite *Prosopis glandulosa*

Also edible autumn

Habitat: desert valleys, open range, grazing lands where plant becomes a pest. **Identification:** shrub or small spreading tree to 8 m (26 ft) with slender ***branches*** that ***have paired, sharp-pointed spines where each leaf is attached. Leaves*** alternate, deciduous, ***twice compound and featherlike*** (bipinnate), with 1 pair of side branches (pinnae), each 12–20 cm (4.7–8 in) long and composed of 7–18 pairs of leaflets. Flowers produced from May to August in densely flowered, narrow, cylinder-shaped clusters; each flower yellow, 3–4 mm (0.2 in) long. ***Fruits long, narrow, straight pods 10–25 cm*** (4–10 in) ***long,*** rounded, slightly constricted between seeds. **Harvest:** Gather flower clusters as they open and fruits as they ripen during summer. **Preparation:** Suck sugar-rich nectar from newly open flowers. The yellowish gelatinous to mealy pulp surrounding seeds in mature pods contains up to 30% sugar and can be eaten fresh. Make beverage by boiling clean pods in fresh water and straining off residue. Dry pods, grind, mix with little water to form dry, mealy dough; bake in sun for several hours or over very low heat. Soak seeds in water several hours for lemon-flavored drink or grind dry seeds into meal and use to make porridge. Discard those seeds infested with larvae of bean beetle. **Related edible species:** SCREWBEAN MESQUITE (preceding) and mesquite *(P. veluntina).* **Poisonous look-alikes:** may resemble CORALBEAN, which has toxic seeds.

American elder, Common elder *(Also edible spring)* *Sambucus canadensis*

Habitat: rich moist soil of woodland margins, stream and river banks, fencerows and right-of-ways. **Identification:** mainly shrubby, 1–4 m (3–13 ft), with upright to spreading branches and broad rounded crown. ***Leaves opposite,*** deciduous, featherlike (pinnate) ***compound,*** 15–25 cm (6–10 in) long, composed of ***5–11 leaflets,*** each leaflet broadest near base to uniformly wide, 6–12 cm (2.4–4.7 in) long, 2–4 cm (0.8–1.6 in) wide, pointed, toothed, dark shiny green. ***Twigs*** with prominent ***white pith. Flowers*** white, small, in branched, ***round or flat-topped showy clusters*** in late spring or early summer. ***Fruits berrylike,*** rounded, 4–6 mm (0.2 in), ***deep purple,*** juicy, 3-seeded, in large clusters, maturing in mid to late summer. **Harvest:** When flowers are fully open, pick whole clusters or shake petals of old flowers into container. When fruit is deep purple, pick whole clusters or strip berries from bush. **Preparation:** Dip flower cluster in fritter batter and deep fry; dust with sugar and sprinkle with orange juice. Add cup of blossoms to pancake batter. Extract juice from fresh fruit by simmering mix of 1 cup water per 1 qt mashed berries 20–25 min; strain through cloth. Sweetened juice good with tart juices. Use juice with (STAGHORN SUMAC) extract in jelly. Fresh fruit has rank flavor; improves with drying; use dried fruit for pies and breads. Dry fruit in sun or warm attic; before using stew in little water. Flowers and fruit excellent for wine. Fruits high in vitamin C. *Warning:* Roots, stems, leaves, unripe fruits are somewhat toxic. **Related edible species:** BLUE ELDER (next). **Poisonous look-alikes:** See BLUE ELDER.

Blue elder *(Also edible spring)* *Sambucus cerulea*

Habitat: stream banks, gullies, washes, margins of fields, woodlands, pastures. **Identification:** shrub or small tree with broad rounded crown. Leaves ***opposite,*** deciduous, featherlike (pinnate) ***compound,*** 12–18 cm (4.7–7.1 in) long, composed of ***5–9 leaflets,*** leaflets broadest at middle to uniformly wide, 5–15 cm (2–6 in) long, 1.2–5 cm (0.5–2 in) wide, long-pointed at tip, coarsely-toothed, bright green above. ***Flowers*** tiny, yellowish-white, produced in spreading, ***many-flowered, flat-topped clusters*** in late spring or early summer. ***Fruits*** in clusters, each ***berrylike,*** rounded, 6–10 mm (0.2–0.4 in), ***dark blue to nearly black,*** juicy, maturing in mid to late summer. **Harvest:** Flowers in early summer, fruits in late summer, can be eaten raw or dried like raisins. **Preparation:** Prepare tea from blossoms. For jelly, wash berries, partially crush, cover with water, cook over low heat 10 min. Pour contents into wet jelly bag and collect juice; do not squeeze bag. Add 2 cups elderberry juice to 3 cups apple juice prepared in similar manner. Heat to boil, add 1 pkg pectin; stir until dissolved. Bring to rapid boil, add 6 cups sugar. Stir until rolling boil cannot be stirred down; continue for 1 min. Pour mix into prepared jelly jars; cover with ¼ in paraffin; seal with lid. For other uses see AMERICAN ELDER (preceding). **Related edible species:** AMERICAN ELDER, Mexican elder *(S. mexicana),* and black berried elder *(S. melanocarpa).* **Poisonous look-alikes:** Red fruited elder *(S. pubens)* of East and Midwest somewhat toxic. Pacific elder *(S. callicarpa)* bitter but not poisonous.

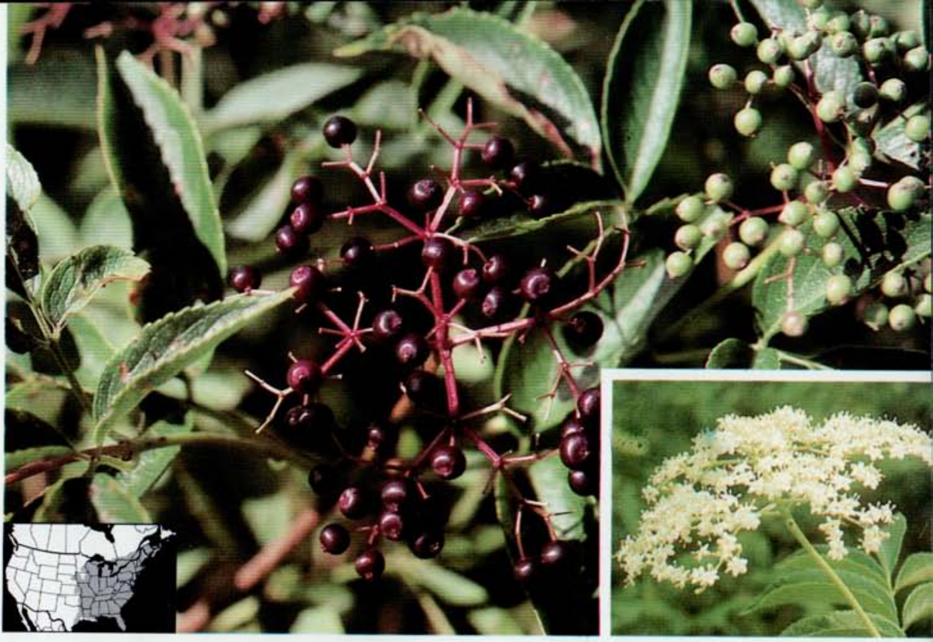

American elder: fruiting branches (Elias). Flowers (NYBG).

Blue elder: fruiting bush and fruits (Kavasch).

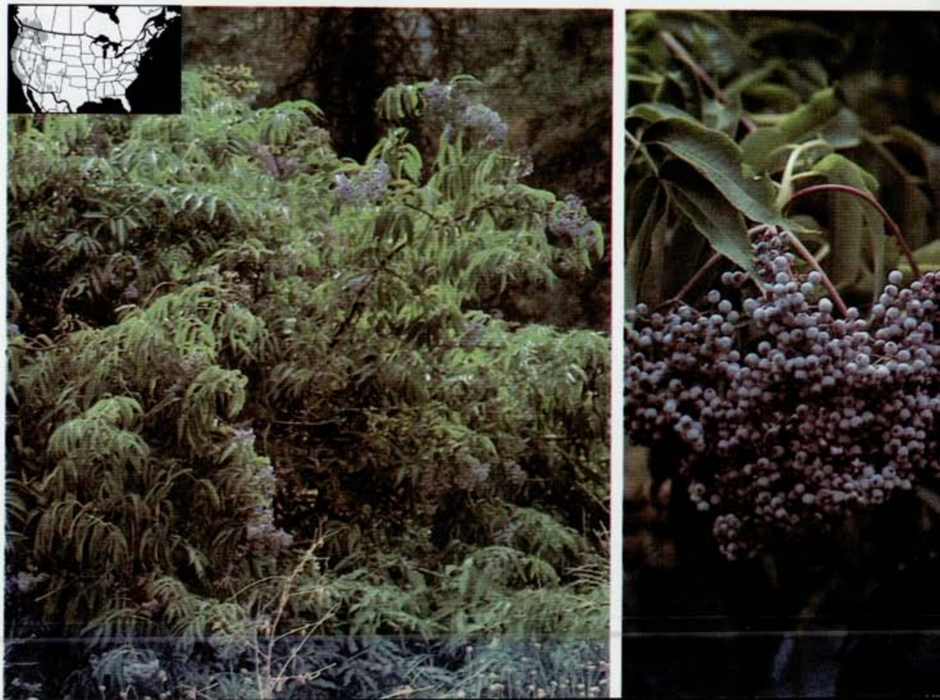

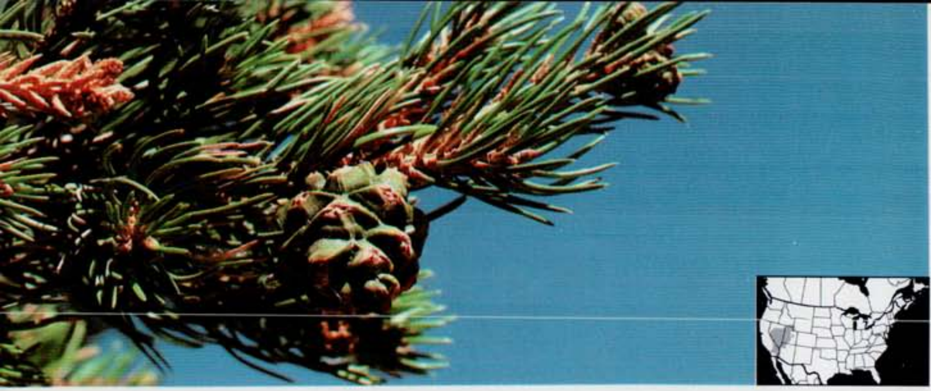

Singleleaf pinyon pine: branch with immature cone (Dykeman).

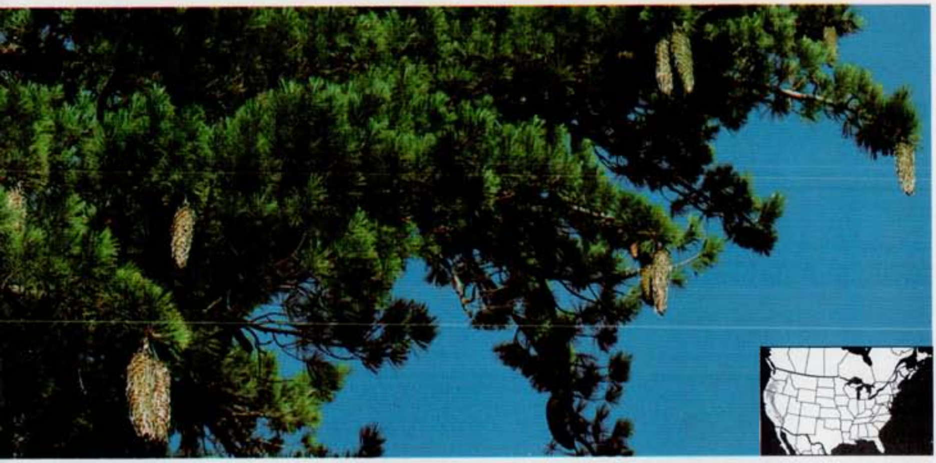

Sugar pine: branches with cones (McNeal).

Digger pine: leader and cones (McNeal).

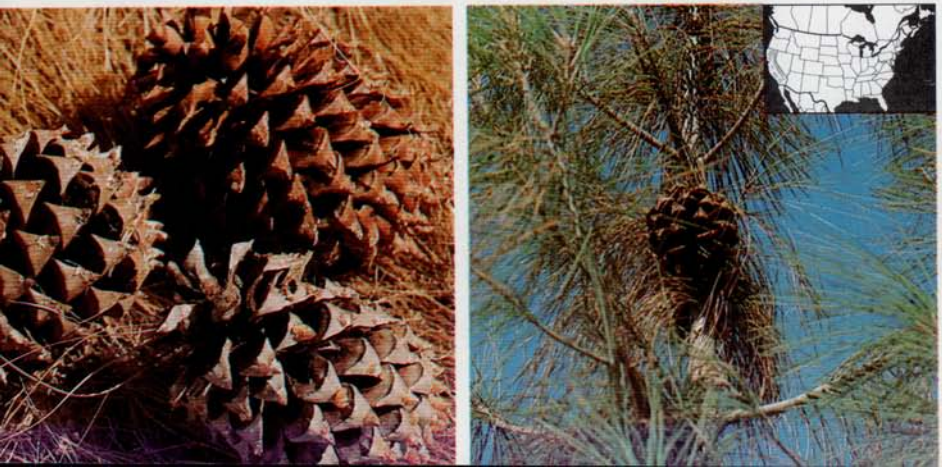

Singleleaf pinyon pine *Pinus monophylla*

Habitat: lower mountain slopes, foothills, canyons from 650–2,350 m (2,100–7,800 ft) elevation; often growing with junipers. **Identification:** small, spreading tree with rough, irregularly furrowed, dark brown bark. ***Leaves (needles) 1 per sheath, 2.5–5 cm (1–2 in) long, sharp-pointed, stiff, often curved toward branches,*** yellowish-green to gray-green; only native North American pine with single leaves. Fruits hard, egg-shaped cones 5–7.6 cm (2–3 in) long, composed of thick, stiff, light brown, blunt scales, each scale bearing 2 large, ***egg-shaped, brown, wingless seeds 1.8–2 cm*** (0.7–0.8 in) ***long.*** Trees produce moderate to large cone crops every 2–3 yr. Each cone requires 2 growing seasons to mature. **Harvest and preparation:** Gather second year cones in August or September. Oily seeds, rich in protein and oils, have nutty taste and can be eaten fresh. Thin shell can be cracked with teeth. Roast nearly ripe cones to facilitate seed removal and improve flavor. Store cones or seeds for later use, especially after roasting. Several cones tossed into campfire coals 10–20 min yield an excellent snack. Carry roasted cones or seeds in backpack for trail snack. If large numbers of seeds are available, roast and grind into meal and use in breads, soups, or mixed with cornmeal or flour. **Related edible species:** SUGAR PINE, DIGGER PINE, PINYON PINE, Mexican pinyon pine *(P. cembroides)*. **Poisonous look-alikes:** none.

Sugar pine *Pinus lambertiana*

Habitat: mountain slopes from 750–3,000 m (2,500–9,850 ft) elevation. **Identification:** tall trees to 65 m (213 ft) with tall, straight trunk and thick, deeply, irregularly furrowed, dark, reddish-brown to purplish-brown bark. ***Leaves*** (needles) ***5 per sheath, 8.8–10 cm*** (3.5–4 cm) ***long, stout, stiff, deep bluish-green and marked with silvery lines. Fruits large woody cones, 27–46 cm*** (10.6–18 in) ***long,*** to 12 cm (4.7 in) in diameter; seeds nearly egg-shaped to football-shaped, 1.2–1.5 cm (0.5–0.6 in) long, dark chocolate-brown, nearly black, winged. **Harvest:** See SINGLELEAF PINYON PINE (preceding). *Note:* Cones with seeds are difficult to obtain because of height of the trees. **Preparation:** same as for SINGLELEAF PINYON PINE. **Related edible species:** Coulter pine *(P. coulteri)*, DIGGER PINE (next), and the PINYON PINES. **Poisonous look-alikes:** none.

Digger pine *Pinus sabiniana*

Habitat: foothills, lower mountain slopes from 20–1,000 m (66–3,300 ft) elevation, often found with BLUE OAK. **Identification:** medium-size spreading tree to 15 m (50 ft) with short, thick trunk, soon branching, with irregular, scaly, dark brown to almost black bark. ***Leaves*** (needles) ***3 per sheath, 17–43 cm*** (6.7–17 in) ***long, slender, flexible, dull blue-green. Fruits large 15–25 cm*** (6–10 in) ***long,*** to 18 cm (7 in) broad, light reddish-brown, ***scales thick, each armed with clawlike spine;*** seeds 1.7–2 cm (0.7–0.8 in) long, egg-shaped, with hard, dark brown to almost black shell. **Harvest and preparation:** same as for SINGLELEAF PINYON PINE. **Related edible species:** PINYON PINES, Coulter pine *(P. coulteri)*, SUGAR PINE (preceding). **Poisonous look-alikes:** none.

JUNIPER GENUS *Juniperus*

The 13 species of junipers native to North America are mostly common widespread trees and shrubs. These evergreen conifers are more abundant in the drier western states, although eastern redcedar and southern redcedar are prevalent in the eastern U.S. All junipers provide excellent cover for small game and food for many birds. Only three species, all western, are recommended for their edible fruits and then primarily as a survival food. The stems, leaves, and fruits contain varying amounts of resins which affect fragrance and taste.

Rocky Mountain juniper *Juniperus scopulorum*

Also edible autumn, winter

Habitat: bluffs, ridges, cliffs, dry rocky hillsides from sea level to 3,000 m (9,850 ft) elevation. **Identification:** large evergreen shrub or small tree to 15 m (49 ft) with reddish-brown to gray scaly bark separating in long narrow strips. ***Leaves tiny, scalelike, opposite each other,*** closely pressed against branchlets, 2–4 mm (0.1 in) long with tiny gland on back, ***pale green to grayish green,*** resinous. Male and female flowers on separate trees in spring, inconspicuous and easily overlooked. ***Fruits small, fleshy, sweet, berrylike cones, nearly rounded, 6–9 mm*** (0.2–4 in) ***in diameter, bright blue, maturing in 2 seasons,*** containing 1–2 seeds. **Harvest:** Gather ripe berries from summer to early winter. Use fresh or dry and store for later use. **Preparation:** Although high in sugars, ripe fruits have an unpleasant resinous taste, so fresh fruits become strictly survival fare. Use a few fresh berries as substitute for sage in flavoring meats. Roast dried fruits, grind, and use as coffee substitute. **Related edible species:** UTAH JUNIPER and CALIFORNIA JUNIPER (next). **Poisonous look-alikes:** Take care in identification of junipers. Fruits of most species have such high resin content that they are unpalatable or may cause stomach upset.

California juniper *Juniperus californica*

Also edible autumn, winter

Habitat: dry slopes, canyons, desert slopes from 125–1,350 m (410–4,500 ft) elevation. **Identification:** large shrub or small tree with fluted and ridged trunk and reddish-brown to ash-gray bark that splits into long, loose, shredding strips. ***Leaves tiny, scalelike, arranged in 3s,*** overlapping and closely pressed against branchlets, 2–4 mm (0.1 in) long, with tiny gland on back, light yellowish-green. Male and female flowers on separate trees in spring, inconspicuous and easily overlooked. ***Fruits*** are ***leathery, berrylike cones, almost rounded, 10–16 mm*** (0.4–0.6 in) ***in diameter, dry, sweet, not resinous, light reddish-brown,*** maturing in 2 seasons, each containing 1–3 seeds. **Harvest and preparation:** same as for ROCKY MOUNTAIN JUNIPER (preceding) except that nonresinous fruits are superior to those of other junipers as wild food source. **Related edible species:** UTAH JUNIPER (next), ROCKY MOUNTAIN JUNIPER. **Poisonous look-alikes:** same as for ROCKY MOUNTAIN JUNIPER.

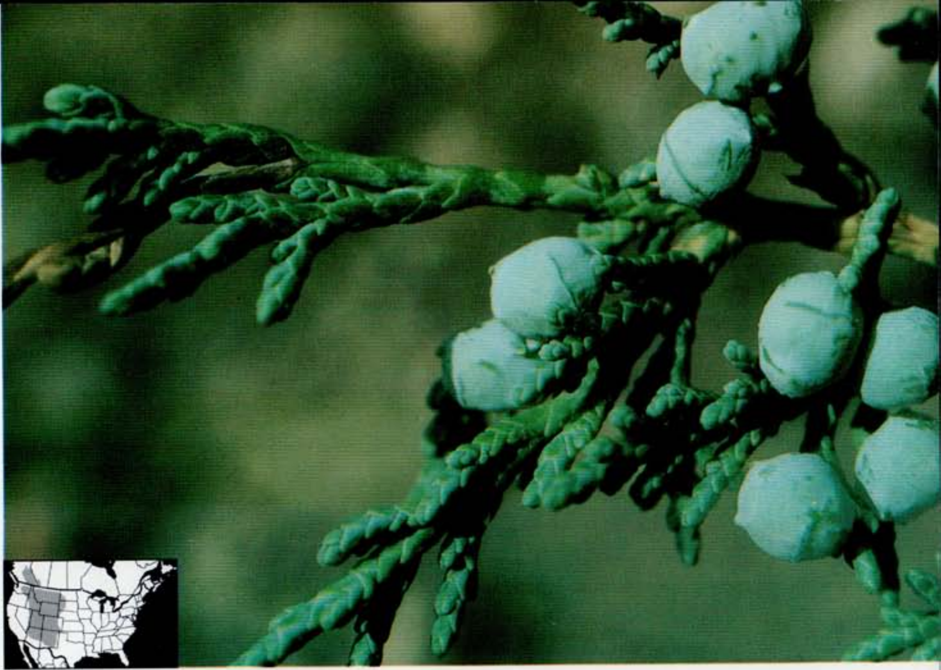

Rocky mountain juniper: fruiting branch (Dykeman).

California juniper: fruiting branch (Webber, Jepson Herbarium).

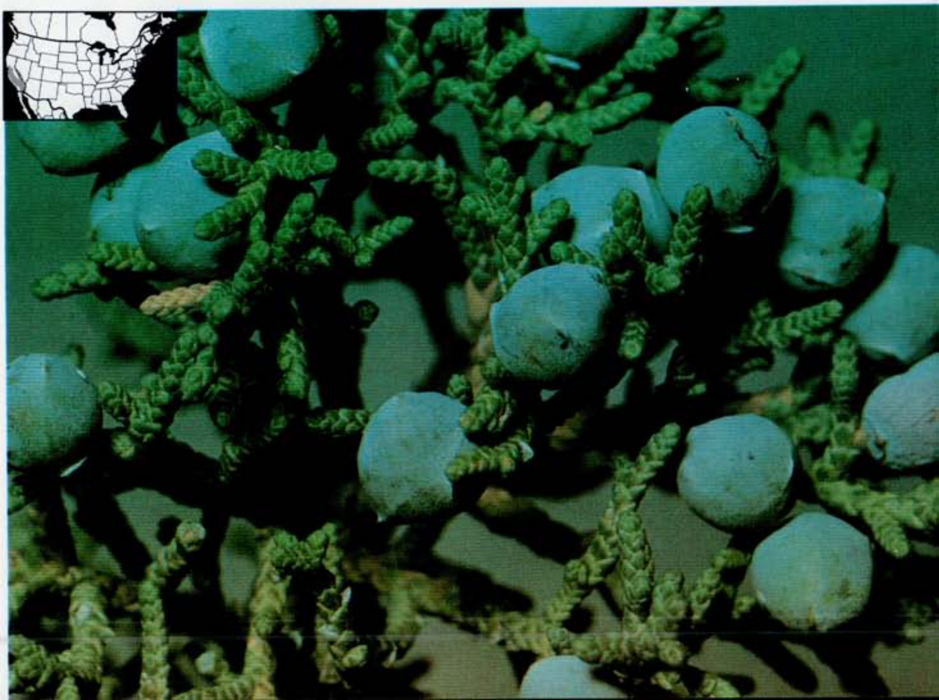

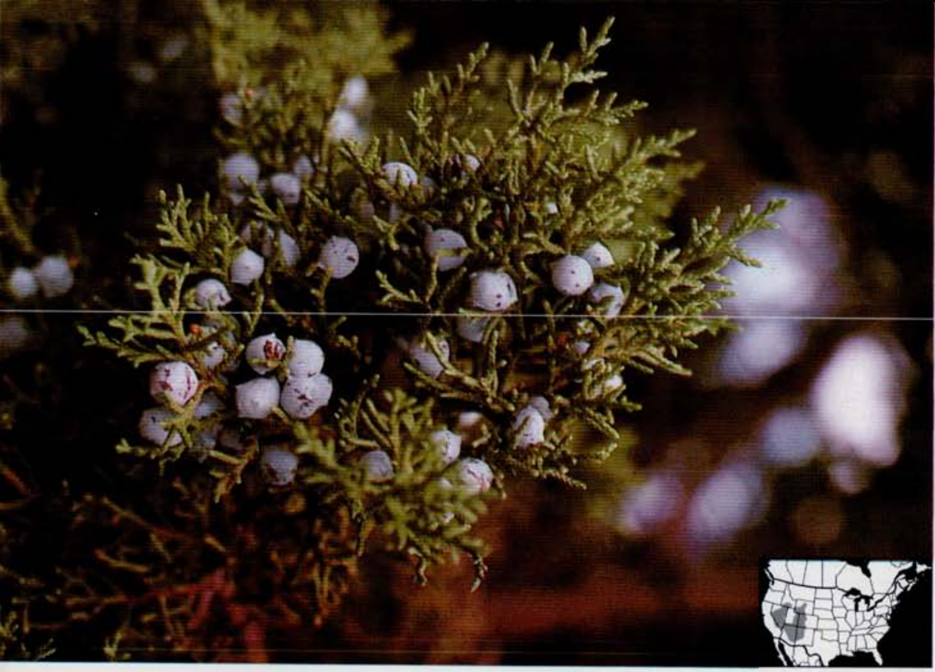

Utah juniper: fruiting branch (Pawloski).

Sea grape: fruiting and flowering branches (Elias).

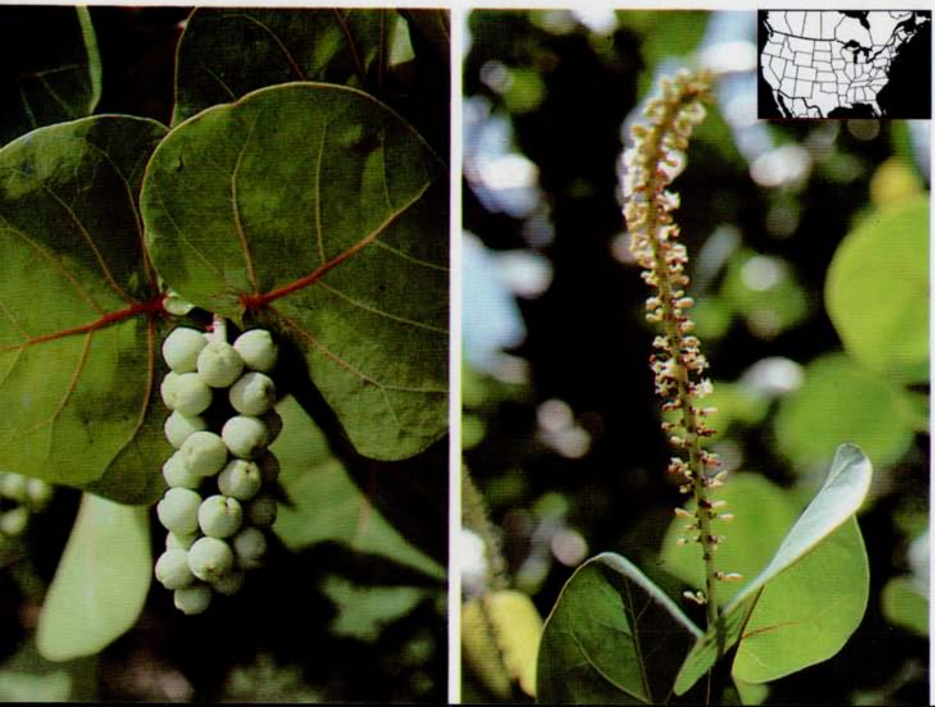

Utah juniper *Juniperus osteosperma*
Also edible autumn, winter (synonym: *J. utahensis*)

Habitat: mesas, mountain slopes, high plains, principally from 1,000–2,650 m (3,300–8,700 ft) elevation in dry rocky or gravelly soils. **Identification:** large ***shrub*** or small ***bushy tree*** to 8 m (26 ft) high with ***furrowed ash-gray bark that shreds in long, loosely attached strips. Leaves*** tiny, ***scalelike, opposite and overlapping one another in 2s, closely pressed against branchlets,*** 3–4 mm (0.1–0.2 in) long, usually gland-dotted on back, ***sharp-pointed at tip, yellowish-green.*** Male and female flowering cones usually produced in spring on separate trees, inconspicuous and usually overlooked. ***Fruits small, leathery berries 6–9 mm*** (0.2–0.4 in) ***in diameter, nearly globe-shaped, reddish-brown, resinous, sweet,*** with mealy texture, containing 1, rarely 2, seeds. **Harvest and preparation:** same as for ROCKY MOUNTAIN JUNIPER. **Related edible species:** CALIFORNIA JUNIPER and ROCKY MOUNTAIN JUNIPER (preceding). **Poisonous look-alikes:** same as for ROCKY MOUNTAIN JUNIPER.

Sea grape *Coccoloba uvifera*

Habitat: beaches, along sea shores; Southern Florida, Bermuda, Bahama Islands, the West Indies. **Identification:** very distinctive large shrub or tree to 15 m (50 ft) with compact rounded crown, trunk short, covered with pale blotchy bark. ***Leaves*** alternate, simple, ***persistent and appearing evergreen, 10–27 cm*** (4–10.6 in) ***long, 12–25 cm*** (4.7–10 in) ***wide, almost circular in outline,*** rounded at tip, entire, heart-shaped at base, ***thick, leathery,*** dark shiny green above. Flowering clusters produced along narrow, elongate stalks to 30 cm (12 in) long. ***Fruits crowded together in long, hanging clusters,*** each fruit nearly rounded, 1.8–2.5 cm (0.7–1 in) long, greenish-white to purple, with thin, juicy flesh; seed hard, nutlike. **Harvest:** Since fruits do not ripen uniformly, spread sheet under bush and shake branches vigorously. Ripe fruits will fall while green ones persist in cluster. **Preparation:** best used for jelly. Partially crush 2 qt clean ripe fruits with ½ cup water and cook for 20 min. Strain juice through jelly bag. Boil juice with equal amount of sugar until jelling point is reached. Skim off foam, pour into sterilized jars, seal with paraffin. **Related edible species:** dove plum *(C. diversifolia)* of coastal southeastern Florida, but not a recommended substitute for sea grape. **Poisonous look-alikes:** none.

California laurel, California bay *Umbellularia californica*

Also edible autumn, spring

Habitat: lower mountain slopes, hillsides, flatlands, in variety of soils; Pacific coastal regions and western slopes of Sierra Nevada. **Identification:** small to medium-size tree to 20 m (66 ft) with thin, dark brown bark. ***Leaves*** alternate, simple, ***evergreen,*** 8.6–13 cm (3.4–5.1 in) long, lance-shaped to broadest near middle, usually pointed at tip, ***entire, dark green and shiny, leathery, with spicy odor when crushed.*** Flowering clusters flat-topped, containing 4–9 flowers, each small, 1–1.4 cm (0.4–0.6 in) across, yellow. ***Fruits resembling an olive,*** globe-shaped to broadest near base, ***2–2.6 cm*** (0.8–1 in) ***long,*** slightly fleshy, green turning yellowish-green or purplish-green at maturity, containing single large seed. **Harvest:** Pick young unblemished leaves, wash and leave in sun until thoroughly dry. Store in cool dry place for later use. Pick ripe fruits in late summer or autumn, remove and clean seeds. **Preparation:** Use dried leaves as substitute for bay leaves. California laurel is more potent, so cut recipe amount by ⅓ unless stronger flavoring is desired. Good in stews, chili, spaghetti sauce, soups. Roast seeds in campfire, split open, and eat or grind into flour for bread and other recipes. Indians of California used leaves in lodges to repel fleas. **Related edible species:** none. **Poisonous look-alikes:** none.

Pawpaw, Papaw *Asimina triloba*

Also edible autumn

Habitat: river valleys and bottomlands, usually in deep, rich, moist soils; grows in association with sweetgum, swamp chestnut, black gum, cherrybark oak. **Identification:** large shrub or small tree with thin, smooth bark (dark brown with gray blotches) 3–12 m (10–40 ft) tall. Leaves alternate, deciduous, simple, 18–30 cm (7–12 in) long, 8–14 cm (3.2–5.5 in) wide, usually broadest near tip, entire along margin, papery textured, light green. Flowers ***green turning deep purple, produced in spring along stems of previous year, bell-shaped,*** with 3 sepals, 6 thick, almost rounded petals 2–2.6 cm (0.8–1 in) long. Fruits irregular, cylinder-shaped, 4–10 cm (1.6–4 in) long, aromatic, green turning yellow then black and fleshy at maturity; mature in late summer or early autumn. **Harvest:** when fruits mature or while still green but full size; store in cool place (safe from small animals) until ripe. **Preparation:** Tasty sweet pulp of fruit can be eaten raw. Cook creamy, custardlike pulp, minus seeds and skin, and use in puddings, breads, ice cream. *Caution:* Fruits disagree with some people. **Related edible species:** dwarf pawpaw *(A. parviflora)*. **Poisonous look-alikes:** none.

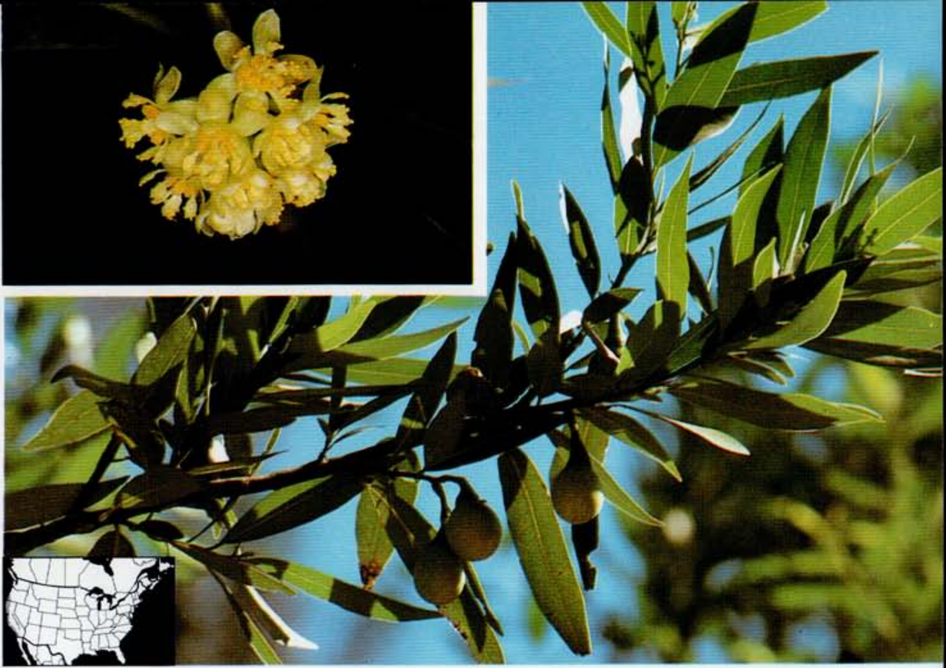

California laurel: flower and fruiting branch (U. of Cal., Davis).

Pawpaw: flowering branch (Folkerts). Fruiting branch (Elias).

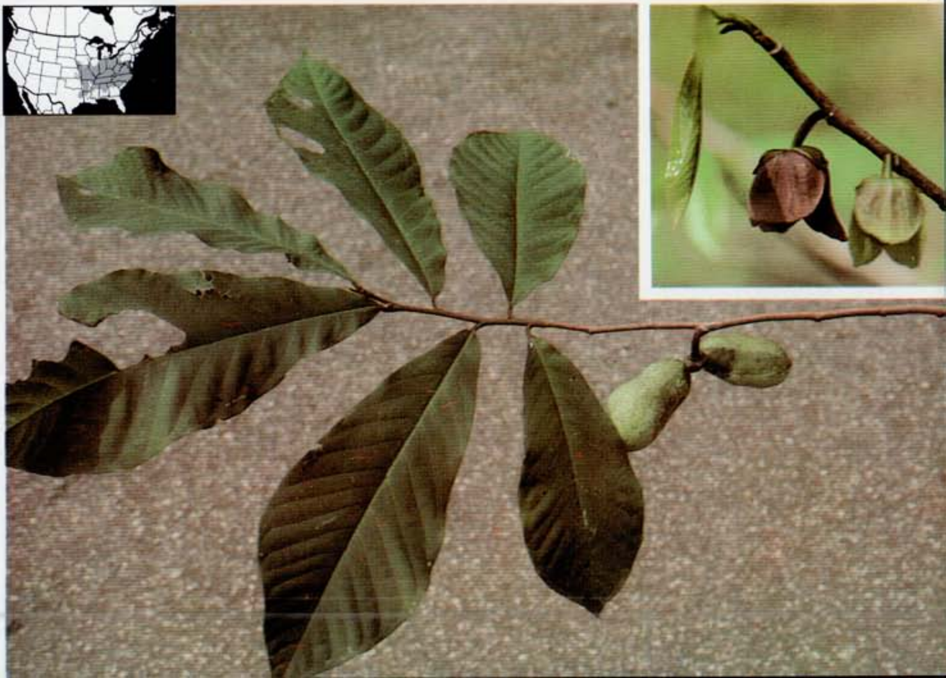

White mulberry *(Also edible spring)* *Morus alba*

Habitat: introduced and cultivated tree that now grows wild along fencerows, roadsides, margins of abandoned fields, and in woodlands. **Identification:** medium-size tree with widespreading, rounded crown and thick, irregularly furrowed, light to dark gray to yellowish-brown bark and ***milky sap. Leaves*** alternate, deciduous, simple, *6–16 cm (2.4–6.3 in)* ***long, 3–8 cm (1.2–3.2 in) wide, broadest near base,*** pointed at tip, ***unlobed or 2–3 lobed, sharply toothed along margin.*** Male flowers tiny, clustered on short spikes; female flowers tiny, clustered in round heads, both appearing in spring on new growth. ***Fruits globe-shaped to short cylindrical, 1–2 cm (0.4–0.8 in) long, blackberrylike, fleshy, white to pinkish at maturity*** (June–July). **Harvest and preparation:** Pick fruits as they ripen from late May through July. Fully ripe berries may be shaken from tree onto plastic sheets but tend to be very syrupy and too sweet for most palates. Dry like raisins for breads, cookies, and puddings. Use plastic or aluminum screen supported so air can reach all sides of fruit. If solar drying, protect from dew condensation at night. Dry until crushed fruit lacks any watery juice. Drying time: bright sun, 2 days; warm attic, 4 days; warm oven, 10–15 hr. Store dried berries in sealed containers. Afghans grind dried fruit with almonds for nutritious combination. Substitute any other type of nut, roll small balls of ground mix in confectioner's sugar, eat as candy. Pick shoots and prepare as RED MULBERRY (next). **Related edible species:** RED MULBERRY, Texas mulberry *(M. microphylla),* black mulberry *(M. nigra).* **Poisonous look-alikes:** none; see RED MULBERRY caution.

Red mulberry *(Also edible spring)* *Morus rubra*

Habitat: river valleys, floodplains, lower slopes of hills, usually in rich, moist soils. **Identification:** small to medium-size tree with broad, rounded crown and irregularly fissured, dark reddish-brown bark and ***milky sap. Leaves*** alternate, deciduous, simple, *7.5–10 cm (3–4 in)* ***long, 5–5.8 cm (2–2.3 in) wide, broadest near base to almost rounded, unlobed or 1–3-lobed, heart-shaped at base, pointed at tip, with coarse teeth along margin, with whitish hairs on lower surface.*** Male flowers tiny, crowded together in slender, hanging clusters; female flowers tiny, crowded in broader, elongated clusters. ***Fruits cylinder-shaped, blackberrylike, 2.2–3 cm*** (1–1.2 in) ***long, fleshy, red turning dark purple to almost black, sweet,*** maturing from May to July. **Harvest and preparation:** Pick ripe fruit from ladder, or spread tarps beneath tree and shake limbs. Delicious as picked or with cream. Do not let persistent small stem and axis of fruit bother you. For beverage mash and squeeze fruit through cheesecloth or use electric juicer or blender. Add sweetener and lemon juice to taste. Serve with ice and, if desired, club soda. Also good for pies, jams, jellies. Use like other berries in breads, muffins, cakes. Pick shoots as leaves are beginning to unfold. Boil 20 min, drain, add butter and seasoning. *Caution:* Raw shoots and unripe fruits contain hallucinogens. **Related edible species:** WHITE MULBERRY (preceding), Texas mulberry *(M. microphylla),* black mulberry *(M. nigra).* **Poisonous look-alikes:** none.

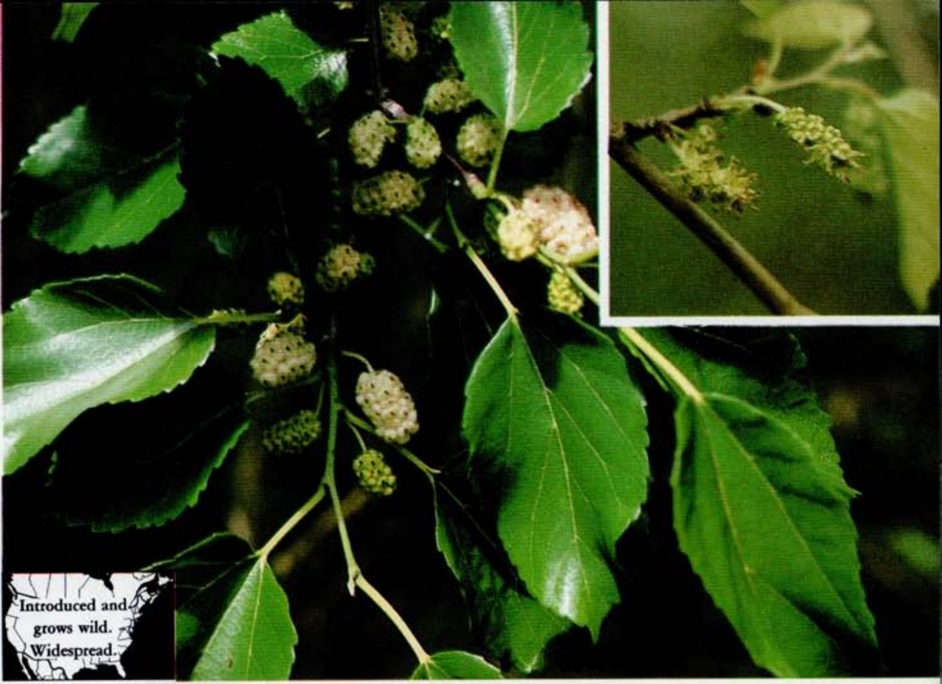
Introduced and grows wild. Widespread.

White mulberry: flowering branch (NYBG). Fruiting branch (Elias).

Red mulberry: fruiting branch (Elias).

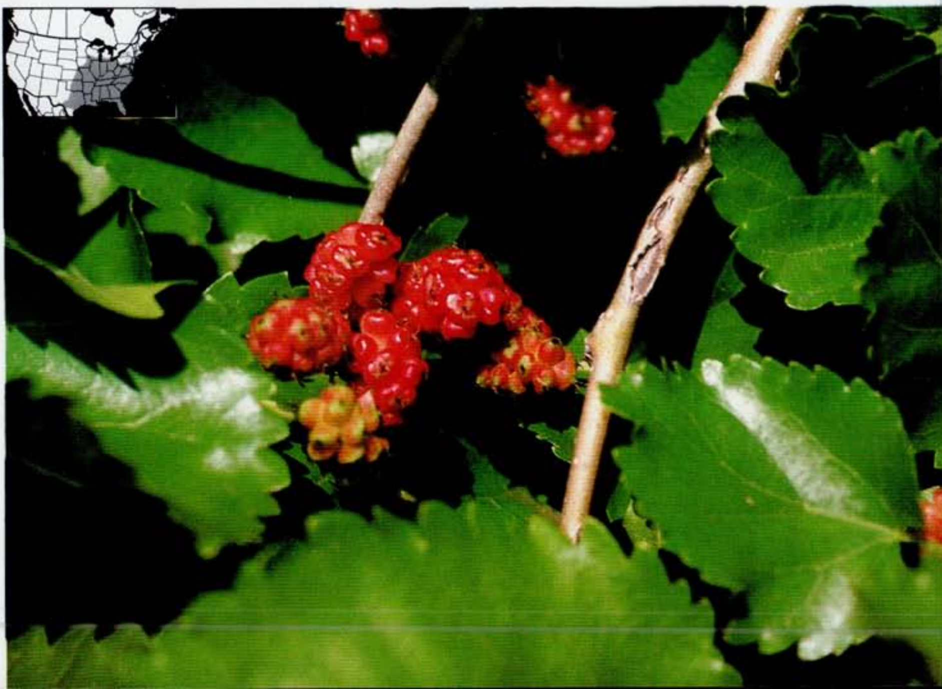

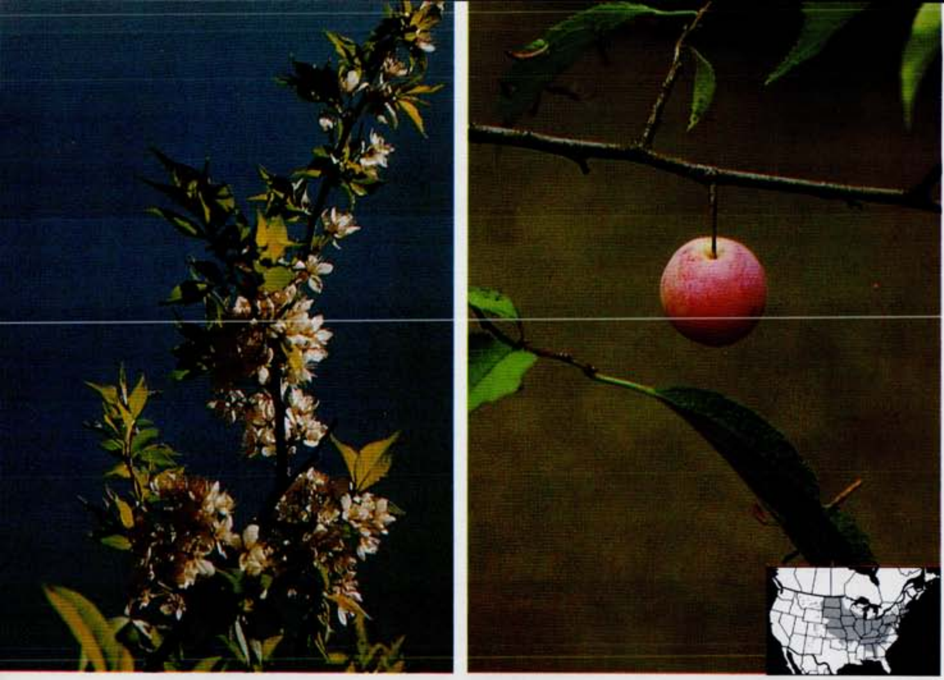

American plum: flowering branch (Weber). Fruiting branch (Elias).

Flatwood plum: flowering and fruiting branches (Anderson).

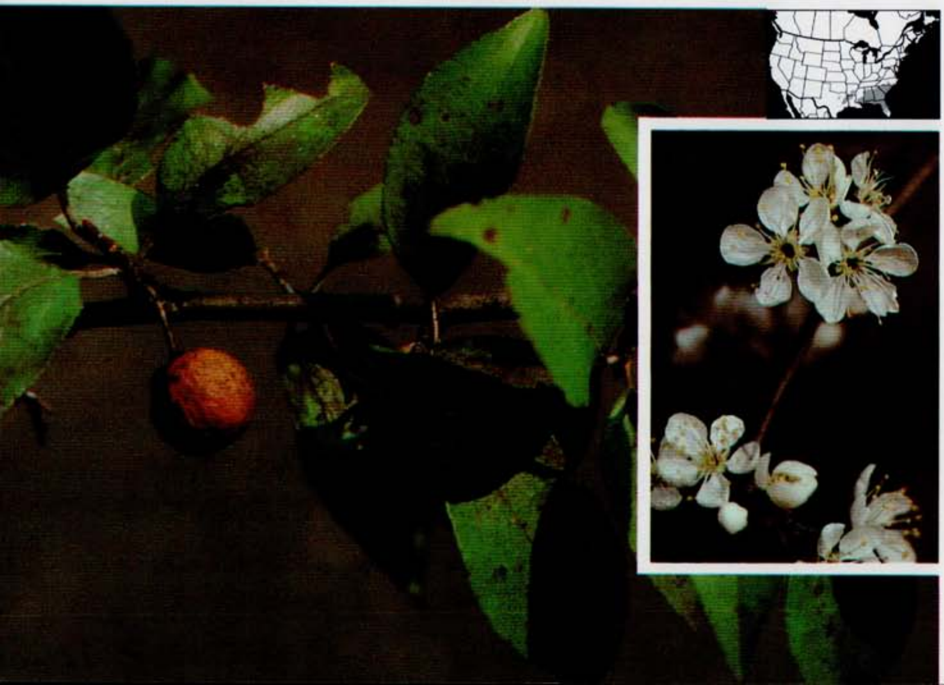

American plum, Wild plum *Prunus americana*

Habitat: common, mixed deciduous woodlands, along streams, fencerows, margins of fields. **Identification:** shrub or small tree to 11 m (36 ft) with spreading crown and ***spiny*** branches. ***Leaves*** alternate, simple, deciduous, ***6–10 cm*** (2.4–4 in) ***long, 3.5–4.5 cm*** (1.4–1.8 in) ***wide, with double row of sharp pointed teeth along margin,*** dark green and smooth. ***Flowers produced 2–5 per cluster*** in early spring, like cherry blossoms, each with 5 rounded, ***white*** petals 9–12 mm (0.4–0.5 in) long. ***Fruits*** solitary or 2–4 per cluster, each ***rounded or nearly so, 1.8–2.5 cm*** (0.7–1 in) ***in diameter, orange to red,*** juicy. **Harvest:** ripe fruits in mid to late summer. **Preparation:** Boil unripe fruits with sugar for emergency food. Use fresh ripe fruits in any cultivated plum recipes. Especially good in jams and jellies. For plum butter, puree pulp and skins after extracting juice for jelly. Blend with equal amount of maple sugar or honey in crock or bean pot and bake at 165°C (330°F) for 3 hr with occasional stirring. Seal in hot, sterilized jars. For jam add ½ cup water to each 2 cups raw pitted fruit. Boil until skins are tender, 10–15 min. Add 1 cup sugar for each cup of fruit pulp. Mix and cook about 15 min to jelly stage with constant stirring. Pour into sterilized jars and seal. For pie, use plums with equal volume of sweeter fruit such as peaches and use basic pie recipe. **Related edible species:** several throughout the country, including BEACH PLUM, FLATWOOD PLUM (next), and cherries. **Poisonous look-alikes:** none. Pits and seeds contain hydrocyanic acid which can react with stomach acids to release highly poisonous cyanin.

Flatwood plum *Prunus umbellata*

Habitat: coastal plain, piedmont; along rivers, in swamps and hammocks. **Identification:** small tree to 6 m (20 ft) with irregular rounded crown and reddish bark. ***Leaves*** alternate, simple, deciduous, ***5–7 cm (2–2.8 in) long, 2.5–4 cm (1–1.6 in) wide, broadest at or above middle,*** rounded to slightly heart-shaped at base, ***with fine, sharp-pointed teeth along margin,*** dark green and smooth. ***Flowers 2–4 per cluster*** in early spring before leaves appear, each with 5 ***white*** rounded petals. ***Fruits*** solitary or in small clusters, ***rounded, 1.2–1.5 cm*** (0.5 in) ***in diameter, dark red to almost black,*** sour, containing flattened stone. **Harvest:** Handpick fruits as they mature from June to September. **Preparation:** See AMERICAN PLUM (preceding), but use additional sugar as with BEACH PLUM. **Related edible species:** BEACH PLUM, AMERICAN PLUM, and several other species. **Poisonous look-alikes:** none.

Black cherry *Prunus serotina*

Habitat: widely distributed; mixed hardwood forests, woodlands. **Identification:** tree to usually 20 m (66 ft), rarely taller, with rounded crown and dark reddish-brown to nearly black bark; inner bark with odor of almond extract. ***Leaves*** alternate, simple, deciduous, 5–15 cm (2–6 in) long, 2.5–4 cm (1–1.6 in) wide, narrow but widest near base to uniformly wide, ***with numerous blunt teeth along margin,*** dark green and shiny. ***Flowers and fruits produced in narrow elongated clusters 10–15 cm*** (4–6 in) ***long,*** each flower with 5 white rounded petals. ***Fruits fleshy, rounded, 8–10 mm*** (0.3–0.4 in) ***in diameter, dark red to black at maturity*** (late summer), juicy, enclosing egg-shaped stone, slightly bitter tasting. **Harvest:** Handpick fruit or if fully ripe shake loose onto tarp August through September. **Preparation:** Eat fresh as trail snack or with sugar and cream as dessert, though pits make eating an effort. Prepare jelly as described under COMMON CHOKECHERRY (next) but use 2 cups each of apple and cherry juice. Also use black cherries for brandied cherries, breads, beverages, pies, wine. For pie add 1½ cups sugar, 2 tbsp cornstarch or flour, and ¼ tsp salt to 1 qt pitted cherries (a pitter makes the task easier). Mix and pour into uncooked 9-in piecrust. Add top crust or lattice of pastry. Bake at 230°C (450°F) for 20 min, then reduce heat to 190°C (375°F) and bake 10 more min. Indians prepared and used dried black cherries as described under COMMON CHOKECHERRY. **Related edible species:** See COMMON CHOKECHERRY. **Poisonous look-alikes:** See COMMON CHOKECHERRY for details.

Common chokecherry *(Also edible autumn)* *Prunus virginiana*

Habitat: common and widespread; moist soils of roadsides, fencerows, woodland margins. **Identification:** shrub or small tree to 8 m (26 ft) with smooth usually reddish-brown bark. ***Leaves*** alternate, simple, deciduous, 5–10 cm (2–4 in) long, 2.5–5 cm (1–2 in) wide, uniformly wide to broadest near base, pointed at tip, ***with numerous sharp teeth along margin,*** dark green and smooth. ***Flowers produced in short cylinder-shaped clusters 8–15 cm*** (3.2–6 in) ***long,*** each with 5 white, rounded petals. ***Fruits fleshy, rounded, 8–10 mm*** (0.3–0.4 in) ***in diameter, usually dark red or black,*** juicy, enclosing egg-shaped stone. **Harvest:** Pick fruit from midsummer to October, depending on location. **Preparation:** Best use, aside from trail thirst quencher, is for jelly. Prepare juice by crushing 1 qt unpitted cherries in 1 cup water. Simmer, stirring about 30 min and strain out juice. Quarter several whole apples, cover with water, simmer until tender. Strain off juice. Combine 2 cups apple juice with 1 cup cherry juice. Add 4 cups sugar and stir until dissolved. Boil rapidly until jelly temperature is reached. Pour into sterilized jelly glasses and seal. Or use sour cherry jelly recipe accompanying pectin package. Indians ground whole fruits, leached poison from seeds, formed pulp into cakes, and dried them in sun for sauce and pemmican. **Related edible species:** BLACK CHERRY (preceding) and several other sweet and sour fruited species, with various ranges. **Poisonous look-alikes:** Autumn and wilted leaves and pits of fruits contain hydrocyanic acid. Do not eat them! Also, Carolina laurelcherry *(P. caroliniana)* fruits are somewhat toxic.

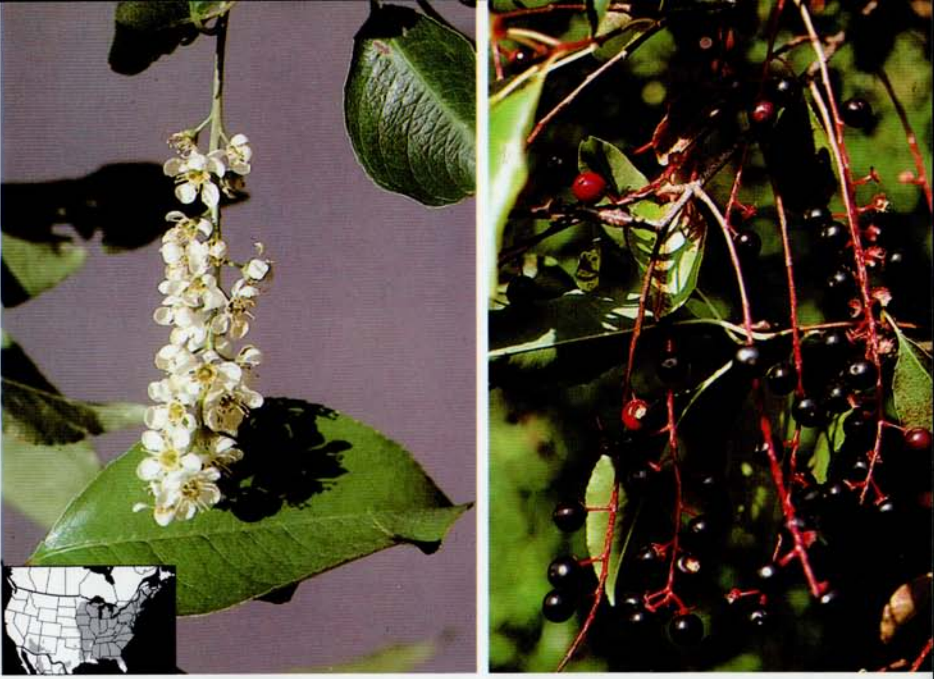

Black cherry: flowering branch (Dykeman). Fruiting branch (Elias).

Common chokecherry: fruiting branch (Elias).

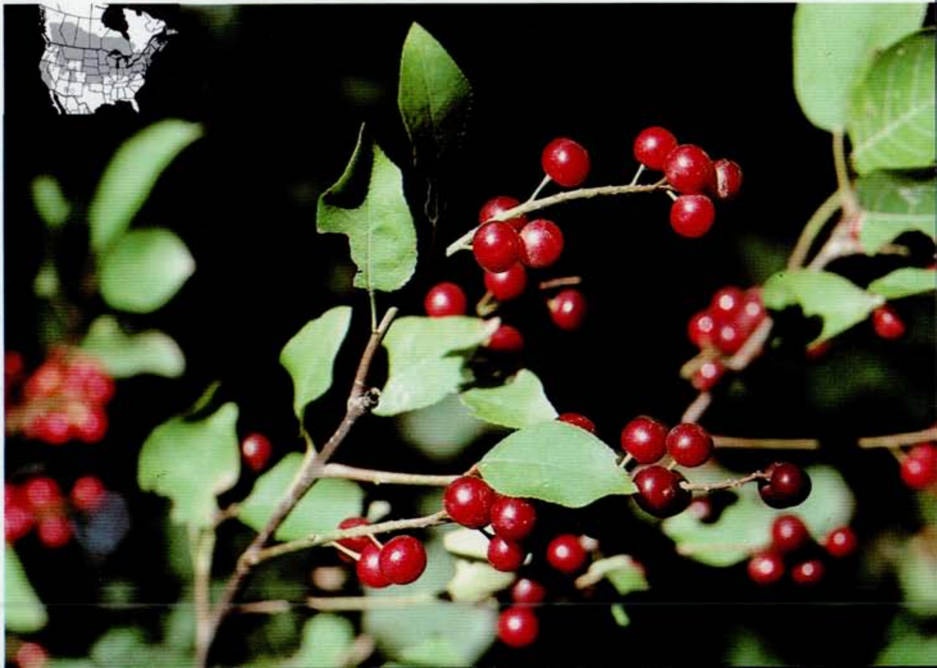

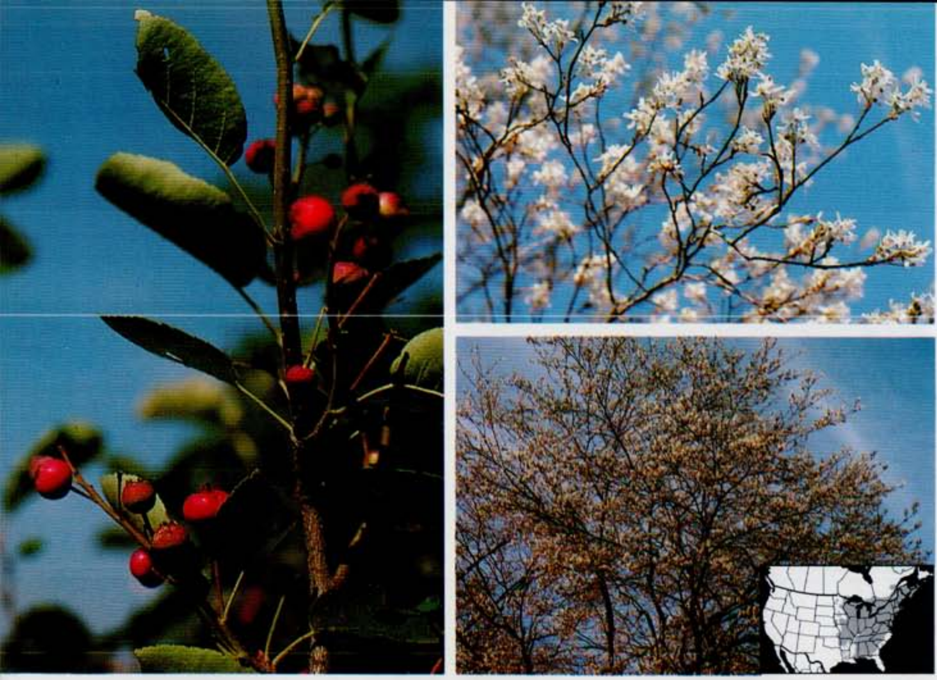

Downy serviceberry: fruiting branch (Elias). Flowering branch and tree (Dykeman).

Saskatoon serviceberry: flowering and fruiting branches (Ross).

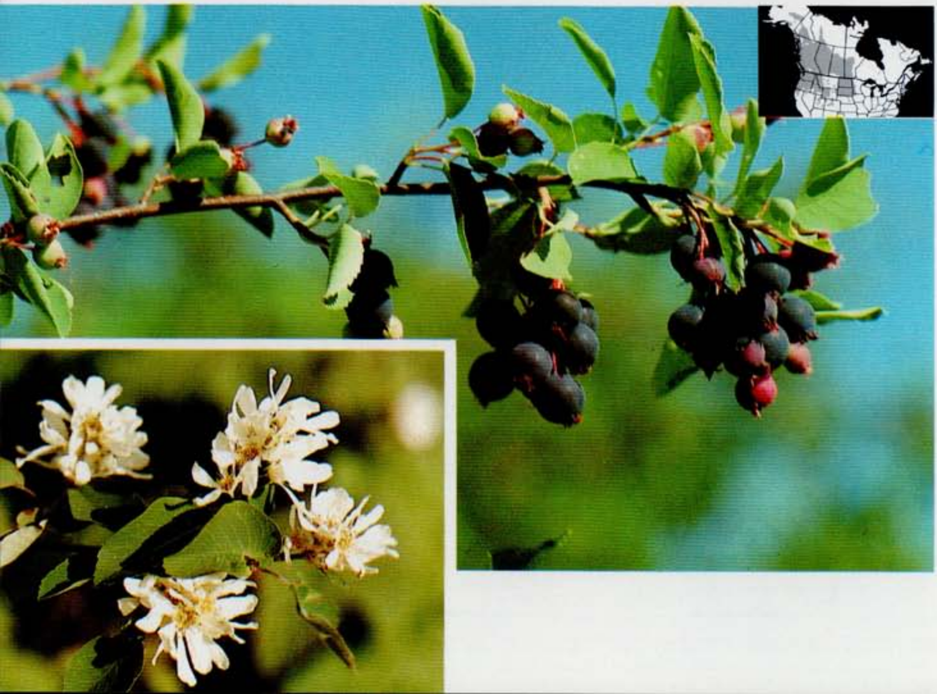

SERVICEBERRY GENUS *Amelanchier*

There are about 25 species of serviceberries native to North America, Europe, Northern Africa, and Asia. In North America, they are small trees or shrubs that occur in deciduous forests and mountains, where their attractive white flowers announce spring. Many birds and wildlife eat the small fruits produced in summer.

Downy serviceberry *Amelanchier arborea*

(synonym: *A. laevis*)

Habitat: woodlands, rocky slopes, river banks, and along swamps; often growing with OAK, hickory, tulip tree, SWEET BIRCH. **Identification:** shrub or ***small tree*** with rounded crown and thick, slightly ridged and furrowed, ash-gray to nearly black bark. ***Leaves*** alternate, deciduous, simple, 5–9 cm (2–3.5 in) long, 2.5–3.5 cm (1–1.4 in) wide, widest at or near middle, ***finely toothed along margin. Flowers white, attractive,*** produced in spring (February in South to May in North), each ***with 5 strap-shaped petals 1–1.4 cm (0.4–0.6 in) long.*** Fruits in clusters, each rounded, 6–10 mm (0.2–0.4 in) in diameter, red to dark purple, fleshy, containing 5–8 black seeds. **Harvest and preparation:** Pick berries when ripe, usually mid to late summer. Eat as picked. Use fresh in recipes for jelly, jam, pie; fresh or dried in breads. For fruit sauce, stew 3 cups berries with 1 cup sugar for about 20 min. Mix dried berries with dried beef and suet to form pemmicanlike product. **Related edible species:** SASKATOON SERVICEBERRY (next) and other less common species. **Poisonous lookalikes:** none.

Saskatoon serviceberry *Amelanchier alnifolia*

(synonym: *A. florida*)

Habitat: mountain slopes, moist hillsides, prairies, margins of lakes and streams. **Identification:** shrub or small tree, sometimes forming thickets, with rounded crown and thin, smooth, or finely furrowed, brown bark. ***Leaves*** alternate, deciduous, simple, 2–6 cm (0.8–2.4 in) long, 2.5–4 cm (1–1.6 in) wide, broadest near base to uniformly wide, ***entire near base, sharply and coarsely toothed along upper margin. Flowers showy, white, produced in early spring,*** each with 5 ***strap-shaped petals 0.9–1.5 cm*** (0.4–0.6 in) long. Fruits in clusters, each rounded to egg-shaped, 6–11 mm (0.2–0.4 in) long, dark reddish-purple to almost black. **Harvest and preparation:** Fruits ripen in mid to late summer; eat raw or dried but better cooked. Use fruits in recipes calling for fruit, as in cakes, pies, jellies, puddings. To make jelly, wash fruits thoroughly, drain, partially crush, cover with water, and cook over low to medium heat until soft. Pour contents into wet jelly bag. For clear jelly do not squeeze bag. Simmer juice 5 min, removing any surface foam. Add sugar while stirring (1 cup juice to ⅔ cup sugar). Simmer 10–15 min or until mix shows sign of jelling. Pour mix into properly prepared jelly jars to within ¼ in of top. Cover with ¼ in melted paraffin. **Related edible species:** DOWNY SERVICEBERRY (preceding) and other less common species. **Poisonous look-alikes:** none.

Autumn

Autumn is the season for fruits and nuts. These come mainly from trees and shrubs that require a full growing season before flowers transform into mature seeds and fruits. Oaks, beech, hickories, walnuts, hawthorns, viburnums, and grapes are common sources and easy to identify. In autumn you can gather large quantities of edibles and consume them then or store them. The oaks, beech, walnuts, and hickories are cyclic in nut production, or mast crop. That is, a large mast crop is normally followed by two or three years of smaller crops. So one year you may encounter a large crop of hickory nuts and the next year run into the acorn bounty of oaks instead.

This is also the time to note or even map populations of native edible plants (other than nuts) for harvest the next year. Keep track of the edible populations, especially immature ones; and avoid overharvesting them. Many plant species have been seriously depleted by greedy collectors. On the other hand, if you harvest fruits and nuts, you don't destroy the trees and shrubs. They will live to produce again, and their offspring will eventually be producers too—especially if you plant a fruit or nut now and then.

Additional autumn edibles

Plants listed below have parts that are edible in autumn and at least one additional season. Descriptions and photos of these plants appear in a seasonal section other than "Autumn," as indicated.

Alpine bistort (*Polygonum viviparum*), p. 154
American bistort (*Polygonum bistortoides*), p. 154
American hazelnut (*Corylus americana*), p. 172
Beach plum (*Prunus maritima*), p. 166
Beaked hazelnut (*Corylus cornuta*), p. 172
Bugleweed (*Lycopus uniflorus*), p. 88
California juniper (*Juniperus californica*), p. 194
California laurel (*Umbellularia californica*), p. 198
Cattail (*Typha latifolia*), p. 68
Chicory (*Cichorium intybus*), p. 114
Chufa (*Cyperus esculentus*), p. 130
Common burdock (*Arctium minus*), p. 112
Common chickweed (*Stellaria media*), p. 84
Common chokecherry (*Prunus virginiana*), p. 204
Common day lily (*Hemerocallis fulva*), p. 136
Cranberry (*Vaccinium macrocarpon*), p. 162
Creeping snowberry (*Gaultheria hispidula*), p. 152
Creeping wintergreen (*Gaultheria humifusa*), p. 160

Dandelion (*Taraxacum officinale*), p. 104
Eastern camass (*Camassia scilloides*), p. 64
Edible valerian (*Valeriana edulis*), p. 88
Evening primrose (*Oenothera biennis*), p. 102
Glandular mesquite (*Prosopis glandulosa*), p. 188
Glasswort (*Salicornia europea*), p. 76
Great bulrush (*Scirpus validus*), p. 70
Green pigweed (*Amaranthus hybridus*), p. 72
Groundnut (*Apios americana*), p. 148
Hard-stem bulrush (*Scirpus acutus*), p. 70
Indian fig (*Opuntia humifusa*), p. 138
James chickweed (*Stellaria jamesiana*), p. 84
Labrador tea (*Ledum groenlandicum*), p. 160
Lamb's quarters (*Chenopodium album*), p. 74
Lance-leaved pickerelweed (*Pontederia lanceolata*), p. 132
Live-forever (*Sedum purpureum*), p. 110
Marsh mallow (*Althaea officinalis*), p. 146
Mountain cranberry (*Vaccinium vitis-idaea*), p. 162
Narrow-leaved cattail (*Typha angustifolia*), p. 68
Pawpaw (*Asimina triloba*), p. 198
Pickerelweed (*Pontederia cordata*), p. 130
Pigweed (*Amaranthus retroflexus*), p. 72
Plantain (*Plantago major*), p. 78
Poverty weed (*Monolepis nuttalliana*), p. 74
Prickly pear (*Opuntia phaeacantha*), p. 140
Reed grass (*Phragmites communis*), p. 128
Rocky Mountain juniper (*Juniperus scopulorum*), p. 194
Salal (*Gaultheria shallon*), p. 160
Sassafras (*Sassafras albidum*), p. 124
Screwbean mesquite (*Prosopis pubescens*), p. 188
Sheep sorrel (*Rumex acetosella*), p. 120
Silver buffaloberry (*Shepherdia argentea*), p. 178
Smooth sumac (*Rhus glabra*), p. 184
Sour dock (*Rumex crispus*), p. 120
Southern bayberry (*Myrica cerifera*), p. 174
Squawbush (*Rhus trilobata*), p. 186
Staghorn sumac (*Rhus typhina*), p. 250
Sunflower (*Helianthus annus*), p. 156
Sweet birch (*Betula lenta*), p. 122, 252
Sweet gale (*Myrica gale*), p. 176
Utah juniper (*Juniperus osteosperma*), p. 196
Watercress (*Nasturtium officinale*), p. 146
Western camass (*Camassia quamash*), p. 136
Wild garlic (*Allium canadense*), p. 60
Wild leek (*Allium tricoccum*), p. 60
Wild onion (*Allium cernuum*), p. 58
Wild rice (*Zizania aquatica*), p. 128
Wild strawberry (*Fragaria virginiana*), p. 90
Wintergreen (*Gaultheria procumbens*), p. 150
Yellow birch (*Betula alleghaniensis*), p. 122, 254

Jerusalem artichoke *Helianthus tuberosus*

Also edible winter, spring

Habitat: damp or rich thickets, waste areas, roadsides; often cultivated today as by Indians in past. **Identification:** perennial herb to 3 m (9.8 ft) tall, forming colonies; stems upright, rough-hairy, from a root system with elongated, fleshy, edible tubers 7.5–12.5 cm (3–5 in) long. ***Leaves opposite or in whorls of 3 on lower part of stems, alternate above,*** broadest near base, ***to 25 cm*** (10 in) ***long, 6–15 cm*** (2.4–6 in) ***wide,*** tapered at tip and base, ***hairy above, rough, leathery. Flower heads yellow*** with rounded central disc 1–2.5 cm (0.4–1 in) across, ***10–20 ray flowers,*** each about 4 cm (1.6 in) long, ***several flower heads per plant, on slender stems from leaf bases on upper part of stems.*** **Harvest:** Identify colonies in summer when in bloom; dig tubers from first frost through winter. **Preparation:** Peel or scrub tubers before using. Slice raw in salad, make potato salad, boil and mash, roast with meat, slice and fry. For pickles, peel and boil 3–4 min, cover with wine vinegar, age 3–5 weeks. Adapt to many potato and carrot recipes. Food value equal to potato. The carbohydrates are mostly inulin and suitable for low starch diets. **Related edible species:** other SUNFLOWERS (*Helianthus* species) produce edible seeds. **Poisonous look-alikes:** none.

Lotus lily, American lotus *Nelumbo lutea*

Also edible spring, summer

Habitat: lakes, ponds, slow-moving streams, bays and tidal waters. **Identification:** ***aquatic perennial herb;*** roots large, thick with swollen tubers. ***Leaves large,*** 20–70 cm (8–28 in) across, circular, often ***rising above water level, center depressed,*** bluish-green; ***leafstalk attached at center of blade,*** variable in length, strong but flexible. ***Flowers*** solitary on long stalks, ***showy,*** 14–26 cm (5.5–10.2 in) across, ***with numerous pale yellow petals*** or petallike structures. ***Fruits produced in large cone-shaped, flat-topped receptacles*** raised above water level; each fruit dry, hard, nutlike, globe to egg-shaped, 1–2 cm (0.4–0.8 in) long. **Harvest:** Some part edible most of year. Pick unopened leaves and young stalks in spring. Pick seed receptacle either before or after seeds ripen during summer through autumn. Dig up tuberous root in autumn or early spring if you can overcome the difficulty of its deep water bed. **Preparation:** Leaves and young stalks are edible, but not exceptional, potherbs. Immature seeds are excellent raw, boiled, or roasted. Mature seeds are hard-shelled; crack shells before boiling to soft chestnutlike texture. Or roast *fresh* mature seeds, crack, and remove shells. Eat like nuts, or grind into meal and substitute for part of flour in bread and muffin recipes. Bake or boil tubers like sweet potatoes until tender. Peel and mash. To every cup add 1 tbsp butter, ¼ tsp salt, and a little hot milk. Beat to smooth texture. **Related edible species:** various species of YELLOW POND LILY (next) and water lily *(Nymphaea)*. **Poisonous look-alikes:** none.

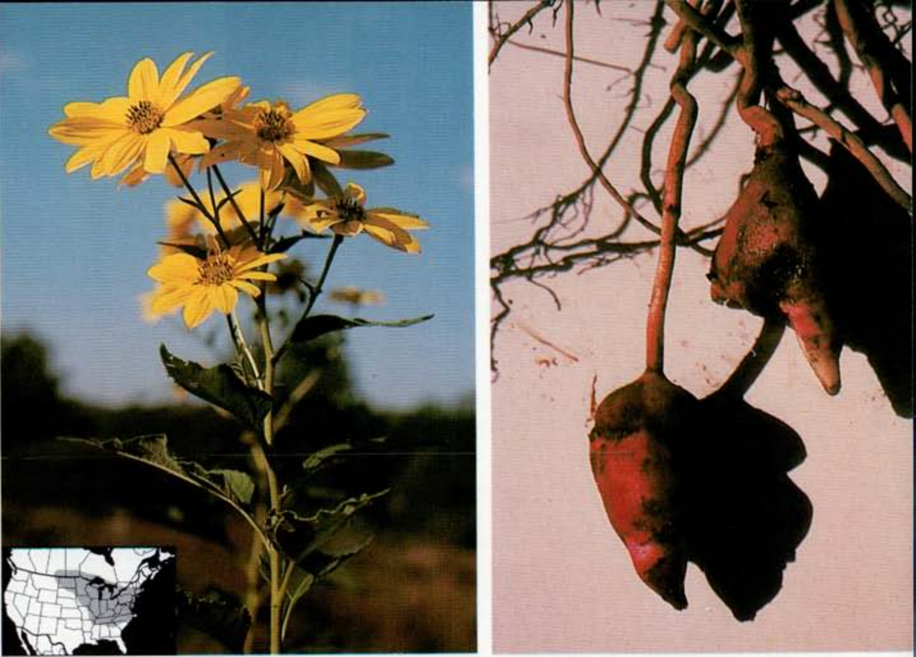

Jerusalem artichoke: flowers and tubers (Elias).

Lotus lily: receptacle with protruding fruits and flower (Feldman).

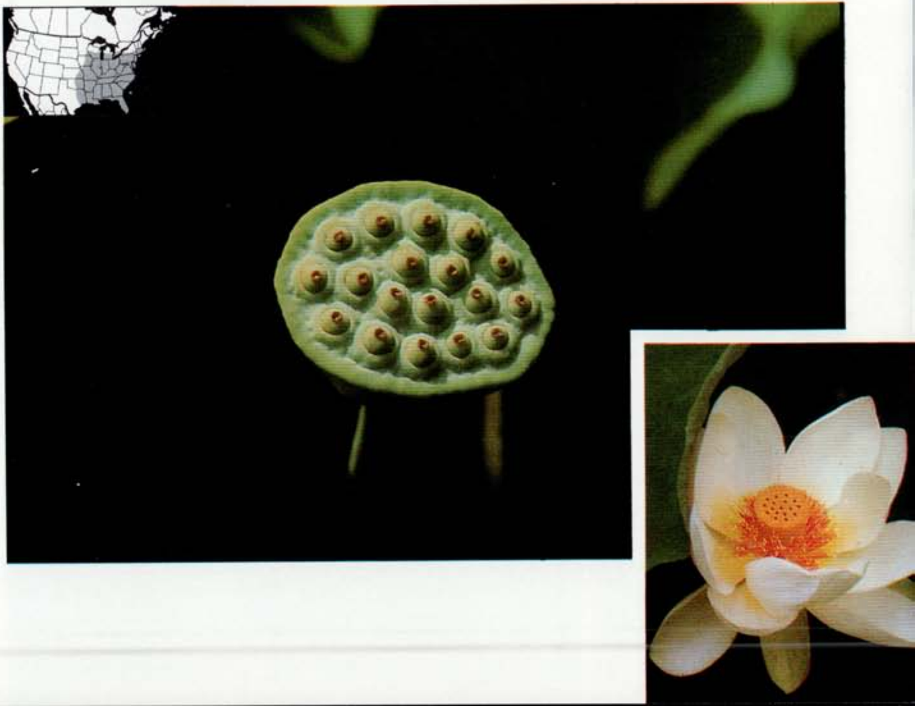

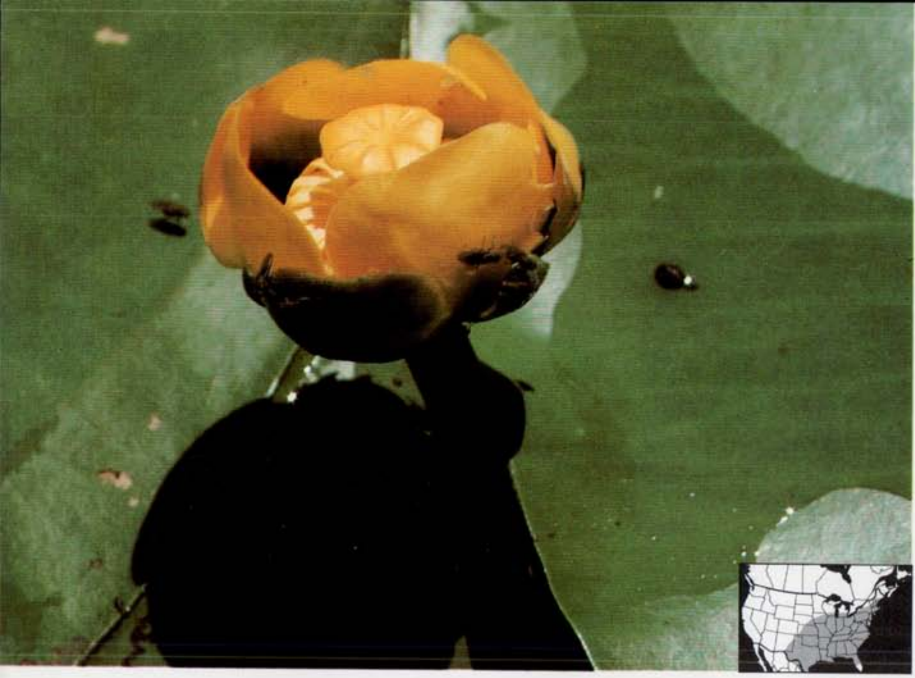

Yellow pond lily: flower prior to seed development (Elias).

Arrowhead: flower and whole plant (NYBG).

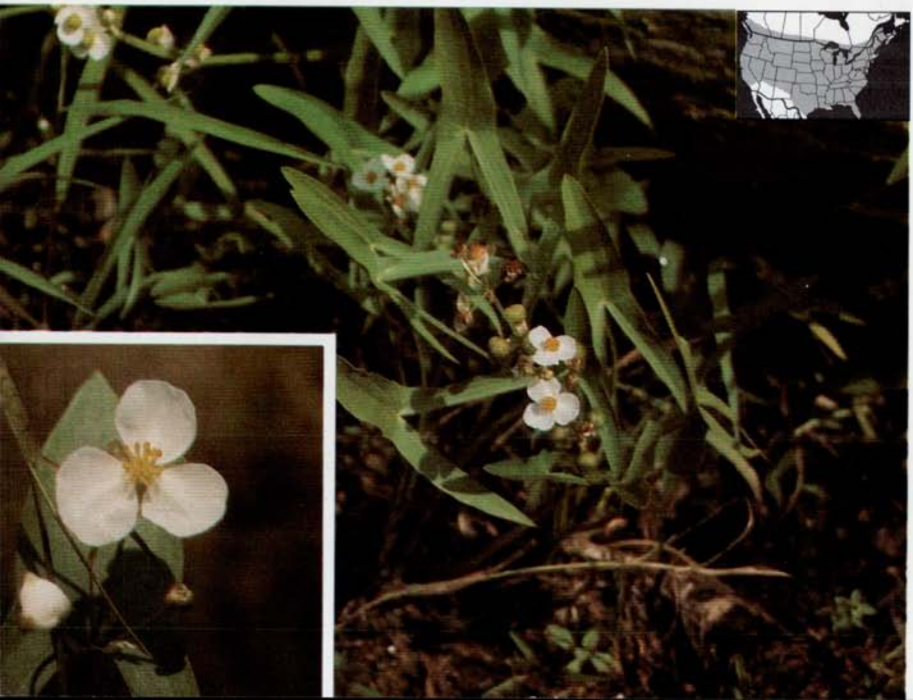

Yellow pond lily, Spatterdock *Nuphar advena*

Also edible winter, spring, summer

Habitat: lakes, ponds, slow streams, swamps, tidal waters. **Identification:** ***aquatic perennial herb;*** roots (actually stems or rhizomes) large, thick, to 13 cm (5 in) in diameter, elongate, branching, covered with leaf scars. ***Leaves large,*** 10–40 cm (4–16 in) long, 5–16 cm (2–6.3 in) wide, broadest below middle, ***deeply lobed at base to form U- or V-shaped cleft,*** shiny green, ***usually raised above water level,*** sometimes floating; leafstalk thick, stout, attached at base of cleft. ***Flowers*** solitary on long stalks, ***showy,*** 3–10 cm (1–4 in) across, ***with numerous yellow, green-tipped petals. Fruits broadly urn-shaped,*** 2–5 cm (0.8–2 in) long, with short thick neck and flattened top, ***ribbed,*** green, containing numerous, yellowish-brown, small, egg-shaped seeds. **Harvest:** Cut fruits free of stalk as they ripen (late summer or fall). Air-dry until fruits pull apart to separate seeds. Dump seeds in bucket of water. After 5–6 hr good ones sink; dry in sun or warm oven. Loosen tuberous roots from mud in bottom of pond from autumn to spring. Indians used their toes. **Preparation:** Seeds store well. Pop seeds like popcorn or remove shells by parching over coals or in oven, pounding lightly to crack shells, and winnowing. Then boil nutmeats like rice or grind and use like cornmeal. Peel, slice, and boil roots in 2 water changes. Mash, season, and serve like potatoes or dry and use like flour. **Related edible species:** several other species of yellow pond lilies *(Nuphar)*, most important is cow lily *(N. polysepalum)* in West. **Poisonous look-alikes:** none.

Arrowhead, Duck potato, Wapato *Sagittaria latifolia*

Also edible winter, spring, summer

Habitat: shores of lakes, ponds, streams, ditches, bogs, wet meadows; common and widespread. **Identification:** ***perennial herb*** 0.6–1.2 m (2–4 ft) high, with fibrous roots developing small tubers (potatolike) 0.5–5 cm (0.2–2 in) in diameter. ***Leaves all at base,*** blades 10–25 cm (4–10 in) long, 1–25 cm (0.4–10 in) wide, ***arrowhead-shaped with long pointed lobes at base, leaf shape highly variable*** from extremely narrow to very broad, smooth or hairy; leafstalks flexible, to 1 m (3.2 ft) long. Flowering stalk taller than leaves; ***flowers in clusters of 3s along upper part of spike, each flower with 3 delicate, white petals.*** Fruiting heads globe-shaped or nearly so, containing numerous flat, winged seeds. **Harvest:** Collect tubers from late summer through early spring, whenever there is no ice. Dislodge tubers from roots in mud beneath water with rake or hoe, or wade in as the Indians did and loosen them with your toes. Tubers will float to surface and are easily collected. **Preparation:** Tubers are edible but not always tasty raw; better cooked like potatoes. Bake or boil about 30 min until tender; peel after cooking. Serve topped with mixture of butter and lemon juice or STAGHORN SUMAC extract. Use in potato recipes. To dry tubers for storage, boil until tender, drain, slice about ½-in thick, dry thoroughly in sun or warm oven. Store in warm dry place. Before using soak 20 min; allow 20 min cooking time. **Related edible species:** several other arrowhead species *(Sagittaria)*; all have arrowhead leaves; PICKERELWEED has similar leaves but lacks tubers. **Poisonous look-alikes:** none.

WILD GRAPES *Vitis*

North America has about 50 native species of grapes along with several introduced ones that now grow wild. While it is easy to identify a grape vine, determining grape species can be challenging. It is best to familiarize yourself with the grapes in your region and the variation in fruit quality. Grapes are climbing vines with large lobed or toothed leaves, numerous small flowers in elongate clusters, and then fruit bunches. Three of the more common grapes are described here. **Harvest:** Gather leaves in spring when young and tender but full grown. Collect fruits between August and October, depending upon species and region. **Preparation:** Boil leaves until tender, and serve with other vegetables. Check cookbook recipes for stuffed grape leaves. Most fresh grapes are too acid to eat in quantity. Best to make juice, jelly, preserves, pie, and wine. For jelly, use mixture of ripe and unripe (for pectin) fruits. Wash, drain, remove stems, and mash in kettle. Add small amount of water and cook until fruit comes apart, about 15–20 min. Strain juice through jelly bag. Prevent crystallization in jelly by letting juice sit overnight or cooking 1 cup tart apples with each 4 cups of grapes. Boil 4

Riverside grape *Vitis riparia*

Also edible spring, summer

Habitat: stream and river banks, margins of woodlands, thickets. **Identification:** *large vine,* usually climbing; branches tough, rounded, becoming smooth, pith chambered. ***Leaves*** alternate, simple, deciduous, ***12–20 cm*** (4.7–8 in) ***long,*** 8–16 cm (3–6 in) wide, nearly circular in outline to broadest near base, ***usually 3-lobed, long-pointed at tip,*** coarsely toothed along margin, heart-shaped at base, smooth above, ***hairy beneath.*** Flowers tiny, produced in many-flowered, elongate clusters to 10 cm (4 in) long. Fruits are dense bunches of ***berries,*** each ***rounded, 8–14 mm*** (0.3–0.6 in) ***in diameter, black with whitish bloom,*** sour tasting. **Harvest and preparation:** See WILD GRAPES (preceding).

Frost grape *Vitis vulpina*

Also edible spring, summer

Habitat: streams and river banks, low wetlands, thickets. **Identification:** ***large vine,*** often climbing; branches stout, rounded, becoming smooth, pith chambered. ***Leaves*** alternate, simple, deciduous, ***10–18 cm*** (4–7 in) ***long, almost as wide,*** very broad, widest near base to almost round, ***usually unlobed or slightly 3-lobed,*** pointed at tip, coarsely toothed along margins, heart-shaped at base, smooth above, smooth below except for junctions of major veins. Flowers tiny, produced in many-flowered, elongated clusters to 18 cm (7 in) long. Fruits are clusters of ***berries,*** each ***rounded, 4–10 mm*** (0.2–0.4 in) ***in diameter, black,*** sour tasting until frost, then sweet. **Harvest and preparation:** See WILD GRAPES (preceding).

cups juice with 4 cups sugar until jelling point is reached. Skim off foam, pour into sterilized jars, seal with paraffin. For pie, squeeze skins from 4 cups ripe grapes. Simmer pulp until it comes apart; sieve to remove seeds. Chop skins and combine with strained pulp, 1½ cups sugar, 2 tbsp flour, ¼ tsp salt, 1 tbsp lemon juice, 2 tbsp melted butter. Mix well and pour into unbaked pie shell. Fit and slit top crust. Bake at 205°C (400°F) for 30–40 min until crust is browned. For juice, use ripe fruit. Crush in kettle, heat and simmer for 30 min, but do not boil. Press through sieve and strain through jelly bag. Sweeten to taste. For wild grape butter, puree remaining pulp. Add equal amount to maple sugar or honey and blend in bean pot. Bake in preheated 160°C (320°F) oven, stirring occasionally, for 3 hr. Seal in sterilized jars. For emergency water source, cut off large vine and drain sap. **Poisonous look-alikes:** moonseed *(Memispermum canadense)* is a vine with simple entire or shallowly lobed leaves and clusters of black fruits resembling a bunch of grapes. The leafstalk is attached on the lower surface of the blade above the base (peltate) and the leaves are not coarsely toothed as those of most grapes. Single seed a flat crescent.

Riverside grape: fruiting branch (Elias).

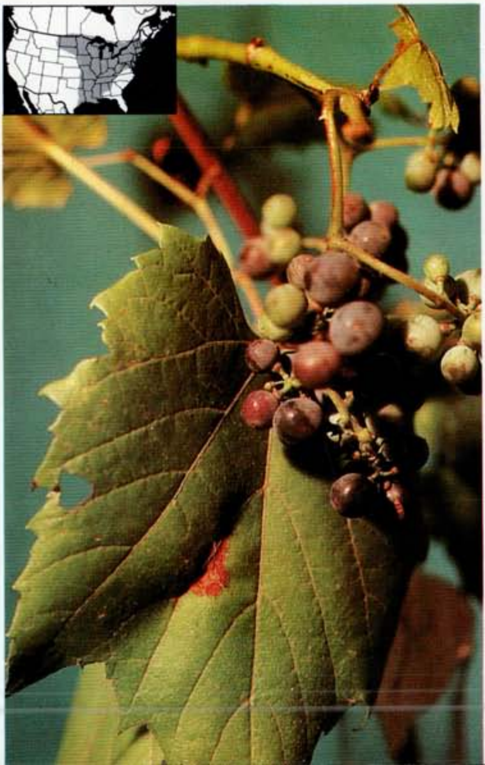

Frost grape: fruiting branch (Elias).

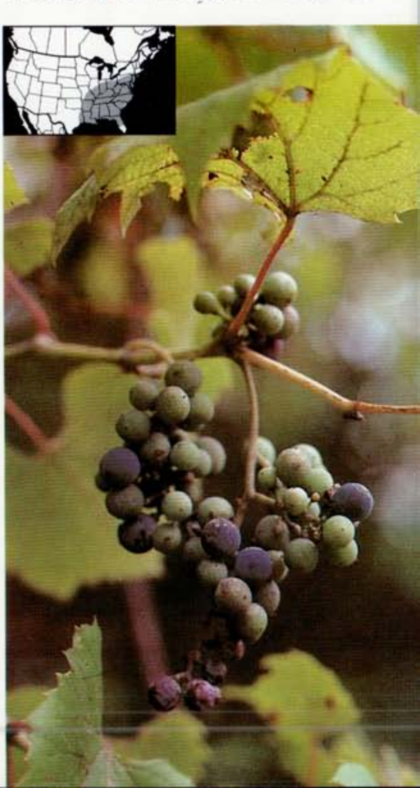

California grape *Vitis californica*
Also edible spring, summer

Habitat: stream and river banks, moist canyons, margins or clearings of forests. **Identification:** large ***woody vine,*** climbing; branches stout, rounded, with shredding bark, pith chambered. ***Leaves*** alternate, simple, deciduous, ***6–14 cm*** (2.4–5.5 in) ***long,*** nearly as wide, ***circular in outline, nearly unlobed to 3-lobed,*** short-pointed at tip, coarsely toothed along margin, deeply lobed at base, hairy above, dense cobweblike hairs beneath. Flowers tiny, produced in many-flowered, branched, elongated clusters. ***Fruits*** are dense bunches of ***berries,*** each rounded, ***6–10 mm*** (0.2–0.4 in) ***in diameter, dark purple and covered with whitish bloom.*** **Harvest and preparation:** See WILD GRAPES (preceding).

Wild potato vine *Ipomoea pandurata*
Also edible spring, summer

Habitat: fields, fencerows, roadsides; dry, open or partly shaded areas. **Identification:** ***perennial*** trailing or slightly twining ***vine,*** 0.9–3.6 m (3–12 ft) long, ***often purplish along stem,*** branching, ***from a large tuberous root to 9 kg*** (20 lbs), ***vertical, deep, slightly milky,*** resemble a yam. ***Leaves 4–8 cm*** (1.6–3.2 in) ***long, 3–8 cm*** (1.2–3.2 in) ***wide, heart-shaped, pointed at the tip, entire along the margins. Flowers*** to 7.5 cm (3 in) across, with a ***white*** (sometimes purple-centered), ***funnel-shaped*** corolla, like a morning glory flower. Fruit a 2-chambered capsule 0.8–1.2 cm (0.3–0.5 in) long, broadest near the base and tapering to a tip. **Harvest:** Dig deep root any time when ground is not frozen, but best not in summer when starch level is low. Young roots are preferred to older, woody ones. **Preparation:** Boil or bake roots like sweet potatoes. When boiling, change water 2 or 3 times for bitter ones; others will not need changing. All roots have tough outer skins and require peeling. Serve with butter and seasonings. Some older roots may be too woody to use. Dry peeled slices of root for later use. **Related edible species:** Bush morning glory *(I. leptophylla)* of the West is variously described from excellent to nearly inedible. Edibility of other species has not been reported. **Poisonous look-alikes:** Flowers may resemble those of JIMSON WEED; however, JIMSON WEED is an upright herb with large, coarsely toothed leaves.

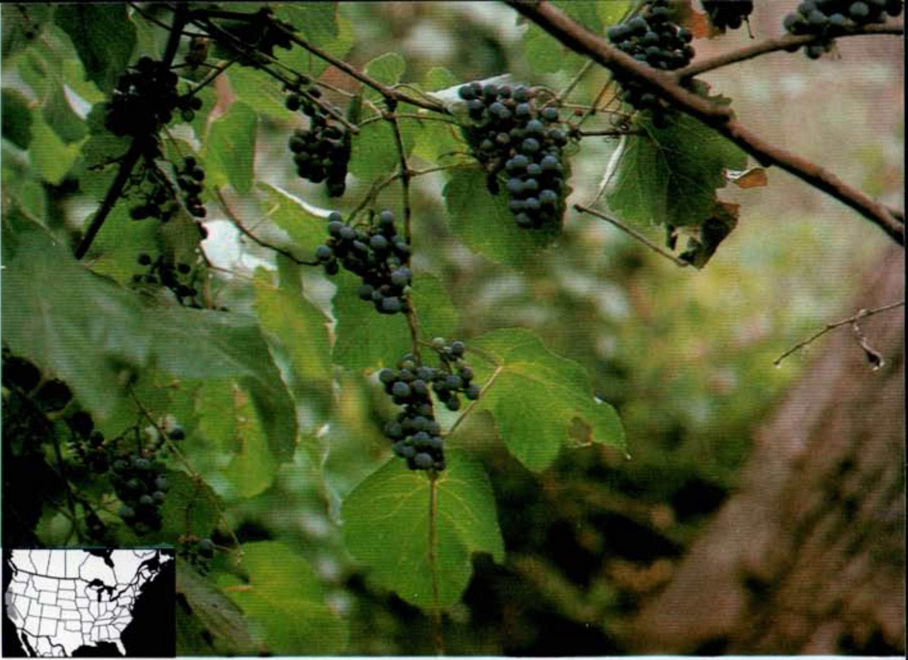

California grape: fruiting branch (Ornduff).

Wild potato vine: flower (Mohlenbeck).

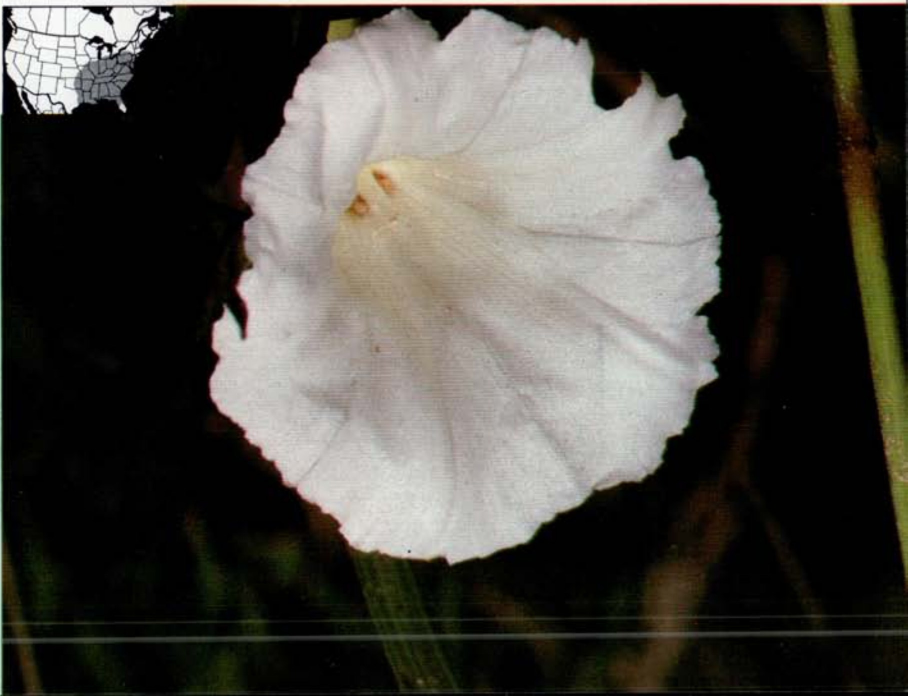

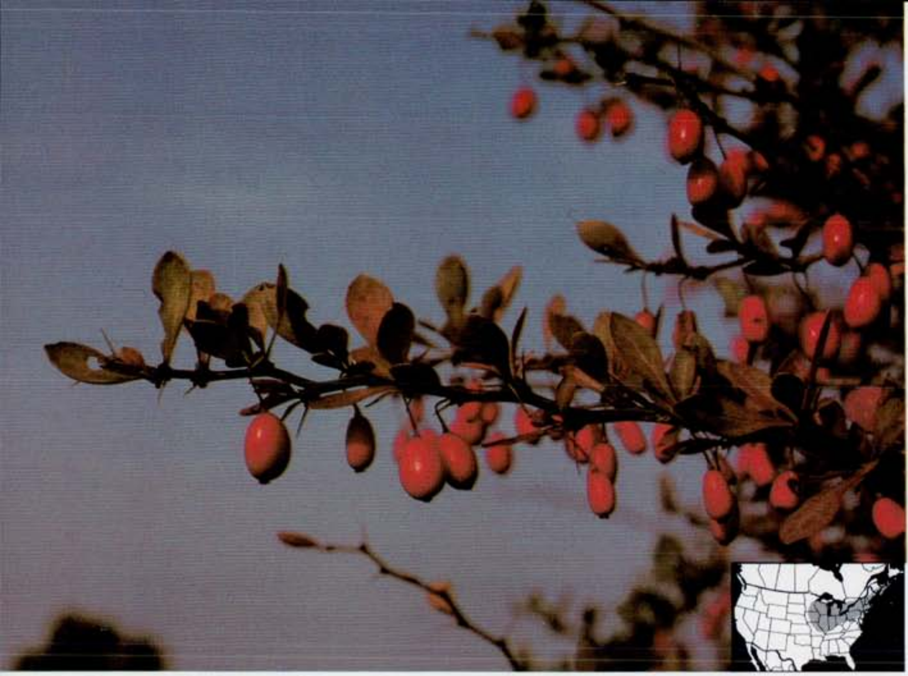

Common barberry: fruiting branch (Dykeman).

Spice bush: fruiting branch (Elias). Flowers (NYBG).

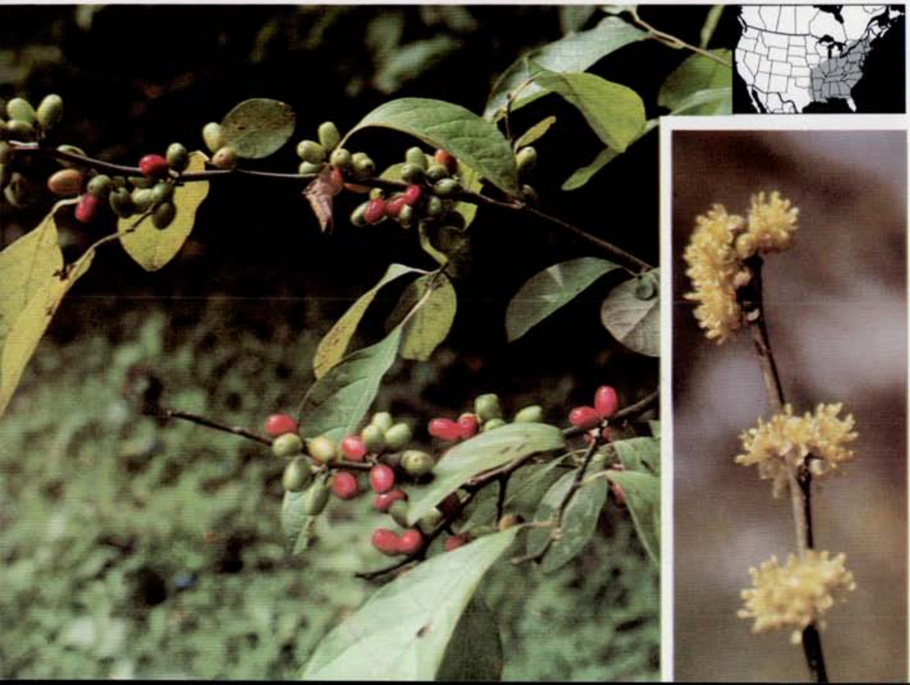

Common barberry *Berberis vulgaris*

Habitat: widely planted European shrub; also grows wild, the seeds spread by birds; disturbed woodlands, fencerows, fields, roadsides. **Identification:** shrub 1–3 m (3–10 ft) high, with rounded crown. Branches grooved, yellowish-red, turning gray by second season, with ***3-forked sharp-spines. Leaves*** alternate, simple, deciduous, ***often crowded together on short shoots,*** 2–4.8 cm (0.8–2 in) long, 0.8–2.4 cm (0.3–1 in) wide, ***broadest above middle, rounded at tip, finely sharp-toothed along margin.*** Flowers produced in spring, stalked, in 10–20-flowered, elongated, narrow clusters, each flower greenish-yellow. ***Fruits in drooping clusters*** in late summer or early autumn, ***berries 8–12 mm*** (0.3–0.5 in) ***long, orange-red to scarlet.*** **Harvest and preparation:** Pick fruits in autumn when full color has developed. For jelly, 4 cups crushed barberries, 2 whole oranges, chopped, 2 cups water, 1½ cups sugar, 1 pkg pectin. Combine berries, oranges, and water. Bring to boil, simmer until tender, about 20 min. Strain juice through cloth or jelly bag. Combine juice and sugar; bring to rolling boil. Add pectin; bring to boil. Pour into hot, sterilized jars; seal. Yields 3 pt. For straight barberry jelly, cook berries and extract juice. Boil juice to 105°C (200°F); pour into hot, sterilized jars; seal. Needs no pectin, very tart. Juice also good in sauces and beverages where tartness is needed. To relieve thirst, chew a few leaves. **Related edible species:** red-fruited barberry *(B. haematocarpa),* southwestern U.S.; American barberry *(B. canadensis)*, southeastern U.S.; and other species. **Poisonous look-alikes:** none.

Spice bush *Lindera benzoin*

Also edible winter, spring, summer

Habitat: rich woodlands, stream banks, especially in moist soils. **Identification:** shrub to 5 m (16.4 ft) high, with numerous, spreading to upright, smooth ***branches, with spicy fragrance when broken.*** Leaves alternate, simple, deciduous, 6–14 cm (2.4–5.5 in) long, 3–6 cm (1.2–2.4 in) wide, broadest near or above middle, long-pointed at tip, entire along margin, thin, smooth, bright green, stalked. ***Flowers in early spring before leaves appear, yellow, small, produced in dense clusters along previous year's twigs. Fruits*** produced in clusters, each ***a berry,*** widest at middle, rounded at end, ***6–10 mm (0.2–0.4 in) long,*** green turning ***bright red at maturity.*** **Harvest:** young leaves in spring; bark and twigs all year; fruits in late autumn before migrating flocks of birds eat them all. **Preparation:** For tea steep 15 leaves or handful of fresh twigs and bark in 4 cups boiling water for 15 min. Strain if necessary; serve with honey and milk for best flavor. Dry fruits thoroughly in warm oven. Grind and use as substitute for allspice. **Related edible species:** none. **Poisonous look-alikes:** none.

Desert hackberry, Spiny hackberry *Celtis pallida*

Habitat: desert foothills and mesas at 450–1,050 m (1,500–3,500 ft). **Identification:** ***shrub or small tree,*** to 5 m (16 ft) tall, single or in thickets, ***densely branched, twigs with many single or paired spines at base of leaves. Leaves simple, alternate, evergreen, broadest near or below the middle,*** 2.5–7.5 cm (1–3 in) long, ***unequal at base, toothed from mid margin to tip, rough upper surface.*** Flowers in spring, small, greenish white, in long clusters at base of younger leaves. ***Fruits single, fleshy,*** 6–8 mm (0.2–0.3 in) diameter, ***yellow to red-orange, sweet, containing 1 large seed.*** **Harvest:** fruits in late September and October; crop frequently abundant if rain was sufficient. **Preparation:** Eat fresh or dried as snack. Dry whole or mashed and formed into small cakes; dry in sun or 100°C (212°F) oven. Store in dry place. For seedy jam, mash 6 cups fresh berries in saucepan with 2 cups sugar, 3 tbsp lemon juice, and ½ cup water. Bring to boil; reduce heat and simmer with stirring until thick. Ladle into sterilized jars and seal with paraffin. If seeds are annoying, grind fruit fine before cooking. Use fresh or dried fruits in baking. **Related edible species:** American hackberry *(C. occidentalis)* and several other species are edible; palatability depends on species and weather. **Poisonous look-alikes:** none.

ROSES *Rosa*

There are about 35 species of roses native to North America and several introduced species that now grow wild on roadsides and disturbed sites. While it's easy to distinguish roses from other groups of plants, the distinctions among individual species are often difficult. Fortunately, all roses have edible parts although quality and size vary greatly. Roses are shrubs normally with spiny branches, and alternate featherlike (pinnate) compound leaves. The large attractive flowers have 5 petals, and many stamens and pistils. Four of the more common species are treated on upcoming pages. **Harvest and preparation:** Gather fruits (hips)

California rose *Rosa californica*

Also edible winter, summer

Habitat: along streams, rivers, washes, and in moist canyons. **Identification:** shrub to 3 m (9.8 ft) tall with coarse, erect to nearly erect branches, armed with stiff, flattened, usually curved spines. Leaves alternate, deciduous, compound, each with 5–7 leaflets, leaflets broadest near base and gradually tapering to a pointed to blunt tip, 1.2–3.6 cm (0.5–1.4 in) long, about half as wide, with 1 or 2 rows of sharp teeth along margin, hairy. ***Flowers in few-flowered clusters*** in spring or summer, each with green smooth floral tube (hypanthium), narrow lance-shaped sepals, ***5 white or pink petals 1.2–2.5 cm*** (0.5–1 in) ***long. Fruits*** (hips), fleshy, ***egg- to globe-shaped, 1.2–1.8 cm*** (0.5–0.7 in) ***in diameter, with distinct neck,*** red, smooth. **Harvest and preparation:** See preceding. **Related edible species:** all native roses. **Poisonous look-alikes:** none.

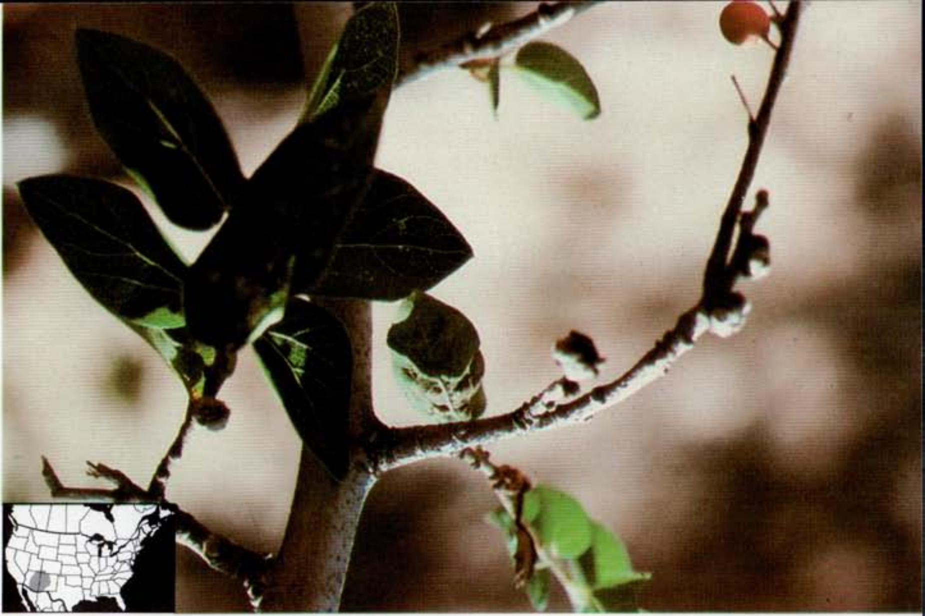

Desert hackberry: fruiting branch (Cramer).

as they ripen in autumn (after frost) or during winter, wash and remove dried persistent flower parts from top of hips, then split open and remove seeds. Eat pulpy portion fresh or in jellies or sauces. Dry whole or halved cleaned fruits for later use (soak overnight in warm water) or finely grate or grind dried hips to yield a slightly fragrant powder rich in vitamin C and essential minerals. Sprinkle on hot breakfast cereals or use to make hot tea. Also wash young leaves, cut into small pieces, and dry for hot rose tea. Flower petals can be used in candy, tea, and jellies, but fruits are more nutritious.

California rose: fruits and flower (McNeal).

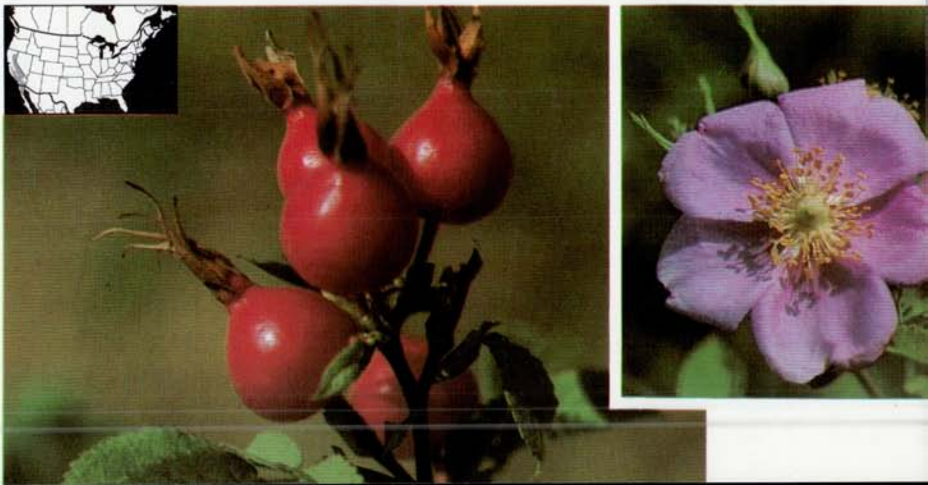

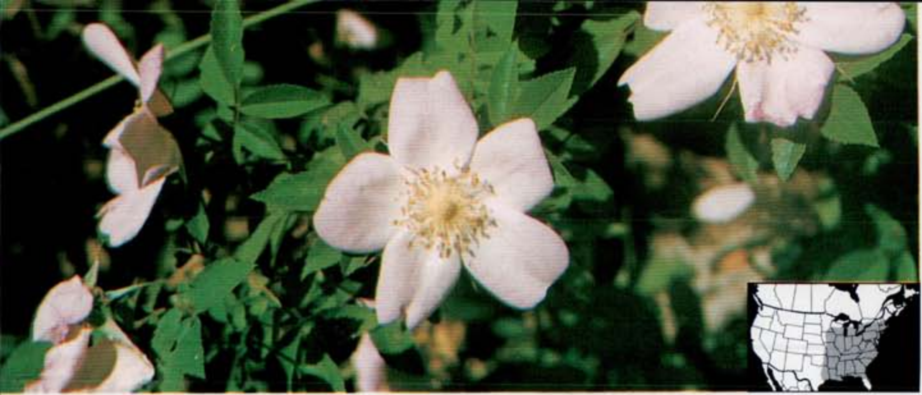

Wild rose: flowers (NYBG).

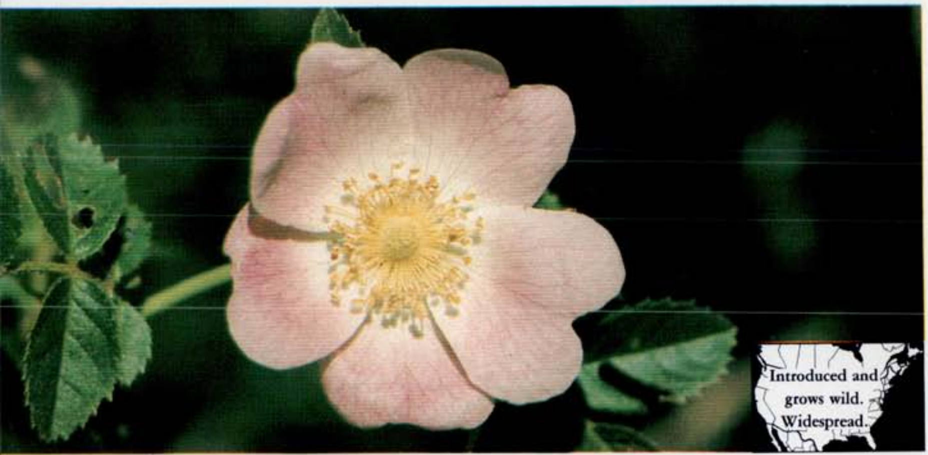
Introduced and grows wild. Widespread.

Sweetbrier rose: flower (Speas).

Wrinkled rose: fruits (Crow). Flower (Hardin).

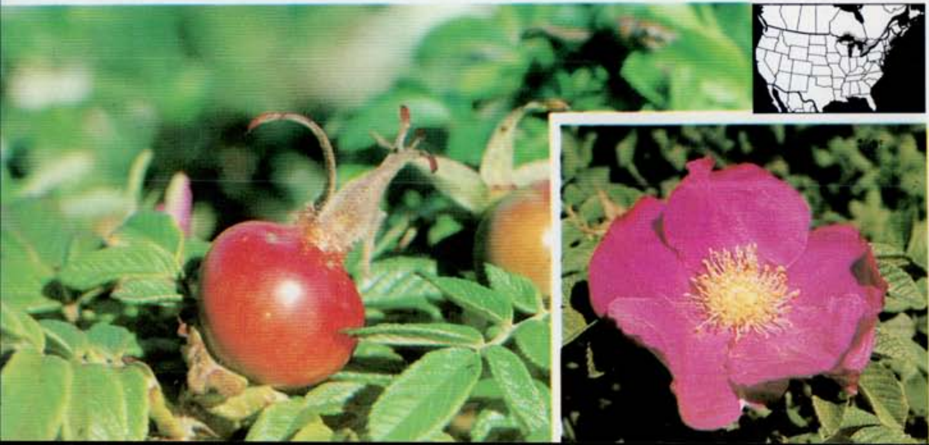

Wild rose, Pasture rose *(Also edible winter, summer)* *Rosa carolina*

Habitat: open woodlands, margins of woodlands, fields, pastures, rocky outcrops. **Identification:** low shrub to 1.5 m (5 ft) high, with slender arching branches, with stiff hairs and armed with broad-based straight or curved spines, especially on lower half. Leaves alternate, deciduous, compound, each with 3–7, rarely 9, leaflets, leaflets usually broadest near middle to almost rounded, 1.6–3.8 cm (0.6–1.5 in) long, about half as wide, pointed to blunt at tip, ***coarsely sharp-toothed along margin. Flowers*** usually solitary and produced in spring or early summer, each ***with glandular-hairy floral tube*** (hypanthium), narrow lance-shaped sepals, and 5 ***pink petals*** 2–3.2 cm (0.8–1.3 in) long. ***Fruits (hips)*** fleshy, ***nearly globe-shaped, 6–10 mm*** (0.2–0.4 in) ***in diameter,*** red, containing several seeds. **Harvest and preparation:** See ROSES. **Related edible species:** all native roses. **Poisonous look-alikes:** none.

Sweetbrier rose *(Also edible winter, summer)* *Rosa eglanteria*

Habitat: disturbed sites such as roadsides, abandoned pastures and fields, edges of woodlands, fencerows. **Identification:** shrub to 3 m (9.8 ft) high, with large arching coarse branches, new branches with laterally flattened, broad-based, spines curved or straight. Leaves alternate, deciduous, compound, apple odor when crushed, each leaf with 7–9 leaflets, ***leaflets*** broadest near or above middle, 1–4 cm (0.4–1.6 in) long, pointed or blunt at tip, ***doubly toothed and glandular along margin,*** hairy. ***Flowers*** solitary or in few-flowered clusters from May to July, each with green, ***smooth or sparsely hairy floral tube*** (hypanthium), ***featherlike*** (pinnate) ***glandular sepals,*** and 5 pink petals 1.4–2.4 cm (0.6–1 in) long. ***Fruits*** (hips) fleshy, ***elongate and broadest near middle, 1.5–2 cm*** (0.6–8 in) ***long,*** red, containing several seeds. **Harvest and preparation:** See ROSES. **Related edible species:** all native roses. **Poisonous look-alikes:** none.

Wrinkled rose *(Also edible winter, summer)* *Rosa rugosa*

Habitat: introduced from Asia; grows wild along roadsides, in dunes, and near seashores. **Identification:** shrub to 2 m (6.6 m) tall, with coarse, arching ***branches, densely covered with stiff hairs and stiff needle-shaped spines.*** Plants forming dense clumps. ***Leaves*** alternate, deciduous, compound, each leaf with 5–9 leaflets, leaflets widest near or above middle, 2–5 cm (0.8–2 in) long, about half as wide, pointed to rounded at tip, with teeth curving downward (recurved) along margin, ***dark green, deeply furrowed or wrinkled. Flowers*** solitary or in few-flowered clusters in spring, summer, or early fall, each ***with green, smooth floral tube*** (hypanthium), ***narrow spoon-shaped sepals, and 5 rose-purple to white petals*** 2.5–5 cm (1–2 in) long. ***Fruits*** (hips) fleshy, ***globe-shaped but depressed at top and bottom, 2–2.5 cm*** (0.8–1 in) ***in diameter,*** red, ***smooth,*** containing many seeds. **Harvest and preparation:** See ROSES. **Related edible species:** all native roses. **Poisonous look-alikes:** none.

Highbush cranberry *Viburnum trilobum*

Also edible winter

Habitat: stream banks, wet thickets, moist woodlands. **Identification:** Not a real cranberry, this is a ***shrub,*** upright, to 5 m (16.4 ft) high; branches upright to spreading, ash-gray with age. ***Leaves*** opposite, simple, deciduous, ***6–12 cm*** (2.4–4.7 in) ***long, almost as wide,*** broadest near base, ***deeply 3-lobed,*** long pointed at tip, coarsely toothed along margin, rounded to squared at base, dark green and sparsely hairy above, hairy beneath; ***leafstalks with 1 or 2 small glands. Flowers produced in dense, many-flowered, round-topped clusters, with showy white sterile flowers on margin of flower cluster;*** small sterile flowers in center. ***Fruits berrylike,*** rounded to egg-shaped, 8–14 mm (0.3–0.6 in) long, ***bright red, translucent*** when ripe, juicy, containing single flattened stone. **Harvest:** Fruits ripen in late summer but persist on shrub and become less sour through winter. To make sure your shrub is highbush cranberry and not its bitter look-alike, the introduced guelder rose *(V. opulus),* make a harmless taste test. Pick in fall and winter. **Preparation:** much the same as CRANBERRY except large seeds must be removed. While cooking add lemon or orange peel shavings to eliminate bad odor of berries. For sauce, cover several quarts of berries with water in kettle. Add outer peels of 2 lemons or oranges. Bring to boil and simmer 5 min. Mash thoroughly and simmer another minute or two. Force juice and pulp through coarse strainer to remove seeds. Add 1 pkg pectin to 4 cups strained fruit, bring to boil, and add 5 cups sugar. Boil again for 1 min and seal in sterile jars or refrigerate. If fruits are picked before first frost, pectin may not be needed. For juice and jelly, cook berries as above but strain through jelly bag. For juice, dilute and sweeten to taste. For jelly, proceed with 4 cups juice as for sauce. **Related edible species:** HOBBLEBUSH, NANNYBERRY, and SQUASHBERRY (next), and in the Pacific Northwest the cranberry tree *(V. pauciflorum).* **Poisonous look-alikes:** none.

Hobblebush *Viburnum alnifolium*

Habitat: stream banks, moist woodlands, ravines. **Identification:** ***shrub,*** often spreading, to 3 m (9.8 ft) high, ***branches spreading, some drooping and rooting in ground. Leaves*** opposite, simple, deciduous, 10–24 cm (4–9.4 in) long, almost as wide, ***broadest near base,*** short-pointed at tip, heart-shaped at base, finely double toothed along margin, hairy beneath. ***Flowers produced in many-flowered, round-topped clusters, with sterile flowers, each with showy pink or white petals 2–3.8 cm (0.8–1.5 in) wide on margin of flower cluster;*** tiny fertile flowers in center. ***Fruits berrylike, slightly flattened,*** egg-shaped, 6–12 mm (0.2–0.5 in) long, ***black,*** containing single flattened stone. **Harvest:** fruits in autumn, generally only on shrubs in open area, not in dense woods. **Preparation:** same as NANNYBERRY (next). **Related edible species:** See NANNYBERRY. **Poisonous look-alikes:** none.

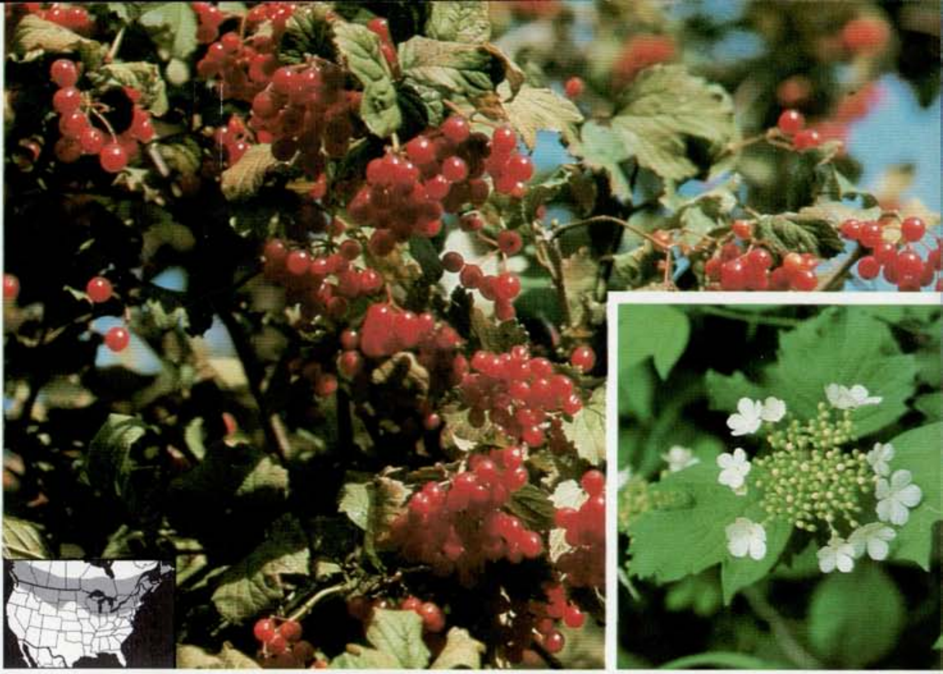

Highbush cranberry: fruiting branches (Elias). Flower cluster (Stevenson).

Hobblebush: flower cluster (NYBG). Fruiting branch (Elias).

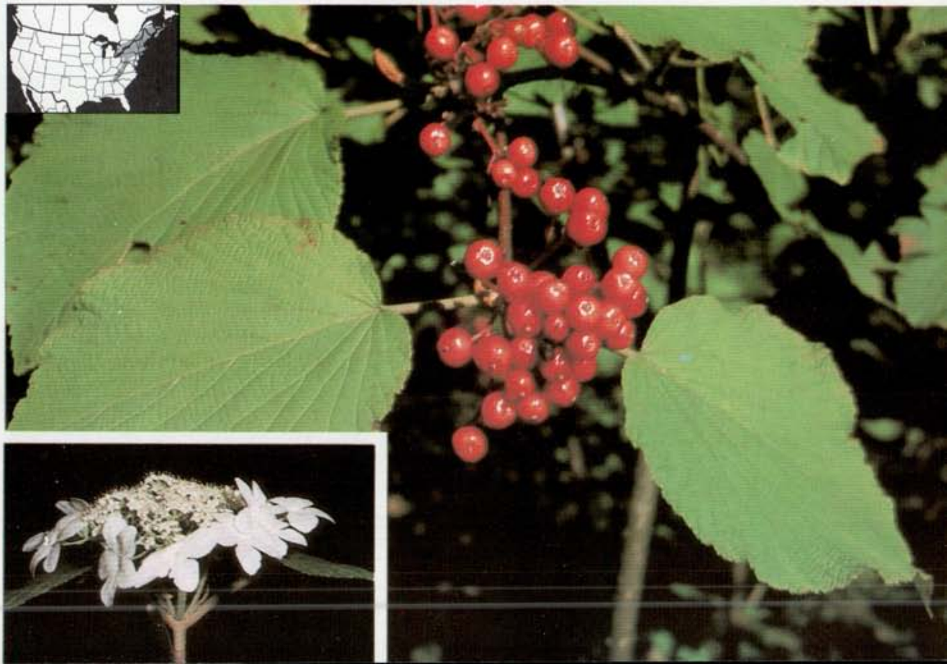

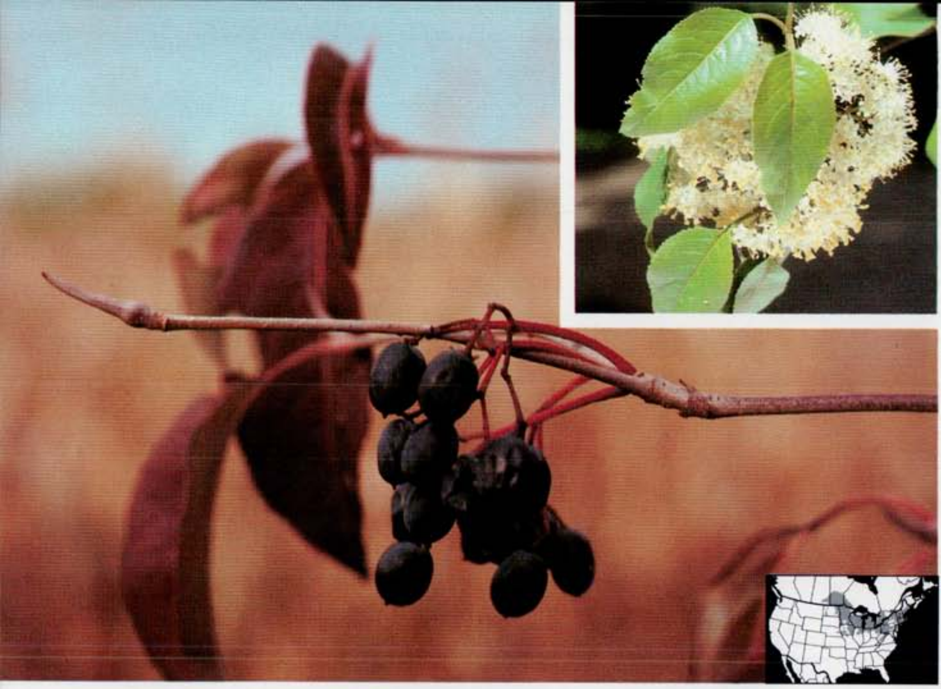

Nannyberry: fruits and flower cluster (Dykeman).

Squashberry: fruiting branch (Clawson).

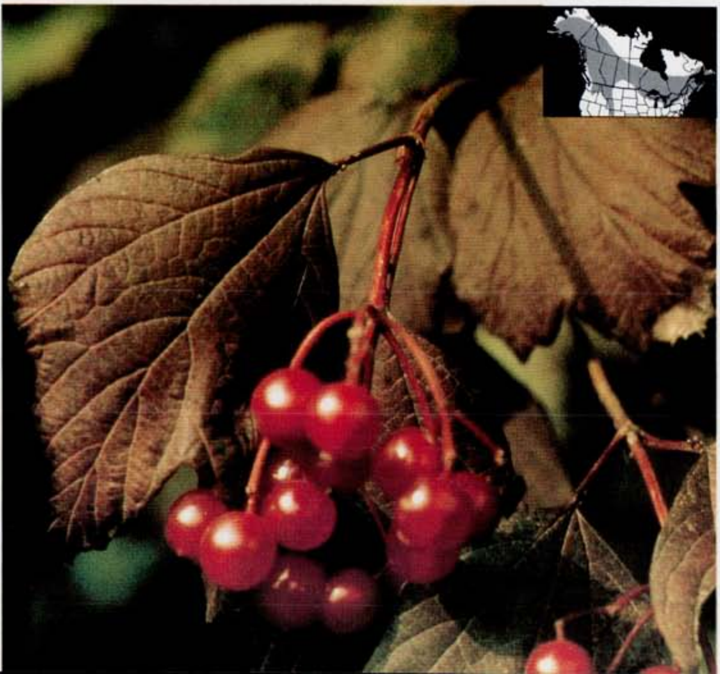

Nannyberry *Viburnum lentago*

Habitat: swamps, stream banks, rocky hillsides, margins of woodlands. **Identification:** ***shrub*** or small tree to 10 m (33 ft) with short branching trunk and slender tough, dark reddish-brown branches. ***Leaves opposite,*** simple, deciduous, 5–10 cm (2–4 in) long, 2.5–4 cm (1–1.6 in) wide, ***broadest near base or middle,*** occasionally rounded, with many fine, sharp teeth along margin, bright shiny green and smooth. ***Flowers produced in dense, many-flowered, rounded heads to 12 cm (4.7 in) in diameter,*** each flower white to cream, 5-lobed. ***Fruits berrylike, rounded,*** 8–12 mm (0.3–0.5 in) in diameter, ***dark blue to black,*** sweet, juicy, containing single flattened seed. **Harvest:** fruits when ripe in autumn as leaves are falling. **Preparation:** fruits good for trail snacks; flavor excellent, seeds a nuisance. For pudding add 1½ cups water to 1 qt fruits, boil for 5 min. Mash and work pulp through food mill. If pulp remains with seeds, scrape residue into small pan and add ½ cup water. Stir, mash, mill again. Yields about 1 cup pulpy syrup. Add 1 tbsp lemon rind, juice of ½ lemon, 1 cup sugar, ⅛ tsp cinnamon, ¾ cup water. Bring to boil and simmer 15 min. In separate pan combine ⅓ cup cornstarch with 1 cup cold water. Blend into fruit mixture. Cook slowly 15 min, pour into mold, and cool. Serve plain or with whipped cream. **Related edible species:** black haw *(V. prunifolium)*, northern wild raisin *(V. cassinoides)*, and HOBBLEBUSH (preceding); use fruits in same ways as for nannyberry. **Poisonous look-alikes:** none.

Squashberry *Viburnum edule*

Also edible winter, summer

Habitat: stream banks, wet thickets, margins of woodlands, moist low woodlands. **Identification:** ***shrub,*** often sprawling, to 2.5 m (8 ft) high; stems usually spreading, smooth. Leaves opposite, simple, deciduous, 4–11 cm (1.6–4.3 in) long and wide, ***broadly circular in outline, with 3 shallow lobes,*** sharply and irregularly toothed along margin, squared to slightly heart-shaped at base, sometimes hairy beneath. ***Flowers produced in loose, round-headed clusters 4–6 cm (1.6–2.4 in) across,*** each flower milky white, 5-lobed. ***Fruits berrylike, rounded,*** 6–10 mm (0.2–0.4 in) in diameter, ***red,*** containing single flattened seed. **Harvest:** fruits similar to HIGHBUSH CRANBERRY but smaller and less acid. Pick in late summer through winter. **Preparation:** same as HIGHBUSH CRANBERRY, except juice does not require dilution for use as beverage. **Related edible species:** See HIGHBUSH CRANBERRY. **Poisonous look-alikes:** none.

Pinyon pine *Pinus edulis*

Habitat: mesas, plateaus, lower mountain slopes, and foothills from 825–2,350 m (2,700–7,700 ft) elevation. Often occurs with JUNIPERS, GAMBEL'S OAK, and ponderosa pine. **Identification:** small, spreading, bushy tree with thin, irregularly furrowed, scaly ridged, gray to reddish-brown bark. ***Leaves (needles) 2 per bundle*** (rarely 1 or 3), ***1.2–5 cm*** (0.6–2 in) ***long,*** stiff, sharp-pointed, curved inward toward branches, bluish-green when young, turning yellowish-green with age. Fruits hard, short, egg-shaped cones 3.8–5 cm (1.5–2 in) long, composed of thick, blunt, yellowish-brown scales, each scale with 2 seeds. ***Seeds egg-shaped, 1.2–1.4 cm*** (0.5–0.6 in) ***long, wingless,*** pale yellow with reddish-brown markings, oily, sweet. **Harvest and preparation:** large cone crops produced every 3–4 yr. Gather ripe, almost open, cones in autumn. Store dry so that scales will separate to release seeds. Eat seeds fresh or after roasting 20–30 min. Roasting dries cones, facilitates seed removal, improves nutty flavor of seeds, and permits longer storage. Grind roasted nuts into meal. Seeds are rich in protein (14.5%) and fats and have high caloric content. **Related edible pines:** SINGLELEAF PINYON PINE, Mexican pinyon *(P. cembroides),* Parry pinyon *(P. quadrifolia)*. **Poisonous look-alikes:** none.

OAKS *Quercus*

This large genus of about 55 native trees and shrubs north of Mexico offers important food sources to wildlife and provides nuts edible to people as well. North American oaks are divided into two subgenera: the white oak group and the red/black oak group.

White Oak Group *Quercus* subgenus *Quercus*

This large group of related oaks occurs in the Americas, Europe, and Asia. There are 32 native species north of Mexico. They are distinguished from other oaks by a combination of characteristics: leaves entire, toothed, or lobed but never bristle-tipped; acorns maturing at end of first growing season on current year's twigs, and inner surface of acorn shell lacking hairs. Four of the most common species are discussed on upcoming pages. **Harvest and preparation:** Gather acorns in autumn as they turn brown and fall. They are usually sweet and need no leaching to remove tannins (if any bitterness exists, process as NORTHERN RED OAK acorns). Store nuts in cool dry place or shell for immediate use. For grits, dry nutmeats in air or warm oven and grind coarsely. Use like chopped nuts. For meal, regrind grits to finer texture and redry in warm oven, stirring occasionally. Use in recipes under RED OAK group. For candy combine 2 cups sugar, 1 cup water, ⅛ tsp cream of tartar, and pinch of salt in small pan and boil until carmelization begins. Maintain liquid by placing in larger pan of boiling water. Dip dried whole nutmeats in syrup and place on wax paper to harden.

White oak *Quercus alba*

Habitat: mixed deciduous forest in sandy, gravelly, loamy soils; widespread, often growing with hickories or other oaks. **Identification:** medium-size to tall

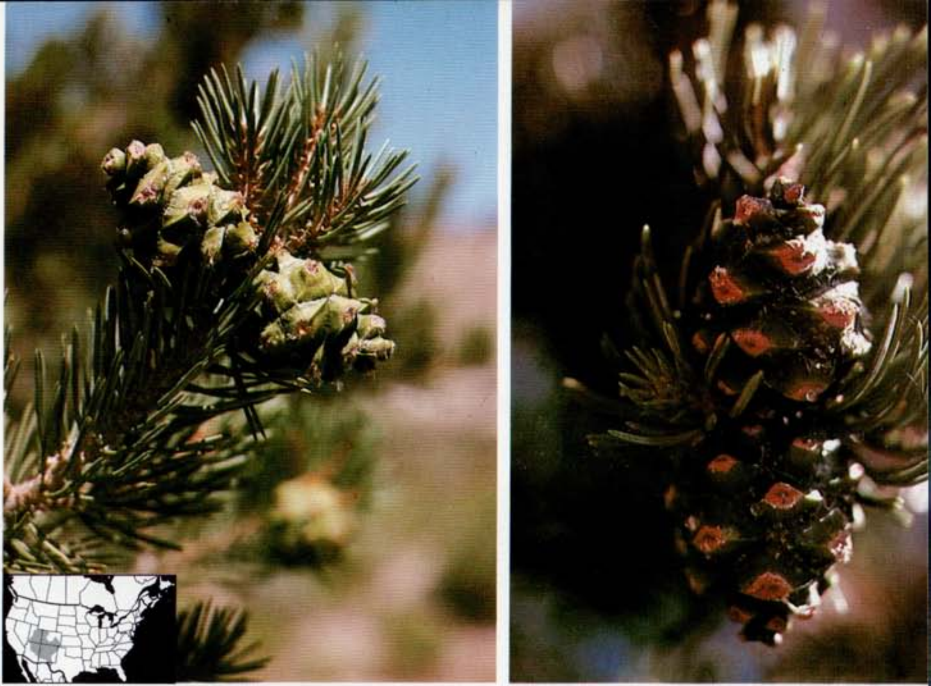

Pinyon pine: immature cones (Pawloski).

White oak: immature acorns (Elias).

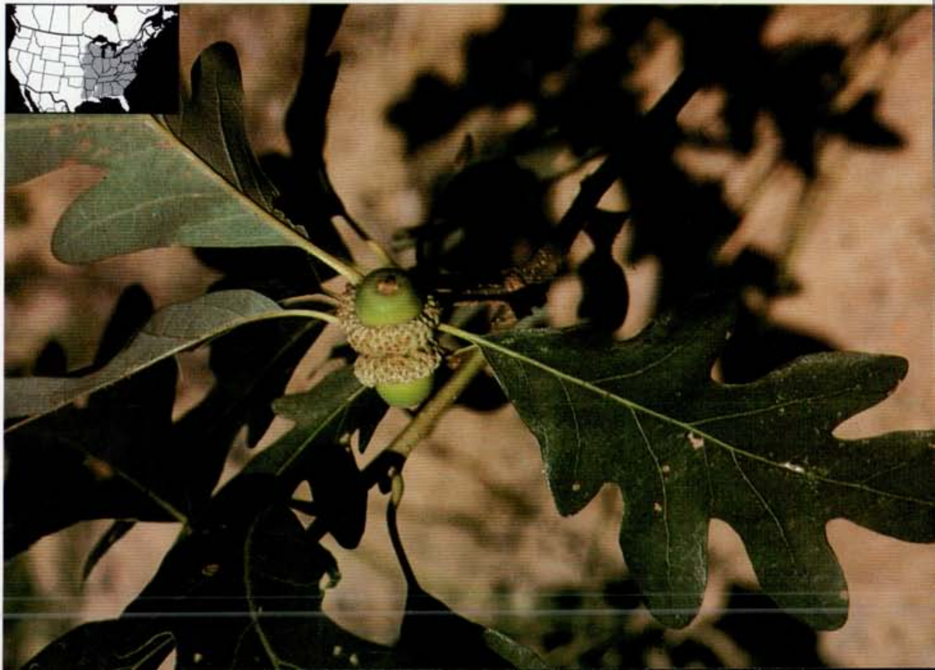

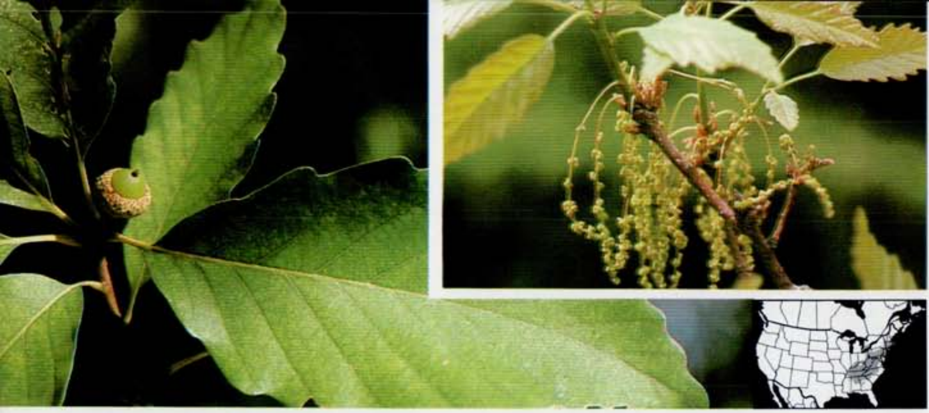

Chestnut oak: immature acorn and flowering branch (Elias).

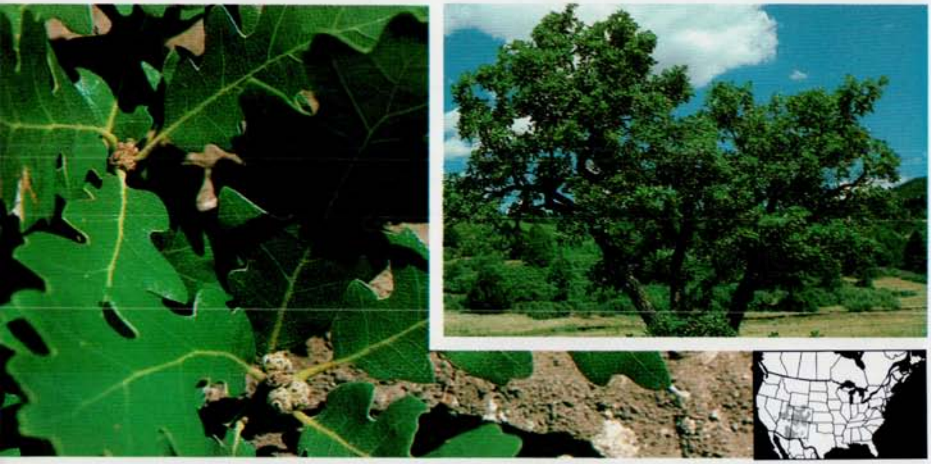

Gambel's oak: fruiting branch and tree (Dykeman).

Blue oak: mature acorns with dried leaves (Dykeman). Normal blue oak form foreground; weeping form background (Kruckeberg).

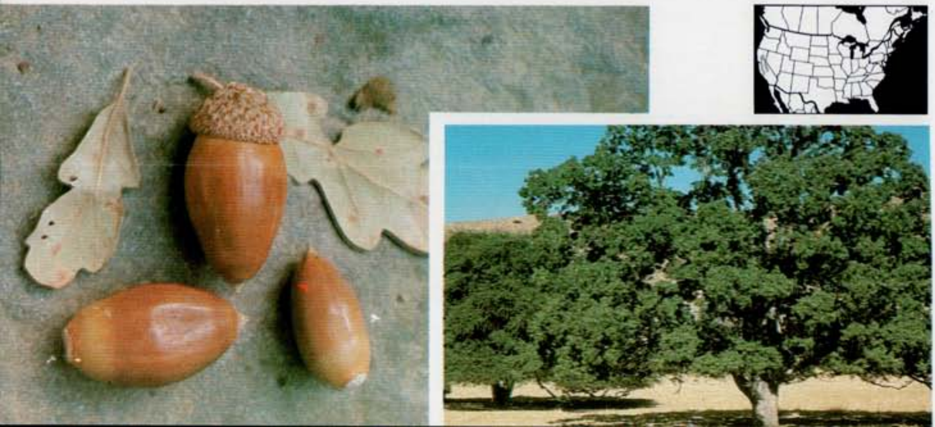

tree with broad round crown and thick, shallow to deeply furrowed, light ash-gray bark. ***Leaves*** alternate, deciduous, simple, ***12–22 cm (4.7–8.7 in) long, 5–10 cm (2–4 in) wide, with 7–10 shallow to deeply rounded lobes along margin, bright green,*** smooth. Fruits are ***acorns*** on current year's twigs, ***1.2–2 cm (0.5–0.8 in) long,*** widest near middle or base, light brown at maturity, ***acorn cup bowlshaped, 8–12 mm (0.3–0.5 mm) deep, enclosing about ¼ of nut,*** cup scales thick, warty; nutmeat sweet. **Harvest and preparation:** See WHITE OAK GROUP (preceding).

Chestnut oak *Quercus prinus*

Habitat: mixed deciduous woodlands, especially dry to moist well-drained slopes and ridges. **Identification:** medium-size tree with broad, open, rounded crown and thick, ***deeply furrowed,*** dark brown to reddish-brown ***bark. Leaves*** alternate, deciduous, simple, ***12–25 cm*** (5–10 in) ***long, 3.6–7.5 cm*** (1.4–3 in) ***wide,*** broadest above middle, ***coarsely and irregularly round-toothed along margin,*** greenish-yellow. Fruits are ***acorns 2.5–3.8 cm*** (1–1.5 in) ***long,*** uniformly wide to broadest near base, acorn ***cup deeply bowl-shaped to top-shaped, enclosing about half of nut;*** cups scales thick, warty; nutmeat sweet. **Harvest and preparation:** See WHITE OAK GROUP (preceding).

Gambel's oak *Quercus gambelii*

Habitat: dry foothills, canyons, lower mountain slopes, usually from 1,350–2,800 m (4,400–9,200 ft) elevation. **Identification:** shrub or small tree with rounded crown and thin, light gray to grayish-brown, rough, scaly bark. ***Leaves*** alternate, deciduous, simple, ***8–16 cm*** (3.2–6.3 in) ***long,*** 4.5–7.6 cm (1.8–3 in) wide, broadest above middle to uniformly wide, ***with 5–9 moderately deep to deep lobes along margin,*** leathery, ***yellowish-green.*** Fruits are ***acorns*** on current year's twigs, ***2–2.5 cm*** (0.8–1 in) ***long,*** nearly globe-shaped, ***acorn cup*** shallow to deeply bowl-shaped, ***enclosing ¼ to ⅓ of nut,*** scales thick; nutmeat sweet. **Harvest and preparation:** See WHITE OAK GROUP (preceding).

Blue oak *Quercus douglasii*

Habitat: interior valleys, rolling hills, lower mountain slopes to 1,350 m (4,400 ft) elevation. **Identification:** small to medium-size tree with rounded crown and thick, checkered, light gray bark. ***Leaves*** alternate, deciduous, simple, 3.8–10 cm (1.5–4 in) long, 1.9–5 cm (0.7–2 in) wide, uniformly wide to broadest above middle, ***entire to shallowly and irregularly lobed along margin, bluish-green.*** Fruits are acorns on current year's branches, 1.6–3.1 cm (0.6–1.2 in) long, broadest near base; ***acorn cup saucer-shaped, enclosing only base of nut,*** cup scales small, warty; nutmeat somewhat sweet. **Harvest and preparation:** See WHITE OAK GROUP (preceding).

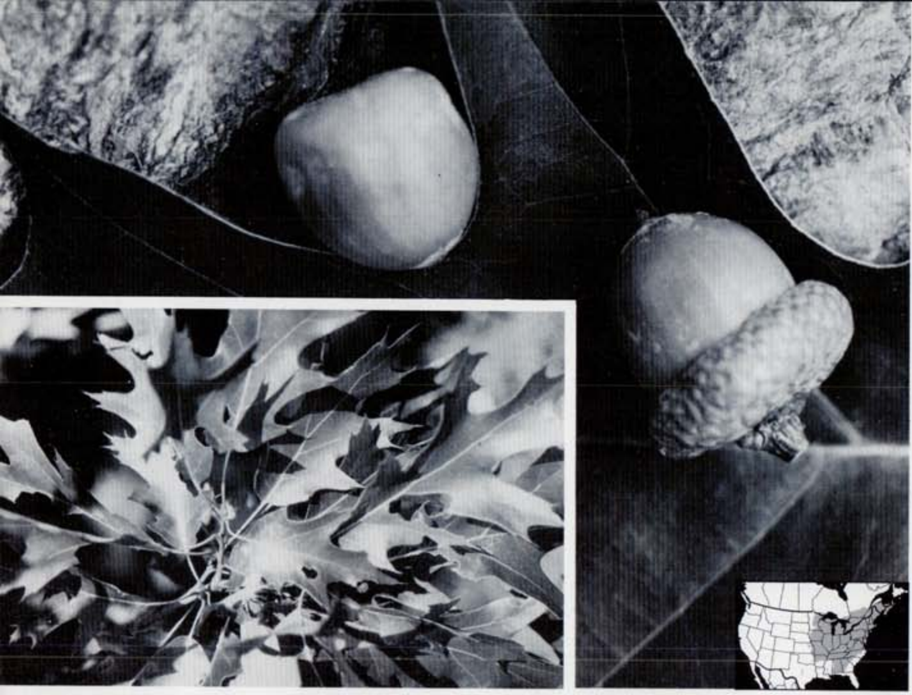

Northern red oak: acorn with cap and without. Summer branch. (Elias photos)

American beech: fruiting branch (Elias).

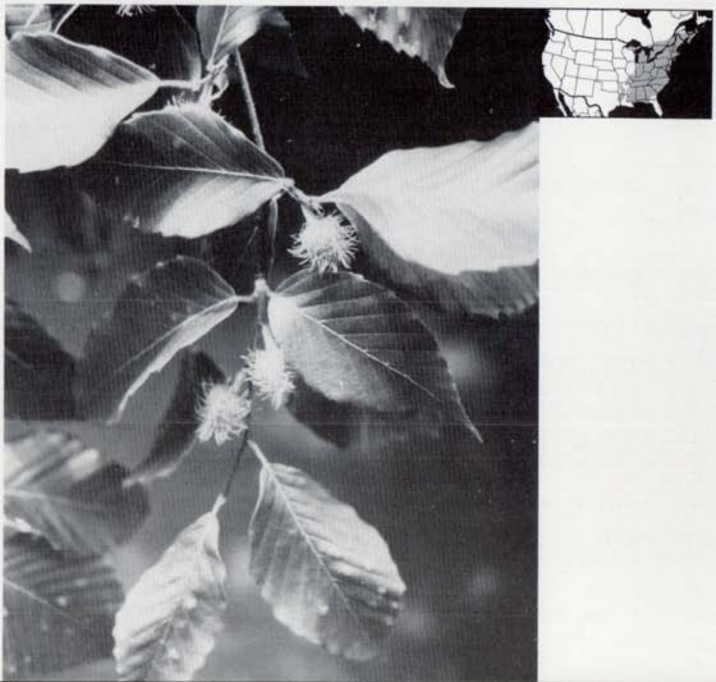

Red/Black Oak Group *Quercus* subgenus *Erythrobalanus*

The red oak species and black oak species and the numerous relatives in this subgenus occur only in the Americas. This group is characterized by leaves entire, toothed, or lobed and always with bristlelike tips extending beyond the margin of the leaf; acorns maturing at end of second year on the previous year's twigs, and the inner surface of the acorn shell hairy. Acorn nutmeats are nearly always bitter. Because of this, only 1 species is discussed here as edible, although most species can be utilized after the bitter properties have been removed.

Northern red oak *Quercus rubra*

Habitat: valleys, ravines, lower and mid slopes of hills and mountains. **Identification:** medium-size tree to 20 m (66 ft), rarely taller, with open, irregular crown and shallowly furrowed, checkered bark. ***Leaves*** alternate, deciduous, simple, ***10–22 cm*** (4–8.7 in) ***long,*** 6–12 cm (2.4–5 in) wide, widest above middle, ***with 7–9 moderately deep lobes,*** dull green and smooth above, paler beneath and with occasional tufts of hairs in junction of principal veins. Fruits are ***acorns on previous year's twigs, 1.2–2.5 cm (0.5–1 in) long, acorn cup*** saucer-shaped, ***enclosing about 1/4 of nut,*** cup scales thin, overlapping; nutmeat bitter. **Harvest and preparation:** Gather acorns in autumn as they turn brown and fall. Store in cool dry place or shell for immediate use. Place whole, chopped, or coarsely ground nutmeats in cloth (clean T-shirt will do) and tie with string. Place in boiling water until water turns brown, drain, add more boiling water. Repeat many times until water remains clear. In the field, place bag filled with nutmeats in clear running stream until nutmeats are no longer bitter, from 1 to several days. Spread leached acorns in pan and dry in sun or warm oven. Use whole nutmeats or prepare and use as grits or meal as described under WHITE OAK GROUP. For acorn muffins, cream ¼ cup butter with ½ cup sugar or molasses. Add alternately and blend 2 eggs, 1 cup milk, and dry mix of 1 cup acorn meal or grits, 2 cups flour, 4 tsp baking powder, ¾ tsp salt. Beat thoroughly. Bake in greased muffin pan at 200°C (390°F) for about 20 min. For pancake batter, repeat muffin recipe but add 1 more egg plus milk. For acorn coffee, roast acorn meal at 175°C (350°F) until dark and crisp, about 30 min. Use ½ cup roasted meal to 4 cups water. Combine, boil 15 min, strain, and serve.

American beech *Fagus grandifolia*

Habitat: river valleys, lower mountain slopes, mixed deciduous forests, usually in rich, moist soils. **Identification:** medium-size to tall tree with narrow to spreading, rounded crown and ***thin, smooth, light gray to bluish-gray bark.*** Leaves alternate, deciduous, simple, 6–14 cm (2.4–5.5 in) long, 2.5–7.2 cm (1–3 in) wide, broadest just below middle, sharply toothed along margin, leathery texture at maturity. Male flowers tiny, in globe-shaped, hanging heads; female flowers tiny, usually paired at or near tips of branches. ***Fruits consist of 2 triangular nuts enclosed in 2 spiny bracts, bracts 1.6–2.2 cm (0.6–1.9 in) long,***

opening in late summer to release oily nuts. **Harvest and preparation:** Gather nuts as they drop from trees following hard frost. Separate burrs from nuts by shaking and winnowing or singly by hand. Dry in open, warm area. Crack brown shells by heating in oven; remove shells by beating or rubbing in hands and winnowing. Nutmeats are edible raw. Grind and air dry nutmeats for flour; crush and boil in water; skim off excellent oil for cooking. Roast nutmeats at 160°C (320°F) 30–40 min and grind for coffee substitute. For beechnut pie, substitute 1½ cups beechnuts for pecans in PECAN pie recipe. Thoroughly whip 3 eggs. Slowly beat into them ½ cup melted butter or margarine, 1 cup white corn syrup, and ¾ cup sugar. Turn into unbaked 9-in pie shell and bake at 150°C (300°F) 35 min. Remove pie long enough to cover quickly with nuts. Return to oven, turn up heat to 175°C (350°F), and bake another 15 min. **Edible related species:** none. **Poisonous look-alikes:** none.

Allegheny chinkapin *Castanea pumila*

Also edible summer

Habitat: widespread; dry woodlands and mountain slopes or along coast in dry sandy soils. **Identification:** shrub or small tree to 8 m (26 ft) with a rounded crown and light reddish-brown bark. Twigs densely hairy when young. ***Leaves*** alternate, simple, deciduous, ***6.8–16 cm*** (2.7–6.3 in) ***long, 3.8–5 cm*** (1.5–2 in) wide, ***broadest near middle, with numerous large teeth along the margin,*** whitish and hairy beneath. Male flowers numerous on the upper ⅔ of semi-erect spikes; female flowers fewer on the lower ⅓ of flowering spike. ***Fruits are spiny burrs 3–4 cm (1.4 in) in diameter, enclosing 1, rarely 2, nuts;*** nuts broadest near the base and pointed at tip, nutmeat sweet. **Harvest and preparation:** Gather nuts as they ripen in late summer or early autumn. Separate nuts from burs and store whole, or shell, blanch, and dry nutmeats. To shell and blanch the nuts, cover with boiling water and boil until shell and skin can be easily removed. Then place nutmeats in shady place about 1 week until dry and shriveled. Store shelled and dried or unshelled nuts in porous bag in dry place. Before using nuts, soak in water for 30 min. Use chestnuts in stuffings, soups, desserts, as a vegetable, and as a coffee substitute. Nuts can be eaten directly by slitting shells and roasting over coals or in hot oven. For soup, combine 1½ cups dried ground chestnuts, 2 sliced carrots, 1 cup sliced acorn or butternut squash, 2 tbsp chopped fresh parsley, ½ tsp oregano or native plant seasonings, and 5 cups water. Cook until tender, about 40–50 min. For coffee, chop freshly shelled chestnuts, spread on foil, and roast at 190°C (375°F) until dark and dry. Grind to a powder as needed. Boil gently 28 grams (1 oz) powder per cup of water for 15 min. Strain and serve. **Related edible species:** American chestnut *(C. dentata)* produces large sweet nuts but the persistent chestnut blight has literally eliminated it from its native range. **Poisonous look-alikes:** Do not confuse poisonous horsechestnuts with true chestnuts. Horsechestnuts (*Aesculus* species) have compound (palmate) leaves and nuts with a large circular scar, and belong to the family Hippocastanaceae. Allegheny chinkapin and true chestnuts have simple leaves and belong to the beech family (Fagaceae).

Allegheny chinkapin: fruiting branch (Elias). Flowering branch (NYBG).

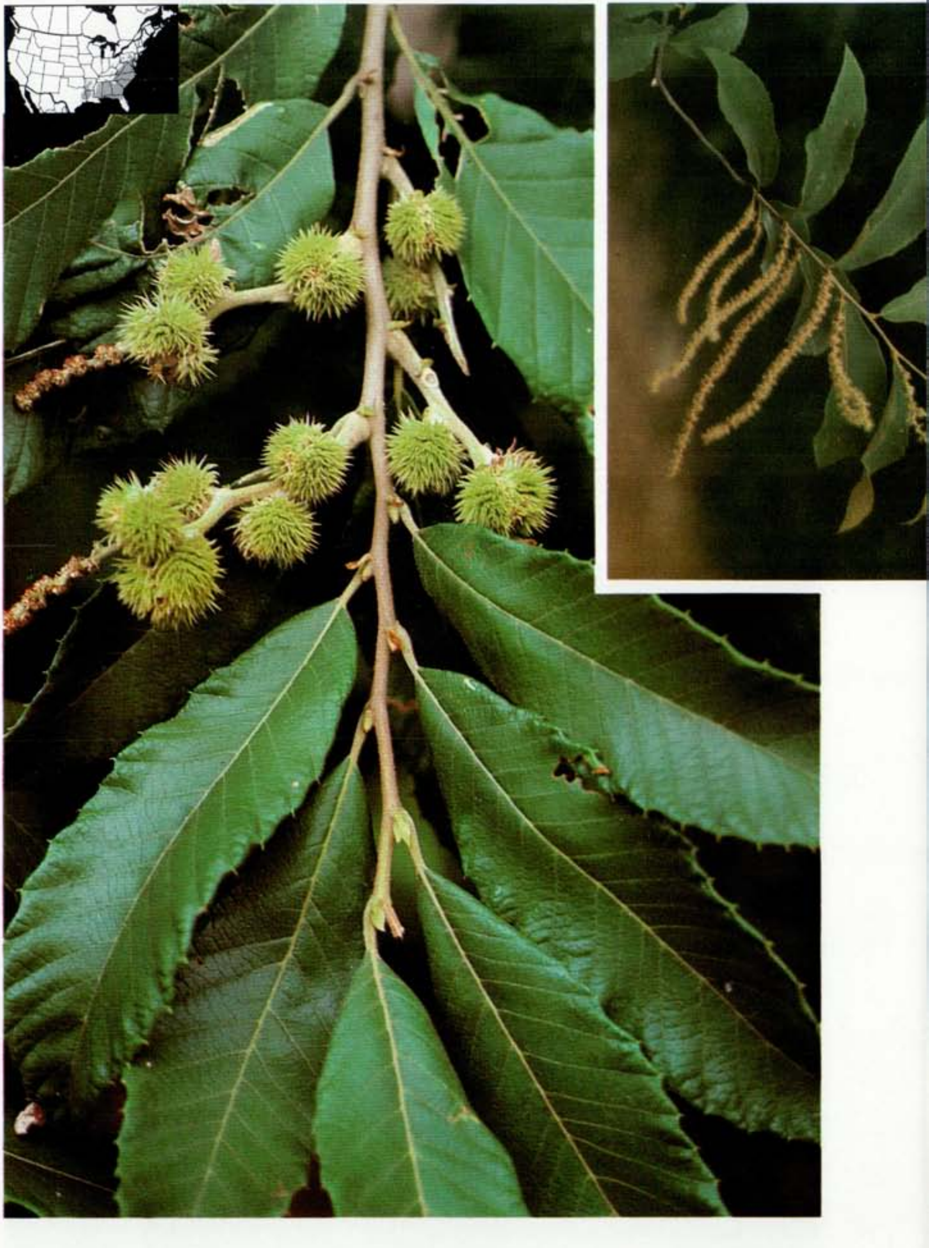

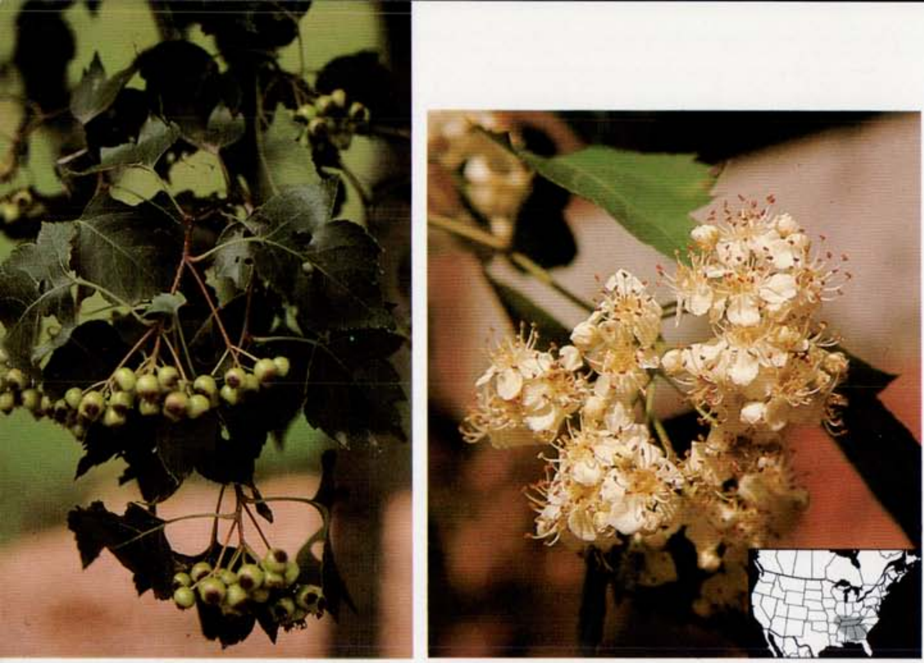

Washington hawthorn: unripe fruits (Elias). Flower clusters (Feldman).

Downy hawthorn: fruiting branch (Speas).

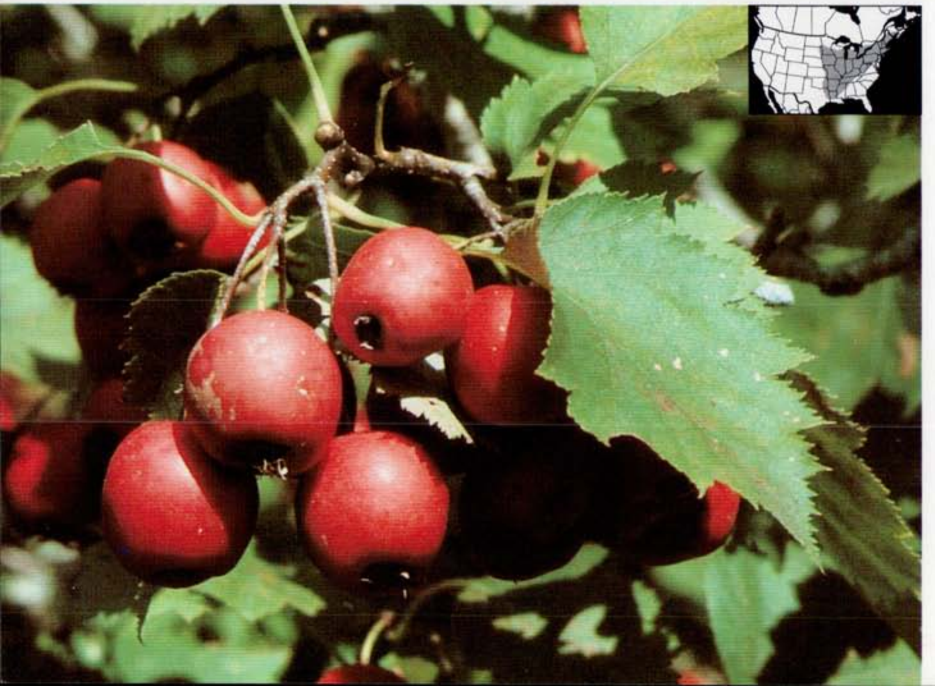

HAWTHORNS *Crataegus*

This large group of trees and shrubs contains highly variable species that are often difficult to distinguish. Nearly all of the 26 distinguishable North American species of hawthorns have edible fruits. They are mostly small spreading trees with dense branches often armed with long sharp spines. The toothed leaves may be entire, shallowly or deeply lobed. Bisexual flowers are generally white, with 5 petals and the rounded to pear-shaped fruits yellow, green, red, orange, or nearly black at maturity. Four common species are discussed on upcoming pages. **Harvest:** Most fruit ripens in autumn and may persist for several months, but some southern species, including May hawthorn *(C. aestivalis)*, have fruit ripening in late spring. The fruits vary greatly in flavor from sweet to bitter. Sample the fruits from each plant and harvest the best, but be careful of the vicious thorns. **Preparation:** Eat sweetest fruits fresh. For jelly, crush 1.4 kg (3 lb) fruit or 1 kg (2 lb) if a juicy type, cover with water, and cook until soft. Add more water if necessary. Strain juice through jelly bag. Bring 4 cups juice to boil. Unless fruits are unripe, add 1 pkg pectin, stir, add 6–7 cups sugar, stir, and bring to rolling boil. If no pectin is added, boil to 115°C (239°F). Remove from heat, skim, pour into sterilized jelly glasses, and seal. If fruits do not seem acid enough, add juice of 1 or 2 lemons. For tea, steep 2–3 tbsp crushed fruits and sprig of mint in cup of boiling water for 5–10 min. Fruits can be dried and stored for winter use. Substitute hawthorn fruits in recipes calling for rose hips. The fruits were used by Indians to some extent in pemmican. **Related edible species:** many *Crataegus* species. **Poisonous look-alikes:** none.

Washington hawthorn *Crataegus phaenopyrum*

Habitat: stream banks, old fields, openings in woodlands. **Identification:** small tree to 12 m (40 ft), with rounded, spreading crown; branches dark reddish-brown, zigzagging, ***armed with slender sharp-pointed spines 3–5 cm*** (1.2–2 in) ***long. Leaves*** alternate, simple, deciduous, 2–6 cm (0.8–2.4 in) long, 2–5 cm (0.8–2 in) wide, broadest near base, ***with 1–4 pairs of shallow lobes,*** pointed at tip, broadly rounded at base, sharply toothed along margin, dark green, shiny, smooth. ***Flowers white, attractive,*** produced in branched, spreading many-flowered clusters. ***Fruits*** are ***dry, berrylike, globe-shaped,*** 5–8 mm (0.3 in) in diameter, ***bright red.*** **Harvest and preparation:** See HAWTHORNS (preceding).

Downy hawthorn *Crataegus mollis* (syn.: *C. arnoldiana, C. canadensis, C. submollis*)

Habitat: along streams and rivers, bottomlands, lowland fields, clearings. **Identification:** small tree to 12 m (40 ft) high, with dense, wide-spreading crown; ***branches*** covered with dense white hairs when young, ***almost thornless,*** occasionally with sharp, slender dark reddish-brown spines to 5 cm (2 in) long. ***Leaves*** alternate, simple, deciduous, 3.5–10 cm (1.4–4 in) long, 3–8 cm (1.2–3.2 in) wide, usually broadest near base, sharp-pointed along margin, ***dark yellowish-green, hairy on lower surface.*** Flowers white, attractive, produced in branched, spreading clusters in spring. Fruits stalked, dry, berrylike, globe-shaped, 9–16 mm (0.4–0.6 in) in diameter, red. **Harvest and preparation:** See HAWTHORNS (preceding).

Fleshy hawthorn *Crataegus succulenta*

Habitat: along coast, hillsides, abandoned fields, usually in rocky or gravelly soils. **Identification:** shrub or small tree to 8 m (26 ft) high, with irregular, usually rounded, crown; branches stout, slightly zigzagging, gray, ***armed with frequent, stiff, slightly curving, dark reddish-brown spines*** 3–4.5 cm (1.2–1.8 in) long. ***Leaves*** alternate, simple, ***3–7 cm*** (1.2–2.8 in) ***long, 2–5.2 cm*** (0.8–2 in) ***wide,*** broadest above or below middle, ***shallowly lobed near tip, pointed*** at tip, tapering at base, ***sharply pointed*** and ***double-toothed along margin,*** dark green, shiny. Flowers white, attractive, produced in branched, spreading clusters in spring. ***Fruits*** stalked, dry, berrylike, ***globe-shaped, 0.8–1.6 cm*** (0.3–0.6 in) ***in diameter, bright red.*** **Harvest and preparation:** See HAWTHORNS (preceding).

Cockspur hawthorn *Crataegus crus-galli*
(synonyms: *C. acutifolia, C. canbyi, C. regalis, C. pyracanthoides*)

Habitat: hillsides, lower mountain slopes, river valleys, abandoned fields. **Identification:** small tree to 8 m (26 ft) high, with broad spreading, ***nearly flat-topped crown;*** branches stiff, light brown to gray, armed with stout, slender, straight to slightly curved, sharp-pointed spines. ***Leaves*** alternate, simple, deciduous, ***2–6 cm*** (0.8–2.4 in) ***long, 1–3.2 cm*** (0.4–1.3 in) ***wide, usually broadest above middle*** or near middle, ***unlobed,*** rounded to pointed at tip, gradually tapering at base, sharply toothed along upper ⅔ of margin, ***glossy, dark green. Flowers white, pink, or red,*** attractive, produced in branched, spreading clusters in spring. ***Fruits*** stalked, ***dry, berrylike,*** globe-shaped, ***0.7–1 cm*** (0.4 in) ***in diameter, dull red to green.*** **Harvest and preparation:** See HAWTHORNS (preceding).

Common persimmon *Diospyros virginiana*
Also edible winter

Habitat: rich bottomlands, especially Mississippi River Valley, margins of woodlands and fields, rocky hillsides, along fencerows and highways. **Identification:** large shrub or small to medium-size tree with thick dark gray to grayish-brown bark that forms small square blocks. ***Leaves*** alternate, deciduous, ***7–13 cm*** (2.8–5.1 in) ***long, 3.5–8 cm*** (1.4–3.2 in) ***wide,*** broadest near base or middle, entire, dark green and shiny. ***Male and female flowers on separate trees,*** produced at bases of leaves along younger branches, ***each flower tubular or urn-shaped,*** greenish-yellow to cream-white. ***Fruits are globe-shaped, 2–6 cm*** (0.8–2.4 in) ***in diameter, orange to orange-purple, fleshy,*** containing a few flattened seeds 1.2–2 cm (0.5–0.8 in) long. **Harvest:** Fruits mature in early autumn and persist on trees into early winter. Best collected after first frost when soft and skin becomes slightly wrinkled. Prior to that fruits will be very sour. To obtain many fruits, shake branches. **Preparation:** Delicious as fresh fruit. The pulp, strained to remove seeds and skin, can be used to make bread, pie, pudding, jam, even frozen for later use. See jam recipe early in book. A tea can be made from clean dried leaves. **Related edible species:** Texas persimmon *(D. texana)*. **Poisonous look-alikes:** none.

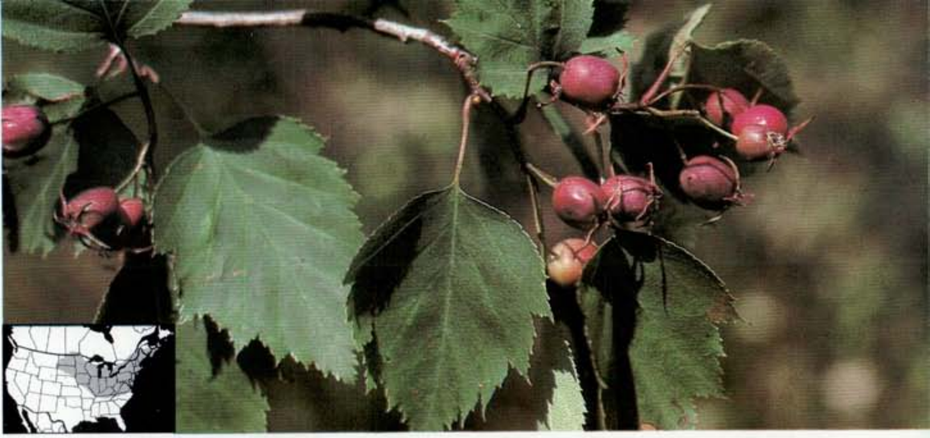

Fleshy hawthorn: fruiting branch (Elias).

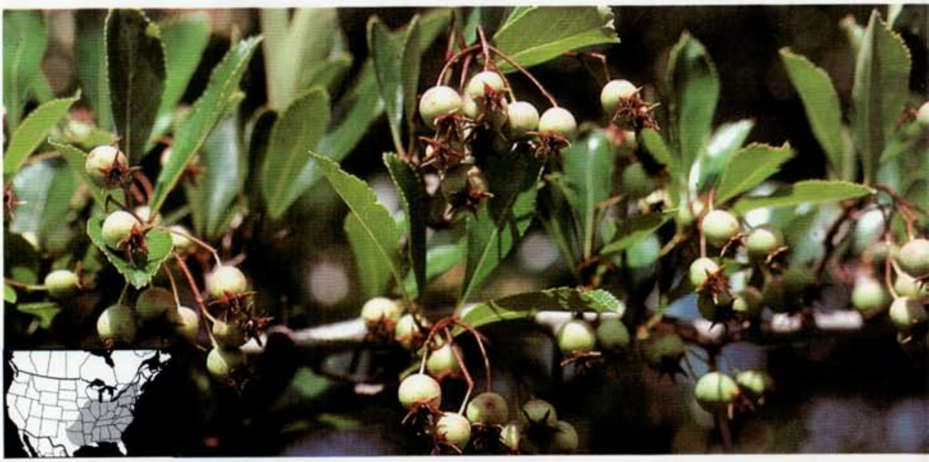

Cockspur hawthorn: fruiting branch (Elias).

Common persimmon: fruiting and flowering branches (Speas).

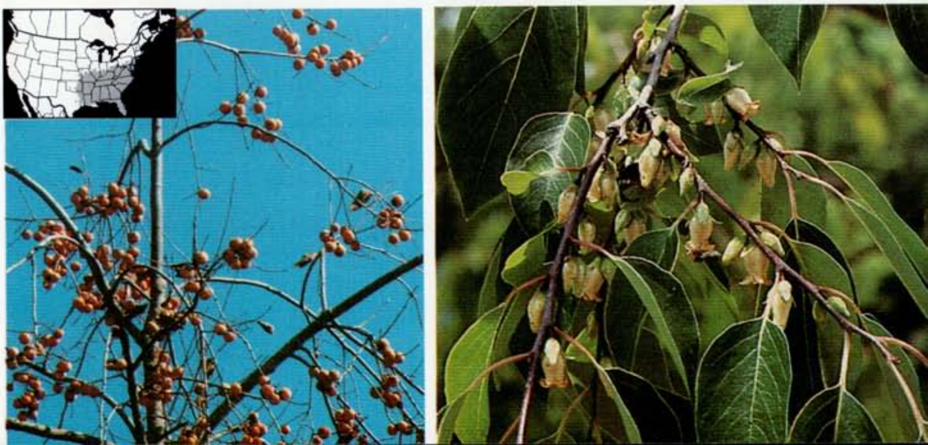

American mountain ash *Sorbus americana*
Also edible winter

Habitat: rocky slopes, seepage areas along ridges, at higher elevations. **Identification:** shrub or small tree to 9 m (30 ft) with open, rounded crown and thin, light gray smooth bark. ***Leaves alternate,*** deciduous, ***featherlike (pinnate) compound, 12–25 cm (5–10 in) long, with 9–17 leaflets, each leaflet*** 3–8 cm (1.2–3.2 in) long, lance-shaped to uniformly wide, ***finely toothed along margin,*** dull green. ***Flowers in dense, many-flowered, flat-topped clusters*** to 15 cm (6 in) across, each flower tiny, with 5 white petals. ***Fruits in large, dense clusters, persisting into winter, each fruit*** 4–8 mm (0.2–0.3 in) in diameter, ***bright orange-red, shiny,*** slightly fleshy, containing 1 or 2 seeds. **Harvest and preparation:** Fruits become palatable in winter after 1 or more freezings and are easily stripped from twigs. Edible raw but better cooked. For sauce, stew like CRANBERRIES; bring to boil 2 lb fruits and 2 cups sugar in 1 cup water, or berries plus 1 cup maple sugar (SUGAR MAPLE) and 1 cup cider. Reduce heat and simmer until berries mash easily and a drop of the sauce gels on cold plate. Pour into mold or storage jars. For pie, to sauce (above) add rind of ½ lemon and ⅛ tsp cinnamon. Line pie plate with crust, put in sauce, cover with crisscross strips of crust, leaving spaces between. Bake at 230°C (450°F) 20 min. Try basic jelly recipe with vinegar replacing about ¼ of water for sauce to serve with game. Also use regular jelly recipe. Berries keep well but must be dry for prolonged storage. **Related edible species:** European mountain ash *(S. aucuparia)*, Greene mountain ash *(S. scopulina)*, Sitka mountain ash *(S. sitchensis)*. **Poisonous look-alikes:** none.

Pecan *Carya illinoensis*

Habitat: bottomlands, rich moist soils; also extensively cultivated. **Identification:** tall tree with open round crown and thick, irregularly furrowed, dark reddish-brown bark. ***Leaves*** alternate, deciduous, ***featherlike (pinnate) compound,*** 32–52 cm (13–21 in) long with ***9–17 leaflets; leaflets*** 8–20 cm (3.2–8 in) long, 2.5–7.5 cm (1–3 in) wide, ***lance-shaped,*** coarsely toothed along margins. Male flowers in slender hanging catkins; female flowers short, inconspicuous, at tips of branches; flowers occur in early spring as leaves unfold. ***Fruits*** in clusters of 3–6, each large, uniformly wide, 3.5–5 cm (1.4–2 in) long, ***husk splitting into 4 thin sections to expose nut.*** Nuts cylinder-shaped to broadest near base, pointed at tip, light brown to reddish-brown, shell thin, meat oily, sweet. **Harvest:** Gather nuts in autumn as leaves turn and begin to fall. **Preparation:** Excellent eaten directly from shell, or use in any recipes calling for nuts. For pecan pie filling, beat 3 eggs, add ½ cup sugar, 1 cup dark corn syrup or molasses, ¼ tsp salt, 1 tsp vanilla, ⅓ cup melted butter. Spread 1 cup pecans over bottom of 9-in pie shell. Pour above ingredients over pecans and bake 10 min in 230°C (450°F) oven, then reduce heat to 175°C (350°F) and bake 35 min longer. **Related edible species:** BIG SHELLBARK HICKORY, SHAGBARK HICKORY, and MOCKERNUT HICKORY (following). **Poisonous look-alikes:** none.

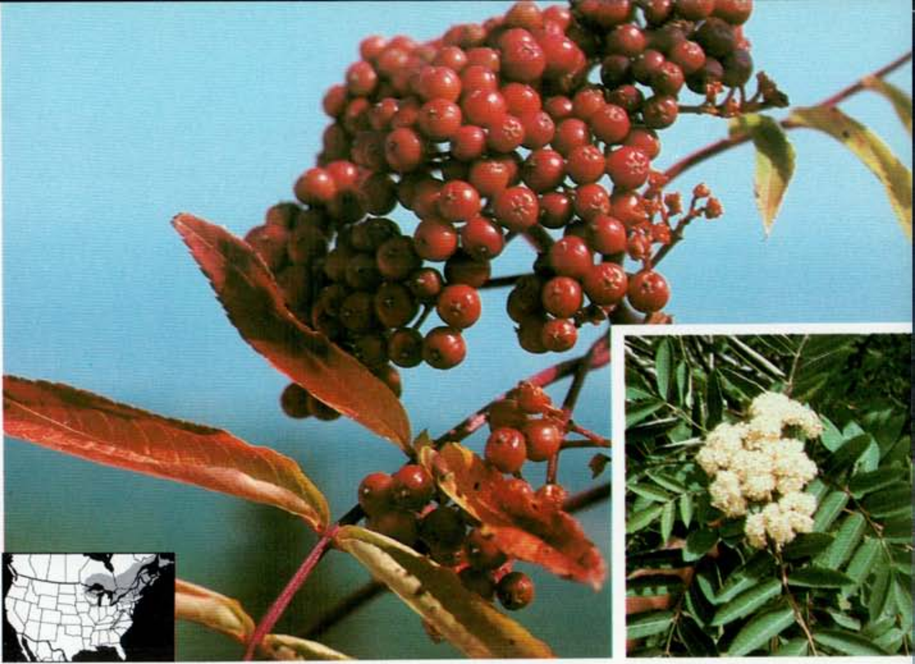

American mountain ash: fruiting and flowering branches (Elias).

Pecan: summer branches and nuts with splitting husks (Elias/Dykeman).

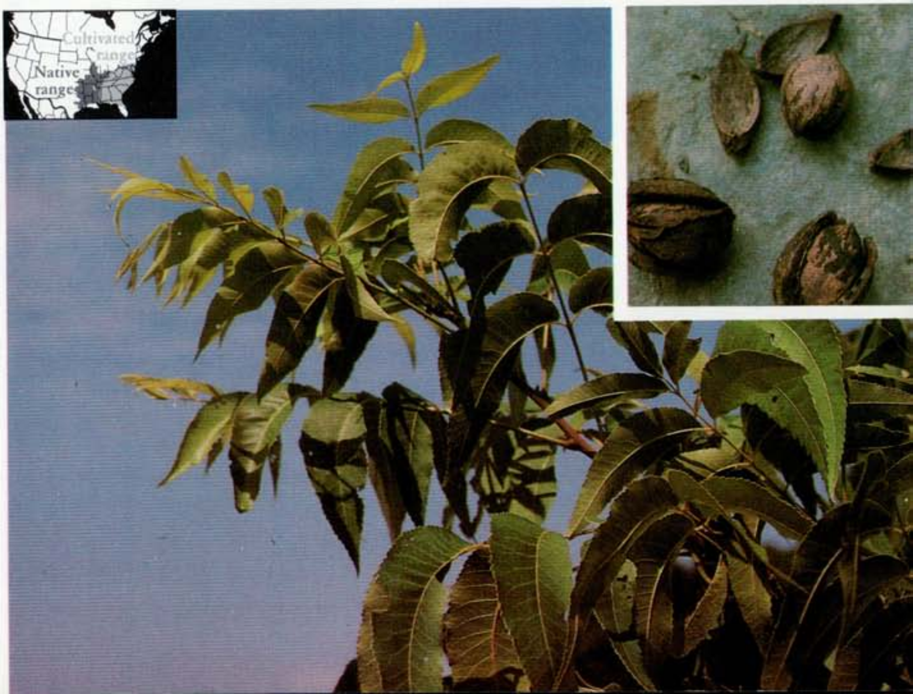
Cultivated range
Native range

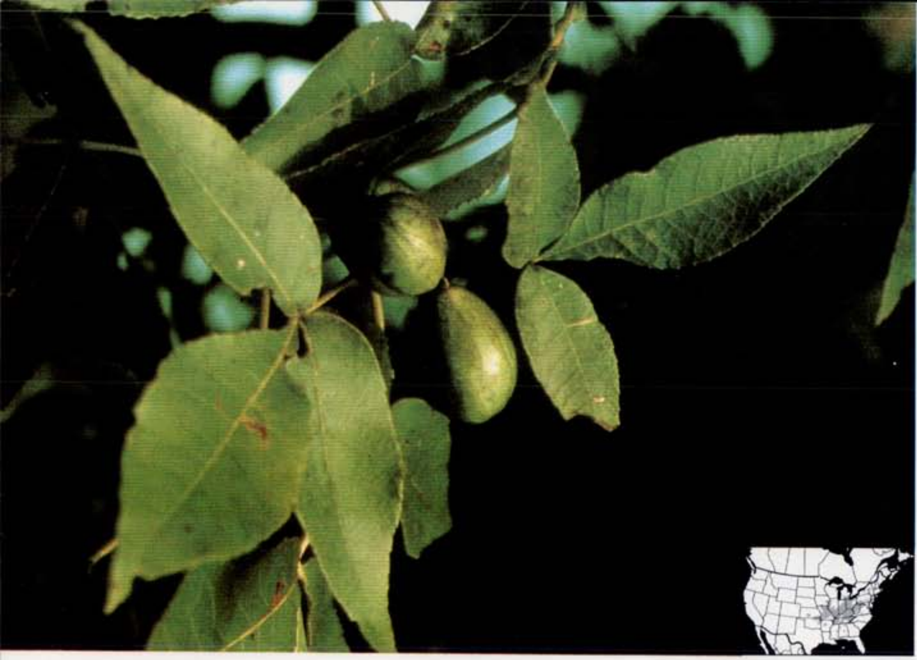

Big shellbark hickory: immature fruits (Clawson).

Shagbark hickory: flowering and fruiting branch (Elias). Bark (Dykeman).

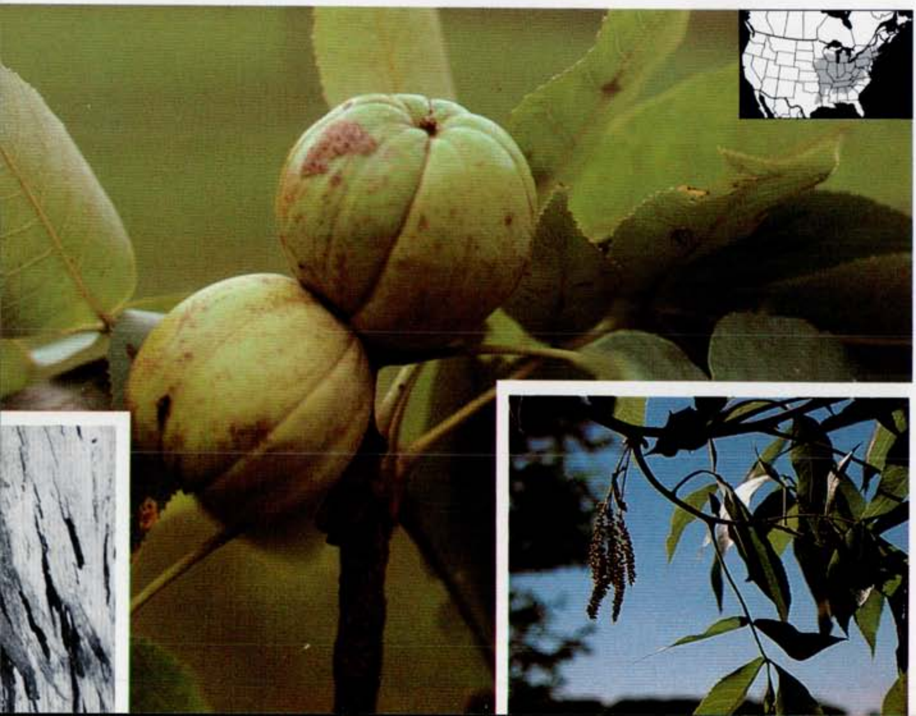

Big shellbark hickory *Carya laciniosa*

Habitat: deep rich soils of floodplains and bottomlands, especially of Ohio and Mississippi River valleys. **Identification:** large tree with rounded crown and ***light gray, somewhat shaggy bark. Leaves*** alternate, deciduous, featherlike (pinnate) compound, 25–60 cm (10–24 in) long, ***with 5–9 leaflets, usually*** *7;* leaflets 6–25 cm (2.4–10 in) long, 3.7–12.5 cm (1.5–5 in) wide, lance-shaped to broadest near base, sharp-pointed at tip, with numerous tiny teeth along margin. Male flowers in slender hanging catkins; female flowers short, inconspicuous, at ends of branches; both produced in early spring as leaves unfold. Fruits solitary or in pairs, globe- to egg-shaped, 3.5–7 cm (1.4–2.8 in) long, thick husk splitting lengthwise into 4 parts to expose nut. ***Nut egg-shaped,*** usually ridged, ***shell thick, hard, nutmeat sweet.*** **Harvest:** Gather nuts in autumn as leaves turn and fall; PECAN (preceding) and SHAGBARK (next) nuts are superior because they have thinner shells. **Preparation:** Remove nutmeats from shells; see SHAGBARK HICKORY. Use in any nut recipes. For nut-corn pudding, combine thoroughly 1½ cups cooked corn, ½ cup shelled, dried hickory nuts, 2 tbsp nut butter or oil, 1 cup boiling water, 2 beaten eggs, 2 tbsp honey, 2 tbsp fine cornmeal, ¼ cup seeds, raisins, or edible blossoms. Pour into well-greased casserole. Sprinkle top with more hickory nutmeats. Bake in preheated 175°C (350°F) oven for 1 hr. Serve hot. See other hickory species for additional uses. Trees can be tapped in late winter for sap. Prepare syrup as with SUGAR MAPLE. **Related edible species:** PECAN, SHAGBARK HICKORY, MOCKERNUT HICKORY. **Poisonous look-alikes:** none.

Shagbark hickory *Carya ovata* (synonym: *C. carolinae-septentrionalis*)

Habitat: dry upland slopes, well-drained soils of lowlands and valleys. **Identification:** medium-size to tall tree with open rounded crown and ***light to dark gray, shaggy bark.*** Leaves alternate, deciduous, featherlike (pinnate) compound, 20–36 cm (8–14.2 in) long, ***usually with 5,*** sometimes 7, ***leaflets;*** leaflets broadest near middle or tip, 8–18 cm (3–7 in) long, 1.5–6 cm (0.6–2.4 in) wide, ***with fine, sharp-pointed teeth along margin.*** Male flowers in slender, hairy catkins 10–13 cm (4–5.1 in) long; female flowers short, inconspicuous, at tips of branches; flowers occur in spring as leaves unfold. Fruits produced singly or in 2s, egg- to globe-shaped, 3–5 cm (1.2–2 in) long, husk thick, splitting lengthwise into 4 equal parts to expose the egg-shaped, ***4-sided nuts.*** Nuts with thin shell and sweet nutmeat. **Harvest and preparation:** Gather nuts in early autumn as they fall with leaves. Crack nuts with hammer and remove meats with nutpick. For efficient extraction, crack nuts enough to identify and eliminate wormy ones. Then smash enough to loosen meats. Place pounded nutmeats in pot, cover with water, and bring to boil. Most of meats will rise and can be skimmed off if desired. Indians boiled mix slowly and skimmed oil from surface to use as butter; then meats were skimmed and mixed with potatoes or meal, or dried into cakes. Boiled mix separated from shells is nutritious soup Indians called *powcohicora*. For hickory pie, use PECAN pie recipe, but increase amount of nuts to 1½ cups. See other hickory recipes. **Related edible species:** PECAN, BIG SHELLBARK HICKORY (preceding), MOCKERNUT HICKORY (next). **Poisonous look-alikes:** none.

Mockernut hickory *Carya tomentosa* (synonym: *C. alba*)

Habitat: ridges, hills, slopes, river valleys. **Identification:** large tree with rounded crown and dark gray, shallowly furrowed bark. ***Leaves*** alternate, deciduous, featherlike (pinnate) compound, 14–32 cm (5.5–12.6 in) long, ***with 5 or 7 leaflets; leaflets 8–22 cm (3–9 in) long, 6–12.5 cm (2.4–5 in) wide, narrow, broadest above middle,*** pointed at tip, toothed along margin. Male flowers in slender, stalked, 3-clustered catkins; female flowers small, 2–5 at tips of branches. Fruits solitary or few-clustered, globe- to egg-shaped, 4–8 cm (1.6–3.2 in) long, almost as wide, husk splitting lengthwise to expose nutmeat. ***Nut nearly rounded, often 4-sided, shell thick, hard, nutmeat sweet.*** **Harvest:** Gather nuts in autumn as leaves turn; PECAN and SHAGBARK HICKORY (preceding) are superior because of thin shells. **Preparation:** Remove nutmeats from shell; see SHAGBARK HICKORY instructions. Use any nut recipe including those for other hickories. For salad dressing, combine ½ cup ground hickory nuts, 1¼ cups vegetable oil, 3 tbsp lemon juice or STAGHORN SUMAC extract, 2 tbsp vinegar, ¼ tsp pepper, 1 chopped clove garlic. Shake well and use on salads. Tap trees in late winter for sap and prepare syrup as with SUGAR MAPLE. **Related edible species:** PECAN, SHAGBARK HICKORY, BIG SHELLBARK HICKORY. **Poisonous look-alikes:** none.

Black walnut *Juglans nigra*

Habitat: bottomlands, floodplains, rich low mixed deciduous forest, especially in well-drained soils. **Identification:** large tree with open rounded crown and thick, deeply furrowed, light to grayish-brown to almost black bark. ***Leaves*** alternate, deciduous, ***featherlike*** (pinnate) ***compound,*** 20–60 cm (8–24 in) long, ***with 15–23 leaflets;*** leaflets 4–9 cm (1.6–3.5 in) long, 2–3.5 cm (0.7–1.4 in) wide, broadest near base to broadly lance-shaped, with numerous fine, sharp-pointed teeth along margin. Male and female flowers separate, appearing in spring. ***Fruits large, rounded, 4–6 cm*** (1.6–2.4 in) ***in diameter, with thick husk enclosing globe-shaped nut,*** nut 3–4 cm (1.2–1.6 in) in diameter, ***shell irregularly and deeply furrowed,*** thick, nutmeat 4-lobed at base, sweet. **Harvest:** Gather nuts in autumn as they fall from trees. Remove thick brown-staining husks immediately by partially crushing, then rubbing with gloved hands or scrubbing with stiff brush to remove all remaining husk material. (Indians first suspended nuts in mesh bag in running water for several days.) Allow hulled nuts to dry thoroughly before using. **Preparation:** Crack shell with hammer, rock, or vise, applying blow or pressure at right angles to seam. Use nutpick. Indians crushed shells well to obtain oil and meats. See SHAGBARK HICKORY. After boiling crushed nuts for 30 min or more, allow shells and nutmeats to settle; then skim off oil. Boil again and use sieve to skim nutmeats as they rise above shells. Remove coarse shell pieces, dry and grind nutmeats. Use oil like that of hickories. Use nutmeats in baking or grind with honey or herbs to make nut butter. Use sap under SUGAR MAPLE. For other uses, see BUTTERNUT, hickory species, and book's Introduction. **Related edible species:** BUTTERNUT, CALIFORNIA WALNUT (next), Texas black walnut *(J. microcarpa)*, Arizona walnut *(J. major)*, Hinds walnut *(J. hindsii)*. **Poisonous look-alikes:** none.

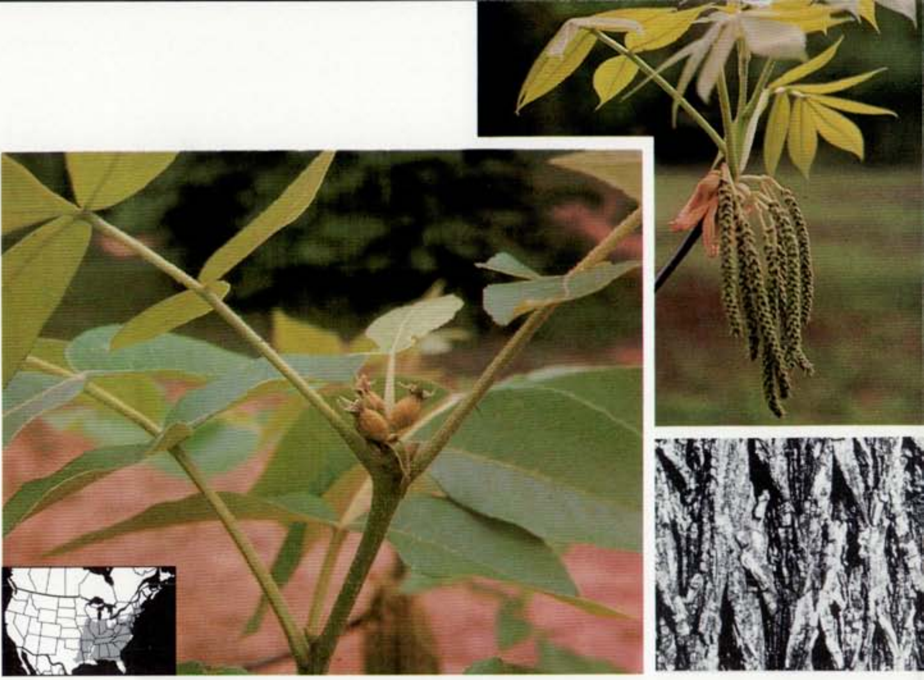

Mockernut hickory: immature fruits, male flowers, bark (Eye).

Black walnut: nearly mature fruits (Elias).

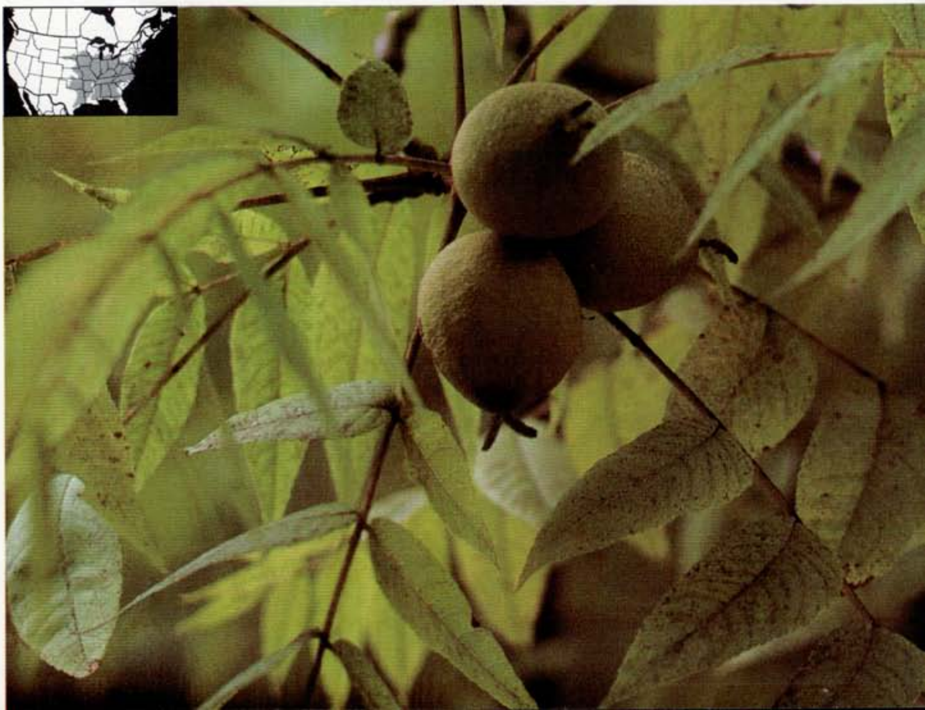

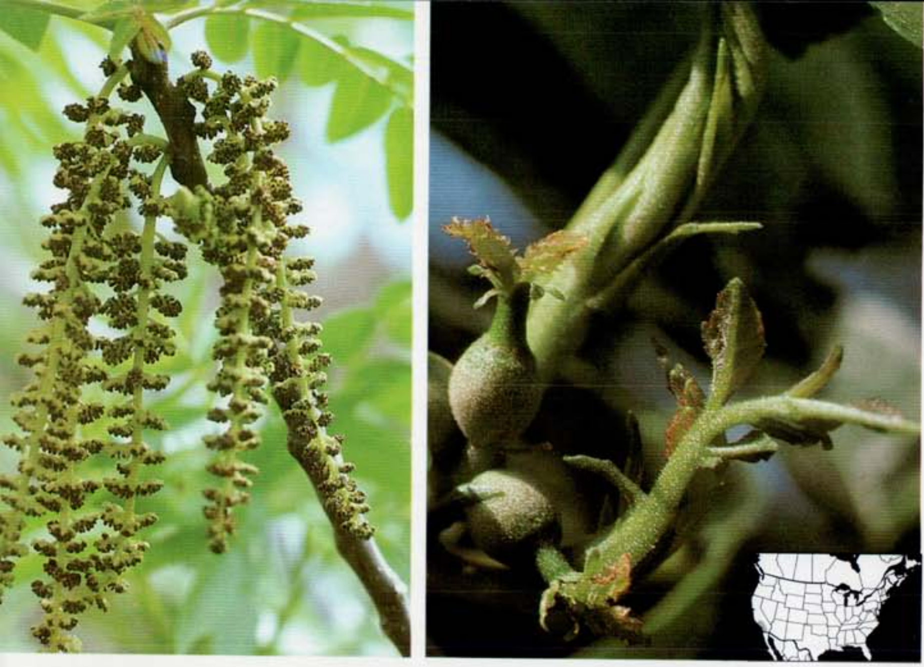

California walnut: male flowers and immature fruits (Shevock).

Butternut: nearly mature fruits (Elias).

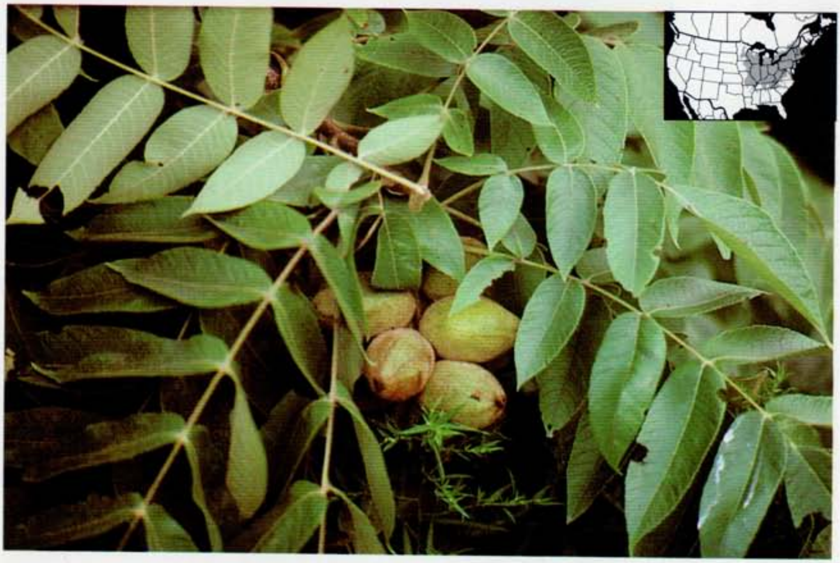

California walnut *Juglans californica*

Habitat: stream and river banks, bottomlands, often in gravelly soils at lower elevation, usually below 1,000 m (3,200 ft). **Identification:** small to medium-size tree with open rounded crown and thick deeply furrowed dark brown to almost black bark. ***Leaves*** alternate, deciduous, ***featherlike (pinnately) compound, 15–25 cm*** (6–10 in) ***long, normally with 9–15 leaflets;*** leaflets 2.5–7.5 cm (1–3 in) long, 1–2 cm (0.4–0.8 in) broad, lance-shaped to uniformly wide, with fine sharp-pointed teeth along margin. Male flowers in slender, hanging catkins; female flowers small, 1–4 at tips of new branches, both produced in early spring as leaves unfold. ***Fruits small, globe-shaped, 1–2 cm (0.4–0.8 in) in diameter, with thin husk surrounding nut; nuts rounded, with few shallow grooves,*** shell thin, nutmeat large, sweet. **Harvest:** Gather ripe nuts in autumn as they fall from tree. Remove brown-staining hulls immediately as described for BLACK WALNUT (preceding). Air dry before using. **Preparation:** Crack shells of dried nuts and remove meats with nutpick; easier to extract than some other walnuts. For Indian method, see BLACK WALNUT instructions. For uses, see other walnuts, hickories, and the book's introductory section. **Related edible species:** BLACK WALNUT. **Poisonous look-alikes:** none.

Butternut *(Also edible summer)* *Juglans cinerea*

Habitat: bottomlands, floodplains, mixed deciduous forests. **Identification:** medium-size tree with open, broadly rounded crown and thick, deeply furrowed, light gray to light brown bark. ***Leaves*** alternate, deciduous, ***featherlike*** (pinnate) ***compound,*** 30–60 cm (12–24 in) long, ***with 11–17 leaflets;*** leaflets 5–11 cm (2–4.3 in) long, 2–6 cm (0.8–2.4 in) wide, lance-shaped to widest above middle, finely toothed along margin. Male and female flowers separate, both appearing in early spring. ***Fruits large, egg-shaped, 5–8 cm*** (2–3.2 in) ***in diameter,*** often 4-sided, husk not splitting open at maturity. Nut egg-shaped, 3–6 cm (1.2–2.4 in) in diameter, 4-ribbed, shell deeply furrowed, thick, nutmeat sweet. **Harvest:** Pick partly grown nuts in summer for pickling when needle will penetrate husk. Gather ripe nuts in autumn as they fall from trees. Remove brown-staining hulls immediately as described for BLACK WALNUTS. Dry thoroughly before using. **Preparation:** Crack like BLACK WALNUTS or hold nut in vertical position and hammer on end until well cracked. Remove nutmeats with nutpick or use Indian method described under BLACK WALNUT to prepare oil and nut butter. For pickling, scald immature nut and remove all outer fuzz. Place nuts in kettle; cover with water and boil until water discolors. Change water and continue boiling. Repeat water changing and boiling until water remains clear. Pack nuts in canning jars with 1 dill flower, 3 walnut leaflets, 1 heaping tsp pickling spices, 1 tsp salt, ¼ tsp alum. Fill jar with boiling cider vinegar and seal. Age at least 1 month. For other uses see BLACK WALNUT, hickories, and book's introductory section. **Related edible species:** BLACK WALNUT, CALIFORNIA WALNUT (preceding), Arizona walnut *(J. major)*, Hinds walnut *(J. hindsii)*, and Texas black walnut *(J. microcarpa)*. **Poisonous look-alikes:** none.

Winter

Winter is hardly the season for wild plant foraging—especially in colder climates. Aboveground, edibles are basically limited to berries that persist into winter, twig bark for tea, and the inner bark of some trees. The most nutritious plant parts are underground; these are the starch-rich roots and tubers of perennials and biennials. But, in winter, roots and tubers are difficult to locate and identify, even if there's no snow cover. Besides, digging them in frozen ground would cost you more calories than the plants would return.

What about survival foods? You can survive many days without food, provided you stay warm and reasonably dry, and drink sufficient water to prevent dehydration. If you decide to stay put until help arrives, choose or fashion a shelter you can heat with a reflected campfire. In selecting a site, consider access to firewood, water, and an open area for distress signalling. Expend as little energy as possible. If you expect a long wait, consider passive means of obtaining meat: fishing with set lines or setting snares on small-game trails. If you have a firearm, you could

watch game trails from your shelter or attempt to call predators to you by imitating the squeal of an injured rabbit.

In winter, Indians sometimes raided the nut caches of squirrels and arrowhead tubers cached by muskrats. But such raids are strenuous. Besides, getting into a muskrat's house can be a wet, cold ordeal.

Some northern Indians made the inner bark of trees their major winter food staple. Since bark stripping kills trees, today it should be done only for survival.

You can also partake of wild edibles in winter by manipulating the seasons. That is, you can feast on wild preserves you harvested earlier. Or, if you are highly dedicated, you can in autumn dig up the whole roots of dandelion, chicory, dock, and pokeweed. Then plant the roots in a deep, soil-filled container until after a hard freeze. Bring the container into a dark cellar. Watered and kept at 13°C (55°F), the roots will produce several crops of blanched, tender, and mild leaves for salads or cooking.

The plant species shown in upcoming pages provide berries you can eat raw, berries and twigs for beverage flavoring, inner bark for gruel and flour, and sap for pure water or boiling down to syrup and sugar. These plants are shown as they look in winter, though most of them are also shown elsewhere in this book. Many other winter-edible plants are difficult to locate and identify in winter and so are not described in this winter section. In some regions, edible plants normally harvested in autumn may also be gathered in early winter.

(Dykeman)

Staghorn sumac *Rhus typhina*

Also edible summer, autumn

Habitat: common; woodland margins, fencerows, roadsides, and along streams. **Identification:** shrub or small tree to 10 m (33 ft) with thick, ***densely hairy branches and twigs with large orange to olive pith.*** Leaf scars C-shaped, nearly surrounding ***bluntly conical, hairy buds.*** No end bud. ***Fruits persist into winter in dense, cone-shaped clusters 10–20 cm (4–8 in) long; each fruit berrylike, rounded, 3–5 mm (0.1–0.2 in) in diameter, densely covered with dark red hairs.*** See also p. 186. **Harvest:** Fruits ripen in late summer to autumn and have strong lemony taste. Dense hairs on fruits prevent rain from penetrating cluster and washing out all flavor, so fruits are edible although less potent in winter. **Preparation:** For beverage, steep berries, with all twigs removed, in near boiling water for at least 15 min (at least 1 cup fruit to 1 qt water). Strain through cloth, dilute and sweeten to taste. Drink hot or cold but in moderation; some people show allergic reactions such as hives and cramps. To make jelly prepare extract as for beverage. Bring to boil 4 cups extract with 1 pkg powdered pectin. Add 5 cups sugar, boil hard 1 min, remove from heat, skim, pour into sterilized jars and seal with paraffin. **Related edible species:** SMOOTH SUMAC and SQUAWBUSH, and lemonade berry *(R. integrifolia)*. **Poisonous look-alikes:** Poison sumac *(R. vernix)* has white fruits and grows in swampy areas.

Highbush cranberry *Viburnum trilobum*

Also edible autumn

Habitat: along streams, wet thickets, moist woodlands. **Identification:** ***shrub,*** upright, to 5 m (16.5 ft) high; branches upright to spreading. Twigs light reddish-brown turning ash-gray with age. Buds opposite; no end bud. ***Fruits berrylike, in clusters of 10 or more, rounded to egg-shaped, 8–14 mm (0.3–0.5 in) long, bright red, translucent*** when ripe, juicy, sour, containing a single flattened stone. See also HIGHBUSH CRANBERRY on page 224. **Harvest:** Fruits ripen in late summer but persist on shrub and become less sour in winter. Make sure your shrub is highbush cranberry and not its bitter look-alike, the introduced guelder rose, *V. opulus.* Tasting will tell you. Best to pick in fall and winter. **Preparation:** much the same as CRANBERRY *(Vaccinium macrocarpon)*, except large seeds must be removed. While cooking add lemon or orange peel shavings to eliminate bad odor of berries. For sauce cover several quarts of berries with water in a kettle. Add outer peels of 2 lemons or oranges. Bring to boil and simmer 5 min. Mash thoroughly and simmer another minute or two. Force juice and pulp through coarse strainer to remove seeds. Add 1 pkg pectin to 4 cups strained fruit, bring to boil, and add 5 cups sugar. Boil again for 1 min and seal in sterile jars or refrigerate. Before first frost, pectin may not be needed. For juice and jelly, cook as for sauce above but strain through jelly bag. For juice, dilute and sweeten to taste. For jelly, proceed with 4 cups juice as for sauce. **Related edible species:** SQUASHBERRY (next), HOBBLEBUSH, NANNYBERRY, and the cranberry tree *(V. pauciflorum)*, of Pacific Northwest. **Poisonous look-alikes:** none, but the guelder rose *(V. opulus)* has bitter fruits.

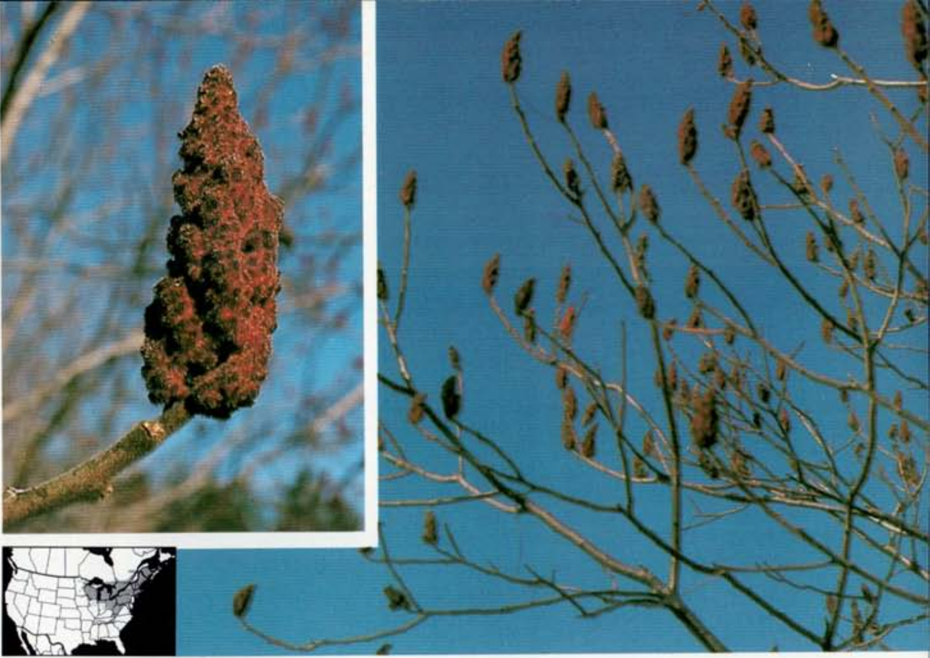

Staghorn sumac: winter fruit cluster and fruiting branches (Dykeman).

Highbush cranberry: early autumn branches (Elias). In winter berries look like this and are sweeter, though leaves will have fallen.

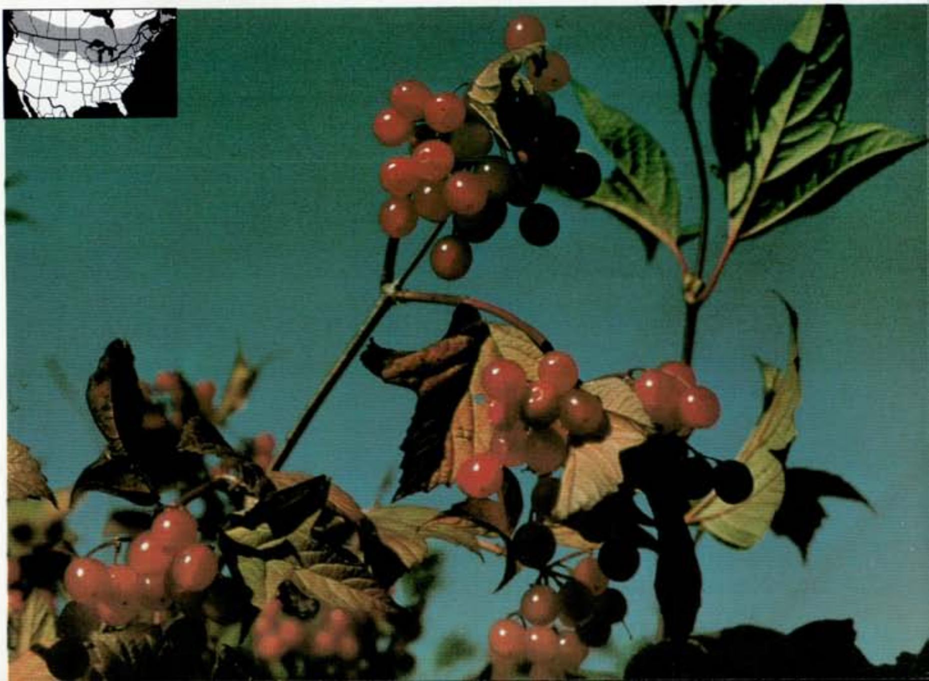

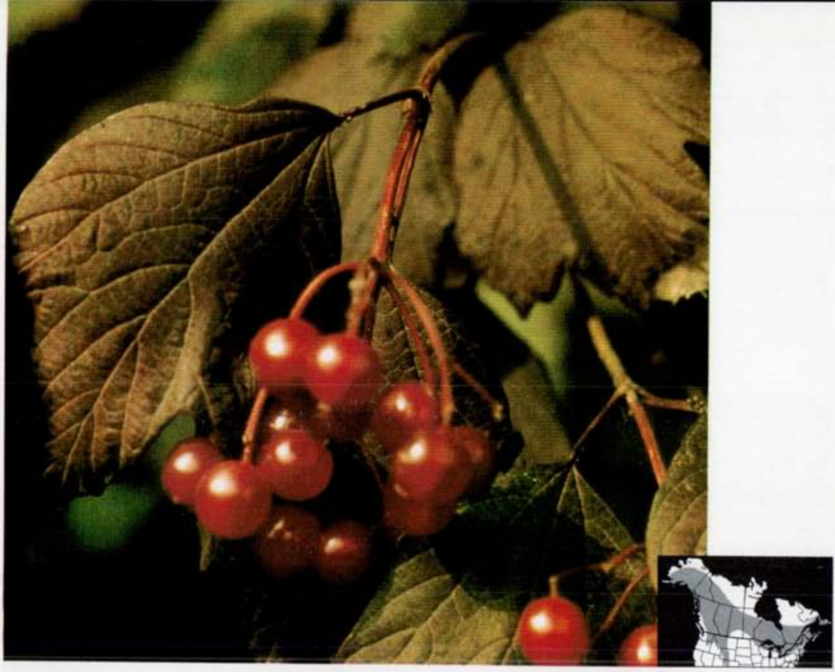

Squashberry: early autumn branch (Clawson). In winter, berries look like this and are sweeter, though leaves will have fallen.

Sweet birch: winter treetop, trunk bark, and immature male catkins (Dykeman).

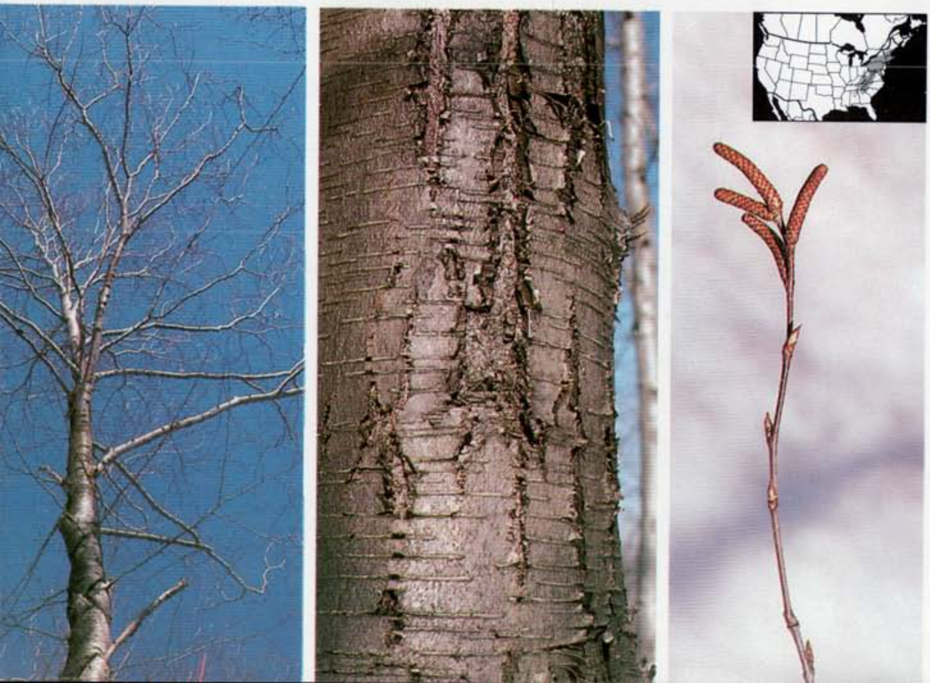

Squashberry *Viburnum edule*

Also edible summer, autumn

Habitat: along streams, wet thickets, margin of woodlands, moist low woodlands. **Identification:** ***shrub,*** often sprawling, to 2.5 m (8.1 ft) high; stems usually spreading, smooth. New branches reddish brown, older ones gray. ***Buds opposite, dark red, shiny, 2-scaled;*** both ***side and end buds present. Fruits in small clusters, berrylike, rounded,*** 6–10 mm (0.2–0.4 in) in diameter, ***red, containing a single flattened seed.*** See also SQUASHBERRY on page 226. **Harvest:** Fruits are similar to HIGHBUSH CRANBERRY but smaller and less sour. They ripen in August but persist through winter so are listed here. **Preparation:** See HIGHBUSH CRANBERRY (preceding), except juice does not require dilution for use as beverage. Makes excellent jelly. **Related edible species:** HIGHBUSH CRANBERRY. **Poisonous look-alikes:** none; the introduced guelder rose *(V. opulus)* is bitter but not toxic.

Sweet birch, Black birch *Betula lenta*

Also edible spring, summer, autumn

Habitat: forest or open woods, especially moist, north facing, protected slopes; in deep, rich, well-drained soils. **Identification:** medium-size tree with rounded crown and smooth, dark red to almost black bark. ***Broken twigs have wintergreen fragrance.*** Buds alternate, both side and end buds present, 5–7 mm (0.2–0.3 in) long, light brown, broadest near base and tapering to a point. Fruits in erect, ***brown cones, 2.5–3 cm (1–1.2 in) long, containing many tiny, winged seeds.*** Fruits mature in late summer and early fall. Cones persist into winter. ***New male flower catkins 2–3 cm (0.8–1.2 in) long at tips of twigs in winter.*** See also p. 122. **Harvest:** twigs, red inner bark, and bark of larger roots year round but best in late winter and spring. Sap in early spring, 3–4 weeks later than SUGAR MAPLE. **Preparation:** Eat inner bark fresh as emergency food, boiled like noodles, or dried and ground into flour. Dry inner bark at room temperature; store in sealed jars for later use. For tea, steep (do not boil) twigs or fresh or dried inner bark in water or, preferably, birch sap. Boiling removes volatile wintergreen oil. Sweeten to taste. For birch beer pour solution of 4 gal birch sap and 1 gal honey (or 5 gal sap and 3 lb sugar), which has been boiled 10 min, over 4 qt fine twigs in a crock. Cool, strain to remove twigs, add 1 cake yeast. Cover, ferment about 1 week, until cloudiness starts to settle. Bottle and cap tightly. For birch syrup, see instructions under YELLOW BIRCH (next). **Related edible species:** YELLOW BIRCH. **Poisonous look-alikes:** none.

Yellow birch *Betula alleghaniensis*

Also edible spring, autumn (synonym: *B. lutea*)

Habitat: rich woodlands, lower slopes and occasionally cool marshlands, usually below 1,000 m (3,300 ft) elevation. **Identification:** medium-size tree with rounded crown and peeling, ragged-edged, reddish-brown bark turning ***grayish dull yellow or yellowish-brown.*** Buds alternate, end bud absent, side buds 5–7 mm (0.2–0.3 in) long, brown, broadest near base and tapering to sharp point, somewhat sticky, with several scales. ***New male flower catkins 2–3 cm (0.8–1.2 in) long at tips of twigs.*** See also YELLOW BIRCH on page 122. **Harvest:** For syrup, tap as with SUGAR MAPLE beginning 3–4 weeks after maple sap flow begins. Drill 7/16-in hole (just over 11 mm) slightly upward about 7 cm (3 in) into a mature tree. Hammer in metal or wood spile only until firmly anchored; driving spile too far may split wood. Hang bucket and wait. Flow of sap copious but sugar content lower than from maple. For use of twigs and inner bark, see SWEET BIRCH (preceding); wintergreen flavor is milder in yellow birch. **Preparation:** Boil sap in open container, outdoors, adding more as volume decreases, until evaporation leaves a viscous, molasses flavored syrup with boiling temperature of about 104°C (220°F). Store in sterilized, completely filled, sealed jars. **Related edible species:** SWEET BIRCH. **Poisonous look-alikes:** none.

American mountain ash *Sorbus americana*

Also edible autumn

Habitat: rocky slopes, seepage areas along ridges at higher elevations. **Identification:** shrub or small tree to 9 m (30 ft) with open, rounded crown and thin, light-gray smooth bark. ***Leaf scars alternate, slightly raised on twig surface.*** Buds rather conical, ***end bud 6–12 mm (0.2–0.5 in) long, curved and gummy;*** side buds smaller, flattened and pressed against twig. ***Fruits in large, dense clusters, persisting into winter,*** each fruit 4–8 mm (0.2–0.3 in) in diameter, ***bright orange-red, shiny, slightly fleshy, containing 1 or 2 seeds.*** See also AMERICAN MOUNTAIN ASH on page 240. **Harvest and preparation:** Fruits become palatable in winter after one or more freezings and become easy to strip from twigs. Edible raw but better cooked. For sauce, stew like CRANBERRIES; bring to boil 4 cups berries and 2 cups sugar in 1 cup water (or berries plus 1 cup maple sugar and 1 cup cider). Reduce heat and simmer until berries mash easily and a drop of sauce gels on a cold plate. Pour into mold or storage jars. For pie, add to sauce (above) a rind of ½ lemon and ⅛ tsp cinnamon. Line pie plate with crust, put in sauce, cover with crisscross strips of crust leaving spaces between them. Bake at 230°C (450°F) for 20 min. For jelly, see the book's introductory section. For meat sauce substitute 1 cup vinegar for 1 cup water in jelly recipe. Berries keep well; but for prolonged storage, dry them to prevent mold. **Related edible species:** European mountain ash *(S. aucuparia)*, Greene mountain ash *(S. scopulina)*, and Sitka mountain ash *(S. sitchensis)*. **Poisonous look-alikes:** none.

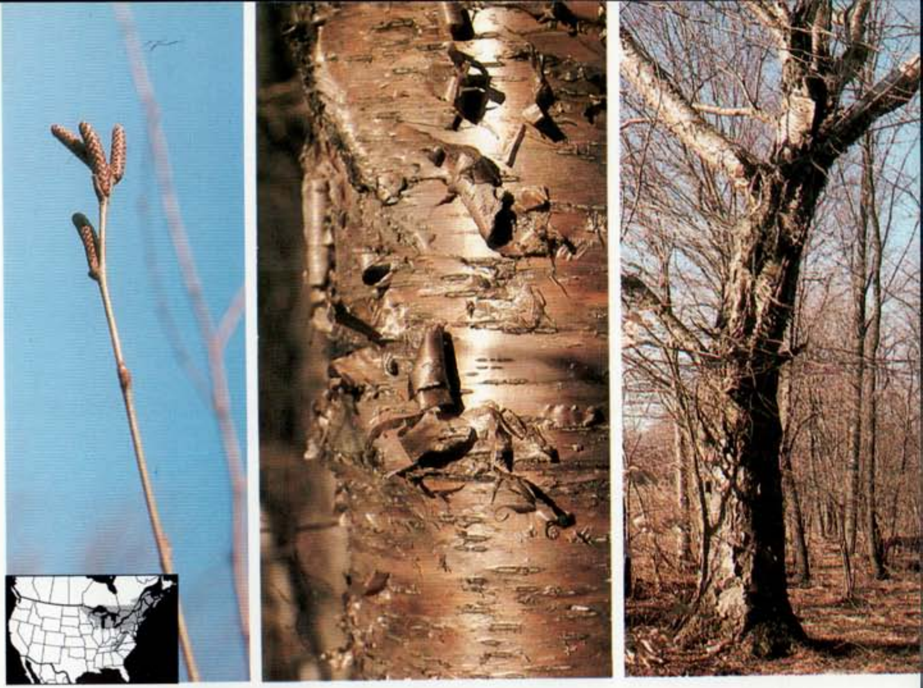

Yellow birch: winter immature male catkins, trunk bark, and tree (Dykeman).

American mountain ash: ripening autumn fruits (Elias). In winter, berries are brighter orange-red, and leaves will have fallen.

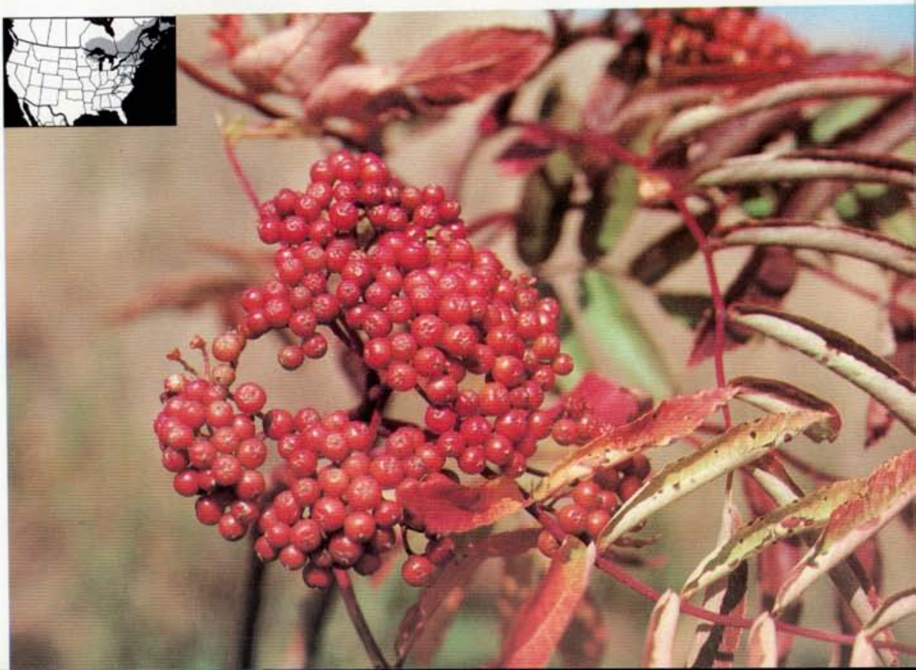

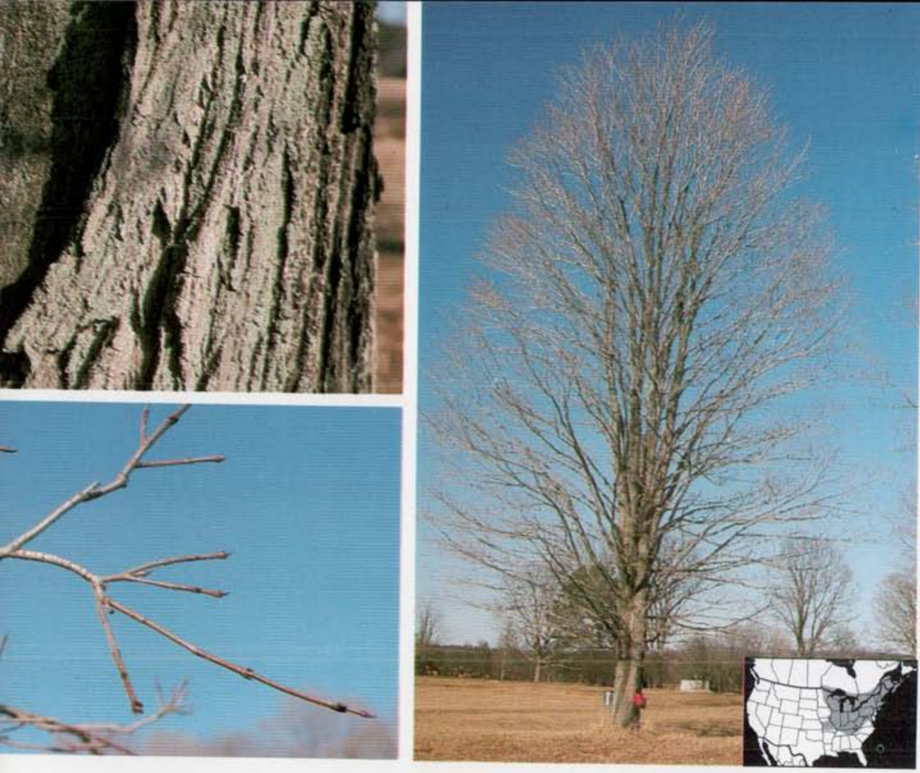

Sugar maple: trunk bark, winter twigs and tree. Metal spile with sap dripping into bucket. (Dykeman photos)

Sugar maple *Acer saccharum*

Also edible spring, summer

Habitat: eastern deciduous forest, common and widespread, especially at lower elevations. **Identification:** medium-size to tall ***trees*** to 30 m (98 ft) with a rounded crown, slightly ***fissured to furrowed dark-gray bark. Buds opposite,*** end and side buds present, ***light brown,*** scaly; ***end bud 6–12 mm (0.2–0.5 in) long,*** cone shaped; side buds smaller. See also SUGAR MAPLE on page 256. **Harvest:** Obtain sap from first spring flow until buds swell. Good sap flow requires freezing nights and warm days. Drill $^{7}/_{16}$-in hole (just over 11 mm) for commercial spile, slightly upward and 5–8 cm (2–3 in) into trunk. Drive spile just far enough to hold bucket, not so hard that you split wood. Collect sap and store cold until you have 25–30 gal. Collect winged seeds before fully ripe, usually from June to September. **Preparation:** Boil sap outdoors, indoors only if you like sticky walls. Use large surfaced, open pan, adding sap as water boils away. When 25 gal or so of sap has been added, continue to boil liquid until temperature is 4°C (7°F) above that of boiling water on that particular day. You may wish to boil for the last two degrees on a regulated stove. Filter syrup through milk filter into sterilized canning jars. Fill jars completely and seal. Yields about 3 qt. For maple sugar, boil 1 pt of syrup in saucepan with frequent stirring until temperature is at least 110°C (230°F). Remove from heat, cool, and stir until surface loses sheen. Pour quickly onto foil or waxed paper. When partially hardened, score into pieces with knife. Use sap fresh as a beverage, fermented to form a mild vinegar, or reduced by boiling to 25% its original volume as a base for teas. Soak seeds and remove wings. Boil until tender, drain, season, and roast for 10–15 min. **Related edible species:** Red maple *(A. rubrum)*; sap of all native maples may be used for syrup but sugar maple is best. **Poisonous look-alikes:** none; but avoid Norway maple *(A. platanoides)*, which has milky sap.

Poisonous plants

Plants that are inedible and even poisonous to varying degrees grow among the edible ones. So it's wise to learn to recognize poisonous plants, especially those that may resemble the edibles. In this section, you'll find full coverage of 20 plants that are particularly toxic, and you'll also find mention of many more related toxic species. To learn more about poisonous plants not fully covered here, consult a guidebook for your region.

Unfortunately, there are no general rules for recognition of poisonous plants. Even though birds or mammals eat a particular plant, that plant may still be poisonous to humans. For example, birds readily consume the fruits of poison ivy with no ill effects. But if you tried this, you'd be in serious trouble. Ironically, sometimes the most toxic plants are among the most beautiful. Some toxic beauties include blue flag, yellow flag, star of Bethlehem, and butterfly weed.

The more field experience you gain identifying plants, the easier you'll sort the edibles from the inedibles. Try to identify at least a few new poisonous plants each season.

Star of Bethlehem *Ornithogalum umbellatum*

Habitat: fields, grassy areas, roadsides, and open woodlands. **Identification:** ***small perennial herb*** to 35 cm (13.8 in) high, from an egg-shaped bulb 3–4 cm (1.2–1.6 in) long. ***Leaves mainly basal, long, narrow, 15–40 cm (5.9–15.7 in) long, 3–6 mm (0.1–0.2 in) wide, channeled, main vein almost white.*** Flowering stem erect, bearing several, short- to long-stalked flowers at the tip; ***each flower showy, bisexual, with 6 spreading petallike segments, white with a green stripe on lower side, each segment 1.5– 2.5 cm (0.6–1 in) long.*** Fruits are a 3-lobed, egg-shaped capsule 5–10 mm (0.2–0.4 in) long containing a few, nearly globe-shaped, black seeds. **Poisonous constituents:** Bulbs contain one or more alkaloids that have killed grazing animals. **Related poisonous species:** An introduced species, *O. nutans*, also has poisonous bulbs.

Death camass *Zigadenus venenosus*

Habitat: meadows, fields, scrubby woodlands along coast, in foothills, and coniferous forests. **Identification:** ***perennial herb from an egg-shaped bulb 1.2–2.5 cm (0.5–1 in) long, with dark outer scales;*** stems upright, 24–60 cm (9.4–24 in) high, slender, smooth. ***Leaves around base and on lower stem; basal leaves very narrow (linear), 15–30 cm (6–12 in) long, 5–10 mm (0.2–0.4 in) wide;*** stem leaves usually shorter and sheathing the stem. ***Stalked flowers numerous along an upright, unbranched, elongate cluster (lower-most flowers opening***

first), each flower with 6 petals or petallike structures, whitish, broadest near base, pointed at tip, bisexual. Fruit a small, egg-shaped capsule 1–1.5 cm (0.4–0.6 in) long, splitting open at maturity to release tiny, brown seeds. **Poisonous constituents:** All plant parts, especially the bulbs, contain alkaloids that can cause gastrointestinal distress, weakness, loss of motor function, and death. This plant and its relatives have killed many horses and cattle. **Related poisonous species:** All 12 native *Zigadenus* species should be considered *extremely* poisonous.

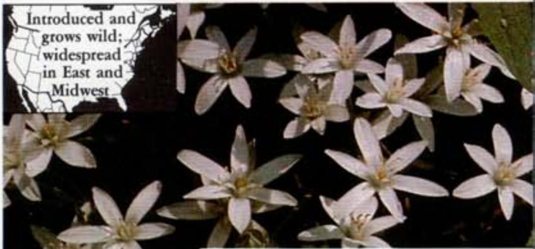

Star of Bethlehem: flowering plants (NYBG).

Death camass: flowering plant (NYBG).

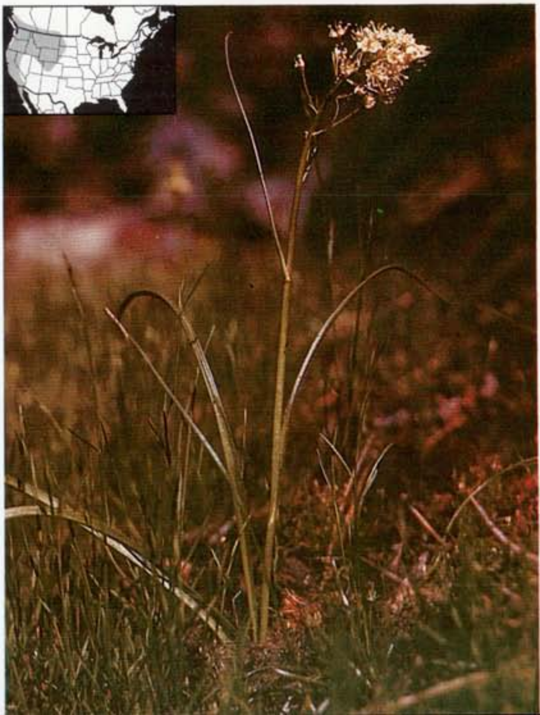

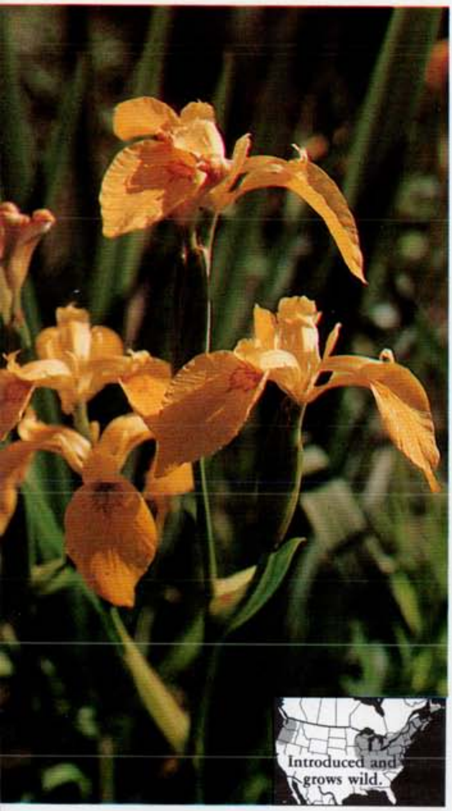
Introduced and grows wild.

Yellow flag: flowers (Elias).

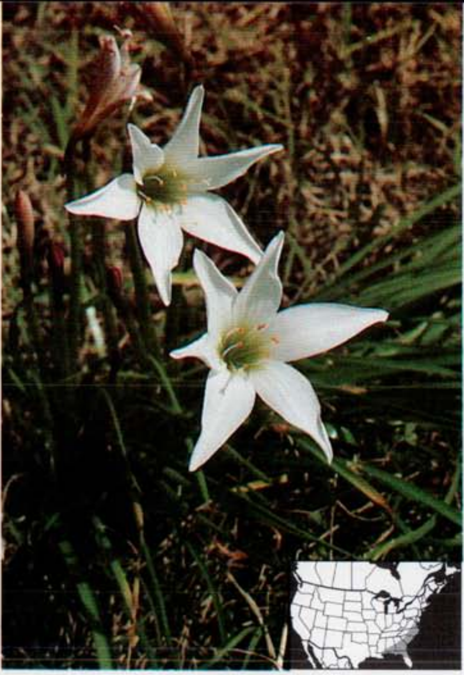

Atamasco lily: flowering plants (Speas).

Blue flag: flower (Elias).

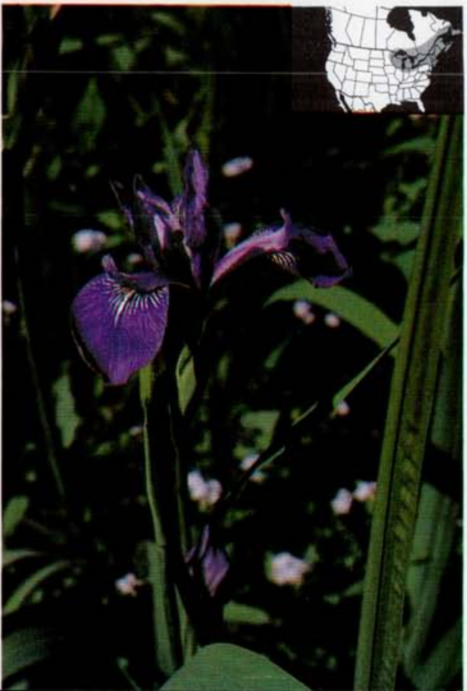

Atamasco lily *Zephyranthes atamasco*

Habitat: floodplain and low wet woodlands, wet fields and meadows, coastal plain and lower piedmont. **Identification:** perennial herb to 30 cm (12 in) high, from a small onionlike bulb 2–3 cm (0.8–1.2 in) in diameter. ***Leaves all basal, long, narrow (linear), 20–40 cm (8–16 in) long, 4–8 mm (0.15–0.3 in) wide,*** channeled, bright-green to bluish-green, shiny. ***Flowering stalk upright, to 30 cm (12 in) high, bearing a single flower at the tip; flowers white, rarely pink, showy, 6–10 cm (2.4–3.9 in) long, funnel-shaped with the 6 lobes spreading.*** Fruits are papery to leathery, nearly globe-shaped capsules, each 3-lobed and splitting open at maturity to release the shiny, black seeds. **Poisonous constituents:** Toxic compounds in the bulbs have produced staggering, collapse, and death in grazing animals. Thus, they should be regarded as *extremely* poisonous. **Related poisonous species:** All 12 native species of *Zephyranthes* should be considered extremely poisonous. ☠

Yellow flag *Iris pseudoacorus*

Habitat: swamps, ditches, borders of ponds, lakes, and streams; common, introduced from Europe. **Identification:** large perennial herb 0.6–1.2 m (2–3.9 ft) high, from spreading, underground stems (rhizomes). ***Leaves long,*** narrow, ***sword-shaped,*** 0.4–0.8 m (1.3–2.7 ft) long, 2–4 cm (0.8–1.6 in) wide, flattened, erect, pale green, sheathing the flowering stalks. Flowering stalks usually shorter than the leaves, bearing a ***large, showy flower;*** flowers usually ***bright yellow,*** with 3 spreading, petallike sepals each crested with brownish, irregular spots and 3 narrower, upright petals. Fruits are elongate, slightly angled capsules 5–10 cm (2–4 in) long, splitting open lengthwise to release the seeds. **Poisonous constituents:** leaves and roots; see BLUE FLAG (next). **Related poisonous species:** All native and introduced species of irises should be considered poisonous. ☠

Blue flag *Iris versicolor*

Habitat: wet meadows, ditches, swamps, marshes, along ponds and streams, native. **Identification:** ***large perennial herb*** to 1 m (3.3 ft) high, ***from thick, spreading, underground stems*** (rhizomes). ***Leaves long, narrow, sword-shaped,*** 0.4–1 m (1.4–3.3 ft) long, 1–3 cm (0.4–1.2 in) wide, flattened, erect, pale green to grayish-green, sheathing the stalks. Flowering stalks with 1 to 3 large, showy flowers; ***flowers bluish-purple to violet,*** with 3 broadly spreading, petallike sepals with conspicuous white veins and yellowish near base, petals 3, erect, narrow, ¼ to ½ as long as sepals. ***Fruits are elongate, 3-angled capsules 3.4–5.6 cm*** (1.4–2.2 in) ***long.*** **Poisonous constituents:** Leaves and roots are toxic and can cause an inflammation of the mucous membrane of the stomach and intestine. Most grazing animals avoid wild irises, but calves have been fatally poisoned by them. **Related poisonous species:** All native and introduced species of irises should be considered poisonous. ☠

Common buttercup *Ranunculus acris*

Habitat: yards, fields, clearings, and other disturbed sites. **Identification:** ***slender perennial herb*** to 1 m (3.3 ft) high, from a short, thick underground stem (rhizome); stem upright, branched, usually with spreading hairs. ***Basal leaves rounded in outline, 4–10 cm (1.6–3.9 in) across, deeply 5–7 lobed,*** each lobe coarsely toothed or with narrow lobes, softly hairy, with long leafstalks; stem leaves similar to other leaves except smaller and gradually becoming stalkless. Flowers several in a branched, open, spreading cluster; ***each flower with 5 bright glossy, yellow to whitish petals*** 7–14 mm (0.25–0.5 in) long, broadest near the tip. Fruits are clusters of flattened, broad, beaked, thin-walled nuts about 2–3 mm (about 0.1 in) long. **Poisonous constituents:** Leaves and stems are bitter tasting, possibly due to oils. These oils can cause abdominal pains from an irritated stomach and intestines. Fresh buttercups have caused diarrhea and blindness in grazing animals. **Related poisonous species:** All native species of buttercups contain toxic or otherwise irritating oils and should not be eaten.

Jimson weed *Datura stramonium*

Habitat: disturbed sites and waste grounds, old fields, roadsides, barnyards. **Identification:** ***coarse annual herb*** to 1.5 m (5 ft) high, from a short taproot; stems upright, branched, usually smooth. ***Leaves broadest near the base, 8–16 cm (3.2–6.4 in) long, 4–12 cm (1.6–4.7 in) wide, coarsely and irregularly large-toothed along the margin. Flowers large, 5–8 cm (2–3.1 in) long, showy, white to light purple, upright,*** produced singly on upper stem; petals fused and narrowly funnel-shaped, the lobes spreading, with a heavy fragrance. ***Fruits upright, egg-shaped capsules 3–5 cm (1.2–2 in) long, usually with short spines,*** splitting open along 4 valves to release dark pitted seeds. **Poisonous constituents:** All plant parts, including seeds, are poisonous. Alkaloids will cause impaired vision, thirst, delirium, unpredictable movements, possibly coma and death. **Related poisonous species:** All *Datura* species.

May apple *Podophyllum peltatum*

Habitat: rich and open woodlands, wet fields and meadows, pastures, occasionally roadsides. **Identification:** attractive perennial herb to 50 cm (20 in) tall, from a fleshy, spreading, underground stem (rhizome); stem single, short, producing single or paired leaves and single blossom between. ***Leaves large, nearly circular in outline, 15–40 cm (5.9–15.7 in) in diameter, deeply 5–9 parted, margins with a few coarse teeth or lobes; leafstalk attached to center of blade.*** Flower somewhat showy, cup-shaped, with 6–9 white petals, each 3–5 cm (1.2–2 in) wide, produced on a nodding stalk. ***Fruit an egg-shaped, fleshy, yellow (rarely red) berry 2.5–5 cm (1–2 in) long,*** containing several seeds. **Poisonous constituents:** Leaves and particularly the roots contain a resinous compound known as podophyllin that can cause violent cathartic reactions. Consumption of small quantities will produce severe gastric upset and vomiting. Death may occur from larger quantities. The fleshy pulp of the fruit is edible, although the seeds should not be eaten. See MAY APPLE on page 140. **Related poisonous species:** none in North America.

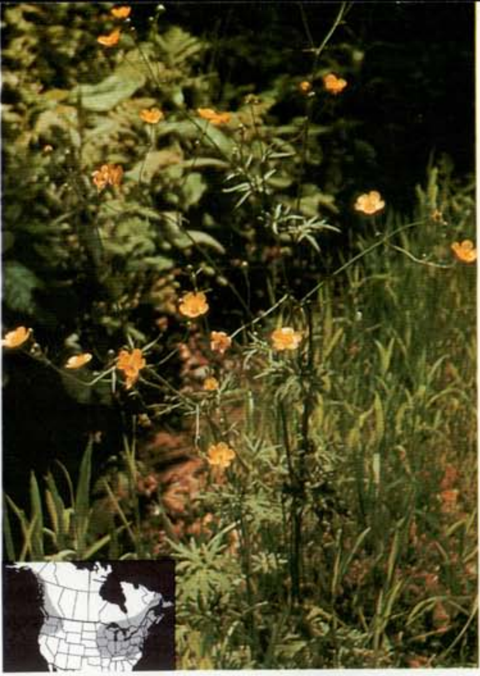

Common buttercup: flowering plants (NYBG).

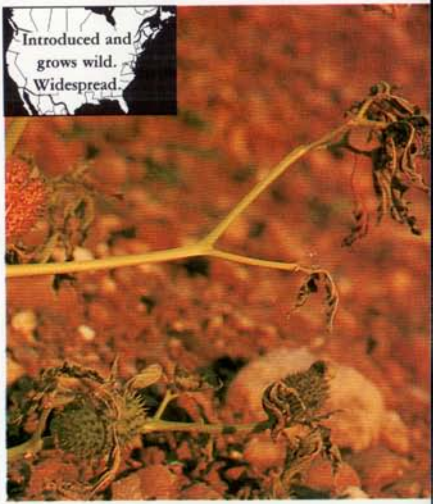
Introduced and
grows wild.
Widespread.

Jimson weed: fruiting branch after frost (Dykeman).

May apple: plants in early spring and flowering plant in May (NYBG).

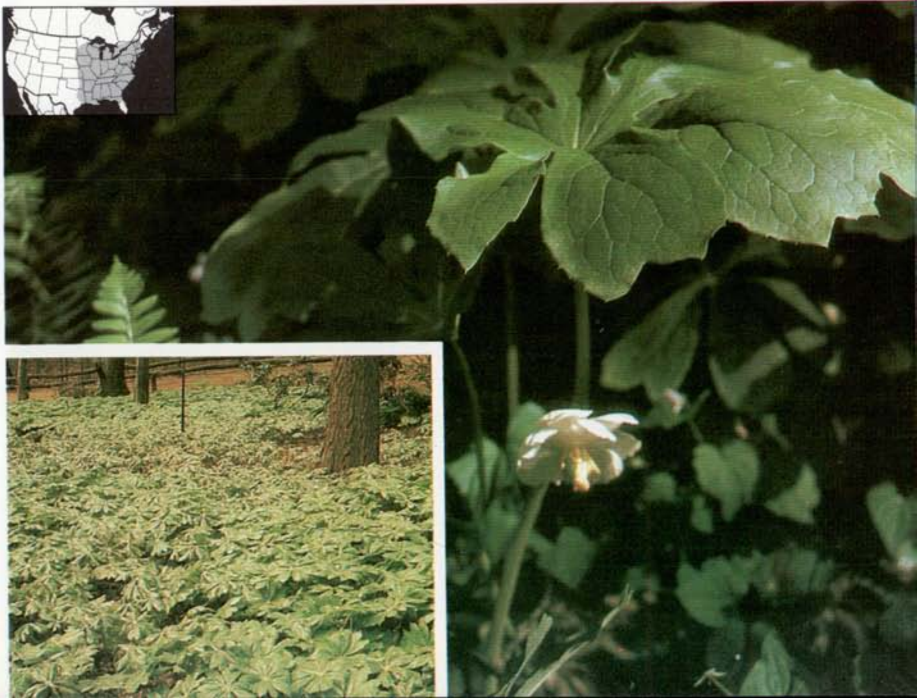

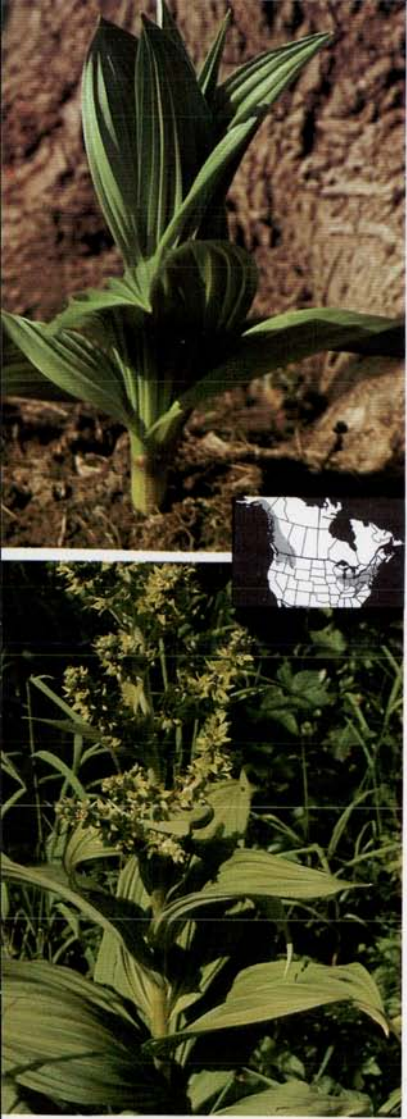

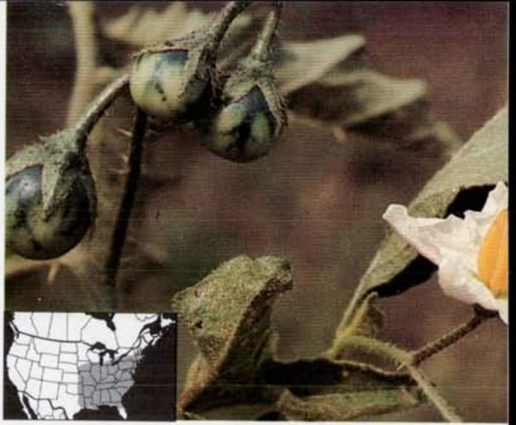

Horse nettle: flower and immature fruits, above (NYBG).

White hellebore: young plant, upper left (Elias). Top half of flowering plant, left (NYBG).

Yellow sweetclover: flower clusters, bottom left. Flower stalk, below. (NYBG photos)

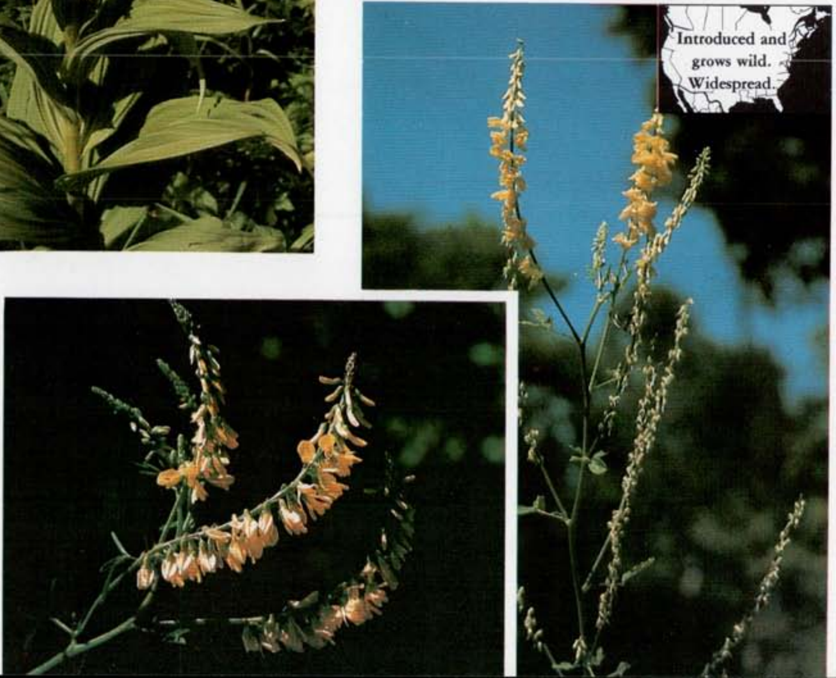
Introduced and grows wild. Widespread.

Horse nettle *Solanum carolinense*

Habitat: disturbed sites, fields, vacant lots, roadsides, and sandy waste areas. **Identification:** coarse perennial herb to 1 m (3.3 ft) high, from spreading underground stems (rhizomes); ***stem and leaves with branched hairs and slender spines. Leaves*** alternate, simple, broadest near base, 6–12 cm (2.4–4.7 in) long, 3–6 cm (1.2–2.4 in) wide, ***coarsely and irregularly large-toothed along margin, rough to touch.*** Flowers few on elongated terminal cluster; ***each flower white to light purple, 2–3 cm (0.8–1.2 in) across,*** five-parted petals spreading and even bending backward, broadest near base. ***Fruit a yellow fleshy berry*** 1–1.6 cm (0.4–0.6 in) in diameter, smooth. **Poisonous constituents:** Leaves and fruits contain the alkaloid solanine, which can cause vomiting, nausea, abdominal pains, and other gastrointestinal problems. **Related poisonous species:** Many species of the genus *Solanum* contain the solanine alkaloid in varying amounts. Among species to avoid are European nightshade *(S. dulcamara)*, black or deadly nightshade *(S. nigrum)*, and silverleaf nightshade *(S. elaeagnifolium)*. ☠

White hellebore, False hellebore *Veratrum viride*

Habitat: wetlands, swamps, floodplains, near streams and rivers, low wet sites in woodlands. **Identification:** ***large perennial herb*** to 2 m (6.6 ft) high, from thick, fibrous roots; stems upright, stout, unbranched. ***Leaves*** numerous along stem, ***broadest near base or middle, 15–32 cm (5.9–12.6 in) long, about half as wide,*** entire along margin and pointed at tip, nearly stalkless and almost clasping stem, with conspicuous, slightly sunken ribs. ***Flowers numerous in large, erect, branched clusters*** 20–50 cm (7.8–19.7 in) high, hairy; each flower yellowish-green, with 6 spreading, petallike segments. Fruits are egg-shaped capsules 1.8–2.6 cm (0.7–1 in) long, containing numerous seeds. **Poisonous constituents:** Several alkaloids from white and other hellebores have been used in medicines and insecticides. While all plant parts contain alkaloids, the roots concentrate them. Eating this plant can cause depressed heart activity, headaches, a burning sensation in the mouth and throat, and prostration. **Related poisonous species:** All 12 native *Veratrum* species should be considered poisonous. ☠

Yellow sweetclover *Melilotus officinalis*

Habitat: disturbed sites and waste areas, roadsides, fencerows, old fields; introduced weed from Europe and Asia. **Identification:** biennial herb to 2 m (6.6 ft) high, stems sparingly branched and weakly upright, usually smooth. ***Leaves compound, composed of three leaflets (trifoliolate);*** leaflets broadest near or above middle, 1.2–2.5 cm (0.5–1 in) long, 5–15 mm (0.2–0.6 in) wide, finely toothed along margin, greenish-yellow, usually smooth; leafstalks with conspicuous leaflike bract (stipule) at junction of stalk and stem. ***Flowers numerous on slender, elongated clusters 4–12 cm (1.6–4.7 in) long, each flower yellow, resembling tiny pea blossom.*** Fruits are egg-shaped pods 2–4 mm (0.1–0.2 in) long, smooth, stalked. **Poisonous constituents:** Improperly dried sweetclover will easily mold and in the process produce coumarin, an anticoagulant that can cause severe internal bleeding and death. Moldy sweetclover mixed in hay has killed many cattle. Thus, avoid this plant when gathering green leafy edibles. **Related** ☠

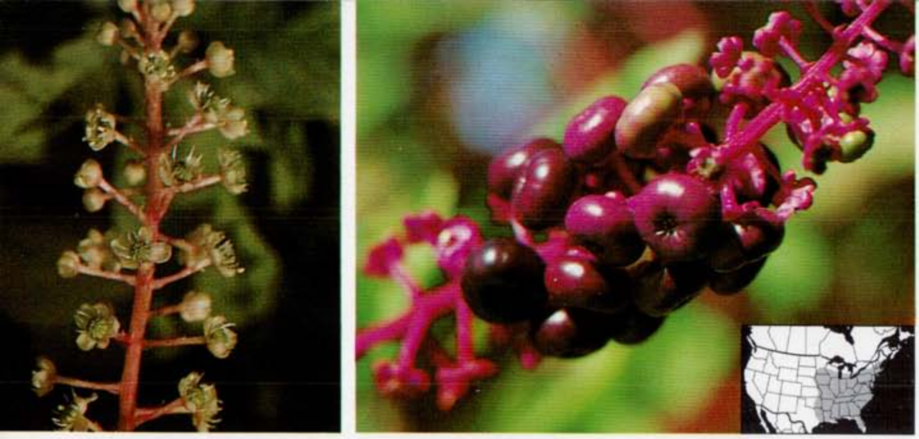

Pokeweed: flower cluster (NYBG). Fruit cluster with immature fruits at tip, bottom third; ripe fruits, middle; and stalks after fruits have fallen, upper third (Crow).

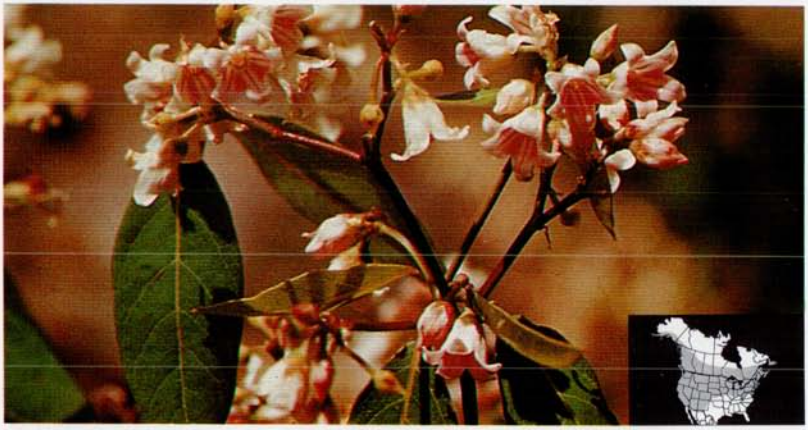

Dogbane: flowering branch (NYBG).

Butterfly weed: flowering plant top and fruit capsules (NYBG).

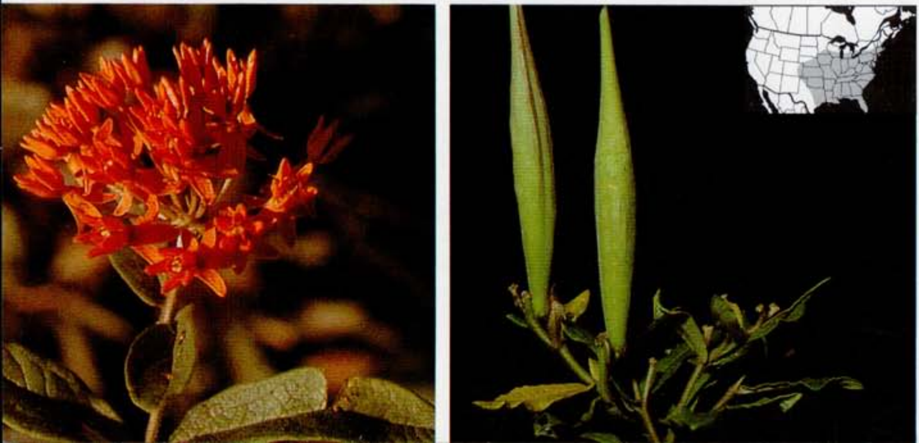

poisonous species: Also introduced from Europe, white sweetclover *(M. alba)* is equally dangerous.

Pokeweed *Phytolacca americana*

Habitat: woodland margins, disturbed sites, waste ground, roadsides, fencerows. **Identification:** ***large, perennial herb*** to 3 m (9.8 ft) tall with thick, fleshy roots; stems branched, upright, smooth, ***with disagreeable odor when broken.*** Leaves alternate, 9–30 cm (3.5–11.8 in) long, 3–12 cm (1.2–4.7 in) wide, broadest near base, pointed at tip, entire along margin; leafstalks to 5 cm (2 in) long. ***Flowers greenish to white, produced in narrow, elongate clusters*** (racemes); each flower 2–4 mm (0.1 in) long, petals absent. ***Fruits are 5–12-parted berries, wider than long, 6–10 mm (0.2–0.4 in) in diameter, green turning purplish-black at maturity.*** **Poisonous constituents:** Toxins are more prevalent in mature leaves, stems, and particularly in roots than in young growth and berries. Diarrhea, vomiting, cramps, sweating, reduced breathing capacity, or even death can occur from eating this plant. You can eat young cooked shoots if you discard the cooking water. See also POKEWEED on page 96. **Related poisonous species:** none. There is only one species of pokeweed in North America. ☠

Dogbane *Apocynum androsaemifolium*

Habitat: open woodlands, margins of woodlands, fields, roadsides, usually drier sites. **Identification:** ***perennial herb*** to 50 cm (19.7 in) high; stems upright, branching, ***with milky juice in all parts. Leaves opposite, broadest near base to almost uniformly wide, 4–8 cm (1.6–3.2 in) long, 2–4.5 cm (0.8–1.8 in) wide,*** entire along margin, pointed at tip, dark green above, hairy beneath, sometimes drooping, with slender leafstalk. Flowers several in branched, round-topped clusters at top or upper part of stems; each flower bell-shaped, 5–10 mm (0.2–0.4 in) long, white to pink, 5-parted, spreading to drooping. ***Fruits are narrow, cylindrical, capsulelike pods 6–18 cm (2.4–7 in) long, slender pointed at tip, containing numerous small seeds, each tipped with long silky hairs.*** **Poisonous constituents:** All plant parts contain resins and glycosides capable of affecting the cardiovascular system of mammals. **Related poisonous species:** All of the 6 or 7 native species of dogbane *(Apocynum)* contain milky sap and probably the glycosides. Thus, all should be considered poisonous. ☠

Butterfly weed *Asclepias tuberosa*

Habitat: fields, meadows, prairies, open woodlands, margins of woodlands. **Identification:** perennial herb, upright, 30–80 cm (11.8–31.5 in) tall; stems single or branched, stout, hairy. ***Sap not milky, leaves alternate on stems and opposite on branches,*** variable in shape but often very narrow and widest near base or tip, 5–10 cm (2–4 in) long, 0.4–2.5 cm (0.1–1 in) wide, pointed at tip, ***usually without leafstalk,*** hairy, especially beneath. ***Flowers produced in showy, branched, flat-topped clusters;*** each flower ***orange to red or yellow,*** 5-parted, petals (corolla) deeply lobed. Fruit a dry, narrowly egg-shaped capsule 8–14 cm (3.1–5.5 in) long, splitting open only along 1 side to release tiny, flattened seeds tipped with numerous, long, silky hairs. **Poisonous constituents:** Plant ☠

parts contain some glycosides, alkaloids, and resinoids that can cause weakness, staggering, and even seizures. The various species of *Asclepias* have killed sheep, cattle, horses, and goats. **Related poisonous species:** All 30 native species should be considered potentially poisonous.

Poison hemlock *Conium maculatum*

Habitat: disturbed sites such as fencerows, roadsides, stream and river banks, old fields and vacant lands; introduced from Europe. **Identification:** ***large, biennial herb*** to 3 m (9.9 ft) high, from large, white, carrotlike taproot; ***stems*** upright, branching, purple-spotted, hollow, ***smooth. Leaves basal,*** repeatedly branched, irregularly divided and highly dissected, ***resembling parsley leaves,*** broadly triangular in outline, smooth, dark green, ***with odor of parsnips when crushed.*** Tiny ***white flowers*** produced ***in*** compound, ***flat-topped clusters (umbels)*** 4–8 cm (1.6–3.1 in) wide; flower cluster stalks 4–10 cm (1.6–3.9 in) long. Fruits dry, egg-shaped, 3–4 mm (0.1 in) long, flattened on inner surface, conspicuously ribbed, grayish-green. **Poisonous constituents:** All plant parts, especially the stem, leaves, fruits, and root are extremely poisonous. The toxic alkaloids cause nervousness, trembling, reduced heartbeat, coma, and even respiratory failure. **Related poisonous species:** A related genus, *Cicuta,* contains several species, commonly called WATER HEMLOCK (next), which are also extremely poisonous. ☠

Water hemlock *Cicuta maculata*

Habitat: wet sites such as swamps, stream and river banks, ditches and bottomlands. **Identification:** ***large biennial or perennial herb*** to 2 m (6.6 ft) high, from one to a cluster of several, thick, fleshy, tuberlike roots; stems upright, branched, often purple spotted, smooth, exuding yellow liquid when cut. ***Leaves large,*** bi- or tripinnately compound on lower stem, upper stem leaves less dissected, leaf segments usually 4–8 cm (1.6–3.1 in) long, 0.5–3 cm (0.2–1.2 in) wide, toothed, smooth, ***with odor of raw parsnips when crushed. Small white flowers numerous in large, compound, flat-topped clusters (umbels).*** Fruits brown, dry, egg-shaped, 3–4 mm (0.15 in) long, with conspicuous ribs. **Poisonous constituents:** All parts of this plant are extremely poisonous. A resinlike cicutoxin causes severe abdominal pains, excessive salivation, vomiting within minutes of consumption, and can cause death. One of the most deadly poisonous North American plants! **Related poisonous species:** All 10 native species of *Cicuta* are extremely poisonous. Also see POISON HEMLOCK (preceding). ☠

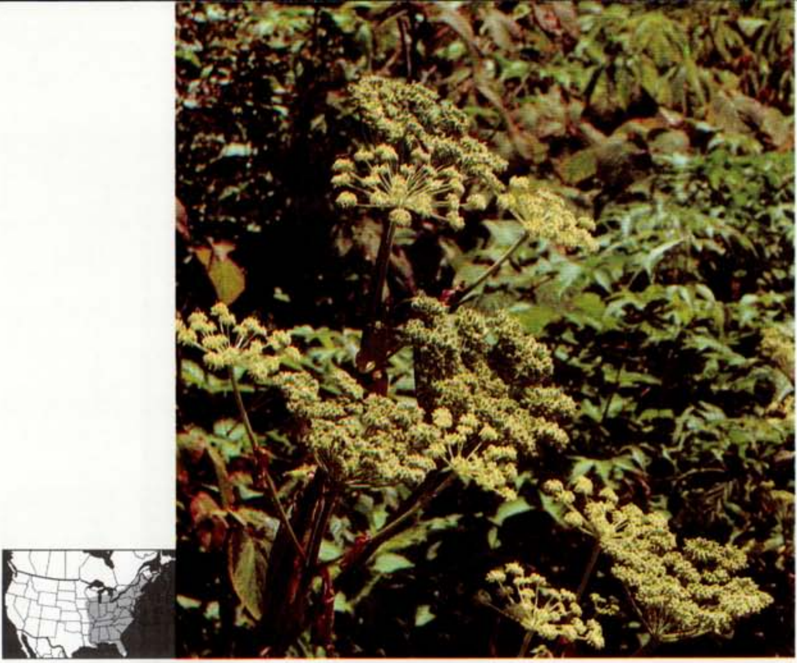

Poison hemlock: flowering stalks (Speas).

Water hemlock: flowering stalk (NYBG).

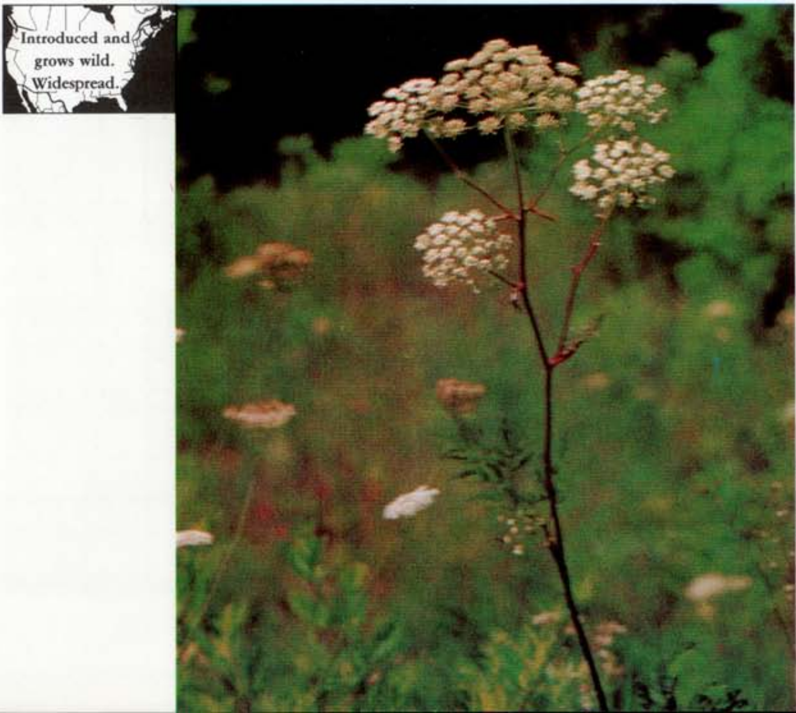
Introduced and grows wild. Widespread.

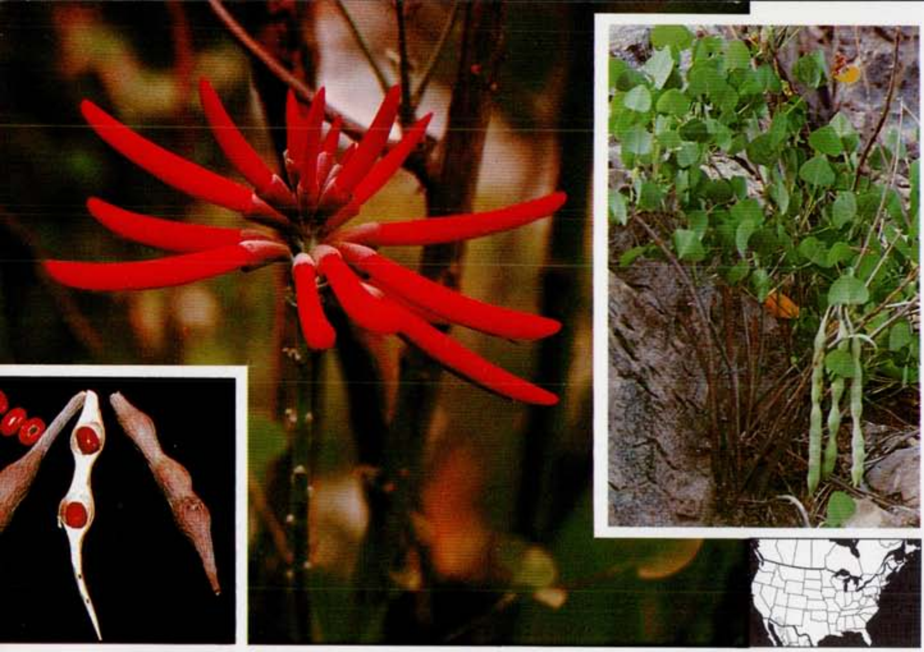

Southwestern coral bean: seeds and dried pods, flower cluster, and fruiting plant (Bernard).

American yew: mature fruits (Bogel). Fruiting branches (NYBG).

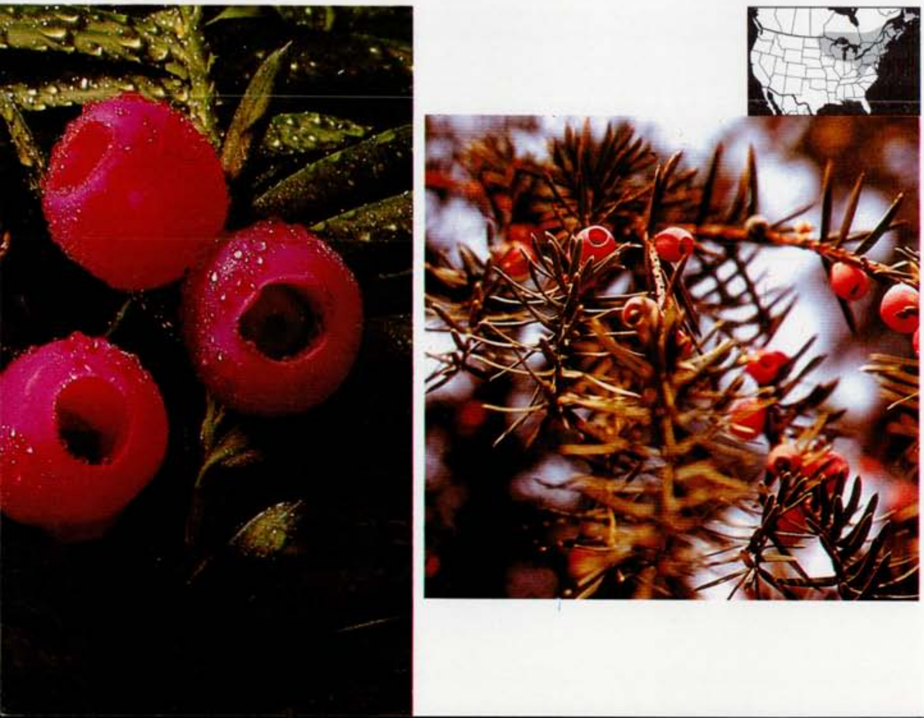

Southwestern coral bean *Erythrina flabelliformis*

Habitat: canyons, washes, desert grasslands, open oak woodlands. **Identification:** large spiny shrub or small tree to 5 m (16.4 ft); ***stems leafless except in summer,*** thick, brittle, armed with short spines. ***Leaves alternate, featherlike (pinnately) compound with 3 leaflets, the leaflets broadly triangular*** and tapering to a point, 2.5–7.5 cm (1–3 in) long, 4–8.7 cm (1.5–3.5 in) wide, grayish green; main leafstalk long, armed with several short, hooked spines. ***Flowers in large, showy clusters usually in spring, each narrow, elongated, pea-shaped, with 5 bright red to scarlet petals.*** Fruits are long, leathery pods 10–25 cm (4–10 in) long, constricted between the seeds; seeds hard, bright red, 1–2 cm (0.4–0.8 in) long. **Poisonous constituents:** Seed wall contains highly poisonous alkaloids that can cause death, even when consumed in small amounts. The seeds are sometimes used to make necklaces, which some wearers unfortunately have played with between their teeth. **Related Poisonous Species:** All species of *Erythrina* are poisonous. Eastern coral bean *(E. herbacea)* is native to the Coastal Plain from North Carolina to Texas.

American yew *Taxus canadensis*

Habitat: rich woodlands, pine forests, bogs, thickets. **Identification:** ***shrub, usually low, spreading and straggly,*** occasionally upright, up to 2 m (6.6 ft) high, branches turning reddish-brown with age. ***Leaves spirally arranged and spreading in 2 ranks along stems, narrow (linear), 1–2 cm (0.4–0.8 in) long, about 2 mm (0.1 in) wide, abruptly fine-pointed at tip, dark yellowish-green above, with 2 pale-green bands beneath,*** leafstalk short. Flowers tiny cones, either male or female. ***Fruits consist of a hard, dark, egg-shaped seed about 1 cm (0.4 in) long, enclosed in bright red to reddish-orange, fleshy disc.*** **Poisonous constituents:** Leaves and seeds contain one or more alkaloids called taxine, which can cause sudden death in people as well as horses and cattle. The fleshy red disc around the seed is edible but is better avoided because of the highly toxic seeds. **Related poisonous species:** Leaves and seeds of all species of yew *(Taxus)* are extremely poisonous. The introduced English yew *(T. baccata)* and the Japanese yew *(T. cuspidata)* are widely planted as ornamentals.

Poison ivy *Toxicodendron radicans*

(synonym: *Rhus radicans*)

Habitat: woodland margins, pastures, streams and river banks, fencerows, and disturbed sites. **Identification:** low shrub or vine, often climbing, with hairy, light-brown twigs. ***Leaves*** alternate, deciduous, ***compound of three leaflets*** (trifoliolate), leaflets broadest near base, 10–16 cm (3.9–6.3 in) long, 5–10 cm (2–3.9 in) wide, pointed at tip, ***entire or with a few coarse teeth along margin,*** yellowish-green, dull, often smooth, ***leafstalks reddish near junction with blade.*** Plants either male or female. Tiny male flowers produced in elongated, branched clusters 5–10 cm (2–3.9 in) long on new growth; small female flowers produced in smaller branched clusters 4–6 cm (1.6–2.4 in) long. ***Fruits are rounded,*** 4–6 mm (0.1 in) in diameter, ***creamy white,*** smooth, ***produced in clusters like grapes.*** **Poisonous constituents:** All parts of the plant have an oily resin containing urushiol. Contact with urushiol can cause minor to severe dermatitis, consisting of a rash and blisters. This affliction can be serious and most uncomfortable in the mouth, throat, and nasal passages as a result of eating berries or breathing smoke produced from burning plants. **Related poisonous species:** Poison oak *(Rhus toxicodendron)*, of sandy soils, pine forest, or barrens in the eastern U.S., and *Rhus diversiloba,* also called poison oak, of the Pacific coastal states, both contain oily resin capable of producing dermatitis like that caused by poison ivy. ☠

Ohio buckeye, Horsechestnut *Aesculus glabra*

Habitat: woodlands, hillsides, especially riverbottoms and stream banks. **Identification:** broad spreading tree with open, rounded crown and dark-brown, deeply furrowed bark. ***Leaves opposite,*** deciduous, ***hand-shaped compound*** (palmate) ***with 5 or 7 leaflets,*** leaflets usually widest near middle, 10–15 cm (4–6 in) long, 3.5–6 cm (1.4–2.4 in) wide, pointed at tip, finely toothed along margin, yellowish-green, leafstalks long. Flowers produced in large, erect clusters 12.5–15 cm (5–6 in) long, pale yellowish green, with 5-lobed calyx, 4 petals sometimes with red stripes. ***Fruits a leathery, 3-parted capsule 2.5–5 cm*** (1–2 in) ***long, with short blunt spines on outer surface*** of husk, ***containing 1 or 2 large, shiny, dark reddish-brown seeds.*** **Poisonous constituents:** All plant parts, especially the large seeds, contain aesculin, a glycoside. If ingested, this toxin can cause vomiting, stupor, twitching, paralysis, and possibly even death. Seeds of buckeyes or horsechestnuts resemble those of edible chestnuts. **Related poisonous species:** All native species of the buckeye genus *(Aesculus)* should be considered poisonous. Compounds from some species have been used medicinally. A coumarin glycoside, esculoside, obtained from the European horse chestnut *(A. hippcastanum)*, has been used to treat hemorrhoids. ☠

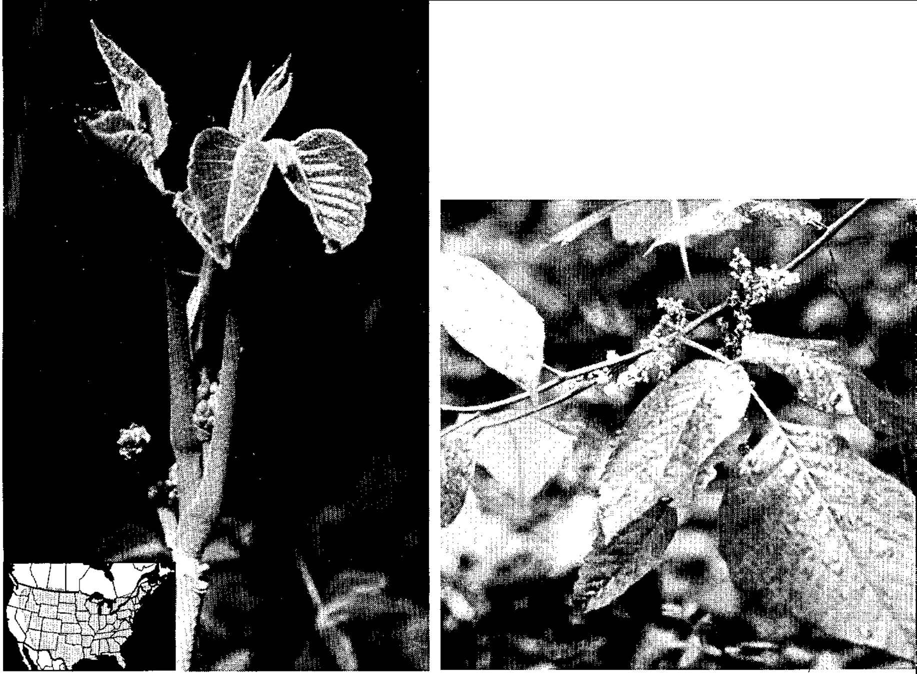

Poison ivy: immature plant with bronze leaves and fruiting branch with shiny green leaves and light-green female flower clusters (NYBG).

Ohio buckeye: Fruiting branch (Clawson).

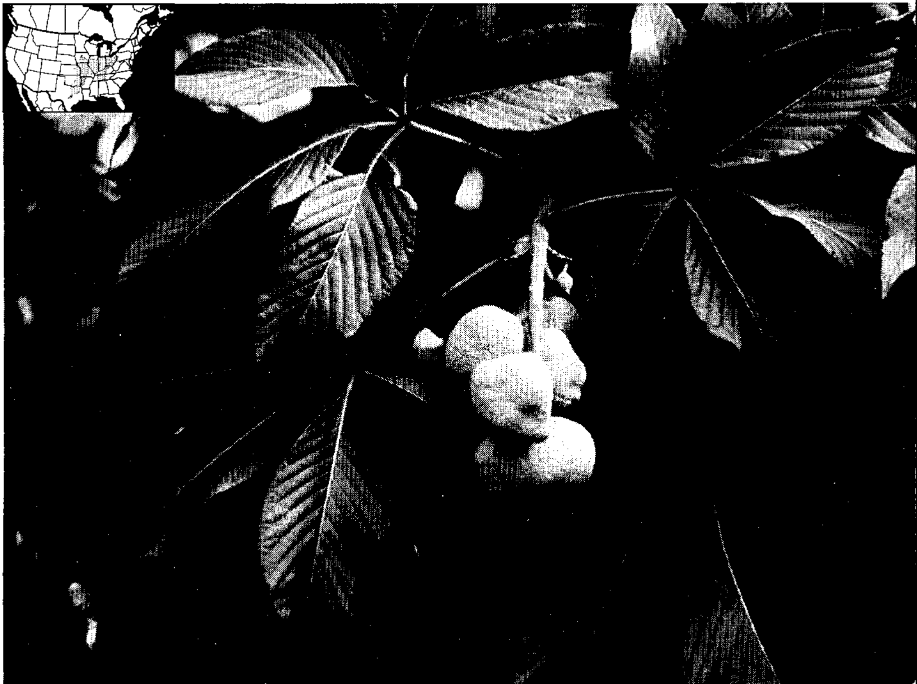

Nutritional Contents of Selected Foods

Per 100 gram portion (about 3.5 ounces)

		ENERGY BUILDING BLOCKS			MINERALS					VITAMINS				
Food	Calories	Protein (g)	Fat (g)	Carbohydrates (g)	Calcium (mg)	Phosphorus (mg)	Iron (mg)	Sodium (mg)	Potassium (mg)	Vitamin A (IU)	Thiamine (mg)	Riboflavin (mg)	Niacin (mg)	Ascorbic acid (mg)
Apples, raw	58	0.2	0.6	14.5	7	10	0.3	1	110	90	.03	.02	.1	4
Asparagus, raw	26	2.5	0.2	5.0	22	62	1.0	2	278	900	.18	.20	1.5	33
Bananas, raw	85	1.1	0.2	22.2	8	26	0.7	1	370	190	.05	.06	.7	10
Beans, lima, boiled, and drained	111	7.6	0.5	19.8	47	121	2.5	1	422	280	.18	.10	1.3	17
Beef, broiled	207	32.2	7.7	0.0	13	261	3.9	60	370	10	.09	.25	6.4	0
Beverages: Beer	42	0.3	0.0	3.8	5	30	tr	7	25	0	tr	.03	.6	0
Wine	85	0.1	0.0	4.2	9	10	0.4	5	92	0	tr	.01	.1	0
Blackberries, raw	58	1.2	0.9	12.9	32	19	0.9	1	170	200	.03	.04	.4	21
Bread, whole wheat	243	10.5	3.0	48.0	99	228	2.3	527	273	tr	.26	.12	2.8	tr
Broccoli, raw	32	3.6	0.3	5.9	103	78	1.1	15	382	2,500	.10	.23	.9	113
Carrots, raw	42	1.1	0.2	9.7	37	36	0.7	47	341	11,000	.06	.05	.6	8
Cauliflower, raw	27	2.7	0.2	5.2	25	56	1.1	13	295	60	.11	.10	.7	78
Chicken, white meat, roasted	182	32.3	4.9	0.0	11	272	1.3	66	422	110	.08	.10	11.8	0
Chicory greens, raw	20	1.8	0.3	3.8	86	40	0.9	0	420	4,000	.06	.10	.5	22
Cucumbers, raw	15	0.9	0.1	3.4	25	27	1.1	6	160	250	.03	.04	.2	11
Dandelion greens, raw	45	2.7	0.7	9.2	187	66	3.1	76	397	14,000	.19	.26	.0	35
Dock & sheep sorrel, raw	28	2.1	0.3	5.6	66	41	1.6	5	338	12,900	.09	.22	.5	119
Eggs, raw or boiled	163	12.9	11.5	0.9	54	205	2.3	122	129	1,180	.11	.30	.1	0
Grapes, raw	69	1.3	1.0	15.7	16	12	0.4	3	158	100	.05	.03	.3	4
Hazelnuts	634	12.6	62.4	16.7	209	337	3.4	2	704	0	.46	.00	.9	tr
Hickorynuts	673	13.2	68.7	12.8	tr	360	2.4	0	0	0	.00	.00	.0	0
Jerusalem artichoke, raw, fresh	7	2.3	0.1	16.7	14	78	3.4	0	0	20	.20	.06	1.3	4

Nutritional Contents of Selected Foods (continued)

Food	Calories	Protein (g)	Fat (g)	Carbohydrates (g)	Calcium (mg)	Phosphorus (mg)	Iron (mg)	Sodium (mg)	Potassium (mg)	Vitamin A (IU)	Thiamine (mg)	Riboflavin (mg)	Niacin (mg)	Ascorbic acid (mg)
Lamb's-quarters (Pigweed), raw	43	4.2	0.8	7.3	309	72	1.2	0	0	11,600	.16	.44	1.2	80
Leeks, bulb & lower leaf, raw	52	2.2	0.3	11.2	52	50	1.1	5	347	40	.11	.06	.5	17
Lettuce, raw	14	1.0	0.2	2.9	35	26	2.0	9	264	970	.06	.06	.3	8
Milk, cows, whole	65	3.5	3.5	4.9	118	93	tr	50	144	140	.03	.17	.1	1
Mustard greens, raw	31	3.0	0.5	5.6	183	50	3.0	32	377	7,000	.11	.22	.8	97
Onions, bulb, raw	38	1.5	0.1	8.7	27	36	0.5	10	157	tr	.03	.04	.2	10
Papaw, raw	85	5.2	0.9	16.8	0	0	0.0	0	0	0	.00	.00	.0	0
Parsnips, raw	76	1.7	0.5	17.5	50	77	0.7	12	541	30	.08	.09	.2	16
Peanuts, with skins	564	26.0	47.5	18.6	69	401	2.1	5	674	0	1.14	.13	17.2	0
Pears, raw	61	0.7	0.4	15.3	8	11	0.3	2	130	20	.02	.04	.1	4
Peas, green, immature, raw	84	6.3	0.4	14.4	26	116	1.9	2	316	640	.35	.14	2.9	27
Pecans	687	9.2	71.2	14.6	73	289	2.4	tr	603	130	.86	.13	.9	2
Persimmons, raw	127	0.8	0.4	33.5	27	26	2.5	1	310	0	.00	.00	.0	66
Plums, raw	66	0.5	0.0	17.8	18	17	0.5	2	299	300	.08	.03	.5	0
Potatoes, raw	76	2.1	0.1	17.1	7	153	0.6	3	407	tr	.10	.04	1.5	20
Prickly-pears, raw	42	0.5	0.1	10.9	20	28	0.3	2	166	60	.01	.03	.4	22
Purslane, leaves & stems, raw	21	1.7	0.4	3.8	103	39	3.5	0	0	2,500	.03	.10	.5	25
Raspberries, raw	57	1.2	0.5	13.6	22	22	0.9	1	168	130	.03	.09	.9	25
Rice, brown, raw	360	7.5	1.9	77.4	32	221	1.6	9	214	0	.34	.05	4.7	0
Salsify, fresh, raw	13	2.9	0.6	18.0	47	66	1.5	0	380	10	.04	.04	.3	11
Strawberries, raw	37	0.7	0.5	8.4	21	21	1.0	1	164	60	.03	.07	.6	59
Sunflower seed kernels, dry	560	24.0	47.3	19.9	120	837	7.1	30	920	50	1.96	.23	5.4	0
Walnuts, black	628	20.5	59.3	14.8	tr	570	6.0	3	460	300	.22	.11	.7	0
Watercress, leaves & stems, raw	19	2.2	0.3	3.0	151	54	1.7	52	282	4,900	.08	.16	.9	79

Source: *Handbook of the Nutritional Contents of Foods,* U.S. Department of Agriculture, Dover Publications Inc., New York 1975.

g = grams, mg = milligrams, IU = international units, tr = trace

General Index

Autumn, edibles characteristics of, 12, 208
Baking soda (tenderizer), 13
Beverages, 14–15
Bitterness. *See* Toxicity
Brambles, blackberries, raspberries, harvest and preparation of, 178–179

Coffee substitutes, 15
Currants and gooseberries, harvest and preparation of, 168–169
Dicots. *See* Monocots and dicots
Diet and wild edibles, 27
Flour substitutes, 14, 22
Food value of wild edibles, 274–275
Foraging for plants, 10–11
Fruit
  drying, 15–16
  handpicking, 15
  solar drying, 16

Gooseberries. *See* Currants
Grains, nuts, and seeds
  as flour substitute, 22
  winnowing methods for, 22
Grapes. *See* Wild grapes

Harvest and preparation of plants, 12–22
  boiling, 12–13
  edibility through growing season, 12
  handpicking, 12, 15
  pit-roasting, 14
  roots and tubers, 13–14
  shoots and leaves, 12–13
Harvest dates, regional calendar for, 12
Hawthorns, harvest and preparation of, 237

Indian: wild food harvesting, 22–24; food preparation, 23–24; survival techniques, 249

Jams, 18–20
  without pectin, 20
  recipes (table), 18–19
  uncooked, 20
Jellies, 16–18
  and pectin, 16
  recipes (table), 17

Leaves. *See* Plant parts, aboveground

Maple sugaring, 24
Mapping populations of edible plants, 208
Mast crop, 208
Medicinal use of plants, 15
Monocots and dicots, contrasting features of (table), 6

Nutrition and wild edibles
  carbohydrates, 26
  fats, 26
  Nutritional Contents of Selected Foods (appendix), 274–275
  vitamins and minerals (table), 26–27
Nuts. *See* Grains, nuts, and seeds

Overharvesting, 9, 208

Pemmican (recipe), 169
Pies, 20–22
  recipes (table), 20–21
Plant guides, 28
Plant habitats, 10–11
Plant names, 6
Plant parts, aboveground, 12
Plant parts, underground, 13–14
  difficulty in harvesting, 13, 248
  as trail nibbles, 14
Plant reproduction, 7–8
Plants, defensive structures of, 8–9
Plants, introduced, 9–10

Plant taste, factors influencing, 9
Poisonous plants, recognition of, 258
Potato substitutes, 13
Powcohicora, 243

Raspberries. *See* Brambles, blackberries, raspberries
Rennet substitute, 79
Roots. *See* Plant parts, underground
Roses, harvest and preparation of, 220–221
Sap, use of in beverages, 16, 249
Seeds, dispersal of, 9. *See also* Grains, nuts, and seeds
Shoots. *See* Plant parts, aboveground
Spring, edibles characteristics of, 11, 56
Stone grinding, 24
Summer, edibles characteristics of, 11, 126
Survival foods, 248–249

Tea from wild edibles, 14–15
Toxicity, removal of, 9
  by boiling, 12–13
  by drying, 14
Vitamin C, 14
Vitamins
  and baking soda, 13
  loss and retention of in cooking, 12, 14
White Oak Group, harvest and preparation of, 228
Wild grapes, harvest and preparation of, 214–215
Winnowing. *See* Grains, nuts, and seeds
Winter, edibles characteristics of, 248
Winter foraging, 12, 249

Index to plant names

Bold face type is used for common names and pages where you'll find plants both described and illustrated. ***Boldface italic*** is used for scientific names of described and illustrated plants. The thinner typeface (normal and *italic*) is used for plants mentioned incidentally, for example, when they resemble a plant covered in full detail.

Acer platanoides, 126, 257
Acer rubrum, 257
Acer saccharum, **125, 257**
Acorus calamus, **133**
Aesculus species, 234
Aesculus glabra, **272**
Aesculus hippocastanum, 272
Agave species, 134
Alaskan knotweed, 99, 155
alfilaria, 109
Allegheny chinkapin, 234
Alliaria officinalis, **85**
Allium canadense, **61**
Allium cernuum, **58**
Allium tricoccum, **61**
Allium vineale, 58
alpine bearberry, 159
alpine bistort, 155
alpine smartweed, 155
alpine spring beauty, 92
Althaea officinalis, **146**
Amaranthus hybridus, **72**
Amaranthus retroflexus, **72**
Amelanchier, **207**
Amelanchier alnifolia, **207**
Amelanchier arborea, **207**
Amelanchier florida, **207**
Amelanchier laevis, **207**
American barberry, 219
American beech, 233
American bistort, 155
American brooklime, 144
American chestnut, 234
American elder, 190
American hackberry, 220
American hazelnut, 173
American lotus, 210
American mountain ash, 240, 254
American plum, 203
American yew, 271
Apios americana, **149**
Apios priceana, 149
Apocynum androsaemifolium, **267**
Arctium lappa, 112
Arctium minus, **112**
Arctostaphylos alpina, 159
Arctostaphylos uva-ursi, **159**
Arizona walnut, 244, 247
arrowhead, 213
artichoke, Jerusalem, 210
Asclepias speciosa, 106
Asclepias syriaca, **106**
Asclepias tuberosa, **267**
Asimina parviflora, 198
Asimina triloba, **198**
asparagus, 62
Asparagus officinalis, **62**
Astragalus species, 149
atamasco lily, 261

Barbarea verna, 99
Barbarea vulgaris, **99**
barberry: American, 219; **common, 219;** red-fruited, 219
basswood, 15
bayberry: northern, 174; **southern, 174**
bay, California, 198
beach plum, 167
beaked hazelnut, 173
bearberry, 159; alpine b., 159
bedstraw, 86; northern b., 86
beech, American, 233
Berberis canadensis, 219
Berberis haematocarpa, 219
Berberis vulgaris, **219**
Berlandier's lamb's quarters, 75
Betula alleghaniensis, **122, 254**
Betula lenta, **122, 253**
Betula lutea, **122, 254**
big shellbark hickory, 243
bilberry, bog, 164
birch: sweet (black), 122, 253; yellow, 122, 254
bistort: alpine, 155; American, 155; western, 155
black berried elder, 190
Blackberries, 178
blackberry: California, 185; common, 183

black birch, 122, 253
black cherry, 204
black haw, 227
black huckleberry, 167
black mulberry, 200
black mustard, 100
black nightshade, 265
black raspberry, 183
black walnut, 244; Texas b.w., 244, 247
Blueberry and Cranberry Genus, 163
blueberry: highbush, 164; late sweet, 164; swamp, 164; velvet-leaved, 163
blue elder, 190
blue flag, 261
blue oak, 231
bog bilberry, 164
box huckleberry, 167
bracken fern, 58
Brambles, Blackberries, Raspberries, 178
Brassica juncea, 100
***Brassica nigra,* 100**
Brassica rapa, 100
breadroot, common, 150
brooklime, American, 144
buckeye, Ohio, 272
buckwheat, tartary, 155
buffaloberry: Canadian, 179; silver, 179
bugleweed, 89
bullbrier, 66
bull thistle, 115
bulrush: great, 70; hard-stem, 70
bunchberry, 143; Swedish b., 143
burdock: common, 112; great, 112
bush morning glory, 216
buttercup, common, 262
butterfly weed, 267
butternut, 247

cactus, saguaro, 5
calamus, 133
California bay, 198
California blackberry, 185
California grape, 216
California juniper, 194
California laurel, 198
California rose, 220
California walnut, 247
***Caltha palustris,* 100**
camass: death, 258; eastern, 65; western, 136
camass lily, 65, 136
***Camassia quamash,* 136**
***Camassia scilloides,* 65**
Canada violet, 95
Canadian buffaloberry, 179
***Capsella bursa-pastoris,* 81**
Capsella gracilis, 81
Capsella rubella, 81
Carolina hemlock, 122
Carolina laurelcherry, 204
Carolina spring beauty, 92
carrion flower, 66
***Carya alba,* 244**
***Carya carolinae-septentrionalis,* 243**
***Carya illinoensis,* 240**
***Carya laciniosa,* 243**
***Carya ovata,* 243**
***Carya tomentosa,* 244**
Castanea dentata, 234
***Castanea pumila,* 234**
catbrier, 65
cattail, 69; narrow-leaved c., 69
***Ceanothus americanus,* 176**
Ceanothus ovatus, 176
Celtis occidentalis, 220
***Celtis pallida,* 220**
Cerastium vulgatum, 85
Cereus giganteus, 5
checkerberry, 150
cheese mallow, 146
***Chenopodium album,* 75**
Chenopodium berlandieri, 75
Chenopodium fremontii, 75
cherry: black, 204; ground, 149
chestnut, American, 234
chestnut oak, 231
chickweed: common, 85; James, 85; mouse-ear, 85; star, 85
chicory, 115
China-root, 66
Chinese mustard, 100
chinkapin, Allegheny, 234
chokecherry, common, 204
chufa, 130
***Cichorium intybus,* 115**
***Cicuta maculata,* 268**
cinnamon fern, 58
***Cirsium vulgare,* 115**
Claviceps species, 129
Claytonia caroliniana, 92
***Claytonia lanceolata,* 92**
***Claytonia megarhiza,* 92**
***Claytonia perfoliata,* 95**
***Claytonia virginica,* 91**
cleavers, 86
***Clintonia borealis,* 62**
Clintonia umbellata, 62
clintonia, white, 62
Clinton's lily, 62
clover, 15

Coccoloba diversifolia, 197
***Coccoloba uvifera*,** 197
cockspur hawthorn, 238
common chokecherry, 204
common barberry, 219
common blackberry, 183
common breadroot, 150
common burdock, 112
common buttercup, 262
common chickweed, 85
common day lily, 136
common elder, 190
common greenbrier, 65
common milkweed, 106
common persimmon, 238
***Comptonia peregrina*,** 174
***Conium maculatum*,** 268
coral bean: eastern, 271; **southeastern, 271**
corn lily, 62
***Cornus canadensis*,** 143
***Cornus suecica*,** 143
***Corylus americana*,** 173
***Corylus cornuta*,** 173
Coulter pine, 193
cow cress, 82
cow lily, 213
cowslip, 100
cranberry, 163; highbush c., 224, 250;
  mountain c., 163; small, 163
cranberry tree, 224, 250
***Crataegus*,** 237
***Crataegus acutifolia*,** 238
Crataegus aestivalis, 237
***Crataegus arnoldiana*,** 237
***Crataegus canadensis*,** 237
***Crataegus canbyi*,** 238
***Crataegus crus-galli*,** 238
***Crataegus mollis*,** 237
***Crataegus phaenopyrum*,** 237
***Crataegus pyracanthoides*,** 238
***Crataegus regalis*,** 238
***Crataegus submollis*,** 237
***Crataegus succulenta*,** 238
creeping snowberry, 153
creeping wintergreen, 160
cress: cow, 82; field, 82
Crotalaria species, 149
Cruciferae, 82
cucumber, Indian, 139
curled dock, 121
currant, fragrant golden, 170; golden, 169;
  sticky, 169; **subalpine prickly, 170;**
  wax, 170
Currant or Gooseberry Genus, 168

***Cyperus esculentus*,** 130

dandelion, 104
dangleberry, 167
datil, 134
***Datura stramonium*,** 262
deadly nightshade, 265
death camass, 258
***Dentaria laciniata*,** 81
desert hackberry, 220
digger pine, 193
Diospyros texana, 238
***Diospyros virginiana*,** 238
dock: curled, 121; sour, 121; yellow, 121
dogbane, 267
dove plum, 197
downy hawthorn, 237
downy serviceberry, 207
duck potato, 213
dwarf fireweed, 109
dwarf huckleberry, 167
dwarf mallow, 146
dwarf pawpaw, 198

early wintercress, 99
eastern camass, 65
eastern coral bean, 271
eastern hemlock, 122
edible valerian, 89
Elaeagnus argentea, 179
elder: American, 190; black berried, 190;
  blue, 190; common, 190; Mexican, 190;
  Pacific, 190; red-fruited, 190
English yew, 271
***Epilobium angustifolium*,** 109
Epilobium latifolium, 109
***Erodium cicutarium*,** 109
***Erythrina flabelliformis*,** 271
Erythrina herbacea, 271
European horse chestnut, 272
European mountain ash, 240, 254
European nightshade, 265
evening primrose, 103

Fagaceae, 234
***Fagus grandifolia*,** 233
false hellebore, 265
fern: bracken, 58; cinnamon, 58; **ostrich, 58;**
  sweet, 174
field cress, 82
field garlic, 58
field mustard, 100

field pennycress, 82
fig, Indian, 139, 140
filaree, 109
fireweed, 109; dwarf f., 109
flag: blue, 261; sweet, 133
flatwood plum, 203
fleshy hawthorn, 238
foxberry, 163
Fragaria virginiana, 91
fragrant golden currant, 170
Fremont's lamb's quarters, 75
Fremont's peppergrass, 82
frost grape, 214

galinsoga, 159
Gallinsoga ciliata, 159
Galinsoga parviflora, 159
Galium aparine, 86
Galium boreale, 86
Gambel's oak, 231
garlic: field, 58; **wild, 61**
garlic mustard, 85
Gaultheria hispidula, 153
Gaultheria humifusa, 160
Gaultheria ovatifolia, 153, 160
Gaultheria procumbens, 150
Gaultheria shallon, 160
Gaylussacia baccata, 167
Gaylussacia brachycera, 167
Gaylussacia dumosa, 167
Gaylussacia frondosa, 167
giant knotweed, 99, 155
glandular mesquite, 189
glasswort, 76
goat's beard, 106; yellow g.b., 106
golden currant, 169
Gooseberry Genus, 168
goosegrass, 86
grape: California, 216; frost, 214; riverside, 214; sea, 197
grass: reed, 129; yellow nut, 130
great bulrush, 70
great burdock, 112
greenbrier, 66; common g., 65
Greene mountain ash, 240, 254
green pigweed, 72
Grindelia species, 8
ground cherry, 149
groundnut _(Apios americana),_ 149
groundnut *(Apios priceana),* 149
guelder rose, 224, 250, 253

hackberry: American, 220; **desert, 220; spiny, 220**

hard-stem bulrush, 70
haw, black, 227
hawthorn: cockspur, 238; downy, 237; fleshy, 238; May, 237; **Washington, 237**
Hawthorns, 237
hazelnut: American, 173; beaked, 173
Helianthus annus, 156
Helianthus tuberosus, 210
hellebore: false, 265; white, 265
Hemerocallis fulva, 136
hemlock: Carolina, 122; **eastern, 122;** mountain, 122; **poison, 268; water, 268;** western, 122
henbit (Lamium amplexicaule), 116
henbit *(Lamium purpureum),* 116
Hibiscus esculentus, 106
hickory: big shellbark, 243; mockernut, 244; shagbark, 243
highbush blueberry, 164
highbush cranberry, 224, 250
Hinds walnut, 244, 247
Hippocastanaceae, 234
hobblebush, 224
horehound, water, 89
horsechestnut, 234, 272
horse chestnut, European, 273
horse nettle, 265
huckleberry: black, 167; box, 167; dwarf, 167

Impatiens biflora, 119
Impatiens capensis, 119
Impatiens pallida, 119
Indian cucumber, 139
Indian fig, 139, 140
Ipomoea leptophylla, 216
Ipomoea pandurata, 216
Iris pseudoacorus, 261
Iris versicolor, 261

jack-in-the-pulpit, 14
James chickweed, 85
Japanese knotweed, 99
Japanese yew, 271
Jerusalem artichoke, 210
jewelweed: orange-flowered, 119; yellow-flowered, 119
jimson weed, 262
Joshua tree, 5
Juglans californica, 247
Juglans cinerea, 247
Juglans hindsii, 244, 247
Juglans major, 244, 247
Juglans microcarpa, 244, 247

Juglans nigra, 244
juniper: California, 194; Rocky Mountain, 194; Utah, 197
Juniper Genus, 194
Juniperus, 194
Juniperus californica, 194
Juniperus osteosperma, 197
Juniperus scopulorum, 194
Juniperus utahensis, 197

kinnikinick, 159
knotweed: Alaskan, 99, 155; Japanese, 99; giant, 99, 155

Labrador tea, 160; northern L.t., 160
Lactuca canadensis, 104
Lactuca scariola, 104
Lactuca serriola, 104
lady's thumb, 110
lamb's quarters, 75; Berlandier's l.q., 75; Fremont's l.q., 75
Lamium amplexicaule, 116
Lamium purpureum, 116
lance-leaved pickerelweed, 133
Laporta canadensis, 79
late sweet blueberry, 164
Lathyrus species, 149
laurel, California, 198
laurelcherry, Carolina, 204
Ledum glandulosum, 160
Ledum groenlandicum, 160
Ledum palustre, 160
leek, wild, 61
lemonade berry, 186, 250
lemonade sumac, 185
Lepidium campestre, 82
Lepidium fremontii, 82
Lepidium virginicum, 82
lettuce: miner's, 95; prickly, 104; wild, 104
lily: atamasco, 261; camass, 65, 136; Clinton's, 62; common day, 136; corn, 62; cow, 213; lotus, 210; water, 210; yellow pond, 213
Lindera benzoin, 219
lingenberry, 163
live-forever, 110
locoweed, point, 149
lotus, American, 210
lotus lily, 210
Lycopus amplecteris, 89
Lycopus uniflorus, 89

mallow: cheese, 146; dwarf, 146; marsh, 146
Malva neglecta, 146
Malva rotundifolia, 146
mandrake, 140
maple: Norway, 126, 257; red, 257; sugar, 125, 257
marigold, marsh, 100
marsh mallow, 146
marsh marigold, 100
Matteucia pensylvanica, 58
Matteucia struthiopteris, 58
May apple, 140, 262
May hawthorn, 237
Medeola virginiana, 139
Melilotus alba, 267
Melilotus officinalis, 265
Memispermum canadense, 215
Mentha piperita, 143
Mentha spicata, 144
mesquite, 189; glandular, 189; screwbean m., 189
Mexican elder, 190
Mexican pinyon, 193, 228
milkweed, common, 106; showy, 106
milk vetch, 149
miner's lettuce, 95; Siberian m.l., 95
mockernut hickory, 244
Monolepis nuttalliana, 75
Montia perfoliata, 95
Montia sibirica, 95
moonseed, 215
morning glory, bush, 216
Morus alba, 200
Morus microphylla, 200
Morus nigra, 200
Morus rubra, 200
mountain ash: American, 240, 254; European, 240, 254; Greene, 240, 254; Sitka, 240, 254
mountain cranberry, 163
mountain hemlock, 122
mountain Labrador tea, 160
mountain sorrel, 153
mouse-ear chickweed, 85
mulberry: black, 200; red, 200; Texas, 200; white, 200
mustard: black, 100; Chinese, 100; field, 100; garlic, 85
Myrica cerifera, 174
Myrica gale, 176
Myrica pensylvanica, 174
myrtle, wax, 174

nannyberry, 227
narrow-leaved cattail, 69
Nasturtium officinale, 146

***Nelumbo lutea,* 210**
nettle, 79; horse n., 265; slender n., 79; **stinging n., 79;** wood n., 79
New Jersey tea, 176
nightshade: black (deadly), 265; European, 265; silverleaf, 265
nodding onion, 58
northern bayberry, 174
northern bedstraw, 86
northern Labrador tea, 160
northern red oak, 233
northern wild raisin, 227
Norway maple, 126, 257
***Nuphar advena,* 213**
Nuphar polysepalum, 213
nut grass, yellow, 130
Nymphaea species, 210

oak: blue, 231; chestnut, 231; Gambel's, 231; northern red, 233; white, 228
Oaks, 228
***Oenothera biennis,* 103**
Ohio buckeye, 272
okra, 106
onion: nodding, 58; wild, 58
***Opuntia humifusa,* 139**
***Opuntia phaeacantha,* 140**
orange-flowered jewelweed, 119
Ornithogalum nutans, 258
***Ornithogalum umbellatum,* 258**
orphine, 110
ostrich fern, 58
Osumunda cinnamomea, 58
***Oxyria digyna,* 153**
Oxytropis species, 149
oyster plant, 112

Pacific elder, 190
pale touch-me-not, 119
pale violet, 96
papaw, 198; dwarf p., 198
Parry pinyon, 228
pasture rose, 223
pawpaw, 198
pear, prickly, 139, 140
pea, scurf, 150
pecan, 240
pennycress, 82; **field p., 82**
peppergrass, 82; Fremont's p., 82
peppermint, 143
persimmon: common, 238; Texas, 238
phragmites, 129
***Phragmites communis,* 129**
***Physalis pubescens,* 149**
***Phytolacca americana,* 96, 267**
pickerelweed, 130; lance-leaved p., 133
pigweed, 72; green p., 72
pine: Coulter, 193; digger, 193; **pinyon, 228; singleleaf pinyon, 193; sugar, 193**
Pinus cembroides, 193, 228
Pinus coulteri, 193
***Pinus edulis,* 228**
***Pinus lambertiana,* 193**
***Pinus monophylla,* 193**
Pinus quadrifolia, 228
***Pinus sabiniana,* 193**
pinyon: Mexican, 193, 228; Parry, 228
pinyon pine, 228
Plantago juncoides, 79
***Plantago major,* 79**
plantain, 79; seaside p., 79
plum: American, 203; beach, 167; dove, 197; **flatwood, 203; wild, 203**
***Podophyllum peltatum,* 140, 262**
point locoweed, 149
poison hemlock, 268
poison ivy, 272
poison oak, 185, 186, 272
poison sumac, 185, 250
pokeweed, 96, 267
Polygonum alaskanum, 99, 155
***Polygonum bistortoides,* 155**
***Polygonum cuspidatum,* 99**
***Polygonum persicaria,* 110**
Polygonum sachalinense, 99, 155
Polygonum tataricum, 155
***Polygonum viviparum,* 155**
***Pontederia cordata,* 130**
***Pontederia lanceolata,* 133**
Portulaca neglecta, 156
***Portulaca oleracea,* 156**
Portulaca retusa, 156
potato, duck, 213
potato vine, wild, 216
poverty weed, 75
prickly lettuce, 104
prickly pear, 139, 140
primrose, evening, 103
***Prosopis glandulosa,* 189**
***Prosopis pubescens,* 189**
Prosopis veluntina, 189
***Prunus americana,* 203**
Prunus caroliniana, 204
***Prunus maritima,* 167**
***Prunus serotina,* 204**
***Prunus umbellata,* 203**
***Prunus virginiana,* 204**
***Psoralea esculenta,* 150**

Psoralea hypogala, 150
Psoralea tenuifolia, 150
Pteretis nodulosa, 58
Pteridium aquilinum, 58
purslane, 156
pusley, 156

Quercus alba, 228
Quercus douglasii, 231
Quercus subgenus _Erythrobalanus_, 233
Quercus gambelii, 231
Quercus prinus, 231
Quercus rubra, 233
quickweed, 159

raisin, northern wild, 227
ramp, 61
Ranunculus acris, 262
Raspberries, 178
raspberry: black, 183; red, 180
rattlebox, 149
Red/Black Oak Group, 233
red-fruited barberry, 219
red-fruited elder, 190
red maple, 257
red mulberry, 200
red oak, northern, 233
red raspberry, 180
redroot (_Amaranthus retroflexus_), 72
red root (*Ceanothus ovatus*), 176
red shepherd's purse, 81
reed grass, 129
Rhus diversiloba, 186, 272
Rhus glabra, 185
Rhus integrifolia, 185, 186, 250
Rhus ovata, 185
Rhus radicans, 272
Rhus toxicodendron, 272
Rhus trilobata, 186
Rhus typhina, 186, 250
Rhus vernix, 250
Ribes, 168
Ribes aureum, 169
Ribes cereum, 170
Ribes montigenum, 170
Ribes odoratum, 170
Ribes viscosissimum, 169
rice, wild, 129
riverside grape, 214
Rocky Mountain juniper, 194
root, red, 176
Rorippa nasturtium-aquaticum, 146
Rosa, 220
Rosa californica, 220
Rosa carolina, 223
Rosa eglanteria, 223
Rosa rugosa, 223
rose: California, 220; guelder, 224, 250, 253; pasture, 223; sweetbrier, 223; wild, 223; wrinkled, 223
roseroot, 110
Roses, 220
Rubus, 178
Rubus allegheniensis, 183
Rubus idaeus var. _strigosus_, 180
Rubus occidentalis, 183
Rubus parviflorus, 180
Rubus phoenicolasius, 180
Rubus strigosus, 180
Rubus ursinus, 185
Rumex acetosella, 121
Rumex crispus, 121
Russian thistle, 76

Sagittaria latifolia, 213
saguaro cactus, 5
salal, 160
Salicornia europaea, 76
Salicornia virginica, 76
salsify, 112
Salsola kali, 76
saltwort, 76
Sambucus callicarpa, 190
Sambucus canadensis, 190
Sambucus cerulea, 190
Sambucus melanocarpa, 190
Sambucus mexicana, 190
Sambucus pubens, 190
Saskatoon serviceberry, 207
sassafras, 125
Sassafras albidum, 125
Scirpus acutus, 70
Scirpus validus, 70
screwbean mesquite, 189
scurf pea, 150
sea grape, 197
seaside plantain, 79
Sedum purpureum, 110
Sedum rosea, 110
Serviceberry: downy, 207; Saskatoon, 207
Serviceberry Genus, 207
shagbark hickory, 243
sheep sorrel, 121
Shepherdia argentea, 179
Shepherdia canadensis, 179
shepherd's purse, 81; red s.p., 81; slender s.p., 81
showy milkweed, 106

Siberian miner's lettuce, 95
silverberry, 179
silver buffaloberry, 179
silverleaf nightshade, 265
singleleaf pinyon pine, 193
Sitka mountain ash, 240, 254
skunk cabbage, 14
slender nettle, 79
slender shepherd's purse, 81
small cranberry, 163
smartweed, alpine, 155
***Smilax bona-nox,* 66**
***Smilax californica,* 66**
***Smilax herbacea,* 66**
***Smilax rotundifolia,* 65**
Smilax tamnoides, 66
smooth sumac, 185
snowberry, creeping, 153
soapweed, 134
***Solanum carolinense,* 265**
Solanum dulcamara, 265
Solanum elaeagnifolium, 265
Solanum nigrum, 265
***Sonchus arvensis,* 103**
***Sorbus americana,* 240, 254**
Sorbus aucuparia, 240, 254
Sorbus scopulina, 240, 254
Sorbus sitchensis, 240, 254
sorrel: mountain, 153; sheep, 121
sour dock, 121
southern bayberry, 174
southern wood violet, 116
southwestern coral bean, 271
sow thistle, 103
Spanish bayonet, 134
spatterdock, 213
spearmint, 144
spice bush, 219
spiny hackberry, 220
spotted touch-me-not, 119
spring beauty, 91; alpine s.b., 92; Carolina
s.b., 92; **western s.b., 92**
squashberry, 227, 253
squawbush, 186
staghorn sumac, 186, 250
star chickweed, 85
starflower, 139
star of Bethlehem, 258
***Stellaria jamesiana,* 85**
***Stellaria media,* 85**
Stellaria pubera, 85
sticky currant, 169
stinging nettle, 79
storksbill, 109
strawberry, wild, 91
subalpine prickly currant, 170
sugar bush, 185
sugar maple, 125, 257
sugar pine, 193
sumac: lemonade, 185; **smooth, 185; staghorn,
186, 250**
sunflower, 156
swamp blueberry, 164
Swedish bunchberry, 143
sweet birch, 122, 253
sweetbrier rose, 223
sweetclover, white, 267; **yellow, 265**
sweet fern, 174
sweet flag, 133
sweet gale, 176

***Taraxacum officinale,* 104**
tartary buckwheat, 155
tarweed, 8
Taxus baccata, 271
***Taxus canadensis,* 271**
Taxus cuspidata, 271
tea: Labrador, 160; mountain Labrador, 160;
New Jersey, 176; northern Labrador, 160
Texas black walnut, 244, 247
Texas mulberry, 200
Texas persimmon, 238
Texas wildrice, 129
thimbleberry, 180
thistle: bull, 115; Russian, 76; sow, 103
***Thlaspi arvense,* 82**
Thlaspi perfoliatum, 82
toothwort, 81
touch-me-not, pale, 119; spotted, 119
Toxicodendron diversiloba, 185
***Toxicodendron radicans,* 272**
Toxicodendron vernix, 185
Tragopogon dubius, 106
***Tragopogon porrifolius,* 112**
***Tragopogon pratensis,* 106**
Trientalis borealis, 139
***Tsuga canadensis,* 122**
Tsuga caroliniana, 122
Tsuga heterophylla, 122
Tsuga mertensiana, 122
tule, 70
***Typha angustifolia,* 69**
***Typha latifolia,* 69**

***Umbellularia californica,* 198**
***Urtica dioica,* 79**
Urtica gracilis, 79
Utah juniper, 197

***Vaccinium*,** 163
***Vaccinium angustifolium*,** 164
***Vaccinium corymbosum*,** 164
***Vaccinium macrocarpon*,** 163
Vaccinium myrtilloides, 163
Vaccinium oxycoccos, 163
***Vaccinium uliginosum*,** 164
***Vaccinium vitis-idaea*,** 163
***Valeriana edulis*,** 89
valerian, edible, 89
velvet-leaved blueberry, 163
***Veratrum viride*,** 265
***Veronica americana*,** 144
vetchling, 149
vetch, milk, 149
***Viburnum alnifolium*,** 224
Viburnum cassinoides, 227
***Viburnum edule*,** 227, 253
***Viburnum lentago*,** 227
Viburnum opulus, 224, 250, 253
Viburnum pauciflorum, 224, 250
Viburnum prunifolium, 227
***Viburnum trilobum*,** 224, 250
***Viola canadensis*,** 95
Viola hirsutula, 116
***Viola palmata*,** 116
Viola purpurea, 116
***Viola striata*,** 96
violet: **Canada, 95; pale, 96;** southern wood, 116; western mountain, 116; **wood, 116**
***Vitis*,** 214
***Vitis californica*,** 216
***Vitis riparia*,** 214
***Vitis vulpina*,** 214

walnut: Arizona, 244, 247; **black, 244; California, 247;** Hinds, 244, 247; Texas black, 244, 247
***wapato*,** 213
Washington hawthorn, 237
watercress, 146
water hemlock, 268
water horehound, 89
water lily, 210
wax currant, 170
wax myrtle, 174
western bistort, 155
western camass, 136
western hemlock, 122
western mountain violet, 116
western spring beauty, 92
western wintergreen, 160
white clintonia, 62
white hellebore, 265
white mulberry, 200
white oak, 228
white sweetclover, 267
wild garlic, 61
Wild Grapes, 214
wild leek, 61
wild lettuce, 104
wild onion, 58
wild plum, 203
wild potato vine, 216
wild raisin, northern, 227
wild rice, 129
wildrice, Texas, 129
wild rose, 223
wild strawberry, 91
wineberry, 180
wintercress, 99; early w., 99
wintergreen, 150; creeping w., 160; western w., 160
wood nettle, 79
wood violet, 116; southern w.v., 116
wrinkled rose, 223

yellow birch, 122, 254
yellow dock, 121
yellow flag, 261
yellow-flowered jewelweed, 119
yellow goat's beard, 106
yellow nut grass, 130
yellow pond lily, 213
yellow sweetclover, 265
yew, American, 271; English, 271; Japanese, 271
***Yucca baccata*,** 134
Yucca brevifolia, 5
***Yucca glauca*,** 134
***Zephyranthes atamasco*,** 261
***Zigadenus venenosus*,** 258
***Zizania aquatica*,** 129
Zizania texana, 129

EYEWITNESS COMPANIONS

Scuba Diving

MONTY HALLS & MIRANDA KRESTOVNIKOFF

AQUA LUNG
FIRST TO DIVE

TECHNIQUES • EQUIPMENT
MARINE LIFE • DIVE SITES

EYEWITNESS COMPANIONS

Scuba Diving

MONTY HALLS &
MIRANDA KRESTOVNIKOFF

DOLPHIN GLAS

"WE FORGET THAT THE WATER CYCLE AND THE LIFE CYCLE ARE ONE."

Jacques Cousteau

LONDON, NEW YORK,
MUNICH, MELBOURNE, DELHI

Project Editor	Richard Gilbert
Project Art Editor	Mark Cavanagh
Managing Editor	Stephanie Farrow
Managing Art Editor	Lee Griffiths
Publishing Director	Jonathan Metcalf
Art Director	Bryn Walls
DTP Designer	Vania Cunha
Production Controller	Melanie Dowland

Produced for Dorling Kindersley by
cobaltid
The Stables, Wood Farm, Deopham Road,
Attleborough, Norfolk NR17 1AJ
www.cobaltid.co.uk

EDITORS
Marek Walisiewicz,
Jamie Dickson, Maddy King,
Kati Dye, Steve Setford

ART EDITORS
Paul Reid, Lloyd Tilbury,
Claire Oldman, Darren Bland,
Rebecca Johns

Commissioned photography by Dan Burton

First American Edition, 2006
Published in the United States by DK Publishing, Inc.,
375 Hudson Street, New York, NY 10014

06 07 08 09 10   10 9 8 7 6 5 4 3 2 1

Copyright © 2006 Dorling Kindersley Limited

This companion guide is intended to help you understand the key principles that enable divers to practice their sport enjoyably and safely. It is not a training manual, and it is recommended that you read this guide in conjunction with an accredited course of diving instruction and official training literature. Diving centers around the world will ask to see evidence of formal qualifications before allowing you to rent equipment, have your tank filled, or join a diving trip. The authors and publishers have made every effort to ensure that the information in this book is correct at press time. However, they accept no responsibility for any injury, accident, or inconvenience sustained by any person using this book or following advice presented in it.

All rights reserved under International and Pan-American Copyright Conventions. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the copyright owner. Published in Great Britain by Dorling Kindersley Limited.

A Cataloging-in-Publication record for this book
is available from the Library of Congress.

ISBN-13: 978-0-7566-1949-7
ISBN-10: 0-7566-1949-1

DK books are available at special discounts for bulk purchases for sales promotions, premiums, fund-raising, or educational use. For details, contact: DK Publishing Special Markets, 375 Hudson Street, New York, NY 10014 or SpecialSales@dk.com
Color reproduction by GRB, Italy
Printed in China by Leo

Discover more at
www.dk.com

Introduction
10

Chapter One

HISTORY OF DIVING
20

The story of diving's evolution, describing the pioneering figures, scientific breakthroughs, and record descents that have helped to shape the modern sport.

Chapter Two

GETTING STARTED
30

How and where to learn diving, including information on training agencies and clubs, advanced and specialized training, and what to expect from instruction.

CONTENTS

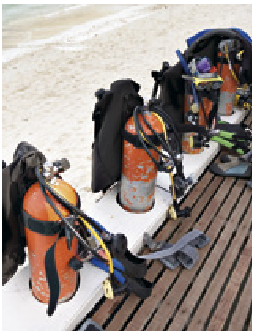

Chapter Three

EQUIPMENT
40

Information on essential items of diving gear, including what you need, how things work, and how to assemble and maintain your scuba equipment.

Chapter Four

THE DIVING ENVIRONMENT
90

The physical principles that underpin the sport of diving, including perception underwater, decompression theory, and how currents and tides work.

Chapter Five

DIVING SKILLS
104

Key diving techniques, procedures, and safety measures, featuring step-by-step instructions on how to master fundamental skills, and advice on fitness for divers.

Chapter Six

GOING DIVING
156

How to research, plan, and conduct independent dives, plus information and advice on popular diving activities and specialty pursuits.

Chapter Seven

UNDERWATER LIFE
198

A natural history of the world's marine and freshwater zones, with detailed information on key environments, animals, and natural processes.

Chapter Eight

PLACES TO DIVE
234

An illustrated directory of the world's best dive destinations, including famous wrecks, marine life hot spots, and adventure dives.

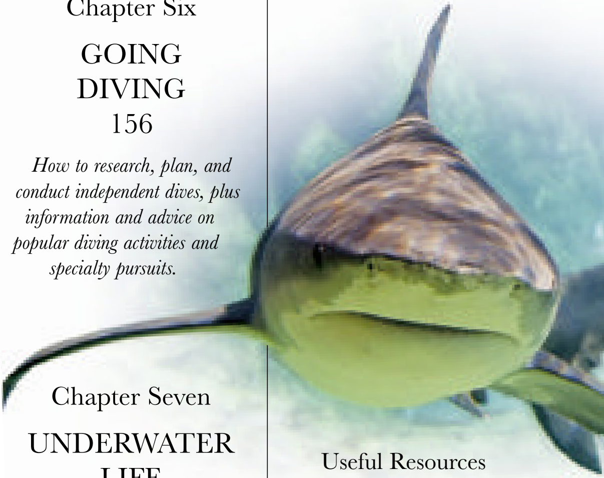

Useful Resources
328

Glossary
331

Index and Acknowledgments
336

THE IDEA THAT PEOPLE COULD BREATHE UNDERWATER WAS ONCE A FANTASY. BUT TODAY, DIVING HAS LEFT THE REALMS OF IMAGINATION AND GROWN INTO A WORLDWIDE SPORT WITH MILLIONS OF FOLLOWERS. THOSE WHO TAKE PART IN IT ARE GRANTED PRIVILEGED ACCESS TO A MYSTERIOUS YET BEAUTIFUL WORLD, AND EVERY DIVE IS AN ADVENTURE INTO THE UNKNOWN.

THE DAWN OF AN ERA
Diving pioneer Jacques Cousteau (*right*) models early scuba equipment.

Since prehistoric times, people have been drawn to explore the incredible world that lies underwater, whether in pursuit of food or lost riches, or from sheer inquisitiveness. There is evidence, for example, that people were holding their breath and diving for shellfish by about 3,000 BCE (*see p.22*).

Despite this long relationship with the undersea world, the human body is not designed to remain underwater for prolonged periods of time. Only comparatively recently, with the development of rudimentary diving bells and helmets from the Middle Ages onward, did it become possible to stay submerged for longer than a single breath allows. During the 18th and 19th centuries, there were remarkable advances in diving technology, such as full diving suits supplied by air pumped from the surface. These advances, which were often spurred on by the lucrative rewards of salvage work, extended dive times and gave access to ever greater depths.

The invention of the Aqua Lung in the mid-20th century permitted divers to explore the depths of the world's oceans freely for the first time. This truly remarkable apparatus enables people to take their own air supply with them when they dive, eliminating the need to be connected to the surface by an airline. The system that evolved from it—scuba—is today widely used in recreational and professional diving.

GLOBAL APPEAL

As scuba technology became cheaper, diving opened up to the general public. What was once the realm of experts and professionals gradually became a sport with global appeal. Even in countries

WRECK DIVING
Divers inspect a wreck in tropical waters. Such underwater relics offer a glimpse into maritime history, and are home to colonies of sea life.

A NEW GENERATION
Trainee instructors receive a group briefing at a dive center in Egypt. For many, a job in diving represents the ultimate career move.

without a coastline, you will find divers who spend time underwater at inland sites, such as lakes and rivers, or who travel abroad to experience the beauty of the oceans. Once you are qualified, you can dive virtually anywhere in the world. Language is no barrier underwater, since divers use standard hand signals that are understood internationally.

It is difficult to estimate how many divers there are worldwide, but well over a million people qualify each year. Since its inception in 1966, over 13 million divers around the world have been certified by PADI (the Professional Association of Diving Instructors), and PADI is just one of many training agencies. Diving is one of the fastest-growing hobbies in the world today.

SILENT ODYSSEY
Clown fish shelter in an anemone while a diver swims by. The underwater world can be a uniquely serene environment.

THE WORLD'S OCEANS

Covering approximately 71 percent of the Earth's surface, or 140 million sq miles (361 million sq km), the vast size of the world's oceans is difficult to comprehend. The total volume of sea water is about 322,300,000 cu miles (1,347,000,000 cu km), equivalent to around 97 percent of all the Earth's water—the rest exists as glaciers, rivers, lakes, groundwater, and atmospheric vapor. The average depth of the world's oceans is 2.3 miles (3.8 km); the deepest point is in the Pacific Ocean's Mariana Trench, where the seabed plunges to a depth of nearly 7 miles (11 km). With only about 5 percent of the ocean floor having been mapped in detail, we know more about the surface of the Moon than the bottom of our oceans.

The oceans teem with wildlife, from giant kelp and the mighty blue whale (the largest animal alive) to microscopic plankton that form the basis of the oceanic food chain. Scientists estimate that there may be between 500,000 and 10 million marine species yet to be discovered. While much of the oceans' biological wealth is concentrated between continental shelves and coastal margins, the open sea also abounds with life. Even the mud on the floor of the Mariana Trench is home to single-celled organisms called foraminiferans.

NATURAL WONDERS
Diving can lead to encounters with many intriguing and beautiful marine creatures, such as this leafy sea dragon.

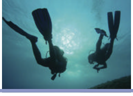

GETTING READY
Buddy pairs prepare for a beach dive. By working in pairs, donning awkward gear becomes considerably easier and safer.

EXPLORING THE UNDERWATER WORLD

What will I see? What will it feel like? Will I feel claustrophobic? These are all questions you will ask as a novice diver entering the underwater world for the first time. Everyone feels different about being underwater; some people take to it immediately, but others need more training and experience before they feel at ease. Either way, it is an unforgettable experience. There is the initial thrill of being able to breathe underwater, and also the novelty of moving freely in three dimensions, thanks to the buoyancy provided by the water, which is a denser medium than air. This same buoyancy can enable those with mobility problems on land to feel relaxed and move more easily in water, and give what some divers call a “womblike” experience.

Imagine the thrill of being the first person to see a sunken ship or coral reef, or of coming face to face with a wild dolphin, or diving under ice. Whatever excites you—whether it be encounters with fascinating creatures, investigating shipwrecks, or just diving into the unknown—you will not fail to be satisfied by what the oceans have to offer.

CAPTURING THE MOMENT
Underwater photography is an absorbing activity in its own right, and offers the chance to record your dive experiences for posterity.

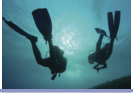

CONSERVATION WORK
A diver jots down his observations of marine life at a conservation site. Divers can play an active role as "stewards of the seas."

BUT IS IT SAFE?

Many people new to the sport worry about the safety aspect. It is true that a small number of divers die every year while pursuing their sport. The majority of these incidents are not due to equipment failure, as some people might suppose, but rather to inadequate training, diving in extreme conditions, or diving beyond the limits of safe practice.

Diving today is safer than it has ever been. If you undertake the correct training and adhere to recommended safety procedures, you will rarely encounter problems. Dive training is thorough, and a great emphasis is placed on rescue skills and the importance of diving in a buddy pair (*see pp.110–11*). Good communication with your buddy is essential, as most minor difficulties can be easily resolved with the help of another diver.

NEW SKILLS, NEW DIRECTIONS

You may find that you enjoy diving so much that you want to pursue it beyond the hobby level. Some people use their underwater knowledge and skills for voluntary or charitable work, or in a full-time career. The directions in which diving can take you are endless, from environmental projects, marine archaeology, historical research, and working with disabled divers, to filming and photography, and even training to be a diving instructor yourself.

RECOGNITION GUIDE
A pair of divers consult a species identification slate. Becoming familiar with marine life, in all its fascinating variety, is part of the fun.

DISCUSSING A DIVE
Diving is a social sport, and the diving community is a supportive one. Experienced divers are often willing to offer advice to novices.

RESPECT AND RESPONSIBILITY

Ask any group of divers what inspired them to take up the sport and you will hear a range of different answers, from a love of the sea to a desire to challenge their personal limits. One thing they all have in common is the fact that they return again and again to the water—the magical environment that Jacques Cousteau (*see p.26*) described as the "Silent World." Divers should be custodians of the world's oceans, and respect what they contain: "take only pictures and leave only bubbles" is a good maxim for low-impact diving.

Many of the marine ecosystems around the world are today in decline. Hopefully education and investment will help to reverse this trend, so the divers of tomorrow will have healthy oceans to explore. By taking part in conservation work and diving responsibly, you can become a force for the good in preserving the oceans' natural treasures.

MARINE LIFE
A diver examines a beautiful and delicate sea pen—a colonial polyp organism—growing out of white sand next to the Great Barrier Reef.

GLIMPSE OF A GIANT
Anticipation of a dive can be almost as thrilling as the dive itself. Sighting large animals, such as rays, from the boat adds to the excitement.

UNDERSEA GARDEN
Coral thrives in clear, shallow water, because it needs plenty of light to live. Such reefs are an ideal environment to explore while snorkeling.

REEF SAFARI
Coral reefs are home to diverse marine species, offering sheltered areas for sedentary animals, and feeding grounds for predators and foragers.

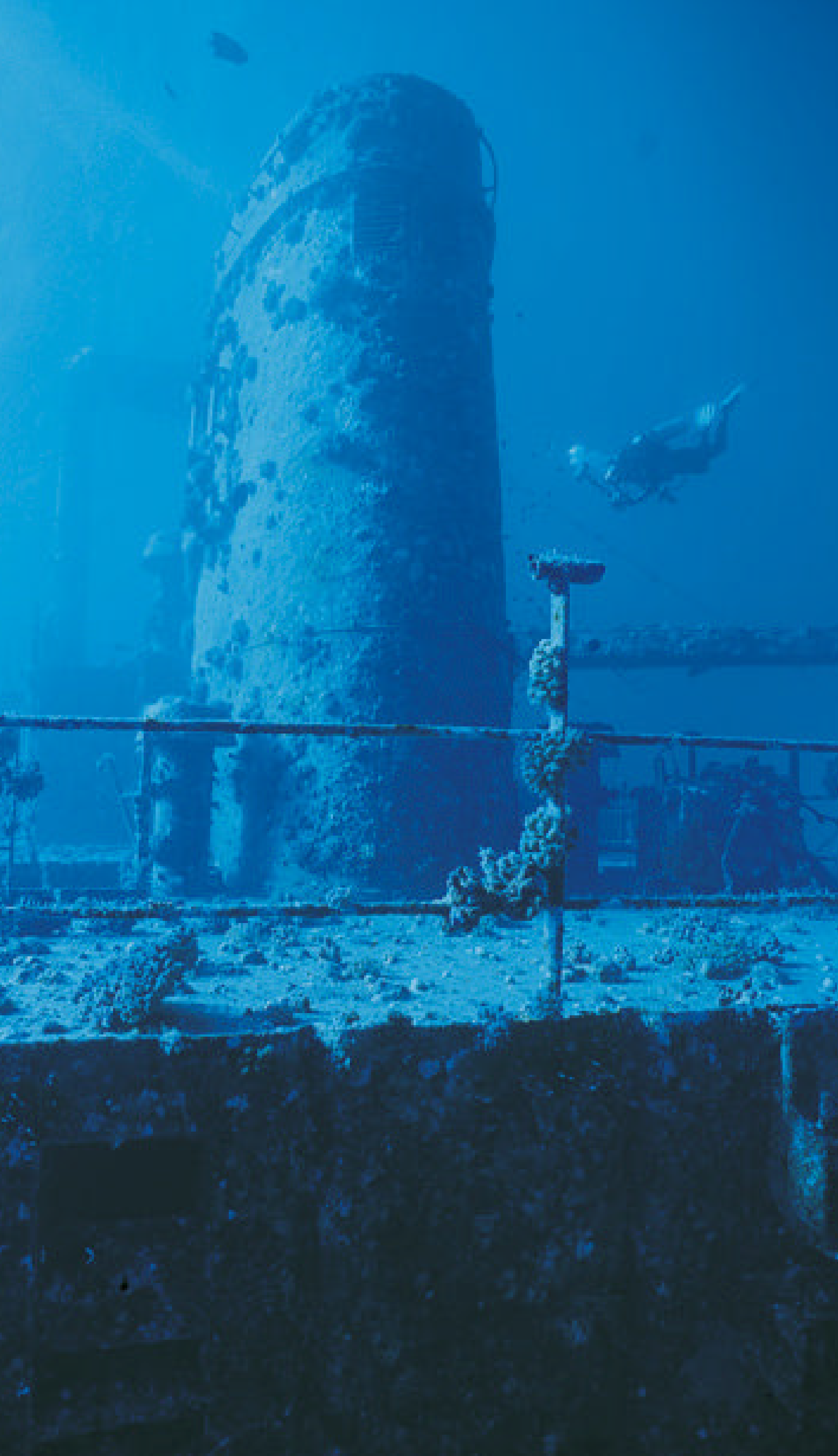

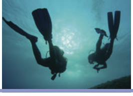

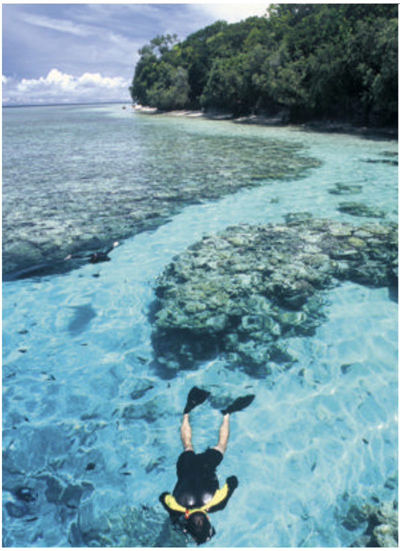

CRYSTAL WATERS
Exploring the shallows above reefs is a perfect pastime between dives. Many thousands of fascinating marine species live in these areas.

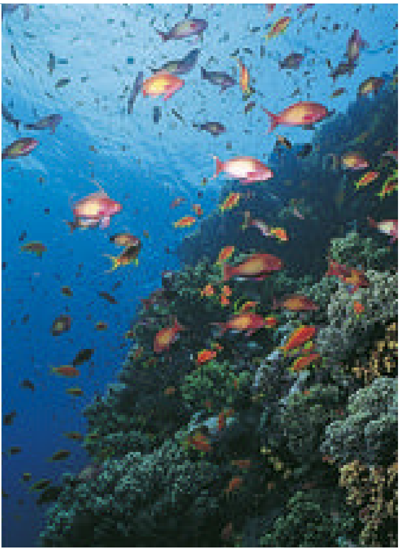

CLOUDS OF FISH
Shoals of orange antheas are a common sight on the reefs of the Indo-Pacific. They appear to move as one when startled.

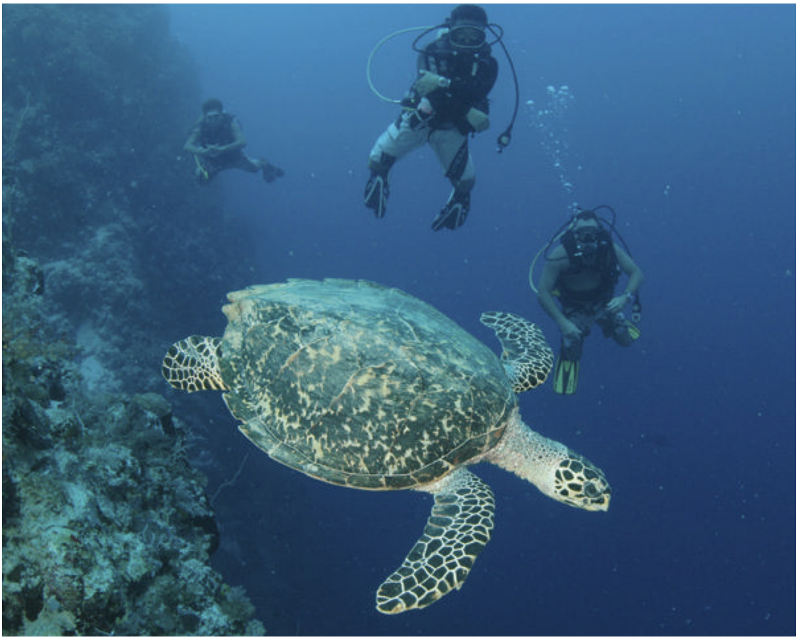

FROM WRECK TO REEF
In tropical waters, a wreck provides a firm platform for coral growth. Wrecks are swiftly colonized by marine life, creating eerie reefs.

LONE VOYAGER
If approached cautiously and respectfully, turtles will happily tolerate the presence of divers. Many species migrate for thousands of miles.

History of diving

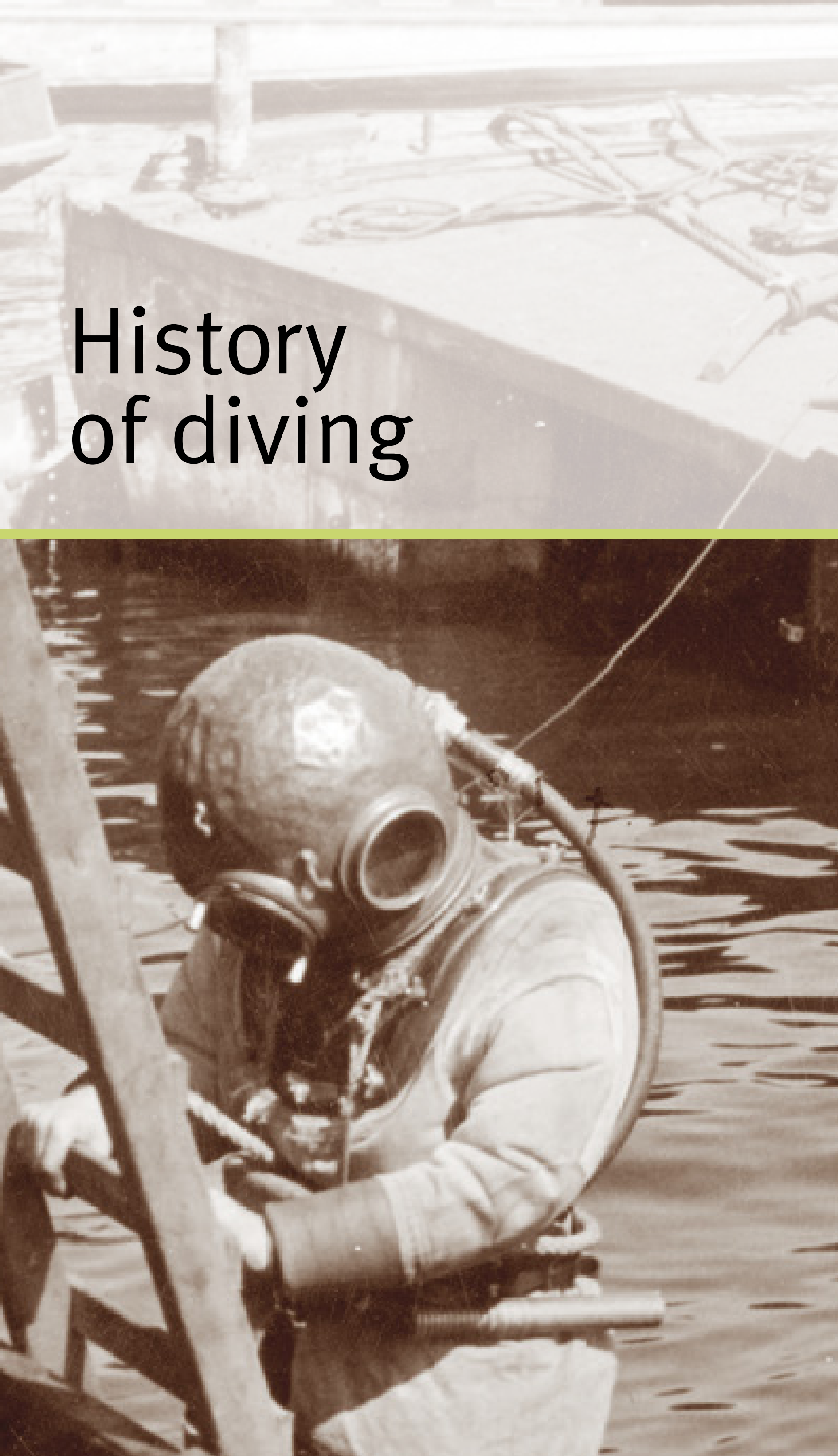

Diving through the ages

The search to free divers from the limitations of breath-holding began in ancient times with the use of reed snorkels. Developing from diving bells and diving suits supplied with air from the surface, it evolved into modern scuba, in which divers carry their own air supply.

EARLY DIVERS AND DIVING BELLS

It is hard to trace the origins of diving, but mummified human remains dating back more than 5,000 years have been found in Chile alongside fish bones and mussel shells. Small growths in the humans' ears, resembling the modern condition called surfer's ear, suggest that these people may have been divers.

Hollow reeds or breathing tubes that extended to the surface were used in ancient times to enable divers to stay submerged for longer. Herodotus recorded that in 500 BCE, a Greek diver named Scyllias used a reed snorkel to swim unnoticed and cut the moorings of Persian ships. Salvage diving was well known in the ancient world, and laws regulating those who dived for treasure were in operation in Greece as early as the 1st century BCE.

An important milestone in helping divers to overcome the restrictions of breath-holding was the invention of an apparatus to hold air underwater. At the siege of Tyre in 332 BCE, Alexander the Great reportedly used such a device, descending beneath the waves in a glass barrel. In the 4th century BCE, Aristotle described a bell-shaped diving chamber that took the form of an overturned cauldron filled with air, in which divers made trips to retrieve objects from the seabed. From the Middle Ages onward, diving equipment became more sophisticated. During the Renaissance, speculative designs for diving suits began to appear, although nearly all

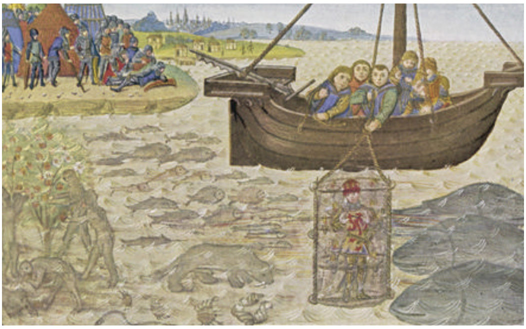

ALEXANDER'S GLASS BARREL Surprisingly, Alexander the Great used his glass barrel to make observations of fish, rather than as an aid to his military campaigns.

KEY DATES

3000 BCE

c. 3000 BCE Early divers collect oysters from the sea bed, developing a bone deformity caused by frequent exposure to cold water—a condition today known as "surfer's ear."

c. 2800 BCE A Sumerian epic detailing the exploits of King Gilgamesh reports that he used weights to sink to the seabed in search of a plant granting immortality.

1000 BCE

9th/8th century BCE Homer's *Iliad* refers to the use of military divers in the Trojan War, *c.* 1250 BCE.

c. 500 BCE Herodotus recounts the story of Scyllias diving to cut ships' anchor lines on the seabed.

c.4th century BCE Aristotle records the first use of a diving bell, made from an overturned cauldron filled with air. The diver within could venture out for short periods before returning for air.

c. 332 BCE Alexander the Great descends to the seabed inside a glass barrel during the siege of Tyre.

0

would have proved ineffective if put to use. Nonetheless, Spaniard Diego Ufano described the use of leather diving helmets during salvage work in his *Treatise on Artillery*, published in 1612.

Diving bell technology also improved during this period. In 1691, English astronomer Edmund Halley produced an advanced bell that was linked by a hose to weighted barrels of air, which were periodically replenished at the surface. Stale air was vented from the top of the bell. Halley and his bells were widely employed for commercial salvage work.

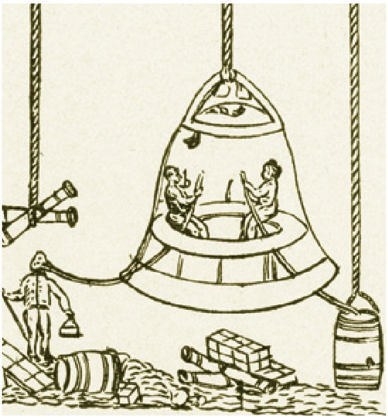

HALLEY'S BELL, 1691
A diver could swim out from under Halley's bell with an air line linked to a smaller bell over his head—a precursor of the modern diving helmet.

DIVING FOR PEARLS
The earliest divers relied only on breath-holding. Using this technique, female divers in Japan have been harvesting pearls for 4,000 years.

1500 Leonardo da Vinci designs equipment for underwater exploration, including snorkel tubes and bags intended to supply a diver with air.

1660 Robert Boyle identifies the relationship between water pressure and gas volumes, furthering the understanding of buoyancy.

1691 Edmund Halley develops a diving bell with a glass top to admit light, and an air replenishment system using barrels lowered from the surface. An umbilical tether supplies air to the bell's operator when venturing outside the bell.

1715 Inventor John Lethbridge uses a modified barrel as a diving vessel to recover salvage from the sea floor. The barrel has a glass viewing port and holes with watertight leather sleeves, through which the operator's arms protrude. The barrel reaches depths of 60 ft (18 m) in trials.

1500 1600 1700 1800

DEVELOPMENT OF THE DIVING SUIT

In 1715, Englishman John Lethbridge devised a pressure-proof wooden diving barrel. Reinforced and clad in leather, the barrel had a glass viewing porthole and arm holes with watertight leather sleeves. Lowered from a ship to depths of 66 ft (20 m), the barrel enabled the diver inside to use his arms freely, pulling his way around a site and retrieving objects. Lethbridge was commissioned by the Dutch East India Company to recover treasure from wrecks around the world using these diving barrels.

The barrel was a forerunner of the armored diving suit. Its usefulness, however, was limited by the inability to replenish the air in the barrel without hauling it back to the surface. Three other developments would be required before diving suits became a reality.

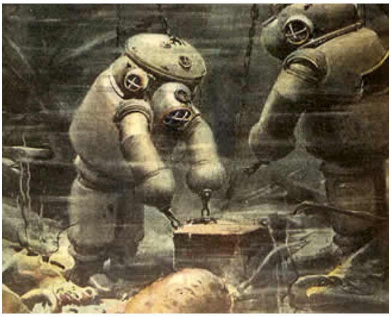

SALVAGE IN ARMORED DIVING SUITS, 1932
Lethbridge's barrel was a precursor of armored diving suits. Initially too clumsy, they now have a niche for specialized work at great depth.

The first of these occurred in 1779, when John Smeaton, another Englishman, used a steam-powered pump to supply air to a diving bell. Air pumps allowed diving bells to stay down considerably longer than before, and would later supply the first diving suits. The next development, by the Scot Charles Mackintosh in 1818, was the production of a rubber-based coating for waterproofing canvas, providing an ideal material for a diving suit. The third innovation, in 1820, was the patenting by British salvage operators John and Charles Dean of a diving helmet—a modified firefighting helmet into which air could be pumped. The helmet rested on the

CAISSON DISEASE

Improved air-pump designs in the mid-19th century made it possible to build underwater chambers in which several men could work at depth. Called caissons, these chambers were ideal for tasks such as building bridge footings and excavating tunnels. Air-pumps delivered sufficient pressure to keep water entirely out of the chamber. However, workers often felt ill at the end of a shift. This was decompression sickness (DCS), or "the bends."

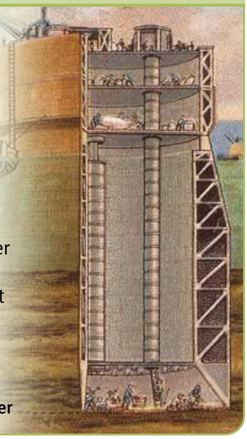

An illustration of a caisson chamber

KEY DATES

1779 Englishman John Smeaton pioneers the use of a pump system to replenish submerged diving bells with fresh air.

1788 German diver Peter Kreeft successfully demonstrates the use of a leather diving helmet in the Baltic sea, witnessed by the king of Sweden.

1800

1818 Naphtha, a coal-tar derivative, is discovered. When mixed with melted rubber, it forms a waterproof coating, and is soon used in the manufacture of lightweight diving suits.

1825

1829 Brothers Charles and John Dean undertake wreck salvage work in Britain using a modified firefighting helmet to supply air during dives to the seabed.

1839 British Army divers develop the practice of diving in pairs during salvage work on the warship *Royal George*.

diver's shoulders, held in place by its own weight, and air passed out from under the edge of the helmet. However, if the diver stumbled, it could easily fill with water.

In the 1840s, German-born inventor Augustus Siebe perfected a "closed dress" design, with the helmet sealed to a waterproof canvas diving suit. An exhaust valve was incorporated into the helmet, which was supplied with air by a pump at the surface. With a few modifications—such as neck and foot weights to help divers stay upright, and thick woolens under the suit to combat cold—Siebe's closed dress design remained standard for deep-sea divers for more than a century.

DECOMPRESSION SICKNESS

In the mid-19th century, the consequences of breathing highly compressed air were being seen among divers and caisson workers, who suffered what we now know as decompression sickness (DCS). In 1878, French physiologist Paul Bert realized that these ill-effects were caused by nitrogen damaging the body during rapid decompression, and recommended that caisson workers and divers return to normal atmospheric conditions slowly.

Nearly 30 years later, Scottish physiologist J. S. Haldane devised the first decompression tables after researching the effects of gas poisoning and DCS. Haldane pushed diving to new limits, sending divers down to 210 ft (65 m) in 1906. The divers avoided DCS by performing decompression stops.

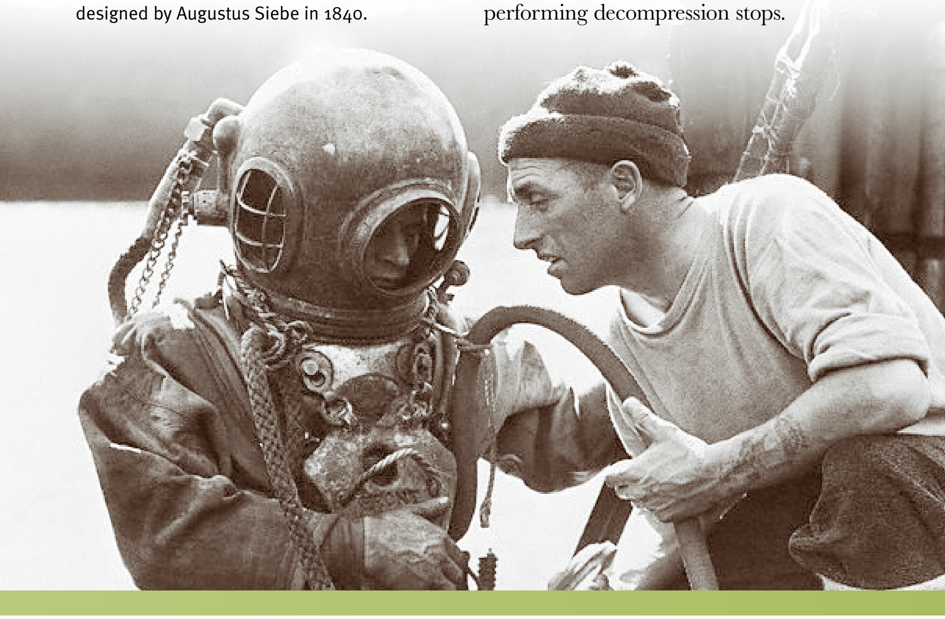

DEEP-SEA DIVING GEAR
The helmet and suit worn by this diver in 1954 are remarkably similar to those designed by Augustus Siebe in 1840.

1869 Author Jules Verne popularizes diving with his novel *20,000 Leagues Under the Sea*, in which characters explore the seabed with compressed air breathing apparatus.

1878 French scientist Paul Bert publishes a paper explaining the causes of "Caisson Disease" (later known as Decompression Sickness), a condition associated with breathing pressurized air for long periods of time.

1850 **1875** **1900**

1865 French inventors Benoit Rouquayrol and August Denayrouze invent compressed-air breathing apparatus that allows short periods of autonomy from surface air supply.

1878 English merchant seaman Henry Fleuss designs the first fully autonomous breathing apparatus. The device uses compressed oxygen as a breathing gas.

1893 The first underwater camera is designed by Frenchman Louis Boutan.

SCUBA DIVING IN THE 1960s
As other manufacturers introduced their own versions of Cousteau and Gagnan's scuba apparatus, the cost of equipment fell, and the 1960s became a boom-time for scuba.

AUTONOMOUS DIVING

In the 19th century, inventors strove to devise self-contained breathing systems that would allow autonomous diving—diving without an air line to the surface. In 1865, Frenchmen Benoit Rouquayrol and Auguste Denayrouze devised a compressed-air tank with a primitive regulator, which adjusted the air flow to meet the diver's breathing requirements. In this "open-circuit" apparatus, air was inhaled from the tank and waste gases vented into the water. Although the diver was tethered to the surface by a hose that pumped air into the tank, he could disconnect the hose for a few minutes, and breathe from the tank alone.

In 1878, Englishman Henry Fleuss produced the first fully autonomous system, using a breathing bag fed by a copper tank of pressurized oxygen. This was a "closed-circuit" system, or "rebreather," in which the exhaled air is recirculated. Carbon dioxide was removed from the exhaled breath and oxygen injected to replace that consumed by the diver, allowing it to be breathed again.

In 1911, the German firm Draeger began producing oxygen rebreathers, soon developing these to use enriched air (compressed air with added oxygen). Rebreathers were used by naval divers during both world wars, in preference to open-circuit systems, which emitted a telltale stream of bubbles.

WWII NAVAL DIVER'S MASK
Air from a rebreather was supplied via a hose to a rubber hood.

SCUBA GEAR

The type of equipment used by divers today first began to take shape in the 1930s and 40s. American Guy Gilpatric created the first modern mask in the early 1930s, and the first diver's snorkel was

KEY DATES

1908 John Scott Haldane, Arthur C. Boycott, and Guybon Damant publish a landmark paper on "compressed air sickness." The first decompression tables are based on this research.

1917 The liner *Laurentic* sinks off the Irish coast after hitting a mine. Its cargo includes 35 tons of gold bullion. Royal Navy salvage divers succeed in recovering 3,189 ingots from the wreck over a period of seven years.

1910

1920

1917 The Draeger company of Germany manufactures an enriched air rebreather (a respirator that recycles exhaled air) that can be operated to a depth of 130 ft (40 m).

1925 French naval officer Yves Le Prieur develops an open-circuit (non-recycling) compressed-air respirator—a forerunner of modern scuba.

JACQUES COUSTEAU
From the 1950s to the 1970s, Cousteau's groundbreaking films and TV programs revealed the wonders of the deep to a worldwide audience, and persuaded many to start diving.

developed by Englishman Steve Butler at about the same time. Fins were patented in France by Louis de Corlieu in 1933.

Open-circuit scuba (Self-Contained Underwater Breathing Apparatus) as we know it today was pioneered by two Frenchmen, engineer Emile Gagnan and naval officer Jacques Cousteau. In 1943, they combined a two-stage regulator, from which the diver breathed, with a high-pressure air tank. The regulator supplied air to the diver on the slightest intake of breath. Other self-contained breathing apparatus used at the time either supplied air continuously or had to be manually turned on and off. (The primitive regulator design of Rouquayrol and Denayrouze had fallen out of use.) Efficient, reliable, and comparatively cheap to produce, Cousteau and Gagnan's "Aqua Lung," as it was known, went on sale in France in 1946. By the early 1950s, Aqua Lungs were available throughout the world.

The first commercially successful buoyancy compensator, which allowed divers to adjust their buoyancy in the water, was patented in 1961 by Frenchman Maurice Fenzy.

1930 American William Beebe descends to a depth of 1,426 ft (435 m) in a bathysphere—a tethered diving vessel launched from a ship.

1933 French inventor Louis de Corlieu patents rubber fins to aid underwater swimming.

1938 American Guy Gilpatric's book *The Complete Goggler* popularizes spearfishing underwater.

1941 Italian navy divers use rebreathers to mount underwater attacks on moored Allied shipping during World War II.

1942 Jacques Cousteau and Emile Gagnan design and test the Aqua Lung compressed-air breathing apparatus. The design's ease of use helps to encourage recreational diving.

1948 Aqua Lung first exported to the US.

1930 1940 1950

DIVING TODAY

The Aqua Lung caused a revolution in diving, making it accessible to the public for the first time. In 1950, there were some 30,000 scuba divers globally; ten years later, there were in excess of one million. There are now estimated to be tens of millions of divers worldwide.

Modern diving equipment seems a world away from that used by Jacques Cousteau, yet the basics remain the same. Like Cousteau, we dive with a tank of compressed air, and use a mask, snorkel, and fins. Today, however, there is also a host of other equipment and gadgets to make diving even safer and more enjoyable. Mixed-gas diving (*see pp.194–95*), which allows deeper and longer dives, is becoming increasingly commonplace, and there are computers that give us data such as depth, water temperature, dive times, and breathing rate. Some dive computers even allow divers to play games during lengthy

DIVING FOR PLEASURE

Two divers prepare to explore beneath the surface of a tropical lagoon. The rebreather units on their backs allow very long dive times.

KEY DATES

1950

1951 *Skin Diver* magazine goes on sale in the US, and plays a central role in popularizing diving.

1953 Dr. Hugh Bradner develops the first neoprene wetsuit for the US Navy.

1956 Jacques Cousteau's diving film *The Silent World* (based on his book of the same name) wins an Academy Award.

1959 The first nationally organized scuba training program is established in the United States by the YMCA.

1960

1961 Frenchman Maurice Fenzy patents an inflatable safety jacket for divers, which becomes the first commercially available buoyancy compensator (BC).

1967 American Bob Croft is the first man to free dive (breath-hold dive) below 200 ft (60 m).

decompression stops. Scooters eliminate the need to use our legs for propulsion, and waterproof MP3 players enable us to listen to music underwater. It may not be long before divers can take cell phones underwater and send text messages and photographs from the bottom of the sea.

HUMAN TORPEDO
A diver clings to a battery-powered propulsion unit. These devices enable divers to explore large areas in a single dive.

THE FUTURE OF DIVING

Many divers believe that the future of diving lies with using rebreathers (*see pp.82–83*). Once used only commercially, these devices are now available to leisure divers and are becoming increasingly commonplace. Rebreathers greatly extend the time divers can spend at the bottom, and reduce the need for multiple tanks of breathing gas to be taken on a dive. Fifty years from now, it is possible that we may all be diving with sophisticated rebreathers.

There is also increasing interest in submersible vehicles for the private market. Billed as the ultimate rich person's toy, such craft may open up extended explorations of the seabed to their owners. Diving limits are also being extended every year. In the sport of free diving, where competitors use no scuba gear but simply hold their breath (*see pp.190–91*), records of more than 500 ft (150 m) have now been set. Scuba divers have reached similar depths breathing compressed air, and have descended to below 1,100 ft (330 m) using mixed gases. Will we one day be able to dive to the bottom of the Pacific Ocean—6.8 miles (11 km) underwater? Only time will tell, but it would be surprising if advances in technology and our understanding of the human body did not grant some extension of today's limits.

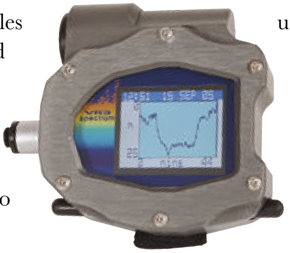

DIGITAL DIVE PLANNING
Sophisticated computers record and process vital dive information.

1968 Americans Neal Watson and John Gruener dive to 437 ft (133 m) breathing compressed air.

1970

1971 Scubapro markets a "stabilization jacket" for divers. Its waistcoat-style design sets the trend for modern BCs.

1979 Sylvia Earl walks untethered on the sea floor at a depth of 1,250 ft (381 m) in a special pressure suit.

1980

1980 Divers Alert Network, a nonprofit organization for promoting dive safety, is founded.

1983 The first commercially available dive computer, the Orca Edge, goes on sale.

1990

1993 Italian free diver Umberto Pelizzari sets a depth record of 403 ft (123 m).

1999 Briton Mark Andrews dives to 500 ft (152 m) on air.

2000

2003 Briton Mark Ellyat dives to 1,024 ft (313 m) on Trimix scuba.

Getting started

Introduction

Learning to dive is an adventure in itself. There are many different training agencies offering basic scuba qualifications, with branches in nearly all parts of the world. Though dive tuition makes demands on both your body and mind, it can be a very rewarding experience.

LEARNING TO DIVE

Diving is a safe sport provided that you have undergone formal training before taking to the water. It would be extremely foolhardy to proceed on your own, as there is no substitute for a proven program of instruction. Moreover, no dive center or charter boat will take you diving without recognized qualifications. Considerations of safety apart, you will feel more confident, relaxed, and prepared for the underwater experience if you are fully trained. Basic "open water" training qualifies you to go diving without supervision. It will not, however, teach you to dive under ice, in a cave, or beyond certain depth limits. For these activities, you need to attain further qualifications and gain more experience.

WHO CAN BE A DIVER?

Practically anyone can learn to dive, but there are a few basic health requirements. For instance, if you suffer from asthma, you will need to consult your doctor beforehand, although the condition need not be a barrier. Diving is not a strenuous sport, but you will encounter situations that require you to be reasonably fit:

EXPERT GUIDANCE
During basic training, the instructor will tutor students in how scuba equipment works, as well as its use and assembly.

lifting yourself out of the water onto a boat, for example. Diving when pregnant is not advisable, but it is possible to pick up the theory and basic snorkeling skills during pregnancy, and continue with full training after your child is born.

There are special junior courses for those under 14 who wish to learn to dive, and advancing years are not necessarily a barrier to learning either, given a fair level of fitness and approval from a doctor. Many disabled people also enjoy the sport of diving, and dive training programs have been adapted to embrace a wide range of disabilities.

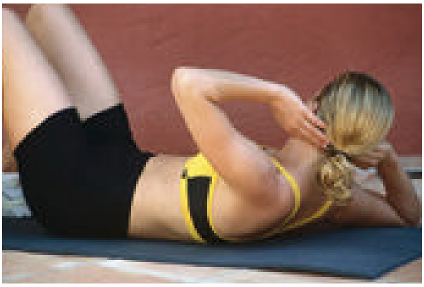

FITNESS FOR DIVING
While it is not necessary to be an athlete to go diving, a basic level of fitness is required (*see p.150*) for both safety and enjoyment.

HOW MUCH WILL IT COST?

Taking up a new hobby always involves a degree of investment, and diving is no exception. New diving equipment can be expensive, but second-hand gear is widely available, though it should be serviced before use. Equipment is almost always available to rent from dive centers and shops. and it is advisable to try out a few different pieces of gear before investing in your own.

Dive training is fairly expensive, but learning with a club can make it cheaper, and most training organizations offer trial lessons to help you decide if you enjoy diving before paying for a full course of instruction.

NOVICE IN SHALLOW WATER
An instructor guides a student through his first dive, maintaining contact the whole time to ensure a safe learning experience.

Where to learn

There are a number of different diving organizations around the world with whom you can train. Other divers may recommend the one they learned with—which is not a bad start—but investigate all the training agencies, since each offers a different approach to tuition.

TAKING THE PLUNGE

The first question you should consider is whether to learn to dive with a school or with a club. A diving school (confusingly, also called a diving club in mainland Europe) is a commercial venture—you pay to learn to dive. A diving club is a nonprofit venture where members donate their time training other members. You might think that a school sounds like a more professional option, but in truth, instructors in clubs are equally highly trained, and the fact that they are giving their time free of charge says much about their love for diving and desire to pass on their knowledge to others. Clubs also add a social dimension to learning, arrange talks and trips both at home and abroad, and may offer a longer-term framework for gaining diving experience. However, clubs tend to run training sessions on just one day of the week, and it can take a few months to become qualified.

By contrast, diving schools offer greater flexibility, and offer you the chance to learn the basics at home,

LEARNING THE BASICS
Be aware that courses abroad may have an instructor whose first language is not yours.

possibly over a couple of weekends, and then complete the rest of your course in warmer waters on vacation. Alternatively, they allow you to complete your training beforehand (and on your own schedule), then depart for your trip abroad as a fully qualified diver.

TRAINING ORGANIZATIONS

The world's largest training organization is PADI (Professional Association of Diving Instructors). They operate all over the world, and most dive sites you visit around the globe will be affiliated to PADI. Several other training organizations are prominent in the US, including NAUI (National Association of Underwater Instructors) and SSI (Scuba Schools International).

In the UK, most diving clubs are affiliated to BSAC (British Sub-Aqua Club), and their training is tailored to the waters around the UK. You may also come across the SAA (Sub-Aqua Association) and the SSAC (Scottish Sub-Aqua Club); the CFT/IUC (Irish Underwater Council) in Ireland; and CMAS (Confédération Mondiale des Activités Subaquatiques) in Europe. There are other organizations for more advanced diving (*see pp.329–30*).

CLUB BOAT WITH DIVERS
Diving clubs usually pool resources to arrange group trips that can cement long-lasting and rewarding dive buddy relationships.

THE CHOICE IS YOURS

For the beginner, the various training agencies offer broadly similar courses in terms of content and standards. The best course of action is to take advice from experienced divers and do what suits you. Always ask to see a trainer's qualifications before starting a course, and try to get a recommendation for a diving school or club before enrolling.

RESORT INSTRUCTOR WITH TRAINEES
Many people choose to learn diving on vacation, and there are thousands of diving schools around the world that offer training at all levels of ability.

Basic training

Learning to dive involves gaining some knowledge of essential theory as well as training in practical skills. Sometimes these skills are practiced in a pool before getting into open water but, depending on where you learn, they can be taught in a lagoon or in the shallows.

MASTERING THE THEORY

Your dive training will always start with theory. You will be taught about basic diving equipment, the principles involved in breathing gas at depth and under pressure, safety skills, communication, navigation, and much, much more. You will learn to carry out simple calculations relating to the physics of diving. Don't worry—none of the theory is difficult, but it does demand a few hours of study.

Dive agencies run this part of the training in different ways—some rely on traditional classroom teaching, others present materials on DVD or online—so choose the way that suits you best. In all cases, you will be provided with a reference manual containing all the material you need to know.

GOING DEEPER INTO DIVING
Learning to dive involves understanding some of the theory that underpins the sport. This training is given either in classes or one-to-one tutorials.

GETTING WET

Basic practical skills are taught in "confined water"—usually a swimming pool or shallow water just off a beach. You will be expected to pass a simple swimming test to show your instructor

TAKING THE PLUNGE
A trainee learns the basics in safe, shallow waters, under the guidance of an instructor.

that you are safe and comfortable in the water before they begin to teach the fundamental skills (*see below*). After this initial stage, you will progress into open water, where you will undertake a few introductory dives in order to get used to diving in deeper water and build your confidence to swim in the open sea.

SUITING UP
Trainees are not expected to buy a full set of equipment. Gear can be rented from most diving schools.

STAYING SAFE

Recreational diving is an extremely safe activity as long as you receive the correct training and keep your skills up to date. Choosing the right training agency, course, and instructor is therefore vital. Most instructors are highly trained and dedicated to their job, but there will always be a few "bad apples" in any profession, so take advice from friends, diving clubs, and specialty magazines and websites before you enroll. Do not be afraid to ask questions during your training, simple as they may seem; you can even request a different instructor if you do not feel comfortable.

Many people take up diving as a vacation activity and only get a chance to practice once or twice a year. If you have not dived for a while, it is a good idea to take a refresher course, and carry out a trial dive (with your buddy) to check all your gear and your skills before venturing into deeper water.

BUILDING CONFIDENCE

Fundamental skills and safety procedures are taught in the security of confined water before you can venture into the open sea. The abilities you need to master include buoyancy control and finning techniques, out-of-air procedures, removing, recovering, and replacing your regulator, and clearing your ears and your mask underwater (*right*)—a common source of apprehension among novice divers.

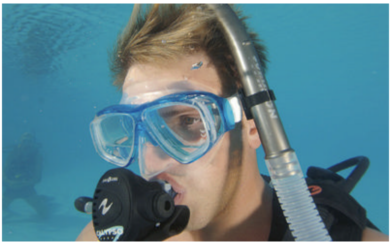

1 **Water can accumulate** in your mask during diving, usually in small amounts. With practice, even a fully flooded mask can be cleared without difficulty.

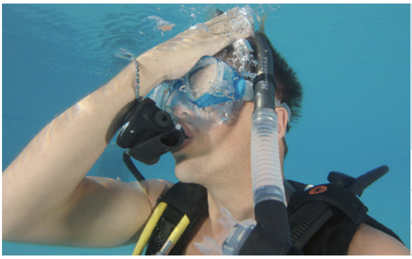

2 **Blow clear** by tilting your head back and pressing in the mask's upper edge. Exhale through your nose. The water should be forced out, escaping under the mask's lower skirt.

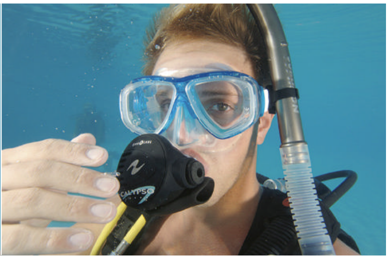

3 **Check that your mask** is free of water. If there is any remaining, repeat the process. This simple operation is quite normal during diving, and will quickly become second nature.

Specialized training

Diving is an inclusive, accessible sport, and there are many training options for children and the disabled. If you are already qualified but wish to undertake further training, there are numerous courses of advanced tuition to make you a better and more confident diver.

TAILORED COURSES

LIBERATING EXPERIENCE
A disabled diver undergoes training in a pool, with guidance from specially trained instructors.

Though standard open-water diving qualifications have a minimum age requirement (usually 14), children are now well catered to in diving. This means it can be a family sport—not one that is just for adults. For younger children, introductory courses are available, such as PADI's Bubble-Maker program, or diving clubs for children, such as the Scuba Ranger organization. These programs are designed to teach basic skills that lead to full open- water diving qualifications. Junior diving permits, available for children between 10 and 14 years old, are similar to basic open-water qualifications, but place special restrictions on the maximum depth to which the holder can dive.

Advances in training techniques mean that more disabled people are able to take up diving than ever before.

BUBBLE-MAKER TRAINING
Special courses for children are taught by most schools. Clubs also offer training for budding junior divers.

BECOMING INDEPENDENT
Your experience and judgment will build as qualifications are gained. Comprehensive training makes for a confident, safe diver.

Diving makes you effectively weightless, and the freedom of movement this offers makes the sport especially beneficial for those with limited mobility. The only condition that is an automatic bar to becoming a diver is epilepsy, due to the risk of seizures while submerged. However, blindness, paraplegia, cerebral palsy, and many other conditions have all been accommodated by adaptive training programs. The general rule is that if you have been cleared to dive by a doctor specializing in dive medicine, then training can begin.

DEEPER INTO DIVING

All the major training organizations offer courses that aim to take you a step onward from the basic open-water qualification. Such courses serve to refine your basic skills and introduce you to new challenges, such as deep diving, or diving in conditions of low visibility. They will also cover rescue and survival skills, and can be the first step on the path toward becoming a diving instructor. In addition to these skills development courses, there are also numerous specialty courses, which teach the technique and theory of activities such as cave diving, wreck diving, and mixed-gas diving. Such courses include dives to put new skills into practice, and help to improve your general diving experience. For more details, visit the websites of the various training organizations (*see pp.329–30*), or ask at your local diving center.

CAVERN DIVING
Specialized and difficult types of diving, such as cavern diving, ice diving, and other skilled activities, can be enjoyed after proper training.

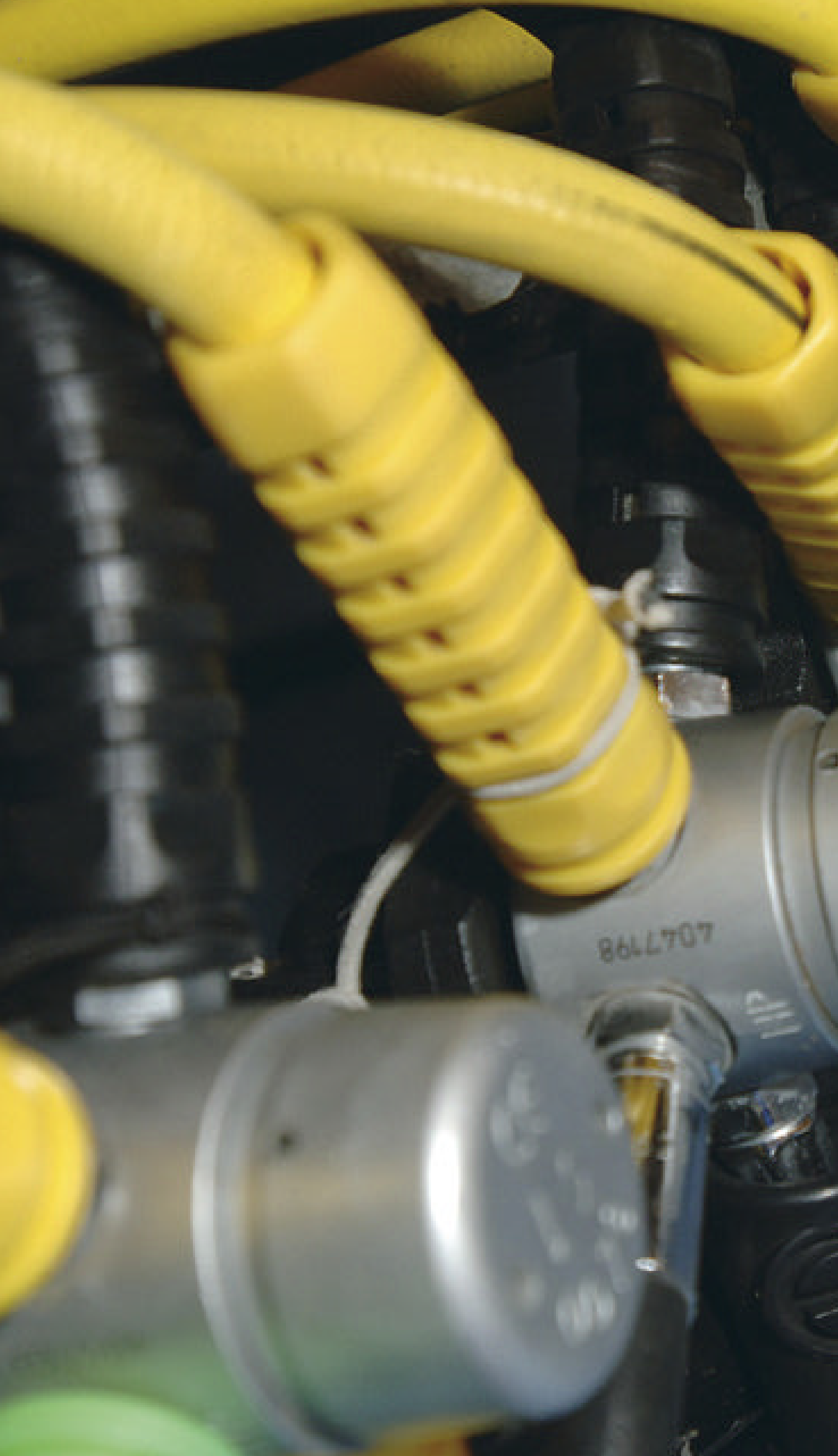

Equipment

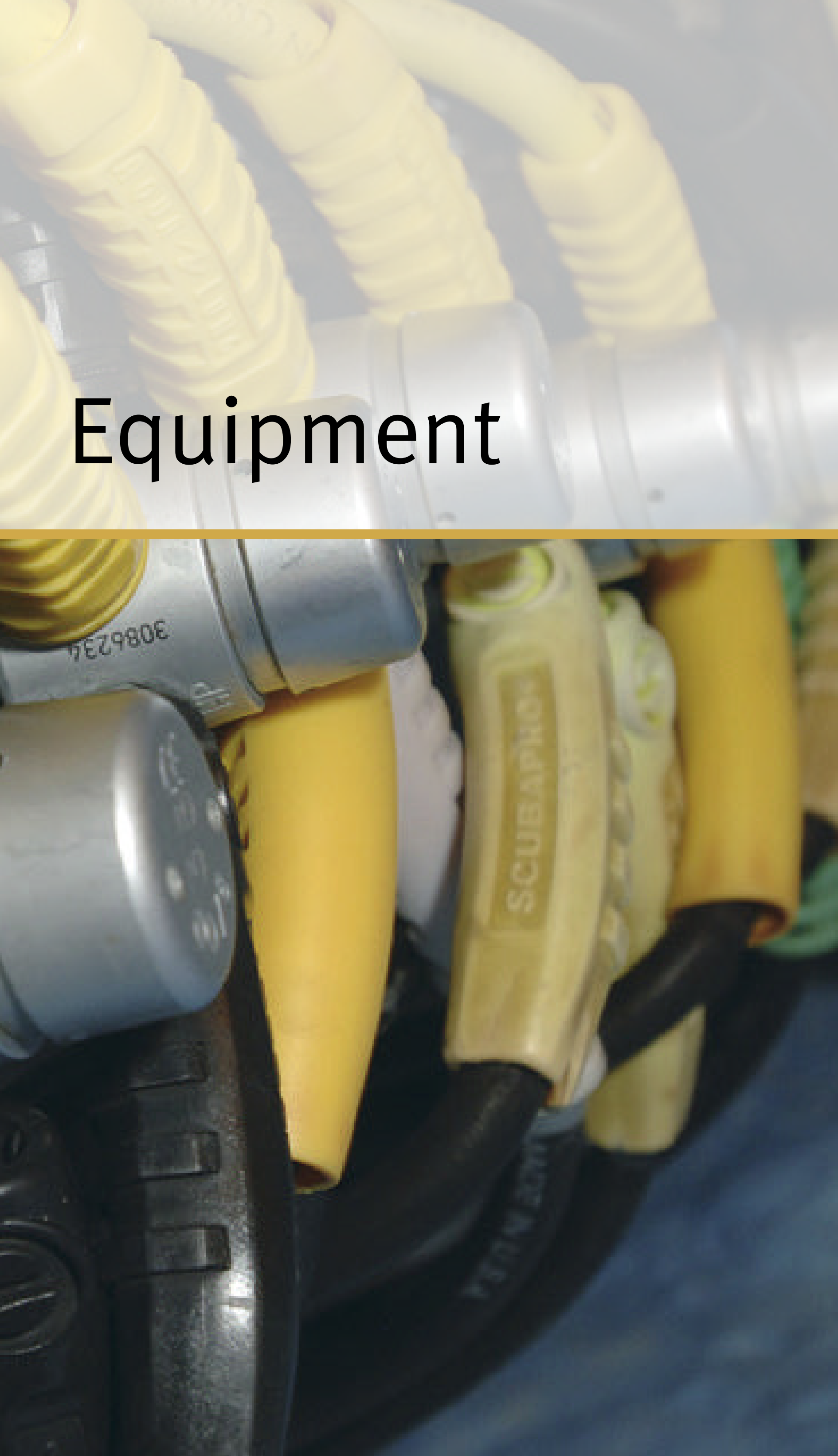
3086234
SCUBAPRO

Introduction

The wide array of equipment available is often bewildering to a novice diver. Some items are essential in all circumstances, while others are used only in specific diving environments. The more research you do and advice you seek, the more likely you are to buy safe, suitable gear.

BUYING AND RENTING GEAR

You could begin dive training with as little as a swimsuit, but it may be advisable to at least obtain a good-quality mask, snorkel, and fins. This means that you always have three key pieces of gear that fit perfectly, and with which you are comfortable.

Many divers own no more than this basic gear, preferring to rent the rest. There are pros and cons to this approach: it frees you from carrying heavy gear around, but it can also mean using ill-fitting equipment that has been worn by many people, and which may not have been well cared for. The alternative is to buy and transport all your own gear. This is recommended if you are likely to be diving regularly in your local area (although a local club or shop may have gear you can rent). It is not so practical if you plan to dive any farther afield—taking your own gear overseas can be expensive, although most dive tour operators are fairly generous with their excess baggage allowances. You can rent a tank and weights from any dive center, so you need never travel with these items.

The best places to buy equipment and obtain advice are dive shows or local dive shops. Second-hand gear is worth investigating, but inspect it for signs of wear and corrosion before making a purchase, and make sure that it is serviced before use.

DIVERS TOURING A REEF
Knowledge and care of gear is vital, since your equipment is your life-support system for exploring the underwater world.

BASIC EQUIPMENT CONFIGURATION

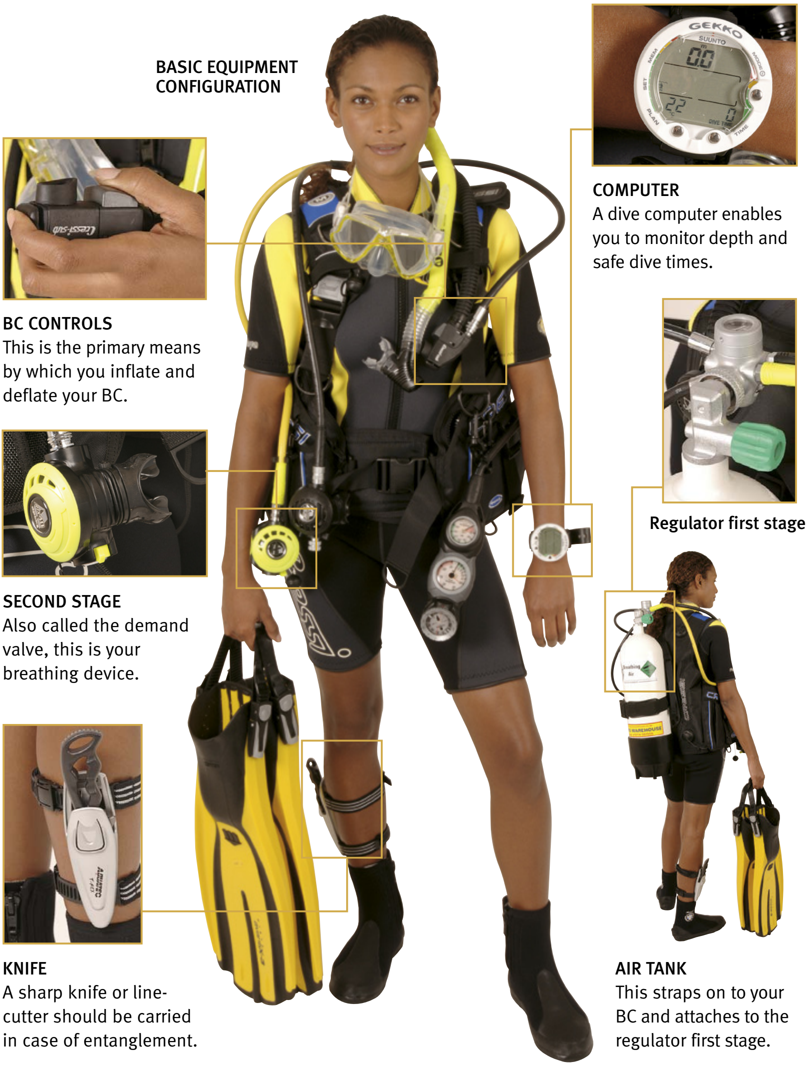

BC CONTROLS
This is the primary means by which you inflate and deflate your BC.

SECOND STAGE
Also called the demand valve, this is your breathing device.

KNIFE
A sharp knife or line-cutter should be carried in case of entanglement.

COMPUTER
A dive computer enables you to monitor depth and safe dive times.

Regulator first stage

AIR TANK
This straps on to your BC and attaches to the regulator first stage.

CORE EQUIPMENT

There are several items of gear that are essential, in addition to your mask, fins, and snorkel. A buoyancy compensator (BC) is an inflatable jacket that keeps you afloat at the surface, helps you adjust your buoyancy underwater, and supports other vital pieces of equipment. Weights, and a belt to mount them on, are another necessity, although weights can also be carried on some BCs (*see p.59*). You will also need some kind of dive suit (*see pp.50–55*) depending on the water temperature—and boots.

A regulator (the device that supplies air to the diver) is also necessary, preferably one that incorporates an alternative air source (an octopus) and an instrument console with a compass. Dive computers are a useful aid to safe diving, and you should also carry a knife to cut yourself free if you become snagged underwater.

DIVING ACCESSORIES

There is a range of equipment and accessory items designed for specialized diving. Try to borrow or rent specialty gear that you may only use infrequently.

Slates are useful accessories, since they allow you to write messages to your buddy, or to make notes from your dive—for example, on species that you do not recognize—or to sketch a site.

Flashlights are essential for diving at night, in murky water, and in low-light environments, such as caverns and wrecks.

If you will be doing a lot of open-water diving, especially in currents, you should purchase a marker buoy, either a surface marker buoy (SMB) or an inflatable surface marker buoy. These indicate your position to observers on the surface, and also aid your ascent.

EQUIPMENT FOR COLD WATER

Cold seas make extra demands on diving equipment, and additional peripheral items may be needed to ensure safe diving. Thermal protection is a priority.

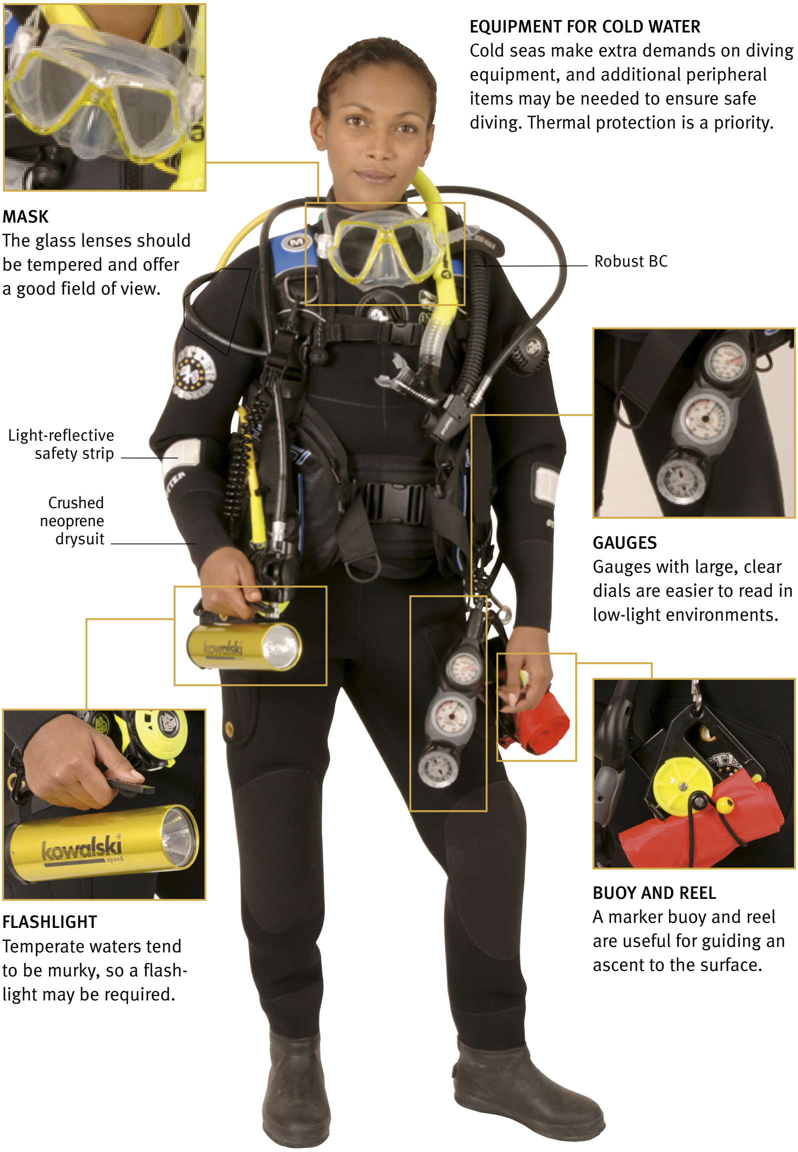

MASK
The glass lenses should be tempered and offer a good field of view.

GAUGES
Gauges with large, clear dials are easier to read in low-light environments.

FLASHLIGHT
Temperate waters tend to be murky, so a flashlight may be required.

BUOY AND REEL
A marker buoy and reel are useful for guiding an ascent to the surface.

DEMANDING ENVIRONMENT
Diving in cold water requires extra equipment, and appropriate training in how to use it. Safety devices are especially important in such waters.

Lastly, you may wish to record your dive experiences with a camera. Digital cameras in housings are now small enough to fit into pockets in your kit.

DIVING IN COLD WATER

If you plan to dive frequently in cold water, you might want to invest in a drysuit (*see pp.54–55*). A good thermal underlayer is essential with a drysuit, since loss of body heat can result in poor decisions and lead to exhaustion, speeding the onset of hypothermia. Dive clothing manufacturers make special thermal undersuits with wicking properties to draw moisture away from the skin. If you are planning on doing any extreme cold water diving or ice diving (*see pp.188–89*) you will need to use an environmentally sealed regulator, because these are more resistant to freezing.

Also necessary in cold water are a hood and gloves, the thickness of which will depend on the water temperature. The thicker the neoprene, the more insulation it offers, but the more it will restrict movement. Thick gloves or mittens reduce dexterity considerably, which is worth bearing in mind if you intend to use smaller items of gear.

MAKING THE RIGHT CHOICE

Don’t rush into buying the first bit of new gear you see—always look at a range of products and compare prices. Take advice from fellow divers and shop staff, but remember that everyone has their favorite brands. Reviews in magazines and on the Internet can give you more information, so that you can weigh the options before you buy.

Masks

A well-fitting mask is arguably your most important piece of diving gear, so take time to choose one that is right for you. There are many styles and designs to choose from, but above all you need one that fits correctly, and suits your personal diving requirements.

SEEING UNDERWATER

Human eyes cannot focus in water, which is why you have to wear a mask when diving. The way masks work is to trap a pocket of air in front of the eyes, allowing them to focus normally. A standard mask consists of a single or double panel of tempered glass inside a frame, with a silicone skirt forming an airtight seal against the face. A strap and buckles hold it all in place.

If you take care of your mask and keep it in a sturdy case to avoid damage, it should last for many years. The lenses of new masks have a protective coating from the manufacturers that should be removed before use. This can be wiped off using a nonabrasive cloth and non-gel toothpaste (or a fine abrasive cleaner) to clean the coating off without scratching the glass. If in doubt, consult the manufacturer's instructions. Make sure you rinse your mask in fresh water after use and store it away from direct sunlight in a dry place.

TESTING FOR FIT

There are lots of shapes and sizes of mask, so take time to find one you like that fits. A simple test is to hold it up to your face and inhale gently through your nose. If you then lean forward, it should remain in place.

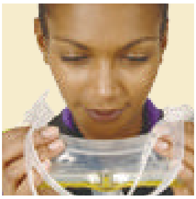

1 **Hold the mask** in front of your face with the strap hanging loose.

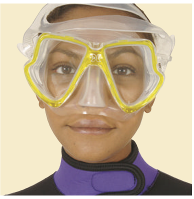

2 **Place it on** your face. Inhale through your nose—the mask should stay on without support.

MASK DESIGNS

CLASSIC This is a slightly old-fashioned design, but it is still popular with photographers because it gives a clear view of the diver's face. It offers a good field of view, but its high internal volume of air makes clearing the mask a little harder.

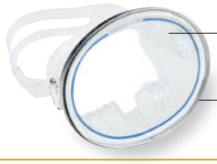

Single tempered glass lens
Oval metal outer frame

EXTENDED VISION Many new designs of mask are constructed to give the diver an extended field of view, especially downward. This model is a multi-lens design, which also allows corrective lenses of different strengths to be fitted.

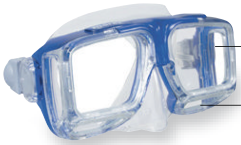

Removable lenses
Lower lenses enhance downward view

MATTE FINISH Some divers prefer to use a mask with a black silicone skirt to avoid confusing reflections on the inner surface of the glass. This is especially useful to marine photographers. However, this feature does reduce peripheral vision.

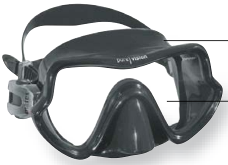

Opaque silicon skirt
Single-lens design

SIDE VIEW Some designs offer improved peripheral vision through glass panels on either side of the main lens, but are still low-volume, so are easy to clear. When choosing, make sure that additional lenses do not distort perspective.

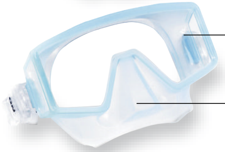

Side lenses widen view
Clear, light-permeable skirt

FULL FACE Mostly used in commercial diving, full-face masks are advantageous in very cold or polluted water. Some recreational models have intercom radio units so divers can communicate.

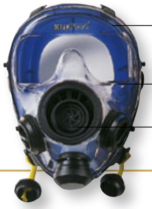

Quick-release faceplate
Primary lens
Radio intercom system

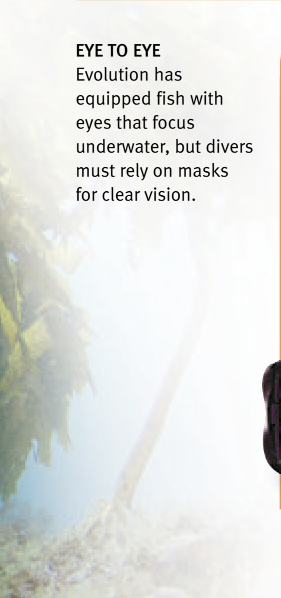

EYE TO EYE
Evolution has equipped fish with eyes that focus underwater, but divers must rely on masks for clear vision.

CORRECTIVE LENSES

Some masks can be fitted with prescription lenses, so if you wear glasses you will still be able to see underwater. Contact lens wearers can use normal masks—they provide an adequate degree of protection from water, unless accidentally removed.

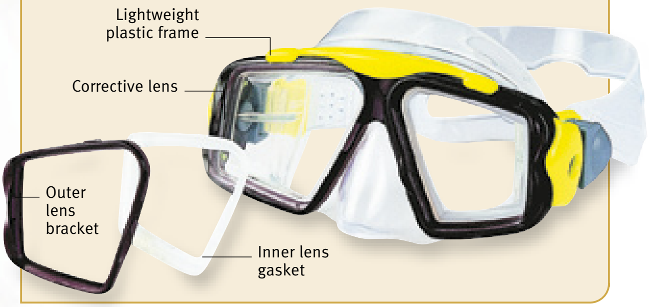

Lightweight plastic frame
Corrective lens
Outer lens bracket
Inner lens gasket

Snorkels and fins

Along with a mask, a snorkel and a pair of fins are probably the first pieces of gear you will buy. Snorkels vary very little, but there are plenty of styles of fins to choose from, depending on your budget. As always, look for a good fit and features that suit your diving needs.

CHOOSING A SNORKEL

A snorkel is simply a tube that allows you to breathe easily at the surface without needing to raise your head from the water. When diving, you do not need a snorkel underwater; however, you may use one before and after submerging. A good snorkel should be easy to breathe through, and should not be too long—ideally no more than 17 in (43 cm)—so it is easy to clear. Attach your snorkel to the left-hand side of your mask, leaving the right-hand side free for your regulator hose (*see pp.60–61*). Always rinse your snorkel after use and store out of direct sunlight.

DESIGN FEATURES

Basic snorkels are little more than a curved tube, but some models have a splash-guard at the top of the tube to prevent ingress of water, and a drainage valve at the lower end. Collapsible types are easy to stow away when not in use.

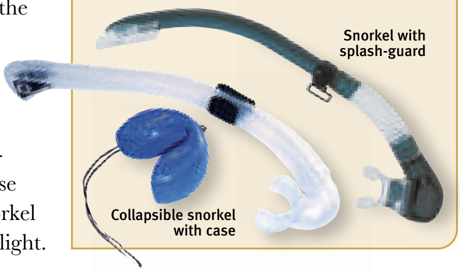

Snorkel with splash-guard

Collapsible snorkel with case

FITTING FINS

Sometimes incorrectly referred to as flippers, fins are used to aid propulsion when diving. A standard fin consists of a shoe made of rubber and a blade made of a stiffer material, often with ridges and vents to enhance performance. The shoe can either fit the whole foot like a slipper or leave the heel exposed. The latter is held on with a strap, and is designed to be worn over neoprene boots. When buying fins of this type, it is advisable to take your boots along with you, since their thickness will need to be taken into account when testing for fit.

SELECTING THE RIGHT PAIR

Advances in fin technology mean that there is a wide range for the diver to choose from, depending on their specific preferences. For example, split fins offer a high degree of agility underwater, but many divers believe traditional models deliver more outright thrust. Before deciding which pair to buy, consider the kind of diving you will be doing, and seek advice from store staff or other divers.

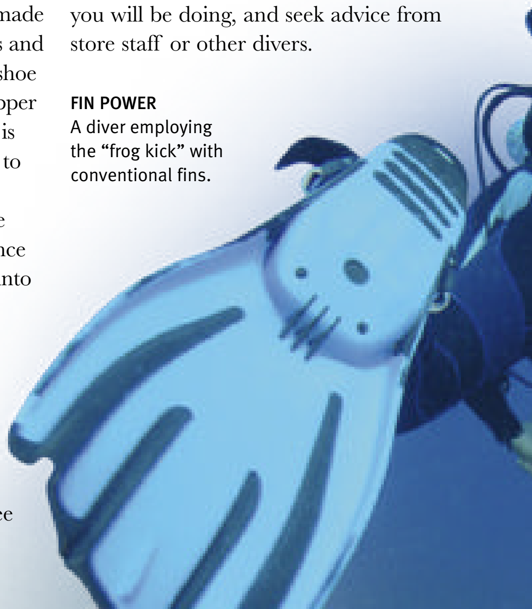

FIN POWER
A diver employing the "frog kick" with conventional fins.

TYPES OF FINS

CONVENTIONAL A basic fin with a single, fairly rigid blade, the conventional design is a good starter fin, but you may want to try other types to see if they offer features that suit you better.

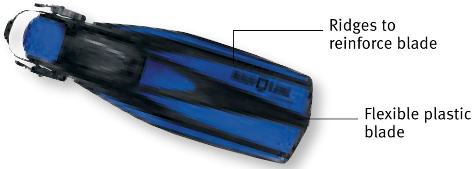

Ridges to reinforce blade
Flexible plastic blade

FORCE FINS Made exclusively by a California manufacturer, this range of fins is designed to minimize the water turbulence produced, and reduce the diver's kicking effort.

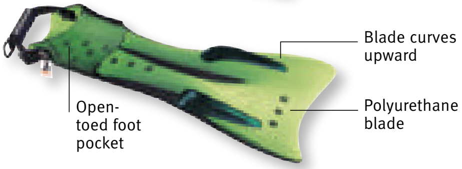

Blade curves upward
Open-toed foot pocket
Polyurethane blade

PIVOTING A design that has been adopted by two manufacturers, this fin features a flex-point in its structure and a V-shaped area of softer plastic to scoop water for better traction and speed.

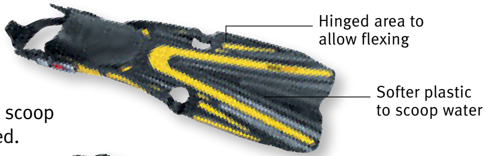

Hinged area to allow flexing
Softer plastic to scoop water

SPLIT A popular style, split fins offer better agility than standard fins. However, some divers believe they are less effective than conventional fins in strong currents.

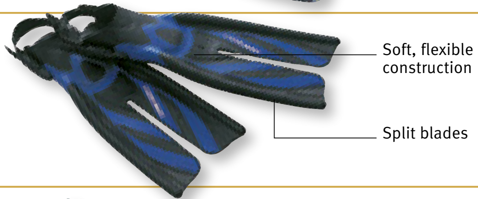

Soft, flexible construction
Split blades

SNORKELING Fins designed for snorkeling are lighter than scuba fins, and of a foot-pocket design, so they can be used with bare feet. Some also have shortened blades.

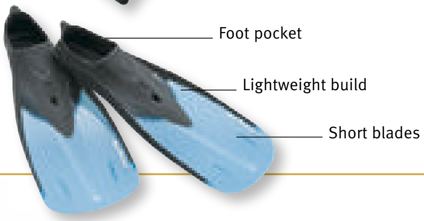

Foot pocket
Lightweight build
Short blades

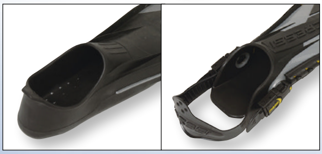

Foot-pocket fin

Heel-strap fin

FIT AND FASTENING
Fins are held on the foot in two basic ways: in a form-fitting foot pocket, or by a strap over the heel.

Diving suits

In all but the very warmest waters, you will need some kind of thermal protection during your dive. Good diving suits (which are technically known as “exposure suits”) are available in every price range, but always look for comfort, fit, and quality of construction.

BEATING THE CHILL

Water conducts heat away from the body much more efficiently than air. Unless the water you are in is the same temperature as the surface of your skin—about 90°F (32°C), which is hotter than most corals can withstand—your body will lose heat, and you will get cold. This means that in almost all diving environments, thermal protection is required. If you are diving in warm waters, a normal wetsuit will be sufficient. Only in the warmest waters should you consider a non-insulating or “skin” suit. Never be tempted to dive in just a swimming suit; even if you don’t need insulation, a thin layer will prevent stings and abrasions. If you are diving in cooler waters, you should consider the heavyweight options: a semi-dry suit or a drysuit.

SECOND SKIN
Modern wetsuits are made of materials that offer highly efficient heat insulation.

WHICH SUIT?

Personal warmth requirements vary. If possible, take various suit options—you can layer these to give you the best protection. It is rare to be too hot underwater, but in tropical climates you may overheat in your suit before you enter the water. This can be prevented by wetting the suit before putting it on.

Buying a suit for every temperature range is too expensive for most divers, but most suits will suffice for a variety of waters. Although in the tropics you would be too warm in a drysuit, you could opt for a full-length wetsuit, as its extra warmth would be welcome in the cooler depths.

In addition to a suit, you may need accessories such as gloves, boots, a hood, and talcum powder, which is useful for easing entry into suits.

EXPOSURE PROTECTION

There are guidelines for the type and thickness of suit needed in different conditions. For the tropics, where the waters rarely drop below 68°F (20°C), a 2–3-mm wetsuit is fine. In the Mediterranean in summer, a 5–6-mm full wetsuit is more appropriate, rising to 7–8 mm during winter months. A semi-dry suit of 10–15-mm thickness, with a hood and thick gloves, will cope with most conditions in colder temperate waters, but in waters below 50°F (10°C), a drysuit is necessary.

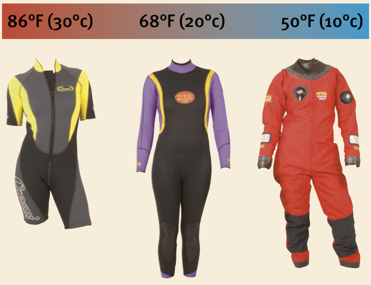

Shortie **Full wetsuit** **Drysuit**

SUIT LAYERS

Wetsuits and drysuits keep the diver warm in different ways. In wetsuits, a thin film of water is trapped next to the diver's skin and acts as an insulator. Drysuits seal a layer of air inside the suit, and this permits the use of thermal undergarments for extra insulation.

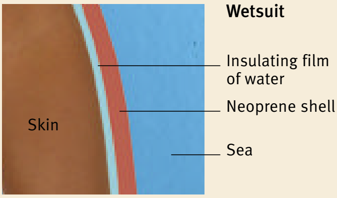
Wetsuit
Skin
Insulating film of water
Neoprene shell
Sea

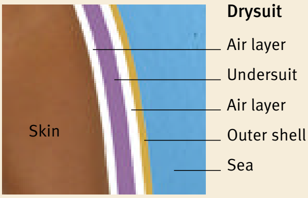
Drysuit
Skin
Air layer
Undersuit
Air layer
Outer shell
Sea

SKINSUITS AND SHORTIES

In warm, tropical waters, you will probably not require any thermal insulation, but a thin layer over your body will prevent your equipment from rubbing on your skin, as well as giving protection from the Sun before and after dives and providing a barrier to stings and scrapes.

The thinnest skinsuits are made of Lycra, but some manufacturers offer suits made from 0.5-mm neoprene. Lycra suits can also be used as a base layer under a heavier suit, so they are a good add-on option if you are not sure how warm or cold your destination will be.

So-called "shortie" wetsuits offer slightly more thermal protection than skinsuits, typically being constructed of 3-mm neoprene. However, they do leave the arms and legs exposed and unprotected from stinging plankton and jellyfish or abrasions on rocks or coral.

Non-chafing collar
Tough plastic zipper
Thin neoprene

SHORTIE WETSUIT This type of wetsuit is a good choice in warm waters, and does not cost as much as a full suit.

Leather palms

Light gloves

Light material

Skinsuit vest

TROPICAL DESCENT
Even in warm waters, a full-length wetsuit may prove invaluable, because water temperatures drop as the diver ventures deeper.

FULL-LENGTH WETSUITS

The next level of thermal protection after shorties and skinsuits is a full-length wetsuit. These offer increased thermal protection for the arms and legs, and are typically made from 2–8-mm neoprene. But thickness is not the only factor in thermal efficiency—a close-fitting suit provides better insulation. A correctly fitting wetsuit should not only be easy to put on, but should follow the contours of your body as closely as possible. Wetsuits rely on trapping a thin film of water next to the body, which gradually becomes warm. If the suit fits badly, water flows more freely between the body and the suit, and chills—rather than insulates—the diver.

CARING FOR YOUR WETSUIT

A simple maintenance regimen is vital to prolong the life of your wetsuit. After every dive, wash your suit to remove salt and debris, which can cause it to rot. It is best to use special wetsuit shampoo, which helps to maintain the suppleness of the neoprene and prevent mildew. Once the suit is fully dry, place it on a hanger and store it in a cool, dry place.

FULL-LENGTH WETSUIT
Scuba wetsuits are available in a range of styles, categorized by the thickness of neoprene they are made from.

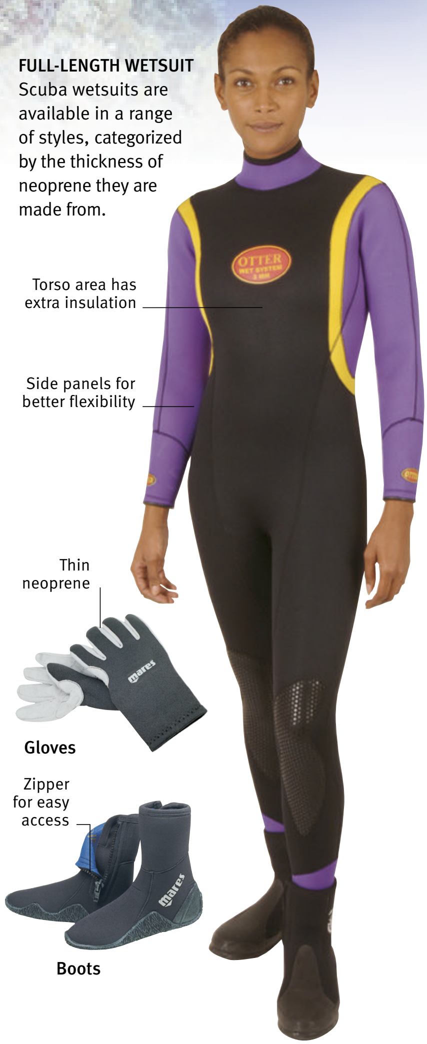
Torso area has extra insulation
Side panels for better flexibility
OTTER WET SYSTEM
Thin neoprene
mares
Gloves
Zipper for easy access
mares
Boots

SEMI-DRY SUITS

Though intended for colder temperate waters, where they are worn with a hood and gloves, semi-dry suits are a versatile style that can be used even in warm Mediterranean seas. Like standard wetsuits, semi-dry suits are made of neoprene and work by trapping an insulating layer of water between the body and the suit. However, they have more efficient seals at the wrists and ankles and around the neck or hood, which stops water from flushing through the suit and allows the trapped water to warm up. They also tend to be made of thicker neoprene and are often designed to be used in layers (with a jacket fitting over a coverall-style base layer).

The thickness of the neoprene in some semi-dry suits has a tendency to restrict movement on land (although not in water). However, suits made of super-flexible material are available that offer an equivalent degree of thermal protection, but are much easier to put on and move around in. Although a semi-dry suit offers protection similar to a drysuit (*see p.54*), some divers prefer a drysuit, because a damp body can be uncomfortable after the dive, especially in cold weather.

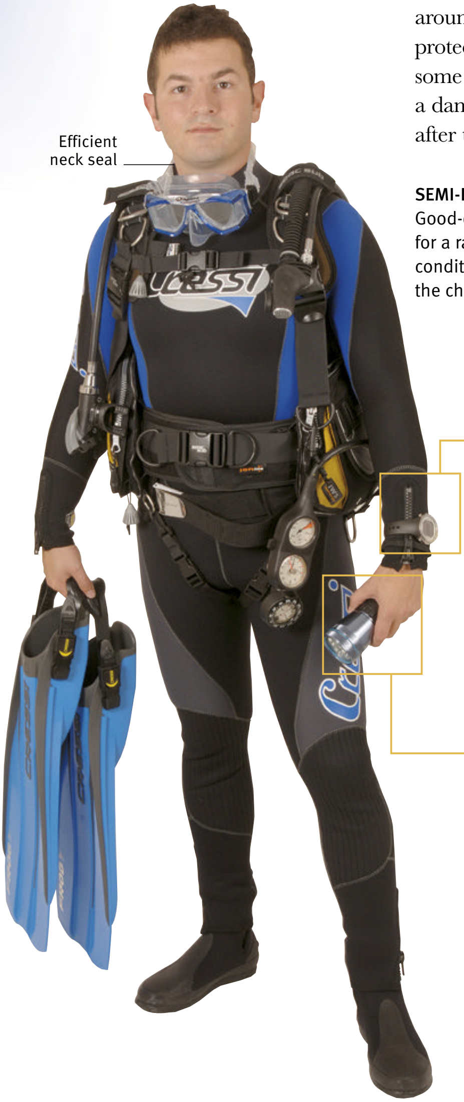

SEMI-DRY SUIT WITH EQUIPMENT
Good-quality semi-dry suits are suitable for a range of dive environments and conditions, and allow you to withstand the chill of greater depths.

DIVE COMPUTER
Often no bulkier than a wrist watch, a dive computer may need a strap extension when worn with thicker suits such as a semi-dry suit.

FLASHLIGHT
In the colder and deeper waters accessible with a semi-dry suit, visibility can be impaired, so a light should always be carried.

DIVING IN DRYSUITS

Drysuits are the most expensive type of diving suit but offer excellent thermal protection. They use trapped air, not water, as an insulating barrier. Drysuits are available off the rack, but it is worth spending a little extra to get a made-to-measure garment, because a well-fitting suit will be more comfortable and more likely to avoid buoyancy problems caused by unequal air distribution.

Drysuits are made from a variety of materials, including crushed neoprene and membrane materials. Crushed neoprene suits are tough and offer a good degree of thermal insulation. Membrane drysuits are made from strong, thin, waterproof fabrics, which offer less inherent thermal insulation than crushed neoprene, and are designed to be used with a thermal undersuit. They are more supple than neoprene types, and so easier to move around in. However, they can be looser-fitting, and allow excess air to move around inside, which can shift buoyancy to the legs, and tip the diver upside down.

KEEPING WARM
Even in conditions as extreme as those encountered diving under ice, drysuits provide good thermal protection, ensuring the diver stays comfortable.

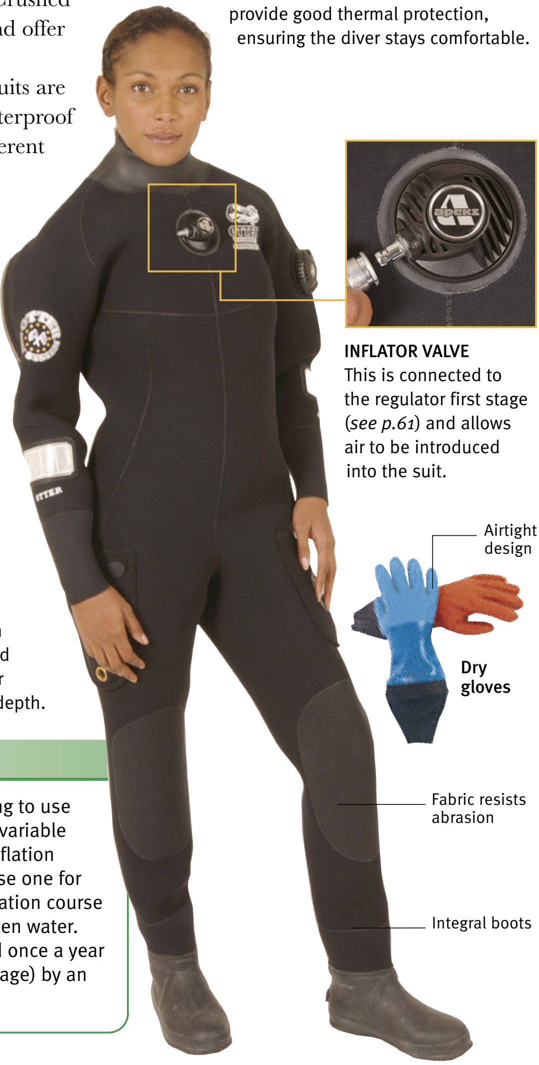

INFLATOR VALVE
This is connected to the regulator first stage (*see p.61*) and allows air to be introduced into the suit.

CRUSHED NEOPRENE SUIT
The inherent buoyancy of a suit made of neoprene that has been "crushed"—treated with heat and pressure to flatten its internal air cells—remains the same at any depth.

USING A DRYSUIT

Drysuits are more challenging to use than wetsuits, as they have variable buoyancy, due to their air-inflation mechanism. If you plan to use one for the first time, take an orientation course in a pool before diving in open water. A drysuit should be serviced once a year (or after long periods of storage) by an approved repair technician.

FASTENINGS AND SEALS

Drysuits generally have an “across the shoulder” zipper to allow entry. You will need help from your buddy to do this up. These zippers need regular coating with wax or “zip slip” to keep them lubricated and protected from corrosion.

To prevent water from entering the suit, drysuits have watertight seals—made of thin neoprene or latex—at the wrist and neck. Latex seals are generally less likely to leak than neoprene versions, but do not offer any insulation and are more fragile, so there is the risk that they may tear. Neoprene seals, by contrast, are tough, easy to repair, and warm, but it can be difficult to get a good fit.

INFLATING WITH AIR

Drysuits have an inflator valve on the chest that injects air to reduce the “shrink wrap” sensation caused by water pressure squeezing the suit against the body. This air expands on ascent and must be dumped via a valve, normally located on the shoulder, which can be manual or automatic. Cuff dumps are also available that work by raising your arm to release excess air.

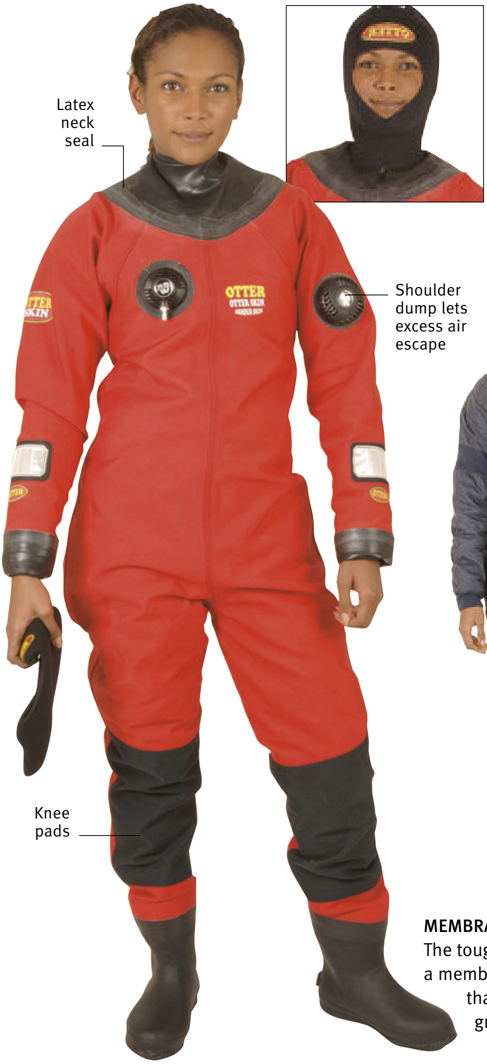

TOP COVER
A neoprene hood is necessary in cold waters to insulate the head.

MEMBRANE SUIT
The tough, lightweight fabric of a membrane suit is more flexible than neoprene and so allows greater ease of movement.

High-tech fibers trap heat inside

THERMAL UNDERSUIT
Specialty garments, undersuits can be worn under drysuits for extra insulation and comfort.

Thick neoprene

Heavy wetsuit gloves

Buoyancy compensators

Unlike fish, humans have no natural way to adjust their buoyancy, or floating ability, when underwater. Divers wear a buoyancy compensator (BC)—a jacket that can be filled and emptied of air—to control their buoyancy underwater and at the surface.

HOW A BC WORKS

The BC (also known as BCD, or buoyancy compensation device) offers flotation at the water's surface and fine adjustment of buoyancy underwater, and acts as a harness for equipment items. At the surface, a fully inflated jacket will keep your head well out of the water. Underwater, buoyancy can be increased by inflating the BC with small amounts of air.

All BCs are connected via an air hose to the regulator first stage (*see p.61*), which supplies pressurized air from the tank for inflating the jacket. BCs can also be inflated orally, via a corrugated hose that hangs down over the left shoulder. Controls for automatic inflation and deflation are usually located at the end of this hose.

Of course, BCs must be able to reduce buoyancy as well as increase it. This is achieved by venting air from the jacket via dump valves at various points. Air can be released through the oral inflation hose by depressing a button, and can also be vented by using a pull-cord dump valve, which acts more quickly. These are often located on the opposite shoulder from the oral inflation hose, and at the bottom edge of the jacket. They should be used cautiously, since they alter buoyancy rapidly. Many BCs also vent air when the oral inflation hose is tugged sharply.

STOWING KIT
Buoyancy compensators (*right*) are fitted with handy stowage points for peripheral gear.

SURFACING IN SAFETY
When fully inflated, BCs allow the diver to float effortlessly at the surface. Here, two divers signal OK after a dive.

FIXINGS AND FASTENINGS

The tank (*see p.68*) is also attached to the BC. A strap attached to a rigid backplate wraps around the tank and holds it in place with a cam-action buckle. It is important that your tank is fastened securely and does not move during the dive. A well-designed BC will have various adjustable straps around the waist and at the shoulders, to ensure that it fits snugly. These can be easily re-adjusted underwater. Steel or plastic D-rings (*see below*) can be used to attach smaller items of gear.

FRONT VIEW OF BC JACKET

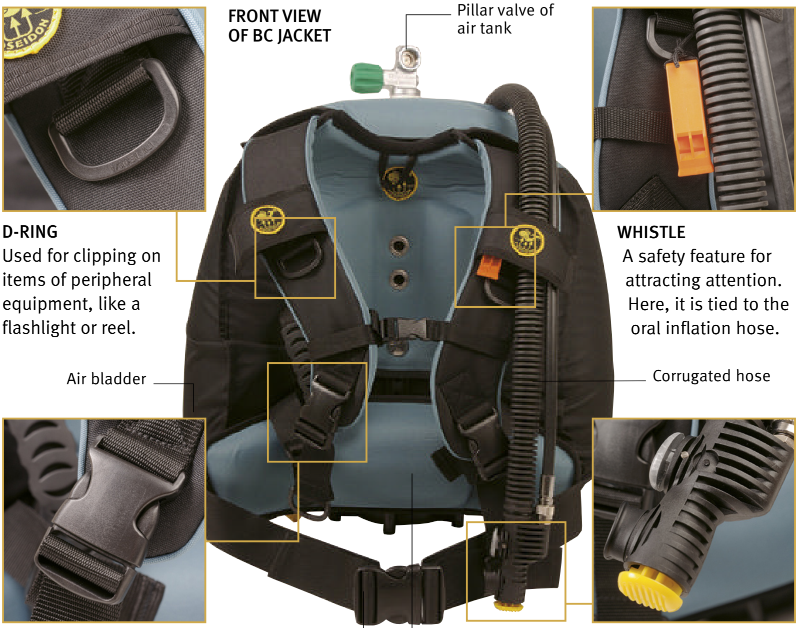
Pillar valve of air tank
Air bladder
Corrugated hose
Waist belt
Lumbar support pad

D-RING
Used for clipping on items of peripheral equipment, like a flashlight or reel.

WHISTLE
A safety feature for attracting attention. Here, it is tied to the oral inflation hose.

HARNESS RELEASE CLIPS
For quick release from the harness, and easy donning of the jacket.

BUOYANCY CONTROL
The inflation and deflation controls and oral inflator are at the end of the hose.

SIDE VIEW OF BC JACKET

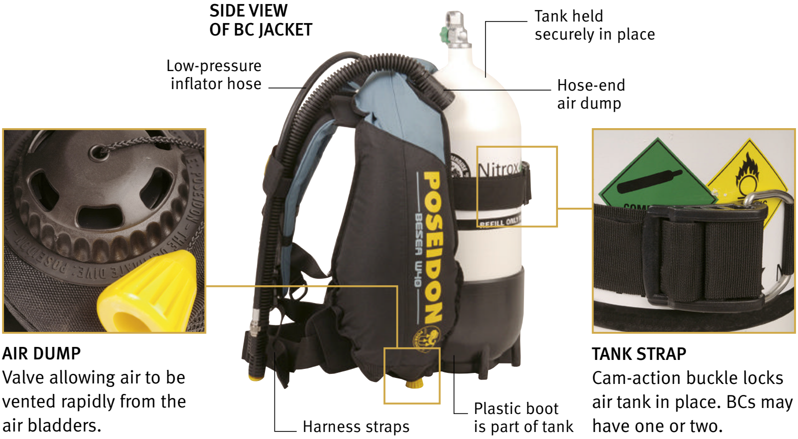
Low-pressure inflator hose
Tank held securely in place
Hose-end air dump
Harness straps
Plastic boot is part of tank

AIR DUMP
Valve allowing air to be vented rapidly from the air bladders.

TANK STRAP
Cam-action buckle locks air tank in place. BCs may have one or two.

SPECIALTY JACKETS
High-volume "wing" style BCs (*top*) are used in advanced diving (*left*). The large air capacity of these jackets provides extra lift for heavily laden divers.

CHOOSING A JACKET

A basic BC is like a padded vest, and these are often only adjustable at the waist. More advanced models have adjustable straps to make sure the jacket fits as snugly as possible underwater. Ideally, you should buy one that fits like a glove and does not roll around with a tank attached. You should also find it easy to take on and off. If possible, you should try it on with your wetsuit (or drysuit) to check the fit, and to ensure that there is room for more adjustment on every strap in case of changes in your suit or your body shape. Beyond the basic models of BCs there are the more specialized features available on some jackets, such as weight integration, or a "wing" style design (*see opposite page*). Whatever the features of the jacket you decide to buy, make sure they suit the type of diving you will be doing the most. Try out several different jackets before deciding.

SERVICING YOUR BC

Careful maintenance of your BC will keep it in good working order for many years. It should be serviced at least once a year, or after it has been stored unused for more than six months. Take it to a recommended dive shop. Only service it yourself if you have been trained to do so. Do not take risks by diving with old or unserviced gear.

CARING FOR YOUR BC

As with all items of equipment, your BC should be rinsed with fresh water after every dive. The inside should also be rinsed by flooding it through the corrugated hose. The jacket should then be emptied again by turning it upside-down and draining it through the same hose. After this, you should reinflate it to help locate and remove any remaining water, which will rattle around inside when the BC is shaken. The jacket should be stored in a cool, dry place in a partially inflated state.

BUOYANCY COMPENSATOR DESIGNS

INTEGRATED WEIGHT BC
Some BCs feature integral weight systems. Many people find these quick-release weight pockets more comfortable to use than a traditional weight belt. The weights should be removed before putting the BC on and taking it off.

Inflator mechanism incorporates a backup regulator

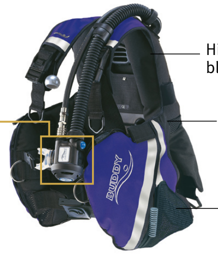
High-lift air bladder
Tough plastic backplate
Quick-release weight pouches

INTEGRATED BREATHING BC
The Mares HUB (Human Underwater Breathing) system was the first to integrate the regulator into the design of the BC. A great advantage of this top-of-the-line product is its simplified profile, which reduces drag in the water.

Regulator integrated with jacket design

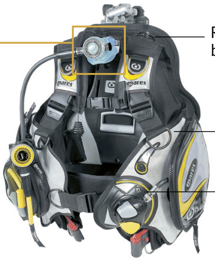
Regulator built into BC
Reflective shell for visibility
Ergonomic inflator controls

WING / SEMI-WING BC
Favored by technical divers (*see p.194*), these offer high lift capacity. The air bladder is rear-mounted: this aids horizontal posture under the water, but can push the diver face-forward at the surface. A semi-wing offers better surface support.

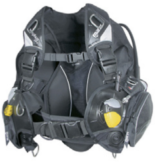

Semi-wing

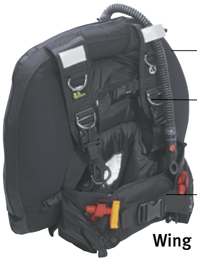
Rear-mounted air bladder
Steel D-rings for attaching gear
Padded waist belt for comfort

Wing

WOMEN'S BC Women divers are increasingly well catered to with special jacket designs to give a comfortable fit without the usual tension points across the chest. Such jackets offer other convenient options like weight integration.

Specially shaped harness

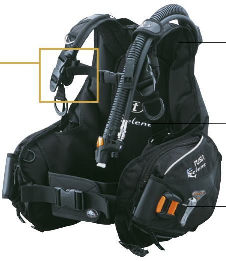
Tough nylon-weave fabric
Padded backplate
Quick-release weight pouches

TRAVEL BC If you travel often, it is a good idea to consider a lightweight BC without a rigid back, which can be rolled up. These designs have a lower lift capacity, as they are mostly used in warm-water sites, where minimal weight is carried by the diver.

Toggle for jettisoning weight pouches

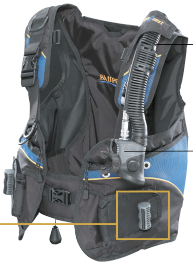
Lightweight, foldable form
Inflator controls

Regulators

One of the most important items of diving gear is a regulator. This two-stage gas-supply system provides you with air from your tank at the correct pressure for you to breathe underwater. Since it is a key life-support device, most divers buy their own regulator.

BREATHING UNDERWATER

Your tank can store large amounts of air in a small space because the air it contains has been highly compressed (*see pp.68–69*). Before you can breathe this stored air, it must be reduced to a safe pressure, which varies according to how deep you are underwater. The deeper you go, the greater the weight of water pressing on your chest cavity from outside (*see pp.94–95*). This is called ambient pressure. Air for breathing must be supplied at an equivalent pressure, otherwise your lungs will not be able to inflate properly.

STORING REGULATORS
Regulators are tough, but they should be well cared for. Hang them up in a cool, dry place away from sunlight.

HOW REGULATORS WORK

Supplying air at the right pressure is the job of the regulator, which is made up of the first stage (which fits on to the tank) and the second stage (the piece you put in your mouth), which is also known as the demand valve. The first stage reduces the air passing out of the tank to a lower pressure of around 116–145 psi (8–10 bar) above ambient. A valve in the first stage allows some of this low-pressure air to pass into a hose, and then closes again. When the air in the hose has been inhaled by the

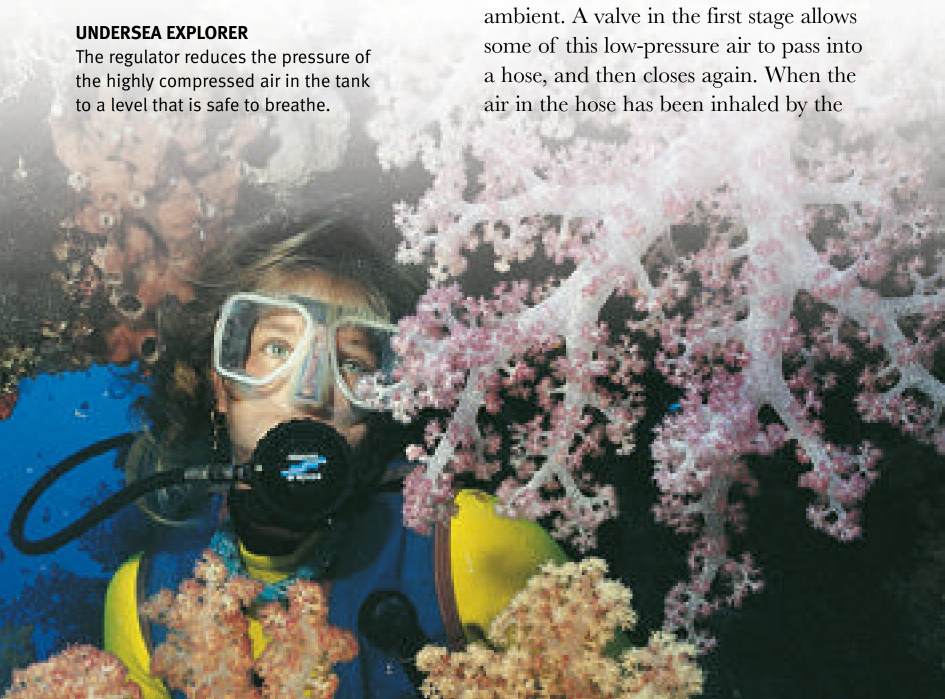

UNDERSEA EXPLORER
The regulator reduces the pressure of the highly compressed air in the tank to a level that is safe to breathe.

diver, the first stage detects a decrease in pressure in the hose, and allows more air to flow in, to replace that which was used. It automatically alters air-pressure in the hose when depth changes occur. The first stage also supplies air for BC and drysuit inflation (via separate hoses).

The second stage reduces the low-pressure air in the hose to the same pressure as the water you are swimming through, allowing normal breathing. The second stage houses a diaphragm that is subject to equal forces of water pressure on one side, and air pressure from your breathing passages on the other. When you inhale, the diaphragm flexes inward, causing an inlet valve to open, and air flows into your mouth. Exhaling causes the diaphragm to flex outward and triggers an exhaust valve to open, allowing used air to be expelled. The unit can be cleared of water by using the purge button, which is located on its front surface.

A-CLAMP FIRST STAGE
A common and convenient clamp-style first stage, suitable for most recreational diving requirements. Less secure than DIN.

DIN-TYPE FIRST STAGE
A screw-in first stage that can handle high tank pressures of more than 3,000 psi (200 bar). Highly secure.

SECOND STAGE
Designed to supply air safely at all depths, your second stage should fit comfortably in your mouth.

INSTRUMENT CONSOLE
A high-pressure hose from the first stage feeds directly into a pressure gauge, indicating remaining air supplies.

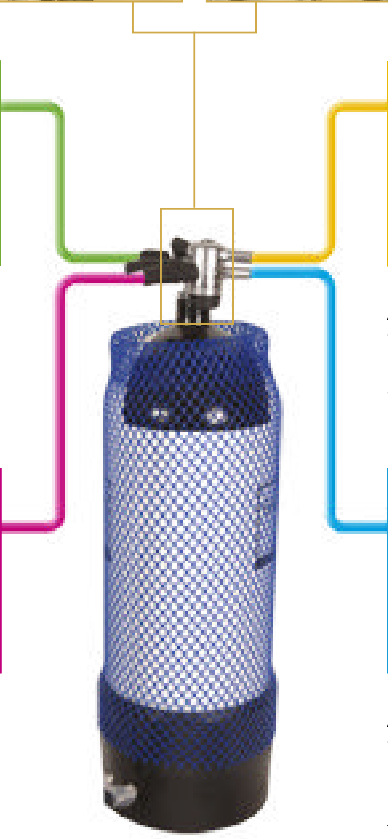

OCTOPUS SECOND STAGE
The octopus is for use in the event that your main second stage fails, or for donation to other divers.

BC JACKET
A low-pressure hose from the first stage supplies air to your BC to inflate the jacket underwater.

THE OCTOPUS SECOND STAGE

Most recreational regulators are set up in an "octopus" configuration (so called because of the multiple hoses emerging from the first stage). This incorporates a backup second stage, which is intended for use if your main one fails, or if your buddy needs a donated air supply. Over time, this backup air source has become known as the octopus second stage, or simply the octopus.

The octopus and the hose to which it is attached are usually finished in a bright yellow color for easy location in an emergency, and to distinguish them from your main second stage (which ordinarily has a black hose). The octopus hangs unused from your BC most of the time, and for this reason it is often detuned (adjusted) to prevent "free-flow"—an accidental discharge of air that occurs more easily when the mouthpiece of the regulator is left continually flooded. Some types of octopus are attached to the hose via a swiveling joint, so that they can be easily and safely donated to your buddy regardless of whether they are beside you, or face-to-face.

The hose supplying an octopus should be longer than the one feeding the main second stage. This will allow it to be offered to a buddy easily and for the buddy pair to remain face-to-face during a shared-air ascent (*see p.140*). Practice air-sharing procedures regularly.

STOWING THE OCTOPUS

Consider where to stow your octopus so that it can be easily located in an emergency by either you or your buddy. You can buy specialized clips of various designs to attach the octopus to your BC, which allow it to be freed easily when required, and replaced again afterward. Many divers hang their octopus near their midriff, but others mount it higher up on the chest, so it remains constantly within sight and close at hand. You can also buy regulator retainers made from rings of surgical tubing. These go around your neck, meaning the octopus is always right under your chin, though some find this arrangement constricting.

SAFE STOWAGE
Attach your octopus to your BC at a point where you can easily access it, and where it will be clearly visible at all times.

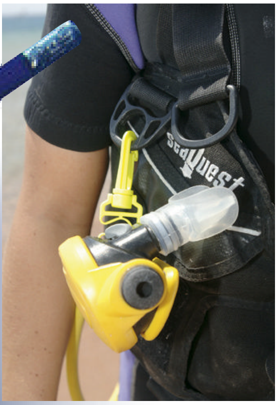

EMERGENCY ACCESS
Special quick-release clips are available for fastening your octopus to your gear.

TYPES OF REGULATORS

BASIC REGULATOR Simple, inexpensive regulators at the budget end of the market can be very reliable and durable. However, some may deliver air a little less smoothly than high-performance models.

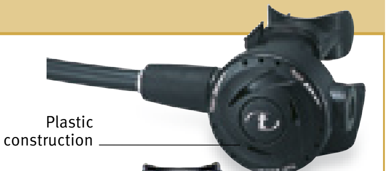

OCTOPUS SECOND STAGE Octopus second stages are often just standard second stages with high-visibility color schemes. However, some designs can be swiveled to any angle for easy access.

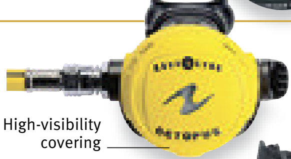

HIGH-PERFORMANCE REGULATOR At the top end of the market are regulators that combine high build quality with low inhalation effort and toughness. Many are environmentally sealed for cold-water use.

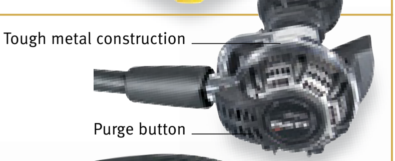

OXYGEN-COMPATIBLE REGULATOR Divers who wish to use nitrox mixes (*see pp.194–95*) that are very rich in oxygen will need a regulator designed to handle high oxygen concentrations without risk of combustion.

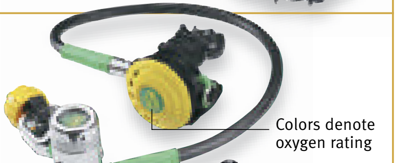

VINTAGE-STYLE REGULATOR In twin-hose regulators, the valve of the second stage is separated from its mouthpiece, and sits behind the diver's head, where it releases bubbles out of the diver's field of vision.

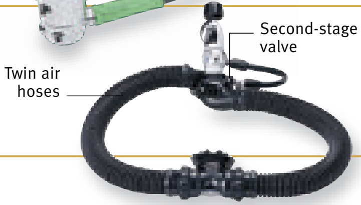

CHOOSING A REGULATOR

As with all items of dive equipment, prices vary enormously, but the regulator is one piece of gear worth investing decent money in. Basic regulators use a piston valve in the first stage and can be very reliable but often feature rather chunky second stages. The more expensive models use a diaphragm in the first stage and tend to have more streamlined second stages. Top-of-the-line models may be designed specifically for cold water diving, or be optimized for efficient air delivery on deep dives. They may also have dials to adjust the inhalation effort required to open the valve and deliver air. These can be used either to tune the unit to free-flow less easily, or maximize air delivery during a normal dive. Always rinse your regulator in fresh water after use, and have it serviced annually by a qualified dive-equipment technician.

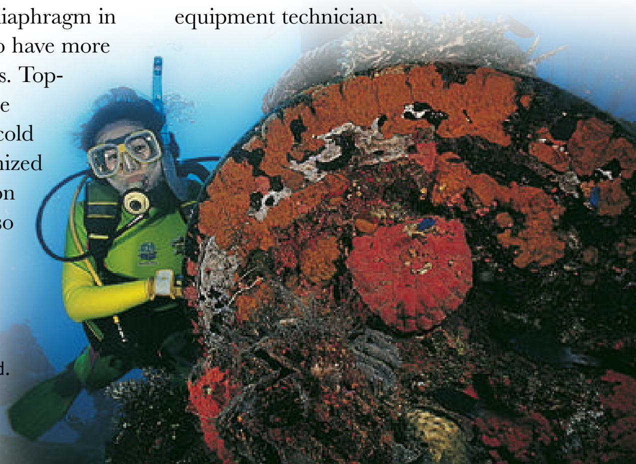

DEEP BREATH
Regulators are designed to stringent standards, and are very reliable if well-maintained.

Gauges and consoles

To stay within safe time and depth limits and avoid running out of air, you must carry instruments that monitor your air consumption, time underwater, and depth. There are various devices on the market to help with this, from simple gauges to complex digital consoles.

BASIC INSTRUMENTS

You should never dive without some means to measure the time you spend underwater, the amount of air you have left in your tank, and your depth. A waterproof watch with a bezel—a rotatable outer ring—allows you to mark significant times, and thus keep track of how much time has elapsed. It also serves as a backup in case of failure of digital equipment used to monitor time, such as a dive computer.

Depth gauges allow you to monitor your depth even if you cannot see the water's surface. They are usually analog (needle-and-dial) devices, although digital versions are also available. Both work in the same way: they measure the pressure of the water around you, and convert this into an accurate reading of your depth. Make sure you buy one that records the maximum depth of your dive, and remember to reset it after every dive. Equally important is an air-pressure gauge, also known as a submersible pressure gauge (SPG) or contents gauge. This is connected to your regulator first stage via a high-pressure hose and measures the pressure of the air in the tank. You should check your air-pressure gauge every few minutes during the dive to carefully monitor your air consumption.

INDICATING REMAINING AIR SUPPLY
The air-pressure gauge allows you to keep track of the amount of air in your tank; let your buddy know (*see p.145*) when your supply is running low.

CHOOSING A CONSOLE

Your console is your dive information center, so you must be happy with the way it handles. Different consoles need to be held at different angles to read the dials; experiment before you buy to find out which is easiest for you to use. Choose one with large, easy-to-read numbers. Some are equipped with other instruments, such as a compass; these can be either on top of the console or on the back.

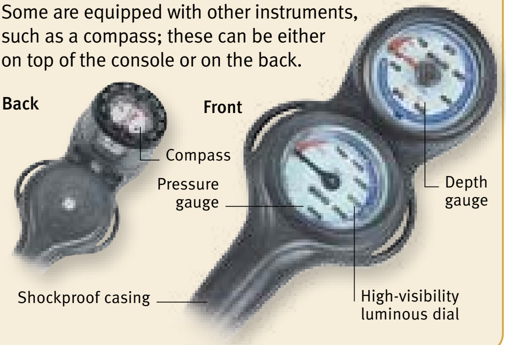

GAUGE CONSOLES

Depth and air-pressure gauges are often combined in a single integrated console that attaches to your regulator first stage and hangs by your side during a dive. Mounted together like this, the gauges are easy to locate and can be

TYPES OF GAUGES AND CONSOLES

STAND-ALONE DEPTH GAUGE A depth gauge is usually oil-filled and reacts to ambient pressure changes. Mostly analog, these gauges are accurate and durable.

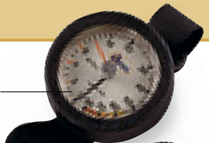

STAND-ALONE AIR-PRESSURE GAUGE Typically analog, this type of gauge contains a tube that expands or contracts according to the air pressure inside your tank.

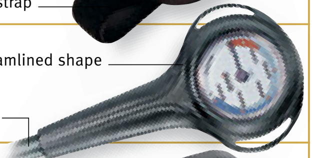

DEPTH AND AIR-PRESSURE CONSOLE This option allows at-a-glance checking of both depth and tank pressure, and also registers your maximum dive depth.

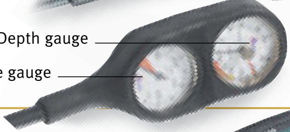

DIGITAL CONSOLE This console has an integrated computer and displays a range of information, such as air consumption, temperature, depth, time, and ascent rate.

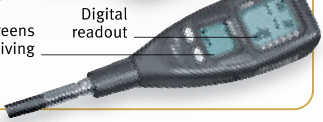

read simultaneously, like a "diver's dashboard." They can sometimes also include other instruments, such as a compass or a thermometer.

Consoles should be rinsed with fresh water after use and stored away from direct sunlight, and kept in a padded or protective case to avoid damage.

ADVANCED CONSOLES

Some manufacturers offer instrument consoles with digital rather than analog gauges. Consoles are also available that perform the same advanced functions as a dive computer (*see pp.66–67*). As well as indicating dive time, depth, and decompression schedules like a normal computer, these also incorporate a digital air-pressure gauge. For divers who use a wrist-mounted computer to monitor dive information, console instruments can provide an essential backup.

ANALOG VERSUS DIGITAL

Although less easy to read, analog gauges sometimes give slightly more accurate readings than digital ones, particularly at shallow depths.

Dive computers

To avoid decompression sickness, it is critical to monitor your depth and time underwater on every dive. In the past, this would have been done using a watch and a depth gauge, but these functions and many others can now be performed by a dive computer.

USING A DIVE COMPUTER

The main functions of a dive computer are to monitor your dive time and depth and calculate safe dive schedules. A basic model uses this data to assess the level of nitrogen absorption in your body while you are under the water, and also displays information about your ascent rate, including any stops you may need to make during an ascent to avoid decompression sickness (*see pp.100–01*). More advanced models also inform you about the temperature of the water and your air consumption (if linked via a transmitter to your regulator first stage), both for normal air and technical gas mixes (*see pp.194–95*).

Computers are more flexible than dive tables (*see pp.100–01*). Dive tables show how long you can spend at different depths before needing decompression stops. Most tables assume that you spend your entire dive at the greatest depth; computers can make more complicated calculations, including the time spent at all the different depths reached on a dive. Some computers will also schedule deep decompression stops. These reduce the risk of nitrogen microbubbles forming in your body, meaning that less time is spent decompressing in shallow water. When making repeat dives, especially in places where you are not limited by times of tides and slack water, using dive

TYPES OF DIVE COMPUTERS

MULTI-GAS This type of computer can calculate dive schedules for several different mixtures of breathing gas, from normal air to various oxygen-enriched blends of nitrox.

Computer can be linked to rebreather

Detailed display, which uses icons as well as numbers

CONSOLE-INTEGRATED This computer is built into the diver's instrument console (*see p.65*), which is attached to the high-pressure hose from the regulator. As such, it also displays remaining air supplies.

Integrated compass

Hand-held device with finger-tip controls

WRIST-MOUNTED This instrument is typical of a basic level, watch-style dive computer. It provides data on depth, dive time, ascent rate, and safe dive schedules, and is for use with normal breathing air.

Data displayed in basic form

Push-button controls alter information displayed

RADIO-UPDATED This wrist-mounted computer also displays how much air remains in the diver's tank by means of radio updates from a module screwed into the regulator's first stage (*see pp.60–61*).

Wrist strap

Radio module

SAFETY CHECKS
As you ascend, keep a close watch on your dive computer to make sure that you are not exceeding the maximum ascent rate it recommends.

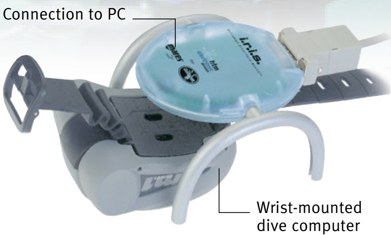

DIGITAL INTERFACE
Peripheral devices are available that link dive computers to PCs, enabling you to download dive profiles and create dive logs.

tables instead of a computer may not give you the maximum number of dives you can safely make each day.

Computers are also able to monitor your ascent rate accurately. In your dive training, you will be taught to ascend no faster than the smallest exhaled bubble, but it is more accurate to follow a digital readout that tells you the safe ascent rate.

BUYING A DIVE COMPUTER

A key consideration when buying a computer is the readability of the display. Too many figures on a small screen are difficult to read—the simplest models are often the best. Choose one with digits that are large enough to read with ease, without any nonessential data on the primary screen. It is also advisable to opt for a brand with a proven track record.

Computers are the most precise way of monitoring your dive, but they are not a foolproof means of avoiding DCS. It is impossible for data from a computer to exactly reflect your body's requirements. All models have built-in safety margins, but to be sure of avoiding problems, you should always dive conservatively, and not push your depths and times to the limit.

MALFUNCTION PRECAUTIONS

Never rely solely on a computer. As a precaution against malfunction, it is always worth practicing how to calculate dive times using tables. Although your computer will give early warning of a low battery, make sure you carry back up devices—either a watch and depth gauge, or a second computer, which can be mounted on your console.

Tanks

Also referred to as “cylinders” and “bottles,” tanks contain pressurized air for breathing underwater and have a valve that delivers this air to the regulator. Requiring regular testing and inspection, they can almost always be rented from dive centers.

SIZE AND MATERIALS

Different sizes of tanks are available, but most commonly they have a volume, or capacity, of 72, 80, or 95 cubic feet. The maximum pressure to which these can be filled varies—it can be as much as 4,350 psi (300 bar), but is more commonly 3,000 psi (206 bar). These figures are stamped on the neck of the tank, along with the dates of manufacture and most recent testing. Visual inspection at a certified dive shop must be carried out annually, and hydrostatic testing (for metal fatigue) is required at different intervals (from one to five years) in different countries. Your test stamps must be in date for your tank to be filled by any reputable dive center.

Many recreational divers use an 80 cubic foot tank, but if you find you usually run low on air before your buddies do, you should consider using a larger size. On deeper, or more lengthy dives, two tanks (a “twin-set”) may be needed. Compact “pony bottles” are often used as a backup air source; mounted next to the main tank, these are equipped with an independent regulator.

SECURED AGAINST MOVEMENT Tanks stored upright must be secured. Always store them out of direct sunlight.

TYPES OF TANKS

Tanks may be made of aluminum or steel. Steel is heavy and corrodes in salt water, but if well maintained it is durable and long-lasting. Aluminum is a lighter metal, but because tank walls must therefore be thicker, this adds bulk, and largely negates any weight advantage; they are also more prone to damage.

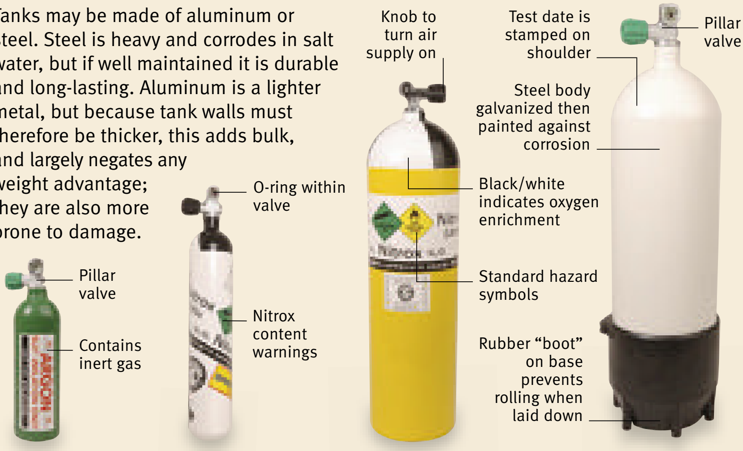
Pillar valve
Contains inert gas
O-ring within valve
Nitrox content warnings
Knob to turn air supply on
Black/white indicates oxygen enrichment
Standard hazard symbols
Test date is stamped on shoulder
Steel body galvanized then painted against corrosion
Rubber “boot” on base prevents rolling when laid down
Pillar valve

Suit inflation tank

3-liter pony bottle

Nitrox tank

A steel tank

Tanks are most commonly filled with compressed air. For technical diving (*see pp.194–95*), however, they can be filled with special gas mixes, such as nitrox and trimix. Drysuit inflation tanks, used mainly by technical divers, are filled with the inert gas argon, which, being denser than air, retains body heat for longer.

Because the air inside the tank is used up during the course of a dive, it gradually becomes lighter. Most people forget that air weighs a certain amount too, and compression packs a great deal of it into one tank. Aluminum tanks may even become positively buoyant as air is used up.

THE TANK VALVE

A valve screwed into the neck of the tank provides an attachment for the regulator first stage, for either an A-clamp or a DIN (*see pp.60–63*), and a means of turning the gas supply on and off. Always check that there is a sound O-ring in the valve before fitting your regulator. Without one, or with a damaged one, the fitting won't seal.

MAINTENANCE AND STORAGE

Fill your tank at reputable dive shops or centers and make sure it is "in date" and regularly serviced. If you are storing your tank on a boat, do not leave it standing up (especially if gear is attached) unless secured in a rack. Take care to lay it down and secure it so that it does not roll around. Tanks are full of highly-pressurized gas, so handle with care. Store tanks out of direct sunlight, especially when full, as this can cause a pressure buildup inside. If you have not used your tank for more than six months, have it refilled to get rid of any stale air.

DOUBLE CAPACITY
A twin-set is ideal for deep dives that require long, scheduled decompression stops.

Weight systems

When diving, your suit, BC, and even your own body all add to your buoyancy. There are many systems for carrying the lead ballast used to offset this, and they can be tailored to the environment you will be exploring. Comfort and ease of use should be a consideration.

STAYING DOWN

When wearing a drysuit or thick wetsuit (*see pp.50–55*), you will need to wear a considerable amount of weight to counteract the buoyancy of the suit. The most inexpensive way of wearing ballast is on a belt carrying hard lead tablets. Other systems use lead shot as the ballast, which is more comfortable to wear around your waist, as it molds to your contours. This is of most relevance to women, who often find that a standard weight belt is uncomfortable on the hips.

PASSING A WEIGHT BELT
When handing over a belt, always hold it with the buckle down, so that the weights cannot slide off.

Many BCs now feature integrated weight pockets, offering a more adjustable approach to weight distribution and trim. The lead carried by such BCs is mostly stored in quick-release pockets, but there are usually secondary stowage areas that are inaccessible during a dive. It is important to make sure the amount of releasable lead in the BC will give you sufficient positive buoyancy to guarantee an unassisted return to the surface if you have to jettison it.

WEIGHT SYSTEMS

STANDARD WEIGHT BELT A basic nylon belt carries lead tablets that have two slits through which the belt is threaded. The entire belt is dumped in an emergency, via a rapid-release cam buckle.

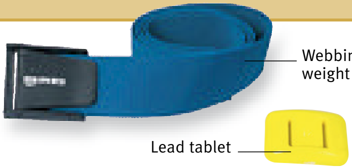

POUCH-STYLE WEIGHT BELT Some belts have pouches that hold bags filled with lead shot, which are easy to pull out and discard. A variation on this is a belt with a single compartment to which you add loose shot.

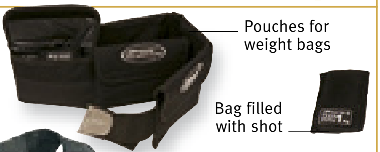

WEIGHT HARNESS Taking weight off your waist and suspending it from both shoulders makes carrying the load more comfortable. Quick-release catches allow dumping in an emergency.

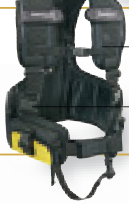

WEIGHT-INTEGRATED BC This is perhaps the most comfortable way of carrying weight. The weight can be distributed around your body, aiding correct posture in the water (*see also* BCs, *pp.56–59*).

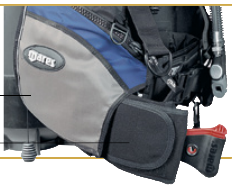

SECURING WEIGHTS TO A BASIC BELT

To prevent the lead tablets from sliding along your belt and altering your stability during a dive, you can secure them in place with plastic keepers. Alternatively, you can twist the belt to hold them in place. Insert the belt through the first slit in the weight, make a half turn, and then run the belt through the second slit.

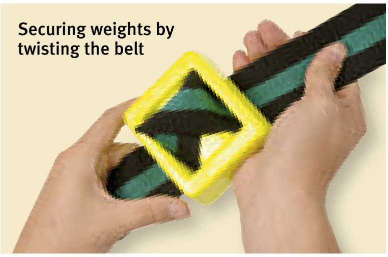

Securing weights by twisting the belt

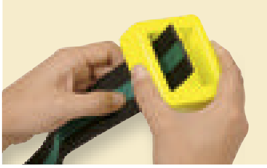

1 **To use a keeper** on a basic weight belt, thread the belt through the weight tablet, and locate in the desired position.

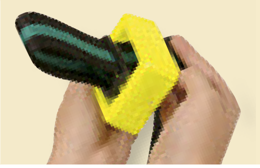

2 **Take the belt** up through the keeper, bend and thread back through the keeper, then through the weight.

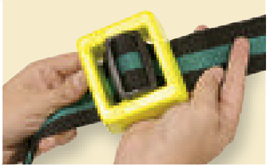

3 **Pull the belt** taut on each side of the weight. Be careful to avoid kinking the webbing as you do so. Repeat as necessary.

REFINING YOUR WEIGHT SYSTEM

Weights are something you will not want to travel far with—and you do not need to, because they can be rented at dive centers. You will find that when you dive in different environments and in different suits, your weight requirements change. Record the configurations that work best for you somewhere handy, such as at the back of your log book. Whichever way you decide to wear your weights, it is vital that you can release them quickly and safely in an emergency.

LEAVING GRAVITY BEHIND
Heavy weights, though awkward on land, will not be a burden underwater.

ANKLE WEIGHTS

When a diver wears a drysuit (*see pp.54–55*), the air within the suit can become trapped around the legs and feet, inverting the diver. Ankle weights will help to counteract this undesirable effect.

Weights mounted in ankle strap

Dive flashlights

A flashlight is an essential item of dive gear. Although primarily used when diving at night and in conditions of reduced visibility, such as in caves or in murky water, flashlights also have other useful roles, including illuminating subjects for underwater photography.

LIGHTING UNDERWATER

Lights vary enormously in design and cost, but the basic function is to provide reliable illumination when submerged. There are good designs in nearly every price range, but flashlights, especially hand-held ones, are easy to lose and damage, so it is worth weighing up carefully how much you want to spend. It is also wise to obtain an equally reliable backup flashlight, such as a smaller pencil-style model, to stow in a pocket, for use if your primary light fails during a dive.

BURN TIME AND BATTERIES

One of the most important factors to consider when buying a flashlight is burn time—you do not want to run out of battery charge halfway through a dive,

USING A DIVE TORCH

Flashlights are easy to lose, especially during dives in strong currents. Wrist straps offer a degree of security, but it is safer to clip them on with a flexible leash. Stow them close to your body to minimize the risk of being hit on rocks. Also, try not to switch the light on and off too often during the dive, because this may cause the bulb to fail.

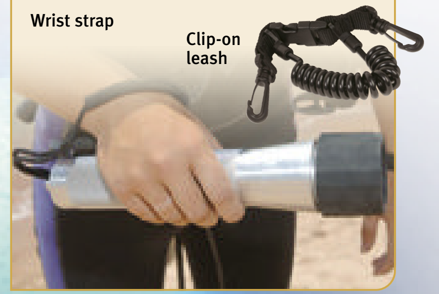

LIGHT IN THE DARKNESS
A cavern diver surveys the inside of a sea cave. Lights are an essential piece of diving safety equipment.

TYPES OF FLASHLIGHTS

HIGH-INTENSITY DISCHARGE This type of light burns brighter than a halogen bulb and, because it requires less power, has a long burn time. Such lights can be heavy and expensive, however.

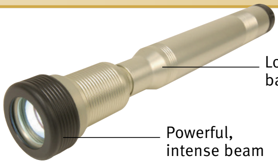
Long casing for batteries
Powerful, intense beam

BACKUP A range of flashlights have evolved that are intended for use if the diver's primary light fails. As such, they are small and lightweight, but only generate enough light for emergency use.

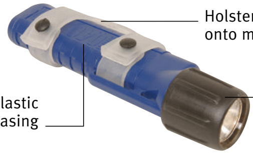
Holster can be clipped onto mask strap
Plastic casing
Twisting collar activates light

PISTOL-GRIP This is probably the most widely used style of light among recreational divers. They are fairly inexpensive and usually use conventional non-halogen filament flashlight bulbs.

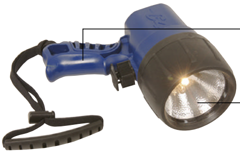
Pistol grip for ease of use
Wide reflector for broad beam

UMBILICAL An umbilical light has the battery housed in a separate case. This model uses high-intensity light-emitting diodes (LEDs) to generate a beam. Umbilicals have long burn times.

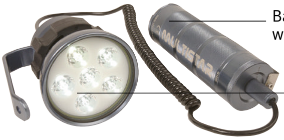
Battery pack worn by diver
Six LEDs for efficient burn

HIGH-TECH RECHARGEABLE This model can be charged via a special "wet" contact that means the case does not have to be opened, minimizing the risk of flooding through careless reassembly.

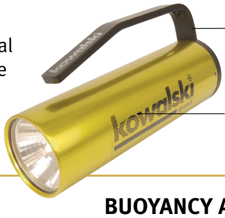
kowalski
kowalski
Rigid metal handle
Precision-engineered case

Rear of light

which could compromise safety. There is usually a trade-off between output and burn time. Some lights are very bright and offer an excellent field of illumination for a short duration. But many lower-output models can outlast much brighter competitors, so consider carefully which would be more useful for you before making a purchase.

Rechargeable flashlights are more cost-effective than those requiring battery replacements, and tend to offer the greatest light output but the shortest burn time. Also consider the recharge time, especially if you are diving from a liveaboard boat and the vessel's generator is run only for short periods of time.

BUOYANCY AND BEAM

Another factor to consider is buoyancy. Some flashlights are positively buoyant, because they have a large airspace within the reflector. This makes them unbalanced in the water and difficult to handle. A positively buoyant light cannot be laid down, for example, if you need to light a subject for photography.

Lamps also vary in the size of the hotspot (central area of beam) and peripheral halo. The halo can be useful because it lights up a wider area than just the main hotspot. The hotspot is the brightest part of the beam, and is best for highlighting specific objects or for illuminating murky surroundings.

Surface marker buoys

Marker buoys may be used either during a whole dive or just before you ascend, to let those at the surface or on the shore know where you are, and also to aid your ascent. These devices offer a sense of security for both diver and surface cover, and aid safe recovery.

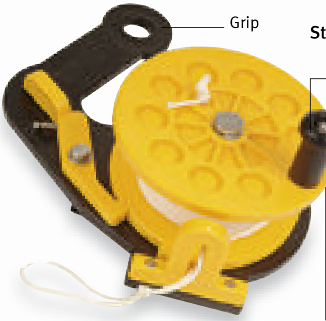

STANDARD REEL AND LINE Reels are available for both right- and left-handers for ease of winding.

REEL AND DSMB Line and buoy can be wound and clipped together for compact storage.

SMB OR DSMB?

A surface marker buoy (SMB) is an inflatable float linked to the diver by a line. It allows an observer at the surface to monitor the diver's position underwater. SMBs should be in a high-visibility color—most are orange.

The line to which the float or buoy is attached is usually stowed on a special reel. The reel should be held in one hand, with not too much slack on the line. You should use an SMB when your boat cover needs to be aware of your position at all times, such as during drift dives (where undersea currents propel the diver along), or where there is a likelihood of heavy boat traffic overhead.

Another type of buoy is deployed only at the end of a dive, or in an emergency. The delayed SMB (or DSMB) is carried deflated during the dive, then inflated underwater (*see pp.130–31*) before ascent, carrying a line from the reel up to the surface. When it pops up,

DIVER WITH DSMB INFLATED By the time you surface, your boat cover should be en route to pick you up.

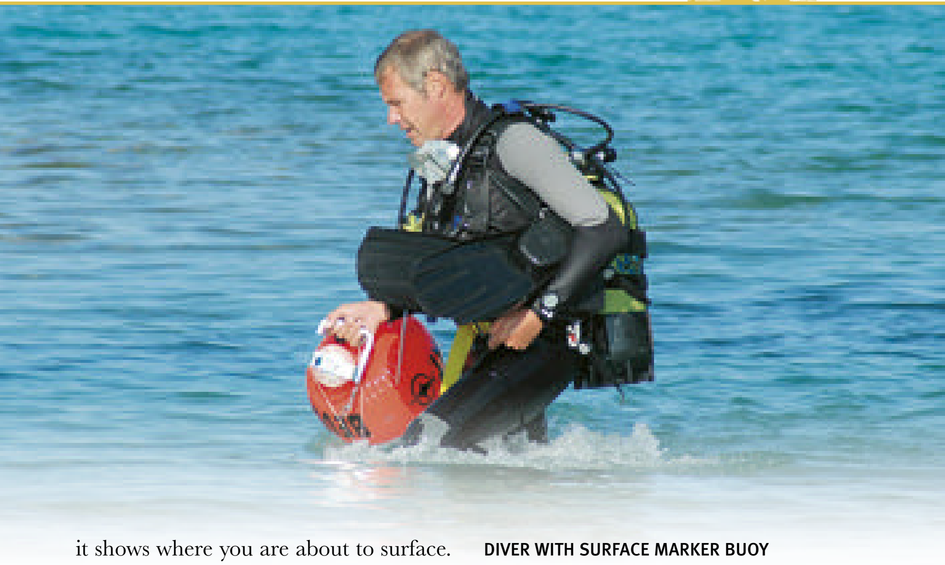

DIVER WITH SURFACE MARKER BUOY
Only one SMB, inflated before the dive, is needed per buddy pair, provided you remain together during the dive.

it shows where you are about to surface. Its line also acts as a guide for making a slow, safe ascent and for decompression stops (*see p.101*). Always make sure you are some distance from other divers using buoys before deploying, or you risk fouling each other's lines.

USING A REEL

Both SMB and DSMB lines should be tied with a bowline knot to the reel. A reel should be compact enough to fit into a pocket or clip onto your BC, but not so small that it is difficult to operate, especially if gloves are worn. Some reels have a ratchet action to stop the line from unwinding on its own, and this helps to avoid tangles. The length of line you use should correspond to the depth you are expecting to dive to; this may be 130 ft (40 m) or even more. If the line is too short, then when it runs out, the DSMB will yank you upward at a dangerously fast rate until the buoy reaches the surface. This can also happen if the reel jams, which is why you must never deploy a DSMB while the reel is still clipped to your BCD: you must hold the reel in your hand so that in an emergency you can simply let it go.

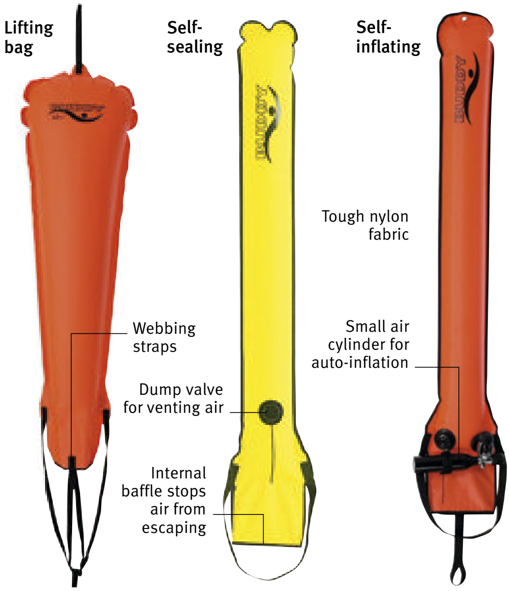
Lifting bag
Self-sealing
Self-inflating
Tough nylon fabric
Webbing straps
Small air cylinder for auto-inflation
Dump valve for venting air
Internal baffle stops air from escaping

TYPES OF DSMBS
The cheapest lifting bags have an open base from which air can escape after inflation. Self-sealing types are more reliable.

Safety and signaling devices

It is rare to be lost at the surface, but you must be equipped for such an eventuality. To guide rescuers to your position, you should carry high-visibility surface location aids. The boat crew must know what color equipment you are wearing, and which aids you are likely to be using.

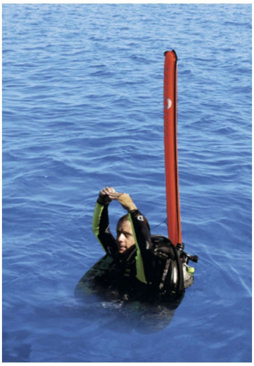

SURFACE OBSERVATION SIGNAL This tall, inflatable signal is easily seen by boats or aircraft. It is inflated using compressed air from your cylinder.

SIGNALING OPTIONS

Even on a relatively straightforward dive, the current can take you a long way away from your boat cover, and heavy swell can make you difficult to locate. If this happens, stay close to your buddy, and, if possible, secure yourself to them. Signaling to the boat will speed up recovery, and there are various gadgets available for this. The devices that are most commonly used are inflatable surface marker buoys (*see pp.74–75*). It is best to use the long "safety sausage" type, which stands up in the water, and remains inflated once deployed, rather than an open-ended one. If you opt for the latter, make sure you also carry an alternative device, such as a collapsible flag. Surface horns and power whistles are also used. They are of limited range, but can help to attract

SIGNALING DEVICES

FLAG Collapsible plastic flags in high-visibility or fluorescent colors are excellent for signaling at the surface, especially in a swell. They can be rolled and stowed alongside your cylinder.

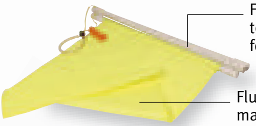

Folding or telescopic handle for easy storage

Fluorescent for maximum visibility

STROBE LIGHT Strobes are particularly effective when attached to the top of a folding flag. If attached to your BC, the flash may only be intermittently visible as waves lap over and around it.

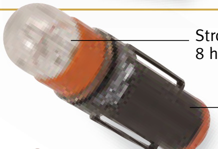

Strobe gives up to 8 hours constant use

Battery housed in waterproof casing

DISTRESS FLARES Aerial flares do not pinpoint your position, so they are best used in conjunction with other devices. They can be difficult to operate when wearing gloves, or with cold hands.

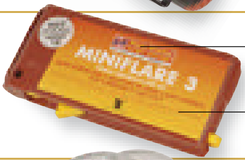

Unbreakable plastic box

Box houses firing mechanism

PERSONAL LOCATOR BEACON (PLB) Devices that broadcast radio distress signals can be used to alert surface and air searchers to your presence. Some are also equipped with bright lights for close-range visual detection.

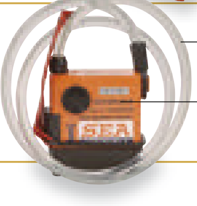

PLB antenna flashes to aid location

Signal from beacon has range of at least 35 miles (55 km) attention if the boat is not too far away. You should always carry a flashlight, since the beam can be easily spotted at night, especially when moved slowly in a scanning motion, both vertically and horizontally (*see pp.184–85*). Make sure you also carry a backup light. Strobe lights are even more effective, since their flashing light travels in all directions, eliminating the need for scanning. Flashlights and strobes must be robust and reliable; regular inspection and maintenance is vital, with special attention to sealing surfaces and O-rings.

Reflective signals are useful in bright weather. It is worth carrying an old CD or mirror to use as a reflective device to attract attention. Reflective patches on your hood and upper body may be visible to aircraft searching overhead. Dye markers stain the water and can help a search from the air, but should only be used once the aircraft is overhead, since the dye rapidly dissipates.

If you find yourself on the surface with no signaling devices whatsoever, waving your arms or even your fins and shouting may help to alert attention if your boat cover is nearby. Keep calm and persist with signaling.

INTEGRATED SAFETY DEVICES

Some safety devices can be integrated into existing gear. Air-horns, for example, attach to the low-pressure hose on your BC. These are far louder than whistles, and are useful if you are out of breath or under stress. Safety sausages can also be inflated with your BC.

BLACK DIAMOND

Standard BC

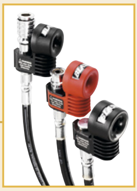

Air horn attachments

RESCUE FLAGS
A brightly colored flag is a cheap, simple-to-carry, easy-to-deploy, and highly visible location device.

Knives

A dive knife is essential for freeing yourself if you get snagged on fishing lines and other obstacles. The simplest dive knife—little more than a basic blade in a plastic mount—will get you out of most tangles, but a good one will prove an invaluable multipurpose tool.

SELECTING A KNIFE

Knives must be rustproof, but the better the rust-proofing of stainless steel, the harder it is to maintain a sharp edge. Titanium knives, while expensive, are light, rust-resistant, and stay sharp. Knives come in various sizes. Small ones can be mounted on your BC or belt, where they are easy to reach and out of the way.

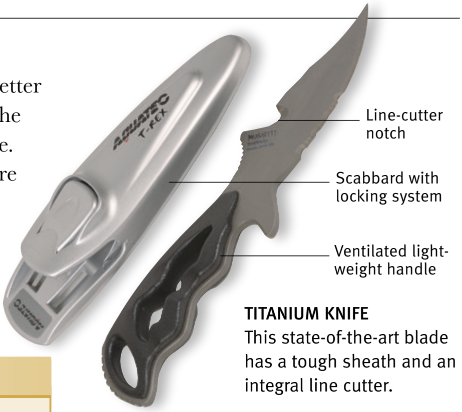

TITANIUM KNIFE
This state-of-the-art blade has a tough sheath and an integral line cutter.

Standard knives with 4½ in (11.5 cm) blades are most popular. Blades may be serrated for sawing rope or smooth for slicing monofilament line, and may have a line-cutter notch. Blunt-ended blades are ideal for tasks that need leverage, such as prying open containers. It is worth choosing a knife with a metal pommel on the end, which can be used as a hammer.

SHEARS AND MULTI-TOOLS

Multi-tools incorporate both types of knife blades, for slicing and sawing—useful for rope and monofilament line. Underwater shears are only really needed if diving in very hazardous conditions or for use as tools in salvage.

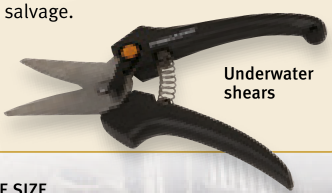

KNIFE SIZE
Large knives are popular, but a smaller blade can be just as useful on a dive.

Bags

Whenever you move to and from a dive location, you will need a bag to carry all your gear. Ideally, your chosen dive bag should be capacious and light, but it must also be durable enough to survive the rigors of transporting heavy equipment.

BUYING A BAG

When choosing a bag, check that it has strong seams, sturdy handles, and no parts that are likely to rust. Also consider security, especially when travel will involve your being separated from your baggage. Thin bags can be cut open, so a padlock will not guarantee that the contents stay safe. You may want to invest in a rigid or semi-rigid bag (*see p.238*).

Many divers use a large duffel bag for travel, and a smaller day bag (or even two) for dive days. Duffel bags range in size from 6,000 to 10,000 cu in (100 to 180 liters), day bags from 1,800 to 5,000 cu in (30 to 80 liters). There are countless “extra” bags, such as fin bags, mask bags, tool pouches, and pony-bottle pouches. While useful, most of these are not essential if you pack your gear carefully into one large duffel bag.

FLIGHTS TO FAR-AWAY LOCATIONS
Dive baggage has to be hard-wearing and tough, especially as several transfers are often needed to reach a dive location.

BAGS FOR DIVERS

MESH BAG Perforated bags can be used to rinse all your gear at once in fresh water, or to carry small items aboard a boat without fear of water ruining the bag.

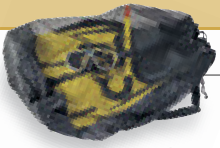

Water-permeable mesh fabric

DUFFEL BAG These can accommodate most of your gear, and also have various compartments where you can stow small items, such as slates and carabiners.

Rustproof plastic zippers

REGULATOR BAG Your regulator is valuable, and needs its own padded bag. It can be put in your main bag, but when flying, most divers stow it safely in their carry-on.

Protective padded sides

DRY BAG When you reach the dive site, a watertight bag or case is essential to protect clothing and electrical equipment, such as cameras, from moisture and salt.

Aperture can be sealed

TRAVEL CASE A secure, rigid-sided case gives extra protection to your gear while in transit. Wheels and a telescoping handle make transportation easier.

Outer pockets for nonvaluable items

Gadgets and accessories

There are a host of diving gadgets on the market, including fun items such as underwater MP3 players. Other accessories, however, have more practical uses. Clips and retractors are very useful for securing equipment, and may save you the loss of a camera or flashlight.

DIVER PROPULSION DEVICES

Underwater scooters, usually referred to as DPVs (diver propulsion vehicles), make great toys, but they can be practical and useful, too—for example, when exploring a large wreck site. There are two basic types: those that you sit astride, and those you hold on to, which tow you along. It is safest to use DPVs only to travel on a horizontal plane. Descending or ascending under power can be very dangerous, due to rapid pressure changes. DPVs are not cheap, but there are less expensive models suitable for general recreational use—that is, where failure of the device would hardly constitute an emergency. More expensive but tougher, more reliable machines are generally not aimed at the leisure diver but intended for expeditions, such as cave explorations, where failure of the unit mid-dive would be a more serious development.

Jetboots, made by a California diving company, also propel the diver along effortlessly: they comprise a pair of propellers that attach to your legs, controlled by a switch at your waist. Jetboots can propel you as fast as some DPVs, and leave your hands free.

DIVER USING SCOOTER
A diver propulsion vehicle, or DPV, takes the effort out of exploring large dive sites.

BANDS, CLIPS, AND FASTENINGS

Cam bands—webbing straps that hold your cylinder to your BC—are available with a number of different buckle mechanisms that ensure that the tank is secure and held close to your body. Retractors are very useful for attaching items, such as a flashlight or slate, to your BC because they pull the item close to your body when not in use. Carabiners and bolt-snap clips (as used on dog leashes) are useful for attaching equipment securely to your BC during the course of a dive.

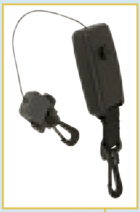

Retractor

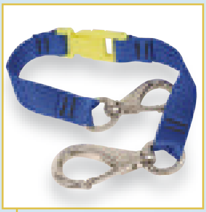

Spring-loaded snap clips

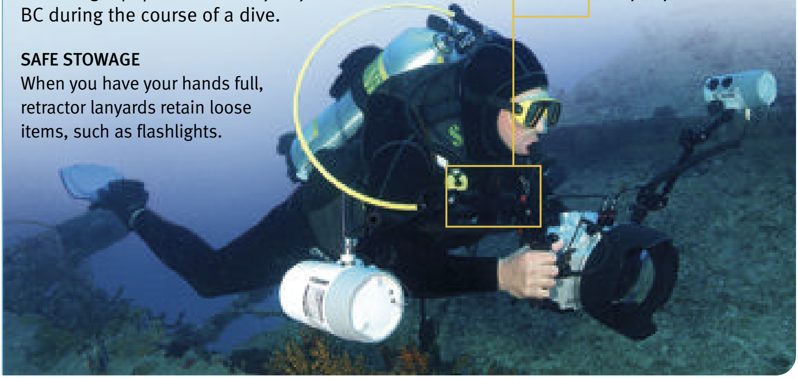

SAFE STOWAGE
When you have your hands full, retractor lanyards retain loose items, such as flashlights.

ATTRACTING ATTENTION

There is a range of gadgets that you can use to attract your buddy's attention underwater. Some BCs have a buzzer that uses a small amount of compressed air to generate a sound. Slightly more low-tech are tank pingers—a piece of elastic with a hard plastic ball attached. This is slipped around the cylinder and can be stretched and released to "ping" against it, creating a metallic sound that travels well underwater. Shakers are tubes filled with metal balls. When agitated, these create a clearly audible rattle.

WRITING UNDERWATER

Slates are handy for mapping dive sites or making notes on any unusual species encountered on your dive. They are made of white plastic, with a matte surface that can be written on with a pencil. Flat slates can be attached to a BC or slipped into a pocket; wrist-mounted slates are contoured to fit around your arm. The latter are especially useful when making technical dives with various decompression stops: you can read your computer and the list of stops from the same arm. A compass can be mounted on a flat slate, enabling you to look at navigational notes while checking your bearings.

AIR-LOCK TANK CAM BAND
Some dive gadgets are labor-saving devices. Tank band buckles that can be tightened using compressed air are a useful example.

New technology

In the 1950s, development of the scuba unit revolutionized diving, and its technology continues to develop apace. Modern equipment is sleek, lightweight, and colorful, and advances in the efficiency of rebreather systems are transforming diving once again.

MASKS AND VEHICLES

One of the most notable recent advances in modern dive equipment is the use of head-up displays in masks, which are similar to those used by military pilots. Essential dive information is sent via radio frequency from the tank to the mask, where it is presented on an LCD. Masks with an integral radio communication system, which allows divers to talk directly to one another, are also available.

Advanced synthetic materials have made dive equipment stronger and lighter, and have been used in the development of diver propulsion vehicles (DPVs; *see p.80*) that are smaller and easier to handle.

MASK TECHNOLOGY New mask developments include head-up displays of critical information (*top*) and masks that correct optical distortion underwater (*left*).

FROM SCUBA TO REBREATHER

In the last decade, rebreather technology has become available to the recreational diver, representing the most dramatic leap forward in diving equipment since the development of scuba itself.

In contrast to scuba—an "open circuit" system that discharges most of a diver's air supply into the water with the exhaled breath—rebreathers recycle the diver's exhaled breath, using chemicals to remove waste (mainly carbon dioxide) and allowing the diver to inhale the gas again. Each breath uses about 4 percent of the oxygen in the inhaled gas mix, so one of the main tasks of a rebreather is to maintain oxygen levels in the breathing loop, generally by adding extra oxygen.

UNDERWATER SCOOTER DPV The breathing observation bubble (BOB) is powered by an electric motor and has a top speed of 2.5 knots.

Rebreathers using pure oxygen were used by frogmen in World War II, for whom the absence of bubbles was vital to hide their presence. Dive depths were limited to 20 ft (6 m), at which point pure oxygen becomes toxic. A revolution in rebreather technology occurred in the mid-1980s, allowing the use of nitrogen–oxygen gas mixes rather than pure oxygen. Modern rebreathers electronically monitor the concentration of oxygen in the mix and adjust it to safe levels for a given depth.

Rebreathers offer longer dive times and fewer decompression stops. They remain costly, and need careful maintenance and specialized training.

REBREATHER TECHNOLOGY

Modern rebreathers are divided into two categories—semi-closed rebreathers (SCRs) and closed-circuit rebreathers (CCRs). SCRs expel a small quantity of exhaled gas into the water, whereas CCRs only dump gas into the water on ascent. Both types clean exhaled gas in a scrubbing unit, which removes CO_2. They also ensure that the oxygen content of the inhaled gas is at a suitable level for the diver, and monitor toxicity, which the diver can check manually on a gauge, or on a computer display. Computer control has dramatically increased the efficiency and safety of rebreathing devices.

HOW A REBREATHER WORKS

A chemical in the scrubbing unit absorbs carbon dioxide from exhaled gas, while pure oxygen is introduced to the breathing loop to replace that consumed by the diver.

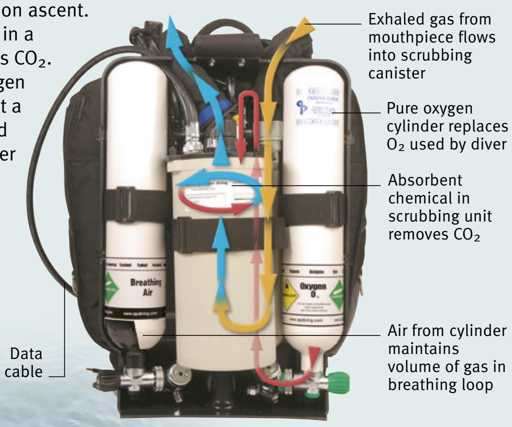
Exhaled gas from mouthpiece flows into scrubbing canister
Pure oxygen cylinder replaces O_2 used by diver
Absorbent chemical in scrubbing unit removes CO_2
Air from cylinder maintains volume of gas in breathing loop
Data cable
Breathing Air
Oxygen O_2

Oxygen flow

Inhaled gas flow

Exhaled gas flow

Mouthpiece

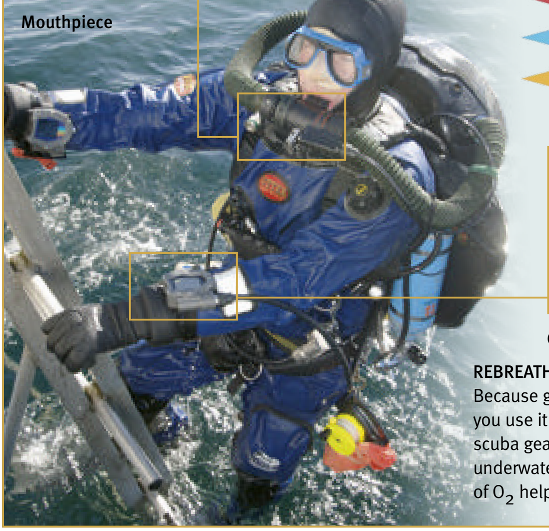

Computer display

REBREATHER BENEFITS

Because gas is recycled in a rebreather, you use it up at a slower rate than with scuba gear, and so can spend longer underwater. The consistently high levels of O_2 help reduce post-dive fatigue.

Equipment assembly

The gear assembly process is a key part of any dive. During your training, you will be taught the basics of assembling dive gear. With experience and a few simple tips, you can perfect this process and make it efficient, rapid, and—most importantly—consistently safe.

THE IMPORTANCE OF PREPARATION

With clear, blue water beckoning, the Sun shining, and the dive site awaiting you, you may find that there is a great temptation to rush the gear assembly process. However, the equipment used for diving is nothing less than life-support gear, and should be treated with a corresponding level of respect. It is reasonable to say that putting dive gear together safely before jumping into the sea is as important as packing a parachute correctly before jumping out of a plane.

CHECKING FOR DAMAGE

The first stage in assembling your gear is to carry out a visual inspection. This is particularly important if the equipment is unfamiliar or rented. It is essential to examine the test dates (*see p.68*) for the

1 **Lay out equipment** on a mat or other clean, grit-free surface where you can assemble the gear easily.

2 **Wet the tank** band on your BC, if possible, since this helps improve its grip on the tank's metal surface.

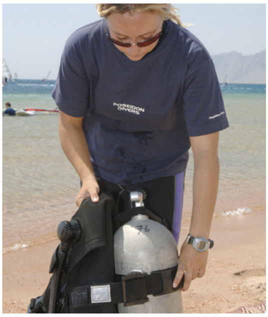

3 **Position the tank** with the pillar valve O-ring facing the backplate of the BC, then fasten the band around it.

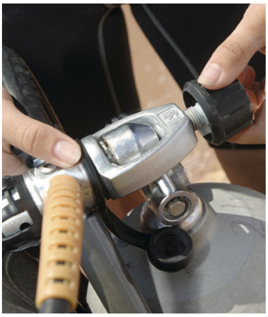

4 **Attach the regulator** first stage to the tank. Do not overtighten the clamp (or screw thread, on a DIN fitting).

5 **Pass the inflator hose** from your regulator's first stage (*see pp.60–61*) through the jacket's guides and straps before connecting it to the BC's inflator valve. Make sure the valve has engaged securely with the hose coupling.

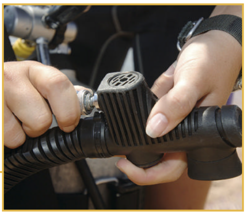

INFLATOR VALVE
The spring-loaded collar on the end of the hose fits on to the inflator valve.

tank and also the integrity of the O-ring (*see p.88*). Any nicks or cuts in the O-ring may lead to serious problems during the dive, yet replacing a damaged O-ring before the dive takes only seconds. Also look for wear and tear on the regulator hoses, particularly where they join the ports of the first stage. This is a common point of strain for regulator hoses, resulting in cracks and splits in the hose with sustained use. If the integrity of a hose has degraded, it may rupture when the air supply is turned on.

Check for signs of damage to the BC. With regular diving, there will be inevitable signs of wear and tear, but be sure to check the integrity of the buckles and load-bearing straps. These must be completely sound.

ASSEMBLING YOUR GEAR

When you are satisfied that your equipment is in good order, follow the steps below to make it ready. Take particular care to ensure that the band securing the tank to the BC is neither too high (resulting in the tank possibly coming free), or too low (resulting in the pillar valve restricting movement of your head during the dive, or actually striking it on entry).

6 **Turn on the air** supply, opening the tank's pillar valve all the way. Then turn the knob a quarter-turn back.

7 **Test air supply** by breathing from both your primary and secondary air sources. Check air quality and delivery.

8 **Secure the gear**, if you are not diving immediately, by tucking any loose parts inside the BC for protection. Most breakages occur in transit.

Disassembly

The end of a dive is an exciting time, and neglecting your dive gear at this stage is an easy trap to fall into. Take a little time, however, to clean and store your equipment properly: not only will it last much longer, but everything will be in order at the start of your next dive.

DEVELOPING A ROUTINE

As with so many things in diving, routine is the key. Many experienced divers attend to their gear the moment the dive is finished, and regard the dismantling, rinsing, and storage of their gear as an integral part of the dive itself. This is particularly true for diving from a boat, where a shortage of space frequently leads to cluttered decks and broken gear.

Diving equipment can cost a considerable amount of money, so it is worth taking a few moments to make sure it is properly dismantled and safely stowed.

The first stage of disassembly is to turn the tank off and purge any remaining air from the hoses using the purge button on the second stage (*see also pp.60–63*). The regulator first stage may then be removed from the pillar

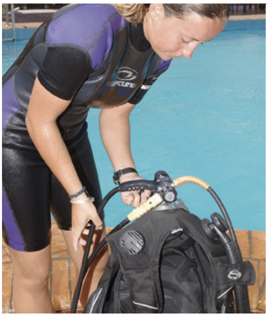

1 **Shut off the gas** supply, turning the knob on the tank's pillar valve all the way off before disassembly.

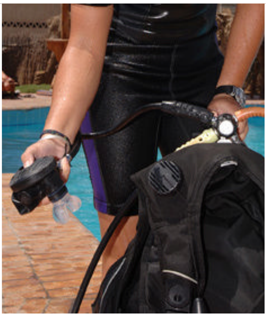

2 **Purge gas from** the system using the purge button on the regulator to vent any air remaining in the hoses.

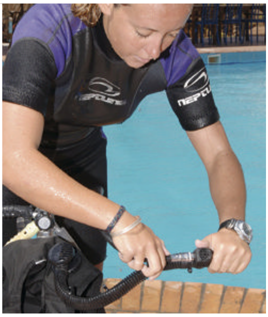

3 **Disconnect the inflator** hose from your BC's inflator valve, and the hose from the drysuit inflator, if worn.

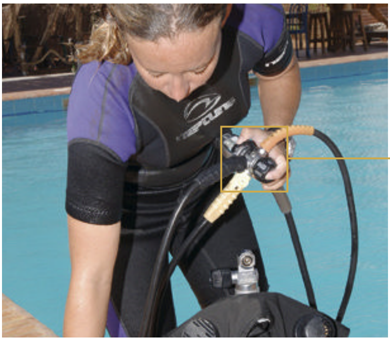

4 **Disengage the regulator first stage** from the pillar valve of the tank, and free the hoses from the BC by undoing all the clips and straps.

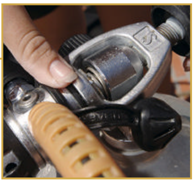

A-CLAMP FIRST STAGE Remove this type of first stage by loosening the clamp first.

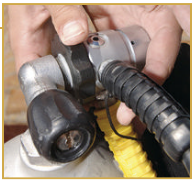

DIN-TYPE FIRST STAGE Turn the knurled collar to unscrew the regulator first stage from the pillar valve.

valve of the tank. It should take no more than finger pressure to turn the clamp or collar; any more resistance indicates that compressed air remains in the system. You will need to check that the tank is completely turned off, then purge the hoses again.

After removing the regulator first stage from the tank, use a blast of air from the tank to dry the dust cap and filter. Fit the dust cap securely, then rinse the regulator thoroughly in fresh water. Never forget to put the dust cap on, otherwise water will enter the delicate mechanism of the first stage: this is to be avoided at all costs. Hang the regulator up to dry, ensuring that the hanging method does not put strain on the hoses at any of the points of attachment to the first stage. You can then detach your BC from the cylinder and drain it of any saltwater. Rinse the BC inside and out with fresh water, and inflate it orally to allow quick drying and prevent the formation of salt crystals inside.

Dive gear must be dried thoroughly between dives, ideally by hanging it up out of direct sunlight. Do not store it where it is compressed or where moisture may accumulate, such as in a boat deck locker or a diving bag.

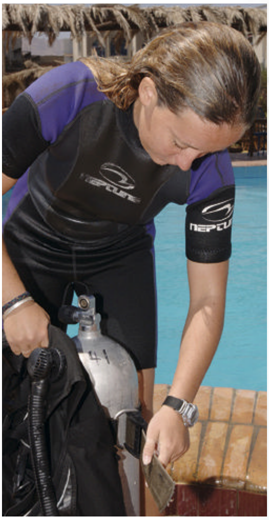

5 **Remove the BC** from the cylinder by undoing the cam buckle on the strap, and then sliding the BC upward until it is clear.

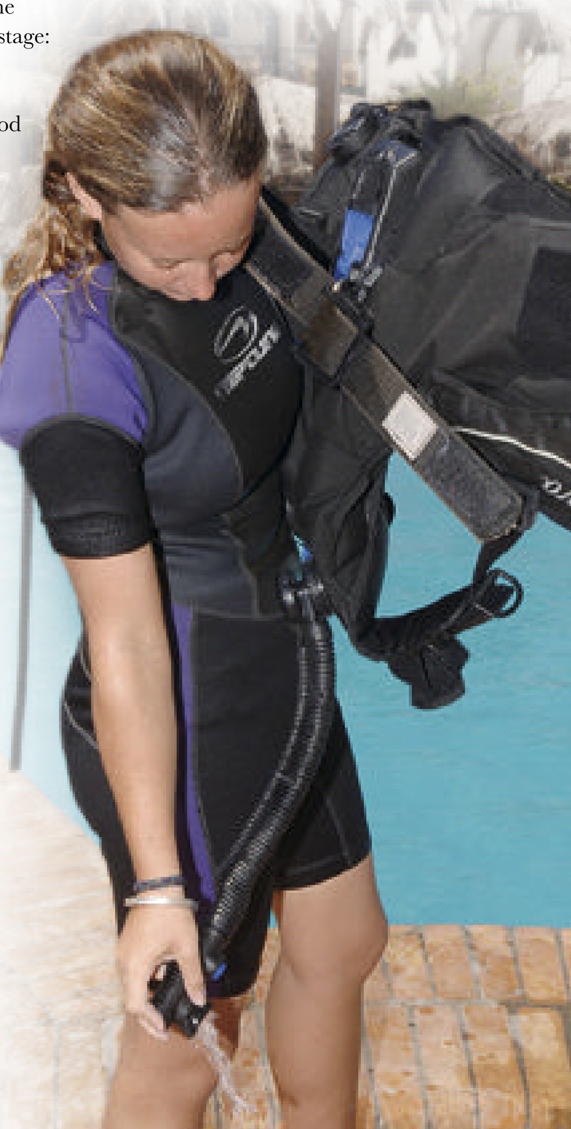

6 **Drain the BC** of saltwater by depressing the deflator button at the end of the corrugated hose. This allows the water to escape through the oral inflation point.

Equipment maintenance

Diving equipment is your life-support system underwater, yet many divers take a worryingly relaxed approach to maintaining their gear. A few simple steps will increase the life span of your dive gear, avoid the expense of constant replacements, and increase your safety.

TOOL KITS

Most divers carry a set of Allen keys, screwdrivers, or adjustable wrenches. Scuba multi-tools are all-in-one alternatives that incorporate virtually every basic tool that a diver might need on a dive boat or at a dive site. These include a range of hexagonal Allen keys designed to fit dive equipment and standard wrenches for regulator adjustment, as well as screwdrivers and a marlin spike.

Diving shops also sell “save-a-dive” kits, which consist of a set of lubricants, parts, and tools that can be used in the majority of equipment-related incidents on site. O-ring replacement and adjustments to hose configuration, for example, are simple, commonly required tasks that can be undertaken by anyone with such a kit.

Experienced divers also carry spare parts with them. This is sensible diving practice, and while the range of items differs from diver to diver, it usually includes items such as spare O-rings, silicone grease, tape, adhesive, an O-ring extractor, spare plugs for the regulator first stage, and any spare buckles or straps that might be necessary.

SCUBA MULTI-TOOL
A useful gift for any diver, this contains tools for most small maintenance jobs.

O-RING MAINTENANCE

The most disappointing reason for having to cancel a dive is the lack of a spare O-ring for a cylinder. O-ring malfunction may be indicated by air escaping once the regulator is fitted, but even in the absence of this you should make regular inspections to check that the O-ring has not degraded. In many cases, simply cleaning the O-ring and the valve groove with a lint-free towel will solve the problem. Pull the O-ring through your fingers—if you can feel cracks, it should be replaced. When removing an O-ring, never use a sharp tool, as this is likely to damage the ring.

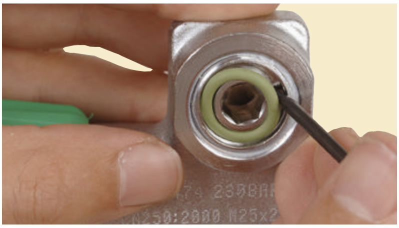

1 **To remove the O-ring** from a cylinder’s pillar valve (*see pp.68–69*), use a specially designed extractor tool. Alternatively, use the rounded end of a tie-wrap or even a credit card.

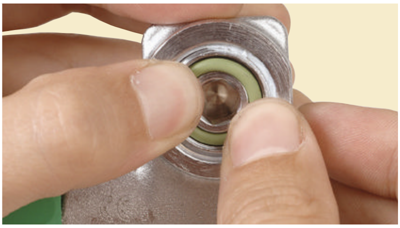

2 **Inspect the O-ring closely** and replace it if any cracks or tears are visible. Clean the valve groove with a lint-free towel, then press the new O-ring in carefully with your fingertips.

ADDING A REGULATOR HOSE

To add a hose to your regulator or change an existing hose is a simple process. You must avoid attaching a low-pressure hose to the high-pressure port, as this will rupture the hose. Most regulators now make this impossible by having different- sized ports, although older models may not have this feature. The high-pressure port is for the pressure gauge, while other ports are for a second stage or the inflator hoses. It is worth taking time to ensure that the hose configuration is practical and comfortable, to avoid strain on the base of the hoses and reduce drag.

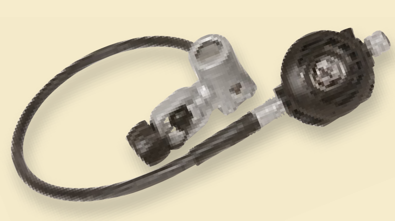

1 **When adding a hose**, be careful to select the correct port (high/low pressure). If the ports are not clearly differentiated, or if you are in any doubt, consult the manufacturer's manual.

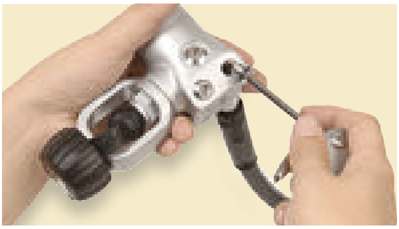

2 **Remove the plug** from the regulator port using the correct-sized Allen key. Some regulators still have raised bolts—these can be loosened with a wrench.

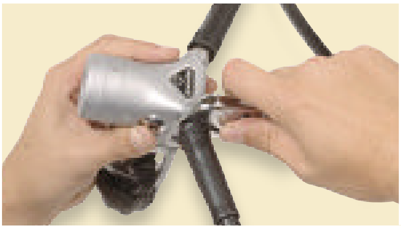

3 **Attach the hose** by simply screwing it into place. Tighten the nut with a wrench of the appropriate size—hose nuts are soft and can easily be damaged by pliers.

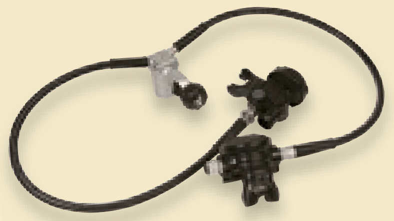

4 **Exactly the same** procedure can be used to attach further hoses to the other ports, and to replace damaged hoses. Always check hoses for cracks, fraying, and abrasion.

EQUIPMENT MAINTENANCE
Post-dive, wash all of your gear in clean, fresh water, and check out any problems you may have noticed.

The diving environment

Introduction

When you go diving, you are introducing your body to an alien environment. Evolution has not equipped our bodies for the underwater world, and you will encounter unusual sensory phenomena when you descend into the depths for the first time.

LIGHT AND VISION

Our eyes are designed to focus in air; underwater, our vision blurs unless we wear a mask, which puts an air space in front of our eyes. However, light travels more slowly in water than it does in air, and as it passes through the mask from water to air it bends, causing refraction—an optical effect that makes objects appear bigger and closer than they are.

Colors also look different to us as depth increases (*see p.183*). Objects near the surface look as colorful as they would above water, but as you go deeper, first reds disappear, then oranges, yellows, greens, and blues, until all light is lost. Colors can be revived by illuminating a submerged object with a flashlight.

SOUND AND HEARING

Above the surface, sound travels at approx 1,150 ft (350 m) per second, but underwater it travels four times faster. This makes it hard to identify the direction of a sound source: if your buddy signals to you by tapping his tank, you will hear the sound easily but may not be able to pinpoint where it is coming from. The range of audible sounds is also reduced, but the underwater environment is not as quiet as is popularly supposed: you can hear the scratching and snapping sounds made by fish and crustaceans, along with the sound of your breathing and any

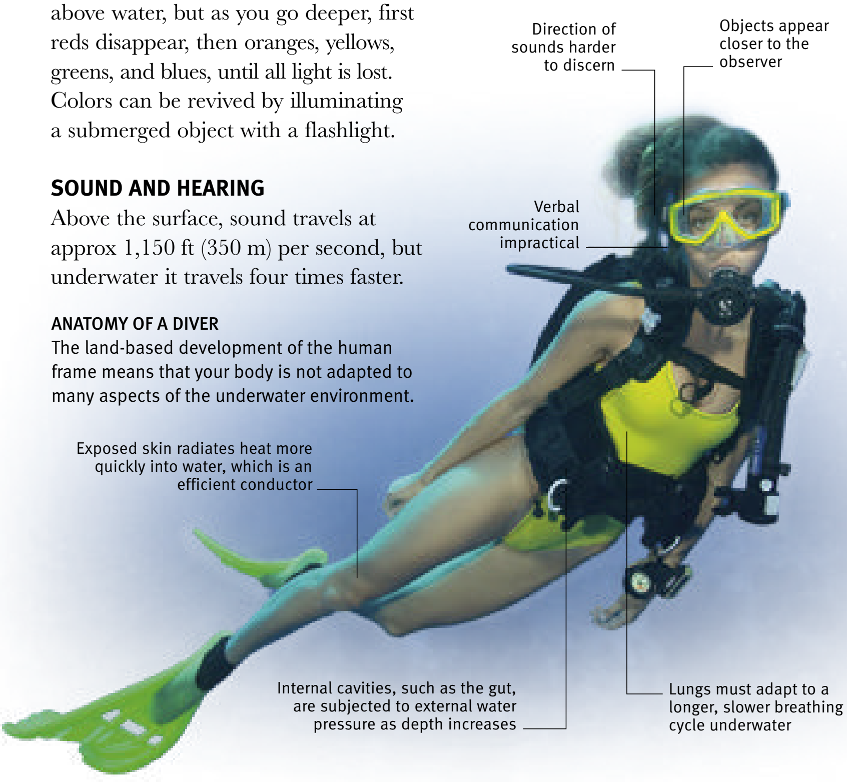

ANATOMY OF A DIVER
The land-based development of the human frame means that your body is not adapted to many aspects of the underwater environment.

Direction of sounds harder to discern
Objects appear closer to the observer
Verbal communication impractical
Exposed skin radiates heat more quickly into water, which is an efficient conductor
Internal cavities, such as the gut, are subjected to external water pressure as depth increases
Lungs must adapt to a longer, slower breathing cycle underwater

boat noise close by. However, you will find it difficult to communicate verbally underwater, even if you remove your regulator and shout, which is why divers use hand signals (*see pp.144–45*).

Sounds made above water do not easily penetrate below the surface, so any noises made to alert divers must be made underwater. A typical recall signal from a dive boat is to bang on the metal ladder on the craft's stern, which, as its lower part is in the water, transmits sound to divers beneath.

SEA CHANGE
Divers patrol a reef wall in the Red Sea. As depth increases, physiological effects become more profound and numerous.

HEAT AND PRESSURE

Water conducts heat 25 times more efficiently than air, so the human body loses heat quickly underwater. In extreme cases, a diver may lose heat faster than it can be generated by their body, leading to hypothermia. When diving, it is important to always wear thermal protection appropriate to the conditions.

The airspaces in the body—not just the lungs, but the inner ear, sinuses, and gut—will be increasingly affected by pressure as you make your descent. This can lead to uncomfortable sensations (*see pp.98–99*) that must be relieved by equalizing internal and external pressure manually using the Valsalva maneuver.

Depth and pressure

To understand the principles of diving, it helps to have a grasp of basic physical laws. When you dive, water exerts pressure on your body, and this pressure increases with depth. This simple fact influences all aspects of diving, from buoyancy control to safety considerations.

UNDERSTANDING PRESSURE

On land, your body is constantly subjected to external pressure from the weight of all the gases in the atmosphere. This is called atmospheric pressure, and it is measured in pounds per square inch (psi) or bar. Atmospheric pressure at sea level is 14.7 psi (1 bar)—that is, 14½ lb pressing on every square inch (1 kg per square centimeter) of your body's surface. This pressure is not fixed; it varies slightly with weather conditions and decreases with altitude. At the summit of Mount Everest (29,028 ft/8,848 m), atmospheric pressure is only 4.5 psi (0.31 bar).

When you dive underwater, pressure is also exerted on your body by the weight of the water above and around you. Because water is denser than air, the pressure generated by just 33 ft (10 m) of seawater is equal to that generated by all the gases in the atmosphere above sea level. Seawater is denser than fresh water, because it contains dissolved salts and minerals. This increased density means that pressure increases slightly faster with depth in seawater than in fresh water.

When you dive, you are subject to the cumulative weight of both atmospheric pressure and water pressure. The combined total is known as absolute pressure. Your depth gauge translates pressure into an accurate depth reading, but it is calibrated to ignore atmospheric pressure, so it indicates zero depth when you are at the surface. This is known as gauge pressure. Always use absolute pressure to calculate the total pressure at a given depth.

HOW PRESSURE RELATES TO DEPTH

Because water is much denser than air, pressure increases quickly with depth underwater. Water cannot be compressed and transmits pressure freely, so pressure increases cumulatively, and at a constant rate, as you dive deeper. At a depth of 33 ft (10 m) of seawater, absolute pressure is 29.4 psi (2 bar). This figure is the sum of the pressure generated by the atmosphere (14.7 psi) and by 33 ft of water (14.7 psi). At 66 ft (20 m) absolute pressure is 44.1 psi (3 bar), and so on, rising by 14.7 psi for every additional 33 ft of water depth.

If diving at a high altitude site, like a mountain lake, atmospheric pressure is reduced; consequently, absolute pressure is lower, too. You must use special dive tables at such sites.

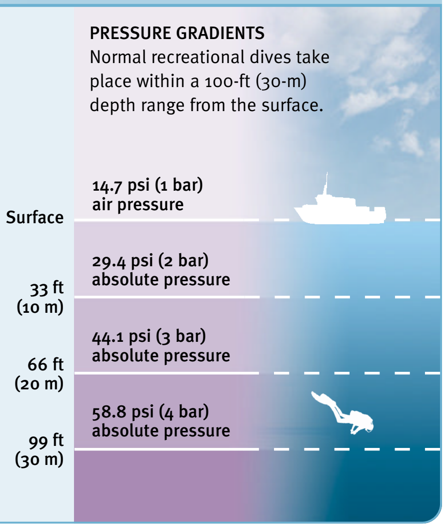

PRESSURE GRADIENTS

Normal recreational dives take place within a 100-ft (30-m) depth range from the surface.

HOW PRESSURE AFFECTS GASES

In the 17th century, scientist Robert Boyle made an important discovery. He found that at a constant temperature, the volume of a gas is inversely proportional to the pressure exerted on it—meaning that as pressure increases, the volume that a gas occupies decreases. This relationship is known as Boyle's law.

Boyle's discovery has important implications for divers. Buoyancy devices such as BCs contain bladders filled with air. These provide upward lift to help you float weightlessly underwater. As you descend, however, increasing water pressure compresses the air inside the BC. This means it displaces a smaller volume of water, and so generates less buoyancy (*see p.97*), causing you to sink. You must then inject more air into the BC to boost volume to its former level. On ascent, this must be released again, because as pressure decreases, the extra air inside the BC expands, causing you to become very buoyant and rise too quickly. For the same reason, you should never hold your breath underwater. Compressed air inhaled at depth expands as you ascend, and if it is not exhaled, it can burst delicate vessels in your lungs.

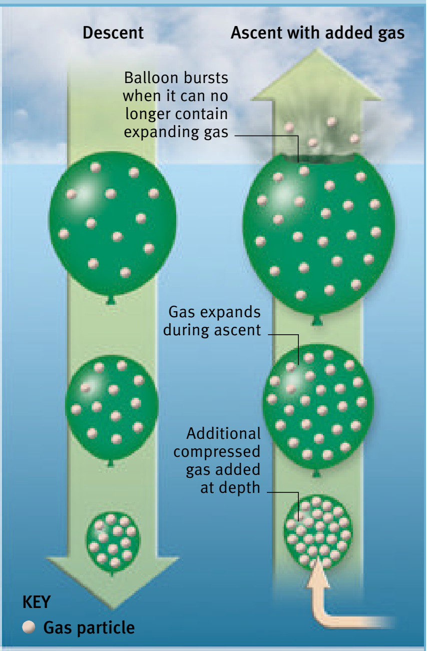
Descent
Ascent with added gas
Balloon bursts when it can no longer contain expanding gas
Gas expands during ascent
Additional compressed gas added at depth
KEY
● Gas particle

If you took a volume of gas, like that contained within a balloon, and submerged it, the space the gas occupied would get smaller with increasing depth.

If you added extra gas to the same balloon at depth, it would expand slightly. During ascent, the gas inside would expand to more than the balloon's original volume.

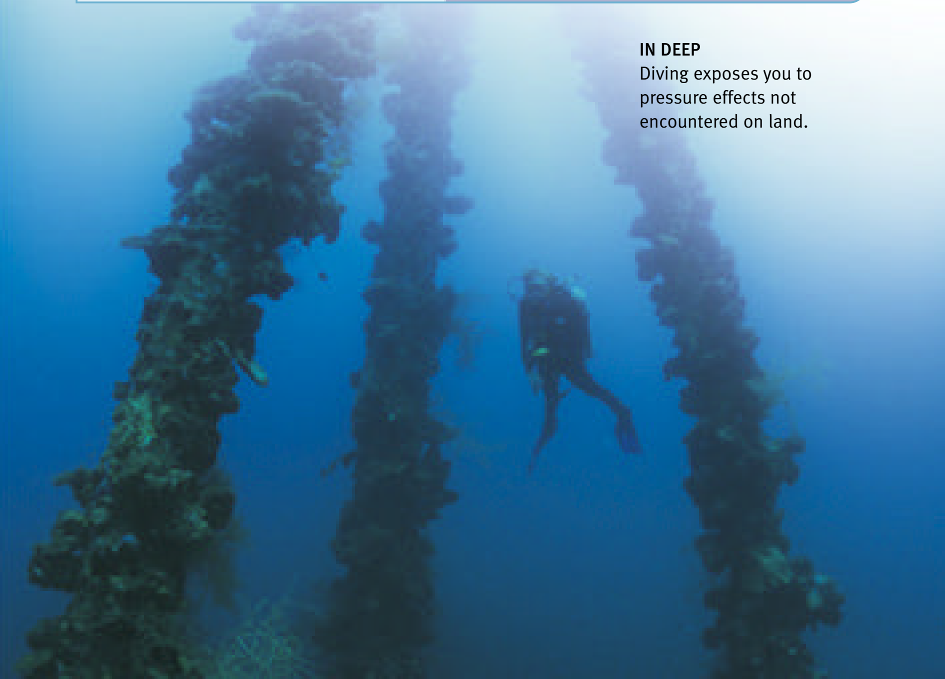

IN DEEP
Diving exposes you to pressure effects not encountered on land.

Breathing compressed gas

On land, we seldom need to worry about the availability of air, its composition, or the pressure it is supplied at. But when you venture underwater, things are very different. Being aware of how the diving environment affects your respiration cycle is of prime importance.

BREATHING UNDERWATER

Scuba divers draw breath from a tank of highly pressurized air, via a regulator. The regulator (*see pp.60–61*) reduces the pressure of the air in the tank to that of the water around you (ambient pressure), allowing it to be inhaled safely. As depth increases, so does ambient pressure, meaning that the regulator must adjust the pressure of the air it delivers to match.

Air, like any gas, is compressible, and becomes denser when pressurized. This means that at depth you will be inhaling more air per breath than you would at the surface (because the gas molecules are packed in more tightly). Because air is mainly composed of oxygen and nitrogen, this means you will be breathing higher concentrations of both gases than you would be at the surface, though the percentage of each in the mixture does not change. Both of these gases can have toxic effects on the body when pressurized. Nitrogen enters your body's tissues, and can have a narcotic effect (*see pp.100–01*). Oxygen is vital for supporting life, but becomes toxic if inhaled in sufficient concentrations. As the pressure of the air you are breathing rises, there are more and more molecules of oxygen packed into each breath you take. Beyond a depth of 200 ft (60 m), there is a risk that the concentration of oxygen in pressurized air could have a

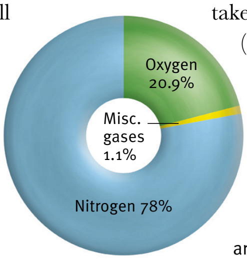

WHAT AIR IS MADE OF
Air is mainly composed of nitrogen, with oxygen forming nearly 21 percent. Inert gases like argon and freon form the remainder.

DEPTH AND AIR CONSUMPTION

Most divers consume approximately 1 cubic foot of air per minute at the surface, though this rate varies between individuals, and is also dependent on how much you exert yourself, and what your mental state is. An anxious or excited diver will consume more air than a calm one.

Because your regulator supplies you with denser air as you go deeper, your air consumption will also increase as you descend. In seawater, your Surface Consumption Rate (SCR) will have doubled by the time you reach a depth of 33 ft (10 m), tripled by 66 ft (20 m), and so on. There are various methods of obtaining an accurate calculation of your SCR, which you will find in your diving manual. Once you know this you will be able to establish how long the air in different sizes of tank (filled to maximum operating pressure) will last at different depths, and plan your dive accordingly.

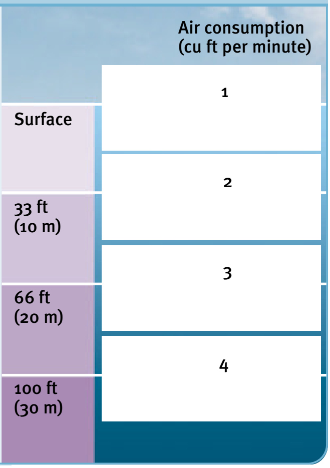

toxic effect on the body, leading to convulsions, loss of consciousness, and even death, in extreme cases.

To get around these problems, “technical” divers going beyond the limits of normal recreational diving (*see pp.194–95*) use special gas mixes. These contain altered percentages of oxygen and nitrogen, and also include helium. These mixes are used during dives in excess of 165 ft (50 m).

WASTE GASES

One by-product of the respiration cycle is carbon dioxide. The more you exert yourself, the more carbon dioxide is produced. Since your body’s breathing reflex is triggered by a build-up of carbon dioxide in the blood, not by falling oxygen levels, overexertion underwater will make you feel short of breath. On the surface, your body would respond to this by breathing faster, thus flushing out carbon dioxide. Your regulator, while efficient, cannot process the same volume of air as you can breathe on land. Always rest for a moment if you begin to feel breathless.

BREATHING AND BUOYANCY

An object immersed in water is subject to an upward thrust equal to the weight of the water it has displaced. You will only float, therefore, if your weight is less than that of the water you displace. If it is greater, you will sink; if the two are equal, you will do neither—you will have neutral buoyancy. The compressed gas you breathe underwater alters your buoyancy. When you inhale, the volume of your chest gets larger—you displace more water without becoming any heavier, and so become more buoyant.

A diver displays buoyancy control

DEEP BREATH
Slow, even breathing patterns promote low air consumption and help prevent hyperventilation.

Pressure and the body

When you dive, your body is subjected to pressure generated by the weight of the water around you. As you descend, you will feel the increase in pressure in some of your body's internal air spaces, chiefly your ears. If you do not act to correct this, you may be injured.

EQUALIZING PRESSURE

The body contains four major air spaces that are affected by pressure underwater: the ears, the sinuses, the airways and lungs, and the stomach and gut. Air spaces in our respiratory and digestive systems are compressible, in that they are contained by soft tissue that can expand or contract to accommodate pressure changes. However, your regulator (*see pp.60–63*) ensures that the air pressure in your lungs and airway is always in equilibrium with external water pressure, so pressure effects are never felt here. You should barely notice any effect on your stomach and gut, although you may break wind, especially on ascent.

The air spaces in your head, however, such as the ears and sinuses, are rigid and cannot be compressed. When the pressure outside them is greater than that within, a "squeeze" is created, which can cause pain and damage, especially to the eardrum. You must equalize the pressure in these cavities with the ambient pressure of the water around you as you descend. Equalizing is simple, using a procedure that clears the ears called the Valsalva

ROUTINE MANEUVER
Equalize pressure in your ears and sinuses every 3 ft (meter) or so, during your descent to the bottom. Don't wait until your ears begin to hurt.

THE IMPORTANCE OF EQUALIZATION

The ear is divided into the outer, middle, and inner ear. The eardrum separates the middle and inner ear from the outside world, and it is this membrane that is most vulnerable to pressure changes. If you do not equalize as you descend to raise the pressure behind your eardrum—by sending air to the ear from your throat via your eustachian tubes—then the eardrum will eventually rupture and let water into your middle ear, leading to possible infection and deafness. Your sinuses are pairs of cavities within the bones of your face. Under normal circumstances they are open, via your nasal passages, but if you have a cold, or congestion caused by allergies, they fill with mucus and become blocked. Failure to equalize will cause blood vessels in the lining of the sinuses to rupture and bleed into the sinuses, which can be quite painful.

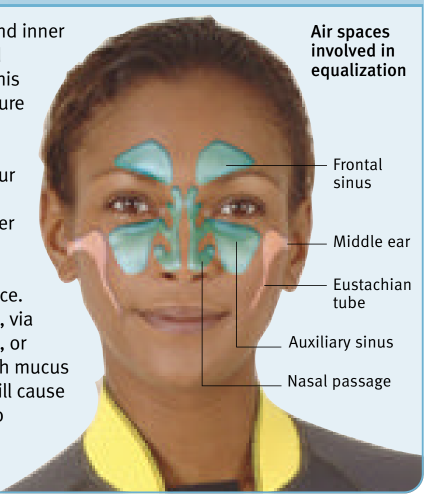

Air spaces involved in equalization

maneuver. You just have to hold your nose and breathe out against your nostrils. This will send air from your throat through the eustachian tubes into the middle ear so that the pressure there becomes equal to that of the water outside, thus easing the discomfort. This maneuver should also equalize the pressure in your sinuses, and in addition resolve any problems of dizziness or vertigo you may be experiencing (as balance is also controlled within your ears). Some people get the same result by simply swallowing repeatedly.

You need to equalize regularly on descent, but on ascent equalization should occur automatically. If you have problems equalizing on the way down, ascend a little to reduce the pressure slightly, and try again. This equalization process is vital for comfort when diving; never dive with ear plugs or with anything tight over your ears. Check that your hood does not fit too tightly, as this can hinder efficient equalization. If you have any problems with your sinuses or ears, do not attempt to dive, since this may make them worse. A cold, or any nasal congestion, may cause blockages in your airways and you will not be able to clear your ears and sinuses. If you notice a slight nosebleed after a dive, this is a common sign that you have a mild sinus blockage and you should not dive again until it has cleared up. Decongestants will not necessarily help once you are at depth, and you should not dive while you are taking them.

SAFE DESCENT
To adapt safely to the underwater world, you must understand how depth affects the human frame, especially internal cavities.

Nitrogen and decompression

When you dive using scuba gear, nitrogen gradually accumulates in your body, and is released when you ascend. In order for this to happen safely, you must plan your dive carefully and employ diving procedures that minimize the risk of decompression sickness (DCS).

GAS ABSORPTION

BOTTOM TIME
The longer you spend at depth, the more nitrogen builds up in your body.

Gases are soluble in liquids, and dissolve more easily when they are pressurized. When you breathe compressed air underwater, the two main gases it is composed of (nitrogen and oxygen) enter and dissolve in your bloodstream. This “gas exchange” takes place in the lungs, via tiny saclike vessels called alveoli.

Nitrogen is particularly easily absorbed into the blood and body tissues. Your regulator supplies denser and denser air as depth increases (*see also p.60*), which means that every breath you take is loaded with more nitrogen molecules than an equivalent breath at the surface. This surplus of nitrogen rapidly accumulates in the tissues of your body. If you ascend too quickly, the nitrogen in your tissues can reenter the blood in the form of harmful bubbles. It is therefore very important for you to manage your dive time and depth,

NITROGEN AND BODY TISSUES

When you breathe compressed air underwater, its nitrogen content enters the blood and begins to penetrate the body’s tissues. When you ascend, ambient pressure drops and nitrogen starts to come out of the tissues again. If your ascent is slow and controlled, the nitrogen reenters the blood and exits the body harmlessly, via your lungs. If your ascent is too fast, the escaping nitrogen can form harmful bubbles, leading to decompression sickness (DCS; *see pp.142–43*). DCS can cause minor discomfort if the bubbles are lodged under the skin or in the joints, but if they are in the brain or nervous system, it can lead to paralysis or even death.

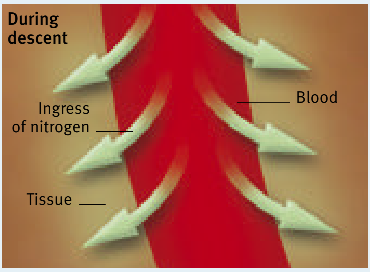

NITROGEN ENTERS BODY TISSUES
Pressurized air dissolves in blood and its nitrogen content enters tissues, especially fatty areas, where it is temporarily stored.

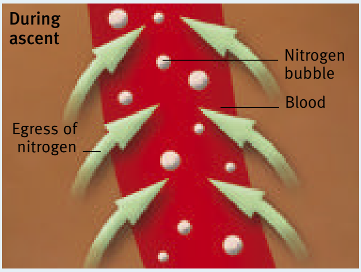

FORMATION OF NITROGEN BUBBLES
If pressure drops too quickly during ascent, nitrogen forms bubbles in the bloodstream, which can lead to decompression sickness.

PERFECT PERCH
Dive boats sometimes lower a “trapeze” for divers to hold on to while making decompression stops, with a spare air tank attached for safety.

using either dive tables or a dive computer, to ensure that nitrogen does not build up to levels that cannot be safely dissipated during ascent.

Dive tables help you to calculate how long you can safely dive at a given depth, and what the safe limits are for any following dives that day (excess nitrogen takes 24 hours to fully leave the body). The tables will also tell you if you need to perform a decompression stop. This is a scheduled pause in your ascent that allows nitrogen to dissipate safely. It is a good idea to stop at 16 ft (5 m) for three minutes during ascent, even if your dive tables do not require it, just as a precautionary measure. This is called a “safety stop.” Dive computers also generate safe diving schedules, by making continuous calculations during the dive (*see pp.66–67*).

NITROGEN NARCOSIS

The nitrogen in compressed air can also have a narcotic effect at depth, ranging from a feeling of euphoria to anxiety. Nitrogen narcosis rarely causes death directly but at depths below 30m (100ft) it causes divers to make increasingly bad judgments. If you feel the effects of narcosis during a dive, ascend to a shallower depth and the narcosis should rapidly disappear. Keep an eye on your buddy for symptoms of narcosis, such as sluggish responses and irrational behaviour.

USING DIVE TABLES
An understanding of the use of dive tables is a key element in all dive training programs.

Tides and currents

The tides have a huge impact on the marine life and conditions of many dive sites, and localized currents can create unseen offshore hazards. Knowing how tides and currents work and how they affect diving procedures is essential for safe diving at any location.

TIDAL MOVEMENT

Water levels at the shoreline can vary considerably over the tidal cycle, in extreme cases by as much as 50 ft (15 m). The region between the highest and lowest tide marks is called the intertidal zone.

Make a thorough check of local tidal conditions and coastal topography when planning a dive, otherwise you may encounter unexpected hazards, like areas of strong tidal flow. Water clarity varies from site to site, but it is often best at high tide, while incoming tides tend to bring marine species closer in to shore with them. The most important tidal event for divers is “slack water,” a short interval when tidal flow halts at the changeover of every tide. Tidal currents mean that some sites are only safe for diving during the slack water period.

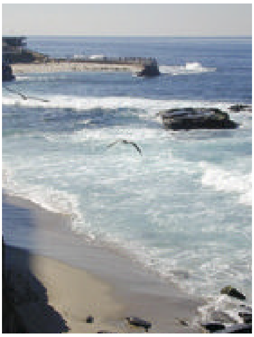

INTERTIDAL ZONE
Sometimes submerged and sometimes exposed, this is a harsh habitat for marine life.

HOW TIDES ARE FORMED

The Earth’s seas are subject to the gravitational pull of both the Sun and the Moon. Although the Moon is smaller than the Sun, it is much closer to the Earth, so its influence is stronger. The gravitational pull of the Moon draws water into a bulge on the side of the planet closest to it. A counter-bulge forms on the opposite side of the Earth because the water there is pulled less toward the Moon than the planet is itself. As the Earth rotates, the two opposing bulges of water move across the face of the planet, creating two high tides and two low tides in just over 24 hours.

NEAP TIDES: FIRST AND THIRD QUARTER MOONS
With the Sun and Moon at right angles, high tides are at their lowest, as they are subject to the gravitational pull of the Moon alone, but low tides are at their highest, pulled by the Sun.

SPRING TIDES: NEW AND FULL MOONS
Twice a month, the Sun and Moon are in line with the Earth. Their combined gravitational pull creates the highest high tides, and low tides are at their lowest.

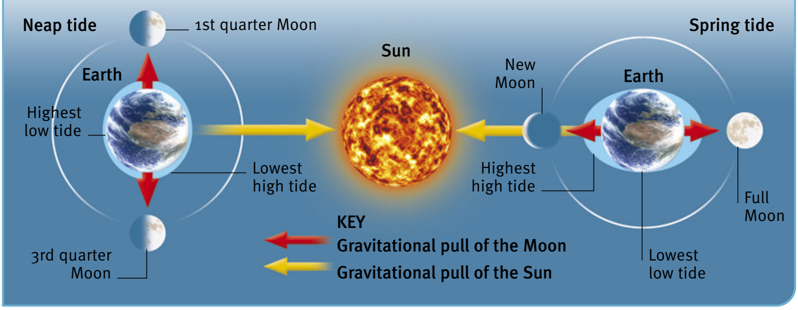

TYPES OF CURRENTS

Currents are created by winds, tides, the rotation of the Earth, and the confluence of bodies of water of differing salinity and temperature. Tidal currents occur when local topography, such as a narrow channel, creates an area of unusually strong flow when the tide is coming in (flooding) or going out (ebbing). These currents can be powerful enough to carry away even strong swimmers.

DIVERS EXITING SEA
Plan your entrance and exit so that you avoid swimming against the tide. The best time to dive is usually at "slack tide"—at high or low tide, when the tidal pull is weakest.

Standing currents are permanent currents independent of tides. Longshore currents, which flow in one direction along the coast, are localized standing currents. There are also standing currents in the open ocean, such as the Gulf Stream, which flows northward from the Gulf of Mexico, driven by the prevailing southwest winds, to the mid-North Atlantic, where it divides into two lesser currents.

Rip currents occur when incoming waves force the backwash from previous waves sideways. The backwash streams along the shore until it finds a route back to the open sea, such as a gap in a sandbar, and surges out in a strong, narrow current, perpendicular to the shore. To escape a rip current, swim across the current, parallel to the shore. You may find that trying to swim out of a rip current leads you into a longshore current, which pushes you back into the rip current. If this happens, swim out of the rip current in the opposite direction.

WAVE SURGE
Underwater topography and weather conditions can combine to create dangerously strong water movement where surf breaks on the shore.

Diving skills

Introduction

Good basic skills are key to being a competent diver. Some techniques, such as buoyancy control, may seem difficult to master at first, but with practice they will become second nature, allowing you to relax and enjoy your dives. Skilled divers are also safer divers.

SUPPLEMENTING YOUR TRAINING

This chapter of the book discusses diving skills, but is not intended to be, or in any way replace, a training manual. While it covers the skills and techniques that are taught in the basic dive training offered by approved agencies, it must not be used in place of any element of that training. There is no substitute for comprehensive dive training with a certified instructor.

In this chapter, therefore, we outline the skills you need in order to be a safe and competent diver. Many people only dive once a year on vacation, or even less frequently, so it is very important to keep up to date with your diving skills. If you have been away from diving for any length of time, take a refresher course or an easy trial dive.

STEPPING OUT
Knowing how to enter the water properly will help you get your dive off to a good start.

USING A COMPASS
Don't forget to practice navigation skills regularly. Getting lost can ruin a dive, and may hinder your return to a safe exit point.

You will find that most of your basic skills are still familiar to you after a break, or will come back in due course. However, if you are put in a stressful situation, you may not be able to react quickly and effectively if your skills are rusty. Even experienced divers can benefit from regular practice of basic skills, and you should always be thinking about how to improve your equipment configuration and technique.

KEEPING SKILLS CURRENT

A great way of keeping your skills up to date is to practice in a pool (*see pp.146–49*). This might not be as much fun as getting out into the open water, but it will certainly provide you with a safe environment in which to perfect key skills such as buoyancy. Regular training in a pool or sheltered water can help you to meet the challenges of open-water diving more confidently. If your local waters are cold and stormy in winter, use the off-season to brush up drills and procedures. Joining a diving club may give you access to a nearby pool.

DEPLOYING A BUOY
Diving requires the use of many items of special equipment that you must learn how to use safely.

Pre-dive skills

Good organizational skills will help ensure a trouble-free diving trip. A successful dive begins before you even leave your home, with checks on weather conditions and equipment. Even if someone else is organizing the dive, make sure you are well prepared and informed.

GETTING OFF TO A GOOD START

On the morning of the dive, get a detailed, up-to-date weather forecast specific to the site you are diving, and double-check tide times if appropriate. If you don't feel well for any reason, make an honest assessment of whether you are fit to dive that day. It is better to sit out a dive than to feel ill underwater and have to abort. Illnesses affecting lung or circulatory efficiency also increase the risk of decompression sickness (*see p.142*)—a very good reason for skipping the dive.

Always check your gear before you set off on a diving trip. Before you leave home, you must be sure that you have everything you need. It is good practice to make a checklist of all the items you will need, and keep this in your bag. When you arrive, be sure to transfer all your gear from the initial mustering point to the actual dive site, or onto the boat you are diving from. All too often, divers leave a vital piece of

BRIEFING FROM DIVE LEADER TO GROUP
While a group leader will take overall charge of the dive plan, you must take responsibility for your own fitness and equipment readiness.

DIVING WITH A NEW BUDDY

Diving abroad often means that you will dive with strangers. Make sure you give the dive organizer an honest assessment of your level of ability and interests, so they can pair you with a suitable buddy. Even if you don't know your buddy, give them all the respect and assistance you would offer a familiar dive partner.

gear on the dock, such as a weight belt or mask, and only realize their mistake when the dive boat reaches its destination—by which time it is too late.

DIVING FROM A BOAT

The boat should have been booked in advance, but even so, it is polite to call the skipper again before you arrive. He or she will be able to confirm that weather and conditions are favorable for diving. When you board, check that the boat has adequate first aid and safety signaling devices on board, and appears to be in a seaworthy condition. Before you leave shore, notify the coastguard or relevant authority—and do not forget to inform them of your safe return when you get back. Ensure that the skipper reviews safety procedures, including the location of lifejackets, flares, and life rafts, and demonstrates the underwater recall signal.

COMFORT AFLOAT
Remember to bring seasickness tablets with you on boat trips. Test them 24 hours beforehand to make sure they have no adverse effect on you.

GETTING READY TO DIVE
If you have organized your day efficiently, you will be able to relax and concentrate on getting yourself ready to dive.

DIVING FROM THE SHORE

If you are diving from the shore, always check for the easiest possible access to your dive site on arrival (or before) to avoid long walks down to the beach with heavy gear. You may need permission to drive to the water's edge to unload.

WATCHING THE WEATHER

Obtain a forecast that includes the wind speed and direction, sea state, and visibility (fog, mist, etc). Do not take any risks with diving in bad weather—it is better to miss the dive than dive in surging seas or in visibility so poor your boat cover will not be able to locate you. You should also check tide tables to establish the timings of slack water. This is critical when diving areas affected by strong tidal currents.

Rough seas

Final checks

Diving with a partner (or buddy) means that there is someone to help if you encounter problems underwater. For the system to work properly, however, you and your buddy need to conduct your own briefing before a dive and check that all equipment is functioning properly.

MAKING BUDDY CHECKS

Just before the start of the dive, get together with your buddy and assemble your gear (*see pp.84–85*) and put it on. There is no particular order in which to do this, but most divers find that to avoid overlooking anything, it is useful to develop a routine. When you are both suited up and before either of you enter the water, you should carry out a buddy check—sit or stand next to your buddy and carefully check each other's gear, following the sequence shown below. Do not be tempted to rush these vital checks—you may regret it later. They ensure that you know how each other's gear is assembled, how it works, and that it is functioning. They also serve as a double-check that neither of you has overlooked anything before you dive.

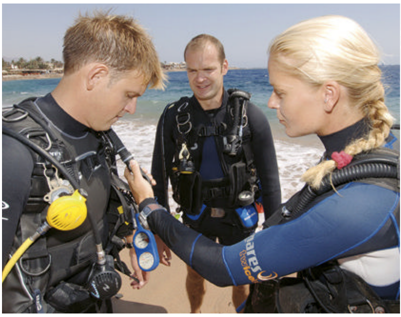

1 **Check BCs** so that you and your buddy know where each other's inflation and deflation points are, and ensure that they are working. Do the same for drysuits, if worn.

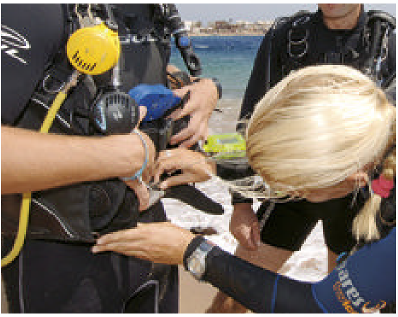

2 **Check that weights** are present and securely fastened. You and your buddy must be especially aware of how each other's weights are released, in case either diver is incapacitated.

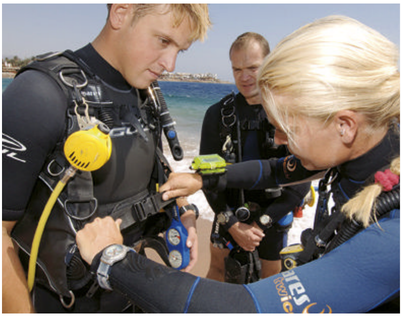

3 **Check harness is secure** and note where, on your buddy's kit, key fastening points and harness release clips are located, and how they are operated.

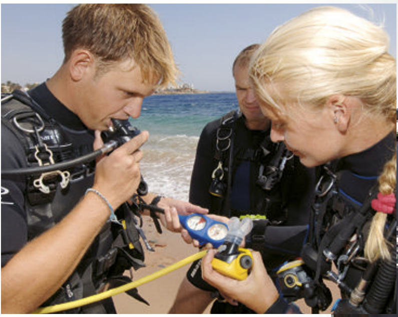

4 **Check air contents** gauges and breathe from your regulators to check they are working. Test each other's octopus second stage.

BUDDY BRIEFING

Suiting up provides a good opportunity to talk over a dive plan with your buddy. Ensure first that you are both agreed, as a pair, on the aim and course of the dive, your entry and exit points, and your predicted maximum depth and time for the dive. Check that you both have the required amount of air for your plan (including a reserve for emergencies). Agree on who will lead the dive and on whether you will dive to the left or right of your buddy. Decide on all communication signals, including how and when you will signal for the end of the dive, and on what you will do if you become separated. Once you have agreed a plan, stick to it unless it becomes impossible to do so. If circumstances change during the dive, use hand signals to discuss how, as a pair, you are going to modify the dive.

PREVENTING FOGGING

Mask fogging is a very common inconvenience, and is caused by oils on the mask's lens allowing moisture to bead. The traditional way to prevent fogging is to rub saliva onto the inside of the lens glass, then lightly rinse clear, before putting on the mask for a dive. Anti-fogging sprays are also available.

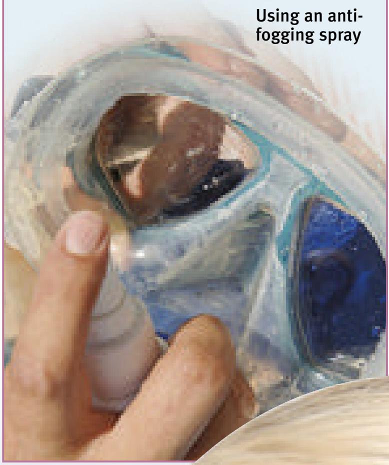

Using an anti-fogging spray

5 **Ready to dive?** Give each other a last once-over to establish who is carrying any miscellaneous pieces of gear, such as reels and slates, and where they are fastened. When you are ready to dive, make a final OK signal.

Entry

Every site will require a different method of entry, and the easiest method possible at each site is the best one to use. In all cases, it is safer to enter without large, heavy equipment, such as cameras, so if conditions allow, have them passed to you once you are in the water.

SHORE ENTRY

When entering the water from a beach, make sure your gear is assembled and on your back, then carry your mask and fins to the water's edge. There, put on your fins and mask with your buddy's help and walk backward into the water (walking forward with your fins on is difficult). As soon as you are deep enough, lower your center of gravity by bending down, then start to swim, or "fin," out into open water. If there is surf, make sure your BC is partially inflated and wait until a big wave approaches; lean into it, and let the return surge take you away from shore.

ENTERING FROM THE SHORE
If entering through surf, be careful not to lose your fins as you put them on. Breathe from your regulator throughout entry.

CONTROLLED SEATED ENTRY

This is the easiest method of entry into a pool or from a platform just above water level, and an especially useful method if you have difficulty with mobility on land or with keeping your balance when standing on a floating platform. It also enables you to put on your gear while seated.

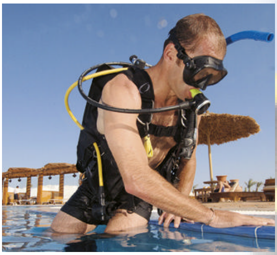

1 **Turn from the waist** to place both hands on the hard surface to one side of your body, ready to take your weight.

2 **Pivot your body around,** using your hands to support you, so that you turn to face the side while easing yourself into the water.

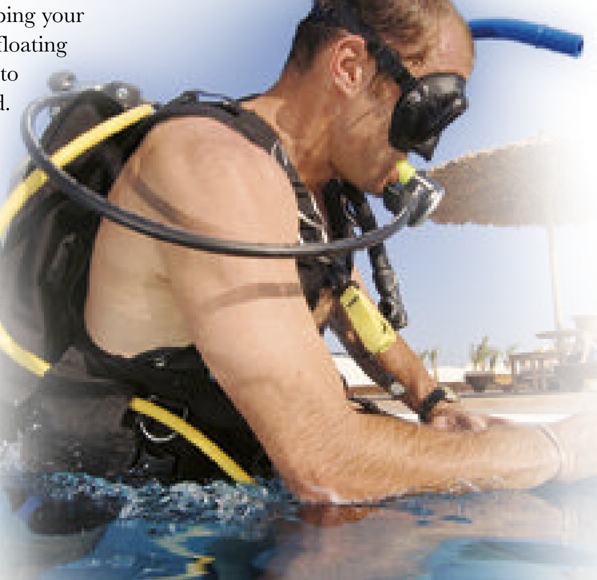

GIANT STRIDE ENTRY

Stepping into the water is a good entry method to use from a large diving boat or off a dock or platform, provided there is sufficient depth of water. Partially inflate your BC, place your regulator in your mouth and hold it and your mask in place with one hand, and take a large stride out into the water.

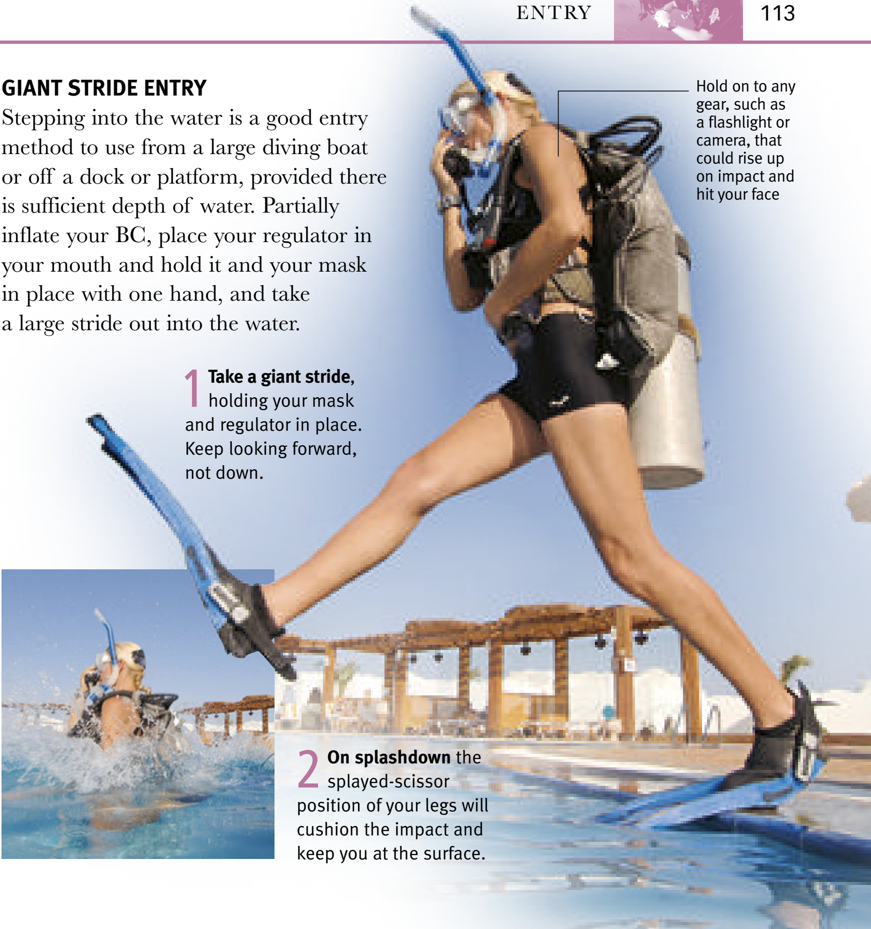

Hold on to any gear, such as a flashlight or camera, that could rise up on impact and hit your face

1 **Take a giant stride,** holding your mask and regulator in place. Keep looking forward, not down.

2 **On splashdown** the splayed-scissor position of your legs will cushion the impact and keep you at the surface.

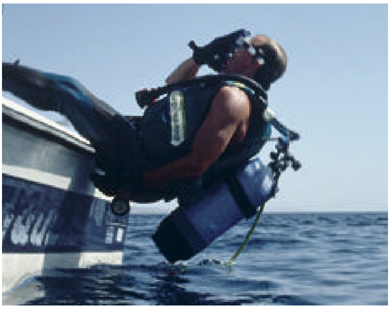

ENTRY POSITION
Hold your mask and regulator in place with one hand and ease yourself backward until your tank is hanging over the water.

BACKWARD ROLL

Use this entry from small boats with a low freeboard (distance from deck to waterline). With your regulator in place, check that your fins are free from obstruction, and that the water behind you is clear. Ease yourself over the edge and roll over backwards, aiming to hit the water first with your upper back.

BUDDY PROCEDURE AND SIGNALS

After entering the water, meet on the surface and make an OK signal (hand on head or fingertips together) to shore or boat cover. One exception to this is when making a negatively buoyant entry through strong surface currents. In this case, descend as quickly as possible, and rejoin your buddy on the seabed.

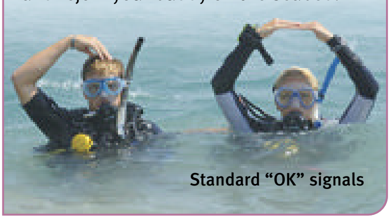

Standard “OK” signals

Descent

Descending to the seabed is always an exciting experience, but while you may want to get there quickly, you must always descend in a controlled manner. The method you use will depend partly on the site and prevailing conditions, and partly on your own preferences.

MAKING YOUR DESCENT

Having met up with your dive buddy in the water, follow the preparatory steps below. If you find that you do not sink, check that no air remains in your BC (and drysuit, if applicable). If the problem persists, you probably do not have sufficient weight on your belt (*see pp.70–71*), and should return to the boat or the shore to get some more. It is a good idea to note how much weight you need each time you try a different equipment configuration—record the details in your logbook for reference.

At a depth of 10 ft (3 m) you should carry out a bubble check. This involves briefly stopping so that you and your buddy can check each other's equipment for signs of air leaks—for example, from an incorrectly fitted hose. This will allow you to return to the surface to resolve the issue before you resume the dive. By fixing minor problems now, you may avoid bigger problems during the dive that could lead to an emergency.

As you descend, you will feel pressure in your ears. Release this regularly by swallowing or holding your nose and

1 **Meet up with your buddy** on the surface, well clear of any boat cover. If using a snorkel, remove it, and switch to breathing from primary regulators.

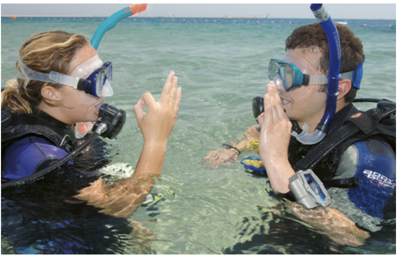

2 **Give an OK signal** to each other when ready to begin the dive. If using a "buddy line" to link yourselves together (useful when one buddy is a novice), ensure now that you are both attached.

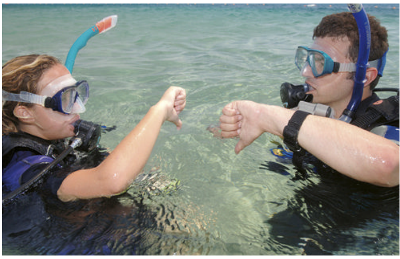

3 **The "down" signal** confirms your intention to descend immediately. If your buddy is a novice or nervous, reassure them by holding their hand and helping them to descend.

4 **Both you and your buddy** should deflate your BCs and exhale together, so that you become negatively buoyant and start to sink simultaneously. If you have a problem sinking, address it now.

BUBBLE CHECK
Buddy pairs check each other's gear for telltale air leaks before descending further.

blowing against the closed nostrils—a process called "equalization" or "ear-clearing" (*see also pp.98–99*). If you feel you are descending too fast, allow a little air into your BC. Your mask will also start to press on your face; relieve this by exhaling gently through your nose. If you are wearing a drysuit, you will feel the water pressing it against your body. Ease this by letting air into the suit, but not so much that it alters your buoyancy too greatly—and remember to release it again on ascent.

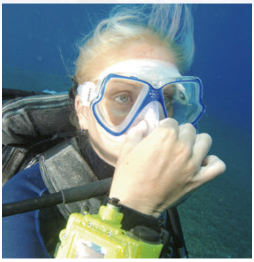

EQUALIZATION
Pinch your nose and blow at intervals to equalize pressure in the skull with that of the water.

DESCENT OPTIONS

Many divers prefer to descend feet-first so that they can remain face to face with their buddy and maintain eye contact. This is good practice, unless you need to descend quickly through a current and therefore have to use your fins.

A free descent is where no line is used and you descend freely (head- or feet-first) to the seabed. If you are using an SMB (*see pp.74–75*), ensure before descent that the buoy is fully inflated and the line is free. Your buddy should stay on the same side throughout the descent to avoid confusion and tangling in the line. On group dives there may already be a weighted guideline in place, attached to a buoy and leading to a wreck or fixed point on the seabed, which will give you and your buddy a handy point of reference. Do not use the line to pull yourself down, since it may not be anchored securely at the base.

FINAL CHECK
A group of divers meeting at the bottom of a guideline give a last "OK" signal to each other before exploring the dive area.

Breathing underwater

Swimming freely underwater for the very first time can be exhilarating, but also potentially unnerving and even frightening. Learning and perfecting good breathing techniques allows the novice diver to become calm and relaxed in this new environment.

CONSERVING AIR
The more relaxed you are, the less air you will use and the longer your dive will last.

LEARNING TO RELAX

The sheer thrill of your very first dive will cause your heart rate to increase and make you breathe faster than normal. You may be alarmed at how quickly your air supply runs out, but after a few dives, you will learn to be calmer and more relaxed, and your air consumption will decrease. When you feel comfortable underwater, take a few moments to concentrate on your breathing and remind yourself of the dos and don’ts of underwater breathing.

First, and most important, never hold your breath underwater, because you will run the risk of bursting a lung. Inhale and exhale more deeply than you would normally do on land. Try to maintain neutral buoyancy (*see pp.118–21*), since this will decrease your need for leg and arm movements to maintain a constant position, and thus reduce air consumption. Furthermore, avoid

BREATHING OUT
If you remove your regulator, exhale slowly to minimize the risk of lung expansion injury.

CLEARING THE REGULATOR
A regulator can be cleared of water by simply exhaling, or by pressing the button on the front to purge the water.

FREQUENT CHECKS
Buddies should regularly check each other's air reserves, to maintain a good awareness of the remaining levels of both members of the team.

swimming unnecessarily against currents—it is better to go with the flow than to waste your air by overexertion.

Always be aware of where your buddy is and what he or she is doing; many novice divers lose sight of their buddy and rapidly become anxious. If there is a problem, stop, breathe, and think before taking action. Acting too quickly and without thought could lead to panic and further problems.

Make sure that you are familiar with your equipment and fully understand how it works. You must know exactly where each item is stored to avoid panic if things go wrong during a dive—for example, if you lose your regulator or your mask begins to fill up. Regularly practice emergency procedure drills (*see pp.136–43*), either in the pool or in open water, to ensure confidence in a real crisis.

PLANNING YOUR AIR CONSUMPTION

On an open-water, fairly shallow dive, the rule of air consumption is to plan your dive so that you can carry out your descent, exploration, and ascent, and still surface with a reserve of 725 psi (50 bar). If you have to come back to a given point to exit, such as a guideline or anchor line, start your return when you have 1,500 psi (100 bar) left. If there is a current, swim out against it and swim back with it, so that the return is never longer than the outward trip. It is also wise to swim back to your point of entry or shotline at a shallower depth than on the outward journey, since this reduces air consumption. In difficult environments, follow the "Rule of Thirds" (*below*).

THE RULE OF THIRDS

When planning a dive in an environment that is covered overhead—such as a wreck, cave, or under ice—or when making deeper dives, always apply the Rule of Thirds. The rule states that you should use a third of your air for descent and exploration, and a third returning to your guideline or point of entry and for ascent. The remaining third is an emergency reserve.

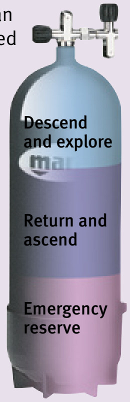

Buoyancy control

Without doubt, the most important skill for a diver to master is that of buoyancy control. Until you achieve this, you will be constantly expending effort in order to maintain your position in the water, wasting air resources and distracting you from your dive.

SUITS AND WEIGHTS

Wearing any type of exposure suit will make you more buoyant in the water, and thus make it more difficult to dive. Drysuits contain an airspace between the suit fabric and the diver's body. Wetsuits have air cells within their neoprene fabric, and so the thicker the neoprene, the more buoyant you will be. The function of weights is to counteract the buoyancy your suit gives you so that, with your lungs full of air but your BC uninflated, you float at the correct level at the water's surface.

Correct assessment of how much weight you need to carry for a given dive is very important. Spend time conducting weight tests for different suit and equipment combinations, and note your findings for future use. Remember, saltwater is denser than fresh water, so you will require more weight when you dive in the sea than in fresh water.

USING YOUR BC

At the start of a dive, you will usually enter the water with your BC inflated, as having positive buoyancy allows you

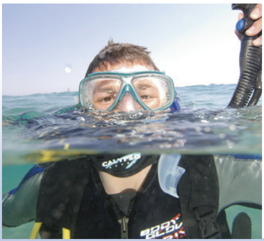

Too high

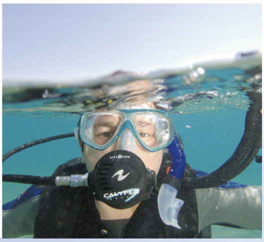

Too low

CONDUCTING A WEIGHT TEST
Empty your BC, and take a deep breath and hold it. Float in an upright position. If you are wearing the right amount of weight, you will float at eye-level to the waterline. Add or discard weights as necessary until this is achieved.

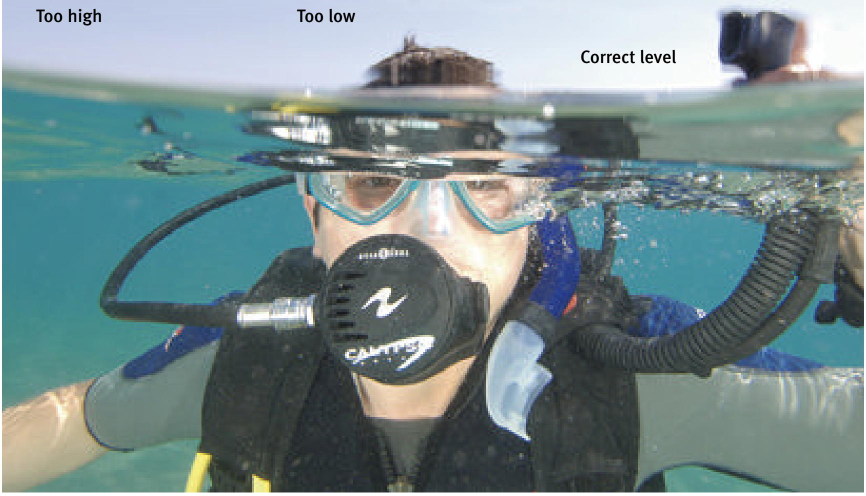

Correct level

to float at the surface effortlessly while you meet up with your buddy—the only exception to this is at sites where there are strong surface currents, when it is important to get to the bottom quickly.

In order to descend, you must be negatively buoyant, which you achieve by deflating your BC and exhaling (*see pp.114–15*). However, as you sink, you will find that your descent rate accelerates as your suit—and the remaining air in your BC—becomes compressed. To counteract this, you must add air to your BC. If this halts you before you have reached the desired level, let air out. Repeat the process, letting small amounts of air in and out of your BC until you become stationary at the desired depth.

When this happens, you will have achieved neutral buoyancy underwater—a state in which you neither float nor sink. Achieving this perfect equilibrium of weight ballast and lifting forces means you do not waste effort and air using your fins to maintain a certain depth.

When you want to ascend, the same control principles apply, but in reverse. As you go up, the air in your suit and BC expands as external water pressure decreases, making you increasingly buoyant, and causing you to rise rapidly. As it is important to ascend at a slow,

LEARNING THE BASICS
An instructor looks on as a student diver practices fin pivots (*see pp.146–49*), an exercise designed to improve basic buoyancy control.

HOVERING OVER A REEF
Buoyancy control is critical when diving over delicate coral that can be damaged by physical contact.

steady rate (*see pp.128–31*), you'll need to vent air accordingly as you rise. To descend and ascend as smoothly as possible, you must become familiar with how your BC responds to inputs of air, and how quickly it vents air.

Drysuits have independent inflation and deflation controls (*see pp.54–55*). If you are diving in a drysuit, the air space between your suit and your body will become compressed, and you will need to inject a little air into the suit to avoid discomfort from being "shrink-wrapped" by external water pressure. This will make you more buoyant, so you will need to vent air from your BC to counteract it. On ascent, you will need to dump the excess air from your drysuit.

FINE CONTROL OF BUOYANCY

Small movements up and down—to look at different parts of a reef, for example—are best achieved through breath control, rather than letting air in and out of your BC. Good buoyancy control only comes with practice, and is worth perfecting. Take time at the start of each dive to practice fine control of buoyancy before you set off exploring. From a neutrally buoyant position on the seabed, slowly inhale—you should rise up and hover just above the seabed. Breathe out a little—you should sink. Practice breathing in and out and become familiar with how it influences your gentle rise and fall. You should be able to hover over the seabed and control your buoyancy by breathing alone, without constantly filling and

JETTISONING YOUR WEIGHTS

In extreme circumstances, such as being caught in a powerful downcurrent, it may be necessary—as a last resort only—to dump your weight belt. You will instantly start to rise up rapidly through the water, so be ready to vent air from your BC to slow your ascent. Breathe out on the way up, to minimize the risk of lung expansion injury (*see p.95*).

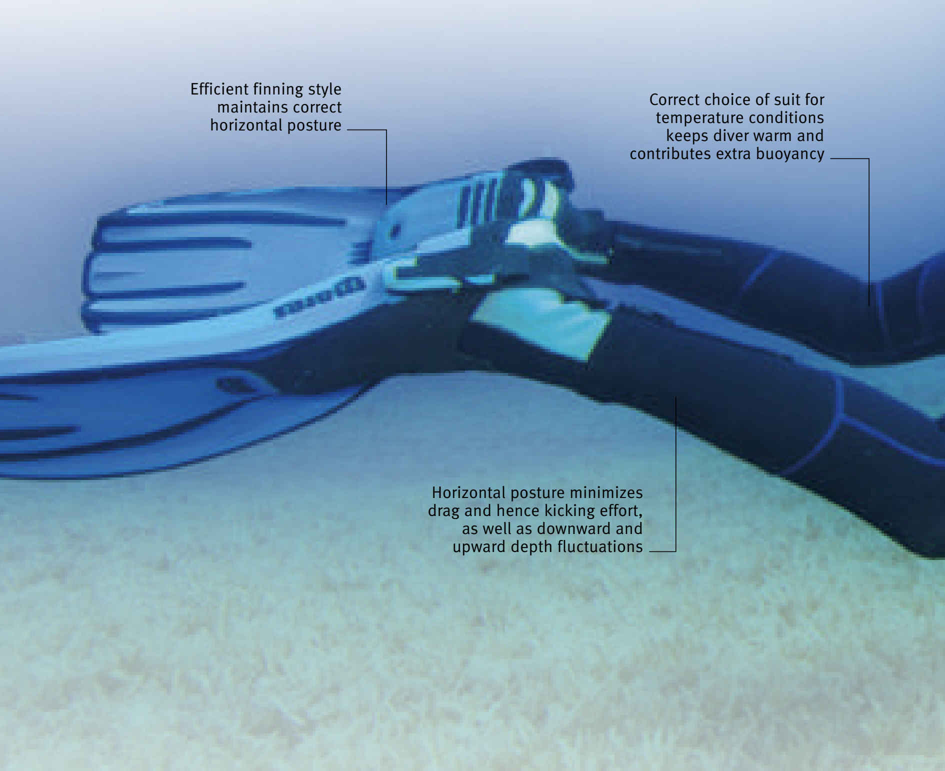

Efficient finning style maintains correct horizontal posture

Correct choice of suit for temperature conditions keeps diver warm and contributes extra buoyancy

Horizontal posture minimizes drag and hence kicking effort, as well as downward and upward depth fluctuations emptying your BC. Once you have mastered this, you will be able to float at any depth without effort.

TRIM AND WEIGHT DISTRIBUTION

The position of your weights and equipment has an important part to play in maintaining good "trim" (correct posture and balance) underwater. Distribution of weight is key in achieving a good horizontal attitude, which makes buoyancy control easier and helps you to propel yourself along with the minimum of drag. A streamlined profile also helps posture and ease of movement, which helps to reduce air consumption. Experiment with different weight systems and stowage points until you find a comfortable combination that helps you achieve the best position in the water. Some designs of BC allow you to stow weight at different points, not just on a belt. For example, some have special quick-release pouches that are designed to take moderate weight loads (*see p.59*). Always check that you are able to jettison the bulk of your weight load in case of emergency.

ACHIEVING PERFECT BUOYANCY
Thinking about buoyancy as a system of influences to be harmonized, rather than just weights and BC, will help you perfect the art of moving underwater.

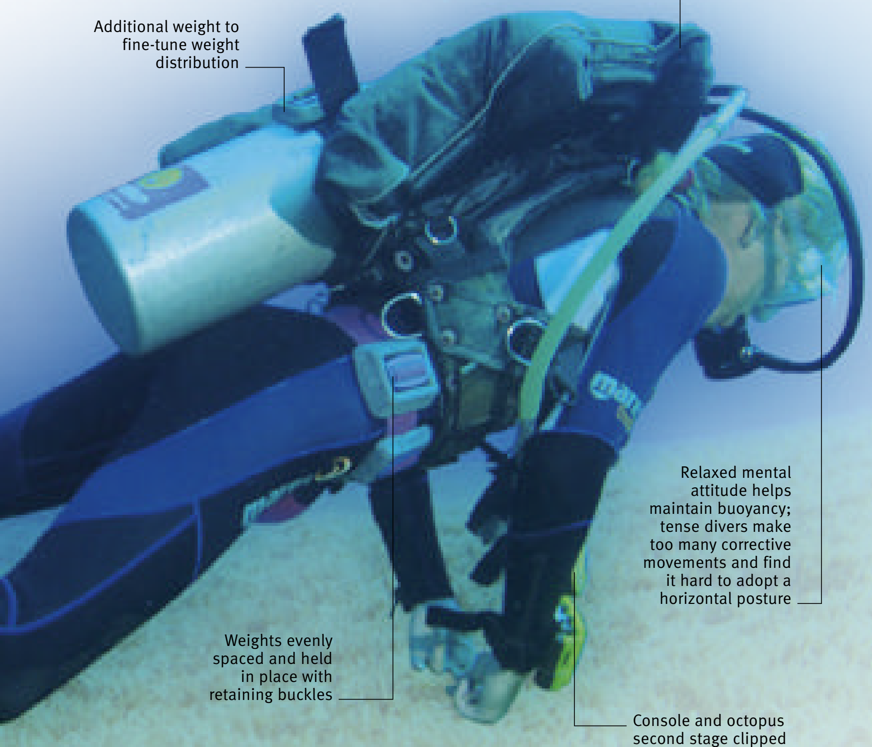

BC filled with correct amount of air for depth and weight load

Additional weight to fine-tune weight distribution

Relaxed mental attitude helps maintain buoyancy; tense divers make too many corrective movements and find it hard to adopt a horizontal posture

Weights evenly spaced and held in place with retaining buckles

Console and octopus second stage clipped close to body to minimize drag and avoid damage or entanglement

Moving underwater

The duration of your dive is limited by your rate of air consumption, which in turn depends on how much energy you expend. Good finning technique, buoyancy control, and a streamlined profile will all make your movement underwater more energy-efficient.

EFFICIENT PROPULSION

Your movement through the water should be driven and controlled by your legs. If you watch experienced divers, you will notice that they use their arms very little, and move effortlessly through the water, saving energy, minimizing air consumption, and avoiding damage to reefs and the marine environment.

Swimming with fins should be an extremely efficient way of propelling yourself through the water, but many divers are let down by a poor finning technique. There are several different strokes you can use, but all should be carried out with strong, rhythmic movements, fully extending the leg (or legs) with each kick. Uncontrolled thrashing and incomplete kicks will simply waste energy. It is worth asking your buddy to give you feedback on your technique—even to photograph or film you. You may be surprised at the results.

Scuba diving is not a race, and the faster you go, the less you see. Learning to pace yourself and move at a leisurely speed will help you to relax, and a relaxed diver is an aware, efficient, and safe one.

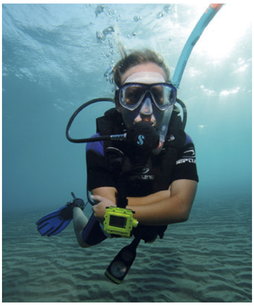

MINIMIZING ARM USE
To practice with your legs only, cross your arms in front of you or let them hang by your side.

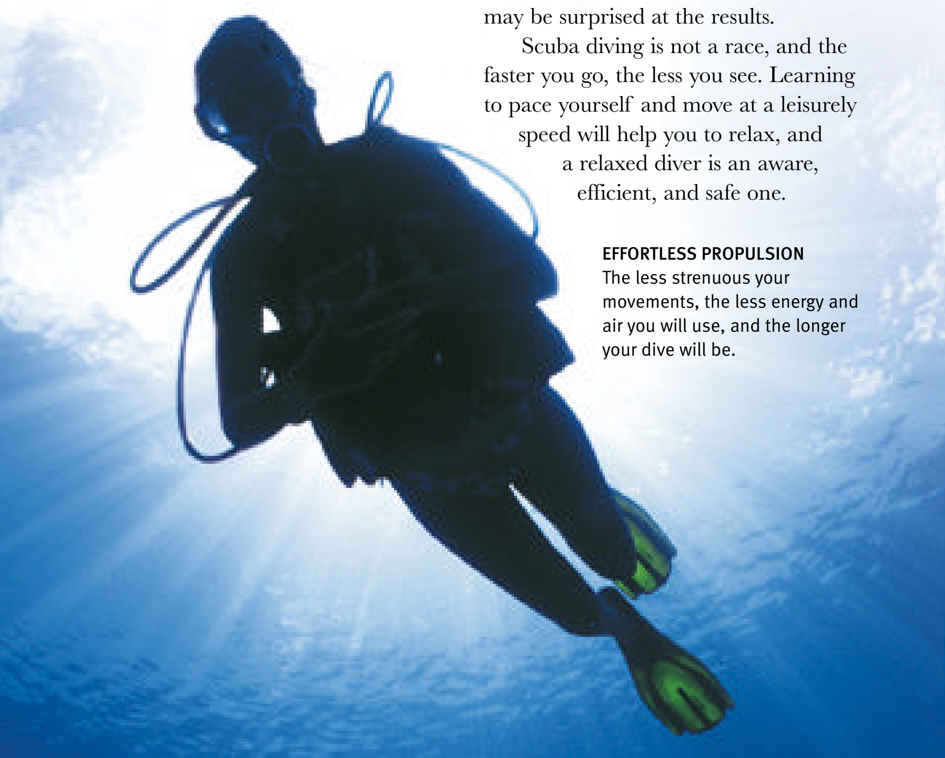

EFFORTLESS PROPULSION
The less strenuous your movements, the less energy and air you will use, and the longer your dive will be.

PROFILE AND BUOYANCY

To achieve the most streamlined (and therefore most energy-efficient) swimming position, you should be almost horizontal in the water. Your hips will be slightly lower than your upper body due to that area being weighted. During a dive, you should look forward and downward most of the time—raising your head too much will lift your upper body and make your profile less streamlined.

POOR PROFILE
Swimming with the wrong profile may mean that your fins kick up silt, impairing visibility.

Carrying the correct amount of weight is important. Being overweighted will drag your hips and lower body downward. Swimming will then require greater effort due to your unnatural angle in the water, which will produce increased drag. You can improve your profile by repositioning the weights in your BC, or refining your weight requirements (*see pp.118–21*).

Achieving neutral buoyancy (*see pp.118–21*) is the key to maintaining a good dive profile with ease, and also increases your efficiency. If you are positively buoyant, you will be constantly swimming downward; if negatively buoyant, you will have to swim upward all the time or crawl along the bottom—all of which require more effort than horizontal, neutrally buoyant, swimming.

TYPES OF KICKING STROKES

With the flutter kick (*below*), all fin movement takes place above the body's centerline, reducing the amount of water directed downwards. This can be useful when you wish to avoid stirring up sediment on the seafloor. The frog kick is also ideal for use where fine silt can ruin visibility if disturbed, because it directs thrust horizontally behind you. The swimmer kick is good for power, but uses up energy quickly. Most divers tend to drift slightly when using the flutter kick or the swimmer's kick, because one leg is slightly stronger than the other.

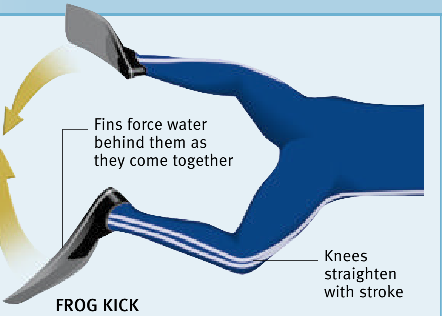
Fins force water behind them as they come together
Knees straighten with stroke

FROG KICK
Spread your legs and bring up your knees. Rotate your fins so that the bottoms of the blades face behind you. Bring the fins together in an arc, thrusting out behind.

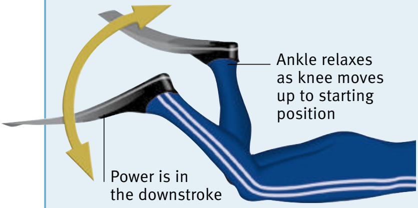
Ankle relaxes as knee moves up to starting position
Power is in the downstroke

FLUTTER KICK
Bend the knees slightly, shorten the stroke of the kick, and keep the hips fairly static. The stroke ends when the knee is straight.

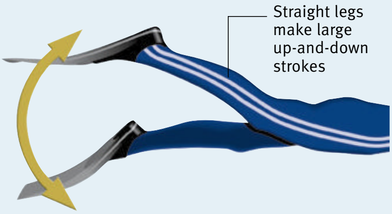
Straight legs make large up-and-down strokes

SWIMMER KICK
Lock your legs straight and kick from the hips with long strokes. This technique exploits the power in your thigh and hip muscles.

Exploring and navigating

Learning how to use instruments and natural features to find your way around underwater is important. Accurate navigation will help you to reach your planned exit point at the end of a dive, and will also help you find and explore features of interest, such as wrecks.

MAPPING THE SITE

It is easy to become disoriented in poor visibility, but good navigation skills will help you find your way, even when conditions are against you. Developing such skills will make you less dependent on other divers for your own safety—if you always rely on your buddy to lead the way, you could find yourself in trouble if you become separated.

On first entering the water, make mental notes of the surrounding features, such as unusual rocks, coral, or your position on a wreck site. As you progress through the dive, try to build up a map of the site in your mind. At every turning point, note nearby features and which direction you turned in relation to them.

FOLLOWING A REEF WALL
In clear conditions, a reef wall makes an easy navigational reference point. Try to remember distinctive features as you go along.

MENTALLY MAPPING THE DIVE SITE
Reef, rock, and seabed features, sunlight, and even shoals of fish can all help you to orient yourself at the dive site.

Look at the seabed geology, too: rocks often have features or show strata that make memorable visual markers. It may be helpful to jot notes down on a slate.

NATURAL NAVIGATION AIDS

A compass (*see pp.126–27*) gives the most accurate directional information, but there are also clues in nature. If the water is clear and shallow enough, or if there are any shadows cast by rocks, you can use the Sun as a directional reference. On a morning dive, for example, the Sun will be in the east. Take note of which way the current is running. (If you are diving close to slack water (*see pp.102–03*), at high or low tide, this may change by as much as 180° as the tide begins to turn.) Even if you cannot feel the current, you can find visual clues to its direction. Plankton, for example, drift along with the current, so shoals of fish station themselves facing into the current so that they can feed on them. Kelp fronds flow in the direction of the current. Exhaust bubbles will rise up toward the surface (unless you are in a down current).

READING SAND RIPPLES

Sand ripples run parallel to the shore, at 90 degrees to the path of incoming and outgoing tidal currents. Sand ripples get closer together nearer to shore, providing a useful reference point for underwater navigation. If the spacing of sand ripples underneath you appears to be getting narrower, you are swimming toward shallow water and a possible exit point.

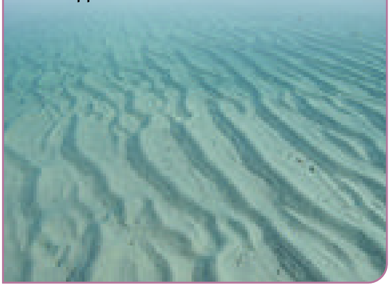

Sand ripples on the seabed

FOLLOWING A REEF

On a basic reef dive, swim out with the reef wall on one side of your body and then return with the wall on your other side. When diving around a wreck, try to maintain either a clockwise or counter-clockwise motion, especially if the wreckage is fragmented. Keep the main part of the wreck to one side of your body during the dive, so that you return to your starting point after one circuit.

USING A HANDHELD COMPASS
In case you are separated for any reason, both you and your buddy should carry and be confident at using a magnetic compass.

USING A COMPASS

A compass has a circular body, often divided into 360-degree increments, and a needle that swivels to point to magnetic north. Many also have a "bezel"—an outer ring, also marked with increments and the points of the compass, that can be rotated independently to the body. This can be used in conjunction with the "direction of travel" arrow (known as the "lubber line"), which is marked on the compass body, to establish the direction you want to travel in relation to magnetic north. Point the lubber line in the direction in which you wish to travel, then turn the bezel so that its "north" mark matches the direction in which the needle is pointing. If you ensure that the needle always points to the north mark, as you move in the direction indicated by the lubber line you will maintain your intended direction. Digital compasses work in the same way, but display the information on an LCD.

A compass is a fairly basic device, but it is easy to use it incorrectly. If it is a wrist-mounted model, ensure that your

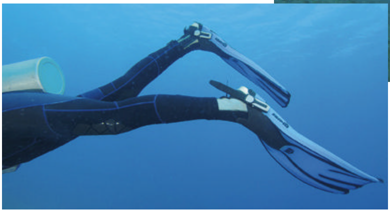

MEASURING DISTANCE
When navigating underwater, you can judge relative distances by counting how many fin strokes you have made in a given direction, or by timing your progress along a course.

wrist is held square to the direction of travel, and not at an angle. It is simpler to use a compass mounted on a console or a retractable lanyard, which can be held out straight in front of you. Use both compass readings and environmental observations, so that you can correct any inaccuracies in either method. Mark readings down on a slate as you go, and try to integrate these into your mental map of the dive site. In slack water, timing how long it takes to get from one point to the next gives you an idea of your position if you return along the same compass course. Combining this information with depth readings taken along the way will help you navigate with more accuracy.

LAYING A DISTANCE LINE

In low-visibility environments—such as murky temperate waters—it may be necessary to lay lines to aid navigation back to a convenient point. Lines can also be used to search an area: if you know the approximate position of a specific site, natural object, or artifact, you can use a line, in arcs of different lengths, to scan a large area of seabed efficiently.

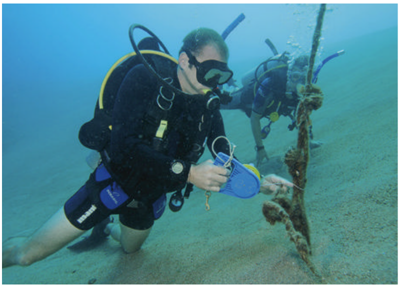

1 **Find a solid anchor point** to attach the line to. Make a couple of turns around the object for extra security. Do not secure a line to jagged edges that will fray or cut through it.

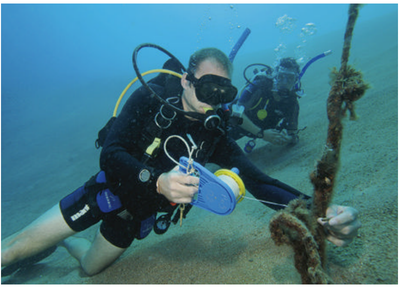

2 **Secure the line** to the anchor point. If you have a reel and line you use regularly, fit a spring clip to the end of the line so that it can wrap around the anchor point and clip back on to itself.

3 **Unspool the line** as you explore, making sure that your buddy keeps clear to one side. The line should be tied off again wherever you change heading.

Ascent

All good things come to an end, and at some point your dive will be over. Ascending at the end of a dive should be carried out with great care. This is the part of the dive where decompression sickness (*see pp.142–43*) can develop, so your ascent must not be hurried.

WHEN TO ASCEND

Dives often end when you reach a prearranged time to surface, either to meet the boat cover or to account for changing tides. However, you may decide to ascend before this if the site proves uninteresting, or if you or your buddy are cold or tired. You may also be forced to ascend if you have only 725 psi (50 bar) left in your tank, if you are making a decompression dive (*see pp.130–31*) and reach the scheduled minimum volume of air in your tank, or if problems cause you to abort the dive.

SIGNALING TO ASCEND
To ask your buddy if they are ready to ascend, or to signal your intent, raise your thumb.

Before ascent, you and your buddy should signal the "thumbs up" to each other that you are both able and ready to ascend to the surface.

If possible, avoid making a "free ascent," with no visual reference to guide you. If you can follow a reef wall or a permanent mooring line, for example, this will give you a natural sense of your ascent rate, in addition to your computer readings, and can help avoid disorientation. Sometimes, though, it is not possible to ascend up a line or along a natural feature, and it can be helpful to deploy an delayed surface

MONITORING A DIVE COMPUTER
Dive computers (*see pp.66–67*) help the diver judge ascent rate. Some devices give audible warnings if safe limits are exceeded.

BUDDY PAIR ASCENDING A GUIDELINE
A buoyed line moored to a wreck or seabed feature offers one of the safest guides for an ascent, especially in a strong current.

marker buoy (DSMB; *see pp.130–31*)—this will give you a vertical reference and also signals your position at the surface. Ideally, both members of the buddy pair should carry a DSMB. If you are diving on a wreck or a site with a guideline, aim to get back to the line for your ascent—not only will boat cover know where to pick you up at the surface, but the line will assist you in making decompression stops and, in a current, can provide something to hold on to. If you do not manage to get back to the guideline, send up a DSMB.

MAKING YOUR ASCENT

To start your ascent, slowly breathe in or let a little air into your BC—this should cause you to rise up. Rise up next to, and ideally face-to-face with, your buddy so that you can check each other's rate of ascent and condition. As ascent continues, carefully monitor the amount of air in your BC (and drysuit if relevant)—the air will expand as the pressure decreases during your ascent, causing your rate of ascent to increase rapidly. Keep one hand on the BC exhaust hose in order to vent air from

FOLLOWING A SLOPE
Ascending next to a natural feature, such as a reef, offers a reassuring reference point.

MAKING A SAFETY STOP
Divers make a precautionary safety stop at 16 ft (5 m) to allow nitrogen to exit the body's tissues at a safe rate.

it as you ascend. You may also need to dump air from your drysuit. You can fin to assist your ascent, especially to gain momentum to begin with, but this will not be necessary if you are correctly weighted and positively buoyant. Keep breathing normally and do not hold your breath.

A conservative estimate of a safe rate of ascent is 50 ft (15 m) per minute (or slower) up to 16 ft (5 m) in depth. At this point many divers make a 3-minute safety stop. Whether you do or not, the last 16 ft (5 m) to the surface should take another full minute. If making a decompression dive, make all your scheduled safety stops and check with your buddy as you do so.

DEPLOYING AN INFLATABLE BUOY

Since ascent must be conducted in a controlled way, it is helpful to carry your own portable guideline—an inflatable buoy, or safety sausage. These highly visible tubes are launched using air from your second stage, and allow you to ascend up an attached line. Using them is a real skill, so seek training before your first attempt. Remember to reel the line in as you ascend, since slack line can get wrapped around your gear.

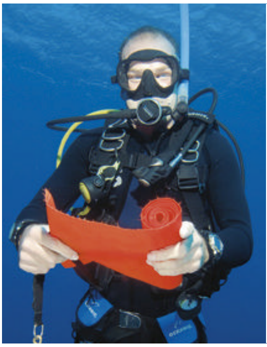

1 **Unfurl the buoy**, which should be attached to a reel. Unclip the reel from yourself, taking care not to drop it.

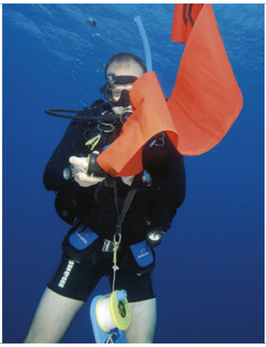

2 **Partially inflate** with a small burst of purged air from your octopus (*see p.61*) to stand the buoy upright.

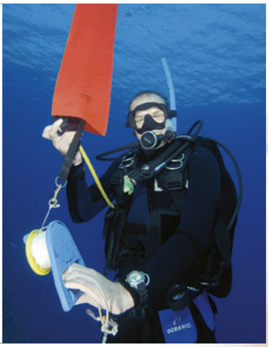

3 **Hold the reel** in one hand, and with the other, hold your octopus valve under the buoy's inflation point.

4 **Launch the** buoy by purging a long blast of air into it with your octopus. Release the spool, holding tight to the reel, and the buoy will shoot upward.

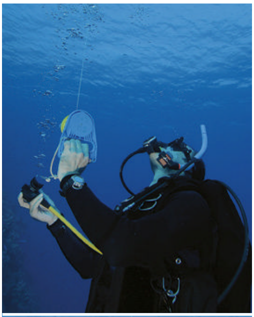

5 **When the** spool stops spinning, start to ascend, reeling in the line as it slackens. Ascend at a safe pace: do not tug on the line to pull yourself up. During ascent, keep one hand ready to vent air from your BC if required.

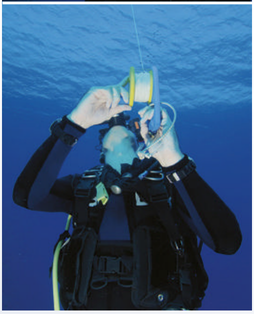

6 **Await recovery** at the surface from your boat cover, which will have sighted the buoy at the start of your ascent. The buoy should be left in its inflated state until you have been recovered.

Surface skills and recovery

Once you and your buddy have completed your dive and are at the surface, you need to head back to shore or wait to be picked up by your boat cover. Always remember that the dive isn't over until you and your buddy are back on dry land or on your dive vessel!

EXITING ON TO THE SHORE

If completing a shore dive, you and your buddy should swim to the shore and find a safe, appropriate method of getting back on to dry land. If you are diving from a sandy beach, it is usually best to keep your fins on for as long as possible and walk backward up the beach until you are well out of the way of any swell. Removing your fins too early can lead to problems if you are swept off your feet or back out to sea, since you will have no propulsion.

If diving from a rocky shore, try to find an easy route back on to the rocks. Again, aim to keep your fins on for as long as possible. As you take off your gear, place it clear of any swell that could carry it back into the water.

EXITING ON TO A BEACH
Only remove your fins and walk forward when exiting the shallows of a calm sea, or once you are out of the way of any swell.

SIGNALING TO A BOAT
Your dive plan should ensure that your boat cover knows roughly when you will surface. If you need help quickly, shout and wave a signaling flag.

SIGNALING TO BOAT COVER

If you are diving from a boat, remain close to your buddy with your regulator in place and signal to your boat that you are OK and awaiting recovery. If there is a problem or emergency (*see p.141*) and you or your buddy require immediate recovery, wave to the boat, or draw extra attention to yourselves by blowing a whistle, shouting, or waving

BOARDING A BOAT
It is usually easiest to hand up any small items first, then hoist yourself up the ladder while still suited up.

a flag. As the boat approaches, move to it with caution, keeping your distance until you have a signal from the boat that it is in neutral and safe to approach.

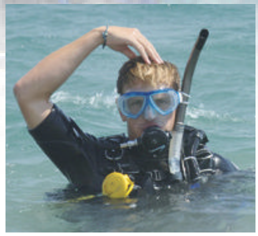

DIVER GIVES "OK" SIGNAL
A standard signal for "I'm OK" at the surface is to put your hand on top of your head, forming a closed arc.

EXITING ON TO A BOAT OR RIB

To get into a rigid inflatable boat (RIB), you should first get a hold on the boat, either by a rope or by a handle on the side. Make sure your BC is inflated, hand your weight belt in first, and then slip out of your BC by unclipping one of its shoulder straps. Keep hold of it until someone from the boat can pull it in, along with your tank. You should then be able to kick with your fins and pull yourself up on to the RIB.

If returning to a hard boat, head for the ladder and hold on to the bottom. Take great care in a swell—the ladder will move in the water. The skipper or dive master will probably tell you whether to de-kit in the water or to enter the boat wearing your gear. You may be able to hand in your weight belt, BC and tank, but often it is easy enough to remove just your fins at the base of the ladder, hand them up to someone in the boat or hold on to them, and simply climb the ladder fully kitted up. If you feel you are not strong enough to do this, let someone on board know (preferably before you dive).

Once on the boat, be aware that others may be waiting in the water to get on board. Stand back to let others on, stow your gear quickly and neatly, and then offer assistance if it is needed.

Post-dive and debriefing

Your dive is over, and there is equipment to be cleaned and stowed, and records to write up while your dive is still fresh in your mind. Problems may have occurred during the dive, and questions may have arisen that need to be addressed. Now is the best time to do this.

PRACTICAL MATTERS

The first post-dive task is to wash your gear. You need to ensure that it is clean and free from mud or sand. If you have been diving in saltwater, then this also needs to be thoroughly rinsed off before the salt starts to corrode metal parts. Large dive boats and liveaboards usually have one facility for washing electronic equipment, such as computers and cameras, and another for your BC, regulator, and other gear. Rinse all items thoroughly to remove excess salt and sand, which can easily work itself into O-rings and regulators. There may also be facilities on board to dry your gear before stowing it away in a bag or box for your next dive. Make sure you follow instructions and do not leave gear lying around—this can lead to loss or damage, which may impede your next dive. If you are diving from a rigid inflatable boat (RIB) or small dive boat, you will need to find a water source back on land to rinse your gear, and then leave it to dry or stow it for transport.

ANALYZING THE DIVE

There will normally be some sort of debriefing with your dive leader after a dive. This gives you an opportunity to ask questions about the dive (for example, to identify fish species that you did not recognize, or discuss problems with gear, navigation, or currents). This is a time when you can learn about the site for a future dive and share with others ways to improve skills.

You should always fill in a log book after every dive, recording the depth, time, and conditions (including visibility, water temperature, and currents), as well as key features of the dive—details of a wreck, species encountered, equipment configuration, the type of suit you were wearing and the weights you used, and any problems you experienced. The log will be a valuable reference for future dives.

Finally, ensure that after every dive you are well rehydrated. You will have expended a high level of energy and will have lost fluids. Choose non-fizzy soft drinks or, ideally, water, to ensure that you rehydrate your body properly. Keep your energy levels up, especially if you are diving again soon, by having a good meal or eating a snack with a high energy content. If you are cold after a dive, get warm and dry—have a hot drink and get out of your wetsuit, if possible.

DIVER CLEANING GEAR
Rinsing, drying, checking, and stowing your gear should become a post-dive routine.

REACHING THE SHORE
Eager as you may be to talk about what you have seen, attend to your gear first.

LOOKING FORWARD

If you are planning to make more than one dive in the same day, consider beforehand what your surface interval will be between dives. Aim to do your deepest dive first, and shallower dives subsequently. Make sure you are in good shape to dive again—if you have become cold or fatigued on the first dive, it is better to spend time recovering than to dive again too soon.

DIVERS DEBRIEFING
Talking to other divers in the post-dive get-together will help broaden your knowledge of diving, and you may pick up some handy tips.

Problem solving

Diving is a safe sport if you follow the rules, avoid taking unnecessary risks, and always dive with a buddy. But that does not mean you will never encounter difficulties during a dive. Learning how to anticipate and solve problems is part of becoming an accomplished diver.

KNOW YOUR LIMITS

One of the most exciting things about diving is that no two dives are ever the same. Even consecutive dives on the same dive site, with the same buddy, will be different due to changes in currents, visibility, and your own attitude. Some days you will feel great, and everything will work out perfectly, and other days you might be tired, have a slight headache, perhaps, and there may be minor problems underwater that you have to deal with. In this unpredictable sport, you must be mentally and physically prepared for every eventuality.

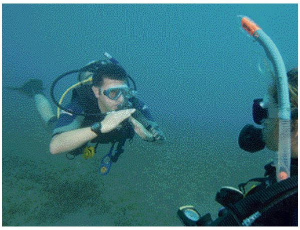

REGULAR SIGNALING
Remain in close communication with your buddy at all times, and signal to each other regularly that you are OK.

To minimize the likelihood of problems and incidents occurring during a dive, do not dive without preparation in conditions with which you are unfamiliar. For example, if you have never dived in a cave, wreck, or under ice, make sure that you undergo the relevant training before attempting to do so. Diving in conditions for which you are unprepared will lead to anxiety, and sometimes to panic—and that is when problems often arise.

The buddy system is important for your safety, but as a diver you must ultimately take responsibility for your own welfare. If you feel uneasy during a dive, tell your buddy, and either change or abort the dive. Never let peer pressure push you beyond your limits and training, even if this means that your buddy's dive may be unsatisfactory. Buddy compatibility is important, since accidents can occur when two egos clash.

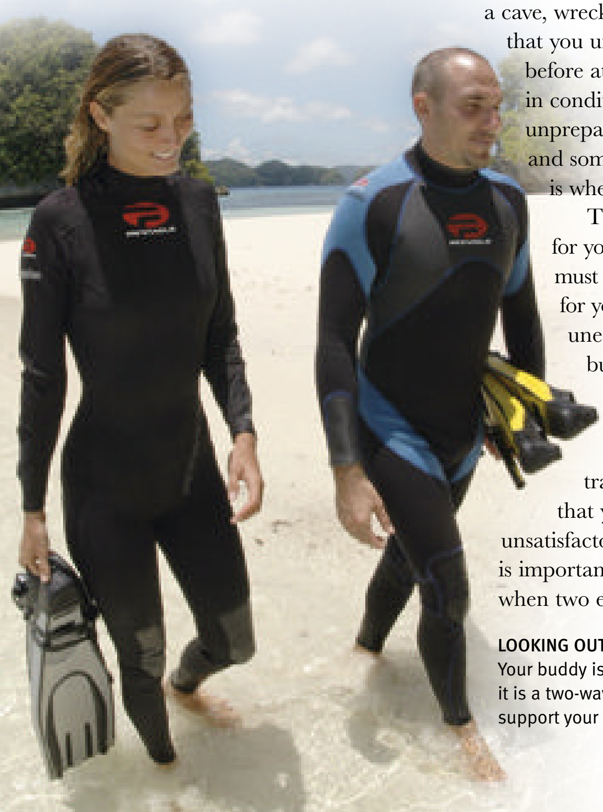

LOOKING OUT FOR YOUR BUDDY
Your buddy is your underwater lifeguard, but it is a two-way relationship—you also have to support your buddy when problems arise.

MENTAL AND PHYSICAL FITNESS
Confidence, a positive mental attitude, good health, and good levels of fitness are essential ingredients for a safe, successful dive.

ANTICIPATING EQUIPMENT PROBLEMS

It is rare that equipment fails during a dive, but you should nevertheless take great care when assembling, storing, and servicing your gear. After all, this is what keeps you alive underwater, so it is worth spending time and money on keeping it in perfect working order.

Consider "what if" situations ahead of a dive, and think through how you will deal with the failure of any element of your gear. Do you have a spare if your buddy is not close by? Your regulator should always have an octopus second stage (*see p.61*), to use as a backup in case your main second stage fails.

HANDLING INCIDENTS CALMLY

Many things can go wrong on a dive, but most are minor issues that can be solved without aborting the dive. Even serious problems can be dealt with if you remain focused and positive. Be persistent, and say "no" to failure. When you become aware of a problem, stop, think, breathe, and then act. Whatever happens, do not panic—remember that your buddy is there to help. Make sure you are there for your buddy, too, and be quick to reassure or assist them if they appear disoriented or anxious.

THE INCIDENT PIT

Minor problems can escalate into major ones if you do not stay calm and deal with them methodically. Imagine that you lose your regulator while trying to ease cramp in your foot, and then fumble as you try to retrieve it. If you panic, you may inhale water, choke, and become disorientated. This type of scenario is known as the "incident pit"—stay out of it by thinking clearly and acting methodically.

TROUBLE-SHOOTING

The equipment problems you will encounter most commonly will involve your regulator, mask, and inflator hose.

If your regulator comes away from your mouth and you cannot retrieve it, reach for your octopus second stage (*see p.61*) and breathe from it. If your octopus is not accessible for any reason, signal to your buddy, who should offer you their own octopus to use. Take a few breaths from this to calm down and, with your buddy's help, locate your main regulator and replace it.

Should air start to flow uncontrollably from your regulator, take the regulator out of your mouth and push the purge button, which should clear any grit that may be jamming the valve open. If this does not work, switch to your octopus second stage and signal to your buddy that your air supply is running down more quickly than normal. Monitor your air supply carefully, and ascend immediately with your buddy. Note that in very cold fresh water, freezing up of the regulator will probably be the cause of any free-flow, in which case purging will not help. In this scenario, you may need to breathe from your buddy's octopus while performing a shared-air ascent together (*see p.140*).

RETRIEVING LOST REGULATOR
A lost regulator is usually easy to retrieve, but if your buddy is finding this difficult, go to their aid and, if necessary, offer them your octopus.

CALMING YOUR BUDDY
Reassuring physical contact can help to calm a disoriented buddy and prevent panic from setting in.

SECURING A BUDDY'S LOOSE TANK
Some problems are difficult to solve alone, and you should not hesitate to alert your buddy if you need assistance.

SOLVING MASK AND HOSE ISSUES

If your mask is dislodged or starts to fill with water, simply clear it in the normal way. Lean your head back, press in the upper edge of the mask, and breathe out through your nose. The air bubbles you exhale will clear any excess water in your mask. At depth, this can seem daunting, so be prepared to practice this procedure until you are comfortable with it and can execute it confidently.

REMOVING YOUR WEIGHT BELT
You must be competent at removing your weight belt underwater, in case you need to adjust your weights or if your belt becomes snagged.

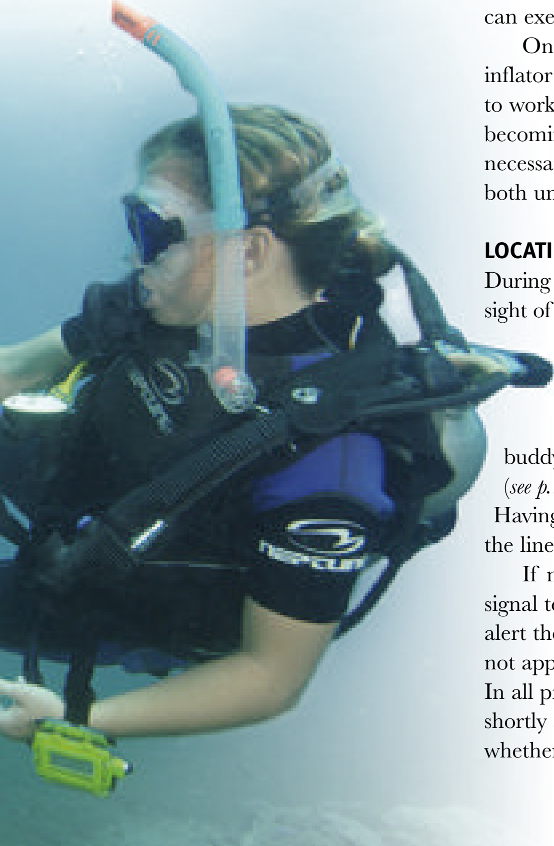

On occasion, you may find that the inflator hose for your BC or drysuit fails to work. This is usually caused by grit becoming lodged in the valve. If necessary, BCs can be inflated orally, both underwater or at the surface.

LOCATING A LOST BUDDY

During a dive, you may find you lose sight of your buddy, especially in low visibility conditions. The correct procedure is to look around for a minute or two, and then surface independently. If your buddy is using a surface marker buoy (*see p.74*), it will indicate their position. Having located it, you can then follow the line down to find your buddy again.

If neither buddy nor buoy is visible, signal to any boat cover to recover you; alert the coastguard if your buddy has not appeared after a few minutes.
In all probability, your buddy will surface shortly after you do. You can then decide whether or not to resume the dive.

FREE AND BUOYANT ASCENTS

If you run out of air, locate your buddy and follow the procedure below. If you cannot find your buddy, you will need to make a rapid "free ascent" by finning to the surface. This can be aided by ditching your weights, but be ready for a sudden increase in buoyancy. Breathe out slowly during ascent to prevent lung expansion injury. If, however, you have a little air left, you can make a more controlled "buoyant ascent." Let some air into your BC to kick-start your ascent, and tilt your head back to watch for the surface. Again, breathe out during the ascent. Control your ascent rate by venting air from the BC, but not so much that you lose buoyancy. Do not rise any faster than

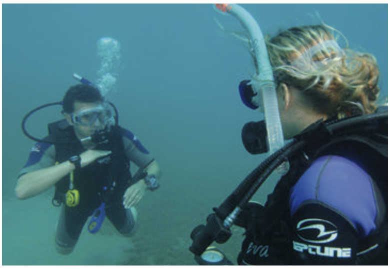

1 **Signal to your buddy** that you are out of air by making a chopping motion at your neck with the flat of your hand.

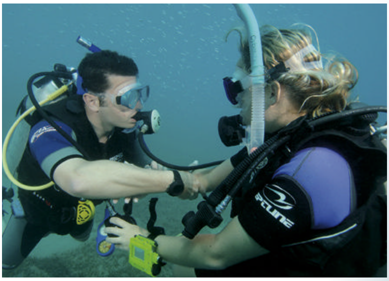

2 **Your buddy will offer** you their octopus second stage, or, if they have one, a separate regulator attached to a pony bottle.

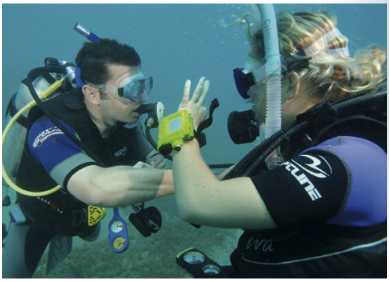

3 **When your buddy is** sure that you are OK and your breathing has settled, you should lock arms and ascend to the surface.

OUT-OF-AIR PROCEDURE WITH BUDDY

If you have not kept track of your air consumption and find that you are out of air, locate your buddy and ascend using their octopus or pony bottle. If no alternate air source is available, then buddy breathe. This involves both divers using the same regulator, with each diver taking two breaths before passing it to their buddy. This is a complex procedure, however, that needs to be practiced if it is to succeed in a real-life crisis.

4 **Hold on to one another** as you rise to the surface. Make any decompression stops as you would do during a normal ascent.

your exhaled bubbles. At the surface, signal to boat cover immediately. If there is no boat, swim to the shore. Your buddy will need to be found and, as you have ascended without safety stops, you will need to be monitored for DCS (*see p.142*).

MAKING EMERGENCY LIFTS

If your buddy is unconscious or injured, keep their regulator in their mouth and perform a buoyant lift to get them to the surface. This means holding on to your buddy's harness as you ascend, using their BC to adjust buoyancy for both of you. Alert any boat cover once at the surface. If your buddy is not breathing, artificial ventilation (AV) may be needed until medical help arrives. If there is no cover, or if you are close to shore, you may decide to tow your buddy to safety. Both towing and AV require special training; if your dive training did not cover them, a lifesaving course is recommended.

OVERCOMING PANIC

If you feel panic coming on, alert your buddy, stop moving, and steady your breathing. If your buddy panics, reassure them with hand signals but observe them from a safe distance, since a flailing arm can knock your regulator from your mouth or injure you. When they have calmed down, hold their hand or arm, make sure their regulator stays in place, and remain close until normal breathing resumes.

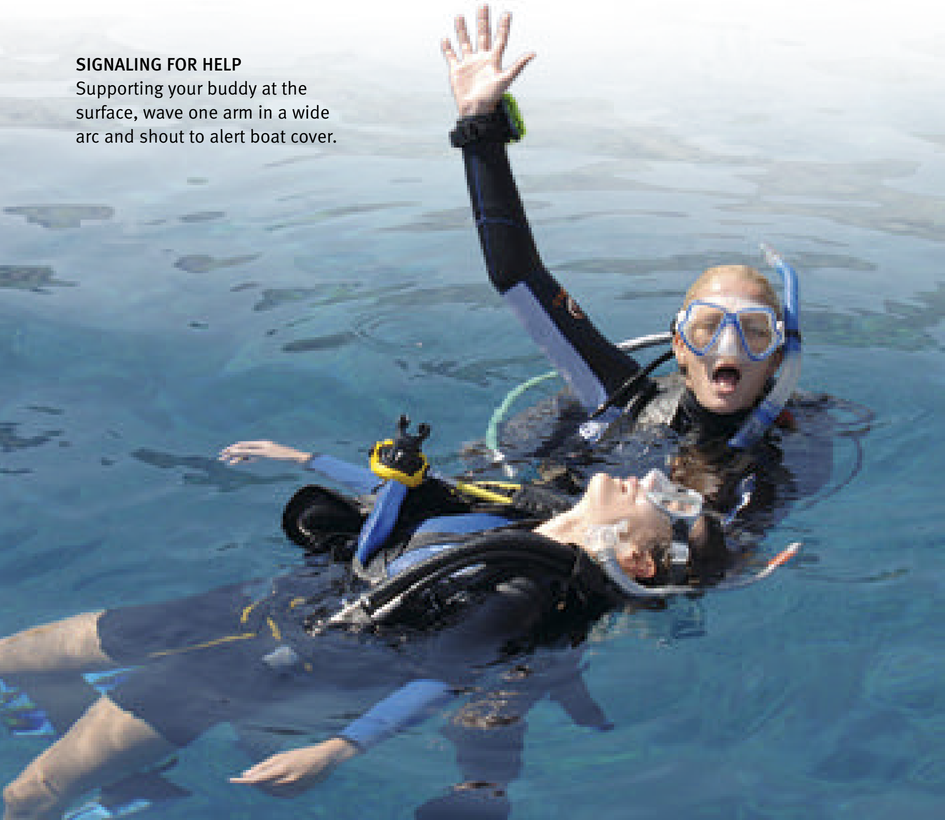

SIGNALING FOR HELP
Supporting your buddy at the surface, wave one arm in a wide arc and shout to alert boat cover.

Decompression sickness

If a diver ascends too quickly, the rapid pressure change causes bubbles of nitrogen to form in tissues and blood vessels *(see pp.100–01)*, resulting in decompression sickness (DCS). This condition must be watched closely and treated quickly, since in extreme cases it is fatal.

DCS SYMPTOMS

Severe DCS may cause impaired vision and balance, convulsions, and even unconsciousness. Nitrogen bubbles can form under the skin, or become trapped in capillaries and restrict the blood supply to the lungs, slowing breathing, lowering blood pressure, and causing shock. If a bubble enters the brain or spinal cord, it can lead to paralysis or even death. Symptoms usually appear between one and six hours after surfacing, but occasionally do not manifest themselves until 24 hours later. They start with minor aches in joints, rashes, and itching. These will disappear after a few hours in mild DCS, but with severe DCS they will be precursors of serious complications. At the first sign of symptoms, request emergency medical attention (from the coast guard, if you are at sea) and give basic first aid: supply the patient with 100 percent oxygen; keep them warm and still (lying down if possible); provide nonalcoholic fluids (ideally sports or rehydrating drinks); and apply pressure to joints to alleviate pain. The patient should be referred for medical attention and may need to be treated in a recompression chamber.

EMERGENCY OXYGEN
It is vital that the dive boats carry an oxygen administration kit in case of emergencies.

If so, provide those in charge of the chamber with the diver's symptoms and details of the dive. In the chamber, the pressure is raised until it causes gas bubbles in the diver's tissues and blood to shrink. The diver is decompressed under controlled conditions, allowing any residual nitrogen to dissipate safely.

LOOKING FOR SYMPTOMS
Skin rashes and joint pains may be precursors of more serious problems.

DCS AT SEA
Record the symptoms, give first aid, and get to shore if feasible—or wait for the coast guard to arrive.

Hand signals

Good communication with your dive buddy is important, not just in an emergency, but also to monitor each other's progress, share underwater experiences, and agree when to end the dive. The most effective method of communication is to use hand signals.

USING HAND SIGNALS

Verbal communication is not an option underwater (*see p.92*) unless you are wearing a full-face mask, but most messages can be conveyed using a combination of hand signals and drawings on a slate. Using hand signals will allow you to exchange information with your dive buddy and even ask each other questions underwater. They can also be reassuring.

Signal to your buddy regularly during a dive, as this will ensure that you remain in close contact and can easily notify each other of any problems that may develop. Repeat any sign that is not clear to your buddy, and acknowledge every signal that he or she makes with an "OK" to show you understand. If your buddy is slow to respond to a signal, check that they are OK. If you are diving at depth, sluggish responses may indicate the onset of nitrogen narcosis (*see pp.100–01*), which impairs reaction times and mental acuity.

Signals are not just limited to showing each other how much air you have left or highlighting problems. They can also be used to inform your buddy of interesting things you have seen, or to alert them to something you are both looking for, such as a particular marine organism. There are numerous signals for marine animals, and you can develop your own within your buddy pair. A common example is the use of a straight, upright hand (representing a fin) on top of your head, meaning "I've seen a shark." Signals for animals like turtles and rays can be readily improvised.

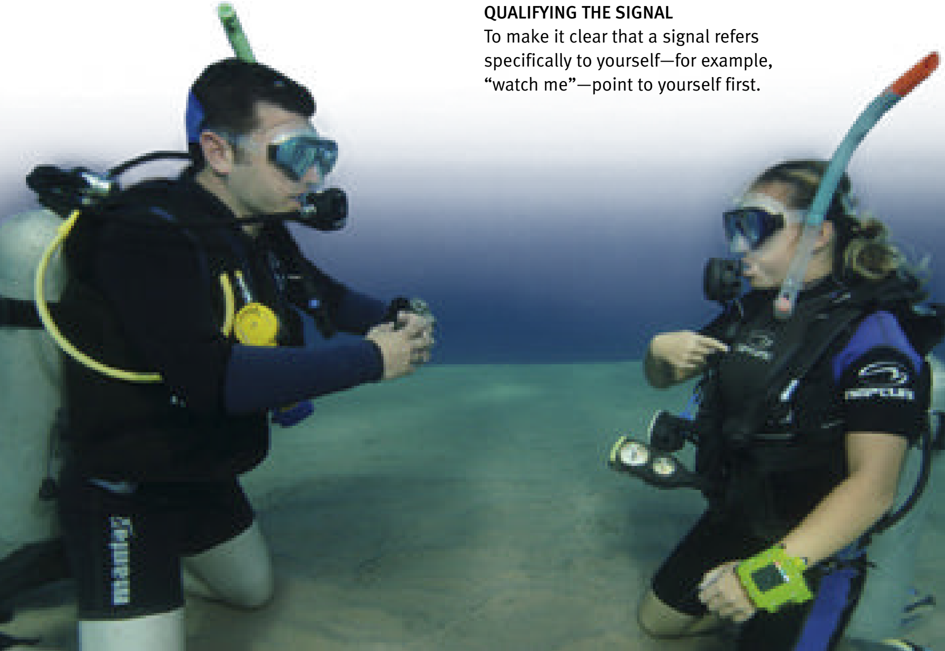

QUALIFYING THE SIGNAL
To make it clear that a signal refers specifically to yourself—for example, "watch me"—point to yourself first.

COMMON HAND SIGNALS

There are a number of standard hand signals that are essential to all divers, and you can also improvise new signals with friends. Always review signals at the start of a dive, especially if you are diving with a new buddy. Give signals to your buddy slowly and clearly, to ensure that they are understood. Use combinations of signals to convey messages more precisely (for example, something is wrong/I can't clear my ears). Below are some of the most important and common diving signals.

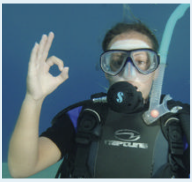

OK/Are you OK?
Form an O with thumb and index finger. Point the other fingers up.

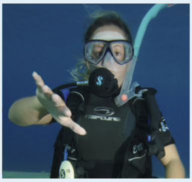

Something is wrong
Tilt the flat of your hand from side to side, palm down.

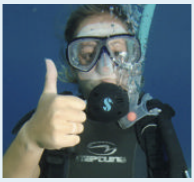

Up/Let's ascend
Point thumb straight upward with fingers clenched in a fist.

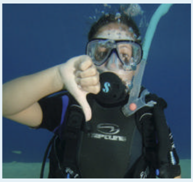

Down/Let's go down
Point thumb downward with fingers clenched in a fist.

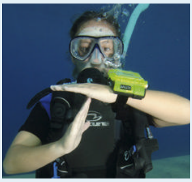

1,500 psi (100 bar) left
Form T-shape with hands to indicate 1,500 psi remaining.

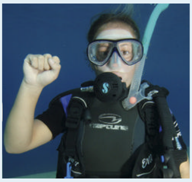

750 psi (50 bar) left
Make a fist, with palm outward, to indicate 750 psi remaining.

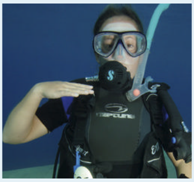

I am out of air
Make a horizontal chopping motion across base of neck.

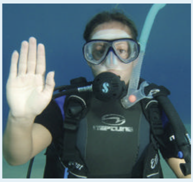

Stop
Present flat of hand, palm outward, to halt buddy or other divers.

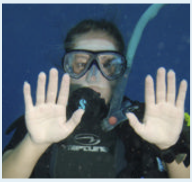

Slow down
Move palms slowly downward together (pivoting at elbows).

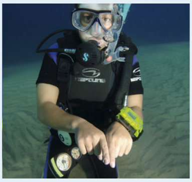

Stay/Move together
Move index fingers together, until touching side by side.

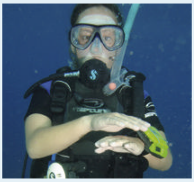

Stay at this depth
Hold both hands horizontally and move one over the other.

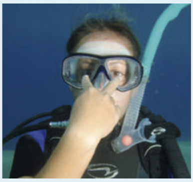

Watch/Look
Point at your eyes, then at subject of interest, or self, or another diver.

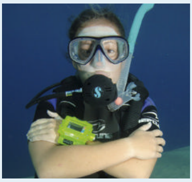

I am cold
Hug yourself with both arms crossed to indicate chill.

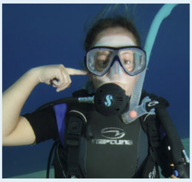

I can't clear my ears
Point at ear to indicate you are having difficulty equalizing.

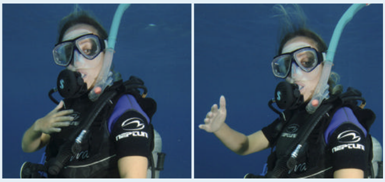

Feeling breathless
If you are feeling out of breath and need to rest for a moment, you should use the flat of your hand to mimic your chest rising and falling.

Developing your skills

With any skilled pastime, it is important to keep your technique sharp. Many divers only dive once a year (or even less frequently), and it is easy to forget even basic skills during your time off from diving. Honing your skills in sheltered water is a good way to stay proficient.

PERFECTING BUOYANCY

Beginners, occasional divers, and even those who dive regularly will benefit from practicing key skills in a pool or sheltered water. Buoyancy is the most fundamental diving skill of all, so it is a good place to start when reviewing your technique. Refamiliarize yourself with buoyancy control with "fin pivots," using inhalations and exhalations to make the upper part of your body rise and fall. Keep your fins on the bottom and your legs apart. They will act as the "hinge" of your upward and downward movements. The value of this exercise is that it trains you to use breathing to adjust your buoyancy, rather than controlling it only with your BC. Also, try swimming through a hoop with your arms folded—this will test your buoyancy control and give you an idea of the space you can squeeze through with your equipment on.

1 **To practice fin pivots,** breathe in gently and feel your upper body rise, keeping your fin tips on the floor. Stop at a 45° angle to the bottom.

2 **Slowly release the air** in your lungs until your body begins falling. Then, before you hit the bottom, begin inhaling again.

CONFIDENCE-BUILDING EXERCISES

Mask clearing is one of the most basic, yet important, skills in diving. Building coordination and confidence by removing and replacing your mask, and even performing tasks without it, will improve your diving. It will also prepare you for minor upsets like your mask being accidentally dislodged. Begin building up your confidence slowly by removing and replacing your mask in shallow water, before repeating the exercise in deeper water. Repeat while swimming, and then try it with your regulator removed. You can also try swapping masks with your buddy. This is simply a variation on the solo exercises but also helps you to develop useful cooperative skills.

Giving yourself other handicaps can also help improve your confidence underwater. Try swimming with just one fin, for example. Losing a fin can happen on an open-water dive, but if you are already prepared, you will be able to cope with the situation calmly and effectively. Practice all of these exercises in a pool or sheltered water.

REDUCED VISIBILITY
Most divers feel uncomfortable swimming with their mask removed, but with practice, you will become confident performing tasks without it.

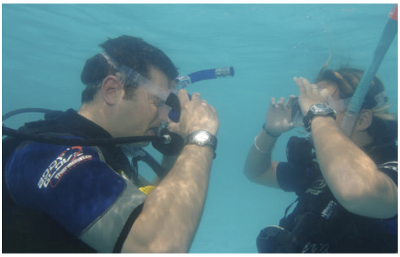

1 **To exchange masks,** kneel on the bottom facing each other. Give each other the “OK” signal, then remove your masks, while continuing to breathe normally.

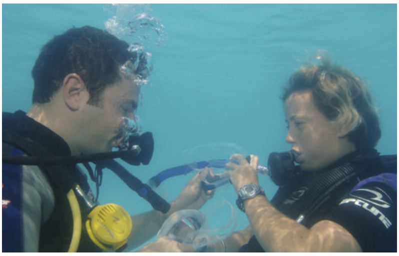

2 **Swap masks** with your partner. It is much easier to keep your eyes open throughout the exchange, but perfecting this by touch, with your eyes closed, is a very useful exercise.

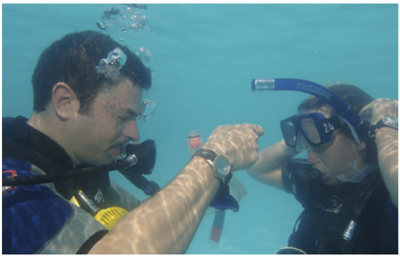

3 **Don each other’s masks,** and blow them clear (*see p.139*). Ignore any urge you feel to hurry when putting the mask on. In particular, make sure the strap is not twisted.

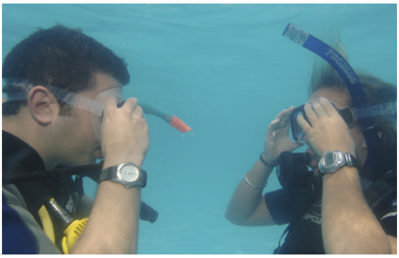

4 **Adjust the mask** for comfort, then open your eyes and signal OK to indicate a successful swap. The masks can then be swapped back again for further practice.

VARYING THE ROUTINE

Skills exercises don't have to be simple drills. Introducing an element of fun or setting an objective can motivate you to train for longer. Games for the pool or sheltered water include "hide the mask," which forces you to navigate and search without your mask; or you can practice breath control by performing simple exercises without breathing apparatus, such as retrieving objects from the bottom, or swimming through a series of hoops, extending the course as your stamina improves.

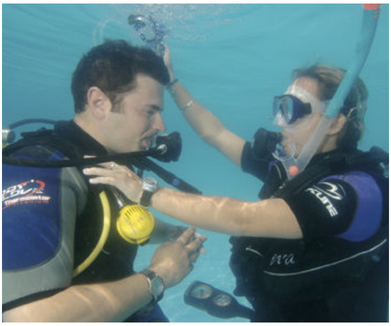

1 **"Hide the mask"** begins with an "OK" signal from the seeker, once the diver who is going to hide the mask has removed it.

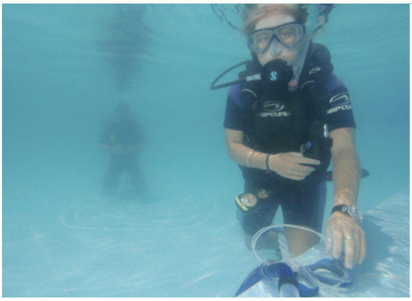

2 **The seeker closes his eyes** while the hider finds a corner of the pool to put the mask in, and returns to signal the seeker to start searching.

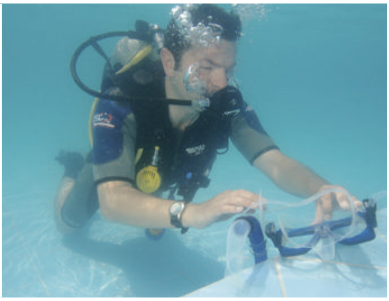

3 **The seeker opens his eyes** and starts to methodically search the pool bottom until the mask is found and replaced.

REMOVING AND REPLACING GEAR

Familiarity with your basic gear is a vital skill. One exercise that helps to foster confidence in this area is to remove your scuba unit—that is, your BC and breathing apparatus, all fully connected—and put it back on underwater, keeping your regulator in your mouth throughout. This exercise should be carried out at the bottom of a pool or in sheltered water, and if you are a beginner or haven't tried this before, it may be useful to have an instructor present who can guide you through the process. Start by undoing all fastening clips on the front of your BC and pull your left arm out of it first (even if you are left-handed), using your right arm to pull it around to your right. The BC jacket should now be in front of you, between you and the tank. Keep the hose for your regulator second stage between your arms (otherwise it may get caught under your shoulder strap when you put the unit back on). Then put the scuba unit back on by reversing the operation.

For an extra challenge, once you have removed your BC, try removing your regulator from your mouth and swimming away to a distance of about 30 ft (10 m), before returning and putting your gear back on again.

Alternatively, try entering the water without your gear, then putting it on while treading water, keeping your head above the surface. This is a much more difficult exercise, and should only be carried out under the guidance of an instructor.

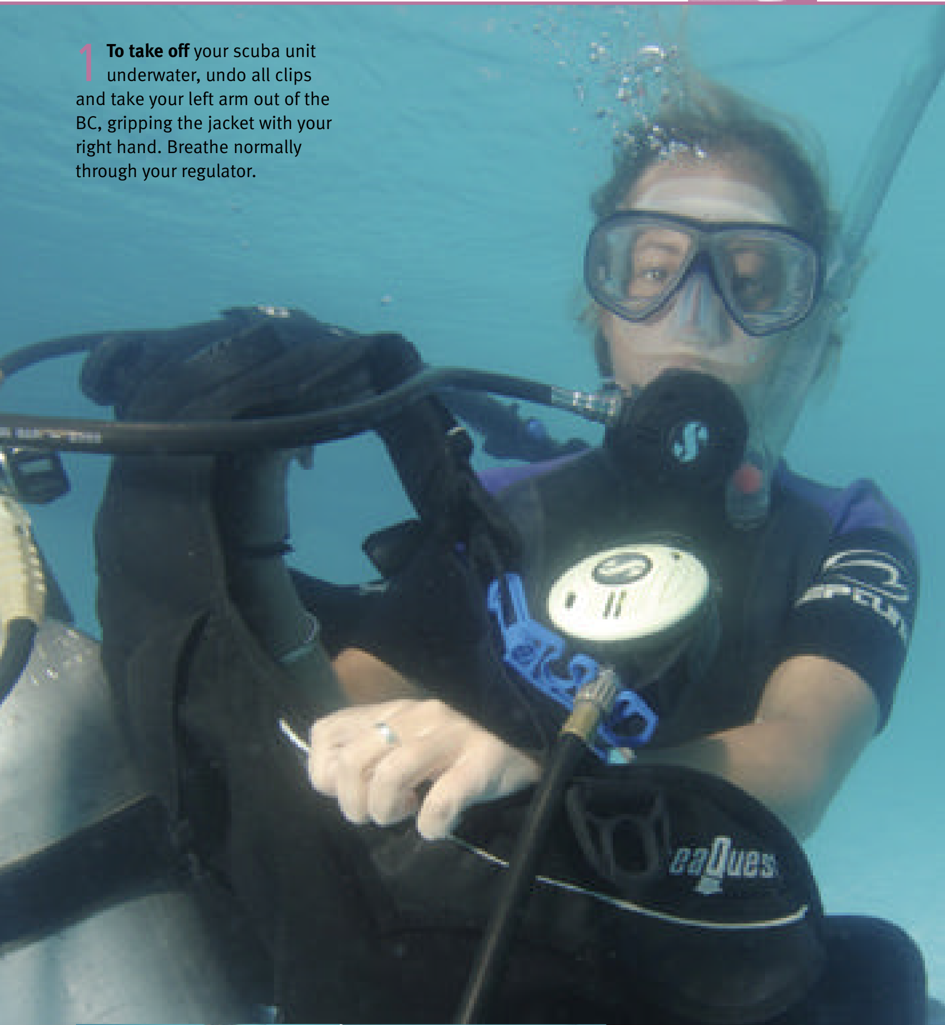

1 **To take off** your scuba unit underwater, undo all clips and take your left arm out of the BC, gripping the jacket with your right hand. Breathe normally through your regulator.

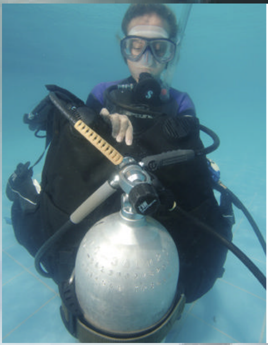

2 **Pull the whole unit around** in front of you. Keep a firm grip on your BC to avoid losing hold of the scuba unit.

3 **To put gear back on,** reverse the operation, making sure all clips and fastenings are reconnected.

Dive fitness

As a diver, you are reliant on your body as well as your gear for a safe and enjoyable dive—so make sure both are well maintained. Diving regularly helps keep you in shape, but it is important to maintain your fitness levels if you go for long periods without diving.

HOW FITNESS AFFECTS DIVING

Staying in shape reduces some of the risks associated with diving, and makes it more enjoyable. Having healthy lungs and an efficient circulatory system means you will use less air, so dives last longer. A good level of cardiovascular fitness will also help to prevent panic attacks, which can be triggered by a buildup of carbon dioxide in the blood. A fit body will expel this by-product of respiration more efficiently, reducing that risk. General strength is required for lifting equipment and other incidental tasks. Maintaining a healthy weight for your size is also advisable. When you dive, nitrogen builds up in the tissues of your body with each intake of breath, and fat retains nitrogen longer than all other tissue types. Therefore, if you are overweight and breathe rapidly because you are out of shape, you expose yourself to an increased risk of DCS (*see pp.142–43*).

Weight loss and gain can also alter your buoyancy. Fat is inherently buoyant, so if you have put on or lost a lot of weight since your last dive, you must adjust how much lead weight you carry as ballast (*see pp.118–21*).

TALKING THINGS THROUGH
If you have any injuries or physically weak areas, let those diving with you know before you dive. Don't be afraid to ask for assistance if this will help prevent injury.

DEVELOPING A FITNESS REGIMEN

There is no quick fix for getting and staying fit. Aim to exercise regularly, tailoring your routine to strengthen the muscles you use most during diving: those of your back, legs, shoulders, and arms. If you do not already exercise regularly, seek medical advice before embarking on a fitness routine. Start gently with easy exercises, and build up slowly. A routine based on the exercises overleaf could start, for example, with ten repetitions of each, building up to 50, three times a week. You should also aim to undertake regular cardiovascular exercise (such as walking or cycling), for 30 minutes two or three times a week to improve your general fitness and respiratory health.

Whatever exercise routine you decide to adopt, try to incorporate it into your normal daily routine wherever possible. A walk to work every day does more good than a monthly gym session. Never exercise after diving. This is a time when it is important not to do any strenuous activities—there is still residual nitrogen in your body, and this may form bubbles if you exercise too hard. Do some gentle stretches and take it easy for the rest of the day. Avoid hot baths or showers after diving, since these can also raise your vulnerability to DCS.

GOING AGAINST THE FLOW
Finning into strong currents places demands on your leg muscles and lungs.

PROTECT YOUR BACK

Be very careful when carrying and loading heavy equipment. A poor lifting technique can result in pulled muscles, and possibly long-term back problems. A back injury can spoil a trip, but diving with a weak back can also be a safety risk: surface maneuvers, climbing onto boats, and exiting surf, for example, are all compromised by weakness in this area. If your lower back is vulnerable, consider transferring your weights from a belt to a harness (*see p.70*).

Loading dive equipment

WARMING UP

Always warm up and stretch—especially your thighs, calves, and upper and lower back—before exercising, and spend a few minutes gently stretching and cooling down afterward. Many sporting injuries are caused by failure to warm up properly, so treat it as an important part of your fitness routine.

ARM AND LEG EXERCISES

Basic squats build up strength in your legs. Standing upright with your feet shoulder-width apart and your knees soft, place your hands on your hips and shift the weight back into your heels. Keeping your head and chest pointing forward and your back straight, bend at the hip and knees, as if you were sitting down onto a chair. Keep your heels on the floor; you should be able to move your toes, since all your weight is on your heels. Straighten and repeat.

BASIC PUSH-UP
The push-up is a great all-around exercise for the upper body and arms. Beginners can start with their knees resting on the ground.

Lunges are a more advanced exercise for legs. Stand upright with one foot a good stride in front of the other, with your hips parallel and your toes forward. Your back heel should be off the ground and your front foot flat on the floor. Both knees should be bent and your back straight. Drop your body weight toward the floor until your front knee is bent at a 90° angle. You may have to tilt your pelvis forward to get this position right. Rise up and repeat.

Biceps curls strengthen the biceps—muscles in the upper arms that are used whenever you lift anything. Using hand weights or a weighted bar, stand upright,

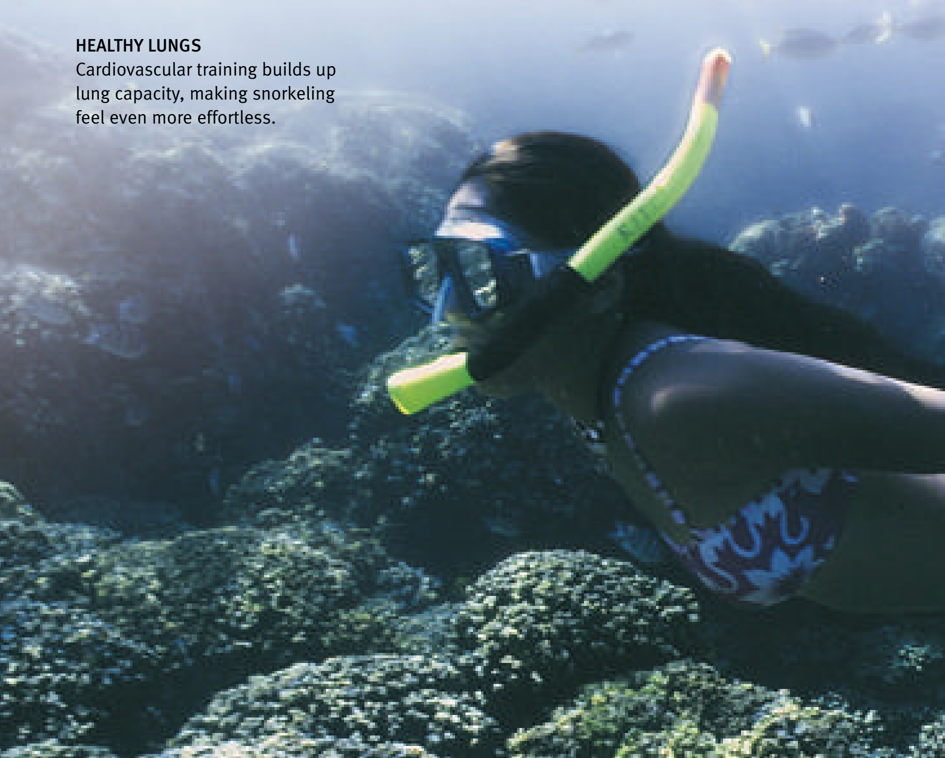

HEALTHY LUNGS
Cardiovascular training builds up lung capacity, making snorkeling feel even more effortless.

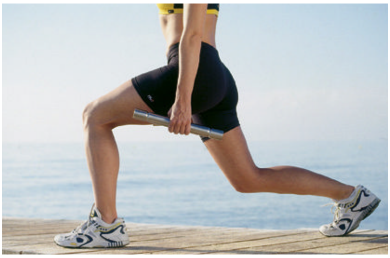

LUNGE EXERCISE
Practicing lunges in front of a mirror will help you to get the position of your front leg right: knee directly over foot, bent to a right angle.

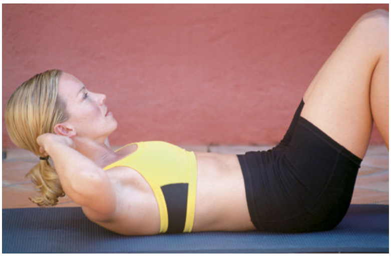

ABDOMINAL CRUNCH
Always keep your lower back flat on the floor and never arched. For a harder workout, aim your right elbow to your left knee, and vice versa.

with your back straight and knees slightly soft, and hold the weights in your hands with your palms facing upward. Without rocking or moving your shoulders, bend the arms toward the chest and then lower, taking care not to lock the elbows.

EXERCISING YOUR TORSO

Push-ups strengthen your shoulders, upper body, and core muscles. Lying on your front, bend your knees, curl your toes under your feet, and place your hands shoulder-width apart below your shoulders. Press your body weight up and down from this position. Keep your back straight at all times and your elbows pointing out to the sides.

Crunches strengthen abdominal muscles, which will help protect your lower back. Lying on your back, bend your knees slightly and place the soles of your feet on the floor. Place your fingertips behind your ears and lift your upper torso and shoulders off the floor. Hold this for three to five seconds while looking forward, then lower and repeat.

HYDRATION AND NUTRITION

During a dive, you can burn as much energy as you would playing an aerobic sport, such as squash or badminton. The precise amount will vary according to your fitness, the temperature of the water, and any prevailing currents. But while it is tempting to stoke up on a big fried breakfast in preparation for a day's diving, better food choices before a dive are complex carbohydrates, which will keep you feeling fuller for longer and provide a steady supply of blood sugar to maximize energy levels while you are diving. Some foods may make you feel uncomfortable on a dive—commonly, greasy or hot and spicy dishes—so these are best avoided.

After a dive, you need to replace the energy you have burned with foods that release energy into your blood relatively quickly, such as refined carbohydrates. Have something to eat within an hour of surfacing, if possible. Energy-replacement drinks can be a quick fix, but avoid those with a high caffeine content.

GOOD DIVING FOODS
Whole-wheat bread, legumes, and unrefined cereals and pasta release energy slowly during a dive. Dried fruit and nuts make good snacks.

Above all, the most important thing is to remain well hydrated before and after a dive. Your physical performance will be reduced if you become even mildly dehydrated, and this is a big issue if you are diving in warm waters where the outside temperature is also high. Aim to drink at least 2 quarts (liters) of water during the course of the day (more if it is hot). This will help to reduce muscle cramps and fatigue, and also helps decrease the risk of DCS.

AVOIDING PROBLEMS

It is important to get plenty of sleep before and after diving, to give your body a chance to recharge. Diving is physically demanding, and if you are run down, you are more susceptible to catching colds. If you do have a cough or a cold, you should not dive, because some symptoms of respiratory illness can prevent you from being able to equalize and may increase your vulnerability to DCS (*see pp.100–01*). Avoid taking decongestants, since these can impair your performance underwater.

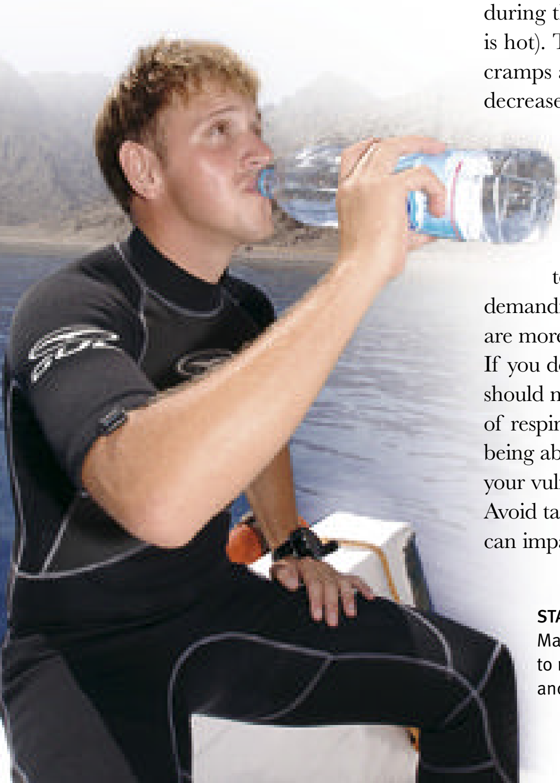

STAYING HYDRATED
Make sure you drink plenty of still water to maintain good hydration. Tea, coffee, and carbonated drinks are no substitute.

FOODS TO EAT WHEN DIVING

The following list gives examples of the kinds of foods it is good to include in meals and snacks before and after a dive:

BEFORE A DIVE Multigrain, whole-wheat, or rye bread; whole-wheat pasta; sweet potatoes; brown rice; oatmeal; bran-based cereals; legumes and nuts; beans; leafy vegetables; and plain yogurt.

AFTER A DIVE Pasta and rice (ideally in a salad or low-fat sauce); potatoes; fresh fruit; white bread or bagels.

So can alcoholic drinks, which are best avoided the night before a dive because they can also contribute to dehydration. Divers sometimes experience minor ear infections, especially after days of continuous diving or if diving in slightly polluted waters. There are various ways of avoiding ear problems, including rinsing the ears with fresh water after every dive, or using an alcohol-based fluid designed for swimmer's ear.

Lastly, you should have a regular dive medical to check that you are in good enough shape to dive. If, between medicals, there have been any changes in your health since you last dived, then it is always best to have a checkup.

AFTER A DIVE
It's tempting to head for the bar after a day's diving, but have something to eat, too. Even standing still in cold water burns calories.

Going diving

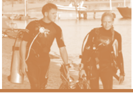

Introduction

Becoming a qualified diver is a tremendously exciting experience, but the range of choices facing you post-qualification can be daunting. The joy of diving is that within one sport lies a host of different specialties, all requiring different skills, knowledge, and planning techniques.

RECREATIONAL DIVING

Upon qualification, the world of diving opens up to you. Whether you become a casual resort diver or an advanced technical diver with a liking for the more extreme elements of the sport, only time and experience will tell.

The broad spectrum of recreational diving encompasses many different environments and experiences, including reefs, caves and caverns, wreck diving, ice diving, freshwater diving, and night diving. There are also various specialty activities in which you can participate, such as filming and photography, wildlife and conservation diving, underwater archaeology, and free diving (diving without an air supply). The majority of divers find that they derive most enjoyment from sampling a range of different environments and activities, rather than limiting themselves to one specific aspect of diving.

BEACH BRIEFING
Becoming confident enough to conduct your own dives without a guide or instructor gives you greater freedom. Always plan and brief dives thoroughly before entering the water.

Many new divers aim to acquire further specializations, pins, and qualifications from the moment they pass their initial training course. This urge, while understandable, should be tempered by the need to gain general diving experience. There is no substitute for time spent in the water—what was once intimidating becomes familiar, helpful routines start to emerge, skills and safety consciousness improve, and knowledge increases. You will also feel more confident and comfortable in the water as you learn to tailor your equipment to your own needs.

Only when you have served your "apprenticeship" by gaining experience of different environments and diving skills should you make any decisions about future specializations. Without putting in the hours underwater, you will find further training and courses more difficult, no matter how well structured and delivered they are. In addition, the more experience you have, the more aware you will be of your own likes, dislikes, and aptitudes, allowing you to make more informed choices for the future direction of your diving.

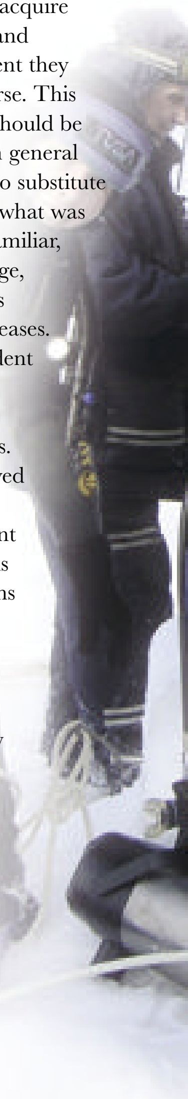

Fortunately, gaining general diving experience is relatively easy, with dive groups and clubs springing up in abundance to meet the sport's growing appeal. Furthermore, dive travel is so well established that almost any region of the world is now accessible, and opportunities to dive exist in virtually every coastal resort around the world. A whole range of diving encounters are available—from a casual beach dive through to full-blown expeditions—all easy to arrange via the Internet or through dedicated dive travel agencies.

PLOTTING YOUR FUTURE COURSE
The more experience you gain, the better informed you will be when deciding whether or not to pursue a particular aspect of diving.

PROFESSIONAL DIVING

The enjoyment of diving leads some people to consider a full-time career in which diving plays an essential part. Diving professionals range from recreational instructors and dive masters, whose main role is to introduce others to diving, through to military experts, commercial specialists, scientific divers, fishermen, and underwater cameramen and photographers.

Further skills are always needed to succeed in your chosen career—an underwater filmmaker, for example, must be an accomplished camera operator as well as a superb diver, and an instructor needs good communication and teaching skills, in addition to first-class abilities in the water.

Professional diving can offer excellent rewards, both personal and financial, but they tend to be well earned. Professionals must dive in all conditions and at all times, and the work is often arduous and sometimes dangerous.

PREPARING FOR AN ICE DIVE
Diving under ice is thrilling, but it is not for beginners. The hazards of enclosed surroundings and extreme cold mean that special training is required.

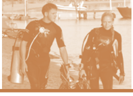

Identifying your aims

Time spent researching a dive is time well spent. Not only does it help you to unearth new attractions and clarify the aims of the dive, but it is also likely to heighten anticipation, enhance safety, and maximize the pleasure and interest you derive from the trip.

ASSESSING AIMS

The nature of dive research will vary from trip to trip, but certain core issues should always be taken into consideration.

Identifying the divers is obviously one of the first items to be addressed. Diving is not, as a rule, an individual sport, so of necessity others will be involved in the dive itself. The experience and qualifications of divers within a group will vary, and what may be a comfortable dive for one person may be extreme for another. Time spent discussing qualifications and experience levels may well prove invaluable in establishing the ideal dive location and activity for the group as a whole.

Along with making sure that you have the right mix of divers, you will need to ensure that their aims are not likely to conflict. Imagine the scenario, for example, of an avid photographer preparing to take an

BEGINNING RESEARCH
Use magazines, books, and maps to highlight basic issues that you can then investigate in depth on the Internet.

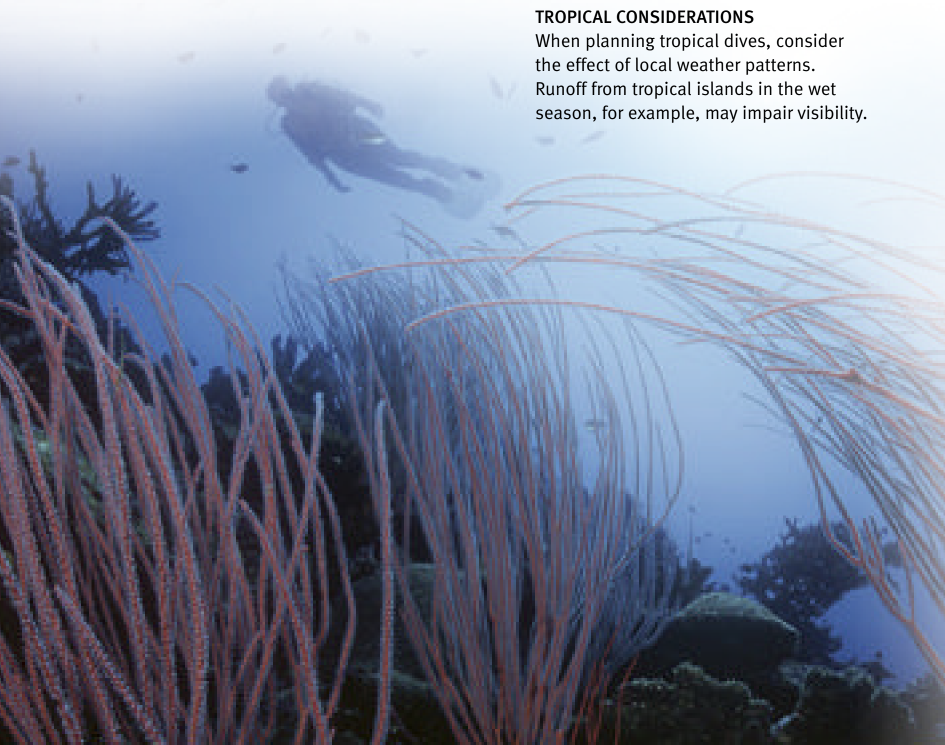

TROPICAL CONSIDERATIONS
When planning tropical dives, consider the effect of local weather patterns. Runoff from tropical islands in the wet season, for example, may impair visibility.

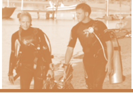

DIVING WEBSITE
The Internet is an invaluable research resource, allowing you to access the opinions and advice of authoritative dive organizations.

award-winning image of a sunken ship, when an impatient wreck enthusiast swims in front of the camera lens and ruins the shot. Establishing the main aim of the dive at the outset, or even allocating buddy teams with the same interests, should prevent such problems.

When diving with a local operator, try to check out their credentials. In many parts of the world there are associations set up by groups of respectable operators to govern the local dive industry. Failing that, many tourist boards and even government agencies require dive operators to register—it is always worth getting an independent view of your potential host from these organizations.

LOCATION AND TIME

Conditions can vary greatly along even the most bland section of coastline or within the most innocuous-looking body of water. Choosing an exact site in advance has the benefit of allowing very specific research into local issues. Key questions to be answered include: what are the hazards involved? Where are the established entry and exit points? What are the divers likely to encounter during the dive? What is the tidal range? What are the prevailing currents?

The timing of the trip is also important. Tides, for example, may affect not only visibility, but also the marine life present, as well as entry and exit considerations. Marine life will also be influenced by seasonal variations, as will water temperatures. Furthermore, a different range of marine animals may well be present at night than during the day, while boat traffic and use of the area for other watersports will vary throughout the day and the week. Diving at the wrong time can make the event a disappointing experience.

Safety is the most important area of your research. You will need to assess the conditions you are likely to encounter—waves, currents, depth, temperature—and the facilities available to you should an accident occur. You should not show up at a dive site unless you know the exact location of the nearest hospital, recompression chamber, and casualty evacuation facilities. Knowing how to contact these agencies in an emergency is also crucial—researching phone numbers in the comfort of your own home is infinitely better than trying to do it on a rocking boat while caring for an injured diver.

UNDER THE ICE
The possibilities for thrilling diving trips around the world are almost endless, limited only by time, budget, and imagination.

Planning a trip

Armed with the fruits of your research, you will have a clear idea of where you want to dive, when you want to go, who you will be diving with, and what you will see when you get there. The final task before your departure is to plan the trip in as much detail as possible.

REALISTIC TIMETABLING

Most trips revolve around a key moment in time—the instant when the divers enter the water. This is often determined by a certain phase in the tide when the diving is most favorable, or by the desire to dive at a specific time of the day or night, such as when the Sun is high for photography, or in the gathering dusk when the reef is at its most active. A realistic, workable timetable is thus vital to achieve such time-specific goals.

Devising a timetable, such as the chart below, by working backward from the moment you aim to enter the water is the best way to plan your diving day. This method is used by many divers to ensure that their diving is as stress-free and as enjoyable as possible.

PACKING YOUR EQUIPMENT Checklists and routines will help to ensure that all your gear makes it to the dive site.

PRESSURE TEST

Before you start packing your equipment for a dive, carry out a pressure test. This involves rigging the gear to a tank, and ensuring that the seals and regulators are all functioning as they ought to be. This is also a useful means of ensuring that your tank is full prior to departing for the dive site. There is no more sickening sensation than journeying to a dive site, looking out over perfect conditions, assembling your dive gear, and turning on the air to be greeted by a cacophony of hissing regulators and seals—or, indeed, the tank itself being empty.

REVERSE TIMETABLING

A diver knows they must enter the water at 3:00 pm to dive during slack water between the tides. They work backward from this to calculate a realistic departure hour, building in a contingency to ensure arrival at the dive site in good time.

Activity	Time Taken	Actual time
Entering the water		3:00 pm
Suiting up and buddy checks	25 min	2:35 pm
Site inspection and briefing	40 min	1:55 pm
Arrival and parking	10 min	1:45 pm
Driving to site	1 hour	12:45 pm
Loading car	30 min	12:15 pm
Checking and packing dive gear	30 min	11:45 am
Contingency time	30 min	11:15 am

CAVE DIVING

For more complicated trips, such as cave dives, special equipment may be necessary, and this must be factored into your planning.

PACKING WITH CARE

The greatest cause of the postponement or cancellation of dives is forgetting pieces of gear. A fool-proof method of guaranteeing that you pack all of your equipment is to load your gear into your dive bag at home in the order that you will be putting it on at the dive site. Visualizing your arrival at the site and suiting up will lead you to load your dive bag in the following sequence: suit, boots, knife, weight belt, tank, BC, regulator, computer, hood, mask, snorkel, fins, gloves, and finally any ancillary items (safety sausage or DSMB, flashlight, slate, and so on).

An alternative packing technique is to run through a checklist from head to toe, and again pack your gear in that order. The list then runs: hood, mask, snorkel, suit, regulator, BC, tank, computer, gloves, weight belt, knife, boots, fins.

There are innumerable ways of trying to ensure that no item of equipment is forgotten, but the most effective are those that are convenient, quickly become a habit, and take little time.

DIVERS ENTERING THE WATER

A realistic timetable will ensure that suiting up and briefing on site is not rushed, and that you do not miss your window of opportunity with the tide.

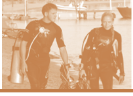

Site-specific planning

Pre-dive research is important, but there is no substitute for a thorough assessment of the site on the day of the dive. This may highlight issues not raised by your research, and allow you to make adjustments to accommodate changes in conditions before the dive starts.

FINALIZING YOUR PLAN

Once you know when and where you are going to dive, you must devise a practical plan for each of the dives you mean to conduct on the day. Whether you are planning for just yourself and a friend, or for a larger group, always take local conditions into account before settling on a final dive plan.

You must address a number of key points in your dive plan. First, it should broadly define the aim of the dive, the dive's entrance and exit points, and its general route, as well as describing an alternative exit in case the main choice is inaccessible. It should also define the dive's maximum depth and time, as well as decompression restrictions, and should specify the minimum reserve of air that divers must surface with at the conclusion of the dive.

Depending on the aim of the dive and the experience of the group, special signals may be required. It is imperative to select diver recall systems—the signals or sounds that will be used to indicate that the dive should be aborted. Your plans for the dive should also identify any potential hazards that may be encountered. If an incident occurs, what procedures should be followed? Define what action to take if divers become separated—in murky waters, this is one of the most common reasons for dives being aborted. Make sure that the divers you intend to pair together as buddies share similar interests or levels of experience, so that there are no conflicts of interest during the dive.

INSPECTING THE DIVE SITE
Resist the temptation to enter the water as quickly as possible. Assess the conditions and inspect the site, noting the dive's entry and exit points.

UNLOADING THE GEAR
Even if you feel sure that every item of gear was packed before the trip, double-check on site that you have everything you need.

Assess local conditions and examine weather reports; worsening weather conditions might dictate a move to a more sheltered site nearby, or rule out the use of a boat. Finally, ensure that the designated shore marshal or dive supervisor understands their duties, including recording tank pressures and the times that divers enter the water, and responding to incidents. Provide them with contact details of local agencies to be informed in the event of an emergency.

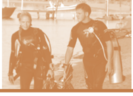

OUTLINING THE PLAN
If it helps, you can always draw simple sand maps to clarify the route to the other divers.

THE DIVE BRIEFING

The final task is to brief the entire dive crew. The briefing, which should ideally take place overlooking the dive site, must cover all aspects of the dive plan, and include not only divers, but also shore marshals, deck hands, and anyone else associated with the dive. At the end of the briefing, ensure that all the divers clearly understand the constraints of the dive plan, and are fully prepared for the dive. After any final questions from the group, the dive can begin.

The simple maxim "plan the dive, and dive the plan" is an old one, but its worth has been proven over time. Following it means that all of the divers will get the most they can from the dive, and, importantly, will enjoy it in safety.

ENTERING THE WATER
By the time you enter the water, everyone involved with the dive should be fully aware of the dive parameters.

Wreck diving

Exploring wrecks is—for many people—the main reason to dive. This activity is particularly popular in countries without coral reefs to dive on. However, it also has a general appeal, because wrecks give a historical context for a dive and demand mastery of special skills.

DIVING ON WRECKS

Sunken ships have an obvious attraction as sites of archaeological and historical interest. A wreck is a time capsule from a given period, and the waters around most countries contain thousands of them.

Although most known wrecks are metal-hulled vessels, unusual conditions will sometimes preserve wooden-hulled ships dating back hundreds of years. These provide an excellent opportunity to examine past shipbuilding techniques, and sometimes hold interesting and unusual artifacts. To explore significant wrecks, you should be properly trained and, preferably, work within an archaeological organization.

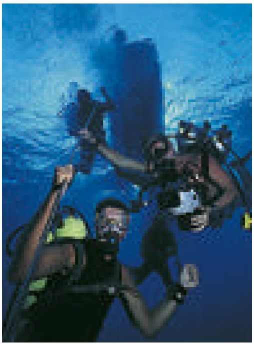

DESCENDING A GUIDELINE
Popular wrecks often have a buoy moored to their structure to make location and descent easier.

Information on recent wrecks, such as warships sunk during the two world wars, is usually readily accessible. It is often possible to research the specifications and history of the vessel, along with details of its crew and mission. Historical records may give details of the events surrounding the sinking of the ship.

Wrecks also make for good diving because they act as artificial reefs, attracting all sorts of marine life. Over time, the structure becomes encrusted with marine growth, especially filter-feeders, if it lies in a tidal stream. Fish exploit the safety and shelter that wrecks offer: smaller fish tend to shoal around wrecks; larger fish, such as moray or conger eels, can hide safely in nooks and crannies; and predators, such as sharks, may be found in the wreck's surroundings, because they feed off the smaller fish.

PREPARATION AND SKILLS

Wreck diving can be a fascinating activity, but it is not without risks. Under no circumstances go inside a wreck if you lack appropriate training and equipment; the danger of entrapment is very real, and specialist knowledge and skills are required to enter wrecks safely. Exploring the outside of wrecks is safer, but always treat such sites with respect.

TANK IN TRUK LAGOON, SOUTH PACIFIC
Wreck diving is not just about ships; crashed aircraft and military vehicles carried by sunken vessels provide equally enthralling alternatives.

SUNKEN LEVIATHAN
The sight of a major wreck, such as the *Giannis D* in the Red Sea, rarely fails to send shivers down the spine.

DIVING IN CONFINED SPACES
Wreck penetration often involves exploring in dark, confined spaces. It is not recommended for those with claustrophobic tendencies.

FINNING TECHNIQUE

Wrecks often contain a lot of static silt and mud, which is easily disturbed by changes in water currents caused by divers finning overhead. Good visibility can suddenly and dramatically change to almost zero, so using a frog kick (*see p.123*), which generates relatively little turbulence, is recommended to minimize disruption of the silt. The best time for wreck diving is at slack water, but a small current can help sweep away any disturbed silt.

USING GUIDELINES

When you arrive at the dive site, ensure that your boat cover is moored to the wreck's marker buoy. If it doesn't have one, an experienced diver needs to swim down and attach a guideline directly to the wreck, to act as a guide for other divers to follow. If you intend to return to the surface via the same guideline, it can be useful to use a distance line (*see p.127*) to help you find your way back to the bottom of the guideline, especially in poor visibility. If you plan to ascend from another part of the wreck, deploy an inflatable marker buoy (*see pp.130–31*).

PENETRATING WRECKS

When wreck diving, you should carry backups of certain pieces of gear, such as your reel and your flashlight, as well as an independent air source, such as a pony bottle (*see p.68*), in case of emergencies. You must also ensure that you follow the "Rule of Thirds" (*see p.117*) with your air consumption—a third of your air is used to get to the site and explore, a third

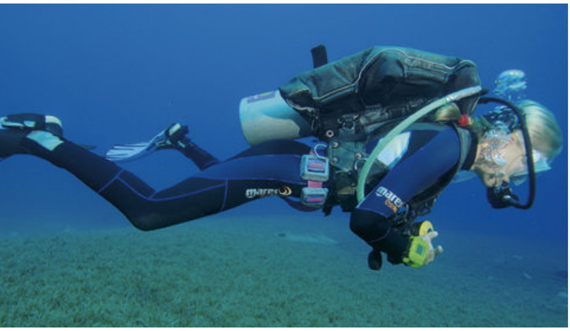

LEVEL POSITIONING
The frog kick is a gentle swimming stoke that mainly causes turbulence in the water behind and not below you, and so does not kick up silt.

to return to the point of entry, and a third is left in reserve as a contingency. Allow extra time at the end of the dive to find the guideline, and do not take any unnecessary risks with your dive time, depth, or penetration.

EXPLORATION TIPS

It is a good idea to investigate the area around a wreck, since artifacts may have spread over time and pieces of wreckage may have been knocked off that might be of interest, but take great care. Always be aware that the structure (both inside and outside a wreck) may be much degraded and potentially dangerous. Wear protective gloves, even in warm waters, as corroded metal can be sharp. Wrecks are often heavily fished, so carry a knife in case you become entangled in fishing line. You also need a powerful flashlight (and a backup) to illuminate your way in poor visibility, and to signal to your buddy. In some very enclosed wrecks, it is advisable to wear a helmet and head lamp.

DIVER USING WRECK KIT
Carrying extra air and nitrox is a failsafe, and can help to extend dive times and shorten decompression stops. Good light sources and dive plan information are also essential.

THE GOLDEN RULE

Never remove items from a wreck. As with any artifact, marine finds are more useful to archaeologists when examined *in situ*, and their context provides more information about the wreck than if they are handed in without details of where they were located. Worse still is if they are lost forever in a diver's personal artifact collection.

US troop ship artifacts

RESPONSIBILITIES WHEN DIVING

Wrecks are historical sites, and just as you would not remove artifacts from a historical site on land, such as a castle or monument, you should also respect a wreck and not damage it in any way. Be considerate of the enjoyment of others who may dive the wreck after you. Dive responsibly and with care, and do not remove anything without being given permission to do so. When you find an artifact, such as a ship's bell, after a long, hard search, you may feel that you deserve to be able to take it away as a trophy, but all such discoveries should remain where they were found, or be raised by trained archaeologists for preservation in a museum.

WOODEN-HULLED VESSELS

Most of the wrecks that you are likely to explore will be metal-hulled, but wooden ships are also preserved under some conditions. Avoid touching the wooden structure, as it is likely to be extremely fragile. Historic wooden vessels may have been lying protected under sand or silt for many hundreds of years. If you want to expose parts of the wreckage for any reason, you should do so by wafting your hand over it to remove the silt, creating a current to take the silt away.

LIVE AMMUNITION

Wrecks from the two world wars can provide fascinating dive experiences for military history buffs, but remember at all times that they may also contain live ammunition that could potentially be very dangerous. Never touch anything that resembles a shell or mine, and research what you are likely to find on the wreck before diving.

ARTIFICIAL REEFS

At some popular dive sites, wrecks have been sunk deliberately to provide a safe, interesting dive at a reasonable depth for recreational divers. Such wrecks quickly become colonized by marine life and provide a fantastic opportunity for diving and marine research. They are generally a good place to start wreck diving, since the structure will be intact and any hazards will have been removed. At some wreck sites there are special diving trails with underwater display boards to

WRECKED BY DESIGN
The former US Coastguard cutter *Keith Tibbetts* was sunk deliberately at Cayman Brac to provide an artificial reef for divers to explore.

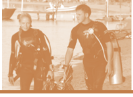

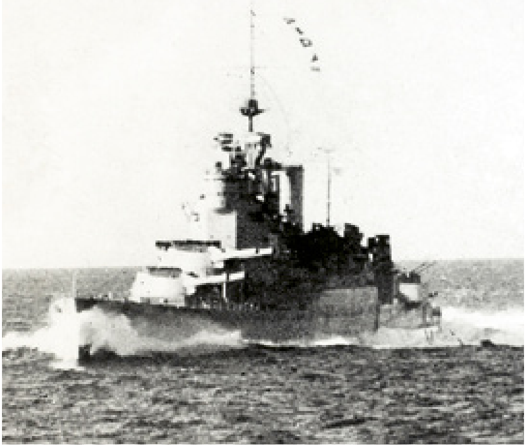

explain the layout. There may also be laminated booklets that you can take with you on the dive, which enable you to read about the most interesting features of the wreck while on site.

VIRGIN WRECKS

Every wreck diver's dream is to find a previously undiscovered wreck—and there are certainly still plenty to be found. Before you embark on such a search, you should familiarize yourself with the law regarding wrecks, and any wreck you find should be reported to the authorities. Generally speaking, you can dive any wreck around the world apart from those that are designated as historic wrecks and war graves, which are restricted and require special permission to dive.

Locating a virgin wreck can be a difficult, lengthy task, and is likely to be expensive. But a modern hardboat with an experienced skipper and high-tech equipment such as an echo sounder, 3-D imaging of the seabed, and a GPS system, combined with charts and some good background research on your part, could help you hit the jackpot. Many undiscovered wrecks are found in very deep water, so make sure you are fully qualified and appropriately trained before undertaking deep wreck dives.

RESPECTING HISTORY
Many wrecks sunk during conflict are official war graves and may not be dived.

FINDING OUT MORE

If your interest in a wreck goes beyond merely diving it and you wish to enhance your dive by finding out more about the wreck's history, there are a number of available sources of information. Good starting points include books on wrecks in the area that you are diving, and using the Internet to do some background research. There are numerous websites and forums where you can discover more about wrecks, and also make contact with other divers interested in finding new wrecks. Furthermore, local museums may have displays of artifacts recovered from wrecks in the area, and local divers and boat skippers may be able to reveal information about the history of a wreck.

EXPLORING VIRGIN WRECKS

If you are searching for a particular undiscovered wreck, it is important to start your research in the archive. Studying the ship's specifications and plans will help you identify any remains that you find, and determine whether they belong to the subject of your search. It is a good idea to take video footage of anything you find on a dive, as this is a nonintrusive method of recording the remains, and the footage will be critical in identifying the wreck and for showing to experts.

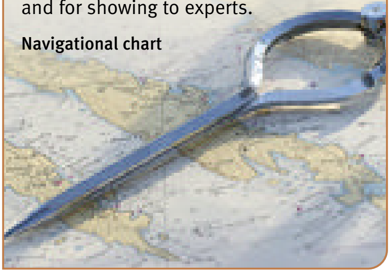

Navigational chart

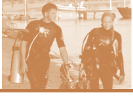

Nature and conservation diving

Recreational divers represent a huge pool of manpower that can potentially contribute to conservation work. Many divers take up the sport due to an inherent fascination with nature, and more and more are now applying their skills to organized conservation projects.

DIVING RESPONSIBLY

Most divers have a certain level of knowledge and enthusiasm for marine life and the ocean environment, and realize that along with the thrill of exploration comes responsibility. Divers are in a unique position to observe the deterioration of busy coastal margins around the world and, with the development of the modern ecotourism market, are now also able to actively contribute to conservation projects in the same regions.

At the most basic level, conservation work simply means responsible diving by anyone taking up the sport in the first place. Always adhere to a "no-touch" rule when diving delicate reef systems; you will find this vigorously enforced by most dive operations around the world. Good buoyancy control, streamlined equipment, and a respect for the varied inhabitants of the sea are now widely taught as the only way to dive, and as a result even heavily dived areas experience less habitat disturbance than they used to.

GOOD PREPARATION
A group of divers at a conservation site discuss the work they will be conducting.

MAKING A DIFFERENCE

Voluntary work, though often conducted at the most basic level, can nonetheless be extremely significant at a local and even international level. The efforts of one voluntary organization alone—Coral Cay Conservation—contributed greatly to the establishment of the Belize Barrier Reef (*see pp.316–17*) as a World Heritage Site, and have seen eight Marine Protected Areas (MPAs) established around the world in the last 20 years.

Formally observing and recording marine life is not just the preserve of the scientific expedition. Significant work has been done by enthusiastic amateurs. Organized conservation and monitoring projects range from casual one-day beach cleanups run by local diving clubs, to multi-national operations setting up nature

USING IDENTIFICATION SLATES
To familiarize themselves with marine life during a dive, divers use species identification slates—waterproof cards with pictures of the species likely to be found at the dive site.

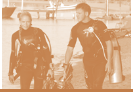

reserves throughout the world. Diving conservation is now big business, with hundreds of organizations around the world claiming (with varying levels of accuracy) to run projects that contribute to scientific knowledge or the establishment of marine reserves and protected areas. Most organizations put volunteers through an intense training period before allowing them to take part in even relatively simple sampling or monitoring programs.

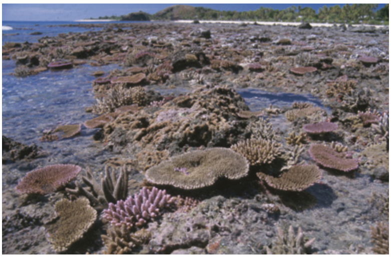

REEF EXPOSED
A healthy reef at low tide. Rising global temperatures are harming coral, and sights like this may become less common in future years.

FIRST PRINCIPLES

There are two principal considerations that must be borne in mind by anyone working as a diver on a marine conservation project. The first is safety. You need good diving skills, and enough experience to be able to keep a watchful eye on the ever-present factors of depth, time, and potential hazards while simultaneously making meticulous observations. Enthusiasm for the work being undertaken must not distract you

PRESERVING HABITATS
Carrying out conservation work not only allows you to give something back to the marine environment, but may give you the opportunity to visit exotic destinations.

from the normal rules of safe diving. The second factor is consideration for marine life. Unless you are working under the direction of a well-recognized conservation group, you must only gather data without touching or harming the marine environment. Both marine algae and animals should be recorded only *in situ*.

WORKING TOWARD OBJECTIVES

Good science starts with a clear means to identify the subject of the research. In the case of most amateur enthusiasts, this means reading up in accurate reference works and guides about the dive environment in question. A further, essential part of any observation program is the need to record data effectively, requiring not only a means of noting down results, but a framework of headings and categories that remain the same for the entire observation program. Only by combining these two important factors—accurate identification and consistent record-keeping—can you accomplish truly effective research and monitoring. Repeated observation of species populations within a given area, for example, can provide invaluable data for analysis by research groups. Your recordings may help local initiatives, or possibly go forward to form part of a much larger-scale project, in which

FINDING A PROJECT

Conservation, to paraphrase an old adage, begins at home. There are marine ecology programs in most countries, and interested divers can usually find groups in their area that contribute to protecting the local marine environment. Search the Internet and specialty diving magazines to find suitable organizations local to you.

CORAL MONITORING PROGRAM
Advanced or potentially invasive techniques should only be used under the direction of a recognized conservation group.

DIVING AT A MARINE RESERVE
Even a little research into how and why an area is protected will add food for thought to your enjoyment of the dive.

complex statistical techniques may be used to draw conclusions with far greater significance than you perhaps imagined when gathering your data.

EQUIPPED FOR THE JOB

Although this is the age of the computer and of instant global communications, most scientific expeditions still use some fairly basic equipment to record data in the field. Observation strategies can begin with nothing more than a sharp pair of eyes, an underwater slate, and a good guidebook. An additional item that can be extremely useful is a magnifying glass. This simple tool greatly enhances enjoyment and has become essential for many divers around the world when diving complex reefs.

Underwater photography and videography, if correctly documented, can provide potentially valuable information. In an organized program, you may receive training in advanced techniques, such as the use of catch bags, nets, and quadrats, or in ways to ensure that the target region is sampled effectively, such as GPS systems and datum lines. Effective sampling and observation does not necessarily require high-tech gadgets, but an effective strategy to ensure scientific integrity in the results you achieve is essential.

A CAREER IN CONSERVATION

Should you wish to develop your interest in conservation techniques, there are many courses and reference works available. Formal qualifications may be sought at undergraduate and postgraduate level, although demand for places is extremely high. Jobs in this area are also relatively difficult to come by. However, the rewards for persisting in either scientific or conservation work can be immense.

The loggerhead turtle, an endangered species

Photography and videography

Since the dawn of diving, there has been interest in capturing images of the remarkable ecosystems and animals of the sea. Although underwater photography was initially a complex process, advances in technology have seen it become an accessible, absorbing pastime.

EARLY PIONEERS

The development of underwater filming and photography has mirrored the evolution of diving itself. Early pioneers made their own waterproof housings for cameras, but the invention of the O-ring (enabling watertight hydraulic seals) in 1937 and the Aqua Lung in 1943 convinced camera companies that they should make and sell their own designs.

What prompted this was the huge appetite for images of the underwater world. The rapid growth of recreational diving, and growing interest in underwater films among movie audiences, spurred the development of underwater film and stills photography equipment. Although the first underwater film was made in 1939 by diving pioneer Hans Hass—an instructional documentary for spear-fishermen called *Stalking Underwater*—the first truly commercial underwater movie was made by Jacques Cousteau in 1956. His feature film *The Silent World* was an Oscar-winning sensation that established him as the world's premier underwater filmmaker and explorer.

MINUTE SUBJECTS
One of the most interesting aspects of underwater photography is macro (close-up) work. Tiny creatures can make dramatic subjects.

GROWING MOVEMENT

Until the early 1980s, underwater stills photography and filming was seen as the preserve of either the dedicated amateur or full-time professional diver. Underwater photography was brought to the masses with the development of the Nikonos series of amphibious cameras—culminating in the Nikonos V, still viewed by many as the best underwater camera ever made. Underwater filming did not become a popular pastime until the mid-1990s, when digital video cameras came onto the market, making moving pictures cheaper and technically easier for amateurs to produce. Digital imaging also revolutionized stills photography underwater. Small, easy-to-use digital cameras with

STALKING A SUBJECT
Photographing marine species underwater requires guile and a good understanding of animal behavior and habitats.

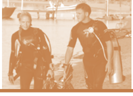

CAMERA CONTROL
Though cameras, especially digital models, have become smaller and easier to use, it still takes great skill to wield one effectively underwater, especially for video footage.

many automated functions have made it simpler for divers to record high-quality images of their underwater experiences.

Despite this, the jump to producing professional-standard material remains as great as ever. Light behaves very differently underwater than in air (*see p.183*), and there are myriad extra factors to take into consideration when aiming for truly professional results. The work that you see on the page or the screen may well have involved lengthy planning, complex equipment, and the specialized skills of a professional post-production studio.

However, the same principles that guide professionals can enhance amateur work. You must have knowledge of the underwater environment and the behavior of target animals, and possess impeccable diving skills. Add to this mix some tips on technique when photographing and filming underwater, and you will be well placed to get the best from the dazzling range of camera and video gear available to divers.

USING LIGHT CREATIVELY
An understanding of how light behaves underwater is essential to capture a mood effectively.

STILLS PHOTOGRAPHY

The advent of the digital age has seen an explosion in the popularity of underwater photography, with more and more divers creating images of their dives for posterity. However, producing good results consistently is a craft, and this is what separates the professional from the amateur photographer.

CAMERA SYSTEMS

Underwater camera systems for stills photography can be divided into two categories. The first consists of truly amphibious cameras that do not require a housing to be used underwater, and can also operate on dry land. The pioneering Nikonos V camera is a famous example of the type. Amphibious cameras often have interchangeable lenses, and are available in both film and digital formats. These highly specialized cameras feature oversized controls for ease of operation underwater.

The second category, housed systems, are conventional cameras in waterproof cases. The cameras themselves can be divided into two subcategories: compact devices, and SLR (single lens reflex) models.

Compacts come in film and digital formats, and are light and convenient to carry and use. Most models offer fully automated focus and exposure control. You cannot change the lens on a compact, but some have variable focal length (zoom) lenses. After-market add-on lenses that attach to the outside of camera housings are also available.

ANIMAL MAGIC
To capture the vibrant colors of marine creatures and environments, you will need a purpose-designed strobe lighting system.

SLR cameras, which are available in both digital and film versions, allow the use of a range of lenses and offer complete creative control over key functions such as aperture, shutter speed, and focus. For this reason they are invariably the choice of professional underwater photographers, although SLR-compatible ports, housings, and lighting systems can be very expensive.

CAMERA HOUSINGS

Both compacts and SLRs require a housing to protect them from water damage. These are normally made of plastic, though aluminum housings are also available for SLRs. Aluminum housings are strong and durable but expensive and bulky, and hide the camera within, making it harder to operate. Plastic housings are cheaper and allow you to see the camera, but can be less durable. They are especially vulnerable to scratches and abrasion.

NIKONOS CAMERA
Amphibious cameras like the Nikonos V have an integral water-resistant housing.

DIGITAL COMPACT
Fully automated functions make digital compacts easy to use, but limit creative control.

SLR CAMERA
Available in film and digital formats, SLRs are bulky but offer superior image control.

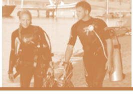

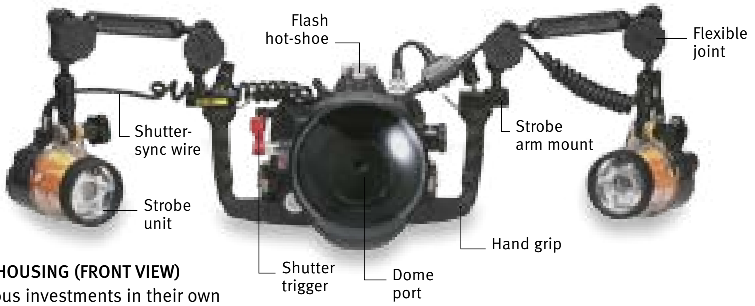
Flash hot-shoe
Flexible joint
Shutter-sync wire
Strobe arm mount
Strobe unit
Hand grip
Shutter trigger
Dome port

SLR HOUSING (FRONT VIEW)
Serious investments in their own right, large SLR housings offer excellent access to key controls.

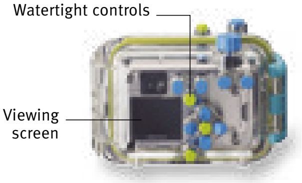
Watertight controls
Viewing screen

COMPACT IN PLEXIGLAS HOUSING
The housing features push-button mechanical controls for taking, reviewing, and even editing images underwater.

COMPACT WITH STROBE
A diffuser lens placed over the strobe softens the light it casts.

Bare strobe unit

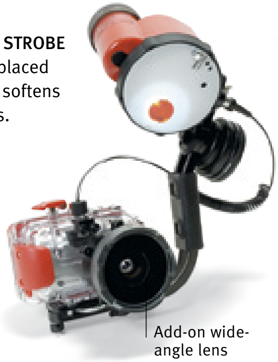
Add-on wide-angle lens

UNDERWATER LIGHTING SYSTEMS

Lighting is one area where compact digital cameras fall short. Their built-in flashes are too close to the lens to avoid backscatter (*see p.183*). You can distance the light source from the lens by using an auxiliary strobe light, bolted to the housing. SLR-based and amphibious systems often use strobes in pairs, mounted on flexible arms to allow the subject to be lit from different angles. Knowing how to position strobes to best advantage and manipulate power settings is essential, though many cameras will automatically trigger the strobe to fire for the correct duration for the exposure.

HOUSING MAINTENANCE

Maintaining a housing is relatively straightforward. O-ring integrity should always be checked and the ring lightly lubricated with silicone grease. Make a final visual check before diving. Housings should be rinsed in fresh water as soon as possible after a dive. Operate the mechanical controls during rinsing to prevent a build-up of salt crystals. If water penetrates the housing, it will ruin the camera, so a few moments taken to prepare and rinse gear can prevent expensive damage.

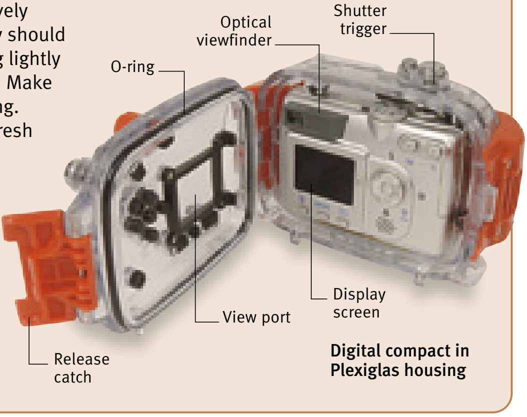
Optical viewfinder
Shutter trigger
O-ring
Display screen
View port
Release catch

Digital compact in Plexiglas housing

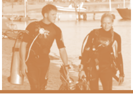

MAKING FILMS UNDERWATER

Like underwater stills photography, videography has been revolutionized by the advent of digital cameras, which are small and easy to use, compared to film or analog video. Basic point-and-shoot underwater videography is quite easy with consumer-grade digital cameras. To advance beyond the novice stage, however, you need special skills, more advanced equipment, and a flair for using video editing software.

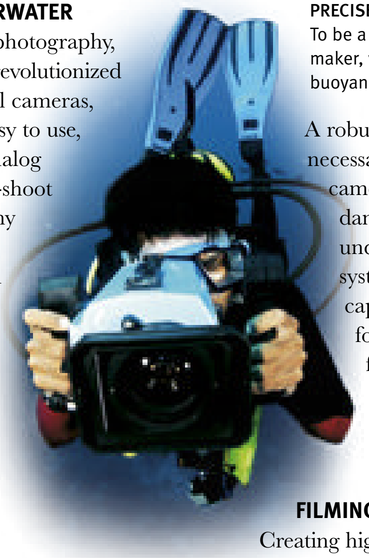

PRECISE POSITIONING
To be a good underwater filmmaker, you need to perfect your buoyancy control and finning.

A robust housing will be necessary to protect the camera from water damage, while a basic underwater lighting system is essential for capturing good color footage, since water filters out various colors from natural light as depth increases (*see p.183*).

CHOOSING A VIDEO SYSTEM

Choice of camera is the first decision for novice videographers. Though analog video cameras are still available, digital is now the dominant technology. Digital cameras are available in digital tape, flash media, HDD (hard disk drive), and DVD recording formats. Look for cameras with three-chip processors (called 3CCD cameras). These divide red, green, and blue color information into separate processing paths, offering richer colors than cheaper one-chip models. At the top end of the market are cameras with HD (high definition) imaging, which produce video of astonishing clarity, due to the high rate at which they scan visual information.

FILMING SKILLS

Creating high-quality underwater video requires you to do more than merely jump in with everything set to automatic, press record, and film throughout the dive. Many automatic systems perform poorly underwater. For example, autofocus becomes inaccurate in low-contrast environments, resulting in hazy images. Careful manual control of focus and manipulation of white balance (which sets the baseline for color values in different light conditions) are key skills.

Your diving skills also must be excellent; you need to be a steady platform for the camera—camera shake is a common cause of ruined footage. The camera and housing should also be neutrally buoyant (they should neither sink nor rise underwater) for ease of

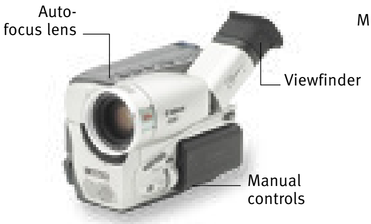

BASIC VIDEO CAMERA
Affordable cameras for home use usually feature automatic focus and exposure control, but may also have some manual options for skilled users.

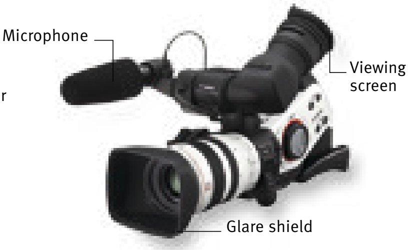

ADVANCED DIGITAL CAMERA
Top-end video cameras have advanced features like high definition, multi-chip processing, plus image stabilization systems for sharp images.

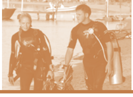

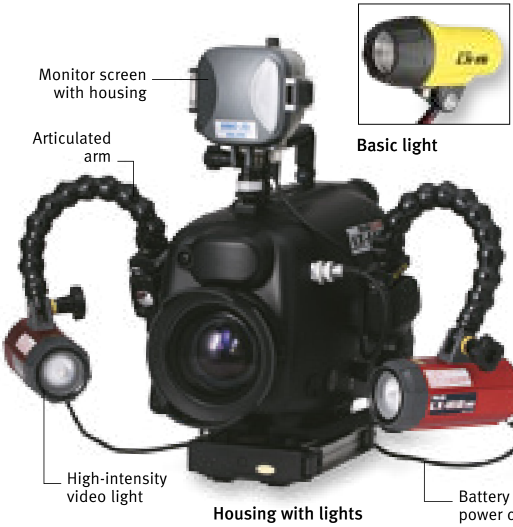

Housing with lights

Basic light

PRESSURE HOUSING AND LIGHTS
Housings fall into two categories—electronic or manually controlled. The latter use plungers, rods, pins, and cogs to allow the camera controls to be manipulated from outside the housing; they require more skill, but have the advantage of avoiding electronic systems that can go wrong or short out in the event that the housing is flooded.

Tough casing
Carrying handle

Bare housing

handling, and you should familiarize yourself completely with how your camera is operated, so its use becomes second nature underwater.

PRODUCTION AND EDITING

Thorough planning is the secret to good filmmaking, and video should be shot with an audience or editor in mind. Building a story around a given dive is not a complex process, but creates a target shoot list for the camera operator to work from, and gives the finished film an engaging narrative. You can place scenes in order and polish the results to a professional standard using video editing software that is commercially available for use on home computers.

KEEPING CLEAR
While filming, be careful to avoid causing physical damage to the organisms you have come to film. Avoid touching, or kneeling on, reefs.

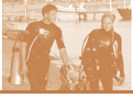

KEY PRINCIPLES

Modern cameras and housings allow even the novice to achieve consistently good photographs and video footage. One key development of the last ten years has been the ability to immediately view the results—even during the course of the dive—and then make adjustments accordingly before the next shot. But throughout the history of underwater photography and videography, certain principles have applied as much to hardened professionals as they have to happy snappers. Following these simple rules can immeasurably improve even your first forays into photography and filmmaking underwater.

LINING UP FOR A SHOT
Good positioning relative to your subject can help make subjects stand out from the background.

GETTING CLOSER

The first principle when shooting underwater is to get as close to the subject as is feasibly possible. Even the most crystal-clear water contains a mass of suspended particles, and even if your eyes cannot pick them out, you can be sure your lighting system or lens will. Getting close to the subject means greater clarity of the resulting image, and this is one of the reasons why "close-up wide-angle" is such a popular approach. This uses a wide-angle lens at close range, allowing the whole subject to be viewed with the lens only a short distance away. For the same reason, "macro" (close-up) photography is often a good place to start if you are a beginner, since clear, well-lit macro images are relatively easy to achieve.

For the novice photographer or videographer, another excellent principle is to remember to take shots with an upward orientation. There is nothing worse than viewing an endless series of dull blue or green shots where the subject is lost in the background of reef or seabed. Shooting upward isolates the

SUNLIGHT THROUGH WATER
The shifting light patterns encountered underwater can make for dramatic photography, but are challenging to expose correctly.

WAITING FOR A SHOT
Many fish are inquisitive, but easily alarmed; good buoyancy control will allow you to hold still until a subject presents itself.

subject against the background of the open water, and creates a more striking contrast. The composition of such shots can still include reef or coral, wall or wreck; however, the subject itself should be consistently easier to spot.

USING LIGHT SOURCES

Effectively lighting a subject is one area where most amateurs struggle. It is also one of the areas where modern compact digital cameras built-in flashes will let a picture down. If the light source is too close to the lens, particles in the water will reflect the flash directly back into the lens. This creates a phenomenon known as “backscatter,” a snowstorm effect of white dots on the photograph. One of the most basic prerequisites of taking good underwater images in anything other than perfect underwater conditions is an ability to position a strobe away from the lens. This allows the flash to come in at an obtuse angle, and helps to prevent backscatter from spoiling a shot.

These simple principles do not, unfortunately, guarantee good images every time, but they are useful for anyone starting out in photography or filming. Trial and error will help to refine your skills over time, and the best advice is to jump in, remember the basics, snap or film away, and enjoy!

DEPTH AND COLOR LOSS

Water absorbs different wavelengths of light as depth increases, causing images taken without artificial illumination to seem washed-out and dominated by blue tones. Reds fade from visibility at a depth of 10–15 ft (3–5 m), oranges at 35 ft (10 m), and yellows around 50 ft (15 m); by 80 ft (25 m) greens have almost disappeared, and by 100 ft (30 m) everything appears deep blue or black to the naked eye.

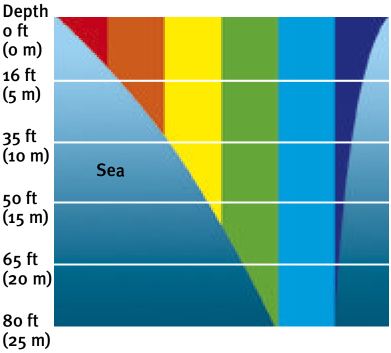

Night diving

To many, the idea of diving at night may seem a little crazy, but there is far more to it than just the thrill of exploring in the dark. Night diving reveals a whole host of underwater organisms that lie hidden during the day and only become active at dusk.

LIGHT AND COMMUNICATION

Night diving is obviously less simple than diving during the day, but when properly organized, it is relatively straightforward. A powerful flashlight will easily light your way and the creatures around you, revealing marine life in its true colors. However, if you cover your flashlight (never switch it off; the bulb may blow when you turn it back on), you will be surprised at how much light there is underwater. Many creatures use phosphorescence at night, and as you move through the water, you will cause plankton to release tiny pulses of light, leaving beautiful glowing wakes trailing behind you.

The main restrictions compared to daytime diving relate to communication, not only between divers, but also—and perhaps more importantly—between divers and boat cover. You will be able to monitor your buddy's location from the position of their flashlight beam. Signaling to your buddy can be done in the normal way, but while shining your light on to your hand as you make the sign. Attract your buddy's attention by waving your light, but take great care not to shine it at their face, since it can take as long as 15 minutes for their eyes to fully readjust to the near-darkness and restore night vision.

At the surface, avoid flashing your light unless you need to alert the surface cover in an emergency. If you are diving from the shore, leave a light source there—for example, a bright boat light, flashlight, or beacon—so you can easily find your way back to land after surfacing. Exit points should be marked with colored lights, to distinguish them from any other light sources.

You will find that most gauges and computers are easily readable by flashlight, and just shining the beam directly on to your gauge will make it luminous for a while.

NIGHT ON THE OCEAN
When diving on a clear, moonlit night such as this, you may be surprised at just how effectively the Moon is able to illuminate the underwater environment.

LIGHTING TIPS

It is vital that all divers carry a flashlight, and at least one other as a backup. Take spare batteries; if using rechargeables, make sure they are fully charged. Use hand-held lights, not head-mounted ones: one look at your buddy with a head-mounted light will ruin their night vision. A glow-stick attached to the back of your suit, while not essential, will increase your own visibility.

Hand-held flashlight

DIVING INTO DISCOVERY
A very different “night shift” of marine life can be seen during nocturnal dives, like this outlandish basket star.

TAKING PRECAUTIONS

Night-diving buddy pairs often link themselves together with a short line. While not essential, this is recommended for novice night divers and in poor visibility. If you are diving on a wreck to which a guideline and buoy are attached, secure a strobing marker light to the base of the line so it can be easily located in the dark. When choosing a dive site, opt for something simple, such as a shallow wreck or reef with plenty of distinctive features. An interesting daytime dive may be worth revisiting to investigate marine life that emerges at night. To minimize the risk of incidents, do not dive too deep, since any problems will be much harder to cope with in the gloom. If diving a more open site, use a surface marker buoy with marker light attached to identify it to any boat cover. Be careful not to lose contact with surface cover: getting lost at night could mean you are not recovered until morning.

FLASHLIGHT SIGNALS

During a dive, your torch beam will alert your buddy to your position. You can also make simple signals with your flashlight beam on the seabed to communicate with your buddy. At the surface, wave your torch vigorously from side to side if you require immediate assistance.

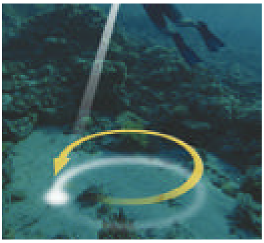

SIGNALING “OK”
Tracing a large circle on the seabed communicates to your buddy that you are OK.

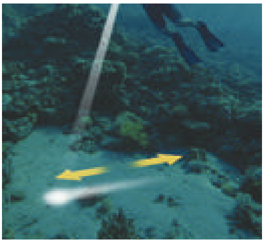

SIGNALING FOR ATTENTION
A rapid side-to-side movement means that you want to show or tell your buddy something.

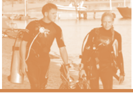

Cavern diving

Rich with opportunities for dramatic dives, caverns offer a bridge between ordinary open-water diving and the far more specialized activity of cave diving. Nonetheless, diving in caverns demands careful preparation and respect for potential risks.

WHAT IS CAVERN DIVING?

A cavern is a large, enclosed area of rock or coral that is of sufficient size to allow a diver to enter and move around. These fascinating environments can be extremely dramatic, are often eerily lit, and may play host to an array of animals not encountered on the open reef.

Unlike the more extreme sport of cave diving, in which divers penetrate beyond the reach of natural light, in cavern diving the exit is always clearly within view. However, cavern diving still carries inherent risks because the diver does not have direct access to the surface in the event of an emergency. If simple safety drills are not followed at all times, there is the potential for very serious problems to develop when diving even in simple caverns.

CAVERN DIVING EQUIPMENT

A flashlight is an essential item for cavern diving, but it is also sensible to bring along a backup light. This need not be a large, bulky model but must be able to provide enough light if your main flashlight malfunctions. Be sure to keep batteries freshly charged. An alternative air source, such as a pony bottle (*see p.68*), is also an excellent idea in caverns, since the diver cannot ascend directly to the surface if the primary air source suffers an unexpected failure.

A backup flashlight

SAFETY AND SKILLS

The first rule of cavern diving is to prevent your natural curiosity from turning a cavern dive into a cave dive! The exit should be clearly visible at all times—if not directly in your line of sight, then at least the natural light of the cavern entrance should remain visible. For many divers, the temptation to explore deep into cave systems is strong; however, the moment the exit is lost from view and natural light fades, the dive becomes a cave exploration, which should not be attempted without appropriate equipment and special safety and skills training.

Certain skills are essential when diving larger caverns. Buoyancy control is important, particularly when there is a silty substrate on the cavern floor.

ON THE THRESHOLD
A diver hovers at the entrance of a freshwater cavern in Australia. Following basic safety rules is essential at such sites.

INNER SPACE
A diver shines light on the interior of a sea cave. Caverns have a mysterious ambience quite different from that of open diving environments.

It is also advisable to use a finning technique known as the frog kick (*see p.123*). This is a gentle, circular fin-stroke that helps you avoid stirring up sediment. One misplaced kick from a fin can turn a cavern into an impenetrable soup of silt and lead to disorientation and panic. For this reason, and as a matter of plain common sense, it may also be appropriate to tie off a line at the entrance of larger caverns (*see p.127*) to provide a direct route to the exit in case of reduced visibility. It is also strongly advisable not to enter a cavern without local knowledge or a guide; there may be specific issues about a cavern system and the hazards it contains—such as seasonal changes in visibility—that do not appear in any standard dive guidebooks.

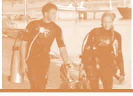

Ice diving

Diving under ice might not sound like an attractive proposition, but it is a uniquely exciting experience. Visibility is typically excellent beneath ice, with soft light filtering through from the world above. Ice diving can also take you to breathtakingly beautiful parts of the world.

UNIQUE ENVIRONMENT

Ice diving can be done in the sea or in frozen lakes. The aquatic life you encounter will depend on where you are in the world and whether you are in a marine or freshwater environment. Visibility is often stunning; the ice provides a protective layer over the water, so it is not churned up by the wind, and low light levels under the ice also help to inhibit algal growth.

COLD-WATER GEAR
A diver waits to enter the water wearing special equipment designed to cope with conditions of extreme cold.

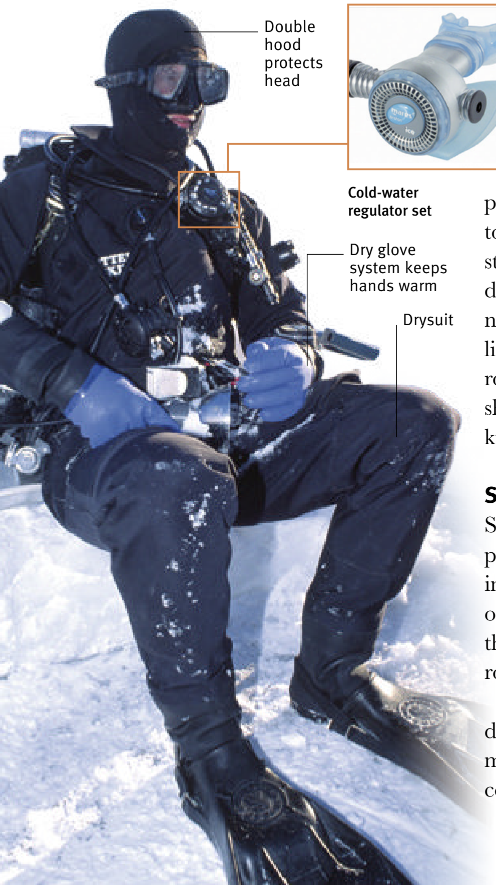

KNOWING THE ROPES

If you are planning to go ice diving, you will need to undergo special training. This will teach you not only the safety aspects of diving while tethered on a rope under ice, but also how to cut the ice hole through which you enter the water, and how to tend safety ropes for other divers while you remain on the surface in a support role.

Ice diving requires a minimum of four to six people per dive. There will normally be two divers operating as a buddy pair (*see pp.110–11*), and two people to tend their lines. It is advisable to have at least one extra person on standby. It is standard practice for each diver to be tethered separately, and it is not advisable to tether two divers on one line, because if anything happens to this rope, then both divers can be lost. Lines should always be tied using a bowline knot, which is very secure.

SIGNALS AND SAFETY

Signaling to your tender is an important part of ice diving. Since you are diving in extreme conditions, it is vital that he or she knows you are OK throughout the dive. You will learn about giving rope signals during training.

Apart from the obvious dangers of diving in an enclosed environment, you must also be careful to monitor your core body temperature. Repeat dives

SPECIAL PROCEDURES
Ice divers rely on a surface support team (*above*) for safe diving under the ice.

should only be undertaken when you are fully warmed up. You should also ensure that you eat high-energy, hot food to replace the calories you will burn during the dive and in keeping warm afterward.

FROZEN DEPTHS
A diver films the fascinating world beneath the surface. Many unique animals live in cold seas.

SPECIAL GEAR

Ice diving requires special equipment. You will need an environmentally sealed regulator set (*see p.63*), which is designed to be less susceptible to freezing. You will be taught to submerge this in the water before diving, to acclimatize it to the temperature. You must only breathe from it when submerged, to reduce the risk of free-flow (an uncontrolled release of air due to a frozen mechanism). It is also recommended that your primary air source and octopus are mounted on separate first stages (*see p.61*) to ensure that one will work if the other freezes. Your cylinder will need a special pillar valve to accommodate this.

You will need to wear warm undergarments beneath your drysuit, thick neoprene gloves or a pair of waterproof dry gloves (*see p.54*), and two thick hoods for extra warmth.

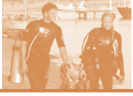

Freediving

Once simply a fishing technique, freediving—diving using only the air carried in the lungs—has become increasingly popular as a sport. Extreme athletes now constantly push the perceived limits of human performance, diving to remarkable depths on a single breath of air.

PURE SPORT

Freediving is not exclusively the preserve of the extreme athlete: anyone who has snorkeled and duck-dived to explore the seabed could describe themselves as a freediver, making this the most accessible of all diving disciplines. Many would argue that it is also the purest form of diving, requiring only the minimum amount of equipment and relying entirely on discipline, courage, and physical conditioning.

The roots of freediving stretch back to when the human race took its first tentative steps into the sea, diving into the water to harvest food and explore the seabed. Modern freediving, by contrast, is a disciplined sport, with record attempts highly regulated by the two governing bodies—A.I.D.A. (Association Internationale pour le Développement de l'Apnée) and F.R.E.E. (Freediving Regulations and Education Entity). Divers compete in a number of categories, two of which usually take place in a swimming pool: static breath-hold—a simple endurance test with a timed immersion in a shallow pool—and dynamic breath-hold, a test of the distance divers can swim using only a single breath. In open water, competitors are judged on the depths they can reach and return from safely.

EQUIPMENT

Freediving does not require much equipment, but specialized items, such as streamlined suits and fins of variable flexibility, are used to maximize the speed of movement through the water. Masks with low internal volume are also useful, as they allow divers to conserve air that would otherwise be used to exhale into a standard mask to counteract mask squeeze (*see p.115*).

Specialist mask

SUUNTO
CSM
sports net global
GUL
BAUER KOMPRESSOREN

STATIC BREATH-HOLD COMPETITION
Divers compete for the length of time they can hold their breath. World-class competitors can achieve times of over seven minutes.

Freediving naturally carries inherent risks—the most advanced and extreme disciplines demand that the body's supply of oxygen is completely used on a dive, and losing consciousness is always a danger. Exploration of freediving limits should never be conducted alone, and only after attending one of the many introductory courses available to those interested in taking up the sport.

TOUCHING THE DEEP
A freediving competitor reaches for the tag that proves that the designated target depth has been reached—this must be presented at the surface for inspection by judges.

DIVING DEPTHS

Competitive depth records are judged in strict categories, defined by precise parameters such as the use of weights or fins. In “No Limits”—the category in which the highest-profile records are set—the diver is dragged down by a weighted sled, and rockets to the surface using an airbag.

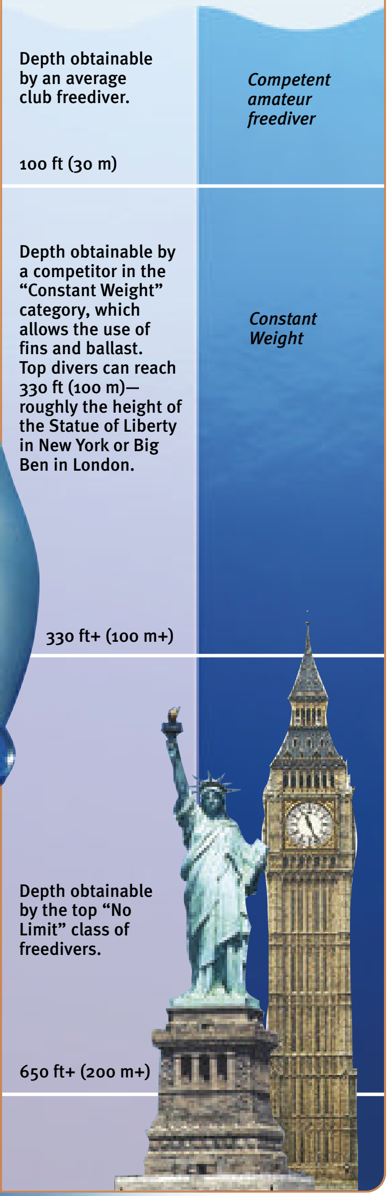

Depth obtainable by an average club freediver.

100 ft (30 m)

Depth obtainable by a competitor in the “Constant Weight” category, which allows the use of fins and ballast. Top divers can reach 330 ft (100 m)—roughly the height of the Statue of Liberty in New York or Big Ben in London.

330 ft+ (100 m+)

Depth obtainable by the top “No Limit” class of freedivers.

650 ft+ (200 m+)

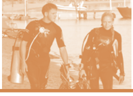

Underwater archaeology

While marine archaeology is mainly confined to the academic and scientific world, it could be said that every diver who explores a wreck has an archaeological interest and responsibility. Both the seabed and many freshwater sites hold compelling historical secrets.

RULES FOR THE AMATEUR

Although every dive has the potential to reveal historic artifacts, targeted methodical surveying is an altogether more serious business, as is the correct means of excavating and recording any finds. Archaeological divers can therefore be placed in one of two distinct categories—the serious professional diving archaeologist, and the enthusiastic amateur.

Although this is a field in which amateurs can make a very valuable contribution, it is essential that they do so only through involvement in an organized project, conducted by a recognized and authorized group. Far too many sites have been destroyed, and artifacts removed, by what amounts to vandalism and theft—the equivalent on land might be digging a historic site or burial ground without permission, which is unthinkable, of course.

There are many organizations that welcome help from amateur volunteers, and will supervise and guide you. The thrill of discovery really does reach a new level when it is shared with individuals who understand the full significance of a find. You may find that more scope for joining an organized project exists if you first undertake some training, perhaps by volunteering on digs on land.

Learning professional marine excavation and recovery techniques can be fascinating for the amateur. Archaeologists go to great lengths to avoid damaging a surveyed location, and even when they locate an item of interest, try to avoid intrusive collecting techniques or digging.

MORE THAN YOU BARGAINED FOR

By taking an artifact from a dive site, you destroy its integrity and hinder efforts to learn about past peoples. And that may not be the end of the matter. In many parts of the world, sites of marine archaeological interest are strictly regulated. The recovery of artifacts is—quite rightly—taken very seriously, and penalties for illegal activity and for removing items can be severe.

Amphorae, used to store oil and wine, recovered from a Turkish site

RELIC OF CONFLICT
Divers inspect the remains of a Japanese fighter aircraft shot down in the Pacific during World War II.

WHAT TO EXPECT

Marine archaeological expeditions are usually initiated because research or local knowledge has pointed to a region that has archaeological potential. Modern GPS (Global Positioning System) equipment has made pinpointing such sites much easier. On reaching the site, a detailed survey is generally first carried out. This may involve using a datum line (a fixed reference line on the seabed from which measurements are taken), or pegging out a grid. Precise recording of the location and state of the site before digging is vital, and is usually conducted through photography or extensive sketching. Small test digs may take place before a full-scale excavation occurs, with the exact site of each find carefully logged, and the finds themselves sketched or photographed. Preservation and restoration techniques are extremely specialized for such finds. Marine archaeology is a long way from the derring-do of Indiana Jones, requiring organization, attention to detail, fine diving skills, and persistence.

CAREFUL SURVEY
Meticulous recording of survey data is one way in which amateurs can help on a professional dig.

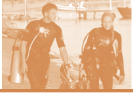

Technical diving

The desire for ever-deeper dives and longer periods underwater has led to the development of special gas mixtures. The use of these in technical diving offers an exciting range of new skills and marine environments for qualified technical divers to explore.

PUSHING BACK BOUNDARIES

Scuba divers are limited, by simple physical laws, to a certain amount of time under water (dependent on the depth of the dive) and an overall maximum depth. Training agencies recommend that recreational divers using compressed air go no deeper than 130–165 ft (40–50 m). Beyond this, nitrogen can build up dangerously quickly in the body, greatly increasing the risk of decompression sickness (DCS; *see pp.142–43*). In addition, the increased levels of nitrogen in the bloodstream below 165 ft (50 m) can lead to severe nitrogen narcosis (*see pp.100–01*). The oxygen content of air, too, becomes increasingly toxic to the nervous system the deeper you dive (*see pp.96–97*).

To extend dive times and allow exploration to greater depths, special mixtures of gas have been developed in which the balance of oxygen and nitrogen has been altered, or that contain new gases to dilute the impact of either oxygen or nitrogen. The use of these mixes is known as technical diving, and involves three types of gas mixes: nitrox, which allows longer dive times, and trimix and heliox, to reach greater depths.

EXTENDING DIVE TIMES

Normal air is composed of around 21 percent oxygen and 79 percent nitrogen. Nitrox is a gas mixture that contains a higher percentage of oxygen and a lower nitrogen content than normal, allowing longer dive times before the diver's body stores up too much nitrogen, and also reducing diver fatigue after the dive. Nitrox can be mixed in various proportions, from

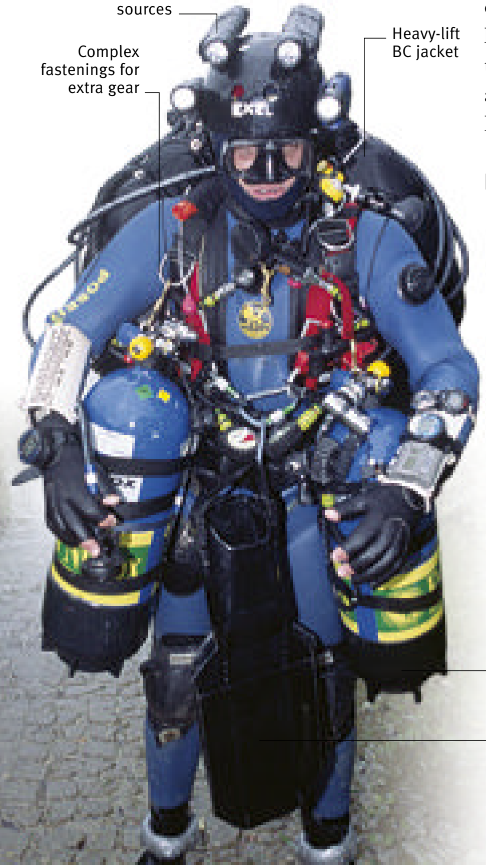
Multiple light sources
Complex fastenings for extra gear
Heavy-lift BC jacket
Tanks contain gas mixtures for specific depth ranges
Heavy-duty drysuit worn over a thermal undersuit insulates the diver from the low temperatures of extreme depths

UNDERGROUND ODYSSEY
A diver emerges from the water wearing the complex equipment required for cave diving, including tanks filled with special gas mixes, and sophisticated dive computers.

BEYOND NORMAL LIMITS
A diver explores deep within a flooded cave system. Technical diving procedures enable divers to go farther than ever before.

22 to 50 percent oxygen. However, beyond a certain depth the increased oxygen in the nitrox mix becomes toxic and can cause seizures. The more oxygen in the blend, the shallower this will occur. For this reason, diving with nitrox requires special training, although courses are readily available to recreational divers.

GOING DEEPER

Divers wishing to travel deeper than 165 ft (50 m) use either heliox or trimix, both of which introduce helium into the breathing mixture. Heliox contains only helium and oxygen, whereas trimix is a three-gas combination of helium, oxygen, and (usually) nitrogen. Heliox is used mainly by military and commercial divers, while trimix is more often used by leisure divers. Both succeed in reducing the likelihood of nitrogen narcosis (*see p.101*), oxygen toxicity, and DCS. However, since helium conducts heat six times faster than air, these mixes can cause the diver to feel very cold. Helium is also absorbed rapidly into the body, so lengthy decompression is necessary. Divers wishing to use these gas mixes must undergo appropriate training.

TECHNICAL KIT

Diving at extreme depths or in extreme environments, like deep caves, requires special gear. Technical divers meticulously analyze their equipment, and try to eliminate weak points—there is little margin of error during a deep technical dive. To support the extra gas tanks used during technical dives, most divers opt for high-lift “wing”-style buoyancy compensators (*see p.59*). They also choose high-performance regulators that can handle the task of supplying breathing gas at extreme depths. Sophisticated dive computers are used to monitor how technical gas mixes are metabolized during a dive.

A multi-gas computer

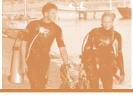

Careers in diving

Lots of people grow to love recreational diving and consider turning their passion into a profession. Work in diving is demanding, and earnings are often modest, but the personal satisfaction to be gained from a career in diving may make the hard work worthwhile.

GOING PROFESSIONAL

There are two common factors in the life of any diving professional: the work is invariably challenging, and there are always many more candidates for jobs than there are vacancies.

As a growth sector, the recreational diving industry offers the most career opportunities. Qualifying as a divemaster or instructor with a recognized training agency means that you will be able to work at dive centers in every corner of the globe. However, qualified instructors are responsible for the safety of their students, and the selection process is necessarily a rigorous one. Being a highly skilled diver is just one of the qualities an instructor must possess; strong communication skills, maturity, and in-depth technical knowledge are also required. Becoming an instructor with a commercial agency can also be an expensive process, but it is a crucial first step towards gaining work in diving, and may open the door to a lifelong career.

COMMERCIAL DIVING

The world of commercial diving covers all activities outside of recreational instruction, and includes scientific, media, and industrial diving. The last of these represents the best-known face of the commercial diver—the hard-hatted aquanaut working on the rigs and pipelines of the oil industry. The demands of the job have changed from

DEEP-SEA EXPLORER
A diver in a mechanized pressure suit works on the deep ocean bed. Such jobs are very demanding, and competition for new positions is high.

EXPERT GUIDANCE
An instructor shepherds a new student through his first dive. Resort guides are sought after in many areas.

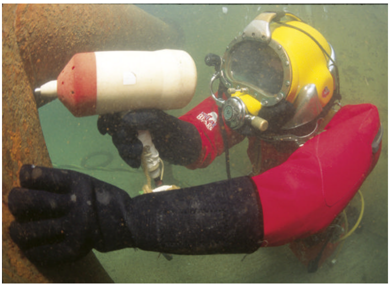

UNDERWATER WORKER
A commercial diver repairs a steel frame on the sea floor. Such jobs require advanced knowledge of marine engineering techniques.

the pioneering days of oil exploration, and today's commercial divers must be able to master advanced surveying and engineering skills while working in the most difficult of environments.

SAFEGUARDING THE SEAS
A navy diver prepares to submerge for mine-clearance duties. Military divers face dangerous tasks and are trained to exacting standards.

JOINING THE SERVICES

Military and police diving present a different set of challenges for the aspiring professional, often requiring diving in appalling conditions, such as in confined spaces or very low visibility. Demand for diving jobs in the military or police is invariably high, and only the best candidates are accepted after a careful and strict selection process.

Underwater life

Introduction

Oceans cover 70 percent of the world, yet we remain ignorant of much of what lives beyond the relatively shallow regions explored by humans. The sea is a vast depository of life holding a bewildering array of animals and plants, from the massive to the microscopic.

PLUMBING THE DEPTHS

Although the majority of our planet is covered in water, only about 16 percent of all known species live in the sea. This is more a reflection of our own inadequate knowledge than of a lack of diversity in the oceans: there are undoubtedly many thousands of unknown species lurking in the vastness of open sea that we have yet to encounter. Much of our exploration of the sea has, by necessity, taken place only in the shallow margins around land. The limits of our own physiology and technology make sustained exploration below 330 ft (100 m) both demanding and expensive, but the average depth of the ocean is in the region of 2.3 miles (3.8 km), leaving vast tracts of our own planet totally unexplored.

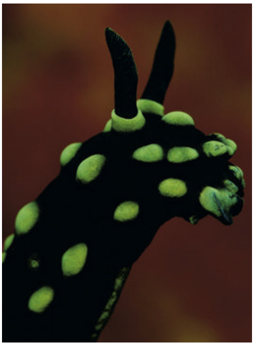

MARINE MELTING POT
The ocean contains a diverse range of life-forms, such as this bizarrely patterned nudibranch.

The coastal fringes that are the realm of the recreational diver reveal a wealth of life. Only two percent of known marine species live in the open ocean, with the remaining 98 percent burrowing, swimming, scuttling, and crawling over reef, rock, and muddy sea bed. Although recreational divers are confined to depths of around only 130 ft (40 m), the first 650 ft (200 m), known by scientists as the neritic province, hosts the greatest diversity of life in the sea.

ZONES OF LIFE

Four temperature bands define the basic zones of marine life on the planet. At the northern and southern extremities of the Earth are the polar seas, capped with ice. These graduate into temperate seas, which mostly range from 39 to 68°F (10–20°C). Subtropical seas form a loosely-defined buffer zone around the tropical seas of the equator, which are the world's warmest, rarely dropping below 68°F (20°C).

THERMAL RANGE
Sea surface temperatures vary from freezing point at the poles to more than 30°C (86°F) in the tropics.

KEY

- 🟦 Polar
- 🟦 Temperate
- 🟨 Subtropical
- 🟧 Tropical

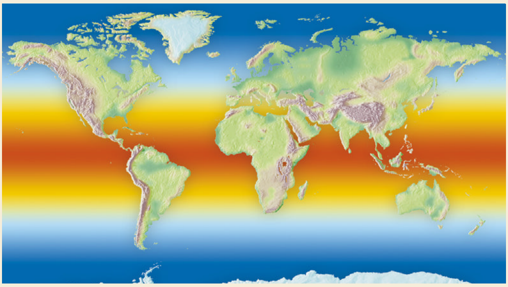

ENDLESS VARIETY

Water is the perfect medium to support life, at once a means of transportation, communication, support, and reproduction. The lives of countless numbers of people worldwide are dependent on the world's oceans and the life they support, which ranges from microscopic plankton to the immense bulk of the largest animal alive today—the blue whale.

Types of marine environments and the species that inhabit them vary with location around the world. The open ocean is roamed by giants, with large migratory species, filter feeders, and opportunistic predators. The latter, bold and inquisitive, provide many of the "big-animal encounters" that divers find so rewarding. Closer to shore, the warm, clear waters around the tropics are home to coral reefs, the most spectacular of all marine systems. More temperate coastal regions have richly populated marine "forests" of kelp—seaweeds that grow to immense lengths to reach the sunlit upper reaches of the water.

ENVIRONMENTAL CONCERNS
The sea supports countless livelihoods, and it is perhaps no surprise that overexploitation occurs. Here, an official apprehends shark-finners.

Colder waters are the province of migratory marine species, traveling the seas in search of new feeding opportunities as the seasons change.

And although we tend to think of migrations taking place over great distances laterally, there is also a great vertical movement in the oceans. As the day and night change, a range of animals migrate through the water column, creating the largest daily movement of animals on Earth.

FLIGHT PATH
With a massive wingspan of up to 10 ft (3 m), the eagle ray is a magnificent marine species.

Tropical seas

If ever there was an image that sums up the magic of diving, it is that of a coral reef. Although coral reefs are the most famous feature of tropical seas, these warm clear waters have much more to offer, such as mangrove swamps, sea-grass beds, and vast tracts of open ocean.

WARM BELT

Tropical waters occupy the region within the "20°C isotherm"—the irregular band of water north and south of the equator that seldom drops below 68°F (20°C). Although undeniably appealing to the diver, they actually present a less attractive environment for marine life. The warm equatorial sun heats the upper layers of the water column, creating a marked temperature difference between the surface water and deeper colder water. Such temperature stratification prevents mixing of the water column, so nutrients are not passed up from deeper water. Marine animals and plants have therefore had to adapt and evolve numerous strategies to cope with the lack of nutrients in the water around them. Perhaps the most remarkable of these is the coral reef. Primary production—the formation of organic compounds from inorganic material—is up to 100 times greater in coral reefs (*see p.204*) than in open tropical waters, and although they cover only 0.2 percent of the ocean environment, they are home to 25–30 percent of all fish species.

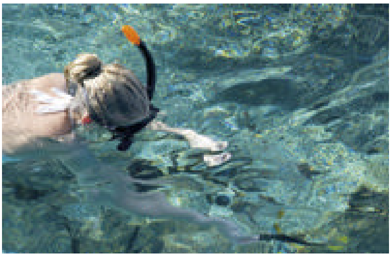

PERFECT SNORKELING CONDITIONS
By happy chance for divers, tropical coral reefs require clear warm water with little seasonal variation to grow at an optimum rate.

LIFE ON THE REEF

The diversity of tropical regions is based to a large extent on coral reefs, although these vary in their diversity depending on location. The reefs with the most vibrant marine life are found in the Indo-Pacific, with up to 700 coral species playing host to 2,000 fish species. In contrast, Atlantic coral reefs have only 70 coral species and about 600 fish species.

Raccoon butterflyfish

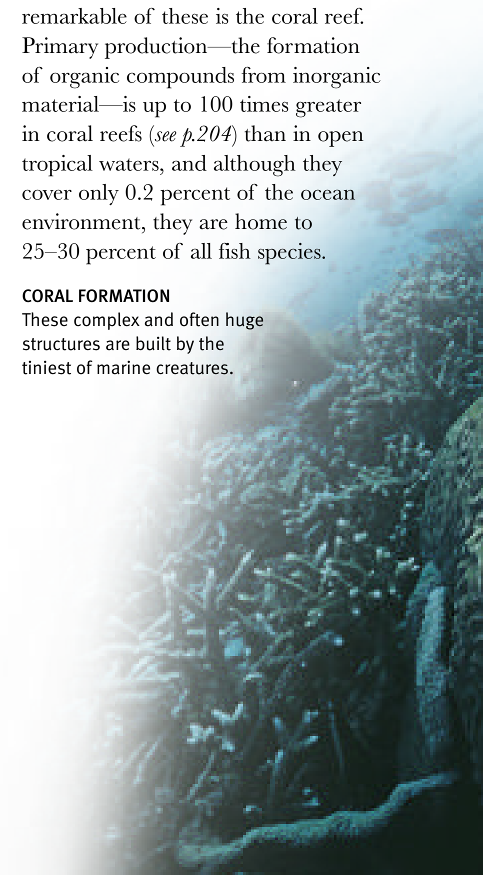

CORAL FORMATION
These complex and often huge structures are built by the tiniest of marine creatures.

The reefs of the Indo-Pacific are the richest marine environments on Earth. Their beautiful structure and bright colors, combined with the splendor and variety of animals that inhabit them, make reefs irresistible to divers, and it is no coincidence that many of the world's best dive sites are found on coral reefs.

PARADISE ISLAND
Tropical beaches draw ever-increasing numbers of tourists, creating intense development pressures which, if uncontrolled, pose a significant threat to tropical seas.

COASTAL NURSERIES

Tropical waters are also home to mangrove swamps and sea grass beds, both arguably as important as coral reefs in the overall health of tropical seas. There are 40 species of mangroves—tropical trees and shrubs that grow in shallow and intertidal coastal waters—and they form flooded forests that act as nurseries for various reef and open-water fish species. The 50 species of sea grass form "meadows" in shallow waters that are feeding grounds and nurseries for many fish species. The eradication of sea grass beds and mangrove swamps around the world is a real concern, with undeniable impacts on coastal ecology as animal populations are denied crucial areas for the growth and development of their young.

REEFS UNDER THREAT

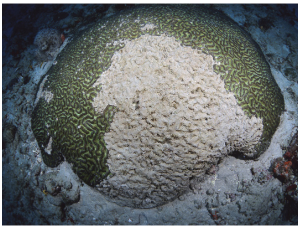

BLEACHED CORAL
Coral bleaching is one of the more vivid consequences of global warming and increased pollution.

Coral reefs worldwide are under intense pressure. The continued development of coastal regions has caused silty water to run into the seas, smothering these delicate systems, and, coupled with the damage caused by destructive fishing methods, it is thought that up to 90 percent of reefs have been impacted by humankind. There is also evidence that rising water temperatures are causing a phenomenon known as bleaching, which is fatal to reefs. This occurs when the coral polyps eject the minute algae that sustain them as a response to stress.

Diving a reef can be the highlight of a diver's life, but we have a very real responsibility when exploring reefs not to harm or disrupt them in any way. A considerate approach is vital if we are to preserve these wonderfully vibrant ecosystems.

THE CORAL CYCLE

Reefs are formed because of the extraordinary symbiotic relationship between tiny animals—polyps—and minute algae. Polyps are colony-forming animals that construct coral reefs. The polyp has a hard exoskeleton made of calcium carbonate, and this forms the physical structure of reef systems. Minute single-celled algae called zooxanthellae live within the polyp, protected by its hard skeleton, and nourished by its waste products. In turn, the zooxanthellae provide the polyp with nutrients and oxygen, which they produce, like plants, through photosynthesis. This process requires light, which is why tropical reefs are found only in waters less than 330 ft (100 m) deep—the so-called "photic zone", penetrated by sunlight. This relationship is vital: warm, clear water tends to be low in oxygen and nutrients, but the partnership of algae and polyp creates a highly efficient nutrient recycling system, and, in turn, sustains a rich array of marine life.

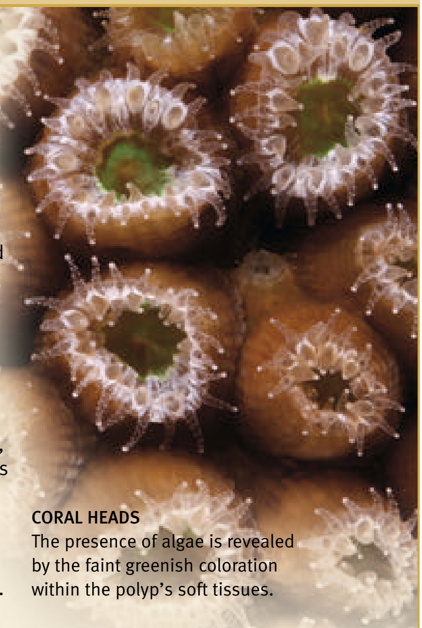

CORAL HEADS
The presence of algae is revealed by the faint greenish coloration within the polyp's soft tissues.

REEF DISTRIBUTION
Over 90 percent of coral reefs are clustered in the Indo-Pacific region. Reefs do not form on coasts that experience cold or strong currents.

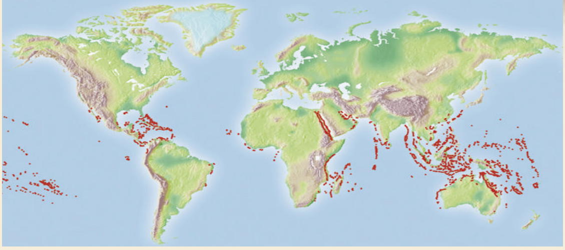

KEY
● Reef

REEFS AND RESPONSIBILITY Low-impact “hands-off” diving techniques are essential to avoid damaging coral reefs.

Subtropical seas

Although there is no strict definition, subtropical seas can be described as waters too consistently warm to support temperate marine life, but too cold for coral reefs to flourish. Also known as warm temperate waters, subtropical seas are found in all of the world's oceans except the Arctic.

BETWEEN TWO WORLDS

Subtropical waters are sandwiched between tropical ecosystems (generally identified by the presence of coral reefs and mangroves), and the temperate zone (identified by cold, rich waters supporting complex systems such as kelp forests). In subtropical marine systems, the water temperature is too high to contain large amounts of oxygen and nutrients (*see pp.208–09*), and yet too low to promote the recycling properties of coral reefs (*see pp.204–05*).

The most popular subtropical areas for diving are the Mediterranean and parts of the Gulf of California (also known as the Sea of Cortez). Other key sites include South Africa's Aliwal Shoal (*p.259*), Fernando de Noronha (*see p.318*) off the Brazilian coast, and the islands of Cocos, Malpelo (*see p.319*), and Roca Partida (*see p.320*) in the eastern Pacific.

RED GROUPER
These large predatory fish are typical of Mediterranean seas, but overfishing has depleted their numbers.

In terms of their marine life, there is a stark contrast between the Gulf of California and the Mediterranean. The former is famed throughout the diving world for its big-animal encounters, while the latter is a somewhat barren system. The cause of this difference is their topography. The Mediterranean is in effect an isolated bowl of deep water,

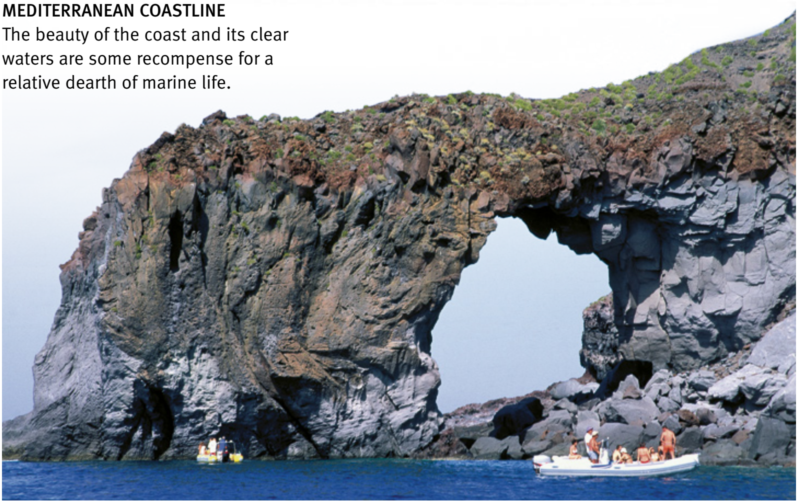

MEDITERRANEAN COASTLINE
The beauty of the coast and its clear waters are some recompense for a relative dearth of marine life.

whereas the Gulf of California is less enclosed and relatively shallow, with water and animals able to move freely to and from the Pacific.

Plankton, which are at the base of the food chain, thrive in waters in which mixing of the water layers brings nutrient-rich deep water to the surface (*see pp.214–15*). Most of the cold, nutrient-rich water from the Atlantic that enters the Mediterranean does so through the narrow Straits of Gibraltar, creating a plume of colder water that extends to the Sicilian Channel. The best dive sites in the Mediterranean lie along this plume. Much of the rest of this sea experiences very little mixing of the water column, and while the result is very clear water, there is little to support life.

The Gulf of California is open at the southern end, with the Baja peninsula to the west and the Mexican mainland over 125 miles (200 km) to the east. Much of it is less than 985 ft (300 m) deep, and these shallow waters are regularly mixed by water entering from the Pacific to the south and are rich in marine life. This is one of the finest locations to encounter large animals such as humpback whales and manta rays.

OCTOPUS ENCOUNTER
Various species of this cephalopod can be encountered in subtropical waters worldwide.

Temperate seas

The richest of all marine ecosystems, temperate waters are the stretches of ocean between the poles and the tropics. These waters are hugely changeable, offering a seasonal range of conditions at any one dive site, often accompanied by a variety of visiting migratory species.

THE TEMPERATE ZONE

It is difficult to define the temperate zone precisely, because major ocean currents can artificially extend temperate waters into tropical regions, or push temperate species into near-polar regions. The best definition is perhaps found in the temperature range of the water: true temperate waters rarely sink below about 39°F (10°C) or rise above 68°F (20°C).

The positioning of the temperate zone means that its waters see a greater change of conditions from season to season than tropical and subtropical seas. In particular, the amount of light hitting the surface of temperate waters varies considerably between winter and summer, and this leads to a seasonal variation in productivity—the sheer amount of life that an ecosystem can support. At the base of the food chain in the marine ecosystem are tiny organisms called plankton, which are dependent on light and nutrients for growth and reproduction. In the winter, when even water near the surface receives little light, conditions are poor for plankton growth, and so productivity is low. However, during the stormy winter months there is considerable mixing of cold, nutrient-rich deep water with the

FILTER-FEEDER
The basking shark moves through the water with its huge mouth open, filtering plankton from the water.

surface layer (*see pp.214–15*), and so in spring, when the amount of light hitting the surface of the water vastly increases, conditions are perfect for the growth of plankton. Huge spring blooms develop, attracting plankton-feeding species and their predators, and the waters become rich in marine life. Productivity drops again in summer, because, even though light conditions are perfect for plankton growth, there is little mixing between the surface layer and deeper water, and so nutrients become scarce.

ROUGH SEAS
Temperate seas can be rougher and colder than tropical and subtropical waters, but those who brave them are rewarded by the wealth of marine life on display.

TEMPERATE DIVING

Diving in temperate waters is not the crystal-clear, warm-water experience beloved of the coral reef diver, and requires more equipment (*see p.45*) than diving in tropical regions. However, some of the localized ecosystems within temperate seas, such as the majestic kelp forests (*see pp.210–11*) found in certain coastal regions, are extremely dramatic. When dived at optimum times in the season, temperate waters are rich in life. The seasonal blooms of plankton draw a range of migratory species, and also attract impressive predatory species, such as shark and tuna.

PLANKTON BLOOMS

Plankton are free-floating animals (zooplankton) and plants (phytoplankton) at the base of the marine food chain. When environmental conditions are optimal for their reproduction, they can multiply very rapidly, creating huge “blooms” of many millions of individual plankton. Where massive blooms of phytoplankton develop, the sea can literally turn green.

Cyclops zooplankton

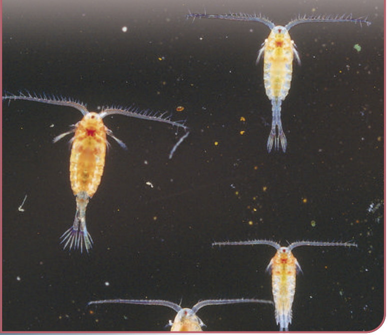

UNDERWATER FORESTS

The forests of giant kelp that grow in temperate seas are considered by some to be the most impressive marine environment of all. Although kelp—a brown alga—is extremely abundant in temperate waters throughout the world, true kelp forests are specific to the areas with marine conditions that support the growth of two kelp genera: *Macrocystis* and *Nereocystis*. These require reasonably clear water, a degree of water movement, a suitable substrate making up the sea bed, and a maximum temperature: most kelp forests will not flourish in water that exceeds 68°F (20°C) on a regular basis. Some species of kelp grow at a phenomenal rate, and to a truly colossal size. *Macrocystis pyrifera*, or giant kelp, can reach in excess of 165 ft (50 m), and can grow up to 25 in (60 cm) in a single day, making it the fastest-growing alga on Earth.

The best areas for the growth of kelp forests are off the west coast of the US and Canada; in the Cape region of South Africa; in Southern Australia; along almost the entire length of the west coast of South America; around much of New Zealand; and in southern Japan. For the diver, the attraction of these biological wonders is not simply the forests themselves, but also their numerous residents. The analogy with a terrestrial forest is a good one: activity takes place from the seabed up to the canopy, from specialist mollusks that bore into the kelp itself, through to pelagic animals such as rays that soar through the upper levels. As well as permanent residents, a kelp forest provides a nursery for many open-ocean species and a hunting ground for impressive predators. To dive a kelp forest on a day with good visibility as the sun dapples through the canopy overhead is an unmatchable experience.

SEA OTTER
Vital residents of kelp forests, sea otters help control populations of sea urchins, which if left to multiply would decimate the kelp.

SUBMERGED CANOPY
Invariably dramatic and teeming with aquatic life, kelp forests are the temperate water equivalent to the bustling reefs of tropical seas.

KELP BEDS
Huge numbers of marine creatures, large and small, make their home among kelp.

Polar waters

Although both the Arctic and the Antarctic conjure up visions of frozen desolation, below the ice and water surface lie regions of staggering abundance. Possibly the most challenging of all marine environments, the polar regions offer many surprises for the intrepid diver.

FROZEN ENVIRONMENT

Polar waters are found in the regions around the Arctic and Antarctic, though the precise lines of demarcation are blurred. Where polar waters start and temperate waters end varies between the poles. For the Arctic it is the 50°F (10°C) above-water July temperature boundary, while for the Antarctic it is seen as the line of the Antarctic Convergence, a zone where cold Antarctic waters sink below warmer northern currents.

Both the Antarctic and Arctic have quite consistent water temperatures, because the low angle of the sun and the reflective qualities of the ice minimize the impact of solar radiation. The temperature of the water hovers around an extremely chilly 32°F (0°C), only just above the freezing point for seawater, which is 29°F (–1.8°C). However, although they may seem similar, the Antarctic and Arctic are actually quite different in terms of their physical geography and marine ecosystems.

BELUGA WHALE
Arctic diving “safaris” provide opportunities to dive with animals such as beluga whales and a number of seal species.

POLES APART

The plants and animals of the Arctic and Antarctic differ considerably, situated as they are on opposite sides of the world. Antarctica has its own species of krill (some seven of them)—the small shrimplike creatures that are the bedrock of the food chain here. Swarming in huge numbers on the surface, they draw in a host of feeders—including the mighty baleen whales—during seasonal migrations.

Krill

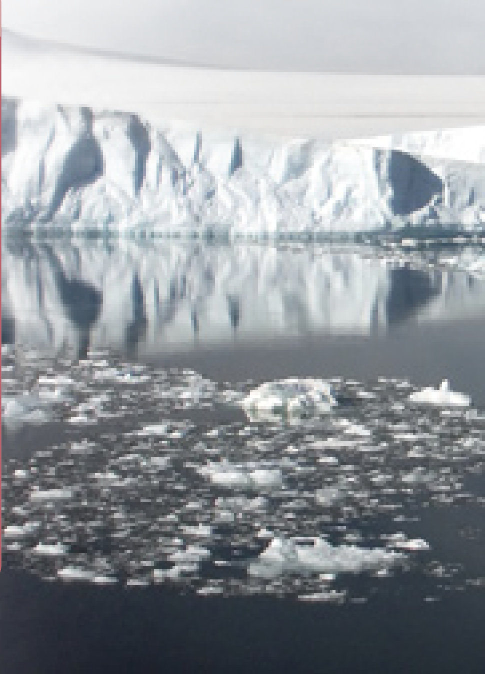

ICE DIVER
There are more diving opportunities in the Arctic than in Antarctic waters. Diving under the ice demands a high level of expertise and special skills (*see pp.188–89*).

The Antarctic is an isolated continent covered largely in ice: its nearest land masses are over 600 miles (1,000 km) away, and the entire continent is circled by an ocean current. The Antarctic Circumpolar Current is extremely significant, as it essentially cuts the Antarctic waters off from warmer waters farther north. The Arctic, by contrast, is mainly an area of ocean covered in a thick layer of ice, bordered by the land masses of North America, Europe, and Asia, and as such its waters are affected by temperature changes, land runoff, and shallow coastal currents.

LIFE ON THE ICE

The adaptations that have evolved in animal life to cope with extreme polar environments are remarkable. There is a tendency toward gigantism—a larger body size keeps the relative body surface area to a minimum, and therefore the effects of heat loss through the skin are less pronounced. Fish have a high body fat content, and mammals have blubber and fur to insulate them, while behavior such as migration in the winter months means the worst conditions are avoided. The most extraordinary adaptations exist, perhaps, in the fish and invertebrates of polar waters: many have special "antifreeze" proteins in their blood to keep it liquid at low temperatures.

ICE SHELF CALVING SMALL ICEBERG
Up to 70 percent of all fresh water on Earth is held in the Antarctic ice cap. It sits on a huge land mass, while the Arctic is largely floating ice.

The open ocean

Only a tiny proportion of the world's oceans cover the shallows around land. Beyond, great tracts of open water plunge to abyssal depths. At first glance a featureless marine desert, this is actually a fascinating ecosystem containing some of the most impressive marine creatures.

A LAYERED WORLD

The oceans of the world are divided not only geographically, but also vertically, descending through the water column in distinct zones defined by temperature. Only the upper 650 ft (200 m) or so are heated by the Sun. Below this warmer surface layer, the temperature of the water can fall dramatically with depth in a layer of water called the thermocline. This layer of rapid temperature change is present in tropical waters year-round and in temperate waters in summer, but is not found in polar waters. Below the thermocline, the temperature continues to fall, but more slowly. Many animals find the dramatic change in temperature at the thermocline stressful, and are confined to either the warmer water above it, or the deep, cold water below it.

The deeper layers of the ocean are richer in nutrients than the surface layer; they are also colder, and so have a greater capacity for holding oxygen. Mixing of the layers—bringing oxygen-and nutrient-rich water into the surface layer—can be

LIFE IN THE OPEN SEA

Even in remote areas of open ocean, communities of living creatures thrive. In the Sargasso Sea in the North Atlantic, a vast rotating current propels floating *Sargassum* weed in an endless circular procession. The weed supports many invertebrate species and juvenile fish.

Sargassum weed

CAUGHT ON CAMERA
Open-water predators, like this blue shark, will investigate any new visitor to the environment.

caused by currents, storms, or other natural patterns of water circulation. When the nutrient-rich water is hit by sunlight, an explosion in plankton growth, or "bloom," occurs, creating regions of abundance as plankton feeders, and in turn their predators, congregate.

BUILT FOR SPEED
With their streamlined bodies and powerful sickle-shaped tails, dolphins are perfectly designed to hunt even the fastest-moving prey.

LIFE AT THE DEEP END

Pelagic animals (those of the open ocean) have developed remarkable physical and behavioral adaptations to suit their demanding environment. Massive vertical migrations take place at night, when marine animals move into the well-stocked surface layers, hidden under cover of darkness from predators that hunt by sight. However, many of the predators of the open ocean boast a battery of senses for detecting prey at long range, and investigate any object in their vast territory for its food potential. These opportunistic hunters include the blue shark, marlin, and tuna, some of the most awesome and beautiful animals in the oceans. The cold-blooded tuna can heat its blood to warm its muscles, allowing deadly bursts of speed.

Freshwater environments

Branching out from the sea to explore freshwater environments opens up a whole new range of dive sites to discover, from Icelandic ravine lakes to Mexican caverns or the Florida swamps, each with fascinating, freshwater-specific animal and plant life.

LAKES AND RIVERS

The first thing any diver accustomed to seawater will notice in a freshwater system is the difference in buoyancy. The salt content of seawater—around 35 parts per thousand—makes it denser, and thus a naturally more buoyant environment for the diver than freshwater. Consequently, you will find that you need to carry considerably less weight. The lower salt content also means that equipment suffers less damage through corrosion.

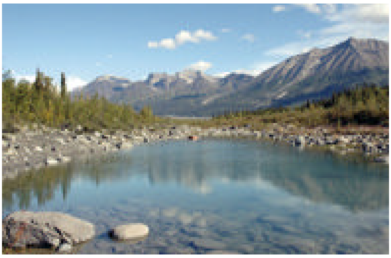

PRIVATE PROPERTY
Some of the most scenic rivers and lakes are privately owned, so it is important to check whether permission is required before you dive.

There are a number of logistical and safety issues specific to freshwater diving. Some of the more adventurous dive sites are high-altitude lakes, and require the use of special tables or a dive computer to avoid decompression issues. Lakes are susceptible to land runoff, and may quickly become murky in poor weather, while rivers may swell rapidly in adverse conditions, creating dangerous currents. Rivers and lakes may also support disease, notably Weil's disease and bilharzia, particularly in regions of dense human population.

Despite these drawbacks, with some research and strategic preparation before a dive, local rivers and lakes can provide a perfectly viable and exciting alternative to sea diving. Freshwater systems become anoxic (lacking in oxygen) at depth, and subsequently support very little life. This means that shipwrecks do not deteriorate as quickly—because the organisms that would usually destroy them cannot function—so deep lakes can offer impressive collections of well-preserved wrecks.

African lake fish

NATURE ON DISPLAY
Freshwater species are generally unaccustomed to divers, creating wonderful opportunities for close observation and encounters.

LAKE DIVING
Exploring a freshwater lake can provide a fascinating and exhilarating alternative to sea diving. Here, in the beautiful Blindsee mountain lake in Austria, divers can explore an underwater forest of fallen trees.

Mammals

Of the 5,000 species of mammals on Earth, around 2.5 percent are found in marine environments. These warm-blooded aquatic animals are perhaps among the most charismatic and approachable creatures you will encounter on any dive.

WARM-BLOODED LIVING

Mammals have a number of features that make them unique in the animal kingdom. First, they are warm-blooded, which means that they are able to regulate their internal body temperature. Most are covered in hair (apart from Cetaceans) and all, except monotremes like the duck-billed platypus, give birth to live young and feed them on milk produced by milk-secreting glands. At least three orders of mammals include species that spend the majority, if not all, of their lives in water: Cetaceae, which includes whales, dolphins, and porpoises; Carnivora, including the pinnipeds (seals, sea lions, and walruses) and otters; and Sirenia—manatees and dugongs.

WHALES, PORPOISES, AND DOLPHINS

Cetaceans spend their entire life in water, even though, like other mammals, they have lungs and need to breathe air to live. They are divided into two main subgroups—Mysticeti and Odontoceti—based on their adaptations for feeding. Mysticeti are filter-feeding baleen whales, such as northern and southern right whales, humpback whales, and blue whales.

SOUTHERN RIGHT WHALE
There are few more impressive sights than a whale's flukes breaking the ocean's surface.

SUPREME PREDATOR
The majestic orca or killer whale is a versatile hunter, catching a range of prey from fish to seals and even whales and birds.

Odontoceti, by contrast, have teeth and include some species of whales and all dolphins and porpoises.

AQUATIC CARNIVORES

Pinnipeds are aquatic mammals that spend some time on land, particularly for breeding. There are 33 species, including seals and fur seals, sea lions, and walruses. These carnivorous mammals have webbed feet, enabling them to move on land and in the water. With their long front flippers, fur seals and sea lions are able to walk on all four flippers on land.

Sea otters are the largest members of the mustelid or weasel family and are most commonly found off the coasts of California, Alaska, and Russia. Unlike other marine mammals, they have no insulating layer of blubber, but have the thickest fur in the animal kingdom. They are one of the only mammals, other than primates, known to use tools, utilizing stones and other items to crack open shellfish and pry prey from cracks in rocks.

THE SIRENIANS

It is rare to encounter manatees (*see also p.309*) and dugongs. These fascinating migratory mammals, belonging to the order Sirenia, typically reach adult sizes of around 10 ft (3 m) in length, and have front limbs in the shape of flippers and no hind limbs. Despite their imposing size, manatees and dugongs are placid, herbivorous creatures that spend most of their time grazing.

SEALS

Commonly seen on dives and at the water surface, seals tend to be found close to shore and in the more remote areas of the world, such as the Arctic and Antarctic, where they can feed and reproduce without disturbance. There are 18 species, the largest of which is the elephant seal. Seals are efficient swimmers, using their rear flippers for propulsion and their front flippers for steering. These lively mammals are fun to swim with but have sharp teeth and can be aggressive during the pupping season.

Harbor seal

Sharks, skates, and rays

Like other cartilaginous fish (also known as the Chondrichthyes), sharks, skates, and rays have a skeleton made of flexible cartilage rather than bone. Collectively known as the elasmobranchs, they have between five and seven paired gill openings, and nostrils beneath the snout.

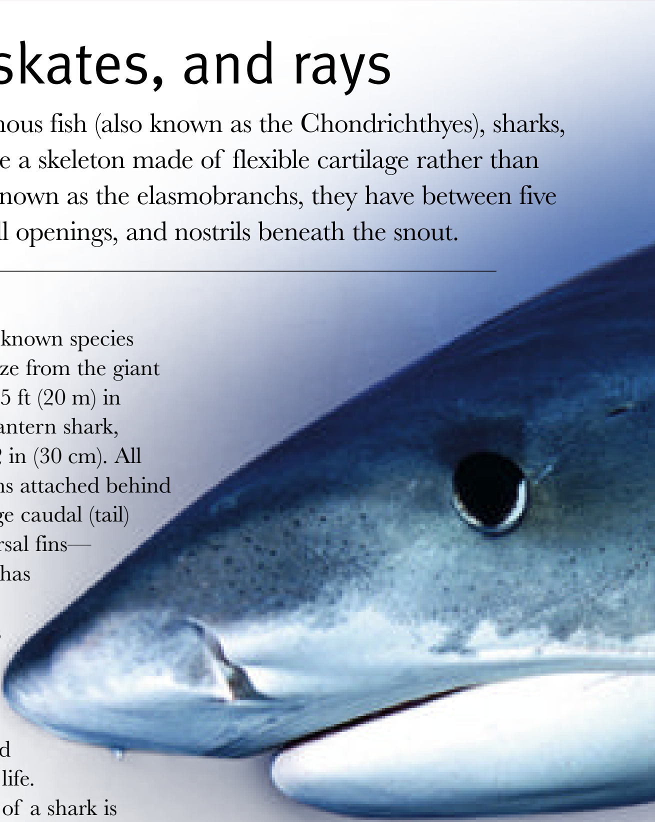

SHARK SPECIES

There are around 500 known species of sharks, ranging in size from the giant whale shark, at up to 65 ft (20 m) in length, to the minute lantern shark, which measures just 12 in (30 cm). All sharks have pectoral fins attached behind the gill openings, a large caudal (tail) fin, and one or two dorsal fins—a basic body plan that has changed very little in hundreds of thousands of years. The jaws are lined with many rows of teeth, which are continually renewed throughout the shark's life. The tough, rough skin of a shark is covered by dermal denticles—tiny scales that minimize drag (friction with the water) when swimming. Sharks propel themselves by sweeping their caudal fin from side to side. Unlike most other fish, sharks have no swim bladder, so they must swim continuously or they will sink. They also need to maintain the flow of seawater over their gills so that they can extract oxygen for respiration.

Sharks have extraordinary sensory powers. Their highly developed senses of smell and taste are used to detect potential prey, and they have excellent vision. In addition, they can sense changes in water pressure caused by the movement of other animals using their lateral line system. This consists of fluid-filled channels along the sides of the body that are lined with sensory hair cells. Sharks can also detect the minute electrical fields produced by other living animals, using electrosensory organs called ampullae of Lorenzi, which are located on the head.

LEOPARD SHARK
The leopard shark can be found at depths of up to 100 ft (30 m) and is often seen lying on sandy bottoms.

TIGER SHARK
Although rarely encountered, the tiger shark is one of the most voracious shark species. It feeds on large fish, turtles, and sea birds.

FOOD AND FEEDING

All sharks are carnivores, but they feed on a wide range of prey and use a variety of feeding techniques. Filter-feeders, such as basking, megamouth, and whale sharks, sift plankton from the water as they swim. Many bottom-dwellers, including wobbegongs and angelsharks, lie camouflaged on the seabed ready to ambush passing fish, while others, such as leopard sharks, feed largely on seabed invertebrates. Some species, such as white-tip reef sharks, feed cooperatively, hunting in packs and actively pursuing prey. The mako and the great white shark tackle large prey such as seals, swordfish, other sharks, porpoises, and sea turtles.

Although sharks are the top marine predators, only a few species are potentially dangerous to people, and attacks on humans are rare. In fact, humans pose far more of a threat to sharks, with many shark species currently endangered, and some even facing extinction, due to overfishing.

SHARK REPRODUCTION

During mating, the male uses his claspers (modified extensions to the pelvic fins) to transfer sperm to the female. The majority of sharks, including dogfish and leopard and tiger sharks, are viviparous, giving birth to fully formed offspring. Oviparous (egg-laying) species lay their eggs on the seabed, where the embryos are nourished by the yolk within the egg case. Catshark egg cases have long tendrils that twist around objects on the seabed. In other species, such as epaulette sharks, sticky filaments glue the egg cases to a secure site. All young sharks fend for themselves from birth or hatching.

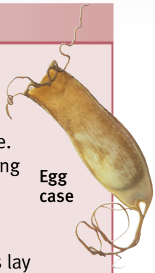

Egg case

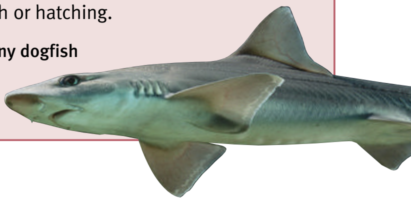

Spiny dogfish

SKATES AND RAYS

The 600 skate and ray species share many of the features of sharks, but have a very different body plan. The gill slits are on the underside of the body. The pectoral fins are positioned above the gill openings, not behind them as in sharks, and are greatly enlarged and flattened to form "wings," while the caudal fin is reduced or absent.

Most skates and rays are bottom-dwellers, and are well adapted for life on the seafloor. Their flat body shape enables them to lie on the seabed unnoticed, and they have a downward-pointing mouth and eyes on top of the head. Spiracles (breathing holes) behind the eyes enable them to "breathe"—take in water to extract oxygen—when half-buried in the substrate. Most rays and skates eat invertebrates, such as mollusks and crustaceans, and small fish, but some are plankton feeders. They swim with strokes from their modified pectoral fins, in an undulating movement similar to a bird flapping its wings. Their tails are used for defense rather than propulsion: they have a stinger at the base of their tail, with short barbs that can inject venom into aggressors. Skates lack this tail barb.

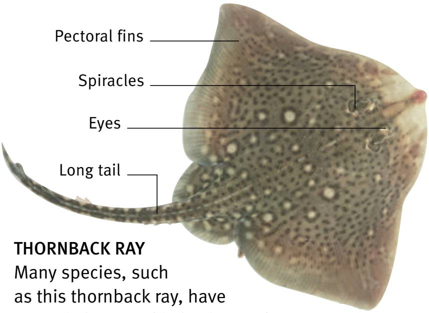

THORNBACK RAY
Many species, such as this thornback ray, have a mottled or speckled coloring that camouflages them against the seabed.

Skates tend to be more common in temperate waters, and rays in the tropics, but the main distinction between them is in their reproduction. Rays produce live young, while skates lay eggs in leathery cases (often called "mermaid's purses").

MANTA RAYS
These large, filter-feeding rays sieve microscopic plankton from the water with their gill-rakers. Manta rays also eat small shoaling fish.

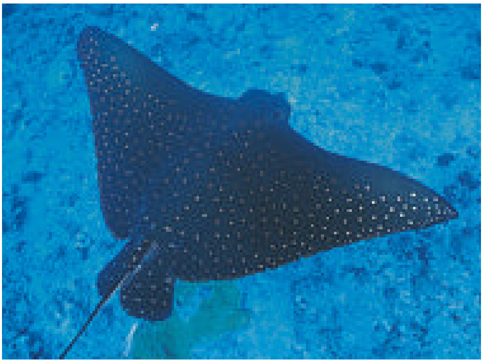

SPOTTED EAGLE RAY
This species uses its flattened, duckbill-like snout to dig up food from the seabed.

DIVING WITH RAYS

Manta rays, with their elegant, dancelike swimming motion, are perhaps the most popular ray species. They are also the largest of the rays, measuring up to 20 ft (6 m) across. They often congregate at special cleaning stations on the reef, where they allow smaller reef fish to pick tiny parasites off their skin. The flaplike lobes on either side of the mouth funnel plankton between their jaws as they swim. When not in use, the lobes are coiled up to make the ray more streamlined.

Other types of rays seen during dives include guitarfish, which have pointed heads and elongated bodies, and electric rays, which possess organs that can deliver an electric shock to incapacitate prey. Torpedo rays can discharge more than 200 volts—enough to stun a diver. Circular-bodied stingray species should also be treated with caution, since they have venomous barbs at or near the base of the tail. The sting's effect can be mildly painful to potentially fatal, depending on the species. Eagle rays, which also have venomous barbs, can be identified by their distinctive beak-shaped snout. They have powerful crushing teeth, large "wings," and a long tail.

Rays are docile animals, and despite their self-defense mechanisms, pose no real threat to the diver if unprovoked.

Bony fish

Bony fish, or Osteichthyes, make up about 95 percent of all fish species—more than 23,500 in number. These vary in every ocean that you will dive in, and there is a wealth of shapes, sizes, and colors, and intriguing behaviors and interactions for you to discover.

COMMON FEATURES

Typical bony fish are torpedo-shaped and streamlined with a muscular tail. They have a caudal fin at the base of the tail for propulsion; a pair of pectoral fins, just behind the gills, for steering; an anal fin just below the tail; a pair of pelvic fins on the lower side; and, normally, one to three dorsal fins on the back. They are covered in a slimy secretion that reduces drag, and nearly all species have a covering of protective scales. They breathe using gills, although a few have evolved to use lungs as well as gills, and they are cold-blooded.

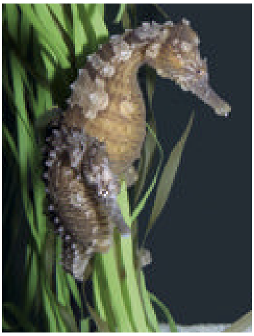

ROLE REVERSAL
Male seahorses have a pouch in which they incubate the eggs produced by the female.

Bony fish are the most diverse and numerous of all the vertebrates. The earliest fossilized fish remains date from around 395 million years ago. All fish have a bony skeleton and a swim

NAPOLEON WRASSE
These lumbering open-water dwellers can become relatively tame in frequently dived sites.

SAFETY IN NUMBERS
Moving as part of a large group means that each member of this shoal of barracuda—a ray-finned species of bony fish—is less likely to be eaten.

bladder—an air-filled sac that aids buoyancy. The sexes are separate; most lay eggs, and fertilization is usually external.

There are two groups of bony fish. The Sarcopterygii (lobe-finned fishes), which have a fleshy lobe at the base of their fins that is leglike in appearance, form a small group that includes the coelacanth and Australian lungfish. These ancient fish probably gave rise to the amphibians. But by far the majority of bony fish belong to the Actinopterygii (ray-finned fishes). These fish have jointed bony rays that support their fins, and a symmetrical caudal fin.

The species you see will vary depending on your location. Fish identification can be rather like bird-watching, and many divers make checklists of what they have seen on each dive. An identification guide to the fish in the area that you are diving in will help to identify them correctly.

A WORLD TO DISCOVER

One of the greatest attractions of diving is the opportunity to observe an amazing range of fish species in their natural habitat. Bony fish display a wide variety of colors and patterns, especially those species found around coral reefs. As in some bird species, the male is often brightly colored and the female is drab in contrast. A male's vivid coloring may be used to attract a female in courtship displays. For example, the vivid markings of the male cuckoo wrasse become even more colorful in spring to attract females.

On reefs, you may also have the opportunity to observe some fascinating interactions between species. Animals such as manta rays and turtles regularly visit "cleaning stations," where smaller fish, such as cleaner wrasse, pick off algae, parasites, and dead tissue from

their skin and even from inside their mouths. Anemonefish (also called clownfish) are able to live among the stinging cells of host sea anemones because they secrete a special mucus that protects them from being stung. In return for their sheltered home, they drop pieces of food and attract other fish that the anemone can feed on.

AVOIDING PREDATORS

You'll need keen eyes to spot some of the many fish species that are camouflaged or have unusual ways of avoiding predators. Some of these species are so elusive that they have only recently been discovered. Masters of disguise include the scorpion fish, which hides on reefs and is virtually indistinguishable from the coral and algae that surround it, but can deliver a nasty sting should you step on it. You may encounter flatfish, such as plaice and flounder, when diving in areas with a sandy bottom, but many are so well camouflaged to match the seabed that they are extremely hard to spot.

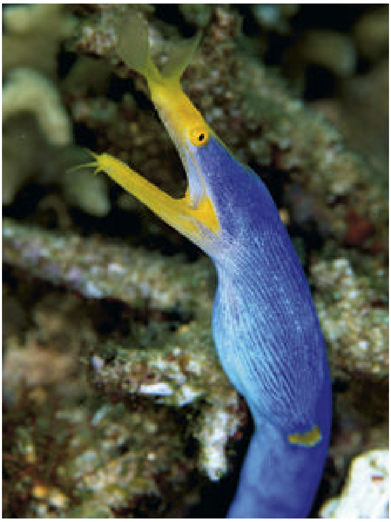

BLUE RIBBON EEL
This small, elegant moray eel from the Indo-Pacific is a shy creature: although it seems to gape threateningly, in fact, it is simply breathing.

The holes and crevices in many reefs and wrecks provide ideal hiding places for fish such as moray and conger eels, with their slender body shape, large mouths, and menacing-looking teeth. Their hiding place is revealed when they gape to breathe (they have restricted gill openings). Eels have no pectoral or pelvic fins and use their long, muscled bodies to dart in and out of their lairs.

ADAPTIVE CAMOUFLAGE

Bizarrely, flatfish start off life with a "normal" fish shape but then undergo a metamorphosis into their characteristic flat body form. One eye migrates to the other side of the head, the skull deforms, and the mouth becomes asymmetrical. The side of the body that sits on the sea floor becomes lighter and the upper side becomes darker, camouflaging itself in the color of the substrate.

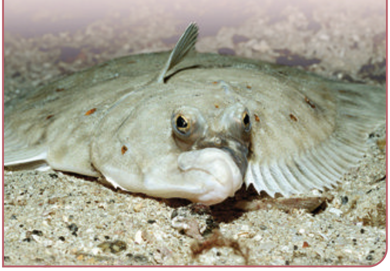

SOCIABLE FISH

In open water you may encounter incredibly large fish like the Napoleon wrasse, which can reach a weight of nearly 300 lb (135 kg). In some areas the wrasse have been hand fed so often that they are friendly toward divers, though feeding them is discouraged. In more open water and around wrecks, you will see great shoals of fish such as sardines, barracuda, glassfish, and mackerel. Such ocean-going, or pelagic, fish tend to have pale bellies, silvery flanks, and dark-colored backs. This makes it hard for prey and predators to spot them—when viewed from above, the darker shade blends in with the color of the deep sea; viewed from below, the pale coloring merges with the light from the sky.

IMMUNE TO STINGS
A covering of protective mucus allows the anemonefish to shelter from danger among its host's tentacles.

Reptiles

Most reptiles live on land, but some are aquatic, notably turtles, terrapins, crocodiles and the related alligators and gavials, and sea snakes. These cold-blooded, or ectothermic, animals are mostly found in tropical areas, but some live in temperate regions.

BIRTH AND HATCHING

Reptiles are typically oviparous (the females lay eggs in which the young develop), although a few species are ovoviviparous (the young hatch from eggs within the mother's body). Egg-laying reptiles often build a nest in which to incubate their clutch.

One of the most prolific reptilian egg-layers is the turtle (order Chelonia), which has a hard shell that encloses and protects its vital organs. The head, tail, and flippers protrude, but many species can draw them in when threatened. Although they are air-breathing creatures, they only leave the water to breed and can lay as many as 200 eggs in nests excavated in sandy beaches.

THE SNEEZING IGUANA

The only lizard that forages for its food in the sea, the marine iguana has a curious habit of "sneezing" salt. The marine algae that forms the main part of its diet contains a lot of salt. A gland connected to the nostrils removes the salt from the body, and the iguana then expels it by sneezing, leaving its face white. This is one of the unique adaptations that allows marine iguanas to live and feed in seawater.

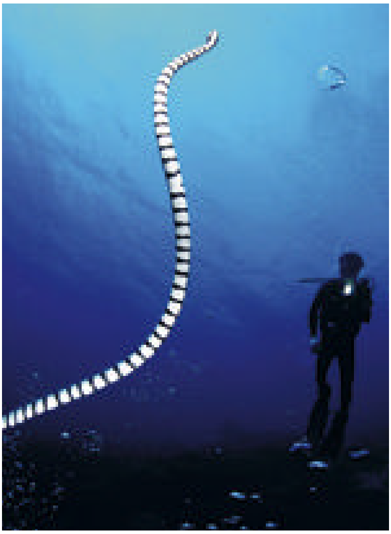

DIVER AND BANDED KRAIT
Related to cobras, sea snakes are venomous, but fortunately they are quite timid and if unprovoked are not inclined to attack humans.

SEA SNAKES AND LIZARDS

Although many snakes, such as the banded krait, swim in the sea, they usually return to land to breed. The yellow-bellied sea snake is the only true sea snake, the female giving birth in the water to up to ten live young, each about 10 in (25 cm) long. It can grow up to 3¼ ft (1 m) in length, and is black or dark brown in color with a bright yellow underside. This sea snake has spread across the Pacific from Southeast Asia to America, and gathers in huge numbers during the breeding season: a really big swarm can stretch for 70 miles (115 km). Marine iguanas, from the Galápagos Islands, are the only lizards that can tolerate salt water. Males may grow to a length of 4 ft (1.2 m) or more (almost half of which is the tail). They are harmless herbivores, feeding exclusively on algae growing on rocks near the shore. When feeding, they can remain submerged for up to an hour, especially in the warmer shallows, though dives of five or ten minutes are more common.

SALTWATER CROCODILES

The saltwater crocodile is the world's largest living crocodile; record lengths have been up to 23 ft (7 m), but 13–16 ft (4–5 m) is more common. This is the most dangerous species of crocodile, responsible for most crocodile attacks on people. It is found from eastern India and Bangladesh through Southeast Asia to Papua New Guinea and Australia, usually in brackish water in tropical estuaries, rivers, and swamps. The female lays eggs under a large mound of vegetation and stands guard over them for about 100 days. When the young hatch, she digs them out and carries them to the water. The hatchlings then live in nursery groups for about eight months, guarded by the adults.

RETURNING TO THE SURFACE
Turtles breathe air, like land-based reptiles. This one has a parasitic remora fish attached.

Marine invertebrates

Of all animal species on Earth, 95 percent are invertebrates (animals with no backbone), forming an incredibly numerous but often overlooked section of the animal world. Every major invertebrate group has an array of marine representatives.

AQUATIC ARTHROPODS

An arthropod is a type of invertebrate that has its body encased in a hard outer casing (the exoskeleton) and jointed appendages. Land-based arthropods include insects and arachnids, whereas most underwater species are crustaceans, a group that includes crabs, shrimp, lobsters, water fleas, copepods, and barnacles. There are about 39,000 species of crustaceans, most of which are marine, though some live in freshwater environments. They vary in size from the massive Japanese spider crab, with its leg span of 12 ft (3.7 m), to the tiny water flea, only about 9/1000 in (0.25 mm) long. But large or small, crustaceans all have certain things in common. They possess two pairs of antennae, three pairs of mouth appendages, and compound eyes on stalks. They also have three body segments—the head and the thorax, which are almost always fused and covered by a carapace (an outer shield), and the abdomen.

DRAWING NOURISHMENT FROM THE SEA
Bivalve mollusks, such as these mussels, feed by filtering organic matter from the water. Mussels live in nutrient-rich tidal areas.

MOLLUSKS

All mollusks, except the octopus, possess some form of calcium carbonate shell. There are three major marine classes of this group of invertebrates. Gastropods, with some 35,000 species, include whelks, limpets, sea slugs, and conchs. Many possess a shell but some, like the sea slug (or nudibranch), have lost theirs completely. Bivalves, such as clams, mussels, oysters, and scallops, have a shell in two pieces. There are 20,000 marine and freshwater species, the largest being giant clams up to 4 ft (1.2 m) across. Cepahalopods comprise around 600 species, including nautilus, squid, cuttlefish, and octopus. The cephalopods

SEA SCAVENGER
Crabs have a varied diet; these crustaceans feed on carrion, mollusks, and even other crabs.

DELICATE DRIFTERS Transparent and serene, jellyfish are members of the cnidarian group of invertebrates.

include the largest invertebrate of all, the giant squid, which grows to a size of 52 ft (16 m). Cephalopods can swim in short bursts by squirting water from their bodies, but many, like the squid, use undulating fins for locomotion. The only shelled species of cephalopod is the nautilus: squid and cuttlefish have internal shells and the octopus has no shell at all.

CREEPING BEAUTY
Nudibranchs are colorful mollusks related to the common garden snail.

ECHINODERMS

Marine bottom-dwelling invertebrates, echinoderms have a hard, spiny skin and five-part radial symmetry. They lack a head, are often brightly colored, and are distributed throughout most of the world's seas. There are approximately 6,000 species of echinoderms, including starfish, brittle stars, sand dollars, sea urchins, and sea cucumbers, among others.

Most species of echinoderms reproduce sexually and have separate sexes. Fertilization is external—the eggs and sperm are shed into the water at spawning time. The floating embryo develops into a planktonic larval form. A few species reproduce asexually by fission (genetic copying of the parent organism into two or more clones).

Starfish may be circular, pentagonal, or the familiar star shape, with radiating arms. Under each arm are tiny tube feet, which the animal uses for walking.

UNDERWATER CHAMELEON
Cuttlefish are mollusks that can camouflage themselves by changing the color and texture of their skin.

The mouth is on the underside and the starfish feeds on living and dead animals as it moves along rocky surfaces and muddy bottoms.

Sea urchins, on the other hand, do not have arms. Instead, they have a shell of tightly fused plates, forming a sphere. They have long, prominent spines that protect them from predators, and have tube feet on their undersides, which are used for locomotion, respiration, and for trapping drifting particles of food.

CNIDARIANS

The cnidarian group of invertebrates, which includes the jellyfish, sea anemones, corals, and hydroids, are species with one of two basic body forms: medusae or polyps. Medusae, such as adult jellyfish, are bell-shaped with tentacles hanging down from the margins. They tend to be free-floating and solitary. Polyps, such as coral, sea anemones, and hydra, are tubular-shaped animals topped with tentacles. They frequently form colonies. Corals (*see also pp.204–05*) are unusual in that they depend on symbiotic algae to provide life-sustaining sugars. Medusae reproduce sexually, whereas most polyps reproduce by asexual fission. Cnidarians are unique in manufacturing cell-based stinging mechanisms, called nematocysts, in their tentacles, which are used to catch prey and in defense.

OTHER INVERTEBRATES

Comb jellies, which are bell-shaped creatures with vertical combs over their bodies, form another class of invertebrates—the ctenophores. The 5,000 species of sponges that live in both salt and fresh water belong to the phylum Porifera. Finally, there are over 8,000 species of segmented worms, which are classed as annelids.

PLANTS OR ANIMALS?

Many marine creatures stay rooted to a single spot for life. They posed a dilemma for early biologists who believed that mobility defined the animal state; this explains the common names of animals like the "lily of the sea" – in fact a crinoid. Other stationary animals include polyps, sponges, feather stars, and annelids such as fanworms. They are filter-feeders, and contain toxins to deter predators.

Feather star

Places to dive

Introduction

With increasing numbers of dive centers opening up around the globe, virtually the entire world is now accessible to the traveling diver. Following a few basic travel tips can stand you in good stead when transporting yourself and your dive gear to your destination.

TRAVEL OPTIONS

Most vacation destinations in coastal regions around the world provide some sort of diving facility. Indeed, the sport is now so convenient that you can turn up on the day, certification card in hand, and be decked out from head to toe with rented gear, before being guided around local dive sites. This is one of the greatest joys of the sport—the ability to visit different parts of the world and explore local sites with relative ease.

A step up from the more casual "just-passing-through" diving experience is the dedicated diving vacation. This is usually arranged through a resort-based dive operator, most of which will run a series of dives to local dive sites and offer a range of equipment for rent.

ICE DIVING
With many divers looking for more adventure in their diving, travel companies are offering challenging experiences, such as ice diving.

Developments in diving equipment over the last decade mean that most dive centers worldwide can offer a reasonable caliber of gear for rent. With the addition of local knowledge from a divemaster

LIVEABOARD DIVE VESSEL
A liveaboard is essentially a floating dive center, and as such should have all the facilities on board to guarantee safety, comfort, and a memorable diving adventure.

who knows the waters well, you should be assured of a reasonable diving experience, regardless of the location. While most dedicated dive operators work to a high standard, there are a few rogue operators around, and you should be wary of any that show a cavalier approach.

Liveaboards provide an even more intense diving vacation. These are dive vessels designed for long-range trips to more inaccessible places, and they provide nonstop diving throughout the voyage. Liveaboards often offer supplementary courses, such as photography or natural history, hosted by resident experts. There are few marine dive destinations on Earth that are not accessible via liveaboards, and for many this is the finest way to experience extensive reef systems, remote sites, or sprawling island groups.

There is also a range of diving expeditions available in the modern travel market. Some are conservation-oriented and others more adventurous, such as seeking out big animal encounters or spectacular wrecks. Take the time to investigate the results and credibility of the group coordinating your project, since some that claim high ethical standards in conservation and research fall well short of expectations. There are others, of course, who perform genuinely valuable work, and for many divers this is the most rewarding and educational aspect of their entire diving career.

LOCAL KNOWLEDGE
Most dive centers will be able to take you out on a boat and show you the best of the local dive spots.

EQUIPMENT RENTAL
Renting gear when you reach your destination is one way of lightening your travel burden and avoiding penalties for carrying excess baggage.

TIPS AND PRECAUTIONS

Diving overseas involves putting a great deal of trust in your host, be it a local guide or dive operator, or a leader of a trip or expedition. Following a few simple rules and precautions will make this a less daunting prospect, and may also ease the logistical problem of transporting expensive equipment across the world.

Deciding on a destination is the first step in planning a successful trip. Finding a location that really appeals to you, yet stays within your budget, will make your trip more rewarding than simply opting for the best-known sites or the cheapest deals. A brief period of research can reap huge rewards, and the range of information sources now available makes this a relatively simple exercise.

After identifying the aim of the dive trip, and the location to be visited, the next stage is to choose a host operator. Many local tourist boards can recommend registered operators who are known to work to a high standard. In addition, the PADI website (*see pp.34–35*) lists dive operators who have met a series of demanding criteria. This independent evaluation is invaluable in identifying the most professional centers and boats.

While rental equipment is generally of a good quality, many divers prefer to travel with their own gear. This is prudent, since unfamiliarity with rented gear can create an uncomfortable (and

CHOOSING A DIVE BAG

While you can pack your gear in any case or bag you wish, most divers use bags specially designed to stand up to the rigors of dive traveling (*see p.79*). Soft bags, such as duffle bags, can fold up and be stashed away at your destination. Semirigid wheeled cases make carrying easier, but they are less flexible. Rigid bags are heavier, but offer better protection for your gear.

mares
mares

Semirigid wheeled bag

occasionally unsafe) diving experience. The diver also has the assurance of knowing that their own gear fits well, is functional, and has been tested recently.

Most airlines do not officially offer excess baggage allowances for dive gear (the dive industry is lobbying hard for a more reasonable approach to this issue), but a phone call in advance of the trip often works wonders—as does a courteous approach when checking in. Carry delicate items, such as regulators and dive computers, in your carry-on; the latter, in particular, may be affected by pressure changes in the hold of the plane.

Special diving insurance is vital, since you will not be covered by general travel policies. You need cover for damage to equipment in transit, medical care abroad (such as recompression treatment) and repatriation in the event of a serious diving incident. Policies usually cover divers to a depth of 100 ft (30 m)—more than adequate for most trips.

ONWARD TRAVEL
If the final leg of your trip or any travel between islands involves flights on small aircraft, check if the carrier imposes their own luggage restrictions.

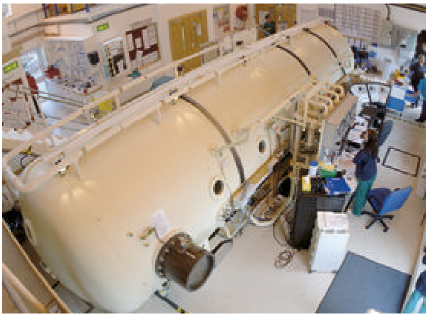

RECOMPRESSION CHAMBER
The costs of recompression—if it is available at all—can be exorbitant, and many chambers will not treat divers without proof of insurance.

On arrival at your target destination, it is always advisable to carry out a few simple checks when diving with the host operator. All dive operators should have emergency oxygen supplies appropriate to the range and scale of diving they provide, as well as first-aid facilities and effective communications. It is worth making a quick visual inspection of the compressor facility—air certificates should be on display, but it is relatively easy to tell a clean, well-maintained compressor from a neglected one.

Europe

As a dive destination, Europe tends to be greatly underestimated by the international diving community. With an excellent diving infrastructure and well-established resorts and travel networks, Europe offers an eclectic mix of dive environments and a host of wrecks to explore.

TEMPERATE ZONE

Europe is, on the whole, a region of temperate waters, and this provides tremendous seasonal variation in the marine life around its coasts. The waters of western Europe are cold and rich in nutrients, and historically have been some of the richest fishing grounds on Earth. Migratory species move around the western edge of the continent, many borne on the undersea highway of the Gulf Stream. Consequently, the same dive site can be unrecognizable from winter to summer.

In diving terms, Europe is a continent of extremes, with Arctic diving to be had in the dramatic Norwegian fjords to the north, while subtropical diving can be enjoyed in the warm waters of the Mediterranean Sea to the south.

FISHING BOATS AT ALBUFEIRA, PORTUGAL
Europe's fish stocks have been an important resource for millennia, but overfishing has now put the populations of some species at risk.

Europe has long been the scene of intense shipping activity, a fact reflected in the thousands of wrecks that today dot the seabed. These wrecks represent a history of seafaring itself—from Viking longships, to all types of fishing vessels, and merchant fleets that foundered in violent storms. Sunken warships from virtually all of the continent's navies bear witness to the many wars that have ravaged Europe through the ages.

NORWEGIAN FJORD
Carved by glaciers millions of years ago, the fjords offer excellent dives in crystal-clear water, against a superb backdrop of snow-capped mountains.

WRECK OF THE *ZENOBIA*, CYPRUS
The Swedish ferry *Zenobia*, which sank in 120 ft (40 m) of water off Cyprus in 1980, is one of Europe's most impressive wrecks.

Wreck of the *Zenobia*

LARNACA HARBOR, CYPRUS

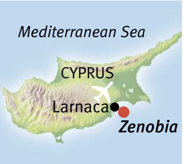

A wreck for all grades of divers, the *Zenobia* offers shallow and deep dives, and a range of penetrations from gentle investigations through to technical explorations. Readily accessible, this wreck is one of the most popular dive sites in Europe.

ACCIDENTAL SINKING

In June 1980, the 10,000-ton Swedish passenger ferry *Zenobia* ran into difficulties off the coast of Cyprus. A problem with the autopilot caused the vessel to heel over and sink over the course of several days, giving plenty of time for the passengers to be evacuated, but leaving the cargo, including 135 trucks, to sink with it.

ESSENTIAL INFORMATION

WATER TEMPERATURE 70–79°F (21–26°C)

ECOSYSTEM Subtropical

WHEN TO VISIT Divers can visit year-round, but diving is at its best between March and October.

MUST-SEES Vehicle hold Number One; the bridge; the accommodations area.

REVERED WRECK

The *Zenobia* has become one of the most popular wrecks in diving because of its accessibility and the range of dives it offers. Divers can access the two main vehicle decks, the bridge, the accommodations area, and the restaurant. More extreme penetrations of the wreck include the engine room, although this is a multilevel, technical dive that requires careful planning and training. While not suitable for the novice, such dives are thrilling challenges. For those seeking a more sedate experience, the exterior of the wreck also provides a dramatic spectacle, with its props, huge hull, and superstructure still in excellent condition.

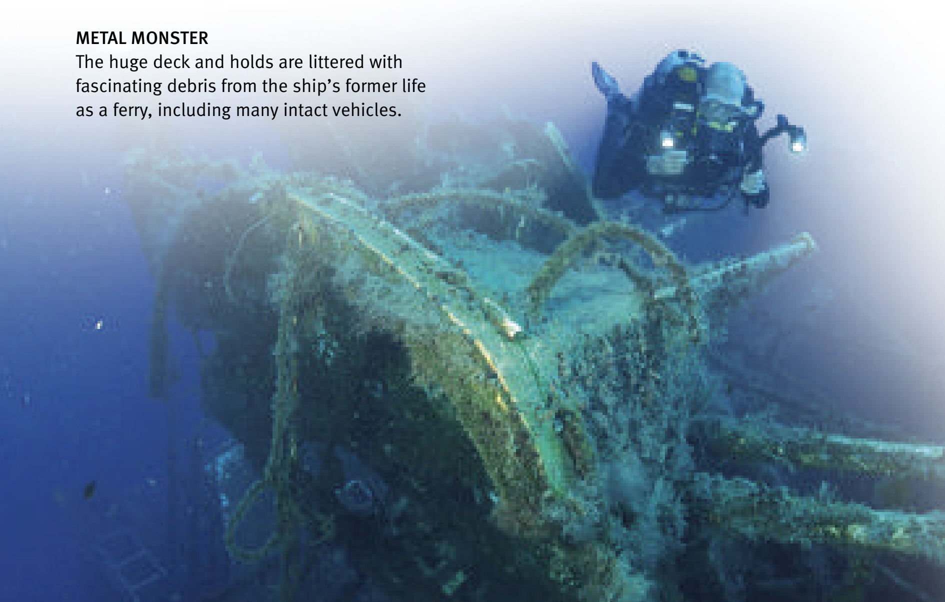

METAL MONSTER
The huge deck and holds are littered with fascinating debris from the ship's former life as a ferry, including many intact vehicles.

Medas Islands

CATALONIA, SPAIN

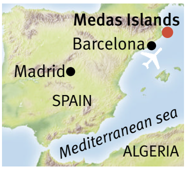

The Medas Islands offer a glimpse of how Mediterranean marine life was in days of old—large fish shoals, passing pelagics, densely colonized reefs, and fine coral formations. Closely protected, the islands have much to offer visiting divers.

CONSERVATION AREA

The Medas Islands were granted the status of a Nature Reserve in 1983, and have therefore enjoyed more than 20 years of official protection. The life around the islands reflects this, with perhaps a greater abundance of marine life and range of species than anywhere else in the western Mediterranean.

OUTWARD BOUND
A dive boat prepares to embark on the short voyage to the islands. Divers flock to the area to enjoy the profusion of marine life sheltering in the clear blue waters of the nature reserve.

DIVING THE ISLANDS

Located only 1 mile (1.6 km) off the north-eastern coast of Spain, the islands benefit from the nutrient rich runoff from the nearby Ter River. In addition, cold water upwellings on the craggy reefs around the islands create an oxygen-rich environment that allows marine life to thrive. As well as the usual Mediterranean species, such as scorpionfish, octopuses, and moray eels, local residents include large shoals of barracuda and bream, and there have even been sightings of eagle rays around the islands. Diving highlights include the famous Dolphin Cave—a 150-ft (46-m) tunnel set in a rock face—and the tiny islet of Carall Bernat. It is also reassuring to note the return of red coral to the seas around the islands, a natural resource that was for many years exploited by the jewelry industry. Though a small fee is levied by the authorities for diving around the Medas Islands, the success of the ongoing conservation policy justifies the extra expense.

ESSENTIAL INFORMATION

WATER TEMPERATURE 57–77°F (14–25°C)

ECOSYSTEM Subtropical

WHEN TO VISIT The best time to visit the Medas Islands is between March and September.

MUST-SEES Dolphin Cave; Carall Bernat; red coral formations; rare sightings of eagle rays.

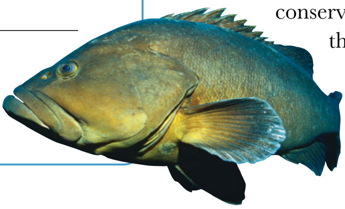

Dusky groupers are just one of the species living under protection around the islands

Sardinia

CENTRAL MEDITERRANEAN, ITALY

ITALY
CORSICA
Rome
Sardinia
Cagliari

A rugged island in the middle of the Mediterranean, Sardinia has a dramatic geological past reflected in the caverns and caves that dot its coast. Although these are impressive features, Sardinia also has unique marine life and fascinating wrecks.

ROCKY COAST

The sharp reefs, dark overhangs, and craggy shoreline of Sardinia play host to an unusual array of marine life for the Mediterranean—a body of water not particularly famed for its marine diversity. Sardinia's geographical position means that it receives species from the Red Sea to the east (thought to be due to animals traveling through the Suez Canal), and the Atlantic to the west. Divers may encounter any of six species of wrasse, two species of groupers, wonderfully colorful sponges and corals encrusting rock walls, and a unique giant mollusk called the noble pen shell.

LOCAL CHARACTER
The common octopus is one of the key species found off the rocky Sardinian shoreline. This highly intelligent invertebrate is a common sight for divers.

NETWORK OF CAVERNS

This coastline is also famed for its caves and caverns. One system—Nereo Cave on the northwest coast—is particularly impressive, stretching for over 1,000 ft (300 m). As with all caves, caverns, and overhangs in Sardinia, such areas of low light are festooned in thick growths of red coral, for which the island is particularly well known.

There are also a number of fine wrecks to be dived. The best known of these is the wreck of the *Angelica*, a 4,400-ton freighter that ran aground on the island's jagged reefs in 1983, at Capo Falcone. The wreck lies at a maximum depth of 65 ft (20 m), so represents a relatively easy dive. The shallow location means that wind and weather have taken their toll on the wreck's structure over the years, but the site is worth a visit should you desire a change of pace from the sea caverns that dominate Sardinian diving. With many attractive beaches and towns along the coast, there is much to enjoy on shore, as well as beneath the waves.

ESSENTIAL INFORMATION

WATER TEMPERATURE 66–77°F (19–25°C)

ECOSYSTEM Subtropical

WHEN TO VISIT Sardinia can be dived year-round, but the summer months offer warmer, if more crowded, seas.

MUST-SEES Red coral; Nereo Cave; noble pen shells; wreck of the *Angelica*.

Ustica Island

SICILY, SOUTHERN ITALY

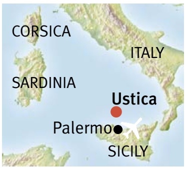

Ustica Island was described by no less an authority than Jacques Cousteau as the most beautiful island dive site in the Mediterranean. A tiny volcanic outcrop off the coast of Sicily, its qualities have seen it established as a Marine Nature Reserve.

SMALL IS BEAUTIFUL

Ustica Island may owe its undisturbed ecosystem to the fact that it was a penal colony until 1950. A tiny island—it has a total surface area of 3 sq miles (8 sq km)—it is nonetheless one of the most significant sites in the Mediterranean, to the extent that Italy's Academy of Underwater Sciences and Techniques is based here. The volcanic origins of the island (seen in the name itself—the Latin *usta* translates as a burnt red color) mean that the reefs around the island are particularly dramatic—with a craggy underwater topography featuring numerous caves, caverns, and overhangs.

ESSENTIAL INFORMATION

WATER TEMPERATURE 59–77°F (15–25°C)

ECOSYSTEM Subtropical

WHEN TO VISIT Diving takes place on Ustica Island year-round, but in winter the seas are relatively chilly.

MUST-SEES Red coral formations; the Colombara Bank; Medico Tunnel.

TOP DIVES

One of the best sites is Medico Tunnel, off the northwest of the island, which consists of a tall pinnacle with a tunnel passing directly through it. Another site of particular note around the island is the Colombara Bank—well known for its beautiful reefs and aggregations of large fish, such as barracuda and tuna.

SILVER SHOAL
Barracuda shoal in large numbers in the waters off Sicily, and are frequently sighted by divers visiting the Marine Reserve at Ustica.

The Hebrides

WESTERN SCOTLAND

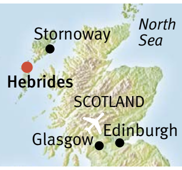

The islands of the Hebrides have the feel of genuine wilderness about them. Precipitous mountains plunge into dark bays and craggy lochs, and white beaches lead to waving kelp forests, creating one of the finest diving regions in Europe.

RICH SEAS

The topography of the Hebrides lends itself readily to abundant marine life, with steep walls in the lochs, convoluted rocky reefs in the open water, and strong tidal movements through numerous narrow channels connecting the islands. Washed in oxygen-rich cold water, the islands are host to a dazzling array of marine life both large and small. Notable sites in the region include dives around Skye (particularly the small island of An Dubh Sgeir), the wrecks and reefs of the Sound of Mull, the famous dives around Oban (home to one of the most established and respected dive operations in the UK), and the newly discovered dive sites of the remote Summer Isles. The last include the wreck of the *Fairweather V*, a fishing vessel that sank in 1991 and is still intact and in good condition.

SURVIVING WILDERNESS
The characteristic rocky coastlines of the Hebrides are perhaps one of the last truly wild places in the United Kingdom.

TEMPERATE-WATER ECOSYSTEM

The Hebridean islands provide one of the richest and most complete insights into the marine life of a temperate-water zone. Visitors to the islands' rocky shores include itinerant whale and dolphin species, ling, conger eel, and a profusion of smaller native fish species. Colonial and encrusting organisms coat the rock walls, with anemones and sponges creating colorful and dramatic reefs. These are patrolled by otters—the Hebrides is one of the few dive sites in the United Kingdom where these fascinating mammals have largely taken to the sea. If conditions do not permit diving, the Scottish Sealife and Marine Sanctuary near Oban is well worth a visit. It has a spectacular aquarium and a busy otter and seal sanctuary.

A diver examines an undersea wall covered in dead men's fingers and plumose anemones

In the winter months, the islands of the Hebrides are subject to wild weather and cold water, with many dives becoming very demanding due to dark water and strong tides. In the summer months, the region really comes to life, with some of the beaches on the more remote islands becoming gateways to absolutely top-class shallow dives.

ANIMAL MAGIC

The shallow reefs just offshore in the Hebrides are some of the busiest temperate reefs anywhere in Europe, playing host to large edible crabs and lobsters, various species of cephalopods including cuttlefish and octopus, colorful nudibranches, and an array of fish species including some spectacular wrasse.

Larger animals encountered in the open seas around the Hebridean archipelago, as well as in its many dramatic coves and bays, include both the common and the Atlantic seal. Bottlenose dolphins are regular visitors to these shores, as are the more rarely seen rissos and white-beaked dolphins. Whale species that can be spotted include minke and pilot whales, orcas, especially around the islands of Coll and Tiree, and, if you are very lucky, sperm whales and humpback whales. There are large gatherings of basking sharks late in the year, although predicting when and where these will take place is almost impossible. Occasionally, sunfish and leatherback turtles also visit the region.

It is not only the animals that provide the excitement for divers visiting this region. The third largest whirlpool on Earth can be seen at the Gulf of Corryvreckan, between the islands of Scarba and Jura.

ESSENTIAL INFORMATION

WATER TEMPERATURE 46—64°F (8–18°C)

ECOSYSTEM Temperate

WHEN TO VISIT Diving is possible all year, but can be stormy in winter.

MUST-SEES Otters; wreck of the *Fairweather V* in the Summer Isles; An Dubh Sgeir in Skye.

Scapa Flow

ORKNEY ISLANDS, GREAT BRITAIN

Orkney Islands
Scapa Flow
North Sea
SCOTLAND
Glasgow
Edinburgh

The most famous wreck-diving location in Europe, Scapa Flow is a huge natural harbor amid the islands of Orkney off northern Scotland. Its fame stems from the fact that it was the site of the largest ship-scuttling action in maritime history.

ANCIENT HARBOR

Although Scapa Flow has been used as an anchorage since Viking times (the name derives from the Old Norse *Skalpeidfloi*, "Bay of the Long Isthmus"), one event ensured its fame in the diving world. After World War I, the German High Seas Fleet was impounded in Scapa Flow by the Allies. The fleet, consisting of 79 vessels, was imprisoned there for over a year. On June 21, 1919, Admiral Ludwig von Reuter, the German commanding officer, misread a newspaper headline and assumed that hostilities were about to resume and ordered his ships to be sunk by their own crews rather than let them fall into the hands of Allied forces.

ESSENTIAL INFORMATION

WATER TEMPERATURE 46–68°F (8–20°C)

ECOSYSTEM Temperate

WHEN TO VISIT Best diving conditions are between March and September.

MUST-SEES Wrecks of *Koln*; *Kronprinz Wilhelm*; *James Barrie*; *Gobernador Bories*; basking sharks; seals.

GRAVEYARD OF SHIPS

The subsequent mass scuttling saw the majority of the fleet sunk, although the Allies managed to save 22 vessels by towing them into the shallows. Several of the German ships were salvaged in the intervening years, but three battleships, four light cruisers, and five torpedo boats still lie on the seabed. In addition to this rich array of wrecks, the British Admiralty sank 50,000 tons of shipping during World War I and II as blockships in the narrow entrances of the Flow to prevent raids by German submarines.

With several other wrecks of varying origins scattered throughout the Flow, the site has become a mecca for serious wreck divers. Although the larger German wrecks can be demanding dives, there are many easier options for the less experienced. The blockships in particular are splendid dives, sitting in shallow tidal channels, washed by cold nutrient-rich waters and teeming with marine life.

Notable wrecks in Scapa are the *Kronprinz Wilhelm*, the *Brummer*, and

RUST IN PEACE
Not all of Scapa Flow's historic wrecks lie beneath the sea. Sandy beaches also bear witness to the past.

the *Koenig*. Amongst the blockships the *Gobernador Bories* has become particularly renowned, and the wreck of the trawler *James Barrie* is a magnificent sight at 131ft (40m)—sitting almost perfectly preserved on the sea floor.

WILD SCAPA

Scapa Flow has an extensive array of marine life, with the shallow wrecks in particular playing host to huge numbers of ballen and cuckoo wrasse, lobsters, conger eels, octopuses, and beautiful nudibranchs. Larger residents include grey seals, dolphins, and basking sharks in the summer months, and also the occasional visiting pod of orcas. Though the wrecks of Scapa Flow are, naturally, its greatest attraction, the marine life here is a welcome bonus.

VALUABLE SALVAGE

During the 1920s, the British salvage firm Cox & Danks raised many of the scuttled German battleships for scrap. Metal from the wrecks is still used today to make sensitive scientific instruments; their steel was forged before the age of atomic weapons tests, and so is a rare source of high-grade metal that is completely free of radioactive particles.

Salvage work at Scapa Flow

GLIMPSE INTO HISTORY
Divers at Scapa Flow can inspect the urchin-covered remains of German battleships scuttled shortly after the close of World War I.

The Isles of Scilly

ENGLISH CHANNEL, GREAT BRITAIN

IRELAND
ENGLAND
WALES
London
Isles of Scilly
English Channel

Something of a well-kept secret among British divers, the Isles of Scilly are a cluster of tiny islands that sit at the tail end of the Gulf Stream. Rich marine life, lively reefs, and an abundance of wrecks make this archipelago a diver's paradise.

AZURE WATERS
The seas around the Scillies have a clarity more typical of southern France than southern England. This makes for excellent diving.

SPARSELY POPULATED

The Isles of Scilly (or the Scillies, as they are universally known) consist of 300 granite islands 28 miles (45 km) off the southwestern tip of England. Only six of the islands are inhabited, and even these could hardly be described as crowded. The islands are low-lying and very beautiful, with white sandy beaches fringing green islands.

The islands' position in one of the world's busiest shipping channels means that wrecks are abundant. The most recent wreck here was the cargo ship *Cita*, which foundered on March 26, 1997, leading to a bonanza of free goods for enterprising locals. This vessel is now an excellent dive, lying in 115 ft (35 m) of clear water. By contrast, the wreck of the *Eagle*, a British warship that sank in 1707, offers a glimpse into naval history, with cannons still visible on the seabed.

ESSENTIAL INFORMATION

WATER TEMPERATURE 46–64°F (8–18°C)

ECOSYSTEM Temperate

WHEN TO VISIT Diving is possible year-round, although in winter there can be violent storms.

MUST-SEES Inner Gilstone Rock; *Cita* and *Eagle* wrecks; gray seals.

OTHER ATTRACTIONS

As well as numerous wrecks to explore, divers also have access to splendidly busy reefs, and may even encounter larger species such as seals, as well as seasonal visitors, including basking sharks.

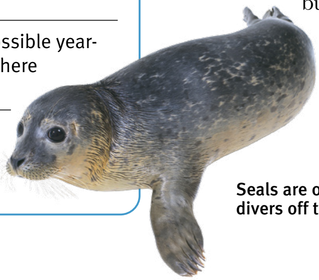

Seals are often encountered by divers off the Scillies

The Skelligs

COUNTY CORK, IRELAND

REPUBLIC OF IRELAND
Dublin
The Skelligs
Cork

Two rocks that jut out of the Atlantic at the eastern end of the Gulf Stream, the Skelligs are a beacon for wildlife in the rich seas off southwestern Ireland. This has created a world-class dive site in one of the finest dive regions to be found anywhere in Europe.

TWIN ISLANDS

The Skelligs sit 11 miles (18 km) off the southwestern coast of Ireland, and rise almost 700 ft (200 m) above the wild Atlantic waters below. They consist of two huge rocks—Skellig Michael and Little Skellig—around 1 mile (1.5 km) apart. The craggy faces and reefs of the rocks are washed by the waters of the North Atlantic Current, a flow of water that carries the tail end of the Gulf Stream. This brings warm waters from the mid-Atlantic, accompanied by occasional large species, such as dolphins. This current is the secret of the area's richness.

ESSENTIAL INFORMATION

WATER TEMPERATURE 48–72°F (9–22°C)

ECOSYSTEM Temperate

WHEN TO VISIT The best time of year to dive on the Skelligs is between April and September.

MUST-SEES The Pinnacle dive site; seal colony; gannets on Little Skellig.

HAVEN FOR MARINE LIFE

The rock faces of the Skelligs are coated in colorful encrusting organisms, such as anemones, sponges, and dead man's fingers. Every nook and cranny has an occupant, with huge edible crabs, lobsters, conger eels, and rocklings all jostling for position. In the kelp forests, wrasse and dogfish lurk, while mackerel and bass cruise the open waters. Larger residents include a healthy gray seal colony, as well as visiting basking sharks.

SHELTERING STONE
The isolated, rugged setting of Skellig Michael may look inhospitable, but beneath the water's surface, the area teems with vibrant marine life.

Lundy

BRISTOL CHANNEL, SOUTHWESTERN ENGLAND

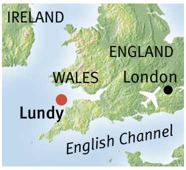

Lundy Island sits amid the powerful tidal stream of the Bristol Channel. Washed constantly by waters rich in nutrients, it is host to wonderfully busy reefs and the occasional larger visitor arriving from the great waters of the Atlantic, to the west.

PIONEERING RESERVE

Lying 10 miles (16 km) off the coast of Devon, Lundy is a craggy granite outcrop approximately 3 miles (5 km) in length. The richness of its reefs and the abundance of marine life saw Lundy established as the first Marine Nature Reserve in the United Kingdom.

There are two distinct faces to Lundy—the more exposed western shore, which faces the swells and storms of the Atlantic, and the more benign eastern shore. The topography of the granite creates wonderfully dramatic cliffs and reefs, with the regular tidal movement in the channel bathing these rock faces in nutrient-rich water.

GENTLE GIANT
The gaping mouth of the basking shark nets only the tiniest plankton for its food. The species is often encountered off Lundy in summer.

REEF POPULATIONS

Divers will encounter an array of British reef life around the island—lobsters, several species of wrasse, huge shoals of bib and pollack, bass, conger eel, and octopus. More unusual Lundy residents include anglerfish and red-banded fish—the latter being unique to the island. The island is also frequently visited by basking sharks and dolphins during summer months.

The wreck of the British warship *Montague* can be found on Shutter Reef at a depth of 15–50 ft (5–15 m). A 14,000-ton battleship, it ran aground in fog in 1906 while on exercises in the Bristol Channel. Intensive efforts to refloat it failed. Although the wreckage is scattered and broken, there remain piles of armor plate amid the kelp. The *Montague* was armed when it sank, so beware of live shells.

ESSENTIAL INFORMATION

WATER TEMPERATURE 46–72°F (8–22°C)

ECOSYSTEM Temperate

WHEN TO VISIT Visit Lundy from April to September for the best diving.

MUST-SEES Basking sharks and dolphins; the wreck of the *Montague*; Gannets Rock; Knoll Pins.

Norwegian Fjords

NORTHWESTERN COAST, NORWAY

Norwegian Fjords
Bergen
Oslo
NORWAY
SWEDEN

Cold water, steep walls, and strong currents may not be everyone's ideal diving experience, but for some marine life, these are ideal conditions. Diving the Norwegian fjords is not for the faint-hearted, but the rewards are well worth the effort.

WILD COASTLINE

Norway's convoluted coast stretches for around 14,000 miles (22,000 km) in a series of jagged inlets, bays, and fjords, and is dotted with over 50,000 islands. Norway also has, in Sognefjord, the longest (140 miles/225 km) and deepest (4,430 ft/1,350 m) fjord on Earth.

There are many great dive sites along this gloriously wild coastline: one of the finest regions—and one that is gaining an international reputation for the quality of its diving—is Ryfylke, to the southwest. Diving the fjords requires previous experience and competence in cold-water diving, along with the relevant cold-water gear. The topography of the fjords creates steep walls colonized by kelp, dead men's fingers, sponges, and some rare deep water corals. Piked dogfish, skate, catfish, and numerous crustaceans are particularly abundant in Ryfylke. Seasonal highlights include migratory visits from salmon and their attendant predators, such as orcas.

ESSENTIAL INFORMATION

WATER TEMPERATURE 36–68°F (2–20°C)

ECOSYSTEM Temperate

WHEN TO VISIT The fjords can be dived year-round, but winter months can present cold, challenging conditions.

MUST-SEES Pulpit Rock; Kjerag Wall; Lysefjord; visiting orcas.

DRAMATIC HORIZON
The plunging walls of the fjords, both above and below the water's surface, set the tone for this rugged and beautiful destination.

Iceland

NORTH ATLANTIC

Akureyri
ICELAND
Höfn
Reykjavík

Born from the depths of the sea itself, Iceland is still a youngster in geological terms. The country's volcanic past provides some unique dive sites, including the opportunity to explore the marine life around a hydrothermal vent.

VOLCANIC BIRTH

Iceland was formed by volcanic action on the Mid-Atlantic Ridge, and has a surface area of just over 38,600 sq miles (100,000 sq km). The island is still experiencing an extraordinary level of volcanic activity, with 30 eruptions recorded in the last 200 years. Iceland's coastline is classically volcanic, bristling with craggy inlets, dramatic cliff and rock formations, and fissured reefs.

ESSENTIAL INFORMATION

WATER TEMPERATURE 34–46°F (1–8°C)

ECOSYSTEM Temperate

WHEN TO VISIT Diving in Iceland is at its best in summer, between June and September.

MUST-SEES Strytan hydrothermal vent, Silfra in Thingvellir National Park.

There are two diving experiences in Iceland that are not found anywhere else in the world. The first is the legendary Strytan hydrothermal vent, which rises from a depth of 230 ft (70 m) to within 50 ft (15 m) of the water's surface. Divers can see—and indeed feel—the hot water rising from the vent while enjoying the diverse invertebrate and fish life drawn into this oasis of warmth in an icy sea. The second is Silfra, a water-filled ravine in Thingvellir National Park. This is another dive that gives a marvelous insight into Iceland's fiery origins. Part of a huge rift in Earth's crust between the Eurasian and American continental plates, Silfra is filled with glacial meltwater that has been filtered through lava for decades, giving famously crystal-clear visibility.

ICELANDIC COASTLINE
The dramatic geological eruptions that gave birth to Iceland are evident in its craggy and forbidding cliffs and wild shores.

White Sea, Russia

NORTHEASTERN RUSSIA

Barents Sea
White Sea
Arkhangelsk
FINLAND
RUSSIA

The White Sea is one of the most beautiful of the seas of northern Russia, and the only place in Russia where you can dive under the ice. The marine environment here is relatively untouched by human activity and the diving is superb.

FROZEN WORLD

Diving beneath ice is exciting and adventurous, and the visibility—anything from 50 to 165 ft (15–50 m)—can be stunning. The water is not affected by waves (due to the protection of the ice), and low light conditions mean that algal growth is minimal. The water in the White Sea, which lies across the Arctic Circle, is frozen from November until April, but the best time to go is in mid-March. Any earlier and the temperatures are very cold, far below zero; any later and the ice has retreated and you cannot get to the best dive sites. The Gulf Stream warms some of the waters farther north, preventing them from freezing over.

DENIZEN OF THE ICY DEEP
A diver comes face to face with a gorgonocephalus—or, more dramatically, “gorgon’s head”—a bizarre form of basket star common in the White Sea.

RARE ORGANISMS

The White Sea offers plenty of life to see and photograph: for example, there are some 80 species of fish, around 30 species of nudibranchs, and five species of seals. Shoals of tiny comb jellyfish also form an extraordinary sight, the light glinting off them in a million tiny rainbows. It is possible to encounter beluga whales, and occasionally orcas and Greenland sharks, which venture into these waters.

In addition to some great diving, there are possibilities of ice snorkeling in fast-flowing rivers and among the icebergs and, looking to the skies, witnessing the stunning beauty of the atmospheric phenomenon known as the aurora borealis, or northern lights. Also within excursion distance of dive sites on the east coast is Star City, the Russian cosmonaut training facility; you can even arrange to dive at the hydrospace laboratory at the MIR Space Center, used to prepare cosmonauts for the weightless conditions of space.

ESSENTIAL INFORMATION

WATER TEMPERATURE 31°F (–0.5°C) at the surface to 27°F (–3°C) deeper down

ECOSYSTEM Polar

WHEN TO VISIT Diving is best in mid-March (*see also, above*).

MUST-SEES Beluga whales; seals; aurora borealis; cosmonaut facility.

Africa and the Middle East

A continent synonymous with adventure and exploration, Africa offers spectacular shark encounters and many miles of unexplored reef. The Middle East boasts the Red Sea, perhaps the most famous of all places to dive, with a highly developed dive industry to match.

AFRICAN HIGHLIGHTS

South Africa is renowned throughout the world for big animal encounters—in particular with the great white sharks of the Cape—and for the epic shark dives to be had off Natal. The latter is also the scene of one of the greatest spectacles in the marine world—the legendary Sardine Run, where gigantic sardine shoals move slowly up the coast, harried by dolphins, whales, sharks, seals, and gannets.

DESERT MEETS SEA
At many sites in Africa, the warm blue sea laps against a harsh, arid landscape of bleached rock.

Lying to the northeast of the African coast are the magnificently wild beaches and reefs of Mozambique and the islands off Kenya and Tanzania. Farther out to sea lie Mauritius, the Seychelles, and the Comoros, all offering diving of real quality on reefs influenced by the abundant marine life of the Indo-Pacific region to the east.

VIBRANT ECOSYSTEM
Africa's varied reefs sport a profusion of colorful soft corals, which play host to a stunning array of marine life.

RED SEA RICHES

The outstanding gem of the Middle East is the Red Sea, perhaps the most heavily dived body of water in the world. From the northern resorts of Eilat, Ras Mohammed, Sharm El Sheik, and Dahab, through to the wilder southern sites like the Brothers and the coast of Sudan, this region provides superb reef diving, as well as varied wrecks, such as the famous *Thistlegorm.*

Despite its popularity, the Red Sea still has unexplored regions, and new dive sites continue to be opened, so it will always have much to attract divers.

CRYSTAL-CLEAR WATERS
One of the saltiest bodies of water in the world, the Red Sea is also renowned for its superb visibility.

Gansbaai

WESTERN CAPE, SOUTH AFRICA

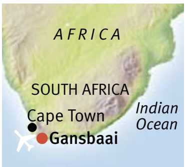

Gansbaai offers a rare chance to see the ultimate marine predator. With an island packed with breeding seals just offshore, Gansbaai is one of the finest locations on Earth to get up close and personal with the great white shark.

AWESOME ANIMAL
Great white sharks cruise the cold waters near Dyer Island, on the hunt for their preferred prey, fur seals.

SHARK CENTER

Gansbaai initially appears to offer little as a dive destination, with a small harbor on a rocky shore that is frequently battered by large swells. It is only when the diver stands on the wharf and squints into the middle distance that the reason for this site's global fame emerges. A low island several miles offshore is the home of 60–90,000 fur seals, and proves irresistible to the great whites that cruise this region of South Africa. Great whites are one of the ocean's most formidable predators, growing up to 20 ft (6 m) in length. The great white's notoriety as a man-eater, however, is a misrepresentation of this fascinating creature, about which surprisingly little is known.

ESSENTIAL INFORMATION

WATER TEMPERATURE 52–63°F (11–17°C)

ECOSYSTEM Temperate

WHEN TO VISIT Year-round shark activity; numbers peak May to September.

MUST-SEES Cage diving with great white sharks at Dyer Island, plus seal dives in a few recommended areas.

MAIN ATTRACTION

The presence of the sharks has spawned a thriving industry, with thrill-seekers coming from all over the world to dive with the great whites. Only cage diving is permitted with the sharks, but it is possible to dive within the seal colonies around the island in certain regions that are considered safe by local dive operators. Farther afield, there are beautiful cold-water reefs covered in swaying fields of kelp, but there is really only one reason why divers visit Gansbaai—that most magnificent of marine predators, the great white shark.

Aliwal Shoal

KWAZULU-NATAL, SOUTH AFRICA

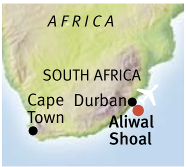

Aliwal Shoal is one of the greatest adventure dive sites on Earth. The excitement begins on the beach, with perhaps the most dramatic surf launch in the diving world, and ends with outstanding shark encounters above a thriving reef system.

CROSSROADS IN THE SEA

Aliwal Shoal is a mountain of sandstone that juts into the underwater highway of the Agulhas Current. The site is renowned for shark encounters, in particular the famous ragged-tooth sharks that lurk within the caves and caverns of the reef. There are also Zambezi sharks in great numbers at the right time of year, as well as more occasional encounters with tiger sharks, and, thrillingly, there are a few rare sightings of great whites. However, Aliwal Shoal has much more to offer than shark encounters, since the mixture of both warm water reef fish and temperate species creates one of the most diverse reef systems in all of southern Africa.

VISITING SPECIES

Added to the array of life are transient visitors, such as migrating humpback and right whales, as well as turtles and huge brindle bass. The dive site is also directly in the path of the Sardine Run, one of the most impressive marine migrations anywhere in the world's oceans. Sardine shoals travel more than 1,000 miles (1,600 km) in this great annual exodus.

ESSENTIAL INFORMATION

WATER TEMPERATURE 66–79°F (19–26°C)

ECOSYSTEM Subtropical

WHEN TO VISIT Ragged-tooth sharks in residence from July to October; Zambezi sharks can be seen October–April.

MUST-SEES Ragged-tooth sharks; Sardine Run; migrating whales.

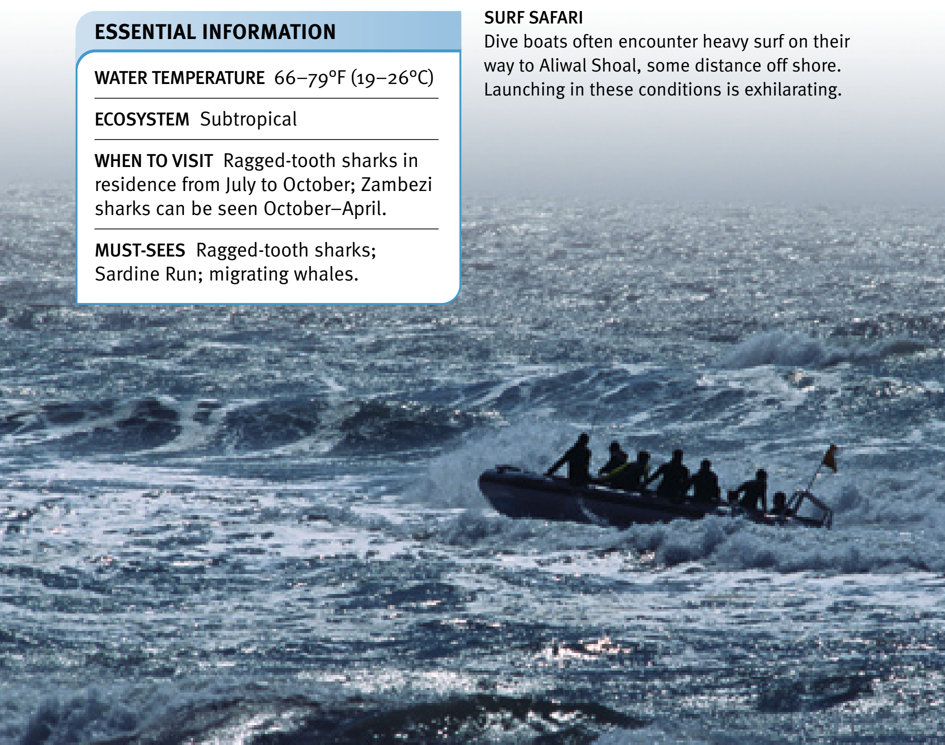

SURF SAFARI
Dive boats often encounter heavy surf on their way to Aliwal Shoal, some distance off shore. Launching in these conditions is exhilarating.

Sodwana Bay

KWAZULU-NATAL, SOUTH AFRICA

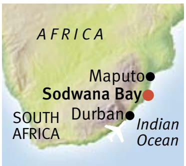

The only coral reef dive site in South Africa, Sodwana is set in the heart of a national park in northern Natal. The reefs host more than 1,000 species of fish, as well as impressive transients, such as whale sharks, billfish, manta rays, and whales.

REMOTE REEFS

Sodwana lies in a beautifully remote region of South Africa, nestling right up against the border with Mozambique to the north. Simply getting to the site is something of an adventure, requiring a journey by four-wheel drive vehicle through wild bush on bumpy dirt tracks. Although the dive industry in Sodwana is well established, it is also closely monitored. Only three dive centers are allowed to operate in the region, to limit the number of divers on the reef and so protect the delicate coral reef ecosystem.

ESSENTIAL INFORMATION

WATER TEMPERATURE 70–84°F (21–29°C)

ECOSYSTEM Tropical reef

WHEN TO VISIT Year-round diving, but visit October to February for a chance of sighting whale sharks at Sodwana.

MUST-SEES Seven Mile Reef; Nine Mile Reef; surf launch; whale sharks, mantas, and other visiting species.

Turtles of several species visit Sodwana

DIVERSE ATTRACTIONS

The four established reefs off Sodwana are named after their distance from the launch site, and are called Two, Five, Seven, and Nine Mile Reef. Each offers something slightly different, ranging from the most delicate of hard corals through to encounters with large ocean-going species including whale sharks. Although the journey to Sodwana is a considerable undertaking, the reward is nothing less than the finest coral reef dives on the continent.

SURF AND SAND
Despite the fact that Sodwana Bay is fairly remote, the site can be busy. Dive boats often launch with an audience of waiting divers.

Protea Banks

KWAZULU-NATAL, SOUTH AFRICA

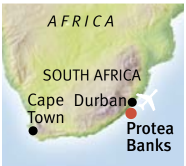

Even within the singularly robust South African diving community, Protea Banks is a dive site famed for thrills. A reef cut with deep gullies, overhangs, and caverns, and notable for its shark encounters, this is not a site for the faint-hearted.

ON THE PROWL
The tiger shark is a species quite regularly sighted around Protea Banks, often with remora fish in tow.

CHALLENGING SITE

Although Protea Banks does not boast the dramatic topography of nearby Aliwal Shoal (*see p.259*), it lacks none of the excitement. Indeed, there are many in the local diving community who consider Protea Banks more extreme than Aliwal—the reasons being the depth, the currents, and the terrific shark encounters. The top of the reef is, on average, deeper than Aliwal Shoal, with several dives in excess of 100 ft (30 m). The currents that sweep over the flat reef crest are also stronger, and several of the dives are exhilarating drifts over a considerable distance of undersea terrain.

SHARK SPOTTING

The final reason Protea Banks is so revered among the dive community in South Africa is its shark encounters. Not only can large numbers of ragged-tooth sharks be found here during their mating season, but there is a strong possibility of encountering tiger sharks or hammerheads. For several months of the year, divers can also be confronted in mid-water by large numbers of the Zambezi sharks. These magnificent predators follow game fish that arrive during the summer months.

ESSENTIAL INFORMATION

WATER TEMPERATURE 66–79°F (19–26°C)

ECOSYSTEM Subtropical

WHEN TO VISIT Ragged-tooth sharks can be seen from July to October; Zambezi sharks from October to April.

MUST-SEES Ragged-tooth sharks; Zambezi sharks; migrating whales.

Mozambique

SOUTHEAST AFRICA

AFRICA
Mozambique
Maputo
Indian Ocean
Cape Town

Wracked by decades of civil war, Mozambique was once a forgotten destination in world diving. The end of the conflict in 1992 prompted the gradual return of divers, and revealed the riches that lie in the blue waters off one of Africa's wildest coastlines.

NURTURING SEAS

The fast-moving waters of the Agulhas Current, which passes through the gap between the east coast of Mozambique and the island of Madagascar, are the source of much of Mozambique's great marine diversity. The reefs in this region are teeming with life, and although the entire coastline has dive sites of merit, there are certain regions that stand out, including Inhaca Island, Ponta do Ouro, the Bazaruto Archipelago, and Cabo San Sebastian.

FAR-FLUNG LOCATION

Mozambique is by no means an easy dive destination to access. Many of the sites to the south of the country require a sturdy four-wheel drive vehicle to cope with the dirt and sand roads, and the isolated operators and resorts are tucked in behind long stretches of deserted beach. However, Mozambique's dive sites offer ample reward. The southern region of the country has become well

ESSENTIAL INFORMATION

WATER TEMPERATURE 70–82°F (21–28°C)

ECOSYSTEM Subtropical

WHEN TO VISIT Year-round, but conditions are best from September to March.

MUST-SEES Potato bass; ragged-tooth sharks; bull sharks; humpback whales; whale sharks; manta rays.

known for encounters with huge potato bass and ragged-tooth sharks, whereas farther north, encounters with bull and tiger sharks are a distinct possibility. Humpback whales move through the area in the summer months, and the occasional dugong is sighted in the shallow estuaries along the coast.

The Bazaruto Archipelago on Mozambique's central coast is an excellent place to see whale sharks, while nearby Vilnaculos is an excellent site to encounter manta rays and dolphins.

BRIGHT FUTURE

Mozambique still has a primitive tourist infrastructure, particularly in the northern regions of the country. Malaria is also a significant factor, with most of the country affected at certain times of the year. Nonetheless, Mozambique is making tremendous progress as a tourist destination, with the country's diving rapidly establishing itself as truly world-class. Diving here can present challenges for the intrepid traveler, but the isolation and low tourist traffic can be seen as a real advantage in one of Africa's most rapidly developing dive destinations.

WAVE-SWEPT WILDERNESS

Mozambique has over 1,500 miles (2,400 km) of coastline, much of it sparsely inhabited. The southern stretches of coast are dominated by sandbars, estuaries, and swamps, while the wilder north is characterized by long beaches and cliffs. The entire coastline is tremendously dramatic and wild, and even the capital Maputo has the famous Inhaca Island—a site famed in the diving world for its large animal encounters—directly offshore. Dolphins, whale sharks, and manta rays all frequent this site.

Surf breaks on a stretch of coast

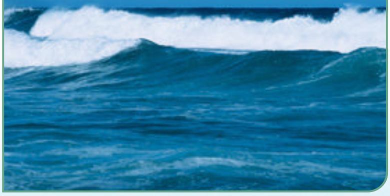

LARGER THAN LIFE
Enormous potato bass are often found patrolling the bustling reefs of Mozambique. This imposing species can be extremely inquisitive, and individuals often approach divers.

Seychelles

WESTERN INDIAN OCEAN

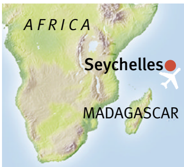

The clear waters of the Seychelles teem with abundant marine life, and offer those prepared to travel a bit farther a range of gentle reef dives on the inner islands. The outer islands have big walls, wild drifts, and large animal encounters.

A PATTERN OF ISLANDS

The Seychelles archipelago numbers 115 islands in total, although these are spread over a vast area of ocean hundreds of miles off the east coast of Africa. The Seychelles is divided into the inner islands and the outer islands, with the two groups over 680 miles (1,100 km) apart. The inner islands—including Mahe, the main island in the Seychelles—sit on a huge natural platform rising from the deep ocean around them. As such, diving here tends to be within the 30–100-ft (10–30-m) range, and takes place on shallow protected reef systems. Farther afield are the outer islands, including the Aldabras, a group of islands known as the “Galápagos of the Indian Ocean” due to their isolation and diversity of marine life. The diving here is more intense and suitable only for experienced divers. The reef cuts and walls create opportunities for tremendous drift-dives, in which divers are carried along by underwater currents.

ESSENTIAL INFORMATION

WATER TEMPERATURE 81–84°F (27–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT May to October.

MUST-SEES Large pelagic species in the Aldabra Group; shallow reefs and marine life around the inner islands, including Mahe.

Powder blue tang

TROPICAL VISTA
Balmy tropical beaches are typical of the Seychelles. Though remote, the islands offer tremendous attractions for the diver.

Mauritius

WESTERN INDIAN OCEAN

AFRICA
Indian Ocean
MADAGASCAR
Mauritius

The islands of Mauritius in the Indian Ocean are a popular vacation destination, but are not well known for diving. This is something of an oversight, since the islands' stark volcanic reefs harbor a rich tropical ecosystem and offer dives of real quality.

PLACE IN THE SUN
The shallow coastal waters and white sands of Mauritius make it a leading vacation destination, but it also offers great diving experiences.

VOLCANIC ORIGINS

Mauritius consists of two islands, and numerous smaller islets, 500 miles (800 km) east of Madagascar. Although relatively small, the islands have nearly 250 miles (400 km) of undulating coastline due to the craggy nature of the volcanic rock that makes up the island group. The seas are colonized by over 200 species of coral, creating reefs inhabited by around 450 fish species—a rich tropical ecosystem.

ESSENTIAL INFORMATION

WATER TEMPERATURE 70–82°F (21–28°C)

ECOSYSTEM Tropical

WHEN TO VISIT Diving takes place in Mauritius year-round.

MUST-SEES The Rempart Serpent and Cathedral dive sites; the unspoiled reefs around Rodrigues Island.

TOP LOCATIONS

Most diving in Mauritius takes place on the western coast of the main island; the eastern end of the island is exposed to the wind and waves of the open ocean. The Cathedral cavern with its resident porcupine fish is a must-see, and the Rempart Serpent site boasts a huge number of moray eels (it is reputed that 32 species have been seen here!). The region north of the main island, however, is becoming known as a good spot for encounters with some of the larger pelagic species, with regular sightings of tuna, wahoo, and marlin. The remote island of Rodrigues to the east offers a glimpse of some of the region's most pristine reefs—making an excursion well worth the effort.

The Brothers

NORTHERN RED SEA, EGYPT

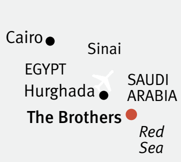

Spectacular coral gardens have grown undisturbed around these two remote pinnacles of rock for thousands of years. Only accessible by liveaboard (*see p.237*), the Brothers dive site offers steep drop-offs and some terrific shark action.

REMOTE ROCK

Big Brother, the larger and more northerly of the two islands, is 1,300 ft (400 m) long and 295 ft (90 m) wide, and dominated by a lighthouse.

ISOLATED OUTCROPS

The Brothers consist of two large rocks, named Little Brother and Big Brother, around 1 mile (1.6 km) apart and 36 miles (58 km) offshore in the northern Red Sea. Rising from deep water, they provide dramatically steep walls as a foundation for bustling reefs, and the upwellings of nutrient-rich waters and swirling currents draw in fish species both large and small.

SHARK SPOTTING

In addition to the diversity and color of the reefs themselves, there are some tremendous shark encounters off the walls of the Brothers. Common species in the area include white-tip and gray reef sharks, although silver-tips, scalloped hammerheads, and oceanic white-tips are also frequently seen. If you are lucky, you may even encounter a thresher shark. Its scythe-shaped tail fin makes this species an unmistakable, if very infrequent, sight.

ESSENTIAL INFORMATION

WATER TEMPERATURE 70–84°F (21–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT Diving takes place on the Brothers all year.

MUST-SEES Oceanic white-tip sharks; scalloped hammerheads; wreck of the *Numidia*; (occasionally) thresher sharks.

WRECKED ON THE REEF

As an additional attraction to the site, two wrecks—the *Aida* and the *Numidia*—cling to the precipitous sides of Big Brother. The *Numidia*, a former cargo vessel, lies at a very steep angle on the northern tip of the rock, with which it collided in 1901. Its bow can be found in just 26 ft (8 m) of water, but its stern lies in the depths, at 262 ft (80 m).

Another victim of Big Brother was the Italian supply ship *Aida*, which sank in 1957 while carrying troops. Almost as impressive as *Numidia* and surprisingly intact, *Aida* is now heavily colonized.

Dahab

NORTHERN RED SEA, EGYPT

EGYPT
Cairo
Sinai
Dahab
Sharm el Sheikh
Red Sea

The coastal town of Dahab in southern Sinai is ringed by mountains and desert. The inhospitable landscape is in striking contrast to the splendid reefs and coral formations to be found just offshore—among which is the infamous Blue Hole.

DESERT HAMLET

Dahab lies on the Gulf of Aqaba, in the northeast of the Red Sea, only 50 miles (80 km) north of the busy dive resort of Sharm El Sheik. Clinging to the shore in a harsh, arid landscape, the town has a faintly frontier feel to it. There are several dive centers in Dahab, offering training to all levels of ability.

DIVING IN DAHAB

Within a short drive of Dahab are many excellent sites accessible from the shore, including the Canyon and the Bells. Both of these sites are dramatic examples of the twisted coral formations that seem to characterize this stretch of coastline, with deep gullies, impressive overhangs, arches, and swim-throughs. The legendary Blue Hole site is actually the top of a massive shaft that sinks 985 ft (300 m) into the reef. The lure of diving through a long passage to the outer reef at a depth of 165 ft (50 m) has caused many divers to perish in the Blue Hole, and the conventional wisdom for diving this deep arch is that it should only be attempted using technical dive techniques and equipment (*see pp.194–95*).

ESSENTIAL INFORMATION

WATER TEMPERATURE 70–82°F (21–28°C)

ECOSYSTEM Tropical

WHEN TO VISIT Diving takes place in Dahab all year, with water temperatures high throughout.

MUST-SEES The Blue Hole; The Canyon and The Bells; Gabr El Bint (by boat only).

SHELL GAME
A diver follows a turtle—at a respectful distance—as it conducts a slow tour of one of Dahab's colorful reefs.

Wreck of the *Thistlegorm*

NORTHERN RED SEA, EGYPT

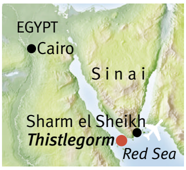

Packed full of artifacts and lying in the clear waters of the Red Sea, the *Thistlegorm* is perhaps the most famous of all shipwrecks for divers. Its fame comes at a price, however, with great numbers of divers clambering through the structure every day.

THE CLASSIC WRECK

The *Thistlegorm* first came to the attention of the global diving community in the 1950s, when it was discovered and filmed by Jacques Cousteau. However, its location remained a secret until 1992, when it was rediscovered by recreational divers, and it has since become one of the most popular diving wrecks in the Red Sea.

The *Thistlegorm* was a World War II armed merchant vessel, 413 ft (126 m) in length, carrying supplies to the British 8th Army in North Africa. Packed with machinery and ammunition, it was spotted in the northern Red Sea by a German bomber on October 6, 1941. Two bombs entered its Number Four hold, causing a huge explosion that cut it in half; it sank immediately with the loss of nine souls.

TIME CAPSULE

A diver examines military hardware inside the wreck of the *Thistlegorm*

For many years after the *Thistlegorm* was sunk, British vessels would lower their flag as they passed its wreck, but today it has a happier role, having become a focus for diving in the Red Sea. The vessel's real treasure is its outstanding cargo of military equipment, which ranges from Lee Enfield Mark 3 rifles, wellington boots, Bedford trucks, and two Rolls Royce armored cars, to—perhaps most famously of all—racks of BSA motorcycles. On the deck stands one locomotive; another lies on the sea bed beside the wreck.

DIVE EXPERIENCE

In many ways, the *Thistlegorm* is the perfect wreck dive. Its bow lies just 50 ft (15 m) below the surface in clear waters, and no part of the vessel is below 100 ft (30 m); however, currents may be strong, and descent is usually made down mooring lines tied by the guide. Once inside the wreck there is a vast, serene, yet poignant underwater museum to explore, and the sheer size of the vessel means that thorough exploration requires several visits.

The *Thistlegorm* holds a unique place in divers' affections, though its popularity has become a drawback. It is only a five-

ESSENTIAL INFORMATION

WATER TEMPERATURE 68–84°F (20–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT Year-round. The wreck is best dived from a liveaboard.

MUST-SEES The captain's bath; BSA motorcycles; locomotive on deck; Bedford trucks; armored cars.

hour cruise away from the Red Sea's busy diving center of Sharm el Sheikh, and in peak season it is not uncommon for more than 20 diving vessels to be moored over it at one time. The best time to explore the *Thistlegorm* is at dawn, when there is a chance to see this grand wreck with few others around, although this means finding a skipper who will travel through the night.

Although not as densely colonized as one might expect, the *Thistlegorm* wreck still has abundant marine life. This region has approximately 1,000 fish species, 17 percent of which are endemic to the Red Sea. Of particular note on the *Thistlegorm* are large numbers of lionfish, which lie in wait ready to ambush shoals of unsuspecting glassy sweepers, while overhead jack and barracuda sweep the superstructure for prey.

WEAPON OF WAR
A diver hovers above the distinctive deck gun on *Thistlegorm's* stern. There are strict penalties for wreck-stripping, so never be tempted to take a souvenir.

Asia

The tropical regions of this vast continent have always been highly praised by divers for the staggering richness of their marine life and their glorious reefs, but adventurous new diving opportunities are also beginning to open up to the north—even in Siberia.

REEFS UNDER THREAT

The geographical span of Asia, the largest continent, is reflected in the vast range of diving available. For many divers and marine biologists, the highlight is the famed "Triangle of Diversity"—a region in the Indo-Pacific that boasts the world's greatest diversity of coral and fish species.

The many thousands of islands of the Philippines, Indonesia, Papua New Guinea, and Malaysia are host to some legendary dive sites, with evocative names such as Sipadan, Lembeh, Tubbutaha, Komodo, and Kimbe. Sadly, while these are all dream destinations for any diver, destructive fishing techniques in the Indo-Pacific region, such as dynamite and cyanide fishing, are damaging some of these delicate ecosystems. In the Philippines, only 4 percent of all reefs remain unaffected by human activity, with up to 30 percent reported to be impacted beyond recovery. Beautiful pristine reefs can still be found in the Indo-Pacific, but when diving on these threatened reef systems, be mindful of your responsibility to minimize any impact your presence could have.

BEACH BASE
A dive base on a sun-kissed tropical beach can provide the ideal "chill-out" spot after an exhilarating dive.

RESPONSIBLE REEF DIVING
With Asia's reefs at risk, it is vital that divers follow the rules of responsible reef diving to conserve them for future generations.

BEYOND THE TROPICS

The classic perception of Asian diving tends to be the archetypal tropical island chain fringed by glorious technicolor reefs. Dive sites throughout the region, including the Maldives, the Andaman Islands, and the Similan Islands, are all justifiably famed for their extraordinary diversity and beauty, both above and below the surface.

However, although much of the diving in Asia tends to be concentrated in the tropics, there are some excellent dive sites to be explored farther north. One of the most notable is Lake Baikal in Siberia. This is the world's deepest lake, and the largest reservoir of fresh water on the planet's surface. Because of its size and varying depths, Lake Baikal contains a huge variety of ecosystems.

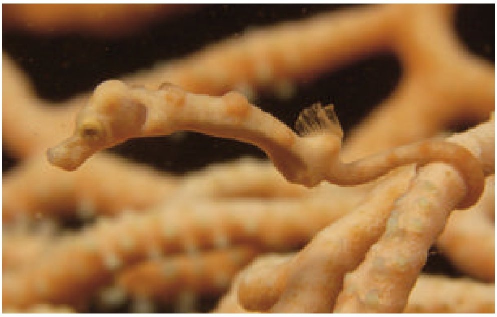

PYGMY SEA HORSE
A rare species found only on *Muricella* coral in Southeast Asia and Australia, the pygmy seahorse wraps its tail around coral branches and lies in wait for tiny crustaceans and larvae.

The Maldives

INDIAN OCEAN

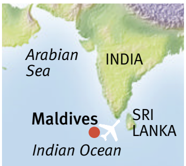

Coming to prominence as a dive destination in the 1970s, the Maldives swiftly became established as a dream diving location. The extent of the archipelago, and its range of dive sites, ensures that it remains an essential stop-off at least once in a diver's lifetime.

RICH ARCHIPELAGO

The islands of the Maldives are the crests of an undersea mountain range that plunges over 1¼ miles (2 km) into the deep waters of the Indian Ocean. The archipelago extends 475 miles (764 km) from north to south, and is 80 miles (129 km) across. This vast collection of islands is loosely divided into 26 major atolls, and although the islands are spread over 38,000 sq miles (98,000 sq km) of ocean, their land mass covers only 115 sq miles (300 sq km).

The Maldives offer a huge range of diving activities for those lucky enough to visit. Such is the abundance and delicacy of the remarkable reef systems throughout the Maldives that there are no less than 25 nature reserves and protected areas within the island group. The marine life around the islands is colorful and diverse, and includes over 700 fish species.

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–84°F (24–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT Diving takes place year-round in the Maldives.

MUST-SEES Gray reef sharks; hammerhead sharks; manta rays; whale sharks; turtles.

TOUGH CHALLENGES

Tourism and diving are vital to the economy of the Maldives, but both industries have had to recover from major natural disasters in the late 20th and early 21st centuries. El Niño—a periodic warming event in the Pacific Ocean—had a severe impact in 1997–98, when the hard coral on many reefs was bleached and killed by the rise in ocean temperatures. The islands were also hit by the 2004 tsunami, which destroyed much of the tourist infrastructure.

Fishing and diving boats ready for business on a Maldives quayside

Diving in the Maldives tends to be divided into drift dives on the outer reef, and dive sites within the atolls. The latter contain thilas, a local word meaning a coral platform rising almost to the water's surface, and kandus, which are smaller isolated coral outcrops and coral heads. The best drift diving is where cuts in the coral walls leading to the open sea channel tidal water through narrow gaps.

Most dives in the Maldives seem to include at least one encounter with white-tip reef sharks, and there are several sites where gray and black-tip reef sharks, as well as hammerheads, may be seen. Eagle rays and manta rays, too, are regularly encountered at the right time of year at certain sites, with whale sharks also returning to the same sites to make the most of seasonal feeding grounds. There are also five different species of turtles in the Maldives, the most common of which are the green and hawksbill turtles.

COLORFUL CORAL

A diver examines soft corals on one of the region's flourishing reef systems. The Maldives offer an incredible variety of quality dives.

Similan Islands

ANDAMAN SEA, INDIAN OCEAN

INDIA
THAILAND
Bangkok
Similan Islands
MALAYSIA
Indian Ocean

A tiny string of granite islands sitting off the coast of Thailand, the Similans are best explored by liveaboard dive boat, but with some of the best beaches in Thailand, they have a great deal to offer both above and below the water.

STONE AND SEA

The Similan Islands consist of nine large granite outcrops sitting 40 miles (64 km) off the coast of Thailand. They rise from the depths of the Andaman Sea, and run for 15 miles (24 km) north to south. Although the islands do have individual names, they are commonly known simply by numbers, starting in the south with Number 1 (Koh Huyong), and continuing to Number 9 (Koh Bangu) in the north.

They are characterized by steep granite undersea cliffs on their western shores, with more gentle slopes to the east, where, sheltered from the summer monsoon winds, the coral growth is particularly lush. Over 1,000 fish species swarm over reefs colonized by a dazzling array of soft corals. Brightly colored species, such as coral grouper, sweetlips, snapper, and blue triggerfish are present in great numbers and, from October to May, whale sharks and manta rays are frequently encountered.

The thriving reef systems around the islands were once threatened by overfishing as well as "dynamite fishing"—the use of explosives to kill and collect large numbers of fish—but the islands were declared a national park in 1982, and commercial fishing is now banned. The sweeping coral fields of Christmas Point are particularly spectacular, although they were slightly damaged in the tsunami of 2004. Other notable attractions include a dramatic maze of huge boulder formations at Elephant's Head Rock. The Similans remain well worth every hour of sea crossing required to visit them.

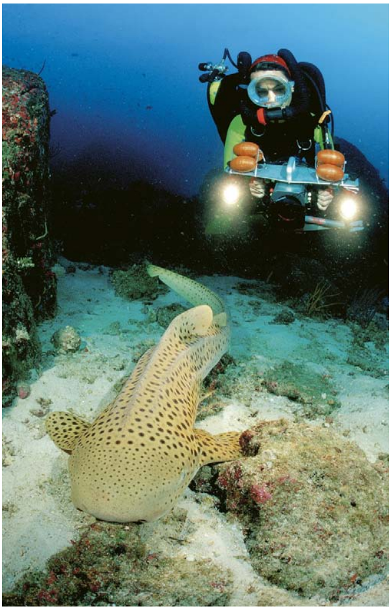

SHARK SPOTTING
A diver observes a leopard shark, hovering above this beautifully patterned predator, which frequents the waters of the Similans.

ESSENTIAL INFORMATION

WATER TEMPERATURE 79–84°F (26–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT Visit the Similan Islands between October and May.

MUST-SEES Whale sharks and manta rays; boulder formations at Elephant's Head Rock; Christmas Point.

Andaman Islands

ANDAMAN SEA, INDIAN OCEAN

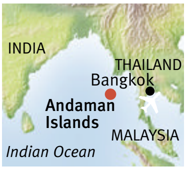

One of the last great diving wildernesses, the Andamans are a truly unspoiled string of volcanic islands in the Bay of Bengal. Cloaked in rain forest and home to some of the last untouched tribes on Earth, they offer genuine diving exploration.

ISOLATED IDYLL

The Andaman Islands are sprinkled over 435 miles (700 km) of ocean, and define the western edge of the Andaman Sea. Although the archipelago consists of 572 islands, only 36 are inhabited. Over 80 percent of the land mass is covered in rain forest, home to tribes that have had very little contact with the outside world—in some cases, none at all.

The islanders do not run any commercial fishing operations to speak of, and therefore the steep walls and rich reefs of the region are home to a range of large marine animals in healthy populations not encountered anywhere else in the Andaman Sea.

ESSENTIAL INFORMATION

WATER TEMPERATURE 82–88°F (28–31°C)

ECOSYSTEM Tropical

WHEN TO VISIT The best time to visit is between November and May.

MUST-SEES Barren Island; Passage Island; large manta ray aggregations; steep lava walls; dogtooth tuna.

Diving takes place from November to May, and can either be from liveaboard dive vessels, or through a small number of operators based in the islands. Notable sites include the outlying islands of Passage Island, Narcondam, and Barren Island, which have barely been dived at all. Expect encounters with a range of shark species, huge dogtooth tuna, and feeding manta rays.

DIVER'S PARADISE
With no large-scale fishing fleet, the Andaman Islands have well-preserved reefs stocked with a fascinating diversity of marine life.

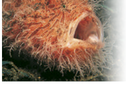

North Sulawesi

CELEBES SEA, INDONESIA

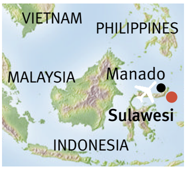

Indonesia has been termed the “Cradle of Diversity” by marine biologists, and with good reason. A recent survey counted more than 3,000 fish species here. Crowning it all is North Sulawesi, and the jewel in that crown is the Lembeh Straits.

CONSERVATION IN ACTION

North Sulawesi is formed by the long, thin peninsula that juts out from Sulawesi Island into the Celebes Sea. The marine species diversity of this region is unparalleled, a fact recognized by the locals, who have set up a number of reserves, and by the divers who flock here. The careful approach to the marine environment here by local government means that 90 percent of the coral cover is alive and in good condition, a remarkable statistic considering the pressure the remainder of the Indonesian archipelago experiences from destructive fishing methods.

ESSENTIAL INFORMATION

WATER TEMPERATURE 74–82°F (23–28°C)

ECOSYSTEM Tropical

WHEN TO VISIT Year-round; optimum period between May and November.

MUST-SEES Panguilingan; Lejuan II; Mandolin; muck diving for ghost pipefish; pygmy sea horses; mimic octopus.

Camera in watertight plastic housing

NATURAL TREASURES

The Lembeh Straits is the channel between the island of Lembeh and the northern tip of Sulawesi. It is quite a shallow body of water—ranging in depth from 16 ft (5 m) to 98 ft (30 m)—so divers need to be aware of the large vessels that pass regularly overhead, as well as the periodic strong currents.

But while at first glance this may seem merely an unprepossessing shipping channel, closer inspection reveals a world of tiny riches.

MACRO LIFE IN SULAWESI

The magic that draws divers to this area is the extraordinary range of tiny but beautiful reef animals that can be found in the shallow coastal seas off North Sulawesi. The sheer diversity of “macro life” (small marine creatures) here, and the sometimes bizarre forms it takes, has led to the area being dubbed “God’s wastebasket.” Outlandish species like the hairy frogfish make amazing photographic subjects.

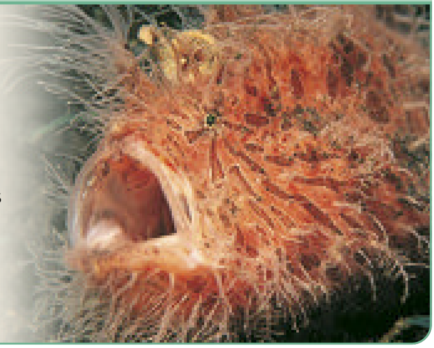

A hairy frogfish gapes in the current

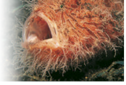

The Straits have a big name for small things—this is the finest site in the world for encountering the most exquisite, delicate, and unusual inhabitants of tropical seas. Diving in the Lembeh Straits goes by the unpromising name of "muck diving"—so called because of the dark sand that covers the seabed and the necessity to examine it closely for target species during the course of a dive. But the effort involved in diving here is well rewarded. Mimic octopus, pygmy sea horses, ghost pipefish, leaf fish, frogfish, and blue-ringed octopus can all be found here.

CONSERVATION AREA

The Bunaken Marine Park is one of the driving forces for conservation on North Sulawesi. It surrounds five islands, on which are a few small fishing villages, and covers over 290 sq miles (750 sq km). The park offers fantastic muck diving and also some tremendous walls, the most notable of which are Lekuan II and Mandolin on Bunaken, and Panguilingan in Mahado Tua.

VOYAGE OF DISCOVERY
An indigenous fisherman setting out in a traditional canoe. Many divers eschew the clear surface waters for the pleasures of muck diving.

Papua New Guinea

WESTERN PACIFIC

Papua
New
Guinea
Port Moresby
Cairns
AUSTRALIA

A location that still epitomizes diving exploration, Papua New Guinea offers a hugely exciting experience for the intrepid dive tourist. A vast region with numerous undived reefs and islands, Papua New Guinea is a real adventure.

OUTER LIMITS

For many years, Papua New Guinea drew only the very boldest pioneers of adventure diving and extreme tourism. Made up of 600 islands, the reputation of this region as a wild and untamed part of the world is well founded. Papua New Guinea has 4 million inhabitants, divided into 1,000 tribes that speak 700 different languages. Spread over the 177,600 sq miles (460,000 sq km) of the island group, many of these indigenous tribes knew nothing of the modern world until very recently, with one tribe only making contact with the outside world in the late 20th century. Doubtless, others—as yet unknown—exist in the rugged Highlands, with much of the main island of New Guinea consisting of mountainous rainforest with a maximum elevation of 14,760 ft (4,500 m). While most visits to Papua New Guinea are trouble-free, travel after dark is not advised, and in towns such as the capital, Port Moresby, the risk of robbery and violent crime must be taken seriously. Trips are best arranged with well-established diving operations who can provide guides and transportation to the coastal diving regions, where the risks are significantly lower than in poor urban areas.

PRISTINE REEFS

The shallower regions of Papua New Guinea's coastline are the site of some fantastically complex reef formations, with perhaps the most famous being Kimbe Bay on the western coast of New Britain. Another dive site that has gained a reputation for providing particularly dramatic reefs and tremendous diversity is Eastern Fields, in the northern limits of the Coral Sea. This remote site can only be reached by ocean-going vessels and remains relatively unexplored.

Vibrant coral reef in Papua New Guinea

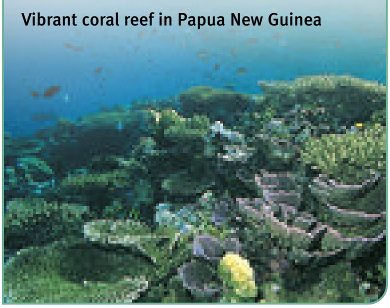

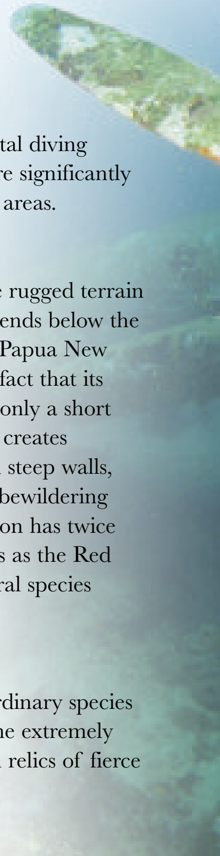

RICH SEAS

The good news is that the rugged terrain of the land frequently extends below the surface of the water, and Papua New Guinea is notable for the fact that its seas reach abyssal depths only a short distance from shore. This creates tremendous drop-offs and steep walls, which are inhabited by a bewildering array of species. The region has twice the number of fish species as the Red Sea, and ten times the coral species of the Caribbean.

SUNKEN HISTORY

In addition to this extraordinary species diversity, the area has some extremely fine wrecks, most of them relics of fierce

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–82°F (24–28°C)

ECOSYSTEM Tropical

WHEN TO VISIT Diving takes place all year in Papua New Guinea.

MUST SEES Submerged atoll at Eastern Fields; marine life in Kimbe Bay; World War II aircraft wrecks in Milne Bay.

fighting in the region during World War II. Rabaul, Kavieng, Loboata, and Madang all provide excellent wreck diving, with aircraft debris littering the seabed in certain areas, like Milne Bay.

Diving throughout the archipelago is becoming more accessible due to the presence of a number of established diving operators in the more renowned dive sites, and the increasing range of liveaboard operators now pushing into some truly remote regions. Nonetheless, there are still underwater areas of this extraordinary island group that remain completely unexplored. In a world that seems to become smaller and smaller with each passing day, it is an exciting thought that such places still exist for us to discover. By taking a responsible approach to venturing off the beaten track, divers can help preserve this region, and by doing so ensure a steady stream of new adventures for many years to come.

GHOSTLY REMAINS
Wreckage of aircraft lost during World War II is a notable sight in the waters off Papua New Guinea. The area was the scene of battles between Allied and Japanese forces.

The Philippines

SOUTHEAST ASIA

Manila
Philippines
VIETNAM
MALAYSIA
BORNEO

The Philippines undoubtedly offer a world of diving possibilities. Although destructive fishing techniques have taken their toll on many reefs, local legislation and a realization of the benefits dive tourism can bring to the local economy offer a chance of recovery.

RESTORING LOST REEFS

Consisting of 7,107 islands spread over 116,000 sq miles (300,000 sq km), the Philippines comprise one of the most bountiful regions on Earth in terms of species diversity, with at least 2,000 fish species recorded throughout the archipelago. However, in many areas, complex and delicate reef systems have been almost completely destroyed by dynamite and cyanide fishing. A recent survey noted that "excellent" reefs—those classified as having more than 75 percent healthy coral cover—are now present in only 4.3 percent of the total reef area surveyed. This is an issue that the Philippines authorities are, thankfully, attempting to address, and certain regions now have energetic conservation programs in place.

VACATION HOTSPOT
Indigenous watercraft pulled up at a palm-fringed beach in the region. The Philippines appeal to both divers and non-divers alike.

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–81°F (24–27°C)

ECOSYSTEM Tropical

WHEN TO VISIT Best diving conditions are between April and November.

MUST-SEES Tubbutaha Reef; Jessie Beazley reef; Coron Bay; Apo Island; Puerto Galera.

LIVEABOARD EXCURSIONS

The remote Tubbutaha and Jessie Beazley Reefs in the Sulu Sea remain relatively undamaged, and provide some world-class diving. All diving at Tubbutaha is from liveaboards, although many of the shallow reef crests provide excellent snorkelling opportunities. The steep walls of Tubbutaha are swept by strong currents and have crystal-clear visibility. Of particular quality is North End, a dive that combines spur and groove reef formations with some classic walls. Jessie Beazley Reef is essentially a massive pillar of rock rising from deep water, providing a beacon for groupers, manta rays, barracuda, and eagle rays.

Quality diving can also be found at the Visayas, a small cluster of islands accessed via Cebu City. Malapuscua

HISTORIC FINDS
A diver inspects a wreck off the Philippines. Palawan Island is the departure point for Coron Bay, with its extensive World War II wrecks.

has become particularly famous for providing one of the most unusual of all shark dives, at a site where thresher sharks congregate to be "groomed" by cleaner fish (*see pp.225–26*). Apo Island is a fine example of a progressive conservation project, where local people have become involved in policing the reefs around the island.

MUCK AND WRECK DIVING

Farther north at Puerto Galera, readily accessible from Manila, is some of the finest diving in the Philippines, with some outstanding "muck diving" (*see p.279*). Palawan also offers some wilder diving, with Coron Bay containing some particularly fine Japanese wrecks sunk during an attack by US aircraft on September 24, 1944.

Barracuda Lake offers a genuinely different diving experience, requiring a short jungle hike that is rewarded by an exploration of a system in which freshwater sits on top of seawater. A large barracuda in the lake is said to guide divers through his territory.

MARINE LIFE

The diversity of this archipelago is truly remarkable, with 1,499 crustacean species, 462 types of mollusks, and 22 different types of beach vegetation. Of the 500 known coral species worldwide, 488 are found here. This biological diversity was recognized in 1993, when the Tubbataha Reef Marine Park was made a World Heritage site.

The mandarin fish, one of over 2,000 fish species recorded in the Philippines

Australasia

The vast coastline of Australia offers superb coral reefs as well as spine-tingling cold-water encounters with giant marine predators. New Zealand, to the southeast, is rapidly emerging as a dive destination of real quality, with some of the world's best temperate dive sites.

BARRIER REEF AND MORE

For the uninitiated, the notion of diving in Australasia is limited to exploring the Great Barrier Reef, which lies on the northeastern coast of Australia. One of the most impressive natural features on the planet, it is a site of remarkable species diversity.

While the fame of the reef is fully justified, there are also many other rewarding dive sites in Australia and New Zealand. East of the tourist hotspots of the Great Barrier Reef lies the Coral Sea, equally beautiful and with fabulous marine life to be seen. To the south of Australia, the cold waters of Dangerous Reef and the Neptune Islands offer thrilling encounters with great white sharks. Farther south still is Tasmania, home to lush kelp forests. On the sparsely populated western coast of Australia is Ningaloo Reef, one of the best locations for encountering the mighty whale shark.

VOLCANIC REEF
The temperate reefs of New Zealand's Poor Knights Islands teem with colorful marine life.

NEW ZEALAND

The jewel in New Zealand's diving crown is the Poor Knights Islands, which for many (including the late Jacques Cousteau) ranks as the finest temperate-water dive site in the world. The unique topography of the site means that this is a breathtaking place to dive, with vast shoals of fish swarming over craggy volcanic reefs and through atmospheric cavern systems. Protected by both traditional Maori law and modern legislation, the islands are a magnet for divers the world over.

New Zealand boasts some impressive wrecks, including the world-famous *Rainbow Warrior* and the dramatic *Mikhail Lermentov*. It is also rapidly developing a reputation for providing big animal encounters with numerous whale species, as well as with iconic sharks such as the mako and the bronze whaler. To top it all off, New Zealand also offers stunning scenery, crystal clear rivers and springs, and magnificent fjords.

SOUTHERN FJORDS
The fauna of New Zealand's fjords includes red and black corals, penguins, fur seals, and bottlenose dolphins.

LIFE ON THE REEF
The Great Barrier Reef hosts a vast array of flora and fauna, including more than 1,500 fish species and 4,000 different types of mollusks.

Great Barrier Reef

QUEENSLAND, AUSTRALIA

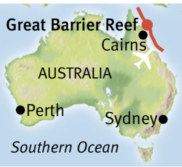

One of the true natural wonders of the world, the Great Barrier Reef is the longest reef of its kind on Earth. Home to a spectacular diversity of animal and plant life, recent legislation has granted even greater protection for this abundant natural habitat.

VAST ECOSYSTEM

The Great Barrier Reef is so extensive that its precise size has never been recorded, although it is estimated that it stretches for 1,400 miles (2,250 km), and covers around 135,000 sq miles (350,000 sq km)—about the same surface area as Ireland. It is made up of 2,000 individual reefs, and around 71 coral islands are spread along its length.

In broad terms, the Great Barrier Reef can be said to be divided into two regions—the inner reef, known as the Great Shallow, and the outer reef, consisting of deeper reef systems and sloping walls leading to the open sea beyond. The reef itself is made up of over 400 species of coral, and supports around 1,500 species of fish. Approximately 4,000 species of mollusks inhabit the flats and nooks and crannies of the reef, while 23 species of marine mammals may be encountered along the vast length of this complex ecosystem.

ESSENTIAL INFORMATION

WATER TEMPERATURE 77–84°F (24–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT The Great Barrier Reef can be dived throughout the year.

MUST-SEES Ribbon Reefs (particularly Ribbon Reef Number 10); Cod Hole; Pixie's Pinnacle.

GLOBAL IMPORTANCE

Certain areas of the Great Barrier Reef have enjoyed protected status since 1973, when it was established as a marine park. This was enhanced in 1981 when the reef was recognized as a World Heritage Site. More recent legislation has seen 71 percent of the total area of the reef covered by some sort of

WORLD OF COLOR
The range of marine life inhabiting the Great Barrier Reef is staggering in its variety and splendor, but divers must be careful not to touch delicate corals.

DEEPER INTO THE REEF

Certain parts of the Great Barrier Reef have been noticeably degraded or damaged by the high volume of dive traffic. This is no great surprise, given the delicacy of the reef's ecosystem and the inexperience of the many divers exploring it. Although there are still some fine dives to be had in these busy areas—especially for those new to the sport—many dive operators have moved further out into the reef system, offering packages on liveaboard boats that allow more experienced divers to explore remote and less spoilt sites.

Eye-stripe surgeonfish

protective status—an optimistic sign for the future of a reef that is larger than any other coral system, but just as vulnerable to misuse and exploitation.

DIVING IN THE REGION

There is a mass market for diving the Great Barrier Reef, operating out of Cairns and Townsville in day boats containing very large numbers of divers. There have been a number of well-publicized incidents where divers have been missed when boats have departed back to shore, and although this is now very well regulated, it is still worth, on the larger vessels at least, ensuring that the correct procedures are in place before diving. The other main diving activity on the reef takes place from liveaboards that venture farther afield to more remote sites.

Divers on the reef are assured of reasonable hard coral and colorful fish life almost anywhere along its length, but there are also a number of notable individual sites. Ribbon Reef Number 10 is particularly rich and diverse, even by the Great Barrier Reef's high standards, and is the location for dive sites such as Cod Hole—famous for encounters with giant potato cod—and Pixie's Pinnacle, renowned for its diversity of large animal encounters.

ARTIFICIAL ISLAND
Some sites on the Great Barrier Reef have a permanently anchored observation platform, which visitors can travel to by tour boat.

The Coral Sea

WESTERN PACIFIC OCEAN

Coral Sea
Cairns
AUSTRALIA
Perth
Sydney
Southern Ocean

In the early 1970s, the first liveaboard operators began to explore the waters to the east of the Great Barrier Reef. Here they found the Coral Sea—a region of beautiful pristine hard coral formations and precipitous drop-offs.

REMOTE LOCATION

The Coral Sea begins about 100 miles (160 km) east of Townsville, in the territory of Queensland, Australia. It stretches for another 100 miles (160 km), covering an area of 38,000 sq miles (98,500 sq km). The dive sites consist of a series of undersea mountaintops that rise from a great depth, creating atolls, ridges, and reefs. The visibility in the region is some of the best in the diving world.

TEEMING WITH LIFE

The crossing to the Coral Sea through the rough waters of the inner reef can be demanding; however, divers with the budget and time to make the journey are rarely disappointed. The area is renowned for its big animal encounters, with tiger sharks and hammerheads mixing with more abundant species, such as gray reef sharks and white-tips. Manta and eagle rays are frequently seen off the steeper reef walls, while sea snakes can be seen on Marion Reef.

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–81°F (24–27°C)

ECOSYSTEM Tropical

WHEN TO VISIT Conditions are suitable for diving year-round, but a lengthy boat trip is required to reach the area.

MUST-SEES Cod Wall; Watnabe Bommie; Marion Reef; manta rays; sea snakes.

REEF DWELLERS
Bumphead wrasse cruise over the sandy seafloor of the Coral Sea. These large fish graze in "herds" over the area's reefs.

Wreck of the *Yongala*

QUEENSLAND, AUSTRALIA

Cairns
Wreck of Yongala
AUSTRALIA
Perth
Sydney
Southern Ocean

The *Yongala* is a biological wonder—one of the most heavily colonized shipwrecks on Earth. Sitting amid miles of featureless plains of sandy seabed next to the Great Barrier Reef, over time it too has become a teeming coral reef.

FATAL STORM

Lost in a typhoon on March 24, 1911, the *Yongala* was a 357-ft (109-m) steamer carrying 121 passengers north along the Queensland coast. All hands were lost as it foundered, and now it lies in 100 ft (30 m) of water, listing to starboard but still resolutely facing its original northerly direction of travel. Its isolated position, and the fact that it lay undiscovered for over 40 years after sinking, means that it is very heavily colonized with marine life.

MARINE SANCTUARY

The entire food chain of the coral reef is packed into the length of the *Yongala's* short hull. Schools of jack and trevally hunt clouds of bait fish on its port hull, while in its superstructure lurk turtles, morays, coral trout, and huge numbers of sea snakes. Potato cod skulk beneath the bow of the wreck, and very large bull rays nestle in the sand beside the wreck, ever watchful for the tiger and bull sharks that circle out in the blue, just beyond sight.

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–82°F (24–28°C)

ECOSYSTEM Tropical reef

WHEN TO VISIT There is year-round diving on the *Yongala*, but conditions can be rough from May to August.

MUST-SEES Sea snakes; potato cod by the wreck's bow; bull rays; bull sharks.

OCEAN WILDERNESS
Divers can encounter potato cod patrolling the bow section of the *Yongala*. The wreck is a magnet for marine life, attracting a range of species from sharks to turtles.

PROTECTED STATUS

Despite the popularity of the *Yongala* wreck among divers, exploration of the wreck's interior is strictly forbidden by Australian authorities under the country's Historic Shipwrecks Act. It may be, however, that the resident population of grouper are protection enough. In 2002 it was reported that a 6½-ft- (2-m-) long giant grouper had seized a diver's head before releasing him, shaken, but without serious injury, shortly afterward.

Tasmania

TASMAN SEA, AUSTRALIA

Cairns
AUSTRALIA
Perth
Sydney
Tasmania
Hobart

Tasmania offers a diving experience far removed from that of the usual coral reefs and warm waters of Australia. Nonetheless, this tiny island off the southern coast holds a host of diving possibilities for those prepared to travel a little farther afield.

UNSUNG ISLAND

Tasmania is a small island dwarfed by the immensity of mainland Australia to the north. It measures only 186 miles (300 km) from north to south, and is frequently overlooked by divers visiting Australia, distracted by the magnitude of the Great Barrier Reef and the big animals of Ningaloo Reef.

Tasmanian divers are delighted at this relative anonymity, since the island offers some of the finest kelp forests on Earth and some gloriously colorful and diverse cold-water dives. Currents rich in oxygen and nutrients sweep up from the icy waters to the south, giving life to wonderfully dense kelp forests and heavily colonized rock walls.

ESSENTIAL INFORMATION

WATER TEMPERATURE 54–64°F (12–18°C)

ECOSYSTEM Temperate

WHEN TO VISIT The best time to visit Tasmania for diving is between June and September.

MUST-SEES Red handfish; kelp forests; Eaglehawk Neck.

SMALL WONDERS

Of particular note is the abundance of smaller animals, such as red handfish, although the island also plays host to numerous penguins, Australian fur seals, a number of different types of sharks, and a variety of visiting whale and dolphin species. This rich diversity means that divers visiting Tasmania come to appreciate the true breadth of Australia's natural heritage.

PERFECT CONCEALMENT
Sea dragons, such as this leafy variety, can be found in the kelp forests off Tasmania and South Australia. Its shape mimics the plant fronds.

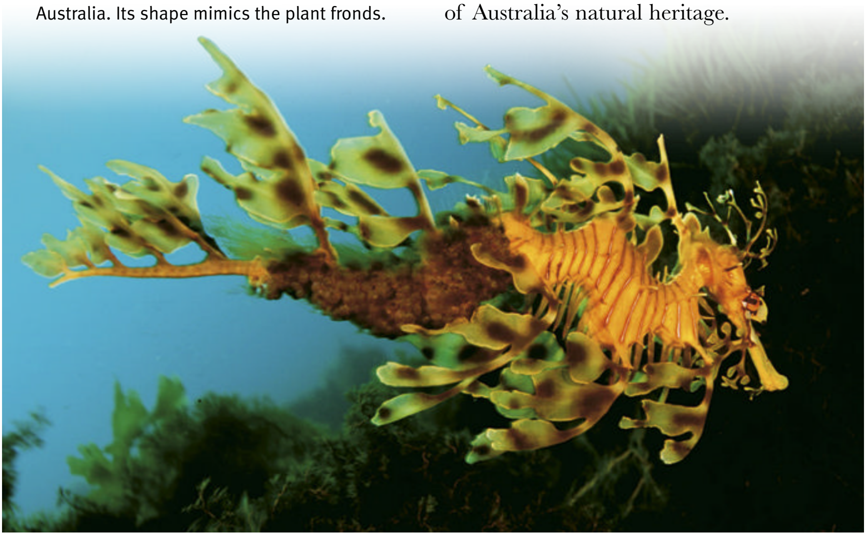

Ningaloo Reef

WESTERN AUSTRALIA

Indian Ocean
AUSTRALIA
Ningaloo Reef
Perth
Sydney
Southern Ocean

A beautiful coral reef stretching for 175 miles (280 km) off the coast of Western Australia, Ningaloo has much to offer the diver. Its fame, however, lies in the abundance of a single animal—Ningaloo is the whale shark capital of the world.

MAGNET FOR DIVERS

Ningaloo Reef is approximately 750 miles (1,200 km) north of Perth, Australia, and lies along a beautifully remote stretch of coastline.

The diving on the reef is centered around the small town of Exmouth—the population of which doubles during whale shark season, as divers flock there from around the world. Ningaloo Reef is splendidly rich in its own right—with 500 fish species and 200 types of coral. Such is the scale of diversity here and the integrity of the ecosystem, that the region was declared a national park in 1987, and now over 1,500 sq miles (3,885 sq km) enjoys protected status.

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–84°F (24–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT Year-round diving, but whale sharks present March to June.

MUST-SEES Whale sharks; manta rays; migrating whales; shore dive at Point Murat Navy Pier.

WEALTH OF WILDLIFE

Although the whale sharks are less abundant from July through to February, they are replaced by other great attractions—between May and October, huge numbers of manta rays can be seen, and from July to November, migrating whales pass along the coast. In addition, a survey in 1994 counted more than 1,000 dugongs along this stretch of coast—one of the most significant populations on Earth.

CLOSE ENCOUNTER
Touching whale sharks may distress them, and only those with special training, such as this marine biologist, should do so.

Poor Knights Islands

NORTH ISLAND, NEW ZEALAND

Poor Knights Islands
Auckland
NEW ZEALAND
Wellington

A mystical set of islands, sacred in Maori legend, the Poor Knights lie 14 miles (23 km) off the coast of North Island in New Zealand. Honeycombed with caves and caverns, they are host to some truly remarkable gatherings of marine animals.

FORBIDDEN LAND

The Poor Knights Islands consist of stark volcanic rock pockmarked with overhangs, ridges, caves, and arches. The dramatic topography of the islands evokes their bloody history, as they were the site of a terrible Maori massacre in 1820, causing them to be declared *tapu* (forbidden) in Maori law. No one has set foot on the islands since, and this, combined with marine park status since 1977, has created a wonderfully pristine environment both above and below the waves, which has drawn divers to this site from all over the world. Boats run to the islands from nearby Whangerai.

REMOTE LOCATION
The islands are only accessible by boat, so visits involve a short voyage aboard one of the many dive charter boats based on the mainland.

ESSENTIAL INFORMATION

WATER TEMPERATURE 59–75°F (15–24°C)

ECOSYSTEM Temperate reef

WHEN TO VISIT Year-round diving, but rainy season runs from July to October.

MUST-SEES Gatherings of southern stingrays; Blue Mao Mao Arch; bronze whaler sharks on deep ridges.

WEALTH OF MARINE LIFE

There are approximately 120 fish species around the Poor Knights, but it is the sheer numbers of fish present that are the main attraction. Vast shoals of species such as blue mao mao inhabit the overhangs and caverns around the islands, while the stark reefs are coated in an extraordinary diversity of

encrusting organisms, with brightly colored sponges jostling for position with sea squirts and bryozoans. Thick carpets of kelp provide perfect hunting grounds for brightly colored wrasse species, scorpion fish, and several different types of moray eels. In March, the dark waters under the Southern Arch host huge mating congregations of short-tailed southern stingrays. As well as this spectacular annual gathering, there is also a resident population of bronze whaler sharks, and regular transient visitors including orcas, dolphins, and several whale species.

Jacques Cousteau was so impressed when he visited the Poor Knights in the 1960s that he described the islands as the finest temperate-water dive destination on Earth, and one of the top ten overall. The range of dive sites and sheer abundance of marine life provide a compelling argument that this applies as much today as it did then. This is an essential dive site if visiting New Zealand.

ODD ONE OUT
A lone wrasse swims at the center of a swirling shoal of blue mao mao. The Poor Knights Islands are famous for their abundance of fish life.

EXPLORING CAVERNS

The marine life of the Poor Knights Islands teems amid dramatic underwater caves and arches, including the largest sea cave in the world—Rikoriko Cave. Visiting divers should carry flashlights, including a reliable backup, and ensure that basic cavern safety rules are followed (*see pp.186–87*).

A diver explores a sea cave

Wreck of the *Rainbow Warrior*

CAVALLI ISLANDS, NEW ZEALAND

Rainbow Warrior
Auckland
NEW ZEALAND
Wellington

Perhaps no other shipwreck on Earth has the poignancy of the *Rainbow Warrior*. Sunk in an act of international espionage that shocked the world, it now lies off the Cavalli Islands, in one of the most beautiful regions of New Zealand.

TRAGIC FATE

The conservation vessel *Rainbow Warrior* was moored in Auckland harbor on the night of July 10, 1985. Shortly before midnight, two explosions ripped through it, killing the vessel's photographer, Fernando Pereiro. This senseless act of terrorism was traced to the French Secret Service, leading to international condemnation and lengthy jail sentences for the two agents who planted the explosives.

PEACEFUL RELOCATION

The *Rainbow Warrior* wreck remained in Auckland for many months, before being refloated and towed farther north to the Cavalli Islands, a beautiful marine sanctuary granted sacred status in Maori law. Here it was sunk as an artificial reef, and has become a popular dive site.

The *Rainbow Warrior* is a small vessel, and so can easily be covered in a single dive. It sits almost upright on the bottom, and its superstructure rises to within 50 ft (15 m) of the surface, with the seabed below at 85 ft (26 m). Its bow section is still largely intact, creating the illusion that it is sailing across the white sands of the seafloor beneath its hull.

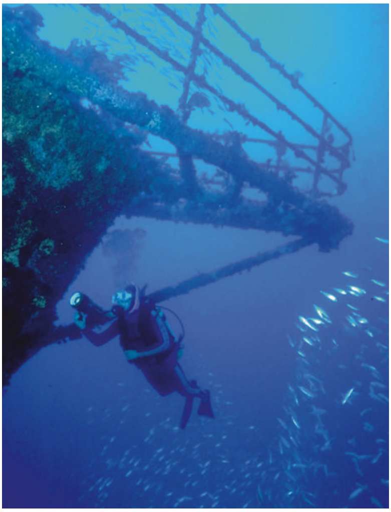

FINAL RESTING PLACE
A diver surveys the wreckage of the former Greenpeace vessel *Rainbow Warrior*, which was badly damaged by a violent act of sabotage.

ESSENTIAL INFORMATION

WATER TEMPERATURE 57–75°F (14–24°C)

ECOSYSTEM Temperate

WHEN TO VISIT Year-round, but September to April is best.

MUST-SEES Scorpionfish; the prow; the Rainbow Warrior Memorial; the ship's propeller and stern.

REFUGE FOR MARINE LIFE

The *Rainbow Warrior's* hull is covered in a multicolored carpet of jewel anemones, and the vibrancy of color on display here is remarkable. Its superstructure is now a home for scorpionfish, blue cod, and shoals of golden snapper, plus kingfish, and the distinctively shaped John Dory. It seems fitting that this former conservation vessel now serves as a home for so many fascinating marine species.

Wreck of the *Mikhail Lermentov*

MARLBOROUGH SOUND, NEW ZEALAND

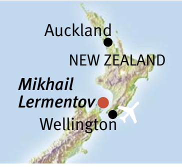

The gloomy passageways and silty spaces of the *Mikhail Lermentov* are mute witnesses to one of the great mysteries of modern sea travel. Sunk under the strangest of circumstances, it now lies in the cold waters of Marlborough Sound, New Zealand.

ENIGMATIC WRECK

The 20,000-ton Russian cruise liner *Mikhail Lermentov* was driven hard onto rocks by an experienced local pilot on February 16, 1986. It was a clear, calm day, and theories abound as to the factors behind the sinking—bribes, alcohol, and espionage have all been blamed at one time or another. The mystery remains to this day, but the wreck represents a tremendous diving experience, although perhaps not one for the faint-hearted.

DIVING THE *LERMENTOV*

It lies on its starboard side in 121 ft (37 m) of water, although the sheer size of the vessel means that its port side can be reached at only 66 ft (20 m). Dives on its external superstructure are recommended, with highlights including the magnificent bow, the bridge, and the stern. For the more adventurous there is the famous ballroom, although this dive should only be undertaken by those with appropriate training and equipment (*see pp.166–71*). Marine life on the wreck includes scorpionfish, conger eels, and blue cod; occasionally pods of orcas hunt in the Sound itself.

ESSENTIAL INFORMATION

WATER TEMPERATURE 57–72°F (14–22°C)

ECOSYSTEM Temperate

WHEN TO VISIT Year-round, but September to April is best.

MUST-SEES Bridge; swimming pool; stern; ballroom, provided you have the right gear and level of experience.

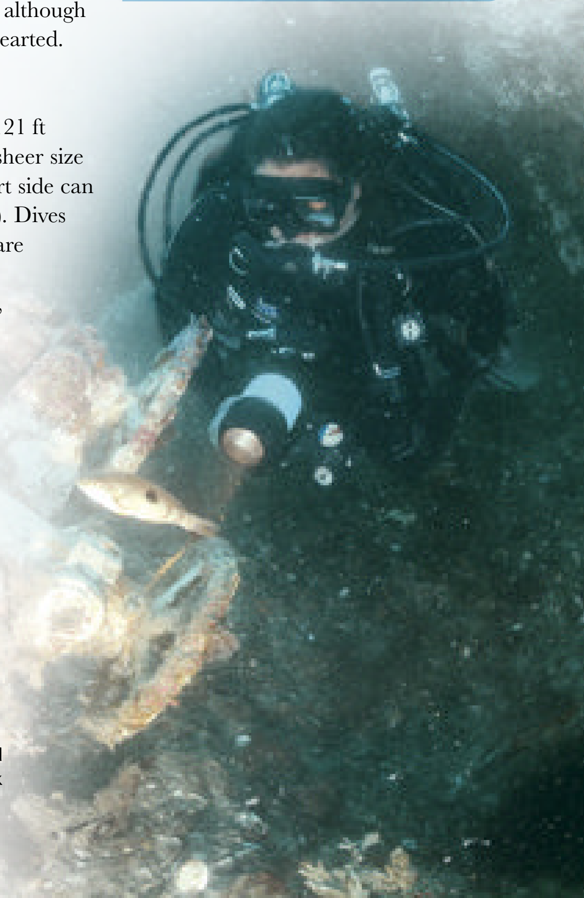

SERIOUS ADVENTURE
A diver explores in conditions of reduced visibility, often encountered on the *Mikhail Lermentov*. The wreck lies in deep coastal waters.

Oceania

The beauty of the estimated 20,000 islands that make up Oceania is unmatched anywhere else on Earth. The region holds tremendous riches, particularly for those confident enough to venture off the beaten track to visit the more far-flung islands and reefs.

WRECK DIVING IN PARADISE

Oceania is the collective name for the islands of the Pacific, generally considered synonymous with the South Sea Islands of Melanesia, Micronesia, and Polynesia. The region is typified by tranquil beaches with sugar-white sand and swaying palms, and, below water, powerful and intense diving experiences. Some of the world's most diverse reefs are in Micronesia, and the fierce fighting that raged through much of the region during World War II has created some impressive wreck sites.

Truk Lagoon in Micronesia, with its array of sunken Japanese ships, was viewed by many as the finest of all wreck diving locations until the establishment of a dive operation on Bikini Atoll in the Marshall Islands in 1996. From Bikini Atoll, divers can explore the remains of the fleet of ships sunk during the testing of hydrogen bombs by the US from 1946.

Palau has many magnificent dive sites, including Blue Corner, a coral ridge that juts out into the ocean and

VOLCANIC CREATION
Many of the world's atolls are located in the Pacific Ocean. These are created as a coral reef builds up around a subsiding volcanic island.

UNDERWATER WONDERS
The coral gardens of Micronesia, among the healthiest on the planet, are home to butterflyfish, wrasse, angelfish, and puffers.

plunges more than 1,000 ft (300 m), and the Ngemelis Wall, one of the best wall dives in the world. There is also Jellyfish Lake, an inland saltwater lake swarming with jellyfish that have lost their ability to sting. The nearby island of Yap boasts one of the finest sites for manta ray encounters. Yap is one of the more popular destinations in Micronesia, yet 80 percent of the reefs remain undived. For those who have the time to explore the remote islands of Oceania, they offer a truly original dive experience.

TRADITIONAL ISLAND LIFE
Unspoiled islands offer the possibility of fascinating cultural encounters and more fulfilling diving experiences.

Palau

MICRONESIA, SOUTH PACIFIC

PHILIPPINES
Palau
INDONESIA
AUSTRALIA

Divers constantly argue about what they consider to be the ultimate dive destination, but Palau will always be a strong contender. Consisting of 307 scattered islands, Palau has it all—wrecks, caves, and some of the busiest coral reefs on Earth.

A WORLD OF VARIETY

One of the great draws of Palau is the vast range of diving available. Like other island groups in Micronesia, it has colorful reefs cruised by large predators and impressive oceanic species. What gives Palau the edge over other sites in the region is that the geology and recent history of the islands have combined to create a diver's paradise. Caves and caverns honeycomb walls and reefs throughout the island chain; and fierce battles between American and Japanese forces during World War II mean that aircraft and shipwrecks abound.

Palau also has a number of rare natural phenomena, the most famous being Jellyfish Lake. This is a large volcanic crater filled with brackish water containing the pulsating forms of over 20 million mastigia jellyfish. To snorkel there is a surreal experience.

DIVING PALAU

Perhaps the most famous of all dive sites in Palau is Blue Corner. Jutting into deep water at the western edge of the island chain, this site has it all: swirling schools of barracuda, patrolling sharks, elegant eagle rays, and bustling Napoleon wrasse. Close to Blue Corner is another highlight, showing off Palau's renowned cavern diving—the Blue Holes. Here, divers enter a vast chamber illuminated only by pillars of light streaming through great gaps in the reef above. The Temple of Doom chamber, found at the rear of the Blue Holes, is a more serious dive that requires considerable experience and skill. The floor here is littered with the shells of sea turtles that were unable to find their way out of the chamber and perished. Chandelier Cave in Malakal Harbor is a more sedate alternative for those wishing to explore an easy cave system.

RELICS OF CONFLICT

In addition to the great natural diversity on show at Palau, there are many fascinating relics from World War II.

One of the best shipwrecks from that era is that of the Japanese fleet tanker *Iro*. At 475 ft (145 m) in length, the *Iro* is large and exceptionally well preserved. It lies in moderate depths in a lagoon near the central town of Koror. Visiting divers should also see the wreck of the Aichi E13-A1 (codename "Jake") seaplane. Discovered a decade ago, this wonderfully intact aircraft lies in shallow water, and is an easy dive with many layers of interest.

The wreck of the "Jake" seaplane

ESSENTIAL INFORMATION

WATER TEMPERATURE 77–82°F (25–28°C)

ECOSYSTEM Tropical

WHEN TO VISIT Year-round diving. Rainy season between July and October.

MUST-SEES Blue Corner; Blue Holes; Jellyfish Lake; the wreck of the *Iro*; the “Jake” seaplane wreck.

PATTERN OF ISLANDS An aerial view of a cluster of islets at Palau, ringed with white coral sand and azure seas.

Yap

CAROLINE ISLANDS, NORTH PACIFIC

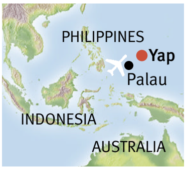

A remote island paradise, Yap has become a magnet for divers seeking close encounters with manta rays. However, there is a great deal more to this tiny Micronesian island, with plenty of sharks on patrol, including gray, white-tip, and black-tip reef species.

PACIFIC IDYLL

With its lush mangroves and rolling green hills, Yap lies in clear waters that abound with tuna and dolphins. It has one of Micronesia's most vibrant indigenous cultures, and retains a genuinely remote island feel in all aspects of daily life.

MEETINGS WITH MANTAS

The real attraction, however, lies beneath the waves. Yap is one of the very few dive destinations on Earth where encounters with manta rays are guaranteed. Two shallow tidal channels cut through the islands, and it is here that the mantas gather at cleaning stations that dot the sandy sea floor. At these sites, divers can observe the activity of smaller fish called remoras, which use suckers to attach themselves to the rays and remove irritating parasites by nibbling them off. Divers crouch by the stations—normally isolated coral heads—and await the arrival of the star attraction. Manta rays can grow to more than 2,200 lb (1,000 kg) in weight, and their wingspan can sometimes exceed 20 ft (6 m) from tip to tip.

ESSENTIAL INFORMATION

WATER TEMPERATURE Around 82°F (28°C) year-round.

ECOSYSTEM Tropical

WHEN TO VISIT Year-round, but December to April is best.

MUST-SEES Manta rays visiting cleaning stations; shark dives; Yap Caverns.

PROTECTED ENVIRONMENT

Other dives in Yap should not be overlooked—there is some tremendous shark action here, and the reefs are in pristine condition due to a strictly-adhered-to agreement to use traditional fishing techniques. Such forward-thinking legislation is of great benefit to both marine life and visiting divers, and is an admirable example of responsible stewardship of the seas. Nearly 80 percent of the reefs around Yap remain undived due to local conservation by-laws, although the remaining 20 percent are well worth exploring.

ON THE WING
Despite their impressive and imposing size, manta rays are a docile species that feed on microscopic plankton life.

Bikini Atoll

MARSHALL ISLANDS, CENTRAL PACIFIC

Bikini Atoll
Pacific Ocean
AUSTRALIA
NEW ZEALAND

Bikini Atoll vies with Truk Lagoon to be the finest wreck diving location on Earth. This tiny collection of coral islands and lagoons is the site of spectacular and dramatic wrecks, with a truly extraordinary story behind their sinking.

ATOMIC TEST SITE

As World War II ended and the Cold War began, the US chose a tiny atoll at the northern end of the Marshall Islands to conduct a massive program of nuclear weapons testing. Ships of varying size, nationality, and vintage made up the 84 vessels of a dummy fleet, the vast majority of which were sunk in tests running from 1946 to 1958. Included in the fleet was the US aircraft carrier *Saratoga*—weighing 33,000 tons and now the largest shipwreck accessible to divers anywhere on Earth. Also at Bikini is the *Nagato*, the flagship of the Japanese fleet, and the first warship in the world to be armed with 16-inch guns. The last active Japanese battleship, the *Nagato* saw action as late as July 1945, when it was attacked by US torpedo bombers.

The ships lie in deep water, with depths averaging 160–180 ft (50–55 m), but with 1,500 or so fish species swarming over the stunning remains, the rewards are well worth it for divers with sufficient training and experience. No dive site can offer a more dramatic insight into history.

ESSENTIAL INFORMATION

WATER TEMPERATURE Around 84°F (29°C) year-round.

ECOSYSTEM Tropical

WHEN TO VISIT Best time to visit is between April and September.

MUST-SEES The US carrier *Saratoga*; Japanese battleship *Nagato*.

GHOSTS FROM THE PAST
The wreckage of a naval aircraft destroyed during US nuclear bomb tests (*inset*) at Bikini Atoll. Residual radiation has since dropped to negligible levels at this dive site.

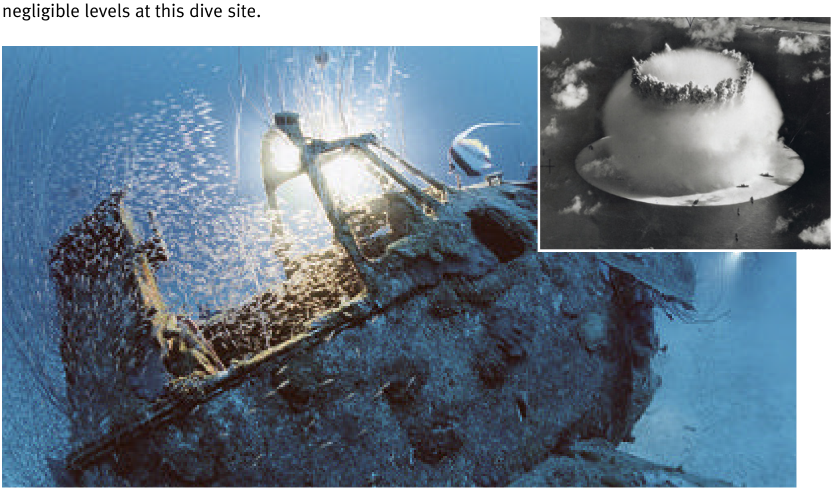

Bora Bora

SOCIETY ISLANDS, SOUTH PACIFIC

Pacific Ocean
Bora Bora
AUSTRALIA
NEW ZEALAND

Renowned as one of the most beautiful islands in the Pacific, Bora Bora is a huge lagoon dominated by an island at its center. The lagoon and the surrounding reef provide a dazzling array of dives against the stunning backdrop of volcanic peaks.

PACIFIC PARADISE

Bora Bora is part of the Society Islands, and is located 170 miles (274 km) northwest of Tahiti. An atoll of sparkling beauty that has long entranced writers and artists alike, the main island is only 6 miles (10 km) long, and sits at the heart of a lagoon three times its size. The reef around the lagoon has only one navigable channel, which is crisscrossed by smaller channels and gullies, most of which provide excellent dive sites.

ABUNDANCE OF LIFE

Bora Bora's lagoon and surrounding reefs are home to a huge variety of fish—a recent survey documented 500 different species, although the real number is almost certainly higher. Larger animals encountered on the outer reef and in the cuts in the atoll walls are gray reef sharks, white-tip reef sharks, groups of black-tip reef sharks, lemon sharks, and silvertips. Numerous species of rays can be encountered in the central lagoon, including the groups of manta rays for which the site has become justifiably famous.

ESSENTIAL INFORMATION

WATER TEMPERATURE 79–84°F (26–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT Thanks to its balmy climate, diving in Bora Bora can be enjoyed year-round.

MUST-SEES Manta ray encounters at Anua; shark dives at Tupitipiti.

ANATOMY OF AN ISLAND
Bora Bora is dominated by the jagged remains of a volcano, and encircled by a great lagoon. This remote atoll is widely considered to be one of the most beautiful spots on Earth.

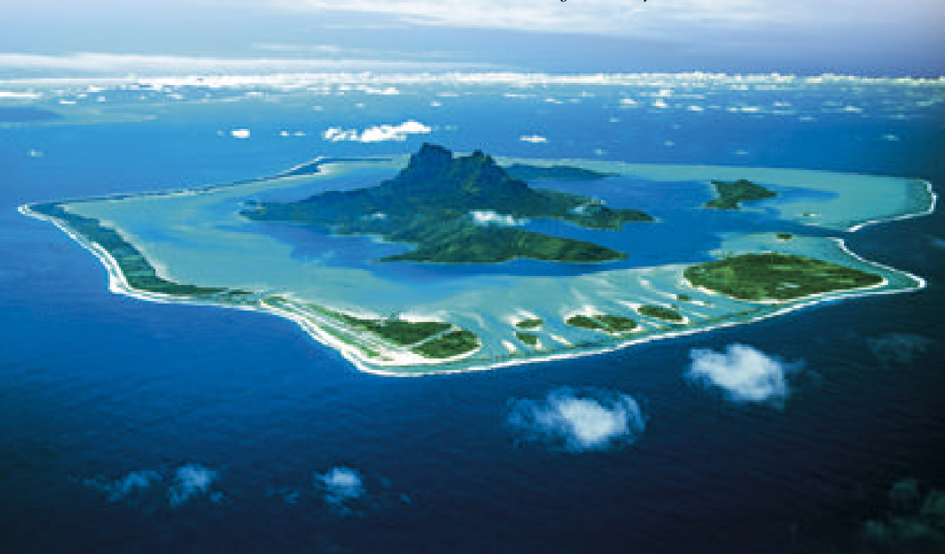

Wreck of the *President Coolidge*

VANUATU, SOUTH PACIFIC

Pacific
Ocean
Vanuatu
AUSTRALIA

A relic of an elegant age of sea travel, the *President Coolidge* is a strong candidate for the most exciting wreck now accessible to recreational divers. Perched on a reef in Vanuatu, it hides one of the most famous icons in diving in its dark interior.

BEACHED BEHEMOTH

The *President Coolidge* sank on October 26, 1942 while carrying 4,800 American troops to the Pacific during World War II. It struck floating mines, laid by "friendly" forces in the area, and when it became clear the ship would not make it into the nearby harbor of Luganville, the captain beached the huge vessel on a reef. Here it remains today, with its stern submerged in 236 ft (72 m) of water, although the bow lies at a depth of just 49 ft (15 m).

DIVING THE *COOLIDGE*

The wreck is accessed directly from the shore. Within its bulk is a network of passageways and enormous holds containing the paraphernalia associated with equipping an army for war. Although converted for duty as a troop ship, there are still traces of its time as one of the world's most luxurious ocean liners.

An essential stop for any diver visiting this magnificent wreck is the statue of the Lady, in one of the first-class dining rooms. This once stood on the mantelpiece of the smoking room, but a small earthquake dislodged it from its perch, and it was later moved farther up the wreck. Other highlights include the promenade deck and a beautiful mosaic fountain, once the centerpiece of a gracious lobby.

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–82°F (24–28°C)

ECOSYSTEM Tropical

WHEN TO VISIT There is year-round diving on the wreck.

MUST-SEES The Lady; the swimming pool; lobby mosaic; the promenade deck; lampfish dive.

CASUALTY OF CONFLICT
The *President Coolidge* ran aground at Vanuatu during World War II. The wreck is still filled with corroding relics from that era.

Truk Lagoon

CAROLINE ISLANDS, NORTH PACIFIC

Truk Lagoon
INDONESIA
Pacific Ocean
AUSTRALIA

Regarded by many as the finest of all wreck sites, Truk Lagoon is the resting place for an eclectic mix of Japanese vessels sunk in a single raid by American aircraft in 1944. With many wrecks still in good condition, Truk Lagoon offers unforgettable diving.

WITNESS TO WAR

Truk Lagoon is a massive natural atoll almost 40 miles (65 km) across; it is ringed by 140 miles (225 km) of reef and small coral islands, forming one of the largest natural harbors of this type on Earth. The presence of 15 larger islands in the center of the lagoon made it an ideal base for military operations, a fact not lost on the Japanese Imperial Navy, which amassed a vast number of warships, merchant vessels, and fleet tenders here in the latter stages of World War II, in readiness for further deployment.

DIVING TRUK LAGOON

Wreck diving in Truk is deep, with the majority of the wrecks lying in depths of 100 ft (30 m) or more. Local operators are, however, well set up to cope with deep air diving, and have established effective drills and techniques to cope with repeated deep dives. Nonetheless, a certain level of qualifications and experience is advisable before diving the fascinating wrecks of Truk Lagoon. Contact local operators for details.

Marine life at Truk

ESSENTIAL INFORMATION

WATER TEMPERATURE 84°F (29°C)

ECOSYSTEM Tropical

WHEN TO VISIT Year-round diving, but rainy season from July to October.

MUST-SEES The wrecks of the *Fujikawa Maru*, *Shinkoku Maru*, *San Francisco Maru*, and the "Betty" bomber aircraft.

Such an assembly of enemy vessels proved irresistible to American forces, which carried out a daring raid on February 17–18, 1944, launching 450 aircraft from nine aircraft carriers beyond the horizon. The raid was a complete success, with over 60 Japanese vessels sunk in a savage 48 hours.

SECRET SITES

The "ghost fleet" of Japanese ships lies on the seabed in Truk Lagoon, although only 38 of them have been charted. The others have been left unmarked for fear of looting and illegal diving (to dive the wrecks in Truk, a local guide is a legal and practical necessity). Highlights include the *Fujikawa Maru*, complete with Zero fighter aircraft still in its holds, and the *San Francisco Maru*—a huge freighter with an array of cargo including two tanks on its deck, so outstanding it is known as the "Million Dollar Wreck."

RELICS FROM THE PAST

A diver hovers over the ghostly remains of a Japanese ship sunk at Truk Lagoon. Aircraft and submarines also fell victim to aerial attack here.

Fiji

SOUTH PACIFIC

Pacific Ocean
Fiji
AUSTRALIA
NEW ZEALAND

Simply hearing the name "Fiji" summons up images of an exquisite tropical paradise. With 800 islands sprawling over 80,000 sq miles (207,000 sq km) of ocean, the reality is just as exciting. Fiji offers a wealth of different experiences for the diver.

IDYLLIC ISLANDS

Perhaps one of the most notable features of Fiji for the visiting diver is that only 12 percent of the islands in the archipelago are inhabited—and 75 percent of the entire population lives on just one island, Viti Levu. Fiji truly offers a wealth of different diving opportunities around a huge variety of reefs, channels, and islands. Coral reefs cover a staggering 6,000 sq miles (15,540 sq km); they include at least 398 species of coral and are home to over 1,200 species of fish.

UNDERSEA GARDENS

Fiji is particularly famous for its beautiful soft-coral gardens, the most renowned of which is the White Wall in the Somosomo Strait. The region's rich waters mean that filter-feeding invertebrates thrive in the coral lagoons. Sea fans, whip corals, and a myriad of other species vie for space on the brilliantly colored reef walls.

REEF RESIDENT
Sharks of many species, including gray reef sharks, patrol the soft-coral reefs of Fiji's scattered islands.

There are also many spectacular fish species in Fijian waters. For the thrill-seeker, there is plenty of large-animal action. The best-known shark dive here is known as "Aqua Trek 3-D," after the operator that runs it, and offers the opportunity to see up to seven species of sharks on a single dive, including bull, black-tip reef, white-tip, and lemon sharks. The dive is one of 100 recognized dive sites in Beqa Lagoon—an indication of the sheer variety of diving in this beautiful tropical archipelago. Happily, the area has been designated a Marine Park and enjoys protection from overfishing; the small fees asked of divers who experience its underwater environment are used to help fund conservation measures. Such careful stewardship ensures that Fiji's diving will remain outstanding for years to come.

ESSENTIAL INFORMATION

WATER TEMPERATURE 79–88°F (26–31°C)

ECOSYSTEM Tropical

WHEN TO VISIT Year-round, but the best time to visit Fiji is between April and November.

MUST-SEES Somosomo Straight and the White Wall; Beqa Lagoon; Naingora Pass.

Anemonefish

Tonga

FRIENDLY ISLANDS, SOUTH PACIFIC

Pacific Ocean
AUSTRALIA
Tonga
NEW ZEALAND

Tonga is a dramatic chain of volcanic islands to the east of Fiji. The grandeur of its scenery is matched by one of the most amazing encounters with a marine mammal that a diver can have anywhere on Earth—snorkeling with humpback whales.

PACIFIC OUTPOST

Diving in Tonga is based mainly in the central island group of Ha'apai, and in Vava'u farther to the north. Port of Refuge in Vava'u, which for many divers is their first experience of Tongan diving, contains a particularly fine wreck in the form of the *Clan McWilliam*, a large merchant vessel sunk in the 1920s.

ESSENTIAL INFORMATION

WATER TEMPERATURE 72–86°F (22–30°C)

ECOSYSTEM Tropical

WHEN TO VISIT July to October for humpback whales; avoid typhoon season from November to April.

MUST-SEES Humpback whales; Hunga Magic; Mariners Cave; *Clan McWilliam*.

STAR SITES

Other sites of note in the region are Tuangasika Island, the location of an exciting cavern called Mariners Cave, and the steep walls of Hunga Magic—an undersea peninsula jutting into the blue water from Hunga island. Tonga's fame in the diving world is, however, based on encounters with humpback whales. From July through to October, around 700 humpbacks move into warm waters around the island group. Although local conservation groups do not allow scuba diving with the whales (to minimize disturbance to them), snorkeling alongside these amazing mammals is a unique experience.

ANNUAL MIGRATION
A humpback whale raises its massive flukes above the surface. Large numbers of this magnificent animal pass by Tonga each year.

The Americas

As befits a landmass that spans virtually all latitudes of the globe, the Americas boasts a staggering range of dive sites. Great ocean currents beset both North and South America, combining with local conditions to create something for divers of every level of expertise.

COASTAL DIVIDE

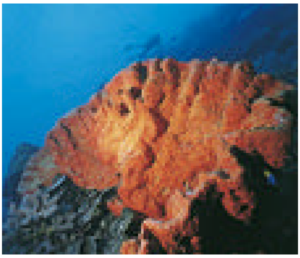

EASILY EXPLORED
The superb visibility and calm, warm waters of the Caribbean provide ideal conditions for reef diving.

The coasts of the North American continent offer very different, but equally thrilling, diving environments. To the west, you can enjoy magnificent wild diving in the fjords of Alaska and British Colombia, explore vast kelp forests off the California coast, and experience glorious encounters with humpback whales and whale sharks in the Gulf of California.

The eastern coast features wrecks and raggedtooth sharks on the Outer Banks off North Carolina—the region where technical diving made its greatest leaps forward. The Florida Keys provide reef diving of the highest quality, and are popular as a training ground for thousands of divers who try the sport while on vacation.

The Caribbean offers classic coral reef diving. With its well-developed diving infrastructure, it has long been a favorite dive destination. To the south, in Central America, is the Belize barrier reef—the longest in the Northern Hemisphere. Off the coast of Venezuela lies Bonaire, in the Dutch Antilles, a beacon of marine conservation. Farther south still is the island of Fernando de Noronha off Brazil, perhaps the finest dive site in the whole of South America.

SEA SAFARI
Even snorkeling over sandbars in the Cayman Islands, western Caribbean, can bring you face to face with fascinating marine animals.

ENCHANTED FOREST
The towering kelp forests in the shallow waters off California are among the most beautiful and biologically rich of all marine habitats.

Florida Keys

FLORIDA, UNITED STATES

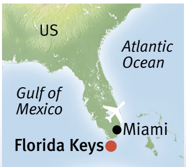

The Florida Keys are a string of coral and limestone islands stretching 120 miles (193 km) into the Gulf of Mexico. For many, this popular vacation destination represents a first introduction to diving, and there are dive sites of real quality and variety here.

SUNKEN TREASURES
The wreck of a tugboat lying in clear blue water. The Keys are famed for good underwater visibility and excellent areas for training.

PROTECTED ECOSYSTEM

The Florida Keys consist of 200 small islands, the largest of which are connected by the 42 bridges of US Highway 1. The importance of the Florida Keys as a vibrant marine ecosystem was recognized as long ago as 1960, with the John Pennecamp Coral Reef State Park established as the first-ever marine park in the US. More followed, and the region now boasts 18 protected areas, providing cover out to 3 miles (5 km) beyond the reef.

UNDERSEA ATTRACTIONS

The reef itself runs parallel to the island chain, roughly 5 miles (8 km) offshore. As such, most of the diving here is from boats, and there are many experienced operators. There are some excellent wrecks, as well as classic spur and groove reef formations, with the outer reef inhabited by a diverse range of Caribbean species. The shallow inner reef also hosts manatees, as well as providing a nursery for smaller species and huge numbers of trainee divers.

ESSENTIAL INFORMATION

WATER TEMPERATURE 70–84°F (21–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT Year-round diving, but hurricane season October to November.

MUST-SEES Molasses Reef, wreck of the *Adolphus Busch Sr.*; the Christ of the Deep statue in Key Largo.

Crystal River

FLORIDA, UNITED STATES

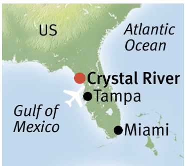

The Crystal River in northwest Florida is parent to a network of channels and tributaries snaking through the swamps. What draws divers to the region is the chance to swim with manatees—one of the great animal encounters in the underwater world.

WEB OF WATERWAYS

The Crystal River stems from a series of inland upwellings through porous rock, creating a natural filtering system that results in some of the clearest water on Earth. What this meeting of rock and water also creates is a honeycomb of caverns, caves, and tunnels. Although these are frequently occupied by huge catfish (a draw in themselves), the real attraction for the diver is the presence of the West Indian manatee.

GENTLE GIANTS

These huge and placid animals—measuring up to 14 ft (3 m) long and weighing up to a ton and a half—have been using the shallow pools and warm streams of the Crystal River to feed and breed for thousands of years. Speedboats and habitat destruction have exacted a terrible toll on the manatees, which are now on the endangered list. Diving with the manatees is not permitted, but nonetheless, snorkeling next to one of these huge animals in crystal-clear water is a tremendous experience.

ESSENTIAL INFORMATION

WATER TEMPERATURE 68–81°F (20–27°C)

ECOSYSTEM Freshwater

WHEN TO VISIT Crystal River can be visited year-round, but for snorkeling with manatees, the best time is between the months of December and April.

MUST-SEES Devil's Den, King's Cavern; swimming with manatees.

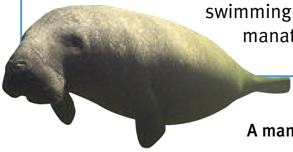

A manatee

CLEAR ADVANTAGE
The stunning clarity of the water at Crystal River stems from geological conditions in the area. Such waters are perfect for observing manatees.

Little Bahamas Bank

BAHAMAS, WESTERN ATLANTIC

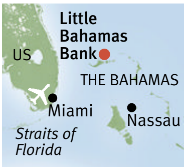

The Little Bahamas Bank would appear to be the most unpromising of dive sites—a vast patch of featureless, shallow sand bank. One animal has, however, made this area irresistible to divers. Dolphin encounters here are uniquely memorable.

WILD ENCOUNTER
Spotted dolphins have frequented the Little Bahamas Bank for many years, and these inquisitive animals freely interact with snorkelers in the shallow waters of the Bank.

SHALLOW SEA

The Little Bahamas Bank lies between the islands of Grand Bahama and Great Abaco, two of the larger islands of the 700 that make up the Bahamas archipelago. The northern end of the bank lies 70 miles (113 km) east of the Florida coast, allowing access to the region from the mainland US as well as the Bahamas. The bank is relatively featureless, consisting of shallow white sand dotted with low-lying reefs and the occasional wreck. Marine life is relatively sparse compared to other regions in the Bahamas, although there is the possibility of encounters with some enormous barracuda, several species of turtles, and a number of shark species.

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–82°F (24–28°C)

ECOSYSTEM Tropical

WHEN TO VISIT Year-round, but hurricanes may strike later in the year.

MUST-SEES Spotted dolphin snorkeling encounter; the Bull Pit (shark dive); the Sugar Wreck.

STARS OF THE SEA

Both bottlenose and spotted dolphins visit the bank, but it is the guaranteed encounters with the latter that have made the site famous. These wild pods of dolphins have been interacting with divers for 20 years, and now a booming industry has grown up around these wonderful animals. Although only snorkeling is permitted with the dolphins, divers who have the privilege to encounter them in these clear, shallow seas take away lifelong memories.

Bonaire

DUTCH ANTILLES

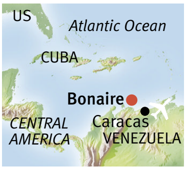

For conservationists, Bonaire is a beacon of hope in the Caribbean. More than 40 years ago, the island's authorities realized that good reef management could promote tourism and boost the local economy. Bonaire is particularly noted for fine shore diving.

FORWARD THINKING

Bonaire is the most eastern island in the group that makes up the Dutch Antilles. It has always demonstrated an enlightened attitude toward conservation, with protective laws first passed in 1961 (to protect turtles), followed by a steady stream of additional measures over subsequent years that eventually resulted in the establishment of a marine park that now encircles the island. This extends down to 200 ft (60 m).

ESSENTIAL INFORMATION

WATER TEMPERATURE 77–84°F (25–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT There is year-round diving on Bonaire, but hurricane season begins in the later part of the year.

MUST-SEES Town Pier at Kralendijk; the wreck of the *Hilma Hooker*.

SANDY SHORES

Bonaire offers a wealth of shore diving possibilities. Many divers simply turn up on the island, rent a car, and go exploring. The spontaneity of this kind of diving has a charm all of its own. As well as splendidly pristine reefs, Bonaire also has two notable wrecks in the *Hilma Hooker*, a freighter lying at 100 ft (30 m), and the *Red Slave*. Diving under the Town Pier at Kralendijk, and Salt Pier (on the other side of the island), is also very rewarding.

LEISURELY PACE
Divers join a turtle in a tour over the lush reefs of Bonaire's shallow coastal waters.

Cayman Islands

SOUTHERN CARIBBEAN

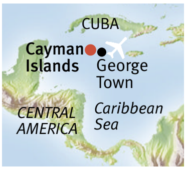

The Caymans are three low-lying islands 150 miles (240 km) south of Cuba. Views of the islands above the water belie the riches that lie beneath—steep walls, fine reefs, and a very shallow dive that is famed among the diving fraternity.

CLEAR WATERS

The Cayman Islands have two key characteristics that make them a classic dive area: they lack any mountains or large rivers, and the limestone bedrock absorbs water. Consequently, there is very little sedimentary runoff from the land when it rains, which ensures that the visibility underwater is excellent. Added to this are steep walls, sandy lagoons, and 18 protected marine life reserves that ensure excellent diving.

ESSENTIAL INFORMATION

WATER TEMPERATURE 81–84°F (27–29°C)

ECOSYSTEM Tropical

WHEN TO VISIT Diving takes place in the Caymans year-round, although hurricane season starts late in the year.

MUST-SEES Stingrays in North Sound; the mantas of Bloody Bay Wall.

REEFS AND RAYS

The underwater environment is typified by shallow fringing reefs extending some way from shore, leading to busy reef crests and steep drop-offs. Large pelagics, such as mantas, eagle rays, and several species of sharks, drift past walls characterized by healthy fish populations and rich coral and sponge growth. The Caymans are famous for one particular stingray feeding site in North Sound, Grand Cayman, where as many as 200 stingrays cluster around divers in the shallows.

CLOSE ENCOUNTER
Stingrays gather to be fed at North Sound. Contact with them should always be on their terms.

Turks and Caicos

BAHAMAS, NORTHERN CARIBBEAN

US
Turks and Caicos
CUBA
Caribbean Sea
CENTRAL AMERICA

These two tiny island groups at the southeastern end of the Bahamas chain offer the diver true variety within a small area—from safe, shallow lagoons to steep walls and strong currents—with big animal encounters always possible.

DIVE PARADISE

The Turks and Caicos consist of two small clusters of islands on either side of the abyssal Turks Island Passage. To the west are the Caicos Islands, epitomized by shallow lagoons and gentle slopes. To the east, the smaller Turks Islands offer access to stunning drop-offs only a short boat ride from shore—although both sets of islands do have a range of dives within these broad categories, catering to the tentative beginner through to the hardened veteran.

PROTECTED STATUS

There are 33 protected areas within the Turks and Caicos, and these islands were the first in the entire Caribbean to ban spearfishing. As such, the marine life here is very impressive indeed, with healthy reefs dotted with huge sponges, and impressive pelagic activity in the form of dolphin encounters, gray reef, bull, and tiger sharks, as well as three resident species of turtles. In addition to these animals, large migrants also create excitement. From December through to April, the entire humpback whale population of the Atlantic passes through Turks Island Passage.

For wreck fans, the remains of the *Endymion*—a British warship that ran aground in 1790—offer a fascinating glimpse into naval history. Cannonballs and other relics can still be found on the seabed around the site. The wreck lies in shallow water 18 miles (30 km) south of Salt Cay; the site is readily accessible by boat.

THRIVING REEFS
A diver inspects growths of colorful tube sponges, which are a common sight on the reefs of the Turks and Caicos region.

ESSENTIAL INFORMATION

WATER TEMPERATURE 81–86°F (27–30°C)

ECOSYSTEM Tropical

WHEN TO VISIT Visit the Turks and Caicos between November and August

MUST-SEES Humpback whale migration, dolphin encounter; Northwest Point dives; the Tunnel; *Endymion* wreck.

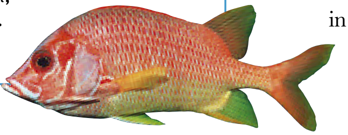

Giant squirrelfish

Cuba

WESTERN ATLANTIC

US
THE BAHAMAS
Havana
CUBA
JAMAICA
CENTRAL AMERICA

It did not take long for Cuba's developing dive industry to draw the attention of the international diving community. Certain classic sites have been well known for years, but Cuba's vast and varied coastline has much more to offer traveling divers.

GIANT COASTLINE

Cuba is a huge island 90 miles (155 km) south of Florida, with 3,500 miles (5,630 km) of coastline bordered by the Gulf of Mexico to the west, the Atlantic to the north, and the Caribbean to the south. The scale of the island is such that each of the three reefs along the coast of Cuba exceeds the length of the entire Florida Keys.

MARINE DIVERSITY

A recent survey noted over 1,000 fish species around Cuba—tremendous diversity for this part of the world. There are also some excellent big-animal encounters to be experienced here, including some classic shark dives and large groups of tarpon.

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–81°F (24–27°C)

ECOSYSTEM Tropical

WHEN TO VISIT Diving continues year-round in Cuba, but hurricane season begins late in the year.

MUST-SEES The Isle of Youth; bull shark dive at Santa Lucía; *Cristóbal* wreck.

CLASSIC CUBAN DIVES

Cuba has an abundance of riches in terms of dive sites, and some consider the diving here to be the best in the world. Diving on the Isle of Youth is particularly popular, but Santa Lucía, Guardalavaca, and Santiago de Cuba also offer excellent dives. The last of these boasts a particularly fine shore dive with the wreck of the *Cristóbal Colón.* This former Spanish warship was attacked by the US navy in 1898, during a brief conflict between the two countries provoked by the sinking of a US ship in Havana harbor. The *Cristóbal Colón* was run aground by its commander, and is now a splendid historical dive site.

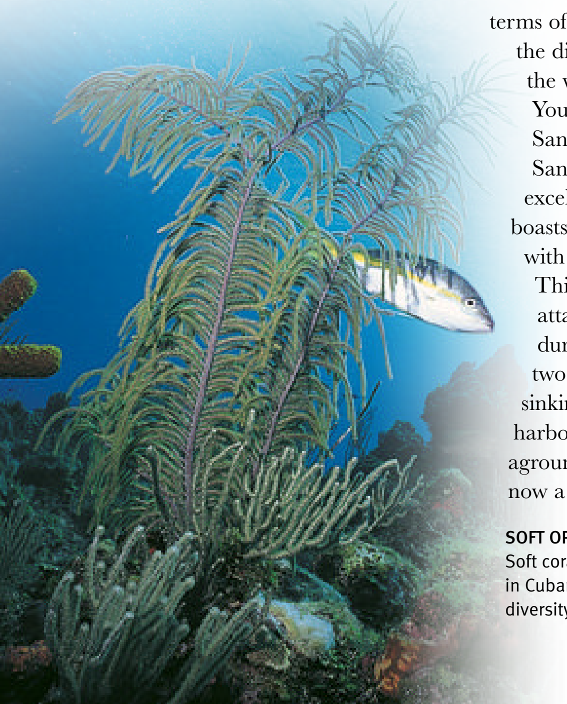

SOFT OPTION
Soft corals and sponges grow in profusion in Cuban waters, which enjoy unusually high diversity in marine species.

Cenotes

YUCATÁN PENINSULA, MEXICO

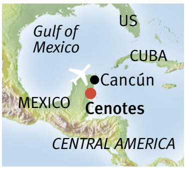

A network of flooded caves filled with crystal-clear water, the Cenotes offer the novice a chance to dive caverns in relative safety. For the more intrepid, and appropriately trained, the honeycomb of tunnels and passages offers advanced exploration opportunities.

INNER SPACE
A diver surveys the eerie interior of the Cenotes cave network. Though the system offers difficult technical challenges, it also has easier dives.

GEOLOGICAL WONDER

In Mexico's Yucatán Peninsula, the best dives are not at the coast but take place some distance inland, within the honeycombed heart of the Earth itself. The term "Cenotes" (from the Mayan word *tzonot*) refers to the entrances to the network of flooded passages and caverns that riddle this region.

Over 60 million years ago, this entire region was beneath the sea, creating rock strata of limestone from compressed coral and marine organisms. In the intervening period, this soft rock has been shaped by rain and rivers, creating one of the finest cave and cavern diving locations on Earth.

DIVING THE CENOTES

Although explorations are still ongoing in the Cenotes, many of the systems are accessible even for novice divers, with guides taking groups into vast caverns where an exit to the surface is always available if required. Sea diving at Cancún and Cozumel is available nearby.

ESSENTIAL INFORMATION

WATER TEMPERATURE 66–77°F (19–25°C)

ECOSYSTEM Freshwater cave

WHEN TO VISIT The best time of year to dive the Cenotes is between November and March.

MUST-SEES Gran Cenote; Car Wash Cenote; coastal resorts nearby.

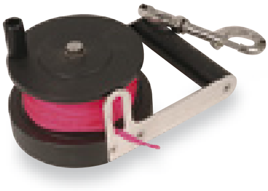

A SAFETY LINE
Special equipment helps ensure safe, enjoyable diving in caverns (*see pp.186–87*).

Belize Barrier Reef

BELIZE, CENTRAL AMERICA

CUBA
Yucatán
Belize City
Belize Barrier Reef
CENTRAL AMERICA

Belize is a small Central American country that has the good fortune to have a true natural marvel off its coastline. The Belize Barrier Reef is the longest in the Northern Hemisphere, and offers a vast range of diving possibilities for all skill levels.

SECLUDED COUNTRY

Belize covers around 8,880 sq miles (23,000 sq km) and has only 220,000 inhabitants. The country is renowned for its marine life, and due to the efforts of various voluntary organizations and its government, over 40 percent of Belize has been classified as a nature reserve.

THE BARRIER REEF

Belize also boasts a World Heritage Site—the Belize Barrier Reef. Extending for 174 miles (280 km) and ranging from 9 miles (14 km) to 19 miles (30 km) offshore, the reef offers a huge variety of divable sites to explore. Three of the four atolls in the Caribbean are found along the length of the reef. An atoll is a circular island that surrounds a lagoon, and is formed when coral reefs grow around the perimeter of a collapsed volcano. Linear sand and coral islands —known locally as cayes—also run

ESSENTIAL INFORMATION

WATER TEMPERATURE 77–82°F (25–28°C)

ECOSYSTEM Tropical

WHEN TO VISIT Diving in Belize takes place year-round, but the hurricane season begins around October.

MUST-SEES Whale sharks at Gladden Spit, Blue Hole, Shark Ray Alley.

along the length of the Barrier Reef, and provide a base for much of the area's dive activity. Of these, Ambergris Caye is the most popular location.

DIVING HIGHLIGHTS

Around 500 fish species swarm over the reef, including five different species of butterfly fish. Some larger animal species can also be seen, including huge jewfish and manta rays on the outer reef, and stingrays and nurse sharks at Shark Ray Alley. There are also healthy populations of reef sharks, as well as manatees (*see p.219*) in the shallow inshore waters and estuaries of the country.

While the most famous site in Belize is undoubtedly the Blue Hole, the area has a great deal more to offer, with classic reef dives, magnificent spur and groove reef formations, and some notable seasonal gatherings of big animals. At Gladden Spit in the south of the country, large numbers of whale sharks meet in feeding aggregations during the early part of the year. These huge ocean-going plankton-feeders grow to as much as 60 ft (18 m) in length.

THE BLUE HOLE

Perhaps the most famous site in Belize is the Blue Hole. Famously explored by Cousteau in 1972, this vast cavern lies 50 miles (80 km) off the east coast of Belize in the middle of Lighthouse Reef. Its perfectly circular entrance measures some 985 ft (300 m) in diameter, with the cavern plunging to a depth of 410 ft (125 m) beneath. The resident population of bull and lemon sharks in the interior of the Blue Hole is an eerie sight for divers visiting the stalactite formations found 130 ft (40 m) below the surface.

Aerial view of the Blue Hole

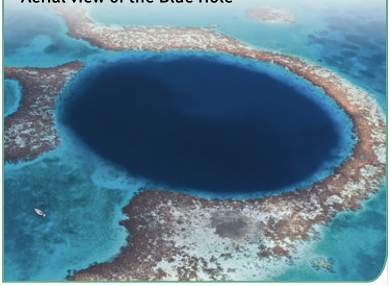

CYCLE OF LIFE
Whale sharks come to Belize to feed on the vast numbers of eggs generated by the mass spawning of a number of reef fish species, such as dog and cubera snappers, in spring.

Fernando de Noronha

WESTERN ATLANTIC

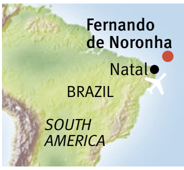

Rising up from the wild waters of the Atlantic, 250 miles (400 km) off the eastern coast of Brazil, is Fernando de Noronha. This tiny group of volcanic islands plays host to one of the largest resident pods of dolphins on Earth.

ATLANTIC TREASURE

The stark island group of Fernando de Noronha was for many years a well-kept secret among the Brazilian diving fraternity. The sea around the islands plunges to depths of over 2,300 ft (700 m), creating rich upwellings and swirling currents.

CLOSE ENCOUNTERS

Although the reefs here are relatively bare—only 95 fish species have been recorded on them—it is the big animal encounters that have made the site famous. Stories of extraordinary interactions with the huge resident pods of spinner dolphins around the islands have attracted divers who now travel here from around the world. Some 14 species of sharks, including nurse sharks and gray reef sharks, cruise over dark ridges, and two turtles are commonly sighted; one—the hawksbill—actually breeds here. The islands have been a World Heritage Site since 1981, a well-deserved status for this exciting site.

ESSENTIAL INFORMATION

WATER TEMPERATURE 70–79°F (21–26°C)

ECOSYSTEM Subtropical

WHEN TO VISIT Diving is best at Fernando de Noronha between September and March.

MUST-SEES Spinner dolphins at Baia dos Golfinhos; gray reef sharks at Laje dos Dois Irmãos.

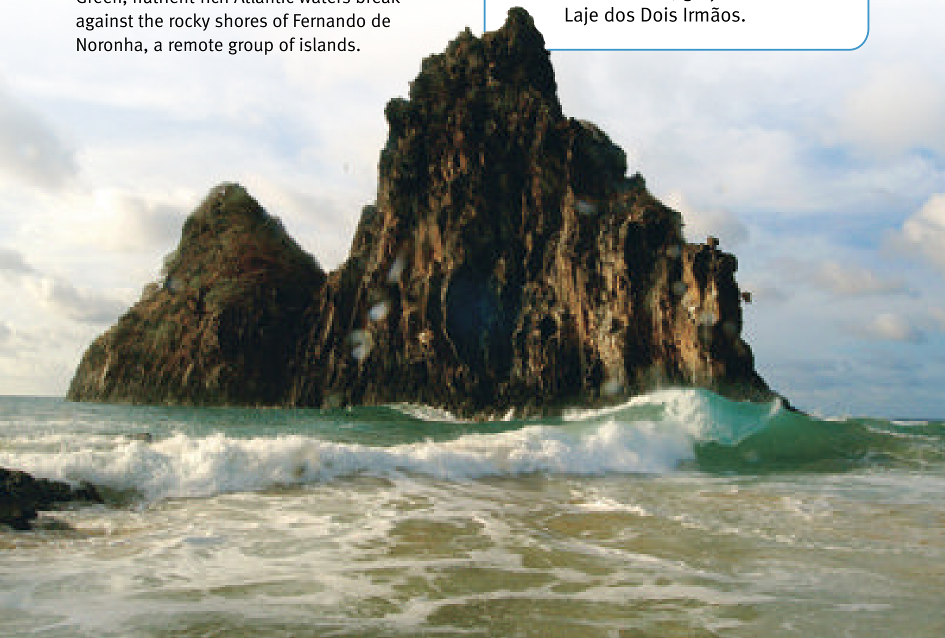

WILD BEAUTY
Green, nutrient-rich Atlantic waters break against the rocky shores of Fernando de Noronha, a remote group of islands.

Cocos and Malpelo

EASTERN PACIFIC

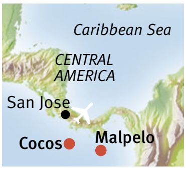

Considered by many to be among the finest big animal dive sites on Earth, the two volcanic islands of Cocos and Malpelo sit several hundred miles from land amid the wild currents and big swells of the eastern Pacific ocean.

DOUBLE IDENTITY

Although Cocos and Malpelo are generally referred to in the same breath by divers, the two islands are actually 250 miles (400 km) apart—Cocos off the coast of Costa Rica, and Malpelo off the coast of Colombia.

MARINE SANCTUARY

Rising from the deep waters of the Pacific in a series of serrated volcanic reefs and craggy pinnacles, the two islands provide a fantastic opportunity to experience larger marine animals at close range. Schooling hammerheads, hunting white-tip reef sharks, patrolling silky and Galápagos sharks—are all virtually guaranteed. Spotted moray eels are present in huge numbers, and itinerant visitors include whale sharks, mantas, and a number of bill-fish species. Both islands have a protected status—Cocos is a World Heritage Site, and Malpelo a Colombian Flora and Fauna Sanctuary.

ESSENTIAL INFORMATION

WATER TEMPERATURE 63–75°F (17–24°C)

ECOSYSTEM Subtropical

WHEN TO VISIT Conditions are best from December to September.

MUST-SEES Hammerhead schools; spotted morays in Malpelo; white-tip reef sharks hunting prey.

MEETING PLACE
Sharks of many species congregate in numbers off Cocos Island. Use of a rebreather (*left*) allows close approaches.

TOUGH CONDITIONS

Diving on Cocos and Malpelo is demanding, with big swells, swirling currents, occasional limited visibility, and stark reefs. The rewards for braving the challenging conditions, however, are immense, given the ample opportunities for big animal encounters. Some shark species, especially hammerheads, are naturally wary of divers, but a careful approach can work wonders. Specialist equipment can help, too—rebreathers (*see pp.82–83*), which do not create bubbles, allow the diver to make an unobtrusive approach and thus give the best chance for close encounters.

Roca Partida

COLIMA STATE, MEXICO

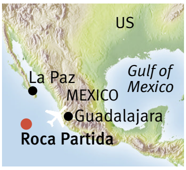

This stark volcanic pinnacle rises from the depths of the Pacific 200 miles (320 km) off the tip of the Baja Peninsula. This is truly wild diving, with powerful swells, big currents, and large marine species at one of the most isolated dive sites in the world.

SHEER DIVING

The most remote outcrop of the Archipelago of Revillagigedo (a group of four islands), Roca Partida rises starkly from the wild surface of the Pacific Ocean, its rugged silhouette reflecting the characteristics of the reefs beneath. Steep walls plunge into deep, dark water, with very little life encrusted on their barren surfaces.

ESSENTIAL INFORMATION

WATER TEMPERATURE 66–77°F (19–25°C)

ECOSYSTEM Subtropical

WHEN TO VISIT Roca Partida has good, year-round diving, but cooler winter water is best for shark diversity.

MUST-SEES Manta rays; sleeping white-tip and silver-tip sharks; black jacks.

BOUNTIFUL SEAS

The rich upwellings from deep water around Roca Partida provide plentiful nutrients and oxygen for circling fish shoals, including huge black jacks and dense schools of barracuda. These in turn attract silky, white-tip, silver-tip, and hammerhead sharks in great numbers. What makes this dive site so special, however, are the manta ray encounters, which are among the best to be had anywhere; these magnificent animals almost seem to seek out divers, resulting in some extraordinary interactions.

The reputation of Roca Partida is spreading, and it is rapidly becoming an iconic site in the diving world. Divers willing to make the journey to this remote and wild location will find themselves amply rewarded.

SHARK ENCOUNTERS
Divers at Roca Partida can count on being buzzed by many different sharks, such as the normally shy scalloped hammerhead.

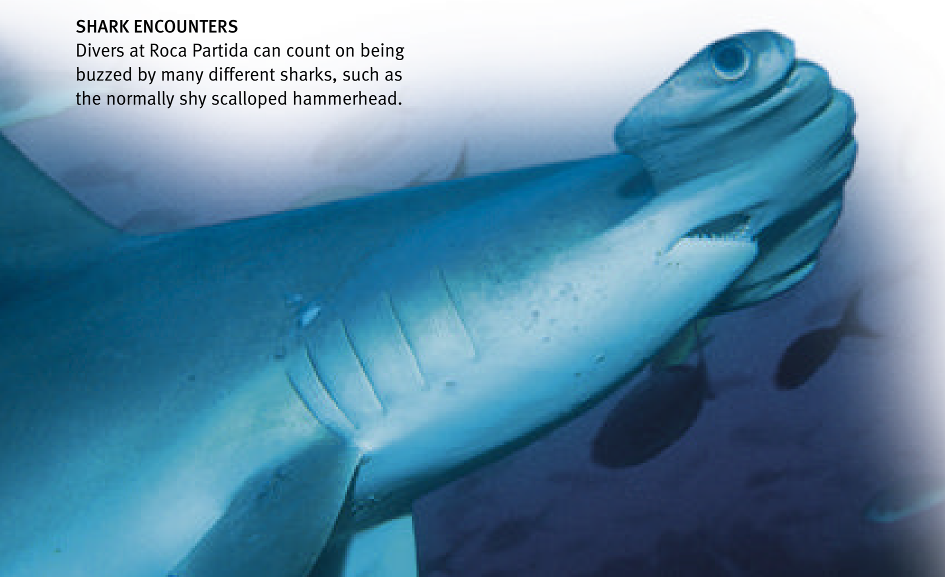

Hawaii

CENTRAL PACIFIC OCEAN

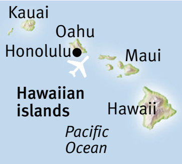

The underwater terrain of Hawaii reflects its volcanic origins, with twisted lava formations and steep drop-offs into abyssal depths. The oceanic island group is world-renowned for its surfing, but offers exceptional diving, too.

ISLAND PARADISE

Hawaii is the world's longest island chain, extending more than 1,500 miles (2,400 km), and situated over 2,500 miles (4,000 km) from the mainland. It comprises eight main islands—each the top of a submerged volcano—as well as another 124 tiny areas of uninhabited land that amount to no more than 3 sq miles (8 sq km). The stark ridges and bizarre rock formations that epitomize Hawaii's terrain above sea level extend beneath the water's surface, offering exciting dive opportunities, and great wildlife encounters. Hawaii's extreme isolation means it is visited by nomadic species from the vast expanse of the Pacific Ocean. The most notable of the pelagic animals are humpback whales, pygmy pilot whales, manta rays, and oceanic white-tip sharks. Humpbacks have been hunted globally almost to extinction, and have become rare even in the waters around Hawaii. Excellent dive sites can be found on the island of Hawaii—or Big Island—at the eastern end of the island chain, as well as Maui, Kauai, and Oahu.

ESSENTIAL INFORMATION

WATER TEMPERATURE 75–81°F (24–27°C)

ECOSYSTEM Tropical

WHEN TO VISIT Year-round, but winter months at this busy site are less crowded.

MUST-SEES Manta rays at Kona Beach; pygmy pilot whales; Shark's Cove, Oahu; Molokai; Long Lava Tube, Big Island.

WAVE-SWEPT COASTLINE
Hawaii lies at the same latitude as central Mexico. Clean waters, pure sands, and a tropical climate add to its appeal as a dive destination.

Catalina Island

CALIFORNIA, UNITED STATES

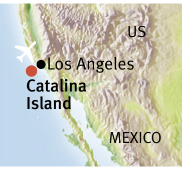

Lying only 22 miles (35 km) off the coast of California is a spectacular island of dramatic scenery above the water, and tremendous beauty beneath. Catalina Island has been a nature reserve since 1974, with huge rewards for the diver.

PROTECTED ISLAND

Catalina Island is 20 miles (32 km) in length and 8 miles (13 km) wide, and has been owned by the Wrigley family since 1919. Avid conservationists, they have consistently refused permission for development on the island, and now 86 percent of it is an established nature reserve. The rugged topography of the island extends below the waterline, with tremendous diving on craggy reefs and steep underwater pinnacles.

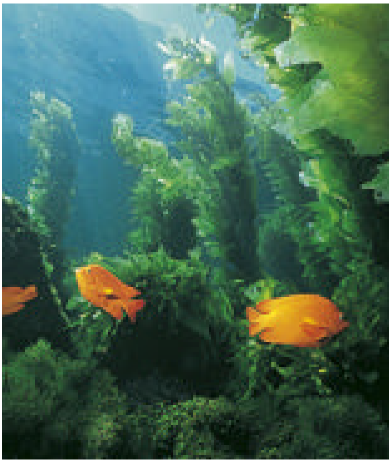

FOREST BENEATH THE WAVES
A rich variety of marine life is supported by the giant kelp stands of Catalina Island. Garibaldi fish are familiar residents of the kelp.

RICH SEAS

The island is swept by the cold waters of the California Current as it moves down from British Colombia. This creates oxygen-rich waters where kelp forests flourish—some of the larger specimens stretch from surface to seabed, 100 ft (30 m) below. Cruising these underwater forests is a wide range of fish species, the most famous of which is the garibaldi. Fiercely territorial, these brightly colored fish will stand their ground as divers approach—a photographer's dream. The kelp forests are also home to huge black sea bass, eagle rays, California morays, and two-spot octopuses. The black sea bass is a particularly welcome sight, since it was once heavily overfished but is now making a comeback. Larger residents include the California sea lion, a species known for its curiosity and willingness to approach divers. The occasional blue shark can be encountered off Catalina's shores, rounding off an exhilarating list of animals waiting to be discovered here.

ESSENTIAL INFORMATION

WATER TEMPERATURE 54–68°F (12–20°C)

ECOSYSTEM Temperate

WHEN TO VISIT Diving takes place year-round off Catalina, but winter diving can be very cold.

MUST-SEES The Casino Point Underwater Park; The Farnsworth Banks.

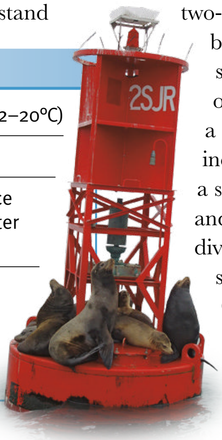

California sea lions

The Outer Banks

NORTH CAROLINA, UNITED STATES

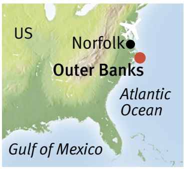

Famed in seafaring lore as the graveyard of ships, the Outer Banks are thought to be the site of up to 5,000 shipwrecks. Big ocean currents and an abundance of large pelagic animals combine to make this region one of the most exciting in the diving world.

HISTORIC WATERS

The Outer Banks are identified by a dramatic chain of islands and sandbars that extend for 130 miles (35 km) up the coast of North Carolina. Within this barrier is a vast area of estuary, shallow lagoons, and enclosed water. This ecological gem is a crucial breeding ground for many fish species, and 70 percent of the region enjoys protected status. The Outer Banks beyond the island chain are the real draw for the diver. With three great ocean currents meeting off this coastline, and the continental shelf only 40 miles (65 km) to the east, this is dramatic diving territory. The Outer Banks are littered with wrecks—many of them sunk by U-boats in World War II. Many of these are home to some of the larger residents of the banks, including numerous sand tiger sharks (also known as ragged-tooth sharks). These fierce-looking animals are in fact quite approachable if not provoked.

ESSENTIAL INFORMATION

WATER TEMPERATURE 54–75°F (12–24°C)

ECOSYSTEM Temperate

WHEN TO VISIT For the most favorable diving conditions, visit the Outer Banks between July and September.

MUST-SEES Wrecks of the *Papoose* and *Proteus*; sand tiger sharks.

FEARSOME APPEARANCE
Despite its looks, the sand tiger shark is relatively docile. However, like all large predators, it should be treated with respect.

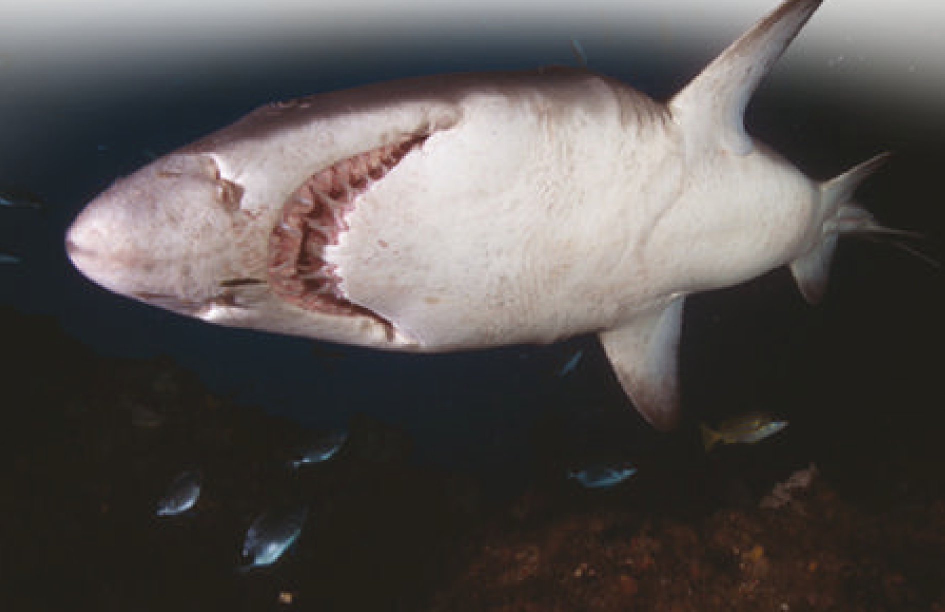

The Great Lakes

CANADA/UNITED STATES

The Great Lakes
CANADA
Toronto
Detroit
Chicago
US

Straddling the border between Canada and the US, the five Great Lakes make up the largest expanse of fresh water on Earth. Conditions in the lakes can be demanding, but the reward is access to some of the best-preserved wrecks in the diving world.

WORLD-CLASS WRECKS

The Great Lakes cover an area of 95,000 sq miles (246,000 sq km), and offer a range of diving experiences and conditions. Many of the more densely populated regions around the lakes have always been significant industrial zones and have suffered through pollution and overexploitation. The water is cold, and visibility can be limited, although in the lower lakes it has improved through the accidental (and in all other respects unwanted) introduction of the zebra mussel—a voracious filter-feeder that has cleaned the lakes of algae. There are many wrecks on the various lake beds. The very low temperatures and the lack of salt in the water have kept metal hulls relatively free of rust, while wooden vessels and even rope can be found in a remarkably good state of preservation.

ESSENTIAL INFORMATION

WATER TEMPERATURE 32–64°F (0–18°C)

ECOSYSTEM Temperate

WHEN TO VISIT Year-round, but northern dives may be under ice in winter.

MUST-SEES Sea caves along Lake Superior coast; the *Arabia* and other well-preserved wrecks.

PRESERVING THE PAST

In the early days of diving the lakes, numerous artifacts were plundered from otherwise pristine wreck sites. This malpractice was quickly stamped out through the establishment of a number of preserves. Lake Michigan alone has nine such special areas, covering 1,900 sq miles (4,920 sq km). Notable wrecks throughout the lakes include the *Arabia*, a sailing vessel sunk in 1884 and still perfectly intact; the *Bermuda*, sunk in 1870; and the more contemporary *Mesquite*, sunk in 1989.

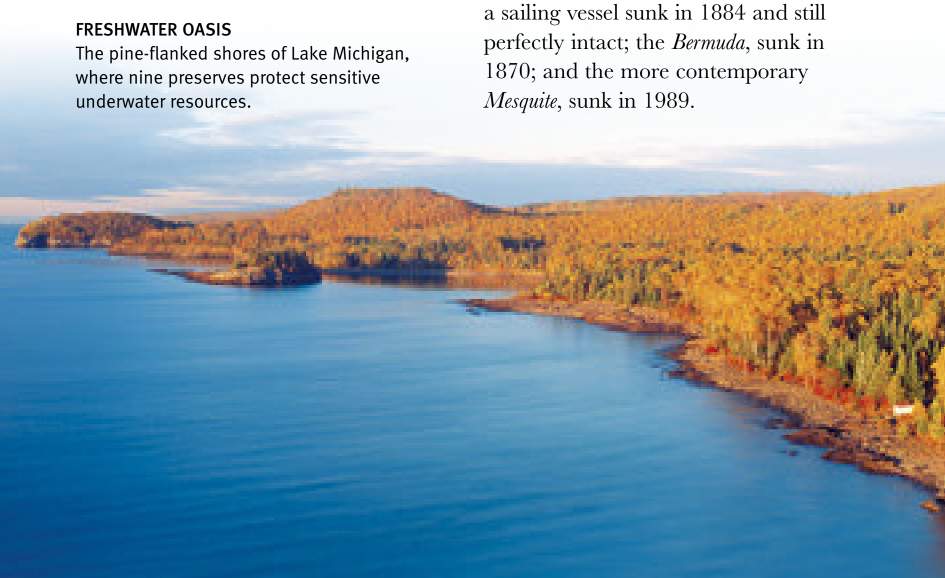

FRESHWATER OASIS
The pine-flanked shores of Lake Michigan, where nine preserves protect sensitive underwater resources.

Vancouver Island

BRITISH COLUMBIA, WESTERN CANADA

CANADA
British Columbia
Vancouver
US

The cold waters and strong tides around Vancouver Island create a glorious mix of wildlife, and there are also vessels sunk as artificial reefs to add extra interest for the visiting diver. Some of the most popular dives lie between the island and the mainland.

CHANNEL CROSSING

The standard means of reaching Vancouver Island is via a ferry from the city itself, a 90-minute journey that gives you a glimpse of the abundant waters of this region. Cold waters are rich in oxygen and nutrients, and the channel between the island and the mainland is no exception. The addition of strong tidal movements creates an irresistible draw for marine plants and animals, and although the diving can be demanding at times—for example, in the fast-running Dodds Narrows channel—there is much to make the experience worthwhile. One of the most famous residents of these waters is the inquisitive giant octopus, although more elusive wolf eels, harbor seals, various whale species (including orcas), and huge numbers of salmon may also be encountered at the right time of year.

ARTIFICIAL REEFS

The vessel *Saskatchewan* was sunk in 1997 as an artificial reef, and has proved a real hit with both local and visiting divers. It is just one of a number of decommissioned warships, all well worth a dive, sunk off the coast by the Artificial Reef Society of British Columbia—a society formed by a group of underwater archaeologists in 1990 to turn warships into dive sites.

EXTRAORDINARY ENCOUNTER
A diver makes contact with a giant octopus, a fascinating species that thrives in the nutrient-rich waters.

ESSENTIAL INFORMATION

WATER TEMPERATURE 48–64°F (9–18°C)

ECOSYSTEM Temperate

WHEN TO VISIT Year-round, but winter can be chilly.

MUST-SEES Giant octopuses; whales; wreck of the *Saskatchewan*; Dodds Narrows channel.

Galápagos Islands

EASTERN PACIFIC

Caribbean Sea
CENTRAL AMERICA
Galápagos Islands
SOUTH AMERICA
Pacific Ocean

The Galápagos were known as the Enchanted Isles by early mariners, baffled by the currents that swirled around them. The name has changed, but the experience remains the same for divers fortunate enough to visit this unique archipelago.

RUGGED AND REMOTE

The Galápagos Islands are named after the Spanish word for "saddle"—*galápago*—after the shape of the shells of the giant tortoises that provided a food source for the early explorers and whaling ships that visited the islands. The island group consists of 13 main islands situated 600 miles (970 km) off the west coast of Ecuador. These stark islands, only a few million years old, still show the contorted rock formations—twisted reefs and pitted ridges—caused by their volcanic birth. These provide an excellent substrate for busy marine communities, replenished by the seven ocean currents that converge on this island group from all directions.

ESSENTIAL INFORMATION

WATER TEMPERATURE 61–81°F (16–27°C)

ECOSYSTEM Temperate

WHEN TO VISIT Diving continues year-round in the Galápagos.

MUST-SEES Hammerheads and white-tips at Wolf and Darwin; whale sharks; sea lions; marine iguanas; penguins.

MARINE LIFE

Although there are only 13 reef-building species of coral in the Galápagos, meaning that the reefs are stark affairs clinging to the dark volcanic rock, the life that swarms over them is spectacular.

The fame of the Galápagos as a dive site lies in its gatherings of large marine animals, including magnificent schools of scalloped hammerheads. Sadly, overfishing has had a serious impact on shark populations in the waters around the islands, and unless circumstances change, such sightings may become increasingly rare. Groups of cow-nosed rays and eagle rays are regularly sighted off the more exposed reefs, as well as gatherings of manta rays and the occasional whale shark.

UNIQUE SPECIES

A visit to the Galápagos offers animal encounters that cannot be experienced elsewhere. The islands have evolved in isolation, and are home to a number of remarkable species. Perhaps the most famous of these is the marine iguana, although the islands also have colonies of the most northerly penguin species on Earth—the tiny Galápagos penguin. It is possible to snorkel with both of these species if accompanied by a knowledgeable local guide and with the right sea conditions.

WILD GALÁPAGOS

Visiting California sea lion

It is not surprising that the Galápagos Islands provided inspiration for Darwin's theory of evolution—they are rich in all kinds of life. There are well over 300 fish species in the region, including 30 species of sharks. Larger residents include nearly 80,000 seals and sea lions. There are also 600 species of mollusks and 100 different types of crabs! Larger visitors include sperm whales, humpbacks, and Bryde's whales, as well as many species of dolphins. The islands are a naturalist's paradise.

BARREN APPEARANCE

The bleak contours of the Galápagos Islands might seem to suggest a region devoid of life. The appearance is deceptive, however—life thrives here both above and below the green waters of the surrounding Pacific Ocean.

Useful resources

PADI: DIVE SAFETY CHECKLIST

- Divers should limit their depths according to their training, certification level, and experience.
- Regardless of level of training, any increase in diving depth should be incremental.
- Further training and/or supervision is needed when diving in a new type of environment.
- Avoid overweighting; carry out a buoyancy check whenever using new equipment. Remember to reduce weighting when going from salt to fresh water.
- Ensure that all weight systems used have a quick release and that each buddy is familiar with its operation.
- Always review emergency procedures prior to each dive.
- Have regular medical checkups and stay in shape for diving.
- If surface conditions, currents, or visibility deteriorate, consider aborting a dive.
- Consult tide tables where appropriate and seek local advice on prevailing diving conditions.
- Use surface marker buoys and ascent/descent lines wherever appropriate.
- Always inform a non-diver where you are going and when you expect to return.

Visit www.padi.com

PROJECT AWARE FOUNDATION: TEN WAYS A DIVER CAN PROTECT THE UNDERWATER ENVIRONMENT

Divers and snorkelers are obvious ambassadors for the underwater environment. Let's make sure we:

1. Dive carefully in fragile aquatic environments
Although at first they may look like rocks or plants, many aquatic organisms are fragile creatures that can be damaged or harmed by the bump of a tank, knee, or camera, a swipe of a fin, or even the touch of a hand. It is also important to know that some aquatic organisms, such as corals, are extremely slow-growing. By breaking off even a small piece, you may be destroying decades of growth. By being careful, you can prevent devastating and long-lasting damage to magnificent dive sites.

2. Streamline yourself
Much damage to the environment is done unknowingly. Keep your gauges and alternate air source secured so they don't drag over the reef or bottom. By controlling your buoyancy and taking care not to touch coral or other fragile organisms with your body, diving equipment, or camera, you will have done your part in preventing injury to aquatic life.

3. Continue your education
If you haven't dived in a while, your skills (particularly buoyancy control) may need sharpening. Before heading to the reefs, seek bottom time with a certified assistant or instructor in a pool or other environment that won't be damaged by a few bumps and scrapes. Better yet, take a diving continuing education course such as PADI Scuba Review, the PADI Adventures in Diving program, or a PADI Specialty Diver course. AWARE continuing education diving courses provide you with the skills practice and practical application of environmentally sound diving techniques.

4. Consider your impact on aquatic life through your interactions
Very few forms of aquatic life pose a threat to us. In fact, some creatures even seem friendly and curious about our presence. As we become bolder and more curious ourselves, we may even feel compelled to touch, handle, feed and even hitch rides on certain aquatic life. However, our actions may cause stress to the animal, interrupt feeding and mating behavior, introduce food items that are not healthy for the species, or even provoke aggressive behavior in normally nonaggressive species. Interact responsibly with the aquatic environment.

5. Understand and respect underwater life; resist the urge to collect souvenirs
Through adaptation to an aquatic environment, underwater life often differs greatly in appearance from life we are used to seeing on land. Many creatures only appear to look like plants or inanimate objects. Using them as "toys" or food for other animals can leave a trail of destruction that can disrupt a local ecosystem and rob other divers of the pleasure of observing or photographing these creatures. Consider taking part in a Project AWARE program to become more familiar with the importance of, and the interdependent nature of, worldwide aquatic

ecosystems. Dive sites that are heavily visited can be depleted of their resources in a short time. Collecting specimens, coral, and shells in these areas can strip their fascination and beauty. If you want to return from your dives with trophies to show friends and family, you may want to consider underwater photography instead.

6. Be an ECO-Tourist
Protected areas, such as parks, reserves, and sanctuaries, are one of the best tools for conserving the aquatic environment. Support the creation of protected areas, follow all local laws, and learn to appreciate that all aquatic habitats (such as grass beds, mangroves, and rubble zones) are important and interesting environments. Do not confine your diving only to sites highlighted in brochures and articles. When planning a diving trip, choose ECO Tour Operators involved with ECO-resorts and ECO-operators. Make informed decisions when selecting a destination, and support the Project AWARE Go ECO environmental campaign.

7. Respect the underwater cultural heritage
Divers and snorkelers have the privilege to access dive sites that are part of our cultural heritage or maritime history. You should help preserve these sites for future generations by obeying local laws, diving responsibly, and treating wrecks with respect. Wrecks can serve as important habitats for fish and other aquatic life.

8. Report environmental disturbances or destruction of your dive sites
Divers and snorkelers are in a unique position to monitor the health of local waterways, lakes, and coastal areas. If you observe an unusual depletion of aquatic life, a rash of injuries to aquatic animals, or strange substances or objects in the water, report them to your department of natural resources or similar entity.

9. Be a role model
As a diver or snorkeler, you realize that when someone throws a plastic wrapper or other debris overboard, it is not out of sight, out of mind. You see the results of such neglect. Set a good example in your own interactions with the environment, and other divers and non-divers will follow suit.

10. Get Involved
There are plenty of opportunities to show your support of a clean aquatic environment, including local beach cleanups, surveys, and attending public hearings on matters that impact local coastal areas and water resources. Divers' skills in particular are always needed and appreciated by many environmental organisations.

Visit www.projectaware.org for more details

Copyright Project AWARE Foundation 2004. All rights reserved.
Used with permission from Project AWARE Foundation.

USEFUL CONTACTS

Training organizations

International groups offering both basic and specialized dive instruction, including training organizations based on an amateur footing.

Professional Association of Dive Instructors (PADI)
www.padi.com

PADI Americas: 30151 Tomas Street, Rancho Santa Margarita, CA 92688-2125

PADI Europe: Oberwilerstrasse 3 CH-8442, Hettlingen, Switzerland

PADI Canada: 107 – 1680 Broadway Street, Port Coquitlam, BC V3C 2M8

National Association of Underwater Instructors (NAUI)
Worldwide headquarters: PO Box 89789,
Tampa, FL 33689-0413
www.naui.org

British Sub Aqua Club (BSAC)
Telford's Quay, South Pier Road,
Ellesmere Port, Cheshire, CH65 4FL, UK
www.bsac.com

Sub Aqua Association
Space Solutions Business Centre, Sefton Lane,
Maghull, Liverpool, L31 8BX, UK
www.saa.org.uk

Scottish Sub Aqua Club
The Cockburn Centre, 40 Bogmore Place,
Glasgow, G51 4TQ, UK
www.scotsac.com

Scuba Schools International
2619 Canton Court, Fort Collins,
CO 80525-4498
www.ssiusa.com

World Underwater Federation
Viale Tiziano, 74 00196, Roma, Italia
www.cmas2000.org

Technical Diving International
18 Elm Street, Topsham, Maine, 04086
www.tdisdi.com

International Association of Nitrox and Technical Divers
5609 Power Road, Ottawa, Ontario,
K1G 3N4
www.iantd.com

USEFUL CONTACTS (CONTINUED)

National Association for Cave Diving (NACD)
P.O. Box 14492, Gainesville, FL 32604
www.safecavediving.com

ACUC International (American Canadian Underwater Certifications)
379 West Street, Brantford, ON
N3R 3V9
www.acuc.ca

Governing bodies

Organizations promoting and regulating recreational diving activities.

World Underwater Federation
Viale Tiziano, 74 00196, Roma, Italy
www.cmas2000.org

International Association for the Development of Freediving (AIDA)
Rue du Petit-Beaulieu 4
CH 1004 Lausanne, Switzerland
www.aida-international.org

Freediving Regulations and Education Entity (FREE)
7480 NW 175 St., Miami, FL 33015
www.divingfree.com

International Association for Handicapped Divers
http://www.iahd.org

Conservation and heritage organizations

Charitable and nonprofit organizations promoting marine wildlife conservation, and groups involved in preserving maritime heritage.

Coral Cay Conservation
40-42 Osnaburgh Street, London,
NW1 3ND, UK
www.coralcay.org

World Wildlife Fund
1250 Twenty-Fourth Street, N.W.
P.O. Box 97180, Washington, DC
20090-7180
www.worldwildlife.org

Marine Conservation Society
Unit 3, Wolf Business Park, Alton Road,
Ross-on-Wye, Herefordshire, HR9 5NB, UK
www.mcsuk.org

British Divers Marine Life Rescue
Lime House, Regency Close Uckfield,
East Sussex, TN22 1DS, UK
www.bdmlr.org.uk/donate

Marine Fish Conservation Network
600 Pennsylvania Avenue, SE, Suite 210,
Washington, DC 20003
www.conservefish.org/site

Nautical Archaeology Society
The Nautical Archaeology Society,
Fort Cumberland, Fort Cumberland Road,
Portsmouth, PO4 9LD, UK
www.nasportsmouth.org.uk

Whale and Dolphin Conservation Society
Brookfield House, 38 St Paul Street,
Chippenham, Wiltshire, SN15 1LJ, UK

The Shark Trust
National Marine Aquarium, The Rope Walk
Coxside, Plymouth, PL4 OLF, UK
www.sharktrust.org/sharkconservation.html

Parks Canada (National Marine Conservation Areas of Canada)
25 Eddy Street, Gatineau, QC
K1A 0M5

Historical Diving Society
Historical Diving Society USA , PO Box 2837,
Santa Maria, CA 93457
www.hds.org

Historical Diving Society UK
www.thehds.com

Maritime Archaeological and Historical Society
P.O. Box 44382, L'Enfant Plaza, Washington,
DC 20026
www.mahsnet.org

Diving medicine and safety organizations

Expert advisory groups on diving medicine and safety issues.

Divers Alert Network (DAN)
The Peter B. Bennett Center, 6 West Colony
Place, Durham, NC 27705
www.diversalertnetwork.org

British Diving Safety Group
West Quay Road, Poole, BH15 1HZ, UK
www.bdsg.org

London Diving Chamber
Hospital of St John and St Elizabeth,
60 Grove End Road, St John's Wood,
London, NW8 9NH, UK

Glossary

ABSOLUTE PRESSURE The total *pressure* acting on a body, equal to the sum of *atmospheric pressure* and *gauge pressure*.

A-CLAMP FITTING A *regulator first stage* that attaches to the *pillar valve* of a *tank* by means of a yoke that clamps it securely in position.

AIR CONSUMPTION The depletion of oxygen in air through respiration; the rate at which air is used by a diver during the course of a dive.

ALTERNATE AIR SOURCE A redundant source of air supply, in the form of either an *octopus second stage*, or a completely independent *tank* and *regulator*.

AMBIENT PRESSURE The total *pressure* acting on a body at a given depth. See also *absolute pressure*.

ANCHOR LINE A rope, cable, or chain that attaches a ship to its anchor.

ANOXIA A medical condition caused by a lack, or severe deficiency of, oxygen in the human body.

AQUA LUNG The first underwater breathing apparatus to use compressed air and a two-stage *regulator*, designed by Emile Gagnan and Jacques Cousteau.

ARCHIPELAGO A group or chain of islands.

ARTIFICIAL REEF A human-made object deliberately scuttled to create a haven for marine life, and for the enjoyment of divers.

ASCENT Returning from depth to the surface at the end of a dive. This must always be performed carefully to avoid *decompression sickness*.

ATMOSPHERIC PRESSURE The *pressure* exerted by the gases in the atmosphere. At sea level, this is equal to 14.7 psi (1 bar).

ATOLL Circular *coral reef*, often surrounding a lagoon, that has formed around a submerged extinct volcano.

BALLAST Lead weights carried to offset the inherent *buoyancy* of the diver's *exposure suit*.

BENDS See *decompression sickness*

BEZEL A rotatable, notched, or marked outer ring on instruments, such as watches and compasses, used to mark key datum points, such as a bearing.

BLOCK SHIP A vessel sunk deliberately across the approaches to a naval harbor to bar entry to hostile submarines and surface ships.

BOMMIE An isolated outcrop of coral growth on the seabed.

BOYLE'S LAW Physical law stating that at a constant temperature, the volume of a gas is inversely proportional to the *pressure* exerted on it.

BRACKISH Water that is not as salty as seawater, but has a higher salt content than fresh water.

BREATH-HOLD DIVING See *freediving*

BUBBLE CHECK A precautionary visual inspection of another diver's gear (conducted underwater) to determine if it has any air leaks.

BUDDY A diver who assumes an informal duty of care over the other member of their assigned "buddy pair" during diving.

BUDDY LINE A safety line for tethering a *buddy* pair together.

BUOYANCY An upward thrust exerted on an immersed object, which is equal to the weight of the water that has been displaced by the object.

BUOYANCY COMPENSATOR (BC) A jacket that can be inflated and deflated to allow the diver to control their *buoyancy*.

BUOYANCY CONTROL Manipulation of *buoyancy* through the use of devices such as a *BC*, and the breathing cycle.

CAISSON A pressurized, watertight chamber used to undertake construction work underwater.

CERTIFICATION CARD Accreditation proving that a diver has achieved a certain level of dive training with one of the recognized agencies.

CORAL HEAD Protrusions on a *coral reef* formed by a colony of living coral polyps feeding and growing.

CORAL ISLAND Portion of a *coral reef* that is permanently out of water, its surface usually eroded to a flat top of white coral rock and sand. May occur in a chain along a reef.

CORAL REEF A massive, marine, rocklike ridge or outcrop created by the gradual accretion of the skeletons of generations of coral polyps.

DATUM LINE A line used as a fixed reference to aid in an accurate survey of an area.

DECOMPRESSION Returning to conditions of normal *atmospheric pressure* by controlled means.

DECOMPRESSION SCHEDULE A plan describing the number, duration, and depth of *decompression stops* required for a given dive.

DECOMPRESSION SICKNESS (DCS) A potentially dangerous medical condition resulting from the formation of bubbles of nitrogen in the bloodstream and tissues of a diver's body. Also known as decompression injury (DCI) and the bends.

DECOMPRESSION STOP A scheduled pause in a diver's ascent to allow nitrogen to pass from body tissues back into the blood at a safe rate.

DEHYDRATION A medical condition resulting from excessive loss of water from the body.

DELAYED SURFACE MARKER BUOY (DSMB) Marker buoy deployed just before ascent. It is inflated underwater, then released to indicate where a diver will surface.

DEMAND VALVE See *second stage*

DEPTH GAUGE An analog or digital gauge that measures *ambient pressure*, and uses this to give a reading of depth.

DESCENT Traveling from the surface to depth.

DIN FITTING A screw-thread fitting for attaching a *regulator first stage* to a compatible *pillar valve*. Allows higher working pressures than *A-clamp* fittings.

DISPLACEMENT The displacement of water that occurs when an object is submerged in it.

DISTANCE LINE A line fastened to a submerged feature and deployed from a *reel* as a navigational guide, especially in areas of low visibility.

DIVE CENTER A commercially run diving school and rental center, frequently offering organized trips and guides to local *dive sites*.

DIVE COMPUTER A digital device that provides the diver with a range of information, such as depth, time, and ascent rate.

DIVE CONSOLE An instrument panel incorporating a *tank contents gauge*, *depth gauge*, and sometimes also a compass, or a *dive computer*.

DIVE GUIDE An employee of a *dive center* or school who guides qualified divers around local *dive sites*.

DIVE LEADER The diver in a *buddy* pair who is designated (by agreement) as the principle pace-setter and decision-maker underwater.

DIVE MARSHAL A senior diver whose role on a dive trip is to allocate *buddy* pairs, approve *dive plans*, ensure that complete records of dives are kept, and initiate and direct rescue efforts in the event of a diving accident.

DIVE PLAN The outline of a proposed dive, including its depth, duration, intended goals, entry and exit points, signaling conventions, and contingency plans.

DIVE SITE Any location where diving is conducted.

DIVE SLATE A plastic tablet used for making notes underwater.

DIVE TIME The duration of a dive, starting at the moment the diver leaves the surface and ending with their return to the surface.

DIVEMASTER A person qualified to act as a *dive guide*, assist diving instructors, and oversee the organization of dives for paying customers.

DIVING CLUB An organization offering diving instruction and social events on an amateur basis. It can also refer to continental European diving centers offering instruction on a commercial basis.

DIVING SCHOOL A center for instruction in diving skills (usually on a commercial basis).

DPV (DIVER PROPULSION VEHICLE) A propulsion device either ridden or clung to by a diver to minimize swimming effort or aid in the exploration of large *dive sites*.

DRAG Frictional resistance to the motion of an object through a fluid.

DRIFT DIVE A dive in which the diver is propelled along a route by undersea currents.

D-RING A D-shaped ring of metal or plastic used as an anchor point for tethering equipment to your *BC* during a dive.

DROP-OFF A falling-away of the ground or seabed, such as a *reef wall* or rock shelf that descends steeply.

DRYSUIT An *exposure suit* designed for use in very cold waters, in which trapped air is the principle insulating medium, and that offers a high degree of thermal protection.

DSMB (DELAYED SURFACE MARKER BUOY) A tubular buoy designed to be deployed at the end of a dive to aid *ascent* and alert observers at the surface.

DUMP VALVE One-way valve used to release air from an inflatable device.

DYE MARKERS A fluorescent dye designed to be released into the water to alert rescue craft to the position of a drifting diver.

ECOSYSTEM The interactions between a community of living organisms and the environment they inhabit.

EL NIÑO A warming of the eastern tropical Pacific occurring every few years that has a pronounced effect on both the local and wider climate. Often occurs around Christmas time—"*El Niño*" in Spanish refers to the Christ child.

EQUALIZATION The process of “clearing” the ears and sinuses until the pressure within them matches the pressure of the water surrounding the body. See also *Valsalva maneuver*.

EXPOSURE SUIT Any suit worn by a diver in order to prevent excessive loss of body heat. See *wetsuits*, *semi-dry suits*, and *drysuits*.

FILTER-FEEDER An aquatic animal that feeds by straining minute organisms and other suspended food particles from the water.

FINNING Kicking the feet rhythmically while wearing fins to achieve forward motion.

FIRST STAGE A major component of the *regulator* that reduces the *pressure* of air from the tank to 8–10 bar above *ambient pressure*, then allows it to pass to the *second stage* for inhalation by the diver. It also supplies air for inflatable items such as the *BC*, and to *pressure gauges* displaying the remaining contents of the *tank*.

FOOD CHAIN A sequence of organisms, each depending on the next as a source of food.

FOOT POCKET A style of fin shoe that covers the whole foot.

FREE ASCENT An *ascent* that is carried out without the use of a line or any other point of reference.

FREEDIVING A form of diving in which divers do not carry an air supply but remain underwater only as long as they can hold their breath.

GAUGE PRESSURE The surrounding water pressure, as measured by a gauge that does not take *atmospheric pressure* into account.

GLOBAL POSITIONING SYSTEM (GPS) A navigation system that uses satellite signals to determine an exact longitude and latitude position.

GUIDELINE A line with a weight at one end and a buoy at the other, chiefly used to mark the position of wrecks and provide a guide for descending and ascending divers.

HELIOX A mixture of helium and oxygen, used for diving beyond *recreational diving* limits.

HYDROSTATIC TESTING A method of testing a *tank* for serviceability by filling it with pressurized water.

HYPERVENTILATION A medical condition caused by rapid or deep breathing, resulting in lowered carbon dioxide levels in the blood.

HYPOXIA A medical condition caused by a deficiency of oxygen in the human body.

INFLATOR VALVE A manually operated valve that admits compressed air into the buoyancy bladders of a *BC*, or into a *drysuit*.

KELP FOREST A marine *ecosystem* based around dense growths of seaweed belonging to the kelp family.

LIVEABOARD A recreational diving vessel equipped to conduct extended diving trips without returning to port.

LOG BOOK A record containing the details of individual dives made by a diver.

LONGSHORE CURRENT A current running parallel to the shoreline.

MANGROVE Assorted coastal tree and shrub species with aerial roots that grow in dense thickets along tidal shores in the tropics.

MARINE PARK An area of sea or coastline in which marine species are protected by government legislation.

MASK FOGGING Misting caused by the accumulation of oils on the lens (or lenses) of a diving mask, permitting moisture to condense there more readily.

MASK SQUEEZE Discomfort caused by increasing water *pressure* acting on the diver’s mask, pressing its frame into the face.

MUCK DIVING Informal term for diving in areas where the bottom sediments are dark and coarse but where small marine species proliferate (such as Sulawesi’s Lembeh Straits).

NEAP TIDE The point in the monthly cycle of tides where the gravitational influences of the Sun and Moon are least aligned, resulting in the lowest high tides and the highest low tides.

NEOPRENE A synthetic rubber fabric with good heat insulation properties used to make *wetsuits*.

NITROGEN NARCOSIS Intoxication caused by breathing nitrogen at elevated *partial pressures* (such as during deep dives), resulting in the impairment of reasoning and motor skills, feelings of irrational anxiety or elation, and perceptual narrowing.

NITROX Any mixture of nitrogen and oxygen prepared for use as a breathing gas, but normally one containing a greater percentage of oxygen than that found in air.

OCTOPUS SECOND STAGE A redundant *second stage*, carried by divers as a contingency against failure of the main second stage, and to enable air-sharing with a diver whose own air supply has been depleted.

OFF-GASSING The gradual release of gases (especially nitrogen) that have accumulated in a diver’s body during the course of a dive. Off-gassing occurs as *ambient pressure* decreases during *ascent* and at the surface.

OPEN WATER A body of water that is not sheltered from the influence of naturally developing weather conditions and currents, and whose depth may exceed safe training limits.

O-RING A ring-shaped gasket, used to seal interfaces between one pressurized piece of equipment and another, such as in the *pillar valve* of a *tank*.

OVERHANG A projection from the face of a reef, cliff, etc.

OVERHEAD ENVIRONMENT Any diving environment where direct access to the surface is restricted, such as inside a cave or under ice.

OXYGEN TOXICITY Acute nervous system dysfunction caused by breathing oxygen at elevated pressures. Can cause loss of consciousness and convulsions.

PARTIAL PRESSURE The *pressure* of a single gas in a mixture of gases, measured as if that gas alone were present (the total pressure of a gas mixture being the sum of the *partial pressures* of its constituent gases).

PELAGIC ANIMALS Creatures that dwell in the *open ocean*, rather than coastal waters.

PILLAR VALVE The valve that transfers gas from the diver's *tank* to the *regulator first stage*.

PLANKTON Minute free-floating plant and animal organisms inhabiting seas and lakes, which form the basis of the *food chain* in many environments.

PLB (personal locator beacon) A submersible device that emits a radio signal on an internationally recognized search-and-rescue homing frequency, enabling emergency services to pinpoint its location.

POLAR SEAS The cold seas at the polar extremities of the planet, which vary in temperature from 29°F to 50°F (–1.8°C to 10°C).

PONY BOTTLE A small *tank* of breathing gas with an independent *regulator*, intended for emergency use if the diver's main breathing apparatus fails.

PRESSURE The exertion of force by one surface, substance, or gas onto another.

REBREATHER A breathing apparatus that recycles exhaled gas by filtering out carbon dioxide and replacing any oxygen metabolized by the diver.

RECOMPRESSION CHAMBER A pressure chamber used in the treatment of divers suffering from *decompression sickness*.

RECREATIONAL DIVING Sports diving that takes place within certain depth and environmental limitations, as defined by the various training agencies.

REEF WALL A vertical face of coral growth, often descending into deep waters.

REEL A spooling device used to store and deploy line.

REFRACTION The change in the velocity of light (or any propagating wave) that occurs when it passes from one medium to another (e.g., from air to water).

REGULATOR The mechanism that supplies breathing gas from the *tank* to the diver via a two-stage process of *pressure* reduction. The regulator is made up of two parts: the *first stage* and the *second stage*.

RIB (rigid inflatable boat) A fast utility boat characterized by a V-shaped hull with an inflatable tube running around its upper edge.

RIP CURRENT A fast-moving current running perpendicular to the shore, caused by the action of incoming waves.

RULE OF THIRDS, THE A rule of thumb used when calculating air requirements for a dive, especially one requiring a conservative approach, such as a cavern exploration.

SAFETY SAUSAGE See *SMB*

SAFETY STOP A precautionary pause in a diver's *ascent* to prevent the formation of nitrogen bubbles in the bloodstream. Its purpose is as a failsafe against *decompression sickness* even when no scheduled *decompression stops* need to be made.

SALINITY The salt content of a measure of water.

SCUBA Acronym for "self-contained underwater breathing apparatus."

SECOND STAGE The part of the *regulator* that regulates the *pressure* of a diver's gas supply to the same pressure as its immediate environment.

SEMI-DRY SUIT A thick *neoprene exposure suit*, similar to a *wetsuit*, designed for use in cold waters.

SHELTERED WATER A shallow body of water protected from currents and weather conditions, used for training divers in a controlled environment.

SILTING A rapid reduction in underwater visibility that occurs when fine sedimentary particles are stirred up.

SKINSUIT A thin *exposure suit*, typically made of Lycra, designed for use in tropical waters.

SLACK WATER A period of reduced water movement that occurs between incoming and outgoing tides.

SMB (surface marker buoy) An inflatable buoy designed to be towed behind a diver to indicate their position to observers on the surface. Long, cylinder-shaped SMBs are known as safety sausages.

SNORKELING Swimming face-down at the surface using a short, curved tube to draw breath.

SPG (SUBMERSIBLE PRESSURE GAUGE) See *tank contents gauge*

SPRING TIDE The point in the monthly cycle of tides when the highest high tides, and lowest low tides occur. Spring tides occur when the gravitational forces of the Sun and the Moon are aligned.

SPUR AND GROOVE REEF FORMATION A typical growth structure of coral reefs, in which outcrops of coral are separated by linear depressions.

STANDING CURRENT A continuous directional movement of water generated by prevailing weather, sea, and climatic conditions.

STROBE A beacon that emits high-intensity flashes of light at short intervals.

SUBTROPICAL SEAS Warm seas within a loosely defined zone between *temperate* and *tropical* climatic conditions.

SWELL The undulating motion of surface waters in areas of open sea.

SWIM-THROUGH A submerged arch, or short passage of rock, where direct access to the surface is restricted, but in which the exit is clearly visible at all times.

TANK A steel or aluminum container designed to hold compressed gas.

TANK CONTENTS GAUGE An instrument that displays the remaining quantity of air or other breathing gas in a diver's *tank*. Also known as a submersible pressure gauge (SPG) or pressure gauge.

TECHNICAL DIVING The use of special breathing-gas mixtures, such as *Nitrox*, *Trimix*, and *Heliox*, to extend depth limits and dive durations beyond *recreational diving* limits.

TEMPERATE SEAS Seas of an intermediate temperature between *tropical* and *polar*, typically 50–68°F (10–20°C).

TEST DATE The date when a compressed gas *tank* was last subjected to *hydrostatic testing*.

THERMOCLINE A layer in a body of water in which temperature changes rapidly with depth. This is usually between a warm surface layer and colder deep water.

TIDAL CURRENT A current generated by the action of the tides.

TIDAL CYCLE The cyclic rise and fall of shoreline sea level, caused by the gravitational influence of the Moon and Sun.

TOPOGRAPHY The shape or form of a region of land or seabed.

TRIMIX A mixture of oxygen, helium, and nitrogen commonly used as a breathing gas for dives deeper than 165 ft (50 m).

TROPICAL SEAS The seas found within the boundaries of the 68°F (20°C) isotherm—a band of warm equatorial water.

VALSALVA MANEUVER A method of manually increasing *pressure* on the inside of the eardrum by closing the nasal passage and exerting expiratory effort, forcing air through the eustachian tubes into the inner ear.

WATER COLUMN The zone of water between the surface and the bottom of a lake, sea, or river.

WEIGHT SYSTEM The means by which a diver's lead *ballast* is secured.

WETSUIT An *exposure suit* made of *neoprene* rubber, which traps a thin layer of water against the skin as an insulating medium.

WICKING The absorption of moisture into a fabric by capillary action.

WORKING PRESSURE The normal *pressure* at which an equipment item is designed to operate.

WORLD HERITAGE SITE An area of natural or cultural importance protected by a United Nations treaty.

WRECK Remains of ships, aircraft, or other vehicles found on the seabed.

INDEX

A

- absolute pressure 94
- accessories 80–81
- Africa 256–65
- aims of dive 160–61
- air, composition 96
- air tanks *see* tanks
- air dumps, buoyancy compensators 56, 57
- air-horns 76–77
- air leaks 114
- air-lock tank bands 81
- air pressure
  - equalizing 98–99, 114–15
  - gauges 64, 65
  - pressure tests 162
  - regulators 60–61
  - tanks 68
- air-pumps 24
- air supply
  - air-sharing procedures 62, 140
  - breathing compressed gas 96–97
  - calculating consumption 96, 117, 168–69
  - problem-solving 138, 140
  - rebreathers 26, 29, 82–83
  - *see also* tanks
- air travel 239
- alcohol 155
- Alexander the Great 22
- Aliwal Shoal 206, 259
- Allen keys 88
- aluminum tanks 68
- ambient pressure, regulators 60
- Americas 306–27
- ammunition, wreck diving 170
- Andaman Sea
  - Andaman Islands 275
  - Similan Islands 274
- ankle weights 71, 123
- Antarctic 212–13
- Aqua Lungs 27–28
- archaeology 192–93
  - wreck diving 166, 169–70
- Arctic 212–13
- argon, suit inflation tanks 69
- Aristotle 22
- arm exercises 152–53
- arthropods 230
- artificial reefs 19, 325
- artificial ventilation (AV) 141
- ascent 128–31
  - buoyancy control 119–20
  - controlled buoyant ascents 140–41
  - decompression dives 128, 130
  - decompression stops 101, 129, 130
- ascent *cont'd.*
  - dive computers 66, 67
  - equalizing pressure 98–99
  - free ascents 128, 140
  - nitrogen in bloodstream 100
  - safe rate of 130
  - water pressure and 95
- Asia 270–81
- asthma 32
- Atlantic Ocean
  - Cuba 314
  - Fernando de Noronha 318
  - Little Bahamas Bank 310
- atmospheric pressure 94
- Australasia 282–93
- Australia
  - Great Barrier Reef 282, 284–85
  - Ningaloo Reef 289
  - Tasmania 288
  - wreck of the *Yongala* 287
- autonomous diving, history 26

B

- back exercises 153
- back injuries 151
- backup flashlights 72, 73
- backward rolls 113
- bags 79, 238
- Bahamas
  - Little Bahamas Bank 310
  - Turks and Caicos 313
- Baikal, Lake 271
- ballast *see* weights
- batteries, flashlights 72, 73, 184
- beaches, exiting on to 132
- beacons, personal locator 76
- Belize Barrier Reef 172, 316–17
- belts, weight 43, 70, 71
- "the bends" *see* decompression sickness
- Bengal, Bay of
  - Andaman Islands 275
  - Similan Islands 274
- Bert, Paul 25
- Bikini Atoll 299
- blood vessels
  - decompression sickness 142
  - nitrogen bubbles 100
- Blue Hole, Belize 317
- boats 109
  - depth indicators 101
  - entering water from 113
  - exiting on to 133
  - liveaboards 237
  - signaling to 132–33
- body
  - effects of water pressure 93, 98–99
  - nitrogen accumulation 100–01
- Bonaire 311
- bony fish 224–27
- boots 43, 48, 50
- Bora Bora 300
- bottles *see* tanks
- Boyle's law 95
- brain, decompression sickness 142
- Breathing Observation Bubble (BOB) 82
- breathing underwater 116–17
  - artificial ventilation 141
  - ascents 130
  - buddy breathing 140
  - buoyancy control 120–21, 146
  - compressed gas 96–97
  - decompression sickness 142
  - freediving 190
  - overcoming panic 141
- briefing
  - buddy 111
  - debriefing 135
  - pre-dive 165
- Bristol Channel, Lundy 252
- Britain
  - Hebrides 246–47
  - Isles of Scilly 250
  - Lundy 252
  - Scapa Flow 248–49
- British Columbia 325
- The Brothers 266
- BSAC (British Sub-Aqua Club) 35
- bubble check 114, 115
- buddies 136
  - air-sharing procedures 62, 140
  - ascents 128, 129
  - attracting attention 81
  - briefing 111
  - buddy breathing 140
  - buddy check 110
  - descents 114–15
  - diving with new buddies 108
  - emergency lifts 141
  - entering water 113
  - exchanging masks with 147
  - hand signals 144–5
  - ice diving 188
  - losing sight of 139
  - night diving 185
- buoyancy
  - breathing and 97
  - buoyancy control 118–21
  - cavern diving 186–87
  - developing skills 146
  - drysuits 54
  - emergency lifts 141
  - flashlights 73
  - freshwater environments 216
  - neutral buoyancy 97, 116, 118, 119
  - swimming profile 123
  - and weight loss or gain 150

buoyancy compensators (BCs) 28, 43, 56–59
  ascents 129
  buddy check 110
  buoyancy control 118–20
  caring for 58
  checking for damage 85
  controlled buoyant ascents 140–41
  designs 59
  disassembly 87
  fixings and fastenings 57
  history 27
  inflating and deflating 56
  integrated safety devices 77
  problem-solving 139
  removing and replacing 148–49
  servicing 58
  signaling devices 81
  technical diving 195
  water pressure and 95
  weight distribution 121
  weight pockets 70
buoys *see* surface marker buoys
Butler, Steve 27
buzzers 81

C

caissons 24, 25
California 322
California, Gulf of 206–207
cameras 45, 176–83
camouflage, flatfish 226
Canada
  Great Lakes 324
  Vancouver Island 325
carbohydrates 154
carbon dioxide
  closed-circuit systems 26
  and overexertion 97
  panic attacks 150
  rebreathers 82, 83
cardiovascular fitness 150, 151
careers in diving 196–97
Caribbean
  Bonaire 311
  Cayman Islands 312
  Little Bahamas Bank 310
  Turks and Caicos 313
Caroline Islands
  Truk Lagoon 302–303
  Yap 298
Catalina Island 322
Cavalli Islands 292
cave diving 80, 186
cavern diving 72, 186–87
  Blue Hole, Belize 317
  Cenotes 315
  Sardina 244
Cayman Islands 312
Celebes Sea 276–77
Cenotes 315
Central America 316–17
cephalopods 230–32
CFT/IUC (Irish Underwater Council) 35
children, learning to dive 38
clips 81
closed-circuit rebreathers (CCRs) 83
closed-circuit systems 26
CMAS (Confédération Mondial des Activités Subaquatiques) 35
cnidarians 233
Cocos 206, 319
cold-water diving
  equipment 44–45
  hypothermia 45, 93
  ice diving 188–89
  Norwegian fjords 253
  regulators 63
colds 99, 154
colors, underwater 92, 183
comb jellies 233
commercial diving 196–97
communication *see* signals
compasses 43, 65, 81, 125, 126–127
compressed air *see* air pressure; air supply
computers *see* dive computers
confidence-building exercises 147
conservation diving 172–75
consoles 64–65
  with dive computers 66
  regulators 61
  types of 65
contact lenses 47
controlled buoyant ascents 140–41
Coral Cay Conservation 172
coral reefs *see* reefs
Coral Sea 286
corals 233
Corlieu, Louis de 27
corrective lenses, masks 47
Cortez, Sea of 206–207
Cousteau, Jacques 15, 27, 28, 176, 291
crocodiles 229
crushed neoprene drysuits 54
crustaceans 230
Crystal River 309
Cuba 314
currents 102–03
  descents 119
  drift dives 74
  guidelines 129
  navigational aids 125
  safety 117
Cyprus 242

D

Dahab 267
damage, checking gear for 84–85
day bags 79
Dean, John and Charles 24–25
debriefing 135
decompression
  ascents 128, 130
  decompression stops 101, 129, 130
  high-altitude diving 94, 216
  technical diving 195
decompression sickness (DCS) 24, 25, 142–43
  deep decompression stops 66
  dive computers 66, 67
  overweight divers 150
decongestants 99, 154
dehydration 154, 155
delayed surface marker buoys (DSMB) 44, 74–75, 76
  ascents without a line 128–129
  deploying 130–31
  reels 75
  types of 75
demand valves *see* second stages
Denarouze, Auguste 26, 27
depths
  depth gauges 64, 65
  depth indicators 101
  distance lines 75
  dive computers 66
  dive tables 66–67
  freediving 29, 191
  mixed-gas diving 29
  rebreathers 83
  technical diving 194, 195
  using compressed air 29
  and water pressure 94–95
descents 114–15
  buoyancy control 118–19
  equalizing pressure 98–99
  water pressure and 95
diaphragms, regulators 61, 63
diet 154, 155
disabled divers 33, 38–39
disassembling gear 86–87
distance, measuring 126
distance lines 75, 127
distress flares 76
dive computers 28–29, 43, 53, 66–7
  ascents 128
  avoiding decompression sickness 101
  choosing 67
  consoles 65
  night diving 184
  technical diving 195
  types of 66

dive operators 236–37, 238, 239
dive tables 66–67, 101
divemasters 236–37
Diver Propulsion Devices (DPVs) 29, 80, 82
diving barrels 24
diving bells 23
diving clubs 34
diving environment 91–103
diving helmets 24–25
diving schools 34–35
diving suits 43, 50–55
  exposure protection 50
  history 22–23, 24–25
  layers 51
dizziness 99
dolphins 215, 218, 219, 310
down signal 114
Draeger company 26
drift dives, surface marker buoys 74
drinks 135, 154
dry bags 79
drysuits 45, 50, 54–55
  ankle weights 71
  buoyancy control 118, 120
  fastenings and seals 55
  inflating with air 55
  layers 51
dugongs 218, 219
dust caps, regulators 87
Dutch Antilles 311
Dutch East India Company 24
dye markers 77

E

ears
  equalizing pressure 98–99, 114–15
  infections 99, 155
  pressure underwater 93
  surfer's ear 22
echinoderms 232–33
ecosystems 201
Egypt
  The Brothers 266
  Dahab 267
  wreck of the *Thistlegorm* 268–69
emergency lifts 141
emergency signals 132, 141
English Channel 250
entry into water 112–13
epilepsy 39
equipment 41–89
  accessories and advanced gear 44
  anticipating problems 137
  bags 79, 238
  buddy check 110
  buoyancy compensation devices 56–59
  cold-water diving 44–45
  confidence-building exercises 147–49
  core equipment 43
  costs 33
  disassembly 86–87
  dive computers 66–67
  diving suits 50–55
  fins 48–49
  flashlights 72–73
  freediving 190
  gadgets and accessories 80–81
  gauges and consoles 64–65
  gear assembly 84–85, 110
  ice diving 189
  knives 78
  maintenance 88–89
  masks 46–47
  new technology 82–83
  packing 163
  post-dive tasks 134
  regulators 60–63
  renting 33, 42, 238–39
  safety and signaling devices 76–77
  snorkels 48
  surface marker buoys 74–75
  technical diving 195
  tanks 68–69
  weight systems 70–71
Europe 240–55
exercise 151–53
expeditions 237
extended vision masks 47
extreme cold-water diving 45
eyes, masks 46–47

F

facemasks *see* masks
fastenings 81
Fernando de Noronha 206, 318
Fiji 304
films 176–77, 180–83
fin pivots 119, 146
fins 27, 48–49
  cavern diving 187
  exiting on to shore 132
  finning technique 122
  freediving 190
  swimming with one fin 147
  types of 49
  wreck diving 168
first stages, regulator 43, 60–61, 63
fish
  bony fish 224–27
  coral reefs 202
  as navigational aids 125
  polar waters 213
  sharks 220–21
  skates and rays 222–23
  wreck diving 166
fitness 32–33, 150–55
fjords, Norwegian 253
flags 76, 77
flares, distress 76
flash, underwater photography 183
flashlights 44, 53, 72–73
  burn time 72–73
  cavern diving 186
  night diving 184, 185
  signaling devices 77
  wreck diving 169
Fleuss, Henry 26
flippers *see* fins
floating *see* buoyancy
Florida
  Crystal River 309
  Florida Keys 308
flutter kick 123
fogging, masks 111
food 135, 154, 155
Force Fins 48, 49
free ascents 128, 140
free descents 115
freediving 29, 190–91
Frenzy, Maurice 27
freshwater environments 216–17
Friendly Islands 305
frog kick 123, 168, 187
frogmen, World War II 83
full-face masks 47

G

gadgets 80–81
Gagnan, Emile 27
Galápagos Islands 326–27
games, confidence-building 148
Gansbaai 258
gases
  effect of water pressure on 95
  see also carbon dioxide; nitrogen; oxygen
gauges 64–65
  cold-water diving 44
  gauge pressure 94
  night diving 184
  types of 65
giant stride entry 113
Gilpatric, Guy 26
glasses, wearing masks 47
gloves 50, 55
  cold-water diving 45
  ice diving 189
  reef gloves 51
Great Barrier Reef 282, 284–85
Great Lakes 324
group dives, guidelines 115
guidelines 115, 129, 168
Gulf Stream 103, 240
gut, pressure underwater 93, 98

H

Haldane, J. S. 25
Halley, Edmund 23
hand signals *see* signals
harnesses 70, 110
Hass, Hans 176
Hawaii 321
hearing, underwater 92–93
heat, underwater 93
Hebrides 246–47
heliox 194, 195
helium, technical diving 195
Herodotus 22
high-altitude lakes 94, 216
high-intensity discharge lights 73
hinged fins 49
history of diving 21–29
hoods 50, 99
  cold-water diving 45, 55, 189
horns, signaling devices 76–77
hoses
  checking for damage 85
  octopus 62
  problem-solving 139
  regulators 89
hydration 154
hydrostatic testing, tanks 68
hydrothermal vents, Iceland 254
hypothermia 45, 93

I

ice diving 188–89
  equipment 45, 189
  polar waters 213
  signals and safety 188–89
  White Sea 255
Iceland 254
iguana, marine 228–29
illness 108, 154
“incident pit” 137
Indian Ocean
  Andaman Islands 275
  Maldives 272–73
  Mauritius 265
  Seychelles 264
  Similan Islands 274
Indonesia 276–77
industrial diving 196–97
inflator valves 54, 55, 84
instructors 34
instrument consoles 43
insulation
  drysuits 51, 54
  semi-dry suits 53
  wetsuits 51, 52
insurance 239
integrated breathing buoyancy compensators 59
integrated weight buoyancy compensators 59
invertebrates 230–33
Ireland 251
Isles of Scilly 250
Italy
  Sardinia 244
  Ustica Island 245

J

jackets *see* buoyancy compensation devices
jellyfish 231, 233
jetboots 80

K

kelp forests 12, 125, 209, 210
kicking strokes 123
knives 43, 78
krill 212

L

ladders, exiting onto 133
lakes 94, 216–17
lamps *see* flashlights
lanyards 72, 81
Larnaca harbor 241, 242
lead weights 70
learning to dive 32–39
  basic training 36–37
  special training courses 38–39
  where to learn 34–35
leg exercises 152
lenses
  camera 182
  mask 47
Lethbridge, John 24
lifting techniques 151
light
  night diving 184
  photography 179, 183
  strobe lights 76, 77
  underwater 92
  see also flashlights
lines
  and ascents 128–29
  cavern diving 187
  decompression stops 129
  delayed surface marker buoys 130–31
  and descents 115
  distance lines 75, 127
  guidelines 115, 129
  ice diving 188
  surface marker buoys 74, 75
  wreck diving 168
Little Bahamas Bank 310
liveaboards 237
lizards 228–29
log books 135
longshore currents 103
Lundy 252
lungs
  decompression sickness 142
  fitness 150
  pressure underwater 98
Lycra skinsuits 51

M

Mackintosh, Charles 24
magnifying glasses 175
maintenance, equipment 88–89
Maldives 272–73
Malpelo 206, 319
mammals 218–19
manatees 218, 219, 309
mangrove swamps 203
mapping sites 124–25
Mariana Trench 12
Marine Protected Areas (MPAs) 172
marine wildlife *see* wildlife
marker buoys *see* surface marker buoys
Marlborough Sound 293
marlin 215
Marshall Islands, Bikini Atoll 299
masks 46–47
  clearing water from 37, 139
  cold-water diving 44
  designs 47
  exchanging with buddy 147
  fogging 111
  freediving 190
  head-up displays 82
  “hide the mask” 148
  history 26
  integral radio systems 82
  removing and replacing 147
matte finish masks 47
Mauritius 265
Medas Islands 243
medical checks 155
Mediterranean 206–207
  Medas Islands 243
  Sardinia 244
  Ustica Island 245
  wreck of the *Zenobia* 241, 242
membrane drysuits 54, 55
mesh bags 79
Mexico
  Cenotes 31
  Roca Partida 206, 320
Micronesia 296–97
Middle East 256, 266–69
Mikhail Lermontov, wreck of 293
military diving 197
mixed-gas diving *see* technical diving
mollusks 230–32
moon, tides 102
moving underwater 122–23

Mozambique 262–63
MP3 players 29
multi-gas dive computers 66
multi-tools 78, 88
muscles, exercises 151–53
music, MP3 players 29

N

narcosis, nitrogen 96, 101, 144, 195
Natal 260, 261
nature diving 172–75
NAUI (National Association of Underwater Instructors) 35
navigation underwater 124–27
neoprene
  drysuits 54, 55
  gloves 45
  semi-dry suits 53
  skinsuits 51
  wetsuits 52
neutral buoyancy 97, 116, 118, 119
New Zealand 282
  Poor Knights Islands 290–91
  wreck of *Mikhail Lermontov* 293
  wreck of *Rainbow Warrior* 292
night diving 184–85
Ningaloo Reef 289
nitrogen
  accumulation in body tissues 100–01
  breathing compressed gas 96
  decompression sickness 25, 142–43
  narcosis 96, 101, 144, 195
  overweight divers 150
  rebreathers 83
  technical diving 97, 194–95
nitrox
  tanks 68, 69
  oxygen-compatible regulators 63
  technical diving 194–95
North America 306
Norwegian fjords 253
nosebleeds 99
nutrition 135, 154, 155

O

O-rings
  cameras 176, 179
  checking for damage 85
  maintenance 88
  tank valves 69
Oceania 294–305
oceans 12
  marine wildlife 200–201
  open oceans 214–15
  tides and currents 102–03
octopus second stages 43, 61, 62, 63
  malfunctioning regulators 138
  out-of-air procedure with buddy 140
OK signal 113, 114, 133
open-circuit systems 26, 27
Orkney Isles 248–49
out-of-air procedure with buddy 140
Outer Banks 323
overweight divers 150
oxygen
  breathing compressed gas 96
  for decompression sickness 142
  freediving 190
  rebreathers 82–83
  technical diving 97, 194–95
  toxic effects 96–7
oxygen-compatible regulators 63

P

Pacific Ocean
  Bikini Atoll 299
  Bora Bora 300
  Cocos 206, 319
  Coral Sea 286
  Fiji 304
  Galápagos Islands 326–27
  Hawaii 321
  Malpelo 206, 319
  Oceania 294–305
  Palau 296–97
  Papua New Guinea 278–79
  Tonga 305
  Truk Lagoon 302–303
  wreck of the *President Coolidge* 301
  Yap 298
packing equipment 163
PADI (Professional Association of Diving Instructors) 11, 35, 238
Palau 296–97
panic attacks 141, 150
Papua New Guinea 278–79
pearl diving 23
personal locator beacons (PLBs) 76
Philippines 280–81
phosphorescence 184
photography 175, 176–83
physics, depths and pressure 94–95
phytoplankton 209, 214
pinnipeds 218, 219
pistol-grip flashlights 73
plankton 201
  as navigational aids 125
  open oceans 214–15
plankton *cont'd.*
  phosphorescence 184
  subtropical seas 207
  temperate seas 208, 209
planning dives 160–65
poison, stingrays 223
polar waters 212–13
police diving 197
pony bottles 68, 186
Poor Knights Islands 290–91
porpoises 218, 219
post-dive tasks 134
pouch-style weight belts 70
pre-dive briefing 111, 165
pre-dive checks 108–11
pregnancy 33
President Coolidge, wreck of 301
pressure *see* air pressure; water pressure
problem-solving 136–41
professional diving 159, 196–97
propulsion underwater 122–23
Protea Banks 261

Q

qualifications 32
Queensland
  Great Barrier Reef 282, 284–85
  wreck of the *Yongala* 287

R

radio
  distress signals 76
  integral systems in masks 82
  radio-updated dive computers 66
Rainbow Warrior, wreck of 292
rays 201, 222–23
rebreathers 26, 29, 82–83
recall signals 93
rechargeable flashlights 73
recreational diving 158–59
Red Sea 256
  The Brothers 266
  Dahab 267
  wreck of the *Thistlegorm* 268–69
reeds, breathing through 22
reefs 16–17, 202–203
  artificial 19, 325
  Asia 270
  Belize Barrier Reef 316–17
  Cayman Islands 312
  destruction of 204–205
  Florida Keys 308
  Galápagos Islands 326–27
  Great Barrier Reef 282, 284–85
  Hebrides 247
  Lundy 252

- reefs *cont'd.*
  - as navigational aids 125
  - Ningaloo Reef 289
  - "no-touch" rule 172
  - Sodwana Bay 260
- reels, surface marker buoys 44, 74, 75
- reflective signals 77
- refraction, light 92
- regulators 43, 60–63
  - air-pressure gauges 64, 65
  - bags 79
  - buddy check 110
  - buoyancy compensators 56
  - checking for damage 85
  - choosing 63
  - clearing 116
  - development of 27
  - disassembly 86–87
  - environmentally sealed 45, 189
  - first stages 43, 60–61, 63
  - functions 96
  - gear assembly 84
  - hoses 89
  - ice diving 45, 189
  - pressure tests 162
  - problem-solving 138
  - second stages 43, 60–66, 69
  - technical diving 63, 195
  - types of 63
  - *see also* octopus second stages
- renting equipment 33, 42, 238–39
- reptiles 228–29
- research
  - nature and conservation diving 174
  - planning a trip 160–61
  - wrecks 171
- rigid inflatable boats (RIBs) 133, 134
- rip currents 103
- rivers 216
- Roca Partida 206, 320
- rocks, exiting onto 132
- Rouquayrol, Benoit 26, 27
- Rule of Thirds, air consumption 117, 168–69
- Russia, White Sea 255

S

- SAA (Sub-Aqua Association) 35
- safety 15
  - ascents 130
  - cavern diving 186
  - dive computers 67
  - ice diving 188–89
  - learning to dive 37
  - night diving 185
  - planning dives 161
  - rays 223
- safety *cont'd.*
  - safety and signaling devices 76–77
- sailfish 215
- salt, washing gear 134
- sand ripples, navigational aids 125
- Sardinia 244
- Sargassum weed 214
- Scapa Flow 248–49
- Scilly Isles 250
- scooters 29, 80
- Scotland
  - Hebrides 246–47
  - Scapa Flow 248–49
- Scuba Rangers 38
- Scyllis 22
- sea grass beds 203
- sea otters 218, 219
- sea urchins 233
- seals, drysuits 55
- seals, wildlife 219
- seasickness tablets 109
- seated entry 112
- seawater, density 94, 118, 216
- seaweed 209
- second stages 43, 60–66, 69
  - octopus 61, 62, 63
- security, bags 79
- semi-closed rebreathers (SCRs) 83
- semi-dry suits 53
- semi wing buoyancy compensators 59
- Seychelles 264
- shakers 81
- sharks 209, 220–21, 258, 261, 266
- shears 78
- ships *see* wreck diving
- shore
  - diving from 109, 112
  - exiting onto 132
- shortie wetsuits 50, 51
- shoulder exercises 153
- Sicily 245
- side view masks 47
- Siebe, Augustus 25
- signals
  - ascents 128
  - attracting buddy's attention 81
  - emergencies 132, 141
  - entering water 113
  - hand signals 93, 144–45
  - hearing underwater 92–93
  - ice diving 188
  - night diving 184, 185
  - OK signal 113, 114, 133
  - out-of-air 140
  - recall signals 93
  - signaling devices 76–77
  - signaling to boats 132–33
- Similan Islands 274
- sinking problems 114
- sinuses, pressure underwater 93, 98–99
- sirenians 218, 219
- sites
  - choosing 161
  - site-specific planning 164–65
- skates 222
- The Skelligs 251
- skinsuits 51
- slack water 102, 125, 127
- slates 44, 81
- Smeaton, John 24
- snakes, sea 228
- snorkeling fins 49
- snorkels 26–27, 48
- Society Islands 300
- Sodwana Bay 260
- sound, underwater 92–93
- South Africa
  - Aliwal Shoal 259
  - Gansbaai 258
  - Protea Banks 261
  - Sodwana Bay 260
- South America 306
- Spain 243
- special gas mixes *see* technical diving
- split fins 48, 49
- sponges 233
- SSAC (Scottish Sub-Aqua Club) 35
- SSI (Scuba Schools International) 35
- standing currents 103
- starfish 232–33
- steel tanks 68
- stills photography 178–79
- stomach, pressure underwater 98
- storage, dive gear 87
- strobe lights 76, 77
- submersible pressure gauges (SPG) 64
- submersible vehicles 29
- subtropical wildlife 206–07
- suit inflation tanks 68, 69
- suits *see* diving suits
- Sulawesi, North 276–77
- Sun, as navigational aid 125
- surf, entering water through 112
- surface marker buoys (SMB) 44, 74–75
  - cold-water diving 44
  - delayed surface marker buoys (DSMBs) 44, 74–75, 76, 128–29, 130–31
  - night diving 185
- surface skills 132–33
- surfer's ear 22
- swimmer kick 123
- swimmer's ear 99, 155
- swimming underwater 122–23

T

- talcum powder 50
- tank pingers 81
- tanks 43, 68–69
  - air-lock tank bands 81
  - air-pressure gauges 64, 65
  - attaching to buoyancy compensators 57
  - cavern diving 186
  - checking for damage 85
  - disassembly 86–87
  - gear assembly 84–85
  - ice diving 189
  - maintenance 69, 88
  - pressure tests 162
  - problem-solving 139
  - storage 69
  - tank pingers 81
  - technical diving 195
  - types of 68
  - valves 69
- Tasmania 288
- technical diving 158
  - equipment 195
  - special gas mixes 69, 97, 194–95
- temperate wildlife 208–11
- temperatures 93
  - choosing diving suits 50
  - cold-water diving 44–45, 63, 253
  - hypothermia 45, 93
  - ice diving 188–89
  - open oceans 214
  - polar waters 212
  - temperate seas 208
- test stamps, tanks 68
- thermal undersuits 45, 55
- thermometers 65
- *Thistlegorm*, wreck of 268–69
- tides 102–03, 125, 161
- time, extending dive times 194–95
- timetabling trips 162
- Tonga 305
- tool kits 88
- torso exercises 153
- tour operators 238
- training *see* learning to dive
- travel, air 239
- travel buoyancy compensators 59
- travel cases 79
- trim, buoyancy control 121
- trimix 69, 194, 195
- tropical wildlife 202–05
- Truk Lagoon 302–3
- tuna 209, 215
- Turks and Caicos 313
- turtles 19, 228, 229
- twin-set tanks 68, 69

UV

- Ufano, Diego 23
- umbilical flashlights 73
- unconsciousness, emergency lifts 141
- undersuits, thermal 45, 55
- United States of America
  - Catalina Island 322
  - Crystal River 309
  - Florida Keys 308
  - Great Lakes 324
  - Outer Banks 323
- Ustica Island 245
- vacations 236–39
- Valsalva maneuver 93, 99
- valves
  - demand valves 43, 60–61, 69
  - inflator valves 54, 55, 84
- Vancouver Island 325
- Vanuatu 301
- vertigo 99
- videography 175, 180–83
- vintage-style regulators 63
- virgin wrecks 170–71
- voluntary work
  - conservation diving 172–74
  - underwater archaeology 192

W

- washing equipment 134
- watches 64
- water
  - density 94, 118, 216
  - entry into 112–13
- water pressure
  - depth and 94
  - effect on body 93, 98–99
  - effect on gases 95
- weather forecasts 108, 109
- weights 43, 70–71
  - buddy check 110
  - buoyancy control 118
  - descents 114
  - harnesses 70
  - jettisoning 120, 121, 140
  - refining 148
  - removing belt underwater 139
  - securing weights 71
  - swimming profile 123
  - weight distribution 121
  - weighted lines 115
- wetsuits 50
  - buoyancy control 118
  - caring for 52
  - full-length 52
  - layers 51
  - shortie 50, 51
- whales 212, 218–19
- wheeled dive bags 238
- whistles 76–77
- White Sea 255
- wildlife 12, 199–233
  - fish 220–27
  - freshwater environments 216
  - invertebrates 230–33
  - mammals 218–19
  - nature and conservation diving 172–75
  - open oceans 214–15
  - polar waters 212–13
  - reptiles 228–29
  - subtropical seas 206–07
  - temperate seas 208–11
  - timing dives 161
  - tropical seas 202–05
  - wreck diving 166
- wing buoyancy compensators 59, 195
- women's buoyancy compensators 59
- wooden ships, wreck diving 170
- World War II, frogmen 83
- worms, segmented 233
- wreck diving 18, 166–71
  - Bikini Atoll 299
  - Bonaire 311
  - The Brothers 266
  - Cuba 314
  - Europe 240
  - freshwater environments 216
  - Great Lakes 324
  - Isles of Scilly 250
  - *Mikhail Lermontov* 293
  - navigational aids 125
  - New Zealand 282, 292–3
  - night diving 185
  - Oceania 294
  - Palau 296
  - Papua New Guinea 278–9
  - Philippines 281
  - *President Coolidge* 301
  - *Rainbow Warrior* 292
  - Sardina 244
  - Scapa Flow 248–9
  - *Thistlegorm* 268–9
  - Truk Lagoon 302
  - *Yongala* 287
  - *Zenobia* 241, 242
- writing underwater 81

XYZ

- Yap 298
- *Yongala*, wreck of 287
- Yucatan Peninsula 315
- *Zenobia*, wreck of 241, 242
- zippers, drysuits 55

ACKNOWLEDGMENTS

Authors' Acknowledgments

Monty Halls (www.montyhalls.co.uk) would like to thank: Oceanic, Gates Housings, Fourth Element, Ed Poore of Poseidon Diving, the Hilton Hotel in Dahab, Dave Moore—the finest diver I have ever met—and in particular Dr. Antje Steinfurth.

Miranda Krestovnikoff would like to thank: Bristol University Sub Aqua club for introducing me to the world of diving, and Damian Wozniak and Tim Walsh for their thorough training and inspiration; DIVER magazine for giving me opportunities to visit exotic diving destinations and for publishing the resulting articles; and Dave Shaw, Dive Services UK, Phil Bullen, and Richard Bull for attentive and solid buddy support when diving commercially. Both Monty and I would like to thank Dan Burton for his spectacular photography in this book and for making initial contacts with DK to make it all possible. Thanks also to the editors for working with Monty and me, taking our abstract scribbles and turning them into this dive guide. Most importantly, a huge thank you to my husband, Nicholas, who has been immensely patient and supportive during book writing, and for my baby daughter, Amélie, who waited to emerge until after my deadlines were met!

Publisher's Acknowledgments

Dorling Kindersley and cobalt id would like to thank the following for their invaluable help with this book: Dr. Kim Bryan and Letitia Luff for editorial assistance; Hilary Bird for indexing; Christine Heilman for Americanization; and Stephen Dye for specialist advice. Special thanks to Poseidon Divers (www.poseidondivers.com) for their generous provision of diving equipment and expert advice during the photographic shoot in Dahab, and for the help of Mark Rogers, Katie Rogers, and Sophia Crook, who kindly agreed to act as models; and to Dan Burton for technical guidance throughout the production of this book. Many thanks also to John Womack of Otter Watersports (www.drysuits.co.uk) for allowing us to use his products and premises for photographic purposes; Vicky Boateng and Catwalk Model Management; the team at PADI UK for kindly arranging permission for us to reprint PADI advisory texts; Coral Cay for providing images of their work in marine conservation; and Fraser Bathgate for advice on, and images of, adaptive training techniques.

We also extend our thanks to the following manufacturers who kindly contributed images and products for the production of this book: Mares, Tabata USA (TUSA), Aqua Lung, Apeks Marine Equipment, AP Valves, Delta P Technology, Seemann, Diving Unlimited International Inc., Bob Evans Designs Inc., Mazin Submersible Technology Inc., SpecialFins Ltd, Northern Diver, Inon Inc., Sea & Sea, Canon Inc., Pentax, DZ Active Systems, HydroOptix LLC, Ocean Reef, Nikon Corporation, McMahon Precision Engineering Ltd, Ambient Pressure Diving Ltd, Bright Weights, Fujifilm, Ikelite, and Underwater Kinetics. Jetboots™ are manufactured by Mazin Submersible Technology, Inc. (www.jetboots.com). Force Fins are manufactured by Bob Evans Designs Inc. (www.forcefin.com).

Picture credits

The publisher would like to thank the following for their kind permission to reproduce their photographs:

Abbreviations key: t = top; b = bottom; l = left; r = right; c = center; rh = running header.

4 Dan Burton; **6–7** Kevin Davidson; **8–9t** Kevin Davidson; **8cr** Dorling Kindersley/Dave King/Courtesy of the North Wind Undersea Institute, New York; **9cl** iStockphoto/Andy Lim; **9cr** iStockphoto/Ian Scott; **10–19rh** Dan Burton; **10bl** Linda Pitkin; **10c** Alamy/POPPERFOTO; **11b** Ross & Diane Armstrong; **11tr** Dan Burton; **12–13b** Tony White; **13br** Ross & Diane Armstrong; **13tl** Dan Burton; **14** Dan Burton; **15bl** Linda Pitkin; **15tr** Dan Burton; **16b** Dan Burton; **16tl** Dan Burton; **16tr** Kevin Davidson; **18** Linda Pitkin; **19b** Kevin Davidson; **19tl** Kevin Davidson; **19tr** Linda Pitkin; **20–21** Getty/Hulton Archive/Fox Photos; **22–29rh** Dorling Kindersley/Dave King/Courtesy of the North Wind Undersea Institute, New York; **22cl** Mary Evans Picture Library; **23c** Alamy/POPPERFOTO; **23tr** Mary Evans Picture Library; **24c** Mary Evans Picture Library; **24tr** Mary Evans Picture Library; **25c** Corbis/Bettmann; **26c** Dorling Kindersley/Alex Wilson/Courtesy of the Charlestown Shipwreck and Heritage Centre, Cornwall; **26tl** Mary Evans Picture Library/Dr David Lewis Hodgson; **27t** Corbis/Bettmann; **28c** Kevin Davidson; **29tr** Dan Burton; **30–31** Dan Burton; **33tr** Dorling Kindersley/John Davis; **34–35b** Dan Burton; **34c** Kevin Davidson; **35tr** Kevin Davidson; **38–39b** Dan Burton; **38tc** Fraser Bathgate; **39br** Alamy/David Fleetham; **40–41** Dan Burton; **45t** Dan Burton; **46–47b** Kevin Davidson; **47b** Mares; **47c1** Mares; **47c2** Northern Diver; **47c3** Ocean Reef; **47c4** Ocean Reef; **47t1** Mares; **47t2** Mares; **48cr1** Tabata USA (TUSA); **48cr2** Aqua Lung; **49c1** Mares; **49c2** Tabata USA (TUSA); **49c3** Tabata USA (TUSA); **49t1** Aqua Lung; **49t2** Bob Evans Designs Inc. (Force Fin); **51bl1** Tabata USA (TUSA); **51bl2** Tabata USA (TUSA); **52bc1** Mares; **52bc2** Mares; **52t** Dan Burton; **54br** Diving Unlimited International Ltd; **54tr** Nataliya Chervyakova; **55cr2** Mares; **56b** Dan Burton; **58c** Kevin Davidson; **58t** Kevin Davidson; **59l1** AP Valves; **59l2** Mares; **59l3** Mares; **59l4** Tabata USA (TUSA); **59l5** Aqua Lung; **59r1** AP Valves; **59r2** Mares; **59r3** Apeks Marine Equipment; **59r4** Tabata USA (TUSA); **59r5** Aqua Lung; **60b** Kevin Davidson; **61br** Aqua Lung; **63br** Kevin Davidson; **63tr1** Tabata USA (TUSA); **63tr2** Aqua Lung; **63tr3** Mares; **63tr4** Apeks Marine Equipment; **63tr5** Aqua Lung; **64bc** Mares; **64bl** Mares; **65b** Corbis/Todd Gipstein; **65tr1** Apeks Marine Equipment; **65tr2** Mares; **65tr3** Apeks Marine Equipment; **65tr4** Mares; **66b2** Mares; **66b3** Tabata USA (TUSA); **67cl** Mares; **68c** Dan Burton; **69b** Kevin Davidson; **70br1** Seemann; **70br2** Mares; **70cr1** Mares; **71bl** Bright Weights; **72tr3** Kevin Davidson; **74b** Alamy/GUILLEN PHOTOGRAPHY; **75bl1** AP Valves; **75bl2** AP Valves; **75bl3** AP Valves; **75t** Dorling Kindersley/Demetrio Carrasco/Rough Guides; **76c** Dan Burton; **77b** Dan Burton; **77cl1** Aqua Lung; **77cl2** AP Valves; **79br1** Mares; **79br2** Mares; **79br5** Apeks Marine Equipment; **79cr** Alamy/Jack Sullivan; **80b** Dan Burton; **81br** Mares; **81c** Alamy/GUILLEN PHOTOGRAPHY; **82bl** Alamy/Stephen Frink Collection; **82c** HydroOptix LLC; **82tr** Ocean Reef; **83bl** Dan Burton; **90-91** Dan Burton; **92–103rh** iStockphoto/Dennis Sabo; **92bc** Alamy/Chris A Crumley; **95bc** Kevin Davidson; **97cr** Kevin Davidson; **98cr** iStockphoto/Daniel Gustavsson; **99c** Kevin Davidson; **100tc** iStockphoto/Dennis Sabo; **101tc** Kevin Davidson; **102tc** iStockphoto/Kevin Bergen; **103bc** Kevin Davidson; **109br** iStockphoto/

Joe Gough; **109tr** Alamy/Nick Hanna; **113bl** Dorling Kindersley/Brian Pitkin; **117br** Mares; **122bl** Kevin Davidson; **125tc** Linda Pitkin; **129tl** Dan Burton; **132tr** Dan Burton; **135tc** Paul Reid; **136bc** Kevin Davidson; **137tc** Kevin Davidson; **141tr** Dan Burton; **152-153c** Kevin Davidson; **152tr** Dorling Kindersley/John Davis; **153tl** Dorling Kindersley/John Davis; **153tr** Dorling Kindersley/John Davis; **154cr** Dorling Kindersley/Sian Irvine; **154tc** Dorling Kindersley/Andrew Whittuck; **156-157** Dan Burton; **159cl** Nataliya Chervyakova; **160b** Ross & Diane Armstrong; **160tc** Marek Walisiewicz; **161br** Nataliya Chervyakova; **161tl** Poseidon Divers; **162c** Marek Walisiewicz; **163t** Dan Burton; **164bl** Dan Burton; **165tl** Dan Burton; **166bl** Kevin Davidson; **166tc** Kevin Davidson; **167** Linda Pitkin; **168t** Kevin Davidson; **169b** Dan Burton; **169tr** Dan Burton; **170b** Alamy/Stephen Frink Collection; **171br** iStockphoto/Lara Seregni; **171tl** Alamy/POPPERFOTO; **172tc** Coral Cay Conservation; **173b** Kevin Davidson; **173tr** Coral Cay Conservation; **174b** Coral Cay Conservation; **175br** Alamy/Sarkis Images; **175t** Dan Burton; **176bl** Ross & Diane Armstrong; **176tr** Tony White; **177b** Dan Burton; **177tr** Kevin Davidson; **178bc** Pentax; **178bl** Nikon Corporation; **178br** Fujifilm; **178tr** iStockphoto/Dennis Sabo; **179c** Inon Inc.; **179cl** Pentax; **179cr** Inon Inc.; **179t** Sea & Sea; **180bl** Canon Inc.; **180br** Canon Inc.; **180tc** Alamy/Roger Munns; **181b** Dan Burton; **181tc** Sea & Sea; **181tl** Sea & Sea; **181tr** Sea & Sea; **182b** Kevin Davidson; **182tc** Kevin Davidson; **184bl** Sea & Sea/Underwater Kinetics; **184tr** iStockphoto/Rob New; **185t** Dan Burton; **186bl** Dan Burton; **186cr** Tabata USA (TUSA); **187t** Ross & Diane Armstrong; **188bl** Dan Burton; **188c** Mares; **189c** Dan Burton; **189t** Dan Burton; **190bl** Dan Burton; **190c** Mares; **190cr** SpecialFins Ltd; **191br2** Dorling Kindersley/Michael Moran; **191** Dan Burton; **192-193b** Dan Burton; **192bl** Alamy/Hugh Sitton Photography; **193tl** Dan Burton; **194bl** Dan Burton; **195br** Dan Burton; **195t** Dan Burton; **196-197b** Dan Burton; **196bl** Alamy/Henry Owen; **197tl** Dan Burton; **197tr** Corbis/The Cover Story; **198-199** Dan Burton; **200–233rh** iStockphoto/Johnny Lye; **200tc** Tony White; **201b** Kevin Davidson; **201tr** Kevin Davidson; **202bl** iStockphoto/Dan Schmitt; **202tr** iStockphoto/Steven Tulissi; **203b** Linda Pitkin; **203tr** Kevin Davidson; **204cr** Alamy/STOCK IMAGE/PIXLAND; **205** Linda Pitkin; **206b** Dorling Kindersley/Demetrio Carrasco; **206tr** Alamy/GUILLEN PHOTOGRAPHY; **207bc** Alamy/Reinhard Dirscherl; **208-209b** Dan Burton; **209br** Dorling Kindersley/Kim Taylor and Jane Burton; **209tr** iStockphoto/Sue Loader; **210b** Ross & Diane Armstrong; **210tr** Alamy/Robert E. Barber; **211** Linda Pitkin; **212-213b** iStockphoto/vera bogaerts; **212bl** Dorling Kindersley/Harry Taylor; **212tr** iStockphoto/sandra minarik; **213tr** Nataliya Chervyakova; **214-215b** Linda Pitkin; **214cl** Alamy/Michael Patrick O'Neill; **215tr** Tony White; **216bl** iStockphoto/George Hoeylaerts; **216c** Dorling Kindersley/Jerry Young; **216tr** iStockphoto/Vera Bogaerts; **217** Alamy/Visual&Written SL; **218b** Dorling Kindersley/Ken Findlay; **219br** Dorling Kindersley/Frank Greenaway; **219t** Alamy/George McCallum Photography; **220-221tc** Tony White; **220bl** Kevin Davidson; **221br2** Dorling Kindersley/Frank Greenaway; **222-223b** Kevin Davidson; **223tr** Kevin Davidson; **224b** Kevin Davidson; **224c** iStockphoto/Steffen Foerster; **225t** Kevin Davidson; **226bl** Alamy/Reinhard Dirscherl; **226tr** Tony White; **227** Kevin Davidson; **228bl** Linda Pitkin; **228cr** Tony White; **229bc** Kevin Davidson; **230bl** Tony White; **230tr** iStockphoto/Simon Edwin; **231** Kevin Davidson; **232-233b** Ross & Diane Armstrong; **232tc** Tony White; **233br** Linda Pitkin; **234-235** Linda Pitkin; **236–327rh** Dan Burton; **236-237b** Dan Burton; **236cr** Nataliya Chervyakova; **237tr** Dan Burton; **238bl** Mares; **238t** Dan Burton; **239b** iStockphoto/Lisa Kyle Young; **239tr** Dan Burton; **240b** Dorling Kindersley/Frits Solvang; **240tr** Dorling Kindersley/Francesca Yorke; **241** Dan Burton; **242b** Dan Burton; **243bc** Alamy/Wolfgang Pölzer; **243cr** Dorling Kindersley/Ian Aitken/Rough Guides; **244cr** Alamy/Reinhard Dirscherl; **245b** iStockphoto/Dan Schmitt; **246bc** Alamy/Worldwide Picture Library; **247tr** Dan Burton; **248bl** Alamy/Worldwide Picture Library; **249bc** Dan Burton; **249tr** Corbis/Bettmann; **250bl** Dorling Kindersley/Frank Greenaway; **250c** Alamy/Trevor Smithers ARPS; **251b** Alamy/David Lyons; **252cr** Dan Burton; **253b** Dorling Kindersley/Linda Whitwam; **254b** iStockphoto/Nancy Barr-Raper; **255cr** Nataliya Chervyakova; **256bl** Dan Burton; **256c** Linda Pitkin; **257** Linda Pitkin; **258c** Tony White; **259b** Dan Burton; **260b** Dan Burton; **260c** Alamy/Peter Pinnock; **261c** Tony White; **262–263b** Dan Burton; **263tr** Dan Burton; **264b** Linda Pitkin; **265c** Alamy/Jon Arnold Images; **266c** Dan Burton; **267b** Dan Burton; **268tr** Dan Burton; **269bc** Dan Burton; **270bl** Dan Burton; **271bc** iStockphoto/Chartchai Meesangnin; **271tr** Ross & Diane Armstrong; **272-273c** Dan Burton; **273tr** Dan Burton; **274tr** Alamy/Reinhard Dirscherl; **275bc** Dan Burton; **276br** Linda Pitkin; **276c** Ikelite; **277bc** Tony White; **278bl** Linda Pitkin; **279bc** Ross & Diane Armstrong; **280bl** iStockphoto/Erik de Graaf; **281br** Kevin Davidson; **281tc** Alamy/F. Jack Jackson; **282bl** Dan Burton; **282tc** Ross & Diane Armstrong; **284bl** Dan Burton; **285bc** Alamy/Iconsinternational.Com; **285tl** Dorling Kindersley/Alan Williams; **286bc** Dan Burton; **287tr** Dan Burton; **288bc** Tony White; **289bc** Andy Salmon; **290bc** Dan Burton; **291br** Ross & Diane Armstrong; **291tc** Ross & Diane Armstrong; **292tr** Corbis/Greenpeace/Sygma; **293br** Dan Burton; **294-295c** Dan Burton; **294bl** Dan Burton; **295tr** Kevin Davidson; **296br** Dan Burton; **297** Kevin Davidson; **298bl** Dan Burton; **299bc** Corbis/Claude Rives/Mer Images; **299br** Alamy/POPPERFOTO; **300b** Alamy/Chad Ehlers; **301br** Dan Burton; **302bl** Dan Burton; **303** Dan Burton; **304bc** Alamy/Stephen Frink Collection; **304cr** Alamy/David Fleetham; **305b** Tony White; **306b** Alamy/Chris A Crumley; **306c** Linda Pitkin; **307** Linda Pitkin; **308c** Dan Burton; **309b** Dan Burton; **309cr** iStockphoto/Wayne Johnson; **310c** Alamy/Stephen Frink Collection; **311b** Linda Pitkin; **312b** Dan Burton; **313bc** Alamy/ImageState; **313cr** Alamy/OceanPix.com; **314bl** Linda Pitkin; **315c** Alamy/Gavin Newman; **316–317b** Alamy/Brandon Cole Marine Photography; **317cr** Alamy/Brandon Cole Marine Photography; **318b** Alamy/Andre Seale; **319c** Ambient Pressure Diving Ltd; **319cr** Alamy/Jeff Rotman; **320b** Alamy/Wolfgang Pölzer; **321b** Dorling Kindersley/Rob Reichenfeld; **322bc** iStockphoto/William Smith; **322cr** Linda Pitkin; **323b** Tony White; **324b** Alamy/Visions of America, LLC; **325br** Alamy/Wolfgang Pölzer; **326–327b** iStockphoto/Dan Cooper; **327tr** Linda Pitkin.

All maps on pages 200, 204, and 242–326 are Mountain High Maps® copyright © 1993 Digital Wisdom, Inc.

Every effort has been made to trace the copyright holders. The publisher apologizes for any unintentional omissions and would be pleased, in such cases, to place an acknowledgment in future editions of this book.

All other images © Dorling Kindersley
For further information, see **www.dkimages.com**

RYA

Sea Survival

HANDBOOK

Keith Colwell

RFD
SURVIVA
12 PERSONS

THE RYA'S COMPLETE GUIDE TO SURVIVAL AT SEA

Keith Colwell

Keith Colwell has sailed all his life, in all sorts of craft from simple sailing dinghies to high-speed motor cruisers.

Although Keith trained as a naval architect and was employed initially as a development engineer for a workboat builder, he has spent most of his career as a yachting journalist. He worked for 22 years as a staff journalist starting as a Technical Reporter for *Motorboat and Yachting*, Technical Editor and then Deputy Editor of *Practical Boat Owner* before jumping ship to become Development Editor of *Sailing Today*.

He is now employed by the Royal National Lifeboat Institution as a Divisional Sea Safety Manager, managing and training a large team of volunteer RNLI safety advisers, who provide a safety advice service to the boating public.

Keith is an RYA Yachtmaster Offshore Sail, an RYA Advanced Powerboat Instructor, a SRC instructor/assessor and a RYA and MNSTCW'95TB Sea Survival Instructor.

RYA Sea Survival Handbook

by Keith Colwell

Foreword by Mike Golding
Illustrations by Steve Lucas

2008

© RYA 2008

First Published 2008
The Royal Yachting Association
RYA House, Ensign Way, Hamble
Southampton SO31 4YA
Tel: 0845 345 0400
Fax: 0845 345 0329
Email: publications@rya.org.uk
Web: www.rya.org.uk
ISBN: 978-1-905104-31-4
RYA Order Code: G43

Totally Chlorine Free *Sustainable Forests*

All rights reserved. No part of this publication or cd-rom may be stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission in writing of the publishers.

A CIP record of this book is available from the British Library. Telephone 0845 345 0400 for a free copy of our Publications Catalogue.

Note: While all reasonable care has been taken in the preparation of this book, the publisher takes no responsibility for the use of the methods or products or contracts described in the book.

Acknowledgements: Ralph Macdonald Boatswain Joint Services Adventurous Sail Training Centre, Steve Judkins JSASTC
Cover Design: Design House
Illustrations: Steve Lucas
Typeset: Creativebyte
Proofreading, index and glossary: Alan Thatcher
Printed in China through World Print

Contents

1 **Before you go out** 5

2 **Weather** 17

3 **Preparing for heavy weather** 22

4 **Handling heavy weather** 27

5 **Lifejackets and buoyancy aids** 33

6 **MOB (Man overboard) prevention** 44

7 **MOB (Man overboard)** 46

8 **Liferafts** 53

9 **Fire control** 69

10 **Raising the alarm** 75

11 **When to abandon ship** 90

12 **Abandoning ship** 93

13 **Survival in cold water** 118

14 **First aid** 125

15 **Life-saving signals** 136

16 **Rescue** 140

Glossary 144

Index 148

Membership pages 157

Foreword

By Mike Golding

'Sea Survival training saves lives, it is essential knowledge for the modern offshore skipper. The stark fact is, if things do go dramatically wrong onboard, this knowledge can save lives. Training for the worst eventuality will make you better prepared and it will give you the confidence to know that you have planned for every conceivable circumstance when you leave port. The sea can be a hostile and brutal place for the ill prepared, you don't need to be in the Southern Ocean, the English Channel or anywhere away from easy access to shore, can quickly become just as dangerous and hostile if things go badly wrong onboard your boat. When it comes to survival, preparation is everything and all good preparation begins with solid information and sound practical training.

This book highlights the importance of having the correct safety equipment and will also give you the knowledge of know how and when to use it. Armed with this information and the right equipment every eventuality onboard, no matter how severe, can be met with the determination and conviction that is the essential mark of every survivor.'

Introduction

Boating is one of the safest leisure sporting activities. However, in any sport that pushes us to the edge of our abilities, there's always a chance of accident or injury. Being properly prepared makes us better able to cope with an emergency and will significantly increase our chances of survival.

This handbook is essential for anyone taking the one-day RYA Basic Sea Survival course and the two-day RYA/ISAF Offshore Safety course. It is also invaluable for anyone who takes a boat offshore.

K. Colwell

For further information and a list of establishments running RYA Training Courses, contact the RYA at RYA House, Ensign Way, Hamble, Southampton, SO31 4YA.
Tel 0845 345 0400 Fax 0845 345 0329

CHAPTER 1 : Before you go out

In the UK, HM Coastguard has a free Voluntary Safety Identification Scheme (CG66). Details about your boat are stored on a database accessible by all HM Coastguard Co-ordination Centres. A contact number is provided for those ashore should they become worried about you. To join, visit www.mcga.gov.uk

Check the weather forecast and tides. Familiarise yourself with the area you're planning to sail. Are you and your crew sufficiently skilled for the planned trip? Tell somebody ashore where you are planning to go, when you expect to get there and how many people are on board. Let them know when you expect to arrive and what to do if you don't contact them to tell them of your safe arrival.

Many other countries have similar voluntary or compulsory registration schemes, so check if the country in which you keep your boat has a scheme and how you can join.

Safety brief

As skipper you should ensure that everyone on board knows where the safety equipment is stowed and how to use it. Talk them through your passage and pilotage as well as your contingency plan should something go wrong. Don't forget the less obvious safety equipment and procedures such as how to drop anchor or how to read your current position off the GPS or chart.

Are your crew familiar with boating? Is there anything you should know about them? Can they swim? Do they have medical problems that you need to be aware of, such as asthma, diabetes or angina? As skipper, their safety is your responsibility.

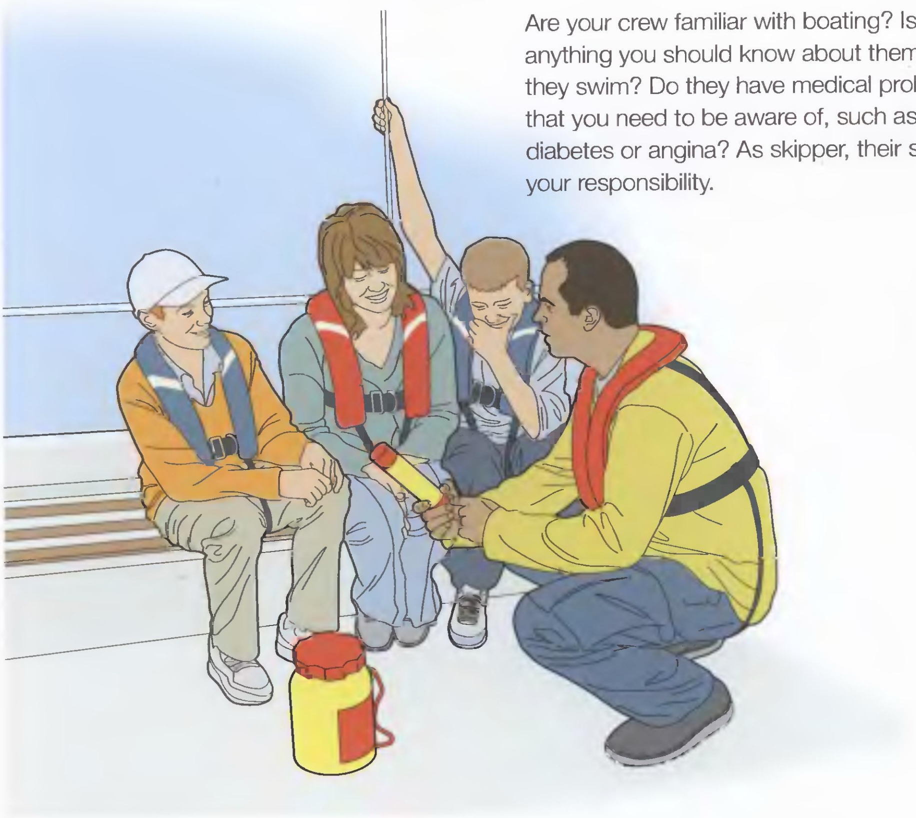

Safety brief checklist

Lifejackets	Check condition, lights, when and how to don and inflate
Harnesses	When and where to clip on
Flares	Where stowed and how to fire
Liferaft	Where stowed and how to launch and inflate
Grab bag	Where it is, also other items such as water, food etc
Onboard hazards	Boom, winches, cleats, ropes etc
Anchoring procedure	
How to start the engine	
Lifebuoy/dan buoy	
VHF radio	Mayday and DSC procedures
GPS	How to read off a position
Bilge pump	Where and how to use it
First-aid kit	Where it is kept
Fire extinguishers	Where, what type and how to use them
Medicines	Are they taking any medication including anti-seasickness tablets
Clothing	Spare clothing, sunscreen
Lookout	Tell me what you see and hear; I may not have seen or heard it!
Passage Plan	Where we are going and contingency plans
MOB recovery equipment	Where and how to use it
Watch rota	When to call the skipper
Can they swim?	
Medical or physical problems?	

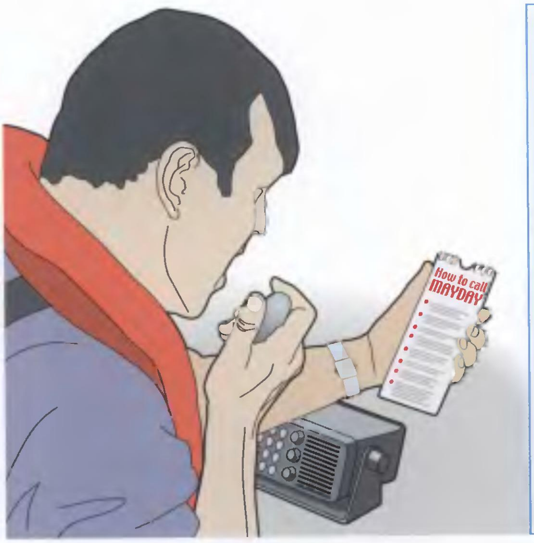

VHF tip

Operating procedures for marine-band radios vary from set to set. Stick a prompt-card in a prominent position close to the set. Make sure the crew know not only how to switch on the radio – including the correct switch on the distribution board – but also how to use the DSC distress button (if fitted) and how to choose Ch 16 and transmit a Mayday.

Your MMSI number and call sign should be prominently displayed beside the radio.

Clear labels on lockers and a plan of where the safety equipment is stowed will save time and help reduce panic during an emergency. Place a copy on the door in the heads, where crew may find the time to study it!

Labelling will assist SAR (Search and Rescue) crew if they need to come aboard during a rescue.

Make a plan

Think about what can go wrong and how you can reduce the severity of an incident.
For example, ask yourself:

- What would you do in a range of emergencies?
- How would you go about abandoning ship?
- What you would do if somebody went overboard and how would you recover them?
- When should you fire flares or make a Mayday call?
- What is your contingency plan?
- Talk your procedures through and, where possible, practise them with your crew.

Clothing

Do you have the right clothing for the conditions? Take extra and spare clothing with you because conditions at sea are more hostile than on land.

Synthetic materials are better at wicking moisture away from the skin. Breathable waterproofs keep you warmer not only by keeping you dry from the outside but also by preventing moisture created by your body from becoming trapped inside the clothing.

Most marine clothing manufacturers now provide layer systems that are better at keeping you comfortable in a range of weather conditions.

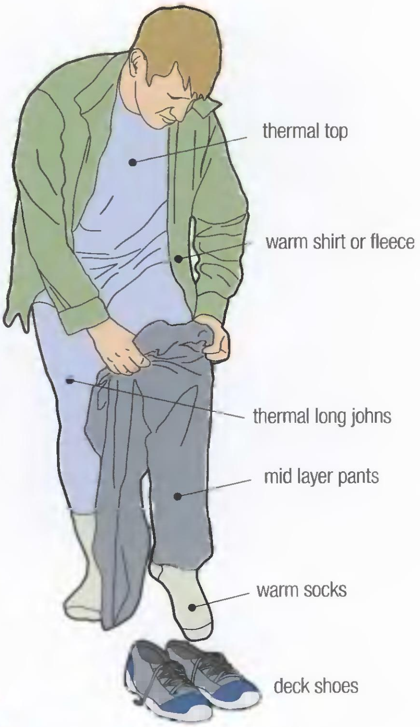
thermal top
warm shirt or fleece
thermal long johns
mid layer pants
warm socks
deck shoes

Cool weather clothing

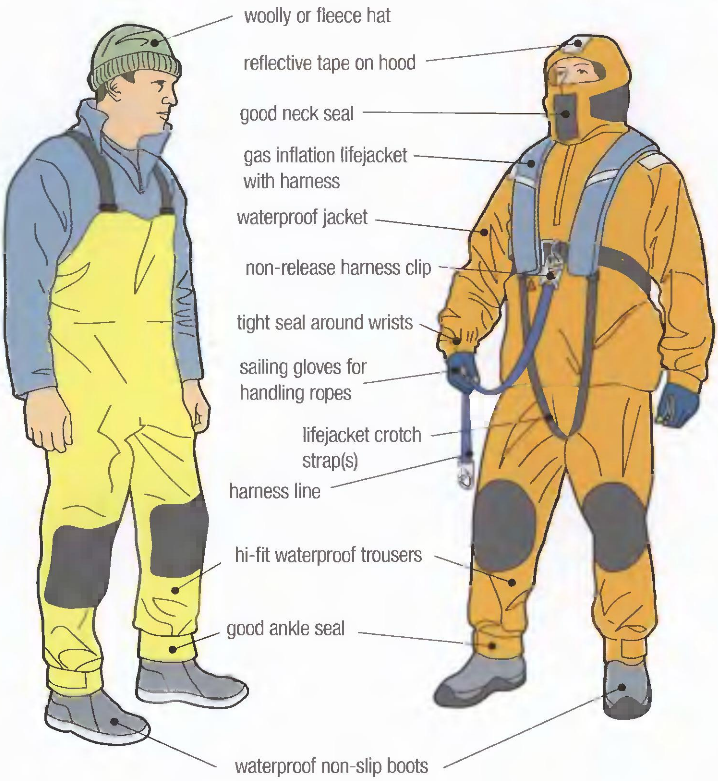
woolly or fleece hat
reflective tape on hood
good neck seal
gas inflation lifejacket with harness
waterproof jacket
non-release harness clip
tight seal around wrists
sailing gloves for handling ropes
lifejacket crotch strap(s)
harness line
hi-fit waterproof trousers
good ankle seal
waterproof non-slip boots

Warm weather clothing

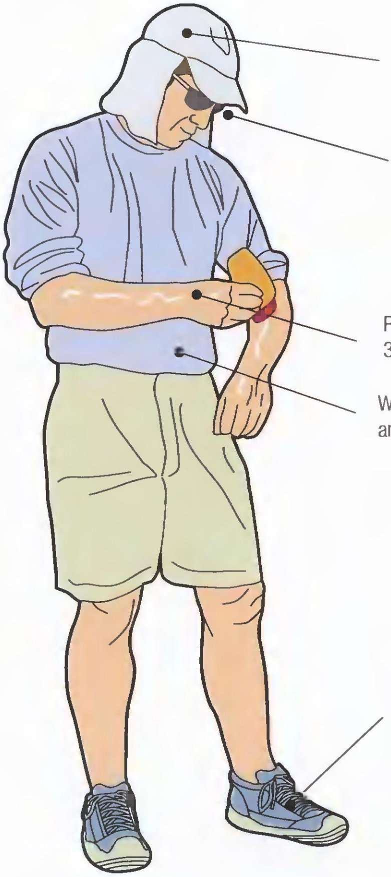

Wear a wide brimmed hat and if possible one with neck protection.

Reflection from water increases the glare effect of the sun. Use sunglasses with 100% UVA protection.

Regularly apply sun block of SPF 30-40 or more on exposed skin.

Wear loose long-sleeved shirts and trousers.

Wear deck shoes to protect feet from hot decks and stubbing toes.

In hot weather, protect yourself from the sun. UV light is reflected by the sea and will increase the possibility of sunburn. Use sun block or high-factor cream. Don't forget vulnerable areas such as tops of your feet, ears and the front of your legs. Long-sleeved shirts and broad-brimmed hats give added protection and wrap-around sunglasses reduce glare.

Recommended safety equipment for sailing yachts

The amount of safety gear you carry depends mainly on where you are planning to sail. Here's a selection of recommended equipment.

Before you start your voyage, leave details with a shore contact. Check requirements for all the countries you are sailing.

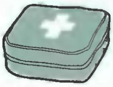

Comprehensive first aid kit and sufficient knowledge to use it correctly

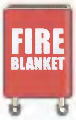

Fire blanket near, but not over cooker

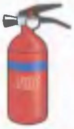

Serviced fire extinguisher for each accommodation space

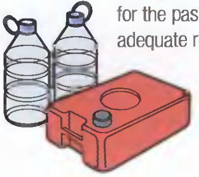

Enough fuel and water for the passage plus an adequate reserve

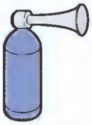

Fog horn

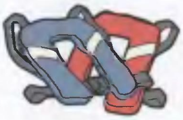

Lifejackets with lights, crotch straps and spray hoods, spares and re-arming kits, harnesses and wet weather gear for all the crew

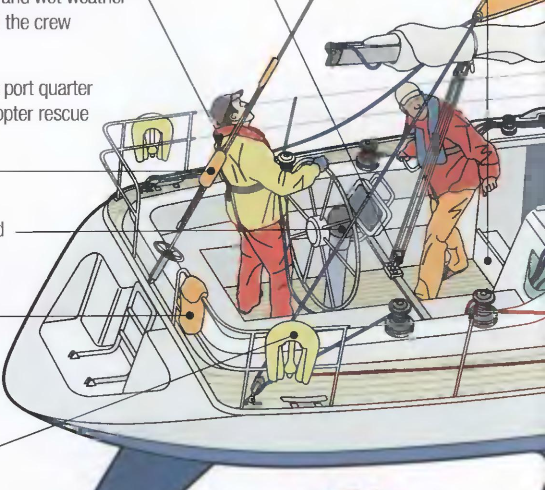

Snap shackle on bottom of main sheet for MOB rescue

System to hold washboards in place in the event of capsize

Try to keep the port quarter clear for helecopter rescue

Dan buoy attached to lifebuoy

Accurate main and hand bearing compasses

Lifebuoy, floating light and drogue

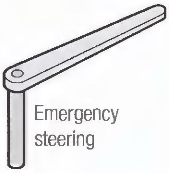

Emergency steering

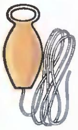

Floating rescue line

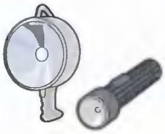

Powerful spotlight and torch with spare batteries and bulbs

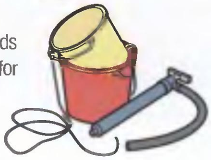

Two buckets with lanyards and a hand bilge pump for use in awkward spaces

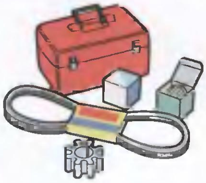

Comprehensive tool kit including sail repair kit, spare navigation light bulbs and fuses

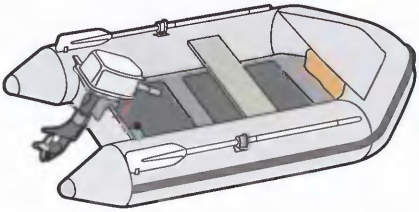

Dinghy with oars, repair kit, pump, anchor, engine with fuel, tools and kill cord

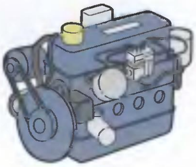

Fully serviced engine, with handbook, adequate spares and sufficient knowledge to effect repairs

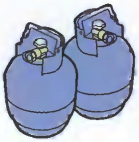

Gas kept in a properly vented locker with all pipes and connections regularly checked for leaks

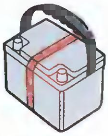

Spare battery capacity

In the UK and Ireland the RNLI offer a free safety equipment advisory service to all boat owners.
www.rnli.org.uk/seasafety

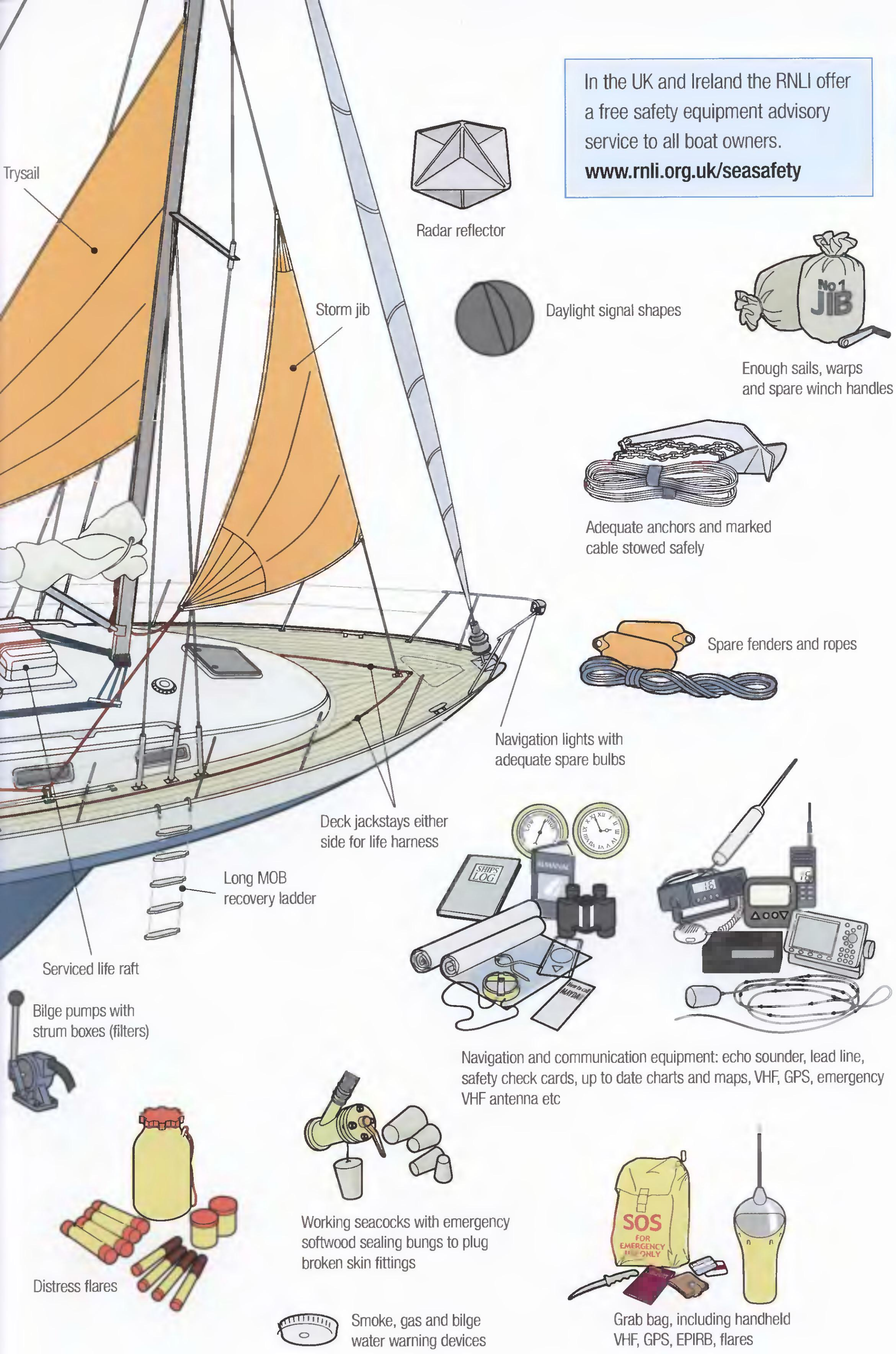
Trysail
Radar reflector
Storm jib
Daylight signal shapes
No 1 JIB
Enough sails, warps and spare winch handles
Adequate anchors and marked cable stowed safely
Spare fenders and ropes
Navigation lights with adequate spare bulbs
Deck jackstays either side for life harness
SHIPS LOG
ALMANAC
Long MOB recovery ladder
Serviced life raft
Bilge pumps with strum boxes (filters)
Navigation and communication equipment: echo sounder, lead line, safety check cards, up to date charts and maps, VHF, GPS, emergency VHF antenna etc
Working seacocks with emergency softwood sealing bungs to plug broken skin fittings
SOS FOR EMERGENCY USE ONLY
Distress flares
Smoke, gas and bilge water warning devices
Grab bag, including handheld VHF, GPS, EPIRB, flares

Equipment for motorboats

Mechanical failure is the main cause of problems in a motorboat. Ensure engines are regularly serviced and checked. Carry sufficient fuel for the proposed voyage. One of the main reasons of crew accidents on a motorboat is the helmsman making a sudden manoeuvre without informing the rest of the crew, who can be thrown off their feet (or worse overboard). While motorboats will carry much of the same equipment as sailing boats, the lack of rigging and the dependence on mechanical power can bring specific problems.

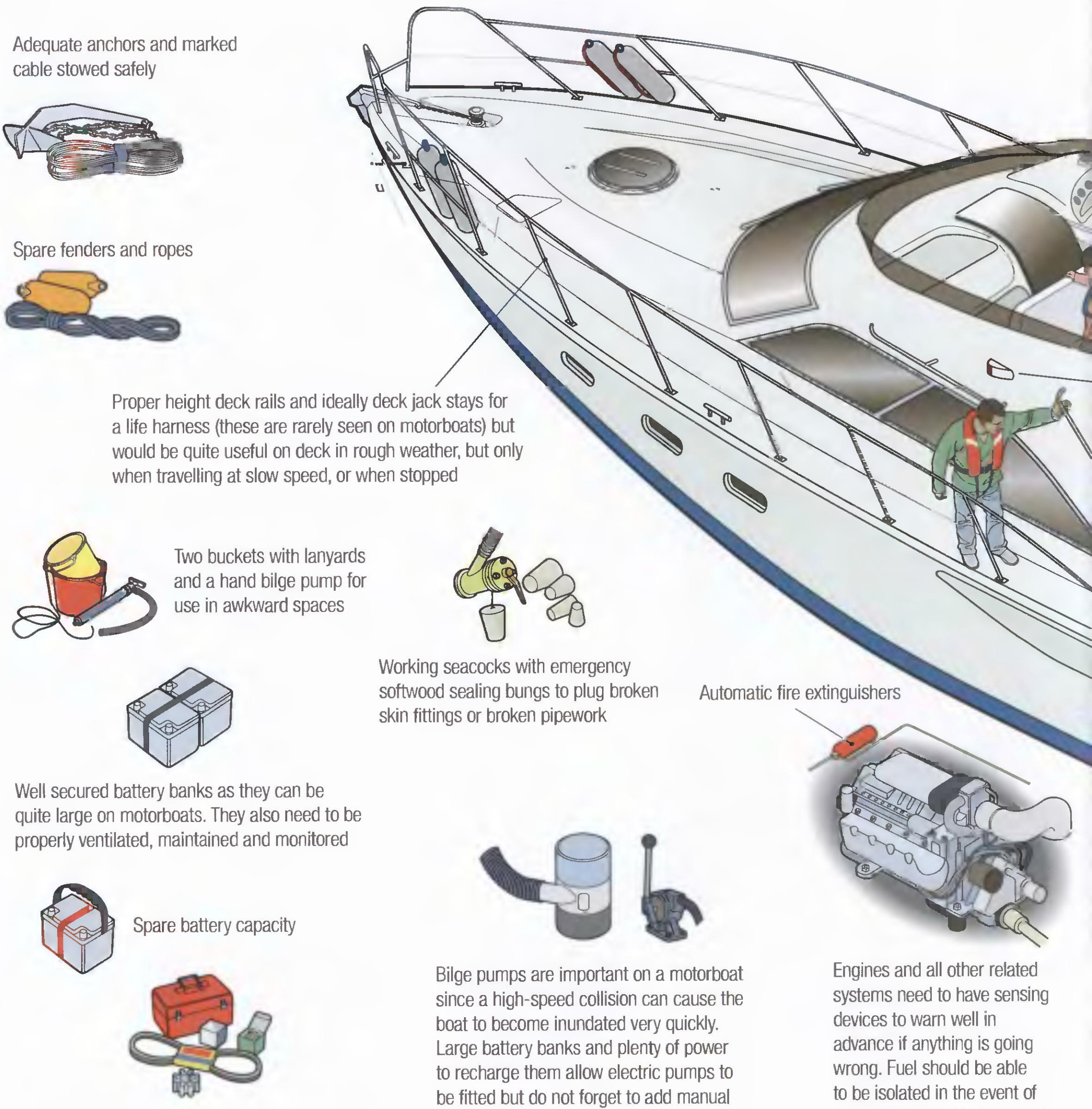

Adequate anchors and marked cable stowed safely

Spare fenders and ropes

Proper height deck rails and ideally deck jack stays for a life harness (these are rarely seen on motorboats) but would be quite useful on deck in rough weather, but only when travelling at slow speed, or when stopped

Two buckets with lanyards and a hand bilge pump for use in awkward spaces

Working seacocks with emergency softwood sealing bungs to plug broken skin fittings or broken pipework

Automatic fire extinguishers

Well secured battery banks as they can be quite large on motorboats. They also need to be properly ventilated, maintained and monitored

Spare battery capacity

Comprehensive tool kit including spare engine parts, spare navigation light bulbs and fuses

Bilge pumps are important on a motorboat since a high-speed collision can cause the boat to become inundated very quickly. Large battery banks and plenty of power to recharge them allow electric pumps to be fitted but do not forget to add manual pump as a backup, in case the electric pump fails

Engines and all other related systems need to have sensing devices to warn well in advance if anything is going wrong. Fuel should be able to be isolated in the event of an emergency with the use of on/off fuel taps in fuel lines

In the UK and Ireland the RNLI offer a free safety equipment advisory service to all boat owners.
www.rnli.org.uk/seasafety

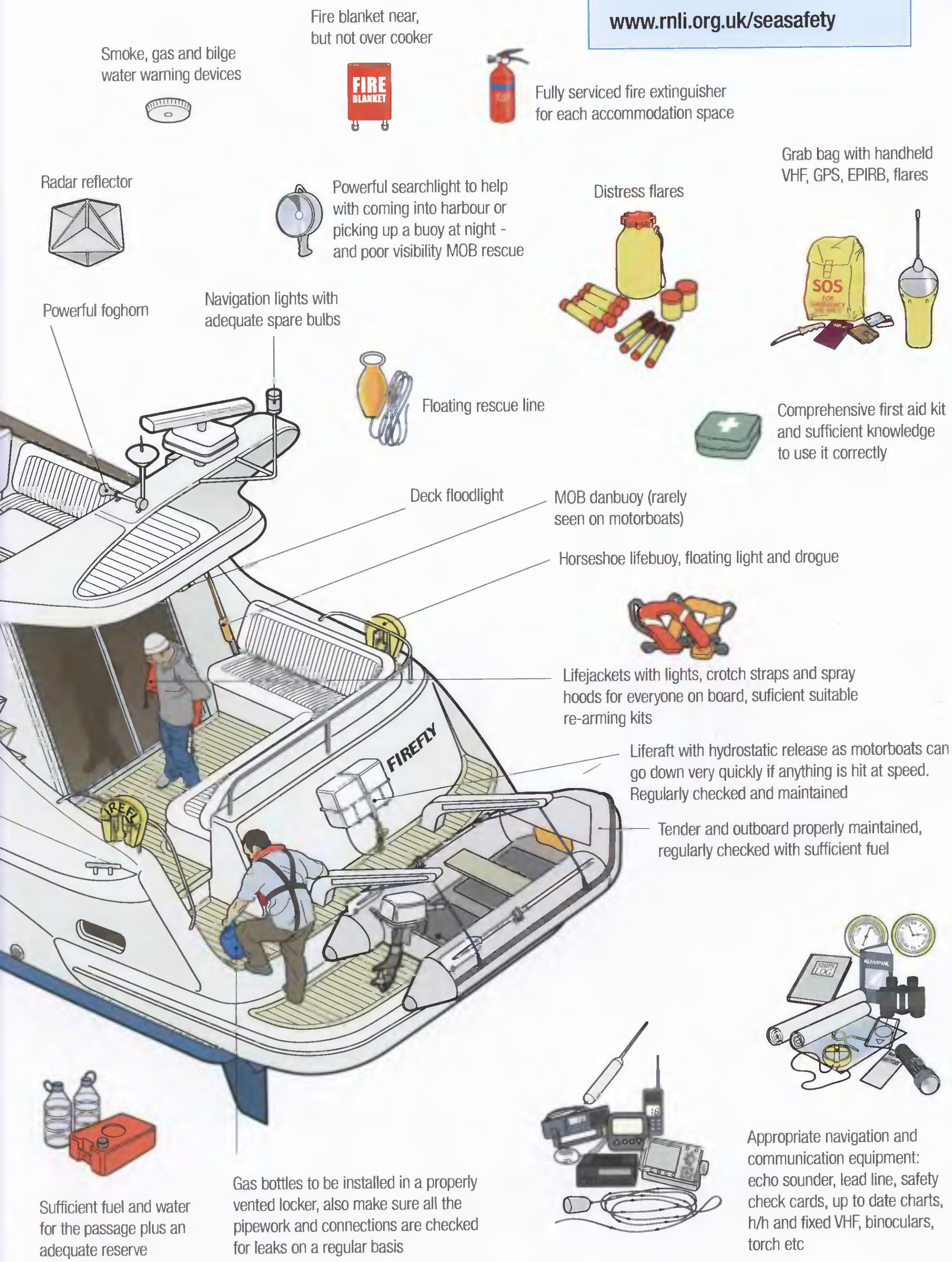

Fire blanket near, but not over cooker

Smoke, gas and bilge water warning devices

Fully serviced fire extinguisher for each accommodation space

Grab bag with handheld VHF, GPS, EPIRB, flares

Radar reflector

Powerful searchlight to help with coming into harbour or picking up a buoy at night - and poor visibility MOB rescue

Distress flares

Powerful foghorn

Navigation lights with adequate spare bulbs

Floating rescue line

Comprehensive first aid kit and sufficient knowledge to use it correctly

Deck floodlight

MOB danbuoy (rarely seen on motorboats)

Horseshoe lifebuoy, floating light and drogue

Lifejackets with lights, crotch straps and spray hoods for everyone on board, suficient suitable re-arming kits

Liferaft with hydrostatic release as motorboats can go down very quickly if anything is hit at speed. Regularly checked and maintained

Tender and outboard properly maintained, regularly checked with sufficient fuel

Sufficient fuel and water for the passage plus an adequate reserve

Gas bottles to be installed in a properly vented locker, also make sure all the pipework and connections are checked for leaks on a regular basis

Appropriate navigation and communication equipment: echo sounder, lead line, safety check cards, up to date charts, h/h and fixed VHF, binoculars, torch etc

Immersion (Survival) suits

Not usually found aboard most cruising yachts, although increasing in popularity with professional yachtsmen and commercial small boat operators, the cold water immersion suit has been designed for larger commercial craft.

Since July 2006, an immersion suit is required for every person on board a cargo vessel that operates in a cold-water area. They are often bulky which is not conductive to easy stowage aboard a small yacht.

There are a broad variety of types from thick neoprene insulated suits to those that are made from PU-coated nylon which are similar to diving or watersports drysuits and require insulated clothing to be worn beneath the suit. Both types usually have insulated hoods to minimise heat loss from the head.

They require practice to be donned quickly.

Unless built-in buoyancy is provided, a 275N lifejacket (see p.33) must be worn to ensure sufficient buoyancy to turn and keep the immersion suit wearer face up.

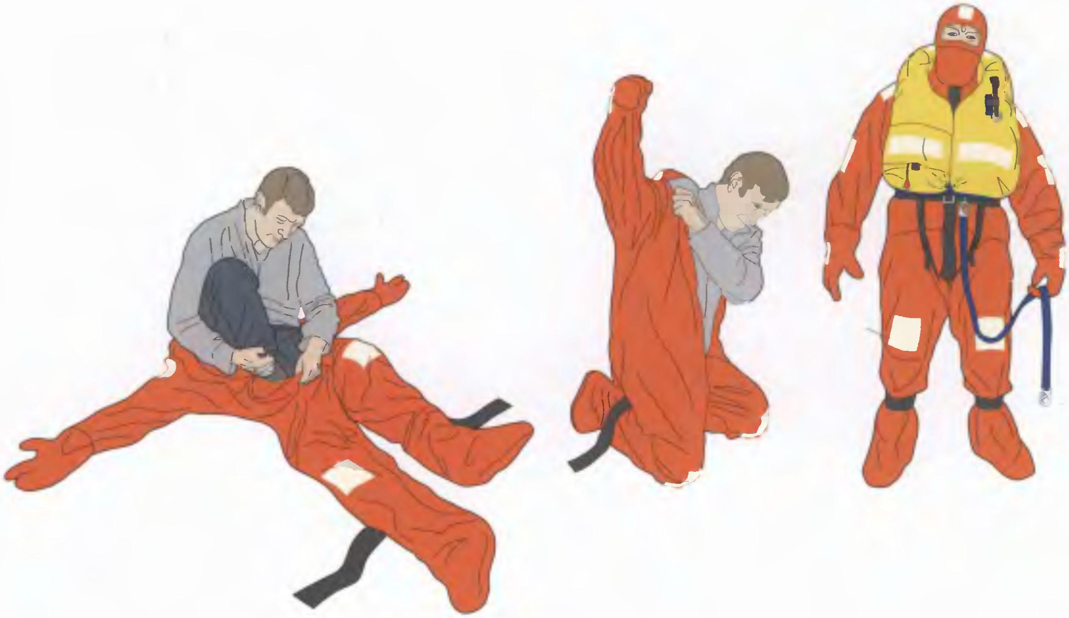

Donning an immersion suit

Care and maintenance of boat and equipment

For the safe operation of a boat at sea, it's essential to inspect and regularly service not only safety equipment but also the boat's machinery, structure and systems. The smallest failure can sometimes be the beginning of a catalogue of problems that can lead to a catastrophe. Prevention is far better than cure.

Follow the example of aircraft pilots and check your boat before departure and put right any problems:

- Do you have enough fuel?
- Does the engine have sufficient oil and water?
- Is the engine's raw-water filter clear of debris?
- Is there water or dirt in the fuel filter/water separator?
- Are the engine mounts secure?
- Is there water or oil or fuel in the bilge – if so investigate where it came from?
- Are the hoses securely clipped to the skin fittings?
- Are the seacocks working smoothly, and are softwood bungs fastened to them in case the skin fitting fractures?
- Is the steering free from stiffness or excessive play?
- Are the bilge pumps operational?
- Are the strum boxes clear of rubbish?
- Are the rigging pins in place at shroud bases?
- Are the lifejackets operational?
- Are the jackstays secure?
- Are the fire extinguishers up to pressure?
- Is the anchor ready?
- Is the liferaft ready?
- Is the lifebuoy ready and the lifebuoy light working?
- Is the dan buoy in place with the flag concertinaed (rather than rolled) so that it deploys quickly?
- Is the handheld VHF charged? Spare charged batteries available?
- Is the fixed VHF working?
- Are all the navigational instruments working?
- Are in-date flares readily to hand?

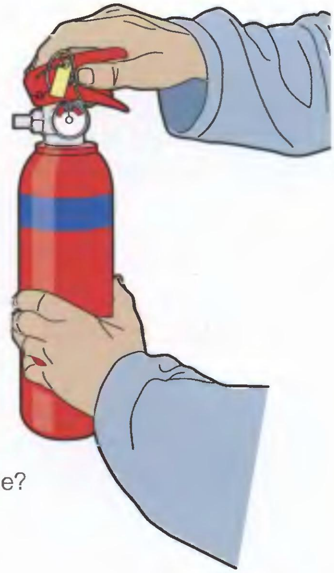

Always carry sufficient spares on board – engine oil, fuel filters, drive belts, water pump impellers, hoses and spare fuses.

Replace any items in the first-aid kit that need replacing. Make sure you have replacement batteries for handheld kit and that you have the right tools on board.

On sailing boats:

- Check the mast for wear or damage.
- Are the spreaders in their sockets?
- Are the navigation lights working properly?
- Are there any loose halyards, damaged rigging wires or bent fittings?

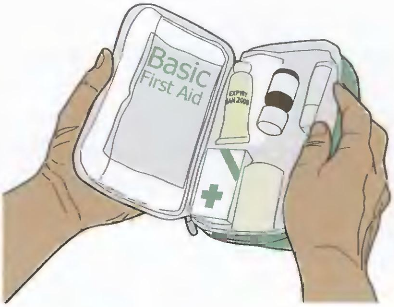

Check the engine instruments. Lift the engine-room and floor hatches and check that all's well. Look around the deck and up the mast. Is there anything becoming loose or that has moved into an unusual position? Listen for unusual noises – they can indicate the beginning of something going wrong.

Remember to continue to make regular checks whilst on passage.

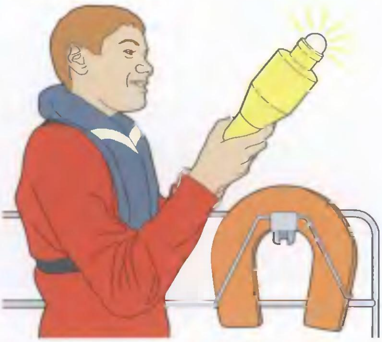

When a skipper is familiar with their boat they are more likely to quickly notice if anything is amiss.

Keep a log of items that need longer-term maintenance, including the dates that equipment such as flares, fire extinguishers, lifejackets and liferaft require servicing or replacing. Regularly check all safety equipment for signs of corrosion and wear and tear and make a note of anything that's likely to need replacing in the near future.

Many items of safety equipment are rarely if ever used. It's easy for them to be forgotten and neglected. However, regular maintenance, servicing and inspections are necessary to ensure the equipment works when you need it.

Engine servicing must not be neglected. Engine failure is one of the main reasons for yachts (power and sail) to founder.

Put the boat's name or call sign on all safety equipment and other items that can float free so it is easily identified if found by the SAR services.

Tip
Add retro-reflective tape to the lifebuoy and danbuoy so that they can be seen more easily at night.

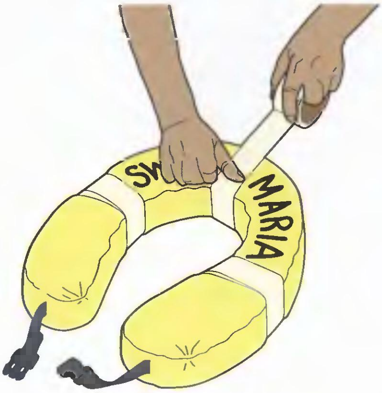

CHAPTER 2 : Weather

Marine weather forecasts are widely available. They are broadcast on public and commercial radio and by coast radio stations, Navtex, the internet and on Inmarsat.

Times for radio broadcasts are shown in an almanac or RYA Weather Forecasts (G5).

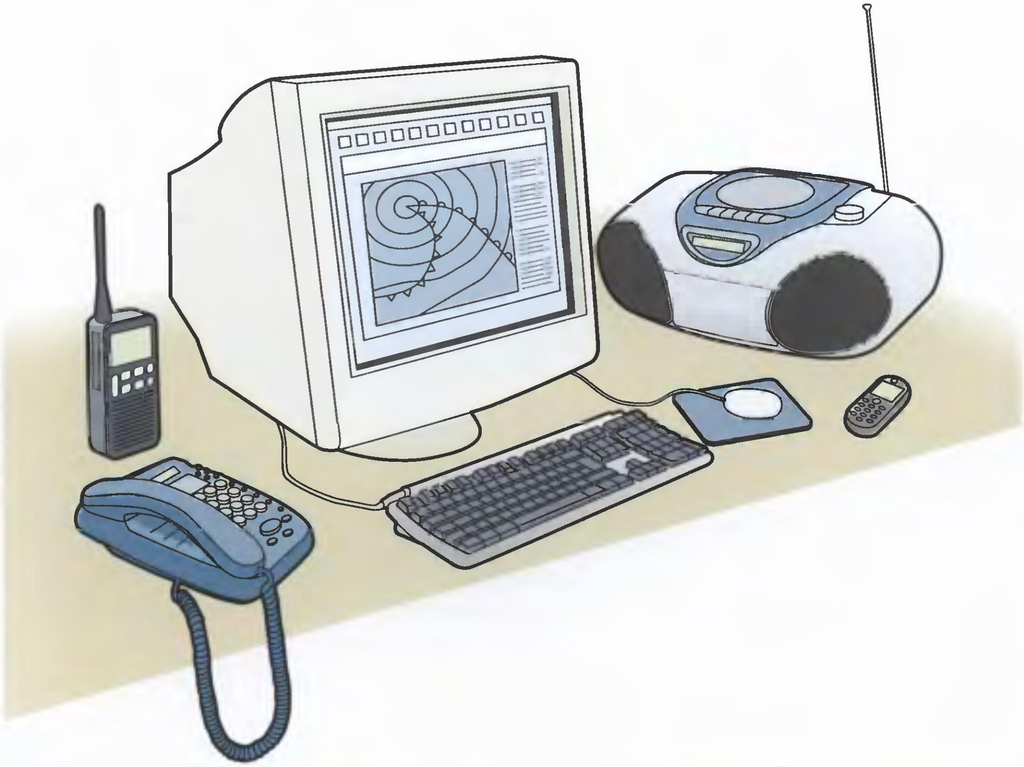

Shipping forecast terms

Gale = Average wind expected to be F8 or more
Strong-wind warning = Average wind expected to be F6 or F7

Timing from the time of issue of warning or forecast

Imminent = within 6 hours , Soon = within 6–12 hours, Later = more than 12 hours

Visibility
Good = greater than 5 miles, Moderate = between 2 and 5 miles, Poor = 1,000 metres to 2 miles, Fog = less than 1,000 metres.
Fair = No significant precipitation
Backing = wind changing in an anticlockwise direction
Veering = wind changing in a clockwise direction
General synopsis = How, when and where the weather systems are moving
Sea states: Smooth = wave height 0.2–0.5m, Slight = wave height 0.5 to 1m, Moderate = wave height 1.25–2.5m, Rough = wave height 2.5–4m, Very rough = wave height 4–6m.

Pressure and tendency
Steady = change less than 0.1mb in 3 hours
Rising or falling slowly = change 0.1–1.5mb in the last 3 hours
Rising or falling = change 1.6–3.5mb in the last 3 hours
Rising or falling quickly = change 3.6–6mb in the last 3 hours
Rising or falling very rapidly = change of more than 6mb in the last 3 hours
Now rising, now falling = change from rising or falling or vice versa in the last 3 hours

Typical depression
In the UK, strong-wind weather is usually determined by Atlantic depressions formed by low pressure. As a depression moves across the country, the warm front is caught by the cold front, creating an occluded front. The lower the pressure in the low, the stronger the winds.

Clouds give a good indication of the approaching weather. High-altitude cirrus clouds give an indication that a depression is approaching. As the warm front approaches, the cloud base descends and pressure falls. Rain becomes heavier with poor visibility, then lightens to drizzle.

Movement of depression

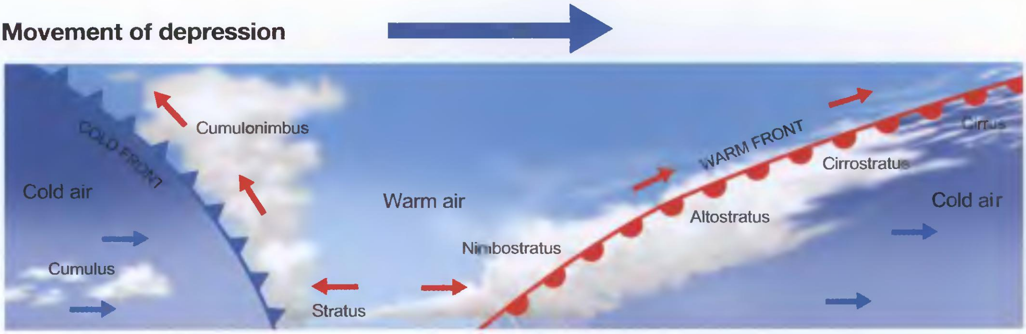

As the warm front passes, the wind veers and becomes steady and the air pressure settles. Visibility is moderate. Low-level fracto and altostratus cloud forms overhead.

A cold front approaches and the wind veers suddenly as the front passes over. Heavy rain, possibly hail and thunder may occur. Visibility is poor.

Behind cold front, the pressure rises then levels. Sunshine and showers, visibility good except in showers. Cumulonimbus and then Cumulus cloud can be seen.

Local winds

Winds often vary from forecast due to local effects.

Wind is slowed by friction with the surface, which is greater over land than sea.

Wind blowing offshore can be variable in direction and strength, especially when blowing off trees, buildings, cliffs, valleys etc.

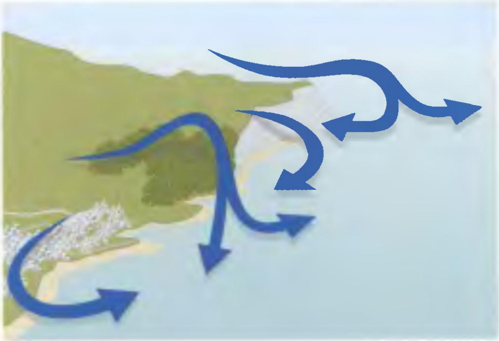

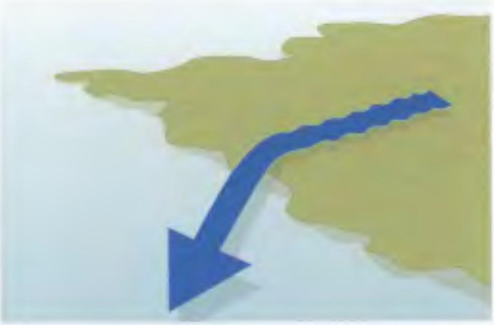

Wind backs more over land than sea. If the wind is blowing off the land, expect it to back as you approach the coast. This may affect the sea state depending on the direction of the tide.

Fog

Advection or sea fog

Created when moist air blows over a colder sea. Usually associated, in the UK, with warm S or SW winds of a depression. Most common in spring when sea temperature is at its lowest.

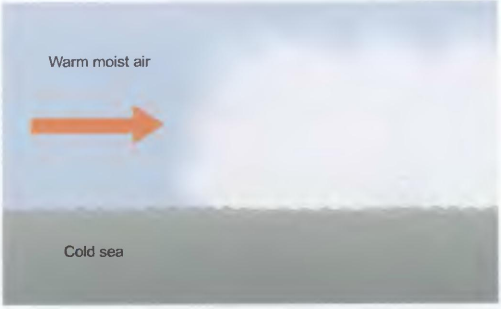

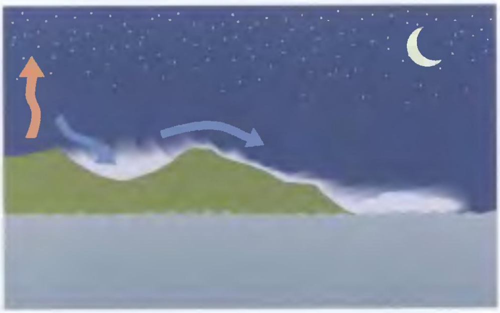

Radiation or land fog

Occurs in settled weather in autumn/winter. Created when land cools quickly at night, causing moisture in the air to condense and form fog over land that will drift out to sea on a land breeze.

Beaufort scale

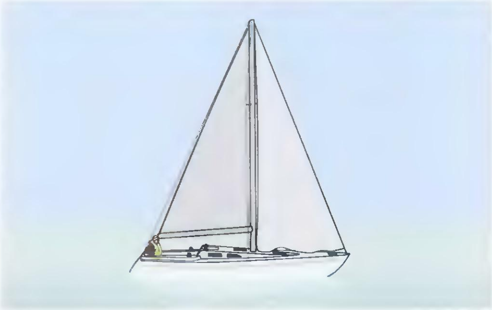

1 Light airs 1–3 knots
Ripples
Sail – drifting conditions
Power – fast, smooth planing conditions

2 Light breeze 4–6 knots
Small wavelets
Sail – full mainsail and large genoa
Power – fast planing conditions

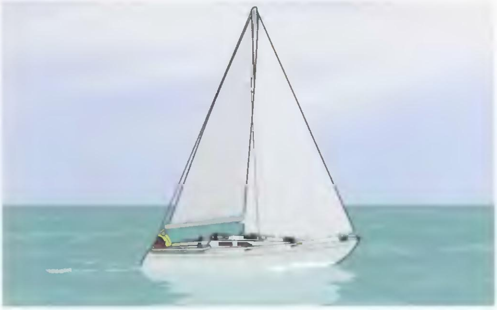

3 Gentle breeze 7–10 knots
Occasional crests
Sail – full sail
Power – fast planing conditions

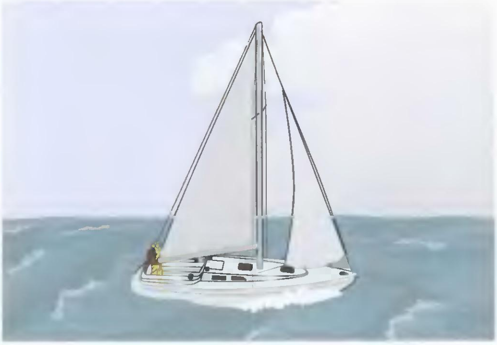

4 Moderate 11–16 knots
Frequent white horses
Sail – reduce headsail size
Power – fast but bumpy planing conditions

5 Fresh breeze 17–21 knots
Moderate waves, many white crests
Sail – reef mainsail
Power – reduce speed to prevent slamming when going upwind

6 Strong breeze 22–27 knots
Large waves, white foam crests
Sail – reef main and reduce headsail
Power – displacement speed

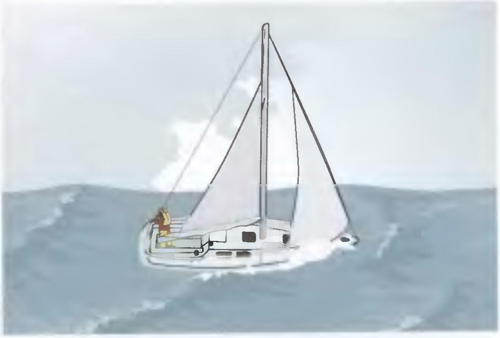

7 Near gale 28–33 knots
Sea heaps up, spray, breaking waves, foam blows in streaks
Sail – deep reefed main, small jib
Power – displacement speed

---

8 Gale 34–40 knots
Moderately high waves, breaking crests
Sail – deep reefed main, storm jib
Power – displacement speed, stern waves

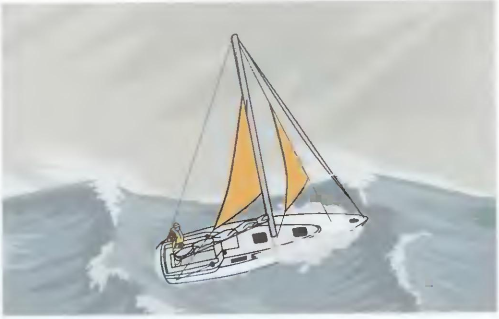

9 Severe gale 41–47 knots
High waves, spray affects visibility
Sail – trysail and storm jib
Power – displacement speed, stern waves

---

10 Storm 48–55 knots
Very high waves, long breaking crests
Survival conditions

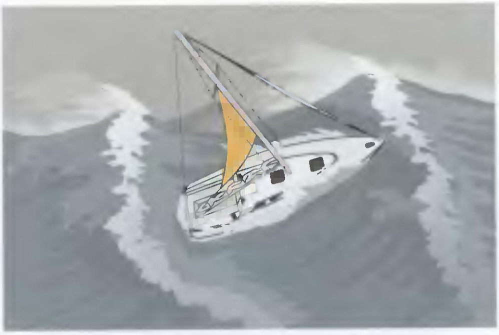

CHAPTER 3 : Preparing for heavy weather

Motorboats are less likely to find themselves caught in heavy weather since most have the speed and agility to find a safe haven quickly or, on longer passages, to route around a storm. Sailing yachts, however, often cannot avoid stormy conditions.

Whether under power or sail the following guidelines are a sensible precursor to heavy weather:

- Stow all gear securely
- Shut lockers and hatches
- Shut and secure portlights
- Stow essentials in plastic bags
- Turn off seacocks to toilet and sink
- Mark fix on chart and secure chart to table
- Update log
- Check contingency plans and, if not already done so, add appropriate waypoints to GPS
- Prepare contingency passage plans
- Sail away from the lee shore
- Print or write out weather forecast times and fasten to the chart area
- Organise a crew roster. Rest the strongest and most able crew first
- Eat a meal and prepare sandwiches and flasks of hot drinks
- Drink plenty of water
- Inform the Coastguard of your position and planned course
- Take anti-seasickness tablets
- Charge engine batteries
- Check bilges and ensure bilge pumps are working
- Check and wear safety equipment
- Secure cockpit lockers to prevent down-flooding
- Secure washboards in companionway
- If appropriate, shut watertight doors
- Stow loose items off deck
- Prepare and secure anchor locker
- Check sea anchor and heavy warps
- Rig and check jackstays
- Close off deck ventilators
- Check cockpit drains are clear
- If required, rig mainsail reefing pennant to third reef
- Prepare trysail and storm jib

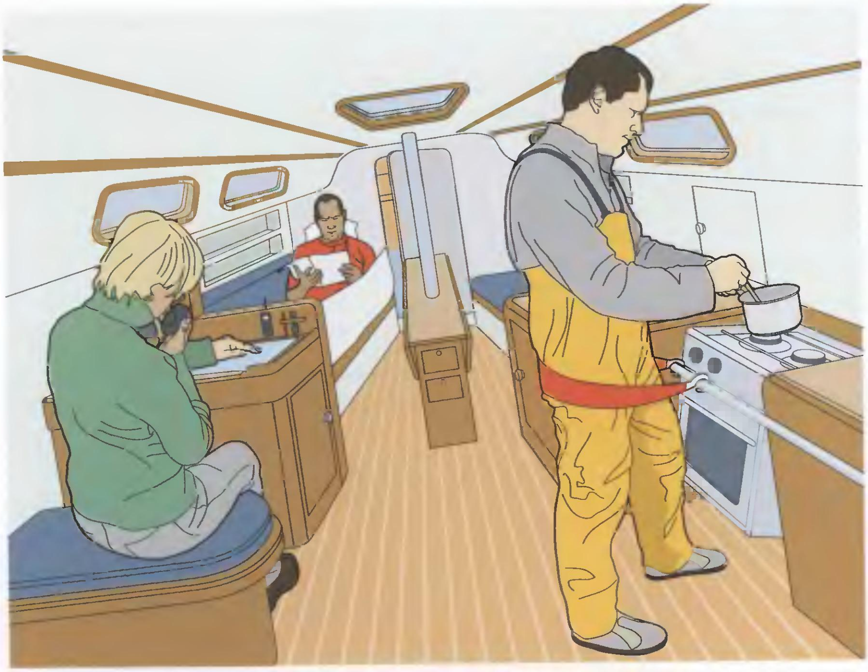

Good seamanship - Prepared for heavy weather

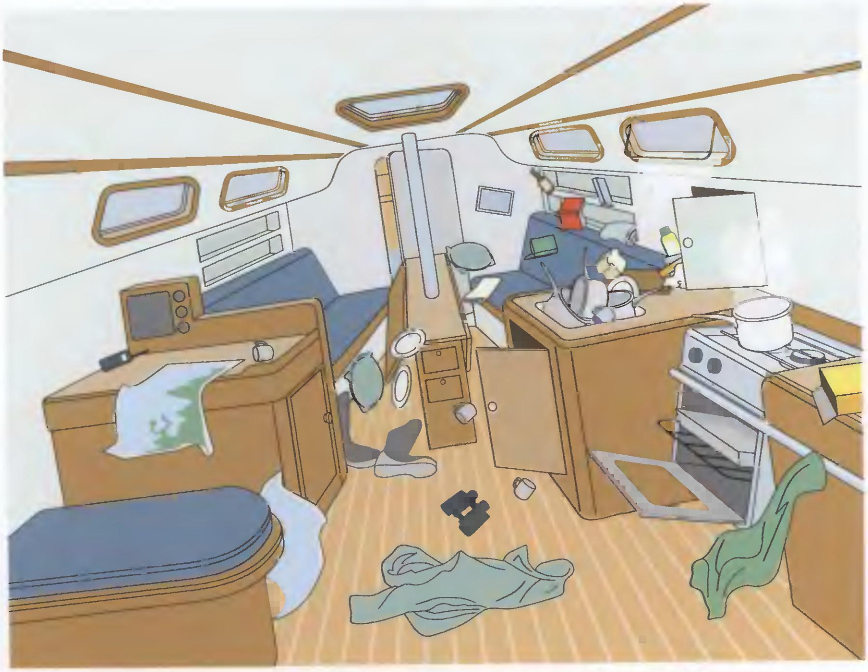

Bad seamanship - Poorly stored gear after heavy weather

Shortening sail

As the wind strengthens you will need to shorten sail. Always shorten sail before you have to because it's easier to shake out a reef if the conditions allow than to take one in when the wind has increased.

The order of shortening sail will depend on your boat and rig. In general, the principle is to shorten sail by gradually reducing the headsail and mainsail so that the centre of effort does not move too far forward or too far aft. This will keep the boat balanced and ensures sufficient weather helm is maintained.

Note that on some boats with slab or jiffy reefing only two reefing pennants are run. This requires the first reefing line to be reattached to the third or uppermost reefing cringle. Three reefs will reduce the mainsail by up to 60% of its total area.

While roller headsails can reduce sail area completely, a partially reefed roller headsail is very inefficient upwind and will not provide the proper balance.

Storm sails

If wind conditions increase, you will need to resort to storm sails comprising of a storm jib and trysail. These lower the centre of effort and may help a yacht to be sailed off a lee shore. They are made from heavy-duty material usually coloured bright-red or luminescent orange so that other shipping can more easily see the boat. A storm jib is best set on an inner forestay, this can be a rigged temporarily or permanently.

A trysail is a loose-footed sail. Its sheet is led aft either to a spinnaker pad eye or to a stern mooring cleat. Occasionally, it may be attached to the boom but this is not recommended.

Depending on type, a trysail can allow for the mainsail to be furled and secured to the boom. The boom can then be dropped and secured to the deck. A trysail can be hoisted in the mainsail's track or preferably in its own track. It must be securely fastened.

Both sails should be pre-rigged with their own sheets.

> Since the rigging of storm sails differs from that of ordinary sails, it's important to practise rigging them in fine weather rather than waiting until you really need them. Rig them in advance of a storm.

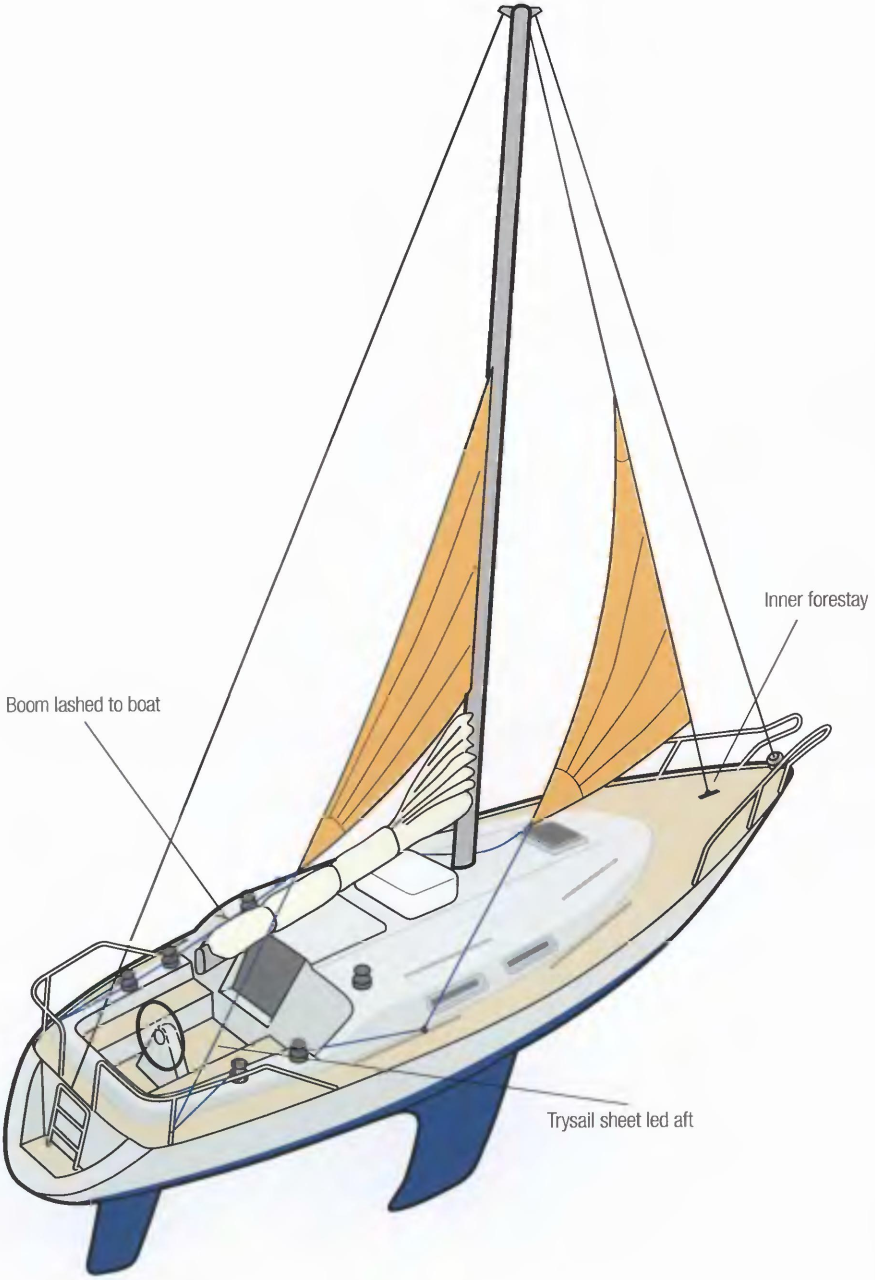
Boom lashed to boat
Inner forestay
Trysail sheet led aft

Rigged for heavy weather
Storm jib is best set on an inner forestay while the trysail replaces the mainsail.

Breaking waves

Waves are generated by the frictional effect of wind on the surface of the water. The stronger the wind, the longer it has blown and the greater the fetch (the distance to shelter from windward), the bigger the waves.

Very steep waves are formed when:

- the wind strength increases rapidly
- wind and current oppose each other – wind against tide
- waves are coming from different directions due to a sudden wind shift
- water becomes shallow

Waves break when they reach a critical steepness and gravity can no longer sustain their shape.

Compared to their size, breaking waves release enormous amounts of energy and they are often sufficient to knock down and capsize a boat.

When caught beam-on, a breaking wave the same height as the beam of the boat is enough to turn the boat upside down.

Owing to the boat's reduction in moment of inertia, a dismasted yacht is often more easily rolled by a breaking wave than one with a mast.

Avoid areas of breaking waves whenever possible.

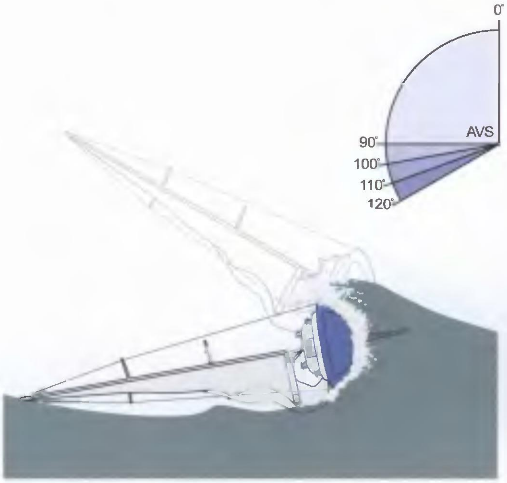

The Angle of Vanishing Stability will depend on the design of the boat. Many will not right themselves once inverted to 120°

A breaking wave the same height as the beam can capsize the boat.

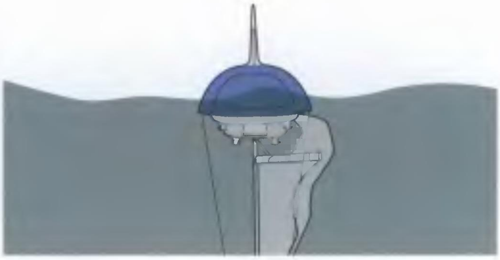

Many boats will stay upside down after capsize.

CHAPTER 4 : Handling heavy weather

Your plan for handling heavy weather needs to take into account:

- capability of the boat
- capability of the crew
- severity of the expected weather conditions
- proximity to land

There are two basic strategies for handling heavy weather:

1. run for shelter
2. ride it out at sea

Running for shelter

Choose a haven that has a sheltered entrance, waves get bigger and steeper as they approach shallow water. An opposing tide will also increase wave size and steepness. Bigger waves travel faster and are more likely to break. When approaching a harbour entrance running with the waves, the increasing size of waves can be deceptive. Look behind you for waves catching you up. Good shelter and a wind with tide can make a flatter sea close in. Radio the Coastguard and harbour master for advice and to tell them of your intention.

Riding it out at sea

Riding it out at sea needs room. There are several methods:

Going to windward (sailing craft)
A good method in strong winds, but only if the crew are relatively strong, is to close reach across troughs, luff (steer) into and over crests and bear away hard to keep the leeward side in the water and avoid a slam. Keep steerage in the troughs or the yacht could end up beam-on to the next wave. Not a method to be used if the crew are tired, in darkness or strengthening winds.

Another option for yachts that do not behave well in stormy winds is to sail slowly upwind. Heavily reef the mainsail and set it as flat as possible and as tightly sheeted in as possible so that it acts as a stadying sail. The headsail is furled. A ketch or yawl would be better using the mizzen rather than the mainsail. Adjust the boat's speed so that it is just moving forward, speeding up to miss breaking waves.

Heaving to
Reduces strain on the yacht and crew by riding waves more easily, making little or no headway but considerable leeway. Boats vary in their ability to heave to, if properly balanced, a yacht will hold its position at about 60 to 70 degrees off the wind and will leave a slick of calm water to windward.

The method is to tack but keep the headsail backed. Hold or fasten the helm to leeward. Some modern designs are difficult to heave to since their efficient rigs continue to drive them forward. However, yachts can heave to under storm sails and have a steadier motion.

Going to windward (powered craft)

Motor cruisers can drive straight into head seas but it will be very uncomfortable for those on board, a better way uses a similar approach as sailing yachts. Rather than heading directly into the waves, dog-leg or tack across the waves, look for gaps between the breaking crests to head into and over the wave. If possible, trim the bow down. Use the throttle continuously and carefully to prevent the boat from flying off the wave or from digging into the trough.

Going downwind (sail craft)

Running before the wind will dramatically reduce the apparent wind speed and greatly reduce slamming.

However, the mainsail boom may dig into the water and slew the yacht around, so using a trysail and a storm jib – or either on their own – may be a better option.

Going downwind (powered craft)

Trim the bow up and position the boat on the back of a wave. After the wave breaks, drive through and on to the back of the next wave.

Big waves travel quickly, so keep a good lookout behind for waves catching you up.
Keep a hand on the throttle and control the boat speed carefully. You may need to slow down to prevent breaking through the wave and then speed up to get on the back of the next wave.

Drogues at sea

Drogues can be used on power or sail craft. They are attached to the stern of the boat and help prevent the yacht from being slewed sideways and rolled over by breaking waves in a following sea.

Use enough length of line – 100 metres or 10 x the boats lenth – to stop the drogue from breaking free of the water and at least two to three times the wave length. Because of the danger of being pooped by waves breaking over the stern, ensure that hatches, lockers and the companionway are secured.

If a drogue is not available, a length of heavy warp either in a line or a loop can be used.

Spread the load over several cleats and pad areas that are likely to chafe.

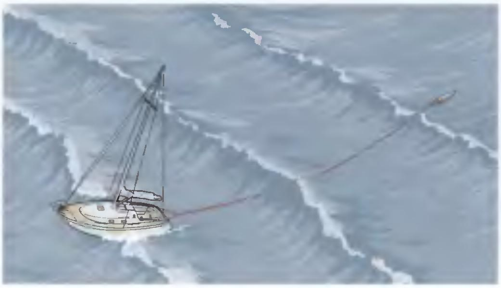

Drogues help prevent broaching and twisting sideways onto the waves.

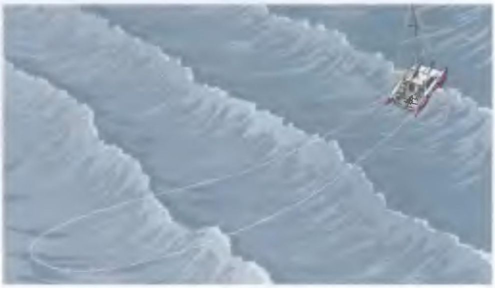

A long loop of trailing warp can also be used in a following sea.

Sea anchors

Deployed from the bow, and similar in size and shape to a parachute, a sea anchor keeps the bow head-to-wind and waves and reduces the possibility of capsize or pitch-poling. Sea anchors can be used on all types of craft but are particularly useful for catamarans and motor cruisers, which would otherwise turn their sterns towards the waves.

A sea anchor for a 10-metre yacht would be about 4 metres in diameter. As with a ground anchor, nylon warp or chain and warp will reduce snatching. A sea anchor should be deployed with sufficient length of warp to ensure it stays submerged, usually about two to three wave lengths, which may require about 100 metres of warp.

As with drogues and when using the normal anchor, spread the load over several strongpoints and pad it in areas likely to chafe.

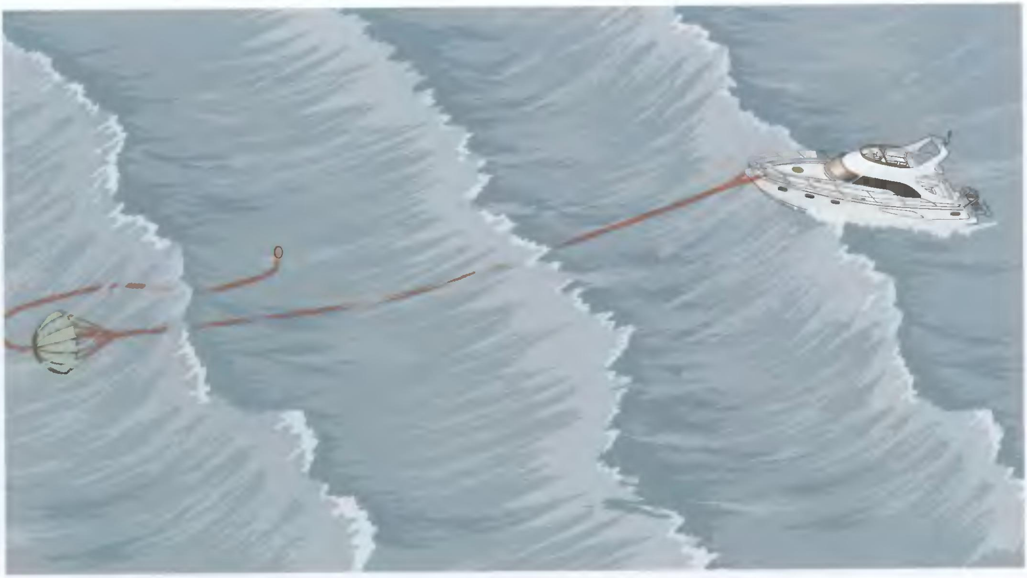

Damage control and repair

Your boat is your best liferaft. Unless it is about to sink, or is on fire, stay with it. It provides the best shelter, a large target for rescuers to see, and contains more provisions and equipment than a liferaft. Be prepared and improvise to keep your boat afloat.

The 1979 Fastnet Race

Winds reached Force 11 during the 1979 Fastnet Race. Over 300 yachts started the 605-mile race to the Fastnet Rock and back. Confused seas added to the ferocious conditions. Of the twenty-four yachts abandoned only five sank. Sadly, fifteen people lost their lives, seven after abandoning to their liferafts.

Steering failure

In heavy seas damage can be sustained to the steering system up to and including complete loss of the rudder, so all boats should carry an emergency steering system. If there is damage to the steering system, try to attach a temporary emergency tiller to the top of the rudder stock, or jury-rig a rope and blocks to the rudder quadrant.

Twin-engine boats can steer by using the engines separately. Auxiliary outboard powered boats can use the engine to steer and if fitted, a bow thruster may be able to point the boat in the right direction.

If the rudder is lost

A pre-drilled locker cover or washboard can be fastened to a spinnaker or whisper pole using U-bolts or lashings to create a steering oar or sweep. Use guys from winches to assist in steering and to stop the rudder floating.

Help in steering may be possible by dragging a drogue or bucket on a bridle to one side or other over the stern.

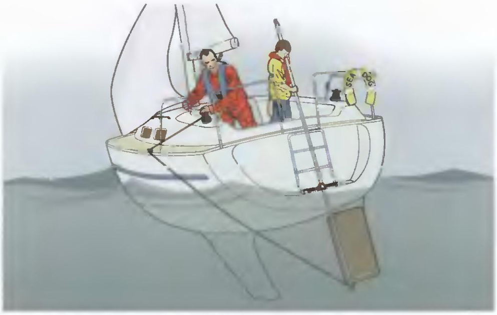

Jury rig a steering oar using the spinnaker pole and washboard or cabin locker cover.

Broken mast

Try to prevent mast breakage by keeping a close eye on standing rigging. The windward rigging is under tension and more likely to break – watch out for parting strands and try tacking to reduce tension on the shroud. Use a halyard to reinforce the shroud. A spinnaker pole may give a better lead. Watch out for leeward shroud turnbuckles that may have vibrated loose before tacking.

If the mast comes down

Clear away any broken spars quickly since they may hole the boat. Use cable cutters and a hacksaw (before going out ensure you have plenty of blades), a cartridge cutter or hydraulic cutter will cut rod rigging or rigging wire of over 10mm diameter. Try to keep the spars for jury-rigging a temporary mast and sail. The motion of the boat will change and the boat will become more susceptible to breaking waves. Advise approaching rescuers if any lines or spars are still in the water.

Jury rig

Use lateral thinking to jury-rig a temporary mast. Bulldog grips can be used to shorten and secure rigging wire and mainsheets, handybillys and vangs used to hoist or tension spars. Recut the sails if necessary.

Emergency VHF antenna

When the mast breaks it is likely the radio antenna will be damaged. Always carry an emergency VHF antenna to ensure communications.

Hatch damage

To seal a broken hatch use locker lids tourniqueted into position with a boat hook or deck brush. Better still, have pre-drilled hatch and portlight covers ready for emergency use.

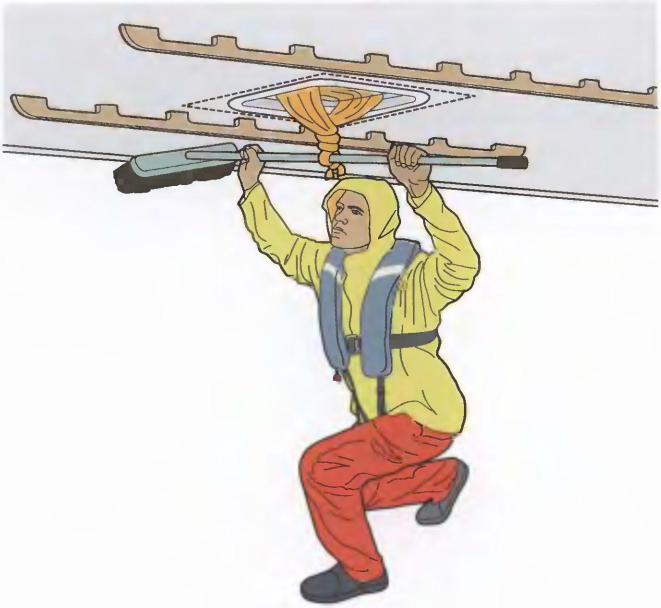

Hull damage

A relatively small hole can let in a large amount of water in a very short time. The amount of water will depend on the size of hole and its depth below the waterline. For example, a 75mm (3in) diameter hole about 30cm (1ft) below the waterline will let in about half a tonne of water in a minute, too much for bilge pumps to cope with.

Reducing the area of the hole even by partially blocking it will significantly reduce the flow rate. Block the hole with any materials you have available, braced against the opposite side of boat.

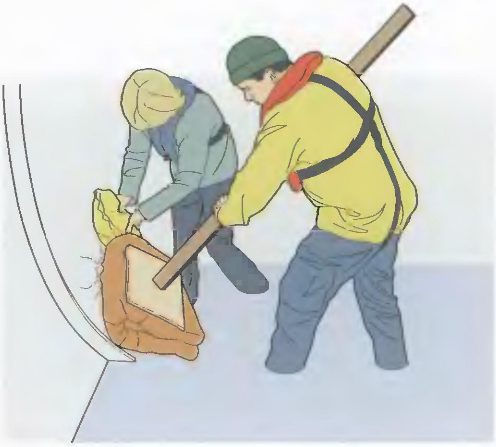

An alternative method is to cover the hole from outside using a sail or other waterproof fabric cover. Whether this is possible will depend on the position of the hole, the shape of the hull and the weather conditions.

Make sure the bilge pumps are able to reach all parts of the hull. Electric pumps will only continue to pump so long as they have electricity. Engine-driven pumps can shift considerable amounts of water. Ensure manual pumps are operational. All pumps should be fitted with strum boxes (filters) to prevent blockage.

Attach an appropriate-sized softwood bung to each seacock. If the skin fitting or hose breaks and a seacock has seized, the bung can be used to stem the flow.

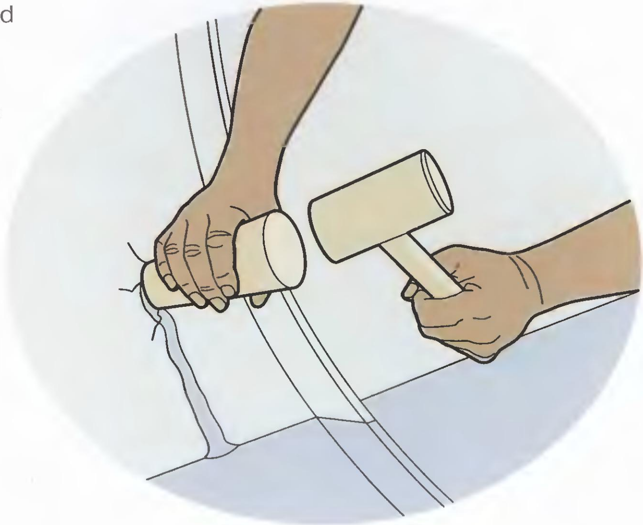

CHAPTER 5 : Lifejackets and Buoyancy Aids

There are a broad range of lifejackets available on the market from stiff buoyant foam models suitable for merchant shipping, to the inflatable bladder types that you will see demonstrated on commercial airlines. Here we have restricted our comments to those commonly available to the boat user.
All new lifejackets sold in Europe must have a CE mark. Other countries often apply their own (similar) standards. An International ISO standard has recently taken effect and will supercede the CE standards as new designs are introduced. The current European standards give buoyancy ratings in Newtons (N):

EN 393 (ISO 12402-5)	50 N buoyancy aid
EN 395 (ISO 12402-4)	100N lifejacket
EN 396 (ISO 12402-3)	150N lifejacket
EN 399 (ISO 12402-2)	275N lifejacket

Buoyancy Aids

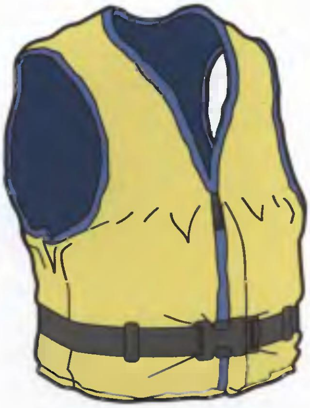

Buoyancy aids are not recommended as life-saving equipment.

- **50N buoyancy aids** are designed for competent swimmers as an aid to flotation in inshore waters where help is close at hand. They are suitable for those who expect to fall into the water, for example dinghy sailors, waterskiers, kayakers etc – allowing them to swim easily back to their craft. This type is comfortable to wear but will not turn the wearer face-up in the water. It provides 5kg (11lb) of buoyancy from, usually, one or several layers of closed-cell foam.

Sold in several sizes from child to XXL, buoyancy aids need to be a snug fit to work properly. A simple test is to try lifting the jacket at the shoulders. If it lifts more than 50mm (2 inches) it's too big.

Lifejackets

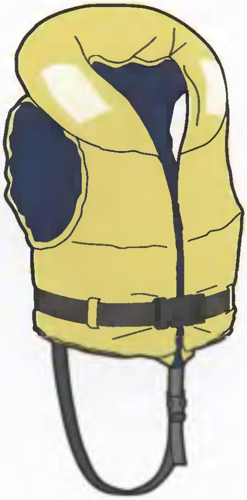

- **100N lifejackets** are designed to keep non-swimmers afloat but will not necessarily turn an unconscious wearer face-up. Usually made from one or several layers of closed-cell foam and provides 10kg or 22lb of buoyancy. Like buoyancy aids, they are sold in several sizes from child to XXL and need to be a tight fit to work properly. Check for fit by lifting the jacket at shoulders. If it lifts more than 50mm (2 inches) it's too big.

- **150N lifejacket** is the standard lifejacket, providing a minimum buoyancy of 15kg (33lb). Suitable for swimmers and non-swimmers and is arguably the best type for offshore boats. Although this type is designed to turn an unconscious wearer face-up, if there is any air trapped in the wearer's waterproof clothing this may prevent the jacket from righting the person immediately.

  The lifejacket will give a reasonable chance of survival but will hamper movement when fully deployed. Most jackets of this size are gas-inflated; however, bulky all-foam or foam-and-air jackets are also available.

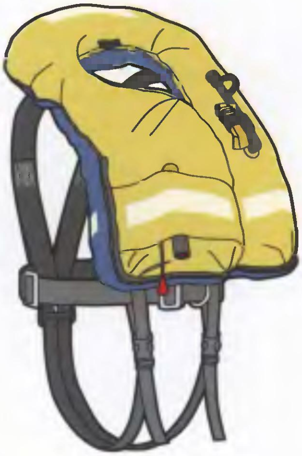

- **275N lifejacket** was originally intended to provide sufficient buoyancy to self right an offshore oil worker wearing an immersion suit and is a high-performance lifejacket, suitable for severe offshore conditions. Usually gas-inflated and consisting of one or, occasionally, two separate bladders, it has sufficient buoyancy – 27.5kg (60.5lb) – to counteract trapped air in waterproof clothing. Sadly, its large deployed size can make it difficult for the wearer to climb into a liferaft.

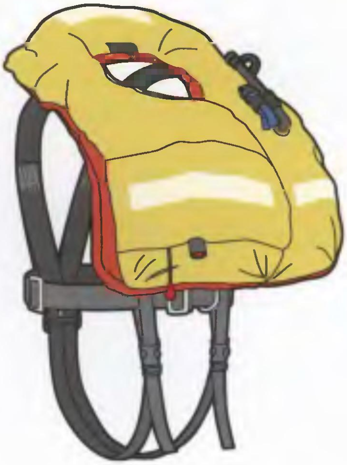

Children's lifejackets have equivalent buoyancy ratings to adults' lifejackets, for example, 50N and 150N, but with reduced buoyancy to suit their smaller size. Most have harnesses and crotch straps built in as standard. It's essential that children wear the correct size lifejacket or buoyancy aid. Don't buy a lifejacket that is too big thinking they will grow into it.

Five top lifejacket tips

- Carry one lifejacket for each member of the crew
- Carry spare lifejackets for any casualties you may pick up
- Always check your lifejacket before donning
- Lifejackets require regular maintenance
- Your lifejacket will only work properly if it is worn correctly

Gas lifejackets

Gas lifejackets are inflated with carbon dioxide (CO_2) stored under pressure in a sealed metal bottle. The gas is released by piercing the seal, either manually by pulling the toggle or automatically on immersion in water.

If fitted with an automatic firing head, it is essential to fit the correct size bottle for the size of lifejacket. Too big a bottle may burst the jacket; too small a bottle will not fully inflate the bladder. The correct size bottle will fully inflate the jacket in under five seconds.

- A lifejacket has to be worn outside of clothing. If it inflates underneath a waterproof jacket, the wearer's chest will be squeezed making it impossible to draw breath.
- Make sure that the manual-inflation toggle is always available – do not tuck it inside the jacket.

Types of inflation

Oral-only inflation (blow air into)
Inflatable lifejackets that can only be inflated through a top-up tube are not recommended. Due to the physiological effects of falling into cold water, you will not be able to inflate the lifejacket.

Manual gas inflation
It will not inflate automatically, the jacket is inflated by the wearer pulling on the toggle. It will not inflate automatically. The firing mechanism should have some means of indicating if the jacket has been fired – often it is as simple as a small green clip over the toggle lever. If this is not present, unscrew and check that the seal on the bottle has not been pierced - if in doubt rearm.

Always carry spare gas bottles and clips so that the lifejacket can be immediately rearmed.

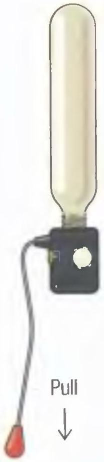

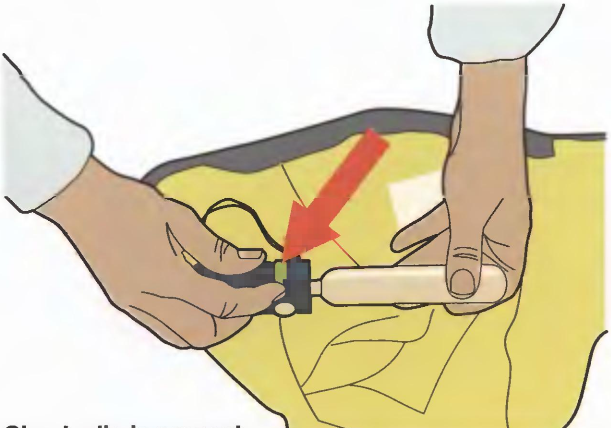

Check clip is present

Automatic gas/cartridge inflation

There are several different designs of automatic gas-inflation firing heads that give various degrees of protection to the cartridge or pill. Automatic operation uses a spring-loaded plunger held in place by either a compacted paper cartridge or a pill. When the cartridge or pill becomes wet it expands or dissolves, releasing the plunger so that it pushes the pin through the gas-bottle seal.

In addition to a manual indicator, automatic firing mechanisms have a separate indicator to show they have not been operated. When donning the lifejacket, check the automatic firing head and the manual indicators. If one or other is not present, unscrew and check that the seal on the bottle has not been pierced.

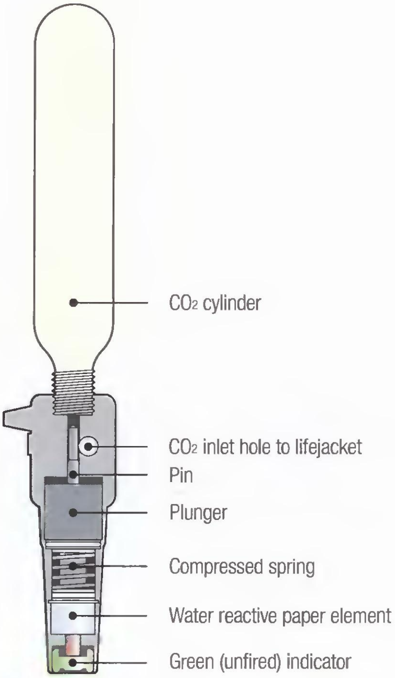
CO$_2$ cylinder
CO$_2$ inlet hole to lifejacket
Pin
Plunger
Compressed spring
Water reactive paper element
Green (unfired) indicator

While some lifejackets use cartridge style autoheads others have a simple replaceable bobbin. Although different makes of lifejackets may appear to have identical firing heads they are often not interchangeable. Some automatic firing heads can fire accidentally if they become excessively wet. Keep appropriate spare automatic firing heads and gas bottles on board so that every lifejacket can be immediately rearmed.

Standard automatic heads have a limited life and will be marked with an expiry date.

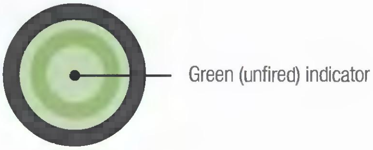
Green (unfired) indicator

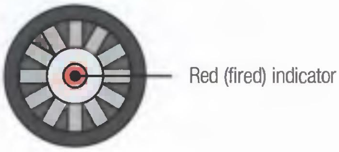
Red (fired) indicator

Hammar automatic inflation

The Hammar system protects the water-sensitive element with a hydrostatic valve that prevents accidental activation by wave splash, heavy rain or spray. The valve opens after two to three seconds' immersion at a depth of 100mm (4 inches). This allows water to reach the water-sensitive element, which dissolves and triggers the pin to pierce the bottle.

When the jacket is fired manually, the toggle detaches from the firing head. The firing head will show a green indicator if it has not been fired or a red one if it has. Hammar automatic inflation heads have a five-year life. The replacement date is printed on the firing head. Keep spare rearming kits on board so that the jacket can be rearmed immediately and check they have not passed their expiry date. Inflate the lifejacket after re-arming, to check the head has sealed properly on to the fabric.

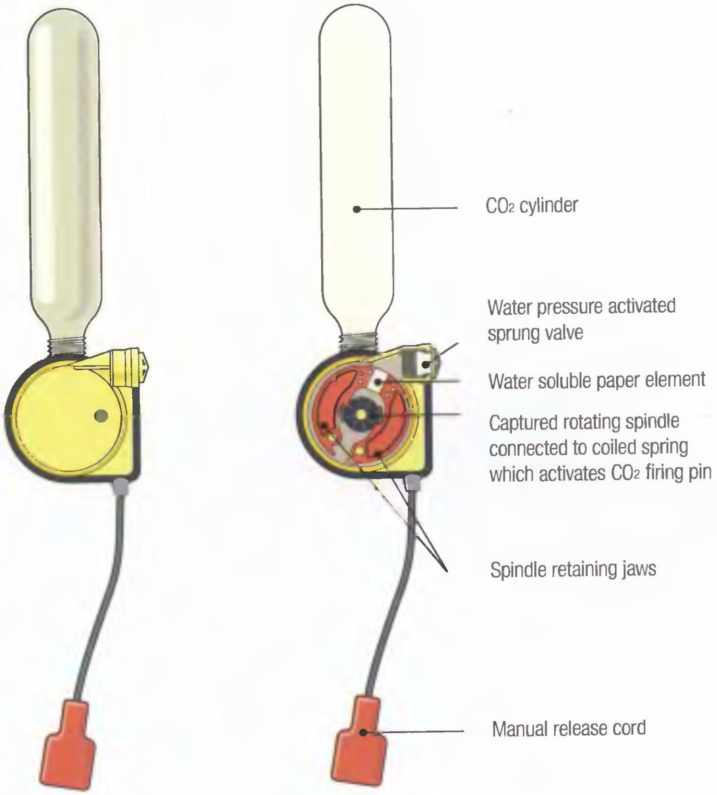
CO_2 cylinder
Water pressure activated sprung valve
Water soluble paper element
Captured rotating spindle connected to coiled spring which activates CO_2 firing pin
Spindle retaining jaws
Manual release cord

Lifejacket maintenance and checks

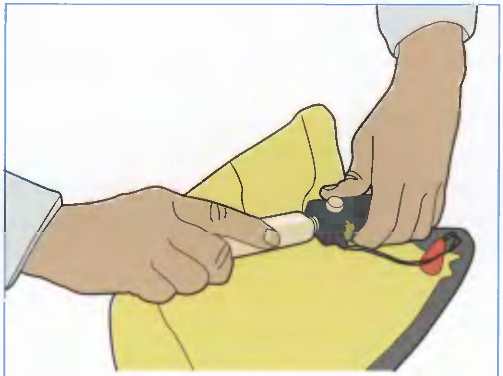

Regularly check that the gas bottle is securely tightened into the firing mechanism. If it is loose, it may not fire and inflate the jacket.

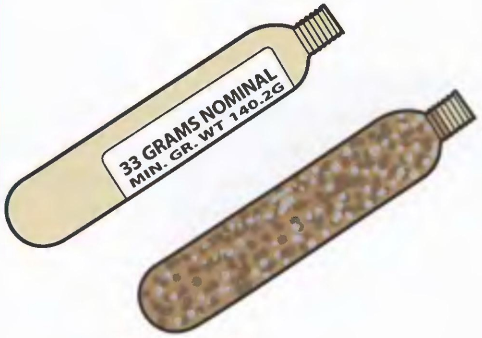

Inspect the gas bottle for corrosion, which can cause small pinholes in the lifejacket bladder. Check that the bottle is the correct capacity for the lifejacket.

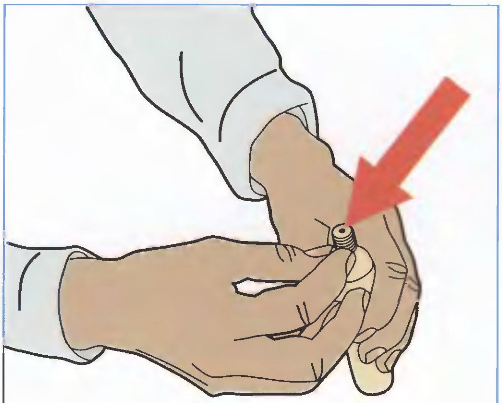

Unscrew the bottle and check the seal is still secure. If it has a hole in it, fit a replacement bottle.

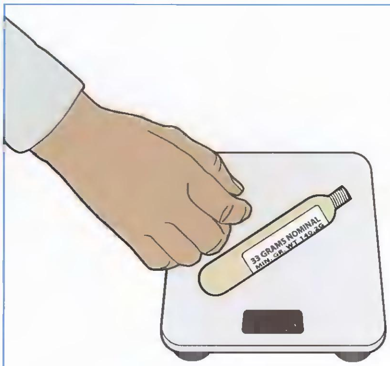

Weigh the bottle on a digital scale to ensure it matches its gross weight as embossed on the bottle. A 33g bottle has a gross weight around 140g.

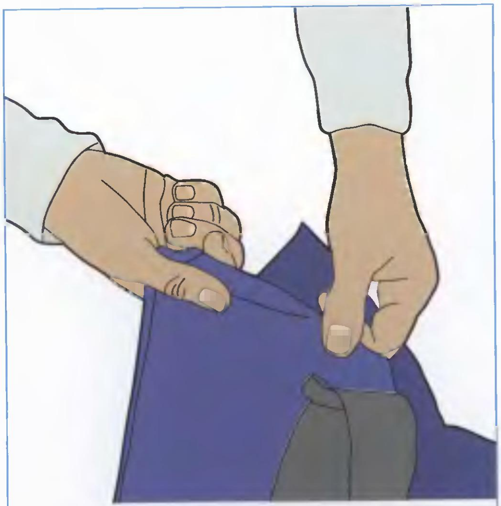

Unfold the lifejacket and inspect the cover and bladder for punctures, cracks and wear and tear.

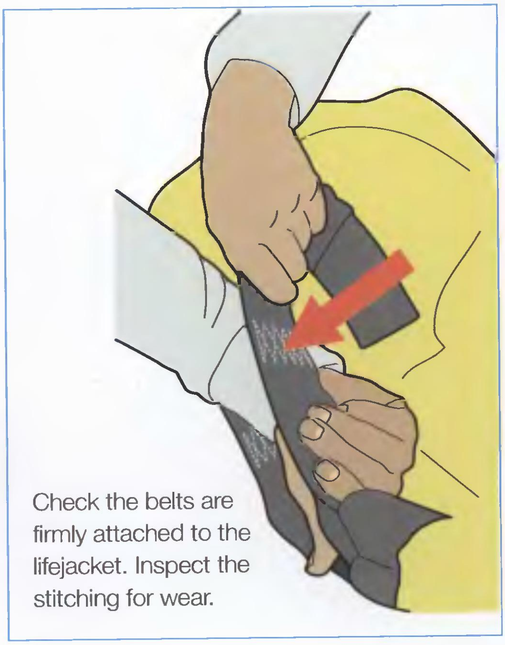

Check the belts are firmly attached to the lifejacket. Inspect the stitching for wear.

To deflate, push in the valve in the oral inflation tube while squeezing the jacket. Make sure all gas is removed before reattaching a new gas cylinder and repacking.

Every three to six months

- Remove the gas bottle and inflate the jacket, preferably with an air pump, through the oral inflation tube.
- Leave the jacket inflated for 24 hours to check that it is airtight.
- When not being used regularly, unfold the jacket and store it on a wooden clothes hanger.

Tip

Not all manufacturers air-test their lifejackets at every appropriate stage in its manufacture, so it's a good practice immediately after purchase to inflate all new jackets, preferably using an air pump through the oral tube, to ensure the bladder does not leak.

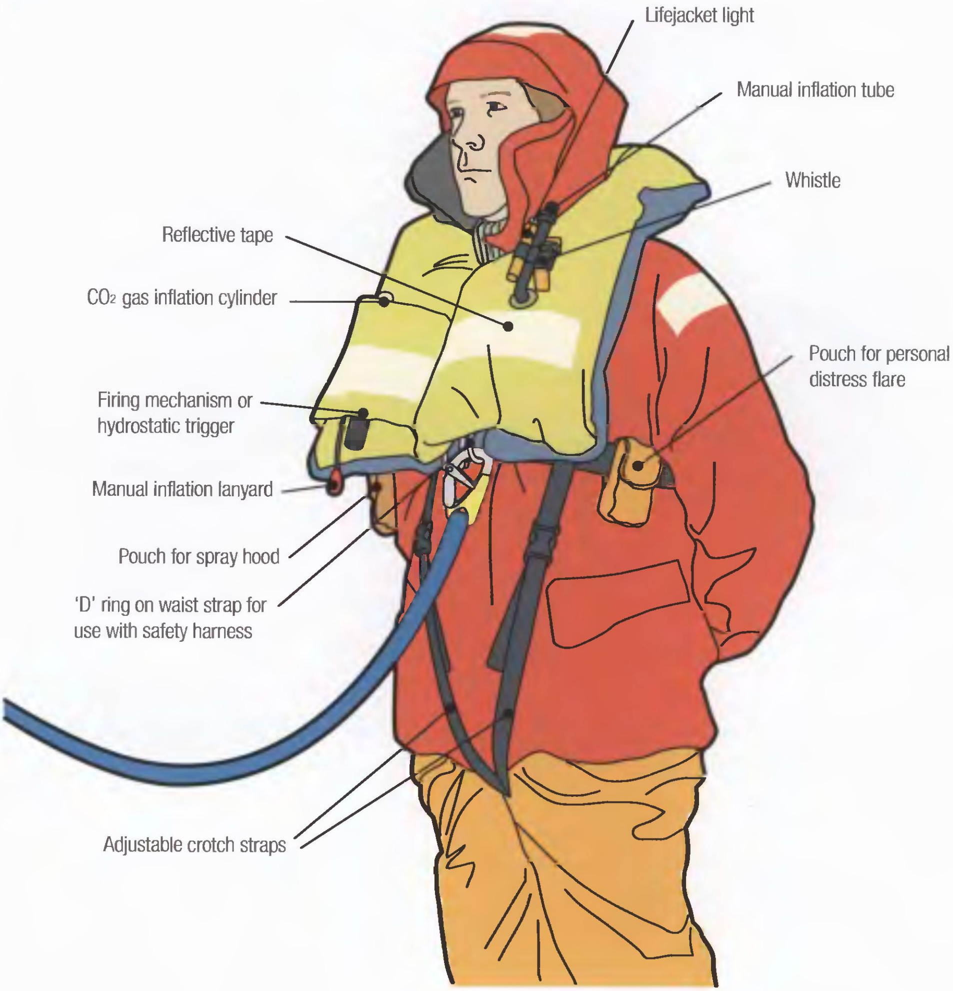
Lifejacket light
Manual inflation tube
Whistle
Reflective tape
CO2 gas inflation cylinder
Pouch for personal distress flare
Firing mechanism or hydrostatic trigger
Manual inflation lanyard
Pouch for spray hood
'D' ring on waist strap for use with safety harness
Adjustable crotch straps

Lifejacket fittings

Some jackets are sold with a number of fittings as standard, for example, the whistle, oral inflation tube, retro-reflective tape and gas inflation system. But other fittings such as the crotch straps, lifejacket light, built-in safety harness, safety harness, spray hood and flare pouch, are often optional extras. It is recommended to always use a lifejacket fitted with a light, spray hood and crotch straps.

The effect of wind and waves

When at rest in the water, the wearer of a lifejacket will be continuously turned to face the wind and waves. While a conscious casualty can cover their airway, an unconscious one will drown from waves and spray continually breaking onto their face.

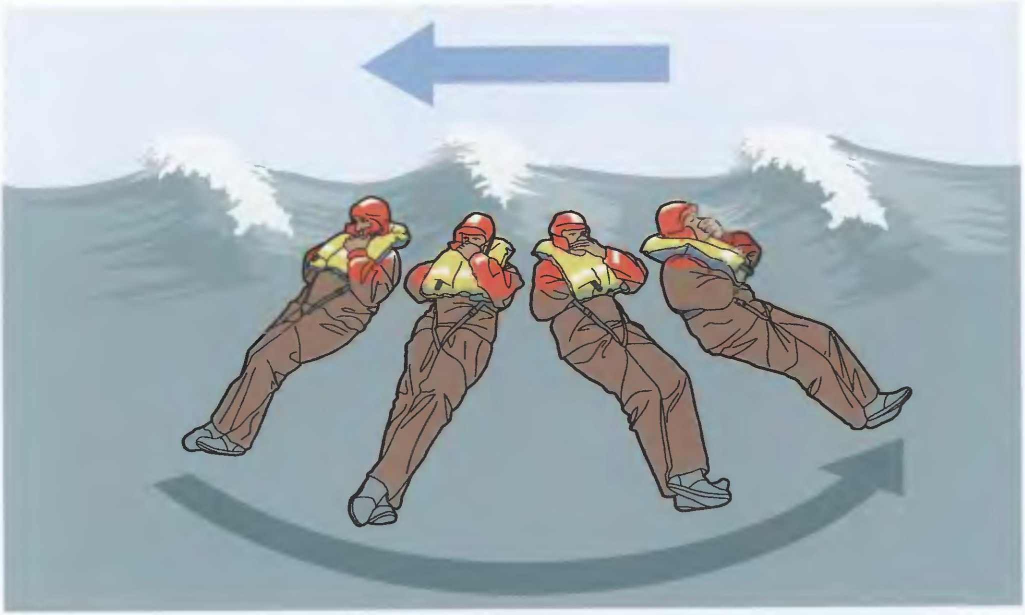

A spray hood, either fitted into the collar of the lifejacket or stored in a separate pouch on the waist belt, can be worn over the head or face and will reduce the effects of breaking waves.

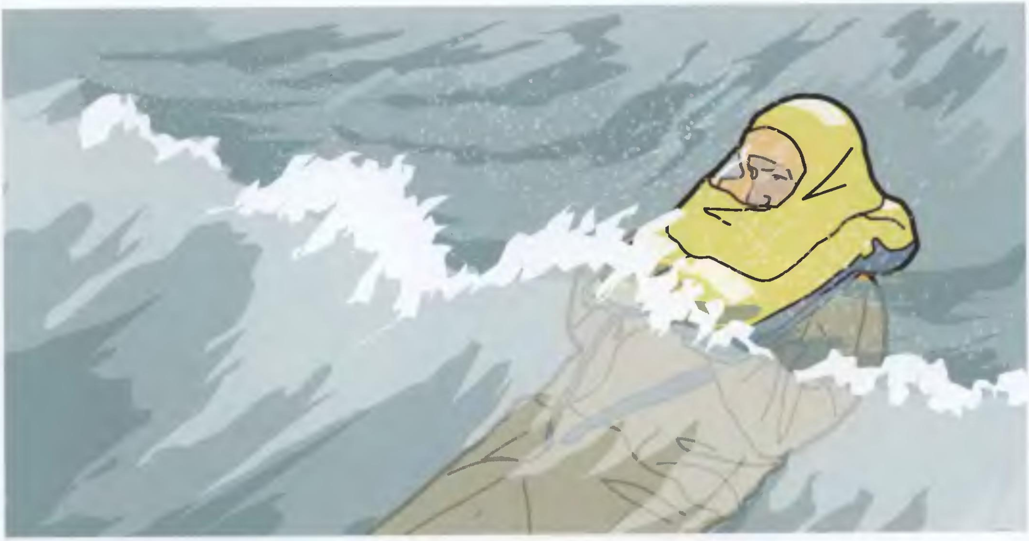

Donning a gas-inflated lifejacket

You should be able to put on and fasten your lifejacket in less than 30 seconds, so practise until you can.

It's essential the lifejacket is worn correctly. If it is too slack the wearer may slip down or even out of the jacket while in the water or being lifted aboard a boat. The lifejacket needs to keep the airway as high above the water as possible. Crotch or thigh straps hold the lifejacket down and lift the body up and, therefore, the airway higher above the water. Wear your lifejacket at all times when on deck. Once it is adjusted to fit you, keep your lifejacket close to hand when below decks, so that it can be donned quickly.

1

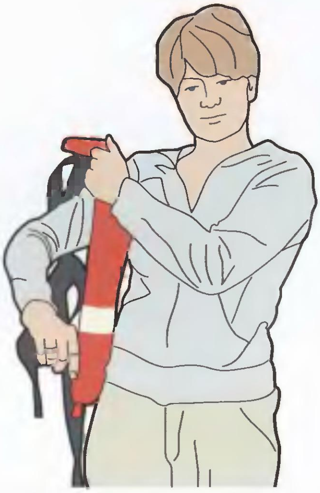

Put the jacket on like a waistcoat. Hold the jacket the right way up and put your right arm between the jacket front and the back strap.

2

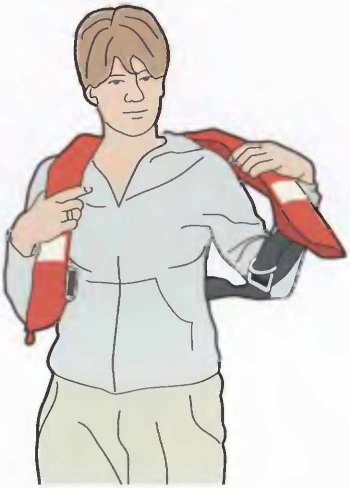

Pull the jacket across your back and put your left arm between left jacket front and the back strap.

Buddy check

It's good practice to ask another crew member to check you are wearing the lifejacket correctly.

3

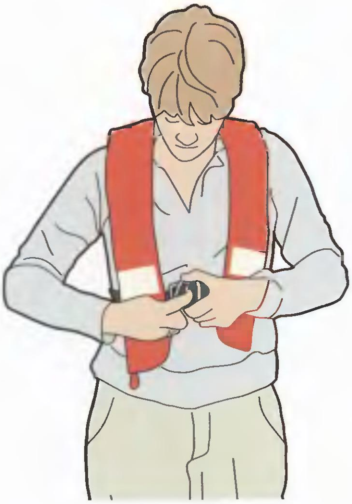

Straighten the waist belt and fasten the clip at the front of the jacket. Check it is secure.

4

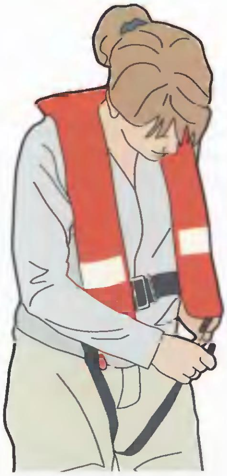

Reach down and pull the crotch strap or thigh straps between your legs and clip it to the waist belt. Shorten the straps to take up the slack but not so tightly as to be uncomfortable.

5

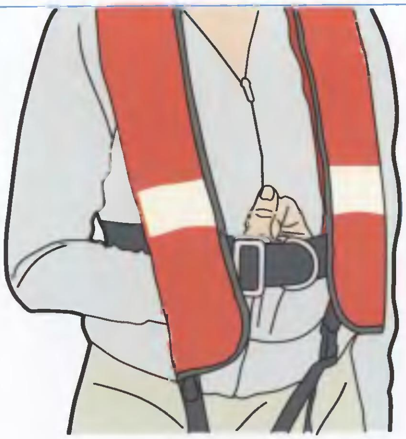

Adjust the waist belt so that it is a tight fit when your fist is placed between the strap and your chest. This will prevent the jacket from squeezing your chest when it inflates and without making it too slack.

Tip

- Stow lifejackets where they can be reached quickly
- Each crew member to have their own lifejacket already adjusted to size
- Add a coloured tag to make the lifejacket easier to identify
- Write your name and the boat's name on the back of the service label in indelible ink
- Don't think when to put your lifejacket on, consider when to take it off!

CHAPTER 6 : MOB prevention

Staying on board your boat is preferable to falling overboard! Clip on. Do not go on deck unnecessarily in heavy conditions and clip on if you must. When moving around remember the rule – one hand for the boat, one for yourself. Prevent accidents by clearing away unnecessary clutter and other hazards. Wear non-slip shoes or boots and avoid standing on ropes or sails since these can roll or slide underfoot. Try to avoid stepping on to smooth wet surfaces such as shiny glassfibre. Wash off oil or diesel spills especially from sidedecks.

Regularly check lifelines and stanchions are firmly attached. Lifelines should be ready to be released or cut easily to aid MOB recovery. A 'Granny bar' at the mast of a sailing yacht helps to steady crew when working at the mast.

Ensure the helmsman warns the crew of manoeuvres – avoid sudden acceleration or surprise tacks or gybes. Look out and warn other crew of larger-than-normal approaching waves.

Clipping on

There are many factors which determine the need to clip on. On displacement sailing yachts, wear a safety harness or, preferably, a lifejacket with a built-in safety harness and 'clip on'. Strong attachment eyes should be fitted in the cockpit just outside the companionway so that crew can clip on before leaving the cabin. Other eyes can be fitted in the cockpit so that the crew can move away from the companionway.

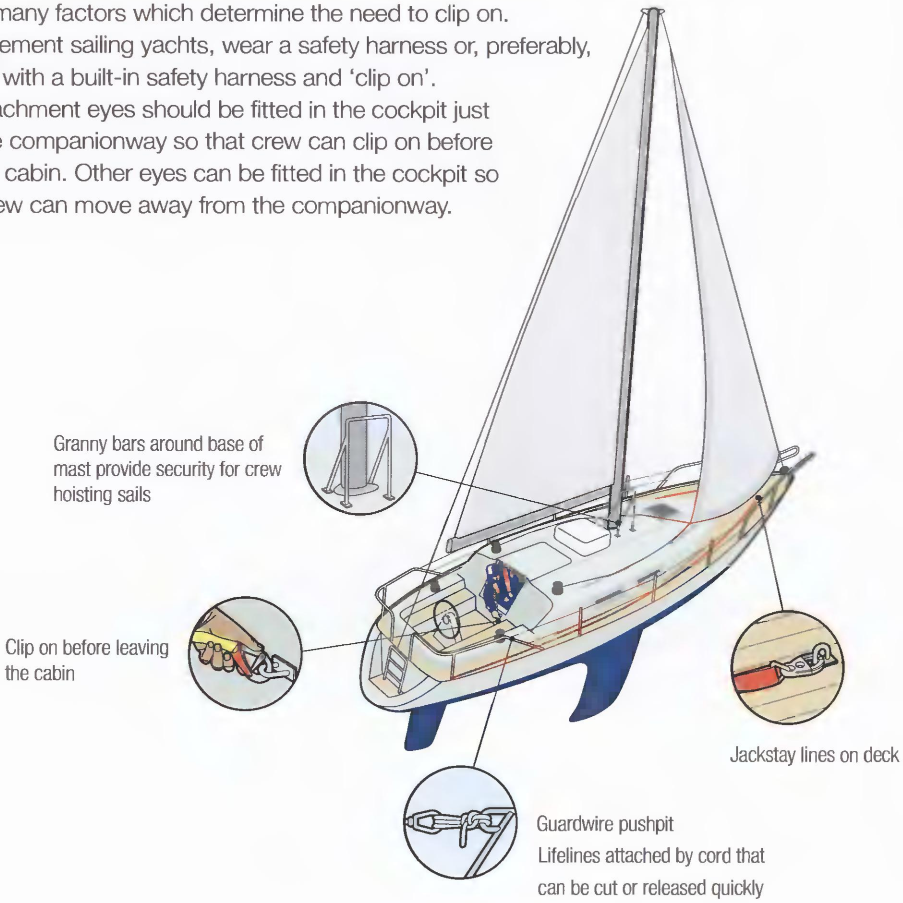
Granny bars around base of mast provide security for crew hoisting sails
Clip on before leaving the cabin
Jackstay lines on deck
Guardwire pushpit
Lifelines attached by cord that can be cut or released quickly

In heavy weather, only go on deck if necessary and clip on to a jackstay. Keep your centre of gravity low – crawl if necessary. Tell the helmsman or skipper before going on deck. On sailing boats go forward on the windward side.

Use a harness with locking snap hooks, such as Gibb Hooks or Wichard Hooks. A three-hook safety harness allows crew to clip onto the next strongpoint before unclipping. Safety lines should have a safety hook at all ends so that the crew can unclip themselves quickly. Jackstays may be made from polyester webbing or stainless-steel wire, webbing jackstays are degraded by sunlight so check and replace them regularly. Wire jackstays can roll underfoot.

Fast boats

It is advisable not to clip-on to planing or semi-displacement boats. It is better to fall clear since being dragged alongside a hull at high speed will cause severe injury.

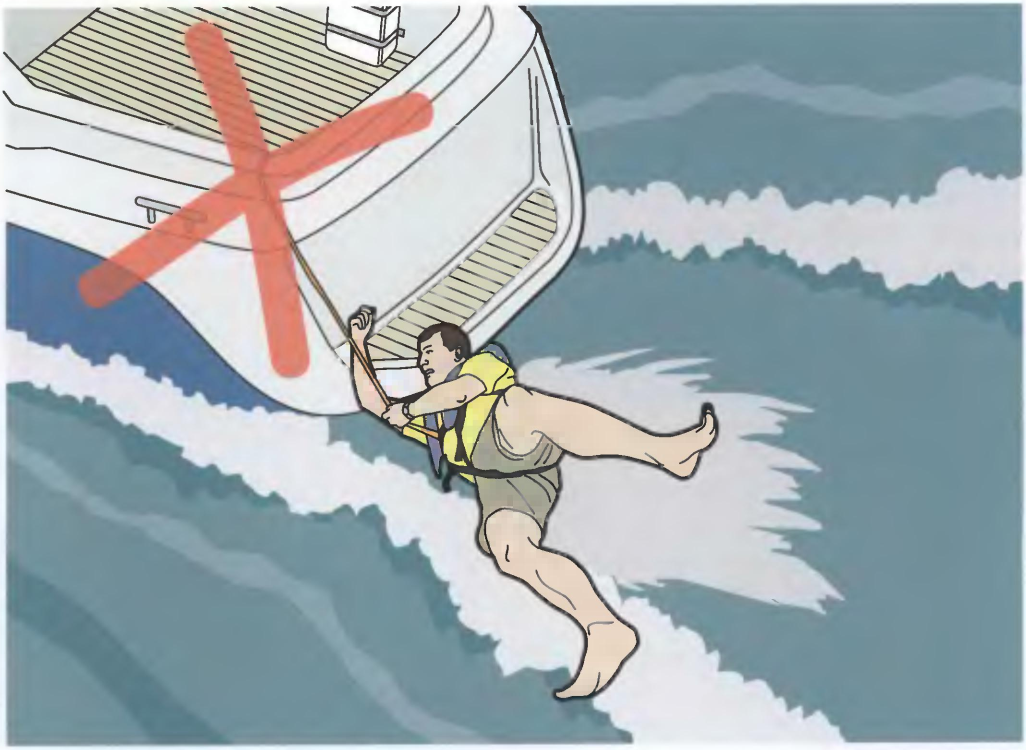

Never clip onto a motorboat when it is moving at speed.

However, if the boat has stopped or is moving at under 5 knots, you can clip on when going on deck, for example, when going forward to drop anchor in heavy weather. Motorboats seldom have jackstays fitted but there is no reason why they cannot be fitted.

CHAPTER 7 : MOB

MOB actions to be taken by crew

1. Shout 'man overboard' to alert the crew.
2. Press the MOB button on the GPS.
3. Throw a life-buoy and dan buoy to the MOB. Mark the MOB with a buoyant smoke flare.
4. Allocate a crew member to point at the MOB in the water.
5. Send a DSC distress alert and a voice Mayday.
6. Keep pointing. Don't lose sight of the MOB.
7. On a sailing boat, the skipper will ask for the jib to be lowered or furled and the engine started.
8. Prepare a throwing line.
9. The skipper will bring the boat alongside the MOB, with the boat pointing into wind and the propeller stopped.
10. Get a line around the MOB and get them aboard.

MOB search patterns

These search patterns are for a boat to follow when looking for a man overboard who has been lost from sight. Press the MOB button on the GPS receiver as soon as possible after the person falls overboard. This will record a geographical position and a bearing and distance to that point.

Press the DSC distress alert, choose 'MOB' from the menu options and send a voice Mayday. If short -handed, there may not be time to send a voice Mayday immediately. The MOB is unlikely to be at the recorded position but it will provide a datum for calculating where they will drift due to wind and current.

Sector search

A good datum is required. Sector search is good for searching small areas. Mark the datum point with a dan buoy, a life buoy with drogue, a weighted fender or similar. The marker should be able to drift with the tide but not be blown by the wind.

Go North (000 degree) first (some prefer to go downdrift first but this can make the sums more complicated). When moving away from the marker, count out loud glancing at the marker. When the marker is visible for only 50% of the time, that is the distance that indicates the Expected Detection Range (EDR). Measure each leg by counting aloud. It is suggested that each leg is three times the EDR. At the end of each leg, turn 120 degrees to starboard and proceed for the same distance.

If the search is not successful after the first circuit, reorientate the search by 30 degrees to starboard.

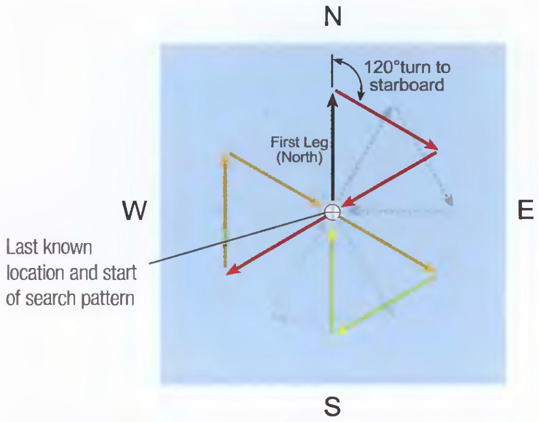

All turns are 120^0 to starboard. On completion of first search pattern (coloured) rotate by 30^0 as shown by dotted line

Sector Search

Expanding box search

If the datum point is a little unreliable or a sector search has not proved successful, an expanding box search should be started.

Usually, the leg length is 75% of the EDR. If a sector search has been completed, this will be the third search of the area, in which case it is acceptable to use a leg length of 100% EDR.

For ease of use, it is suggested that the headings are North, East, South and West. Count aloud to measure the distance.

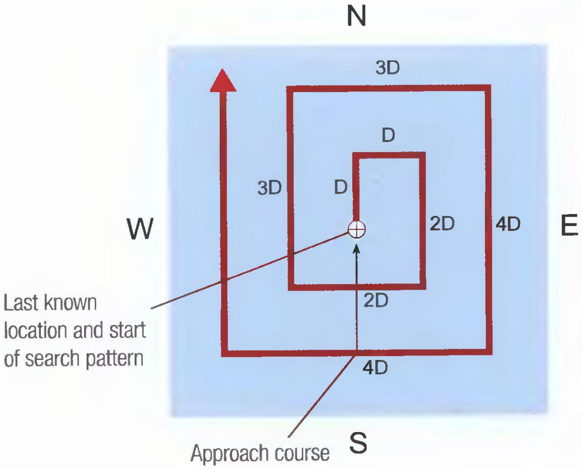
N
3D
D
3D
D
W
2D
4D
E
Last known location and start of search pattern
2D
4D
Approach course
S

Expanding Box Search

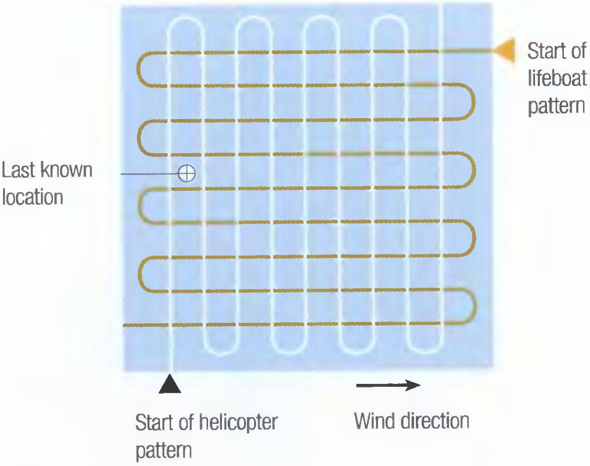
Start of lifeboat pattern
Last known location
Start of helicopter pattern
Wind direction

SAR Search Pattern

SAR search pattern

To search large areas where the position of the casualty is unknown, the Rescue Co-ordination Centre (RCC) will task the SAR units to undertake a parallel search pattern. If lifeboat and helicopter are both involved, their tracks are likely to be at 90 degrees to each other to increase the chances of detection.

Sailing yacht MOB manoeuvre

It will be quicker and safer to use the engine to return to the MOB. But if it won't start or the propeller is fouled, use the sails. Make sure the crew are practised in MOB recovery procedures.

Under sail

1. Sheet in the mainsail and heave to to slow the boat. Pass buoyancy to the casualty and mark with a dan buoy. Instruct a crew member to point at the MOB. Retrieve any warps in the water and start the engine.
2. If the engine is not working turn onto beam/broad reach and sail away.
3. Sail about five to six boat-lengths from the MOB. Do not lose sight of MOB.
4. Tack. Aim the leeward side of yacht at the MOB or the marker. Let out the headsail and mainsail sheets. The mainsail should flap; if not, bear off downwind to change the angle of approach. Point the boat back at the MOB until the mainsail flaps.
5. The angle of approach should be a close reach so that the sails can be powered and depowered. Drop the headsail if there is sufficient power from the mainsail alone.
6. Fill and spill the mainsail and slowly approach the MOB. Pick up the MOB to leeward, aft of the mast.

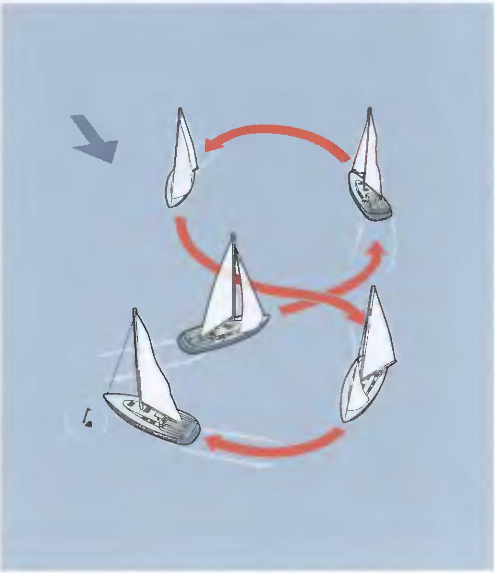

Under power

1. Sheet in the mainsail and heave to to slow the boat. Pass buoyancy to the casualty and mark with a dan buoy. Instruct a crew member to point at the MOB. Retrieve any warps in the water and start the engine.
2. Furl or drop the headsail.
3. Make ready the throwing line.
4. Manoeuvre the boat downwind of the MOB, keeping the MOB in sight.
5. Approach the MOB into the wind so that the mainsail is depowered. Pick up the MOB on the leeward side, aft of the mast.

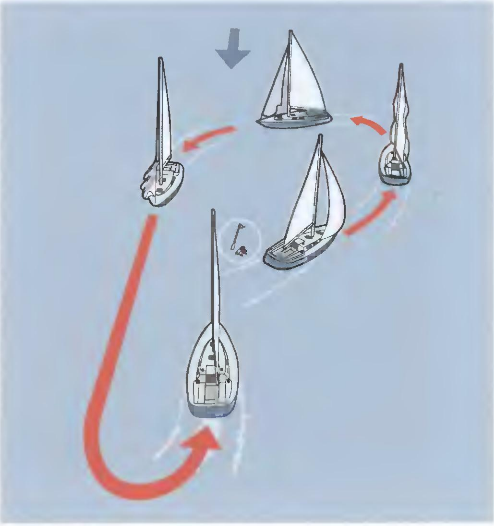

Powerboat MOB manoeuvre

The Williamson Turn is designed to turn the motorboat onto a reciprocal track.

1. Note the compass heading and add 60 degrees.
2. Turn to starboard on to the new course; note the amount of helm used.
3. When you reach the heading, turn the helm in the opposite direction at the same rate as before.
4. As you reach the reciprocal heading or you see your original track in the water, straighten the wheel.
5. Slow down and look for the MOB.

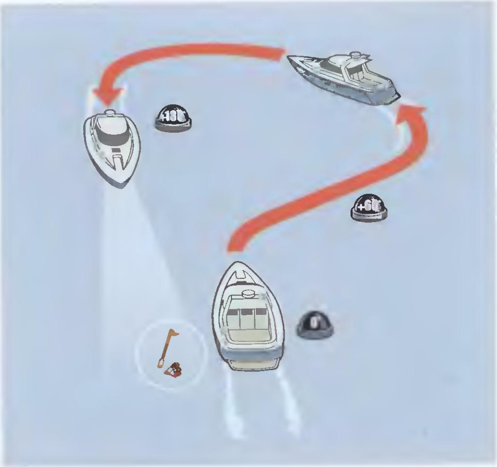

Approaching the MOB

In rough weather, it will be best to approach from downwind, keeping the bow into the wind, taking care not to run over the MOB.

1. Pick the MOB up on the leeward side.
2. Stop engines when the MOB is alongside.
3. Use a throw line if unable to get close to a conscious MOB and pull them to the boat.

Note

If the helm position is on the starboard side,try to leave the MOB to starboard so that they can be seen more easily.

In acceptable conditions, position the boat across the wind, upwind of the MOB.

1. Keep the boat sideways-on to the wind and allow it to blow the boat towards the MOB.
2. Pick the MOB up on leeward side.
3. Stop the engines when the MOB is alongside.

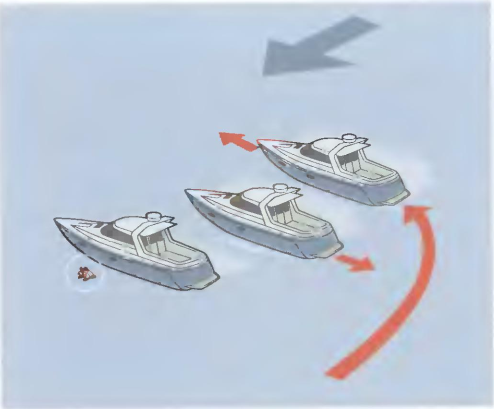

MOB recovery methods

Lifting a person out of the water requires planning and practice. It is a difficult task, especially for short-handed crews.

It's essential to get the person out of the water as quickly as possible. In very calm conditions it may be possible to retrieve them using the bathing platform or the ladder at the transom. However, in rougher seas this will be a dangerous place to attempt to recover a MOB, as the casualty is likely to be drawn under the stern of the boat and injured.

How you go about the task will depend on how your boat is set up and how long the person has been in the water. Hypothermic casualties should be recovered in a horizontal position and handled with care.

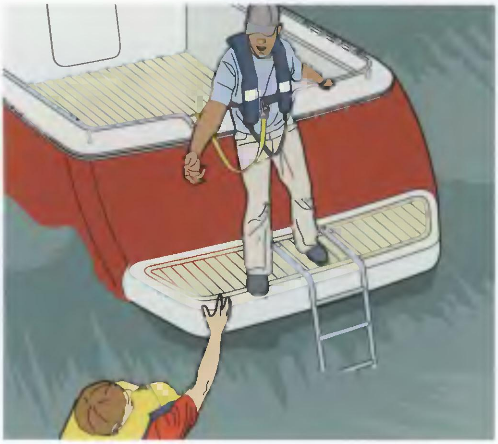

It may be easiest and quickest to launch the liferaft, and for a crew member to enter the raft dry and pull the MOB into the raft. This method is not without risk and the raft will not be available should you need it later in the voyage. Lower either the lowest or both guardrail wires along the side of the boat to ease recovery.

Crew dealing with MOB recovery should ensure that they are attached to the boat with a safety line. One MOB is enough.

Recovery methods for sailing yachts

Parbuckle

The Parbuckle is set on the side of the boat. The foot is fastened along the gunwale and the head attached either to a handybilly (block and tackle) that is attached to a halyard or to the halyard alone.

Lower the Parbuckle into the water and position the MOB.

Heave in on the handybilly – use a sheet winch for greater mechanical advantage.

The casualty will roll inside the parbuckle up and on to the sidedeck.

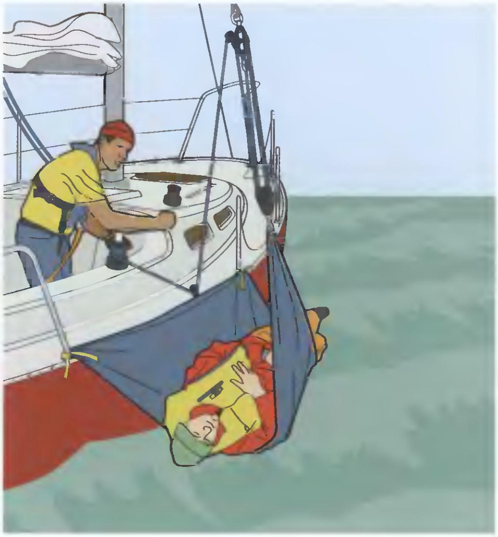

A small headsail can be used in place of a purpose-designed parbuckle. However, it will be more difficult to get the sail to sink sufficiently to position the MOB in it. Also there's a strong possibility that the MOB will slip out of the sail as they are being brought aboard.

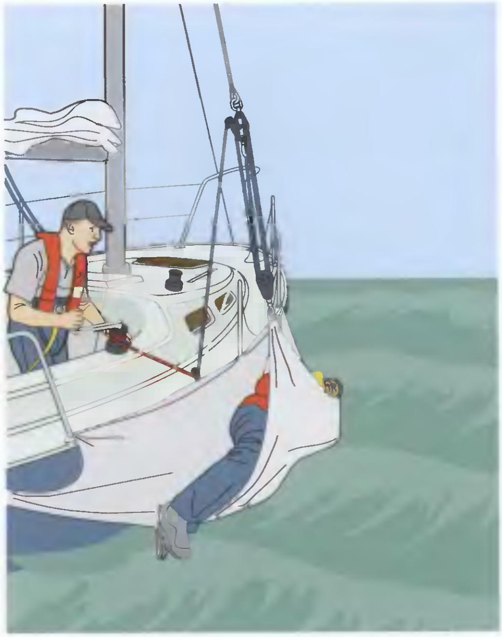

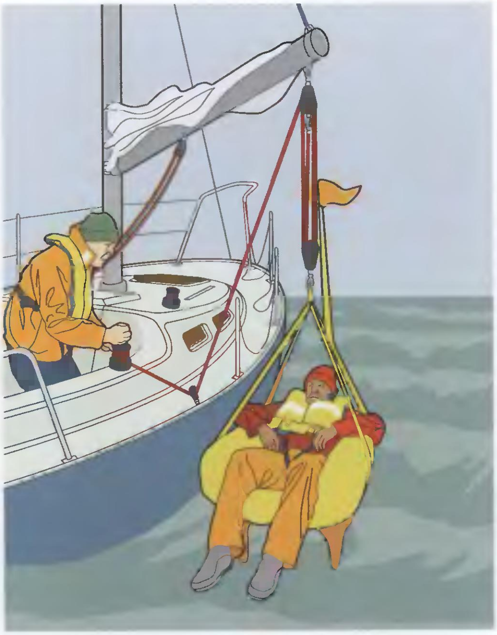

MOB recovery raft
Proprietary MOB recovery rafts can also be lifted on board using a handybilly attached to the halyard or the boom lift.

Fastening to the top of the raft is easier than attaching to a MOB in the water. The rafts are easy to get into and minimise hydrostatic squeeze.

Boom lift

On yachts with booms not fitted with solid kickers, it may be possible to use the boom as a derrick. The yacht needs to be sufficiently large to withstand lifting a MOB on the boom.

The main sheet, attached with snap shackles at both ends, is reversed and freed from the deck fitting.

Slacken off the kicker and scandalise the boom to an angle of about 30 degrees from the horizontal. It may be necessary to back up the topping lift with the halyard.
Attach the bottom of the sheet either to the MOB's lifejacket or harness or to a rescue strop. Place the rescue strop around the MOB's chest.

Heave on the mainsheet or redirect the free end to a sheet winch and lift the MOB into the yacht.

A horizontal lift can be achieved by placing an additional safety lanyard or line behind the knees of the MOB.

Elevator method

Developed in the United States, the elevator method provides a quick way of helping a conscious MOB out of the water.

A line is attached to a forward cleat and led aft outside of the guard wires, fed through the headsail fairlead to the sheet winch. The MOB stands on the line as it is winched tight, thereby lifting the MOB up the side of the boat. In most cases a second line will need to be rigged over the side of the boat for the MOB to hold on to.

Some motorboats may be able to adapt this method by rigging the line from a stern cleat, through the bow roller and on to the windlass.

Recovery method for motorboats

MOB recovery on a motorboat is more difficult than on sailing yachts due to the lack of rigging.

Many motor cruisers carry a small inflatable dinghy in davits or on snap davits attached to the transom bathing ladder. A quick method of MOB recovery is to lower the dinghy and recover the MOB into the dinghy. Make sure helping crew are attached to the boat.

If the MOB is too heavy to lift into the dinghy, one of the sponsons can be deflated and the MOB rolled into the dinghy.

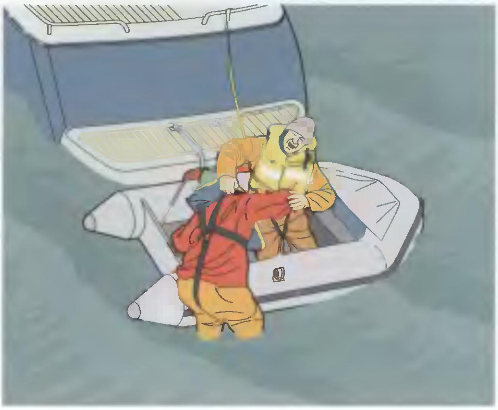

CHAPTER 8 : Liferafts

Liferafts are designed should your boat sink or catch fire to get you out of the water and provide shelter for a limited period. Whenever possible, let the SAR services know you are abandoning to a liferaft so that they know what they are looking for.

The quality of liferafts varies enormously. Some products marked as a liferaft may be no more than a couple of gas-inflated floats: others provide an acceptable, stable, sheltered environment. Always look inside the canister if buying a secondhand raft – it has been known for unscrupulous sellers to replace the raft with bricks!

Basic leisure rafts

Usually at the lower end of the price scale, leisure liferafts are not made to any national or International standard or specification. Nor are they approved for any commercial operation, yachts sailing under racing rules or Class XII craft. The leisure liferaft is light in weight (easier for launching), and may or may not have a canopy. Some may have only one buoyancy chamber and minimal survival equipment as standard. But this is better than nothing in an emergency situation.

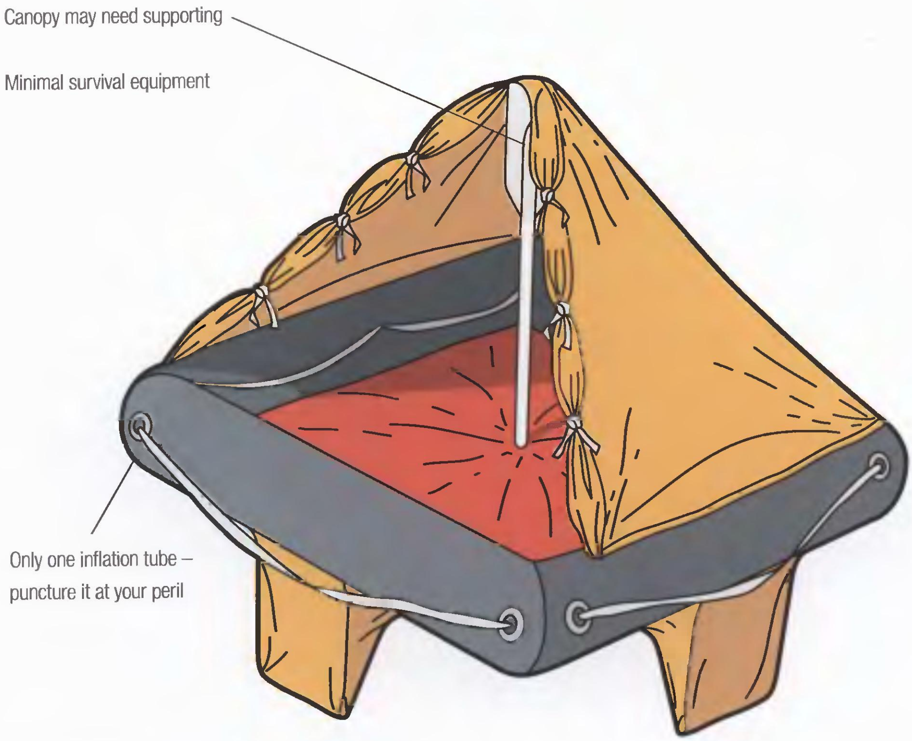
Canopy may need supporting
Minimal survival equipment
Only one inflation tube – puncture it at your peril

Basic Leisure Liferaft

Inflatable Active Survival Craft

A number of offshore racing associations approve some makes of active survival craft as an alternative for a liferaft. However, most national authorities do not accept them as suitable safety equipment because there is currently no recognised standard for active survival craft.

Many are based on an inflatable dinghy fitted with a liferaft gas inflation system and a canopy for shelter. While the craft can be rowed or sailed to safety, the lack of ballast pockets make them unstable in a seaway.

Since some areas have minimal rescue services, many bluewater sailors design their own survival craft, based on an inflatable or rigid dinghy with a small rig that allows them to effect their own rescue. However, these may not always be able to be readied and launched as quickly as a liferaft.

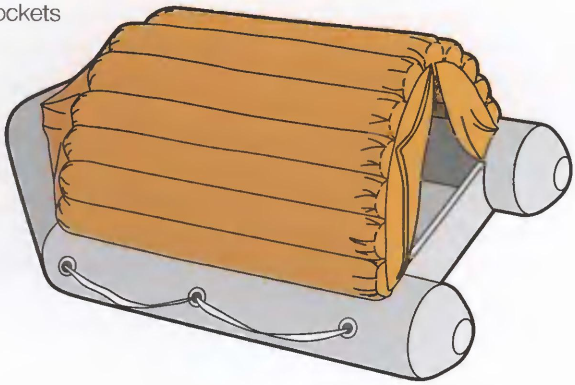

Liferaft standards

SOLAS

Commercial vessels have been required to carry liferafts built to SOLAS Standards, as detailed in Chapter IV of the Life Saving Appliances (LSA) code, since 1959. This standard has been periodically upgraded and SOLAS rafts are made and tested to high standards and extremes of operating temperatures. They are approximately 30% heavier, tougher and significantly more expensive than ISAF or ISO rafts. The Standard covers rafts from six-person capacity upwards.

ORC

The Offshore Racing Congress specification was introduced after the 1979 Fastnet Race inquiry. Yachts racing under their rules were required to carry ORC rafts. This has now been superseded by the ISAF Part II specification up until June 2007. ORC rafts built before December 2002 could still be fitted to ISAF racing yachts. However, this is no longer allowed. ORC rafts may still be found on many cruising yachts. ORC is also known as ISAF Part I and is in the ISAF Offshore Special Regulations.

ISAF

Following the 1998 Sydney–Hobart Race, the International Sailing Federation, (which now oversees the majority of offshore sail yacht racing in place of the ORC), developed a specification for liferafts similar to the ISO standard, principally as an interim measure while the new ISO standard was being completed. Although similar there are differences between ISAF and ISO. The ISAF specification came into force on 1st January 2003 and covers rafts of maximum capacities of 4 to 12 persons.

> Note that ISAF specifies that Category 0 (zero) yachts should carry SOLAS rafts. The specification is included in Appendix A part 2 of ISAF's Offshore Special Regulations. (www.sailing.org/)

ISO 9650 Small Craft – Inflatable Liferafts
Introduced by the International Standards Organisation during 2005, ISO 9650 comprises of three parts:

Part 1:	Type 1 Liferafts for offshore use subdivided into: Group A rafts designed to inflate in temperatures of between –15 and +65 degrees C. Group B rafts that will inflate between 0 and +65 degrees C (air temperature).
Part 2:	Type 2 Liferafts for coastal use. All Part 2 rafts are expected to inflate between 0 and +65 degrees C.
Part 3:	Material to be used in liferaft manufacture

The main differences between Part 1: Type 1 and Part 2: Type 2 rafts are shown in the table. Type 2 rafts are, size for size, slightly smaller than Type 1 rafts. Type 1 rafts are more robust than Type 2 rafts. Type 1 Group A liferafts are recommended for northern European waters.

	ISO 9650 Part 1: Type 1	**ISO 9650 Part 2: Type 2**
Maximum Capacities	4 person to 12 person	4 person to 10 person
Painter length	9 metres	6 metres
Withstand drop of	6 metres	4 metres
Minimum Freeboard	4 person = 250mm 6 person+ = 300mm	4 person = 200mm 6 person+ = 250mm
Floor	Grp A rafts include thermal floor	Optional

Liferafts built to ISO 9650 Part 1: Type 1 are given a choice of two equipment packs depending on expected duration at sea – less than 24 hours and more than 24 hours (see Liferaft packs on p.62).

ORC, ISAF, ISO 9650 and SOLAS specify that the liferaft's main buoyancy is divided into two separate chambers. Each chamber must have sufficient buoyancy to support the maximum capacity of the raft. The chambers must be able to be topped up with a bellows and have pressure-relief valves. Each standard also specifies that the liferaft must have a canopy.

Which size raft?

The capacity of the liferaft/s should be sufficient to carry all persons on board. However, if regularly sailing short-handed, it may be advisable to obtain two smaller rafts instead of one large raft. An added advantage being that the smaller rafts will be lighter and easier to handle. In addition to the ballast pockets, a liferaft uses the weight of its crew to keep it stable in waves. For example a boat that can carry eight people would have two four-person rafts instead of one eight-person raft if the boat was regularly sailed by a crew of only two or four.

Typical leisure raft

There are a number of mid-range yacht liferafts that do not meet ORC, ISAF, ISO or SOLAS standards. Their prices are lower but the quality of construction is also lower and there are often very few fittings included.

ISAF, ISO 9650 Part 1 and SOLAS

Externally there are few differences between liferafts built to SOLAS, ISO 9650 Part 1:Type 1 and ISAF. However, as a guide, SOLAS rafts are more robust than ISO/ISAF rafts, which in turn are more robust than non-standard rafts.

SOLAS rafts are designed to:

- withstand exposure for 30 days afloat (compared with 20 days for ISO/ISAF).
- withstand stowage in temperatures from –30 to +60 degrees C.
- withstand a drop from 18 mtrs and withstand persons jumping repeatedly into them from 4.5 mtrs.

SOLAS rafts have lined canopies to provide insulation. Fig. 1 is a raft that includes the correct fittings to meet these standards.

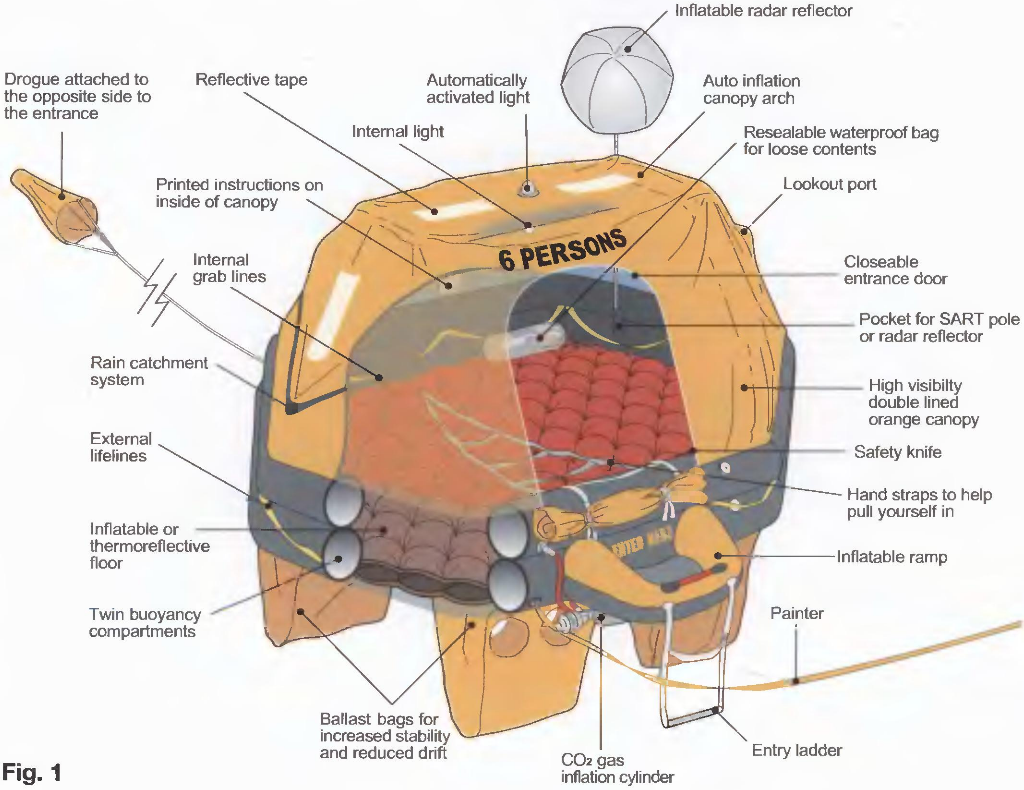
Inflatable radar reflector
Drogue attached to the opposite side to the entrance
Reflective tape
Automatically activated light
Auto inflation canopy arch
Internal light
Resealable waterproof bag for loose contents
Printed instructions on inside of canopy
Lookout port
Internal grab lines
6 PERSONS
Closeable entrance door
Pocket for SART pole or radar reflector
Rain catchment system
High visibilty double lined orange canopy
External lifelines
Safety knife
Hand straps to help pull yourself in
Inflatable or thermoreflective floor
Inflatable ramp
Twin buoyancy compartments
Painter
Ballast bags for increased stability and reduced drift
CO$_2$ gas inflation cylinder
Entry ladder

Fig. 1

Canister v Valise

Liferafts are supplied in either a plastic/glassfibre box or in a fabric valise. To protect from water damage, most rafts – canister and valise – are vacuum-packed in large polythene bags that rip open when the raft inflates.

Canister

The canister is usually made from two parts held together with straps and provides a hard case to protect the raft from knocks and sharp objects. The canister is not waterproof but the packaging inside should prevent ingress of water.

Pros

- Usually easily accessible
- Contents protected from damage
- Can be used with an HRU (see p.60)

- ***Cons***
- More likely to be stolen or washed overboard
- More expensive to buy
- Prone to damage from the elements

Valise rafts

These need to be stowed in a dedicated weatherproof locker. A valise should never be left strapped to an exposed deck or buried deep in a locker under piles of sails, fenders or warps. Make sure it is always readily available.

Pros

- Cheaper to buy
- Easier to move because they have handles and are lighter in weight than a rigid canister
- Less likely to be washed overboard

Cons

- Likely to be damaged by rough handling – do not throw them on to pontoons
- Harder to deploy because they have to be lifted out of a locker
- Cannot be automatically deployed
- Cannot be left in an exposed position

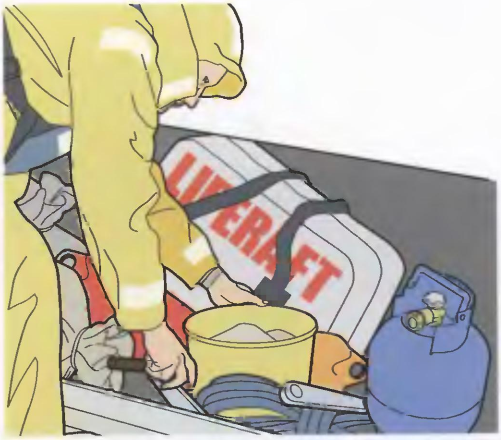

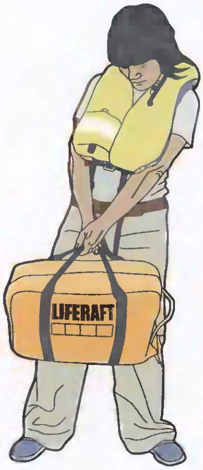

Liferaft stowage

A liferaft is a heavy and sometimes awkward piece of kit to carry. Apart from the smallest rafts, most will require a two-man lift. As you are likely to need it in a hurry, you need to give careful consideration as to where to stow it.

Ideally, the raft needs to be positioned so that it is readily accessible yet is protected from heavy weather.

Canister rafts can be deck-mounted on the coachroof, in the cockpit, on the bathing platform, on the transom or in a purpose-made bracket on the pushpit.

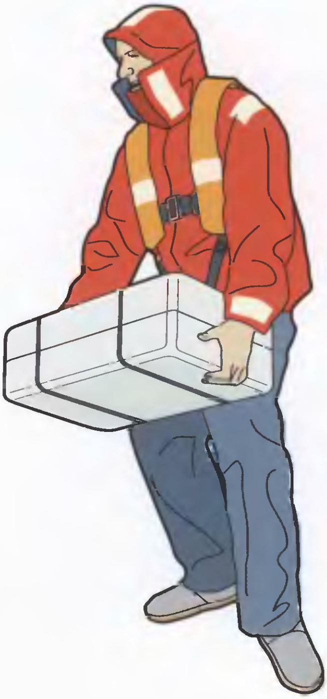

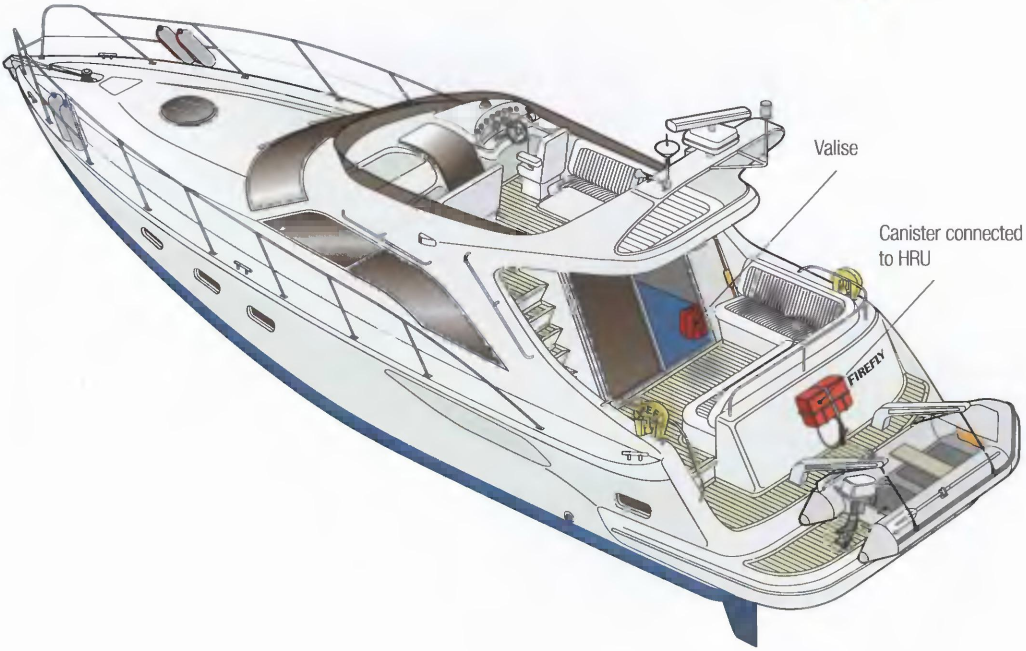

If stowed vertically in a bracket the raft may need to be specially packed to prevent the weight of the cylinder damaging the raft material. The canister's drain holes may also need to be repositioned.

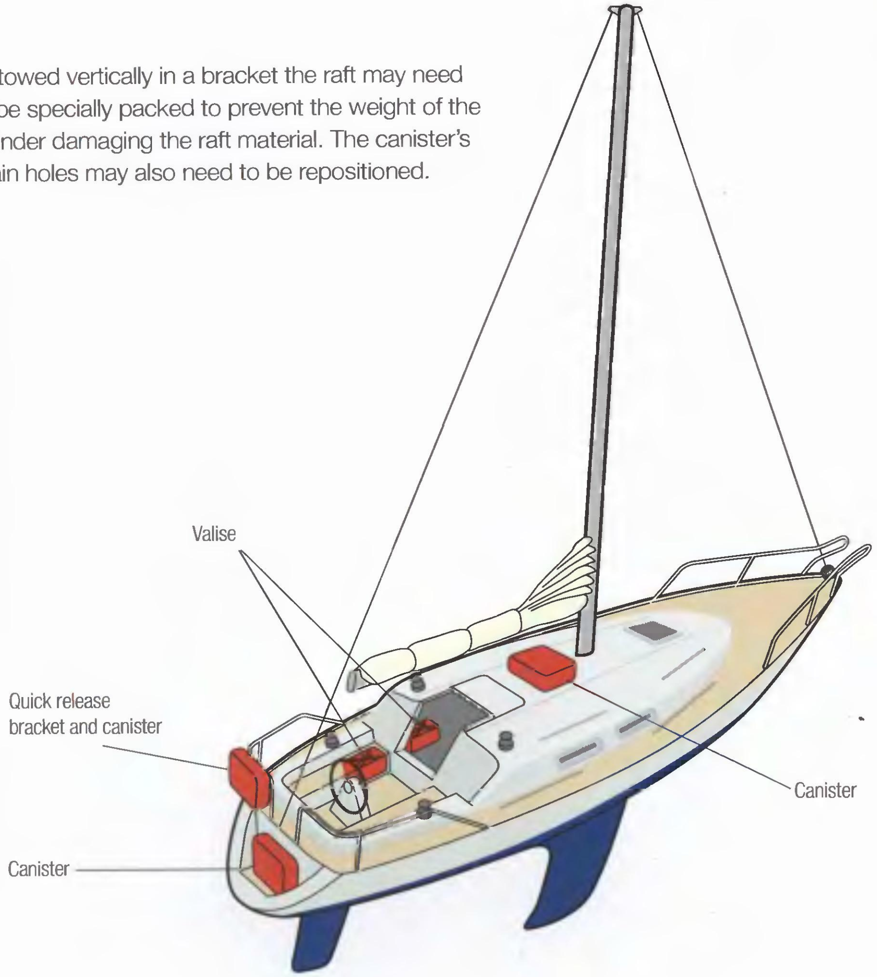

Maintenance

Make sure that a competent service agent services your liferaft at the manufacturer's required periods. Most modern rafts are vacuum-packed in large plastic bags, and if stored correctly should extend the service period to three years. However, there are still many rafts that need to be serviced annually. Good service centres will let you view the liferaft inflated and check the quality and amount of equipment packed with the raft.

Hiring

In most countries, it is possible to hire a liferaft from a liferaft hire centre. Rafts can be rented for periods from as little as a week right through to a whole season or more. This may be particularly useful for those that sail short-handed with one raft for most of the year and only occasionally with a full crew, when two rafts are needed.

Hydrostatic release units (HRUs)

Boats can sink within minutes if their hull is breached by even a relatively small hole. Fast craft are especially susceptible to high-speed impacts and have been known to sink in under 30 seconds! Commercial craft are required to fit HRUs to their liferaft.

How an HRU works

The HRU automatically cuts the securing strap holding the raft in place when the boat sinks to a depth between 1 and 4 metres (3 to 13 feet). The painter remains attached to the HRU weak link and the raft has sufficient buoyancy to float to the surface.

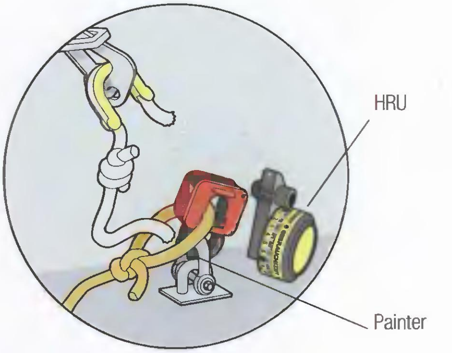

As the boat sinks further, it tugs on the painter to trigger the inflation system.

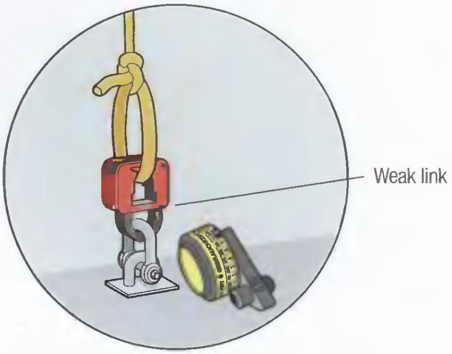

The greater buoyancy of the inflated raft is sufficient to break the weak link and releases the painter from the boat.

If the boat sinks in shallow water it may be necessary to release the liferaft manually.

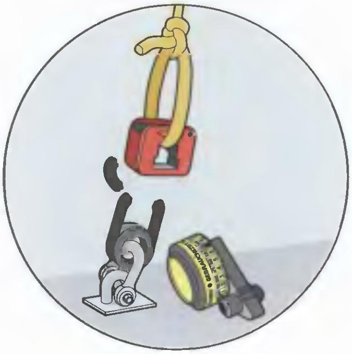

It's essential that the HRU is properly installed, because otherwise the raft will not break free and inflate. Illustrated below are new and old versions of the disposable Hammar HRU often used on boats. More expensive serviceable units are also available.

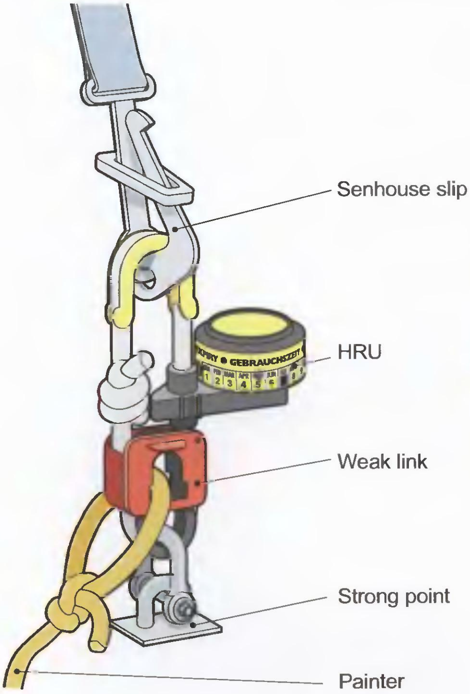
Senhouse slip
HRU
Weak link
Strong point
Painter

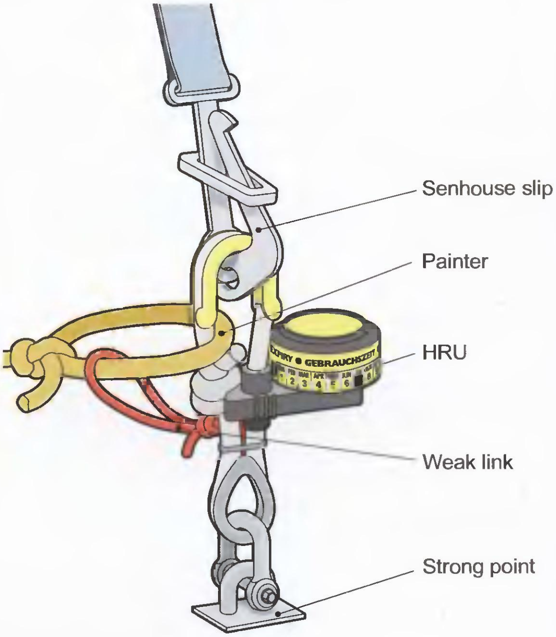
Senhouse slip
Painter
HRU
Weak link
Strong point

HRU safety checks

The Hammar HRU is a replaceable unit with a two-year lifespan. The expiry date should be marked on the unit. Replace it when required.

Check that no other straps (or padlocks) are used to secure the liferaft.

The ring on the Senhouse slip should be checked periodically to make sure it is free as vibration can cause the ring to jam in place, making it difficult to launch the raft manually.

Liferaft packs

Generally, the contents of a liferaft will depend on the type of raft: leisure, ORC, ISAF, ISO 9650 or SOLAS. Listed on page 63 are the minimum contents required under each specification. Note that ISO 9650 Part 1: Type 1 rafts have a choice of packs suitable for survival for under 24 hours or over 24 hours. You specify which pack when you purchase or service the liferaft.

The standard pack can be supplemented with extra equipment as long as it can fit in. Check contents with your service agent. If there is no room for the item you need, add it to your grab bag. Carefully consider what you may need for the area you will be sailing.

Find out what you have in your liferaft before you need it.

Many items inside the raft will have a limited life – flares, water, food etc – and need to be replaced when the raft is serviced. When possible, check the quality of equipment in your raft – cheap rafts may have poor-quality equipment - for example childrens' beach spades for paddles!.

Some raft packs pack equipment that you will need immediately in a separate pack.

Buoyant safety knife – attached with a lanyard and usually found in a sheath next to the raft entrance. Use it to safely cut the painter as close to the yacht as possible; keep the line, it may be useful later.

Bailer – it may be no more than a fabric dish, boots are much more effective but only as a last resort because you will suffer heat loss if they are removed. It's difficult to bail with the door closed, a stirrup pump is more effective, so pack one in your grab bag. The bailer can also be used as a sick bucket or a toilet – this is safer than leaning out of the doorway. Some liferafts may be fitted with an integral bailer, check manufacturer's usage instructions.

Sponges – the number supplied will depend on the raft pack. They usually come 'flat-packed' and dry. Use one to help dry out the raft. The other should be kept uncontaminated and used to collect drinking water by wiping condensation from inside the raft canopy.

Paddles – are essential for moving clear of a burning yacht or other hazard and very useful for manoeuvring to pick up survivors from the water. They are usually but not always in two parts and can also be used as a splint for broken limbs.

Whistle – this is more effective than shouting.

Item	Solas A	Solas B	Iso 9650 24 hours or more	Iso 9650 less than 24hrs	ISAF	ORC	Typical leisure raft (D-pack)
Buoyant knife	1	1	1	1	1	1	1
Bailer	1	1	1	1	1	1	1
Sponges	2	2	2	2	1pp	2	2
Paddles	2	2	2	2	2	2	2
Whistle	1	1	1	1	1	1	-
Torches	1	1	1	1	2	1	-
Heliograph	1	1	1	1	1	-	-
Anti seasick pills	6pp	6pp	6pp	6pp	6pp	-	-
Seasick bag	1pp	1pp	1pp	1pp	1pp	-	-
Red h/h flares	6	3	6	3	3	3	-
Para rocket flares	4	2	2	2	Grab bag	-	-
Buoyant smoke flare	2	1	-	-	-	-	-
TPA	2	2	2	-	2	-	-
Repair kit	1	1	1	1	1	1	1
Water	0.5lt pp	-	1.5lt pp	-	Grab bag	-	-
Food	10,000kJ pp		10,000kJ pp	-	Grab bag	-	-
First Aid kit	1	1	1	-	1	-	-
Bellows	1	1	1	1	1	1	1
Throwing line – 30m	1	1	1	1	1		1
Drogue	2	2	1	1	2	1	1
Waterproof notebook	-	-	-	-	1	-	-
Signal card	1	1	-	-	-	1	1
Sea Survival instructions	-	-	-	-	1	-	1
Leak stoppers	Set	Set	-	-	Set	Set	-
Radar reflector	1	1	-	-	-	-	-
Fishing kit	1	-	-	-	-	-	-
Tin opener	3	-	-	-	-	-	-
Graduated drinking cup	1	-	-	-	-	-	-

pp = per person

Torches – these usually come packed with only one set of spare batteries and bulb or there may be two sealed torches so make sure you pack sufficient spares. They are essential for looking for other casualties and for signalling in the dark.

Heliograph (signalling mirror) – easy to use and effective for daytime signalling. If no sight is provided, use your fingers. Hold out your arm and put up two fingers between you and the target. Direct the sunlight on to your fingers. You can use a CD to practise with, but be aware that water may dissolve the shiny surface, so they are not a suitable long-term alternative.

Anti-seasickness pills – should be taken as soon as possible, preferably before entering the raft.

Seasickness bag – only one per person is supplied in rafts made to the Standards,using them keeps the interior a little more habitable.

Flares – are arguably the most important pieces of equipment on the liferaft. The number and type will depend on the type of pack. Make sure you and your crew are familiar with the triggering methods and use of the different types because what are supplied in the raft may be different to those on your boat. Keep them dry and do not use them until you are confident they will be seen.

TPA (thermal protective aid) – not found on all rafts. Used to reduce heat loss of wet casualties, use by wringing out their clothes and redressing the casualty, then place the casualty in the TPA. Some TPAs are big enough for two people at one time.

Repair kit – the standard repair kit consists of patches and glue not dissimilar to those used to repair bicycle inner tubes, needs to be used in dry conditions. Many rafts are now also supplied with liferaft repair clamps and/or leak-stopper bungs that can seal holes quickly in wet or dry conditions.

Water – if included, will be only half a litre per person. Supplement this with extra rations or keep a desalinator in the grab bag.

Food – the type varies but it will be carbohydrate-based and should provide 10,000 joules per person.

First-aid kit – the contents will vary with the quality of the raft. The SOLAS C kit provides a comprehensive selection of equipment. Supplement your grab bag with extra medical supplies or any specific medications you need.

Bellows – used to top up the sponsons and for inflating the inflatable floor. Take care not to suck up water. Tie the bellows in because if you lose it you will need to top up the tubes by mouth. If possible take an extra pump from boat.

Throwing line and quoit – usually 30 metres long. If you are unable to paddle to other casualties, throw the line. If a casualty is unconscious, tie the line to a rescue swimmer.

Drogue/sea anchor – Deploy this immediately you are clear of the stricken vessel. The drogue will improve stability, turn the liferaft entrance away from the wind and waves and slow your rate of drift. Be aware that some drogues deploy automatically.

Waterproof notebook – useful for working out a watch system, keeping a tab of rationing and medication.

Signal card – this shows what the signals from aircraft and ships mean.

Sea survival instructions – a useful reminder of what to do. Supplement them with this book kept in your grab bag.

Radar reflector – deploy this as high as possible. **Do not use with a SART since they can conflict with each other.**

Fishing kit – this gives you something to do. Eat fish if you have plenty of water.

Tin opener – for opening supplies taken from the boat.

Graduated drinking cup – for ensuring fair sharing of water.

Grab bag

The grab bag is designed to supplement the contents of the liferaft. Ideally, it should float and be watertight. The contents should be stowed in waterproof containers that can be opened with cold wet fingers.

Some sailing-yacht racing rules specify what you must have in the grab bag.

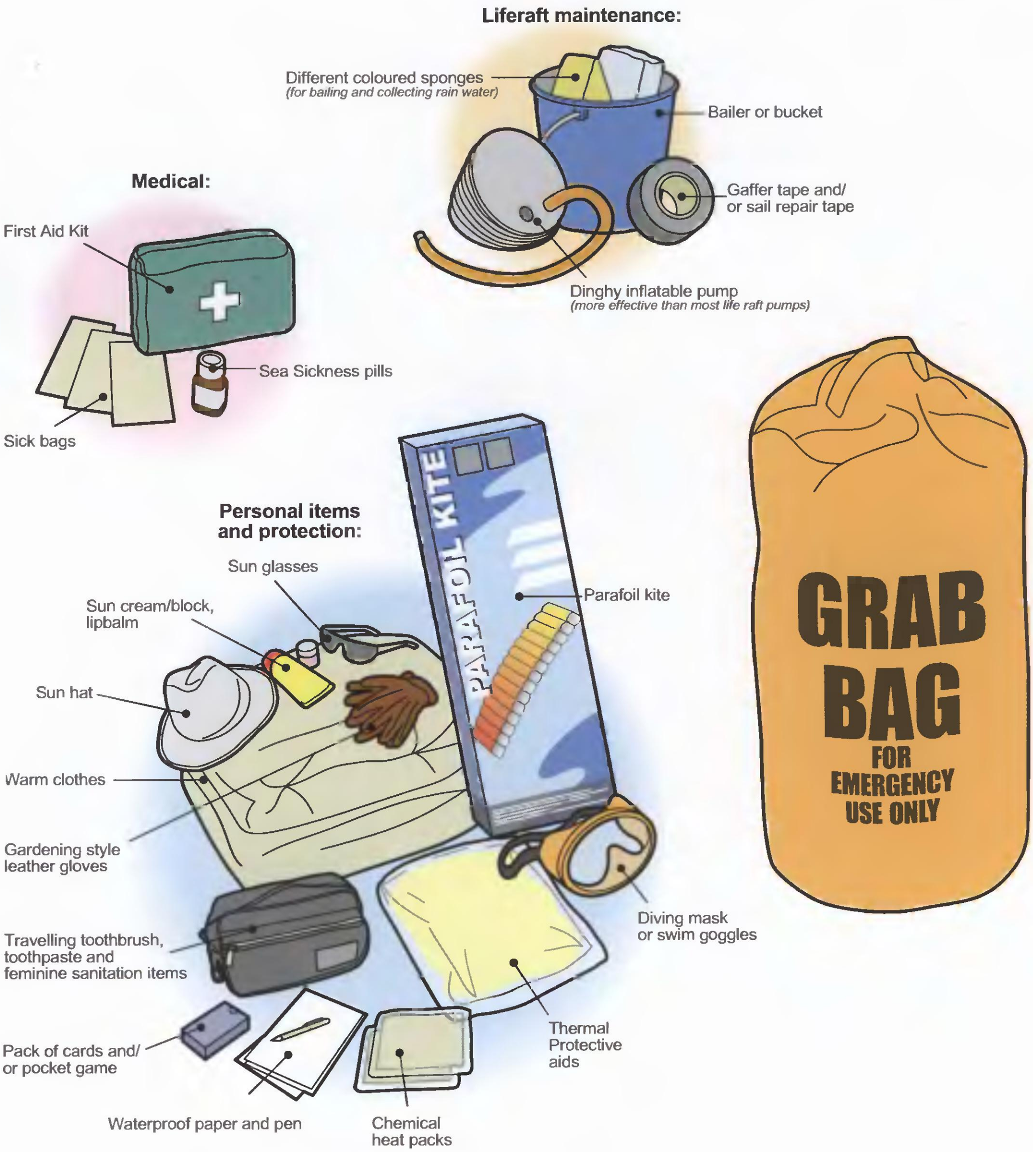
Liferaft maintenance:
Different coloured sponges
(for bailing and collecting rain water)
Bailer or bucket
Gaffer tape and/ or sail repair tape
Dinghy inflatable pump
(more effective than most life raft pumps)
Medical:
First Aid Kit
Sea Sickness pills
Sick bags
Personal items and protection:
Sun glasses
Sun cream/block, lipbalm
Sun hat
Warm clothes
Gardening style leather gloves
Travelling toothbrush, toothpaste and feminine sanitation items
Pack of cards and/ or pocket game
Waterproof paper and pen
Chemical heat packs
Thermal Protective aids
Diving mask or swim goggles
Parafoil kite
PARAFOIL KITE
GRAB BAG
FOR EMERGENCY USE ONLY

Survival and maintenance items:

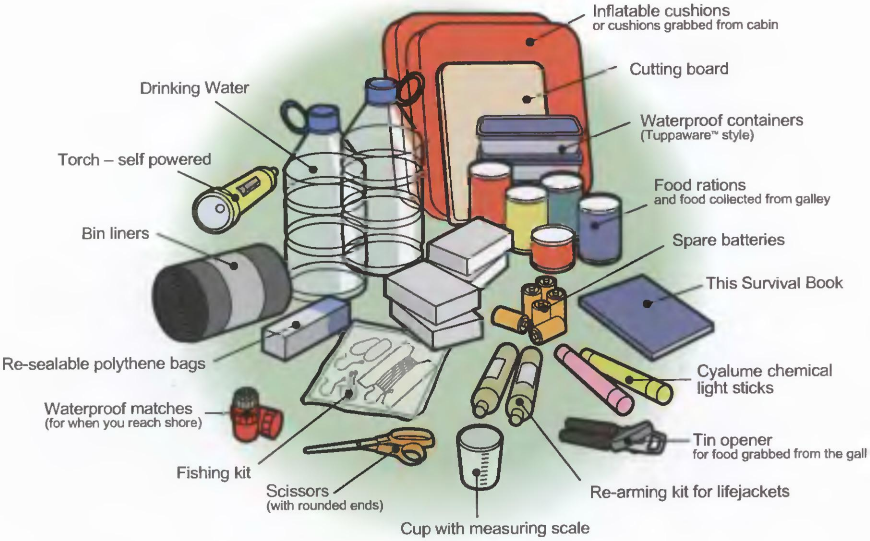
Inflatable cushions
or cushions grabbed from cabin
Cutting board
Drinking Water
Waterproof containers
(Tuppaware™ style)
Torch – self powered
Food rations
and food collected from galley
Bin liners
Spare batteries
This Survival Book
Re-sealable polythene bags
Cyalume chemical
light sticks
Waterproof matches
(for when you reach shore)
Tin opener
for food grabbed from the gall
Fishing kit
Scissors
(with rounded ends)
Re-arming kit for lifejackets
Cup with measuring scale

Rescue and navigation:

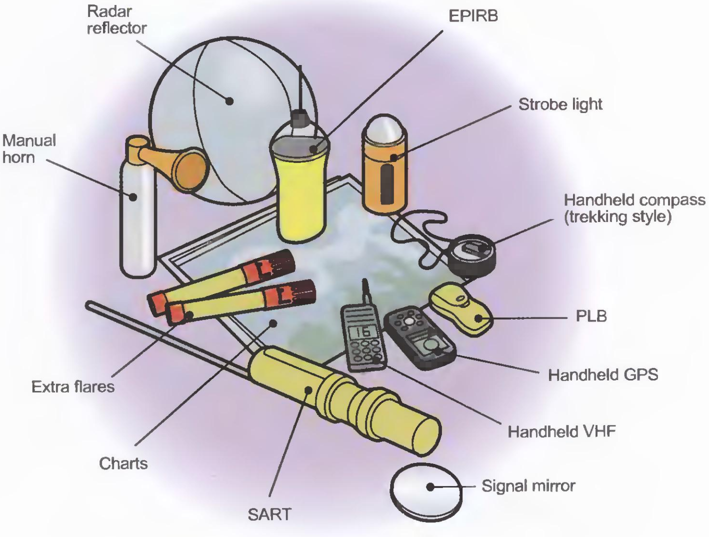
Radar
reflector
EPIRB
Strobe light
Manual
horn
Handheld compass
(trekking style)
16
PLB
Handheld GPS
Extra flares
Handheld VHF
Charts
Signal mirror
SART

If some items are normally in use on the boat, make up a laminated 'grab bag list' of what to take with you. Stow grab bag extras where you know you will be able to find them quickly.

If you have time, take the dinghy as well.

Two-tier system

Pack a waist-bag (bum-bag) with small items that will make your eventual return to shore easier: money, credit card, spare spectacles, personal medication, the boat's papers, passports, house- and car-keys etc. If rescued by helicopter, you will not be able to take a large case with you. It's bad enough losing your boat but not being able to get home, or having to break into your car and house will make it worse.

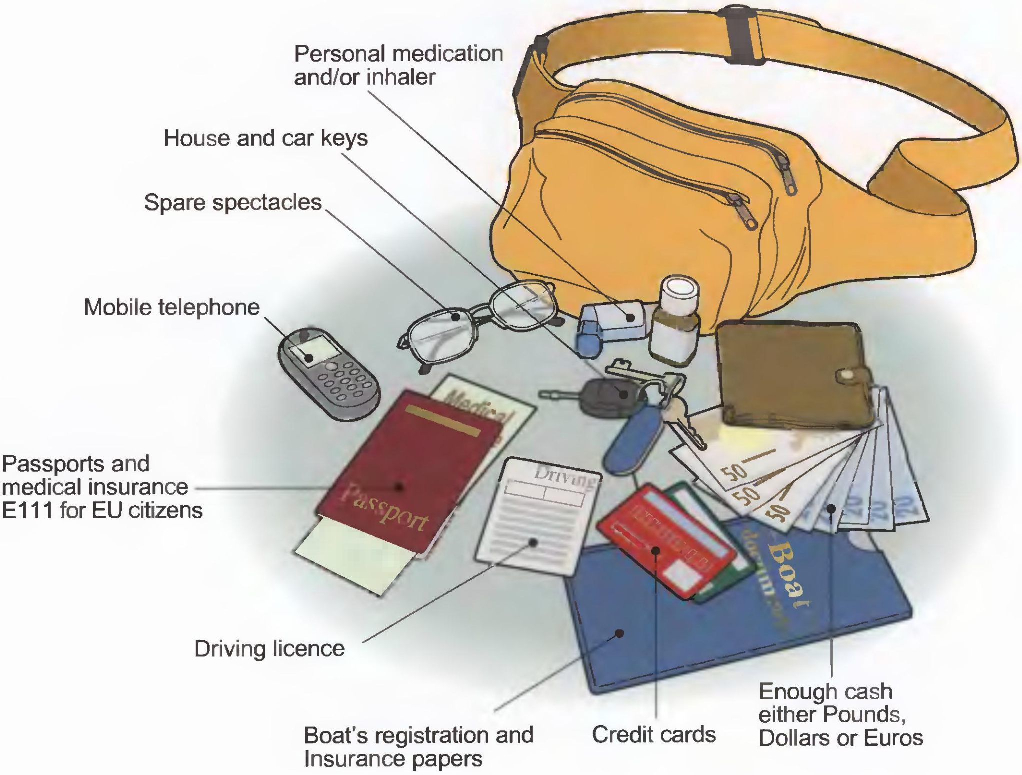
Personal medication and/or inhaler
House and car keys
Spare spectacles
Mobile telephone
Passports and medical insurance E111 for EU citizens
Medical
Passport
Driving
Driving licence
Boat's registration and Insurance papers
Credit cards
50
50
50
Boat documents
Enough cash either Pounds, Dollars or Euros

CHAPTER 9 : Fire Control

Fire prevention

- Keep the engine bay and electrics clean and tidy
- Install the correct size wiring
- Do not smoke below decks
- Take care with cooking fats and solvents
- Vent the engine compartment (petrol engines)
- Turn off the gas at source when not needed
- Check the gas system and use a gas alarm
- Ensure all crew know the gas routine
- Fit an automatic extinguisher in the engine compartment

Fire triangle

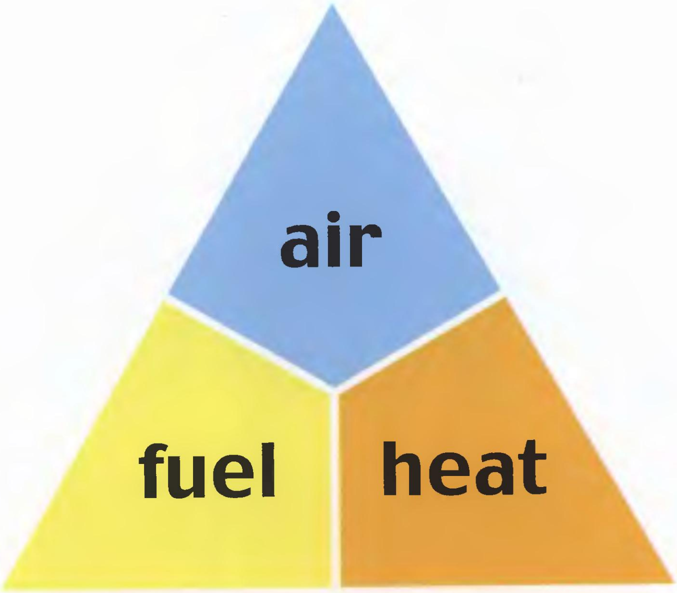

Fires require just three elements to burn. Remove just one of these and the fire will go out.

Fire blanket
Often overlooked but a very useful piece of safety equipment. Use to smother flames on liquid (cooking oil) or people, thereby depriving the fire of air. The blanket also makes a good protective shield when passing close to a fire. DO NOT store directly above a cooker.

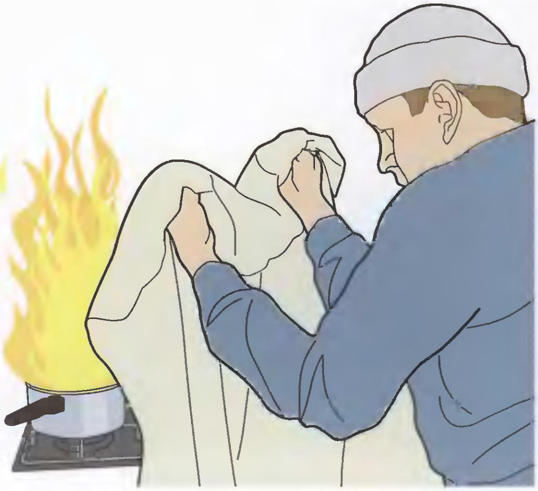

Fire extinguishers

All new extinguishers are red with a colour-coded area to show the type of extinguishant.

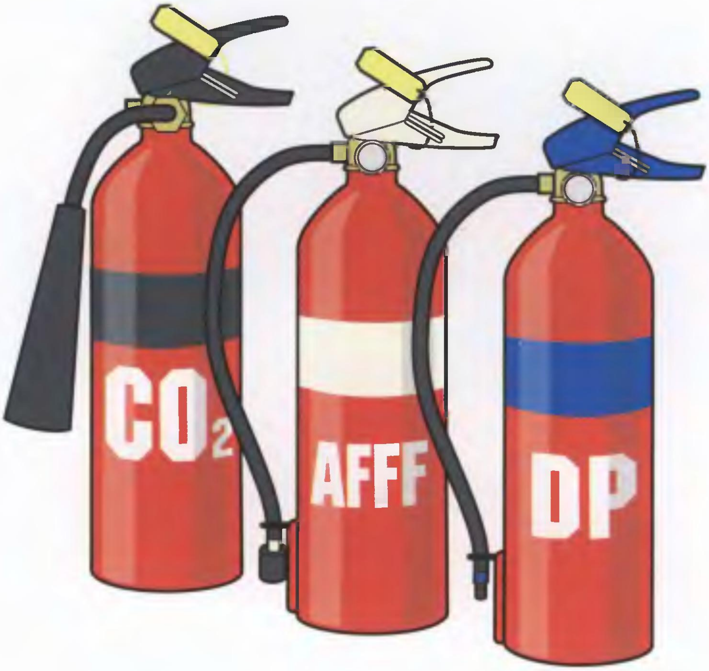

They are rated by the types and size of fire for which they are suitable.

A – carbonaceous (wood, glassfibre, upholstery)
B – flammable liquids
C – flammable gases

Sizes are related to a crib of wood for A-type fires and to a pan of flammable liquid for B-type fires, a typical rating is 5A/34B. The higher the number, the bigger the fire the extinguisher can tackle.

Water (A)

There is plenty of it around (sea) – collect it in a bucket (remembering to first tie a lanyard to the bucket). Water fire extinguishers are available, the water will cool the fire but they are large and heavy. Do not use water on liquid, gas, electrical or cooking-fat fires.

Dry powder (ABC) Blue

Use on solids, flammable liquid (petrol/diesel engine fires) and gas fires **but not cooking oil**. It is messy but very effective at knocking down flames, however it is suffocating when used in enclosed spaces, so care must be taken.

AFFF foam (AB) Cream

Use on solids, flammable liquid (petrol/diesel engine fires) **but not cooking-oil, gas or electrical fires**. Foam can run off solid objects and allow fire to reignite.

CO^2 (B) Black

Used in manual and automatic extinguishers although not often found on boats. Use on flammable liquid fires but not cooking-oil fires since the blast of gas can cause the oil to splash. **Do not use on people** – expanding gas is cold and can cause severe cold burns.

Automatic extinguishers for engine compartments

Use a single extinguisher of the correct size for the compartment.

Halocarbon gas

A number of environmentally friendly gas extinguishants are now available (FM 200, FE 36) for engine compartments. They are clean to use.

Micro-powder fine powder extinguishant

Chemically reacts with the fire.

Dry-powder (ABC)

Similar to a manual extinguisher. Can be used in naturally aspirated engine compartments. Dry powder will damage engines which have turbo - or super-chargers.

In the UK boats over 13.7 metres (45 feet) are required to fit minimum firefighting equipment, as prescribed by the Merchant Shipping Regulations Class XII Vessels.

Position extinguishers at exits from cabins so that they can be used to provide a means of escape. Fire blankets need to be easily accessible but not stored above a cooker. Fit an automatic extinguisher in the engine compartment.

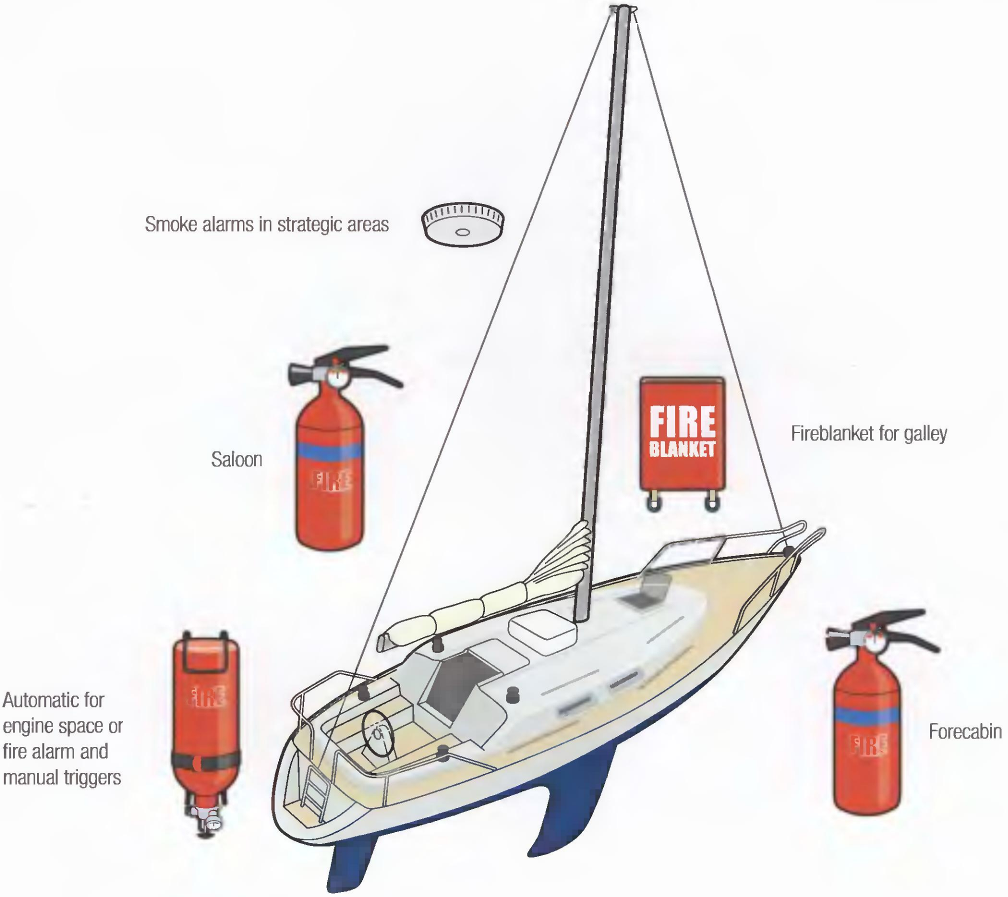
Smoke alarms in strategic areas
Saloon
FIRE
FIRE
BLANKET
Fireblanket for galley
Automatic for
engine space or
fire alarm and
manual triggers
FIRE
FIRE
Forecabin

Small extinguishers provide a very limited operating time. A 1kg dry powder extinguisher will give about 5 to 6 seconds of extinguisher.

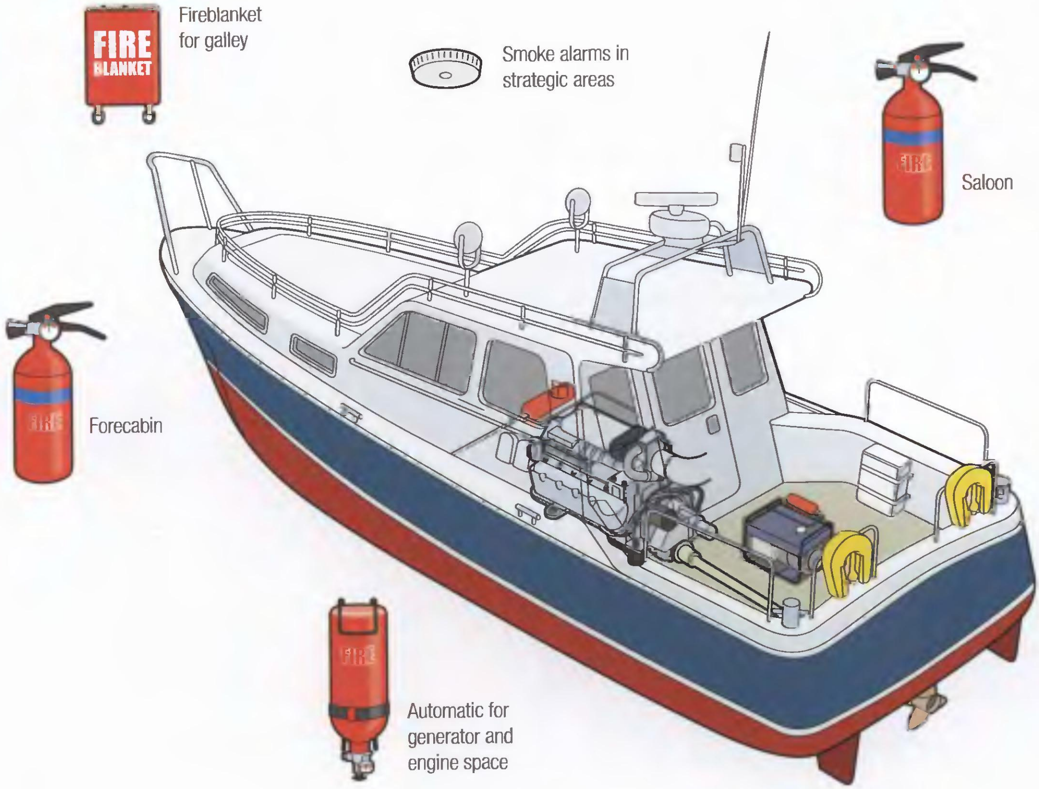
FIRE BLANKET
Fireblanket for galley
Smoke alarms in strategic areas
FIRE
Saloon
FIRE
Forecabin
FIRE
Automatic for generator and engine space

Maintenance

Regularly shake dry-powder extinguishers to prevent the powder clumping. Service all extinguishers at the prescribed periods.

Firefighting

Always keep a clear exit behind you when fighting a fire:

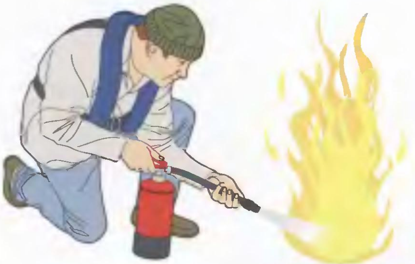

Aim the jet from a manual extinguisher at the base of the fire. Keep the extinguisher upright.

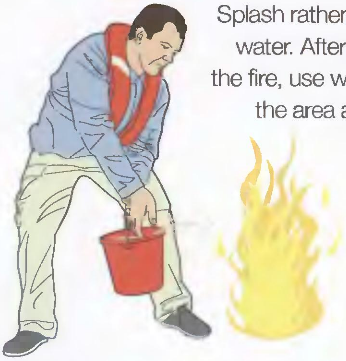

Splash rather than throw water. After putting out the fire, use water to cool the area and prevent re-ignition.

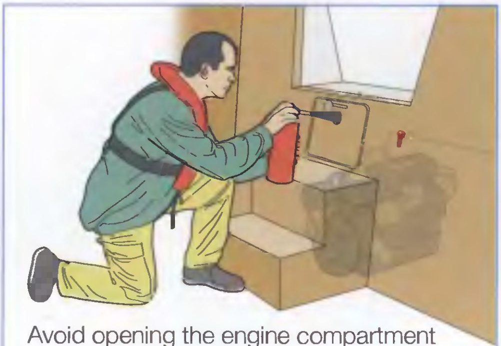

Avoid opening the engine compartment hatch unless absolutely necessary. If you have to, crack open the edge of the hatch and fire a manual extinguisher into the compartment.

Protect your hands behind the edges of the fire blanket. Smother clothing using a fire blanket.

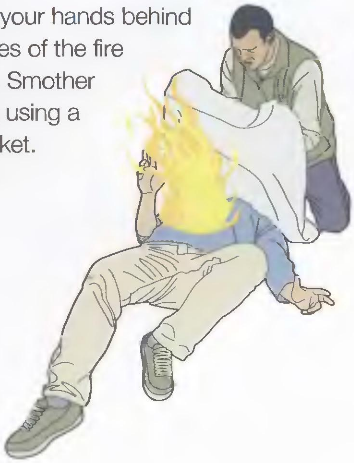

If you cannot extinguish the fire, prepare to abandon ship:

- If you are not already wearing lifejackets, put yours on and instruct the crew to do the same
- Send a Mayday
- Use flares
- Launch the liferaft

CHAPTER 10 : Raising the alarm

Are you in distress?
You are in distress when, in the opinion of the skipper, a vessel (boat, yacht, ship), or a person is in grave and imminent danger and requires immediate assistance.

Flares

Flares need to be readily accessible. In a sailing yacht, this usually means just inside the companionway (fig.1) ; in a motor cruiser, beneath or next to the helmsman's seat with some more in the wheelhouse and some on the flybridge (fig. 2). The number and type of flares to carry will depend on where you plan to sail. Race rules often specify a minimum pack that must be carried.

- Stow flares in a waterproof container lined with sponge foam or bubble wrap.
- Do not fasten the container's lid so tightly that others cannot open it.
- Handheld white anti-collision flares need to be immediately available and can be stowed in clips in the companionway or next to the helm. Keep flares out of reach of young children.

Fig. 1

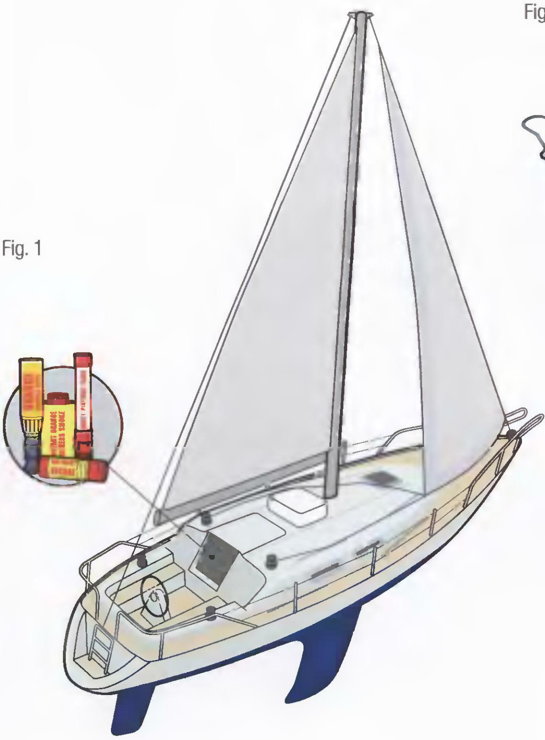

Fig. 2

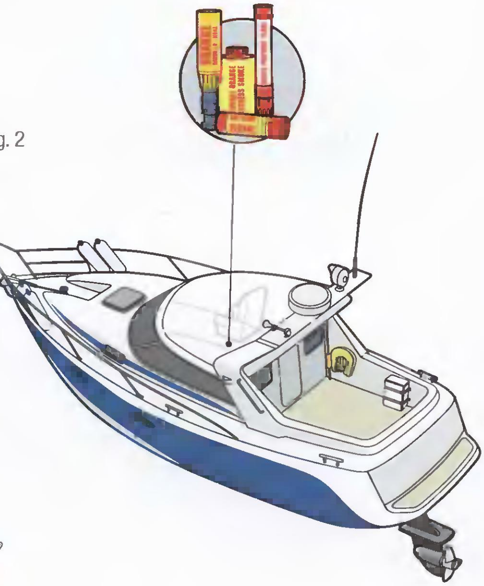

Stow flares where you can get to them quickly and easily.

Firing flares

Different makes of flares operate in different ways. Find out how to use your flares before you need them – Read Before You Need!

Which distress flare?

Red handheld flare

Burns for at least one minute at 15,000 candela. Visible up to 7 miles. Use day or night.

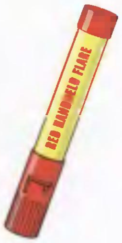

Red parachute rocket flare

Fires to a height of 300 metres or more and burns for at least 40 seconds at 30,000 candela. Visible for up to 28 miles. Use day or night.

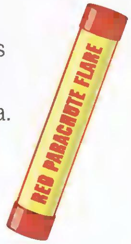

Handheld orange smoke

Produces bright orange smoke for at least one minute. Visible up to 7 miles. Some handheld smoke flares float and can be used to mark a MOB. Use during daylight only.
Not to SOLAS standard.

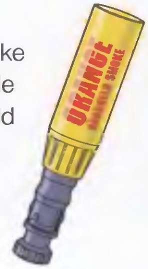

Buoyant orange smoke canister

Produces bright-orange smoke for over three minutes. Visible up to 7 miles. Use during daylight only. Can be used to mark a MOB.

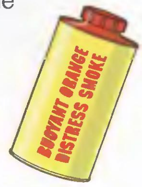

Day/night flare

Personal flare, waterproof. Used by divers and RNLI crew. Has a smoke flare at one end and a red flare at the other. Each end burns for 18 to 20 seconds. Visible up to 7 miles.
Not to SOLAS standard.

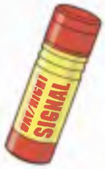

Mini-flare

Personal aerial flare fires up to a height of 60 metres. Comes with up to ten cartridges. Burns for up to 8 seconds. Fire two cartridges in succession. Visible up to 15 miles at night, 10 miles during day. **Not to SOLAS standard.**

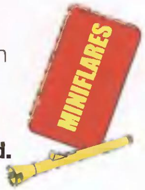

Aerial distress flares raise the alarm. Handheld distress flares can be used to raise the alarm but also to pinpoint your position.

Fire all flares with your back to the wind. Find a firm footing before operating the flare.

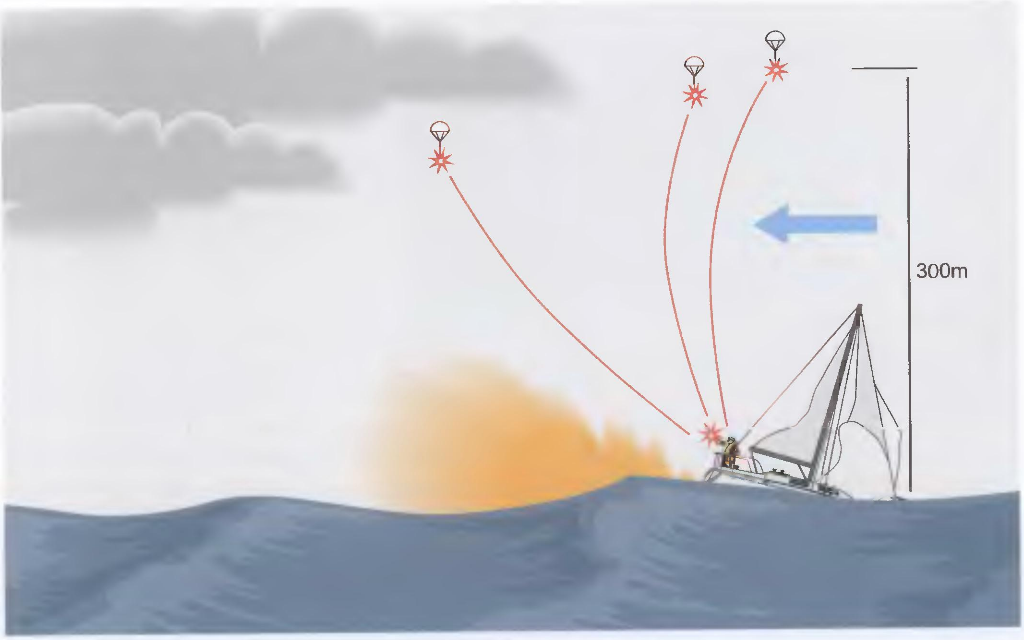

Parachute rocket flares reach a height of over 300m. For low cloud conditions it may be possible to fire the flare at an angle so that it deploys under the clouds. Orange smoke flares help to pinpoint your position.

Firing a parachute rocket flare:

1. Brace yourself in a firm position
2. Back to the wind
3. Remove any caps or covers
4. Check aloft to make sure it's clear of obstructions such as rigging, the flybridge or helicopters
5. Note which way up the flare needs to be fired
6. Remove the safety pin (if fitted)
7. Grip the flare firmly about two-thirds of the way up, the flare will recoil when fired
8. Gently squeeze the trigger upwards until the flare fires
9. Rocket flares tend to seek the wind. If the cloud level is low, fire at an angle downwind so that the flare ignites below cloud

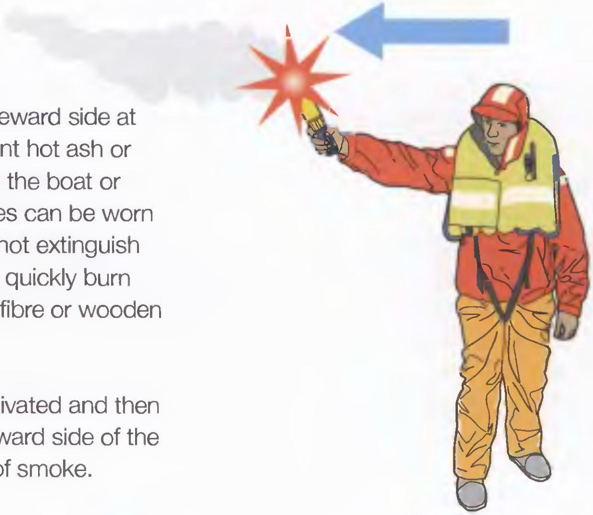

Handheld flares

Hold handheld flares over the leeward side at an angle of 45 degrees to prevent hot ash or burning material from damaging the boat or liferaft. Leather (gardening) gloves can be worn to protect the hands. Flares do not extinguish in water. If you drop a flare it will quickly burn through a liferaft or even a glassfibre or wooden hull.

Buoyant orange smokes are activated and then dropped into the sea on the leeward side of the boat or raft to build up a cloud of smoke.

Non-distress flares

Handheld white flares raise awareness and pinpoint your position to other shipping to prevent collision.

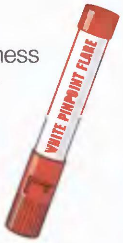

White parachute rocket flare

Used to illuminate the local area at night. Good for searching for MOB.

GMDSS

The Global Maritime Distress and Safety System was fully implemented in February 1999. It provides an International co-ordinated approach to calling for help and requires merchant vessels to carry at least two independent means of radio distress alerting. Leisure craft carry calling equipment voluntarily. It also provides for urgency and safety communications and for the provision of navigational and weather warnings.

The type of equipment suggested depends on where you sail. GMDSS divides the world into four areas:
Area A1 – within range of a shore-based VHF Coast Station fitted with DSC (up to 50 miles depending on antenna height).
Area A2 – within range of a shore-based MF Coast Station fitted with DSC (up to between 100 and 300 miles).
Area A3 – between latitudes 70 degrees N and 70 degrees S.
Area A4 – north of 70 degrees N and south of 70 degrees S (the polar regions).

Equipment covered by GMDSS includes – VHF, MF/HF SSB (radio tranceivers), Inmarsat A/B, C and M (satellite communications systems), Navtex, EPIRBs, PLBs and SARTs.

Raising the alarm

Mayday is the internationally recognised word to indicate distress.

DSC – Digital Selective Calling is a major part of the GMDSS. In a matter of seconds, pressing one button (correctly) will alert the rescue services and other shipping in the vicinity. The alert will tell them who you are, where you are and that you need help. DSC should also reduce routine radio traffic on Ch 16 making it easier to hear voice distress calls.

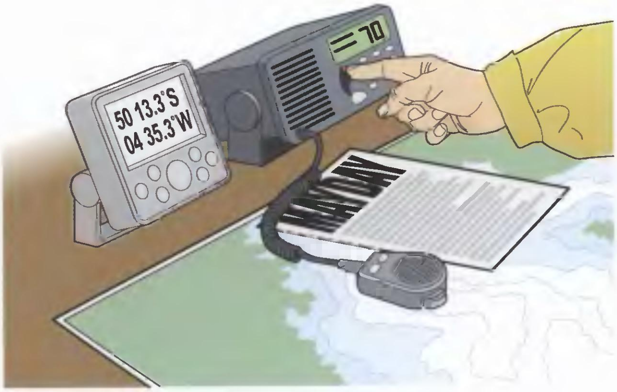

Using a VHF DSC radio

Make sure the radio is switched on and receiving position information from GPS. Different makes of radio transceiver operate in slightly different ways but the general approach shown below will operate most sets. However, make sure everyone on board knows how the radio works.

Lift the cover over the distress button and press and release the distress button once. This will (usually) provide a menu of types of distress.

If there's time, select the type of distress that suits your circumstances.

Press and hold the distress button again for five seconds (the radio will count down) until the call is sent. The radio may sound an alarm. Press CLR to stop the audio alarm. The radio will switch automatically to Ch 16. Make your voice Mayday call.

DSC distress calls will be sent automatically every four minutes, for as long as the radio can do so, until an acknowledgement is received.

Mayday by voice

Non-DSC VHF radios will need to be manually switched to Ch 16. Press the Ch 16 button and check that the radio is switched to high power. If not, push the H/L or 1/25 button.

Now make your voice Mayday call.

A reminder card next to the radio can help to make sure the right information is transmitted. In the panic of the moment it's easy to make mistakes. Try to speak slowly.

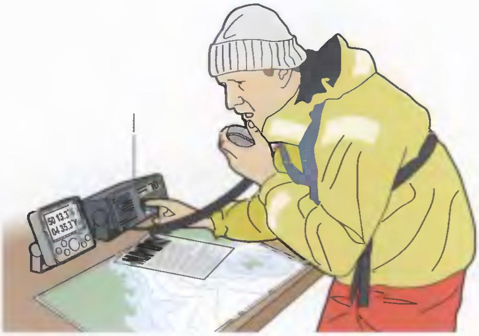

MAYDAY MAYDAY MAYDAY

THIS IS... yacht name spoken three times.

DSC - say your MMSI number.

MAYDAY... yacht name

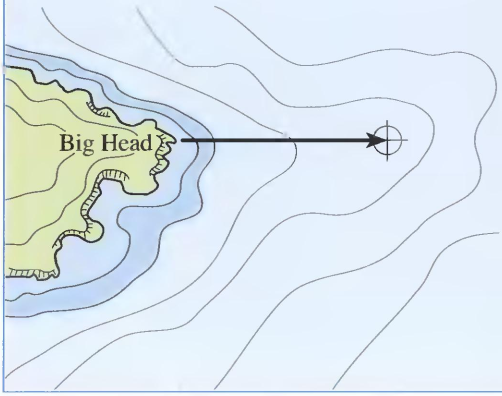

POSITION IS... give position in either Lat and Long or distance and bearing from a charted object

NATURE OF DISTRESS - describe briefly what the problem is, for example, sinking, MOB, boat on fire, stranded, broken down close to rocky lee shore etc

I REQUIRE IMMEDIATE ASSISTANCE

NUMBER OF PEOPLE ON BOARD
– don't forget to include yourself

FURTHER INFORMATION - anything else that may help rescuers, such as abandoning to liferaft, triggered EPIRB or PLB etc

OVER - the invitation to reply

You should receive a digital acknowledgement from the rescue services or ship, followed by a voice reply. If no acknowledgement or voice reply is received, check the batteries are on, the set is switched on and that Ch 16 and high power are selected. Repeat the procedure.

If still no reply, try a voice call on another Channel – Ch 6 or, if you are in shipping lanes, Ch 13. Prefixing the boat's name with the words 'Yacht' or 'Motor Cruiser' helps rescuers know what they are looking for.

VHF versus mobile phone

A mobile phone can be used to call for help but is a poor substitute for a marine VHF radio, the reasons being:

- VHF radios have a greater range.
- They raise the alarm with all in range who are listening.
- SAR services can determine and home in on VHF transmission, they have a longer battery life and can send distress calls in seconds (DSC).
- Mobile phones are designed for land use.
- They have poor coverage at sea and reduced range.
- They only alert the person you call. But in a real emergency a mobile phone is better than nothing.
- The emergency services operator will be the first one available and maybe located anywhere in the country.

Use the National Emergency Services number, in the UK, dial 999 (or 112 throughout Europe, 911 in the USA) and ask for the Coastguard.

Satellite mobile phones

Satellite phones are becoming a popular alternative to SSB radio.

In January 2006, experienced yachtsman Richard Wood and his partner Jetti were on passage from Nicaragua to Mexico aboard his catamaran *Eclipse* when they were caught in stormforce conditions. Wave heights were estimated at approaching 30ft (9m) with winds of over 70 knots. The sea anchor had been torn to shreds and they tried running downwind trailing warps and anchors but there was increasing danger of capsize. By late afternoon they decided they needed to call for help before things got worse and so set off their EPIRB and telephoned HM Coastguard in Falmouth (Cornwall, England) on their satellite telephone. The HM Coastguard contacted the Mexican authorities. Mr Woods continued to telephone the HM Coastguard every hour to check on progress and give a position and weather update. At 11pm, a helicopter from the USS *Ford* contacted *Eclipse* and the pair were lifted to safety.

Other Distress signals

Raising and lowering outstretched arms

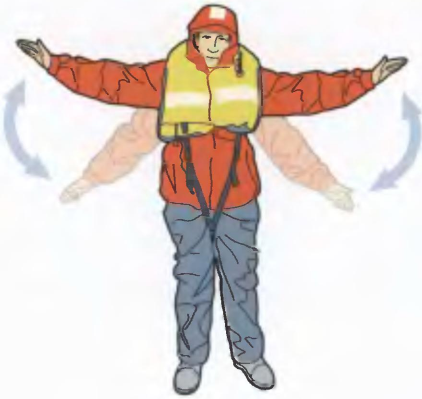

Continuous sounding of horn

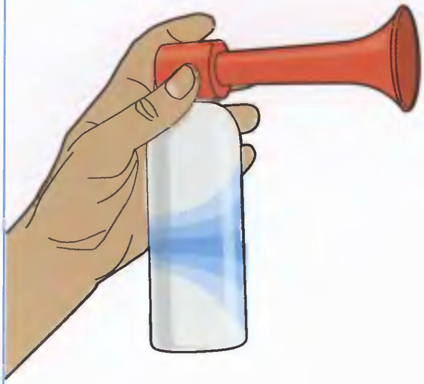

Morse SOS by light

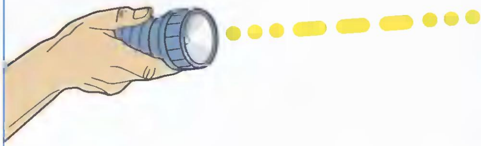

Black ball over a black square

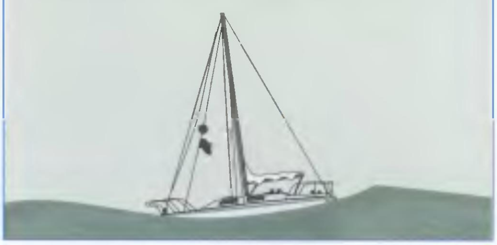

Smoke

Code flags N over C

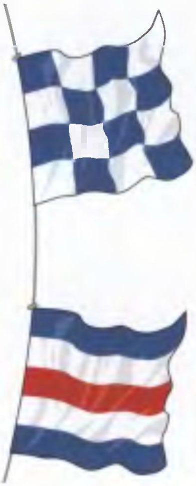

Upside down ensign

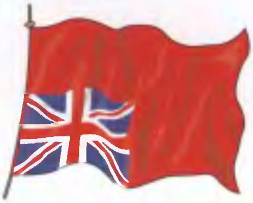

Divers

A diver on the surface will indicate distress by waving. They may also have raised their goggles onto the top of their head. Don't simply wave back and sail by – they are asking for help.

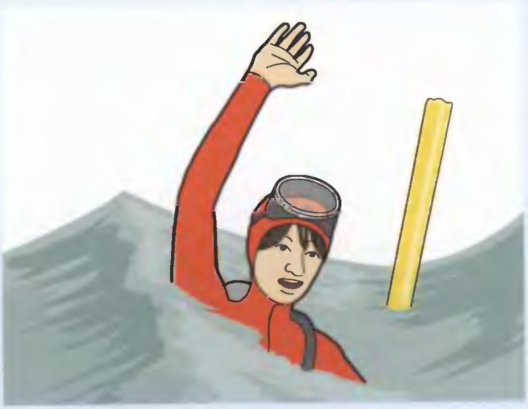

Using a heliograph (mirror or CD)

A heliograph is a simple way of drawing attention. Some come with a short stick and string but you can also use your hand. Using two fingers as a sight, reflect the light onto the two fingers. The target will then see the reflection off the mirror. Practise with a CD but be aware that its shiny surface will dissolve in water.

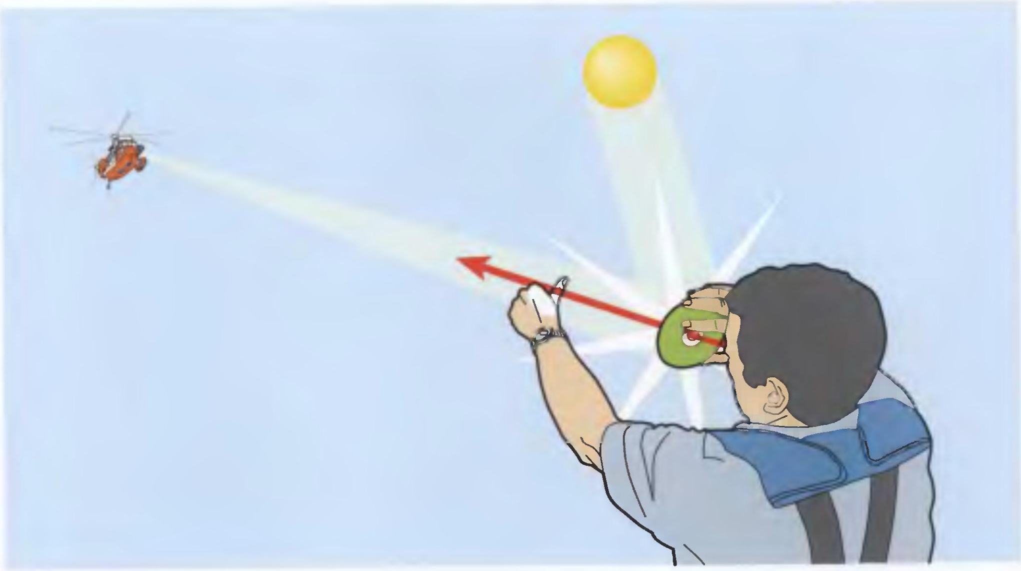

Need help but you are not in distress

You can also ask for assistance by making a Pan Pan call on the radio preceded by an Urgency Alert on DSC (if fitted).

By Voice

PAN PAN (Three times)
ALL SHIPS (Three times)
THIS IS… name of boat x 3 (+ MMSI if fitted with DSC)
POSITION IS… give position in either Latitude and Longditude or distance and bearing from a charted object
NATURE OF PROBLEM I am broken down and require a tow
NO OF PERSONS ON BOARD don't forget to include yourself
OVER

Let the SAR authorities know as soon as something goes wrong. Don't wait until the last minute.

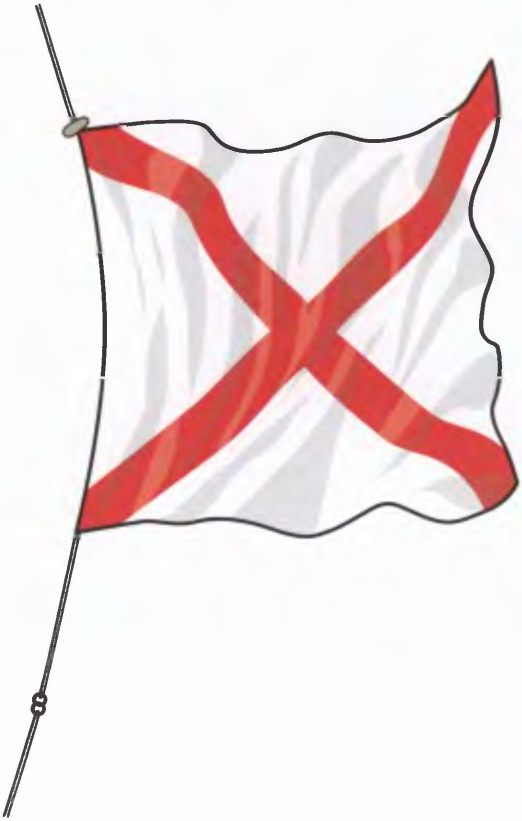

Code flag V indicates 'I require assistance'.

EPIRBs

An Emergency Position Indicating Radio Beacon will alert the Rescue Services by sending a distress signal via satellites. Two sets of satellites pick up the signals: Inmarsat and Cospas-Sarsat. Inmarsat satellites are in geostationary orbit. Cospas-Sarsat satellites are in a lower polar orbit and circle the earth approximately every 12 hours.

There are two types of EPIRB: standard and GPS. Both transmit a signal on 406MHz, which includes an identification code.

GPS-enabled EPIRBs also transmit position information that Inmarsat satellites will immediately download to an Earth station. The position information may come from a built-in receiver or from a link to the vessel's own GPS receiver.

Standard EPIRBs' positions are found using Doppler Frequency Shift theory by Cospas-Sarsat satellites. They require two satellite passes before an approximate position can be found. Each satellite will hold and then download data as soon as a ground station comes into range.

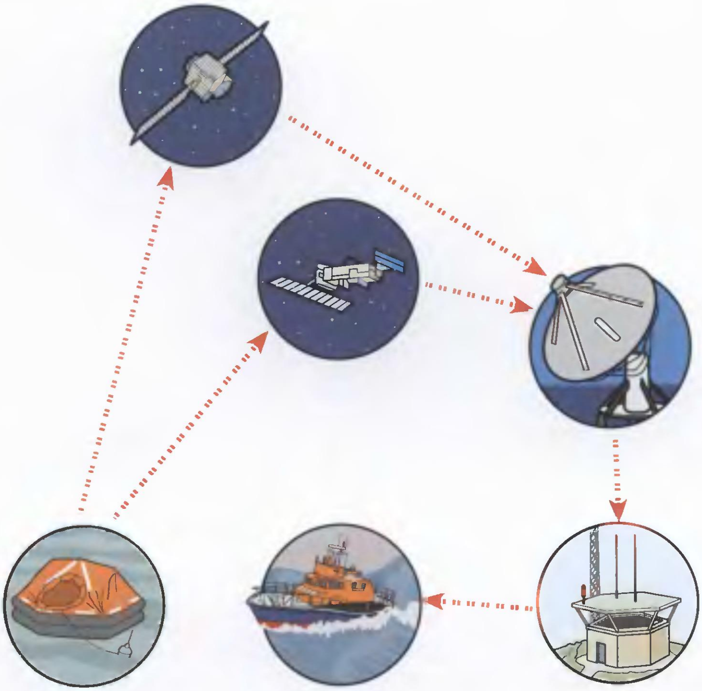

When the EPIRB is activated, it transmits a signal via satellite to an Earth station. The Earth station sends it to the appropriate rescue co-ordination centre, who will task a ship or local rescue services to your position.

GPS EPIRB – provides a distress alert immediately and a position within five minutes accurate to 100 metres.

Standard EPIRB – provides a distress alert immediately, if the Earth station is within the satellite range, and a position to within 3 nautical miles (5km) within 40 to 120 minutes depending on your location. EPIRBs can be activated manually or automatically on contact with the sea. Float-free versions are also available. Built to SOLAS standards, they have a battery life of 48 hours and will float upright.

Also built into the EPIRB is a 121.5MHz homing beacon and SAR craft use radio direction finders to home into the signal.

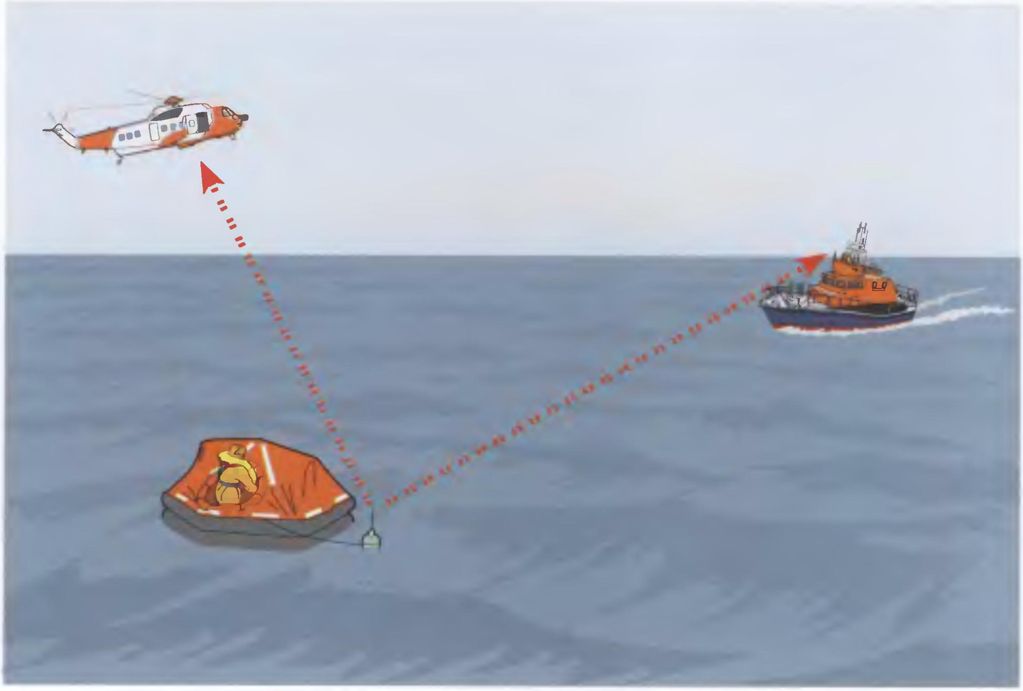

All 406MHz EPIRBs are required to be registered in the country of their vessel's flag. These details are used to check the authenticity of a distress alert. Batteries need to be replaced at the manufacturer's required date. Check the instructions and make a note of the date they need to be replaced - always replace the batteries.

PLBs

Personal Locator Beacons, sometimes referred to as PEPIRBs, are similar to EPIRBs but they are smaller and are designed to be worn by an individual crew member. Most countries require 406MHz PLBs to be registered.

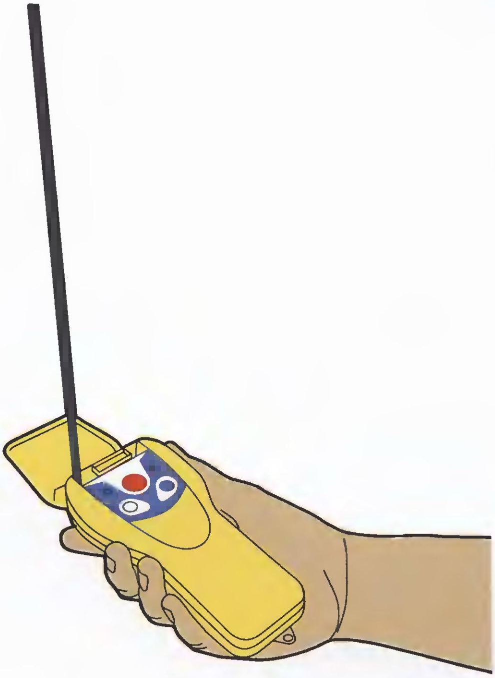

As with EPIRBs, there are two 406MHz PLB types. Both operate with the Inmarsat and Cospas-Sarsat satellites in the same way as EPIRBs:

- GPS PLB – provides a distress alert immediately and a position within five minutes accurate to 100 metres.
- Standard PLB – provides distress alert immediately and a position to within 3 nautical miles (5km) within 40 to 120 minutes depending on your location.

Also built into the PLB is a 121.5MHz homing beacon. SAR craft use radio direction finders to home into the signal.

Unlike EPIRBs, PLBs do not float upright. They have a battery life of 24 hours and can only be activated manually.

121.5 PLBs

Designed to be worn by individual crew members, 121.5 PLBs can trigger an onboard alarm to warn crew that somebody has gone over the side. A radio direction finder can also be used to home in on the MOB. Always tell the SAR services if a MOB may be using a 121.5 PLB.

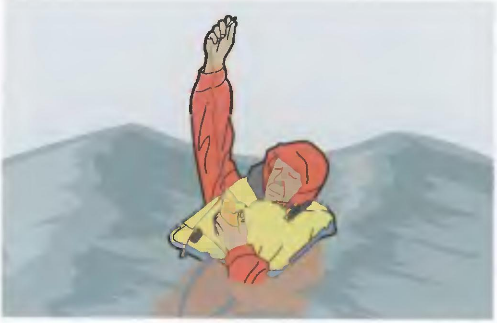

The MOB should hold the antenna as high as possible to give the greatest range.

MOB alerting devices

Proprietary MOB alerting devices that work on non-SAR rescue frequencies are also available. Some are capable of being linked to the GPS MOB button or an autopilot or may come with an appropriate radio direction finder.

RNLI MOB Guardian

This is a new worldwide system, developed by the UK-based Royal National Lifeboat Institution, which continually monitors boat and crew. An onboard transceiver sends regular position and condition data updates via the Iridium satellite system to the RNLI Operations Centre. If data is not received, the RNLI alerts the local rescue co-ordination centre to the vessel's last known position.

The transceiver also includes personal proximity alarms to be worn by each crew member. If crew goes out of range of the transceiver, it activates an alarm and alerts the RNLI Operations Centre.

SARTs

A Search-and-Rescue Radar Transponder is an alert and position-finding electronic device that produces a distinctive 'echo' on the screen of any 9GHz Radar. Previously only used on commercial vessels' liferafts, SARTs are now available sufficiently small enough to be fitted as an optional extra in most leisure liferafts.

Fig. 1

Fig. 2

Fig. 3

On initial contact with a SART, the radar shows a line of dots giving the range and direction of the casualty (fig. 1). As the radar closes on the SART, the dots broaden to arcs (fig. 2). When the arcs make an almost complete circle the SART is within 1 mile (fig. 3).

Range will depend on the height of the SART and the height of the Radar scanner. A SAR helicopter at 1000m (3000ft) can pick up a SART 48 miles away. A radar scanner 7m above the waterline will 'see' a SART at about 6 miles.

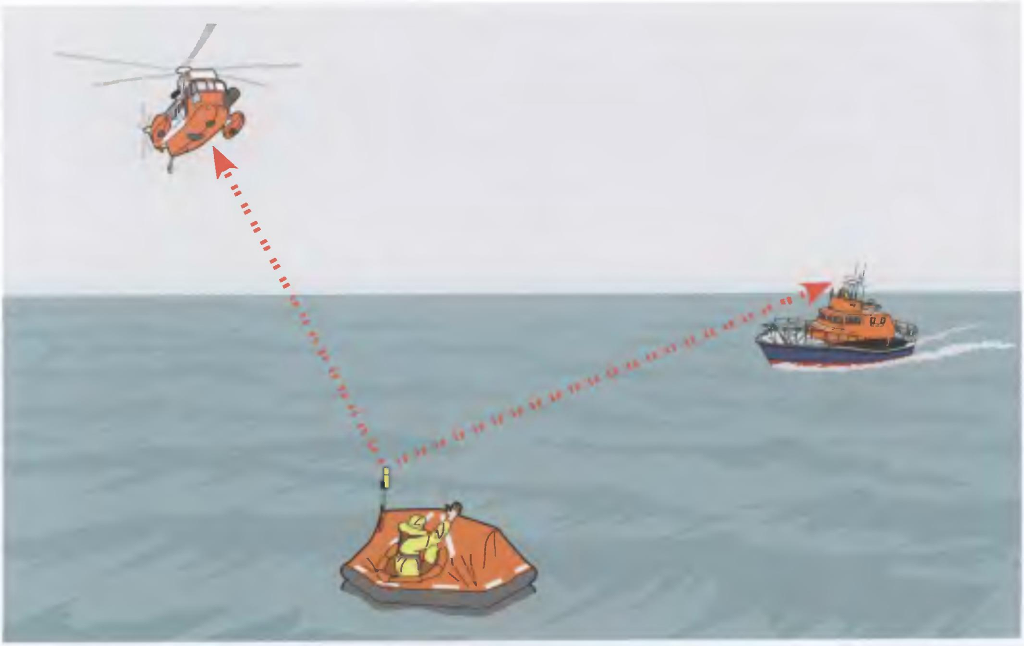

The SART starts transmitting when it receives a radar signal. Switch off your vessel's radar before abandoning ship, to prevent your SART from transmitting before rescuers are in range.

Position the SART as high as possible to achieve the greatest range.

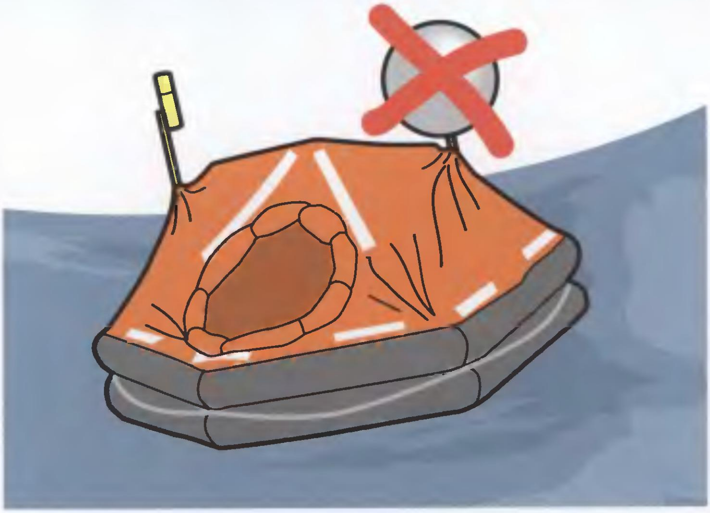

SARTs and inflatable radar reflectors conflict – use only one at a time.

CHAPTER 11 : When to abandon ship

There are two main reasons to abandon ship:

1. If your boat sinks.
2. If your boat catches fire and the fire cannot be controlled.

If there is no grave danger of the boat immediately sinking, then stay with your boat because:

- It carries far more supplies than your liferaft
- It will probably offer better shelter from the elements
- You may be able to find a way to jury-rig a repair and improve your conditions
- It makes a much larger target for rescuers to see

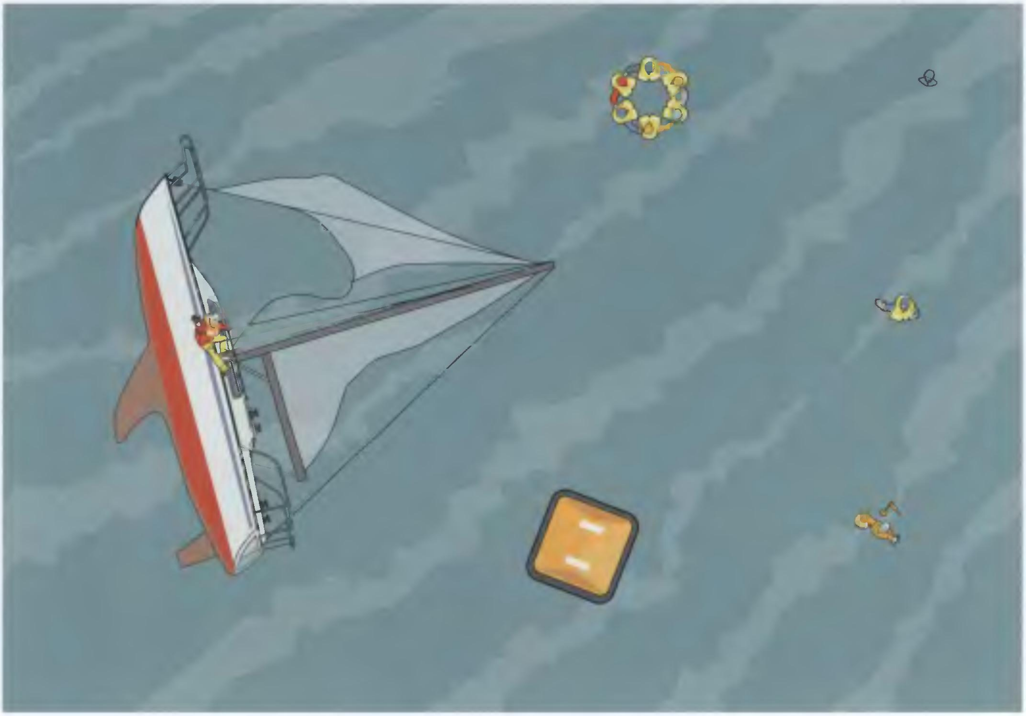

Your boat is full of supplies – only leave it if it sinks or catches fire. It makes a much bigger target for rescuers to find compared to a liferaft, a huddle of survivors, a single person with a lifejacket and/or a dan buoy or a person without a lifejacket.

What you need to survive

To ensure your survival you need the correct equipment with the right knowledge and, most importantly, a belief and will to survive. The four tenets of survival are protection, location, water and food. To this end, the survivor requires:

Equipment

- Personal survival aids
- Life-saving appliances
- Grab bag

Knowledge

- What equipment you have
- How to use it properly
- Have a strategy ready

Will to Survive

- Believe you will be rescued
- Avoid blame
- Positive attitude

I will survive!

The greatest factor in any survival situation is the positive mental attitude of the would-be survivor, some call it the will to survive. There are numerous examples where a sheer determination not to be beaten, to keep going despite the odds has made the difference between life and death. The belief that, no matter how bad the situation, you will pull through has to be emphasised to every casualty; a liferaft or semi-submerged boat is no place for negative thinking.

Obviously, a will to survive is not enough on its own. Training, preparation and a good understanding of the factors that will affect your survival will bolster that belief. Positive leadership will also improve a crew's chances of survival.

Be aware that tiredness, cold, seasickness and injury add to demoralising a crew and that the loss or death of a crew member can have devastating effects on the survivors' will to live.

Tony Bullimore's 60-foot yacht capsized in 60-knot winds in the Southern Ocean when it lost its keel during the Vendée Globe single-handed round-the-world race in January 1997, 1700 miles from Perth, Western Australia.

Without a keel, the boat remained inverted (upside down) with Bullimore caught inside the hull, waist-deep in water. Realising that he would have sufficient air to breathe and that the boat was unlikely to sink, he assessed his situation and put together a survival strategy.

He activated his EPIRB knowing that help would take several days to reach him. Although he was wearing a drysuit, he realised that he had to get out of the freezing water. In total darkness, he managed to jury-rig a hammock from cargo nets. He was aware that spotter planes may not realise that he was still alive inside the hull so he made several attempts to swim out of the cabin to release his liferaft so that it would float free next to the yacht. Each attempt sapped his strength and resulted in two hours of intense shivering back in his hammock. But his will to survive forced him to keep trying. For four days he lived on limited supplies of water and chocolate. The Royal Australian Navy rescued him, hypothermic and dehydrated but alive.

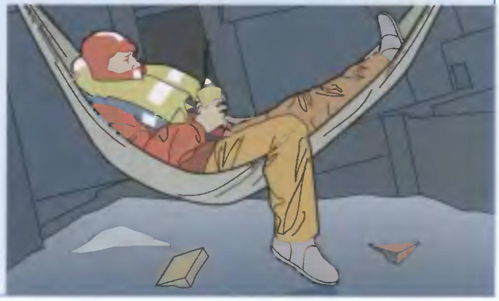

When disaster strikes

Crew will probably react in one of three ways when a yacht founders:

- 10 to 20% will remain calm and be aware of the situation and will respond appropriately
- 75% will be stunned and bewildered, they will be unable to judge what to do for the best and be unable to concentrate on a task
- 10 to 15% will react inappropriately and may not realise the seriousness of their predicament, they may scream, weep or be unable to move

Applied to a crew of six, these figures suggest that there will be;

- one person who is able to react effectively
- four who will need goading to undertake tasks
- one who may completely lose touch with reality and could endanger themself or their colleagues.

A good leader, who may or may not be the skipper, will stay positive, prevent recriminations, take stock of the situation and organise and allocate tasks to the survivors. Good planning and preparation, training and knowledge of what to expect will increase the number of crew who react appropriately.

CHAPTER 12 : Abandoning ship

Don't leave your boat until your boat leaves you! It is packed full of supplies and useful equipment that will aid your survival.

A few tasks, carried out before you abandon, may save your boat: emptying water tanks, therefore turning them into buoyancy tanks, closing the engine-cooling-water seacock, and disconnecting the raw-water feed and using the engine-cooling system to pump out the bilge.

Drills for abandoning ship

Prior to abandonment

1. Make everybody aware that you are preparing to abandon ship. **Stay calm.**
2. Activate the EPIRB. Press DSC distress button on VHF / HF radio. Make a voice distress call.
3. Fire parachute rockets if someone is likely to be within 30 miles.
4. Get everyone on board to put on layers of warm clothing. Extra clothing will not weigh you down in the water. Initially, it will help you float.
5. Waterproof clothing will reduce cold-water shock and heat-sapping water flow around your body. Use an immersion suit. **Keeping warm is a priority.**
6. Check everyone has donned lifejackets correctly – use a buddy system to speed things up. Inflate lifejackets before entering the water.
7. Have a drink (non-alcoholic) and take anti-seasickness tablets.
8. Find the grab bag and as many useful extras as you can lay your hands on.
9. Standby to launch the liferaft in case the vessel has to be abandoned quickly.

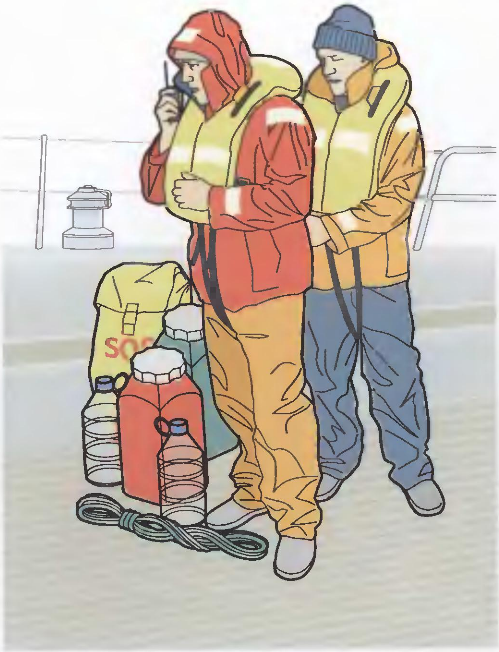

Priorities are donning lifejackets, sending Mayday, initiating the EPIRB and launching the liferaft. Anything else is a bonus.

Do not launch the liferaft earlier than you need it because it is difficult to hold a raft alongside in a rough sea.

Launching and entering liferaft

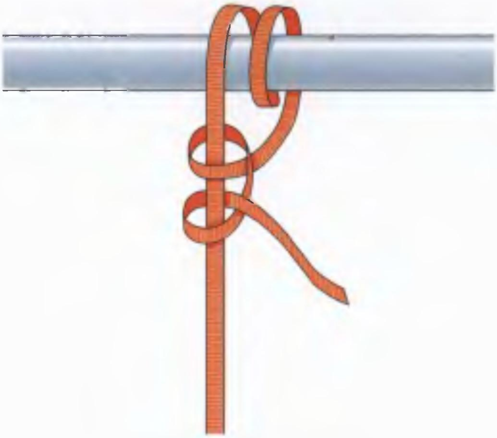

Prior to launching the liferaft

1. Secure the painter to a strongpoint – ask another crew member to check the knot. A round turn and two half-hitches is sufficient but any knot that holds will do.

2. Check that the water is clear of debris or the raft may be damaged as soon as it inflates.

3. Launch the liferaft to the leeward (downwind) side of boat.

Liferafts are heavy and difficult to manhandle – a six-man raft can weigh between 25kg (55lbs) and 75kg (165lbs) depending on its quality, type and contents. Launch them from as near to the stowage site as possible. It may need two people to lift the liferaft canister.

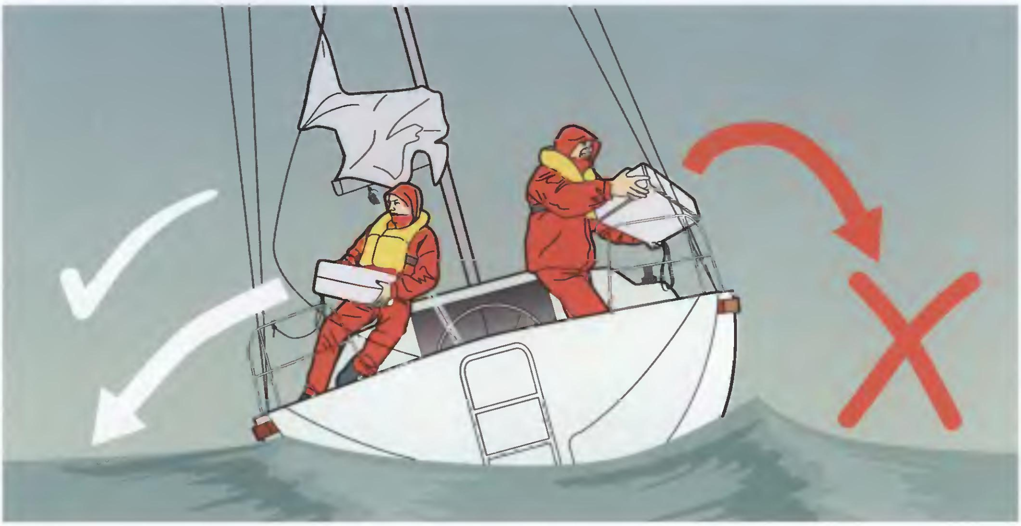

Launch the raft to leeward

Launching the raft to leeward will provide some protection from the prevailing conditions when boarding it.

The freeboard is likely to be lower on the leeward side, making the raft easier to board and reducing the chances of the raft damaging itself on the side of the boat.

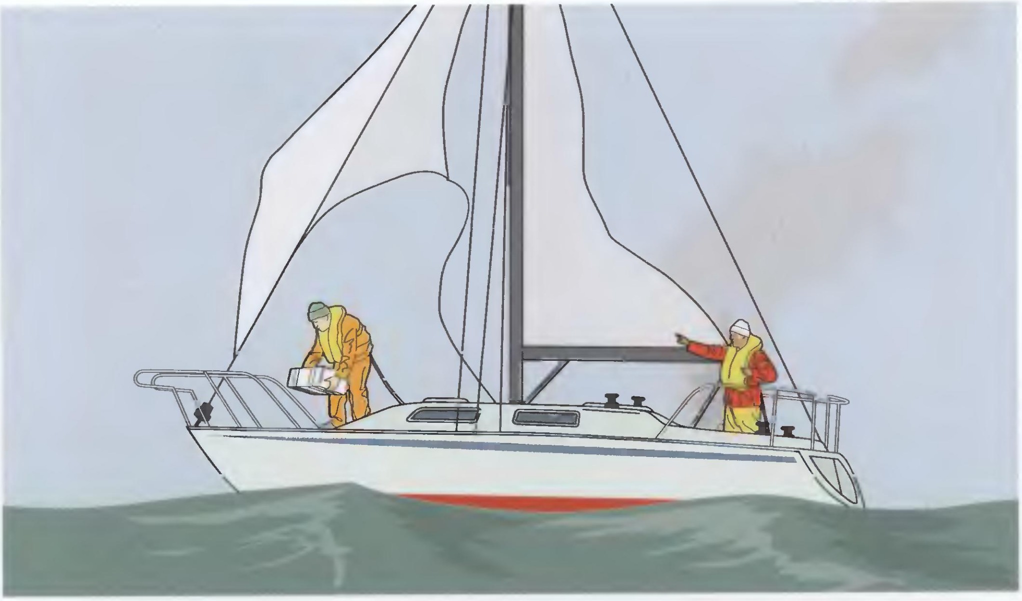

However, you may have to move the raft to a safer position – for example, because of a fire. Take care not to slip and injure yourself or lose yourself or the raft overboard. Be aware of other dangers such as flogging sails.

Launching the liferaft

4. Pull the painter. It's likely to be up to 10 metres (33ft) long.
5. When the painter is fully extended, a sharp tug will trigger the gas inflation mechanism.
6. The raft will inflate in 30 to 60 seconds.

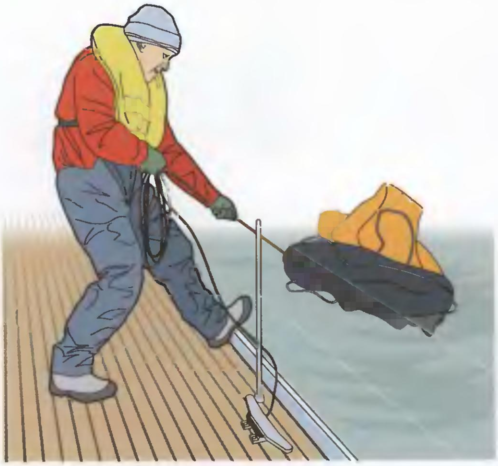

You may hear gas escaping during and shortly after inflation. This is excess CO_2 being expelled through the pressure-relief valves and is normal.

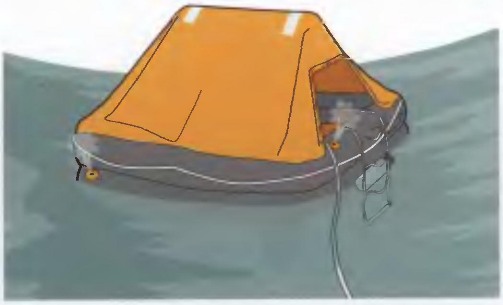

Launching the liferaft to windward may make it more difficult to board. Sharp barnacles on the bottom of the boat may puncture the liferaft tubes.

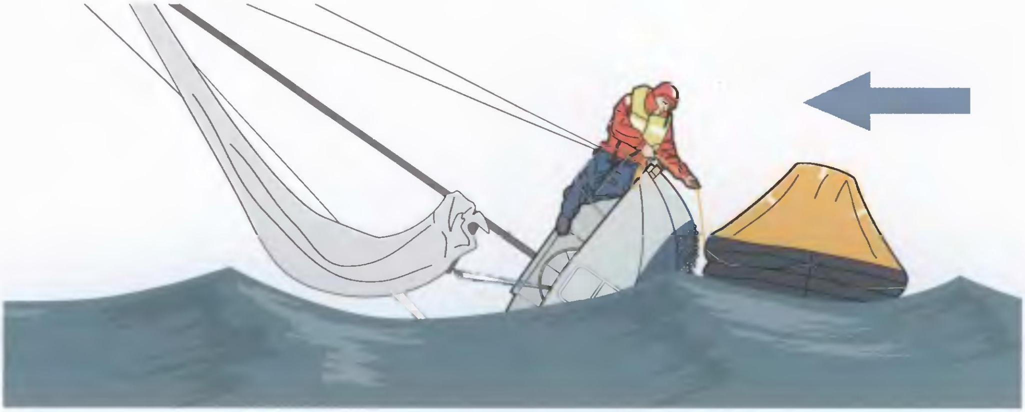

Don't let the liferaft drift away. In strong winds it may be too difficult to pull it or even winch it back to the boat, leaving you with no raft.

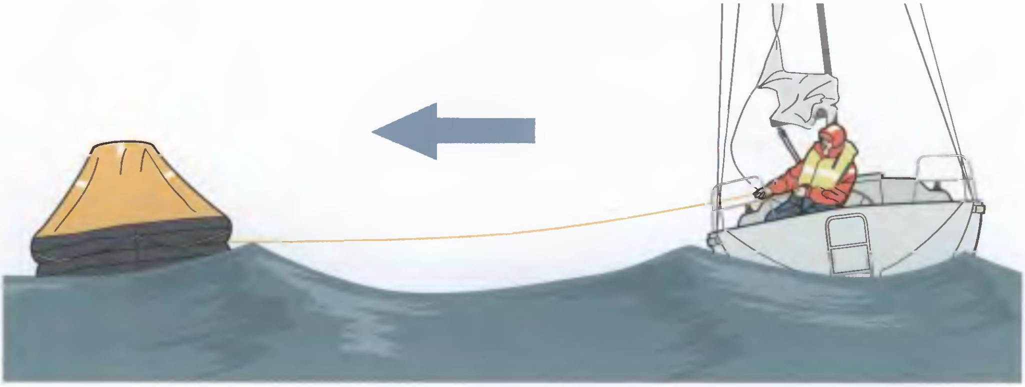

If the raft inflates upside down or capsizes before you can board it, try to right it while still onboard the vessel.

STAY OUT OF THE WATER – COLD KILLS

Boarding the liferaft

Remove sharp objects before entering the raft from your person such as keys attached to a belt clip.

Try to climb directly into the liferaft from the boat. The strongest and heaviest person should board the raft first to stabilise it and to help others in. As each person gets in they should move out of the way of the entrance and balance the raft as the others board.

Avoid jumping if at all possible. You risk injuring yourself or others already in the raft. Do not jump onto the canopy. If you have to jump, the maximum recommended height is 2 metres.

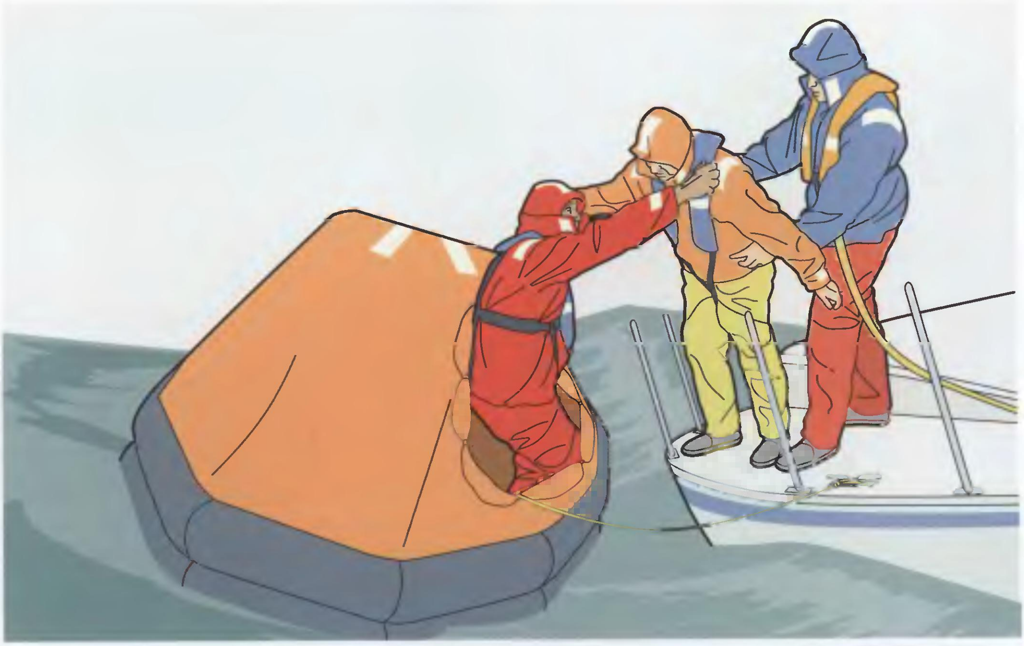

Spread the impact load by holding your feet slightly apart and your arms out.

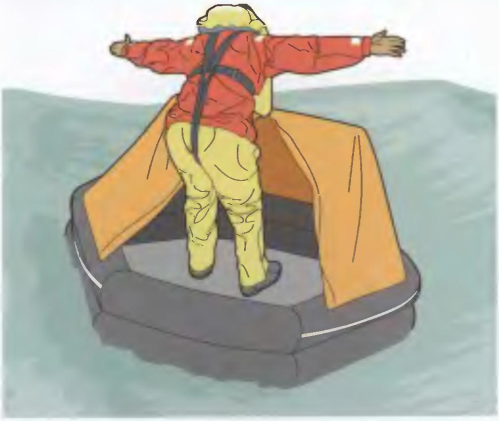

REMEMBER TO LOAD THE GRAB BAG AND EXTRAS

Leave any injured to the last but one. If they enter early, other people may make their injuries worse if they land on them. Get the injured aboard the raft before the last able person leaves the boat.

Entering the water

Only enter the water if you are unable to board the raft direct from the boat.

If possible, enter the water slowly, either by lowering yourself on a rope or by using the boat's fixed or emergency boarding ladder, this will minimise the effects of cold-water shock.

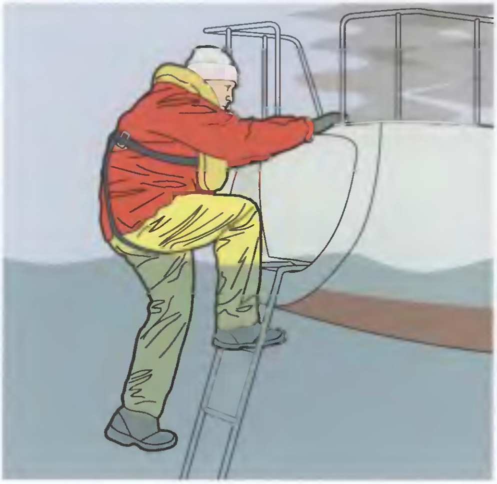

If a slow entry is not possible:

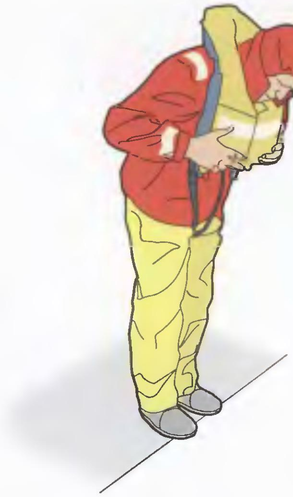

1. Go to the side of the boat where the freeboard is least or upwind of fire and smoke.
2. Steady yourself and check below for debris and other people.

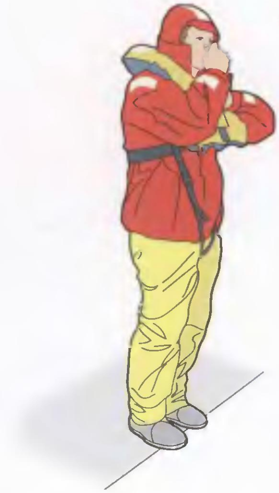

3. Inflate your lifejacket.
4. Place one arm over the top of the lifejacket to prevent it from jumping up when you enter the water.

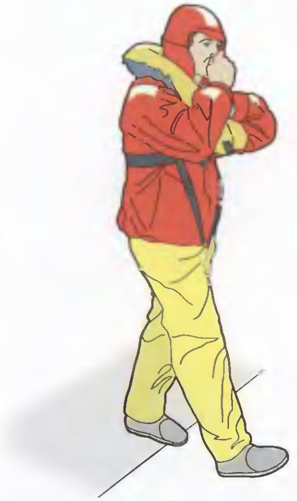

5. Use your other hand to cover your airway.

6. Look ahead to stop you from toppling forward – check the area is clear and step off the side of boat.

7. Put your feet together and enter the water in an upright position.

8. Hold this position until you have surfaced and regained control of your breathing.

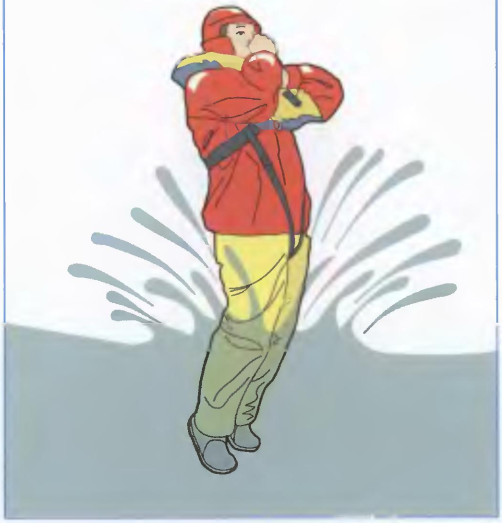

At night or in heavy weather, it may be prudent to clip your safety harness to the liferaft painter and pull yourself along it until you reach the raft.

While waiting to get into the raft, hook your arm through the lifelines fixed around the outside of the raft. Cold numb hands will be unable to grip the lifeline. Do not let go until you are on board the raft.

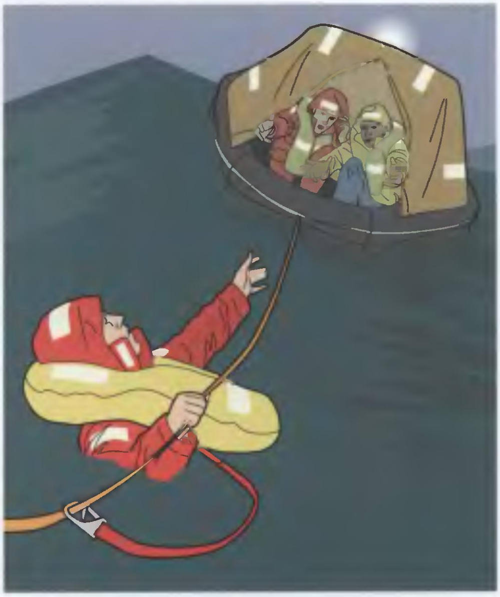

Getting into the liferaft from the water

The cumbersome shape of an inflated lifejacket, the weight of wet clothing and the strength-sapping effects of cold water make it very difficult to climb unaided into a liferaft from the water. The strongest and fittest person should enter the raft first. They can then help others to board.

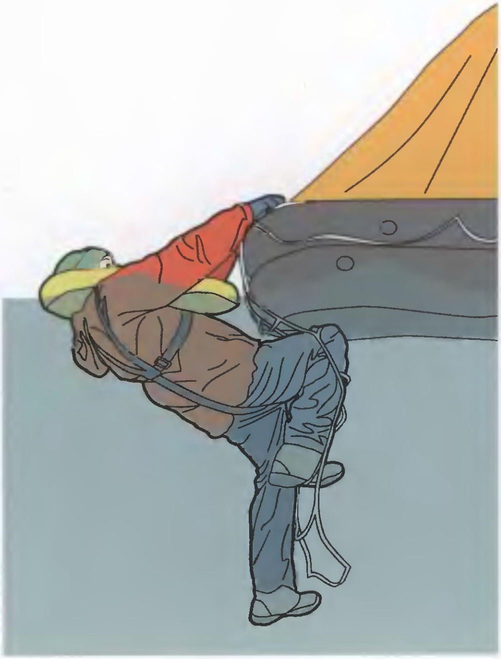

All liferafts have ramps or ladders to aid entry. Get a foothold on the ladder and a firm grip on the handhold on top or just inside the top tube.

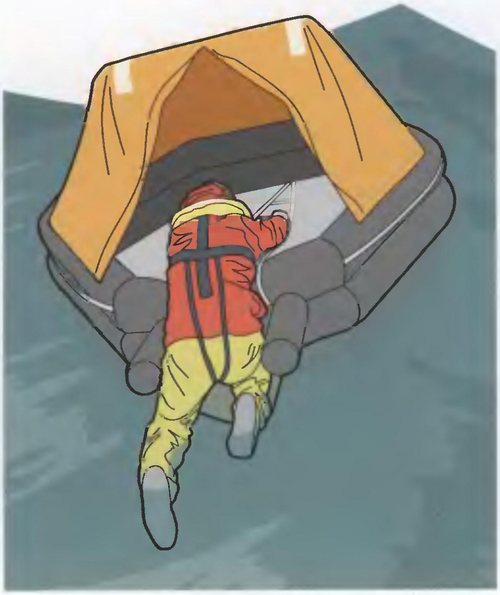

Partially deflating your lifejacket may make it easier to reach handholds inside the raft, especially if wearing a 275N lifejacket.

Rescuing survivors

Get casualties into the raft quickly. The two strongest survivors should be on either side of the entrance and grab hold of the casualty's lifejacket straps and pull them face-first up and into the raft.

GET EVERYONE OUT OF THE WATER FAST.

Make sure that semi-conscious casualties are not lying face-down in the water in the bottom of the raft. Look out for flailing legs.

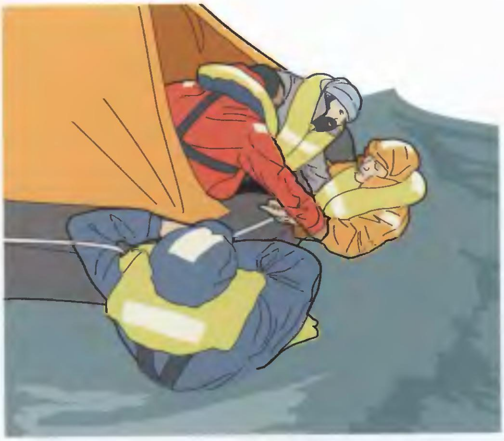

Once you are in the raft, move to the opposite side to help keep the raft balanced.

Keep a count of survivors. Continually check that none drift away. Communicate with those in the water. Help each other.

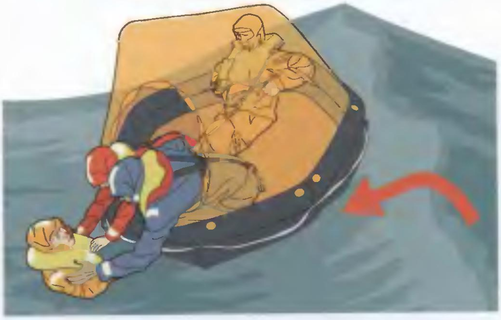

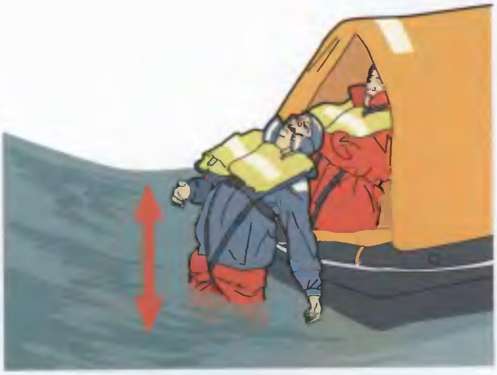

Pulling casualties in backwards by using their inherent buoyancy and dunking them first is slow and difficult and may also injure their back.

Righting a capsized raft

With the correct technique, one person can right any raft with a capacity of up to 12 persons.

Whenever possible right the liferaft while still on the boat

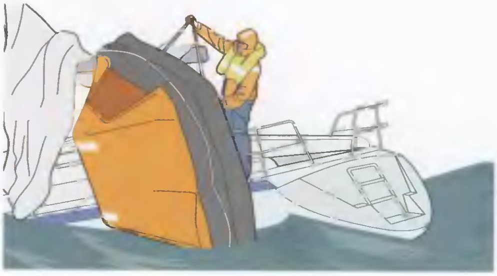

1. Using the lifelines, move around the outside of the raft to the gas bottle. There should be a sign or symbol saying 'right here'.
2. Reach up and find the line or webbing handles on the bottom of the raft.
3. Pull yourself up on the edge of the raft.
4. Draw up your knees and place them in the fold of the tubes. Use them as a fulcrum. Lean back while holding on the righting line.

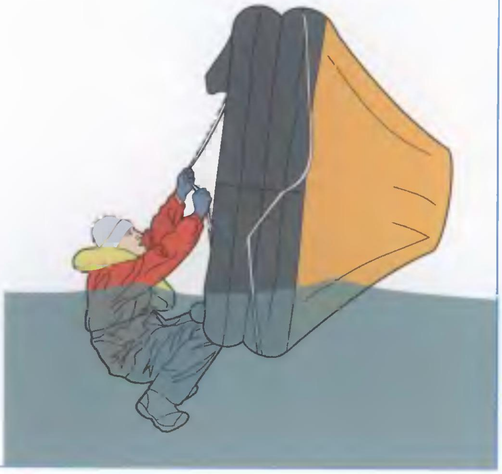

5. Pull the raft over on top of you. It's much easier if you can get the wind underneath the raft as you begin to lift.
6. Hold on to the line and stay on your back. As the raft comes over, keep your hand up to create an air pocket under the raft.

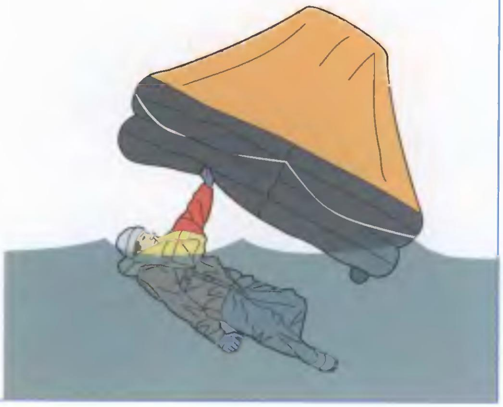

7. Use the line to pull and guide yourself out from under the raft. Mind your head on the gas bottle.

8. Do not let go of the raft. If the wind catches it, it will be blown away faster than you can swim.

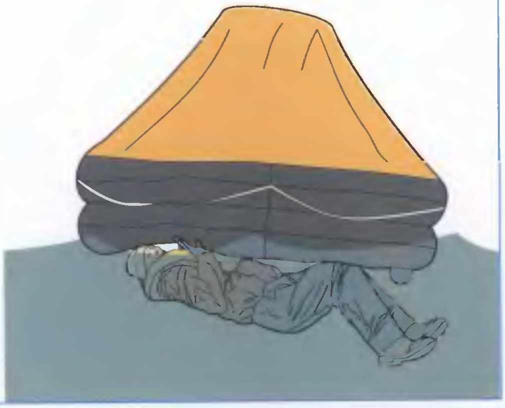

9. Do not turn on to your front when under the raft. The buoyancy in the lifejacket will trap you against the floor.

Note that self-righting rafts are also available.

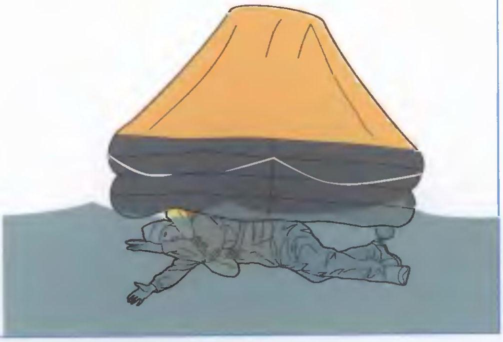

Initial actions

- Cut the painter with the buoyant safety knife (held in a sheath close to the liferaft entrance).
- Cut it as close to the point of attachment to the boat as possible.
- Pull in the line, it may prove to be useful later.

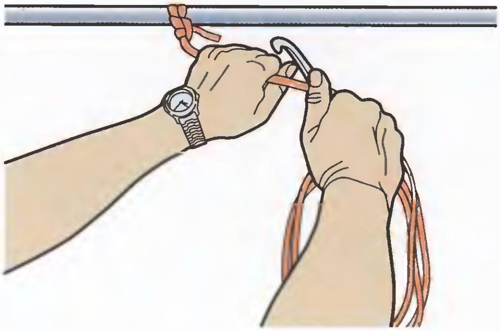

- Paddle clear of the sinking or burning vessel to prevent the raft from becoming damaged.
- Pick up other survivors from the water.
- Congregate with other liferafts and fasten them together.
- Salvage anything that might be useful.

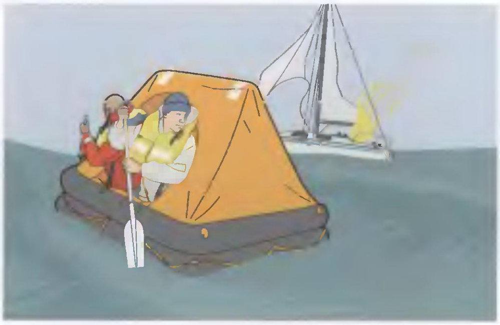

- Stream the sea anchor when you are clear of the vessel. It will turn the raft's entrance away from the wind and waves. It will also improve the raft's stability and slow its rate of drift, keeping you closer to the area where the rescuers will be searching.
- Retrieve the canister lid and any packaging.
- Pick up any useful flotsam.

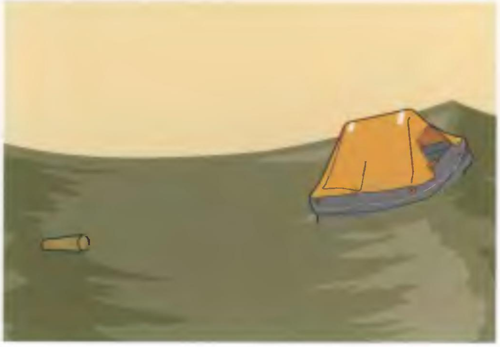

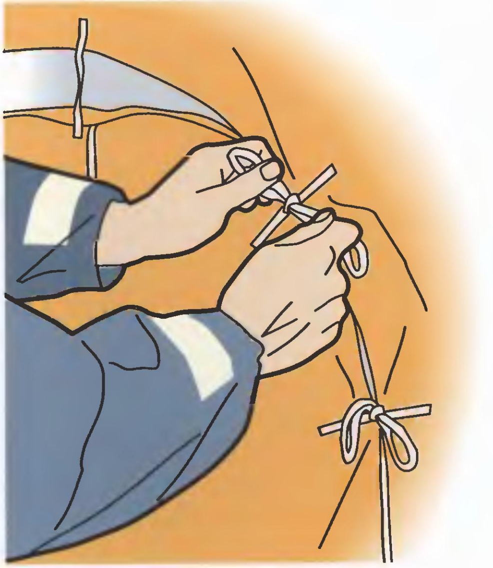

Close the raft's entrance to prevent further ingress of wind and water. It will also keep the warmth in. Entrance closures range from heavy-duty zips to Velcro fastening to webbing ties. Make sure that, whichever system is used, the raft entrance can be opened quickly. Use slipknots on ties.

Open the entrance every 20 minutes to allow fresh air in. Most pressure-relief valves vent outside the raft. If the PRVs vent inside the raft, open the door to let out the fumes. Inhaling CO_2 will give you a headache.

Maintain the raft by checking for leaks and bailing out. Check that the equipment pack is still secured inside the raft, if not check outside the raft; it may still be afloat and retrievable.

Remember...

1. CUT the painter
2. STREAM the sea anchor
3. CLOSE the door
4. MAINTAIN the raft

Secondary actions

What to do next

Be aware that there will be a great temptation to collapse into the liferaft and relax. Now's not the time to rest. Get organised before the cold saps your strength and your ability to use your hands.

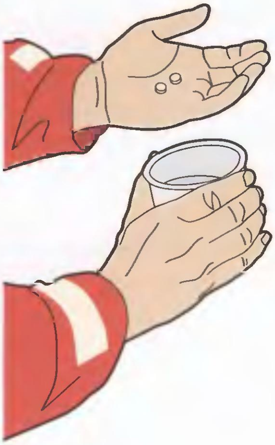

If anti-seasickness medication has not been taken before boarding the raft, open the liferaft pack, find anti-seasickness tablets and take them now. Everyone must take one. Even the hardiest of sailors is likely to be sick inside a raft.

Be prepared – take out one or two seasickness bags as well. Once someone is sick, everyone else is likely to be affected.

Seasickness will cause not only physical and psychological debility but also loss of valuable body heat and fluid. It may also mean the sufferer loses their medication before it takes effect.

The priority now is to get as warm and as dry as possible.

Bail out the raft. The bailer provided may be no more than a fabric scoop. Some rafts have a self-bailer fitted in the floor that will drain the water automatically.

If not, bail out by hand using the bailer and other containers – a boot, a bucket or the salvaged liferaft canister lid. Use the portable bilge pump from your grab bag. Be resourceful.

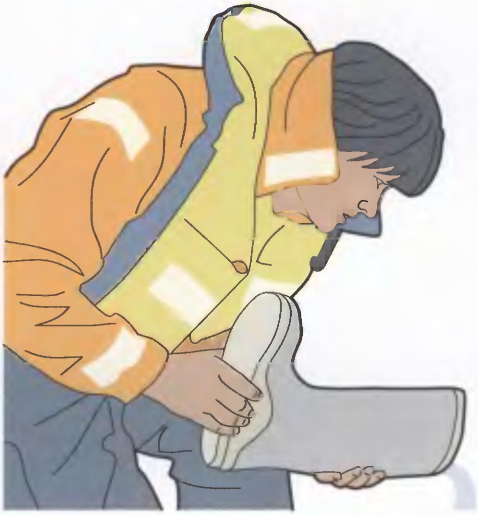

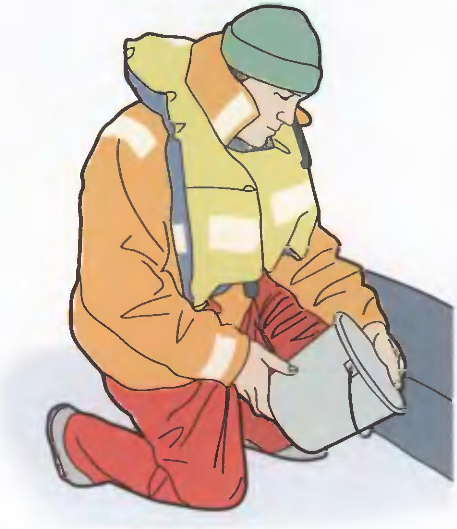

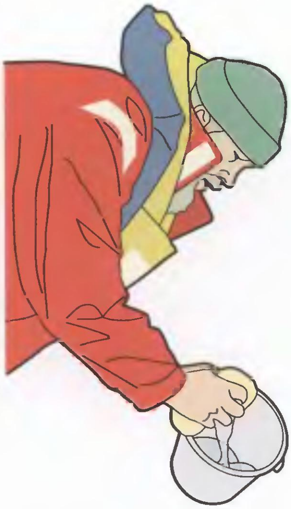

Use one of the sponges from the liferaft pack to dry the floor of the raft. Remember to keep another sponge clean for fresh-water collection.

Apply first-aid to those that require it. Everyone is likely to be suffering from varying degrees of hypothermia and shock in addition to injuries. Look out for the quiet ones – the quietest casualties are likely to be the worst-injured. Calm those that are acting inappropriately.

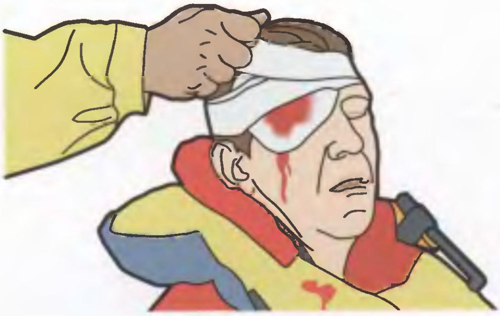

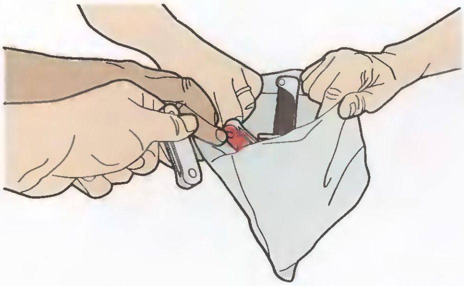

Collect sharp objects and weapons – knives, multi-tools etc – that could damage the raft or be used to harm or threaten anyone when rations become short and tempers fray at a later date.

Wring out wet clothingand put it back on. Huddle together for warmth – don't be shy!

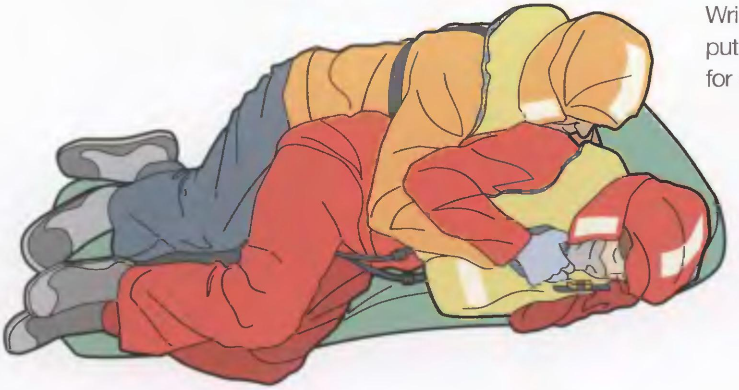

Your main heat loss is by conduction through the raft floor. Some floors have thermally reflective surfaces but others need to be inflated with the bellows. Take care not to pump water into the floor. Insulation can also be provided by sitting on salvaged bunk cushions, fenders or, conditions permitting, inflated lifejackets.

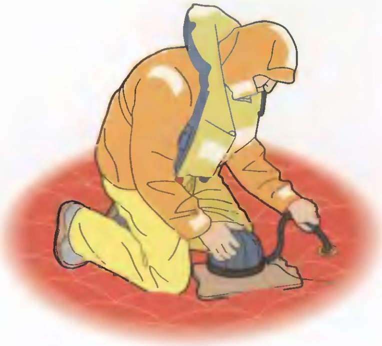

ISAF, ISO 9650 Part 1: Type 1 and SOLAS raft packs include a number of TPAs (thermal protective aids) that will reduce the body's evaporative heat losses. Thick plastic bags, large bin liners or the salvaged liferaft polythene wrapping could also be used.

KEEP YOUR HEAD WARM.

Locate the pressure-relief valve stoppers and put them in place to prevent leakage through the valves. Check for leaks. Top up the tubes with the pump if necessary. Repair kits will only work if you can dry the area to be repaired. Use liferaft repair clamps if the hole is large, and then reinflate the tube.

A leader, who may not be the skipper, will emerge. It may be the fittest person with the most sea survival knowledge. Their job is to make sure everyone believes they are going to be rescued and to organise the crew and resources effectively.

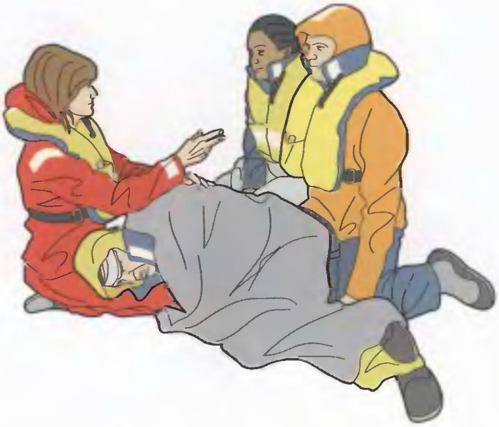

It's important that everyone works as part of a team to increase the chances of survival and keep each other's morale up.

No recriminations – there will be plenty of time for that after you've been rescued.

THE WILL TO SURVIVE IS PARAMOUNT.
KEEP MORALE HIGH.

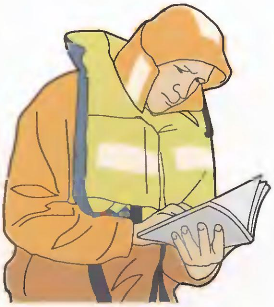

Read the survival notes in the liferaft instruction booklet to refresh your memory. Stow equipment that needs to be kept dry in resealable bags. Some rafts will have stowage pockets on the canopy tube high above the water.

Establish a routine as soon as possible. If there are sufficient crew, put two people on watch for between 20 minutes and 2 hours depending on the weather conditions. Make sure the watch system is fair.

Outside watch should keep a lookout for ships, aircraft, other survivors and dangers. They must be fully briefed on signalling with flares, heliograph and VHF.

Inside watch should keep the raft inflated, treat and look after casualties, collect water, bail out etc.

Switch off the liferaft light during the day to preserve its power. It may seem very dim but to an SAR helicopter's or lifeboat's night-vision goggles it will stand out for miles. Switch off and save lifejacket lights – you may be able to use them when the liferaft light fails.

SOLAS rafts have an inflatable radar reflector. Wetting the canopy may also increase the radar reflection range.

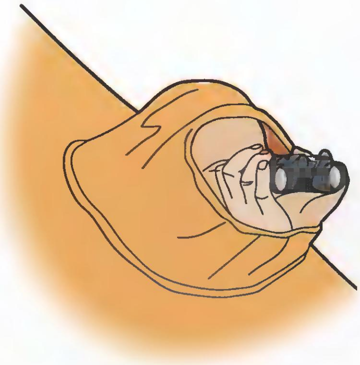

Do not use a radar reflector and SART at the same time.

Look after the handheld VHF, EPIRB, PLB and flares since they are probably your best means of location by rescuers. Save the VHF's batteries – remember its range is limited. If conditions allow, stand up when using the VHF to increase its range.

Avoid urine retention because it could have serious consequences later. Urinate within two hours of boarding. Female survivors may prefer to use the bailer. Take care not to fall in if urinating out of the entrance. Ask other survivors to hold your lifejacket straps.

You cannot reabsorb urine from the bladder. Urinating will not increase your rate of dehydration. Try to get rid of urine overboard. The smell can cause nausea. Avoid urinating in your clothes since cross-infection could cause problems later.

A bowel movement is not uncommon. You will probably not need to do it again if you are not ingesting food or fluid.

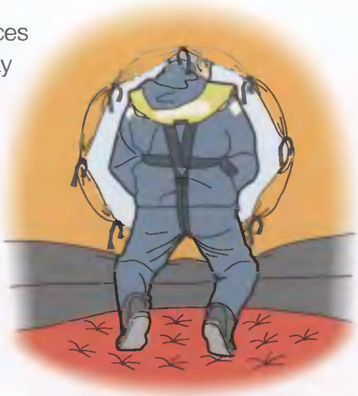

Protection

In northern European and southern Australian waters, the cold is your greatest enemy. The liferaft is designed to help save you from drowning and reduce heat loss by getting you out of the water as well as providing protection from wind chill.

As long as the canopy is providing a good seal from wind and waves, the air inside the raft will be warmed by the body heat of the occupants. Therefore the greatest heat loss is from conduction through the floor of the raft.

Dry the raft floor thoroughly. Pools of water will intensify conductive heat loss. Inflatable floors will need considerable air pressure before providing an air gap beneath seated survivors. Better insulation may be provided by sitting on lifejackets, fenders, bunk cushions or anything else you have salvaged. Only remove your lifejackets if conditions allow and you are confident that the raft is stable.

Put on a hat, raise your hood, cover your head with a scarf or other fabric. Your head loses more heat than any other part of your body.

Once you are insulated from the water, the greatest heat loss will be due to evaporation of water from wet clothes. To stop this, remove and wring dry your clothes. Put the damp clothes back on and don a thermal protective aid (TPA).

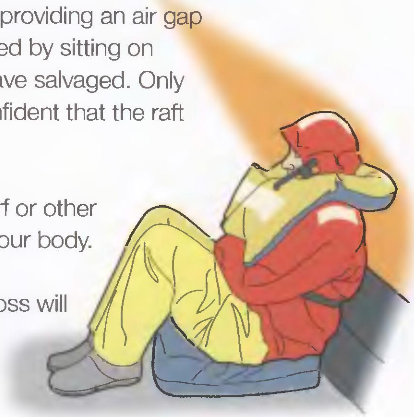

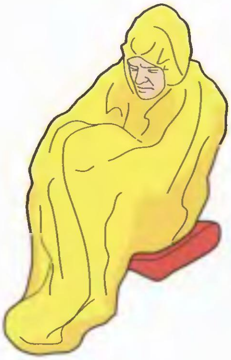

TPAs reduce evaporative heat losses by containing the water vapour in the bag. Once the air inside the bag reaches saturation, no more water can be absorbed and evaporation stops. If TPAs are not available, use bin liners or plastic sheeting.

Reflective blankets are not effective in a cold, wet environment. Radiant heat losses from a cold casualty are minimal and the blanket's reflectivity is reduced to virtually nothing as water vapour condenses on its cool inner surface.

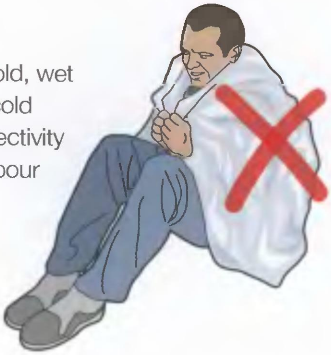

Survival or immersion suits are for extreme conditions and provide the best protection against cold. They tend to be bulky and awkward to put on, particularly in a hurry in a confined space, however many people have survived by using one.

Water and food rationing

For survival, the human body needs water more than food. Tests have shown that a person will live for only seven to ten days without water assuming there is no excessive fluid loss but can survive for 20 to 30 days without food.

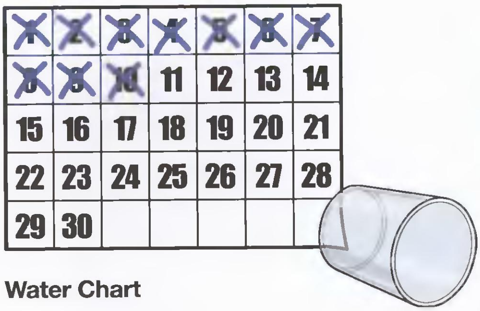

Water Chart

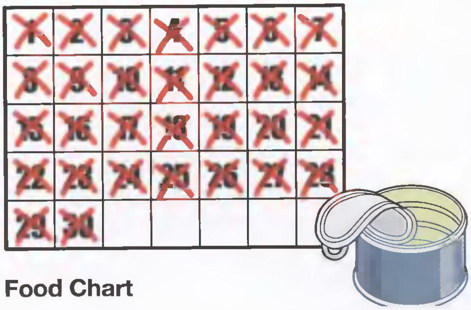

Food Chart

In busy waterways such as those found in northern Europe, it is unlikely a raft will be adrift for more than 48 hours before being rescued. The greatest dangers are drowning and hypothermia rather than dehydration or starvation.

Since the human body has a reserve of water and energy within its tissues, no food or water should be taken in the first 24 hours except by children or conscious injured casualties, who will experience greater fluid loss.

If only minimal water supplies are available, the ration for water, after 24 hours, is half a litre per person per day. (0.5lt/person/day = just under 1 pint/person/day)

Take a third of a ration at sunrise, a third at midday and a third at sunset. When drinking water, keep it in your mouth for as long as possible to moisten the membranes.

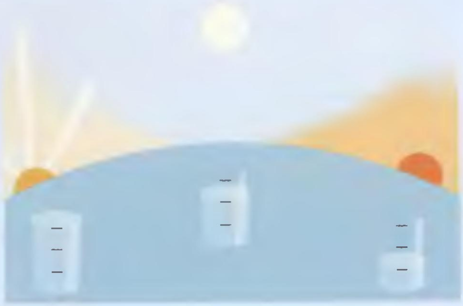

To maintain morale and good harmony, it is vital to ensure that food and water rations are issued fairly.

Water may be supplied in the liferaft pack. It will be in sealed sachets. Use the scissors in the first-aid kit and open the sachet carefully. Decant any leftover water into a clean bottle or sealable container. Stow an empty bottle in your grab bag to collect leftover water.

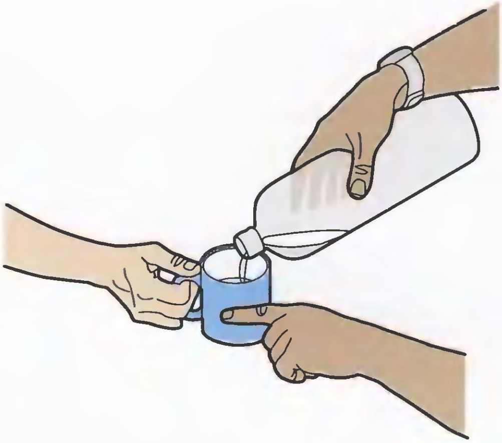

If you have plenty of water, each person may drink up to 1 litre per day.
(1lt/person/day = about 2 pints/person/day)

Whether you have plenty or virtually none, start to collect water from the outset – you don't know when it might rain again.

Drink collected water first – because bottled water and sealed-sachet water will stay fresh for longer than rainwater.

Do not drink urine. Do not drink seawater, even in a diluted state. Death has been found to occur quicker when you drink seawater rather than nothing at all.

Reduce the body's water loss by making the best use of the shade and cooling from the breeze.

Avoid eating protein-based food (such as fish or meat) because digesting them uses more fluid than they provide. Carbohydrates and sugars are better.

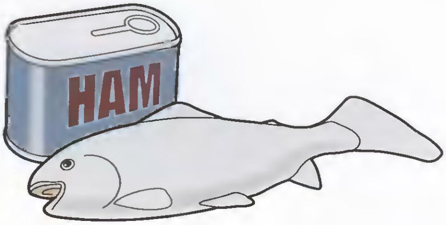

On 30th July 1945, the USS *Indianapolis* was attacked by a Japanese submarine. The ship sank in 12 minutes and many survivors ended up in the water. Rescuers took several days to reach them. Desperately thirsty, some survivors drank seawater.

They became delirious and started to hallucinate. Some thought that there was an island nearby and would try to swim off to it. Some claimed that the ship was not sunk but was just below the surface and convinced others to remove their lifejackets and swim down with them to drink from a water fountain.

Of a crew of 1196, it's estimated 900 went into the water. Four days later, only 316 men remained alive. While some were lost to sharks, many were lost due to the effects of drinking seawater.

Long-term survival techniques

Sustaining the will to survive and the belief that you will be rescued is essential.
Keep the survivors' morale high by running a disciplined routine. Fear, panic, despair and recriminations will undermine your belief that you will pull through. Knowledge of survival techniques and confidence in equipment will help you remain calm and positive.

Lack of water and food will cause fatigue and exhaustion and increase depression. Catching food can be a great morale-booster and will supply much-needed energy. Rest as much as you can but when you are not asleep, keep busy with routine raft duties. Stay optimistic and look forward to the future. Discuss what you plan to do when you get ashore.

Long-term survival techniques revolve around collecting or making water to drink, fishing, keeping the raft inflated and in good order and keeping watch for potential rescuers.

COLLECT RAINWATER AT EVERY OPPORTUNITY.

Better-quality liferafts have a small rain-catcher fitted to the canopy of the raft. It's important to prepare the canopy and the collection bottle before the rain arrives. Wash built-up salt crystals off the canopy and rinse containers with fresh seawater. Use the first downpour to wash off the remaining salt – anything you collect during the initial shower will be contaminated and undrinkable. Collect the rainwater once the salt has been washed off.

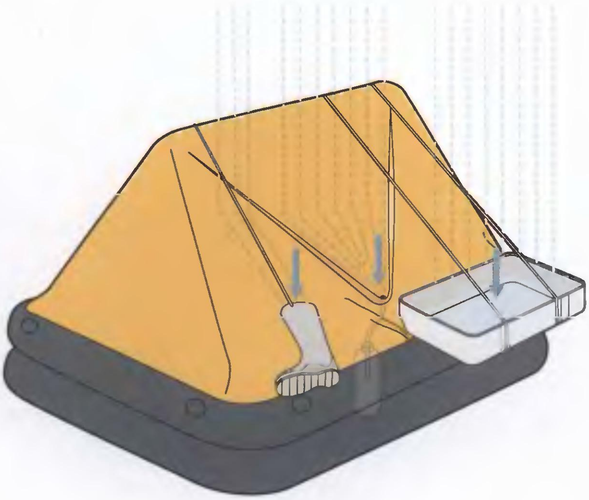

Improvise other means of collecting rainwater – the canopy's collection system is not large.

Unpalatable rainwater can be ingested by enema. Lyn Robertson, a qualified nurse, used the raft's bellows as a method for introducing poor-quality rainwater for absorption through the gut on her family when they were shipwrecked for 37 days in June 1972.

However, tests have shown that saltwater enemas are not a solution to dehydration. The gut does not filter salt out of the water – only use rainwater.

Solar still

An inflatable solar still can be filled and fastened alongside the raft. Salt water held in the bottom of the still evaporates and potable water condenses on the inside of the tent and runs down to a collecting bottle. It can produce about 0.75 to 1 litre per day in calm conditions. However, rough conditions can cause the seawater in the bottom of the still to splash into the distilled water making it undrinkable.

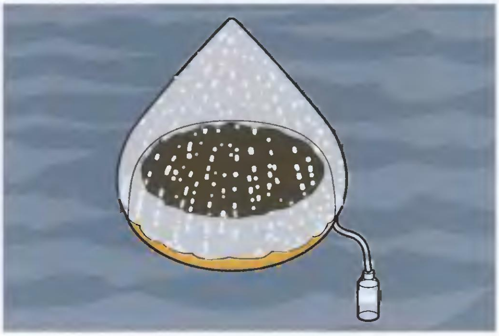

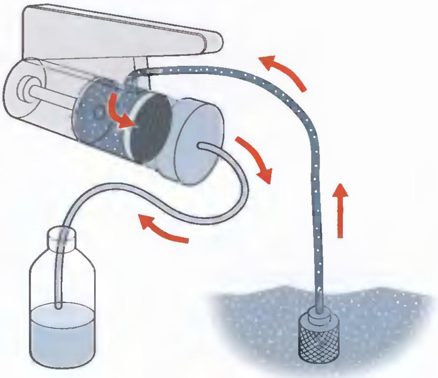

Hand-operated desalinator

Using reverse osmosis, a desalinator or water purifier pump will remove 98% of the salt from seawater. It can produce as much as 4 to 5 litres of drinking water per hour but is hard physical work to operate. It may, in the heat of the day, cause the operator to sweat out more water than they can make. To make the most of a desalinator, use it at night when the weather is cooler.

Lone sailor Steve Callahan's yacht sank after a collision with a whale while on a transatlantic passage from the Canaries. He abandoned ship into his liferaft. In his grab bag were 8 pints of water and two solar stills. Seventy-six days and 1500 miles later he was picked up by fishermen within sight of land.

Bill and Simone Butler survived for 66 days in a liferaft by using a PUR Surviva 34 manual reverse-osmosis water filter after whales sank their 38ft yacht *Siboney* 1200 miles west of Panama.

What's on the menu

Catching food boosts survivors' emotional and physical well-being. Foods high in complex carbohydrates are more suitable than protein food such as meat and fish. Protein will draw water from the body while it is being digested. Ideally, the body needs just over a litre of water a day to digest food and dispose of waste matter. If no water is available, the survivor should not eat, as this will accelerate dehydration. If food is contained in the liferaft pack it will be high-energy carbohydrate, similar to a block of dried porridge.

- Plankton can be caught using an improvised net made from a pair of stockings or tights.
- Barnacles will grow on the bottom of the raft and the meat can be extracted and eaten.
- Fish can be caught using a harpoon, spear or gaff, or by hook and line.

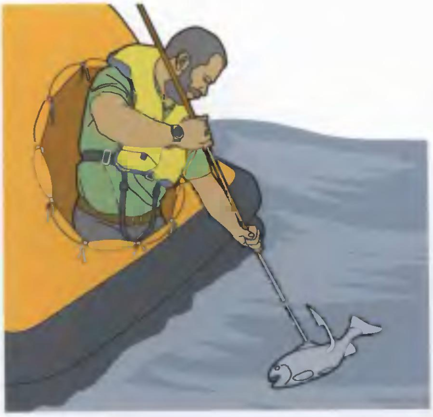

An improvised spear can be formed by lashing a knife to a paddle. Steve Callahan included a spear gun in his grab bag, but if you do the same, take care not to puncture the raft.

Flying fish may also land on or in the raft.

Maurice and Maralyn Bailey adapted the safety pins in their first-aid kit to make fish hooks. On average, they landed forty fish a day.

Alain Bombard created hooks from a bone found just behind the gill of the Dorado fish.

Initially, lures can be made from silver paper. After that, most survivors have used the heads of the fish they caught as their bait. Poon Lim, who survived for 133 days on a lifeboat, used a dried paste of biscuit and saliva for his initial bait.

Fluid can be extracted from large fish by sucking their bones and eyes. The fish guts may be more digestible than the flesh.

***Turtles* –** Dougal and Lyn Robertson and their family and crew caught turtles. Taking care that the creature's beak didn't damage their inflatable, they slit its throat and drank its blood immediately to rehydrate. The meat was dried and the blood used as a 'sauce'.

***Seabirds* –** Poon Lim, Steve Callahan and the Baileys all tell of catching and eating seabirds. Seagulls and boobys will often land on the liferaft or can be attracted by a baited hook.

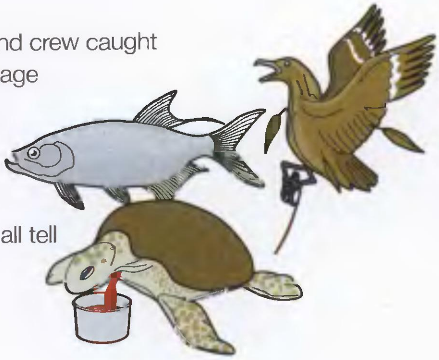

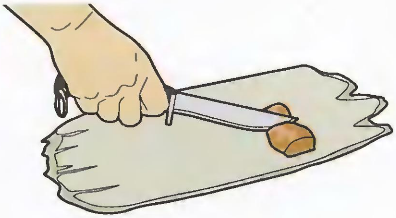

Some form of chopping board is useful and can be improvised from the liferaft canister, flare box or salvaged flotsam.

Keeping cool

While those in cold-water climates will be concerned about keeping warm, sailors in the tropics, are more concerned about staying cool and reducing dehydration caused by excessive sweating.

Methods of keeping cool

Obviously, stay under the canopy or create shade so that you can stay out of the sun. Rest during the heat of the day and undertake routine tasks in the morning and evening.

Cool the raft by wetting the canopy, and keep your clothes damp. Regularly rinse them in the sea to prevent a build-up of salt crystals, which would exacerbate sores.

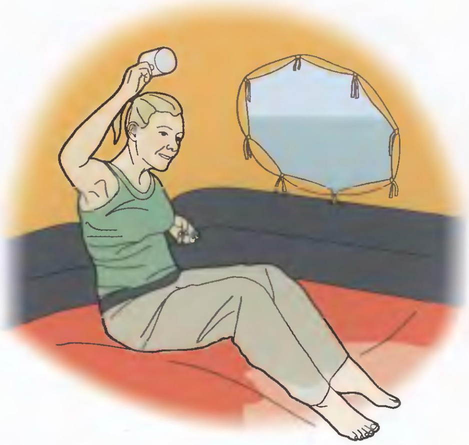

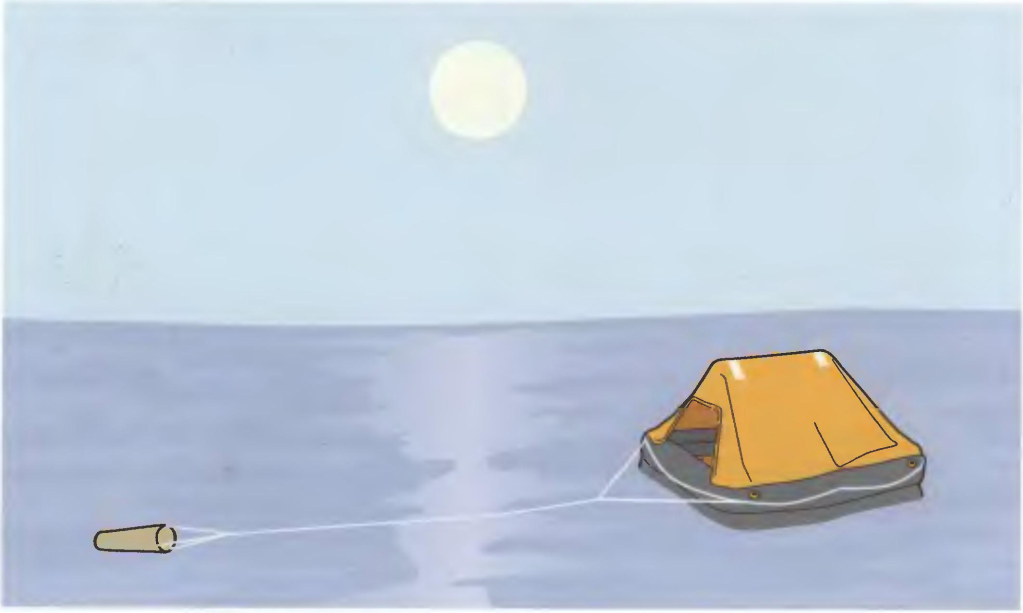

In hot, calm weather, reattach the sea anchor to the entrance-side of the raft so that the wind blows into the raft.

During the day, deflate the inflatable floor and sit directly on the raft floor. Reinflate it at night to prevent getting cold.

If you are in a group of two or more rafts, spread the survivors equally between them so that each individual has more space to stretch.

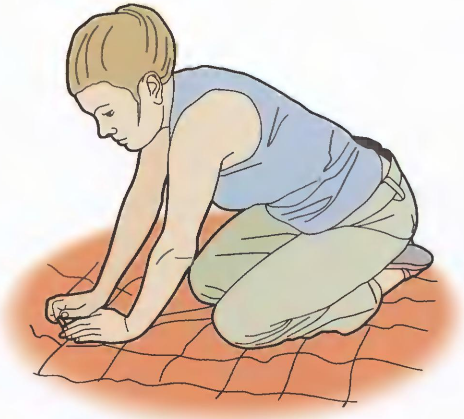

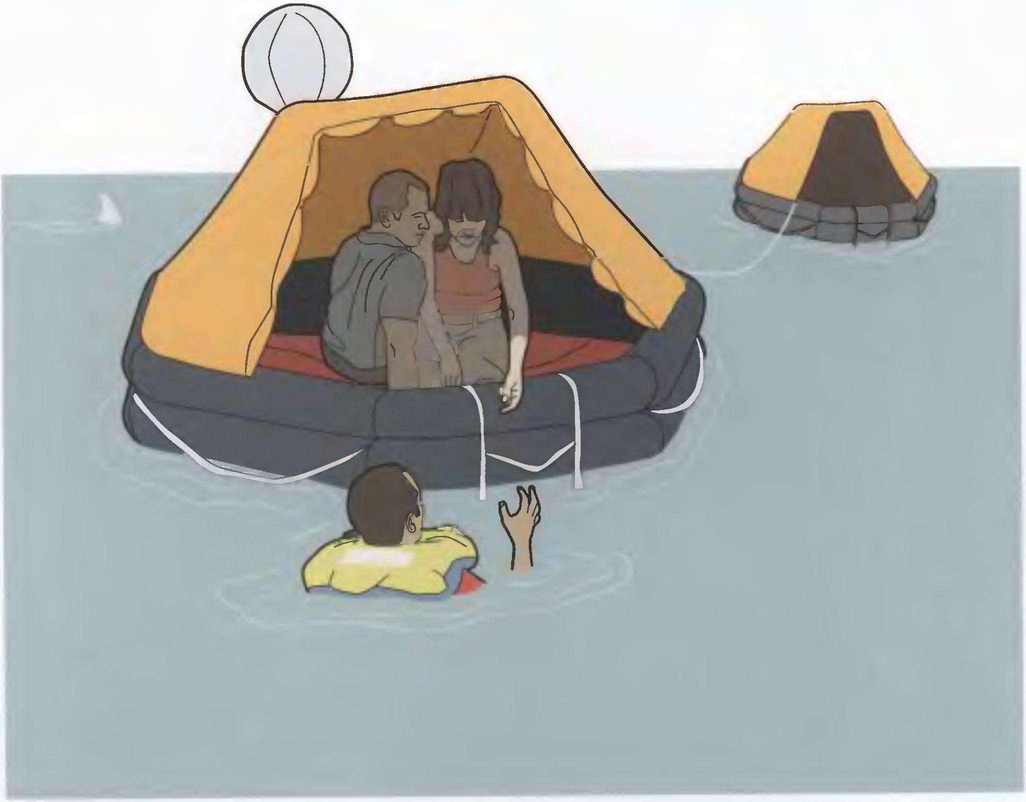

Don't get into the water to cool down; you may not be able to get back into the raft.

CHAPTER 13 : Survival in cold water

Research by Golden and Hervey shows that the body goes through four stages when immersed in cold water:

Length of Time Immersed	Risks
0–3 minutes	Diving reflex and cold shock
3–30 minutes	Swim failure
More than 30 minutes	Hypothermia
Post-immersion	Hydrostatic squeeze, secondary drowning, post-rescue collapse (during and after rescue)

Because the human body loses heat 26 times faster in water than in air, immersion in cold water can be extremely dangerous. The graph below shows the possible effects of immersion in 10 degrees C water. Around the UK, sea temperature varies between 5 degrees C and 18 degrees C. The temperature of the Mediterranean during the summer rarely gets above 26 degrees C.

Typical change in deep body temperature during immersion in stirred water at 10˚C *(50˚F)* for an average clothed individual. *(F. Golden)*

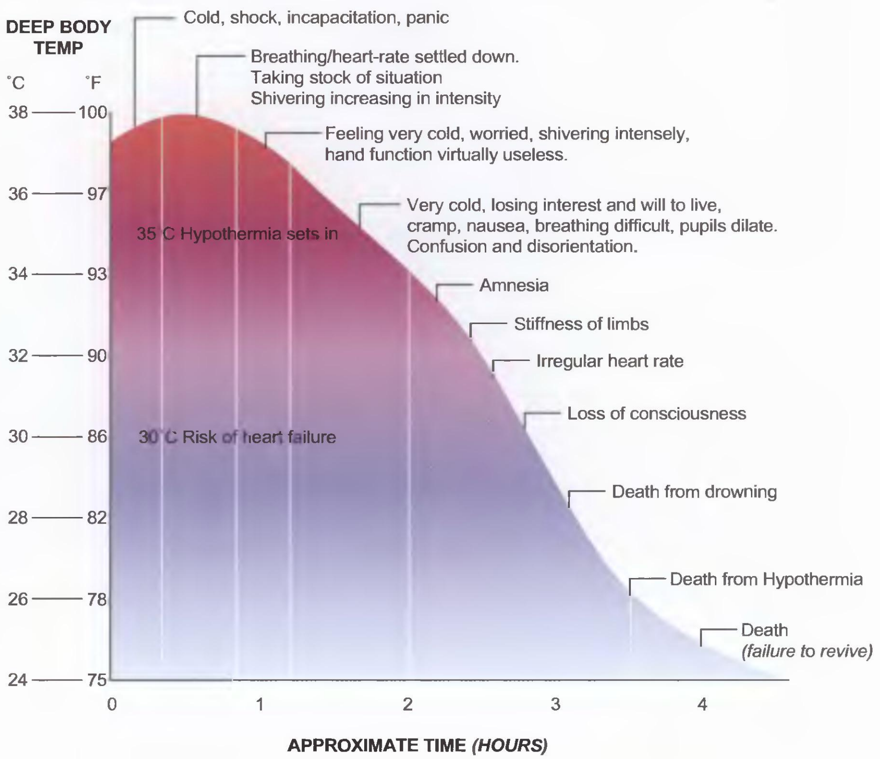

Diving reflex

The diving reflex occurs when you first make contact with the water. If the water is not too cold you involuntarily hold your breath and your blood circulation and heartbeat slows. However, in water 15 degrees C and below, cold water shock may overcome the diving reflex.

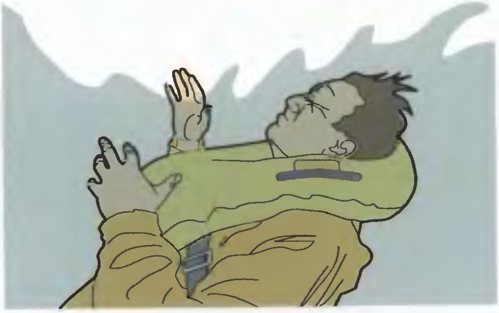

Cold shock

Cold shock takes place after thirty seconds to three minutes of immersion. Your breathing rate increases from around 10 breaths per minute to 60 breaths per minute (hyperventilation, which can cause dizziness and confusion). Breath-holding times are likely to reduce from over a minute to less than 10 seconds.

In water of less than 15 degrees C, it is probable that you will involuntarily inhale (gasp) even when underwater.

Blood vessels close to the skin shut down causing a sharp increase in blood pressure, which may be sufficient, depending on your fitness levels, to initiate a stroke or a heart attack.

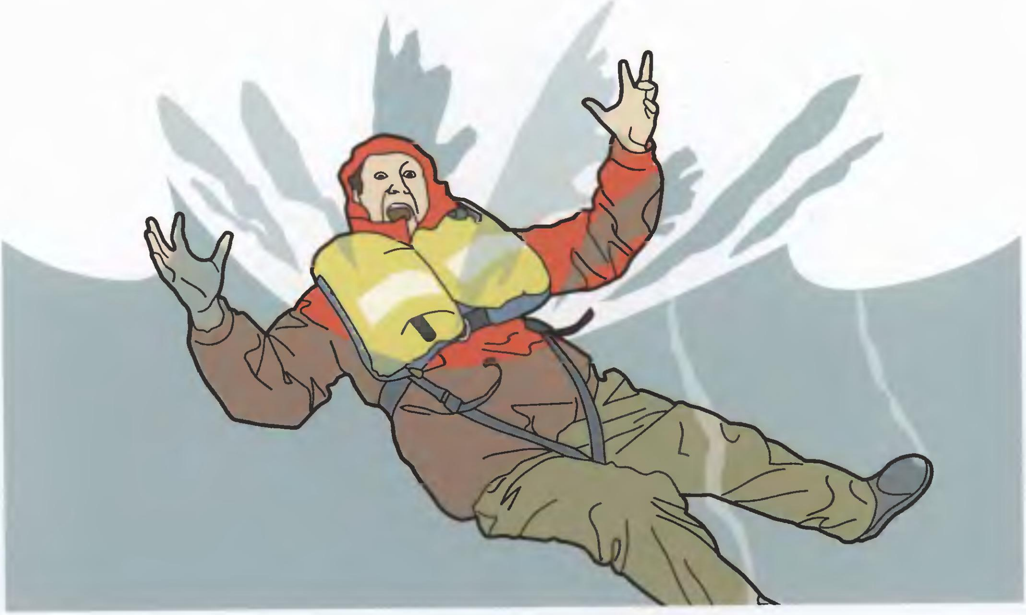

Protecting against cold shock
To reduce the likelihood of cold shock, wear a lifejacket, waterproof clothing and, if possible, enter the water slowly. Do not try to swim – adopt the HELP position (see p.66).

Regular cold showers, baths or sea swimming can help to prepare your body for sudden immersion in cold water. Tests have shown that after just one week of daily cold showers, the effects of cold-water shock can be reduced by 50%.

The fitter you are, the better able you will be to withstand cold-water shock.

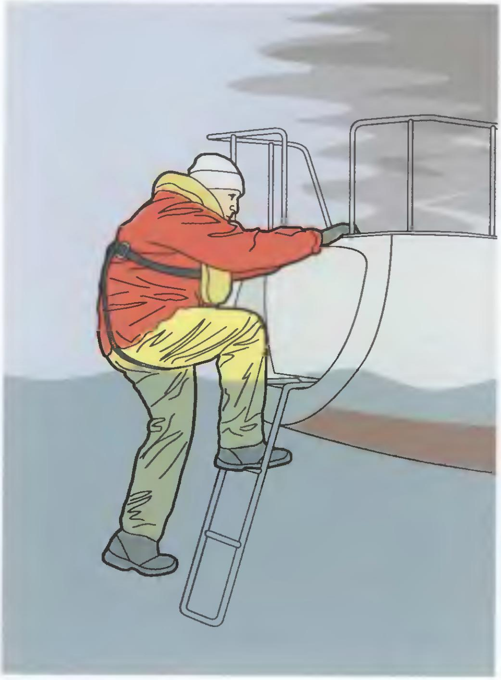

You are more likely to drown as a result of entering cold water quickly than you are from hypothermia.

Swim failure

After three to thirty minutes of being immersed in cold water, your arms and legs become numb and your muscles will not function properly. Shivering increases and your heart and breathing rate decreases.

Grip strength and co-ordination are significantly reduced. You may find it difficult to use your hands for firing flares, donning a spray hood, setting off a PLB or using a handheld VHF.

Even fit adult men can find it impossible to lift themselves up the side of a small boat and normally strong swimmers may only manage 100 metres before collapsing. This emphasises the importance of wearing a lifejacket and being realistic about your swimming ability.

Hypothermia

Hypothermia can occur in water of up to 24 degrees C and sets in as the core body temperature drops from 37 to 35 degrees C. In northern European waters in the summer, your survival time with clothing and lifejacket is between two and twelve hours. In winter, it ranges from several minutes to two hours. The exact length of time depends upon on the water temperature, sea state, clothing, gender, fitness, health, shivering rate and age of the person.

What happens?

The casualty will feel cold and their body becomes numb. Intense shivering may stop as the muscles become cramped and rigid. Speech will become slurred, nausea will set in and behaviour will become irrational (similar symptoms to being drunk). The casualty will slip into unconsciousness, increasing their chances of drowning. Heart and breathing rates decrease and the pupils will dilate. Heart failure is possible when body temperature drops below 30 degrees C.

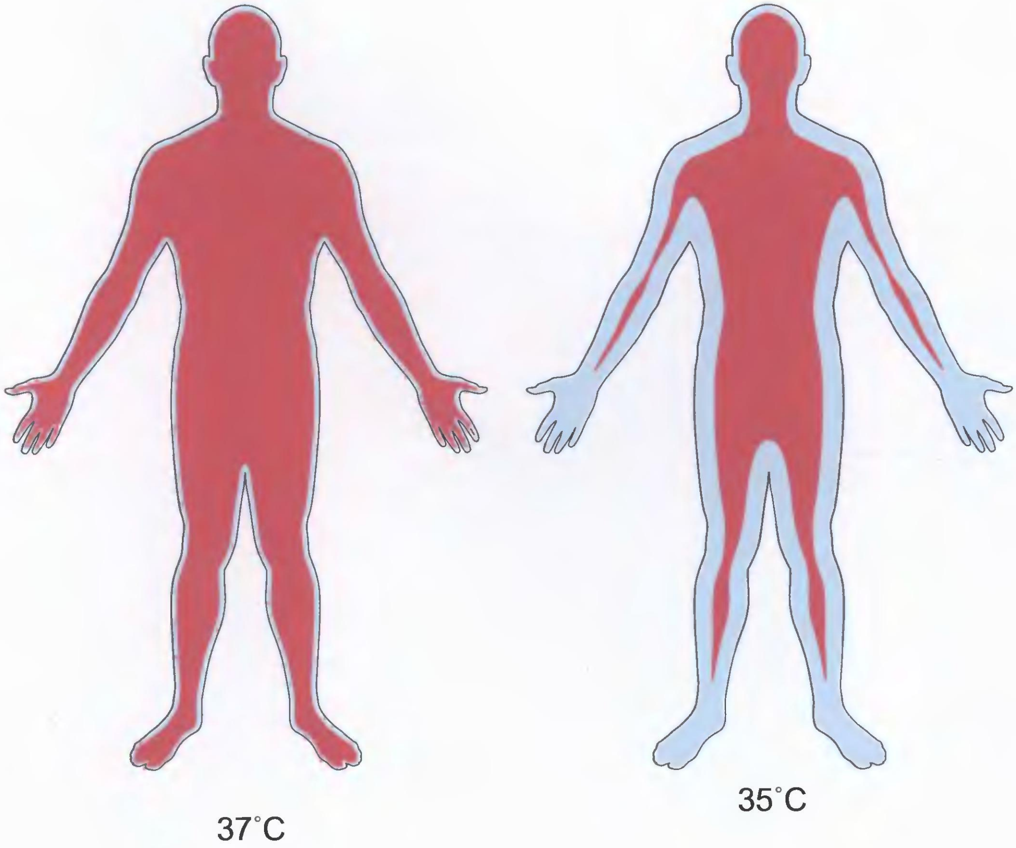

Methods to help prevent the onset of hypothermia

Before entering the water put on an immersion/survival suit, drysuit, wetsuit or more clothing and a lifejacket.

Wear a hat and cover your head with a hood (this also makes you more visible to rescuers). Once in the water, do not kick off boots and try to remain as still as possible, adopting the HELP position (see fig. 1).

Get out of the water if at all possible.

Avoid alcohol.

Heat Escape Lessening Posture (HELP position)

A static body position reduces heat loss and increases your survival time. Therefore, assuming the HELP position is strongly recommended. However, holding this position may not be possible in anything other than calm conditions.

Lie with ankles crossed, knees together and legs slightly bent and elbows held close to the body with one hand under the front of the lifejacket. Hold the other hand up with the palm towards your face, covering the mouth and nose to protect the airway from waves and spray.

In wind and waves, a lifejacket wearer will be turned to face the waves, causing water and spray to break continually into their face. The natural reaction is to stabilise their position using their arms and legs and to turn their body so that the back of the head faces the waves.

A spray hood helps protect the face when both conscious and unconscious and is highly recommended. Put it on as soon as possible.

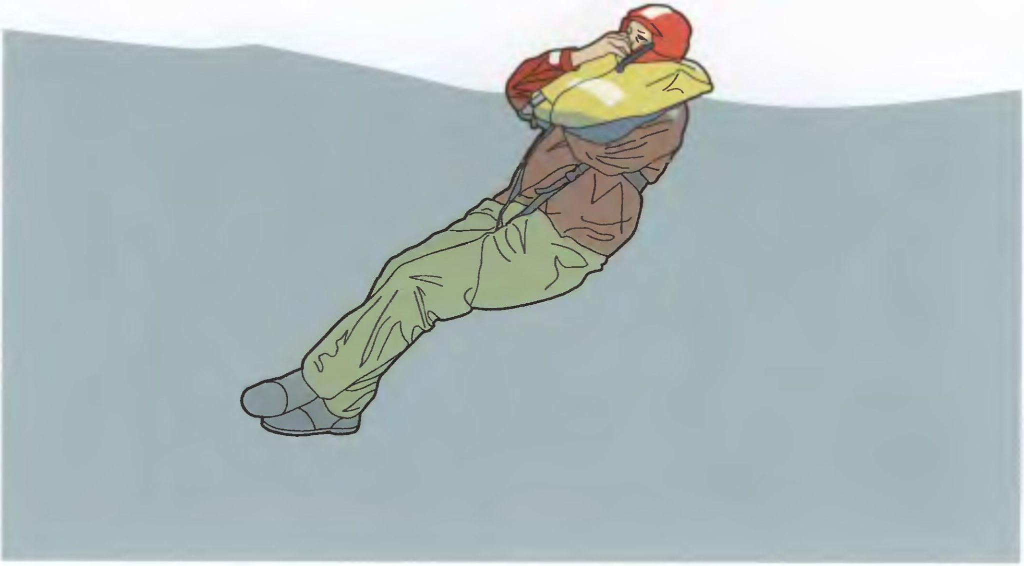

fig. 1

Swimming with a lifejacket

Avoid swimming if possible, since it increases heat loss, however if it is essential, swim on your back. Trying to swim on your front when wearing a lifejacket wastes energy. Swimming by kicking your legs loses less body heat than swimming with your arms but may not be as effective.

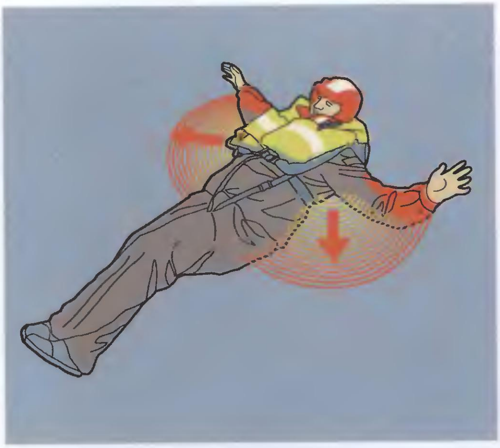

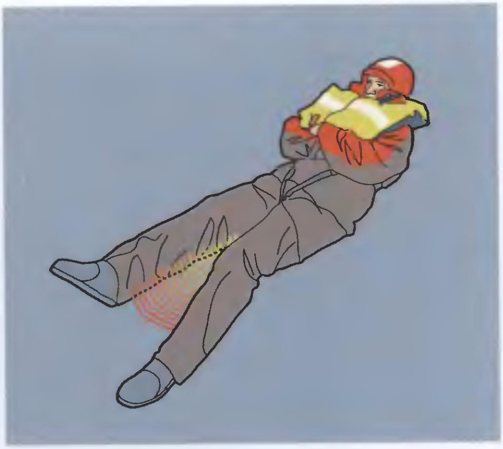

To reach other casualties, a liferaft or a survival craft in the water:

- Send a strong swimmer ahead, connected to the group by a line, or swim as a group.
- The 'crocodile' may be used to swim as a group but it can be very difficult to achieve in waves.
- If a casualty is injured or unconscious, they should be placed at the tail of the crocodile.

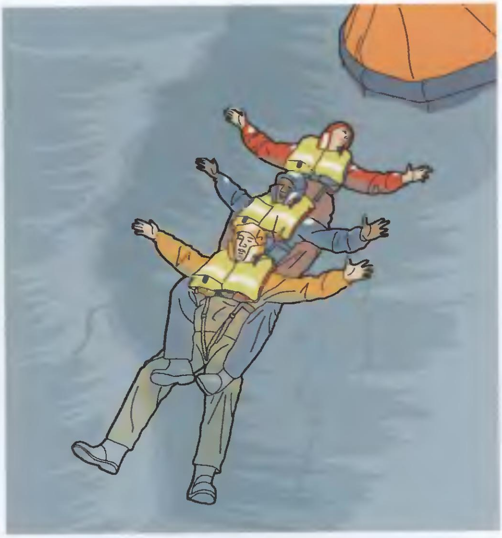

Huddle

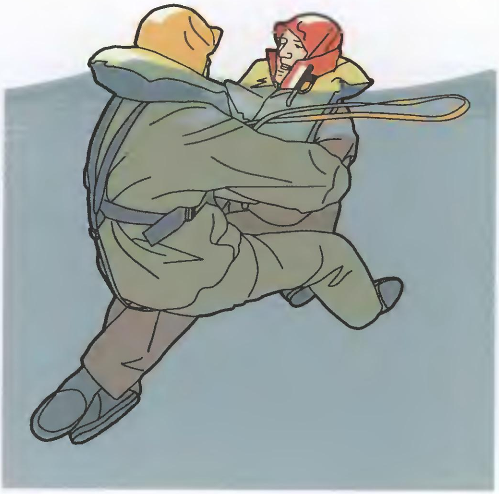

The huddle shares body heat, improves morale, supports weaker members of the group and increases your chances of being seen by rescuers. Use safety harnesses or rope to clip yourselves together as soon as possible, while your hands are still functional.

Be aware that in waves some members of the huddle will be facing wind and waves, making it impossible for them to hold their position.

Two person huddle – one person places their legs around the outside of the other. Take turns to be the person on the outside.

Group huddle – organise the group so that all use their right hand to hold on and their left hand to cover their airway or vice versa. Place injured or weak casualties in the middle of the huddle and rotate the huddle so that no one has to face wind and waves continuosly. Keep your ankles crossed and knees together to preserve body heat.

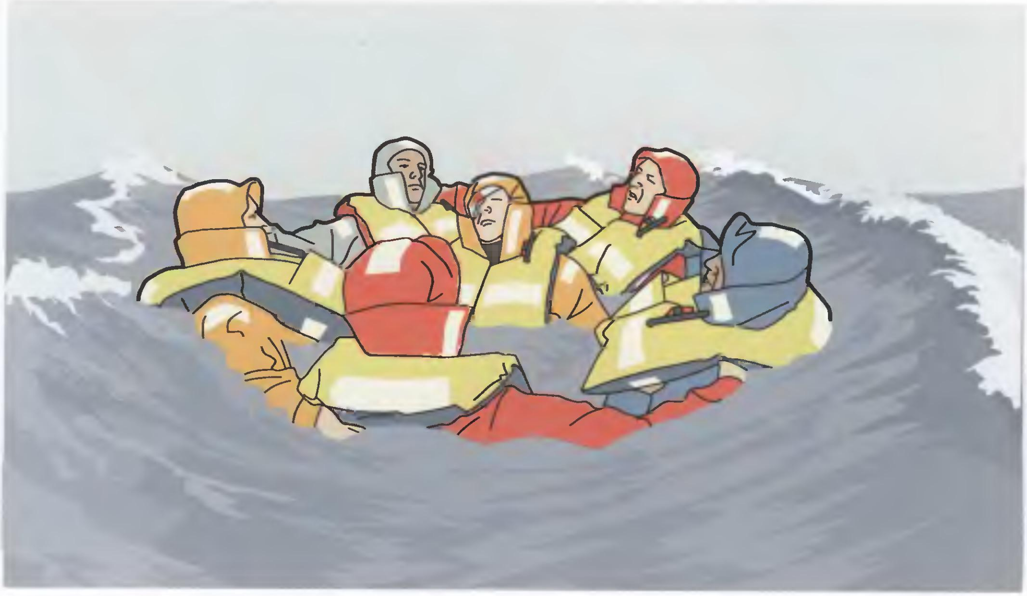

CHAPTER 14 : First Aid

The quality of the first-aid kit supplied will vary depending on the quality of the liferaft. The SOLAS C first-aid kit provides a broad range of medical items and will be found on many rafts, or can be specified for inclusion.

The contents are shown below.

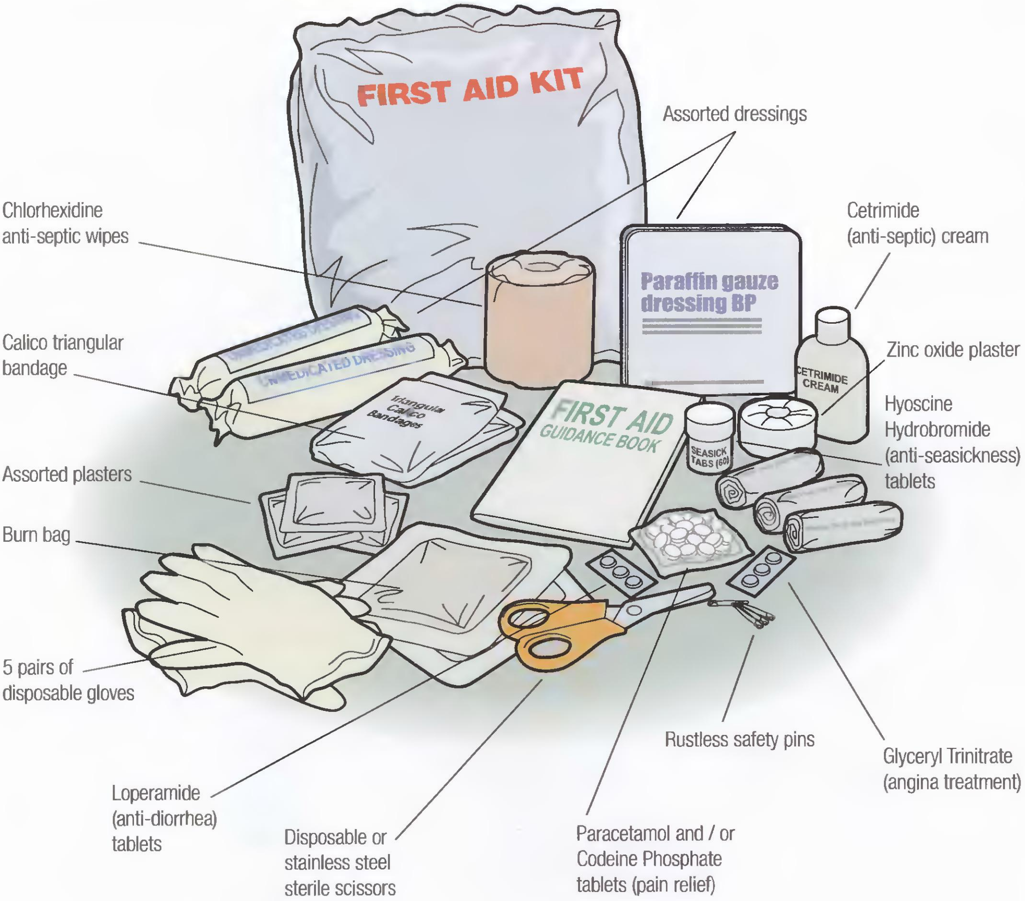
FIRST AID KIT
Assorted dressings
Chlorhexidine anti-septic wipes
Cetrimide (anti-septic) cream
Paraffin gauze dressing BP
UNMEDICATED DRESSING
UNMEDICATED DRESSING
Calico triangular bandage
Triangular Calico Bandages
CETRIMIDE CREAM
Zinc oxide plaster
FIRST AID GUIDANCE BOOK
SEASICK TABS (60)
Hyoscine Hydrobromide (anti-seasickness) tablets
Assorted plasters
Burn bag
5 pairs of disposable gloves
Rustless safety pins
Glyceryl Trinitrate (angina treatment)
Loperamide (anti-diorrhea) tablets
Disposable or stainless steel sterile scissors
Paracetamol and / or Codeine Phosphate tablets (pain relief)

Not shown: Laerdal-type pocket mask (for expelled air resusitation)

In addition to conventional first-aid equipment, consider what else may be improvised. A lifejacket, for example, can be made into an inflatable splint to immobilise the lower leg, or alternatively, use two paddles or a plank of salvaged wood.

For general advice on diagnosis and treatment of casualties refer to the St John's First Aid Book (available from the RYA webshop). The RYA also has a one-day first-aid course for yachtsmen.

Heat exhaustion

This is a condition caused by the loss of water and salt from the body, usually as a result of physical exercise in a hot, moist environment.

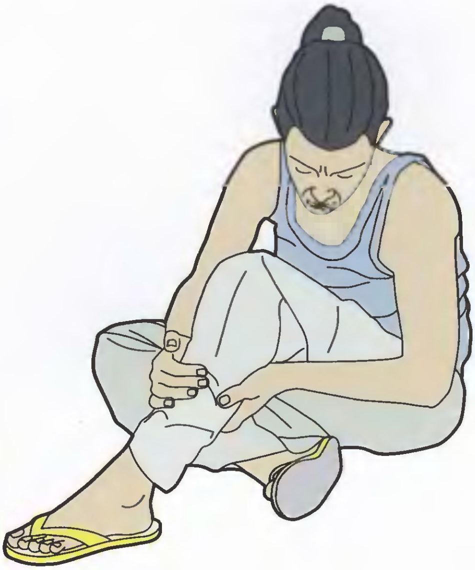

Symptoms
Fatigue, headache, dizziness, nausea, possibly vomiting and muscular cramps.

The skin will be cold, pale and clammy, the pulse weak and rapid, breathing fast and shallow.

Treatment
Remove the sufferer from direct heat and ventilate the liferaft. Sponge down their body, and give sips of cold water with a little salt.

Heat stroke

Caused by prolonged exposure to extreme heat or high humidity. The body temperature rises due to its inability to lose heat through sweating.

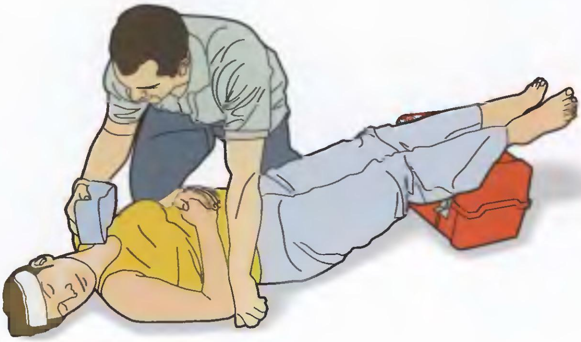

Symptoms
Similar to heat exhaustion except that the skin will look flushed and dry, the pulse is rapid and strong and breathing full and rapid.

Treatment
Sponge down the body, and give frequent sips of cool, slightly salted drinks. Help cool the sufferer by fanning or exposing them to a cooling breeze.

Hypothermia

You can become hypothermic in water of up to 24 degrees C. Around the UK, the water temperature varies between 5 degrees C and 18 degrees C. The temperature of the Mediterranean in the summer rarely rises above 26 degrees C.

Normal core body temperature is 37 degrees C. Hypothermia is defined to have occurred when the core temperature drops to 35 degrees C or below.

Hypothermia may occur in or out of the water. Liferaft occupants are susceptible due to the thin material insulating them from the cold water. Casualties that become cold relatively slowly – chronic hypothermia – such as those in a liferaft or those in water of temperatures between 20 and 28 degrees C will require different treatment from those who are suffering acute hypothermia from becoming cold quickly, such as falling into water below 15 degrees C.

Chronic hypothermic victims are more likely to be exhausted and unable to regain normal circulation on rewarming. Acute hypothermic casualties tend to respond well to rewarming and are soon back to normal.

Symptoms

The graph below shows the typical symptoms of a casualty who has fallen into 10-degree water as the body temperature falls.

Typical change in deep body temperature during immersion in stirred water at 10˚C *(50˚F)* for an average clothed individual. *(F. Golden)*

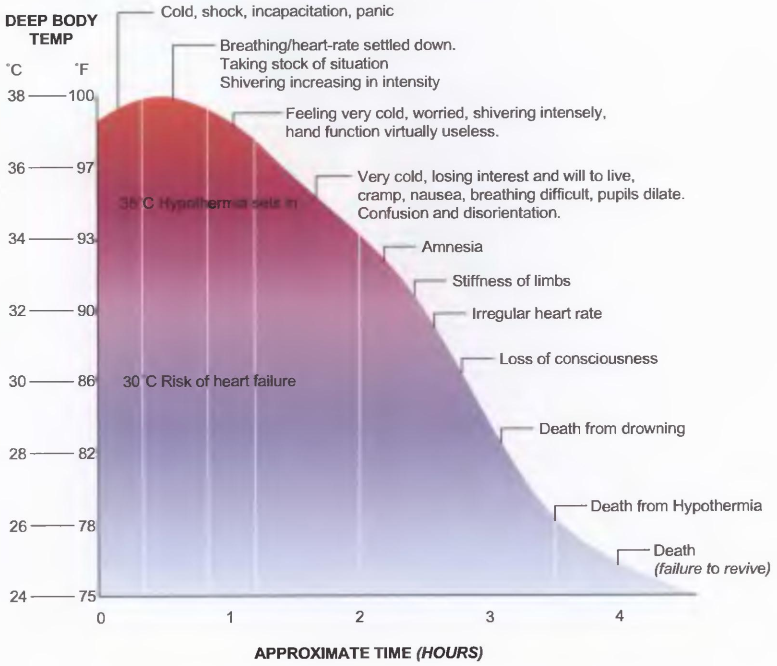

Without a thermometer it can be difficult to tell whether a person is cold or hypothermic. According to Golden and Tipton, one test is to feel the casualty's deep armpit; if it feels like cold marble, they are probably hypothermic.

Initially, the core temperature will rise as the blood vessels in the arms and legs contract, forcing blood back to the core. After the effects of cold-water shock have worn off, heart and breathing rates will slow and shivering will start and become more intense as the core temperature falls. Extremities become numb. Those in the water will suffer from swim failure as they become unable to control their limbs.

Speech will become slurred and the casualty may appear to be quiet and introverted. Nausea may be present and behaviour may become irrational. They will lose co-ordination and become mentally slow. Chronic hypothermics will stop shivering as the body runs out of energy, and may suffer cramps. As cooling continues, the victim will become unconscious and, if in water without a lifejacket, will drown. In a raft, they will continue to cool, the heart and breathing rate will slow and pupils dilate. When their body temperature drops below 30 degrees C, they are in danger of heart failure.

Treatment

Hypothermia is difficult to treat in a survival situation. Passive rewarming is the suggested treatment. The aim is to stop further heat loss and give the body a chance to warm up from the inside.

If the casualty is conscious and co-operative, remove wet clothing. Insulate them with blankets and sleeping bags and wrap in plastic or a TPA. Leave the face exposed. Ensure the casualty is well insulated from the ground, deck or the floor of the raft. The body will slowly rewarm itself as the internal organs start to get back to normal.

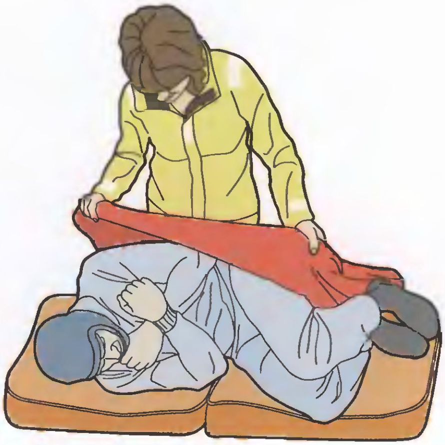

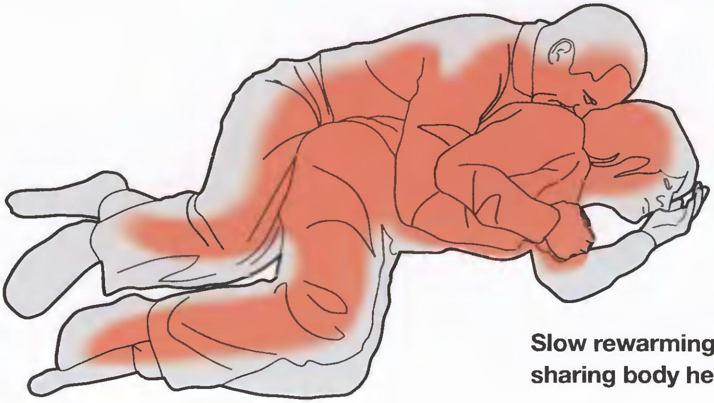

Slow rewarming may also be possible by sharing body heat in a 'buddy system'.

Warm drinks may be given to a conscious casualty. The hypothermic person can have lost as much as 2000 calories of heat.

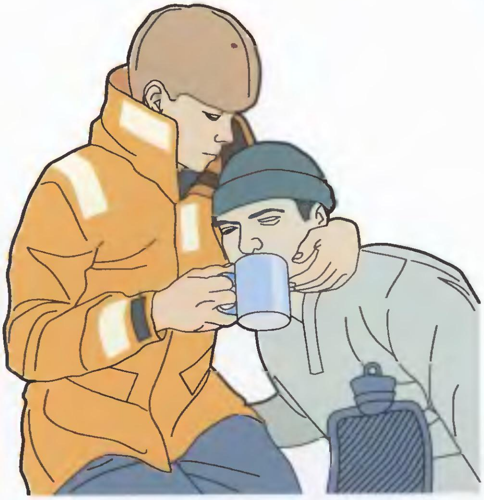

DO NOT:

- Give alcohol
- Use hot-water bottles
- Massage/rub areas
- Handle roughly
- Administer fluid to a semi-/ unconscious person

Unconscious hypothermic casualties must be handled very carefully – otherwise they may collapse and die.

Active rewarming should only be undertaken by medical professionals. Even with the right knowledge, it is unlikely to be possible in a liferaft or small yacht.

Note: it can be difficult to tell if a hypothermic casualty's heart has stopped. Attempting CPR may fibrillate the heart and cause death.

The risk of drowning is reduced if you are wearing a good lifejacket with a spray hood.

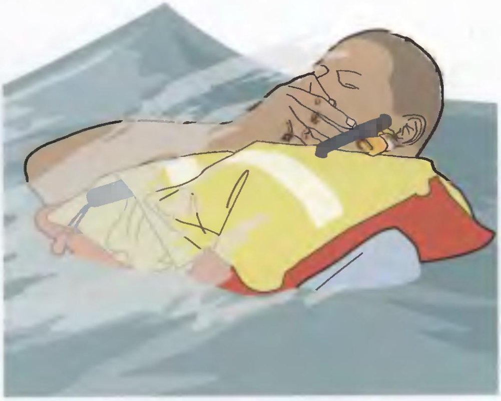

Near-drowning

Symptoms

Breathing may have stopped or be very shallow; lips blue, bluish pallor of the face. Froth around the mouth, vomiting. If conscious, the patient may complain of chest pains and may be coughing.

Treatment

Use Expelled Air Resuscitation as soon as possible (see p134). Water will be automatically coughed up from the lungs when breathing restarts. If available, apply oxygen. Seek further medical help.

Most diving boats carry oxygen and may be closer than medical services.

Secondary drowning

This occurs when a small amount of dirty water or seawater is inhaled into the lungs. A cupful is enough to do it. The water inflames the membranes of the lungs, which react by releasing fluid, filling the lungs thereby causing death by secondary drowning.

Symptoms
Increased chest pain and coughing (possibly with pink froth). Some people may show signs of asthma-type symptoms but many do not. It is likely to manifest itself up to 12 hours after the incident. Seek medical attention as soon as possible (and certainly before 12 hours) if you suspect a casulaty has ingested water into the lungs.

Heart attack

Heart attack is always possible for those in a stressful situation. It is also a possible effect of cold-water shock and hypothermia. It is difficult to tell if a hypothermic casualty's heart is beating, if necessary, use CPR or an automatic heart defibrillator.

Symptoms
Vice-like central chest pain and pain in left or both arms that does not ease when the casualty rests; breathlessness, gasping for air, collapse, sudden dizziness, grey skin, sweating.

Treatment
Make the casualty as comfortable as possible, apply oxygen if available, give angina medicine (glyceryl trinitrate) and/or aspirin (300mg). Monitor and record vital signs. Seek medical advice.

Panic attack

Caused when the casualty has a sudden bout of extreme anxiety. This may be caused by the situation the casualty is in or brought on by symptoms caused by other illness. Either way, the effects on the casualty are very real and frightening.

Symptoms
Rapid breathing (hyperventilation) and a feeling of a racing heart beat. Muscular tension leading to headache, backache and a feeling of tightness across the chest. The casualty may believe they are dying and experience trembling, sweating and a dry mouth.

Treatment
If possible remove the cause of the fear. Reassure them and explain that they are having a panic attack. Get them to slow their breathing. If they are hyperventilating ask them to breathe into a paper bag (use an unused sick bag). Continue to reassure the casualty until they are recovered.

Shock

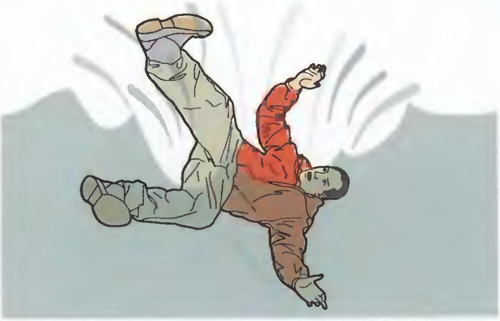

Caused by injury, burns, severe vomiting/ diarrhoea, fear, pain or traumatic experience.

- The effect of shock must not be underestimated, it can be fatal. It is caused by reduced blood flow to the vital organs. Movement can cause fainting and even heart failure.
- Watch out for delayed shock. MOB casualties may say they are OK but may go into shock 20 to 60 minutes after recovery.

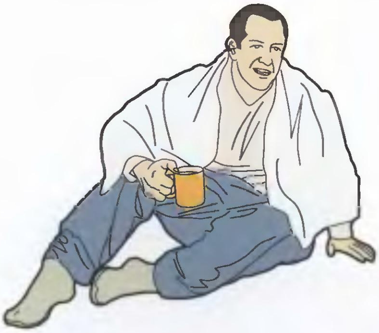

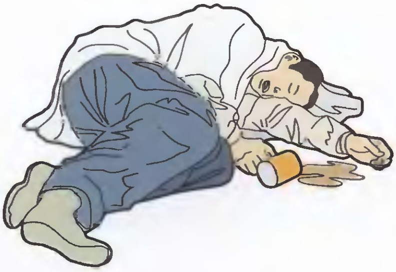

- Liferaft survivors are also susceptible – always check on quiet, introverted survivors, who may be suffering from shock.

Symptoms

Grey or pale skin that is cold but moist with sweat, feeling weak and faint, nausea, weak but rapid pulse, shallow breathing and perhaps gasping; thirst, possible unconsciousness.

Treatment

Most of the crew will have varying degrees of shock so there is little you can do in a liferaft.

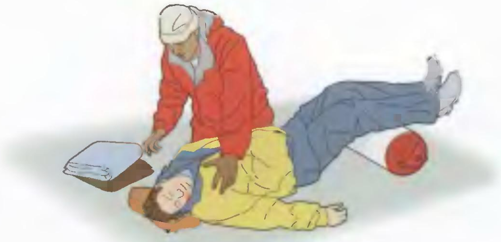

Treat obvious wounds such as external bleeding and broken limbs. Lie the casualty down with their feet higher than their head. Loosen their clothing and if possible put them into the recovery position. Moisten lips with water. Do not apply hot water bottles.

If the casualty stops breathing commence ABC of resuscitation.

Severe bleeding

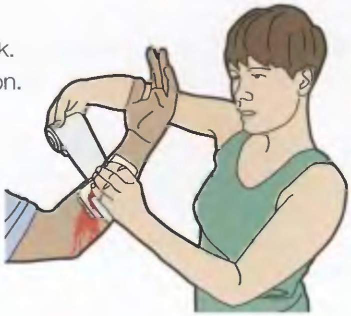

External bleeding is obvious and distressing, and can lead to shock. The aims are to stem the flow and minimise the chances of infection.

Treatment
Cover the wound with a dressing and apply direct pressure. Raise the wound to reduce blood loss, and bandage the area. Keep applying dressings until the bleeding stops. Support the injury. If an object is embedded in the wound, do not remove it. Provide padding either side of object. Apply dressings.

Frostbite

On watch, crew may be in danger of frostbite, where local tissue becomes frozen. The areas affected are usually the fingers, toes, ears or nose. Prevention is better than cure. Wear gloves, cover the face and ears and to be aware of the length of watch duties. Exercise and massage can delay the onset of cold injury.

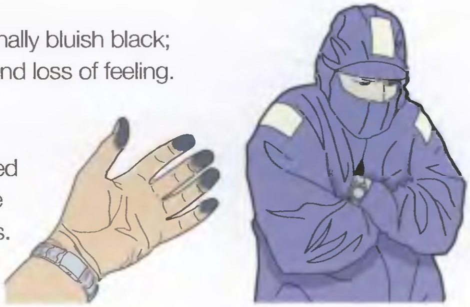

Symptoms
The area becomes pale then white, mottled blue and finally bluish black; there is swelling and fluid release from damaged cells and loss of feeling.

Treatment
Remove constricting garments, raise and dry the affected area. Warm the hands or feet by putting them under the armpits, in the groin or between another person's hands.

Do not massage or burst blisters or apply pressure.

Immersion foot

Known during World War One as Trench Foot, immersion foot is caused by continuous contact with a cold and wet environment.

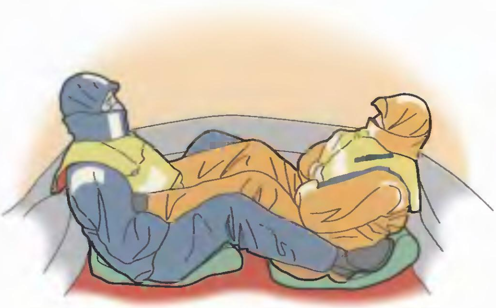

Symptoms
Feet become white as local circulation is reduced, causing swelling of tissues and numbness.

Treatment
Dry feet, use rapid rewarming, encourage gentle exercise and, if possible, give warm drinks.

As the foot warms, it will become painful.

Fractures and dislocations

Broken bones can be either open or closed fractures. If it is an open fracture, treat as for bleeding (p.133) and then immobilise the affected limb. If it is a closed fracture or a dislocation, the broad aim is to immobilise the limb to make it more comfortable.

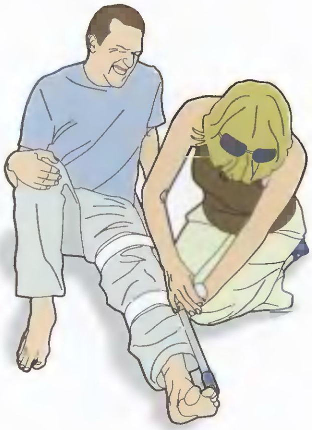

For leg fractures, the leg can be immobilised by improvising a splint from paddles and bandages, rope or strapping.

A lifejacket can be used to immobilise the lower leg. Place the ankle into the neck of the lifejacket with the lifejacket's body either side of the calf. Wrap it gently over the leg and apply strapping at the top, middle and bottom. Gently inflate the jacket by mouth until it is firm.

Seasickness

Seasickness is a debilitating illness that comes on rapidly, especially in the confines of a liferaft, at a time when the crew need to be as alert and able as they can be. If possible, take anti-seasickness medication before entering the raft or as soon as possible after boarding the raft.

Symptoms
Nausea and vomiting, pale clammy skin. Wishing never to go afloat again!

Treatment
Take anti-seasickness medication in the form of hyoscine hydrobromide. If vomiting has already started there is little point in taking more, because they are unlikely to be ingested. It's also possible that, if all the tablets are kept down, serious side effects will be experienced. Side effects can be drowsiness, dry mouth and blurred vision. Excessive doses of seasickness pills can produce hallucinations.

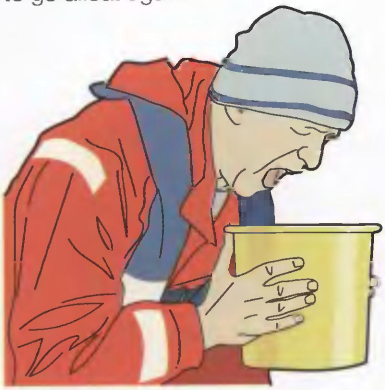

Do not attempt cardiac massage in a raft unless it has a wooden floor.

Resuscitation technique

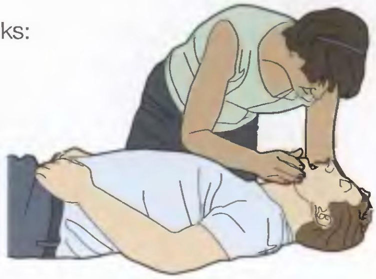

Before attempting resuscitation, make the following checks:

1. **Is the casualty responding to you?**

   If not, check the airway for breathing.

2. **Look at the casualty's chest. Is it moving?**

   If not, tilt the head back to open the casualty's airway.
   Check and remove any obstructions in the mouth.

Listen at the casualty's mouth for sounds of breathing. Hold your cheek close to their nose and mouth and feel for warm air movement. Alternatively, condensation forming on a cold mirror or metal spoon held close to their mouth and nose can also indicate breathing.

It is now accepted that if breathing is not present, circulation is not present either and that resuscitation should now be attempted. It is assumed that the casualty has suffered a cardiac arrest.

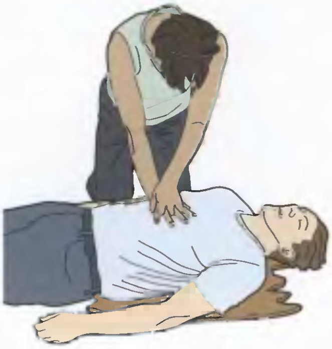

Place the heel of your hand in the centre of their chest. Place the heel of your other hand on top of the first hand. Interlock the fingers. Do not apply pressure over the upper abdomen or the bottom end of the breastbone (sternum).

Apply 30 chest compressions at the rate of 100 per minute.

Tilt the head back by placing a finger under the chin to open the casualty's airway. Allow the mouth to open but maintain chin lift.

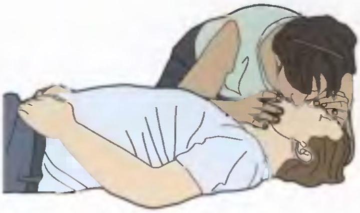

Place your mouth over casualty's mouth to achieve a good seal and blow into mouth.

Watch for the chest rising. Remove your mouth and allow the chest to fall.

Repeat the rescue breath. Start chest compressions again.

For non-drowning – generally cardiac arrest – casualties, give 30 chest compressions to two chest compressions.

For suspected drowning casualties: Check for breathing as for cardiac arrest. If not apparent, give five rescue breaths followed by one minute of chest compressions, continue until help arrives.

Pre-rescue collapse

Hypothermic casualties can suffer collapse at the thought of imminent rescue. Although not proven, it's believed that the feeling of overwhelming relief causes chemical changes within the body, which suddenly reduces blood pressure and so causes the casualty to collapse.

Circum-rescue collapse and hydrostatic squeeze

A combination of hypothermia with the sudden onset of physical action can result in some survivors collapsing and dying while being rescued.

Immersed casualties are at higher risk. In addition to the effects of hypothermia, water pressure on the lower limbs squeezes blood flow back into the body and is helping to support circulation.

When the casualty is lifted vertically from the water, the full effect of gravity creates extra strain on the heart. This may lead to unconsciousness or even death.

Immersed and hypothermic casualties should be treated with great care. If recovering from water they should be lifted horizontally to minimise the gravitational effects on the heart.

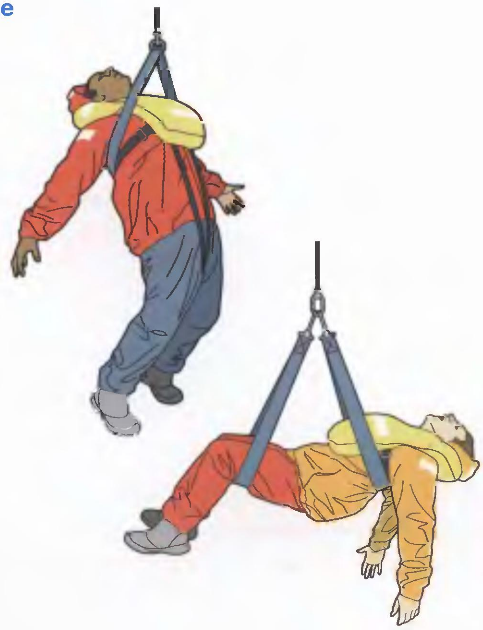

Post-rescue

Survivors are not survivors until they have survived.

Even after rescue there is a chance that the survivor will succumb to the effects of their ordeal.

This can be due to secondary drowning. Even those that show no immediate signs of aspirating seawater should be regularly checked during the hours immediately after immersion.

Rewarming survivors too quickly may result in a catastrophic drop in blood pressure. Active rewarming needs to be monitored professionally. Take care when taking hypothermic survivors into a warm cabin. Avoid warming by a shower or hot bath unless monitored by a medical professional.

Internal bleeding caused when the casualty fell into the water may not be obvious when a hypothermic MOB is recovered. If the MOB fell heavily, they should be monitored while the body rewarms. Seek medical advice.

CHAPTER 15 : Life-saving signals

These signals are used by ships, aircraft or persons in distress to communicate with rescue service stations, lifeboats, vessels and aircraft engaged in search-and-rescue operations. Use the most suitable signal for the situation and taking into account the prevailing conditions.

Search-and-rescue unit replies

These indicate that you have been seen and assistance will be given as soon as possible.

Orange smoke

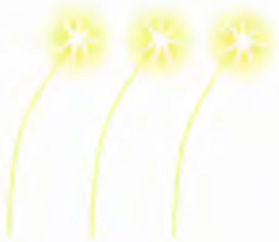

Three white pocket flares

Surface-to-air signals

These are shown by means of lights or flags or by laying out the symbol on the ground or deck in highly contrasting colours.

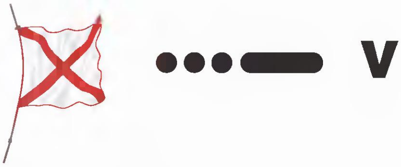

I require assistance

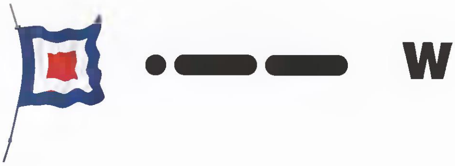

I require medical assistance

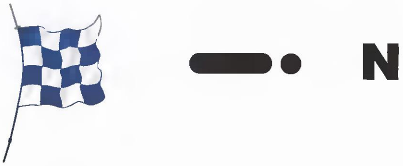

No or negative

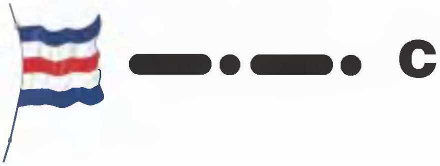

Yes or affirmative

Air-to-surface direction signals

Sequence of three manoeuvres meaning **'Go in this direction'**. May be used to show which way to go to assist another vessel or to indicate direction to a safe haven.

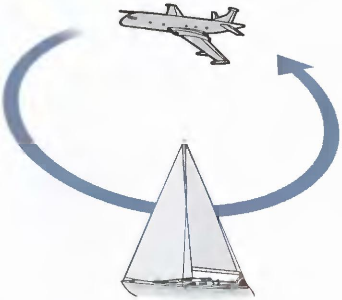

1. plane circles vessel at least once

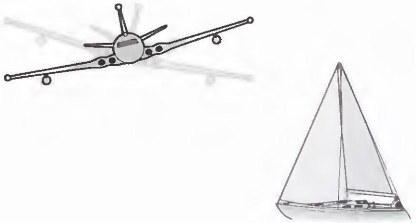

2. plane crosses low, ahead of boat, rocking wings

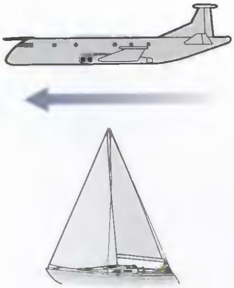

3. plane flies over boat in the direction to go

Your assistance is no longer required

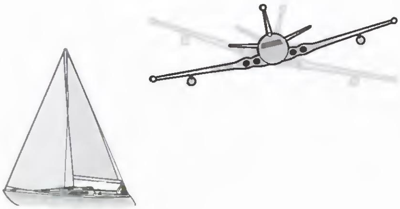

Plane crosses low, astern of vessel, rocking wings

Shore-to-ship signals

Safe to land here

Vertical waving of arms, white flag, light or flare.
Morse code letter K – dah dit dah

K – ▬ ● ▬

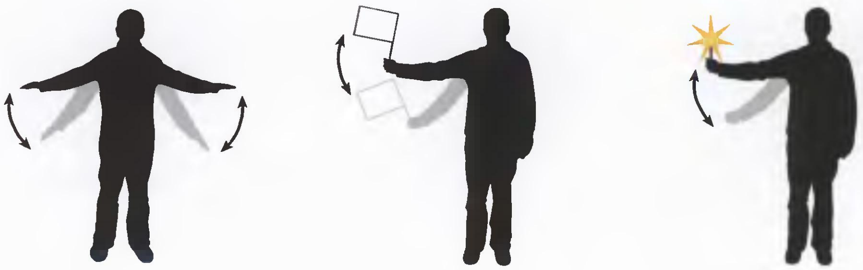

Landing here is dangerous with additional signals that indicate direction of safer landing place

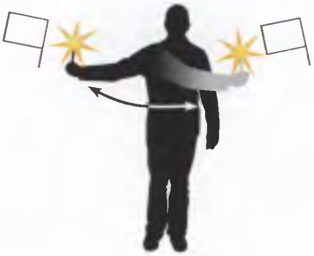

Horizontal waving of white flag, light or flare.

Go this way

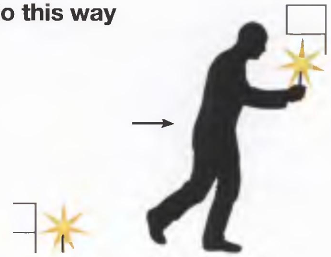

Putting one flag or flare on ground and moving in direction of safer landing area with the other indicates direction of safer landing

May also be shown by Morse code light or sound

S – ● ● ● = landing here is dangerous

R – ● ▬ ● = land to the right of your current heading

L – ● ▬ ● ● = land to the left of your current heading

Air-to-surface replies

Message understood

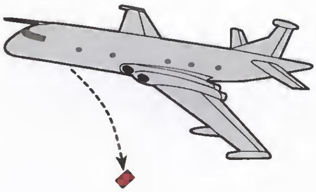

Drop a message

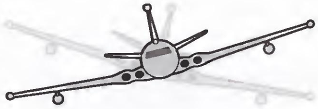

or rocking wings

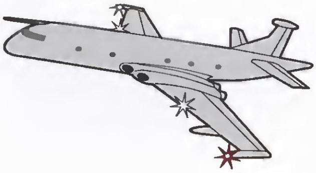

or flashing landing lights on and off twice

▬ ●▬●

T R

or Morse code by light T = dah or R = dit dah dit

Message not understood

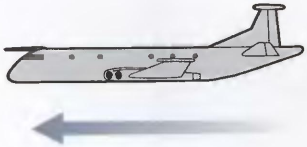

Straight and level flight

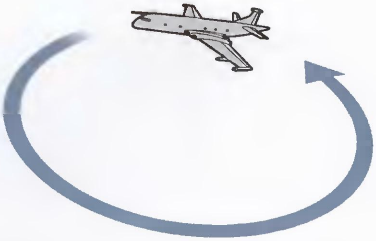

or Circling

Or Morse code by light R P T

●▬● ●▬▬● ▬

R P T

Surface-to-air replies

Message understood – I will comply

Change course to required direction

Or Morse code by light T = ▬ or code and answering pendant

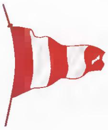

I am unable to comply

Morse code by light N = ▬● or code flag N

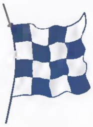

CHAPTER 16 : Rescue

Make yourself as visible as possible. An EPIRB will tell the SAR authorities where you are and who you are. But it's wise not to rely on any one system.

You will see potential rescuers before they see you. Try to make contact with them. Use flares only when there's a chance of rescuers seeing them. Use a handheld VHF and / or a signalling miror.

But be prepared for disappointment. A common theme of all survival stories is that they could have been rescued earlier if only the passing ship had seen them.

Don't celebrate early – they may not be able to pick you up and you might have to wait a few more hours. You haven't been rescued until you are safely on board the ship or helicopter.

Rescue by lifeboat

If rescued by lifeboat, you'll probably see the boat searching for you before they see you. Use a handheld flare to help pinpoint your position.

If it's dark, switch on the liferaft light and use a torch to signal to them. Talk to them on a handheld VHF – they will use their direction finder to home in on you – but help confirm their bearing by telling them where you are using a hand held GPS or relative to their direction of travel. Use nautical terms. 'I am on your port bow, half a mile…' etc.

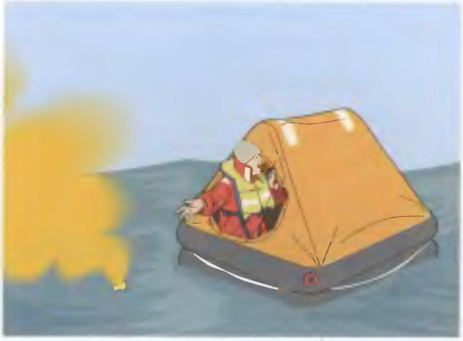

Tell them if you have injured casualties or children and the condition of the survivors. This will help the coxswain make the best plan for their evacuation. Tell them if there are any ropes or hazards around the raft – it may affect their approach.

As the lifeboat approaches, show the crew that you have a painter – assuming you cut it at the correct point when abandoning ship. The crew will then be able to secure the raft alongside during the rescue. If they throw you a heaving line, find a strongpoint and attach it – you will not have the strength to hold it while they pull the raft in.

The lifeboat will approach either head- or stern-to-wind to avoid becoming entangled in the sea anchor or other loose lines.

Rescue by ship

While many professional ships' crews practise emergency recovery, there are some that may have no idea how to go about retrieving survivors from a small boat or liferaft.

It's important to communicate effectively with them because it will be difficult for the captain and crew to appreciate the size and effects of the waves on a small vessel or liferaft.

Beware – the liferaft (or boat) can be sucked in to a lightly loaded ship's propellers.

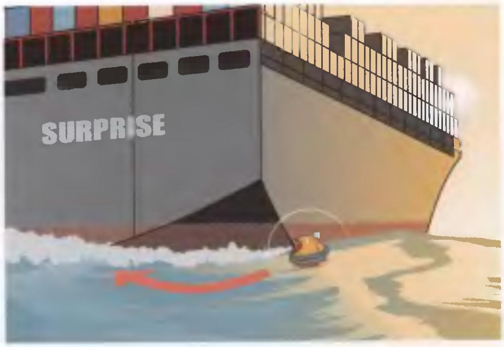

Due to exhaustion and the effects of the cold, you will not be in a fit state to climb a ship's ladder. In the excitement of the rescue, it's easy to think you can. There are many reported cases where survivors have tried to climb the ladder only to be lost when they fall off it after climbing only a few rungs.

The ship may be able to lower a basket or a cargo net to lift everyone in together. A large net can lift the whole raft with the survivors inside.

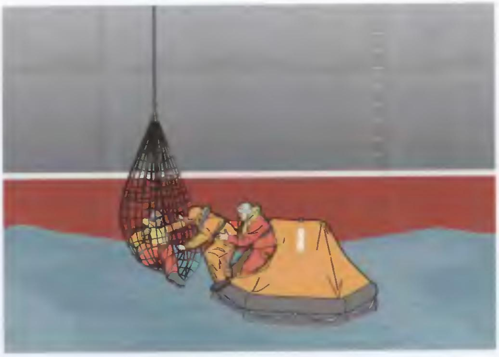

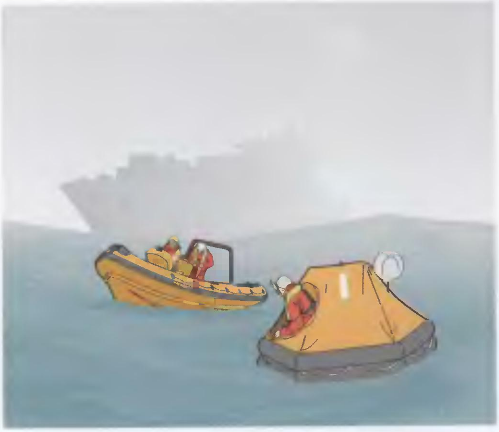

The best option is for the ship to lower a fast rescue craft that will come to the raft and pick up survivors. The ship will stand off and create a lee for the rescue boat to work. The rescue craft, its crew and the survivors will then be lifted back on board. This avoids exerting the survivors. Remember to take your small grab bag with ID documents etc.

Hypothermic casualties should lie down in the rescue boat with their heads towards the stern of the boat.

Rescue by helicopter

As with lifeboat rescue, the helicopter may have difficulty seeing you. If a 406/121.5 EPIRB or PLB is available, make sure it's switched on and operating. Switch on the liferaft lights – they may seem dim to you but a pilot using night-vision goggles will see them clearly, use your torch to signal to the helicopter.

If you have one, use a handheld VHF to communicate with the helicopter. Pilots use clock notation. For example, 'We are lying at your 4 o'clock'.

Use orange smoke – during the day – or handheld red flares to pinpoint your position.

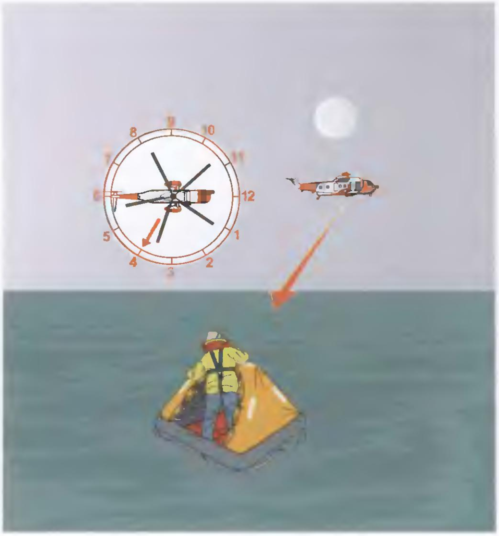

Do not fire parachute rocket flares if the helicopter is within a short distance.
A rocket flare has the capability to down a helicopter.
At night, a bright flare can blind a pilot wearing night-vision goggles. If in radio contact, ask before firing a hand-held flare. It probably will not be necessary.

The helicopter crew will instruct you what to do. Do as they say – they are the experts.

Take care not to all crowd to one side of the raft or dinghy – the prevailing wind and the downdraught of the helicopter may be enough to turn the raft over.

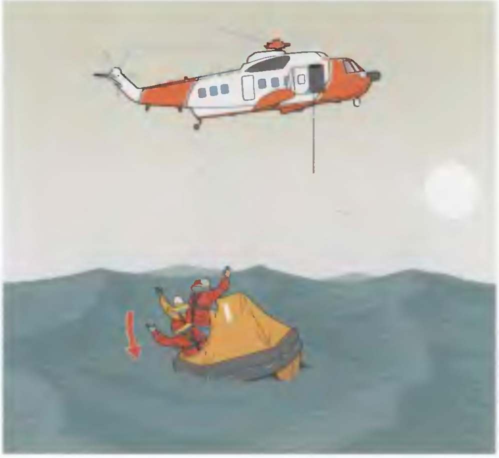

Most SAR helicopters will lower a winch man into the raft. Be aware that some countries' SAR units use different methods. Some may drop a rescue diver first, others may use a rescue basket.

Helicopters quickly build a charge of static electricity. Do not grab the winch man until he has 'earthed' into the water.

If the helicopter is using a 'hi-line' to help guide and steady the winch man, hold on to the line while the winch man is pulled back up to the helicopter. Do not tie it to the raft, and make sure it does not become tangled with the remaining casualties.

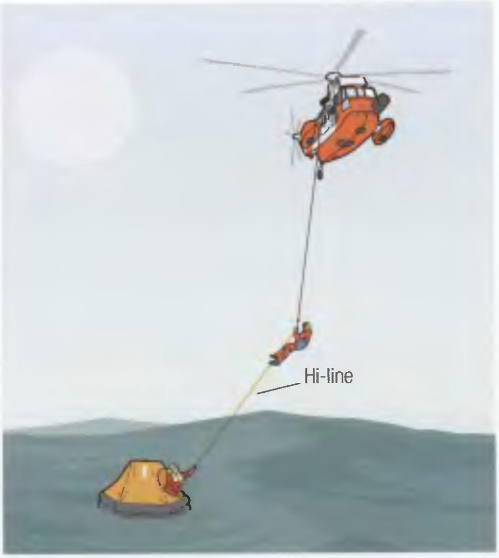

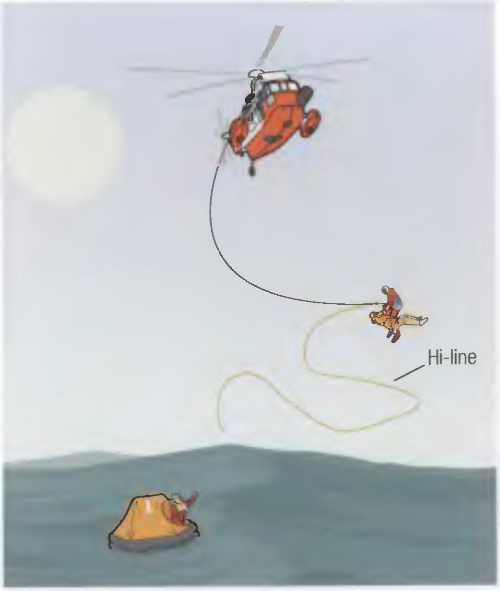

If the winch man suspects a casualty is hypothermic he will lift them in a horizontal position either in a stretcher or use a double strop – one underneath the arms and one under the back of the knees.

GLOSSARY

bear away	turn the boat downwind
bilge pump	pump for removing water from inside hull
bow thruster	an impellor mounted sideways in a tube below the waterline in the bow to assist manoeuvring
cleat	a fitting to which ropes are attached
coachroof	cabin roof
cold-water shock	physiological reaction to immersion in cold water
companionway	entrance to cabin, generally down steps
dan buoy	highly visible float thrown to a man overboard
datum point	marked position on which sector search is based
diving reflex	involuntary breathing reaction when falling into cold water
drogue	funnel shaped canvas device trailed from stern to help prevent craft from slewing sideways in following sea
DSC	Digital Selective Calling – automated VHF radio facility
EDR	Expected Detection Range
EPIRB	Emergency Position Indicating Radio Beacon – satellite-based location system
fairlead	metal fitting used to guide rope
fulcrum	pivot point
GMDSS	Global Maritime Distress and Safety System
GPS	Global Positioning System – satellite-based navigation system
grab bag	watertight buoyant bag for stowing essential items for emergency use
Granny bar	crew support bar around mast
gunwale	top edge of hull

halyard	rope for raising or lowering sail
handybilly	block and tackle
heaving to	bring boat to a standstill
heliograph	signalling system using mirror to reflect sun's rays
HELP position	Heat Escape Lessening Posture – position adopted in water
'hi-line'	winching technique used by SAR helicopters
HRU (hydrostatic release unit)	device triggered by immersion in water
hydrostatic squeeze	condition where water pressure on lower limbs squeezing blood back into body is removed, thereby placing strain on the heart
hypothermia	condition of having body temperature reduced to dangerously low level
immersion foot	condition caused by continuous contact with cold and wet environment
immersion suit	waterproof insulated emergency survival suit
ISAF	International Sailing Federation
ISO	International Standards Organisation
jackstay	line at deck level for attaching lifelines
jury-rig	makeshift repair, e.g. to sails or steering
kicker	control pulling boom downwards (also vang)
leeward	downwind direction
luffing	heading into wind
Mayday call	distress call
MMSI number	Maritime Mobile Service Identities number
MOB	man overboard

moment of inertia	force of rotation of boat dependent on weight of mast, keel etc.
ORC	Offshore Racing Congress
painter	mooring line attached to dinghy or liferaft
Pan Pan call	radio call seeking assistance
parbuckle	sling of canvas or rope used for raising objects
PEPIRB	Personal Emergency Position Indicating Radio Beacon – satellite-based location system (see PLB)
pitch-poling	capsizing with bow going over stern
PLB	Personal Locator Beacon – satellite-based or surface-based emergency location system
PRV	pressure-relief valve
reefing pennant	rope used to assist in reducing sail
RNLI	Royal National Lifeboat Institution
roller headsail	sail in front of mast set or reduced from a roller device
SAR	search and rescue
SART	Search-and-Rescue Radar Transponder – alert and position-finding electronic device
scandalise the boom	raise the outer end of the boom
sea anchor	parachute-like device deployed from bow to keep head-to-wind and reduce possibility of capsize or pitch-poling
seacock	valve in hull controlling seawater flow into boat
secondary drowning	physiological reaction caused by presence of water in lungs
Senhouse slip	release attachment for straps etc
shrouds	wires holding up the mast at sides of boat

signal card	card showing meanings of signals from ships and aircraft
snap shackle	fitting allowing quick attachment/removal of rope
solar still	solar-operated device for extracting potable water from seawater
SOLAS	Safety of Life at Sea Convention
sponson	inflated hull section of inflatable craft
spreaders	mast fittings guiding rigging from hull to mast head
stanchion	post supporting guardrail
steerage	steering effectiveness dependent on flow of water past rudder
storm jib	small sail set before mast for storm conditions
strum box	bilge pump filter
tack	change direction by bringing the wind around the front of the boat
topping lift	rope that holds the boom up when the sail is down
TPA	thermal protective aid – insulated bag-like suit for casualties
trysail	small loose-footed sail set in place of mainsail for storm conditions
turnbuckle	device for adjusting tension in rigging
vang	control pulling boom downwards (also kicker)
warp	mooring line
washboard	removable board at top of companionway to minimise ingress of water into cabin
Williamson turn	manoeuvre by powered craft to turn onto reciprocal track

INDEX

A

abandoning ship 82-105
  cold, protection against 98
  collecting sharp objects and weapons 95
  crew reaction to 36
  drills 82
  entering the water 86-87
  fire, because of 31, 34
  huddle for warmth 68, 95, 109
  keeping cool 104-105
  liferaft, boarding 85
    initial actions after 92-93
    secondary actions after 94-97
    from water 88
  liferaft, launching 83-84
  liferaft, righting capsized 90-91
  location, means of 97
  reasons for 34-36
  rescuing survivors, in liferaft 89
  routine, establishing 97
  saving boat 82
  survival, requirements for 35
advising person ashore 5, 10
anchors 11, 12, 14
anchors, sea 23
  in liferaft 79, 81, 92, 104
Australian Navy, Royal 35

B

Bailey, Maurice and Maralyn 103
bailing equipment 10, 12, 78, 79, 94
batteries 10, 12, 14
Beaufort scale 132-133
bellows 79, 80
bilge pumps 11, 12, 14, 26
birds, catching and eating 103
blankets, reflective 98
bleeding, severe 113
boat checklist, pre-voyage 14
  sailing boats 15
Bombard, Alain 103
bowel movements 97
Bullimore, Tony 35
bum-bags 81
bungs, sealing 11, 12, 14, 26
buoyancy aids 52
buoys, dan 10, 13, 14
  retro-reflective tape on 15
Butler, Bill and Simone 102

C

Callahan, Steve 102, 103
capsizing 20
care of boat and equipment 14-15
CDs for signalling 45, 80
clothing, cool weather 8
clothing, warm weather 9
Coastguard, HM 16
  informing 43
  Rescue Co-ordination
  Centre search procedure 123
  Voluntary Safety Identification
  Scheme (CG66) 5
collapse, pre- and circum-rescue 117
communication equipment 11, 13
see also distress situations: alarm, raising; radio, VHF

compasses 10
cool, keeping 104-105
crew, knowledge of 5, 6
crew reaction to yacht foundering 36

D

damage control and repair 24-26
  hatch damage 25
  hull damage 26
  jury-rig 24, 25, 35
  mast, broken 25
  rigging, standing 25
  rudder, lost 24
  steering failure 24
  VHF antenna, emergency 25
dan buoys 10, 13, 14
  retro-reflective tape on 15
desalinators, hand-operated 102
Digital Selective Calling (DSC) 41, 82, 122
dinghies 10, 13, 128
dislocations 114
distress situations
  alarm, raising 41-45
    divers 45
    DSC 41, 82, 122
    EPIRBs 47-48, 82, 97, 118, 120
    flares *see flares*
    heliograph (mirror or CD) 45, 79, 80
    Mayday by voice 42-43, 82, 122
    MOB alerting devices 50
    mobile phones 43
    PLBs 49-50, 97, 120
    SARTs 50-51, 97
    satellite phones 43
    signals 44
  GMDSS (Global Maritime Distress and Safety System) 40, 41
divers in distress 45
drinking cups, graduated 79, 81
drinking seawater and urine 100
drogues, liferaft 79, 81
drogues at sea, use of 22, 24
drowning, near 110
drowning, secondary 111, 117

E

Eclipse 43
Emergency Position Indicating Radio Beacons (EPIRBs) 47-48, 82, 97, 118, 120
enemas, water 80, 101
engine problems 10, 12, 14, 15
EPIRBs (Emergency Position Indicating Radio Beacons) 47-48, 82, 97, 118, 120

F

Fastnet Race (1979) 24, 70
fire blankets 10, 13, 29, 31
fire control 27-31
  fire blankets 10, 13, 29, 31
  fire extinguishers *see fire extinguishers*
  firefighting 31
  prevention of fire 27
fire extinguishers 10, 13, 14, 28-30
  location of 30
  maintenance 31

types of 28, 29
automatic 29, 30
fire prevention 27
first aid 106-117
bleeding, severe 113
collapse, pre- and circum-rescue 117
dislocations 114
fractures 114
frostbite 113
heart attacks 111, 115, 116
heat exhaustion 107
heat stroke 107
hydrostatic squeeze 117, 127
hypothermia *see hypothermia*
immersion foot 114
near-drowning 110
post-rescue 117
resuscitation technique 115-116
seasickness 94, 115
see also seasickness, tablets preventing
secondary drowning 111, 117
shock 112
splints, lifejackets as 106, 114
first aid kit 10, 13, 106
liferaft 79, 80
fish, catching and eating 103
fishing kit 79, 81
flares 37-40, 118, 120
accessibility 37
firing 39
handheld 38, 40, 118, 120
in liferaft pack 79, 80
non-distress 40
types of 38-39
food, catching 101, 102, 103
suggested diet 103
food in liferaft 79, 80, 102
food rationing 99
foods to avoid 100, 102
foot, immersion 114
Ford, USS 43
fractures 114
frostbite 113
fuel, sufficient 12, 13, 14

G

gas storage 10, 13
Gibb Hooks 33
Global Maritime Distress and
Safety System (GMDSS) 40, 41
Golden, F. 62, 108, 109
GPS, MOB button 122
grab bags 11, 13, 81, 82, 85
small 81, 119

H

Hammar hydrostatic release unit 77
Hammar lifejacket inflation system 55
harbour entrance, approaching 21
hatch damage 25
headsails, roller 18
heart attacks 111, 115, 116
Heat Escape Lessening Posture
(HELP) position 64, 66
heat exhaustion 107
heat stroke 107

helicopter, rescue by 120-121
helicopter, search pattern by 123
help required (Pan Pan call) 46
Hervey 62
huddling for warmth 68, 95, 109
hull damage 26
hydrostatic release units (HRUs), liferaft 76-77
hydrostatic squeeze 117, 127
hypothermia 62, 65-66, 95, 108-110
see also survival in cold water: body heat loss
  acute 108
  chronic 108, 109
  circum-rescue collapse 117, 126
  pre-rescue collapse 117
  preventing 66, 68, 82, 95, 96, 98
  recovering casualties with 119, 126
  symptoms 62, 95, 108-109
  treatment 109-110, 117

I

immersion (survival) suits 14
Indianpolis, USS 100
injured people in liferafts 85, 95
instructions, sea survival 79, 81, 97
International Sailing Federation (ISAF) 70, 71
International Standards Organisation
(ISO) liferaft standards 71, 78, 79

J

jackstays 11, 12, 32, 33
jury-rigs 24, 25, 35

K

knives, buoyant 78, 79, 92

L

labelling lockers 7
ladders, MOB recovery 11
lanyards, safety 33, 87
leadership 36, 96
lifeboats
see also Royal National Lifeboat Institution
  rescue by 118
  search pattern 123
lifebuoys 14
  retro-reflective tape on 15
lifejackets 52-61
  checks 56-57
  donning 60-61, 82
  effect on, in water 59
  fittings 58
  gas-inflated 54-55, 56
    donning 60-61
  inflation types 53-55
  lights 58, 97
  maintenance 56-57
  as splints 106, 114
  spray hoods 59, 66
  swimming in 67
  tips 52
  types of 52-54
lifelines 32, 126
liferaft packs 78-81, 96
  minimum contents list 79
liferafts 69-81

  boarding from boat 85
  boarding from water 88
  canisters 73, 74, 92
  floors 94, 96, 98, 105
  hiring 75
  hydrostatic release units (HRUs) 13, 76-77
  inflatable active survival craft 70
  injured people in 85, 95
  launching 83-84
  leisure liferafts 69, 72
  lights 97, 120
  maintenance 75
  painters 76, 83, 84, 87, 92, 118
  pressure-relief valves and stoppers 84, 93, 96
  protection against cold 98
  raincatchers 101
  for recovering MOB 126
  repair kit 79, 80, 96
  rescuing survivors 89
  righting capsized 90-91
  secondhand 69
  size required 72
  standards 70-71, 72
  stowage 73, 74-75, 97
  urinating in/from 97
  valises 73
  ventilation 93
Lim, Poon 103
lockers, labelling 7

M

maintenance of boat and equipment 14-15
  safety equipment 15
man overboard (MOB) 122-128
  actions to be taken by crew 122
  alerting devices 50
  parbuckle 126-127
  powerboat manoeuvre 125
  recovery methods 126-128
    boom lift 128
    elevator method 128
    for motorboats 128
  recovery raft 127
  sailing yacht maneouvre 124
  search patterns 122-123
man overboard prevention 32-33
  clipping on 33
  safety lanyards 33, 87
manoeuvres, sudden 12
mast, broken 25
mast 'Granny bar' 32
Maydays
see distress situations: alarm, raising
mirrors, signalling (heliographs) 45, 79, 80
MMSI number 7
MOB *see man overboard*
mobile phones 43
morale 36, 96, 101

N

navigation equipment 11, 13
navigation lights 11
notebooks, waterproof 79, 81

O

Offshore Racing Congress (ORC) 70, 71

P

paddles 114
paddles, liferaft 78, 79
Pan Pan call 46
parbuckles 126-127
Personal Locator Beacons (PLBs) 49-50, 97, 120
phones, mobile and satellite 43
planning, forward 5, 7
PLBs (Personal Locator Beacons) 49-50, 97, 120
Poon Lim 103
pumps, bilge 11, 12, 14, 26
pumps, water purifier 102
PUR Surviva 34
reverse-osmosis water filter 102

R

radar reflectors 51, 79, 81, 97
radios, VHF 14
- antenna, emergency 25
- Digital Selective Calling (DSC) 41, 82, 122
- handheld 97, 118
- usage 7
- versus mobile phones 43

reefing pennants 16, 18
registration schemes 5
rescue 118-121
- by helicopter 120-121
- by lifeboat 118
- *see also Royal National Lifeboat Institution*
- by ship 119

resuscitation technique 115-116
rigging, standing 25
Robertson, Dougal, Lyn, and family 80, 101
Royal National Lifeboat Institution
see also lifeboats
- MOB Guardian 50

rudder, lost 24
RYA book of Weather Forecasts (G5) 129
RYA first-aid course 106

S

safety brief 5-6
safety equipment, marking boat's name on 15
safety equipment for motorboats 12-13
safety equipment for sailing yachts 10-11
safety equipment maintenance 15
safety equipment stowage, plan of 7
sail, shortening 18
sailing yachts, preparing for heavy weather 16, 18
sails, storm 18-19
St John's First Aid Book 106
SARTs (Search-and-Rescue Radar Transponders) 50-51, 97
satellite phones 43
satellites, Cospas-Sarsat and Inmarsat 47, 49, 129
saving boat 82
sea anchors 23
- in liferaft 79, 81, 92, 104

seabirds, catching and eating 103
seacocks 11, 12, 14, 26
Search-and-Rescue Radar
Transponders (SARTs) 50-51, 97
searchlights 10, 13
seasickness 94, 115
  symptoms 115
  tablets preventing 79, 80, 82, 94, 115
  treatment 115
seasickness bags 80, 94
shelter, running for 21
ship, rescue by 119
shipping forecast terms 129-130
shock 112
shock, cold water 63-64, 82, 86
Siboney 102
signals, life-saving 134-138
  air-to-surface direction 135
  air-to-surface replies 137
  SAR unit replies 134
  shore-to-ship 136
  signal cards 79, 81
  surface-to-air 134
  surface-to-air replies 138
snap hooks, locking 33
solar stills 102
SOLAS Standards (liferafts) 70, 71, 72, 79
spares 10, 14
splints from paddles and lifejackets 106, 114
sponges 78, 79, 94
spotlights 10, 13
steering failure 14, 24
storm sails 18-19
stowage of safety equipment, plan of 7
stowing gear 16, 17
sun protection 9
sunglasses, wrap-around 9
survival craft, inflatable active 70
survival in cold water 62-68
  body heat loss 62, 95, 108
  see also hypothermia
    Heat Escape Lessening Posture
    (HELP) position 64, 66
    preventing 66, 68, 82, 95, 96, 98
  cold water shock 63-64, 82, 86
  diving reflex 63
  huddle 68
  swim failure 64
  swimming with lifejacket 67
survival (immersion) suits 82, 98
survival techniques, long-term 101-103
survive, will to 35-36, 96, 101
survivors, rescuing, in liferafts 89

T

thermal protective aid
(TPA) 79, 80, 96, 98,109
throwing line and quoit 79, 80, 124
tin openers 79, 81
Tipton 109
tool kits 10, 12
torches 10, 79, 80, 120
trysails 18, 19
turtles, catching and eating 103

U

urinating in/from liferaft 97
UV light 9

V

valuables, care of 81, 119
Vendée Globe race (1996-97) 35
VHF radios 14
  antenna, emergency 25
  Digital Selective Calling (DSC) 41, 82, 122
  handheld 97, 118
  radio usage 7
  versus mobile phone 43

W

waist-bags 81
washboards 10
watches on liferafts 97
water, collecting 100, 101, 102
water, drinking 100
  in liferaft 79, 80
  rationing 99
water ingestion by enema 80, 101
water purifier pumps 102
waves, breaking 20
weather 129-133
  clouds 130
  depression, typical 130
  fog 131
  forecasts 129
  heavy, preparing for 16-20
  pressure in forecasts 130
  shipping forecast terms 129-130
  visibility in forecasts 130
  wind scale, Beaufort 132-133
  winds, local 131
weather, heavy, handling 21-26
  damage control and repair
  see damage control and repair
  riding it out at sea 21-24
    drogues 22
    going downwind 22
    going to windward 21, 22
    heaving to 21
    sea anchors 23
  running for shelter 21
whistles 58, 78, 79
Wichard Hooks 33
Williamson turn 125
Wood, Richard, and partner Jetti 43

Useful Contacts

UK

Royal Yachting Association
RYA House, Ensign Way, Hamble,
Hampshire SO31 4YA
Telephone: 0845 345 0400
Email: admin@rya.org.uk
Web: www.rya.org.uk

Maritime and Coastguard Agency
Spring Place, 105 Commercial Road,
Southampton, Hampshire SO15 1EG
Telephone: 023 8032 9100
Email: infoline@mcga.gov.uk
Web: www.mcga.gov.uk

Royal National Lifeboat Institution
West Quay Road, Poole, Dorset, BH15 1HZ
Telephone: 0800 328 0600
Email: seasafety@rnli.org.uk
Web: www.rnli.org.uk/seasafety

British Marine Federation
Marine House, Thorpe Lea Road, Egham,
Surrey TW20 8BF
Telephone: 01784 473377
Email: info@britishmarine.co.uk

The Radio Licensing Centre
PO Box 1495, Bristol BS99 3QS
Telephone: 0870 243 4433
Web: www.radiolicencecentre.co.uk

CHIRP - Confidential Human Factors Incident Reporting Programme
FREEPOST (GI3439), Building Y20E, Room G15,
Cody Technology Park, Ively Road, Farnborough,
Hampshire GU14 0LX
Telephone: Freefone (UK only) 0808 100 3237

International Maritime Organisation
4 Albert Embankment, London SE1 7SR
Telephone: (0)20 7735 7611
Email: publications-sales@imo.org
Web: www.imo.org

USA

US Sailing
PO Box 1260, 15 Maritime Drive, Portsmouth,
RI 02871-0907
Telephone: 1-800-USSAIL 1 Ext. 1
Email: info@ussailing.org
Web: www.ussailing.org/safety

United States Coast Guard Boating Safety
Web: www.uscgboating.org

EPIRB registration

International
International Beacon Registration Database
Web: www.406registration.com

UK
The EPIRB Registry
Maritime and Coastguard Agency
Pendennis Point, Castle Drive, Falmouth,
Cornwall TR11 4WZ
Telephone: 01326 211569
Email: epirb@mcga.gov.uk

USA
SARSAT Beacon Registration
National Oceanic and Atmospheric Administration,
NSOF, E/SP3, 4231 Suitland Road,
Suitland, MD 20746
Telephone: 301.817.4515
Web: www.beaconregistration.noaa.gov
or by calling 1-888-212-SAVE

Australia
Beacon Registration Section
AusSAR, Australian Maritime Safety Authority (AMSA),
GPO Box 2181, CANBERRA CITY, ACT 2601
Telephone: 1800 641 792
Email: aus_beacon_query@amsa.gov.au

New Zealand
Rescue Co-ordination Centre New Zealand
PO Box 30050, Lower Hutt 6009
Telephone: +64 4 914 8383
Email: 406registry@msa.govt.nz

Useful Websites

Safety of Life at Sea Convention
Web: www.imo.org

Marine Accident Investigation Branch
Web: www.maib.gov.uk

First Aid
St John Ambulance
Web: www.sja.org.uk

BBC Health - First Aid
Web: www.bbc.co.uk/health/first_aid/index.shtml

RYA Sea Survival Handbook

Boating is one of the safest leisure sporting activities. However, in any sport that pushes us to the edge of our abilities, there's always a chance of accident or injury. Being properly prepared makes us better able to cope with an emergency and will significantly increase our chances of survival.

This handbook is essential for anyone taking the one-day RYA Basic Sea Survival course and the two-day RYA/ISAF Offshore Safety course. It is also invaluable for anyone who takes a boat offshore.

The RYA

The RYA is the UK governing body representing sailing, windsurfing, motor boating, power boating and personal watercraft, at sea and on inland waters. It works for the good of all who enjoy these activities, campaigning for their interests at local, regional, national, European and world level.

RYA training is recognised throughout the world and every year over 140,000 recreational and professional sailors take RYA courses and exams.

This book is one of over 80 published by the RYA. Some relate to specific training courses and are a valuable study and learning aid. Others provide technical, legal or general boating advice. All are expertly written, informative and authoritative.

RYA Publications 2007

For information about the RYA's training courses or publications, and the valuable benefits of being an RYA member, please visit our website at www.rya.org.uk.

RYA, the RYA tick logo and YACHTMASTER are registered trade marks of the Royal Yachting Association (registered in the United Kingdom and selected marketing territories)

ISBN 978-1-905104-31-4

www.rya.org.uk

RYA code G43

Cover design by Tony Collins

The official manual used by the AMGA to teach single pitch rock climbing

> "I've been waiting for this book for a long time. Professionalism, programming, pedagogy, anchors . . . this is a comprehensive resource for understanding the complexities of teaching in the single pitch environment. Highly recommended!"
>
> **—Arno Ilgner, AMGA Single Pitch Instructor and author of *The Rock Warrior's Way: Mental Training for Climbers***

> "If you want a thorough treatise on what to expect in the [AMGA SPI] course, as well as a ton of information on climbing gear and its evolution, this book will fill your needs. Even if you don't intend to take the course and just want to know what are considered best practices by the AMGA, the answers are here."
>
> **—Russ Clune, AMGA Single Pitch Instructor and standard-setting Gunks first ascentionist**

Rock Climbing: The AMGA Single Pitch Manual is intended to serve as a textbook for past and future participants of the American Mountain Guide Association's Single Pitch Instructor program. The book builds upon Bob Gaines' book *Toproping* to more specifically address the needs of the professional climbing instructor and advanced recreational climbers. It presents the most current, internationally recognized standards for technical climbing systems used in single pitch rock terrain. Included are chapters on effective teaching in the outdoor environment, risk management, professionalism, and rescue.

Bob Gaines has been a professional rock climbing guide since 1983. He is the owner of Vertical Adventures Rock Climbing School, voted the top rock climbing school in America by *Outside* magazine. An AMGA Certified Rock Instructor and AMGA Single Pitch Instructor Program Provider, he is the author of *Best Climbs Joshua Tree National Park*, *Best Climbs Tahquitz and Suicide Rocks*, *Rappelling*, and *Toproping*, and the coauthor of *Climbing Anchors* (all FalconGuides). Visit him at verticaladventures.com.

Jason D. Martin is a professional guide, writer, and the director of operations at the American Alpine Institute. He has guided in Red Rock Canyon, Joshua Tree National Park, the Cascades, the Sierra, the Alaska Range, Canada's Coast Range, and the Andes of Ecuador, Bolivia, and Peru. An AMGA Certified Rock Guide and an AMGA Single Pitch Instructor Program Provider, he is also a playwright, film critic, and outdoor adventure writer. He coauthored *Washington Ice: A Climbing Guide* and authored *Fun Climbs Red Rocks: Topropes and Moderates*. Visit him at alpineinstitute.com.

Front cover photo by Bob Gaines

facebook.com/AMGAUSA

@amgausa

made in the USA

Join our community

$21.95 / Canadian $23.95
ISBN 978-0-7627-9004-3
52195
9 780762 790043

Essential gear for every adventure: **Outfit Your Mind®**

FALCONGUIDES®

Rock Climbing

FALCON GUIDES®
falcon.com

FalconGuides®
is an imprint of
Globe Pequot Press

A
FALCON
GUIDE®

HOW TO CLIMB® SERIES

Rock Climbing
The AMGA Single Pitch Manual

BOB GAINES AND
JASON D. MARTIN

FALCON GUIDES

GUILFORD, CONNECTICUT
HELENA, MONTANA

AN IMPRINT OF GLOBE PEQUOT PRESS

HOW TO CLIMB™ SERIES

Rock Climbing

The AMGA Single Pitch Manual

Bob Gaines and Jason D. Martin

FALCON GUIDES

GUILFORD, CONNECTICUT
HELENA, MONTANA

AN IMPRINT OF GLOBE PEQUOT PRESS

To buy books in quantity for corporate use
or incentives, call **(800) 962–0973**
or e-mail **premiums@GlobePequot.com.**

FALCONGUIDES®

Copyright © 2014 by Bob Gaines and Jason D. Martin

ALL RIGHTS RESERVED. No part of this book may be reproduced or transmitted in any form by any means, electronic or mechanical, including photocopying and recording, or by any information storage and retrieval system, except as may be expressly permitted in writing by the publisher. Requests for permission should be made in writing to Globe Pequot Press, Attn: Rights and Permissions Department, PO Box 480, Guilford, CT 06437.

FalconGuides is an imprint of Globe Pequot Press.
Falcon, FalconGuides, and Outfit Your Mind are registered trademarks of Morris Book Publishing, LLC.

Photos: Bob Gaines unless credited otherwise.
Illustrations: Mike Clelland
Text and layout designer: Casey Shain
Project editor: Julie Marsh

Library of Congress Cataloging-in-Publication data is on file.

ISBN 978-0-7627-9004-3

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Warning: **Climbing is a dangerous sport. You can be seriously injured or die. Read the following before you use this book.**

This is an instruction book about rock climbing, a sport that is inherently dangerous. Do not depend solely on information from this book for your personal safety. Your climbing safety depends on your own judgment based on competent instruction, experience, and a realistic assessment of your climbing ability.

The training advice given in this book is based on the authors' opinions. Consult your physician before engaging in any part of the training program described by the authors.

There are no warranties, either expressed or implied, that this instruction book contains accurate and reliable information. There are no warranties as to fitness for a particular purpose or that this book is merchantable. Your use of this book indicates your assumption of the risk of death or serious injury as a result of climbing's risks and is an acknowledgment of your own sole responsibility for your safety in climbing or in training for climbing.

Globe Pequot Press, the authors, and the American Mountain Guides Association (AMGA) assume no liability for accidents happening to, or injuries sustained by, readers who engage in the activities described in this book.

Contents

Acknowledgments ... vii
Introduction ... viii
About the American Mountain Guides Association (AMGA) ... viii
AMGA Certification ... x
Concept Driven Techniques and the AMGA ... xii
How to Use This Book ... xiii

Chapter 1: The Role of a Climbing Instructor ... 1

Chapter 2: Professionalism

Pre-Program Preparations ... 3
At the Instructional Site ... 4

Chapter 3: Programming

Permitting ... 7
Risk Management and Liability ... 9
Risk Management in the Field ... 12
Stewardship ... 14
Instructor Meetings ... 16

Chapter 4: Pedagogy: The Art of Teaching

Learning Styles ... 17
Lesson Planning ... 20
Learning from Mistakes ... 22
Teachable Moments ... 23
Taking Kids Climbing ... 24
Teaching Climbing Movement ... 26

Chapter 5: Baseline Equipment

Harness ... 28
Helmets ... 31
Ropes ... 32
Knife ... 44
Slings and Webbing ... 44
Cords and Cordelettes ... 50
Carabiners ... 52
The Instructor's Pack ... 61

Chapter 6: Knots and Hitches

Loop Knots ... 62
Double Loop Knots ... 72
Knots for Webbing ... 78
Bends ... 80
Knots for Joining Two Ropes ... 83
Hitches ... 85
Stopper Knot ... 101

Chapter 7: Protection and Anchoring

Rock Assessment 102
Natural Anchors 108
Passive Protection 112
Active Protection 123
Fixed Anchors 135

Chapter 8: Anchor Systems

Fundamental Anchor Anatomy and Principles 145
Notes on Redundancy 146
Notes on Load Distribution 148
Vectors 162

Chapter 9: Advanced Anchor Rigging Systems and Institutional Anchors

Using a Setup Rope: Components Far Apart 164
Instructor Tether: Managing Risk at the Cliff's Edge 165
Extensions: Creating a Master Point Over the Cliff's Edge 167
Systems Using a Setup Rope 167
The Joshua Tree System—V Rigging with Extension Rope 168
The Lifeline System—V Rigging with Instructor Tether 173
The Backside System—Pre-Equalized Extensions 175
The Three-in-One System 179

Chapter 10: Risk Management at the Crag

Rockfall 183
Terrain Hazards 184
Closing the System 184
Systems Checks 184
Falling 185
Other Considerations 186

Chapter 11: Top- or Base-Site Management for a Single Pitch

Top-Managed Sites 187
Base-Managed Sites 190
Belaying 192
Belaying in a Base-Managed Site 196
Backup Belayers 199
Belaying from the Top 201
Lowering from a Top-Managed Site 204
Lowering with a Munter Hitch 204
Lowering with an Assisted Braking Device 204
Using a Backup When Lowering 205

Chapter 12: Rappelling

Teaching Rappelling 208
Student Rappel Belays 208
The Instructor Rappel Backup 210
Managing the Risks—
Avoiding Rappelling Accidents 216

Chapter 13: Basic Assistance and Rescue Skills

Assistance from the Base 221
Assistance from the Top 227

Chapter 14: Fixed Lines 231

Chapter 15: Leading

Student Lead Belay Training 234
Ground Anchors, Student Belay
Backups, and Knots 236
Descent Options 237
Teaching a Student to Lead 237

Appendix A:
Leave No Trace Ethics 240

Appendix B:
How to Determine a Fall Factor 242

Appendix C:
Autoblocking Devices 244

Appendix D:
Wilderness Medical Certifications 247

Appendix E:
Recommended Reading 248

Appendix F:
Climbing Instructor
Employment Resources 249

Index 252

About the Authors 257

Acknowledgments

The creation of a book like this is a grand effort with many parties involved. First, we would like to thank those who assisted the most in this process: Ron Funderburke and Ed Crothers of the AMGA, and John Burbidge and Julie Marsh from Globe Pequot Press. Ed and Ron provided most of the technical support for this text from the AMGA, and John and Julie patiently worked to put it all together. Special thanks to Casey Shain for the excellent job on the layout and design.

Thanks to the instructors, guides, and volunteers who took time out of their busy schedules to pose for, shoot, and donate photos for this project. In particular, Bob would like to thank Patty Kline, Mark O'Brien, Erik Kramer-Webb, Tony Grice, Dave Mayville, Ryan Murphy, Reggie Bulman, Kevin Jackson, Bryan Baez, Robin Depke, Ryan Bennet, William Jaques, Adam Radford, Wills Young, Nicholas Giblin, Chris Idiart, Dan Richter, Frank Bentwood, Terri Condon, David Berlin, Ben Shackelford, Andrew Jones, Eric Ludwig, Adam Hufford, Kris Jackson, Todd Gordon, and Chris Ramirez. Jason would like to thank Tim Page, James Pierson, Richard Riquleme, Jeremy Wilson, and Betsy Winter. We would also like to thank those who gave us feedback throughout this process. These include John Belanger, Tom Kirby, Casey O'Brien, and Dunham Gooding.

Thanks to our fellow guides and mentors who have taught us so much over the years. Jason would like to specifically thank Michael Powers, Art Mooney, Tom Hargis, Marcus Jollif, and Adam Fox. Bob would like to thank Jon Tierney, Alan Jolley, Peter Croft, Scott Cosgrove, Marcus Jollif, Dylan Taylor, Josh Jackson, Chris Baumann, Todd Vogel, Erik Kramer-Webb, Tony Grice, Tony Sartin, Pat Dennis, and Dave Mayville. Bob would like to give a sincere thanks to John Long for the many insightful conversations while working together on the Climbing Anchors books.

And finally, we would like to thank our families. Jason would like to thank Krista, Holly, and Caden for supporting all of his endeavors. And Bob would like to thank his wife, Yvonne, for her help with the photography.

Introduction

Rock climbing is a sport unlike any other. Hanging on by your fingertips, high above the ground, gives us a sense of focus and freedom that few other activities provide. We love the feeling of the warm sun on our backs while we stick the perfect jam or pull the crux move. We relish in our movement over the stone and the beauty of our surroundings. We love the element of exploration that comes with the sport and the anticipation of a trip to a new crag or the ascent of a new route. And we adore both the acquaintances that are made on the rock as well as the enduring friendships that are built within a climbing partnership.

We love it all so much that we have a desire to share it with others.

And while we wish to share this with others, we have to understand that the role of the climbing instructor comes with a great deal of responsibility. People can get hurt or even killed in this sport. With that in mind, it's important that those who wish to teach have the proper training before instructing any level of climber.

This text was designed to be a supplement to the American Mountain Guides Association Single Pitch Instructor (SPI) Course. The SPI course was designed to help capable recreational climbers transition into capable and effective climbing instructors. The course—and this text—focuses on the technical skills required of an instructor as they are applied in all forms of single pitch climbing instruction. In addition to this, the course addresses the essential educational and environmental tenets required to teach climbing. AMGA Single Pitch Instructors are expected to demonstrate the technical and educational proficiencies necessary to instruct a variety of single pitch rock climbing skills in a secure and effective manner to both groups and individuals.

The SPI course is intended for recreational climbers who are already proficient in both toprope and lead climbing. It was designed to benefit those who wish to facilitate outdoor climbing programs for groups such as those offered by guide services, camps, schools, universities, climbing clubs, therapeutic groups, churches, and climbing gyms.

Those who wish to teach climbing in the outdoors tend to have a higher success rate obtaining employment as climbing instructors with this training and certification. Additionally, the AMGA recommends this program as a precursor to upper-level AMGA climbing instruction and guide programs.

About the American Mountain Guides Association (AMGA)

The American Mountain Guides Association is the organization that oversees guide and instructor training in the United States, while also supporting and advocating for climbing instructors, guides, and accredited guide services. The organization is dedicated to supporting these groups through excellence in instructor and guide education, through the development of standards to ensure the quality of services provided to the public, and by serving as a resource for access to and the protection of the natural environment. As a group, the AMGA

AMGA presidents at the 2010 AMGA annual meeting. Back row, from left to right: Alan Pietrasanta, Ian Wade, John Cleary, Mark Houston, Dunham Gooding, Phil Powers, Dick Jackson, and John Bicknell. Front row, kneeling: Margaret Wheeler and Doug Robinson.

BETSY WINTER

presents a strong and unified voice for high standards of professionalism in guiding and climbing instruction.

The AMGA is grounded in a powerful tradition that continues to evolve with the ever-changing arena of mountain guiding and climbing instruction. The organization offers a series of training courses and exams designed to certify guides and climbing instructors to highly respected and internationally recognized standards.

A Brief History of the AMGA

The profession of instructing and guiding in the mountains spans back to the start of the United States as a nation. However, it wasn't until 1979 that a group of twelve guides got together in Jackson, Wyoming, and decided that it was time to formalize an organization to represent the greater guiding community. The goals these guides developed were written in a bar on the back of a napkin and became known as the Moose Bar Charter. As a result, the American Professional Mountain Guides Association was born. Over the next few years, the "P" was dropped, and the AMGA developed and grew throughout the United States.

In the AMGA's formative years, there was a great deal of wrangling. Instructors and guides are independent minded, and an organization overseeing standards wasn't always popular. Early presidents such as Doug Robinson, Alan Pietrasanta, Ian Wade, and Dunham Gooding spent much of their time trying to expand the ranks of the guides involved, while also overseeing the development of certification courses for guides and accreditation programs for guide services.

Unity among guides and guide services became especially important in the mid-1980s. At that time a worldwide liability insurance crisis shook the guiding industry to its core. After steps were taken to secure affordable insurance appropriate for climbing schools and guide services, independent-minded guides began to see that there was great value in working together as an association.

The next major goal of the AMGA was to gain acceptance into the International Federation of Mountain Guides Associations (IFMGA). Admission into this organization would allow American guides access to the European Alps while recognizing that AMGA programs met an international standard. Between 1990 and 1997, AMGA presidents Steve Young, Dunham Gooding, Matt Culbertson, and John Cleary worked tirelessly to achieve this goal. On November 22, 1997, delegates admitted the AMGA into the IFMGA after a unanimous vote from the existing member countries. Today the AMGA is the sole representative of the United States to the twenty-six-member IFMGA, the international governing body responsible for guiding standards and education around the world.

In the late 1990s, under the supervision of Mark Houston, both as technical director and then as president, and with help from Jon Tierney and others, the AMGA developed the Toprope Site Manager program. Unfortunately, the TRSM program was limited in its scope and didn't adequately meet the needs of instructors looking to advance to higher-level AMGA courses. It wasn't until 2008, under the supervision of Adam Fox, that the Single Pitch Instructor program became the nation's premier entry-level program for those who sought to be climbing instructors. In 2009, after a third party review, the AMGA SPI received the endorsement of the UIAA (Union Internationale des Associations d'Alpinisme) as meeting the international standards for the training of climbing instructors.

Today the AMGA continues to develop its programs to meet changing standards and to support the growing community of guides and climbing instructors in the United States. In addition, the organization has become a resource for land managers and outdoor industry leaders by promoting land stewardship, world-class training, and sustainable practices to protect our natural resources.

The AMGA is the post-graduate school of guiding, and it is where professional climbing instructors and guides go to develop their skills in order to attain the highest level of training and certification.

AMGA Certification

There are three disciplines of certification in the AMGA: Rock, Alpine, and Ski. Some of the disciplines have multiple levels of certification. For example, in the Rock discipline one can be certified at the Single Pitch Instructor level, the Rock Instructor level, or the Rock Guide level. In each of the disciplines, the guide level is the highest level of certification. If an individual obtains AMGA certification at the guide level in Rock, Alpine, and Ski, then that person will be granted the International Federation of Mountain Guides Associations (IFMGA) certification, which is the highest level of certification available for a guide.

The Single Pitch Instructor certification was designed for those who wish to work in a single pitch setting and as an entry-level program for those who wish to continue on to gain a higher level of training and certification.

If you are interested in working solely in a single pitch setting, then there is no reason to continue on to a higher level of training. However, if you do wish to work on multi-pitch rock terrain, in the alpine, or on skis, then additional courses will be required in order for you to be adequately trained.

Following is a breakdown of the different types of certification available from the AMGA and the terrain guidelines for the certifications:

AMGA Single Pitch Instructor

AMGA Single Pitch Instructors are trained and assessed to operate on all routes that:

- The anchors are accessed by either nontechnical terrain or by leading technical 5th-class terrain.
- Are climbed without intermediate belays.
- Present no difficulties on approach or retreat from the area, such as routefinding, scrambling, or navigating.

AMGA Climbing Wall Instructor—Toprope

AMGA Climbing Wall Instructors (Toprope) are trained and assessed to operate on indoor climbing and bouldering walls and climbing structures that:

- Do not involve lead climbing.
- Do not include the instruction of lead climbing.

AMGA Climbing Wall Instructor—Lead

AMGA Climbing Wall Instructors (Lead) are trained and assessed to operate on indoor climbing and bouldering walls and climbing structures that:

- Involve lead climbing.
- Include the instruction of lead climbing.

AMGA Rock Instructor

Rock Instructor certification is designed to apply to most "cragging" rock climbing areas in the United States. AMGA Rock Instructors are trained and assessed to operate on all rock routes that:

- Are Yosemite Decimal System Grade III or shorter in length.
- Do not have approaches or descents where extensive use of short-rope techniques is appropriate for security.
- Are not remote in nature.
- Do not involve technical snow, glacier travel, the use of skis, or exposure to avalanche risk.

AMGA Rock Guide

Rock Guide certification is designed to apply to all Rock Instructor terrain, but it also includes climbs of much greater length, including big wall and aid climbs, and climbs where significant short-roping is appropriate for the safeguarding of clients on either approaches or descents. AMGA Rock Guide certification is not, however, applicable to terrain that is alpine in nature (i.e., remote or involves technical snow or technical glaciated or icy terrain). AMGA Rock Guides are trained and assessed to operate on all rock routes that:

- Are Yosemite Decimal System Grade IV or Grade V in length.
- Include approaches and descents that may require extensive use of short-roping technique for security.
- May be remote in nature.
- Do not involve glacier travel, the use of skis, or exposure to avalanche risk.

AMGA Alpine Guide

AMGA Alpine Guide certification is designed for guiding glaciated and non-glaciated peaks, approaches, and climbs, with no limitation with respect to season or elevation. It includes rock climbs, peak ascents, waterfall ice climbs, and expeditionary climbing. AMGA Alpine Guides are trained and assessed to operate in all alpine and rock terrains that:

- Are Yosemite Decimal System Grade V or shorter (for rock routes).
- Are not conducted on skis where the main objective of the outing is either a ski tour or the enjoyment of downhill skiing.
- AMGA Alpine Guides who hold both Alpine and Rock Guide certifications are allowed to guide alpine rock routes that are YDS Grade V or longer.

AMGA Ski Guide

AMGA Ski Guide certification is designed for guiding either ski tours or ski ascents/descents on alpine touring or free-heel equipment. AMGA Certified Ski Guides are trained and assessed to operate in all terrains that:

- Include as the main objective either ski touring, off-piste, and/or mechanized skiing.
- Are on slopes up to 50 degrees.
- May employ the use of rope including short-roping techniques with the use of ice axe and crampons.
- Require travel on crevassed glaciers.
- Have significant avalanche hazard.
- Involve multi-pitch ice or rock.

IFMGA American Mountain Guide

IFMGA American Mountain Guide certification is designed for individuals certified as AMGA Rock, Alpine, and Ski Guides. IFMGA/American Mountain Guides are trained and assessed to operate in all alpine, rock, and ski terrain.

Concept Driven Techniques and the AMGA

"I want to learn the AMGA way."

The preceding comment is incredibly common. But it is a misperception of what the AMGA is and what it does. There is no "AMGA way."

The American Mountain Guides Association teaches students a series of different techniques in each of the disciplines. Association guidelines seldom dictate which technique an individual is required to use. Instead, AMGA program instructors evaluate their students on the application of the techniques that they have learned. In other words, one student might accomplish the exact same goal using a completely different technique than another student, but both might still be acceptable.

There are a lot of right ways to do things. Some ways may be more efficient and some ways may be more effective. The goal of an AMGA instructor is to find the best and most efficient way based on the circumstance.

In the Single Pitch Instructor curriculum, there are a handful of prescribed techniques that are taught. However, the bulk of the material is driven by concept. A concept is a general idea of how something should work. For example, you may wish to set up a toprope with an anchor that self-adjusts to a variable direction of load. There are a number of ways to do this. You could use a quad or a magic-x with load limiting knots, or an equalette. What you use doesn't matter. Instead, what matters is that you achieved a self-adjusting system, which is conceptual in nature.

In the SPI curriculum there are a number of concepts. An SPI provider's goal is not necessarily to have a student ape exactly what is in this text, or exactly what an individual instructor has taught, but instead to demonstrate an understanding of each concept and how to apply it.

How to Use This Book

This book is broken into two major sections. The first part (chapters 1 through 4) covers instructor knowledge. The information here, concerning everything from professionalism to pedagogy, combines to provide a foundation of base-level knowledge that every climbing instructor should have. The second part (chapters 5 through 15) delves into the specific technical skills that are needed to effectively run programs in a single pitch setting. Like the first part, this second part also provides a foundation, though this foundation is built of technical skills.

To fully understand and appreciate the material presented in this text, readers should have a solid understanding of toproped climbing and have experience with traditional climbing. This is not a "how-to-toprope" book or a "how-to-lead-climb" book, but a book that assumes you already have those skills and wish to take them to the next level.

This text was specifically designed as a supplement to the AMGA SPI program. It was not designed to be used by independent practitioners without the oversight of a qualified instructor. Readers of this book are strongly advised to work through the material herein with an AMGA Single Pitch Instructor Program Provider.

AMGA Single Pitch Instructors Tom Kirby and Erin Smart toprope at Mount Erie in Washington State.

JASON D. MARTIN

CHAPTER 1

The Role of a Climbing Instructor

Working as a climbing instructor is a very special job. The dream is alluring: work at the crag every day under cloudless skies, become a master climber, explore new places, meet fascinating people, and teach them the art of rock craft. And while the dream is alluring, it doesn't reflect the complete picture. The reality is that when you become a climbing instructor you are no longer able to follow your climbing whims or operate the way you may have in the past as a recreational climber.

Working as a climbing instructor and climbing recreationally are two very different things. Recreational climbers assume a limited amount of responsibility for one another. There is an implied assumption between partners that each person will be the primary manager of his or her own safety.

Professional climbing instructors are not climbing recreationally and are not climbing for themselves. The job requires managing risk first, followed closely by the quality of the student experience.

The most critical role that a climbing instructor provides is that of a risk manager. First, we have to manage our own exposure to the inherent risks of working in a vertical environment. There is no student security without instructor security. And second, we must manage the risk to our students.

While risk management is the most important aspect of the climbing instructor's job, the quality of the student experience comes in a close second. When working in a climbing program, personal goals and ambitions are put aside. It is not about you "living the dream," but instead is about the students, their goals, and their ambitions.

At the outset one of the best ways to achieve the dual objectives of managing risk and student experience is to choose an appropriate crag. If the crag is appropriate to the student(s), then both the objective and subjective dangers will be manageable. Additionally, the student climber will feel comfortable in the setting and will be more emotionally open to learning the art of climbing.

The other major difference between recreational climbing and climbing instruction is the obvious fact that a climbing program is an instructional setting where students look for advice and training. This means, for example, that once an appropriate crag has been chosen, then the next step is to choose an appropriate curriculum. Think about where a student is at the beginning of the day, and what the desired goal is at the end. Don't skip or rush through instruction to accommodate instructor desires. Focus on where the student is and where he or she needs to be.

The most effective instructors are those who recognize these distinctions early in their instructional career. The most ineffective are those who attempt to teach while still acting as a recreational climber.

Terri Condon leads a trad climb during her AMGA Single Pitch Instructor Assessment at Joshua Tree National Park, California.

CHAPTER 2

Professionalism

Doctors, lawyers, and teachers don't have too many conversations about professionalism. They don't have to. They act the part without anybody telling them to. They dress appropriately for their workday, they use the language of their respective professions, they make concrete plans about how they are going to handle the challenges of their day, and they are respectful of those they're working with. Additionally, professionals are up to date with the best and most current practices in their field, they undergo formal training, they demonstrate a level of mastery through peer assessment and certification, they pursue continuing education, and they adhere to a code of ethics.

When climbing instructors act professionally, they represent three different interests. First, they represent themselves. If an instructor would like to be thought of as a respected professional, then she has to act as one. Second, they represent the companies, schools, or programs that they work for. In this day and age, a lack of professionalism can lead to negative press online, which can have a serious and long-term impact. And third, instructors represent all climbing instructors. If a recreational climber has a negative experience with one climbing instructor, that often creates an impression of climbing instructors in general.

Clearly, if an instructor develops a reputation for a lack of professionalism, it will affect his or her bottom line. Program directors don't want to hear anything negative about their instructors. What they want to hear is that everybody had a successful and enjoyable day.

Presenting oneself as a professional engenders trust and confidence from one's students. But there's a great deal more to professionalism than just attitude. Professionalism is something that should permeate every instructional day from the moment a student has contact with an organization to the moment he or she leaves.

Pre-Program Preparations

Professionalism starts in the office. The best way to present a professional program is to be prepared. Following is a short list of questions to consider before meeting students at the crag:

- Are there land-use permits in place or has permission to operate at the location been granted? If charging for a program on public lands, a land-use permit will be required. If operating on private lands, you may need to obtain permission.
- What are the rules and regulations regarding the area?
- Do you have liability insurance, or do you have coverage from the organization that you're working for?
- Do the marketing materials accurately reflect the experience that you will provide your students?
- If working a course for an organization, do you know what the marketing materials say about the course or program that you intend to run?
- Do the students have any medical conditions?

- Do you have student emergency contact information?
- Did you collect climbing biographies from your students?
- Have you or a representative from your organization been made available to the students to answer any questions that they might have?
- Have you researched the area where you plan to operate? Do you have a Plan B, in case your initial plan for the day falls through due to crowding or some other unforeseen circumstance?
- If it's going to be an instructional day, do you have lessons prepared?
- Do you have a release or acknowledgement of risk form for your students?
- Have you developed an emergency plan?
- Do you have appropriate first-aid training for the location that you intend to work?
- Do you have other required credentials? Some areas require instructors to obtain a state guide's license, and others require individuals to hold a Leave No Trace Trainer or Master Educator certification. If you're providing meals, some areas may require that you hold a food handler's license.

Obviously, some of these issues will be the responsibility of the organization an instructor works for. However, if the organization doesn't provide information on the questions that seem to be its responsibility, then it's the instructor's responsibility to make sure that each of these questions has been answered.

At the Instructional Site

Professionalism continues from the office and into the field. Once in the field, an instructor is a role model not only for the students, but also for other recreational climbers in the area. With that in mind, instructors should model the way that they believe people should act at the crag by being considerate and having a low impact.

Managing both the students and the scene is an important part of the day. This management doesn't necessarily relate only to the technical and pedagogical aspects, but it also pertains to how the instructor relates to the group, how the group acts, and how the instructor interacts with recreational climbers.

The crag needs to be a place for people to experience climbing without feeling put down or threatened. Threatening language in particular is sometimes not obvious and can come from inappropriate comments or jokes that might seem minor to an instructor, but can have an impact on one or more of the students. It is important to continually monitor the language, jokes, and conversations taking place so that all members feel respected.

When in contact with recreational climbers or other instructors, be professional, courteous, and friendly. Try not to tie up routes when other parties are waiting and make sure that students park themselves and their gear in unobtrusive places.

The ropes, harnesses, and equipment used on a program should be in good repair. The instructor's rack and rope should be organized in the instructor's pack before arriving at the crag. When unpacked, everything should remain organized. Avoid creating a "yard sale" at the base of the crag with equipment and packs scattered all over the place.

Most instructors choose to carry a notebook with student information, lesson plans, and time management strategies. This helps an instructor stay on track and organized. An organically run day without set lessons, strategies, and notes is feasible for advanced-level instructors, but should be avoided by those new to the profession.

The way that each individual instructs is different, and it's often difficult to determine how long it will take to teach certain lessons or to cycle through students at a toprope or rappel site. For new instructors it's important to keep track of how long each activity takes in the instructor's notebook. This will help with time management strategies on future programs.

Managing student time is only one aspect of time management. Often the most time-consuming element of the day is setting up and taking down systems. It is imperative that instructors have mastery of these foundational skills. Additionally, when planning for the day, instructors need to remember to account for this time.

AMGA Single Pitch Instructors are required to be able to lead 5.6 traditional routes and to toprope 5.8, both in good style. These are minimum standards, and most instructors climb harder than the standard. Though an ability to "climb hard" is not a prerequisite for an instructor, an ability to model good technique is. Professional instructors should constantly work on their climbing skills in order to demonstrate the best technique possible.

Professionalism at the crag is essential. A climbing instructor's professional demeanor should be a model for students, for recreational climbers, and for other instructors from before the trip starts to the moment that the last student leaves.

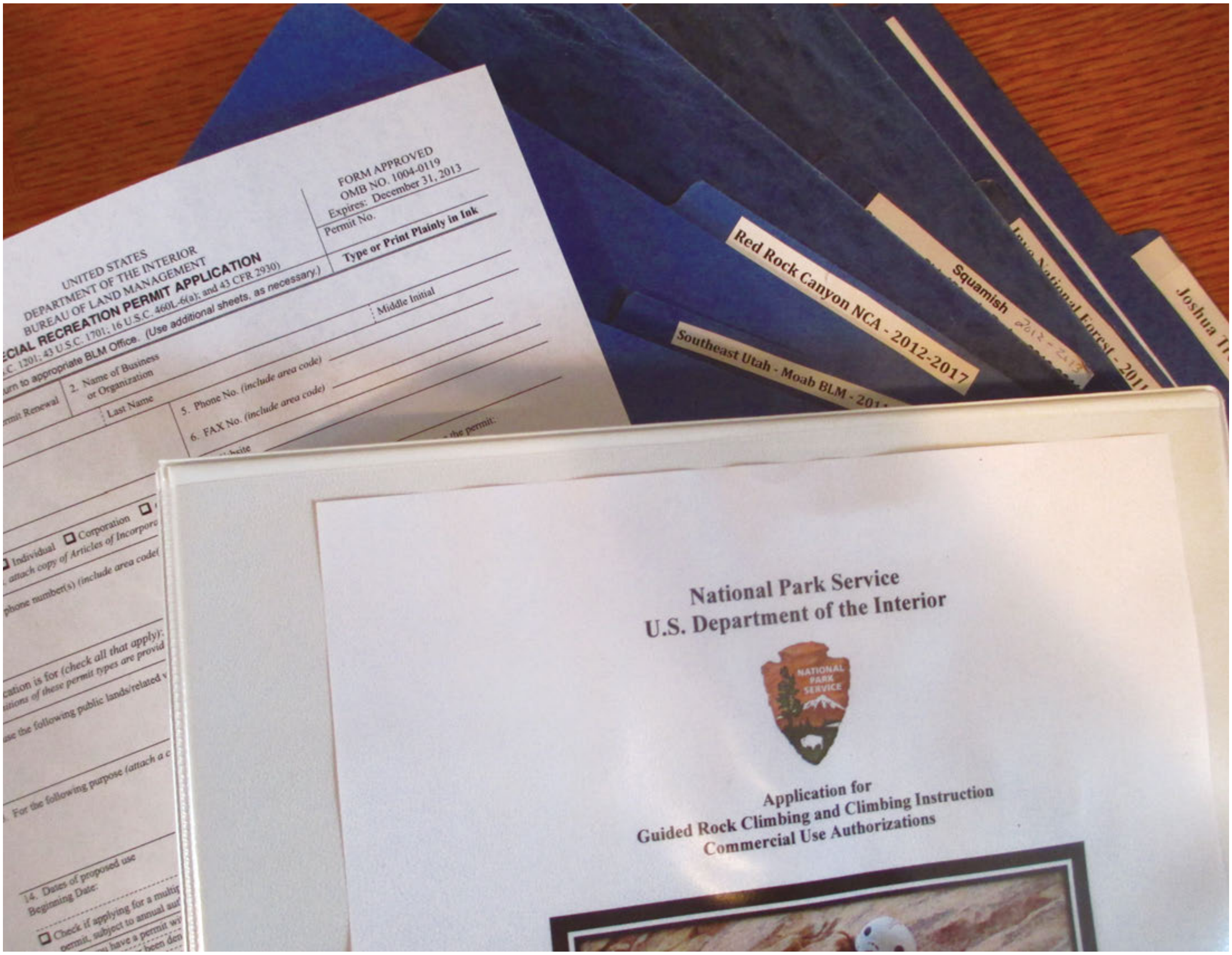
UNITED STATES
DEPARTMENT OF THE INTERIOR
BUREAU OF LAND MANAGEMENT
ECIAL RECREATION PERMIT APPLICATION
FORM APPROVED
OMB NO. 1004-0119
Expires: December 31, 2013
Permit No.
Type or Print Plainly in Ink
Red Rock Canyon NCA - 2012-2017
Southeast Utah - Moab BLM - 201
Squamish
Joshua Tr
National Park Service
U.S. Department of the Interior
NATIONAL PARK SERVICE
Application for
Guided Rock Climbing and Climbing Instruction
Commercial Use Authorizations

A stack of permits and permit applications at the American Alpine Institute.

JASON D. MARTIN

CHAPTER 3

Programming

Most climbing schools offer a variety of courses. Within the simple beginner to advanced-level framework, there might be any number of courses that focus on a specific skill or a specific series of skills. These could include anything from beginner courses for people who have never put on a harness before, to advanced courses in high-end movement or traditional lead climbing.

Regardless of the course, the one thing that instructors must keep in mind is that students signed up for a specific curriculum, and aside from managing risk, the most important instructor goal of the day should be to honor that curriculum.

To provide the curriculum that has been sold, instructors will need to find an appropriate crag for the program. To operate at the crag, they will need a land-use permit and an understanding of liability. They will need to be aware of risk management issues and develop an emergency response plan. They should be conscious of crag etiquette and advocacy. And they will need to have functional equipment for their programs and foster good communication with other staff members. In other words, to provide the curriculum that has been sold, the instructor will have to do a lot of prep work.

Permitting

In the United States, climbing destinations are managed by a patchwork of different agencies that operate independently of one another. On the federal level the primary land managers include the National Park Service, the USDA Forest Service, and the Bureau of Land Management. Regional land managers include state, county, and city parks and private lands. This variety and the many different and sometimes contradictory rules and regulations pose serious administrative challenges. In addition to all this, the resource is not growing, but the demands on the resource are.

The popularity of recreation on public lands has exploded over the last decade. This has resulted in more and more commercial operators vying for a limited number of permits. The combination of this and the mandate that land managers preserve and protect wild places has created an often-frustrating challenge for independent instructors and organizations that seek land-use permits.

Legal access to public lands in particular is important because most climbing instruction takes place on such lands. If the students pay a fee of any sort to participate—whether for a commercial operation, a nonprofit, an individual instructor, a school, or a university—then a permit is required to operate legally on public lands.

Operating on public lands without a permit is breaking the law and could result in a wide array of consequences. These range from a minor citation to a significant fine or even an arrest. In some cases it is possible to be permanently banned from public lands for willfully disregarding land-use regulations. Indeed, there are some insurance providers that will deny coverage if they discover that an operator was working without a permit at the time of an incident that resulted in a lawsuit. Working on public lands without a permit is serious and should never be taken lightly.

There are two main questions that an individual or an organization should ask to determine whether or not a land-use permit is required:

(1) Did the participants pay to engage in the activity?

Did they pay any fee at all? Were they asked to pay tuition for the course or guided days? Were they asked to submit a donation to the program? Are they required to rent equipment from the operation in order to participate?

If the answer to any of these questions is yes, then the land manager should be contacted to determine if a commercial-use permit is required. Even if you are a volunteer instructor for a club, but the club is taking in funds, the land manager should be contacted to determine if a permit is required.

(2) Are the instructors getting paid to run the program or to provide the instruction?

If the instructors receive any kind of compensation, then they or the organization will need to obtain a permit. Compensation could mean an hourly or daily wage, a stipend, meals, or even "free" equipment. Nearly any type of compensation outside of a thank-you constitutes commercial use.

There is a bit of a gray area when it comes to salaried teachers or professors who work with students from their school at a crag. Some land managers will require such a group to obtain a permit, whereas others will allow it to operate without one. Teachers and professors should contact the land manager at the desired location for more information.

Some public lands have special permits for educational institutions and not-for-profits. Educators are encouraged to contact land managers to see if they have programs like this available. If an area does not offer "educational permits," it is encouraged to suggest such a program to the manager.

The first step to obtaining use authorization is to contact the land management agency and get in touch with the person who handles recreational commercial or nonprofit use.

If the land manager is accepting new permit holders, he or she will provide an application. It is not uncommon for the application process to take a significant amount of time, so plan adequately.

Nearly all permit applications will require some kind of application fee. It is not uncommon for land managers to also ask for additional monies prior to the issuance of a permit. These include, but are not limited to, monitoring fees, reservation fees, or administrative fees. The costs for all of these vary based on the land management agency and the type of permit.

Nearly every land manager will require commercial users to hold general liability insurance before issuing a permit. The land manager is usually included in the policy as an "additional insured." This liability policy—often between $500,000 and $1,000,000—provides them some cover if an accident takes place during a program.

At the end of the year, the permit holder will be required to submit reports that include revenues and user days. Many contracts include a clause that states that the land manager has the right to audit a permit holder at any time. Therefore it is important for organizations to keep excellent financial and user day records, as failure to do so could lead to the loss of the permit. It should be noted that many land managers compare your reports to any other reports that they have on hand. For example, if an instructor has to sign in at a ranger station before operating each day, the land manager will compare the final report to the information gathered at the station.

Fees are calculated in different ways. Some land managers charge a flat fee for operations, whereas others charge anywhere from 2 percent to 4.5 percent of gross income. Most charge 3 percent of gross.

Guides, instructors, and guide services spend a great deal of time trying to obtain and then maintain their permits. For many, this is a major administrative duty. As a result, the way that the majority of new guides and instructors choose to access public or private lands is through guide services or other organizations that have laid the groundwork and developed relationships with local land managers. In many cases it is much easier to get hired by a guide service than it is to obtain an individual permit.

Risk Management and Liability

In the context of climbing instruction, risk management is broadly defined as the identification of the risks that pose a threat to students, instructors, or a program, and to find ways to avoid or mitigate those risks.

A recommended practice for avoiding and mitigating risk is to systematically assess potential sources of harm, outline ways to avoid or reduce exposure to harm, and determine how to respond in the event of an incident or accident. When compiled into a document, a risk management plan is created. No matter the size of the program, from the very smallest to the very largest, the development of a written plan is essential. Additionally, everyone involved with the program should have a copy of the plan and be familiar with the contents.

An intelligent and well thought out risk management plan accomplishes the following:

- It can greatly reduce the number and severity of incidents or injuries.
- It protects the instructor and the program from potential negligence lawsuits.
- In the event of a negligence lawsuit, it can provide a layer of protection to the instructor and program during the defense against the suit.

An emergency response plan is typically a part of the overall risk management plan.

There are a number of considerations that should be addressed in the plan.

First-Aid Requirements

Climbing instructors must have appropriate first-aid training. The AMGA terrain standards state that Single Pitch Instructors are required to hold and maintain a Wilderness First Aid (WFA) certificate if they intend to work in the frontcountry, or a Wilderness First Responder (WFR) certificate if they intend to work in the backcountry. If it will take less than 2 hours to evacuate a student to definitive medical care, then an instructor is working in the frontcountry. However, if it will take more than 2 hours, then an instructor is working in the backcountry and will need to hold a WFR certification. Additionally, instructors must hold a CPR certification. Most American guide services require their instructors to maintain WFR and CPR certifications.

Participant Screening

Collect both physical and medical information on the trip's participants. Be sure to be aware of any conditions that might need to be considered before the trip, such as allergies, injuries, or medical conditions that could hinder a student's ability to participate. It is important to have this information in the field in case something happens, but it is equally important to protect each student's privacy in these matters.

Communications

If an incident occurs, it is imperative to have a communications plan. If the plan is to use a cell phone for this, be sure to know where there is reception and to have all appropriate phone numbers. If using a radio, a sat phone, or a personal locator beacon, understand how each of the devices work. Additionally, if there is a possibility that students will be required to operate one of these devices, they will need to be briefed on their use.

Evacuation

Instructors need to have a plan for an injured climber. Obviously there are levels of injury—some will only require an easy hike out, whereas others may involve a carryout or even an air evacuation. If evacuating the climber, an instructor should know where the nearest hospital is located. If calling for a rescue, have an understanding of what will happen and who will show up.

Minors

In the United States, by law, a minor child cannot enter into a contractual agreement. A release of liability is a contract. Additionally, some states allow legal guardians of minor children to sign away their right to sue; most do not. There are strategies for dealing with this dilemma, but they are beyond the scope of this book. The bottom line for the Single Pitch Instructor is to be aware of the protocols put in place by the organization one works for and to follow those protocols.

Student Safety Briefing

Before the trip starts, explain safety procedures to your students. Brief them on the communications plan and where the nearest hospital is located. If carpooling to the crag, it will be important for everyone to note where the keys are stored, whether in a pack, a pocket, or hidden near the vehicle. And lastly, the instructor should show the students where he or she keeps the first-aid kit and check to see who else in the group has first-aid training.

Post-Trip Contact

Establish somebody to check in with at the end of each program. If the instructor does not check in on time, that person needs to have instructions on what he or she should do to facilitate a rescue.

Incident Debrief

If something happens and a student or an instructor is hurt, or if there is a "close call," it's important to debrief the incident with the organization's management and with other instructors. Climbing instructors can learn a great deal from discussions following an incident.

Assumption of Risk and Liability Release Forms

Another important risk management tool is the use of release of liability or assumption of risk forms. When well crafted, these documents serve two purposes. First, they inform participants of the "inherent risks" that are or could be involved in rock climbing. Even the best risk management planning cannot eliminate all risk. Inherent risks are risks that are an integral part of an activity, which, if removed, completely alter the nature of the activity. By signing one of these documents, a participant is acknowledging that they have been made aware of the risks and that they are choosing to participate despite the risks.

Second, in the event of a participant's injury or death, these documents protect the instructor and program from charges of negligence. Simply put, negligence means that the instructor and the program have a "duty" to manage the "foreseeable" risk of harm to participants. A claim of negligence means that they failed in that duty, an incident occurred, and the participant suffered a loss. With a properly executed release of liability form, a participant is agreeing to give up the right to sue for negligence.

It is essential that all participants sign a release of liability or an assumption of risk form. Making sure that happens is often the responsibility of the climbing instructor. A recommended procedure for executing this process properly is to:

- Inform participants to read and, if they agree to the terms, sign and date the document.
- Allow the participants time to read the document. This process should not be rushed.

AMGA_i-258_3pp_CS55.indd 10 2/27/14 11:48 AM

- Collect the documents, making sure they have not been altered and are signed and dated. If the form uses boxes or lines after paragraphs, be sure they are checked or initialed.
- Store in a secure location.

If a participant will not sign the document as it is written, they cannot be allowed to participate in the activity. Also, it is not the responsibility of the instructor to interpret language in the document for participants. A well-written release should use language that a "reasonable" person can understand. If the same questions come up time and again, it is time to have the document reviewed by an attorney familiar with recreation law.

Liability

A question often asked by climbing instructors is "Can I be held personally liable and sued in the event of an accident?"

The bad news is that the answer is yes. If an accident involving serious injury or the death of a participant occurs, a lawsuit may follow. Along with the program, it is common that the instructor involved will be named as a defendant. The good news is that having general liability insurance will assist in the defense of the suit, and if a settlement is either reached between the parties or is handed down by the court, it will cover all or most of the settlement costs and, in some cases, but not all, attorney fees. That said, it is important to understand that not all liability insurance policies are created equally. These types of policies tend to be written in technical language that is difficult for the layperson to understand, consist of many pages, include exclusions, and don't make for the most exciting reading. Regardless, there are several questions a climbing instructor should ask:

- Are all employees covered? This may depend on whether the instructor is an employee or an independent contractor.
- Does it cover all activities? For example, some policies will cover rock climbing, but not ice climbing.
- Is it in effect for all areas of operation? The policy may not be in effect if the instructor or program is guiding without a permit.

For the climbing instructor, the name of the game is to reduce exposure to liability. Basically, this means conducting business at the "standard of a reasonable professional." This includes, but is not limited to, following a program's policies and procedures, being in step with current industry standards, using appropriate and well-maintained equipment, being respectful and building a good rapport with students, adhering to the rules and regulations of land management agencies, and getting training and, ideally, certification for the terrain worked on.

Even when following the best practices, there are a couple of subtle ways that climbing instructors can inadvertently increase their exposure to liability. The first is misrepresentation. An example of this would be an instructor who markets services as an "AMGA Guide," "trained and certified by the AMGA," or any of a number of other examples, but in reality is an AMGA Single Pitch Instructor. The problem here is that it would be pretty easy for a participant to make the case that when he purchased the services of an "AMGA Guide," he thought was hiring a fully certified IFMGA/AMGA Guide; in other words, he paid for one thing and got something else. In legal terms, this is referred to as "fraudulent inducement." If a court were to uphold the accusation, one likely outcome would be that all contracts between the instructor and the participant would be nullified; this includes the release of liability. Should that occur in a negligence suit, one of the most important tools for the defense is gone. The solution to this problem is easy: Instructors need to be specific about who they are and what they do when promoting their services.

The other subtle way to increase exposure to liability is to tell participants they will be kept safe. A legal definition of the word *safe* is "free from harm." As mentioned earlier, a "reasonable person" understands or has been informed that there are inherent risks in rock climbing. However, if marketing materials, operations manuals, a website, or the instructor tell participants that they will be kept safe and someone is injured or killed, any defense that might have been provided by the common law doctrine of inherent risk will have been severely compromised, if not lost. Again, the solution is fairly simple: Climbing is not safe, don't say that it is.

Risk Management in the Field

For AMGA Single Pitch Instructors, there are four types of risk that students might encounter on a program. They are psychological, sociological, financial, and physical. It's important to recognize each of these risks to students and to understand how to mitigate them.

Psychological

The fear of falling and the fear of heights are simultaneously attractive and repellent to students. Many new climbers are fascinated by the idea that they can climb to a point where a fall would result in their death, whereas others seek out climbing to combat a fear of heights.

There are two common ways to alleviate a student's fear of heights. The first is to spend a significant amount of time early in the day discussing "the system." In other words, explain the strengths of the ropes, the slings, and the other equipment. The idea is to assure students intellectually that the system will not fail. The problem is that an intellectual discussion is often not enough to alleviate the emotional nature of their fears.

The second way is through exposure therapy. The more often a climber is lowered on a climbing rope and caught by a belay, the more secure he will feel. Additionally, the more that a climber actually climbs, the more secure he will feel. For this reason, it's always a good idea to have a beginner climb a route more than once.

In extreme situations where a climber has a severe fear of heights, instructors might consider using incremental exposure therapy. In other words, have the student climb up 5 feet and then lower him. Then have the student climb up 8 feet and lower again. Every time the climber goes up, there will be an incremental decrease in fear. This is because the climber is becoming familiar with both the moves on the route and the security offered by the system.

Sociological

A second risk is fear of failure in front of other students, or in front of the instructor. Some may even be afraid of ridicule for their failure. It's not uncommon for students to apologize for their inability to complete a route. Some are embarrassed, and others believe that the fact they didn't get up somehow damaged the day for other students or the instructor. This, of course, couldn't be further from the truth.

Students need to be comfortable with you and with the others in your program. Instructors should strive to create a place where it is acceptable for their students to fail. At times this may mean moderating student comments to ensure that negativity is kept to a minimum.

Another excellent tool is for the students to see the instructor challenged on a route. Once a student has displayed competency with belaying, climb a route that is hard for you and let the students watch as you try to work out the moves. This will help them to understand that working a route, taking falls, and failing is part of the game.

Financial

Anytime a person spends money on a program, there is financial risk. A lot can go wrong that could make an individual feel like she wasted her money. These include the following:

A professional search and rescue deputy is short-hauled to an accident victim in Red Rock Canyon, Nevada.

JASON D. MARTIN

- Program goals were not met.
- Course was poorly or unprofessionally run.
- Student was placed in a course that didn't match her skill level.
- Personal equipment is lost or broken.

These four financial risks are easily averted. First and foremost, it's important to run a professional program. Instructors should meet and attempt to exceed all program goals.

Second, when screening someone for a program, ask probing questions to make sure that the student is placed in the proper course. Occasionally an instructor is placed in the awkward position of working with a student who has been assigned to the wrong program. A skilled instructor does her best to accommodate the person, and then makes sure that program administrators are made aware of the mistake.

And finally, be careful with student gear. Make sure that everyone leaves with the same equipment they arrived with.

Physical

The most obvious risk in rock climbing is that of an injury or a fatality. This is likely the number one risk that students think about when they sign up for a course.

The best instructors are those that are a little bit paranoid. Managing risk is paramount in climbing instruction, and an instructor who can imagine a wide array of dangers is an instructor who is likely to keep her students from injury.

There are a handful of things a climbing instructor can do to mitigate physical risk:

- Orient students to safety procedures around the crag. These include everything from checking knots to belay backups to climbing commands, etc.
- Create boundaries at the climbing site. Some crags are exposed, and there are places where students shouldn't be without a rope. Other crags may have excellent resting spots where students can relax away from the base without their helmets on. Make it clear to students what is appropriate at the crag.
- If working with other instructors, meet regularly to discuss risks and other programming issues. If working alone, set aside time to speak to other instructors and program administrators about their experiences.

Risk management and liability in adventure sports and adventure tourism is a massive subject. Seminars are held every year on this topic at the Association for Outdoor Recreation and Education (AORE) Conference, at the Wilderness Risk Management Conference, and at the Association for Experiential Education (AEE) Conference. Additionally, dozens of books, blogs, and websites are devoted to the subject.

Understanding liability is something that instructors—especially independent instructors—need to invest themselves in just as deeply as the rest of their training. Knowledge of liability issues is not the sexiest part of the climbing world, but can go a long way toward keeping an instructor in the field and out of the courtroom.

Stewardship

Because climbing instructors visit the same crags over and over again, they are in the unique position to see problems as they arise. Whether there are issues with graffiti, trail-braiding, litter, or something else, they are often the first to notice. Noticing is not enough. Climbing instructors should be advocates for the places where they work.

It is not uncommon for instructors to go to work and then go home, leaving the issues surrounding the crag at the crag. Perhaps they clean up some trash while they're in the field, but they don't make much of an effort beyond that. It is essential that instructors work to be more involved in stewardship, local climbing politics, and climbing advocacy.

AMGA Access director Scott Massey removes and replaces an old bolt on Rebel Without a Pause (5.11a) in Red Rock Canyon, Nevada. Prior to working for the AMGA, Scott guided in Red Rock, where he oversaw numerous crag-related stewardship projects.

JASON D. MARTIN

Many climbing areas have a local climbing stewardship organization. Sometimes these are connected to the Access Fund, and sometimes they're stand-alone organizations. Many advocacy groups focus on a very specific area, like the Friends of Indian Creek, the Las Vegas Climbers Liaison Council, or the Friends of Joshua Tree; whereas others have a much broader scope, like the Washington Climbers Coalition, the Northern Colorado Climbers Coalition, or the Southeastern Climbers Coalition. Regardless of an organization's purview, all of these organizations have one thing in common: They need volunteers.

If you don't have the time to volunteer for an advocacy group, make sure to participate in "Adopt-a-Crag"–style activity days. These usually involve litter cleanup and trail-building activities. They may also include re-bolting projects or the removal of invasive plant species. These events are excellent places to meet other local climbers and to give back to the place that provides you an opportunity to work.

Crag stewardship is not just good for the climber's soul; it's just the right thing to do.

Instructor Meetings

If working with another instructor, plan for informal meetings throughout the instructional day. This allows everyone to stay on the same page. Issues that come up during the day can include anything from safety concerns to instructional problems, from inclement weather strategies to dealing with a difficult student. In other words, these meetings cover the gamut.

If working for an organization or guide service, administrators may have formal meetings or debriefs after each course. If they don't actively provide a debrief session, then take the initiative and offer to meet with them. Most administrators appreciate it when new instructors come to them to talk about their programs.

Finally, if you have the opportunity to meet with the instructional staff on a weekly or monthly basis, take advantage of that. Instructors often work alone or with one or two other people. The more exposure you have to other instructors, their hardships and victories in the field, the better an instructor you will become.

CHAPTER 4

Pedagogy: The Art of Teaching

As climbing instructors, it's easy to get focused on the technical aspects of the job: A cam goes here, a rope goes there, a belay works like this, a lower works like that, etc. When we're with other instructors, we tend to talk about how to run a program at a given crag or how to deal with a certain technical situation. Indeed, our students come to us in part because we're experts in the technical aspects of climbing, but that's not the main reason why they climb with us.

They climb with us because they want to learn.

Musicians, actors, writers, painters, and photographers are artists. They spend years refining every aspect of their craft, becoming better as each day marches by. After decades of criticism and self-assessment, these people become masters.

Pedagogy is the art of teaching. And like the musician or the painter, one must dedicate time and energy to the art in order to become a master.

This chapter provides an introduction to the art of teaching as it pertains to climbing. But this is only an introduction, and it is recommended that aspiring instructors seek out other resources—both master-level instructor mentors as well as written materials on educational theory and technique—to continue moving personal development toward mastery.

Learning Styles

Effective instructors not only demonstrate a wealth of knowledge and expertise, but are also able to convey information about what they know in a way that others can readily comprehend, understand, remember, and then go out and use themselves.

When teaching a figure eight follow-through knot to a group of six novices, something like this typically happens: About half the group gets it right the first time, two people get it right after a few tries, and one person just can't get it right even after multiple tries. Why is this? People have differing learning styles, and it is usually not a question of intelligence, but rather a problem with the way the information is being presented that does not suit their particular learning style. Often, taking a different approach with the presentation of the information will help in comprehension and retention.

The theory of learning styles presented here is based on New Zealand educator Neil Fleming's VAK model. There are a number of different theories out there, but this one is a good introduction to the concept of how people learn.

The three main learning styles are visual, auditory (verbal), and kinesthetic (VAK). And although most people embody some combination of different learning styles, one is usually predominant.

Learning knots at Joshua Tree National Park, California.

Visual learners pick up information most effectively by reading the written word and viewing photos or illustrations. They need to see it to learn it. This group represents the highest percentage of the population (roughly 65 percent), and they learn most effectively from written communication. Visual learners think in pictures and words. They prefer their information presented in a written format as opposed to the spoken word, and usually prefer to take detailed notes when presented with verbal information. They can have difficulty with spoken directions. For example, a verbal description of how to get to some place is likely ineffective to them, but a written description along with

a map would prove useful. For this group charts, diagrams, graphs, and visual schematics are highly effective. When teaching a knot to this group, a verbal description is of little help, but presenting the sequence of tying the knot like a schematic diagram, with step 1, step 2, and so forth, may prove more successful.

Auditory, or verbal, learners represent another large segment of the population (around 30 percent) and learn most effectively from the spoken word. They need to hear it to learn it. This group is composed of the most sophisticated speakers, and they relate most effectively to spoken words. They are good listeners and will listen attentively to a verbal presentation and take notes afterward. They benefit from hearing lectures and participating in group discussions. They prefer their information presented in a verbal format, and for them information is not tangible unless it is spoken. They can have difficulty with written directions.

Another, much smaller, percentage of people are kinesthetic learners, who learn by doing, by getting the feel and tactile sensation of performing the steps required to learn a task. They are usually gifted athletes and can learn a knot just by seeing it tied, then tying it once or twice. Kinesthetic learners prefer hands-on learning and can appear slow if information is not presented to them in a style that suits their learning method.

Most people learn most effectively by watching, doing, and reflecting on a particular technique. Around 450 BC the great Chinese philosopher Confucius said: "Tell me and I will forget; show me and I may remember; involve me and I will understand." Various research studies point to the fact that people remember best by watching a demonstration, having a discussion, practicing the skill, then teaching it to others. Military trainers use the EDIP principle (explain, demonstrate, imitate, practice).

When teaching climbing techniques, the climbing instructor should present the topic in a way that most broadly targets all the differing learning styles. The approach can be simple:

1. Tell (auditory or verbal)
2. Show (visual)
3. Do (kinesthetic)

Using this approach, here is an example of how to teach an abstract concept, such as a complicated knot:

1. Tell. Verbally describe the steps to tie the knot.
2. Show. Demonstrate the knot. Facing in the same direction as the person being taught is helpful.
3. Do. Have the student practice tying the knot.

This is the basic approach, although there are additional teaching methods that will help in memory retention and learning, namely imagery and mnemonic devices. An example of using imagery when teaching the bowline knot: "The snake comes up through the hole, around the tree, and back down through the hole." Attaching a verbal image of an animal or thing is one of the most effective ways to remember an abstract concept.

Mnemonic devices are another form of memory aid. The most common one is the acronym. The military is big on acronyms, like IED (improvised explosive device), SOTG (Special Operations Training Group), etc. In teaching climbing, some common acronyms are ABC (anchor, belayer, climber) or BARCK (buckle, anchor, rope, carabiner, knot).

To most effectively teach climbing techniques, it's best to start with an approach that presents the information for a broad range of learning styles, then try to identify each person's particular learning style in order to aim directly at his or her individual style. If someone is not getting the information, change the approach and present it another way. Use imagery and mnemonics to help students remember key principles (like safety checks or anchoring fundamentals), and chances are they will retain these key concepts throughout their climbing lifetime.

Lesson Planning

Many climbing instructors are natural born teachers. But even the best teachers do a better job when they have a structure to work within. Creating concrete unit and lesson plans provides that structure.

Remember when you were in high school and there was a unit on the Civil War or a unit on *Romeo and Juliet*? And remember when during that unit, you studied the battle of Gettysburg or Shakespearean language? These were individual lessons inside of a larger learning arc, an arc that is referred to as a unit.

In rock climbing instruction a unit could cover something as vast as anchor building, while an individual lesson within that unit might cover something like placing a cam. In other words, a lesson is a bite-size piece of a unit, and a series of lessons together combine to provide the knowledge required to complete a unit.

Student-Learning Objective—This is the overriding goal of the lesson or unit, what students should be able to show after they have received instruction.

Examples of student-learning objectives for units and individual lessons include the following:

Unit Objectives

- Students will demonstrate an ability to pass a climbing gym test.
- Students will demonstrate an ability to build an anchor.
- Students will demonstrate an ability to set up a toprope.

Lesson Objectives

- Students will demonstrate an ability to tie a figure eight follow-through.
- Students will demonstrate an ability to belay effectively.
- Students will demonstrate an ability to use appropriate climbing commands.

Please note that each of the preceding sentences starts with "students will demonstrate an ability to . . ." It's incredibly important for every learning objective to start with this, because it provides a concrete goal that can be measured.

In a lesson plan, start with the student-learning objective, work into the specifics, and then culminate in an evaluation. There are many examples of lesson plans out there, but they all start with a learning objective and finish with an evaluation. Following is a simple outline of a lesson plan and how each part of the lesson functions:

Student-Learning Objective—"Students will demonstrate an ability to . . ."

1. **Set**—Preparation for what the students are about to learn. It could be a complex and creative opening, or a very simple, "this is what you're going to learn today."
2. **Input**—The actual teaching of the lesson.
3. **Guided Practice**—Working with the students until they achieve mastery.
4. **Evaluation**—Evaluating whether the student has learned the requisite material.

Many instructors do a very good job of identifying a learning objective, setting up a lesson, teaching it, and practicing a skill, but often don't evaluate students when they're done. Evaluation is just as important as any other part of the lesson.

Following is a sample plan for a knot-tying lesson:

Student-Learning Objective—Students will demonstrate an ability to tie a figure eight follow-through.

1. **Set**—Explain to students that this is one of the most important knots they will learn. They will need this knot to tie into the rope, and if they want to join a climbing gym, they will be tested on it.

2. **Input**—Demonstrate how to tie the knot. With certain populations you might choose to use a mnemonic (a memory trick) to help the students remember.

Single Pitch Instructor Sample Lesson Plan

Lesson Title: **Timeframe:** Overall timeframe for the lesson:

Student-Learning Objective: Students will demonstrate an ability to . . .

Audience: Brief description of the audience including age, skill level, limitation, etc.

Prerequisites: List any prerequisite training or skills.

Lesson Location and Special Notes: Identify where the lesson will take place. Have a back-up plan in place if the location is being used. Add reminders such as equipment needs, classroom prep, and "don't forget" notes.

Time	**Materials**
____min	**Set:** Prep the students for what they are going to learn.
____min	**Input:** Provide the content of the lesson.
____min	**Guided Practice:** Work with the students to achieve mastery.
____min	**Evaluation:** Test the students on their mastery of the subject.

Does the lesson reach multiple learning styles? **Yes** **No**

Instructors new to lesson plans should photocopy this sample plan and fill it out for each lesson they intend to teach.

3. **Guided Practice**—Practice the knot with the students.

4. **Evaluation**—Have all the students show that they know how to tie the knot.

Many less effective climbing instructors don't take the time to build lesson plans. The result is that they learn how to teach through an organic process of trial and error, which is terrible for the students who are forced to work with an instructor on an error day.

A review of the day with input from the students provides valuable feedback from their perspective of how they perceived an instructor's presentation of the topics. This allows the instructor to improve and modify the approach for the next time. Getting feedback from students is key to becoming a better instructor.

The best teachers prepare lesson plans on paper ahead of time and then continue to tweak them over a period of months and years. They use a structured approach through student feedback and self-assessment and keep track of what works and what doesn't work. They make note of all student-learning objectives and do their best to make sure that their students achieve those objectives.

Master teachers don't just appear out of the woodwork. They work hard every day to be the best.

Harsh Lessons

In my (BG) career as a rock climbing instructor, I've received training and undergone testing and evaluation of my guiding skills. The toughest lessons, but the most memorable ones, came from the mistakes I made that were critiqued by my mentors and peers. I've also found, in my role as an examiner for the AMGA Single Pitch Instructor exam, that candidates learn more and remember best from their mistakes when it really counts, when they're being scored in a pass or fail test situation. The harshest lessons in life are the ones we don't soon forget.

Learning from Mistakes

Asking a student with some experience to perform a task can tell an instructor a great deal about the student and where to begin. For example, one quick test that can be given to new students is to simply hand them the end of the rope with the instructions: "Go ahead and tie into the rope." The way that they accomplish this easy task tells more about where they're at than even a brief interview.

Another example: When teaching anchoring fundamentals to a group class in a ground-school setting, present a concept with a verbal presentation, a demonstration, and then a hands-on practice (tell-show-do). Give each of the students a scenario (e.g., this is the edge, here are the crack systems you can work with, build your anchor system so that your master point is here, etc.), then let the students build an anchor system to their satisfaction, without constant coaching by the instructor along every step in the process. By letting them finish without instructor input, they are likely to make some mistakes. Then, as a group, go around and critique each anchor system setup. In this way everyone learns from everyone's mistakes. The big plus is that it gives the instructor valuable and instant feedback on whether or not the students have grasped the concepts just presented and shows what they (or we) need to work on to achieve our objective.

This can be a very effective strategy when teaching a lesson to one or two students who have some modicum of anchoring experience but have asked for a refresher course. Start by giving them

Anchor rigging session during AMGA Single Pitch Instructor course at Joshua Tree National Park, California.

a specific scenario, like rigging a toprope anchor, and then let them rig it without any coaching. What is seen in the final product of their anchor system helps the instructor decide if he or she needs to start with more basic concepts, or can simply move on and build from a solid foundation to more advanced topics. At this point the instructor can prepare a lesson plan specifically targeted at the students' skill level and build upon their current foundation to reach the objective for the day.

Teachable Moments

A less formal way to present information is to explain and describe the technique as it comes up during the course of the outing or class. These are often referred to as teachable moments. For example, when setting up a climbing situation for students, explain to them as they are tying in to the rope the importance of closing the system by tying a stopper knot in the other end of the rope. If they are on a slab climb and their heels come up, tightening their leg muscles, explaining the correct technique of relaxing the calf muscles and dropping the heels down is a teachable moment. Try to keep it positive. Many teachable moments occur when a simple mistake has been made, and addressing it at that very moment is a great way to teach someone the correct way to do something. The trick is to do it in a positive way that encourages the student to try something new and instantly improve or react positively to a potentially negative situation. Often, the ideal time to address a question is at the moment a particular event that illustrates a key point is occurring. This adds valuable context to the concept.

Taking Kids Climbing

Many kids are natural climbers with an instinctive curiosity and tendency to want to climb things. If you take kids to a rock climbing environment, they'll spontaneously want to explore, climb boulders, and scramble around on the rocks. The obvious danger, however, is that in a fall or slip they can get seriously scraped up, or worse.

Climbing develops muscular strength, balance, and kinesthetic awareness and introduces kids to valuable concepts like problem solving, teamwork, and the success that comes from perseverance and determination.

How Young Is Too Young?

I (BG) have seen, over the years, cases of precocious children with great balance (as young as age 3) who were physically talented enough to climb and be lowered on a toprope.

Toproping is the best method to use to take kids climbing. By the time they are in elementary school, most kids will have reached the level of physical and emotional development to allow them to enjoyably experience climbing in a toprope setting. Nevertheless, there are definitely challenges among this population. It can be difficult to sustain the attention of small children for any sort of sustained lesson. Try to reserve lengthy lessons for those children who seem to show an active interest in acquiring new skills.

It can be very difficult to initiate a lower on a small child. Many of them are so small, the friction of the toprope system is almost as great as the child is heavy. Low-friction belay devices, skinny ropes, and smooth, round-stocked carabiners at the master point can help, but in some cases, the other end of the climbing rope can be used to help retrieve these small climbers.

Selecting climbing equipment for small children can also be challenging. Many small children do not have well-defined hips, so a standard sit harness would be inappropriate for them. There are a number of full body harnesses designed for children, and these should be seriously considered if programming for children becomes frequent. Similarly, small climbing shoes are not always the first thing an instructor considers when selecting shoes for a fleet, but climbing shoes are just as helpful for a 6-year-old as they are for a 60-year-old. Climbing helmets are also just as important for children as for adults, so having those sizes available will be essential.

Route selection is also a challenge when designing programming for children. Because they are so much shorter than even the shortest adults, they will often use drastically different movement strategies than adults. Smaller holds are more realistic options for smaller hands. Be careful, therefore, of feature-oriented climbs, like face climbs. Sometimes the distances between holds are just unrealistic. Instead, try to find climbs that have a homogenous character for the entire pitch, like crack climbs, chimneys, or slabs. A climb with a uniform character can be climbed by people of all sizes. Set up a short, easy climb on either a low-angle slab or a short wall with abundant hand- and footholds. Teach them the basic climbing signals. For their first time, have them climb up a short distance (say 10 feet above the ground), then go through the signals and have them lean back and weight the rope, get into the lowering position, and get lowered back down to the ground. If a child does not possess the requisite balance to successfully be lowered down the climb, or if they are so small they will require some sort of retrieval system, it is better to know that *before* they reach the top of a climb. Take on a climb in small increments, and don't allow them

AMGA Single Pitch Instructor Tim Page assists as a child lowers. If a child is too light to lower, you may be able to attach the other end of the rope to his harness and pull down to provide more weight.

JEREMY WILSON

What Makes a Good Teacher?

Just because you can run 100 meters in 9.7 seconds doesn't necessarily mean you'll be a great track-and-field coach, but it helps. If you have great talent and are able to effectively communicate, you have a winning combination. But I (BG) have seen many world-class climbers who were poor teachers, and many competent but not exceptional climbers who were excellent teachers. What I always look for in my climbing school are instructors who have both traits: talented climbers and patient teachers. Climbers who have mastered all the various climbing techniques at a high level can ultimately be the best teachers, as their mechanics and fundamentals are so sound that they have the innate ability to easily demonstrate key concepts. But they also must be effective communicators and have enthusiasm, patience, empathy, and a desire to share their knowledge in a way that truly benefits their students.

to climb too high above the ground if they're not ready for it. You want the experience to be fun, not terrifying. If they're not yet able to be lowered down, they can swing around on the rope and play on the rock a little bit. There will be a next time. Just familiarizing them with the rope, harness, and equipment can be a good start.

Teaching Climbing Movement

Perhaps the greatest teaching challenge instructors will face involves actually showing people how to use their bodies on the rock, how to actually climb things. The magnitude of the subject is unfathomable. If an instructor looks at a single climb, for example, how many different moves are on that climb? Are there patterns? Are there consistent strategies? The variability of the moves on a single climb is of course compounded by the variety of bodies that will attempt to climb it. This person is tall, this person is short, this person has great upper body strength, this person has none. Even if the bodies were identical, this person is aggressive, that person is contemplative, this person is fearful, that person climbs with reckless abandon, and so on. Of course, instructors don't get to teach one climb; they are trying to teach dozens of climbs, on dozens of crags, with dozens of rock types, in dozens of climates. It is an immensely dense subject.

This text cannot reasonably cover a topic that large, and it is probably folly for an instructor to attempt to do so as well. Instead, try to focus on what is absolutely essential for a successful experience. When teaching beginners, instructors should first focus on the fundamental difficulties all beginners have: It is difficult to make a bipedal creature move in a quadrupedal manner. Second, teach the specific moves that your students will need to climb in a given area.

Bipeds Becoming Quadrupeds

Most mammals move through terrain on all four limbs, and the coordination of those four limbs does not change when they begin to climb trees, mountainsides, or even rock cliffs. Humans have evolved to move through the world on their hind legs, leaving the forelimbs free to use tools, care for children, etc. It is a strange and counterintuitive proposition to tell a bipedal brain that it must forget its evolutionary programming and move like a quadruped. That is the fundamental difficulty that all beginners have. Even though bipedal locomotion has conditioned them to shift their center of gravity from one foot to the next as they move through terrain, the presence of handholds—the proposition of quadrupedal locomotion—short-circuits their brains. They consistently fail to transfer their center of gravity from one foot to the next when they are holding something in their hands. That is why beginners get pumped when they climb 5.5. They are not necessarily weak, but their upper bodies must exert enormous energy to execute a skill that is normally reserved for their legs.

Unfortunately, knowing and understanding this predicament, even carefully explaining it to a student, doesn't make the learning any less difficult. But, if the predicament is carefully explained to the student, the subsequent coaching will be more precise, more focused on the real problem. Instead of shouting useless and ambiguous feedback to a student, "Trust your feet! Trust your feet! Just stand up!" an instructor can be more precise, "Shift your center of gravity from your low step to your high step!"

Often experienced climbers do not understand how perfectly unconscious their understanding of this fundamental problem is. Their bipedal brains have fully adjusted to quadrupedal locomotion. Climbing instructors need to help their students catch up.

Terrain-Specific Movement

Most likely there will be moves that are specific to the terrain, and instructors should concentrate their coaching on those specific skills. It will be useless to tell students about nuanced and esoteric concepts like flagging or heel hooking if the terrain they are climbing is low angled. There is no need to teach crack climbing if there are no cracks, and so on. Try to find the three or four techniques that are consistent in the course area and teach those skills. On a low-angled climb, for example, students will need to know how to smear. But backstepping and edging are probably extraneous pieces of information. Students will need to focus attention on dynamic weight transfers from a low step to a high step, but they will not need to toe hook or do foot switches. Students will need to mantle, but they will never need to dyno. Be specific; find the movement patterns at the crag and teach those skills.

CHAPTER 5

Baseline Equipment

An office worker can't do anything without a functional computer. A carpenter can't build a house without a hammer. A taxi driver is useless without a working taxi. And a climbing instructor is not a climbing instructor without certain essential tools. An instructor's baseline equipment is composed of the instruments crucial to a successful day of climbing instruction.

Following is a breakdown of the essential baseline equipment required by a professional climbing instructor.

Harness

The modern climbing harness is not only more comfortable than the swami belts of climbing's golden era, but they also are manufactured and sold with incredible variety. A climbing instructor selects two types of harnesses—the instructor harness and the participant harness—from a wide array of options for two essential purposes. In the course of a climbing outing, an instructor will need to use the harness to rack lead protection, carry an array of assistance and anchoring tools, tie into a rope (or two ropes) while also using the belay loop, haul a rope, rappel on one device and transfer to another, and so on. A participant, on the other hand, may never need to carry an extraordinary amount of equipment, may never do any sort of technical multitasking, and may only actually sit in the harness while being lowered. As a result of these two fundamentally different harness applications, a climbing instructor should deploy two different sets of criteria when selecting the best harness for the job. In both cases a harness with a waist belt, leg loops, and a belay loop will be the most comfortable and useful.

When selecting an instructor harness, look for a model that has

- a separate space for belaying/clipping to the harness (belay loop) and tying-in (the hard points of the harness)

Swami Days

I (BG) began climbing before the advent of the modern climbing harness. Instead of a harness, I used a swami belt. I took a 20-foot length of 2-inch-wide nylon webbing (rated at 8,000 lbs.), wrapped it a bunch of times around my waist, and tied it with a water knot. Then I tied my climbing rope around the swami belt with a figure eight follow-through. This setup discouraged **hangdogging** and was rib jarring in a leader fall! Soon we figured out how to add leg loops, which made the rig more comfortable, to say the least!

- gear loops to hold all tools and a rack
- adjustable leg loops (if you'll be climbing in different seasons, with different amounts of clothing)
- comfortable for all-day use

For a participant harness, look for a model that has

- durability to sustain repeated novice use
- simple functionality to expedite instruction and supervision
- adjustability to fit a large of number of body types and sizes

Modern, lightweight climbing harness.

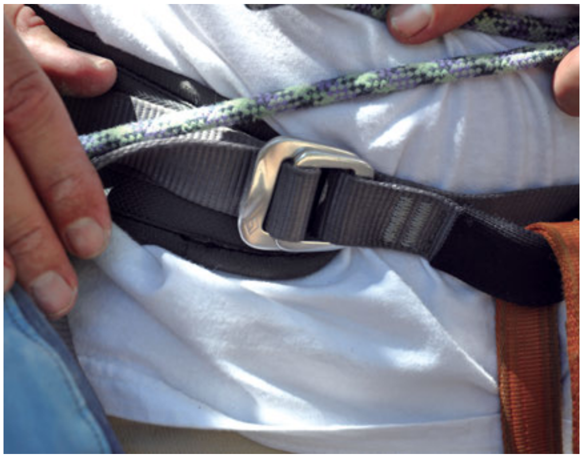

Modern speed buckle.

Top brands include Arc'teryx, Black Diamond, Petzl, Metolius, Trango, Mammut, Misty Mountain, and Camp USA.

Some newer models have webbing with "speed buckles" that are pre-threaded and already doubled back. Other harnesses have the old-school "double pass" buckle, where the webbing belt must be doubled back through the buckle. The instructor should be able to demonstrate how to properly buckle and use each student's harness.

For students, fully adjustable models are convenient and efficient, especially for outfitting groups. Black Diamond, for example, offers a fully adjustable model with speed buckles for sizes extra small through medium, and another model for large to extra large.

Harnesses should be retired when the belay loop becomes frayed or shows signs of wear. For recreational climbers, Petzl recommends using nylon products no longer than seven years, even with minimal use. In an institutional setting, harnesses are retired sooner than that.

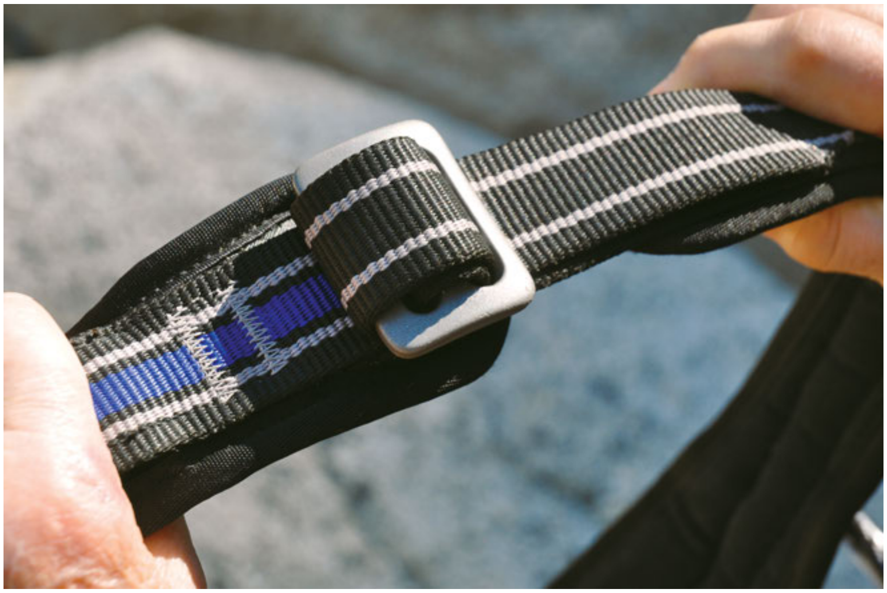

Traditional, doubled-back buckle.

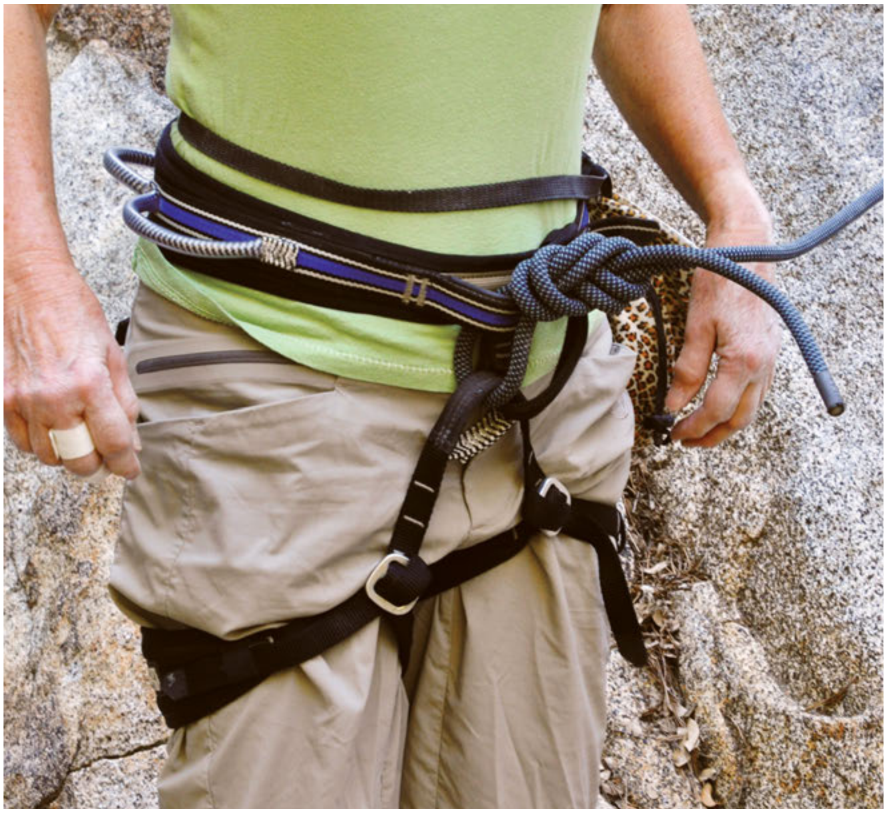

Harness with adjustable leg loops.

Helmets

Back in the 1960s and 1970s, very few climbers (probably less than 10 percent) wore helmets. Today it's the opposite. Most climbers wear helmets, particularly in trad climbing areas. The professional standard is for all guides and instructors to wear a helmet and to enforce a helmet policy for their students.

Even in the seemingly benign single pitch environment, rockfall can be a real hazard, especially if there is loose rock at the top of the cliff. Rockfall can occur naturally (from wind or melting snow or ice) but more commonly is caused by other climbers. The most common scenarios are a climber pulling off a handhold or breaking off a foothold while climbing, inadvertently kicking off a rock from the top of the cliff, or causing a rockfall by manipulation of the rope. Be especially aware when other climbers are at the top, rigging anchors and setting ropes, as rocks are easily dislodged by ropes being pulled around. When teaching groups, establish a mandatory helmet zone at the base of the cliff, and then watch for students who are lounging at the base without a helmet.

Students should be taught that while on the cliff or at the top of the cliff, if a rock (or any object) is dislodged, the universal signal is "ROCK!" How loud the signal is yelled can signify the size of the rock. If the warning comes from 100 feet or more away, students can be taught to look up, judge

Helmets help manage risk in the single pitch environment.

the trajectory, and move aside accordingly. If the signal comes from just above, students should be instructed to hunker down and not look up, so as not to get hit in the face. Obviously, wearing a helmet is most important when climbing, belaying from the base of the climb, and hanging out at the bottom of the cliff.

Be particularly alert for rockfall when rappelling, as it can be caused by the rope coming into contact with the cliff, particularly if the climber swings from side to side and especially if there is loose rock on the cliff.

Top brands of climbing helmets include Petzl, Black Diamond, Mammut, and Camp USA.

Ropes

History

Mountaineering's first golden age ended abruptly on July 14, 1865. For Edward Whymper, who led a team of seven climbers up the first ascent of Switzerland's Matterhorn—at the time the last great unclimbed peak in the Alps—both triumph and tragedy were his fate on that great and dreadful summer's day. After reaching the summit the climbers, all tied together with manila ropes, began a tedious descent. One of the party suddenly slipped, pulling several of the others off. While Whymper held fast, clinging to a rock outcrop to avoid being yanked into the abyss, the rope broke, sending four climbers to their death. Later buried in the Zermatt cemetery, the Matterhorn soars above them as their tombstone for all time. Some speculated that the rope had been cut with a knife, but a formal investigation and inspection of the rope revealed the cord had simply parted—broken under the strain.

Through the late nineteenth and early twentieth century, climbing and rappelling ropes were made from natural fibers like hemp, manila, and sisal. While strong enough for many applications, these ropes were ill-suited to the rigors of climbing and rappelling, and subject to mildew and rot.

Nylon, also known by the generic name polyamide, was developed by the Dupont company in 1935. Modern nylon climbing and rappelling ropes that could actually hold up to the forces generated from leader falls were first manufactured in the 1940s, coinciding with the availability of high-quality "nylon 6," allowing the construction of lighter weight ropes that could stretch to absorb great forces and hold more than 2 tons. Nylon is still by far the best material for rock climbing and mountaineering ropes, due to its ability to stretch and absorb the forces created in a fall, making it superior to any other yarns currently available for this application.

Polyester, patented in 1941, has less stretch than nylon and is today widely used in the manufacture of low-stretch ropes used in rescue, fixed lines, rope systems, and rappelling, where dynamic properties are not required.

Early nylon climbing ropes were a three-strand, woven rope with the brand name Goldline. The three-strand twisted construction (called laid construction) consisted of three strands of twisted filaments spiraled into one singular rope. These ropes stretched considerably under body weight, and if you were rappelling or prusiking without contact to the rock, the spiral construction would result in a dizzying spin on your way down.

Kernmantle Ropes

The big step up in rope technology came in 1953 with the advent of the first kernmantle rope, made in Europe. The kernmantle design consists of an outer, tightly woven sheath (mantle) over a core of twisted, parallel fibers (kern). The core of the rope provides most of the rope's strength, and the sheath protects the core from abrasion and damage and reduces friction as the rope runs through carabiners and rappel devices. The kernmantle rope handles better and is more durable than a rope with laid construction.

During the 1950s and into the golden age of big wall climbing in Yosemite during the 1960s,

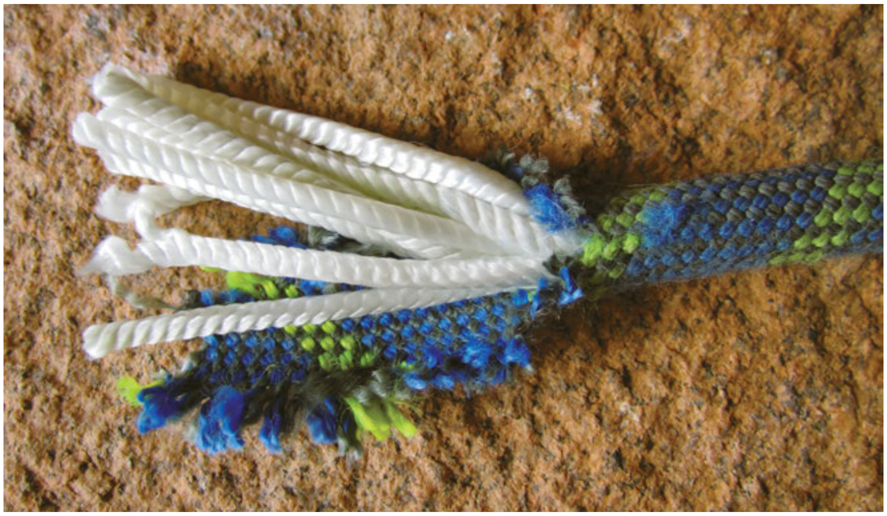

Deconstructed kernmantle rope showing the braided white interior core under the more tightly woven sheath.

MAXIM
MAXIM
10.1

Selection of dynamic ropes from Nomad Ventures climbing shop, Idyllwild, California.

Dynamic Rope Types

Dynamic ropes are sold in single, double, and twin configurations.

A **single rope** (marked on the label with a 1 inside a circle) is the most commonly used rope for rock climbing and rappelling and is used as a single strand.

A **half rope** (marked on the label with a ½ inside a circle) is used primarily for alpine rock climbing and mountaineering and consists of a pair of ropes both tied in to the leader, who alternates clipping one strand at a time into protection. This configuration reduces rope drag through protection and provides two full-length ropes to facilitate a rappel descent.

A **twin rope** (marked with two overlapping circles) is used primarily in alpine climbing and should only be used with the two strands together; each climber ties into both strands and both strands are clipped into points of protection. The big advantage is that it allows for retrievable rappels the length of the rope, and it is lighter than a half rope. A single strand of a twin should never be used alone for any purpose.

Goldline ropes were still the standard, as they were about a third of the price of kernmantle ropes. Once they became more widely manufactured and distributed, with a correspondingly better price, kernmantle ropes became the standard climbing rope worldwide by the late 1960s and early 1970s.

Today the kernmantle rope is the standard design used in climbing, rappelling, caving, canyoneering, vertical rope access, fixed lines, rescue, and life safety ropes. These ropes are made from nylon, polyester, or a combination of these and other synthetic yarns. There are three basic types of kernmantle ropes: dynamic, low-stretch, and static. A dynamic rope is required for situations where a high-impact force can be generated—such as in a leader fall—and therefore is the standard rope for rock climbing, ice climbing, and mountaineering, where the rope will be used for belayed climbing and rappelling. A dynamic rope will typically stretch about 26 to 36 percent during a leader fall (dynamic elongation) and around 7 to 11 percent under body weight (static elongation).

UIAA and CE Certification for Dynamic Ropes

Dynamic ropes sold for rappelling, climbing, and mountaineering are tested and certified to UIAA (Union Internationale des Associations d'Alpinisme) and CE (Conformité Européenne) standards and should bear the UIAA or CE certification on the label. This means that the rope has been tested and certified by a "third party" at one of three UIAA-approved testing facilities to meet the European Norm (EN) 892 and UIAA-101 standard for dynamic rope. To receive this certification, the rope sample must survive at least five UIAA drop tests. This test is done by taking an 80 kg (176 lb.) weight, attaching it to one end of a 9-foot length of rope, then raising it 8½ feet above the anchor and dropping it 15 feet over a 10mm diameter bar (similar to a carabiner) that is anchored approximately 1 foot above where the rope end is anchored. This simulates a fall factor of 1.7 (total distance of the fall divided by the length of rope in the system), which is a very severe fall in climbing situations.

For a single rope, during the first drop, the peak impact force on the rope is measured and must be less than or equal to 12 kN (8 kN for a half rope with a weight of 55 kg) with a corresponding stretch of less than or equal to 40 percent. Twin ropes have the same requirements as a single rope but must survive at least twelve drops without breakage. In addition to the stringent drop test requirements, sheath slippage can be no more than 1 percent, and static elongation under an 80 kg (176 lb.) load can be no more than 10 percent for a single rope and no more than 12 percent for a half or twin rope.

To the best of our knowledge, the only documented rope failures among UIAA-certified ropes were ropes that were cut over sharp edges and one rope with known pre-exposure to sulfuric acid.

Static and Low-Stretch Ropes

For many years the term *static rope* was used to define any low-stretch rope typically used as a rescue rope or as a fixed line for rope access, rappelling, and life safety. As rope manufacturers developed new techniques and technologies to create better low-stretch ropes, the terms *static rope, low-stretch rope, low-elongation rope,* and *semi-static*

STERLING
ROPE

Climbing shops sell static and low-stretch ropes from spools, cut to the length you desire.

rope, often used interchangeably, became somewhat ambiguous, especially since all ropes have some stretch, so a more precise definition was needed.

The Cordage Institute, an international rope industry association that disseminates industry standards, defines two categories of non-dynamic ropes: static and low-stretch.

STATIC ROPES

Static is defined as rope with a maximum elongation of less than 6 percent at 10 percent of the rope's minimum breaking strength (MBS). New technology allows manufacturers to create static ropes with virtually no stretch—think of them almost like a wire cable. For example, the Sterling Rope company's ½-inch diameter HTP Static rope stretches only 0.8 percent with a 300-pound load and has a safe working load (SWL) of 908 pounds, which is 1⁄10 of its MBS of 9,081 pounds.

These ropes should not be used for toprope belaying, but can be used for anchor rigging and fixed lines, although low-stretch is also preferable for those applications since even a slight amount of stretch means a lower impact force on the anchors.

Static ropes generally have poor handling characteristics due to their stiffness and are typically used only for haul lines, high lines, zip lines, and rappelling, where dynamic properties are not required. A static rope should never be used for lead climbing or for belaying, where it may be subject to a high-impact force.

LOW-STRETCH ROPES

The Cordage Institute defines low-stretch as a rope with an elongation between 6 and 10 percent at 10 percent of the rope's MBS. Since a low-stretch rope has relatively little stretch (usually 3 to 4 percent under body weight) compared to a dynamic rope, it is a great choice if you're using it just for toproping and rappelling, but it should not be used for leading. Low-stretch ropes may also be a good choice for rigging anchor extensions and fixed lines because they are easy to tie knots with, and they reduce potential impact forces when used as a tether or belay.

Dynamic versus Low-Stretch

When buying ropes, the following criteria will help the instructor assess which rope is best for a particular application.

A dynamic rope is required for situations where you'll be using the rope for leading or belaying, where any impact force is anticipated. A low-stretch rope should not be used for lead climbing, because rope stretch is the key to absorbing the energy generated in a leader fall.

For a program involving just toproping, using a low-stretch rope makes a lot of sense, since dynamic rope stretch is often a hazard, or at least a nuisance, in a toprope situation. Picture a 100-foot-high cliff set up as a toprope. At the start of the climb, you'll have 200 feet of rope in the system. Let's say you're belaying a climber using a dynamic rope, with a bit of slack, and the climber falls 15 feet up the route. Let's assume the rope has a static elongation of 9 percent. Do the math: 185 feet times 9 percent equals 16.7 feet. Chances are your climber will find herself on a 16.7-foot bungee cord that eventually returns her to the ground (the thing she is most worried about impacting).

Diameter and Sheath Percentage

Ropes sold for climbing and rappelling commonly range from 9 to 11mm; 10mm is a very popular diameter among climbing instructors, as it's durable yet light and supple. Recently there has been a trend toward ever-thinner ropes. Ropes as thin as 9.4mm in diameter are UIAA rated for lead climbing. But these thinner ropes are generally stretchier and will wear out far more quickly. For toproping, selecting a rope that is thinner than 10mm in diameter often results in having to purchase new ropes more frequently.

Low-Stretch Ropes

In my (BG) climbing school program, we've used low-stretch ropes for our toproping classes for more than twenty-five years. These low-stretch ropes stretch about 3 to 4 percent under body weight, slightly more in a toprope fall with a bit of slack. From a risk management standpoint, it just makes sense to use a low-stretch rope for toprope situations, especially if you're setting up relatively long routes (up to 100 feet high). I look for a rope with either a CE EN 1891 Type A certification or a National Fire Protection Agency (NFPA) light use certification, with a diameter of between 10 and 10.5mm, to be compatible with various belaying and rappelling devices. I also check the rope's suppleness to ensure the rope will hold knots firmly and handle well. Sterling makes an excellent low-stretch polyamide (nylon) rope called the Safety Pro. I prefer the 10.5mm diameter for good handling and durability for toproping and rappelling.

If using a dynamic rope for toprope belaying, beware of the dangers of rope stretch, and keep the rope taut in situations where the climber is just off the ground or just above a ledge.

Thinner ropes generally stretch more and cut more easily over sharp edges. A thicker diameter rope also affords more friction when used with rappelling devices.

Many manufacturers now provide information on the sheath's percentage of the total weight. A rope with a higher percentage of sheath (40 percent or more) will generally be more durable for rappelling use than one with a lower sheath percentage.

Sharp Edge Resistance

The UIAA has developed a new, optional test for manufacturers to receive a "sharp edge resistant" certification. The test is very similar to the UIAA drop test, although instead of dropping the rope over a rounded bar simulating a carabiner, a sharp edge is used. This is a pass or fail test, and although not a true indicator of durability, it's a good measure of the rope's ability to resist slicing over sharp rock edges. As mentioned earlier, some of the only documented cases of modern climbing ropes breaking in the field have been when the rope was cut over a sharp edge.

Static Elongation

This is a measure of how much the rope stretches under a weight of 80 kg (176 lbs.), telling you how much your rope will stretch during a normal rappel or when someone hangs on the rope. For toproping, rappelling, and fixed rope applications, some instructors prefer a low-stretch rope with a static elongation of around 3 to 4 percent. Most dynamic climbing ropes have a static elongation of about 8 or 9 percent.

Dry or Non-Dry

Nylon, when it gets wet, absorbs water, which weakens the fibers. Nylon ropes can lose much of their strength when wet (at least 30 percent; some studies show over 50 percent loss of strength), so manufacturers sell ropes with a "dry coating" to keep the rope from absorbing water and make it more abrasion resistant. If using a dynamic nylon rope in snow, ice, or wet conditions, a rope with a dry coating is recommended.

Marking the Middle of Your Rope

Many climbers use a black felt tip marking pen to mark the midpoint of their rope. In 2002 the UIAA Safety Commission issued a warning based on testing done by the UIAA and by some rope manufacturers that showed the ink from some marking pens decreased the strength (more specifically, the rope's ability to hold repeated falls in accordance with the EN 892 testing standard) by as much as 50 percent. While this may seem to be a shocking figure, the UIAA president pointed out that "such a marked rope can only break in practice when the 2 or 3 centimeters, which are marked, are placed over a sharp rock edge when the rope is loaded by a fall." While this is a very remote possibility, an instructor should consider other alternatives to identify the midpoint on a rope, or at least use only marking pens sold or recommended by the manufacturer of your rope. Tape is not a good option, as it can slide on the rope or, more likely, become gummy and stick in rappel and belay devices. A good option is to use a "bipattern" rope, which is a rope that changes pattern at the middle without a change in yarns or color. Another option is a "bicolor" rope, which has a color change at the midpoint of the rope. The rope manufacturer changes yarns and joins the yarns together with what is known as an "air splice" (forcing the ends to entwine around each other using extremely high air pressure). The process creates a cosmetic blemish at the yarn change, which the manufacturers say is actually stronger than continuous fibers because of the extra fibers at the splice. Without a middle mark on your rope, to find the middle simply start with both ends and flake the rope out until you reach the middle.

For pure rock climbing in dry conditions, it might not be worth the added cost, as dry ropes are generally more expensive and the coating wears off with repeated toproping.

Rope Length

In the 1970s the standard length for a dynamic climbing rope was 50 meters (165 feet). Today the standard is 60 meters (200 feet), and many climbers use 70-meter ropes (230 feet). These are the standard pre-cut lengths you can buy from a climbing shop. Static and low-stretch ropes are commonly sold in pre-cut lengths and are also sold directly from spools, cut to your desired length. Climbing instructors should select a length that is appropriate for the climbs they intend to instruct, choosing a length that will allow them to work at both the bottom (where the rope will need to be twice as long as the climb is tall) and the top of a given climb. It seems particularly wasteful to carry 70 meters of rope if the average climb in a given area is only 50 feet tall.

Rope Care and Use

Many instructors have access to climbing equipment, including ropes, through pro-purchase programs. Top brands include Sterling, Maxim, Bluewater, Mammut, Edelrid, Edelweiss, Petzl, Millet, and Metolius.

Avoid setting up a climb or rappel where a rope might abrade or cut over an edge. This can severely weaken or ruin your sheath in just one outing! Avoid standing or stepping on the rope, as this can grind sharp pebbles and grit through the sheath and into the core. Minimize the rope's exposure to UV light, as this will weaken the fibers over time. Store ropes in a shaded, dry place.

A dirty rope can be washed in a tub by hand or in a washing machine (preferably a front-loading washing machine, because a top-loading machine's agitator will abrade the rope) with hot water and a soap suitable for nylon. It's not a bad idea to run the machine once with nothing in it before washing the rope to make sure that there is no residual bleach inside. If washing a rope in a bathtub, make sure the tub is free from any chemicals that may damage the rope. It's best to daisy chain the full length of the rope before washing it in a machine to keep it from getting tangled. Let the rope dry by hanging it in a shaded area.

Be vigilant and protect all rope from coming into contact with any chemicals that contain acids, bleaching or oxidizing agents, and alkalines. Acid is the arch enemy of nylon and can severely weaken nylon and polyester fibers. Be extremely cautious to avoid exposing your rope to battery acid or any type of acid that may be encountered in your garage or the trunk of your car. It is wise to store your rope in a rope bag.

It is not a good idea to use a rope without knowing its history. Keep track of its history and age. Most manufacturers recommend keeping a rope for no longer than five to seven years even with minimal use, and no longer than ten years even if the rope has been stored and never used. In an institutional setting, ropes are usually retired well before the manufacturer's recommendation.

Inspect a rope by running your hand over the entire length of the sheath when coiling and uncoiling the rope. Visually inspect for excessively worn areas on the sheath, and feel for irregularities (voids, flat spots, etc.) in the core. The rope should be retired (or cut to a shorter length) if the sheath is excessively worn or frayed, exposing the core, or if there are any anomalies in the core. Multiple fast rappels can burn the sheath of a rope, a result of heat generated by the friction between the rope and rappel device. If the sheath feels glazed or melted, the rope should be retired.

Coiling and Uncoiling Your Rope

With a new rope, take extra care the first time it is uncoiled to prevent kinking. The best method is to simply unroll the rope from the coil, as if pulling it off a spool, holding the rope and rotating the coil until the entire rope is stacked on the ground, keeping the rope free from any twists. Once the rope is in a loose pile, inspect the rope by running it foot by foot through your hands from one end to the other, then coil it with the butterfly coil method.

BACKPACKER OR BUTTERFLY COIL

The backpacker, or butterfly, coiling method puts fewer kinks in your rope. It is also the fastest way to coil a rope, since you start with both ends or from the middle and coil a doubled rope. When coiling from the ends, a good idea is measuring two and a half arm lengths (both arms extended), then begin the butterfly.

MOUNTAINEER'S COIL

Another standard coiling method is called the mountaineer's coil. This is a traditional method that makes for a classic, round coil that can be easily carried over the shoulder or strapped onto the top of a pack.

Backpacker or Butterfly Coil

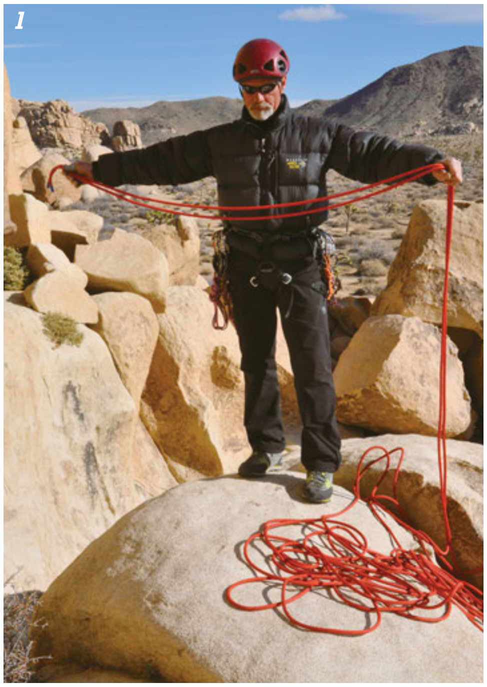
1

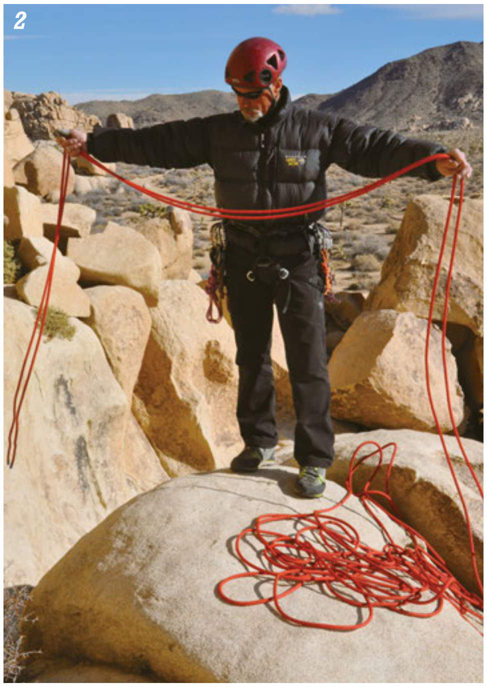
2

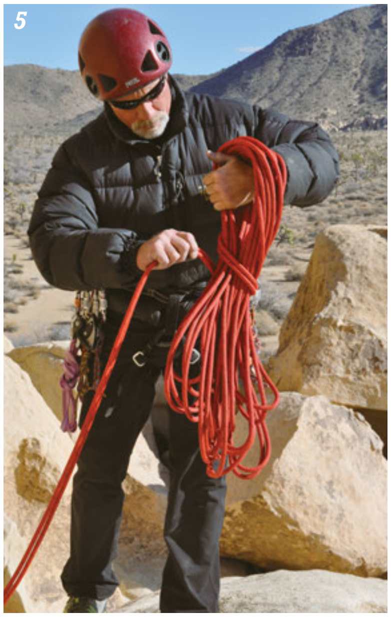
5

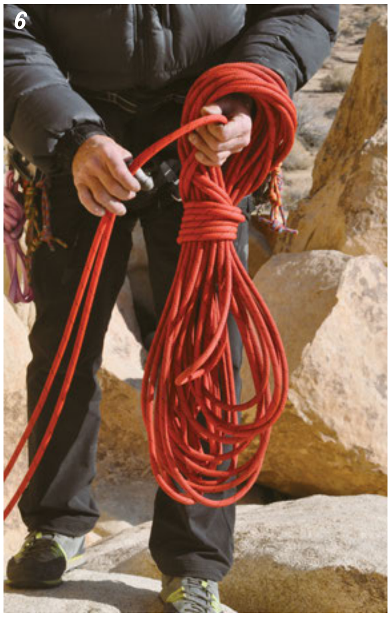
6

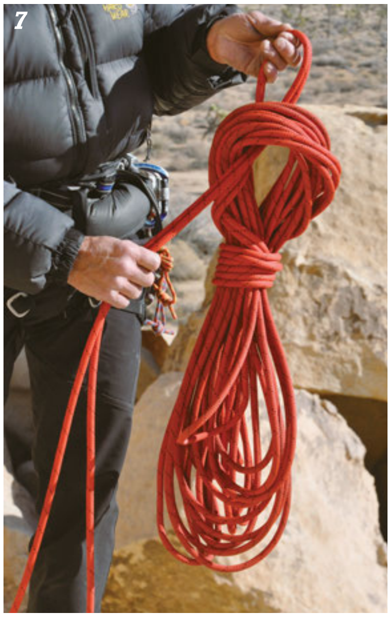
7

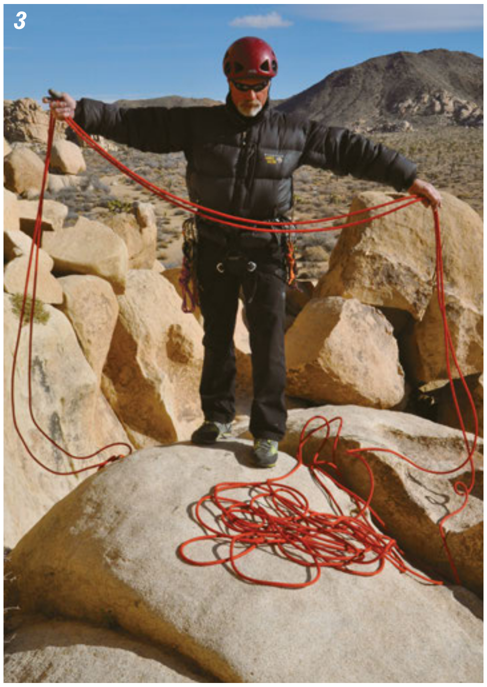
3

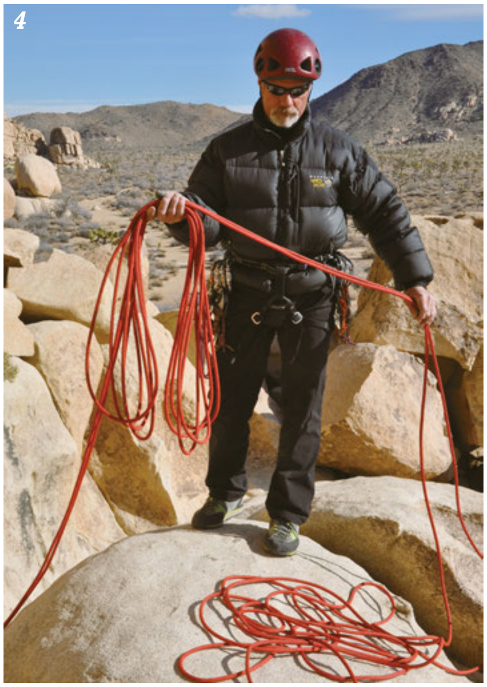
4

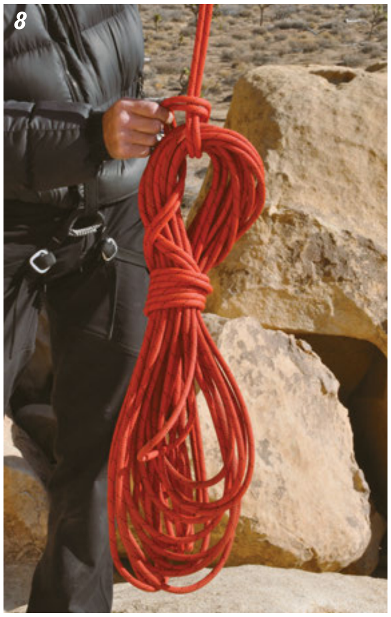
8

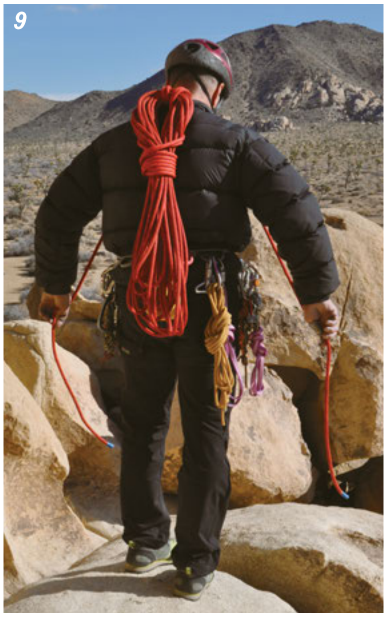
9

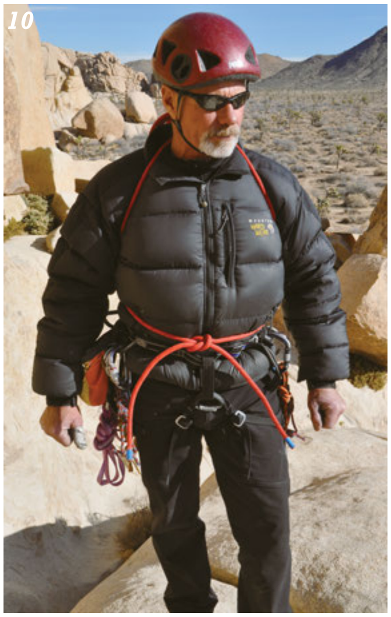
10

The Mountaineer's Coil

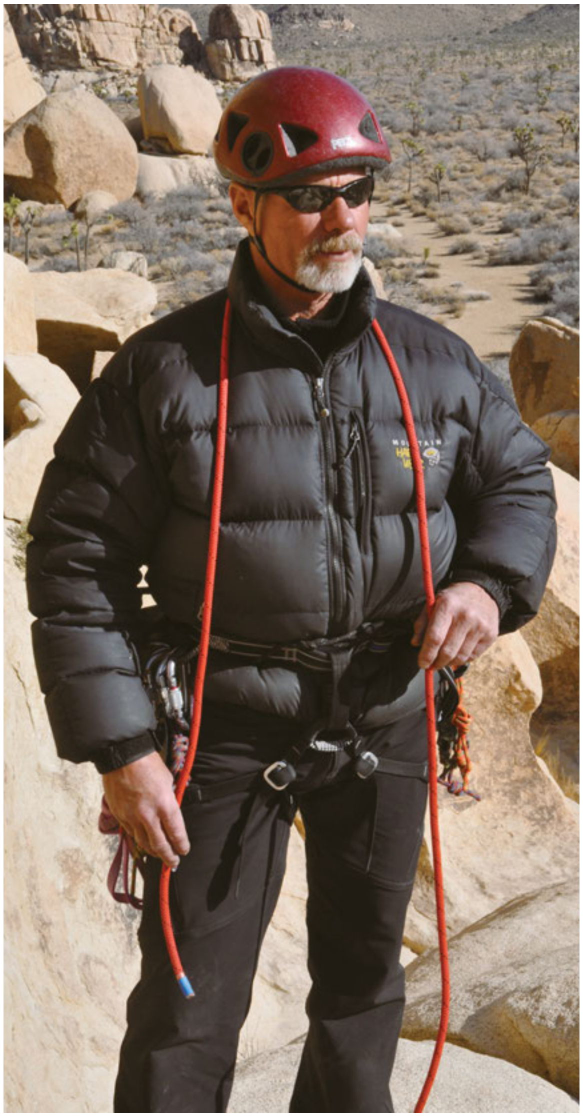

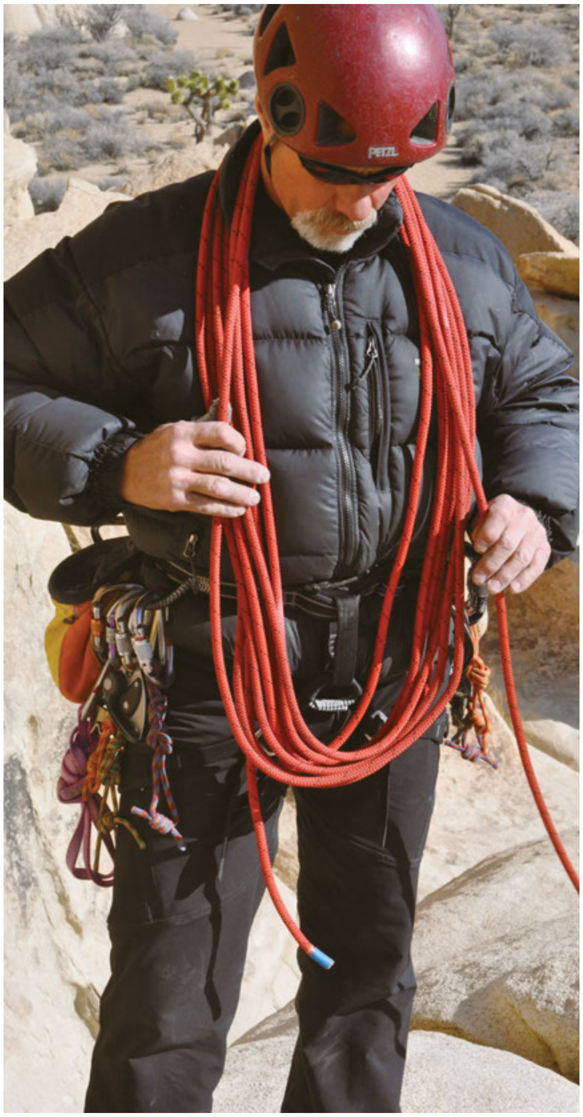

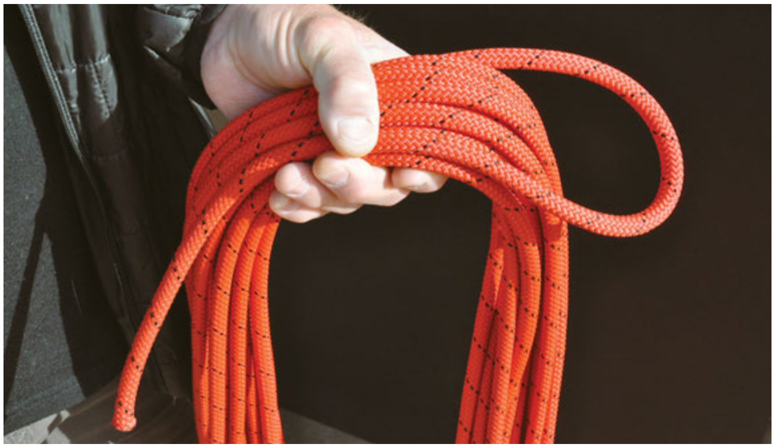

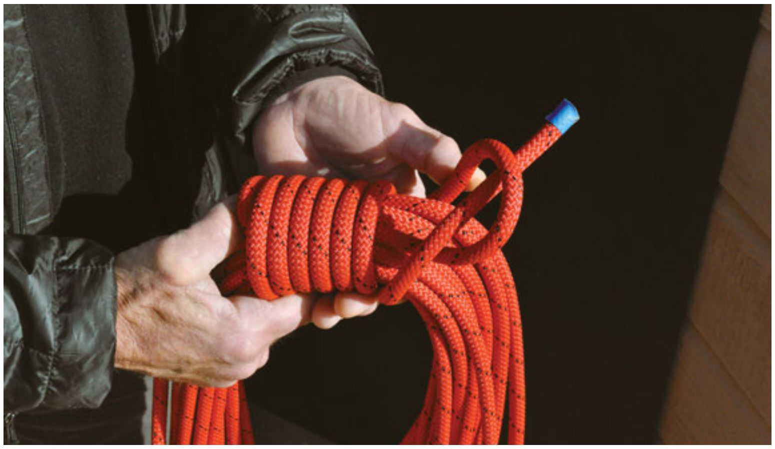

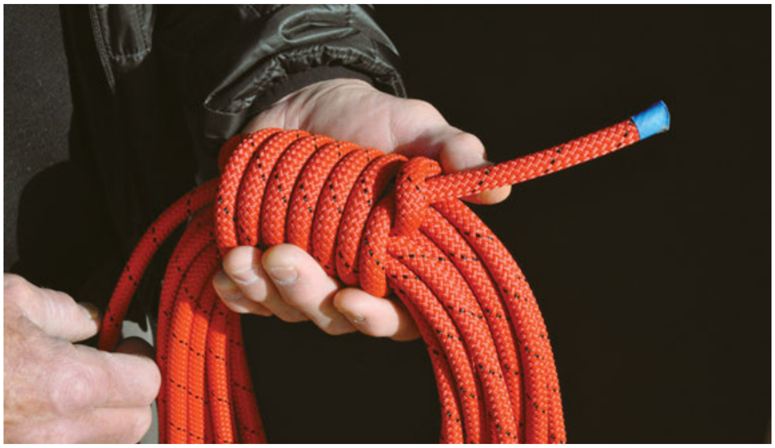

Knife

Many instructors carry a knife, but it's important to be very careful with when and how it is used. A rope under tension cuts more easily than one not under tension.

Slings and Webbing

In the 1960s and 1970s, 1-inch-wide tubular nylon webbing was the standard sling material, tied into a loop with a water knot or double fisherman's knot. Eventually, sewn slings with bartacked stitching came onto the market and were actually stronger than the same material tied with a knot. Sewn slings are not only stronger but also more secure in that there is no need to worry about the knot loosening and coming untied.

Flat Webbing

Flat webbing is woven solid, as opposed to tubular webbing, which is woven into a hose-like shape. It is stiffer and more abrasion resistant than softer tubular webbing, with a higher tensile breaking strength (Sterling 1-inch flat webbing is rated at 43.5 kN or 9,800 lbs.), which makes it useful for high-strength applications. Its stiffness, however, makes it more difficult to knot and gives it poor handling characteristics, making it unpopular with climbers.

Tubular Webbing

There are two types of tubular nylon webbing: mil-spec and climb-spec.

Since webbing was originally manufactured for military applications, mil-spec means that the webbing meets the standards demanded by the military. Mil-spec has a coarser, rougher-textured weave, with a more pronounced ribbing across the width of the webbing. Climb-spec is a finer, more high-quality weave, without the noticeable ribbing and with a more tightly woven edge. Climb-spec usually tests slightly stronger than mil-spec webbing and is generally more abrasion resistant and more impervious to tearing or slicing over a sharp edge, but both are suitable for climbing applications and both are roughly the same price.

The Bluewater company, known for manufacturing high-quality webbing, says its climb-spec tubular nylon webbing "outperforms normal Mil-spec webbing in strength, flexibility, knotability, and durability. There is minimum exposure to individual fibers as a result of high thread count and fine weave pattern." Bluewater's 1-inch climb-spec

What Kind of Knife Is Best?

In thirty years of climbing, I (BG) have never had to cut anything out of a rappel device. I have used a knife to cut off old nylon webbing and cord from rappel anchors, to replace it with new material. In fact, if I'm doing a climb that requires a rappel using a natural anchor such as a tree or block of rock, I'll make sure I have a small knife and extra webbing. I prefer a stainless steel knife with a serrated blade that easily cuts both rope and webbing. Petzl makes several models specifically for climbing, with a folding design that has a round hole on one end to clip into a carabiner. The small and lightweight Trango Pirana has a unique design that can't open up on you when it's folded closed and clipped into a regular-size carabiner.

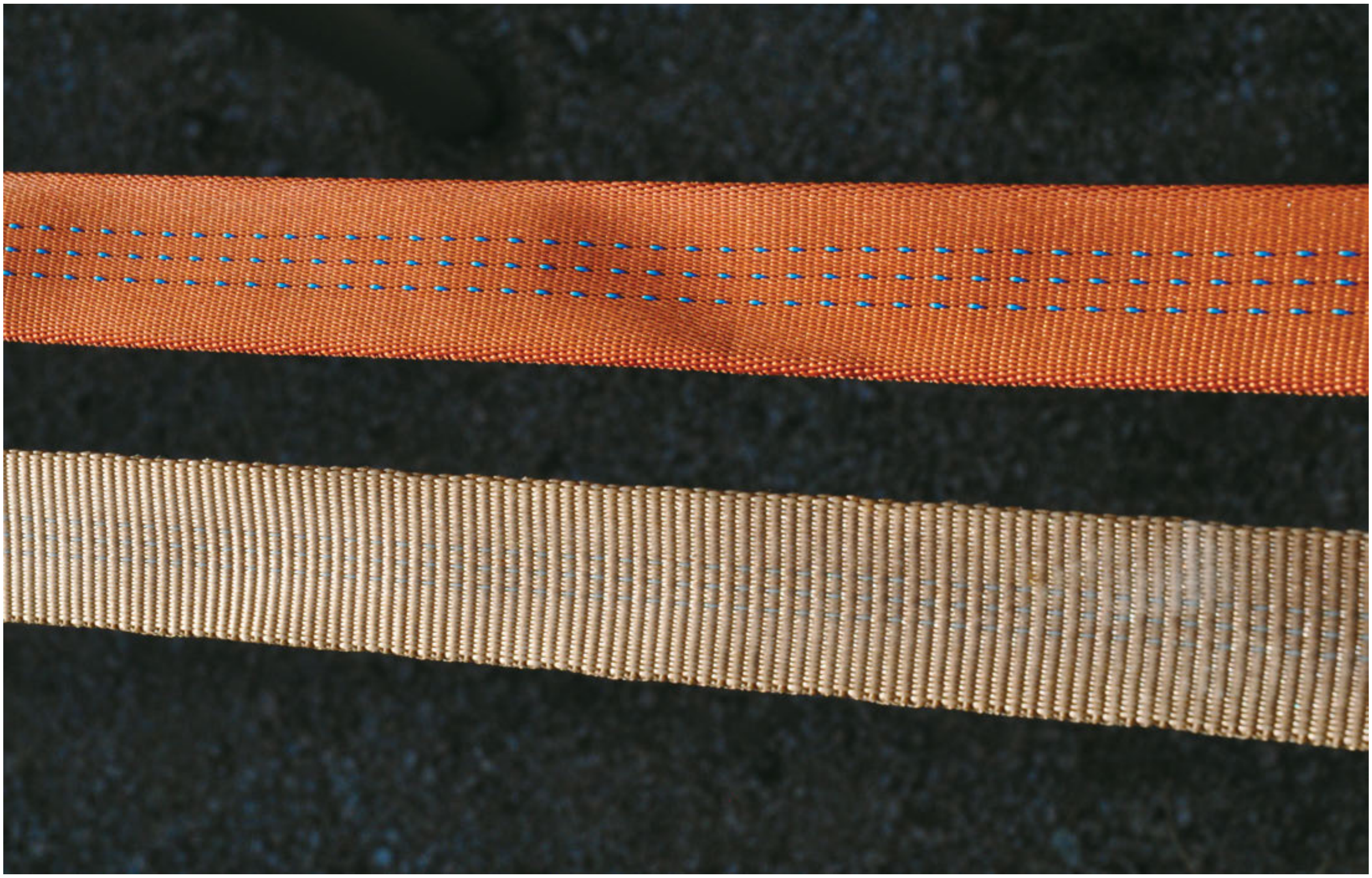

Top: Bluewater 1-inch climb-spec nylon webbing, rated at 18 kN tensile strength (4,047 lbs.).
Bottom: REI 1-inch mil-spec nylon webbing, rated at 17.8 kN tensile strength (4,002 lbs.).

webbing has been tested to over 6,000 pounds loop strength when tied with a water knot.

The Sterling Rope Company's version of climb-spec webbing is called Tech Tape, with a "smoother, denser weave and higher tensile strength" than their mil-spec webbing and a 4,300-pound tensile strength. Sterling's mil-spec webbing is rated to a minimum breaking strength of 4,000 pounds tensile strength and 6,129 pounds in a bartacked sewn loop.

Pre-sewn nylon slings are typically sold in $^{11}\!/_{16}$-inch (18mm) width, bartacked into 24-inch or 48-inch loops with a rating of 22 kN (4,946 lbs.) loop strength.

Climbing shops sell both mil-spec and climb-spec 1-inch tubular nylon webbing from spools, cut to any length desired. Be aware that these spools of webbing contain taped splices where the webbing ends have been joined together with masking tape.

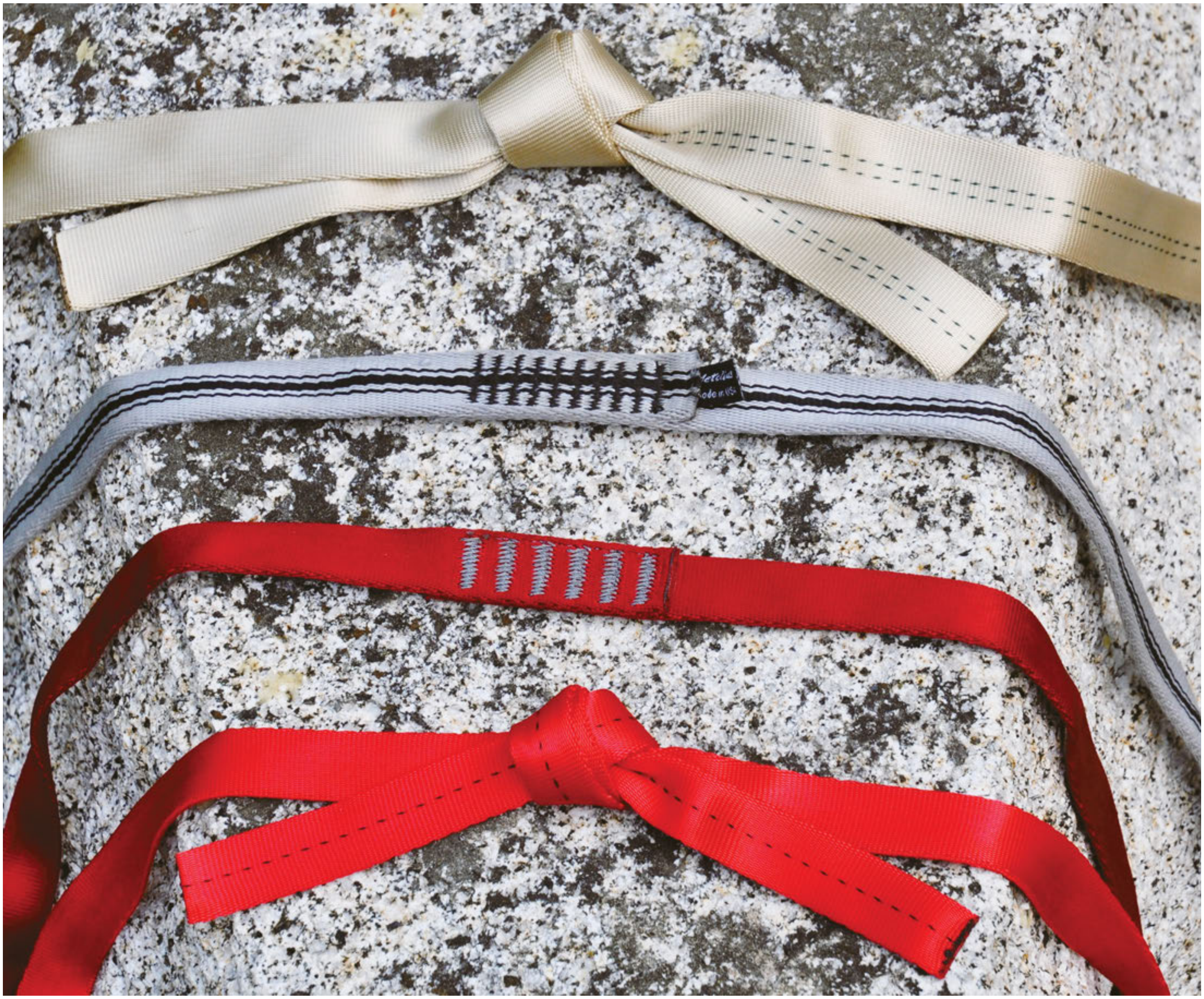

Nylon Webbing Comparison. Top to bottom: 1-inch tubular nylon webbing tied with water knot (Sterling Tech Tape, rated at 4,300 lbs. tensile strength); 18mm Metolius Nylon Sling, rated at 22 kN loop strength (4,946 lbs.); 18mm Black Diamond Runner, rated at 22 kN loop strength; 11⁄16-inch Sterling tubular webbing tied with a water knot (rated at 3,000 lbs. tensile strength).

Tape Won't Hold Body Weight

It seems impossible for it to happen, but I (BG) know of two cases where spliced webbing was sold to customers who then used it with only the masking tape joining the webbing together, in one case with devastating results.

One advantage of cut to length nylon webbing is that it can be untied and re-tied around a tree, through a tunnel, or threaded through bolt hangers for a rappel anchor.

The Bluewater company recommends the maximum lifespan of its nylon webbing to be no more than five years, and also recommends retiring a nylon sling if it has been subjected to temperatures above 176 degrees Fahrenheit, is scorched or glazed from a rope being pulled across it, shows signs of UV degradation from being left out in the elements (faded color and/or stiffness), or if it has been exposed to acid or bleach. Like a nylon rope, nylon webbing can lose an appreciable amount of strength when wet or frozen.

Spectra and Dyneema Slings

Spectra slings, introduced in the late 1980s, were lighter, less bulky, and stronger than nylon. Dyneema is a more recent innovation, typically sold in various-length loops sewn with bartacked stitching in 10mm width. Dyneema and Spectra both have almost the identical chemical makeup of high-molecular-weight polyethylene, which, pound for pound, is stronger than wire cable. Most experts say that the manufacturer of Dyneema consistently produces more high-quality fibers than the manufacturer of Spectra material, and most of the climbing and rappelling slings on the market today are made from Dyneema.

Both Spectra and Dyneema slings are constructed from parallel fibers—very strong but with

Tensile Strength vs. Loop Strength

Strength ratings are often given as *tensile strength* and *loop strength*. Tensile strength is tested by a straight pull on a single strand of the material with no knots, done by wrapping the material around a smooth bar (4-inch diameter gives the most accurate test) on both ends and pulling until it breaks. Loop strength is the material tested in a loop configuration, either tied with a knot (in the case of webbing, usually the water knot) or sewn with bartacked stitching. In general, webbing loop strength when tied with a water knot is about 80 percent of twice the tensile breaking strength, and bartacked sewn webbing loop strength is generally about 15 percent stronger than the same material tied with a water knot, depending upon the quality and number of bartacks.

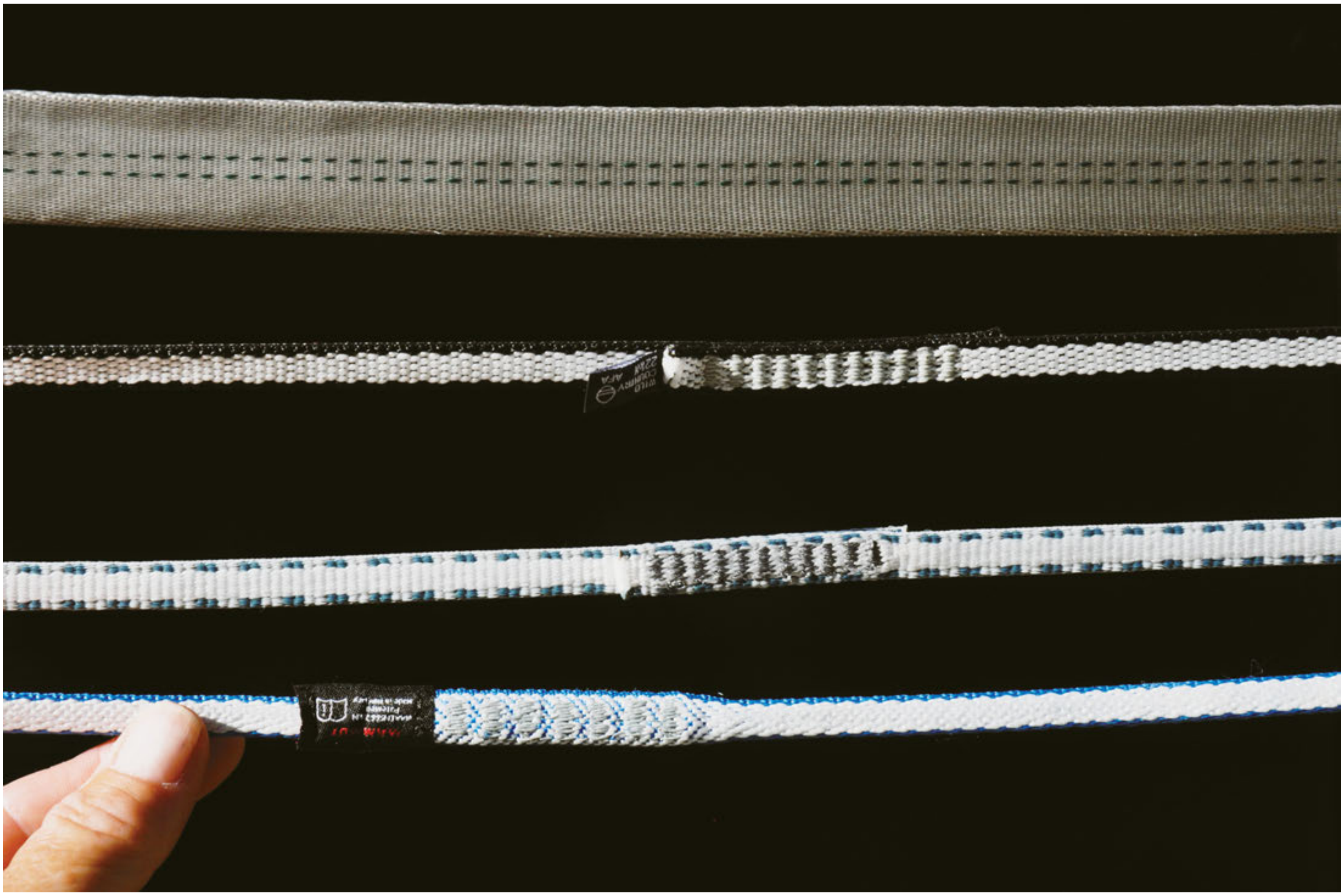

Dyneema Sling Comparison. Top to bottom: 1-inch tubular nylon webbing (for comparison); Wild Country 10mm Dyneema sling (22 kN or 4,946 lbs.); Black Diamond 10mm Dynex Runner (22 kN); Mammut 8mm Dyneema Contact Sling (22 kN).

high lubricity, which means the material itself is inherently slick. That is the reason it can only be purchased in sewn loops—it does not hold knots well. *Do not cut a Spectra or Dyneema sling and re-tie it with a water knot!*

Both Spectra and Dyneema have a lower melting point than nylon (around 300 degrees Fahrenheit for Dyneema/Spectra compared to nylon's melting point of around 480 degrees). The lower melting point, along with the inherent slipperiness, make Spectra and Dyneema slings a poor choice for tying friction hitches like the prusik, klemheist, or autoblock, compared to 5mm or 6mm diameter nylon cord.

In a pinch, if required to use a sling to tie a

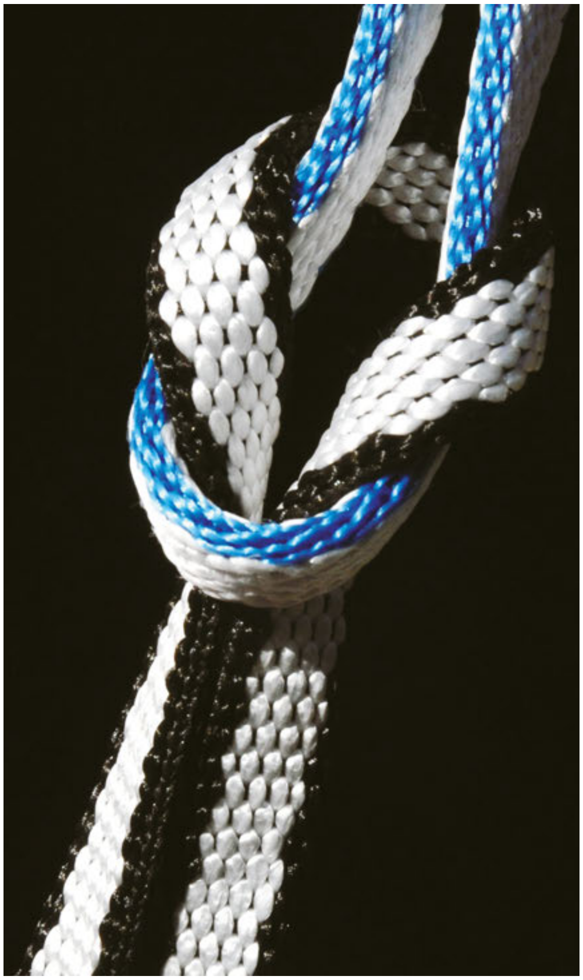

Girth-hitching two Dyneema slings together can decrease their strength by 50 percent, but for most climbing situations, this is not a concern since the loop strength is 5,000 pounds to begin with.

friction hitch, use a nylon one over a Dyneema or Spectra sling, as nylon will grip better. The newer, thinner (10mm width) Dyneema slings will work for friction hitches, and they do possess some nylon in their construction, but if they start to slide on a rope when under load, the friction will generate heat, which could potentially weaken the sling.

Both Spectra and Dyneema fibers do not retain dye and cannot be colored, so the fiber is distinctive in that it is always white. Manufacturers add a blend of nylon to Spectra and Dyneema, usually in a distinctive border pattern, and it is likely that in the future there will be more nylon in the mix. The Metolius company recently came out with 13mm width slings that are a blend of 36 percent Dyneema and 64 percent nylon.

When using Spectra or Dyneema slings, think of them like a wire cable—they have no stretch, even with nylon blended into the weave. Avoid tying knots with them—it can be almost impossible to untie a simple overhand knot in the newer, thinner Dyneema after it has been seriously weighted. Wild Country warns that the material loses a hefty percentage of its strength (around 50 percent) when tied in a simple overhand knot or girth-hitch—a property that nylon does not possess. The best way to use a Spectra or Dyneema sling is clipped to carabiners. If using them in a sling-to-sling configuration, either basket one sling over another, or use a properly tied girth-hitch.

When buying slings for rappelling, 1-inch or 11⁄16-inch width tubular nylon webbing will be the most versatile material for rigging rappel anchors, as it can be cut to a desired length and tied with a water knot. Double-length (48-inch) sewn nylon slings are also handy for tethering into anchors and extending a rappel device away from the harness. The UIAA states that sewn slings should have a minimum load-bearing capacity of 18kN.

Recent studies show that dirty slings are weaker than clean ones. The Mammut company suggests that "to maintain the quality and safety of your slings, you need to clean them regularly." Mammut recommends to "clean soiled slings in hand-hot water with a small amount of mild detergent or in a delicates machine cycle up to 30 degrees centigrade (86 degrees Fahrenheit). Rinse in clear water. Leave to dry in shade."

Cords and Cordelettes

A good all-purpose cordelette is 7mm diameter nylon cord, about an 18- to 20-foot length tied into a giant loop with a double fisherman's knot or a double overhand flat bend.

Cordelettes made with a Spectra or Dyneema core and nylon sheath have incredibly high strength and low stretch. Pound for pound, Spectra and Dyneema are stronger than steel (and is the material used in body armor for the military), but both Spectra and Dyneema lose an appreciable amount of strength when tied with knots. Because these cords are so light and strong, with less bulk to carry, they have become popular, especially for rock climbing. The Bluewater company markets the 5.5mm diameter Titan Cord, with a Dyneema core and nylon sheath, rated at 13.7 kN (3,080 lbs.). They say its "combination of high strength, low elongation and light weight provides superior characteristics over other combinations. Dyneema does not lose significant strength with repetitive flexing and offers a huge increase in abrasion and cut resistance over other materials. Bluewater Titan Cord can be cut and sealed with a hot knife. We recommend a triple fisherman's knot for tying 5.5 Titan into loops."

Ideal Cordelette

I (BG) prefer a length that allows me to double the cordelette within the span of my outstretched arms. My favorite brand is Sterling, whose 7mm diameter nylon cord is rated at 12.4 kN (2,788 lbs.) and tests over 5,000 pounds when tied into a loop with a double fisherman's knot.

In recent years high-tech cords utilizing aramid fibers (namely Technora) for the core, with a nylon sheath, have become popular. Aramid fiber has extraordinary tensile strength (stronger than Spectra or Dyneema) with low stretch and an extremely high melting point (900 degrees Fahrenheit), making it difficult to cut and melt. The best way to cut it is with wire cable cutters. Then milk the nylon sheath over the end and seal it by melting the nylon with a lighter. The Sterling 6mm Powercord has a Technora core and nylon sheath, with a single strand breaking strength of around 19 kN (4,271 lbs.); and the 5mm Tech Cord, sold by Maxim/New England Ropes, with a 100 percent Technora core and polyester sheath, rates at a whopping 5,000 pounds tensile strength.

However, at the 2000 International Technical Rescue Symposium, Tom Moyer presented a paper titled "Comparative Testing of High Strength Cord" that revealed some startling deficiencies in Technora and other high-tech cords. Testing showed that with repeated flexing, aramid fibers break down much more quickly (losing strength) than good old-fashioned nylon. In his study a flex cycle test was performed on various cordelettes. The cord sample was passed through a hole in a steel fixture, flexed 90 degrees over an edge, and loaded with a 40-pound weight. The steel fixture was rotated back and forth 180 degrees for 1,000 bending cycles, then the cord's tensile strength was tested (single strand pull test) at the section that had been flexed. The Technora sample showed a remarkable loss of nearly 60 percent of its strength, while Sterling 7mm nylon cord and 1-inch tubular nylon webbing showed no strength loss at all. Bluewater Titan Cord (Spectra core/nylon sheath) showed a few hundred pounds of strength loss, but was nowhere near the drastic loss of Technora. Further research is warranted. The big advantage of these high-tech cords is their low weight, high strength, and low bulk, which is advantageous for situations like multi-pitch rock climbing and canyoneering.

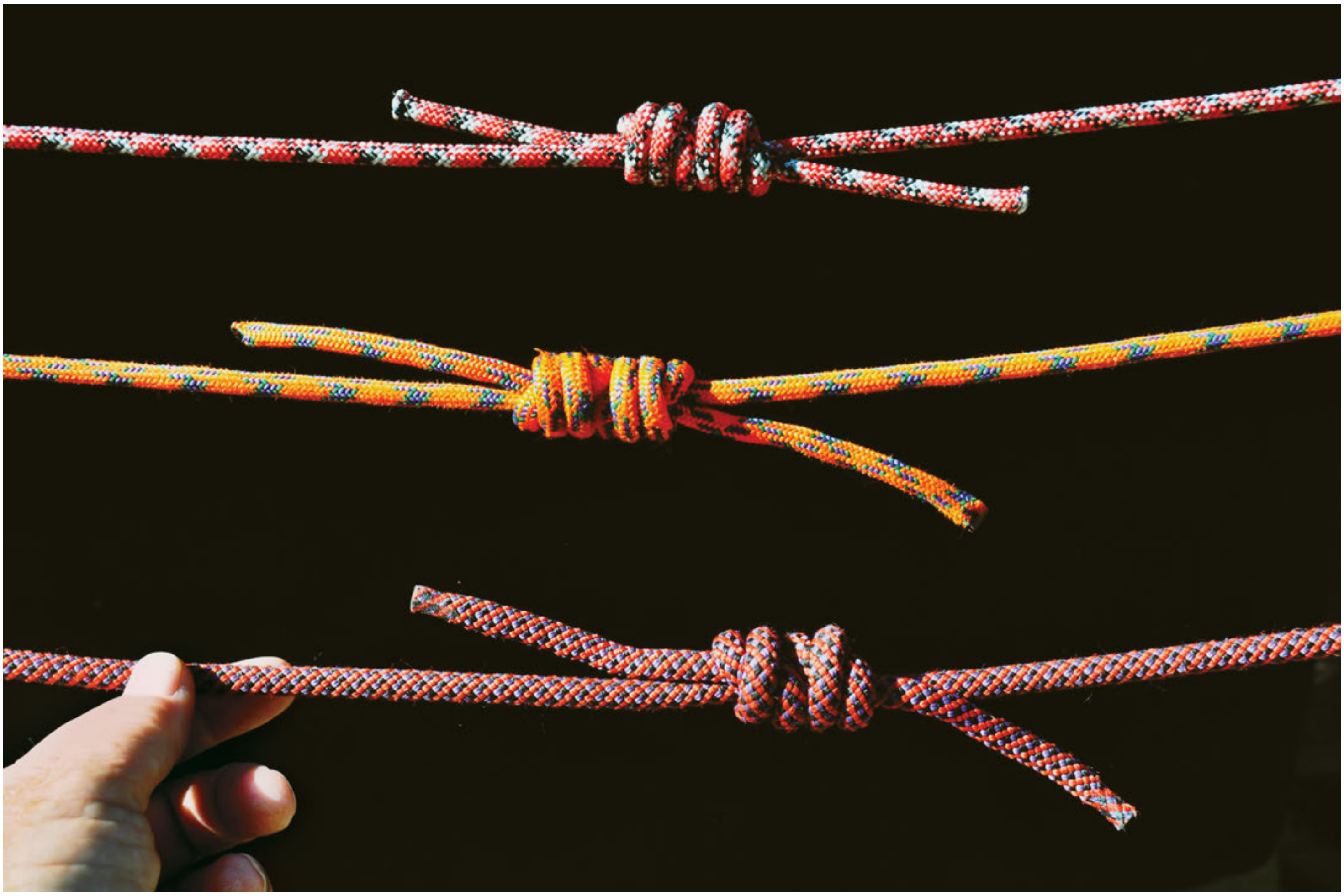

Cord Comparison. Top to bottom: Bluewater 5mm Titan Cord tied with triple fisherman's knot (Dyneema core/nylon sheath, tensile strength 13.7 kN or 3,080 lbs.); Sterling 6mm Powercord tied with triple fisherman's knot (Technora core/nylon sheath, tensile strength 19 kN or 4,271 lbs.); Sterling 7mm Nylon Cordelette tied with double fisherman's knot (nylon core/nylon sheath, tensile strength 12.4 kN or 2,788 lbs.).

If using these high-tech cords, be sure to tie the cordelette with a triple fisherman's knot, and consider replacing them more often with high use. Keep in mind that the price tag on the high-tech cords is roughly twice as much as nylon. The bottom line is this: For an all-purpose cordelette, it is hard to beat old-school nylon—a 7mm diameter nylon cord is a good choice because it is thick enough to be used in a single loop application, like a personal tether. Narrower 5mm and 6mm cord are great to tie friction hitches like the prusik, klemheist, and autoblock, but these smaller-diameter nylon cords don't provide enough tensile strength to be used as a critical link. When buying this accessory cord to be used primarily for friction hitches, buy the softest, most pliable cord that you can find; a stiff cord won't grip as well. Also, be aware of the difference between 5mm nylon accessory cord (typically rated at 5.2 kN or 1,169 lbs.) and 5mm high-tenacity cord (like Bluewater Titan Cord rated at 13.7 kN (3,080 lbs.) As mentioned, 5mm nylon is less useful as a cordelette, while a 5mm high-tenacity cord offers a much greater tensile strength for anchor building and critical link applications.

Carabiners

Carabiners are used primarily to attach various links (like slings and rope) together in the anchor or belay chain. Locking carabiners are used in critical applications and in conjunction with belay and rappel devices. Carabiners come in a variety of shapes: oval, D-shaped, and pear-shaped.

The basic design is of aluminum alloy, with a spring-loaded gate on one side. The spine of the carabiner is the solid bar stock opposite the gate. The small protrusion on one end of the gate is called the nose, and this will tell you which way

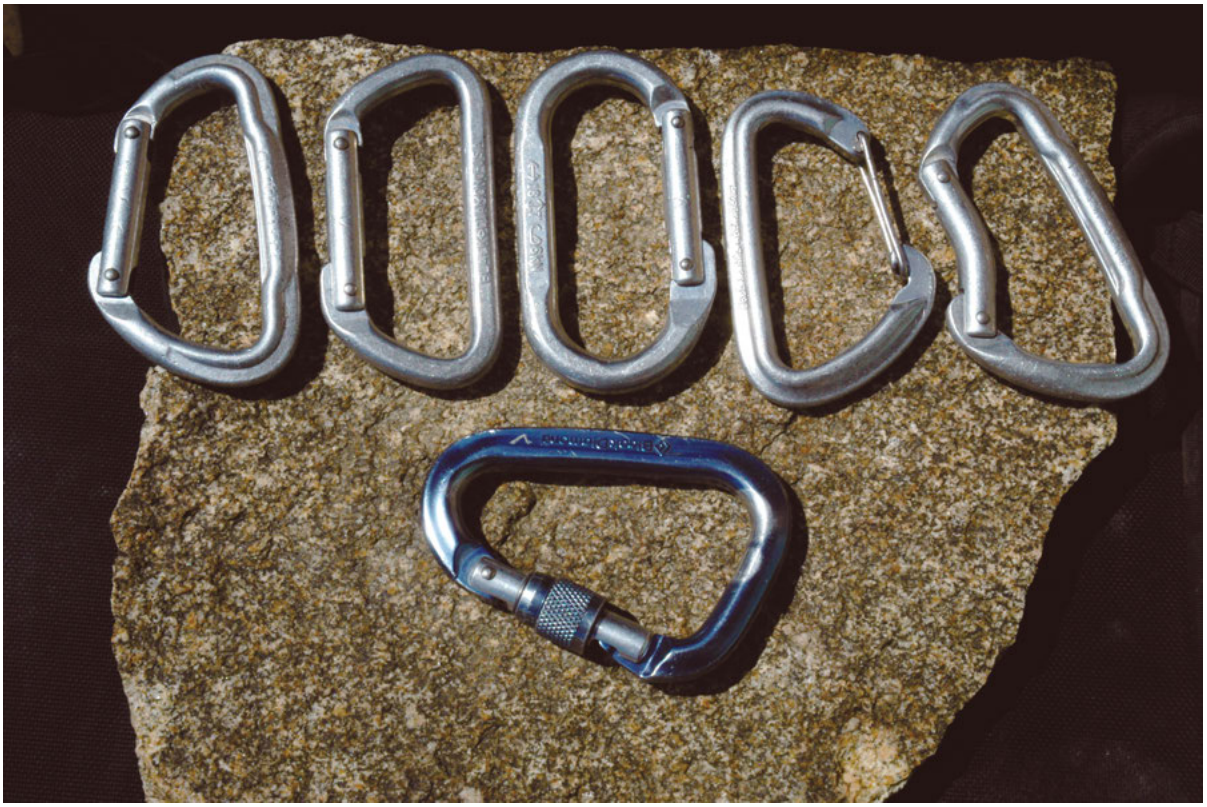

Carabiners come in a dazzling array of designs for various applications. Top row (left to right): asymmetrical D, regular D, oval, wiregate D, bentgate D. Bottom: pear-shaped locking. The most useful carabiners for toproping are ovals, Ds, and pear-shaped locking. Bentgate carabiners are used primarily for sport climbing (on the rope-clipping end of a quickdraw).

the gate opens by visual observation. The basic design has a small pin on the gate that latches into a groove on the nose end. The preferable "keylock" design eliminates the pin, and the gate and bar come together in a machined notch. A wiregate carabiner has a wire under tension serving as the gate, providing a wide opening because of its slim mass and eliminating "gate flutter," which is a vibration of a normal solid gate during a fall or peak loading of the carabiner.

Oval carabiners are useful for racking gear, and for use in sets of two or three for connecting the climbing rope to the toprope anchor master point. Because of their symmetry, the gates can be opposed and reversed, and the carabiner configuration retains its oval shape. Two opposed and reversed ovals can also be used in lieu of one locking carabiner at any critical junction in the anchor system, in situations where there are no additional locking carabiners and extra security at a key point is needed. Three opposite and reversed ovals are the equivalent of two opposite and reversed locking carabiners.

Locking carabiners are used for critical links and applications where it is absolutely imperative that

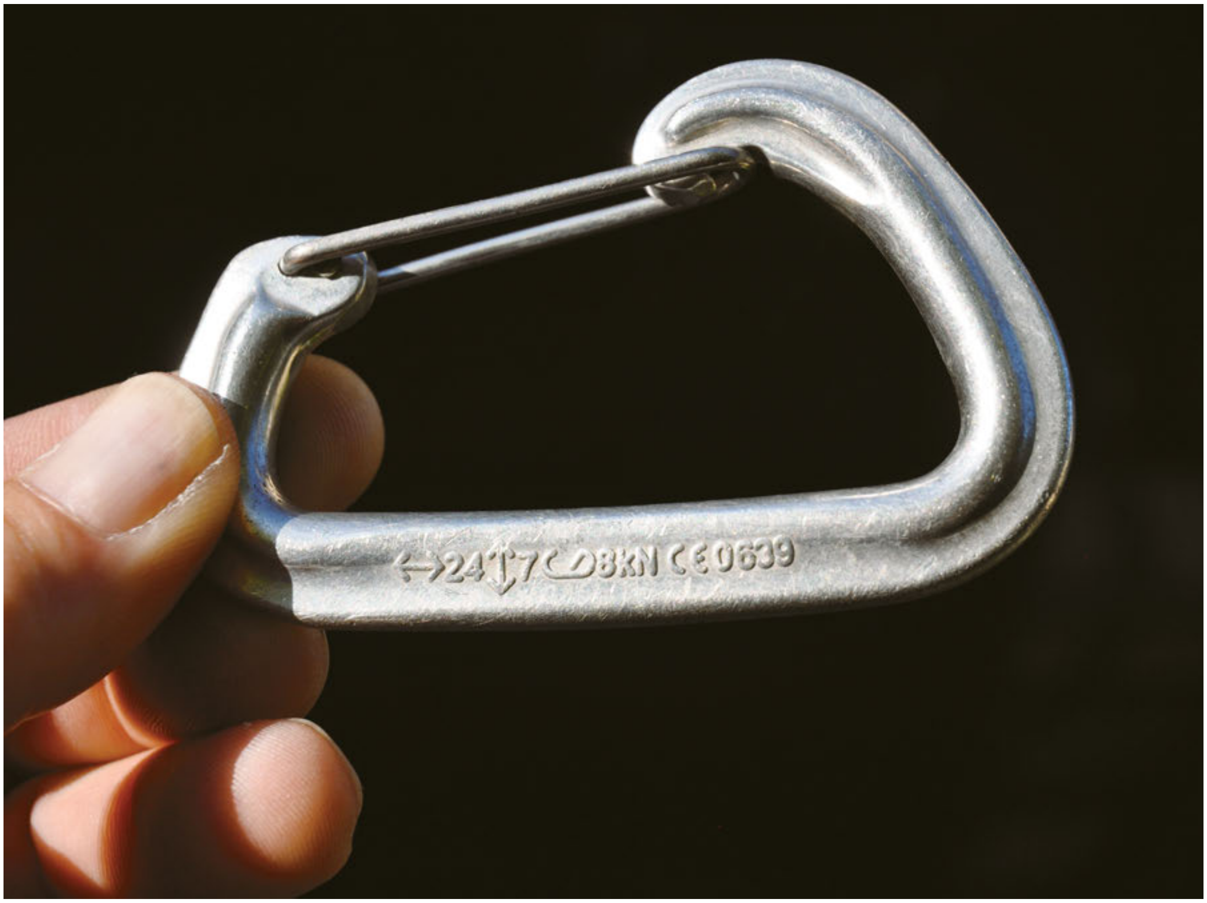

Wiregate carabiner. Every carabiner you buy should have the UIAA breaking strength rating stamped on the spine.

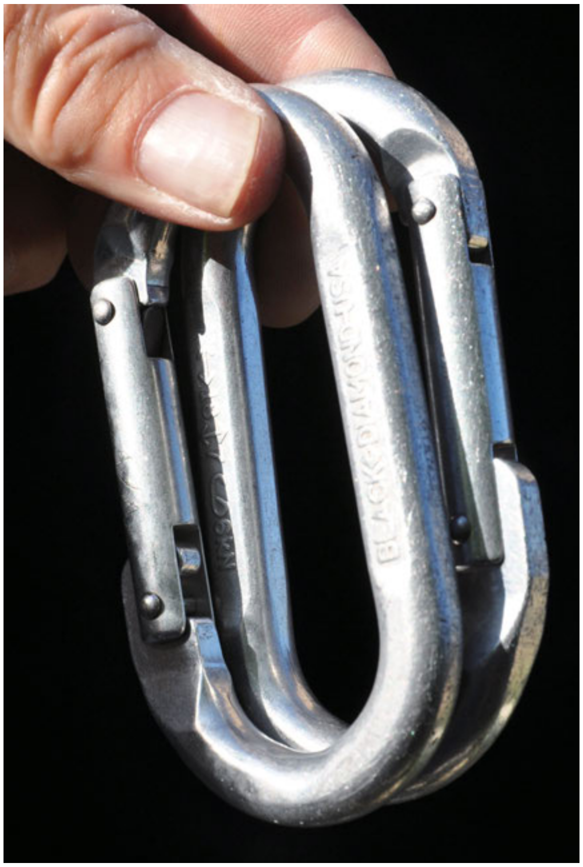

Two oval carabiners with the gates properly opposed and reversed.

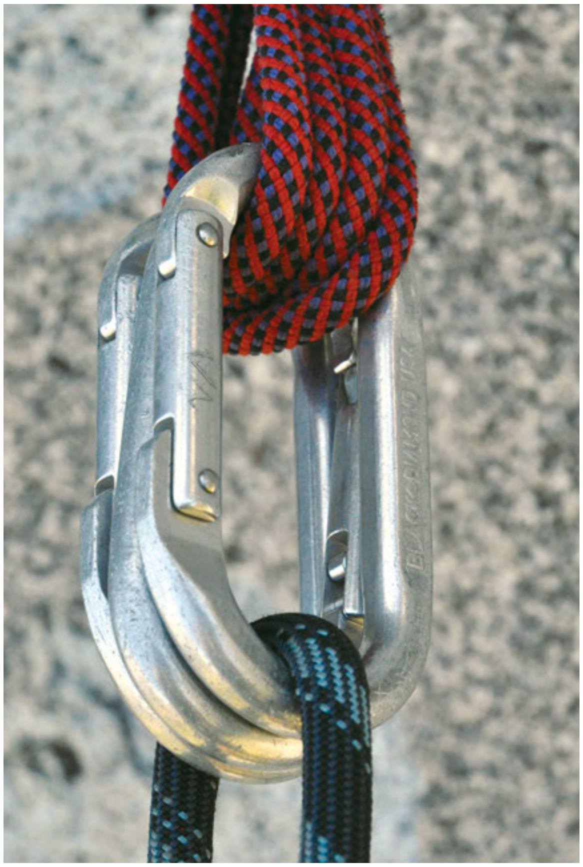

Three ovals opposed and reversed at a toprope anchor master point.

the carabiner gate stays closed, like on a rappel or belay device, at a critical link in the anchor system, or when attaching the belayer's climbing rope to the anchor.

D-shaped carabiners have the strongest configuration, since when the carabiner is loaded on the major (long) axis the weight naturally is loaded closest to the spine. For this reason a locking D is a good choice for a belay/rappel carabiner. A locking pear-shaped carabiner is useful for many applications because of its wide aperture on one side. It's a good carabiner to use with a Munter hitch, and a great carabiner to pair up for use at the toprope anchor master point. When using two opposed and reversed pear-shaped locking carabiners, the symmetry is maintained (unlike an asymmetrical D shape), and the climbing rope runs smoothly through the carabiners.

The most common locking carabiner is the screwgate. The screwgate locking carabiner is just that, a mechanism with a collar that screws shut over the nose of the carabiner. Obviously, with a screwgate locking carabiner, it is important to remember to lock it, and it's an essential habit to always check locking carabiners to make sure they are locked. Check them with a close visual

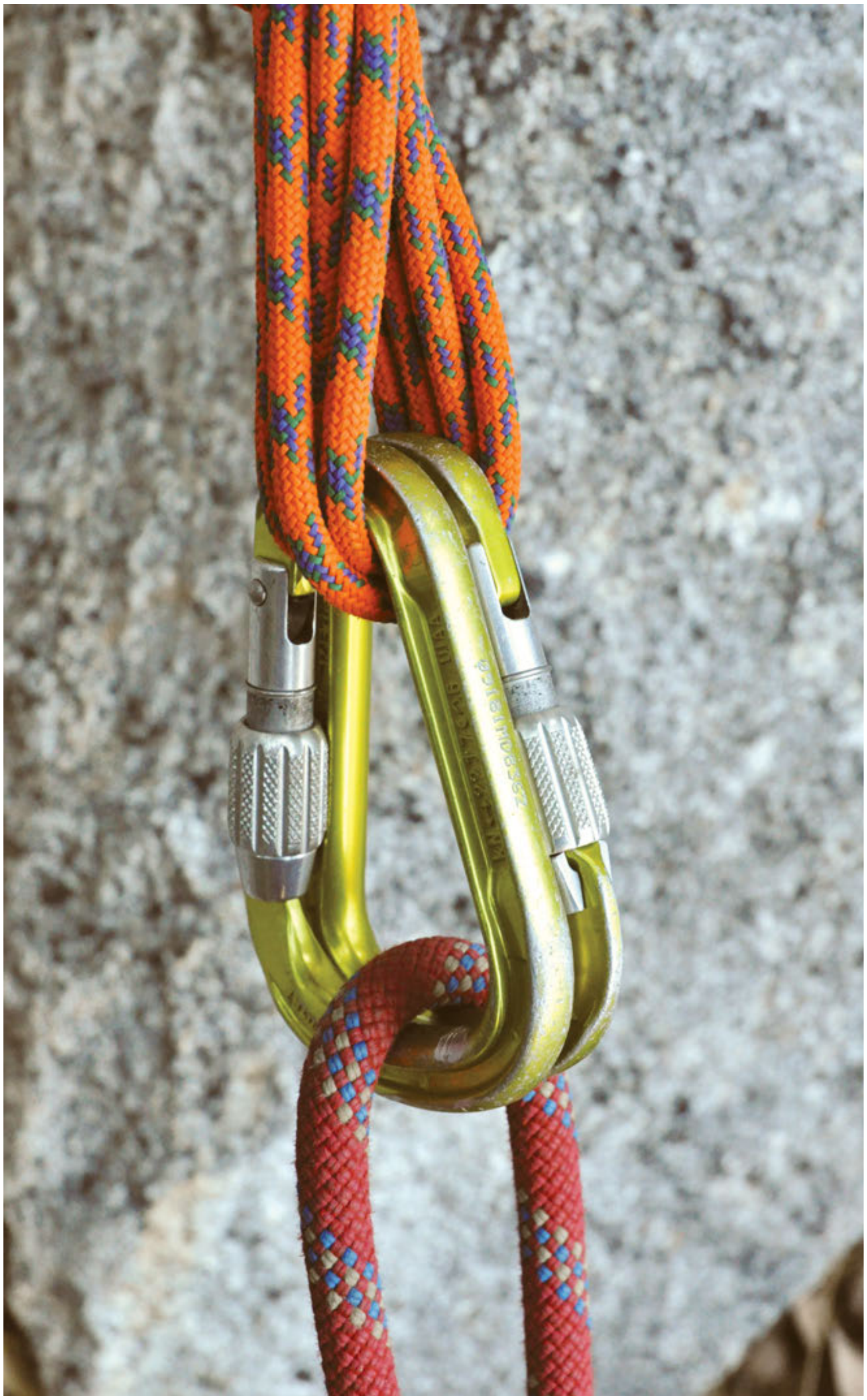

Two pear-shaped locking carabiners with the gates opposed and reversed on a toprope anchor master point.

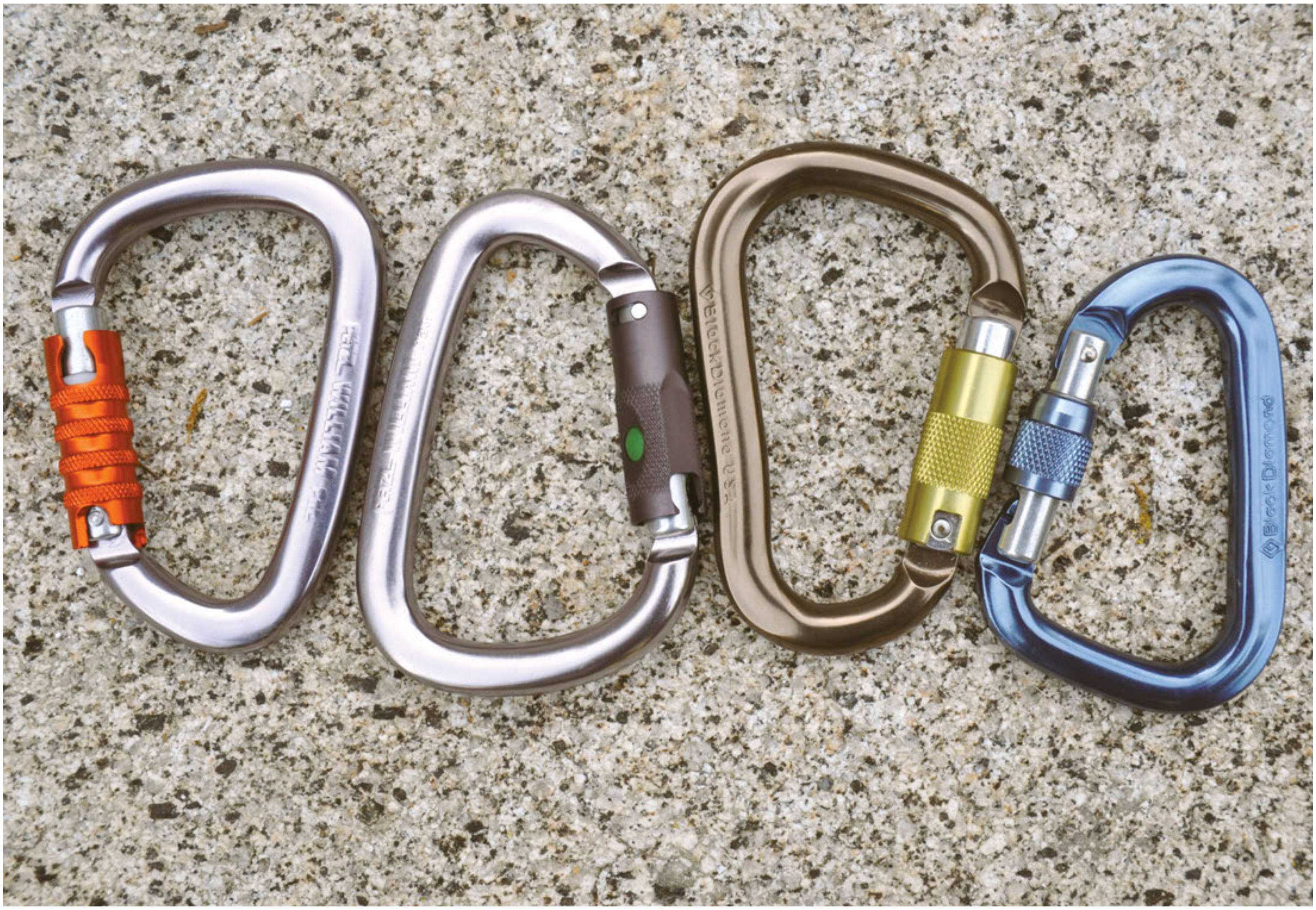

Locking carabiners (left to right): Petzl William Triac, Petzl William Ball Lock, Black Diamond Twistlock, Black Diamond Screwgate.

inspection, and also by pressing on the gate (squeeze test) for an additional precaution.

If you're absentminded, or occasionally find yourself not locking a screwgate carabiner, consider buying an autolock or twistlock carabiner. The twistlock design has a spring-loaded gate that locks automatically, and there are several autolocking designs on the market that have a tri-action mechanism that must be manipulated (like pushing the gate upward, then twisting the gate to lock it; or pressing a button, then twisting open the gate), but some climbers find these difficult to use. Interestingly, for industrial workers in the Industrial Rope Access environment (rappelling and rope ascending on the faces of dams, buildings, and bridges), OSHA standards require autolocking carabiners, as does the tree-trimming industry.

The UIAA has determined the following strength ratings for a carabiner to be CE certified.

For oval carabiners:

- Closed gate strength, major (long) axis: 18kN (4,047 lbs.)
- Minor axis (cross-loading) strength, closed gate: 7kN (1,574 lbs.)
- Major axis strength, gate open: 5kN (1,124 lbs.)

For regular D-shaped carabiners:

- Closed gate strength, major (long) axis: 20 kN (4,496 lbs.)
- Minor axis (cross-loading) strength, closed gate: 7kN (1,574 lbs.)
- Major axis strength, gate open: 7kN (1,574 lbs.)

For locking carabiners:

- Closed gate strength, major (long) axis: 20 kN (4,496 lbs.)
- Minor axis (cross-loading) strength, closed gate: 7kN (1,574 lbs.)
- Major axis strength, gate open: 6kN (1,349 lbs.)

An important thing to remember with carabiners is that a carabiner is only about one-third as strong if it's loaded with the gate open. For this reason, keep a few things in mind when using a carabiner:

Carabiners should always be loaded on the major axis.

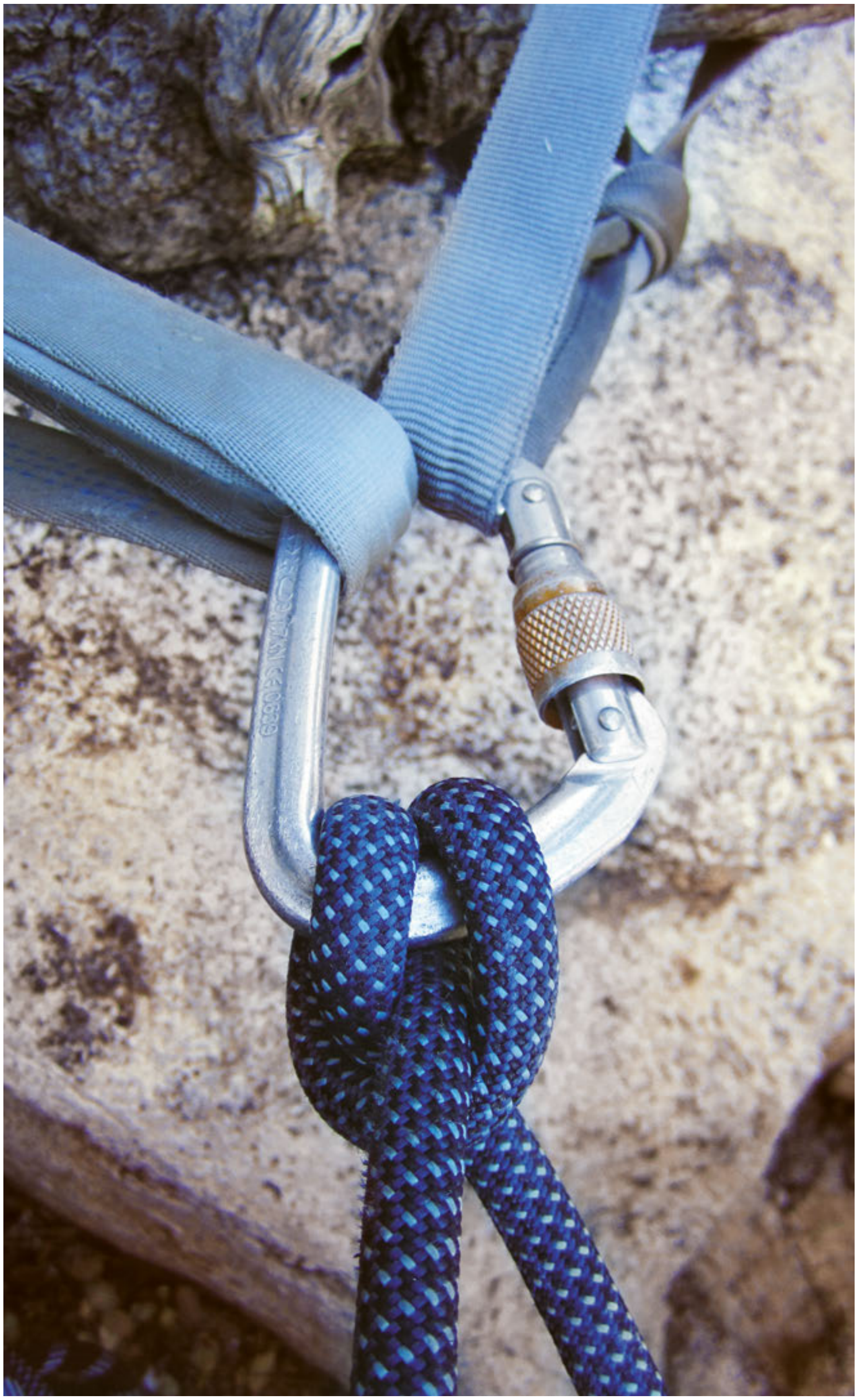

Bad. Never load a carabiner in three directions as shown here.

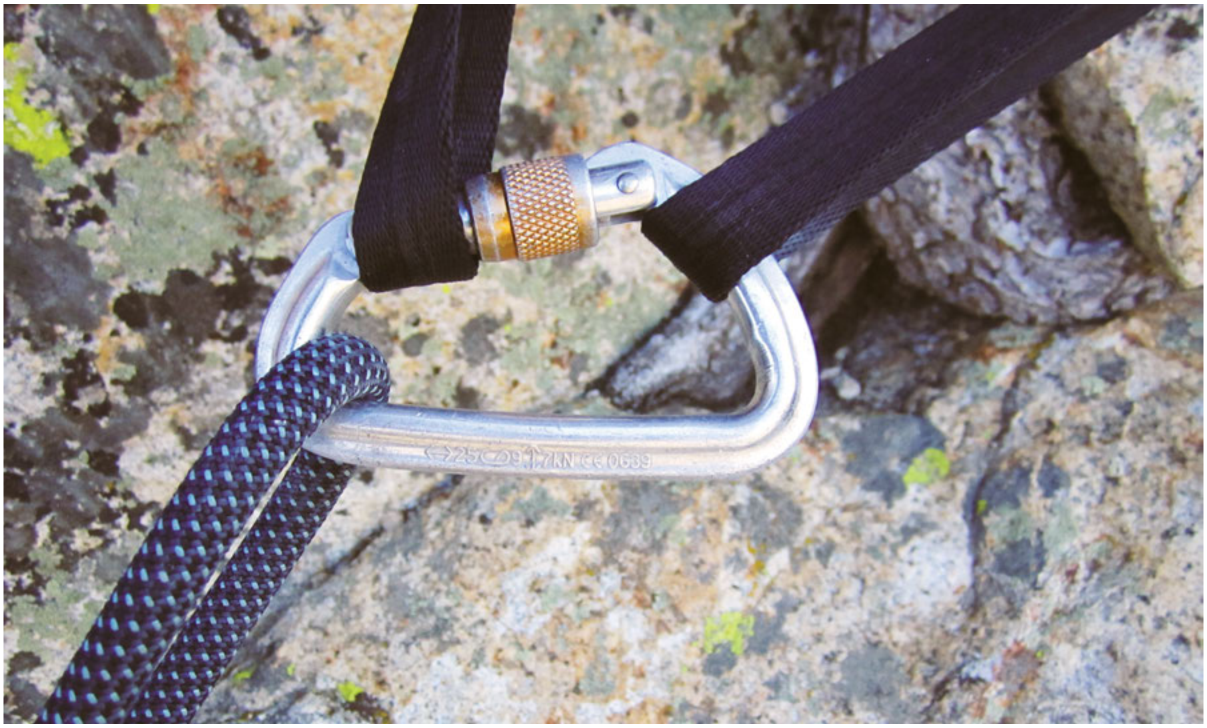

Very bad. Never load a carabiner on the minor axis or with a force outward on the gate.

- Always load the carabiner in the proper direction—on the major, or long, axis.
- Do not cross-load a carabiner (on the minor axis) or load it in three directions (called triaxial loading).
- Do not load a carabiner over an edge of rock—this can open the gate when the carabiner is loaded, and two-thirds of the carabiner's strength will be lost.

Ovals

I (BG) have used three ovals for thousands of client days without incident. Simply oppose and reverse the outside carabiners to the middle one. The wide radius created by the three carabiners provides a stable platform for the rope and tends not to flip sideways as often as two locking, which can pin the rope against the rock while lowering if the climber's (weighted) strand is on the outside, away from the rock.

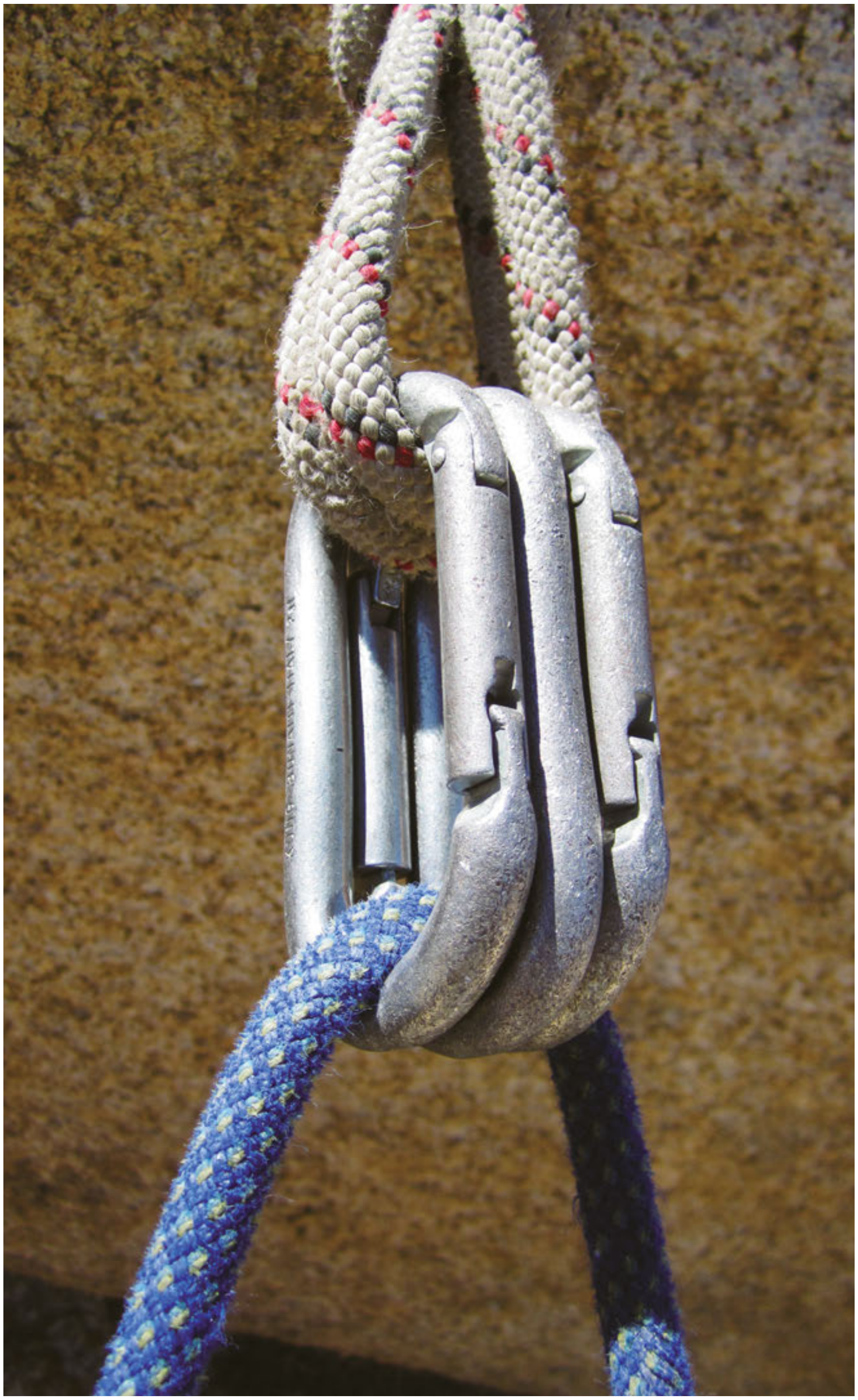

Three steel ovals with the gates opposed and reversed at a toprope anchor.

What's in the Pack?

I (BG) prefer a pack that has accessory straps and buckles so I can strap a rope on the outside of the pack. In addition to my rack, slings, cordelettes, belay devices, helmet, shoes, harness, food, water, headlamp, knife, and extra clothing, I carry two first-aid kits. One is a small ''patch kit'' with small bandages, tape, and a blister kit for minor cuts, scrapes, and blisters. Inside this one I'll also have NSAIDs (aspirin and ibuprofen), acetaminophen, and Benadryl. I also carry a roll of athletic tape. For client care I carry extra sunscreen, lip balm, a trash bag, and a squirrel-proof container for food storage.

My larger first-aid kit is a trauma kit complete with trauma shears, heavy absorbent pads, tape, large bandages, roller gauze, a pocket mask, a couple of Sam splints, and two epi pens.

An instructor's kit should include a communication device (e.g., cell phone, satellite phone, or walkie-talkie), pen and paper (SOAP notes if you're a WFR), and a written emergency response plan with standard operating procedures and appropriate actions for specific scenarios.

Retire a carabiner if it shows a groove from excessive rope wear, or if the gate is no longer in alignment with the body of the carabiner after it has been dropped a considerable distance. If the gate is sticky, washing it with soap and water and using some graphite lubricant will usually take care of the problem.

In the professional realm, the industry standard for attaching the climbing rope to the toprope anchor master point is either two locking carabiners or three oval carabiners with the gates opposed and reversed.

When doing a lot of toproping, aluminum carabiners actually wear rather quickly, developing noticeable grooves. When this happens, they should be retired. Petzl recommends retiring a carabiner with 1mm of wear, about the thickness of a dime. Also, worn-off aluminum particles get on the rope and on the belayer's hands. Using steel ovals solves this problem, as steel is far more durable and wears much more slowly than aluminum.

The Instructor's Pack

A climbing instructor is constantly striving to select the perfect tool for a given task. The instructor's pack is no exception. It should be large enough to hold everything an instructor will need to run an outing and be prepared for emergencies. For most, the best pack is large, about 4,000 to 5,000 cubic inches, and any items that do not fit inside the pack (like a rope or a helmet) are strapped down firmly. A climbing instructor should be able to move through terrain with grace and balance, without swinging or dangling encumbrances, prepared to deal with all contingencies.

CHAPTER 6

Knots and Hitches

What knot should I use here? What hitch should I use there? The answer isn't always obvious. Many knots and hitches accomplish the same thing and work equally well in similar applications. Others do not.

An instructor's job is first to determine which knot or hitch is appropriate for the technical task at hand and, second, to develop strategies for teaching knots and hitches appropriate to a given lesson.

The following pages discuss a wide array of baseline knots and hitches used regularly by Single Pitch Instructors. Certainly there are many more that could be acceptably used in a single pitch setting or taught to students with specific learning objectives, but these represent a fundamental standard of knowledge that every instructor should hold.

Loop Knots

Loop knots are tied by taking two strands of rope (called a bight) and wrapping them back over themselves so that the knot does not slide, or by taking the end of the rope and tying it back over the standing part so the knot does not slide. Loop knots are used to clip the rope into a carabiner, or to tie around an object.

Overhand Loop

This is the simplest knot used to form a loop. It requires less rope to tie than the figure eight. For most applications, however, the figure eight loop is preferable because it tests slightly stronger than the overhand loop and is easier to untie in small-diameter cord.

Knot Terminology

Bend: Two ropes tied together by their ends.

Bight: Two strands of rope where the rope is doubled back on itself.

Load strand: The strand of the rope that bears all the weight.

Hitch: A knot that is tied around another object (such as a carabiner or rope).

Standing end: The part of the rope that the end of the rope crosses to form a knot.

Tag end: The very end of a rope, or the tail end that protrudes from a knot.

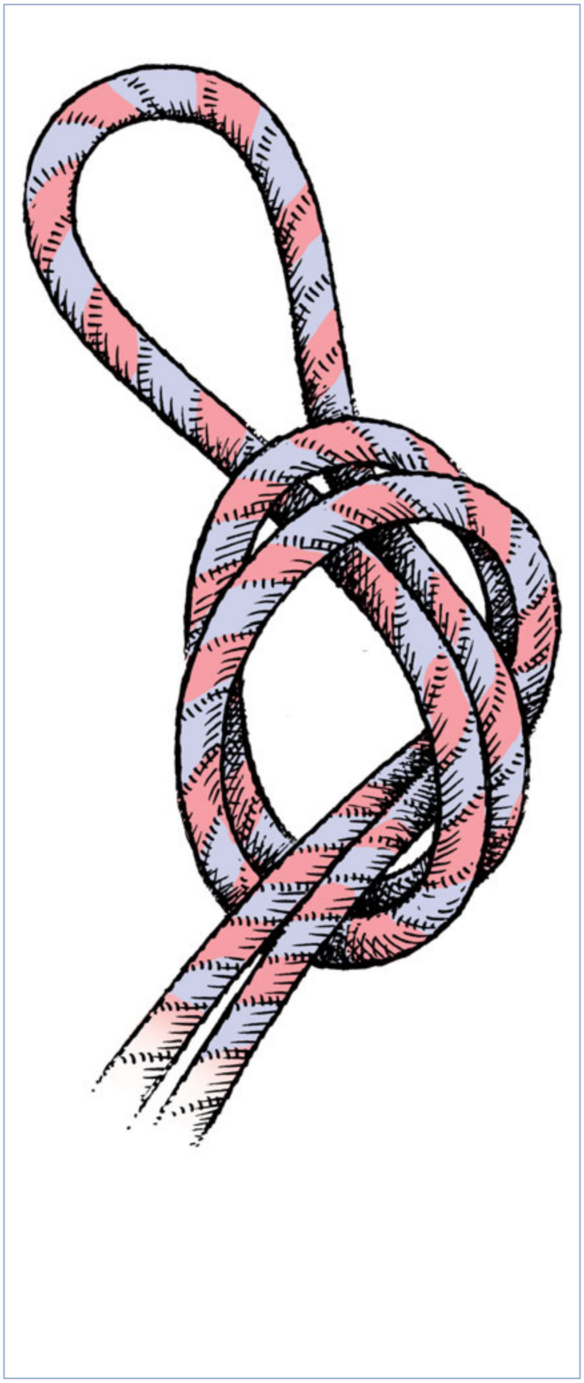

Overhand loop.

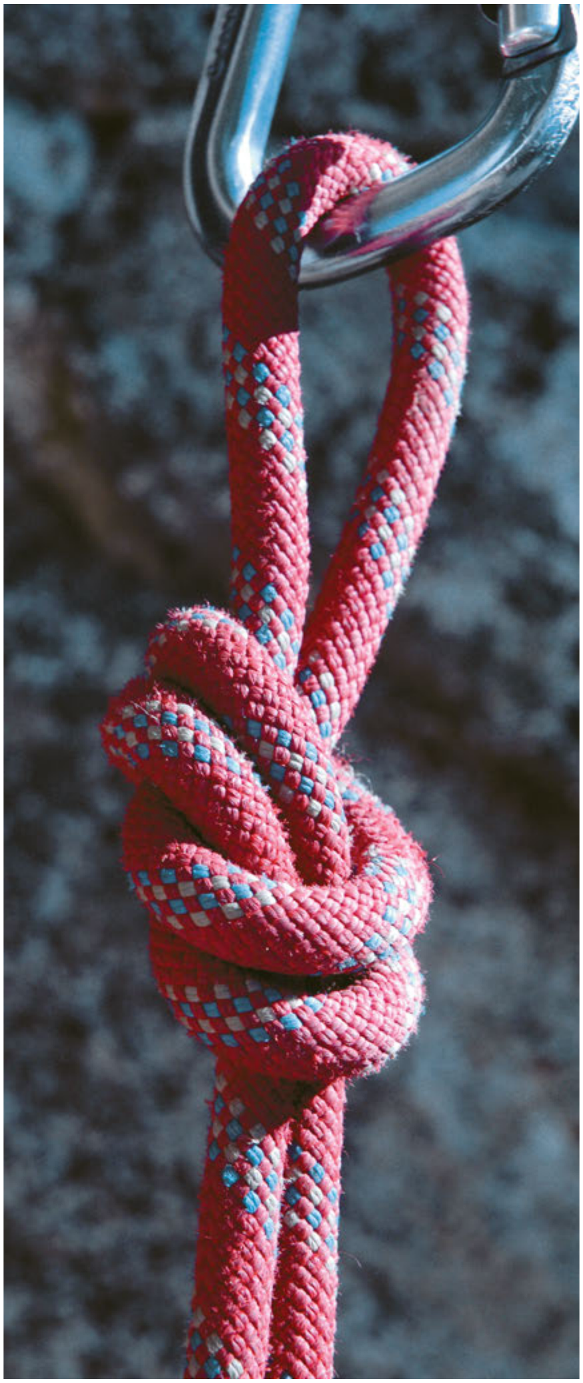

Finished overhand loop.

Figure Eight Follow-Through

This is the standard knot that climbers use for tying the rope to their harness. It can also be used to tie a rope around an object like a tree or through a tunnel. Climbers have gravitated toward this knot over the years because it is strong, secure, and easy to recognize when tied correctly. Instructors like it because it is easy to teach, and the uniform nature of the knot enables quick inspection and supervision. When tied correctly, the knot is tight, with a 5- to 8-inch tail, and it has bilateral symmetry because the initial figure eight and its follow-through strand are always parallel.

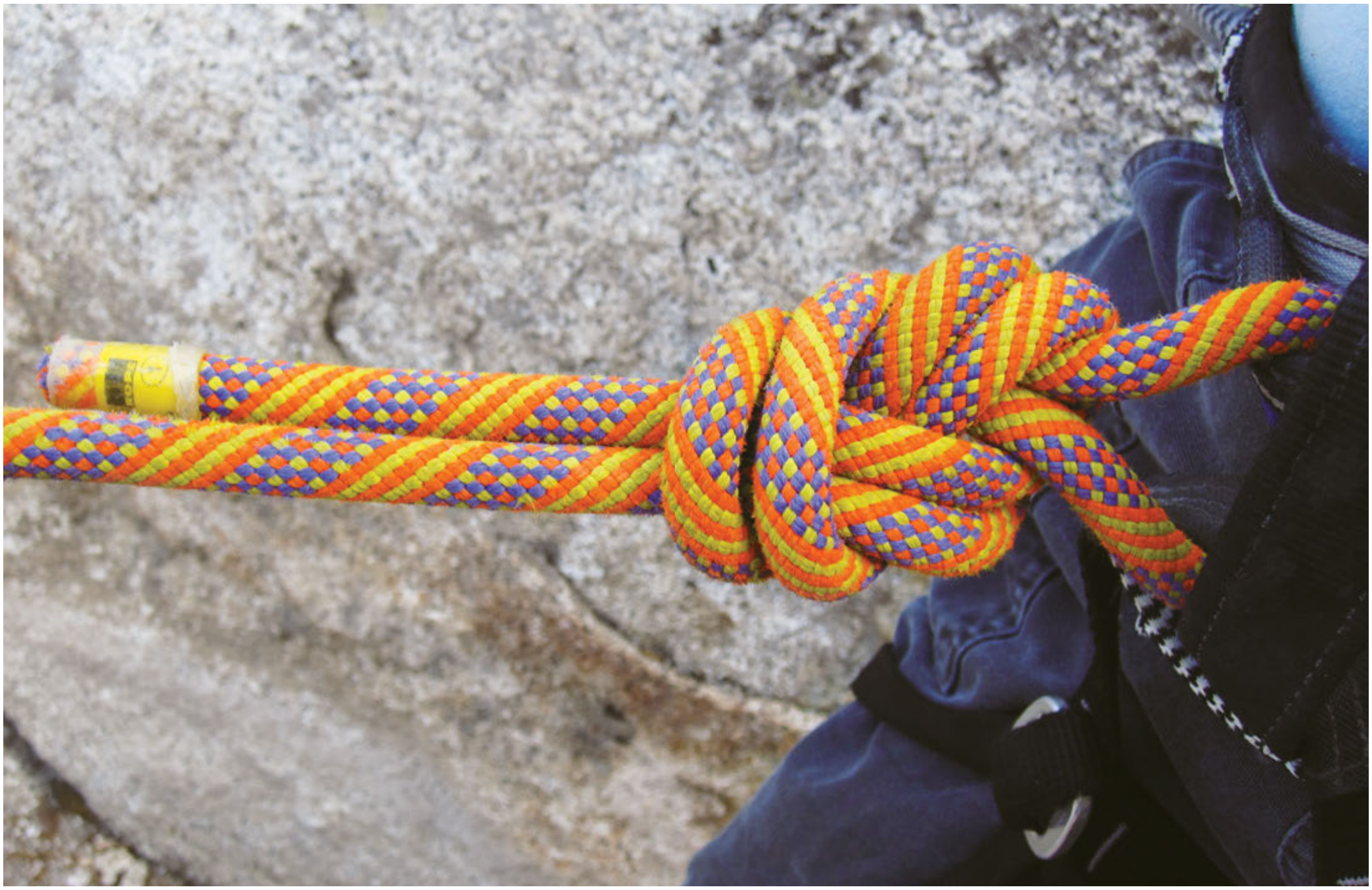

A well-dressed figure eight follow-through has bilateral symmetry. The original strand of rope from the initial figure eight and the follow-through strand are parallel through each curve of the knot. As a result, if a viewer were to draw an imaginary line through the middle of the knot, vertically or horizontally, the halves of the knot would reflect each other in shape, size, and general appearance. This bilateral symmetry gives the knot its signature uniformity, making it easy to recognize from a distance.

Tying In with a Figure Eight Follow-Through

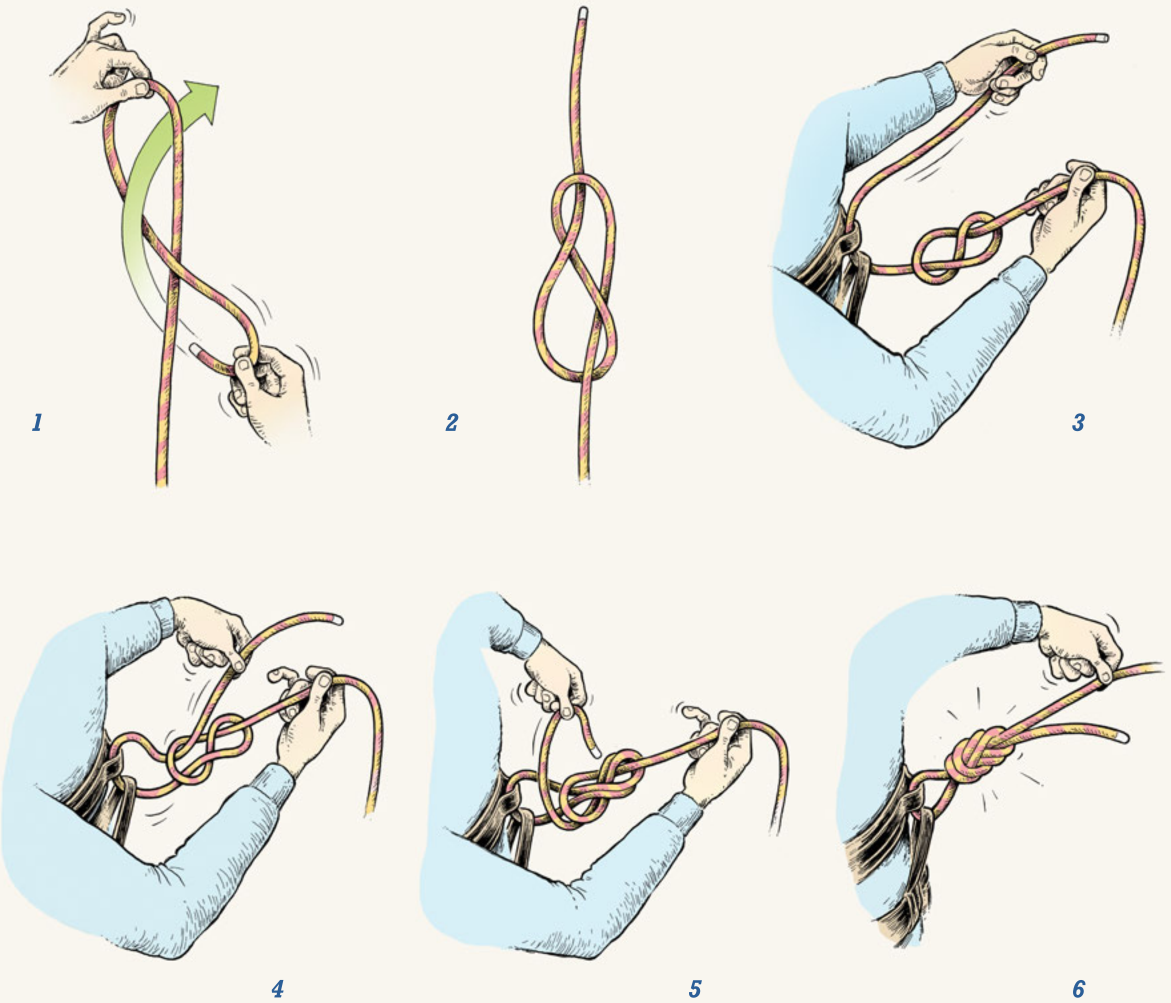

Check the harness manufacturer's guidelines for information on how to properly tie the rope to the harness. For harnesses with belay loops, generally follow the same path as the belay loop, which goes through two tie-in points on the harness. Tie the figure eight so that its loop is about the same diameter as your belay loop. The figure eight knot does not require a backup knot.

Figure Eight Loop

Another standard climbing knot, the figure eight loop is used for tying off the end of a rope, or for tying a loop in the middle, or "bight," of a rope. It is also commonly referred to as a "figure eight on a bight."

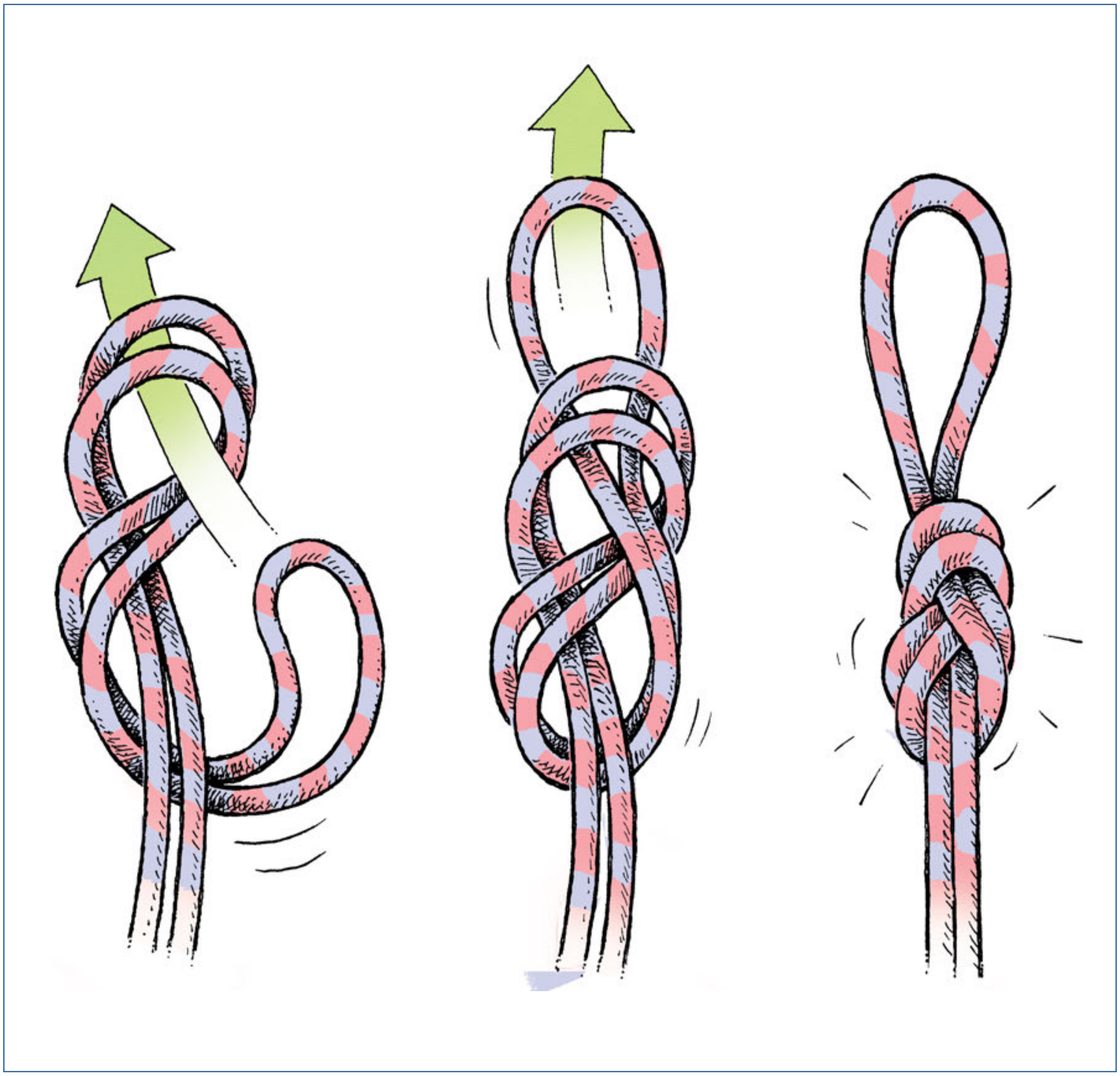

How to tie a figure eight loop.

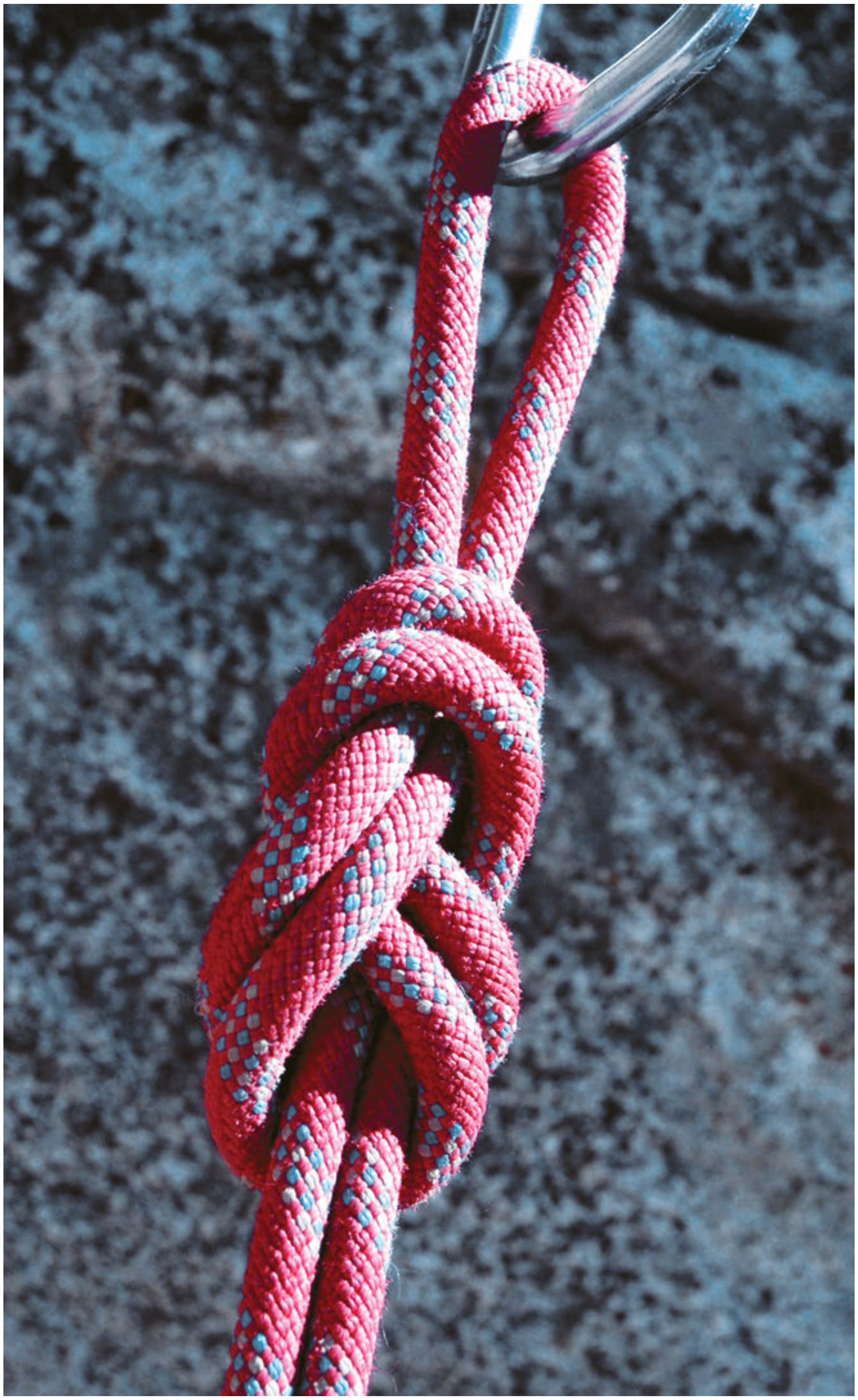

Finished figure eight loop clipped to an anchor.

Tying the In-line Figure Eight (aka Directional Figure Eight)

In-line Figure Eight

Also known as a directional figure eight, this knot, like the clove hitch, can be used to tie off a series of anchors in a line. It takes some practice to master this knot, but some find it easier to use than a clove hitch.

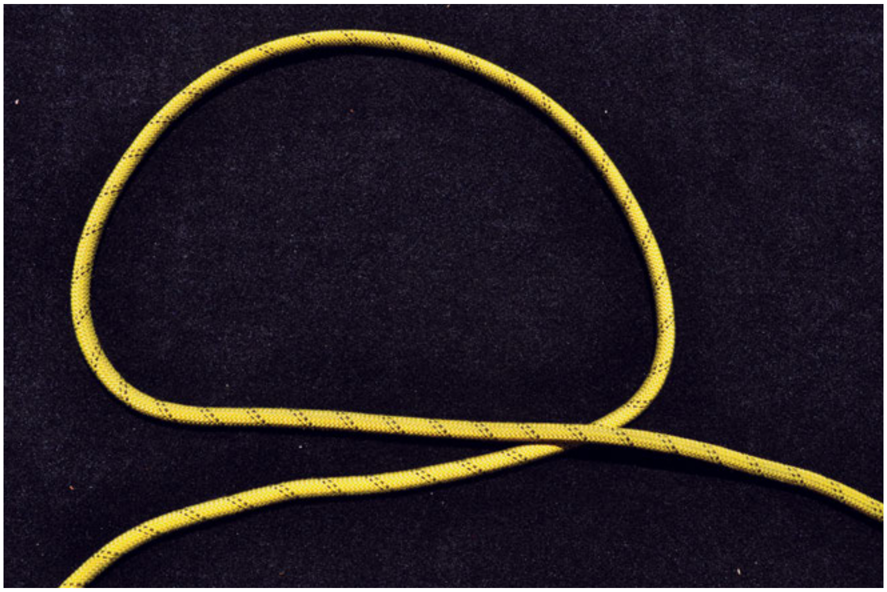

Cross the strands to form a simple loop.

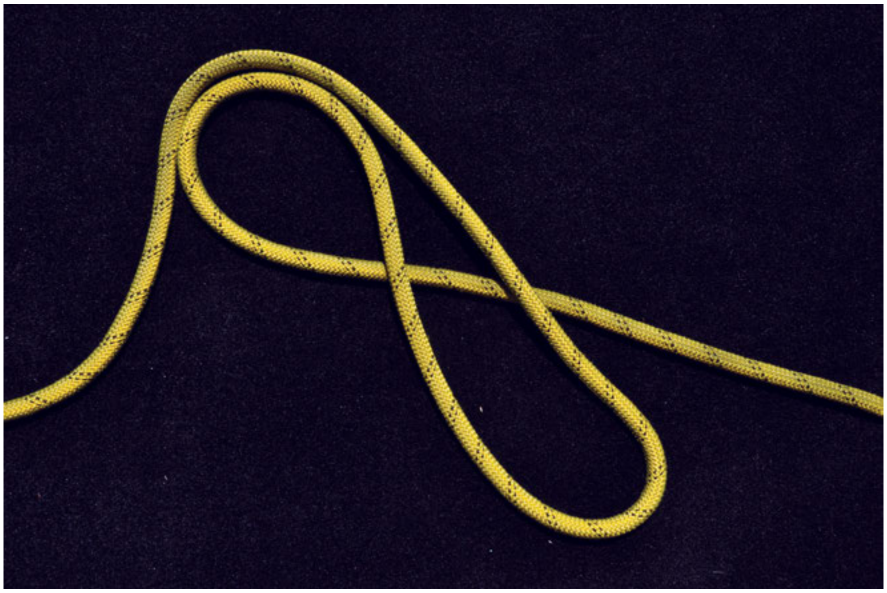

Cross a bight over the single strand.

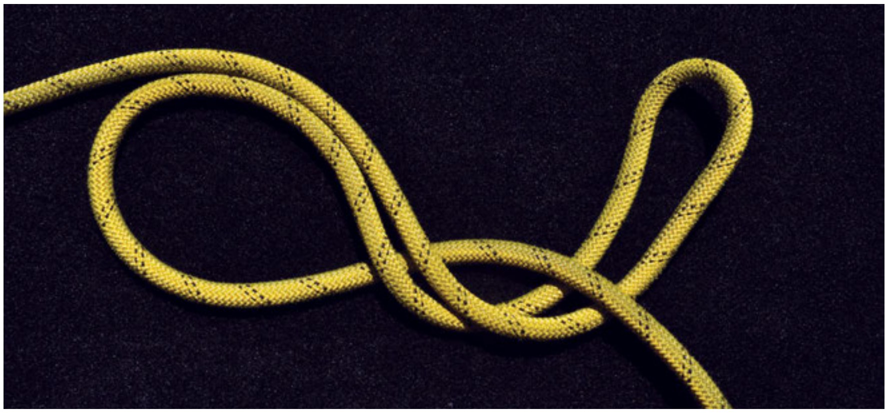

Cross the bight under the strand.

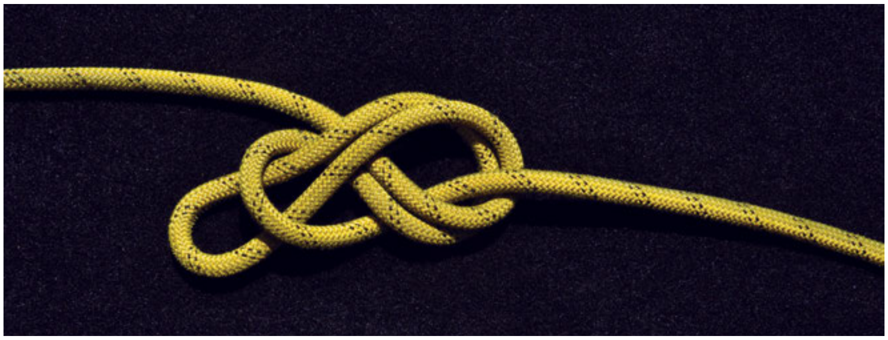

Thread the bight back through the loop you've just formed.

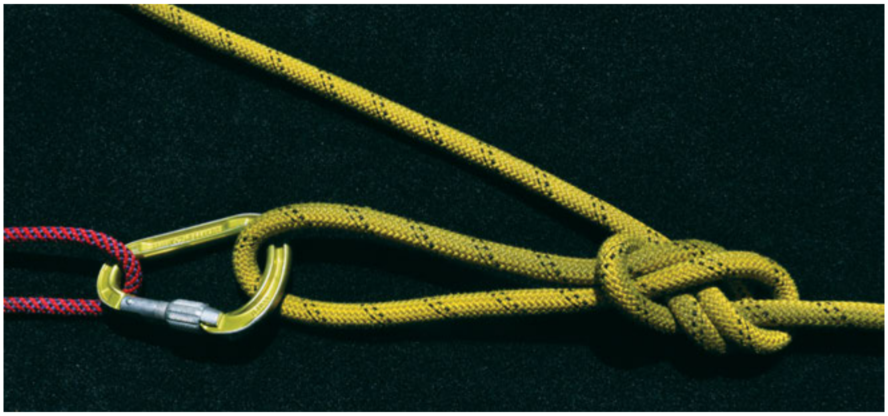

Finished in-line figure eight.

Bowline

Boy Scouts were taught this knot with the saying "the rabbit comes up through the hole, around the tree, and back down through the hole." The bowline is very useful to tie the rope around something, like a tree, block of rock, or tunnel in the rock. It is important to note that a bowline knot requires a backup, as weighting and unweighting the knot easily loosens it. Always tie half of a double fisherman's knot to back it up. One advantage of the bowline is this same feature—it is very easy to untie after it has been weighted, so it is used regularly by professional riggers.

Tying the Bowline

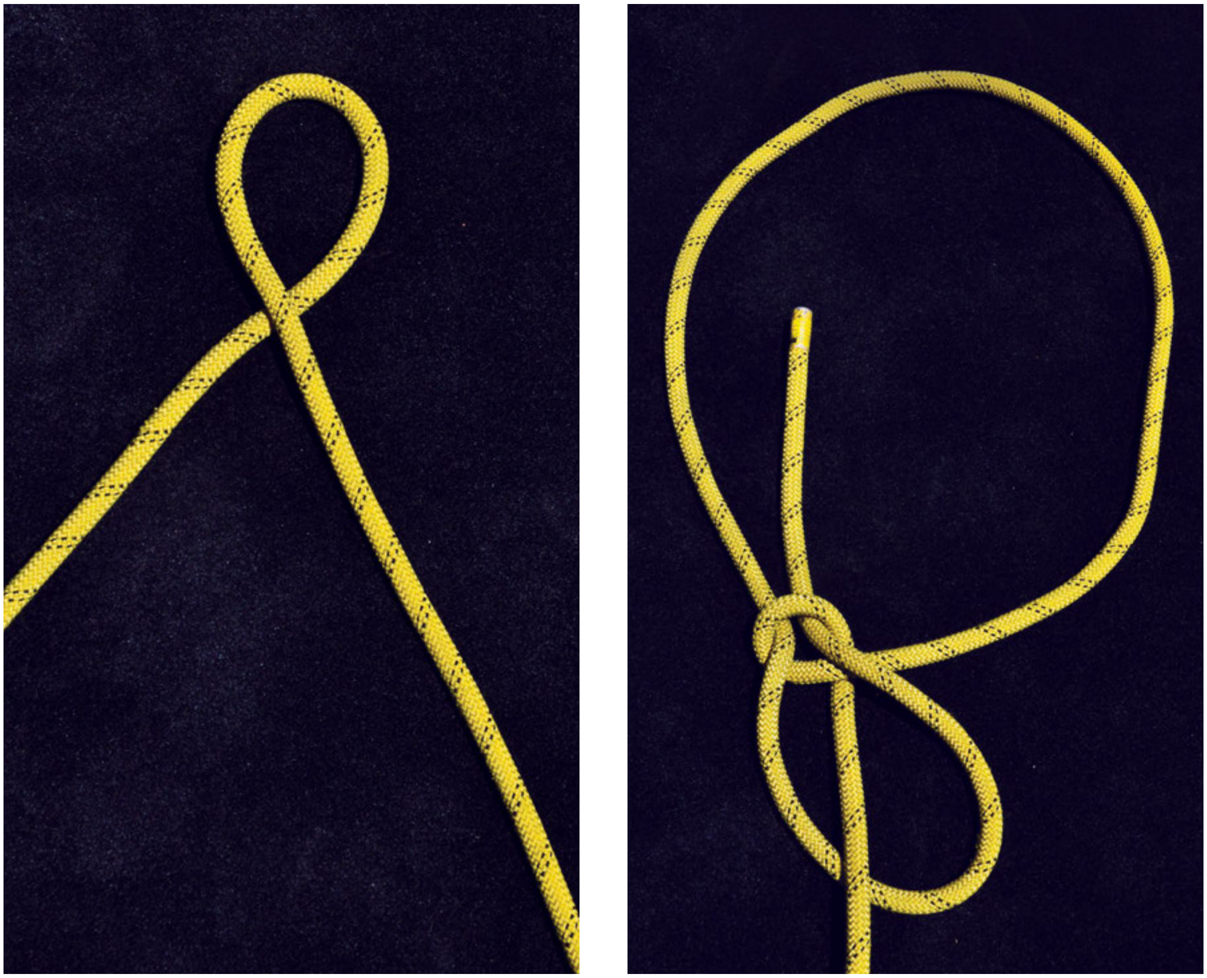

Tying the bowline. The bowline should always be tied with a backup, shown here with half a double fisherman's for the backup knot (final photo).

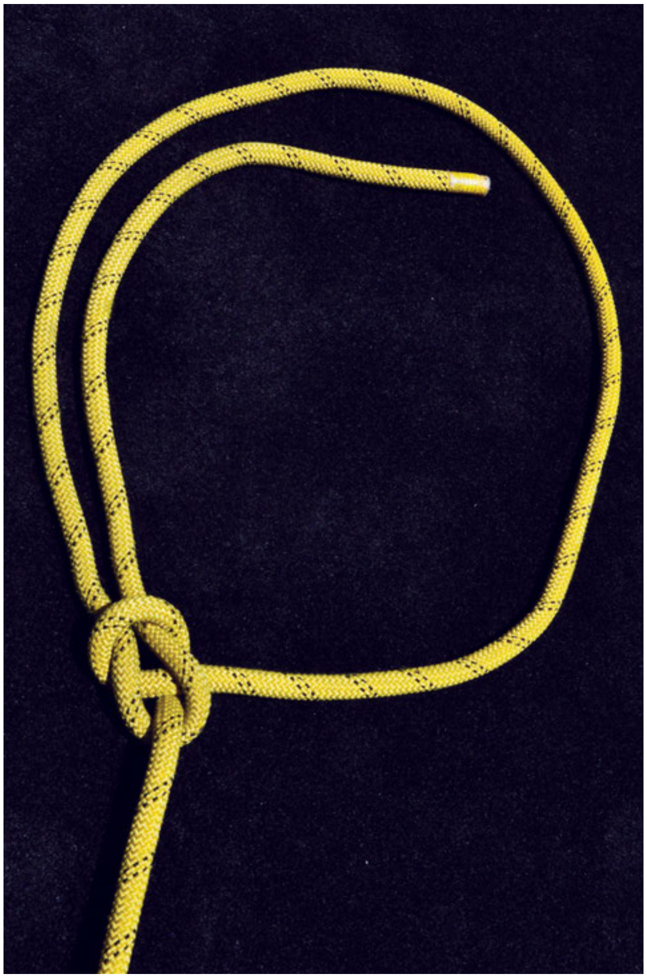

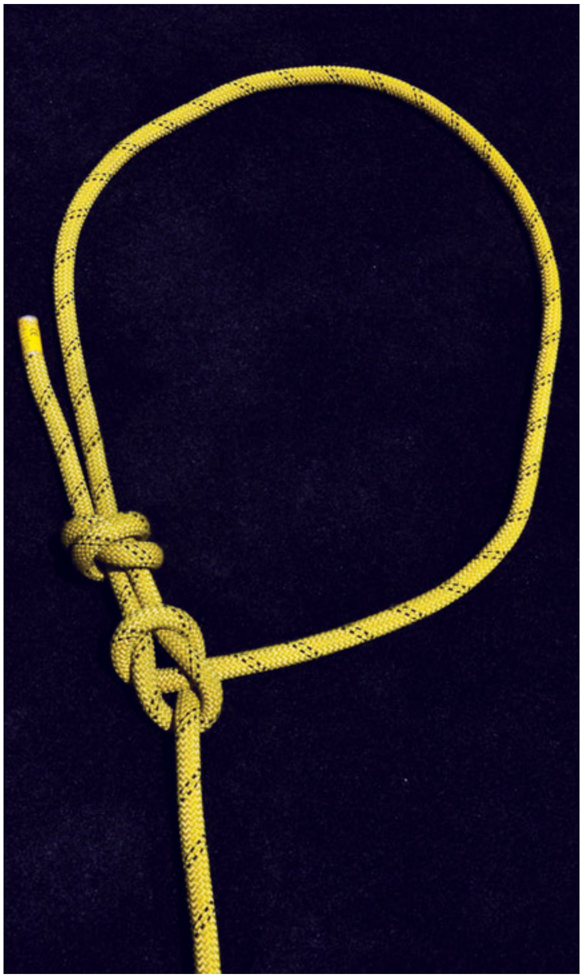

Double Loop Knots

BHK

The BHK, an acronym for “big honking knot,” is technically a double overhand on a bight, used to create a redundant master point. It is commonly used for toprope anchor extensions and master points using rope.

Tying the BHK (“Big Honking Knot”)

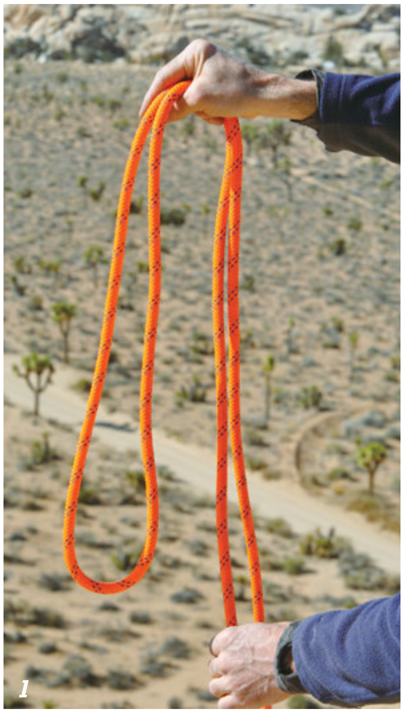

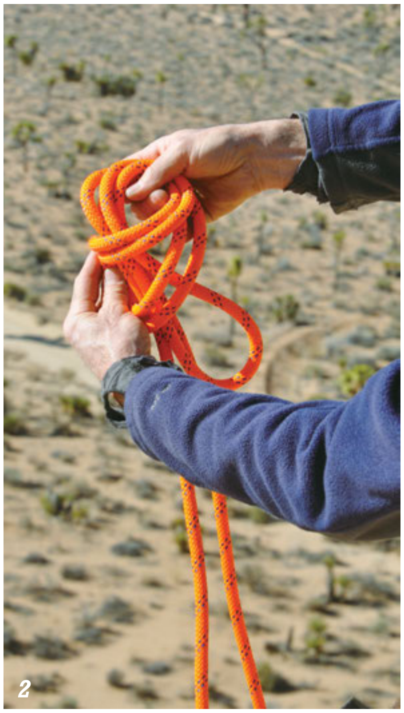

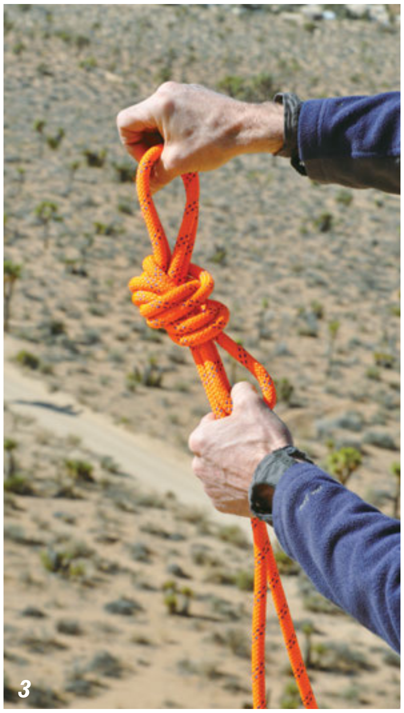

1 Start by taking a doubled bight about 4 feet long.

2 & 3 Then tie an overhand loop on all four strands.

Master point toprope extension using a BHK and three steel oval carabiners with the gates opposed and reversed.

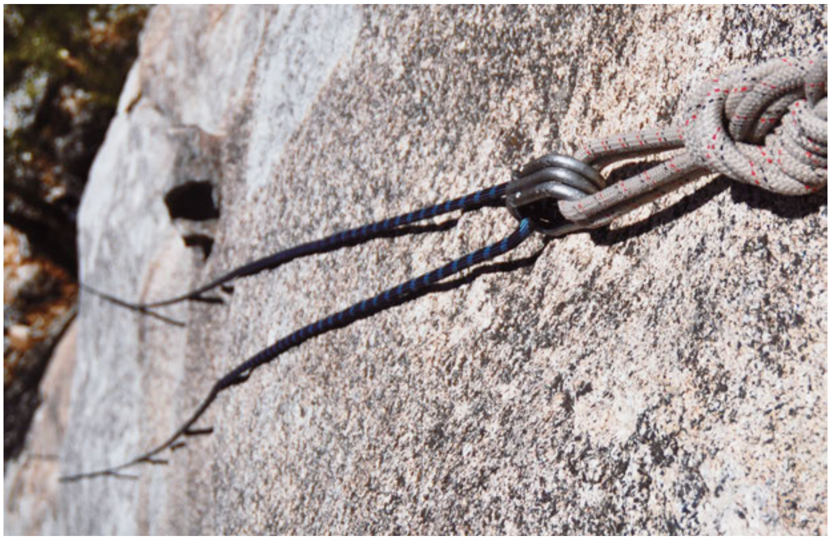

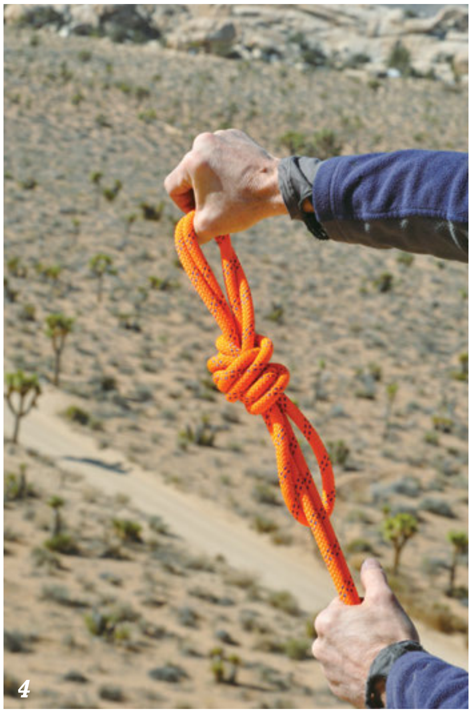

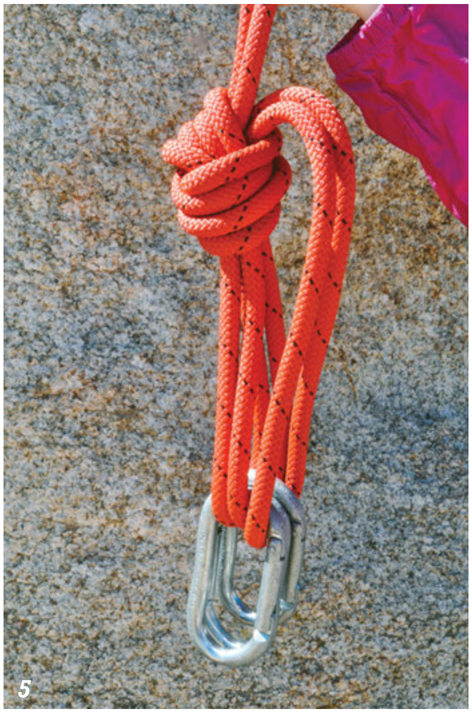

4 & 5 For greater security, take the single loop and hitch it all the way back around the body of the two-loop knot or clip it into the carabiners on the master point.

Double Loop Figure Eight

This knot is used primarily to clip in to or equalize two anchor points. If the two loops are used together at a master point, however, they are not redundant, due to the configuration of the knot.

Tying the Double Loop Figure Eight

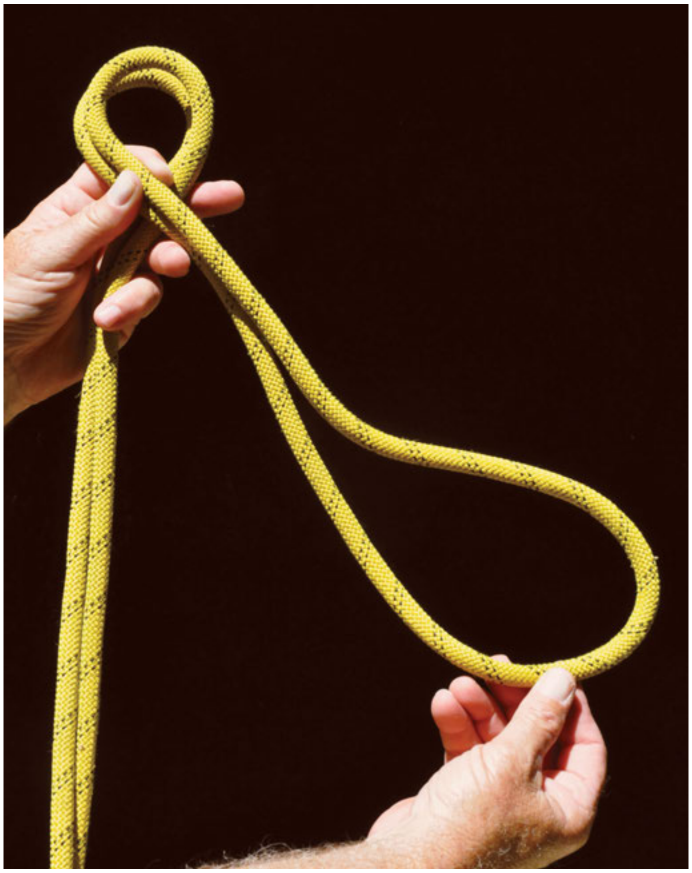

Take a bight of rope and cross it back over itself, forming a loop.

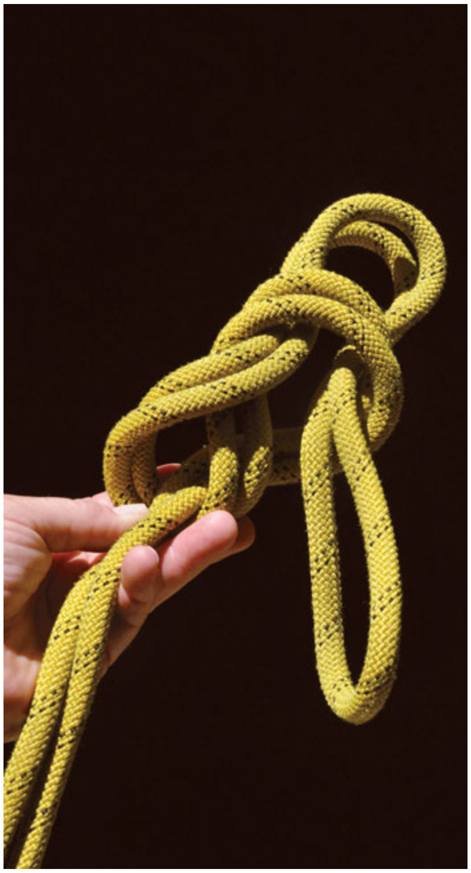

Take two strands of the bight and wrap them around the standing part, then poke them through the loop.

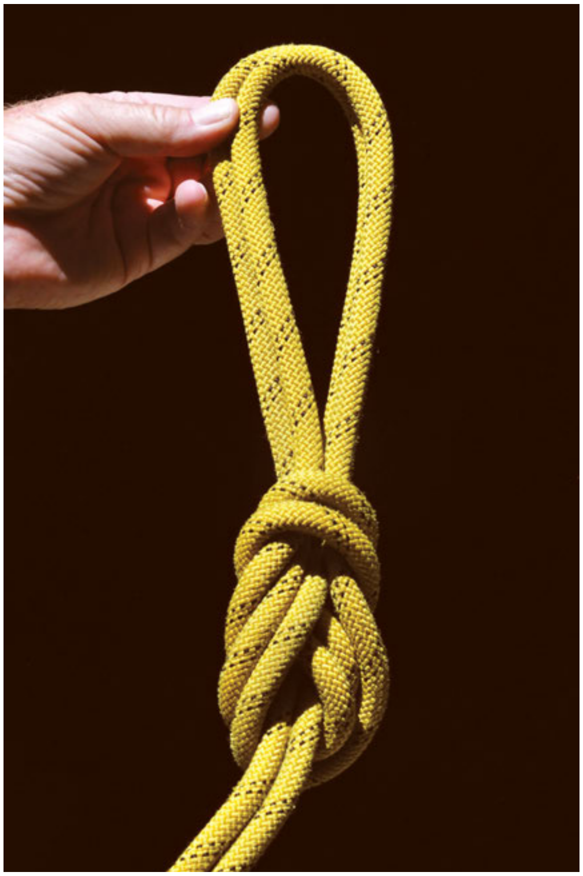

To finish, take the loop at the very end of the bight and fold it down and around the entire knot you've just formed.

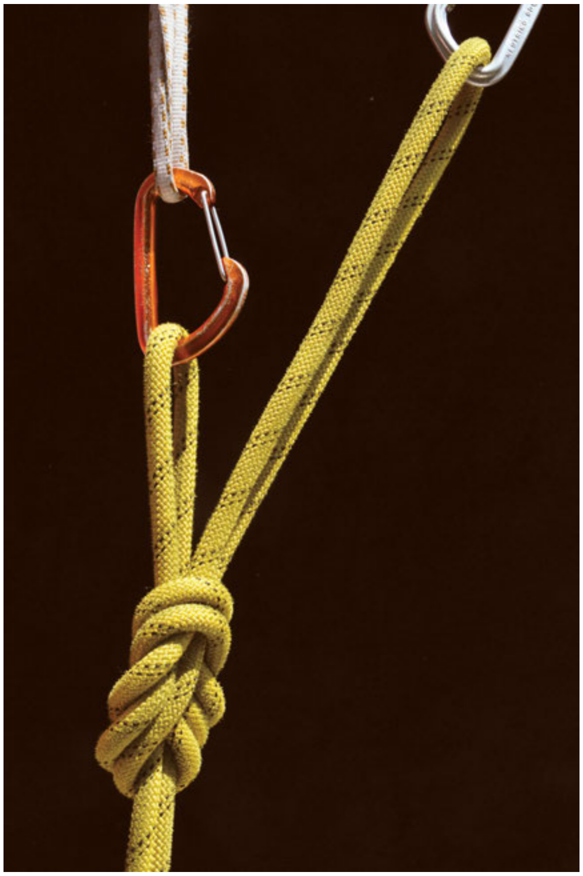

The double loop figure eight is a great knot to use to equalize two gear placements. You can manipulate the knot by loosening one strand and feeding it through the body of the knot, shortening one loop, which makes the other loop larger.

Double Loop Bowline

This knot is handy for equalizing two anchor points or clipping in to a two-point anchor. It can shift when weighted, so always back it up with a half a double fisherman's if the tail end is near the body of the knot.

Tying the Double Loop Bowline

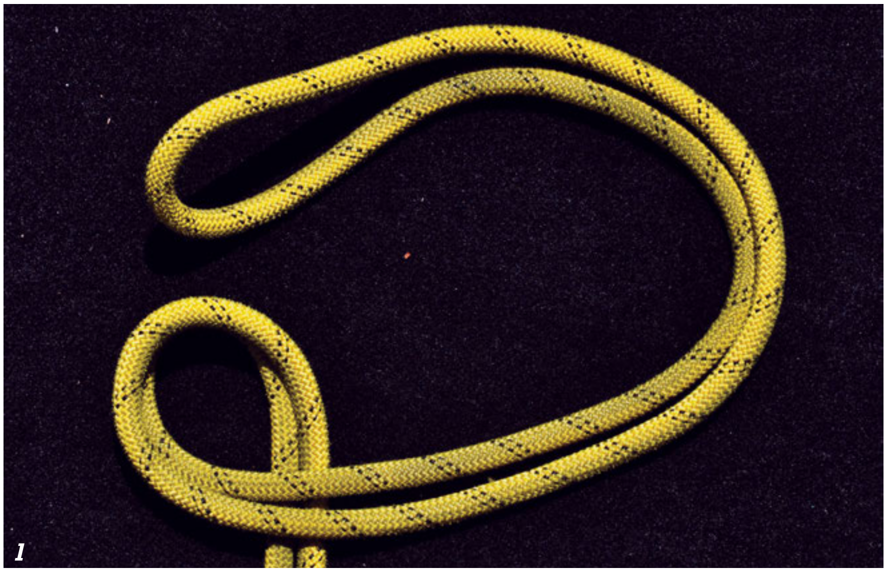

Take a bight of rope and cross it over the standing part.

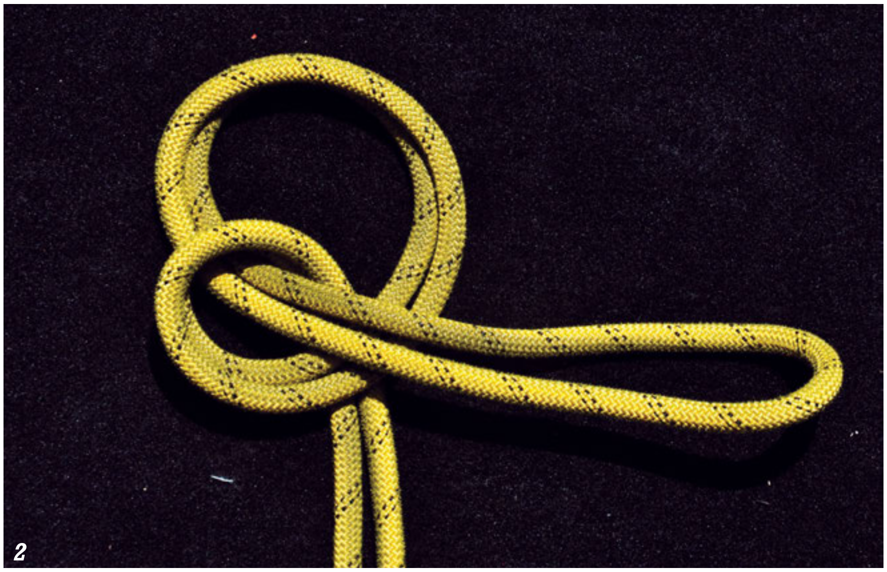

Thread the bight through the loop you've just formed.

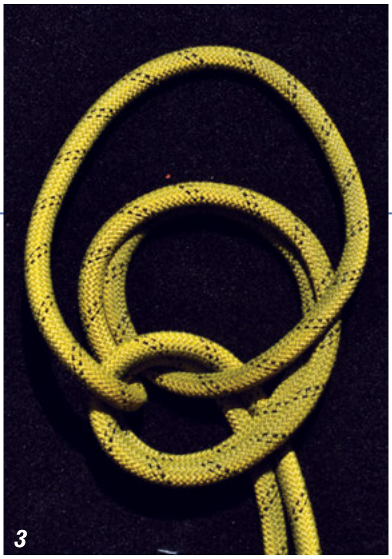

Configure the end of the bight in a loop above the rest of the knot.

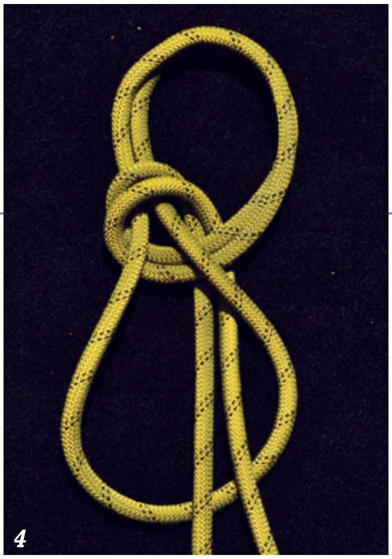

Flip the loop down like a hinge behind the rest of the knot.

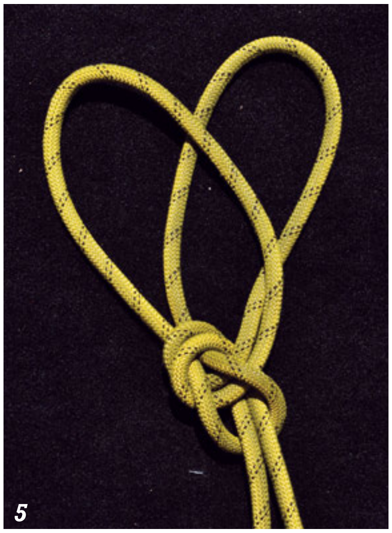

Pull on the two loops until the end of the bight tightens at the base of the knot.

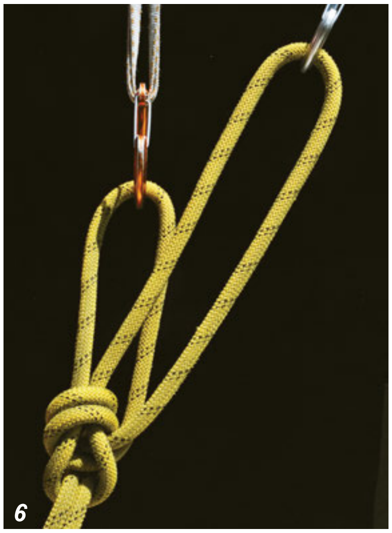

The two loops can be adjusted by feeding one strand into the body of the knot, which alternately shortens one loop and lengthens the other.

Knots for Webbing

Nylon webbing is a slick material that should be tied with caution: There have been many accidents where poorly tied knots in nylon webbing have failed. The two recommended knots for tying nylon webbing into a loop are the water knot (also known as the ring bend) and the double fisherman's knot (also known as the grapevine knot). When tying the water knot, your finished tails should be a minimum of 3 inches in length. It is important you tighten the water knot properly, as it has a tendency to loosen if tied slackly in a sling that is being used over time.

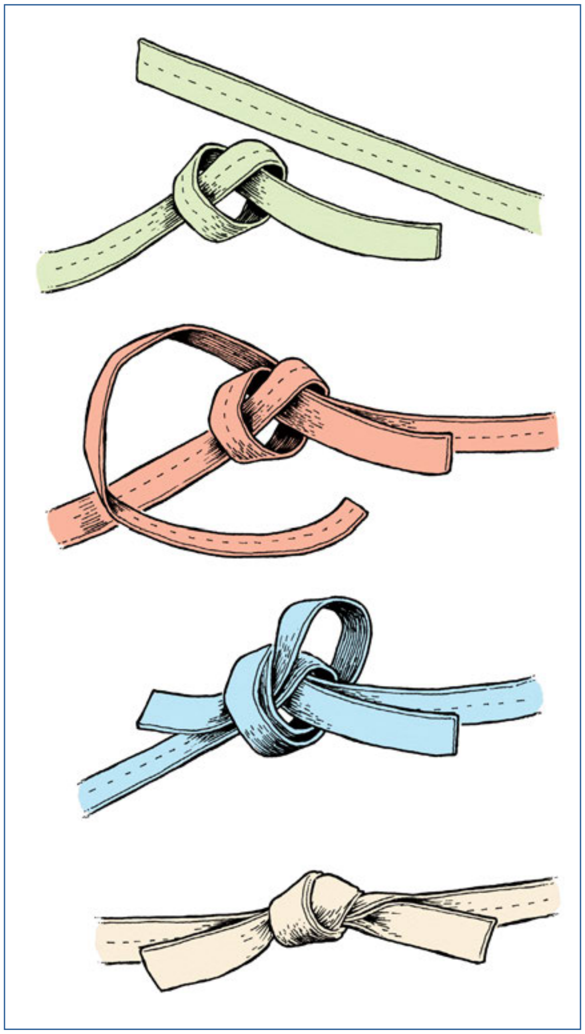

Tying the water knot (ring bend).

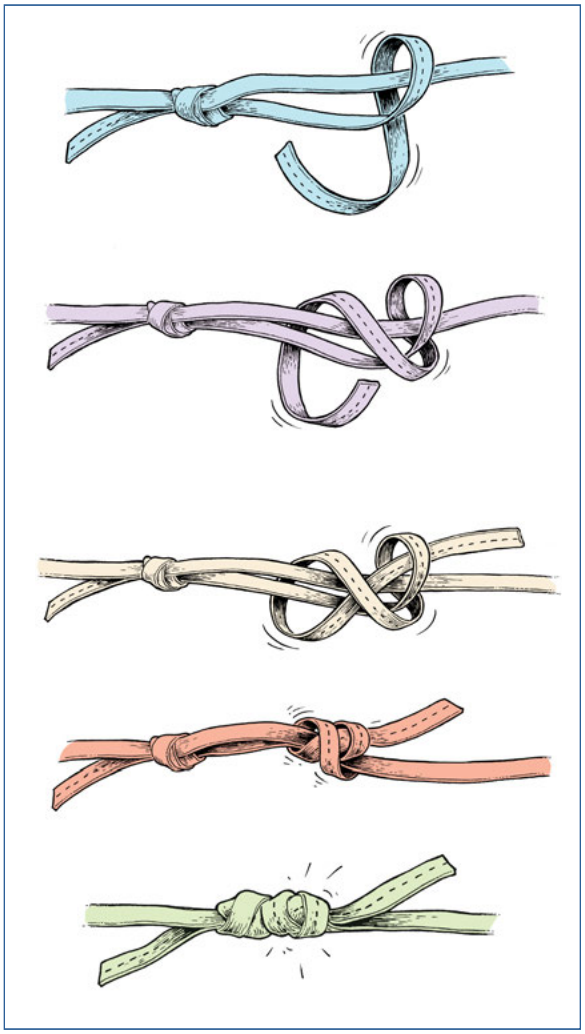

Tying nylon webbing with a double fisherman's (grapevine) knot.

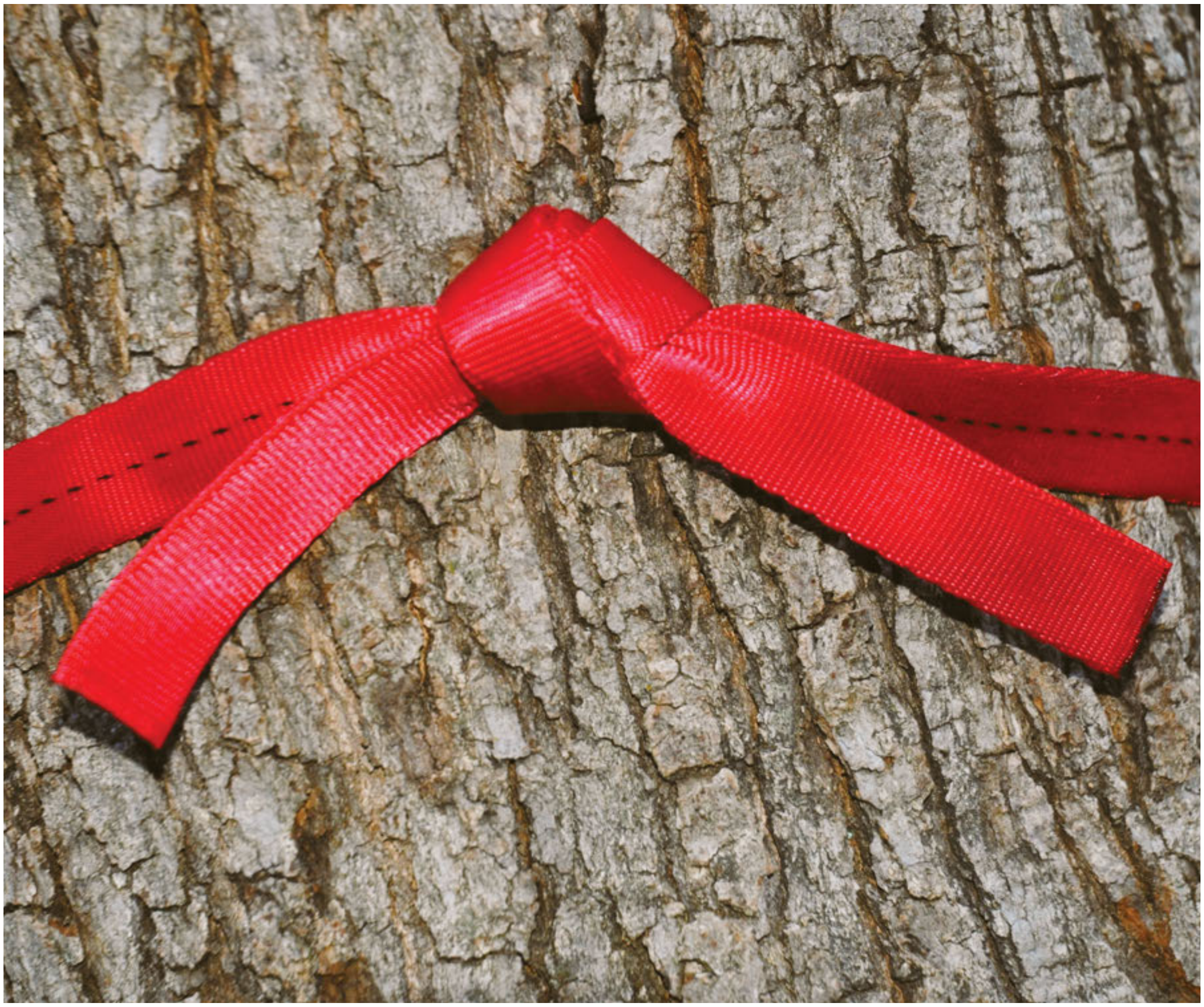

The water knot.

Why would you ever use nylon webbing tied with a knot as opposed to a sewn runner? A sewn nylon runner is stronger than the same material tied with a knot. The answer is for rappel anchors when the slings are tied around a tree or through bolt hangers. It is also sometimes useful to untie the knot, thread it through something (like a tunnel), and re-tie it.

The double fisherman's knot is also a good knot to use to tie nylon webbing into a loop, although it does require more length of material to tie and is very difficult to untie after it has been seriously weighted.

Bends

A bend is a knot that joins two ropes or lengths of cord together. These knots are used to tie your cordelette into a loop, and also to tie two ropes together for rappelling.

Figure Eight Bend (aka Flemish Bend)

A variation of the figure eight follow-through, this knot can be used to tie two ropes together. It has superior strength and is easy to untie after it has been weighted. It is simply a retraced figure eight. On 9mm to 11mm diameter rope, tie it with the tails a minimum of 5 inches long.

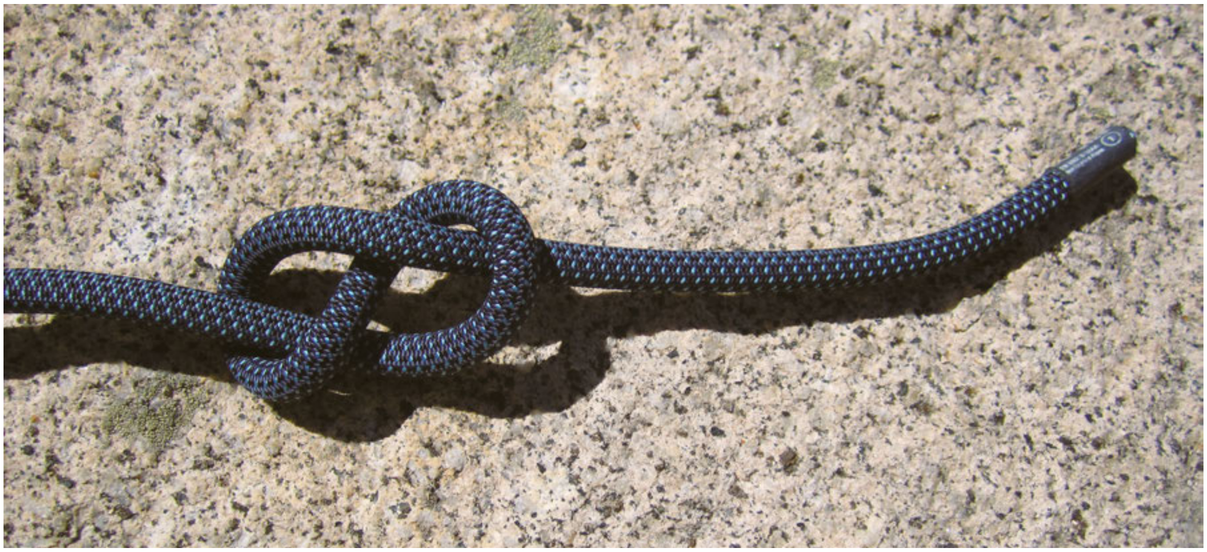

The figure eight bend.

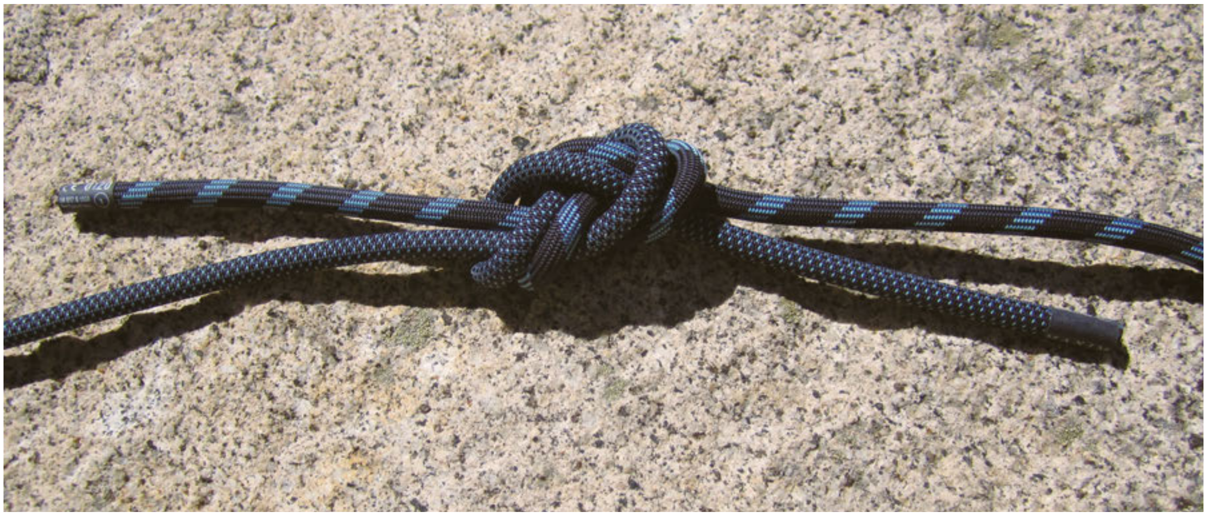

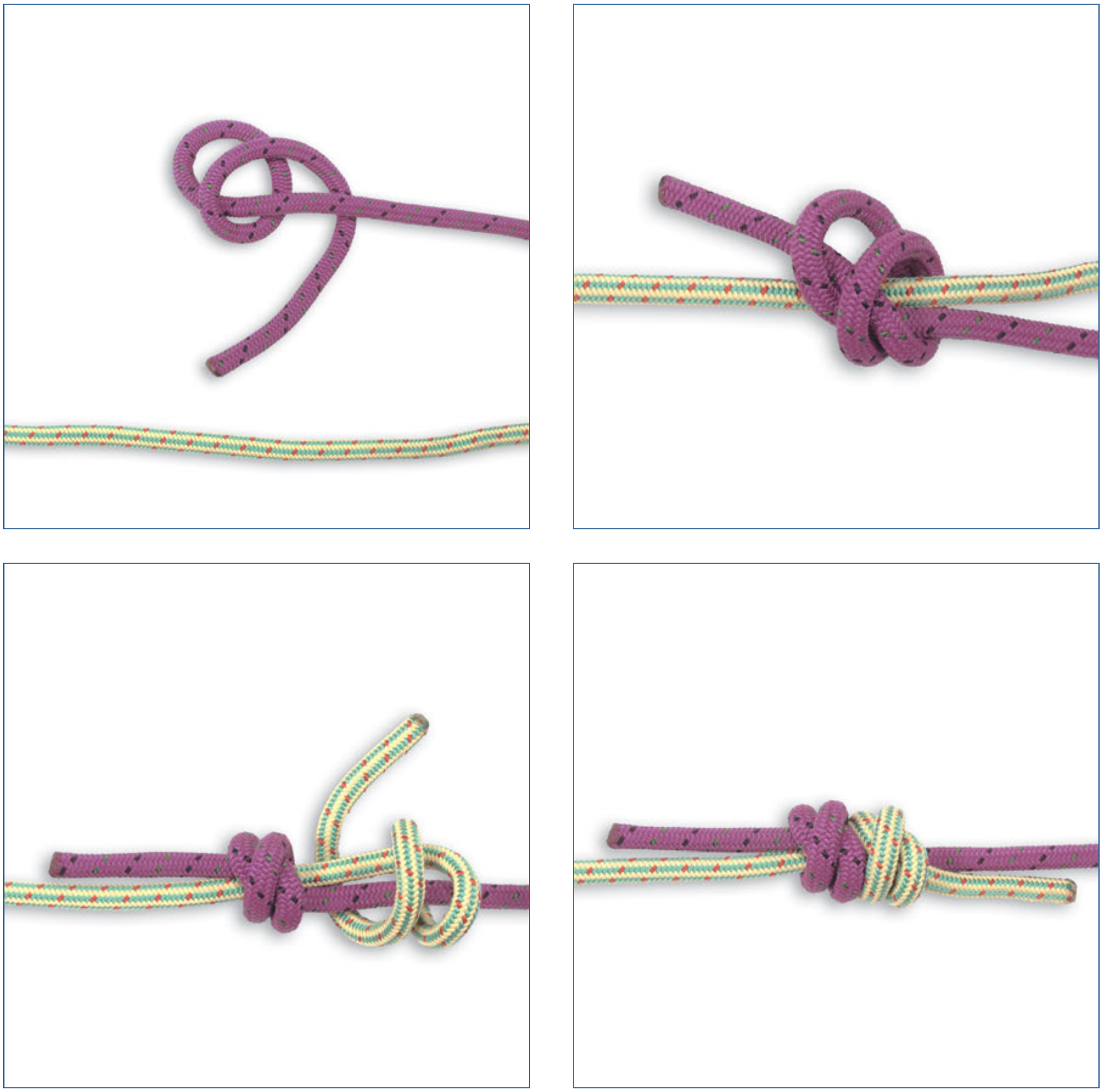

Tying the double fisherman's knot (aka grapevine knot). When tying 7mm nylon cord, leave the tails about 3 inches long.

Double Fisherman's Knot

This is the preferred knot to use for joining nylon cord into a loop to make a cordelette or prusik loop. It is also a very secure knot to tie two ropes together for a double-rope rappel, but can be difficult to untie.

Triple Fisherman's Knot

For 5mm and 6mm diameter high-tech cord (i.e., Spectra, Dyneema, Technora), a triple fisherman's knot tests slightly stronger than the double fisherman's. (Moyer, Tom, Paul Tusting, and Chris Harmston [2000]. http://user:xmission.com/~tmoyer/testing/High_Strength_Cord.pdf. 2000 International Technical Rescue Symposium.)

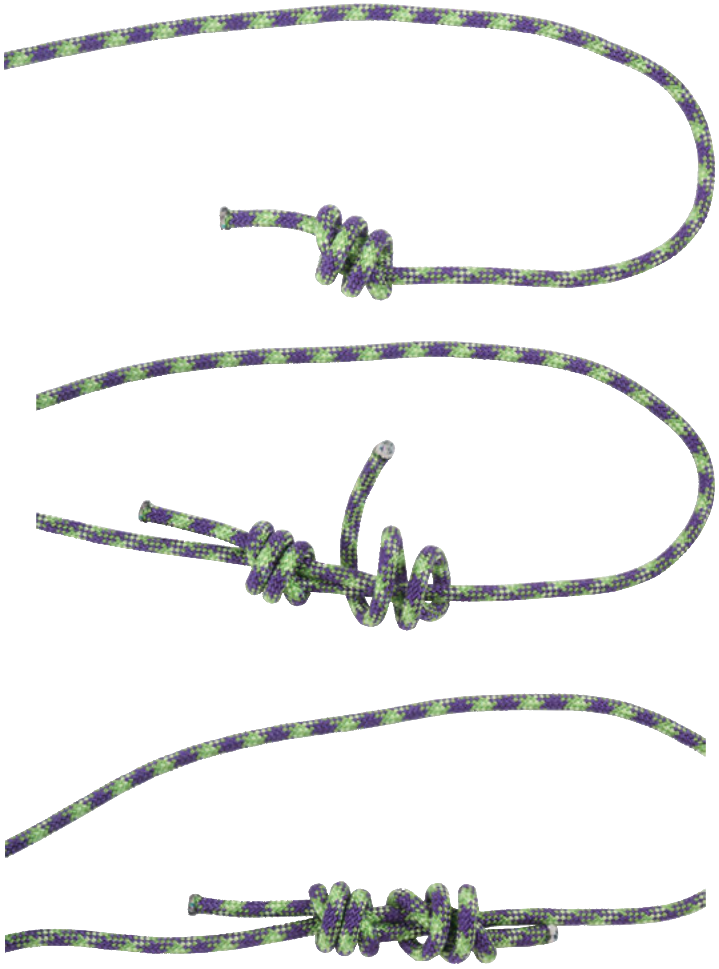

To tie a triple fisherman's, make three wraps before feeding the cord back through.

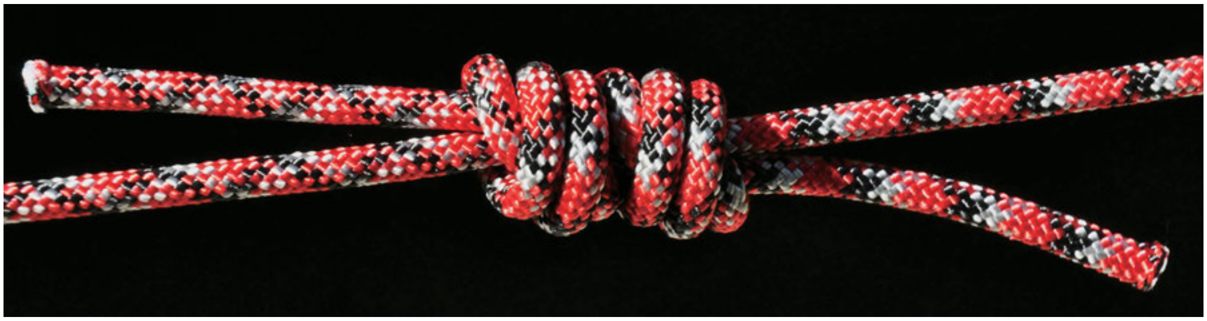

The triple fisherman's knot.

Knots for Joining Two Ropes

Standard knots for joining two ropes include the double fisherman's knot and the figure eight bend. The double fisherman's is more difficult to untie than the figure eight bend once weighted; the figure eight bend, while relatively easy to untie, is bulky. Tie these knots with a minimum of 3 inches of tail, and carefully tighten the knots before using them. A stiff rope makes it harder to cinch the knots tight, so be especially careful with a stiffer rope.

Which knot to use should be based on several variables. If the ropes differ drastically in diameter, or are very stiff, the most foolproof knot is the figure eight bend, backed up with half a double fisherman's on each side. This is a bulky knot, but it provides a real sense of security.

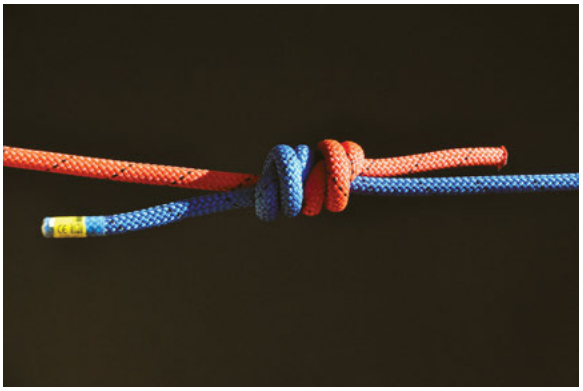

Double fisherman's knot.

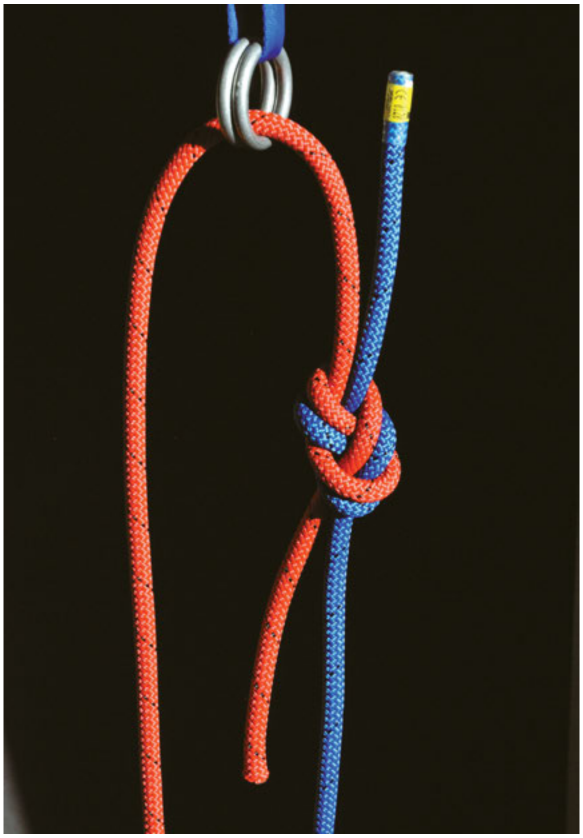

Figure eight bend.

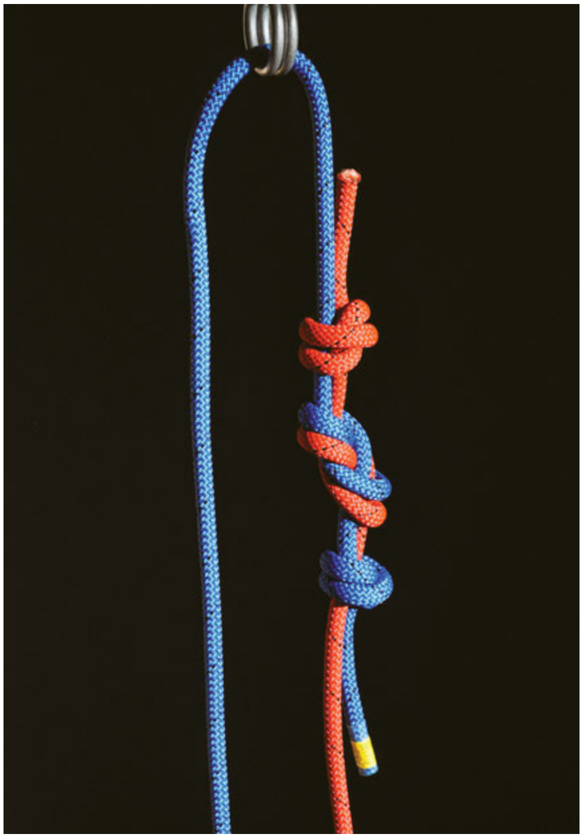

Figure eight bend with fisherman's backups.

Flat Overhand (aka Euro Death Knot)

How this knot received the "Euro Death Knot" moniker is unclear. Most likely the knot was initially adopted by Europeans and deemed unsafe when first seen by American climbers unfamiliar with its use. In reality, it is rarely responsible for incident or injury. In one such rappelling accident in recent times (in the Tetons, September 1997), the flat overhand failed when it was sloppily tied with too short of a tail. Ironically, former *Rock and Ice Magazine* editor George Bracksieck wrote in July of that year that "the one-sided overhand knot (tails parallel and together) remains the best knot for rappelling. . . . Be sure to leave plenty of tail and to set it snugly." After analyzing the accident, Grand Teton ranger Mark Magnusun wrote: "I intend to do some additional research in an effort to gain information on the overhand knot used for joining ropes, the origin of the 'Euro-death' nickname, and incidents of other failures."

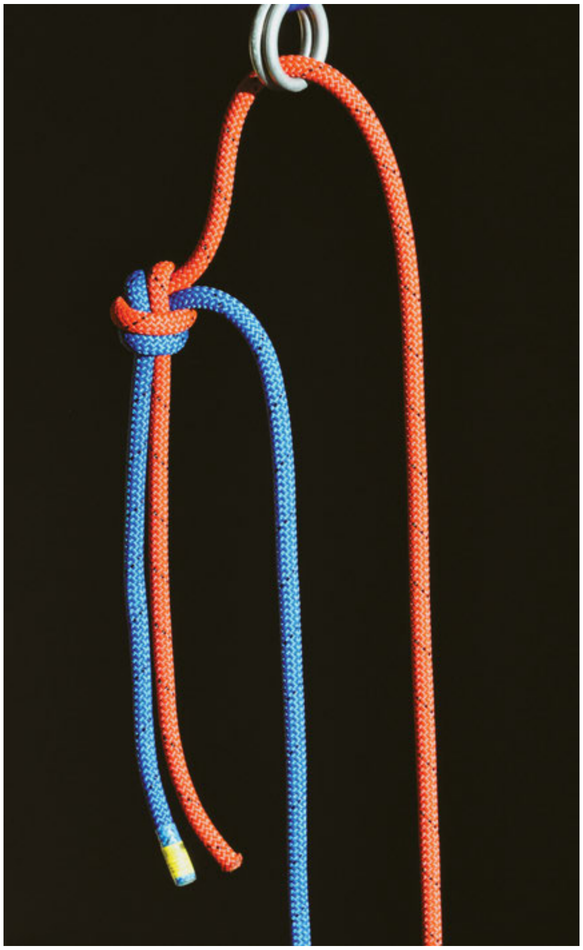

Flat overhand knot.

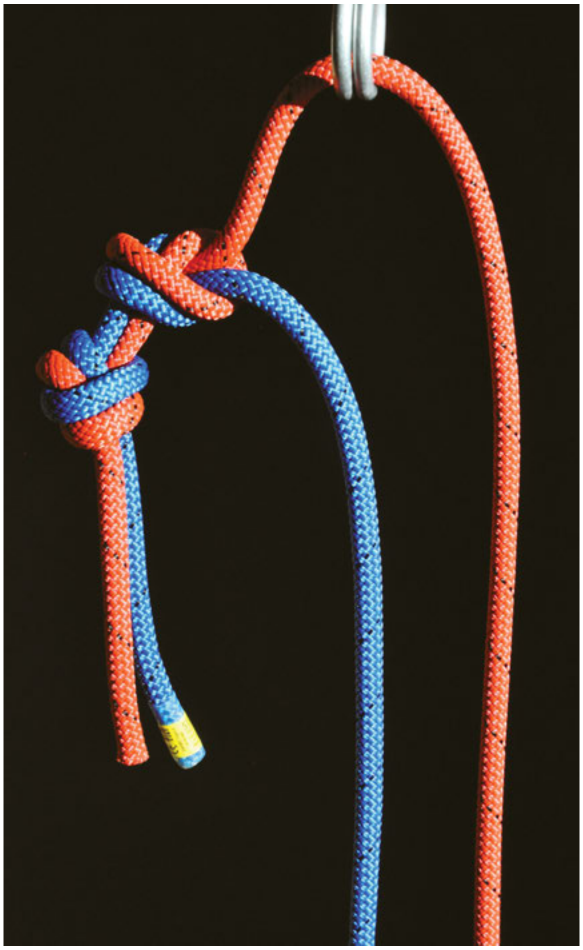

Flat overhand with overhand backup.

From 1999 to 2009 various tests revealed the flat overhand to be roughly 30 percent weaker than the double fisherman's for tying two ropes together, but still plenty strong for rappelling situations. Testing also revealed that it was virtually impossible to get the knot to fail, as long as it was tied with a suitable-length tail and properly tightened.

Petzl, a leading manufacturer of rappelling devices, recommends the flat overhand as the knot to use for joining two rappel ropes together, as long as the ropes are of similar diameter and the tail is a minimum of 20 centimeters (8 inches).

The flat overhand has become widely adopted as the knot for joining two rappel ropes of similar diameters because it is easy to tie and easy to untie after it has been weighted, and it presents a clean profile when pulled down the cliff as the ropes are retrieved, thus less likely to jam in a crack. For added security it can be easily backed up simply by tying another flat overhand above the first one, although this adds bulk.

That being said, there are a few cautions: It is not recommended for tying together two ropes of drastically differing diameters (e.g., 7mm to 11mm), or for use on very stiff ropes. The bottom line is that the knot should be used with discretion, well tightened (pull hard on all four strands), and tied with a long tail (minimum of 8 inches).

The flat overhand is a poor choice for use with nylon webbing, and it has been responsible for rappel anchor failures where it was tied in webbing with a very short tail. The flat eight is even more dangerous if tied with short tails, especially in nylon webbing, and has been responsible for numerous accidents, as it inverts at shockingly low loads when the knot rolls inward and collapses.

Hitches

A hitch is a knot that is tied around something. A friction hitch is a knot tied with a cord or sling around another rope, utilizing friction to make the knot hold when it is weighted, but releasable and moveable without untying when it is unweighted.

Clove Hitch

The clove hitch is used to fasten a rope to a carabiner, tied around the wide base of a carabiner. The beauty of the clove hitch is easy rope-length adjustment without unclipping from the carabiner, making it a truly versatile knot for anchoring purposes: for anchoring a belayer, tying off an anchoring extension rope, or tying off the arms of a cordelette.

Get in the habit of tying the load-bearing strand on the spine side of the carabiner. Doing so loads the carabiner in the strongest configuration. Be sure to tighten the clove hitch properly by cranking down on both strands.

Slip Hitch

This is a simple knot used to tie off a knob or spike of rock.

Munter Hitch

The Munter hitch can be used for belaying, lowering, and rappelling. The mule knot, backed up with an overhand, is used to tie off a Munter hitch. The great advantage of the Munter/mule combination is that it can be tied off and released when the rope is weighted and under tension, making it one of the key knot combinations for many systems used in single pitch terrain.

The Clove Hitch

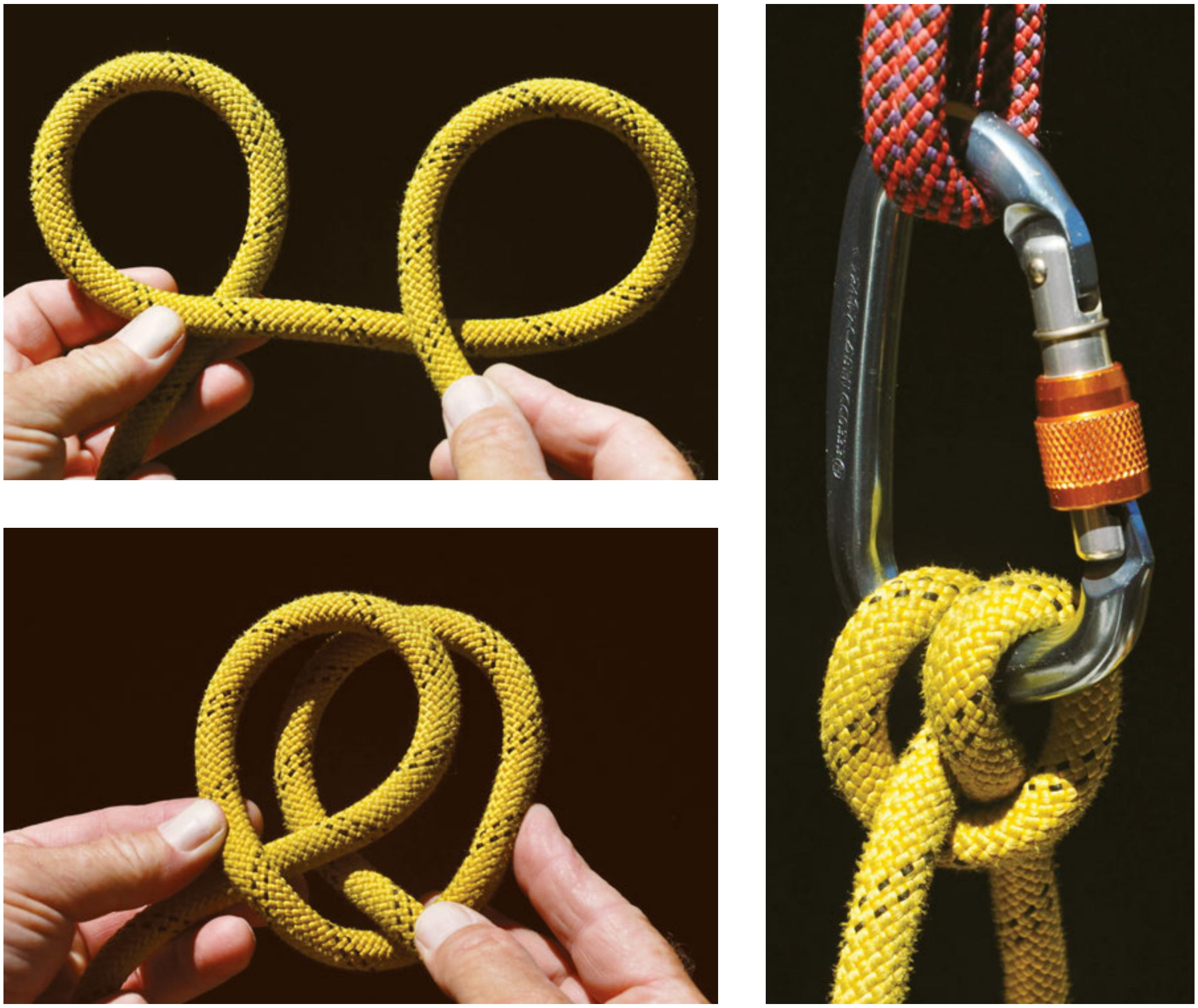

Tying the clove hitch.

The Slip Hitch

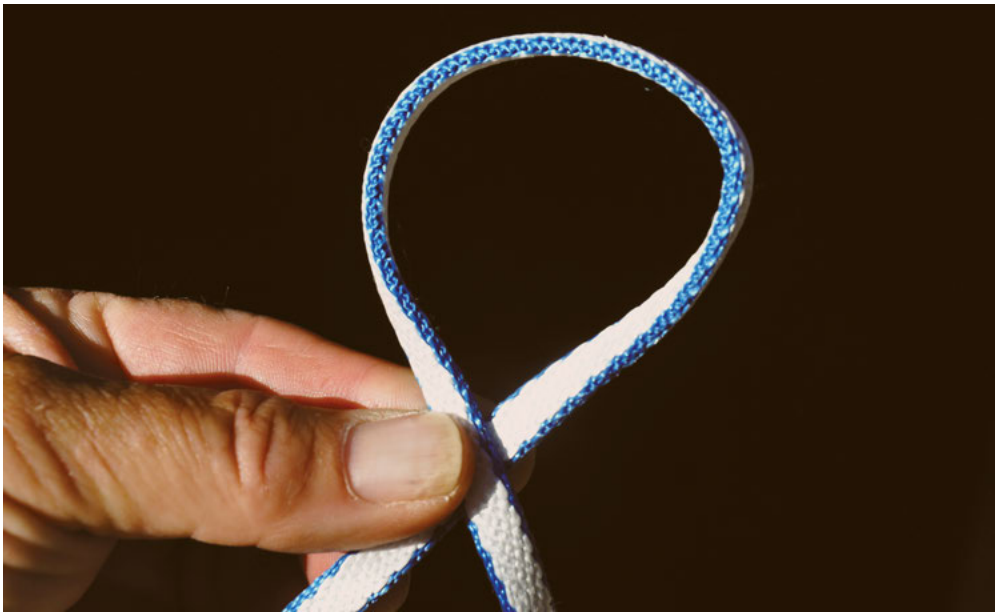

How to tie a slip hitch. The slip hitch can be tightened by pulling on one strand, making it more secure than a girth-hitch for tying off knobs of rock.

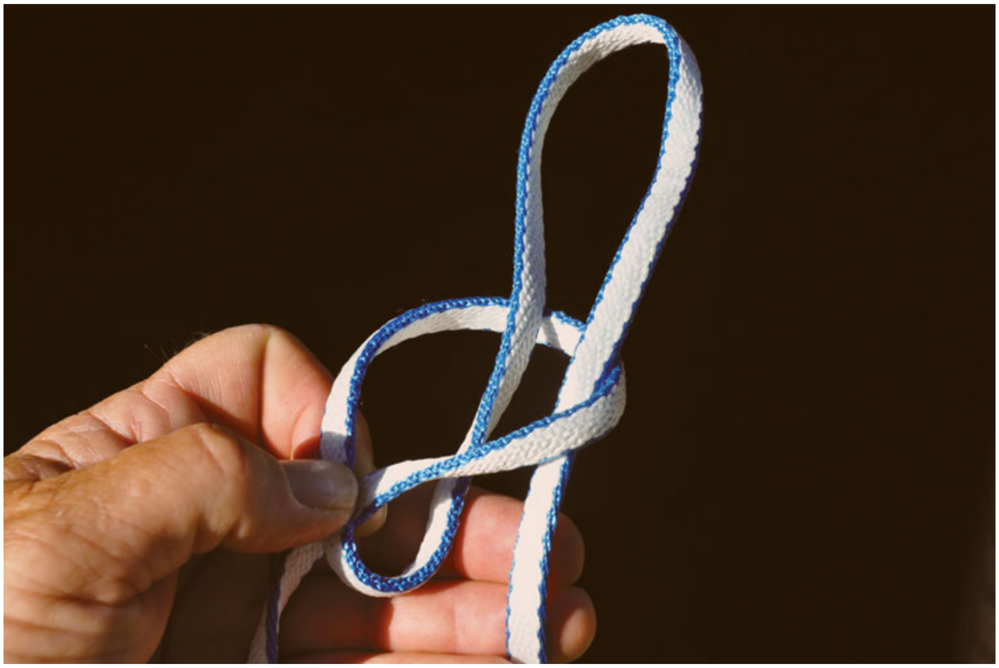

Tying the Munter Hitch on a Carabiner

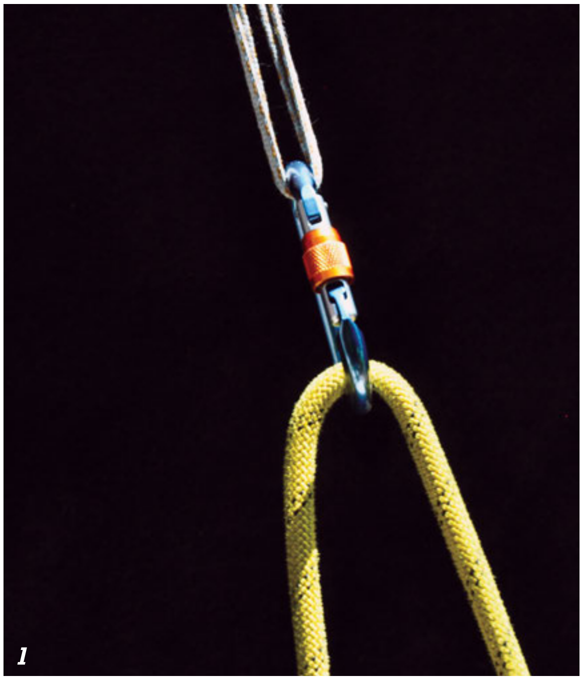

1

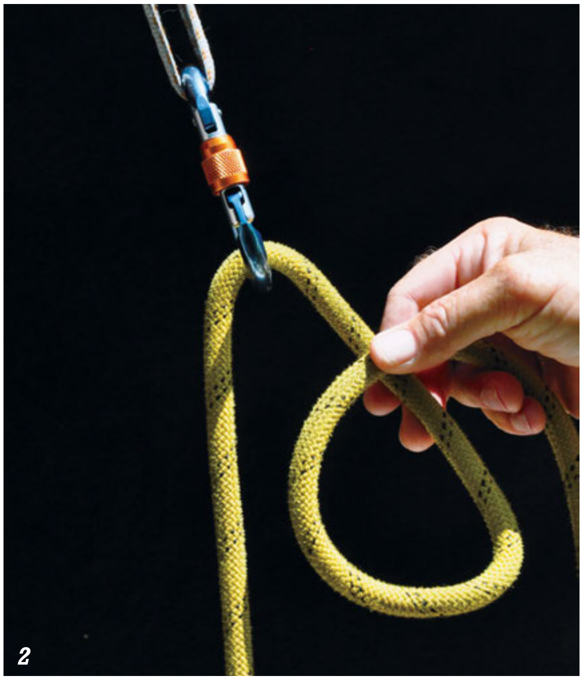

2

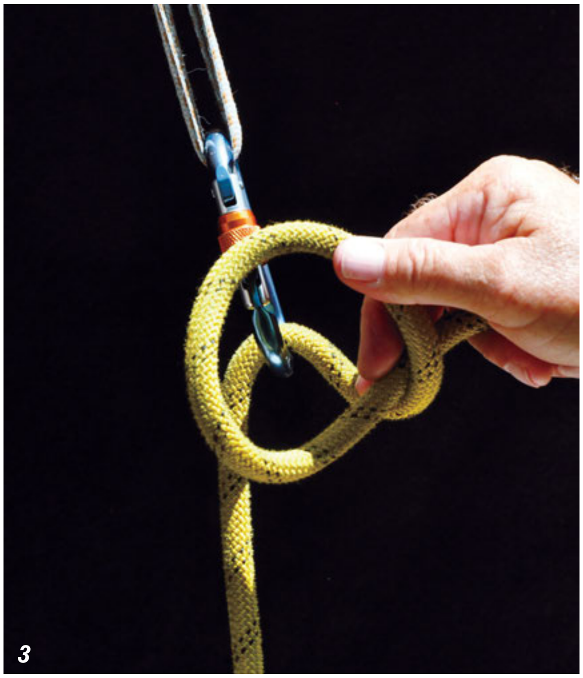

3

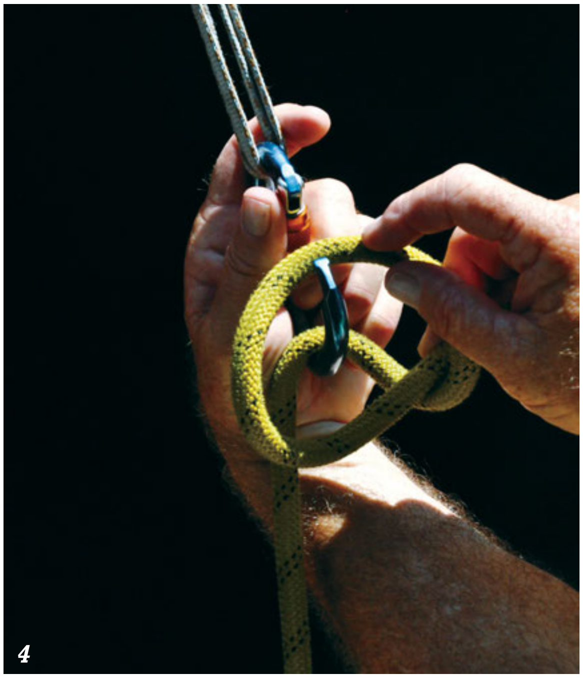

4

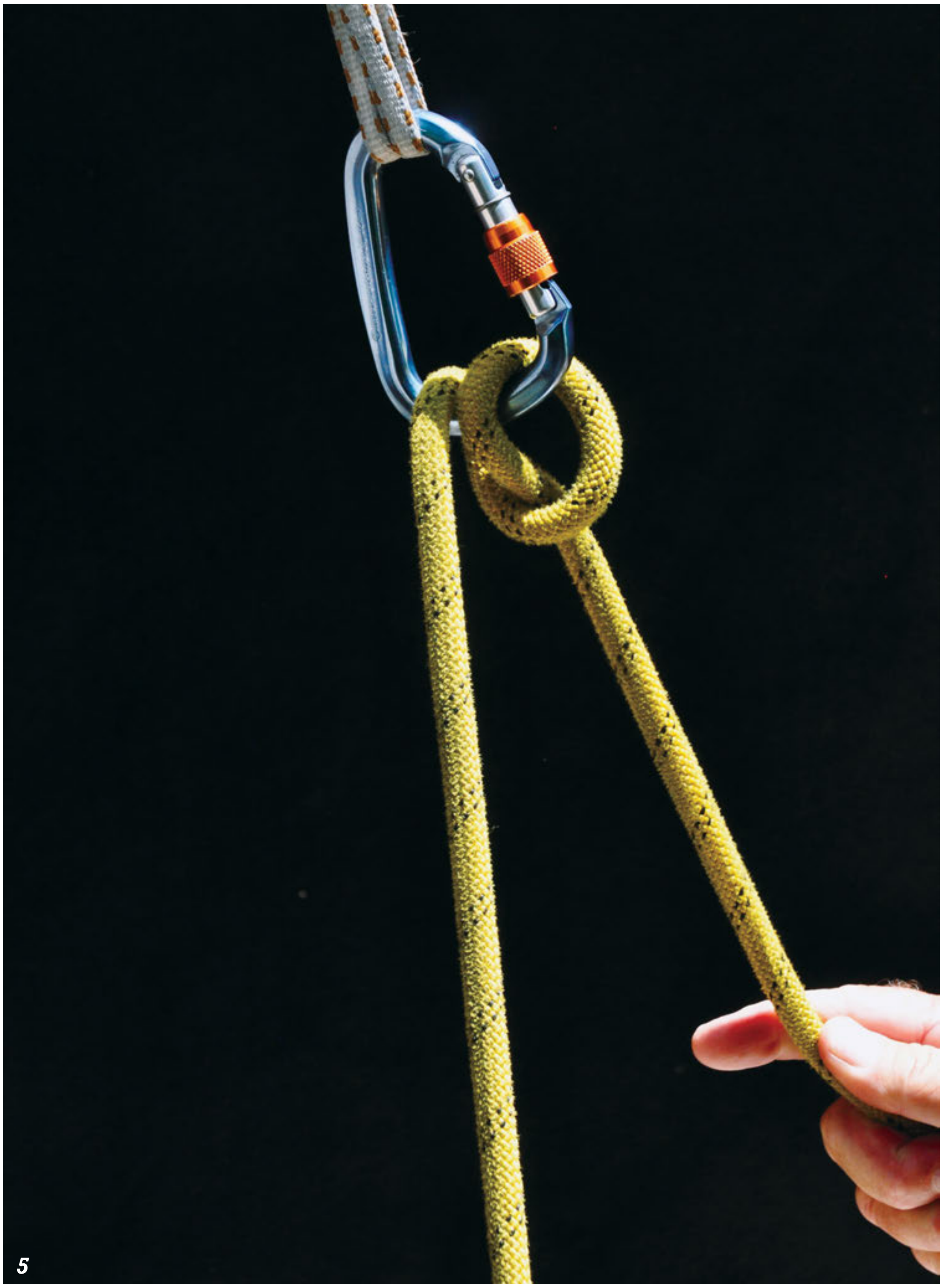
5

Tying the Munter Hitch

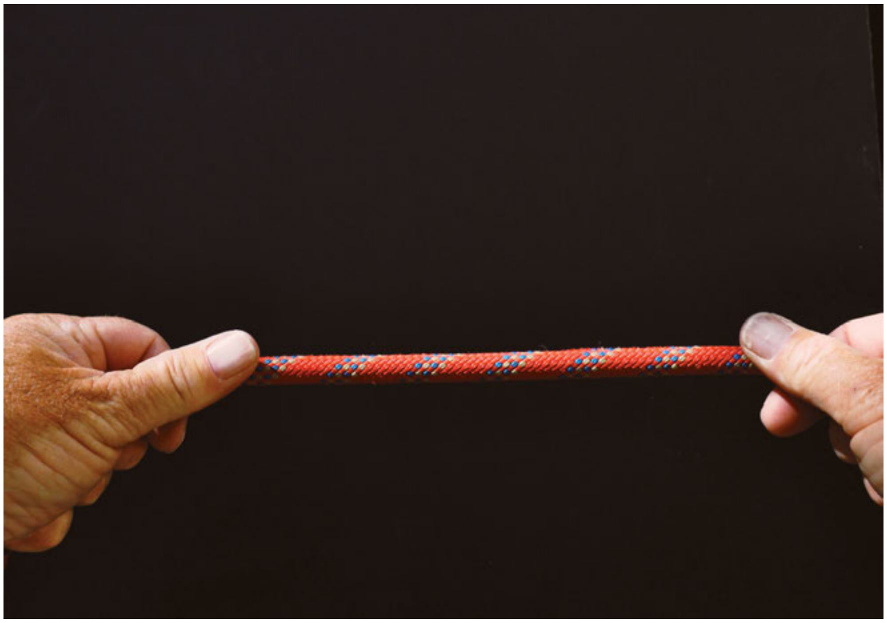

Hold a single strand of rope with both hands, thumbs pointing toward each other.

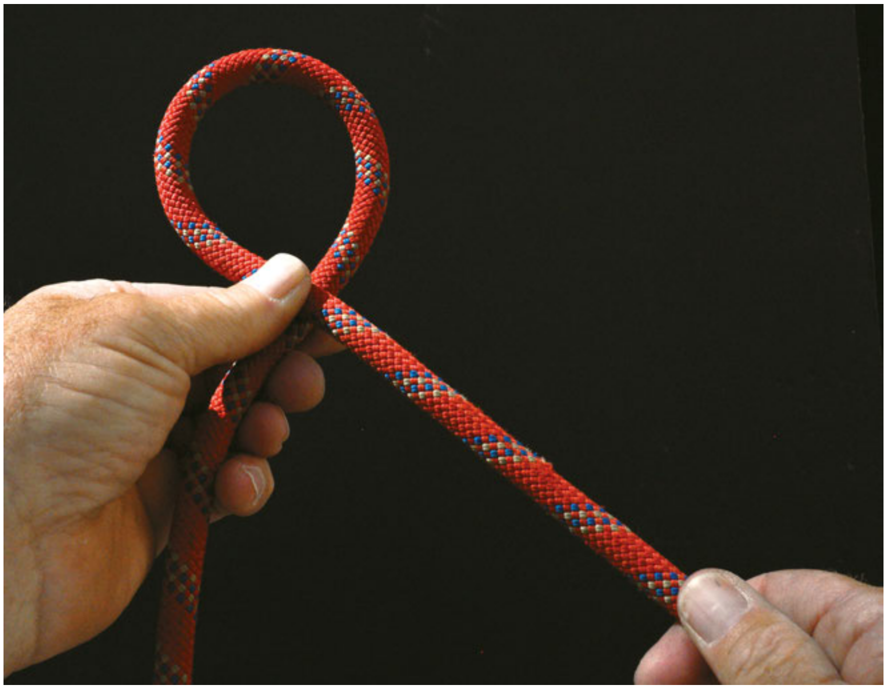

Cross the right-hand strand in front of the left-hand strand and hold the two strands where they cross with the left thumb and forefinger, then slide the right hand down about 6 inches.

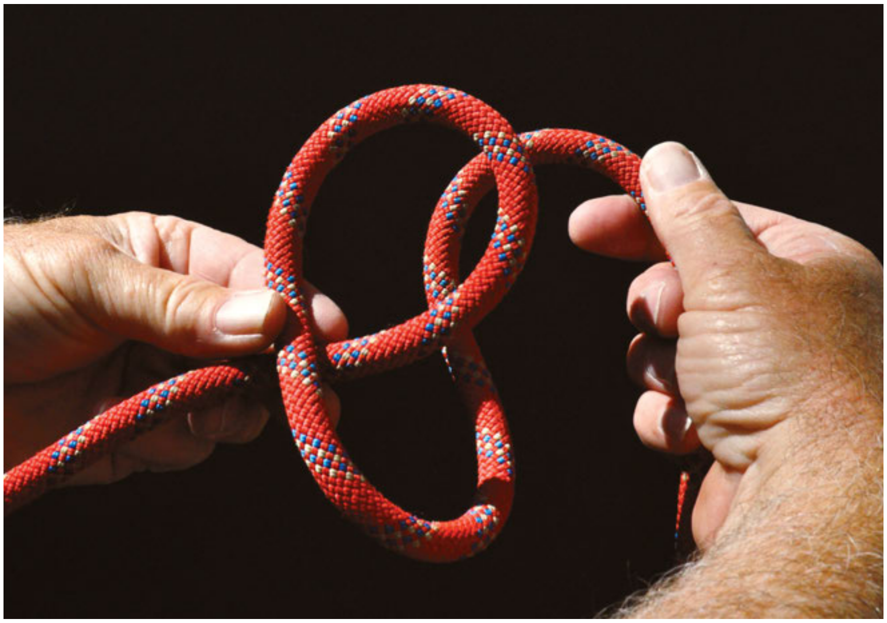

Bring the right strand up and behind the loop.

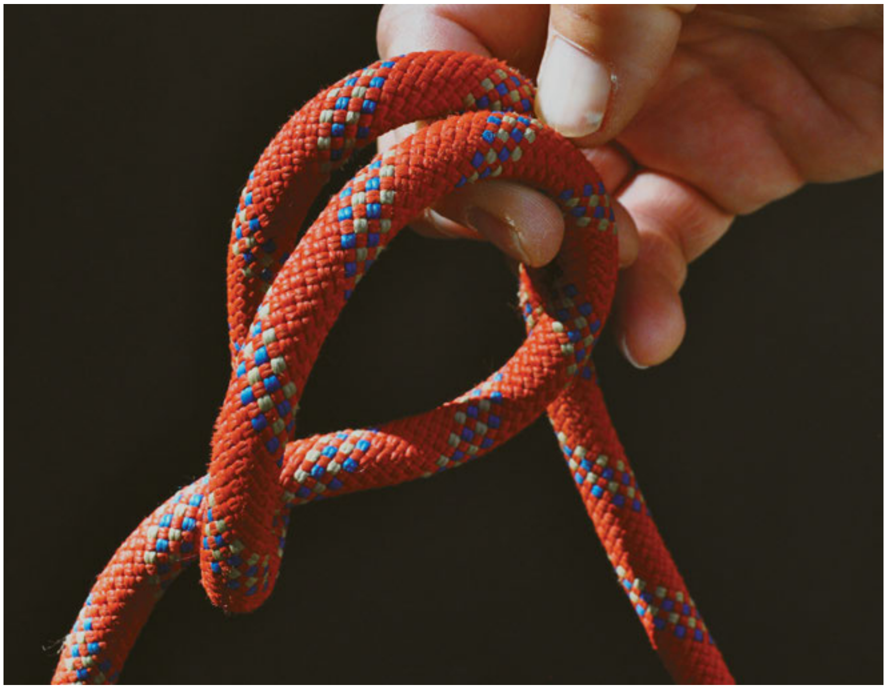

Clip a locking carabiner where the forefinger is shown here, below the two top strands.

Tying a Mule Knot

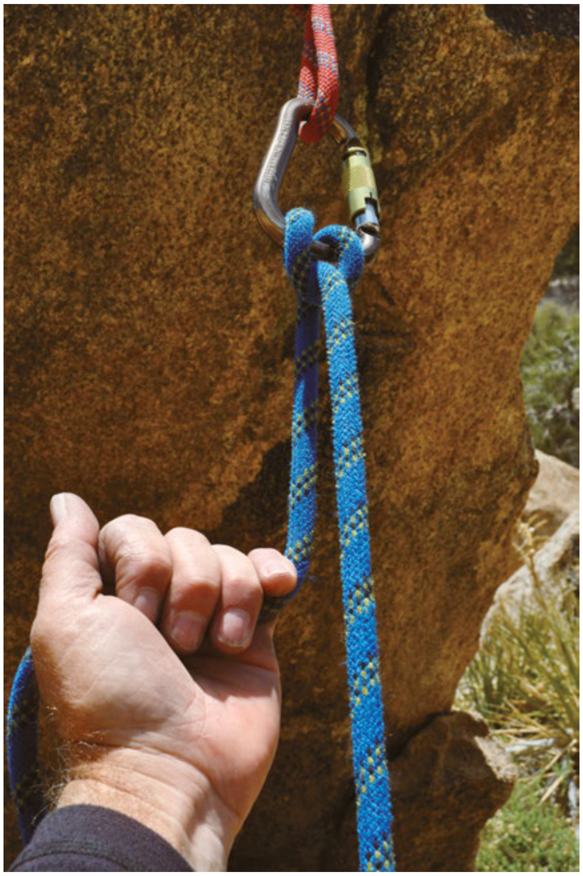

When tying a mule, be aware that when the rope is under tension (holding a climber's weight), it is necessary to maintain control of the brake strand.

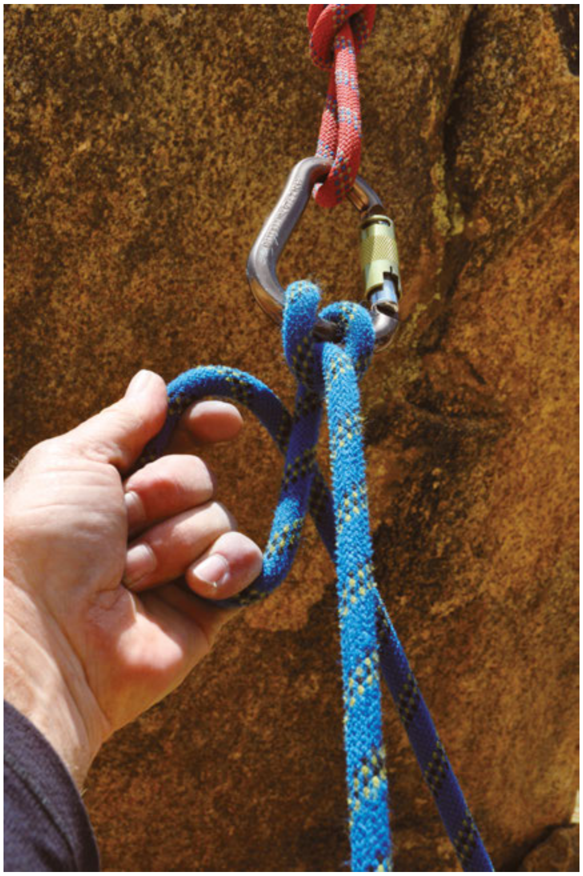

Keeping the load and brake strands parallel, form a loop on the brake strand by crossing it behind while still maintaining your grip with your brake hand.

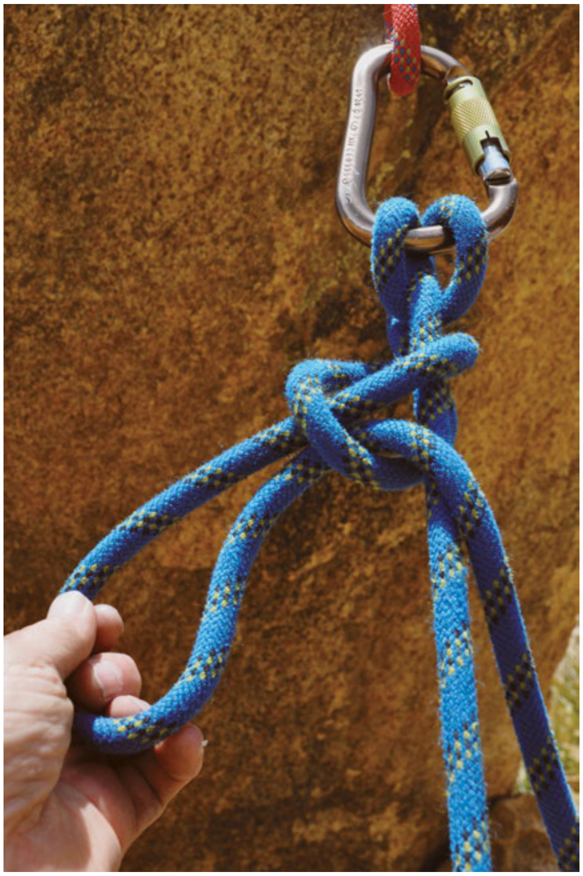

With the non-brake hand, take a bight of rope and pass it through the loop you've created, with the load strand in between the loop and the bight. Snug the mule knot up tight against the Munter hitch.

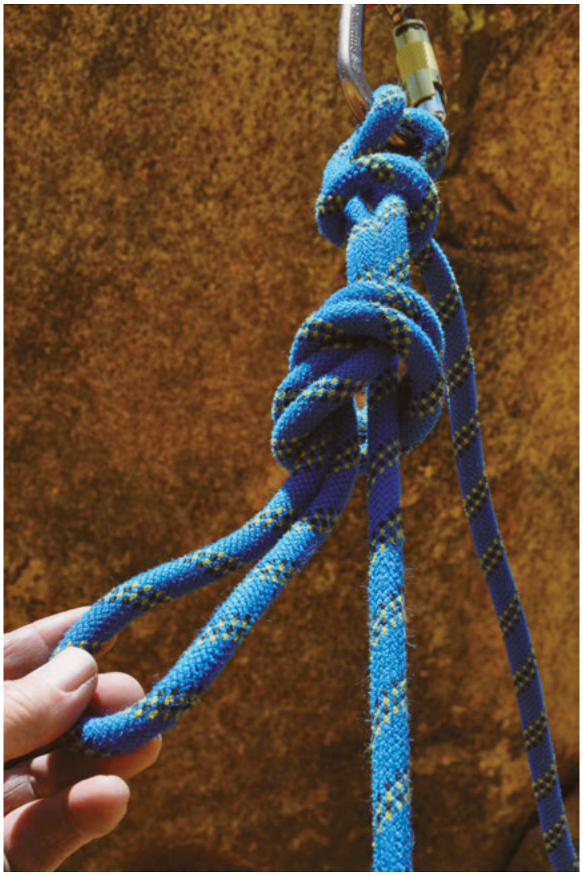

Pull some slack and finish with an overhand loop backup.

Tying Off a Tube Device with a Mule Knot in Front of the Device

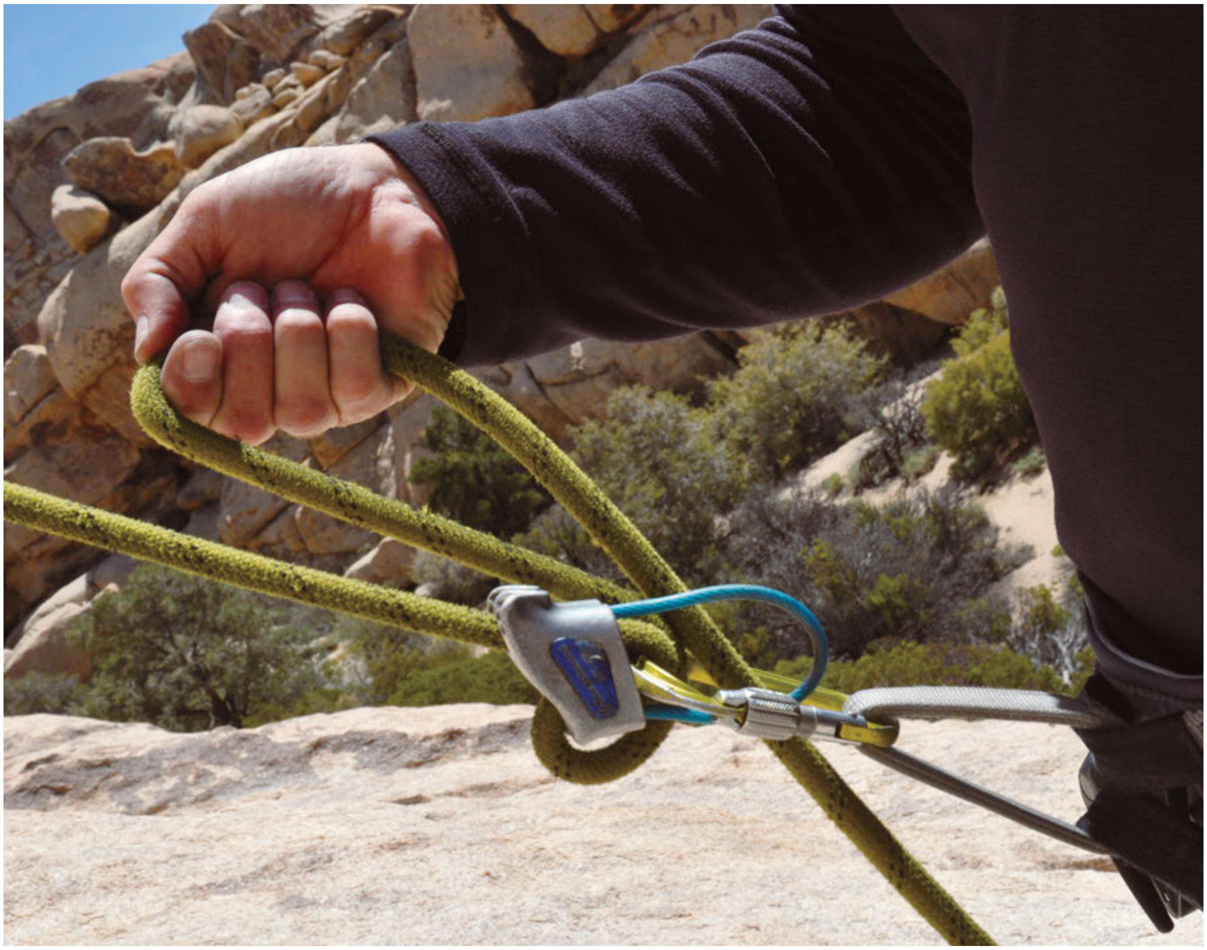

First pass the bight of rope (on the brake hand side) through the belay carabiner.

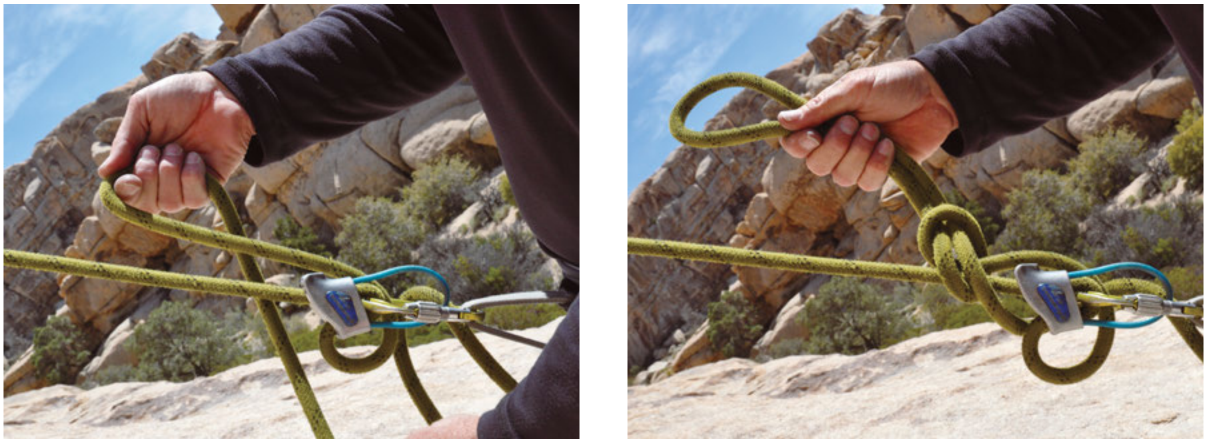

Then tie the mule knot above your device on the load strand of rope going to the climber.

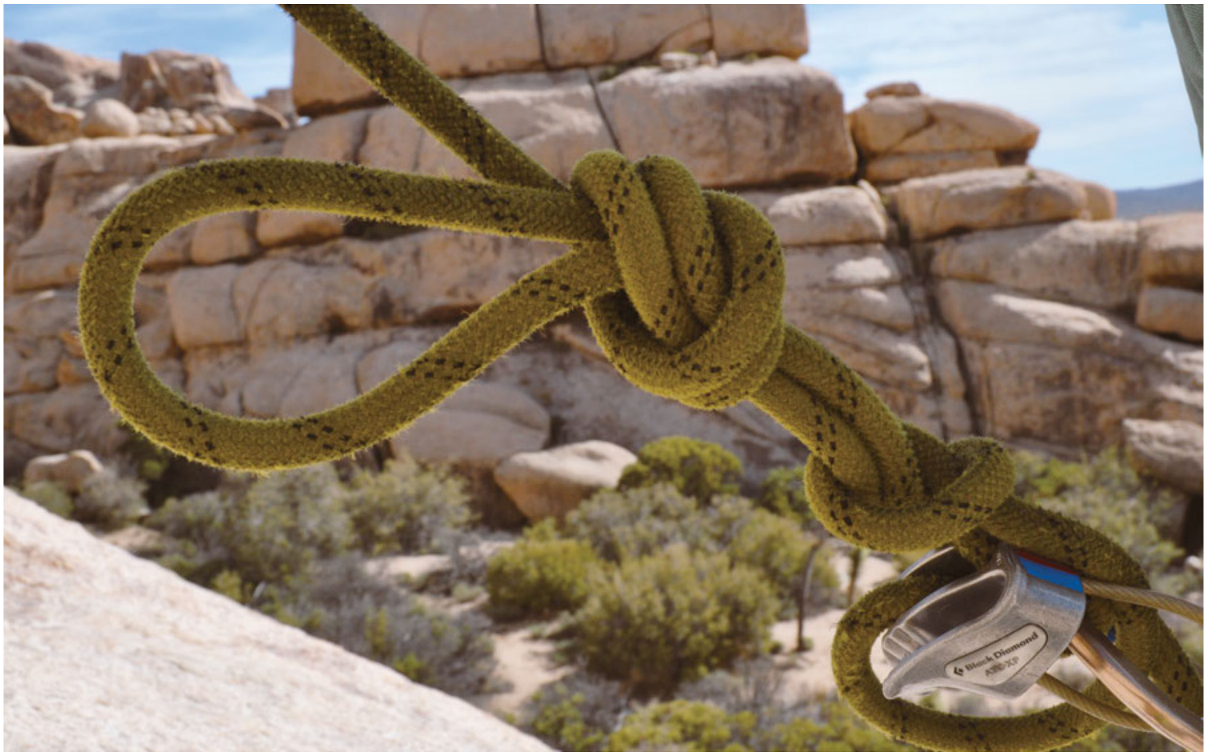
Black Diamond
ATC-XP

Finish with an overhand backup.

Tying Off a Tube Device on the Carabiner Spine

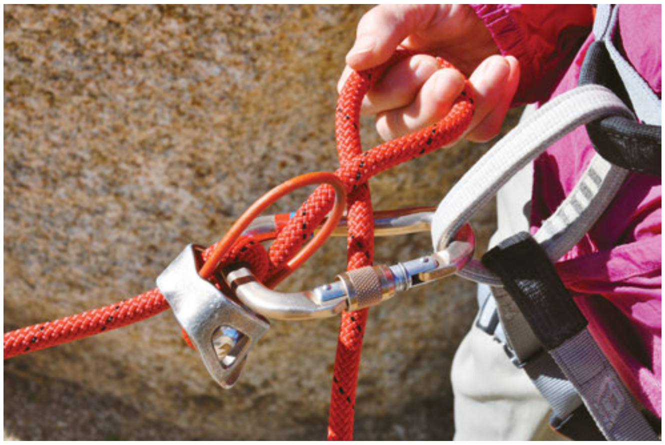

First pass a bight of rope through the carabiner and form a loop. If the device is under tension, pinching the rope against the device with the opposite hand will help lock it off.

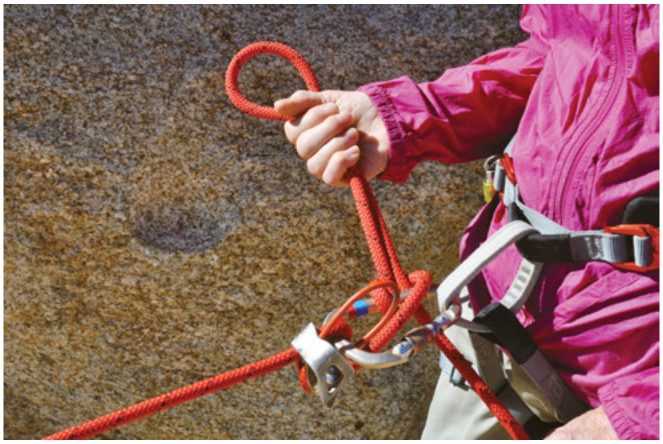

Then pass a bight of rope through the loop you've created, with the spine between the loop and the bight.

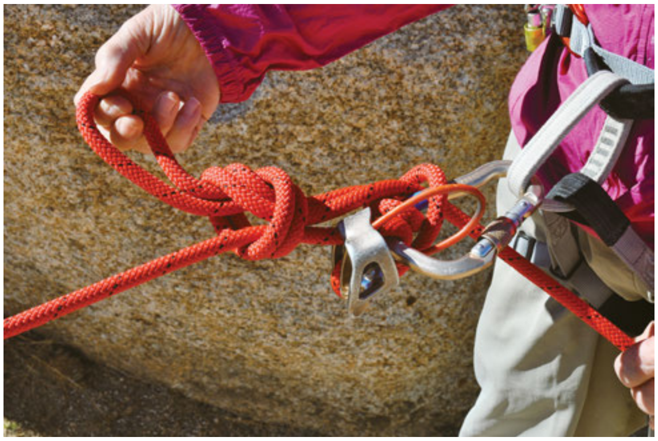

Finish by tying an overhand backup on the load strand.

Prusik Knot

A prusik knot is used for rope ascending and as a component in many load transfer systems. It can be loaded in either direction. To tie a prusik, first make a "prusik cord" out of a 5-foot length of 6mm diameter nylon cord tied into a loop with a double fisherman's knot. Buy the softest, most pliable nylon cord available, because a softer cord will grip best. To tie the prusik, simply make a girth-hitch around the rope with the cord, then pass the loop of cord back through the original girth-hitch two or three more times. Dress the knot to make sure all the strands are even and not twisted—a sloppy friction hitch will not grip as well. Test the knot before using it. A thinner cord will grip better, but below 6mm in diameter, the cord will be too weak for many rescue applications. To slide the prusik after it has been weighted, loosen the "tongue," which is the one strand opposite all the wraps.

Tests on various friction hitches reveal that the prusik consistently has the most holding power in a wide array of cord and rope combinations, so use the prusik in scenarios (like 3:1 raising systems) where it will be loaded with more than body weight.

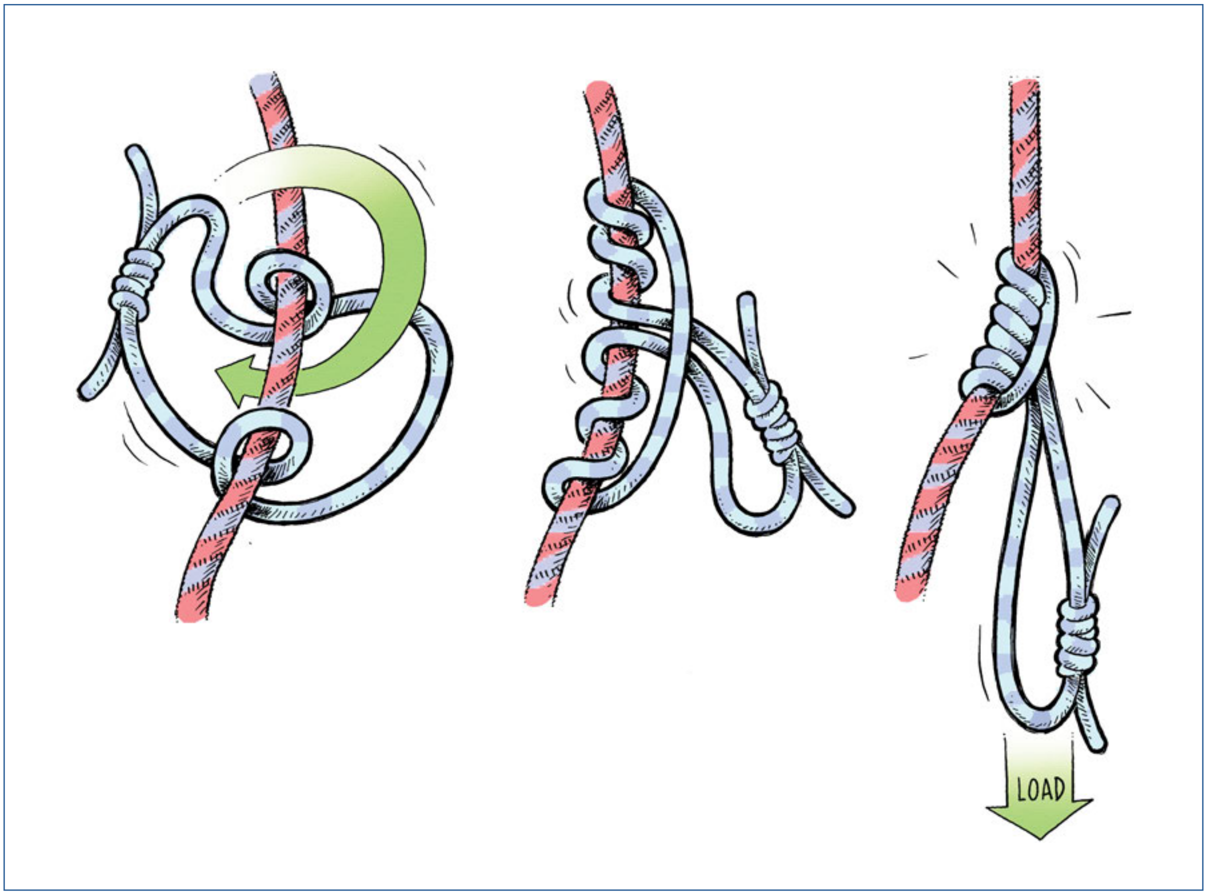

Tying a prusik knot.

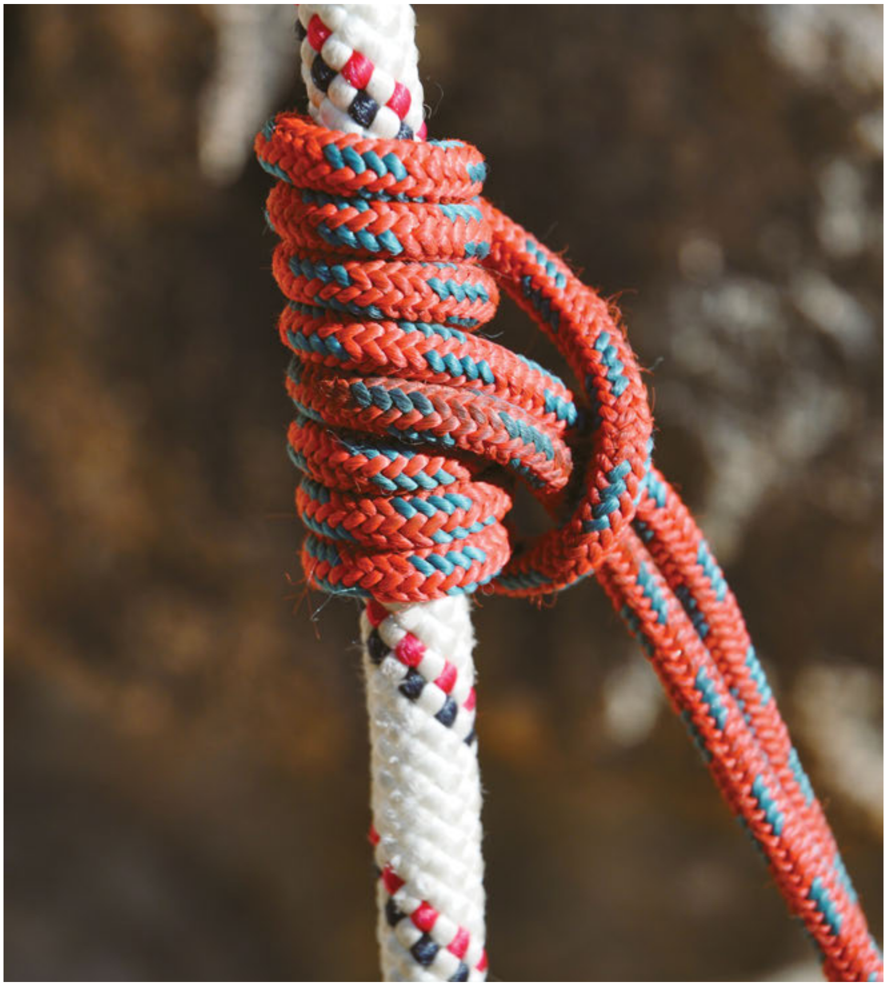

A four-wrap prusik.

Klemheist Knot

This is another useful friction hitch that is quick and easy to tie, and is a good choice as a rope-ascending knot or if you're forced to use a sling rather than a piece of cord to tie a friction hitch. If using a sling, pick a nylon one over a Spectra or Dyneema sling, because it grips better and is less susceptible to weakening if it gets hot (nylon has a higher melting point). Four wraps of 6mm cord tied on a single 10mm diameter rope usually work well. After the hitch has been weighted, loosen the tongue (the one strand opposite all the wraps) to slide it more easily. The klemheist also can be loaded in either direction. If tied the opposite way, with the tongue at the base of the wraps, it is commonly referred to as the hedden knot. The hedden knot can be unreliable if it is formed by simply reversing the direction of a klemheist.

Tying the klemheist knot

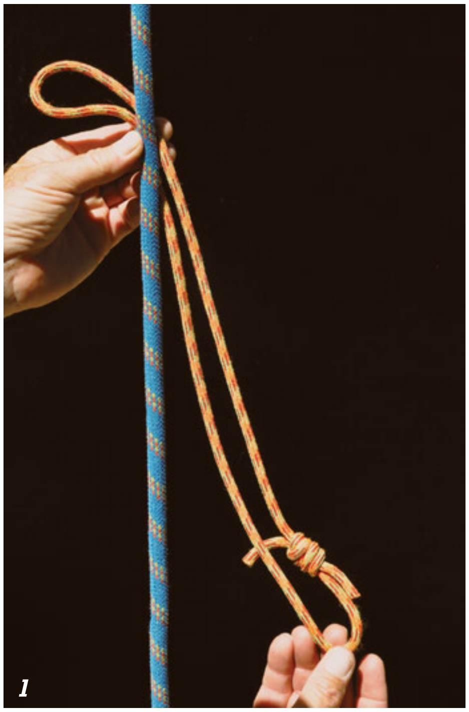
1

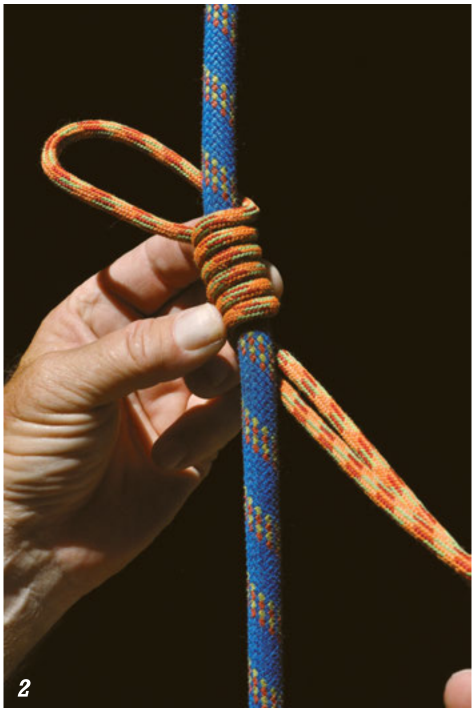
2

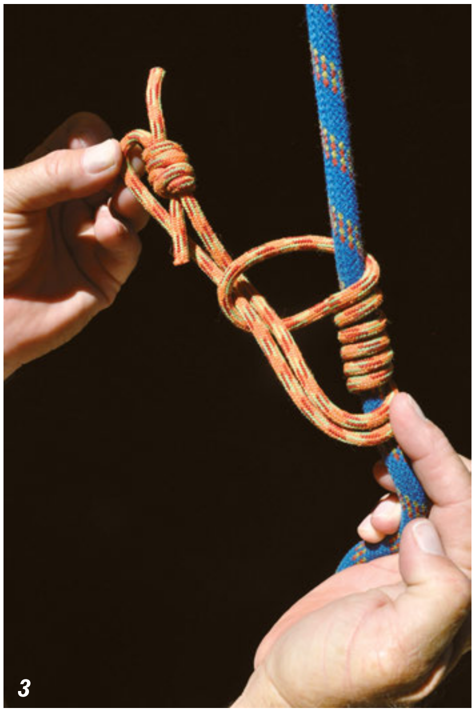
3

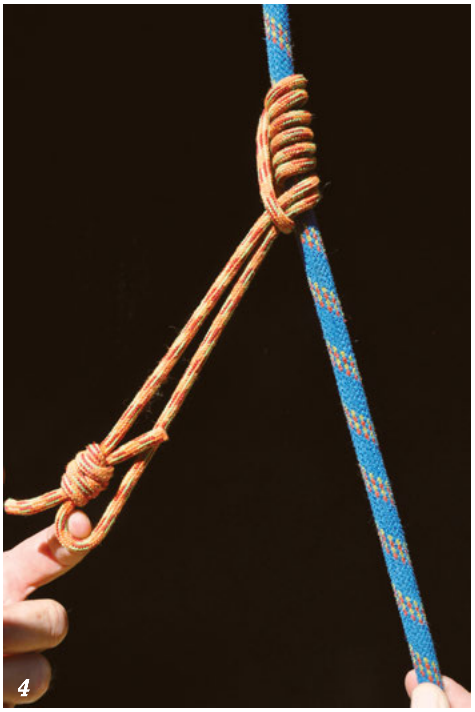
4

Autoblock

Sometimes called the "third hand," the autoblock is used to back up your brake hand when lowering someone, or to back up the brake hand when rappelling. Tie your autoblock loop from a 4-foot 9-inch length of 5mm- or 6mm-diameter nylon cord tied into a loop with a double fisherman's knot. On a doubled rope, three wraps usually works best, and on a single strand of rope, four wraps.

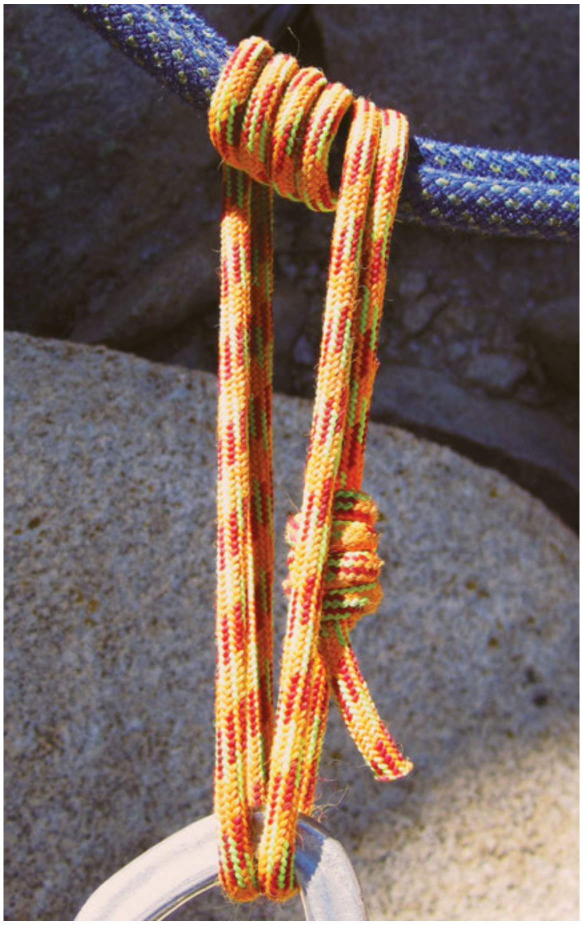

The autoblock with nylon cord.

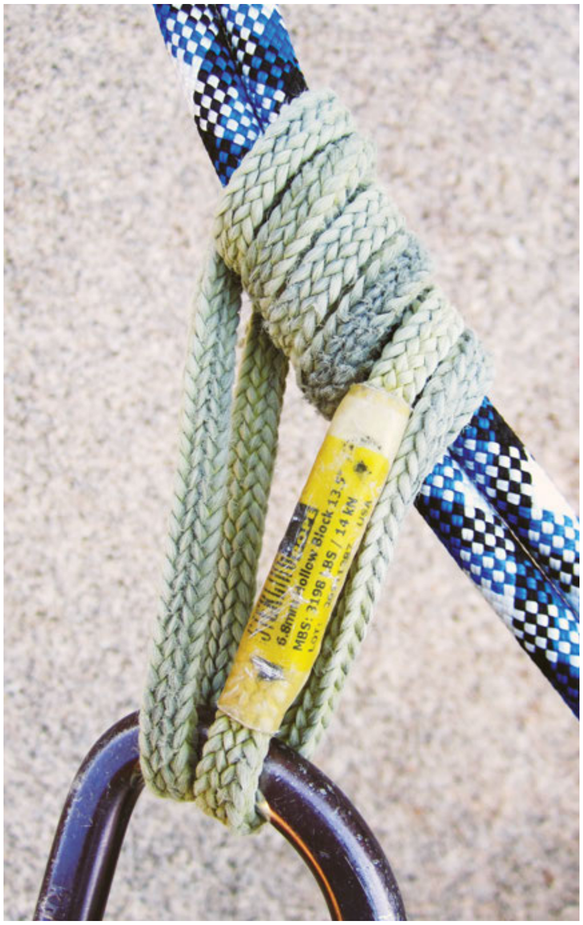

The Sterling Hollow Block, shown here wrapped in an autoblock configuration, is a 100 percent Technora sling designed specifically for use with friction hitches.

Stopper Knot

This knot is used as a safety knot in the end of a rope to "close the system." It is essentially half of a triple fisherman's knot tied on one strand of rope. A stopper knot prevents lowering accidents when the rope length is too short and the end of the rope travels through a belay device. A stopper knot also prevents rappelling off the end of the rope if rappelling with a plate, tube, or assisted braking (e.g., Grigri) device. When using two ropes, tie a separate knot at the end of each rope, as tying both ropes together can cause the ropes to twist around each other.

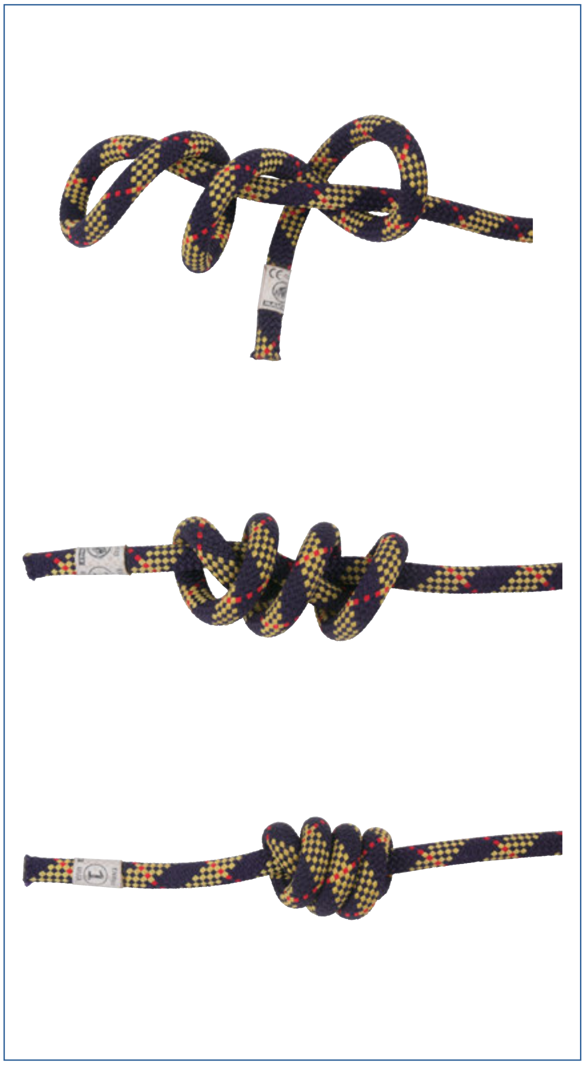

Tying a stopper knot.

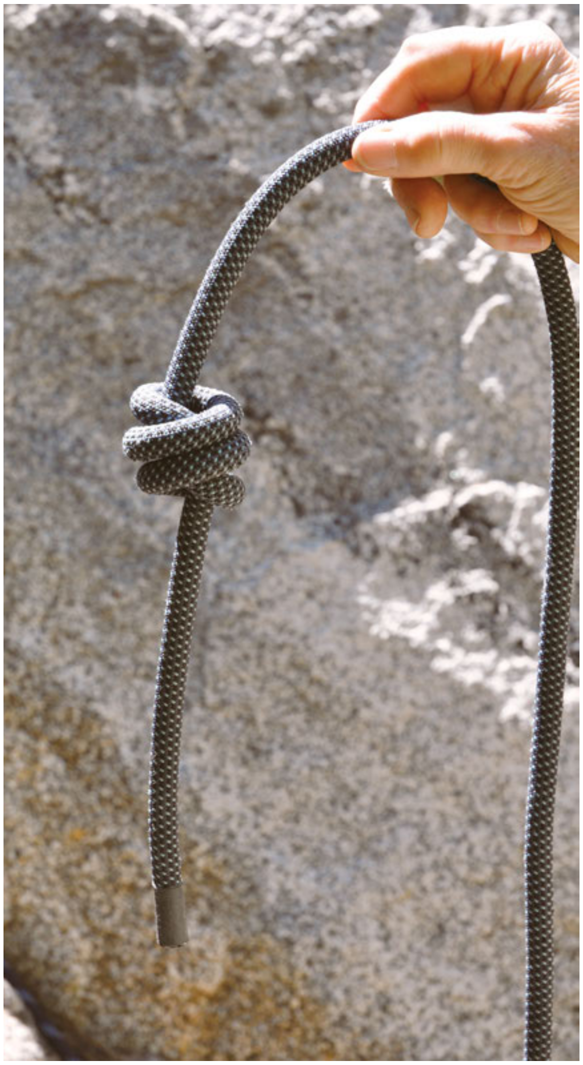

Always close the system by either tying into the end of the rope or tying a stopper knot.

CHAPTER 7

Protection and Anchoring

Nearly every system that a climber can build starts either with good protection placed in solid rock or with strong independent natural features. Many recreational climbers make placements that are "good enough" but could be better. It's an instructor's job to not only make excellent placements, but also to evaluate student placements.

"Good enough" isn't good enough. An instructor should strive to make the best placements possible given the circumstances, while developing strategies to help students understand the dual arts of protecting a climb and building anchors.

Rock Assessment

The first thing to think about when assessing an anchor is the integrity and structure of the rock itself. Catastrophic anchor failures have occurred not because the gear placements were bad, or the rigging was flawed, but because the rock itself was unsound. Determining good rock structure and knowing what to watch out for are fundamental requirements to build solid anchors.

When placing gear, the ideal crack is what instructors call "a crack in the planet," a deep fissure that runs perpendicular (i.e., at a right angle) to the plane of the rock face, cleaving a massive, solid face of granite.

In general, there are two things to avoid: detached blocks and flakes. A detached block is just that—a chunk of rock that is not attached to the main rock structure, but is either sitting on top of the cliff like a boulder, or is part of the main rock face but completely fractured with cracks on all sides.

To assess a block, start by looking at its size. How big is it? Is it the size of a refrigerator, a car, or a house? Putting a piece of gear in the crack beneath a smaller block is a very bad idea. When the piece is weighted, it has a prying effect outward on the block. Even large blocks can shift easily. Look at how the block is situated. Is it perched down low, where it cannot slide out? Does it rest on a flat surface, or is it resting on an inclined slab? When tying off blocks, watch for sharp edges that may fray or cut the rigging rope, and use padding or an edge protector when needed. Generally, be very skeptical of using detached blocks as part of an anchor system, especially smaller blocks.

Flakes should also be avoided. A flake is formed by a crack in the rock that runs parallel to the main rock face. It can be wafer thin, or several feet thick. A flake is inherently weak, since any gear placement, when loaded, will exert a prying effect outward on the structure of the flake, which can fracture the flake if it's not strong enough to bear the force. In a naturally weak rock, like sandstone, a thin flake of rock can be extremely weak.

Exfoliation is a natural process of granite formations and is the key in the formation of domes. Flakes of granite are layered, like layers of an onion, and the outer layer peels off from time to time due to the effects of weathering and gravity, exposing a new layer beneath.

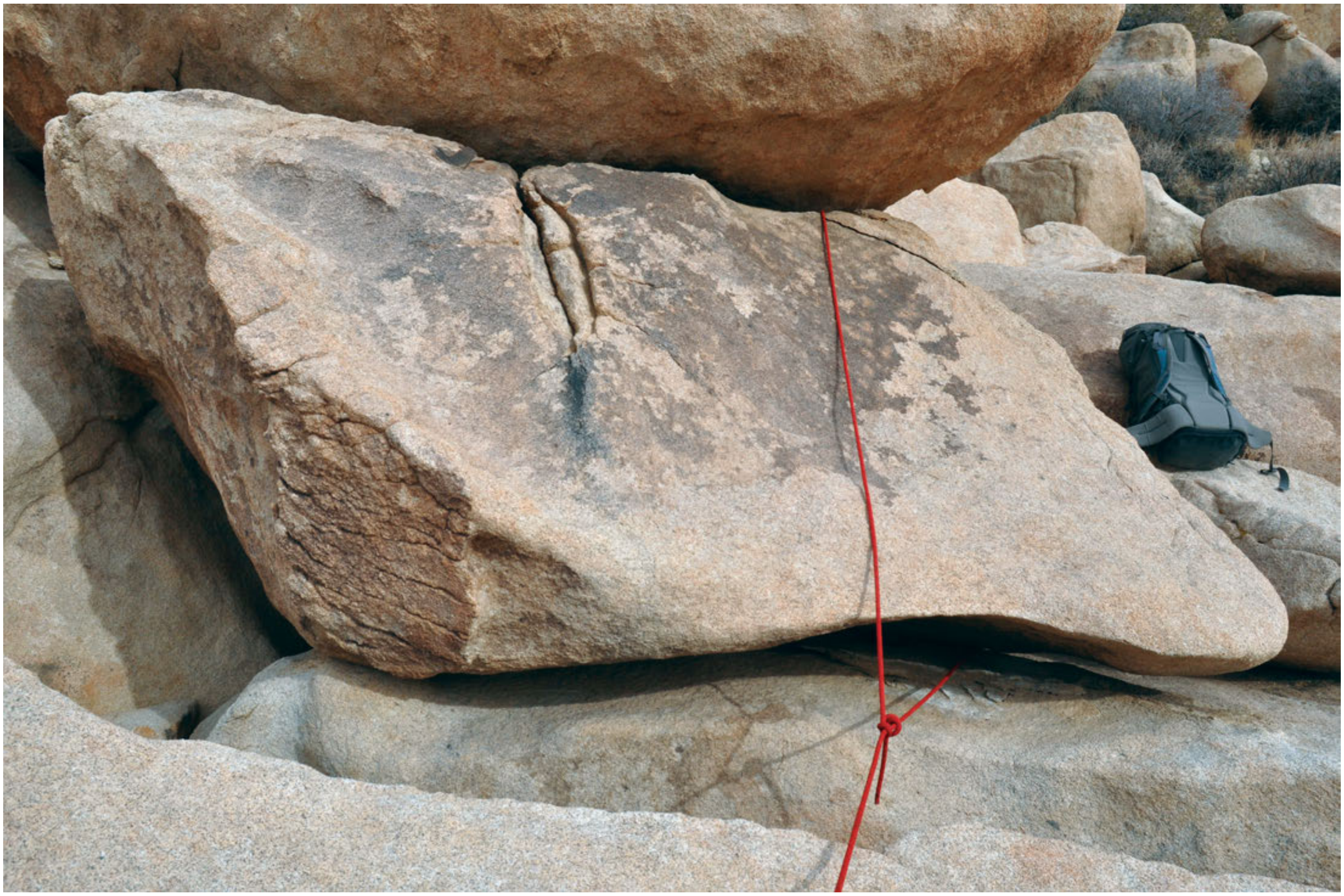

Use good judgment if anchoring to detached blocks. Avoid small ones and those situated on an inclined slab or those precariously perched. A good rule of thumb when using detached blocks is this: for a sole monolithic anchor, the block should be at least as big as a full-sized refrigerator, resting on a flat surface.

Rockslide!

One of the largest examples of exfoliation occurred in Yosemite Valley on a hot July day in 1996 at the Glacier Point Apron. An enormous flake, roughly the size of a football field and about 4 feet thick, detached from a point high on the cliff, shearing off in one gigantic piece. After a 2,000-foot free fall, the impact resulted in a massive explosion, creating a 300 mile per hour shock wave of wind that felled a thousand pine trees in a wide swath. A tourist, in line at the Happy Isles snack bar over a quarter mile away, was killed when hit by a piece of the shrapnel from the blast.

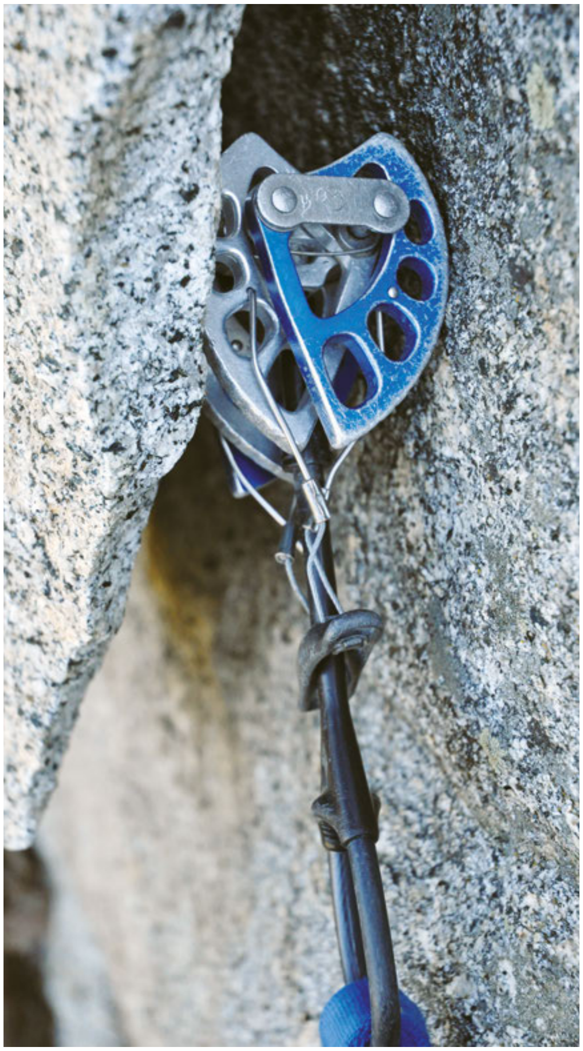

This camming device has been placed behind a flake of rock (left). If the cam is loaded, it will pry outward, potentially breaking the flake. This can cause the anchor to fail and also send the chunk of rock tumbling down onto people below.

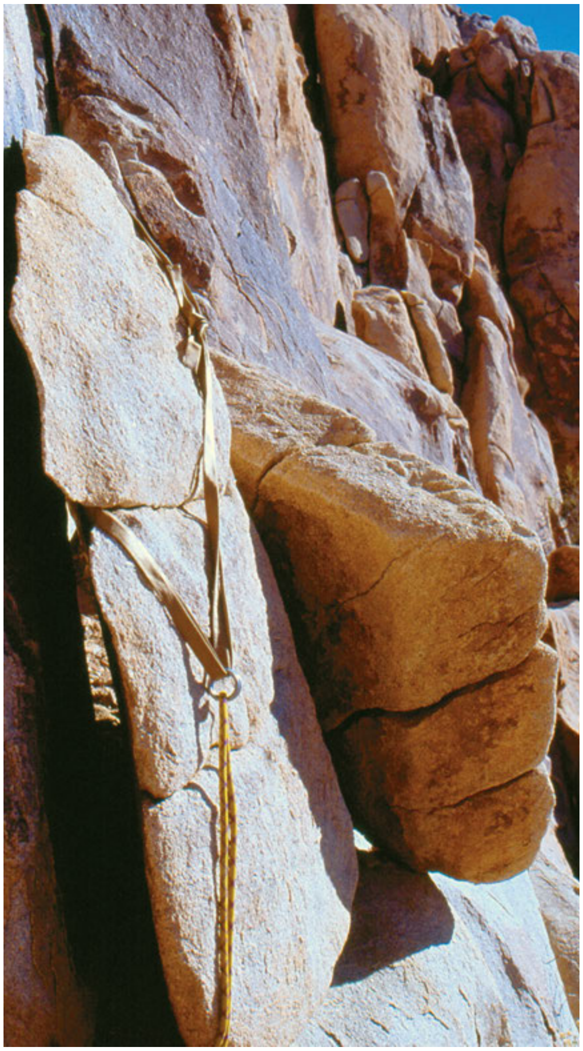

This fabricated anchor is set up only to illustrate a point. Leaning back on this rappel anchor would most likely pull this detached flake right off the cliff.

When building anchors, look with skepticism at any flake. How thick is it, and how well attached to the main rock structure or cliff face? Test its soundness by thumping on it with the palm of the hand. Does it vibrate? Is there a hollow sound? When analyzing rock structure, act like a geologist and scrutinize the rock and its various formations very carefully.

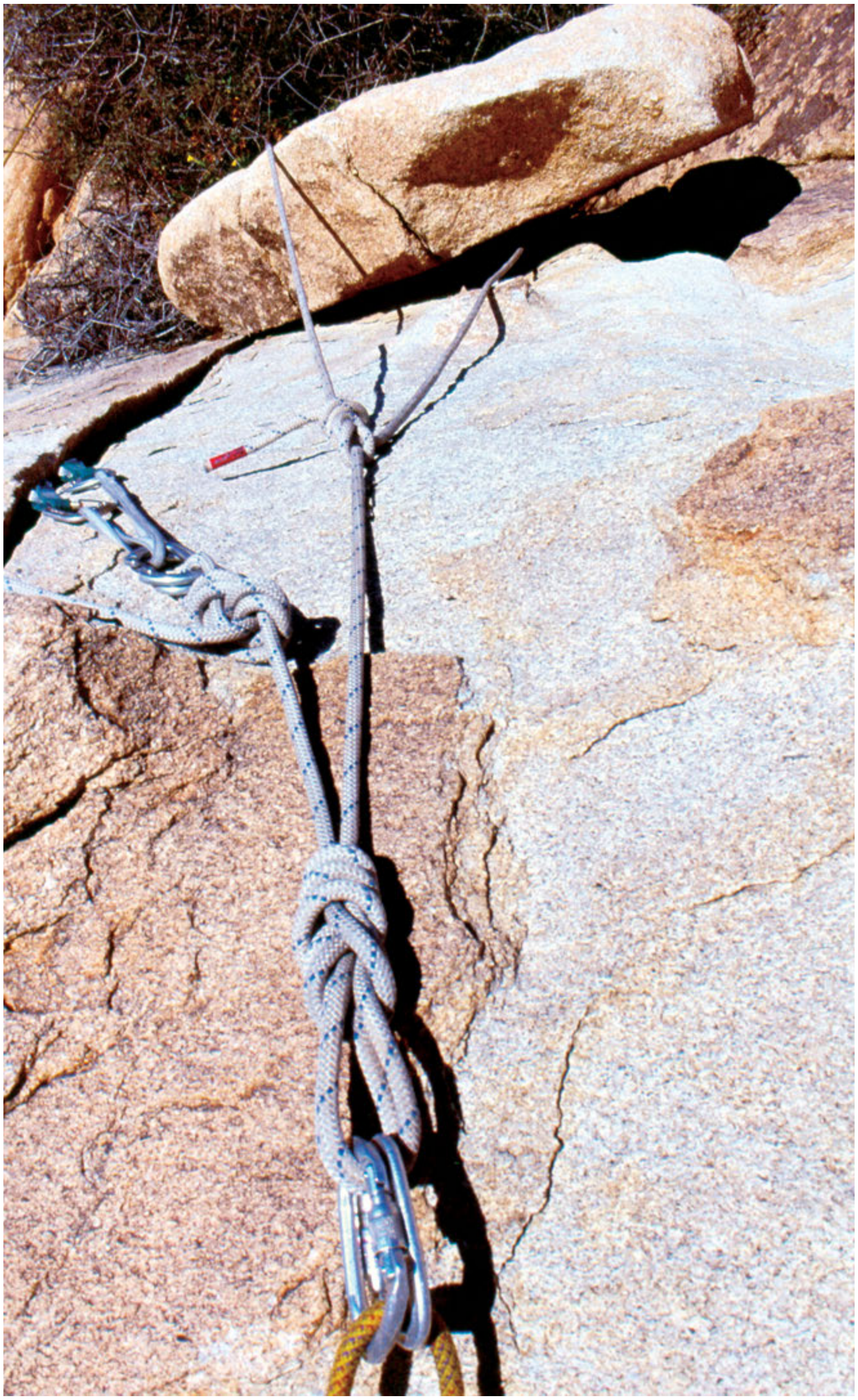

Another dangerous anchor rig using a very questionable block. The block sits on a severely inclined slab, and a pull on this toprope anchor setup may be just enough to send it sliding down the slab and over the brink.

Macro to Micro Rock Assessment

When assessing rock structure, evaluate from macro to micro. Macro is the big picture. Look at the main rock face. Is there a massive, solid rock structure? Is there a crack in the planet? Or are the cracks an intricate matrix where no real massive piece of completely solid rock exists. Are there block or flakes? Is it possible to avoid using them? These are questions

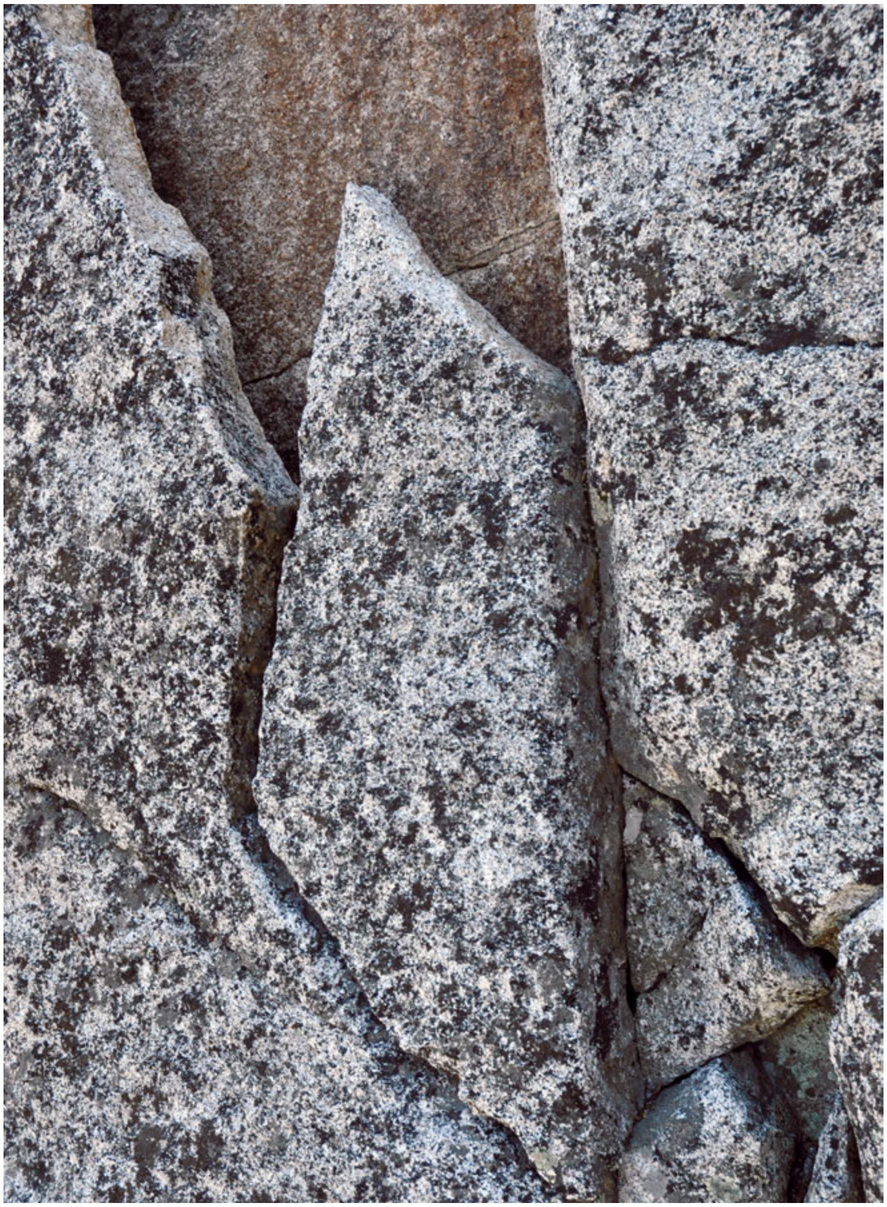

This detached flake is a great example of bad rock structure.

that need to be asked. Never blindly place gear in cracks without first scrutinizing the big picture: the overall structure and integrity of the rock itself.

Microstructure is what's inside the crack. Is the surface of the rock rotten, grainy, dirty, or flaky? Are there hollow spots or hollow flakes inside the crack itself? Microstructure can affect the integrity of gear placements as much as the overall macrostructure.

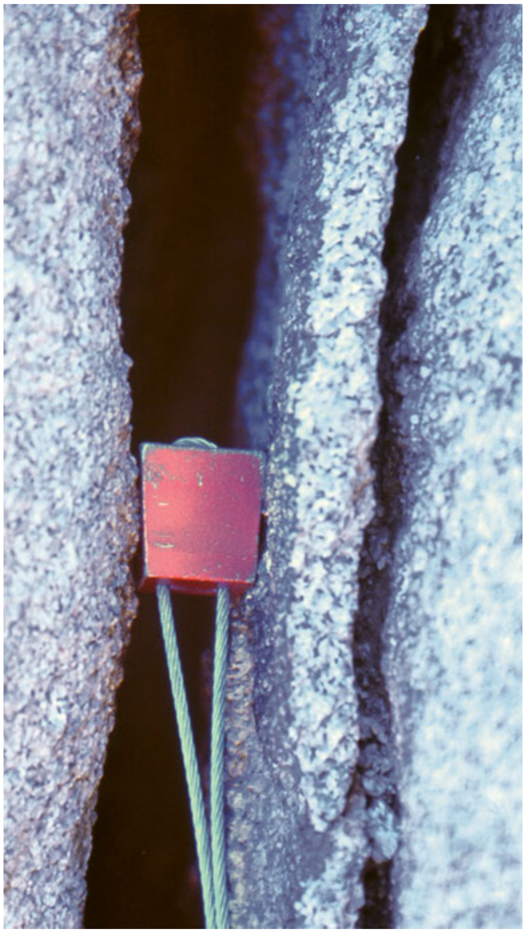

Bad microstructure: The right side of this nut rests on a fragile flake.

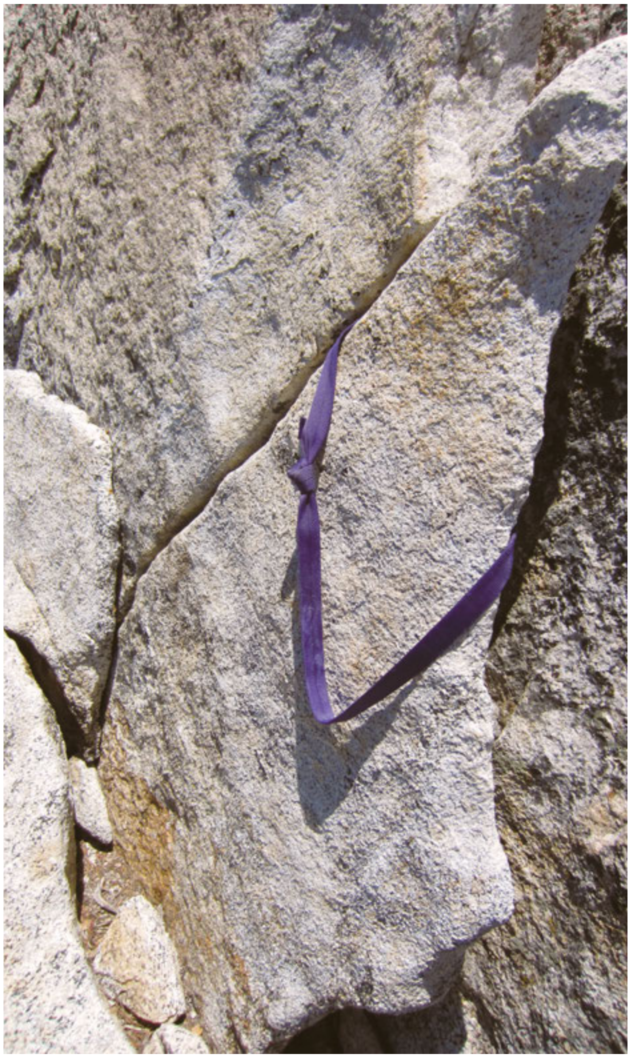

You don't have to be a geologist to figure out that this flake is ready to exfoliate. Many catastrophic anchor failures are due to poor rock structure.

Natural Anchors

Natural anchors utilize the natural features found at the crag environment, such as trees, and the configuration of the rock itself. Trees are plentiful in some areas, rare in others, like in a desert environment. When assessing the reliability of a tree, there are several considerations. Is the tree alive or dead? What is the environment (dry or wet)? What is the diameter of the tree's trunk? How deeply rooted is the tree? When using a tree as part of an anchor system, a good rule of thumb is to choose a live, healthy tree with a minimum trunk diameter of approximately 12 inches. Generally speaking, trees rooted in a thin carpet of topsoil tend to be less reliable as anchor points than those deeply rooted in a thicker layer of topsoil. Climate should also be taken into account when considering the use of trees as anchor points. Trees in areas where the climate is wet and humid and the topsoil is thin can be less desirable as anchors than those in an area of thin topsoil but a dry climate.

The rock itself can be used for anchoring. Look for large spikes or horns of rock attached to the main rock structure to tie off as part of the anchor. A tunnel in a solid rock structure is called a thread and is utilized by threading a sling or cord, or tying a rope, through the tunnel. Limestone is a rock type with many threads, whereas threads in granite are a rarity.

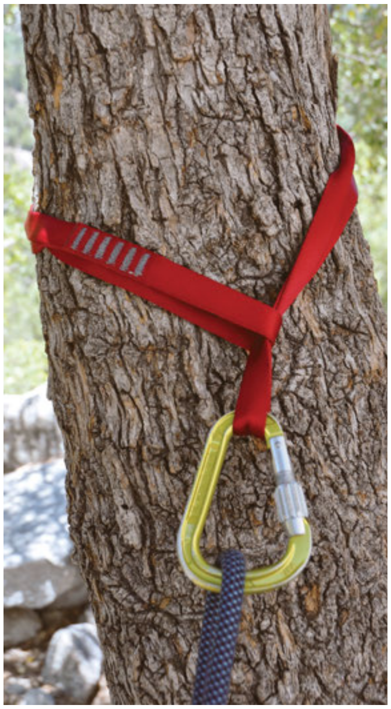

OK. A properly girth-hitched nylon sling.

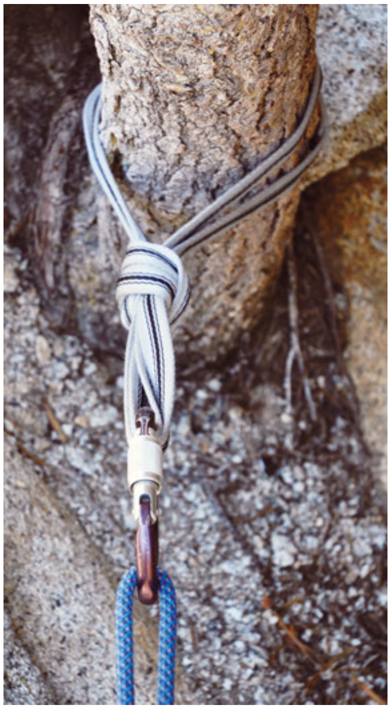

Good. A double-length (48-inch) nylon sling tied with an overhand knot makes the sling itself redundant.

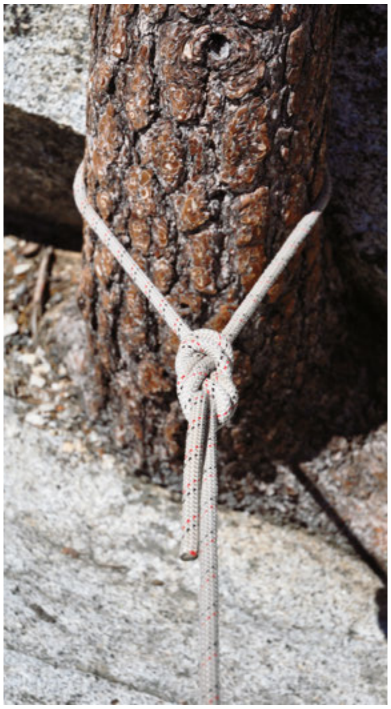

Good. A figure eight follow-through knot used to tie the anchor rope directly to the tree.

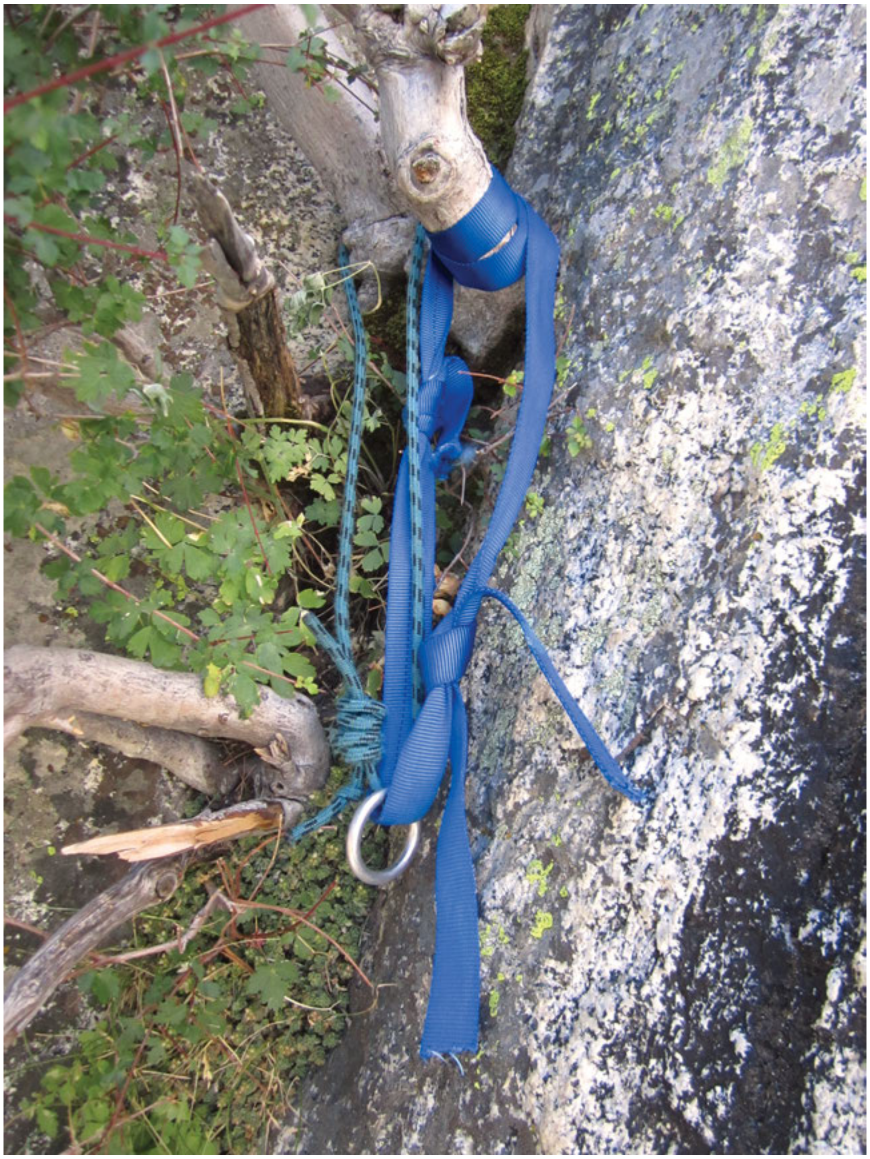

I (BG) came across this natural anchor several hundred feet up a route on Tahquitz Rock, near Idyllwild, California. Someone had obviously rappelled from it, probably to escape an afternoon thunderstorm. Although the anchor has two basic components, each trunk is less than 3 inches in diameter, and the master point where you'd thread your rappel rope is a nonredundant, single aluminum rappel ring (albeit rated at 3,000 lbs.). The real problem I have with this rappel anchor is simply the size of the "tree" itself. I'm glad I didn't have to rappel from it!

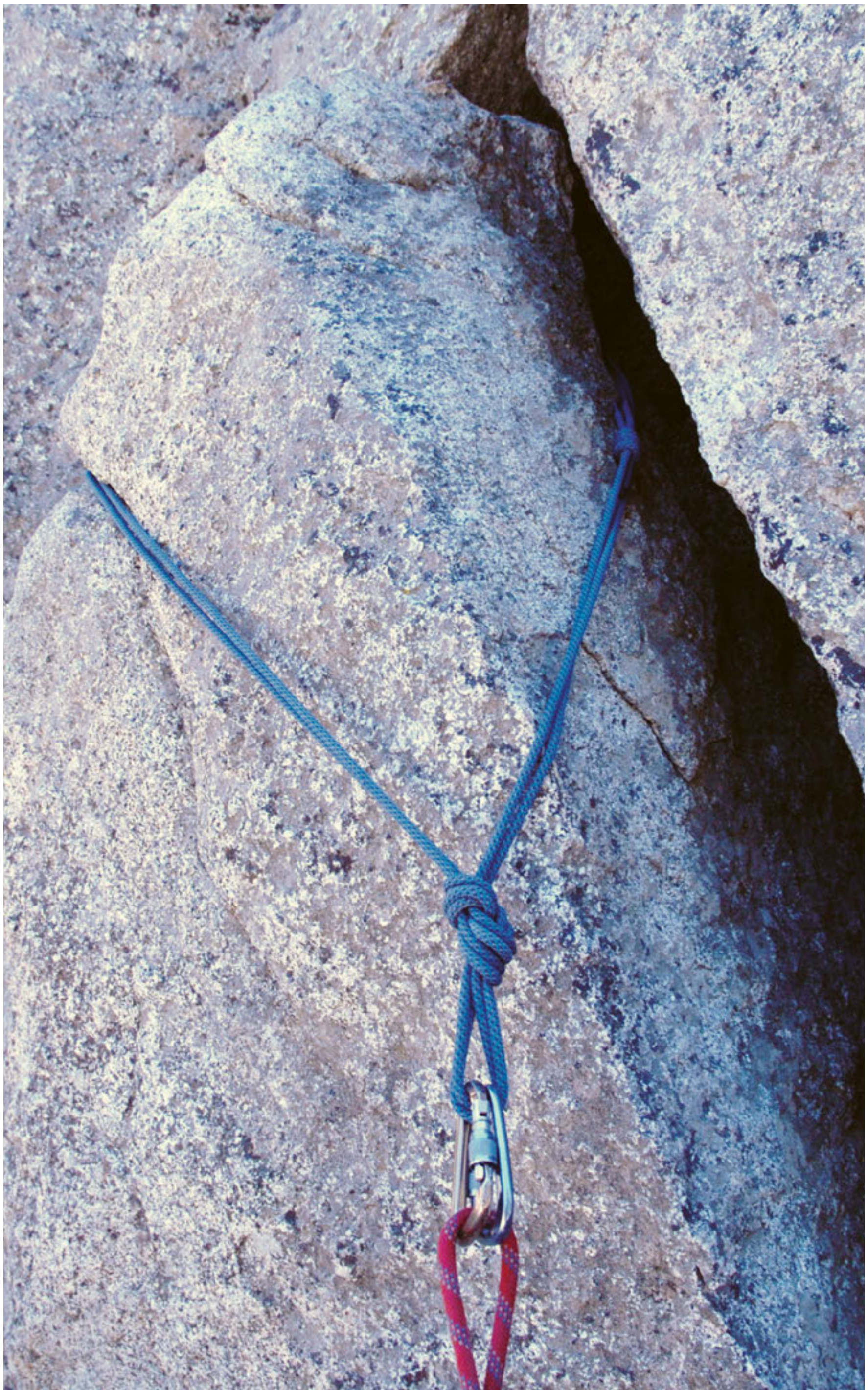

A solid block tied off with a doubled cordelette using a figure eight knot, making the cordelette itself redundant.

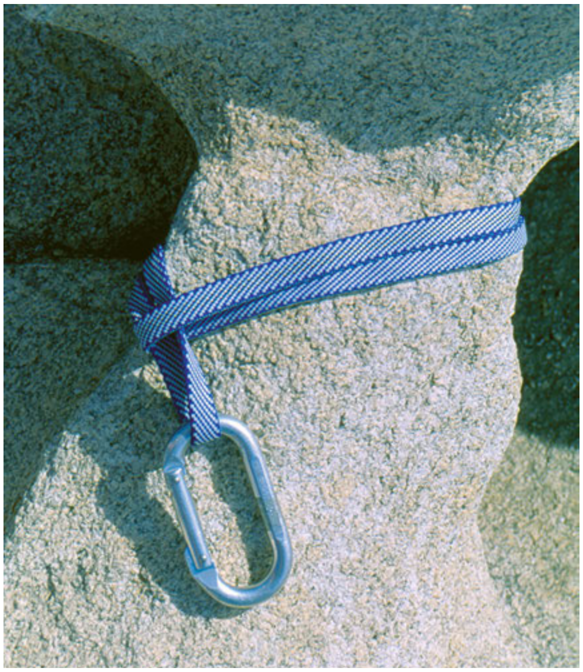

A sling threaded through a tunnel with a girth-hitch. This is called a "thread."

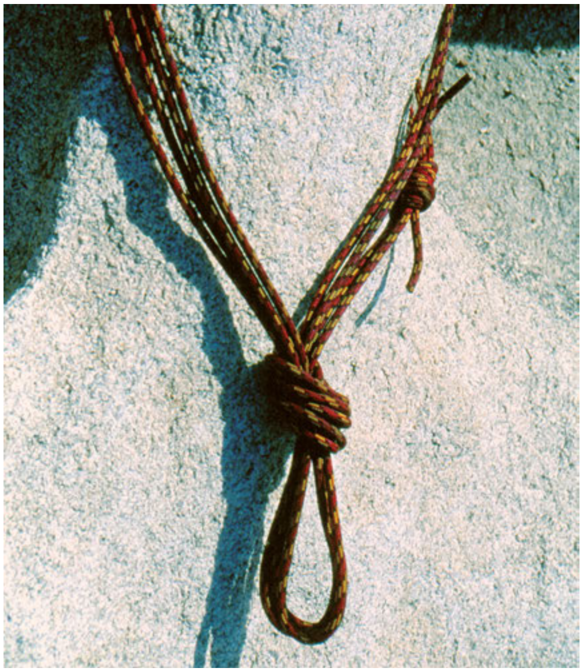

The same thread with a doubled cordelette tied with a figure eight. This adds redundancy—if one loop is cut, three loops back it up.

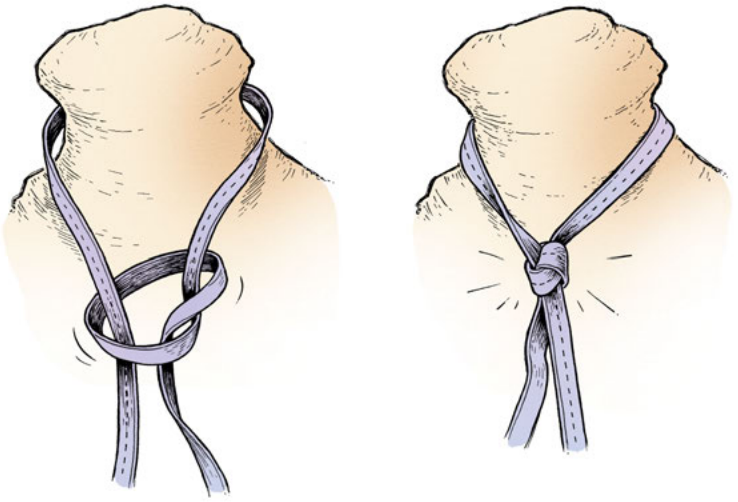

A slip hitch used to tie off a knob of rock.

Monolithic Anchors

Instructors will often encounter natural anchors that are so large and reliable, they may be trusted as the sole component in an anchoring system. Titanic boulders and large trees are often unquestionably reliable, and that level of absolute certainty is how the merit of a monolith should be evaluated. If, after careful examination and consideration, there is any doubt of the integrity, security, and strength of a monolithic component, it is not monolithic and should not be used, or should be backed up appropriately. Common examples include living, deeply rooted trees that are larger in diameter than the entire climbing team; or a boulder the size of a small vehicle, firmly attached and embedded in the landscape.

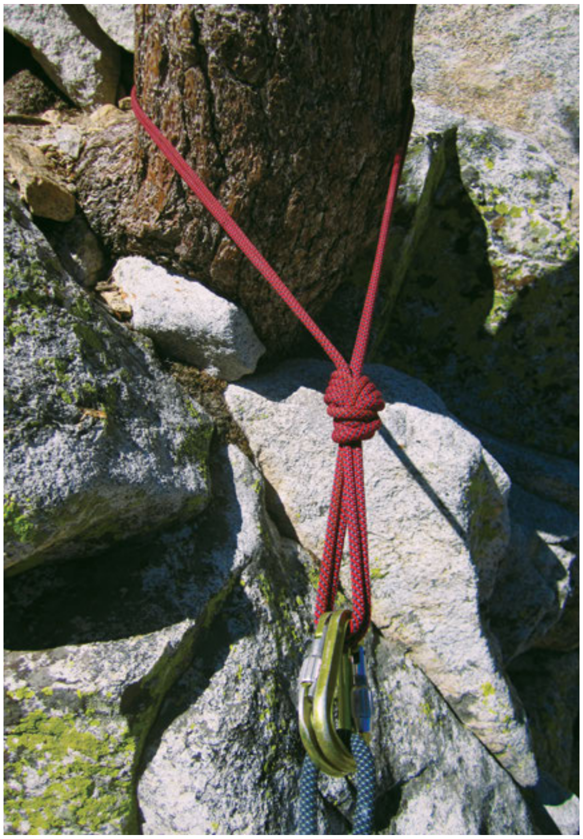

A single pine tree used for a toprope anchor. Both the cord and the carabiners are doubled for redundancy in the rigging.

Just make sure the sling or rope around the anchor is redundant. For example, when rigging a rappel anchor around a tree, use two separate slings with two rappel rings to gain redundancy in the anchor system, at least in the rigging. When rigging a belay or toprope anchor, loop two strands of the cordelette around the tree, then tie a figure eight knot for a two-loop master point. Clip in with two locking carabiners, opposed and reversed, to achieve redundancy in the anchor rigging (although technically, one single tree is nonredundant). Use caution and sound judgment when using a nonredundant natural anchor.

Passive Protection

The Evolution of Chockcraft

A chockstone is simply a rock wedged in a crack. Naturally occurring chockstones can be as small as a pebble or as big as a house. The notion of using a chockstone for an anchor dates back to the origins of the sport. In the late 1800s, in the British Isles, rock climbers began using natural chockstones for anchors by slinging a cord around them and attaching their rope to the sling with a carabiner. The use of artificial chockstones—called chocks, or more commonly, nuts—began in the early 1960s, at a cliff in North Wales of all places, at a crag named Clogwyn du'r Arddu. The hike up to the crag followed a railroad track, and some ambling climber picked up a nut along the way and pocketed it. Up on the cliff he threaded a small cord through the nut before wedging it in a constriction in a thin crack. Thus the subtle art of chockcraft was born.

In American rock climbing, pitons were used almost exclusively for protection and anchors until the 1970s. In Europe, pitons were made of soft iron and once hammered into a crack were nearly impossible to remove and reuse. Legendary

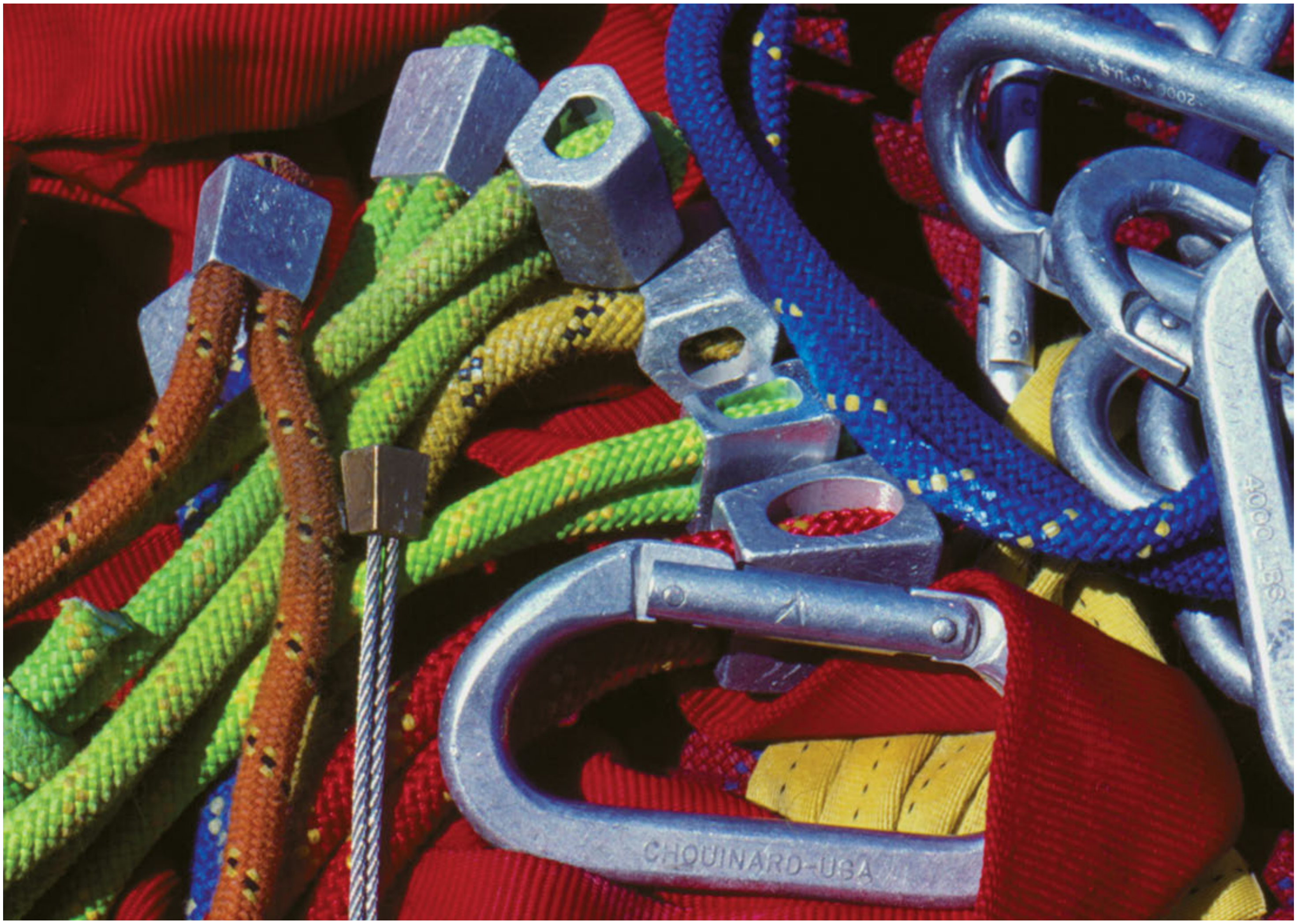

A selection of chocks from the 1970s. Nuts have evolved over the years but are still based on the same original basic designs.

Now You See It, Now You Don't

A friend of mine (BG) put up a new route at Joshua Tree—a 40-foot-high sport climb with five bolts—up the face of a massive block that was a facet of a larger cliff. One day I got a phone call: ''Tony's route fell down!'' I didn't believe it until I walked out there and saw it with my own eyes. The gigantic block was top heavy and had simply toppled over, with the side where Tony's route was now straight down in the dirt, leaving behind a void in the cliff the size of a small house. I got down on my hands and knees and peered underneath. I could see one of the bolt hangers! Bouldering legend Chris Sharma visited the site shortly thereafter, climbing what is now one of Joshua Tree's most difficult boulder problems, up the newly exposed overhanging face of one side of the block.

NEUTRINO
NEUTRINO BD

A selection of modern-day nuts.

American climber John Salathe, a wrought-iron worker by trade, developed the first hard steel pitons, forged from an old Ford Model A axle, which he used for his famous ascents in Yosemite Valley during the 1940s. These high carbon steel pitons could be driven and then removed, over and over again.

Yvon Chouinard refined and innovated the design of chrome molly steel pitons from 1957 to 1965, improving on Salathe's designs with the introduction of knifeblade, horizontal (called the Lost Arrow), and angle pitons. These pitons revolutionized big wall climbing in Yosemite during the "Golden Age" of the 1960s, where hundreds of placements were required for the ultimate big wall climbs in Yosemite, like El Capitan. Once placed, they could be removed by the second, leaving the climbing route in the same condition for the next climbing team. Climbing standards in Yosemite led the world at the time.

But it came with a price. On popular climbs in Yosemite, the repeated pounding and removal of hard steel pitons began to permanently damage the cracks, leaving ugly "pin scars" every few feet up

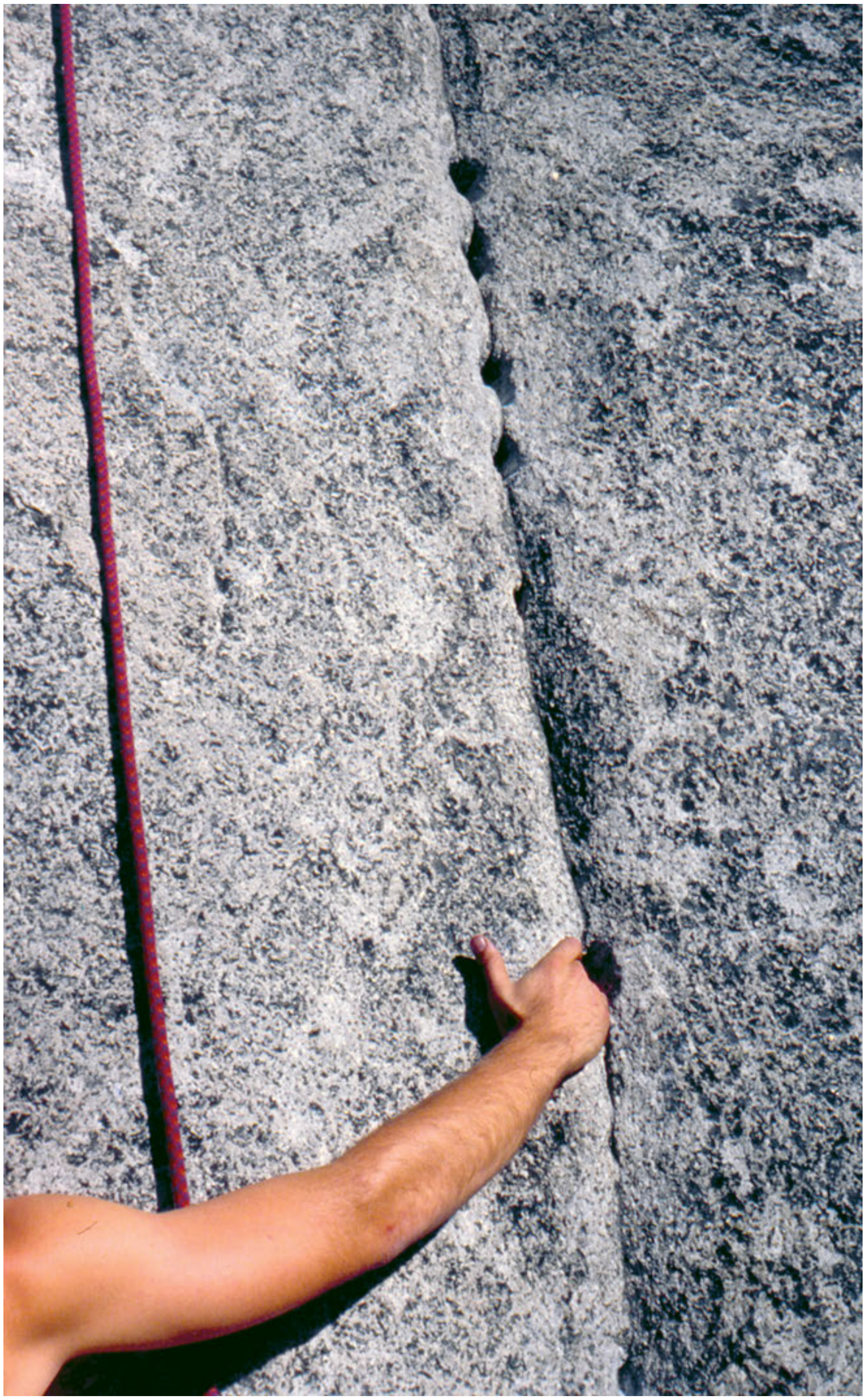

Piton scars on a Yosemite crack.

crack systems. Cracks were getting "beat out," and something had to be done. In Yosemite the National Park Service actually closed down a few climbs because of piton damage.

When the great American climber Royal Robbins made a trip to England in the 1960s, he saw how effective nuts could be, and he imported the idea back to Yosemite. His 1967 ascent of The Nutcracker, one of Yosemite's most popular climbs, was done entirely with nuts, Royal's way of showing that nuts were a viable alternative to the destructive pitons. Today there still are piton scars on the route, a testament to how slow American climbers were to embrace the new and more gentle technology of chockcraft—a big change from bashing hard steel pitons into cracks with heavy blows from a hammer.

The change was finally precipitated by the fact that many cracks were simply being destroyed. Even granite is relatively soft when compared to cold hard steel. But it wasn't until Yvon Chouinard introduced chocks to American rock climbers in his 1972 equipment catalog, and Doug Robinson espoused the virtues of nuts in his seminal treatise, *The Whole Natural Art of Protection,* that the American climbing community firmly embraced the idea of "clean climbing," a new ethic where climbing anchors were placed and removed without scarring or damaging the rock.

Today there are thin crack climbs in Yosemite where for hundreds of feet every finger jam is in an ancient piton scar, although now instead of using pitons, nuts can be slotted into the V-shaped bottom of the old pin scars.

Artificial chocks now come in a dazzling array of shapes and sizes, the largest ones capable of holding over 3,000 pounds, and the tiniest micro-nuts designed to hold body weight only. The Hexentric, commonly called a Hex, is a unique, six-sided nut with four distinct orientations of placement, first introduced by Chouinard Equipment in 1971. It was followed by the Stopper in 1972, with its simple, but effective, tapered trapezoidal shape. Although there have been many new designs

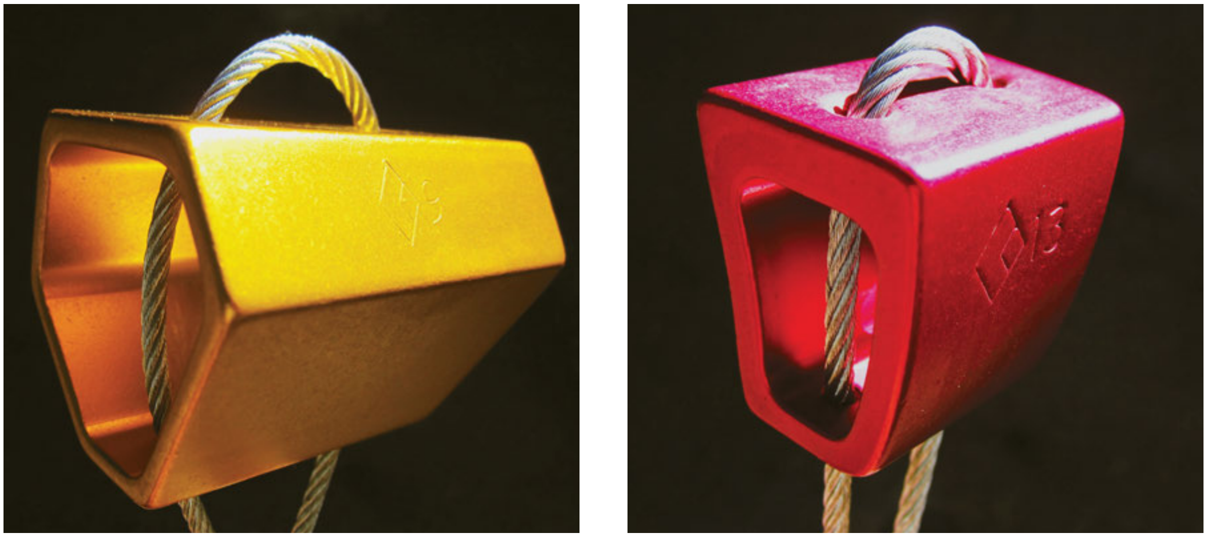

The classic designs of the Hex (left) and the Stopper (right) have changed little since their inception in the early 1970s.

introduced since then, they are basically variations on a theme to these classic and timeless designs, which are still as viable today as they were more than forty years ago.

Another ingenious design, called the Tricam, invented by Greg Lowe in 1973 and available to the public in 1981, is essentially a single cam that can be used either passively or actively. Since it has a tapered design, with a point on one end, it can be wedged like a nut (called a passive placement) or used like a cam (called an active placement), where a mechanical action (i.e., camming) takes place. The camming action occurs when the sling is loaded on the back, or spine, of the cam, between two rails that contact the rock on one side of the crack, creating a force that pivots like a fulcrum onto the

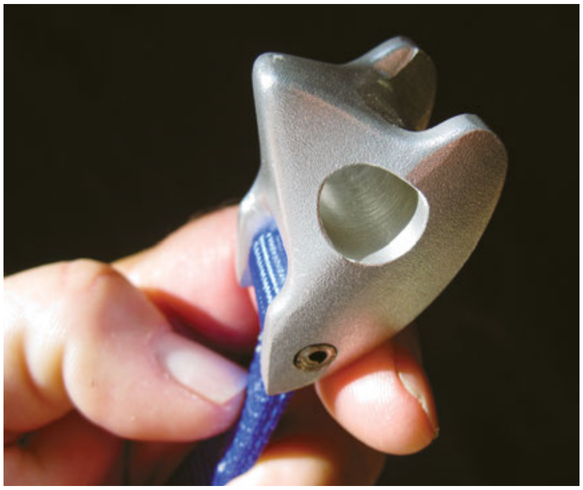

The Lowe Tricam.

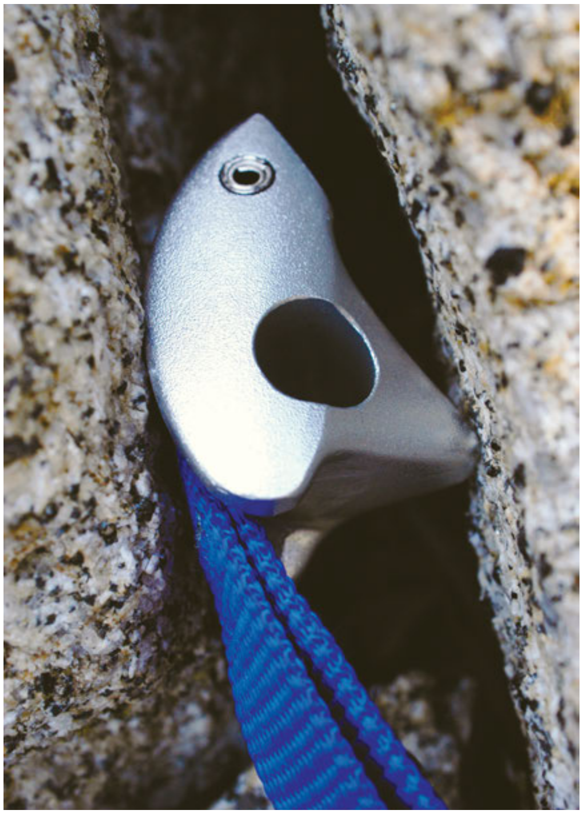

Tricam in camming mode.

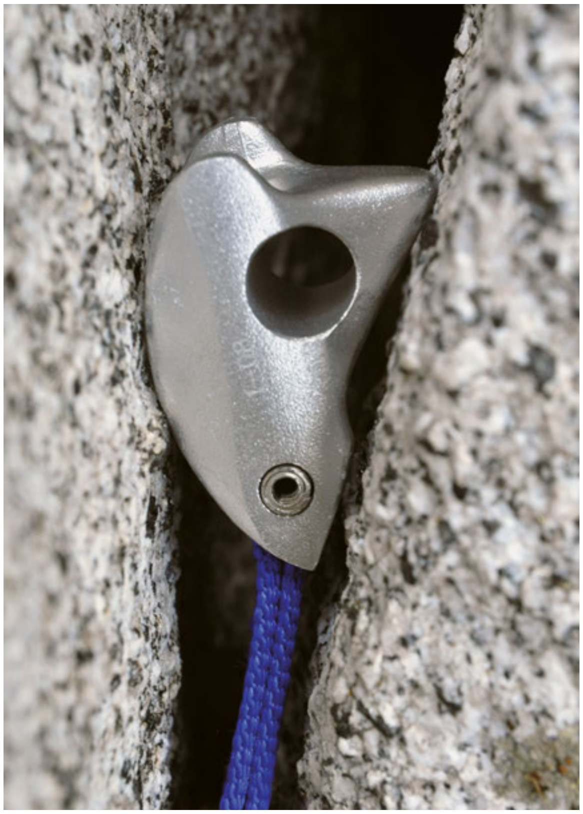

Tricam in passive mode.

Assessing Nut Placement

To assess a nut placement, remember these three key elements:

1. Rock structure
2. Direction of pull
3. Surface contact

A good acronym is SOS:

- **S** Structural integrity of the rock
- **O** Orientation (direction of pull)
- **S** Surface contact

pointed end on the other side of the crack. The design is useful for many horizontal crack situations, but can be somewhat difficult to remove with one hand or once it is weighted.

Be aware that in the smaller-size Tricams, when set in the passive mode, a Stopper of similar size is considerably stronger. The reason is because on the Tricam, the load is being placed directly on the scroll pin that connects the webbing to the body of the Tricam. (A pink Tricam in passive mode is rated at 6 kN; a similarly sized Stopper, 10 kN. In camming mode the pink Tricam is rated at 9 kN.)

When placing a nut, or any other piece of gear for that matter, again the first thing to consider is the overall integrity of the rock itself. The importance of rock assessment cannot be overemphasized. Nuts have very low holding power in soft sandstone, rotten, or flaky rock. Avoid placing nuts in cracks under or around detached blocks, or in cracks behind loose flakes. Look for "straight-in" cracks in massive rock structure, where the crack runs perpendicular to the plane of the rock face.

Once a good crack system is found, look for obvious constrictions in the crack itself. A "bottleneck" placement is found where the crack tapers drastically, and the proper size nut is fitted in the narrowing constriction.

With a basic tapered nut, like the Stopper, the preferred placement is in the narrow configuration, since this setting has the most surface contact and stability. The wider endwise configuration is an option for narrow slots and shallow cracks, but ultimately has less surface contact and generally less stability.

The typical nut placement is in a vertical crack, but horizontal cracks will work if there is a narrowing at the lip of the crack and you can slide a nut in

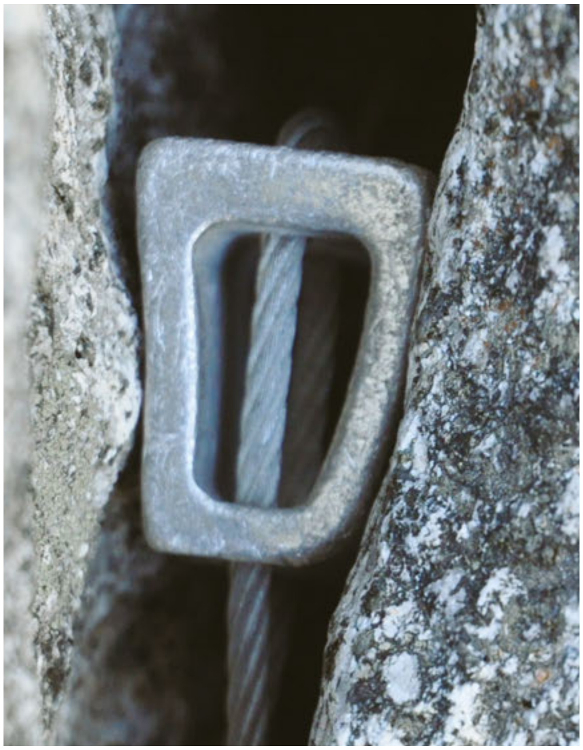

Stopper in a bottleneck placement. There is simply no way that in a downward pull the nut could be pulled through the bottleneck—something would have to give, either the rock itself or the nut or wire cable breaking.

AMGA_i-258_3pp_CS55.indd 118 2/27/14 11:54 AM

from the side, then pull it into the constriction.

The real art of chockcraft comes into play with the more subtle placements. Look for any slight variations in the walls of the crack. When placing a nut, aim for maximum surface contact between the metal faces of the chock and the walls of the crack.

When the walls of the crack are virtually parallel sided, using the camming action of a Tricam or Hex is the best option for a nut placement, although this is territory that specific camming devices were designed for.

After rock quality has been assessed, the next concern when placing a nut is the direction of pull. In what direction will the chock be loaded? Most placements can withstand a pull in only *one* direction. While the nut may be able to withstand a load of 2,000 pounds in that one direction, the slightest tug in the opposite direction might jerk the nut right out of its placement. When incorporating a nut placement into an overall anchor system, look at the ultimate direction the anchor system will be loaded and orient placements in a line toward this focal point (called the master point).

Setting a nut properly is also important. Many novice climbers make a great nut placement but fail to set it properly, which makes the nut susceptible

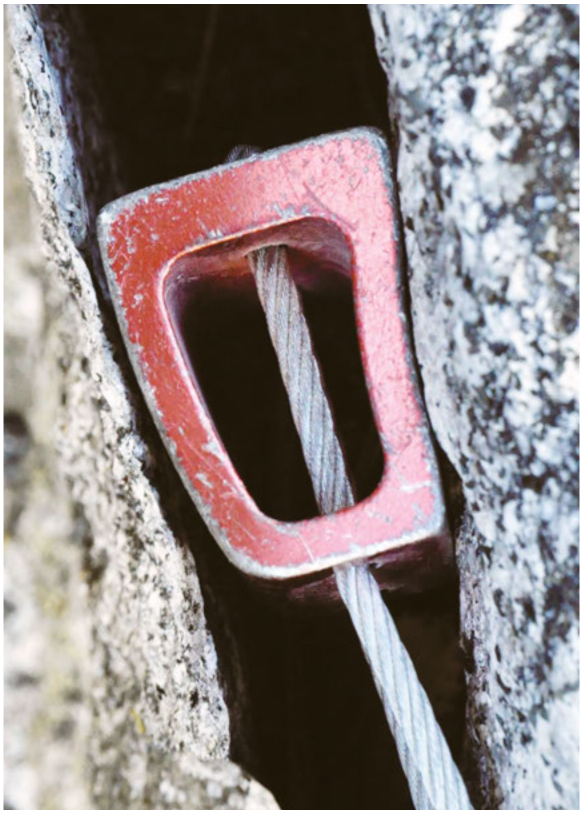

Excellent. This Stopper placement is in good, solid rock and has flush surface contact on both sides of the nut.

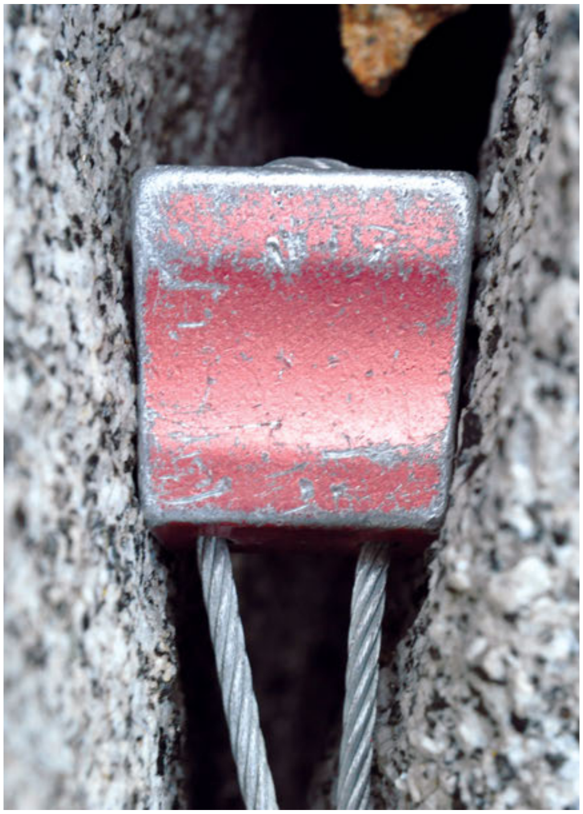

Good. This endwise Stopper placement has good surface contact on both sides.

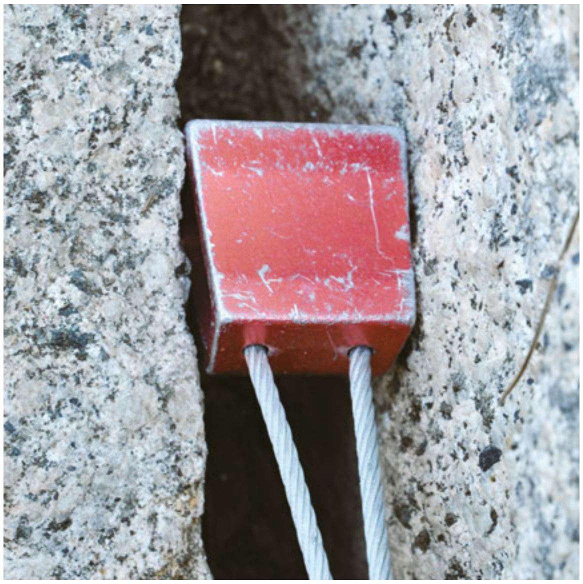

Bad. The left side of this nut lacks surface contact with the rock.

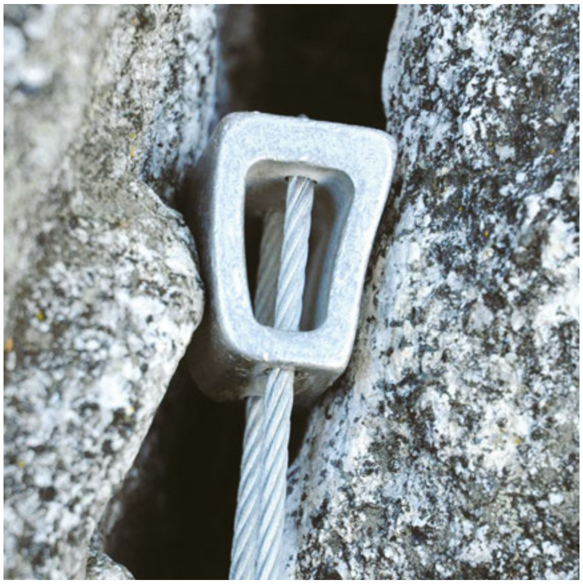

Marginal. This Stopper is in a good bottleneck, but since it lacks flush contact on the right side, a slightly outward pull will pluck it from its placement.

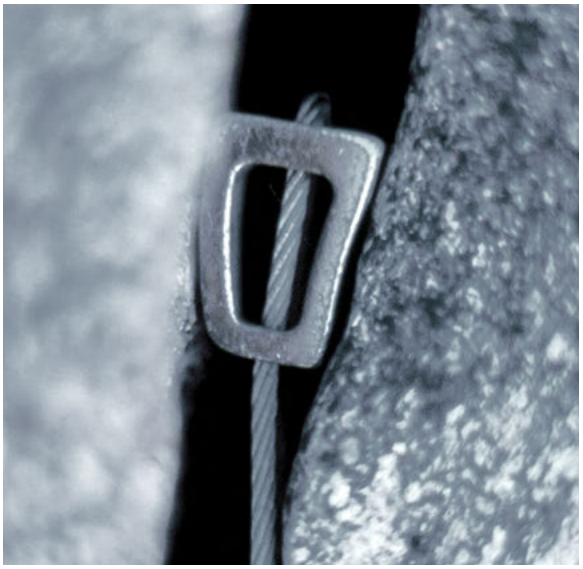

Good. The left side of this nut is nearly 100 percent flush, and the curve of the nut on its right side fits the curve of the crack.

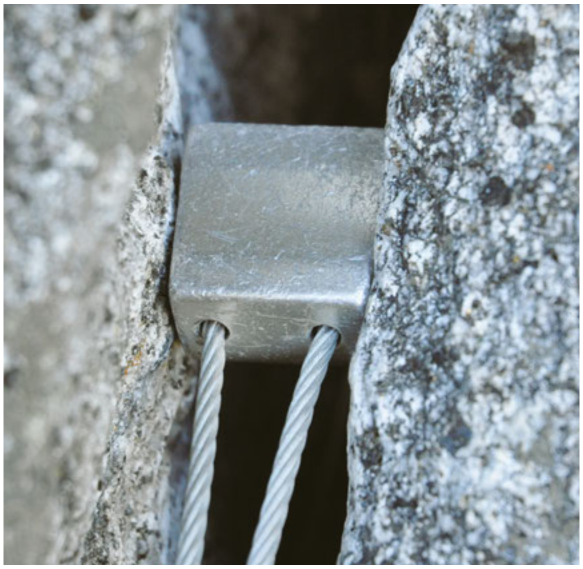

Excellent. This nut has good surface contact on both sides, plus the lip on the right side of the crack protects against any outward force.

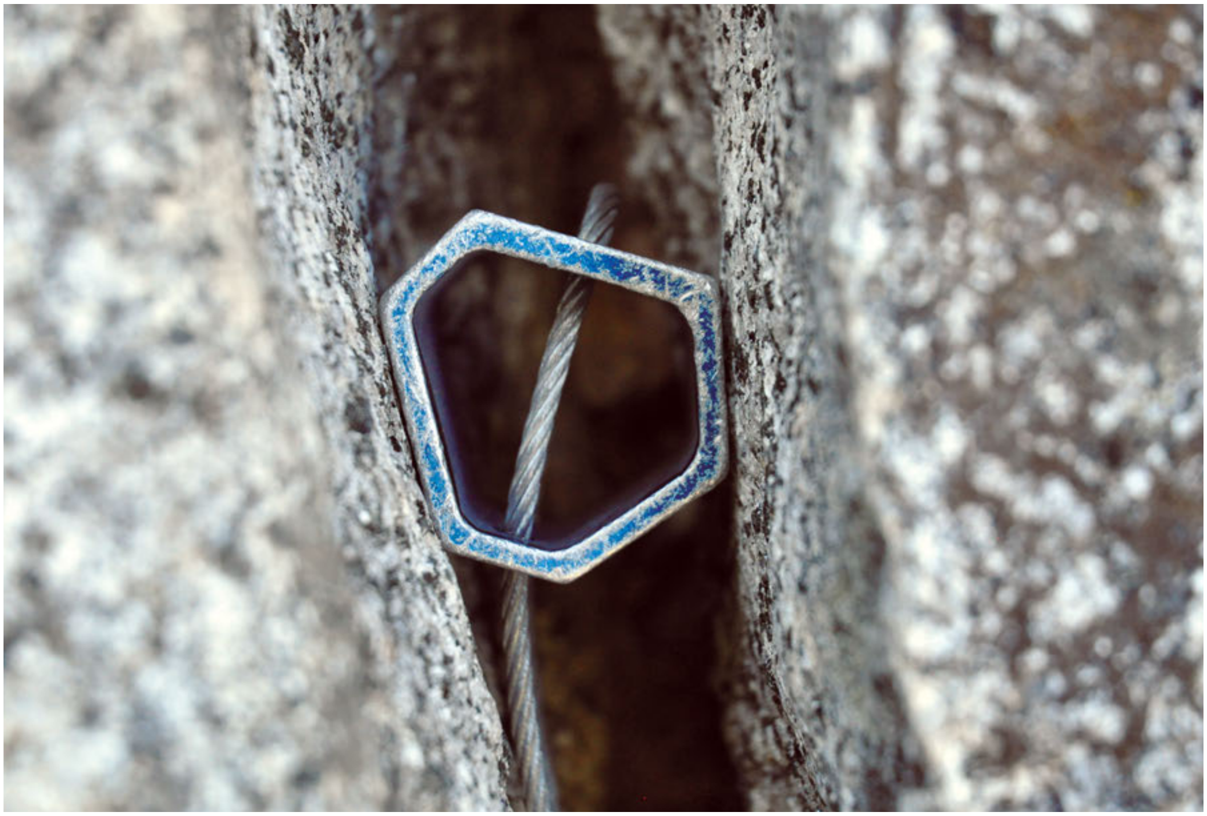

Excellent. This Hex placement has near complete contact on both faces of the nut in a solid straight-in crack. Loading the nut's cable will kick in the camming action of the Hex.

to levering out of its placement if pulled from a different angle than intended. Setting the piece is accomplished by simply applying several stout tugs in the direction the piece will be loaded, most easily accomplished by attaching a sling to the nut with a carabiner and yanking on the sling, firmly wedging the nut in its intended placement. While this definitely makes the nut more difficult to remove, it is an important concept that many novices miss.

Carrying a nut tool, which is a metal pick designed for nut removal, facilitates "cleaning" nuts. To clean a nut can be as easy as yanking it in the opposite direction from the intended direction of pull, but be careful with recalcitrant nuts that can suddenly pop out and hit your face or teeth. Yanking a piece out can also send a hand bashing into the rock, scraping knuckles. A better approach to removing a nut is to use the nut tool, giving the nut a tap opposite from the direction of loading. For larger nuts an easy way to loosen them is to tap the nut with a carabiner, metal to metal.

To become skilled at chockcraft takes practice. Look for good rock structure, preferably a straight-in crack. Then look for any obvious V-shaped constrictions in the crack. If there is nothing obvious, look for any subtle narrowing of the width of the crack. Practice nut placements, aiming for maximum surface contact between metal and rock, keeping in

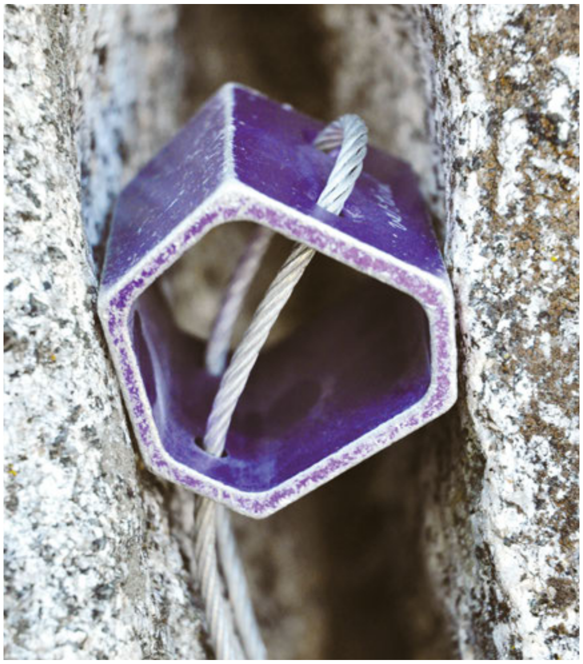

Good. This Hex is in a narrowing pocket, and both sides of the nut have good surface contact.

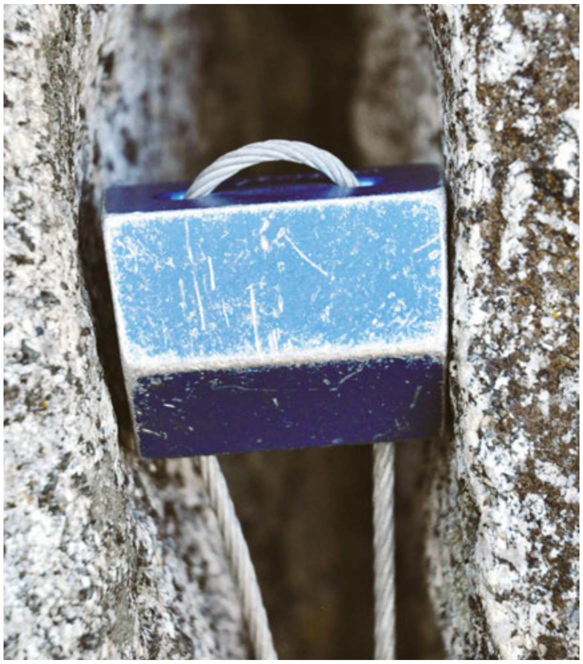

Bad. No surface contact on the left side makes this Hex placement likely to fail with the slightest outward force.

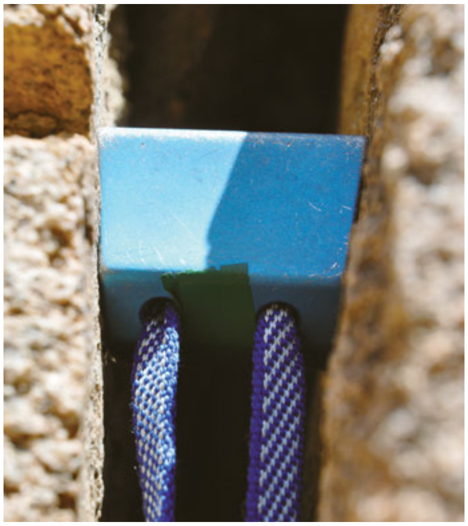

Excellent. This endwise placement has great surface contact on both sides.

A nut tool is indispensable for removing chocks. Here are two models: Black Diamond (top) and Metolius (bottom).

To remove a nut with a nut tool, inspect the placement and determine the intended direction of pull, then tap in the opposite direction.

mind the paramount importance of direction of pull. Set the nut with a sharp tug. Once set, it should not move, pivot, or wiggle in its placement when you test it with a slight tug in the opposite direction of the intended direction of loading. Practice will help to develop the knack for seeing what size nut is needed for a particular placement, then selecting the proper nut and fitting it into the crack on the first try. Every nut placement is different, some less than perfect, some great, some worthless. It is important to have enough knowledge to know what's good and what's not, and what constitutes a placement that can be trusted.

Active Protection

In the mid-1970s a stout, muscular fellow by the name of Ray Jardine could often be seen peering through binoculars, gazing upward at the various nooks and crannies on the walls of Yosemite Valley. With his thick beard and glasses, he looked like a bird watcher, but Ray wasn't looking for birds. The bulging forearms gave it away—Ray was a climber, and he was looking for the ultimate crack: one of those perfectly straight cracks that split Yosemite's steep walls like a surgeon's incision, shooting upward for a hundred feet, uninterrupted.

Ray had invented a new technology—the spring-loaded camming device, or SLCD—that allowed him to place reliable protection in even perfectly parallel-sided cracks. When he found his ultimate crack climb, he swore his partners to secrecy and set out on a mission: to climb the most difficult crack ever climbed in Yosemite. He named it the *Phoenix*—a fingertip to hand-size crack on a gently overhanging wall high above Cascade Falls in the lower valley. After dozens of attempts using his new-fangled technology, he finally succeeded in climbing Yosemite's first 5.13. Ray called his miracle invention the "Friend," and soon the word was out. Some climbers called it "cheating," while others claimed it was "the greatest invention since the nylon rope."

Marketed by Wild Country, the Friend soon became an integral part of every rock climber's rack. Ray soon retired from climbing and, financed by his proceeds from the licensing of the Friend, went on to sail around the world, hike the Pacific Crest Trail, row across the Atlantic, and ski to the South Pole.

The idea of the SLCD, or "camming device" for short, is simple in concept yet complex in design. Jardine's original design consisted of a unit with a rigid aluminum shaft connected by an axle to four independent, spring-loaded aluminum cams (called "lobes"). The cams retracted via a trigger bar that slid up and down a slot in the shaft. The unit was fitted into a parallel-sided crack with the cams retracted; when weight was applied to a sling tied

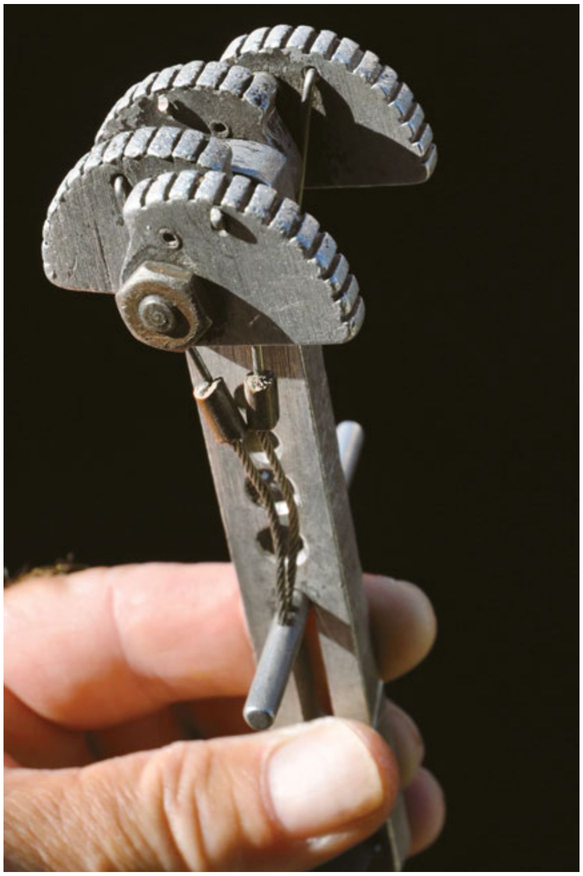

The original Wild Country Friend was one of the greatest innovations in rock climbing.

Spring-loaded camming devices have become an integral part of every climber's rack.

into a hole in the bottom of the shaft, the cams were activated in response to the load. To keep the unit from being pulled out of the crack, a corresponding force held it in place. The downward force in the direction of the shaft was transferred outward at the cams, which generated an outward force against the walls of the crack.

The disadvantage of Ray's design was that a rigid shaft could not flex or bend in the direction of pull, an especially troubling problem for placements in horizontal cracks.

Today there is a huge array of SLCDs on the market, and the majority of these designs have flexible wire cable shafts instead of rigid ones. One of the biggest improvements since the invention of the Friend was the first double-axle design, called the Camalot, introduced by Black Diamond Equipment, which allows for a much greater range of placement of the cams. Now, in addition to units with four cam lobes, there are TCUs (three-cam units) and offset cams (for flared cracks).

The Black Diamond Camalot was the first double-axle design.

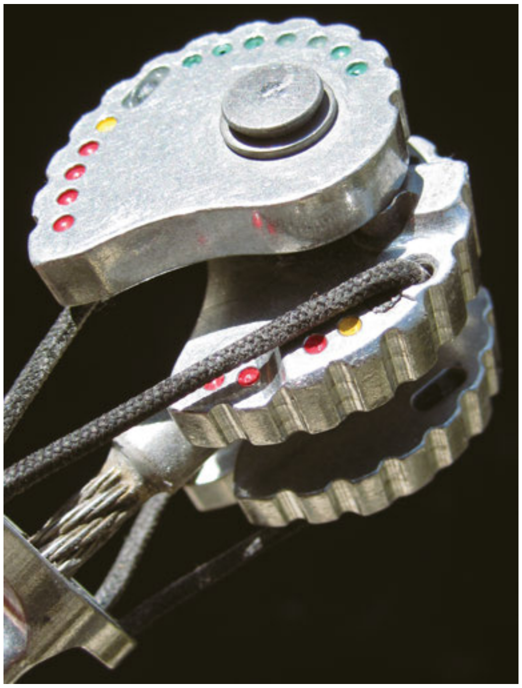

The Metolius Power Cam has color-coded dots that help you assess your placement.

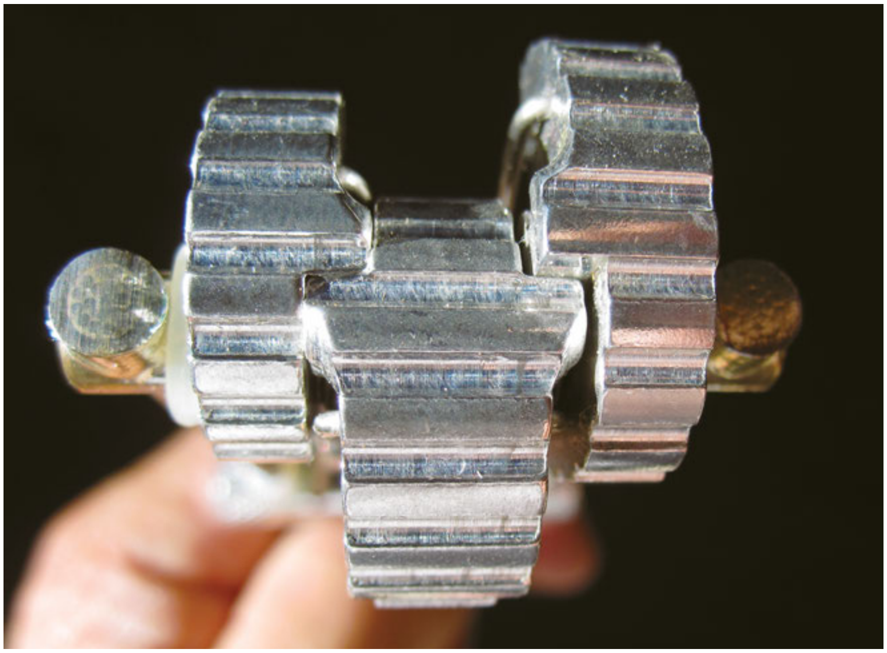

The Metolius offset TCU (three-cam unit) works well in slightly flaring cracks.

Placing an SLCD

When placing an SLCD, the first thing to consider is rock quality. SLCDs can fail if the rock is soft, brittle, or loose. They can easily pull out if placed behind a small, loose block or thin flake of rock. In solid granite, in an ideal placement, a Black Diamond Camalot can hold as much as 14 kN (3,147 lbs.) Do not rely on a camming device to hold in very soft sandstone, or in rotten or flaky rock. Cam manufacturer Metolius advises: "Rock fails in two basic ways: either a relatively large piece breaks off or the surface layer is crushed under the pressure of the cam lobe, allowing the cam to 'track out.' Instructors must assess the integrity of the rock and choose the soundest possible location for your placements. Look for fractures in and around the walls of a potential placement that could denote weakness, as well as pebbles, crystals, or micro-flakes that could snap off. Be extremely suspicious of placements behind flakes or blocks."

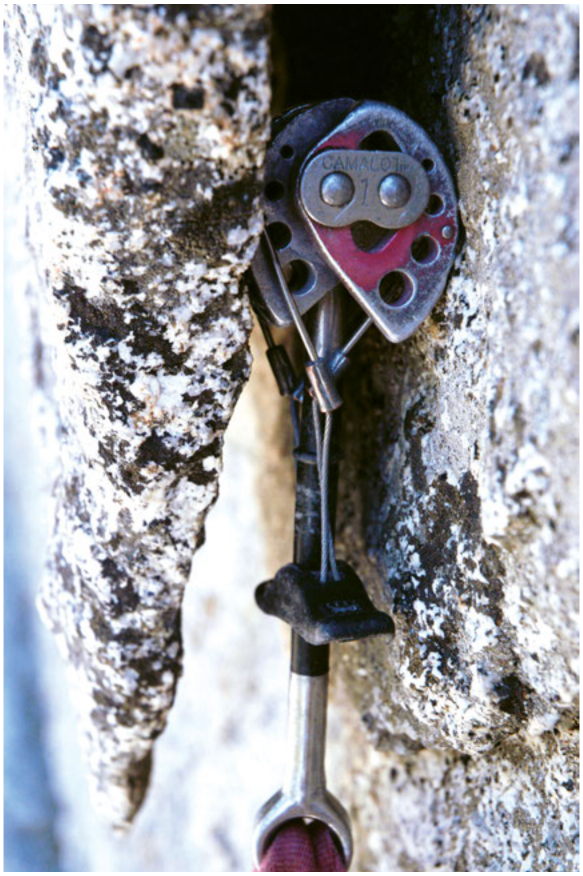

Bad. Any force applied to this Black Diamond Camalot will be converted to an outward force that can pry out and potentially break the flake of rock it's placed behind.

Since they rely on friction to a certain extent, camming devices are not as strong in exceptionally slick or polished rock, or rock that is wet or icy. Again, avoid placements behind detached blocks and loose flakes—the outward expansion of the cams can generate a tremendous force that can pry the rock loose. Look for straight-in cracks in solid rock. A straight-in crack is one that runs perpendicular to the face of the rock, bisecting the rock at a right angle.

When placing a camming device, look for a section of the crack where the walls are uniformly parallel, or where they form a subtle pocket. Avoid widening cracks, where the crack is wider above the cams, as the camming device, due to its spring-loaded design, will naturally have a tendency to wiggle upward as the cam is activated. This phenomenon is known as "walking." This walking movement is most exaggerated when the cam's sling and cable are manipulated from side to side, but it can also happen when the piece is repeatedly weighted and unweighted, as in toproping. In a crack where the walls are uniformly parallel, or where the crack narrows slightly above the cams, if there is any walking, the cams will not open any wider and will stay within acceptable retraction range. As a test, pull the device from side to side to see what walking, if any, occurs. There is only one definitive solution to walking: Cams must be placed in a slot or crack that dead-ends or closes off. Otherwise, walking can be somewhat mitigated by slinging, supervising the placement, and creating anchor systems that self-adjust as direction of load

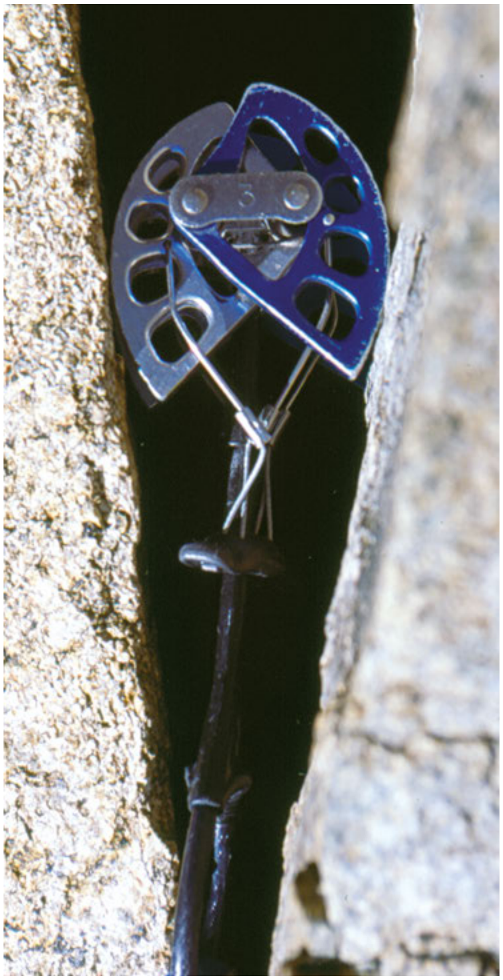

Bad. Even though the cams on this Black Diamond Camalot are within the acceptable range (right around 50 percent retracted), the widening crack above will allow the cams to easily "walk" into the wider section, even with minimal loading and unloading of the device. Avoid this situation, since the cams can potentially walk to an open and unstable position.

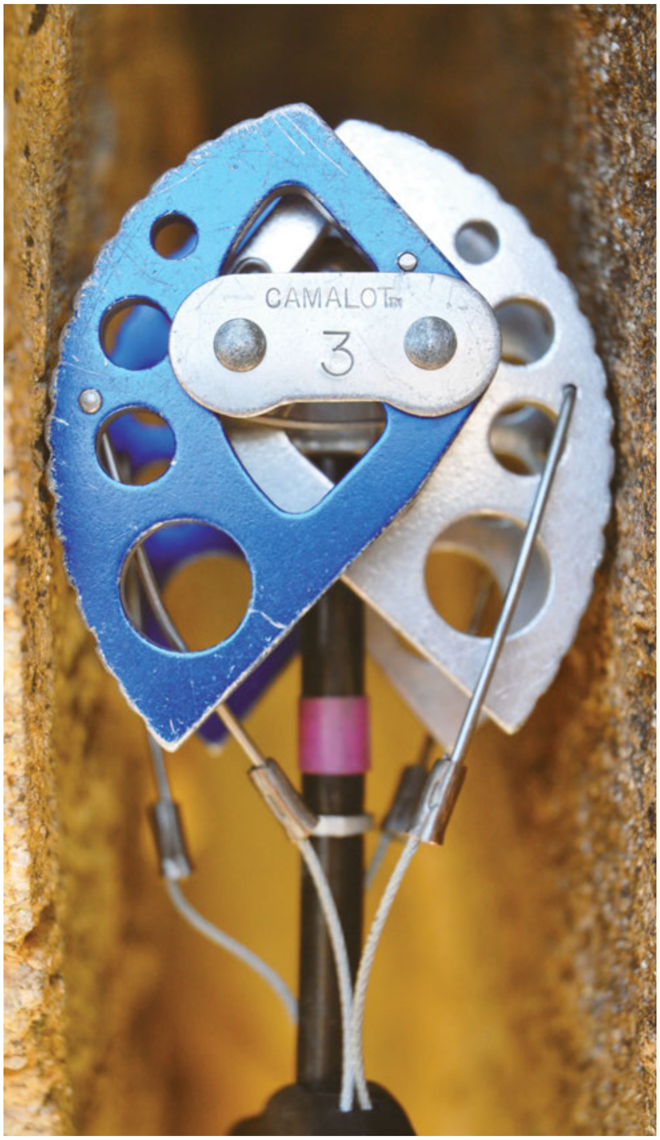

Very good placement. All the cams have good surface contact in a solid, straight-in crack, and the cams are in the recommended range for a Camalot (50 to 90 percent retraction).

varies. The key point here is that walking is something to be aware of and watch for.

Another key to a good placement is the range of retraction on the cams. Black Diamond recommends that the Camalot be placed in the lower to mid-expansion range (50 to 90 percent retraction), while Wild Country advises the following for its single-axle designs: "It is vitally important that all the cams make contact with the sides of the rock, preferably in the middle half of their expansion range (i.e., the cams should be one-quarter to three-quarters open)." Metolius recommends to "select the largest size cam that will fit without getting stuck. Cams should not be placed near the wide end of their expansion range. When a unit is loaded, it expands as the slack is removed from the system and the cams and rock compress. A nearly tipped-out cam won't have enough expansion left to accommodate this process. A loose cam is also more prone to walking and has little range left to adjust."

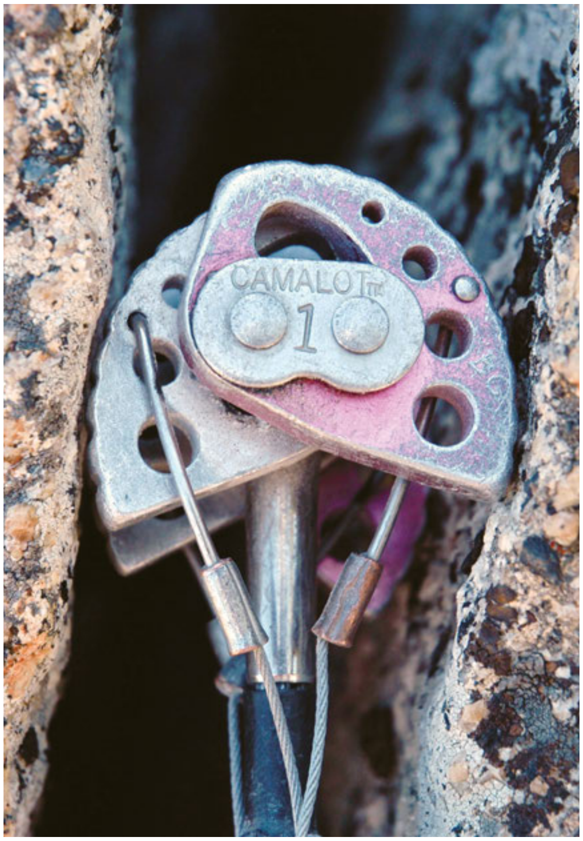

Bad. The two outside cams are not in the acceptable range for a Camalot (too wide), and they don't have flush contact with the walls of the crack.

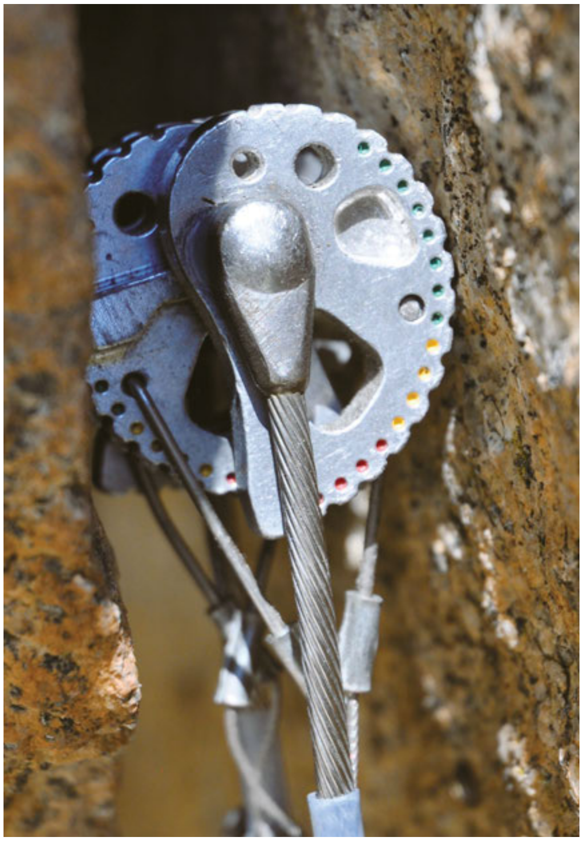

Good. This Metolius Power Cam displays optimal green "range finder" dots in a solid, parallel-sided crack.

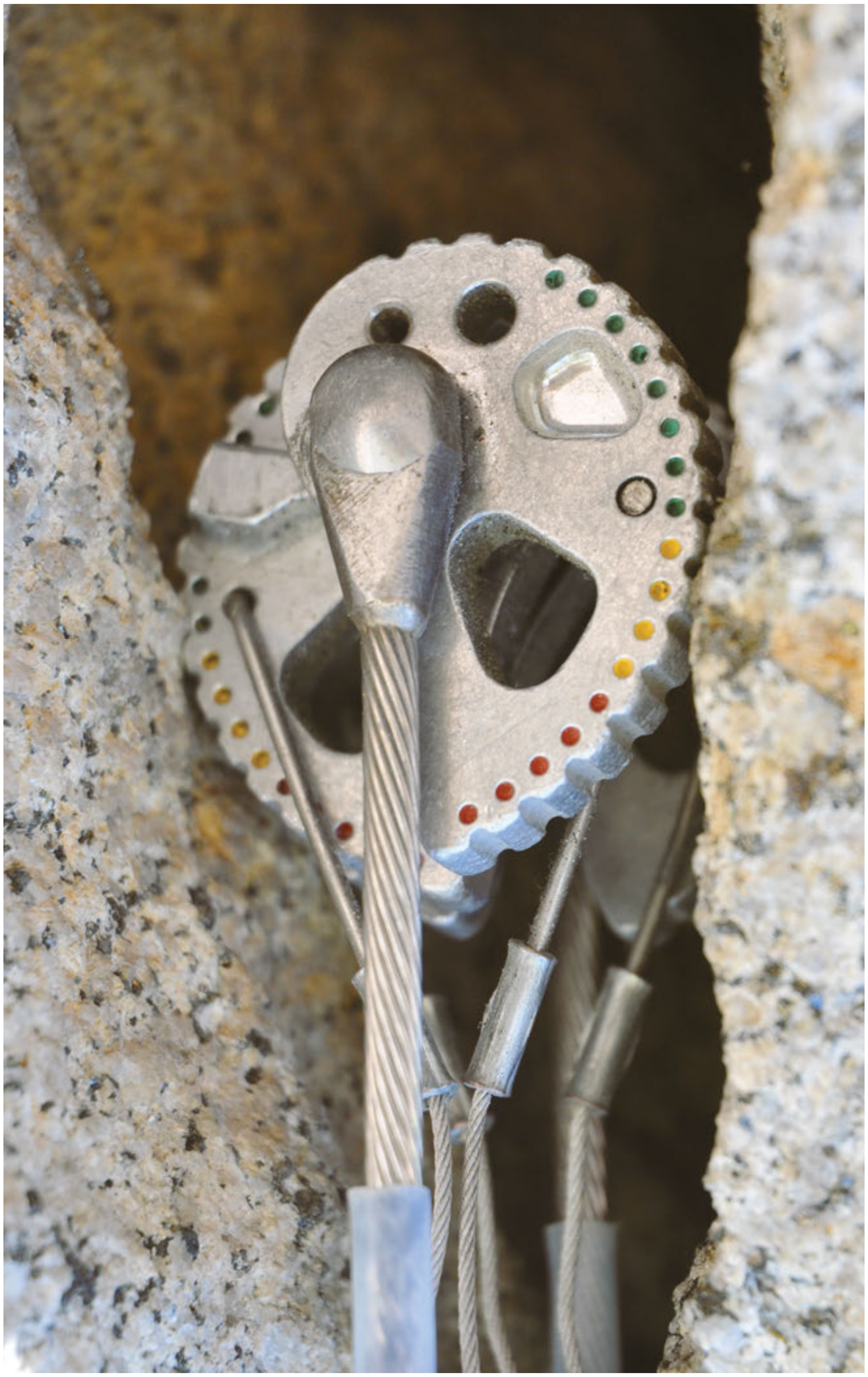

Poor. Although the range of retraction is acceptable, this Metolius Power Cam could easily walk up into the wider pod in the crack above the cams, rendering the placement unstable. Also, the outside right cam has poor surface contact and is too close to the edge of the crack.

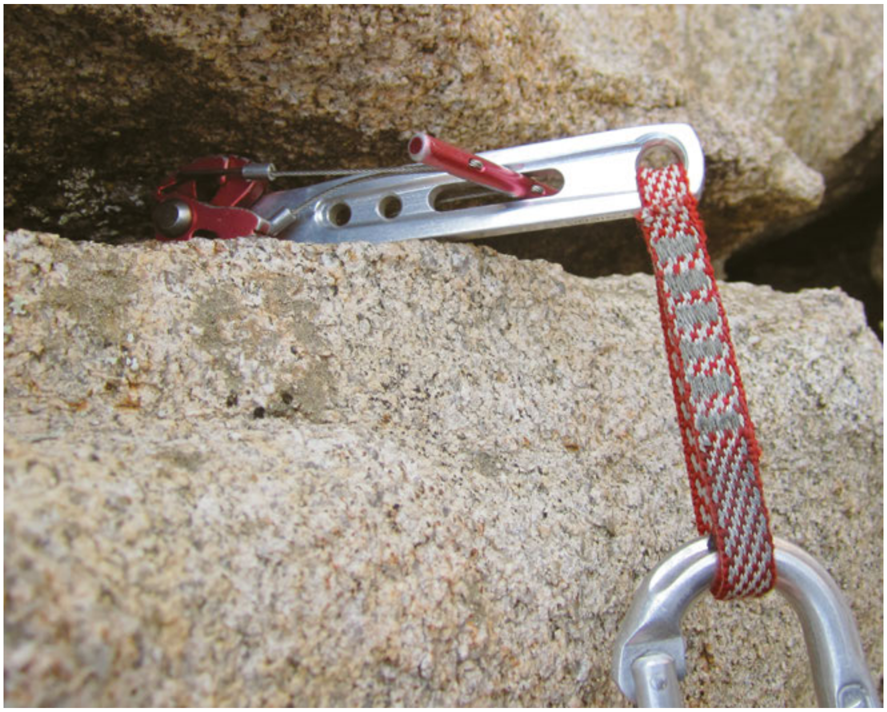

Placing a camming device with a rigid shaft in a horizontal crack with the shaft protruding can be dangerous because the shaft can break if loaded at a 90-degree angle during a leader fall.

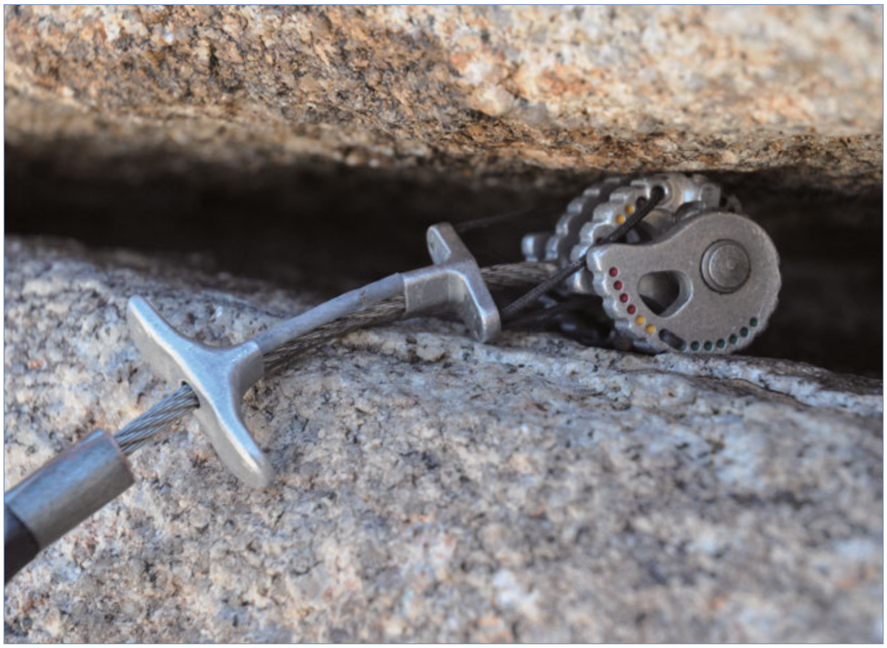

Camming units with flexible shafts are the best option for horizontal placement. Metolious recommends that in a horizontal crack, the outside cams should be placed on the bottom of the crack for maximum stability.

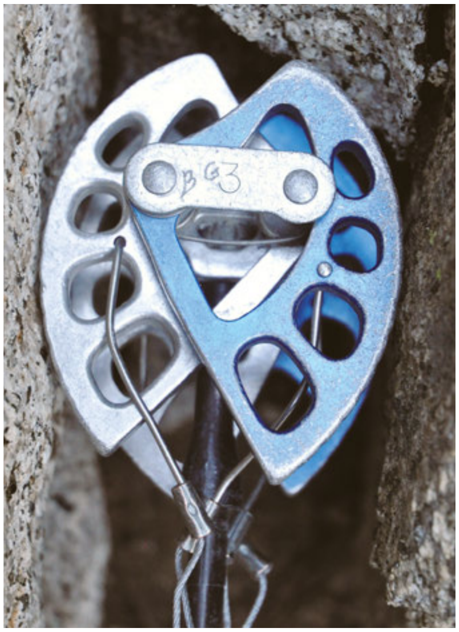

Good. This Camalot, placed in a solid horizontal slot, has all four cams tighter than 50 percent retraction with flush surface contact.

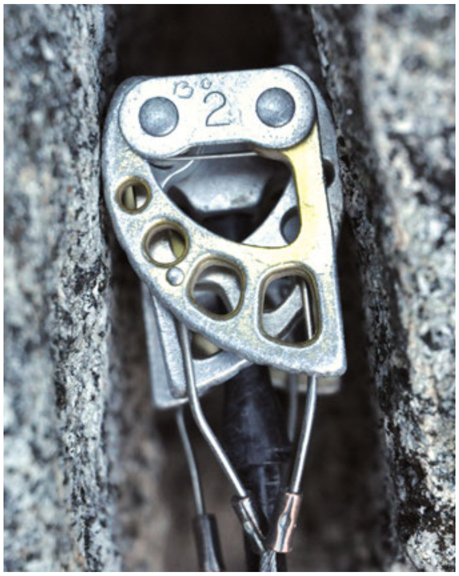

Borderline too tight. This Camalot is around 90 percent retracted. Any tighter and it may be very difficult to remove. There is also some loss of holding power in the last 10 percent (90 to 100 percent retracted) on a Camalot.

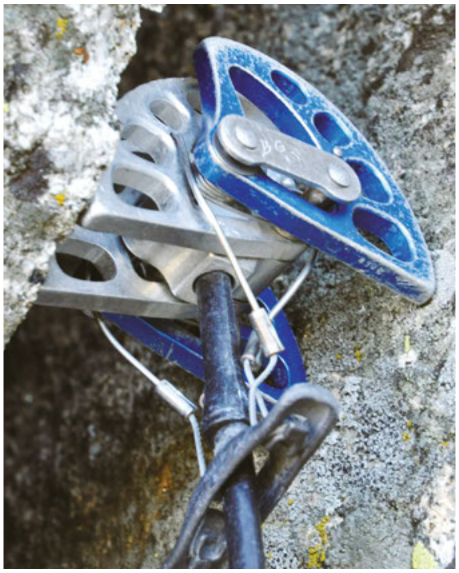

Very poor. The cams are barely retracted and nowhere near the recommended range. The piece can easily walk and might fail completely.

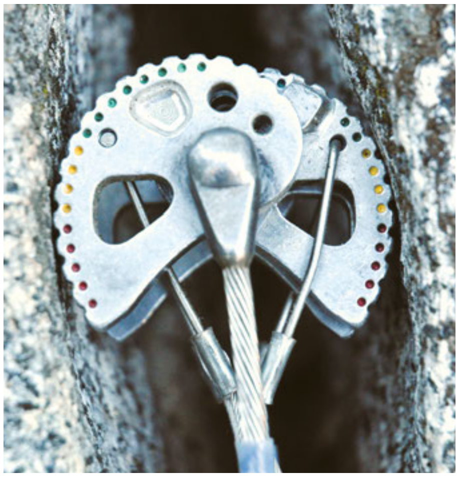

Borderline marginal. While this Metolius Power Cam is in a pocket that lends some stability to the placement, the range of retraction is between the yellow and red "range finder" dots; the red dots signify a marginal placement. The next larger unit would fit nicely.

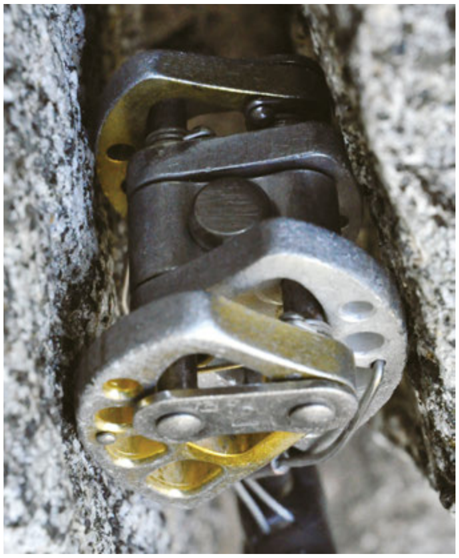

Fair. This Camalot is in a slightly flaring crack, with the inside cams retracted tighter than the outside ones, although each set of cams (inside and outside) are within a suitable range and all the cams have flush contact with the rock.

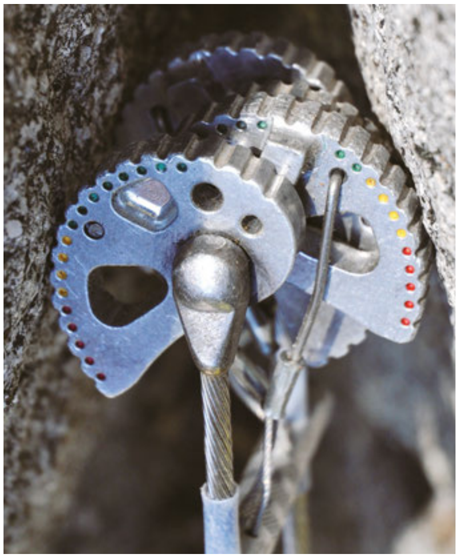

Bad. The crack is way too flared for this Metolius Power Cam, and the cam on the right side has very poor surface contact with the rock.

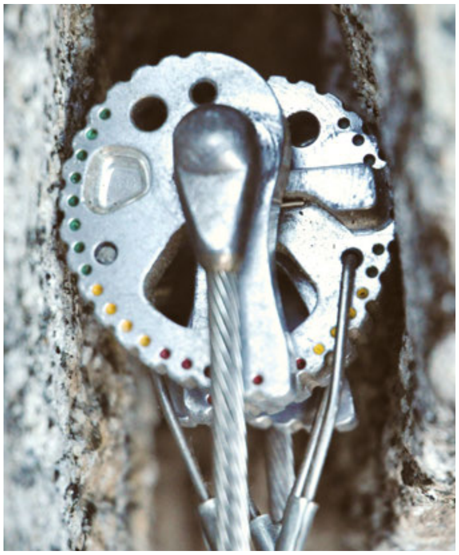

Good. This Metolius cam is in the tighter aspect of its range. Green means good to go.

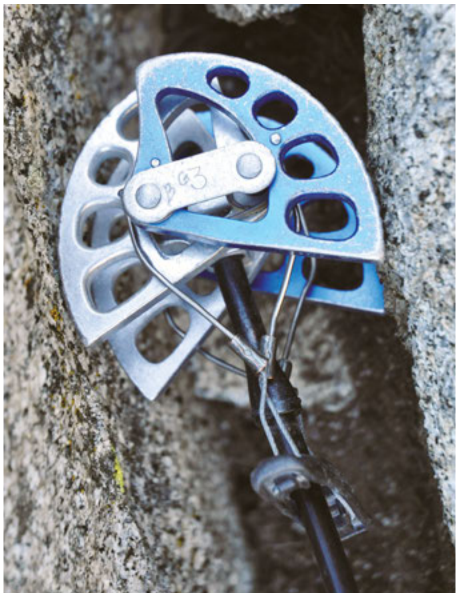

Bad. The cams are too open, rendering the placement unstable. Shoot for at least halfway tight on the cams.

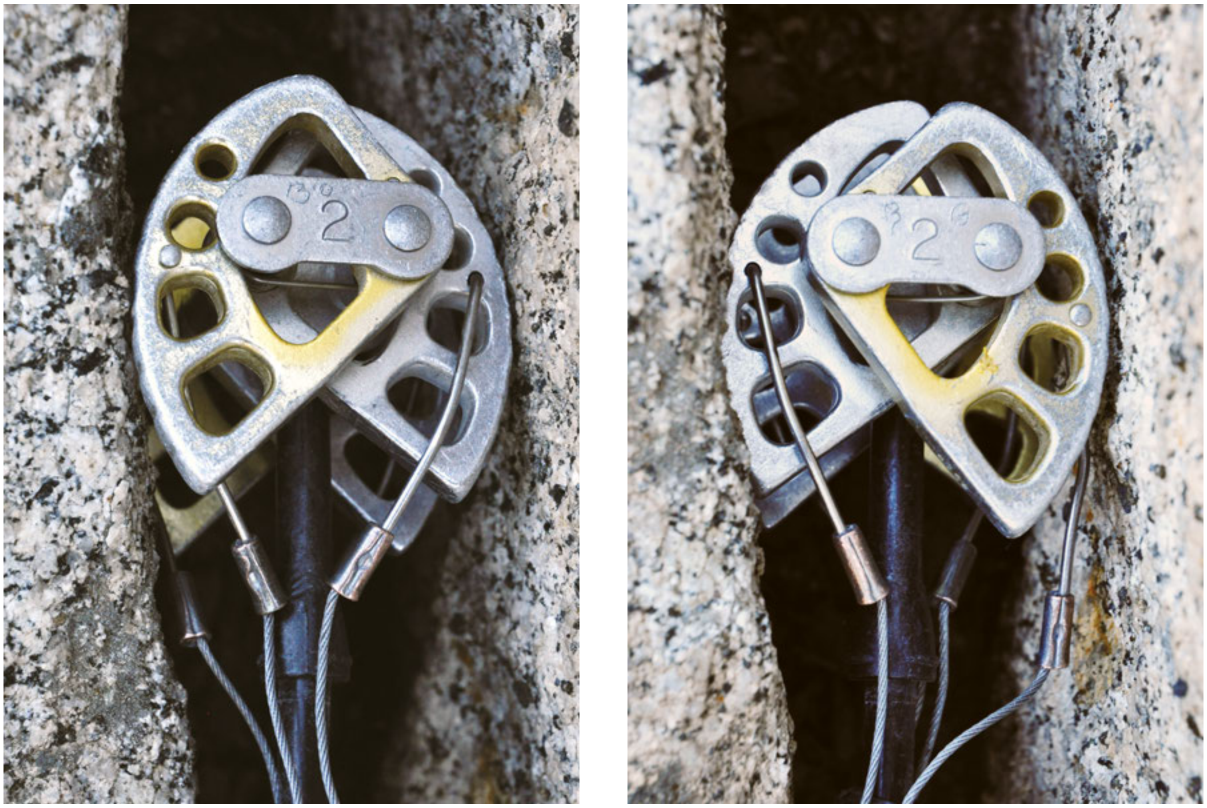

The same crack with two different placements. In the left-hand photo, the left outside cam has poor contact and is too close to the edge of the crack. By flipping the cam around (right photo), the gold cam now has flush surface contact with the rock. Since the inside and outside sets of cams are offset, flipping the cam one way or the other can often afford a better placement, particularly in shallow cracks in corners.

To illustrate what constitutes an acceptable range of retraction for the cams of a camming device, let's look at the Black Diamond Camalot in greater detail.

What is 50 to 90 percent retracted for a double-axle camming device like the Camalot? When you're looking at the Camalot without pulling on the trigger, it's at 0 percent retraction. Squeezing the trigger mechanism so that the cams are as tight as possible is 100 percent retracted. At 100 percent retracted, in a very tight placement, the Camalot will likely be very difficult to remove. In the last 10 percent of the tightest aspect of the range (90 to 100 percent retracted), the Camalot also loses some of its holding power, another reason not to go too tight on a placement. The starting point for a good placement is at 50 percent retraction, which is when the cams are pulled at least halfway tight. Looking at the base of the cams, 50 percent retraction is when the base of each cam is at a 45-degree angle relative to the vertical axis of the Camalot. If the cams are symmetrically retracted, they will be at a 90-degree angle relative to each other. A common mistake is to place a Camalot near the outer limit

The innovative Link Cam, by Omega Pacific, covers the range of four standard camming devices in a single unit.

of its range (0 to 50 percent retraction). This can prove to be a very unstable placement if the unit moves at all in the crack, which can easily happen if the Camalot is placed in a crack that widens above the cams and the piece is repeatedly weighted and unweighted. Again, the optimal Camalot placement is when the cams are at least halfway tight (50 percent retracted). From the beginning position, pull the trigger mechanism until the range on the cams is half the starting size, then go only smaller and tighter from there. Scrutinize your placement after the camming device has been placed in a crack to make sure the cams are in the acceptable range.

Metolius cams have a unique color coding that assists in their assessment. The company gives this advice: "Verify that you have chosen the best size by making sure that the green Range Finder dots are lined up where the cam lobes touch the walls of the placement. Yellow dot alignment is okay too, but you must exercise more caution with the placement, because the cam will be less stable, hence more prone to walking, and it will have less expansion range left to accommodate walking to a wider position. If the cam you choose aligns in the yellow zone, the next larger size will align perfectly in the green zone. Use that cam instead, if it's still on your rack. Never use a placement in the red zone unless it's the only placement available."

Study the literature that comes with any camming device and learn what the manufacturer recommends for the acceptable range of retraction and the various placement criteria. Most manufacturers also have informative PDF files on camming device guidelines that can be downloaded from the company's website.

To become proficient in the use of camming devices takes thought and practice. To develop confidence quickly, hire an AMGA certified guide or climbing instructor to critique placements. Metolius suggests to "practice placing cams in a safe venue,

Ground Practice

When I (BG) teach camming device placements, I first demonstrate the fundamentals, then let students make a variety of placements with a critique on each one. Working with an instructor allows them to learn from their mistakes before venturing out on their own. Better to learn in a "ground-school" setting than on their first toprope anchor that their "Friend" was no friend at all.

at ground level, before you trust your life to a cam placement. This process can teach you a lot, but written guidelines and practice are no substitute for qualified instruction. We strongly recommend that you learn to place cams under the supervision of a certified guide."

Fixed Anchors

Pitons

A piton is a metal spike that is hammered into a crack for an anchor. The blade of the piton is the part hammered into the crack, leaving the protruding eye into which a carabiner can be clipped. Piton anchors are something of a rarity these days, but occasionally fixed pitons (also called pins) can be found at the top of a crag. Follow these steps before using any fixed pin. First, assess the rock structure and look at the crack where the piton resides. Is it behind a block or flake, or is it in a straight-in crack with good structure? A good piton should be driven in all the way to the eye, and should not wiggle when clipping into it with a sling and pulling on it to test it. The piton itself should not be excessively corroded or cracked. (Look closely at the eye of the piton, as this is usually where the piton will be cracked.) To effectively

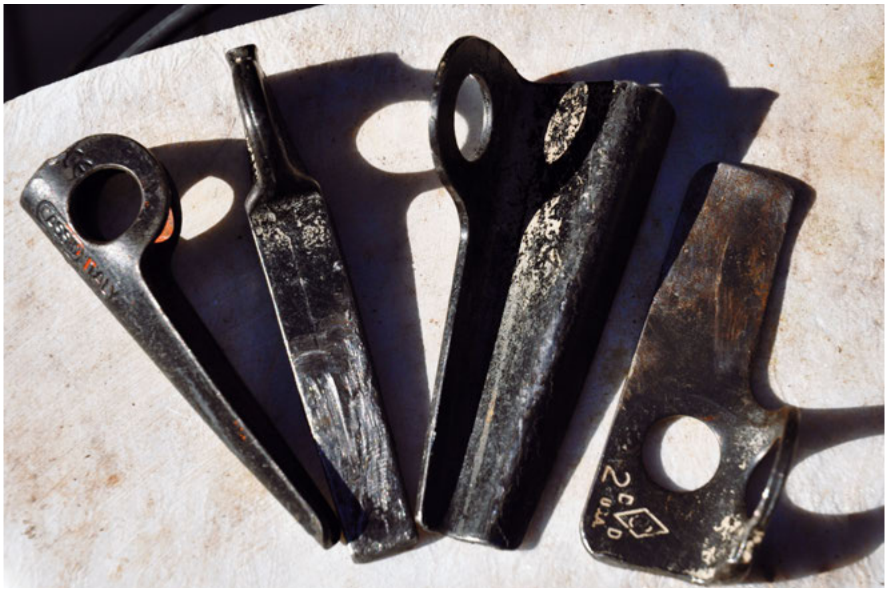

Pitons (left to right): Angle, horizontal, Leeper Z, knifeblade.

An angle piton, driven all the way to the eye—a good placement.

test a fixed pin, a hammer is needed. Give the piton a light tap—it should have a high-pitched ring to it, and the hammer should spring off the piton. Without a hammer, tap it with a carabiner or small rock. The best test is to clip a sling into it and give it a vigorous yank in the direction it will be loaded. Over time, pitons suffer from the vagaries of thermal expansion and contraction, particularly in winter, as water expands when it freezes, prying and loosening the piton. Often a piton can be easily plucked out by hand after only a few seasons. If utilizing fixed pitons as part of a toprope anchor system, always back them up, and use them with skepticism.

Bolts

The most common fixed anchor is a two-bolt anchor. Some knowledge of the history, characteristics, and specifications of bolts used for rock climbing will improve one's ability to assess the reliability of bolt anchors.

In the 1960s and 1970s, bolts were placed by hand drilling—an arduous process where a drill bit was inserted into a drill holder, then a hammer was used to pound on the holder to painstakingly drill into the rock. Once the hole was deep enough, a bolt, with a hanger attached, was hammered into the hole. The most common bolt during that era was the ubiquitous ¼-inch contraction bolt, called the Rawl Drive, manufactured by the Rawl Company and designed for the construction industry for anchoring in masonry or concrete. A contraction bolt has a split shaft that is wider than the diameter of the hole. When pounded into the hole, the two bowed shaft pieces are forced to straighten slightly, contracting under tension in the hole. This works fine for hard granite, but in soft rock, like sandstone, the split shaft doesn't really contract all that much, and there is little tension to keep it in the hole, resulting in very weak pullout strength (i.e., pulling straight out on the bolt).

Another problem with ¼-inch bolts is that they came in various lengths, some as short as ¾ inch long, and once placed in the rock, there was no way for future climbers to determine the length of the bolt merely by inspection.

There are two basic styles of ¼-inch Rawl Drive bolts. The buttonhead design has a mushroom-like head and is pounded into the hole with the hanger preattached. The threaded Rawl Drive has threads with a nut on the end to hold the hanger in place, a weaker configuration since the threads can weaken the shear strength of the shaft if the hanger is at the level of the threads. But more significantly, the threaded design has a serious flaw: Pulling straight out on the bolt hanger will only be as strong as the holding power of the nut on the threads, a dangerous problem if the nut is at the very end of the threads.

The shear strength on a brand-new ¼-inch Rawl Drive bolt is roughly 2,000 pounds, but the problem with contraction bolts is not shear strength but pullout strength, which varies drastically depending on the quality and hardness of the rock. In very soft sandstone the pullout strength of a ¼-inch contraction bolt is extremely low, rendering the bolt unsafe.

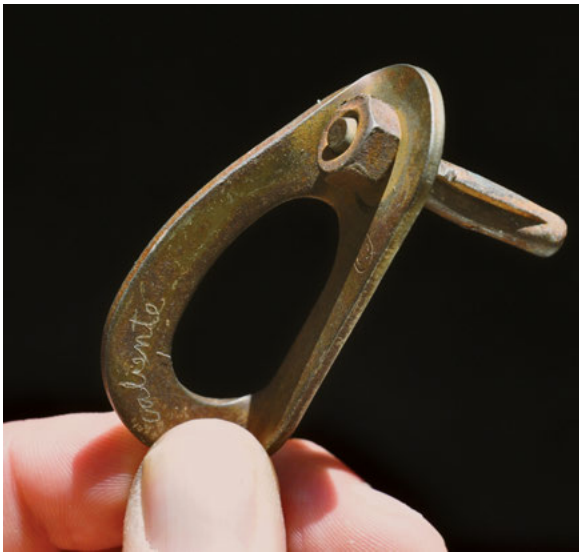

The infamous ¼-inch threaded Rawl Drive contraction bolt, complete with the SMC "death hanger." This ticking time bomb was removed and replaced from a route on Suicide Rock, California.

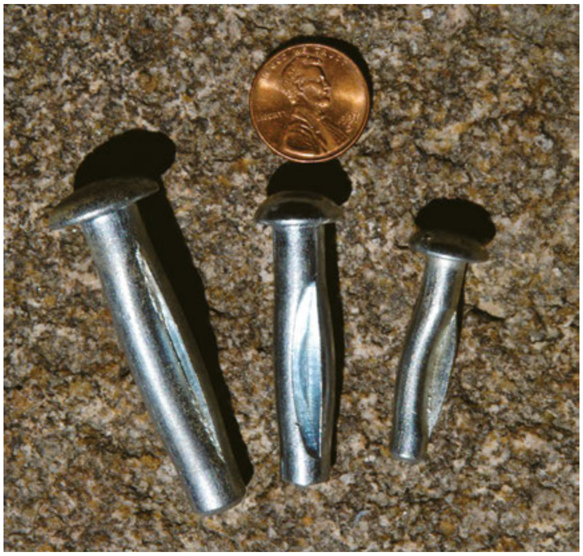

Buttonhead Rawl Drive contraction bolts (left to right): ⅜-, 5⁄16-, and ¼-inch sizes.

The buttonhead Rawl Drive bolts were also sold in 5⁄16-inch diameters, these being far more reliable as long as they were placed in good, hard, fine-grained granite. The 5⁄16-inch buttonhead, for example, has a shear and pullout strength in excess of 4,000 pounds, and for many years was the bolt of choice for first ascensionists who were hand drilling bolts. The 5⁄16-inch buttonhead Rawl Drive was discontinued, but the ⅜-inch buttonhead is still on the market today, with a shear strength of 7,000 pounds and a pullout strength of over 4,000 pounds in the best granite.

Probably the most disconcerting problem associated with bolts from the ¼-inch era is not the bolts themselves but the hangers. During that time, hangers made for rock climbing were manufactured primarily by the SMC company. Thankfully, the hangers are easily identified, as the "SMC" brand is stamped on them. There were two series of hangers: one good, and one very bad. The "bad" hangers

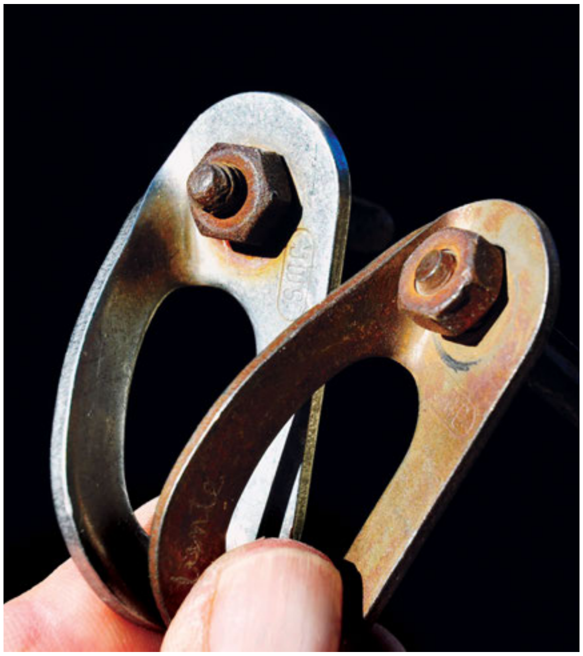

¼-inch threaded Rawl Drive bolts with "good" (left) versus "bad" (right) SMC hangers.

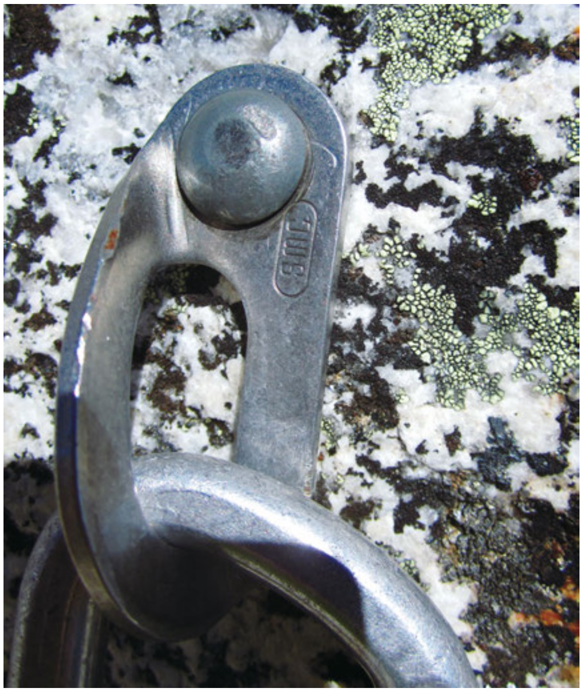

5⁄16-inch buttonhead Rawl contraction bolt with "good" SMC hanger. In a good placement in solid granite, these bolts are rated at over 4,000 pounds shear strength.

were nicknamed the SMC "death hangers," since some of them failed under body weight after only a few seasons of exposure to the elements. These hangers are identifiable by a distinctive corrosive discoloration—a yellowish or bronze tint, whereas the "good" SMC hangers, made from stainless steel, show no signs of corrosion or rust and still appear silvery bright, even after twenty-five years. Another noticeable difference is in the thickness of the hangers—the "bad" hangers roughly the thickness of a dime, and the "good" ones the thickness of a quarter.

Another dangerous relic from the 1970s is the Leeper hanger. Over 9,000 of these hangers were manufactured by Ed Leeper of Colorado, and subsequently recalled because of stress corrosion

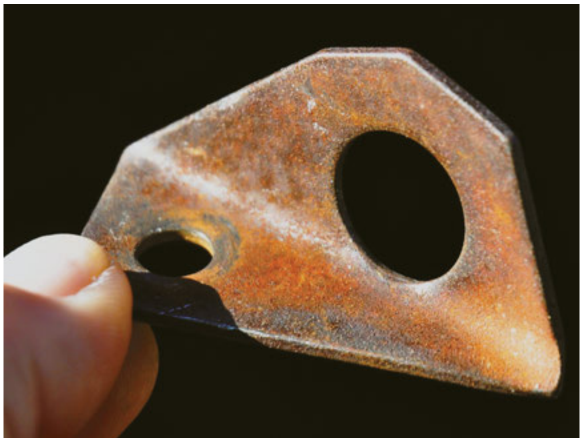

The recalled Leeper hanger can easily be identified by its unique shape and rusty condition.

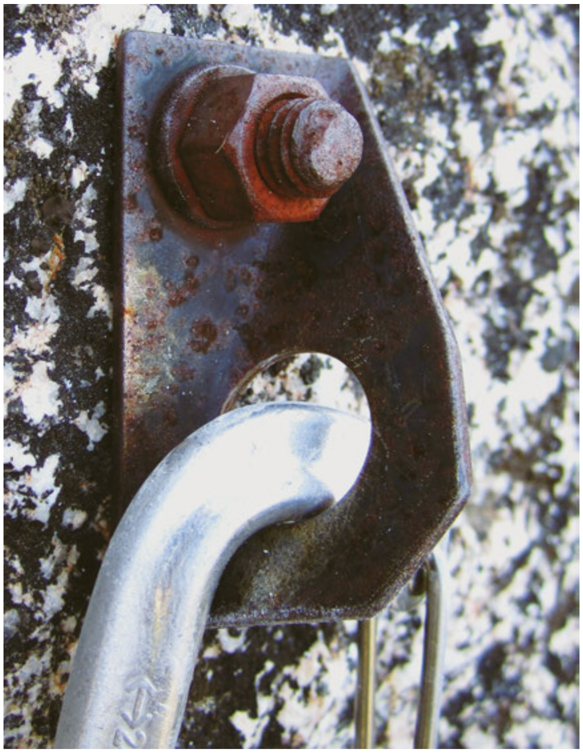

Bad corrosion on a 3⁄8-inch diameter threaded Rawl Drive bolt with a badly corroded Leeper hanger to match.

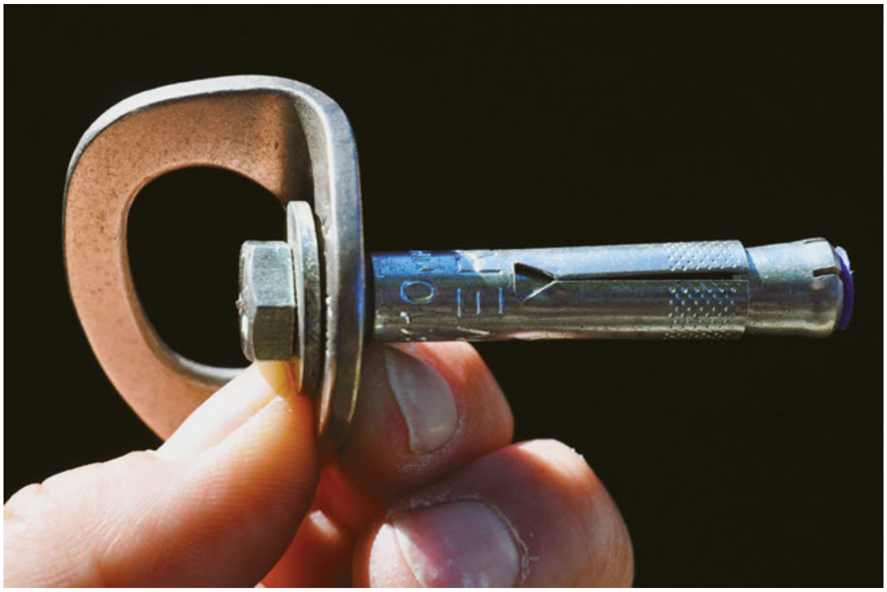

The ⅜-inch diameter Powers Power Bolt expansion bolt with a stainless steel hanger has become the minimum standard for climbing anchor bolts.

problems with the metal, which rusted badly since it was not made of stainless steel. These hangers are easily identifiable due to their strange geometric shape and their rusty condition.

In the 1980s sport climbing was ushered into the United States, and climbers began to place bolts on rappel using cordless rotary hammer power drills. Since these bolts would now have to absorb numerous falls, climbers began to look for the strongest bolts available, and the standard became ⅜-inch diameter for good, solid rock (like granite) and ½-inch diameter for softer rock (like sandstone)—standards that are still prevalent today.

Although there are numerous types of bolts used in rock climbing today, the gold standard has long been the "5-piece Rawl" expansion bolt (now sold as the Powers Power Bolt). This expansion bolt has a shaft with a hex head on one end and threads on the other end (the end that goes in the hole), with a cone-shaped piece screwed onto the threads. The shaft has a two-part split sleeve, and as the hex head is tightened, the cone climbs up the threads and under the sleeves, which presses the sleeves outward, "expanding" the bolt in the hole. The more you tighten it, the wider the sleeve gets. The performance and strength of the bolt relies, to a great extent, on two things: the tolerance (diameter) of the hole, and the strength of the rock itself. In good rock the ⅜-inch Power Bolt is rated at over 7,000 pounds shear strength, with a pullout strength of roughly 5,000 pounds.

Since these bolts are really designed for the construction business, the Powers Fastener company lists strength ratings based on the density of the concrete they are placed in. Concrete is given a psi (pounds per square inch) rating. For example, "2,000 psi concrete" means that it would take a weight of 2,000 pounds to crush a square inch of concrete. Hard, dense granite is analogous to 6,000 psi concrete, and soft sandstone is more like 1,000 psi concrete.

Once a bolt has been installed, it's impossible to see what's going on beneath the surface (like with the length of the bolt), and all that can be seen is the head of the bolt, again making identification of the type of bolt more difficult.

For more in-depth information, peruse "mechanical anchors" on the Powers company

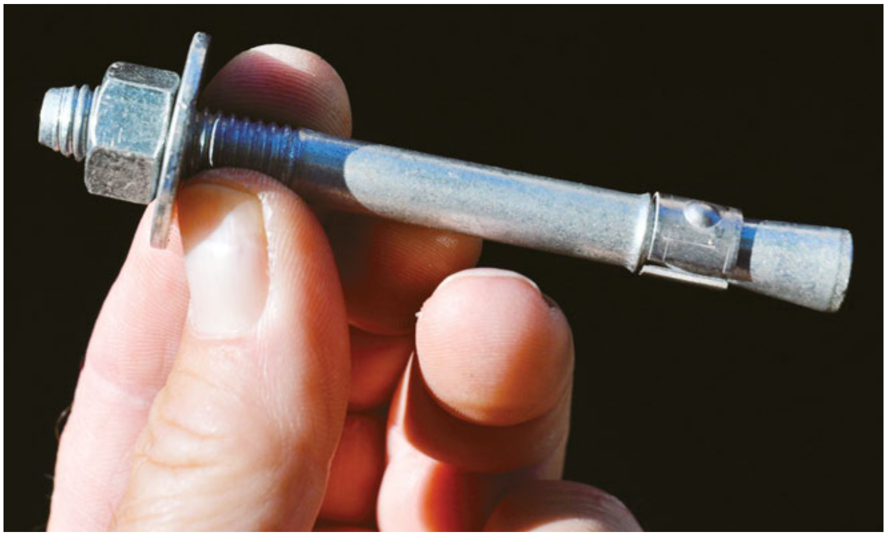

⅜-inch threaded expansion bolt.

website (www.powers.com); there is an excellent tutorial on the various types of bolts and how strong they are in differing rock types.

Even if lacking expertise in mechanical engineering or in identifying bolt design and type, it is important to know what to watch for when inspecting a bolt anchor. An obvious red flag is rust. SMC "death hangers," Leeper hangers, homemade aluminum hangers, and any bolt or hanger with obvious signs of corrosion should not be trusted. Look closely and identify the diameter of the bolt. A ⅜-inch diameter bolt has become the minimum

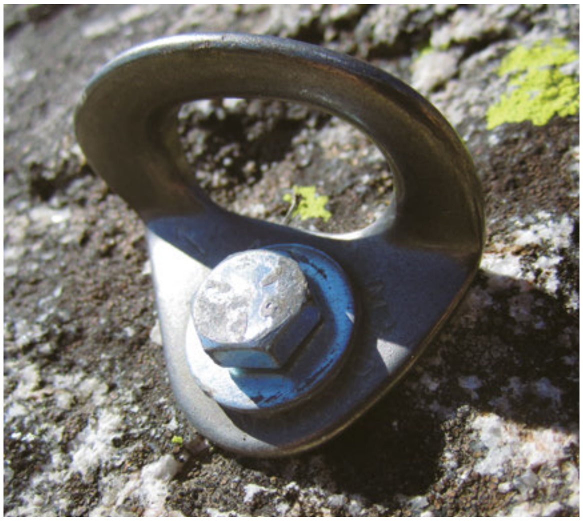

⅜-inch Powers expansion bolt with stainless steel Mad Rock hanger.

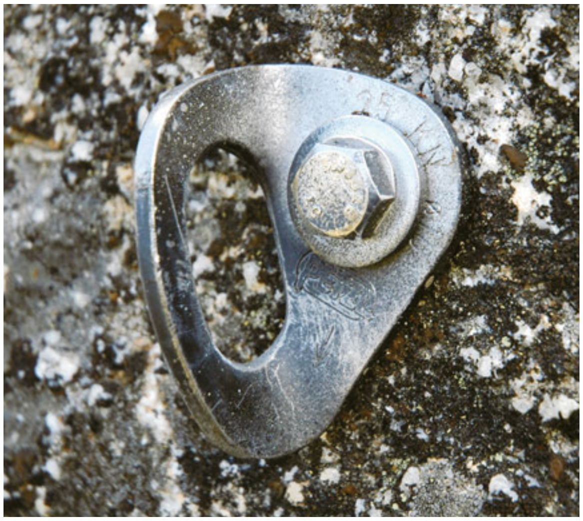

⅜-inch stainless steel Powers bolt with stainless steel Petzl hanger, painted to match the rock color.

Beware the ¼-inch Bolts

As someone who has replaced many bolts over the years, I (BG) can tell you that any ¼-inch bolt should be considered suspect, particularly in less than perfect rock. I've plucked out many ¼-inch contraction bolts that came out with about the same resistance as a nail being pulled out of plywood. To replace a ¼-inch bolt, the best method is to pry it out of its hole, then re-drill the same hole to a ½-inch diameter and install a ½-inch diameter stainless steel Powers Power Bolt (10,000 lbs. shear strength) with a stainless steel hanger. I like to paint the hanger (before I install it) the same color as the rock so that the bolt is visually unobtrusive. It's a good feeling to replace a ticking time bomb with a solid anchor that will last a lifetime.

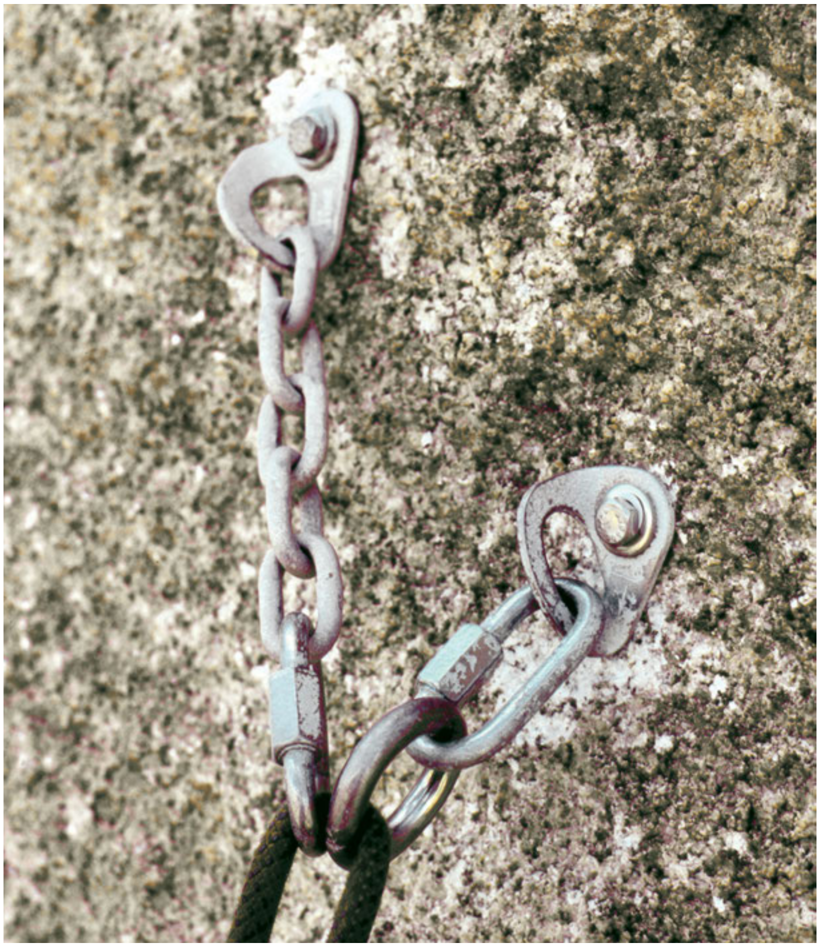

A well-engineered rappel anchor. Both bolts are ⅜-inch stainless Powers bolts with stainless steel Petzl hangers, along with a stainless steel chain, quick link, and ring. Everything was painted before installation to match the color of the rock.

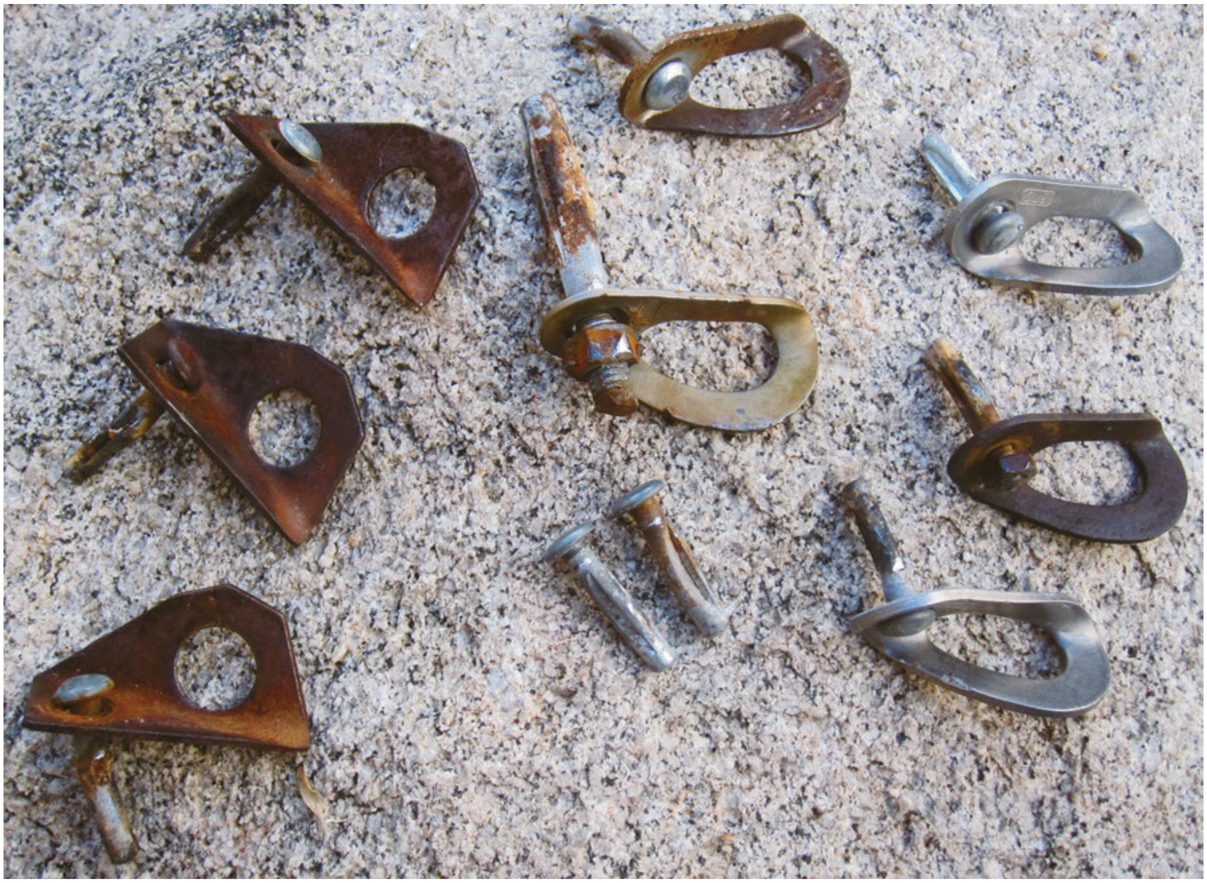

All these old bolts at Joshua Tree were replaced with brand-new stainless steel hardware, courtesy of the American Safe Climbing Association (ASCA).

standard, along with a stainless steel hanger. A bolt with threads and a nut holding the hanger in place is generally not as strong as the hex head types.

The rock should not show cracks emanating from the bolt placement—a more common problem with contraction bolts than expansion bolts.

In a good placement, the hanger should be flush against the rock and should not budge or be deformed in any way. A "spinner" is a bolt that protrudes enough so that the hanger can be easily spun around 360 degrees. This generally means that when the bolt was installed, the hole was not drilled deeply enough, and the bolt contacted the bottom of the hole before the hanger could be drawn flush against the rock.

If the bolt wiggles slightly when pulled or the hanger is loose, and the bolt has a hex head or a nut on threads, tightening the bolt with a wrench may help, but most likely the bolt has a problem that can't be fixed. If, while trying to tighten it, no increasing resistance is felt, and it won't tighten any further, then the bolt has serious problems—usually this means the tolerance (diameter) of the hole is too big for the bolt, or the rock is too soft.

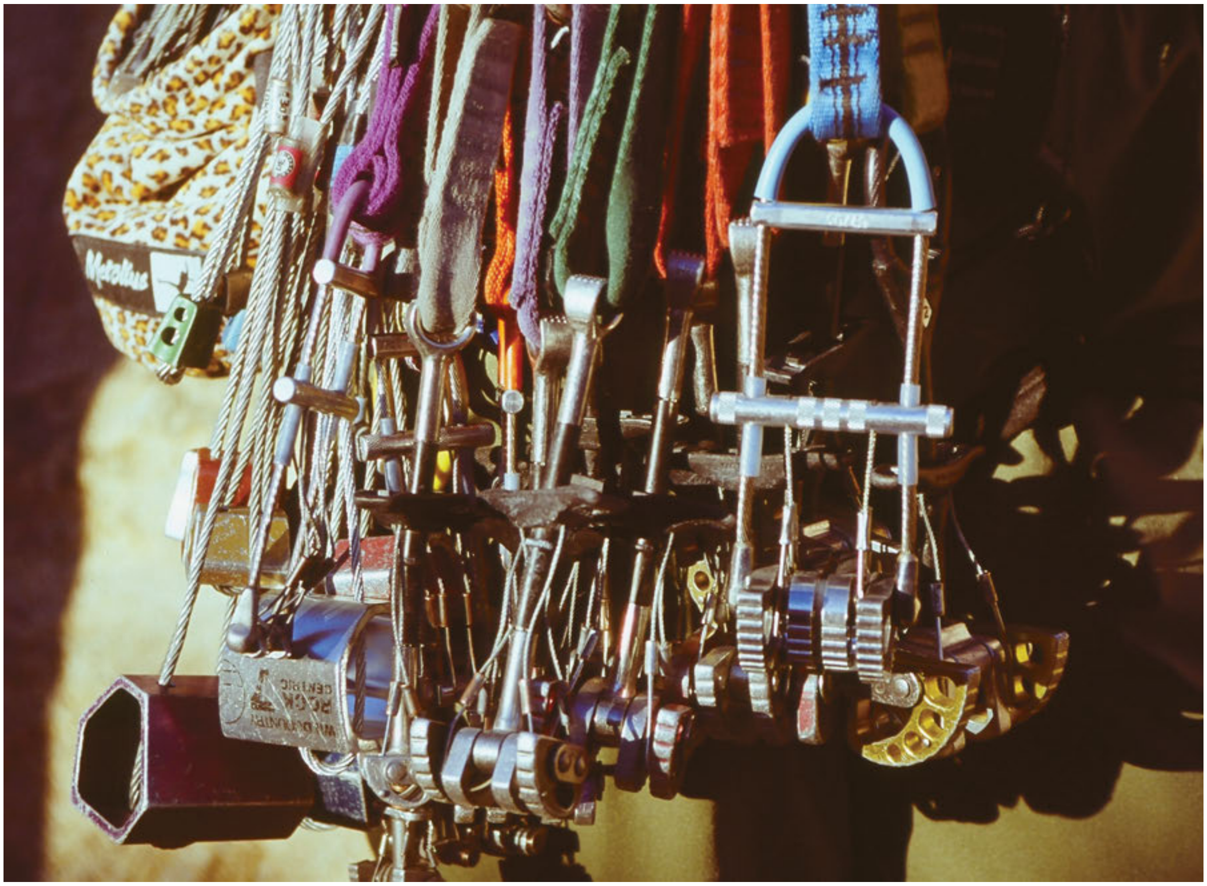

This rack has a good assortment of both nuts and camming devices.

The American Safe Climbing Association (ASCA) has been very active in donating the necessary (and expensive) hardware to climbers who take on the task of upgrading unsafe bolt anchors with modern, stainless steel bolts and hangers. To support their efforts and donate to the ASCA, visit www.safeclimbing.org.

CHAPTER 8

Anchor Systems

If the hundreds of pages of how-to literature, magazine articles, blogs, lengthy message board posts, and hours of instruction from professionals and climbing partners teach us nothing else, it's that anchors are something climbers pay attention to. And yet, much of what is said and has been said is facile. Most climbing anchors involve the simple combination of obvious components, using strong materials to make an unquestionably effective master point. As climbing instructors, and as climbers too, there is a tendency to obsess over the outliers, conflate simple problems and complex solutions, and worry ourselves. This section will attempt to differentiate between the simple tools used to solve simple problems that instructors are likely to encounter, and the more complex tools used to solve complicated problems that instructors will rarely encounter.

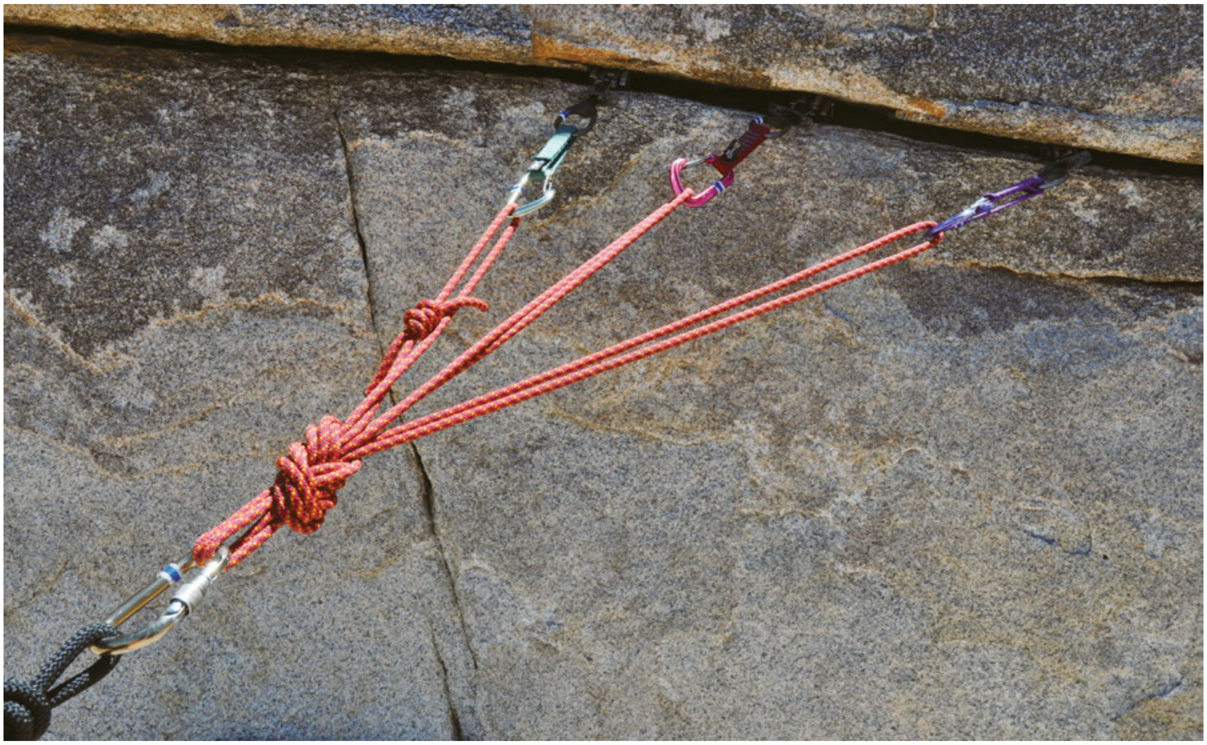

A simple three-point cordelette anchor.

For every anchor, simple or complex, the need for an unquestionably reliable, strong, and secure attachment point is the goal. The responsibility for managing risk of harm to others should inspire an anchor builder to create an anchor that can do the work of the climbing system, can easily sustain any potential loads that emergency scenarios may create, and can still give the entire climbing team a margin of error. So how strong is that? How strong would something have to be to eliminate all doubt?

Let us begin by considering the base-managed toprope, the system where the toprope has been redirected through the anchor and the climber and the belayer are on the ground. If the climber weighs 100 pounds, the belayer must match that weight in order to arrest the climber; therefore, we'll call the belayer a counterweight. At any point in a taut climbing system, the anchor must be able to hold the climber (100 lbs.) and the counterweight (100 lbs.); that's 200 pounds. But the climbing system is not always perfectly taut. The latent elongation of the climbing rope means that the climber is always a mass in motion, a mass on a sort of bungee cord when he or she falls. With a perfect belay, let's say that the climber's dynamic weight is more like 200 pounds. Now the counterweight is 200 pounds too, and the anchor must hold 400 pounds. Now let's say that the climber moves faster than the belayer can belay, or the belayer is a bit inattentive, and the falling climber now generates an even bigger dynamic load. Four hundred pounds is not inconceivable, and matched by a 400-pound counterweight, the anchor needs to hold 800 pounds. Now, if we could conceivably create an 800-pound load, how strong would the anchor need to be to create an unquestionably reliable, strong, secure attachment point? Now consider that climbing instructors are often taking all kinds of people climbing, people who weigh more than 100 pounds, for example. How strong would the anchor need to be? More than 800 pounds, surely. More than 2,000 pounds? More than 3,000? There is no definite answer. "Unquestionably reliable, strong, and secure" might mean different things to all of us. That is both the challenge and the beauty of anchor building.

The goal of this chapter is to demonstrate skills that will allow an instructor to build an anchor that not only solves this problem effectively, but provides a reasonable margin of safety. Since all climbing gear, attachment tools, carabiners, protection, and ropes are overly strong, we should be able to build anchors that solve the problem unquestionably well.

Fundamental Anchor Anatomy and Principles

All anchors have the same fundamental anatomy. They have components, attachments, and a master point. Most anchors also have a shelf. Together these create the acronym, CAMS.

C = Components are the features or tools that ground an anchor to the landscape. They are often organized into natural (trees, boulders, terrain features) and artificial (bolts, pitons, removable protection) categories.

A = Attachments are the materials that combine the relative strength of each component and consolidate their strength in a focal point.

M = Master point is the point of the anchor where all critical attachments occur.

S = Shelf. Some anchors also have a shelf, an auxiliary attachment point with all the same load-bearing and security qualities as the master point. It should only be used in an auxiliary fashion.

All anchors should strive to adhere to the following fundamental principles:

1. Eliminate or limit extension: Given the failure of any individual part of the anchor, the attachment systems will not extend enough to severely shock load the remaining components.
2. Redundancy: Given the failure of any feature of the anchor system (with the exception of

a monolithic component), there will be other features to reinforce or back up that initial failure.

3. Intelligent load distribution: Given any load the anchor may sustain during its use, the attachment system should distribute load to the components in a manner that optimizes the relative strength and security of each component.
4. Strength: All attachments, components, knots, rock, or natural features must have more than adequate strength to sustain any load that could be applied to it during a climbing activity.
5. Simplicity: Given the anchoring tools available, the available components, and the prevailing conditions, an anchor should be efficiently constructed and deconstructed; while accomplishing all the other principles of anchoring, it should make efficient use of time and equipment.

Notes on Redundancy

By itself, redundancy does not guarantee safety—rappelling off a bunch of old weathered slings or a couple of old rusty chain links, for example—but redundancy in anchor systems is something to always strive for as a baseline and a starting point for evaluation. In rock climbing, a nonredundant anchor is considered marginal by professional standards. Many canyoneering enthusiasts argue that redundancy at the anchor system is superfluous, as there will be a point in the system where redundancy can no longer be maintained—rappelling on only one rope, with one rappel device, on one locking carabiner. This is absolutely true, so it's important to remember why redundancy is important to climbers. Unlike the canyoneering participant on rappel, or the military unit launching out of a helicopter, climbers are intentionally challenging themselves by climbing unknown difficulties. There is nothing unpredictable about a rappel; the line will be loaded consistently and in the same direction. But a rock climber's movements are erratic and unpredictable, with wide variation from one climber to the next, especially on a challenging climb. Systemic redundancy is a way to create a satisfactory margin of error to account for that unpredictability. Sometimes instructors might intentionally modify the concept of redundancy if they can substitute supervision to provide the security that redundancy might otherwise offer. The rope direct belay is a good example. A thoughtful instructor knows that the principle of redundancy is not simply being ignored, but that redundancy is not necessarily the goal. Security is the goal; redundancy is just one of the most common ways to create that security.

The Rope Direct Belay

If the belay anchor is initially built well back from the edge, and you want to belay from the edge to maintain a visual on the climber who's climbing up from below, you can use the climbing rope and a rope direct belay technique, which is essentially belaying off an extended master point. Attach your climbing rope to the master point on the anchor with a clove hitch on a locking carabiner, so you can adjust the length and position yourself just where you'd like at a stance near the edge where you can see the climber. Off the back side of the clove hitch, leave a little slack, then tie a figure eight loop and clip it into a separate locking carabiner to the master point. On this strand of rope, tie a figure eight loop and clip in your assisted braking device (ABD), at an ergonomic position slightly above you (toward the anchor) but not so far away that you can't reach the device and manipulate the handle if need be. This method is referred to as "rope direct" because you are essentially belaying directly off the anchor, albeit extended whatever distance is required to position you at the edge. Using a device like a Grigri, you'll have the benefit of a quick and

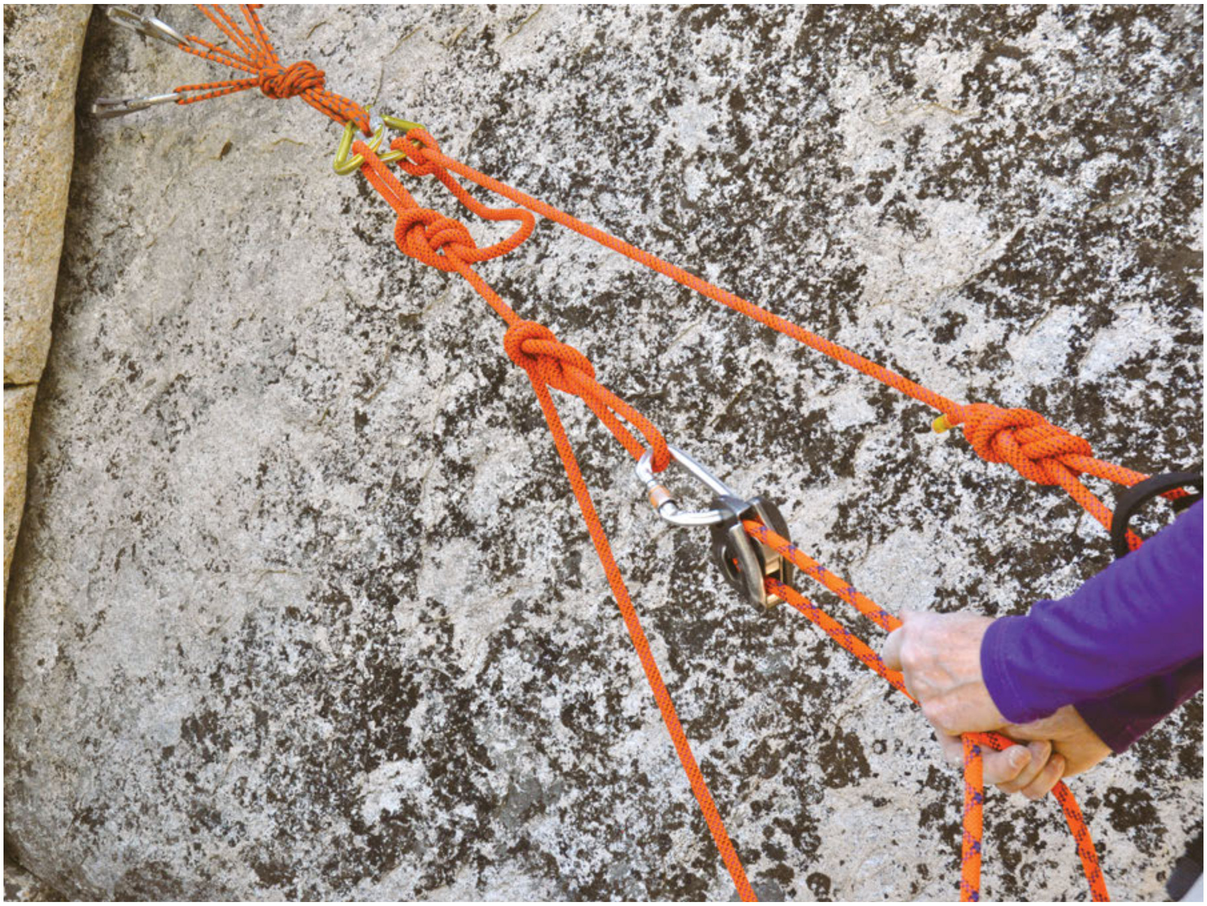

The rope direct belay rigging. This rigging technique allows you to position yourself away from the anchor and still use the direct belay technique, albeit on an extended master point. The belayer is attached to the master point with a clove hitch to a locking carabiner. Off the back side of the clove hitch, the rope is clipped back to a separate carabiner with a figure eight loop. Off this strand, the Grigri is clipped to another figure eight loop. Here the distance is fairly close to the anchor, but this rigging technique is most useful for greater distances between the anchor and your desired belay stance, limited only by the length of available rope. The big advantage is visual contact with your climber in situations where the anchor is some distance back from the edge.

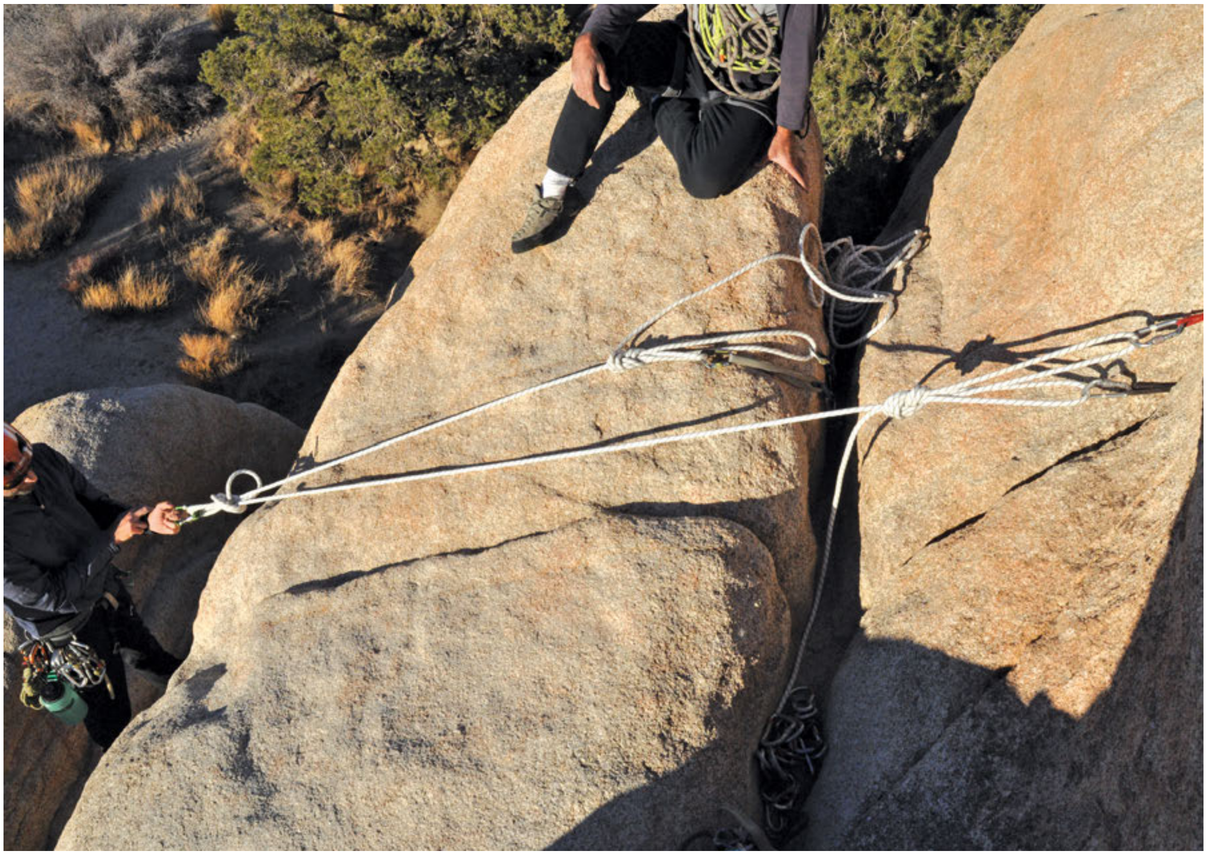

Students practice rigging a uniform load anchor system using double loop figure eights. The master point is tied with a BHK.

easy conversion to a raising system if required, and it's easy to lower someone on this setup by redirecting the brake strand on the Grigri.

Notes on Load Distribution

Most toproping situations involve an anchor that is uniformly loaded in a single direction. This is true of top-managed and base-managed climbs. Most anchor systems involve components that have comparable strength: a pair of bolts, a trio of cams in a similar size range, a combination of strong natural and removable components. In these cases, load distribution is fairly simple. A cordelette and 48-inch slings allow an anchor builder to pull tension on the components and tie a strong overhand or figure eight knot to ensure that the load is theoretically distributed to all the components. But in this configuration these anchors have been proven to have misleading load distributions. They do not in fact equalize the total load between the components. John Long argued this point compellingly in *Climbing Anchors.* John maintained that even if one could achieve equalization, the slightest shift in the direction of load would

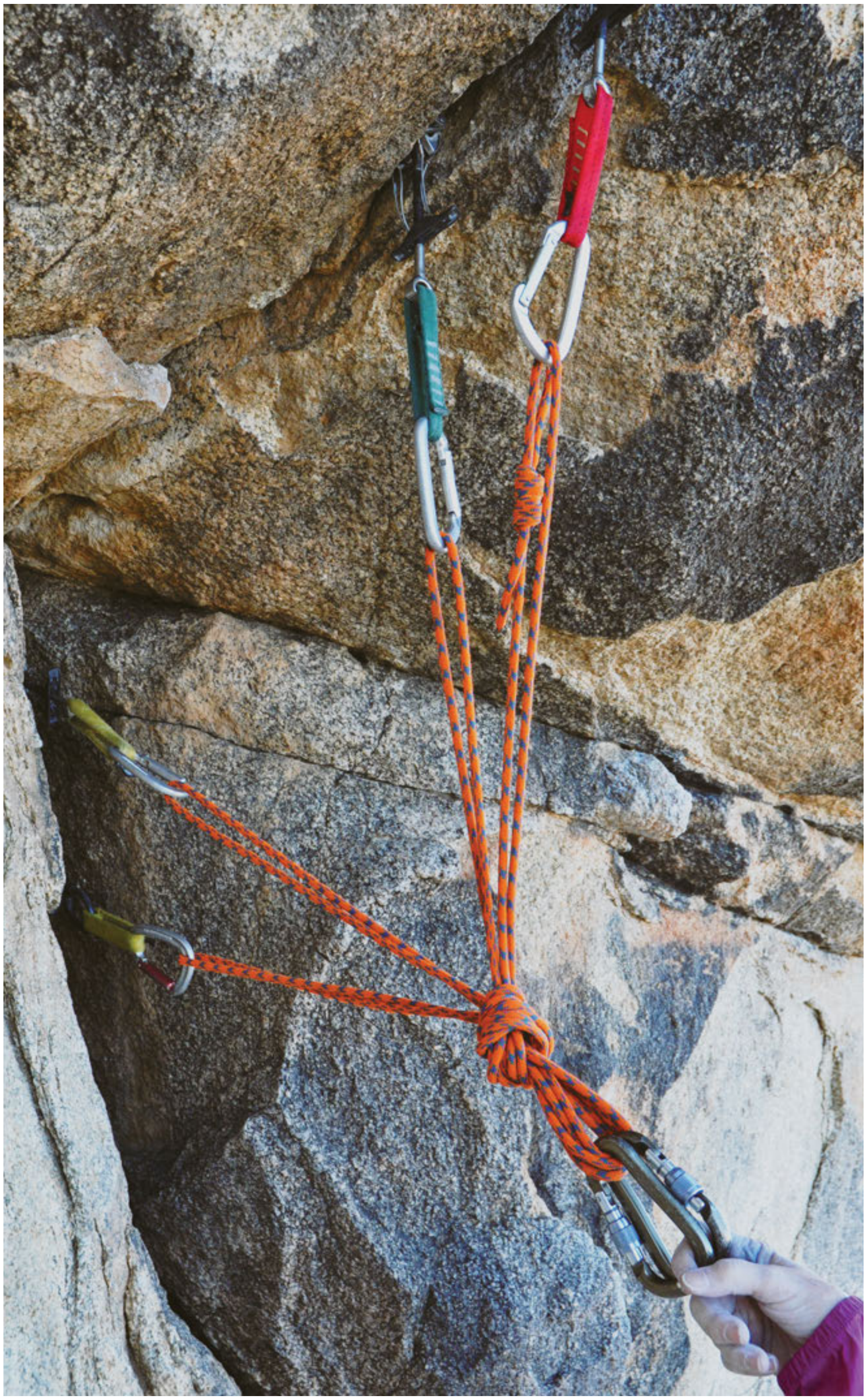

Four-piece uniform load anchor pre-equalized with a cordelette in the anticipated direction of load. If the load shifts even slightly, all of the load will come to bear on a single piece. If that piece were to fail, the vector would shift to the next piece to load, albeit with minimal extension from the cordelette. The cordelette system is essentially a series of backups rigged with minimal extension.

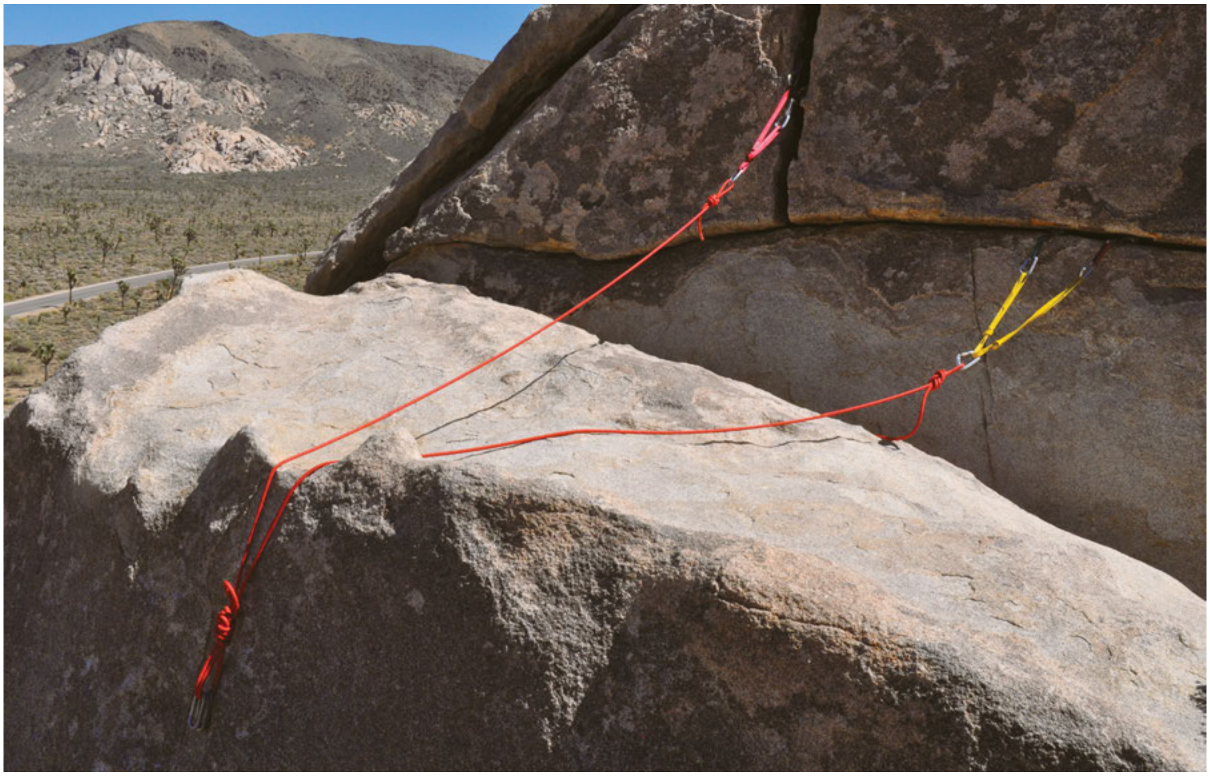

This toprope anchor, rigged using the Joshua Tree System, is an example of a variable load (aka self-equalizing) anchor system.

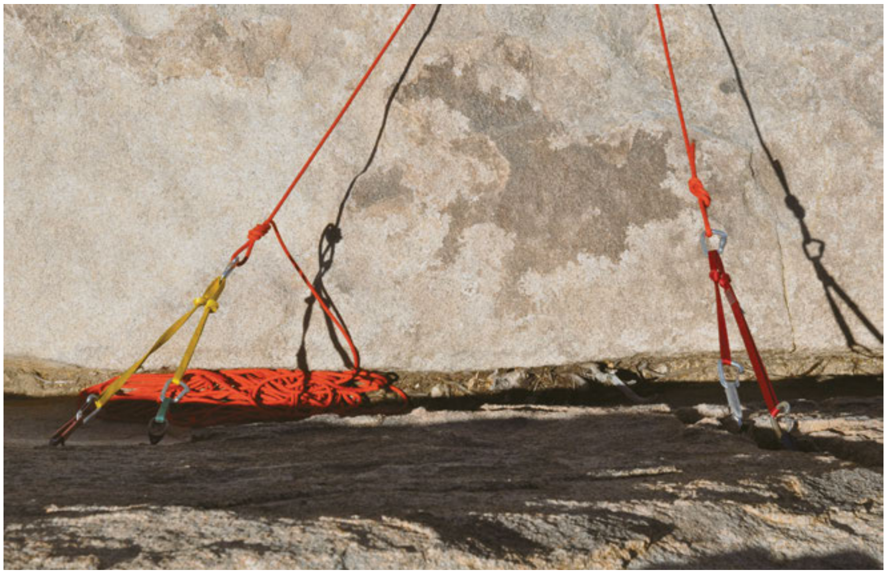

Each leg of the V is connected to two camming devices rigged with magic Xs and load-limiting knots on 48-inch nylon slings. If the direction of load shifts slightly, the magic Xs will adjust accordingly to distribute the load on each leg of the V.

distort that equalization. He further noted that even if the load was 100 percent uniform and equalization was 100 percent perfect, he would still suggest to the anchor builder that not all the components are of equal merit, deserving an equal share of the load. When necessary, a more effective anchor builder ignores the idea of equalization and intentionally manipulates the distribution of the total load. In the following examples we will describe simple, timely systems that are most appropriate when complex load distributions are not necessary, as well as more complex systems used to intelligently distribute load to components when the load varies in direction and the components are not equally strong.

The Cordelette: A simple solution to a common anchoring problem

The cordelette system is an attachment system where the anchor builder ties off the cordelette in the anticipated direction of load. If the direction of load shifts slightly in any direction, all the load goes onto one placement (albeit with minimal extension). For toprope anchor systems, in most cases it is simple to determine the direction the anchor system will be loaded in, so a simple solution like the cordelette makes sense most of the time. Using the cordelette in this manner, the anchor builder essentially creates a system of backups. If one piece fails, the load transfers instantly to the remaining pieces with minimal shock loading, since the rigging limits extension.

The cordelette can be used to attach two, three, four, or even five placements to a master point, so it is versatile as well as simple. For two placements, like a two-bolt anchor, start by doubling the cordelette, then clip the doubled strand into both bolts with a pair of carabiners. Pull down between the bolts in a manner that targets the anticipated load, gather all the strands together, and tie an overhand or figure eight on a bight. There should be four strands of cord at the master point.

To rig three or four placements, clip the cordelette into all the placements, pull down between the

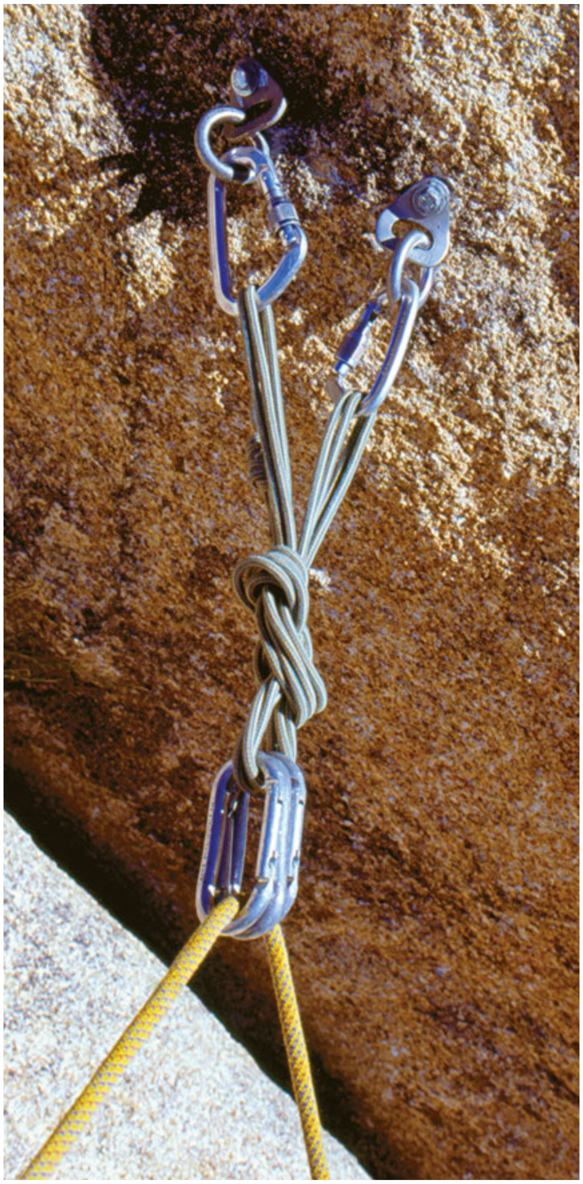

Simple two-bolt anchor rigged with a Tech cord (5,000 pounds tensile strength) cordelette. The cordelette is doubled to start with, producing four strands at the master point loop, and the climbing rope is clipped into three oval carabiners opposed and reversed.

pieces in a manner that targets the anticipated load, gather all the loops together, and tie an overhand or figure eight on a bight. Many instructors like to clip a carabiner into all the gathered loops and pull in the anticipated loading direction, then tie a figure eight knot with the carabiner attached to help even out all the strands.

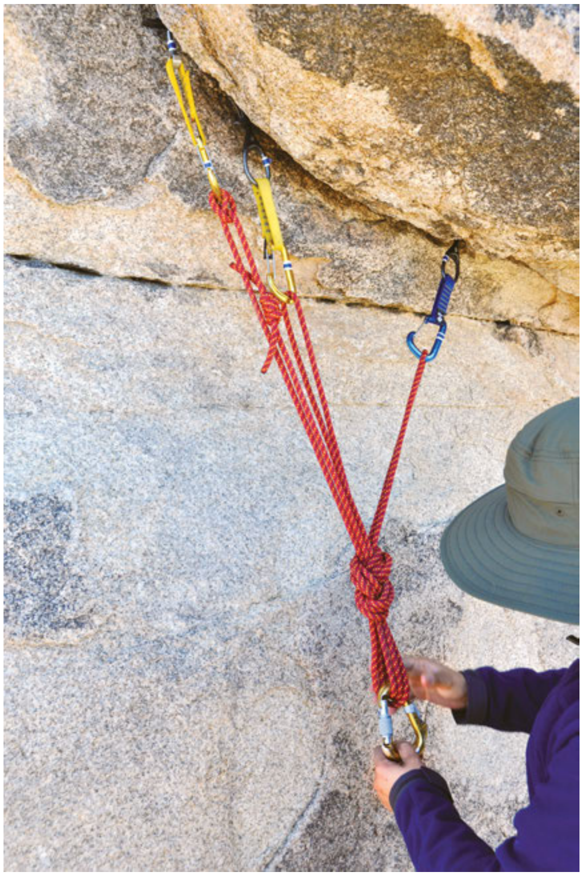

Demonstration of pre-equalized cordelette with three anchor placements, tied with a 7mm nylon cordelette. A clove hitch has been tied to the top left piece to keep the double fisherman's knot away from the end loops. This is a simple and effective rig as long as the direction of load is predetermined, which is most often the case when toproping.

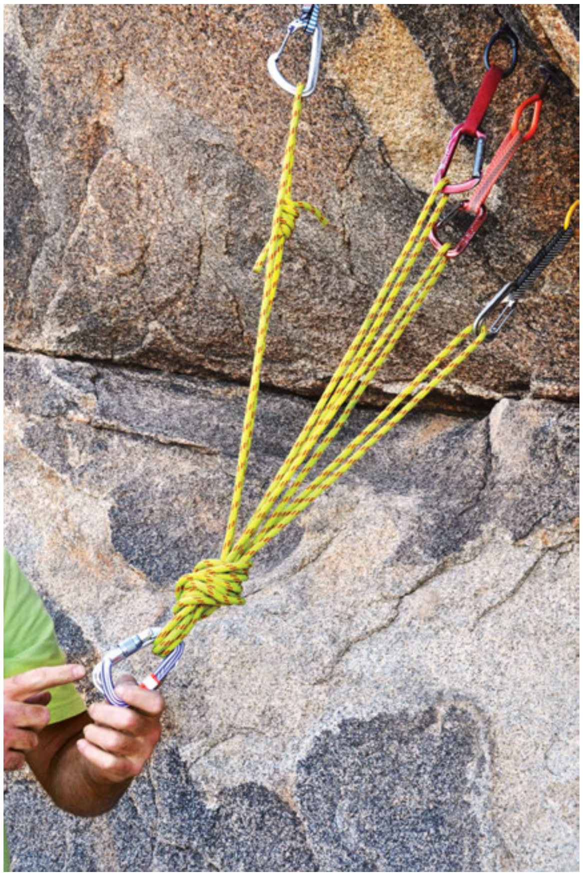

This four-piece anchor is pre-equalized with a 7mm nylon cordelette. The overhand knot creates a four-loop master point to which the two opposed and reversed locking carabiners are clipped.

Rigging a Cordelette

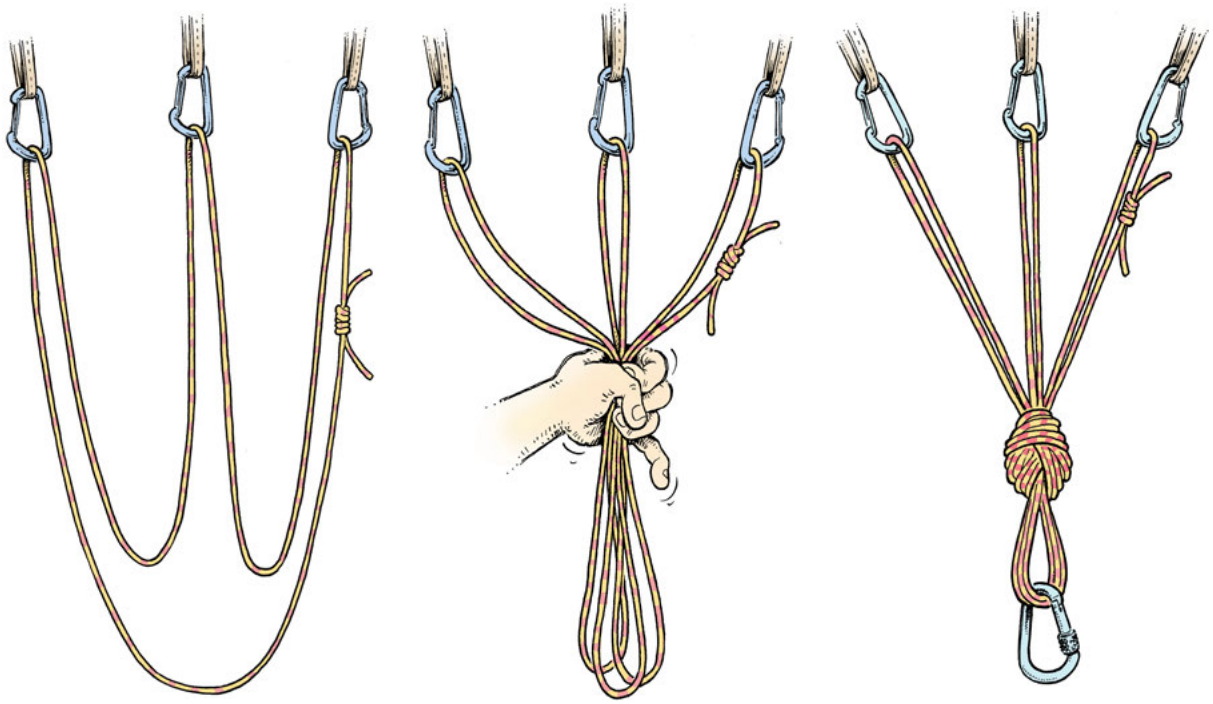

An 18- to 20-foot-long cordelette is usually long enough to equalize three or four anchor points, as long as they are not spaced too far apart. Use a sling or two if necessary to get all the carabiners you'll be clipping into within a workable range. Clip a single strand of the cordelette into each carabiner, then pull down between the pieces and gather the loops (with three pieces you'll have three loops). Clipping a carabiner into the loops before tying the knot will make it easier to equalize all the strands. Tie a figure eight knot to create your master point, which should be roughly 3 to 4 inches in diameter. If you don't have enough cord to tie a figure eight, an overhand knot takes up less cord.

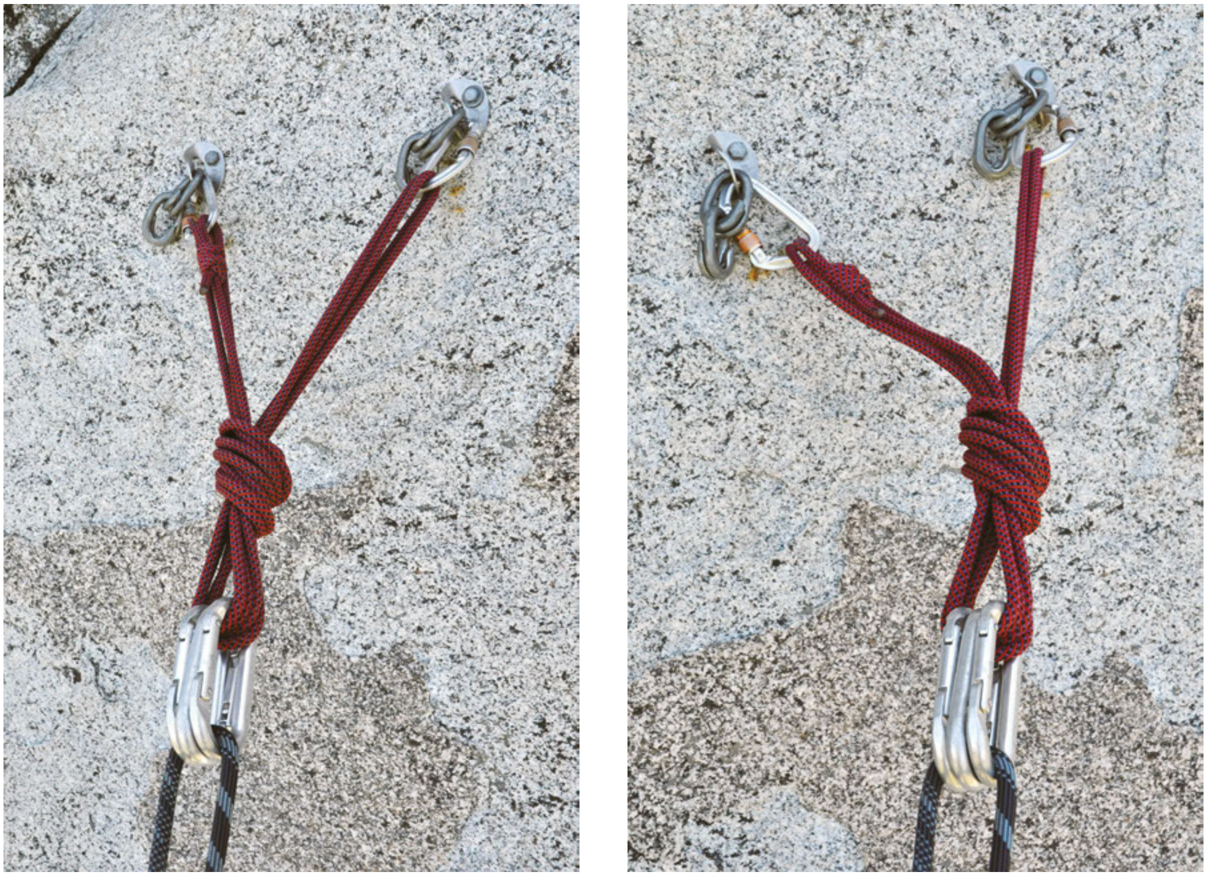

The drawback of the cordelette system (left) is that if the direction of the anticipated load changes, one piece in the anchor takes all the load (right). Think of the cordelette system as a system of backups: If the one piece that is loaded fails, the load goes onto the next piece with relatively minimal extension in the system. For toproping anchors, since the load on the anchor system is relatively low, the cordelette system has the advantage of being easy to use and simple to rig, negating any potential for shock loading.

The Sliding X

The sliding X (aka magic X) is a simple way to distribute load to two anchor points with a sling, creating an anchor system that adjusts as the load shifts in one direction or another. If using the sliding X with a long sling (like a sewn, 48-inch double-length sling), it is possible to minimize extension by tying overhand knots just above the clip-in point. This allows the system to adjust, but limits any extension if one piece fails. Or two single-length slings can be used together with a sliding X, creating a redundant rig with minimal extension.

Rigging a Sliding X

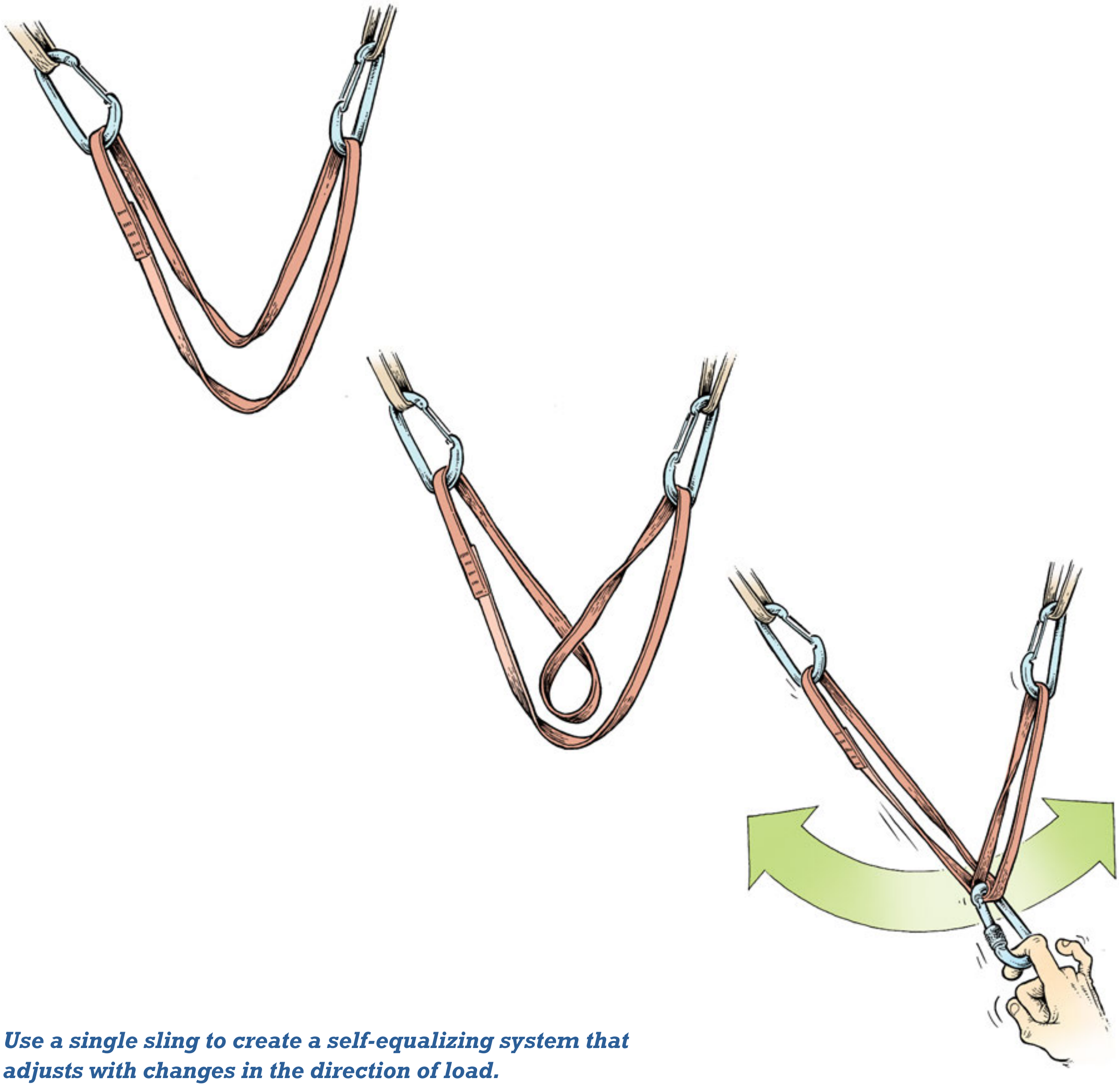

Use a single sling to create a self-equalizing system that adjusts with changes in the direction of load.

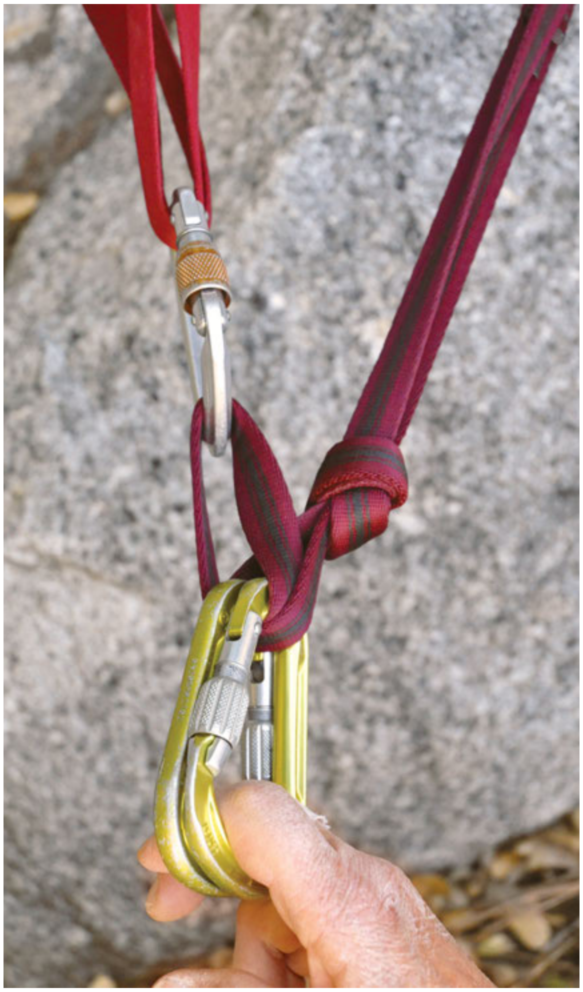

As an anchor system, this rigging is not redundant. For the sling itself to be redundant, there must be an overhand knot tied on both sides of the sliding X.

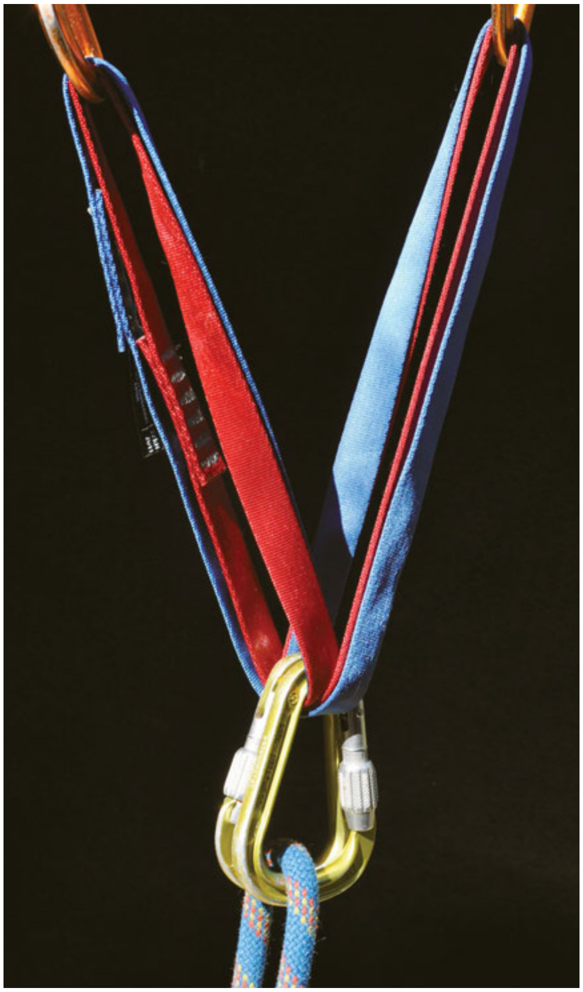

A simple two-bolt anchor can be rigged with a sliding X using two slings and two locking carabiners at the master point for a redundant, self-equalizing system. With a two-bolt anchor, many instructors use locking carabiners on the bolt hangers too. The drawback of this rig is that if one bolt were to fail, the system would extend to the length of the slings. As a general rule of thumb, limit the maximum extension in your anchor system to half the length of a single (24-inch) sling.

The Quad

The quad is a slightly more complicated system for a variable load direction using a cordelette. It's a great system to use for toprope anchors. It gives near perfect equalization with minimal extension and great strength. To rig the quad, start by doubling a cordelette, then grab the middle with the fist. Tie an overhand knot on each side of the fist, and it is ready to rig. Clip the double-strand loops into the bolts with locking carabiners, then clip only three of the four strands at the master point, leaving one loop outside the master point carabiners. This ensures that if one bolt fails, carabiners remain clipped into a pocket on the master point.

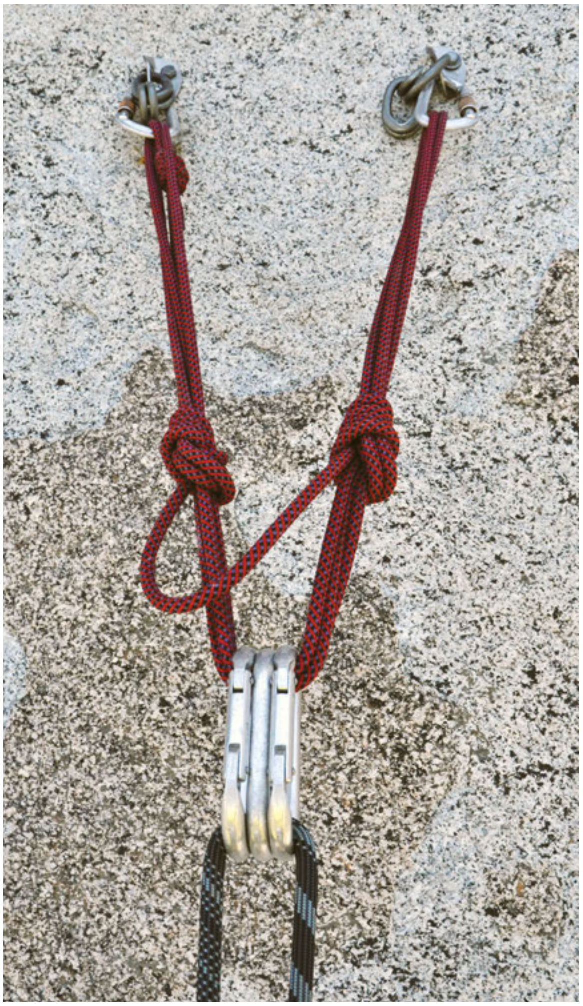

A two-bolt anchor rigged with a 7mm cordelette and the quad system.

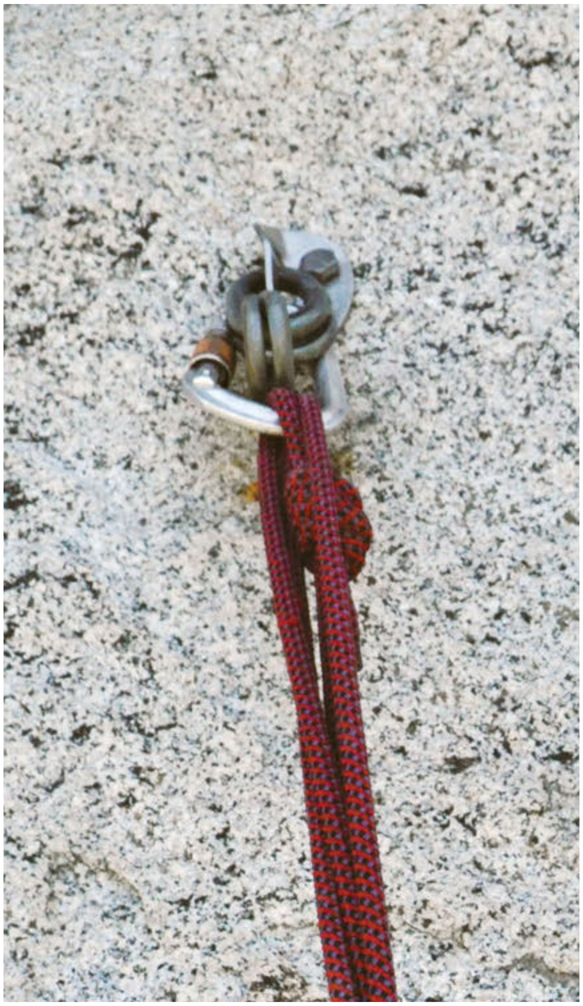

The cordelette is clipped directly to the bolt hangers with locking carabiners, bypassing the cheap hardware store lap links (which are only rated at around 1,000 lbs.).

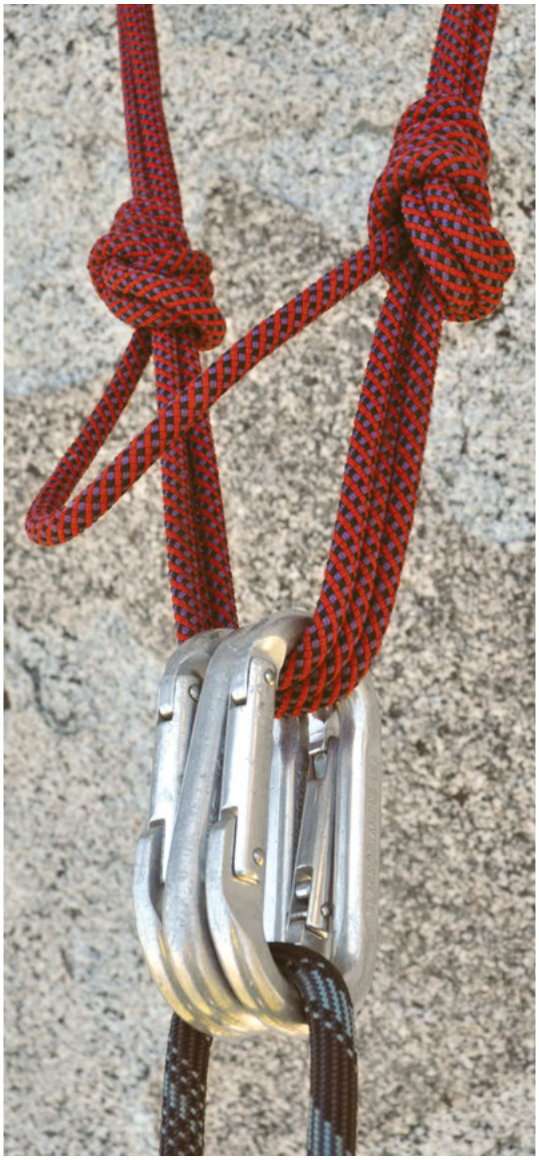

Detail of quad rig master point with three ovals opposed and reversed.

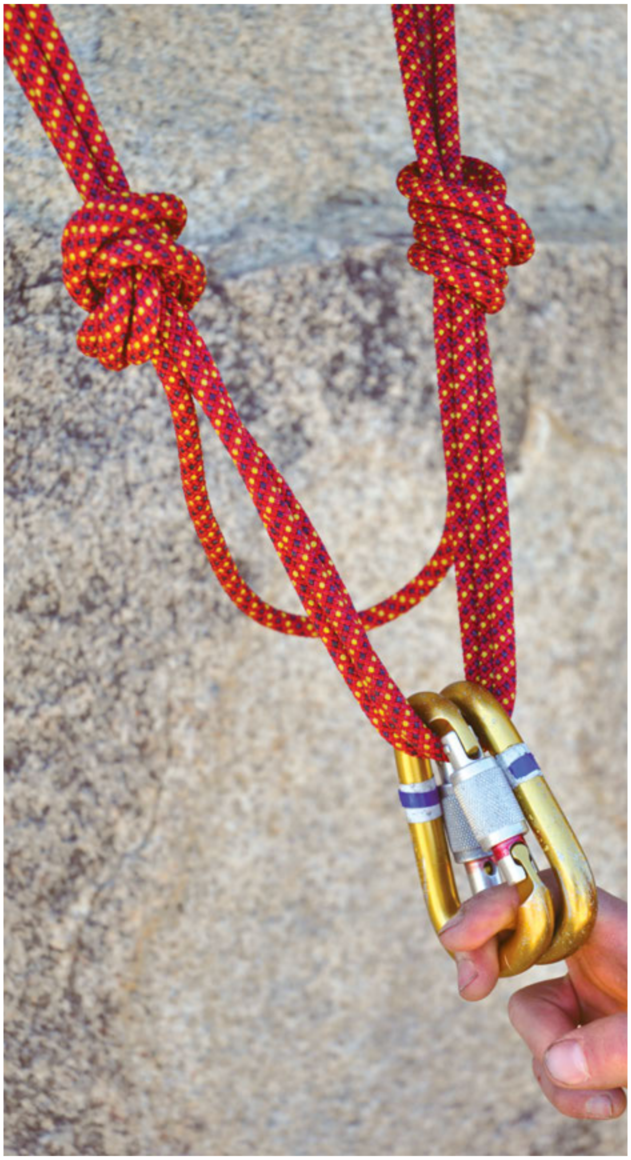

Detail of quad rig with two locking carabiners opposed and reversed.

Stacked Sliding Xs

This gives you a complex rig for a variable load with slings. Occasionally an anchor builder encounters anchor configurations that have components with very different strengths. Imagine a narrow 4-inch crack that immediately closes to a ½-inch crack running along the top of the crag. An effective anchor builder understands that a 4-inch SLCD is much stronger and has a much higher margin of error in its placement (due to its large camming range). By comparison, ½-inch SLCDs are weaker and have a minuscule camming range. These components are not equally strong, and therefore distributing load to them in an equal manner would not take advantage of their relative strengths and weaknesses. In other words, equalizing the pieces would not be an intelligent distribution of the load. An effective alternative would be to build a sliding

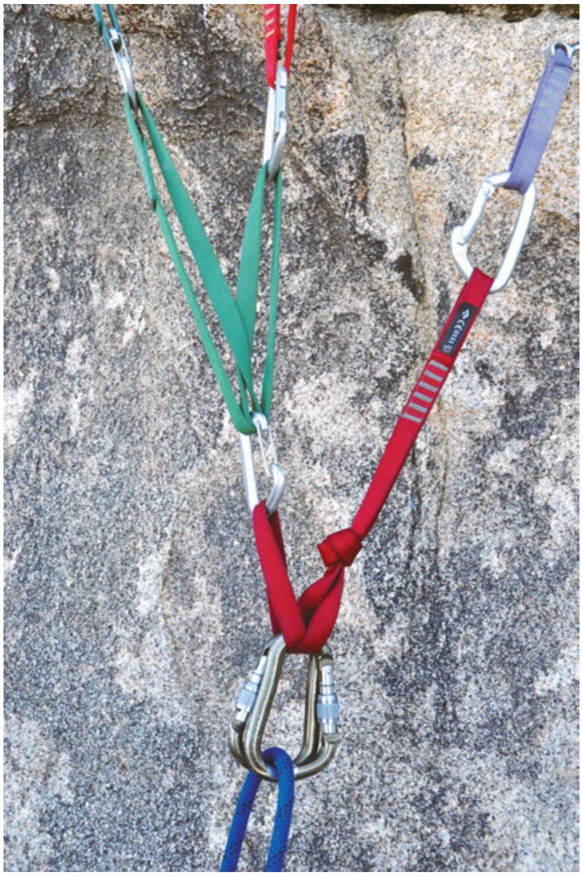

Here the green sling equalizes two components of lesser strength than the top right piece. Although this anchor system achieves good equalization and logical load distribution, it is not redundant because the red sling is not redundant at the master point.

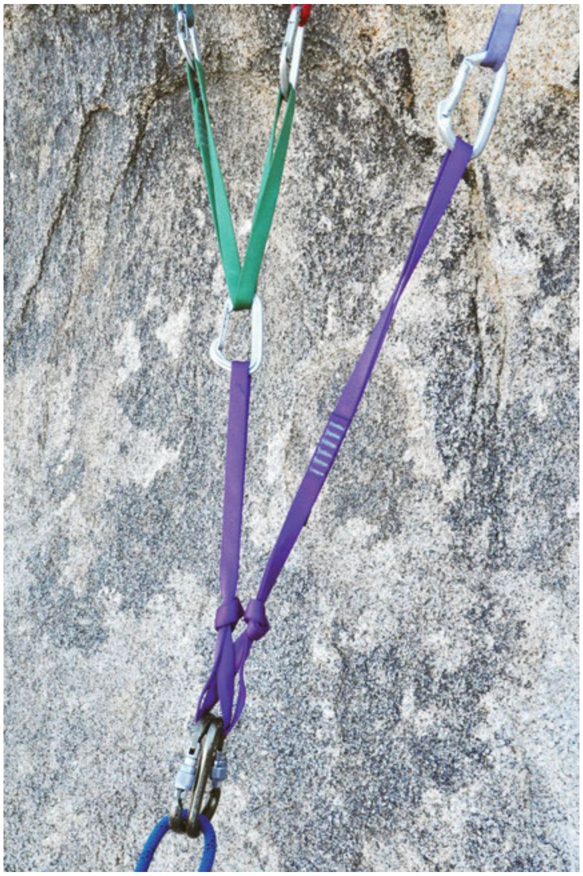

The rigging is now made redundant by tying an extension-limiting overhand knot on each side of the sliding X, although the green sling will not minimize extension.

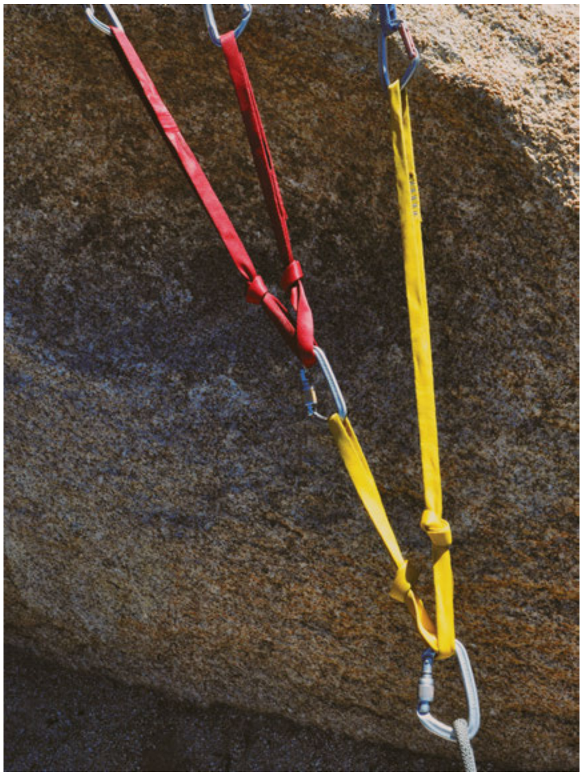

Here the rigging is redundant with minimal extension. The two pieces on the left have been equalized with the red 48-inch nylon sling, using a sliding X and extension-limiting overhand knots. The yellow 48-inch nylon sling equalizes this point to the single, strongest piece in the anchor, again with a sliding X and extension-limiting overhand knots.

X with load-limiting knots between the two ½-inch SLCDs, creating a mini–master point. Then a second sliding X with load-limiting knots would connect the 4-inch SLCD to the first sliding X, to that mini–master point.

The Equalette

The equalette is a complex rig for a variable load direction using a cordelette. It gives the anchor builder four strands, or "legs," running from the master point to the various pieces in the anchor

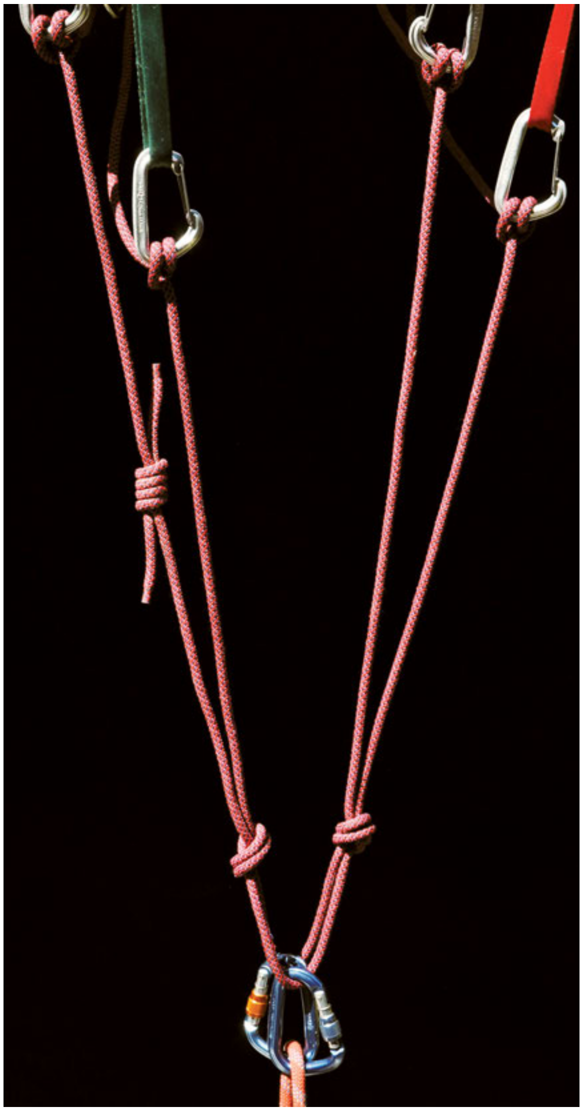

The equalette rigged with a 7mm nylon cordelette. This is a versatile system to use for equalizing three or four placements, giving you redundancy, equalization, and no extension. The only drawback is in its complexity and the fact that it does not have one singular master point to clip into. Here the equalette is rigged to four placements, using the various "arms" of the cordelette attached with clove hitches for easy adjustment.

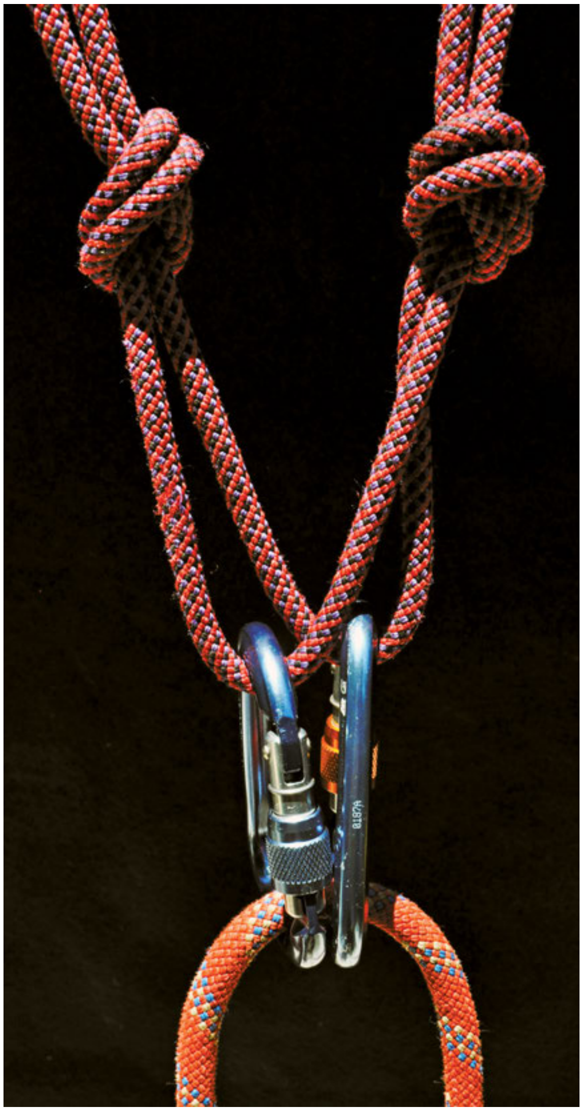

Detail of the equalette master point using two locking carabiners opposed and reversed.

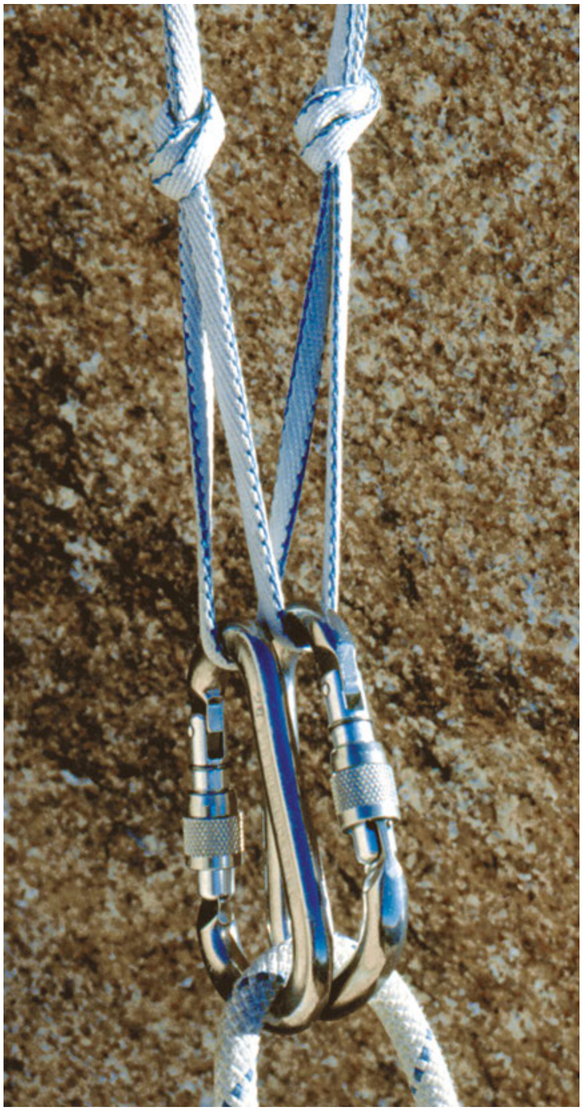

Detail of equalette master point rigged with a Dyneema sling. Tying an overhand knot in a Dyneema sling reduces its strength by 50 percent; it's better to use a 7mm nylon cordelette when rigging the equalette, as you'll be tying lots of knots.

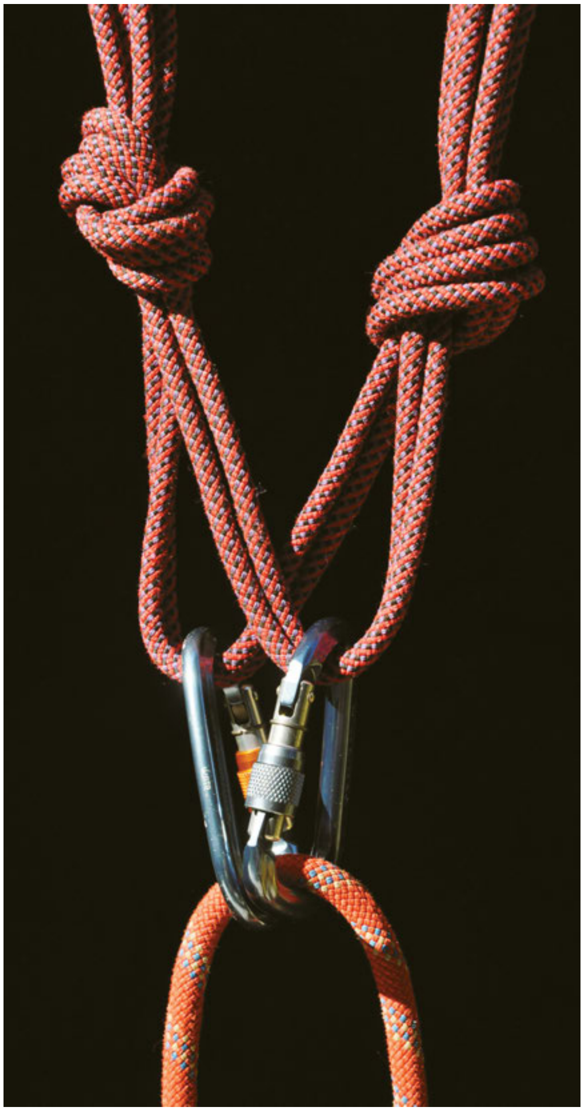

Detail of an equalette toproping rig tied with a doubled cordelette for a two-bolt anchor.

matrix. These four legs can be tied to the pieces with figure eights, clove hitches, or double loop knots like the double loop eight or double loop bowline.

To tie an equalette rig, form a U shape with the cordelette and grab the bottom of the U, positioning the fisherman's knot on the cordelette about 18 inches away from the bottom of the U. Tie an overhand knot on both sides of your fist, about 10 inches apart.

At the master point there will be two separate loops. Clip into each loop with a separate locking carabiner.

Vectors

A vector is a quantity that incorporates both direction and magnitude. Picture a tightrope walker balancing out on the middle of a wire. If he weighs 200 pounds, the load at each end where the wire is attached will be roughly 1,000 pounds. Why is this? When two anchor points are equalized, as the angle of the wire, sling, cord, or rope approaches 180 degrees, the forces at the anchor points increase drastically. When the angle is narrow, the load is distributed at around 50 percent to each anchor.

Keep this in mind when building toprope anchor systems. If the angle between two anchor points reaches 120 degrees, the load at each anchor is 100 percent. Strive to keep all the angles under 60 degrees to split the load roughly 50/50. A good way to help guarantee an intelligent load distribution is to keep the angles in an attachment system under 90 degrees. Also, avoid rigging a sling between two anchors in a triangular configuration (called the American Triangle), which, even at 90 degrees, places 1.3 times the force at each anchor point. An American Triangle rigged at 120 degrees would almost double the load at each anchor point!

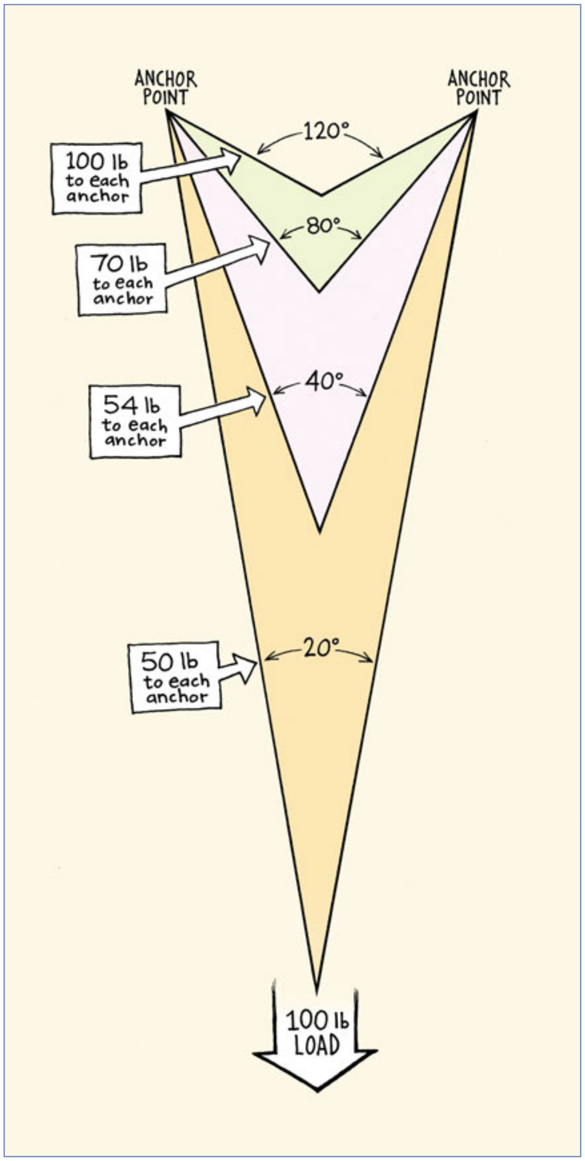
ANCHOR POINT
ANCHOR POINT
120°
100 lb to each anchor
80°
70 lb to each anchor
40°
54 lb to each anchor
20°
50 lb to each anchor
100 lb LOAD

This diagram illustrates how a 100-pound load is distributed between two anchor points at various angles. Keep the angle between two anchors as narrow as possible, striving to keep it under 60 degrees. At 120 degrees the load is 100 percent at each anchor! Think of 0 to 60 degrees as ideal, 60 to 90 degrees a caution zone, and over 90 degrees a danger zone.

American Triangle

Load per anchor with 100 lbs. of force

Bottom Angle	**V Rigging**	**Triangle Rigging**
30 degrees	52 lbs.	82 lbs.
60 degrees	58 lbs.	100 lbs.
90 degrees	71 lbs.	131 lbs.
120 degrees	100 lbs.	193 lbs.
150 degrees	193 lbs.	380 lbs.

The American Triangle rigged at a rappel anchor. Avoid rigging with a triangle configuration—it adds unnecessary forces to your anchor points. Stick to a V configuration for lower loads.

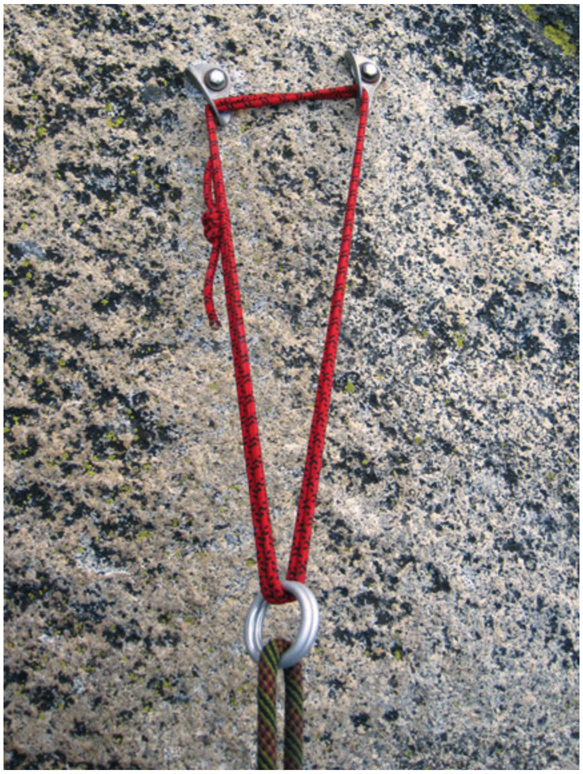

CHAPTER 9

Advanced Anchor Rigging Systems and Institutional Anchors

In the previous chapter, fundamental anchoring principles were discussed with strategies to create an unquestionably effective attachment point. Those principles will carry an anchor builder through all disciplines in climbing. The Single Pitch Instructor, however, will also require further skills for establishing an anchor and setting a toprope, because there are a few challenges that are unique to a Single Pitch Instructor's terrain:

1. Sometimes the components in an anchor, natural or otherwise, are far apart from each other, too far for a cordelette or a 48-inch sling.
2. The cliff's edge is an especially dangerous place for an instructor who is trying to keep safe and supervise participants.
3. When working at the base of a climb, a master point of the anchor has to be positioned in a place that is hazardous, such as over the cliff's edge.

Using a Setup Rope: Components Far Apart

Thankfully, a Single Pitch Instructor can bring along extra tools for anchor building. Unlike a multi-pitch or alpine climbing team, a single pitch team can afford to carry a little bit of extra weight. A length of static or low-stretch rope is worth every ounce of extra weight. To use it effectively and efficiently, an instructor should be able to stand at the site of the intended master point, look back into the terrain, select potential components, then use the setup rope to connect those components to the master point. A common sequence might proceed as follows:

1. Determine the climb and a distance from the cliff's edge directly above the climb where there is little risk of falling over the edge.
2. Face the terrain and use raised arms to target potential components (trees, boulders, removable and permanent protection). Let the arms forecast the vectors the setup rope will eventually create to make sure component selections are not too far apart.
3. Flake out the rope and take the end of the rope out to the most distant component. Attach the rope to that component in whatever manner is most efficient and effective.
4. Leave enough slack at the site of the intended master point to tie a BHK, then take the remaining bight of rope back to the other component.
5. Once both components are connected, a closed U shape awaits the anchor builder at the site of the intended master point. Use this bight of rope to tie a BHK that distributes load to the components along the anticipated direction of pull (toward the climb).

This sequence of connecting a setup rope to components and tying a BHK has limitless variations. Becoming efficient takes practice.

Instructor Tether: Managing Risk at the Cliff's Edge

Most climbers and climbing instructors alike have become familiar enough with the danger of the cliff's edge that their balance, their attentiveness, and their caution allow them to monitor and regulate proximity to a drop-off. Climbing instructors, however, cannot rely entirely on these skills because they are also supervising their students, observing the environment, and trying to complete setups efficiently. In other words, they are more distracted,

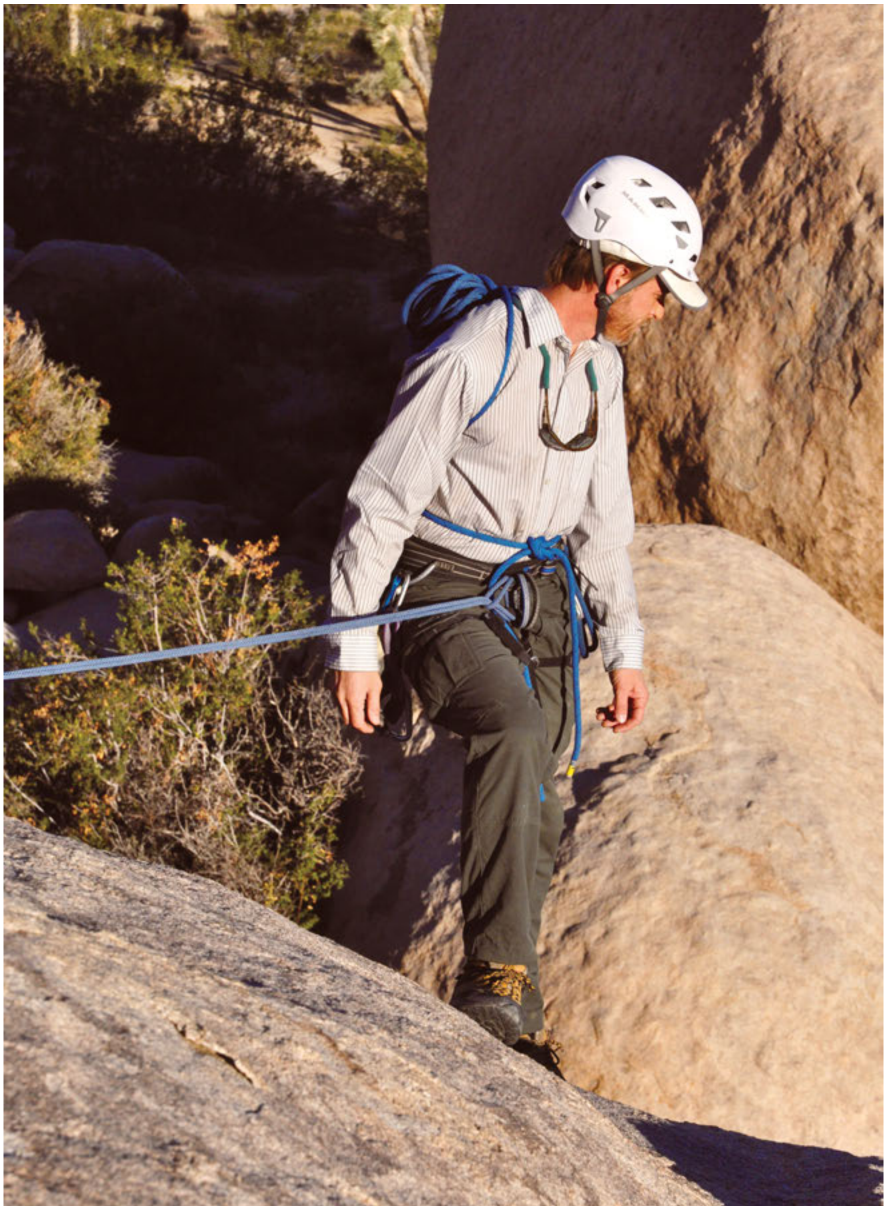

A simple tether using a cordelette. The instructor has girth-hitched the 7mm nylon cordelette through both tie-in points on the harness (then clipped to an anchor with a locking carabiner.

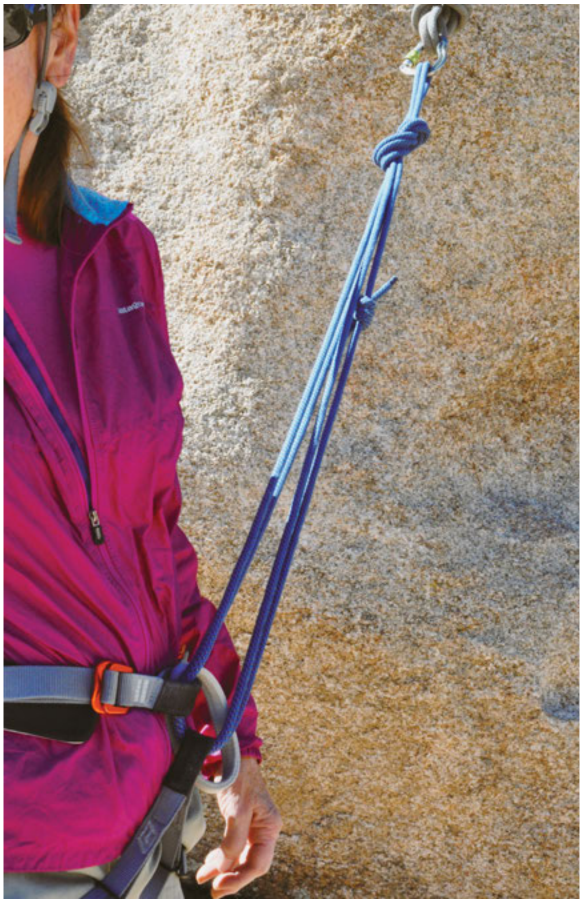

Here the cordelette has been threaded through both tie-in attachment points and tied with an overhand knot for redundancy to create a simple tether.

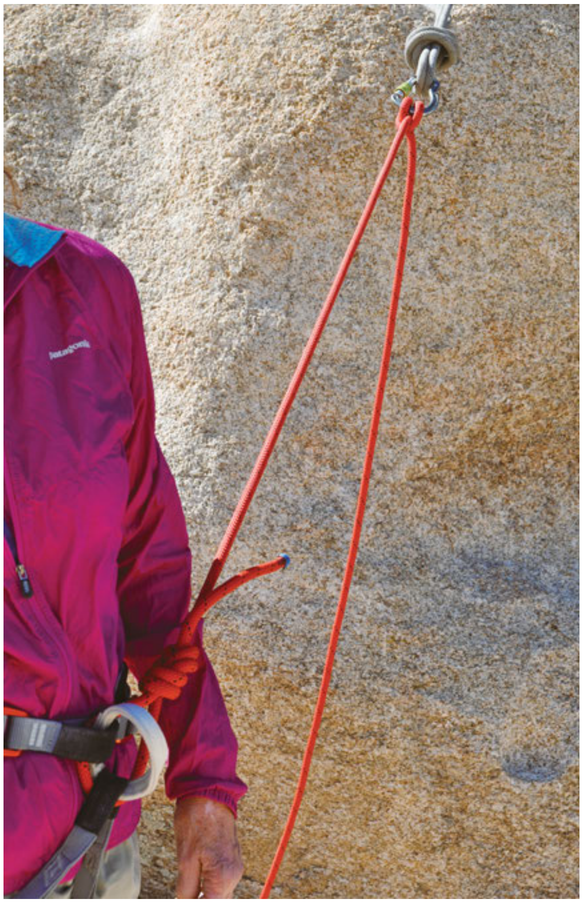

A simple tether using the rope tied with a clove hitch at the anchor.

and distraction is very dangerous indeed. Fortunately, climbing instructors have many options to secure themselves when in proximity to a cliff's edge; they can construct some form of instructor tether.

- Fixing a section of rope to an anchor allows an instructor many tethering options—using a Grigri, an ATC with friction hitch backup, a clove hitch, a full-strength sling with a klemheist, or even tying in to the rope.
- Any number of slings and cordelettes might work nicely to create a long leash. Instructors can girth-hitch one end of a closed loop to the harness and attach the opposite end of the loop to the master point.
- The climbing rope itself is always an obvious tethering tool. Instructors can easily tie in with a figure eight follow-through and then attach the climbing rope to the anchor with a clove hitch.

When using a Grigri on an instructor tether, tie an overhand knot backup whenever you take your brake hand off the rope. Always tie a stopper knot in the end of a rope tether to close the system.

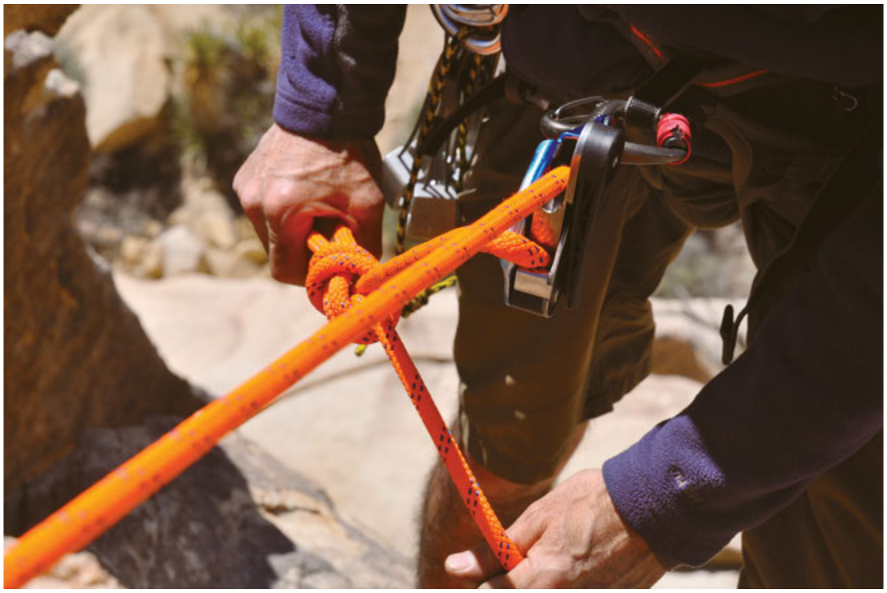

Extensions: Creating a Master Point Over the Cliff's Edge

Once an instructor has a master point and a tethering method, it may be necessary to create an extension for a secondary master point, which will allow a base-managed toprope to run through a pair of carabiners free and clear of the cliff's edge. Many instructors refer to this secondary master point as the action point, or hot point. Either way, since it is a part of the anchor, the same fundamental principles that govern the master point should also govern the extension. For a typical extension, the instructor might attach a section of static or low-stretch rope to the initial master point with an overhand or figure eight loop then feed out a loop of slack that hangs over the cliff's edge, closing the long loop with another overhand or figure eight loop. Now, using the tether, an instructor can comfortably and securely set up a position at or just over the edge to tie a BHK. This is the point where the climbing rope will finally be placed.

Systems Using a Setup Rope

Aside from being the best tool for the individual tasks mentioned so far, a setup rope further proves itself as an elegant and efficient anchoring tool because it can be used to accomplish all three tasks—combine components, build tethers, and create extensions. When the full range of possibilities are understood by a climbing instructor, certain patterns emerge in the use of the tool; we call those patterns "systems." It should be understood, however, that the most elegant and efficient use of the tool often involves a combination of the techniques demonstrated. An instructor will want to explore the most applicable terrain, the programmatic options, and the economy of materials used for each technique when deciding how to use the setup rope to build a given anchor.

The Joshua Tree System—V Rigging with Extension Rope

Most anchors in Joshua Tree National Park in California require gear placements set well back from the cliff's edge, and bolted anchors are a rarity. Out of necessity, instructors at Joshua Tree developed a system to rig toprope anchors that is both efficient and redundant, using a length of low-stretch rope.

To rig the Joshua Tree System, visualize a V configuration, with two separate sets of anchors on the top of the V, and the point, or bottom, of the V being the toprope master point over the edge of the cliff. Ideally, the angle of the V should be as narrow as possible—at least less than 90 degrees. Once you have determined where the climb is and where you want your master point, picture the V in your mind and begin to set your anchors. If using natural anchors, it could be as simple as two trees. For simplicity let's say we have two anchors: anchor A and anchor B. Start by attaching one end of your static rigging rope to anchor A. This can be done either by using a sling or cordelette and a locking carabiner to a figure eight loop, or by tying the rope directly around the tree. (The re-threaded bowline works well for tying a rope to a tree.)

Now, to increase security as you approach the cliff's edge, protect yourself by taking a 48-inch nylon sling and attaching it to your rigging rope with a klemheist knot. The 48-inch sling with klemheist is an excellent full-strength tether option in this system; be sure to attach the sling to the belay loop with a locking carabiner. A nylon sling is better for this application because nylon has some stretch, whereas Spectra or Dyneema are static (like a wire cable) and have no stretch. Now you can slide the klemheist knot up and down the rigging rope to safeguard yourself as you work near the edge. Tie a BHK so that your master point dangles just over the lip of the cliff's edge, positioned directly above your chosen climb. Attach your climbing rope with carabiners (either two opposed and reversed locking, or three opposed and reversed ovals) and run the rope back to anchor B, attaching it with a clove hitch to a locking carabiner. This will allow you to adjust the tension and fine-tune the load distribution.

The Joshua Tree System allows the anchor builder to tether to a leg of the anchoring system. That's why the klemheist and 48-inch nylon sling is an important selection for tethering—any other method would not allow the anchor builder to precisely position the master point. But it is almost exclusively a system used for base-site climbing anchors, since neither of the component anchors are positioned directly above the climb and the master point itself is positioned so far below the edge of the climb that top-site management would be impractical. The Joshua Tree System also boasts excellent material economy. It only requires enough material to extend out to the cliff's edge; that is, two times the distance between the component anchors and the edge (with a little extra for knot tying).

Personal Preference

I (BG) prefer to tie both legs of the V into anchor points using figure eight loops. I know from experience that the BHK takes about 4 feet of rope to tie. So I'll tie both ends of the V first, with about 4 feet of the extension rope dangling over the edge, weighted down by a few carabiners to find my equalization point. Then I'll pull the point of the V back up and tie my BHK last.

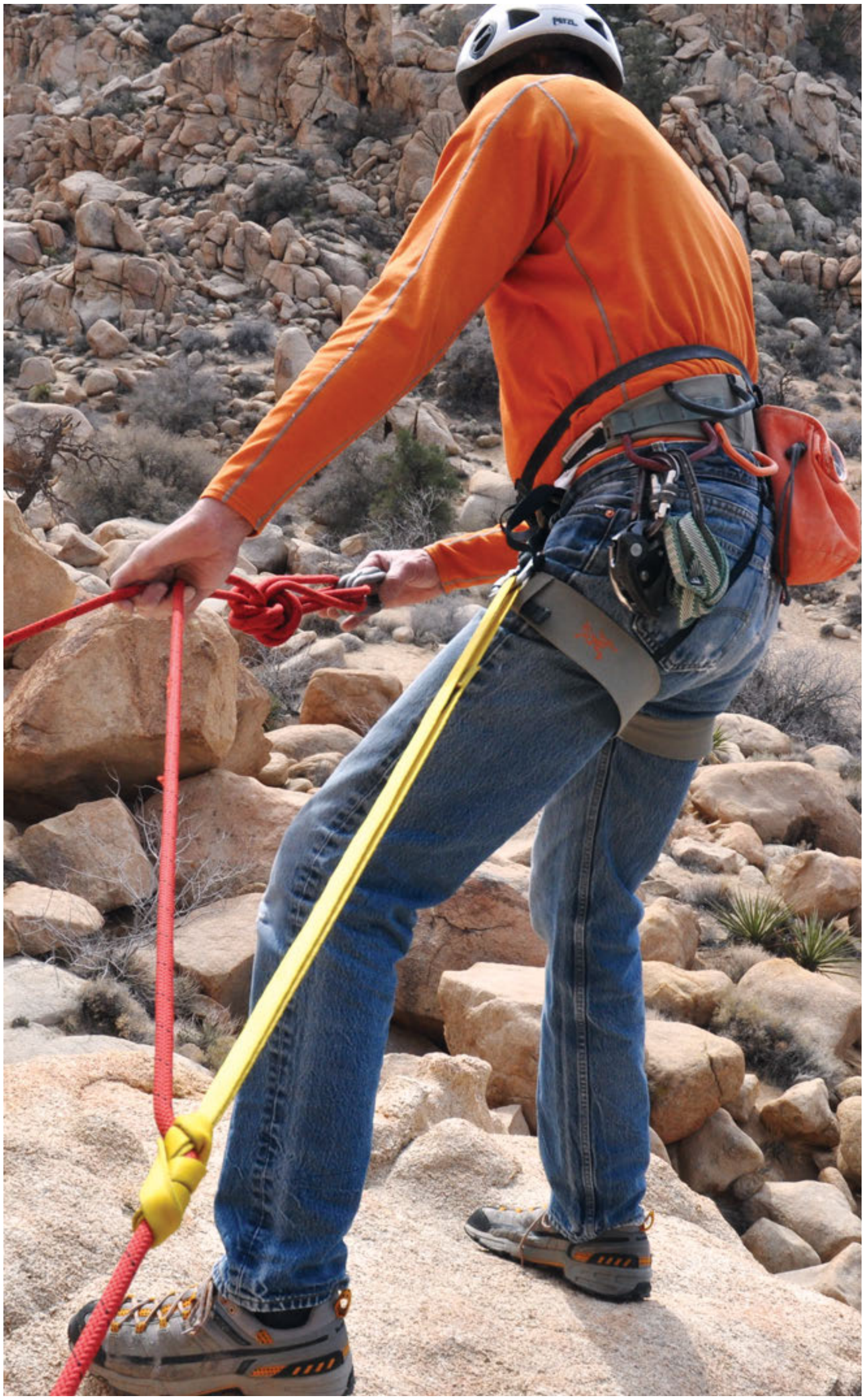

Proper use of a friction hitch for security while rigging a toprope. Here the instructor is tethered with a 48-inch nylon sling, attached to one leg of the setup rope with a klemheist knot and clipped into the harness belay loop with a locking carabiner.

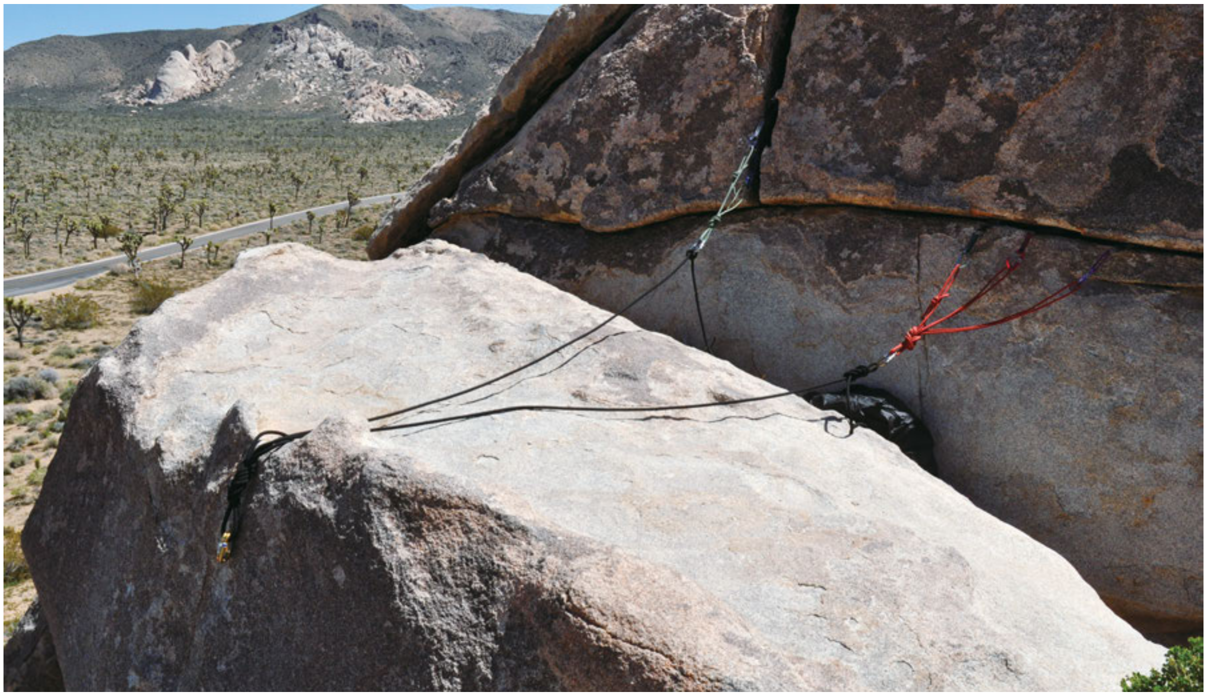

Overview of the Joshua Tree System. The right "leg" of the extension rope is attached with a figure eight loop to a locking carabiner clipped to a master point on a three-piece anchor pre-equalized with a cordelette. The left leg of the extension rope goes to a two-piece anchor pre-equalized with another cordelette, attached with a clove hitch on a locking carabiner for adjustment of the extension rope. A BHK is tied for the master point, with two opposed and reversed locking carabiners ready for the climbing rope.

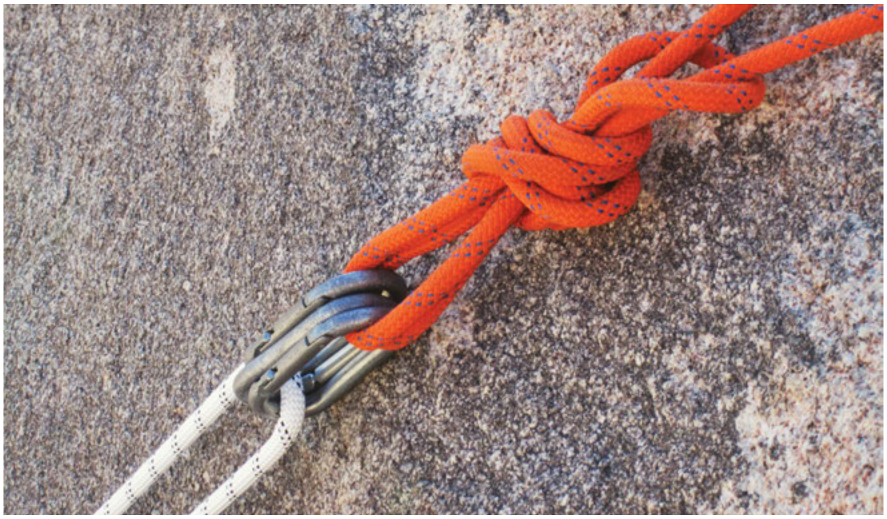

Close-up of the master point on the Joshua Tree System using a BHK and three steel oval carabiners with the gates opposed and reversed.

Toprope all day long with your extension rope rubbing on a sharp edge . . .

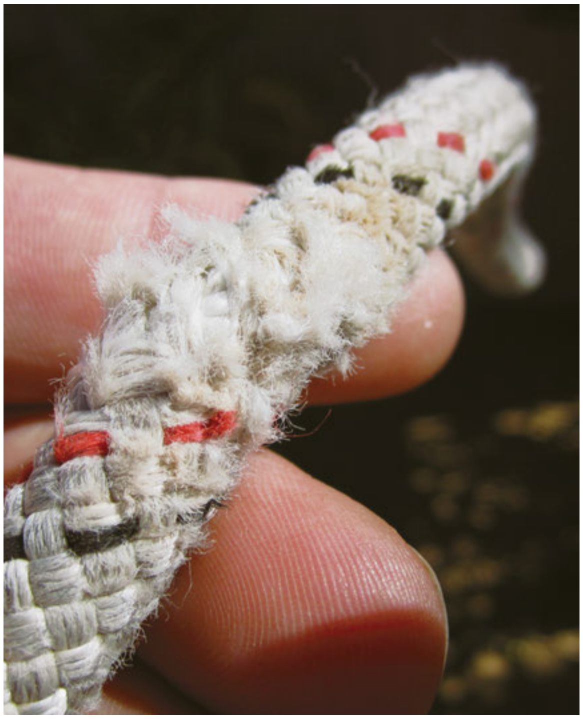

. . . and you'll end up with a seriously abraded rope like this one.

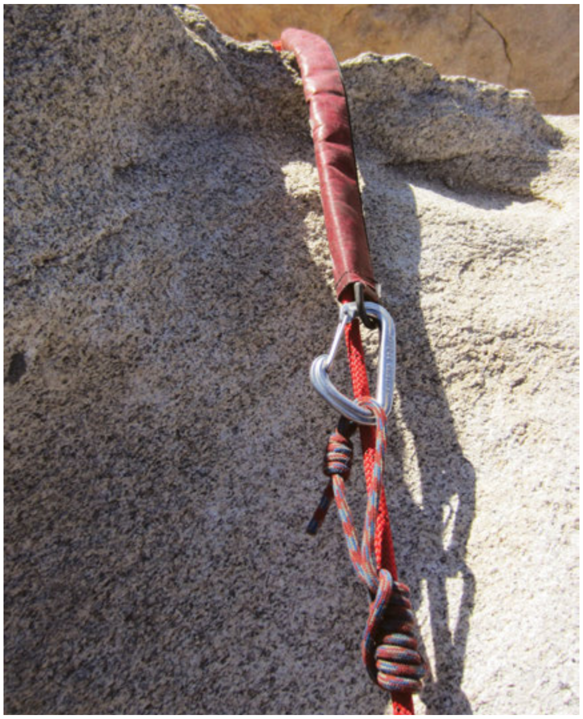

A commercially made edge protector, like this one sold by Petzl, is a wise investment. Attach it to the rigging rope with a friction hitch.

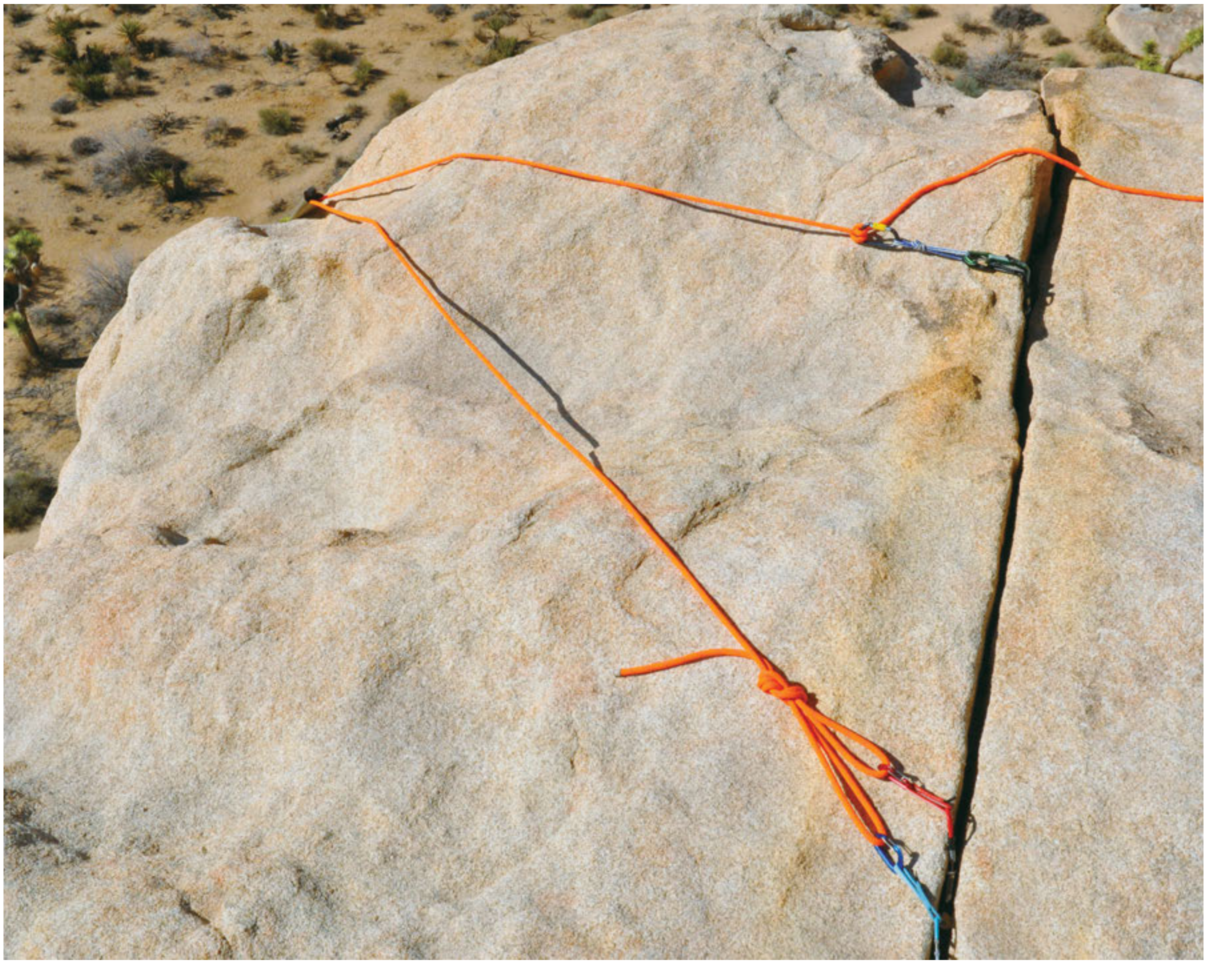

This finished rig illustrates the simplicity of the Joshua Tree System, with the lower leg connected to two anchors using a double loop figure eight. The upper leg was connected last, using a clove hitch to a locking carabiner for equalization, attached to two pieces and a doubled sling. Note the edge protector guarding both legs of the extension rope from abrasion.

Take care to make sure the extension rope is not resting over sharp edges at the lip of the cliff. This setup is an "unmonitored" anchor system, which means that once rigged, the climbing instructor will be at the base and not able to watch what is happening at the anchor—like the extension rope abrading over an edge, so take special care to prevent this by padding the edge; a pack or rope bag will work or, better yet, use commercially made edge protectors.

If you learn to tie double loop knots like the double loop figure eight and double loop bowline, along with the in-line figure eight, you'll be able to eliminate many slings and cordelettes from your anchor system and become more efficient in your rigging.

The Lifeline System—V Rigging with Instructor Tether

This is a variation of the Joshua Tree rigging method. Instead of tethering with a klemheist and a 48-inch sling, an instructor may find it advantageous to tether with a fixed section of rope. While the Joshua Tree System attaches the end of the rope to one of the component anchors, the Lifeline System attaches a figure eight with a bight so that two load-bearing strands are available for use. The one strand is only long enough

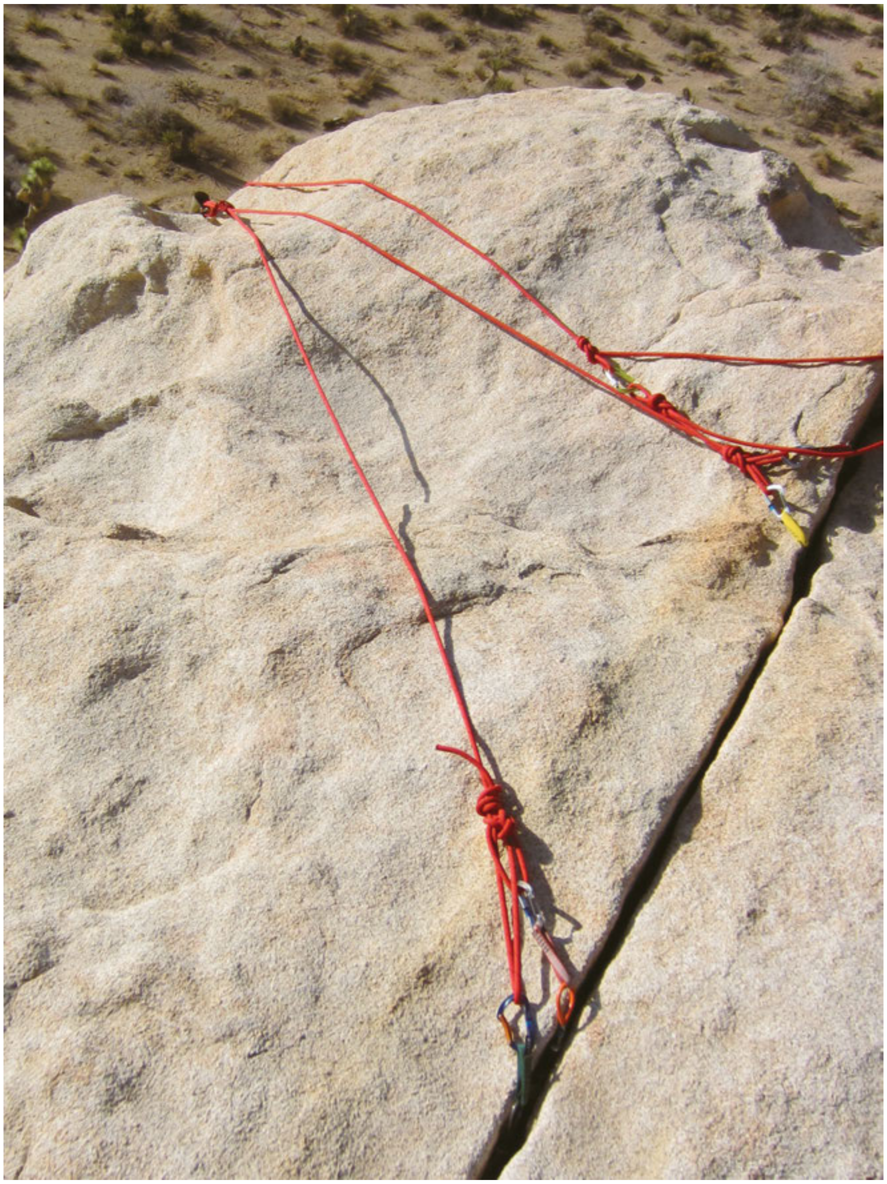

Example of V rigging with a tether. Each leg of the V consists of two bomber pieces. The bottom leg is equalized to the two placements with a double loop bowline, and the top leg (with tether) is equalized with a double loop figure eight.

to access the cliff's edge and serves as an instructor tether, while the second becomes the same leg of the anchor system that the Joshua Tree System utilizes. The important concept to remember here is that the tether should be clipped into a bona fide, redundant anchor—not just a single piece of gear.

An astute instructor should immediately conclude that this variation of the Joshua Tree System does not have the same material economy as its forbearer. While the Joshua Tree System only requires two lengths of material to connect the component anchors to the master point, the Lifeline will

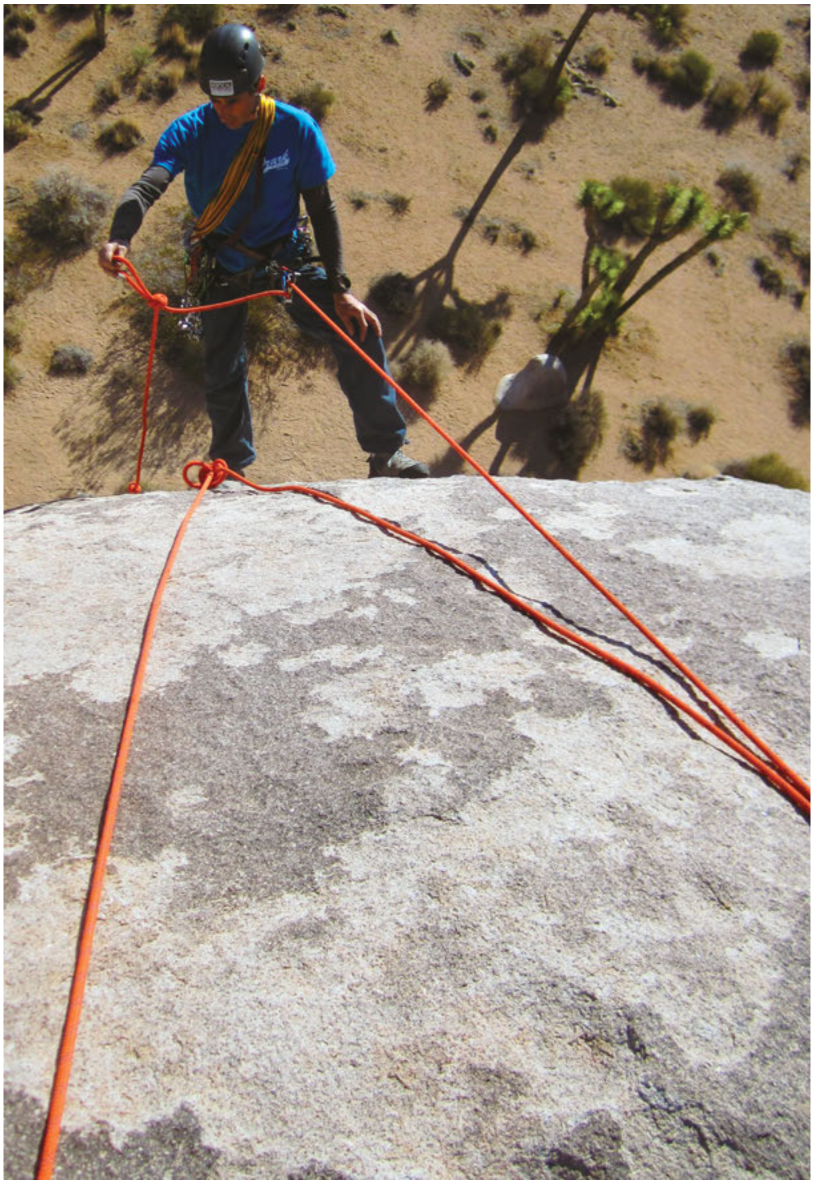

Here the instructor has rappelled using an instructor tether to approach the edge in low-angle but exposed terrain to make some adjustments on his BHK master point toprope rig. Whenever you take your brake hand off the Grigri in low-angle terrain, tie a backup knot (overhand loop) as shown here, because if you unweight the Grigri, the rope can slide through. Always tie a stopper knot on the end of your tether before you begin using it.

require three, with a little extra for knot tying, of course.

Instructors may find the Lifeline to be an effective system if the other half of the anchor, the second component anchor, is in a precarious or inaccessible place. The Lifeline would provide access to those components so that the anchor can be completed. The Lifeline is also an excellent tool for probing an ambiguous cliff's edge to pinpoint the location of the climb. Lastly, the Lifeline might be an excellent tool to access anchoring hardware normally intended for sport climbing use. Such anchors are typically installed over the cliff's edge, where access from above may not only be atypical, but dangerous without the Lifeline System.

The Backside System—Pre-Equalized Extensions

Unlike the Joshua Tree or Lifeline Systems, the Backside System preestablishes a space for both a top-managed system and a bottom-managed system at the same time. In the simplest terms, this is done by setting up a V configuration and tying two BHKs into it. The top BHK becomes the master point for top management. The backside of that top BHK should be long enough to droop over the edge of the cliff, where a second BHK is tied. The bottom BHK then becomes the master point for bottom-managed operations.

Similar to the Joshua Tree System, an instructor will be tethered to a leg of the anchor system. A 48-inch nylon sling tied as a klemheist is a logical choice for tethering.

The Backside System can boast certain material economies that the Joshua Tree and Lifeline Systems cannot because it does require entire component anchor points. It presumes that the initial master point will be positioned above the climb. But its greatest strength is its programmatic versatility. As mentioned, it can easily be used for top-managed sites as well as base-managed sites. Furthermore, when built in anticipation of both kinds of activities, it allows the instructor to quickly transition from one to the other.

Rigging the Backside System

1. Configure a Joshua Tree V with a loop of rope hanging over the edge of the cliff. Make sure to estimate for the amount of rope that will be eaten up by the dual BHKs.
2. Determine where you would like your top-managed high master point to be and tie a BHK there. The backside of the BHK should be long enough to hang over the edge of the cliff.
3. Tie a second BHK in the loop that hangs over the edge. There are two options to secure the toprope BHK: Pass the BHK through the backside of the knot or clip all three loops of the BHK into the toprope carabiners.
4. Use a klemheist with a 48-inch nylon sling to tether to one of the ropes directly below the high master point. Do not clip into this friction hitch until you are ready to make a transition to the bottom.
5. If operating from the high master point, pull the backside of the high master point up and stack it out of the way.

Making the Transition from Rigging to Rappelling

When using any of the aforementioned systems, if you decide to rappel to the base, you will need to transition from some tethering tool to rappelling. In all three systems, it is reasonable to tether over the cliff's edge to access the climbing rope. Fixed sections of rope with ABDs make this transition quite easy, because the instructor can rappel over the edge instead of downclimbing while adjusting a klemheist. Once in position over the cliff's edge, pull up your doubled rappel rope, rig your rappel device, and back it up with an autoblock. Once the rappel is correctly rigged and double-checked, the original tethering system may be deconstructed.

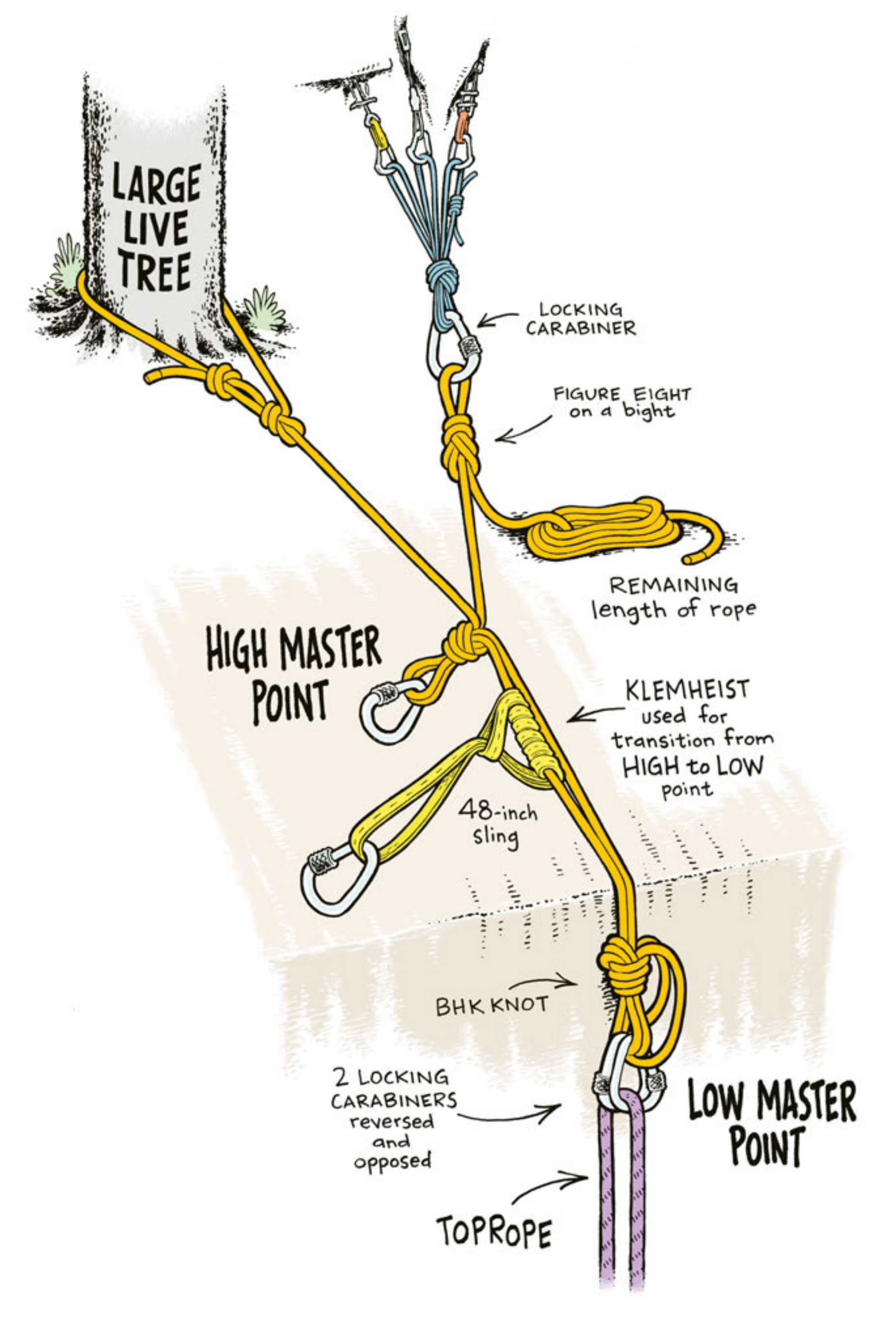
LARGE LIVE TREE
LOCKING CARABINER
FIGURE EIGHT on a bight
REMAINING length of rope
HIGH MASTER POINT
KLEMHEIST used for transition from HIGH to LOW point
48-inch sling
BHK KNOT
2 LOCKING CARABINERS reversed and opposed
LOW MASTER POINT
TOPROPE

The Backside System.

MIKE CLELLAND

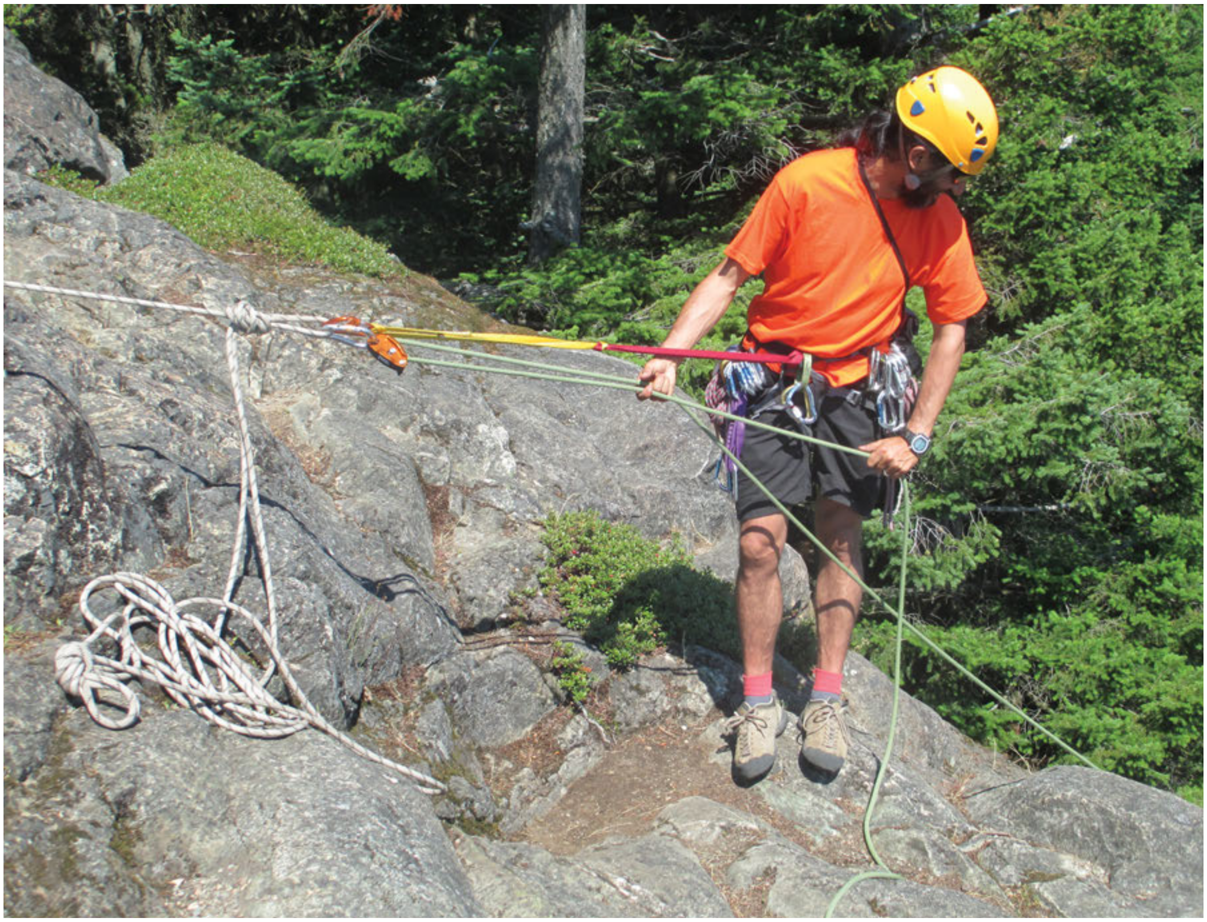

AMGA Single Pitch Instructor Richard Riquelme managing a climber on a Backside System. Note that the pre-equalized strands are uphill and out of the way.

JASON D. MARTIN

Loading the Backside of a BHK

It may strike an instructor as unusual to load the backside of a BHK in this manner. Although this application does not conform to the way instructors have typically seen a BHK in use, it is an acceptable way to load a BHK. It is essentially a double flat overhand. As long as the resultant BHK loop is either in use or adequately long enough, it is highly unlikely that a climbing team could create a load large enough to compromise the backside use of the BHK.

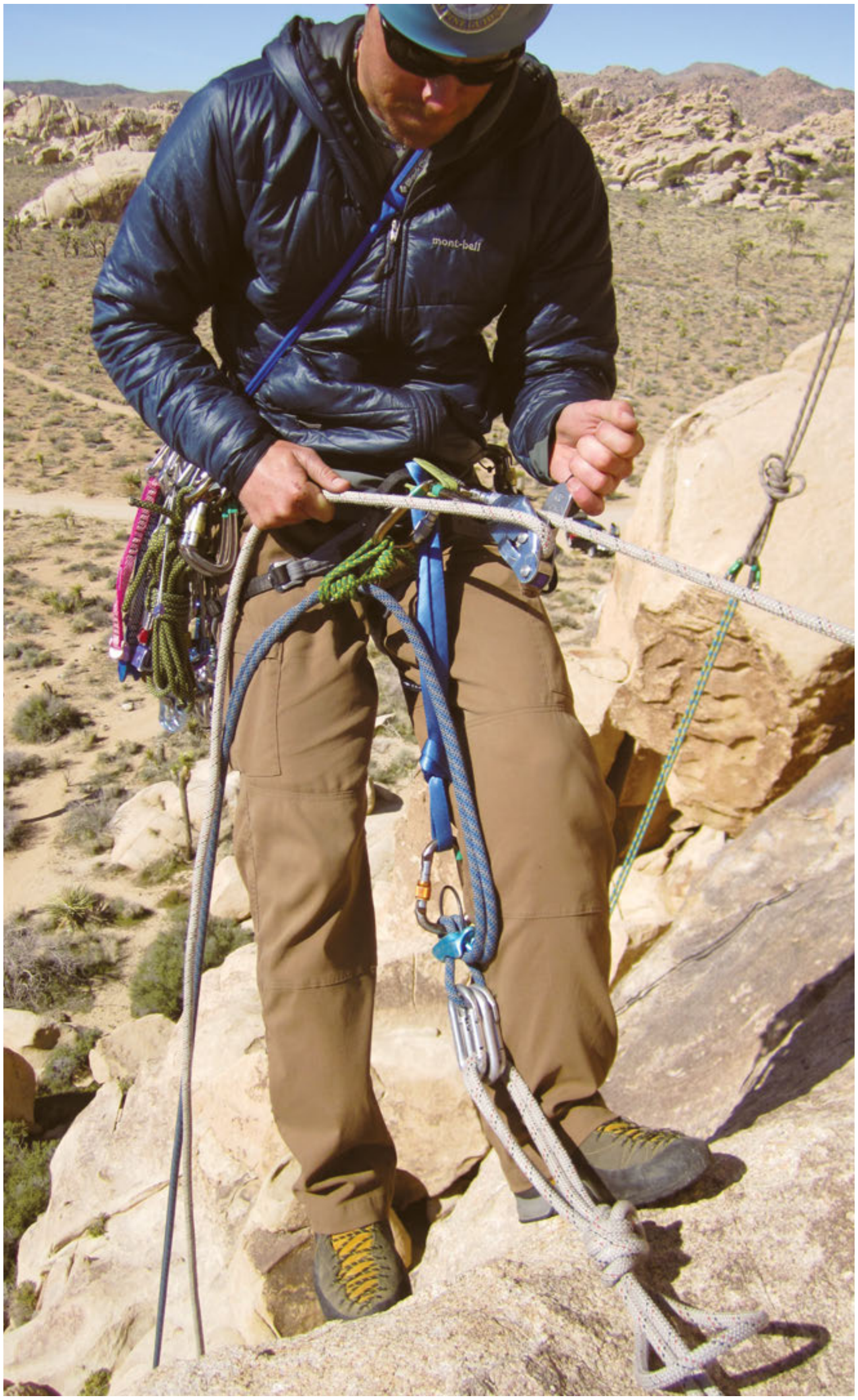

Making the transition on an instructor tether with an ABD. The instructor pre-rigged his rappel device—extended on a doubled nylon sling and backed up with an autoblock knot attached with a locking carabiner to his belay loop. Once he descends on the tether below the master point and weights his pre-rigged system, he can double-check everything before unclipping the Grigri.

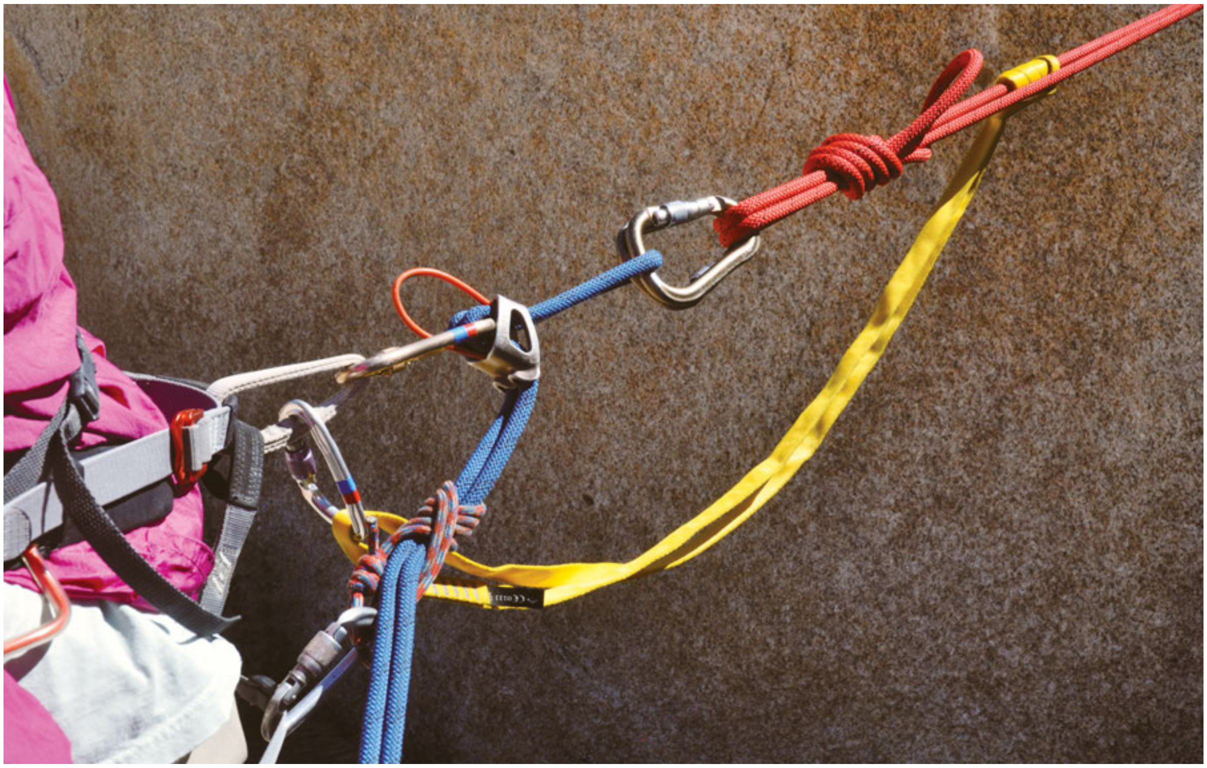

Making a transition with a 48-inch sling attached to one strand of the rigging rope with a klemheist knot. The rappel device is backed up with an autoblock. This technique could be used to transition to a rappel using the Backside or Joshua Tree Systems.

The Three-in-One System

Popularized by Adam Fox during his tenure as the AMGA's SPI Discipline Coordinator, the Three-in-One System was conceived as an efficient rigging system for professional instructors that addresses the needs for instructor security, top-managed scenarios, and easy conversion to a bottom-managed site. Instructors around the country have learned to appreciate its signature elegance, since it manages to anchor, create an instructor tether, and create an extension with the same rope. Three tasks with one rope: Three-in-One.

Even in its simplest configuration, the Three-in-One System will require enough rope to accomplish all these tasks. An instructor can anticipate the amount of rope needed by multiplying the distance between the components and the cliff's edge by three. If the components are farther from the cliff's edge than a third of the setup rope length, the Three-in-One System may not be effective.

Start by building a high master point in much the same sequence as previously discussed:

1. Locate the climb and a distance from the cliff's edge directly above the climb that reduces the risk of falling over the edge.
2. Face the terrain and use raised arms to target potential components (trees, boulders, removable and permanent protection). Let the

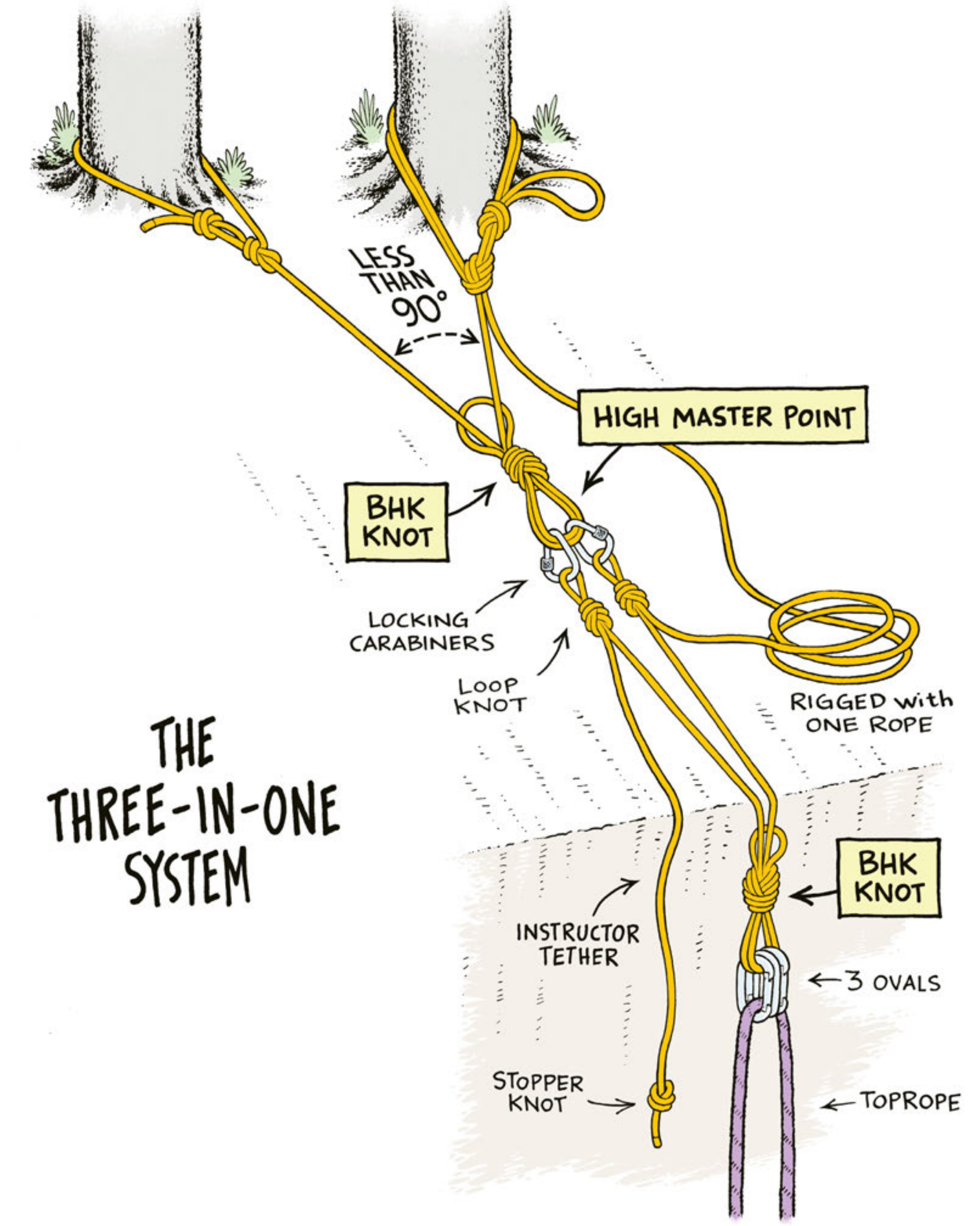
LESS THAN 90°
HIGH MASTER POINT
BHK KNOT
LOCKING CARABINERS
LOOP KNOT
RIGGED with ONE ROPE
THE THREE-IN-ONE SYSTEM
BHK KNOT
INSTRUCTOR TETHER
3 OVALS
STOPPER KNOT
TOPROPE

The Three-in-One System.

MIKE CLELLAND

arms forecast the vectors the setup rope will eventually create to make sure the component selections are not too far apart. Keep in mind that a narrow-angle V configuration (or at least less than 90 degrees) is the goal.

3. Flake out the rope and take the end of the rope out to the most distant component. Strive to use the setup rope alone to attach the component. Bowlines can be very helpful.
4. Leave enough slack at the site of the intended master point to tie a BHK, then take the remaining bight of rope back to the other component. Again, try to attach the bight directly to the component to avoid using some other tool; a bowline with a bight works well for this application.
5. Once both components are connected, a closed U shape awaits the anchor builder at the site of the intended master point. Use this bight of rope to tie a BHK that distributes the load to the components along the anticipated direction of pull (toward the climb).

Anchor Redundancy

In both of the illustrations included in this chapter, the rigging rope is tied off to two different bombproof anchor points. This is not a requirement. The rigging rope could go to two different anchor points, or it could go to a single solid anchor. For example, if you have a giant tree or a solid multi-piece traditional anchor, there is no need to build a second anchor. Both strands of the rigging rope could be attached to the same master point.

Now that the master point is built, use what is left of the setup rope to construct the rest of the system. Start with the instructor tether:

1. Since one end of the setup rope is tied to the most distant component, take the other end and tie a stopper knot.
2. Toss the stopper-knotted tail of rope so it hangs down about 10 feet below the cliff's edge.
3. With this distance measured, affix the rope to the master point using a figure eight loop and a locking carabiner.
4. With the attachment of a Grigri or assisted braking device, this fixed length of rope will be the instructor tether.

Next, an extension can easily be created using the rope that remains.

1. From the backside of the instructor tether, feed out a loop of slack that hangs over the cliff's edge.
2. Close the long loop with another overhand or figure eight loop.
3. Now, using the tether, an instructor can comfortably and securely position themselves at or just over the edge to tie a BHK.

Stepping back, an instructor should see that the entire system—anchor, tether, and extension—has been tied with the same setup rope.

Making the Transition from Rigging to Rappelling

To rappel back to the base, use the instructor tether to facilitate a secure, easy transition to a pre-rigged rappel. Simply pull up the doubled rappel rope, rig a rappel device, back it up with an autoblock, then rappel on the instructor tether with a Grigri until the rappel is tight to the pre-rigged rappel setup. Double-check the rappel device and make sure the autoblock is grabbing, then finally unclip from the Grigri and proceed with the rappel.

CHAPTER 10

Risk Management at the Crag

Every year the American Alpine Club publishes *Accidents in North American Mountaineering,* detailing significant climbing accidents for the year in a comprehensive analysis. Studying what happened to other climbers can heighten an instructor's awareness of what to watch out for to avoid a mishap. John Dill, head of Yosemite's Search and Rescue (SAR) team, studied the most serious climbing accidents that happened in Yosemite Valley from 1970 to 1990. During that time fifty-one climbers died in accidents, 80 percent of them, Dill estimates, "easily preventable." In his article "Staying Alive," Dill points out that state of mind is the key to effective risk management: "It's impossible to know how many climbers were killed by haste or overconfidence, but many survivors will tell you that they somehow lost their good judgment long enough to get hurt. It's a complex subject and

INSTRUCTOR
AREA

Sign at Mountain Warfare Training Center, Leavitt Meadows, California.

Inattentional Blindness

Inattentional blindness is a term psychologists use to describe the neurological phenomenon that occurs when the brain fails to see something obvious, when attention is distracted or focused on something else. Psychologists who study multitasking have found that most of us aren't the multitaskers we think we are: Our brain is simply switching back and forth from one activity to another, deactivating one area of focus to process the other task. In my (BG) role as an examiner for guides' certification exams, I've seen examples of inattentional blindness many times during complicated technical scenarios, where the instructor has moved his focus without seeing an obvious error, like a carabiner unlocked at a key belay or rappel device. It's as if the mind skipped a step, or the brain said that everything was correct and complete, when in fact it wasn't—a cognitive blind spot.

Psychologists theorize that once the brain determines what is important, it fills in the picture with whatever your expectations believe should be there. These failures of awareness happen to all of us, at one time or another, but we're not aware of them, so we don't realize what we've missed! A systematic and routine checklist is helpful, but what we really need to learn to look for is what might be wrong, not what looks right.

Pat Ament, the great Colorado climber of the 1960s and 1970s, writes in his book *Rock Wise:* "As with all of climbing, it is attitude that saves or kills. There is no better beginning than within the mind, in the form of complete concentration. There is no room for oversight or for dismissing what is logically understood. Keen intuition must evaluate all which strikes both mind and eye. Preparing . . . is a process of reasoning and advancing through a regular chain of tests until a synthesis reads go."

sometimes a touchy one. Nevertheless . . . at least three states of mind frequently contribute to accidents: ignorance, casualness, and distraction."

Proper risk management in the single pitch environment involves identifying and assessing hazards, making the right decisions to avoid these hazards, then implementing controls and supervision to minimize the risks.

Rockfall

One environmental hazard is rockfall, which can be naturally occurring (caused by melting ice, wind, etc.) or more likely man-made, caused either by someone pulling off or stepping on a loose hold, or in most cases by a rope being pulled or dragged across the top edge or the face of the cliff. Dropped equipment is also a hazard. Setting up a helmet perimeter zone at the base of the cliff minimizes the danger, as does requiring students to not hang out at the very base of the cliff in the rockfall zone unless they are belaying or climbing. An explanation of the risk is appropriate (i.e., why you need to wear your helmet at the base and while belaying and climbing), as is an explanation of the universal verbal signal for a falling rock or dropped piece of equipment: "ROCK!" Being vigilant is the best strategy, especially when other climbers are directly above or at the top of the cliff.

Terrain Hazards

Another environmental hazard is the terrain itself. If working at the top of the cliff, falling off the edge is a real hazard, as is any steep and exposed terrain involved in scrambling up to the top. In very exposed situations a fixed line can be rigged. See chapter 14, Fixed Lines, for more information.

The base of the cliff also has terrain hazards. The ideal base for group classes would be perfectly flat, but this is often not the case or even possible. Showing students the proper approaches to the climb and belay spots and supervising them is important. Often the most dangerous aspect of taking a large group climbing in a toprope crag setting is their scrambling around on uneven terrain at the base of the cliff, unroped, where they aren't protected by a toprope. The use of ground anchors is important for the belayers in situations where there are chasms and drop-offs adjacent to the belay stances.

Closing the System

In the single pitch environment, the rope system should always be closed. This simple protocol will prevent many accidents during lowering and rappelling. A closed system means that both ends of the rope have a knot in them—either the end is tied into someone's harness, or a stopper knot is tied on a free end. This simple habit prevents the end of the rope from ever going through a belay or rappel device. Always close the system.

Systems Checks

The instructor is responsible for risk management for the entire group. A protocol for systematically checking every climber and belayer before every climb should be standard procedure. A commonly used system is to start with ABC: Check the ground anchor (if used), check the belayer, then check the climber. A proper systems check should be both visual and verbal. The instructor should be close enough to clearly see the harness buckle and rope tie-in. A verbal check is important because it lets the student recognize and learn what to check for. The best method is to have the students themselves go through the checks, with instructor supervision, to get into the habit of a mandatory systems check before every climb.

Anchor

Check the ground anchor to make sure the belayer is in a line between the direction she will be pulled in the event of a fall and the anchor itself. A good way to attach the belayer to the ground anchor is with a clove hitch to a locking carabiner (the end is tied to the front of the belayer's harness with the standard figure eight follow-through knot).

Belayer

Check the belayer's harness, to make sure it is buckled properly. Check the figure eight follow-through knot, to make sure that it is (1) tied properly and (2) threaded through the correct tie-in points at the front of the harness. Check the belayer's belay device, to make sure the rope is properly threaded through the device, and lastly, check the belayer's locking carabiner on the belay device, to make sure that it is locked. Check that the belayer is wearing her helmet.

Climber

Check the climber's harness, to make sure it is buckled properly. Check the climber's figure eight follow-through knot, to make sure that it is (1) tied properly and (2) threaded through the correct tie-in points at the front of the harness. Check to ensure the climber is wearing his helmet.

These checks are simple, but it is surprising how many times students are caught making a mistake somewhere along these lines. The instructor needs to do these checks before every climb. It's that simple. Go through the checks with the students until they learn what to look for, and have the belayer/

climber pair go through the checks themselves with instructor supervision.

New instructors might want to come up with an acronym to help remember what to look for, such as CRASH:

C = carabiner (squeeze check the locking carabiners)

R = rope (check to make sure the rope is not twisted at the top anchor, and check the knots)

A = attitude (check to make sure the climber is ready or has any questions)

S = stuff (check for any extraneous stuff on the climber that should be removed)

H = helmet and harness

Another common one is BARCK:

B = buckles on the harness

A = anchor (check the ground anchor)

R = rope

C = carabiners locked

K = knots

Whatever system is used, know what to check for, and be methodical.

Falling

With a sound toprope anchor, a good belayer, a proper tie-in, and good systems check, the biggest risks the climber faces during the climb are from falling. If the belayer is attentive and there is minimal slack in the system, the fall will be short and uneventful. For novices a demo on how to fall is important, showing them the proper position if a fall does occur: a wide stance with the legs, leaning back to weight the harness, not grabbing onto the rope, extending both arms outward against the rock, not clutching the rock or grabbing handholds.

The falls to guard against and watch out for are falls when the climber is too far to the right or left in relationship to the toprope anchor, which result in a swing or pendulum across the rock face; and falls when there is too much slack in the rope. Remember, if using a dynamic rope with a lot of rope in the system, rope stretch can be substantial (dynamic ropes stretch approximately 10 percent even in a toprope fall). When a climber is directly above the ground or a ledge, take particular care to

A discussion and demonstration on proper falling technique is appropriate for novices—feet wide, hanging in the harness, arms outstretched to brace against the rock, not grabbing the rope.

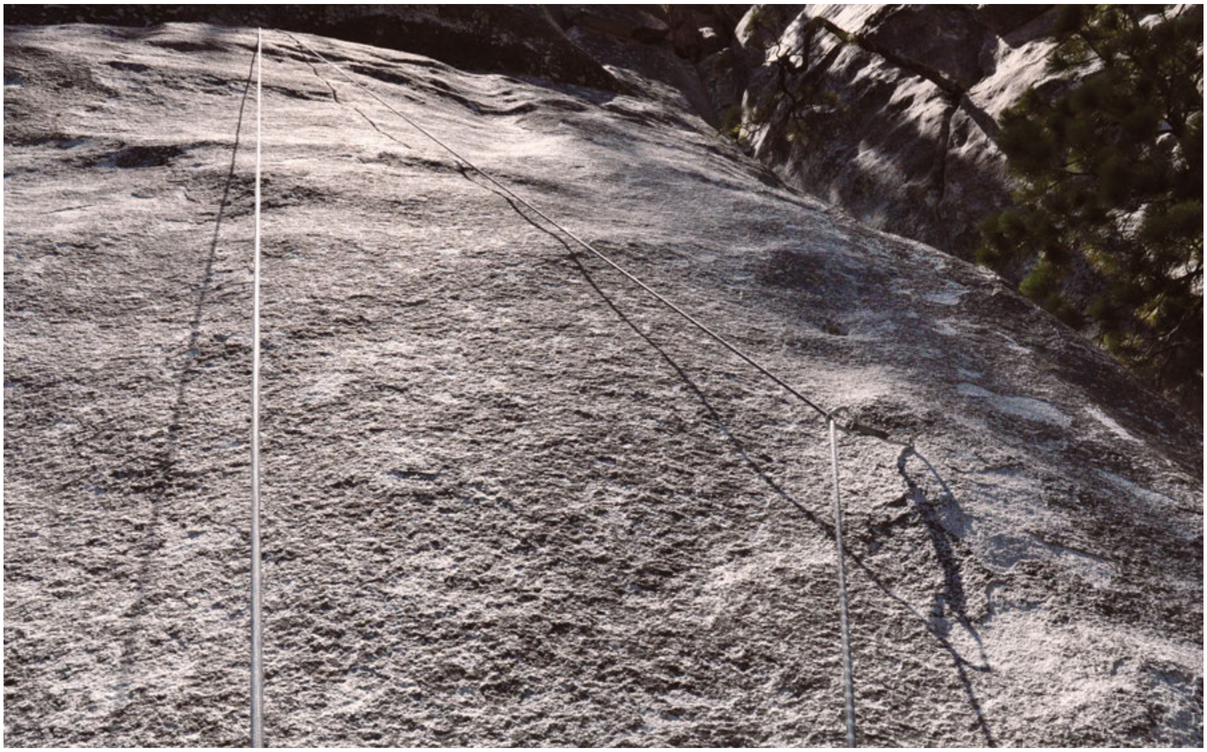

The right side of this toprope setup is clipped into a bolt that will act as a directional to prevent a falling climber from swinging.

ensure the rope is slightly under tension (especially if you are using a dynamic rope), to guard against rope stretch in the event of a fall.

A directional can be used to prevent a swing in the event of a fall. A directional is a separate piece of gear placed below and to the right or left of the main anchor. For example, if the start of the route is 20 feet to the right of the toprope anchor, and the climber falls near the bottom, he will swing 20 feet to the left during the fall (or even hit the ground). A solid piece placed directly in line above the start would be a directional, preventing the swing. When the climber reaches the piece, he simply unclips and continues to the top. As the climber is being lowered, he re-clips into the directional so that the climb is ready for the next climber.

Other Considerations

Climbing sites can be the habitat for things that sting and bite, both flora and fauna. Students should be informed of the hazard and given strategies to avoid an unpleasant encounter.

Weather is another hazard that must be considered. Instructors should check the forecast prior to the work day to ensure they and their students are adequately prepared for whatever weather conditions the day may bring. An excellent resource for regional and spot forecasts anywhere in the United States is the National Weather Service website: www.weather.gov.

CHAPTER 11

Top- or Base-Site Management for a Single Pitch

If an observer took a snapshot of recreational climbing teams around the country, he might begin to think that single pitch climbing has a default pattern. First, most climbers establish the climb; it is either led or the anchor is assembled and the toprope is tossed down from the top. The climber and the belayer are on the ground at the beginning and the end of the climb. Belaying seems to universally happen with the rope having been redirected through the anchor, and the belayer uses his or her own body as a counterweight, belaying directly off the belay loop. After the climb is over, someone has to clean the anchor.

It is inarguable that this single pitch paradigm predominates the climbing community, but instructors should appreciate that just because this arrangement is typical, it is not the only way to work a climb. In fact, many veteran instructors find the opposite arrangement, where the belayer lowers the climber down from the top of the climb, to be the most common arrangement in single pitch outings. For a professional instructor, neither arrangement should become the default setting. Every single outing, every crag, and every student requires a thorough instructor to make a choice: "Shall I work at the top of this climb, or shall I work at the bottom?" The goal of this chapter is to inventory all the options an instructor has when working in a single pitch setting, divulge the tools that are necessary to enable those options, and explore the criteria that an instructor would use to make an appropriate choice for his or her students.

Top-Managed Sites

Working at the top of a climb has many obvious advantages. For one thing, it may be the only way to work the crag. Many climbs are located along sea cliffs and riverbanks, on the rims of a canyon or cave, or at the top of a multi-pitch climbing area. For these kinds of climbing areas, the base of the climb is either inaccessible (hundreds of feet away), or there is water lapping against the bottom of the cliff. In each case, the climbing team will have to approach the top of the climb, build an anchor, lower a climber down, and have the climber climb back to the top.

When the instructor does have a choice, however, between the top and the bottom of a climb, there are still clear advantages to the top.

- The entire climbing team is rarely in danger of rockfall in an area accessed from the top.
- There is 50 percent less rope in the system, compared to a base site, minimizing rope elongation.
- Some climbers like climbing to the instructor.
- Belaying from the top allows an instructor to use an entire rope length when leading, opening longer pitches and huge stretches of terrain.
- The instructor can usually see the anchor system the entire time, monitoring components, sharp edges, and other users.
- An instructor already at the top of the climb can immediately clean the anchor, making the top site a more efficient way to work.
- The top site opens up programming options like rappelling.

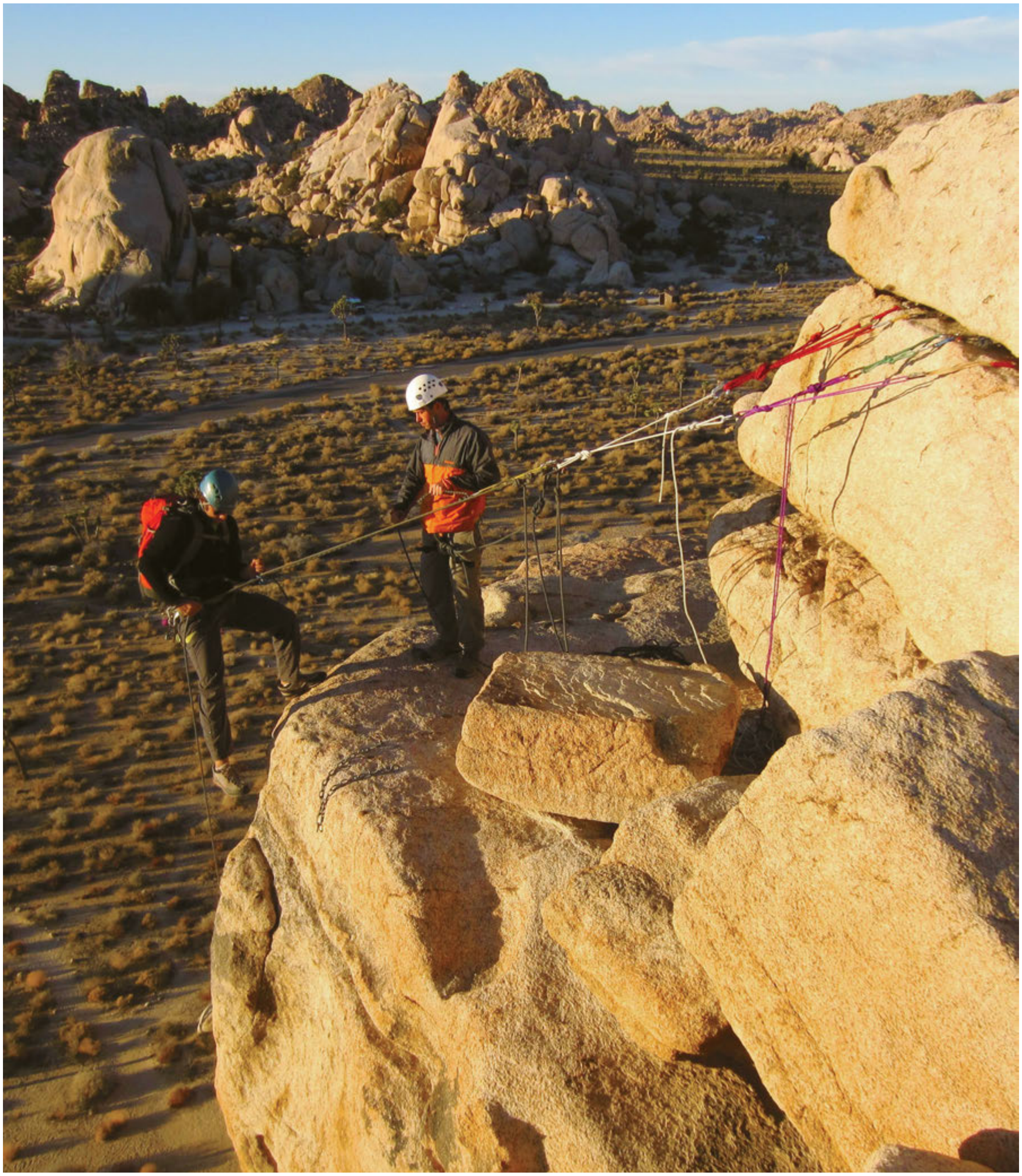

A top-managed site at Joshua Tree National Park, California.

AMGA Single Pitch Instructor James Pierson top-manages a climber from a cliff in Washington's Larrabee State Park. Instructors working on a sea cliff have no choice but to work from the top.

TIM PAGE

There are clearly disadvantages to the top site as well. It is difficult or impossible to see climbers from above on many crags, and coaching individual moves is difficult. The climber's rope always goes over the edge when belaying from above, so sharp edges are particularly troublesome. Many cliff-top ecosystems are fragile and should not be unnecessarily impacted. Lastly, the cliff's edge is always close by, so instructors have to build technical systems to protect themselves and closely supervise participants who may try to venture too close to a precipice.

Because of these advantages and disadvantages in top-site work, instructors may find themselves preferring the top site for low-ratio work, especially in areas where lead climbing is the most common way to access the anchor. If there are only one or two students to manage, it is an efficient way to work.

When establishing a top site, instructors need to select an anchor and a stance that allows them to mitigate risks when approaching the cliff's edge (with a tether of some sort), recognizing that the space between the anchor and the cliff's edge is about to become a busy place. Instructors will be moving back and forth, students will be entering the system, rappels will be rigged, ropes will be stacked, etc. Instructors come to think of this area in front of the anchor as the "workspace," and that is an appropriate way to think of it. As a workspace, it should be clear of all superfluous equipment and personal items. As a workspace, tools should be logically arranged for easy access, or intentionally stowed. As a workspace, the things that happen there should have the tone of work: intentional, careful, and diligent. Students will need to be carefully instructed about where to stand and when and how it's okay to approach the cliff's edge.

Base-Managed Sites

Working from the base of the cliff has distinct advantages: There is no danger of anyone falling off the edge of the cliff, belayers are easily supervised, and climbers can be seen and coached on their way up the route. Recreational climbers will be most familiar with this arrangement, and it is most likely the system that they will default to using on their own, with friends, or in climbing gyms.

But hazards do exist for base-managed sites, such as rockfall and dropped gear. If the climbs are long, there is more rope in the system, so rope stretch can be a big hazard, particularly if a dynamic rope is being used for toproping.

Because of these advantages and disadvantages in base-site work, instructors may find themselves preferring the base site for high-ratio instruction. Groups can cheer on their friends, the outing has a social atmosphere, and multiple climbing teams can be supervised simultaneously. Furthermore, each climbing team can employ up to four student roles, so no one needs to sit idly by, waiting for their turn to do something. For any given climbing rope, one student will climb, one will belay, one can be a backup belayer, and one can serve as a ground anchor.

When should a ground anchor be used? If both the climber and belayer are roughly the same body weight and the terrain at the base of the cliff is flat, a ground anchor is unnecessary. But if the climber outweighs the belayer by more than 40 percent, a ground anchor should always be considered. A severe weight discrepancy will be even more obvious in low-friction setups, like overhangs. But friction, from a climber's body on a low-angled slab, or directional pieces that redirect the toprope, or the rope running over the rock itself, will help mitigate

The Ground Anchor

The worst accident I (BG) have had in thirty-five years of rock climbing happened when I was belaying in a toprope situation without a ground anchor. I was belaying from the base of the climb in a relaxed position about 15 feet out from the base of the cliff. Between the ledge I was belaying from and the cliff itself was a deep chimney. My partner was climbing on a toprope and suddenly fell while attempting an overhang about 50 feet above me. He swung wildly in the air. I easily caught the fall and locked off my belay device in the brake position, but I was pulled off my stance, swinging 15 feet straight into the wall like a pendulum. I braced for the impact with an outstretched leg and sustained a severely sprained ankle. A simple ground anchor would have prevented this accident. Belaying accidents are common, and in almost every case, they have the same element: no ground anchor.

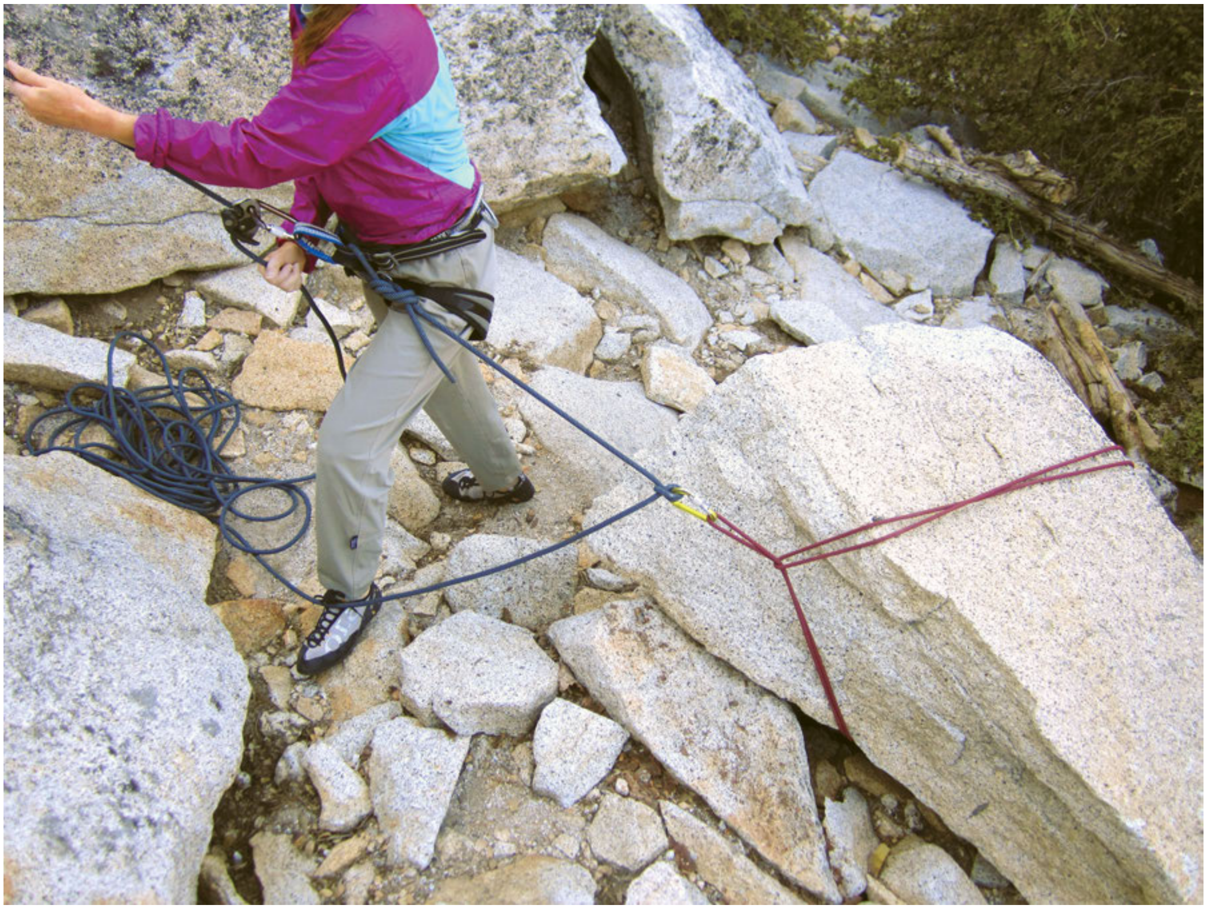

This belayer is using a large block of rock for a ground anchor, ensuring she won't budge if the climber falls.

a severe difference in weight. Ground anchors rarely hurt anything. It is especially important to establish a ground anchor for the belayer in uneven terrain, particularly if the belay stance is perched high on top of boulders, or is some distance away from the base of the cliff.

A good system to rig a ground anchor is to start with the belayer tying in to the end of the rope. Not only does this "close" the rope system, but it allows the belayer to use the climbing rope to connect to a ground anchor with a clove hitch, which can be easily adjusted to suit the stance.

Natural anchors are obvious choices for ground anchors, like a sling or cordelette around a tree or a large block of rock. A single bomber cam or nut in a crack will also suffice. A seated student is an excellent choice for a ground anchor, especially if vegetation has been heavily impacted or other options are not available. The ground anchor student should be given a comfortable seat, like on top of a backpack.

The belayer in a toprope scenario will be pulled in a line directly to the toprope anchor master point, so he or she should be anchored and braced accordingly. Ideally the ground anchor will be low and directly behind or beneath the belayer or just slightly to the side. Remembering the ABCs may be helpful: anchor, belayer, climber. There should be a straight line between the anchor, the belayer, and the direction of pull created by the climber.

Belaying

Whether at the top-managed site or the base-managed site, belaying is a skill performed by students and instructors alike. It doesn't matter who is belaying or what they are belaying with, belaying is a very serious responsibility. It is a skill that instructors must understand, teach, and supervise.

All belaying has three irreducible principles that comprise the grounding theory of belaying; all the techniques discussed in this text will invoke these three key principles.

- The brake hand never leaves the rope.
- The hand transition should be made in a position of strength.
- The body should be positioned comfortably and sustainably.

Even before the popularization of manufactured belay devices, these principles were apparent. From device to device, the fundamental principles never changed. The hip belay was one of the original forms of the belay, and the principles used to make it work also apply to modern belay devices.

BELAY DEVICES

Today there is a wide variety of belay devices available. The most commonly used belay device is a tube or slot device (with two slots so it can be used for both belaying and rappelling on a doubled rope). A bight of rope is threaded through one of the slots in the belay device and clipped into a locking carabiner attached to the belay loop on the front of the belayer's harness. When the two strands of rope (one going to the climber, one to

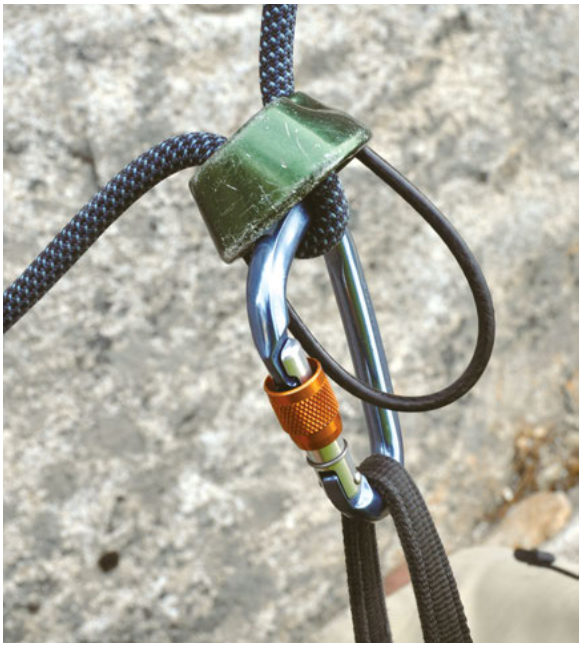

Black Diamond ATC belay device.

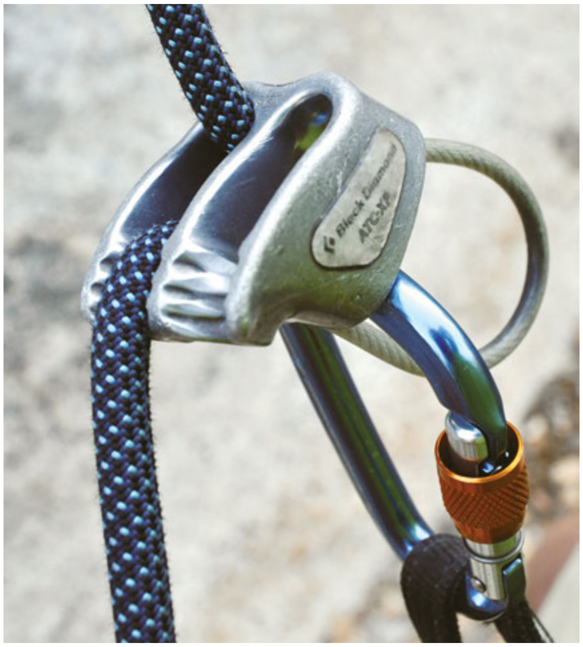

Black Diamond ATC XP in high-friction mode with the teeth on the braking side.

The "Dreaded" Hip Belay

I (BG) began climbing before the advent of the belay device. Back in the 1970s we used the "hip belay" technique to catch a fall, which was simply wrapping the rope around your waist to generate enough friction to stop a fall. Catching a climber on a big "whipper" was painful indeed, and I often ended a hard day of climbing with a black streak singed across the back of my waistline.

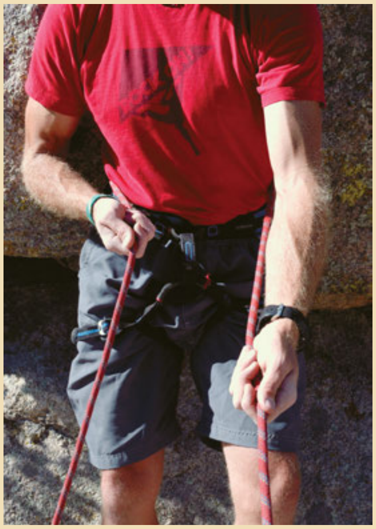

To take in rope, start with the brake hand at your hip and the guide, or "feel," hand extended.

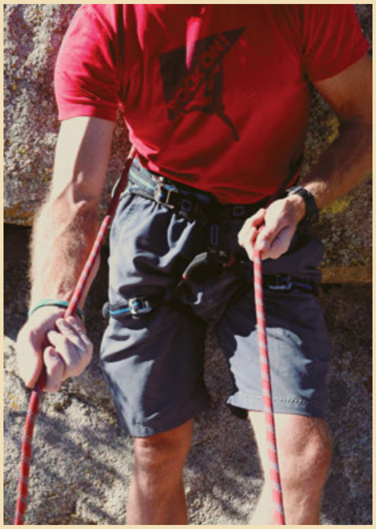

To take up slack, the brake hand goes out as the guide hand comes in.

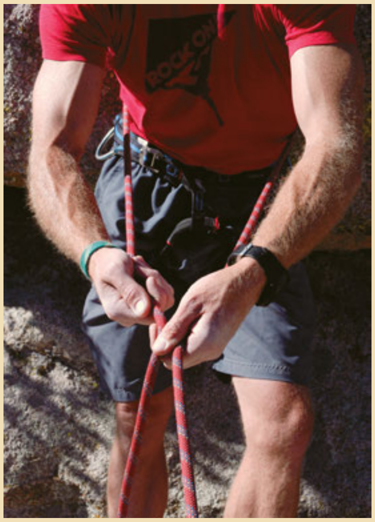

The guide hand reaches above the brake hand and pinches the rope . . .

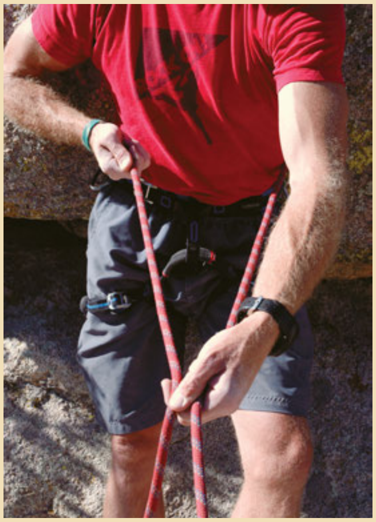

. . . so that the brake hand can slide back.

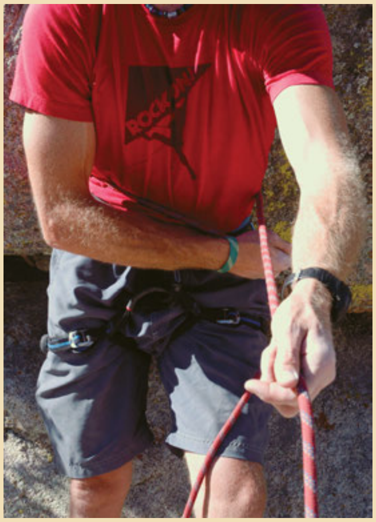

In the event of a fall, the brake hand brings the rope in front of the waist for maximum friction.

Standard Climbing Signals

A methodical systems check along with proper use of the universal climbing signals are integral parts of climbing. Ambiguity in the use of climbing signals has led to many tragic accidents, simply because of lack of communication between the climber and belayer. One infamous tragedy occurred at a popular ice climbing area, in a toprope setup, when the climber reached the top of the climb (at the top of the cliff) and the anchor. The climber yelled, "I'm OK!" but the belayer thought he heard, "Off belay." The belayer unclipped the rope from the belay device and took the climber off belay, thinking he was going to walk off the top. The climber leaned back to be lowered and fell to his death.

On a base-managed toprope climb, it's important to be vigilant at the transition from the climb up to the lower down. This is where most accidents due to improper communication and climbing signals occur. There should be no ambiguity. Many instructors hold onto the strand of rope that goes back down to the belayer until sure that the belayer has heard the command and is in the brake position and ready to hold weight. In most cases the instructor will be within visual contact, so in addition to hearing the verbal commands, the instructor should also look down and visually verify that the belayer is being attentive, with the brake hand in the proper position, alert and ready to lower in a controlled manner. In situations where climbing with other parties around, it's best to include the belayer's name in the signal (e.g., "Off belay, Bob.") to prevent confusion.

Here are the standardized climbing signals used in many climbing schools:

On belay?: Climber to belayer, "Am I on belay?"

Belay on: Belayer to climber, "The belay is on."

Climbing: Climber to belayer, "I'm beginning the climb."

Climb on: Belayer to climber, "Go ahead and start climbing; I have you on belay."

Up rope: Climber to belayer, "There is too much slack in my rope. Take up some of the slack." (Too much slack in the belay rope will mean a longer fall. Remember that rope stretch also contributes to the total distance of a fall, especially when there is a lot of rope out in a toprope scenario.)

Slack: Climber to belayer, "Give me some slack, the rope is too tight."

Tension (or Take): Climber to belayer, "Take all the slack out of the rope and pull it tight; I am going to hang all my body weight on the rope." (This could be a situation where the climber simply wants to rest by hanging in the harness while weighting the rope, or a toprope situation where the climber is getting ready to be lowered back down a climb.) Be careful with the word "take" as it can be misheard as "slack" or "safe."

Tension on (or I've got you): Belayer to climber, "I've taken the rope tight, and my brake hand is now locked off in the brake position, ready to hold all your weight."

Lower me (or Ready to lower): Climber to belayer, "I'm in the lowering position [feet wide, good stance, sitting in the harness, weighting the rope, and leaning back], and I'm ready to be lowered.”

Lowering: Belayer to climber, “I'm proceeding to lower you.”

Off belay: Climber to belayer, “I'm safe. You can unclip the rope from your belay device and take me off belay.” (Never take someone off belay unless this signal has been heard. The universal contract between belayer and climber is that the belayer must never take the climber off belay unless the climber gives the belayer the “off belay” command.)

Belay off: Belayer to climber, “I've unclipped the rope from my belay device and have taken you off belay.”

That's me!: Climber to belayer, “You've taken up all the slack in the rope and the rope is now tight to my harness.”

Watch me!: Climber to belayer, “Heads up! Be attentive with the belay—there is a good chance I'm going to fall right here!”

Falling!: Climber to belayer, “I'm actually falling, go to your brake position and lock off the rope to catch my fall!” (A fall can happen so fast that the climber might not be able to shout this signal during a short fall, but it helps the belayer react more quickly, especially in situations where the belayer can't see the climber.)

ROCK!: Climber to belayer and others below, “I've dislodged a rock and it's now freefalling below me—watch out below!” (The equivalent signal to “fore!” in golf, “ROCK!” should also be yelled when the climber drops a piece of equipment.)

the belayer's brake hand) are held parallel, in front of the belay device, there is little friction, but when the brake strand is held at a 180-degree angle relative to the strand going to the climber, the device affords maximum friction, making it relatively easy to hold the force of a falling climber.

So long as a belay method adheres to the three fundamental principles of belay, there are many techniques acceptable for a sound belay. Along with loyalty to the fundamental principles, good belaying technique effectively manages rope slack, requires maintaining a brake hand on the brake strand side of the rope, and generates enough friction to stop a fall and properly lower a climber.

Over the last decade advances in technology have allowed manufacturers to produce thinner ropes, and belay devices have evolved along with the ropes. When buying a belay device, check the manufacturer's specifications and make sure it's appropriate for the diameter of the climbing ropes being used. The most popular tube device is the Black Diamond ATC (tongue in cheek for Air Traffic Controller), which also comes in a more versatile version with teeth on one side (the ATC XP) that gives the belayer two options: a regular friction mode when the brake strand is on the non-teeth side, or roughly twice the amount of friction when the brake strand is on the teeth side.

Assisted braking devices (ABDs) have become popular for toprope belaying. The most popular models include the Petzl Grigri and the Trango Cinch. These devices have a mechanical lever that

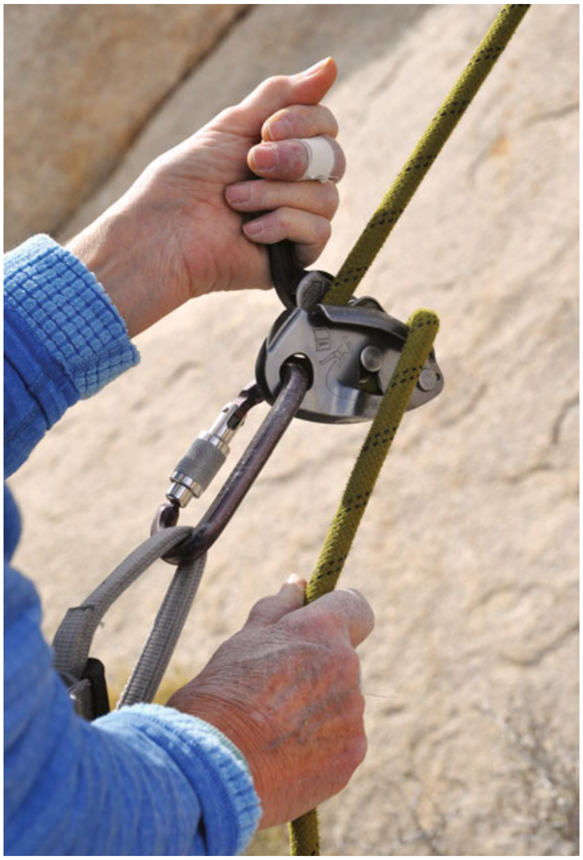

Petzl Grigri in lowering mode.

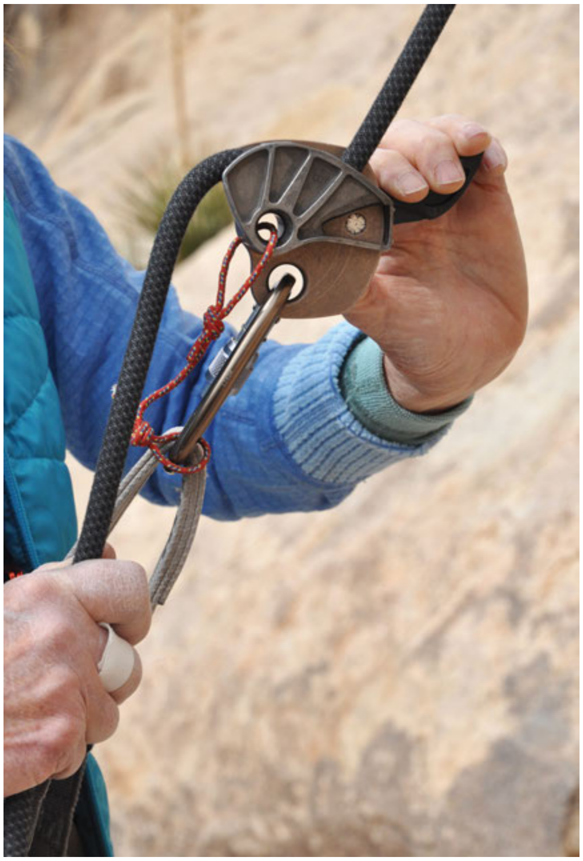

Trango Cinch in lowering mode.

should engage during a fall. Once weighted, the braking lever pinches the brake strand quite hard, so the device has a handle that must be opened to release the tension. When lowering a climber, the belayer must simultaneously manipulate the mechanical lever with the handle with one hand, while maintaining friction and controlling the speed of the lower with the brake hand. These coordinated movements mean that ABDs are far from foolproof, and many accidents have occurred with ABDs, typically when someone is being lowered. The cardinal rule to teach with any belay device is this: Never take your brake hand off the rope! After all, it's the first fundamental principle of belay.

With any belay device, read the manufacturer's guidelines carefully, and seek proper instruction from an AMGA certified instructor if you have any doubts about how to use the device or the proper technique to use when belaying with the device.

Belaying in a Base-Managed Site

Whether using an ATC or an ABD, the following sequence best adheres to the fundamental principles of belay in a base-managed site.

The PBUS (pull, brake, under, slide) method of belaying on a toprope. Start by clipping the rope into the slot in the belay device closest to the spine side of the carabiner and orient the rope so that the brake side is down.

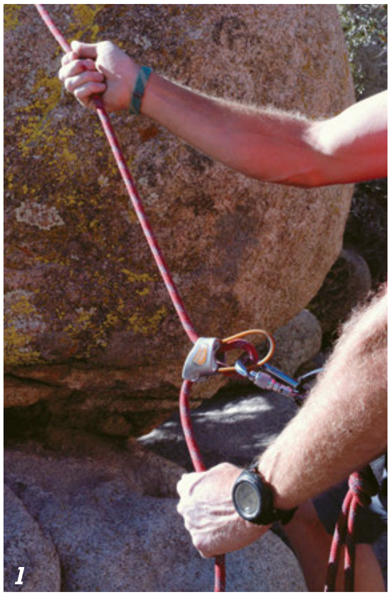

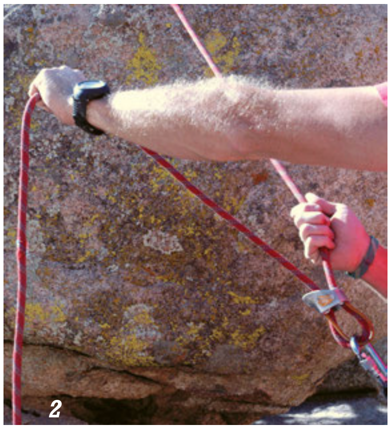

To take up rope, pull the rope up with your brake hand (palm down) as you simultaneously pull the rope down with the other hand . . .

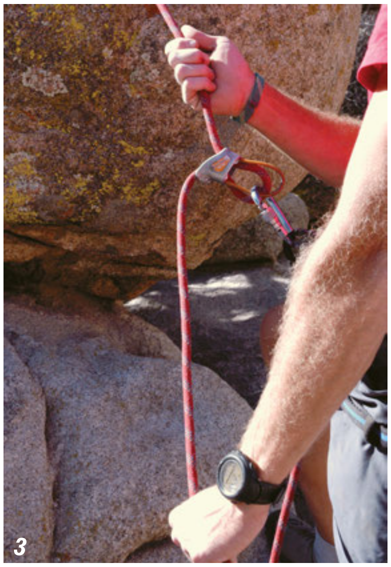

. . . then brake the rope down under your belay device.

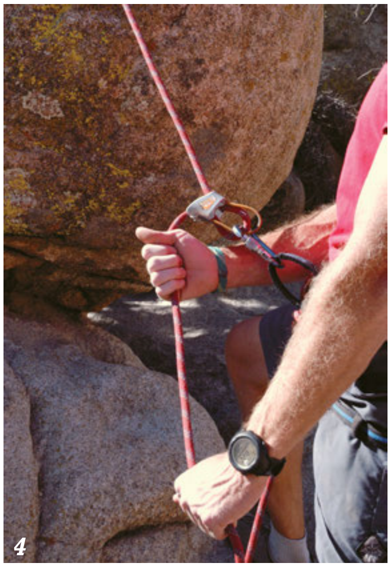

Take your non-brake hand and firmly grasp the rope directly under the belay device . . .

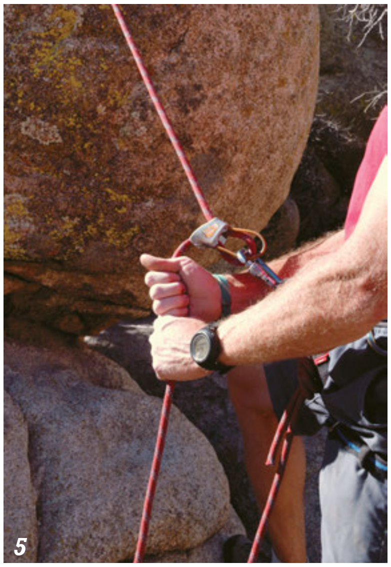

. . . then slide your brake hand up against that hand and repeat the process. This technique is easy to learn and maintains a firm brake position on the rope at all times. In a fall, remember that the brake position is down.

After you've mastered the PBUS method, a more advanced technique is to brake under and switch the brake hand, alternating back and forth with either hand.

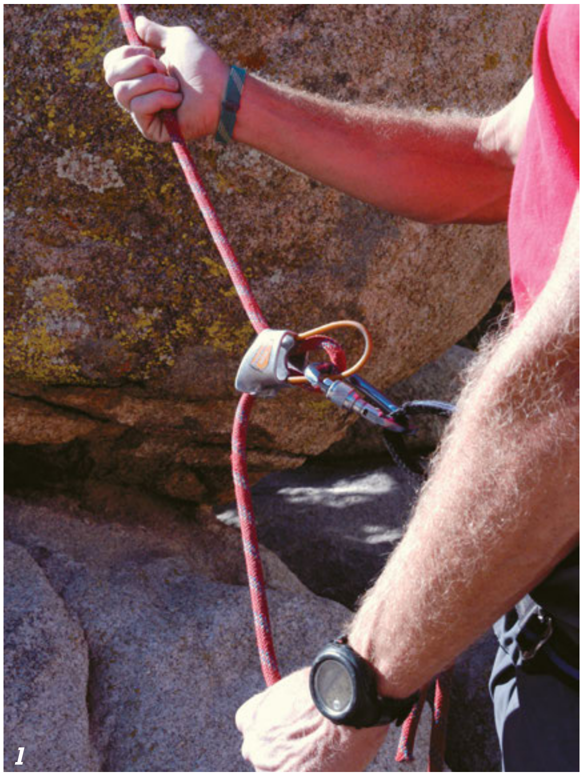
1

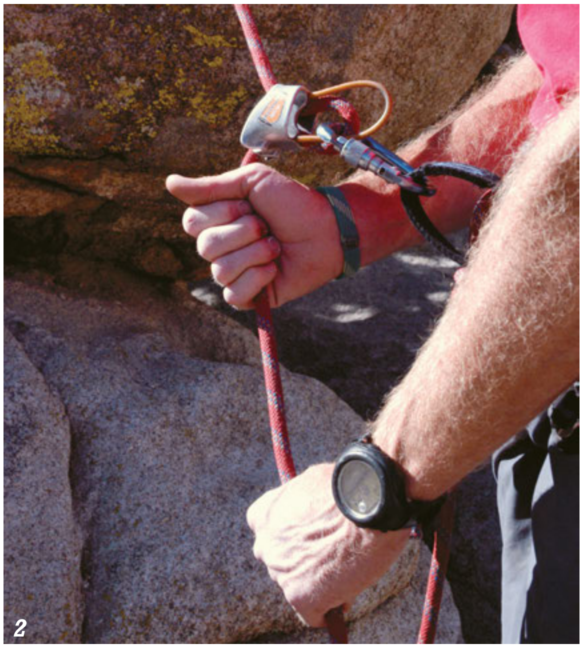
2

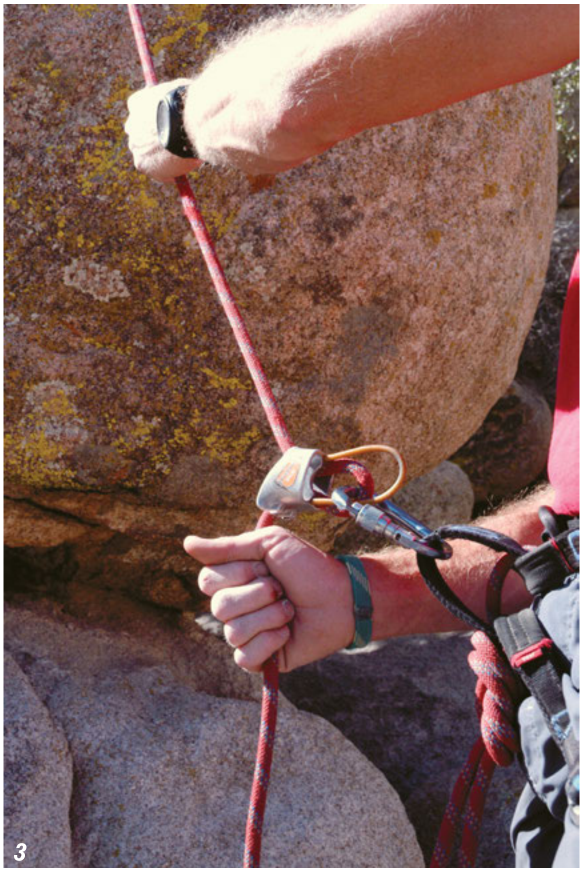
3

Two-Rope Toprope Setups

When rigging long topropes of more than half a single rope length, two ropes can be tied together using knots like the double fisherman's knot or figure eight bend knot. With such a huge amount of rope out between the climber and the belayer, rope stretch is a major concern, especially when using dynamic ropes. Remember that even a short fall in a toprope situation will stretch a dynamic rope about 10 percent, so tighten up the rope when belaying someone just off the ground, or just above a ledge.

Here are two methods that can be used to deal with the knot joining the two ropes. The simple solution, and also the best if there will not be a stance for the climber to stop at, avoids the knot pass altogether. With the knot joining the two ropes at the anchor, tie a figure eight loop and attach it to the climber's belay loop with two locking carabiners (gates opposed and reversed). When the climber reaches the anchor, the knot will be just above the belayer's device, so no knot pass is required.

Another solution is to use two belay devices. The climber ties in to the end of the rope as usual. The belayer anticipates the knot pass and has a second belay device clipped to his belay loop, at the ready. When the knot reaches the belayer, the belayer alerts the climber to find a good stance, then ties a backup knot (overhand loop) on the brake hand side of the belay device. The belayer steps forward to create a bit of slack, then clips the rope into the second belay device on the climber's side of the knot, leaving the first belay device clipped in. If another person is available, that person can assist the belayer simply by holding the rope with both hands under a little tension above the belay device as the belayer accomplishes this. When the climber reaches the anchor, the belayer lowers the climber until the knot is almost to the belay device, and the process is reversed: The climber takes a stance, the belayer unclips the second belay device (first double-checking that the first belay device is still clipped in and has the backup knot), then unties the backup knot and lowers the climber as normal. Knot-passing techniques are something that should be practiced with a qualified instructor.

Backup Belayers

The purpose of the backup belayer is simple: In case the primary belayer loses control of the brake, the backup belayer, holding the brake strand side of the rope, holds the rope to prevent a fall, essentially backing up the belayer's brake hand. For first-time belayers it is appropriate for the instructor to be the backup belayer. For intermediate-level students who have demonstrated proficiency in belaying techniques, other students are commonly used for backup belayers. As an instructor, when you demonstrate a climb with a student belayer, adding another student as a backup belayer gives you an added layer of protection.

To properly back up a belay, the backup belayer should be positioned behind and, if possible, below the belayer's brake hand, so that the backup belayer's hands can activate the brake position of the belayer's device if needed. The backup belayer takes in or feeds out rope as needed, leaving enough slack so that the rope is not being tugged from the belayer's brake hand, or in any way impeding the belayer's ability to manage the belay. If the belayer

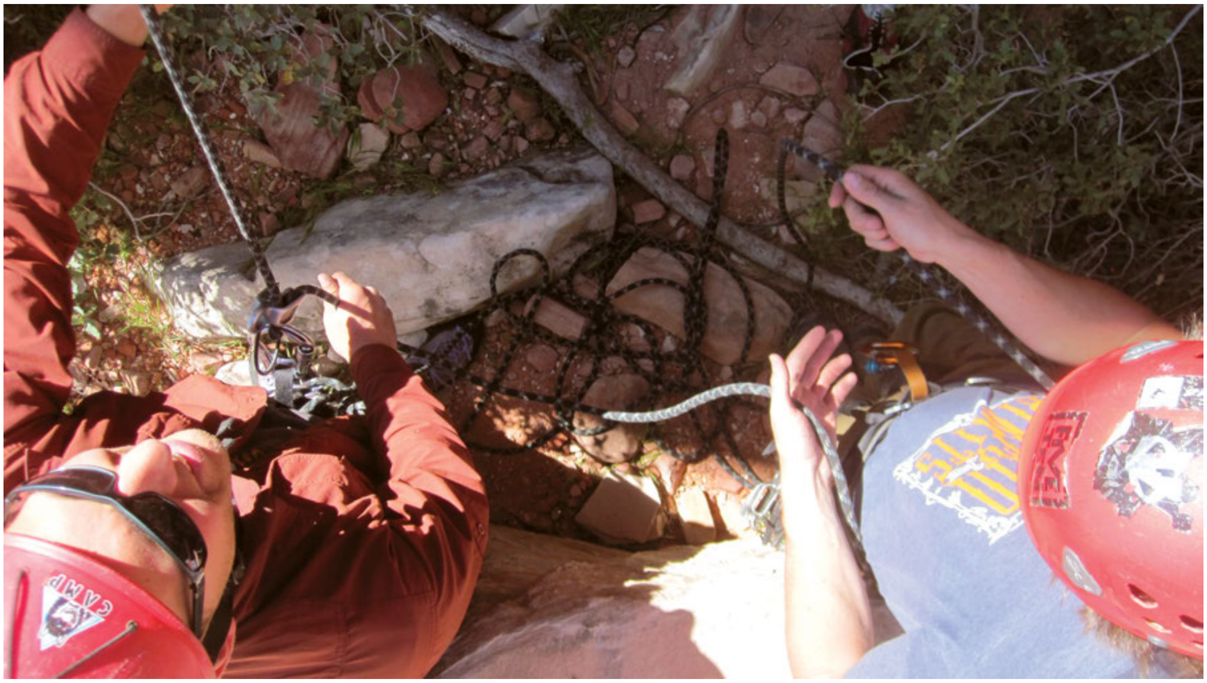

Backup belayer in action. Note that the backup belayer is providing a hip belay for extra security. This may be required in places where an in-line belay backup is difficult to achieve.

JASON D. MARTIN

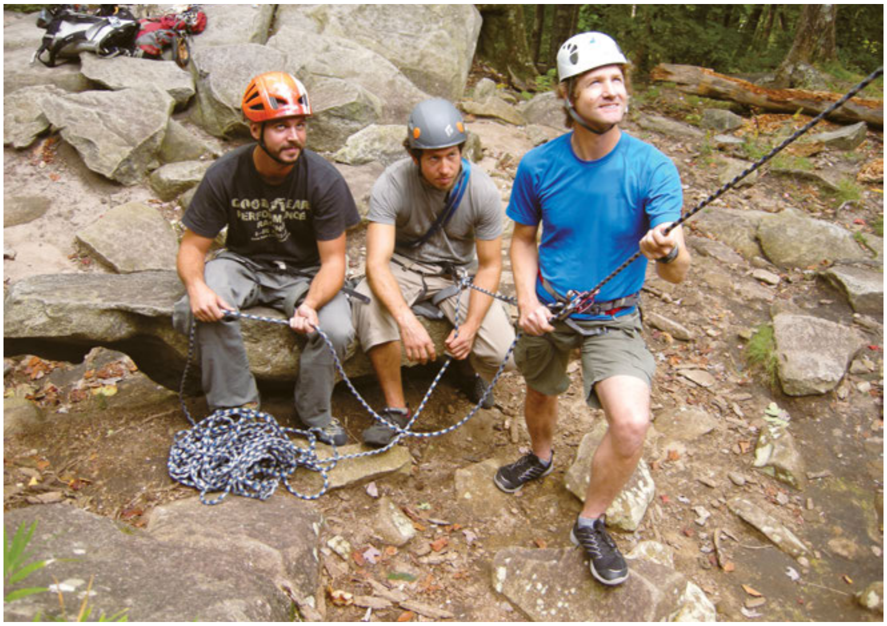

Here a climber provides a "meat anchor" for the belayer. The backup belayer provides an in-line backup belay.

RON FUNDERBURKE

loses control of the brake hand, the backup belayer essentially takes over the braking role. The critical juncture for the backup belayer is when the climber reaches the anchor and transitions into being lowered. The backup belayer can also help the belayer by managing the free end of the rope, preventing any tangles from reaching the belayer.

Some instructors run the rope from the belayer's brake strand through a carabiner on a ground anchor behind the belayer, redirecting the rope back up to a backup belayer. When done properly, this keeps the rope low and in the braking plane, which allows the backup belayer to stand or change locations without compromising the effectiveness of the backup. This redirect technique is also helpful if the instructor is the backup belayer. The instructor can stand in front of the belayer to visually monitor the belayer and climber at the same time while keeping the brake strand in the braking plane.

Another backup technique is the use of a "catastrophe knot." This is simply an overhand knot tied on the rope on the brake strand side of the belayer's belay device. The catastrophe knot can be used if, for some reason, the instructor needs to walk away from the backup belayer role momentarily. If the climber falls and the belayer loses control, the overhand knot will jam in the belay device, preventing a catastrophic fall.

Belaying from the Top

The method preferred by most professional instructors when belaying from the top of the cliff is the direct belay. In a direct belay the belay device is

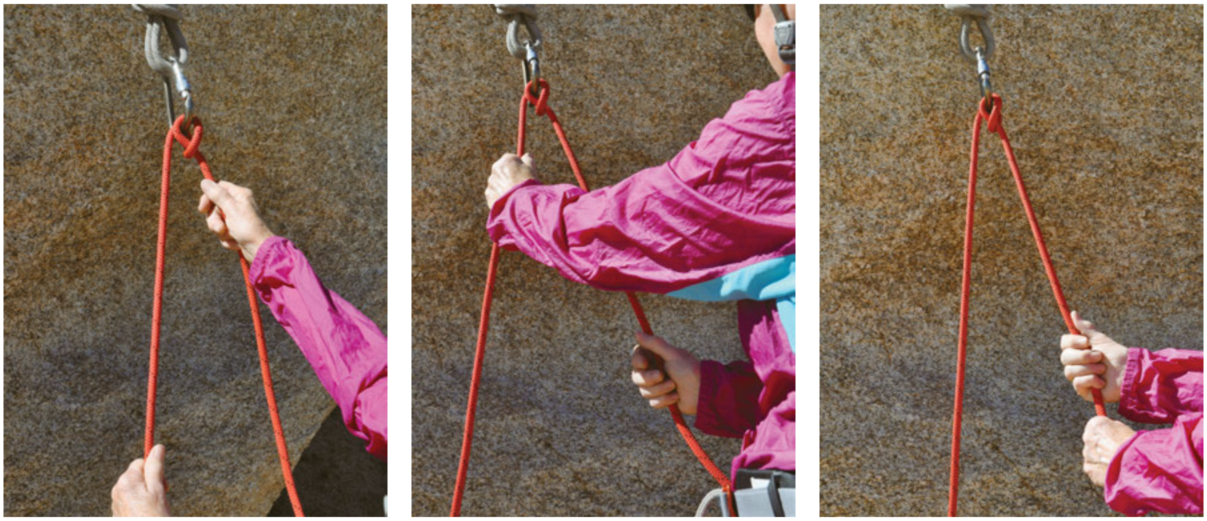

A direct belay with a Munter hitch. In this sequence the belayer's right hand is the brake hand. The starting position (position 1) is with the brake hand at the top and the guide (or feel) hand lower (left photo). Pull the rope down with the brake hand as the guide hand goes up (center). Feeding the rope up toward the carabiner with the guide hand, rather than pulling the rope down tightly with just the brake hand, lessens kinking of the rope. Now take the non-brake hand and grasp the rope below the brake hand (right). This is both the brake position and lowering position. Unlike a tube device, the maximum friction with a Munter hitch is when the two rope strands are parallel to each other. To continue the hand sequence, return to position 1.

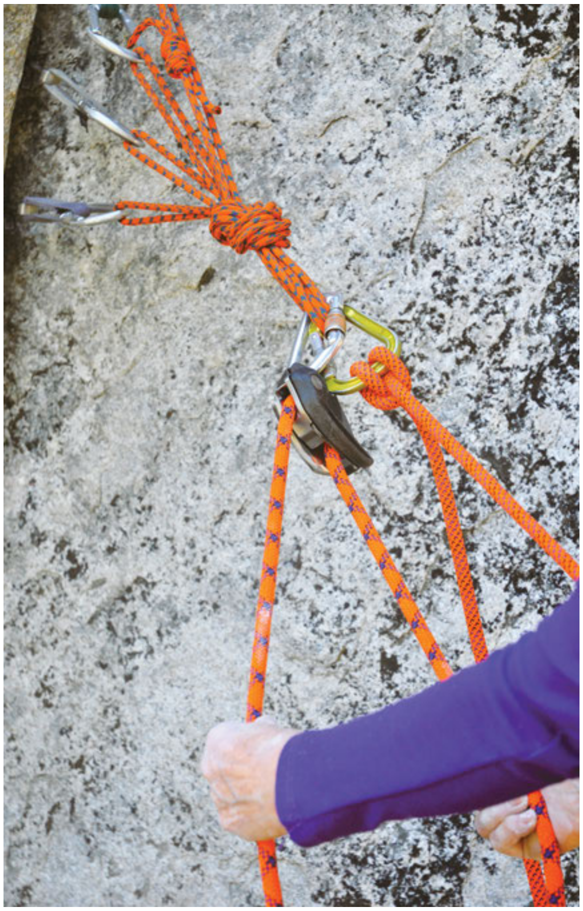

The direct belay. A Grigri is clipped directly to the master point.

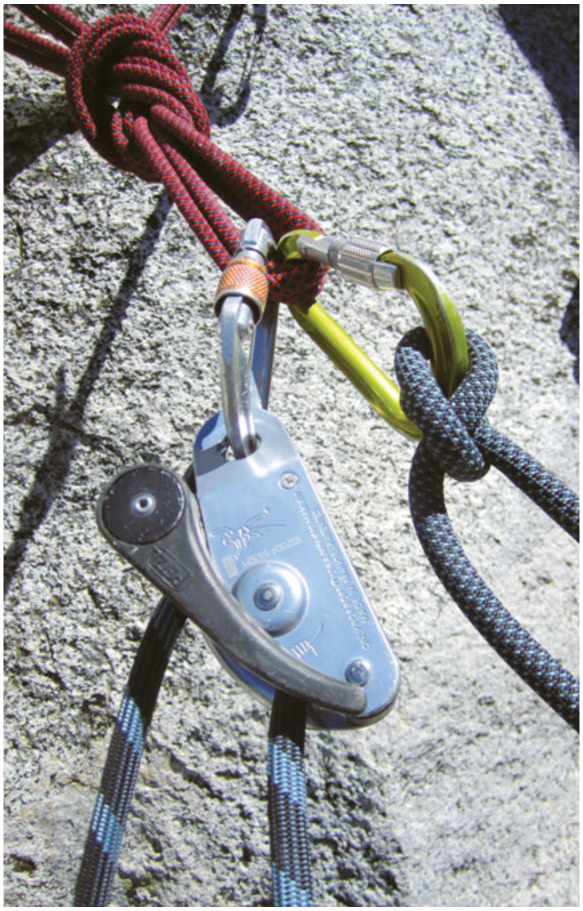

Detail of a direct belay setup. The belayer is anchored with a clove hitch, and the Grigri is clipped directly to the anchor master point.

clipped directly to the anchor, and in the event of a fall, the anchor, not the belayer, bears the brunt of the fall and holds the climber's weight.

Using a standard belaying device, like an ATC, is not recommended for use in a direct belay, because unless the device is positioned below waist level, the braking position will be very awkward, and the instructor will be in a weak and dangerous position to hold a fall. If the master point is above waist level, the Munter hitch works well, since the braking position for maximum friction is when the two strands of rope are parallel to each other, with the brake position down below the carabiner, not above it. Belaying with a Munter also allows instructors the option of positioning themselves some distance away from the Munter belay carabiner due to the maximum braking power achieved with both rope strands parallel to each other.

A commonly used method for a direct belay is to use an assisted braking device (e.g., Petzl Grigri

Not for Instructors

Recreational climbers tend to use two systems that have limited applications for professional instructors. The first is the indirect belay, which is a fancy way of saying a belay directly off the harness. The second is the redirected belay, a technique where the rope runs from the belayer at the top through carabiners at the anchor and then back down to the climber.

The indirect belay is considered to be indirect because the force of a fall is not directly placed on the anchor. Instead, the belayer's body takes a portion of the force, which depending on the position and stance could be a large portion or a small portion.

An indirect belay is a poor choice for an instructor because you are "in the system," frozen in the position that you set up. This makes it difficult to see and coach your climber. Additionally, if something goes wrong, it is difficult, technical, and time consuming to escape the belay.

A redirected belay is a similar technique, but instead of running directly to the climber's belay loop, the rope is first redirected through carabiners in the anchor. The problems with the redirected belay are similar to those with the indirect belay. The instructor can become stuck in the system, unable to coach a climber below, and this belay also is difficult to escape in the event of an emergency.

In addition to the preceding, there are two other problems implicit in a redirected belay. First, when a heavy climber lowers, the belayer can be pulled up into the system. And second, this system puts more force on the anchor.

A redirected belay puts more force on the anchor by creating a 2:1 pulley system at the master point. Such a system requires an equal amount of weight on both strands of rope to remain static. In other words, if a 200-pound person falls on a redirected belay, you will have to counteract that weight with an additional 200 pounds, placing 400 pounds directly onto the anchor.

While some might argue that there are legitimate reasons to use an indirect or a redirect belay, there are few legitimate reasons to use these techniques in a single pitch instructional scenario.

or Trango Cinch) or autoblocking device (e.g., ATC Guide or Petzl Reverso) clipped directly to the master point of the anchor.

The advantage of a Grigri or similar device is that in the event of a fall, the Grigri simply locks off, and the anchor holds the climber's weight. When using a Grigri in the direct belay mode, take care when the device is close to the rock, as anything that presses against the handle (i.e., the rock) will release the braking mechanism. Be sure to position the handle away from the rock.

Remember, a regular ATC or similar non-autoblocking device is not recommended for use in a direct belay if the master point is at or above the belayer's waist level, as the braking position would be very awkward.

Lowering from a Top-Managed Site

There are several lowering methods that are quick and easy to rig. Before lowering someone, make sure the system is closed and the rope is properly flaked at the instructor's feet. Climbing instructors should choose an ergonomic stance that will not put them in an awkward position, and will allow visual contact with students all the way to the ground. If the anchor is not close enough to the edge to allow this, a tether or anchor extension should be incorporated into the system.

Lowering with a Munter Hitch

The rope is flaked at the instructor's feet, so if there are any tangles, they can be handily dealt with. Note that the tail of the rope sitting on top of the rope stack will soon be tied to a student. Therefore, wherever that rope goes, the student will have to follow that same path while approaching the cliff's edge to be lowered. Make sure the student has a clear avenue to the cliff's edge once on belay. Toss the rope to the student and instruct them to tie in. Next, use a locking carabiner designed for use with a Munter hitch on the anchor. Back it up with an autoblock clipped to a locking carabiner at the instructor's belay loop, and begin formal belay commands with the student. One advantage of the Munter hitch for lowering is that an instructor does not have to be at the Munter carabiner to effectively manage the braking position, since the optimal friction is when both rope strands are parallel to each other. This allows an instructor some flexibility; he or she can tether as far away from the Munter as desired while maintaining an effective belay.

Lowering with an Assisted Braking Device

When given the option, many Single Pitch Instructors will lower a student using a Grigri. Since it has a built-in braking system, there is no need to back it up with an autoblock. Petzl actually sells a carabiner, called the Freino, that has a special gate on the side of the carabiner for the brake strand to be clipped into, to facilitate lowering. Without the special carabiner, an instructor can redirect the brake strand back up through a separate carabiner clipped to the master point (or, on a cordelette anchor, up to the shelf). The big advantage of the Grigri is that once the rope is clipped in, it can be used for lowering (just remember to redirect the brake strand!) or belaying (as the climber climbs back up), and it's all set to rig a 3:1 hauling system if a climber needs some assistance on the way back up.

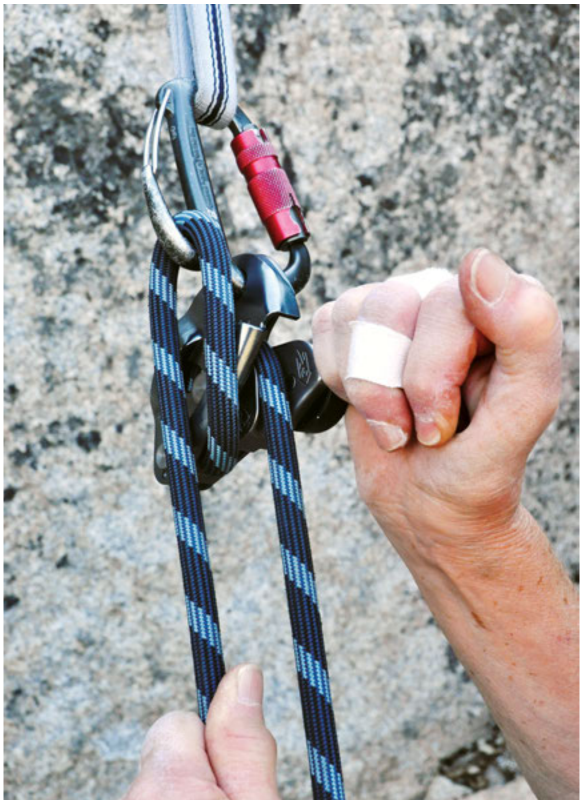

A Grigri in lowering mode with the brake strand redirected using the Petzl Freino carabiner, which is specifically designed for this application.

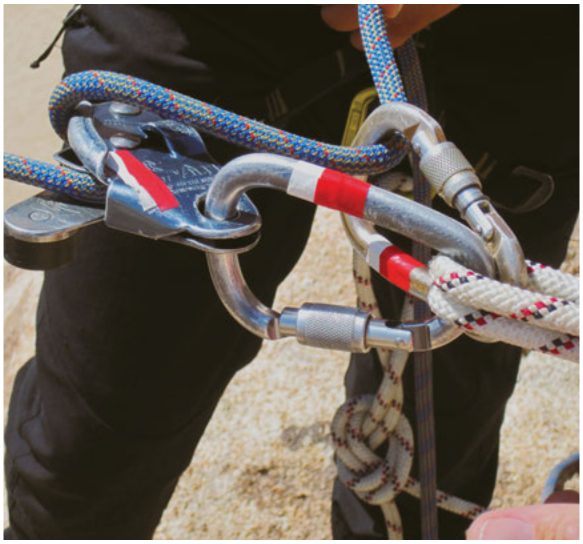

A Grigri rigged for lowering with the brake strand redirected through a carabiner clipped to the master point.

Using an Autoblock Knot as a Backup When Lowering

Whenever lowering a student with a belay device other than an assisted braking device, it's best to back up the brake hand with an autoblock knot clipped to a locking carabiner attached to the instructor's belay loop.

Some instructors call the autoblock the "third hand," because if the brake hand releases when lowering or rappelling, the autoblock grabs the brake strand of the rope. The autoblock adds an extra level of security, especially when there are tangles in the rope that require untangling.

If the stance at the anchor does not allow a clear field of vision down the cliff, to watch climbers while lowering, then instructors can rig the rope direct belay system to position themselves at the edge to maintain visual contact. Instructors should always strive for visual contact with their students whenever belaying or lowering.

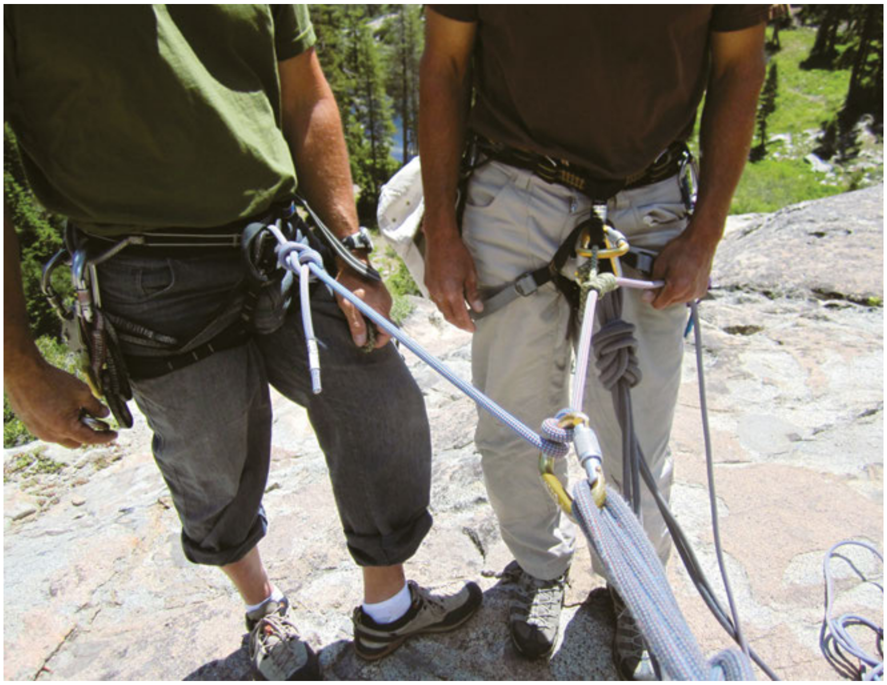

The belayer is ready to lower the climber with a Munter hitch on a locking carabiner clipped to the master point, backed up with an autoblock clipped to a locking carabiner attached to the belayer's belay loop.

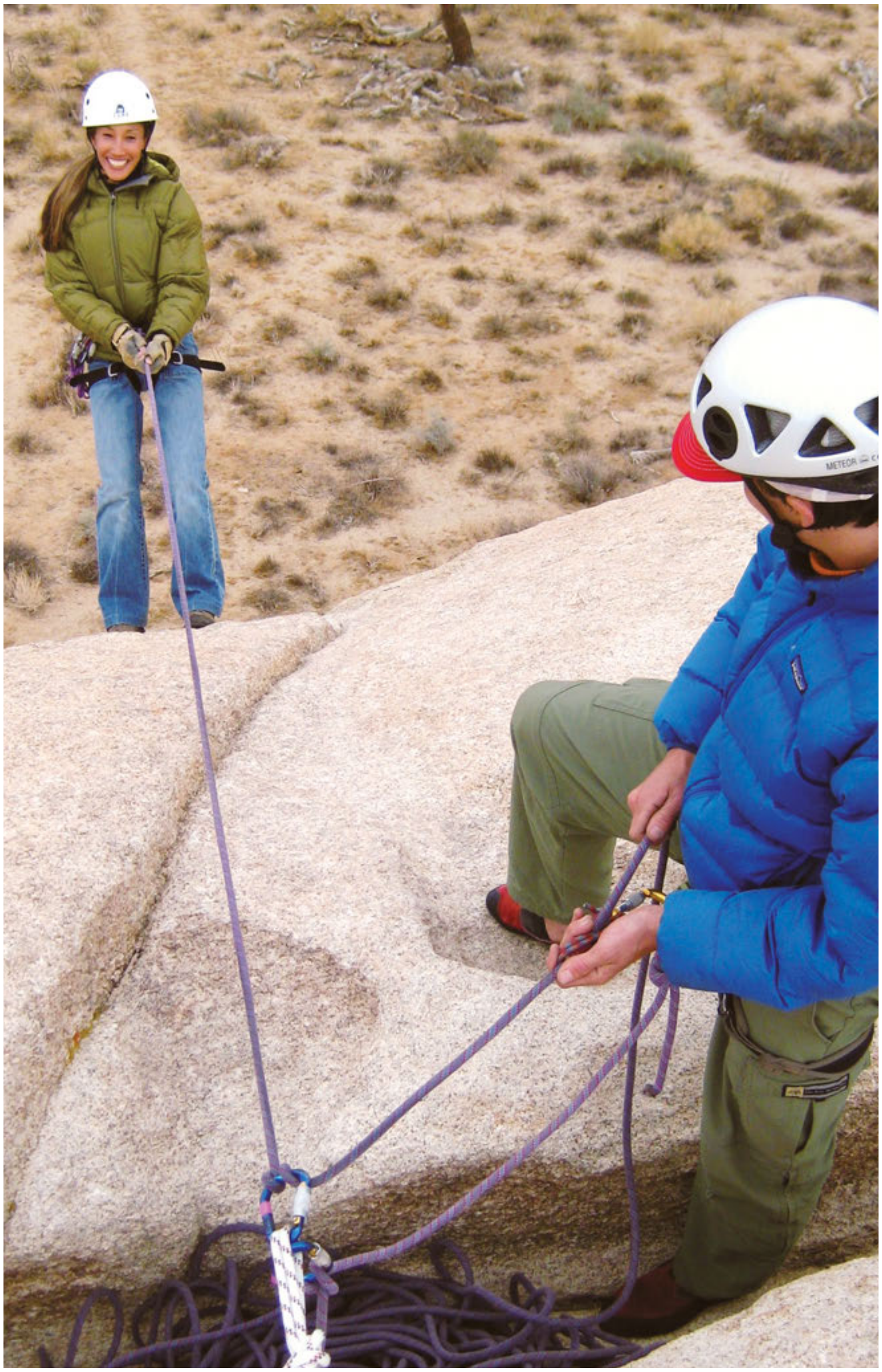

Lowering with a Munter hitch and an autoblock. The belayer is holding an autoblock backup that's clipped with a locking carabiner to his belay loop.

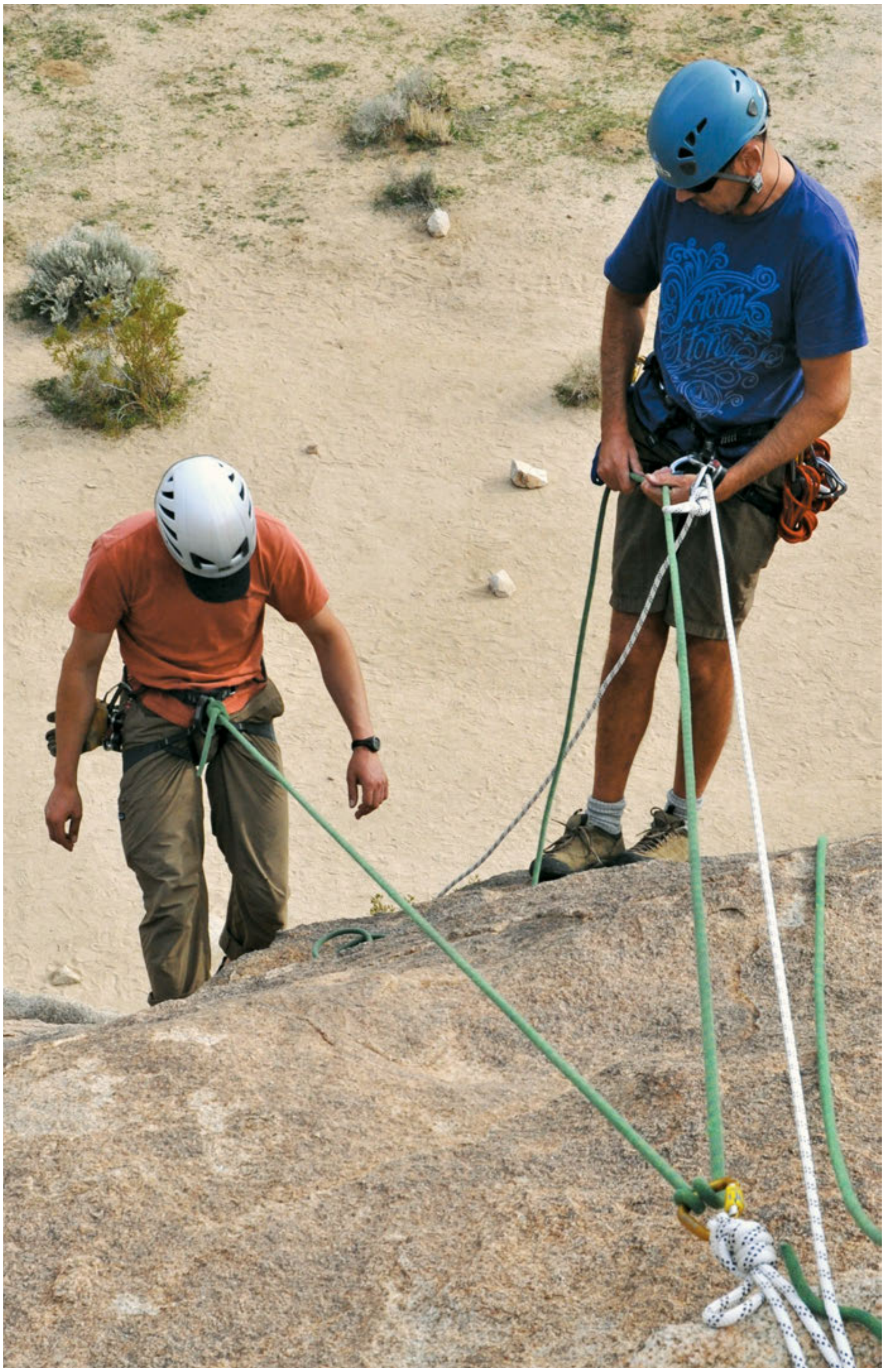

Here the instructor is positioned at the cliff's edge to maintain visual contact with the climber being lowered using a munter hitch at the anchor's BHK master point. The instructor is using a Grigri on the white fixed line (instructor tether) to allow for adjustable positioning. Note the backup knot tied on the fixed rope on the brake strand side of the Grigri. The instructor's brake hand on the lowering rope is backed up with an autoblock knot attached to his belay loop.

CHAPTER 12

Rappelling

The first question the instructor needs to answer is "why rappel?" Is there an easier way to descend from the top of the cliff? Would lowering be more efficient? Can the risks be properly managed? The most common reasons for rappelling are to teach novices the fundamental skills of rappelling or to provide students with an experience. Every recreational climber needs to know the fundamentals of rappelling; it is one of the cornerstone skills of technical rock climbing. Additionally, rappelling can be an impactful activity. It is an activity where students take complete responsibility for managing their own risk. Properly facilitated, it can be a great tool for building self-confidence and generating a sense of accomplishment. The instructor should have clear goals for conducting the activity and should avoid treating rappelling like an amusement ride.

Teaching Rappelling

Site Selection

Site selection is important to teach a novice rappelling technique. Select a site that has a comfortable, flat area on top without a drastic transition from the horizontal to the vertical. The ideal site should have a high master point for the rappel and belay anchor and a flat "desktop" area that extends at least 10 feet back from the edge of the cliff, be free of loose rocks, and have a rounded, gradual edge for a comfortable transition to a wall that is angled slightly less than vertical. The site should also provide a good stance for the instructor, where he or she can coach and model while maintaining visual and verbal contact with the student during the entire rappel. The cliff face should be free of loose rock and uniformly flat, not riddled with cracks, and without corners or overhangs. The base should have a nice flat landing area without boulders, bushes, or trees.

Rigging for Contingencies

In the institutional setting all students are belayed when they are learning to rappel. A good instructional system anticipates any potential problems that may occur during a rappel and uses a rigging system that is ready to remedy any problem.

Student Rappel Belays

Belaying students while rappelling can be done from the base or from the top of the cliff. The technique used at the base is called the fireman's belay. This technique requires the instructor to be at the bottom of the rappel. Holding the rappel rope, the instructor can stop the rappeller's progress by forcefully pulling down on the rope. When the instructor pulls down on the rope and applies tension, the rappeller will stop—it's not possible to move down the rope when it is under tension. This technique requires the instructor to be on high alert, prepared to arrest a fall the moment it happens. Failure to react quickly can lead to loss of control of the belay, with catastrophic consequences. The fireman's belay should be limited to situations where there are two instructors present: one to perform the belay and one to monitor and coach the students on top. It is

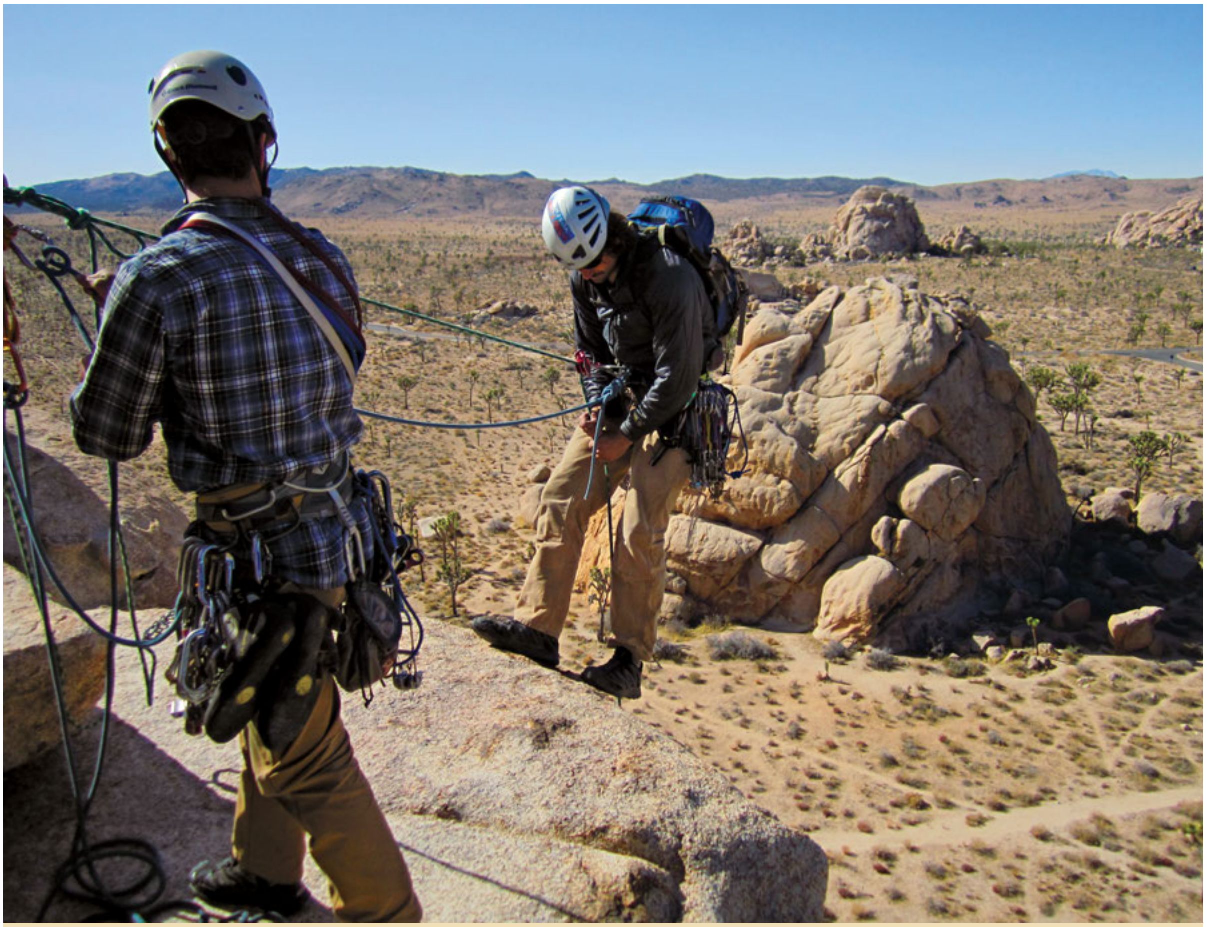

Instructor-belayed rappel rigged from a high master point.

Elevators

Care needs to be taken to avoid having the master point pinned to the ground once the student has weighted the system and gone over the edge. A low master point has two disadvantages. It makes the system difficult to manage for the instructor, and it makes the transition over the edge far more demanding for the novice rappeller. The SAR community uses a ''high redirect'' (aka an elevator) to deal with this; it's also an elegant solution to the problem for the climbing instructor. A simple way to create an elevator is to use a pack to prop up the master point; however, the use of trees or terrain features are a much more effective way. With a little creativity there is no reason to grovel in the dirt while managing a belayed rappel.

not a recommended belay technique when dealing with novices or when facilitating an experience. Assisting or rescuing a student in mid-rappel is highly problematic with this technique.

The preferred method for students who will be rappelling for the first time is to provide a belay from above. The system can be rigged using one or two ropes. One line is the rappel line; the other is the belay line. If the length of the rappel is less than half the rope length, the system can be rigged with a single rope. One end is tied to the belayer, and the other end goes to the ground and includes a stopper knot. If the rappel is more than half a rope length, two ropes are required. The rappel rope is attached to the anchor with a Munter/mule combination, which is releasable under tension. The rappeller is tied into the belay rope and belayed on a direct belay with a Munter hitch off the anchor. The belay rope is neatly stacked at the instructor's feet.

It is recommended that an assisted braking device *not* be used in this situation, for a couple of reasons. First, it is difficult to manage the ABD and maintain visual contact with the rappeller. Second, and more importantly, it is nearly impossible to provide slack to the belay line smoothly as the rappeller progresses. Under this circumstance, the student isn't actually rappelling; it becomes more of a lower.

Sometimes the rappeller encounters a problem (like long hair or clothing stuck in the rappel device, or a knot in the rappel rope) and can't continue. As always, seek the nontechnical solution first. If one cannot be found, fix the problem by doing the following:

1. Tie off the Munter on the belay line with a mule knot and overhand backup.
2. Release the mule knot on the rappel line and give the rappeller some slack.
3. Have the rappeller fix the problem (e.g., take the hair out of the device).
4. Re-tie the mule knot with an overhand backup.
5. Release the mule knot on the belay line and continue belaying the rappel.

The Instructor Rappel Backup

Instructors should manage personal risk by using a backup when rappelling. In the old days the most common method for a rappel backup was using a prusik knot on the ropes above the rappel device, connected to the harness with a sling. The nonbrake hand would cup the prusik knot and hold it in a loosened position during the rappel, allowing it to slide down the rope. Letting go of the knot allowed it to slide up and grab onto the ropes,

Extended Master Point for Belay Strand

The belay line on a releasable rappel should always be extended so that the carabiners don't pinch the rope between one another. One way to do this is to use a runner basketed through the master point. By tying an overhand knot in the basketed sling, two additional redundant master points are created. Another way is to mule off the rappel line on one locking carabiner, while belaying the student off two enchained locking carabiners. And yet another way is to use a double loop figure eight knot, making each loop a different length, provided an appropriate diameter anchor rope is used and that it is not placed on an abrasive surface.

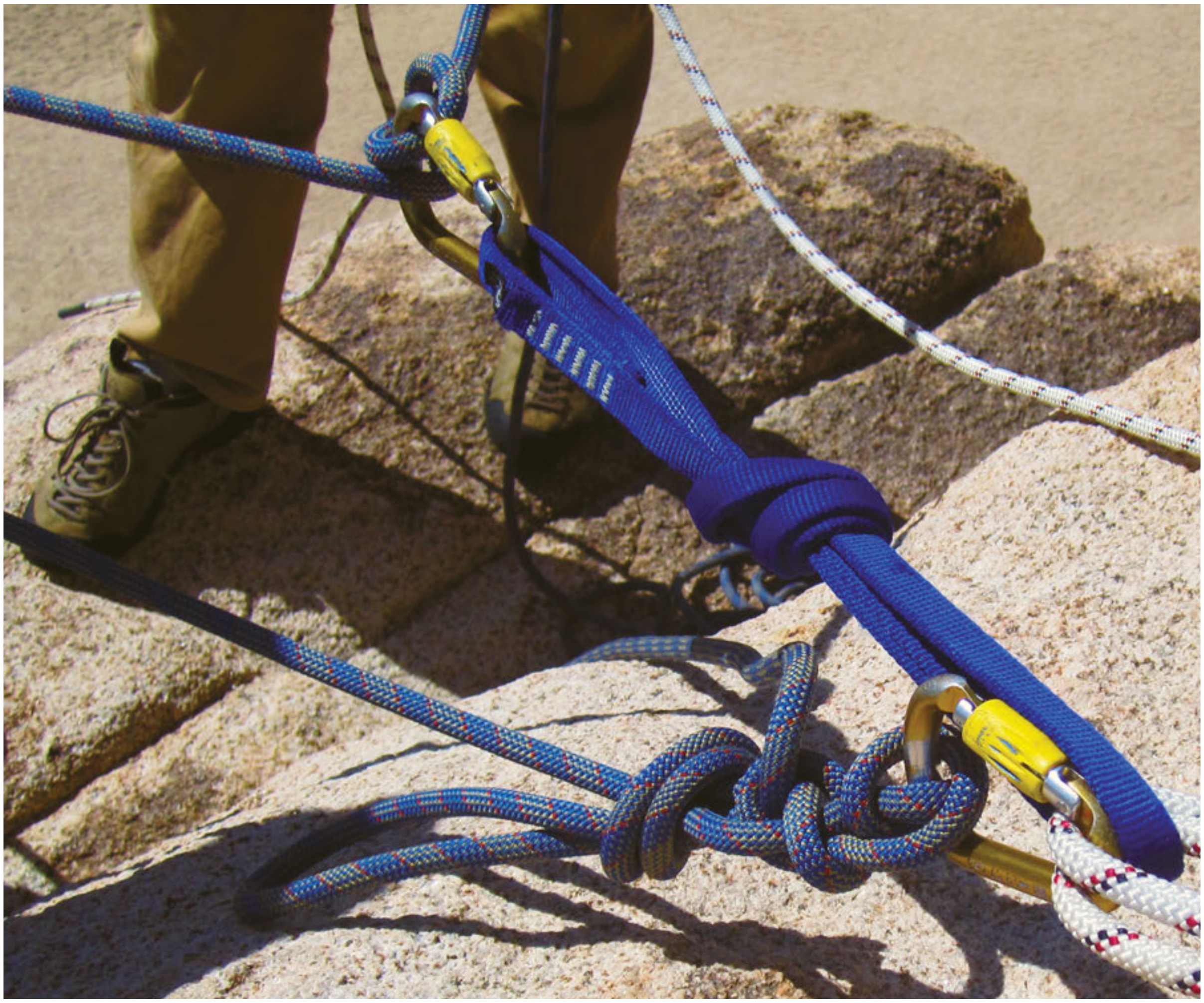

Detail of belayed rappel setup, rigged for contingencies. The rappeller is rappelling on the rope in the foreground, tied to the master point with a Munter/mule combination, which is releasable under tension. The belayer is belaying with the direct belay technique using a Munter hitch from an extended master point, created by basketing a double-length (48-inch) nylon runner through the anchor's two-loop master point, then tying it with an overhand knot for redundancy. The Munter/mule/overhand combo is one ugly knot, but it's easily releasable under tension, and that's the beauty of it.

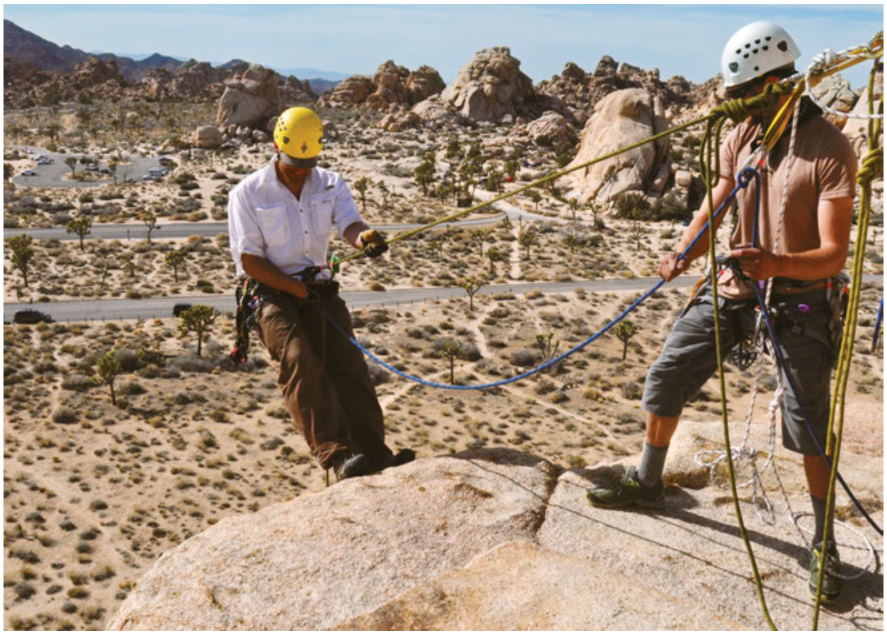

At this site the distance of the rappel is over 100 feet, so the system is rigged with two ropes. The yellow rope is the rappel rope, attached to the anchor with a Munter/mule/overhand combination. The blue rope is the belay line. Note how the instructor has created multiple master points with the doubled yellow nylon sling.

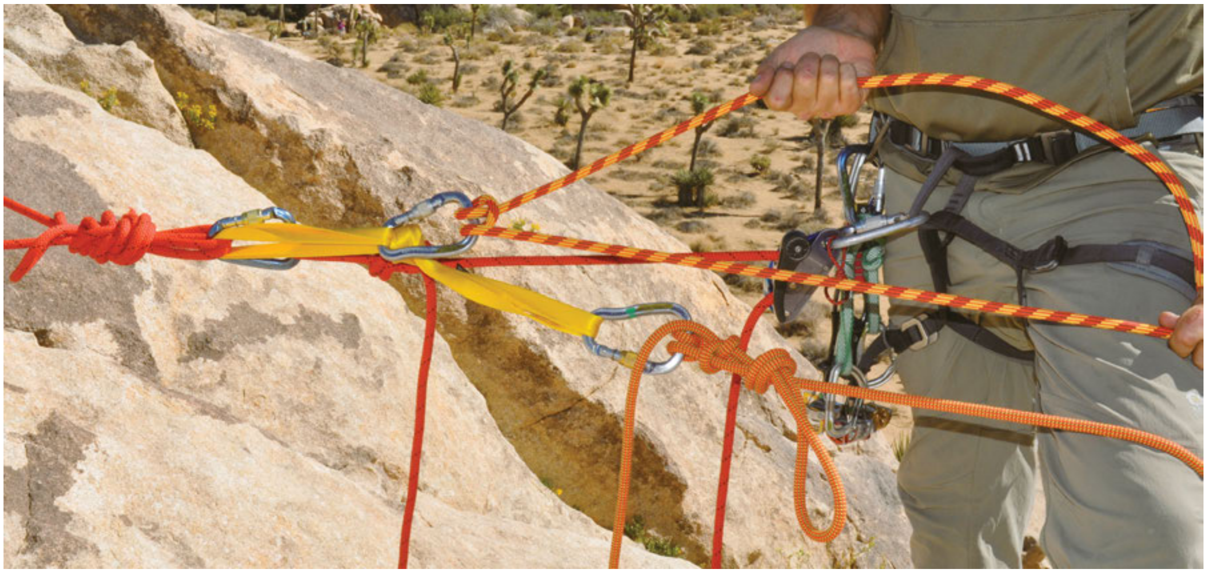

Here one bi-pattern rope is used for a releasable belayed rappel scenario. A BHK master point is rigged with a low-stretch (red) rope. The instructor is attached to a tether (red rope) with a Grigri. The striped orange dynamic rope held by the instructor is the belay line, and the rope at the bottom is the other half of the bi-pattern dynamic rope, serving as the rappel rope, attached to the anchor with a Munter/mule/overhand combination. By basketing the yellow nylon sling through the two-loop BHK master point and tying an overhand knot, the instructor has created two additional master points.

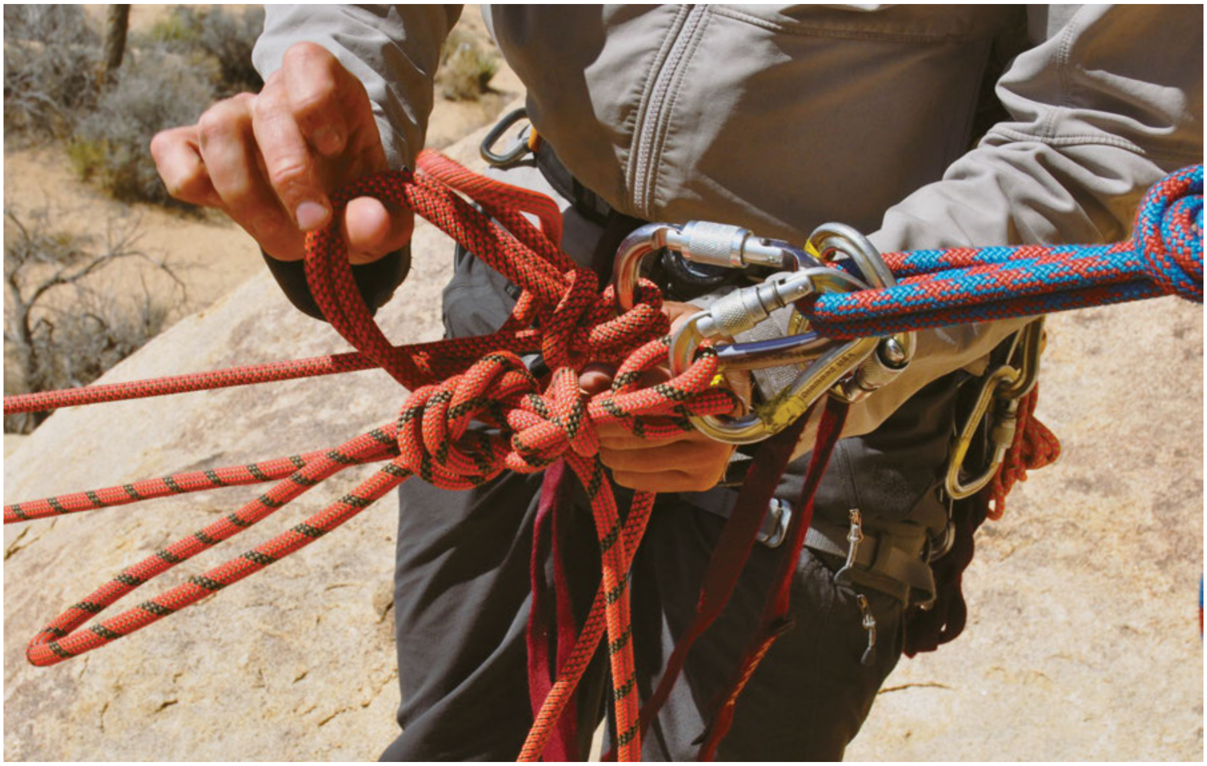

Without staggered master points, things can get clustered in a releasable rappel scenario.

stopping the rappel. There are two drawbacks with this method. One is that for the prusik to lock off, it must hold all the rappeller's weight. The second is that once it is weighted, the rappeller must remove all body weight from the prusik knot in order for it to be released, not an easy task on a free-hanging rappel.

The modern rappel backup utilizes the autoblock knot, rigged below the rappel device. There are two distinct advantages with this method. One is that for the autoblock knot to grab, it only needs to hold a very small percentage of the rappeller's weight, since it is on the braking side of the device, and the device itself is holding most of the weight and providing most of the friction. It is essentially like a brake hand squeezing and gripping the rope, and for that reason some instructors refer to it as the "third hand." The second big advantage of the autoblock method is that it is releasable under tension (i.e., when weighted and grabbing onto the rope). While rappelling down, simply form a circle with the thumb and forefinger (like the OK sign) and push the autoblock down while descending, allowing the rope to freely slide through the knot. When released, the autoblock knot rides up and grabs onto the rope, like the brake hand squeezing the rope. Releasing the autoblock, even when weighted, is as simple as sliding it back down and holding it in the "open" position. It's a beautiful thing, and easy to rig.

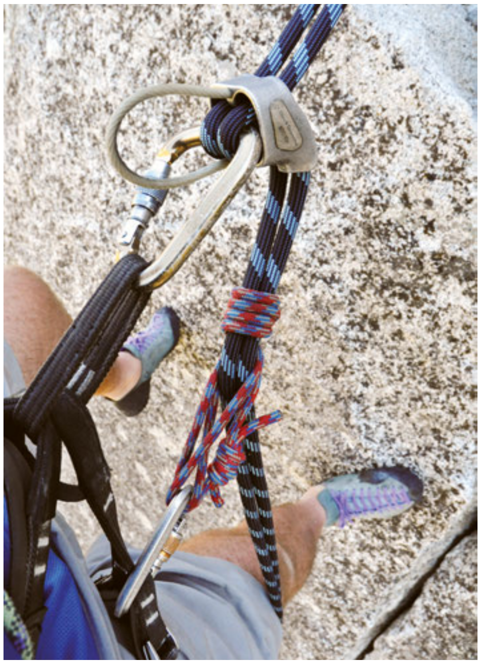

Left: Black Diamond ATC XP rigged for rappelling with a three-wrap autoblock backup clipped to the leg loop with a locking carabiner.

Bottom: another view of how to rig an autoblock backup, here using a Black Diamond ATC rappel device and a three-wrap autoblock tied with 6mm nylon cord attached to the leg loop with a locking carabiner.

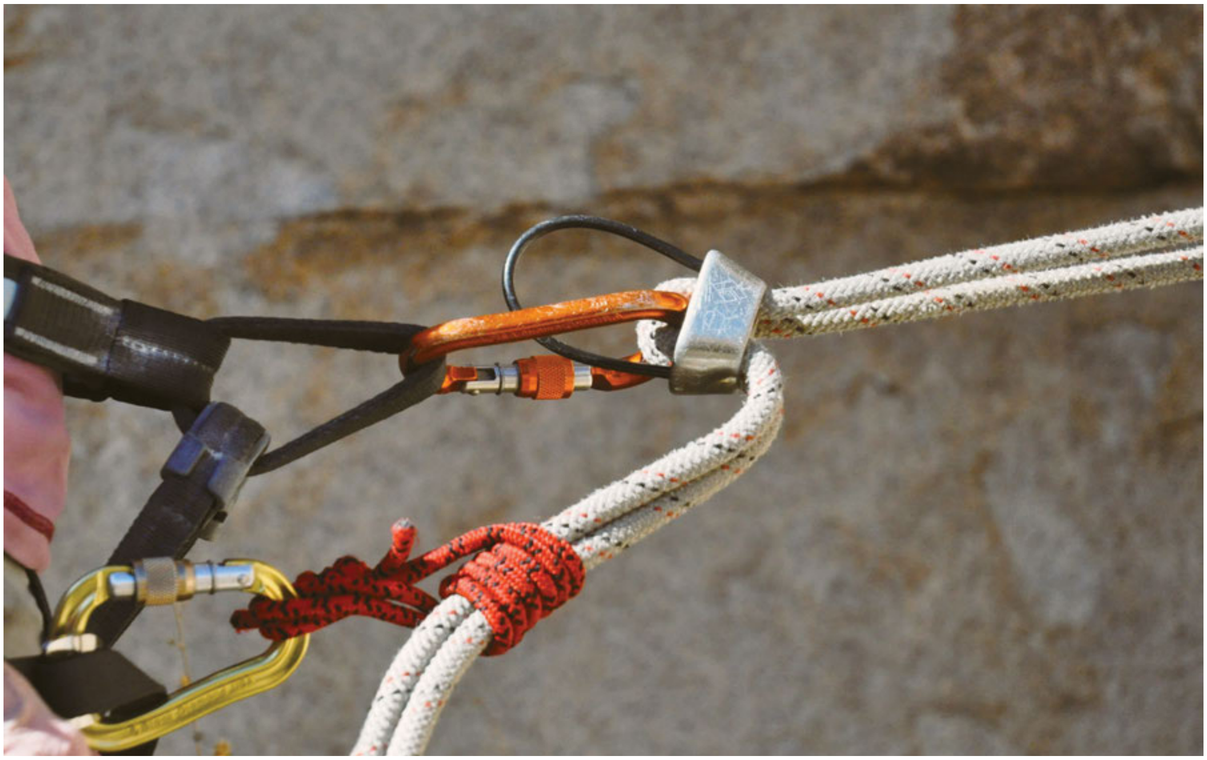

Not Foolproof

I (BG) conducted several tests over the last year with my autoblock rigged at my leg loop and my rappel device (ATC) at my belay loop. With someone giving me a fireman's belay to back me up, I flipped upside down to see if the autoblock would grab and hold. In three out of four tests, it failed to grip since it had ridden up against my ATC, and the rope kept sliding right through. It's not a foolproof backup.

When the technique was first introduced, the autoblock was attached to the rappeller's leg loop with a carabiner. The disadvantage of clipping the autoblock to a leg loop is that if for some reason the rappeller were to go unconscious and flip upside down, the autoblock would ride up and come in contact with the rappel device, which would prevent it from grabbing.

Also, recent harness leg loop buckle designs (particularly the Black Diamond trakFIT and Petzl "doubleback" buckle systems) can loosen when a carabiner is clipped into them and pulled outward. If using an autoblock on a leg loop, make sure the harness buckle cannot be defeated by an outward pull, or better yet, use a harness with a fixed (no buckle) leg loop design. If you have a harness with one of these buckles and need to rappel off the belay loop, consider clipping your autoblock to the inside of the leg loop, opposite the buckle.

In recent years a more secure method has been developed: extending the rappel device with a sling attached to the harness, and rigging the autoblock clipped into the belay loop. One configuration is to rig a double-length sewn nylon sling threaded

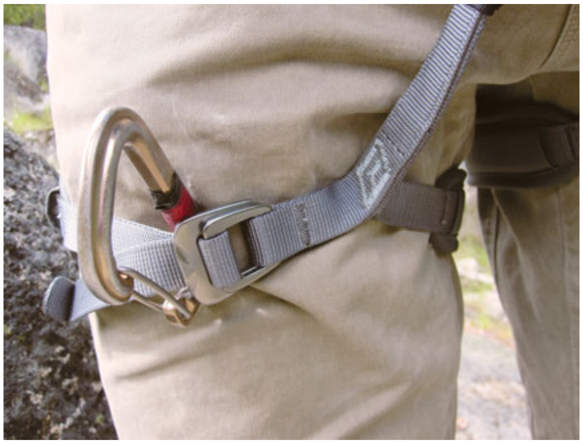

Note where the carabiner sits on the harness. Rigging an autoblock backup from this position has the potential to unbuckle the leg loop.

JASON D. MARTIN

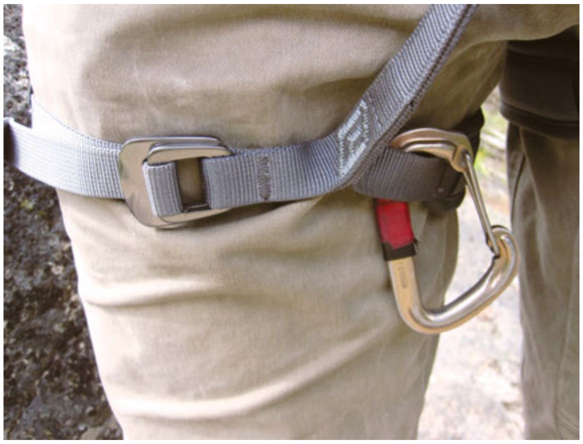

If you have a harness with suspect buckles, rig the autoblock backup from the inside of the leg loop.

JASON D. MARTIN

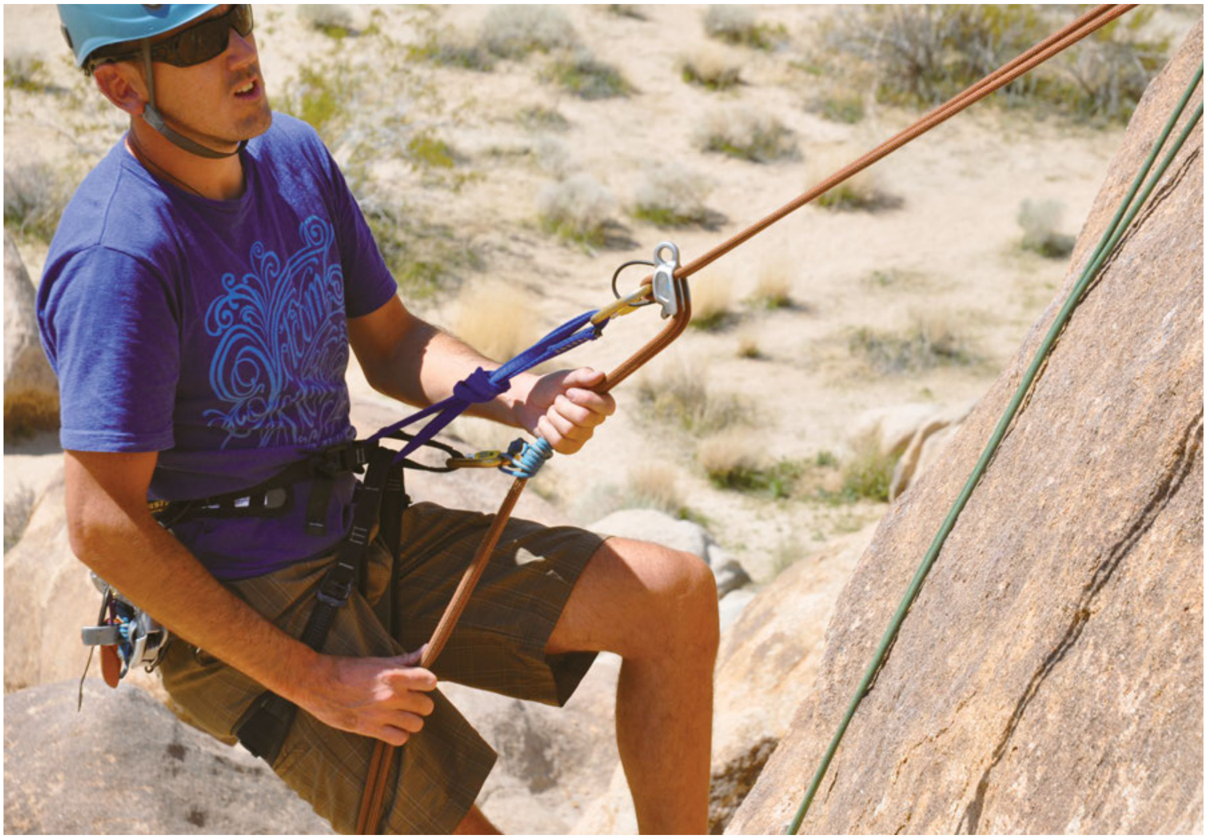

Adam Fox demonstrates the proper use of an extended rappel device with autoblock backup. A double-length nylon runner is threaded through the harness tie-in points and tied with an overhand knot for redundancy. The rappel device is the Black Diamond ATC Guide in rappel mode. The autoblock knot is clipped to a locking carabiner attached to the belay loop.

through both tie-in points at the front of the harness and tied with an overhand knot, to gain redundancy at the sling. The preferred cordage is a nylon sling (18mm or $^{11}\!/_{16}$-inch width) over a thin (10mm) Dyneema sling for this application because nylon has a higher melting point. If the rappel rope is running across the sling, it could potentially create some heat due to the friction, which could damage the sling and reduce its breaking strength in later applications.

Managing the Risks—Avoiding Rappelling Accidents

When teaching rappelling to novice climbers, it is important to stress risk management. Although rappelling is a simple technique, statistically a high percentage of rappelling accidents end in a fatality. Analyzing rappelling accidents helps to understand what can go wrong and what to do to keep it from happening. Here is a look at the most common rappelling accident scenarios.

Scenario 1: Rappelling Off One or Both Ends of the Rope

This happens with some regularity, and almost every year there are several fatal rappelling accidents in America where someone simply rappelled off the end (or ends) of a rope. In the single pitch environment, it can happen when the ends are uneven on a double-rope rappel, especially if there is no middle mark. When the short end passes through the rappelling device, only one strand of the doubled rope remains in the device, and the climber's body weight rapidly pulls the rope through the rappel anchor, quickly dispatching the climber to the ground. To prevent this, *close the system* by tying stopper knots separately in both ends of the rope. Make sure that the stopper knot will jam in the rappel device that is being used.

Scenario 2: Not Clipping Both Strands of the Rope into the Carabiner on a Double-Rope Rappel

This is an easy mistake to make when not alert and double-checking the system. When visually inspecting the top of a tube-style device, it will appear as if both strands are properly threaded because the device holds the bight of rope inside itself, even if the rope hasn't been clipped into the carabiner behind it. If only one strand makes it into the locking carabiner, when leaning back and weighting the rope, the climber will descend as rapidly as in the first scenario—and with equally injurious or fatal results. A good habit is to first tether in with a sling to the rappel anchor, rig the rappel device, then weight the rappel system and double-check everything before unclipping the tether. Rigging an autoblock backup is also a good idea.

Check and Check Again

I (BG) have done thousands of rappels in my climbing career, and one time I caught myself not clipping both strands on a double-rope rappel. Luckily for me I'd rigged an autoblock backup, and as I weighted the system, I noticed my error. If not for that autoblock backup, I probably wouldn't be writing this today.

Rappelling Systems Check

It is important to teach students how to go through a systematic mental checklist before rappelling. An acronym that facilitates the process is ABCDE. A is the rappel anchor. Take a look at the anchor, slings, chains, etc., and make sure the rappel rope is threaded properly through the anchor. The anchor should be redundant all the way to the point where the rope is threaded through the anchor. Do not rely on a single piece of gear in the anchor system, whether it is a single cord, sling, or rappel ring. B is for buckles on your harness—double-check to make sure they are buckled properly and doubled back appropriately. C is for carabiner. Make sure the locking carabiner that attaches the rappel device to the harness is being loaded properly on the long axis—and squeeze check to make sure that it is locked! D is for look down and see where the rope goes. Does it reach the ground? Or the next ledge? E is for the rope ends. Do they have stopper knots? Is the system closed?

Scenario 3: Not Tethering to the Anchor

A sewn sling clipped to the anchor with a locking carabiner for a tether is all it takes to prevent this scenario. Using a daisy chain as a personal tether is not recommended, because if a carabiner is clipped into two adjacent loops, a weak and dangerous connection has been created. Better than a standard daisy chain with sewn pockets is the newer, modern version that consists of full runner-strength sewn loops, like the Metolius PAS (personal anchor system) or the Sterling Chain Reactor. These more versatile tethers also allow the rappel device to be extended away from the harness as described in the Instructor Rappel Backup. Again, a good habit is to weight the rappel device and check to make sure it's rigged properly and supports body weight before unclipping the tether from the anchor.

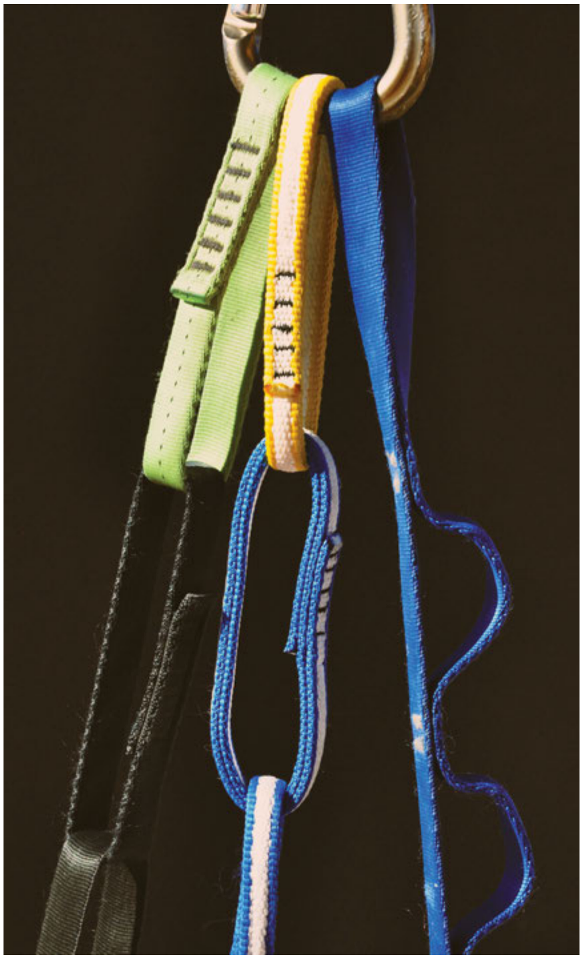

Tether comparison. Left to right: Sterling Chain Reactor, Metolius PAS, Black Diamond Daisy Chain.

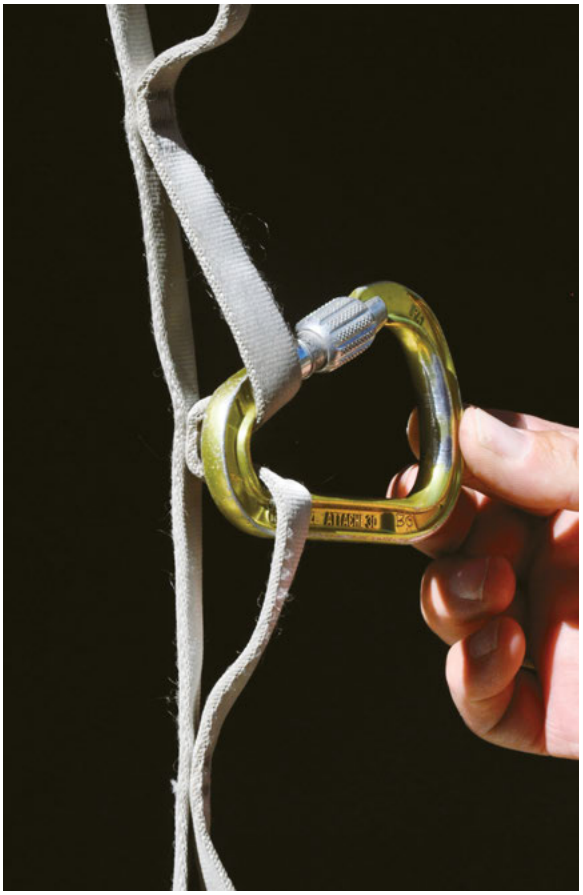

A traditional daisy chain with bartacked pockets is not recommended for rappel extensions or as a personal tether, because if a carabiner is clipped into two loops, an extremely weak connection is created (3 kN or 674 lbs.).

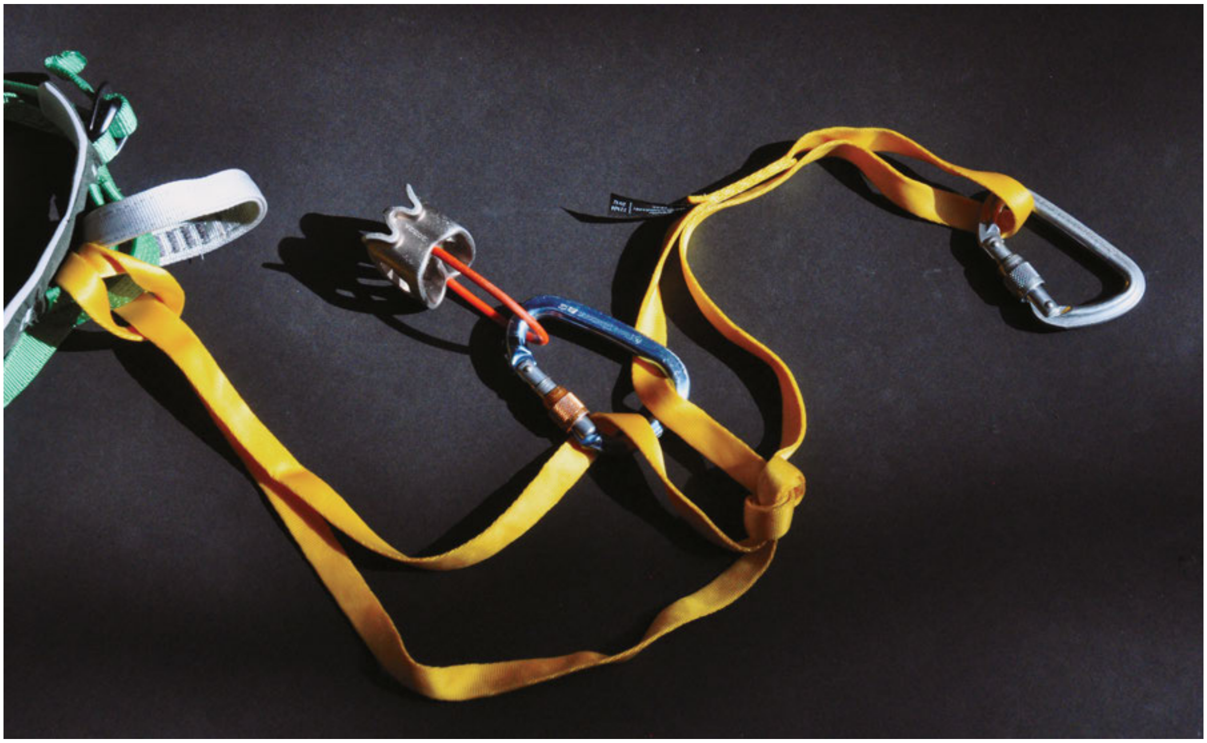

To rig a tether/rappel extension with a sling, start with a double-length (48-inch/120cm) sling and girth-hitch it through both tiein points on the harness. Tie an overhand knot about halfway down the length of the sling, adjusting the knot so that when it is clipped back into the harness belay loop with a locking carabiner, the length to the overhand knot is equalized. Clip the locking carabiner of the rappel device into both loops in the sling to make the sling itself redundant when it's clipped back into the harness belay loop.

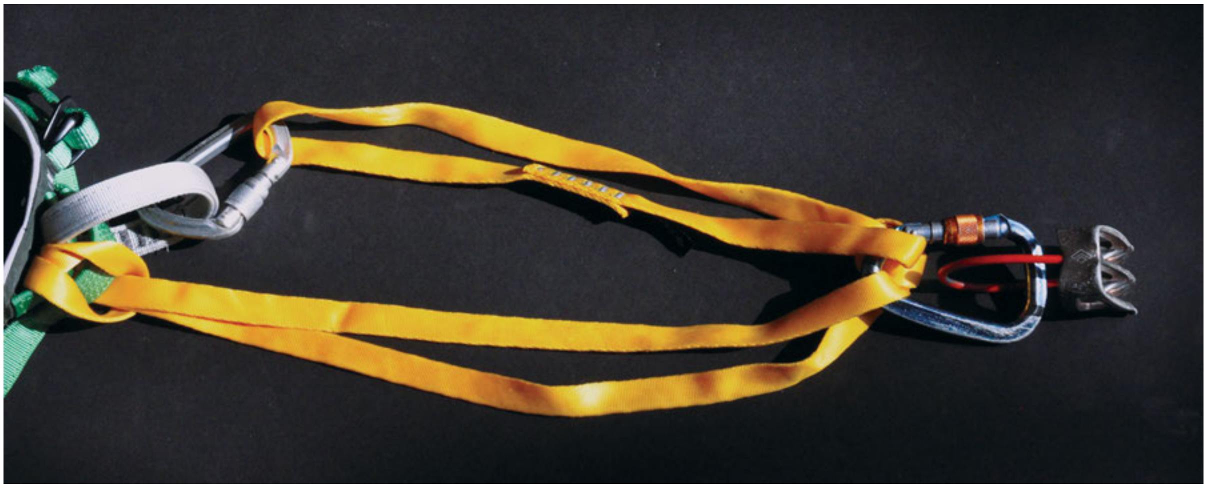

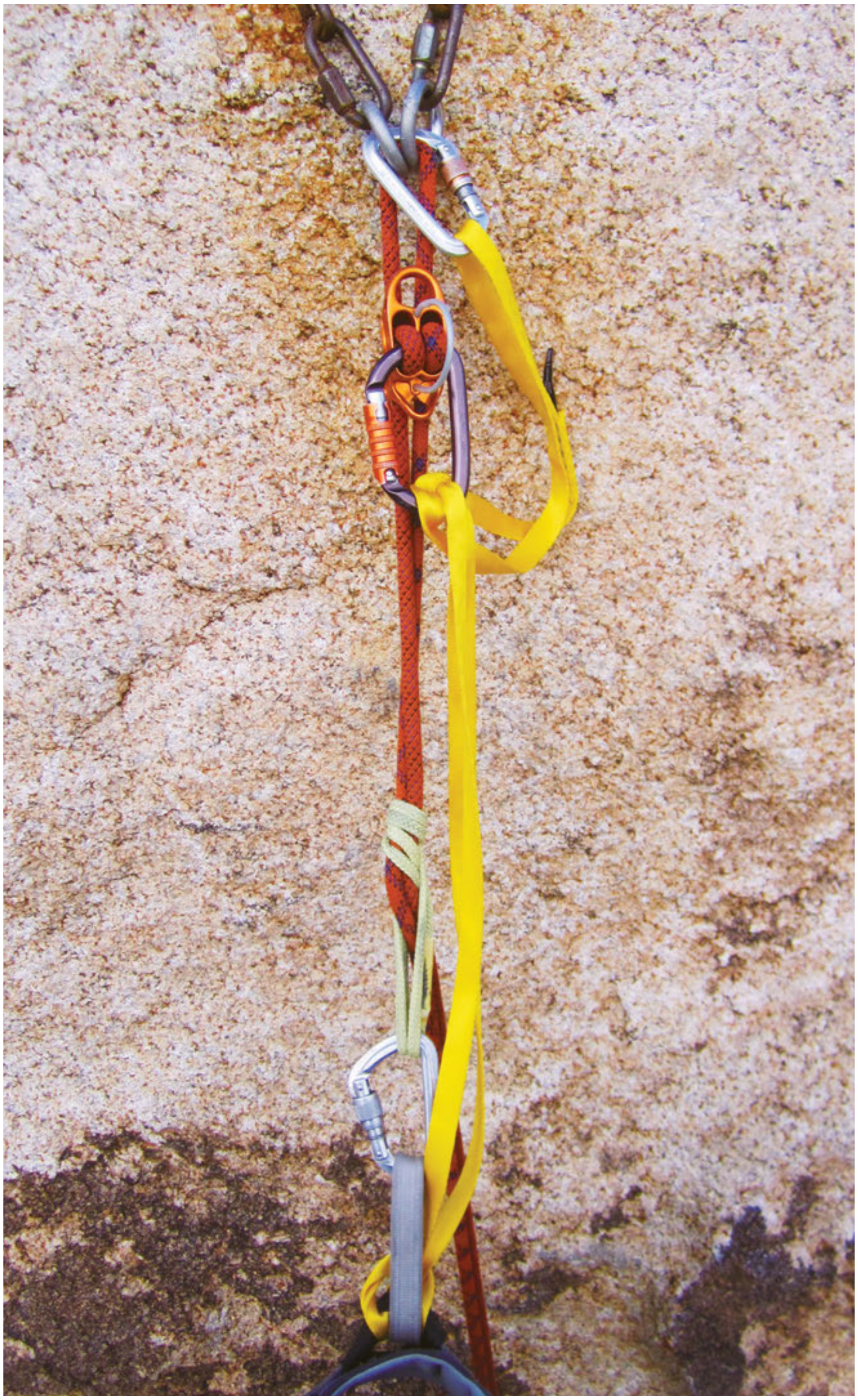

Here my double-length (48-inch) Black Diamond 18mm nylon sling is rigged as a tether/extension. I (BG) girth-hitched the yellow sling through both harness tie-in points. I'd previously tied an overhand knot in the sling, and I then clipped the locking carabiner of my rappel device (a Petzl Reverso) into both loops in the sling. I rigged an autoblock backup with a Sterling Hollowblock and clipped it to my harness belay loop with a locking carabiner. After I double-check everything, I'll unclip the locking carabiner from the anchor and clip it back to my harness belay loop, adding redundancy to the sling.

CHAPTER 13

Basic Assistance and Rescue Skills

One of the many reasons that a student elects to take a course from a certified climbing instructor is that there is an unwritten agreement between the instructor and the student. The agreement is that if the student gets into trouble on the course, the instructor will be there to help the student get back out of trouble.

Certainly the word "trouble" could mean many different things. Examples include anything from forgetting one's water bottle in the car to suffering an injury. From a technical perspective in a single pitch environment, trouble commonly means that an individual is stuck somehow on a route. Following are several ways to get people unstuck and out of trouble.

Assistance from the Base

The first step in any assist scenario is to seek out the nontechnical solution. Coaching, encouraging, directing, or allowing gravity to take over are more efficient and expose both the instructor and student to fewer risks. When it becomes apparent that a nontechnical solution will not resolve the problem, it is time to incorporate more advanced techniques.

The basic assist skills presented below are most easily accomplished with an ABD, like a Grigri, and it is recommended that an ABD be used for belaying from both the base and top of the cliff, as it streamlines and facilitates these assist techniques. It is the right tool for the job and worth every extra ounce that it weighs over a traditional belay device. It is recommended that instructors learn the climber pickoff and the 3:1 raising system with a Grigri before trying it without one.

The Climber Pickoff

The "climber pickoff" is essentially the same technique used by the belayer to rescue a fallen lead climber who is injured and can't be lowered. The difference in a toprope situation is that there is a solid anchor system, not just the one piece of gear that held the leader's fall.

The climber pickoff utilizes a counterbalance ascension and rappel technique and requires minimal equipment. Basically the instructor ascends the rope to the climber, then rappels down with them.

All that is needed is the following equipment:

- 1 double-length (48-inch) sling
- 1 locking carabiner
- 1 prusik cord (6mm or 7mm soft nylon cord, 5 feet long, tied into a sling with a double fisherman's knot)

When belaying with a Grigri, follow these steps:

1. Start by tying a backup knot (overhand loop) on the brake strand side of the Grigri.
2. Tie a friction hitch (klemheist or prusik) with the prusik cord on the load strand going up to the climber, and attach the double-length sling to it with a locking carabiner.
3. Ascend the rope by sliding the friction hitch up, so that when the foot is placed in the sling the knee is bent at a 90-degree angle. A knot can be tied in the sling to shorten it as needed. Stand up in the sling, and as you straighten the leg, simultaneously pull the slack through the Grigri (pulling straight up on the brake strand) and hang in the harness off the Grigri. Alternate

standing up in the sling, then hanging in the harness, tight to the Grigri. About every 10 feet or so, tie a backup knot (overhand loop) on the brake strand side of the Grigri.

4. Upon reaching the climber, transfer the friction hitch from the instructor side of the rope onto the rope above the climber. The instructor then takes the double-length sling and threads it (basket style) through his or her belay loop and clips both ends of the sling to the locking carabiner attached to the friction hitch. Where the friction hitch is positioned will affect where

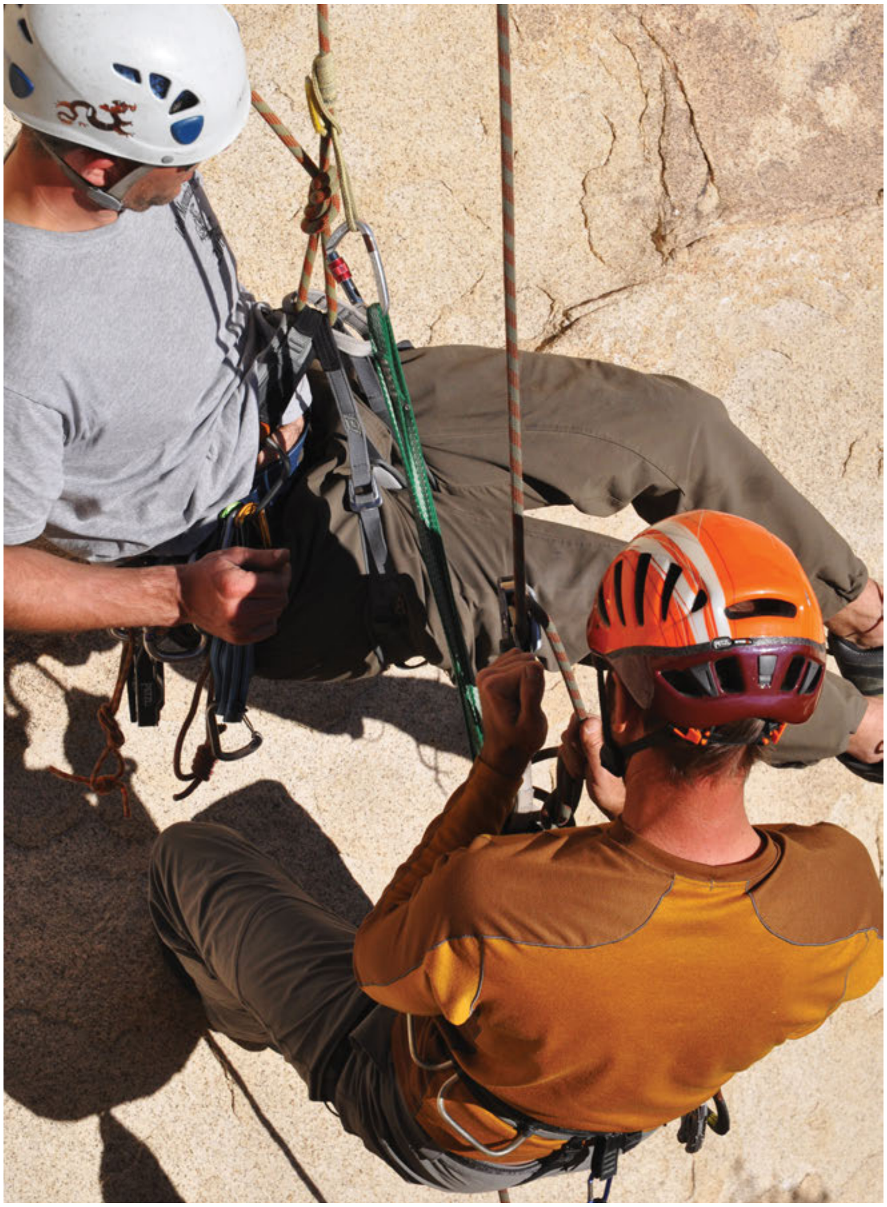

Counterbalance rappel using a Grigri after "picking off" a climber from the cliff.

the assisted climber will be positioned. If the friction hitch is high, the climber will remain at that position while rappelling. If the position of the friction hitch (klemheist or prusik) is just above the climber tie-in knot, the climber will move up and slightly above the instructor while descending.

5. Rappel with the Grigri, untying the backup knots as you descend.

Basic assistance skills will allow an instructor to assist a student if needed.

Rare Rescues

In over thirty years of professionally guiding thousands of clients in toprope situations, I (BG) have only had a few instances where I actually had to go up on the rock and bring a climber down—all of them being young kids who were overcome by fear and mentally lost control, afraid to lean back and weight the rope so they could be lowered back down. Having the knowledge and skills to go up the rope and bring a student down is important in case the instructor ever has to assist an injured student, even in a toprope situation.

If taking over someone else's belay and the belay strand is weighted, the sequence would be as follows:

1. Tie a backup knot (overhand loop) on the brake strand side of the belayer's belay device.
2. Clip a Grigri onto the rope between the backup knot and the belayer's belay device.

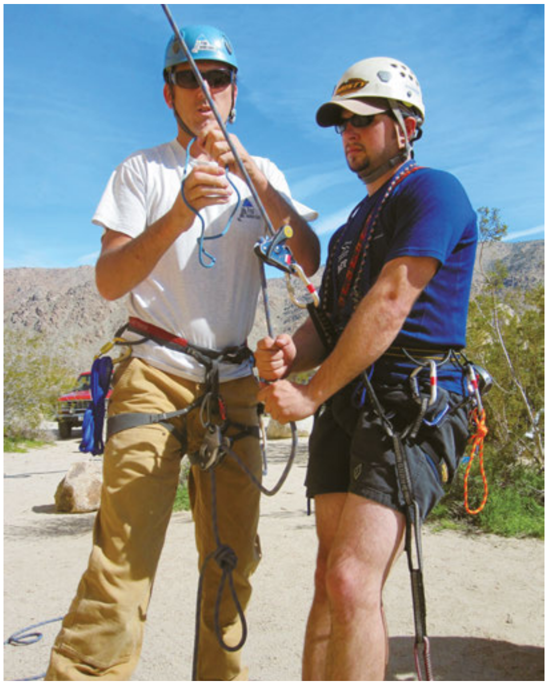

3. Tie a friction hitch (prusik or klemheist) on the load strand (going up to the climber) of the belayer's rope.

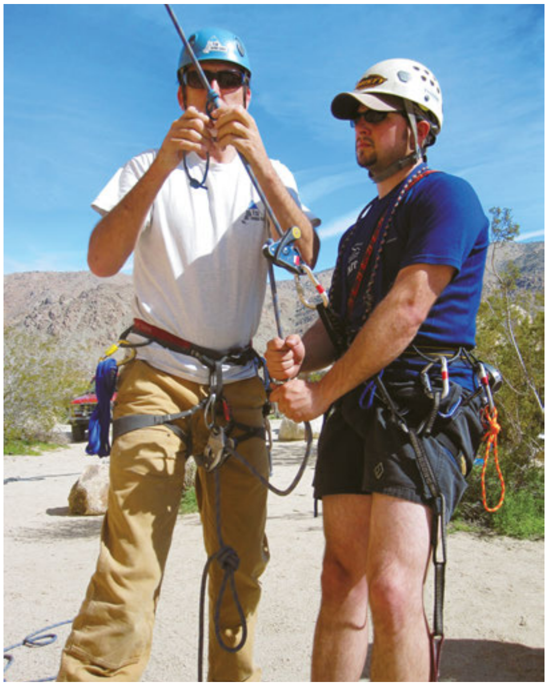

4. Thread (basket style) the double-length sling through the belayer's belay loop and attach it to the friction hitch with a locking carabiner.

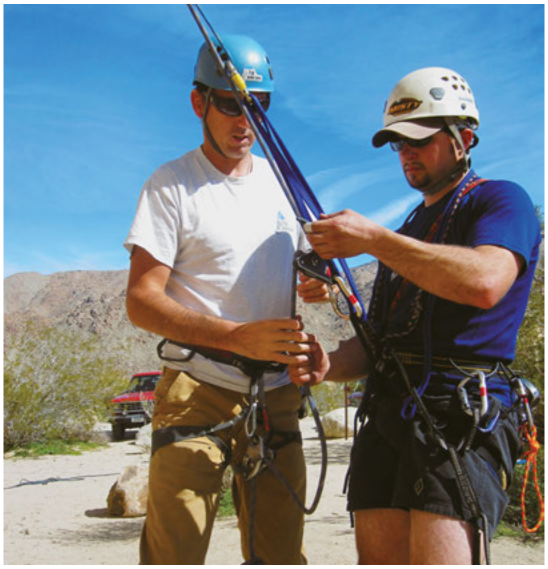

5. Have the belayer give slack at the belay device—the tension will now be transferred to the friction hitch, with the belayer essentially becoming a ground anchor; the belayer then can unclip and remove the belay device.

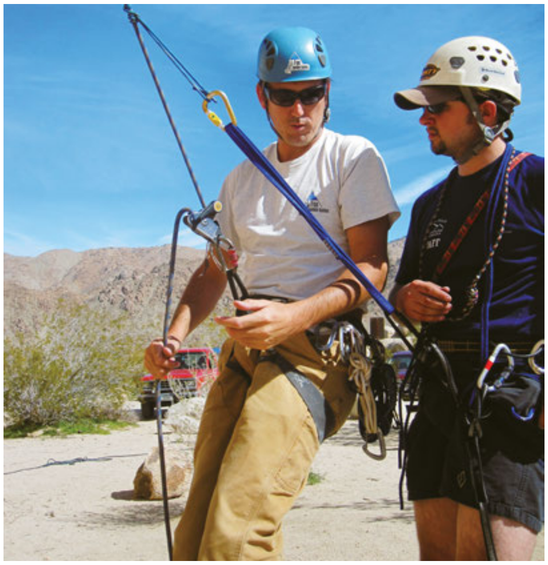

6. Transfer the weight of the hanging climber from the belayer to the instructor's Grigri, by taking up all the slack in the rope and pulling upward on the brake strand side of the Grigri, using body weight as the anchor.
7. Unclip one end of the sling from the locking carabiner at the friction hitch.
8. Ascend the rope by sliding the friction hitch up, so that when the foot is placed in the sling the knee is bent at a 90-degree angle. A knot can be tied in the sling to shorten it as needed. Stand up in the sling, and as you straighten the leg, simultaneously pull the slack through the Grigri (pulling straight up on the brake strand) and hang in the harness off the Grigri. Alternate standing up in the sling, then hanging in the harness, tight to the Grigri. About every 10 feet

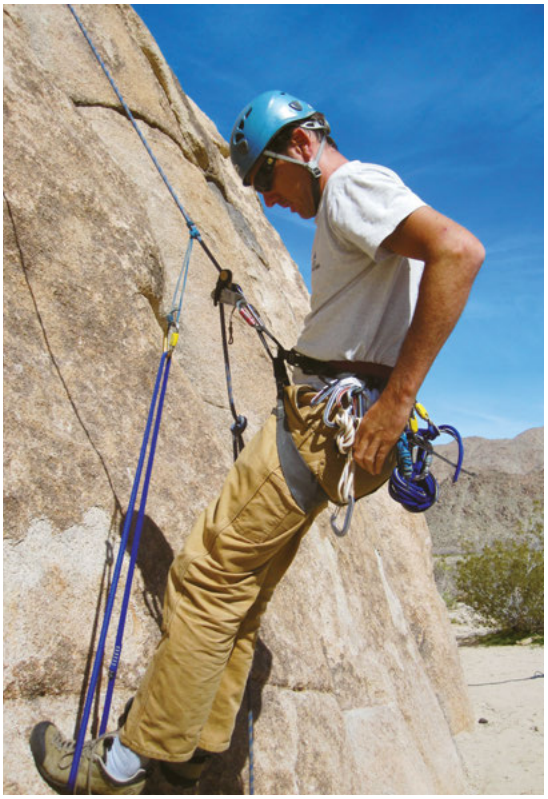

or so, tie a backup knot (overhand loop) on the brake strand side of the Grigri.

9. Upon reaching the climber, transfer the friction hitch from the instructor side of the rope onto the rope above the climber. The instructor then takes the double-length sling and threads it (basket style) through his or her belay loop, and clips both ends of the sling to the locking carabiner attached to the friction hitch. Where the friction hitch is positioned will affect where the assisted climber will be positioned. If the friction hitch is high, the climber will remain at that position while rappelling. If the position of the friction hitch (klemheist or prusik) is just above the climber tie-in knot, the climber will move up and slightly above the instructor while descending.

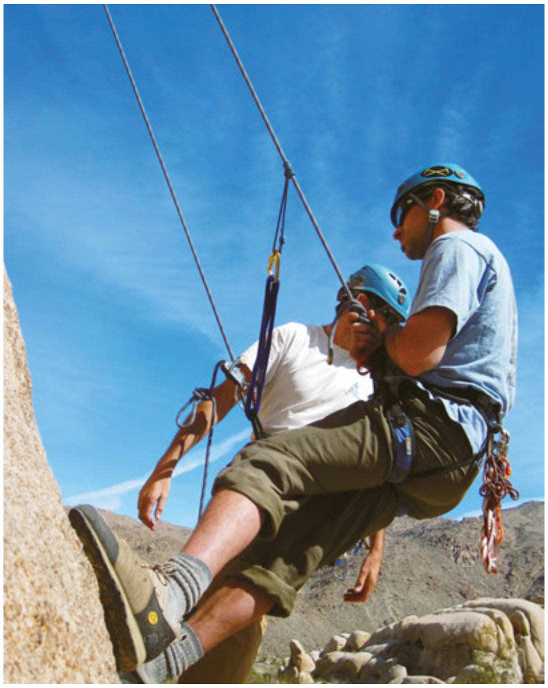

10. Rappel with the Grigri, untying the backup knots as you descend.

Assistance from the Top

Coaching

The best and easiest way to assist a climber who's climbing up to you from below is simply to position your belay stance where you'll be able to watch them. By having visual contact you'll be able to manage the belay more effectively, taking in slack in sync with the movement of the climber. Giving advice to a climber who encounters a challenging section may prove ineffective if the instructor can't actually see them.

Vector Pull

If someone just physically can't do a move or is too tired to climb up a tough crux section, some assistance from the belayer may be enough to get them over the impasse. A simple method is the vector pull.

Imagine a rope strung tight across a chasm, anchored on both ends, rigged for a climber to clip in with a pulley to slide across. This is known as a Tyrolean traverse. If the climber weighs 200 pounds and is hanging from the middle of the rope, with an angle slightly less than 180 degrees, the force at each anchor is roughly 1,000 pounds. As the angle approaches 180 degrees, the force on the anchor points increases even more. Less angle, less force. For example, when the angle is relatively narrow (22 degrees or less), the 200-pound climber weights each anchor with only 100 pounds. Let's say the instructor is belaying from the top of the cliff, and the climber is unable to move up past a tough section of the climb. With a tight rope (180-degree angle), by reaching down and pulling perpendicular to the line of the rope, a surprising amount of force can be generated just by creating an angle slightly less than 180 degrees. By pulling on the rope at a right angle, it is like the climber hanging in the middle of the Tyrolean traverse who is generating ten times his bodyweight at the anchor points. This is often enough to boost someone past a move, especially if he is able to assist by climbing.

3:1 Raising System

One reason to use the direct belay technique with an ABD when belaying from the top of the cliff is that it is easily converted to a 3:1 raising system (aka the Z system) in a matter of seconds. For most climbing instructors, the direct belay is always the first choice, providing the anchor is solid, because it allows the instructor to anticipate and prepare for any eventuality, like a quick lowering or raising of the climber. Most instructors use a Grigri, clipped directly to the master point or extended master point (using the climbing rope). To set up a 3:1 raise, follow these steps:

1. Tie a backup knot (overhand loop) on the brake strand side of the Grigri. This allows the

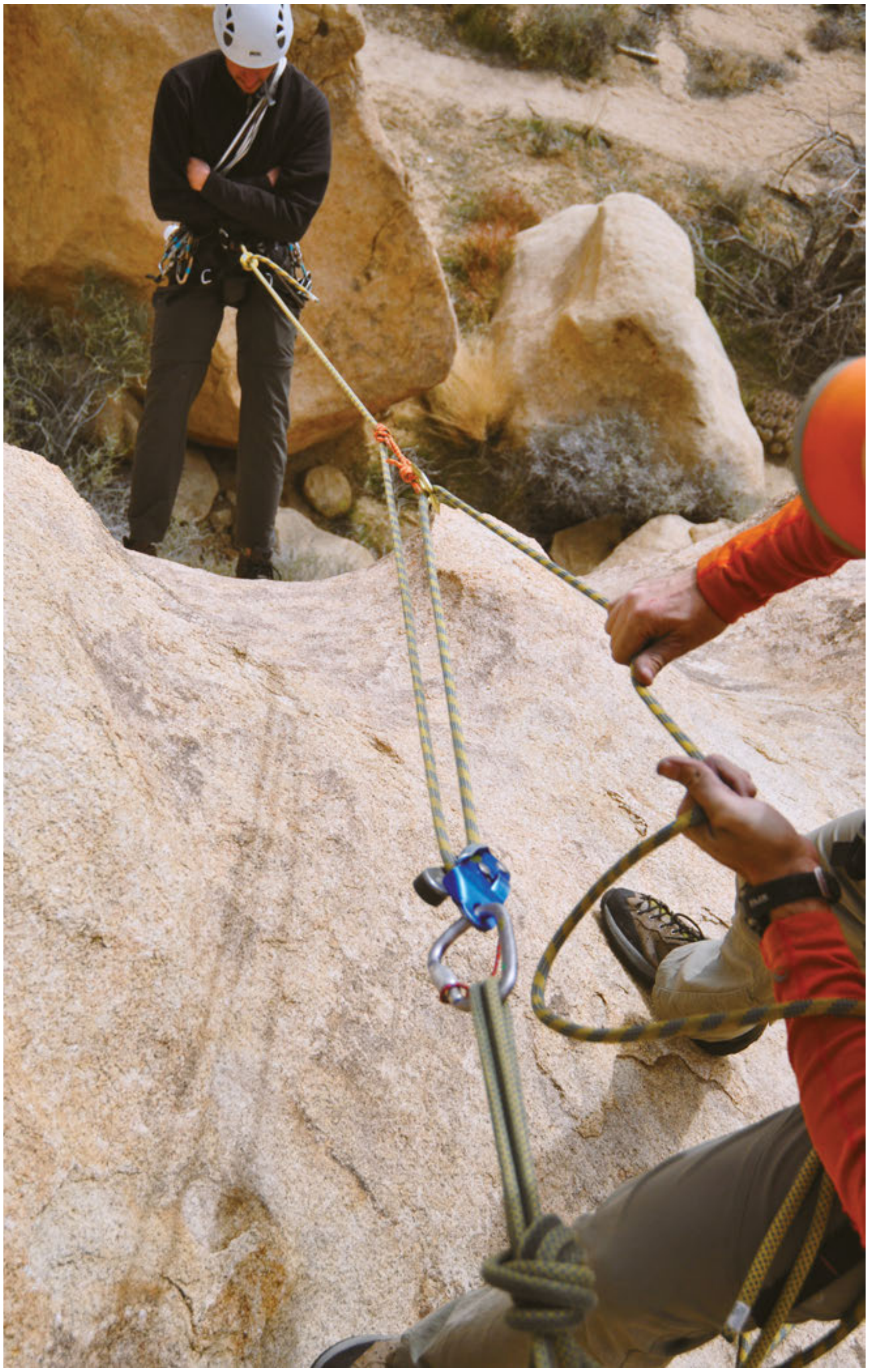

Also known as the Z pulley system, the 3:1 raising system is easy to rig if a Grigri is being used for a direct belay. The Grigri is the ratchet (which locks off when needed to reset), and the friction hitch is the tractor (which moves up and down the field). The distance the rope travels between resets is called the throw.

instructor to go "hands free" and release the brake strand.

2. Tie a friction hitch (prusik or klemheist) on the load strand going down to the climber.
3. Clip the brake strand side of the rope (from the Grigri) to a locking carabiner clipped to the friction hitch and push the friction hitch as far as possible toward the climber.
4. Next, untie the backup knot and pull up on the brake strand side of the rope. For every 3 feet pulled through the GriGri, the load is raised 1 foot. Friction is the enemy in any raising system. If the rope going to the climber is in contact with a large surface area of rock, the raise will be correspondingly more difficult. A pulley at the friction hitch carabiner reduces friction and makes it easier to pull. Remember, this technique is for assisting a climber, helping them get past a tough spot, not to haul up a severely injured or unconscious climber.
5. When the friction hitch is all the way to the Grigri, reset the friction hitch by sliding it back down toward the climber. The Grigri's built-in ratchet will lock off and hold the load as this is done. Then continue the raise.

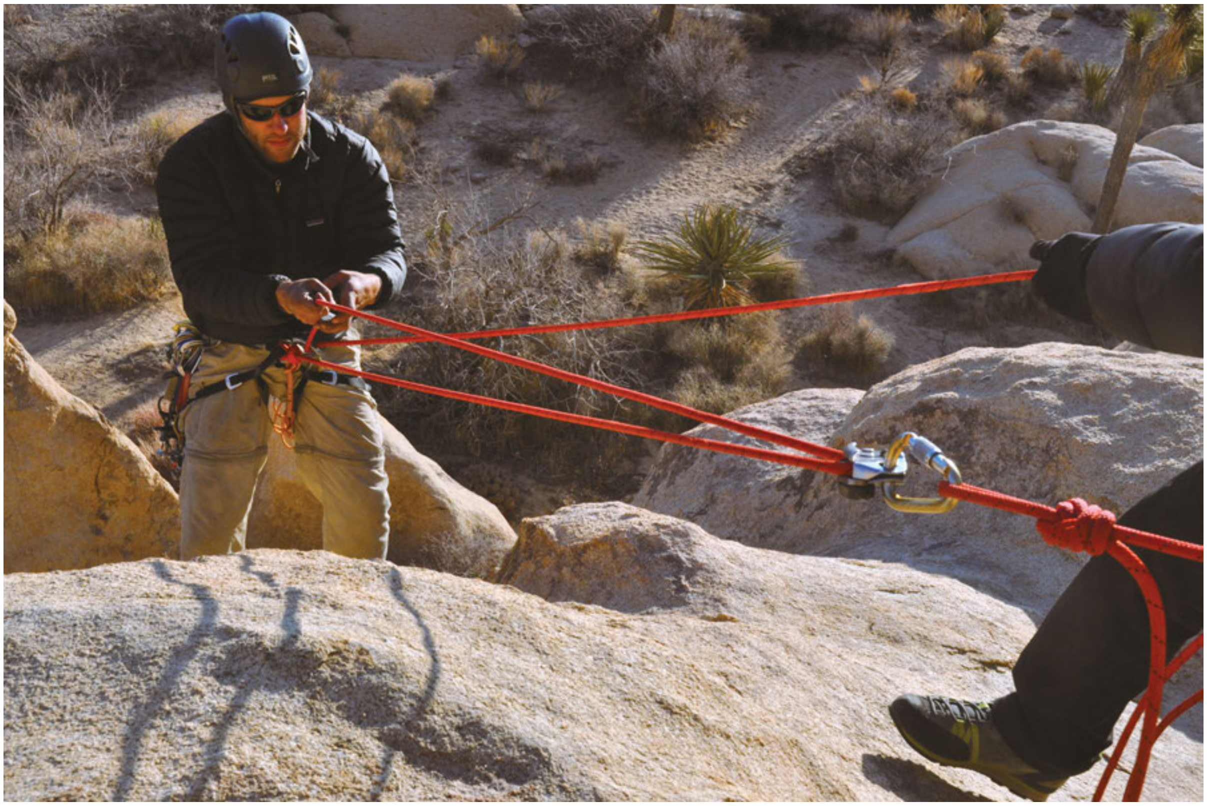

A 3:1 assisted raise. With the climber also pulling, there is a tremendous mechanical advantage, making it far easier to raise someone than with a 3:1 raise.

3:1 Assisted Raise

For this method the climber must be close enough that it is possible to throw him a bight of rope. The climber clips the rope into his harness to assist in the raise. Using this system allows both the climber and instructor to work together, and makes it much easier for the instructor to raise the climber.

The steps are as follows:

1. Tie a backup knot (overhand loop) on the brake strand side of the belay device.
2. Toss a bight of rope down to the climber and have him clip it into his belay loop with a locking carabiner. If he already has a locking carabiner on his harness, have him clip that to his belay loop and to the bight of rope. If not, carefully lower the carabiner down. Don't throw it.
3. Identify which strand the climber should pull on by shaking it.
4. Untie the backup knot.
5. The climber pulls down as the instructor pulls up on the brake strand side of the rope. Warn the climber to watch his hands so they are not pinched in the carabiner when pulling.

CHAPTER 14

Fixed Lines

A fixed line is a rope that is literally fixed in one location to provide additional security to climbers, by providing a line that they can be attached to while moving. In an instructional setting, fixed lines provide a means to protect an individual in a group who is required to move over small sections of exposed 2nd, 3rd, or even 4th-class terrain.

Fixed lines should not be employed to protect long sections of exposed or 5th-class terrain. They should only be installed in areas where the likelihood of a fall is relatively low, but where the consequences of a fall are high enough to warrant additional protection. The first line of defense in these areas is that the terrain is easy enough that the student is unlikely to fall. The fixed line is a backup to the student's ability to effectively scramble. If the consequences of a fall are high, a belay is necessary.

Fixed lines usually take time to rig. If possible, place them before the students arrive. If the local ethic doesn't allow for this, then scout the area thoroughly and make a fixed line plan before the instructional day starts.

Following are the steps for fixed line installation:

1. Scout the crag or the approach to the instructional area for exposure and determine where a fixed line is needed.
2. Build an anchor at the top of the crag and clip the end of the rope into the master point with a locking carabiner.
3. Work down through the exposed area, pulling the rope down and placing gear along the way.

   At each piece of gear, the fixed line should be clipped in with a loop knot. The rope should *not* run through the carabiners freely. Each section of the rope should be isolated.

Depending on the fixed line's orientation and the terrain, the means by which the line is secured at the beginning varies. A catastrophe knot or a single anchor point may be sufficient. For traverses or anywhere else that an unfixed end could lead to catastrophe, such as an injury or fatality, an anchor system should be used.

There are three ways commonly used to travel on a fixed line. The method chosen is determined by the terrain, the orientation of the rope, and the amount of security desired. The first, and simplest, method is a hand line. As the name implies, the students simply hold the line. In this case the likelihood of a fall is low, and the individual or group simply needs a little bit of additional security.

Second is to use lobster claws. This technique involves girth-hitching two slings to the tie-in point on the climber's harness with a locking carabiner clipped to the end of each sling. The climber clips and locks both carabiners on the fixed line, then moves along the line. The climber bypasses anchor points by clipping past each point with one lobster claw and then the other. This allows the climber to stay attached to the rope at all times

The lobster claw technique is good for exposed traverses, but is not acceptable when the fixed line is in a vertical orientation where a student could fall back to the previous anchor point. Doing so could generate forces exceeding a factor 2 fall.

The third technique, which offers the greatest security, is to use a friction hitch on the fixed line,

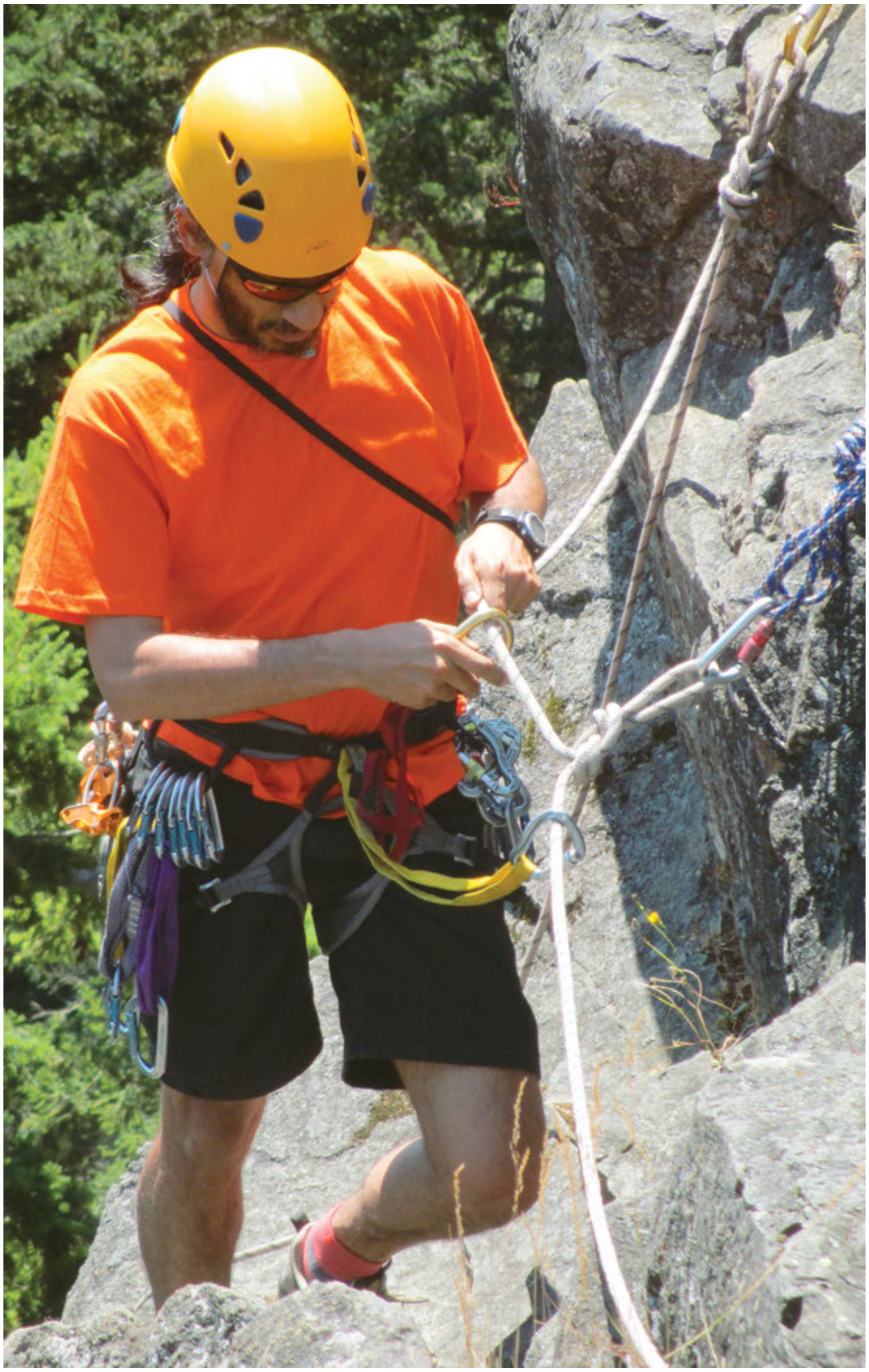

A climber using the lobster claw technique to pass an anchor point.

JASON D. MARTIN

which the climber clips into with a lobster claw. A friction hitch offers the most security because it won't allow the climber to fall anywhere if he or she slips (unless the hitch itself slips). If a section requires such tactics, it's a good idea to pre-rig the friction hitches, to allow the climber to smoothly clip in and continue moving. When the climber arrives at a friction hitch, one lobster claw is clipped to the friction hitch and the other to the rope.

The friction hitch fixed line technique should be used sparingly. If you believe that you need to resort to this technique, first ask yourself if it would be possible to belay the section. If you cannot belay the section, then consider the consequences of a friction hitch slipping. If the consequences are high, you may need to choose a different crag. No matter which technique you choose to employ, only one climber should ever be attached to a given part of the line. You should never have two climbers in the same section of the system.

Fixed lines are a useful tool, but they should not take the place of a real belay. Before exposing students to a fixed line, make sure to weigh all the options. Make sure that it is the best solution to the problem. And make sure that everybody knows exactly what they're supposed to do as they move up or down the line.

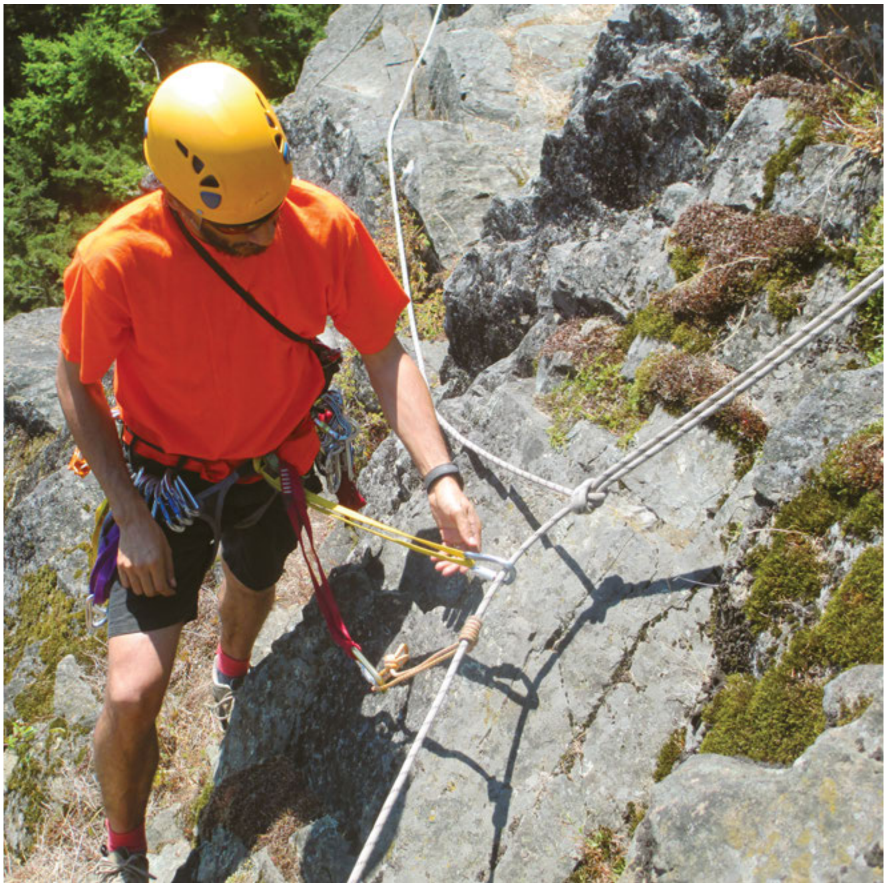

A climber is using a friction hitch (prusik) on the rope for more security. Note the bight of rope coming down from above. When a piece is not available, it is often possible to use a long bight of rope to extend a less convenient placement in order to continue the logical route of the fixed line.

JASON D. MARTIN

CHAPTER 15

Leading

Leading is dangerous, whether the instructor is on the sharp end with an inexperienced belayer, or watching a student lead for the first time. The risk management mantra of "What is the likelihood of a fall and what are the consequences?" should be in the forefront of the instructor's mind when lead climbing with students.

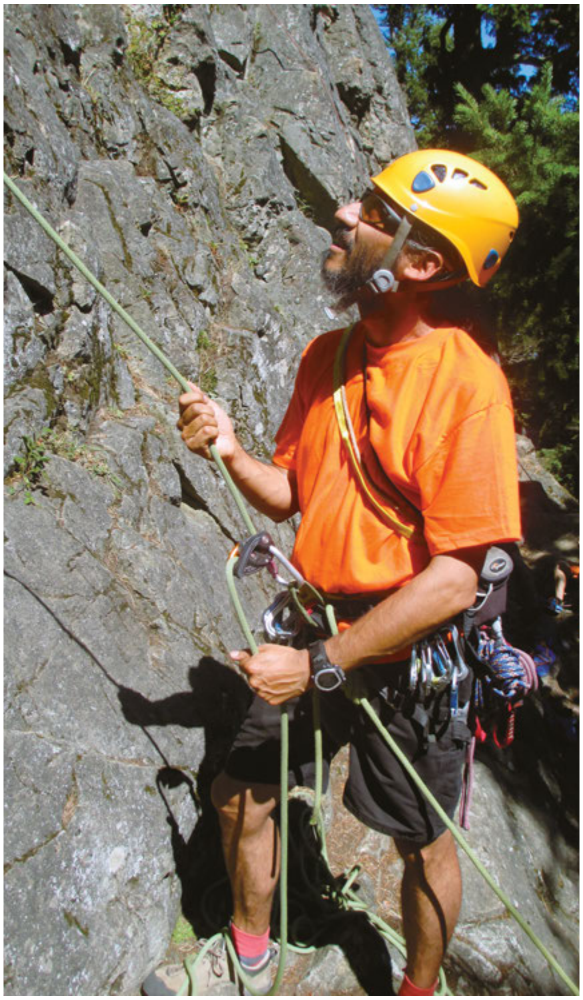

AMGA Single Pitch Instructor Richard Riquelme belays a leader using the principal belay position for a Grigri.

JASON D. MARTIN

Student Lead Belay Training

When an instructor works at a venue that requires a lead to access the anchors, it is important to add a lead belay contingent to the belay lesson. Once the PBUS (pull, brake, under, slide) technique has been taught and the student demonstrates proficiency, then it is appropriate for an instructor to move into a lead belay lesson.

The orientation of the student's hands while belaying a leader should reflect the posture taken in the braking position of the PBUS. The student pays out rope with the guide hand above the device, while the brake hand remains in the same position below the device. If the student needs to bring rope back in, he or she simply reverts back to the PBUS toproping technique.

There are instructors who prefer to have students belay them with an assisted braking device. The advantage to these devices is that they reduce the likelihood of catastrophic failure of the system. The problem with them is they are far from foolproof and require specialized instruction and technique.

There are a number of devices on the market, and they all have their own idiosyncrasies. It's important to read all associated instructions before

Leading with a Student Belayer

When you're leading, you should just assume that you're soloing.''

It was my (JDM) first year working as a climbing instructor, and this was one of the first things another young instructor shared with me. At the time it made sense. Indeed, in my first two years of instructing, I led literally hundreds of pitches with terrible student belays.

But then it came to me.

The student belays weren't terrible because the students were inexperienced belayers. The student belays were terrible because I didn't invest the time and energy in the students to ensure that their lead belay technique was flawless. From that time forward, I invested myself in ensuring that my lead belays were solid.

Approximately a year after I started spending more time with students on belay technique, I was teaching an aid-climbing clinic in Red Rock Canyon. I placed a piece behind a massive flake and bounce tested it.

It held.

As I worked my way up to the top of my aiders, the flake shifted and the piece popped. I dropped about 6 feet before the rope arrested my fall. The energy that I put into the belay lesson was not wasted. My student easily caught me.

using a new device, follow the manufacturer's recommendations, heed the manufacturer's warnings, and practice with the device prior to using in an institutional setting.

The Petzl Grigri is one of the more common devices on the market. As a result, lead belay technique with the Grigri is described here.

The primary belaying position for the Grigri is the PBUS position, with a guide hand above the device on the rope and a brake hand below. As a leader moves up the rock, the belayer slowly feeds rope through the device, gently pulling with the guide hand, while pushing rope through with the brake hand. If the rope is fed at an appropriate speed, the cam in the Grigri will not engage.

In this principal belay position, the belayer's brake hand never leaves the rope. If there is a need to bring in slack, the belayer reverts to the PBUS toproping technique.

Because the cam automatically engages with a sudden acceleration of the rope, it can be difficult to pay out slack quickly. The simplest solution to this problem is to never allow the rope to suddenly accelerate. This may be accomplished by the leader placing gear at chest level or lower and extending the protection with runners. Doing so allows the leader to clip into the protection without having to give a quick tug on the rope.

If the curricular goal is to teach a student the finer points of lead belaying, then there are two ways to give slack to a climber who needs it quickly. The first and easiest way is to simply step in toward the wall. This will immediately put slack into the system and works well. However, this technique is not recommended for novice belayers as they are often tied down to ground anchors and cannot easily move.

The second way is to shift the brake hand, sliding it up the rope to the device, and brace the index

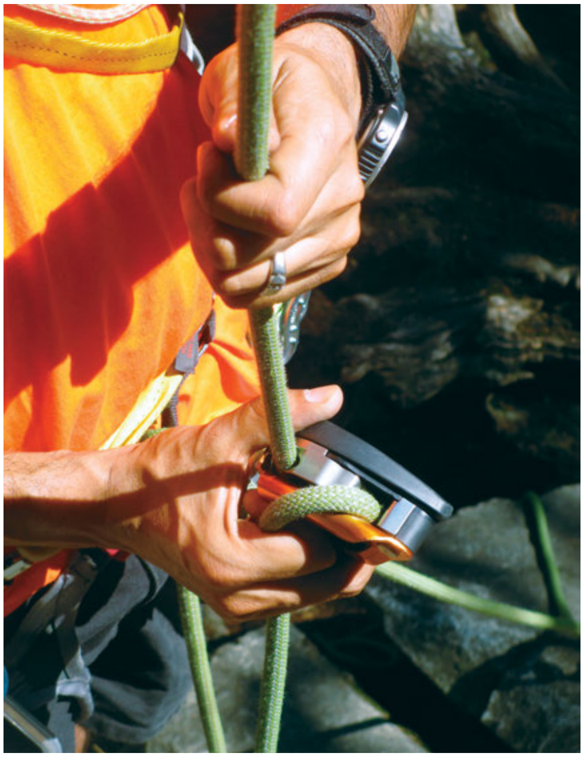

The proper way to give slack quickly with a Petzl Grigri.

finger against the lip of the moving sideplate. Press the thumb of the brake hand down on the cam where the handle is attached while continuing to hold the brake strand of the rope. Pull slack with the guide hand. Once finished, immediately return to the principal belay position.

Petzl recommends that you:

1. Always keep the brake strand in the brake hand. There is never a valid reason to let go of the brake strand.
2. Never grip the device with the entire hand.
3. Anticipate the lead climber's movement, including when additional rope is needed to make a clip.

In a toprope setting, a rope is generally set up early in the day and may be used to practice belaying. In a lead setting, practicing this skill requires some creativity. One method is to clip the first bolt of a sport route, or place a piece of gear about 10 feet up. Clip the rope and then have the student practice belaying a leader on this short mock setup.

Ground Anchors, Student Belay Backups, and Knots

In addition to using a Grigri and placing a lot of protection, here are three other ways to increase instructor security during a lead. First, use a ground anchor. Second, employ a backup belayer. And third, tie knots in the rope behind the belayer and the backup belayer.

A ground anchor keeps the belayer under control. The belayer is fixed to a given spot. If the belayer is anchored, the opportunity to trip, fall over, and pull the instructor off is greatly reduced. The belayer will remain in the designated stance.

With two or more students, a backup belayer will increase security. It is far less likely that both students will drop the leader. To add even greater security, put a friction hitch on the rope behind the belayer and attach it to the backup belayer's belay loop. Rather than being dependent on a hand belay, the backup belayer manages the rope with the assistance of a "third hand."

Some instructors tie knots in the rope behind the belayer and the backup belayer. As the instructor leads and the knots approach the belay team, either the backup belayer or, ideally, a third student unties them. Even if there are a series of mistakes, the leader will still have a reasonable margin of error.

No matter what steps are taken to increase instructor security, it remains important to regularly look down and check on the belayer. Make sure that the belay system is employed appropriately and communicate error corrections as needed.

Descent Options

If walking off or downclimbing is not possible for the climbing instructor, the other descent options from the top of a route are either to rappel or be lowered.

The most secure method is to rappel. When being lowered, the instructor is completely reliant on the belay system and at the greatest exposure to risk of system failure. If there are any doubts about the security of the system (i.e., the belayer), the instructor should rappel.

However, if the curricular goal of the program is to teach the student how to operate as an independent climber, then the student will have to learn how to lower. When faced with that situation, the instructor can mitigate the risk by placing a friction hitch on the belay strand of the rope as a backup, clipping the friction hitch to a sling that is then clipped into the instructor's belay loop with a locking carabiner. While being lowered, the instructor manages the friction hitch, releasing it if the belayer loses control of the brake strand.

Teaching a Student to Lead

When teaching lead climbing skills, it is important to know whether or not the organization offering the course permits students to lead. It is also important to find out if the insurance policy permits student leads.

Regardless of whether the organization or insurance policy allows leading, prospective leaders should spend a great deal of time mock leading. A mock leader is a climber who is on toprope while simultaneously dragging a lead line. The toprope protects the student from a real fall, while the lead line provides the illusion of leading. The student can place gear or clip bolts just as if leading for real, but without the consequences.

After a student has completed a mock trad lead, the instructor can climb the route and inspect the pieces. Ideally, the instructor will provide the student with feedback on the quality of each placement while removing the gear.

An instructor may also choose to have a student belay the mock leader while the instructor ascends a fixed line beside the leader, which will allow the instructor to give feedback in the moment. The instructor may use the same rope climbing technique as described for a climber pickoff (chapter 13, Basic Assistance and Rescue Skills).

If only one rope is available, it is still possible to set up a mock lead. Simply tie both ends of the rope into the climber. One end will come from the climber's harness, run up through the toprope anchor, and drop back down to the belayer. The rest of the rope will lay in a pile between the climber and belayer.

The advantage of the single line mock lead system is that only one rope has to be carried. However, there are two disadvantages with a single line. First, students will have to unclip each piece as they are lowered. There won't be enough rope to leave it running through the pieces. And second, with a single line system, the student may have a harder time distinguishing between the belay line and the lead line. When climbers have two ropes in a mock lead scenario, they're usually different colors. This makes it easier for students to distinguish between the lead line and the belay line.

Most instructors will spend a great deal of time in ground school, teaching students how to place gear or clip bolts before moving into a mock lead sequence. Sport leaders should understand how to clip correctly before getting on the rock, while trad leaders should have a solid understanding of gear placement. Both sport climbers and trad climbers need to be educated in placement or clipping strategies before working their skills on a mock lead.

If teaching a clinic on sport leading, be sure to spend significant time in ground school practicing the transition from a toprope to a rappel. This can be done using traditional pieces close to the ground. Clip locking carabiners to two solid pieces and treat them like bolts. The student will then be

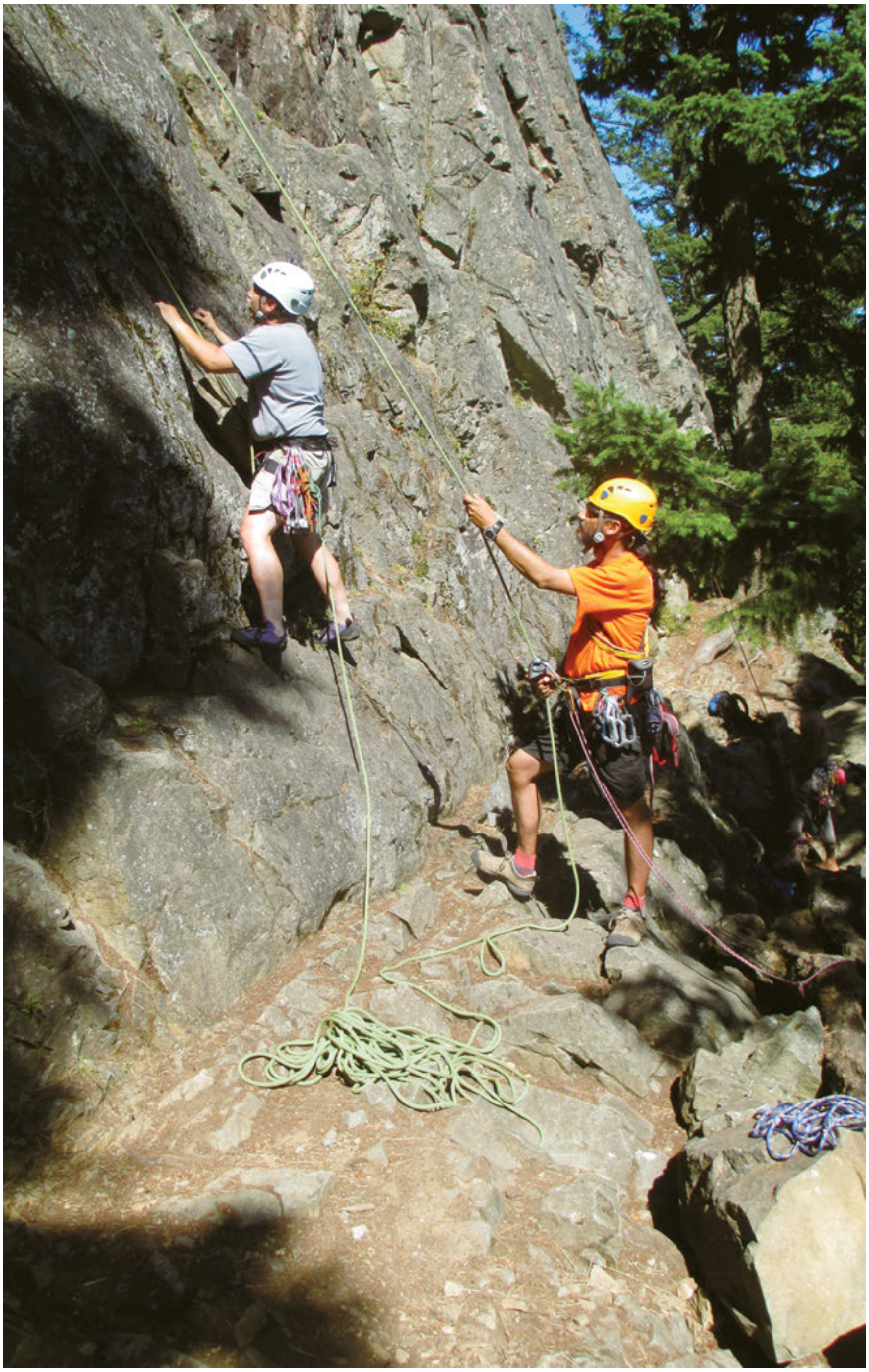

A climber mock leads on a single rope. Note the rope running up to the anchor above the climber, back down to the belayer, into the pile on the ground, and then back to the climber.

JASON D. MARTIN

able to practice the transition from start to finish.

If the instructor is in a position to allow students to lead for real, be extremely conservative. Beginning leaders should not fall. By choosing an appropriate crag—a crag where the student feels confident, or a very easy route, or even a 4th-class route—the exposure to potential student injury decreases dramatically.

Occasionally, an instructor may have a student who is an experienced climber and would like to lead. If organization policy and insurance allow this, then the instructor has to be the one to decide whether or not it's appropriate for the student to lead. This generally requires some assessment of both the student's movement skills and ability to clip or place gear appropriately. If an instructor doesn't like what he or she sees in a student assessment, then the instructor should not be afraid to say no. It's better to have an upset student than a seriously injured student.

Appendix A: Leave No Trace Ethics

Climbers should practice Leave No Trace principles from the moment they step out of the car. The following simple steps will help keep the climbing sites we all share as clean as possible, with minimal degradation to the climbing area and the surrounding environment:

- At popular climbing areas, use the outhouses located at most parking areas before you embark on your approach to that day's chosen cliff.
- Always use marked climber's access trails where they are available. If there is no marked trail to the cliff, minimize your impact by walking on durable surfaces (e.g., a sandy wash, rock slab, or barren ground).
- At popular, easy-access crags, avoid making a beeline from the parking lot straight to the crag without first looking for an established path or trail. Walking off-trail can significantly impact vegetation and cause soil erosion if enough people do it over a period of time.
- If traveling in a group in more remote, pristine areas where no trail exists, fan out instead of walking in single file and try to walk on the most durable surfaces, avoiding fragile vegetation. Don't leave rock cairns to mark the path, as this takes away the challenge of routefinding from those who prefer to experience it on their own terms.
- If nature calls and you're far from any outhouse, deposit solid human waste well away from the base of any climbing site, water source, or wash by digging a cathole 4 to 6 inches deep. Cover and disguise the cathole when you're done. Pack out all toilet paper and tampons in a ziplock bag. Better yet, use "blue bags" and pack out all human waste. Urinate on bare ground or rock, not plants. Urine contains salt, and animals will dig into plants to get at it.
- Leave no trace means just that—pack out everything you bring in, including all trash and food waste (that means apple cores and orange peels, too). Set an example for your group by picking up any trash you find; plan ahead and always carry a trash bag with you when you go out to the crag.

Principles of Leave No Trace

1. Plan ahead and prepare
2. Travel and camp on durable surfaces
3. Dispose of waste properly
4. Leave what you find
5. Minimize campfire impacts
6. Respect wildlife
7. Be considerate of other visitors

- Don't monopolize popular routes by setting up a toprope and then leaving your rope hanging on the climb, unused. If your climb begins from a campsite, ask permission to climb from the campers if the site is occupied. Minimize your use of chalk, and if you're working a route, clean off any tic marks with a soft brush after you're done. Protect everyone's access to a climbing area by being courteous, beginning with parking only in designated areas and carpooling whenever possible. Noise pollution can be a problem, from blasting tunes on a boom box to yelling and screaming while attempting a hard climb. Be considerate and aware of those around you, and limit your noise production to a reasonable level.
- Pick up all food crumbs and don't feed any wild critters—this habituates them to human food and encourages them to beg and scavenge, sometimes even chewing holes in backpacks to get at food.
- Leave all natural and cultural objects so that they can be experienced by everyone in their natural setting. If you are climbing in a national forest or national park, obey all regulations concerning the gathering of firewood and other objects.

For more information on Leave No Trace ethics, visit www.LNT.org.

Appendix B:
How to Determine a Fall Factor

Fall factors are determined by a simple formula. Divide the length of the fall by the length of rope out between the belayer and climber.

The greater the fall factor, the greater the force on the rope, on the belayer, and on the anchor. A factor 2 fall is the maximum that may be attained in a standard climbing fall, since the length of a fall can't exceed twice the length of the rope.

In a multi-pitch setting, a factor 2 fall only takes place when a leader who has not placed any protection takes an unobstructed fall, drops past the belayer, and then falls directly onto the belayer's system. If the climber places protection above the belayer, it is no longer possible to generate a factor 2 fall, as the length of the fall can no longer be twice the length of the rope out.

Many other factors play into the calculation of a true fall factor. For example, if an individual tumbles down a slab, the impact will be quite different than if he fell unobstructed. If the belayer is pulled off the ground or a piece fails, all of these things change the true math.

A factor 2 fall is extremely serious. The forces generated can exceed the strength rating of gear, potentially resulting in anchor failure. Also, the impact force on the belayer is tremendous. Loss of control of the belay or injury to the belayer are both possible outcomes. As such, it is imperative

Fall Factors

Following are some simple word problems that you can use to determine if you understand how to calculate fall factors:

1. A lead climber places a piece 15 feet off a belay. As he reaches 20 feet, he falls, dropping 10 feet. What type of fall factor did he generate?
2. A lead climber on a multi-pitch climb leads 50 feet up above his belayer without placing any gear. He slips and drops 50 feet back to his belayer and then continues to fall another 50 feet, for a grand total of 100 feet. What type of fall factor did he generate?
3. A lead climber is 50 feet up. His last piece is 10 feet below him. He falls. What kind of fall factor did he generate?

Answers: (1) Factor 0.5, (2) Factor 2, (3) Factor 0.4

that multi-pitch climbers use the anchor as the first piece of protection or place gear immediately after leaving the belay station. This will limit the possibility of a high-impact fall.

Single Pitch Instructors are not as concerned about fall factors as much as multi-pitch guides, but there is still cause for concern.

A common practice among climbers is to clip into the anchor with a personal tether. Personal tethers include slings, a PAS, or a daisy chain. If the climber is positioned above the anchor, clipped in with a personal tether, and falls, a high-impact fall—potentially as high as a factor 2 fall—will occur. Due to the static nature of tether material, there is a possibility that the personal tether will break. Some tether materials, like nylon, do better in testing than others, but a fall onto a tether is still a concern that has to be kept in mind. Additionally, the impact force generated to the climber's body can easily result in an injury. The takeaway point here is that a climber does not need to be tied into a rope to suffer a factor 2 fall.

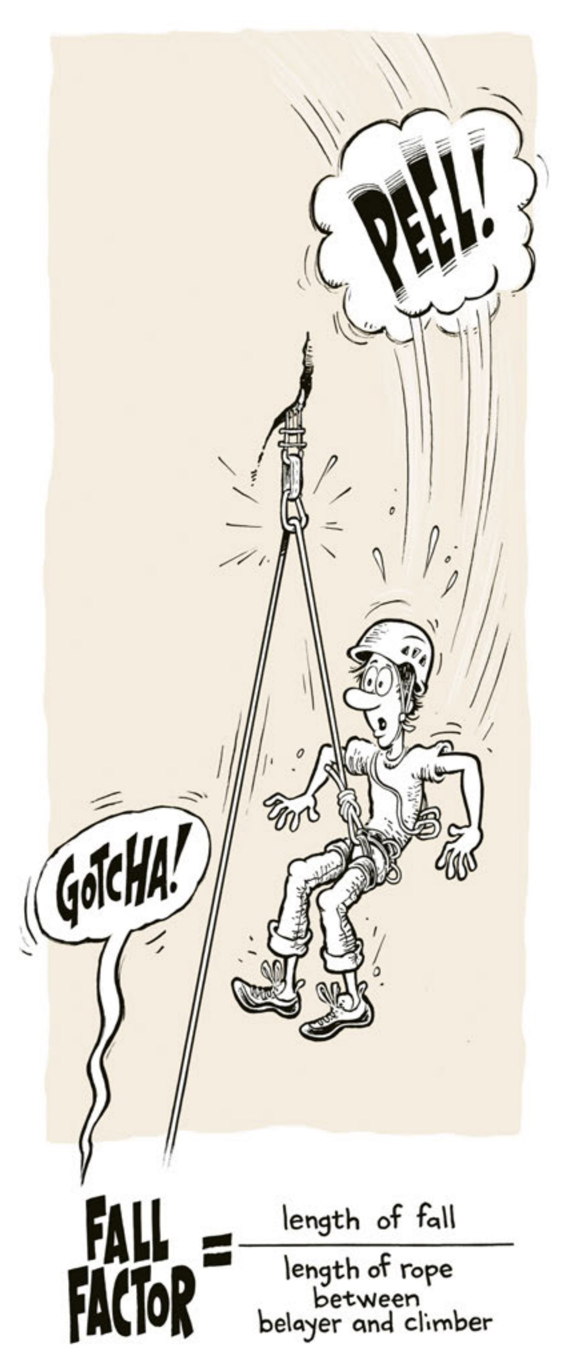
PEEL!
GOTCHA!
FALL FACTOR = length of fall / length of rope between belayer and climber

Calculating the fall factor.

MIKE CLELLAND

Appendix C: Autoblocking Devices

Autoblocking devices like the Petzl Reverso, the Black Diamond ATC Guide, the Mammut Bionic Alpine, the CAMP PIU 2, and the myriad of others on the market are not covered in the Single Pitch Instructor curriculum. These devices are excellent multi-pitch devices and work well in a single pitch environment. However, many new instructors come into the AMGA Single Pitch Instructor Course with limited experience working with assisted braking devices and the Munter hitch. It is extremely important for instructors to understand how to use these tools, since they are the best tools for Single Pitch Instructor terrain. Students who move into the upper-level AMGA courses spend a great deal of time working with autoblocking devices, but at the beginner level, instructors must first have a baseline understanding of essential tools.

Though autoblocking devices are not part of the curriculum and candidates will not be tested on them, it was felt that they should be covered in this text. This is primarily because autoblocking devices are increasingly common at both single and multi-pitch crags throughout the country, and most students have been exposed to them.

Autoblocking devices are a hybrid between standard tube-style devices and plaquette-style devices. They can be used to belay a leader or to toprope, just like any other tube-style device. However, they have one additional feature. When belaying from above, they can be hung directly from the anchor by an additional carabiner hole and rigged to securely belay one or two climbers.

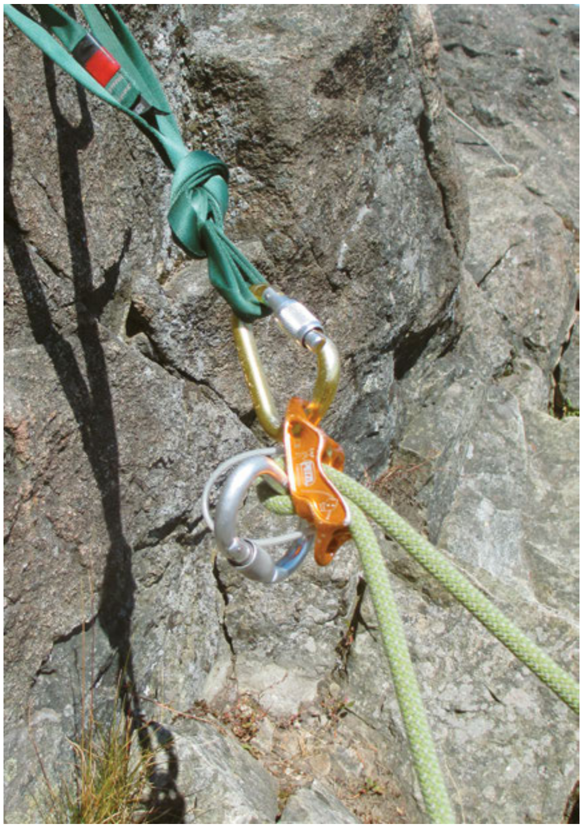

An autoblocking device set up in "guide mode," loaded with the climber strand on the top and the belay strand on the bottom.

As with all climbing equipment, it's important to read the manufacturer's instructions. These devices vary and there are small differences between them. But the concept is the same. A bight of rope runs through the vertical slot on the front of the device, with the strand on top going to the climber and the belay strand on the bottom. This bight is secured to the back of the device with a locking carabiner. When weighted, the strand going to the climber squeezes the belay line, holding it in position and locking it off.

If these devices are loaded improperly, the autoblocking function will not engage. Always pull down on the climber's line after loading the device to ensure that it locks. If the climber's line moves easily, the device is not loaded properly.

Another reason this device is not taught in the SPI curriculum is that, when set up in autoblocking mode, it requires specialized techniques to unlock the device when it is loaded. It is important to understand that these devices, when in autoblock mode, were not designed for long lowers and are a poor choice for lowering a climber in single pitch terrain. Situations may arise that require unlocking an autoblocking device to lower a climber a short distance, and users should understand how that can be accomplished. Following are three methods.

Simple Lower by Ratcheting Carabiner

The first type of lower requires that a belayer crank the locking carabiner that secures the bight of rope back and forth. This allows the load to lower incrementally. This method is used for very short lowers. It is important for the belayer to keep the brake hand on the brake strand of the rope during this process.

Simple Lower with a Nut Tool or Carabiner

In the second lower, a nut tool or small carabiner is used to leverage the device. This method works well for lowers of one or two body lengths. Following is a breakdown of the lower:

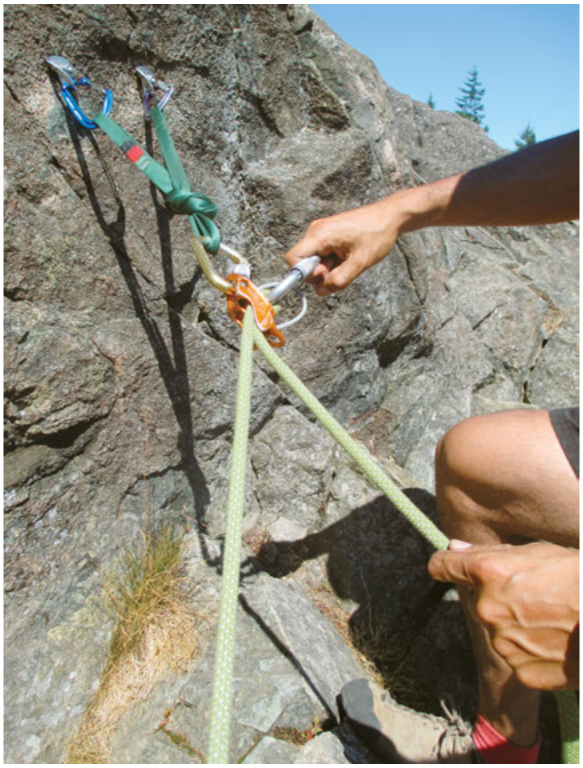

A belayer ratchets the carabiner up and down to complete a short lower.

JASON D. MARTIN

1. Attach a friction hitch to the brake strand and clip it to the belayer's belay loop.
2. Put a nut tool or carabiner in the small hole in the bottom of the autoblocking device.
3. Crank the device up and back, "breaking" the autoblock's hold on the rope.
4. Maintain control of the brake strand while allowing the rope to pass slowly through the device.
5. Once the lower is complete, return the device to its original position.

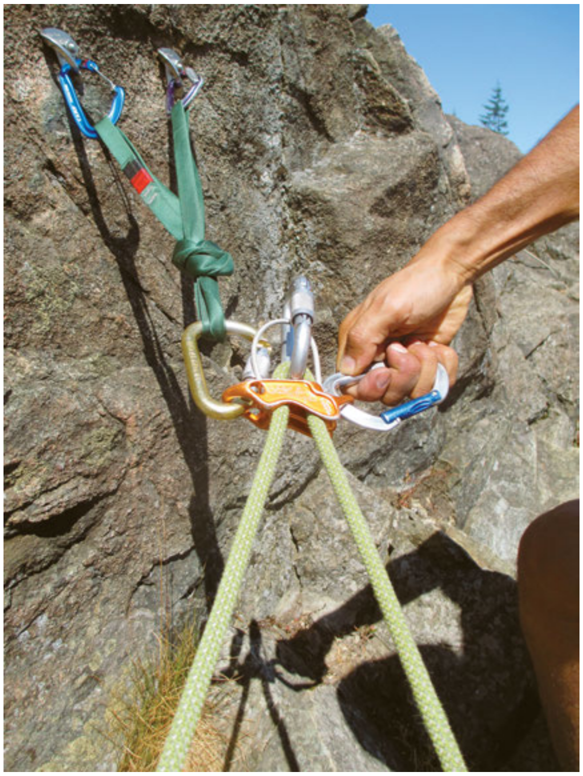

Using a carabiner to open up the device in order to make a short lower.

JASON D. MARTIN

Complex Lower with a Redirect

This method can be used for moderate-length lowers. Again, if a longer lower is required, the system should be transferred to a more secure system. Following is a breakdown of the lower.

1. Tie off the backside of the rope with a catastrophe knot.
2. Attach a friction hitch to the brake strand and clip it to the belayer's belay loop.
3. Place a carabiner somewhere high in the anchor system, above the autoblocking device.
4. Redirect the brake strand through the carabiner high in the system.
5. Run a cord or a sling through the small hole in the bottom of the autoblocking device.
6. Thread the cord or sling through a second carabiner high in the system and attach it to the belayer's belay loop.
7. Untie the catastrophe knot.
8. Lean back on the redirected cord, "breaking" the system.
9. Maintain control of the brake strand while allowing the rope to pass slowly through the device.
10. Once the lower is complete, return the device to its original position.

Appendix D: Wilderness Medical Certifications

AMGA Single Pitch Instructors are required to hold and maintain a Wilderness First Aid (WFA) certificate if they intend to work in the frontcountry, or a Wilderness First Responder (WFR) certificate if they intend to work in the backcountry. If it will take you less than 2 hours to evacuate a student to definitive medical care, then you are working in the frontcountry. However, if it will take more than 2 hours, then you are working in the backcountry and you will need to hold a WFR certification.

A number of organizations throughout the United States offer WFA and WFR certifications, though not all certifications are the same. Some require recertification every two years, while others require it every three. Some companies have reciprocity with one another, while others do not. Be sure to do your research before committing to a certification course. You should know what you're getting into ahead of time and should be aware of how easy—or how hard—it will be to find recertification courses when your certification expires.

In addition to a WFA or WFR certification, instructors are required to hold and maintain a CPR certification.

Following is a breakdown of common wilderness first-aid medical certifications.

Wilderness First Aid (WFA)

This 16- to 30-hour program provides a baseline introduction to wilderness medicine. This course is recognized by the AMGA as a minimum requirement for individuals working in a frontcountry environment.

Wilderness First Responder (WFR)

This 70- to 90-hour certification course is considered a baseline requirement for most backcountry professionals in the United States. The AMGA recognizes this as a minimum level of training for those working in a backcountry wilderness environment.

Outdoor Emergency Care (OEC)

The National Ski Patrol developed the OEC as a first-aid program for members of the ski patrol who work at resorts. This program requires students to devote 80 to 100 hours to the course of study. The AMGA recognizes this as an equivalency to the Wilderness First Responder, as do many land managers.

Wilderness Emergency Medical Technician (WEMT)

The WEMT course meets all the requirements of an EMT course as well as the requirements of a WFR course. Most programs require four weeks of full-time intensive study. The AMGA recognizes this as a higher level of training than the WFR program.

Appendix E: Recommended Reading

Climbing instructors should have a strong understanding of the contents of a wide array of instructional climbing texts. The following are excellent resources for the Single Pitch Instructor candidate. The seasoned certified instructor may find these texts useful to supplement lessons.

Baseline Climbing Skills

Climbing Anchors. John Long and Bob Gaines. Globe Pequot Press/FalconGuides. Guilford, CT. 2013.

Rock Climbing Anchors: A Comprehensive Guide. Craig Luebben. The Mountaineers Books. Seattle, WA. 2007.

Rock Climbing: Mastering the Basic Skills. Craig Luebben. The Mountaineers Books. Seattle, WA. 2004.

Toproping. Bob Gaines. Globe Pequot Press/FalconGuides. Guilford, CT. 2012.

Rappelling. Bob Gaines. Globe Pequot Press/FalconGuides. Guilford, CT. 2013.

Advanced Climbing and Instructional Skills

Coaching Climbing. Michelle Hurni. Globe Pequot Press. Guilford, CT. 2002.

The Rock Warrior's Way: Mental Training for Climbers. Arno Ilgner. Desiderata Institute. La Vergne, TN. 2006.

Self Rescue, 2nd ed. David Fasulo. Globe Pequot Press/FalconGuides. Guilford, CT. 2011.

Technical Handbook for Professional Mountain Guides. American Mountain Guides Association. Boulder, CO. 1998.

Adventure Activities / Outdoor Leadership / Education

The Backcountry Classroom: Lessons, Tools and Activities for Teaching Outdoor Leaders. Jack Drury, Bruce Bonney, Dene Berman, and Mark Wagstaff. ICS Books. Merrillville, IN. 1992.

Best New Games. Dale N. LeFevre. Human Kinetics. Champagne, IL. 2002.

Cowstails and Cobras II: A Guide to Games, Initiatives, Ropes Courses & Adventure Curriculum. Karl Rohnke. Project Adventure, Inc. Hamilton, MA. 1998.

Islands of Healing: A Guide to Adventure Based Counseling. Jim Schoel, Dick Prouty, and Paul Radcliffe. Project Adventure. Beverly, MA. 1989.

Outdoor Leadership: Techniques, Common Sense and Self-Confidence. John Graham. The Mountaineers Books. Seattle, WA. 1997.

Soft Paths. David Cole, Bruce Hampton, and Dana Watts. Stackpole Books. Harrisburg, PA. 1998.

Appendix F: Climbing Instructor Employment Resources

The following resources may be helpful when searching for a job in the outdoor industry. Many of these provide a variety of job opportunities for outdoor educators, instructors, and guides, so pay close attention to find the job that is the right fit.

AMGA Newsletter

www.amga.com

After completing the AMGA Single Pitch Instructor Course, members will be placed on the American Mountain Guides Association mailing and e-mail lists. The AMGA's monthly newsletter includes listings for mountain guide and climbing instructor jobs.

The AMGA job board tends to lean heavily toward for-profit guide services.

AORE Listserv

www.aore.org

The Association of Outdoor Recreation and Education was designed to bring together outdoor professionals in college recreation and outdoor pursuits programs and community and military recreation programs for an annual conference. The yearly conference was designed for academic outdoor professionals to network and trade ideas.

AORE members have access to a listserv that has an academic focus, and many of the jobs are at universities or community colleges.

There is an annual fee to become a member of the AORE community.

Association for Experiential Education

www.aee.org

This is another association that has an academic focus. Like the AORE job announcements, this organization posts a wide array of jobs, from those at schools, to challenge courses, to climbing walls.

Backdoor Jobs

www.backdoorjobs.com/adventure.html

The Backdoor Jobs website is devoted to adventure tourism jobs including climbing and mountaineering, but also covers rafting, backpacking, and mountain biking guides.

Certified Guides Co-op

www.certifiedguidescooperative.com

The Certified Guides Co-op provides Single Pitch Instructors with significant professional work history an opportunity to gain access and insurance. The Co-op is designed for instructors with an established client base who wish to work independently of a guide service, school, or other organization. Instructors must have at least 200 days of single pitch instruction after they become certified instructors to apply.

Chronicle of Higher Education

www.chronicle.com

This is a university job board that focuses heavily on academic jobs.

Military Recreation Jobs

www.armymwr.com;
www.navymwr.org;
www.nafjobs.org;
www.usajobs.gov

As the military is extremely broad, so are the job resources. These are a select few websites where you may find outdoor recreation job opportunities on and around military bases.

Mountain Guide Jobs

www.mountainguidejobs.com

This is a site that provides a very short list of organizations and companies that are looking for climbing and mountaineering guides.

NOLS Jobs Announcements

www.nols.edu/alumni/contact/listservices.php

The NOLS Jobs Announcements is a listserv designed for outdoor professionals. It provides information on some NOLS-related jobs, but the majority of the listings are for adventure education jobs throughout the United States and Canada.

This listserv was designed with NOLS and Wilderness Medicine Institute alumni in mind. However, it does not discriminate against those who have not taken a NOLS course.

This is an extremely active job board, with dozens of new postings coming through every week.

Outdoor Adventure Professional Network

http://adventurejobs.co/

This particular site posts jobs in all the outdoor disciplines.

Outdoor Ed

www.outdoored.com

The Outdoor Ed website is a job-listing website that provides information on both seasonal and year-round adventure tourism jobs. Mountaineering and rock instruction jobs are regularly highlighted on the site.

Outdoor Industry Jobs

www.outdoorindustryjobs.com

This website is a dumping ground of outdoor industry–related jobs. There are some adventure educator and guide jobs on the site, but many of the jobs are retail and gear representation.

Rock Climbing Jobs

www.rockclimbingjobs.com

This site focuses on rock climbing jobs at adventure summer camps.

Northwest Information Exchange

http://groups.google.com/group/nw-information-exchange?pli=1

The Northwest Information Exchange is a place where instructors and guides in the Pacific Northwest trade information about current conditions in the mountains. However, guide service owners and staffing managers regularly post open positions on the listserv.

You will need to be actively working as a guide or instructor for a guide service to get access to the listserv.

San Juan Information Exchange

https://groups.google.com/forum/?hl=en-US#!forum/San-Juan-information-exchange

This is a guide conditions listserv like the Northwest Information Exchange. However, the focus of this exchange is the Ouray/Telluride/Silverton/Durango area of Colorado. Occasionally employment requests are sent out through the listserv.

Like the Northwest Exchange, you must be actively working in the field for a guide service to obtain access to the site.

Index

American Triangle, 162–63
anchoring
  to detached blocks, 102–3, 118, 126
  rock assessment and, 102–7
anchoring systems, 144–63
  about: overview of, 144–45
  anchor anatomy/principles (CAMS), 145–46
  Backside System, 175–79
  BARCK check, 185
  checking, 184–85
  CRASH check, 185
  creating master point over cliff edge, 167
  extensions with tether, 167
  instructor tether, 165–67
  Joshua Tree System, 168–72
  Lifeline System, 173–75
  redundancy and, 146–48, 181
  Three-in-One System, 179–81
  using setup rope, 164, 167
  V rigging with extension rope, 168–72
  V rigging with instructor tether, 173–75
  vectors and, 162–63
  See also load distribution
anchors
  anatomy and principles (CAMS), 145–46
  bolts, 136–43
  checking, 184
  fixed, 135–43
  monolithic, 112
  natural, 108–12
  pitons, 135–36
  tethering to, 218–20
assessing rocks, 102–7
assistance and rescue skills
  assistance from base, 221–27
  assistance from top, 227–30
  climber pickoff, 221–27
  coaching, 227
  3:1 assisted raise, 230
  3:1 raising system, 227–29
  vector pull, 227
assisted braking devices (ABDs), 195–96, 202–3, 204–5, 210
auditory (verbal) learners, 19
autoblock (“third hand”)
  as backup when lowering, 205–7
  function and tying, 100
  as rappel backup, 213–16
  rigged into belay loop, 215–16

backpacker (butterfly) coil, 39, 40–41
Backside System, 175–79
backup belayers, 199–201
BARCK check, 185
base-managed sites, 190–92
  advantages and disadvantages of, 190
  assistance from. *See* assistance and rescue skills
  backup belayers, 199–201
  belaying in, 196–98
  ground anchors and, 190–92, 236
  for high-ratio instruction, 190
  perspective on top-managed and, 187
“Belay off” signal, 195
“Belay on” signal, 194
belaying
  autoblock as backup when lowering, 205–7
  backup belayers, 199–201
  in base-managed sites, 196–98. *See also* base-managed sites
  devices for, 192–96
  fundamental principles of, 192
  hip belay, 193
  indirect belay, 203
  leading guidelines. *See* leading
  lowering from top-managed site, 204
  lowering with assisted braking device, 204–5
  lowering with Munter hitch, 204, 205–7
  with Munter hitch, 201, 202–3
  PBUS method, 234–35
  redirected belay, 203
  standard signals for, 194–95
  systems checks, 184–85
  in top-managed sites, 201–4. *See also* top-managed sites
  two-rope toprope setups, 199
  See also rappelling
bends, 80–82
  about: purpose of, 80
  double fisherman’s knot, 81
  figure eight bend (Flemish bend), 80, 83
  triple fisherman’s knot, 82
BHK (“Big Honking Knot”)
  Backside System and, 175–77
  function and tying, 72–73
  loading backside of, 177
bolts, 136–43

bowline, 70–71
bowline, double loop, 76–77
butterfly (backpacker) coil, 39, 40–41

camming devices (SLCDs), 123–35
CAMS, 145–46
carabiners, 52–61
  array of designs/applications, 52
  autolock or twistlock, 56–57
  basic design, 52–53
  D-shaped, 54, 57
  loading guidelines, 57–60
  locking, 52, 53–57
  oval, 52, 53, 57, 59
  pear-shaped, 52, 54, 55
  screwgate, 54–56
  strength ratings, 57
  tying Munter hitch on, 88–89
  when to retire, 61
  wiregate, 52, 53
catastrophe knot, 201, 246
children
  liability and minors, 10
  minimum age for, 24
  taking kids climbing, 24–26
chockcraft evolution, 112–23
"Climb on" signal, 194
climber pickoff, 221–27
climbing instructor
  instructor meetings, 16
  professionalism, 3–5
  risk management. *See risk management references*
  role of, 1
  stewardship, 14–16
  teaching skills. *See* teaching, art of
climbing movement
  bipeds becoming quadrupeds, 26–27
  teaching, 26–27. *See also* teaching, art of
  terrain-specific movement, 27
"Climbing" signal, 194
climbing signals, standard, 194–95
closing the system, 184
clove hitch, 85, 86
coaching, 227
coiling and uncoiling ropes, 39–43
communications plan, 9
contacts, post-trip, 10
cordelettes and cords
  anchoring solution, 151–54
  instructor tether, 165–67
  the quad, 157–58
  rigging, 153
  types and characteristics, 50–52
  See also ropes
CRASH check, 185

debrief, incident, 10
descent options, 237
directional (in-line) figure eight, 68–69
double fisherman's knot, 78–79, 81, 83
double loop knots, 72–78
  BHK ("Big Honking Knot"), 72–73
  double loop bowline, 76–77
  double loop figure eight, 74–75
dry coating ropes, 37–38
dynamic ropes, 34–35, 36
Dyneema and Spectra slings, 47–49

elevators, 209
equalette, 160–62
equipment basics, 28–61
  autoblocking devices, 244–46
  belay devices, 192–96
  harnesses, 28–30
  helmets, 31–32
  knives, 44
  pack and its contents, 61
  slings and webbing, 44–49
  See also carabiners; ropes
ethics, leave no trace, 240–41
Euro Death Knot (flat overhand), 84–85
evacuation plan, 9
extensions
  creating master point over cliff edge, 167
  pre-equalized (Backside System), 175–79
  V rigging with, 168–72

failure, fear of, 12
fall factor, determining, 242–43
falling hazards, 185–86
"Falling!" signal, 195
fear, of heights/of failure, 12
figure eight bend (Flemish bend), 80, 83
figure eight follow-through, 64–65
figure eight loop, 66–67. *See also* double loop knots; in-line (directional) figure eight
financial risk, 12–14
fireman's belay, 208
first-aid requirements, 9
fisherman's knots
  double, 78–79, 81, 83
  triple, 82
fixed lines, 231–33
flakes, avoiding, 102–7, 118, 126
flat overhand (Euro Death Knot), 84–85
Flemish bend (figure eight bend), 80, 83

grapevine knot. *See* double fisherman's knot
ground anchors
  base-managed sites and, 190–92
  checking, 184
  student belay backups, knots and, 236

harnesses
  about, 28–30

- checking, 184–85
heights, fear of, 12
helmets, 31–32
Hexcentric (Hex), 116, 119, 121, 122
hip belay, 193
hitches. *See knots and hitches references*

inattentional blindness, 183
incident debrief, 10
indirect belay, 203
in-line (directional) figure eight, 68–69
instructional site
- base-managed. *See* base-managed sites
- professionalism at, 4–5
- for rappelling, selecting, 208
- top-managed. *See* top-managed sites

instructor meetings, 16
instructor tether, 165–67

Joshua Tree System, 168–72

kernmantle ropes, 32–34
kinesthetic learners, 19
klemheist knot, 98–99
knives, 44
knots and hitches, categories
- about: knot terminology definitions, 62
- bends, 80–82
- double loop knots, 72–78
- hitches, 85–100
- for joining two ropes, 83–85
- loop knots, 62–71
- stopper knot, 101
- for webbing (water knot and double fisherman's knot), 78–79

knots and hitches, specific
- BHK ("Big Honking Knot"), 72–73
- bowline, 70–71
- catastrophe knot, 201, 246
- clove hitch, 85, 86
- double fisherman's knot, 78–79, 81, 83
- double loop bowline, 76–77
- double loop figure eight, 74–75
- figure eight bend (Flemish bend), 80, 83
- figure eight follow-through, 64–65
- figure eight loop, 66–67
- flat overhand (Euro Death Knot), 84–85
- in-line (directional) figure eight, 68–69
- klemheist knot, 98–99
- mule knot, 92–97
- overhand loop, 62–63
- prusik knot, 97–98
- slip hitch, 85, 87
- stopper knot, 101
- triple fisherman's knot, 82
- tying off tube device on carabiner spine, 96
- tying off tube device with mule knot in front of device, 94–95
- water knot, 78, 79
- *See also* autoblock ("third hand"); Munter hitch

leading, 234–39
- decent options, 236
- ground anchors, student belay backups, and knots, 236
- student-lead belay training, 234–36
- teaching student to lead, 237–39

learning from mistakes, 22–23
learning styles, 17–19
leave no trace ethics, 240–41
lesson planning, 20–22
lessons, harsh, 22
liability
- guidelines/exposure to, 11–12
- minors and, 3
- release forms, 10–11

Lifeline System, 173–75
load distribution, 148–62
- about: overview of, 148–51
- cordelette solution, 151–54
- equalette for, 160–62
- the quad for, 157–58
- sliding X for, 154–56
- stacked sliding Xs for, 159–60

loop knots, 62–71
- bowline, 70–71
- figure eight follow-through, 64–65
- figure eight loop, 66–67
- in-line (directional) figure eight, 68–69
- overhand loop, 62–63

loop strength, 47
"Lower me (or Ready to lower)" signal, 194–95
lowering
- with assisted braking device, 204–5
- autoblock as backup when, 205–7
- with Munter hitch, 204
- from top-managed sites, 204

"Lowering" signal, 195
low-stretch ropes, 36, 37

magic X. *See* sliding X
medical certifications, wilderness, 247
minors. *See* children
mistakes
- avoiding. *See risk management references*
- learning from, 22–23

mountaineer's coil, 39, 42–43
movement. *See* climbing movement
mule knot, 92–97

Munter hitch
  belaying with, 201, 202–3
  lowering with, 204, 205–7
  Munter/mule/overhand combination, 211, 212
  tying, 85, 88–91

natural anchors, 108–12
nut placement, assessing, 118

objectives (unit and student-learning), 20, 21
"Off belay" signal, 195
"On belay?" signal, 194
Outdoor Emergency Care (OEC) program, 247
overhand loop, 62–63

pack, of instructor, 61
PBUS method, 197, 234–35
pedagogy. *See* teaching, art of
permitting, 7–9
physical risk, 14. *See also risk management references*
pitons, 135–36
post-trip contact, 10
professionalism, 3–5
  at instructional site, 4–5
  pre-program preparations, 3–4
programming, 7–16
  about: overview of, 7
  instructor meetings, 16
  permitting, 7–9
  stewardship, 14–16
  See also risk management references
protection
  active, 123–35
  avoiding flakes, 102–7, 118, 126
  camming devices (SLCDs) for, 123–35
  chockcraft evolution, 112–23
  passive, 112–23
  rock assessment, 102–7
  See also anchoring/anchors references
prusik knot, 97–98
psychological risk, 12

quad, the, 157–58

rappelling, 208–20
  avoiding accidents (3 scenarios), 216–20
  extended master point for belay strand, 210
  fireman's belay and, 208
  instructor backup, 210–16
  not clipping both strands on double-rope rappel, 217
  not tethering to anchor, 218–20
  off one/both ends of rope, 217
  rigging for contingencies, 208
  site selection, 208
  student rappel belays, 208–10
  systems check, 217
  teaching, 208
  transitioning from rigging to, 175–79, 181
  troubleshooting device/rope problems, 210
reading materials, additional, 248
redirected belay, 203
redundancy, 146–48, 181
release forms, 10–11
rescue skills. *See* assistance and rescue skills
resources, other
  instructor employment resources, 249–51
  reading materials, 248
risk management
  communications plan, 9
  evacuation plan, 9
  first-aid requirements, 9
  incident debrief, 10
  minors precaution, 10
  post-trip contact, 10
  risk/liability release forms, 10–11
  screening participants, 9
  at site. *See* risk management, in field
  student safety briefing, 10
risk management, in field, 182–86
  avoiding repelling accidents, 216–20
  closing the system, 184
  falling hazards, 185–86
  fear of failure, 12
  fear of heights, 12
  financial risk, 12–14
  importance of, 182–83
  inattentional blindness and, 183
  other considerations, 186
  physical risk, 14
  psychological risk, 12
  rockfall hazards, 183
  sociological risk, 12
  standard climbing signals, 194–95
  systems checks, 184–85, 194, 217
  terrain hazards, 184
rock assessment, 102–7
"ROCK!" signal, 195
rockfall hazards, 183
rockslide, 103
role of climbing instructor, 1
rope direct belay, 146–47
ropes, 32–43
  care and use, 38–39
  coiling and uncoiling, 39–43
  diameter and sheath percentage, 36–37
  dry coating, 37–38
  dynamic, 34–35
  dynamic vs. low-stretch, 36
  fixed lines, 231–33
  history of, 32

  - kernmantle, 32–34
  - knots for joining, 83–85
  - length of, 38
  - low-stretch, 36, 37
  - sharp edge resistance, 37
  - static, 35–36
  - static elongation, 37
  - *See also* cordelettes and cords

- safety. *See* risk management
- screening participants, 9
- setup rope
  - with components far apart, 164
  - systems using, 167
  - two-rope toprope setups, 199
- signals, standard, 194–95
- site. *See* instructional site
- "Slack" signal, 194
- SLCDs (camming devices), 123–35
- sliding X, 154–56
- sliding Xs, stacked, 159–60
- slings and webbing, 44–49
  - flat webbing, 44
  - girth-hitching slings together, 49
  - knots for webbing (water knot and double fisherman's knot), 78–79
  - masking tape splice precaution, 47
  - Spectra and Dyneema slings, 47–49
  - strength ratings (tensile vs. loop strength), 47
  - tubular webbing, 44–47
- slip hitch, 85, 87
- sociological risk, 12
- Spectra and Dyneema slings, 47–49
- stacked sliding Xs, 159–60
- static elongation, of ropes, 37
- static ropes, 36
- stewardship, 14–16
- stopper knot, 101
- strength ratings
  - carabiners, 57
  - ropes (tensile vs. loop strength), 47
- student lead belay training, 234–36
- student rappel belays, 208–10
- student safety briefing, 10
- swami belt, 28
- systems checks, 184–85, 194, 217

- "Take (or tension)" signals, 194
- teachable moments, 23
- teaching, art of, 17–27
  - about: overview of, 17
  - climbing movement, 26–27
  - learning from mistakes, 22–23
  - learning styles and, 17–19
  - lesson planning, 20–22
  - objectives (unit and student-learning), 20, 21
  - qualities of good teachers, 26
  - taking kids climbing, 24–26
  - teachable moments and, 23
- tensile strength, 47
- "Tension (or take)" signals, 194
- "Tension on (or I've got you)" signal, 194
- terrain hazards, 184
- tether, instructor, 165–67
- tethering to anchor, 218–20
- "That's me!" signal, 195
- "third hand." *See* autoblock ("third hand")
- 3:1 assisted raise, 230
- 3:1 raising system, 227–29
- Three-in-One System, 179–81
- top-managed sites, 187–90
  - advantages of, 187
  - assistance from. *See* assistance and rescue skills
  - belaying in, 201–4
  - disadvantages of, 189
  - establishing site, 189–90
  - lowering from, 204
  - for low-ratio instruction, 189
  - mitigating risks, 189–90
  - perspective on base-managed and, 187
- toprope setups, two-rope, 199
- Tricam, 117–18, 119
- triple fisherman's knot, 82

- "Up rope" signal, 194

- V rigging with extension rope, 168–72
- V rigging with instructor tether, 173–75
- vector pull, 227
- vectors, 162–63
- verbal (auditory) learners, 19
- visual learners, 18–19

- "Watch me!" signal, 195
- water knot, 78, 79
- webbing. *See* slings and webbing
- Wilderness Emergency Medical Technician (WEMT), 247
- Wilderness First Aid (WFA) certificate, 9, 247
- Wilderness First Responder (WFR) certificate, 247

- Z system (3:1 raising system), 227–29

About the Authors

Bob Gaines began rock climbing at Joshua Tree in the 1970s. Since then he has pioneered over 400 first ascents in the park. Bob began his career as a professional rock climbing guide in 1983 and is the owner of Vertical Adventures Rock Climbing School, which offers classes and guided climbs at Joshua Tree. In 2008 Vertical Adventures was voted the number 1 rock climbing school in America by *Outside* magazine.

Bob has worked extensively in the film business as a climbing stunt coordinator. He has coordinated over forty television commercials, and he was Sylvester Stallone's climbing instructor for the movie *Cliffhanger.* Bob is an AMGA Certified Rock Climbing Instructor and currently teaches the AMGA Single Pitch Instructor Course at Joshua Tree. He has worked extensively training US military special forces, including US Navy Seal Team 6, and is known for his technical expertise in anchoring and rescue techniques.

Bob is also the author of *Best Climbs Joshua Tree, Best Climbs Tahquitz and Suicide Rocks, Toproping, Rappelling,* and the co-author of *Climbing Anchors* (with John Long).

Jason D. Martin is an AMGA Certified Rock Guide, the operations director at the American Alpine Institute, and a freelance writer. He has been teaching the AMGA Single Pitch Instructor Course since the program's inception in 2008. Much of Jason's adventure writing revolves around the work he does in the mountains. Jason has professionally guided in the Cascades, the Sierra, Red Rock Canyon, Joshua Tree National Park, the Alaska Range, the Coast Mountains of Canada, and in the Andes of Bolivia, Ecuador, and Peru. He wrote the guidebook *Fun Climbs Red Rocks: Topropes and Moderates* and co-authored *Washington Ice: A Climbing Guide.* His writing has appeared in the *Seattle Post Intelligencer, San Antonio Current, Mt. Baker Experience, Northwest Mountaineering Journal, Climbing* magazine, and in numerous other publications. He currently manages the American Alpine Institute blog (blog.alpineinstitute.com), generating numerous climbing and outdoor articles every week. In addition to his work as an outdoor writer and guide, Jason is also a playwright and film critic.

FIND A GUIDE » BECOME ONE
AMERICAN MOUNTAIN GUIDES ASSOCIATION
1979
AMGA.COM

The official manual used by the AMGA to teach single pitch rock climbing

"I've been waiting for this book for a long time. Professionalism, programming, pedagogy, anchors . . . this is a comprehensive resource for understanding the complexities of teaching in the single pitch environment. Highly recommended!"
—Arno Ilgner, AMGA Single Pitch Instructor and author of *The Rock Warrior's Way: Mental Training for Climbers*

"If you want a thorough treatise on what to expect in the [AMGA SPI] course, as well as a ton of information on climbing gear and its evolution, this book will fill your needs. Even if you don't intend to take the course and just want to know what are considered best practices by the AMGA, the answers are here."
—Russ Clune, AMGA Single Pitch Instructor and standard-setting Gunks first ascentionist

Rock Climbing: The AMGA Single Pitch Manual is intended to serve as a textbook for past and future participants of the American Mountain Guide Association's Single Pitch Instructor program. The book builds upon Bob Gaines' book *Toproping* to more specifically address the needs of the professional climbing instructor and advanced recreational climbers. It presents the most current, internationally recognized standards for technical climbing systems used in single pitch rock terrain. Included are chapters on effective teaching in the outdoor environment, risk management, professionalism, and rescue.

Bob Gaines has been a professional rock climbing guide since 1983. He is the owner of Vertical Adventures Rock Climbing School, voted the top rock climbing school in America by *Outside* magazine. An AMGA Certified Rock Instructor and AMGA Single Pitch Instructor Program Provider, he is the author of *Best Climbs Joshua Tree National Park, Best Climbs Tahquitz and Suicide Rocks, Rappelling,* and *Toproping*, and the coauthor of *Climbing Anchors* (all FalconGuides). Visit him at verticaladventures.com.

Jason D. Martin is a professional guide, writer, and the director of operations at the American Alpine Institute. He has guided in Red Rock Canyon, Joshua Tree National Park, the Cascades, the Sierra, the Alaska Range, Canada's Coast Range, and the Andes of Ecuador, Bolivia, and Peru. An AMGA Certified Rock Guide and an AMGA Single Pitch Instructor Program Provider, he is also a playwright, film critic, and outdoor adventure writer. He coauthored *Washington Ice: A Climbing Guide* and authored *Fun Climbs Red Rocks: Topropes and Moderates.* Visit him at alpineinstitute.com.

Front cover photo by Bob Gaines

facebook.com/AMGAUSA

@amgausa

made in the USA

Join our community

$21.95 / Canadian $23.95
ISBN 978-0-7627-9004-3
52195
9 780762 790043

Essential gear for every adventure: Outfit Your Mind®

FALCONGUIDES®

Rock Climbing

FALCON GUIDES®
falcon.com

FalconGuides®
is an imprint of
Globe Pequot Press

A FALCON GUIDE®

Rock Climbing

The AMGA Single Pitch Manual

HOW TO CLIMB® SERIES

BOB GAINES AND
JASON D. MARTIN

The official manual used by the AMGA to teach single pitch rock climbing

"I've been waiting for this book for a long time. Professionalism, programming, pedagogy, anchors . . . this is a comprehensive resource for understanding the complexities of teaching in the single pitch environment. Highly recommended!"
—Arno Ilgner, AMGA Single Pitch Instructor and author of *The Rock Warrior's Way: Mental Training for Climbers*

"If you want a thorough treatise on what to expect in the [AMGA SPI] course, as well as a ton of information on climbing gear and its evolution, this book will fill your needs. Even if you don't intend to take the course and just want to know what are considered best practices by the AMGA, the answers are here."
—Russ Clune, AMGA Single Pitch Instructor and standard-setting Gunks first ascentionist

Rock Climbing: The AMGA Single Pitch Manual is intended to serve as a textbook for past and future participants of the American Mountain Guide Association's Single Pitch Instructor program. The book builds upon Bob Gaines' book *Toproping* to more specifically address the needs of the professional climbing instructor and advanced recreational climbers. It presents the most current, internationally recognized standards for technical climbing systems used in single pitch rock terrain. Included are chapters on effective teaching in the outdoor environment, risk management, professionalism, and rescue.

Bob Gaines has been a professional rock climbing guide since 1983. He is the owner of Vertical Adventures Rock Climbing School, voted the top rock climbing school in America by *Outside* magazine. An AMGA Certified Rock Instructor and AMGA Single Pitch Instructor Program Provider, he is the author of *Best Climbs Joshua Tree National Park*, *Best Climbs Tahquitz and Suicide Rocks*, *Rappelling*, and *Toproping*, and the coauthor of *Climbing Anchors* (all FalconGuides). Visit him at verticaladventures.com.

Jason D. Martin is a professional guide, writer, and the director of operations at the American Alpine Institute. He has guided in Red Rock Canyon, Joshua Tree National Park, the Cascades, the Sierra, the Alaska Range, Canada's Coast Range, and the Andes of Ecuador, Bolivia, and Peru. An AMGA Certified Rock Guide and an AMGA Single Pitch Instructor Program Provider, he is also a playwright, film critic, and outdoor adventure writer. He coauthored *Washington Ice: A Climbing Guide* and authored *Fun Climbs Red Rocks: Topropes and Moderates*. Visit him at alpineinstitute.com.

Front cover photo by Bob Gaines

facebook.com/AMGAUSA

@amgausa

made in the USA

Join our community

$21.95 / Canadian $23.95
ISBN 978-0-7627-9004-3
52195
9 780762 790043

Essential gear for every adventure: **Outfit Your Mind**®

FALCONGUIDES®

Rock Climbing

FALCON GUIDES®
falcon.com

FalconGuides®
is an imprint of
Globe Pequot Press

A FALCON GUIDE®

Rock Climbing
The AMGA Single Pitch Manual

HOW TO CLIMB® SERIES

American Mountain Guides Association 1979

BOB GAINES AND
JASON D. MARTIN

Fifth Edition

SCUBA DIVING

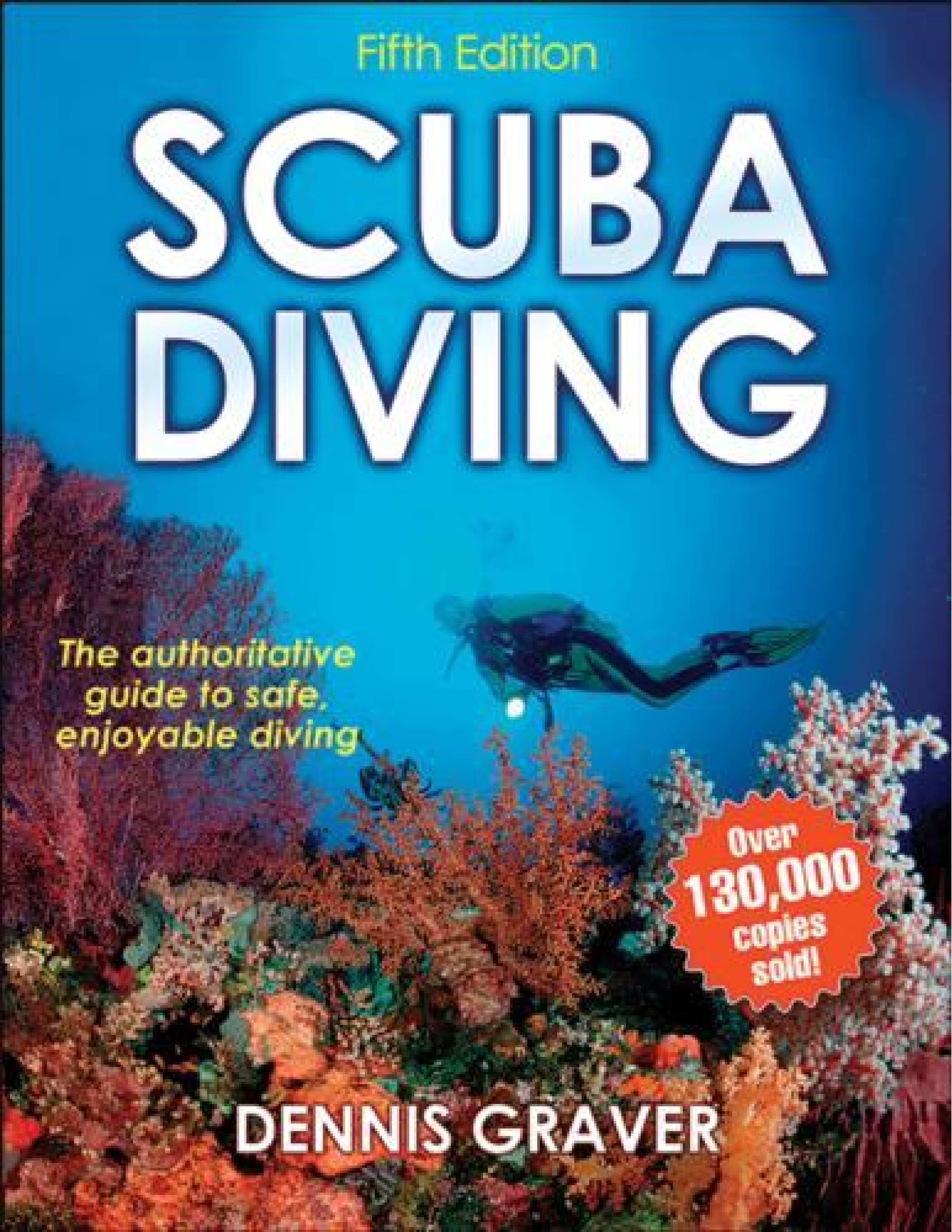

The authoritative guide to safe, enjoyable diving

Over 130,000 copies sold!

DENNIS GRAVER

HUMAN KINETICS

Scuba Diving

FIFTH EDITION

Dennis K. Graver, EMT, SEI

HUMAN KINETICS

Library of Congress Cataloging-in-Publication Data

Names: Graver, Dennis, author.
Title: Scuba diving / Dennis K. Graver, EMT, SEI.
Description: Fifth Edition. | Champaign, IL : Human Kinetics, [2016] |
Includes bibliographical references and index.
Identifiers: LCCN 2016018110 (print) | LCCN 2016031644 (ebook) | ISBN
9781492525769 (print) | ISBN 9781492542032 (ebook)
Subjects: LCSH: Scuba diving.
Classification: LCC GV838.672 .G74 2016 (print) | LCC GV838.672 (ebook) | DDC
797.2/34--dc23
LC record available at https://lccn.loc.gov/2016018110

ISBN: 978-1-4925-2576-9 (print)

Copyright © 2017, 2010, 2003, 1999, 1993 by Dennis K. Graver

All rights reserved. Except for use in a review, the reproduction or utilization of this work in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including xerography, photocopying, and recording, and in any information storage and retrieval system, is forbidden without the written permission of the publisher.

This publication is written and published to provide accurate and authoritative information relevant to the subject matter presented. It is published and sold with the understanding that the author and publisher are not engaged in rendering legal, medical, or other professional services by reason of their authorship or publication of this work. If medical or other expert assistance is required, the services of a competent professional person should be sought.

The web addresses cited in this text were current as of June 2016, unless otherwise noted.

Acquisitions Editor: Tom Heine; **Developmental Editor:** Anne Hall; **Managing Editor:** Nicole Moore; **Copyeditor:** Jan Feeney; **Permissions Manager:** Martha Gullo; **Senior Graphic Designer:** Nancy Rasmus; **Cover Designer:** Keith Blomberg; **Photograph (cover):** Dirscherl Reinhard/Prisma/age fotostock; **Photographs (interior):** Dennis Graver, unless otherwise noted; **Photo Asset Manager:** Laura Fitch; **Visual Production Assistant:** Joyce Brumfield; **Photo Production Manager:** Jason Allen; **Art Manager:** Kelly Hendren; **Associate Art Manager:** Heidi Richter; **Illustrations:** © Human Kinetics; **Printer:** Walsworth

Human Kinetics books are available at special discounts for bulk purchase. Special editions or book excerpts can also be created to specification. For details, contact the Special Sales Manager at Human Kinetics.

Printed in the United States of America 10 9 8 7 6 5 4 3 2 1

The paper in this book is certified under a sustainable forestry program.

Human Kinetics
Website: www.HumanKinetics.com

United States: Human Kinetics
P.O. Box 5076
Champaign, IL 61825-5076
800-747-4457
e-mail: info@hkusa.com

Canada: Human Kinetics
475 Devonshire Road Unit 100
Windsor, ON N8Y 2L5
800-465-7301 (in Canada only)
e-mail: info@hkcanada.com

Europe: Human Kinetics
107 Bradford Road
Stanningley
Leeds LS28 6AT, United Kingdom
+44 (0) 113 255 5665
e-mail: hk@hkeurope.com

Australia: Human Kinetics
57A Price Avenue
Lower Mitcham, South Australia 5062
08 8372 0999
e-mail: info@hkaustralia.com

New Zealand: Human Kinetics
P.O. Box 80
Mitcham Shopping Centre, South Australia 5062
0800 222 062
e-mail: info@hknewzealand.com

E6733

Scuba Diving

FIFTH EDITION

Contents

Foreword vii
Acknowledgments ix

1 **Diving Into Scuba** 1
- The Joys of Diving 2
- How Diving Evolved 3
- The Diving Community 3
- Diver Training 5
- Proficiency Testing 5
- Diving Prerequisites 5
- Diving Risks 6
- Selecting a Dive Course 8
- Diving Responsibilities 9

2 **Diving Science** 11
- Anatomy for the Diver 12
- Respiration and Circulation 15
- Contrasts in Air and Water Environments 19
- Gas Laws 23
- Air Consumption 28
- Heat, Humidity, Light, and Sound 28

3 **Diving Adjustments** 37
- Thermal Adjustments 38
- Buoyancy Adjustments 41
- Pressure Equalization 43
- Breathing Adjustments 46
- Heart and Lungs 50
- Equilibrium Adjustments and Seasickness 52
- Visual Adjustments 53
- Ingassing and Outgassing 53
- Dehydration Prevention 58
- Mobility Adjustments 59

4 **Diving Equipment** 63
- Basic Equipment for All Divers 65
- Exposure Suits 71
- Weighting Systems 78
- Buoyancy Compensators 80
- Scuba Cylinders 83
- Cylinder Valves 87
- Scuba Regulators 89
- Alternate Air Sources 94
- Instrumentation 96
- Dive Knives and Accessories 100
- Specialized Equipment 104

5 Diving Environment 109
Aquatic Biology 110
Potentially Dangerous Aquatic Animals 114
Conservation and Preservation 117
Aquatic Conditions 121
Boat Diving 132

6 Diving Skills 137
Preparing to Skin Dive 138
Skin Diving Skills 141
Preparing to Scuba Dive 148
Scuba Diving Skills 152
Managing Physiological Problems 169
Managing Physical Problems 173
Managing Emergencies 178

7 Dive Planning 183
Dive-Planning Factors 184
Advance Planning and Preparation 185
Short-Term Planning and Preparation 187
On-Site Planning and Preparation 187
Area Orientations 189
Dive Profile Planning 190
Dive Tables 191
Dive Calculators and Dive Computers 204
Postdive Review 206

8 Diving Opportunities 209
Continuing Education 210
Local Opportunities 212
Dive Travel 213
Dive Destinations 216
Career Opportunities 218
Leadership Opportunities 219

Appendix A: Associations and Periodicals 221
Appendix B: AOK Answers 223
Glossary of Scuba Diving Terms 231
Bibliography 239
Index 241
About the Author 245

This page intentionally left blank.

Foreword

Recreational scuba diving has progressed from its beginnings in the 1950s. Dennis Graver's fifth edition of *Scuba Diving* is uniquely designed toward a level of education that provides new divers with an exceptional level of knowledge while promoting safety and enjoyment. With our training agencies, Scuba Educators International (the former YMCA scuba program) and PDIC International, our goal is to provide a full education for new divers. This text accomplishes that goal and perhaps can also be used for advanced programs as well.

At a time when instant gratification leads many to seek recreational activities that can be accomplished with little effort in a brief time, scuba takes new divers into an environment that remains challenging. Therefore, the need for a full education has not changed. Perhaps the delivery of information will change in the future; however, the need to understand the complexities of the undersea world will not.

In addition to bringing this fifth edition up to current levels of information, Dennis has provided end-of-chapter questions that result in additional learning. This edition is among the best of texts in the scuba industry for new divers.

Dave Barry wrote, "There is nothing wrong with looking at the surface of the ocean itself, except that when you finally see what goes on underwater, you realize that you've been missing the whole point." I couldn't agree more.

Tom Leaird, CEO
Scuba Educators International
PDIC International

This page intentionally left blank.

Acknowledgments

My appreciation includes the review of the Application of Knowledge (AOK) questions by Dr. Terry Beattie, former diving medical services officer for the Washington, DC diving unit. His physiology expertise was also helpful.

Wall diving in Haiti

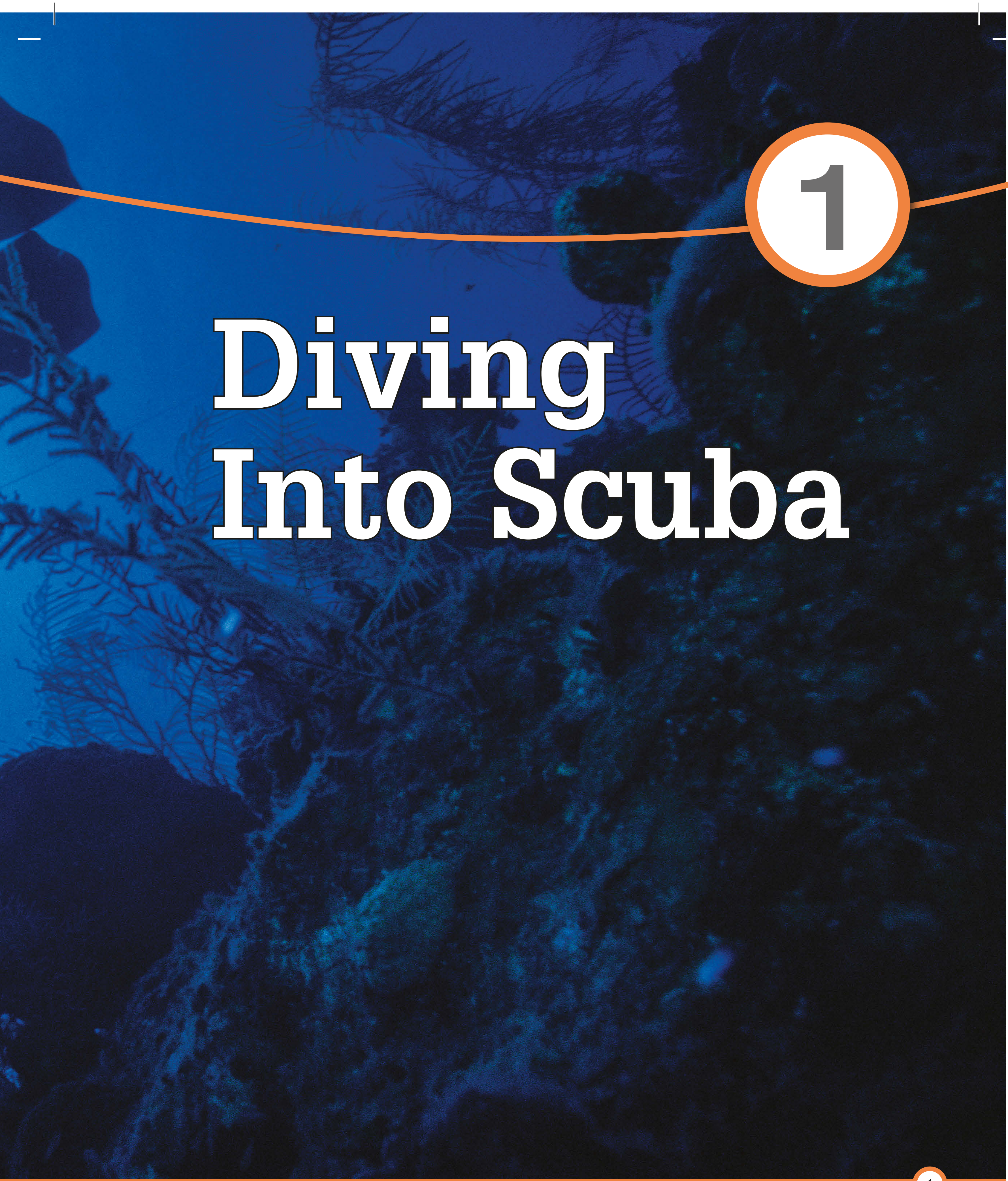

1

Diving Into Scuba

Dive In and Discover

By the end of this chapter, you will be able to do the following:

- Describe the joys of diving.
- Define the terms *closed circuit*, *open circuit*, *scuba*, *semi-closed circuit*, and *C-card*.
- Explain the training requirements for participation in scuba diving activities.
- List two medical conditions that disqualify an individual from scuba diving.
- Describe three risks associated with scuba diving.
- List several questions to ask when selecting an entry-level scuba diver training course.
- Explain several responsibilities that you assume when you become a certified scuba diver.

The Joys of Diving

As a diver, you are weightless and can move in all directions. Your freedom approaches that of a bird as you move in three dimensions in a fluid environment. Diving weightless in clear water in a forest of underwater plants with sunlight streaming down is only one of many unforgettable experiences awaiting you.

Just as there are mountains, plains, and various environments above water, there are various environments for you to experience underwater. Coral reefs, kelp forests, incredible rock formations, and other natural wonders await divers in various geographical regions. In addition, divers can explore piers, jetties, quarries, mysterious shipwrecks, and other artificial structures. The variety of underwater sights is limitless. There is more to view underwater than anyone could possibly see in an entire lifetime. A range of activities—such as photography, hunting, and collecting—make diving challenging and rewarding. Scuba diving includes an activity to interest everyone.

Divers are friendly and easy to get to know; their camaraderie is well known. Diving is a sharing activity, and there is much to share. If you enjoy traveling, you will probably love diving. Dive travel is the number one business of recreational diving. Reasonably priced dive vacations to exotic islands abound. Most divers plan one or more diving vacations each year.

The sensations of diving are fantastic but are difficult to explain. Words cannot describe the peaceful solitude of inner space. Diving contributes to good health, can help reduce work-related stress, can increase self-esteem, and can make you feel great. You need to experience the emotions and sensations for yourself. You will then begin to know the joys of diving.

A giant grouper

When you descend beneath the surface of the water, you enter an entirely new and beautiful world. You have opportunities to see incredible life-forms that only a few people ever see. Imagine swimming in a giant aquarium, and you'll get a glimpse of what you can expect to experience in the underwater world.

How Diving Evolved

Interest in the underwater world has always existed. Driven by curiosity and the need for food, people have ventured beneath the surface of Earth's seas for thousands of years. Records exist of sponge divers, oyster pearl divers, military divers, and even salvage regulations dating back to 3000 BC. In ancient times, the two methods employed for diving were breath holding and the use of an inverted, air-filled bell (which was large and heavy). John Lethbridge, an Englishman, developed an oak cylinder diving engine in 1715 to pump air into a bell. The hand-operated air compressor, invented in 1770, allowed fresh air to be pumped into a submerged diving bell. In 1772, Sieur Freminet of France invented a helmet-hose system that permitted divers to work without a bell. The greatest early advancement in underwater equipment occurred in 1837, when Augustus Siebe (a German living in England) invented the first closed-dress diving suit, which became known as the hard-hat system. Siebe's system was so effective that it dominated underwater work for 100 years, and it is still in use today.

The origin of recreational diving can be traced to 1825. That year, William H. James, an Englishman, invented the first **open-circuit** self-contained underwater breathing apparatus **(scuba)** system. Although this system was not very practical, it did solve the problems of the air hose tether and the depth restrictions related to oxygen poisoning. Benoit Rouquayrol (a French mining engineer) and Auguste Denayrouze (a French naval officer) invented the aerophore in 1865. Their creation is considered the source of modern scuba equipment.

Later, two Englishmen, Henry Fleuss and Robert Davis, developed a **closed-circuit** oxygen rebreather system in 1878. This system eliminated the air hose that had previously tethered divers and restricted their movement. The system had great military application because bubbles were not exhausted into the water. Oxygen rebreathers were used extensively by Italian and British frogmen during World War II. However, the use of oxygen rebreather systems is limited to a depth of 20 feet (about 7.6 m) because breathing pure oxygen at greater pressures causes convulsions, which can be fatal during submersion.

Yves Le Prieur of France developed a manually controlled open-circuit scuba system in the early 1920's (year disputed); however, modern scuba diving was launched with the development of a "demand" scuba system that was perfected by Frenchmen Emile Gagnan and Jacques Cousteau in 1943. With this system, a diver could demand compressed air from a steel cylinder by simply inhaling from a pressure regulator held in the mouth. Scuba was introduced in the United States in 1950 and has been popular ever since.

The military has continued to improve underwater breathing systems that have found their way into use by recreational divers. The latest scuba systems are **semi-closed-circuit** rebreathers. These systems use sophisticated electronics to control the oxygen mixture that a diver breathes, periodically releasing only a small amount of bubbles. High levels of training, frequent use, and dedicated maintenance are required for the safe use of these expensive rebreathers. Table 1.1 outlines the basic features of each type of breathing system in more detail.

The Diving Community

The recreational diving community consists of equipment manufacturers, diving retailers, diving educators, diver training organizations, dive resorts, diving

Table 1.1 Scuba System Comparisons

Open-circuit scuba	Semi-closed-circuit scuba	Closed-circuit scuba
User inhales compressed air	User inhales oxygen and inert gas	User inhales pure oxygen
Exhaled air exhausted into water	Exhaled gases enter closed system	Exhaled gases enter closed system
Air supply duration varies with depth	Carbon dioxide absorbed within system	Carbon dioxide absorbed within system
Components simple	Oxygen added as needed	Oxygen added as needed
Unit easily maintained	Air supply duration unaffected by depth	Depth limit of 25 feet
Affordable	Components sophisticated and complex	Convulsions and death possible
Basic training required for use	Unit requires high maintenance	Components simple
	Expensive	High maintenance
	Highly technical specialty training required	Not available for recreational use
		A predecessor to semi-closed-circuit scuba

supervisors, dive guides, dive clubs and associations, publishing companies, and certified divers. Commercial, scientific, and professional diving are not considered recreational pursuits; separate communities are involved in these types of diving. This book addresses only recreational skin diving and scuba diving. A scuba diver breathes compressed air underwater, while a skin diver holds his or her breath while submerged.

Few laws pertain to recreational scuba diving. The laws that exist do not govern who may dive. The diving industry is self-regulating. The diving community realizes that it is dangerous when people who have not completed a sanctioned course of instruction attempt scuba diving. Dive businesses require proof of completion of training before they will allow you to have your scuba tanks filled or allow you to participate in diving activities. Many dive operations also require proof of recent experience documented in a diving logbook. If you have not been diving for a year or more, you may be required to complete at least one dive under the supervision of a diving professional. The supervised dive requirement helps increase the safety of divers whose skills may need to be refreshed.

When you complete your training requirements as a scuba diver, you receive a certification card called a **C-card**. Most C-cards do not require renewal, but the recreational diving community universally recommends the completion of refresher training after periods of inactivity in excess of six months.

Certified divers may dive without supervision or may employ the services of a diving guide. Just because a divemaster or diving supervisor is aboard a dive boat, you should not assume that this person is a guide who will dive with you. Guide services are not necessarily included with diving trips. If you want a guide to lead you about underwater and show you the sights, you should arrange for guide services in advance.

You will learn more about the community as your diving experience increases. Many opportunities for adventure and enjoyment can be found within this community. Get actively involved in the community when you complete your training and officially qualify as a scuba diver.

Diver Training

A national diver training organization must sanction your training. The training organization establishes standards of training that you must meet before the organization will issue a certification card. The Appendix A includes a list of national diver training organizations. Your instructor should have credentials that identify her as a qualified instructor. The instructor's membership in the training organization must be current in order for the person to be qualified to teach and certify divers. Be sure to confirm your instructor's qualifications.

An entry-level training course usually consists of a series of academic sessions, pool or confined-water (pool-like conditions in open water) sessions, and open-water training (in actual diving locations). You will learn theory in the classroom, learn skills in controlled conditions, and then apply your skills in an actual diving environment. This logical progression is common for all approved diver training courses.

The minimum requirements for your training are as follows: You should have four or more academic sessions, four or more pool sessions, and at least four scuba dives in open water. A skin dive in open water may also be part of your training.

Your initial training should involve a total of 30 to 40 hours of instruction. The instruction should occur over a period of several weeks instead of a few days. The time between class sessions allows you to reflect on your training and helps you absorb and retain the knowledge and skills better than a concentrated training schedule would.

Proficiency Testing

After you have learned and practiced the skills of skin and scuba diving, you must demonstrate your competence at a level established by the agency sponsoring your training. Proficiency testing may include diving exercises that are challenging and fun. Examples include mask recovery and clearing for skin diving, simulated boat exit for scuba diving, alternate air source breathing, buddy breathing, equipment handling, and a sequence of surface entry and equipment donning known as a bailout.

Diving Prerequisites

Scuba diving can be undertaken by anyone over 12 years of age who is in normal health and has a reasonable degree of physical fitness. People younger than this should not participate in scuba diving (even when supervised by adults) because they do not have the mental and emotional maturity to deal with the problems that might arise. Skin diving is a good activity for youngsters if they are well supervised.

For scuba diving, you need to have swimming ability, but you do not need to be a competitive swimmer. At the beginning of your training, you should be able to swim 200 yards (183 m) nonstop at the surface using any combination of strokes. There is no time requirement for the swim. Being comfortable in the water is more important than being able to swim fast. You also need to be able to swim 25 feet (7.6 m) underwater with no push-off. By the end of the course, you should be able to swim 300 yards (274 m) nonstop at the surface using any combination of strokes; you should also be able to swim 50 feet (15.2 m) underwater with no push-off. The goal is to increase your aquatic proficiency during the course.

Good health means your heart, lungs, and circulation are functional and that you do not have any serious diseases. Any medical conditions—even if controllable under normal conditions—that might incapacitate you in the water could cause you to drown while scuba diving. Some individuals with asthma or diabetes may be able to dive if they have obtained special medical approval. People with physical disabilities may also dive if they have medical approval from a physician. The air spaces in your body—sinuses, ears, and lungs—must be normal because changes in pressure affect them. Other medical conditions, such as seizure disorders, absolutely preclude a person's involvement in diving; a seizure while diving can be fatal. Women who are pregnant should not scuba dive. Increased pressure can adversely affect an unborn child. Pregnant women may choose to participate in snorkeling as an alternative to scuba diving. Many women ask whether they may dive during menstruation. Menstruation does not preclude a woman from diving if her health permits participation in other sports during that time.

You need to be emotionally fit as well as physically fit for diving. If you are terrified of water or of feeling confined, diving is probably an activity you should avoid. Normal concerns are to be expected, but stark terror is unacceptable.

You should have a physical examination before you begin your training, especially if it has been more than a year since your last exam. Ask your instructor to recommend a diving physician. Physicians who do not understand the physiology of scuba diving sometimes inappropriately grant approval to people who have medical conditions that place them at great risk in and under the water. Your instructor can likely recommend a physician who understands medical issues related to scuba.

Diving Risks

All activities present some risk. There is risk involved in walking across the street or driving a car. To avoid injury while participating in an activity, people take precautions for their safety. Precautions must be taken for scuba diving just as for any other pursuit. The level of risk in diving is similar to that of flying in an

SCUBA WISE

I nearly drowned when I was 4 years old, and I became terrified of water. When I was 8, I spent a summer with my uncle in Ohio. He would take me to Lake Erie and give me pennies if I would wade into water deep enough to cover my navel. I learned to swim as a Boy Scout at age 11. Although I completed a lifesaving class at age 16, I was still apprehensive about water. When I learned skin and scuba diving and discovered that I could actually relax in water, the water became my ally. For the first time in my life, I enjoyed water and was able to rid myself of my childhood fears. Just because someone feels anxiety about water does not mean that the person can't enjoy scuba diving. If you can swim 200 yards, you can learn to dive and to love being in and under the water.

airplane. Both are low-risk activities when done with well-maintained equipment according to established rules and in good environmental conditions. Unfortunately, both activities are unforgiving if you ignore the rules and recommendations designed to minimize the risks.

The following information (and the information throughout this book) will make you aware of injuries that scuba divers can incur. This information alerts you to potential hazards and, more important, helps you learn to avoid injury. If you do what you are taught to do as a diver, your risk will be minimal, and all of your diving experiences will likely be pleasant ones.

Pressure changes with depth. Changes in pressure can severely injure bodily air spaces if you are not in good health or if you fail to equalize the pressure in the bodily air spaces with the surrounding pressure. You will learn equalizing techniques as part of your training. Gases are normally dissolved in the fluids and tissues of your body. Increased pressure increases the amount of gas dissolved in your body. If you ascend too rapidly from a dive, the gases in your system can form bubbles and produce a serious illness known as **decompression illness**. By regulating your depth, the duration of your dive, and your rate of ascent, you can avoid decompression illness. Failure to heed depth and time schedules and ascent rates can result in serious, permanent injuries.

Diving can be strenuous at times. You need sufficient physical fitness and stamina to handle long swims, currents, and other situations that may arise. If you become winded from climbing a flight of stairs, you may need to improve your level of fitness before learning to scuba dive. Exhaustion in and under the water is hazardous. A good exercise to improve fitness for diving is swimming with fins while breathing through a snorkel.

Diving takes place in water, an alien environment. You use life-support equipment to dive, but you cannot depend entirely on the equipment for your well-being. Aquatic skills are essential in and around water. People with very weak aquatic ability can drown when minor equipment problems occur—problems that could be handled easily by a person with good water skills. To be a scuba diver, you must be comfortable in the water.

You should not be overly concerned with the potential risks of diving because the possible

Red Sea coral cave

injuries are preventable. Learning how to dive as safely as possible is the purpose of your training. You will learn to minimize the risk of injury and maximize your enjoyment of the underwater world.

Selecting a Dive Course

There are many diver training organizations and thousands of professional diving educators. Your phone book might list diving businesses that offer sanctioned courses. Many universities, community colleges, and recreational departments also offer scuba courses. (See Appendix A for a list of diver training organizations.) Ask about the qualifications, experience, and reputation of several diving instructors in your area to select the course that can provide you with the best possible training. Here are some questions you should ask:

- Is this training sanctioned by a diver training agency?
- How long has the instructor been teaching scuba diving?
- Which levels of training is the instructor qualified to teach?
- May I speak with the graduates of a recent class?
- Why is this course better than others in the area?
- Are assisting and rescue techniques taught in the course?
- How many instructor-supervised open-water dives are included?

The tuition for diving instruction is usually between $200 and $300. The lowest-priced course may not necessarily be a bargain. Find out what is included with the course fee and, more important, what the total cost will be for you to become certified as a scuba diver. You do not have to purchase all the equipment needed to scuba dive, but you need to have a mask, snorkel, fins, and usually boots and gloves for your training (see figure 1.1). Use of the additional required equipment is typically part of the course tuition.

You should find out whether the price of the course includes the costs of educational materials and certification. There may be additional costs for travel, lodging, parking, boat fees, and equipment rental for open-water training. Determine the complete cost before enrolling in a course.

When you have selected the best program for you and have enrolled in a course, you should receive a reading assignment for your first session. If you are not given an assignment, speak with the instructor; your learning will be enhanced if you read in advance about the topics to be presented in class. Good diving instructors provide a handout with reading assignments.

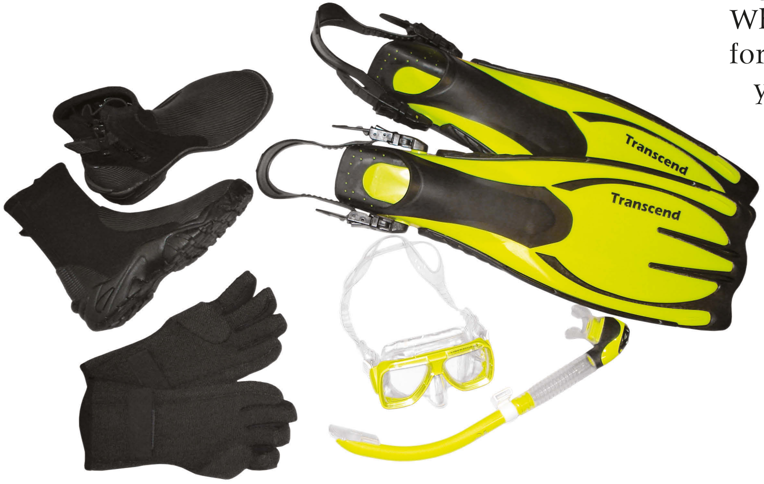

Figure 1.1 Required scuba diving training equipment.

Diving Responsibilities

When you qualify as a scuba diver, you assume many responsibilities. You are responsible for your safety, for the safety of those you dive with, for the image of scuba divers, and for the preservation of the diving environment. The diving community encourages divers to accept responsibility for their actions. To be part of the diving community, you need to be a responsible diver. Learn what you should do, then do what you learn.

Summary

Diving can be a source of great joy. Many exciting experiences await the trained diver. You need dive credentials to participate in diving activities. You must complete diver training to obtain your C-card and logbook. But diving is not for everyone. You must have normal health, good swimming skills, and reasonable physical fitness. Diving poses risks that a well-trained, wise, and fit diver can minimize. Compare training programs and choose the best education, which may not be the quickest or the least expensive. Remember that you accept a great deal of responsibility when you become a diver. Do not assume that you can transfer the responsibility for a dive accident to someone else. Ultimately, you control your actions underwater. Become a competent, self-reliant diver who adheres to recommended safety practices, and you will discover the joy of diving.

Application-of-Knowledge (AOK) Questions

The following questions will enhance your understanding of what you have learned in this and every chapter. Take time to consider each question before you look at the answer, which is at the back of the book in Appendix B. When you apply the basic knowledge you have learned and correctly respond to the following questions, you demonstrate understanding, which is a higher level of learning than mere knowledge. Do not be concerned if you do not have a correct response. Your analysis of the question and the revelation of the answer will increase your understanding.

1. What is your primary reason for learning to scuba dive? What do you think you will do while scuba diving after completion of your initial training?
2. Why do beginning recreational divers use open-circuit breathing systems instead of closed-circuit systems?
3. Why should you document your dives in a logbook?
4. What are the advantages of completing a 40-hour scuba diving course rather than a course that is much shorter in duration?
5. Why is a medical examination recommended for those who want to learn scuba diving? What type of doctor is the best choice for a scuba diving medical exam?
6. List three actions you can take to minimize your risk of injury when scuba diving.
7. What are some of the factors that you should consider when selecting a scuba diving course?
8. List three actions you can take to demonstrate that you are a responsible diver.

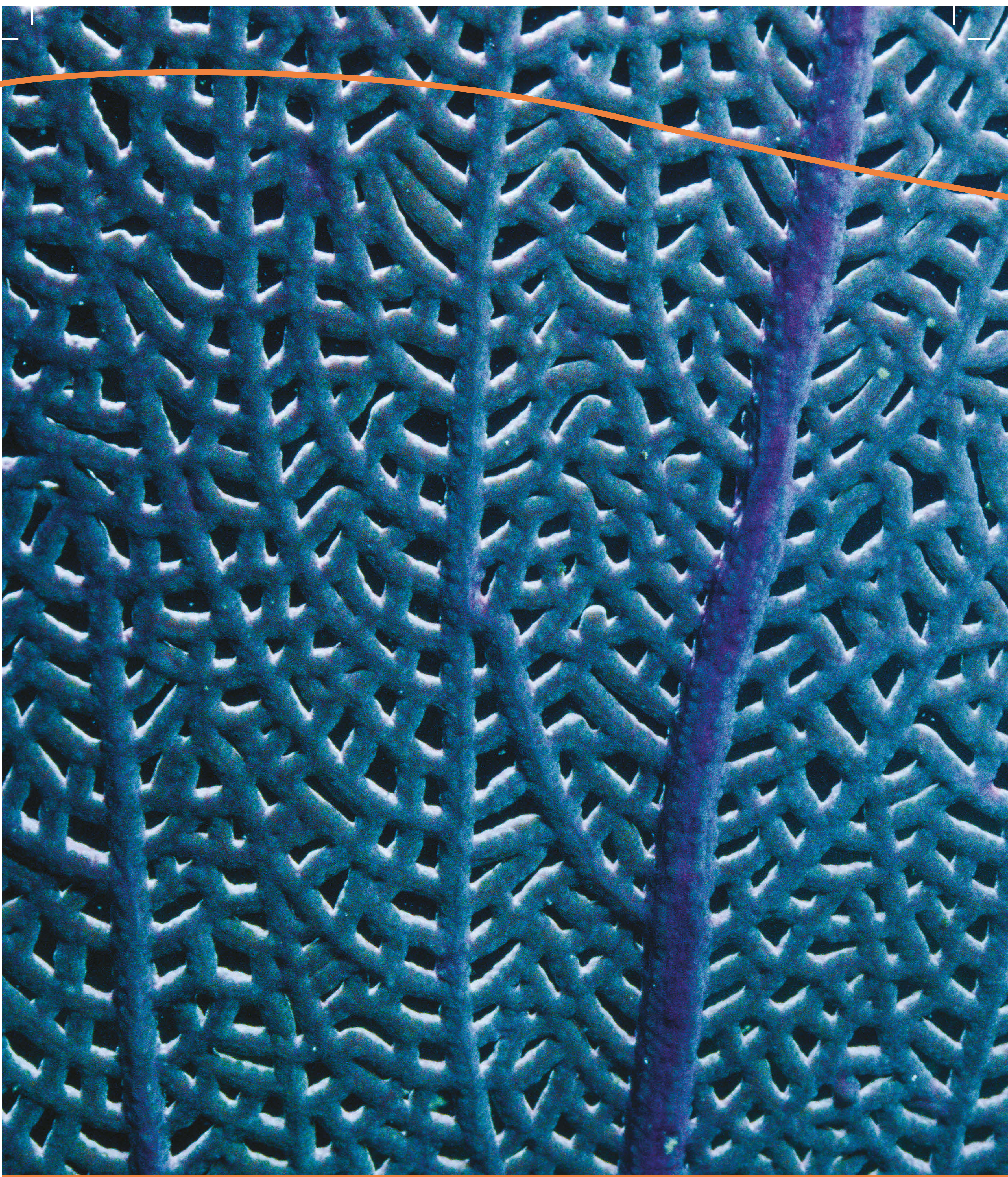

A sea fan

2

Diving Science

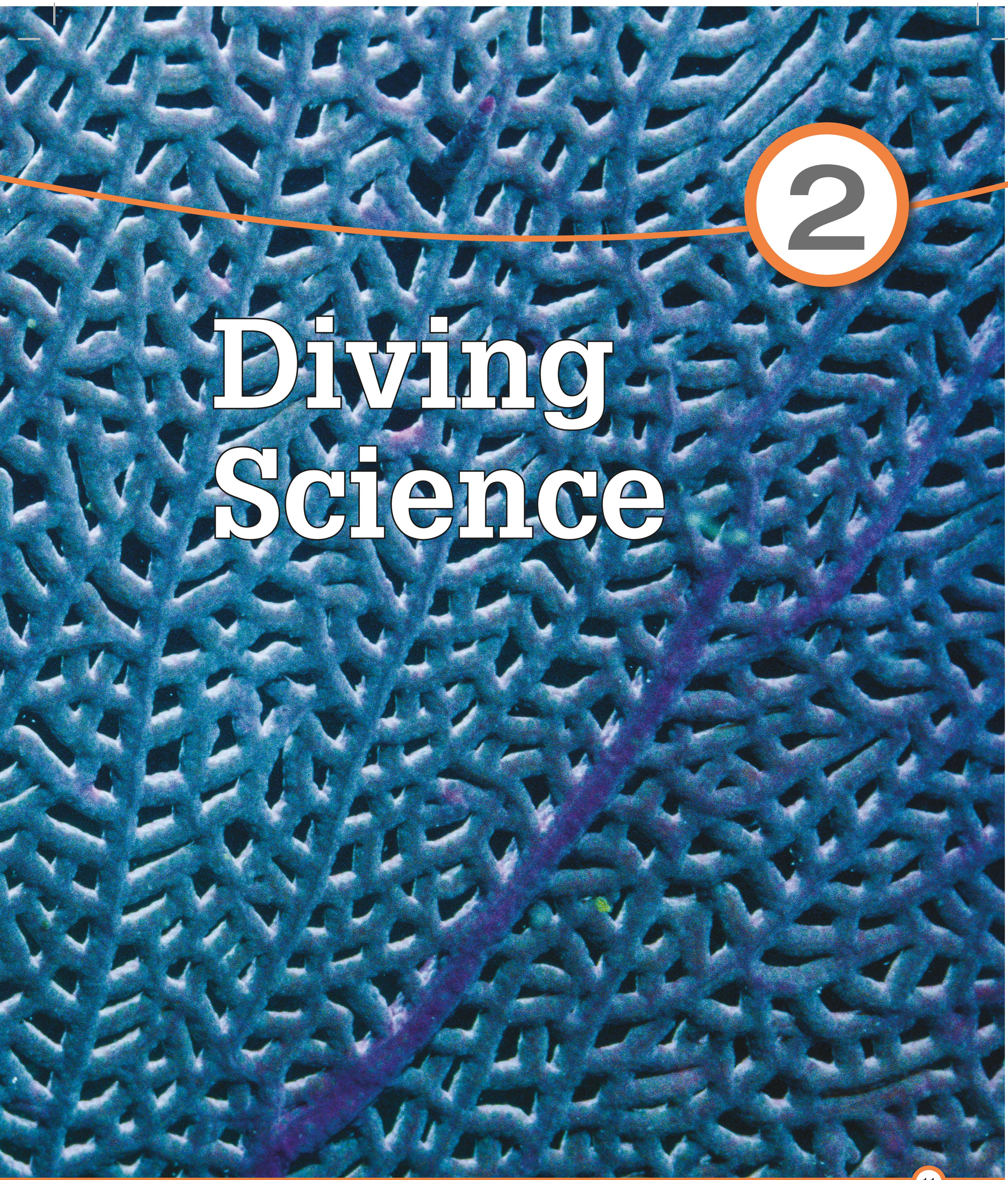

Dive In and Discover

By the end of this chapter, you will be able to do the following:

- List three body air spaces of concern to divers.
- Describe the process of hearing in air.
- Describe the cardiorespiratory process.
- Explain the effect of carbon dioxide on breathing.
- Define the terms *eardrum*, *middle ear*, *eustachian tube*, *nitrogen narcosis*, *hyperventilation*, *hypoventilation*, *density*, *buoyancy*, *pressure*, *gauge pressure*, *absolute pressure*, *Boyle's law*, *Gay-Lussac's law*, *Dalton's law*, *squeeze*, *reverse block*, *ingassing*, *outgassing*, and *partial pressure*.
- Explain the principle of buoyancy and the key to controlling buoyancy.
- Explain the effects of pressure and temperature on a volume of air in a flexible container.
- Explain why it is important to vent your lungs when you ascend in water.
- Convert any temperature to absolute temperature.
- Explain the process of ingassing and outgassing.
- State the two primary factors that affect the air consumption of a scuba diver.
- List four methods of heat loss.
- Describe three potential problems for scuba divers that may be caused by humidity.
- Describe the effects of water on vision and hearing.

Anatomy for the Diver

Inside your body are air-filled spaces that are affected by changes in pressure. The three body air spaces of primary concern to you as a diver are the lungs, the ears, and the sinuses. Understanding the structure and function of your throat will also help you as a diver. Figure 2.1 illustrates the structure and functions of the sinuses, throat, and lungs.

Sinuses

The sinuses warm and humidify inspired air. They secrete mucus to help protect the body by trapping airborne germs. The small airways that connect the sinuses to the nasal passages are normally open. Congested sinuses pose problems for divers. In the next chapter, you will learn more about sinus problems and how to prevent them.

Throat

In addition to being the organ of voice, the throat and larynx help prevent foreign matter from entering the lungs. If something foreign, such as food or water, comes into contact with the larynx, a reflex action causes a spasm of the larynx. Coughing expels the foreign substance. You have experienced this sensation when something has "gone down the wrong pipe." Review the throat section of figure 2.1. During your scuba diving training, you will learn how to keep water out of your larynx to avoid coughing and choking in and under the water.

Lungs

Healthy lungs are essential for scuba diving. The lungs are large organs that contain millions of microscopic air sacs. Your lungs have a maximum capacity and a minimum capacity. When you exhale completely, your lungs are not empty. They contain about 2 pints (1L) of air. The air remaining in your lungs after you have exhaled completely is your residual volume. The amount of air you move in and out of your lungs is your tidal volume. When you are at rest, your tidal volume is small. When you exert yourself, your tidal volume increases until you reach both your maximum lung volume and your residual volume with each breath. Your vital capacity is the difference between the volume of air for a maximum inhalation and the volume of air for a maximum exhalation—typically about 6 to 8 pints (2.8 to 3.8L). In the next chapter, you will learn several reasons why your lungs are the most critical air spaces when diving.

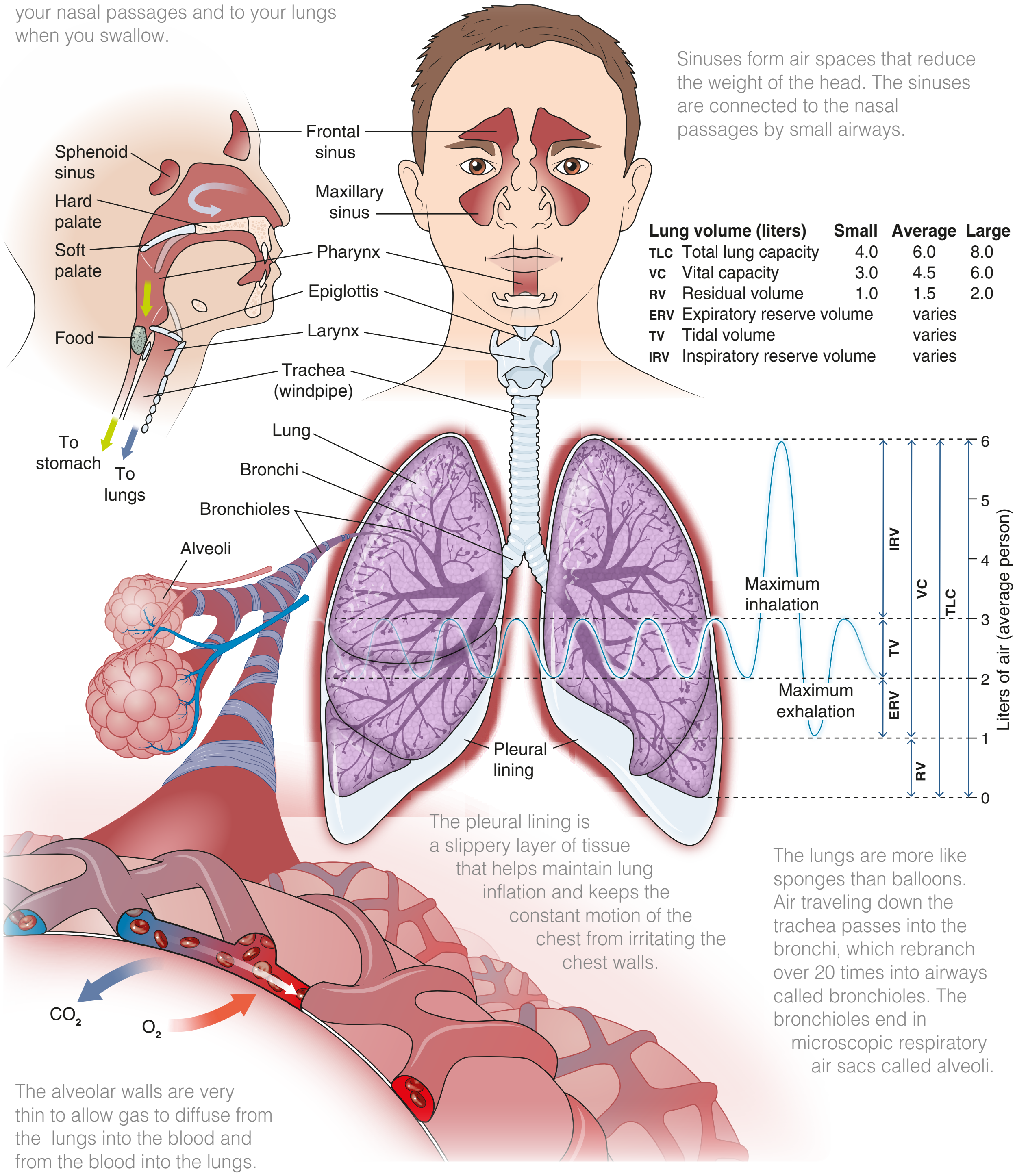
Sphenoid sinus
Hard palate
Soft palate
Food
Frontal sinus
Maxillary sinus
Pharynx
Epiglottis
Larynx
Trachea (windpipe)
To stomach
To lungs
Lung
Bronchi
Bronchioles
Alveoli
Pleural lining
Maximum inhalation
Maximum exhalation
IRV
TV
ERV
RV
VC
TLC
6
5
4
3
2
1
0
Liters of air (average person)
CO₂
O₂

The epiglottis and the soft palate in the back of your pharynx seal the airways to your nasal passages and to your lungs when you swallow.

Sinuses form air spaces that reduce the weight of the head. The sinuses are connected to the nasal passages by small airways.

Lung volume (liters)		Small	Average	Large
TLC	Total lung capacity	4.0	6.0	8.0
VC	Vital capacity	3.0	4.5	6.0
RV	Residual volume	1.0	1.5	2.0
ERV	Expiratory reserve volume		varies	
TV	Tidal volume		varies	
IRV	Inspiratory reserve volume		varies	

The pleural lining is a slippery layer of tissue that helps maintain lung inflation and keeps the constant motion of the chest from irritating the chest walls.

The lungs are more like sponges than balloons. Air traveling down the trachea passes into the bronchi, which rebranch over 20 times into airways called bronchioles. The bronchioles end in microscopic respiratory air sacs called alveoli.

The alveolar walls are very thin to allow gas to diffuse from the lungs into the blood and from the blood into the lungs.

Figure 2.1 Functions of the sinuses, throat, and lungs.

Colorful fish in a Florida coral cave

Your body is a marvelous machine. It performs many complex functions automatically. Your body is well adapted to an air environment, but it can also adjust in many ways to the aquatic environment. In this chapter, you will become familiar with some of the structures and functions of your body that are important for scuba diving. You will learn the differences between the air and water environments and how changes in pressure affect your body. As a diver, you face many challenges in the underwater environment, but these challenges can be managed.

Ears

Behind your **eardrum** is an air space called the **middle ear**, which is illustrated in figure 2.2. The pressure in the middle ear must equal the pressure in the outer ear; otherwise, the eardrum cannot move freely. The next chapter explains how to keep pressure inside your ears equal to external pressure. Your **eustachian tube** allows the equalization of pressure in the middle ear. The liquid-filled cochlea contains hairlike projections called cilia, which convert mechanical movement to electrical signals for the brain. The movement of the oval window by the tiny bones of the middle ear causes the liquid and the cilia in the cochlea to move back and forth. The oval window movement could not take place without a second window in the hearing organ—the round window. When the oval window moves inward, the round window moves outward, and vice versa.

If the motion sensed by your semicircular canals and the visual cues received by your eyes are not in harmony, motion sickness can result. Sudden changes in temperature or pressure in the middle ear can affect your semicircular canals

Abbreviations

✓	**ATA**	atmospheres absolute
✓	**atm**	atmospheres
✓	**CO**	carbon monoxide
✓	**CO_2**	carbon dioxide
✓	**°C**	degrees Celsius
✓	**°F**	degrees Fahrenheit
✓	**FFW**	feet of freshwater
✓	**FSW**	feet of seawater
✓	**ft**	feet
✓	**m**	meters
✓	**O_2**	oxygen
✓	**psia**	pounds per square inch absolute
✓	**psig**	pounds per square inch gauge

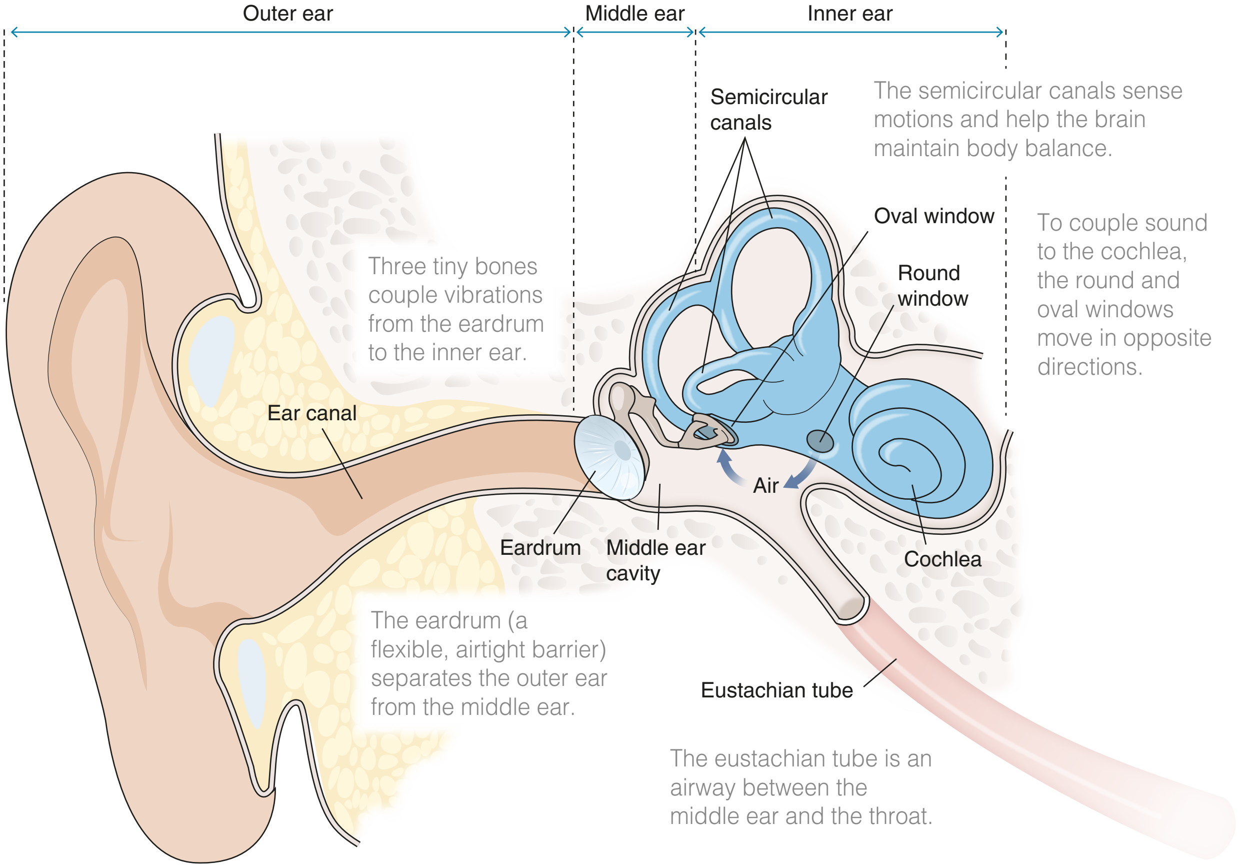
Outer ear
Middle ear
Inner ear
Semicircular canals
The semicircular canals sense motions and help the brain maintain body balance.
Oval window
Round window
To couple sound to the cochlea, the round and oval windows move in opposite directions.
Three tiny bones couple vibrations from the eardrum to the inner ear.
Ear canal
Air
Eardrum
Middle ear cavity
Cochlea
The eardrum (a flexible, airtight barrier) separates the outer ear from the middle ear.
Eustachian tube
The eustachian tube is an airway between the middle ear and the throat.

Figure 2.2 Process of hearing in air.

and cause temporary disorientation. (The next chapter presents potential ear problems for divers, how to avoid them, and how to handle them if they happen.)

Teeth

You may be surprised to learn that there are dental concerns for divers. Pressure can affect air pockets in improperly filled teeth and can cause tooth pain. If a tooth hurts only under pressure or only after a dive, see your dentist and tell him what you suspect. The roots of some upper molars extend into the sinus cavities. You should postpone diving for several weeks after you have had a tooth extracted.

Your mouth and jaws are designed for an even bite. If you bite hard on a mouthpiece with only your front teeth for prolonged periods, your jaws will become sore. Special mouthpieces designed for a proper bite can help reduce the problem. You should not have to bite hard on a mouthpiece to hold it in place. If you find biting necessary, get lighter equipment. Prolonged, improper biting that irritates your jaws can lead to serious inflammation of your jaws and ears.

Respiration and Circulation

One of the fascinating processes within the human body is your ability to breathe in air and circulate oxygen to the tissues with no conscious effort. As your level of exertion increases, your heart and lungs automatically adjust to meet the increased demands for oxygen and nourishment. An understanding of the gases involved in respiration and the basics of respiration and circulation can help you understand the effects and the demands of diving on your lungs and heart.

Gases We Breathe

Several gases affect recreational divers. You need to know about their effects on your body. About 80 percent of air is nitrogen (N^2). At sea-level pressures, nitrogen has no effect on your body. At a depth of about 100 feet (30 m), the increased pressure of the gas has a detrimental effect, which is called **nitrogen narcosis**. Excessive nitrogen in your body at the end of a dive can produce a serious illness known as decompression sickness. You will learn more about nitrogen narcosis and decompression sickness in the next chapter.

Oxygen (O_2) is the gas that supports human life. Any other gas mixed with oxygen serves only as a vehicle for oxygen to be inspired. Approximately 21 percent of air is oxygen (see figure 2.3). You need to breathe at least 10 percent oxygen to remain conscious. However, oxygen breathed under high pressure is poisonous and causes convulsions because oxygen at increased pressure affects your nervous system. You usually have compressed air—not pure oxygen—in your scuba tanks. A specialty form of diving uses a nitrogen and oxygen mixture with a higher percentage of oxygen than is found in air. The mixture, which reduces the effects of nitrogen at depth, is called **nitrox**. The use of special mixed gases, including nitrox, requires special training, equipment, and procedures.

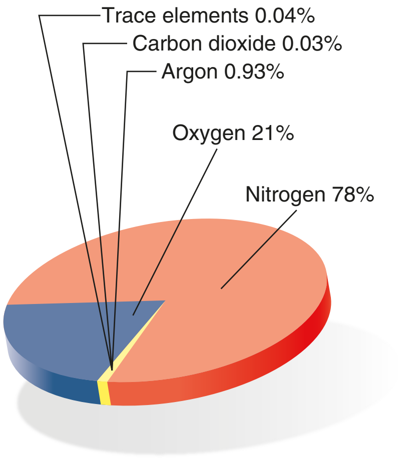

Figure 2.3 Composition of air.

As your tissues use oxygen, they produce carbon dioxide (CO_2). Carbon dioxide is the primary stimulus for respiration. The greater the level of carbon dioxide in your body, the greater your urge to breathe will become. If the level of carbon dioxide in your body becomes too great, unconsciousness will result.

Carbon monoxide (CO) is a poisonous gas produced by the incomplete combustion of gas or oil. The exhaust from an internal combustion engine contains carbon monoxide. An oil-lubricated air compressor that overheats can produce carbon monoxide. Even a minute amount of carbon monoxide in your scuba tank can poison you and lead to unconsciousness or death. Air filling stations must take care to avoid contamination of air with carbon monoxide.

Breathing and Circulation Mechanics

When you need to breathe, sensors at the base of your brain send a signal that stimulates your diaphragm to contract and your chest to expand. This draws air into your lungs in the same way that an old-fashioned bellows draws in air when you expand it. Your diaphragm contracts and increases the volume of the chest cavity, while the muscles of your chest expand your chest cavity to inspire air. Figure 2.4 illustrates how the heart, lungs, and circulatory system work in the process of respiration.

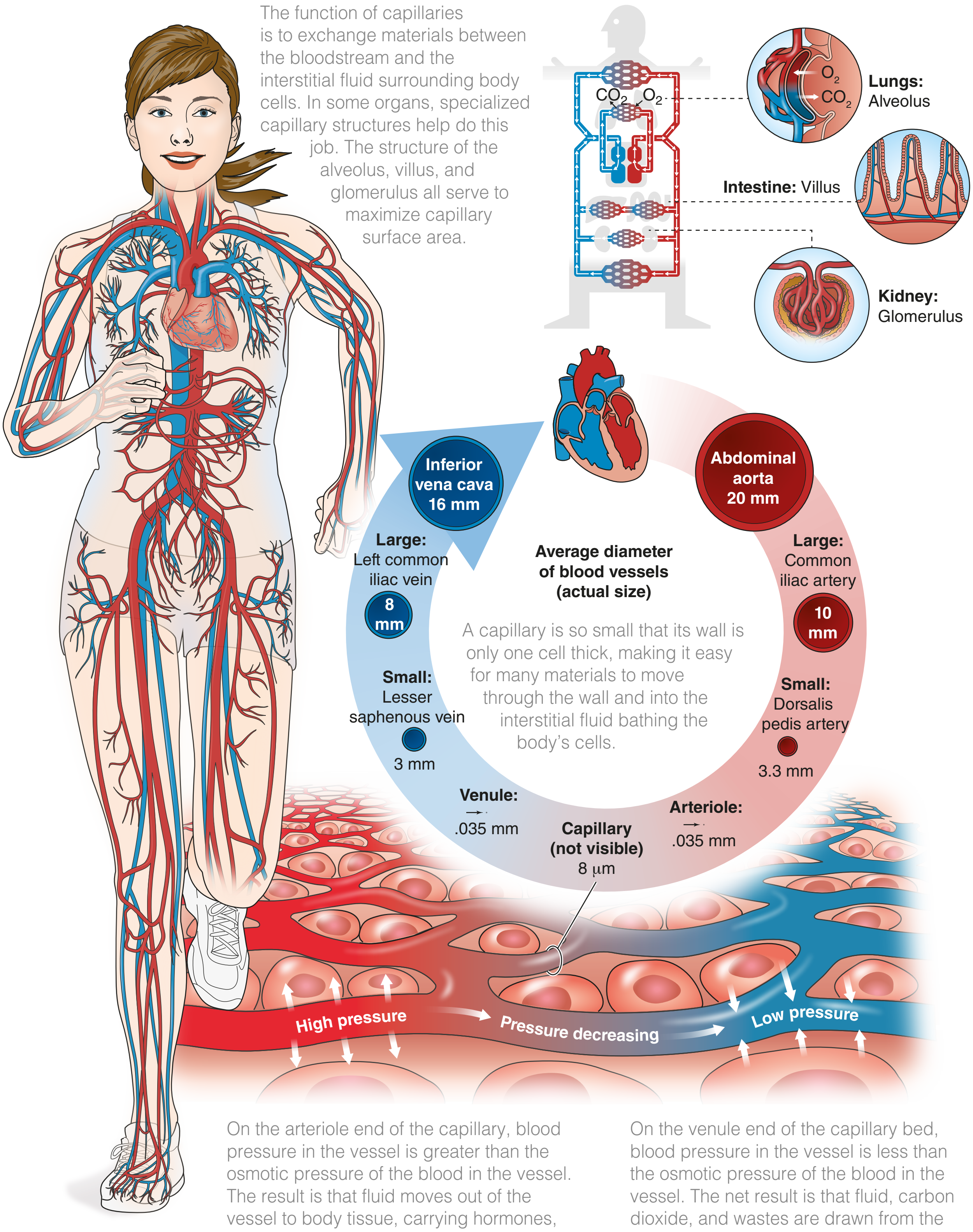

The function of capillaries is to exchange materials between the bloodstream and the interstitial fluid surrounding body cells. In some organs, specialized capillary structures help do this job. The structure of the alveolus, villus, and glomerulus all serve to maximize capillary surface area.

A capillary is so small that its wall is only one cell thick, making it easy for many materials to move through the wall and into the interstitial fluid bathing the body's cells.

On the arteriole end of the capillary, blood pressure in the vessel is greater than the osmotic pressure of the blood in the vessel. The result is that fluid moves out of the vessel to body tissue, carrying hormones, small proteins, oxygen, and nutrients to cells.

On the venule end of the capillary bed, blood pressure in the vessel is less than the osmotic pressure of the blood in the vessel. The net result is that fluid, carbon dioxide, and wastes are drawn from the body tissue into the capillary vessel.

CO_2
O_2
O_2
CO_2
Lungs: Alveolus
Intestine: Villus
Kidney: Glomerulus
Inferior vena cava 16 mm
Abdominal aorta 20 mm
Average diameter of blood vessels (actual size)
Large: Left common iliac vein
8 mm
Large: Common iliac artery
10 mm
Small: Lesser saphenous vein
3 mm
Small: Dorsalis pedis artery
3.3 mm
Venule:
.035 mm
Capillary (not visible)
8 µm
Arteriole:
.035 mm
High pressure
Pressure decreasing
Low pressure

Figure 2.4 The cardiorespiratory system.

Blood consists of plasma (which is a colorless liquid) and a variety of cells. Hemoglobin, a blood component, is the primary oxygen-carrying mechanism in the blood. Approximately 45 percent of the blood is hemoglobin. Hemoglobin releases oxygen when it reaches tissues that need oxygen.

When the tissues use oxygen, they produce carbon dioxide. The carbon dioxide diffuses into the venous system and into the lungs in exchange for oxygen. This completes a circulatory cycle, which takes about 30 seconds.

Carotid sinuses on each side of the neck sense blood pressure within the circulatory system. Excessive pressure on the carotid sinuses during exercise causes the heart to slow when it should be working hard to meet the oxygen demands of the body. Decreased output from the heart can lead to insufficient oxygen for the brain, which can cause unconsciousness. A blackout caused by pressure on the carotid sinuses (see figure 2.5) is a carotid sinus reflex. Therefore, beware of diving equipment that is tight around your neck.

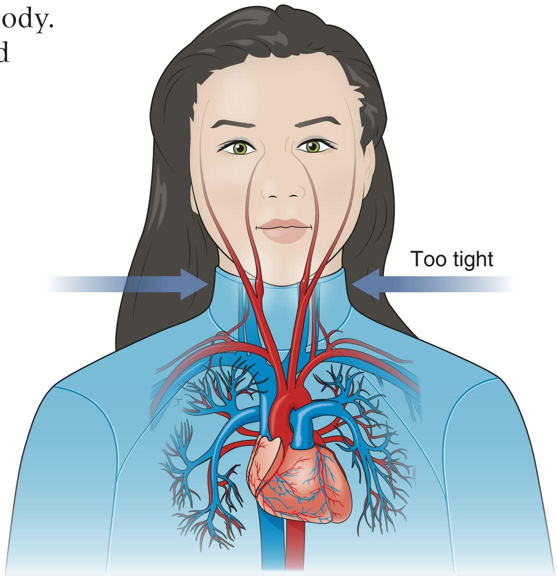

Figure 2.5 Excessive pressure on the carotid sinuses can lead to unconsciousness.

Exhalation is usually a passive process. To exhale carbon-dioxide-laden air from the lungs, the diaphragm relaxes, and the elasticity of the chest cavity forces air from the lungs. You ventilate your lungs approximately 12 to 20 times per minute when at rest. Respiration functions automatically. The key to respiration is the level of carbon dioxide in your circulatory system. When the carbon dioxide in your body reaches a certain level, your brain stimulates respiration. When you voluntarily hold your breath, the buildup of carbon dioxide within your body urges you to breathe. Many people believe that the amount of oxygen in the body controls respiration, but it is primarily the level of carbon dioxide that regulates breathing.

Hyperventilation is rapid, deep breathing in excess of the body's needs. Limited hyperventilation—three or four breaths—enhances breath holding (see figure 2.6*a*). But if you hold your breath after excessive hyperventilation, you may lose consciousness without warning before being stimulated to breathe (see figure 2.6*b*). A breath-holding diver who loses consciousness from lack of oxygen usually blacks out near the surface during ascent. The sudden loss of consciousness near the surface is called **shallow-water blackout**. Loss of consciousness while in the water can cause drowning. You should avoid excessive hyperventilation.

If you breathe rapidly and shallowly, carbon dioxide continues to build in your system, but you do not expel it from your lungs. Inadequate breathing is **hypoventilation**. Shallow breathing is dangerous, especially when you exert yourself, because you can lose consciousness from lack of oxygen. You need to breathe sufficiently to exchange the air in your lungs.

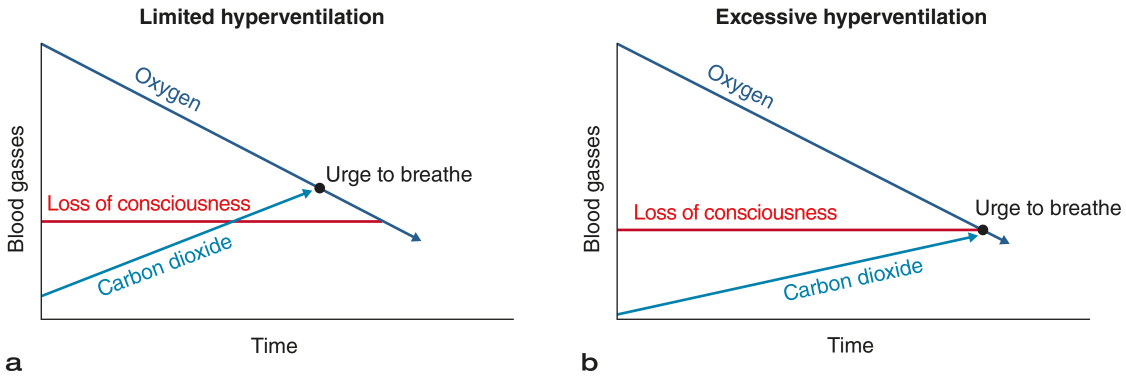

Figure 2.6 Charts showing (*a*) limited hyperventilation and (*b*) excessive hyperventilation.

Contrasts in Air and Water Environments

We live immersed in air, which is a fluid. Air has weight and takes up space. We don't pay much attention to our immersion in air because we are adapted to this environment. We have lived in it all of our lives, and we cannot see the air. The weight of the atmosphere does affect us, however.

Air weighs about 0.08 pound per cubic foot (1.28 mg per cubic cm) at sea level. As altitude increases, air becomes thinner, so its weight per volume is less in the mountains than it is at the seashore (see figure 2.7). The change in the

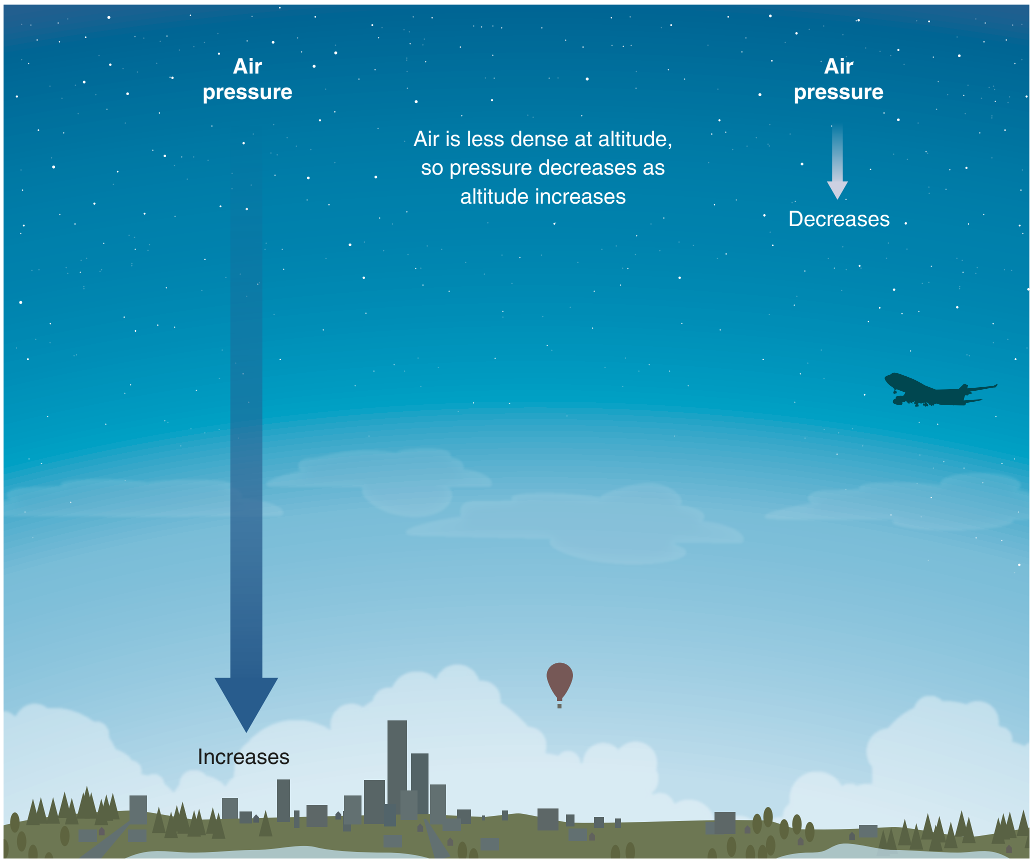
Air pressure
Air is less dense at altitude, so pressure decreases as altitude increases
Air pressure
Decreases
Increases

Figure 2.7 Air density and pressure are affected by altitude.

weight of air affects the air spaces in our ears when we fly or when we drive in the mountains.

Density

Density is weight per unit volume (for example, pounds per cubic foot). Water is a fluid, but it is much heavier than air. Seawater weighs about 64 pounds per cubic foot (1.025 g per cubic cm), which makes it about 800 times denser than air. Freshwater, because it does not contain salt, weighs a little less than seawater: 62.4 pounds per cubic foot (1 g per cubic cm). Temperature affects the density of water, air, and all fluids; cold water is slightly denser than warm water.

Air can be compressed, but water is essentially incompressible. Air becomes thinner as altitude increases, whereas water density remains constant throughout the water column.

Drag

Drag is a force that slows movement. Resistance to movement is much greater in water than in air. Factors affecting drag include the viscosity of the fluid, the speed of motion, and the size and shape of the object moving through the fluid. Drag is greater when the fluid is denser, the motion is faster, the object is larger, or the surface of the object is more irregular (see figure 3.9 in chapter 3).

Because of water's higher density, the molecules of water are much closer together than those of air. The tightly packed molecules affect the transmission of light, sound, and heat (see table 2.1). Light travels about 27 percent slower in water than in air. Sound travels about 4 times faster in water than in air. Still water conducts heat nearly 25 times faster than air does and moving water can

Table 2.1 Water Compared With Air

Property	Air	Water	Comparison	Effects
Density	0.08 lb/ft^3 (1.3 kg/m^3)	62.4 to 64 lb/ft^3 (1,000 kg/m^3)	Water is 800 times denser than air.	Resistance to movement
Compressibility	Yes	No	Air density varies; water density is constant (at dive pressures).	Affects body & attached air spaces
Speed of light	186,000 mi/s (300,000 km/s)	140,000 mi/s (225,400 km/s)	Light travels 27% more slowly in water.	Affects vision
Light absorption	Low	High	Water absorbs color quickly.	Light & color loss
Speed of sound	1,125 ft/s (340 m/s)	4,900 ft/s (1,400 m/s)	Sound travels 4 times faster in water.	Unable to determine source
Conductivity	0.17	3.86 to 4.12	Heat loss is 22 to 24 times faster in water than in air.	Rapid loss of body heat
Heat capacity	0.24	0.94 to 1.0	The heat capacity of water is 4 times greater than air.	Absorbs heat quickly

conduct heat hundreds of times faster than air. Water has an enormous capacity for absorbing heat with little change in its temperature. The higher density of water affects you in many ways when you dive. The next chapter explains how to deal with the effects of water density.

Buoyancy

An object's ability to float in a liquid depends on the density of the object compared with the density of the fluid in which the object is immersed. Water exerts pressure equally in all directions, even upward. You can feel the upward force (**buoyancy**) of water when you try to push something under the water. Buoyancy results from the difference in pressures on the upper and lower surfaces of an object. The weight of an object plus the weight of the fluid (air, water, or both) above the object exert a downward force. Fluid pressure pushes upward from below. The difference between these two forces is the buoyancy of the object.

Archimedes, an ancient Greek scientist, discovered that the force of buoyancy acting on a submerged object equals the weight of the water displaced (this is known as **Archimedes' principle**). A hot-air balloon floats in air because the hot air inside the balloon weighs less than the volume of cooler air the balloon occupies. A diver is buoyed upward with a force equal to the weight of the water that the diver displaces (see figure 2.8). If you and your equipment weigh less than the weight of the water you displace, you will float, or have positive buoyancy. If you and your equipment weigh more than the water being displaced, you will sink. An object that sinks has negative buoyancy. If you and your equipment weigh exactly the same as the water displaced, you neither float nor sink. Instead, you remain at the depth where you are; you have **neutral buoyancy**.

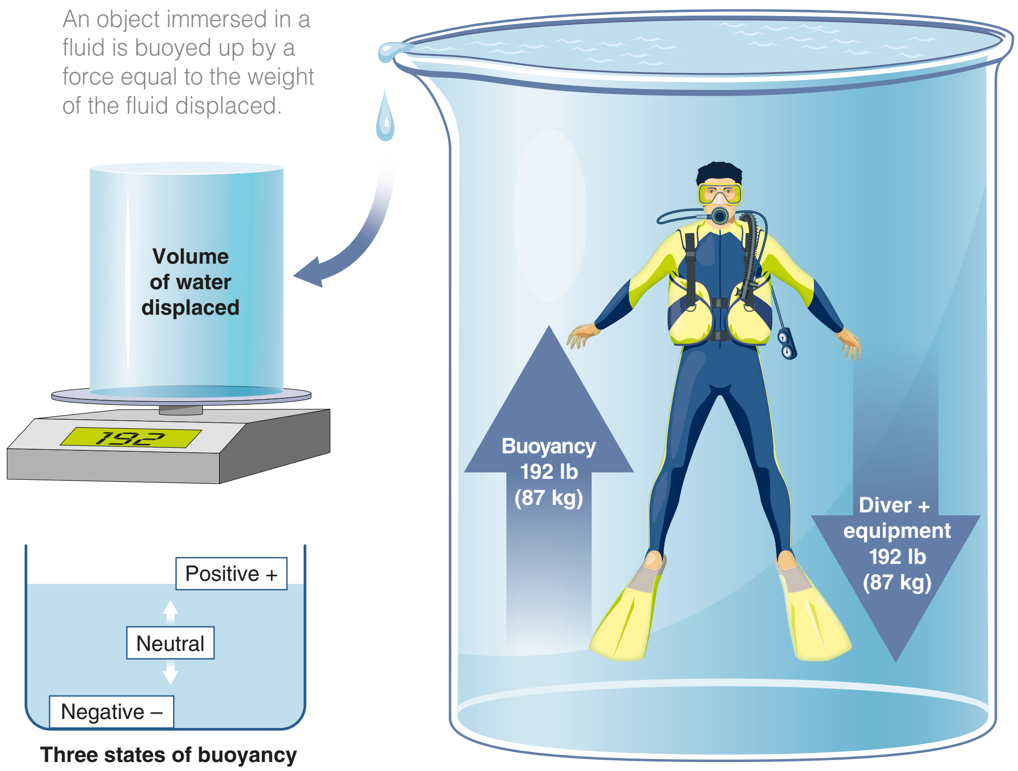
An object immersed in a fluid is buoyed up by a force equal to the weight of the fluid displaced.

Three states of buoyancy

Figure 2.8 Principles of buoyancy.

As a diver, you can float at the surface, sink to the bottom, or hang suspended between the bottom and the surface. If the volume of an object increases with very little change in its weight, buoyancy increases. This happens when you add air to an inflatable jacket or vest. The next chapter addresses the factors affecting buoyancy and the principles of buoyancy control.

Pressure Measurement

Force (often weight) per unit area—such as pounds per square inch or grams per square centimeter—is **pressure**. The envelope of air surrounding the earth is the atmosphere. The weight of one square inch of the atmosphere at sea level is 14.7 pounds (1.03 kg per square cm), or 1 atmosphere (atm) of pressure. As you descend in water, the weight of the fluid—the pressure—exerted on each square inch of your body increases. One square inch of saltwater that is 33 feet (10.1 m) in height weighs 14.7 pounds, 1 atm, or 1.01 bar. One square inch of freshwater 34 feet (10.36 m) in height also exerts a pressure equivalent to 1 atm. Because 1 bar is almost equal to 1 atm, we'll consider them identical in this book. Because water does not compress (at the pressures involved with recreational diving), it follows that water pressure increases by 1 atm for every 33 feet of saltwater (33 FSW) and for every 34 feet of freshwater (34 FFW). Figure 2.9 shows how **atmospheric pressure** and water pressure are measured.

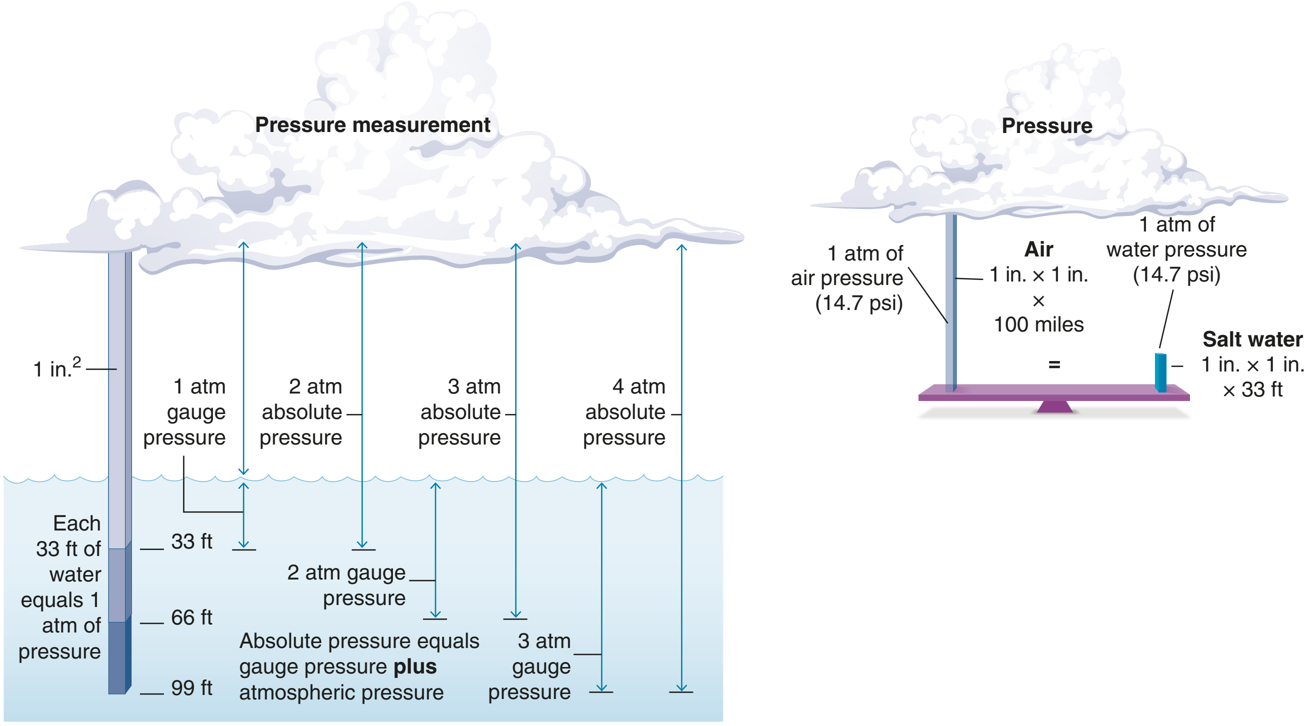
Pressure measurement
1 in.²
1 atm gauge pressure
2 atm absolute pressure
3 atm absolute pressure
4 atm absolute pressure
Each 33 ft of water equals 1 atm of pressure
33 ft
66 ft
99 ft
2 atm gauge pressure
Absolute pressure equals gauge pressure **plus** atmospheric pressure
3 atm gauge pressure
Pressure
1 atm of air pressure (14.7 psi)
Air
1 in. × 1 in. × 100 miles
=
1 atm of water pressure (14.7 psi)
Salt water
1 in. × 1 in. × 33 ft

Depth		Gauge pressure			Absolute pressure		
FSW	FFW	In atm		In psig	In ATA		In psia
0	0	0	=	0.0	1 ATA	=	14.7
33	34	1	=	14.7	2 ATA	=	29.4
66	68	2	=	29.4	3 ATA	=	44.1
99	102	3	=	44.1	4 ATA	=	58.8
132	136	4	=	58.8	5 ATA	=	73.5

Gauge pressure compared to absolute pressure in fresh and salt water.

Figure 2.9 Principles of pressure measurement.

The reference for pressure is either the atmospheric pressure at sea level or zero pressure (outer space). A pressure gauge that reads zero at sea level displays only the pressure in excess of one atmosphere. Tire gauges and depth gauges are good examples of instruments that indicate **gauge pressure**, which may be measured in pounds per square inch gauge (psig), feet or meters of seawater, or bars of pressure.

The total pressure exerted is what matters to divers. The pressures of both the atmosphere and the water apply to diving. The reference for the total pressure is zero, as in a vacuum. The total pressure is called **absolute pressure**, which is measured in pounds per square inch absolute (psia). When people express absolute pressure increments in atmospheres, they use atmospheres absolute (ATA).

You obtain absolute pressure by adding atmospheric pressure to gauge pressure. Be sure you understand the concept because we'll use absolute pressure when dealing with the effects of pressure in this and later chapters.

Gas Laws

When you compress a quantity of gas, you reduce its volume and increase its density and temperature. Allowing a quantity of gas to expand increases its volume, decreases its density, and decreases its temperature. Because divers use gases, you need to understand the relationships between these properties (pressure, volume, density, and temperature) of gases.

Boyle's Law

Boyle's law states that for any gas at a constant temperature, the volume varies inversely with the absolute pressure, while the density varies directly with the absolute pressure. If you compress a closed, flexible air space (such as a balloon), you reduce its volume in proportion to the increase in pressure. When you double the pressure, a closed, flexible air space occupies only half the volume that it did originally. No air is lost. The molecules compress into a smaller area. The density of the air is twice as great as it was originally. When you return the compressed air space to its original pressure, the air inside expands until the object reaches its original volume. You compress your lungs during a breath-hold descent, and they return to normal volume when you return to the surface, provided that you do not expel air underwater.

Scuba equipment provides air to you at the exact pressure of the surrounding water. This allows you to expand your lungs to their normal volume regardless of the depth. The density of the air inside the lungs increases in proportion to the water pressure. If the water pressure doubles, the density of the air in your lungs also doubles. Figure 2.10 shows the relationships between pressure, volume, and density for a gas-filled, flexible-walled container.

Boyle's law also applies when you reduce the surrounding pressure. As outside pressure decreases, compressed air in a closed, gas-filled, flexible container expands in proportion to the reduction in pressure; for instance, if the pressure halves, the volume doubles. If a container filled with compressed air at depth is vented correctly during ascent, expanding air escapes through the vent, and the container remains full throughout the ascent. If the container is not vented, pressure inside increases when the container reaches its maximum volume. If the container is weak, the increase in pressure will rupture the container. This concept is important to scuba divers, who have many air spaces filled with

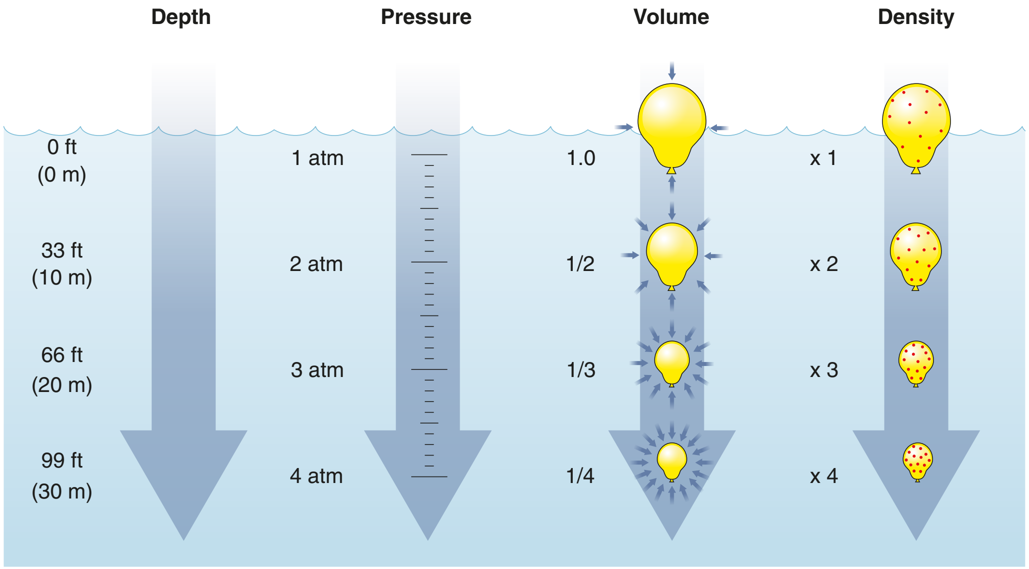

Figure 2.10 Pressure, volume, and density relationships.

compressed air. Vented air spaces do not pose a hazard. But if your lungs are not vented during ascent, life-threatening injuries will result. If you do not vent air from a flotation jacket during ascent, control of buoyancy will be lost as air expands and the jacket volume increases.

Figure 2.10 shows an interesting point about the rate of change of pressure (and volume) in water. The pressure doubles from 1 atm to 2 atm in 33 feet (10 m) of seawater. Doubling the pressure again requires a depth of 99 FSW (30 m). Note that you must ascend from 99 feet to a depth of 33 feet—a distance of 66 feet (20 m)—to experience the same rate of change of pressure that you experience when you ascend from 33 feet to the surface. In other words, the closer you get to the surface, the greater the rate of change of pressure (and of the volume of an air space). You must be more attentive to compressed air in air spaces the nearer you are to the surface.

The change in pressure that you experience during descents and ascents in water is perhaps the most significant challenge of diving. Changes in pressure have direct, mechanical effects on your body. Pressure imbalance in your body air spaces can cause discomfort. In the air environment, you feel pressure changes as a result of changes in altitude, but pressure changes in water occur at a much greater rate than in air. You can sustain serious injury unless you keep the pressure in air spaces inside and attached to you equalized with the surrounding water pressure. Boyle's law causes **squeezes** and **reverse blocks**, which figure 2.11 illustrates. Knowledgeable and experienced divers routinely equalize pressures to avoid squeezes and blocks. Keeping pressure in air spaces equalized with the surrounding pressure is one of the main subjects of the next chapter.

Gay-Lussac's Law

Boyle's law addresses gas at a constant temperature because the temperature affects the pressure and volume of a gas. Jacques Charles, a French chemist, discovered that the volume of gas at a constant pressure changes with temperature. Joseph-Louis Gay-Lussac, a French scientist, defined the effect of

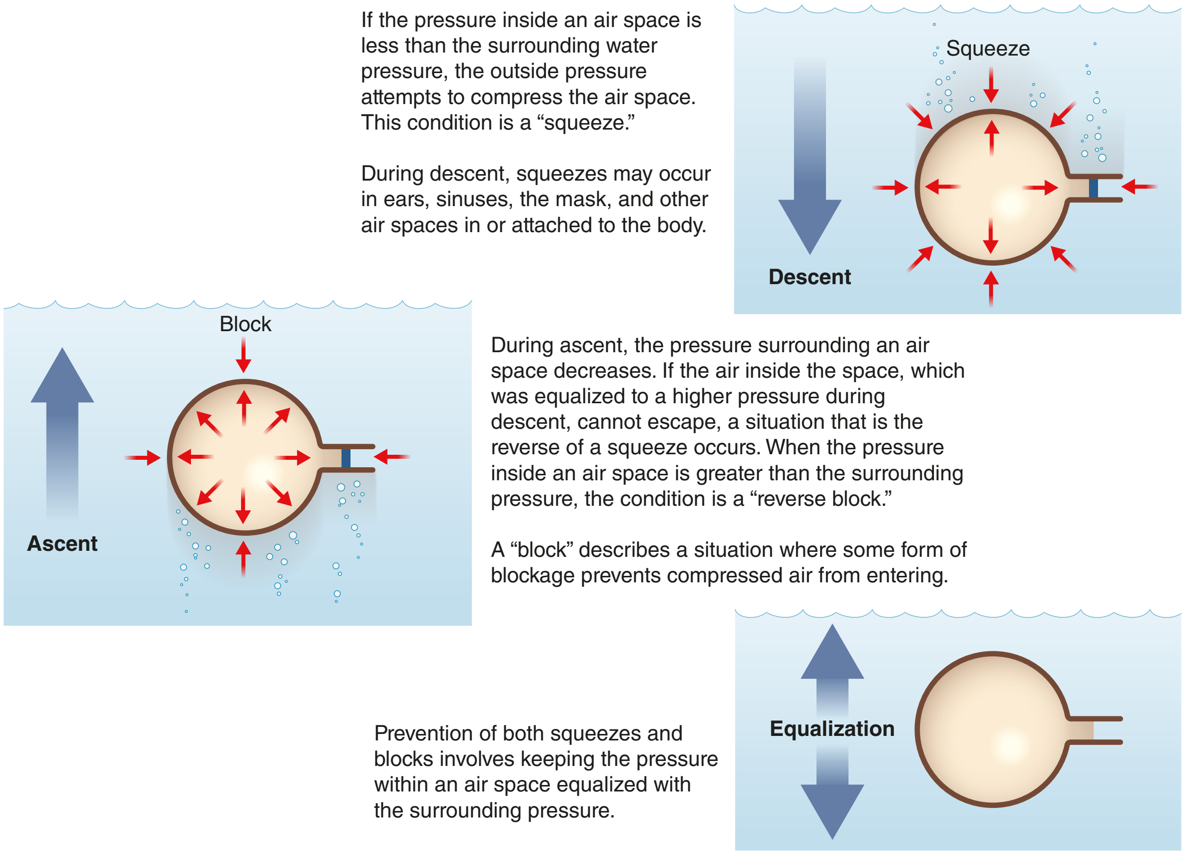

If the pressure inside an air space is less than the surrounding water pressure, the outside pressure attempts to compress the air space. This condition is a "squeeze."

During descent, squeezes may occur in ears, sinuses, the mask, and other air spaces in or attached to the body.

During ascent, the pressure surrounding an air space decreases. If the air inside the space, which was equalized to a higher pressure during descent, cannot escape, a situation that is the reverse of a squeeze occurs. When the pressure inside an air space is greater than the surrounding pressure, the condition is a "reverse block."

A "block" describes a situation where some form of blockage prevents compressed air from entering.

Prevention of both squeezes and blocks involves keeping the pressure within an air space equalized with the surrounding pressure.

Figure 2.11 An explanation of squeezes and reverse blocks when equalizing pressure.

temperature: For any gas at a constant volume, the pressure of the gas varies directly with the absolute temperature. Just as absolute pressure must be used for pressure calculations, absolute temperature must be used for temperature computations. The absolute temperature scale for Fahrenheit temperatures is Rankine. To convert a Fahrenheit temperature to Rankine, add 460 degrees. The absolute temperature scale for a Celsius temperature is Kelvin. To convert a Celsius temperature to Kelvin, add 273 degrees.

You can observe the effect of **Gay-Lussac's law** with a scuba tank, which has a constant volume. Decreasing the temperature of the air in a tank causes the pressure to decrease. Increasing the temperature of the air in a tank causes the pressure to increase. A scuba tank taken from the trunk of a hot car and cooled in water experiences a drop in pressure, although no air leaves the cylinder. Pressure increases or decreases in a standard 80-cubic-foot (2,265 L) aluminum scuba cylinder at a rate of about 6 psig per degree Fahrenheit temperature change. Pressure increases or decreases in a standard 71.2-cubic-foot (2,016 L) steel scuba cylinder at a rate of about 5 psig per degree Fahrenheit temperature change.

Dalton's and Henry's Laws

A gas can diffuse into or out of a substance. When a gas comes into contact with a liquid, the gas dissolves into the liquid. The amount of gas that diffuses into the liquid depends on the density and temperature of the liquid, the pressure of the gas in contact with the liquid, and the length of time that the gas is in contact with the liquid. Another term for the process of gas diffusion into a liquid is **ingassing**. Because the human body is primarily liquid, the gases you breathe diffuse into your body tissues.

In a mixture of gases, such as air, the percentage of the total pressure exerted by each gas is the **partial pressure** of that gas. **Dalton's law** states that the total pressure exerted by a mixture of gases is the sum of the pressures that would be exerted by each gas if it alone were present and occupied the total volume. The partial pressure of a gas determines the amount of that gas that dissolves into a liquid. **Henry's law** expresses gas absorption: The amount of a gas that dissolves in a liquid at a given temperature is directly proportional to the partial pressure of that gas. Table 2.2 shows the partial pressures of gases in the air at 1 ATA.

The partial pressure of a gas is the percentage of the gas in the mixture times the absolute pressure of the mixture. The partial pressure of a gas at depth has the same effect as a higher percentage of that gas at the surface. If a mixture of gas contains 2 percent carbon dioxide (CO_2) at sea level (14.7 psia, or 1.03 kg per square cm), the partial pressure of the CO_2 is 0.294 psia (0.021 kg per square cm). If the absolute pressure of the same mixture of gases increases to the pressure found at a depth of 99 feet (30 m) in the ocean (58.8 psia, or 4.12 kg per square cm), the partial pressure of the CO_2 is 1.176 psia (0.082 kg per square cm). The amount of CO_2 sensed by the body at 99 feet is four times greater than the amount that is sensed at the surface. Breathing 2 percent CO_2 at a depth of 99 feet is the same as breathing 8 percent CO_2 at the surface! A high level of CO_2 has a profound effect on respiration. The surface-equivalent effect of partial pressures makes minute amounts of contaminants in breathing gases unsafe at depth. Table 2.3 shows the surface-equivalent effect of partial pressures at various depths.

Table 2.2 Partial Pressures of Gases in Air at 1 ATA

	Percentage	Partial pressure at 1 ATA	Partial pressure of gas at 1 ATA (metric)
Nitrogen	78%	11.466 psia	0.803 kg/cm^2
Oxygen	21%	3.087 psia	0.2163 kg/cm^2
Argon	0.93%	0.137 psia	0.0095 kg/cm^2
Trace gases	0.04%	0.006 psia	0.0004 kg/cm^2
Carbon dioxide	0.03%	0.004 psia	0.0003 kg/cm^2
Totals	100%	14.7 psia	1.03 kg/cm^2

Table 2.3 Surface-Equivalent Effect of Partial Pressures

Depth	Pressure	O_2*a	CO*b	CO_2*c
0 ft (0 m)	1 ATA	20%	20 ppm	2%
33 ft (10 m)	2 ATA	40%	40 ppm	4%
66 ft (20 m)	3 ATA	60%	60 ppm	6%
99 ft (30 m)	4 ATA	80%	80 ppm	8%
132 ft (40 m)	5 ATA	100%	100 ppm	10%

*Breathing 20% oxygen at 132 ft (40 m) has the same effect as breathing 100% oxygen at the surface.

*Breathing a mixture containing 40 ppm CO at 66 ft (20 m) is the same as breathing 120 ppm CO at the surface!

*Breathing a mixture containing 2% CO_2 at 99 ft (30 m) is the same as breathing 8% CO_2 at the surface and causes shortness of breath, rapid breathing, and headache.

Additional Information About Gas Laws

The following gas formulas can be used to make precise mathematical calculations of pressure, volume, and temperature.

Boyle's law: $P_1V_1 = P_2V_2$

P_1 = Initial pressure (psia or ATA)

P_2 = Final pressure (psia or ATA)

V_1 = Initial volume

V_2 = Final volume

Example: A balloon with 2 pints of air floats from 2 ATA to the surface (1 ATA). What is the volume of the balloon at the surface?

$P_1 = 2$ ATA

$P_2 = 1$ ATA

$V_1 = 2$ pints

V_2 = Unknown

Rearranging the formula to solve for V_2, we find that:

$$V_2 = \frac{P_1V_1}{P_2} = \frac{(2\times 2)}{1} = 4 \text{ pints}$$

Partial pressure = absolute pressure 3 percentage of gas

Example: What is the partial pressure of oxygen if the gas constitutes 20 percent of a gas mixture that has an absolute pressure of 58.8 psi?

$$PP = 58.8 \times 0.2 = 11.76 \text{ psia}$$

Gay-Lussac's law: $\frac{P_1}{T_1} = \frac{P_2}{T_2}$

P_1 = Initial pressure (psia or ATA)

P_2 = Final pressure (psia or ATA)

T_1 = Initial temperature (°R or °K)

T_2 = Final temperature (°R or °K)

Example: A scuba tank with a pressure of 2,250 psig and a temperature of 70 °F is heated to a temperature of 150 °F. What is the pressure of the scuba tank at the higher temperature? First, convert readings for pressure and temperature to absolute measures.

P_1 = 2,250 psig + 14.7 psia = 2,265 psia

P_2 = Unknown

T_1 = 70 °F + 460 = 530 °R

T_2 = 150 °F + 460 = 610 °R

Rearranging the formula to solve for P_2, we find that

$$P_2 = \frac{P_1T_2}{T_1} = \frac{2,265\times 610}{530} = 2,607 \text{ psia}$$

2,607 psia – 14.7 psia = 2,592 psig

When a liquid has absorbed all of a gas that it can hold, the liquid is saturated. When you reduce the partial pressure of the gas in contact with the liquid, gas diffuses out of the liquid. This process is **outgassing**. Ingassing and outgassing provide a foundation for the dive computers and tables used to prevent decompression sickness; these tables are presented in chapter 7.

Air Consumption

The volume of air you breathe per minute during exertion is much more than the volume you breathe at rest—up to 17 times more on land and about 14 times more in the water. In the water, pressure on the torso allows only 85 percent of normal inhalation.

Because the density of the air breathed increases with depth, depth significantly affects the rate at which you consume air. For a given level of exertion, a supply of air lasts only half as long at a pressure of 2 ATA as it does at a pressure of 1 ATA. With heavy exertion at a pressure of 4 ATA (at a depth of 99 feet, or 30 m), you exhaust an air supply over 40 times faster than you would when at rest at the surface! The rapid depletion of your air supply is one reason that you must avoid heavy exertion while diving.

The rate of air consumption is expressed in cubic feet per minute (or liters per minute) or psig (or atmospheres or bars) per minute. By knowing your consumption rate for various levels of activity, you can plan your dives. When you know your consumption rate and the amount of air available, you can calculate air supply duration for future dives.

Heat, Humidity, Light, and Sound

You experience many changes when you enter water. You lose body heat faster, and you lose body moisture when you use scuba equipment. In addition, what you see is deceiving, and what you hear can cause confusion. When you understand what happens to you in water (and why it happens), you can better manage the differences between the water and air environments.

Heat Transfer

The net effect of the various forms of heat transfer is that you can chill quickly while diving. Radiation, convection, and conduction transfer heat from one medium to another (see figure 2.12). Heat waves radiate from exposed surfaces, heat travels upward through fluids by convection, and heat is transferred directly via conduction between substances in contact with each other. Metals are good conductors. Water is a poor conductor compared with metal, but water conducts heat about 25 times faster than air (depending on density). Conduction and convection are the primary means by which heat is transferred from a diver to the surrounding water. Heat rises from the skin, and water carries the heat away. You also lose body heat through the process of evaporation. Moisture evaporates from your lungs when you are breathing underwater and from the surface of your skin when you perspire above water. Scuba equipment expands high-pressure air and cools it. Your body heat warms the air you breathe, and you lose the heat with each exhalation.

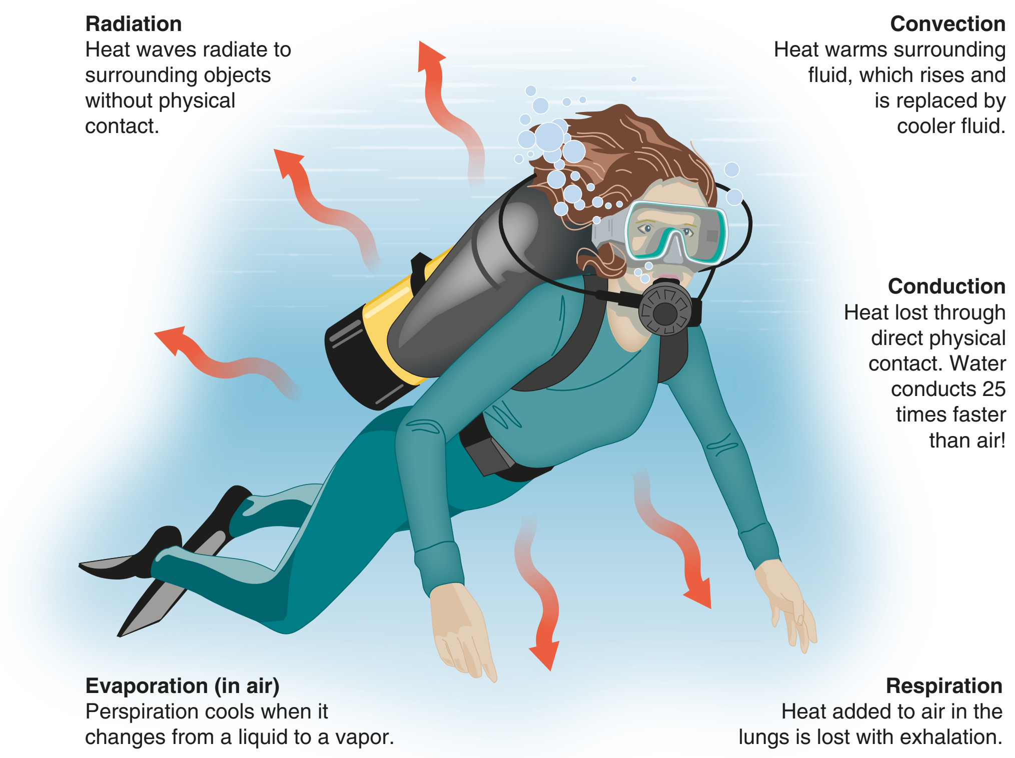

Losing or retaining excessive heat is dangerous!

Radiation
Heat waves radiate to surrounding objects without physical contact.

Convection
Heat warms surrounding fluid, which rises and is replaced by cooler fluid.

Conduction
Heat lost through direct physical contact. Water conducts 25 times faster than air!

Evaporation (in air)
Perspiration cools when it changes from a liquid to a vapor.

Respiration
Heat added to air in the lungs is lost with exhalation.

Figure 2.12 Heat transfer and loss.

You can slow the transfer of heat by insulating yourself with a material that is a poor conductor of heat. Exposure suits help insulate you from the environment, but insulation does not help reduce heat lost through respiration. The next chapter presents ways to manage the problems of heat loss.

Humidity

Scuba divers must be aware of the effects of humidity, or the amount of water vapor present in a gas. The temperature of the gas determines the amount of water vapor a gas can absorb and retain. The warmer the gas, the more humidity the gas can contain.

You humidify inspired air. The process of compressing the air that is put into scuba tanks dehumidifies the air in the cylinder. You lose moisture from body tissues when you breathe dry scuba air. The resulting fluid loss can cause partial dehydration, an undesirable condition, especially for a scuba diver. In the next chapter, you will learn how to avoid the problems of dehydration.

Diving poses other humidity challenges that you must manage. Moisture in the air inside your mask condenses on the faceplate as the air cools. Unless you thoroughly clean your mask lens in advance so that the condensation runs off in a thin sheet, foggy beads of condensation will form and blur your vision. Chapter 6 presents the process for cleaning, or defogging, your mask.

Additional Information About Air Consumption

The following air consumption formulas can be used to calculate consumption rates and air supply duration. To calculate the rate of air consumption, you need three items of information: the depth at which you have remained for a period of time, the length of time you have remained at that depth, and the amount of air you have used during that time.

1. Determine your depth air consumption rate (DACR). This is simply the amount of air used divided by the time at depth. For example, the DACR for a diver who uses 1,000 psi (68 atm) in 10 minutes is 100 psi (6.8 atm) per minute.

$$\text{DACR} = \frac{\text{Air used}}{\text{Time at depth}}$$

2. Convert the DACR to the surface air consumption rate (SACR). You need to express the rate in terms of volume rather than pressure. After you do this, you can apply the air consumption rate to any depth and to a cylinder of a different size from the one used initially to calculate the air consumption rate. Obtain the surface consumption rate by multiplying the DACR by the ratio of the pressure at the surface to the pressure at depth. Because you can express pressure in terms of depth, you can use the following formula:

$$\text{SACR} = \text{DACR} \times \frac{33\text{ ft (or 10 m)}}{\text{Diving depth} + 33\text{ ft (or 10 m)}}$$

   If, for example, your depth consumption rate for a depth of 33 feet (10 m) is 30 psi (2 atm) per minute, your SACR is 30 (33/66) = 15 psi per minute, or 2 (10/20) = 1.0 atm per minute.

3. Convert the rate to volume, establish a ratio of the tank volume and pressure (when the tank is full) to the breathing rate volume and pressure, then solve for the breathing rate volume (BRV) as follows:

$$\frac{V_1}{P_1} = \frac{V_2}{P_2}$$

   where

   V_1 = Full tank volume

   V_2 = Breathing rate volume (BRV)

   P_1 = Full tank pressure

   P_2 = Breathing rate pressure

   so

$$\text{BRV} = \frac{V_1 \times P_2}{P_1}$$

   For example, the breathing rate volume (BRV) for a diver with an 80-cubic-foot (2,265 L), 3,000 psi (204 atm) tank and an SACR of 30 psig (2.04 atm) per minute is

   BRV = 80 ft^3 3 30 psi/min/3,000 psi = 0.8 ft^3/min

   Metric BRV = 2,265 L 3 2.04 atm/min/204 atm = 22.65 L/min

4. For the same level of activity, you can calculate the approximate duration (in minutes) of any amount of air from a tank of any size used at any depth. Here is an example: How long will 1,750 psi (119 atm) of air from a 71.2-cubic-foot (2,016 L), 2,475 psi (168 atm) tank last at a depth of 70 feet (21.3 m) for a diver with a breathing rate volume (BRV) of 0.8 cubic feet (22.7 L) per minute?

First, determine the volume of air in the tank at a pressure of 1,750 psi (119 atm). The formula for determining the volume of air in the tank is

$$V_2 = \frac{V_1 \times P_2}{P_1}$$

where

V_1 = Full tank volume

V_2 = Partially filled tank volume

P_1 = Full tank pressure

P_2 = Partially filled tank pressure

The air supply volume for the partially filled tank is therefore

$$V_2 = \frac{71.2 \times 1{,}750 \text{ psi}}{2{,}475 \text{ psi}} = 50.3 \text{ ft}^3\text{, or}$$

$$V_2 = \frac{2{,}016 \text{ L} \times 119 \text{ atm}}{168 \text{ atm}} = 1{,}428 \text{ L}$$

The formula for air supply duration (ASD) is

$$\text{ASD} = \frac{\text{Air supply volume}}{\text{BRV}} \div \frac{\text{Diving depth} + 33 \text{ ft (or 10 m)}}{33 \text{ ft (or 10 m)}}$$

The air supply duration (ASD) for the question posed earlier is

$$\text{ASD} = \frac{50.3 \text{ ft}^3}{0.8 \text{ ft}^3 / \text{min}} \div \frac{70 \text{ ft} + 33 \text{ ft}}{33 \text{ ft}} = 20 \text{ min}$$

$$\text{Metric ASD} = \frac{1{,}428 \text{ L}}{22.7 \text{ L / min}} \div \frac{21.3 \text{ m} + 10 \text{ m}}{10 \text{ m}} = 20 \text{ min}$$

These calculations may seem complicated at first, but the ideas are simple. The calculations become easy with practice. Let's review the four steps of air consumption calculations: (1) Determine your depth air consumption rate (DACR); (2) determine your surface air consumption rate (SACR); (3) determine your breathing rate volume (BRV); and (4) determine the air supply duration (ASD) for a quantity of air. The abbreviated formulas for the calculations are as follows:

$$\text{DACR} = \frac{\text{Air used}}{\text{Time at depth}}$$

$$\text{SACR} = \text{DACR} = \frac{33 \text{ ft (or 10 m)}}{\text{Diving depth} + 33 \text{ ft (or 10 m)}}$$

$$\text{BRV} = \frac{V_1 \times P_2}{P_1}$$

$$\text{ASD} = \frac{\text{Air supply volume}}{\text{BRV}} \div \frac{\text{Diving depth} + 33 \text{ ft (or 10 m)}}{33 \text{ ft (or 10 m)}}$$

In freezing temperatures, moisture from your exhaled breath can cause a scuba regulator to freeze. Water in other items of diving equipment can also freeze. If you intend to dive in a cold environment, you should complete special training and should know how to prepare and use your equipment in those conditions.

Light and Vision

The density of water makes it challenging for divers to interpret what they see and hear. Light travels faster in air than in water. When rays of light traveling in water pass through the lens of your mask, they accelerate and bend (refract). The effect is that what you see underwater is magnified. Objects appear to be at three-fourths of their actual distance (25 percent closer) and four-thirds of their actual size (33 percent larger). The visual distortion requires adjustments, and you will learn to make these with experience. An object that is 12 feet (3.7 m) away appears to be only 9 feet (2.7 m) away. A fish that appears to be about 2 feet (0.6 m) long is actually only 1.5 feet (0.5 m) in length. Many new divers discover that items they bring back from diving are much smaller than they perceived them to be underwater. Figure 2.13 illustrates how light is perceived differently in water.

One difficulty caused by the refraction of light is that distant objects appear closer than they are. This can create a hazard in clear water when you look downward from a drop-off. You may be tempted to go to a point that appears to be at a safe distance but might actually take you beyond your planned maximum

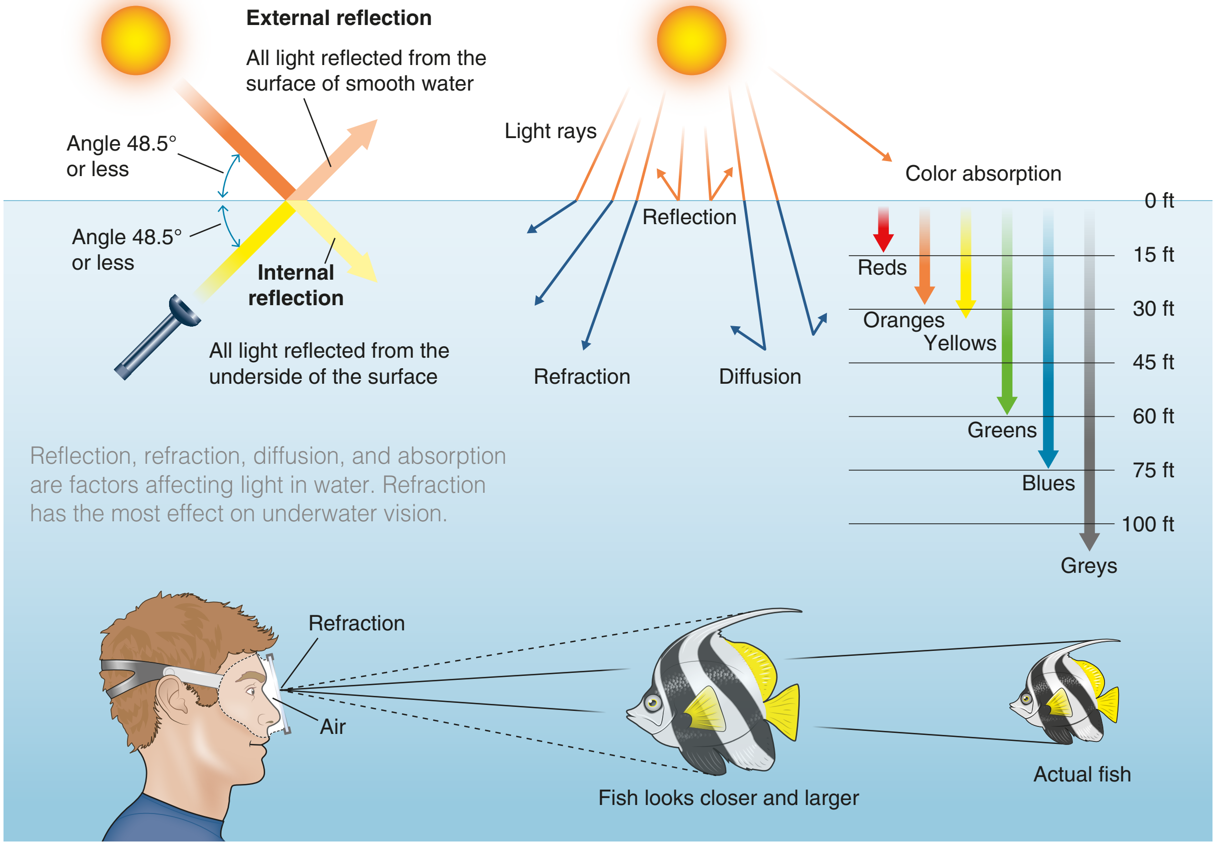
External reflection
All light reflected from the surface of smooth water
Angle 48.5° or less
Angle 48.5° or less
Internal reflection
All light reflected from the underside of the surface
Light rays
Reflection
Refraction
Diffusion
Color absorption
0 ft
15 ft
30 ft
45 ft
60 ft
75 ft
100 ft
Reds
Oranges
Yellows
Greens
Blues
Greys
Refraction
Air
Fish looks closer and larger
Actual fish

Reflection, refraction, diffusion, and absorption are factors affecting light in water. Refraction has the most effect on underwater vision.

Figure 2.13 Visual perception is distorted underwater.

SCUBA WISE

There is an old joke about a young woman who returned home from her first session of scuba diving training and was unenthusiastic when asked whether she enjoyed the class. Her reply was “First the instructor told us all the ways that we could die, then he worked us in the water until I thought that I was going to die.” Fortunately, instructors do not teach classes that way today. The reason for learning about potential injuries is to understand how to prevent them. Armed with knowledge that you can convert to wisdom through proper application, you can avoid all of the potential problems identified in this chapter. The use of scuba diving instruction for fitness training is also outdated. Reasonable physical fitness is a prerequisite for scuba diving. Any healthy person who can swim 200 yards (61 m) continuously is fit enough to dive. The objective of modern scuba training is to help you learn to relax in and under the water, not to whip you into shape.

depth. You must realize that distance perception is inaccurate, and you must rely on your depth gauge instead of your vision.

You have two types of vision: day vision and night vision. You use different parts of your eyes for each type. When you move from a brightly lit area into a dimly lit area, your vision needs 15 to 30 minutes to adapt to the lower level of light. Even after the adaptation, your ability to see fine details is much less than your ability with day vision. In addition, particles in water diffuse, scatter, and attenuate light. The deeper you go, the less light there is. In **turbid** water, the amount of light decreases very quickly with depth. A dive in turbid water involves a change from day vision to night vision.

Water affects light in many ways. The surface of water reflects light. When light strikes the surface of calm water at an angle less than 48 degrees, the water reflects all the light. Early morning and late afternoon sunlight do not penetrate calm water. Light rays going toward the surface from underwater also reflect from the interface, making the underside of the surface appear like a mirror when viewed at the proper angle.

Objects you view underwater often lack their natural colors. White light, such as sunlight, comprises all the colors of the spectrum. Increasing depth absorbs various colors. The water absorbs warm colors, such as red and orange, with only 30 feet (9 m) of depth. Cooler colors, such as blue, penetrate deepest. This is why deep, clear water is blue. The underwater scene appears drab at depths below 100 feet (30 m). Fortunately, you can restore all the colors of the spectrum underwater by using an artificial light at close range.

Another visual challenge is disorientation. When you are weightless in water and do not have a visual reference, **vertigo** can result. Vertigo is the feeling of movement perceived as a spinning sensation. This problem can occur when the water is turbid and visibility is poor, when the water is crystal clear but there are no visual clues for orientation, and when waters of different densities mix together. You will learn how to manage the challenges of disorientation in the next chapter.

A final visual problem is poor visibility in water compared to air. In air environments, visibility is measured in miles, but in water it is measured only in feet. Particles suspended in water cause turbidity, which hampers vision. Reduced visibility makes it difficult to locate objects underwater, to maintain contact with your buddy, and to find your way. You must develop skills to cope with these challenges, including search techniques, buddy system techniques, and underwater navigation skills.

Sound and Hearing

Sound travels well in water. You can hear many sounds while diving. On land you can determine the direction of a sound by the difference in the time it takes the sound to reach one ear and then the other. This interval is brief but sufficient for your brain to discern it. Underwater, you hear sound conducted through water and the bones of your head to your inner ears. Sound travels about four times faster in water than in air. The time delay between a sound reaching one ear and then the other is so small that directional discernment is difficult underwater.

Sound does not transfer well from air to water. Only 0.01 percent of sound can travel directly between air and water. You must use special devices to make voice communications effective underwater.

Summary

The aquatic environment affects your body in several ways. Pressure affects your air spaces: your ears, sinuses, and lungs. The rate of change of pressure in water is many times greater than in air and increases as you approach the surface during ascent. Changing pressures also affect the diffusion of gas into and out of liquids. The level of carbon dioxide in your body controls your respiration, and the partial pressure of the carbon dioxide is affected by depth. Temperature also affects the pressure of a constant volume of gas. Depth and activity have the greatest effect on air consumption. Water absorbs heat from the body, so divers need insulation to help prevent excessive heat loss. Humidity can cause several problems for scuba divers. The difference in density between air and water affects your buoyancy, mobility, heat loss, vision, and hearing. Now that you are aware of the effects of the aquatic environment on your body, you are prepared to learn to adjust to the underwater environment.

Application-of-Knowledge (AOK) Questions

1. If you are exerting underwater and find yourself breathing rapidly and deeply and suddenly feel that you are not receiving enough air, what should you do?
2. A sudden change of conditions while you are diving increases your anxiety, and you notice that your breathing is rapid and shallow. Your anxiety level increases because you feel unable to manage the situation in which you find yourself. What action will help alleviate your anxiety and abnormal breathing?
3. While diving with an experienced diver, you compare tank pressures and discover that you are using twice as much air as your buddy, who is the same size as you. You are using equipment that is identical to your buddy's. What could be causing you to have the high rate of air consumption?
4. Maintained scuba equipment is extremely reliable, but if your regulator started leaking water when you inhaled and you were at a depth of 60 feet (18.3 m), what should you do?
5. While swimming in a pool, you notice a swimmer breathing rapidly and deeply before swimming underwater. You keep an eye on the swimmer because he may lose consciousness without warning while submerged. What is the cause of this problem, and what could you tell the swimmer to help him understand how to prevent an accident?
6. You are diving in water that has a temperature of 60 °F. (15.6 °C) and begin shivering. You decide to increase your activity level to produce more body heat. Is this a good course of action?

Cup corals in the Red Sea

3

Diving Adjustments

Dive In and Discover

By the end of this chapter, you will be able to do the following:

- Define the terms *hypothermia*, *hyperthermia*, *vasoconstriction*, *heat exhaustion*, *heatstroke*, *equalization*, *Toynbee maneuver*, *Valsalva maneuver*, *trapdoor effect*, *skip breathing*, *barotrauma*, *pulmonary barotrauma*, *arterial gas embolism*, *mediastinal emphysema*, *subcutaneous emphysema*, *pneumothorax*, *vertigo*, *perfusion*, *ambient pressure*, *half-time*, *compartment*, *controlling compartment*, *decompression illness*, *dehydration*, and *nitrogen narcosis*.
- State the cause, effect, signs and symptoms, first aid treatment, and prevention of hyperthermia, hypothermia, squeezes, reverse blocks, respiratory distress, pulmonary barotrauma, vertigo, seasickness, decompression illness, and nitrogen narcosis.
- Explain three ways to control buoyancy while scuba diving.
- Explain when to equalize pressure in your ears during descent and what to do if the pressure does not equalize.
- Explain how to minimize resistance to movement for underwater swimming.

To function effectively underwater, you need specific attitudes, equipment, knowledge, and skills. This chapter provides the fundamental knowledge and begins shaping the attitudes required to minimize the risk of injury. You will apply the basic aspects of anatomy, physiology, and physics you have learned to a typical person descending into the depths and ascending back to the surface.

Thermal Adjustments

Maintaining your body's core temperature within a few degrees of normal is challenging in water. When you are immersed in water, you lose body heat rapidly. A water temperature of 50 °F (10 °C) can incapacitate an unprotected diver within 15 minutes. Even water at a temperature of 80 °F (27 °C), which feels relatively warm, can chill a diver within an hour. Wearing only a bathing suit in 80 °F (27 °C) water is the same as being without any clothing in air that is 42 °F (6 °C).

Your brain regulates your body functions to maintain your body temperature. If your core temperature is less than 95 °F (35 °C), you will suffer from **hypothermia**. You need to guard against mild and severe hypothermia, both of which can be dangerous. If your core temperature is higher than normal, you experience the effects of **hyperthermia**. You need to understand the effects of two types of hyperthermia—heat exhaustion and heatstroke—and you must try to prevent them. They both can be dangerous.

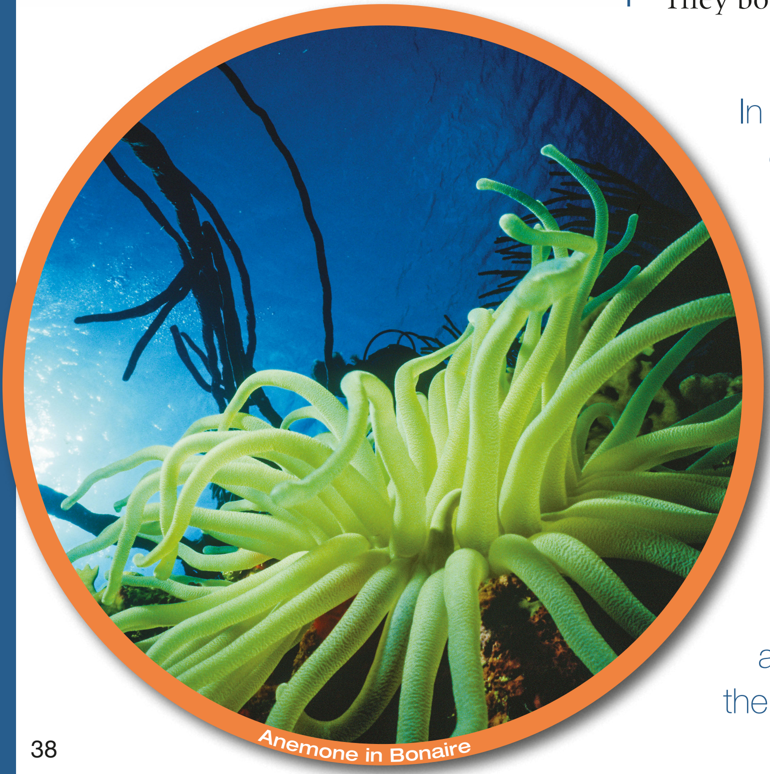
Anemone in Bonaire

In chapter 2, you learned about the effects of water on your body. In this chapter, you will learn how to deal with those effects. Most people with average intelligence and normal health can make adjustments to the aquatic environment, such as buoyancy control and the equalization of pressure. You make many adjustments automatically, but some you must make consciously. Learning to make the transition from an air environment to an underwater environment requires professional instruction and guidance. You need to learn what to do, then do what you have learned.

Heat Loss

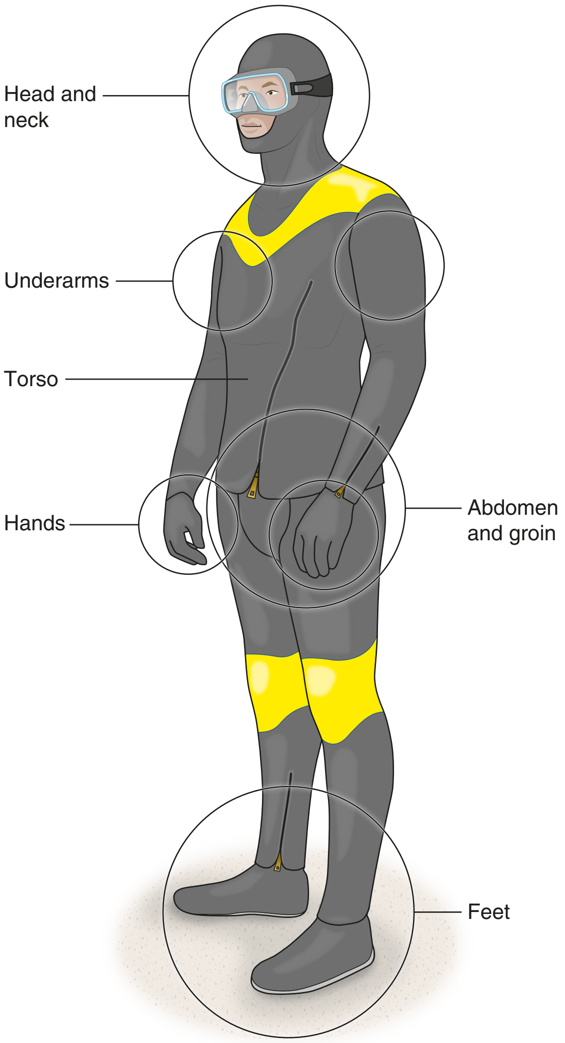

Figure 3.1 Areas prone to high heat loss.

Your body has a variety of physiological responses to the loss of heat. Respiration increases automatically when you get chilled. This is undesirable because you heat and moisturize inspired air, and you lose the heat and moisture with each exhalation. The more you breathe, the more heat and moisture you lose to the environment (and the shorter your dive time). Water depth compounds the problem because the greater the surrounding pressure, the greater the density of the air that you breathe. Denser air absorbs more heat than air that is less dense. The deeper you dive, the quicker you get cold.

Anything that affects the function of your body—excitement, fear, seasickness, and other forms of illness—can increase heat loss. This is why good health and a confident state of mind are safety recommendations.

One way your body responds to cold is to shunt blood from the extremities through **vasoconstriction**. The circulatory shunting reduces heat loss because it keeps warm blood from passing through areas of your body that have little insulation. Your head, neck, underarms and sides, groin, hands, and feet are the areas of your body most prone to heat loss underwater (see figure 3.1). Fortunately, you can insulate these areas easily. In cold water, you can lose considerable heat from your head and neck because the head receives a large supply of blood and lacks natural insulation. Your body does not shunt blood from your head as it does from other body extremities. In water at a temperature of 70 °F (21 °C) or less, you must be sure to insulate your head and neck.

Hands have large surface areas relative to their volume. To prevent excessive heat loss from your hands when you get cold, your body shunts blood from them until they reach a temperature of 50 °F (10 °C). At this temperature, your body restores circulation to your hands to partially rewarm them. Hands quickly lose their warmth to the water. If you dive without hand protection in cold water, you lose body heat through your hands. Because your hands may become numb and lack strength when they get cold, you should insulate them to keep them functioning and to conserve body heat.

Small people chill more quickly in water than large people do. Small people have less muscle mass to generate and store heat. Insulation is important for all divers, but protection against heat loss is more critical for those of smaller stature.

Failure to wear adequate insulation leads to hypothermia. Hypothermia also results from repeated or prolonged exposure. Slow chilling of your body is undesirable. You lose muscle strength and feeling, and your muscles may cramp. Severe heat loss also affects your ability to reason. Another body response to heat loss is shivering, which restores heat through muscular activity. Shivering, which generates about five times as much heat as your body produces at rest, is helpful on land, but is not beneficial in water. Water conducts away the heat you produce by shivering, and you get colder. Uncontrollable shivering indicates that you have lost too much heat from the core of your body and that you cannot rewarm yourself without getting out of the water. When you are shivering, you should terminate the dive. Rewarm yourself thoroughly before diving again. To help your body temperature return to normal, you should put on clothes that are warm and dry, stay in warm surroundings, and consume warm nonalcoholic drinks. You can become so cold that shivering ceases. This occurs because severe hypothermia has incapacitated the body's ability to produce heat. A cold person who is not shivering requires medical care.

You must understand that warming the surface of your body is different from warming the core of your body. You may feel warm, but your deep core temperature may remain below normal. If you return to the water in this condition, you will quickly become chilled. The only way to be sure that you are thoroughly warmed is to keep warming yourself until you begin to perspire. Perspiration occurs when the core temperature begins to rise above normal. To warm yourself, get dry, wrap everything except your face in blankets, and stay that way until you perspire. Most people will shed the blanket when their skin is warm, not being aware that the core of their body is still below normal temperature.

Overheating

You can prevent excessive loss of body heat by insulating your body with an exposure suit, but insulation can cause another problem. When you insulate your body to reduce heat loss in water, you reduce your body's ability to rid itself of excess heat above water. The evaporation of perspiration helps cool your body, but if you cover your body, perspiration cannot evaporate. You may become overheated in warm climates when you are preparing to dive. Maintaining your body temperature within acceptable limits before, during, and after dives can be a challenge. Figure 3.2 describes some of the causes and effects of overheating.

When a person is unable to stop the rise of core temperature, **heat exhaustion** occurs. This condition is serious. A person affected by heat exhaustion becomes weak and may collapse. The victim looks pale and feels sweaty. Place a person with this condition in a cool place, remove the exposure suit, and take steps to lower the person's body temperature. If the victim is alert and not nauseous, have him or her drink a diluted electrolyte fluid or water to replace lost body fluids.

A more serious form of hyperthermia is **heatstroke**, which occurs when the body temperature becomes so high that the body's temperature-regulating ability shuts down. A victim of heatstroke looks flushed and has hot, dry skin. The mental status of a heat stroke victim is not normal. The victim may be delirious, or unresponsive, or have seizures. Heat stroke is extremely serious. Cool the patient's body immediately and summon medical assistance.

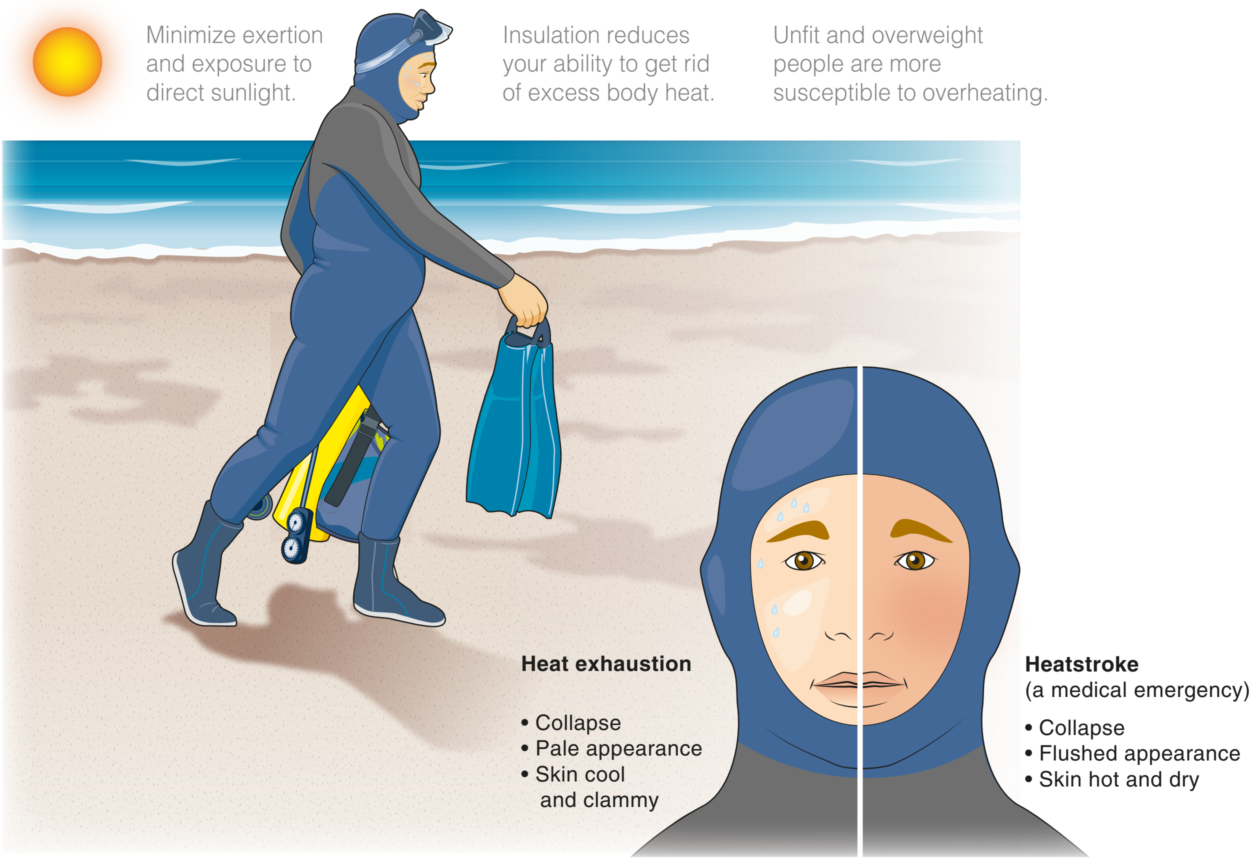
Minimize exertion and exposure to direct sunlight.
Insulation reduces your ability to get rid of excess body heat.
Unfit and overweight people are more susceptible to overheating.
Heat exhaustion
• Collapse
• Pale appearance
• Skin cool and clammy
Heatstroke (a medical emergency)
• Collapse
• Flushed appearance
• Skin hot and dry

Figure 3.2 Take precautions against overheating by learning the symptoms of heat stroke and exhaustion.

Preventing hyperthermia is much better than treating it after it occurs. You should avoid prolonged exposure to warm temperatures when wearing insulation. If the air temperature is warm, douse yourself with water after donning your exposure suit and before donning the remainder of your diving equipment. Stay out of direct sunlight if possible. All thermal considerations for diving are especially important for divers whose physical fitness is marginal.

Buoyancy Adjustments

Exposure suits and other equipment affect your buoyancy. You must adjust the amount of weight you wear to control your depth in the aquatic environment. When you are too buoyant, you must fight to remain submerged; when you carry excess weight, you must work hard to keep from sinking or to stay off the bottom. You need to maintain neutral buoyancy underwater and positive buoyancy at the surface. In chapter 2, you learned the principle of buoyancy and the three states of buoyancy: positive, negative, and neutral. Now you will learn some practical applications of buoyancy.

Your body—which is composed of solids, liquids, and air spaces—has an average density nearly the same as water. A typical human body immersed and relaxed in water has a positive buoyancy of a few pounds when the lungs are filled with air; it has a negative buoyancy of a couple of pounds when the lungs contain the minimum amount of air. Fortunately, the density of water and resistance to movement prevent you from moving up and down rapidly when you breathe.

Factors Affecting Buoyancy

People usually wear some type of exposure suit when diving. Most exposure suits increase buoyancy, so divers wear weights to offset the buoyancy of the suit and achieve neutral buoyancy. These weights are made of lead, which is about 12 times denser than water.

Your initial state of buoyancy in water will depend on the volume of water you and your equipment displace. You can vary your volume—with a negligible increase or decrease in your weight—by adding air to or venting air from your buoyancy compensator (BC). Increasing BC volume increases buoyancy, while decreasing BC volume decreases buoyancy.

Your buoyancy is affected by your physical size and weight, your lung capacity, the equipment you wear, and the items you carry (see table 3.1). Exposure suits use air or small bubbles of gas for insulation. When you descend while wearing an exposure suit, pressure compresses your suit and reduces its volume, so you become less buoyant. You must add air to your BC to compensate for buoyancy lost from suit compression. On the other hand, buoyancy increases as you consume air from your scuba cylinder. Air weighs 0.08 pounds per cubic foot (1.3 g per L). A typical scuba tank contains 80 cubic feet (2,265 L) of air. A full tank weighs 6 pounds (2.7 kg) more than an empty tank. As you consume air from your scuba tank, you must vent some of the air that you added to your BC to compensate for suit compression. The trade-off helps you keep buoyancy constant during a dive.

Inflating your lungs increases your buoyancy, while deflating them reduces buoyancy. A high average lung volume makes you float; a low average volume makes you sink. When you become excited or begin moving quickly, your respiration increases, which affects your buoyancy. For optimal control of buoyancy, you should maintain a calm, relaxed state.

The density of water also affects buoyancy. Saltwater is denser than freshwater, so you are more buoyant in the ocean than in a lake. This means that if you

Table 3.1 Factors Affecting Buoyancy

Factor	Effect
Size and weight of diver	Obese divers are more buoyant.
Type and amount of equipment	Larger equipment is more buoyant (because of increased volume).
Amount of weight worn	Weights decrease buoyancy.
Amount of air in BC	Increasing volume increases buoyancy.
Amount of air in tank	Buoyancy increases as air decreases.
Amount of air in lungs	Exertion or excitement increases volume and buoyancy.
Suit compression	Pressure decreases volume and buoyancy.
Items carried	Added weight decreases buoyancy.
Type of water (salt or fresh)	Denser water increases buoyancy.

SCUBA WISE

At some point, many divers discover a special feeling of becoming part of the underwater world. I remember vividly an overwhelming feeling—a strange combination of peace and exhilaration—that I experienced during a dive in the Red Sea. The goal of training is to help you adjust to a new environment. When you adjust, you can relax, and when you can relax, you can focus more on your exciting new surroundings than on yourself. Meeting the demands of diving is challenging, but when you succeed, the exuberance is worth every adjustment you have to make. People can adjust well to new situations. Training helps people adjust more quickly, easily, and safely than they can by trial and error. Study the remaining chapters carefully to learn how to adjust to the subaquatic environment. I want you to feel what I felt in the Red Sea and have felt many times since while diving. The feeling of being one with the sea is powerful, moving, and unforgettable.

are weighted for neutral buoyancy in the ocean, you must remove some weight to achieve neutral buoyancy for freshwater diving. The amount of weight you remove is about 3 percent of the combined dry weight of you and your equipment. For example, if a neutrally weighted 160-pound (73 kg) diver with 60 pounds (27 kg) of equipment, including 16 pounds (7 kg) of weights, wants to dive in freshwater instead of seawater, the diver must remove about 7 pounds (3 kg) of weight in order to be weighted correctly.

Ways to Control Buoyancy

As a diver, you can control buoyancy in three ways: by the amount of weight you wear and carry, by the amount of air in your BC, and by the amount of air in your lungs. These means of control are coarse, medium, and fine adjustments, respectively. The skills you need to learn in order to adjust to the aquatic environment include determining the correct amount of weight to be worn, regulating the amount of air in your BC, and varying your breathing for minor buoyancy adjustments. Chapter 6 describes these skills.

Pressure Equalization

One of the most important adjustments you must learn is how to handle the effects of pressure changes in water. Pressure changes rapidly as you descend and ascend. To avoid discomfort and injury, you must keep the pressure in the air spaces inside and on your body equalized with the outside pressure. This section identifies the procedures for pressure equalization.

Equalizing Pressure in the Sinuses

Your sinuses equalize pressure automatically as long as they are healthy. But when you have a cold or respiratory illness, the membranes lining your sinuses become swollen. The swelling can close the narrow air passages leading to the sinuses. If you descend with swollen sinus membranes, a sinus squeeze will result. Because the sinuses are formed in bone, they do not compress as a flexible container does. When the pressure inside the sinuses is less than the surrounding pressure, the outside pressure causes the linings of the sinuses to swell and ooze fluids into the cavities. The swelling and fluid reduce the volume and compress the air that is there. Do not attempt this painful method of equalization because your problems are not over after you descend. When you ascend in this condition, the compressed air in the sinuses expands to its original volume. The expansion can force the fluid in your sinuses out through the openings, which are swollen shut. Avoid this painful process; don't dive unless your sinuses are clear and normal. Figure 3.3 shows what happens to both healthy and congested sinuses under pressure.

Do not use medications to relieve stuffiness and congestion caused by an illness before you go diving. Increased pressure can reduce the medication's effects and the duration of its effectiveness. Decongestants do not cure an illness; they simply mask its symptoms. When medication taken to open swollen airways wears off, a rebound effect occurs. The airways become more swollen than they were before you took the medication. If the rebound occurs while you are diving, you can trap high-pressure air in your sinuses. If you are not well enough to dive without medications, you should not dive.

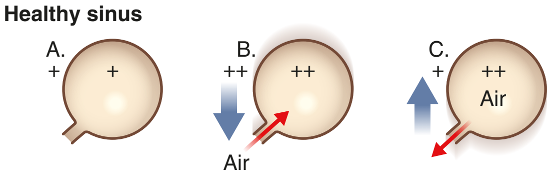
Healthy sinus

Air can move freely in and out of a healthy sinus (A), so equalization is automatic during descent (B) and ascent (C).

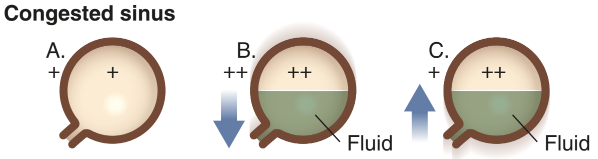
Congested sinus

Air cannot get into or out of a congested sinus (A). The body forces fluids into the sinus at depth to equalize the pressure (B). The air in the sinus tries to expand to its original volume during ascent (C).

Figure 3.3 The mechanisms of equalizing pressure in the sinuses.

Equalizing Pressure in the Ears

You must equalize the pressure in your ears more often than any other air space. Whereas the air passages to your sinuses are normally open, the **eustachian tubes** from your throat to your ears are normally closed. You must learn how to open the tubes at will to allow air to pass through them and equalize pressure in the air spaces of your middle ears.

You can make several movements to open the ends of the eustachian tubes where they connect with the throat. Swallowing, yawning, lifting the base of your tongue, and jutting the jaw forward (either individually or in combination) should produce a cracking sound in your ears. The opening of the tubes causes the sound. Some divers are fortunate because they can use simple movements to equalize pressure in their ears during descent. Most people require a more forceful means of equalization. Many divers use a technique known as the **Toynbee maneuver**, which you do by blocking your nostrils, closing your mouth, and swallowing. The **Valsalva maneuver** is even more forceful and is done by blocking your nostrils, closing your mouth, and gently attempting to exhale.

You must avoid excessive force when you use the Valsalva method because you could damage your ears permanently by rupturing the round window. When pressure in the outer ear increases, the eardrum bulges inward. The bones of hearing in the middle ear transmit the movement to the oval window in the inner ear. When you attempt to exhale against closed airways, you create an internal pressure on your inner ears. The attempted exhalation pressure in conjunction with the external pressure exerted on the oval window can cause the round window to rupture; this is a serious injury that can result in a permanent high-frequency hearing loss and constant ringing in the ear. Because you control how hard you attempt to exhale in a Valsalva maneuver, you can prevent this injury. Be careful!

You must equalize pressure in the air spaces of your middle ear early (before descent) and frequently. If you delay equalization during descent, increasing pressure will hold your eustachian tubes closed—and if you attempt to force air through them, this will only close them tighter. This is the **trapdoor effect**, a difficulty you can avoid by keeping the pressure in your middle ears and your throat equal. You can then open your eustachian tubes and allow air to pass through them. Figure 3.4 shows the trapdoor effect.

You should equalize pressure in your ears before descending and about every 2 feet (0.6 m) for the first 15 feet (4.6 m) of descent, about every 3 feet (0.9 m) from 15 to 30 feet (4.6 to 9.1 m), and as needed thereafter. You can feel and hear air enter your ears when you eliminate any difference in pressure. If you attempt to equalize pressure but cannot get air into your ears, you should ascend several feet to reduce the pressure on your eustachian tubes and try again. If that works, continue your descent. If that does not work, ascend a few more feet and try again. Your initial descents may be somewhat jerky until you become accustomed to equalization techniques.

Failure to equalize pressure in your middle ears is as bad as trying to equalize forcefully. In an unequalized middle ear, pressure bows the eardrum inward. If the squeeze continues, pressure forces blood and fluid into the middle ear. This

Methods to equalize pressure:

- Blowing gently with mouth closed and nostrils blocked.
- Swallowing with mouth closed and nostrils blocked.
- Jutting the jaw forward while yawning with mouth closed.

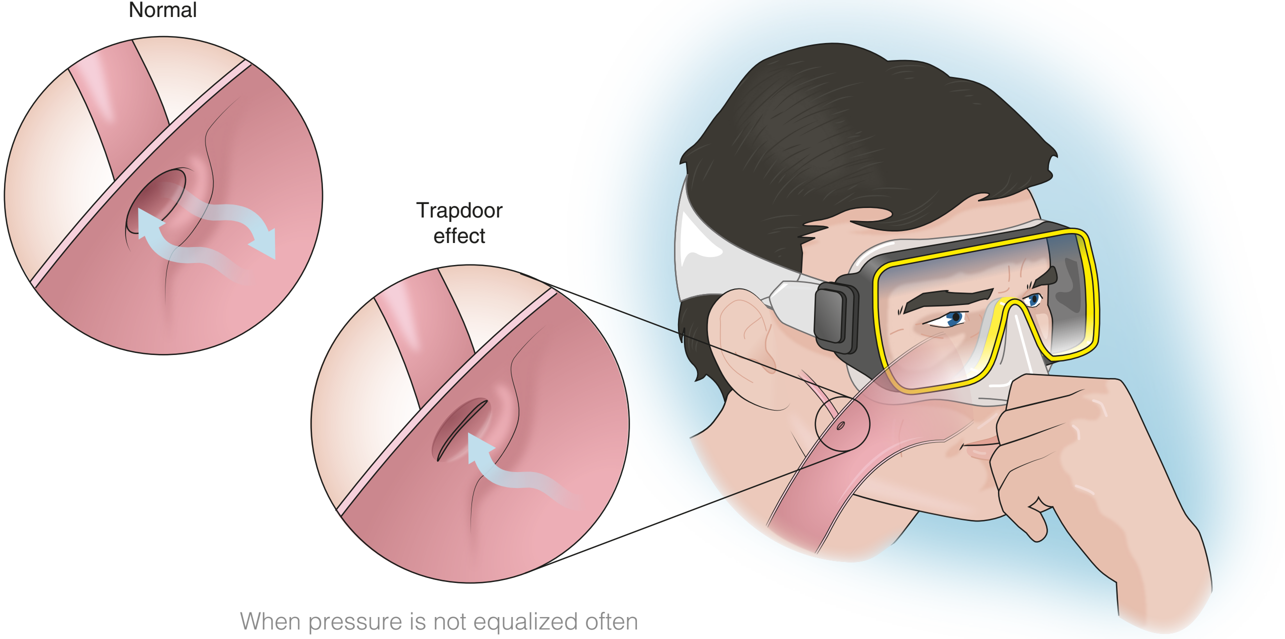

When pressure is not equalized often during descent, the pressure difference holds the end of the eustachian tube closed and prevents any further equalization until the diver ascends enough to relieve the pressure.

Figure 3.4 Helpful techniques for pressure equalization.

process takes time, but it is usually painful. It also damages the ear. If you feel discomfort or pressure in your ears during descent, you should ascend until the pain is gone, and then ascend another couple of feet before trying to equalize.

If you ignore an ear squeeze and continue your descent, the pressure differential can rupture your eardrum. The rupture instantly equalizes the pressure in your middle ear but damages your ear in the process. A rupture of your eardrum causes a temporary loss of hearing and a feeling of fullness in your ear. Cold water rushing into the middle ear can also cause temporary disorientation (see the section on Equilibrium Adjustments and Seasickness later in this chapter). If you suspect an ear injury, see a physician; prompt treatment can minimize the risk of permanent injury.

You can cause an outward rupture of an eardrum if you block the ear canal with a plug or if you cover your ears with a watertight covering. When you obstruct an ear canal, the air in your outer ear remains at surface pressure while the pressure in your middle ear increases with equalization during descent. The difference in pressure between the middle and outer ear pushes the eardrum outward until it breaks. You can prevent such an injury. Do not wear earplugs while diving, and avoid waterproof seals over your ears.

Equalizing is easier when you descend in an upright position as opposed to a head-down position. Membranes line the airways in your head, and gravity affects blood in the vessels within the membranes. When you are upside down in water, the membranes of your air passages swell and narrow.

Difficulties with ear equalization during ascent are uncommon. Air expanding inside the middle ear escapes through the eustachian tube. Air passes out through the tube much more easily than it goes in; you do not have to do maneuvers to open your tubes so the air can escape. But if a plug of mucus happens to block a tube, pressure can build up in the middle ear and cause an uncomfortable reverse block. If you feel pain or pressure in an ear during ascent, stop the ascent. Excess pressure inside the ear usually works its way out if given time. If you are forced to surface, the pain will increase, and an injury can result. Have the ear examined by a physician, especially if you have recurring reverse blocks.

Equalizing Pressure in the Mask

The air space inside the dive mask is affected by changes in pressure. If you descend without increasing the amount of air inside your mask, pressure pulls your face and eyes into the mask slightly. The pulling sensation, if ignored, ruptures capillaries in your eyes and on your face and causes a mask squeeze. After a mask squeeze, the whites of your eyes will be red, and your face will be red and puffy. There is no excuse for a mask squeeze. You can prevent the problem by exhaling through your nose as needed during descent to keep the pressure inside your mask equal to the surrounding pressure.

Breathing Adjustments

Breathing while in water differs from breathing while on land in several ways. You cannot breathe as freely in water as you can on land. Pressure on the lungs from immersion in water prevents you from breathing as fully as you can in air. Also, it is not as easy to get large quantities of air from a scuba system as it is to breathe deeply above water.

When you ascend, compressed air in your lungs expands, which can cause your lungs to rupture unless you allow the excess air to escape. A normal breathing pattern allows expanding air to escape; breath holding does not. A primary rule of scuba diving is to always breathe continuously.

There is a chance that you will inhale some water. This may cause you to choke and gasp when the water strikes your vocal cords. You need to learn the correct methods of breathing while in and under the water.

Lung Overexpansion

The most important aquatic breathing adjustment is overcoming the instinct to hold your breath underwater. When you breathe compressed air at depth, the density of the air in your lungs is greater than it is at the surface. When you ascend, the air in your lungs expands as the surrounding pressure decreases. If you hold your breath, your lungs expand until they reach their maximum volume, then rupture with an ascent of as little as 4 feet (1.2 m). You must avoid breath holding when breathing compressed air underwater. You should breathe continuously or exhale a small amount of air continuously, and you should never hold your breath during ascent after breathing compressed air. Figure 3.5 illustrates why you must breathe continuously, and the section following the figure describes the possible consequences of breath holding while ascending with compressed air in your lungs.

Skip Breathing

Some divers attempt to extend their air supplies by holding each breath for several seconds. This dangerous practice is called **skip breathing**. When you hold your breath, you increase the amount of carbon dioxide in your circulatory system, which increases your urge to breathe and reduces your breath-holding ability. A high level of carbon dioxide in your body will reduce your ability to cope with difficulties that may arise. And if you skip breathe, you may forget to exhale while ascending, or you may have a headache after the dive. You must be sure to breathe continuously when breathing compressed air.

Breathing Problems

Divers can encounter respiratory difficulties if they overexert themselves, breathe contaminated air, inhale water, or run out of air. Fortunately, you can prevent all of these problems.

Overexertion

Scuba equipment allows you to breathe comfortably while underwater. A regulator delivers air with little respiratory effort, but breathing underwater requires more effort than breathing above water. If you do not maintain your scuba equipment properly, the effort required to inhale and exhale can be excessive and can cause respiratory distress.

Scuba regulators have a limited capacity to supply air. Regulators are not designed for activities involving heavy exertion. Commercial divers use helmets with air hoses to supply large amounts of air to meet their requirements while working underwater. You must avoid strenuous activities underwater because

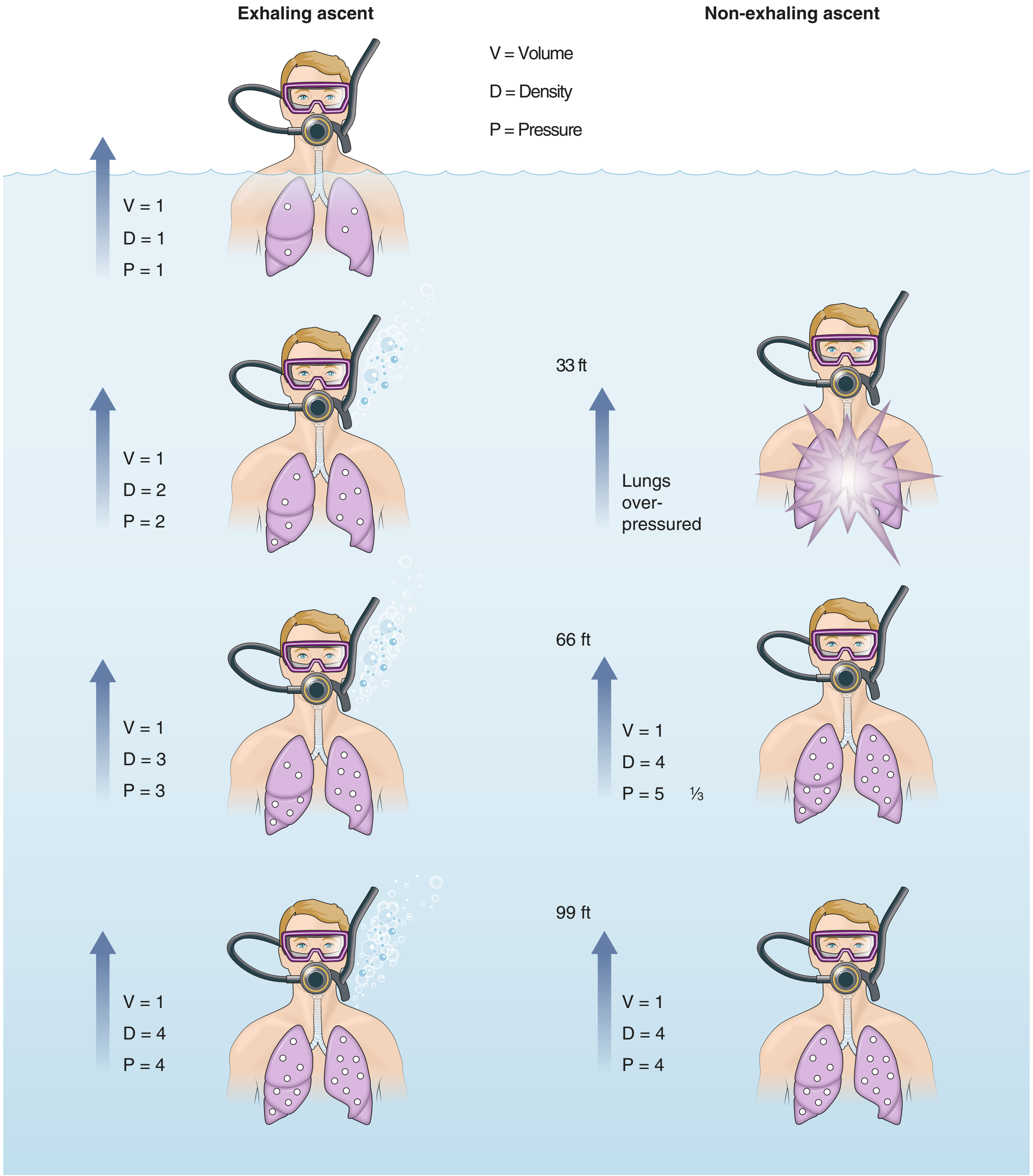
Exhaling ascent
Non-exhaling ascent
V = Volume
D = Density
P = Pressure
V = 1
D = 1
P = 1
V = 1
D = 2
P = 2
V = 1
D = 3
P = 3
V = 1
D = 4
P = 4
33 ft
Lungs over-pressured
66 ft
V = 1
D = 4
P = 5 ⅓
99 ft
V = 1
D = 4
P = 4

Figure 3.5 A breath-holding ascent after breathing compressed air can cause lung rupture. Breathe continuously while ascending.

you can overbreathe your equipment and experience air starvation, a suffocating feeling of not being able to get enough air.

While diving, you should make your breaths longer and slower than your breaths on land. Pace your activity to keep respiration at a slow, controlled rate. If breathing becomes rapid or labored, cease all activity immediately and focus on exhaling deeply until your respiration returns to a controlled rate.

Contaminated Air

If the compressed air in your tank comes from an air compressor that is not operated or maintained properly, your tank can contain contaminated air. The contamination is likely to be carbon monoxide (CO), a gas produced by incomplete combustion. CO in the body impedes the blood's ability to transport oxygen. Blood hemoglobin's affinity for CO is 210 times greater than its affinity for oxygen. Hemoglobin normally exchanges oxygen for CO_2 about every half minute, but when CO attaches to hemoglobin, the contaminant remains bonded to the hemoglobin for hours. Breathing contaminated air under pressure can make you ill and render you unconscious. Avoid contaminated air by having your tank filled only with pure air. CO is an odorless, tasteless gas, but it is usually accompanied by other gases that have a foul taste and odor. If the air from a scuba tank smells or tastes foul, do not use it. Report the situation immediately to the facility that filled the cylinder. Good air stations have their air tested regularly to ensure purity.

Water Inspiration

If you inhale water, your larynx spasms as a reflex action to keep the liquid out of your lungs. Avoid coughing and choking in water; not only are these reflexes unpleasant, but they can cause you to inspire more water and lose buoyancy. You need to breathe differently underwater than you do on land. Avoid quick inhalations. Begin a breath with a light, slow inhalation to ensure that you are inhaling air and not water. Once air begins flowing, the remainder of your inhalation can be normal.

You can also prevent droplets of water from going down your throat by placing the tip of your tongue on the roof of your mouth behind your upper teeth to act as a splashboard. If, in spite of all you do to avoid it, you inhale some water and start coughing, try swallowing hard three times in rapid succession. The swallowing helps you overcome your coughing reflex.

If water does enter your lungs, it interferes with respiration. If an extensive area of the lungs is irritated by water, you can drown. This is a risk you face in any aquatic pastime. The purpose of your training is to allow you to enjoy the aquatic environment with the minimum risk of injury, especially of drowning. When you follow the rules and practices that you are taught, the risk of injury while diving is negligible.

An automatic response to the inhalation of water is to swallow it in an effort to keep your airways dry. When you swallow saltwater, you can experience illness, nausea, or diarrhea. If you swallow several mouthfuls of saltwater while diving, you should terminate the dive.

Running Out of Air

Some divers do not monitor their air supplies and run out of air at depth. Running out of air underwater is no more excusable than running out of gas on the freeway when you have a working gauge. There are ways to manage an out-of-air situation (which you will learn in chapter 6), but it is much better to avoid the problem by monitoring your instruments and keeping them in working order.

Potential Lung Injuries

Pulmonary Barotrauma

Barotrauma is trauma or injury caused by pressure. **Pulmonary barotrauma** is any lung injury caused by pressure. Failure to allow expanding air to escape from the lungs during ascent can cause several forms of pulmonary barotrauma, either singularly or in combination.

Arterial Gas Embolism

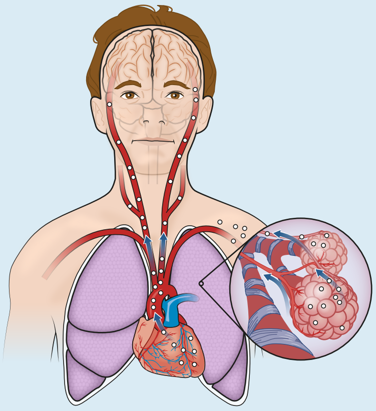

Air embolism

An **embolism** is a blockage of circulation. An embolism resulting from an air bubble blocking the arterial circulation is an **arterial gas embolism** (AGE). This occurs when air expanding in the lungs forces bubbles of air into the circulation. Air bubbles enter the capillary beds of the lungs and pass through the heart, which pumps the bubbles into arteries supplying blood to the body. The diameter of an artery decreases as the distance from the heart increases. At some point, a bubble lodges in an artery and becomes an embolus (plug). It is common for an arterial gas embolism to occur in an artery leading to the brain. The embolus has the effect of a stroke, causes unconsciousness, and is an extremely serious injury. Anytime a diver loses consciousness after a dive, you should suspect AGE. The temporary obstruction of an airway, such as that caused by a cold, increases the risk of AGE. Healthy lungs are a prerequisite for diving.

Mediastinal Emphysema

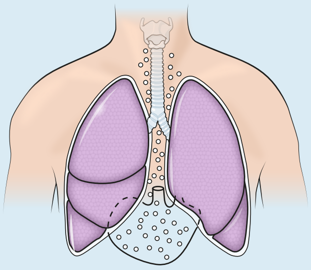

Mediastinal emphysema

If a rupture of the lung does not force air into the circulation, the air may travel along the bronchi and enter the middle area of the chest (called the mediastinum). This results in a **mediastinal emphysema**, which means that air is in the tissues in the middle of the chest. The injury causes a dull ache or tightness that worsens with coughing, swallowing, or taking a deep breath. Expanding air can interfere with the circulation of the heart and cause loss of consciousness.

Heart and Lungs

Immersion changes your cardiac workload. When you are upright in water, there is greater pressure on your lower limbs than on the upper part of your body. The pressure differential shifts more blood than normal into the upper part of your body. Your heart pumps a greater volume with each beat and works from 1.3 to over 1.5 times harder in this situation than it does on land.

Water temperature alters your heart's rate and rhythm. In addition, physical and emotional stress make your heart work harder. The combined effects of water pressure, exercise, cold, and stress can cause serious problems if your heart is not healthy. If you have heart problems and wish to dive, you should first obtain medical approval from a diving physician. Even minor heart problems can cause you to suddenly lose consciousness in the water, and you could drown. Fitness for diving is an important safety issue.

Subcutaneous Emphysema

Expanding air in the mediastinum can migrate upward along the breastbone. The air then swells the tissues around the neck, producing an injury known as **subcutaneous emphysema**, which means that air is in the tissues under the skin. The injury can cause changes in the voice, crackling of the skin, and a feeling of fullness in the neck.

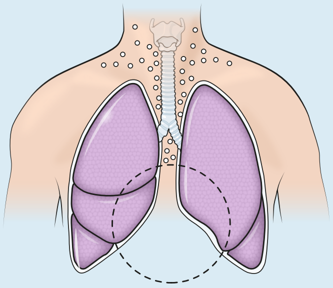

Subcutaneous emphysema

Pneumothorax

If a lung rupture forces air into the space between the lungs and the lining of the chest wall, a condition known as **pneumothorax** occurs. This means that air is trapped in the chest cavity. As the air trapped in the pleural space expands during ascent, it collapses the lung and can affect the action of the heart. Symptoms include severe pain and breathing difficulty.

Lung injuries are serious and can be life threatening. Life support may be required. This is one reason that you should complete a course in cardiopulmonary (heart and lung) resuscitation (CPR). You may have to administer first aid until professional medical treatment (which all lung injuries require) is available.

Failing to allow excess air to escape causes nearly all lung overexpansion injuries. You can prevent injuries by breathing continuously. If you remove the scuba regulator from your mouth for any reason, exhale lightly and continuously to avoid breath holding.

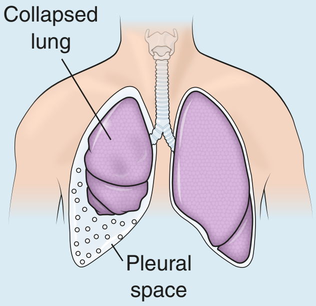

Pneumothorax

Cold water increases the effort required by your heart because cold causes the body to reduce circulation to the extremities in order to conserve heat. With more blood in the core of the body and an increased workload from pressure, your heart may not be able to pump the increased blood volume. The result can be excess blood in the small blood vessels of the lungs. These small blood vessels will transfer fluid to the air spaces of the lungs to reduce pressure. The accumulation of fluid in the lungs—known as pulmonary edema—causes severe breathing distress and can lead to panic or even loss of consciousness. You should test your circulatory ability gradually. Avoid strenuous dives in cold water until you have gradually adjusted to being underwater.

Your ability to exert effort in water is not as good as it is on land. When under pressure, your heart and lungs do not function as well as they normally do. A situation in which you are trying to work hard underwater can quickly result in a frightening sensation of suffocation when your equipment is unable to deliver the amount of air that your body requires. To stay within your cardiopulmonary limitations, you must learn to limit your activities and to pace yourself.

Equilibrium Adjustments and Seasickness

If you rupture an eardrum, water that is colder than body temperature can come into contact with the semicircular canals of your inner ear. The canals provide information for your brain to maintain equilibrium and are sensitive to temperature and pressure changes. Cold water cooling the semicircular canals can cause vertigo, a subjective feeling of movement perceived as a spinning sensation. This disorientation will pass when the water in the middle ear warms to body temperature. Obviously, it is better to prevent vertigo than to deal with it.

When a sudden change in pressure occurs in the air space of the middle ear of one ear (a change that affects the semicircular canals), this can also produce a disorienting sensation called alternobaric vertigo. The disorientation from alternobaric vertigo passes quickly when the brain recovers from the sudden pressure change.

Vertigo has many causes. Fortunately, instances of severe disorientation are rare in diving. If you experience disorientation while underwater, try to grasp a solid object for a point of reference until the feeling passes. If suspended in the water, close your eyes and hug yourself to reduce the effects of the vertigo. Avoid panic by telling yourself that the sensation will not last long.

Motions detected in your inner ears, visual references, and pressures on the joints of your limbs all affect your equilibrium. When your brain receives mixed signals from your inner ears, eyes, and body, you can experience motion sickness. You must avoid seasickness—motion sickness experienced aboard a vessel—because vomiting in or under the water is hazardous.

Medications can help reduce the tendency to be seasick. The medicine dulls the senses of the organs of balance in your ears. Unfortunately, the medications can have other undesirable effects. If you are prone to seasickness, consult a diving physician about the type of medication you should try. Take some of the medicine several days before you plan to dive, and note the effects, if any. If the medication produces drowsiness or blurred vision, do not use it while diving. Seek an alternative that does not produce side effects. Medications that cause dizziness, drowsiness, changes in heart rhythm, or blurred vision can cause you to lose consciousness under pressure. Many divers do successfully use medications to prevent seasickness; you need to find a type that works for you. Take medication for motion sickness at least 30 minutes before you are exposed to motion.

If you do not want to use medication, you may use other techniques to reduce the likelihood of seasickness. Eat a good, non-spicy meal before diving. An empty, acid-filled stomach becomes upset more easily than a full one. People whose breakfast before diving consists of coffee and orange juice are good candidates for motion sickness. When aboard a vessel, you should position yourself as near the center of the boat as possible. Avoid the front end of the boat. You should also avoid breathing engine fumes and avoid reading. You can reduce your reaction to the boat's motion by lying down for a while with your eyes closed. Being still allows your inner ears to adjust to the motion without visual signals confusing your brain.

You can become disoriented when you are weightless in a dimly lit environment. Under some conditions, it can be difficult to tell which way is up if you rely only on your sense of balance. To prevent disorientation, you must learn to recognize clues about your orientation in the water. Water in your mask settles to the lowest point, bubbles ascend, and heavy objects you hold (such as your weight belt) orient themselves up and down.

Visual Adjustments

Experience will help you adjust to magnified vision underwater. You can adjust so quickly to distance corrections in water that you have to readjust when you surface from a dive. At the end of a dive, the distance to a boat or to the shore can look much greater than it is. You may be surprised to find that you require less time than you think to swim to a destination.

Your vision also compensates for color differences. If you know an object's true color, the object will look more like that color. The use of artificial light at close range makes it easier to view the rich and magnificent colors in the underwater world.

Your vision adjusts to low light levels, but the process takes time. Short, deep dives in turbid water do not allow complete visual adjustment, and details will not be clear. You cannot see well without artificial light. You can improve your ability to see while diving by avoiding bright light and glare before a dive. Wear good, dark sunglasses above water during the day. When you complete training for night diving, you will learn other techniques to help your eyes adjust for diving at night.

Humidity can cause condensation to form on the lens of your mask and obscure your vision. A clean glass surface will not fog, so make sure you clean the lens or lenses of your mask thoroughly before diving.

Ingassing and Outgassing

Your body absorbs nitrogen underwater, and there are limits to the amount of nitrogen that you can safely eliminate at the end of a dive. If you exceed these limits, you can be injured. In this section, you will learn the theory of decompression, which will help you understand how dive tables and dive computers function to help you avoid injuries.

Decompression Theory

Gases diffuse by moving from areas of greater concentration to areas of lesser concentration. When external pressure increases, gases diffuse from your lungs into your blood, and then from your blood into your tissues. When the **ambient (surrounding) pressure** decreases, diffusion occurs in the reverse sequence.

Two factors that affect diffusion of gases in your body are time and **perfusion** (the circulation in a tissue). The greater the circulation in a tissue, the sooner the pressures of the gases in that tissue come into balance with the pressures of the gases you breathe. Reaching this state of equilibrium takes time. The amount of time it takes a tissue to accumulate half of the gas it can hold at a given pressure is a **half-time**. A tissue is saturated (holds all the gas it can at a given pressure) after six half-times. If perfusion permits 50 percent of a gas to diffuse into a tissue in 5 minutes, the tissue saturates in 30 minutes. Outgassing also occurs in six half-times.

Air is primarily nitrogen and oxygen. The oxygen in the air you breathe is of no concern during decompression within the limits of recreational diving (130 feet or 39.6 m maximum) because you use the oxygen. Nitrogen, the primary component of air, is inert. Your body cannot use the nitrogen, so when you ascend, you must eliminate the excess nitrogen that you absorbed at depth. Because nitrogen diffuses from your body, there is no problem unless the reduction in pressure is so great that you cannot eliminate the nitrogen fast enough. When you reduce the pressure on a liquid rapidly and there is sufficient gas dissolved in the liquid, the gas forms bubbles. An excellent example of this is CO2 dissolved in carbonated beverages (see figure 3.6). The gas remains dissolved in a sealed, pressurized container. When you reduce the pressure suddenly by opening the container, the gas forms bubbles because it cannot diffuse out of solution slowly. If a beverage container has a tiny leak, however, the carbon dioxide comes out of solution slowly without bubbling. There is no bubbling when you open a container that has a slow leak because the gas has already diffused out of solution.

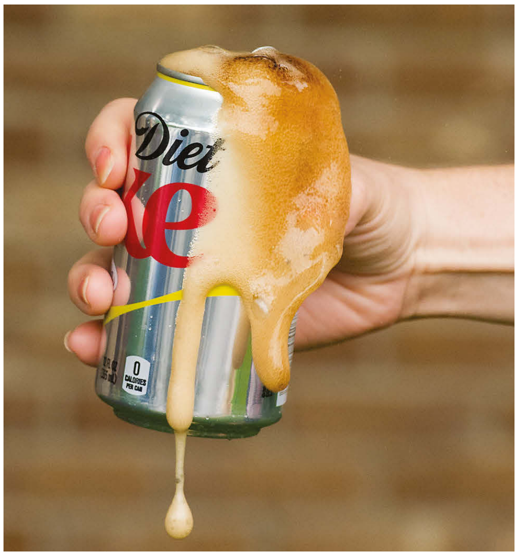

Figure 3.6 Rapid pressure reduction causes dissolved gases to form bubbles within the decompressed liquid.

You must be concerned about the amount of nitrogen in solution in your body and the rate at which you eliminate it. If you absorb too much nitrogen while diving and do not ascend in a manner that allows the excess nitrogen to be eliminated without forming bubbles, the bubbles are likely to cause decompression sickness, a serious diving illness.

Mathematical models provide estimates of the amount of nitrogen in different parts of your body. Because circulation varies in different parts of your body, tissues absorb nitrogen at differing rates. Decompression experts use mathematical models, called **compartments**, to estimate gas absorption and elimination by various areas of the body. A compartment is identified by its half-time. A 5-minute compartment has a half-time of 5 minutes. Experts use compartments ranging from 5 minutes to as long as 960 minutes when calculating gas absorption and elimination.

A compartment that has absorbed gas can withstand some lowering of pressure before bubbling occurs. Originally, scientists believed that a reduction in pressure by a ratio greater than two to one would cause bubbling to occur in divers whose tissues were saturated at a particular depth. This ratio was considered a surfacing limit that could not be exceeded. Then scientists discovered that the surfacing ratios are different for various compartments because of differences in perfusion. For example, brain tissue absorbs and eliminates inert gas faster than bone tissue. Body areas that eliminate gas quickly have a higher surfacing ratio than compartments that eliminate gas slowly. The difference in ratios poses an interesting and complex problem for divers because one tissue controls how long a diver can remain at one depth, while another tissue controls how long the diver can remain at another depth. The **controlling compartment** was used to establish our modern time limits for diving. Do not remain at any depth longer than the time it takes the controlling compartment to exceed its surfacing ratio. If you do, you must prevent bubble formation by stopping during the ascent to allow time to eliminate the excess gas. The time limits for various depths have been conveniently arranged into tables and included in dive computer software. These tools help divers keep the amount of nitrogen in solution in the various tissues of the body within the surfacing ratio of each compartment.

It takes time to eliminate excess gas from tissues. The rate at which you reduce pressure (ascend) is important. The dive tables presented in chapter 7 use an ascent rate, or pressure reduction rate, no greater than 30 feet (9 m) per minute. The mathematical models for some dive computers use rates two or three times slower than the dive tables. You must ascend slowly to prevent the formation of bubbles in your body tissues.

Partial ascents with stops before continuing the ascent can help with outgassing and the prevention of bubble formation. These delays are called safety stops or rest stops, and they are recommended when ascending from any depth in excess of 30 feet. The deeper you dive, the more safety stops you should take when ascending. The recommended practice is to stop for five minutes at half of the deepest depth to which you have been diving and to stop again at a depth of 15 to 20 feet (4.5 to 6.1 m) for five minutes before surfacing.

Decompression Illness

Commonly known as the bends, decompression illness (DCI) is the result of a reduction in pressure (decompression) that is too rapid for the amount of gas in solution in body tissues. The gas forms bubbles in the tissues or in the blood before it can diffuse into the lungs and be eliminated. Scientists do not fully understand what occurs when DCI strikes. The symptoms may appear immediately after surfacing from a dive or days after diving. About half of all DCI cases occur within one hour after diving. The symptoms vary depending on the amount and location of the bubbles. Severe cases of decompression illness can have severe neurological effects and produce permanent paralysis. Common symptoms include the following:

- A mottled skin rash
- Joint pain
- Numbness
- Tingling
- Weakness
- Paralysis

The following factors can increase the chances of DCI:

- Lack of sleep
- Tight-fitting apparel

- Alcohol and its aftereffects
- Exercise during and after diving
- Dehydration
- Cramped position
- Illness
- Altitude after diving
- Age
- Scar tissue or injuries
- Cold water
- Medical conditions

Some experts believe that you should avoid postdive activities that stimulate circulation because increased perfusion may contribute to DCI. Specific activities to avoid include physical exercise, drinking alcoholic beverages, and taking hot showers or baths. Reduced pressure at altitude, such as flying in a plane or driving into the mountains too soon after diving, also poses an additional risk factor that can cause DCI. Avoid all activities that increase the likelihood of decompression illness.

A person suffering from DCI requires prompt first aid and medical treatment because it is an extremely serious condition. The illness worsens with time. The best first aid measure is to administer oxygen in the highest concentration possible. Breathing oxygen eliminates nitrogen in inspired air and enhances the diffusion of nitrogen from the body. The patient should remain still and should sip water. Make arrangements to have the patient transported to the nearest medical facility. After initial medical treatment, the medical staff will arrange to transfer the patient to a facility that has a hyperbaric (high-pressure) chamber. The patient is placed inside a large vessel called a recompression chamber (see figure 3.7). Hyperbaric treatment consists of increasing the pressure in the chamber to reduce or eliminate the symptoms, administering medications, and then slowly decompressing the patient. Recompression must be done in a chamber.

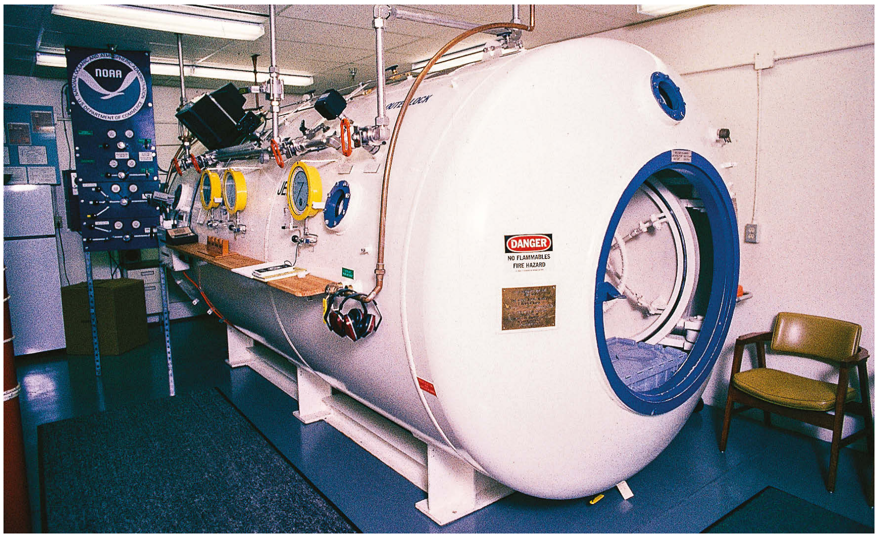

Figure 3.7 Recompression chambers are the only safe method to restore pressure after becoming sick.

Decompression illness is extremely serious because it can cause permanent injury. DCI can cause joint pain and neurological deficits that cannot be cured. To reduce the likelihood of DCI, you should remain well within the established time and depth limits for diving. Surface at a rate no faster than that specified by the dive computer you use. Stop at least twice during ascent (as previously recommended) and eliminate excess nitrogen before surfacing (see figure 3.8). Ascending at the proper rate and stopping are forms of decompression that reduce the likelihood of DCI. (The details about planning your ascent and time delays are provided in chapter 7.) An additional precaution that can help you avoid DCI is to delay excursions to altitude after diving.

Another point to keep in mind is that it is often difficult to distinguish between the symptoms of an arterial gas embolism and DCI. Any diver who has neurological symptoms after a dive should seek medical attention immediately.

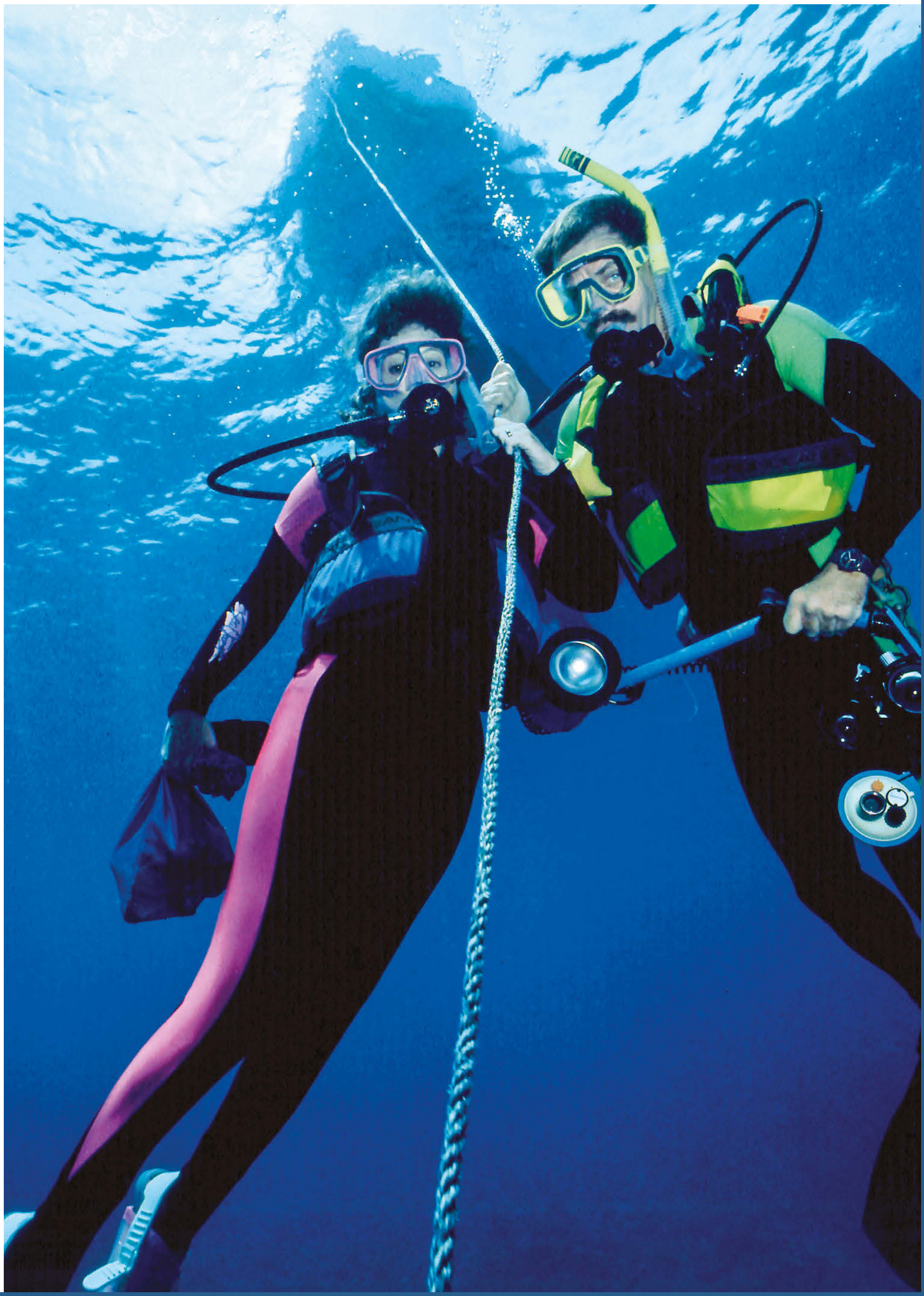

Figure 3.8 Divers completing a decompression stop on their ascent to the surface.

Nitrogen Narcosis

The increased partial pressure of nitrogen can cause a condition known as nitrogen narcosis, or rapture of the deep, at depths of about 100 feet (30 m) and deeper. Scientists do not know the exact mechanism of nitrogen narcosis, but its effects are similar to those of anesthetic gases. The feelings associated with narcosis range from euphoria to overconfidence to terror, although some divers do not experience any unusual feelings. Narcosis affects you whether or not you experience abnormal feelings. And it impairs thinking, judgment, reasoning, memory, and the ability to do physical tasks. Narcosis is hazardous; it reduces your awareness and your ability to respond to an emergency. Susceptibility to narcosis varies from person to person and within an individual from day to day.

Narcosis begins suddenly at depth—usually around 100 feet (30m). You can relieve its symptoms rapidly by ascending to a shallower depth to reduce the narcotic effect.

The following factors predispose a person to narcosis:

- A high level of systemic CO2 caused by exertion
- Alcohol or its aftereffects
- Anxiety
- Cold
- Medications
- Social drugs

Experience, frequent diving, and concentration will reduce susceptibility to narcosis. Preventing narcosis is always better than attempting to cope with it.

Specialty training can qualify you to dive with special gas mixtures that reduce the effects of nitrogen. Diving with mixed gases without completing the required training is dangerous and should never be attempted.

Dehydration Prevention

You need to preserve body fluids to prevent **dehydration**. Getting cold makes you produce more urine than normal. Each breath you take from your scuba cylinder must be humidified as explained in the previous chapter. Breathing dehumidified air causes dehydration. You receive air underwater by creating an inhalation pressure that opens valves in your air delivery system. The inhalation pressure is slight, but it is greater than normal. Inhaling harder than normal is negative-pressure breathing, which also has the physiological effect of increasing urine production. Breathing underwater compounds the problem of dehydration.

Diuretic beverages (such as coffee and alcohol) and some types of medications cause increased urine production. Avoid ingesting anything that makes you urinate more than normal. You must prevent excessive dehydration because the condition predisposes you to diving injuries. To prevent dehydration, you should do the following:

- Insulate yourself to stay as warm as possible.
- Keep your regulator well maintained so you can breathe as easily as possible.
- Avoid diuretic drinks and medications.

- Replenish body fluids frequently; drink fluids before and between dives.

  Note: If you wait until you are thirsty to drink when diving, you are already dehydrated. Be proactive and drink at least 1 pint (500 ml) of appropriate fluid after every dive. A good after-dive beverage is a sports drink diluted with an equal amount of water. Avoid alcoholic beverages because they increase dehydration.

Mobility Adjustments

The equipment that you wear for diving limits your mobility. It reduces your range of motion and makes it difficult to walk. The colder the water, the thicker your exposure suit, so the less your range of movement.

Diving equipment is fairly heavy, and it is challenging to lift and move. Improper lifting techniques can cause back injuries. Squat down and lift with your legs to pick up a tank or weight belt instead of bending over. The weight of the equipment changes your center of gravity and affects your balance. Keeping your balance during entries and exits can be challenging on a rocking boat or an uneven surface for your feet. Move carefully and hold on to something or someone for support when you move around out of the water.

You wear fins so that you can use the large muscles of your legs for propulsion. When diving, you need to learn to use your legs for swimming and your body angle to control direction. These techniques free your hands for other uses underwater. Minimize the use of your hands for propulsion. Fins make it difficult for you to walk when you're out of the water. Shuffle your feet while moving backward or sideways, keep your knees bent, and be careful not to fall.

Drag slows your movement in the water. One factor affecting drag is the speed of motion. The greater the speed, the greater the resistance to movement. The average diver can sustain a speed of a little more than 1 mile an hour (0.9 knot, or 1.6 km per hour). A knot equals 1.15 miles per hour (1.85 km per hour) and is the method of measuring speed in water. Doubling the speed increases the energy requirement fourfold. Trying to move in water as if it were air causes exhaustion quickly. Use a slow, steady pace and slow, deliberate motions to reduce the effects of drag.

Another factor affecting drag is the size of the object in motion. The larger the surface area exposed, the greater the resistance. A swimming diver is under the influence of four forces: Weight (negative buoyancy) pulls the diver downward, lift (positive buoyancy) pulls the diver upward, thrust moves the diver forward, and drag slows forward progress (see figure 3.9). Divers usually wear weight around the waist, which pulls the lower half of the body downward. Buoyancy from the BC lifts the upper half of the body. The effect of these forces increases the surface area of the diver and therefore the drag. You should adjust the amount of weight you wear so that your body is as horizontal as possible in the water. Correct weighting minimizes drag and the effort required to swim. Also consider placing a small amount of weight near the top of your body to adjust your trim—that is, your position in the water.

Water flows smoothly across a smooth and rounded surface but flows turbulently across an irregular surface. When the surface is irregular, the turbulent flow increases drag. Just as vehicles designed to travel through fluids are streamlined to reduce the drag caused by turbulent flow, you can choose and configure your equipment so that it presents the smoothest surface possible to the flow of water.

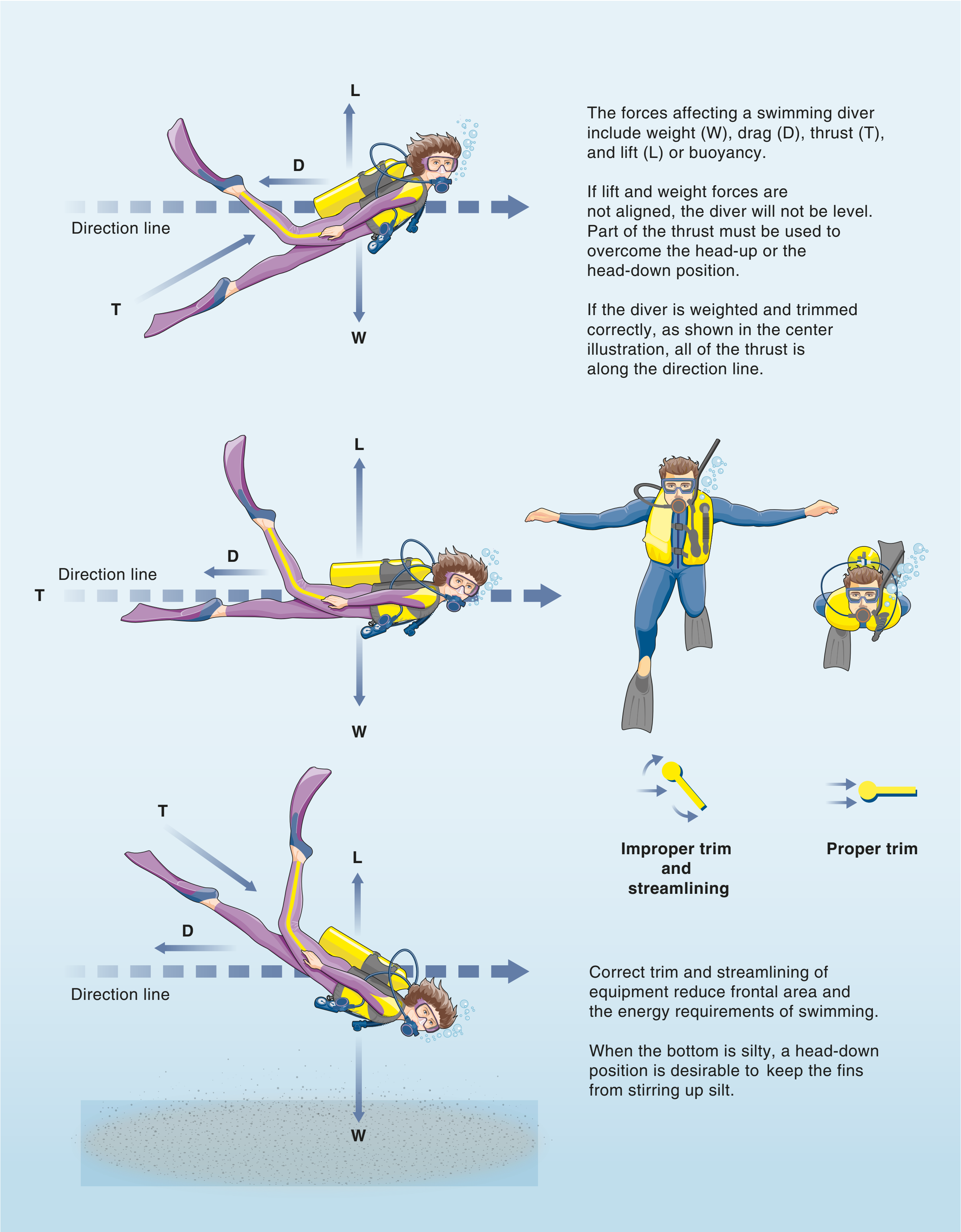
L
D
Direction line
T
W
L
D
Direction line
T
W
T
L
D
Direction line
W
Improper trim and streamlining
Proper trim

The forces affecting a swimming diver include weight (W), drag (D), thrust (T), and lift (L) or buoyancy.

If lift and weight forces are not aligned, the diver will not be level. Part of the thrust must be used to overcome the head-up or the head-down position.

If the diver is weighted and trimmed correctly, as shown in the center illustration, all of the thrust is along the direction line.

Correct trim and streamlining of equipment reduce frontal area and the energy requirements of swimming.

When the bottom is silty, a head-down position is desirable to keep the fins from stirring up silt.

Figure 3.9 Weight, drag, thrust, and lift (buoyancy) affect divers as they move through the water.

Summary

You have to make many changes in your normal behavior to adjust to the underwater environment. The way you breathe is the most important adaptation. You must breathe continuously and avoid breath holding when you use scuba equipment. You must also limit your activity and pace yourself to avoid overexertion. Managing the mechanical and physiological effects of pressure requires major adjustments. You must keep pressure equalized in your air spaces to avoid pressure injuries, and you must limit your depth and time at depth to avoid nitrogen narcosis and decompression illness. Being in the underwater environment seems strange at first, but you can conform to weightlessness and other strange feelings. The sensations of diving become exhilarating as you gain experience in the new world beneath the surface.

Application-of-Knowledge (AOK) Questions

Remember that this is an exercise in thinking, not a quiz. If you carefully consider each situation and how you would respond to it, your time will be well invested even if your thoughts are different from the solution provided. Imagine each situation as vividly as possible.

1. You are wearing a wetsuit and swimming against a current (which is a good idea under most circumstances) in cold water. You are becoming colder with each passing minute. What action, other than ending the dive, can you take to reduce your heat loss?
2. After a dive during which you shivered, you exit the water on a cool, cloudy, windy day. What actions can you take to get warm that are not likely to cause physiological problems? Hint: Rapid rewarming can affect your heart rhythm and may also cause decompression illness.
3. You are wearing 17 pounds of lead weights while diving at a depth of 40 feet (12 m) when your weight belt suddenly slips off. How rapidly do you think you will float upward, and what actions can you take to recover from this situation?
4. You are wearing a wetsuit, a BC, and a weight belt while scuba diving at a depth of 33 feet (10 m). Your buoyancy is adjusted to be perfectly neutral at this depth. You have to ascend 10 feet (3.3 m) to clear a reef. What actions will you need to take to control your buoyancy as you swim over the reef?
5. You are pulling yourself hand over hand down an anchor line to descend to a dive site and are having difficulty equalizing the pressure in your ears. Your normal equalization methods are not working. What should you do to equalize the pressure in your ears?
6. Why do you need to equalize the pressure in your body air spaces and your mask more frequently at the beginning of the descent than when you are approaching your diving depth?
7. You are having trouble keeping up with your buddy, who is swimming ahead. You are breathing heavily and beginning to feel starved for air. What actions should you take?
8. When you switched tanks between dives, you accidently selected one that is almost empty. You were also distracted and failed to check the pressure before you started the dive. You are at a depth of 50 feet (15 m) when it becomes hard to breathe. You look at your pressure gauge and find a reading of zero. Based on what you know at this point in your training, what action do you think is appropriate?

Large, beautiful angelfish at a coral reef in Cozumel, Mexico

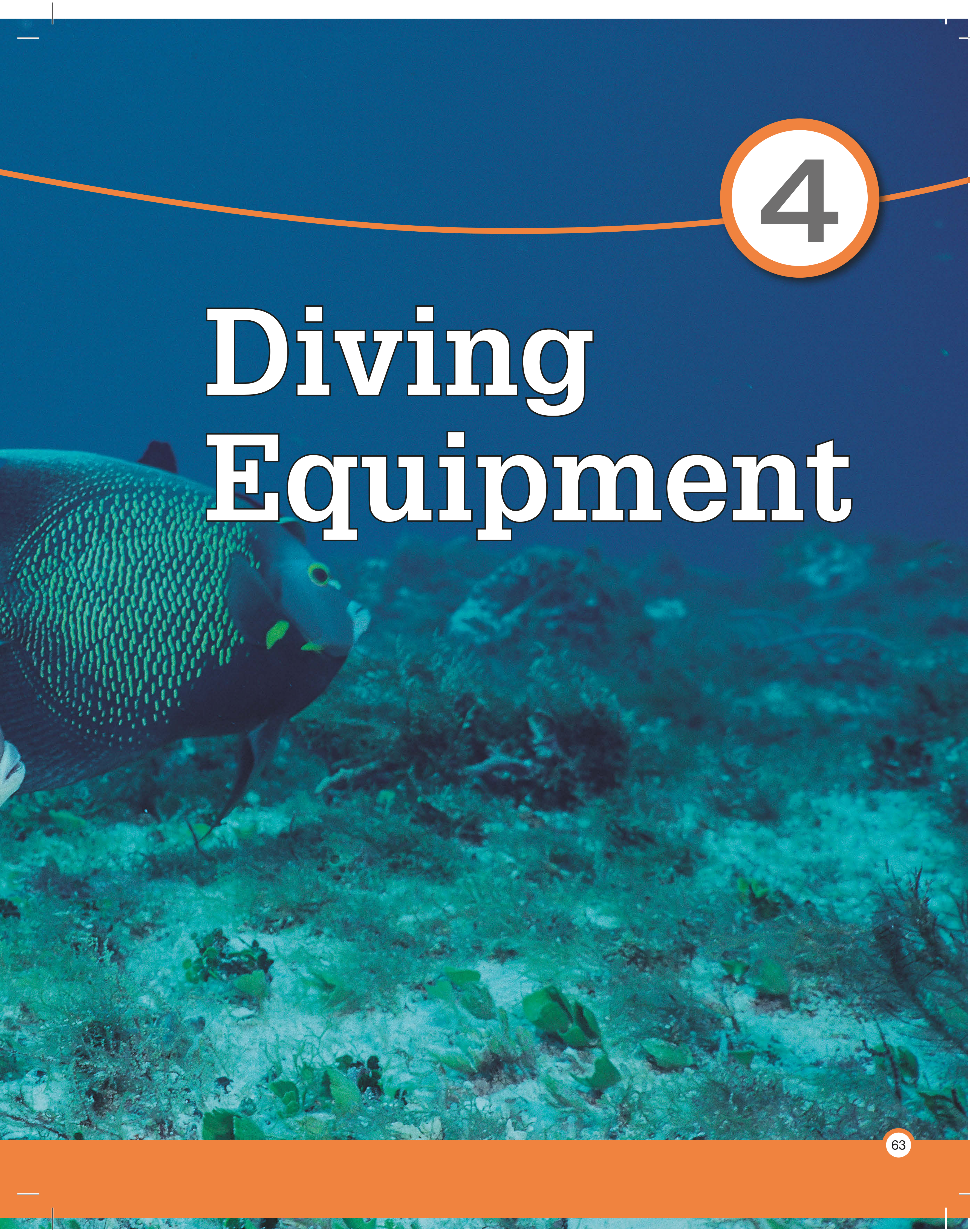

4

Diving Equipment

Dive In and Discover

By the end of this chapter, you will be able to do the following:

- Describe the purpose, types, features, selection criteria, and care and maintenance of skin and scuba diving equipment.
- Contrast steel and aluminum scuba cylinders, wet and dry suits, integrated and nonintegrated weighting systems, and reserve and nonreserve valves.
- Define the terms *buoyancy*, *compensator*, *defogging*, *shorty*, *jumpsuit*, *Farmer Johns*, *K-valve*, *J-valve*, *DIN valve*, *valve seat*, *burst disk*, *O-ring*, *port*, *stage*, *octopus*, *capillary gauge*, *Bourdon tube*, *Boyle's law*, *outgassing*, *bezel*, *submersible pressure gauge (SPG)*, *blowout plug*, *alternate air source (AAS)*, *pony tank*, *Spare Air unit*, *lubber line*, *console*, and *ceiling*.

You will become familiar with the following items in this chapter:

- Masks
- Snorkels
- Fins
- Skin diving vests
- Exposure suits
- Weighting systems
- Buoyancy compensators
- Scuba cylinders
- Cylinder valves
- Scuba regulators
- Alternate air sources
- Instrumentation
- Dive knives and other accessories
- Specialized equipment

You need to be equipped properly to dive in open water. A snorkeler (a diver who remains at the surface) should wear a mask, a snorkel, fins, and a skin diving vest. A skin diver (a breath-holding diver who dives beneath the surface) uses snorkeling equipment and may wear an exposure suit and a weight belt (if the suit requires weights). Scuba equipment includes, at a minimum, a mask, a snorkel, fins, an exposure suit, a weighting system (if needed), a **buoyancy compensator** (BC), a scuba unit (cylinder, valve, regulator, **alternate air source**), instrumentation, and a dive knife. For cold water, a diver also needs a hood, boots, and gloves. Figure 4.1 shows fully equipped divers.

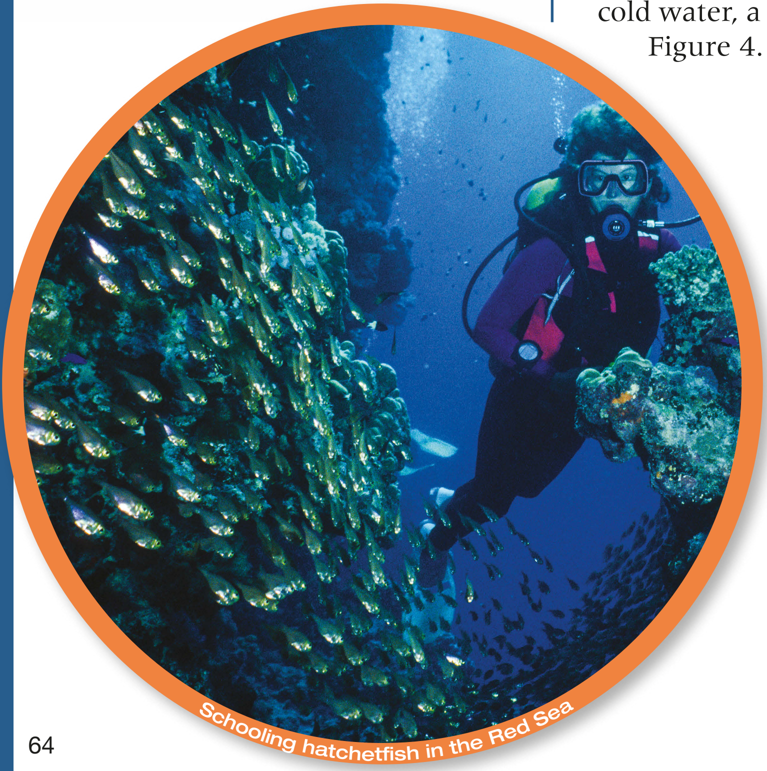
Schooling hatchetfish in the Red Sea

Equipment helps people adapt to the underwater environment. It allows them to see, breathe, move, monitor their position, remain comfortable, have fun, and rest. Diving is an equipment-intensive activity. In this chapter, you will learn what equipment you need, how to select the best equipment for your needs, and how to care for your equipment.

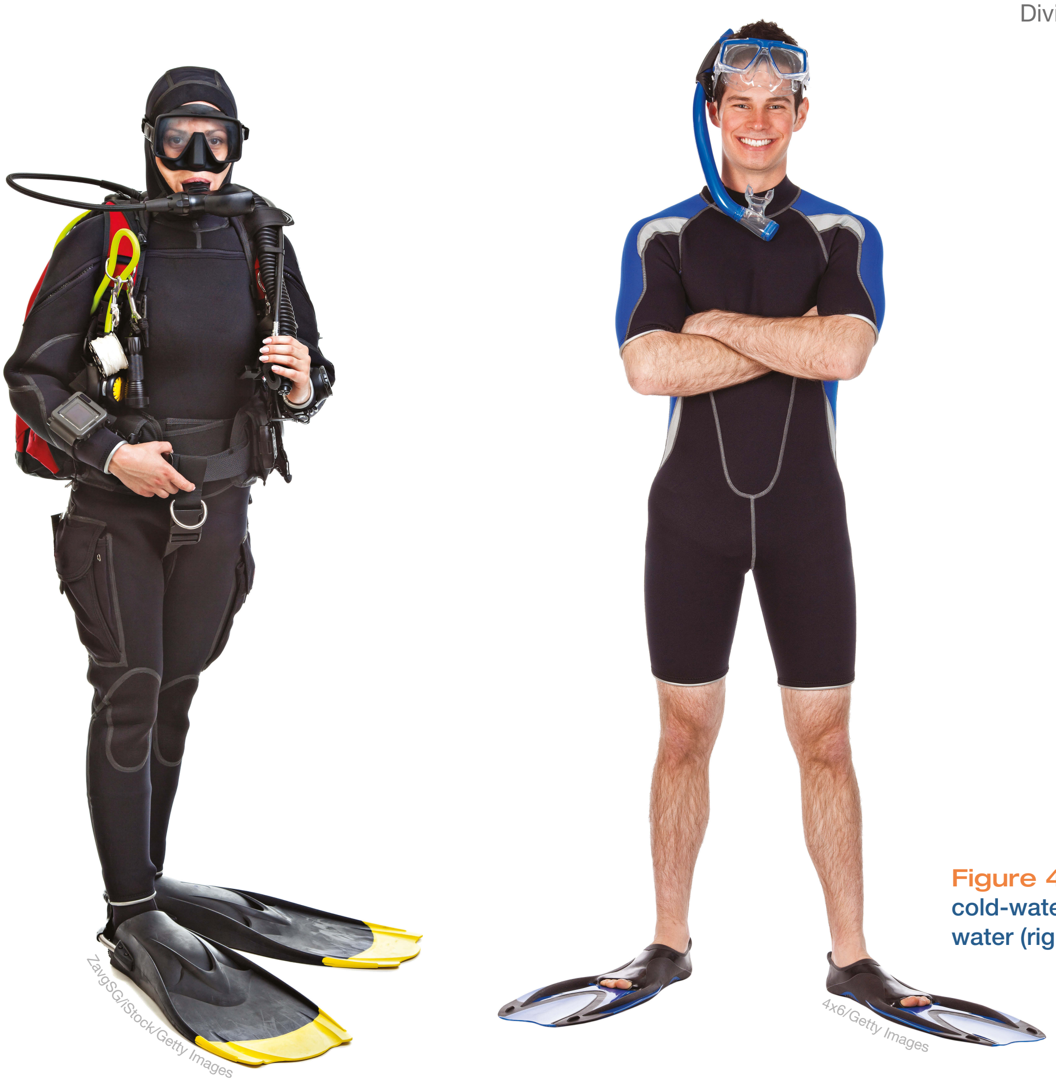

ZavgSG/iStock/Getty Images

4x6/Getty Images

Figure 4.1 Fully equipped cold-water (left) and warm-water (right) divers.

Basic Equipment for All Divers

The mask, snorkel, fins, and some flotation device are basic equipment for all types of recreational diving (snorkeling, skin diving, and scuba diving).

Masks

The eyes require an air space in front of them to focus sharply. Your mask provides an air space and a window to another world. Many styles of masks are available, but only two basic types are used for recreational diving: purge and nonpurge. A purge is a one-way valve through which you can expel water that enters the mask. As you will learn in chapter 6, you can remove water from a mask without a valve, so many masks do not feature a purge valve. A third type of mask, a full-face mask, is for commercial and specialty applications only. Figure 4.2 shows the most common types of masks.

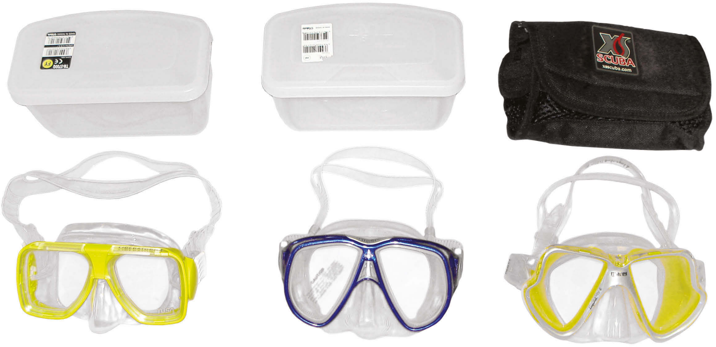

Figure 4.2 Test dive masks before you buy to ensure a snug and comfortable fit.

The type of mask you choose is not nearly as important as the fit of the mask on your face. The mask must fit your facial contours perfectly so it will remain comfortable and watertight throughout a dive. Fit and comfort are the most important features to consider when you select a mask. To test the fit of a mask, remove the strap or position it on the front of the mask. Tilt your head back and lay the mask (do not push it) on your face. Make sure your hair is not under the sealing edge of the mask; then inhale gently. If the mask pulls onto your face snugly from the partial vacuum created by your inhalation, the mask fits. If you have to push the mask to get it to seal on your face, the mask will probably leak when you use it underwater.

Factors that affect the fit and comfort of a mask include the style, the type of material, and the type of sealing edge of the mask. The best masks are made from silicone, which is soft, pliable, and nonallergenic and resists deterioration better than rubber compounds (see figure 4.3). The buoyancy of smaller masks poses no problems, but the tug of buoyancy of a larger-volume mask can affect the seal. Low-volume masks are easier to clear of water and provide excellent visibility. A wide, double-edged seal does a better job of keeping out water than a single-edge seal.

A film of oil from production is on the surface of the glass lenses of new masks. You must remove the film completely—a process called **defogging**—or the mask will fog continuously underwater. Clean your mask thoroughly with scouring powder. The glass (which is tempered) is too hard to be scratched by the abrasive, so do not be timid when cleaning your mask. Commercial defogging solutions help keep your mask clear while you dive. If a clean mask fogs slightly during use, you can allow a small amount of water into the mask and wash it across the fogged area to resolve the problem.

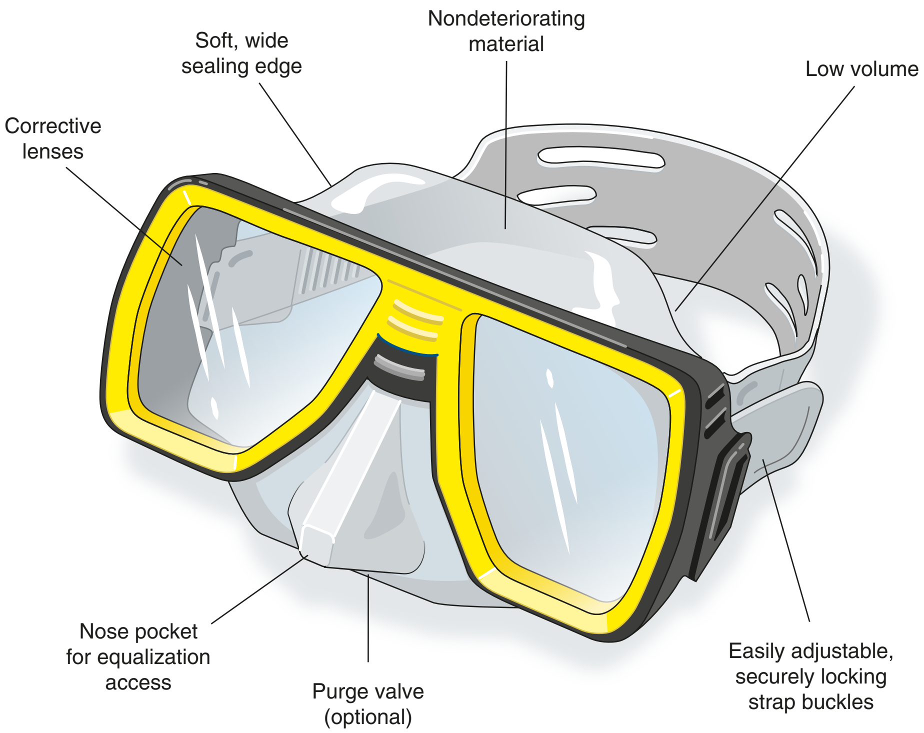
Soft, wide sealing edge
Nondeteriorating material
Low volume
Corrective lenses
Nose pocket for equalization access
Purge valve (optional)
Easily adjustable, securely locking strap buckles

Figure 4.3 Mask features.

Several options are available for visual correction if you need it. Hard contact lenses cannot be used, because gases that are released from the body during ascent can be trapped between the lens and the surface of the eye. This can injure the eye. If you wear soft contact lenses, you can use them for diving after you complete your training; however, you should not use them during training because you may lose them when you are learning how to clear water from your mask. Some contact lens wearers prefer a mask with a small purge valve because they can expel water from the mask with little risk of losing a contact lens. Several companies prepare and bond corrective lenses into any diving mask. Interchangeable corrective lenses are available for some masks. If you require only a simple correction, you may be able to obtain corrective lenses for your mask when you purchase it.

Store your mask in a mask box when you are not using it. The box helps keep the mask from getting damaged and helps prevent discoloration of the silicone mask skirt.

Snorkels

A human head weighs about as much as a bowling ball. If you had to swim while holding a bowling ball out of water, you would become exhausted quickly. If you swim while holding your head above water, you will tire rapidly. But if you allow buoyancy to support your head in the water, you can relax and swim for hours. A snorkel allows you to breathe while water supports your head—allowing you to conserve energy and enjoy continuous underwater viewing.

In its simplest form, a snorkel is nothing more than a breathing tube that extends from a diver's mouth to a point above the waterline. A basic diving snorkel is a J-shaped tube with a mouthpiece on one end. Just as with masks, there are two types of snorkels: purge and nonpurge (see figure 4.4). A purge snorkel has a one-way valve through which you can expel water that enters the tube. As you will learn in chapter 6, you can clear water from a snorkel that lacks a purge valve, so some snorkels do not have a valve. Some types of purge snorkels are self-draining; gravity drains water from the tube automatically when you are at the surface of the water.

Additional snorkel options include a swivel mouthpiece and a flexible hose for the lower half of the tube. With the flexible hose, the lower part of the tube—which is usually curved—hangs straight down when you are not using your snorkel (see figure 4.5). The flex-hose snorkel reduces interference between the snorkel and your scuba regulator. Special mouthpieces can be used to maximize comfort, and devices are available to prevent water from entering the top end of the snorkel, although water exclusion devices are not essential.

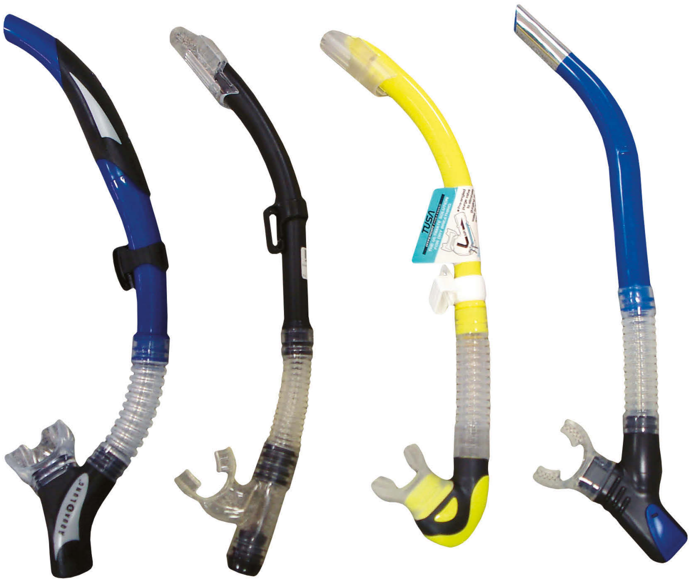

Figure 4.4 Three common types of snorkel are purge (two at left and far right) and nonpurge (center right).

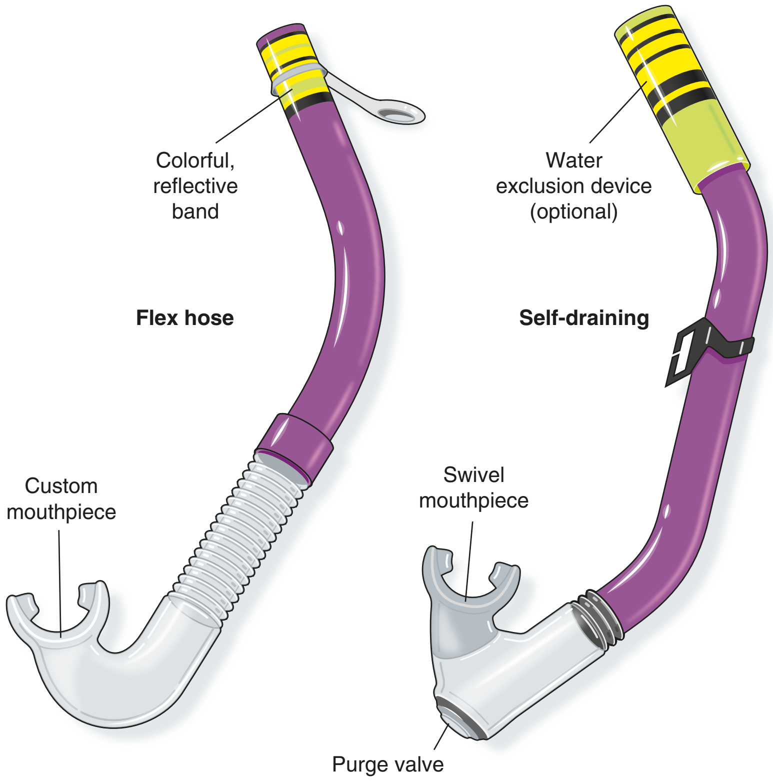
Colorful,
reflective
band
Water
exclusion device
(optional)
Flex hose
Self-draining
Custom
mouthpiece
Swivel
mouthpiece
Purge valve

Figure 4.5 Features of a flex hose and self-draining snorkel.

The fit of your snorkel is the most important consideration, much more important than the type. The mouthpiece must not irritate your mouth, gums, or jaw when in place for extended periods of time. The angle of the mouthpiece in your mouth should not require you to bite hard to hold the mouthpiece in place. A snorkel that fits poorly may cause sore gums or jaws. A snorkel tube should have an inside diameter of approximately 3/4 inch (1.9 cm) so that resistance to airflow through the tube does not make breathing difficult.

Attach the snorkel to your mask strap on the left side using a snorkel keeper because your regulator hose goes over your right shoulder. Several types of snorkel keepers are available. Popular keepers are the simple ones depicted in figure 4.6. The adjustment is correct if the snorkel mouthpiece remains in your mouth when you open your mouth wide.

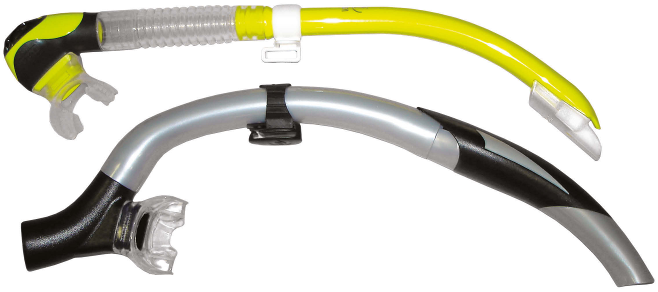

Figure 4.6 Two types of snorkel keepers.

Fins

Diving would be much less enjoyable without fins. Fins increase your ability to move in the water and free your hands for other activities. You move in water by pushing against the water. Fins increase your ability to move by presenting a larger surface area than your hands or feet alone. The muscles of your legs are much larger and stronger than your arm muscles. Your arms would tire quickly if fins were attached to your hands, but your legs are strong enough to handle the load. Fins also help stabilize you in water by providing a large surface to resist movement. The resistance provides leverage for countermovement and directional control.

The two basic types of fins are shoe (full-foot) fins and open-heel fins. Shoe fins slip on to bare feet, so they are good snorkeling fins for tropical climates. You wear open-heel fins with foot coverings called boots. Open-heel fins, generally used for scuba diving, are usually larger and stiffer than shoe fins (see figure 4.7). Small, flexible snorkeling fins may be inadequate for the harder work of scuba diving.

Special materials and designs abound. The fundamental features of a fin are the size and stiffness of the blade (see figure 4.8). The larger and stiffer the blade, the greater the propulsion but also the greater the physical demand when you move the fin through the water. A blade that is too stiff can cause you to cramp and become fatigued. You should begin with fins of a moderate size and stiffness. When you can use those fins for extended periods without difficulty, you may consider fins that can provide greater propulsion.

The most important criteria for the selection of any item of diving equipment are fit and comfort, and this is especially true of fins. To help ensure proper fit, sit down and try on a fin. Wear a boot if you are trying on an open-heel fin. Hold your foot in the air and wiggle it up and down and from side to side. The fin and your foot should move as a single unit. If your foot moves inside the foot pocket, the fin is too large. Diving equipment should fit snugly but not tightly.

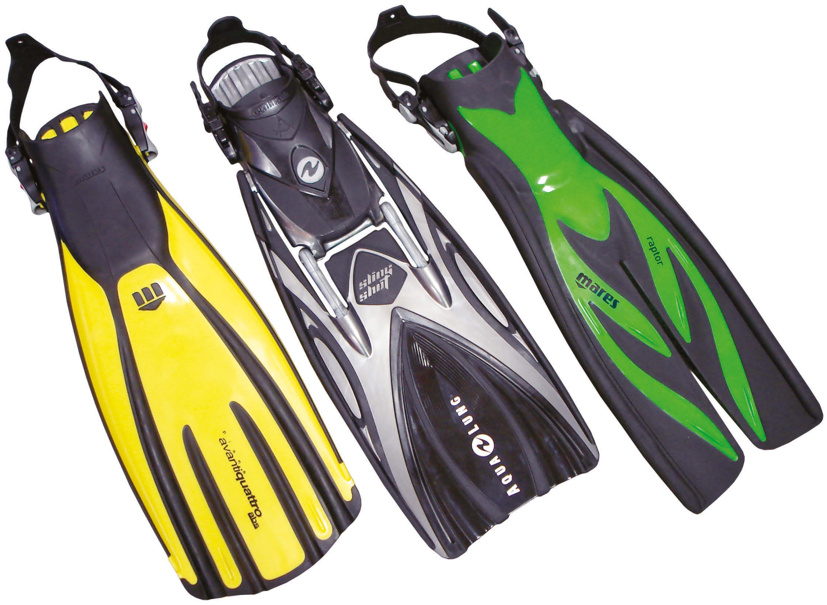

Figure 4.7 Three styles of open-heel fins.

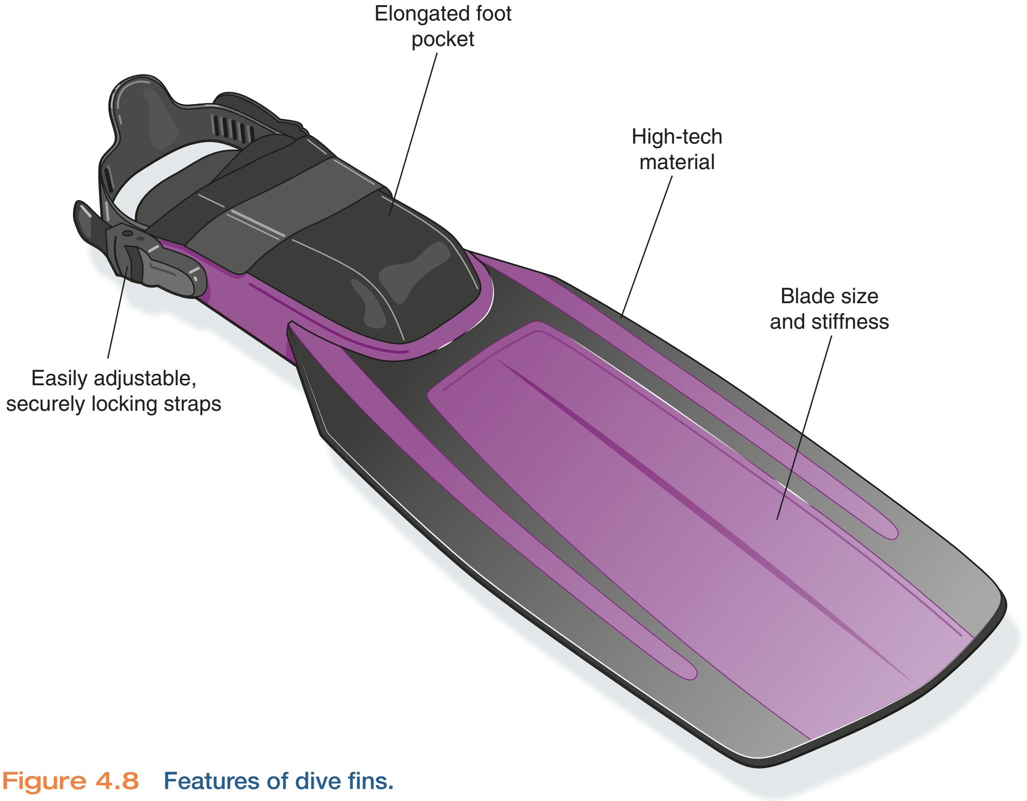
Elongated foot pocket
High-tech material
Blade size and stiffness
Easily adjustable, securely locking straps

Figure 4.8 Features of dive fins.

A fin that is too tight can cause your foot to cramp. The foot pocket of the fin should fit snugly but should not exert pressure on your foot. Select your fins using fit, comfort, and blade size and stiffness as your primary criteria. Features, fashion, and cost should be secondary considerations.

Skin Diving Vests

When you blow air into a skin diving vest (which is smaller and less complex than a buoyancy compensator), you can achieve positive buoyancy (floating). Positive buoyancy is invaluable when you want to rest at the surface or carry items you have collected while diving. Wear a buoyancy device whenever you dive. You can achieve negative buoyancy (sinking) by venting air from the skin diving vest.

Divers use one basic type of skin diving vest; this vest fits around the neck and secures at the waist (see figure 4.9). A standard feature of a skin diving vest is an oral inflation tube, which comes in various sizes. It is easier to use larger tubes with mouthpieces than smaller tubes without mouthpieces.

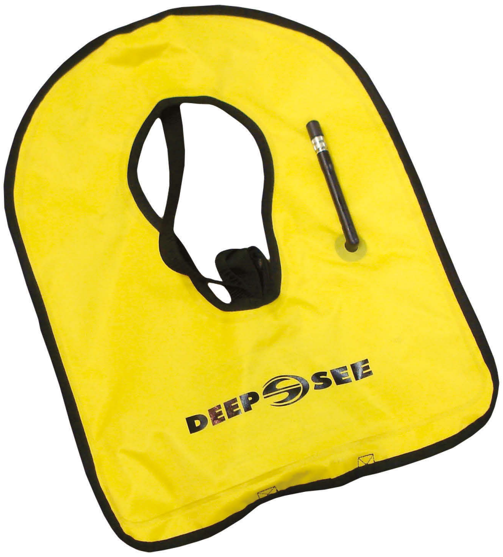

Figure 4.9 Skin diving vest.

Care and Maintenance of Skin Diving Equipment

Your skin diving equipment will provide years of service if you take care of it properly. Rinse the equipment with clean, freshwater after use; dry it in the shade; and store it in a cool, dark, smog-free location. Prolonged sunlight, smog, salt crystals, and swimming pool chlorine will harm your equipment. Fill a skin diving vest partially with freshwater after use, and swish the water inside. Drain the vest completely, and then inflate it for drying and for storage. Periodically inspect the straps on your mask and fins for drying and cracking, and replace straps when they begin to deteriorate. After your mask has dried, rub a drop of defogging solution on the lens to help keep the lens clean during storage.

Exposure Suits

You should wear some type of exposure protection when diving regardless of the water temperature. An exposure suit provides protection against scrapes and stings in addition to providing insulation.

The two basic types of exposure suits are wet suits and dry suits. Generally speaking, the colder the water, the thicker your suit should be and the more of your body should be covered by the suit. Suits come in many configurations, but most are one piece or a combination of a jacket and pants. Vests are accessories that help conserve heat in vital organs. Hoods, boots, and gloves help reduce heat loss from the head, feet, and hands, respectively.

Wet Suits

A wet suit is a good exposure suit for water temperatures ranging from 60 to 85 °F (15.6 to 29.4 °C). A wet suit allows water inside, but any water that enters can also exit and carry heat with it. The better the fit of the suit, the less water inside and the warmer the suit. Table 4.1 provides a comparison of the various types of wet suits. The following descriptions provide some additional information.

Table 4.1 Wet Suits

Type	Warmth	Features	Temperature range
Spandex	45% more than bare skin	Light, compact; useful as undergarments for thicker suits	78 °F (25.6 °C) +
Thermoplastic	30% warmer than spandex	Neutrally buoyant; no weights required; wicks perspiration; windproof	75 °F (23.9 °C) +
Plush-lined thermoplastic	10% warmer than unlined	Same as thermoplastic	72 °F (22.2 °C) +
Foam neoprene 1⁄8 in. (2–3 mm) 3⁄16 in. (4 mm) 1⁄4 in. (5–6.5 mm)	20-100% warmer than plush-lined thermoplastic	Buoyant; weights required; long drying time; evaporation chills wearer; minor repairs easy to do	Down to 60 °F (15.6 °C)
Hoods, vests, boots, and gloves or mitts	16-66% more warmth	Reduces water circulation; layering allows flexibility for various temperatures	Down to 60 °F (15.6 °C)

Spandex suits are thin, stretchy, attractive, full-body garments that are popular for diving in tropical waters. Spandex garments provide 45 percent more insulation than bare skin and provide protection against stings, scrapes, and sunshine. Spandex suits are light and compact for travel. They are also useful as undergarments for thicker neoprene wet suits, adding an extra layer of insulation and allowing you to slip into the heavier suit more easily. If you live in a temperate climate, you can use a spandex body suit beneath a neoprene wet suit for local diving and use the spandex suit by itself for vacation diving in the tropics (see figure 4.10).

Thermoplastic—sandwiched between two layers of spandex—is another type of wet suit material. A suit made from this three-ply material is about 30 percent warmer than a spandex suit. Thermoplastic suits are thin (1.2 to 1.4 mm) and designed for use in tropical waters (75 to 85 °F [24 to 29 °C]). Some suits feature soft, plush lining for extra warmth. The plush lining increases the warmth of the suit by about 10 percent. You can wear a thermoplastic suit beneath a neoprene wet suit for extra warmth in colder water.

Thermoplastic material offers several advantages: The balance between the weight and volume of this unique material makes it neutrally buoyant, so you may not need a weight belt. The stretchy fabric allows you to move freely. The material draws perspiration away from your body but is windproof.

A third type of wet suit material is foam neoprene, as in the suits shown in figure 4.11. The foam contains tiny bubbles of inert gas, which provide insulation. The thicker the wet suit material, the greater the insulation quality of the suit. A 1/8-inch (about 3 mm) neoprene wet suit is about 20 percent warmer than a plush thermoplastic suit.

The thickness of foam neoprene in wet suits ranges from 1/8 inch for warm-water diving to 3/8 inch (about 9 mm) for extreme cold-water diving. The most common wet suit thicknesses are 1/8 inch, 3/16 inch (about 4 mm), and 1/4 inch (about 6 mm). You can layer wet suit material on critical areas of your body to

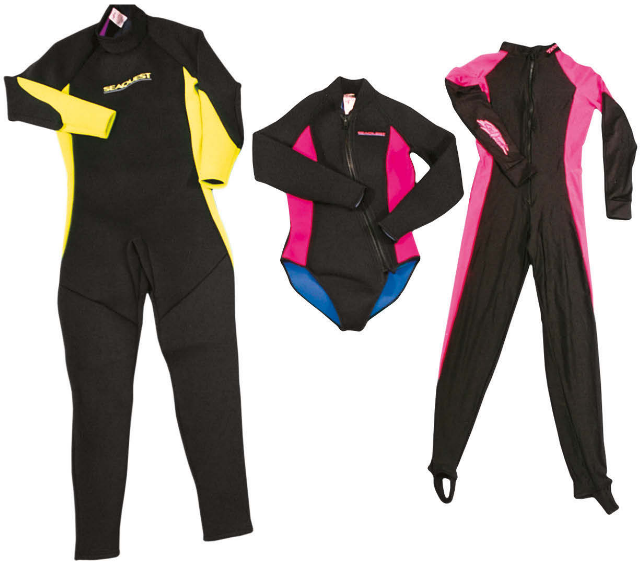

Figure 4.10 Common styles of warm-water wet suits: foam neoprene jumpsuit (left), shorty (center), and nylon (right).

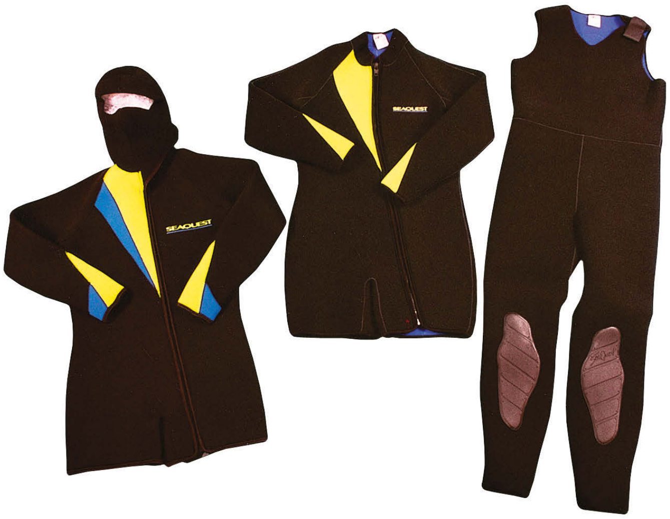

Figure 4.11 Common styles of cold-water wet suits: Jackets at left and center (worn with or without an attached hood) are worn over Farmer Johns at right to form layers.

reduce heat loss, but the thicker the insulation you wear, the more difficult it is for you to control buoyancy. Choose the thickness of wet suit material used by experienced divers in the area where you plan to dive.

Nylon usually covers both sides of the neoprene used for wet suits. The nylon increases the strength and durability of the suit, which is glued and sewed together. You can make minor repairs with wet suit cement, but you should have a wet suit manufacturer do extensive repairs.

Neoprene wet suits require weights to achieve neutral buoyancy. The suit provides immediate buoyancy when you release the weights, but the buoyancy of the suit can be either a benefit or a hazard. Neoprene is not windproof, and the evaporation of water from the suit between dives can chill you. You can wear a wet suit overgarment to retain warmth between dives in colder climates. Neoprene suits take longer to dry than other types of exposure suits. Mobility is good with thin neoprene wet suits but decreases as the thickness of the material increases.

Numerous wet suit designs are available, including the **shorty**, the one-piece **jumpsuit**, and **Farmer Johns** (see figures 4.10 and 4.11). You must consider several features when selecting a wet suit design. The more zippers a suit has, the more water circulates inside the suit and the greater the loss of heat. You can get zipperless suits for cold-water diving. Good wet suits feature a spine pad to minimize water circulation along the channel formed by your spine, and some suits for cold-water diving have attached hoods to minimize water circulation at the neck. You may spend time kneeling around and in the water, so knee pads are a desirable feature.

Heat packs are available for wet suits. These packs contain a nontoxic, reusable chemical that heats to about 130 °F (54 °C) for half an hour or more, depending on conditions. The packs fit into special pockets that are an optional suit feature.

You can buy wet suits in standard sizes, or you can have a suit tailored for a custom fit. The fit of a wet suit is its most important feature. The suit must fit snugly all over, but it must not fit so tightly that it hampers your breathing and circulation. A suit that fits well may feel slightly restrictive out of the water. The true test of the fit of a suit is to dive with it. You may be able to rent a wet suit identical to one you would like to purchase.

Wet Suit Accessories

Scuba divers wear footwear for warmth and foot protection. Several types of footwear—called boots or booties—are available for various needs. Boots, usually made from neoprene, may cover only the foot or both the foot and the ankle (see figure 4.12). Boots range from inexpensive neoprene socks to sturdy footwear with durable, molded soles. They may or may not have zippers. Zippered boots are easier to don and remove, but boots without zippers are warmer.

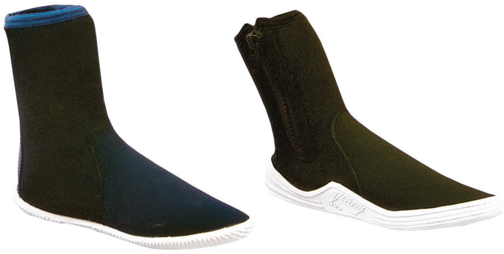

Figure 4.12 Boots are available with or without zippers.

You should wear hand coverings when the water temperature is less than 70 °F (21 °C). Some divers wear gloves for protection, such as when catching lobsters. Types of hand coverings include gloves, mitts, and gauntlets (see figure

4.13). Wear mitts in cold water because they are thicker and have less surface area for heat dissipation than gloves have. Wear gauntlets (neoprene mitts with long cuffs) when the water is extremely cold. Thin neoprene gloves provide sufficient insulation in temperate water. Do not wear gloves in tropical areas. You are more likely to touch things when wearing gloves, and grabbing delicate coral reefs and marine animals while wearing gloves may harm the animals.

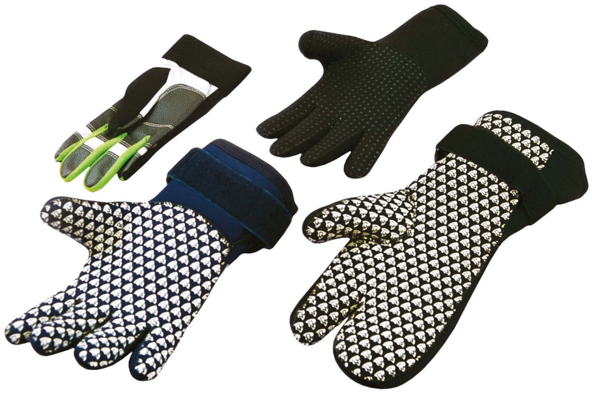

Figure 4.13 Gloves and mitts protect divers' hands from cold temperatures and injuries.

A hood is an important warmth accessory that can reduce your heat loss from 20 to 50 percent, depending on the temperature of the water. Two basic types of hoods—attached hoods and separate hoods—are available for wet suits (see figure 4.14). Cold-water divers like attached hoods because they restrict water circulation in a suit more than separate hoods do. Some separate hoods have skirts that end at the base of the neck; other cold-water hoods have large bibs that cover the neck and shoulder area. Thin hoods are used for protection in warm water, and thicker hoods are used for insulation and protection in cold water.

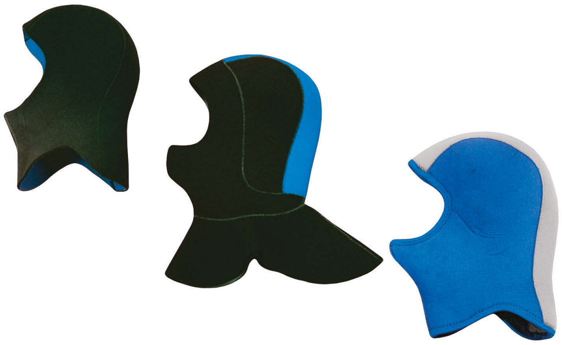

Figure 4.14 Hoods conserve warmth in cold waters.

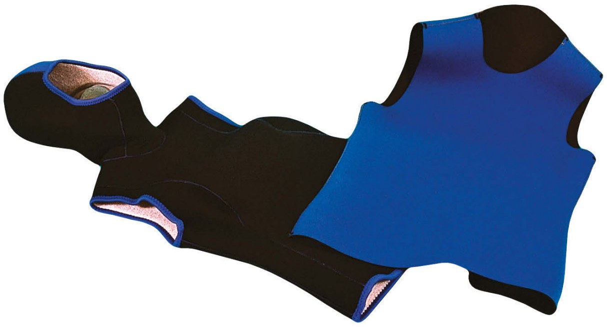

Figure 4.15 The addition of a vest can increase the warmth of your wet suit by as much as 16 percent.

You can increase the warmth of your wet suit by as much as 16 percent by wearing a vest. Wet-suit vests come in all types of material and provide an additional layer of insulation (see figure 4.15). Layering is an effective technique that reduces water circulation while increasing insulation. Some vests have attached hoods.

Wet Suit Care and Maintenance

With proper care, wet suits can provide years of service. Soak your wet suit in clean, warm freshwater after each use. If you can't soak your suit, rinse it. Hang your suit on a wet suit hanger, which is wider than a typical clothes hanger; dry the suit in the shade; and store it on the hanger in a cool, dark, smog-free location. A garage is a poor environment for storing equipment because of emissions from automobile exhaust. Do not fold your wet suit for storage because folds can form permanent creases, which insulate poorly. Inspect your suit regularly for seam integrity and tears. Make repairs or have them made as needed.

Dry Suits

When the water temperature is less than 60 °F (15.6 °C), consider using a dry suit, which is an exposure suit that keeps out water. Dry suits are much warmer than wet suits for three reasons:

1. Air is in contact with your skin instead of water. Because air conducts less heat than water, you lose less heat via conduction.
2. You wear undergarments beneath a dry suit. The undergarments trap a layer of air, which is a good insulator.
3. Suit compression affects two of the three types of dry suits only slightly, so the insulating ability of those types of dry suits remains nearly constant regardless of depth. Constant volume in a dry suit provides an advantage over a wet suit, which compresses with depth and expands during ascent.

Dry suits are identified by the material from which they are made. Table 4.2 compares the three most common types of dry suits. The foam neoprene dry suit is made of the same material as a wet suit, but the dry suit has seals at the wrists and neck, attached dry boots, and a waterproof zipper (see figure 4.16).

A dry suit can cost two to five times as much as a wet suit, but with proper care and maintenance, a good dry suit will last many times longer than a wet suit. If you dive primarily in cold water, a dry suit is a good investment.

A dry suit creates a closed air space that is subject to squeezing. A low-pressure inflator valve, which is a standard feature, allows you to add air to the suit during descent to prevent suit squeeze. Because air expands during ascent, an exhaust valve is another standard feature. Get a dry suit that has a constant-volume exhaust valve that automatically maintains a constant state of buoyancy during ascent.

Diving is a diuretic activity—that is, it increases the amount of urine you produce. Urination in a wet suit is a regular practice by divers in open water, but the problem is more complex when you wear a dry suit. An optional feature for a dry suit is a relief zipper, which is useful only when you are out of the water.

Table 4.2 Dry Suits

Type	Advantages	Disadvantages
Foam neoprene	Form fitting and streamlined	Long drying time; hard to locate and repair leaks; buoyancy control difficult
Crushed neoprene	Durable; easy to repair; long lasting; less buoyant than foam neoprene	Less insulation than foam neoprene; expensive; somewhat bulky
Shell (two kinds) • Coated nylon • Rubberized fabric	Fast drying; easy to repair; nylon suits inexpensive; rubberized suits long lasting	Easily punctured; bulky; nylon suits do not last long; rubberized suits expensive

Divers wearing a dry suit have to exit the water to urinate, and those without a relief zipper must disrobe.

Several types of undergarments for dry suits are available. Some are inexpensive but compress with depth and lose insulating ability when wet. More expensive undergarments are highly resistant to compression and retain most of their insulating quality even if wet. Many divers who wear dry suits prefer to wear two layers of undergarments: a thin garment against the skin to carry perspiration away from the body and a thicker overgarment to provide the bulk of the insulation. Moisture conducts heat, so you are warmer when perspiration is wicked away from your body.

Dry suits have several drawbacks. Controlling buoyancy is more difficult with a dry suit than with a wet suit. Rapid, uncontrolled ascents can occur unless you control the suit. Training is essential before attempting to dive in a dry suit. Dry suits are bulkier than wet suits. It is easier to don and remove a dry suit than a wet suit, but the bulkiness of a dry suit makes surface swimming difficult. A dry suit also restricts your mobility more than a wet suit does. The inconveniences of a dry suit are of little consequence, however, when warmth is your primary concern.

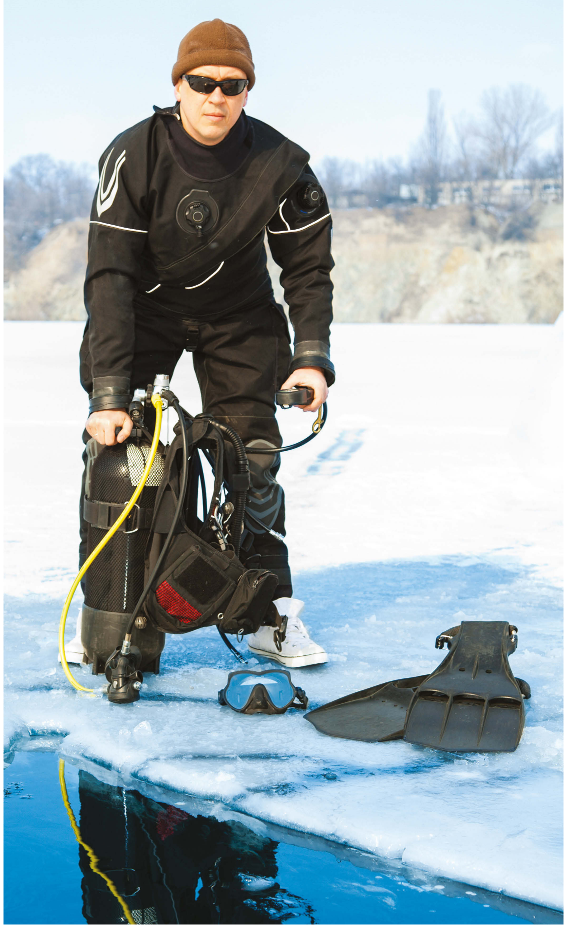

Figure 4.16 **Dry suits can be made from foam neoprene, crushed neoprene, or nylon.**

Berenika Lychak/iStock/Getty Images

Dry Suit Accessories

Most dry suits have attached boots, a desirable feature. Thin latex booties are found on less expensive suits; more expensive suits have hard-soled boots. You should wear heavy socks for insulation with latex booties, and you must wear wet suit boots over the latex booties to protect them. The end result is that divers who wear dry suits usually need fins with large foot pockets.

You can wear neoprene gloves or mitts with a dry suit. If the water is extremely cold, you can get dry gloves with insulating liners that attach to some dry suits. An attached latex dry hood is an option for some models, but most divers use a separate neoprene dry suit hood. Figure 4.17 shows examples of dry suit accessories.

Dry Suit Care and Maintenance

Dry suits require more care than wet suits. The zipper and the control valves are expensive to replace, so do not allow salt crystals to form in the zipper or in the valves. Soak and rinse the zipper and valves in clean freshwater as soon as possible after a saltwater dive. Wash the neck and wrist seals with soapy water, and then rinse them. Coat latex seals with pure talc after they dry. The talc helps protect the rubber from the elements. Fold dry suits in half over a wide hanger for drying. Lubricate the suit zipper according to the manufacturer's instructions, then store the suit with the zipper open. If the suit needs to be repaired, have an authorized dealer make the repairs. As with wet suits, store your dry

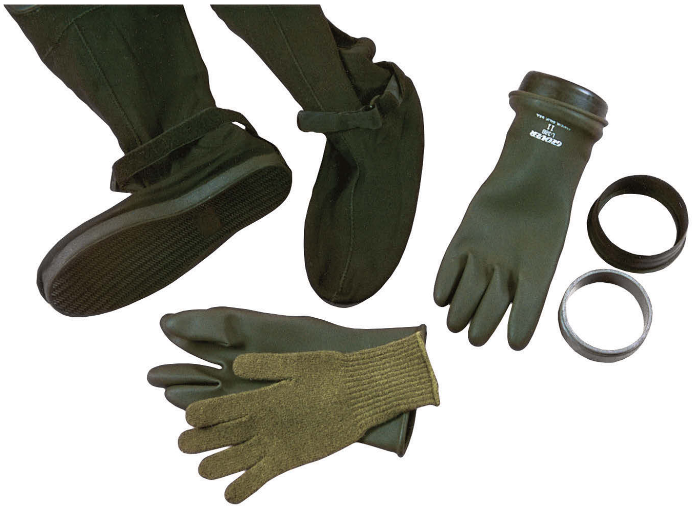

Figure 4.17 Dry-suit boots and gloves help keep divers warm.

suit inside your home. The rubber seals on dry suits are adversely affected by exhaust emissions.

Exposure Suit Selection

You should consider many factors when selecting an exposure suit and accessories: your physical characteristics, where you intend to dive, how you intend to dive, how much diving you intend to do, and what you intend to do while diving. The amount of money you want to invest is also a factor, but keep in mind that buying an inexpensive suit may be false economy. If the suit does not meet your needs, you will have to spend more money for another suit.

The amount of diving you intend to do is an important factor. If you plan to make only one dive per day, your insulation requirements are not as great as they are if you plan to dive several times per day. The more time you plan to spend in the water, the warmer your suit needs to be.

If you are a thin person who gets chilled easily, you need more insulation than the average person for a given water temperature. People with more-than-average body fat may not require as much insulation as those with average body fat; fat is a good, natural insulator. Thermal comfort is essential for diving safety and enjoyment.

If you intend to do most of your diving in one area, the most popular type of suit in the area is probably the best type for you. If the local waters are cold, you have to choose between a wet suit and a dry suit. If you choose to get a wet suit as your first exposure suit for cold-water diving, a layered design with Farmer John pants, a vest, and a step-in jacket with an attached hood retains more warmth than high-waisted pants, a regular jacket, and a separate hood. A custom fit retains more warmth than a suit of a standard size.

Dry suits are not desirable for long surface swims because you can overheat, and the drag caused by the bulkiness of some dry suits may cause you to tire or cramp. On the other hand, dry suits retain much more warmth at depth than wet suits. If most of your diving will be deeper than 40 feet (12 m), if the water is 60 °F (15.6 °C) or colder, and if you can avoid long surface swims, a dry suit is a good choice—as long as you obtain training before using the suit.

If you plan to dive in a variety of climates, a spandex or thermoplastic suit combined with a neoprene wet suit may be a good option. You can wear different parts of the suits to meet different warmth requirements. Your diving activity affects your needs. An underwater hunter looking for game generates more body heat than an underwater photographer whose movements are minimal. The less active you are while diving, the more insulation you need. Your desires, needs, and budget determine the features you select for your exposure suit. When choosing features for your suit, you should consider the features that are popular with local experienced divers and diving leaders.

The accessories you select for your suit depend on the type of suit you choose, the temperature of the water, the activity you intend to pursue, and your budget. A hood may be thin and short for warmer water, thick with a long skirt for colder water, or a dry suit type. Foot coverings may be low cut, ankle high, or attached to the suit. The soles may be soft or hard. Hand coverings range from nothing to thick mitts, gauntlets, or dry gloves.

Local diving professionals can help with your selection of an exposure suit and accessories. No matter what type of exposure suit you choose, keep in mind that it is an investment in your enjoyment of diving. Diving is not fun if you get cold.

Weighting Systems

Exposure suits increase your buoyancy. You need weights for ballast in order to achieve neutral buoyancy. One type of weighting system is a weight belt (see figure 4.18); another type integrates the weights into the scuba unit.

Weights can be attached to or inserted into a belt that you wear around your waist. The belt is heavy nylon webbing two inches (five cm) wide. You can thread the belt through lead weights, wrap pouches of lead shot around the belt, or put weights or pouches of lead shot into pockets on pocket-type belts. A hollow fabric belt that you can fill with lead shot is more comfortable on your hips than hard weights. Lead shot, which comes in different sizes, causes less damage than hard lead weights if you drop the belt accidentally. Smaller-diameter shot weighs more per volume than larger shot, so the more weight you need, the smaller the shot you should use.

Because exposure suits compress with depth, a weight belt around your waist loosens unless it has a means of compensating for the suit compression. An

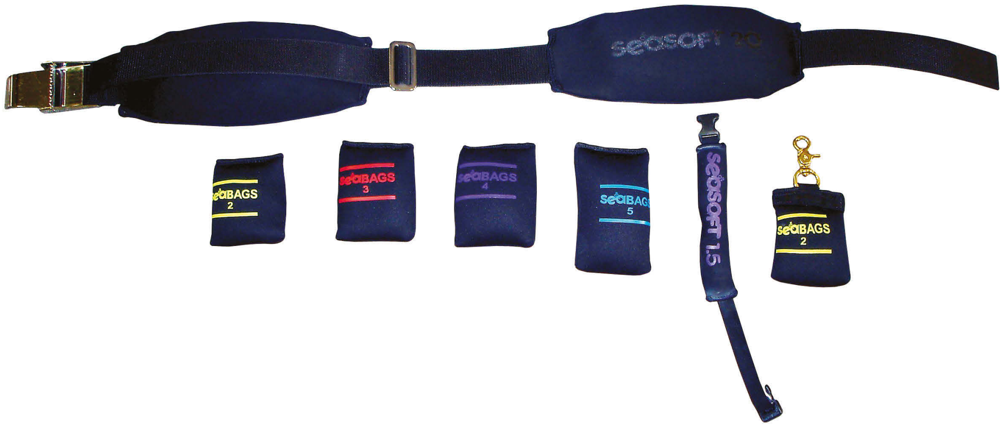

Figure 4.18 Weight belt with various sizes of weights.

elastic compensator is a desirable feature for a weight belt; you can select from a variety of designs.

Secure the weights on your weight belt so that they will not shift position. Pocket-type belts are good in this respect. When you thread separate weights onto a belt, you should secure the first and last weights with retainers, which are available as accessories.

The most important feature of any weighting system is the quick release. In the event of an emergency, you must discard weights quickly to establish positive buoyancy. No matter what type of system you choose, a reliable, easily located, and easily operated quick release is essential.

Several types of weights are available. Large, curved hip weights help offset the buoyancy of cold-water exposure suits. Smaller, rectangular weights find widespread use because they are economical. Coated weights are attractive and practical. Manufacturers mold weights into various shapes and offer a heavy vinyl coating as an option. The coating reduces pollution from lead, improves appearance, makes the weights easier to see in the water, and reduces suit abrasion. Fabric mesh packets filled with lead shot are popular. You can get bulk lead shot for hollow fabric belts. Coated lead shot, although slightly more expensive, is better than uncoated shot. Shot-filled tubular ankle weights are also available, although some diving experts believe ankle weights are unnecessary.

An integrated weighting system eliminates the need for a weight belt by holding lead weight in the **backpack** that holds the scuba cylinder or in the buoyancy compensator. An integrated weighting system makes the scuba unit heavier, but it also allows better distribution of weight than a separate system can provide. Some divers consider an integrated weighting system advantageous because weights are less likely to shift, a suit compression compensator is not needed, and having the weight above the waist improves the diver's trim in the water and reduces strain on the lower back. However, when you raise your center of gravity, you are more likely to lose your balance and fall when you are out of the water. You need strength, good balance, and caution to use an integrated weighting system. Integrated systems usually use lead shot, either in bulk or in pouches.

Weighting System Care and Maintenance

Weighting systems do not require as much care and maintenance as other items of diving equipment. Do not soak or rinse raw lead after use because lead in the runoff pollutes the environment. If gray water seeps from your weight system, replace the lead with new, coated lead. (You can recycle old lead, so do not throw it away.) Pocket-type weighting systems allow you to remove the weights and rinse the remainder of the system.

Regularly inspect the functional aspects, such as the quick release and the compensator. If you use a belt, inspect the weight retainers to make sure they are not broken, and make sure the end of the belt is clean and neat for easy insertion into the buckle. If the end of your belt is frayed, trim and singe it. When you fix a frayed end, be careful that you do not cut or burn yourself.

Weighting System Selection

When selecting a weighting system, you should consider your physical characteristics, the amount of weight you need, and how frequently you will need to change the amount of weight you use.

If you have a large build, your waist is larger than your hips, and your exposure suit is thick, a weight belt may not work well—especially if you need more than 30 pounds (13.6 kg) of weight. In chapter 6, you will learn how to test your buoyancy to determine the amount of weight that you need. Consult with divers whose physical characteristics are similar to yours for advice about weighting systems. If you require more than 30 pounds of weight, you may need weights on a weight belt plus an integrated system. If you require only a small amount of weight, nearly any type of weight system is acceptable.

If the type of diving you do varies, the amount of weight you need also varies. If you dive in freshwater and in saltwater, you need to adjust your weights. When you vary your exposure suit configuration, you also need to vary the amount of weight you wear. The more you need to change weights, the more you need a weighting system that allows changes to be made easily.

Buoyancy Compensators

A buoyancy compensator (BC) helps you control buoyancy. You can inflate your BC at the surface to increase buoyancy, deflate it to reduce buoyancy for descent, and add air to it to achieve neutral buoyancy underwater. Most BCs also contain a backpack to hold your scuba cylinder.

The three types of BCs are jacket style, back mounted, and front mounted (see table 4.3 for style comparison). Most BCs used in diving today have a wraparound jacket design, which provides front and rear buoyancy. There are two basic jacket designs: One style has inflation tubes over the shoulders; a newer style has straps over the shoulders. The straps have convenient, adjustable releases (see figure 4.19).

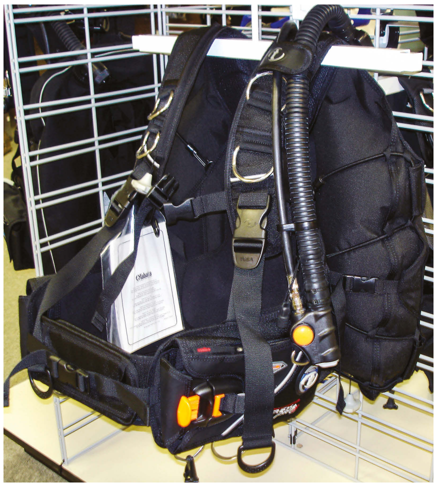

Figure 4.19 Jacket-style buoyancy compensator with shoulder straps and quick-release clasps.

Table 4.3 Buoyancy Compensators

Type	Location of buoyancy chamber	Advantages	Disadvantages
Jacket style	Front and rear	Even lift; diver can remain upright	Not suitable for skin diving
Back mounted	Rear	Does not interfere with valve operation for dry suits	Pushes diver forward; difficult for diver to remain upright
Front mounted	Front	Suitable for skin diving; allows diver to remain upright	Need separate backpack for cylinders; requires disconnection of inflator hose before removal; does not provide as much lift as jacket

Back-mounted BCs place buoyancy chambers behind you. These systems are useful for underwater modeling and for specialty diving activities. Models look better without bulky BCs covering them, and back-mounted units do not interfere with valve operation on dry suits as much as other types of BCs do.

Front-mounted BCs fit around your neck and cover your chest area. This was the first type of BC, but few divers use front-mounted BCs today. Most divers today believe that BC jackets are superior to front-mounted ones. You can use a front-mounted BC for both skin diving and scuba diving. Jacket-style and back-mounted BCs are not suitable for skin diving.

In its simplest form, a BC is a bladder with attachments (see figure 4.20). The bladder may be coated material that constitutes the BC itself, or a separate bladder may be inside a fabric shell. The seams of a BC bladder are glued or electronically welded, depending on the type of material.

Buoyancy Compensator Selection

When selecting a BC, you should consider your physical characteristics, where you plan to dive, and the type of diving you plan to do. Some BCs provide more buoyancy than others. A cold-water diver wearing a wet suit needs more lift than a warm-water diver wearing a thin exposure suit, although a large amount of lift is not necessarily desirable.

If you are a large, strong individual, the size and bulk of a BC may not be of much concern. But if you are small, you may be wise to choose a small, compact BC. The length of the BC is important. If the BC extends downward too far, this makes donning and removing your weight belt difficult. If you are shortwaisted, consider an integrated weighting system. The fit is important because a BC needs

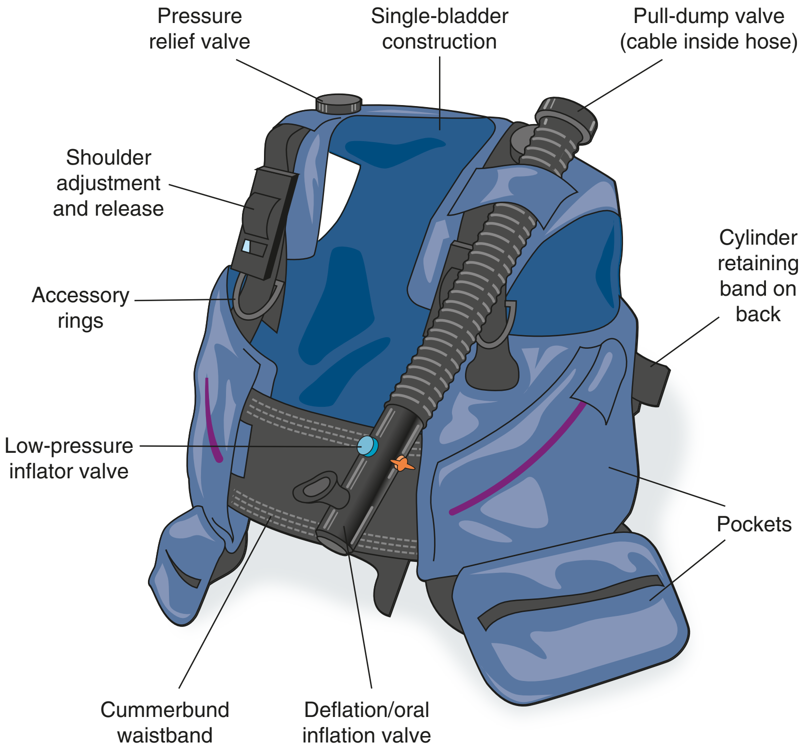
Pressure relief valve
Single-bladder construction
Pull-dump valve (cable inside hose)
Shoulder adjustment and release
Cylinder retaining band on back
Accessory rings
Low-pressure inflator valve
Pockets
Cummerbund waistband
Deflation/oral inflation valve

Figure 4.20 Desirable BC features.

to support you in the water. The BC should fit snugly and not ride up on your body. Models that allow adjustment for a custom fit are desirable.

Another desirable feature is a single-bladder construction, which is less expensive and requires less maintenance than a separate bladder. Renting and using different types of BCs will aid you in making a selection. Talk to experienced divers and diving professionals, and observe BC preferences in the area where you dive.

Buoyancy Compensator Care and Maintenance

Your investment in a BC is large. But, like most diving equipment, your BC will provide years of service if you take care of it. You need to rinse your BC inside and out after use, especially after use in a swimming pool or in the ocean. Chlorine in pool water and salt crystals from seawater are harmful to your BC. Drain the water from your BC after use, fill it about a third full with freshwater, swish the water around, and then drain the bladder. Rinse the inflator assembly thoroughly, and leave the BC fully inflated until it dries. Inflation will test the airtight integrity of the bladder and valves. If the BC does not remain firm for

Diving Equipment Checklist

- ✓ Mask, snorkel, and snorkel keeper
- ✓ Fins and boots
- ✓ Scuba tank (filled)
- ✓ Buoyancy compensator
- ✓ Exposure suit, hood, and gloves
- ✓ Weight system
- ✓ Regulator with pressure gauge
- ✓ Alternate air source
- ✓ Instruments to monitor depth, time, and direction (separate or integrated)
- ✓ Signaling devices (whistle, mirror, safety tube)
- ✓ Dive knife
- ✓ Float, dive flag, and anchor
- ✓ Dive tables
- ✓ Dive light
- ✓ Slate and pencil
- ✓ Marker buoy
- ✓ Collecting bag
- ✓ Gear bag

Spare equipment

- ✓ Scuba tank(s)
- ✓ Weights
- ✓ Straps
- ✓ O-rings
- ✓ Snorkel keeper

Secondary equipment

- ✓ First aid kit
- ✓ Emergency phone numbers and radio frequencies
- ✓ Logbook
- ✓ Swimsuit
- ✓ Towel
- ✓ Jacket
- ✓ Hat or visor
- ✓ Sunglasses
- ✓ Dive kit
- ✓ Save-a-dive kit
- ✓ Drinking water

at least an hour, take it to a professional repair facility. Anytime your BC fails to function properly, have it professionally repaired. It is hazardous to attempt BC repairs without special training, tools, and parts.

Scuba Cylinders

A scuba cylinder stores compressed air at high pressure. The container must be strong and free of corrosion. Scuba cylinders (also called tanks) are made of either aluminum or steel. Each type has advantages and disadvantages. Figure 4.21 shows examples of various steel and aluminum tanks.

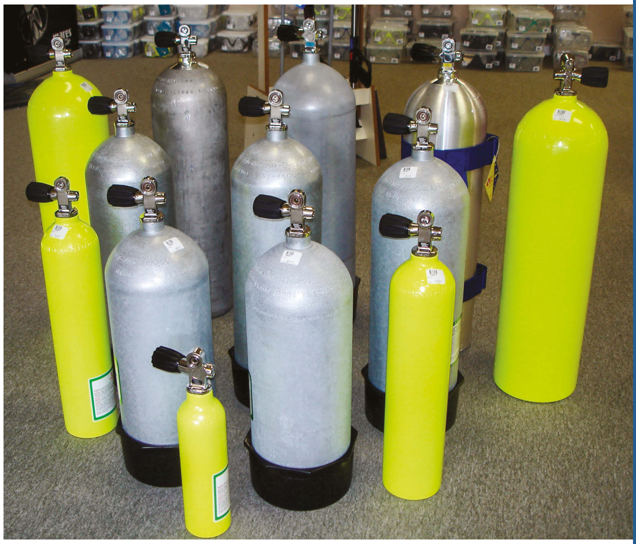

Figure 4.21 Scuba cylinders come in many sizes and are made of steel or aluminum.

Steel Cylinders

Steel scuba cylinders come in various sizes with various pressure ratings. Common sizes are 63, 71.2, 80, and 100 cubic feet (1,784, 2,016, 2,265, and 2,832 L). The pressure to which the cylinders can be filled, called the working pressure, ranges from 1,800 to more than 4,000 pounds per square inch (psi), or 122 to 272 standard atmospheres (atm).

Some compressors pump air only to 2,500 psi (170 atm). When you cannot get air at a pressure higher than 2,500 psi, a steel tank may be more desirable than an aluminum tank. A steel 71.2-cubic-foot scuba tank filled to 2,250 psi (153 atm) contains about 5 cubic feet (142 L) more air than an 80-cubic-foot aluminum tank filled to the same pressure. The aluminum tank must be filled to 3,000 psi (204 atm) to obtain 80 cubic feet.

The main disadvantage of a steel cylinder is that it can rust, which can render a tank unsafe and unusable. Do not allow water inside a scuba cylinder. The high-pressure atmosphere has a large amount of oxygen to fuel corrosion. You can keep the inside of a steel scuba tank dry, but the outside is exposed to moisture. Galvanizing inhibits rust on the outside of steel tanks, but the inside may not be galvanized because galvanization affects air purity. Painting a galvanized surface may improve the appearance of a cylinder, but paint alone is an inadequate finish because cracks or chips in the paint allow moisture to reach the steel. The tank will begin to rust unless there is a galvanized coating beneath the paint.

Another disadvantage of a steel scuba tank is that it has a rounded bottom because of the manufacturing process. The tank cannot stand by itself unless you place a rubber or plastic boot, called a tank boot, on the end of the cylinder. The boot makes the base of the tank flat so it will stand. Some boots have flat sides to help keep a tank from rolling when you lay the cylinder on its side. Moisture and salt trapped between the tank boot and the cylinder can cause corrosion. Boots with internal ridges—the preferred type of boot—are self-draining.

Aluminum Cylinders

Aluminum alloy cylinders also come in various sizes with various pressure ratings. Common sizes are 63, 80, and 100 cubic feet (1,784, 2,265, and 2,832 L). The working pressure for aluminum cylinders is 3,300 psi (224 atm).

Aluminum corrodes, but the oxide that forms arrests the corrosive process—this is a significant advantage over the corrosive process of rust in steel tanks. Rust is an accelerating process, but corrosion in an aluminum cylinder is a self-arresting process.

The bottom of an aluminum cylinder is flat. You do not need a tank boot on the cylinder to allow it to stand by itself, but many divers put boots on aluminum tanks to protect the tank and any objects struck by the bottoms of the cylinders.

Aluminum cylinders also have drawbacks. Aluminum is softer than steel, so aluminum tanks can be dented and gouged more easily than steel tanks. On aluminum tanks, brass cylinder valves control the flow of air. Electrolytic action between the dissimilar metals of the cylinder and the valve can cause the valve to seize in the aluminum cylinder threads unless the valve is removed periodically and coated with a special compound. Valve seizing seldom is a problem with steel cylinders.

Aluminum cylinders do not need to be galvanized. You can paint them to improve appearance, but do not bake the paint finish. Temperatures hotter than 180 °F (82.2 °C) reduce the strength of an aluminum cylinder and can cause it to explode when filled. If you would like your cylinder painted, have it done by a professional tank-painting service.

Cylinder Markings

Several rows of markings on the neck of a scuba cylinder provide useful information about the cylinder. You should be able to determine the meaning of several of the marks. Figure 4.22 shows an example of cylinder neck markings.

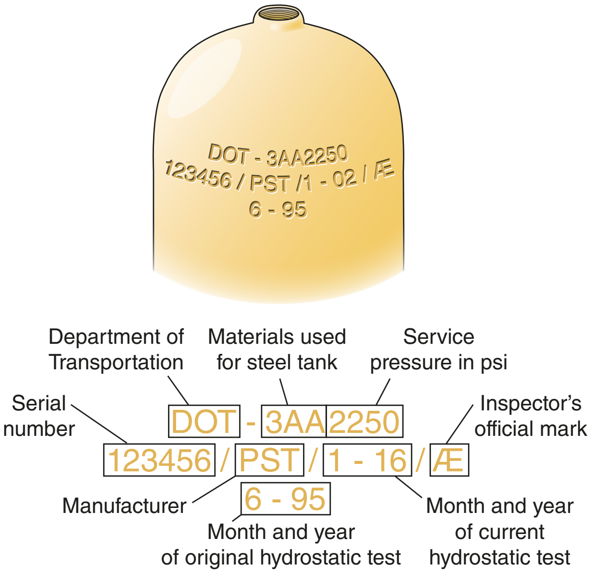
DOT - 3AA2250
123456 / PST /1 - 02 / Æ
6 - 95
Department of Transportation
Materials used for steel tank
Service pressure in psi
Serial number
Inspector's official mark
DOT - 3AA 2250
123456 / PST / 1 - 16 / Æ
6 - 95
Manufacturer
Month and year of original hydrostatic test
Month and year of current hydrostatic test

Figure 4.22 Cylinder neck markings.

The first row of marks on tanks manufactured in the United States discloses the government agency that sanctioned the manufacture of the tank, the type of metal from which the tank was made, from, and the working pressure. The first letters in the row identify the government agency, such as DOT (U.S. Department of Transportation), CTC/DOT (Canadian Transportation Commission and the U.S. Department of Transportation), or ICC (the former U.S. Interstate Commerce Commission). The next characters in the first row identify the type of metal that the tank was made from. The marks 3A and 3AA are designations for steel cylinders. The marks 3AL, E 6498, and SP6498 are designations for aluminum cylinders. The final figures in the first row—the important ones for you to remember—are numbers indicating the working pressure of the cylinder in pounds per square inch.

The second row of markings includes the serial number of the tank (which you should record for identification purposes) and letters or numbers identifying the manufacturer of the cylinder. The month and year of the tank's first pressure test are located somewhere below the second row. U.S. regulations require compressed gas cylinders to be pressure-tested before being put into service and every five years thereafter. Subsequent test dates can be stamped anywhere on the neck of the tank. A registered symbol between the month and year of a pressure test identifies the facility that did the testing.

Tanks are filled at an air (or "fill") station (see figure 4.23).When you take your cylinder to an air station for filling, the facility personnel will examine the tank markings to determine whether the cylinder test date is current and to identify the pressure to which the tank can be filled.

Cylinder Selection

The main criteria for the selection of diving equipment—fit and comfort—apply to scuba tanks as well. A small person should use a small tank. Larger divers may require larger tanks because they have larger lungs and use more air.

The material from which a tank is made affects its capacity, size, and working pressure. Although steel is heavier than aluminum, aluminum tanks are larger and heavier than steel tanks with similar capacities. Because aluminum is not as strong as steel, the walls of an aluminum tank are thicker than the walls of a steel tank of approximately the same capacity. The higher the pressure rating, the greater the capacity of a tank of a given size. Modern high-pressure steel scuba tanks are high-capacity cylinders, but they are quite heavy. Good sizes for small divers to consider for an initial scuba tank are 50 and 63 cubic feet (1,416 and 1,784 L). Popular sizes for divers of average size are 71.2 and 80 cubic feet (2,016 and 2,265 L).

Another important factor to consider when selecting a scuba tank is the buoyancy of the cylinder, which is determined by the volume and weight of the tank. Aluminum tanks are more buoyant than steel tanks. High-capacity tanks have a greater change in buoyancy than smaller cylinders as you use air from them. The change in buoyancy between a full and an empty tank can vary by more than 8 pounds (3.6 kg; see table 4.4). Some tanks are negatively buoyant

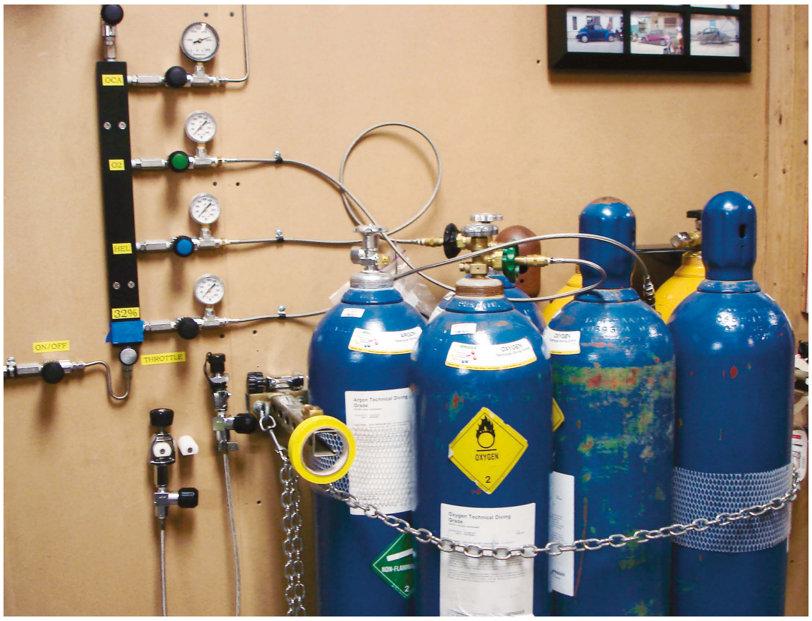

Figure 4.23 Scuba cylinder filling station.

Table 4.4 Cylinder Size, Working Pressure, and Buoyancy

Capacity (ft³/L)	Working pressure (psi)	Buoyancy (lb) from full to empty
Aluminum 50/1,416	3,000	–2.7 to +1.3
Aluminum 63/1,784	3,000	–2.3 to +2.7
Steel 71.2/2,016	2,250	–2.0 to +3.6
Aluminum 80/2,265	3,000	–2.0 to +4.4
Steel 76/2,152	2,400	–6.5 to –0.1
Steel 80/2,265	3,500	–7.4 to –1.0
Steel 102/2,888	3,500	–7.6 to +0.5

whether full or empty, but most tanks sink when full and float when empty. Buoyancy for tanks varies so much that you should select the tank most often used for diving in your area or try diving with several different tanks to determine which is easiest to manage.

You need to decide whether you are going to dive using compressed air, enriched-air nitrox, or mixed gases. If your breathing gas is anything other than compressed air, your cylinder, cylinder valve, and regulator must be dedicated to the specialized gas. It is unsafe to use a compressed air cylinder and regulator with mixed gases. Regulators, valves, and cylinders must be “oxygen clean” and “oxygen safe” when used with mixed gases.

Multiple-tank scuba units are used for specialty applications. As a beginning diver, you do not need double or triple scuba tanks. A single tank is adequate for most diving activities.

Cylinder Accessories

Fabric or plastic sleeves can help protect the exterior of your scuba cylinder. Some sleeves have places to attach items within easy reach. Tank bands, which secure a cylinder to a buoyancy compensator, vary. Some allow you to attach a small, backup scuba cylinder to your main cylinder. Additional accessories are available to help you carry or transport your scuba cylinders. Tank boots are a desirable cylinder accessory.

Cylinder Care and Maintenance

Scuba tanks are high-pressure vessels. They are strong, but you should handle them with care. Exterior damage can render your scuba tank useless. Avoid throwing scuba tanks or allowing them to roll about on the deck of a boat or in the trunk of your car. Secure cylinders for transportation or storage. Unless you are holding a scuba cylinder, do not leave it standing in an upright position, especially at a dive site. If the cylinder falls, it can injure someone or damage the tank valve or regulator. However, you should store your scuba tanks in an upright position so that any moisture inside will settle to the bottom where inspectors can detect it with relative ease. Rinse the outside of your cylinder with clean, fresh water after use, and pay special attention to the boot area of steel tanks.

Corrosion can ruin a cylinder rapidly, and pieces of corrosion can damage a tank valve or scuba regulator. One way to prevent moisture from entering a tank is to keep air in the tank. Water can get into an empty scuba cylinder while you are diving, so avoid using all the air in your tank. Moisture can enter an empty tank if you store it with the valve open. Store your scuba cylinder with a few hundred pounds (about 20 atm) of pressure inside. A low pressure keeps moisture out but provides little oxygen to aid corrosion if there is moisture in the tank.

The filling process can force water into the tank. A water trap in an air compressor is supposed to remove moisture from air, but if the moisture-removal system does not function properly, water can be pumped into your tank along with air. The filling hose attachments for scuba tanks can get wet, and water inside the filling attachments can be forced into your cylinder. Using a high-quality air station is important.

The diving industry requires an annual visual inspection of scuba cylinders. The examination consists of an external inspection, the removal of the valve from a cylinder, an internal inspection using a special light, special electrical testing for aluminum cylinders, the replacement of the valve, and the attachment of a decal indicating the inspection date. Most air stations require a current inspection sticker on a tank before they will fill it. When you handle your scuba tank, listen for sounds of anything moving inside the cylinder. If you hear anything, have the tank visually inspected.

U.S. government regulations specify that compressed gas cylinders must be pressure-tested every five years. Some countries require pressure testing every year or two. The test is hydrostatic because it takes place in water. An inspector fills a scuba tank with water and submerges it in a closed container that is completely filled with water. The inspector applies pressure to the scuba tank hydraulically, and the tank expands slightly from the pressure. The expansion displaces water from the container holding the scuba tank. The inspector measures the expansion and then releases the pressure. The tank must return to within 10 percent of its original volume within a specified period of time. If the tank is too brittle to expand and contract correctly, the inspector condemns it.

You may transport a scuba tank on an airplane only if the tank is completely empty and the valve is open—a situation that is not good for scuba tanks. Do not transport your cylinder by air. Diving destinations have tanks readily available, so there is no need to take a tank on a dive trip.

Cylinder Valves

Cylinder valves control the flow of a liquid or gas. Four types of valves are available for scuba tanks: the simple valve, the reserve valve, the high-pressure valve, and the multiple-cylinder valve. Because multiple tanks are for advanced specialty diving activities, this section addresses only the simple and high-pressure valves.

Simple Valve

A simple valve is an on-off valve that operates like a faucet. You turn the valve handle counterclockwise to open it and clockwise to close it. The first catalog of diving equipment listed this type of valve as item K, and the valve has been identified as a **K-valve** ever since (see figure 4.24).

The **valve seat** is a soft-sealing surface. It is the portion of the valve that closes and stops the flow of air. You can damage the seat with excessive closing pressure.

Valves for scuba tanks have several features, one of which is a snorkel tube that extends from the bottom of the valve into the scuba cylinder. The valve snorkel prevents moisture or particles from entering the valve when you invert the tank. Another standard feature of tank valves is a thin, metal disk called a **burst disk**. If a tank is overfilled or the heat from a fire causes the tank pressure to increase to a hazardous level, the disk will burst and vent the tank to prevent an explosion. The disks corrode over time, and occasionally a burst disk ruptures. The failure makes a loud noise, and the tank hisses loudly, but the situation is not dangerous (even if it occurs while you are diving, which is unlikely). If the burst disk in your tank valve ruptures, you need to have the valve serviced professionally. Manufacturers rate burst disks for various pressures, and the correct disk must be used. Keep the pressure rating of the burst disk in mind if you want to change a valve from one scuba tank to another. A valve with a low-pressure burst disk will rupture if you use it on a tank with a higher pressure rating.

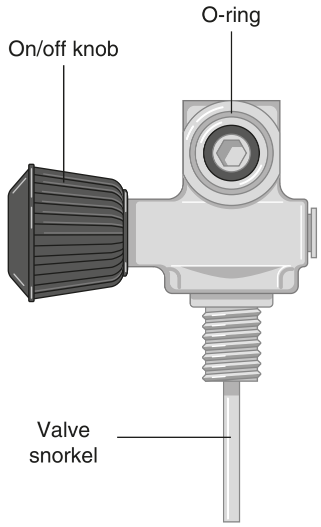

Figure 4.24 K-valve.

There are two types of outlets for scuba tank valves. The traditional outlet is nearly flush with the surface of the valve and surrounded by a soft, circular ring called an **O-ring**. The ring forms the high-pressure seal between the valve and the scuba regulator, so the ring must be clean and free of nicks or cuts. A newer type of threaded outlet with a recessed O-ring seal is a **DIN valve**, which withstands higher pressures than a traditional O-ring valve (see figure 4.25). Tank pressures in excess of 3,000 psi (204 atm) require a DIN fitting.

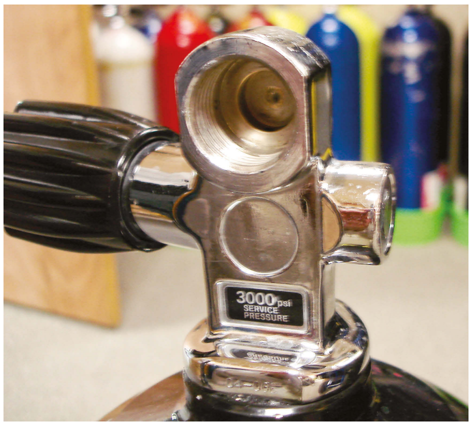

Figure 4.25 Close-up of a DIN valve.

Reserve Valve

A **J-valve** (the valve was listed as item J in the first equipment catalog) was designed to maintain a reserve of air to permit a normal ascent. The introduction of **submersible pressure gauges** (SPGs) for scuba tanks has rendered the J-valve obsolete.

The reserve lever on a J-valve created problems. If the lever was in the incorrect position (down), the valve would not maintain a reserve. If the diver failed to put the reserve lever in the up position before a dive or if the lever was bumped during a dive, the diver relied on a reserve of air that was not available. Safety concerns also contributed to the demise of the J-valve.

Valve Care and Maintenance

Cylinder valves are made of soft metal and have thin areas, so physical abuse can ruin them. A protective cap is a good accessory for a tank valve. The cap helps prevent loss of the O-ring, helps keep dirt out of the valve, and helps prevent physical damage to the high-pressure sealing surface around the valve opening. Protection of the valve is one reason that you should secure scuba tanks when you store them and should not leave them standing unattended. If a tank rolls about or falls over, this can damage the valve and render it inoperable.

When you open a tank valve, you should turn it slowly, open it all the way, and then close it one half turn. If something strikes the handle, a valve seat that is fully open will be damaged more than if it were closed slightly. When you close the valve, avoid excessive force, which shortens the life of the valve seat.

Rinsing your tank valve after use is a good idea, but soaking the valve by inverting your scuba tank in a container of warm water is better than rinsing. Water remains in the valve opening after rinsing or soaking. Open the valve momentarily to blow the water from the opening. If you leave the water in the opening and have the tank filled before the water evaporates, moisture will be forced into the tank.

Have your tank valve serviced annually by a professional. Also have your valve professionally serviced anytime it fails to operate easily or when the burst disk needs to be replaced. Valves receive partial servicing (lubrication) during the annual visual inspection of your scuba tank. The large O-ring that seals the valve to the scuba tank can be replaced at the time of the visual inspection. But partial servicing of your valve during a tank inspection is not the annual servicing of the valve. In a complete valve servicing, a qualified repair technician completely disassembles the valve, cleans the parts, replaces various parts, reassembles the valve, and then tests the unit.

Scuba Regulators

The function of a regulator is to reduce high-pressure air to a breathable level. Most scuba regulators use two **stages** of pressure reduction. The first stage of a regulator attaches to the valve of a scuba cylinder and reduces the high pressure to an intermediate pressure of about 140 psi (9.5 atm). The first stage of the regulator connects via a hose to the second stage, which contains the mouthpiece. The second stage reduces the pressure from the intermediate level to the surrounding pressure. A scuba regulator is a demand system; it delivers air only when you demand it by inhaling, unlike a constant-delivery, free-flow system, which commercial divers use. Scuba regulators are highly reliable and have a fail-safe design that turns the demand system into a free-flow system in the event of a component failure. Figure 4.26 provides more information about scuba regulator terms.

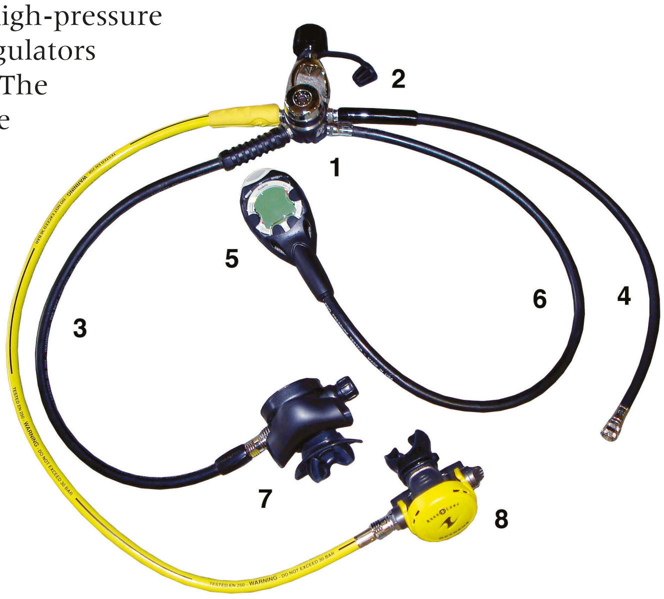

Figure 4.26 Scuba regulator nomenclature: (1) first stage, (2) dust cover, (3) low-pressure hose, (4) low-pressure inflator hose, (5) console, (6) high-pressure hose, (7) primary second stage, (8) extra second stage.

First Stages

The first stages on scuba regulators are either balanced or unbalanced. Changes in tank pressure affect the performance of a balanced first stage only slightly. With an unbalanced first stage, the performance of the regulator changes as tank pressure changes, so a balanced first stage is desirable.

The two main types of valves for the first stages of regulators are diaphragm and piston (see figure 4.27). A diaphragm first stage has a diaphragm that keeps water and dirt from the working parts inside. A bias spring combined with water pressure pushes the valve open. Tank pressure closes the valve when the first-stage pressure equals intermediate pressure plus water pressure. A diaphragm valve has more parts than a piston first stage, so it is more expensive to manufacture and service. The diaphragm valve's exclusion of water and dirt from the mechanics allows high performance for longer periods of time than a piston valve.

Piston first stages have an open, simple design with few moving parts. Water pressure in direct contact with the piston combines with the force of a bias spring to open the piston valve. Tank pressure causes the piston to move and the valve to close when the first-stage pressure equals intermediate pressure plus water pressure. Piston first stages are easier and less expensive to service than diaphragm first stages. But dirt, salt crystals, and mineral deposits that accumulate inside a piston regulator can affect its performance. Each type of first stage has advantages and disadvantages, so either type is acceptable. A diaphragm-piston regulator combines the diaphragm and piston concepts. The diaphragm excludes water and dirt and transfers pressure to a piston.

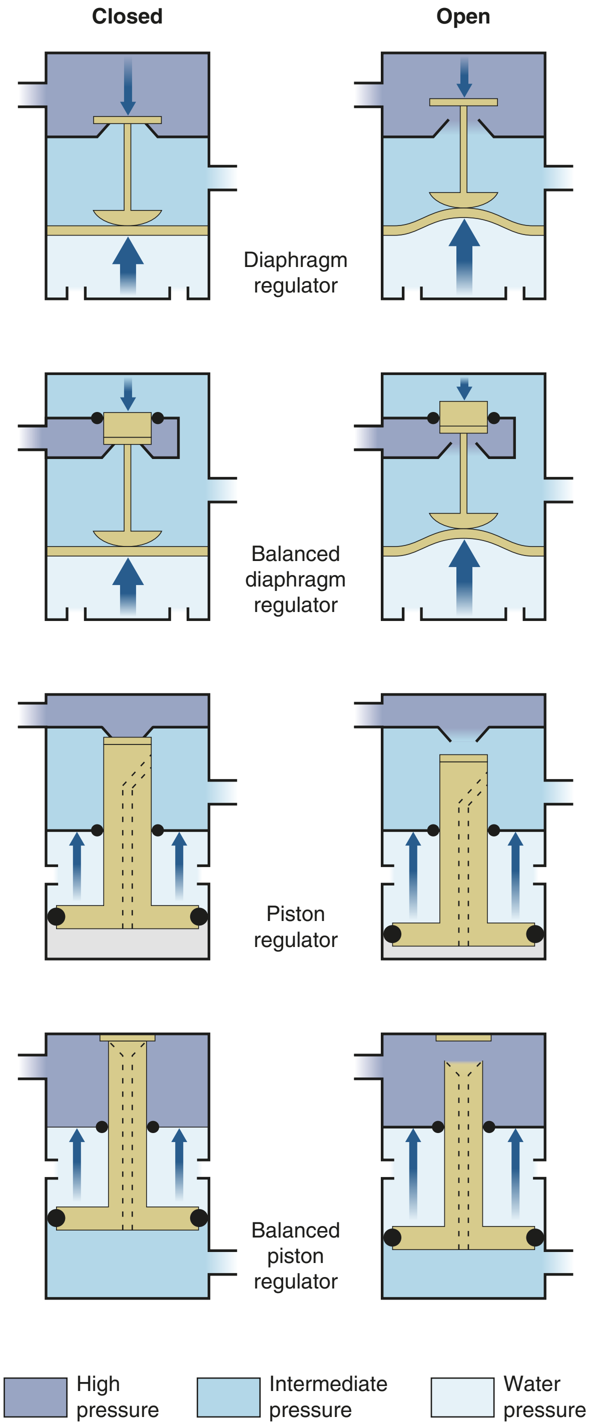

Figure 4.27 Typical regulator first stages. Red arrows represent spring pressure. Black dots represent O-rings.

A first stage must have a means of attachment to the tank valve. A typical regulator has a yoke that surrounds the valve and mates the regulator to the tank valve. The regulator inlet, which has an inlet filter, secures to a tank with a yoke screw. Scuba regulators that operate at above-average pressures use a DIN fitting instead of a yoke screw. A DIN fitting screws directly into a DIN valve and does not have a yoke.

First stages include multiple openings called **ports**. One of the ports is for high-pressure air measurement with an SPG. The remaining ports are for low-pressure air. A regulator should have several low-pressure ports to supply air to the primary second stage, an alternate second stage, a BC inflator, and possibly a dry suit inflator. The sizes of ports vary. The high-pressure port is usually larger than the low-pressure openings. This helps prevent connection errors; a low-pressure hose inadvertently connected to the high-pressure port would rupture.

Some regulators feature environmental shielding by sealing special fluid inside a flexible chamber attached to the first stage. The sealed, flexible chamber transmits water pressure to the regulator, but no water, salt, or dirt can enter the first stage. Extremely cold water can cause an unshielded regulator to freeze, but the fluid in an environmental chamber does not freeze.

Second Stages

The most common second stage is shaped like a cup lying on one side (see figure 4.28). Imagine a pliable diaphragm across the top of the cup, a mouthpiece attached to the bottom of the cup, and an exhaust valve attached to the lower side. A lever that activates a valve inside the container is in contact with a static diaphragm. An inhalation through the mouthpiece creates a partial vacuum inside the second stage. The pressure reduction pulls in the diaphragm, which moves the lever and opens the valve, allowing air to flow from the first stage

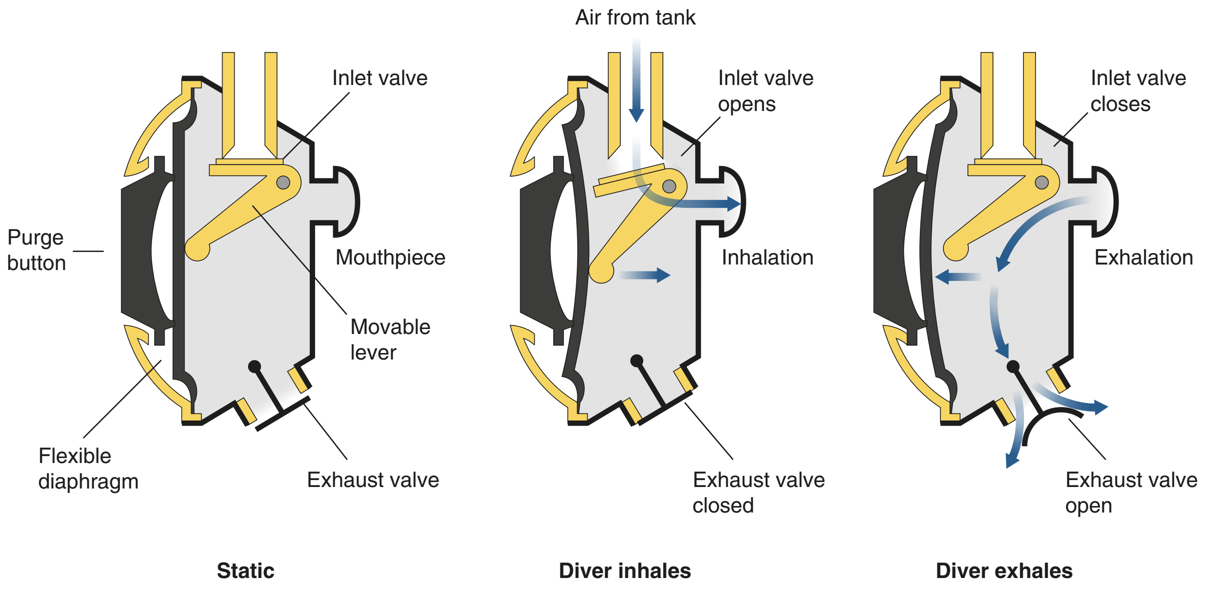
Inlet valve
Purge button
Mouthpiece
Movable lever
Flexible diaphragm
Exhaust valve
Static
Air from tank
Inlet valve opens
Inhalation
Exhaust valve closed
Diver inhales
Inlet valve closes
Exhalation
Exhaust valve open
Diver exhales

Do not depress purge button when rinsing a regulator.

Have regulator serviced professionally every year.

Figure 4.28 Regulator operation.

of the regulator into the second stage. When you stop inhaling, the buildup of pressure returns the diaphragm and lever to their normal positions, closing the valve and stopping the flow of air. When you begin to exhale, increased pressure inside the second stage opens the exhaust valve, allowing air to escape.

The two types of second-stage valves for scuba regulators are the downstream valve and the pilot valve. In a downstream-valve regulator, a small bias spring holds a valve closed. Inhalation moves the diaphragm, which moves a lever. The lever movement overcomes the resistance of the bias spring and opens a valve, allowing intermediate-pressure air to enter the second stage. After inhalation, air flows until the diaphragm moves outward, allowing the bias spring to close the valve. Downstream second-stage valves are simple and inexpensive, and they tolerate more sand and dirt than pilot valves do.

In a pilot-valve regulator, the movement of the diaphragm in the second stage opens a small valve that in turn opens a larger valve. When you stop inhaling from the regulator, the diaphragm returns to its normal position, and the valves close. A pilot valve delivers air up to four times more easily than a downstream valve. But pilot valves are more expensive to manufacture and service than downstream valves, and in shallow water, you may experience the effect of shuddering air movement with a pilot-valve regulator.

All regulators have a purge button or area on the regulator that you depress to manually open the second-stage valve. Use the purge to test the regulator, to expel water and debris from inside the second stage, and to relieve the pressure in the regulator after you close the tank valve.

The location of the exhaust valve varies. The exhaust valve may be at the bottom, the side, or the front of the second stage. The position of the exhaust affects the bubble pattern and the clearing of the regulator when it has water inside and the mouthpiece is in your mouth. Some regulators direct exhaust bubbles by means of an exhaust tee. You will learn more about regulator positioning for clearing in chapter 6.

A regulator is either right-handed, left-handed, or bidirectional, referring to the direction from which the regulator hose must come when the regulator is in your mouth. For example, the hose must come from the right side when you are using a right-handed regulator. The hose may come from either side when you are using a bidirectional regulator. The directional configuration is important only for knowing how to orient the regulator when you place it in your mouth. Figure 4.29 shows the possible directional configurations.

Some second-stage casings feature strong, light, durable materials that do not bend or corrode as metal does. Several types and styles of mouthpieces are available. Use a soft, comfortable mouthpiece that does not cause jaw fatigue. A repair technician can replace mouthpieces quickly and easily.

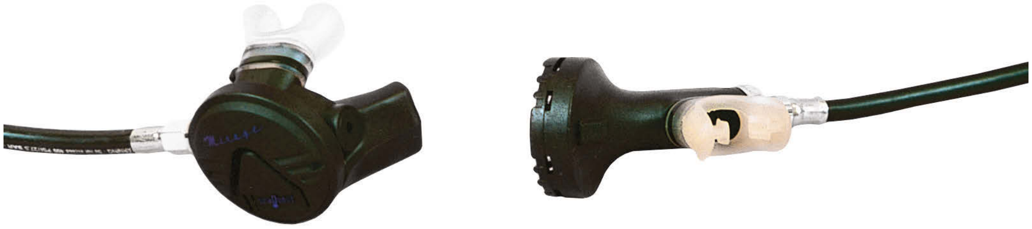

Figure 4.29 Scuba regulators have different configurations.

Regulator Accessories

A regulator hose is flexible but has rigid metal connectors crimped onto the ends. The points where the hose and the metal meet are stress points because the hose fibers strain against an unyielding surface. To prevent the breakdown of the hose fibers at the stress points, you should use hose protector sleeves. At a minimum, equip all regulator hoses with hose protectors at the first-stage end of each hose.

Padded regulator bags help protect your regulator during transportation or storage. The bag should be large enough to accommodate the regulator and all hoses without bending the hoses sharply.

Adapters allow you to use DIN-fitting regulators on standard cylinder valves. Use protective covers for the threads on DIN fittings.

Purge depressors are a built-in feature on some regulators and an accessory for regulators that do not have them. A purge depressor partially depresses the purge button to remove bias spring pressure from the second-stage valve when you are not using the regulator, thus extending the life of the valve seat.

Colored second-stage covers allow you to color-coordinate your regulator with your other equipment. Other accessories include various mouthpieces and hose adapters (as previously mentioned).

Regulator Selection

It takes effort to inhale and exhale through a scuba regulator. Effort is needed to overcome resistance; good regulators have minimal breathing resistance. Compare the performance data for various regulators, and select one that lets you breathe easily over a wide range of tank pressures, which implies a balanced first stage.

Choose a widely used and easily serviced regulator. You should get a regulator that can be serviced by facilities anywhere you happen to be and that uses readily available parts.

The type of diving you do should affect your selection. If you plan to do most of your diving from shore, you should avoid a pilot-valve second stage, which is adversely affected by sand and dirt. Diaphragm first stages are a better choice if most of your diving is from shore.

If you do not have a scuba tank, consider purchasing the tank and regulator at the same time so that you can match the fittings of the regulator and the cylinder valve. If you will be diving in water that is near freezing, you should select an environmentally shielded regulator. Remember that enriched-air nitrox and mixed gases require specially cleaned and dedicated regulators.

Regulator Care and Maintenance

Your regulator is a precision instrument; it requires care and maintenance to ensure the best possible performance. Keep sand and dirt out of your regulator. Do not allow salt crystals to form inside. Soak your regulator in clean, fresh, warm freshwater as soon as possible after diving in the ocean. You need to remove salt before it dries. If you cannot soak your regulator, you should at least rinse it. A combination of soaking and rinsing is best. Follow these rules when rinsing or soaking a regulator:

- *Keep the inside of the first stage dry.* The purpose of the dust cover on the first stage is to keep out water and dust. Develop the habit of replacing the dust cover and securing it in place with the yoke screw anytime you do not have the regulator attached to a tank. Make certain the dust cover is in place before you rinse or soak a regulator.

- *Allow low-pressure water to flow gently through the second stage and also into the openings on the first stage.* High-pressure water can force dirt or grit into crevices, where it causes damage. Gentle pressure washes the dirt away.
- *Do not press the purge button when you rinse the second stage unless the regulator is pressurized.* If you do not pressurize the regulator and you press the purge button while water is inside the second stage, you open the second-stage valve and allow water to flow through the hose into the first stage.

When the regulator has dried thoroughly, lay it flat for storage. For prolonged storage, place the regulator in a plastic bag to help protect it against the harmful effects of smog. Do not bend the hoses sharply because bending damages hose fibers. Replace hoses that are cut, bulging, or leaking.

You can avoid most problems with your regulator by having it serviced annually. Have your regulator serviced even if it seems to be functioning properly. Failure to invest in regular service can affect your safety and shorten the life of your regulator.

Alternate Air Sources

Several equipment options can help if you run out of air underwater (although running out of air is due only to sheer negligence). Your best option is an alternate air source (AAS), which is a source of compressed air other than your primary scuba regulator. An AAS is valuable if your primary source of air begins free-flowing or leaking during a dive; you can switch to the AAS and make a normal ascent. The two primary types of alternate air sources are extra second stages and backup scuba units. An extra second stage allows two divers to share air without passing a single mouthpiece back and forth. Backup scuba units are fully redundant scuba systems that provide an independent source of air in an emergency. You are not dependent on a buddy to provide air when you are equipped with a backup system. Extra second stages are less expensive than backup scuba units but do not provide the benefits of an independent scuba system.

Extra Second Stages

Two types of extra second stages are available. The first is an extra second stage for your regulator; the second is a BC low-pressure inflator that has an integrated regulator second stage (see figure 4.30). An extra second stage, or **octopus**, should meet the following criteria:

- The first stage of the regulator should be capable of meeting the airflow demands of two second stages.

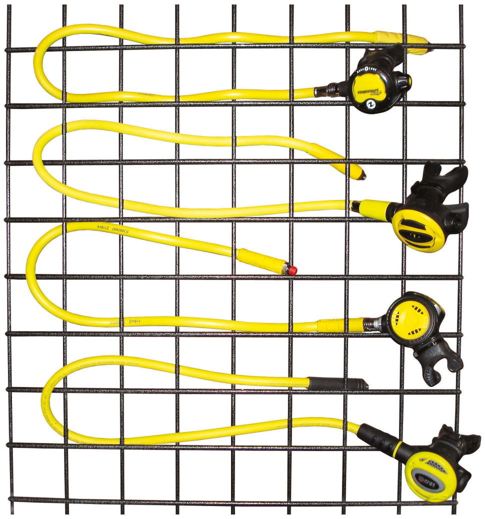

Figure 4.30 Extra second stages.

- The hose on the extra second stage should be several inches longer than the hose on the primary second stage.
- The extra second stage should attach to your chest area in such a way that your buddy can remove it quickly and easily. Do not allow the extra second stage to dangle.
- The attachment device should cover the mouthpiece opening of the extra second stage to prevent the regulator from free-flowing and to keep dirt and debris from getting inside.
- The extra second stage should be brightly colored for easy identification.

A regulator second stage may be integrated into a BC low-pressure inflator in one of two ways. The BC inflator can have a built-in regulator second stage, or quick-release hose fittings on the regulator second stage can allow it to be connected in series in the low-pressure hose leading to the inflator assembly. Both types of extra second stages have advantages and disadvantages. An integrated second stage requires one less hose on your regulator because a single hose provides air for both the extra second stage and the BC inflator. When you must share air and you have an integrated second stage, you must breathe from the integrated second stage because the hose is too short for your buddy to use. With an extra second stage, you or your buddy can use either air source. Extra second stages integrated into the BC inflator may leak air. To stop a leak, you must disconnect the low-pressure air, thereby losing the functions of the low-pressure inflator and the extra second stage.

Backup Scuba Units

Two types of backup scuba units are available. A **pony tank** is a small (about 13- to 20-cubic-foot, or 368 to 566 L) scuba cylinder with a separate, standard regulator. You clamp a pony tank to the side of your main scuba cylinder. A **Spare Air unit** is a smaller (2 to 4 cubic foot, or 57 to 113 L) scuba cylinder with a special regulator integrated directly into the valve. A pony tank provides an adequate supply of air for many situations, but a Spare Air unit provides only enough air to permit an ascent from shallow depths. On the other hand, a Spare Air unit is small and light, whereas a pony tank is bulky and heavy (see figure 4.31).

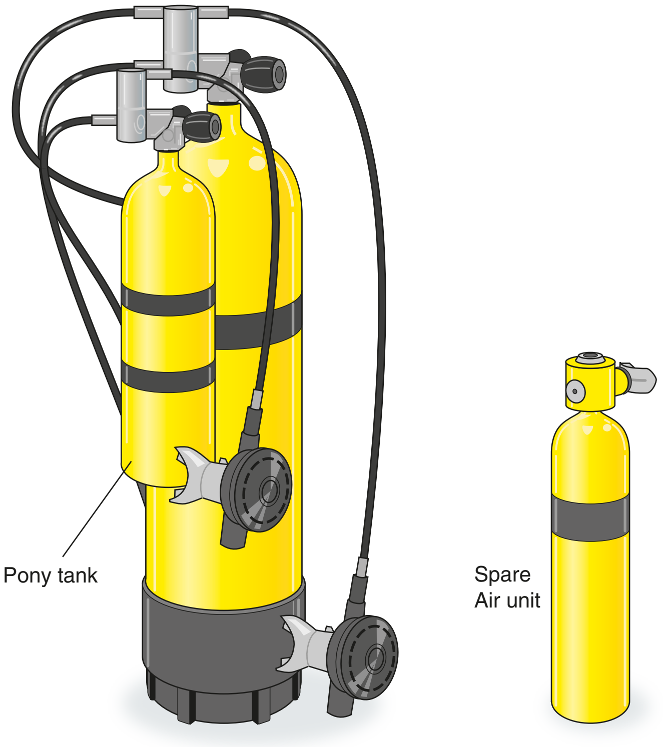

Figure 4.31 A pony tank and a Spare Air unit.

Alternate Air Source Care and Maintenance

AAS equipment should receive the same care and maintenance as your primary scuba equipment. Some divers attempt to save money by having only their primary scuba equipment serviced annually. This is false economy. AAS equipment needs servicing just as much as primary scuba equipment does.

Instrumentation

You move in three dimensions in water the way a pilot does in air. Instruments are important safety requirements for flying. A pilot needs information about altitude, direction, time, and amount of fuel remaining. Similarly, you need information about depth, direction, time, and amount of air remaining. The minimum instrumentation you need is a depth gauge, an underwater timer, a submersible pressure gauge (SPG), and a compass. A dive computer is highly recommended. All required instrumentation can be provided by a single integrated device.

Depth Gauges

To tell how deep you have descended, you can use one of four types of depth gauges: a **capillary gauge**, a **Bourdon tube**, a diaphragm gauge, or an electronic gauge. Manufacturers calibrate depth gauges in seawater. Most mechanical gauges do not indicate the correct depth in freshwater, but they indicate the equivalent seawater depth if you use them starting at sea level in freshwater or if they adjust for altitude pressure changes. The reading of gauges in freshwater is acceptable because tables for time limits at various depths are based on sea-level depths. Table 4.5 compares the four types of depth gauges.

A capillary depth gauge is a simple instrument. It is a hollow, air-filled, transparent plastic tube sealed at one end and placed around a circular dial. The open end of the tube aligns with zero on the gauge dial. A capillary gauge uses the principle of **Boyle's law**. Water pressure compresses the air inside the tube during descent. The position of the air–water interface inside the tube relative

Table 4.5 Depth Gauges

Type	Accuracy	Advantages	Disadvantages
Capillary	Accurate only to depth of about 40 ft (12 m)	Rugged, inexpensive	Can clog with debris or air bubbles
Bourdon tube	Reading accurate within 1% to 2%	Accurate	Can be damaged by reduced pressure at altitude
Diaphragm	Very accurate	Can adjust zero setting for pressure changes at altitude	Expensive
Electronic	Reading accurate within 6 in. (15 cm)	Gauge may zero itself to compensate for changes in atmospheric pressure	Must have sufficient battery power; expensive

to markings on the dial indicates the depth. At two atmospheres of pressure, the air column in the tube compresses to one half its original length.

A Bourdon tube is a thin metal tube formed into a spiral. The tube may be open to the water, or it may be closed and placed inside a housing filled with oil. Oil-filled Bourdon gauges are more popular than open-tube gauges. Pressure on an open-tube gauge begins to straighten the tube and increases the coil diameter. The straightening produces a spiral movement of the tube, which is linked mechanically to a needle to indicate the depth. Water pressure causes the coil of a closed-tube gauge to decrease in diameter. The movement of the coil, linked mechanically to a needle, indicates the amount of pressure exerted on the gauge.

The accurate but expensive diaphragm gauge uses elaborate mechanics to connect a thin, movable diaphragm to an indicating needle. Electronic depth gauges, also accurate and expensive, use a pressure sensor (transducer), electrical circuitry, a display, and a battery to indicate depth.

A maximum-depth indicator is a desirable feature for depth gauges. As you will learn in chapter 7, you must know the depth of a dive for planning purposes. A digital depth gauge retains the maximum depth you attain. The instrument displays the information until the next dive or for 12 or more hours after a dive; it then resets automatically. Many modern depth gauges with needle displays have a thin indicating wire that the gauge needle pushes along the dial face. When the needle retreats, the wire remains at the highest point reached on the dial. You can reset the indicating wire by turning a screw on the dial face. When you use this type of gauge, you must remember to reset the maximum-depth indicator before each dive.

Underwater Timers

You can use either an automatic or a manual underwater timer to keep track of time during a dive. Either type may indicate time with hands on a dial or with a digital display. Pressure activates automatic timers, which start timing at a depth of about 3 to 5 feet (1 to 1.5 m) and stop timing when the depth is less than that. Automatic timers are better than watches because you do not have to remember to start or stop the timing of your dive, although you have to reset some watch-type automatic timers before a dive. Waterproof watches that you can use as underwater timers usually feature a rotating **bezel**, a movable ring that you can set to indicate elapsed time. Digital watches are accurate, but their small buttons can make them difficult to operate. The best timers are electronic automatic timers, which can keep track of how long you dive, how long you are at the surface between dives, and how many dives you make. You do not have to remember to reset or activate anything when you use an automatic electronic timer. Dive computers track underwater time precisely. Some computers can display a timed profile of the dives that you make.

Dive Compasses

Rarely can you see more than 100 feet (30 m) underwater, so a navigational aid can be valuable. If you dive without a directional reference, you can end a dive a long distance from your planned exit point. A dive compass can help you avoid long surface swims or swims through thick surface canopies of water plants. You can use a compass to navigate beneath the canopies, where there are passages through the plants. A compass is also useful for relocating a precise area underwater and as a surface navigation device if fog arises.

Three types of compasses are card types, needle types, and electronic. Card and needle compasses are mechanical. Magnetic deposits in the earth near the North Pole attract either a magnetized disk or a magnetized needle to provide a directional reference. The compass card or needle can deviate from its correct alignment if it is in close proximity to ferrous metal or a source of magnetic forces such as a magnet or electrical motor.

The earth's magnetic field varies in different locations, and the variations affect the balance of the compass card or needle. Some manufacturers offer compasses balanced for different zones of the earth. A good compass has a good tilt potential of plus or minus 30 degrees. Be sure to ask about balancing when selecting a compass.

Diving compasses have liquid inside to dampen the swinging of the needle or disk. To be useful for diving, a compass needs a reference line, called a **lubber line**, to indicate the direction of travel relative to the needle reference. Another desirable feature is a rotating bezel with bracketing index marks that allows you to mark the needle position for a specific direction.

You view some compasses from the top and some from the side. Side-reading compasses display the course in a window on the side of the instrument. You look across a top-reading compass. You will learn how to read and use a compass in chapter 6.

Digital compasses are usually an integrated component of a dive computer (see figure 4.32). The direction of travel is displayed both graphically and numerically. A heading (desired direction) can be set and the device will indicate any correction needed to remain on course. Bearings can be stored in a memory device.

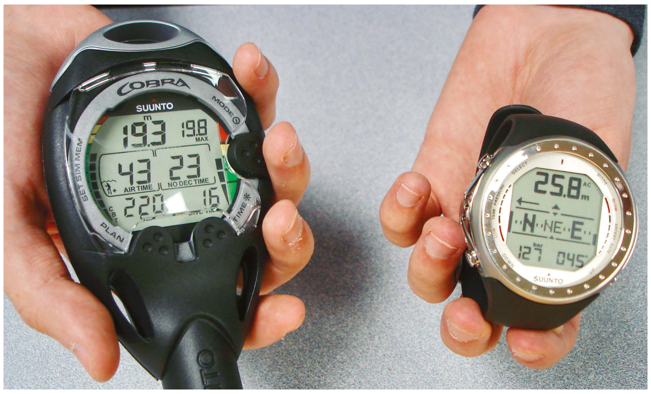

Figure 4.32 Console with computer and SPG (left) and electronic dive compass (right).

Submersible Pressure Gauges

An SPG is analogous to the gas gauge of a car and is just as essential. The SPG measures scuba cylinder pressure. You can use either a mechanical or an electronic SPG. A mechanical SPG is a high-pressure Bourdon tube. High-pressure air from the cylinder passes through the regulator first stage, through a high-pressure hose, and into a Bourdon tube inside a housing at the end of the high-pressure hose. The pressure tries to straighten the spiral tube, which moves a needle on a dial to indicate the tank pressure. Physical shock can damage a mechanical SPG.

An electronic SPG has a pressure sensor (transducer), circuitry, a battery, and a display. It is a form of a high-pressure depth gauge. The display may be digital or graphic. Either a symbol or numbers (in pounds per square inch, atmospheres, or bars) indicate the amount of air in your tank. If the electronics get wet or if the battery dies, an electronic SPG will not function. Some electronic SPGs are hoseless—they do not require an air hose between the cylinder and the gauge, which usually serves multiple functions (dive computer, compass, and SPG). You should retain at least 300 to 500 psi (20 to 34 atm) in your tank at the end of a dive. Mechanical SPGs typically have a red area on the dial for the last 500 psi. When you dive, you should monitor your air supply and make sure that you surface before the needle gets into the red area. Electronic depth gauges usually warn of a low supply of air by blinking the display.

An SPG has a **blowout plug** to relieve pressure in the housing in the event of a high-pressure leak. Identify the blowout plug on your SPG, and do not place anything over the plug that will prevent it from functioning. If the plug cannot come out to release high pressure inside the SPG housing, the face of the instrument can explode.

Instrument Consoles

You can purchase diving instruments individually or in combination. It is convenient to combine several gauges into a display unit called a **console**. An instrument console attaches to the high-pressure hose coming from your regulator first stage. When your instruments are in a console, your arms are free of gauges, and dive preparations are quicker.

The two types of instrument consoles are mechanical and electronic. A mechanical console contains an SPG and a depth gauge. Some also contain an underwater timer, a compass, and a thermometer. The instruments usually feature luminous displays, which are easy to read in low light.

With an electronic console, all instrument information is in a single display (see figure 4.32). If one gauge fails in a mechanical console, the remainder can still function; but when an electronic console fails, all the information provided by the unit is lost. Electronic displays are difficult to read in the dark unless they feature some type of illumination.

Dive Computers

A dive computer is an electronic instrument with a pressure sensor, electronic circuitry, a battery, and a display. A programmed computer inside the instrument uses pressure and time information to continuously calculate the uptake of nitrogen by various compartments that have different half-times. A half-time is the length of time required for a mathematical model (compartment) to increase or decrease its gas absorption or elimination by one-half. A compartment—which resembles, but does not duplicate body tissues—is completely saturated or desaturated in six half-times. When the absorption by any one of a computer's compartments reaches a predetermined level, the device indicates that you are approaching the time limit after which a direct ascent to the surface will no longer be possible. On reaching that time limit, the computer indicates a minimum depth—a **ceiling**—that you cannot exceed during ascent. You risk decompression sickness unless you wait until the computer indicates that sufficient outgassing has occurred to allow you to continue your ascent. A dive computer provides extremely accurate time and depth information. Other common features are a low-battery warning, a rapid-ascent warning, a dive log mode, a dive-planning mode, and information about flying after diving. Additional information about dive computers is provided in chapter 7.

Instrument Care and Maintenance

Physical abuse can damage instruments, so protect your instruments from shock. Secure your console instead of allowing it to swing freely.

Heat or prolonged exposure to hot sunlight can cause oil in a liquid-filled gauge to expand and break the seal on the housing encasing the instrument. If you break the seal, you must have the gauge repaired. Hot water in a whirlpool or a shower can cause an underwater timer to expand, break a seal, and allow water inside. Do not subject diving instruments to high temperatures.

Have air leaks in SPGs repaired at the first opportunity. Have your depth gauge tested for accuracy from time to time by a professional repair facility, or compare your gauge with an extremely accurate instrument, such as a digital depth gauge. Follow all the manufacturer's recommendations. Reduced pressure at elevations above sea level can damage some instruments. Unless an instrument is designed for use at altitude, pack it in an airtight container for flying. Soak and rinse instruments with clean, fresh water after use. Give special attention to pressure-sensing areas to prevent them from becoming clogged with dirt or salt crystals.

Electronic gauges require batteries. Some gauges must be returned to the manufacturer for battery replacement. Gauges that permit the consumer to replace batteries will flood if the batteries are not replaced properly.

Dive Knives and Accessories

This section addresses some additional required and optional equipment. Figure 4.33 shows examples of a dive knife and other dive accessories.

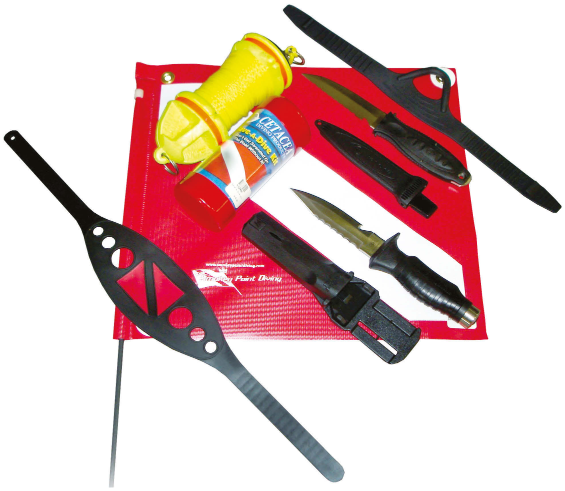

Figure 4.33 Dive knives and accessories.

Dive Knives

A dive knife is a required equipment item. Lines and cords in water can cause entanglements, so you must have a knife to cut yourself or your buddy free if necessary.

The many designs of dive knives include large knives and small knives. A small knife positioned where you can reach it easily is better than a large knife that you cannot reach. The most important feature of a dive knife is an effective cutting edge. Good blades are corrosion resistant and hold a sharp edge; serrated blades cut lines more effectively than straight blades.

Some knives are multipurpose tools designed for prying, digging, pounding, and measuring in addition to cutting. If you have a multipurpose knife as a diving tool, you should also have a small, separate dive knife.

A dive knife comes with a sheath, which has some means to lock the knife in place. Make sure the sheath lock is reliable to prevent losing a knife. You can attach a small knife to your leg, your arm, your console, or your BC. Wear a large knife on the inside of your leg and secure it with straps that stretch to compensate for exposure suit compression.

To prevent corrosion, rinse your dive knife after diving. Inspect the edge for sharpness, remove any rust, and coat the blade with grease.

Accessories

You are likely to need several small, but important, accessory items. These items include a gear bag, a dive flag and float, dive lights, dive kits, signaling devices, dive slates, and a diver's first aid kit.

Gear Bags

You need so much equipment for diving that you also need a means to carry it. Gear bags can be simple or complex. They can feature multiple compartments, padding, novel ways for carrying or moving, sealed fabric edges, and various fabrics. The best type for you depends on your needs and how much you want to invest. Be sure to get a bag with webbing handles that completely surround the bag to provide full, durable support. No matter what your budget, you need a gear bag for your equipment.

Dive Flags and Floats

In many areas, local law requires the use of a dive flag while diving. In the United States, the traditional dive flag is a red flag with a white diagonal stripe (see figure 4.34). The flag is usually vinyl, mounted on a fiberglass staff, and stiffened with

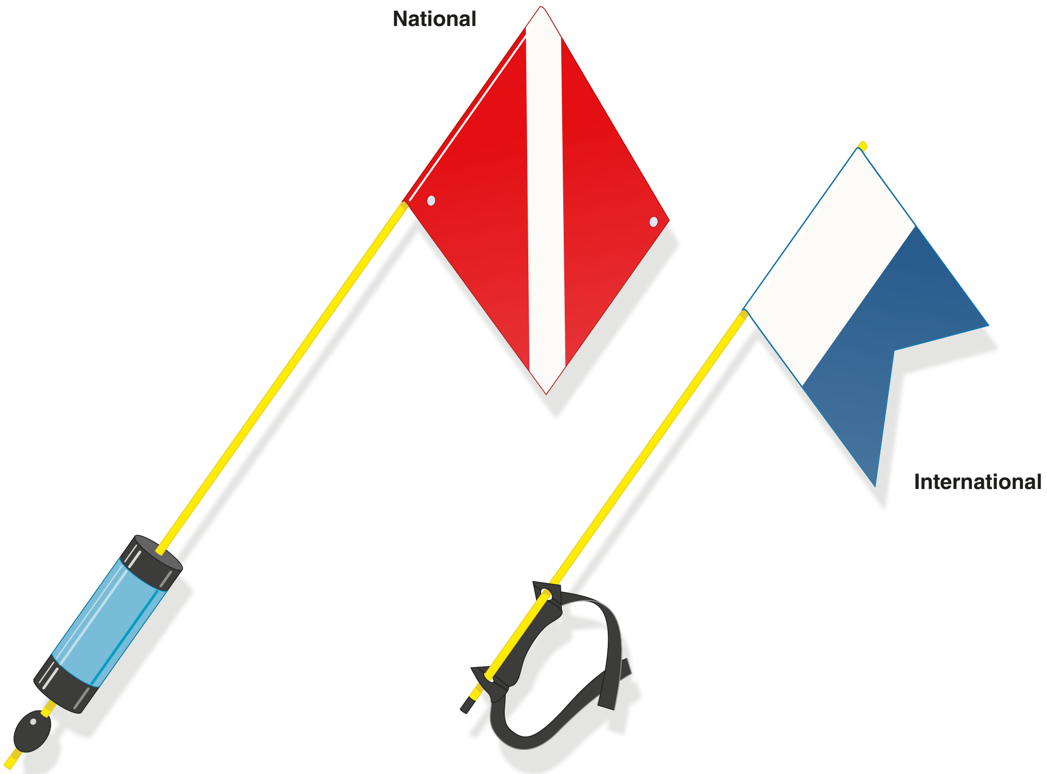

Figure 4.34 Dive flags.

a wire so that it stands out from the staff at all times. In addition to the red and white flag, you should use the international Alpha flag—a swallow-tailed blue and white flag—when diving from a vessel. The Alpha flag is a general dive flag in countries other than the United States.

Unless you are diving from a boat, you need a float to support your dive flag. Some flagstaffs have a float attached. You can also get attachments to secure a dive flag to an inner tube. Another option is a flag-holding canvas covering for an inner tube.

Dive Lights

A dive light can increase your diving enjoyment. Light levels are low underwater, and a light restores color to objects at depth and allows you to peer under ledges and into holes. You can see and enjoy much more when you use a dive light.

The many types of dive lights available include large, powerful, rechargeable lights and small lights that use disposable batteries. Consider a small dive light initially. A large light is for night diving, which is an advanced activity. A small light is easy to carry and is useful both for day dives and as a backup light for night diving. Many of today's small dive lights are bright and compact (see figure 4.35). Maintain your dive light according to the manufacturer's instructions.

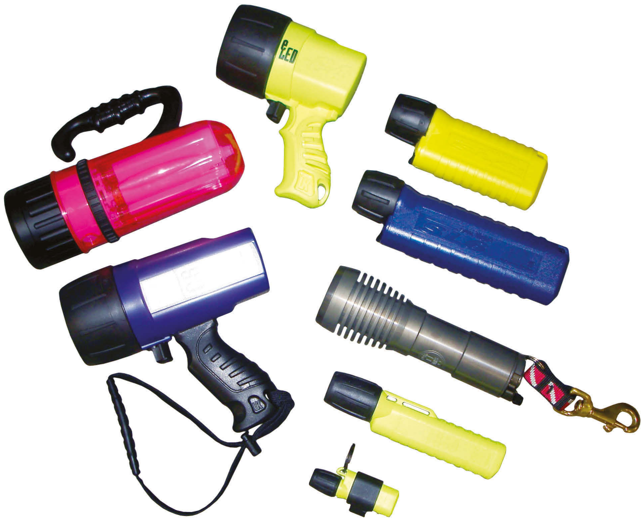

Figure 4.35 Dive lights are useful during the day or night to restore color visibility and illuminate dark places.

Kits

Two equipment kits are recommended: a dive kit and a save-a-dive kit. The dive kit contains items you use frequently for diving. The save-a-dive kit contains items you may need to salvage a dive. See the Equipment Kits sidebar for items that you might include in each kit. Because the items in your dive kit are small and get wet, you may want to keep them in a container by themselves so you can find them easily.

Once you get yourself and your equipment to a dive site, you don't want a minor equipment problem to keep you from diving. A broken strap, a missing O-ring, and a torn mouthpiece are examples of problems that can stop you from diving unless you have spare parts. Keep save-a-dive items in a waterproof box. Do not mix wet items from your dive kit with dry items in your save-a-dive kit.

Dive Slates

You need to record and refer to information around and in the water when you dive. Plastic slates are better than paper because water does not affect the slates. You can write on dive slates with a standard lead pencil or a grease pen.

Equipment Kits

Dive Kit

- Defogging solution for your mask
- Lip balm
- Wet suit shampoo
- Seasickness medication
- Sunscreen

Save-a-Dive Kit

- Mask strap
- Regulator mouthpiece
- Fin strap
- Crescent wrench
- Snorkel keeper
- Screwdrivers
- Tank valve O-rings
- Cable ties
- Quick-release buckle
- Twine
- Weight keepers
- Duct tape
- Superglue
- Extra batteries
- Heavy nylon thread and heavy-duty needles
- Neoprene cement
- Fiberglass cast tape for temporary repairs

Dive slates include checklist slates, reference slates, logbook information transfer slates, and writing slates. All types have value. You will probably have several slates when you become an experienced diver. Initially, you should have an equipment checklist, a dive-planning slate, and an underwater writing slate.

Signaling Devices

Long-range signaling devices are invaluable if you become adrift, especially in areas where currents are strong. You should have a whistle readily available. A whistle is more effective than shouting to gain attention. The sound from a whistle does not require much energy to produce, and it travels well over water.

A diver's air horn, powered by low-pressure air, works with as little as 50 psi (3.4 atm) of tank pressure and produces a sound that can be heard up to a mile away. Although this device is small, it is so loud that you must point the horn away from yourself to avoid hearing damage.

You can get long, bright, thin, inflatable safety tubes that fit easily into your BC pocket. Inflate a safety tube at the surface to make yourself easier to spot in the water.

You can use a signal mirror to flash sunlight long distances over water. Other forms of signaling devices for divers are flashlights, strobe lights, and waterproof flares. The greater the chance of being caught in a current, the greater your need for signaling devices.

Diver's First Aid Kit

Diving takes place in remote areas, and because it is a physical activity, someone may be injured. You need to be prepared for an emergency. You should have a first aid kit at the dive site. See the First Aid Kit sidebar for items that should be included in a standard first aid kit. Pack the items in a waterproof container. Chapter 6 addresses the use of the first aid items.

First Aid Kit

- Rescue breathing mask
- Tweezers
- Seasickness medication
- Bandage scissors
- Isopropyl alcohol
- Penlight
- Hydrogen peroxide
- Coins for emergency phone calls
- White vinegar
- Emergency contact information for diving accidents
- Baking soda
- Diving first aid book
- Analgesic and antiseptic ointment
- Space blanket
- Hot packs
- Pen and small notebook
- Eyewash
- Rubber band tourniquet
- Triangular bandages

Specialized Equipment

Other specialized equipment for scuba diving includes nitrox (oxygen-enriched air) equipment, mixed-gas equipment, and rebreathers. This equipment permits divers with specialty training to obtain longer dive times. The risks associated with the use of specialized equipment make training and adherence to the rules pertaining to its use absolutely essential.

Enriched-Air Nitrox (EANx)

Breathing gas that contains levels of oxygen greater than the 21 percent found in air provides benefits for divers. Higher oxygen levels reduce the amount of nitrogen absorbed, thereby reducing the problems posed by decompression illness. However, divers must know the technical requirements and dangers associated with the use of EANx. Dedicated equipment and specialty training are essential.

Filling standard cylinders with EANx is not permitted. Nitrox equipment is clearly marked to differentiate it from standard scuba equipment (see figure 4.36). Nitrox blending stations are available at many modern dive facilities. Standard blends of nitrox are EAN32 and EAN36, which contain 32 and 36 percent oxygen respectively (also called Nitrox32 and Nitrox36). Nitrogen and oxygen are blended together during the filling process, but the final mixture may not be exact. Divers need to test the oxygen percentage of the gas in their cylinder before diving. The divers use a handheld oxygen analyzer to ensure that the oxygen level is acceptable.

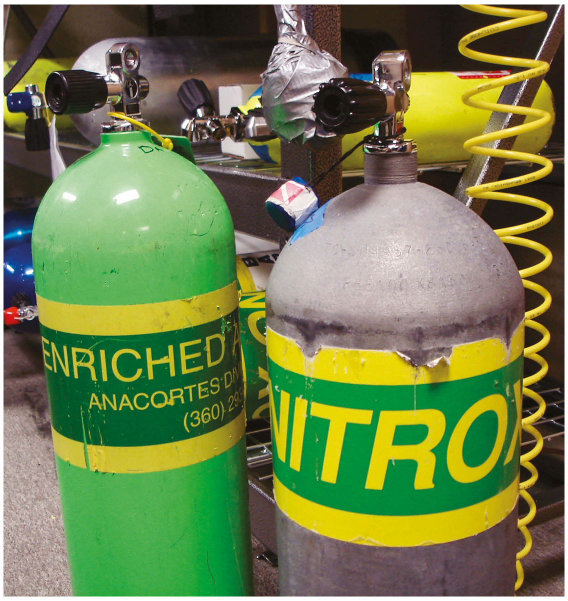

Figure 4.36 Nitrox tanks have explicit markings to differentiate them from standard cylinders.

If you breathe oxygen at a partial pressure that is too high, convulsions can occur. Partial pressures of oxygen must be kept within safe limits (1.4 to 1.6 atm or bars). For this reason, there are maximum operating depths (MODs) for nitrox diving. The MODs for EAN32 and EAN36 are 110 feet (33.5 m) and 95 feet (29 m) respectively. These are absolute depths and must not be exceeded. Additionally, divers must determine equivalent air depths (EADs) for decompression. Modern dive computers allow divers to select the type of breathing gas being used. The computers automatically calculate EADs and are recommended for mixed-gas diving.

Dive operators and gas blending facilities require proof of specialty training for the use of EANx. Training is widely available.

Technical Diving

This form of diving is even more specialized than the use of enriched-air nitrox because it involves the use of mixed gases and highly technical equipment. Breathing gases include helium, neon, nitrogen, and oxygen; these gases are mixed in various quantities depending on the depth. Obviously, divers need substantial training before they attempt this method of diving. Some gas mixtures will not sustain life until the partial pressure of the oxygen in the mix increases with depth. In this instance, technical divers must use travel tanks to reach a safe depth before switching to the low-oxygen mixture in the primary unit. The principal goal of technical diving is to allow divers to exceed recreational depth and time limits. Technical diving has greater risks than recreational diving, but many consider the rewards worth the risk.

Some technical divers use multiple large-capacity cylinders for open-circuit scuba diving. Others use complex and expensive closed-circuit rebreathers (CCRs) to obtain greatly extended dive times. Rebreathers, which use a mixture of oxygen and inert gas, allow exhaled oxygen to be breathed repeatedly until it is metabolized. Additional oxygen is added as needed to the inert gas in the system to maintain required levels. Exhaled carbon dioxide is absorbed by a canister called a scrubber. Oxygen replacement may be controlled automatically by a central processing unit or may be controlled manually by a valve. The result, which is not affected by depth, allows a diver to remain submerged for hours at a time or to make multiple dives with a single fill of breathing gas. Figure 4.37 illustrates a modern CCR.

Dive preparation and predive procedures vary greatly from open-circuit scuba diving. For example, a diver needs to prebreathe a CCR to activate the absorption process of the carbon dioxide scrubber. This procedure can take several minutes and is affected by the temperature of the absorbent. The diver also needs to check the unit's calibration at a depth of 20 feet (6 m) during descent. Changes in the gas mixture that are made during ascent must be managed by a central processing unit or managed manually by the diver. This truly is technical diving.

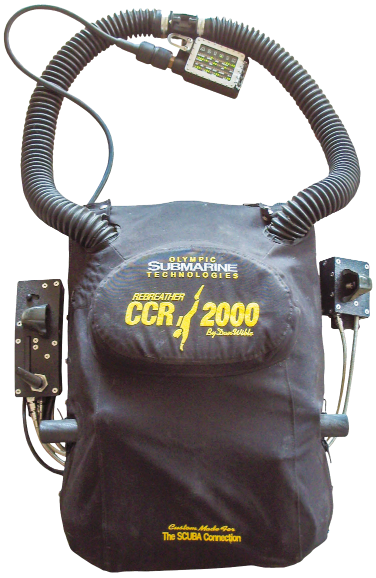
OLYMPIC SUBMARINE TECHNOLOGIES
REBREATHER
CCR 2000
By Dan Wible
Custom Made For The SCUBA Connection

Figure 4.37 **CCR 2000 rebreather.**

Courtesy of Dan Wible.

SCUBA WISE

If you want to avoid diving accidents, you should use familiar, correctly fitting, well-maintained dive equipment the way that it is intended. As I look back over several decades of dive experience, I can recall many instances when divers got into trouble because they did not follow the essential equipment safety procedures summarized in the previous sentence. Diving allows us to explore an alien environment using life-support equipment. Modern scuba equipment is manufactured well, but you must take care of it and have it serviced periodically to keep it functioning properly. Would you even consider going into outer space with life-support equipment that was not carefully maintained and serviced? We go into inner space when we dive. Our life-support equipment is just as vital to us as an astronaut's.

Equipment problems while diving do arise, but equipment failure is rarely the cause of an accident. Problems cause anxiety and stress. A diver experiencing stress from environmental factors and fear may panic from the added stress that a minor equipment problem might pose. Learning to manage yourself and your equipment helps you reduce stress, avoid panic, and prevent an accident. One of the most important rules of accident-free diving is to have good, familiar equipment that is in good condition and to use it properly.

Oxygen levels are managed by employing multiple sensors, which are accurate devices when maintained and calibrated. Current CCRs lack reliable carbon dioxide sensing devices. Carbon dioxide toxicity can be fatal, and a diver's symptoms are not a reliable indicator. This problem requires the usage time of the carbon dioxide absorbent to be monitored and the absorbent must be discarded when time limits (based on statistical tests) have been attained.

Because CCRs are expensive, must be used frequently, and require extensive maintenance, they are used by relatively few divers. But the advantage of greatly extended dive times appeals to many divers, so the popularity of rebreathers is increasing.

Summary

Diving involves a great deal of equipment. You are beginning to understand the equipment you need to have, how to select it, and how to care for it. You will learn more about equipment from your instructor, from retailers, from magazines, and from other divers. Get good equipment, and give it the best care possible. Diving is not enjoyable if you have constant equipment problems.

Application-of-Knowledge (AOK) Questions

1. Diving equipment is expensive. What are the best procedures to follow that will increase the life of your equipment?
2. What two concerns are most important when selecting diving equipment?
3. You were unaware that your regulator was free-flowing behind you at the surface at the end of a dive, and your tank lost all of its air. When you move your tank after the dive, you feel water sloshing inside. How was the water able to get into your cylinder?
4. You have weighted yourself to be weightless in water and are able to sink with your BC deflated just by exhaling. At the end of the dive you are unable to stop your ascent as you near the surface, even when you exhale completely. What is causing this undesirable change in buoyancy?
5. You are diving in a dry suit. You have added air to the suit to maintain a constant volume as you descend. You find a small anchor on the bottom and decide to salvage it. You pick up the anchor and carry it with you. To offset the weight of the anchor, you add more air to your dry suit. Why is this action a dangerous practice?
6. You and your buddy are practicing compass navigation by swimming a compass course. You place your compass next to your buddy's to compare headings and proceed along your planned course, but you missed your mark and had to surface. You find that you swam in the wrong direction. What caused this error?
7. What is the best way to remove salt, debris, or chlorine from you diving equipment?
8. A scuba regulator can have four or five hoses attached to the first stage. Can you think of the purposes of these hoses?
9. What are the best reasons for having good dive equipment and maintaining it properly?

School of barracuda in the Florida Keys

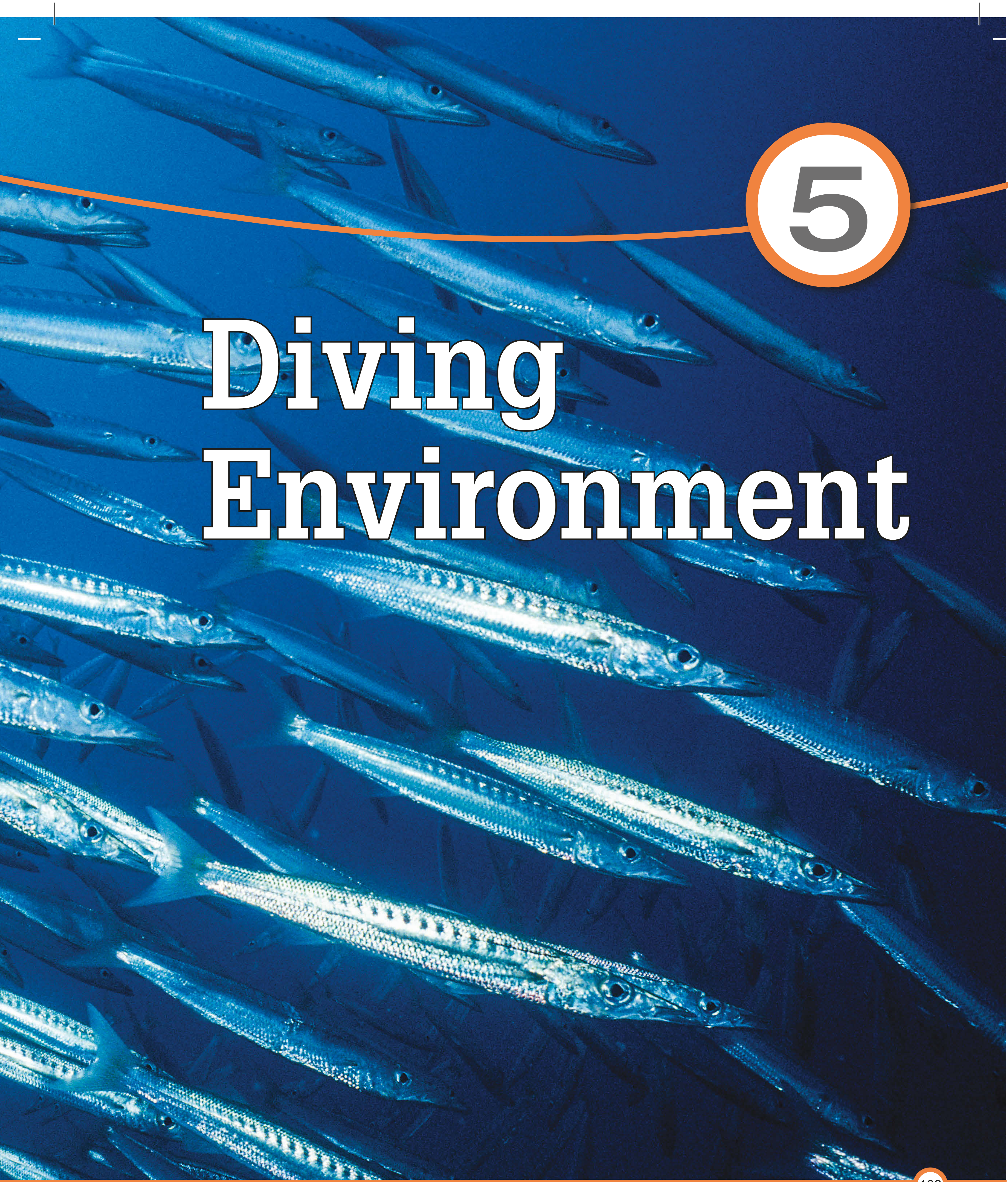

5

Diving Environment

Dive In and Discover

By the end of this chapter, you will be able to do the following:

- Describe the aquatic food cycle and explain how the process contributes to red tide.
- List five types of aquatic life injuries and three ways to avoid such injuries.
- List four types of pollutants and describe the effects they have on the underwater environment.
- List three actions that divers can take to help preserve aquatic life.
- List five ways to keep from damaging the environment while you are underwater.
- Contrast the general diving conditions for freshwater diving and saltwater diving.
- Describe the hazards of cavern diving, cave diving, and ice diving.
- Explain the cause of tides and their effects on diving activities.
- Explain the cause of waves and surf and their effects on diving activities.
- Explain the causes of currents and their effects on diving activities.
- Describe a rip current and explain how to escape from one.
- Define the terms *plankton*, *bloom*, *red tide*, *kelp*, *stipes*, *fronds*, *ciguatera*, *scombroid*, *tetrodotoxin*, *thermocline*, *overturn*, *reverse thermocline*, *upwelling*, *eddies*, *cavern*, *cave*, *sink*, *siphon*, *turbid*, *continental shelf*, *tsunami*, *stand*, *spring tide*, *neap tide*, *flood*, *ebb*, *slack water*, *fetch*, *swell*, *crest*, *trough*, *wave height*, *wavelength*, *wave period*, *wave train*, *surge*, *surf*, *gyres*, *downwelling*, *rip current*, *drift*, *set*, *drift dive*, and *trail line*.

Aquatic Biology

The plants and animals of the underwater world are wondrous and diverse. The millions of animals in the aquatic realm range from those that are microscopic to those weighing tons. To appreciate, respect, and enjoy aquatic life, you need to learn a bit about biology. This section familiarizes you with the flora and fauna of the aquatic realm. Aquatic life fits into three categories: life-forms that drift with the currents, those that swim freely and are able to move against the currents, and those that dwell on the bottom.

Drifters

The drifters are called **plankton**. Animals that drift are called zooplankton. Drifting plants are called phytoplankton. Plankton begins the food cycle in the waters (see figure 5.1). Small animals eat plankton, and larger animals eat the smaller animals. When the large animals die, their remains sink to the bottom and decompose. The decomposed material rises to the surface, where it becomes food for the plankton. Warm water and nutrients cause some plankton to multiply. Overpopulations of plankton, called **blooms**, can color the water, destroy underwater visibility, and form toxins in animals that feed by filtering water. Toxin from blooms makes clams and mussels unsafe for consumption during summer months in some areas. One type of red phytoplankton often creates blooms, known as **red tide**, in the seas. Diving conditions are poor in areas affected by plankton blooms.

Another type of aquatic plant, called algae, is an important part of the aquatic world. Plants use light to produce their own food and become food for animals. They convert water and carbon dioxide to oxygen and carbohydrates through the process of photosynthesis. Various types of algae are found in underwater areas where light is available. Most algae grow in shallow water where light is most abundant. Thick moss drapes objects in some freshwater areas. Turtle grass is a grasslike, green, tropical saltwater plant that provides a habitat for many forms of life. Some types of long, flowing algae in cold and shallow water, such as surf grass or eel grass, can cause you to trip if you try to walk in them. Slippery plants cover rocks and can cause you to slip and fall unless you move cautiously.

Giant algae, called **kelp**, produce long strands, called **stipes**, in which you can become entangled. You can learn how to avoid and how to deal with this problem. A rootlike structure, called a holdfast, anchors kelp to the bottom; numerous gas bladders, called floats, lift

Learning to dive gives you an opportunity to become more familiar with the aquatic environment, which covers over 70 percent of the surface of the earth. The underwater world is fascinating. This chapter introduces you to biological and physical conditions of the diving environment. You will learn the effect of people on the environment and the effect of the environment on people. You affect the underwater world more than you might imagine.

A juvenile octopus in California

the top parts toward the surface. What appear to be the leaves of kelp are **fronds**. Large areas of kelp are known as kelp beds, which have thick canopies that blanket the surface of the water (see figure 5.2). It is difficult to swim through a kelp canopy at the surface, but it is easy to swim between the clumps of stipes beneath the surface. Underwater navigational skills are important when diving in areas where kelp is dense. Kelp forests are popular diving areas because they contain great quantities of life.

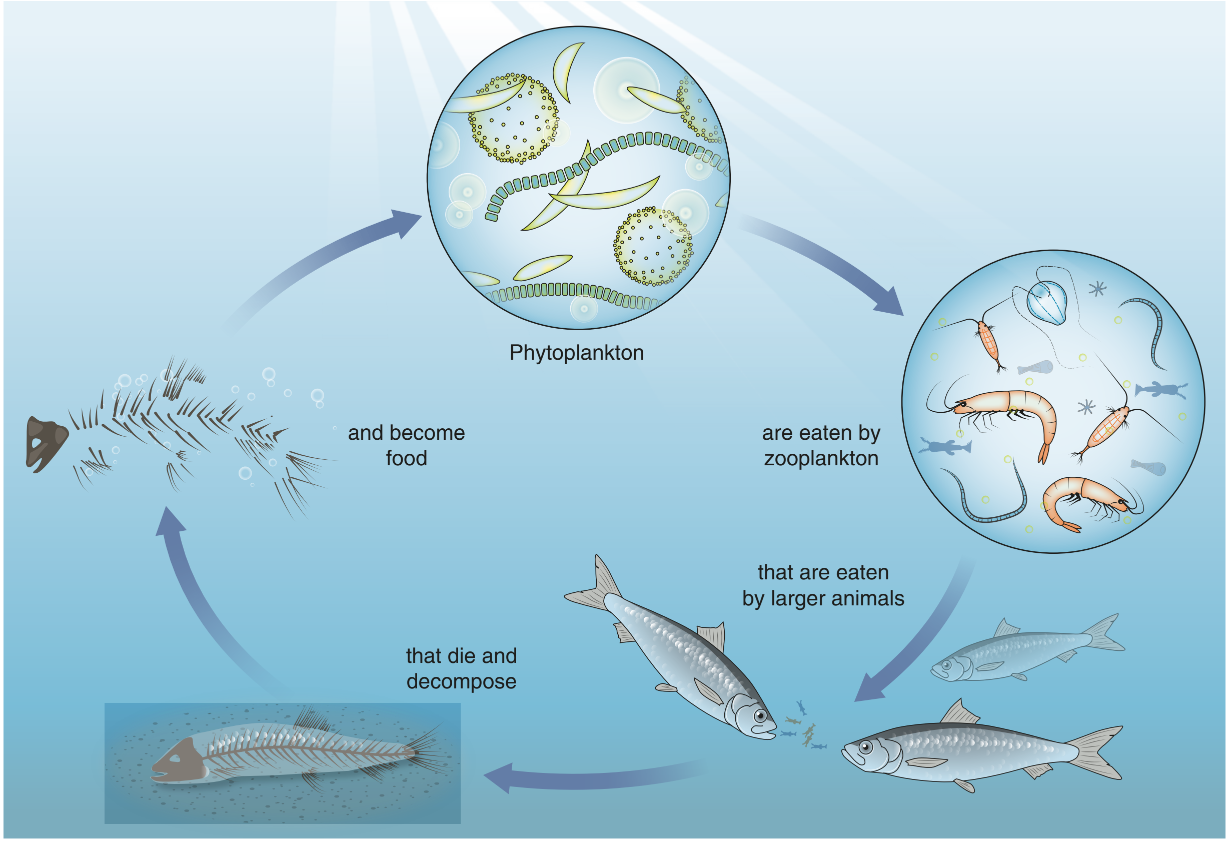

Figure 5.1 The underwater food cycle.

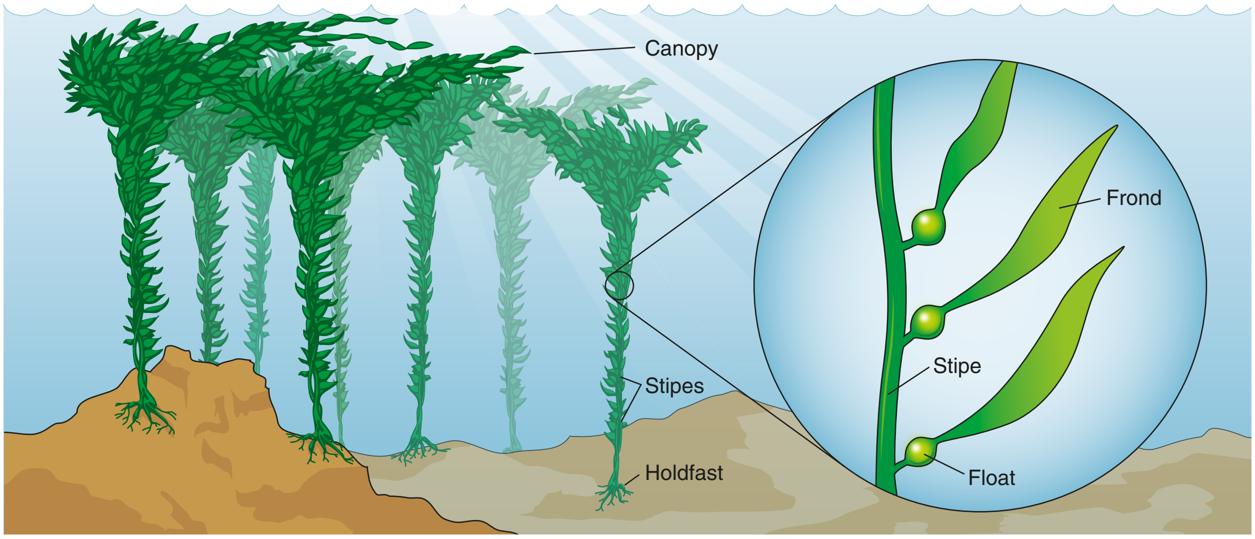

Figure 5.2 A typical kelp bed.

Swimmers

One of the rewards of diving is seeing fish. Fish can be found in nearly all the waters of the world. You cannot outswim even the slowest fish, so do not chase them. If you want to observe fish closely, you must blend into the environment. Fish will get closer to you than you could ever get to them.

Collecting for aquariums requires specialized knowledge and procedures. Most fish have an internal air bladder for buoyancy control. If a fish is taken to the surface too quickly, the air bladder expands and kills the animal. Avoid handling fish because the experience can be traumatic as well as physically damaging to the animal. And the fish's spine could injure you.

Eating some types of fish can harm you, as some fish are poisonous. Types of fish poisoning include **ciguatera**, **scombroid**, and **tetrodotoxin**. Ciguatera results from eating fish that consume a certain species of algae. Ciguatera poisoning causes gastrointestinal problems within 6 to 12 hours. Scombroid poisoning, which produces nausea and vomiting within an hour, can result if you eat fish that have not been kept chilled. Tetrodotoxin, the most serious fish poisoning, results from eating exotic fish such as blowfish or puffer fish (figure 5.3). Tetrodotoxin poisoning can cause death within minutes. Avoid eating large and unusual-looking fish. Check with local fishermen to determine which fish are safe for consumption.

Many large mammals—sea lions, seals, dolphins, whales, and manatees—inhabit the water. They are graceful, beautiful, and sometimes awesome in

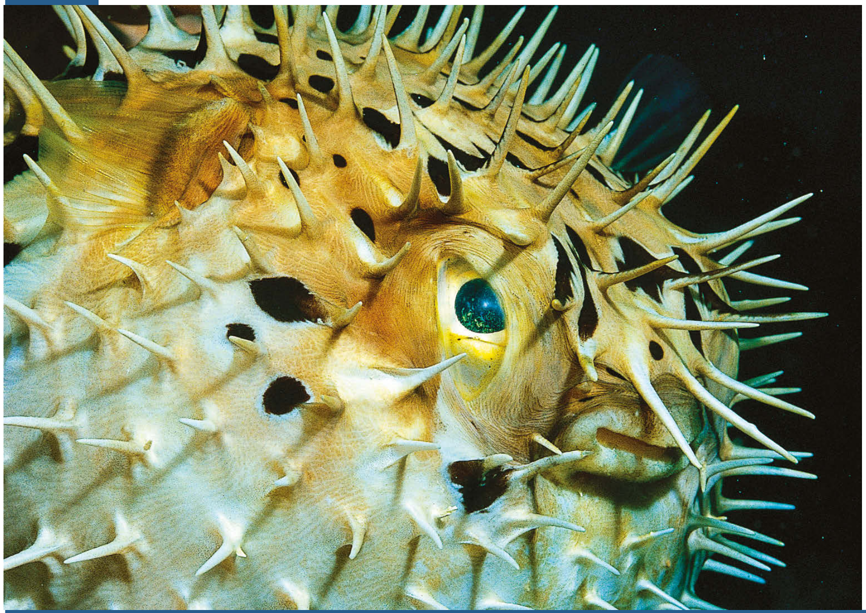

Figure 5.3 Eating improperly prepared puffer fish can result in potentially fatal poisoning.

appearance. Viewing them is an exciting experience. Some are curious and may approach you. Aquatic mammals will usually not harm you in the water if you leave them alone. But sea lions and seals are defensive on land and might bite if you get too close.

Bottom Dwellers

Bottom dwellers include animals that are stationary, such as coral and sea fans, as well as animals that move about, such as crabs and lobsters. Living, stationary bottom dwellers are usually not included in fish and game regulations, and you should not take them. Do not take coral, sea fans, and animals that appear stationary, such as starfish. If you hunt crabs and lobsters for food, you should know how to determine the gender of the animals, how to catch them without harming them, and how to measure them for minimum size. Do not take females, particularly those bearing eggs. Some divers take only one claw from a crab to conserve the species. A crab can feed and defend itself with one claw and can regenerate a new claw to replace the one taken.

Hydroids

If you dive in saltwater, you should know about hydroids, a category of animal that includes bottom dwellers, such as coral, and swimmers, such as jellyfish (see figure 5.4). Some hydroids, such as the beautiful sea anemone, have a round, columnlike body with a mouth surrounded by tentacles. This is a polyp form of hydroid. Another type, called the colonial form, can assume many different

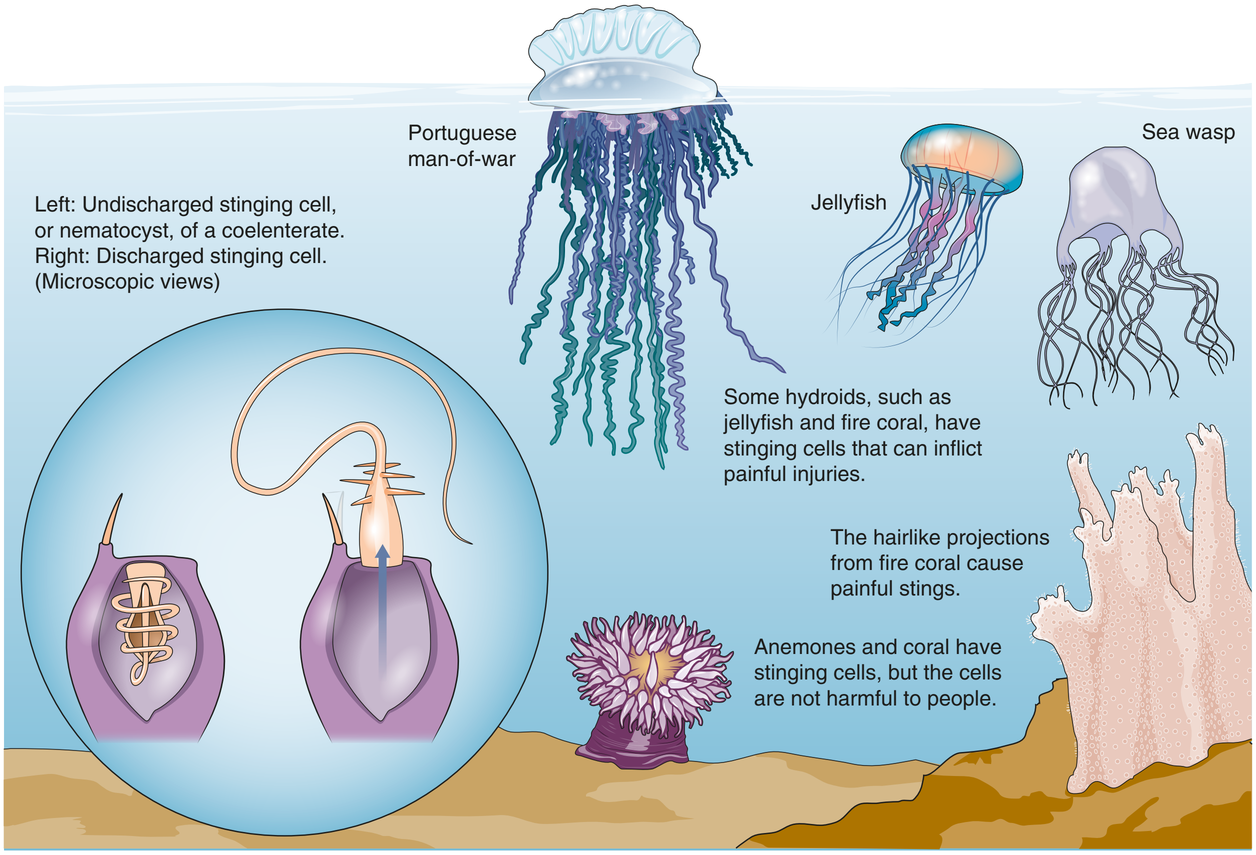
Portuguese man-of-war
Sea wasp
Jellyfish
Left: Undischarged stinging cell, or nematocyst, of a coelenterate. Right: Discharged stinging cell. (Microscopic views)
Some hydroids, such as jellyfish and fire coral, have stinging cells that can inflict painful injuries.
The hairlike projections from fire coral cause painful stings.
Anemones and coral have stinging cells, but the cells are not harmful to people.

Figure 5.4 Several types of hydroids have stinging cells.

shapes. Colonies of hydroids can encrust a surface, and groups of colonies form jellyfish. Learn to recognize and avoid hydroids that can injure you.

Potentially Dangerous Aquatic Animals

Aquatic animals use a variety of mechanisms to obtain food and to defend themselves from attack. You can minimize your chances of injury by being familiar with the ways in which animals can inflict injury. Aquatic animals are rarely aggressive toward humans; they flee, hide, or stand still as you approach. If you do not touch, threaten, or provoke an animal, the animal is unlikely to injure you intentionally. But remember, aquatic animals are wild animals (figure 5.5). If you feed them, and they bite you in the process, you should not blame the animals. Table 5.1 lists types of injuries that you could sustain from aquatic animals and recommended first aid.

Animals That Cause Abrasions or Cuts

Many animals, such as some types of coral and barnacles, have sharp, hard structures that easily cut flesh. Such cuts can be painful, are slow to heal, and can get infected. Avoid contact with reefs and rocks covered with sharp animals. Some fish have knifelike protrusions on their gill plates or at the base of their tails. They defend themselves by flailing rapidly back and forth and slashing anything near them.

Animals That Puncture or Lacerate

Sea urchins are the porcupines of the seas. Cold-water urchins have short, thick spines, while warm-water urchins have spines that are long and thin. The spines of all urchins can penetrate an exposure suit if you bump into them. The spines break off in your flesh and produce a painful, red, swollen wound. The spines can be difficult to remove, so if you are unable to avoid a sea urchin injury, see a physician to have the spines removed.

Some fish have a row of long, sharp spines along their back. The scorpionfish has hollow dorsal spines that have venom-filled sacs at the base. If you are punctured by the spines and compress the venom sacs, you inject yourself with a toxin. The toxin of some species, such as the stonefish, can cause serious symp-

Figure 5.5 Potentially dangerous aquatic animals: (*a*) White tip shark; (*b*) cone snail; (*c*) sea snake; and (*d*) jellyfish with jacks.

Scott Frier: www.wildlifephotoworkshops.com

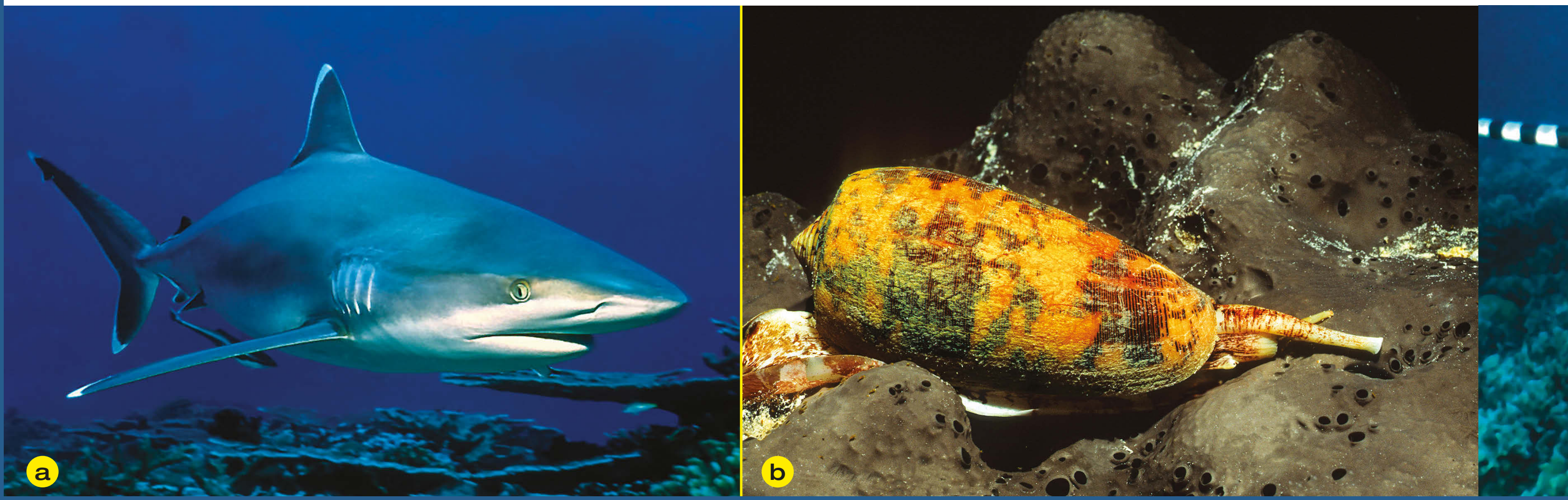

toms. Some fish, such as the lionfish (also called the turkey fish or zebra fish), have spines in other fins. The toxin of a lionfish produces serious symptoms in humans. Freshwater catfish also have venomous spines.

A shellfish called the cone shell uses venom to kill animals for food. The venom is highly toxic, and these animals can inject their venom into humans. Do not handle conical shells in tropical waters.

Rays are bottom dwellers that are round and flat. Some rays have a sharp, serrated barb at the base of the tail. The rays blend into the bottom. When someone disturbs a ray, it defends itself by arching its back and thrusting its barb into the attacker. A sheath, which often remains in the laceration wound, covers the barb and contains a toxin. Stingray wounds in the ankles of divers, waders, and swimmers are common in some areas. Avoid stingray wounds by shuffling your feet along the bottom as you walk. The torpedo ray, an animal found on the West Coast of the United States, can generate electricity. This ray can stun a diver, so you need to be able to recognize it. Avoid contact if you encounter one.

Animals That Bite

Use discretion to avoid being bitten by aquatic animals. One large freshwater fish that can bite is the gar. Turtles can inflict serious wounds. Muskrats may attack in defense, and moray eels in the ocean can deliver a serious bite. Alligators also have the potential to inflict an injury but have not been known to hurt divers.

Sharks can bite, but attacks on scuba divers are nearly nonexistent. Hollywood has exaggerated the danger of sharks; only a few species of sharks are aggressive, and it is rare to encounter these types in the waters visited by divers. Most divers would be delighted to see a shark because these animals usually retreat from an area frequented by scuba divers.

Retreat from snakes if you encounter them. Do not handle snakes; they can be extremely venomous. A sea snake's bite can produce a life-threatening emergency. Some freshwater snakes, such as the cottonmouth, are also venomous. The blue-ringed octopus, while only the size of a golf ball, is one of the most toxic creatures on earth. Don't allow curiosity to endanger your life.

Animals That Sting

Many aquatic animals can sting. Learn to identify and avoid contact with jellyfish, featherlike or whiplike hydroids, certain worms, and even some sponges.

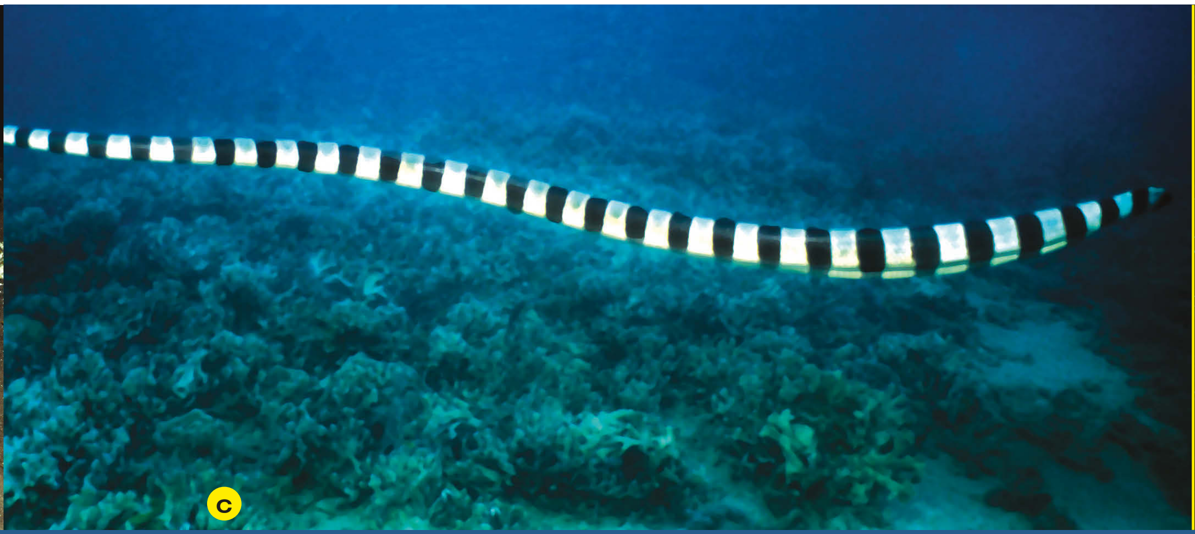

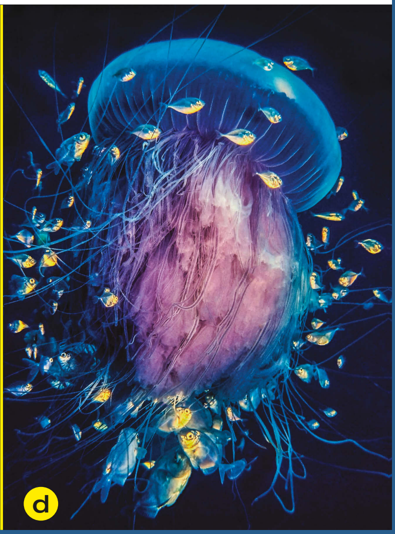

Table 5.1 Potentially Dangerous Aquatic Animals

Injury	First aid
Abrasions or cuts	
Barnacle	Scrub and disinfect the wound.
Coral	
Punctures or Lacerations	
Cone shell	Soak the injured area in hot water. For a cone shell sting, apply a venous tourniquet and immobilize the injured area. It is advisable to seek medical attention for injuries caused by the spines of the scorpionfish family.
Sea urchin	
Venomous fish (e.g., scorpionfish, lionfish)	
Stingray	
Bites	
Barracuda	Stop the bleeding; clean and disinfect the wound. For a blue-ringed octopus bite, apply pressure to the wound and seek immediate medical attention—it may be necessary to administer rescue breathing. For a sea snake bite, apply pressure, immobilize the area, and get prompt medical attention.
Blue-ringed octopus	
Moray eel	
Sea snake	
Shark	
Turtle, muskrat, alligator	
Stings	
Bristle worm	Soak the injured area in vinegar.
Fire coral	
Fire sponge	
Jellyfish	

Seek medical attention as needed for any injury.

Hydroids have tiny stinging cells, called nematocysts, that they use to kill food. The sting of some animals, such as encrusting fire coral, is merely annoying; but the sting of others, such as the Portuguese man-of-war or the box jellyfish, can cause a medical emergency. A tropical marine worm, called a bristle worm, has tufted, silky bristles along the sides of its body. The bristles, which are fine and brittle, penetrate the skin easily, are difficult to remove, and cause a burning sensation. Do not touch or handle these worms. You should wear an exposure suit at all times while diving in the ocean. The suit provides protection from stings, but be careful when you remove it. The stinging cells of marine animals remain active even if they are not on the animal. Parts of jellyfish and other animals that are on your equipment can sting if they come into contact with your skin when you remove the equipment. Avoid contact with any unusual debris clinging to your equipment after diving in salt water because it could be part of a jellyfish. Some jellyfish come to the surface at night in tropical waters. Dangerous stinging animals are seasonal in some regions. Check with local divers to find out what to avoid and when to be on the alert.

Avoiding Danger

Many aquatic animals are potentially dangerous. Tropical waters have the most dangerous animals. Serious injuries to divers are not common because divers avoid animals that can hurt them. You should respect animals, but do not be overly concerned about them. Do not panic or flee just because you see a potentially dangerous animal; simply avoid contact with it. Learn to recognize the dangerous animals in an area, know where to look for them, be alert for them, and keep clear of them. Move slowly and look carefully. Shuffle your feet when wading. Don't touch anything unless you know it is safe to touch it. Neutral buoyancy—the ability to hover—is an excellent defense against aquatic injuries.

The biology of the underwater world is so interesting that many divers study it as a hobby. It is fun to learn about animals and plants and then observe them in their natural habitats. Some divers, fascinated with aquatic life, pursue biology as a career.

Conservation and Preservation

Life in the waters is beautiful and precious, but pollution from people jeopardizes life in the lakes and seas. Divers are physically ruining beautiful reefs. Unless we take action immediately, there is a real danger that many areas of the underwater world will become barren and lifeless. We must preserve and conserve the resources of our waters.

Because the waters are so vast, people often take them for granted. Lakes and seas seem too big to be harmed, but that is not the case. The underwater environment is fragile; the balance of nature can be upset more easily than many people realize. We think of the seas as being great and powerful because we see big waves and pounding surf. People who do not see beneath the surface may not realize the delicate nature of the animals that inhabit the waters. Some living things in the seas grow slowly, only a fraction of an inch (or centimeter) per year.

As a diver, you can help reduce damage to the underwater world. Every diver should take two actions: (1) Do everything possible personally to preserve the diving environment, and (2) do everything possible to educate others and help them learn to preserve the aquatic environment. You will get closer to the aquatic environment than most of your friends. You will see firsthand the effects of pollution, litter, and exploitation. You will also see the beauty and abundance that are possible when the environment is clean and unmolested. Your influence in society can make a difference. If you do not become part of the solution to problems in the aquatic environment, you are part of the problem.

Pollution

Stemming pollution is one of the greatest challenges facing the world. Humans are incredible polluters of the environment; we have polluted the air, the land, and the waters. People dump billions of gallons and pounds of waste into water every day. People seem to think that when something is out of sight, it is no longer a problem. This is not true of pollutants. Sewage, industrial waste, garbage, and sediment have killed and are continuing to kill many underwater environments.

If we stopped all pollution today, the aquatic environment would continue to suffer for decades from the waste materials that are already in the water.

Runoff from land that enters rivers and streams and flows into the ocean also causes pollution. Chemicals used in agriculture, on lawns, and in gardens cause death and destruction in the aquatic environment. Sediment from construction and drilling finds its way into water. The sediment blocks out life-giving sunlight and smothers bottom-dwelling creatures.

Lakes and oceans have been viewed as bottomless toilets for waste disposal. There are two problems with this narrow view: (1) Animals and plants live in the water and are killed by the pollution; (2) the water cannot be replaced because large bodies of water cannot be flushed.

Diver Impact

As a diver, you can harm the environment in several ways. You can remove living things from the environment, you can smash and kill living things while moving about underwater, and you can stir up clouds of silt on the bottom. The silt can choke and kill some organisms. Good intentions can also cause problems; handling and feeding animals can kill them.

You can be an effective predator underwater. Many animals are available for you to hunt and take. A few callous individuals kill things for sport, but responsible people take only what they will eat. Although the impact of divers is of little significance compared with commercial fishing, you should keep in mind that you do have an effect. If you spearfish on a reef, the fish in the area will soon be unapproachable by divers. If you want to be a predator, you must do so in a responsible manner. Conservation and preservation of aquatic resources are important topics that are discussed later in this section.

It is difficult to resist touching animals underwater, but you should refrain until you know what animals may be touched and how to touch them without harming them. Many animals are delicate; rough handling will kill them. Predators will eat animals that a diver removes from a protected area for viewing. Some animals, including fish, have a protective coating of mucus. If the coating is removed by handling, the animal can develop an infection and die. The stress of being handled by a gigantic, bubble-blowing monster may be more than some aquatic animals can take. Sea turtles may abort their eggs after being harassed by a playful diver. Do not be guilty of killing things for your amusement. After

SCUBA WISE

There are similarities between the environments above and below water. Temperature, weather, plants, and animals change when you travel from region to region on land. We enjoy the variety of mountain, plain, forest, and desert regions. Similarly, the world beneath the water varies from region to region. The enjoyment that the various underwater environments provide is a major reason why dive travel is so popular.

When you go to places on land that are rugged, remote, and hazardous, you need special training and equipment. There are dangers everywhere you go. Careful planning and preparation help minimize risks and are essential. A guide makes exploration safer and more enjoyable. When you go diving in places that are rugged, remote, and hazardous, the same requirements apply. Enjoy the wonderful differences beneath the waters—coral reefs, vertical walls, wrecks, and so forth—but always be properly prepared and equipped and always obtain an orientation to a new area.

you learn how, you can get extremely close to animals underwater. You can interact with them and enjoy them without handling them.

Feeding animals underwater was a popular activity until environmentalists showed that this is harmful (not to mention dangerous). Several potential problems are associated with feeding aquatic animals. Unnatural food given to animals can interfere with their digestion. The animals may become dependent on the food fed to them by divers and may not be able to forage if the food supply is discontinued. Food can make animals overcome their natural fear of divers. When an animal that would normally take cover at the sight of a diver becomes accustomed to divers providing food, the animal will readily approach a diver who is a hunter.

A diver underwater has the potential to be a bull in a china shop. Overweighted divers plow along the bottom with their fins pointed downward, stirring up great clouds of silt. Buoyancy difficulties cause divers to hold on to and crash into reefs and other living things. Divers who rest on the bottom crush plants and animals without realizing the damage. Divers who swim too close to a reef often kick animals to death.

Prevention of Diver Impact

One of the most important reasons for developing good buoyancy control skills is to prevent damage to the underwater environment. An environmentally responsible diver is properly weighted and in control of buoyancy at all times. Make your diving no-contact diving. You should be able to hover above a reef, move your mask within inches of an animal, and view it without touching anything but water and without stirring up silt. Learn to use your hands to scull into position while your fins remain still. Sculling is positioning achieved with short movements of the hands (not the arms). Buoyancy control and sculling are excellent skills for reducing diver impact. Figure 5.6 shows a diver sculling.

Another way to protect the environment is to keep your equipment secured close to you so that it does not dangle and drag. Equipment that drags along the bottom for an entire dive can do a lot of damage. Streamline your equipment. Moving slowly underwater conserves energy and air and makes you less likely to come into contact with animals. This helps you avoid injuring both yourself and animals.

If you must hold on to something underwater or push yourself away from something, you should look before you touch, and you should avoid touching anything that is living. If you must settle onto the bottom for some reason, select an area where there is no visible life. If you are weightless, keep in mind that one finger may be all you need to provide the leverage to move. Leave no evidence that you ever visited an underwater area.

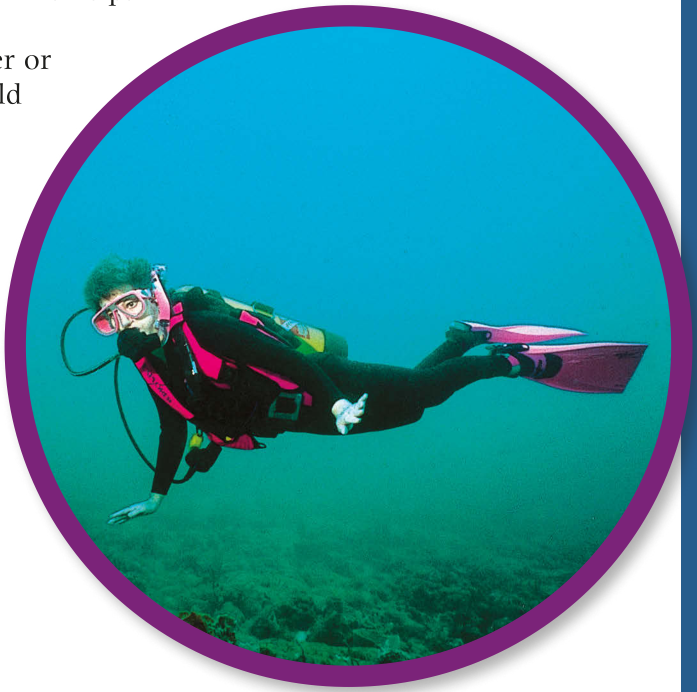

Figure 5.6 Sculling helps minimize the diver's impact on the aquatic environment.

Conservation

Many animals that were once plentiful are nearly extinct because of a lack of conservation. Buffalo and passenger pigeons are good examples of this happening on land. In some areas, this is happening to aquatic animals. Bait fish that once swarmed in enormous schools in some areas no longer exist. With no food to eat, larger predatory fish no longer frequent the areas.

Fish and game regulations are designed to conserve natural resources. Rules regarding sizes, seasons, limits, and the means by which game may be taken have been established to help ensure an ongoing supply of a resource. Obey fish and game regulations, and encourage others to abide by them as well. In most areas you need a license to harvest fish and seafood. There are written regulations you need to be familiar with. Size, sex, limit, and season laws can impose a stiff fine if you fail to adhere to the regulations. Check with your local Department of Fish and Game. The rules can benefit everyone in the long run.

If you take living things from the water, do it in a responsible manner. Avoid taking animals from areas that are popular dive sites. Limit hunting and collecting to remote areas where the impact of divers is much less. Take only what you need, not what you can get or what you are allowed. The two types of hunters most harmful to the environment are the quantity hunter and the trophy hunter. The quantity hunter seeks to take as much and as many of everything as possible to build an image as a mighty hunter. The trophy hunter seeks the largest animals, which destroys the fittest of the breeding stock. If you kill an animal, you have a duty to know how large one must be to breed and what the maximum size of the animal is. You should take animals that have had the opportunity to reproduce, but do not take the largest ones. It is difficult at times to be selective, but you should attempt to conserve life.

Nature left undisturbed maintains a balance of life. Animals are both predator and prey. An animal eats another animal and is in turn eaten by others. If there are too many predators, their population diminishes because of an inadequate food supply. If there is a temporary overpopulation of prey, the number of predators increases. People upset nature's balance. We are the most reckless predators of all. We disturb the food cycle with pollution, hunting, fishing, boating, and every other way that we affect the aquatic environment. It takes nature longer to recover from our impact than from any natural disaster. People must lessen their effect on the environment to reduce interference with the natural management of the life within the waters.

Preservation

Preservation is everyone's business, but it is more your business now than it was before you became a diver. You probably know more than your friends and neighbors about the aquatic environment. You must be an ambassador for preservation. You must educate and motivate people to help preserve the diving environment.

What people do above water affects life beneath the surface. What people put down their drains and toilets winds up in the aquatic environment, as do the chemicals they use on their lawns and gardens. Litter in and around the water kills animals, birds, fish, and shellfish Think twice about the products you buy, how you use what you purchase, and how you dispose of your wastes. Where will the toxic chemicals and waste you use end up? Be environmentally conscientious; then teach others to be so, too. A simple act, such as the use of detergents without phosphates, can make a difference. Phosphates are powerful nutrients that upset the balance of nature when dumped into the aquatic environment.

Be involved in your community. Be concerned about issues such as sewage treatment, toxic wastes, and construction, which can be extremely harmful. Waste products from manufacturing are also harmful. Help others understand the seriousness of pollution.

Become more informed and keep informed. Join groups that are working to preserve the environment. Organizations such as the Ocean Conservancy, the Reef Environmental Education Foundation (REEF), and the Oceanic Society merit your membership. These groups provide up-to-date information and details on how you can help. Various organizations sponsor underwater cleanups from time to time; these are enjoyable and worthwhile. (For a list of environmental organizations you can join, see Appendix A.)

There are actions you can and should take when you go diving. Dispose of trash properly, and encourage others to do the same. If you have a boat for diving, anchor it away from reefs to avoid reef damage caused by the anchor and chain. Retrieve trash you find in the environment, especially plastic. Collect plastic, monofilament, lead, and stainless steel leaders. Not all trash is bad. Bottles and cans provide homes for animals. Report unlawful dumping and lost or discarded fishing nets and traps, which continue to catch and kill after abandonment.

Aquatic Conditions

The particular state of the environment and the physical situations in which divers find themselves are referred to as the aquatic conditions. The conditions of concern include the temperature, visibility, and the degree of movement of the water. You need to be familiar with aquatic conditions in general and local aquatic conditions specifically.

The Big Picture

Many factors affect the seas. The sun bears down on the earth at the equator more directly than on other parts of the earth. The climate and waters near the equator are warm; the greater the distance from the equator, the colder the water. The difference in water temperatures in different parts of the world causes the movement of air and water by convection currents. Winds move from areas of high pressure to areas of low pressure. Winds and the turning of the earth move water. On a global scale, water currents move in a clockwise direction in the northern hemisphere of the earth and in a counterclockwise direction in the southern hemisphere. Weather generally moves from west to east. The gravitational attraction between the earth and other planetary bodies produces changes in the water level called tides. Storms at sea produce energy in the form of waves that travel thousands of miles before giving up their energy in the form of surf.

Seasonal changes, winds, and storms affect inland bodies of water. Water from snow and rain in the mountains and hills flows into streams, rivers, and lakes. Water seeping into the ground resurfaces in quarries and springs.

Weather, seasons, geography, and other factors affect both surface and underwater diving conditions. The water can be rough at the surface but calm at depth. There may be a current at the surface but none on the bottom. The visibility may be good at the surface but poor on the bottom, or vice-versa. The temperature is usually warmer at the surface than it is at depth. You should become familiar with both surface and underwater conditions and their effect on your diving.

Perhaps the most important fact to remember about diving conditions is that they vary. Thus, an environmental orientation is important when you dive in a new region. The diving conditions usually dictate the way you dive. What works well in one area may be totally ineffective for another. An understanding of the effects of diving conditions in an area is important. You should know what conditions to expect, how the conditions affect your approach to diving, and how to manage the effects of the conditions. For example, in one area, it may be safe to wade into the water and put on your fins; however, in another area, you may need to put your fins on before you wade in.

General Freshwater Diving Conditions

Freshwater is 2.5 percent less dense than saltwater, so you are less buoyant in freshwater than in saltwater. The density of water varies with temperature. Freshwater is most dense at a temperature of 39.2 °F (4 °C)—the usual temperature and density of the water at depths greater than 60 feet (18 m) in freshwater lakes and quarries. The water becomes lighter when it is either warmer or cooler than this. Freshwater often forms layers: a layer of warmer, lighter water on top of a layer of colder, denser water. The change from the warm water to the colder water, called a **thermocline**, is abrupt. When the water is calm, the thermocline appears to have wisps of smoke on top of it when viewed from above. The refraction of light is different when the density of the water is different, so there is a slight visual blurring at the interface of a thermocline. When you dive in freshwater, you must insulate yourself for the water temperature below the thermocline. Although the surface of a lake may be warm and sunny, the water at the bottom may be close to freezing.

The best season for diving in a freshwater lake or quarry depends on the body of water. Spring and fall are often good times because the water temperature is the same from surface to bottom, there is oxygen for fish at all depths, populations of plankton are low, and visibility is usually good.

In later spring, the sun warms the surface water, and the lack of wind keeps the surface water from mixing with the colder water at depth. The water stratifies, and a thermocline forms that remains until fall. The layer of water below the thermocline stagnates in most lakes during the summer. Decaying matter depletes oxygen and creates toxins. Fish seek refuge above the thermocline. Sunlight and warm water often lead to plankton blooms, which are aggravated by pollution.

In the fall, the surface water in lakes and quarries cools until it equals the temperature of the water at depth. Colder water and reduced sunlight stifle plankton growth, and visibility improves. Winds cause water circulation. The movement of water from the top to a depth of about 60 feet, called an **overturn**, carries oxygenated water to the bottom and leads to what is called spring diving conditions. Visibility tends to be poor during an overturn.

In the winter, the surface water in a lake or quarry is colder than the water at the bottom. A **reverse thermocline** exists and remains until the temperature of the surface water warms to that of the water at the bottom. This allows the spring overturn to begin. Water becomes 10 percent lighter when it freezes. If ice sank when it froze, bodies of water could become solid ice from top to bottom.

Ice that forms over water insulates the water beneath it. Figure 5.7 shows the annual cycle of freshwater lakes and the concept of thermoclines.

When a strong wind blows along a shore for a sustained period over a body of water, the surface water is pushed away from shore. This water is replaced by colder water flowing up from the depths. This is called an **upwelling**, which can also occur in the ocean. If the wind conditions persist, the water temperature can become constant throughout the water column, even in the summer. An upwelling carries nutrients from the depths into shallow water and brings animals into the area. After an upwelling, a bloom can occur because of the increase in nutrients in warm, shallow water.

Specific Freshwater Diving Conditions

People can have good diving experiences in many kinds of freshwater environments. Springs, low-altitude lakes, and quarries are good sites for divers who do

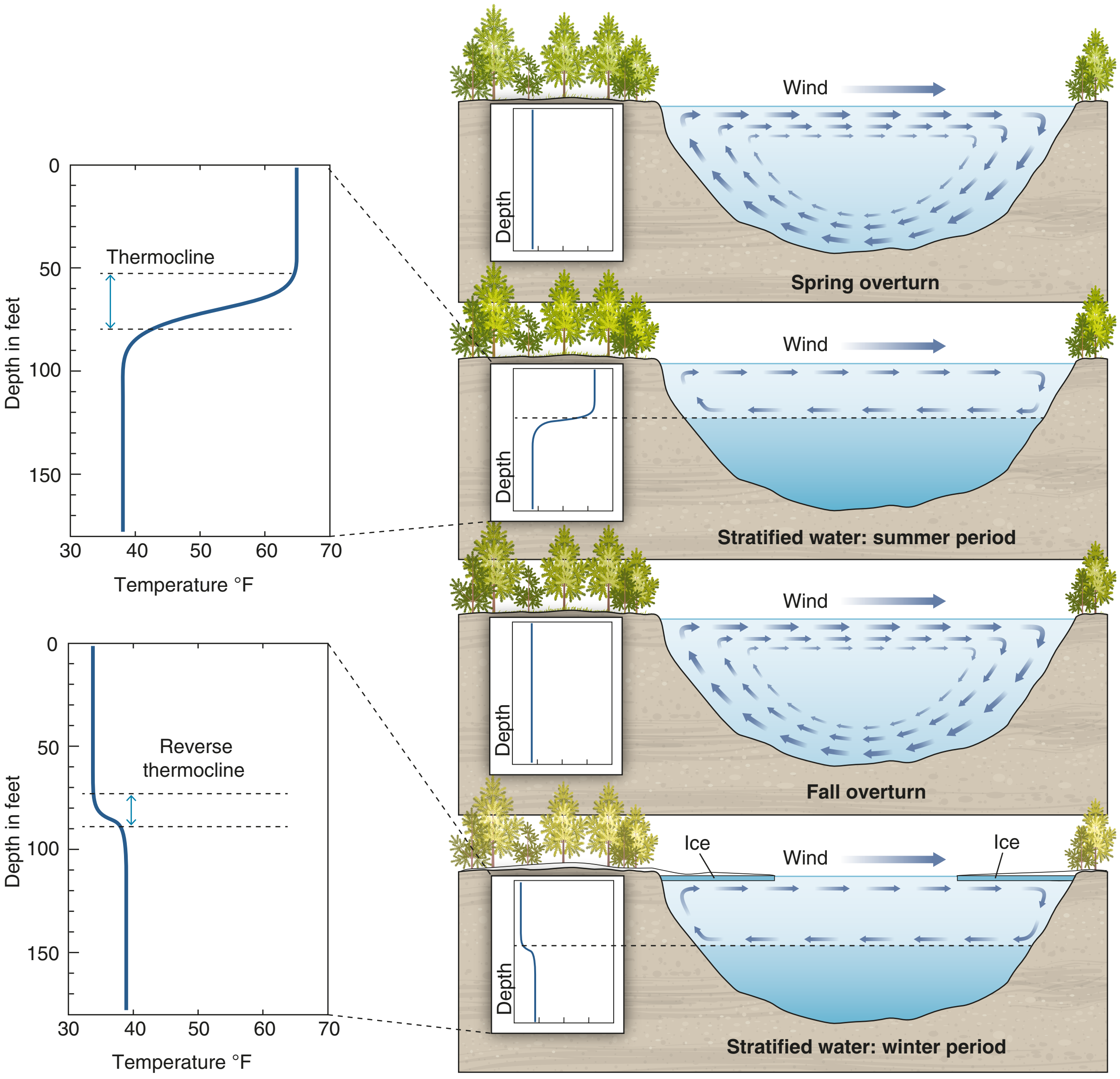
0
50
100
150
Depth in feet
Thermocline
30
40
50
60
70
Temperature °F
0
50
100
150
Depth in feet
Reverse thermocline
30
40
50
60
70
Temperature °F
Wind
Depth
Spring overturn
Wind
Depth
Stratified water: summer period
Wind
Depth
Fall overturn
Ice
Wind
Ice
Depth
Stratified water: winter period

Figure 5.7 Thermocline, reverse thermocline, and the annual cycle of overturn in lakes.

not live near a coast. Other freshwater environments—such as rivers, caverns and caves, high-altitude and frozen lakes, and areas with submerged wreckage—can be dangerous. You should complete specialty training courses before diving in any hazardous area.

You may encounter strong currents in rivers (see figure 5.8). The currents are strongest at the surface and on the outside of bends. Countercurrents and swirling currents, called **eddies**, are common, and entanglements are likely in many rivers. Water flowing through a porous obstruction, such as a pile of logs, can pin a diver against the debris. Currents can undercut the bank of a river, so direct access to the surface can be impossible at times. Rivers are especially prone to seasonal changes and are often unpredictable.

Freshwater springs can provide beautiful diving environments. The flow of clean, calm water, which is usually at a moderate temperature (65 to 78 °F, or 18 to 26 °C), provides excellent visibility. The water often flows through underground limestone cave systems that extend for thousands of feet. Diving in open-water basins in a spring is appropriate for certified divers, but a diver should not enter any overhead area where a direct, vertical ascent to the surface is not possible. A large, roomlike opening where light from the surface can be seen is called a **cavern**. Areas that extend farther than a cavern where no surface light can be seen are called **caves** (see figure 5.9). Divers must complete specialty courses and must meet several requirements to dive in caverns and caves. Entering such an environment—even for a short distance—without the required training and equipment can be fatal. No amount of diving experience without cave diving training qualifies a diver to enter a cave. It is easy to

Figure 5.8 Because many rivers have strong currents and pose other unique diving challenges, specialty training is advised.

Berenika/iStock/Getty Images

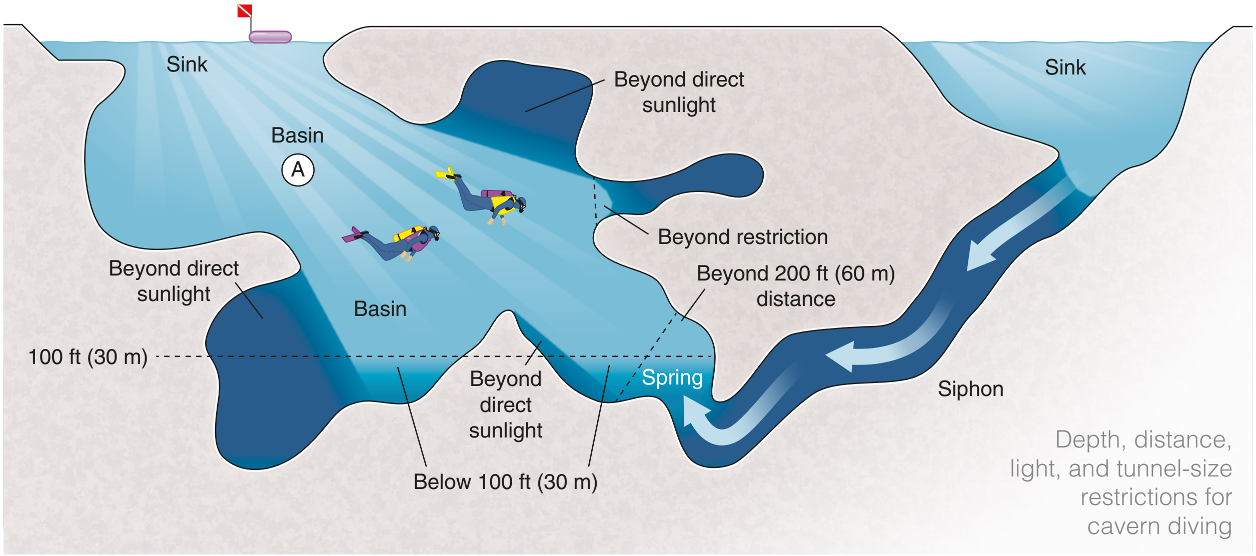
Sink
Basin
A
Beyond direct sunlight
Beyond restriction
Beyond 200 ft (60 m) distance
Sink
Beyond direct sunlight
Basin
100 ft (30 m)
Beyond direct sunlight
Spring
Siphon
Below 100 ft (30 m)
Depth, distance, light, and tunnel-size restrictions for cavern diving

Figure 5.9 Dark blue areas are caves; light blue areas are caverns.

become disoriented, to stir up thick clouds of silt that instantly reduce visibility to nothing, and to panic and drown. You may dive in spring basins, but stay out of caverns and caves unless you meet the requirements to dive in them. Appendix A contains contact information for organizations that offer training in cavern and cave diving.

Sometimes the earth collapses into an underground cave system and forms a **sink**. The water flowing into the sink forms a basin. An opening called a **siphon** channels water from a sink back into the system. Diving in siphons can be dangerous. The amount of water moving through a system depends on the amount of rainfall in the area. Under extreme conditions, the normal flow of water can reverse.

People dig pits to excavate sand, gravel, and stone. At some depth, they encounter the water table, and the pit floods, forming a quarry or sand pit. At some of these sites, you can see submerged construction equipment that was abandoned when the pit flooded. Sand pits and quarries tend to have fair to good visibility, although the disturbance of silt that has settled to the bottom can ruin the visibility quickly. Gravel quarries usually have more silt than other types of quarries, so the water may be turbid.

Lakes can be excellent dive sites. There are freshwater lakes at altitudes above 10,000 feet (3,048 m). Diving at altitudes above 1,000 feet (305 m) is a specialty called altitude diving. This type of diving presents many hazards because the rate of pressure change is greater when you descend into water at an atmospheric pressure that is less than the pressure at sea level. You must follow special procedures to avoid decompression sickness. Problems can also result from the thinner air, which provides less oxygen with each breath taken above water.

When ice forms over water in the winter, diving becomes hazardous. Perils of ice diving include hypothermia, the freezing of regulator and buoyancy control valves, and getting lost beneath the ice. Ice diving can be beautiful and adventurous, but it is dangerous when attempted without the proper training, equipment, and procedures.

Potential hazards when diving in freshwater include body heat loss, submerged trees, wire, fishing line, hooks and lures, log jams, debris, currents, rapids, whirlpools, poor visibility, silt, and overhead restrictions. You can minimize these hazards with training, experience, equipment, correct diving techniques, area orientations, and good judgment.

General Saltwater Diving Conditions

The earth contains five oceans. Seas, gulfs, and bays are smaller sections of the oceans. The land beneath the oceans is not flat. The **continental shelf** is an underwater area that extends from land and slopes gradually to a depth of about 600 feet (183 m). Beyond the continental shelf are great underwater canyons and mountains, as well as hills, valleys, and great plains. In some areas, the tips of mountains and volcanoes extend above the surface of the ocean from great depths, forming islands. In other areas, deep canyons sever the continental slope, creating deep-water conditions close to shore.

The saltwater environment varies greatly from one part of the earth to another. The clear waters of the Caribbean, with temperatures that can exceed 85 °F (29 °C), feature coral reefs covered with beautiful, lush animals that look like plants. Schools of colorful tropical fish abound. Temperate waters, with temperatures ranging from 55 to 70 °F (13 to 21 °C), may have great forests of kelp, which contain more life per volume than a tropical rain forest. Cold ocean waters in northern latitudes are nutrient rich and teeming with life. Beautiful sights and wonders can be experienced in all the oceans of our planet.

Oceans are always in motion. Tides, winds, and currents cause water movement. An underwater earthquake can move water by creating gigantic seismic waves known as a **tsunami**. Although these giant waves are also called tidal waves, the tide has nothing to do with them. Tsunamis can cause great destruction, but they are rare and can be forecasted.

The energy source for movement of ocean water may be local or may originate thousands of miles away. You need to understand what causes water to move, how it moves, and how to dive in ocean water that is in motion.

Tides

The gravitational attraction between the sun, the moon and the earth pulls water toward these heavenly bodies. The resulting increase in water depth is a high tide. Water pulled away from the sides of the earth in the process produces a decrease in water depth called low tide. A high tide forms on the side of the earth opposite the moon because the attraction of the moon is least at that point and because of the centrifugal effect created by the rotation of the earth. At any given time, two areas on the earth are experiencing high tides and two areas are experiencing low tides. During a brief period of time called a **stand**, the tide neither rises nor falls; four tidal stands can occur per day—two high and two low. Geographical formations interfere with or enhance the rising and falling of the water in some areas.

The sun also affects the tides, but only about half as much as the moon because the moon is much closer to the earth. When the moon and the sun are aligned with the earth, which occurs twice monthly (during the new and full moons), tides are highest and are called **spring tides** (see figure 5.10). When the moon and sun are at right angles to each other in relation to the earth, tides are lowest and are called **neap tides**.

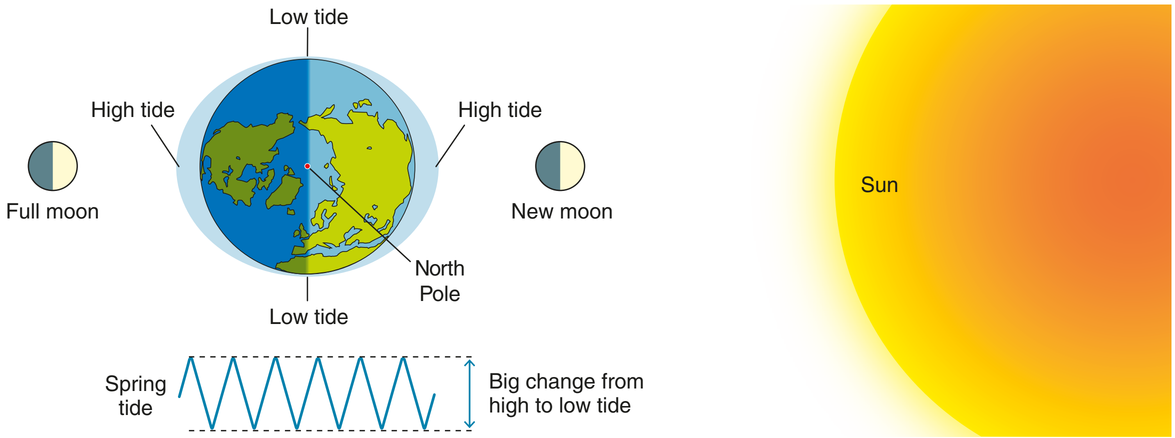

Figure 5.10 The effect of the sun and moon on tidal patterns.

The moon rotates around the earth in the same direction that the earth spins; therefore, the duration of a lunar day (moonrise to moonrise) is nearly 25 hours. This causes tides to occur at different times each day and explains why the heights of subsequent tides vary.

During a tidal change, water flows from one area to another. The water movements caused by the tides are called tidal currents, and they vary in velocity between tides. The flow of water into an area because of high tide is called a **flood**. The flow of water away from an area because of low tide is called an **ebb**. Between flood and ebb, a period of time occurs when the water has minimal movement. The term for this period is **slack water**. Because of geographical features of the earth, water cannot move instantly in many areas. There is usually a delay between the predicted time of a tide and the time when the water is slack.

Tidal changes affect diving activities and operations. Large differences in the water level between tides affect visibility and other conditions. Water height also affects entries and exits, shoal areas when boating, the mooring of boats, and the loading and unloading of vessels. Divers should know how high the tide will be. Tidal changes are small (less than 2 feet, or 0.6 m) in some areas or on some days, but they are large (more than 6 feet, or 1.8 m) in other areas or on other days. The greater the difference between high and low tide, the greater the effect of the tides.

Because the earth, moon, and sun follow set courses, the tides are predictable. Tides vary because the distance between planetary bodies and their positions relative to one another change constantly. Despite all the variables, scientists can predict with accuracy when the tides will occur and the heights of the tides. The U.S. government publishes tide tables, and national weather channels broadcast tidal information continuously. Local correction tables and current tables can provide more precise information. Tide tables refer to the height of the tide in relation to an average of the low tides in an area (see figure 5.11). Wind and barometric pressure affect the height of tides, aiding or opposing the movement of tidal waters.

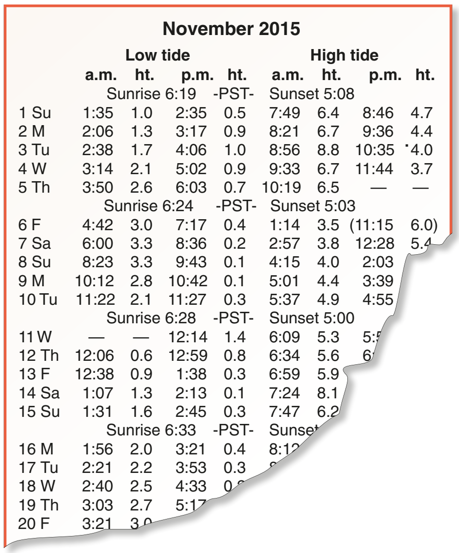

November 2015

	Low tide				High tide			
	a.m.	ht.	p.m.	ht.	a.m.	ht.	p.m.	ht.
	Sunrise 6:19			-PST-	Sunset 5:08			
1 Su	1:35	1.0	2:35	0.5	7:49	6.4	8:46	4.7
2 M	2:06	1.3	3:17	0.9	8:21	6.7	9:36	4.4
3 Tu	2:38	1.7	4:06	1.0	8:56	8.8	10:35	4.0
4 W	3:14	2.1	5:02	0.9	9:33	6.7	11:44	3.7
5 Th	3:50	2.6	6:03	0.7	10:19	6.5	—	—
	Sunrise 6:24			-PST-	Sunset 5:03			
6 F	4:42	3.0	7:17	0.4	1:14	3.5	(11:15	6.0)
7 Sa	6:00	3.3	8:36	0.2	2:57	3.8	12:28	5.4
8 Su	8:23	3.3	9:43	0.1	4:15	4.0	2:03	
9 M	10:12	2.8	10:42	0.1	5:01	4.4	3:39	
10 Tu	11:22	2.1	11:27	0.3	5:37	4.9	4:55	
	Sunrise 6:28			-PST-	Sunset 5:00			
11 W	—	—	12:14	1.4	6:09	5.3	5:5	
12 Th	12:06	0.6	12:59	0.8	6:34	5.6	6	
13 F	12:38	0.9	1:38	0.3	6:59	5.9		
14 Sa	1:07	1.3	2:13	0.1	7:24	8.1		
15 Su	1:31	1.6	2:45	0.3	7:47	6.2		
	Sunrise 6:33			-PST-	Sunset			
16 M	1:56	2.0	3:21	0.4	8:1			
17 Tu	2:21	2.2	3:53	0.3				
18 W	2:40	2.5	4:33	0				
19 Th	3:03	2.7	5:1					
20 F	3:21	3.						

Figure 5.11 Tide table.

Generally, it is best to dive at high tide, but you may need to time your diving to coincide with slack water in areas with strong tidal currents. The timing may not be important in areas where tidal changes are small. Consider local knowledge about the effects of tides when planning your dives.

Waves and Surf

Ripples form when wind blows across water. You can see this in a puddle on a windy day. When ripples form on a large body of water and the wind continues to blow, the sides of the ripples form a surface against which the wind can push to set the water in motion. The longer and harder the wind blows in a constant direction, the larger the waves that form. As waves move away from the area in which they were created—an area called the **fetch**—the tops of the waves become rounded, undulating forms called **swells**. The water within a swell moves in a circular motion but has little forward motion. The effect is similar to the transmission of a waveform along a rope that is tossed up and down at one end. Waves of energy travel along the rope, but the rope itself does not move forward.

Swells are energy forms that can travel thousands of miles through water and still contain a great deal of energy. The top of a wave is called the **crest**. The bottom is called the **trough**. The distance from the crest to the trough is the **wave height**. The distance between waves is a **wavelength**. The time it takes two waves to pass a given point is a **wave period**. A long series of waves is a **wave train**. The greater the height of a wave and the greater the wavelength of a train of waves, the greater the energy contained in the waves. Two trains of waves can have phases during which the waves of each train reinforce one another and produce larger waves, called wave sets. The formation of larger and smaller waves can occur in phases. By timing the rhythm of the surf, you can make entries and exits easier by moving through the surf zone when the wave sets are small. Figure 5.12 illustrates fundamental concepts of waves and surf.

The water within a passing wave moves in a circular motion (this motion is called an orb). The diameter of the motion equals the height of the wave at the surface and diminishes with depth. Motion from a passing wave can be felt to a depth equal to half the wavelength for a series of waves. If there is 100 feet (30 m) between waves, you can feel the effect of a passing wave to a depth of 50 feet (15 m).

When waves enter shallow water, contact with the bottom interrupts the circular motion of the water within the waves. The circular movements flatten and eventually become a back-and-forth subsurface motion called **surge**. Wave contact with the bottom causes waves to slow and to become steeper. The wave height increases as the depth of the water decreases. When the depth of the water is about the same as the wave height, the wave becomes unstable and tumbles forward. At this point, the water within the wave moves forward and gives up its energy in breaking waves known as **surf**. In areas with offshore reefs, sand bars, and underwater obstructions, waves can break in shallow water, pass over the obstruction, reform, and break again in shallow water near the shore. Waves that break offshore indicate the presence of shallow water in the area where the waves are breaking.

Waves can break and release their energy all at once, or they can spill forward and expend their energy over a wide area. Waves that build quickly and break suddenly form plunging breakers. Waves that spill forward over some distance to shore are spilling breakers. Plunging breakers occur on steeply sloping shores,

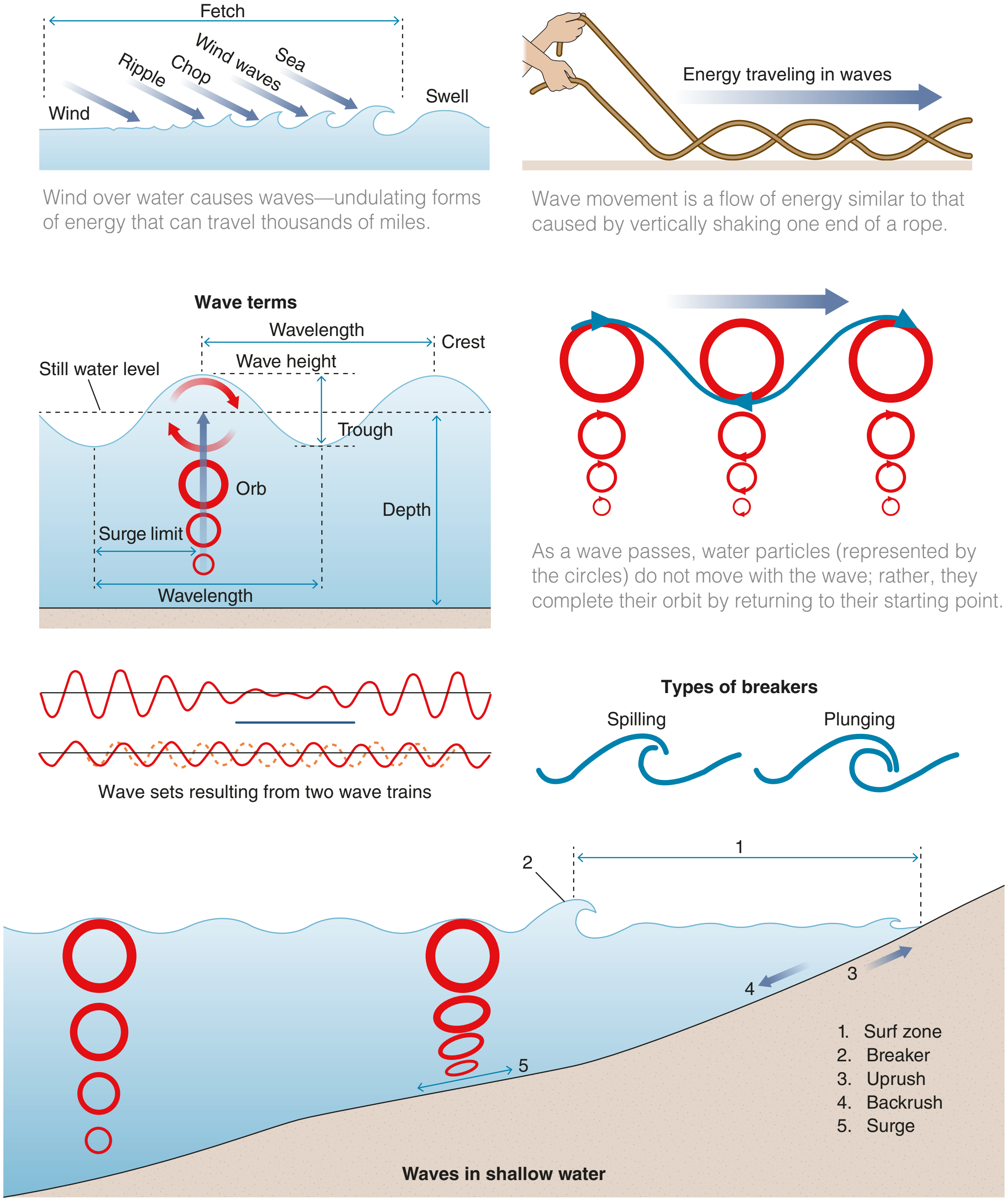
Fetch
Wind
Ripple
Chop
Wind waves
Sea
Swell
Energy traveling in waves
Wave terms
Wavelength
Crest
Wave height
Still water level
Trough
Orb
Depth
Surge limit
Wavelength
Types of breakers
Spilling
Plunging
1
2
3
4
5
1. Surf zone
2. Breaker
3. Uprush
4. Backrush
5. Surge
Waves in shallow water

Wind over water causes waves—undulating forms of energy that can travel thousands of miles.

Wave movement is a flow of energy similar to that caused by vertically shaking one end of a rope.

As a wave passes, water particles (represented by the circles) do not move with the wave; rather, they complete their orbit by returning to their starting point.

Wave sets resulting from two wave trains

Figure 5.12 **Wind speed, the absence or presence of obstacles, and the shape and depth of the ocean floor influence the formation of waves and surf.**

while spilling breakers occur on bottoms with shallow slopes. Spilling breakers tend to reduce visibility more than plunging breakers.

The crashing water of surf contains air. The white water and foam in the surf zone do not provide as much buoyancy as water that does not contain air. It can be difficult to remain above water in the foam of surf, but you should not attempt to rise above the breakers because you can be picked up and tossed forward by the moving water. You should remain low in the wave and breathe from your scuba regulator.

Surf rushes up the face of a beach and then flows back again to the still water level. The return flow of water is called the backrush. This countercurrent, sometimes called an undertow, does not extend beyond a depth of three feet (about a meter). Although backrush can be strong when surf is large on a steep beach, the belief that an undertow can carry a swimmer out to sea is a misconception.

The surf's crashing onto the shore moves sand. Gentle summer waves carry sand onto the beach. Large, rough winter waves carry the sand offshore, where it forms sand bars. This action explains why some beaches are rocky in the winter and smooth in the summer.

Currents

Currents are to water what wind is to air—fluid in motion. Because water is 800 times denser than air, water poses much more resistance to movement than air. You need to know how and why water moves and how to move with it. The force of water in motion is too great to resist. You must learn to use the flow of water to your advantage.

Wind, gravity, tides, and convection cause currents; the most common currents are wind-generated surface currents. Because of the rotation of the earth, currents flow at an angle to the wind that generates them. When the wind pushes water away from an area, a compensating current replaces the water displaced. Overall, the effect on the earth is to produce large circulating currents, called **gyres**, that move clockwise in the northern hemisphere and counterclockwise in the southern hemisphere.

When wind blows along a coast, a vertical compensating current—an upwelling—occurs. Another vertical current, the opposite of an upwelling, is a **downwelling**. A strong downwelling current is rare, but it can occur under the right conditions where there is a steep drop-off near shore. It is frightening to be pulled downward in water. You should find out if and where a downwelling occurs in an area, and you should avoid that location. If you get caught in a downwelling, swim horizontally until you are clear of the current. Trying to swim up against the current can lead to exhaustion and panic.

Water flows in layers (see figure 5.13). The water at the surface moves quickly with the wind, whereas the water a few feet beneath the surface moves slower. The greater the depth, the less the effect of a surface current. After 12 hours of a sustained wind in a given direction, the speed of a surface current is approximately 2 percent of the wind that drives it. Because water flows in layers, it is possible for a surface current to flow in one direction and for a subsurface current to flow in the opposite direction only a few feet beneath the surface.

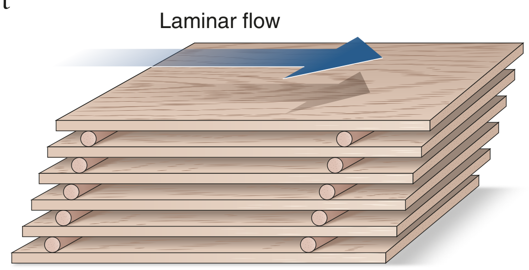

Figure 5.13 The laminar flow of water is similar to a stack of boards with rollers between them. When the stack is tipped, the top board moves much farther than the bottom board. Surface water moves more than water at depth.

Surface water at the equator expands as it is warmed and flows slowly toward the poles of the earth. At the same time, water cooled near the poles increases in density and sinks. Slow-moving convection currents result. These currents do not affect diving procedures, but they greatly affect the oceans by moving nutrients and pollution.

Water can move swiftly, and water movement intensifies whenever an amount of water passes through a restriction. When water moves past irregular formations, turbulence produces dangerous eddies and whirlpools. Swift-water diving is a specialty that is considered hazardous even with experience and training.

Incoming waves can pass over an underwater obstruction, such as a sand bar, and trap water on a shore. If the backrush of water flows back to sea through a narrow opening in the obstruction, a **rip current** exists. The current is narrow and strong and moves away from shore. The current dissipates shortly after passing through the restricted area. You can identify rips by a fan-shaped area of water on the shore, by muddy and foamy water in the rip area, and by a section of waves that breaks before the other waves (see figure 5.14). Rip currents can be stationary or moving. You must learn to recognize and avoid rip currents. If you are unable to swim toward shore, you are probably in a rip current. Swim parallel to the shore for about 60 feet (18 m) to get clear of the rip; then turn and proceed toward shore.

Waves breaking on a shore at a slight angle move water along the shore. The current is called a longshore current, and it affects diving. On steep beaches, the current can cut a shallow trough in the bottom close to shore. The trough, called a longshore trench, is a sudden drop-off and can cause an unsuspecting wading diver to fall. Items dropped into the surf zone move along the shoreline in the direction of the longshore current. The movement of sand on beaches occurs because of wave action combined with longshore currents.

Two types of currents result from all the forces of nature: standing currents, which are constant, and transitory currents, which occur briefly. The Gulf Stream that flows around Florida and up the eastern coast of the United States and the Antilles Current that flows through the Caribbean are examples of standing

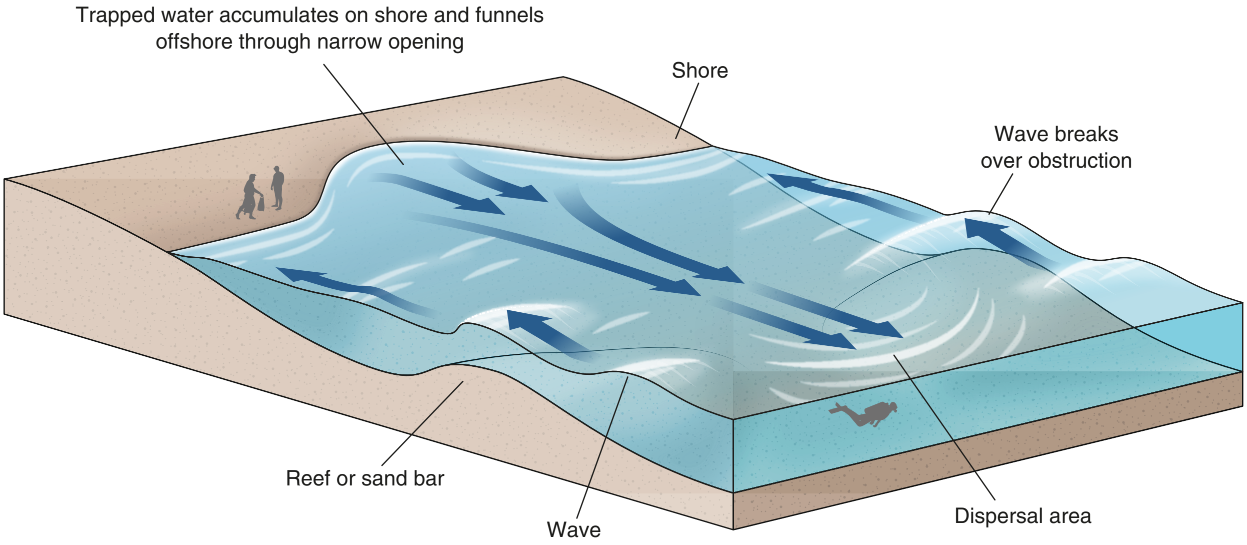
Trapped water accumulates on shore and funnels offshore through narrow opening
Shore
Wave breaks over obstruction
Reef or sand bar
Wave
Dispersal area

Figure 5.14 A classic rip current.

currents. A longshore current is a transitory current. A rip current can be either type. You should be familiar with the currents you will encounter or may encounter in an area.

Currents can flow at a rate of hundreds of feet per minute. The speed of a current is its **drift**, and the direction assumed by a current is its **set**. A fully equipped diver can swim at a rate of 60 to 100 feet (18 to 30 m) per minute (0.6 to 1.0 knots). Attempting to fight even a mild current is a futile waste of energy. Strong water movement can dislodge equipment, cause regulators to free-flow, and cause rapid heat loss. You must learn to recognize and estimate currents. If the currents are strong, you must avoid them. Even if they are mild, you should avoid struggling against them. The primary rule for diving in currents—when you begin and end your dive at the same point—is to dive in a direction opposite the flow at the surface; then use the surface current to aid you when you return at the end of the dive. (See chapter 7 for techniques to estimate currents.)

You also need to know some other general rules for diving in currents. Water movement is minimal at the bottom. Whenever possible, descend along a line from the surface to the bottom. If the current is still strong when you reach the bottom, ascend on the line and abort the dive. When you are unsure of your location during a dive, it may help to surface, determine the direction to the exit point, redescend, and move along the bottom toward your destination.

When there is a surface current, you should minimize the time spent at the surface. Whenever possible, hold on to a fixed object at the surface to keep from drifting downcurrent at a dive site. When a surface current is mild, swim against it to maintain your position at the surface. Develop the habit of noting your surface position in relation to a fixed reference. If you are caught in a sudden, strong current, swim across it to escape from it.

A dive in which the diver moves with the flow of a current is a **drift dive**. Simple drift dives along a coast are acceptable when planned. You enter the water at one point, move along the shore with the current, and exit at another point downstream. Another type of drift diving, which is a specialty activity, involves diving in currents from a boat. This type of diving is hazardous because the boat must be operated while divers are in the water. Do not attempt drift diving from boats unless you have completed specialty training and have a trained, experienced captain for the vessel.

Potential saltwater diving hazards include body heat loss, currents, surf, marine life, poor visibility, and fishing nets and equipment. You can minimize the hazards of ocean diving by following the recommendations that were described in this section.

Boat Diving

All divers eventually dive from boats, either small private boats or large charter vessels. Space is limited aboard boats, so you should organize your equipment compactly and pack it in the order in which you will need it. You may need to carry equipment for some distance to a boat, so plan to wear your scuba tank and weight belt and carry a bag in each hand. One bag should contain your dive gear, and the other bag should hold your lunch, camera, and any items that you want to keep dry.

Charter Boat Diving

You need to be well rested, so you should avoid partying the night before boat diving. Eat a meal that is free of acid and fat before you depart. Arrive at least 30 minutes before the scheduled departure time. Check in with a crew member. Usually, you must sign a roster and complete some forms. Ask where to stow your gear, and pick out a suitable space to store your dry items. If you are prone to seasickness, take preventive medication before the boat gets underway.

Tour the vessel. Locate the head (bathroom) and the emergency equipment. Stay out of the engine room and the wheelhouse (see example in figure 5.15). If you did not come with a buddy, you should find one before reaching the dive site. Listen carefully to the orientation provided by the captain or crewmember.

If you begin to feel queasy after the boat gets underway, you should go outside to the center of the boat and concentrate on the horizon. Sleep may help you acclimate to the boat's motion. If you are nauseated or vomiting when it is time to dive, avoid scuba diving. Vomiting underwater is hazardous. Mildly seasick divers may recover by snorkeling or by swimming to a nearby shore and waiting until they feel better.

Diving operations begin only when indicated by the captain. When you are ready to enter the water, check out with the divemaster. Typically, you will enter the water by stepping off the side of the boat, and you will exit the water by boarding a swim step at the stern (rear) of the vessel. Using the anchor line for your descent is a good practice.

Begin your dive against any current so that the water movement can assist you back to the boat at the end of the dive. Crew members often deploy a **trail line**—a long line with a float attached—to the stern for you to use in pulling yourself to the boat against the current (see figure 5.16). You should dive upcurrent; however, if you end up downcurrent from the boat and are unable to reach the trail line, you should make yourself buoyant, signal the boat, remain calm, and wait to be picked up after all the other divers have boarded the vessel.

If you hear a yelping, sirenlike sound while underwater, you should surface. The sound is from an electronic diver recall system. Remain where you are, and

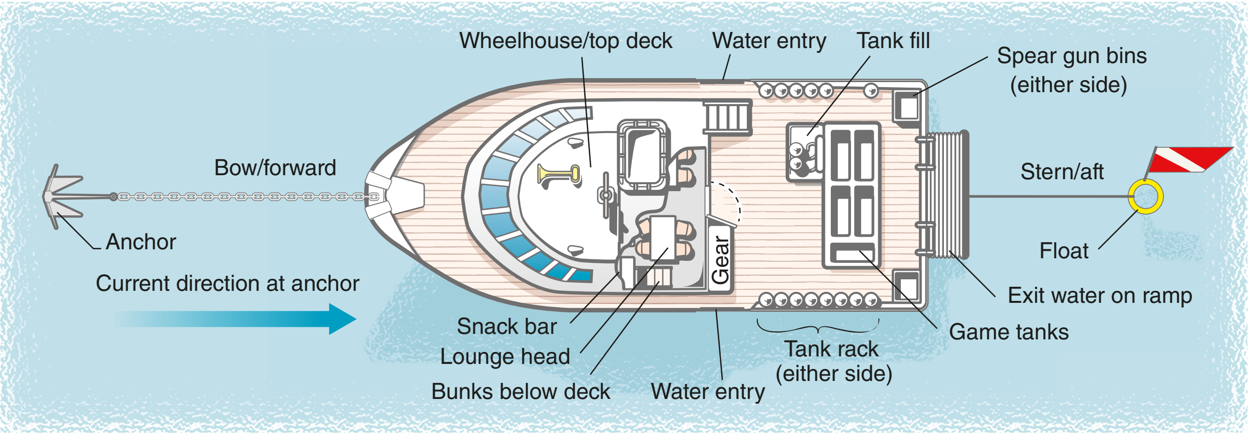

Figure 5.15 Typical charter boat configuration.

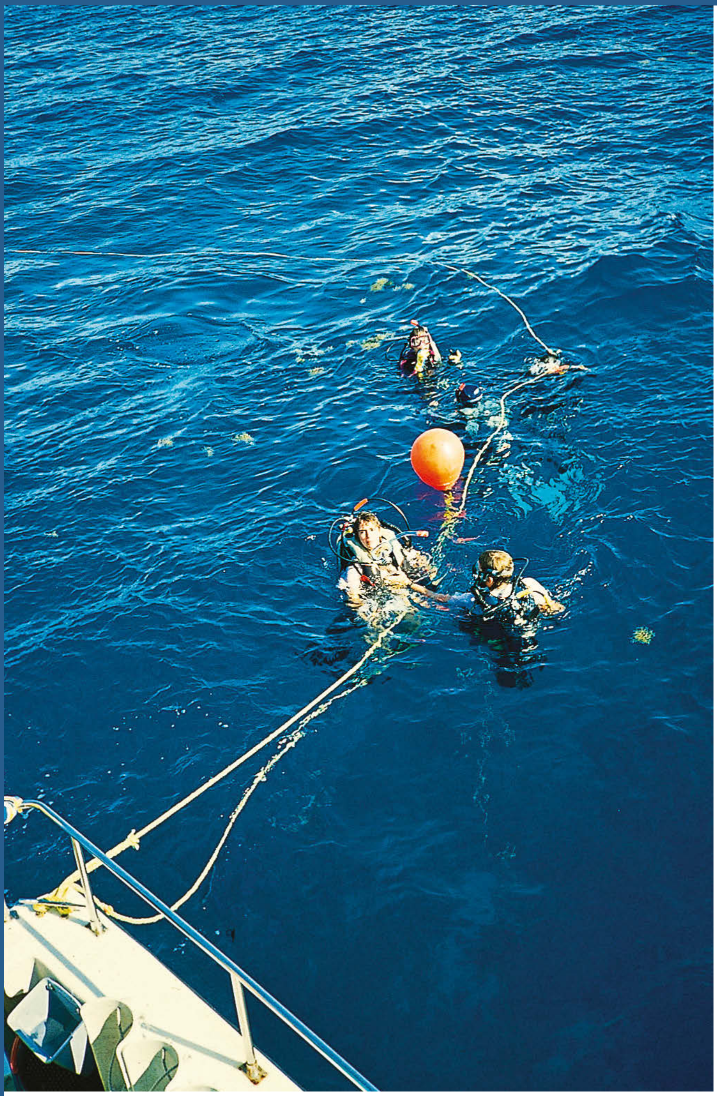

Figure 5.16 Divers can use a trail line to pull themselves to the boat if they surface downcurrent.

look to the boat for instructions. Learn the vessel's recall procedures in advance.

At the end of the dive, return to the boat by swimming along the bottom or at the surface. You should avoid skimming (swimming just beneath the surface), because you cannot be seen and you could be struck by a passing boat. Also avoid crowding the exit area. Divers should exit one or two at a time and remain clear of divers climbing the boarding ladder. Retain your equipment until immediately before exiting the water. If you do not know what equipment to remove, ask for instructions. Immediately after exiting, pick up any equipment that you handed to someone during your exit, and clear the exit area to make room for others.

After your last dive, pack your equipment before the boat gets underway. Be available for a visual roll call, which is required. Watch and learn from experienced divers.

Charter boat dives provide memorable diving experiences. Learn the proper techniques for diving from charter boats and then do what you have learned.

Small-Boat Diving

When you plan to dive from a small boat, make sure the boat is equipped for safety and has all the equipment required by the Coast Guard. The boat should also have a first aid kit. Leave a float plan—information about where you are going and when you plan to return—with someone who can summon assistance if you are delayed. Have a radio or cell phone aboard for emergency communications. Check the weather and water conditions in advance, and reschedule the dive if conditions will be unfavorable or are marginal on the planned dive day. Diving is not good when the weather is bad.

Small boats pitch, roll, and pound, so stow your equipment securely and expect it to get wet. If the trip to the dive site is short, you should suit up on shore. Doing so is easier and safer. If you must suit up in a small boat, remain seated and help one another. Before you enter the water, make sure that there is a means to get back into the boat. More than one diver has been stranded alongside a boat, unable to exit the water. Display the dive flag, but only while diving. Deploy a trail line and float.

The usual entry from a small boat is a seated back roll made by two divers at the same time. This method enables the divers to avoid rocking the boat and causing one diver to lose balance and fall in when unprepared. The descent should be down the anchor line to ensure that the anchor is set and clear. If possible, you should leave an operator in the boat. A sudden weather change, a slipping anchor, or a change of water movement could pose serious problems at the end of a dive if the boat is unattended.

At the end of the dive, remove your weights and scuba unit before exiting. If you cannot hand them to someone in the boat, clip them to gear lines that have been hung over the side. Exit the water and pull in your equipment.

Boat diving is good diving. Be prepared, conserve space, plan ahead, learn the rules, and follow those rules. Learn more about boating to make boat diving more enjoyable. Complete a course in seamanship and small-boat handling.

Summary

The aquatic environment is vast and wonderful. We depend on it for our existence. We need to learn about it, care for it, and respect it. We must protect the environment and encourage everyone we know to assist in that effort. Because the environment varies greatly from place to place, you need an area orientation before diving. There are hazards in the underwater world, just as there are hazards on land. You can minimize aquatic hazards with training, experience, planning, proper procedures, and common sense.

Application-of-Knowledge (AOK) Questions

This chapter provides knowledge but does not explain how to apply it. In the remaining chapters you will learn the skills for diving in various environments. Right now you need to digest what you know about the diving environment and consider how it might affect your diving.

1. In what ways will the temperature of the water affect your diving?
2. In what ways will the visibility under water be affected by the environment and by you?
3. What ways can you recall that can help you manage currents while diving?
4. When you encounter strange animals while diving, how should you respond?
5. How much training and experience are required until you are able to dive anywhere at any time?
6. When you are an experienced scuba diver, where is the safest and most comfortable place to be when diving?
7. If you were equipped for diving in the ocean and wanted to dive in fresh water, what equipment changes would you need to make?
8. What do you think you need to know about diving conditions when you intend to go diving?

Beautiful coral reefs in the Red Sea

6

Diving Skills

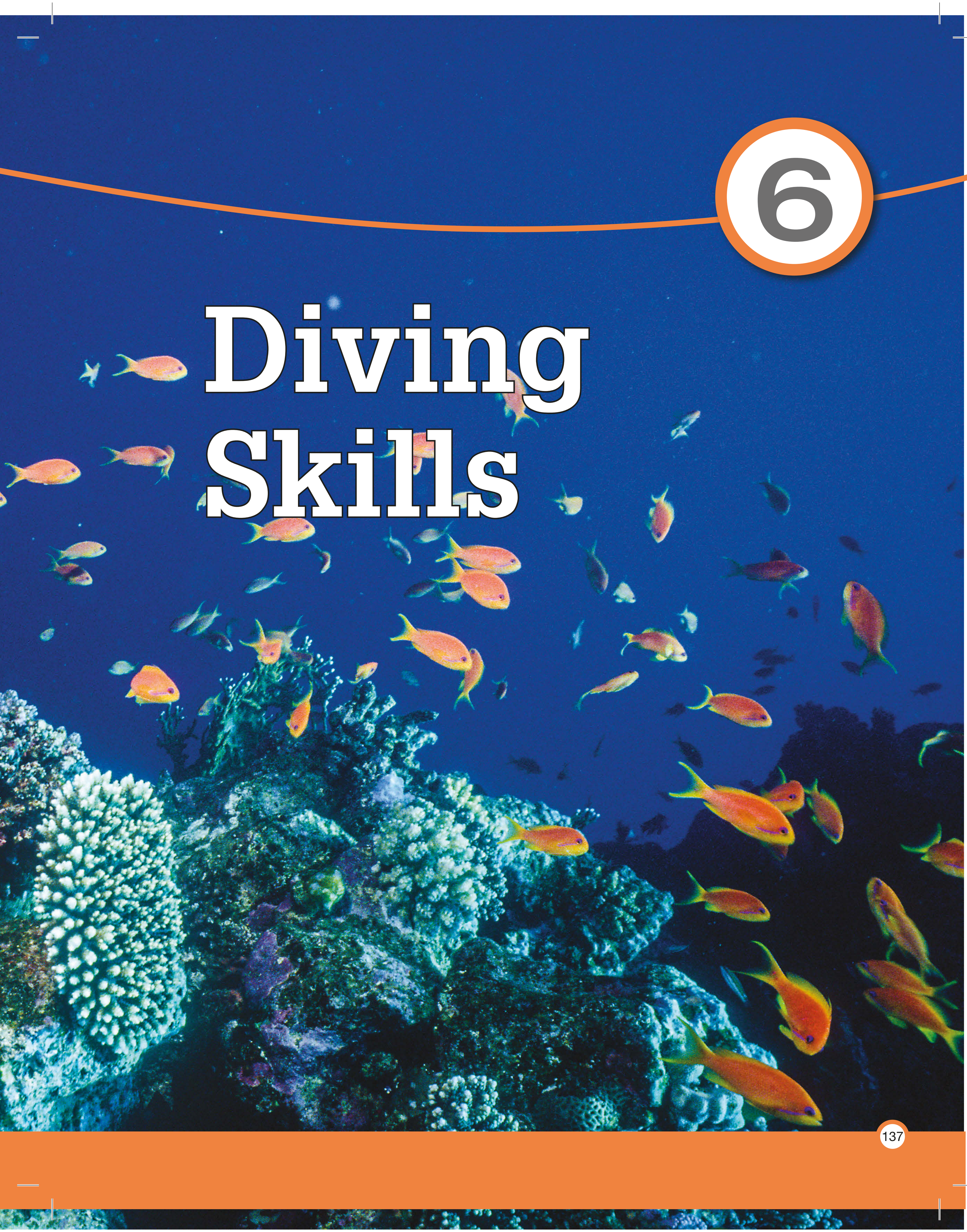

Dive In and Discover

By the end of this chapter, you will be able to do the following:

- Prepare, assemble, don, adjust, inspect, and disassemble your equipment for skin and scuba diving.
- Demonstrate 16 skin and scuba diving hand signals.
- Explain how to test and control buoyancy.
- Explain how to clear water from a mask, a snorkel, and a scuba regulator.
- Explain how to descend, swim with fins, and ascend when skin and scuba diving.
- Explain how to handle, remove, and replace skin and scuba diving equipment.
- Describe various entry and exit techniques for skin and scuba diving.
- Explain how to use dive flags.
- Explain the buddy system and lost buddy procedures.
- Explain how to use a compass for underwater navigation.
- Explain how to prevent and manage seasickness, dizziness, stress and panic, overexertion, coughing, cramps, entanglement, loss of buoyancy control, loss of air supply, and dive emergencies.
- Explain how to rescue a diver who is not breathing and the first aid procedures for an injured scuba diver.
- Define the terms *buddy system*, *flutter kick*, *scissors kick*, *modified frog kick*, *dolphin kick*, *surf*, *neutral buoyancy*, *diver's push-ups*, *open-valve ascent*, *reference descent*, *nonreference descent*, *buddy line*, *compass heading*, *compass course*, *reciprocal compass course*, *flaring*, and *buddy breathing*.

You will learn most of the diving skills in controlled conditions; when you have developed the basic skills, you will apply them in open water. (Open water is any body of water representative of dive sites in the local area.) When you can demonstrate the skills of diving in open water, you are ready to receive your diving certification.

Preparing to Skin Dive

This section covers the skills of skin diving, many of which apply to scuba diving. When you swim at the surface and breathe through a snorkel, you are snorkeling. Scuba divers snorkel to conserve the air in their cylinder for underwater activities. When you breathe through a snorkel at the surface, hold your breath, and dive beneath the surface, you are skin diving. As a diver, you need to master the skills of snorkeling, skin diving, and scuba diving in order to enjoy all that the underwater world has to offer.

Preparing Equipment

Be ready to start your lesson when you come to your first water session. Prepare your mask, snorkel, and fins in advance. Clean your mask lens thoroughly so it will not fog (a description of the cleaning procedure is provided in chapter 4). Adjust the mask strap so it is snug but not tight. Attach the snorkel to the mask strap on the left side, and adjust the snorkel so it is comfortable in your mouth. If your fins have heel straps, adjust them so they are snug but not tight. Complete all adjustments before your first water session.

Inspect your skin diving vest or buoyancy compensator by inflating it, making sure it does not leak, and then deflating it. Put it on, reinflate it, and adjust the strap or straps so it will stay in position in the water. You may need to use a strap that runs between your legs and attaches to the front and back of the flotation device to keep it from riding up when you are in the water.

If you wear an exposure suit, your instructor will suggest an initial amount of weight for your weight belt. Adjust the length of the belt at the buckle end so the excess strap at the opposite end does not exceed 6 inches (15 cm). Allow 2 inches (5 cm) for each two-slot weight that you thread onto the belt. Distribute the weight on both sides of the belt so that you will be balanced in the water. Lock the weights in place with weight retainers.

Mark your personal equipment with your initials so that you can identify it. You can use special paint or markers (available at dive stores) or colored tape. Pack

Diving requires many skills, and you must repeat these skills correctly until you can execute them automatically. When you have mastered the skills of diving, your enjoyment can begin because you will be able to devote more attention to things of interest. The basic skills of skin diving and scuba diving are introduced in this chapter.

Anemones and clownfish in the Red Sea

your equipment—with the exception of your weight belt—in a gear bag. Place the items you will don last, such as your fins, on the bottom. Place the items you will don first, such as your exposure suit, on the top.

Donning and Inspecting Skin Diving Equipment

Don the pants or legs of your exposure suit first, then your boots, and then the top part of your suit. Donning a snug-fitting wet suit is easier if you wet the inside of your suit with water (to which you have added some mild shampoo) or if you wear a spandex suit as an undergarment. Place the ends of the legs of your wet suit over the tops of your boots so that water can drain from your suit when you exit from a dive. If the boots are outside the legs of your wet suit, water from the suit will balloon your boots when you get out of the water. If you become warm while donning the exposure suit, cool yourself before proceeding with your preparations. For open-water diving, don a cold-water hood before donning your wet suit jacket so that the skirt of the hood is underneath your jacket.

After you put on your exposure suit, you should clean and defog your dive mask and then set it aside until you are ready to don it (see the description of the cleaning procedure in chapter 4). Leave about 1/2 inch (1.27 cm) of water inside the mask so it will not dry before you put it on, and place the mask where it will not get damaged.

Don your skin diving vest or BC next. Place the skin diving vest over your head, and then secure the straps. Inflate the vest fully to make sure it will not be too tight. The straps should be as snug as possible without being uncomfortable or interfering with breathing.

Don your weight belt after the skin diving vest so the belt will be clear of the vest straps. Grasp the free end of the weight belt (the end without the buckle) in your right hand, and grasp the buckle end in your left hand. Pick up the belt, step through it, pull it into position across your back, and bend forward so that gravity supports the weight of the belt. Then tighten the belt and secure the quick release. Always wear the weight belt with a right-hand release, even if you are left-handed, so that a rescuer will know how to release your belt in an emergency. If you always hold the free end of the belt in your right hand when donning it, you will always have a right-hand quick release.

Don the remainder of your skin diving equipment—mask, snorkel, and fins—at the water's edge or, if it is calm, in the water. Place the mask on your forehead and pull the mask strap over the back of your head using both hands to position the strap. Position the mask on your face, clearing any hair from beneath the sealing edge of the mask. Reposition the mask strap so it lies flat. If the mask has a split strap, position the split above and below the crown of your head. Don your snorkel when you don your mask. Check the adjustment of the snorkel when your mask is in place.

Stabilize yourself when putting on fins so that you don't lose your balance and fall. Hold on to your buddy (see "The Buddy Diving" section later in this chapter) or an object for support, or sit at the water's edge. Hold one side of a fin, bend one leg into a figure-four position, push the fin onto your foot, and pull the strap or heel pocket into place (see figure 6.1). Repeat the process for the other fin. Avoid walking with fins. If you must walk a few steps while wearing fins, shuffle your feet and walk backward. If you try to walk forward, you can lose your balance or damage your fins.

After you have donned your equipment, you should inspect it for completeness, correct positioning, and adjustment. When you are satisfied with your

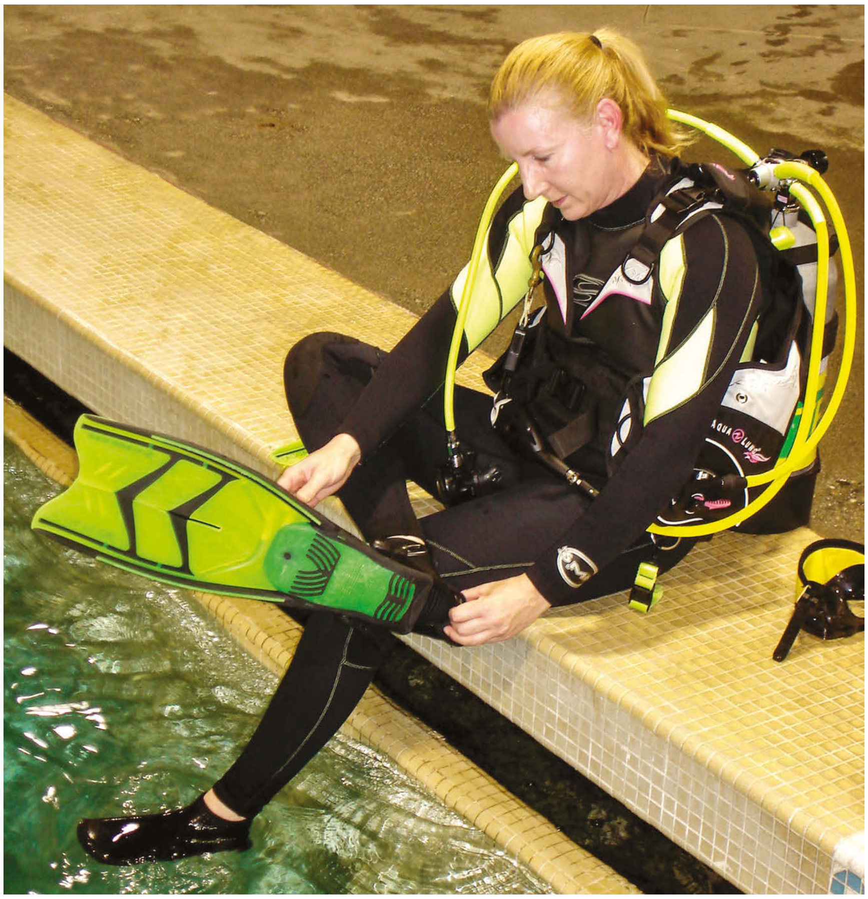

Figure 6.1 Donning fins.

equipment, inspect your buddy's equipment while he inspects yours. Inspect from head to toe. You may be able to see something that your buddy could not see, or vice versa. Develop the habit of inspecting each other's equipment before every dive.

Skin Diving Skills

Diving skills can be divided into three categories: skin diving skills, scuba diving skills, and problem management skills. This section introduces skin diving skills.

Using Skin Diving Hand Signals

You cannot talk with a snorkel in your mouth and your face in the water, so you must use hand signals as a primary means of communication. You need to learn and use the standard hand signals shown in figure 6.2. You will learn additional hand signals for scuba diving. Display hand signals clearly and deliberately when you send them, and acknowledge all hand signals you receive.

Figure 6.2 Skin diving hand signals.

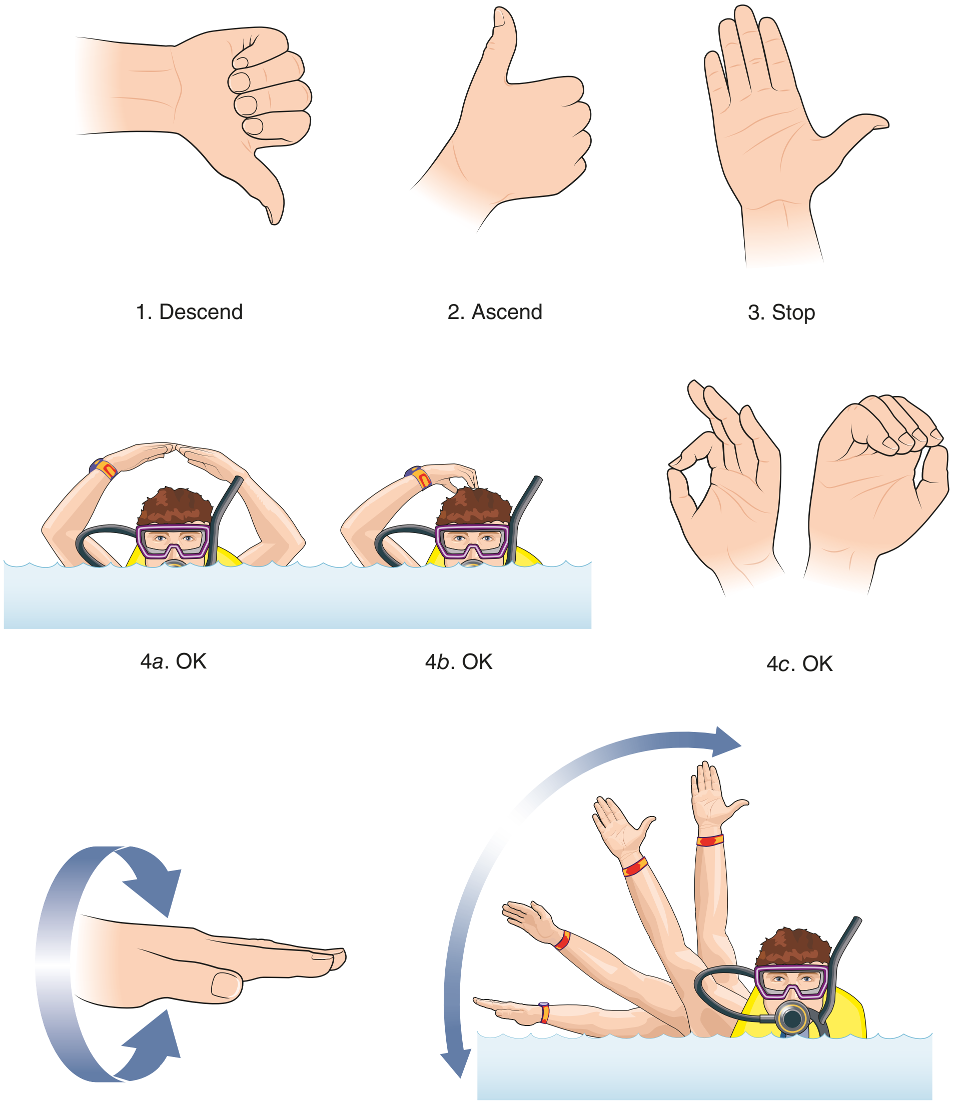

1. Descend

2. Ascend

3. Stop

4*a*. OK

4*b*. OK

4*c*. OK

5. Something's not right

6. Emergency

Using a Skin Diving Vest

You inflated and deflated your skin diving vest as part of your diving preparations. You also need to learn how to deflate and inflate the vest in the water. To deflate the vest, position your body so the vest's exhaust port is the highest point. Hold the deflation valve open while you sink lower in the water. The water pressure helps force the air from the vest. You may have to hold the collar of the vest down with one hand to remove the air from that area. When you have vented all of the air from the vest, close the valve to prevent water from entering. Get your buddy to confirm that all the air is out of your vest.

Inflating the vest in the water is easier with your head underwater instead of above water. Take a breath of air, duck your face beneath the surface, insert the oral inflator into your mouth, open the valve, exhale into the vest, and close the valve. Repeat this process until you have the desired buoyancy.

Testing Buoyancy

Your buoyancy should be neutral at the surface. If you are too buoyant, you will have to struggle to descend and remain underwater. If you are not buoyant enough, you will tire quickly at the surface and sink while swimming underwater. You must be able to adjust your buoyancy correctly.

If you are wearing an exposure suit, you need weights to offset the buoyancy of the suit. With all equipment in place, position yourself in chest-deep water. Exhaust all air from your skin diving vest. Take a full breath, hold it, lift your feet from the bottom, and remain motionless while you slowly count to 10. If you sink, remove some weight and try again. If you exhale half of the air in your fully inflated lungs and still can't submerge, you need to add weight until you can. When your weighting is correct, you remain at the surface while holding a full breath and sink when exhaling. Adjust your weighting until you achieve neutral buoyancy, which means that you neither sink nor float when holding an average breath. Have your buddy watch while you make the final test. With your lungs about half full of air, the top of your head should remain even with the surface of the water when you are motionless in a vertical position.

Clearing the Mask

When your head is above the surface, water inside a mask will run out if you pull the bottom of the mask away from your face. Water inside a mask underwater also flows out the bottom of the mask if you displace the water with air. Putting air inside the mask is easy: Just exhale lightly through your nose. A long, light exhalation is better than a short, forceful one because a strong exhalation blows air past the seal of the mask and does not displace water effectively.

To clear a mask that has a purge valve, seal the mask against your face, tilt your head downward to make the purge the lowest point in the mask, and exhale through your nose until the mask is clear of water. To clear a mask without a purge, hold the top of the mask against your forehead, take a breath, and start exhaling slowly (see figure 6.3). When the level of the water is below your eyes, tilt your head back while continuing to exhale, and the remainder of the water will flow out the bottom of the mask. You must be exhaling when you tilt your head back or water will run up your nose. Bubbles from the bottom of the mask indicate when you have cleared all the water from your mask. It may

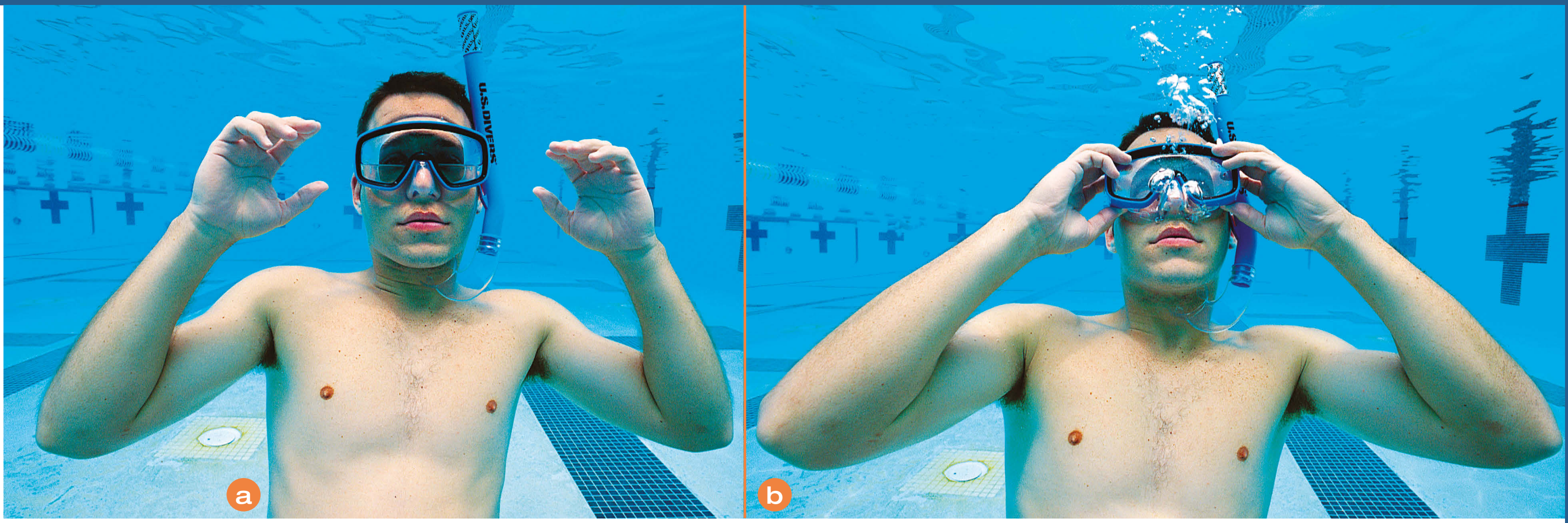

Figure 6.3 Clearing a mask.

sound as though it takes a long time and a lot of air to clear a mask, but with practice you will be able to flood and clear your mask several times after a single inhalation. Clearing requires only a few seconds.

Using the Snorkel

Stand in the water and lean forward with the snorkel in your mouth. Place your face in the water and inhale gently. If you inhale forcefully at first, you might inhale water. When you are sure the tube is clear, you may breathe more forcefully through your snorkel. When you descend beneath the surface, the snorkel tube can fill with water. Self-draining snorkels drain nearly all the water from the tube when you surface.

If your snorkel is not the self-draining or water exclusion type, you have to blast the water from the tube with a strong, sharp exhalation. To do this, blow hard into the snorkel to force the water out. Inhale cautiously after exhaling. If a little water remains in the tube, you can breathe past it if you inhale gently. After you fill your lungs again, expel the remaining water with another forceful exhalation.

If you have a simple snorkel without valves, you can clear the tube while ascending by using the displacement method. Begin by looking up to invert the snorkel; then exhale a small amount of air into the tube. As you ascend, the air expands according to Boyle's law and displaces the water in the tube, which will be clear of water when you reach the surface. The tube is inverted at this point, however, so when you turn it to the upright position, water will flow into the tube unless you exhale while rolling your head forward at the surface. Continuous exhalation prevents water from entering the tube. When you use the displacement method, which is easier than the blast method, you can take a breath the moment you reach the surface. Figure 6.4 shows the displacement method.

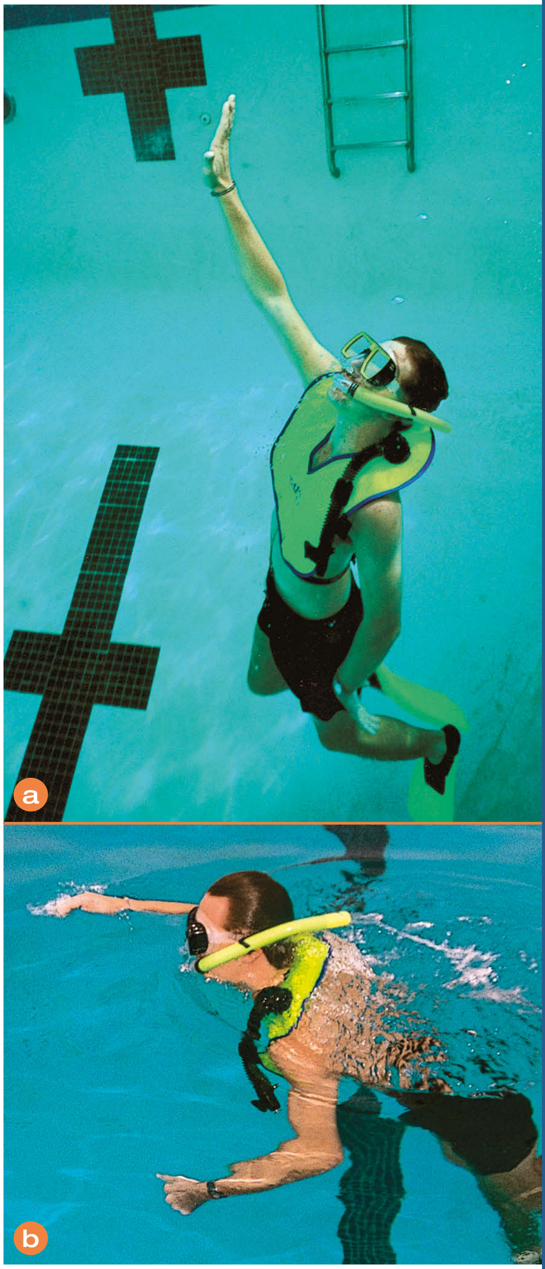

Figure 6.4 Displacement method of clearing a snorkel.

Learn to keep the snorkel in your mouth when you are in the water. Avoid the temptation to remove the mouthpiece and shake the water from your snorkel. You do not have to remove the mouthpiece when you can clear your snorkel proficiently. When your hands are occupied, you will not be able to remove and replace the snorkel to clear it.

Using Fins

Fins provide large surfaces that you can push—using the strong muscles of your legs—against the water for propulsion. The blades of the fins need to push the water the way a broom sweeps a floor. A broom cannot sweep if you move it up and down, and your fins will not provide propulsion if you move them lengthwise in the water. You must kick them back and forth in wide, sweeping kicks. Small fins allow short, quick kicks, but larger fins require wider, slower kicks. Moving efficiently, not speedily, is your objective when wearing fins.

The most common fin kick is the **flutter kick**, which is an up-and-down kick that you can do facing down, up, or to the side. Your fins need to be underwater to provide propulsion. It is easier to keep your fins submerged at the surface when you are on your back or on your side instead of your belly. Move your legs up and down from the hip while bending your knees only slightly. Extend your legs almost fully throughout each stroke, and use wide, slow kicks. Keep your hands at your sides or extend them in front. Do not use your hands and arms to swim. Figure 6.5 shows the flutter kick.

The **scissors kick**, shown in figure 6.6, is a resting kick. The scissors kick is similar to a flutter kick, but it requires less energy. While lying on your side, slowly extend one leg backward while you extend the other leg forward, and then pull your legs together quickly. Hold a streamlined position with your toes pointed while you glide through the water. When you come to a stop, repeat the stroke. Extend your legs almost fully throughout the stroke.

A **modified frog kick** is useful because it uses different muscles than the flutter and scissors kicks. Changing kicks is helpful if you become tired or develop leg cramps. The modified frog kick is different from a swimmer's frog kick. Assume a facedown position with your fins perpendicular to the bottom. Rotate your ankles so that the tips of the fins point outward, slide the fins tip first in an outward direction, then pull the bottoms of the fins together quickly in a wide, sweeping arc. Extend your legs almost fully throughout the stroke, and then hold the final position with your feet together and your toes pointed while you glide (see figure 6.7). When you come to a stop, repeat the stroke.

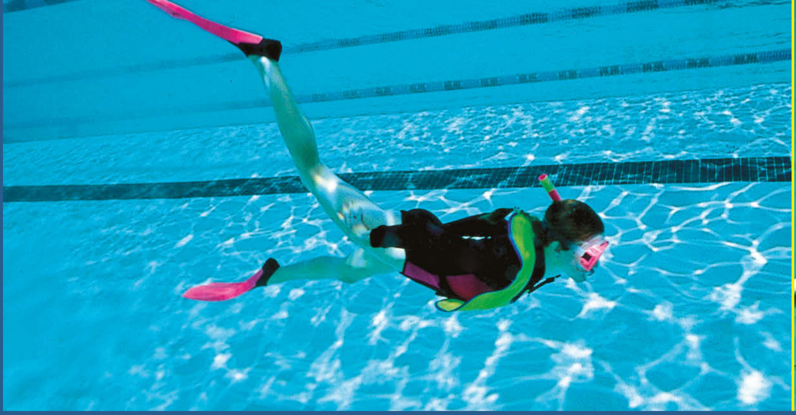

Figure 6.5 Flutter kick.

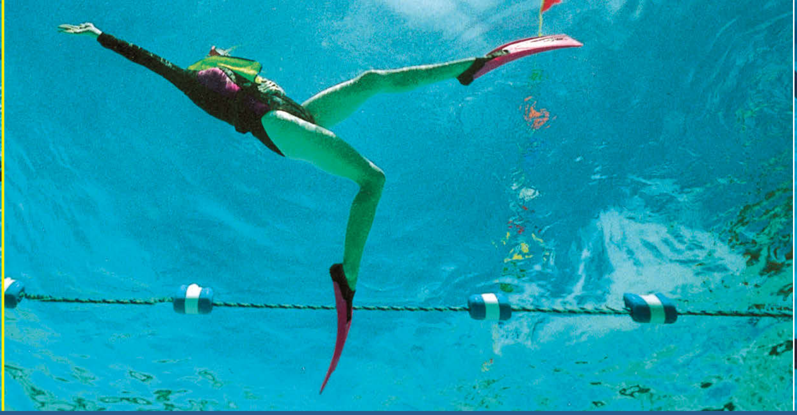

Figure 6.6 Scissors kick.

The **dolphin kick** is also useful for a change of pace or if you lose a fin. Hold both feet together continuously. Exert force against the water with a wavelike up-and-down motion of your body. To begin, straighten your body, extend your feet together, bend your shoulders forward, and pull your feet slightly backward. Next, lift your shoulders, thrust forward forcefully with your hips, and then pull down with your fins. When you see this kick demonstrated, you will be able to learn it by imitation. If you lose a fin, you can use the dolphin kick to propel yourself. Cross your legs and put the leg without a fin behind the one that has a fin. Figure 6.8 shows the dolphin kick.

Performing Surface Dives

You need to be able to dive down to see the beautiful world beneath the surface. The initial dive beneath the surface of the water is a surface dive. The principle of a surface dive is simple: You raise part of your body above the water and point yourself straight down; the part of your weight above the water drives you downward. Once you are submerged, you start swimming to continue your descent.

You should learn three types of surface dives: the pike, the tuck, and the feet-first dive (see figure 6.9). You do the pike dive while moving forward at the surface. Bend forward at the waist to make your trunk vertical in the water. Next, quickly lift your legs out of the water to a vertical position. The more of your legs you can lift out of the water, the farther your surface dive will push you under the water. For shallow dives, you can do the pike dive by lifting only one leg.

When you have learned the pike dive, you are ready for the tuck dive, which is similar. Do the tuck dive from a stationary position. Begin in an upright position in the water, pull your knees to your chest, and sweep backward with your arms to roll yourself forward in the water. When you are inverted, extend yourself fully into a vertical position. The procedure is a coordinated movement that you must do quickly to get your legs above water. After you extend your legs, the remainder of the dive is the same as the pike dive.

Use a feet-first dive in areas where surface plant growth is dense because you are less likely to get entangled. Begin in a stationary, upright position, and use a strong scissors kick to propel yourself up and out of the water as far as possible. Pull your arms downward to your sides for added lift. When you reach the highest point, point your toes downward, and hold your arms to your sides. The weight of your body will push you below the surface. Raise your arms upward when you are below the surface to push yourself even lower. When your downward momentum ends, do a tuck dive to invert yourself, then continue your descent.

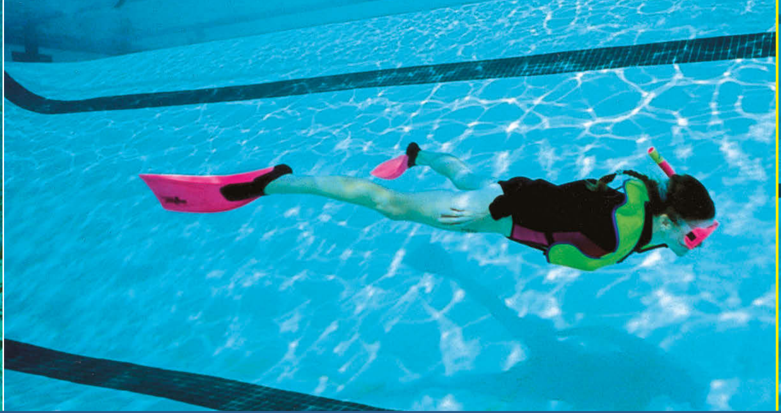
Figure 6.7 Modified frog kick.

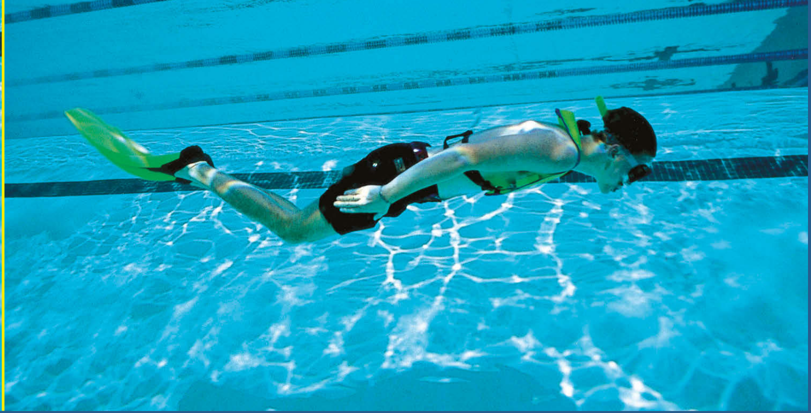
Figure 6.8 Dolphin kick.

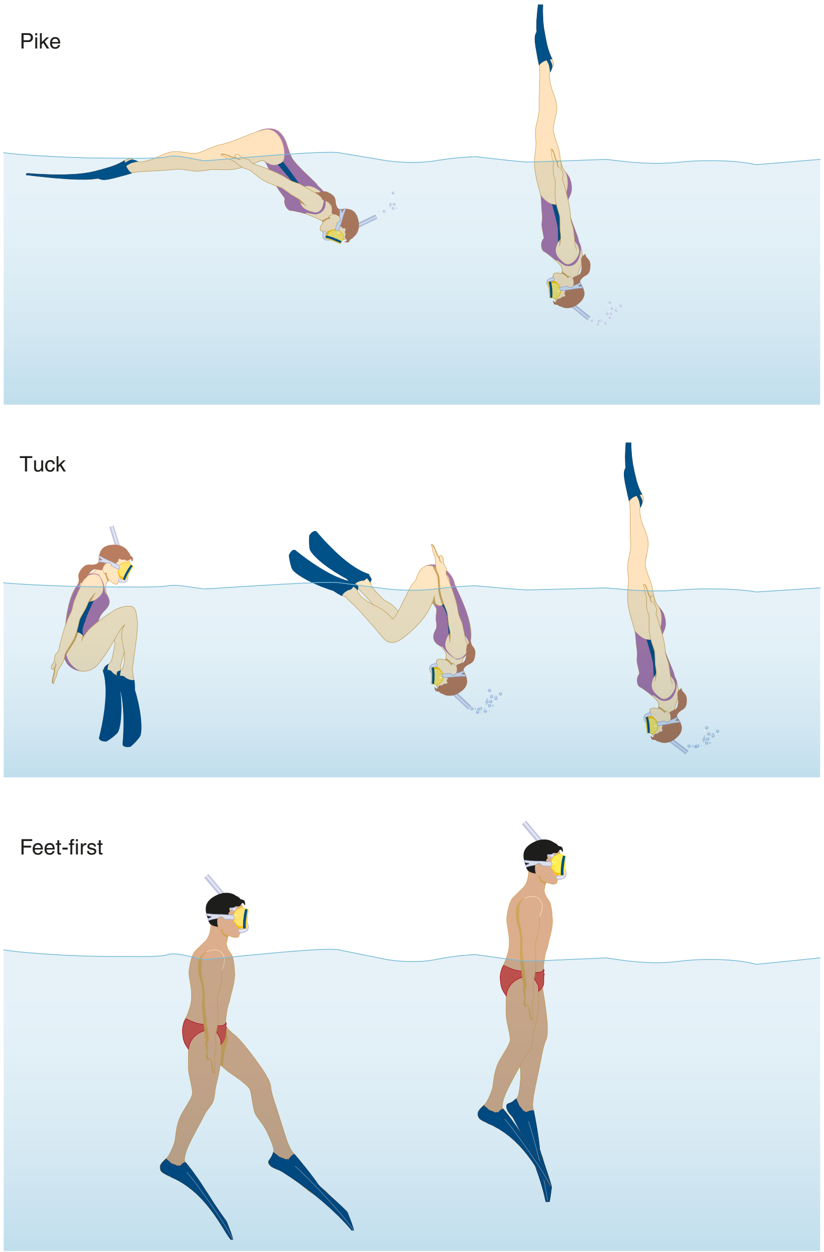

Figure 6.9 Surface dives.

Performing Descents, Underwater Swimming, and Ascents

To prepare to descend, vent all air from your flotation device. Use an ear-clearing maneuver (described in chapter 3) at the surface so that it will be easy to equalize pressure in your ears while you descend. Hyperventilate three times, inhale as fully as possible, and don't exhale until you return to the surface.

Initiate your descent with a good surface dive. Equalize the pressure in your ears and mask every couple of feet during your descent. Use wide, slow, powerful kicks to propel yourself to your desired depth, and then relax as much as possible.

Swimming underwater sounds simple, but some people have trouble controlling direction. Your head is your rudder. Pull your head backward while swimming facedown to go up. Bend your head forward to go down. Bend your trunk to the left or right to turn. With practice, you can go in any direction you choose without using your hands.

Follow these four rules for ascents:

1. Reach up; extend one arm over your head for protection (see figure 6.10).
2. Look up to avoid obstructions and to determine when the surface is near.
3. Come up slowly. A slow ascent is better than a rapid ascent because swimming rapidly uses more oxygen.
4. Revolve once or twice during your ascent to check your surroundings.

Alternate breath-hold dives with your buddy. One diver should remain at the surface, watching, while the other diver is underwater. Make sure your buddy knows where you are at all times.

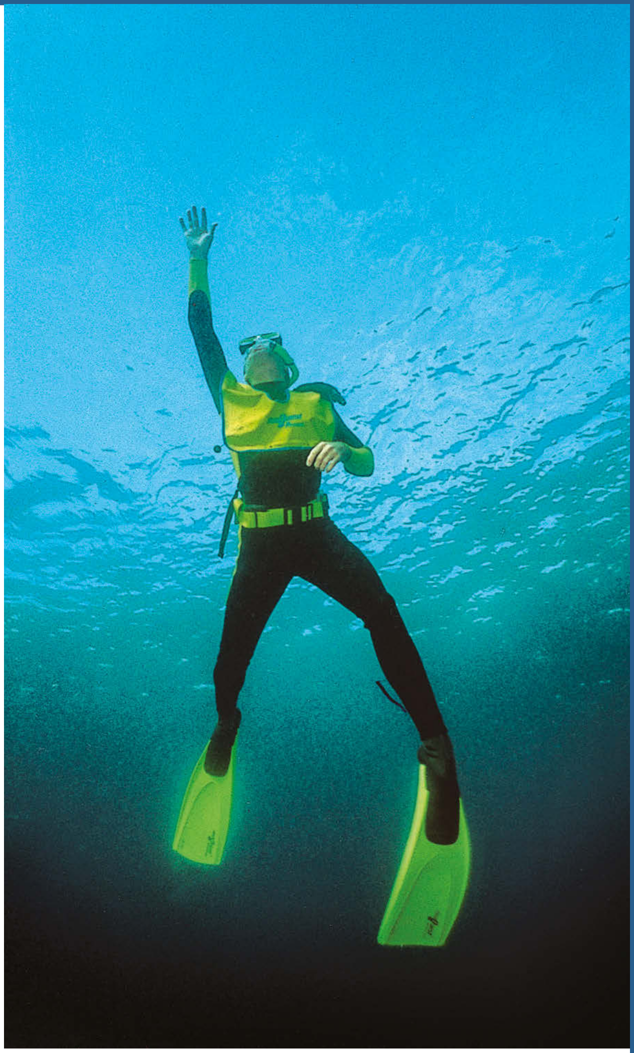

Figure 6.10 Extend your arm for protection during a skin diving ascent.

Handling Equipment

Occasionally you may need to remove, adjust, and replace skin diving equipment while you are in the water. You should be able to handle your equipment easily with training and practice.

To remove, adjust, and replace your mask, begin by inflating your skin diving vest. Because you will make the adjustment above water, you should use buoyancy to reduce the effort. After you make the adjustment, replace your mask using the procedure for donning the mask described earlier in this chapter.

You may need to adjust a fin strap or remove sand or gravel from your fins while diving. The fins are easier to remove, adjust, and replace than the mask is. You do not need as much buoyancy to work with your fins because you can look into the water and make the adjustment underwater. In fact, too much air in your skin diving vest can be a nuisance. Keep your face in the water, breathe through your snorkel, and work with one fin at a time.

You may need to tighten, adjust, or replace your weight belt. The weights may slip, or the belt may fall off. To tighten or replace your weight belt, you need to get the belt across the small of your back while you are in a facedown position breathing through your snorkel. If you try to replace or tighten the belt in an upright position, you will have a constant fight with gravity. When you are in the correct, facedown position, gravity becomes your ally while you secure the buckle. To get the belt into position across your back, begin in an upright position, holding the free end of the weight belt against the outside of your right thigh. Lean back into a horizontal, faceup position momentarily, and then roll to your left to a facedown position while continuing to hold the free end of the belt against your right thigh. At this point, the belt will be draped across the backs of your thighs. Clear your snorkel so that you can breathe. While holding the free end of the belt in your right hand, reach down with your left hand, grab the buckle end, pull the belt into position across your back, and secure the buckle.

Removing Equipment

Your equipment helps you adapt to the underwater environment. Develop the habit of keeping your equipment in place while you are in the water. You may need to defog your mask or make an adjustment, but other than that you should wear your equipment continuously. Avoid the tendency to prop your mask on your forehead when you are at the surface. Mask propping is a sign of distress and a good way to lose your mask and snorkel. If you must remove your mask while in the water, pull it down around your neck where it will be secure.

Bottom conditions at the dive site determine whether you should remove your fins when leaving the water. In some areas, you can remove your fins in waist-deep water and wade out of the water. In other areas, you need to wear your fins until you are clear of the water. Fins can be removed in several ways. When you prepare to climb a boat ladder, hold the ladder continuously while crossing your legs in a figure-four position to remove each fin. If the boat has a platform at the rear, swim onto the swim step and remove your fins in a kneeling position. When you remove fins in waist-deep water or on land, you should use your buddy for support. On steep beaches, you may choose to crawl from the water. Your buddy can remove your fins while you are on your hands and knees. Your buddy then crawls ahead of you, and you remove your buddy's fins. When you stand up, you exchange fins.

Weight belt removal also varies. When you can walk or climb out of the water, keep your belt on while doing so. When you remove a weight belt, lower it gently instead of dropping it. You can cause damage or injury if you develop the habit of dropping your weight belt. If you must pull yourself onto a dock or into a small boat, remove your belt first and hand it up.

Keep your skin diving vest in place until you are clear of the water. Remove your skin diving vest, and then your wet suit. Remove your wet suit by turning it inside out. Be sure to remove your boots before your wet suit pants. You will need assistance from your buddy to remove a pullover jacket or a jumpsuit.

Preparing to Scuba Dive

Nearly all of the skills of skin diving are used in scuba diving, but you need to learn many additional skills for scuba diving. This section introduces the preparatory skills of scuba diving; the next section covers basic and postdive skills. The more familiar you become with the skills by studying them, the better you will be able to perform them. You must learn the skills correctly from the outset in order to develop good habits for diving safety.

Packing Equipment

Nothing is more inconvenient than getting ready to dive and discovering that something is missing or not right. You should take steps to prevent equipment inconveniences. Begin by using a checklist when you pack your diving equipment to ensure that you have everything you need (see the diving equipment checklist). After taking inventory, inspect your equipment as you pack it. If you have not used your equipment for a while, assemble it and test it first. Make sure everything works properly.

Pack your equipment in your gear bag—except for your weight belt, tank, and BC. When you have to carry your equipment, put your BC on your tank, and carry the tank on your back. Carry your gear bag in one hand, and carry your weight belt in the other hand. Equipment-carrying devices with roller wheels can reduce the work of moving your diving equipment (see figure 6.11).

Figure 6.11 A gear bag with wheels can make transporting equipment easier.

Assembling Equipment

Today's backpacks are part of the BC and have one or more bands to secure the scuba cylinder. Many of the bands are made of webbing, which stretches when wet. Soak a fabric tank band in water for a couple of minutes before securing it around a scuba tank. The soaking softens the webbing and allows it to stretch when you tighten the band. A dry fabric belt may allow the tank to slip when the band gets wet and stretches. A tank that slips from a pack can be a hazard.

Orient your backpack with the opening of the tank valve facing the backpack. Stand the tank up with the valve handle facing to your right. Slide the tank band over the tank with the tank between you and the pack. Tighten the band so the valve opening points directly toward the pack (see figure 6.12).

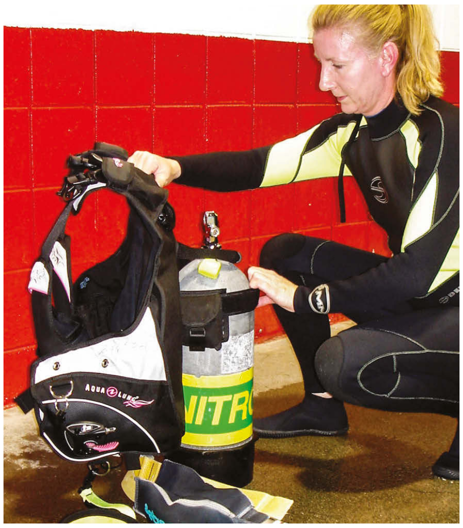

Figure 6.12 Strapping the backpack to the tank.

The height of the tank band on the tank varies with the type of backpack. Generally, the top of the pack is even with the base of the tank valve. After you attach the pack and before you attach the regulator, don the scuba tank and check the height adjustment by slowly tilting your head backward. If your head hits the tank valve, the tank is too high in the pack. Reach over your shoulder for the tank valve. If you cannot touch it, the tank is too low in the pack. Adjust the height as needed. When you become familiar with the correct height adjustment, you will not need to test it before completing the assembly of the scuba unit, but you should check it before entering the water.

Tighten the band as much as possible before securing it, and test the tightness by grasping the tank valve with one hand and the top of the pack with the other hand. Try to move the pack up and down on the tank. If there is movement, the band needs to be tighter.

Attach the regulator assembly to the scuba tank. The tank valve should have a valve protector or a piece of tape over the valve opening. Remove the cover or tape (but do not litter—put the tape in a trash container or in your gear bag). Make sure that the O-ring remains in place and is free of debris. Loosen the regulator yoke screw, and remove the dust cover. If your regulator second stages have purge depressors, release the purges. Expel any water or dirt from the valve by opening the tank valve slightly and momentarily before you attach the regulator to the tank.

Your regulator has several hoses. Orient the hose to your primary second stage toward the same side of the tank as the handle of the tank valve so that the hose will come over your right shoulder. When you orient the hose this way, the other hoses are properly oriented automatically as long as they are untangled and can hang freely.

Carefully seat the inlet opening of the regulator on the tank valve outlet, and then turn the yoke screw or DIN fitting until it is snug but not tight. Tighten the fittings using only your fingers and thumb. See figure 6.13 for an example of a regulator attachment.

Attach the various regulator hoses to the scuba unit. It is easier to attach the low-pressure hose to the BC before you pressurize the hose. Attach your instrument console to the BC, but do not attach your extra second stage yet.

Turn on the air. Hold the SPG with the front facing away from you and others so it will not cause an injury if it fails when pressurized. The SPG blowout plug should prevent an explosion from occurring, but holding the gauge is a good precaution. Open the tank valve slowly in a counterclockwise direction. Open the valve all the way; then close it one-quarter turn. You should feel the hoses stiffen under pressure. Listen for leaks in the system. If the regulator free-flows, cover the mouthpiece opening with your thumb to stop the flow. If air leaks

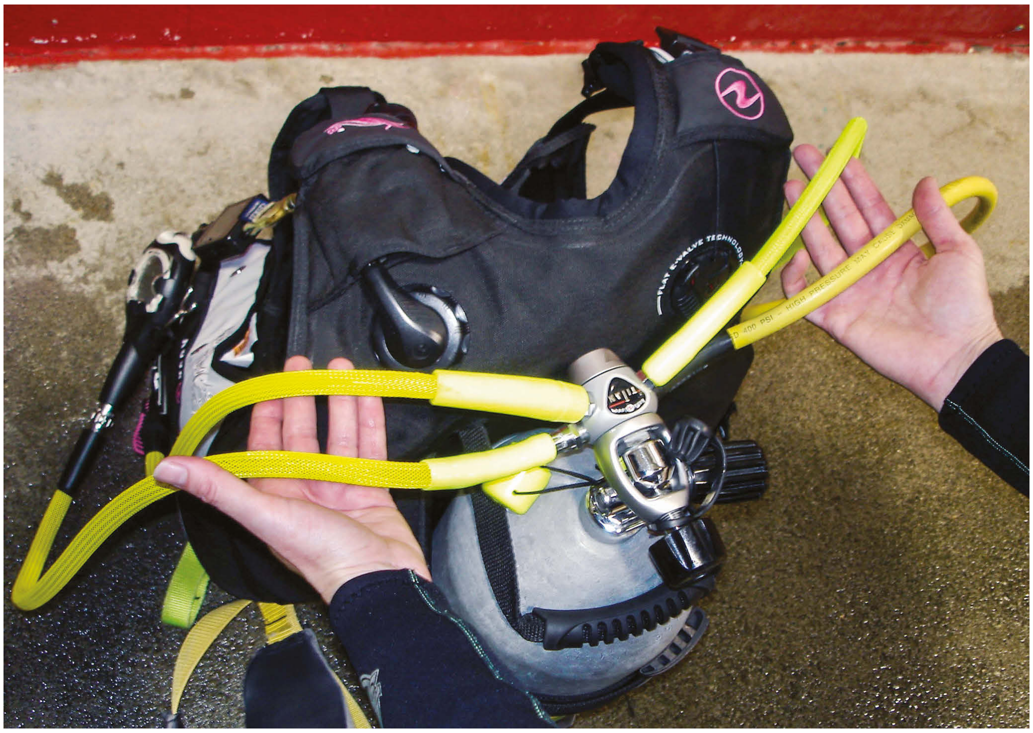

Figure 6.13 Your regulator and hoses should look like this if attached correctly to the cylinder.

from the tank valve seal, turn the air off, remove the regulator, and inspect the seal. You may have to replace the O-ring. Solve all air leakage problems before you use the scuba unit.

Testing the unit is the final step. Use your SPG to make sure the tank is full. Reset instruments in your console as needed. While looking at the SPG, depress the regulator purges momentarily to test the second stages and clear them of any debris; then breathe deeply through each second stage several times. The tank pressure reading should remain constant while you breathe from the regulator. If the pressure reading drops when you inhale, you have not opened the tank valve sufficiently. Do not open the valve partially to check the pressure of the tank. When you have finished testing the regulators, attach the extra second stage to the BC. Depress the low-pressure inflation valve on the BC for one or two seconds to make sure the valve functions correctly. When you have completed the assembly and testing of the unit, lay it on its side with the regulator and instruments on top.

Equipment assembly procedures are the same for both new and experienced divers. Always assemble and test your scuba equipment before you suit up. Practice will help you recall the assembly procedures.

Donning and Inspecting Scuba Equipment

With skin diving equipment, you don the weight belt after the skin diving vest, but with a jacket-type BC, you should don the weight belt before the scuba unit.

A good way to don a scuba unit out of the water is for your buddy to hold the system and help you into it. Once your arms are through the shoulder straps, you should bend forward and balance the unit on your back while you secure the waistband. Place the regulator hoses over your shoulders while you fasten the waistband so the hoses are not trapped beneath the waistband. Some dive boats have vertical tank racks on seats to allow you to sit while donning the system. Do not sit on the deck or ground to don a scuba tank, however, because the tanks of other scuba divers can strike you in the head.

Some divers prefer to don the scuba unit in the water—this is a good practice if you dive from a small boat or have a minor back problem. Put the scuba unit into the water first. Make sure that it does not float away! To don a jacket BC unit in the water the way you would don a coat, inflate it fully, sit on it, put your arms through the arm holes, and slide off and into the unit. Or you may put your scuba unit on over your head. Maneuver the tank in front of you with the valve facing you and the backpack facing up. Place the regulator in your mouth and keep the hose of the primary second stage between your arms. (If you place your right arm inside the loop of the hose, the hose will wrap around your arm when you try to lift the tank over your head.) Place your forearms completely through the arm holes just past your elbows, and begin to lift the scuba unit over your head. Rather than lift the unit, you should push yourself down in the water and duck beneath the scuba unit. Lower the tank into position on your back. Use your left hand to pull your snorkel clear of your BC as the tank lowers into position. When the tank is in place, lean forward and secure the waistband.

When you have donned your scuba equipment and the remainder of your diving equipment, inspect all your equipment to make sure it is positioned, adjusted, and functioning correctly. Your buddy should do likewise. Then inspect each other's equipment.

Do a head-to-toe buddy check. Specific checks for each item of equipment are listed in the Head-to-Toe Buddy Check sidebar.

Equipment inspections are important. Solving problems is easier before you enter the water than afterward. It does not take much time to inspect each other's equipment, but the practice can save a lot of time later. Make predive inspections a habit.

Scuba Diving Skills

To be a scuba diver and enjoy diving with minimum risk of injury, you need to learn the proper procedures for entering and exiting the water, controlling your buoyancy, descending and ascending, monitoring your instruments, and coordinating with a buddy. This section provides a helpful introduction to these skills, but you must learn them by doing. You will learn the skills in pool-like conditions and then apply them in open water.

Using Entry Techniques

The four basic types of water entries for divers are wading, seated, feetfirst, and back roll. You need to learn when to use each type and the procedures for each. The objective of any entry is to get into the water as easily as possible without injuring yourself or losing any equipment. After you enter the water and are under control, switch from your regulator to your snorkel if you are going to swim or remain at the surface.

You can make open-water entries from the shore or from a boat. You can wade in from the shore, or you can enter from an artificial structure such as a dock, pier, or jetty. The conditions vary greatly, and so do the entry techniques. Surf may be a factor at some locations. Bottom conditions can range from smooth and soft to rough and firm. The bottom may slope gradually or steeply, and there

Head-to-Toe Buddy Check

Mask	In place, flush on face, no hair under skirt, not fogged
Snorkel	Attached to left side of mask
Scuba cylinder	Valve open all but a quarter turn, valve facing user's right side, height OK, snug in pack
SPG	Full pressure, not damaged or leaking
Scuba regulator	Primary and alternate air sources function OK, both buddies know the placement of the alternate air systems, regulator hoses oriented properly and free for use
Buoyancy compensator	Inflates and deflates properly
Weighting system	In place, proper amount of weight, right-hand quick release accessible, weights clear to drop, both buddies familiar with weight release system
Exposure suit	Properly positioned and zipped
Dive knife	Attached properly, easy to access, secured
Fins	Securely strapped or worn
Instruments	Watch, dive timer, or dive computer ready for use

may be holes and drop-offs. Plants, animals, and rocks may be present in the entry area. A good entry technique for one location may be inappropriate for another. It takes experience and knowledge of the area to determine an effective entry procedure.

You need to know some general techniques when making wading entries at an open-water site. A wading entry sounds simple, but keep in mind that diving equipment affects your center of gravity, your mobility, and your peripheral vision. You must walk backward or sideways when wearing fins.

If there is little or no surf, you can wade in, don your fins, and begin your dive. Breathe through your regulator and have your BC inflated partially. Shuffle your feet to detect holes and rocks and to chase away bottom-dwelling creatures (see figure 6.14). When the water reaches your thighs, lie down in the water and start swimming. In some areas with muddy bottoms, you should not wade because you can sink deeply into the mud and lose a fin when you try to extract your foot. When the bottom is firm and the water is calm, you can wade into the water without your fins and don them in the water. It is helpful to have information about the bottom conditions of a dive site.

In most areas where there is surf, you should don your fins before you enter the water, and you should not remove them until you are clear of the water after the dive. Time your entry to coincide with small waves (see the discussion of waves and surf in chapter 5). Keep all equipment in place, and breathe through your regulator. Deflate your BC because you need to duck beneath breaking waves when they are higher than your waist. If you inflate your BC, you will be unable to duck beneath the waves, and a large wave could lift and toss you. Hold your mask with one hand at all times; spread your fingers and curl them over the top of the mask so you can see. Keep your knees bent and shuffle sideways into the waves to minimize your profile to the moving water. Stop moving just before a wave hits you, allow the wave to pass, and then resume your shuffling until the water is deep enough for you to swim. Allow incoming waves to pass over you, and move through the surf zone quickly.

From a commercial dive boat, you can enter the water from the side or from a water-level platform at the back of the boat. Have all equipment in place, breathe

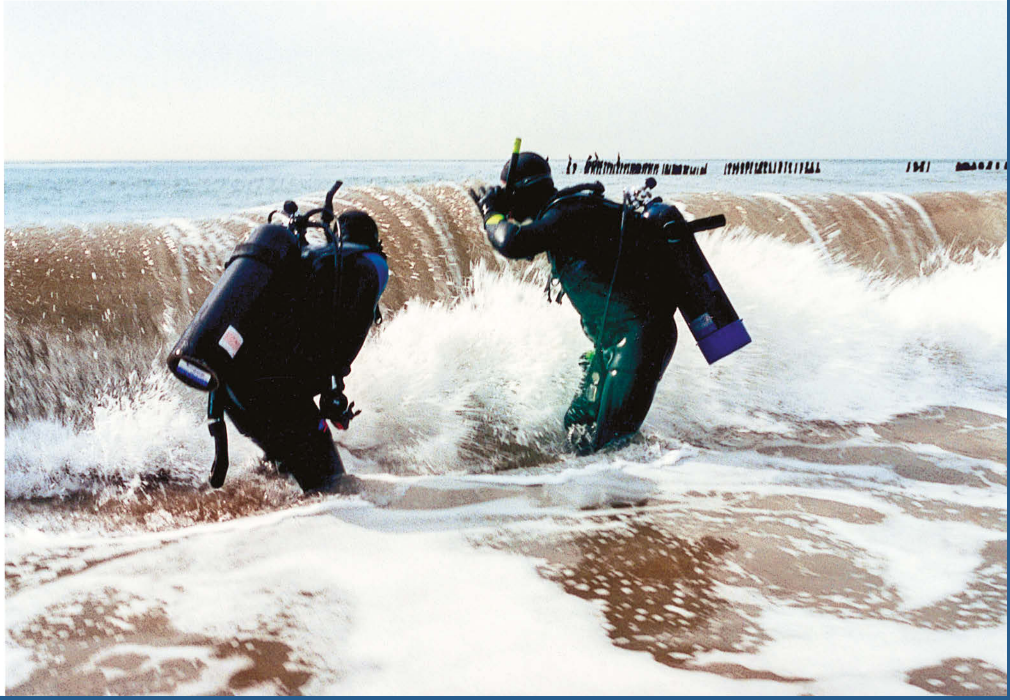

Figure 6.14 When entering surf, shuffle sideways (keeping knees bent), hold your mask, and lean into waves.

Courtesy of Fred Humphrey.

from your regulator, and hold your mask securely. Have any specialty items, such as a camera, handed to you after you are in the water; another option is to retrieve these items from an equipment line. When entering from the side of a boat, note the direction of boat movement. Wind will cause an anchored boat to swing from side to side. If you enter the water on the side of the boat when the boat is moving in that direction, the boat may pass over you after you enter.

You can do a controlled, seated entry from a dock, from a swimming platform on a boat, or from any surface where you can sit close to the water, whether the water is only a few feet deep or too deep for standing. With all equipment in place, turn and place both hands on one side of your body on the surface that you are seated on. Then lift yourself slightly, move your body out over the water, and lower yourself into the water. A controlled, seated entry is a simple and easy entry (see figure 6.15).

Use feet-first entries when the distance to the water is too high for a seated entry, such as when entering from a commercial charter boat. Two types of feet-first entries can be used: giant stride and feet together.

Use the giant-stride entry, shown in figure 6.16, when the distance to the water is about 3 to 5 feet (1 to 1.5 m) and you want to remain at the surface during the entry. Stand at the entry point with your equipment in place and your BC partially inflated. Observe the point of entry and make sure the area is clear. Hold your mask firmly with one hand, spreading your fingers so you can see. Look straight ahead while stepping out with one leg. The entry is a step, not a hop or a jump. If you step out as far as you can, your trailing leg follows automatically. Keep one leg extended forward and the other leg extended backward until you hit the water; then pull your legs together quickly to stop your downward momentum. As soon as you stabilize at the surface, turn and signal that you are OK, and then move away from the entry point so that the next diver can enter.

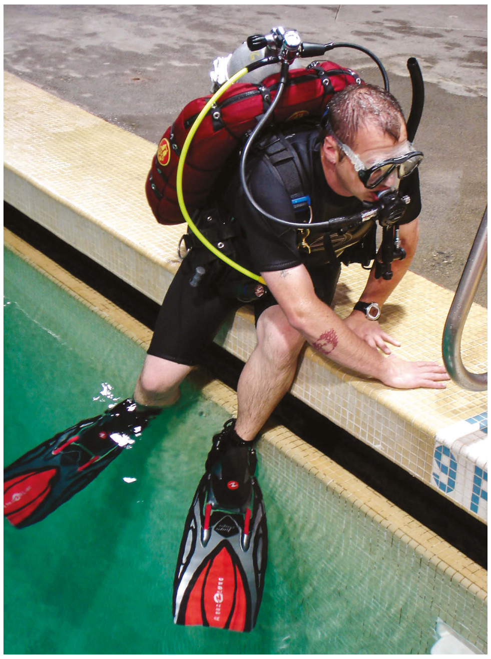

Figure 6.15 Entering the water from a seated position.

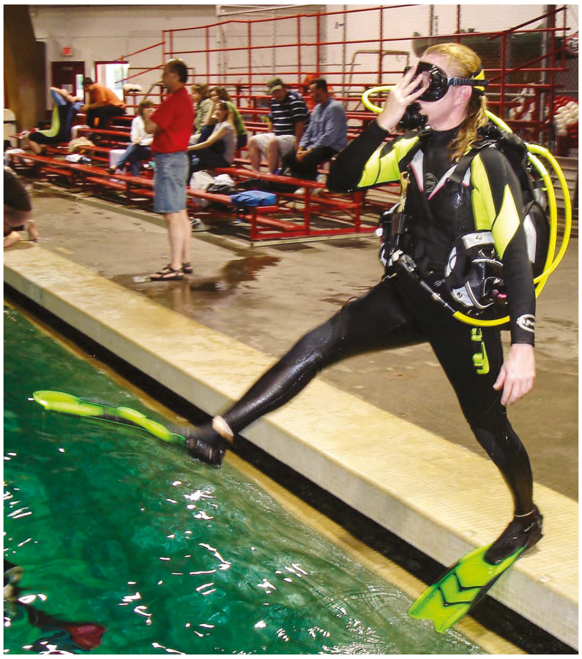

Figure 6.16 Giant-stride entry.

Use the feet-together entry to avoid discomfort from the impact when the distance to the water is too high for a giant-stride entry. The procedures are the same as for the giant-stride entry, except that after you step out from the entry point, you bring your legs together before you hit the water. A feet-together entry submerges you. After you bob back to the surface and stabilize, signal the next diver and clear the entry area.

A back-roll entry can be done from either a seated or squatting position. Use the seated back-roll entry from a low, unstable platform, such as a small boat. You can also use a back-roll entry when the distance to the water is too high for a regular seated entry and the platform is too unstable for you to stand. To do a seated back roll, sit with all equipment in place and your back to the water. Move your bottom to the very edge of the surface on which you are sitting. Have someone make sure the entry area is clear. Hold your mask with one hand and your mask strap with the other hand. If you do not hold the mask strap, the force of the water may wash it up over your head, and your mask may fall off. Lean backward to begin the entry. Hold your knees to your chest as you roll backward; if you don't, you may clip your heels on the edge of the surface you were seated on (see figure 6.17). You are likely to do a backward, disorienting somersault in the water with this entry. You can reorient yourself when you bob back to the surface.

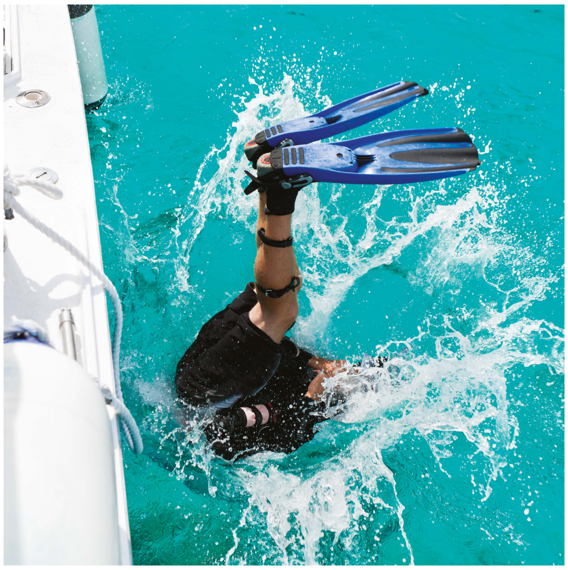

Figure 6.17 **Entering with legs extended risks clipping the heels on the boat. Tuck the legs into the chest instead.**

Christian Wheatley/Getty Images

A squatting back roll is used when there is no suitable surface on which to sit. The thin side of a small, rocking boat is a good example of an unsuitable seat. For a squatting back-roll entry, prepare yourself while sitting on an adequate seat immediately adjacent to the entry area. Make sure the entry area is clear, then stand partially, turn your back to the entry area, and literally sit down into the water (see figure 6.18). Pull your knees to your chest as you enter so that you do not catch your heels.

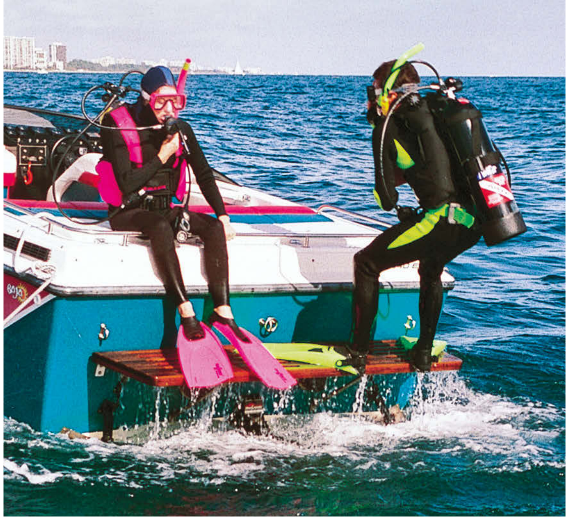

Figure 6.18 **Squatting back-roll entry.**

Surface Snorkeling With Scuba Equipment

You may use your snorkel for surface swimming in order to conserve the air in your scuba tank. This procedure is used in areas where divers need to swim from shore to reach the dive area. If you use air from your cylinder for surface swimming, you will reduce the amount of time you can remain underwater. When snorkeling while wearing scuba equipment, you should inflate your BC partially. Excessive air in your BC combined with the weight of your scuba tank

Three Important Scuba Diving Rules

1. Equalize pressure in your ears early and often.
2. Always breathe; avoid breath holding.
3. If the scuba regulator is out of your mouth, exhale continuously.

can cause you to roll sideways when swimming facedown. It is better to swim on your back or side; these positions allow wider kicks than when swimming facedown, and the tank weight is not a factor. If you swim on your back, you can breathe without the snorkel. Swim beside—not behind or in front of—your buddy. Move slowly and steadily, pacing yourself. If you feel short of breath after snorkeling, rest before beginning your descent.

Recovering and Clearing the Regulator

When you are in the water wearing scuba equipment and have removed your regulator second stage from your mouth—such as when snorkeling or talking at the surface—you need to recover it from behind your right shoulder (where it falls naturally). The second stage will have water inside, and you must clear the water before you breathe. If you start with the regulator in your mouth, you need to place the regulator in the water. The task is more difficult than it sounds because if you place the second stage in the water with the mouthpiece facing up, the regulator free-flows. Place the second stage in the water with the mouthpiece facing downward to prevent free-flow.

Two methods can be used to recover the regulator second stage from behind your shoulder. The most popular technique is the sweep method. Lean to the right side so gravity swings the second stage away from you. Reach back with your right hand until you touch the bottom of your scuba tank, and then extend your arm and sweep it forward in a large arc. The hose will lie across your arm, where you can retrieve the second stage easily.

The second method of recovering the second stage is the over-the-shoulder reach. Reach back toward your regulator first stage with your right hand while lifting the bottom of your scuba tank with your left hand. Grasp the second-stage hose where it attaches to the first stage and follow it down to the second-stage end. Some divers find this recovery method difficult or impossible.

Clearing the regulator can be as simple as exhaling into it. As long as the exhaust valve is at the lowest point, the water inside the second stage will be displaced. If the exhaust valve is not at the lowest point when you exhale, only part of the water may be exhausted, and you may inspire some water when you inhale. To avoid choking on inhaled water, find out where the exhaust valve is on your regulator, be sure to make it the lowest point when you clear your regulator, and inhale cautiously after clearing the regulator.

Another way to clear a regulator second stage—the purge method—involves using low-pressure air to clear the water from the chamber. If you insert the mouthpiece into your mouth and depress the purge, you may blow water down your throat. You can prevent this in two ways. One way is to depress the purge lightly as you place the bubbling regulator in your mouth, and to use back pressure from your lungs to keep water out of your mouth and throat. The other method is to place the regulator into your mouth, block the opening with your tongue, and depress the purge momentarily to clear the chamber. Either method is acceptable. If you are purging the regulator while placing it in your mouth, release the purge the moment the regulator is in place to avoid overinflating your lungs. Develop the habit of blowing a continuous stream of tiny bubbles any time the regulator is out of your mouth to avoid holding your breath while breathing compressed air.

Performing Snorkel–Regulator Exchanges

You must be able to switch from breathing from your snorkel to breathing from your regulator and vice versa. When you prepare to descend for diving, you exchange your snorkel for your regulator; when you surface after a dive, you exchange your regulator for your snorkel. Do both exchanges with your face in the water. Take a breath, exchange one mouthpiece for the other, and clear the new breathing device.

Inflating and Deflating the BC

You should be familiar with two ways to inflate a BC and two ways to deflate it. The easiest and most commonly used means of inflation is the low-pressure inflator, which is used to add air to the BC in short bursts. Inflating the BC for several seconds can cause serious buoyancy control problems if the inflator valve sticks. By adding air a little at a time, you have better control of your buoyancy.

If your low-pressure inflator or your integrated second stage develops a problem while you are diving, you may need to disconnect the low-pressure hose. When you disconnect the hose, the low-pressure inflator no longer functions, so you have to control buoyancy by orally inflating the BC. The mouthpiece of a BC is usually more elaborate than the mouthpiece of a skin diving vest. To help keep water out of a BC, the mouthpiece has a purge so that the water inside can be cleared before you open the valve to the BC. To clear the BC mouthpiece and orally inflate the BC, follow this procedure:

1. Insert the mouthpiece into your mouth.
2. Exhale a small amount of air into the mouthpiece to clear it.
3. Keep the mouthpiece in your mouth.
4. Depress the manual inflator-deflator valve.
5. Exhale into the BC.
6. Release the manual inflator-deflator valve.
7. Repeat the procedure until you achieve the desired amount of buoyancy.

You can use the same bobbing technique at the surface that you learned for the skin diving vest (see the previous explanation in this chapter), but oral inflation procedures are different underwater. To inflate a BC orally underwater, follow these steps:

1. Grasp the inflation valve with your left hand, and grasp your regulator second stage with your right hand.
2. Take a breath.
3. Insert the BC mouthpiece into your mouth.
4. Clear the mouthpiece.
5. Exhale most of the air in your lungs into the BC.

6. Save enough air to clear the regulator, which fills with water when you remove it from your mouth.
7. Repeat the procedure until you achieve the desired amount of buoyancy.

You can deflate your BC with the manual inflator-deflator valve or with a dump valve. If you do not use a dump valve, you must open the deflator valve at the lower end of the BC hose and hold it higher than the highest point of the BC. Using a dump valve is more convenient than using the deflator valve. Note that you can deflate your BC only when the exhaust port is the highest point. Air cannot escape if you try to deflate a BC in a horizontal or inverted position, so you need to be in an upright position to deflate your BC.

Testing Buoyancy

Similarities exist between buoyancy testing for skin diving and buoyancy testing for scuba diving, but there are also some important differences. Buoyancy varies more when you scuba dive than when you skin dive. The volume of air in your lungs varies more in scuba diving. You dive deeper, so suit compression affects your buoyancy more. As you use air from your scuba tank, your buoyancy changes. You need to sense buoyancy changes quickly to maintain control of your buoyancy.

Your initial buoyancy test for scuba diving can be the same as your buoyancy test for skin diving. Begin by testing your BC at the surface. Inflate it fully and then deflate it. Make sure the low-pressure inflator and all deflation valves function correctly. With your regulator in your mouth and your BC completely deflated, relax and breathe slowly. When your lung volume is high, you should remain at the surface with your eyes just below the surface–water interface. If you sink with your lungs full of air, you need to remove some weight. When you exhale completely, you should sink. If you cannot sink after a complete exhalation, you need additional weight. Test your buoyancy while you are close to your point of entry, and correct any buoyancy problems before you attempt to dive.

As you descend in open water, you may need to add air to your BC in order to maintain neutral buoyancy. Strive to maintain neutral buoyancy continuously. Add air to your BC in small amounts, and test your buoyancy by stopping all motion and observing what happens. If you are sinking, inhale and add air to your BC. If you are rising in the water, exhale and release air from your BC. With experience you will know when to add air to your BC, when to vent air from your BC, and how much air to add or release.

You can learn buoyancy control in a swimming pool by assuming a rigid, face-down position with your arms at your sides. If your buoyancy is correct, a slow, full inhalation raises your shoulders while your fin tips remain on the bottom; a slow, complete exhalation causes your shoulders to sink. Some people call this buoyancy evaluation **diver's push-ups** (although you do not use your hands). If you do not rise with a full inhalation, you should add a small amount of air to your BC and try again. Figure 6.19 illustrates a diver doing diver's push-ups.

As a scuba diver, you need enough weight to get down at the beginning of a dive and enough weight to allow you to remain in control when you ascend at the end of the dive. The precise weight for scuba diving is the amount you need to hover at a depth of 15 feet (4.6 m) with 300 psi (20 atm) of air in your tank and no air in your BC. This amount may overweight you slightly at the beginning of a dive. Test your buoyancy at a depth of 15 feet at the end of a dive to see if you need to make an adjustment for your next dive.

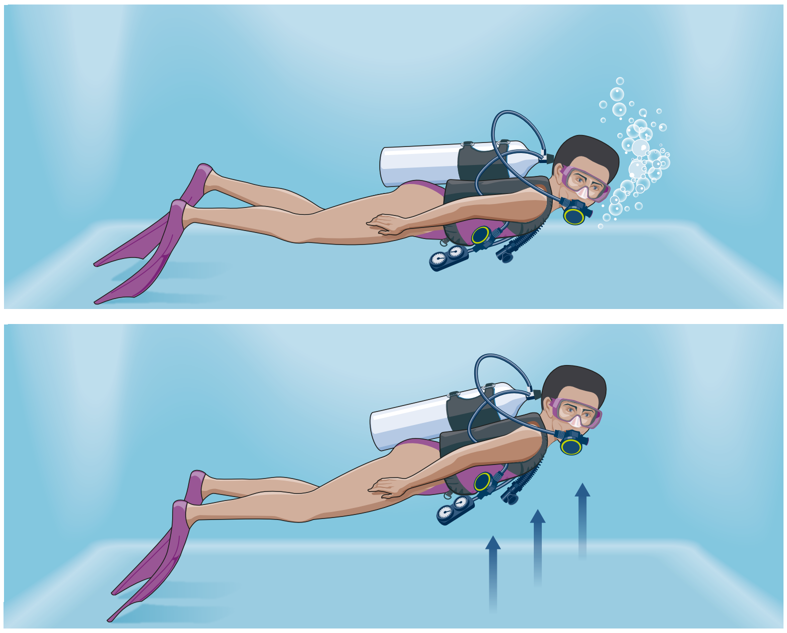

Figure 6.19 Executing proper diver's push-ups ensures good buoyancy control.

Controlling Buoyancy

No skill identifies a scuba diver's ability as much as buoyancy control. The ability to finely control buoyancy is important for safety, for enjoyment, and for the welfare of the environment. When your buoyancy is out of control, a hazard exists for both you and the environment around you.

With your buoyancy adjusted to the point where you can pivot slightly up and down on your fins on the bottom (in a pool) while inhaling and exhaling, push yourself about 2 feet (0.6 m) off the bottom and remain motionless. You may or may not remain in a horizontal position, but that is not important. Maintain your depth by controlling your average lung volume, but remember to breathe continuously. If you are sinking, keep more air in your lungs. If you begin rising, reduce the amount of air in your lungs. With practice, you will be able to hover motionless just off the bottom.

To demonstrate mastery of hovering, assume a vertical position in the water. Cross your legs at the ankles, grasp one wrist with the opposite hand, and remain motionless. Find an eye-level reference and develop the ability to hover motionless while upright in the water. Pay attention to the buoyancy effects of breathing. Once you master hovering both horizontally and vertically, you will realize several benefits. Your air will last longer, swimming will not tire you as much, you will need fewer buoyancy adjustments, and you will do less damage to the diving environment. Your buoyancy control continues to improve with practice until you become a highly experienced diver.

A useful skill for maintaining neutral buoyancy during ascent is a BC venting technique called an **open-valve ascent**. When you ascend, air in your BC expands and affects your buoyancy. You must release the expanding air to control buoyancy. If you vent the air from time to time, your buoyancy is in a constant state of change: You may not release enough air, or you may release too much. The open-valve ascent is a better alternative. Hold the BC inflator-deflator valve just below the level of your shoulder, point the mouthpiece of the inflator-deflator valve downward, and open the deflator valve. Air cannot escape because the opening is lower than the exhaust port on the BC. While you hold the mouthpiece downward and the deflation valve open, raise the inflator-deflator slowly until air just begins to bubble from the mouthpiece. Bubbling occurs when the mouthpiece and the exhaust port on the BC are at the same level. If you hold the inflator-deflator valve at the bubbling level while ascending, the expanding air inside the BC bubbles out through the open valve, and your buoyancy remains constant. If you need to release additional air to control buoyancy, raise the inflator-deflator slightly; if you need to decrease the amount of air you release, lower the assembly slightly.

Using the Dive Flag

You should display the appropriate dive flag or flags when diving. Authorities in some areas require you to use a flag, and flag use is encouraged in all areas. Follow these conventions when you use a dive flag: Display the flag only when divers are in the water, and surface within 100 feet (30 m) of your flag—the closer the better. The flag does not guarantee your safety from boaters, who are supposed to give the flag a berth of at least 100 feet; however, the flag does serve as a signal to many boaters who recognize that the flag means there are divers in the vicinity. Figure 6.20 shows a boat anchored with a dive flag indicating there are divers in the area.

If you support your flag with a surface float, you need to tow your float to the area where you wish to dive. In areas lacking underwater plants, you can tow the float while you dive. In areas with plants, you need to anchor the float to keep it from drifting away, and you must navigate back to the flag at the end of your dive. When entering through surf, you should tow the surface float behind you. Push the float ahead of you when exiting through surf.

Figure 6.20 Surface as close as possible to the dive flag to keep you safe from other boaters.

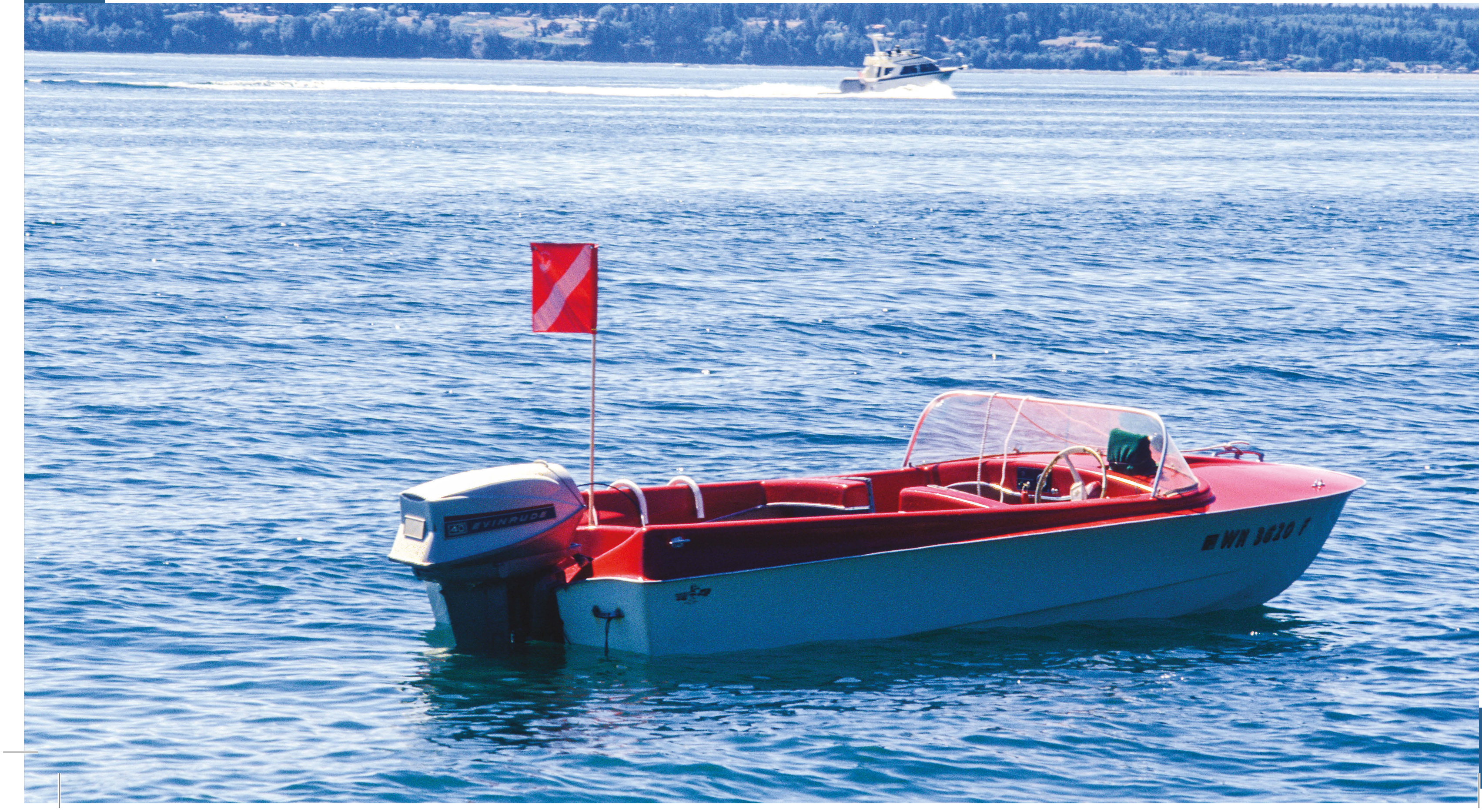

Descending

Scuba descents are different from skin diving descents. You usually descend headfirst as a skin diver, but scuba divers descend feet-first. A feet-first descent allows better control of buoyancy, provides better orientation, allows buddies to maintain contact during descent, helps prevent the swallowing of air, and allows easier pressure equalization of air spaces.

Descending involves several procedures, which seem complex at first but become routine with practice. The process of descending includes preparing to descend, initiating the descent, and completing the descent.

1. Preparing to descend
    - Check your instruments. Orient yourself at the surface and set a reference on your compass. Be sure your underwater timer and depth gauge are zeroed.
    - Confirm that your buddy is ready to begin.
    - Exchange your snorkel for your regulator.
    - Pressurize your ears slightly to begin the equalization process (see chapter 3).
    - Hold your BC deflator valve in your left hand.
    - Give or acknowledge the signal to descend.
2. Initiating the descent
    - Begin the descent by venting your BC. It is better to do this with the dump valve than by holding the deflator valve above your head. Hold the inflator-deflator valve in your left hand throughout the descent so you can add or release air from your BC at any time. Exhale fully to help get started downward.
    - Breathe shallowly for the first 10 feet (3 m).
    - Equalize pressure in your ears about every 2 feet (0.6 m) for the first 15 feet (4.6 m). If you experience equalizing problems, ascend a few feet to reduce the pressure, equalize again, and descend again. Exhale some air into your mask to prevent a mask squeeze. (See chapter 3 for a description of pressure equalization.)
    - Keep your fins still while you descend so that you do not stir up silt on the bottom.
    - Control your rate of descent by the average amount of air you keep in your lungs. When you begin sinking while your lungs are full, add a short burst of air to your BC to regain neutral buoyancy.
3. Completing the descent
    - Remain with your buddy throughout the descent.
    - Avoid contact with the bottom.
    - Hover above the bottom, level off to a swimming position, agree on a direction with your buddy, and begin your exploration.

Two types of descents are performed in open water: **reference descents** and **nonreference descents**. A reference descent is one that you control by following a line or the slope of the bottom. You should do a reference descent whenever possible. A descent made vertically in water without any external reference is a nonreference descent, which is more difficult to control than a reference descent.

Clearing the Mask

To clear a mask while scuba diving, you need to develop the skill of breathing through your mouth with your nose exposed to water. With concentration and practice, you can master this skill quickly. First, try inhaling through your mouth and exhaling through your nose. Next, try inhaling and exhaling through your mouth. If you feel any water going up your nose, exhale immediately to keep the water out.

To practice clearing a mask, you need to flood it with water. This is not difficult, but a few tips make it easier. If you exhale lightly while tilting the mask forward on your face to break the seal at the top, the mask releases from your face easily, and the air escapes at the highest point. When you reseat your mask to begin clearing it, be sure to hold back any strands of hair with one hand while you reseat the mask with the other. Hair under the mask causes leakage. If you are wearing a hood, make sure that the hood is clear of the mask before you attempt to clear water from the mask.

Buddy Diving

You should always dive with a companion, using the **buddy system**. A buddy provides reminders and assistance and sees things that you might not see. Buddies inspect each other's equipment, provide feedback based on observations, and work as a team. Dive buddies should remain close enough to each other that one can immediately assist the other in the event of an emergency. The more turbid the water and the greater the depth, the closer buddies should remain to each other. During your training, you should strive to remain within touching distance of your buddy at all times; learn to keep track of your buddy. Maintaining contact with a dive buddy in open water is not difficult when you follow a few standard procedures.

Agree on a position relative to one another and maintain that position as much as possible. That way, your buddy will know where to look for you, and you will know where to look for your buddy. The best dive team configuration is side by side; the least desirable is for one diver to be above and behind the other. Looking up and back is difficult when you are hampered by your equipment.

Agree on a direction of movement. Both parties should maintain that direction until both agree to proceed in a different direction. When you follow this practice, there are fewer directions to consider if you and your buddy become separated.

Confirm your buddy's position every few seconds. If you scan the areas ahead from side to side while swimming, you should get a glimpse of your buddy each time you turn your head in your buddy's direction.

When visibility is poor, physical contact can keep you and your buddy together. Holding hands is appropriate. Or you can use a short line—a **buddy line**—to keep in contact with each other. If you become separated from your buddy underwater, look for your buddy for up to one minute. If you are unable to locate your buddy, ascend slightly and turn in a circle while looking for bubbles. The visibility is often better a few feet above the bottom than it is at the bottom. If you do not see your buddy's bubbles, ascend to the surface and wait for your buddy, who should also duplicate this procedure. When you have reunited at the surface, you can descend again and continue your dive. Obviously, it is better to remain together underwater than to surface to reunite.

The Buddy System

Teaming up with a qualified partner is appropriate for activities that involve risk, such as swimming, rock climbing, and skin and scuba diving. Safety is the primary purpose of the buddy system, but having a dive buddy also makes diving more enjoyable. The following list describes the duties and responsibilities of a good dive buddy:

- Helps to plan the dive in advance and after arrival at the dive site
- Reviews signals and emergency procedures
- Assists with the donning of the scuba unit
- Inspects equipment
- Maintains contact while diving and follows procedures for reuniting if separation occurs
- Provides reminders concerning depth, direction, time, air pressure, and ascent rate
- Points out items of interest
- Identifies problems that you are unaware of, such as an air leak
- Provides reassurance and assistance as needed
- Summons additional assistance as needed
- Provides first aid as needed
- Assists with exits and the removal of the scuba unit

If you are unable to relocate your buddy underwater and your buddy does not surface promptly, look at your surroundings to mark your position so that you will know the approximate location where your buddy was last seen. If someone is overseeing the diving operations, notify this person that your buddy is missing so that a search can be initiated. If you are alone, try to locate bubbles that could indicate your buddy's position.

Underwater Swimming

When you swim along the bottom, your fins can raise a cloud of silt that harms the environment and reduces visibility. Silting is more of a problem when you are overweighted because the excess weight angles your fins toward the bottom when you swim (see figure 6.21). The first step in reducing silt is to weight yourself properly. In areas where the bottom silt is thick, add air to your BC to make yourself slightly buoyant underwater. The buoyancy forces you to swim at a slight downward angle and keeps the thrust of your fins directed upward. Another way to reduce silt is to remain far enough from the bottom to keep from disturbing it.

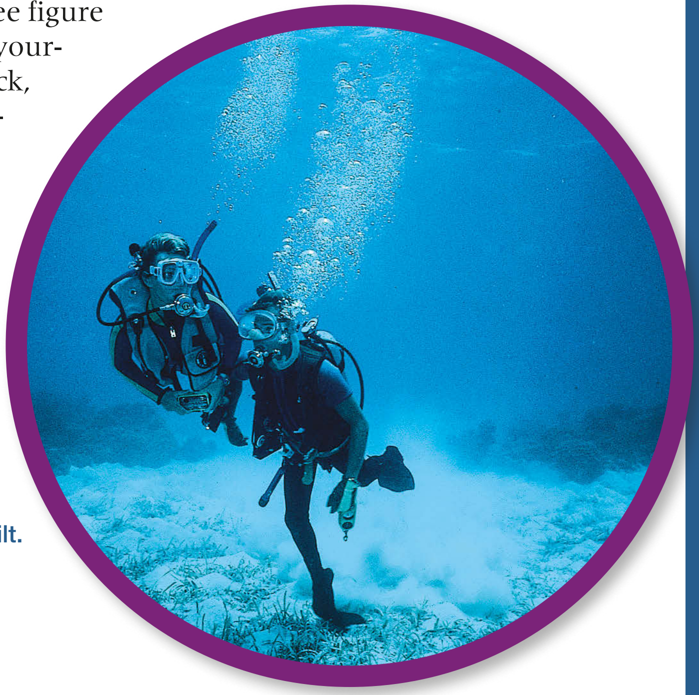

Figure 6.21 Overweighting can cause divers to stir up a cloud of silt.

Finally, consider changing your kick if you boil up silt with your kick strokes. You should also consider redistributing your weight by placing a small portion of it higher on your body to achieve level trim as described in the "Mobility Adjustments" section of chapter 3.

When you kick something while swimming, you must overcome the tendency to want to get away from whatever you kick. Your kicks are strong and can damage the environment or another diver. As soon as you feel something with a fin, stop kicking, look back to see what you have hit, and maneuver yourself clear before proceeding.

Navigating

To find your way underwater, you can use natural navigation and compass navigation. You can best determine your relative position with a combination of both types of navigation.

With natural navigation, you use your natural surroundings to determine where you are. Light, shadows, plants, formations, water movement, depth, and other indicators can help you navigate. As you move, note your surroundings. Observe which way you are going relative to the movement of the water, the sand ripples on the bottom, the depth contour, and the angle of the sun. By noting natural aids to navigation, you can find your way underwater.

A dive compass increases the accuracy of navigation. Figure 6.22 shows two types of dive compass. You need to be able to set a direction (called a **compass heading**) and to determine which way you are going relative to the directional reference you have set. Your compass should have a reference line, called a lubber line, that you point in the direction of travel. The north-seeking needle or card of the compass establishes a position relative to the lubber line as long as you hold the compass in a level position. Many dive compasses have index marks on a movable bezel. You set the index marks to indicate the heading. Electronic compasses allow you to set a heading and use arrows to indicate the direction that you need to turn to remain on course.

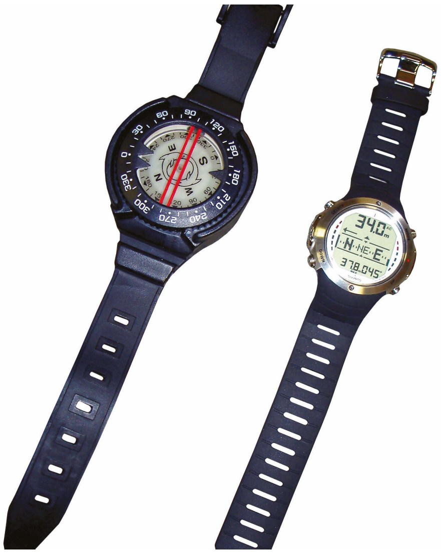

Figure 6.22 Traditional dive compass (left) and electronic dive compass (right).

To go in the direction set on the compass, you must hold the compass so the lubber line is directly in line with the centerline of your body. If the lubber line points to one side, you will not be on course even though the north reference is at the correct point on the dial. Figure 6.23 shows a compass lock position.

A **compass course** is a series of headings that leads to a destination. Many types of compass courses may be used. One frequently used course is the **square compass course**. To navigate a square course, set your initial heading and proceed in that direction for a given distance, which can be measured by time, tank pressure, or fin kicks. Stop and turn 90 degrees to the right while continuing to keep the lubber line aligned with the centerline of your body. Note the

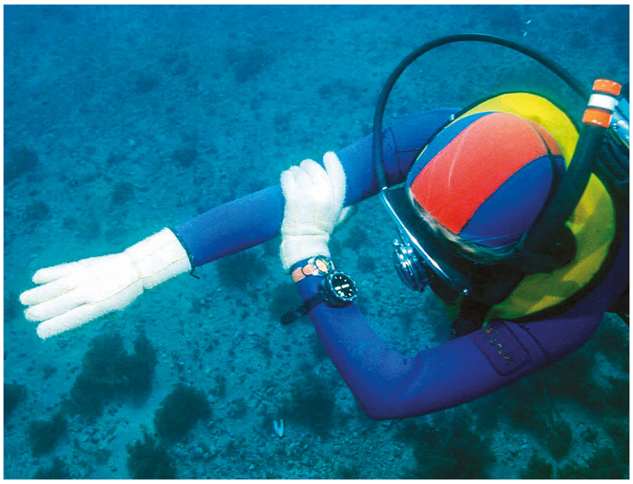

Figure 6.23 The compass-lock position.

relative position of north on the compass, and proceed in the new direction for the same distance as you covered on the first leg of the course. Stop again, and turn another 90 degrees to the right. Note the position of north on the compass, which should be opposite your initial heading. Proceed along the third leg of the course for the same distance as before. Stop once more, turn again 90 degrees to the right, note the relative position of north, and follow the new heading back to your starting point.

Divers also frequently follow a **reciprocal compass course**—an out-and-back course. Set the initial heading on the compass. At the midpoint of the dive, turn 180 degrees until north on the compass is directly opposite the original heading. Then follow the reciprocal heading back to your starting point. Advanced scuba diving courses provide additional compass navigation training.

If you do not know precisely where you are when the end of a dive approaches, you may need to surface, find a reference for your exit, and set a compass heading that leads directly to the end-of-dive location. Be especially careful if you surface more than 100 feet (30 m) from your dive flag.

A compass provides correct directional reference information when it is not affected by nearby objects. Metal objects, other compasses, and electrical fields within a couple of feet of a compass can cause it to deviate from its correct reading. Keep metal, magnets, dive lights, and other compasses away from your compass to help ensure accuracy.

Monitoring Instruments

Most diving instruments are passive; that is, they do not provide information unless you look at them. Some instruments emit an audible beep, but most require observation to provide information. Develop the habit of checking your instrumentation frequently while diving so that you can control your depth, dive time, and direction. This will also help you avoid running out of air. You should be able to accurately estimate your tank pressure at any time during training. If you cannot estimate the pressure within 300 psi (20 atm) at any time, you need to monitor your SPG more frequently.

When you are planning to dive in open water, you should look at your instruments when you assemble your equipment. Look at your instruments again when you inspect your equipment. Look at your instruments again before you descend and again while you are descending. Refer to your compass for directional reference before you begin moving underwater. Monitor your gauges every few minutes while diving, and compare your air pressure with your buddy's several times during a dive. At any given time during a dive, you should be able to accurately estimate your depth, your dive time, your direction, your tank pressure, and your buddy's tank pressure. If you cannot do this, you need to improve your instrument-monitoring skills.

Using Scuba Diving Hand Signals

Scuba divers use several hand signals that are not used for skin diving. The signals specific to scuba diving relate to air supply. Learn and use the standard hand signals shown in figure 6.24. Remember to display hand signals clearly and deliberately and also to acknowledge the hand signals you receive.

Ascending

Scuba ascents are different from skin diving ascents, although there are a few similarities. The procedures you need to learn become automatic with practice and experience.

To initiate an ascent, one member of a buddy team gives the ascent signal, which the other acknowledges. Always obey the ascent signal. Prepare to ascend by noting your time, depth, and remaining air. Locate your BC inflator-deflator assembly, and hold it in your left hand. Begin the procedures for the open-valve ascent previously described (see "Controlling Buoyancy").

Begin ascending slowly with your buddy while breathing continuously. Monitor your depth gauge and keep tabs on your buddy. The maximum rate of ascent is 0.5 feet (0.15 m) per second, which is quite slow. Some instruments warn you when your rate of ascent is too rapid. You need training, practice, and awareness to avoid exceeding the maximum rate of ascent.

Stop and decompress (outgas) when indicated by your dive computer. Whether or not you make a "deep stop," you should stop for one to three minutes at a depth of 15 feet (4.6 m) to help prevent decompression sickness (DCS). The procedures for decompression are described in chapter 7. As you ascend, look up and around. Extend one hand above your head for protection against overhead obstacles. Make one full rotation to view the surrounding area as you near the surface. When you reach the surface, make another rotation to view the area, and then inflate your BC to establish buoyancy. Exchange your regulator mouthpiece for your snorkel.

Handling Equipment

Sometimes you will need to remove, adjust, and replace scuba equipment while you are in the water. You may need to remove equipment to exit the water onto a boat, to make an adjustment, or to free the equipment from an entanglement. With training and practice, you should be able to handle your equipment easily.

Removal of the scuba unit is easy because it is similar to removing a coat. Open the releases, slip your left arm free, swing the scuba tank forward under your right arm, hold the scuba unit with your left hand, and pull your right arm free. Freeing your left arm is easier if you insert your hand and wrist through the armhole of the BC first and remove it hand first (instead of pulling your arm through first). If you are at the surface in water too deep for you to stand when you want to remove your scuba unit, you should remove your weight belt first and place it on a surface float or support station. Don the scuba unit in the water according to the in-water donning procedures previously presented in this chapter.

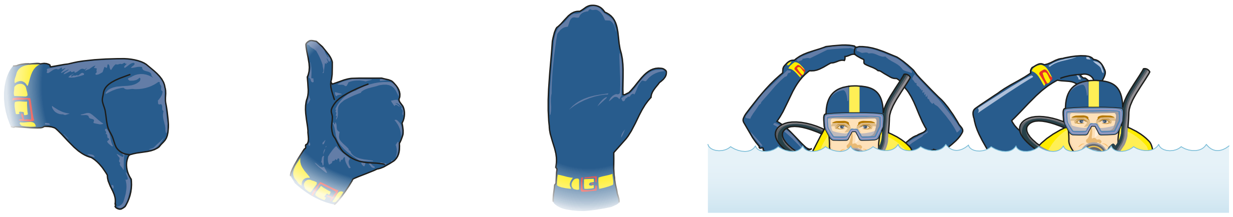

1. Descend

2. Ascend

3. Stop

4*a*. OK

4*b*. OK

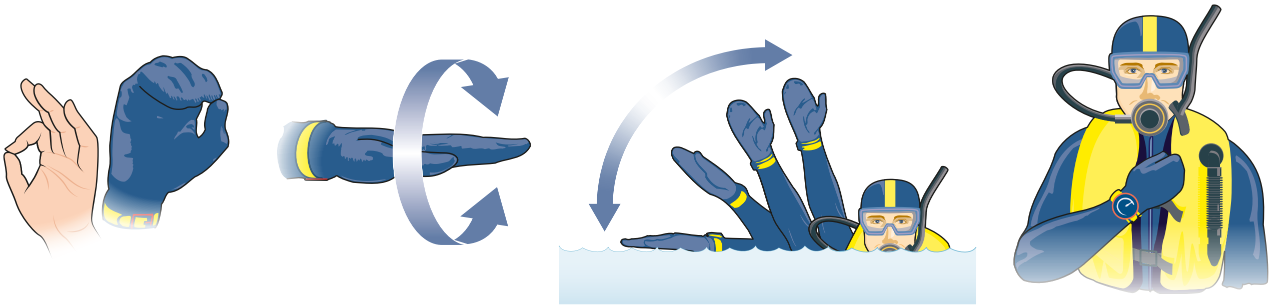

4*c*. OK

5. Something's not right

6. Emergency

7. Low on air

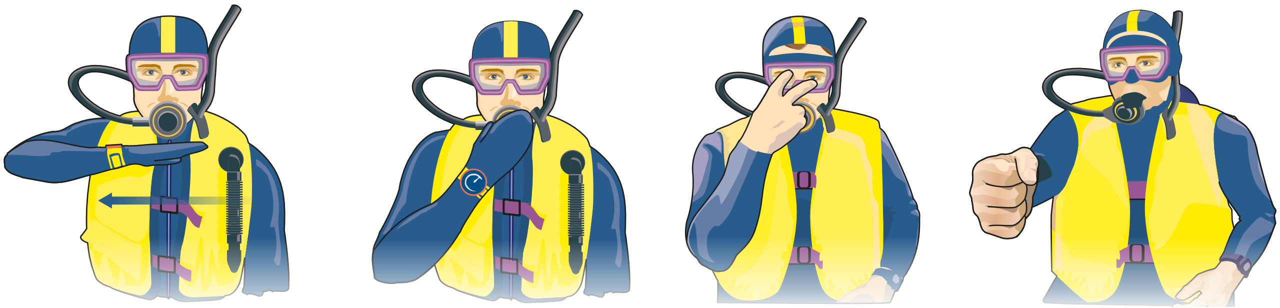

8. Out of air

9. Give me air

10. Look

11. Danger

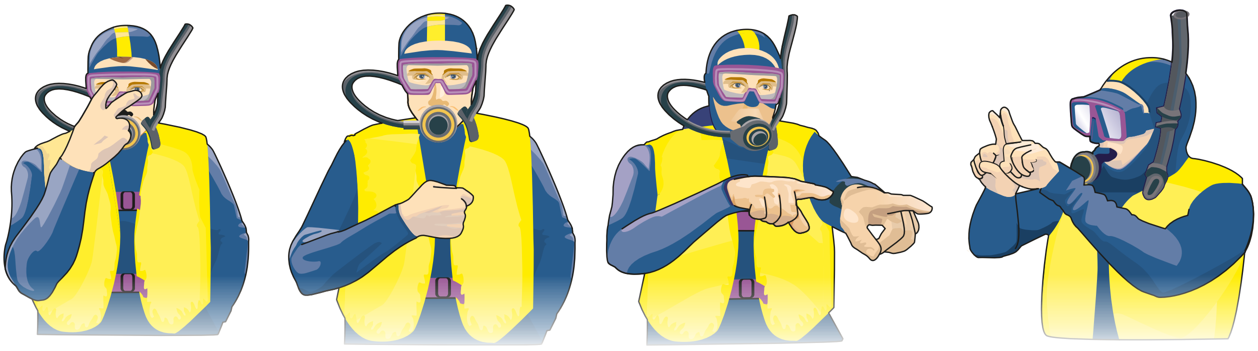

12. Watch me
(finger pointing to chest)

13. You lead, I follow

14. Get with your buddy

Figure 6.24 Scuba diving hand signals.

Using Exit Techniques

The technique you use to exit the water depends on the situation. To exit from shallow water in a swimming pool, begin by removing your weight belt, tank, and fins, in that order. Carefully place the equipment on the side of the pool; then climb out by using the ladder or lifting yourself up onto the edge of the pool.

To exit from the deep end of a swimming pool (as shown in figure 6.25), begin with your BC inflated partially. If you exit on a ladder, grasp the ladder with one hand and remove your fins with the other; maintain contact with the ladder at all times. Place your fins on the edge of the pool, or slide the heel straps over your wrists; then climb the ladder to exit the water. Clear the exit area at once, and take your fins with you. Remove your scuba system and weight belt when you are well clear of the exit area.

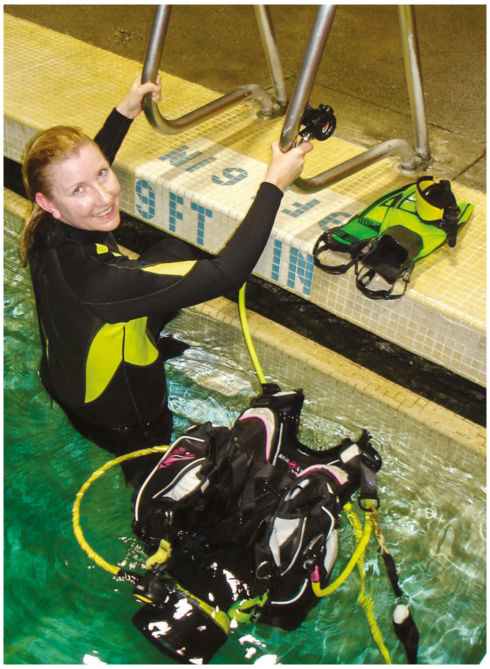
Figure 6.25 Exiting from the deep end using a ladder.

To exit from deep water without a ladder, begin by removing your weight belt and carefully placing it out of the water. Remove the scuba unit and use one hand to trap the regulator hose against the surface onto which you will exit. Use your other hand to remove your fins, and place them out of the water. Place both hands on the exit edge. With the regulator hose trapped under one hand, lower yourself to about chin level in the water while you extend one leg forward and one leg backward. Pull yourself upward with your arms, and pull your legs together forcefully in a strong scissors kick to provide upward momentum. Pull with your arms until you are far enough out of the water to push downward and lift yourself from the water (see figure 6.26). Immediately after your exit, turn around and pull your scuba unit from the water carefully.

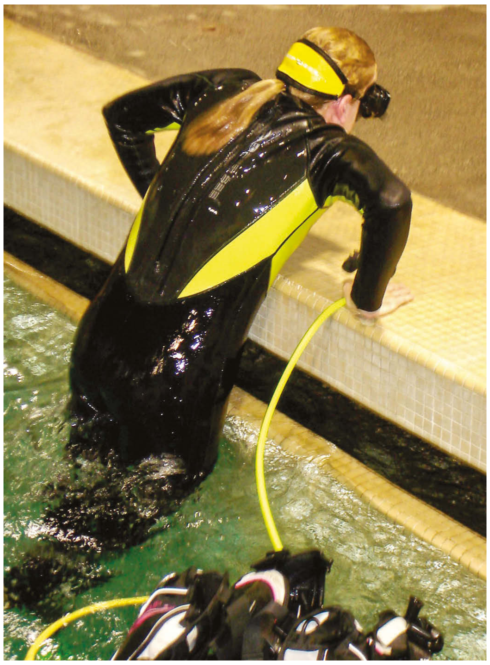
Figure 6.26 Exiting from the deep end without a ladder.

If you are boarding a boat that has a ladder, you should keep your tank and mask on and keep your regulator in your mouth in case you fall back into the water (see figure 6.27). Use the ladder exit technique described for exiting the deep end of the pool. Maintain contact with the ladder at all times when you are in the water. If the boat has a platform at the rear, you usually swim onto the platform, remove your fins, and then stand on the platform and board the vessel.

Techniques for wading exits in open water vary with the environment. You should usually wear all your equipment until you are clear of the water. Shuffle your feet along the bottom while moving backward. Surf exits require training and practice. Stop outside the breaking waves and evaluate the surf. Approach the surf zone with your regulator in your mouth and your BC deflated. Hold your mask continually in the surf zone. Follow a breaking wave, and allow additional waves to pass over you until the water is only a couple of feet deep. If the surf is mild, you can stand at that point and back out of the water. If the surf is strong, swim until you can crawl; then crawl clear of the water (see figure 6.28).

When you are clear of the water, work with your buddy to remove your fins. The buddy system is in effect even when you are out of the water.

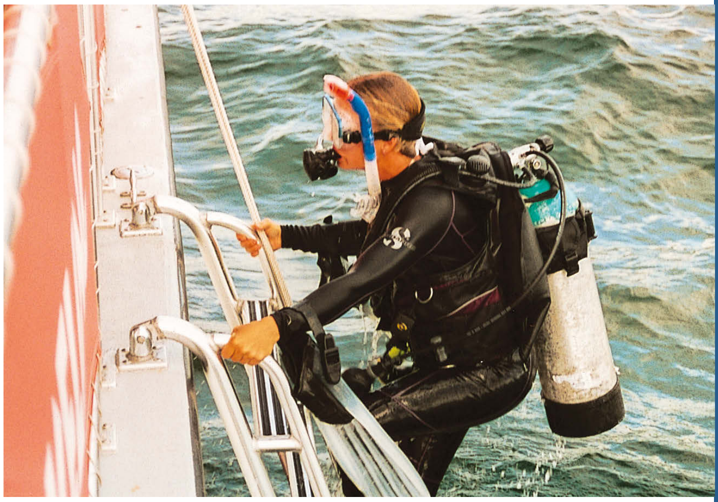

Figure 6.27 Boarding a boat with a ladder.

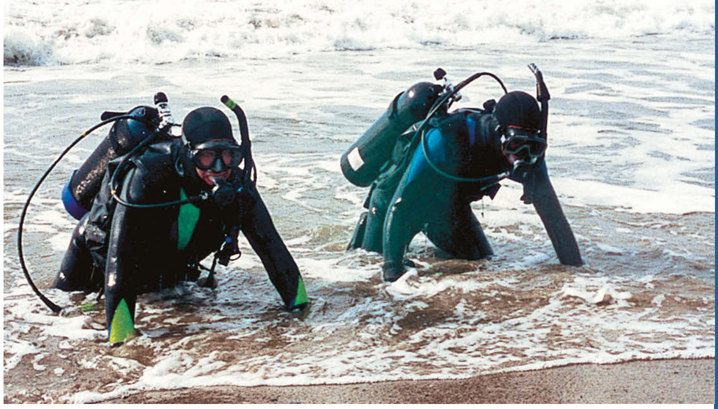

Figure 6.28 Using a crawling exit in strong surf.

Courtesy of Fred Humphrey.

Disassembling Equipment

The first step in disassembling your scuba equipment is to turn off the air by turning the valve in a clockwise direction. Release the pressure in the hoses by depressing the purge on the regulator second stage. Keep the purge depressed until you bleed all the air from the system. Next, disconnect all hoses that are connected to the scuba unit: the low-pressure inflator, the extra second stage, and the SPG. Loosen the yoke screw and remove the regulator from the tank. Dry the first-stage dust cover thoroughly and replace it. Loosen the tank band and remove the BC from the scuba tank. Be sure to rinse and care for your equipment as soon as possible (following the procedures described in chapter 4).

Managing Physiological Problems

If you do everything you are trained to do as a scuba diver, you can avoid problems. But it is not a perfect world. If you fail to pay attention or forget to do something, a problem can occur. Good divers can deal with nearly any problem. This section introduces you to proven ways of dealing with potential diving difficulties. Do not be overly concerned about the problems presented. You can prevent them, but knowing how to deal with them helps reduce your apprehension.

Difficulties affecting your physiology include seasickness, dizziness, stress and panic, overexertion, coughing, and cramping. When you have one of these problems, your body sends messages that something is wrong. You need to know the messages and the actions you can take to overcome physiological difficulties.

Seasickness

The best strategy is to try to prevent seasickness (see "Equilibrium Adjustments and Seasickness" in chapter 3) because taking medication after you are seasick is usually ineffective. If you do get seasick, you are likely to vomit. Vomiting underwater can be dangerous because of involuntary gasping that can cause you to choke. Only you can determine your degree of nausea. If you throw up or feel as if you are on the verge of doing so, do not dive. If you feel queasy, getting into the water may help you overcome the feeling. Some divers who feel slightly nauseated find that they feel better if they get into the water quickly and dive. After that first dive, they are fine for the remainder of the day.

You should surface if you feel nauseated while diving. If you must vomit underwater, do not vomit through your regulator. Hold the second stage against one corner of your mouth, and depress the purge fully while you vomit. You should get air instead of water if you gasp. When you have finished throwing up, place the second stage in your mouth, clear it, and resume breathing. The purge method should be a last resort. Vomiting through the regulator is not recommended because this method sometimes results in clogging.

If you are seasick and have to throw up while aboard a boat, do it over the rail on the side of the vessel opposite the wind. Do not use the restroom (or head, as it is called on a boat). The best remedy is to get to land, rest until you feel better, and take seasickness medication before returning to the vessel. If you are ill from motion sickness—and if the dive boat has a dinghy and there is land nearby—you should request to be taken to shore for a while.

Dizziness

The absence of visual clues in a weightless environment can cause temporary dizziness. Visual references can help you prevent disorientation (see chapter 3). Injury, temperature changes, and pressure changes affecting the inner ear can also cause a whirling feeling called vertigo, which may be more difficult to overcome than dizziness caused by disorientation.

To cope with either dizziness or vertigo, you should first seek a fixed visual reference. If possible, make physical contact with something solid for a point of reference. If there is nothing to see or to grasp, close your eyes and hug yourself. In most cases, dizziness passes in a minute or two. If you then move slowly and keep your head still, you should be able to surface. A good buddy can recognize your difficulty and assist you.

Stress and Panic

Stress is the perception of a substantial imbalance between environmental demand and a person's response capability, and it occurs under conditions where failure to meet demand is perceived as having serious consequences. Stressors are conditions or attitudes that cause stress. Dive stressors, which can be internal or external, include cold, illness, exhaustion, injury, fear, equipment problems, loss of air supply, buddy separation, depth, darkness, currents, and disorientation. Stress is not always bad. Moderate stress can cause a feeling of exhilaration, improve performance, and lead to a positive condition called eustress. Excessive stress causes anxiety, decreases performance, and leads to a negative condition called distress.

Knowledge and appraisal of a situation affect your reaction to stress. Training, experience, and your predisposition toward a situation affect your knowledge and appraisal. Thus, your perception of the circumstances is determined by what you know and are able to do, by what you have done, and by any inherent fears about the circumstances that you might have.

The problem with stress in diving is that it can lead to panic—a sudden, uncontrolled, irrational reaction to a perceived danger. Divers who panic often perish. You must manage stress in order to prevent panic. Dr. Tom Griffiths, who has researched stress, said the following: "The most critical factor in the progression of panic after stress increases is whether or not a problem arises." Divers who can recognize and manage stress can overcome problems and are far safer than those who cannot cope with the effects of stress.

Stress has both physiological and psychological components. Anxiety causes involuntary physiological changes. When stress leads to anxiety, your breathing rate and heart rate increase, your nervous system becomes more active, and your awareness decreases. These factors decrease your performance and increase your anxiety. Then heightened anxiety begins the cycle again. Unless you interrupt the cycle, anxiety escalates into panic. Psychological difficulty is every bit as critical as the physical problems. Frank Pia, a chief lifeguard, says, "Much of the distress that a person experiences when difficulty arises stems from what the person tells himself about the situation." You can be exhilarated by a situation or allow it to cause anxiety and distress. Pia continues, "The difference between panic and a heightened physiological state is the thought process."

Michael J. Asken, who has written books about stress, encourages the use of task-relevant instructional self-talk (TRIST). You determine your emotional state when in distress by what you tell yourself. Asken says that in a stressful situation, "Success is not achieved by focusing on the outcome, hoping for a good outcome, or even telling yourself that the outcome will be OK. The most effective self-talk involves imagining that your instructor is sitting on your shoulder guiding you through your response."

With the previous concepts in mind, you can recognize and manage stress. The first steps are to sense your breathing rate and identify your self-talk when a problem occurs. If your breathing rate is fast and you are having negative thoughts about the situation, it is time to break the stress cycle. Stop all physical activity, establish buoyancy, and breathe deeply. Divers in distress usually have an adequate supply of air! As you gain control of respiration, you can think more clearly. Take control of your thoughts before they take control of you. Imagine that you are telling someone else how to deal with this situation. Assess your options, determine the best course of action, and then take deliberate action. As you begin to overcome the difficulty, your confidence increases and your physiological condition begins to return to normal. Figure 6.29 shows how to break the stress cycle that leads to panic.

Now you can understand why some divers are thrilled by the same situation that causes another diver to panic. Remember that stress is simply a matter of perception and that you can change your perception with training and experience.

You can also learn to manage difficult situations by imagining them vividly in your mind. Your body cannot distinguish the difference between an actual event and one that you imagine. As you learn to solve problems in your mind, you can learn to recognize and manage stress.

Overexertion

If you work too hard while scuba diving, the equipment may not be able to supply enough air to meet your respiratory needs. You experience a sudden feeling of suffocation, and you may suspect that your equipment has malfunctioned. It is unlikely that scuba equipment will cease to function suddenly in a way that restricts airflow. More likely than not, overexertion is causing the feeling that you cannot get enough air. You can manage overexertion similarly to the way you manage stress. Stop all physical activity and breathe deeply; you should overcome your respiratory problem within a minute or two.

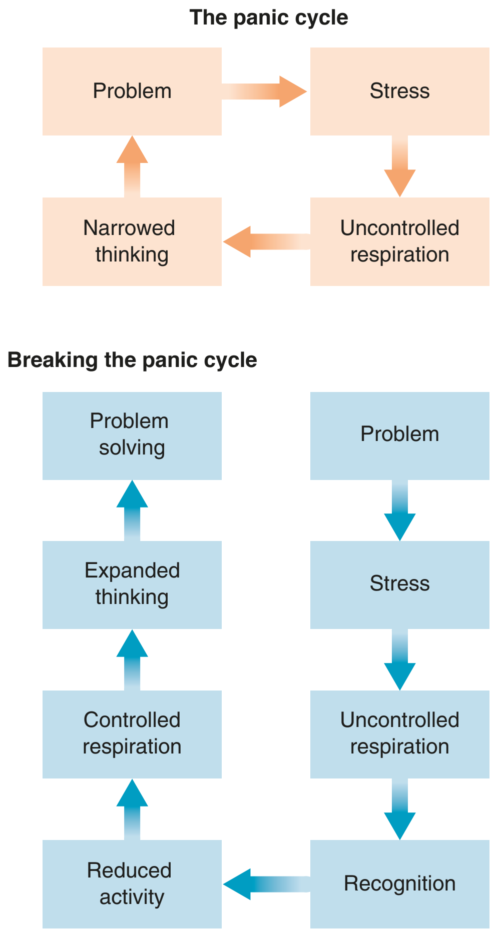

Figure 6.29 The panic cycle and how to break it.

Coughing

When water strikes your larynx, a reflex coughing action tries to clear the foreign matter from your airway. When you cough while in water, you may inhale additional water, which can complicate the situation. You need to overcome the reflex as quickly as possible. The best way is to swallow hard three times in rapid succession. If you must cough, try to do so through a regulator so that you inhale air rather than water if you gasp. You can lose buoyancy as you expel air when coughing, so you may need to establish positive buoyancy if you cough excessively.

Cramps

When your muscles get cold or when circulation to your muscles is inadequate, cramping can occur. A cramp is a sudden, strong, involuntary, persistent, painful muscle contraction. Divers tend to get cramps in the lower legs and the feet. To alleviate a cramp, stretch the affected muscle and rub it to increase circulation. Pounding a cramped muscle is ineffective and causes tissue damage. If you experience a cramp in your leg or foot in the water, you may be able to get rid of it by grasping the tip of your fin and pulling it toward you. Another technique to help release a leg cramp is to contract the muscles of the opposite leg. This action can inhibit nerve impulses to the cramped muscles. Buddies should assist one another with getting rid of cramps. Figure 6.30 shows a diver relieving a cramp.

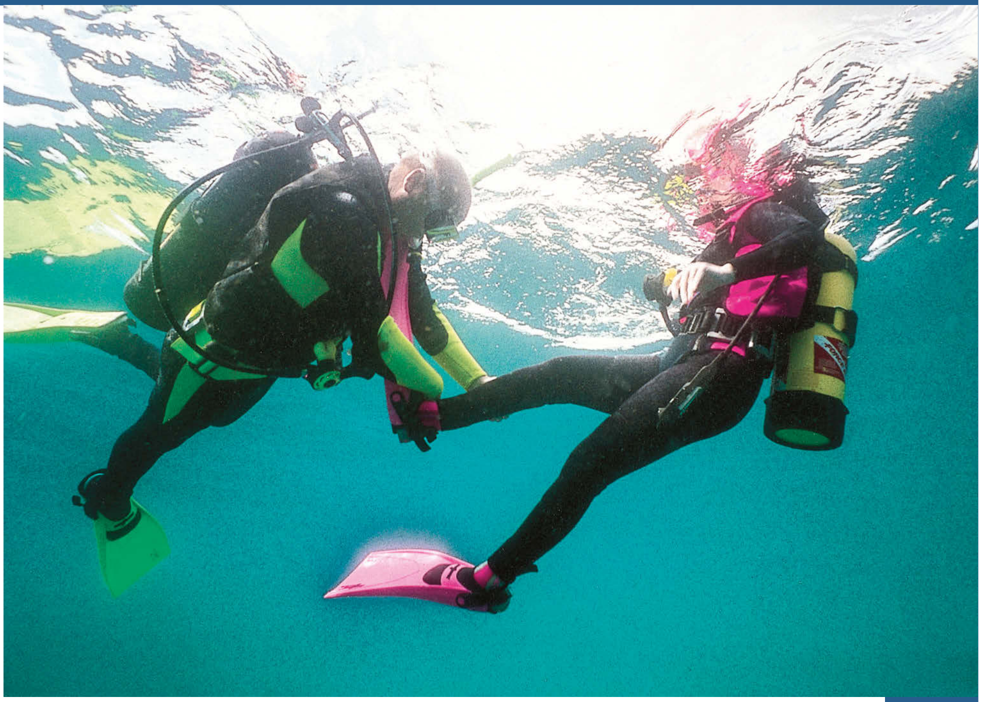

Figure 6.30 Your diving buddy can help you relieve a cramp by rubbing and stretching the cramped muscle.

Managing Physical Problems

Potential physical difficulties in diving include entanglement; loss of buoyancy control; loss of air supply; and a distressed, injured, or incapacitated buddy. As with physiological problems, you can overcome these problems, but it is better to avoid them.

Entanglement

You will probably encounter fishing line, nets, wire, string, and rope in the water—and these items can entangle you. Underwater plants can also entangle divers in some areas. Streamlining your equipment to minimize places where things can get caught helps reduce the chance of entanglement. Being aware also helps. When you encounter something that can entangle you, swim around it or push it beneath you and swim over it. Avoid swimming beneath things that can cause entanglement. The area of your tank valve and regulator first stage is one of the easiest places to become entangled and also one of the most difficult areas to free from entanglement.

If you become entangled underwater, you should stop. Then try to examine the problem without turning because turning often compounds the problem. If you can reach the entanglement, you should free yourself. If not, signal your buddy to help you get free. If you can see what you are doing, you can use your dive knife to cut yourself free. Trying to cut yourself free from something in the area of your tank valve is an obvious mistake. If your tank is entangled and you can't reach the entanglement—and your buddy is not nearby to lend assistance—you should remove your scuba unit, free it from the entanglement, and put it back on.

Loss of Buoyancy Control

You can lose control of buoyancy underwater if you lose weights or if a low-pressure inflator on your BC or dry suit sticks. You can take steps to prevent loss of buoyancy control. Check your weight system from time to time while you are diving to make sure it is secure. Inspect your low-pressure inflators before each use, and have them serviced at the first sign of unusual operation.

If your weights fall off while you are underwater, immediately maneuver yourself into an inverted position, swim down forcefully, and retrieve the weights. If you are successful, you can avoid an uncontrolled ascent; if you are unsuccessful, you will float to the surface. Your rate of ascent depends on how buoyant you are without your weights, the type of exposure suit you are wearing, the amount of air in your BC at the time, your depth, and the amount of surface area you can expose to the direction of motion. You learned in chapters 2 and 3 that the greater the cross-sectional area of an object moving in a given direction, the greater the resistance to movement in that direction. If you lose control of buoyancy and your ascent is uncontrolled, you can slow your rate of ascent by **flaring**—arching your back, extending your arms and legs, and positioning your fins parallel to the surface. Flaring (shown in figure 6.31) is the method recommended to slow a buoyant ascent.

Unless you care for your BC inflator carefully and have it serviced annually, the inflator valve will eventually stick. If your low-pressure inflator sticks in the open position, you should first hold your BC deflator valve in the open position. Modern BCs vent air faster than the low-pressure inflator admits air. If the inflator valve remains stuck, disconnect the low-pressure hose. If an uncontrolled ascent results, flare to slow your rate of ascent.

Loss of Air Supply

You are unlikely to have air supply difficulties if you have your regulator serviced annually, if you maintain your regulator properly, and if you monitor your SPG. Potential problems include a regulator that free-flows, low air pressure, and no air to breathe. Divers have ways to deal with each of these difficulties.

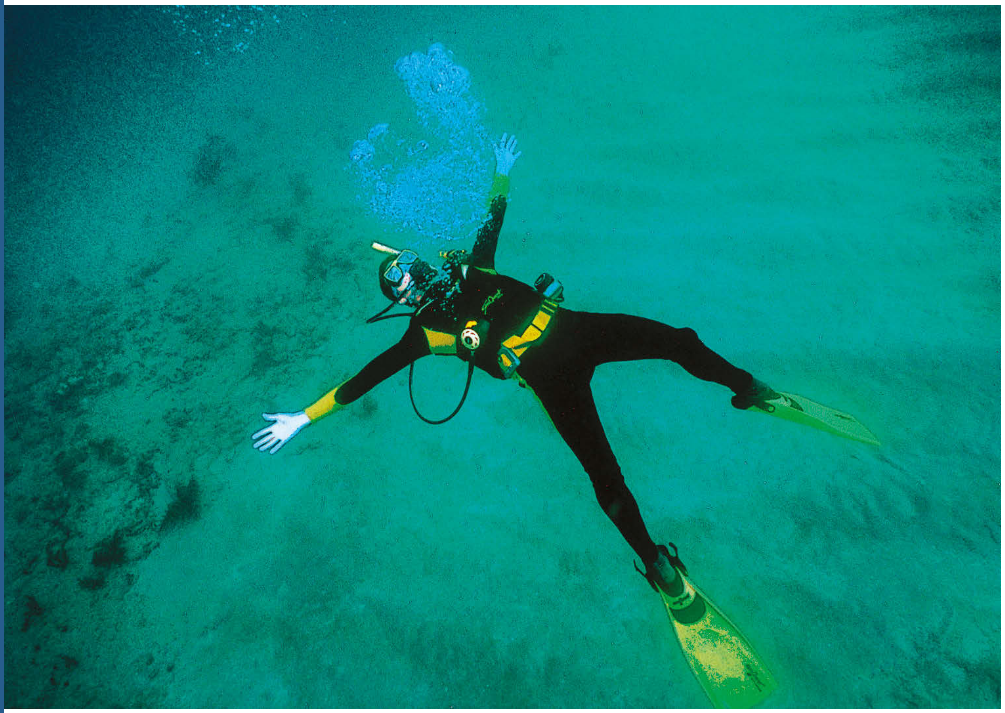

Figure 6.31 Flaring is an excellent way to slow a rapid ascent.

Sand, dirt, vomit, and freezing can cause regulator free-flow. If free-flow occurs, your best course of action is to switch to your extra second stage. You could also use your buddy's extra second stage. If there is no source of air except the regulator that is free-flowing, you can breathe from it by pressing your lips lightly against the mouthpiece, taking the air you need, and allowing the excess air to escape. Look down while you breathe from the regulator to keep the escaping air from causing your mask to leak.

You are supposed to end a dive with at least 500 psi (34 atm) of air in your tank. If you are inattentive and breathe nearly all the air from your cylinder while diving, it will become difficult to get air from your regulator. Divers often refer to this situation as being "out of air," but, in reality, they are out of air only at the depth at which breathing is difficult. As you ascend, the lower ambient pressure allows you to obtain additional air from your tank. When breathing becomes difficult and your tank is nearly empty, you should use your buddy's alternate air source (AAS) or ascend while continuing to breathe shallowly through your regulator.

In the rare event that you should completely lose your primary source of air while underwater, you have five ascent options. Figure 6.32 shows the order of preference of these options.

An extra-second-stage-assisted ascent closely approximates a normal ascent. You ascend while breathing from your buddy's extra second stage. When you require air underwater, get your buddy's attention and give the signals for "out of air" and "give me air," if possible. If your buddy's extra second stage and primary second stage are similar, your buddy will hand you the extra second stage, or you may take it. If your buddy's extra second stage is integrated into the BC low-pressure inflator, your buddy will hand you the primary second stage and breathe from the integrated second stage. If you cannot get your buddy's attention, take the extra second stage, begin breathing, and then signal your buddy that you have no air. After you establish a breathing rhythm, grasp each other's right forearm or BC and ascend normally (see figure 6.33).

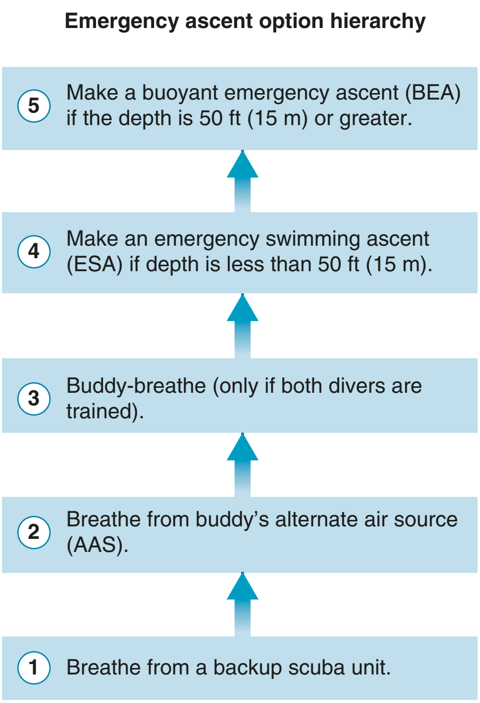

5 Make a buoyant emergency ascent (BEA) if the depth is 50 ft (15 m) or greater.

4 Make an emergency swimming ascent (ESA) if depth is less than 50 ft (15 m).

3 Buddy-breathe (only if both divers are trained).

2 Breathe from buddy's alternate air source (AAS).

1 Breathe from a backup scuba unit.

Figure 6.32 Loss-of-air options.

An emergency swimming ascent (ESA) is a scuba ascent you do using only the air in your lungs. The ascent rate of an ESA is faster than a normal ascent, but it is not rapid. Retain your regulator in your mouth, and try to breathe from it from time to time. Do not hold your breath or you risk a lung overexpansion injury. If you exhale too much air, you will have a strong urge to inhale. The key to a successful ESA is to exhale enough air that your lungs remain at a comfortable volume. When you do an ESA correctly, you can ascend 50 feet (15 m) easily without an overwhelming desire for air. If the depth is over 50 feet, discard your weights to initiate a buoyant emergency ascent (BEA). Swim

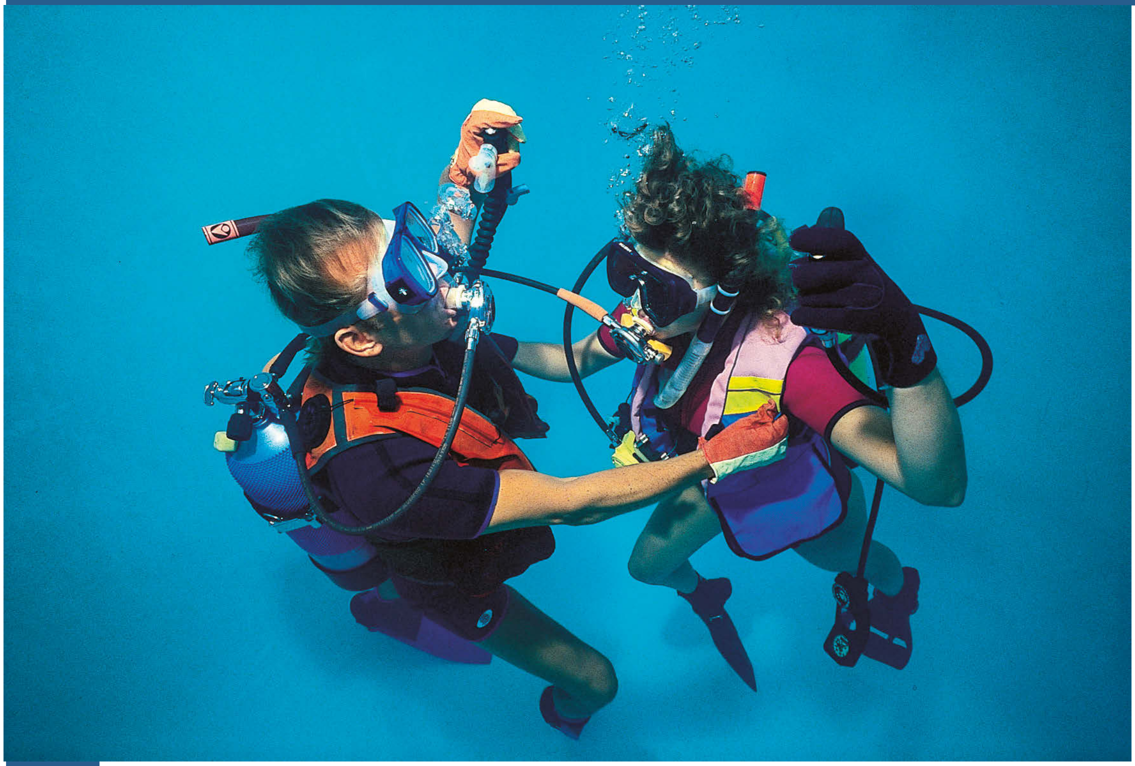

Figure 6.33 Air sharing involves cooperation and coordination with your buddy.

for the first portion of the ascent, but allow yourself to drift up when buoyancy can replace swimming. Flare during the last 15 feet (4.6 m) of a BEA, and keep your lungs at a comfortable—but not maximum—volume.

Buddy breathing—the sharing of a single regulator second stage by two divers—is not a desirable loss-of-air option because it jeopardizes the safety of two people. You and your buddy should practice buddy breathing at the surface before beginning a dive if buddy breathing is a loss-of-air option for the dive. Not all divers are trained to buddy-breathe, and buddy breathing attempted by two divers who are not proficient with the skill can result in disaster. However, those who are proficient buddy breathers can make a loss-of-air situation a mere nuisance.

Initiate buddy breathing with the "out-of-air" signal followed by the signal for "give me air." Your buddy holds the regulator second stage in the right hand and grasps your shoulder strap with the left hand. Your buddy extends the second stage toward you and holds it in such a way that you have access to the purge. You grasp your buddy's wrist (not the regulator) with your left hand and grasp your buddy's shoulder strap with your right hand. Guide the second stage to your mouth and push your lips against the mouthpiece to make a seal instead of inserting the mouthpiece into your mouth. By not putting the mouthpiece into your mouth to breathe, you can exchange the regulator quickly and reduce mask leakage caused by facial movements. Take several quick breaths initially; then pass the regulator back to your buddy. Exhale a small amount of air continuously when you are not breathing from the regulator. The exhalation helps prevent a lung overexpansion injury during ascent. After the initial contact, you and your buddy each take two breaths before passing the regulator. Do not inhale fully when buddy breathing because full breaths can cause buoyancy problems or lung injuries. A medium inhalation is adequate because you receive air every few seconds. You should exhale fully after each first breath that you take. The complete exhalation helps expel carbon dioxide and reduces your urge to breathe.

As soon as you and your buddy establish a breathing rhythm, you should swim to the surface, holding on to each other throughout the ascent. Blow bubbles continuously when the regulator is not in your mouth. Remember to control your buoyancy.

Skills for loss-of-air situations require proficiency, so they must be learned well and renewed periodically. Discuss the procedures for a loss-of-air situation with your buddy, and agree on the options you will use. You should both be familiar with the signals, positions, and techniques.

Assisting Your Buddy

You have read about many ways to assist your buddy and how your buddy can assist you, so you should realize the importance of the buddy system while diving.

In addition to helping your buddy handle entanglements, cramps, equipment problems, and loss-of-air situations, you may need to provide assistance if your buddy is incapacitated from exhaustion, illness, or injury.

A buddy who becomes agitated at the surface needs assistance in regaining control. Help such a buddy establish buoyancy, calm down, and breathe slowly and deeply. When the situation is under control, you may be able to help your buddy resolve the difficulty that caused the agitation.

If your buddy becomes exhausted, provide assistance at the surface with a biceps push, a fin push, or a do-si-do (arm-under-arm) push. The three types of pushes are illustrated in figure 6.34. Use the biceps push when your buddy can help, and use the fin push when your buddy is too exhausted to help at all. Monitor your buddy and offer encouragement while you provide assistance. (Note: Scuba rescues require techniques different from surface assistance.)

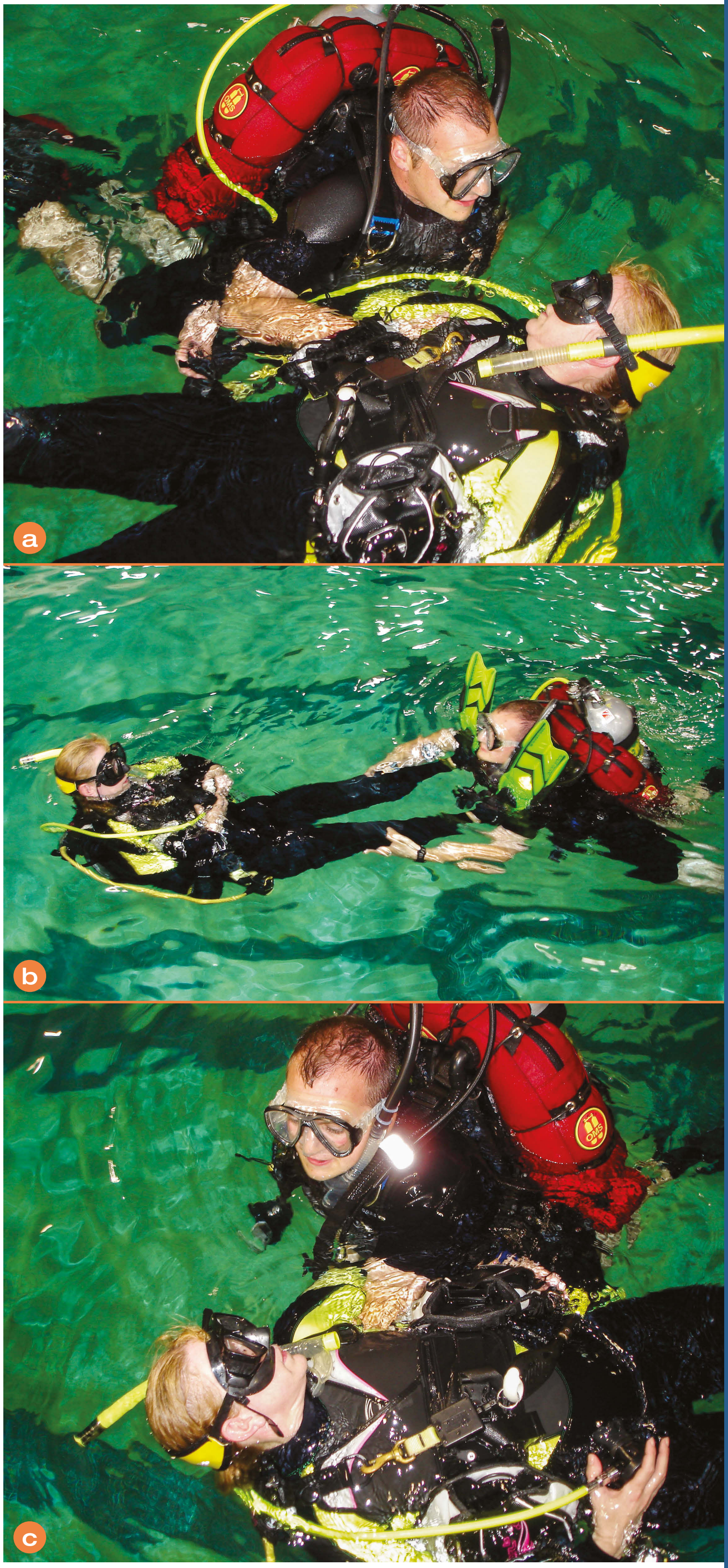

Figure 6.34 Providing assistance to an exhausted buddy using the (*a*) biceps push, (*b*) fin push, and (*c*) do-si-do push.

Managing Emergencies

Diving accidents occur when divers do not exercise good judgment or when they fail to follow recommended practices. If you do what you are supposed to do, the chances of a diving accident are extremely small. But you may have to render aid to someone else who violates safety rules. This section identifies the aid you should be capable of providing.

Training and Preparation

Three types of emergency preparedness training are recommended for all divers: first aid, cardiopulmonary resuscitation (CPR), and diving rescue techniques. You can get first aid and CPR training from various public service organizations. You'll learn some diving rescue techniques in your entry-level course, but you should also complete a rescue specialty course.

Emergency preparedness includes having emergency equipment and information available. The emergency equipment available at a dive site should include these items:

- Diving first aid kit (see the checklist of kit items in chapter 4)
- Oxygen delivery system
- Blanket (if appropriate)
- Drinking water

You may not have all of the emergency equipment yourself, but you can determine whether it is available aboard a boat or as part of an organized dive.

Have a means of communication—a telephone, cellular phone, CB radio, or marine radio—to summon assistance. Have contact information for local emergency medical assistance and for the emergency treatment of divers at the dive site. Have phone numbers and radio frequencies for local emergency support services. Examples of numbers to have include the Coast Guard, paramedics, hospital, ambulance, police or sheriff's office, recompression facility, and the Divers Alert Network (DAN). The Divers Alert Network has a 24-hour emergency number 919-684-8111 to assist with the coordination of responses to diving accidents.

Scuba Lifesaving

The principal elements of scuba lifesaving are cognizance, assessment, rescue, and evacuation (CARE). Cognizance is the ability to detect signs that identify or predict an accident. Trained and experienced divers can recognize situations that can lead to accidents before divers enter the water, when they enter the water, at the surface, during descents, underwater, during ascents, and during exits. As your experience increases, you will become more cognizant of potential difficulties.

Assessment is an ongoing evaluation of various factors that can affect a lifesaving situation. Rescuers need to consider their own ability to rescue the victim, the victim's condition, the environmental conditions, the equipment available, and additional resources.

Rescue techniques include approaches, extensions, throws, assists, establishing buoyancy, defenses, towing, deep-water rescue breathing, equipment removal, and removing the victim from the water. The ability to rescue both conscious and unconscious divers at the surface and underwater is an important skill to develop; you should take a scuba lifesaving or rescue specialty course as soon as possible after completing basic scuba training.

An unconscious diver underwater can drown unless rescued immediately. Illness, drugs, and blows to the head can cause loss of consciousness. If a diver loses consciousness underwater, make the diver buoyant and get her to the surface immediately. You do not need to concern yourself about expanding air if the diver is not breathing because an unconscious person exhales automatically regardless of head position. Do not concern yourself about the diver's decompression status; she can be treated for DCS but will die after about four minutes without air. Do not jeopardize your own safety when attempting to rescue another diver.

A diver who is not breathing needs air quickly and must get it at the surface. This is where your CPR skills and diving rescue techniques are invaluable. Open the victim's airway. Often this is all an unconscious person needs to be able to breathe. To open the airway, tilt the head and lift the chin. Turn the person's head to the side to drain water from the mouth and throat. Vomiting is common, so be prepared for it. Clear vomit from the victim's mouth and throat at once or the person may inhale it and choke.

A person who is not breathing after you have opened the airway requires rescue breathing and medical assistance. Shout for help. If you can remove the victim from the water quickly, you should do so. If removal will be delayed by even a couple of minutes, begin rescue breathing while you remain in the water. Keep the victim's airway open and lightly pinch the victim's nostrils to seal them. Seal your mouth over the other person's mouth and fill the person's lungs with air until the chest rises gently. Give the victim one breath every 5 seconds or two breaths every 10 seconds while swimming to safety. A gurgling sound from the victim indicates water or vomit in the airway. Roll the person to the side and drain the fluid before continuing rescue breathing. The preferred method of in-water artificial respiration involves using a rescue-breathing mask, which you can carry in the pocket of your BC. Figure 6.35 shows a person doing rescue breathing using such a mask.

You will probably not be able to detect a pulse in the water, so don't bother trying. You cannot administer CPR in the water; you must remove the victim from the water and position her on a firm surface.

Evacuation is the final element of scuba lifesaving. Evacuation procedures include accident preparedness, communications, and transportation options. Advance planning is essential to evacuate an injured diver to a medical facility.

First Aid

Some aspects of diving first aid are not taught in standard first

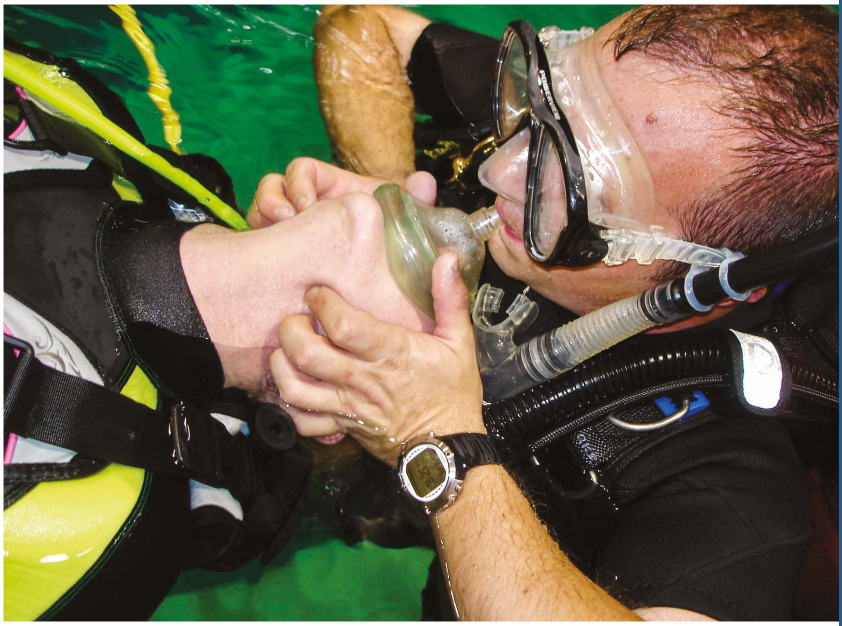

Figure 6.35 Rescue breathing with a breathing mask.

aid courses. This section touches on those aspects, but you need the additional study and training that you can get in rescue specialty and oxygen administration courses.

Nothing is more important than attending to basic life support: airway, breathing, and circulation. The next priority for any serious diving injury is treatment for shock. Lay an injured diver who is breathing on his side, keep the person warm (but avoid overheating), and administer sips of water if he is conscious. If you suspect an air embolism, DCS, or near drowning, have the person breathe oxygen in the highest possible concentration. Keep any diver who has lost consciousness or who has symptoms of DCS lying down until he can be evaluated at a medical facility. Monitor the victim continuously.

You should be able to recognize signs and symptoms that indicate a serious diving illness. The following signs and symptoms indicate an injury that requires the prompt administration of oxygen and medical treatment:

- Confusion
- Seizure
- Loss of consciousness
- Nausea or vomiting
- Shortness of breath
- Sudden, extreme weakness
- Numbness or a pins-and-needles sensation
- Inability to do simple motor skills
- Paralysis
- Unequal pupils

Venomous marine animals may be encountered in some parts of the world. A few venomous animals can inflict life-threatening wounds (see chapter 5). The wounds can cause pain, weakness, nausea, shock, mental confusion, paralysis, convulsions, depression, arrest of breathing, and even cardiac arrest. Fortunately, such wounds are rare.

Venomous injuries are either punctures or stings. First aid for venomous puncture wounds involves removing all foreign matter from the wound, applying hot packs to the injured area for half an hour, and keeping the injured area below the level of the heart. The injured person should obtain medical attention.

First aid for a venomous sting includes killing any stinging cells that are in contact with the skin, removing any residue, cleansing the area, applying an analgesic ointment for pain relief, and seeking medical attention. Vinegar is a good solution to apply to all stings to neutralize stinging cells initially.

You may not recall the appropriate first aid procedures in the event of an accident; therefore, you should bring along a diving first aid book to help you identify an injury and administer the appropriate first aid. A wilderness first aid book is also strongly recommended because diving often occurs in remote areas.

Managing Accidents

If a serious diving accident occurs, and no supervisory personnel are available to take charge, you have to manage the situation to the best of your ability. Summon help, but do not leave a seriously injured diver unattended. Enlist the aid of others. Try to locate the injured diver's identification and medical information. Write down what happened, as well as the person's dive profile,

SCUBA WISE

People are creatures of habit, and habits are the result of repetition. If you repeat an action correctly enough times, you form a good habit. On the other hand, if you repeat an action incorrectly enough times, you create a bad habit. Divers need good habits to avoid accidents, but sometimes they fail to take the time to develop them. I often watch divers on charter boats and at dive sites as they prepare to dive. I notice that some fail to inspect their equipment adequately, some fail to plan their dive properly, and some fail to follow the practices outlined in this chapter. They usually know what to do, but because they have bypassed many steps every time they go diving, they have a habit of skipping important items. How safe is a pilot if he ignores the preflight checklist? Taking the correct action requires concentration initially, but when repeated until the action becomes a habit, the process becomes automatic. People can make complex processes simple by repeating them. When a skill is executed properly every time, it not only becomes habit but also is easier to recall when you really need it, such as in an emergency. I strongly encourage you to take the time to form good diving habits. If you do, I assure you that the odds of being injured are extremely small.

symptoms, dive times, and so forth. Pin the information in a conspicuous place, and send it with the injured diver to the medical facility. Accompany the diver to the medical facility, if possible.

Summary

The skills of diving range from simple skin diving procedures to complex scuba skills to problem management. You need to learn the skills correctly the first time, practice them until you can do them easily, and renew them frequently to stay proficient. You also need to be trained and prepared to handle a diving emergency. As mentioned previously, visualization of skills can help you develop skills. This is especially important for problem-management skills. Some problems are not common, and you may never encounter them; but you need to be prepared to manage them. If you visualize a problem vividly in your mind, you will be able to remain relatively calm when that problem occurs because you will know what actions to take. Knowing what to do helps you remain relatively calm, which allows you to think and to better manage your situation.

Application-of-Knowledge (AOK) Questions

1. What are ways in which you can control buoyancy?
2. What do you need to remember to do while ascending from a scuba dive?
3. What can you do to minimize the chances of becoming separated from your buddy while scuba diving?
4. The inflator valve on your BC sticks when you open it to add air. What actions can you take to prevent a rapid ascent?
5. What should you do if your regulator begins free flowing while you are underwater and cannot be stopped?
6. When preparing to exit the water into a small boat, what is the correct order for the removal of your equipment?
7. While swimming against a current at the bottom in 40 feet (12.2 m) of water, you begin to experience air starvation. What is the proper way to manage this problem?
8. If you become entangled underwater, what can you do to get free of the entanglement?

Sunset at Grand Cayman

7

Dive Planning

Dive In and Discover

By the end of this chapter, you will be able to do the following:

- List at least nine factors to consider in dive planning.
- Explain the why, who, where, when, how, and what of advance planning.
- Explain the on-site planning procedures for a dive.
- List at least five methods of obtaining area orientations for scuba diving.
- Use dive tables to plan repetitive dives that do not require decompression.
- Explain the planning procedures for cold or strenuous dives, variations in ascent rate, multilevel dives, omitted decompression, diving after required decompression, going to altitude after diving, exceeding maximum time when doing precautionary decompression, and a repetitive-dive residual nitrogen time that exceeds the absolute bottom time of the previous dive.
- Define the terms *residual nitrogen*, *repetitive dive*, *surface interval time*, *maximum bottom time*, *decompression stop*, *repetitive group*, *safety stop*, *residual nitrogen time*, *actual bottom time*, *total bottom time*, *emergency decompression*, *no-decompression-stop limit*, *dive profile*, *step dive*, *spike dive*, *multilevel dive profile*, *sawtooth dive profile*, *scrolling*, and *contingency plan*.
- Explain contingency planning for scuba diving.
- Compare the advantages and disadvantages of dive computers.

In this chapter, you will learn about all phases of dive planning: advance planning, short-term planning, on-site planning, and postdive planning. You will also learn about area orientations and how to do dive profile planning. An essential part of dive planning is scheduling your time and depth to avoid DCS (decompression sickness).

A well-planned dive increases enjoyment and satisfaction and decreases the risk of injury. A poorly planned dive can result in disappointment, embarrassment, and discomfort. By the end of this chapter, you will understand the significance of the expression “Plan your dive, and then dive your plan.”

Dive-Planning Factors

Many factors affect your plans for a dive or a diving trip. Keep the following considerations in mind when you are looking ahead to a dive outing:

- Health and fitness are important. Illnesses, required medications, and recent operations probably disqualify you for diving. If your health is not normal, consult a diving physician. If there is any doubt about your physical condition, refrain from diving until you are in good health. If you are prone to motion sickness, take steps to try to prevent it.
- Climate is a big factor affecting dive planning. If you dive close to where you live, dive planning is easier than if you intend to dive thousands of miles away. A difference in climate usually means a big difference in diving conditions, which means a difference in your equipment requirements.
- The distance you travel to a diving destination affects your planning. If you travel far to reach the destination, allow a day to rest and recover from travel before you dive. After even one day of repetitive diving, wait one full day before flying home.
- Weather affects diving conditions significantly. Storms and sudden changes in the weather can make diving dangerous. Know the weather forecast, and reschedule your dive if poor weather is predicted. Know the expected wind speed, air temperature, and water conditions.
- Seasonal changes affect water movement, water visibility, air and water temperatures, entry and exit areas, and the presence of certain types of animals. You

First you learn the theory of diving, then the skills of diving, and then you apply what you have learned. Your training objective is to qualify to dive without supervision, which involves planning your underwater excursions.

Fish in a Red Sea coral cave

should know what to expect at a dive site at different times of the year. It helps to know the visibility, water temperatures, tides, surf, surge, currents, bottom composition, silt conditions, plants, and animals.

You need to be physically and mentally fit for diving. Fitness for diving implies that you are well rested, are well nourished, have the physical strength and stamina to meet the requirements of the environment and the activity, are qualified for the activity, are not apprehensive about your plans, are not goaded into doing something you are not prepared to do, and do not allow pride to affect good judgment.

Your objective for the dive affects your planning. Different diving activities require different plans and different equipment. The planning of an underwater photography dive is not the same as the planning of a dive where you intend to hunt for game.

You must know and observe laws, regulations, and customs. Some areas have laws that require the use of a dive flag. Obey fish and game regulations. Some diving professionals discourage the taking of any living thing in an area. You need to know the behavior expected of you. Knowing the expectations in advance can help you avoid being embarrassed at the dive site.

Etiquette is important. Will early-morning diving activities be offensive to residents near the dive site? Will the parking of vehicles at a site irritate people? Be considerate of others who may be in the area where you intend to dive, including those who are fishing nearby. Consider the impact of noise, changing clothes, and dive site access. Then make your plans using good etiquette.

Advance Planning and Preparation

The first phase of dive planning is the determination of why, who, where, when, how, and what.

Why?	Determine the objective of the dive. What do you want to do? Take photos? Explore? Look for artifacts?
Who?	Determine with whom you want to dive. Select a buddy who is interested in your dive objective.
Where?	Determine a primary and an alternative site. If conditions at the primary site are unfavorable, go to the alternate site.
When?	Determine the best time to dive. The water at most areas is usually calmer in the morning than it is in the afternoon. Tidal currents and height may affect the best time to dive.
How?	Decide how to reach the dive site. Who will drive? What are the directions?
What?	Determine what equipment is needed for the dive. Who will bring the float and flag? How many tanks do you need? Are there any special needs for the intended activity?

Advance preparation can include

- making reservations,
- paying deposits,
- buying or renting equipment,
- having equipment serviced or repaired,
- getting tanks filled,
- obtaining a fishing license or permit,
- preparing equipment for photography, and
- obtaining emergency contact information.

Fresh water diving at Lake Mead, Nevada

Your preparations will usually include a trip to your local dive facility. Inspect your equipment before you go. You may discover a needed repair that requires some time to complete. Identify your equipment needs early.

Short-Term Planning and Preparation

The day and evening before you intend to go diving, you need to take three actions. First, you should find out the weather forecast and current water conditions so you can determine if conditions will be acceptable for your diving activities. Call your dive buddy to discuss and confirm your plans. Last-minute revisions, such as going to the alternative site, may be necessary. If you anticipate poor diving or weather conditions, reschedule the dive.

The second step of short-term preparation is packing your diving equipment and your personal items. (See chapter 4 for an equipment checklist.)

The third step of short-term planning is to write down your dive plans and schedule. Leave this information with a friend. Instruct your friend to notify the authorities if you fail to return by a certain time.

On-Site Planning and Preparation

When you and your buddy arrive at the dive site, you must determine whether the conditions are acceptable for diving. If not, go to an alternative site. If the conditions at the alternative site are also unacceptable, abort the dive.

An important step in the assessment of a dive site is the estimation of the current. Look for telltale signs such as kelp bent over from water movement, a wake around the anchor line or behind an anchored boat, or objects drifting on the surface. Determine the velocity of moving water by measuring how long it takes a floating object to move a known distance, such as the length of your boat. When an object moves 100 feet (30 m) in one minute, the speed of the object is approximately 1 knot (1.15 miles or 1.85 km per hour). When a current exceeds about 1/3 knot (0.4 miles or 0.6 km per hour), you must heed it because you can swim at a speed of only about 3/4 knot (0.86 miles or 1.4 km per hour). Plan the dive so that the current assists you in swimming to your exit point at the end of the dive. Figure 7.1 includes a table that can help you estimate current velocity.

If the diving conditions are favorable, you should then determine the diving area. Select the entry and exit areas for the dive, and discuss the entry and exit procedures. Agree on the course to be followed during the dive. Agree on time, minimum air pressure, and landmarks for changes in direction. You and your buddy should know in advance approximately where you will be at any time during the dive.

An important part of your planning is discussing and agreeing on the procedures for the buddy system. Decide who is in charge of the team, where you will position yourselves relative to each other, how you will move (steadily or start and stop), and what reunion procedures you will follow in the event of separation. Remember that communication is much easier on land than it is underwater, so take advantage of the opportunity you have to communicate and coordinate while preparing for a dive.

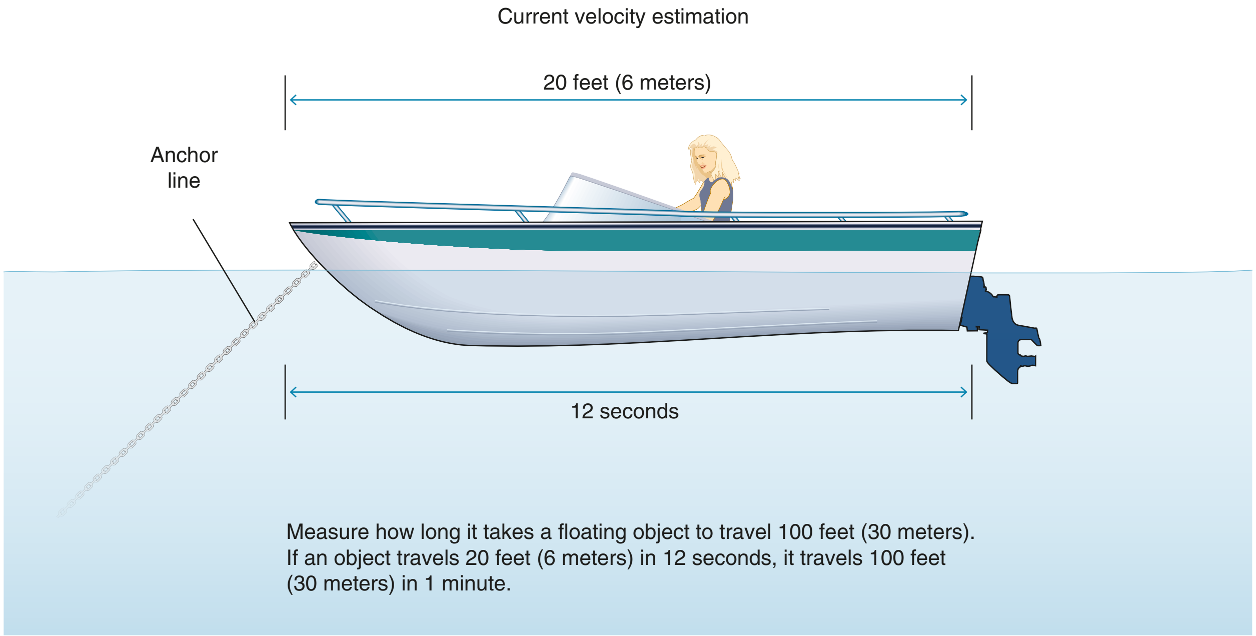

Measure how long it takes a floating object to travel 100 feet (30 meters). If an object travels 20 feet (6 meters) in 12 seconds, it travels 100 feet (30 meters) in 1 minute.

Current velocity table
(time to travel 100 feet or 30 meters)

Time (seconds)	Speed (knots)	Time (seconds)	Speed (knots)
5	12.0	95	0.62
10	6.0	100	0.59
20	3.0	110	0.54
30	2.0	120	0.49
40	1.5	130	0.46
50	1.2	140	0.42
60	1.0	150	0.39
70	0.84	160	0.37
80	0.74	170	0.35
90	0.66	180	0.33

Figure 7.1 You can estimate the velocity of a current if you time how long it takes a floating object to move the length of your boat.

Always review your plans for emergencies. Agree on air-sharing procedures. Discuss what to do in the event of a serious diving emergency. Know where, how, and whom to call for help. Make sure you both have access to a first aid kit and other emergency equipment. A few minutes spent coordinating procedures before an emergency can save precious seconds if an accident occurs.

Scuba diving requires dive profile planning. You and your buddy need to agree on the maximum time and depth for your dive. You must limit time and depth to avoid decompression illness. Dive profile planning is discussed later in this chapter.

Area Orientations

You should recognize the importance of learning about a dive site before diving there. Because orientation is vital and because you want to be a responsible diver, you need to learn how to obtain orientations. Area orientations can be formal or informal. A formal orientation is provided as a service by a diving professional. The professional will tell you what to look for and what to look out for in the area and will lead you on a dive. A professional dive guide provides suggestions and points out items of interest and potential hazards. When you have completed a formal orientation, ask the professional to sign and stamp your logbook. A continuing education dive course is another excellent form of formal orientation to a new area.

Formal orientations are ideal, but if you cannot arrange to have a professional introduce you to a new region, consider some or all of the following options for an informal orientation:

- Read books, articles, and brochures about diving in the area. Learn as much about an area as you can before you go there.
- Write to dive stores in the region where you intend to dive. Ask if you can participate in a dive class session for your orientation to the area.
- Write to dive clubs in a region where you intend to dive. Ask if you can participate in a club-sponsored dive when you are in the area. Ask for contact information for several club members who dive regularly and may be willing to allow you to go diving with them.
- When you arrive in a new area, find local dive sites and visit them when divers are likely to be there. Ask the divers about the sites while they are preparing to dive or after they exit from a dive. If you have your equipment ready, you may be able to accompany them on a dive. Make sure they have experience in diving at the site.

SCUBA WISE

Having reviewed dive accident reports for many years, I have found that failure to adequately plan or to carry out the plan for a dive is a common cause of accidents and injuries. I have been able to avoid serious injury during decades of diving. However, I have had some bad experiences. These experiences occurred when I failed to plan adequately, when I attempted activities without first completing training for the activity, or when I did not have adequate knowledge of a new dive site.

I dived from shore in California on a beach with a steep incline without taking the time to research the dive. I was used to donning my fins, wading into the water, and swimming beneath the oncoming waves. When I attempted this procedure on this type of beach, I found that there was a trough created by plunging waves, and that the trough was a dropoff. The trough is also where the waves break suddenly and violently. As I waded into the water, I lost my footing at the trough at the same time that a large breaker suddenly formed and pounded me into the trough. I managed to kick past the trough before the next wave hit, but I was humbled by the experience. A key part of dive planning is to understand the terrain and learn the local procedures. After the dive, I watched a local diver literally run into the water after a wave broke, jump over the trough, turn onto his back, pull on one fin, and kick beyond the area where the waves broke. I have used that technique on steep beaches ever since that day when the plunging surf taught me a lesson. Avoid surprises that could be dangerous by learning about any new location before diving there.

- Purchase a space on a diving charter boat. When you board the vessel, tell the crew you are new to the area. Ask for advice about diving procedures, and ask to be introduced to an experienced diver who can provide additional information.

When you dive with local divers, allow them to go first. Do as they say and do as they do. Procedures vary from region to region. A procedure you use in your area may be inappropriate in a different area. For example, in your normal diving environment, you may be able to enter the water from a boat in a current without holding on to a line. But if you tried to do so in a different area, you might be swept away at once. Be humble, listen to others, and follow their example to avoid embarrassment.

Dive Profile Planning

There are limits on how long you may remain at various depths. The amount of nitrogen absorbed by your body determines the limits. You need to be aware of the effects of **residual nitrogen**, which is nitrogen remaining in your system from a dive made within the past 12 hours. Whether you do one dive or a series of dives, it takes time to eliminate nitrogen from your body. If you dive again before the excess nitrogen has had time to outgas, you add to the nitrogen already in your body, and you reach critical nitrogen levels faster than if you had not already absorbed excess nitrogen. A **repetitive dive** is any dive made within 6 to 24 hours (depending on the dive-planning device) of a previous dive. Figure 7.2 shows how the amount of residual nitrogen in the body builds from repetitive dives. A **rest stop**—a precautionary decompression stop during ascent—reduces the risk of decompression illness.

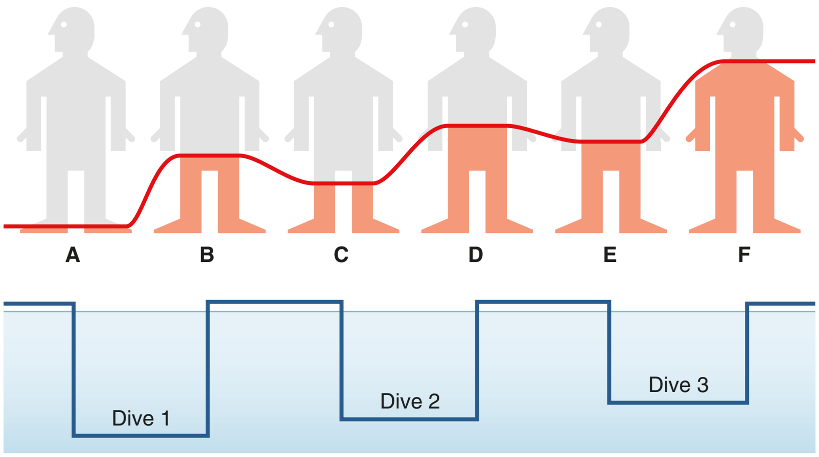

A—Normal amount of nitrogen in body
B—Amount of nitrogen in body after first dive
C—Amount of nitrogen in body after first surface interval
D—Amount of nitrogen in body after second dive
E—Amount of nitrogen in body after second surface interval
F—Amount of nitrogen after third dive

Figure 7.2 If you make a repetitive dive before allowing the residual nitrogen to leave your body, the nitrogen continues to accumulate.

Decompression experts use complex mathematical calculations and field testing to establish time limits for various depths for single and repetitive dives. The time limits are published in tables and programmed into calculators and dive computers. You need to know how to use these dive-planning devices to plan your dive profile so that you can minimize the risk of decompression illness.

No dive-planning device can guarantee that you will not develop decompression illness. Dive tables, calculators, and computers provide information based on statistics acquired through testing. If a diver adheres to the profile limits of a device, the statistical probability that the diver will develop decompression illness is small. The devices assume that you are in good health, that you do not get cold during the dive, that you do not exert strenuously, and that you ascend at the correct rate.

If you dive to the maximum time limits indicated by any dive-planning device, you increase the likelihood of decompression illness. Reducing your exposure to pressure reduces the likelihood of decompression illness.

Dive Tables

The U.S. Navy (USN) developed tables for dive planning. The recreational diving community adopted the military tables in 1980 and has used modified versions successfully. Although the tables were designed for military diving rather than recreational diving, many decompression experts still consider the USN dive tables—with reduced maximum dive times—appropriate tables for recreational use. New Doppler-tested dive tables developed by the USN are the basis for the dive tables presented in this chapter.

The USN tables use compartments with half-times of 5 minutes, 10 minutes, 20 minutes, 40 minutes, 80 minutes, and 120 minutes. The amount of nitrogen remaining in the 120-minute (2-hour) compartment determines the **repetitive group designations** of the tables, which are represented by letters. Because outgassing of a compartment takes six half-times, you can see why the USN tables define a repetitive dive as any dive within 12 hours (6 × 2 hours) of a previous dive.

The tables include four sets of dive tables:

A. Total bottom timetable

B. Surface interval timetable

C. Residual nitrogen timetable

D. Decompression timetable

Dive Table Modifications

Diving organizations and manufacturers have modified the dive tables to make them more appropriate for recreational diving. Here are some typical differences between standard dive tables and the modified dive tables presented in this chapter:

- Reduced time limits—Dive-planning devices today include maximum time limits that are less than the no-decompression-stop limits of the USN dive tables.
- Reduced depth limits—The USN dive tables provide dive-planning information for depths to 190 feet (58 m). The recommended maximum depth limit for recreational diving is 100 feet (30.5 m). New divers should limit diving to

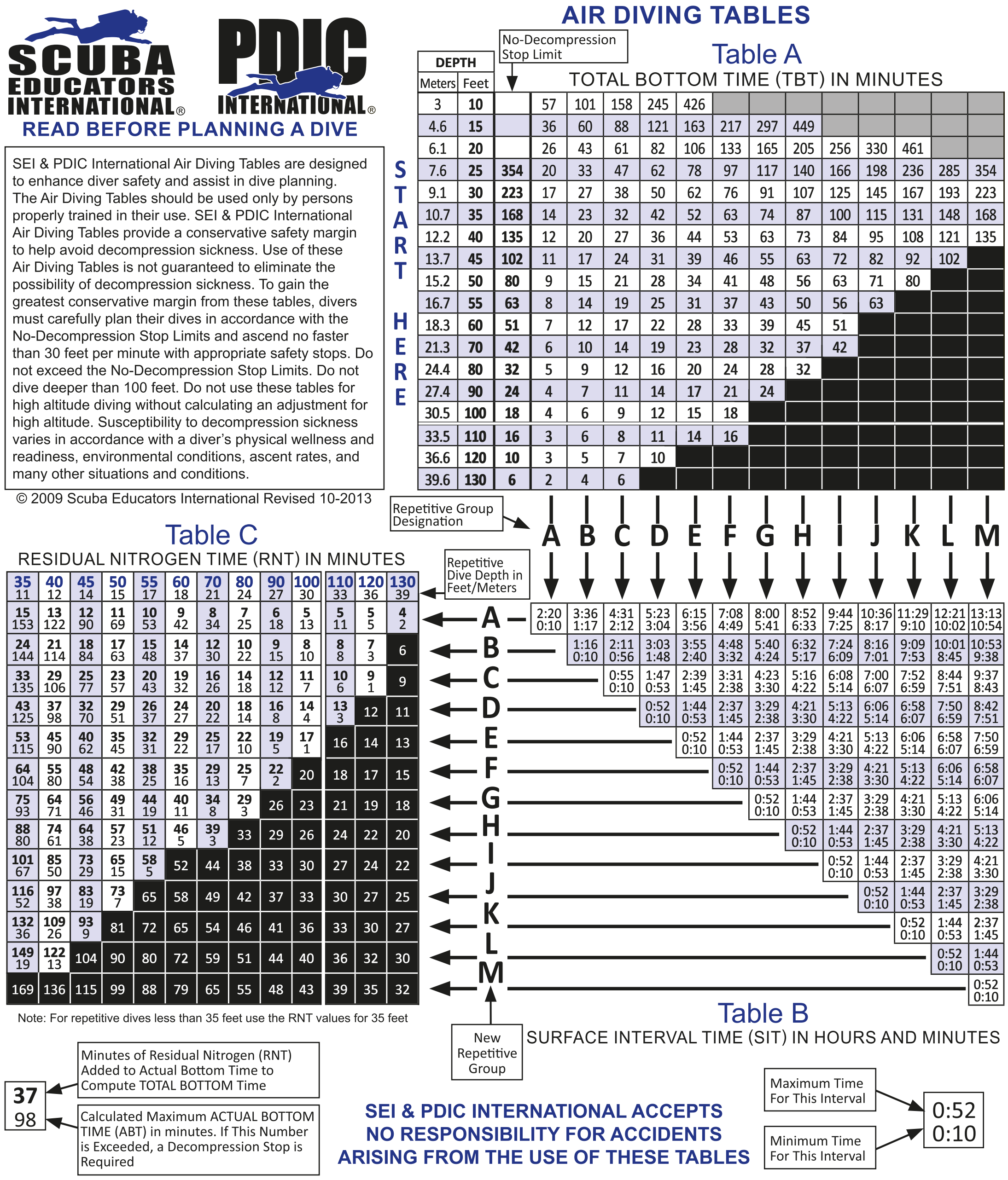

SCUBA EDUCATORS INTERNATIONAL®

PDIC INTERNATIONAL®

READ BEFORE PLANNING A DIVE

SEI & PDIC International Air Diving Tables are designed to enhance diver safety and assist in dive planning. The Air Diving Tables should be used only by persons properly trained in their use. SEI & PDIC International Air Diving Tables provide a conservative safety margin to help avoid decompression sickness. Use of these Air Diving Tables is not guaranteed to eliminate the possibility of decompression sickness. To gain the greatest conservative margin from these tables, divers must carefully plan their dives in accordance with the No-Decompression Stop Limits and ascend no faster than 30 feet per minute with appropriate safety stops. Do not exceed the No-Decompression Stop Limits. Do not dive deeper than 100 feet. Do not use these tables for high altitude diving without calculating an adjustment for high altitude. Susceptibility to decompression sickness varies in accordance with a diver's physical wellness and readiness, environmental conditions, ascent rates, and many other situations and conditions.

© 2009 Scuba Educators International Revised 10-2013

AIR DIVING TABLES

Table A

TOTAL BOTTOM TIME (TBT) IN MINUTES

START HERE

DEPTH Meters	DEPTH Feet	No-Decompression Stop Limit	A	B	C	D	E	F	G	H	I	J	K	L	M
3	**10**		57	101	158	245	426								
4.6	**15**		36	60	88	121	163	217	297	449					
6.1	**20**		26	43	61	82	106	133	165	205	256	330	461		
7.6	**25**	**354**	20	33	47	62	78	97	117	140	166	198	236	285	354
9.1	**30**	**223**	17	27	38	50	62	76	91	107	125	145	167	193	223
10.7	**35**	**168**	14	23	32	42	52	63	74	87	100	115	131	148	168
12.2	**40**	**135**	12	20	27	36	44	53	63	73	84	95	108	121	135
13.7	**45**	**102**	11	17	24	31	39	46	55	63	72	82	92	102	
15.2	**50**	**80**	9	15	21	28	34	41	48	56	63	71	80		
16.7	**55**	**63**	8	14	19	25	31	37	43	50	56	63			
18.3	**60**	**51**	7	12	17	22	28	33	39	45	51				
21.3	**70**	**42**	6	10	14	19	23	28	32	37	42				
24.4	**80**	**32**	5	9	12	16	20	24	28	32					
27.4	**90**	**24**	4	7	11	14	17	21	24						
30.5	**100**	**18**	4	6	9	12	15	18							
33.5	**110**	**16**	3	6	8	11	14	16							
36.6	**120**	**10**	3	5	7	10									
39.6	**130**	**6**	2	4	6										

Repetitive Group Designation

A B C D E F G H I J K L M

Table C

RESIDUAL NITROGEN TIME (RNT) IN MINUTES

Repetitive Dive Depth in Feet/Meters

	35 11	**40** 12	**45** 14	**50** 15	**55** 17	**60** 18	**70** 21	**80** 24	**90** 27	**100** 30	**110** 33	**120** 36	**130** 39
A	**15** 153	**13** 122	**12** 90	**11** 69	**10** 53	**9** 42	**8** 34	**7** 25	**6** 18	**5** 13	**5** 11	**5** 5	**4** 2
B	**24** 144	**21** 114	**18** 84	**17** 63	**15** 48	**14** 37	**12** 30	**10** 22	**9** 15	**8** 10	**8** 8	**7** 3	6
C	**33** 135	**29** 106	**25** 77	**23** 57	**20** 43	**19** 32	**16** 26	**14** 18	**12** 12	**11** 7	**10** 6	**9** 1	9
D	**43** 125	**37** 98	**32** 70	**29** 51	**26** 37	**24** 27	**20** 22	**18** 14	**16** 8	**14** 4	**13** 3	12	11
E	**53** 115	**45** 90	**40** 62	**35** 45	**32** 31	**29** 22	**25** 17	**22** 10	**19** 5	**17** 1	16	14	13
F	**64** 104	**55** 80	**48** 54	**42** 38	**38** 25	**35** 16	**29** 13	**25** 7	**22** 2	20	18	17	15
G	**75** 93	**64** 71	**56** 46	**49** 31	**44** 19	**40** 11	**34** 8	**29** 3	26	23	21	19	18
H	**88** 80	**74** 61	**64** 38	**57** 23	**51** 12	**46** 5	**39** 3	33	29	26	24	22	20
I	**101** 67	**85** 50	**73** 29	**65** 15	**58** 5	52	44	38	33	30	27	24	22
J	**116** 52	**97** 38	**83** 19	**73** 7	65	58	49	42	37	33	30	27	25
K	**132** 36	**109** 26	**93** 9	81	72	65	54	46	41	36	33	30	27
L	**149** 19	**122** 13	104	90	80	72	59	51	44	40	36	32	30
M	169	136	115	99	88	79	65	55	48	43	39	35	32

Note: For repetitive dives less than 35 feet use the RNT values for 35 feet

Table B

SURFACE INTERVAL TIME (SIT) IN HOURS AND MINUTES

	A	B	C	D	E	F	G	H	I	J	K	L	M
A	2:20 0:10	3:36 1:17	4:31 2:12	5:23 3:04	6:15 3:56	7:08 4:49	8:00 5:41	8:52 6:33	9:44 7:25	10:36 8:17	11:29 9:10	12:21 10:02	13:13 10:54
B		1:16 0:10	2:11 0:56	3:03 1:48	3:55 2:40	4:48 3:32	5:40 4:24	6:32 5:17	7:24 6:09	8:16 7:01	9:09 7:53	10:01 8:45	10:53 9:38
C			0:55 0:10	1:47 0:53	2:39 1:45	3:31 2:38	4:23 3:30	5:16 4:22	6:08 5:14	7:00 6:07	7:52 6:59	8:44 7:51	9:37 8:43
D				0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22	6:06 5:14	6:58 6:07	7:50 6:59	8:42 7:51
E					0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22	6:06 5:14	6:58 6:07	7:50 6:59
F						0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22	6:06 5:14	6:58 6:07
G							0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22	6:06 5:14
H								0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22
I									0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30
J										0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38
K											0:52 0:10	1:44 0:53	2:37 1:45
L												0:52 0:10	1:44 0:53
M													0:52 0:10

New Repetitive Group

37
98

Minutes of Residual Nitrogen (RNT) Added to Actual Bottom Time to Compute TOTAL BOTTOM Time

Calculated Maximum ACTUAL BOTTOM TIME (ABT) in minutes. If This Number is Exceeded, a Decompression Stop is Required

SEI & PDIC INTERNATIONAL ACCEPTS NO RESPONSIBILITY FOR ACCIDENTS ARISING FROM THE USE OF THESE TABLES

Maximum Time For This Interval

Minimum Time For This Interval

0:52
0:10

Courtesy of Scuba Educators International.

REPETITIVE DIVE PROFILES

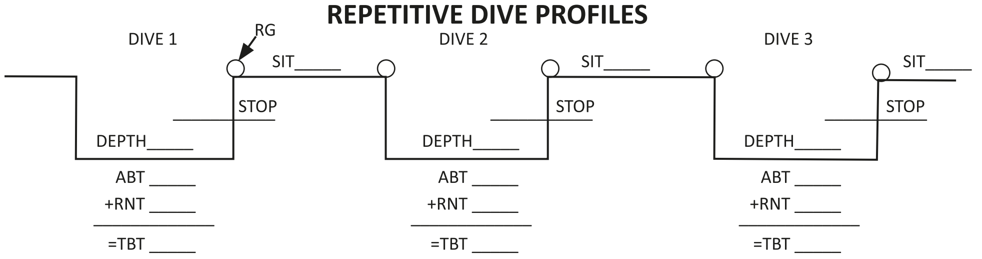

DEFINITIONS

BOTTOM TIME (BT). The elapsed time in minutes, starts when the diver leaves the surface in descent and stops when the diver begins safety stop or surfaces (round up to next whole minute). On repetitive dives this is Actual Bottom Time (ABT).

DEPTH. The depth of a dive; use the maximum depth attained during the dive in feet or meters of seawater. When depth is not indicated on the table use the next greater depth available.

NO-DECOMPRESSION STOP LIMIT. The longest amount of time a diver can spend at depth without requiring a decompression stop. Table A lists these for all depths. For repetitive dives the Residual Nitrogen Time (RNT) must be added to the Bottom Time (BT) and the Total Bottom Time (TBT) must not exceed the No-Decompression Stop Limit.

REPETITIVE DIVE. Any dive conducted within the time period specified by the surface interval table (Table B) of a previous dive with a minimum surface interval of 10 minutes. For repetitive dives of less than 35 feet, use the RNT (Table C) values for 35 feet to calculate Total Bottom Time (TBT).

REPETITIVE GROUP DESIGNATION (RG). Letters that relate to the amount of residual nitrogen in a diver's tissues for the time specified in the Suface Interval Table.

RESIDUAL NITROGEN TIME (RNT). Located in Table C, an amount of time, in minutes, which must be added to the Bottom Time (BT) of a repetitive dive to calculate Total Bottom Time (TBT). Represents residual nitrogen still in a diver's tissues from previous dives.

SAFETY STOP. A 3 minute minimum safety stop at 15-25 feet is recommended at the end of all dives.

SURFACE INTERVAL TIME (SIT). Located in Table B, the elapsed time from when a diver surfaces following a dive to the time (next whole minute) that diver starts the descent of the next dive. SIT must be a minimum of 10 minutes.

TOTAL BOTTOM TIME (TBT). Calculated by adding the Bottom Time (BT) and the Residual Nitrogen Time (RNT). For times not indicated at a particular depth, use the next longer time available for that depth.

RULES

ASCENT RATE. Divers should ascend at 30 feet per minute or slower.

DIVES LESS THAN 25 FEET. Dives less than 25 feet do not have a specific No-Decompression Stop Limit. However, if conducted as repetitive dives, they must be calculated as a 35 foot dive for Residual Nitrogen Times (RNT), Table C.

EXCEEDING THE NO-DECOMPRESSION STOP LIMITS. If divers err and stay longer than the times allowed on the No-Decompression Stop Limit, Table A, the divers have exposed themselves to an increased decompression requirement and must refer to the Decompression Stops Table, Table D, for required decompression stops to complete the dive. Once on the surface these divers should not dive again for at least 12 hours or fly for at least 24 hours.

FLYING AFTER DIVING. The longer the surface interval before flying the less likely a diver will experience decompression sickness. A minimum surface interval of 12 hours is required before ascent to altitude in a pressurized aircraft (or a ground altitude of 1000 feet above sea level). Divers who plan to make multiple dives for several days or computer assisted dives should take extra precaution and wait an extended surface interval of greater than 12 hours before flying or ascending to altitude. Divers who have made dives requiring decompression stops should wait a minimum of 24 hours before flying or ascending to altitude.

HIGH ALTITUDE DIVING. These Air Diving Tables are not to be used for diving at altitudes greater than 1000 feet above sea level without calculating altitude adjustments.

ORDER OF DIVES. All repetitive dives should be conducted with each successive dive equal in depth or shallower than the previous dive.

SPORT DIVING MAXIMUM DEPTH. Sport divers should not exceed 100 feet of depth.

SPORT DIVING TIME LIMIT. On any dive the Total Bottom Time (TBT) should not exceed the No-Decompression Stop Limit. Dives requiring mandatory decompression stops should not be planned.

Table D

DECOMPRESSION STOPS TABLE

Depth Feet Meters	Bottom Time (min.)	Time to First Stop (min:sec)	Decompression Stops (min) @ 20ft.	Total Assent (min:sec)	Repetitive Group
40 12.2	135		0		M
	150	0.40	6	7:20	O
	160	0.40	14	15:20	Z
	170	0.40	21	22:20	Z
	180	0.40	27	28:20	Z
50 15.2	80		0		K
	90	1:00	2	3:40	M
	95	1:00	4	5:40	N
	100	1:00	8	9:40	O
	110	1:00	21	22:40	O
60 18.3	51		0		I
	60	1:20	2	4:00	L
	65	1:20	7	9:00	L
	70	1:20	14	16:00	N
	80	1:20	23	25:00	O
70 21.3	42		0		I
	50	1:40	9	11:20	L
	55	1:40	14	16:20	M
	60	1:40	24	26:20	N
	70	1:40	44	46:20	O
80 24.4	32		0		H
	35	2:00	1	3:40	J
	40	2:00	10	12:40	K
	45	2:00	17	19:40	M
	50	2:00	24	26:40	M
90 27.4	24		0		G
	30	2:20	4	7:00	J
	35	2:20	14	17:00	L
	40	2:20	23	26:00	M
	45	2:20	31	34:00	N
100 30.5	18		0		F
	25	2:40	3	6:20	J
	30	2:40	15	18:20	L
	35	2:40	26	29:20	M
	40	2:40	36	39:20	N
110 33.5	16		0		F
	20	3:00	3	6:40	I
	25	3:00	14	17:40	K
	30	3:00	27	30:40	M
	35	3:00	39	42:40	N
120 36.6	10		0		D
	15	3:20	2	6:00	H
	20	3:20	8	12:00	J
	25	3:20	24	28:00	L
	30	3:20	38	42:00	N
130 39.6	6		0		C
	10	3:40	1	5:20	G
	15	3:40	4	8:20	I
	20	3:40	17	21:20	K
	25	3:40	34	38:20	M

Courtesy of Scuba Educators International.

Dive Table Terms

Actual Bottom Time (ABT)—The elapsed time in minutes (rounded up to the next whole minute) from when a diver leaves the surface in descent until the diver begins a rest stop or surfaces.

No Decompression Stop Limit (NDSL)—The maximum time that a diver may stay at a specified depth (indicated in table A). For repetitive dives, the residual nitrogen time (RNT) from table C must be added to the actual bottom time (ABT) to obtain the total bottom time (TBT), which must not exceed the NDSL specified in table A. Table C also indicates the maximum ABT for various depth and repetitive group combinations that will prevent the TBT from exceeding the NDSL.

Repetitive Dive—Any dive conducted within 12 hours of a previous dive with a minimum 10-minute surface interval. For repetitive dives of less than 40 feet, use the 40-foot (12 m) values in table C (RNT table) to determine the TBT.

Repetitive Group Designation (RGD or RG)—A letter that indicates the amount of residual nitrogen in a diver's tissues for 12 hours after diving.

Residual Nitrogen Time (RNT)—An amount of time (in minutes) that must be added to the actual bottom time (ABT) of a repetitive dive to determine the total bottom time (TBT). RNT compensates for residual nitrogen remaining in a diver's tissues from previous dives.

Surface Interval Time (SIT)—The elapsed time (rounded up to the next whole minute) from when a diver surfaces after a dive to the time the diver begins the descent of a repetitive dive. The minimum SIT is 10 minutes.

Total Bottom Time (TBT)—The sum of ABT and RNT. For times not indicated at a particular depth, use the next longest time available for that depth.

depths of about 60 feet (18 m). Advanced divers qualify to dive to approximately 100 feet. Divers who complete a deep-diving specialty course qualify to dive to 130 feet (39 m). Technical divers qualify to dive at even greater depths. Depth in water is similar to speed on land. When you qualify for a driver's license, you may drive at speeds up to the legal limit. Driving at much higher speeds is hazardous and requires specialized training, special equipment, and controlled situations to minimize the risk of serious injury or death. Deep diving (beyond 100 feet) is similar to race car driving in many respects. Deep diving is a professional endeavor that unqualified recreational divers should not attempt.

- Revised **surface interval times**—In 1983, a USN study reported a number of errors in the surface interval timetable. The errors usually do not affect the type of diving that recreational divers do, but some modified versions of the tables include the corrections.

- Combined tables—Modified versions of the USN dive tables often combine information to make the tables easier to use. Tables may include the **total bottom time** (TBT) for repetitive dives along with the **residual nitrogen time** (RNT). Tables may include required decompression information to eliminate a separate decompression timetable.

Dive Table Use

In this section we'll take a closer look at some of the individual elements comprising the combined dive tables.

Table A provides the NDSL, without required decompression stops for various depths. This table also provides a letter designation for various dive profiles. The letter designation indicates the amount of excess nitrogen remaining in your body after a dive.

Table A

TOTAL BOTTOM TIME (TBT) IN MINUTES

START HERE

DEPTH Meters	DEPTH Feet	No-Decompression Stop Limit	A	B	C	D	E	F	G	H	I	J	K	L	M
3	**10**		57	101	158	245	426								
4.6	**15**		36	60	88	121	163	217	297	449					
6.1	**20**		26	43	61	82	106	133	165	205	256	330	461		
7.6	**25**	**354**	20	33	47	62	78	97	117	140	166	198	236	285	354
9.1	**30**	**223**	17	27	38	50	62	76	91	107	125	145	167	193	223
10.7	**35**	**168**	14	23	32	42	52	63	74	87	100	115	131	148	168
12.2	**40**	**135**	12	20	27	36	44	53	63	73	84	95	108	121	135
13.7	**45**	**102**	11	17	24	31	39	46	55	63	72	82	92	102	
15.2	**50**	**80**	9	15	21	28	34	41	48	56	63	71	80		
16.7	**55**	**63**	8	14	19	25	31	37	43	50	56	63			
18.3	**60**	**51**	7	12	17	22	28	33	39	45	51				
21.3	**70**	**42**	6	10	14	19	23	28	32	37	42				
24.4	**80**	**32**	5	9	12	16	20	24	28	32					
27.4	**90**	**24**	4	7	11	14	17	21	24						
30.5	**100**	**18**	4	6	9	12	15	18							
33.5	**110**	**16**	3	6	8	11	14	16							
36.6	**120**	**10**	3	5	7	10									
39.6	**130**	**6**	2	4	6										

Repetitive Group Designation: A B C D E F G H I J K L M

Dive table courtesy of Scuba Educators International.

Table B, the surface interval timetable, provides letter group information based on your surface interval time (SIT). You begin a surface interval with one letter group designation; as you outgas nitrogen, you acquire lower letter group designations.

Table B

SURFACE INTERVAL TIME (SIT) IN HOURS AND MINUTES

Repetitive Dive Depth in Feet/Meters: 130 39	New Repetitive Group	Repetitive Group Designation: A	B	C	D	E	F	G	H	I	J	K	L	M
4 2	A	2:20 0:10	3:36 1:17	4:31 2:12	5:23 3:04	6:15 3:56	7:08 4:49	8:00 5:41	8:52 6:33	9:44 7:25	10:36 8:17	11:29 9:10	12:21 10:02	13:13 10:54
6	B		1:16 0:10	2:11 0:56	3:03 1:48	3:55 2:40	4:48 3:32	5:40 4:24	6:32 5:17	7:24 6:09	8:16 7:01	9:09 7:53	10:01 8:45	10:53 9:38
9	C			0:55 0:10	1:47 0:53	2:39 1:45	3:31 2:38	4:23 3:30	5:16 4:22	6:08 5:14	7:00 6:07	7:52 6:59	8:44 7:51	9:37 8:43
11	D				0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22	6:06 5:14	6:58 6:07	7:50 6:59	8:42 7:51
13	E					0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22	6:06 5:14	6:58 6:07	7:50 6:59
15	F						0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22	6:06 5:14	6:58 6:07
18	G							0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22	6:06 5:14
20	H								0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30	5:13 4:22
22	I									0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38	4:21 3:30
25	J										0:52 0:10	1:44 0:53	2:37 1:45	3:29 2:38
27	K											0:52 0:10	1:44 0:53	2:37 1:45
30	L												0:52 0:10	1:44 0:53
32	M													0:52 0:10

Dive table courtesy of Scuba Educators International.

Table C, the residual nitrogen timetable, provides adjusted (reduced) no-stop time limits based on the amount of nitrogen in your body from a previous dive or dives. The table also converts your residual nitrogen to an amount of time for each depth. This residual nitrogen time (RNT) must be added to your actual dive time to obtain the total bottom time (TBT) for the repetitive dive.

Table C

RESIDUAL NITROGEN TIME (RNT) IN MINUTES

35 11	**40** 12	**45** 14	**50** 15	**55** 17	**60** 18	**70** 21	**80** 24	**90** 27	**100** 30	**110** 33	**120** 36	**130** 39	Repetitive Dive Depth in Feet/Meters / New Repetitive Group
15 153	**13** 122	**12** 90	**11** 69	**10** 53	**9** 42	**8** 34	**7** 25	**6** 18	**5** 13	**5** 11	**5** 5	**4** 2	A
24 144	**21** 114	**18** 84	**17** 63	**15** 48	**14** 37	**12** 30	**10** 22	**9** 15	**8** 10	**8** 8	**7** 3	6	B
33 135	**29** 106	**25** 77	**23** 57	**20** 43	**19** 32	**16** 26	**14** 18	**12** 12	**11** 7	**10** 6	**9** 1	9	C
43 125	**37** 98	**32** 70	**29** 51	**26** 37	**24** 27	**20** 22	**18** 14	**16** 8	**14** 4	**13** 3	12	11	D
53 115	**45** 90	**40** 62	**35** 45	**32** 31	**29** 22	**25** 17	**22** 10	**19** 5	**17** 1	16	14	13	E
64 104	**55** 80	**48** 54	**42** 38	**38** 25	**35** 16	**29** 13	**25** 7	**22** 2	20	18	17	15	F
75 93	**64** 71	**56** 46	**49** 31	**44** 19	**40** 11	**34** 8	**29** 3	26	23	21	19	18	G
88 80	**74** 61	**64** 38	**57** 23	**51** 12	**46** 5	**39** 3	33	29	26	24	22	20	H
101 67	**85** 50	**73** 29	**65** 15	**58** 5	52	44	38	33	30	27	24	22	I
116 52	**97** 38	**83** 19	**73** 7	65	58	49	42	37	33	30	27	25	J
132 36	**109** 26	**93** 9	81	72	65	54	46	41	36	33	30	27	K
149 19	**122** 13	104	90	80	72	59	51	44	40	36	32	30	L
169	136	115	99	88	79	65	55	48	43	39	35	32	M

Note: For repetitive dives less than 35 feet use the RNT values for 35 feet

Dive table courtesy of Scuba Educators International.

Table D, the decompression stops table, provides the amount of time that you must decompress if your total bottom time (actual dive time plus residual nitrogen time) exceeds the maximum no-stop time (NDSL) limit.

Table D

DECOMPRESSION STOPS TABLE

Depth Feet Meters	Bottom Time (min.)	Time to First Stop (min:sec)	Decompression Stops (min) @ 20ft.	Total Assent (min:sec)	Repetitive Group
40 12.2	135		0		M
	150	0.40	6	7:20	O
	160	0.40	14	15:20	Z
	170	0.40	21	22:20	Z
	180	0.40	27	28:20	Z
50 15.2	80		0		K
	90	1:00	2	3:40	M
	95	1:00	4	5:40	N
	100	1:00	8	9:40	O
	110	1:00	21	22:40	O
60 18.3	51		0		I
	60	1:20	2	4:00	L
	65	1:20	7	9:00	L
	70	1:20	14	16:00	N
	80	1:20	23	25:00	O
70 21.3	42		0		I
	50	1:40	9	11:20	L
	55	1:40	14	16:20	M
	60	1:40	24	26:20	N
	70	1:40	44	46:20	O
80 24.4	32		0		H
	35	2:00	1	3:40	J
	40	2:00	10	12:40	K
	45	2:00	17	19:40	M
	50	2:00	24	26:40	M
90 27.4	24		0		G
	30	2:20	4	7:00	J
	35	2:20	14	17:00	L
	40	2:20	23	26:00	M
	45	2:20	31	34:00	N
100 30.5	18		0		F
	25	2:40	3	6:20	J
	30	2:40	15	18:20	L
	35	2:40	26	29:20	M
	40	2:40	36	39:20	N
110 33.5	16		0		F
	20	3:00	3	6:40	I
	25	3:00	14	17:40	K
	30	3:00	27	30:40	M
	35	3:00	39	42:40	N
120 36.6	10		0		D
	15	3:20	2	6:00	H
	20	3:20	8	12:00	J
	25	3:20	24	28:00	L
	30	3:20	38	42:00	N
130 39.6	6		0		C
	10	3:40	1	5:20	G
	15	3:40	4	8:20	I
	20	3:40	17	21:20	K
	25	3:40	34	38:20	M

Dive table courtesy of Scuba Educators International.

Now you are ready to learn how to plan dives using the air diving tables. Use table A to determine the maximum no-stop time (NDSL) for a dive. This is the maximum time you can spend underwater for the deepest depth you reach during a dive. If your total bottom time exceeds the NDSL time, you must complete mandatory decompression using table D (this will be discussed later). Your goal is to avoid exceeding the NDSL.

Rows are horizontal, and columns are vertical. Enter table A on the row corresponding to the deepest depth you plan to reach during a dive, and then move across to the end of the row to find your maximum NDSL. After your dive, reenter the row and find the first number (in minutes) that does not exceed your actual dive time. Proceed downward on the column containing your dive time and locate the repetitive group designation letter (from A to M) for your dive. For example, a dive of 50 feet (15 m) for 30 minutes assigns you to group E. Remember that whenever the depth or duration of your dive exceeds a listed number in the table, you must use the next larger number.

Next, you use table B, the surface interval table, to determine your letter group for a repetitive dive. Your group designation depends on the amount of time that you remain at the surface until you dive again. The longer your surface interval, the closer to the beginning of the alphabet your letter group will be.

Enter table B using the group designation you obtained from table A. Move downward along the column until you find the time range (expressed as hours: minutes; for example, 1:26 is 1 hour and 26 minutes) that your surface interval time (SIT) falls into. Remember that when you exceed a number on the dive tables, you must use the next larger number. Follow the row to the left to find your ending letter group designation. For example, if your letter group was E at the beginning of a 2:00 surface interval, you would be in group C at the end of the surface interval. You would be in group C after a surface interval ranging from 1:45 to 2:39.

Use table C, the residual nitrogen time (RNT) table, to determine both your adjusted maximum no-stop time for a repetitive dive and your residual nitrogen time (which must be added to your actual dive time) for your planned depth. For example, if your repetitive group letter from a previous dive is C and your planned depth for your next dive is 50 feet, your actual bottom time (the smaller, lower number for the depth column coordinate) must not exceed 57 minutes. Additionally, the residual nitrogen time (RNT) for the depth (the larger, upper number for the depth column coordinate) must be added to your actual bottom time. If you dived to 50 feet for 40 minutes as a group C diver, you would add 23 minutes of RNT to your actual bottom time to obtain a total bottom time (TBT) of 63 minutes.

After a repetitive dive, you return to table A and use your total bottom time (TBT) to obtain a new repetitive group designation. For example, your dive to 50 feet for 40 minutes as a group C diver yielded a TBT of 63 minutes. According to table A, 63 minutes of TBT at a maximum depth of 50 feet would assign you to repetitive group I at the beginning of your next surface interval.

Table D is for emergency use only when you exceed the maximum no-stop time (NDSL) for a depth. Avoid dives that require planned decompression. If you unintentionally allowed your total bottom time (TBT) to exceed the NDSL, you would refer to table D to determine the required decompression. For example, if your TBT for a depth of 60 feet (18 m) was more than 51 minutes (the NDSL), but less than 61 minutes, you would need to decompress at a depth of 20 feet (6 m) for 2 minutes. Your total ascent time, including the decompression, would require 4 minutes. Your repetitive group after the decompression dive would be L. This example is for emergency purposes only. You should never plan a decompression dive unless you complete specialty training and meet every requirement for safety.

Special-Equipment Dive Planning

You must use special dive tables if you dive using oxygen-enriched air (nitrox); you learn how to use these tables when you take a nitrox specialty course. Do not attempt specialized diving without proper training. Exceeding the 130-foot (39 m) depth limit by a few feet when breathing compressed air is not particularly dangerous, but exceeding the maximum depth limit (which varies with the gas mixture) when breathing mixed gases can cause seizures and drowning. Gas mixtures other than compressed air are only for divers who have the prerequisite training and equipment.

Dive Profile Terms and Rules

Divers use several types of **dive profiles**, which are created by plotting the time and depth of a dive. Figure 7.3 shows a diagram of a standard profile. A dive to a constant depth for a given period of time is typically depicted as a profile with square corners. A **multilevel dive profile** is a dive that progresses from deep to shallow during a given period of time. When you make a dive in a series of steps, you can refer to it as a **step dive**. A **sawtooth dive profile** is a dive that progresses from deep to shallow and back to deep. Avoid this type of dive. Another profile is the bounce profile. This represents a dive with a short ABT,

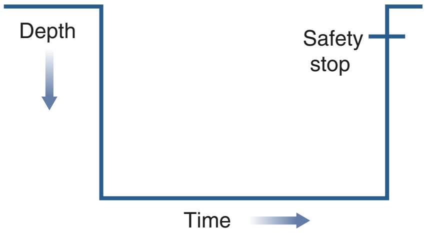
Depth
Safety stop
Time

Figure 7.3 Standard dive profile.

such as a dive to free a fouled anchor. You should also avoid this profile, which is depicted graphically as a spike. Bounce profiles are also called **spike dives**. Figure 7.4 graphically depicts these four profiles.

Dive Profile Diagramming

Diagram your dive profiles when planning and recording your dives. Include planned and actual depths, rest decompression stops, bottom times, repetitive group letter designations, and surface interval times. For repetitive dives, include RNTs and TBTs. A simple method for diagramming dive tables is to use a worksheet like the one shown in figure 7.5. Some dive computers automatically record a profile of your dive that may be displayed for review.

Try the following exercise, which combines diagramming with the dive table procedures you have learned. Use the blank diagramming worksheet in figure 7.5 as a template. If you have any difficulties, refer to the preceding section about the use of the dive tables. Assume that all dive times include three-minute rest

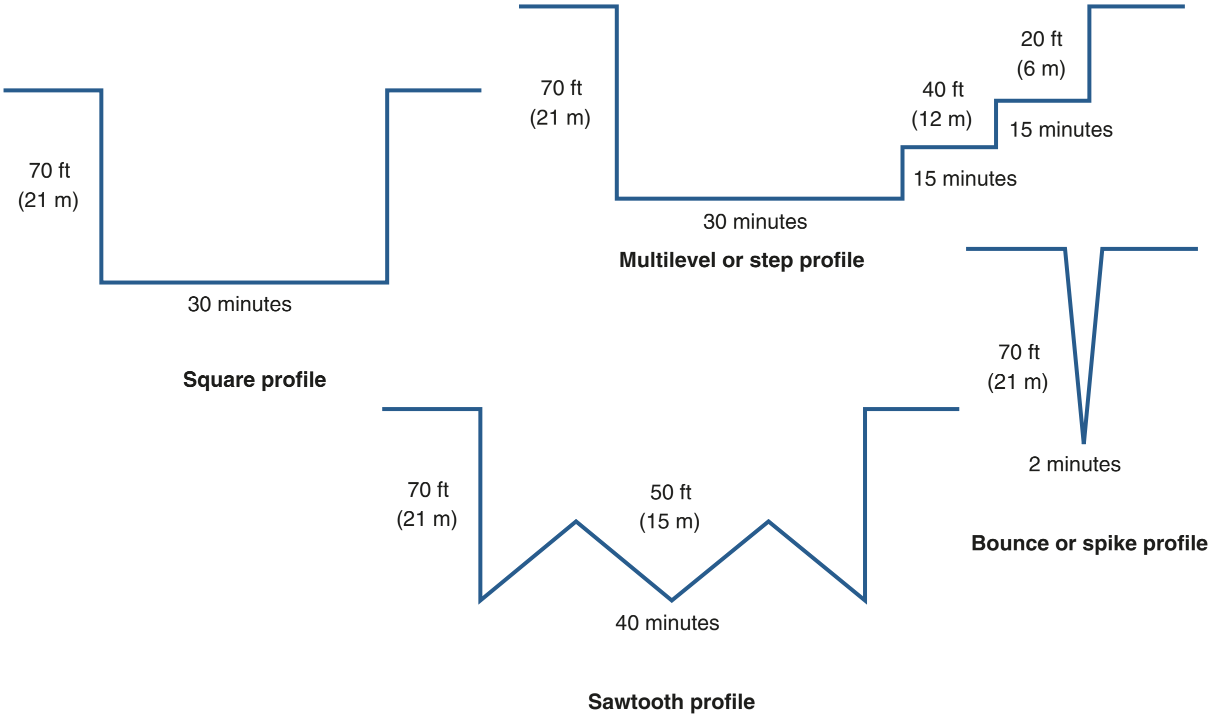
70 ft
(21 m)
30 minutes
Square profile
70 ft
(21 m)
30 minutes
40 ft
(12 m)
15 minutes
20 ft
(6 m)
15 minutes
Multilevel or step profile
70 ft
(21 m)
50 ft
(15 m)
40 minutes
Sawtooth profile
70 ft
(21 m)
2 minutes
Bounce or spike profile

Figure 7.4 Types of dive profiles.

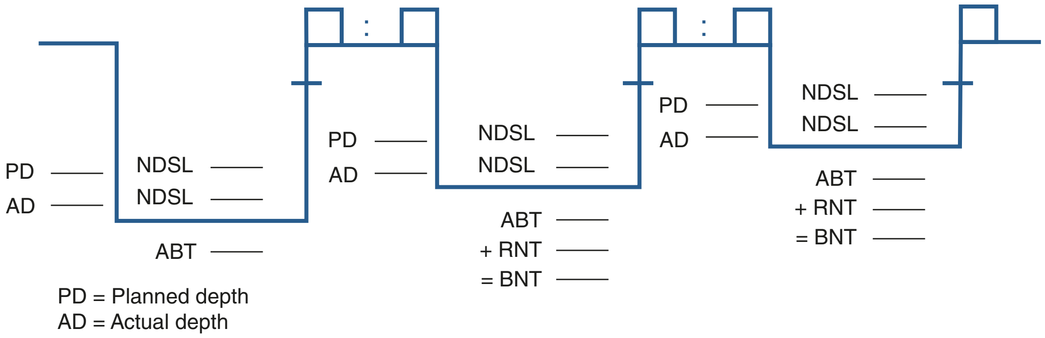
:
:
PD
AD
NDSL
NDSL
ABT
PD
AD
NDSL
NDSL
ABT
+ RNT
= BNT
PD
AD
NDSL
NDSL
ABT
+ RNT
= BNT
PD = Planned depth
AD = Actual depth

Figure 7.5 Use a worksheet like this to plan your dives.

stops. Calculate and diagram the following series of dives (the answers to the diagramming problem are in the next paragraph and in figure 7.6):

- The first dive is to 78 feet (24 m) with an ABT of 20 minutes, followed by a surface interval of 1.5 hours.
- The second dive is to 55 feet (16.8 m) with an ABT of 25 minutes, followed by a surface interval of 2 hours.
- The third dive is to 40 feet (12 m) with an ABT of 25 minutes.

Solution: The repetitive group after the first dive is E. After the surface interval, the repetitive group changes to D. The RNT for a D diver at 55 feet (16.8 m) is 26 minutes. The TBT (ABT + RNT) for the second dive is 51 minutes (25 + 26). The repetitive group after the second dive is I. After the second surface interval, the group changes to G. The RNT for a group G diver at 40 feet is 64 minutes. The TBT for the third dive is 89 minutes (25 + 64). The repetitive group after the third dive is J.

The following tips help make dive profile diagramming easier:

- Enter the repetitive group letter at every upper corner of a profile except the first corner.
- Add ABT and RNT to obtain TBT for every repetitive dive. Use the recall word ART to help remember that you add ABT and RNT to obtain TBT.

Now combine the dive table procedures to plan a series of dives. Use the blank diagram shown in figure 7.5. Assume that the depth of the first dive—the deepest dive—is 60 feet. The NDSL, according to table A, is 51 minutes. Plan an ABT

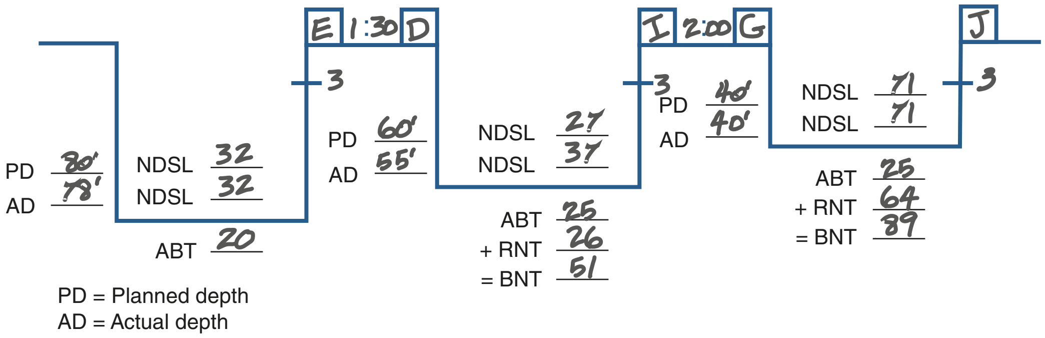
E
1:30
D
I
2:00
G
J
3
3
3
PD 80'
AD 78'
NDSL 32
NDSL 32
ABT 20
PD 60'
AD 55'
NDSL 27
NDSL 37
ABT 25
+ RNT 26
= BNT 51
PD 40'
AD 40'
NDSL 71
NDSL 71
ABT 25
+ RNT 64
= BNT 89
PD = Planned depth
AD = Actual depth

Figure 7.6 Completed diagramming worksheet.

Rules

Ascent Rate. Divers should ascend at 30 feet per minute or slower.

Dives Less Than 25 Feet. Dives less than 25 feet do not have a specific No Decompression Stop Limit. However, if conducted as repetitive dives, they must be calculated as a 35 foot dive for Residual Nitrogen Times (RNT), Table C.

Exceeding the No Decompression Stop Limits. If divers err and stay longer than the times allowed on the No Decompression Stop Limit, Table A, the divers have exposed themselves to an increased decompression requirement and must refer to the Decompression Stops Table, Table D, for required decompression stops to complete the dive. Once on the surface, these divers should not dive again for at least 12 hours or fly for at least 24 hours.

Flying After Diving. The longer the surface interval before flying, the less likely a diver will experience decompression sickness. A minimum surface interval of 12 hours is required before ascent to altitude in a pressurized aircraft (or a ground altitude of 1000 feet above sea level). Divers who plan to make multiple dives for several days or computer-assisted dives should take extra precaution and wait an extended surface interval of greater than 12 hours before flying or ascending to altitude. Divers who have made dives requiring decompression stops should wait a minimum of 24 hours before flying or ascending to altitude.

High Altitude Diving. These Air Diving Tables are not to be used for diving at altitudes greater than 1000 feet above sea level without calculating altitude adjustments.

Order of Dives. All repetitive dives should be conducted with each successive dive equal in depth or shallower than the previous dive.

Sport Diving Maximum Depth. Sport divers should not exceed 100 feet of depth.

Sport Diving Time Limit. On any dive, the Total Bottom Time (TBT) should not exceed the **No Decompression Stop Limit.** Dives requiring mandatory decompression stops should not be planned.

Definitions

Bottom Time (BT). The elapsed time in minutes starts when the diver leaves the surface in descent and stops when the diver begins safety stop or surfaces (round up to next whole minute). On repetitive dives, this is Actual Bottom Time (ABT).

Depth. The depth of a dive; use the maximum depth attained during the dive in feet or meters of seawater. When depth is not indicated on the table, use the next greater depth available.

No Decompression Stop Limit. The longest amount of time a diver can spend at depth without requiring a decompression stop. Table A lists these for all depths. For repetitive dives, the Residual Nitrogen Time (RNT) must be added to the Bottom Time (BT) and the total bottom time (TBT) must not exceed the No Decompression Stop Limit.

Repetitive Dive. Any dive conducted within the time period specified by the surface interval table (Table B) of a previous dive with a minimum surface interval of 10 minutes. For repetitive dives of less than 35 feet, use the RNT (Table C) values for 35 feet to calculate Total Bottom Time (TBT).

Repetitive Group Designation (RG). Letters that relate to the amount of residual nitrogen in a diver's tissues for the time specified in the Surface Interval Table.

Residual Nitrogen Time (RNT). Located in Table C, an amount of time, in minutes, which must be added to the Bottom Time (BT) of a repetitive dive to calculate Total Bottom Time (TBT). Represents residual nitrogen still in a diver's tissues from previous dives.

Safety Stop. A 3-minute minimum safety stop at 15-25 feet is recommended at the end of all dives.

Surface Interval Time (SIT). Located in Table B, the elapsed time from when a diver surfaces following a dive to the time (next whole minute) that diver starts the descent of the next dive. SIT must be a minimum of 10 minutes.

Total Bottom Time (TBT). Calculated by adding the Bottom Time (BT) and the Residual Nitrogen Time (RNT). For times not indicated at a particular depth, use the next longer time available for that depth.

of 28 minutes (including rest stop) for the dive. Your repetitive group designation after the first dive is E. Now plan a repetitive dive to the same depth. Refer to table C to plan the dive because table C provides the maximum allowable times for various letter group designations. If your surface interval is less than 52 minutes, you remain in the E group, your RNT is 29 minutes, and your ABT must not exceed 22 minutes. Your surface interval should be at least 1 hour, as recommended. If you wait an hour between the first and second dive, your repetitive group changes to D, your RNT is 24 minutes, and your ABT cannot exceed 27 minutes. If you want to dive for 28 minutes, you need to extend your surface interval to at least 1 hour and 45 minutes to obtain repetitive group C.

Assume that you repeat the first dive as a group C diver. Your RNT is 19 minutes and your ABT is 28 minutes, so your TBT is 47 minutes. Your repetitive group after the second dive is I.

Assume that the dive site is so good that you want to make a third dive to 60 feet. No ABT minutes are allowed for a diver with a group I designation. Proceed up the column of the table. If you want to dive to 60 feet again for 28 minutes, you need to attain repetitive group C, which allows 32 minutes of diving without required decompression. When you know the group that you need to attain as well as your starting group, you can plan your surface interval. Table B tells you that you must wait at least 5 hours and 14 minutes to move from group I to group C. It would be wise to consider a shallower dive at this point.

Use the tables to plan dives that enable you to avoid emergency decompression. You have three options for planning repetitive dives that do not require mandatory decompression. If your RNT prevents you from making a desired dive, you can (1) reduce the duration of the dive, (2) reduce the depth of the dive, or (3) increase the surface interval preceding the dive.

Special Procedures

Unusual circumstances may arise that require special procedures. You should know the procedures for dive profile planning for each of the following situations:

- A cold or strenuous dive—When a dive is particularly cold or strenuous, use the next greater time for the dive. If the dive is both cold and strenuous, use the next greater time and depth.
- Variations in the rate of ascent—If you ascend faster than 30 feet (9 m) per minute, extend your rest stop by at least two minutes. The faster you ascend, the more you should increase the stop time.
- Multilevel dives—Treat dives to multiple levels as square-profile dives, as if all the time of the dive was at the deepest depth of the dive. Do not attempt to extrapolate the dive tables.
- Omitted decompression—If you need to decompress but fail to do so, use the following procedure for omitted decompression. If you have no symptoms of decompression illness after the dive, remain out of the water, breathe oxygen in the highest concentration possible, rest, drink water, and be alert for symptoms of decompression illness. Wait 24 hours before diving again. If you suspect decompression illness, proceed at once to a hyperbaric facility for a medical examination. The USN has a procedure for in-water decompression, but diving medical experts agree that this procedure is inappropriate unless there is no alternative.

- A rest stop that causes your ABT or TBT to exceed NDSL—If a rest stop causes ABT or TBT to exceed the maximum time limits, determine your repetitive group letter designation using the NDSL for the profile.
- Diving after mandatory decompression—After a dive that requires decompression, wait at least 24 hours before diving again.
- A repetitive dive with an RNT that exceeds the ABT of the previous dive—If the RNT for a repetitive dive exceeds the ABT of the previous dive, use the RNT for planning the repetitive dive.
- Altitude after diving—Ascending to altitude after diving increases the likelihood of decompression illness because of further reduction in pressure. Driving in the mountains or flying after diving can cause decompression illness that would not occur if you remained at sea level until you eliminated the excess nitrogen.

The Divers Alert Network (DAN) recommends that divers wait a minimum of 12 hours before ascent to altitude in a commercial jet airliner (higher than 8,000 feet, or 2,438 m) after a single no-decompression dive. If you make multiple dives for several days, wait at least 18 hours before flight. If your dive requires decompression, a preflight interval substantially longer than 18 hours is required. Lengthening your preflight time interval reduces the chance that you will experience decompression illness.

DAN does not have a recommendation for flying or driving at lower altitudes. The most extensively tested tables for altitudes are the Swiss dive tables. The Swiss tables use a compartment with a much longer half-time than the USN tables. A reasonable approach to altitude delays after diving makes the surface intervals of the USN dive tables equivalent to those of the Swiss tables. Converted USN minimum surface intervals in table 7.1 specify the minimum time required to attain permissible nitrogen levels for various altitudes.

Altitude is any elevation above 1,000 feet (305 m). The altitude delay timetable provides recommended time delays for altitudes up to 10,000 feet (3,048 m). To use the altitude delay timetable, enter the table horizontally on the top line and

Table 7.1 Altitude Delay Timetable

Starting repetitive group										
Altitude (ft/m)	**ABC**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**Group***
2,000/610	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	2:26	K
3,000/914	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	2:37	4:08	J
4,000/1,219	0:00	0:00	0:00	0:00	0:00	0:00	0:00	2:53	4:30	5:51	I
5,000/1,524	0:00	0:00	0:00	0:00	0:00	0:00	3:04	4:57	6:29	7:44	H
6,000/1,829	0:00	0:00	0:00	0:00	0:00	3:20	5:24	7:12	8:38	9:54	G
7,000/2,134	0:00	0:00	0:00	0:00	3:41	6:02	8:06	9:43	11:10	12:36	F
8,000/2,438	0:00	0:00	0:00	4:08	6:50	9:11	11:04	12:41	14:19	15:40	E
9,000/2,743	0:00	0:00	4:57	8:06	10:48	12:58	14:51	16:39	18:11	23:09	D
10,000/3,048	0:00	6:18	10:37	13:25	15:56	18:05	20:10	21:18	23:24	24:50	C

Note: Times represent the minimum recommended time delay before ascending to the listed altitude and are USN surface interval times with a delay factor of 5.4. Times are in hours:minutes (for example, 5:24 is 5 hours 24 minutes).

*Recommended minimum repetitive groups for indicated elevations.

find your starting repetitive group. Next, find the altitude to which you wish to ascend. If you exceed a number, use the next greater one. The time at the coordinates of the desired altitude and your starting repetitive group indicates the minimum time delay recommended before ascending to the altitude selected.

Profile Contingency Planning

When you plan a dive profile, you should also plan for contingencies. You should know what to do in case you unintentionally exceed your planned depth or time or both. A simple matrix, such as that depicted in figure 7.7, is helpful. When diving using dive tables, you should prepare a contingency matrix in advance of your dives and carry it with you while diving. Dive computers automatically provide contingency information.

Not all dive-planning information displayed by a computer can be accepted at face value. You must learn to apply the dive computer guidelines included in this section.

After you complete your plans, you need to implement them. The first rule of dive planning is to plan your dive, and then dive your plan. You and your buddy should make every effort to do as you agreed to do before the dive. When circumstances force changes in your plans, it helps to have **contingency plans**.

Times in boxes are emergency decompression times.

Sample

	Planned time 30 minutes	Next greater time 40 minutes	Next greater time 50 minutes
Planned depth 60 ft	0	0	0
Next greater depth 70 ft	0	0	5
Next greater depth 80 ft	0	5	10

	Planned time ____	Next greater time ____	Next greater time ____
Planned depth ____			
Next greater depth ____			
Next greater depth ____			

Figure 7.7 A contingency matrix helps you plan ahead for unexpected events.

These plans need to address many possible contingencies for a dive. For example, you should know what to do if you

- surface downcurrent from a boat when you had planned to surface in front of it,
- are unable to reach or use the exit point you had selected on shore,
- end the dive a long distance from your exit location,
- exceed the maximum allowable bottom time (ABT or TBT) for a dive,
- experience a failure of your dive computer while diving, or
- ascend directly to the surface without precautionary decompression.

Dive Calculators and Dive Computers

The benefits of a dive computer outweigh the potential problems associated with the use of these devices. Because dive computers simplify dive planning, you should obtain a dive computer as soon as you can. However, make sure you learn how to use dive tables and dive calculators so that you will know how to plan a dive if you do not have a dive computer. Although dive computers have minimized the use of dive tables, divers must be able to use dive tables as a backup in case the dive computer fails to work.

Dive calculators are circular planning devices that eliminate the mathematics required with the dive tables. Dive calculators provide precalculated numbers, and they eliminate the need to add and subtract times. Guides on the calculators help eliminate line-jumping errors frequently made when using dive tables.

Dive tables have 5- or 10-foot (1.5 or 3 m) increments and require all the time of a dive to be counted at maximum depth of the dive. If the first part of your dive is deeper than the remainder of the dive, you suffer a penalty because the tables treat the entire dive as if it took place at the deepest depth (see figure 7.8). At the end of the dive, you receive a repetitive group designation that is higher than you deserve. Dive tables are advance-planning devices, while dive computers provide constant information on decompression status.

Dive computers use 1-foot (0.3 m) increments for profile planning, and they calculate nitrogen absorption continuously. As you vary depth during a multilevel dive, you are charged only for the nitrogen you absorb. You do not incur the maximum-depth penalty (see figure 7.8) of the dive tables, so your residual nitrogen time after a multilevel dive is less than when calculated using dive tables. The penalty avoidance is the primary advantage of a dive computer as a planning device. Dive computers also provide advance-planning information by **scrolling** a sequential display of time limits for various depths. You

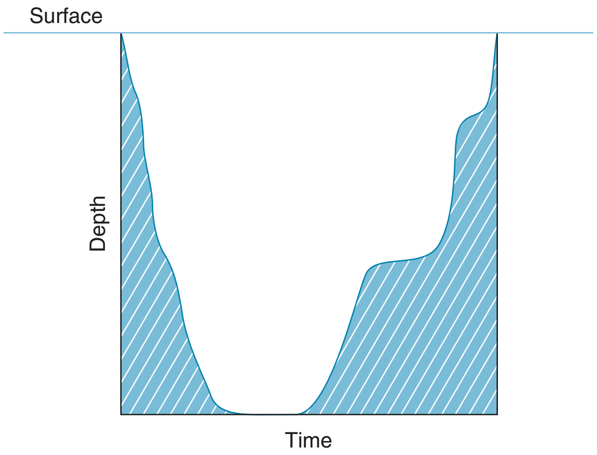

Figure 7.8 The striped area represents penalty time imposed by dive table procedures.

must carry your dive computer with you when you dive because it provides information about your decompression status continuously and keeps track of your residual nitrogen.

Regardless of the type of dive computer you choose, you need to understand some basic principles. First, be sure to read the instruction manual that comes with your computer. Wait 24 hours after diving with dive tables before using a dive computer to ensure that you eliminate all residual nitrogen from your body. Do not exceed the ascent rate specified by the manufacturer, and do a five-minute rest stop at a depth of 15 to 20 feet (4.6 to 6.1 m) at the end of every dive. If you exceed the rate of ascent specified for your computer, extend the duration of your precautionary stop by at least the amount of time it should have taken you to ascend to the stop depth. Do not make repetitive dives deeper than 80 feet (24 m). Keep your dive computer activated until its outgassing is complete. If your computer fails at any time while you are diving at a depth in excess of 30 feet (9 m), terminate the dive immediately with a rest stop. If your computer fails or if you switch it off accidentally, discontinue diving for 24 hours.

If you exceed the maximum time limit for a dive, decompression is mandatory. The computer displays a ceiling—the minimum depth to which you may ascend. As you decompress, the ceiling becomes shallower until the computer indicates that you may surface. After a minimum interval at the surface, the computer begins scrolling—displaying the time limits for various depths sequentially. Wait 24 hours before diving again after any dive for which a ceiling was displayed.

The concept of backup planning with dive tables is good for square profiles; however, for multilevel dives planned with a computer, the use of the dive tables for backup planning is not feasible.

Modern dive computers offer many features. Here are some examples of the features that a computer may include:

- Operating modes for various breathing gases
- Decompression stop data
- A wireless transmitter that provides both cylinder pressure and remaining air data
- Dive profile lifetime history memories

Dive Computers

Advantages

- The maximum-depth penalty of the dive tables is avoided.
- Dive profile information is accurate.
- Dive computers store a record of the dive profile.
- You can eliminate common errors made when using manual dive planners.
- Dive computers offer additional features, such as an ascent rate indicator.

Disadvantages

- Dive computers are electronic instruments that can fail.
- Dive computers are expensive to purchase and service.
- Each diver must have a separate computer.
- The mathematical model varies from one type of computer to another, so some confusion results when each diver of a buddy team uses a different type of computer.

- A multiple-step ascent rate indicator
- Altitude adjustment
- A dive-planning mode
- Backlighted displays
- Audible alarms

An integrated, wireless-transmitting, digital-compass dive computer provides dive-planning information before, during, and after your dives. This type of computer is a highly desirable device. Central processing units (CPUs) in closed-circuit rebreathers (CCRs) are also sophisticated dive computers that provide highly technical dive-planning information.

The advantages of dive computers greatly outweigh the disadvantages, and computers have become an almost essential item for scuba divers.

Postdive Review

After a dive, you and your buddy should reflect on your experience. How closely did the actual dive match the dive that you planned? If there were deviations from the plan, what caused them? Could you have prevented the deviations with a different plan or approach? What changes can you make to improve the next dive? Some problems require research, or you may need to ask the advice of a diving professional. The experience of each dive should affect your plans for future dives. Your dives with your buddy should progress more smoothly each time you dive together. Each time you visit a dive site, your dive procedures should improve. A review of each dive with your buddy and a discussion about future diving are valuable parts of dive planning. Even if you have a new dive buddy for a dive, you should plan your dives together and discuss the experience afterward. It is a good idea to keep notes about your dives on a waterproof slate and transfer unusual experiences to your dive log along with the usual data. Reflecting on past experiences from time to time will help you become a better diver.

Summary

The saying “If you fail to plan, you plan to fail” is true for scuba divers. All phases of dive planning are important and help ensure enjoyable and successful underwater experiences. Follow the recommended steps for planning, get area orientations when appropriate, have contingency plans, and discuss your dives with your buddy.

Dive profiles are a large portion of dive planning. Plan to dive conservatively. No dive-planning device can guarantee that you will not develop decompression illness after diving. The deeper, longer, and more frequently you dive, the greater your risk of decompression illness. Limit multiple-day and multilevel dives because repetitive multiple-depth profiles make you more susceptible to decompression illness. Make a rest stop at the end of every dive, avoid improper profiles, and have surface intervals of an hour or more. After three consecutive days of repetitive diving, wait a full day before diving again. Exercise good judgment and common sense.

Application-of-Knowledge (AOK) Questions

1. What is the most difficult task of dive planning?
2. How can you make dive planning easier and be more likely to accomplish it?
3. You have a dive planned, the weather and water conditions are good, and you catch a cold virus the day before your dive. Cold medicine minimizes your congestion. What should you do?
4. A friend who has more diving experience than you invites you to dive on a shipwreck that is located at a depth of 120 feet (36.6 m). How should you respond to his invitation?
5. Would you rather plan a dive by using a dive computer or by using dive tables? What is the advantage of each approach?
6. You have been diving at a resort for a week. None of your dives have required decompression other than safety stops. Your plane leaves the day after tomorrow at 10 a.m. When should you stop diving?
7. If you have to go over a mountain pass with an altitude of 5,000 feet (1,525 m) to return home after diving, what action do you need to take?

Sergeant major fish in Cozumel

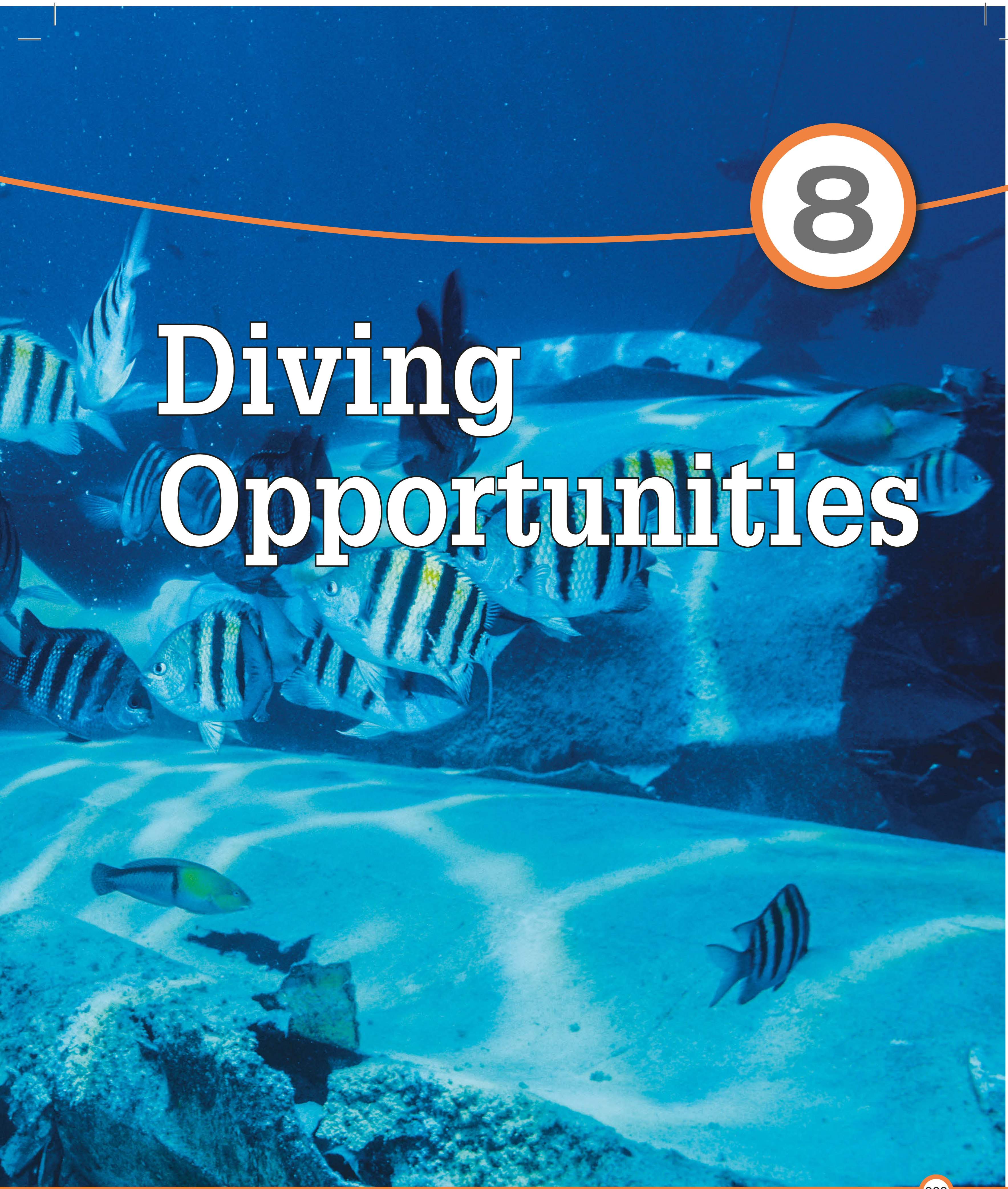

8

Diving Opportunities

Dive In and Discover

By the end of this chapter, you will be able to do the following:

- List 11 special-interest areas for novice divers and 6 special-interest areas for advanced divers.
- List at least five ways to continue your scuba diving education.
- List at least three ways to get involved with the local diving community.
- List several actions to take when preparing for a dive travel trip.

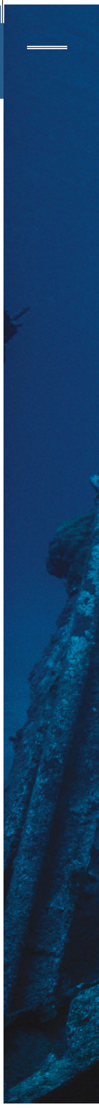

Continuing Education

A good diver never stops learning. You cannot learn everything you need to know about diving during a single course of instruction. After you complete your entry-level course, you can enroll in an intermediate or advanced scuba course immediately. These courses allow you to gain additional experience under supervision, help you develop important skills (such as navigation), and introduce you to special interest areas of diving. When you have identified a diving specialty you would like to pursue, you should complete a specialty course for that topic. A specialty course helps you begin enjoying the special area from the outset and helps you avoid mistakes and injury. Specialty areas that are of interest to new divers include the following:

- Underwater photography
- Underwater hunting and collecting
- Underwater environment
- Scuba lifesaving and rescue
- Night diving
- Boat diving
- Dry suit diving
- Drift diving
- Cavern diving
- Altitude diving
- Nitrox diving (enriched air)

Chestnut cowry by a starfish in California

As a qualified scuba diver, you can do many things. You can learn more, dive in different areas, be recognized for expertise, help others, or earn money, just to name a few. Opportunities abound.

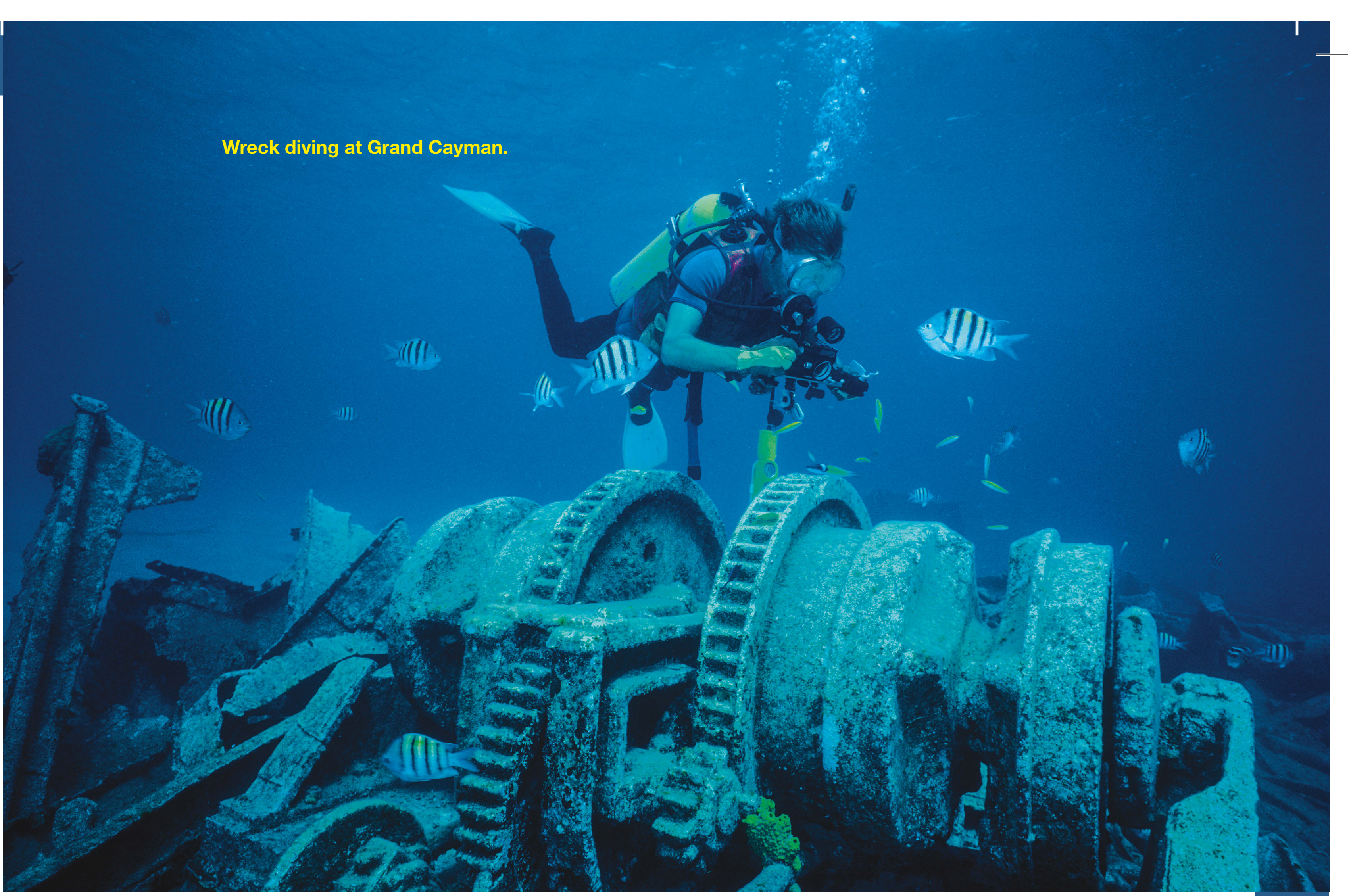

Wreck diving at Grand Cayman.

The following specialty courses may be of interest to experienced scuba divers with advanced training:

- Wreck diving (penetration)
- Ice diving
- Cave diving
- Deep and wall diving
- Light salvage
- River diving
- Search and recovery
- Rebreather diving
- Mixed-gas diving

Courses are only one way to learn about diving. Many diving seminars, workshops, and conferences are also available. Check the calendar section of diving publications to learn of events scheduled in your area. Diving changes constantly, so you need to keep updating your knowledge of diving medicine, diving equipment, and diving procedures. Continuing education programs give you the opportunity to learn from professionals who are continually involved in diving.

You can learn more about scuba diving by reading books, magazines, and newspapers. Subscribe to a scuba periodical and read all you can about diving. The appendix contains a list of scuba periodicals. You can also visit websites by and for scuba divers.

Another good way to continue your diving education is to join a dive club. Dive clubs offer many benefits to divers, including education. A talk on an interesting subject by a good speaker is usually part of the monthly meeting of a local club.

Continuing your education is important for your safety and enjoyment. When you have increased your knowledge about diving, you may become interested in helping others to learn.

Local Opportunities

You do not need to live in a coastal area to dive regularly or to get involved in diving. You can dive nearly anywhere you can find water, and there are dive stores, dive clubs, and diving events in noncoastal areas. Get involved in diving in your area right away. Find all the diving-related businesses and groups in your area: dive clubs, dive stores, dive boats, diving events, and diving publications. Join a dive club, attend the meetings, and participate in club dives. Find a local diving website to make contact with other divers in the area. Complete continuing education courses. Attend local diving seminars, workshops, con-

Underwater photography at Cozumel.

ferences, and shows. Talk to other divers when you attend diving events. Seize every opportunity to learn and every opportunity to dive with those who are more experienced than you.

Dive Travel

Many divers enjoy taking dive trips to exotic destinations, and there are thousands of beautiful and exciting diving locations to choose from around the world. If you plan to travel somewhere to dive, you need to know how to arrange dive travel, how to prepare to travel, and how to enjoy your trip.

Travel agencies that specialize in dive trips provide valuable information, so it is a good idea to book your trip through one of them. You might want to consider a dive package that includes accommodations and diving. Dive resorts advertise these packages in diving publications and often offer packages for bargain prices at dive shows. Many resorts have toll-free numbers you can call to obtain information about dive packages. To ensure the best value for your money, compare several packages before you book a trip. If a brochure from a dive resort is out of date, check with the resort for current offers.

When planning a dive travel trip, you need to decide whether you want to dive from a live-aboard dive boat or from a shore-based dive resort. A live-aboard trip allows you to dive in a variety of locations. The vessel moves from place to place throughout the day. You dive in remote areas that only a few divers visit. You may find a live-aboard dive trip to be a wonderful experience if diving is the only objective of your trip and you are not affected by motion sickness.

Consider a shore-based dive trip if you do not care for confinement and if you want to participate in activities other than diving. Many beautiful islands and resorts offer a variety of things to do in addition to diving. Nondiving members of your travel party can usually find many sources of enjoyment at a land-based diving destination.

Many diving publications and diving websites are available to help you decide where to dive, although making a choice from all the fantastic destinations available can be difficult. Recommendations from other divers, trips sponsored by your dive club or dive store, and travel presentations can help you decide.

When you have decided on the region for your dive trip, get as much information as possible about diving in the area. Obtain and review brochures, books, articles, videos, and any other materials about the region. The Internet is another excellent resource for information on dive locations. Talk to divers who have been to the location where you plan to go. The more you know in advance, the more enjoyable you can make your experience.

After you select your diving destination, you should make your reservations well in advance, confirm all arrangements in writing, and get clarification concerning cancellations, refunds, and so forth.

Research and plan your dive trip by learning about the water temperature for the time of year when you will be diving. Then obtain the appropriate exposure protection. Remember that too much insulation is better than too little. Also make sure that all your equipment is in good working order.

Plan to avoid sunburn. Sunlight in tropical areas is more intense than in other climates. Use a sunscreen, and unless you are well tanned, keep yourself covered at all times, even while in the water. Use lip balm with a sunscreen ingredient.

Underwater salvage operations.

You cannot develop a tan in a few days, so it is silly to try if you do not already have one. Sunburn can ruin an expensive vacation, and donning an exposure suit is painful when you are burned. It takes only a few minutes for an untanned person to get burned in the tropics, so be careful!

Obtain all required documentation well in advance. Obtaining passports and visas can take months. Be sure to find out whether the country to which you are traveling requires immunizations.

Learn what to expect when you arrive at your destination. Know the frequency and voltage of the electricity; take converters if you are bringing electrical devices that operate on a different voltage. Know the monetary exchange rate. Consider taking a small pocket calculator to help compute the conversion of money.

If you have an expensive camera, video equipment, or jewelry, take your property to a customs office before you leave the country and have it documented to avoid being charged duty on the items when you return home.

Do not procrastinate in making your preparations. Procrastination can cause a cancellation of your trip, a forfeiture of your deposit, and a great deal of frustration. Consider travel insurance.

Limit what you pack for your trip. You do not need much clothing for a dive trip unless you plan to attend formal events. A few pairs of shorts and T-shirts, some swimsuits, and a couple of sets of nice clothes for dinner should suffice. Experienced dive travelers travel light when it comes to clothing and personal effects.

Diving destinations provide tanks and weights, so you do not need to take those items with you. When you pack for your trip, keep in mind that there may be weight limitations for your baggage. Excess baggage costs can be very high. To guard against theft, avoid flaunting expensive diving, photography, and video equipment. Use inconspicuous containers to ship your equipment, and it will be less likely to disappear. Insure luggage that contains expensive equipment. Items I always take with me are mask, snorkel, fins, exposure suit, boots, gloves, regulator, buoyancy compensator, and dive computer. I rent tanks and weights. My gear can be packed in a dive bag, but I prefer to use a suitcase for better protection of the equipment. I pack a dive gear bag in the suitcase because I don't want to use a suitcase at the dive site. There are also dive gear bags that have wheels, which are useful when you can roll the bag.

Be prepared for travel illnesses. Obtaining medications for nausea, diarrhea, and colds is easier before a trip than it is if you become ill in a foreign land.

Flying is the most common method of transportation for dive travel. Air travel usually produces jet lag and dehydration. A viral illness a couple of days after reaching your destination is common because many virus germs are concentrated inside the cabins of airliners. Here are some tips and suggestions concerning air travel:

Schedule your flight to arrive the day before you start diving. It is unwise to plan to dive the same day you arrive.

Drink a full glass of water or fruit juice every hour during your flight to avoid dehydration. Avoid alcohol, milk, and drinks containing sugar. The humidity in an aircraft is about 8 percent, so you will become dehydrated unless you drink plenty of fluids.

Avoid eating heavy meals and salty foods when flying. Light foods cause less dehydration than fatty foods. Consider ordering a special meal of salad and fruit, or bring your own food if meals aren't served on the flight.

Wash your hands frequently when traveling. Germs from your hands can get into your eyes and nose and lead to viral infections.

Remember that baggage compartments on airliners are unpressurized and that the reduced atmospheric pressure can damage your gauges. Therefore, you should either bring your gauges inside the cabin in your carry-on luggage or store them in airtight containers.

As soon as possible after reaching your destination, get some exercise and drink plenty of liquids. Limit your consumption of alcohol, which causes dehydration. Getting intoxicated at the outset of a diving vacation is one of the worst things you can do because you increase your chances of getting DCS.

Confirm your return air reservations as soon as you arrive, especially if you travel to another country. If you fail to do this, you are likely to lose the reservations. You could end up arriving at the airport to go home and discovering that your tickets are not valid.

Your first dive at your destination should be an orientation dive. In addition to learning about diving in the area, you should use the dive to check your buoyancy and make any needed adjustments. Work out any equipment problems during a dive in shallow, still, calm water. Avoid deep dives or moving water until you have acclimated to the area.

After three days of repetitive diving, refrain from diving for a day so that your body can outgas. Go shopping or take a land tour for a refreshing change of pace. Allow one day between your final scuba dive and your scheduled flight home.

SCUBA WISE

Scuba equipment is a means to an end—a mechanism that enables you to go to areas that are otherwise unreachable by the average person. I think of scuba diving as a means of transporting myself to a destination where I can explore or participate in a special activity. Scuba diving without an objective loses its thrill after only a few dives. You need to develop an interest area, which can be anything from bottle collecting to wreck diving. When you have an objective for your dives, diving can be extremely rewarding. Over the years, I have participated in many different diving activities. My favorite underwater pursuit is photography. I can collect priceless memories without removing anything from the environment. In my opinion, underwater photography is the most challenging and the most satisfying dive activity. Select a diving activity that appeals to you, complete training for the activity, and then pursue it. You may choose several different activities. No matter what you choose to do while diving, I hope that your experiences will be as wonderful as mine have been.

The nondiving day before flying helps prevent DCS caused by altitude, and it gives you the opportunity to rinse, dry, and pack your diving equipment. Snorkeling is a good last-day activity, and your mask, snorkel, and fins dry quickly for last-minute packing.

Dive travel can be fun, exciting, and memorable when you plan it well. You can avoid unpleasant, frustrating, and disappointing experiences by researching your destination and preparing properly. Good diving to you!

Dive Destinations

One of the great pleasures of scuba diving is visiting different regions of the country and the world. Fantastic dive sites can be found nationally and internationally. The following areas are popular dive destinations that you might want to visit. The pictures throughout this book were taken at these destinations.

California

Dense kelp beds that form surface canopies make diving in California an unforgettable experience. The kelp provides shelter for many forms of life. Rock formations provide anchorage for the buoyant kelp and support additional life-forms, such as abalone, lobster, and moray eels. An expedition on a live-aboard dive boat to one or more of California's offshore islands during the summer months is an extremely worthwhile scuba experience. Shore diving in California can be difficult because of surf and low visibility. An area orientation is essential before attempting to dive through surf.

Washington

Puget Sound, which was carved into the landscape by glaciers, forms a gigantic body of ocean water that is much more protected than the coastal waters. The Sound contains many shipwrecks and artificial reefs, which are havens for many life-forms. Puget Sound is the home of the world's largest octopus as well as giant lingcod, monkey-faced eels, and beautiful orange and white anemones called metridians. The water is cold, but the visibility is 15 to 20 feet (4.5 to 6.1 m) most of the time. Dry suits are recommended.

Florida

Florida features some of the best and most diverse diving in the United States. From crystal clear springs and rivers to the Keys, there are many spectacular dive sites. Cavern and cave diving are popular, but training for these specialties is mandatory. The Florida coast has abundant shipwrecks. The Keys boast coral reefs and colorful fish. All scuba divers should definitely have a dive excursion to Florida included in their dive log.

New England

The cold waters of the New England area offer a great variety of diving opportunities. Wrecks, reefs, kelp, and creatures captivate your attention. Lobsters with claws and other interesting life-forms are great attractions. If you are able to schedule some time to dive when you travel to this part of the country, you should try to include the adventure in your itinerary.

Great Lakes

These great bodies of freshwater boast the most well-preserved shipwrecks in the United States. Many of these wrecks are in deep water, so you must have specialty training to explore them. Divers who are unfamiliar with the Great Lakes should use the services of a divemaster. If you like wreck diving, this is one of the places to include on your must-see list.

Hawaii

Hawaii offers exciting diving that includes good visibility, spectacular underwater formations, beautiful reef fish, large turtles, and giant marine mammals. Hawaii is a popular diving destination that you should visit if you get the chance. Aloha!

Caribbean

The Caribbean is one of the world's most popular places to dive, and for good reason. The area includes reefs, wrecks, drop-offs, and life-forms that are unique and breathtaking. Visibility is usually excellent. Many dive resorts in the area offer trips to phenomenal locations. Nearly every experienced diver talks about Caribbean diving experiences.

South Pacific

The islands in the South Pacific offer some of the most beautiful diving to be found anywhere on Earth. Divers can see fantastic soft corals, sea fans, and incredible fish. Visibility is outstanding. Many of the island waters feature wrecks and war artifacts. The beauty found above and beneath the water in this part of the world makes trips to the South Pacific highly desirable for scuba divers.

Underwater photography in Southern California

Middle East

The diving destination in this region is the Red Sea, which is unparalleled for scuba diving. The shear walls, the abundant and colorful life-forms, and the crystal clear waters make this one of the most desirable of all diving destinations. If you can afford a trip on a live-aboard charter boat to offshore islands, you will be treated to unforgettable sights.

To learn details of the world's most popular diving destinations, you can do a search of the topic on the Internet. Many websites provide excellent information. Traveling to exotic dive destinations is one of the most enjoyable aspects of scuba diving.

Career Opportunities

Some people love diving so much that they want it to be their vocation. Career opportunities exist for those who are willing to put forth the effort to achieve their desires. Jobs within the recreational dive industry include resort dive guide, instructor, journalist, dive travel coordinator, retail salesperson, and sales representative for equipment manufacturers. People can also work in nonrecreational scuba jobs, such as scientific research, archaeology, engineering studies, hull cleaning, salvage, and underwater repairs. Two of the same rules that apply to recreational pleasure diving also apply to underwater work: (1) Complete training for an activity before you attempt it, and (2) use complete, proper equipment for the activity. Career training opportunities abound. If you are interested in diving as a vocation, you can find information about training in dive periodicals.

Night diving preparations at Catalina Island, California.

Leadership Opportunities

After you have become an experienced diver, you may want to help others learn and enjoy scuba diving. You can take courses to become qualified as a divemaster, assistant instructor, and instructor. Many people who love diving work part-time in the industry in leadership positions. When you are ready for leadership training, contact your scuba training agency for information.

Summary

Always remember that a good diver never stops learning. You should complete additional training, make sure you are trained for any special activity before you attempt the specialty, read as much as possible, join a dive club, and attend educational events. Take advantage of every opportunity to learn more about diving.

Increase your diving abilities and experience by applying what you learn. Participate in dive trips and charters. Research and plan your excursions thoroughly to maximize your enjoyment and minimize your disappointment. Many opportunities are available for diving adventures throughout the world.

Another opportunity for you as a diver is to make a positive contribution to the diving community. You can do this by conducting yourself in a responsible manner at all times and encouraging other divers to act responsibly. Help establish a good image of diving, get involved in issues that pertain to the diving environment, and help educate people who do not dive. Whether your contribution is small and personal or vast and international, you can make a difference. Decide now that your involvement in the wonderful pursuit of scuba diving will be a positive one.

Application-of-Knowledge (AOK) Questions

1. Why are you learning to dive? Do you just want to investigate a new activity, or are you interested in an activity that requires scuba diving?
2. What do you visualize as your progression of training as a scuba diver?
3. How often do you think you will dive? Will most of your diving be local or related to dive travel?
4. If a nondiving friend asked you what can be done to reduce the risk of injury while scuba diving, what advice would you provide?
5. What causes you the greatest concern about scuba diving? What can you do to minimize your concerns?
6. How do you see yourself as a scuba diver in 5 years? In 10 years?

This page intentionally left blank.

Appendix A

Associations and Periodicals

Diver Training Organizations

International Diving Educators Association (IDEA): www.ideascubausa.com/home_15.html

Multinational Diving Educators Association (MDEA): http://neptuneshop.wix.com/mdea#!

National Association of Scuba Educators (NASE): http://naseworldwide.org

National Association of Underwater Instructors (NAUI): www.naui.org

Professional Association of Diving Instructors (PADI): www.padi.com/Scuba-Diving

Scuba Diving International: www.tdisdi.com

Scuba Educators International (SEI): www.seidiving.org

Scuba Schools International (SSI): www.divessi.com

Environmental Organizations

Blue Frontier Campaign: http://bluefront.org

Cousteau Society: www.cousteau.org

Earthecho International: http://earthecho.org

Living Oceans Society: www.livingoceans.org

Marine Conservation Society: www.mcsuk.org

Oceana: http://oceana.org

Ocean Defense International: www.oceandefense.org

Oceanic Preservation Society: www.opsociety.org

Reef Check: www.reefcheck.org

Reef Relief: www.reefrelief.org/mission

Save Our Seas: http://saveourseas.com

Sea Shepherd Conservation Society: www.seashepherd.org

Coral Reef Alliance: http://coral.org

Ocean Conservancy: www.oceanconservancy.org

The Ocean Project: http://theoceanproject.org

NOAA National Marine Sanctuary Project: http://sanctuaries.noaa.gov

Scuba Diving Magazines

Advanced Diver: www.advanceddivermagazine.com
Alert Diver: www.alertdiver.com
Dive Training: https://dtmag.com
Diver: http://divermag.com
Scuba Diving: www.scubadiving.com
Undercurrent: www.undercurrent.org

Appendix B

AOK Answers

Chapter 1

1. Some people learn to scuba dive because they seek a new experience or adventure. Scuba diving training opens the gate to a beautiful underwater world, which can provide many pursuits. Scuba diving should be considered a mode of transportation that takes you to places where you can pursue underwater interests such as exploring, photography, and wreck diving. After completion of your basic scuba diver training, continue your education and obtain an advanced scuba diver rating. With the additional skills acquired with advanced training, you will be ready to learn specialty activities and seek the rewards of your chosen endeavors.
2. An open-circuit scuba system allows a diver to breathe compressed air from a pressurized cylinder and exhaust the used gases into the water. It is a simple, reliable, easily maintained air delivery system, which makes it the safest and most popular scuba system. A closed-circuit scuba system allows a diver to rebreathe exhaled air by removing carbon dioxide and adding the correct amount of oxygen without allowing gases to escape from the system. It is a complex air delivery system that requires specialty training and extremely careful maintenance. Closed-circuit scuba systems are very expensive compared to open-circuit systems.
3. A scuba certification card (C-card) indicates that you have completed initial training, but it does not reflect how recently you have been diving. Certification agencies recommend that you complete simple refresher training if you have not dived for a year or longer. This training is a review of your basic skills under the supervision of an instructor during an open-water dive. The confidence you gain from knowing that your skills are refreshed will make your diving safer. A good reason to log your dives is to document your diving activity. Some dive resorts and charter boats will want to review your log book to determine your diving experience, especially your most recent dives. Logged dives also can help you recall many facts that you would otherwise forget.
4. It is logical that the more time you can use your knowledge and skills of scuba diving under the supervision of a certified diving instructor, the better your knowledge and skills will become. Extended training also allows you to ponder your learning between training sessions. You have time to think of questions that will clarify concepts and increase your understanding of principles. Short scuba training courses do not provide these benefits. The goal of learning is to transfer knowledge and skills acquired in a controlled environment to the actual environment in which they will be used. Longer scuba courses develop a greater ability to transfer your

knowledge and skills, which increases your safety for your independent diving after training.

5. A medical exam for scuba diving should be performed by a scuba diving physician who understands medical issues that could be problematic for those considering scuba training. Knowledge of how pressure affects medical conditions can help prevent pain, suffering, and even death.
6. Actions you can take to minimize your risk of injury while scuba diving are to dive only when you are in good health, use well-maintained equipment, dive in good environmental conditions, and adhere to the safety rules of scuba diving.
7. When you are in the process of selecting an instructional course for scuba diving, you need to know the sanctioning national agency, the instructor's experience, the total cost of the course, the course curriculum and duration, and how much open-water (real environment) training is included.
8. You can show that you are a responsible diver if you do only what you are qualified to do, dive within your personal limitations, adhere to safety rules, and remain aware of how your conduct will be viewed by others—divers and nondivers—when you go diving.

Chapter 2

1. Your exertion has used much oxygen, which has produced much carbon dioxide (CO_2), which stimulates you to breathe more to expel the CO_2. The feeling of a lack of air is caused by the need to reduce the amount of CO_2 dissolved in your blood. You need to minimize your activity immediately and breathe deeply until your breathing returns to normal.
2. Rapid, shallow breathing (hypoventilation) moves only a small amount of air in and out of your body. The structures from your mouth to your lungs do not allow gas exchange, so your body is not getting oxygen or ridding itself of CO_2. Gaining control of your breathing by stopping or reducing activity and breathing deeply until you feel better is the correct course of action and can help prevent panic. When your breathing is under control, you can think clearly and manage the difficulty that caused your anxiety.
3. Many factors affect air consumption, but the primary cause of high air consumption is exertion. Relax and move slowly. If you are too buoyant and have to exert to maintain your position in the water, your air supply will be depleted rapidly. If you are wearing too much weight, your attitude in the water will increase your frontal surface area and drag will increase air consumption. Adjust your weight and trim so you can hang suspended in a level position, and then relax and enjoy diving.
4. The first step is to tell yourself that you have time to manage this problem. The next step is to think of the options available to you. The quickest and easiest solution is to switch to your extra second stage or alternate air source. Your buddy's extra second stage is another good choice. The worst way to manage this problem is to spit out your regulator and swim to the surface.
5. Rapid and deep breathing before holding your breath is hyperventilation, which flushes carbon dioxide from your body. The buildup of carbon

dioxide when you hold your breath is what stimulates you to breathe. If you eliminate too much carbon dioxide from your body, you may lose consciousness while breath holding before you are stimulated to breathe. Loss of consciousness underwater can cause drowning. Hyperventilation should be limited to a few breaths before underwater breath holding to avoid what is called shallow-water blackout.

6. Your body loses heat rapidly in water. When you increase your heat production by exercising in cold water, you lose more heat than you would if you had not increased exertion. If you are wearing a wet suit, ascending to shallower water will increase your insulation as the water pressure decreases. Avoid moving directly into a current, which will push cold water through your suit. Your best action is to obtain better insulation for your entire body for future dives.

Chapter 3

1. The current is pushing cold water into and through your wet suit. If you swim at an angle to the current, you will reduce the flow of cold water through your suit and not get as cold. You can also look for places where you can be shielded from the current. You do not need to swim constantly when diving. Most of the life in the ocean is less than 1 inch (2.45 cm), so it is a good idea to stop and look closely at any underwater formations.
2. Some of the best and safest ways to warm yourself after diving are to remove your wet suit, get dry, get out of the wind and into a warmer environment, wrap yourself in a blanket or sleeping bag, eat high-energy food, drink warm drinks, and avoid stimulants. The principle is to warm yourself from the inside out instead of using external heat, which is dangerous. Remember that your core temperature will not be normal until you begin to perspire.
3. Did you think about the density of water and how much resistance to movement it causes? People usually think they will ascend rapidly if they lose their weights, but if you respond quickly, this situation is not difficult to manage. Quickly deflate any air from your BC. Then invert yourself to get your fins above your head, swim downward forcefully for a few kicks, grasp your weight belt, and replace it. (You will learn how to do this skill in chapter 6.)
4. When you ascend 10 feet (3.3 m) to swim over the reef, your buoyancy will increase because your wet suit and the air in your BC will expand. If the upward excursion is brief, you may be able to control your buoyancy just by keeping the amount of air in your lungs minimal. If the excursion is long or you are unable to control your buoyancy by breathing, vent some air from your BC to maintain neutral buoyancy until you descend again.
5. Remember that it is easier to equalize pressure in your ears if you descend feet first. It is good to use an anchor line for reference and control during a descent, but remain upright as you make your way down the line.
6. You should recall that the rate of change of pressure is greatest near the surface. Therefore, you need to equalize pressure in your ears and mask more frequently near the surface than will be required as the depth of

descent increases. If you can keep your air spaces equalized for the first 33 feet (10 m), you are not likely to have problems beyond that depth.

7. If you don't stop swimming, you will experience air starvation. And if you do stop swimming, your buddy will disappear and you may become separated. Prevention is best. Get your buddy to agree that you will swim side by side at all times and that when one of you stops, you both stop. If your buddy is ahead of you and you want or need to stop, use a rock or the butt of your dive knife to bang on your tank to get your buddy's attention. If your efforts fail, stop and get your breathing under control and then surface to reunite with your buddy if he does not return to you.
8. You are out of air, but only at a depth of 50 feet (15 m). As amazing as it sounds, if you start swimming slowly to the surface, there will be additional air to breathe from your tank as the surrounding pressure decreases with ascent. This is not the best option, however. In chapter 6 you will learn other ways to deal with situations involving loss of air supply.

Chapter 4

1. Minimize sun exposure, soak and rinse after use, store in a dark and dry location that is free of smog, inspect your gear regularly, and have your equipment serviced regularly or as needed.
2. The most important concerns for most diving equipment are fit and comfort. Features, colors, and price do not matter if the equipment does not fit well or is uncomfortable to use. If possible, rent equipment before you make a purchase so you will know if an item is right for you.
3. If all of the air escapes from your tank and the valve in your second stage is open (or if you happen to depress the purge button and open the valve), water can enter your tank. You will learn how to minimize the chances that your regulator will free-flow. If you detect water inside you scuba cylinder, take it to a facility for maintenance as soon as possible.
4. Air has weight. The air in a typical scuba tank weighs five to six pounds (~2.25-2.7 kg) when the tank is full. As you deplete the air, your buoyancy increases. The loss of weight may not be apparent at depth if you wear an exposure suit because pressure compresses the suit. You simply add less air to your BC than you would if you were not wearing an exposure suit. Problems arise as you near the surface with a scuba tank that is nearly empty because your suit expands and you are unable to overcome the positive buoyancy caused by the loss of weight from your cylinder. To avoid this problem, you need to adjust your weight to be neutrally buoyant just below the surface while using a tank with approximately 500 psi of air remaining in the tank.
5. You need just enough air in a dry suit to keep it from squeezing you. Any excess air forms a bubble inside the suit, and the bubble rises to the highest point. If you tip upside down to do something, the bubble can rise to your feet, may loosen your fins, and can cause you to float upward in an inverted position. This problem is just one example of why you need special training in the use of a dry suit.

6. A compass is a magnet, and magnets attract one another. Placing your compass near your buddy's compass causes the needles to shift position and the readings to be inaccurate. You need to keep your compass away from any items that can cause the needle to deviate from its natural direction.
7. The best way to clean your diving equipment is to soak it in warm water, rinse it gently with flowing water, and dry it thoroughly before storing it. Some equipment parts, such as dry suit zippers, require scrubbing. Follow the manufacturer's instructions.
8. The attachments to a scuba regulator first stage may include a cylinder pressure gauge, a primary second stage regulator, an alternate air source second stage, a BC inflator hose, and possibly a dry suit inflator hose.
9. The best reasons for having good dive equipment and maintaining it properly are safety and comfort. You will know when your equipment is right when you are not concerned about it while you are diving.

Chapter 5

1. The water temperature affects the equipment you use, your body temperature, your diving procedures, the visibility underwater, and the underwater life, among other things. Water temperature has a profound impact on your diving adventures. Knowing what to expect and how to be prepared for the temperature of the water is essential.
2. Water visibility is affected by depth, pollution, water movement, weather conditions, algae, temperature, your movements, and other factors. Check present diving conditions with your local dive facility and postpone diving when visibility is poor.
3. A good way to manage currents is to know in advance what to expect. Obtaining training for diving in various currents is important. Being aware that the current is weaker at the bottom than at the surface is helpful. Swimming at right angles to a current may help you get out of some currents. Diving against the current is a common practice, but in some situations it is best to drift with the flow of the current.
4. When you encounter a strange animal underwater, respond by stopping movements, respecting the animal, observing it from a distance, and refraining from provoking or feeding the animal. Nearly all marine animals are defensive, not aggressive. Think and respond. The worst action to take when you see a strange or large animal is to panic and ascend rapidly to the surface.
5. No matter how much training or experience you obtain, you will always need to acquire local knowledge about diving in a new area. Watching and mimicking experienced local divers are a good way to avoid embarrassment, but learning in advance about why the divers use the techniques they use is even better.
6. While it may be difficult to accept until you gain experience, the safest and most comfortable place to be while diving is near the bottom. It is much more uncomfortable and dangerous at the surface of the water.

7. The transition from saltwater diving to freshwater diving requires changes that include the amount of weight you wear and perhaps the type and thickness of your exposure suit.
8. When you plan to dive in a new area, you need information about the area. Weather conditions, seasons, forecasts of diving conditions, area hazards, equipment requirements, and local diving procedures are important to know in advance. Chapter 8 provides resources to help you obtain the information that you need.

Chapter 6

1. Buoyancy can be controlled by the amount of weight worn, the amount of air in your BC, and the amount of air in your lungs. Think of these methods as large, medium, and small ways of control, respectively.
2. To ascend properly from a scuba dive, alert your buddy, hold your BC inflator-deflator with the mouthpiece facing down and the valve open, reach up, swim up, ascend slowly, monitor your rate of ascent and your buddy, turn slowly, stop at a depth of 10 feet (3.05 m) for precautionary decompression, ascend to the surface, and make a complete turn to assess your surroundings. Ascending is a complex skill. Learn it well.
3. Successful buddy contact requires predive planning and communications, agreeing on a direction and maintaining that direction until you both agree on a new direction, maintaining the same position (side by side) relative to one another, holding hands or using a buddy line when visibility is poor, and committing to reunite using standard procedures if you become separated.
4. This situation can be frightening unless you respond quickly. All you have to do is raise and open the deflator valve with one hand and disconnect the low-pressure hose with the other hand. You do not need to terminate your dive because you should have the skills needed for manual inflation of your BC.
5. When a regulator free-flows and will not stop, simply switch to your extra second stage and ascend. Remove your tank at the surface and turn off the air supply. If you allow the tank to empty and the second-stage valve is stuck in the open position, water can get inside your cylinder.
6. To exit the water into a small boat, remove your weight belt and place it in the boat, inflate your BC about half full, and remove your scuba unit. Next, grab the rail of the boat and, while holding the regulator hose from your tank under one hand, kick yourself up and into the boat. Remove remaining equipment after you are in the boat.
7. Swimming against a current can be tiring. If there is a nearby obstruction, such as a rock formation, that you can get behind to minimize the current and rest, seek shelter. If there is an outcropping you can grasp that will allow you to rest and catch your breath, use it. If the bottom is sandy, plunge your knife into the bottom to maintain your position while you recover. The most important action is to reduce your exertion and get your breathing under control. If the situation is unmanageable, surface and drift back to your exit point.

8. The assistance of your buddy is the best way to get free of an entanglement, but you can also remove your scuba unit, clear the entanglement, and replace your equipment. Be sure you learn the skills of equipment removal and replacement well so you can manage minor problems that may occur while scuba diving.

Chapter 7

1. The most difficult task of dive planning is getting started! Most people agree that planning is a good idea, but they procrastinate. You can be successful with almost any task if you push yourself to get started. If you keep a list of things to do, consider adding dive planning to that list.
2. To make dive planning easier, find a buddy who would like to dive with you and agree on which items each of you will plan. Set a date to talk again about your plan. By taking responsibility and setting a date, you will have less work to do and will be more likely to meet your goal of developing a good dive plan.
3. If your health is not normal, do not dive. Congestion from a cold can make it difficult to equalize pressure in your ears and sinuses. Pressure can cause the medicine to terminate sooner than it normally would. When decongestants stop working, air passages can swell and close rapidly and trap air. Notify your dive buddy that you need to postpone your dive trip until you are well again.
4. Assuming you are a novice scuba diver, it would be unwise to attempt a dive on a wreck in deep water. Shipwrecks often occur where conditions, such as currents, can be treacherous. Tell your friend that you are extremely interested but that you need to complete advanced and specialty training so you will be adequately prepared for the dive.
5. When you dive using a dive computer, you avoid the dive table penalty rule of having to assume that all of your dive time was spent at the maximum depth. The computer also stores valuable information about your dive, helps you control your rate of ascent, and indicates when you need to stop for decompression. The disadvantage of a dive computer is that it is an electronic instrument that can fail. Dive tables do not need batteries. The best approach to planning time and depth limits is to develop a contingency plan using dive tables, write the plan on an underwater slate, and take it with you when you dive with your computer. If your computer stops working or you inadvertently turn it off, your contingency plan can help keep you safe.
6. You need to plan. After multiple dives, you should wait 18 hours before flying, so your diving needs to end today.
7. When driving decreases pressure with altitude after you have been diving, you need to consult the altitude delay timetable. If your driving altitude reaches 5,000 feet and your starting repetitive-dive group is 1, you need to wait 3 hours 4 minutes before driving over that pass. This requires planning. If your altitude delay timetable is at home in your book, you won't know how long you need to delay your journey.

Chapter 8

1. Some people learn to dive just to experience the activity. They complete training, purchase equipment, make a few dives, and stop diving. Other people continue training, get involved in the diving community, pursue an underwater hobby, and dive for many years. If you think of scuba diving as a means to an end, it can provide wonderful experiences that will keep you involved for many years.
2. You might see yourself completing an advanced diver course and several specialty courses. It helps to gain some dive experience in addition to your training dives. Just be sure that your excursions are limited to the conditions and depths to which you are qualified.
3. If you are not certain that you will dive regularly or that you will dive only locally, consider renting the major equipment (cylinder, regulator, exposure suit, buoyancy compensator, and weight belt). If you find yourself diving frequently, buy your gear. Renting equipment will help you make a good selection. If you plan to dive primarily at dive destinations, rent the major equipment at the destination. When you become an avid scuba diver, you are likely to have full sets of equipment for both cold- and warm-water diving.
4. The best ways to reduce the risk of injury from scuba diving are to be trained for what you do while diving, have the proper equipment for your dive, and gain experience gradually. Making a dive that greatly exceeds your training and limitations without first gaining several experiences that gradually increase your training and limitations is unwise.
5. Phobias are natural, but if something terrifies you and it is related to diving, learn how to deal with the fear before you are exposed to it. For example, many people fear sharks because of what they have seen in media. What you should know is that dangerous shark encounters are rare and most likely occur at the surface of the water. If you see a shark, remain calm, stay on the bottom, and enjoy the exciting experience.
6. When you look to the future, do you envision yourself diving regularly? Are you just going to wait and see what happens, or do you intend to continue your training, get involved, and become a frequent diver with many friends who also enjoy diving? You can dive for a lifetime and not even begin to see everything the underwater world has to offer. May all your diving experiences be enjoyable and memorable!

Glossary of Scuba Diving Terms

absolute pressure—The total pressure exerted on a diver, which is measured by adding atmospheric pressure to gauge pressure.

actual bottom time—The elapsed time in minutes from when a diver leaves the surface in descent until the diver begins a rest stop or surfaces.

alternate air source—A source of compressed air other than the diver's primary scuba regulator. The two primary types of alternate air sources are extra second stages and backup scuba units.

ambient pressure—The surrounding pressure.

Archimedes' principle—This principle states that the force of buoyancy acting on a submerged object equals the weight of the water displaced.

arterial gas embolism (AGE)—An embolism resulting from an air bubble blocking the arterial circulation.

atmospheric pressure—The pressure exerted by the atmosphere.

backpack—The equipment that holds the diver's scuba cylinder.

barotrauma—Trauma or injury caused by pressure.

bends—See decompression illness.

bezel—A movable ring on a waterproof watch (being used as an underwater timer) or compass that you can set to indicate elapsed time or direction.

blooms—Overpopulations of plankton that can color the water, destroy underwater visibility, and form toxins in animals that feed by filtering water.

blowout plug—A plug included in a submersible pressure gauge to relieve pressure in the housing in the event of a high-pressure leak.

Bourdon tube—A type of depth gauge that uses a thin metal tube formed into a spiral. The movement of the coil, linked mechanically to a needle, indicates the amount of pressure exerted on the gauge.

Boyle's law—This law states that for any gas at a constant temperature, the volume varies inversely with the absolute pressure, while the density varies directly with the absolute pressure.

buddy breathing—The sharing of a single regulator second stage by two divers.

buddy line—A short line used by two divers to keep in contact with each other when visibility is poor.

buddy system—The practice of teaming up with a qualified partner for activities that involve risk, such as skin and scuba diving. A buddy provides reminders and assistance and sees things that you might not see.

buoyancy—The upward force of water.

buoyancy compensator—An item of equipment that helps a diver control buoyancy. You can inflate your buoyancy compensator at the surface to increase buoyancy, deflate it to reduce buoyancy for descent, and add air to it to achieve neutral buoyancy underwater.

burst disk—A thin metal disk that is a standard feature of the valves for scuba tanks. If a tank is overfilled or the heat from a fire causes the tank pressure to increase to a hazardous level, the disk will burst and vent the tank to prevent an explosion.

capillary gauge—A type of depth gauge that is made up of a hollow, air-filled, transparent plastic tube sealed at one end and placed around a circular dial. Water pressure compresses the air inside the tube during descent (based on Boyle's law), and the position of the air–water interface inside the tube relative to markings on the dial indicates the depth.

cavern—A large, roomlike opening in a natural formation where light from the surface can be seen.

caves—Openings in a natural formation that extend farther than a cavern where no surface light can be seen.

C-card—The certification card you receive when you complete your training requirements as a scuba diver.

ceiling—The minimum depth that a diver cannot rise above without the risk of DCS.

ciguatera—Fish poisoning that results from eating fish that consume a certain species of algae.

closed-circuit—Refers to a scuba system that eliminates carbon dioxide from exhaled breath and resupplies oxygen without allowing gases to escape from the system.

compartments—Mathematical models used to estimate gas absorption and elimination by various areas of the body.

compass course—A series of headings on a compass that leads to a destination.

compass heading—A direction that is set on a compass.

console—A display unit into which several instruments can be combined. An instrument console attaches to the high-pressure hose coming from the regulator first stage.

continental shelf—An underwater area that extends from land and slopes gradually to a depth of about 600 feet (183 meters).

contingency plans—Plans that address possible circumstances that could force changes in the original plans for a dive.

controlling compartment—The area of the body that determines how long a diver can remain at a given depth. The determination is made by how quickly a gas diffuses from that compartment.

crest—The top of a wave.

Dalton's law—A law stating that the total pressure exerted by a mixture of gases is the sum of the pressures that would be exerted by each gas if it alone were present and occupied the total volume.

decompression illness (DCI)—A serious medical condition in which a diver has neurological symptoms on surfacing. Also known as the bends.

decompression stops—Precautionary stops made during ascent in order to reduce the risk of decompression illness.

defogging—The process of removing the film of oil from the surface of the glass lenses on a scuba mask in order to prevent the mask from fogging underwater. Commercial defogging solutions can help keep your mask clear while you dive.

dehydration—A physical condition resulting from the excess loss of body fluids.

density—Weight per unit volume.

DIN valve—A newer type of threaded outlet for a scuba tank valve that has a recessed O-ring seal and withstands higher pressures than a traditional O-ring valve. Tank pressures in excess of 3,000 psi (204 atmospheres) require a DIN fitting.

dive profiles—Diagrams used to plan a dive by plotting the time and depth of the dive.

diver's push-ups—A buoyancy evaluation used to learn buoyancy control. When the diver is weighted properly, a full inhalation will raise the shoulders while the fin tips remain on the bottom, and an exhalation will cause the shoulders to sink.

dolphin kick—A type of fin kick in which a diver holds both feet together continuously and exerts force against the water using a wavelike up-and-down motion of the body.

downwelling—A compensating downward current that occurs when wind blows along a coast where there is a steep drop-off near shore.

drag—A force that slows movement.

drift—The speed of a current.

drift dive—A dive in which the divers move with the flow of the current.

eardrum—A thin membrane that separates the ear canal from the middle ear and transmits vibrations to the inner ear via a series of small bones that are attached to the membrane.

ebb—The flow of water away from an area because of low tide.

eddies—Swirling currents that result from non-linear water flow

embolism—A blockage of circulation.

emergency decompression—A required delay in ascent that a diver must take if the actual or total bottom time (ABT or TBT) exceeds the no-decompression-stop limit for a dive.

equalization—The process of keeping the pressure inside an air space in or attached to the body equal to the ambient pressure.

eustachian tube—The part of the ear that allows the equalization of pressure in the middle ear.

Farmer Johns—A type of wet suit design where the leg material extends upward to cover the chest and is held in place with shoulder straps

fetch—The area in which waves were created.

fetch length—The distance wind travels unobstructed. The length of water over which a given wind has blown.

flaring—A method of slowing an uncontrolled ascent by arching the back, extending the arms and legs, and positioning the fins so they are parallel to the surface.

flood—The flow of water into an area because of high tide.

flutter kick—The most common type of fin kick. The flutter kick is an up-and-down kick that you can do facing down, up, or to the side.

fronds—The portions of kelp that appear to be the leaves.

gauge pressure—The pressure indicated by a pressure gauge that reads zero at sea level (this type of gauge displays only the pressure in excess of one atmosphere).

Gay-Lussac's law—This law states that for any gas at a constant volume, the pressure of the gas varies directly with the absolute temperature.

gyres—Large circulating currents that move clockwise in the northern hemisphere and counterclockwise in the southern hemisphere.

half-time—The amount of time it takes a tissue to accumulate half of the gas it can hold at a given pressure.

heat exhaustion—A medical condition that occurs when the core temperature of the body is above normal and the victim has become dehydrated.

heatstroke—An emergency medical condition that occurs when the body temperature becomes so high that the body's temperature-regulating ability shuts down.

Henry's law—This law states that the amount of a gas that dissolves in a liquid at a given temperature is directly proportional to the partial pressure of that gas.

hyperthermia—Higher-than-normal body core temperature.

hyperventilation—Rapid, deep breathing in excess of the body's needs.

hypothermia—Excessive loss of heat from the core of the body.

hypoventilation—Rapid and shallow breathing that does not allow you to expel carbon dioxide from your lungs.

ingassing—The process of gas diffusion into a liquid.

jumpsuit—A one-piece dive suit.

J-valve—A cylinder valve that was designed to maintain a reserve of air to permit a normal ascent. The introduction of submersible pressure gauges for scuba tanks has rendered the J-valve obsolete.

kelp—Giant algae that produce long strands that a diver can become entangled in.

Kelvin—The absolute temperature scale for Celsius temperatures. To convert a Celsius temperature to Kelvin, add 273 degrees.

K-valve—A cylinder valve that is a simple on-off valve that operates like a faucet. You turn the valve handle counterclockwise to open it and clockwise to close it.

lubber line—A reference line on a compass that indicates the direction of travel.

mediastinal emphysema—A lung injury in which air is present in the tissues in the middle of the chest.

middle ear—The air space behind the eardrum.

modified frog kick—A fin kick in which the diver rotates the ankles so that the tips of the fins point outward, slides the fins tip first in an outward direction, then pulls the bottoms of the fins together quickly in a wide, sweeping arc.

multilevel dive profile—A dive that progresses from deep to shallow during a given period of time.

neap tide—The twice-monthly period of lowest tides, which occurs when the moon and sun are at right angles to each other in relation to the earth.

neutral buoyancy—A state of buoyancy that a diver achieves when the diver neither sinks nor floats when holding an average breath.

nitrogen narcosis—A detrimental effect caused by the increased pressure of nitrogen that occurs at a depth of about 100 feet (30 meters).

nitrox—A nitrogen and oxygen mixture with a higher percentage of oxygen than is found in air. The mixture reduces the effects of nitrogen at depth.

no decompression stop limits—The maximum time that a diver may stay at a specified depth.

nonreference descent—A descent made vertically in water without any external reference.

octopus—A term used to describe an extra second stage.

open-circuit—Refers to a scuba system that permits a diver to breathe compressed air from a pressurized cylinder and exhausts the used gases into the water. Open-circuit systems are the safest and most popular form of scuba system.

open-valve ascent—A method for maintaining neutral buoyancy during an ascent by keeping the BC inflator-deflator valve open while holding the valve in a special way.

O-ring—A soft, circular ring that surrounds the outlet for scuba tank valves.

outgassing—The process of gas diffusing out of a liquid. For diving, it pertains to the diffusing of nitrogen absorbed under pressure.

overturn—The movement of water from the top of a lake to a depth of about 60 feet as a result of water circulation caused by winds. This water movement carries oxygenated water to the bottom and leads to what is called spring diving conditions.

partial pressure—The percentage of the total pressure exerted by each gas in a mixture of gases.

perfusion—The circulation in a tissue.

plankton—Aquatic life-forms that drift with the currents.

pneumothorax—A lung injury in which air is trapped in the chest cavity.

pony tank—A backup scuba unit consisting of a small scuba cylinder with a separate, standard regulator.

ports—The openings in the first stage of a regulator. One of the ports is for high-pressure air measurement with an SPG. The remaining ports are for low-pressure air.

pressure—Force (often weight) per unit area.

pulmonary barotrauma—Any lung injury caused by pressure.

Rankine—The absolute temperature scale for Fahrenheit temperatures. To convert a Fahrenheit temperature to Rankine, add 460 degrees.

reciprocal compass course—A course where the diver moves in a given heading for a given distance and returns to the origination point by reversing the direction of travel.

red tide—Blooms created by a type of red phytoplankton.

reference descent—A descent that the diver controls by following a line or the slope of the bottom.

repetitive dive—Any dive made within 6 to 24 hours (depending on the dive-planning device) of a previous dive.

repetitive group or repetitive group designation—The letter on a dive table that indicates the amount of nitrogen in a diver's body.

residual nitrogen—Nitrogen remaining in your system from a dive made within the past 12 hours.

residual nitrogen time—An amount of time that must be added to the actual bottom time (ABT) of a repetitive dive to determine the total bottom time (TBT). RNT compensates for residual nitrogen remaining in a diver's tissues from previous dives.

rest stop—A precautionary decompression stop during ascent.

reverse block—Discomfort that occurs when the pressure in the middle ear is above normal during an ascent due to blockage of the Eustachian tube.

reverse thermocline—The point of transition from a layer of cold water toward the surface of a lake to a layer of warmer water toward the bottom. This state exists in winter.

rip current—The narrow, strong current that moves away from shore as a result of water flowing back to sea through a narrow opening in an underwater obstruction.

sawtooth dive profile—A dive that progresses from deep to shallow and back to deep.

scissors kick—A fin kick that can be used as a resting kick. While lying on the side, the diver slowly extends one leg backward while extending the other leg forward, and then pulls the legs together quickly.

scombroid—A type of fish poisoning that can result if you eat fish that have not been kept chilled. It produces nausea and vomiting within an hour.

scrolling—A method used by dive computers to provide advance-planning information through a sequential display of time limits for various depths.

scuba—Self-contained underwater breathing apparatus.

semi-closed-circuit—Refers to a scuba system that combines the benefits of closed-circuit and open-circuit systems. Small amounts of used gases are released periodically.

set—The direction assumed by a current.

shallow-water blackout—The sudden loss of consciousness near the surface of the water.

shorty—A one-piece dive suit with short arms and legs used for diving in warm water.

sink—The area formed when the ground collapses into an underground cave system.

siphon—An opening where water is channeled from a sink back into the system.

skip breathing—An attempt to extend one's air supply by holding each breath for several seconds.

slack water—The period of time between flow and ebb when the water has minimal movement.

Spare Air unit—A backup scuba unit consisting of a small scuba cylinder with a special regulator integrated directly into the valve. A Spare Air unit is smaller than a pony tank and provides only enough air to permit an ascent from shallow depths.

spike dives—A dive with a short actual bottom time (ABT), such as a dive to free a fouled anchor.

spring tides—The twice-monthly period of highest tides, which occurs when the moon and sun are aligned with the earth.

square compass course—A series of headings that allows the diver to travel in a square pattern so that the dive ends where it began.

squeeze—A condition that occurs when the pressure on the outside of an air space in or on the body is greater than the pressure inside the air space. This condition may cause pain and injury and may be avoided by keeping the pressures equal (see equalization).

stages—The two parts of a scuba regulator that perform the two steps in the pressure reduction process.

stand—A brief period of time during which the tide neither rises nor falls.

step dive—A dive that is made in a series of steps.

stipes—Long strands of kelp.

subcutaneous emphysema—A lung injury in which air is present in the tissues under the skin. The air swells the tissues around the neck.

submersible pressure gauges—An instrument that measures scuba cylinder pressure.

surf—Breaking waves that occur when the water within the wave moves forward and gives up its energy. This occurs when the depth of the water is about the same as the wave height.

surface interval time—The elapsed time from when a diver surfaces after a dive to the time the diver begins the descent of a repetitive dive.

surge—The back-and-forth subsurface motion of the water within waves when the waves enter shallow water.

swells—The rounded, undulating forms that waves take as they move away from the area in which they were created.

tetrodotoxin—A type of fish poisoning that results from eating exotic fish such as puffer fish or blowfish. Tetrodotoxin is the most serious form of fish poisoning and can cause death within minutes.

thermocline—An abrupt transition from a warmer layer of water to a colder layer of water.

total bottom time—The sum of actual bottom time and residual nitrogen time.

Toynbee maneuver—A method of opening the eustachian tubes by blocking the nostrils, closing the mouth, and swallowing.

trail line—A long line with a float attached that is deployed from the stern of a boat. Divers use the trail line to pull themselves to the boat against the current.

trapdoor effect—A condition that exists when pressure prevents the Eustachian tube from opening. This condition may be overcome by ascending to reduce pressure.

trough—The bottom of a wave.

tsunami—Gigantic seismic waves caused by an underwater earthquake.

turbid—Water that is unclear or murky because of stirred-up sediment.

upwelling—The replacement of warm water at the surface with colder water flowing up from the depths as a result of a strong wind blowing along the shore for a sustained time.

Valsalva maneuver—A method of opening the eustachian tubes by blocking the nostrils, closing the mouth, and gently trying to exhale.

valve seat—The portion of the cylinder valve that closes and stops the flow of air.

vasoconstriction—Narrowing of the blood vessels.

vertigo—A subjective feeling of movement perceived as a spinning sensation.

wave height—The distance from a wave crest to the trough.

wave period—The time it takes two waves to pass a given point.

wave train—A long series of waves.

wavelength—The distance between waves.

Bibliography

Auerbach, P. (1987). *A medical guide to hazardous marine life.* Jacksonville, FL: Progressive Printing.

Barnhart, R., & Steinmetz, S. (1986). *Dictionary of science.* Maplewood, NJ: Hammond.

Bascom, W. (1964). *Waves and beaches.* Garden City, NY: Anchor Books.

Bove, A., & Davis, J. (1990). *Diving medicine.* Philadelphia: W.B. Saunders.

Divers Alert Network. (1989). *Medical requirements for scuba divers.* Durham, NC: Author.

Edmonds, C., Lowry, C., & Pennefather, J. (1981). *Diving and subaquatic medicine.* Mosman, NSW, Australia: Diving Medical Centre.

Foley, B. (1989). *Physics made simple.* New York: Doubleday.

Graver, D. (2004). *Aquatic Rescue and Safety.* Champaign, IL: Human Kinetics.

Lee, P., Lidov, M., & Tyberg, T. (1986). *The sourcebook of medical science.* New York: Torstar.

Lehrman, R. (1990). *Physics the easy way.* Hauppauge, NY: Barron's Educational Series.

Lippmann, J. (1992). *The essentials of deeper sport diving.* Locust Valley, NY: Aqua Quest.

Maloney, E. (1983). *Chapman piloting seamanship and small boat handling.* New York: Hearst Marine Books.

McGraw-Hill Professional Publishers (1998). Concise encyclopedia of science. Camden, Maine: McGraw-Hill.

Miller, J. (Ed.). (1979). *NOAA diving manual* (2nd ed.). Washington, DC: U.S. Government Printing Office, U.S. Department of Commerce.

National Association of Underwater Instructors (NAUI). (1991). *Advanced diving technology and techniques.* Montclair, CA: Author.

Professional Association of Diving Instructors (PADI). (1998). *The encyclopedia of recreational diving.* Santa Ana, CA: Author.

Sebel, P., Stoddart, D., Waldhorn, R., Waldmann, C., & Whitfield, P. (1985). *Respiration: The breath of life.* New York: Torstar.

Taylor, E. (Ed.). (1985). *Dorland's illustrated medical dictionary.* Philadelphia: W.B. Saunders.

Whitfield, P., & Stoddart, D. (1984). *Hearing, taste, and smell: Pathways of perception.* New York: Torstar.

This page intentionally left blank.

Index

Note: The italicized *f* and *t* following page numbers refer to figures and tables, respectively.

A
AAS. *See* alternate air source
abrasions 114, 116*t*
absolute pressure 23, 26-27
ABT. *See* actual bottom time
accessories 101-104, 101*f*-102*f*
accidents 180-181, 189
actual bottom time (ABT) 194, 199-202
adjustments. *See* diving adjustments
advance preparation 185-187
AGE (arterial gas embolism) 50, 57
air
  composition 16, 16*f*
  consumption 28, 30-31, 34
  environment 19-23, 19*f*, 20*t*, 21*f*-22*f*, 118
  supply, loss of 49, 174-177, 175*f*-176*f*
alternate air source (AAS) 64, 94-96, 94*f*-95*f*
alternobaric vertigo 52
altitude 200, 202-203, 202*t*
aluminum cylinders 83-85
ambient (surrounding) pressure 53
anatomy 12-15, 13*f*, 15*f*
animals, dangerous 114-117, 114*f*-115*f*, 116*t*, 180
aquatic biology 110-117, 111*f*-115*f*, 116*t*
aquatic conditions 121-132, 123*f*-125*f*, 127*f*, 129*f*-131*f*
Archimedes' principle 21
area orientations 189-190
arterial gas embolism (AGE) 50, 57
ascents
  emergency 175-176, 175*f*
  scuba diving 159-160, 166, 174-176, 174*f*-175*f*, 200-201
  skin diving 146-147, 147*f*
atmospheric pressure 22-23, 22*f*

B
backpack 79, 149, 149*f*
back-roll entry 155, 155*f*
backup scuba units 94-95, 95*f*
barotrauma 50
BC. *See* buoyancy compensator
BEA. *See* buoyant emergency ascent
the bends. *See* decompression illness
bezel 97
biology, aquatic 110-117, 111*f*-115*f*, 116*t*
bites, animal 115, 116*t*
blooms 110
blowout plug 99
boat diving 132-135, 133*f*-134*f*, 153-154
boots 73, 73*f*, 76, 77*f*, 78
bottom dwellers 113
bottom time (BT) 200
bounce dive 198, 198*f*
Bourdon tube 96-97, 96*t*
Boyle's law 23-24, 24*f*, 27, 96
breathing 15-18, 16*f*-17*f*, 19*f*, 28, 47
  adjustments 46-51, 48*f*
  buddy 175-177, 175*f*-176*f*
BT (bottom time) 200
buddy breathing 175-177, 175*f*-176*f*
buddy check 152
buddy line 162
buddy system 152, 162-163, 172, 173*f*, 175-177, 176*f*-177*f*
buoyancy 21-22, 21*f*, 59, 60*f*
  adjustments 41-43, 42*t*
  control 119, 159-160, 173-174
  equipment and 66, 70, 73, 75-76, 78-83, 80*f*-81*f*, 80*t*, 85-86, 86*t*
  neutral 21, 142, 158-159, 161
  testing 142, 158
buoyancy compensator (BC)
  equipment 64, 80-83, 80*f*-81*f*, 80*t*, 138-139, 150-151, 153, 157-160, 174
  inflating and deflating 152-153, 157-160, 174
buoyant emergency ascent (BEA) 175-176
burst disk 88

C
calculators, dive 204-206, 204*t*
California 216
capillary gauge 96-97, 96*t*
cardiorespiratory system 16, 17*f*
CARE. *See* cognizance, assessment, rescue, and evacuation
career opportunities 218
Caribbean 217
carotid sinuses 18, 18*f*
cavern 124, 125*f*
caves 124, 125*f*
C-card 4
CCR. *See* closed-circuit rebreather
ceiling 99
certification 4
charter boat diving 133-134, 133*f*-134*f*, 190
ciguatera 112
circulation 15-18, 17*f*-18*f*
closed-circuit rebreather (CCR) 3, 4*t*, 105-106, 105*f*
cognizance, assessment, rescue, and evacuation (CARE) 178-179
cold 32, 51, 201
collapsed lung 51
communication 178
community, diving 3-5
compartments 54-55
compass course 164-165
compasses, dive 97-98, 98*f*, 164-165, 164*f*-165*f*
compass heading 164
computers, dive 99, 204-206, 204*t*
congested sinuses 44, 44*f*
conservation 117-121, 119*f*
consoles 98*f*, 99-100
continental shelf 126
contingency plans 203-204, 203*f*
continuing education 210-212, 219
controlling compartment 55
coughing 172
cramps 172, 173*f*
crest 128
currents 124, 124*f*, 130-132, 130*f*-131*f*, 187, 188*f*
cuts 114, 116*t*
cylinders 83-87, 83*f*-85*f*, 86*t*, 149-150, 149*f*-150*f*
cylinder valves 87-89, 88*f*

D
Dalton's law 25-26, 26*t*
dangerous aquatic animals 114-117, 114*f*-115*f*, 116*t*, 180
decompression illness (DCI) 7, 55-57, 56*f*-57*f*, 190, 200
decompression stops 55, 57*f*, 190, 196, 196*t*, 200-201
decompression theory 53-55, 54*f*
defogging 29, 66, 139
dehydration 29, 58-59
density 19*f*, 20, 20*t*, 23-24, 24*f*
depth 200
depth gauges 96-97, 96*t*, 100
descents 146-147, 161
destinations 216-218
diaphragm gauge 96-97, 96*t*
DIN valve 88, 88*f*

dive calculators 204-206, 204*t*
dive compasses 97-98, 98*f*, 164-165, 164*f*-165*f*
dive computers 99, 204-206, 204*t*
dive courses 8, 8*f*
dive flags 101-102, 101*f*, 160, 160*f*
dive floats 102
dive kits 102-104
dive knives 100-101, 100*f*, 152
dive lights 102, 102*f*
dive planning
  advance preparation 185-187
  area orientations 189-190
  calculators for 204-206, 204*t*
  computers for 99, 204-206, 204*t*
  contingency 203-204, 203*f*
  dive profile 188, 190-191, 190*f*, 197-204, 198*f*-199*f*, 202*t*, 203*f*
  dive tables for 191-197, 192*t*-193*t*, 195*t*-196*t*, 204, 204*f*
  factors 184-185
  on-site 187-188, 188*f*
  postdive review 206
  short-term 187
  special-equipment 197
  special procedures 201-203, 202*t*
dive profile 188, 190-191, 190*f*, 197-206, 198*f*-199*f*, 202*t*, 203*f*
diver impact 118-119, 119*f*
diver's push-ups 158, 159*f*
dive slates 102-103
dive tables 191-197, 192*t*-193*t*, 195*t*-196*t*, 204, 204*f*
diving. *See also* equipment; scuba diving; skin diving
  altitude and 200, 202-203, 202*t*
  boat 132-135, 133*f*-134*f*, 153-154
  buddy 162-163, 172, 173*f*, 175-177, 176*f*-177*f*
  community 3-5
  conditions 121-132, 123*f*-125*f*, 127*f*, 129*f*-131*f*
  drift 132
  evolution of 3
  guides 4
  joys of 2, 9
  periodicals 222
  prerequisites 5-6
  risks 6-8
  surface 145, 146*f*
  training for 5-8, 8*f*, 221
  travel 213-218
diving adjustments
  breathing 46-51, 48*f*
  buoyancy 41-43, 42*t*
  dehydration prevention 58-59
  equilibrium 52-53
  heart and lungs 50-52
  importance of 38, 61
  ingassing and outgassing 53-58, 54*f*, 56*f*-57*f*
  mobility 59, 60*f*
  thermal 38-41, 39*f*, 41*f*
  visual 53
diving environment
  aquatic biology 110-117, 111*f*-115*f*, 116*t*
  aquatic conditions 121-132, 123*f*-125*f*, 127*f*, 129*f*-131*f*
  for boat diving 132-135, 133*f*-134*f*
  conservation and preservation 117-121, 119*f*
diving opportunities
  career 218
  continuing education 210-212, 219
  leadership 219
  local 212-213
diving skills
  automatic execution of 139, 191
  for emergencies 178-181, 179*f*
  physical problem management 173-177, 174*f*-177*f*
  physiological problem management 169-172, 172*f*-173*f*
  scuba diving 152-169, 153*f*-155*f*, 159*f*-160*f*, 163*f*-165*f*, 167*f*-169*f*
  skin diving 141-148, 141*f*, 143*f*-147*f*
dizziness 170
dolphin kick 145, 145*f*
downwelling 130
drag 20-21, 20*t*, 59, 60*f*
drift 132
drift dive 132
drifters 110-111, 111*f*-112*f*
dry suits 75-77, 75*t*, 76*f*-77*f*

E

EANx. *See* enriched-air nitrox
eardrum 14, 15*f*, 52
ears 14-15, 15*f*, 44-46, 45*f*, 52
ebb 127
eddies 124
education, continuing 210-212, 219
electronic gauge 96-97, 96*t*
embolism 50
emergencies 178-181, 179*f*, 188
emergency decompression 201, 203*t*
emergency swimming ascent (ESA) 175-176, 175*f*
emphysema 50-51
enriched-air nitrox (EANx) 104-105, 104*f*
entanglement 173
entry techniques 152-155, 153*f*-155*f*
environment. *See also* diving environment
  air 19-23, 19*f*, 20*t*, 21*f*-22*f*, 118
  water 19-23, 19*f*, 20*t*, 21*f*-22*f*, 28-29, 29*f*, 32-34, 32*f*
environmental organizations 121, 221
equalization 7, 14, 24, 25*f*, 43-46, 44*f*-45*f*
equilibrium adjustments 52-53
equipment. *See also* exposure suits; fins; mask
  AAS 64, 94-96, 94*f*-95*f*
  accessories 101-104, 101*f*-102*f*
  assembling 149-151, 149*f*-150*f*
  BC 64, 80-83, 80*f*-81*f*, 80*t*, 138-139, 150-151, 153, 157-160, 174
  buoyancy and 66, 70, 73, 75-76, 78-83, 80*f*-81*f*, 80*t*, 85-86, 86*t*
  checklist 82
  cylinder valves 87-89, 88*f*
  disassembling 169
  dive knives 100-101, 100*f*
  diver impact and 119
  donning 139-141, 140*f*, 151-152
  handling 147-148, 166
  inspecting 139-140, 151-152
  instrumentation 96-100, 96*f*, 98*f*, 152
  introduction to 64, 65*f*
  nitrox 104-105, 104*f*, 197
  packing 148-149, 149*f*
  preparing 138-141, 140*f*
  removing 148
  scuba cylinders 83-87, 83*f*-85*f*, 86*t*, 149-150, 149*f*-150*f*
  scuba regulators 89-94, 89*f*-92*f*, 149-152, 150*f*, 156-157
  skin diving 70-71, 70*f*, 138-141, 140*f*, 147-148
  snorkel 67-68, 67*f*-68*f*, 138, 140, 143-144, 143*f*, 152, 155-157
  specialized 104-106, 104*f*-105*f*, 197
  surface snorkeling with 155-156
  for technical diving 105-106, 105*f*
  training 8, 8*f*
  weighting systems 78-80, 78*f*, 138-140, 147-148, 152
  weight of 59, 60*f*
ESA. *See* emergency swimming ascent
eustachian tubes 14, 15*f*, 44-46, 45*f*
exit techniques 168-169, 168*f*-169*f*
exposure suits
  dry suits 75-77, 75*t*, 76*f*-77*f*
  preparing 138-139, 152
  wet suits 71-75, 71*t*, 72*f*-74*f*, 77-78
extra second stages 94-95, 94*f*

F

Farmer Johns 72*f*, 73
feet-together entry 155
fetch 128
fins 69-70, 69*f*-70*f*, 152
  donning 138, 140, 140*f*
  removing 148
  using 144-145, 144*f*-145*f*, 147
first aid 103-104, 116*t*, 178-180
flags 101-102, 101*f*, 160, 160*f*
flaring 174, 174*f*, 176
floats 102

flood 127
Florida 217
flutter kick 144, 144*f*
flying 200, 202, 202*t*, 215
food cycle, underwater 110-111, 111*f*
freshwater diving conditions 122-126, 123*f*-125*f*
fronds 111

G
gases 16, 16*f*, 18, 19*f*, 26-27, 26*t*
gas laws 23-28, 24*f*-25*f*, 26*t*, 96
Gay-Lussac's law 24-25, 27
gear bags 101
giant-stride entry 154, 154*f*
Great Lakes 217
guides 4
gyres 130

H
habits 191
half-time 54, 99
hand coverings 73-74, 74*f*, 76, 77*f*, 78
hand signals
  scuba diving 166, 167*f*
  skin diving 141, 141*f*
Hawaii 217
head-to-toe buddy check 152
hearing 14, 15*f*, 34
heart 50-52
heat
  exhaustion 40, 41*f*
  loss 28-29, 29*f*, 39-40, 39*f*
heatstroke 40, 41*f*
Henry's law 25-26, 26*t*
high altitude diving 200, 202, 202*t*
hoods 74, 74*f*, 76, 78, 139
humidity 29
hydroids 113-114, 113*f*, 116
hyperthermia 38
hyperventilation 18, 19*f*
hypothermia 38, 40
hypoventilation 18

I
ingassing 26, 53-58, 54*f*, 56*f*-57*f*
injuries 50, 114-117, 116*t*, 179-181
instrumentation
  equipment 96-100, 96*f*, 98*f*, 152
  monitoring 165-166

J
job opportunities 218
jumpsuit 72*f*, 73
J-valve 88

K
kelp 110-111, 112*f*, 216
kicks 144-145, 144*f*-145*f*, 164
kits 102-104
knives 100-101, 100*f*, 152
K-valve 87, 88*f*

L
lacerations 114-115, 116*t*
laws 4
leadership opportunities 219
lifesaving scuba 178-179, 179*f*
lift 59, 60*f*
light 32-33, 32*f*
lights, dive 102, 102*f*
local diving opportunities 212-213
lubber line 98
lungs 12, 13*f*, 47, 48*f*, 50-52, 176

M
magazines 222
mask 46, 65-67, 65*f*-66*f*, 138-140, 147-148, 152
  clearing 142-143, 143*f*, 162
  defogging 29, 66, 139
maximum-depth indicator 97
mediastinal emphysema 50
menstruation 6
middle ear 14, 15*f*, 52
Middle East 218
mobility adjustments 59, 60*f*
modified frog kick 144, 145*f*
multilevel dive profile 197, 198*f*

N
navigation 97-98, 98*f*, 164-165, 164*f*-165*f*
NDSL. *See* no decompression stop limit
neap tides 126
neutral buoyancy 21, 142, 158-159, 161
New England 217
nitrogen narcosis 16, 58
nitrox equipment 104-105, 104*f*, 197
no decompression stop limit (NDSL) 191, 194-197, 195*t*, 199-200, 202
nonreference descents 161

O
octopus 94
on-site planning and preparation 187-188, 188*f*
open-circuit scuba 3, 4*t*
open-valve ascent 159-160, 166
organizations 121, 221
O-ring 88-89, 150-151
outgassing 26, 28, 53-58, 54*f*, 56*f*-57*f*, 166
overexertion 47-48, 172
overheating 40-41, 41*f*
overturn 122, 123*f*
overweighting 163, 163*f*

P
packing 148-149, 149*f*
panic 170-172, 172*f*
partial pressure 26-27, 26*t*
periodicals 222
physical problems, managing 173-177, 174*f*-177*f*
physiological problems, managing 169-172, 172*f*-173*f*
plankton 110, 111*f*
planning. *See* dive planning
pneumothorax 51
pollution 117-118
pony tank 95, 95*f*
ports 91
postdive review 206
preparation
  scuba diving 148-152, 149*f*-150*f*, 185-188, 188*f*
  skin diving 138-141, 140*f*
prerequisites, diving 5-6
preservation 117-121, 119*f*
pressure
  absolute 23, 26-27
  ambient 53
  atmospheric 22-23, 22*f*
  equalization 7, 14, 24, 25*f*, 43-46, 44*f*-45*f*
  gas laws and 23-24, 24*f*
  injury from 50
  measurement 22-23, 22*f*
  partial 26-27, 26*t*
  reduction, stages of 89-95, 90*f*-92*f*, 94*f*
proficiency testing 5
Puget Sound 216
pulmonary barotrauma 50
punctures 114-115, 116*t*

R
Rankine 25
reciprocal compass course 165
recompression chamber 56, 56*f*
recreational diving 3-4
Red Sea 218
red tide 110
reference descents 161
regulators, scuba 89-94, 89*f*-92*f*, 149-152, 150*f*, 156-157
repetitive dive 190-191, 190*f*, 194, 200
repetitive group designation (RGD or RG) 191, 194, 200
reserve valve 88
residual nitrogen 190, 190*f*, 194-197, 195*t*
residual nitrogen time (RNT) 194-202, 195*t*
respiration 15-18, 16*f*-17*f*, 19*f*
responsibilities 9
rest stops 55, 57*f*, 190, 196, 196*t*, 200-201
reverse blocks 24, 25*f*
reverse thermocline 122, 123*f*
RG. *See* repetitive group designation
RGD. *See* repetitive group designation
rip current 131, 131*f*

risks 6-8
RNT. *See* residual nitrogen time

S

safety stops 55, 57*f*, 190, 196, 196*t*, 200-201
saltwater diving conditions 126-132, 127*f*, 129*f*-131*f*
save-a-dive kit 102-103
sawtooth dive profile 197, 198*f*
scissors kick 144, 144*f*
scombroid 112
scrolling 204-205
scuba (self-contained underwater breathing apparatus)
  backup 94-95, 95*f*
  cylinders 83-87, 83*f*-85*f*, 86*t*, 149-150, 149*f*-150*f*
  lifesaving 178-179, 179*f*
  regulators 89-94, 89*f*-92*f*, 149-152, 150*f*, 156-157
  systems 3, 4*t*
scuba diving
  ascents 159-160, 166, 174-176, 174*f*-175*f*, 200-201
  descents 161
  entry techniques 152-155, 153*f*-155*f*
  exit techniques 168-169, 168*f*-169*f*
  hand signals 166, 167*f*
  at high altitude 200, 202, 202*t*
  laws 4
  navigation 97-98, 98*f*, 164-165, 164*f*-165*f*
  organizations 221
  preparation for 148-152, 149*f*-150*f*, 185-188, 188*f*
  prerequisites 5-6
  recreational 3-4
  responsibilities 9
  rules 156, 200
  skills 152-169, 153*f*-155*f*, 159*f*-160*f*, 163*f*-165*f*, 167*f*-169*f*
  technical 105-106, 105*f*
  underwater swimming in 163-164, 163*f*
  wreck diving 217
sculling 119, 119*f*
seasickness 52-53, 170
seated entry 154, 154*f*
self-contained underwater breathing apparatus. *See* scuba
semi-closed-circuit scuba 3, 4*t*
set, of current 132
shallow-water blackout 18
short-term planning and preparation 187
shorty 72*f*, 73
signaling devices 102-103
signals. *See* hand signals
silting 163, 163*f*
simple valve 87-88, 88*f*
sink 125
sinuses 12, 13*f*, 43-44, 44*f*
siphon 125
SIT. *See* surface interval time
skills. *See* diving skills
skin diving
  ascents 146-147, 147*f*
  equipment 70-71, 70*f*, 138-141, 140*f*, 147-148
  hand signals 141, 141*f*
  preparation 138-141, 140*f*
  prerequisites 5-6
  recreational 4
  skills 141-148, 141*f*, 143*f*-147*f*
skip breathing 47
slack water 127
slates 102-103
small-boat diving 134-135
snorkel 67-68, 67*f*-68*f*, 138, 140, 143-144, 143*f*, 152, 155-157
snorkeling 155-156
snorkel–regulator exchanges 157
sound 34
South Pacific 217
Spare Air unit 95, 95*f*
specialized equipment 104-106, 104*f*-105*f*, 197
SPGs. *See* submersible pressure gauges
spike dive 198, 198*f*
spring tides 126, 127*f*
square compass course 164-165
squatting back-roll entry 155, 155*f*
squeezes 24, 25*f*
stages, of pressure reduction 89-95, 90*f*-92*f*, 94*f*
stand 126
steel cylinders 83-85
step dive 197, 198*f*
stings, animal 115-116, 116*t*
stipes 110
stress 170-172, 172*f*
subcutaneous emphysema 51
submersible pressure gauges (SPGs) 88, 98-100, 98*f*, 150, 152, 165
surf 128-130, 129*f*, 153, 153*f*, 169, 169*f*
surface dives 145, 146*f*
surface interval time (SIT) 194-196, 195*t*, 200
surface snorkeling 155-156
surge 128
surrounding pressure. *See* ambient pressure
swells 128
swimmers 112-113, 112*f*
swimming
  ESA 175-176, 175*f*
  kicks 144-145, 144*f*-145*f*, 164
  surface dives 145, 146*f*
  underwater 146-147, 147*f*, 163-164, 163*f*

T

tanks. *See* cylinders
task-relevant instructional self-talk (TRIST) 171
TBT. *See* total bottom time
technical diving 105-106, 105*f*
teeth 15
temperature 20, 23-27, 32, 34. *See also* thermal adjustments
tetrodotoxin 112
thermal adjustments 38-41, 39*f*, 41*f*
thermocline 122, 123*f*
throat 12, 13*f*
thrust 59, 60*f*
tides 126-128, 127*f*
total bottom time (TBT) 194-202
Toynbee maneuver 44
trail line 133, 134*f*
training 5-8, 8*f*, 210-212, 221
trapdoor effect 45, 45*f*
travel 213-218
TRIST (task-relevant instructional self-talk) 171
trough 128
tsunami 126
turbid water 33

U

underwater food cycle 110-111, 111*f*
underwater swimming 146-147, 147*f*, 163-164, 163*f*
underwater timers 97
upwelling 123

V

Valsalva maneuver 44
valves, cylinder 87-89, 88*f*
valve seat 87
vasoconstriction 39
venomous injuries 114-115, 116*t*, 180
vertigo 33, 52
vests 74, 74*f*, 138-139, 142, 148
visibility 33, 162
vision 32-33, 32*f*, 67
visual adjustments 53
volume 23-24, 24*f*
vomit 52, 133, 170

W

wading entry 153, 153*f*
Washington 216
water
  environment 19-23, 19*f*, 20*t*, 21*f*-22*f*, 28-29, 29*f*, 32-34, 32*f*
  slack 127
wave height 128
wavelength 128
wave period 128
waves 128-130, 129*f*
wave train 128
weighting systems 78-80, 78*f*, 138-140, 147-148, 152
wet suits 71-75, 71*t*, 72*f*-74*f*, 77-78, 139, 148
wreck diving 217

About the Author

Dennis Graver has more than 30 years of experience as a scuba instructor and instructor trainer. He has authored 27 books and manuals on scuba diving, including three previous editions of *Scuba Diving* from Human Kinetics. In his position as director of training for the Professional Association of Diving Instructors (PADI), he designed the PADI modular scuba course and wrote the *PADI Dive Manual*, which revolutionized scuba instruction. During his tenure as director of education for the National Association of Underwater Instructors (NAUI), Graver wrote several diving texts, including the *NAUI Openwater I Scuba Diver Course Instructor Guide*. He has also contributed hundreds of articles on diving to such magazines as *Skin Diver*, *Sources*, and *Undercurrents* as well as several NAUI technical publications.

Since 1977 Graver has been photographing many wonders under the water, from the Red Sea off the coast of Egypt to the barrier reefs of Australia. He has won numerous awards from the Underwater Photographic Society, and his photos have graced the covers of numerous magazines and illustrated several diving texts and audiovisual educational programs.

Graver and his wife, Barbara, live in Camano Island, Washington.

You'll find other outstanding
outdoor and adventure resources at

www.HumanKinetics.com/outdoorandadventure

In the U.S. call 1-800-747-4457

Australia 08 8372 0999 • Canada 1-800-465-7301
Europe +44 (0) 113 255 5665 • New Zealand 0064 9 448 1207

HUMAN KINETICS
The Premier Publisher for Sports & Fitness
P.O. Box 5076 • Champaign, IL 61825-5076 USA

eBook
available at
HumanKinetics.com

THE COMPLETE HUNTER™

THE COMPLETE GUIDE TO HUNTING

Basic Techniques for Gun & Bow Hunters

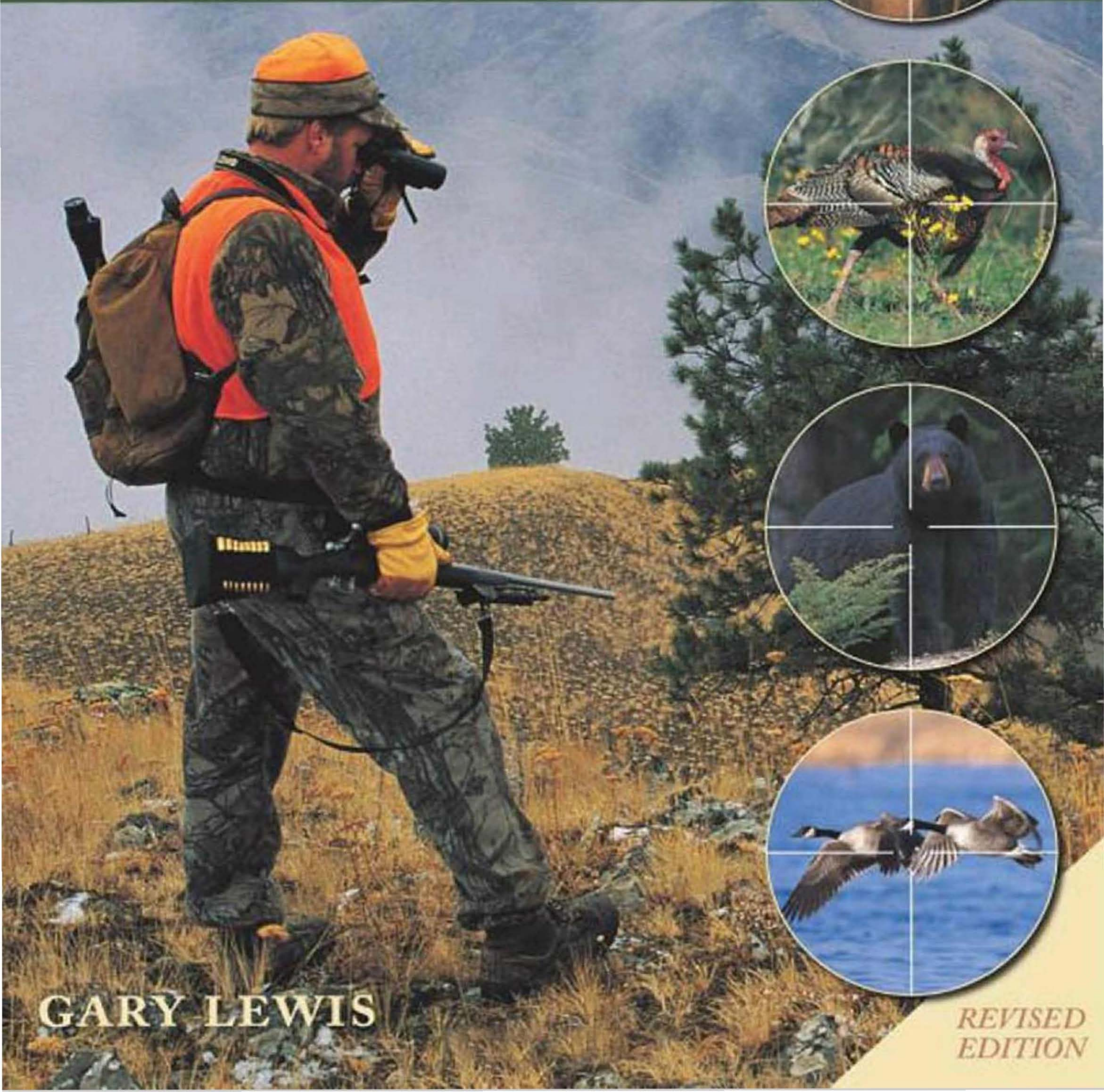

GARY LEWIS

REVISED EDITION

THE COMPLETE HUNTER™

THE COMPLETE GUIDE TO HUNTING

Basic Techniques for Gun & Bow Hunters

Gary Lewis

Creative Publishing
international

Minneapolis, Minnesota

Gary Lewis *is an outdoor writer, speaker and photographer. He has hunted and fished in five countries on two continents and much of the American West. He is a past President of the* Northwest Outdoor Writers Association (NOWA) *and is Associate Editor for* Hunting the West *magazine. Lewis is the author or co-author of seven books, including Creative Publishing International's* Black Bear Hunting *with Lee Van Tassell. Four of his books have won NOWA's Excellence in Craft awards and he has received numerous awards for his newspaper columns and magazine articles. His stories have appeared in many prestigious outdoor magazines such as* Sports Afield, Rifle, Successful Hunter, African Sporting Gazette *and* Traditional Bowhunter. *He lives in Bend, Oregon.*

Creative Publishing
international

Copyright © 2008
Creative Publishing international, Inc.
400 First Avenue North
Suite 300
Minneapolis, MN 55401
Chanhassen, Minnesota 55317
1-800-328-0590
www.creativepub.com
All rights reserved

President/CEO: Ken Fund
VP/Sales & Marketing: Kevin Hamric
Publisher: Bryan Trandem
Acquisition Editor: Barbara Harold
Production Managers: Laura Hokkanen, Linda Halls

Creative Director: Michele Lanci-Altomare
Senior Design Managers: Brad Springer, Jon Simpson
Design Managers: Sara Holle, James Kegley
Book & Cover Design: Danielle Smith
Page Layout: Danielle Smith

Contributing Photographers: Phil Aarrestad, Grady Allen, Charles Alsheimer, Erwin Bauer, Craig Blacklock, Les Blacklock, Tom Brakefield, Browning, Glenn D. Chambers, Tim Christie, Herbert Clarke, Daniel J. Cox, Gary Kramer, John Ebeling, Michael Francis, Michael Furtman, Eric J. Hansen, Larry D. Jones, Gary Kramer, Lon E. Lauber, Gary Lewis, Tom Mangelson, Bill Marchel, Tom Martinson, Jay Massey, Worth Mathewson, Bill McRae, Wyman Meinzer, Arthur Morris/Birds as Art, William H. Mullins, Doug Murphy, Scott Nielsen, Jerome B. Robinson, Lynn Rogers, Leonard Lee Rue, III, Dwight R. Schuh, Ron Shade, Jerry Smith, Dale C. Spartas, Ron Spomer, Norm Strung, Syl Strung, Ken Thommes, Charles Waterman, Chuck Wechsler, Ron Winch, James Zacks, Gary Zahm
Contributing Illustrator: Jon Q. Wright

Library of Congress Cataloging-in-Publication Data
Lewis, Gary
The complete guide to hunting : basic techniques for gun and bow hunters / Gary Lewis.
p. cm.
ISBN-13: 978-1-58923-373-7 (hard cover)
ISBN-10: 1-58923-373-5 (hard cover)
1. Hunting. I. Title.
SK33.L49 2008
799.2--dc22 2007033514

Printed in China

10 9 8 7 6 5 4 3 2 1

All rights reserved. No part of this work covered by the copyrights hereon may be reproduced or used in any form or by any means—graphic, electronic or mechanical, including photocopying, recording, taping of information on storage and retrieval systems—without the written permission of the publisher.

CONTENTS

INTRODUCTION 4

Chapter 1

ALL ABOUT WILD GAME . . . 6

Three Senses 8
Wildlife Habitat 12

Chapter 2

HUNTING SKILLS AND EQUIPMENT 20

The Hunting Rifle. 22
Rifle Ammunition. 24
Sights 26
Sighting in Your Rifle 28
Shooting a Rifle. 30
Shotguns. 32
Bowhunting 35
Muzzleloaders 40
Hunting Safety 42

Chapter 3

HUNTING STRATEGIES . . . 44

Scouting for Game 46
Still-hunting 48
Stand-hunting. 49
Stalking. 50
Driving 52
Flushing 53
Hunting Dogs. 54
Float-hunting 55
How to Find Downed Game. 56

Chapter 4

HUNTING BIG GAME. 58

White-tailed Deer 60
Mule Deer. 68
Black-tailed Deer. 71
Elk 74
Moose. 78
Pronghorn. 80
Black Bear. 82
Wild Boar. 84

Chapter 5

HUNTING SMALL GAME . . . 86

Cottontail Rabbit 88
Jackrabbit 92
Snowshoe Hare 93
Gray Squirrel. 94
Fox Squirrel. 95
Prairie Dog 99
Ground Squirrel 100
Raccoon 101
Coyote 103

Chapter 6

HUNTING UPLAND GAME BIRDS 106

Ring-necked Pheasant. 108
Bobwhite Quail. 114
California Quail 117
Gambel's Quail 118
Mountain Quail 119
Scaled Quail 120
Ruffed Grouse. 121
Sharp-tailed Grouse. 124
Sage Grouse. 125
Chukar Partridge. 126
Hungarian Partridge 127
Wild Turkey 128
Mourning Dove 135
Woodcock 137

Chapter 7

HUNTING WATERFOWL . . . 138

Puddle Ducks 142
Diving Ducks 146
Geese. 149

INDEX. 155

INTRODUCTION

Whether the hunt takes you after Rocky Mountain elk along the western skyline or into a Carolina swamp behind a pack of beagles, the chase offers a unique challenge. Rabbit, deer, elk, moose or bear, whatever the quarry, the hunter must outsmart a wild animal with a better knowledge of its environment and keener senses.

This challenge, along with the camaraderie of the hunt, accounts for the sport's tremendous popularity. And a successful hunter can enjoy a year-round supply of lean, nutritious meat.

The Complete Guide to Hunting is intended to make the reader a better hunter. Each section of the book deals with a different facet of the sport. Together, they will help you develop the skills and savvy needed to consistently bag wild game.

The first chapter, "All About Wild Game," provides an understanding of wildlife behavior and helps you recognize good habitat. By knowing how animals sense danger, react to threats and find food, water and cover, you can improve your chances of being in the right spot at the right time.

Chapter 2, "Hunting Skills and Equipment," provides all the information needed to select rifles, shotguns, bows and muzzleloaders. Learn how to choose the best ammunition for the game and how to become a proficient shooter. This chapter goes far beyond the basics, explaining complex principles of bullet and shot performance with descriptive photos and easy-to-understand charts.

The third chapter, "Hunting Strategies," teaches how to plan the hunt and how to scout a potential spot to make sure it holds game. This chapter details the most popular hunting techniques, from stalking to driving. Learn how to tailor a hunting strategy to the terrain and time of day. It also provides valuable tips for recovering game.

The final four chapters acquaint you with North America's most popular game animals. Hundreds of informative how-to photos, along with a fact-filled text, give all the information needed to locate, pursue and bag each of these species.

Astounding photography puts you in the scene. Face-off with a huge bull moose; walk a fencerow and see a ring-necked pheasant burst from cover at your feet; experience the nerve-wracking anticipation of waiting for a wild turkey to strut close enough for a shot.

The hunting techniques found in this book are those considered most effective by the nation's top hunting authorities. Hunting regulations and legal equipment differ in every state and province, so it is possible that some of the procedures described are illegal in your area. Baiting bears, for instance, is a popular method in some states, but is against the law in others. Check the local hunting laws before going afield.

In addition to the proven high-percentage hunting methods, this book also reveals dozens of little-known but effective tips that help the experts take their game. You will learn how to attract diving ducks by waving a black flag and how to imitate a scolding squirrel by clicking quarters.

This book is a unique blend of straightforward writing and captivating photography. Never before has so much how-to hunting information and so many dramatic wildlife photographs been packed into one volume. *The Complete Guide to Hunting* is sure to make your days afield more enjoyable—and more successful.

Chapter 1

ALL ABOUT WILD GAME

From the moment of birth, a game animal faces threats from predators, weather, disease and competition. Only the strongest and wariest live long enough to reproduce. In this way, nature selects the best breeding stock to ensure that the species thrives.

All game animals have the potential to produce many more offspring than the habitat can support. Reproductive rates are high among small game and upland birds. High mortality rates are nature's way of keeping game populations in check. When too many animals survive, the population explosion causes disease, stunted growth and eventually starvation.

Among most upland birds, waterfowl and small game, 60 to 80 percent of the population dies each year. Individuals over one year old comprise only 20 to 40 percent of the fall population. Because young animals are more abundant and because they lack the savvy of older animals, they make up the bulk of the hunter's bag.

Regulations are based on the concept that game can be harvested as long as breeding stock remains to produce another crop of similar size the following year. Seasons for small game, waterfowl and upland birds are usually long and bag limits generous. Liberal regulations are possible because of the high reproductive rates of these animals. Regulations for big game are stricter because the animals produce fewer young.

Resource agencies monitor game populations and set regulations to achieve the desired harvest. Scientific wildlife management has eliminated the problems of mass slaughter associated with market hunting in years past. In fact, some of the continent's most popular game species, such as white-tailed deer and wild turkey, are more numerous today than at any time in recorded history.

THREE SENSES

Game animals have an amazing ability to elude hunters. Their keen senses of sight, smell and hearing enable them to detect danger far in advance and take evasive action. Many animals also possess unique survival adaptations that far exceed the capabilities of humans.

SIGHT. Game birds have color vision that may surpass that of humans. Their retinas have more color receptors, or cones, than most other game animals. Colored oil droplets in the retina work like a camera filter, enhancing the visibility of certain hues. The cones also give them excellent visual acuity, or sharpness.

Game mammals have retinas that contain mostly light receptors, or rods. The rods enable them to see well in dim light. The vision of most game mammals is not as sharp as that of birds or humans and they have limited color vision, if any.

Most mammals have poor accommodation capabilities; they cannot change the shape of their lenses to focus on both close and distant objects. As a result, images are often fuzzy. This explains why many types of big game animals do not seem to notice a stationary hunter.

SMELL. Game mammals collect most of the information about their environment through their sense of smell. In a laboratory test, rabbits were able to detect the smell of acetone at a concentration of only 1/50 of that detectable by humans. Hunters who pursue big or small game should try to approach from downwind to keep their scent from drifting toward the animals.

Game birds have a poor sense of smell. In one study, wild turkey were given two piles of grain, one of which contained a highly repulsive scent. The birds paid no attention to the odor, feeding equally at each pile. To a bird in flight, a highly-developed sense of smell would be of little value because odors produced by predators quickly sink to the ground.

HEARING. Because game birds lack external ears and because their ears have fewer small bones than those of mammals, they are not as well equipped for detecting faint sounds. The range of pitches they can recognize is slightly less than that of humans and substantially less than that of many game mammals.

In a test of their ability to detect various pitches, mallards recognized sounds up to a frequency of 8,000 cycles per second (cps). Humans detected higher pitched

FACTORS THAT AFFECT GAME POPULATIONS

Big game populations can reach high levels despite low reproductive rates. Predators take only an occasional young animal. In areas with good habitat and little hunting pressure, herds may grow so large that food shortages result. Hunting regulations are set to control the harvest, yet prevent overpopulation.

Game populations tend to remain stable if the habitat is not disturbed and hunting pressure is regulated. In a study of a ruffed grouse population in New York state, there were 174 adult grouse at the beginning of the breeding season. These birds produced 935 fertile eggs. As the months passed, 24 adult grouse succumbed to predators and disease, 26 to hunting and 54 to severe weather. By the following spring, 70 birds remained to breed again. Of the 935 eggs, 374 were destroyed by predators. The rest hatched, but 337 chicks were killed by predators and disease. Hunters claimed 38 of the young birds, 81 perished in winter storms and 105 survived to the breeding season. These first-year breeders, combined with the 70 second-year birds, resulted in a total of 175 adult grouse, nearly identical to the breeding population of the previous year.

sounds, up to 20,000 cps and raccoon heard sounds of up to 85,000 cps.

SPECIAL ADAPTATIONS. The process of evolution has provided almost every species of game animal with some unique physical feature or ability that gives it an advantage over other animals. These adaptations enable the species to survive the threats posed by predators, including humans.

Many of these adaptations are so foreign to humans that we fail to consider them when hunting. For example, a mallard can stay submerged for as long as 16 minutes. It accomplishes this seemingly impossible feat by slowing its heart rate by more than 50 percent while underwater. This substantially reduces its oxygen needs.

Many a hunter has winged a duck, marked it down, spent 5 to 10 minutes looking for it, then given up in frustration. After the hunter returns to the blind, the live duck reappears in the very spot it went down.

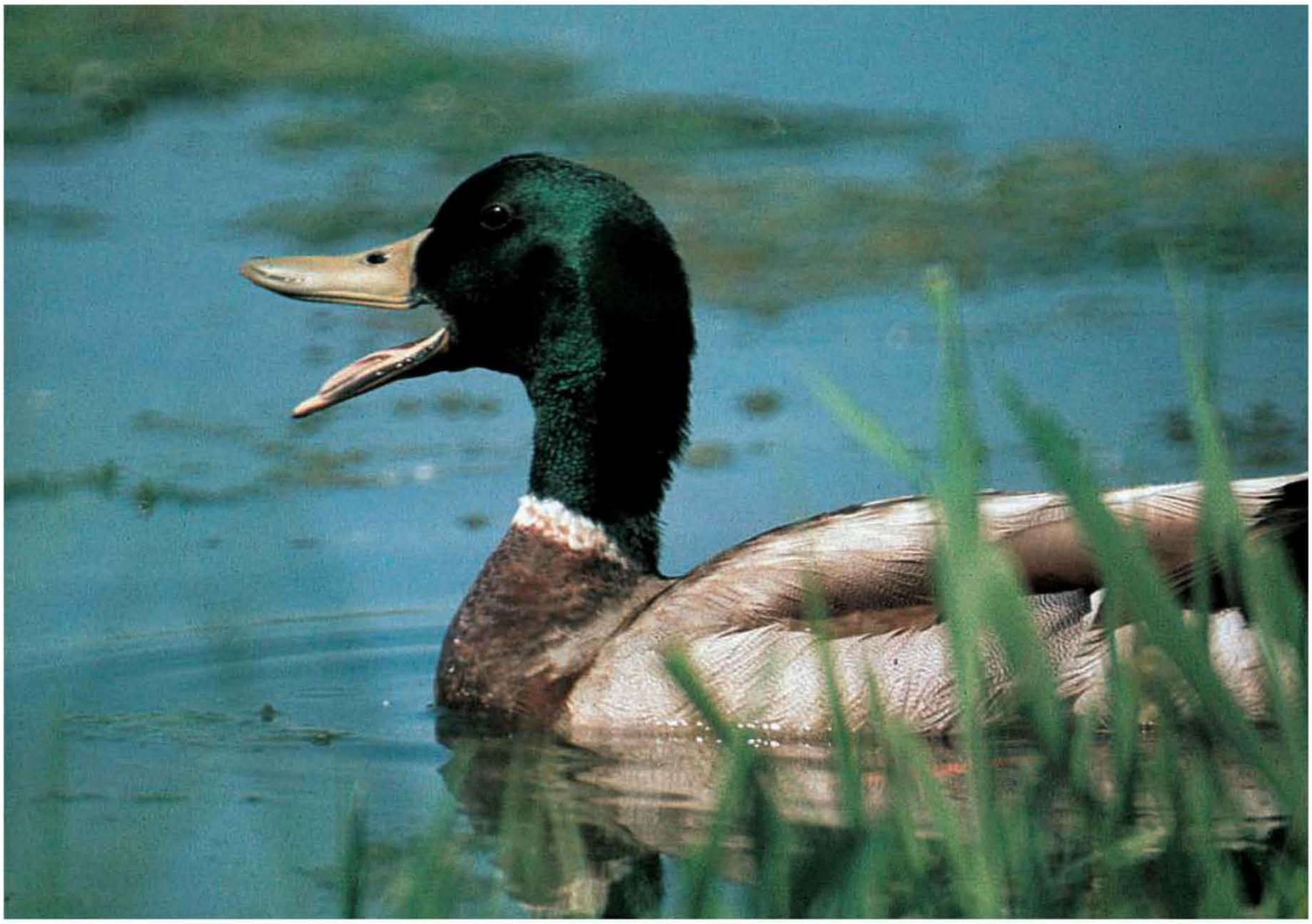

Hunters who understand the senses and special adaptations of the game they pursue stand the best chance of success.

COMPARING HUMAN AND GAME SENSES

	Vision	**Hearing**	**Smell**
Human	*Good*	*Good*	*Fair*
White-tailed Deer	*Fair*	*Excellent*	*Excellent*
Cottontail Rabbit	*Good*	*Excellent*	*Good*
Wild Turkey	*Excellent*	*Excellent*	*Poor*
Ring-necked Pheasant	*Good*	*Good*	*Poor*
Mallard duck	*Excellent*	*Good*	*Fair*

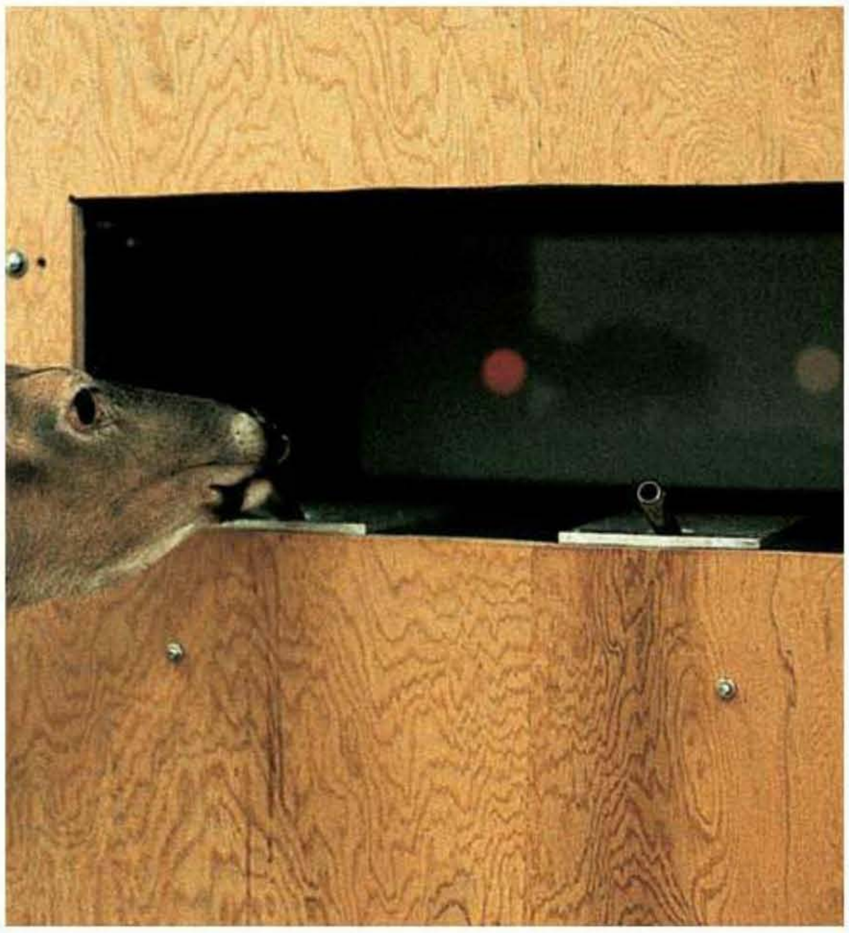

Color vision exists in some big game. In one study, white-tailed deer were rewarded with a drink from a tube when a red light went on; a white light gave no reward. Even though the lights switched between two tubes, deer consistently made the right choice.

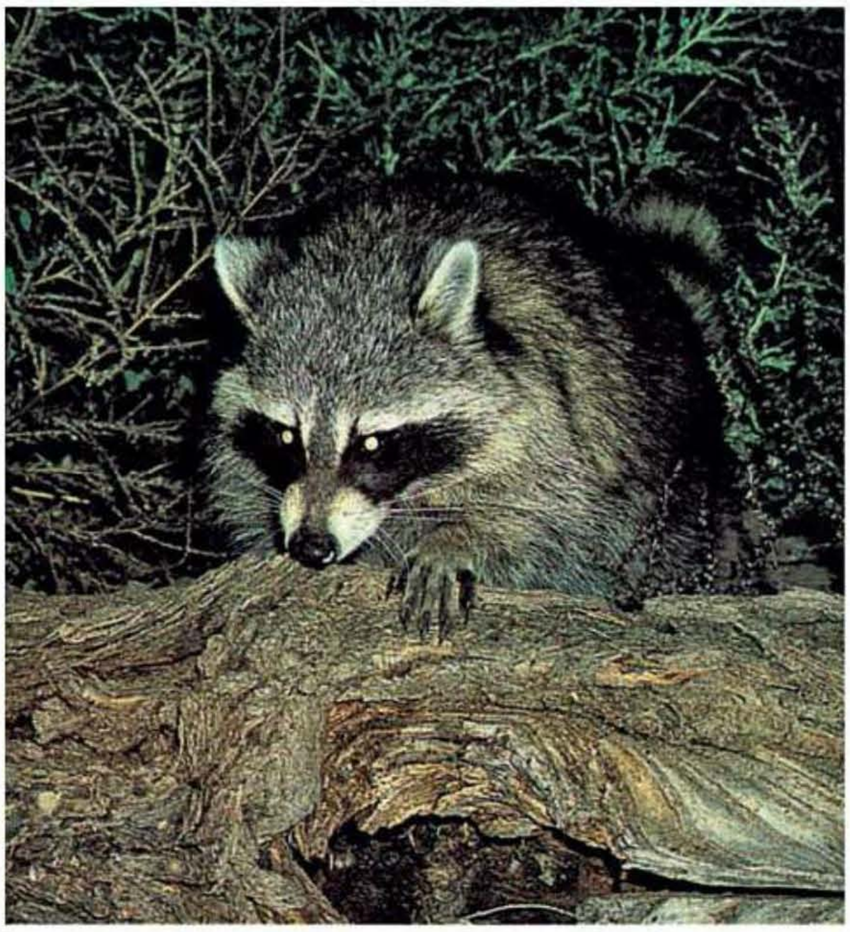

Night vision in nocturnal animals is enhanced by the tapetum lucidum, a layer of reflective pigment below the surface of the retina. Light passes through the receptors on the retina's surface, hits the tapetum, then bounces back to stimulate the receptors again.

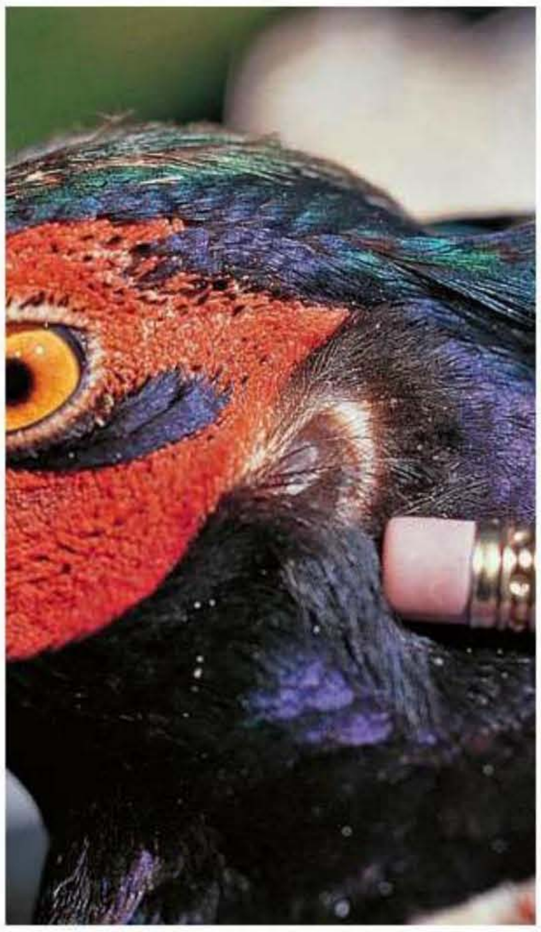

Internal ears of birds do not collect sounds as efficiently as exposed ears. And because the ears are close together, they are not as well suited to locating the direction of sounds.

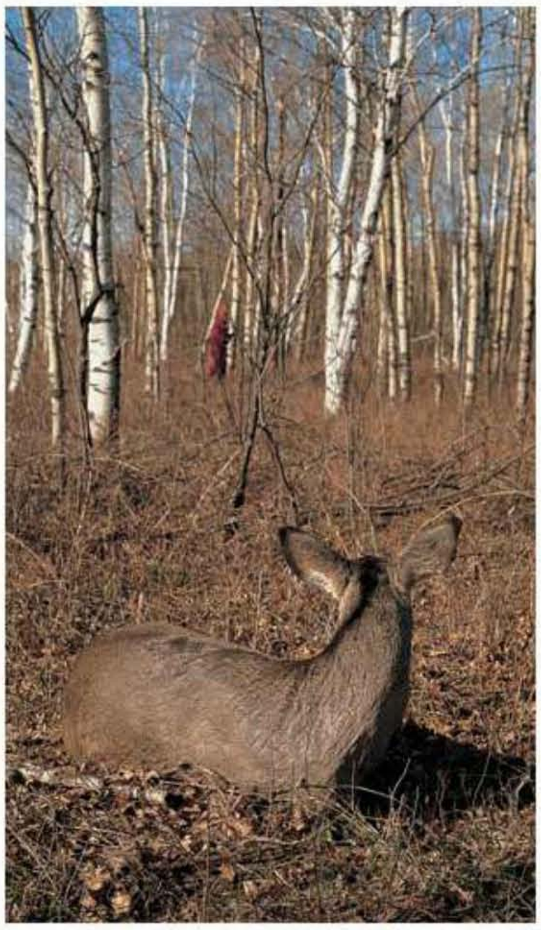

Exposed ears gather sound waves. They also help an animal pinpoint the source of a sound by the difference in the time it takes for the signal to reach each ear.

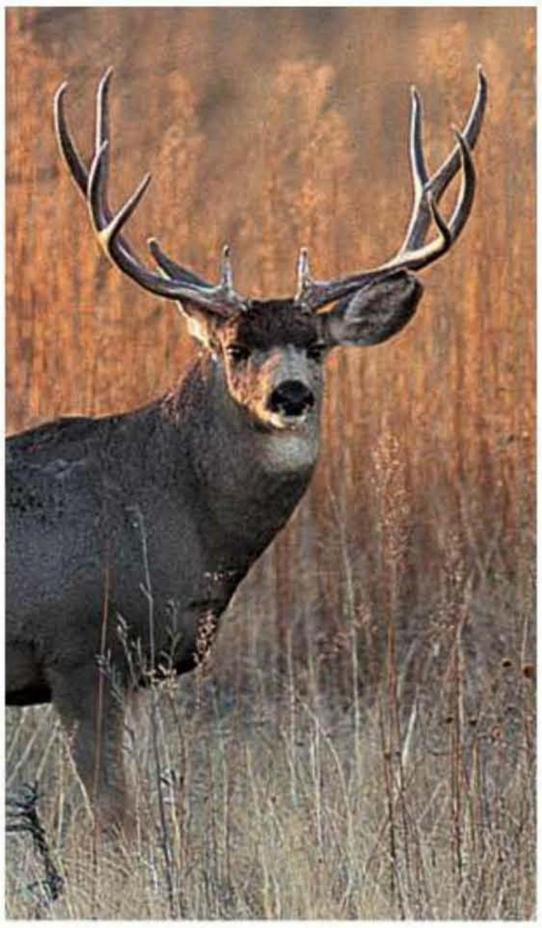

Movable ears enable mammals to locate sounds from any direction. While one ear detects a sound from one direction, the other picks up a sound from a different source.

Positioning themselves below the crest of a wind-swept ridge enables some big game to smell anything approaching from above. They watch for anything coming from below. Scent glands warn other animals of danger. When a mule deer faces a threat, metatarsal glands on the hind legs produce an odor that alerts other herd members.

Freezing in position and pressing its body to the ground (right) reduces the amount of scent an animal gives off. Some game birds also compress their feathers so less scent escapes.

OTHER SPECIAL ADAPTATIONS

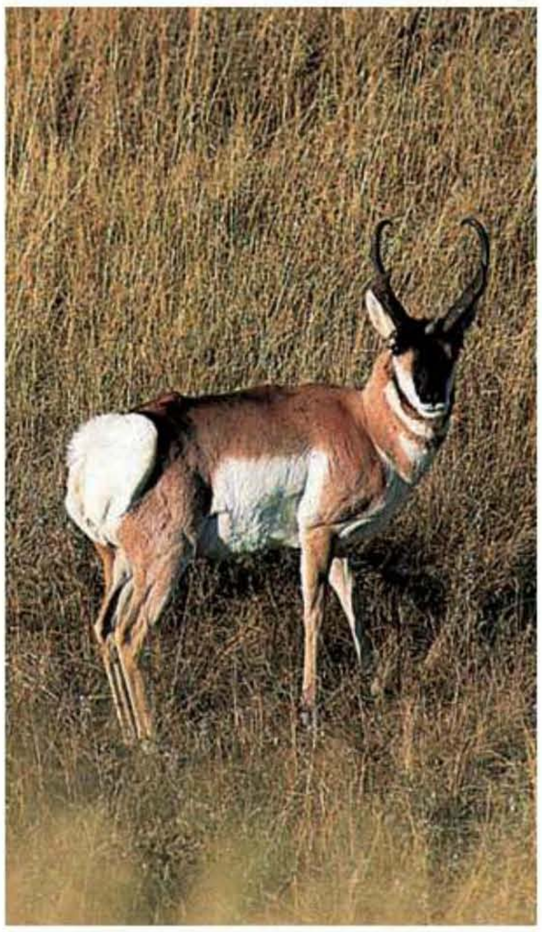

Flared rump hair alerts other animals to a threat. By flaring its rump hairs to the side, the pronghorn can greatly increase the size and visibility of its white rump patch.

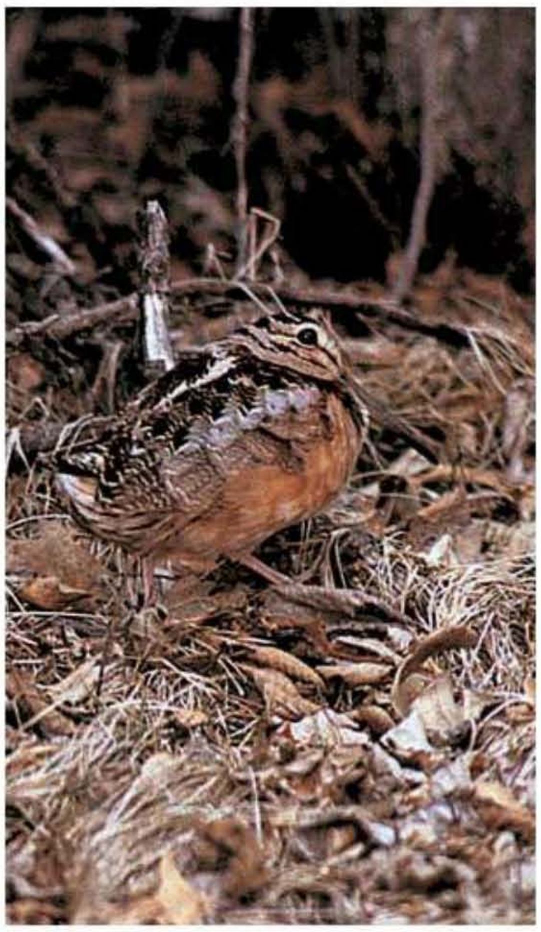

Camouflage enables game animals to escape detection by predators. The woodcock, for example, blends in so well with the vegetation that it can hide in open terrain with no overhead cover.

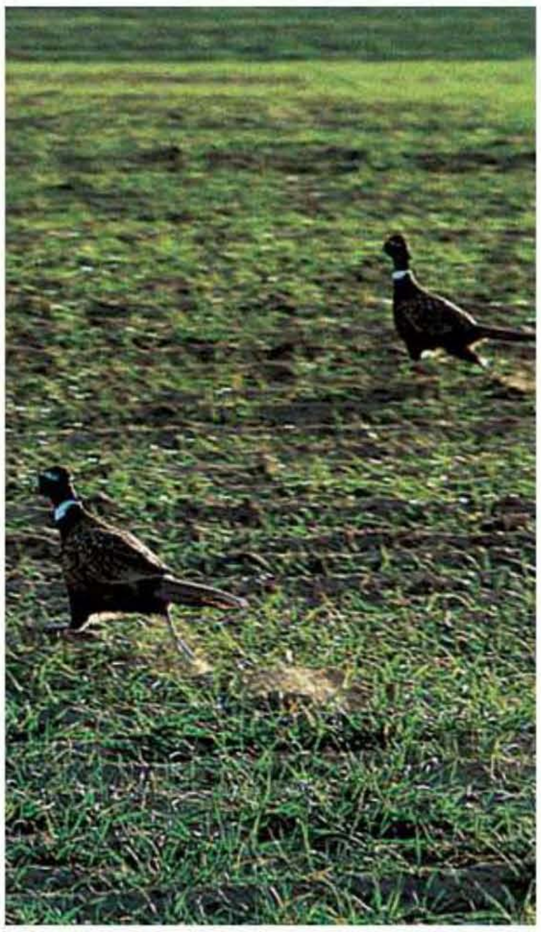

Genetic changes may result from hunting. The tendency to run has increased among pheasants. Many hunters attribute this to the fact that fliers get shot, leaving runners to breed.

WILDLIFE HABITAT

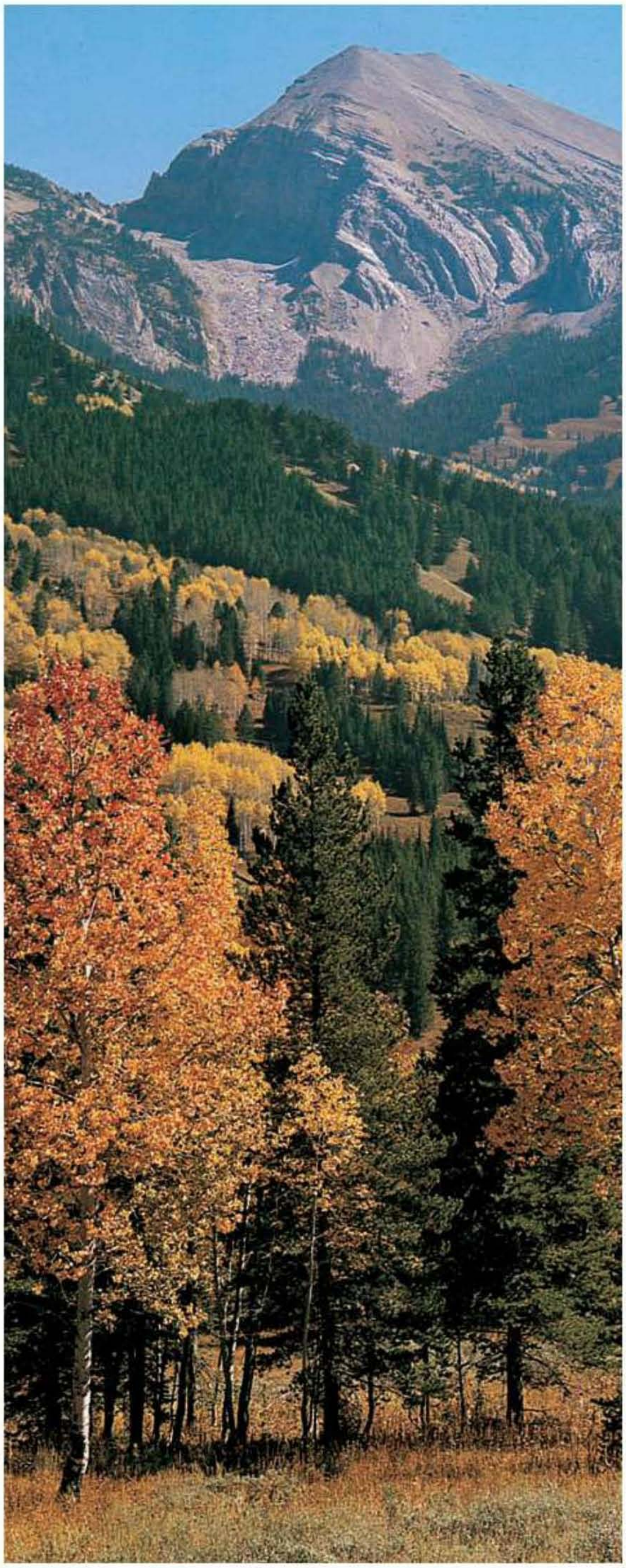

The best wildlife environment has a variety of plant life that provides food and cover. A mixed plant community generally supports more species and higher numbers of game than an expanse of the same type of vegetation.

Good habitat is the key to an abundant wildlife population. To survive and thrive, a game animal needs food, shelter cover, room to escape and water. A reliable supply of drinking water is important, although some animals find sufficient moisture in their foods.

Most game species find the necessary plant variety along the edge between two vegetative types. Where a forest meets a marsh, for instance, the mixture of grasses, berry bushes, low-growing leafy plants and young trees provide an excellent supply of food and cover. The area where the two types of vegetation meet usually holds more kinds and higher numbers of game than either type by itself. Hunters call this "edge habitat."

Another factor that influences plant variety is succession. For example, after a forest fire or logging operation, the bare ground almost immediately begins to grow grasses, shrubs and trees. As the trees grow taller, they form a canopy that shades the forest floor. As the canopy grows denser, shrubs and grasses disappear because of the lack of sunlight. Eventually, only large trees remain. These changes in the plant community invariably affect the type and amount of game the habitat can support.

Succession begins after trees have been removed. New grasses, shrubs and trees support animals like white-tailed deer. As time goes on, trees grow taller, yet enough sunlight penetrates to promote a dense growth of underbrush ideal for animals like ruffed grouse. As the trees mature and shade the forest floor, underbrush disappears and the forest becomes best suited for animals like black bear.

Wildlife managers often set back the process of succession by periodically cutting or burning forested areas to promote growth of new vegetation. These techniques enable them to increase the production of important game species.

In most agricultural areas, the trend is toward less habitat variety. Many farms once had small, weedy crop fields combined with brushy fence lines, unmowed roadside ditches, large wetlands and dense groves. Today's clean farms have vast acreages of crops unbroken by fence lines, wetlands or trees. Even roadside ditches are often mowed for hay. These intensive agricultural practices severely reduce or eliminate populations of farmland game like pheasants and quail.

Cover Requirements

Escape cover is vital to the survival of almost every game species. Wildlife also needs other kinds of cover for bedding, loafing, protection from the elements and producing young.

A ring-necked pheasant, for example, often escapes from hunters by hiding in dense brush. It may roost in open grasslands, loaf along the grassy margin of a crop field and burrow under a clump of slough grass during a storm. Experienced hunters know where to look at different times of the day and under different weather conditions.

ESCAPE COVER. Predators pose a constant threat to most game animals. Wildlife must learn to contend with these threats or perish. Most elude predators by hiding in dense cover; others scamper to a burrow or den; in light cover, some crouch motionless or rely on camouflage.

BEDDING COVER. Some game animals bed in dens or burrows or roost in trees. Others prefer thick grassy cover for bedding sites. The grass offers concealment and warns animals of danger. Even the stealthiest predator finds it difficult to move through thick grass without making some noise.

Adequate shelter cover is crucial to ensure winter survival. Bobwhite quail roost on the ground to exchange body heat. They may be found in low-growing cover on agricultural lands and around farm equipment on old homesteads.

WHERE TO FIND EDGE HABITAT

Forest clearings, like powerline corridors, logging roads and logged or burned areas, usually hold more game than the forest interior. Marshes or grassy meadows in the forests also attract wildlife.

Farmlands abutting forests or brushlands have an abundance of edge habitat. Ample sunlight reaches the forest or brushland margin, resulting in a lush growth of seed- and berry-producing plants.

Mountain slopes often have edge habitat where grassy foothills meet mid-elevation hardwoods; where hardwoods grade into high-elevation conifers; and at the treeline, where conifers end and lichens begin.

LOAFING COVER. Most game animals feed in early morning and late afternoon, then loaf during mid day. Loafing cover is thick enough so that animals are not easily visible, but not so dense that they cannot see.

COVER FROM THE ELEMENTS. All game animals need some type of dense cover to insulate them from extremely cold temperatures. They also use cover as shelter from precipitation, strong winds and the hot summer sun.

Many birds and small mammals crawl under vegetation to escape harsh winter storms. Some burrow under the snow, creating a cozy enclosure. The animal's body heat may keep the temperature as much as 50 degrees warmer than the outside air. Big game animals may bed under conifer limbs in a heavy rain or on a cold night. The boughs shed precipitation and act like a blanket to slow the loss of body heat into the atmosphere.

Abandoned buildings serve as windbreaks for many types of game. Old farmsteads often have discarded machinery, brush piles, tall weeds and evergreen groves that offer cover from the elements.

PRODUCTION COVER. Predators and adverse weather quickly wipe out nesting adults and newly produced young unless they have good production cover. This type of cover may not hold animals during the hunting season, but it often serves as an indicator of high game populations.

BASIC COVER TYPES

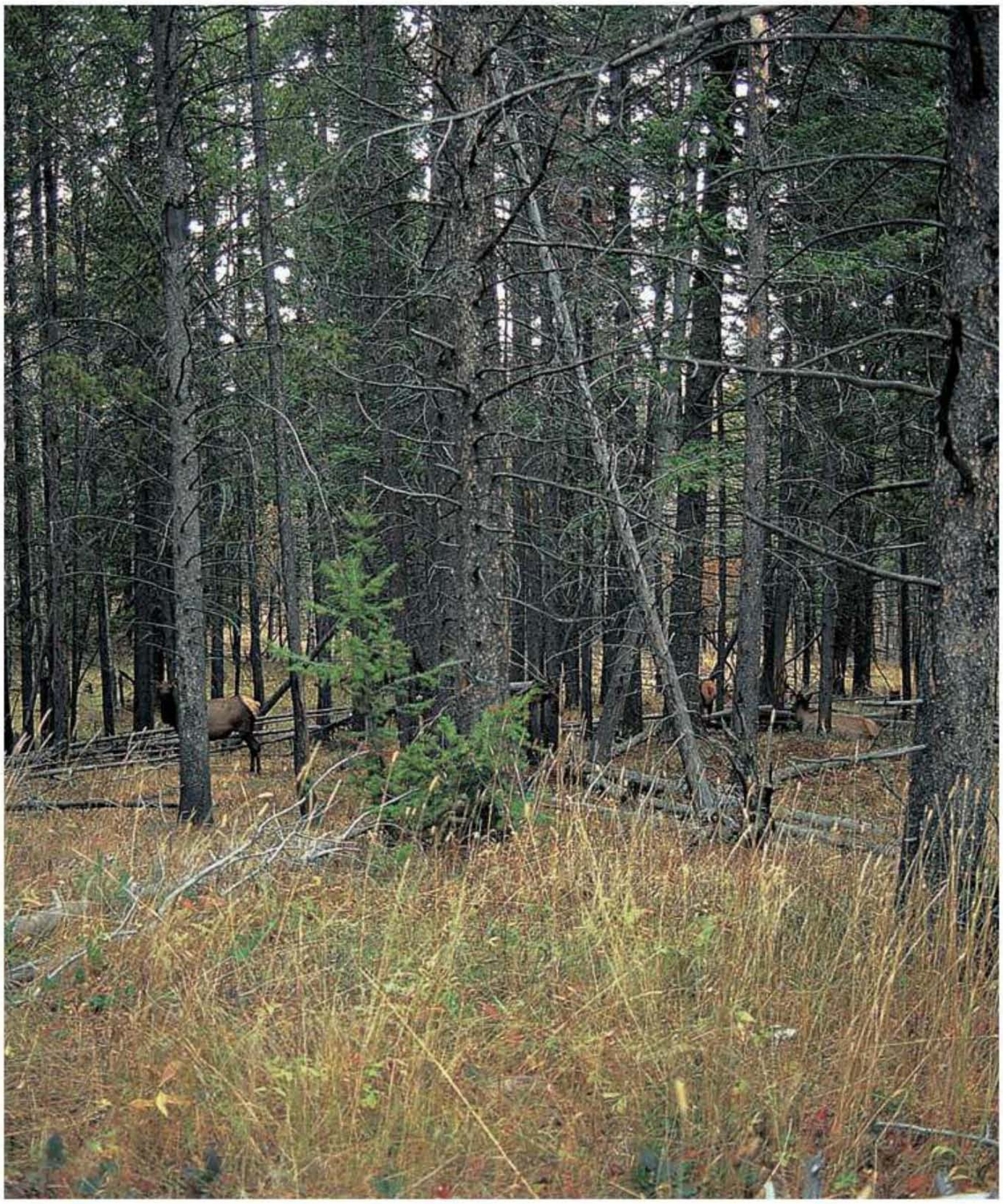

Open timber (left) is used as loafing cover by some types of big game. Small game and upland birds generally loaf in areas with denser ground cover.

Thickets make good escape cover for almost every type of wild game. Brush breaks up the animal's outline and offers shade.

Tall grass makes excellent bedding cover. It provides a comfortable resting spot and offers animals camouflage from predators.

Moderately dense vegetation provides nesting cover for many upland birds. They prefer cover open enough so they can see and escape if necessary.

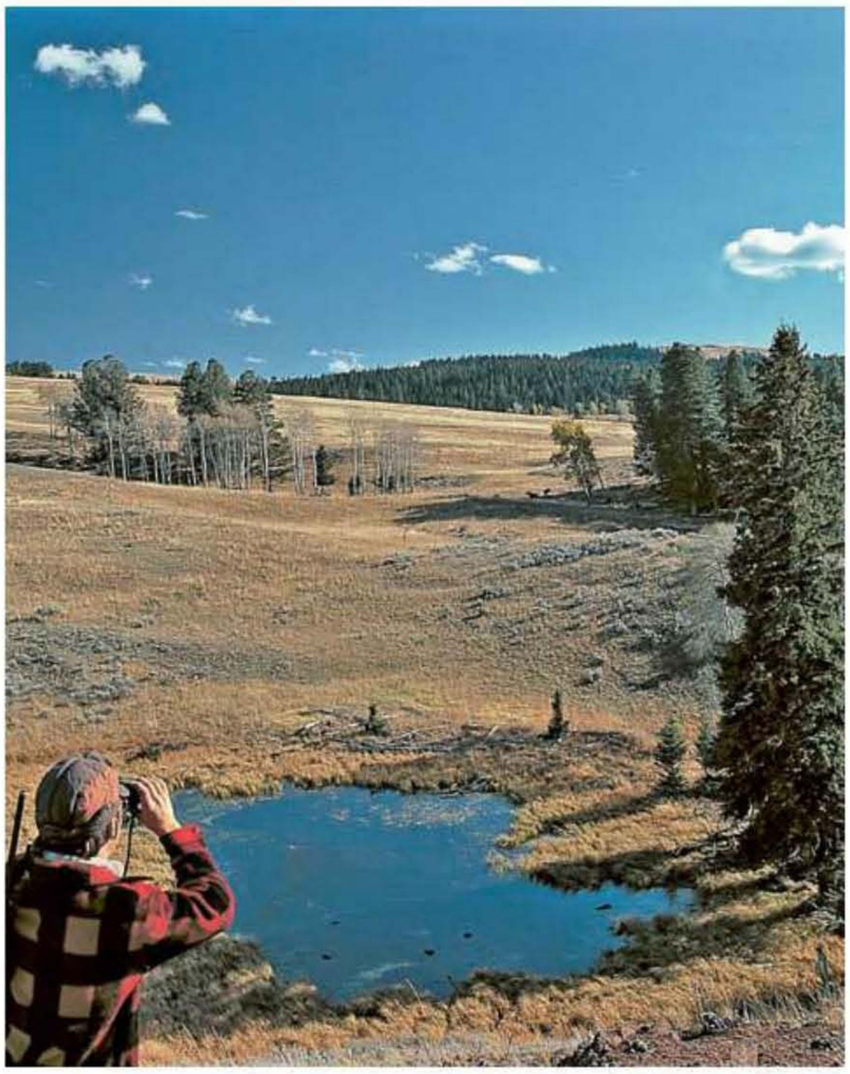

Glass for moose along forest edges, in willow swamps or near other feeding areas. Get to our vantage point early, so you can complete your stalk before the animals retreat to bedding areas.

Fertile croplands offer a good food supply. But expanses of corn and other row crops with little cover support few birds.

Look for early-season ringnecks scattered throughout crop fields, along grassy field edges and in brush patches.

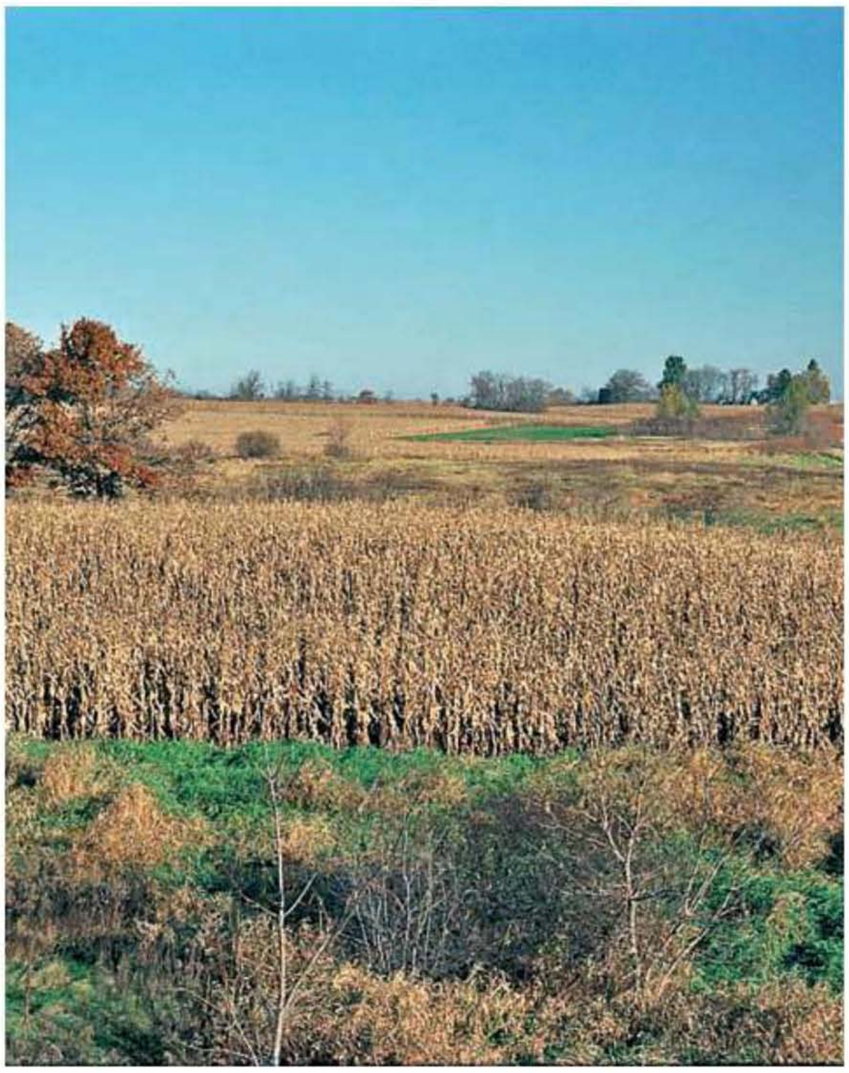

In late season game concentrates in heavy cover like brush patches and farm groves.

Food and Water Needs

Hunters who understand the feeding habits and water needs of their quarry can better predict its daily movements. This makes it easier to select a good hunting location.

An elk may travel miles (km) to find food or quench its thirst, but most wildlife needs food and water in close proximity to cover. If an animal has to travel a long distance across open terrain to fulfill these basic needs, it is more visible to predators.

FOOD. Most game animals can adapt to a wide variety of foods. Researchers have found that bobwhite quail consume over 1,000 different foods including seeds, plants, nuts and insects.

Despite their ability to use such a wide range of foods, most animals select items that they can digest easily and that are high in proteins and calories. They seem to instinctively know which foods are most nutritious. In one study, squirrels gained weight when fed a diet consisting solely of white oak acorns. But they lost weight when fed acorns of red oak. In nature, squirrels commonly feed on acorns of white oak, but ignore those of red oak.

Like humans, game animals must have vitamins and minerals in their diet. Most animals obtain these vital elements from their food and water, but some need additional salt. They frequent springs and natural soil deposits with high salt contents. Many birds pick up grit not only to grind their food, but to provide needed minerals.

WATER. Birds generally require less water than mammals, because they can reuse water as it passes through their digestive tract. Gambel's quail can go a month or two without visiting a water hole. Pronghorns, on the other hand, need water each day.

Many game species can survive in areas with no lakes, streams or other sources of surface water. They obtain water from foods or dew.

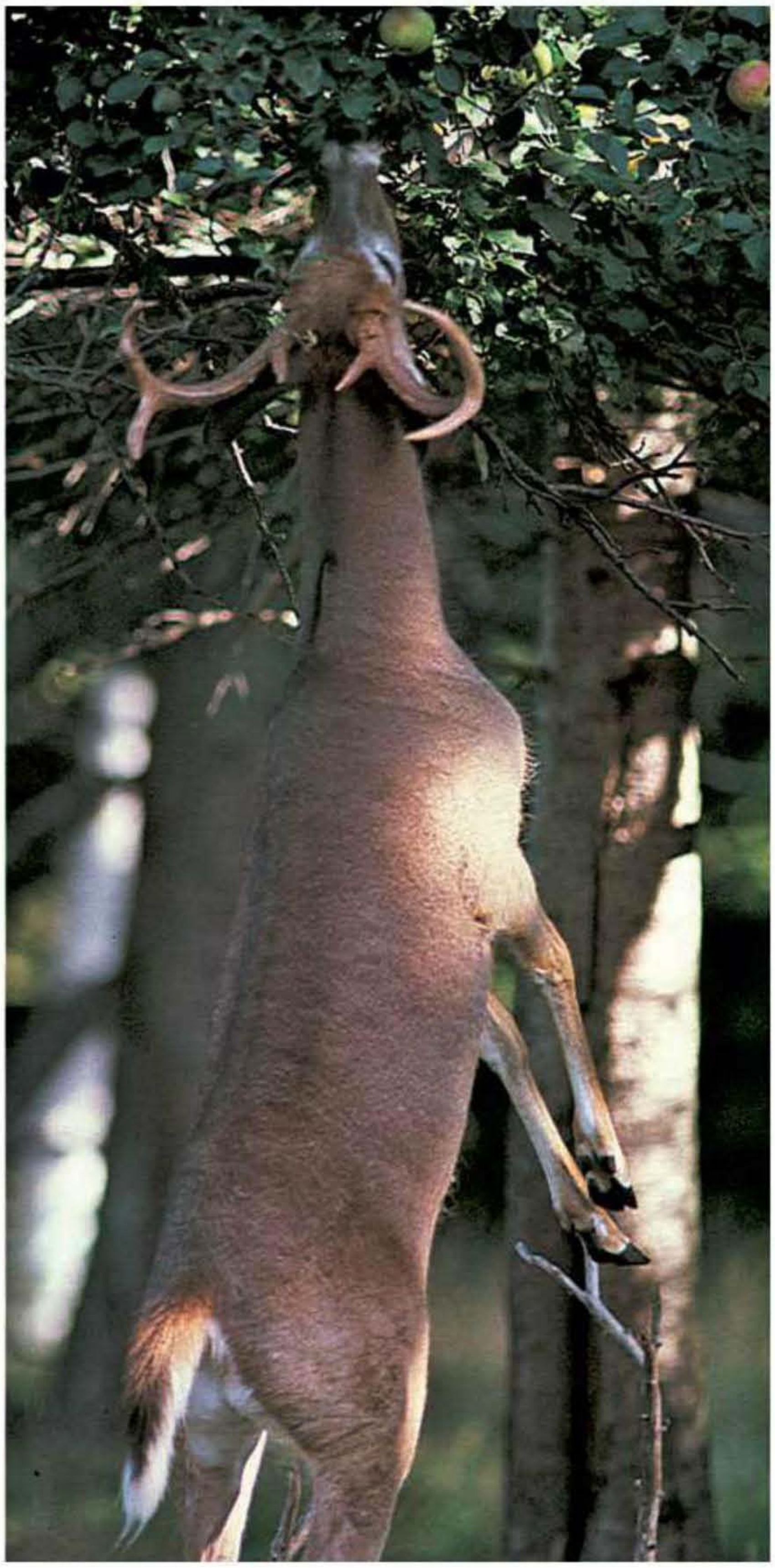

A deer's diet varies throughout the year. Deer feed on twiggy browse, forbs (broad-leafed plants), grass, fruit, nuts, and moss. On scouting trips, identify and anticipate the food sources available during the season.

FOOD SOURCES

Natural foods are the mainstay in the diets of most game animals. A good supply of winter food is especially critical. Survival rates are generally low in years when winter food is scarce.

Food plots planted by wildlife agencies provide an additional source of food during winter, when natural food may be difficult to find. Some farmers leave a few rows of corn to sustain wildlife on their land.

Waste grain remains after farmers harvest their crops. It provides a temporary food supply for many animals like geese, ducks, pheasants and deer. They gather in large numbers to take advantage of this easy source of food. Waste grain may provide food throughout the winter where fields are left unplowed.

Common Types of Habitat

Some game animals can adapt to almost any kind of habitat. White-tailed deer, for instance, live in places as diverse as Florida swamps, southwestern deserts and river valleys of the Pacific Northwest. But most types of wildlife have more specific habitat requirements.

When hunting in unfamiliar territory, it pays to spend a good deal of time scouting. Conditions may not be right for spotting game, but you can always recognize good habitat.

Many types of wildlife have cover, food or water needs so specific that merely finding a broad habitat type is not enough. For example, moose prefer young conifer-hardwood forests. But within these forests, they seldom stray far from stream banks, lakeshores or willow bogs, their main feeding sites. Sage grouse live in semi-arid brushlands, but rarely venture more than 1.2 mile (1.93 km) from a water hole or other source of water.

As the seasons change, many species of game migrate to different types of habitat. In late fall, waterfowl leave northern marshes en masse. Most winter in the South, resting and feeding on large lakes and in crop fields. As winter approaches, some types of mountain wildlife migrate to lower elevations, wintering in grasslands that offer more food and a less severe climate.

Hunting pressure may also force animals from their preferred habitat. Elk favor grassy mountain meadows, but when hunters begin to invade their territory, they retreat to higher, more rugged terrain where a person on foot could not follow. Waterfowl abandon many small lakes and ponds once the season opens. They seek refuge in the open water of large lakes and reservoirs.

In agricultural areas, wildlife may have to use other cover once crops are harvested. Pheasant hunters usually find the birds in crop fields during early season. But when the crops are removed, the birds move to permanent heavy cover like cattail swamps and woodlots.

Throughout this book, I identify the habitats most commonly associated with each species of game. Following are the major habitat types to which I refer.

WATER SOURCES

Surface water is a requirement for many species. Popular watering sites (left) include streams, lakes, human-made ponds and natural springs.

Succulent foods, like prickly pear cactus, provide water in arid climates. Rabbits and grouse often get water from grasses and buds.

Frost or dew also provides water in dry climates. To use this water source, animals must feed in early morning before the plants dry off.

HABITAT TYPES

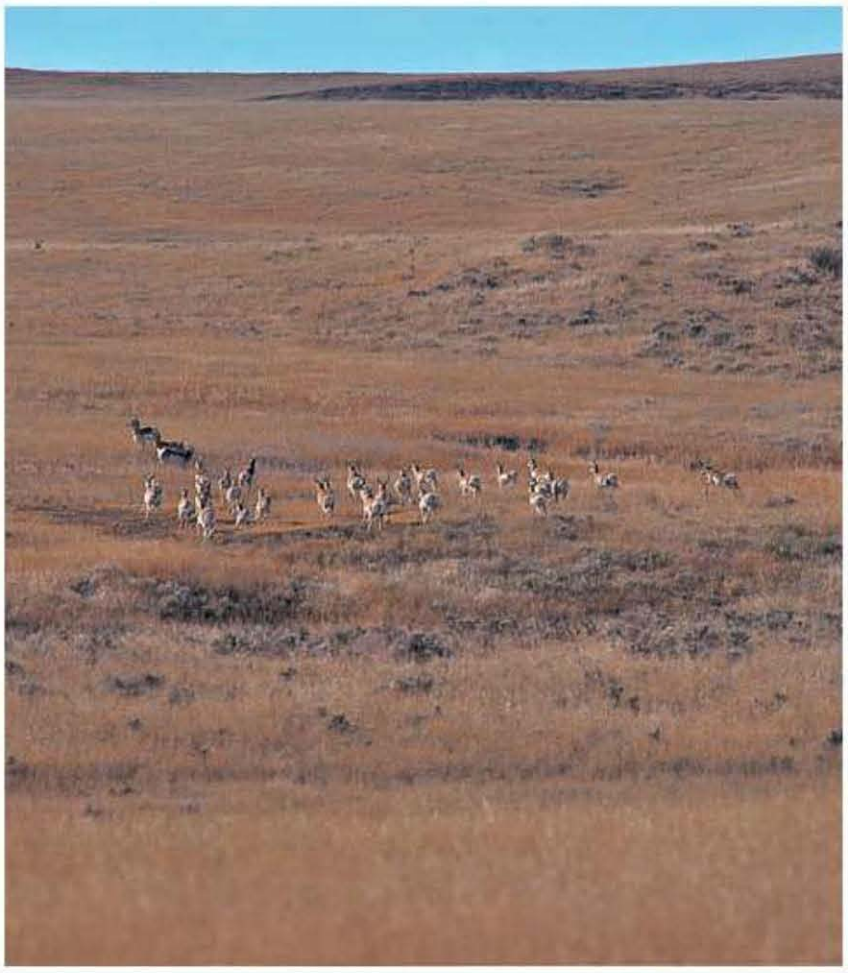

Grasslands usually have grasses growing in combination with broad-leaved green plants or forbs. The height of the vegetation varies from less than 18 inches (46 cm) in arid habitats to 10 feet (3 m) where the soil is moist.

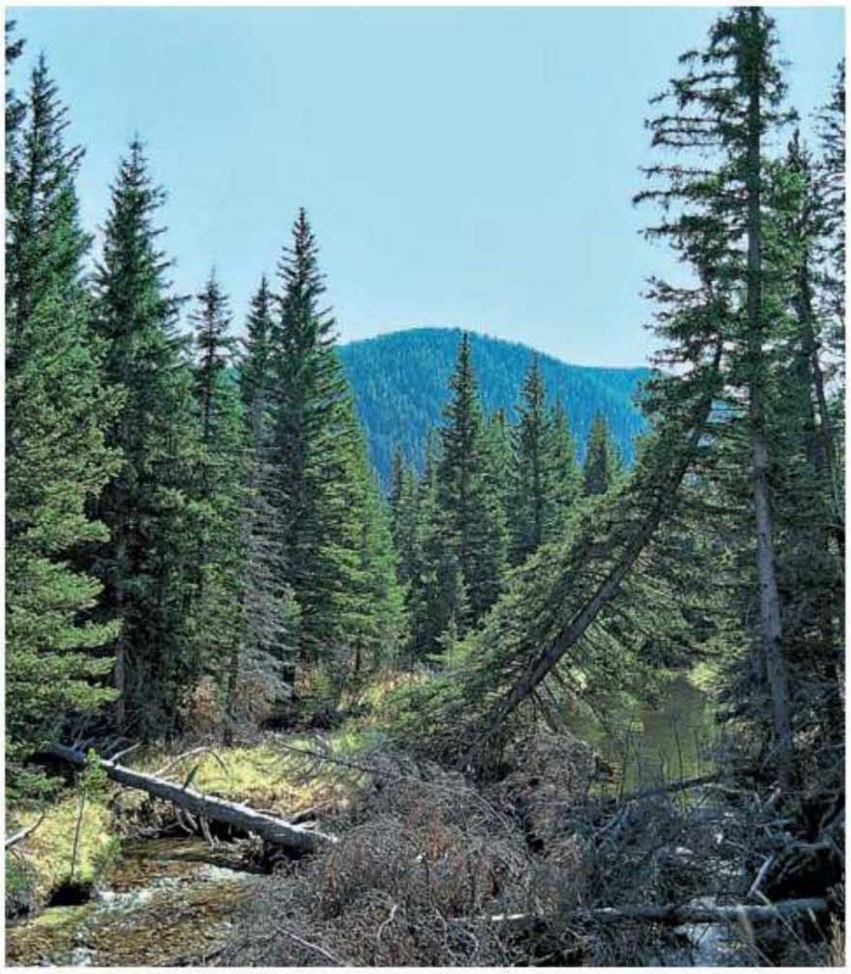

Coniferous forests have dense stands of needle- or scale-leaved evergreen trees that provide some cover. Needles that fall to the ground decompose slowly, forming acidic soil which produces few wildlife foods.

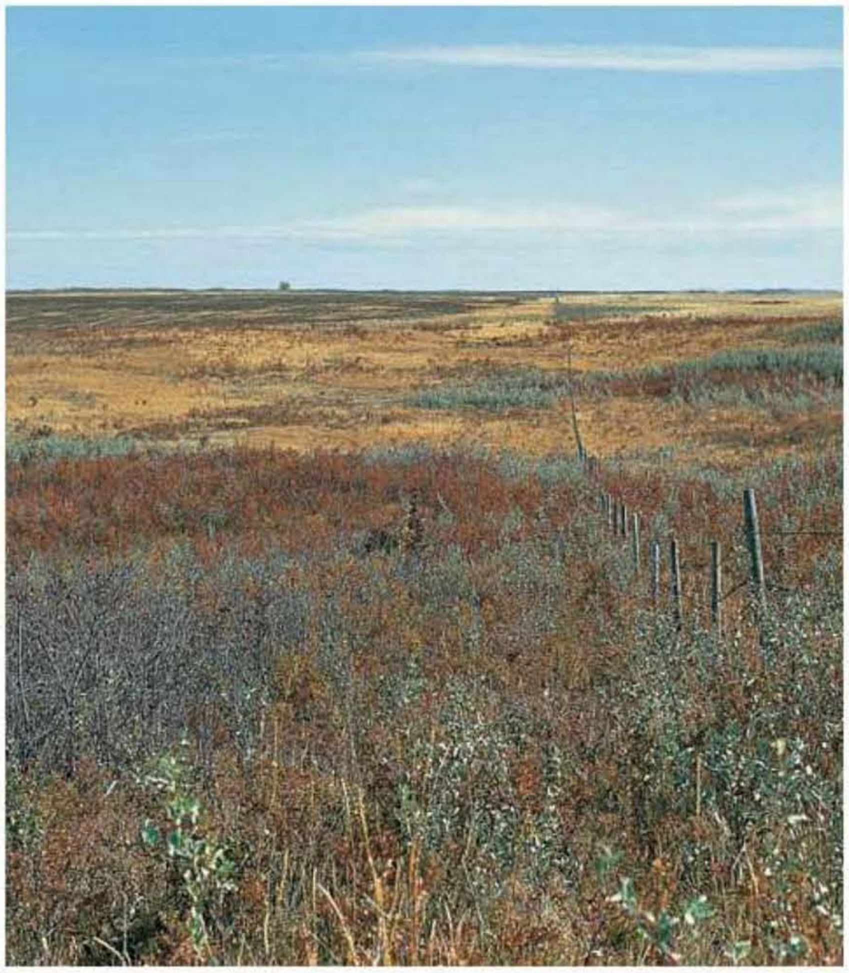

Brushlands have shrubs mixed with grasses and forbs; some have widely-spaced trees. The vegetation may be so thick that it is impenetrable. Brushlands are often an intermediate stage between grassland and forest.

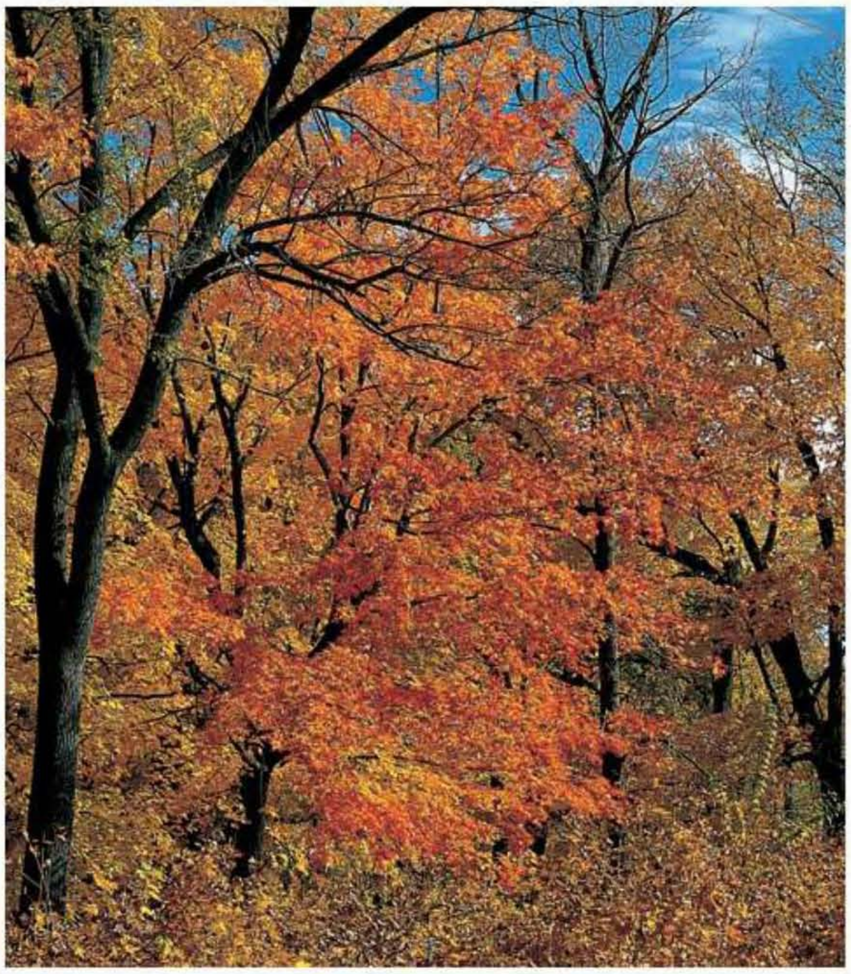

Deciduous forests, also called hardwood forests, contain trees that lose their leaves in the fall. The leaves rot quickly, forming rich humus. Buds, nuts, shrubs and green plants provide food for wildlife.

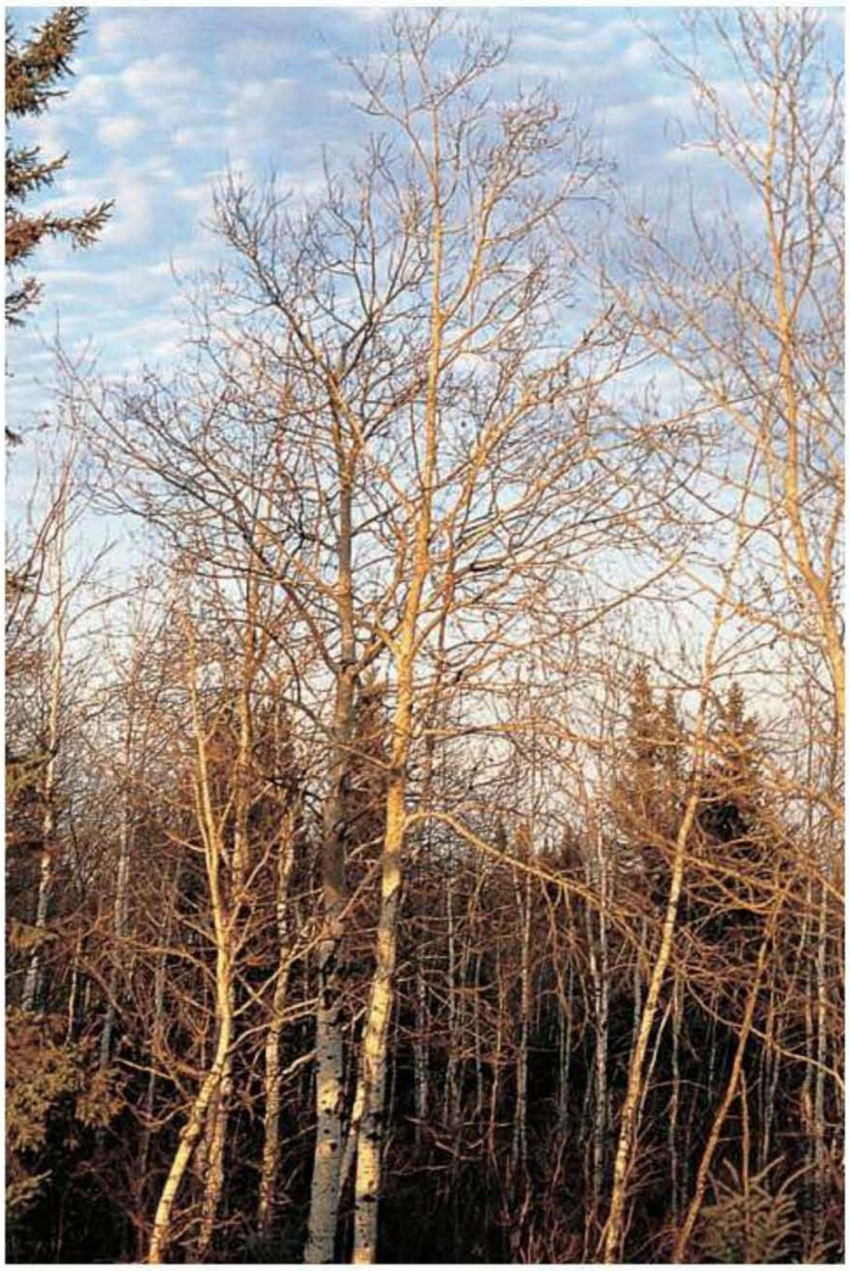

Mixed forests have a blend of conifers and hardwoods. The combination of a good food supply and abundant cover enables these forests to produce more wildlife than coniferous or deciduous forests alone.

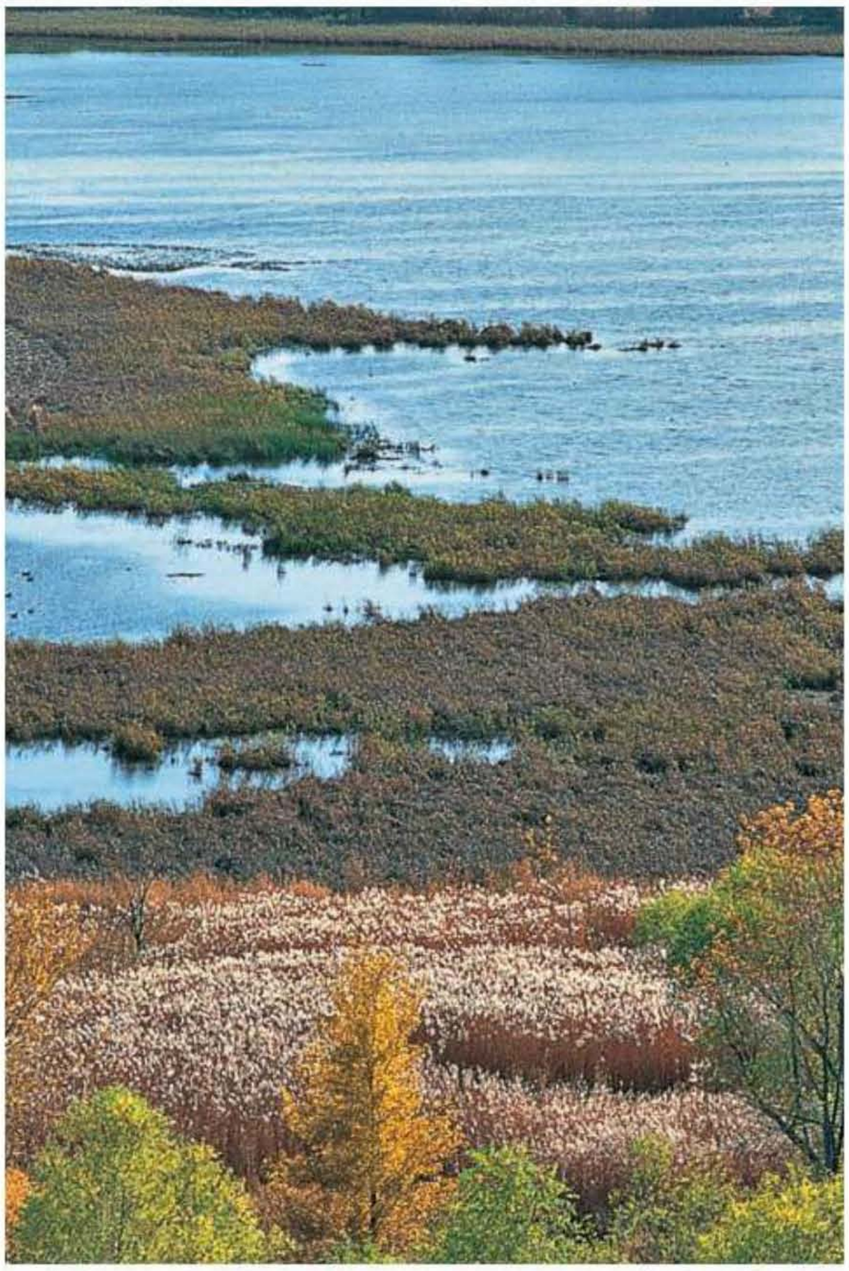

Wetlands have pockets of open water and dense stands of emergent vegetation like cattails and cane. The open water produces food for waterfowl and the fringes provide cover for a wide variety of game.

Semi-arid deserts contain shrubs, grasses and cactuses and may have some trees. They support more types of game than arid deserts, which have only scattered shrubs and cactuses.

Agricultural lands with a variety of cover types support a diversity of wildlife. The fertile soil produces abundant food and larger animals than most other types of habitat.

Chapter 2

HUNTING SKILLS AND EQUIPMENT

When you choose the path of the hunter, you adopt a lifestyle and an outlook uniquely suited to the chase. It is a pursuit that quickens the heart and sharpens the focus. And the life skills and attitudes formed cross over into other facets of life. Hunting skills are gained over the course of several seasons and there is always more to learn, keeping you fit, vital and interested.

Gear to fit the small game hunter differs from that suited to the waterfowler. And the way of the traditional bowhunter is widely divergent from the style of the rifleman. Yet all share a bond and an ethos.

The complete hunter is versed in many skills and knows the operating systems and the nomenclature of a variety of tools. Most learn to hunt with a .22 rifle and then graduate to a centerfire rifle and shotgun. With knowledge of rifle and shotgun comes a passing acquaintance with ammunition and ballistics.

Some hunters opt to shoot a bow. The traditional bow (recurve or longbow) is a good place to start to learn the mechanics of archery. Next comes the compound bow, with its unique technical features and shooting style.

Hunters may opt to hunt with a muzzleloader rifle for a variety of reasons. It too comes with the need to gather knowledge and insight. There are several ignition systems, various powders and an array of projectiles designed to fit specific situations.

There is a lot to learn, but the process is rewarding. Along the way, the hunter faces challenges, but also finds satisfaction, building on successes in the field. Soon, success begins to show with meat on the table and trophies in the living room.

THE HUNTING RIFLE

A rifle that suits your style of hunting greatly increases your chances of success. When selecting a rifle, consider its action, weight and caliber.

ACTION. The action of a rifle refers to the design of the mechanism for chambering ammunition and ejecting spent cartridges. Single-shot actions, like the falling block, are extremely reliable because they have few moving parts. Their rigid mechanisms hold the cartridge firmly, resulting in a high degree of accuracy. These rifles must be reloaded after each round is fired, so they teach shooters to make the first shot count.

Repeating actions hold several cartridges in the magazine, making it possible to fire more than one shot without reloading. Like single shots, bolt actions have a rigid design and few moving parts. Many hunters consider the bolt action to be the most accurate and reliable rifle. Lever, pump and semi-automatic actions are designed for faster firing. You may need a quick second shot if you fail to kill the animal or if a branch deflects your bullet. But fast-action rifles have more moving parts, increasing the chance of mechanical failure, especially in cold weather or when dirt gets into the action.

WEIGHT. Rifles commonly weigh between 6 and 9 pounds (2.7 and 4 kg) without cartridges, slings or scopes. The recoil or kick, of a rifle depends mainly on its weight and its cartridge/caliber combination. Within a given combination, a heavy rifle has the least recoil, because the weight absorbs much of the energy that would otherwise be transmitted to your shoulder. Heavy rifles are easiest to hold steady and generally result in more accurate shots. Light rifles are easier to carry over a long distance.

The term carbine denotes a light rifle with a short barrel. Some hunters prefer carbines when hunting in heavy cover, where a longer rifle would tend to catch on brush or tree limbs.

CALIBER. The caliber of a rifle refers to the diameter of the barrel opening or bore. Caliber is measured in hundredths or thousandths of an inch or in millimeters. Hunters use rifles as small as .17 caliber for small animals and rifles up to .458 caliber for big game. Many rifle models are available in a choice of several calibers. Within a particular model of bolt action there may be fifteen different calibers.

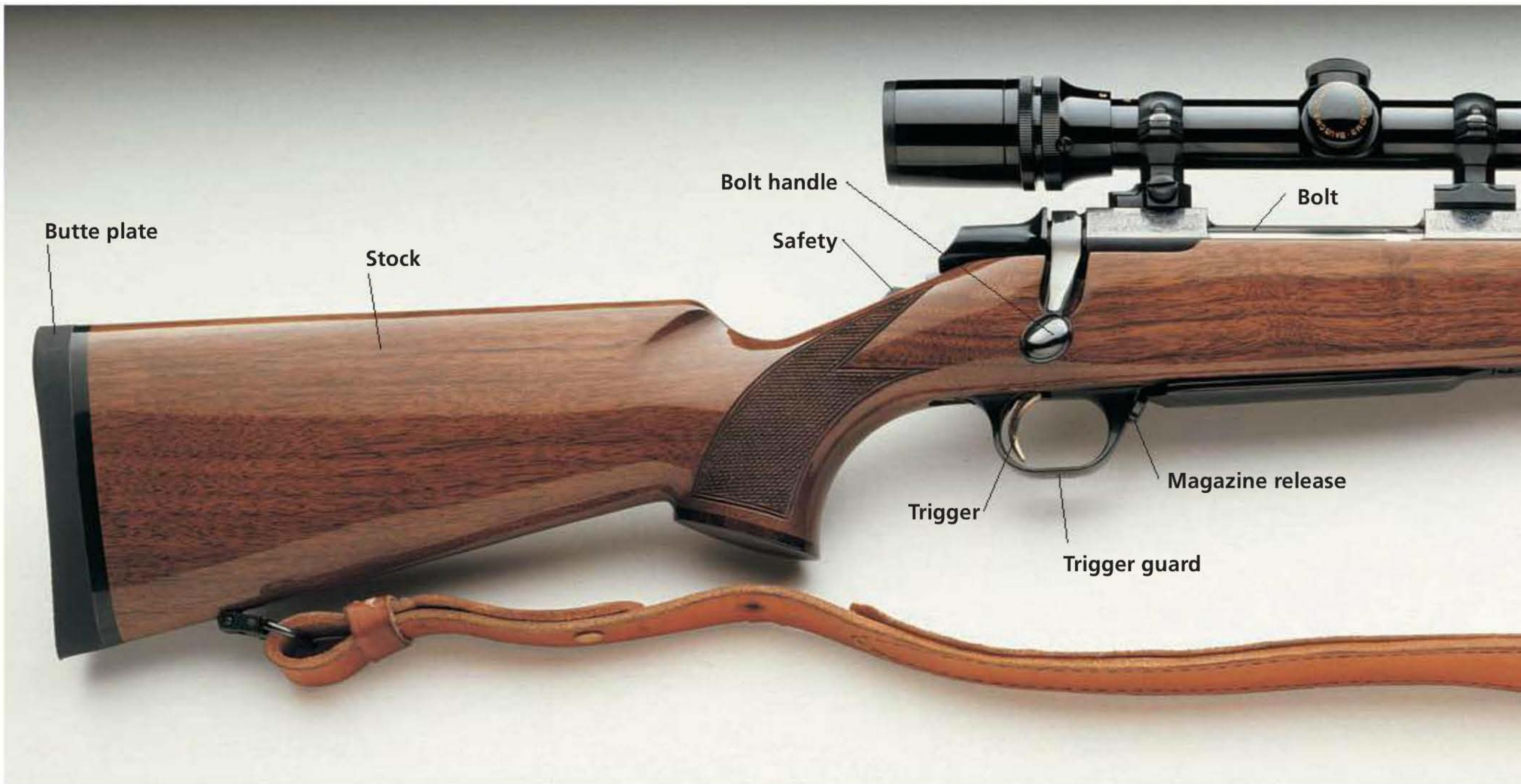

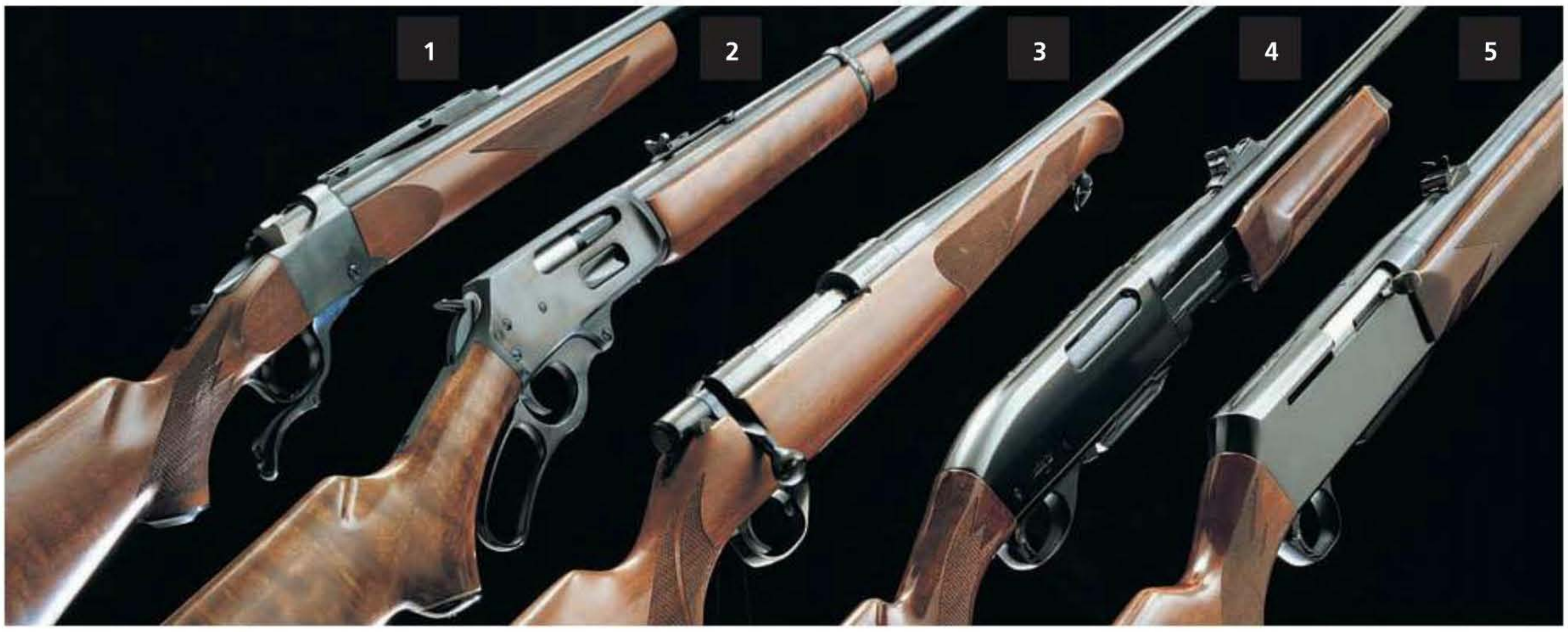

Popular actions include: (1) falling block and (2) lever actions, which chamber and eject cartridges by means of a lever under the trigger guard; (3) bolt action, in which the cartridge locks in place by moving the bolt forward and downward, much like a door bolt; (4) pump or slide action, which chambers and ejects by sliding the forearm back and forth; (5) semi-automatic or autoloader action which chambers and ejects cartridges automatically after each shot. The action is powered by gases from the burning powder or by the recoil.

Hunters are often confused by traditional caliber designations. A .30-06 rifle has a 0.3-inch-diameter (7.6 mm) bore. The 06 refers to 1906, the year the rifle was introduced. A .30-30 rifle also shoots a .30 caliber bullet. The bullet was originally propelled by 30 grains of smokeless powder. To further complicate the matter, some caliber designations do not refer to the diameter of the bore. For example, the bore of a .308 Winchester rifle does not measure 0.308 inch (7.8 mm). The .308 refers to the groove diameter or the diameter to the outside of the rifling grooves. The bore diameter is only 0.3 inch (7.6 mm).

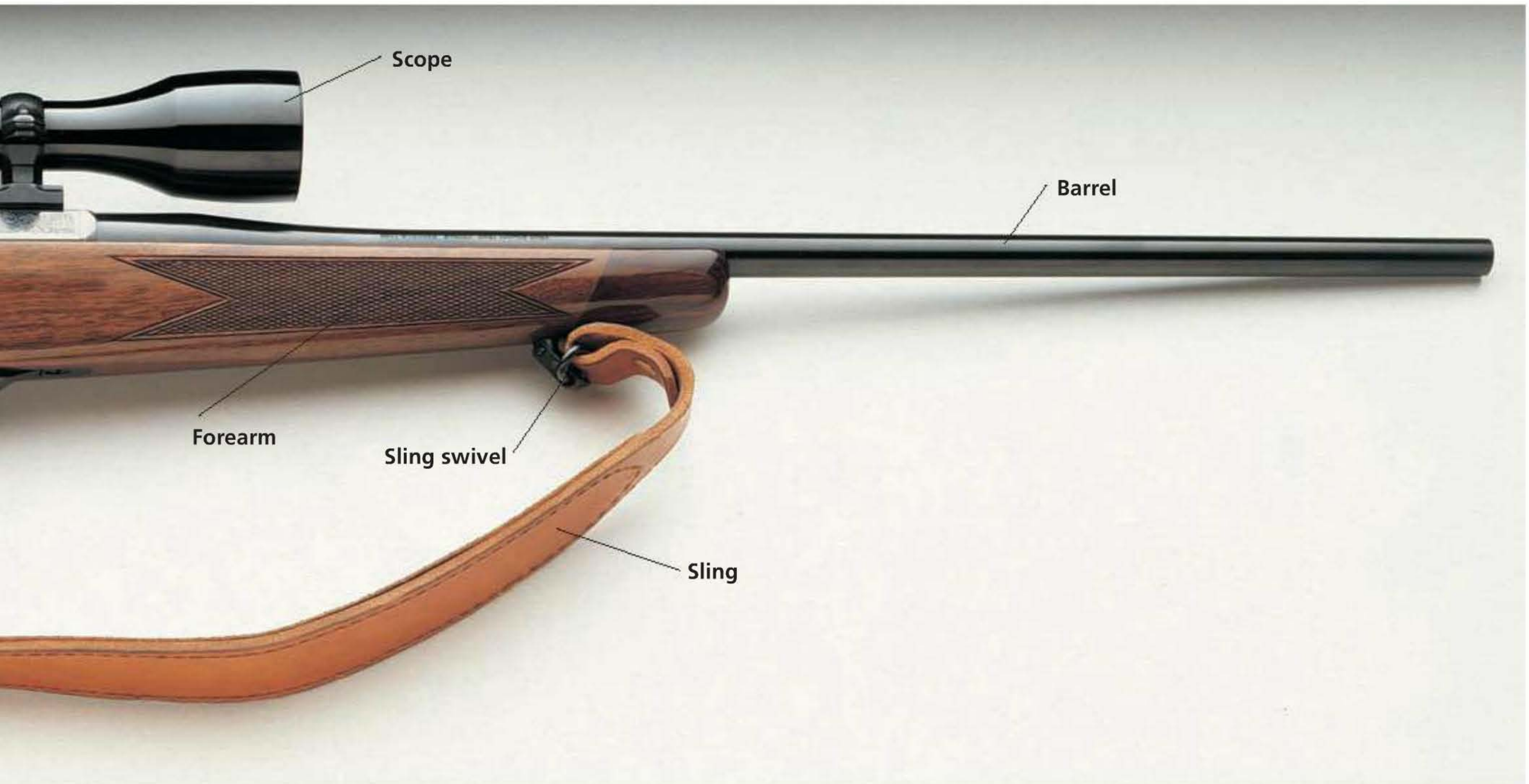

RIFLE AMMUNITION

A hunter's choice of ammunition depends mainly on the size of the animal hunted and the distance at which it may be encountered.

To determine the ammunition best-suited for the game, consult a ballistics table. Most ammunition catalogs contain ballistic information including bullet energy, muzzle velocity and drop, all at different ranges.

A bullet's hitting power, or energy, is determined by its weight and velocity. The heavier and faster the bullet, the more energy it delivers to the target. At normal shooting ranges, a light, fast bullet generally delivers as much energy and kills as effectively as a heavier but slower bullet. Hitting power is measured in foot-pounds (Newton-meters).

Muzzle velocity or bullet speed is measured as distance traveled per second. The faster the bullet is, for a given weight, the more energy it carries. It also flies flatter. When comparing bullets of the same size, the more powder the case holds, the more velocity a bullet has. The size of the brass case affects how much powder is behind the bullet.

The bullet's performance is affected by its size, weight and shape. The bullet's size or diameter is referred to as its caliber. This diameter corresponds to the diameter of the rifling in the gun. Caliber is measured in decimal fractions of an inch, i.e., .300 or .30 is a 30 caliber. Some calibers are expressed in millimeters. So a 7 mm is approximately a .280; 6 mm is a .240; and 7.5 mm is approximately a .300.

Bullet weight is measured in grains.

The shape of a bullet determines its friction or resistance in the air, affecting its trajectory or flight path. A long, thin, streamlined pointed bullet retains its velocity at longer ranges better than a short, stout bullet. So the longer bullet has a flatter trajectory. Trajectory is also affected by the bullet's velocity. A fast bullet flies flatter than a slow one.

The construction of a bullet also determines its ability to penetrate upon impact. The bullet should mushroom on impact and not break apart. Mushrooming transfers all the bullet's energy into the target.

Certain ammunition shoot better in your rifle; each gun is a little different. Today's factory loads can perform up to the highest standards. Try different brands and loads until you find a particular cartridge that meets your needs and with which you can shoot tight groups.

Never use ammunition other than that recommended for the rifle by the manufacturer. The result could be a damaged rifle and serious injury. The correct cartridge designation for the rifle is stamped on the barrel next to the receiver or chamber.

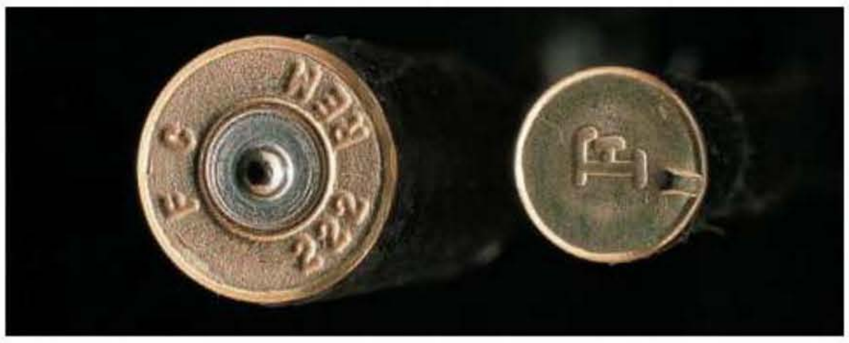

Center-fire cartridges (far left and below) consist of: a brass case; a primer containing a highly explosive compound that ignites when struck by the firing pin; powder, which is ignited by the priming compound; a lead bullet, which may have a copper-zinc jacket to control the bullet's expansion. With rimfires (near left), the firing pin strikes the edge of the rim, igniting the priming compound.

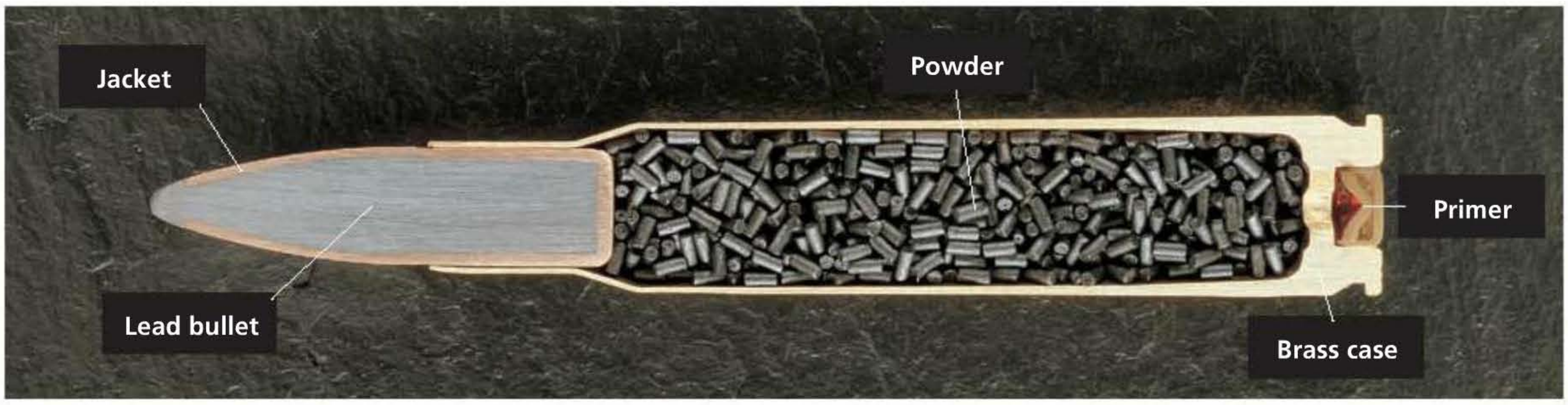

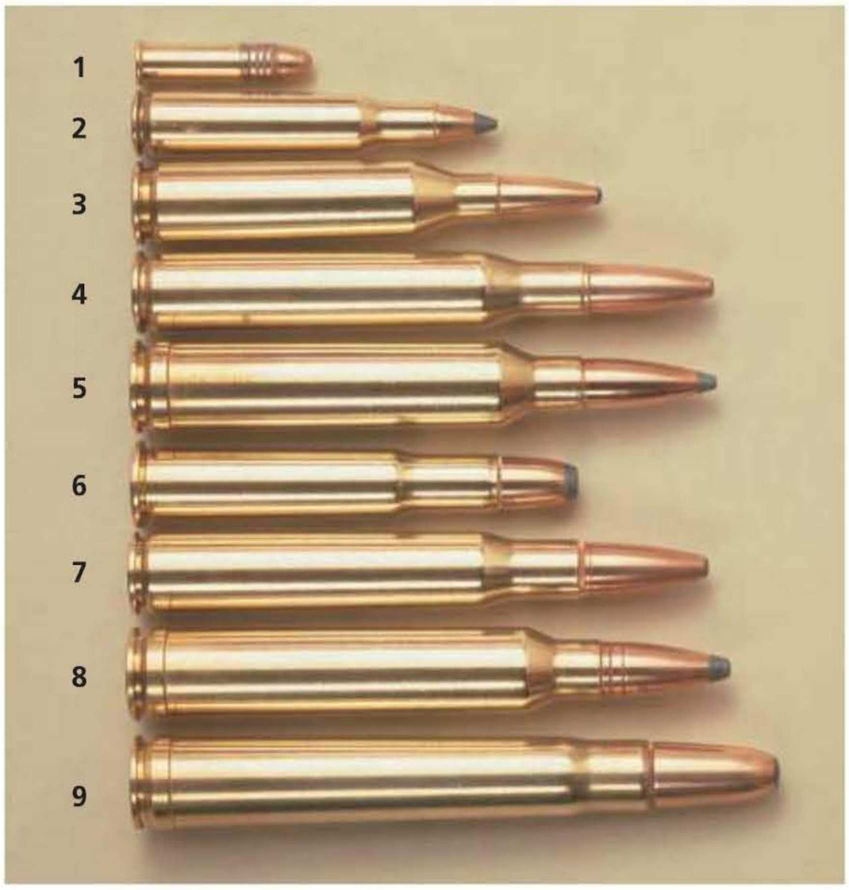

Cartridge types include: (1) .22 Long Rifle 40 Grain, (2) .222 Remington 50 Grain, (3) .243 Winchester 100 Grain, (4) .270 Winchester 130 Grain, (5) 7 mm Remington Magnum 150 Grain, (6) .30-30 Winchester 170 Grain, (7) .30-06 Springfield 150 Grain, (8) .300 Winchester Magnum 180 Grain, (9) .375 H&H Magnum 270 Grain.

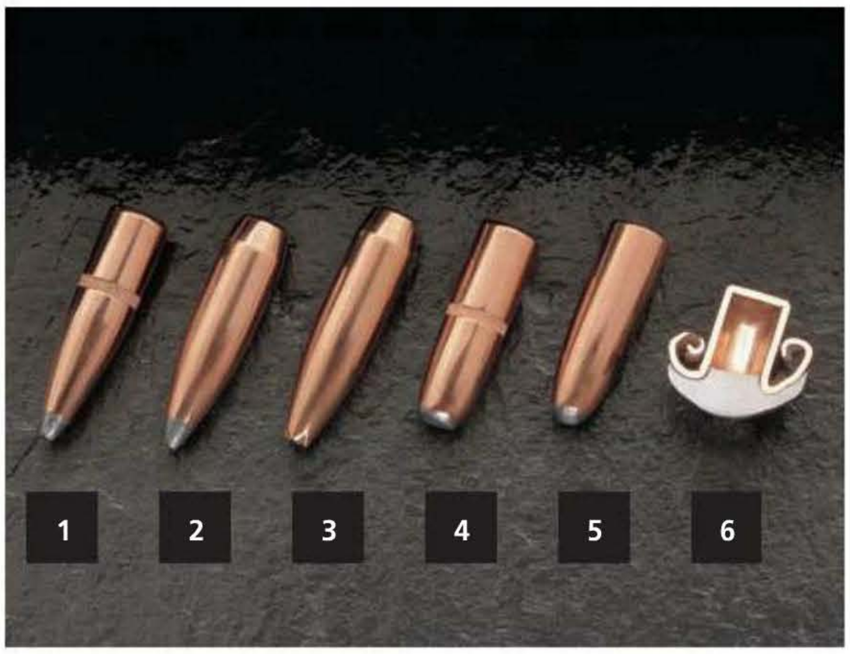

Bullet styles include (1) pointed, streamlined for high velocity and flat trajecto-ry. (2) Pointed boat tail has a tapered base to reduce drag even more. (3) Hollow points mushroom rapidly, but no faster than (4) flat and (5) round points, often used in rifles with tubular magazines. Their blunt points do not detonate cartridges ahead of them, but result in relatively slow flight and short range. (6) Mushroomed bullet shows expansion.

TYPE OF CARTRIDGE (not actual size)	**Muzzle Velocity (in fps)**	**ENERGY (in foot pounds)**				**TRAJECTORY**				
		100 yds	**200 yds**	**300 yds**	**400 yds**	**100 yds**	**150 yds**	**200 yds**	**300 yds**	**400 yds**
.22 Long Rifle 40 Grain	1255	92	—	—	—	0.0	–11.5	—	—	—
.222 Remington 50 Grain	3140	752	500	321	202	+2.2	+1.9	0.0	–10.0	–32.3
.243 Winchester 100 Grain	2960	1615	1332	1089	882	+1.9	+1.6	0.0	–7.8	–22.6
.270 Winchester 130 Grain	3060	2225	1818	1472	1180	+1.8	+1.5	0.0	–7.4	–21.6
7 mm Remington Magnum 150 Grain	3110	2667	2196	1792	1448	+1.7	+1.5	0.0	–7.0	–20.5
.30-30 Winchester 170 Grain	2200	1355	989	720	535	+2.0	0.0	–4.8	–25.1	–63.6
.30-06 Springfield 150 Grain	2910	2281	1827	1445	1131	+2.1	+1.8	0.0	–8.5	–25.0
.300 Winchester Magnum 180 Grain	2960	3011	2578	2196	1859	+1.9	+1.6	0.0	–7.3	–20.9
.375 H&H Magnum 270 Grain	2690	3510	2812	2228	1747	+2.5	+2.1	0.0	–10.0	–29.4

Ballistics tables help determine the effective killing range of the cartridge and bullet drop at various ranges. This table lists muzzle velocity and the energy in foot-pounds delivered at ranges from 100 to 400 yards (91 to 364 m). As a rule, use a cartridge with at least 900 foot-pounds (1,220 N-m) to kill a deer, 1,500 (2,031) for an elk and 2,100 (2846) for a moose. Trajectory figures show how far the bullet strikes above (+) or below (-) the point of aim at ranges from 100 to 400 yards (91 to 364 m).

SIGHTS

The sighting system you use is as important as the rifle and ammunition. Some hunters spend as much for a rifle scope as they do for the rifle itself.

Many rifles come with iron sights. To aim, center a post at the end of the barrel in a notch or peephole in the rear sight. Iron sights are inexpensive, lightweight, durable and best-suited for short-range shooting.

For more accurate, precise shooting or for shots at longer distances, hunters prefer telescopic sights or scopes. A scope consists of a metal tube containing a system of lenses that magnify the target. The reticle, a network of lines or crosshairs inside the scope, enables you to aim precisely. The optics allow you to focus your eye on the reticle and target at the same time, so you can aim quickly. Scopes are not as durable as iron sights and are easier to knock out of adjustment.

Scopes vary in magnification power, from lx, which magnifies the target 1 time, to 12× or more. Fixed power scopes have one magnification power. Variable power scopes allow you to change the magnification with the twist of an adjustment ring. Most hunters use scopes ranging in power from 2× to 9×. Low-power scopes work best for close-range shooting. They have a larger field of view, making it easy to find your target. High-power scopes narrow the field of view, but help you see a distant animal.

A scope is attached to the gun using mounts or bases and rings. Whatever the mounting system, it must fit the scope, attach solidly to the gun and be adjusted to provide enough distance between the scope and your eye. Scopes should be at least 3 inches (7.6 cm) away from your eye, when the gun is shouldered, to allow for clearance when the gun recoils.

Most manufacturers offer a selection of special-purpose scopes designed specifically for shotguns and muzzleloaders.

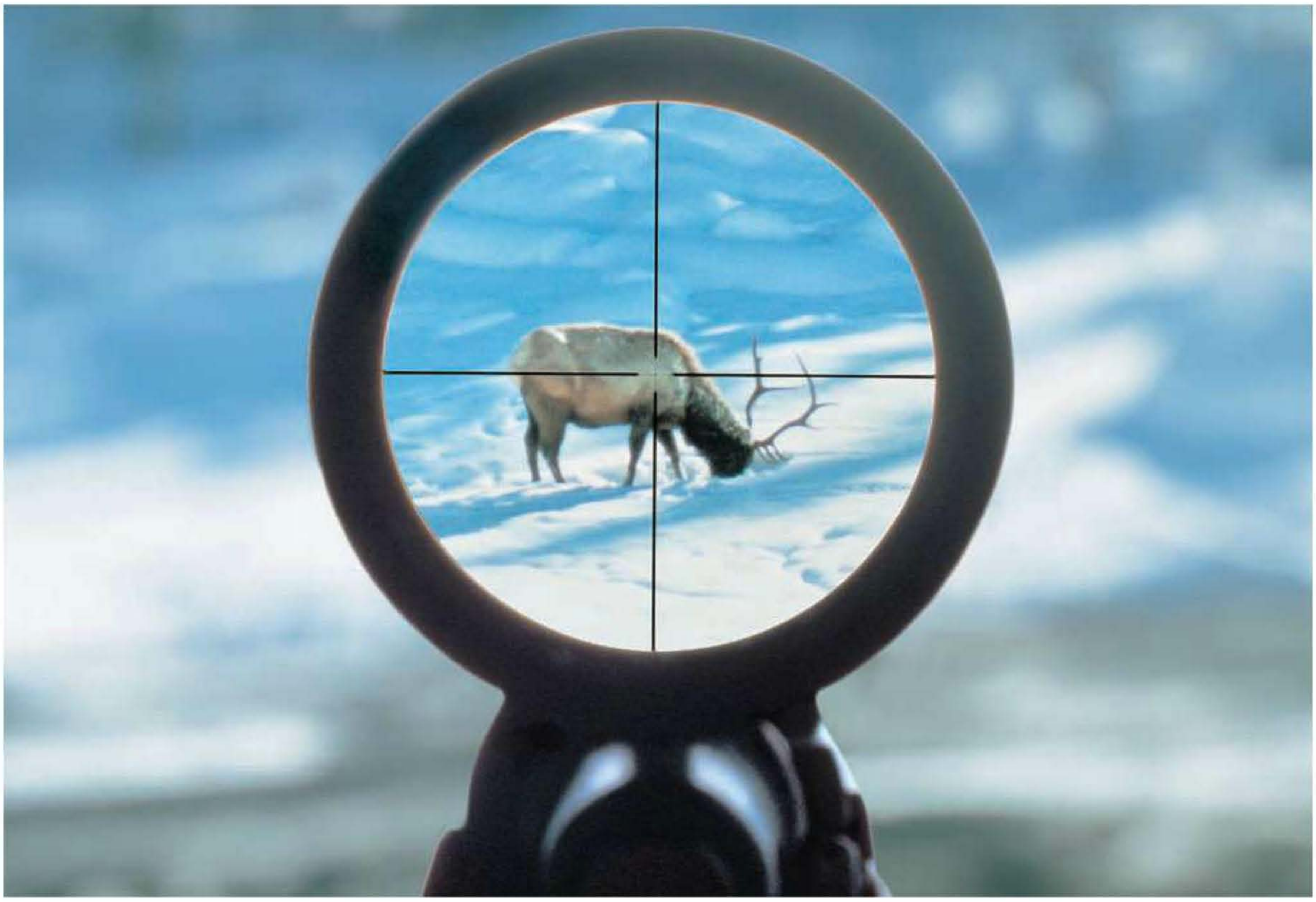

Four-plex reticles have wires that are thick toward the outside. These wires stand out against the background and draw your eye to the target. The thin crosshairs enable you to aim precisely, without obscuring the animal.

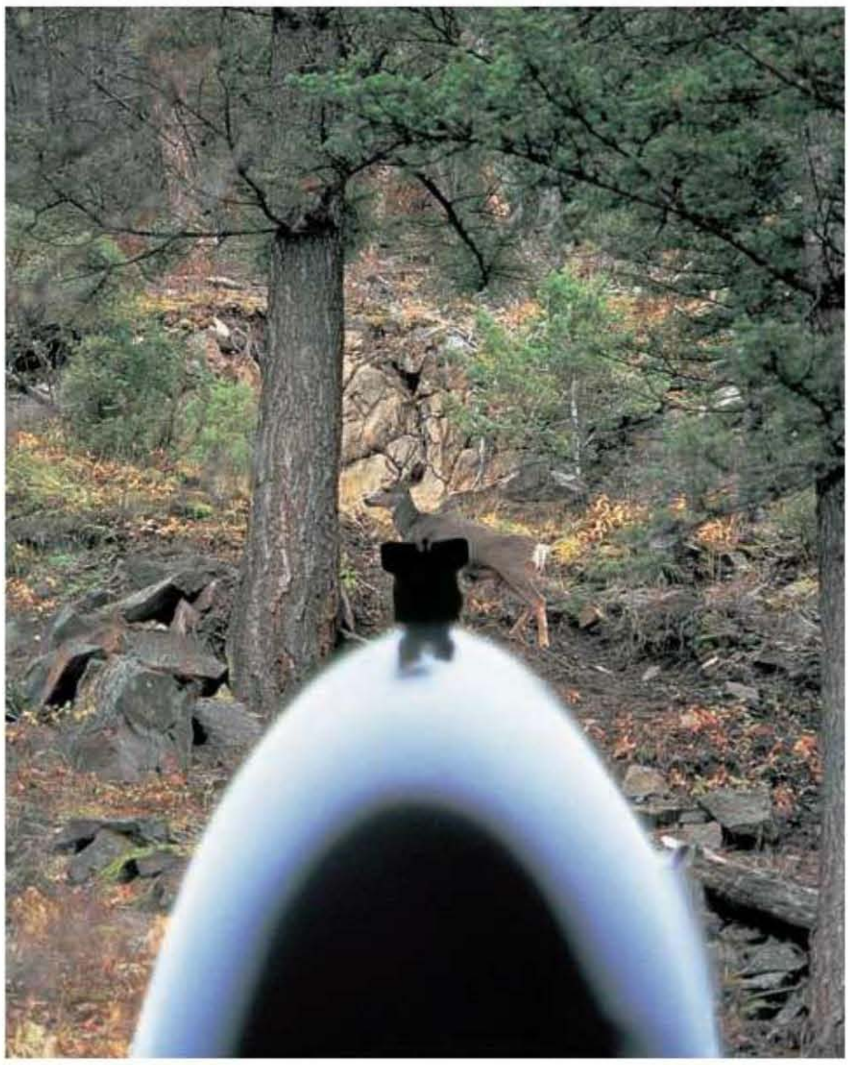

Open sights have a blade or post, which must be centered in the notched rear sight. They are the most difficult of the sights to align and the rear sight covers up part of the target.

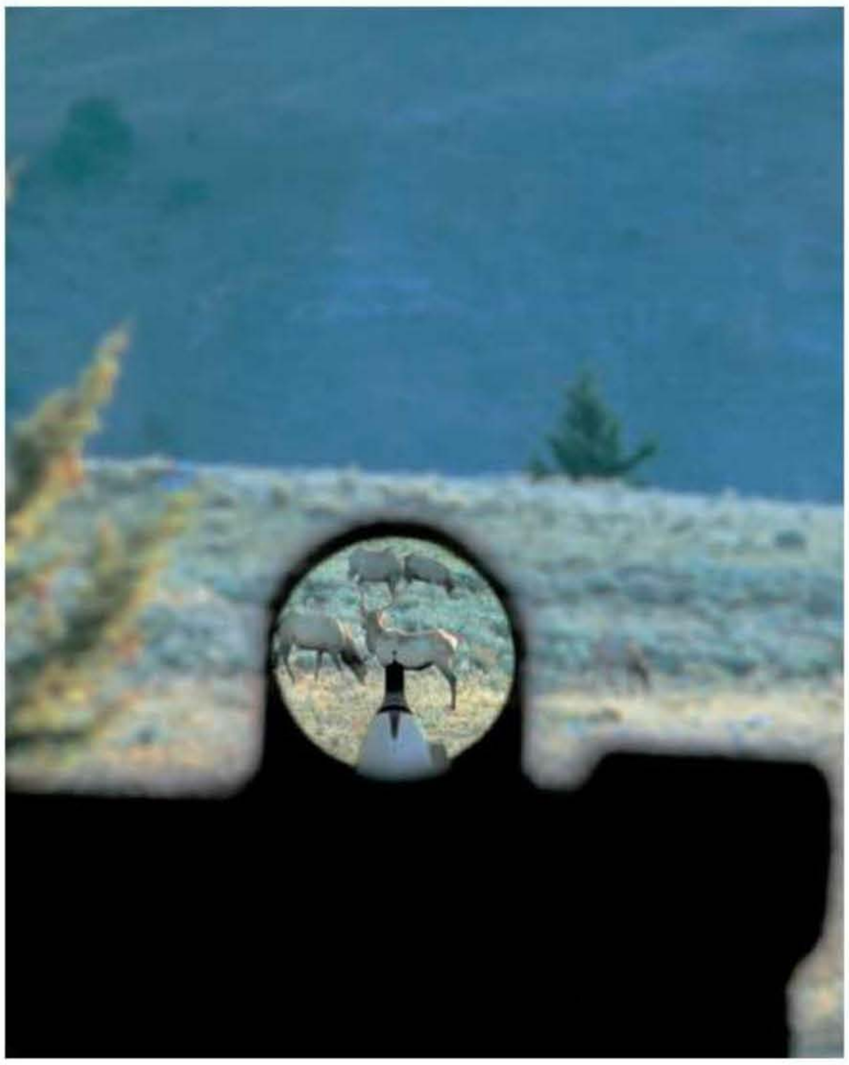

Peep sights require you to center the target in a hole in the rear sight, then place the bead directly on the animal. Hunters prefer large peep-holes so they can see most of the target.

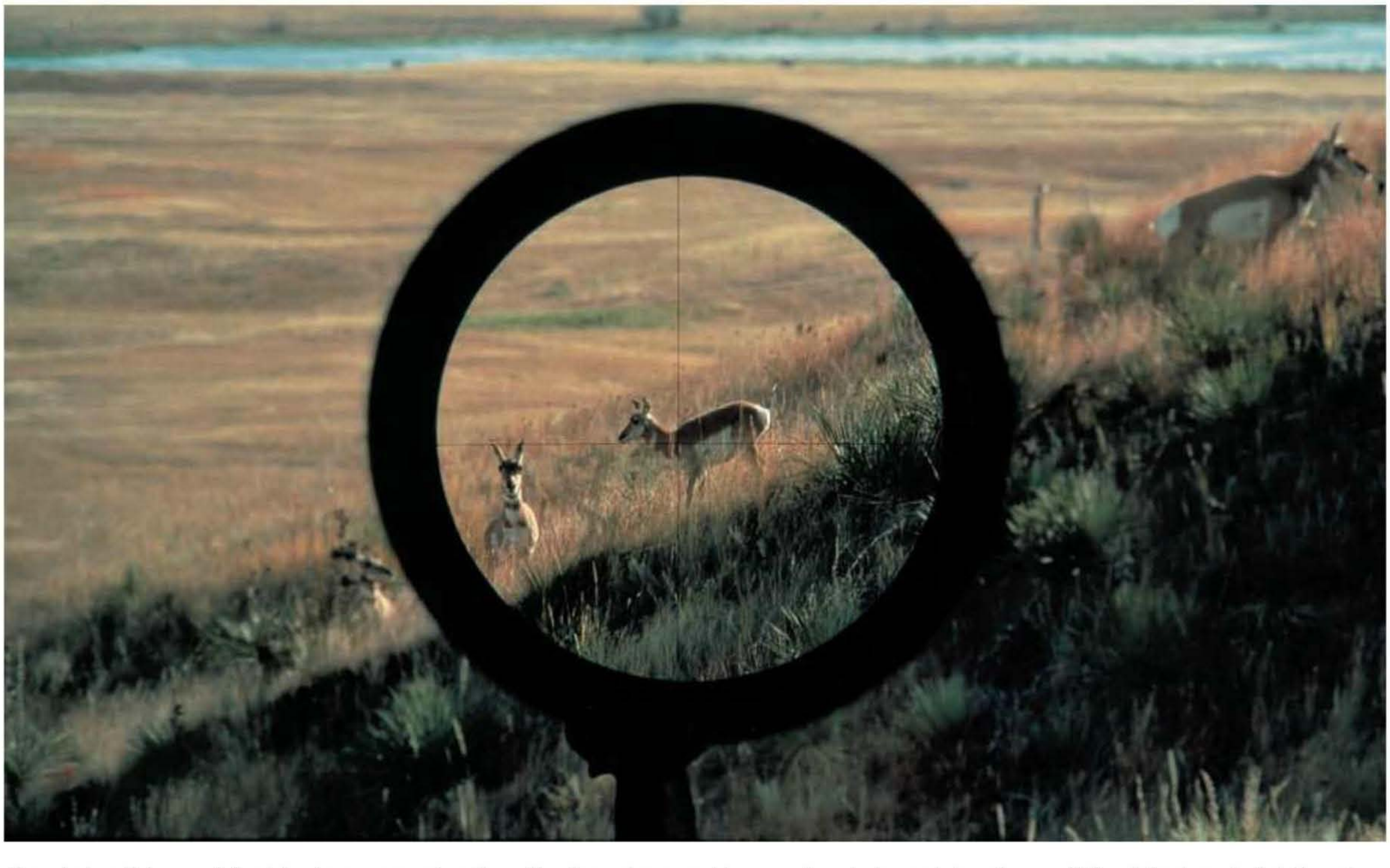

Crosshair reticles work best for long-range shooting. The fine wires provide a precise aiming point and cover little of the target. But the wires may be difficult to see in dim light.

SIGHTING IN YOUR RIFLE

Every season, hunters miss their chance at a trophy because their rifles were not sighted in. Many people think that bore-sighting is sufficient. It is not. A bore-sighted rifle may shoot 20 inches (50.8 cm) or more from point of aim. Before hunting, take several practice sessions with your hunting ammunition.

You can sight in most easily at a shooting range. Most ranges have bench rests and sandbags to provide a steady rest and minimize human error. If sighting in your rifle in the field, make sure you shoot into a solid backstop.

To shoot accurately, take a deep breath, exhale halfway, then hold your breath as you squeeze the trigger. Do not jerk the trigger as you shoot.

Sights on a new rifle may be so far out of adjustment that you miss the target completely. To solve this problem, bore sight your rifle. Look directly through the bore of the bolt or falling block action. You cannot look directly through the bore of most other actions, so you must use a bore-sighting tool.

You can also rough-sight your rifle by simply aiming at a close target through the sights, firing, then making any adjustments needed to hit the bull's-eye.

Once you have rough-sighted at 25 yards (23 m), back off to 100 yards (91 m). Continue to fine-tune the sights until you can center the shot group on the bull's-eye. If you plan to shoot at longer ranges, sight in a few inches (cm) high at 100 yards. Then try a few long-range shots and make any necessary adjustments.

Always hunt with the same type of ammunition you used to sight in. Changing brands or bullet weights often makes it necessary to realign your sights.

Rest the forearm and butt stock on sandbags when sighting in. Do not support the barrel; the result will be inaccurate shooting. Hold the stock firmly against your shoulder. Always wear eye and ear protection when on the range.

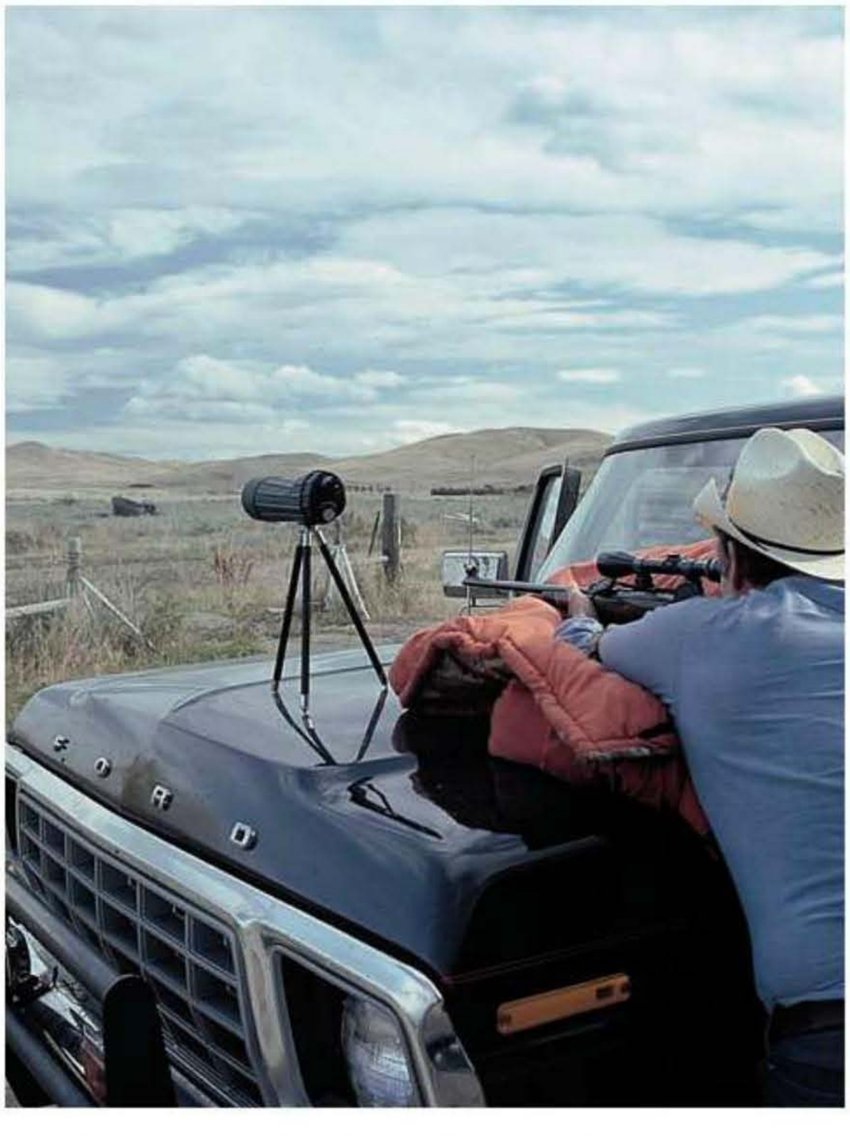

Sight in your rifle so the bullet hits just above the point of aim at 100 yards (91 m). For most flat-shooting cartridges, you should zero-in 2 to 3 inches (5 to 7.6 cm) high; the exact distance depends on the trajectory of your bullet. By sighting in this way, you can aim dead-on at your target at any range up to 300 yards (274 m). The bullet hits slightly high or low, depending on the range, but strikes somewhere in the animal's vital area.

HOW TO ROUGH-SIGHT A RIFLE

Move the rifle barrel until the center of the bore lines up with the bull's-eye of a target 25 yards (23 m) away. Fix the rifle's position so it cannot move, then adjust the sights to aim at the bull's-eye.

Fire a three-shot group. The center of the group is the average point of impact. If the center is not in the bull's-eye, adjust the sights. With a scope you should strive to get a group size of 1 inch (2.5 cm) wide at 100 yards (91 m).

HOW TO ADJUST RIFLE SIGHTS

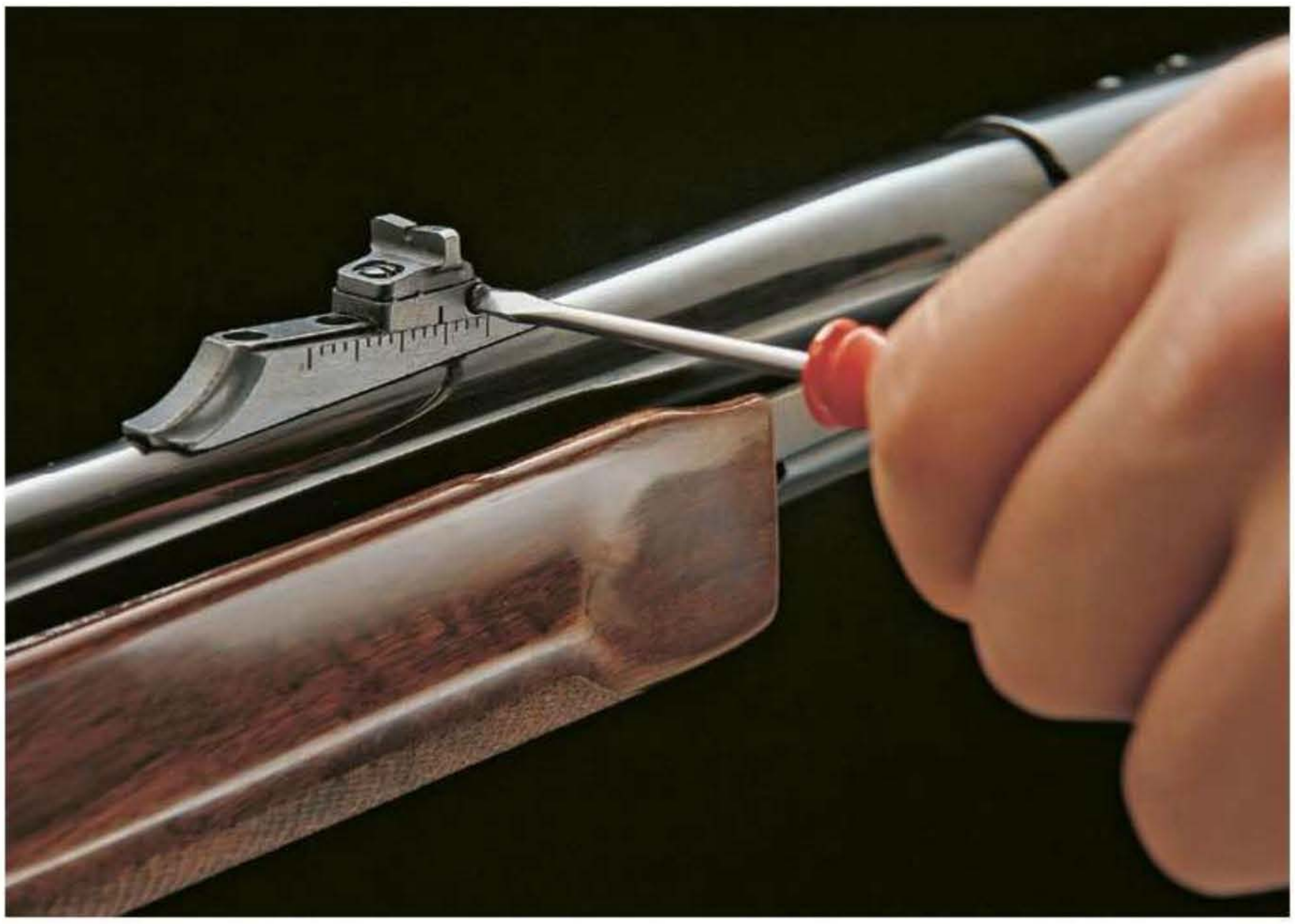

Adjust iron sights by moving the rear sight in the direction you want the rifle to shoot. To make the rifle shoot higher, raise the rear sight. To make it shoot to the left, move the rear sight to the left.

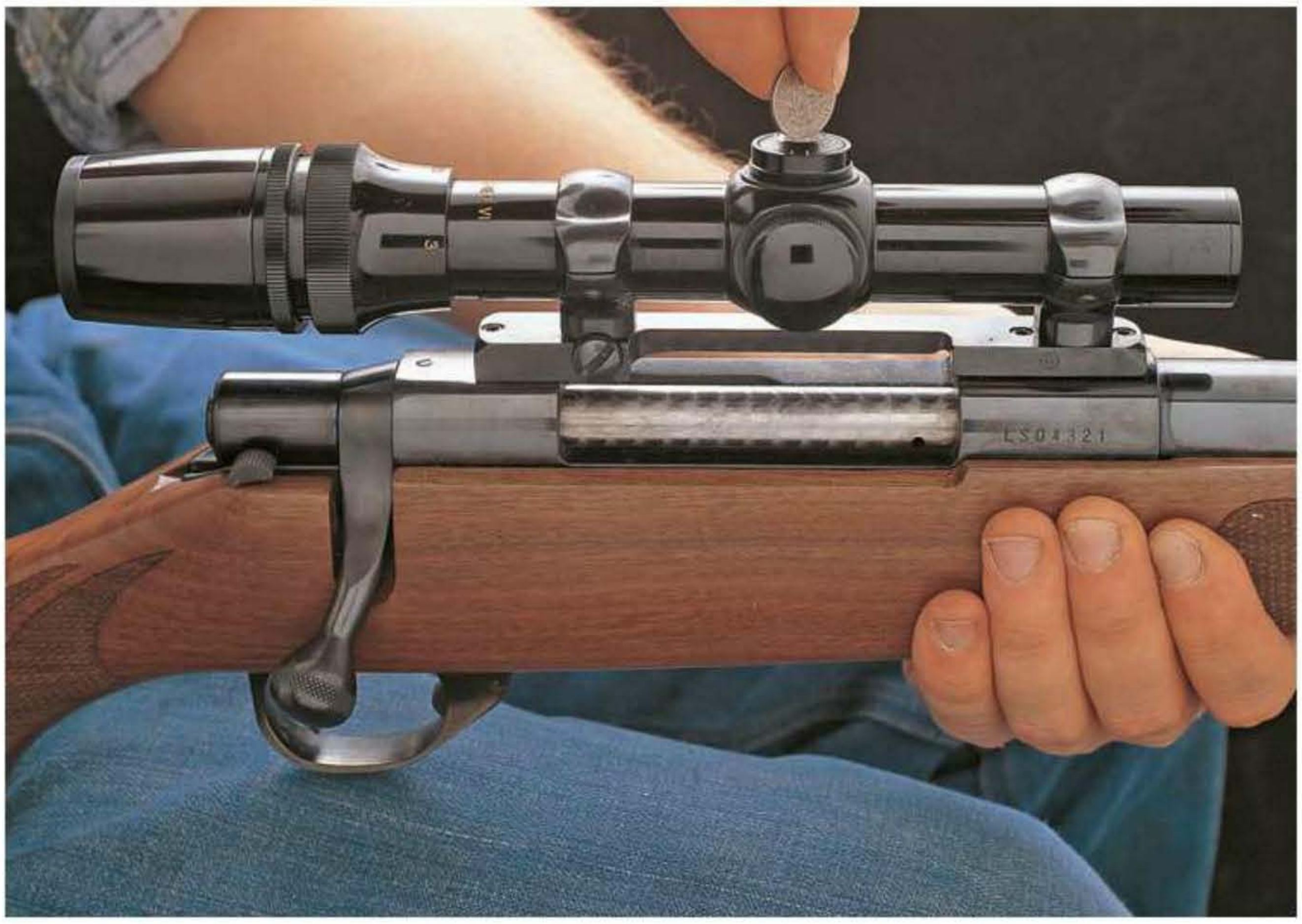

Move scope sights by turning elevation (up-down) and windage (right-left) screws the way you want the bullet impact to move. Read the instruction manual to see how much each click or mark changes the sight at 100 yards (91 m).

SHOOTING A RIFLE

Shooting a rifle under hunting conditions is more difficult than shooting at a target range. To shoot accurately, hunters must assume a stable position and know how to compensate for wind, uphill or downhill angles and moving animals.

Whenever possible, try to find a stable object to provide a steady rest for your rifle. If you cannot find such an object, the prone position is your next best choice. But shooting from the prone position may be impossible if ground cover or rolling terrain obscures your line of sight. Under these conditions, experienced hunters use the sitting position, which offers good stability because both elbows rest on the knees. The kneeling position is less stable, because only one elbow rests on the knee. The offhand position is the least stable. However, it may be the only choice when a moving animal offers a brief chance for a shot.

Learn how to adjust for natural forces like wind drift and the slant range effect. A strong crosswind dramatically affects a bullet's flight path. To hit a standing target, you must compensate by aiming slightly upwind.

The slant range effect causes your bullet to hit high whenever you shoot uphill or downhill. This confusing phenomenon results from the fact that gravity acts at a right angle to the horizontal. When a bullet is shot on a horizontal plane, gravity pulls at a right angle over the entire distance to the target, resulting in a curved flight path. But when a bullet is shot vertically, either straight up or straight down, gravity does not bend the flight path at all. The closer the angle to the vertical, the less the flight path curves.

A rifle is usually sighted in on the horizontal plane, so the sights are adjusted to compensate for maximum bullet drop. Since the bullet does not drop as much when shooting uphill or downhill, it always hits high. To compensate, aim low.

Most hunters tend to shoot behind running animals. A hunter using 150-grain .30-06 bullets would maintain a 4- to 5-foot (1.2 to 1.5 m) lead at a deer running at a right angle 100 yards (91 m) away. At 200 yards (183 m), he would increase the lead to about 10 feet (3 m) and at 300 yards (274 m), to about 16 feet (5 m).

Practice your lead by shooting at a rolling tire with a cardboard target wedged inside. Position yourself on a hill, then have someone roll the tire past you.

BULLET DRIFT

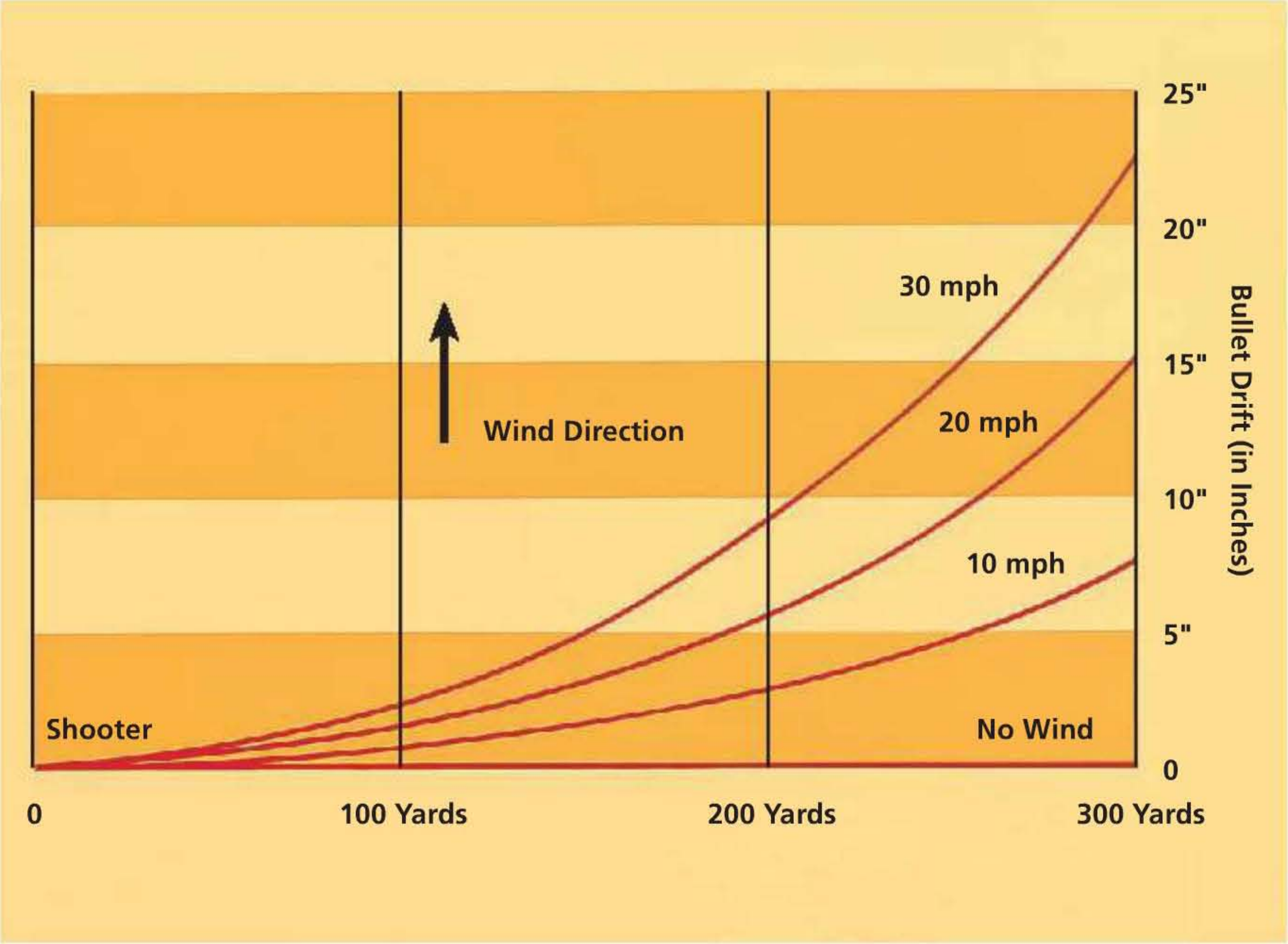

Wind drift increases as wind velocity and range increase. The chart shows how crosswinds of 10, 20 and 30 mph (16, 32, 48 km/h) would affect the path of a .30-06, 150 grain bullet at 100, 200 and 300 yards (91, 183, 274 m). As the chart indicates, a 10 mph crosswind would cause the bullet to drift less than 1 inch (2.5 cm) at 100 yards. But in a 30 mph crosswind, the bullet would drift over 22 inches (56 cm) at 300 yards.

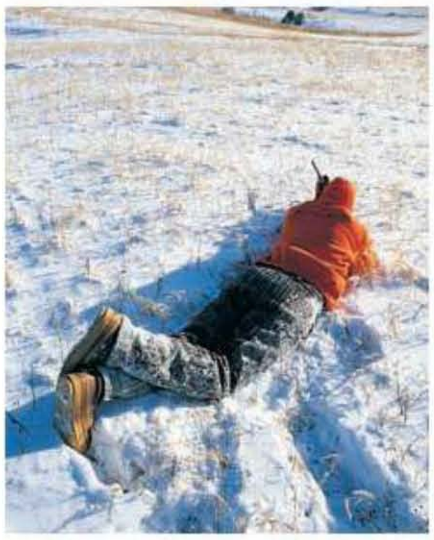

Prone. Lie with your body about 30 degrees left of your line of aim. Place your left elbow just left of the rifle. Pull your right leg forward to lift your stomach off the ground, so your breathing does not affect the shot.

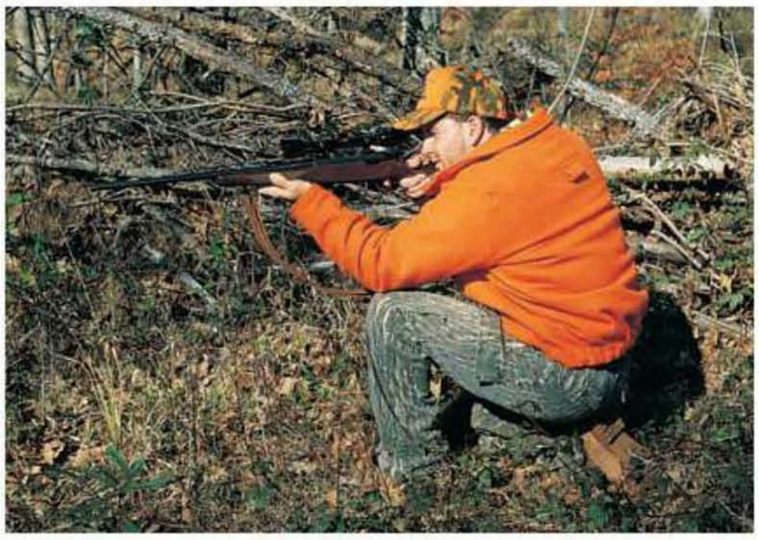

Kneeling. Sit on your right foot with your body 45 degrees left of the line of aim. Place your left foot forward and your left elbow on the knee.

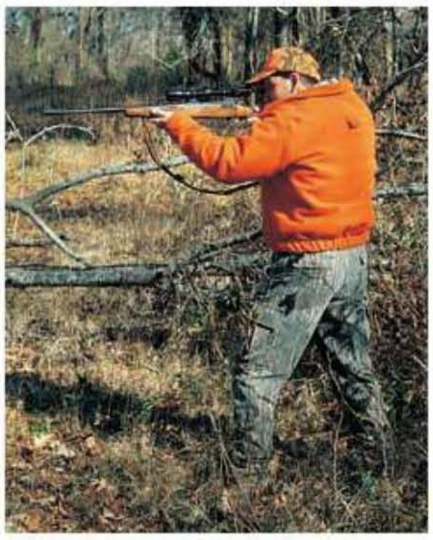

Offhand. Stand sideways with feet perpendicular to the line of aim. Spread your legs to shoulder width. Keep your left elbow close to your body.

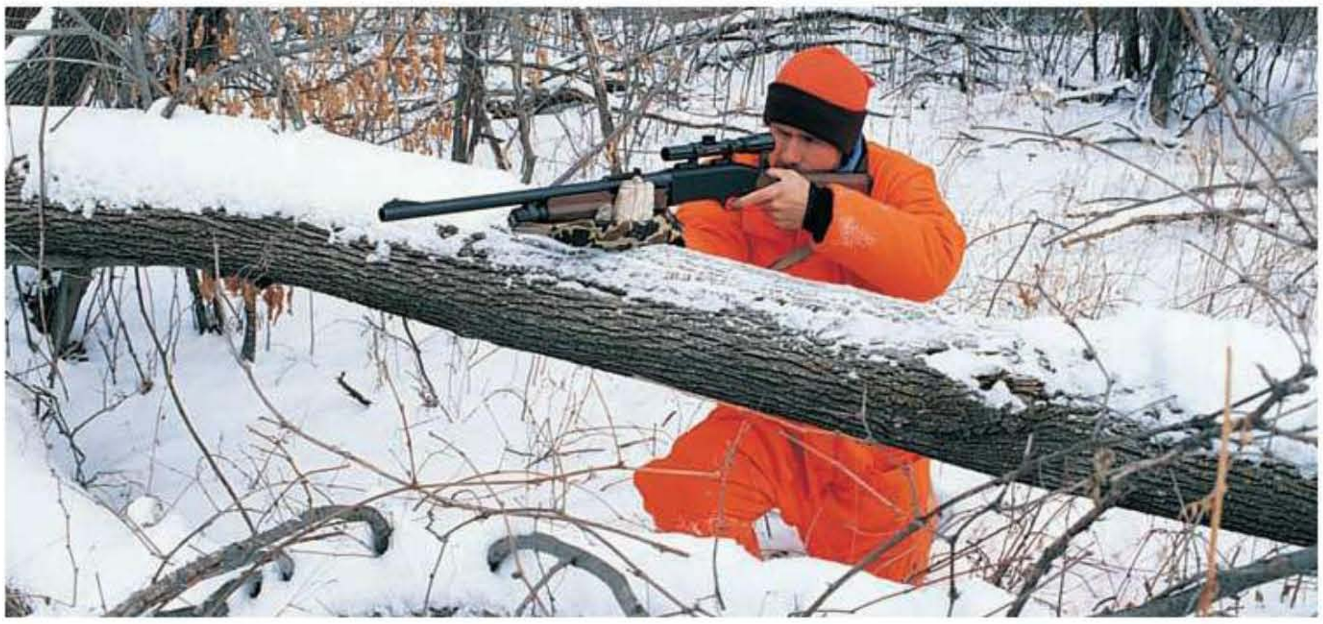

Solid rest. Rest your left hand or elbow on a solid object, like a tree or rock. Do not rest your rifle directly on the object.

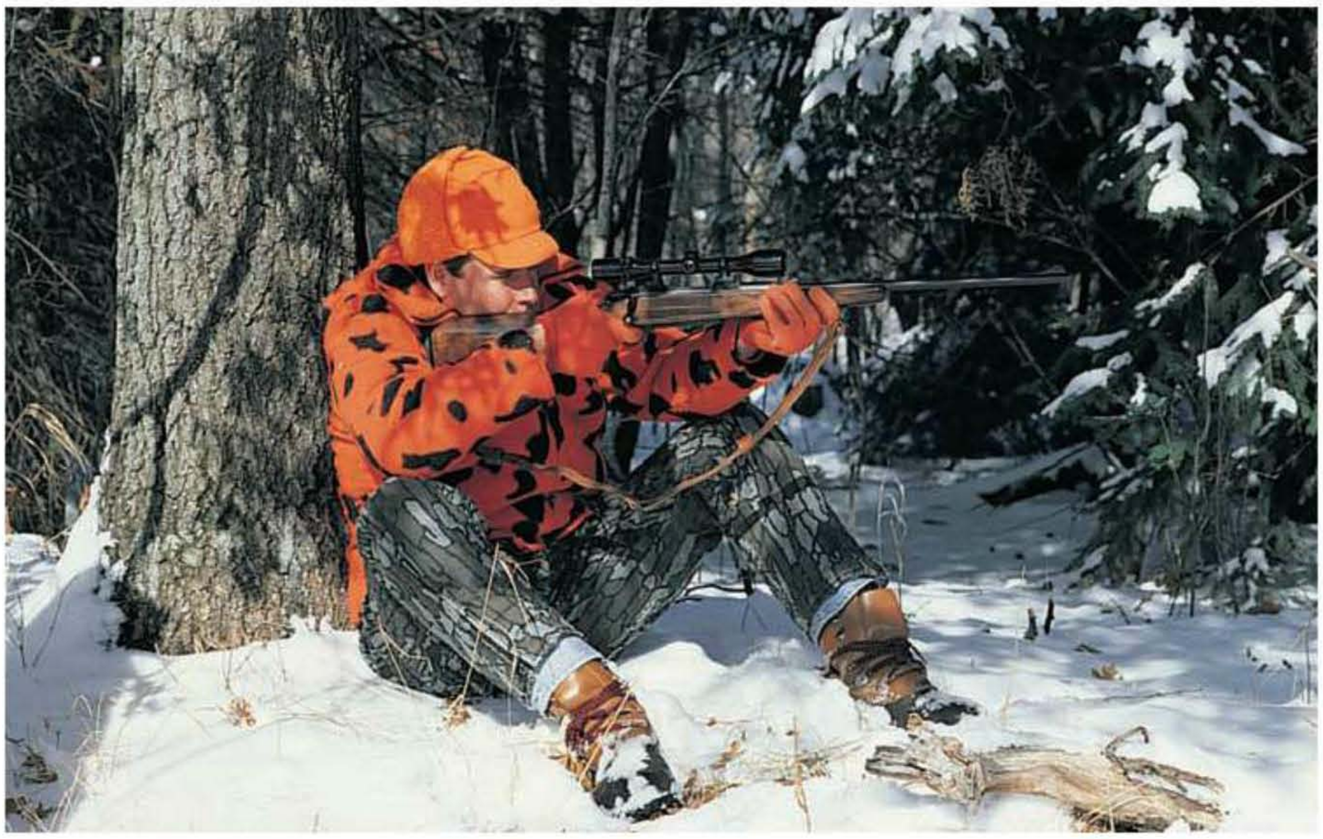

Sitting. If you shoot right-handed, sit with your legs about 30 degrees to the right of your line of aim and rest your elbows firmly on your knees. Sitting is the most useful shooting position. You can use the position almost anywhere, assume it quickly and shoot accurately.

SHOTGUNS

Frontiersmen referred to the shotgun as a scattergun because it sprayed a swarm of lead pellets. A shot swarm is more effective than a single bullet for hitting moving targets.

When selecting a shotgun, consider the following: size of the bore, the chokes you need or the amount of barrel constriction, action, chamber length, weight and barrel type.

BORE SIZE. Shotgun bores are measured in gauge or in inches. As the gauge number increases, the size of the bore decreases. The most common gauges are 12 and 20, but gauges range from as small as 28 to as large as 10. The smallest bore is the .410, the only bore measured in inches. The larger the bore, the more pellets a gun can shoot. The denser shot pattern increases the chances for a long-range kill.

CHOKE. The amount of constriction or choke, at the end of the barrel affects the diameter and density of your shot pattern. Common chokes, from widest (most open) to narrowest (tightest), include: cylinder, skeet, improved-cylinder, modified, improved-modified and full. Open chokes are best for close-range shooting; tight chokes, long-range. The improved-cylinder is a good all-around choice.

Many modern shotguns have interchangeable screw-in choke tubes that enable you to quickly change chokes to suit the type of hunting. A gunsmith can also install screw-in chokes in some older guns.

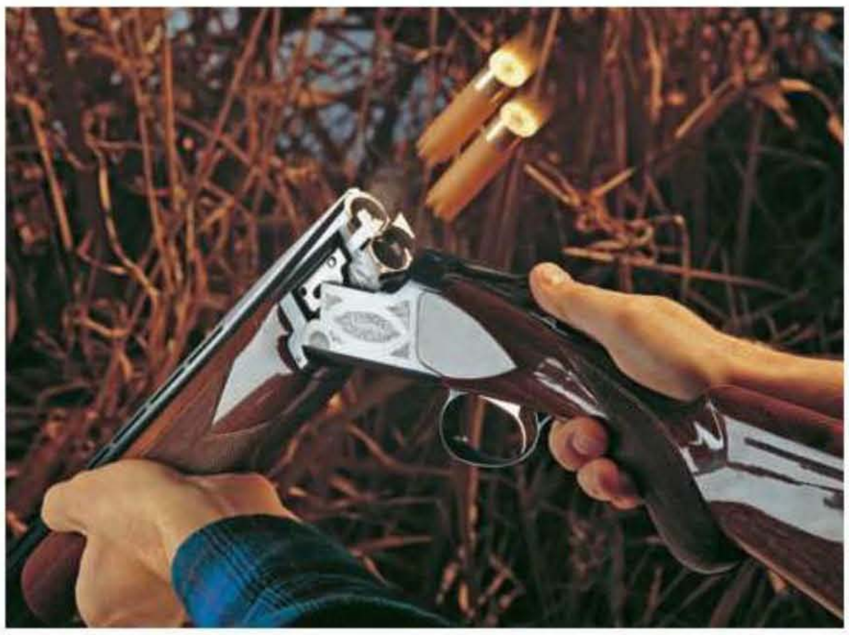

Hinge action. The action opens when you push a lever or button at the rear of the receiver, allowing you to manually insert the shells. After firing, you break the action again. If your gun has ejectors, the spent hulls pop out automatically; otherwise, you must remove them yourself.

ACTION. Most single-shot and double-barreled shotguns have a hinge action. The simple hinge design is more reliable than repeating actions, like the pump and semi-automatic. Pump shotguns are more dependable than semi-automatics, which can malfunction in cold weather.

A pump gun ejects the spent shell and chambers another one each time you slide the fore-end back and forth. The majority hold five shells.

A semi-automatic fires one shell, ejects it and chambers another with each pull of the trigger. Most hold five shells, although a few hold up to seven.

CHAMBER LENGTH. Many shotguns are chambered for standard 2¾ shells. To shoot many types of magnum loads, you need a gun with a longer chamber. The proper shell length is usually engraved on the barrel. Never shoot a shell that exceeds that length.

WEIGHT. A light shotgun works best for quick shots in heavy cover. When you have more time to shoot, a heavier gun works better. You can hold the barrel steadier and swing on a target more smoothly.

BARREL TYPE. A long barrel provides a long sighting plane and can improve shooting accuracy. But contrary to popular belief, a longer barrel does not mean a noticeable increase in shooting range.

Many guns come with ventilated ribs. A rib makes it easier to sight down the barrel and cools it quickly, an advantage for rapid-fire shooting.

Field loads are adequate for close-range shooting at squirrels and rabbits and for small to medium-sized birds. Standard loads work better at longer ranges and for larger animals. Many hunters believe that magnum loads greatly increase their shooting range. But a magnum of the same length as a standard load may actually have a slightly lower velocity and a shorter range. The advantage of a magnum is a denser shot pattern.

Some magnum shells have longer cases that hold even more shot and powder. A 12-gauge, 3-inch (7.5 cm) magnum, for example, contains 1⅞ ounces (53 gm) of shot and 4 dr. equivalents of powder. Its effective range is about 10 yards (9 m) longer than a standard 2¾-inch (7 cm) load. Never attempt to shoot a long-cased magnum shell in a gun chambered for standard shells.

HOW CHOKE AFFECTS THE SHOT PATTERN

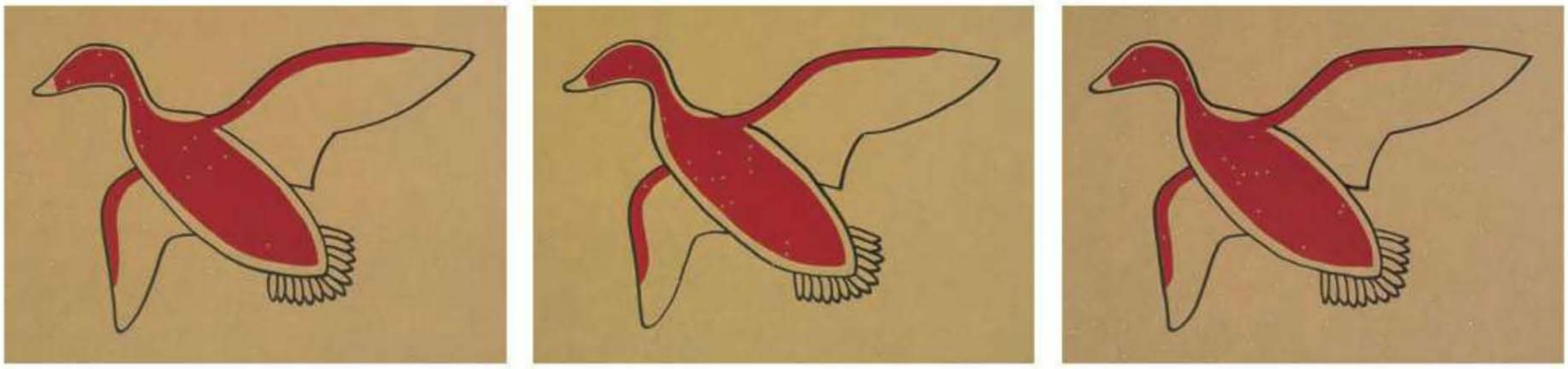

Choke determines how quickly your shot spreads. With a 12-gauge, improved cylinder choke and a standard-load No. 6 shell, about 13 pellets hit the vulnerable area of a stationary mallard silhouette at 40 yards (36.5 m). With a modified choke, about 16 pellets strike the target and with a full choke, about 20. Fewer pellets would strike a flying bird, because not all pellets in the shot swarm reach the target at the same time.

Proper fit should be checked while wearing your hunting clothes. You should be able to bring the gun to your shoulder in one motion, without the butt catching on your clothing. When you point the gun, the heel of your thumb should be 2 to 3 inches (5 to 7.6 cm) from your nose.

Screw-in chokes enable you to change your shot pattern for different types of hunting. These chokes screw into the barrel, so they have little effect on the gun's appearance. Use a specially designed wrench to install and remove the chokes.

RECOMMENDED SHOT SIZES

Game	*Shot Sizes*
Cottontail Rabbit	*6-7½*
Snowshoe Hare	*4-6*
Squirrel	*5-7½*
Raccoon	*6-7½*
Ring-necked Pheasant	*4-6*
Quail	*7½-8*
Grouse	*6-8*
Turkey	*4-6*
Dove	*6-8*
Woodcock	*7½-9*
Small Duck	*1-5*
Medium Duck	*1-4*
Large Duck	*BB-3*
Medium Goose	*BBB-1*
Large Goose	*T-BB*

Most often used for big-game hunting, slug loads are available in two basic styles: Sabot-type and Foster-style. While slugs are available for almost all gauges, 12-gauge slugs are the most popular.

Shooting a Shotgun

It takes little skill to hit a standing target with a shotgun, but moving game is a challenge. Expert shotgunners develop their skills through practice. You must adjust for different angles, ranges and flight speeds, all within seconds.

Because a moving animal offers only a brief shot opportunity, learn to mount the gun quickly and consistently. Place the butt against your shoulder and press your cheek against the stock. Keep both eyes open and sight down the barrel with your dominant eye, which for most right-handed shooters is the right eye.

Practice mounting the gun and operating the safety at home. Wear your hunting coat and make sure that you can bring the gun to your shoulder without it catching on your clothing. Quickly draw a bead on stationary objects. Rotate your shoulders and hips as if following a moving target. Be sure your shotgun is unloaded before you practice.

Shotgunners use three basic techniques for moving game. Snap-shooting works well at close range. But for crossing targets at longer distances, use the swing-through or sustained-lead method.

To sharpen your accuracy, shoot at practice targets or clay pigeons. Practice at a shooting range or have a friend throw clay pigeons with a hand trap. Fire at crossing, overhead and straight-away targets, so you learn how to shoot at different angles.

Most hunters tend to shoot behind crossing targets and below those that are rising, especially at long ranges. If you miss consistently, double your lead after each shot until you hit your mark.

THE SWING-THROUGH TECHNIQUE

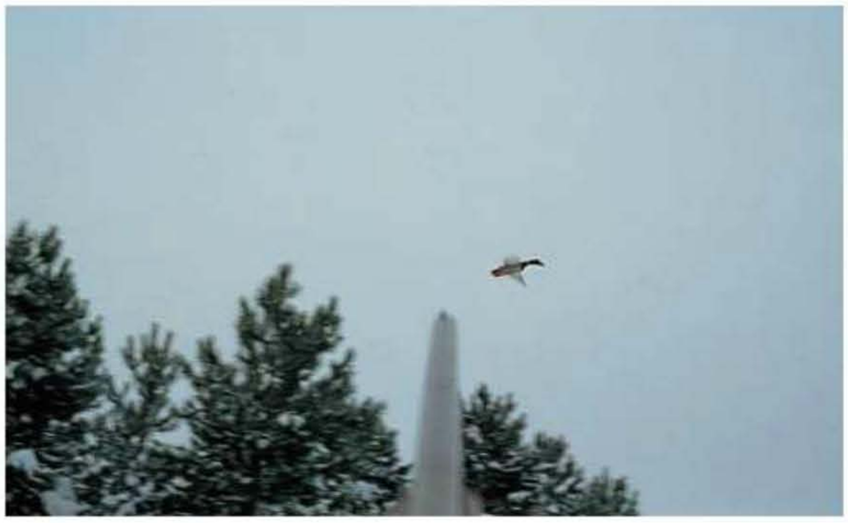

Start your swing with the barrel behind the bird. Move the barrel smoothly and steadily, so it starts to catch up with the bird's tail.

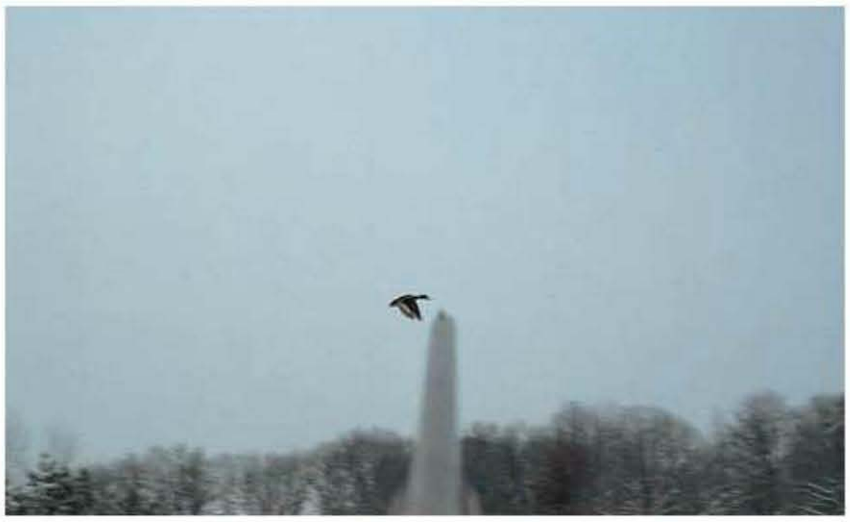

Continue swinging so the barrel moves ahead of the bird. How far ahead depends on the distance and the speed and angle of the bird's flight.

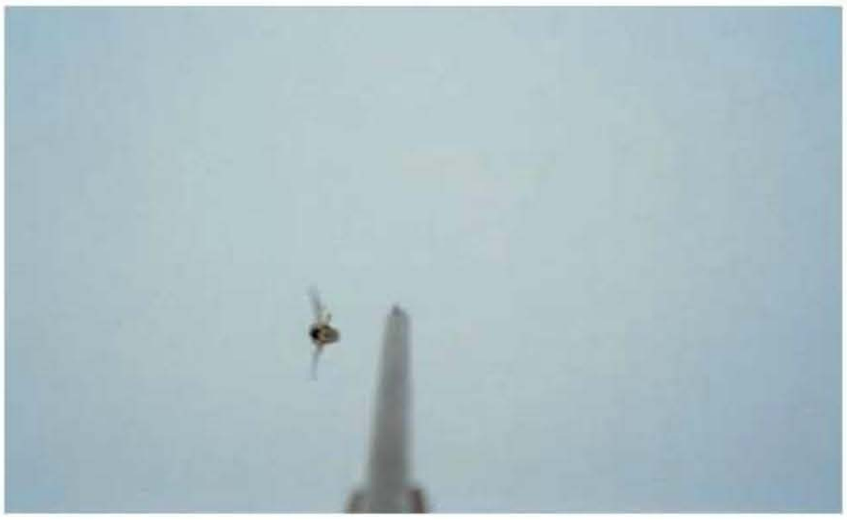

Pull the trigger when you think you have reached the proper lead. Do not hesitate, flinch or slow down your swing as you pull the trigger.

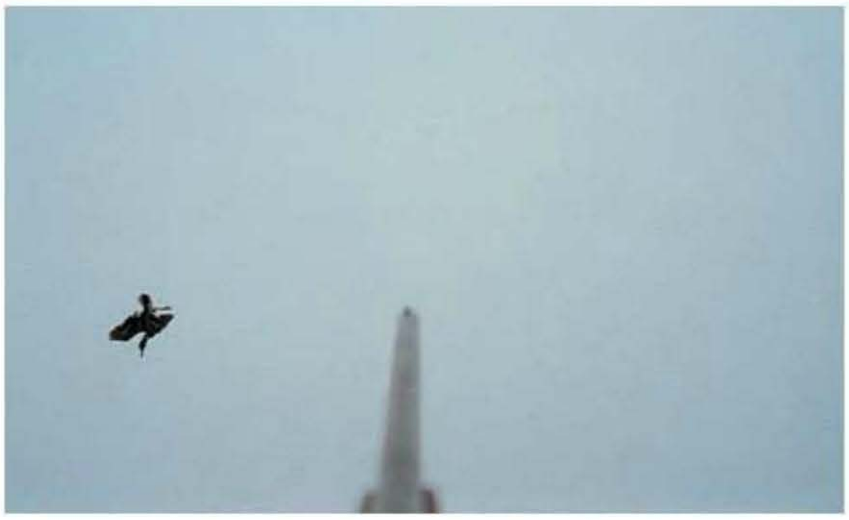

Follow through until the barrel is well past the bird. If you stop swinging too soon, you will shoot behind the target.

BOWHUNTING

Taking a big game animal with a bow and arrow is one of hunting's biggest challenges. The hunter must practice many hours before the season begins; to learn to estimate range, draw the arrow smoothly, anchor it and place each shot in a target the size of a pie plate at 10, 20, 30, 40 and 50 yards (9, 18, 27, 36 and 46 m). And the bowhunter needs to refine his or her hunting skills as well as shooting skills.

Bowhunting is popular for many reasons: Hunters using a bow are generally allowed longer hunting seasons; this form of hunting is a big challenge; it is a relaxing and peaceful form of hunting. And it is fun!

There are basically three types of bows: the compound bow, the recurve bow and the longbow. Your choice of bow type is largely a matter of personal preference, each having its advantages and disadvantages.

The most popular hunting bow is the compound, which uses cams and cables to give the bow let-off, a reduction in the amount of force needed to hold the bow at full draw. This is an advantage when holding the bow back while waiting for a shot opportunity.

The peak draw weight, or maximum amount of weight needed to draw the bow, can be adjusted on a compound bow, whereas on a longbow or recurve, the draw weight is determined by the design of the bow. For a given draw weight a compound bow produces more energy than a recurve or longbow.

To determine draw weight, start by trying different bows. Select a bow with a draw weight that you can draw very comfortably. One of the biggest problems bowhunters have is trying to hunt with bows that have excessive draw weight.

Bows with a 60-pound (27 kg) draw weight are heavy enough for all North American game animals. In truth, it is the kinetic energy—not a bow's poundage—that matters most. Kinetic energy is a measurement based on the speed and weight of the arrow used. The higher the kinetic energy of an arrow, the more penetrating power or killing ability the arrow will have.

Kinetic energy is also affected by your draw length. This is the distance between the bowstring and the grip, when you hold a bow at full draw. Do not confuse draw length with arrow length; arrows are often shorter or longer than your draw length. For a given bow type, a longer draw length produces more energy than a short draw length.

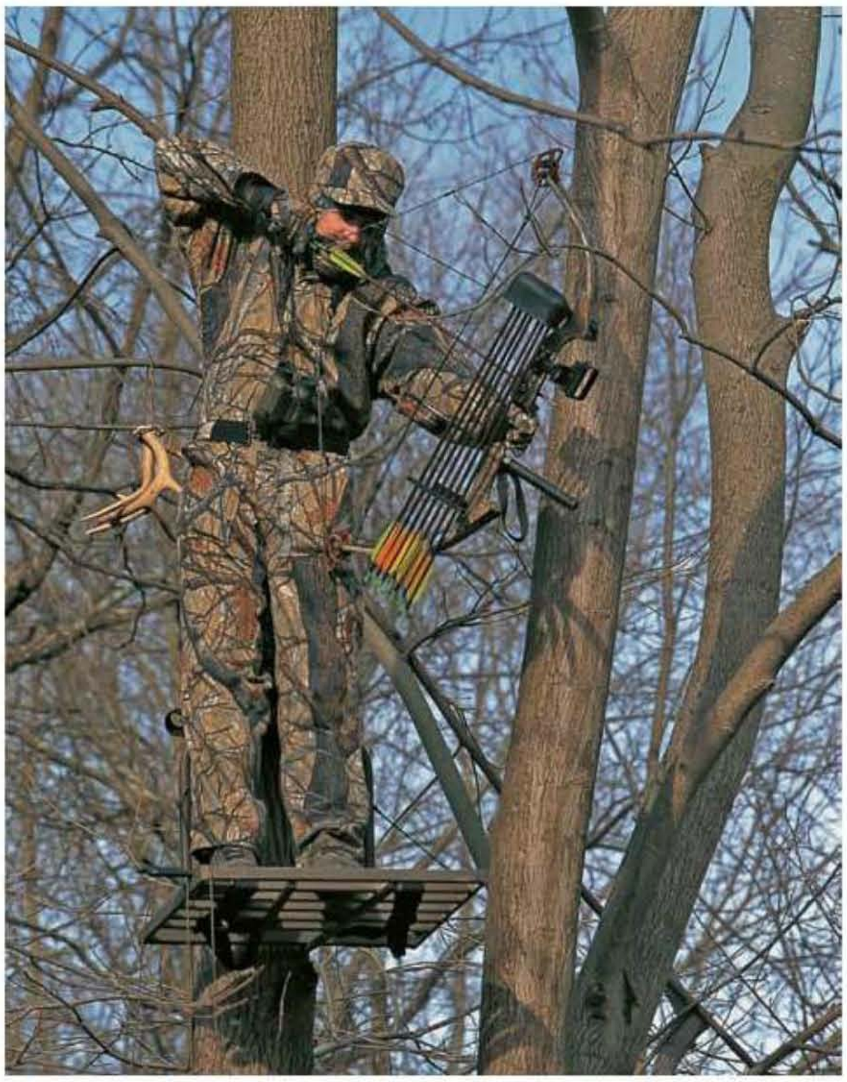

Consistent shooting practice includes knowing how the bow and arrow combination performs from different angles and at varied distances. Practice while wearing the clothing that will be worn on the hunt.

The design of the cams or wheels on compound bows also affects the bow's ability to produce energy. The more round the cam or cams are, the less arrow speed the bow is able to produce. However, they are easier to draw, more forgiving to shoot and are more quiet. Radical cams produce faster arrow speed, flatter trajectory and more energy, but they are also harder to shoot, more noisy and more difficult to maintain.

Arrow weight also affects the amount of energy that you can deliver into a target. A heavier arrow produces more energy. However, an arrow that is too heavy for a particular bow will not perform well. Consult an arrow selection chart or archery shop to help you choose the correct shaft.

Some hunters prefer to hunt with recurve and longbows. They are lighter, quieter to shoot and less prone to mechanical failure. These more traditional bows increase the challenge and add nostalgia to the hunt.

EQUIPMENT FOR BOWHUNTING

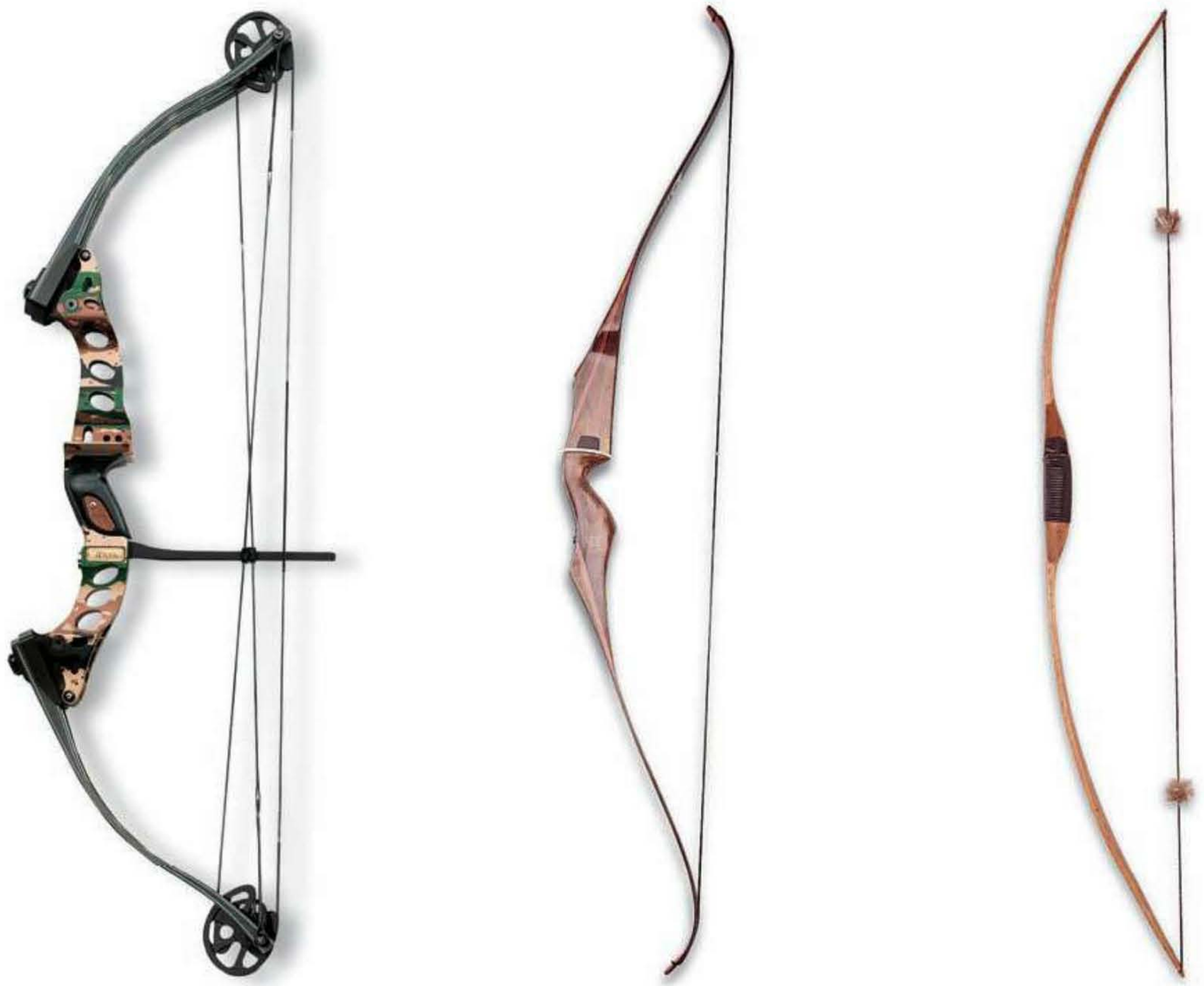

Bows include: (left) compound bows, (middle) recurve bows, (right) longbows. Compounds deliver the most energy for a given draw weight; longbows the least. A compound's limbs are made of manufactured materials. Recurves and longbows are usually made of fiberglass-wood laminates, though some longbows are solid wood.

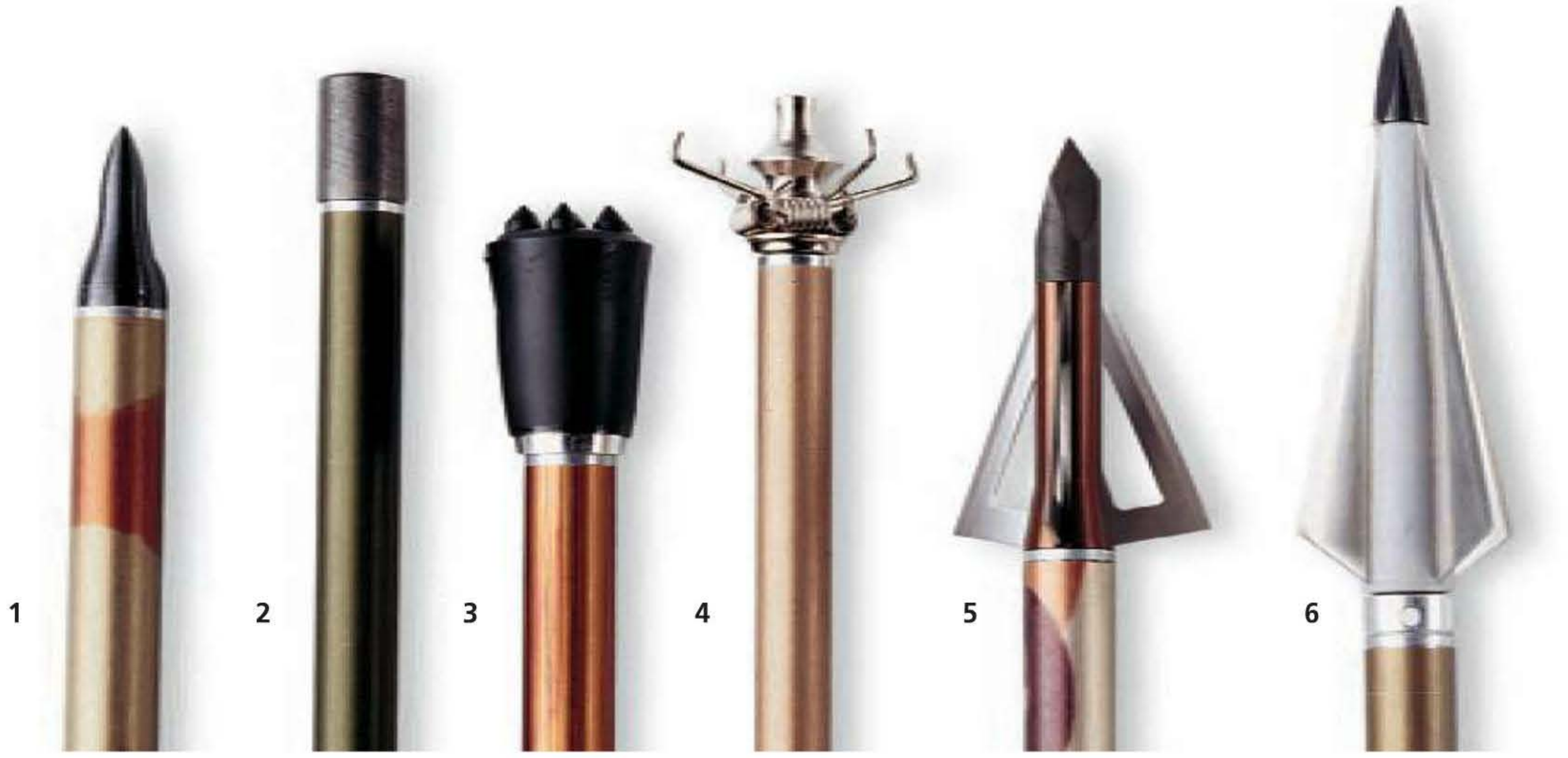

Arrowheads include: (1) field points for practice, (2) steel blunts, (3) rubber blunts, (4) judo points for small game and field practice and (5) broadheads for hunting larger game.

Shooting a Bow and Arrow

In the hands of an experienced shooter, the bow and arrow can be very accurate. But the average hunter stands little chance of killing an animal with a bow, unless he or she is willing to take the time to practice.

Shooting a bow and arrow accurately is more difficult than shooting a firearm. You must build up the muscles you use to shoot, use the correct shooting mechanics and be able to concentrate. You also need to practice and develop your ability to judge distances accurately.

The first key to good shooting is the ability to relax. Learn to let the bow shoot the arrow. If you try to force the arrow into the target you will not shoot consistently.

A good stance is critical for accurate shooting. If your upper body moves around, your sights and bow will be moving, too. It is your legs that hold it all still.

Begin by standing with your feet spread apart at shoulder width, 90 degrees to the target. Then take a half-step back with the front foot and pivot slightly toward the target for a mildly open stance. Keep your weight evenly distributed on both feet and stand straight up, with your head directly over the center of your body. Maintain this stance as you raise the bow to shoot. Don't lean forward or backward to pull the bow and don't cock your head to the side to line up your sights. In some cases, you will not be able to shoot standing straight up, but try to position yourself in a way that gives you good upper body stability.

Your bow hand is placed on the handle slightly differently depending on the type of bow that you shoot. After selecting the correct grip for your bow type, remember that your hand should stay very relaxed as you draw the bow. You must be consistent with your hand position. Even slight variations affect the arrow flight.

Your string hand, like the bow hand, should stay relaxed throughout the drawing process. There are two basic styles that hunters use to hold and release bowstrings. The more traditional method is to use your fingers. If you release with your fingers, start by grasping the string at the first joints of your first three fingers, with the index finger above the nock, the other two fingers below the nock. As you draw, the middle finger should hold most of the weight and the other two fingers should float on the string. The other method for releasing the bow involves using a mechanical release aid. Choose one that has a rotating head, so it does not torque the string as you draw. With a wrist-strap release, you should pull only on the strap and your fingers should remain loose. With a finger-held release, your wrist should stay straight and relaxed.

When you draw a bow, hold it at arm's length, roughly aiming at the target and begin to draw, pulling only with the muscles of your back. When you reach full draw, anchor solidly and aim at the target.

The anchor point for finger shooters is typically fairly high, with the tip of the index or middle finger planted solidly in the corner of the mouth. There is not necessarily a best way to anchor when using a release aid. With a wrist-strap release, many hunters anchor with the big knuckle of the index finger pressed behind the jaw. With a finger-held release aid, experienced shooters commonly anchor with the back of the hand pressed against the jaw.

No matter what type of equipment you use or the release method you use, the release should happen through complete relaxation. Focus your attention on pulling the bowstring with your back muscles. Release shooters pull the release trigger as their back muscles tighten up; finger shooters should let the string slip away as their hand relaxes. The shot should happen as a surprise.

Follow through a shot by holding your bow arm and release hand in the same position they were in when you released, until the arrow hits the target.

Always practice from the positions in which you hunt. If you use a tree stand, practice from an elevated platform. If you sit in your stand, practice in the sitting position. Wear your hunting clothes when practicing; they may affect your shooting. Use an arm guard to prevent the string from catching on heavy clothes.

Foot position and weight distribution are a critical part of a consistent shooting routine. Relax, spread feet to shoulder width, and orient the body toward the target in an open stance. The stance, the draw, and the release will become automatic with regular practice.

PROPER GRIP FOR SHOOTING COMPOUND, RECURVE AND LONGBOW

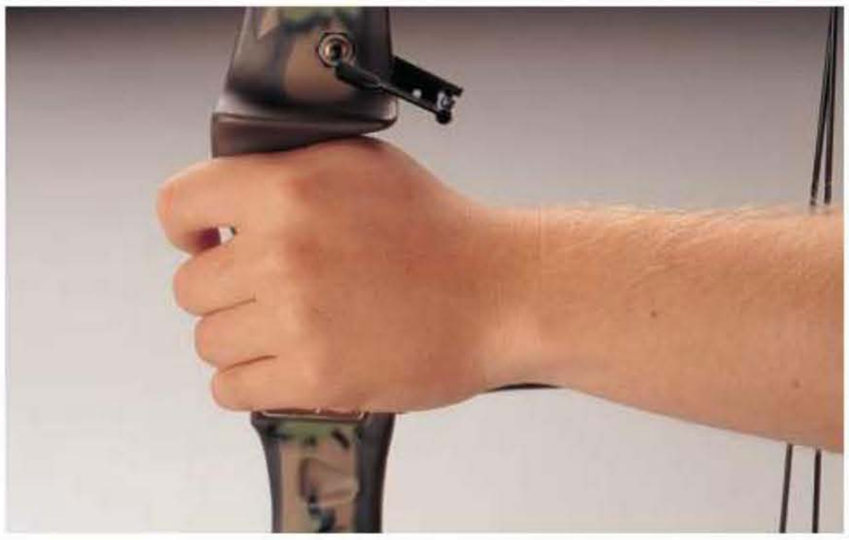

Low wrist position places the meaty party of the thumb on the bow handle. Maximum pressure falls about 1 inch (2.5 cm) below the big joint of the thumb. In this position, the wrist is more stable and less likely to move side to side.

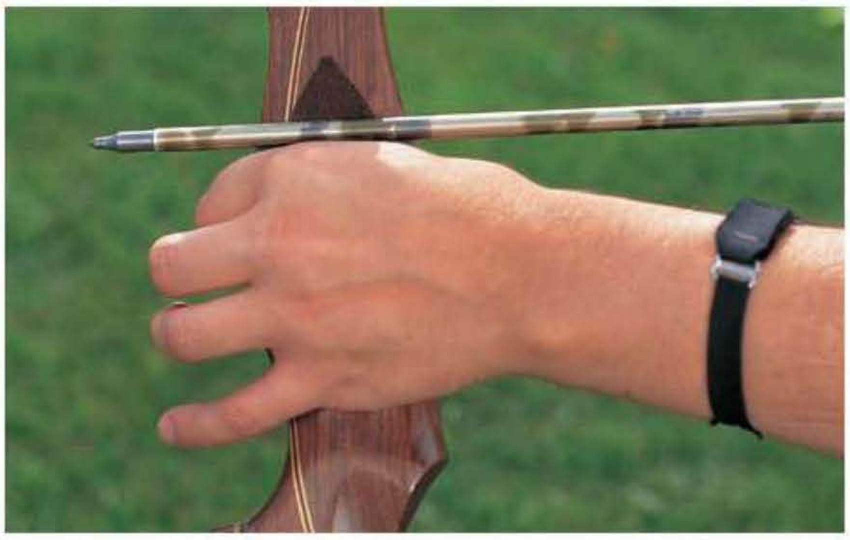

Straight wrist grip is used with the recurve bow. At full draw, all pressure is between the thumb and forefinger.

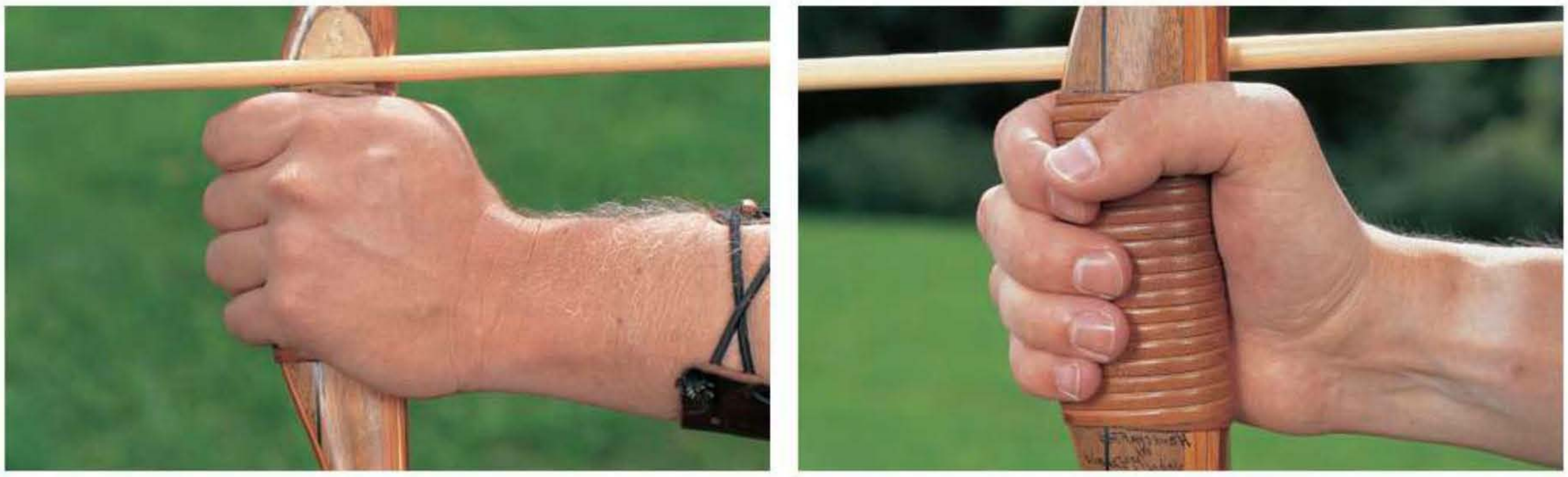

Low wrist grip (left) is standard form when shooting a longbow. Pressure should fall on the back edge of the heel of the hand (right). Gripping the bow with too much of the heel on the handle can cause the bow to twist during relase, leading to poor arrow flight.

TYPICAL ANCHOR POINTS

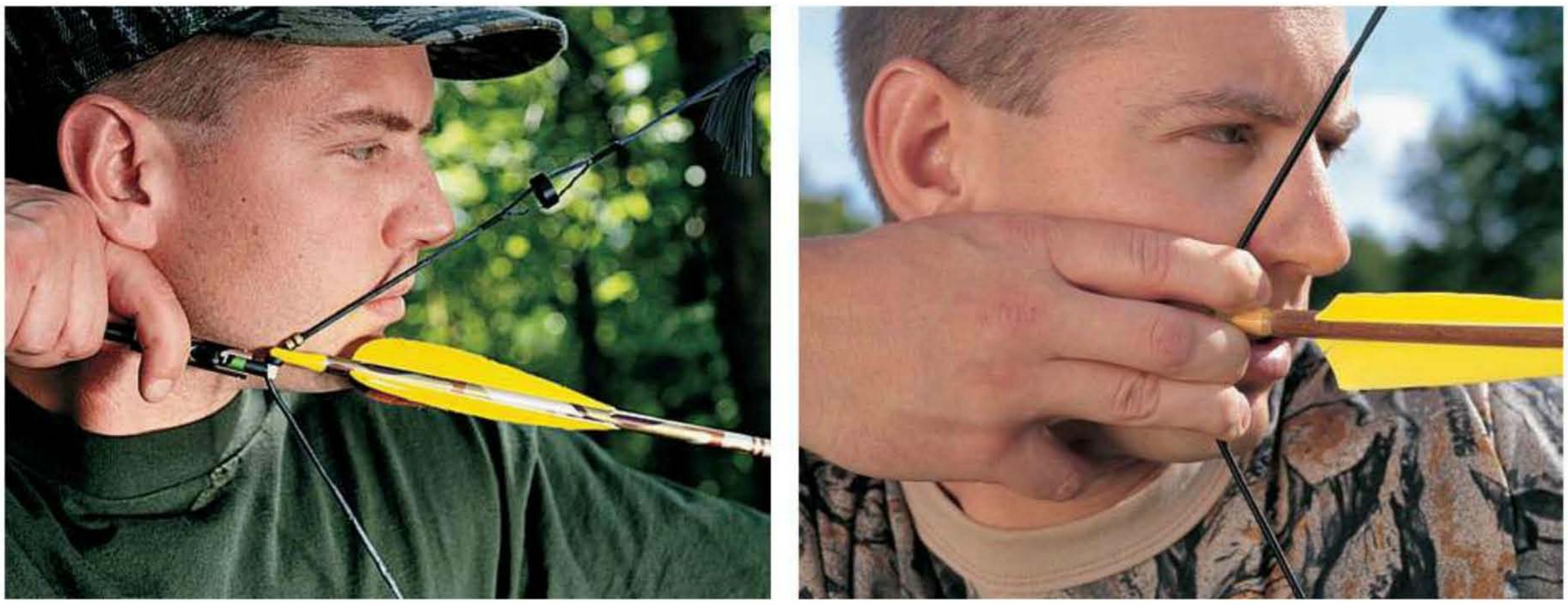

Using a wrist-strap release aid (left), many archers anchor with the big knuckle behind the jaw. With a finger-held release (right), most archers anchor so the tip of the index or middle finger is at the corner of the mouth.

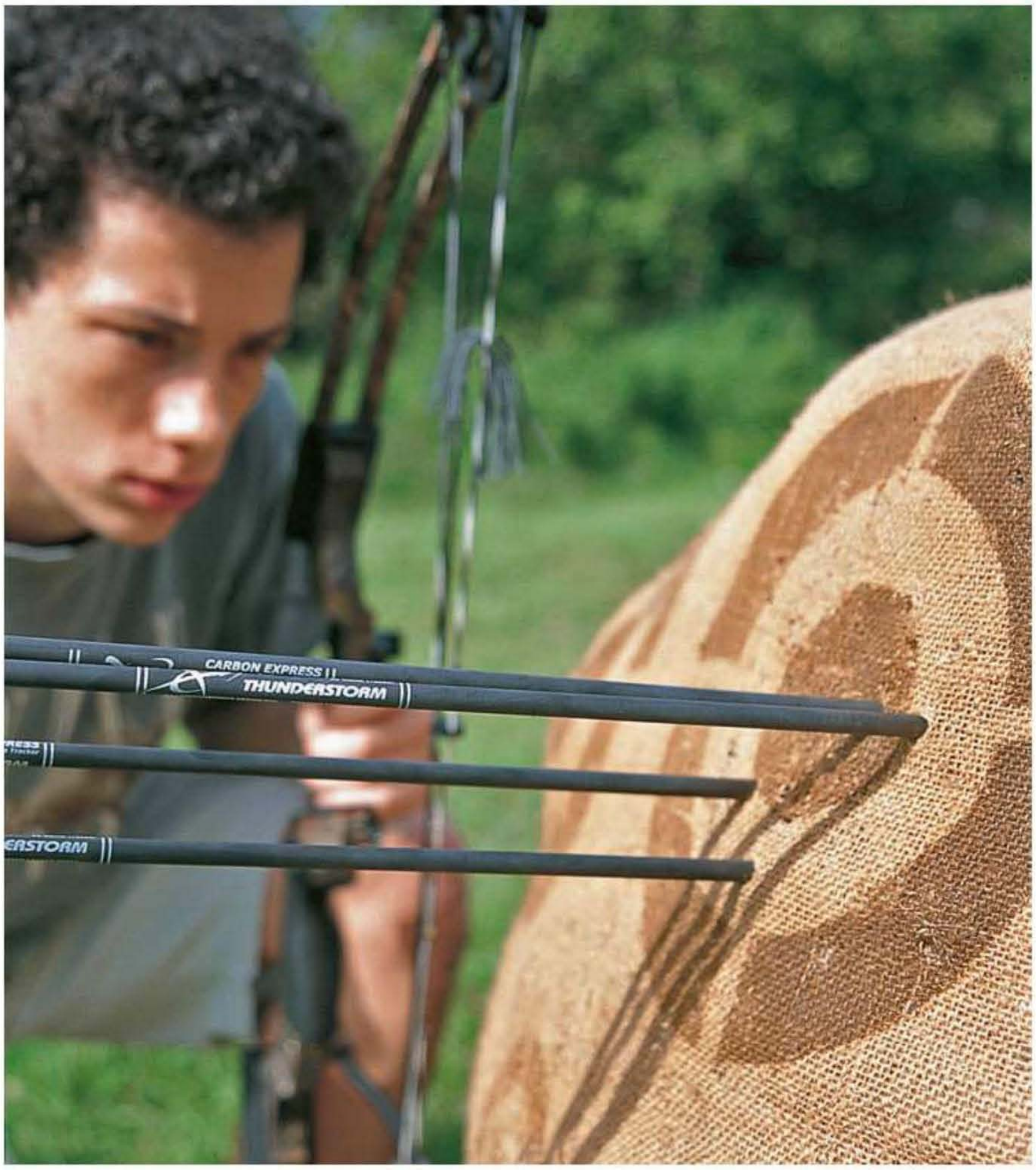
CARBON EXPRESS
THUNDERSTORM
ERSTORM

Practice is critical to ensuring a clean kill. Skills honed on the range translate well to the woods.

BOW CARE TIPS

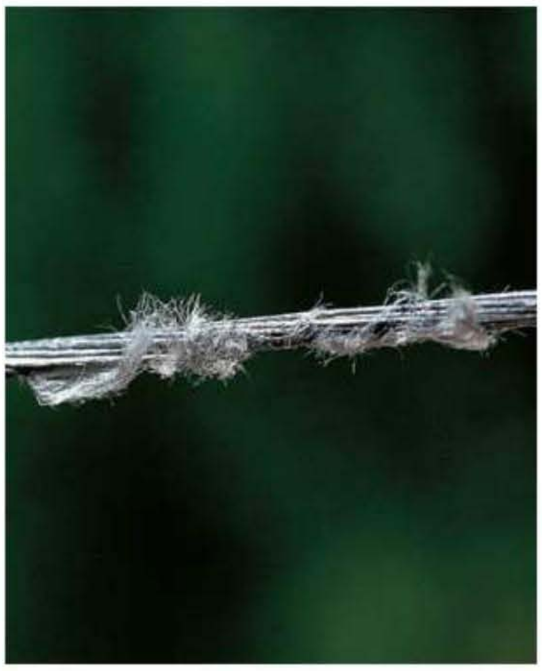

Inspect the bowstring before shooting; frayed strings should be replaced.

Oil axles every few days when in the field.

Keep your bow away from heat. Laminated bow limbs can come apart after only a few hours in a hot vehicle.

MUZZLELOADERS

Hunters who accept the challenge of making one shot count can hunt additional seasons when they start carrying a muzzleloader. Most states have special muzzleloader regulations that extend big-game hunting opportunities beyond normal gun seasons. Muzzleloaders or black powder guns must be loaded by inserting powder and bullet or shot into the muzzle, then pressing the load down the barrel toward the breech. Reloading after a shot can take a minute or more, requiring the first shot to be a good one.

Hunters have a choice between muzzleloading shotguns and rifles. Each one of these is available in different ignition systems. The three most popular include: in-line percussion caplock, percussion caplock and flintlock. Flintlock, the oldest ignition system, has never been very reliable for the hunter. The percussion caplock, developed in 1820, increased the muzzleloader's reliability. It wasn't until the mid-1980s that the first in-line muzzleloaders were developed. These frontloaders became a very reliable hunting weapon. The in-line design allows the percussion cap to be located very close to the powder charge, promoting quicker and more positive ignition.

The most popular muzzleloading rifles for big game are 50 or 54 caliber. The projectile may be a patched round ball or a conical or a saboted bullet. An in-line rifle firing a saboted bullet can achieve good accuracy at 100 yards (91 m), comparable to modern center-fires and have a killing range on whitetail-size animals out to 150 yards (137 m). The practical accuracy range of a patched round ball is approximately 80 yards (73 m); a conical, 100 yards. A modern in-line muzzleloader properly loaded with a heavy, well-constructed bullet is up to the task of downing any North American game animal.

A hunter today has a choice of types of powder to use in his gun. He or she can use the centuries-old black powder or a synthetic black powder called Pyrodex. Either type performs well and can be tested to see which one performs best.

Muzzleloading shotguns work well on waterfowl, upland game birds and turkeys. Small caliber muzzleloading rifles are used for small game and varmints. Be sure to check the regulations in the area you plan to hunt when using these muzzleloaders.

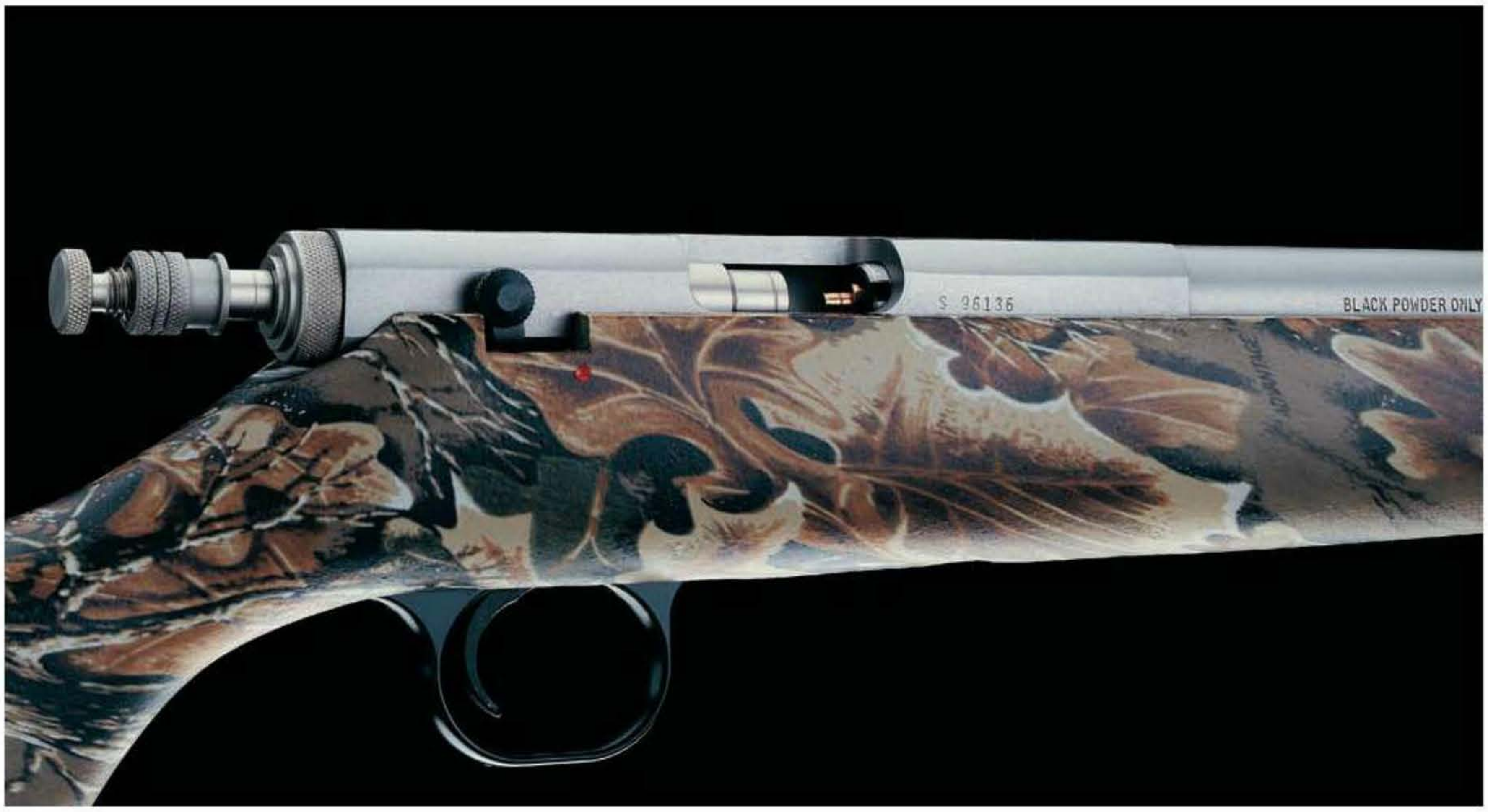

In-line percussion caplocks have an internal striker or hammer that is cocked by pulling it rearward. When the trigger is pulled, the striker slams forward and hits the percussion cap, which is positioned on a nipple. The nipple is positioned directly behind the powder charge, so the fire from the cap reaches it instantly.

OTHER COMMON IGNITION SYSTEMS

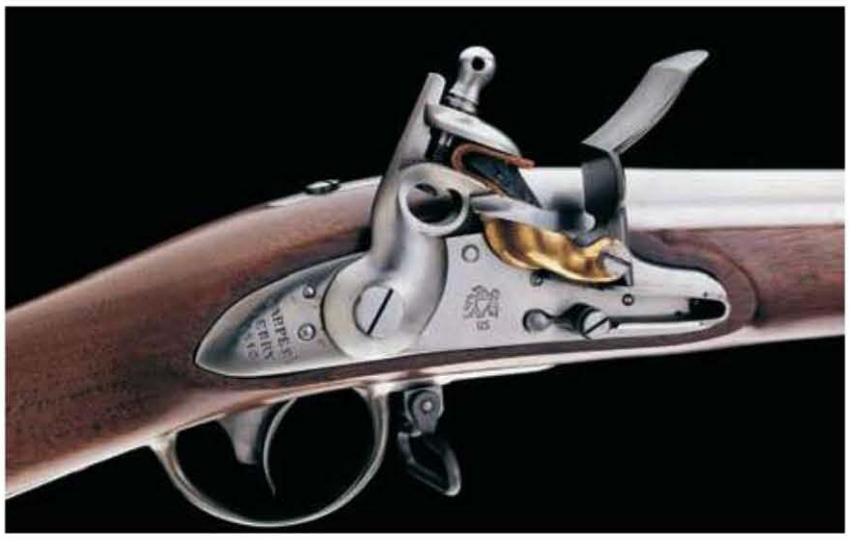

On a percussion caplock, the hammer strikes a percussion cap, similar to the primer of a modern cartridge. A flash travels through the nipple to ignite the powder charge.

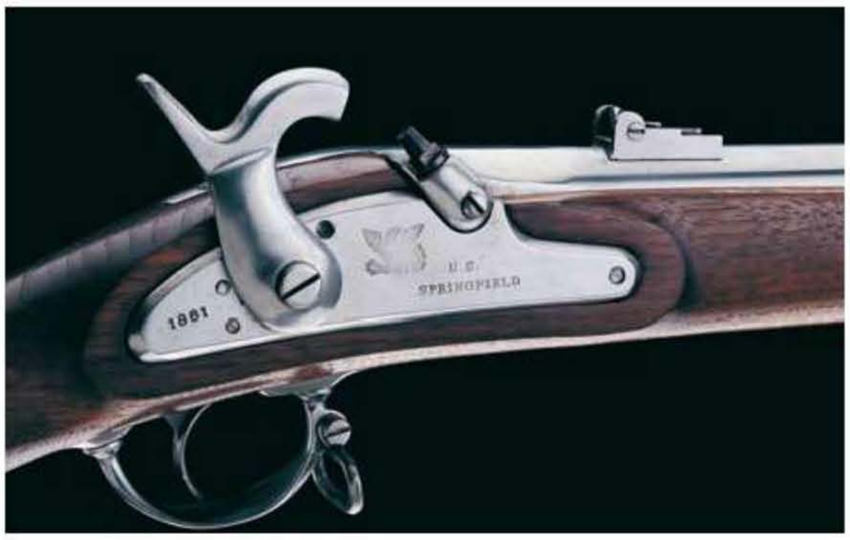

Flintlocks have a flint in the hammer. When the flint strikes a frizzen, sparks ignite the primer powder. Flame passes through a vent hole to set off the powder charge.

PROJECTILES

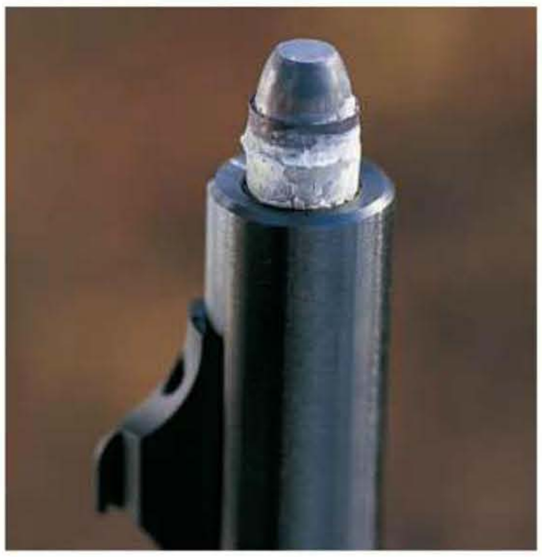

Plastic sabots grip the bullet securely as they are pushed into the muzzle. At the same time, the lands of the rifling groove the sabot. When the gun is fired the sabot transfers the spin of the rifling to the bullet.

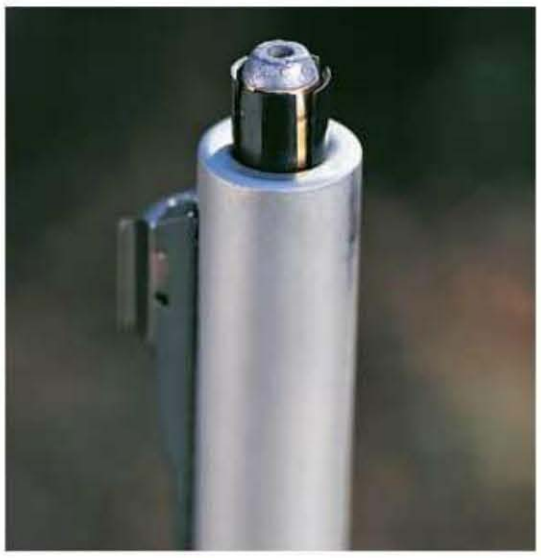

Conicals fit tightly into the barrel without the aid of patching materials. As a conical is pushed down the barrel, the lands of the rifling engrave the soft lead, forming a tight seal.

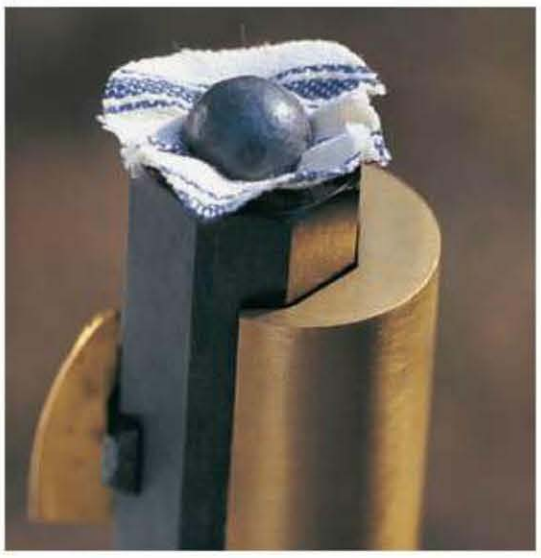

Patched round balls must fit correctly in the barrel to achieve accuracy. The ball never touches the inside of the barrel. The patch grips both the soft lead ball and the grooves of the rifling.

BULLET TRAJECTORY

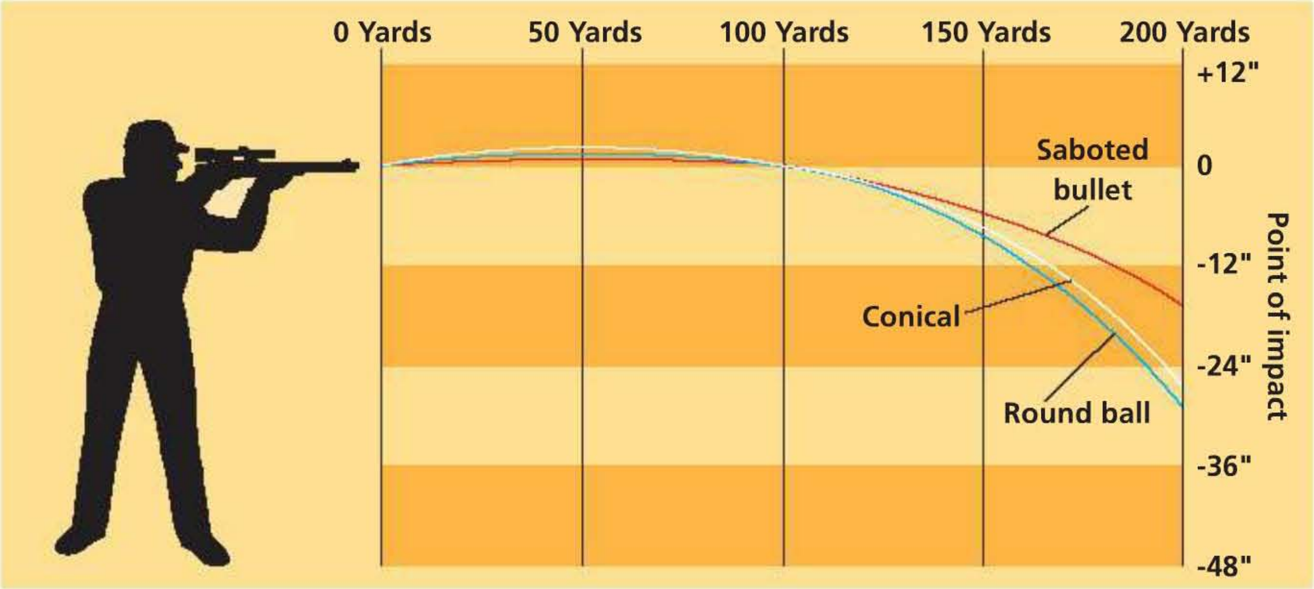

Bullet trajectory for saboted bullets is flatter than that of conicals and round balls. As the illustration shows, a saboted bullet sighted in at 100 yards (91 m) hits only 1½ inches (3.8 cm) high at 50 yards (46 m) and 6 inches (15 cm) low at 150 yards (137 m).

HUNTING SAFETY

Most people would be surprised to learn that hunting is not among the most dangerous sports. A National Safety Council study showed the fatality rate for hunting to be less than half that of boating or swimming.

Improved hunter education and increased use of fluorescent orange clothing account for the low accident rate. Every state and province sponsors some type of firearms safety or hunter education program. Many states require beginning hunters to pass such courses before they can purchase a license.

Nevertheless, the potential for a serious accident always exists. To avoid an accident, follow these safety rules.

- Treat every firearm as if it is loaded. Never assume a gun is unloaded because someone said so.
- Never point a weapon at anything you do not mean to shoot. This includes glassing other hunters with your rifle scope.
- Make sure your safety is on at all times, unless you intend to shoot.
- Positively identify your target before shooting. Never fire at a silhouette, a vague form or an area where you saw or heard something move. Fluorescent orange clothing greatly improves your own visibility.
- Control the direction of your muzzle at all times. If you start to fall, point the barrel away from yourself and other hunters. After a fall, check the barrel for obstructions like dirt or snow. A plugged barrel could rupture when you shoot, possibly causing serious injury.
- Never lean a gun against a tree, fencepost, vehicle or any place where it could fall over and accidentally discharge.
- Never shoot at hard surfaces or water with bullets or slugs. They could ricochet and strike another hunter or a building.
- Never drink alcoholic beverages before or during a hunt. Alcohol does not keep you warm; instead, it speeds the loss of body heat.
- Use only the ammunition recommended for your firearm. Do not carry two different types of ammunition in your pocket at the same time.
- When not hunting, keep the gun unloaded and the action open.
- Keep all firearms and ammunition out of the reach of children.
- Refuse to hunt with anyone who does not observe the basic rules of firearms safety.

Keep the safety on and your finger away from the trigger until you are ready to shoot. This eliminates the possibility of discharging the gun accidentally, especially if you stumble or fall.

PLANNING FIRING ZONES

Establish safe firing zones when hunting with others. When driving game, hunters should shoot only within their zone area.

Position your boat sideways to the shooting area. Do not position the boat lengthwise; this places one hunter in the other's shooting zone.

Split the area around your blind into 180-degree shooting zones. Each hunter scans for birds in the semicircle area around his or her half of the blind.

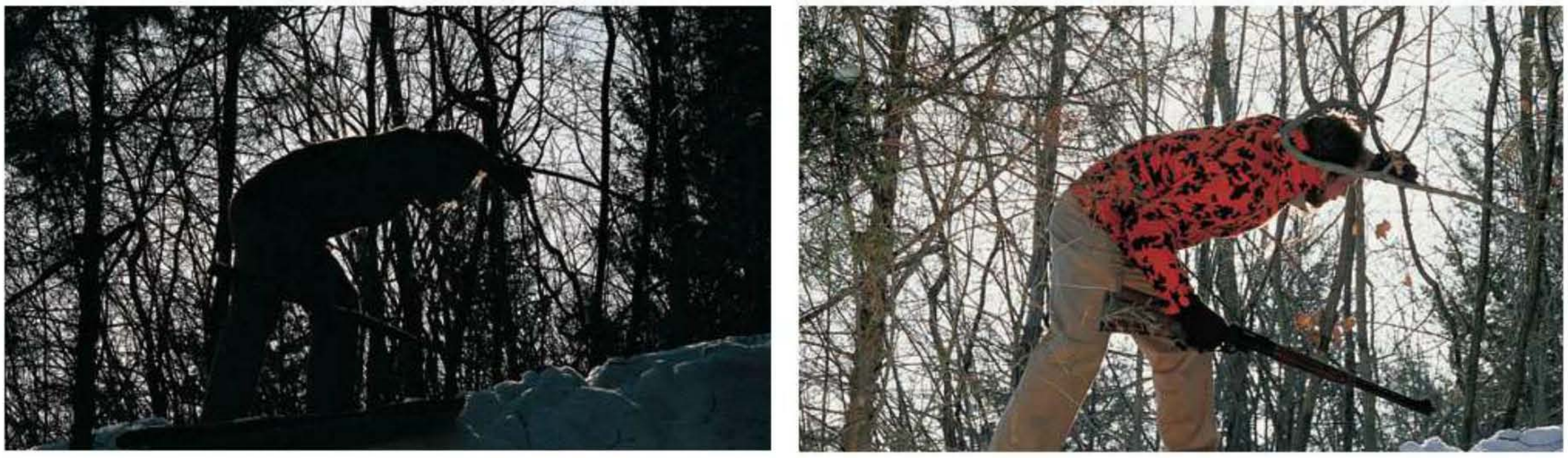

Never shoot unless you can see your target clearly. In dim light you may see a form that resembles a game animal (left). The right photo reveals that the form is actually that of a hunter.

OTHER SAFETY TIPS

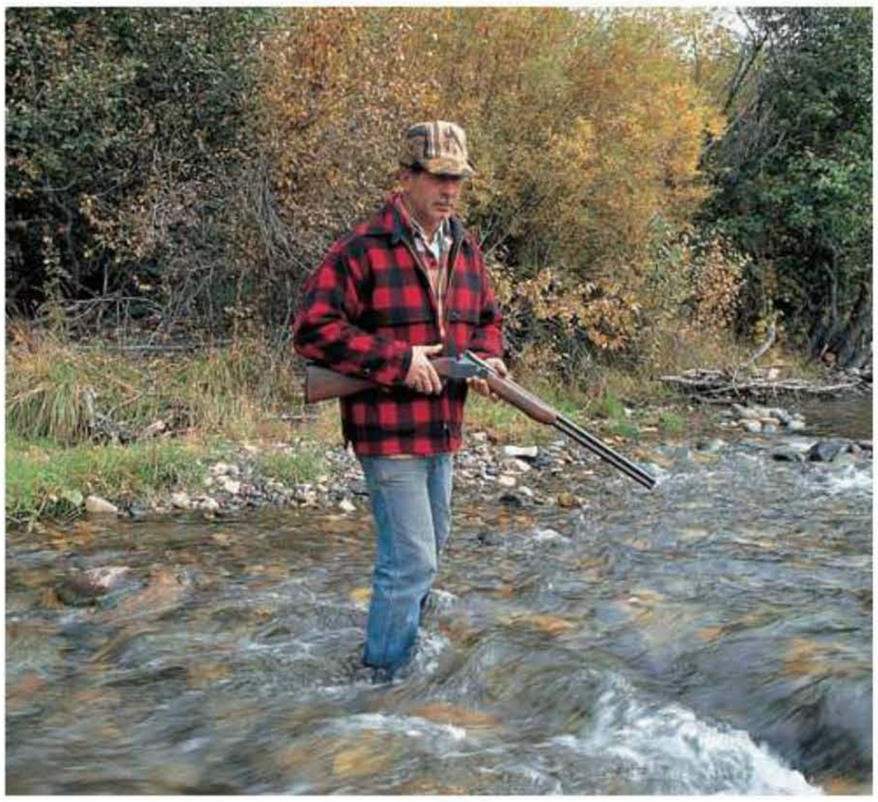

Open your action before crossing a creek (left), climbing over a fence or in any other situation where you are unsure of your footing.

Never shoot if there is a chance of hitting buildings, livestock or any other unintended target. Avoid shooting over the tops of hills and ridges.

Carry your firearm so the muzzle points away from others. Safe positions include: shoulder carry, cradle carry and trail carry.

Chapter 3

HUNTING STRATEGIES

Whether you travel to another state or hunt the back forty, advance planning is the key to a good hunting trip. Lay the groundwork for your trip early.

Preparations for an out-of-state hunt should begin at least a year ahead. Some states conduct drawings for non-resident permits to hunt animals like elk, deer, pronghorn or moose. Most states accept applications until spring, but some set deadlines as early as mid-winter.

If you do plenty of scouting, it will pay off by helping you choose the right hunting method for the game and the area you hunt.

Begin planning your trip by requesting non-resident hunting information from the wildlife agency in the state or province where you plan to hunt. Be sure to specify the type of game you will be hunting. Ask about drawing application deadlines and information on state and federal wildlife areas that may require advance reservations. Licenses may be offered on a first-come, first-served basis. Some states require hunters to show evidence of having passed a firearms safety or hunter education course.

Reservations and permits may also be necessary when hunting in your own state. Wildlife managers and conservation officers can provide information on special hunts that require advance preparation. They can also supply current game census data to help you choose a hunting area. Managers may know of local farmers or ranchers who want hunters to thin over-abundant game populations.

Sportsmen's clubs like the Safari Club International, Rocky Mountain Elk Foundation, Mule Deer Foundation, and state hunting organizations are great places where hunters can network and share information. Some clubs lease private land where members can hunt.

SCOUTING FOR GAME

Pre-season scouting improves your odds of finding game once the hunting season begins. Best of all, scouting is enjoyable and saves valuable hunting time.

Scouting enables you to become familiar with the terrain and to identify heavily used game trails. It also helps determine any changes in habitat conditions or fluctuations in game populations that could affect the hunting.

Severe winter storms, wildlife food shortages, drought and cold or wet weather during nesting can drastically reduce game populations. New roads, housing developments and wetland drainage can permanently eliminate good wildlife habitat. So, there is no guarantee that last year's prime hunting spot will produce game this year.

Binoculars and spotting scopes help to locate game from a distant vantage point. But even if you do not see animals, you can detect their presence by the signs they leave. For instance, squirrels build nests in trees and scatter nut shells on the ground. Waterfowl preen themselves and leave feathers around resting areas. Bull elk bugle and roll in wallows during the mating period. All animals can be identified from their tracks and droppings. Damaged trees and brush result from animals rubbing velvet from their antlers, scratching trees to mark their territory or chewing bark.

Glass a potential hunting area from a nearby hill or tree. Look closely at edges between openings and cover to detect movement. You are most likely to see game early in the morning, at dusk or on cloudy days. When you spot an animal, note the time and identify a landmark to pinpoint the exact location. If the animal is not disturbed, it may appear in the same area at about the same time on subsequent days.

PLANNING TIPS

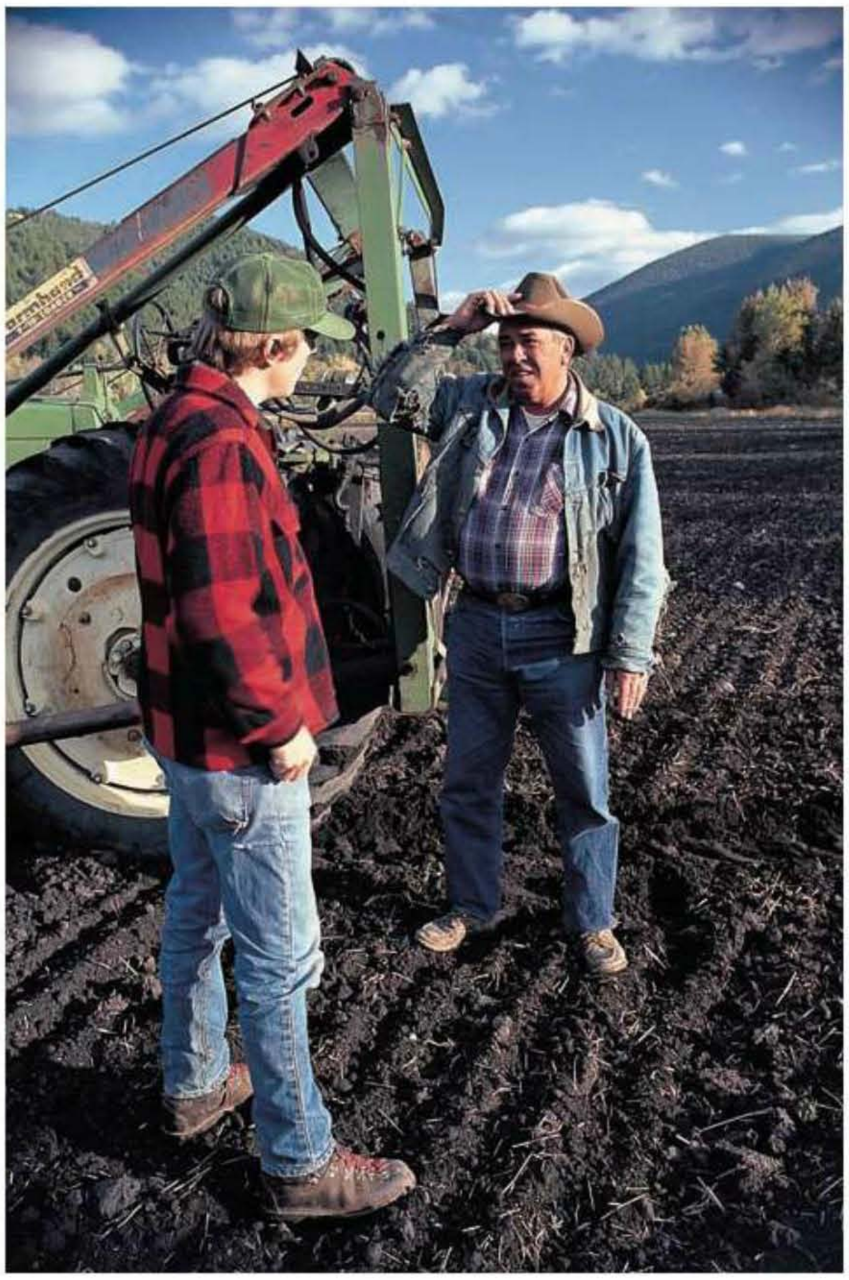

Always ask for permission to hunt on private land. Introduce yourself and specify the type of game you wish to hunt. Some landowners grant permission to a courteous hunter, even though their land is posted.

Talk to a clerk at a sporting goods store for local hunting information. Most carry maps, permit applications and other items for planning your hunt and many employ knowledgeable hunters.

Gather information from state and federal wildlife agencies. They supply maps, brochures on public hunting areas, public access lists and regulations. Tourist bureaus can recommend motels, resorts and campgrounds.

County maps identify paved and unpaved roads. They also show section lines, buildings, public lands, lakes and streams. Plat books detail property lines and identify landowners. The maps and plat books are usually available at county offices.

Aerial photographs reveal important details lacking on maps. They show isolated ponds and marshes, small streams, vegetation types, logging areas, meadows and forest trails. They can be obtained from the U.S. Department of Agriculture and private survey firms.

Topographic maps, like the U.S. Forest Service map and the U.S. Geological Survey map, provide information on land elevation and detail other features of the landscape, such as wetlands. They also show buildings, roads, trails, portages and other features useful for finding your way on a hunting trip. Geological Survey maps show forested areas in green. They generally cover a larger area than Forest Service maps.

SIGNS OF GAME ACTIVITY

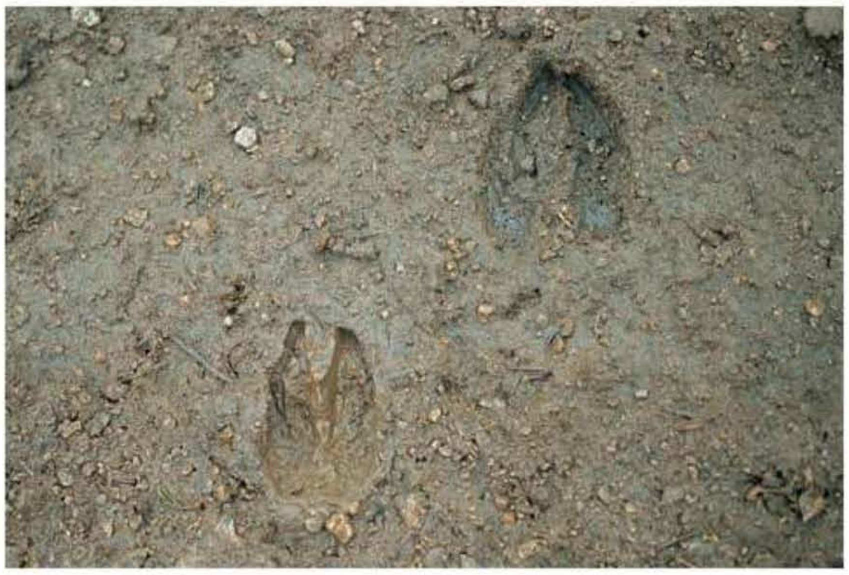

Tracks in damp soil may reveal how long ago an animal passed. Fresh tracks (left) have sharp edges. Older tracks (right) have fuzzy edges. In snow, old tracks have a hard crust.

Droppings help hunters find game. Large quantities indicate feeding, loafing or bedding areas. Some animals do not digest their food completely, so you can check droppings to determine what they have been eating.

Feeding signs include damaged crops; scratched earth; nipped ends on grasses or twigs; and pits, husks and shells of fruits or nuts. Experienced hunters can often identify an animal by these signs.

Damaged trees and brush result fom animals rubbing velvet from their antlers, scratching trees to mark their territory, or chewing bark. Each species has distinctive rubs, scrapes and chews.

Beds and roosts in tall grass or on soft earth reveal the resting spots of mammals and birds. You can identify the animal from the size of its bed or roost and from nearby droppings, hair or feathers.

Game trails lead from cover to food or water. Several kinds of game often use the same trail. Inspect it closely and look for fresh tracks made by the animals you are hunting.

STILL-HUNTING

More people blow it while they're still-hunting than with any other method. The problem is that we move too fast. It's called still-hunting because most of the time the hunter should remain still, processing every sound, every movement, every smell.

Take a step, look and listen. Wait 30 seconds, 60 seconds or 5 minutes before taking the next step. Don't make a sound. Remember, you're in country you scouted. You know the game is close by. The success of the hunt hinges on the confidence you've placed in this piece of real estate.

Hunt into the wind. If the air movement shifts direction, change with it, keeping the breeze in your face. Move too fast and deer see you before you spot them. Slow down. Slow way down. Every time you take a step, a new window opens in the habitat. Look for the horizontal line of a back, the black of a nose, the flick of a tail, the crook of a leg or sunlight glinting from a nut-brown antler. Carry binoculars on a harness around the shoulders, not on the belt or in a daypack. And use them more than your boots.

In heavy cover, crouch down to look beneath the bushes. Sometimes that affords a better view through the stalks and the trunks. Maybe you'll spot a deer standing, listening to make sense of the sound you made when last you took a step. Sometimes you'll spot a deer looking back.

Place each footstep carefully to prevent snapping twigs, crunching leaves or brushing against branches. Avoid crossing open areas and places where animals could see your form above the horizon. Try to stay near cover where game will not notice your movements.

Still-hunting is usually a one-hunter method, but it also works well with two patient hunters. Game surprised by one person may run toward the other. Or an animal may be so intent on watching one human that it does not detect another approaching.

Successful still-hunting demands patience and confidence. If you lose patience and begin moving too fast, animals will spook. If you are confident of seeing game, it is easier to stay focused.

TIPS FOR STILL-HUNTING

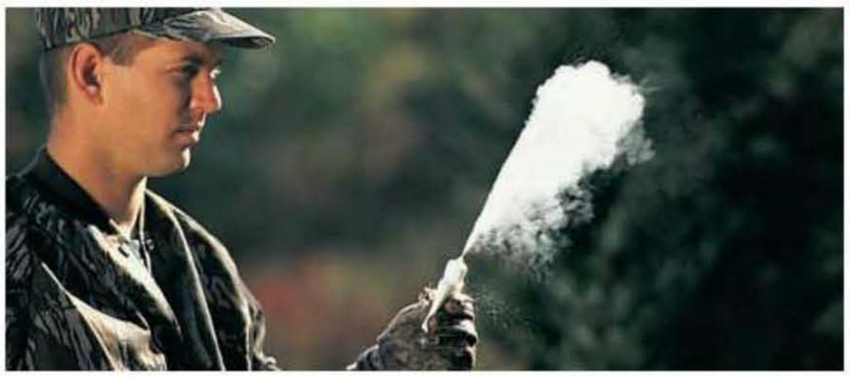

Walk slowly and quietly (top). Avoid turning your head quickly, swinging your arms or making other movements that animals associate with humans.

Detect wind direction and rising and falling thermals (bottom) by squeezing unscented powder into the air or by tossing a small amount of down into the air.

Wear soft clothing to reduce noise. Hard-finish materials like nylon make swishing sounds as they brush the cover. Gummy boot soles enable you to walk quietly.

Look for visual clues to find game. A horizontal line among trees could be an animal's back. A glinting eye or twitching ear could also reveal game.

Avoid walking or waiting in direct sunlight. Animals are quick to notice the glare off your face and clothing. They are less likely to see a hunter in the shade.

STAND-HUNTING

To hunt from a stand, you must be positioned where you are likely to see game, but the game is not likely to see you. Concealment, awareness and confidence are the keys to success.

Hunters conceal themselves on stands or in blinds. A stand may be nothing more than a large tree that obscures your form. Or it may be an elevated platform, either free-standing or attached to a tree trunk or limb. A blind provides more cover. Many have walls of camouflage material or vegetation.

Whether hunting from a blind or stand, choose a spot where game is likely to pass. Take a scouting trip before the season to find first the water, then the feed, then the bedding ground. A deer will not make its living more than ½ mile (0.8 km) from its water source. Scout the area to find heavily-used game trails or escape routes. Take position well before game begins moving. You may have to find your spot in the dark.

Place a stand to take advantage of the view into well-used trails that lead from bedding areas to feeding areas. Or set up on the approach from bedding cover to water.

Take care to position the stand downwind from the trail and if the wind changes direction, find a different place to hunt until conditions improve. Select alternative locations for different wind conditions. To ensure that animals do not detect human scent, many hunters use odor eliminating products or masking scents.

Wind direction is also important when hunting ducks or geese on water. Choose a blind on the lee side of natural cover. Waterfowl usually land into the wind in the calmer water.

Stand-hunters sometimes wear camouflage outfits to reduce their visibility. But many types of hunting require high-visibility clothing for safety purposes. Even if you wear fluorescent orange clothing, game is less likely to see you if you keep motion to a minimum. If you must move, do so very slowly. Keep your face hidden; look at game from the side of your eye or from behind a hat brim.

Comfort is important when stand-hunting. You cannot remain quiet and motionless if you are cold, wet or in an uncomfortable position. You need warm clothing, because you must remain stationary for long periods. Some hunters build roofs to shed rain or use padded seats and stoves. Waterfowl hunters sometimes build blinds that are completely enclosed except for shooting windows.

Stand-hunting works best early or late in the day, when animals move between resting and feeding areas.

HOW TO SELECT A STAND-HUNTING LOCATION

Portable stands enable you to change location quickly. Some types clamp to the trunk in seconds. They have a comfortable seat and, when placed well above ground level, offer a wide field of vision. An elevated stand keeps your scent above the ground and places you higher than the usual sight plane of the game.

Locate your stand or blind near a watering site or stream crossing. Dove and antelope hunters often hunt near water holes. Moose hunters carefully check stream banks for heavily used trails leading to water.

Choose a site near a feeding area. Goose hunters build blinds or dig pits in harvest fields. Deer hunters take stands along the edges of corn, hay or milo fields. Bear hunters sometimes bait an area, then hunt nearby.

Select a stand site in a location where you are not silhouetted against the skyline. Game will detect your movement more easily if you do not have a backdrop.

Avoid stands that do not offer a clear field of fire. A tree or branch too close to your stand will narrow your shooting zone. A stand that restricts hip movement makes it difficult to swing your gun.

STALKING

To stalk an animal, you must see it first. Binoculars or spotting scopes help hunters to locate the quarry from far away and plan a strategy. A spotter who remains in a fixed location can signal to the stalker and direct the hunt from afar.

Like still-hunting, stalking demands fine-tuned hunting skills, because you must close the distance on game without being detected. Sneaking within shooting range can be extremely difficult, but expert stalkers may be able to approach within a few feet (m) of game.

When you spot an animal, watch it to determine whether it is likely to remain in the same area long enough for a stalk. If the animal is moving or appears nervous, you will probably not be able to get close enough for a shot. The odds are much better when the animal is feeding or is bedded.

Stalking works best where the topography or cover conceals the approach. Before beginning the stalk, plot a course that takes you into the wind, but keeps you behind hills, trees, fence lines or other natural or man-made features.

If there is no cover, use a clump of vegetation to break up your outline. Camouflage clothing makes you less visible. Move only when the animal faces away from you or when its head is down. Be sure that sunlight does not reflect off your gun, scope or wristwatch.

Move silently during the final stages of the stalk. Noise may not be important if you shoot from 300 yards (274 m). But for the bowhunter who must approach within 30 yards (27 m), the slightest noise can ruin his or her chances.

It is usually best to stay concealed until after the shot. But waterfowl hunters sometimes rush the birds before they shoot. They may gain an extra 5 to 10 yards (4.5 to 9 m) before the birds take off.

Plan your stalk so you move directly into or quarter into the wind. This prevents the wind from carrying your scent and the sound of your approach to the animal.

HOW TO STALK GAME

Pinpoint game by identifying nearby landmarks like a tree, boulder or fencepost. You can stay hidden while using the landmark to guide your approach.

Use natural cover to conceal your approach. A fenceline, a drainage ditch or a field of tall crops of grass enables you to sneak within gun range.

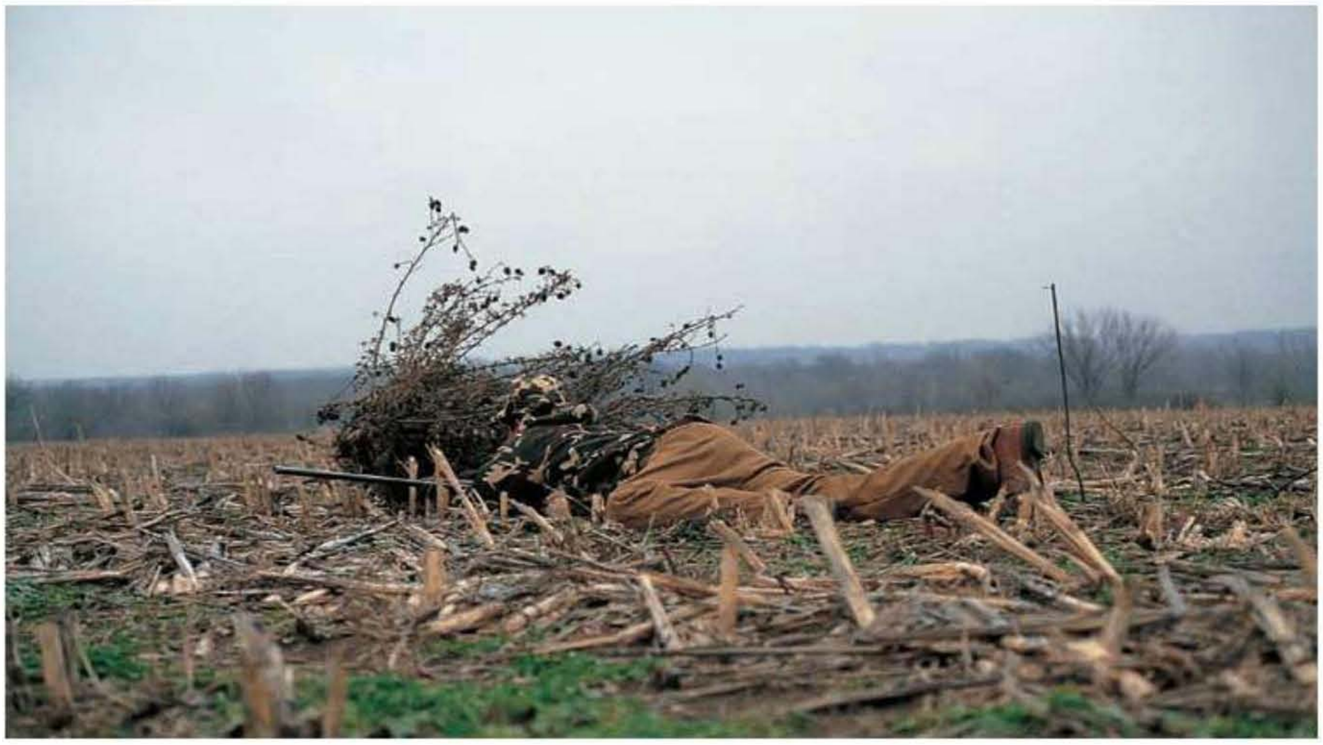

Bring your own cover when hunting in flat, open country. Push a clump of weeds ahead of you and stay low so game cannot spot your form above the horizon.

DRIVING

A lone hunter has little chance of rousting game from a large expanse of cover. Most animals sit tight or move off to the side rather than run or fly.

Before the drive begins, posters sneak to positions at the end of cover, where they intercept game pushed to them by drivers. The drivers spread out across the field or woods. The denser the vegetation, the closer they must be to discourage game from doubling back between them. It is generally better to have more drivers than posters.

Drives work best in a corridor or block of cover surrounded by open land. Game usually stays in the cover until pushed to the end, assuring someone of a shot. If possible, start at the widest end and work toward the narrowest. This concentrates the animals in a relatively small area, increasing your chances of a close shot.

Hunters should always know the position of other drivers and posters. Wear fluorescent orange clothing for maximum visibility and never shoot in the direction of another hunter.

Driving is an effective way to push game out of cover. You can make a drive with as few as two or more than a dozen hunters, depending on the situation.

TIPS FOR DRIVING

Drivers should stay within sight of each other. Outside drivers often move ahead to prevent game from escaping out the sides. Be especially alert as drivers approach posters; cornered game may spring from cover.

Islands of cover are ideal for drives. A patch of high brush in an open field, a shelterbelt or a woodlot is likely to hold game. If you attempt to hunt these areas alone, game may escape out the opposite side.

Corridors of cover, like roadside ditches, canyons, stream courses and railroad tracks, make good places for a drive. Game is less likely to double back in a narrow strip of cover than in a wide expanse.

FLUSHING

For many types of game, the surest way to evade hunters is to hide in dense cover. Even animals as large as a deer or as brightly colored as a rooster pheasant can hide within a few feet (m) of hunters without being noticed.

Hunters who use dogs stand the best chance of flushing tight-holding animals. Even in thick brush or other dense cover, a good dog detects and follows an animal's scent. Without a dog, you are sure to walk by some game.

A lone hunter can unnerve animals, causing them to flush. Unusual sounds and erratic movements may cause game to become nervous and burst from cover. Many hunters yell, clap their hands, crack brush or even run a few steps to convince an animal that it has been discovered.

When hunting with a partner, try to flush game toward each other. Start at opposite ends of the cover in an attempt to trap game in the middle. This cuts off possible escape routes. And because the animal detects danger from two directions, it is more likely to flush.

Stop frequently when attempting to flush game. If you walk at a steady pace, animals usually sit tight because they are confident they have not been detected. When you stop, they lose confidence and attempt to escape.

TIPS FOR FLUSHING GAME

Throw sticks or rocks to flush game from dense brush piles or other cover too thick to walk through. Game usually attempts to slip out the opposite side, so watch carefully and be ready for a shot.

Kick or step on clumps of vegetation that could hold birds or small game. With a hunter this close, game may be reluctant to flush. Tracks, droppings, feathers or hair may reveal the animal's presence.

HUNTING DOGS

One of the hunter's most valuable assets is a well-trained hunting dog. Watching a good dog in action adds to the thrill of the hunt. And a dog with flushing, pointing or retrieving skills greatly improves your hunting success.

Before selecting a dog, consider the type of game you intend to hunt, the terrain and climate. Pointing and flushing breeds were originally developed for hunting upland birds, most retrievers for waterfowl and most hounds for small game. Many breeds work well for more than one type of game. The Labrador, for example, is an excellent waterfowl retriever, but is also favored by many upland bird hunters for its flushing skills.

When hunting upland birds in a large expanse of light cover, use a dog that ranges widely. Pointing breeds work beyond gun range. When they detect a strong scent, they freeze in position or point. This gives the hunter time to walk in and flush the bird. For finding birds in thicker cover, a flushing dog may work better. These dogs work the cover slowly and thoroughly, usually staying within gun range.

Thorny cover can penetrate the fur and cut a short-haired dog. Long-haired breeds can tolerate thorns, but their fur often becomes matted with burrs.

In cold weather, a heavy-coated breed retains body heat longer than a dog with a thin coat. A thick coat is especially important for a dog that must retrieve in icy water. In extremely hot weather, a thin-coated breed works best. A dog with thick fur would overheat quickly.

Learn to read your dog, because every dog behaves somewhat differently when it smells game. With a pointing dog, a loose point usually means that the game has slipped away. A staunch point generally means game is close. Retrieving and flushing breeds perk up their ears or wag their tails rapidly. Hounds bay when they detect game.

Proper training and conditioning are the keys to developing a good dog. A poorly trained dog is worse than no dog at all. If it does not obey basic commands, it may flush game out of range. Or the dog may run off, costing you hours of hunting time.

Pre-season conditioning helps a dog maintain its stamina and toughens its feet. Run your dog at least one-half hour several times a week during summer.

POPULAR HUNTING DOG BREEDS

Retrievers include the Labrador retriever, golden retriever, Chesapeake Bay retriever and American water spaniel. These breeds are well-suited for retrieving waterfowl, because their skin secretes an oily substance that sheds water. Most retrievers also flush and retrieve upland game birds.

Pointing breeds include the English pointer, German shorthair, Brittany spaniel, English setter, Gordon setter and weimaraner. These breeds are used for upland birds, but a few hunters use them for rabbits and squirrels. Many pointers also retrieve downed game.

Flushers include several types of spaniels, like the springer, cocker and Boykin. These breeds pursue game until it flushes from cover, so they should be trained to work within gun range. The springer is especially popular among pheasant hunters.

Hounds, like the black-and-tan and redtick, have highly sensitive noses and are used mainly for trailing small game. Some, like the beagle, also work well for game birds. Other hounds include the bluetick, Walker and redbone.

FLOAT-HUNTING

Float-hunting works best for waterfowl, squirrels, deer, moose and other animals that frequent stream banks and lakeshores. You can also float up on flocks or rafts of waterfowl in open water.

Most float-hunters use small, low-profile watercraft including jon boats, canoes, semi-Vs and sculling boats. Boats are usually painted camouflage colors or draped with natural vegetation, netting or camouflage cloth.

The best boat depends on the area you hunt and the type of hunting. A jon boat is very stable, but the square bow would not slide through dense cattails. A canoe or double-pointed duck boat is less stable, but would slip through easily. A deep semi-V is best for rough water, but because of its high profile would not be a good choice for floating up on waterfowl. The birds would be less likely to notice a shallow-draft, sculling boat.

Silence is the key to float-hunting success. You can paddle very quietly in a canoe. If you row a boat, make sure the oars do not squeak or scrape the gunwales. Keep loose items tied down and try not to move around in the boat. Wood and fiberglass hulls are quieter than aluminum hulls. Many hunters glue carpet or rubber matting to the bottom, seats and gunwales to muffle sounds.

Keep safety in mind when float-hunting. Wear a life preserver. Never hunt from a tippy boat. It may flip over if you stand up to shoot or when your dog plunges into the water to retrieve game. To stabilize a canoe or other small watercraft, add pontoons or bands of buoyant material to the side of the hull. Low-profile boats make you less visible, but they can be dangerous on large, wind-swept waters.

To float-hunt a stream, float from one point to the next. The points conceal your approach, enabling you to slip up on game downstream.

You can pole through heavy cover, such as cane, cattails, bulrushes or other high vegetation to jump waterfowl and other game. Some hunters use push-poles as long as 16 feet (5 m) to propel narrow watercraft through the weeds. Or, you can use a duck-bill on your push-pole.

KEVLAR 45
Mad River Canoe

Water offers hunters an excellent means for a silent approach. A float-hunter can cover a large area with relatively little effort.

HOW TO FIND DOWNED GAME

Losing a wounded animal is one of the greatest frustrations in hunting. Finding downed game requires patience and persistence. But conscientious hunters make every effort to recover game that has been hit.

After the shot, watch the animal closely and try to determine if you made a clean kill, wounded it or missed completely. Clean kills and misses are usually obvious, but it may be more difficult to recognize a wounded animal. Even if you see no sign of a hit, look for hair or feathers, erratic movement or unusual behavior. You might hear the impact when your bullet strikes a big game animal.

Dogs with good retrieving skills greatly improve your chances of finding downed birds and small game. A dog marks a bird down and promptly retrieves it. If it runs, the dog circles to pick up the scent, then follows the scent trail.

Follow up any shot at big game, because large animals may show little evidence of being hit. Despite its size, a wounded deer can be as difficult to find as a cottontail. Unless hit in a vital area, it may run a long distance, especially if pursued by hunters.

If you are sure you hit a vital area, begin your pursuit immediately. An animal shot in the heart-lung area seldom runs farther than 100 yards (91 m). If you suspect a less damaging hit, it may be better to wait before following. A bleeding animal may lie down if not pursued. After 30 to 60 minutes, it will probably be too weak to run. Do not wait to follow if rain or snow threatens to obscure the blood trail.

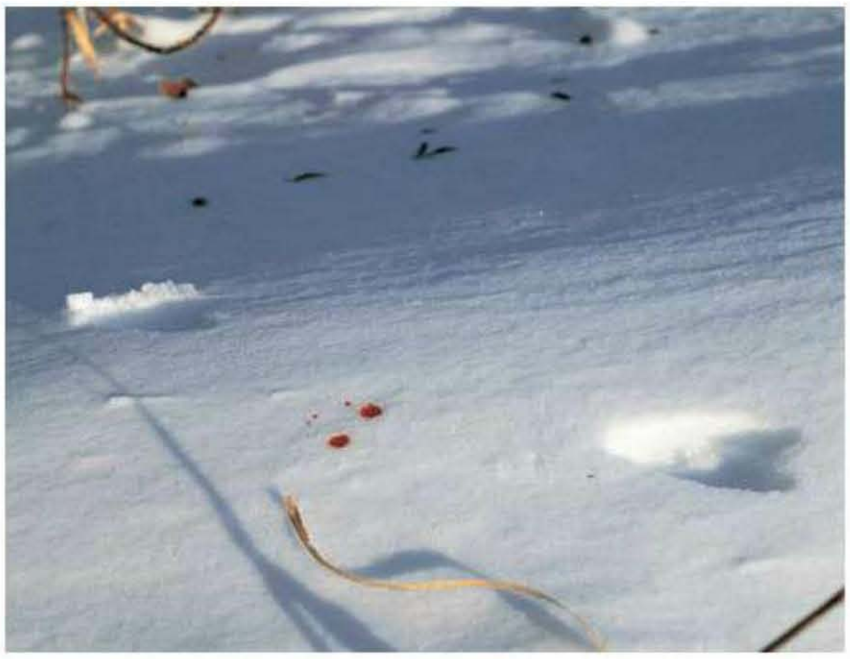

Blood trails may provide clues for pursuing wounded game. Bright red, frothy blood indicates a lung shot and tells you to follow immediately. Darker blood usually means a less damaging hit; wait before following.

Mark the spot where you last saw the animal. Look for an obvious landmark, like a tree or leave a piece of clothing at the spot. Work outward in widening circles, but return if you fail to find the animal.

Look for feathers, hair or blood where you suspect the animal to be. Stand still, listen for movement, then inspect nearby cover. Hunters may fail to find game by leaving too soon, thinking the animal has run off.

Move quietly and watch ahead for movement when trailing big game. If you fail to see the animal and approach too closely, you may frighten it off, making it more difficult to find. Locating downed game is easiest with two or more hunters. While one inspects the ground for blood, the other looks for any signs of movement. If you lose the trail, mark the spot where you last saw blood. Then you can resume the search at that point.

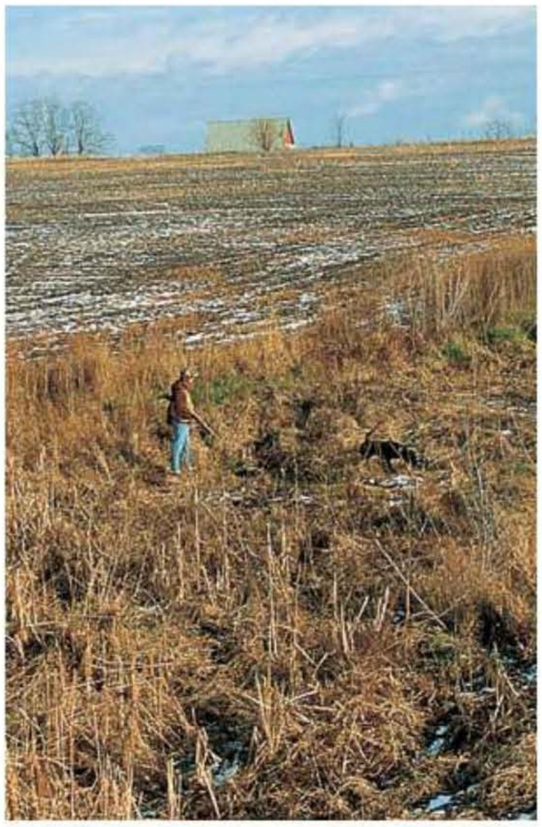

Check likely escape routes if you cannot find the animal. Wounded game may slip into a strip of grass connected to the main cover area.

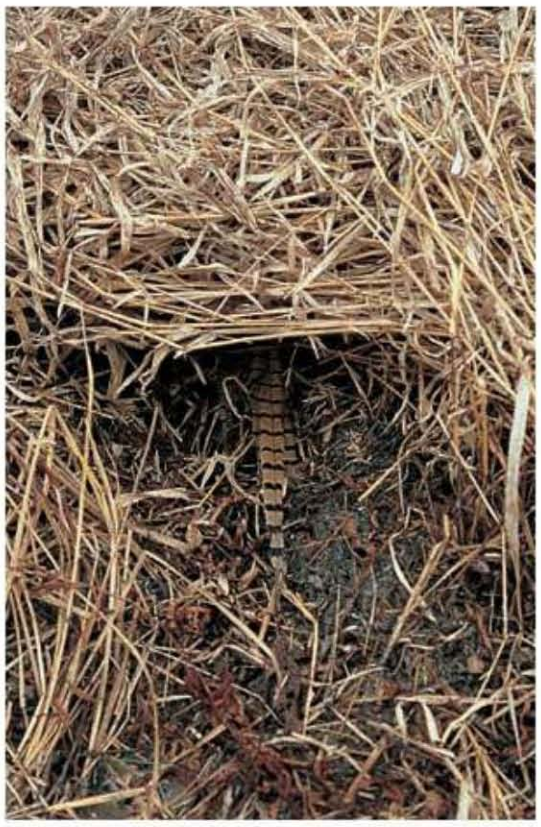

Examine thick clumps of grass or brush patches for signs of the animal. A protruding tail reveals the location of this wounded pheasant.

Work your dog just downwind of downed game. If the dog does not pick up the scent, call it to the spot where you last saw the animal.

Chapter 4

HUNTING BIG GAME

The concept of giving game "a sporting chance" is almost as old as hunting itself. It springs from the ethic of a skilled hunter who appreciates the challenge of the hunt as much as the meat or the trophy.

There is no sport with a richer tradition and no sport as demanding or rewarding as the hunt for big game. To be successful, the hunter has to develop an understanding of the quarry.

An animal's daily movements may be predictable or unpredictable, based on the species, the terrain, the weather and available feed. Scouting helps the hunter learn which trails animals are using between bedding areas and feeding areas. Such information allows the hunter to make an educated decision about where to still hunt, spot and stalk or set a tree stand.

Big game hunters should know how to use a map, a compass and survival gear. It is easy to become disoriented in the mountains, in the plains or in the fog. Study aerial photos or topographic maps before hunting and carry maps while hunting. Identify prominent landmarks in the field. Study of maps and photos will also open up new potential hunting spots.

Proper physical conditioning can make hunting more enjoyable, especially in steep or mountainous terrain. If you are out of condition or have blisters on your feet, it can be a chore to walk up a hill with a rifle and backpack, let alone drag out a large animal.

There is no sport more challenging, demanding or rewarding than the hunt for big game. The individual must match wits with an animal perfectly suited to its habitat. And the reward is great: a sense of accomplishment, perhaps a trophy and a supply of prime, high-quality meat.

WHITE-TAILED DEER

More people hunt for whitetail than for any other big game animal in North America. For many hunters, taking a trophy whitetail buck is the ultimate challenge.

The typical whitetail doe or buck is a high-strung, elusive creature that is very aware of its surroundings, whether it is at rest, feeding or on the move.

In a Michigan experiment, 39 deer including 7 bucks, 14 does and 18 fawns were fenced inside a 1-square-mile (2.6 sq km) area. Six experienced hunters attempted to find one deer. On the fourth day, one hunter finally spotted a buck. After one month, the average amount of time needed for a hunter to spot a buck was 51 hours. It took an average of 14 hours to spot any deer.

The deer's home range is small, so its very survival is based on its keen senses, its ability to find concealment and its intimate knowledge of its home range.

Whitetails rely on their sense of smell to detect danger. They can catch a whiff of human scent from hundreds of yards (m) away. Deer also hear extremely well. They do not have particularly sharp eyesight, but they are quick to detect lateral movement. Deer are not color-blind. Researchers have discovered that deer can perceive some color.

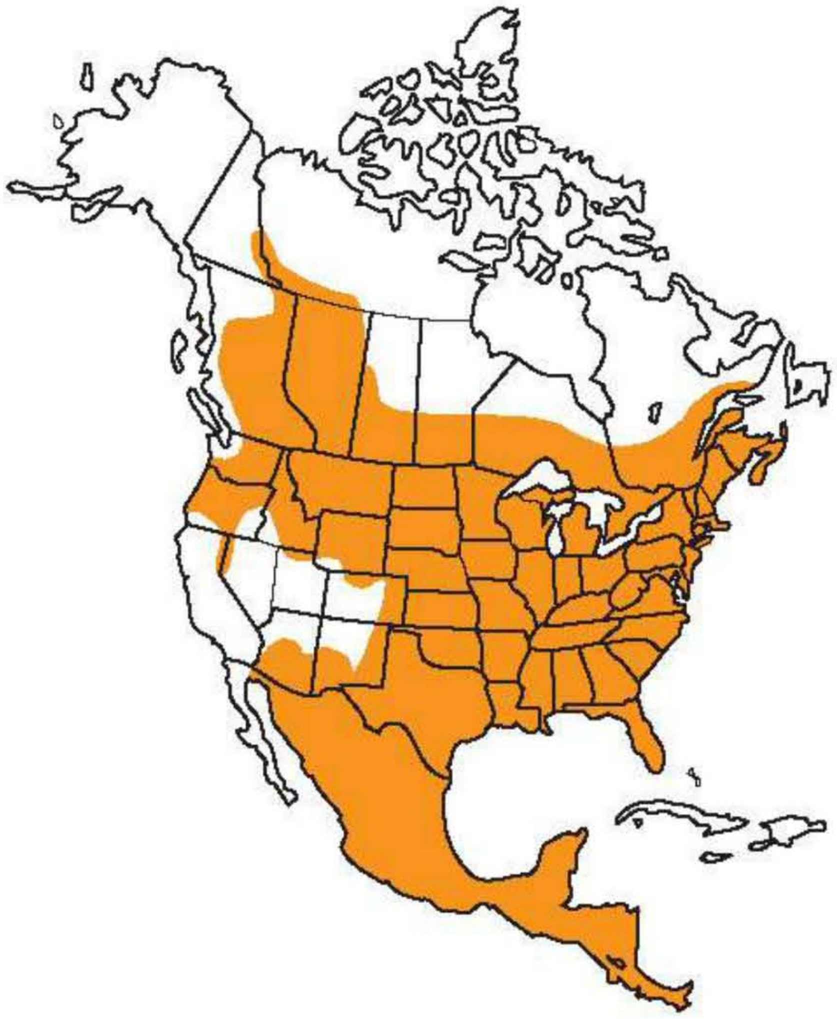

Whitetails often elude hunters by sitting tight in grassy or brushy cover. Their coats blend in perfectly. A hunter may pass within a few feet (meters) of a deer and not see it.

Deer are so familiar with their surroundings that they can quickly find an escape route or hiding place. And they are sure to notice any change in their surroundings, like a new deer stand. They avoid the area for several days until they get used to the new feature.

Whitetails usually attempt to escape danger by sneaking away unnoticed. When alarmed, bucks and does may snort loudly and stamp their feet. This serves as a warning to other deer. If threatened, a deer bounds away with its snowy white tail or flag, erect. But they normally run only a short distance, then look back to see if they are being pursued. They may stand for many minutes to ascertain if danger is approaching.

A whitetail's coat is reddish brown in spring and summer and brown or gray from fall through winter. Although the underside of the tail is white, the outside is the same color as the rest of the coat and covers up the white rump.

More than two dozen varieties of whitetails inhabit North America. The smallest variety is the Key deer, which generally weighs from 45 to 65 pounds (20 to 29 kg). It lives only in the Florida Keys. Another deer, the Coues whitetail (pronounced "cows"), is found in Arizona, New Mexico and Mexico. It reaches an average live weight of 120 pounds (54 kg) and a shoulder height of 32 inches (81 cm). The largest variety, the northern white-tailed deer, usually weighs from 130 to 190 pounds (58.5 to 85.5 kg). It is found in the northeastern states and into southern Canada. The heaviest whitetail on record, 511 pounds (230 kg), was shot in Minnesota in 1926.

The life span of a whitetail seldom exceeds 8 years. Most deer taken by hunters are 1½ or 2½ years old. You cannot tell a whitetail's age by its antlers. A 1½-year-old buck may have only spikes or it may have three or four points per side. Older bucks usually have four points on a side, but sometimes as many as seven.

The record whitetail rack came from a Saskatchewan deer shot in 1993. Both of the main beams measure 28½ inches (72.4 cm) long. The inside spread is 27¼ inches (69.2 cm). Coues deer have an average antler spread of 11.2 inches (28.5 cm) with a main beam length of 10.6 inches (26.9 cm).

Whitetails can attain speeds of 35 to 40 mph (56 to 64 km/h) and can easily jump an 8-foot (2.4 m) barrier. If they cannot escape by land, they take to water. Whitetails are strong swimmers and have been observed crossing lakes several miles (km) wide.

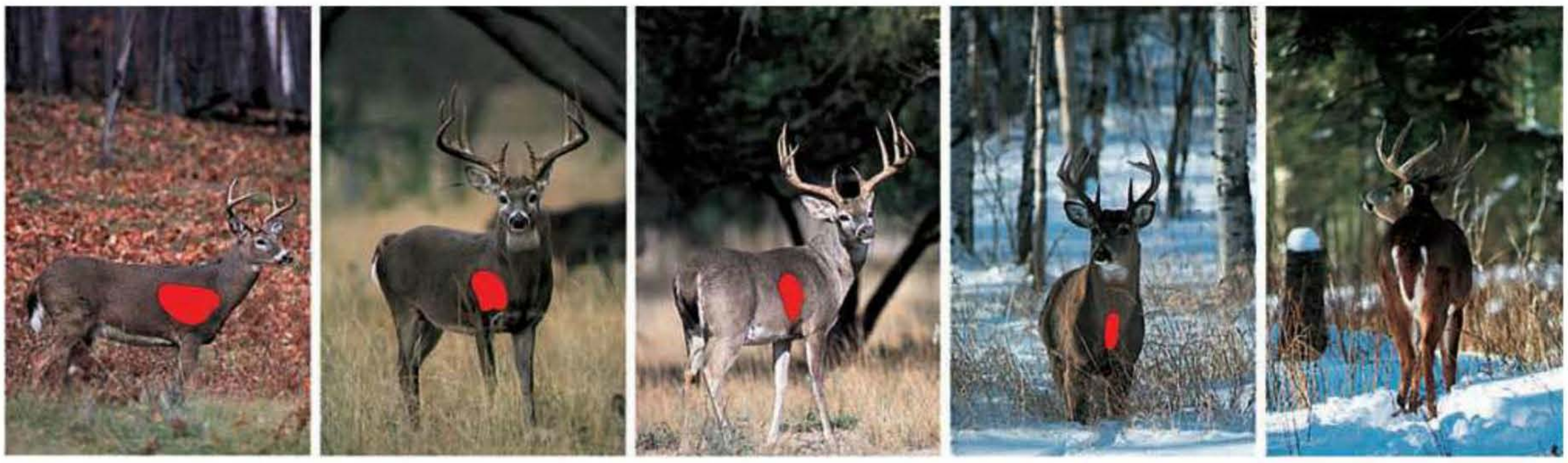

Shoot for the heart-lung area (red) of a big game animal. The best shot is directly from the side, where the most vital area is exposed. When an animal is quartering toward the hunter, slightly more than half as much vital area is vulnerable; slightly less than half when quartering away. An animal facing the hunter has only a small part of the vital area exposed. Facing away, none of the vital area is vulnerable and the shot would damage too much meat.

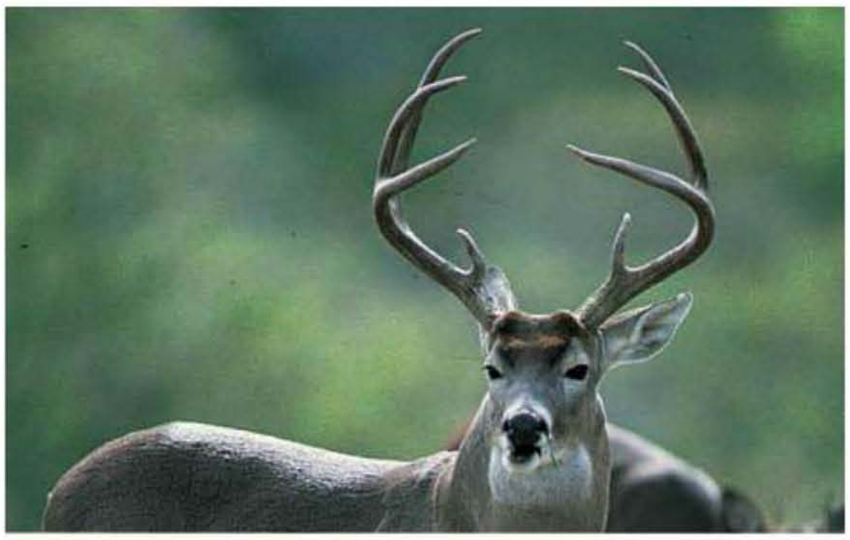

Typical antlers are fairly symmetrical, although they sometimes have a different number of tines on each side. But abnormal tines and differences between the two antlers in tine count or size of the main beam lowers the official score of the rack.

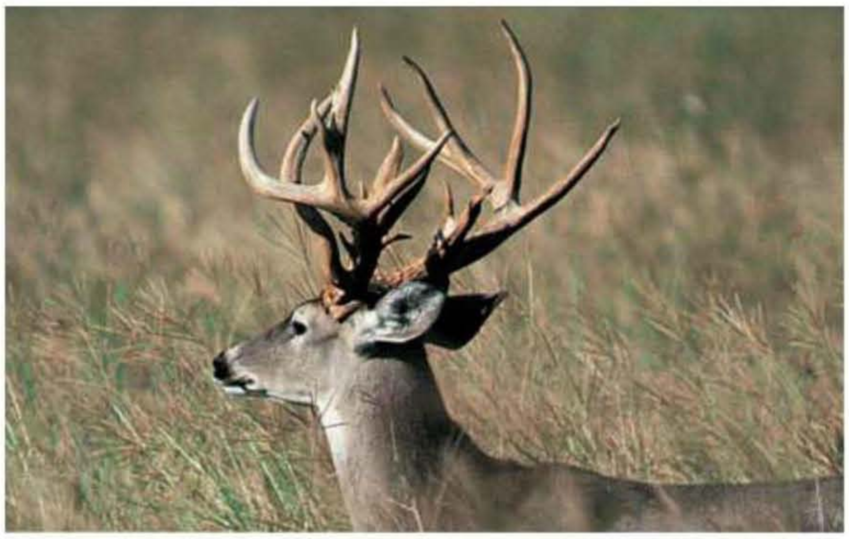

Non-typical antlers can be almost symmetrical or extremely asymmetrical. Many have a large number of oddly-shaped tines in unusual locations. Some have tines that point downward. Abnormal points add to the score of non-typical racks.

Whitetails feed primarily on buds and twigs from shrubs and small trees. They also graze on grasses, clover and other green plants. In agricultural areas, they commonly feed on alfalfa, oats, wheat and corn, but eat almost any crop available.

Does spend the summer and fall with one or two fawns. Mature bucks live alone, except during the breeding season. Whitetails mate between October and January. The rutting period lasts longer in the South that it does in the North. During the early stages of the rut, bucks and does become less cautious. Bucks are busy making scrapes to advertise their presence to does. The does move from one scrape to another looking for a suitable mate.

Where to Find Whitetails

No other big game animal has adapted to such a diversity of habitat. Whitetails can be found from the conifer forests of Canada to the chaparral plains of Mexico and from the shoreline of the Atlantic ocean to the river valleys of western Washington and Oregon. The whitetail thrives in farmlands, in suburbs and other areas where human development has severely reduced or eliminated populations of other big game.

Young hardwood forests make prime habitat. Whitetails prefer large woodlands, but can survive in smaller areas like woodlots and tree-lined stream corridors. Deer also live in grasslands, brushlands and swamps.

If you locate a good whitetail area when scouting before the season, chances are the deer will be there when the season begins. Whitetails have a surprisingly small home range. In a Texas study, marked deer were observed over a five-year period. On the average, does remained within a ⅙-square-mile (0.4 sq km) area. Bucks stayed within an area averaging 1⅔ square miles (4.3 sq km).

Your best chance of seeing whitetails is during feeding periods or when they move between resting and feeding areas. Deer feed most heavily before sunrise and after sunset. They bed down in midday. But season, weather, moon phase and hunting pressure can alter daily movements.

The changing seasons can affect whitetail movements in several ways. As the weather begins to cool in fall, deer feed for longer periods to build up their fat reserves for winter. When the acorns start to drop, deer often feed in the woods rather than moving to their usual feeding areas. Movement increases during the rutting period, as deer wander about in search of a mate.

Scouting for deer continues year-round. Use springtime turkey hunts and early bird seasons to keep track of big game movements and habits.

Changing weather increases deer movement. The animals sense impending weather changes, so they feed heavily while they can. On hot, sunny days, deer spend more time bedded down. They do not necessarily seek cover during a light drizzle. In fact, they often stand out in the open rather than lie in wet grass. But a heavy rain forces them into dense cover.

During a full moon, whitetails may not come out to feed until after dark. In the dark phase, deer begin feeding earlier. They do not feed as heavily at night, because the lack of light curtails their activity.

Hunting pressure can have a dramatic effect on deer movements. In heavily-hunted areas, whitetails change their feeding schedule once the season begins. They feed earlier and later in the day or even at night, to avoid exposing themselves during shooting light.

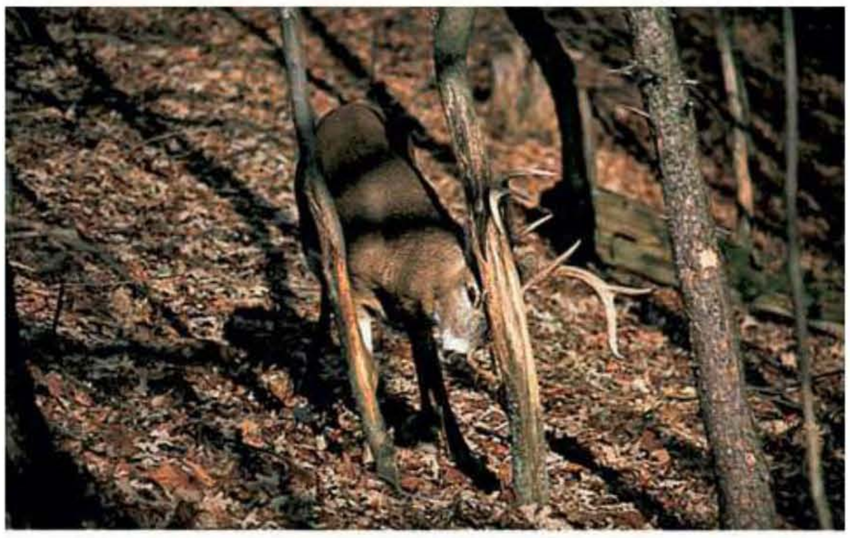

Rubs on small saplings result from bucks marking their territories by rubbing scent from glands on top of their heads. Bucks rub dozens of trees during the rut, seldom returning to the same ones. They also rub before the rut to remove velvet from their antlers.

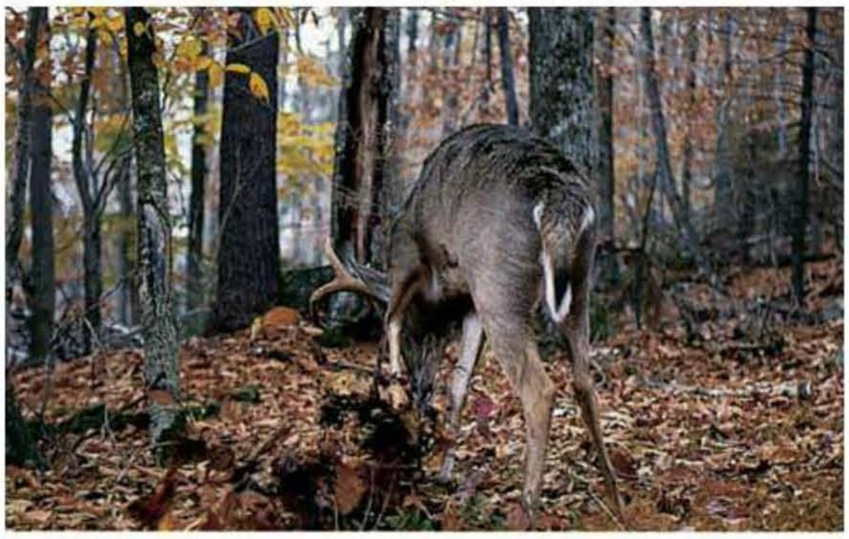

Scrapes on the ground mean that a buck is attempting to attract does. As they paw the ground, whitetail bucks often thrash nearby saplings or overhanging branches with their antlers. They check their scrapes regularly, especially those visited by does.

TIPS FOR FINDING WHITETAILS

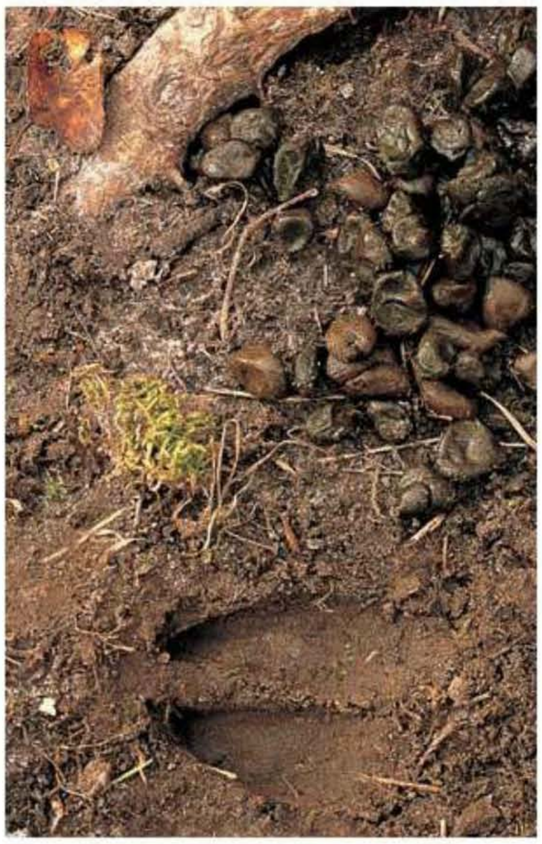

Tracks and droppings reveal how many deer are using an area. Deer tracks range from $2\frac{1}{2}$ to $3\frac{1}{2}$ inches (5 to 9 cm) long. Whitetail pellets are more elongated than those of rabbits and hares.

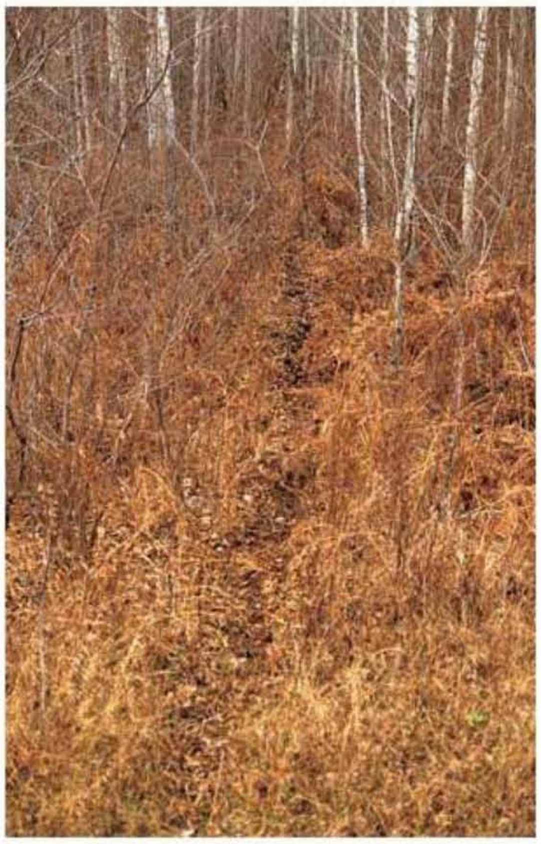

Day trails wind through thick brush and trees, but seldom cross clearings. Hunters wait along these trails when deer are most active.

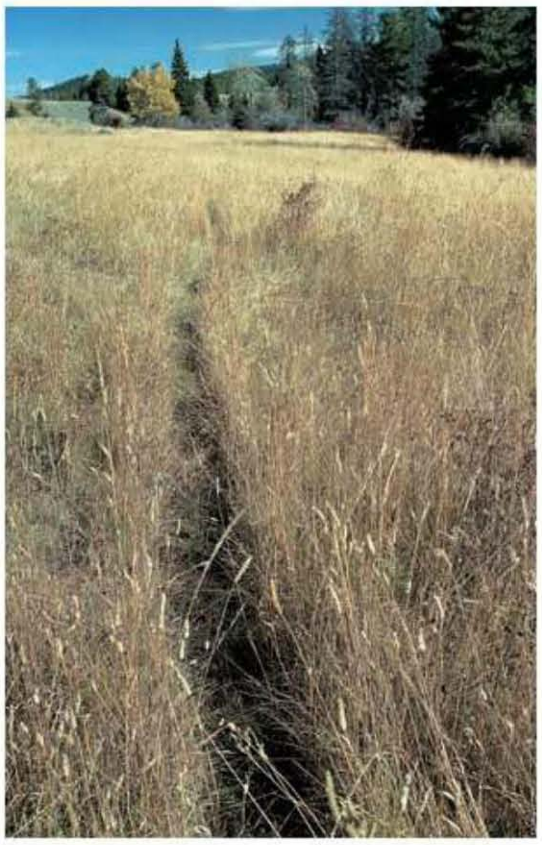

Night trails lead through meadows or open croplands. Do not choose a stand along this type of trail, because deer seldom use them in daylight.

TYPICAL WHITETAIL HABITAT

Hardwood forests provide food and cover. A young forest is best, because sunlight can reach the forest floor to grow shrubs and grasses.

Farmlands offer ample food supplies. They can support large numbers of deer if they have cover like brushy draws or stream corridors.

Lowlands like swamps, bogs and river bottoms provide dense cover. Whitetails usually feed in surrounding fields or woodlands.

HUNTING FOR WHITETAILS

Stand-hunting, still-hunting and driving account for the vast majority of deer taken each year, but hunters use many techniques to outwit whitetails. Antler rattling, stalking, and float-hunting can also be effective.

You can greatly improve your chances by planning your hunt weeks before the season opens. It is possible for an opening-day hunter to leisurely walk into the forest, find a likely-looking spot and bag a trophy buck within minutes. But the odds against such a chance encounter are staggering.

Hunters who enjoy consistent success invest a great deal of time in pre-season scouting. Regardless of how good a spot was in previous years, make sure it still holds good numbers of deer. If you find little sign, look somewhere else. Once you locate a likely area, scout it to determine movement paths and escape routes.

Expert deer hunters know not only where, but when to hunt. Wind is an important consideration. You can approach deer more easily if a light breeze rustles the leaves. This background noise makes the sound of your footsteps less noticeable. In a strong wind, deer bed down in cover and stay extra-alert.

Some hunters prefer a light rain because it softens the leaves and twigs so they do not crackle underfoot. A light rain provides a low level of background noise, but does not reduce deer activity. In a heavy rain, the animals bed down under dense, overhead cover. Powdery snow makes for quiet walking and good tracking. But when a hard crust develops, deer can hear you coming. Snow-covered tree boughs muffle your sound and may block a deer's vision.

Temperature and cloud cover also affect hunting success, but not the way many hunters think. The common belief is that hunting is best on cool or cloudy days. But after thousands of hours of observation, members of a nationwide hunting club found that deer move about more when the weather is warm rather than cold; clear rather than cloudy.

Some hunters use bottled scents, either to mask their own odor or to attract deer. Masking scents are made with skunk or fox urine and sprinkled around the stand. Attractants, made from the urine of a doe in heat or from various fruits, are spread in the area where you want to shoot your deer.

Rattling draws bucks during the rut. Take a stand near a dense thicket where you spotted deer the previous day. Or rattle near a fresh scrape, if the spot offers a clear shot. Rub, knock and rattle two antlers together to imitate the sound of fighting bucks. If nothing appears within 15 to 30 minutes, move quietly to another spot.

You do not need high-velocity cartridges for whitetails, because most shots are at close range. Cartridges should be a minimum of .240 caliber with bullets at least 100 grains.

Stand-hunting for Whitetails

Whitetails are creatures of habit. If you have scouted an area thoroughly and selected a stand near signs of recent deer activity, you can be sure that whitetails will eventually pass your way. If you lose confidence, become impatient and decide to go after the deer, you significantly reduce your odds.

Choose a stand that offers good visibility. It should be located where the wind will not blow your scent toward a trail or other spot where you expect to see deer. You should also select a spot where the sun will not shine in your eyes. Be sure you are concealed on the sides from which deer will most likely approach.

But deer may not come from the direction you expect, so you must slowly scan in a complete circle around your stand. When you spot a deer at a distance, stay motionless and be patient. If you see a doe, watch closely because there may be a buck trailing behind.

If you hunt in the morning, walk to your stand very quietly. Get there while it is still dark. Most hunters stay at their stand until about two hours after sunrise. It often pays to wait a little longer. The commotion caused by other hunters leaving their stands may spook deer to you. If you hunt in the afternoon, stay until the close of shooting hours. Hunters often see more deer in the final minutes than during the rest of the day.

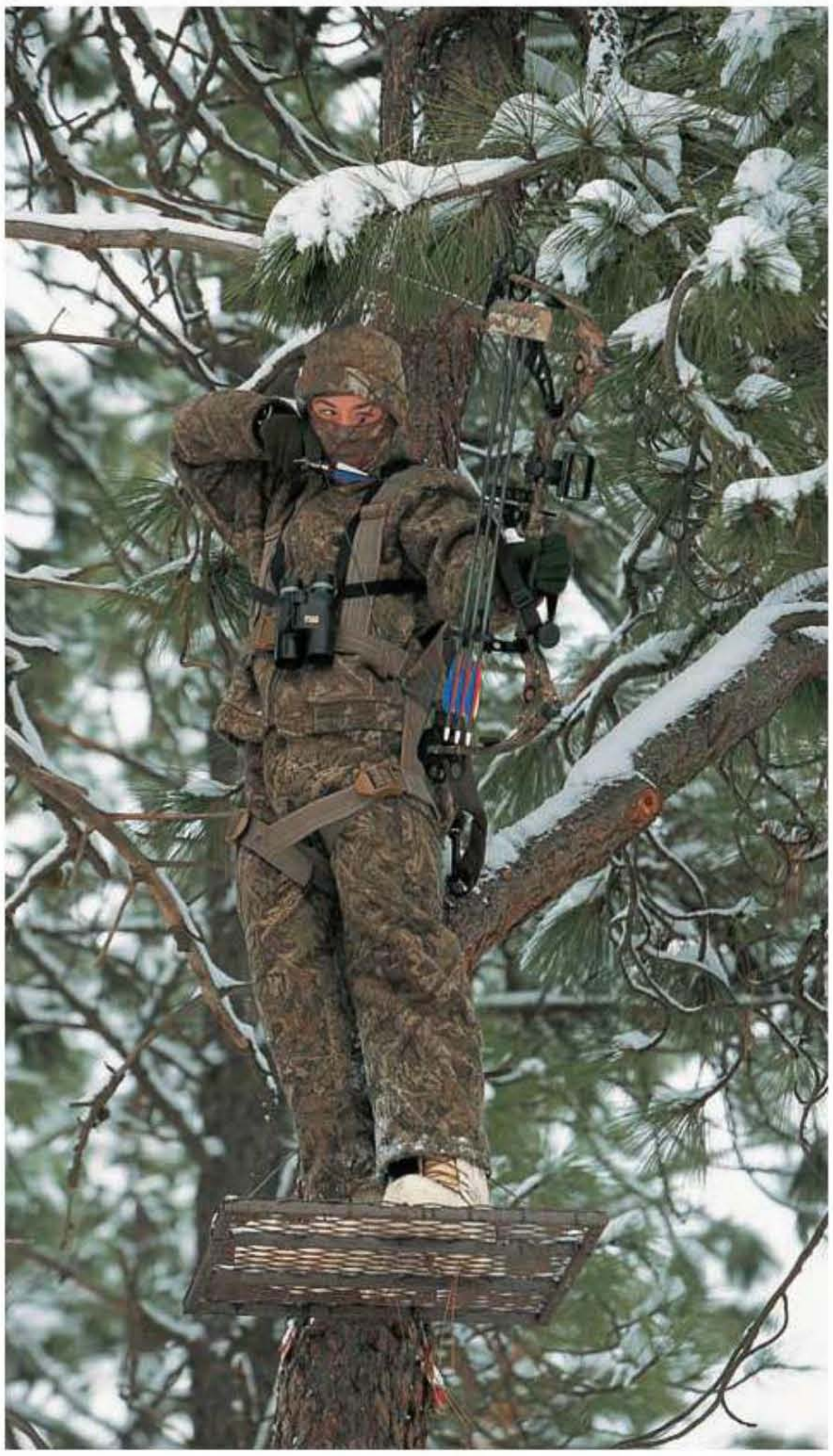

Scouting is critical to the best stand placement. Position a stand to take advantage of daily movements between bedding cover, feed, and water. Ensure that there are clear shooting lanes. On stand, keep your own movements to a minimum and be ready to stay for several hours.

TIPS FOR STAND-HUNTING

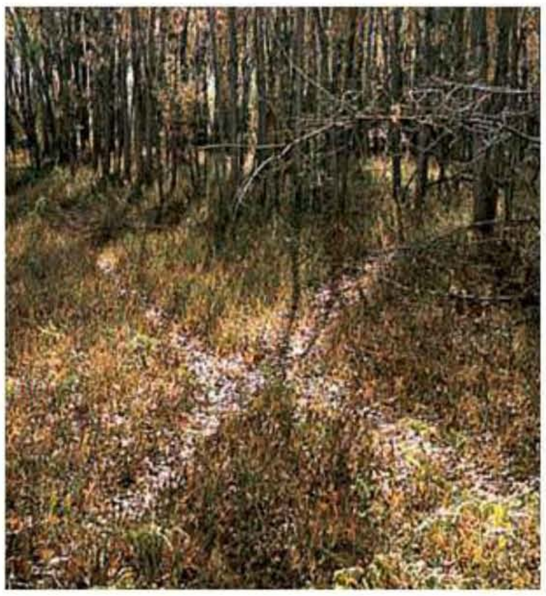

Intersections of two or more heavily used trails (left) make prime stand-hunting locations. Anther good spot is an area with many fresh scrapes.

Elevated stands place you above a deer's usual vision level. They also expand your field of view and keep your scent above the ground. But many hunters simply hide behind a tree or a pile of logs or brush.

Raise and lower your unloaded gun or bow with a rope when hunting from an elevated stand. Make sure the barrel is pointed away from you.

Tower blinds make it possible to hunt from a high elevation where there are no tall trees. Some tower blinds are permanent.

Move slower than you ever imagined you could. Expect to see a deer. If you move carefully and cautiously, you will see them before they see you.

Still-hunting for Whitetails

Technology has turned man, the ultimate predator, into man, the consumer. We consume power, petroleum, potato chips, polyester and prescriptions. We run around in circles, from home to school to job to grocery store and home again in the relentless pursuit of more goods to consume. Every year we get better at it, whirling faster and faster.

It is easy to bring our lifestyle to the woods, but it is not welcome there. In fact, it is counter-productive if the goal is to tie a tag on a deer. To succeed, we must become predators again.

The slower you move, the more deer you will see. Place each step carefully; plant your toe first, then gradually lower your heel. Let the sole of your foot feel for a stick that might snap or gravel that might grind or dry leaves that might crunch. Still hunters should wear shoes or boots with a light sole that allow the hunter to detect noisemakers before they alert game. Stop after a few steps then slowly move your head to examine the terrain. Bend down from time to time to scan the ground below the leaf line.

If you accidentally make a noise, stop. If they do not detect motion, whitetails usually forget the disturbance within a few minutes.

Wait 30 seconds or 1 minute or 5 minutes. Then take another step and scan the woods from this new vantage. You are seeing a new view of the world framed in a different series of trees, looking for the horizontal line of a deer's back or the crook of a leg against the vertical world of trees and brush.

Wind can be a hunter's ally or his worst enemy. Your scent is frightening to most deer. A deer's ability to smell is the one sense that they never question. If the wind is steady from one direction, hunt into it. A steady wind is the friend of the still-hunter.

Successful still-hunting is as much about attitude as it is about stealth. Tell yourself that you will see the deer before they see you. Convince yourself. Believe that the slower you move, the more deer you will see.

Driving for Whitetails

A deer drive's success depends on good organization. A group of hunters wandering haphazardly through the woods has little hope of shooting deer.

Every drive should have a leader who is familiar with the terrain. Before the drive, the leader gives clear instructions to each hunter. Posters take their stands first; each should wait in a spot with a good view, preferably from an elevated stand. Drivers synchronize their watches, then spread out across the upwind side of cover. Distance between the drivers may be only 15 yards (13.7 m) in dense cover or more than 50 yards (45.7 m) if the cover is sparse.

At the appointed time, drivers begin walking downwind. Deer soon detect the hunter's scent. They flush closer if drivers move quietly. Some deer move ahead, some double back and others remain bedded down.

Driving can work anytime deer are in cover. But if you drive a block of cover too large, deer slip to the side and let the drivers pass.

Deer drives can be dangerous. Limit the number of hunters so you can keep track of everyone's location. Posters in elevated stands also make a drive safer. Their shots angle toward the ground and they are above the normal shooting plane of the drivers.

Watch closely for whitetails doubling back through the driving line. Deer may sneak back though drivers are visible on both sides. As drivers approach posters, deer must double back or break into the open.

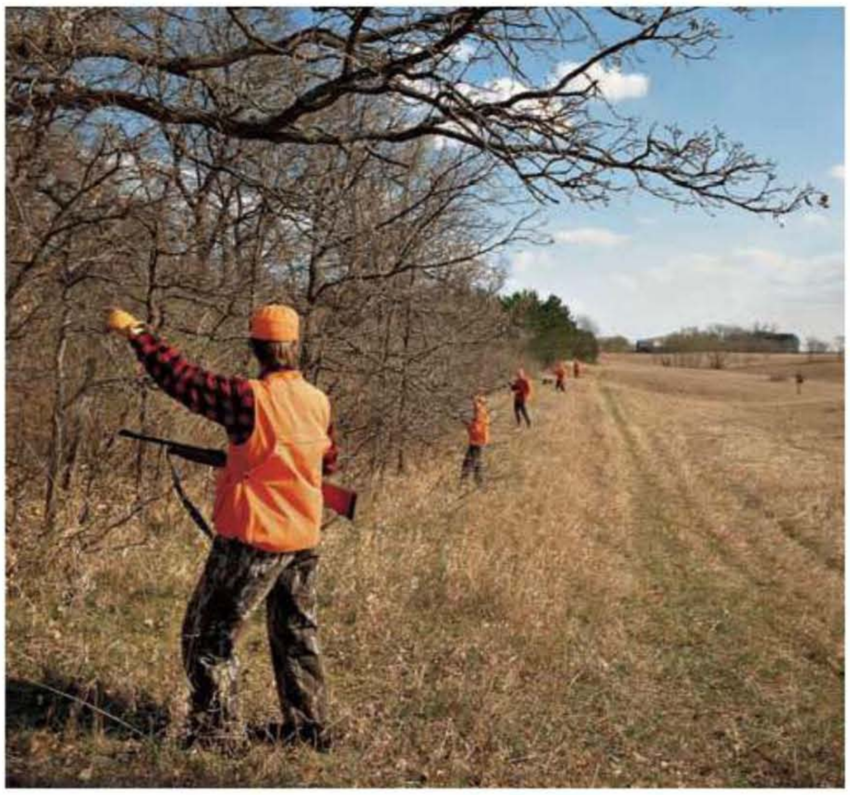

Keep adjacent drivers in sight at all times. This prevents a hunter from moving too far ahead of the others and into the firing zone.

Post near a known escape route. Deer often move from one block of woods to another by sneaking through a connecting patch of lighter cover.

MULE DEER

The mule deer is named for its long, mule-like ears, but you can tell a mule deer from whitetail by the black tip on its tail. Most of the muley's narrow tail is white, making the rump patch more evident than it is on whitetails. Their antlers also differ. The main beam of a mule deer antler is forked; whitetails have a continuous main beam.

The muley is a much sought-after trophy of the western United States, Canada and Mexico. The record mule deer rack has a spread of 30⅞ inches (78.4 cm). Its right antler has six points and a main beam that measures 30⅛ inches (76.5 cm) long; its left has five points and totals 28¾ inches (73 cm). The animal was shot in Colorado in 1972.

Like whitetails, mule deer have excellent senses of smell and hearing. But muleys have better long-distance vision. They usually bed down where they have a good view of the surrounding terrain. Unlike whitetails, they rarely sit tight and let hunters pass only yards (m) away.

When they detect something unusual, mule deer cock their large ears to pinpoint the direction of the disturbance. Then they bound off in pogo-stick fashion, with all four feet touching the ground simultaneously. This distinctive gait, called stotting, enables them to survey the terrain better and to change direction instantly.

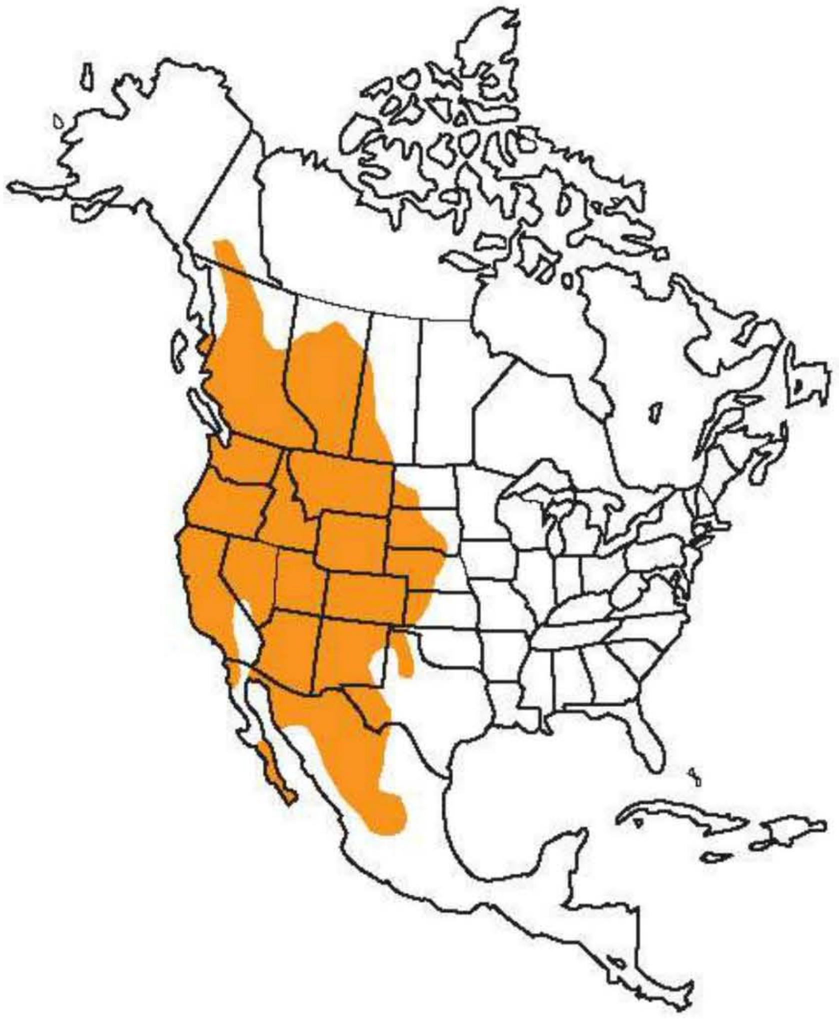

A startled mule deer runs much farther than a whitetail, sometimes up to 4 miles (6.4 km). They usually bound uphill, often pausing for a last look before slipping over a ridge. Muleys normally run with their tails down.

Compared to whitetails, mule deer have a calm disposition. They show little fear of humans as long as enough distance separates them. But they become nervous and often slip away when a hunter disappears from sight.

Mule deer prefer terrain more open than that used by whitetails. Mountains and foothills with sparse stands of timber, rolling prairies broken by canyons and coulees and low brushlands make ideal habitat.

Where mule deer and whitetails co-exist, they eat many of the same foods. But the mule deer's diet usually differs because of the rougher terrain. Common foods include bitterbrush, mountain mahogany, serviceberry, chokecherry and sagebrush. The Pious Report, completed in 1989, found that a mule deer's diet consisted of 55 percent browse (the new growth of trees, shrubs and vines with woody stems), 22 percent forbs (broadleaf plants), 10 percent grasses, 7 percent nuts and 6 percent other materials.

Browse is abundant in mule deer habitat, but broadleaf plants are mule deer favorites. Locate these plants, confirm the deer are eating them and spend time watching in the morning and evening. Early morning and late afternoon are the prime feeding periods. Favorite areas are brushy hillsides, meadows, croplands and pastures, generally at lower elevations than bedding areas.

Mule deer move as much as 2 miles (3.2 km) from feeding to bedding areas. In midday, they often bed down on the lee side of a break. They may rest just below the crest of a hill or the lip of a ravine. They watch only the downhill side, relying on the wind to bring them the scent of anything approaching from behind. Mule deer also bed on grassy mountain terrain, in the bottom of a dry wash or near a tree on a hillside. In remote areas, they may feed in the open rather than bed down during the day.

Daily movements usually begin with mule deer moving to feeding areas like a grassy stream margin in a ravine. They feed until 1 to 2 hours after sunrise, then bed in areas like a ledge near the top of the ravine. Large bucks may bed on rocky terraces at higher elevation. Deer return to feeding areas about an hour before sunset.

Deer take their water from ponds, puddles, creeks, rivers, lakes, seeps and succulent plants. Researchers in another study found that they travel an average of 800 yards (731 m) to get it. They approach, ever watchful for predators, because coyotes, cougar and other carnivores need water too.

After a rainy period, look for mule deer on sunny hillsides. Heavy snow drives them from the mountains to lower ground. Herds migrate as far as 50 miles (80 km), often wintering in brushy draws and canyons that are blown free of snow.

Mule deer often form large herds, especially during the breeding season. Small and medium-sized bucks mix with does and fawns, but the largest bucks are usually loners.

Most mule deer breed in November and December, but the breeding period may be as early as October or as last as March. During the rut, the bucks thrash or horn bushes, poles, branches and tree trunks in a display of dominance. They do not make scrapes.

Seven varieties of mule deer inhabit the western third of North America. The desert mule deer ranges as far south as central Mexico. The most numerous variety, the Rocky Mountain mule deer, is found as far north as the Northwest Territories.

Rocky Mountain mule deer reach the largest size, generally weighing from 140 to 200 pounds (63 to 90 kg). The largest on record, 453 pounds (203.85 kg), was taken in Montana. The southern mule deer is the smallest variety; it weighs from 85 to 110 pounds (38.25 to 49.5 kg).

Hunting for Mule Deer

Mule deer live in rugged mountainous terrain and high desert landscapes. They may feed in the valley on a rancher's alfalfa, but they most often bed on the high ground where they can spot danger a long way off.

A mule deer's survival strategy includes identifying the predator, then putting distance and obstacles between himself and danger. Many hunters have scored on a buck because it stopped to take one last look before it went out of sight. Some make the mistake of thinking that mule deer are easier to outsmart than whitetails, but the truth is a big mule deer is one of the most sought-after trophies on the continent.

Mule deer hunting in open country lends itself to the spot-and-stalk method. Get to the highest part of your hunting area early in the morning, preferably before shooting light. The chances of spotting muleys in the open are best in the morning and the evening. The high elevation affords a good view.

Glass every detail of the landscape, first with binoculars then with a spotting scope. Mule deer blend in well with their surroundings and are masters at concealment in a small amount of cover. Sometimes it takes movement of the sun and shadow to reveal a feeding or bedded animal. If you spot one deer, look closely for others because mule deer often feed and bed in groups.

Glass potential feeding and bedding areas with binoculars. A spotting scope works better for long distances. Look for a reflection off shiny antlers; buff-white patches of throat or rump hair; or the symmetrical shape of a mule deer's face and ears, which resembles a three-bladed propeller.

If the animal is too far away for a shot, plan a stalk. Have a partner stay at the spotting scope and direct the approach using hand signals. Use natural features of the landscape, like ravines, for concealment. Try to approach from downwind and above, because muleys expect danger from below.

Should glassing prove futile, still-hunt your way downhill. Some hunters take stands along trails where muleys are likely to pass on the way to bed, feed or water. When hunting with companions, driving can be effective.

A running buck makes a difficult target because of its stotting gait and unpredictable turns. But it may stop and look back if you shout, whistle or make a deer call.

Mule deer hunters take shots at up to 300 yards (274 m), so sight in your rifle for long-range shooting. Use a high-velocity cartridge of at least .240 caliber with a minimum bullet weight of 95 grains.

Still-hunt a series of ravines by starting at the head of one ravine, then walking the edge. Look over the crest periodically to spot bedded deer. Cross the ravine at the lower end and walk the other side to check for deer you may have missed. Then, cross to the head of the next ravine and repeat the procedure.

Stand-hunt along corridors between feeding and bedding areas. Conceal yourself behind a natural feature like a rock outcrop or large tree or use a tree stand. Mule deer trails are usually less distinct than those of whitetails. But the animals often follow natural passes like a saddle between two ridges.

HOW TO DRIVE FOR MULE DEER

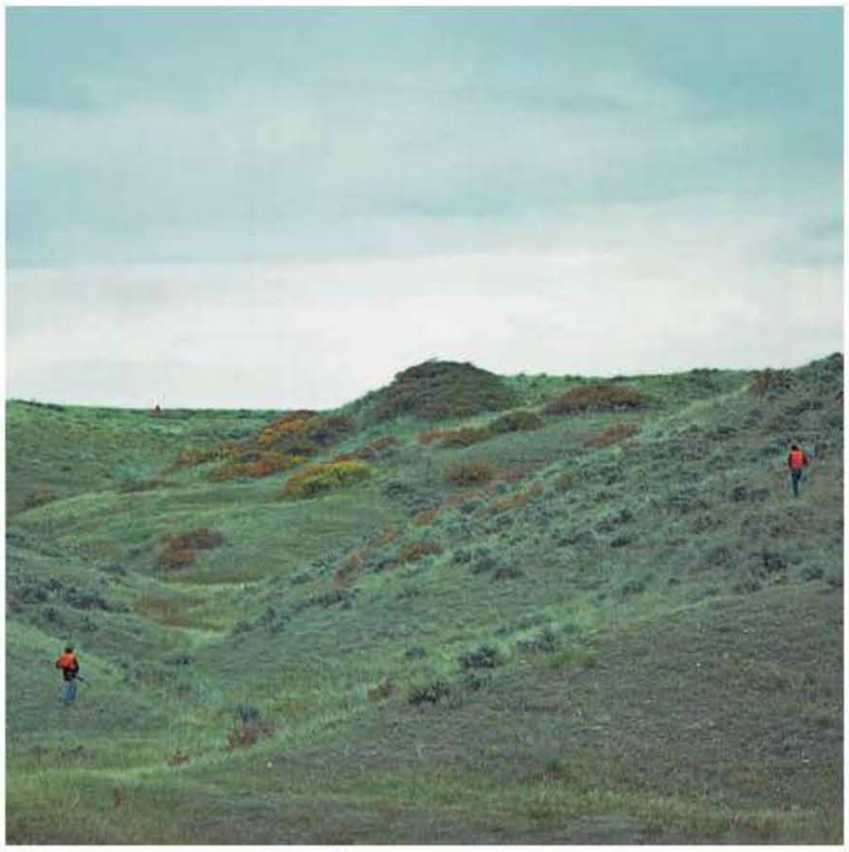

Drive for mule deer in ravines. One hunter walks up the bottom and another halfway up the side. A third hunter posts at the head of the ravine. Driving also works well along wooded or brushy stream corridors.

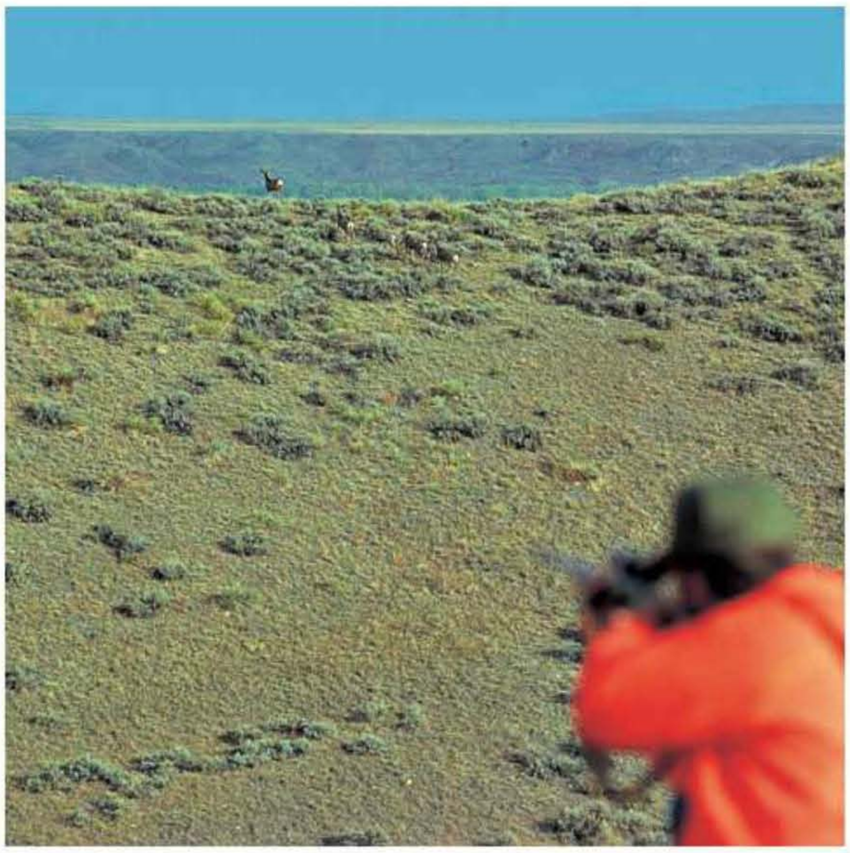

Never shoot at a deer when it is on top of a ridge. Always be sure there is a solid backstop behind the animal before shooting. Bullets travel completely through them.

BLACK-TAILED DEER

Due in part to their narrow geographic range, black-tailed deer are largely ignored in the popular hunting press, but this animal is one of America's most prized trophies. Blacktails are a subspecies of mule deer and their territories overlap along the eastern boundaries of the smaller deer's range. A doe weighs 70 to 140 pounds (30 to 66 kg). Bucks weigh between 120 and 250 pounds (57 to 120 kg). A buck may measure up to 36 inches (90 cm) at the shoulder.

Antler growth is similar to mule deer. Males grow branching tined antlers. Typical antlers are four-points (western count), but many blacktails mature with three-points per side (not counting eye guards).

They are called blacktails because the bottom two-thirds of the tail is black. The tail is wider than a mule deer's tail. At the base, it is brown. The underside is white. Blacktails are reddish-brown in the summer and tend to go gray-brown in the fall.

Black-tailed deer can be found throughout western Canada, western Washington, western Oregon and northern California, but certain types of habitat hold them in greater numbers. If you want to see blacktails, watch clearcuts and burns. Logging benefits these animals because it opens up the forest canopy to let in the sunshine. New growth springs up and deer can find most of what they need all in one place.

Deer need food, water, shelter and space. Sometimes they get all four in a clearcut. They also become vulnerable to hunters. That's why you only find does, fawns and immature bucks in large openings.

If you want to fill your tag with a doe (when legal) or a young buck, hunt the clearcuts in the morning, bedding areas at noon and paths leading from bedding areas in late afternoon. You will find the deer.

However, clearcuts are not as easy to find as they once were. Now, chemicals are used to control weeds, what we hunters refer to as forbs and what deer call groceries. Today's super-fast-growing tree seedlings turn clearcuts into tall timber again, a lot faster than they used to. Also, fear of fire may keep gates closed to timber company lands in low-water years, with "No Access" signs posted until the rains come.

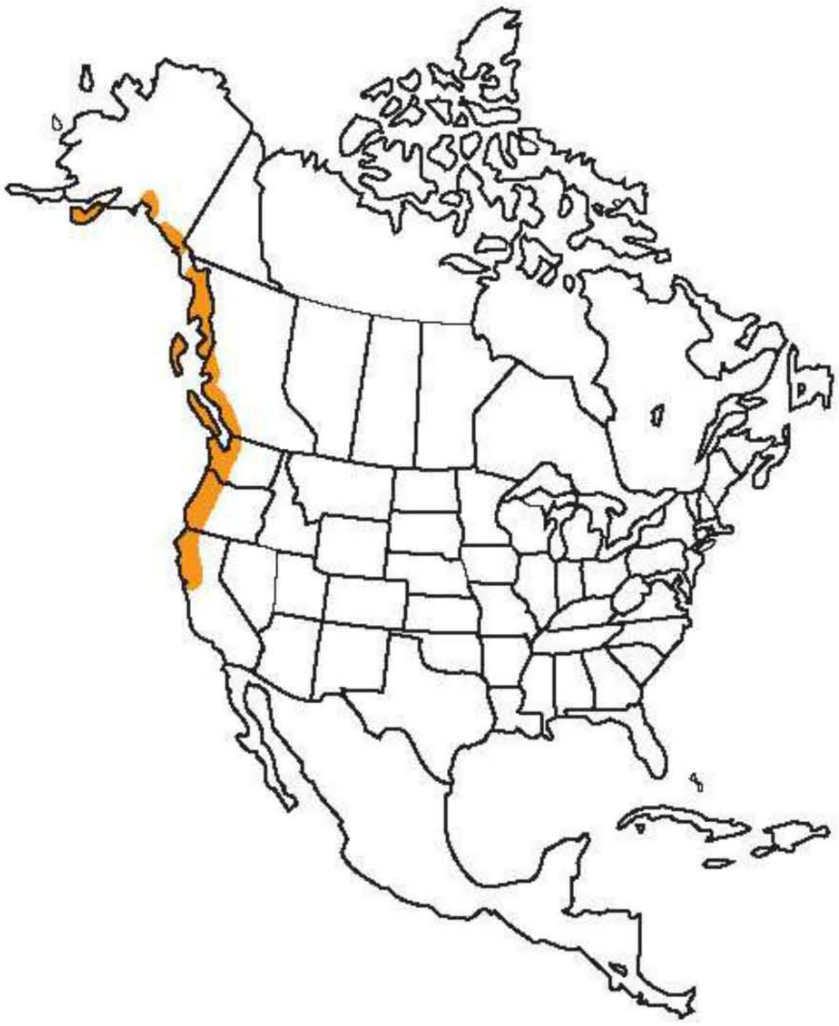

It doesn't take much habitat to hide a deer, and many big blacktail bucks grow old without ever being seen by a hunter. Their sanctuaries provide both cover and food. Water can be reached after dark. During the general rifle season, bucks may spend most of their time in an area of just a few acres (ha).

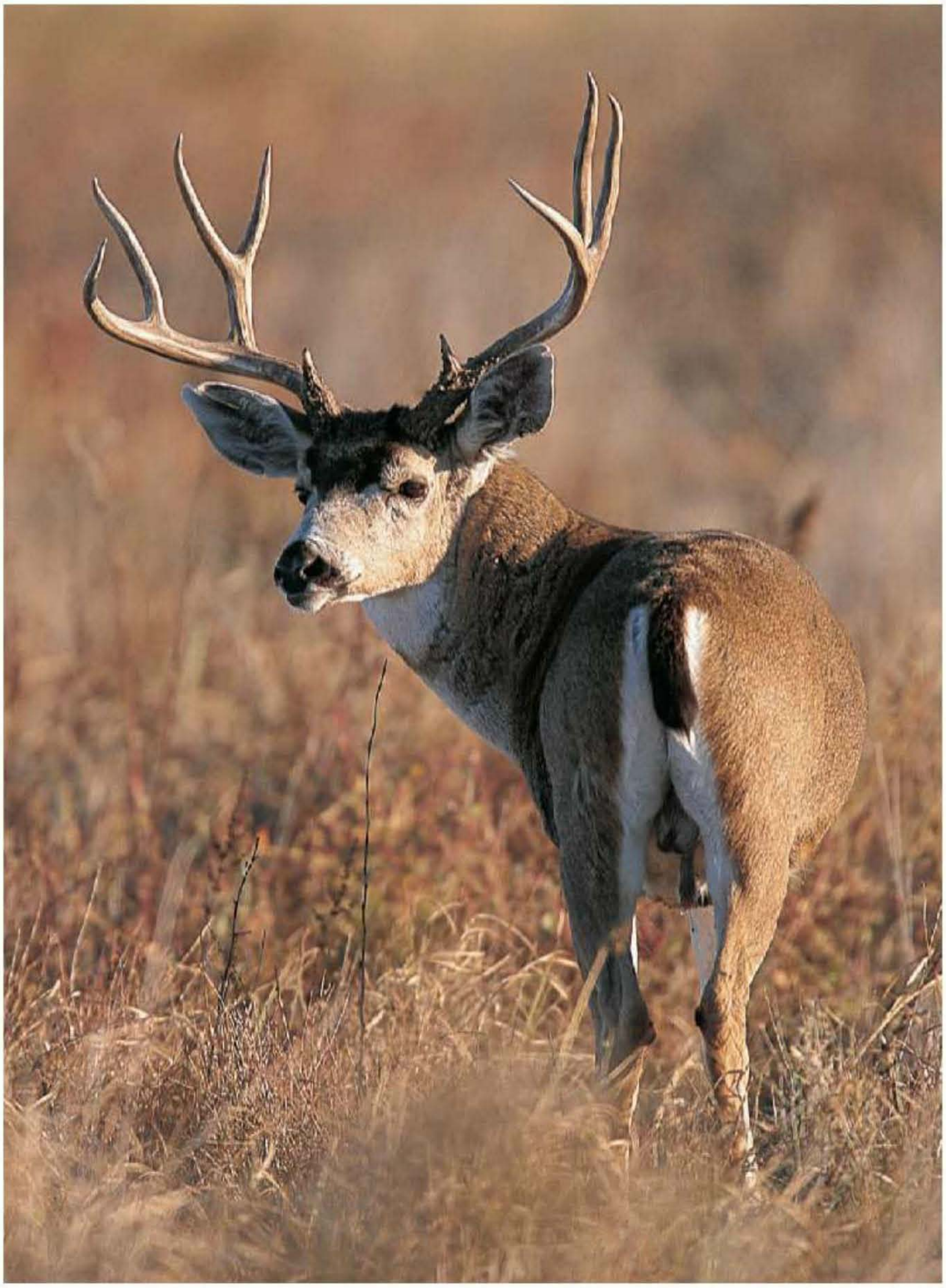

In open country and oak savannah, black-tailed deer are vulnerable to spot-and-stalk tactics. Take up a position before first light and employ a spotting scope to cover a lot of distance from one place. An afternoon stand can pay off with a glimpse of a big buck right before dark.

WHERE TO FIND BLACKTAILS

A good strategy is to look for the pockets the crowds pass by. Sometimes they may be so obvious that no one would think to hunt there. Sometimes they are so far back in the timber or the brush is so thick that other hunters detour around it.

In oak savannah, the hills are often bare, especially the south-facing exposures. Blackberry bushes, poison oak, oak trees and scattered pines grow up the sides, with bedding cover found in the run-off washes and along the creek bottoms. Old road beds provide dependable travel corridors and trails are visible from hundreds of yards (meters) away, showing where a hunter might expect to see a deer emerge from the heavy cover.

Deer are easily seen from the tops of the hills, but usually, glimpses of does and forked horns are the pay-off. Wiser bucks know to bed in the sticker patches when the orange-clad horde takes to the ridges. The bigger bucks have seen it all before.

Finding Unpressured Deer in Pockets of Cover

To find pockets that other people pass by, use a 7.5 Minute Series, 1:24000 U.S. Geological Survey map. With the contour intervals of 40 feet (12 m) on a 7.5 minute topo, it's easy to pinpoint roadless areas, identify swamps, spot north-facing ridges and locate most of the springs and seeps. You can also identify escape routes and make an estimate at where to find the dense cover where deer bed.

In these hideouts, most big black-tailed bucks grow old without ever being seen by a hunter. Their sanctuaries provide both cover and food. Water can be reached after dark. During the general rifle season, bucks spend most of their time in an area of just a few acres (ha) and feed after the sun goes down.

Private land may be the key to access, even to public ground. Get permission to cross a neighbor's land. If parking pull-outs are few and far between, have someone drop you off along the road. If a stream is the boundary, use waders or float a boat. Whatever it takes, hunt the ground other hunters can't or won't get to. And don't fall into their patterns.

Mature blacktails seldom come easy. Examine aerial photos and topo maps to find the places where the big boys live. Locate natural escape routes such as saddles and canyons. Look for brushy shelves where a buck can watch his backtrail from his bed. Locate a trail from a nearby feeding area or water and you have found a buck's living room.

Hunt into the wind and watch every step. Move too fast and deer will see you before you spot them. Slow down. Every time you take a step, a new window opens in the heavy cover. Look for the horizontal line of a back, the black of a nose, the flick of a tail, the crook of a leg or sunlight glinting from a nut-brown antler. Carry binoculars on a harness against your chest, not in a daypack. And use them.

Learning to recognize feed, water, bedding ground and escape cover are the fundamentals of hunting these most challenging of our western deer.

Carry your lunch and stay in the woods all day. Watch a trail leading into a bedding area and let other hunters push the deer your way. In the timber, hunt from above, whether from a tree stand or looking down from a high cliff. Such stands afford greater visibility, keep you above the line of sight and keep your scent stream above the game.

In open country, use a spotting scope to scout bucks on distant hilltops. It's always better to see them first and plan a strategy.

When scouting for blacktails, think small. A strip of trees between roads can hold a deer. A half-acre (0.2 ha) behind the barn may hold a buck, as will an island in a river or a lake. You may find such a spot anywhere in blacktail country.

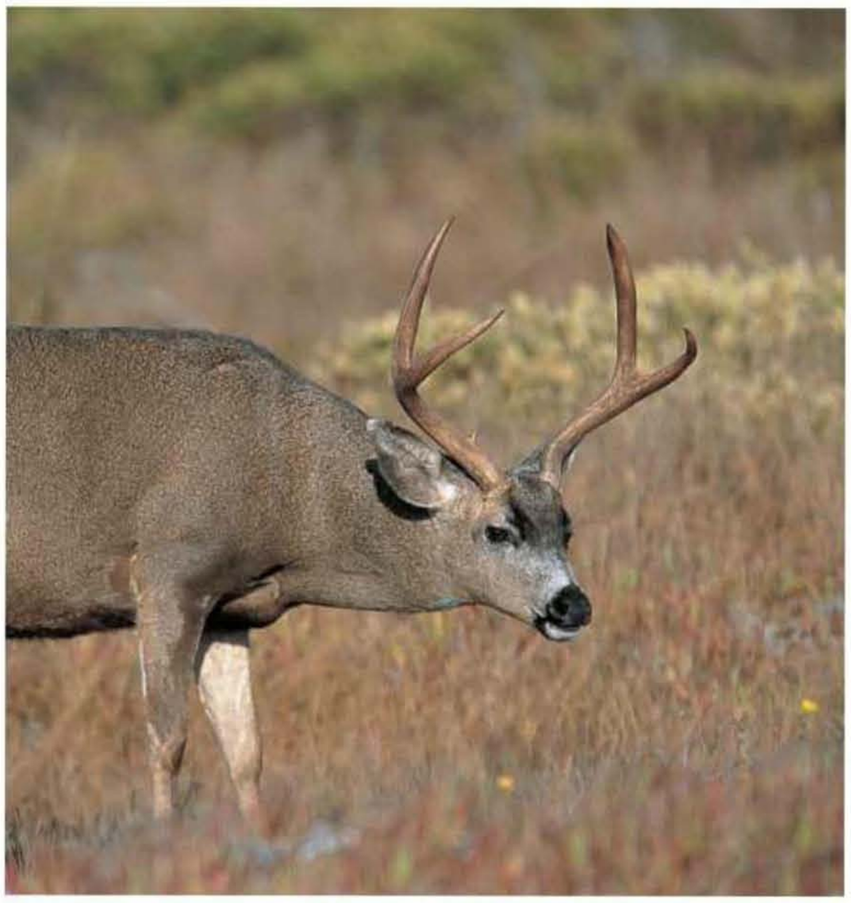

Black-tailed deer are among the most wary of all big game animals. Learn their habits, their preferred foods, and their bedding areas. Scouting pays off with more deer sightings during the season.

ELK

Sought after and prized for its magnificent antlers and savory meat, a mature bull elk is the big game hunter's greatest trophy. A large animal may have antlers 5 feet (1.5 m) long.

Antlers of the largest bulls, called monarchs, have eight points on a side. Imperial elk have seven points per side and royal elk six. The record elk rack has eight points on one side and seven on the other. The main beam of one antler measures 59⅝ inches (151.5 cm); the other 55⅝ inches (141.3 cm). The rack, which has a 45½ inch (115.5 cm) spread, came from an elk taken in Colorado in 1899.

Two varieties of elk provide the vast majority of hunting. Most numerous is the Rocky Mountain elk. The slightly larger Roosevelt elk lives in the coastal mountains of the Pacific Northwest. Elk are brownish-gray with long, chestnut-brown hair on the neck. The tail and rump patch are buff-white. Bulls average about 750 pounds (337.5 kg) and stand 5 feet (1.5 m) high at the shoulder. Cows weigh about one-fourth less. The smaller tule elk are found in California.

Elk prefer heavily timbered country broken by clearcuts, burns and meadows. The best habitat is in remote, mountainous terrain laced with streams and small glacial lakes. Elk also make their home in desert environs. In dry country, elk have large territories and do not tolerate human disturbance.

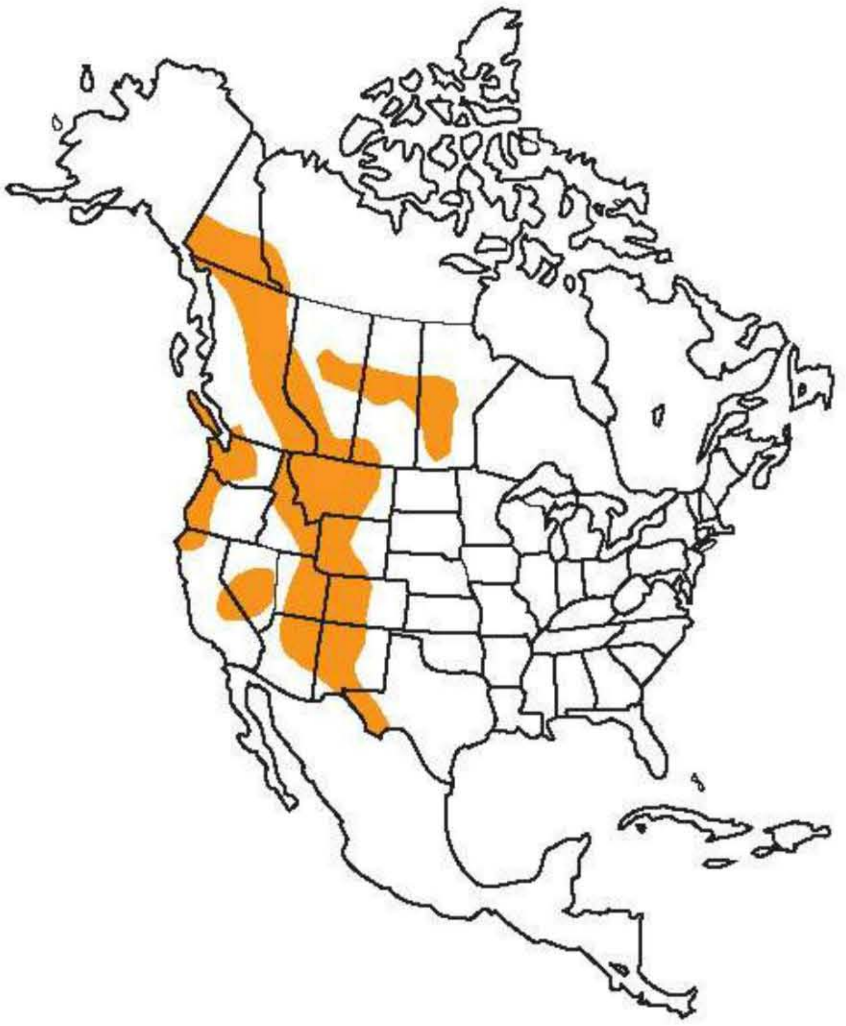

Cow elk live in large herds which also include calves and an occasional spike bull. An old cow leads the herd, alerting the others to danger with a sharp bark. Older bulls live alone or in small groups of up to six. Bulls wander more than cows, shunning cow herds until the mating season.

The breeding season or rut begins sometime between late August and mid-September. Bulls break out of their summer bachelor herds and locate on ridgelines. A dominant bull, called a herd bull, bugles to advertise for a harem that may number up to 30 cows. He protects them from the advances of younger males, also called satellites or of other herd bulls.

To detect danger, elk rely mainly on a keen sense of smell. They also have excellent hearing. On windy days, when swishing tree limbs would obscure the sound of a hunter's approach, they become nervous and retreat to heavy cover. Elk quickly notice movement, but usually ignore stationary objects.

An elk can run 35 mph (56 km/h) in a short burst and can maintain a 15 to 20 mph (24 to 32 km/h) trot over a long distance. A running bull carries his nose high, so his antlers lay back along his body and do not tangle in branches. Elk are strong swimmers and can jump obstructions up to 10 feet high (3 m).

Elk feed mainly on grasses. As winter nears, they browse on twigs and leaves from shrubs and trees. The morning feeding period begins about one hour before sunrise and lasts until one hour after. In late afternoon, they begin feeding about two hours before sunset and continue until dark. Elk usually have four or five shorter feeding periods during the day, each lasting from 15 minutes to one hour. Because they eat so often, elk usually bed within 1 mile (1.6 km) of where they feed. They prefer bedding areas with a good view, like a grassy terrace or bench about half of the way up on a hillside.

In summer and early fall, elk scatter over a large area at high elevation. The rugged terrain prevents intrusion by humans. In late fall, heavy snow and extreme cold push elk to lower elevations. But with a break in the weather, they may return to high altitudes. Some herds move 100 miles (161 km) to find the right conditions.

Daily movement patterns depend on hunting pressure. When not disturbed, elk feed in open meadows or areas with young trees and shrubs. They bed down at the edge of timber nearby. If threatened, they move up to 5 miles (8 km) into the timber after feeding. Or they may retreat to a steep, conifer-studded slope where they sniff rising air currents to detect danger from below.

Migrations in late fall begin when snow depth reaches 19 inches (48.25 cm) or more and temperatures plunge below zero Fahrenheit (–17°C). Elk move to snow-free, south-facing slopes at lower elevations.

Bugling consists of a series of melodious whistles, progressing from a low to high pitch. Herd bulls bugle to intimidate competing males. Challengers bugle and chuckle to lure the herd bull away from his harem. In addition to bugling, herd bulls bellow, rake trees with their antlers, display and spar with other bulls to drive them off. Occasionally, two bulls fight each other in a test of strength.

SIGNS OF ELK ACTIVITY

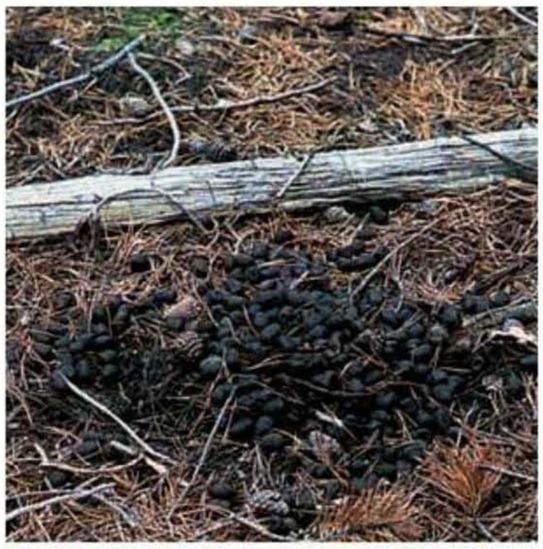

Droppings that are elongated and ¾ to 1½ inches (1.9 to 3.8 cm) in length mean that elk have been browsing on twigs and leaves. Droppings in a large mass are from elk that have eaten green grass.

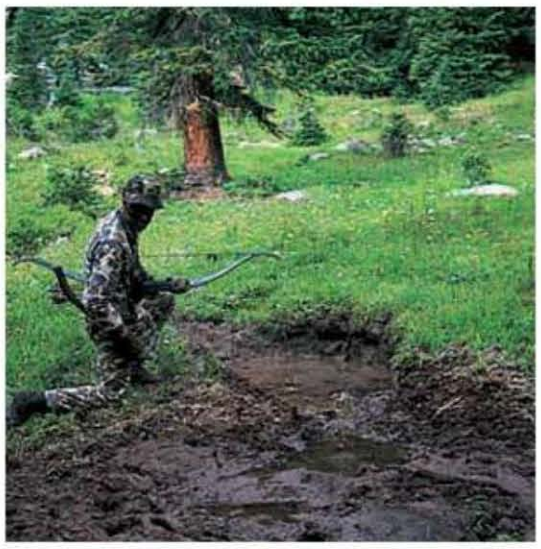

Wallows are made by big bulls to announce their presence to cows. A bull scrapes out a depression at a spring seep. He urinates in the mud, then rolls in it, plastering his body.

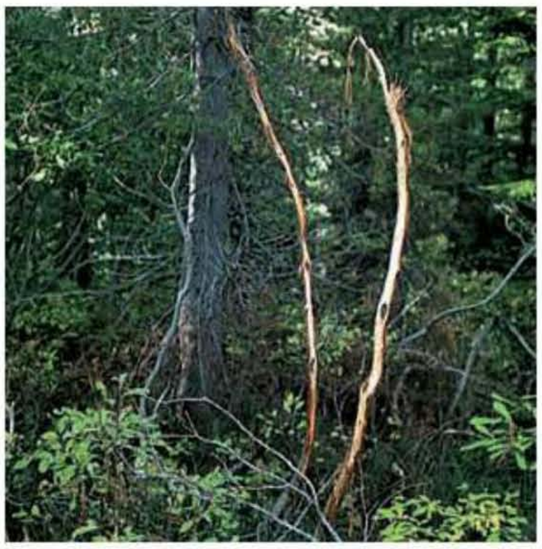

Fresh rubs during the rut also advertise a large bull's presence. Elk rubs are higher than those made by deer and the bark of the sapling is stripped over a greater length.

Hunting for Elk

A successful elk hunter must earn his or her trophy. Unlike most other big game animals, elk retreat deep into the forest or climb to extreme elevations to escape hunting pressure.

Plan to spend several days on the hunt. Scout a prospective area to find fresh elk sign, then set up camp at least 1 mile (1.6 km) away.

Scent can help you locate areas used by elk. The animals emit a strong, musky odor. The scent lingers in bedding or wallowing areas long after the elk leave.

One of the most productive techniques is stand-hunting in early morning and late afternoon, especially when the stand is on or near a saddle that the elk use to move between one drainage and the next. For a morning hunt, walk to your stand in the dark, moving quietly to avoid spooking any elk in the vicinity. Remain on your stand until about two hours after sunrise, glassing to find elk that you could stalk. In the afternoon, be on your stand at least two hours before sunset.

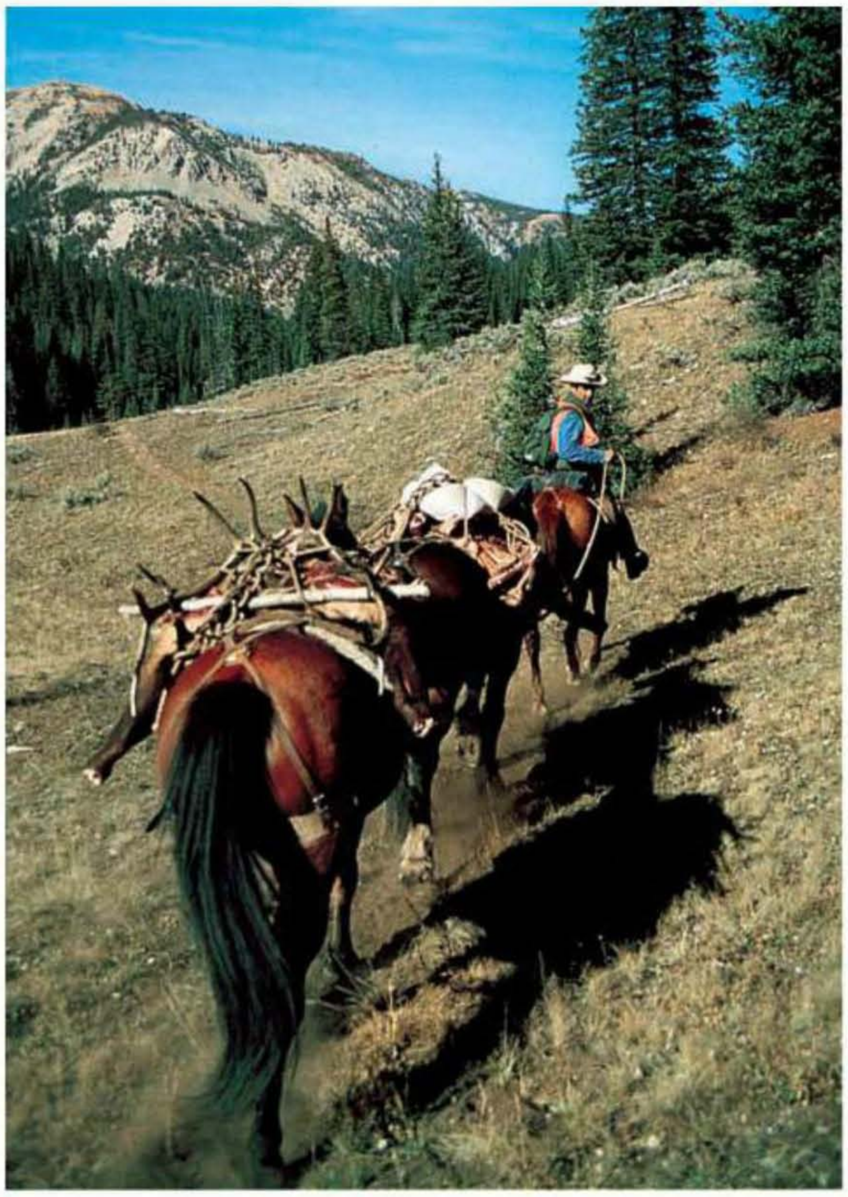

Pack trains enable hunters to reach the remote, mountainous areas which offer the best elk hunting. Horses or mules also simplify the task of carrying out the antlers and cut-up carcasses.

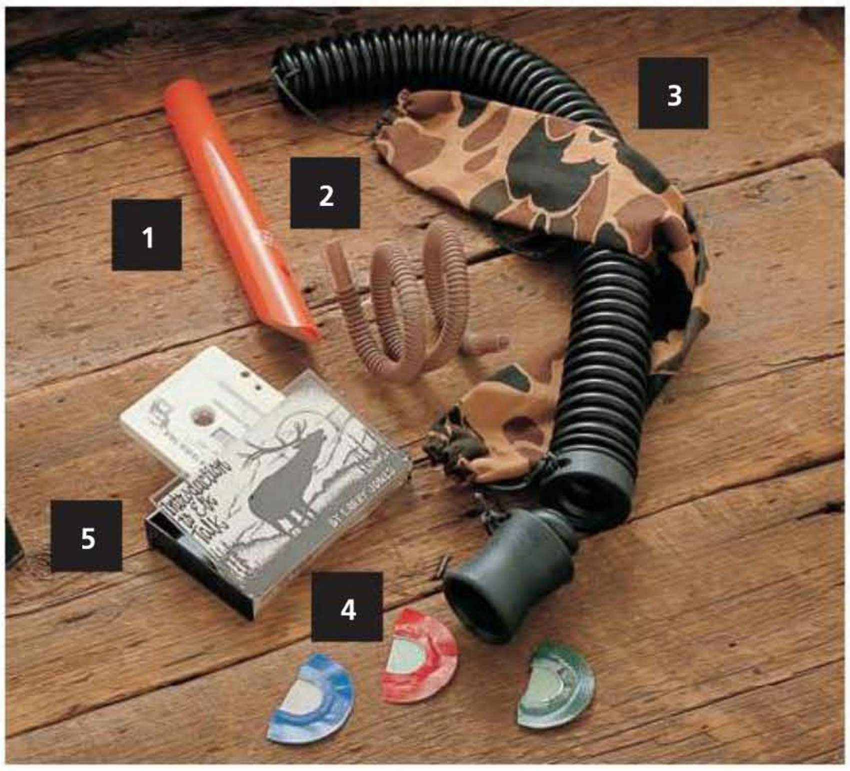

Calls include: (1) flute, (2) pigtail and (3) tube types. Diaphragm calls (4) mimic small, medium and old bulls. Tapes, CDs and DVDs (5) help you learn the best calls.

In midday, when elk are bedded down, you are more likely to see them by still-hunting. The technique works best when the ground is damp or covered with soft snow. Noise doesn't spook the elk as much as scent does. Use scent control to keep the human odor in check.

Driving can be effective, but only if the hunting party is familiar with the terrain. Drivers approach from below a known bedding area and push the elk uphill to posters. The posters station themselves along game trails in thick timber or near clearings where elk are likely to break into the open. Startled elk make plenty of noise. But more often, they slip away silently, so posters must watch closely.

During the rut, hunters can bugle in bull elk. Many elk calls mimic the high-pitched, squeaky whistle of a spike bull. This call infuriates the herd bull. He thinks an unworthy youngster is making a play for his harem, so he moves toward the caller, ready to do battle. If a bull answers but does not move, he is probably guarding his harem and reluctant to leave. In this case, try stalking close enough for a shot.

Because elk are so large and shooting distances so long, most hunters prefer high-velocity cartridges of .270 caliber or larger. Use a bullet weighing at least 150 grains.

A bow hunter should use a bow set up with a pull weight of at least 50 pounds (22.5 kg) and a razor-sharp, strong, fixed-blade broadhead. Muzzleloader hunters are best equipped when using a .50-caliber or larger, stoked with a premium bullet.

Bugle for elk starting at dawn. When a bull answers, move toward him, staying downwind and calling about every five minutes as long as he continues to respond. Keep approaching until you get within about 400 yards (366 m). Then, select a blind and continue calling to lure him within shooting range.

Stand-hunt above trails or wallows or at the edge of a meadow where elk feed. Trails and feeding areas are most productive in early morning and late afternoon. Bulls usually visit wallows late in the day.

OTHER ELK HUNTING TECHNIQUES

Still-hunt to within shooting range. Bowhunters often get as close as 20 yards (18 m). Start by walking a ridge, looking for elk on terraces or hillsides. Approach elk from above because they usually watch the downhill side.

Follow fresh elk tracks in the snow. The tracks generally measure 3½ to 5 inches (9 to 13 cm). Stay on the trail as long as it goes downhill; stay above the trail if it moves across the slope. Constantly look ahead so you see the animal before it sees you.

MOOSE

Most hunters are awestruck by the sight of their first bull moose.

With antlers towering 10 feet (3 m) above the ground, a bull moose is truly a magnificent sight. A large bull weighs 1,200 pounds (540 kg) and moose up to 1,800 pounds (810 kg) have been recorded. The world-record rack has a 65⅛-inch (165.4 cm) spread. The right palm measures 54½ inches (138.4 cm) and has 19 points; the left 53¾ inches (136.5 cm) with 15 points.

Like other members of the deer family, moose have an excellent sense of smell and good hearing. But their eyesight is poorer than that of deer. Moose seldom notice a nearby hunter if he or she does not move.

Despite their size, moose can run up to 30 mph (50.8 km/h). When spooked, they crash through brush and small trees, ignoring trails. But moose can also slip quietly through cover to elude a hunter.

Young hardwood forests with scattered conifers and brushy lowlands make ideal moose habitat. The dense undergrowth provides ample food and bedding cover. But they can also live in older forests with little underbrush.

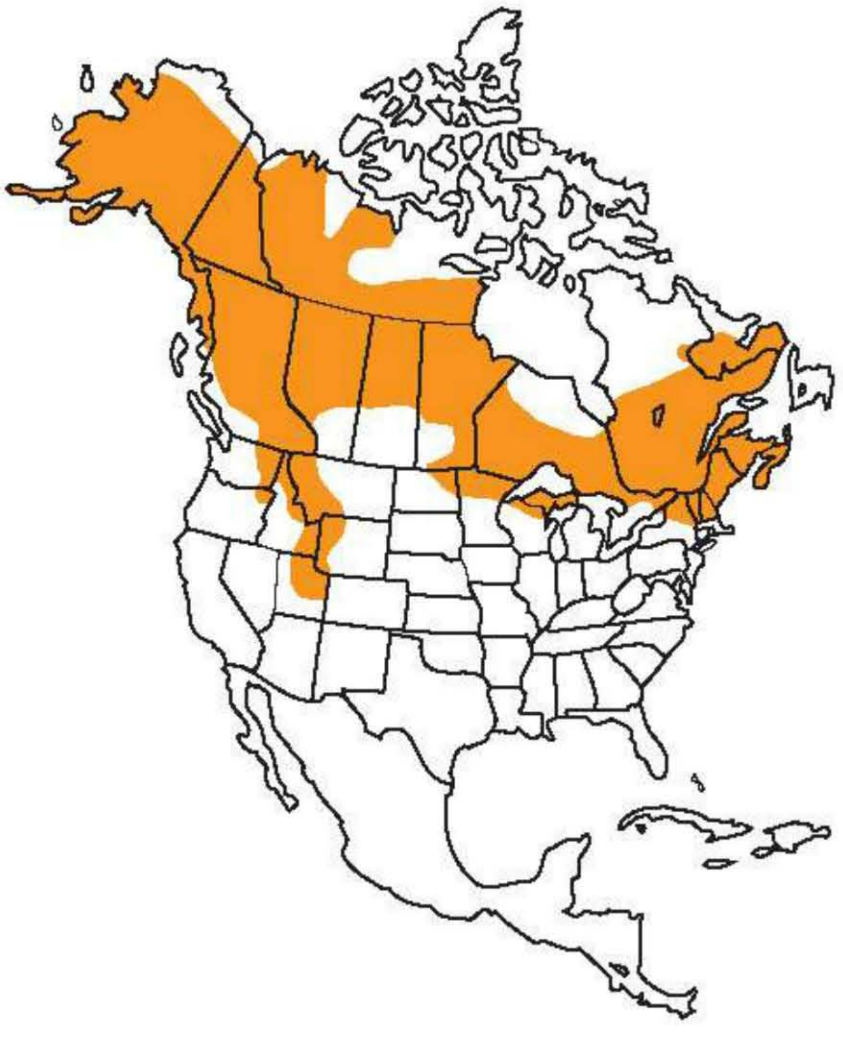

A big bull can browse on vegetation up to 11 feet (3.35 m) off the ground. If it cannot reach the upper portion of a small tree, a moose straddles the trunk, then starts walking. The animal bends down the tree, feeding on leaves and twigs as it walks. Moose feed most heavily from just before sunrise to about two hours after, and again in late afternoon. But they may feed anytime during the day or night. When not feeding, moose bed down in the thickets.

Most of the year, moose live by themselves. They lead docile lives, seldom moving more than ½ mile (0.8 km/h) in a day, but during the rut, which begins in September, bulls become ill-tempered. They have been known to attack cars and even trains. Rutting bulls regularly visit wallows. Both sexes roam widely during the rut, sometimes wandering over 10 miles (16 km) from their usual home range.

Hunters see the most moose on clear, calm days. Heavy overcast, rain, snow or high winds keep the animals bedded down.

Before you hunt, scout the area for sign. With animals of this size, the evidence will be obvious. Select a stand near a wallow, a well-used stream crossing or any spot where the topography funnels moose through a small area. Get to your stand before daylight, remain until mid-morning, then return in late afternoon.

Still-hunting can be effective at midday, especially if there is snow to quiet your footsteps. If you attempt to sneak through heavy timber and brush, you will make too much noise. During the rut, antler rattling or calling may lure bulls from dense cover. Using cupped hands or a birchbark megaphone, make a series of short grunts to imitate a cow in heat. If you hear a response, pour water in a puddle to imitate a cow urinating.

In hilly or mountainous country, hunters stalk moose after spotting them with binoculars. Float-hunting works well along streams and lakeshores.

Most moose hunters use large-caliber rifles and high-velocity ammunition, similar to those used for elk. A moose may not drop immediately after the shot. To prevent losing a wounded animal, wait 10 to 15 minutes for it to lie down, then begin to track it.

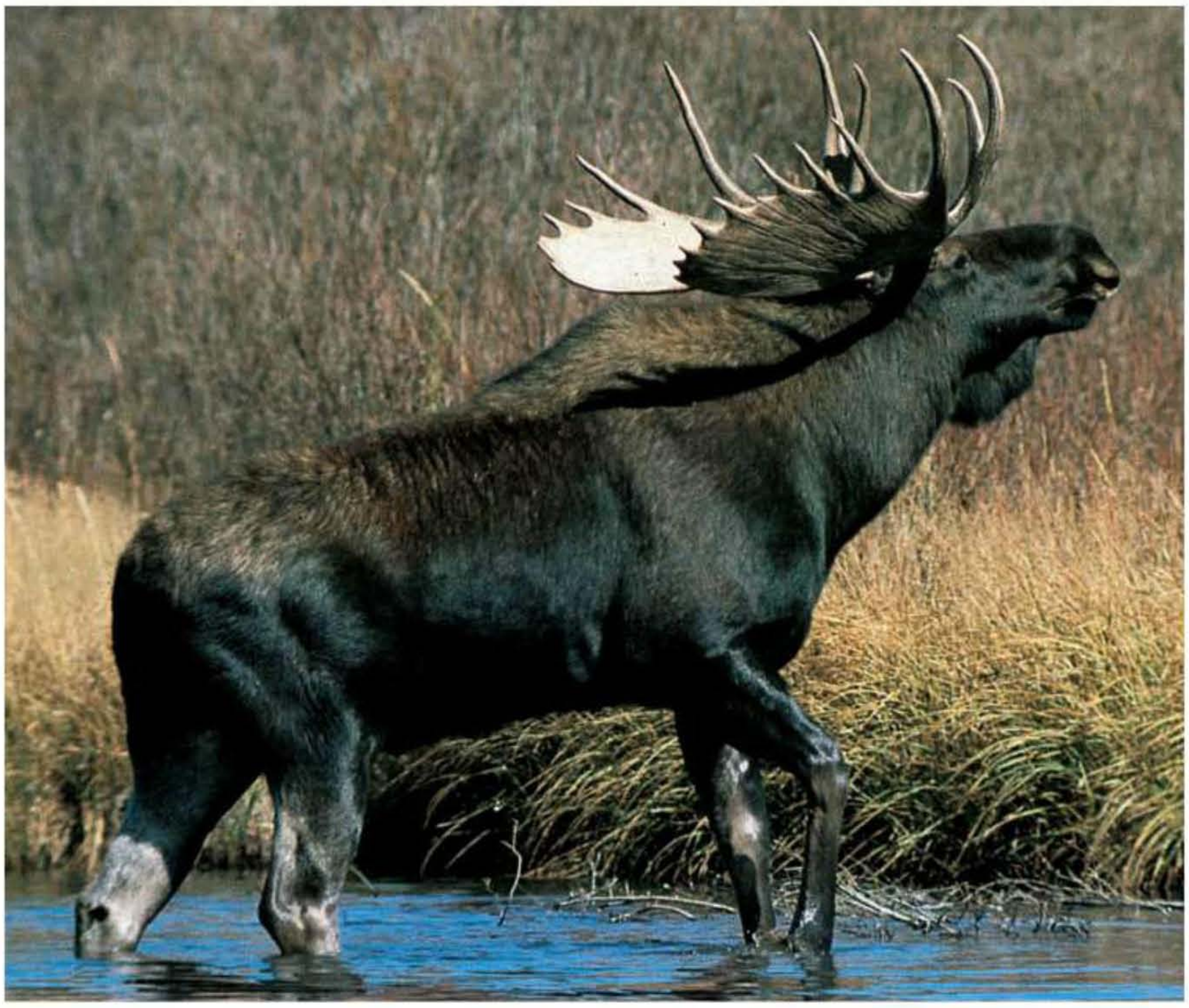

Moose have a dark brown, almost black, coat. A large bull stands 7 feet (2 m) tall at the shoulder and measures 10 feet (3 m) in length. Each antler has a large palm with numerous points along the outer edge. Moose usually stay near water. Bogs and lakes provide a source of food, a place to cool off and a refuge from swarms of insects. Excellent swimmers, moose will not hesitate to cross a fast river or even a large lake. Moose tracks measure about 5 to 7 inches (13 to 18 cm) long.

TIPS FOR FINDING MOOSE

Broken saplings indicate a bull moose in rut. A bull twists off saplings with his antlers to advertise his presence to cows.

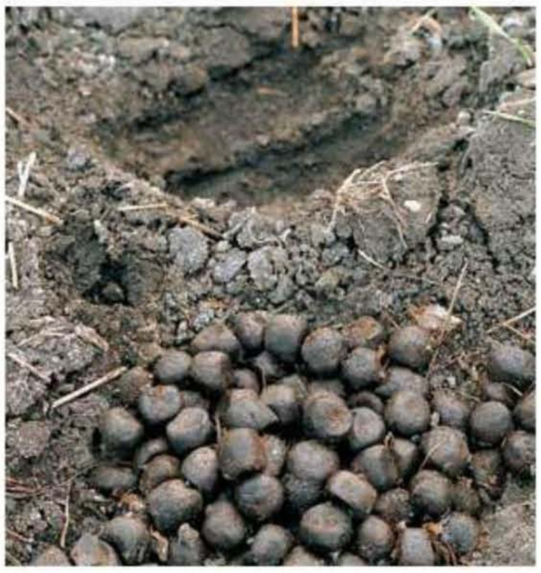

Droppings measure 1 to 1½ inches (2.5 to 3.8 cm) long. When moose are eating browse, droppings have a consistency similar to compressed sawdust.

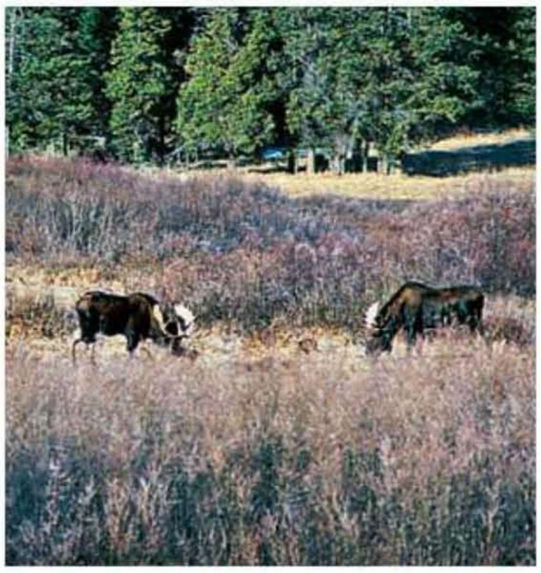

Brushy lowlands are favorite feeding areas. Moose prefer red-osier dogwood, willow, aspen, birch, mountain ash and aquatic plants.

MOOSE HUNTING TECHNIQUES

Glass for moose along forest edges, in willow swamps or near other feeding areas. Get to your vantage point early, so you can complete your stalk before the animals retreat to bedding areas.

Float-hunt for moose if preliminary scouting shows abundant sign along the stream bank. The technique works best in early morning and late afternoon, when moose come to drink and to feed on shoreline willows.

PRONGHORN

The fleet-footed pronghorn can reach a speed of 60 mph (96.5 km/h), unmatched among North American big game animals. According to fossil records, they are the only true native big game species on the North American continent. It is thought that all other big game are descended from animals that crossed the Bering Strait from Asia.

Hunters refer to pronghorns as antelope but they are not related to the antelope of Africa. Pronghorns get their name from the sharp prongs that project forward on the horns of bucks. Their upper body is tan; the underside and rump white. The record pronghorn rack has a right horn measuring 17¾ inches (45.1 cm) in length. Its left horn totals 17¼ inches (43.8 cm) long. The animal was shot in Arizona in 1985.

Special adaptations for stamina such as an oversized windpipe and large lung capacity compared to their body size enable pronghorns to maintain high speeds for several miles (km). Their large eyeballs are approximately 2 inches (5 cm) in diameter and their vision is unequaled.

Pronghorns prefer open plains, prairies and treeless foothills. The best habitat has rolling hills sprinkled with water holes and ample sagebrush for food. One study showed that the typical diet consisted of 55 percent sagebrush, 30 percent rabbitbrush, 7 percent grass and 5 percent forbs (broadleaf plants). They feed most heavily in early morning and late afternoon, but may graze anytime day or night. During the spring and early summer pronghorns get most of their water needs met from succulent vegetation in their diet, but in the dry seasons of midsummer and fall they go to a water hole at least once a day.

This animal is wary and nervous and constantly on the lookout for any signs of danger. The pronghorn's sense of smell is good; however, sight and speed are their primary defenses. When alarmed, an antelope flares its white rump hairs or makes a barking sound, alerting others in the herd. Contrary, however, to their skittish nature, pronghorns are remarkably curious and may move a considerable distance to investigate an unusual sight.

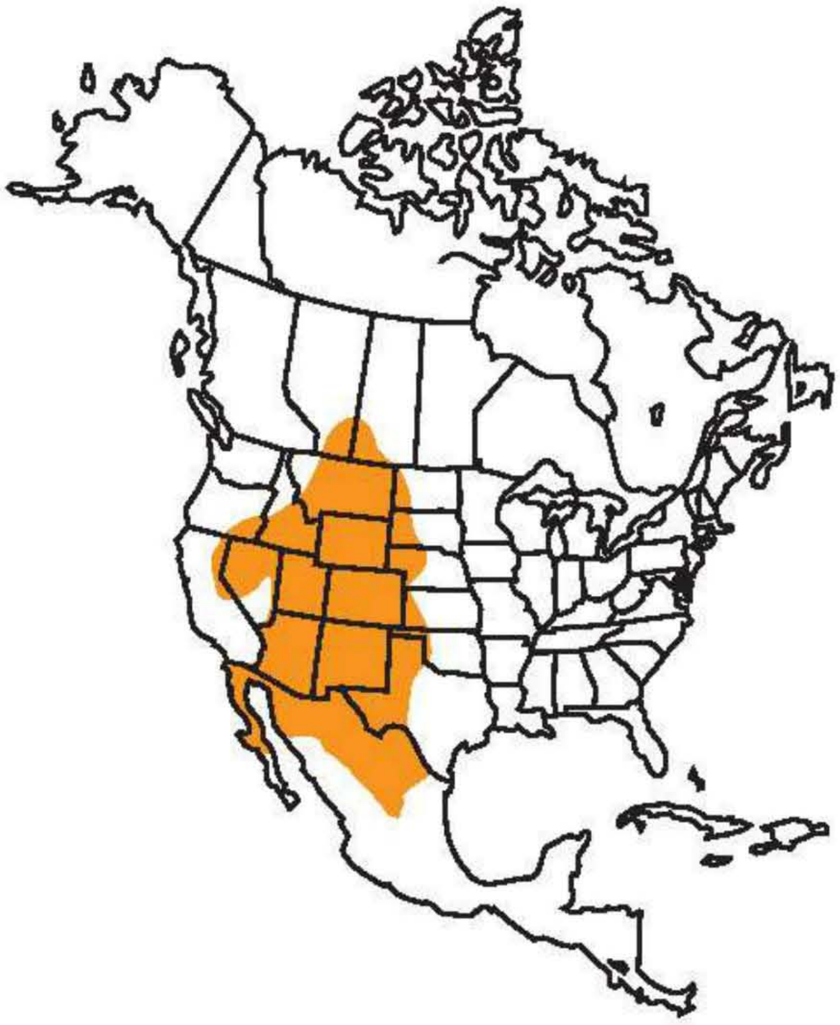

Most antelope weigh 80 to 130 pounds (36 to 58.5 kg), with the largest bucks weighing up to 140 (63 kg).

The breeding season begins in August in northern pronghorn range and may occur as late as November in the south. In spring, mature dominant bucks claim territories that contain the best food and water. They vigorously defend the territory against other intruding bucks. As the rut approaches, the dominant or herd buck attempts to gather and keep a large harem of does in his territory, breeding each when they come into estrus. Pronghorn buck fights, as the herd buck chases intruding bucks away, seldom result in serious injury. These distractions can last long enough for a hunter to stalk within range.

Spot-and-stalk is the primary tactic when hunting pronghorns. The hunter glasses, with binoculars or a spotting scope, from a high point of ground. After a good buck is spotted, a stalk, utilizing any available cover, is planned to get within shooting range.

Stand-hunting tactics can be used if a herd shows repeated daily movement patterns. Some hunters set up a blind and wait near a well-used waterhole or fence crossing. Pronghorns prefer to go under or through fences. They habitually cross at the same places, leaving hair on the barbed wire or on the ground. Stand-hunting near these locations is often productive.

Most hunters opt for a flat-shooting rifle, but using the contour of the land a gun hunter may be able to approach within 100 yards (91 m). Bowhunters must get much closer and have developed methods to overcome the limited range of their weapons. Decoying can be an effective method when the bucks are rutting. The hunter puts up the decoy within sight of a dominant buck, hiding behind it, hoping that the antelope will challenge this intruder. Antelope calls may also be used, when decoying, to challenge aggressive bucks. Calls may be used by a gun hunter.

TIPS FOR HUNTING PRONGHORN

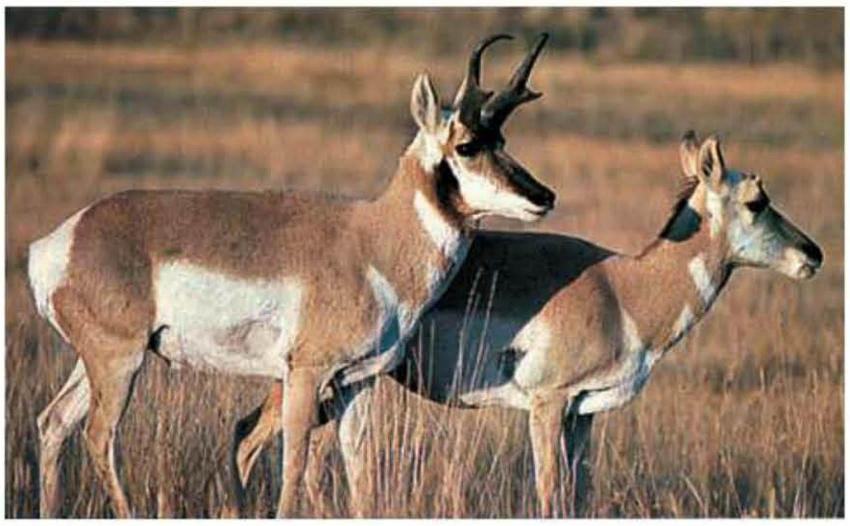

Identify bucks by their black cheek patch. Mature bucks have horns that extend above the ears. Some does have buttons or spikes, but they are shorter than the ears.

Challenge a rutting buck during the archery season using a decoy. Never use a decoy during gun season.

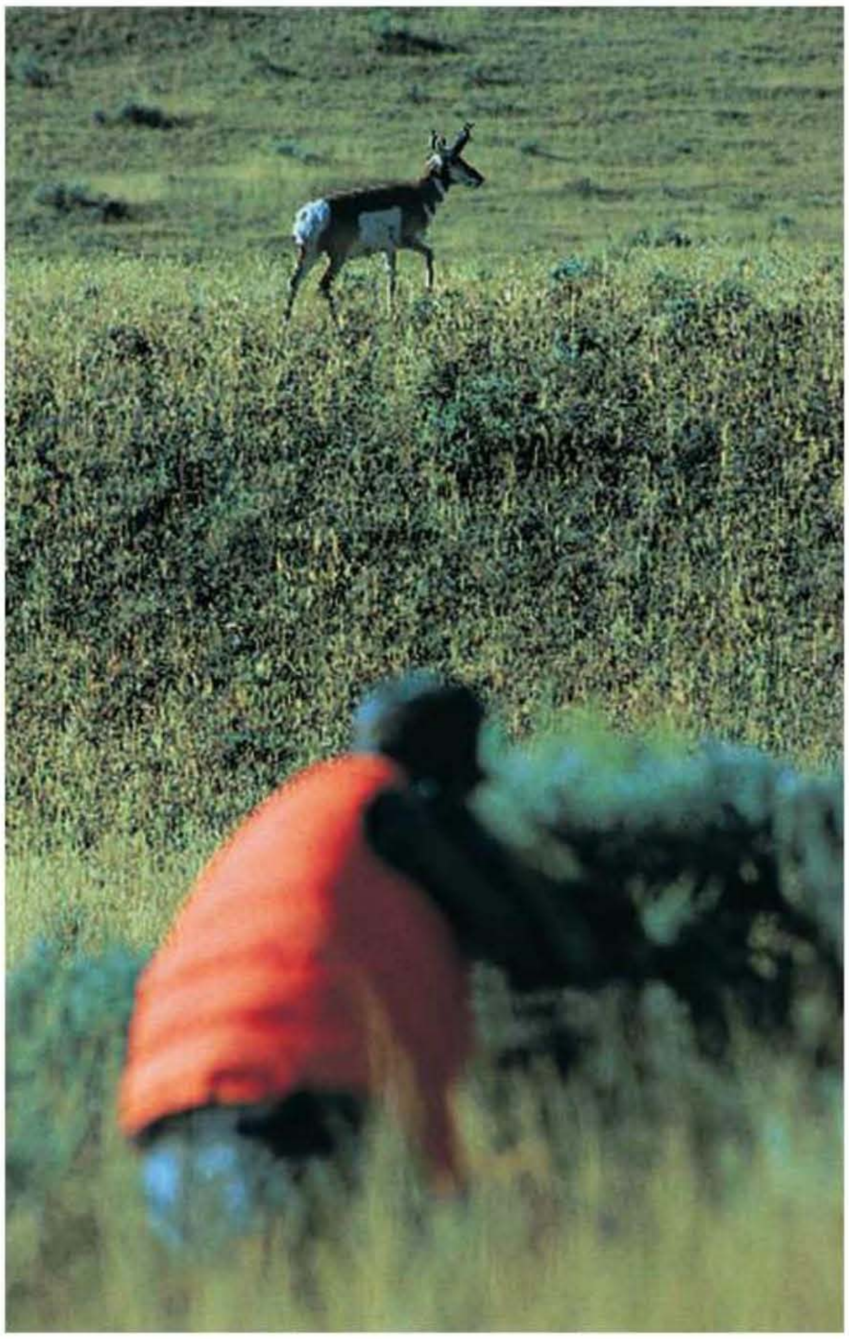

Stand-hunt from a blind near a water hole or fence crossing. Or simply hide in a depression.

BLACK BEAR

The complete hunter not only hunts for deer, elk, caribou and moose, but does his or her part to control predator populations. Spring and fall bear hunts offer another chance to pursue your passion.

The black bear is a predator and an omnivore. To some it is a pest, to others a trophy. To all it is an icon of wilderness and wild country. Of all big game animals, bear are the most misunderstood. Black bear are dangerous game because they are unpredictable.

Black bear favor a habitat of mixed conifers and hardwoods, with clearings that produce food. A bear's diet consists mainly of berries, fruits, nuts, grasses and crops. They also eat insects, small and large mammals and fish. In spring, they may feed on carcasses of animals killed by severe winter weather. Bear have an enormous appetite in late summer and fall. A mature boar can gain over 100 pounds (45 kg) in preparation for hibernation.

Bear feed most heavily during cool morning and evening hours. They remain in the shade on hot days because their fur absorbs too much heat from the sun. Windy or rainy weather limits their activity.

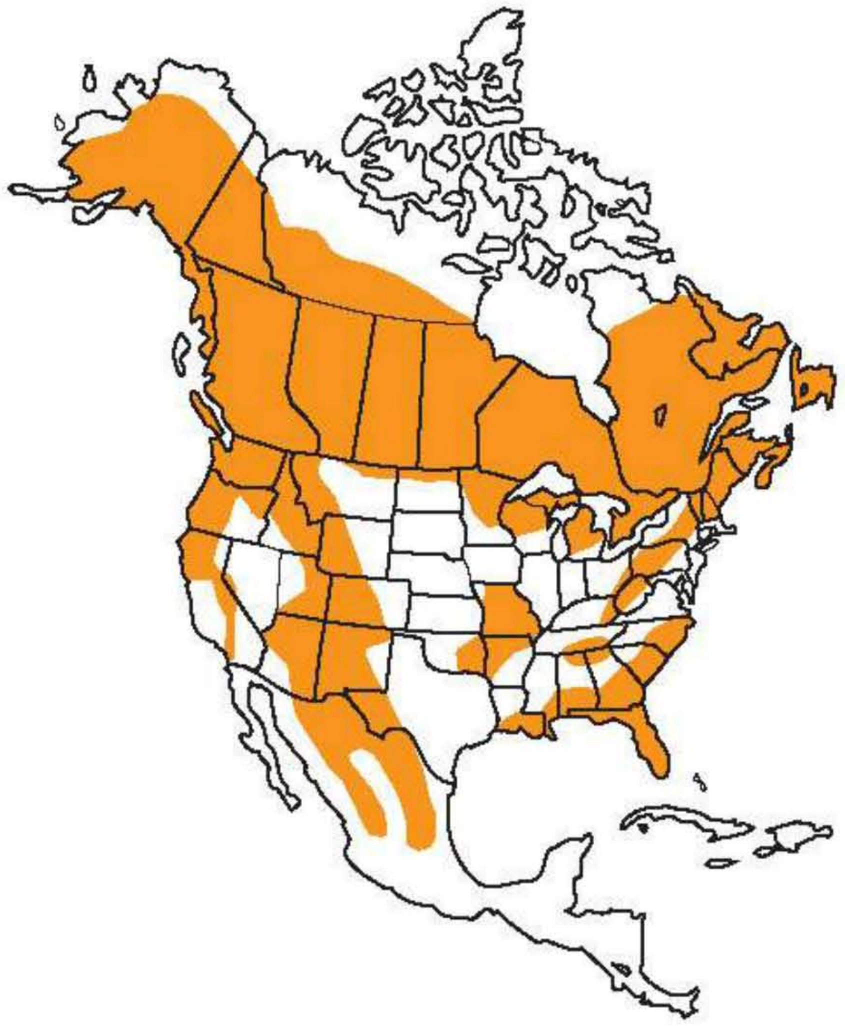

With the exception of the June breeding season, males or boars, lead solitary lives. Sows stay with cubs for about 1½ years. Boars have a much larger territory than sows, often over 100 square miles (260 sq km).

They may go into hibernation between October and December depending on the latitude and weather conditions. They remain in their dens until April, although they occasionally wander about during a warm spell. In warmer climates, bear may not hibernate. If they do, hibernation begins later and does not last as long.

Black bear come in color phases. Most are glossy black, brown or cinnamon, but they may be white or even bluish. A large boar weighs 300 to 400 pounds (135 to 180 kg); some exceed 600 (270 kg). Sows weigh about one-third less. The largest black bear on record weighed 802 pounds (360.9 kg). It was taken in Wisconsin in 1885.

Biologists rate the bear among the most intelligent game animals. A bear can run 25 mph (40 km/h) and climb a tree in seconds. They rely mainly on a keen sense of smell to find food and detect danger. They have excellent hearing and fair eyesight.

Even in good bear country, the animals are not abundant, averaging only one bear for every 2 square miles (5.2 sq km). Because bear are scattered, hunters employ techniques seldom used for other big game.

To many, the music of a pack of hunting hounds is what the chase is about. Walkers, black-and-tans or Plotts are most often employed for bear hunting. A bear can run for miles (km) and often loses the dogs. When cornered, it can quickly kill or injure a hound. Some hunters pursue bear solely for the sport of the chase. Once an animal is treed or brought to bay, they call off the dogs and let it go.

Others like the intrigue and challenge of luring a bear with baits. Baiting takes advantage of the bear's highly developed sense of smell. A bear can detect the scent of bait at up to 3 miles (4.8 km) away. This method works in spring when hungry bear emerge from dens and in fall when they feed heavily to build up fat reserves.

Patient spot-and-stalk hunting pays off where the use of dogs or bait is illegal. Hunters choose a canyon where they can watch the edge of a meadow or a patch of berries to spot feeding bears.

Calling works as well. A rodent or a fawn-in-distress call can bring in a hungry or curious bear. Use an electronic call to bring the bear within range. Let the call play constantly for up to an hour or more.

In some regions, damage complaints from bears are at an all-time high. Wildlife biologists have fewer tools today to deal with problem bears. The biologist's best friend is the hunter who has mastered the art of hunting the black bear.

Bear hunters use rifles of at least .30 caliber and bullets that weigh at least 165 grains (10.7 gm). You can kill a bear with a smaller caliber, but a larger caliber is most effective.

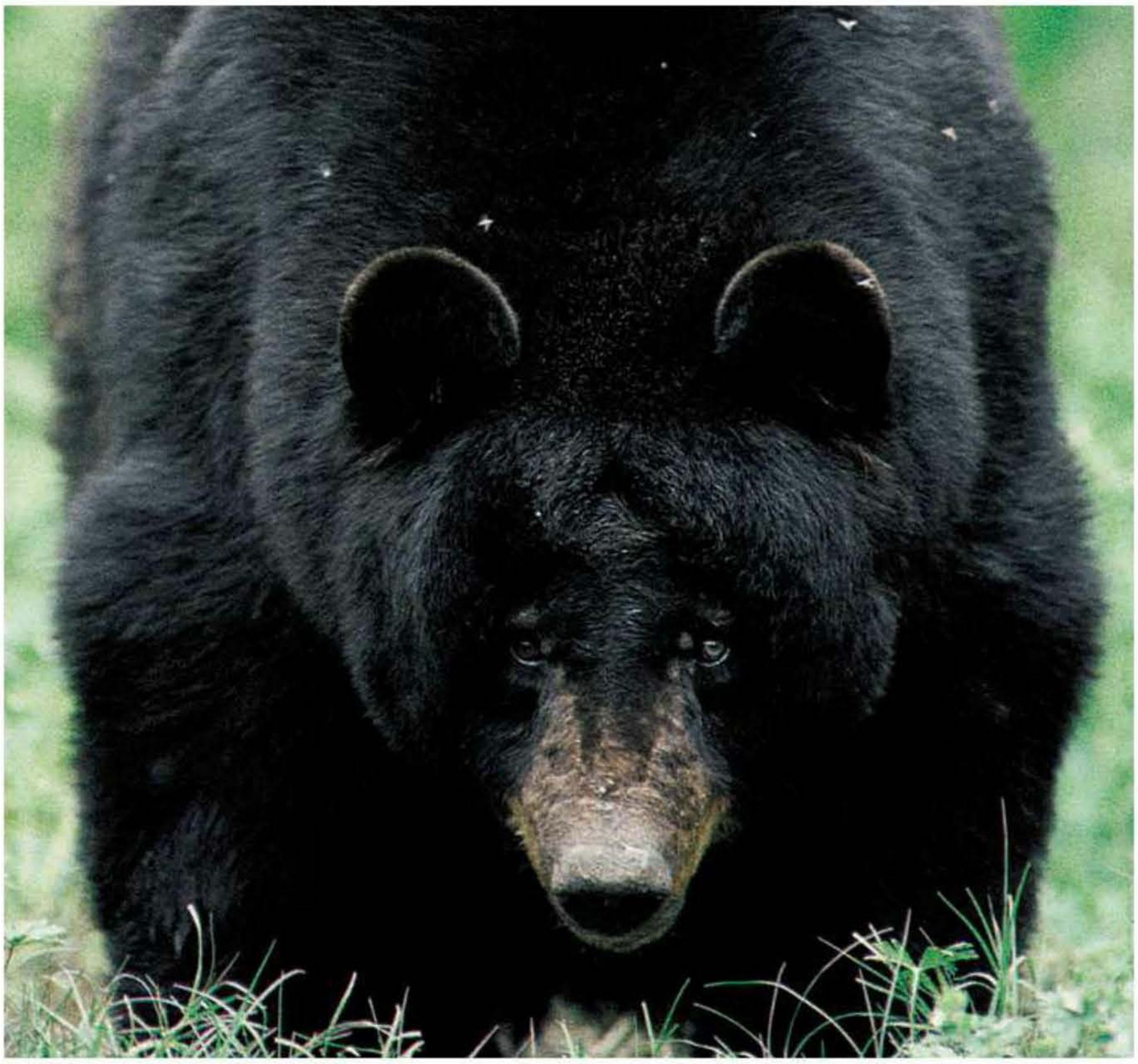

If there's one thing a hungry bear wants more than anything else, it's an easy meal. And he's used to taking food away from smaller predators. Black bear will come to a call, but bears have a short attention span. On the way in, he may stumble across a berry patch or a spawning salmon. Keep the sound rolling to keep him on the move. And give him time. Depending on how far he's got to travel, you may see the bruin in a few minutes or an hour. Keep the wind in your favor and your confidence high. Commit to spending an hour at each call set.

TIPS FOR FINDING BEARS

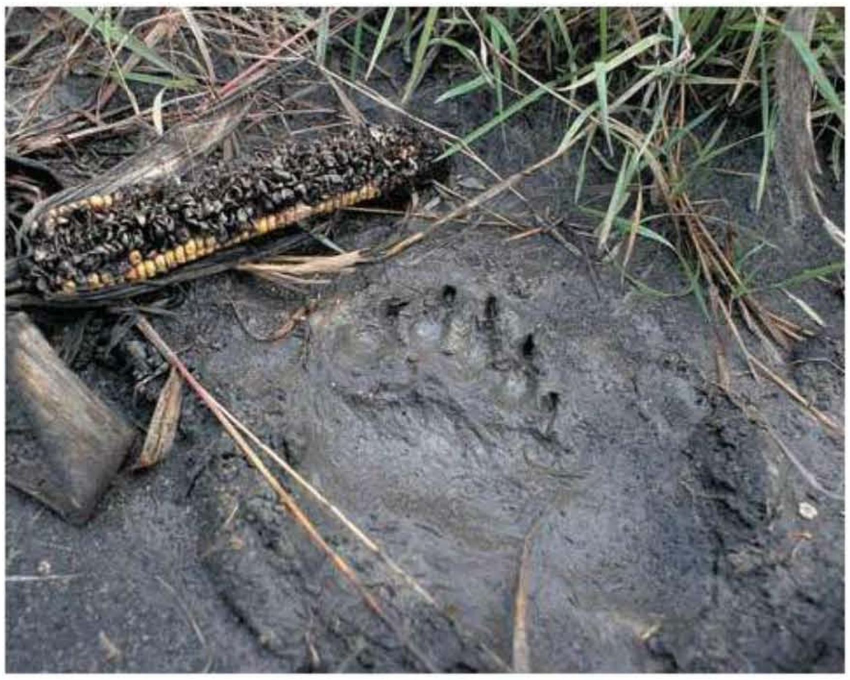

Tracks (left) have distinct claw marks and measure 3½ to 5 inches (9 to 12.7 cm) wide. The front foot is 4 to 5 inches (10 to 12.7 cm) long; the hind foot 6 to 7 inches (15.25 to 17.75 cm).

Damaged trees or bushes are common feeding signs. A bear rips down a branch with its claws, then strips off the fruit or berries.

Droppings reveal what bear have been eating. They often contain berries, grass, hair and wood eaten along with insects.

Bait black bear with food scraps. Cover each bait station with logs to keep out smaller scavengers. Bait several different locations before the season, then select a stand near a bait-site that bear visit regularly.

Follow hounds on foot or horseback until they tree or corner a bear. Hunters scout a potential area to locate fresh tracks, then release the dogs. Even if the scent is cold, good dogs pick up the bear's trail.

WILD BOAR

The Spanish brought the first domestic hogs to California in the sixteenth century. Today, their descendants can be hunted in many places in the South and in California and other western states.

This isn't your average agricultural species. Adult wild hogs tip the scales between 100 and 200 pounds (45 and 90 kg). Boars can reach 375 (169 kg). Most European/feral crosses have a coat of long, coarse, dark hair. Boars grow wicked tusks and are not the kind of animals you'd like to run into on a narrow trail.

Wild hogs come in all shapes and sizes. Long, coarse, dark hair is a characteristic of wild swine, wherever they are found. The ears are small, erect and hairy and tufted at the tips. Tails are long, straight and mule-like, in comparison with domestic pigs. They are lean, with shoulders higher and wider than their hindquarters. A wild boar may stand up to 35 inches (89 cm) at the shoulder. Wild hogs also have longer snouts than domestic swine and longer, sharper tusks (up to about 3 inches/7.6 cm in length).

Wild hogs are also faster runners than their farm-bound kin. They can hit speeds of up to 35 mph (56 km/h).

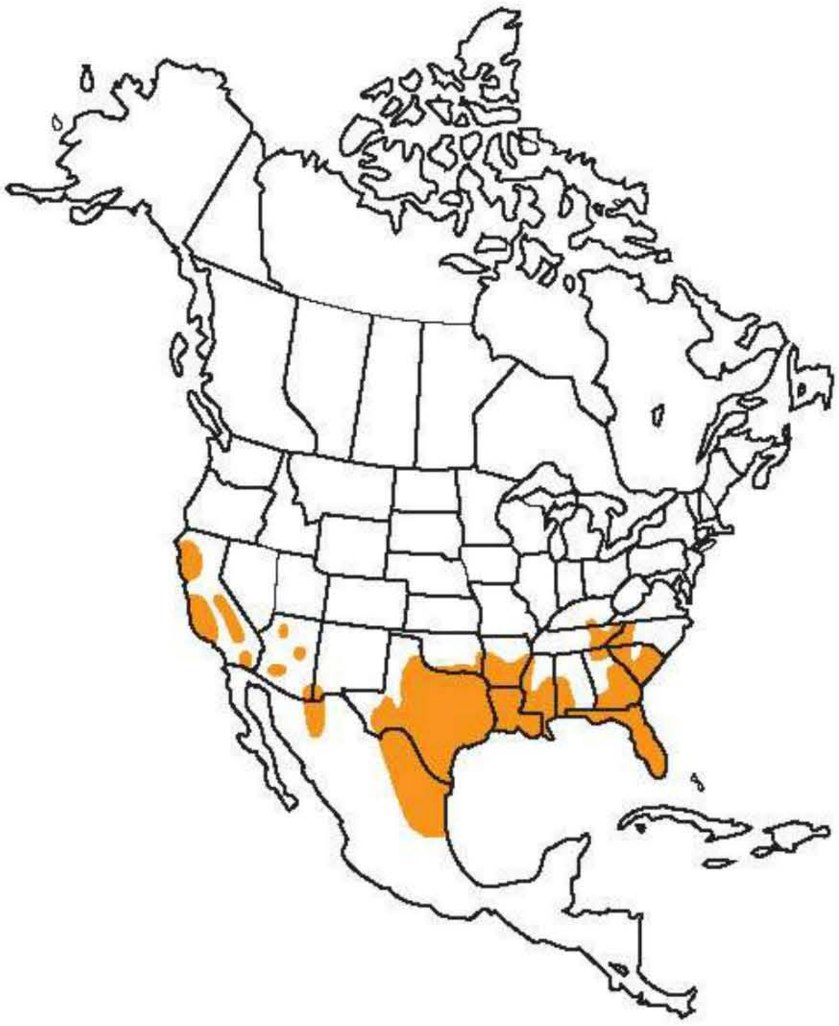

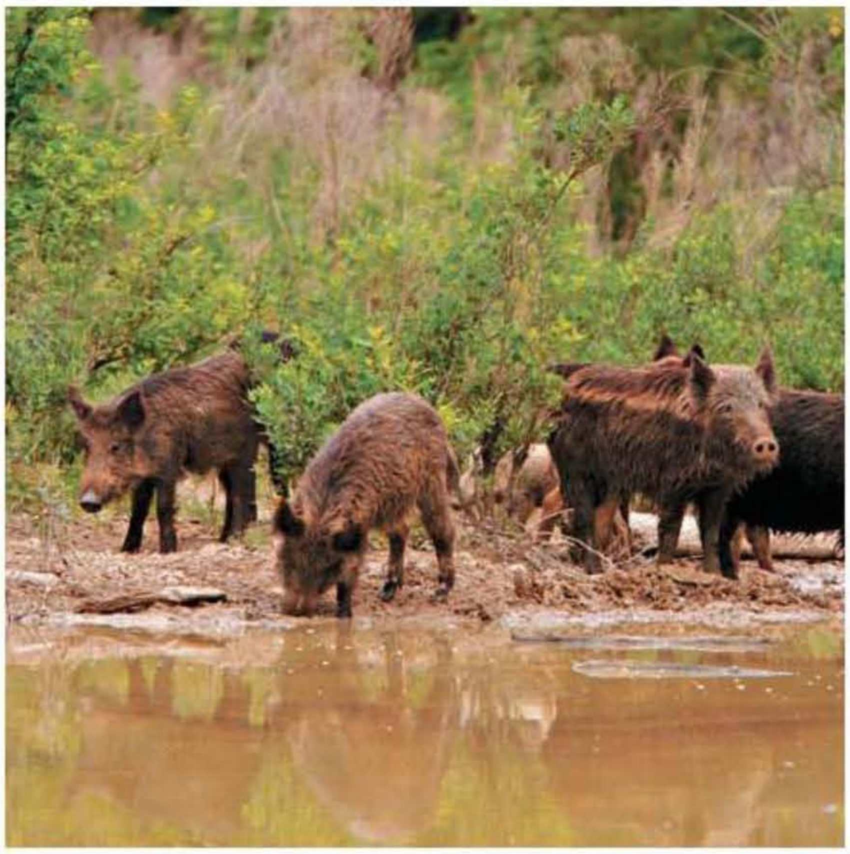

In periods between rains, you can scout for hogs at watering holes and fresh wallows. Such places will be visited by hogs at least once a day. A word of warning—bag your boar before he's all covered with fresh mud.

They eat everything from acorns to alfalfa, to rattlesnakes and bird eggs. And they multiply. Starting at six months old, a sow can produce piglets at the rate of two litters per year.

There is no limit on hogs in many states. You need a hunting license and you may need a tag. And all your hunting skill.

Spot-and-Stalk for Swine

A spotting scope can help locate hogs, but often, binoculars are all you'll need. You can watch from one vantage point or move from hill to hill. Your best bet is in the first hour and the last hour of the day. Generally, the herd is bedded down for the day, but may get up to move around a little at noon.

Feeding swine don't stay in one place very long. If you spot animals a long way out, watch them long enough to get an idea where they're headed. Then meet them there.

For two or more hunters, it's often better to split up to cover more ground. But hunting should be conducted in a low-impact manner. Stay on roads and established trails and out of bedding areas. Spook the herd and they're likely to leave the area or go nocturnal. When you crest a hill, go easy. Take a few steps then glass the habitat that has just opened up to view.

Hogs aren't known for superior visual acuity, but too often hunters don't give a hog's eyes enough credit. As much as possible, use the terrain and the cover of the trees to get in close enough for a shot.

Wind direction is the biggest consideration. Let them smell you and those hogs will be heading for cover in a hurry. Better to constantly check the way the breeze is blowing and plan your hunt and your stalks to keep the wind in your face.

Hogs are creatures of habit. The same trails are used day after day, season after season, dependent on the food sources. In ranchland, you'll see dozens of broad trails leading from bedding areas to feeding areas. You'll find coarse black hair on the barbs of a fence and deep split-toed tracks at watering areas. Choose a stand where you can keep track of two or more trails at once.

In periods between rains, scout for fresh wallows. Such places are visited by hogs at least once a day.

Guns and Loads for Wild Boar

The wild boar is not easy to kill. It has a tough shoulder hide called a shield that protects it from thorns. Your bullet must penetrate this armor and retain enough energy to destroy the heart or lungs. Often the hog doesn't drop at the first shot.

Many shots have to be taken on the run. A well-balanced rifle can help you get on target fast. Keep the scope dialed down to 3× for quick acquisition of the target. Many hunters choose a 30-caliber like the .308 or the .30-06 while handgun hunters prefer the 44 Magnum.

TYPICAL BOAR BIOLOGY

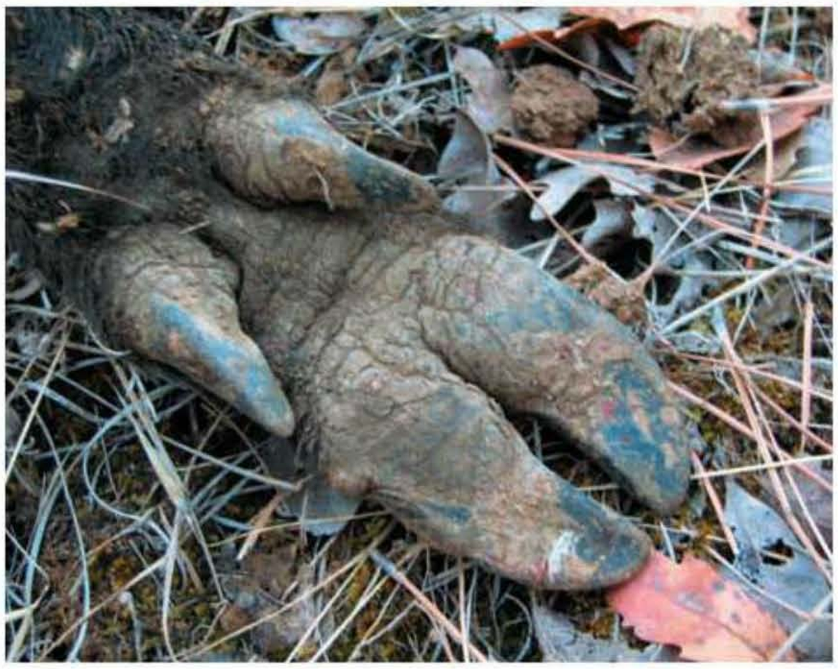

Males and females both have tusks. They may grow to 5 to 9 inches (13 to 23 cm) long.

Wild hogs have a split toe (right). Large hogs leave the imprint of both the hooves and the dew claws.

SIGNS OF WILD BOAR ACTIVITY

Wallows (left) are used by hogs to coat themselves with mud to keep insects at bay. Look for tracks around water holes and the presence of stirred-up mud in the water.

Two signs of hog activity: rooting for grubs and dried mud on the side of the tree where a big boar scratched.

Between rains, hogs head for water every day. Look for tracks around water holes.

Chapter 5

HUNTING SMALL GAME

The popularity of small game hunting stems from the relative ease of finding game and access. With a little scouting, you can probably locate rabbits, squirrels, raccoons or woodchucks within a few miles (km) of your home. Even with a minimum of equipment, you can find great sport and bring home a tasty meal.

Even the novice has a reasonable chance of bagging rabbits or squirrels. But this does not mean that small game hunting is easy. Your success will improve as your level of skill increases. You can apply the skills you gain to other types of hunting.

Small game hunters have many opportunities to enjoy their sport. Most hunting seasons last at least six months and some are continuous. Because you can find these animals close to home, you can make frequent short trips, which is difficult when hunting most other types of game. Almost all states and provinces have liberal bag limits.

Most hunters use shotguns or rimfire rifles chambered at .17 or .22 caliber. But some prefer a combination gun with a .22 caliber rifle barrel on top and a 20-gauge shotgun barrel on the bottom. Combination guns are ideal for small game because you can fire one barrel at standing animals and the other at running targets. Other hunters employ handguns to bring home the game. You can also use small-caliber center-fire rifles, muzzleloading rifles, shotguns and bows with blunt-tipped arrows.

A pair of waterproof leather boots is a good investment. With the possible exceptions of brush pants, a burr-proof jacket with a game pouch, binoculars and a small knife for field dressing, you need little other equipment.

Many hunters compare rabbit and squirrel meat to chicken. Raccoon tends to be oily and has a distinctive flavor of its own. Small game tastes better if you remove the entrails soon after killing the animal.

COTTONTAIL RABBIT

Rabbits don't require a lot to thrive: a little food, water, shelter and escape cover. They can find everything they need in a few acres (ha) of ground. Best of all, rabbits may be hunted close to home.

Each year, hunters throughout North America bag more than 40 million cottontail rabbits. This staggering total results from the cottontail's tremendous reproductive rate and its ability to adapt to a wide range of habitats and foods.

Cottontails breed in the spring and summer, producing up to eight litters, each with three to six young. They prefer brushy edges and woodlots, but can live almost anywhere, with the exception of dense forests. They eat practically any type of green plant. When green vegetation dies back, they switch to twigs and bark.

Rabbits start feeding before dawn and continue for two or three hours. They resume feeding at sunset. They move about most on calm, sunny days. Rain or wind drives them into heavy cover.

A cottontail spends most of the day sitting in a form, a shallow depression in grass or snow. The grass eventually wears away or the snow melts down and compacts. Often a form is concealed by overhanging grass or other type of overhead cover.

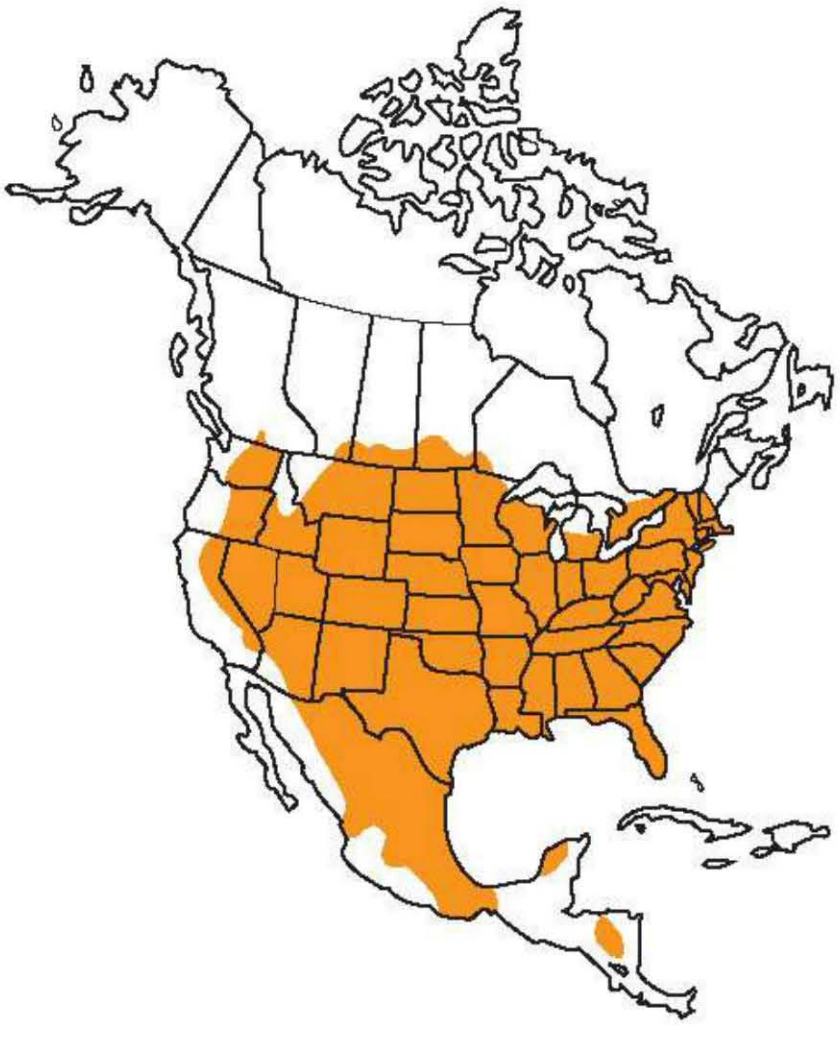

Rabbits use their superb hearing to sense impending danger. To escape, they bolt away on established travel lanes. Cottontails run in an elusive, zig-zag pattern, but their speed is not as fast as a jackrabbit. They average 12 to 15 mph (19 to 24 km/h), but can reach 20 mph (32 km/h).

Normally, cottontails do not run far. Rather than run straight away, a rabbit circles so it can stay in familiar territory. When frightened, it often slips into a woodchuck burrow or brush pile.

You can often bag cottontails by walking through cover, looking for rabbits in their forms. Upland bird hunters flush rabbits by moving in typical walk-and-wait fashion. Like deer, a cottontail becomes nervous when a nearby hunter stands motionless and bounds from its resting spot.

Hunting with dogs offers an interesting and effective alternative. Most hunters use slow-moving hounds such as beagles and bassets. When the dogs get close, the cottontail begins moving in a large circle and eventually passes within shooting range of the hunter. Almost any dog will chase cottontails, but if it works too fast, the rabbit dashes under a brush pile or down a hole.

Rabbits may contract a bacterial disease called tularemia, which causes them to behave listlessly and eventually kills them. But the disease is rare. One researcher found it in only 2 of 12,000 rabbits he examined. Nevertheless, refrain from shooting rabbits that move slowly or otherwise behave unusually. Tularemia can be transmitted to humans who eat or handle the flesh of infected animals.

To hit zig-zagging cottontails, most hunters use shotguns with modified or improved cylinder chokes and No. 6 shot. But when rabbits are in their forms or feeding in the open, reach for a .22 rifle.

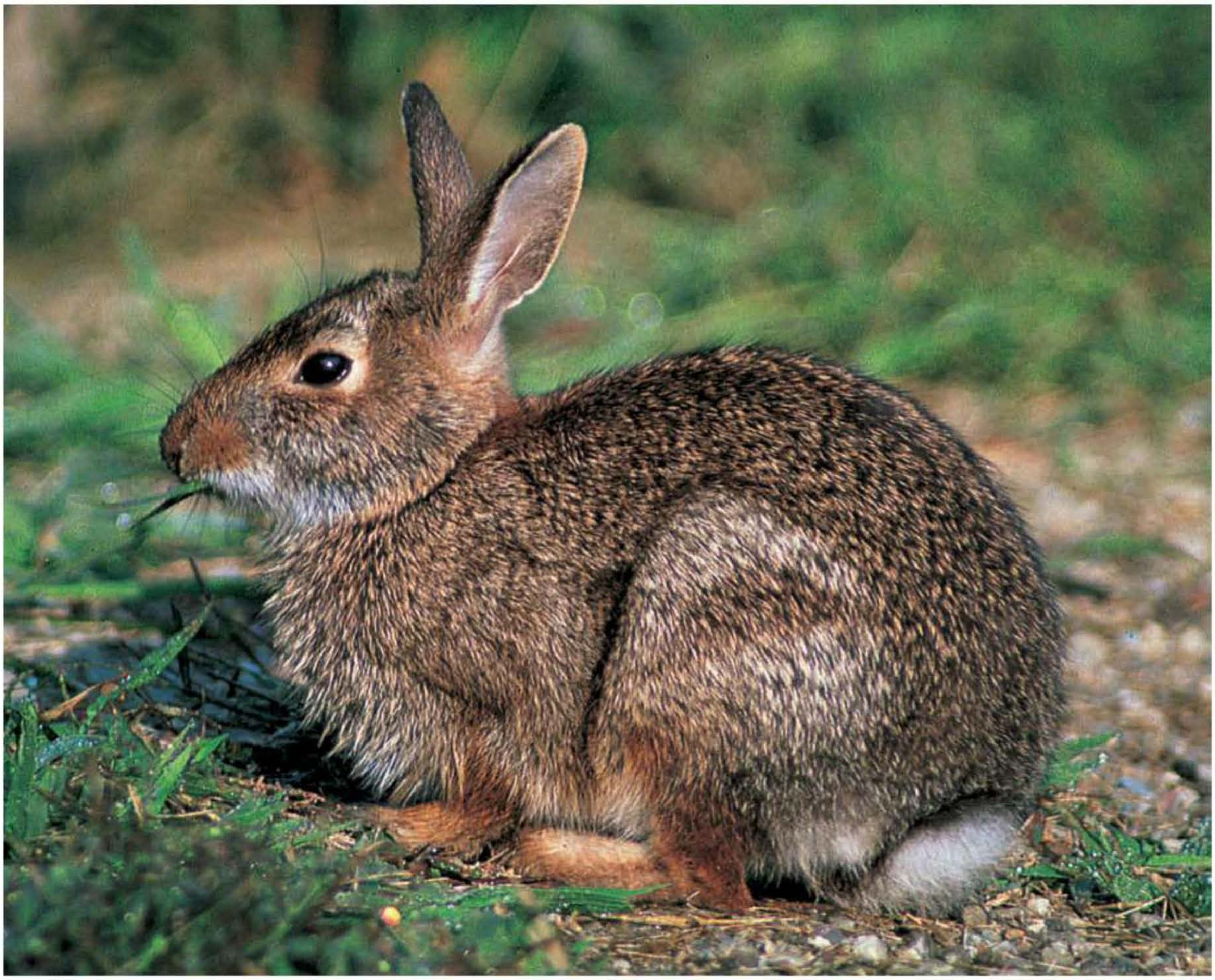

Cottontails are named for their fluffy white tail. The fur on the upper part of the body is grayish-brown with black tips. The undersides are white. Adults measure 14 to 19 inches (35.5 to 48.25 cm) long and weigh 2½ to 3½ pounds (1 to 1.5 kg).

WHERE TO FIND COTTONTAILS

Strip cover, like brushy fencelines and hedgerows, makes ideal cottontail habitat. The brush and high grass provide food and cover.

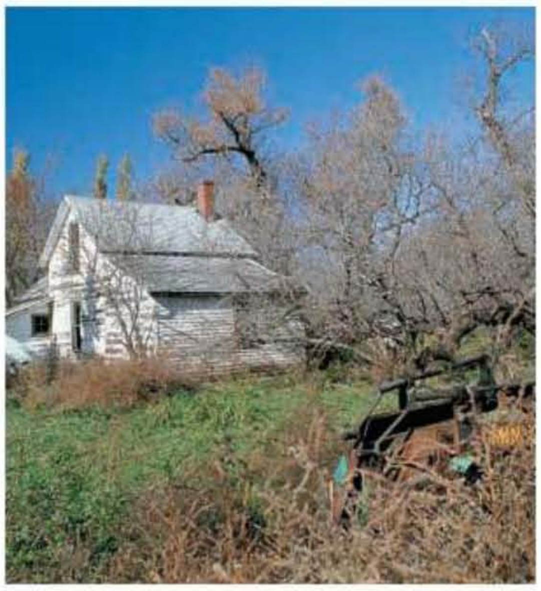

Abandoned farms offer a variety of hiding spots. Look for rabbits around groves, under old machinery or in tall grass or brush.

Brush piles provide good escape cover. A fox or owl would have little chance of reaching a rabbit beneath the logs and sticks.

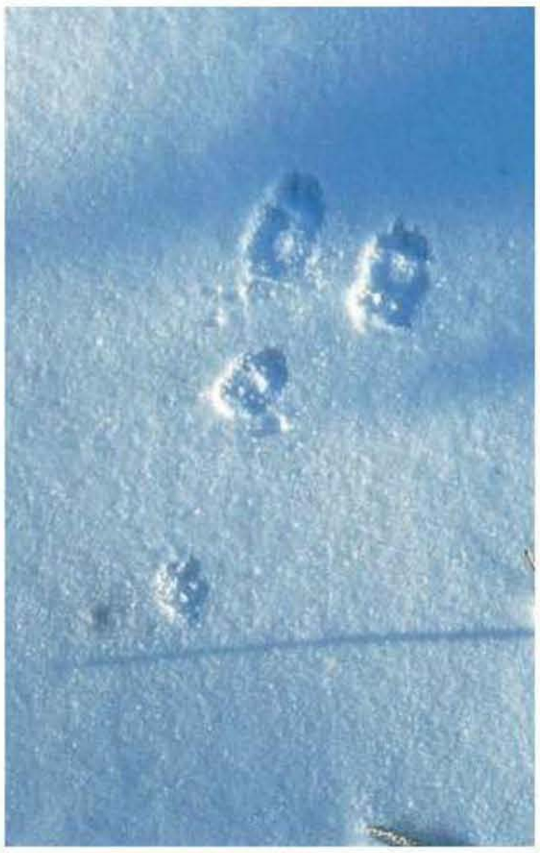

Tracks of cottontails have side-by-side hind prints ahead of smaller front prints. One front foot falls ahead of the other.

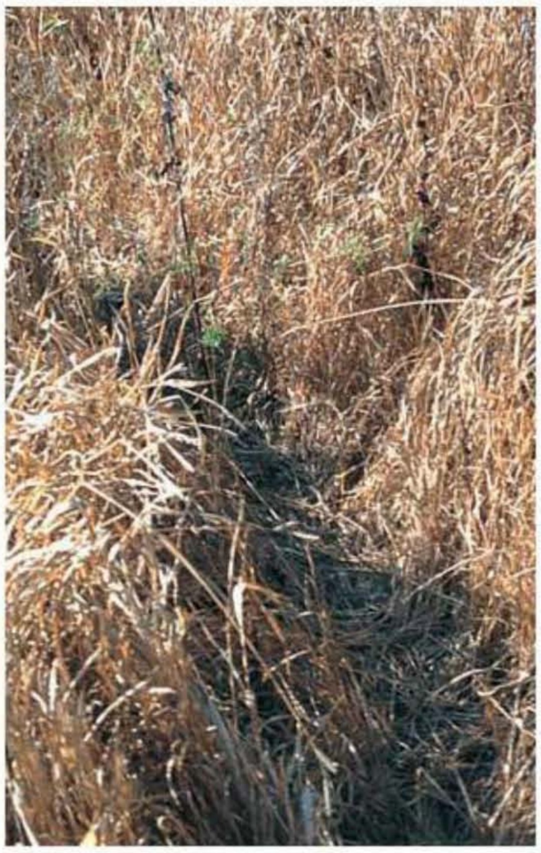

Runways in tall grass serve as escape routes. Rabbits usually follow the same paths through cover, eventually matting down the vegetation.

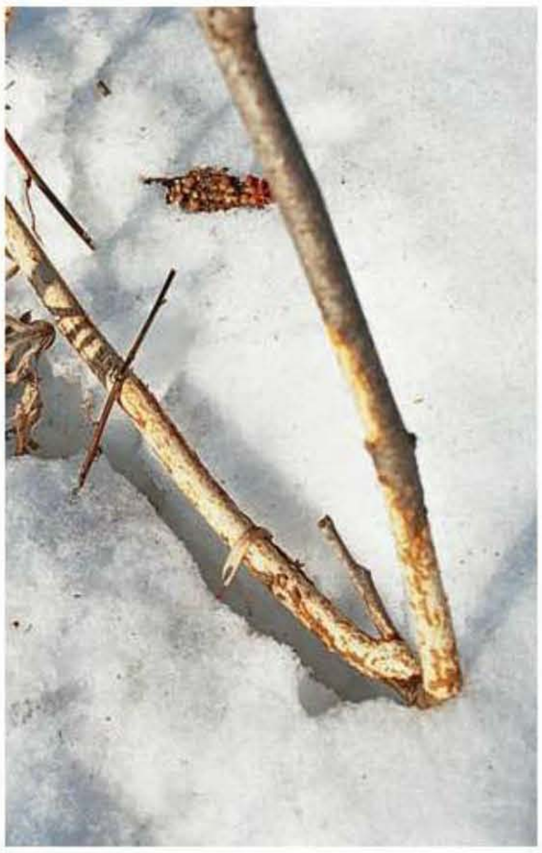

Feeding signs include girdled shrubs or saplings and cleanly-snipped twigs. Cottontails prefer the bark of sumac and fruit trees.

HOW TO HUNT WITH HOUNDS

Release the dogs where you find plenty of sign. Beagles have excellent noses and scour the ground thoroughly to find fresh scent. They methodically follow the trail, slowly pushing the rabbit ahead.

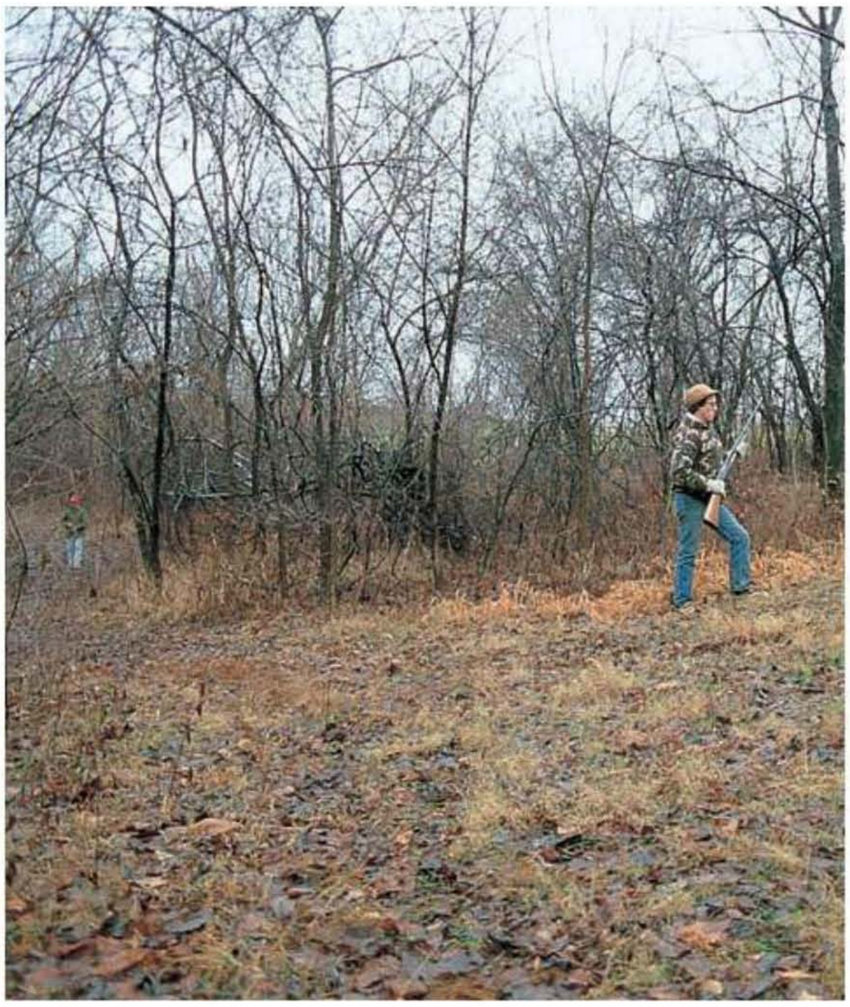

Wait in the area where the dogs first detect fresh scent. A rabbit usually circles and returns to the spot where it was flushed. If it does not circle, try to predict its escape route, then attempt to intercept it.

OTHER COTTONTAIL HUNTING TECHNIQUES

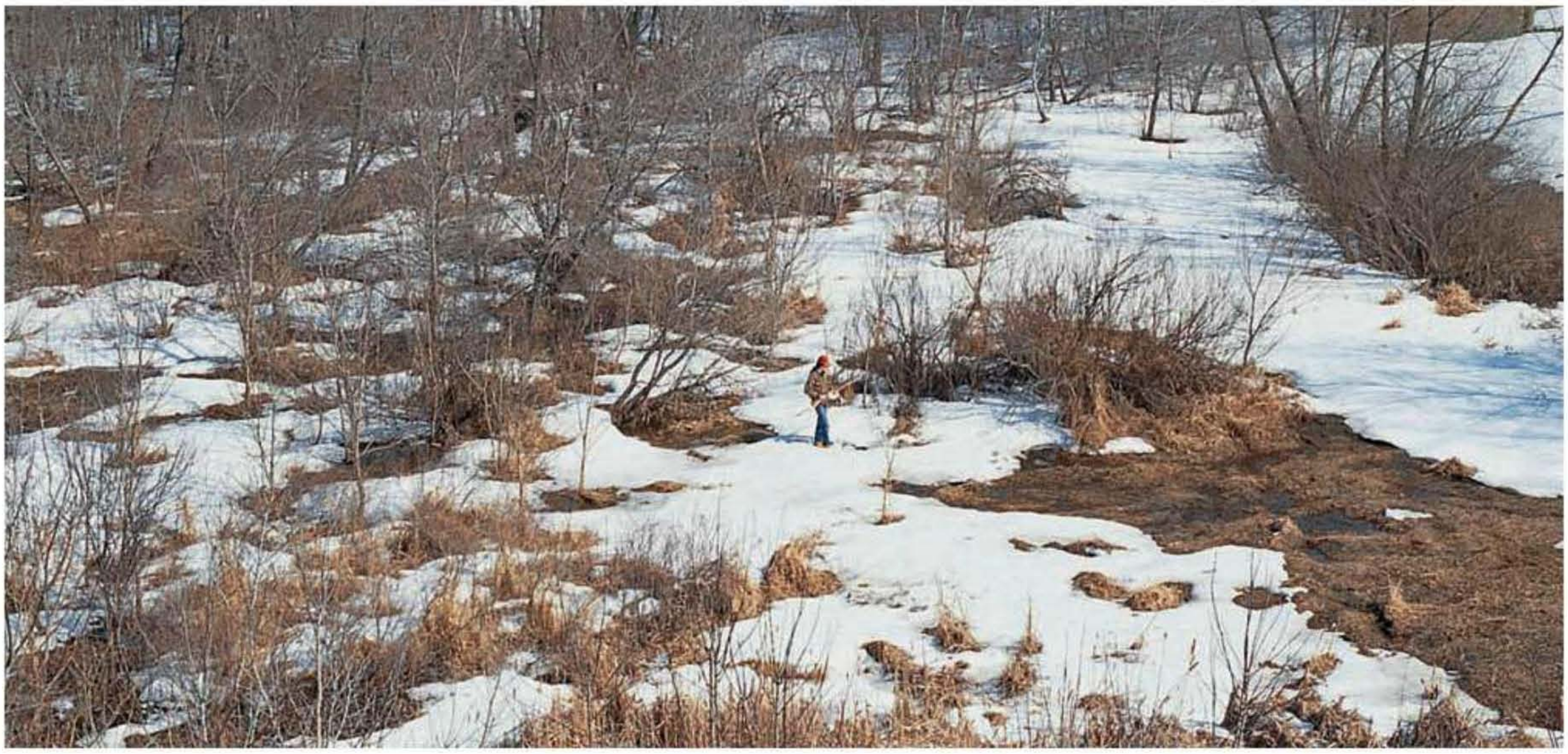

Walk around the form in ever-widening circles. Look for the rabbit in clumps of brush, around the bases of trees, in tall grass or in any dense cover. Continue walking until you cover the entire area within a 100-foot (30 m) radius of the cottontail's form. Follow any fresh tracks you encounter.

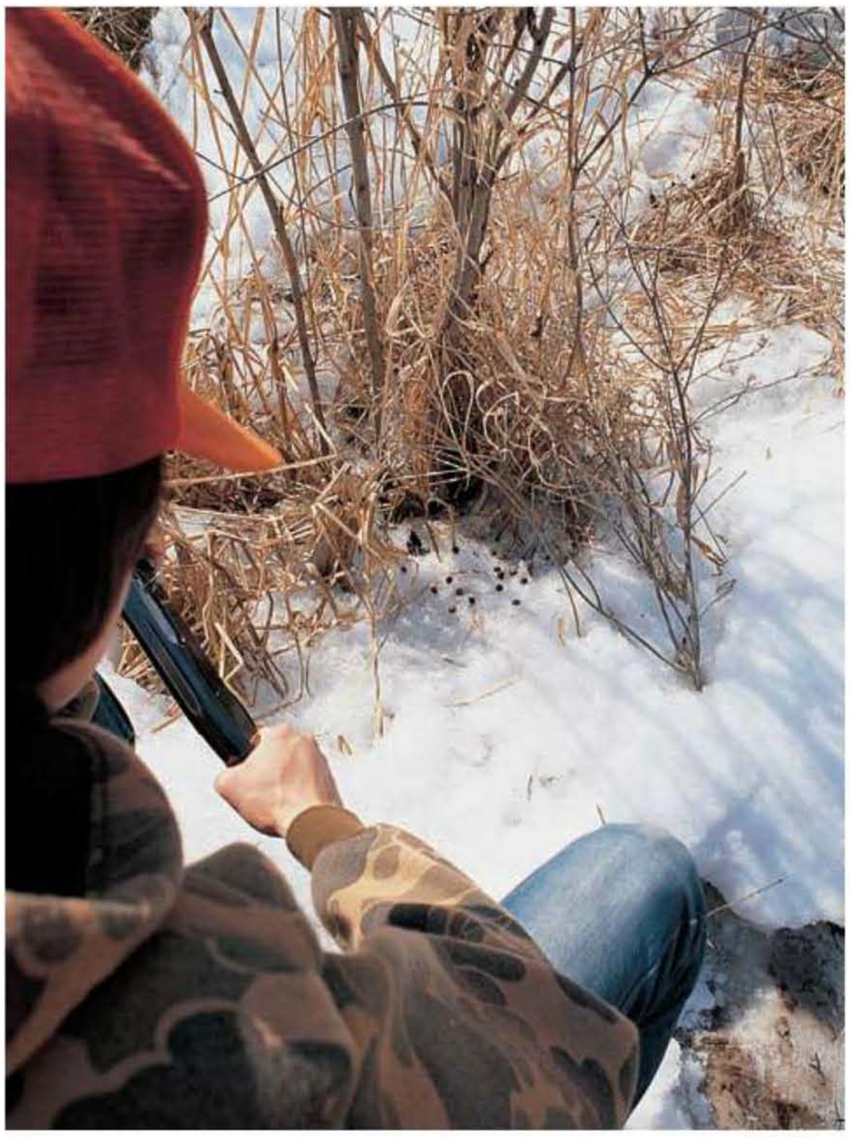

Locate a form with fresh sign. Droppings are round and about ⅜ inch (95 mm) across. A rabbit seldom moves far from its form. Watch for rabbit tracks as you walk a strip of cover like a brushy fence line. A retriever or flushing dog will scare up tight-holding rabbits.

Hunt around abandoned farmsteads. Look for rabbits as you round the corner of a building and be ready for a quick shot. Flush cottontails by beating a brush pile with a stick. Or climb on top of the brush pile, jump up and down and yell to scare out rabbits.

JACKRABBIT

Black-tailed jackrabbits are found in low-lying semi-desert areas of the United States and Mexico. Jackrabbits (which are technically hares, not rabbits) may also be found in typical cottontail country and around tree plantations in more temperate climates. Night feeders, jackrabbits are best hunted in the low light of early morning and in the evening. Landowners are often happy to allow hunting to keep populations from destroying valuable alfalfa and other crops.

The black-tailed jackrabbit grows to a length of about 24 inches (60 cm) and adults weigh between 1 and 2 pounds (0.45 to 0.9 kg). They can reach speeds of up to 35 mph (55 km/h) and can leap up to 5 yards (6m) in a single jump.

One of the most productive methods for hunting jackrabbits is the drive. Groups of three to five hunters work the best. The hunters walk in a line through good rabbit habitat, spread out with about 10 yards (9 m) in between. The preferred tools are .22 rifles and 20-gauge shotguns. Rabbits will be pushed out ahead of the group or hold tight to run the opposite direction after the drive has moved past.

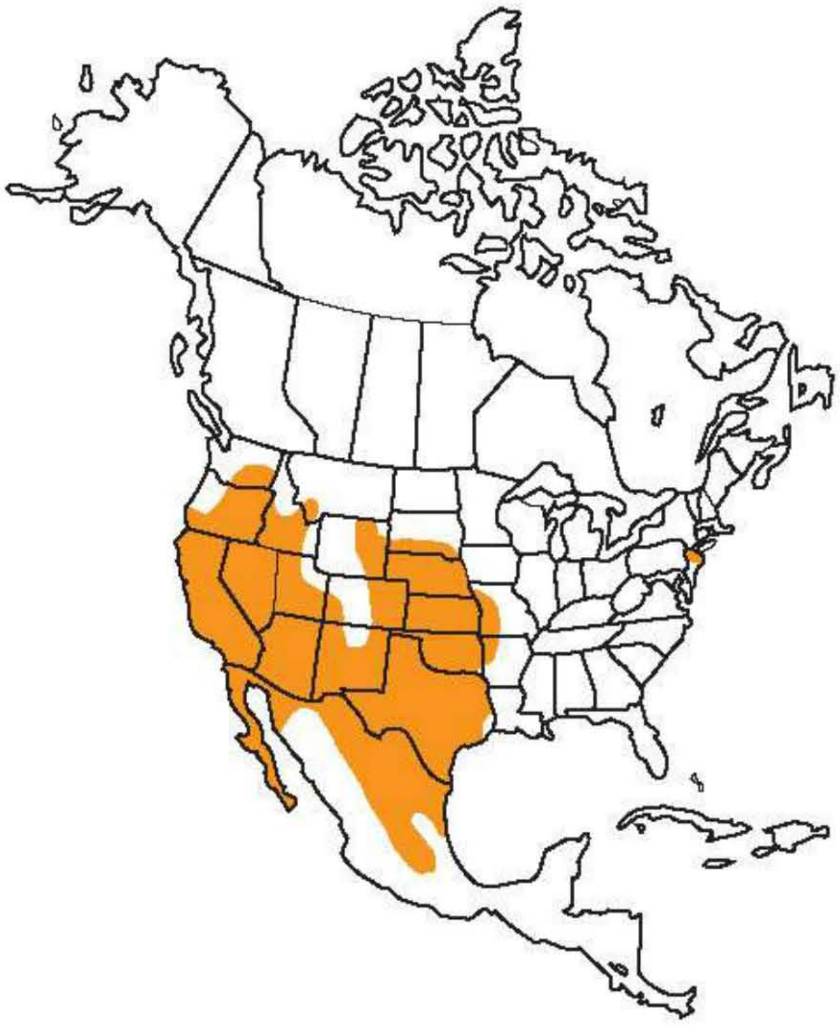

For safety's sake, high-visibility blaze orange is essential on these hunts and the group leader should ensure that the line stays intact.

When flushed, jackrabbits zig and zag, putting brush and distance between the hunter and hunted. Unlike cottontails, jackrabbits don't often circle back, they're headed for the horizon. Be ready, because you'll probably not get another chance.

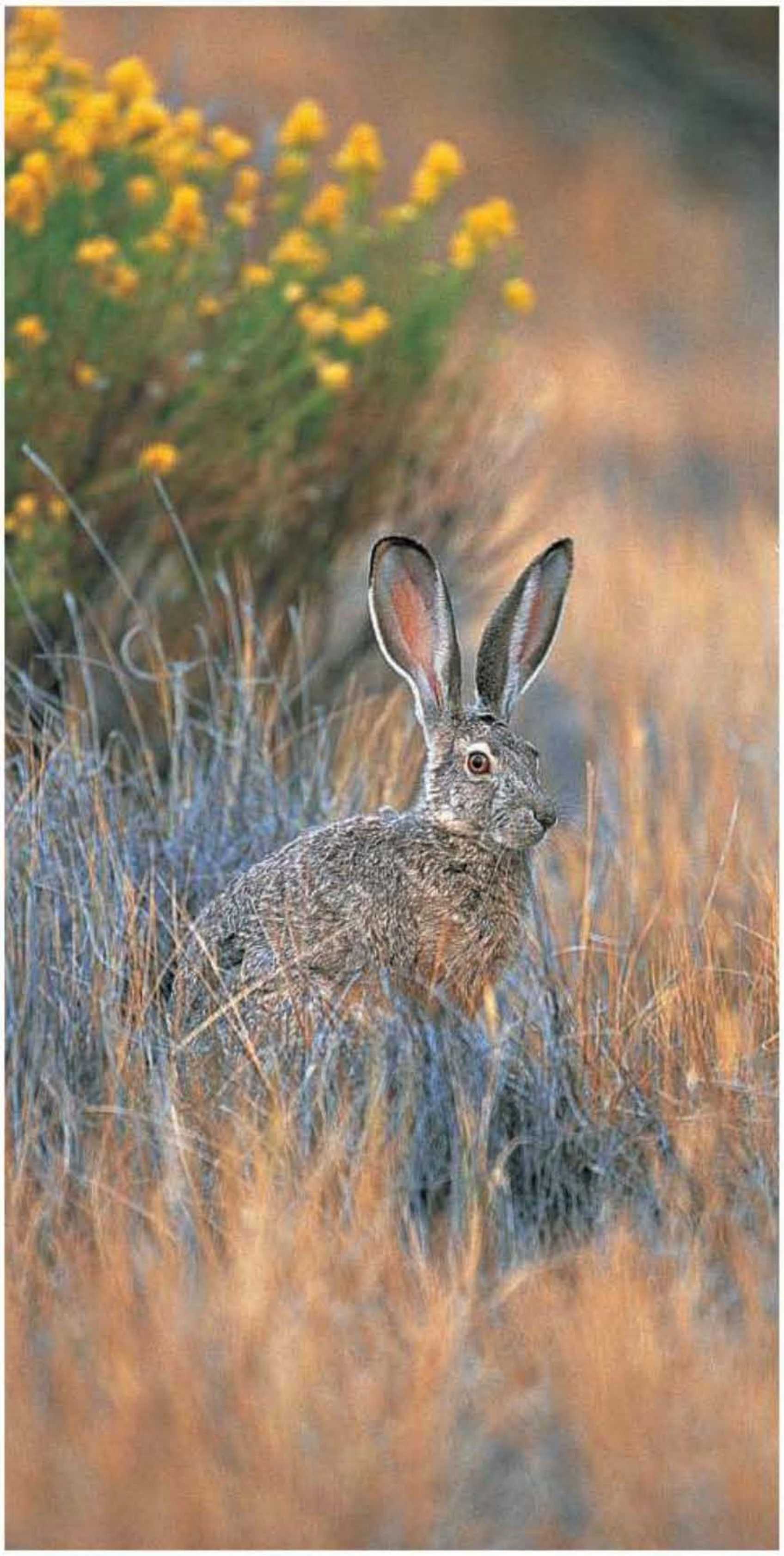

Although blacktails are largely solitary animals, they may be found in large groups at certain times of the year, especially when feeding on a plot of alfalfa.

SNOWSHOE HARE

The snowshoe is named for its oversized hind feet, which provide a large surface area to support the animal on soft snow.

Snowshoes behave much like cottontails and eat similar foods. But they are more likely to feed at night. They prefer conifer swamps and young hardwood forests rather than brushy edges. Snowshoes are larger, faster and better jumpers.

A snowshoe can run up to 30 mph (48 km/h) and change direction in the middle of a leap, so a tight choke is less effective. To increase their shooting range, many still-hunters use .22 rifles.

Look for a snowshoe's black eye and black-tipped ears when hunting in snow. You can spot the animals more easily after a mid-winter thaw, because their white bodies stand out against bare ground.

Hunting with hounds works as well for snowshoes as it does for cottontails. Like cottontails, snowshoes will stay ahead of a dog and circle back toward a waiting hunter.

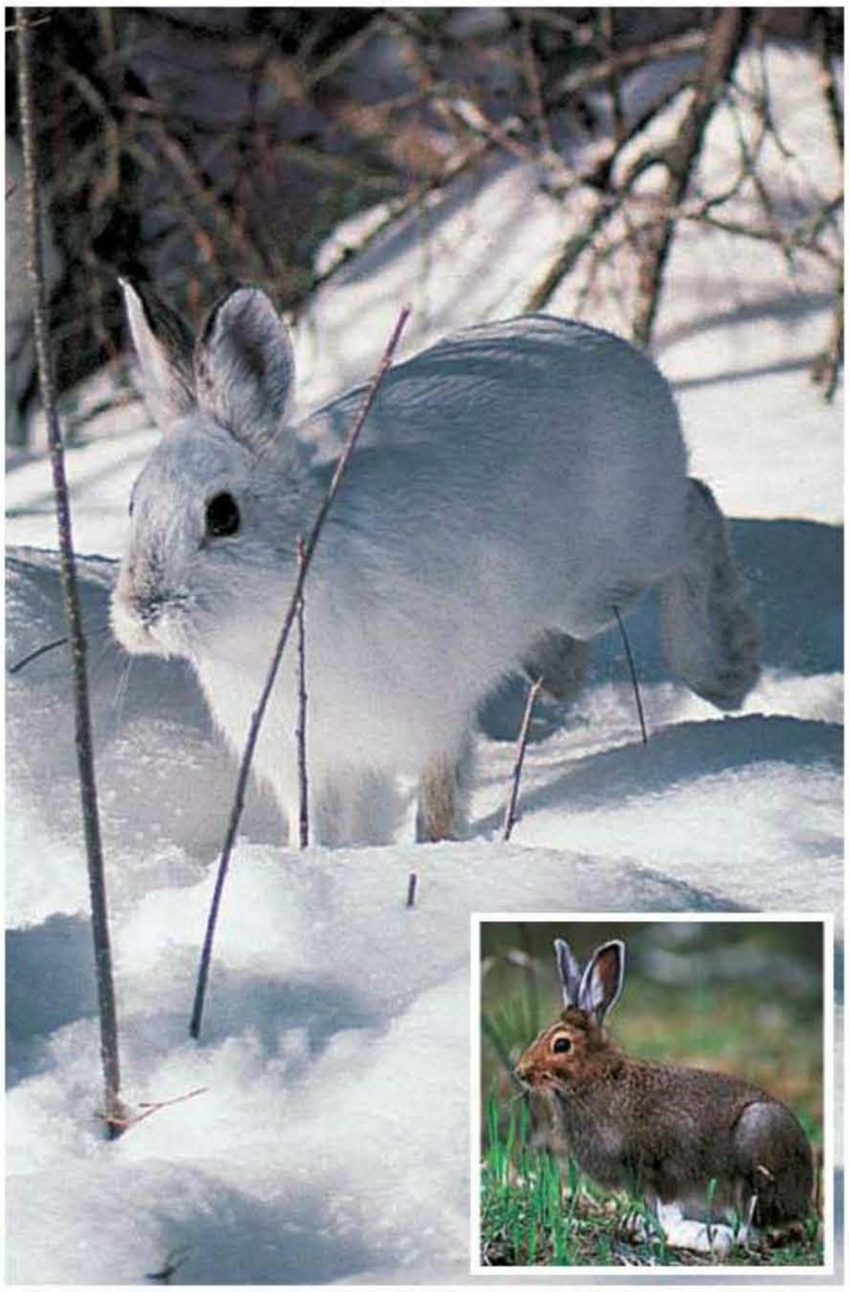

Snowshoes, also called varying hares, vary in color depending on the season. The coat is white in winter and brown in summer. Snowshoes are 16 to 21 inches (40.6 to 53.3 cm) long and weigh 3 to 4 pounds (1.3 to 1.8 kg).

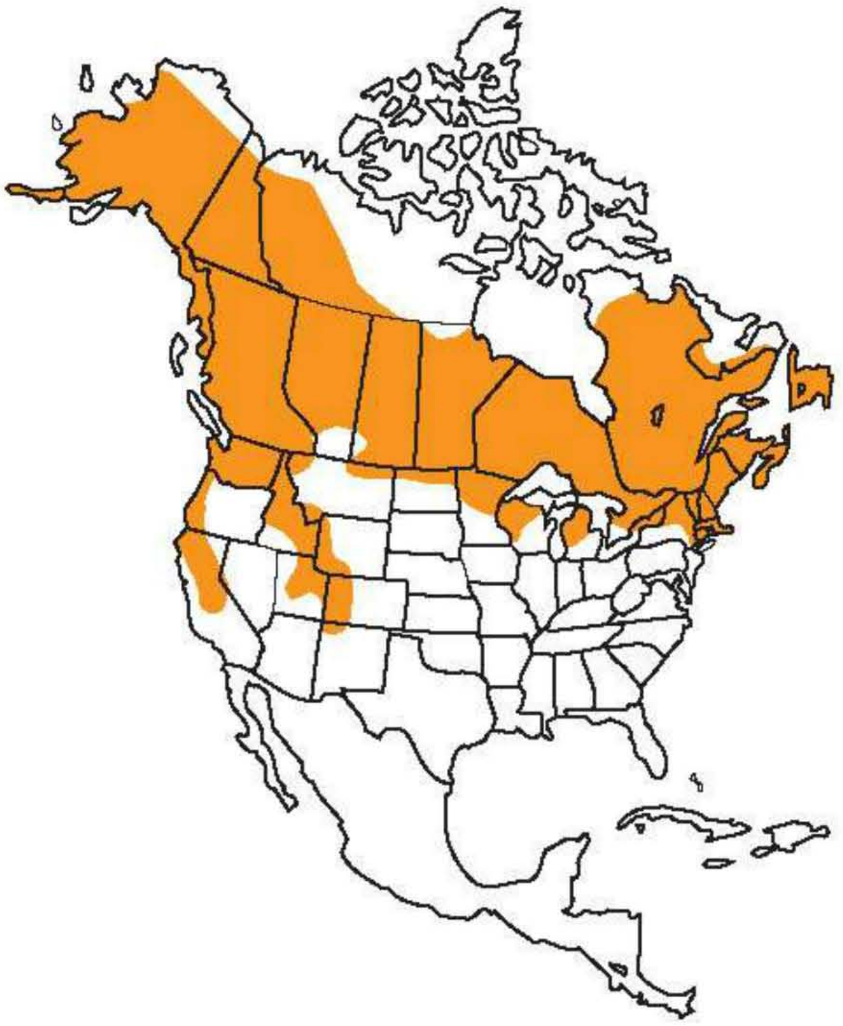

Follow tracks after a fresh snowfall. About 2 inches (5 cm) of new snow over a hard crust makes for ideal tracking. If the snow is more than 1 foot (30 cm) deep, a pair of snowshoes make walking much easier.

GRAY SQUIRREL

Because the squirrel can provide a good meal for hawks, owls, weasels, minks, bobcats, foxes, coyotes or humans, he must be alert at all times—whether on the ground or high in a tree. A gray squirrel hunt is a test of marksmanship for any hunter. Using their wide-angle vision and sharp hearing, squirrels quickly detect danger.

To avoid being seen, a squirrel moves to the opposite side of a tree trunk or limb. Or it flattens its body against the tree. Once the hunter has walked past, it scampers to its den and stays there until it feels safe.

Good habitat produces one gray squirrel per acre (ha). Thus, 1 square mile (2.6 sq km) of prime woodlands could hold over 600 squirrels. How plentiful gray squirrels are in a given year depends mainly on the previous year's acorn crop. Other foods include walnuts, hickory nuts, pecans, berries and corn.

Gray squirrels begin feeding just before sunrise and they continue for two or three hours. They resume feeding in late afternoon and may remain active until after sunset. Squirrels move about most on calm, sunny days. Extreme cold or windy weather keeps them denned up. They seldom travel more than 300 yards (274 m) from their dens.

Mature deciduous forests throughout eastern North America hold gray squirrels. They favor mixed hardwoods with an abundance of mature oaks, dense undergrowth and few open areas.

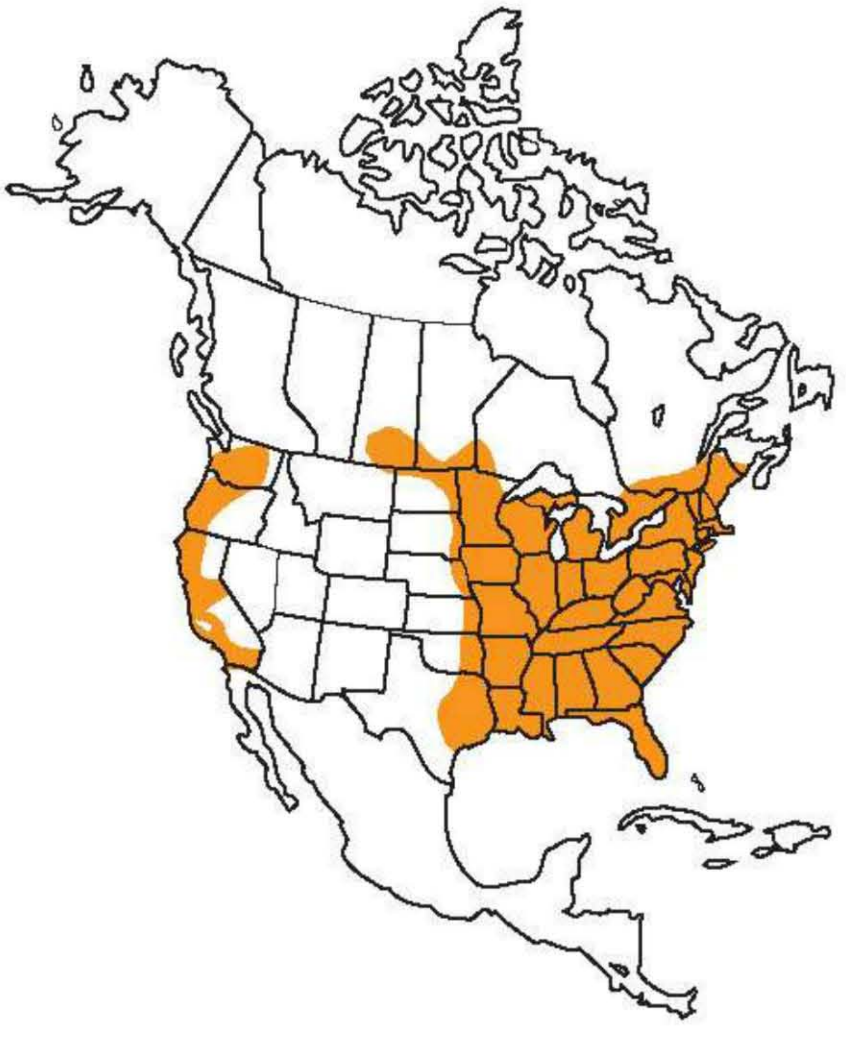

Gray squirrels have a grayish back and sides and whitish or brownish undersides. A solid black color phase predominates in some areas. Adults measure 14 to 21 inches (35.5 to 53.3 cm) from head to tail and weigh ¾ to 1½ pounds (0.3 to 0.7 kg).

FOX SQUIRREL

Named for its reddish, fox-colored fur, the fox squirrel differs from the gray in behavior and habitat. Some hunters maintain that it is less wary and easier to outwit.

Fox squirrels spend more time on the ground than grays and rely less on treetops as escape routes. When threatened, they run straight for their dens or hide behind trunks or limbs. Unlike grays, they seldom feed in early morning. They are more active during midday and stray farther from their dens. Fox squirrels can often be seen lying on a limb basking in the sun.

Many hunters refer to fox squirrels as red squirrels. But the true red squirrel is much smaller and is not considered a game animal.

Fox squirrels prefer woodlots, farm groves and strips of timber rather than forests. Like grays, they eat acorns and other nuts. In agricultural areas, corn and other crops make up a high percentage of the diet. Squirrels get most of their water from their food.

Both gray and fox squirrels make a variety of calls. A series of rapid cherks or barks serves as a warning signal to other squirrels. The familiar chatter means squirrels are approaching one another. A low-pitched chuckle signifies contentment.

Adult fox squirrels measure 19 to 29 inches (48.25 to 73.75 cm) from head to tail and weigh 1 to 2¾ pounds (0.45 to 1.2 kg).

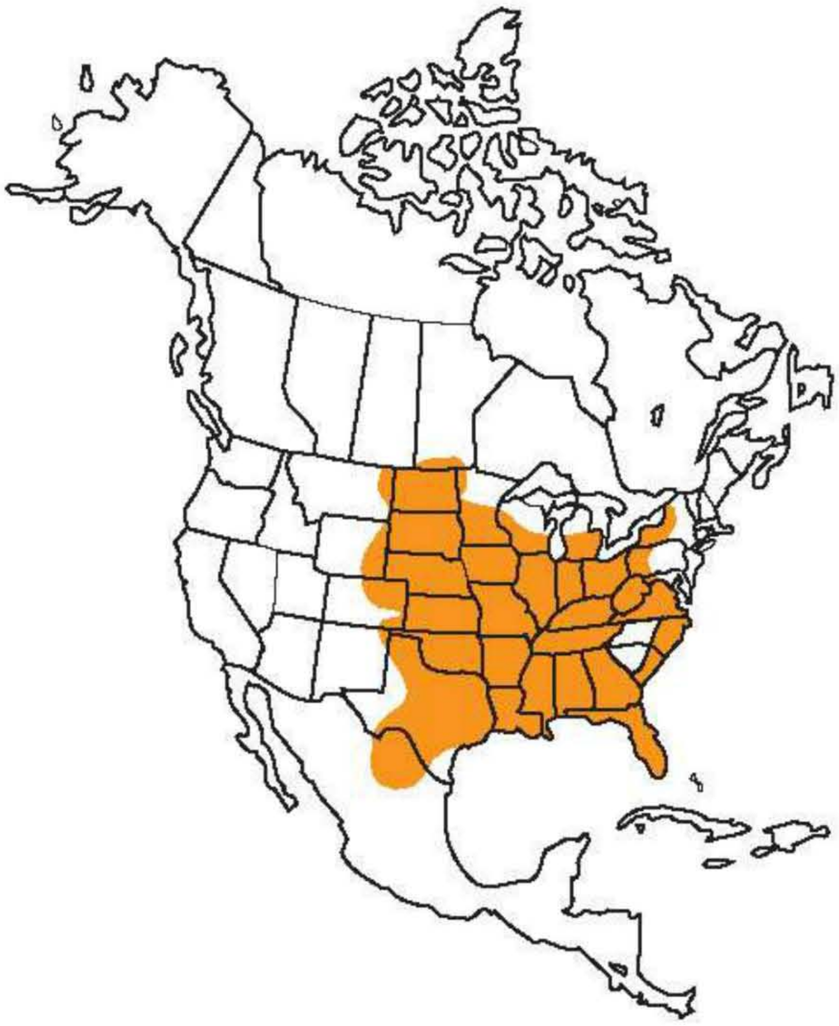

Hickory trees produce nuts that draw squirrels. You can easily identify hickories in early fall; they turn yellow earlier than other trees.

SIGNS OF SQUIRREL ACTIVITY

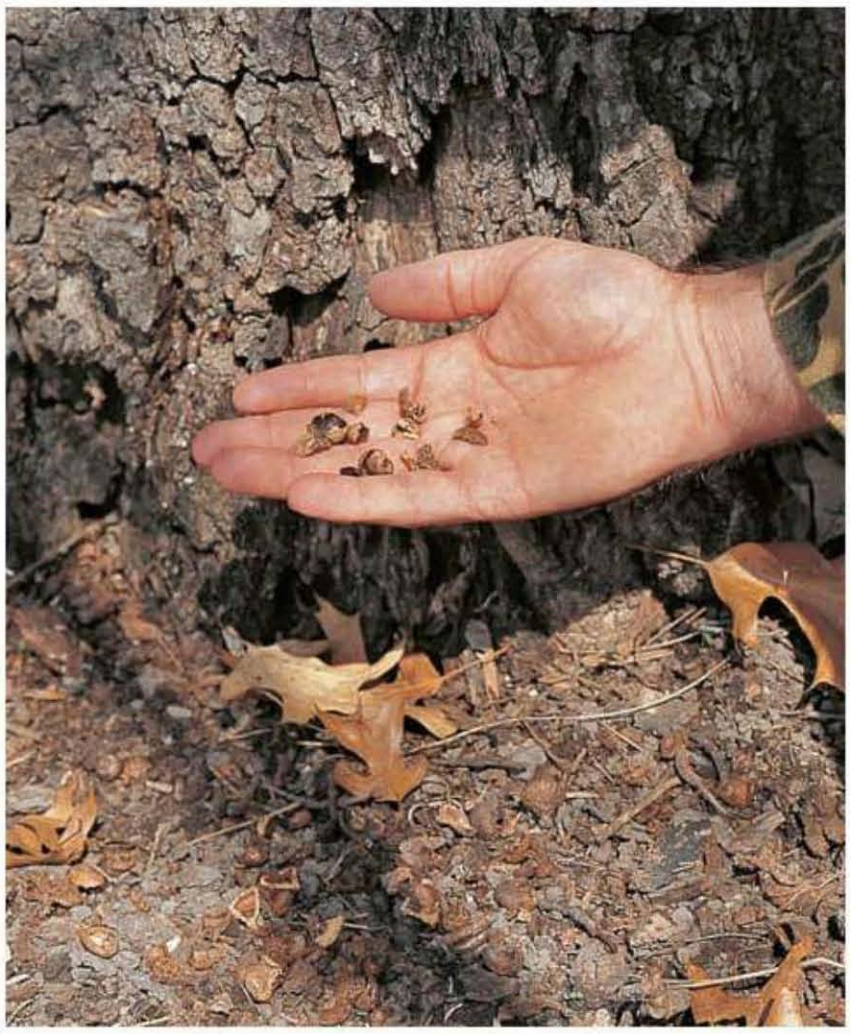

Acorn shells at the base of a tree indicate a squirrel den somewhere above. Sign in a large forest is usually left by grays.

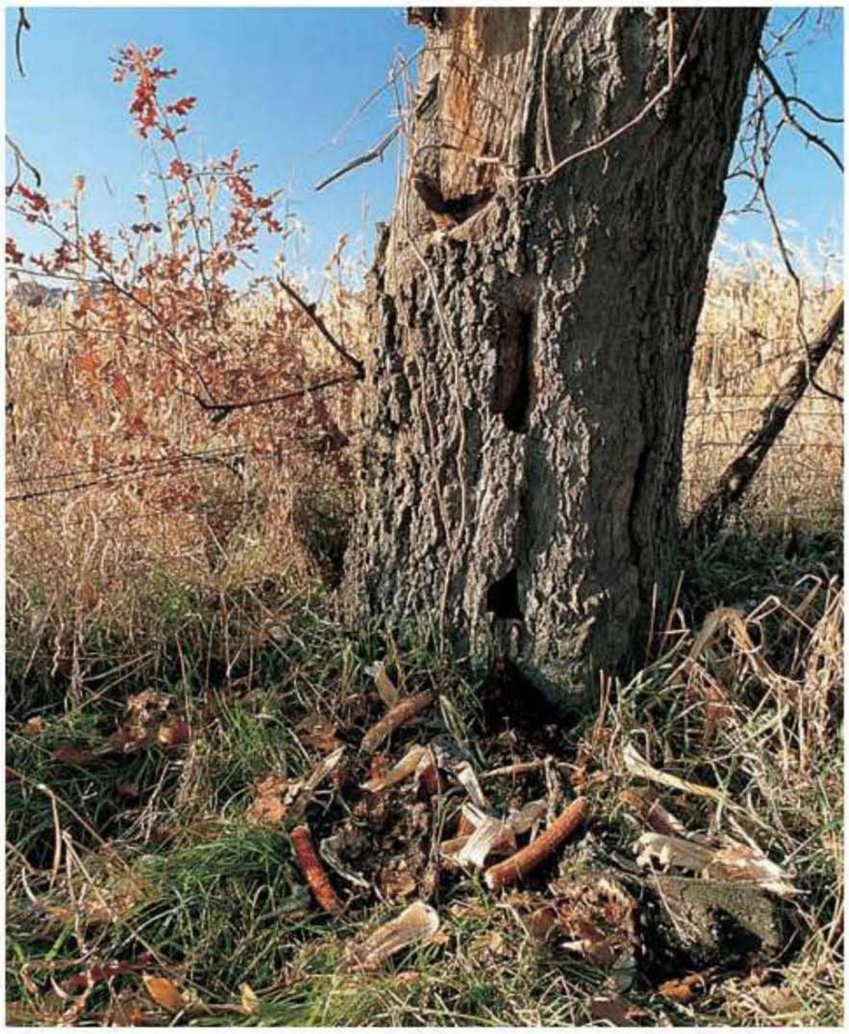

Stripped corncobs also reveal squirrel activity. Sign along a fence line or in a small woodlot is generally that of fox squirrels.

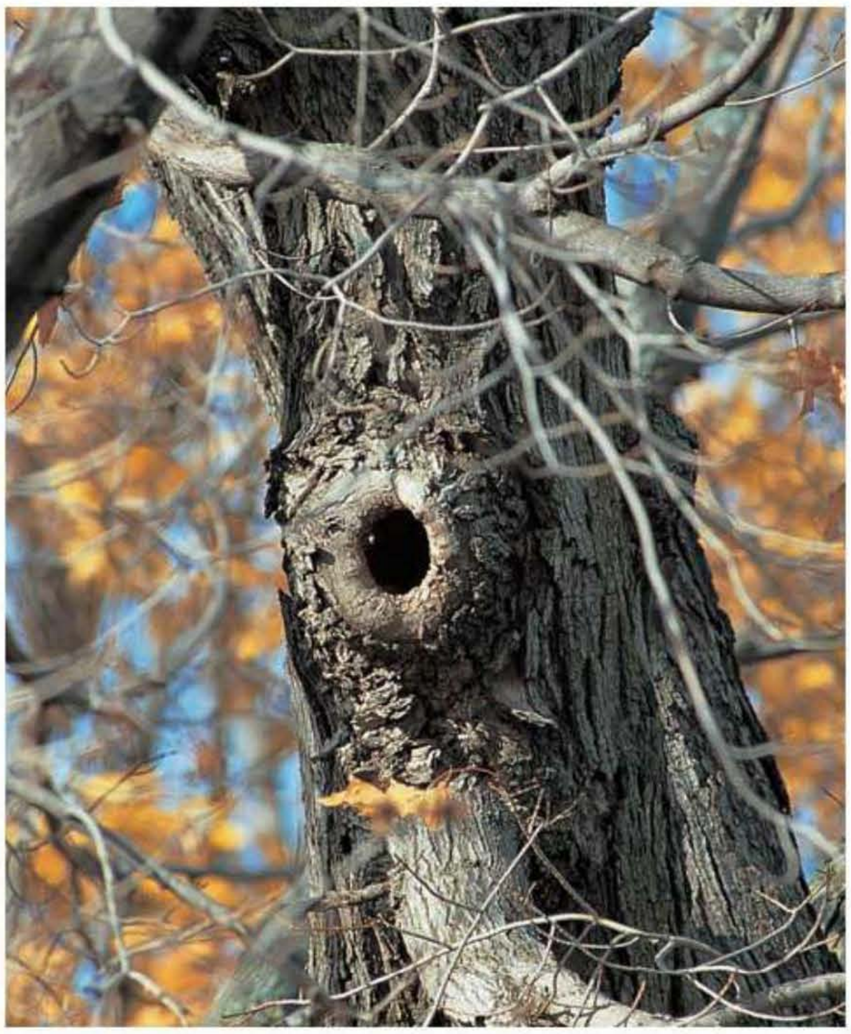

Dens are made by squirrels gnawing at small openings like woodpecker holes. Dens are about 4 inches (10 cm) wide and at least 10 feet (3 m) off the ground.

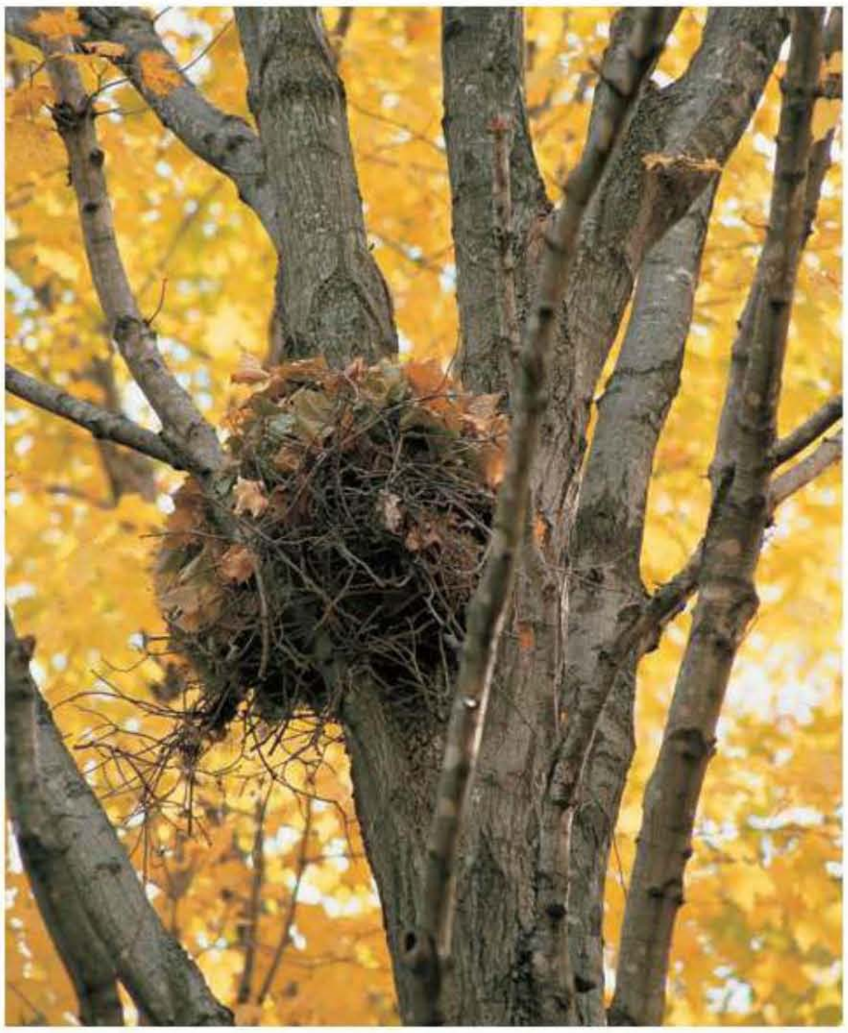

Nests consist of twigs and leaves piled into a crotch or woven into the branches. Nests are 1 to 2 feet (30 to 60 cm) across and 20 feet (6 m) or more above ground.

Hunting for Squirrels

Look for stands of oak trees or other nut-producing trees. An old apple orchard close to timberlands may also produce good hunting.

Mast crops very from year to year and squirrels locate near this food. During late summer squirrels eat wild cherries, poplar buds, maple seeds, mulberries, gum berries and fruits, or the seeds of trees that ripen early.

As fall progresses squirrels start their heaviest feeding of the year, preparing for winter. They cut or feed on the nuts of pecan, hickory, walnut, oak or beech trees. Piles of small pieces of chewed nut hulls dropped on the ground under these trees, called cuttings, are evidence of feeding.

In agricultural areas, squirrels concentrate near cornfields; look for sign where squirrels have climbed cornstalks to eat the kernels off the cob or where they have pulled down the cobs.

During the warm days of early and mid-fall squirrels may start feeding at night and continue until after daylight, resting during the warm part of the day and feeding again in late evening. Feeding and storing nuts for winter lasts longer into the daylight hours as the season progresses. Later in the season squirrels spend more time on the ground gathering, digging and storing.

Hunting can be good during the squirrel's major breeding period, which starts in December or January. Squirrels lose their normal caution as they begin courtship, paying less attention to hunters.

The still-hunter watches for the slightest movement, the twitch of an ear, the flicker of a tail, the shine of an eye and listens for the cutting of teeth on nuts or the skittering of claws on bark. Freeze and investigate each noise. Windy or rainy weather covers your sound, allowing you to move through the woods with more stealth.

Stand-hunting can be the most productive tactic when the woods are too noisy to still-hunt. Sit quietly in a feeding area, watching for movement. Remain still after shooting one; sometimes squirrels resume normal activity after a few minutes.

Most squirrel hunters use .22 rifles with scopes up to 4×. Shotguns are a good choice; carry a .410 or a 20 or a 12-gauge with No. 6 shot.

Hiding on the opposite side of a tree trunk is the squirrel's favorite method of eluding the hunter. No matter where the hunter is, the squirrel is always on the other side of the tree.

HOW TO STAND-HUNT FOR SQUIRRELS

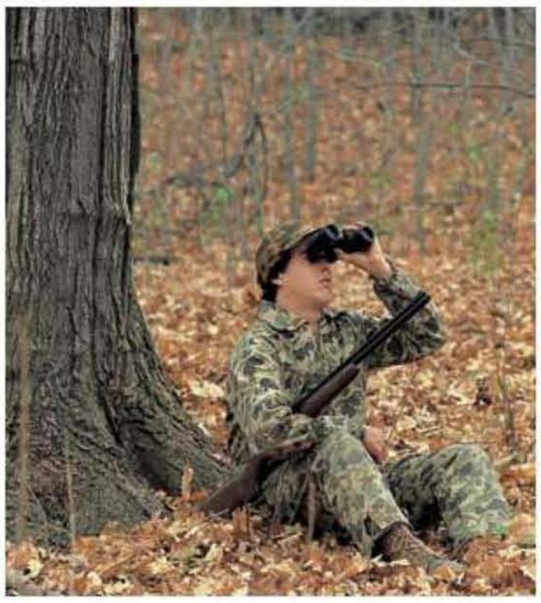

Sit quietly, glassing trees with binoculars. Choose a spot with the sun at your back. That way, you can see squirrels but they may not see you.

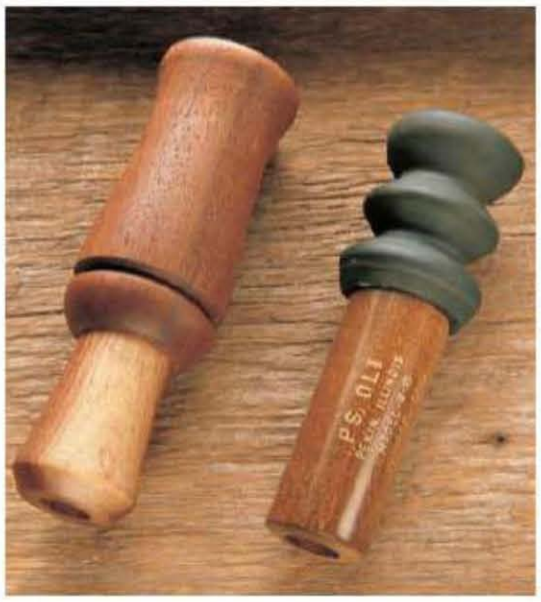

Call to coax squirrels into revealing themselves. Calls include a blow call (left) and a bellows call (right). Either can be used to produce barks and chatters.

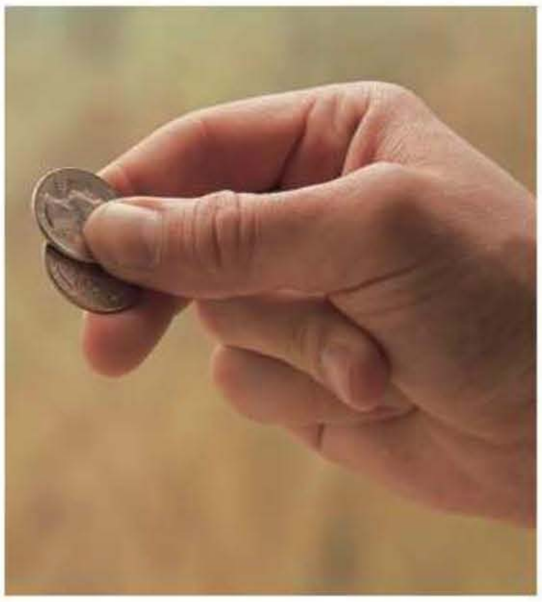

Click coins together to imitate a scolding squirrel. Place one coin slightly lower than the other and rapidly snap the lower coin.

OTHER SQUIRREL HUNTING TECHNIQUES

Team-hunt with a companion. With this method, squirrels cannot hide on the opposite side of a tree without being seen. Be ready to shoot the moment a squirrel moves to your side of the tree.

Toss a stick or rock to the opposite side of a tree. Or tie a string to a bush on one side of the tree, then move to the other side and tug it sharply. The motion will frighten squirrels to your side.

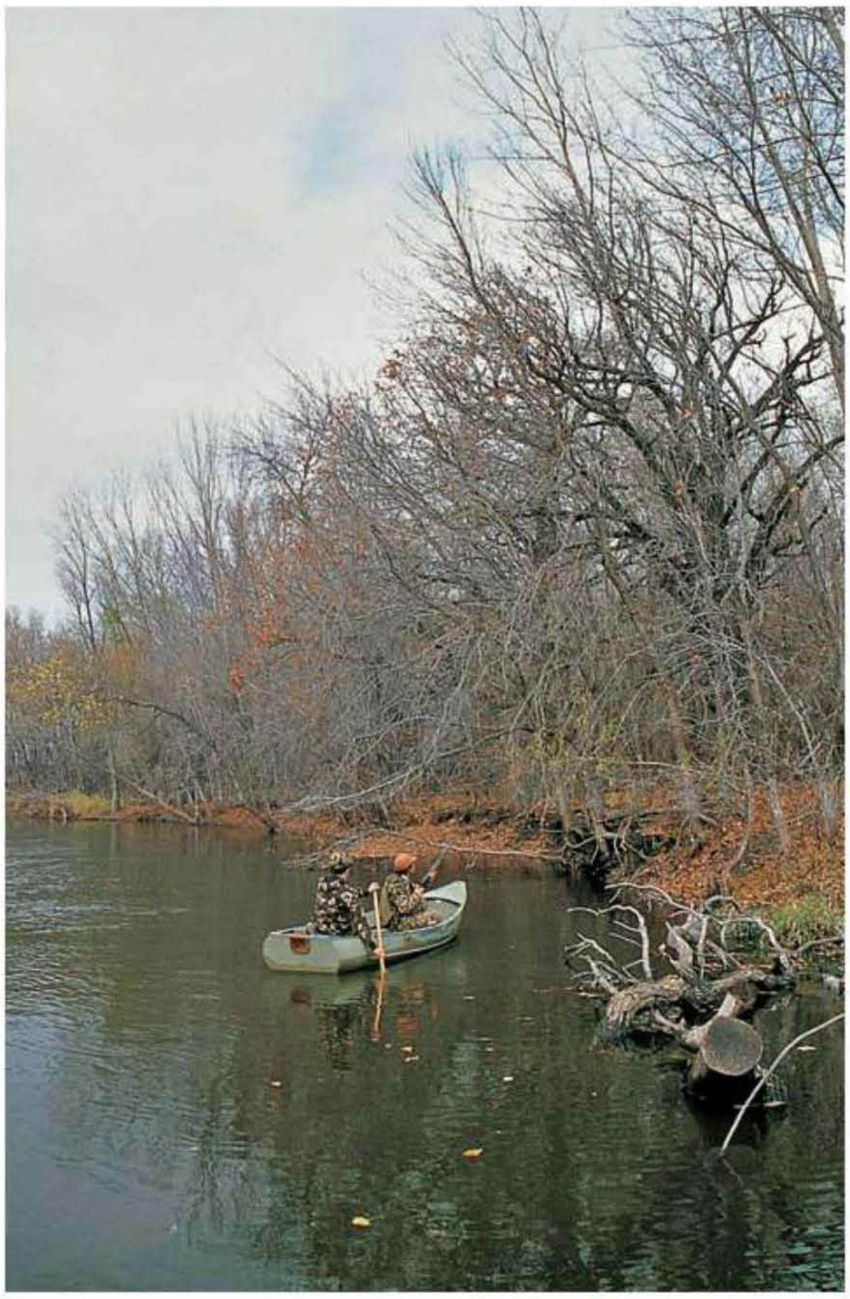

Float-hunt along streams bordered by nut trees. The extra sunlight and moisture result in good nut crops that attract squirrels.

PRAIRIE DOG

The prairie dog is a creature of the western plains, found in the United States, Canada and Mexico. It grows to a length of 12 to 16 inches (30 to 40 cm). Prairie dogs live in large colonies or dog towns that can cover hundreds of acres (ha) of land.

They live in groups that consist of one male and two to four females. Females have one litter a year and give birth to an average of three or four pups. Prairie dogs feed on grass, insects and broadleaf plants. A great tunneler, this creature may dig 15 feet (4.5 m) downward and 100 feet (30 m) in a lateral direction. A tunnel system generally has several escape tunnels. The prairie dog has color vision and can detect danger at a great distance. To warn each other, prairie dogs use vocalizations.

Hunters can use rimfire rifles, but most hunters find that a center-fire rifle is a better choice to make shots on animals out to 400 yards (366 m). The 204 Ruger, 223 Remington and 22-250 Remington, stoked with thin-jacketed bullets are great options for prairie dog hunting.

A variable rifle scope is a big help out on the fields. And a bipod or shooting sticks can make the difference when every little bit of accuracy counts. Often the quarry is almost hidden behind a mound of dirt, leaving you a target no bigger than a silver dollar.

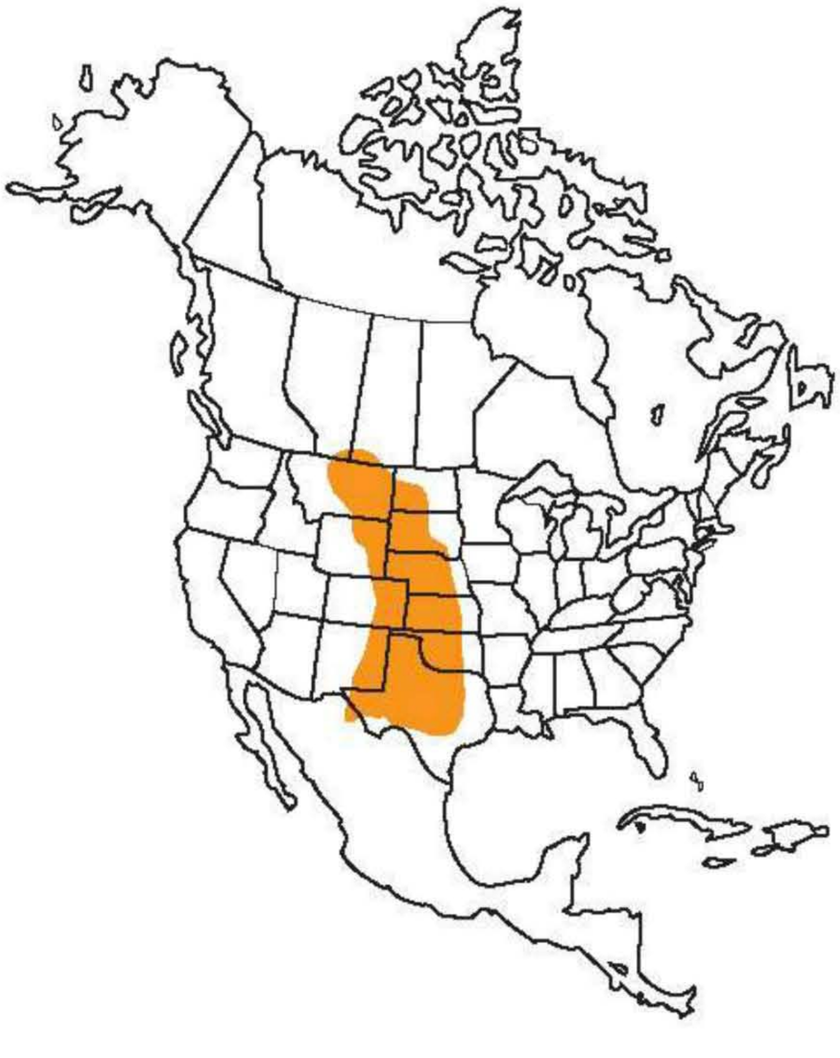

Prairie dogs may emerge from their dens during late winter warming trends, and are quite visible against the snow. Prairie dogs have dichromatic color vision, which allows them to detect predators over a long distance. A system of vocal communication enables them to quickly spread the alarm.

Herbivorous, the prairie dog thrives on a diet of grasses and broadleaf plants. Grasses are trimmed around the colony to remove cover that might hide a coyote, fox, or ferret.

GROUND SQUIRREL

Without food cultivated by humans, ground squirrels do not often grow out of balance. They live in small groups and are kept in check by birds of prey and sharp-toothed predators. But give a rodent an alfalfa field and everything changes. Before long, food becomes so plentiful that they don't have anything left to work at except procreation. Soon there are more ground squirrels than a rancher can feed.

Ground squirrels are members of the family Sciuridae, of the Marmotini tribe. Several different types of ground squirrels can be found. In some cases, the ground squirrel is found in small numbers and can be still-hunted. In cultivated fields, a hunter might spend all day in one place, shooting from an elevated stand.

As a team, two hunters can take turns spotting and shooting. Binoculars can help the spotter direct the shots. Such work makes both hunters more proficient. For the young shooter, looking forward to his or her first big game season, the practice is invaluable. April and May are the prime months for shooting as crops sprout and grow through last year's stubble. By mid-June, the grass may be too tall for spotting.

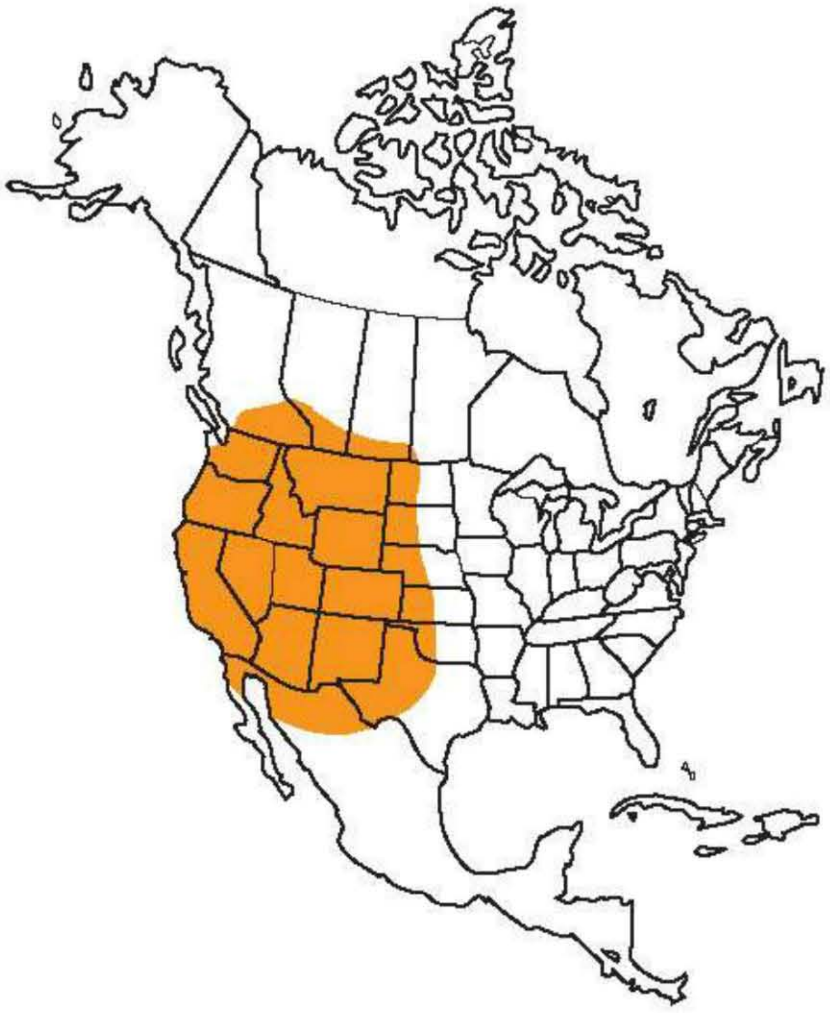

Property owners are often happy to let you shoot, but permission must be obtained. Keep in mind that well-traveled highways carry the most hunters and access might be easier to find off the beaten path. Establish relationships with landowners prior to the hunt or contact an outfitter for a guided shoot on private land.

For years, hunters have used rimfire .22s to good effect, but today there are better choices. The .22s, except in the most practiced hands, are rarely accurate past 100 yards (91 m).

An alternative is one of the 17 caliber rimfires. Based on the .22 Long Rifle or the .22 Magnum case, the 17-caliber bullet tips the scale at half the weight of the .22 projectile and can be accurate to 200 yards (183 m).

Abundant ground squirrel populations offer hunters the chance to improve long-range shooting skills. Over the course of a long day in the field, a marksman may fire 500 rounds or more.

RACCOON

Equipped with needle-sharp teeth and a quick temper, the raccoon is a fighter and a challenge to the houndsman/hunter. They may run for miles (km) and when cornered, can kill or maim hunting dogs twice their size.

Hardwood forests near water make good raccoon habitat. They also live in marshy lowlands and can adapt to habitats ranging from the Florida mangroves to the arid plains of New Mexico.

Raccoon eat almost any kind of food including berries, fruits, nuts, frogs, crayfish and insects. Sweet corn is a favorite. They hold their food with small, nimble paws.

An adult raccoon may travel up to 5 miles (8 km) on its nightly feeding rounds, especially in warm weather. During the day they often den up in hollow trees. But they may bed in the tall vegetation along the edge of a marsh, in a culvert or in the ground burrow of another animal. Raccoon frequently use different dens or beds on successive days. They den up for a period after heavy snow; in cold climates they may hibernate.

Ranked among the most intelligent game animals, raccoon also have excellent hearing and good eyesight.

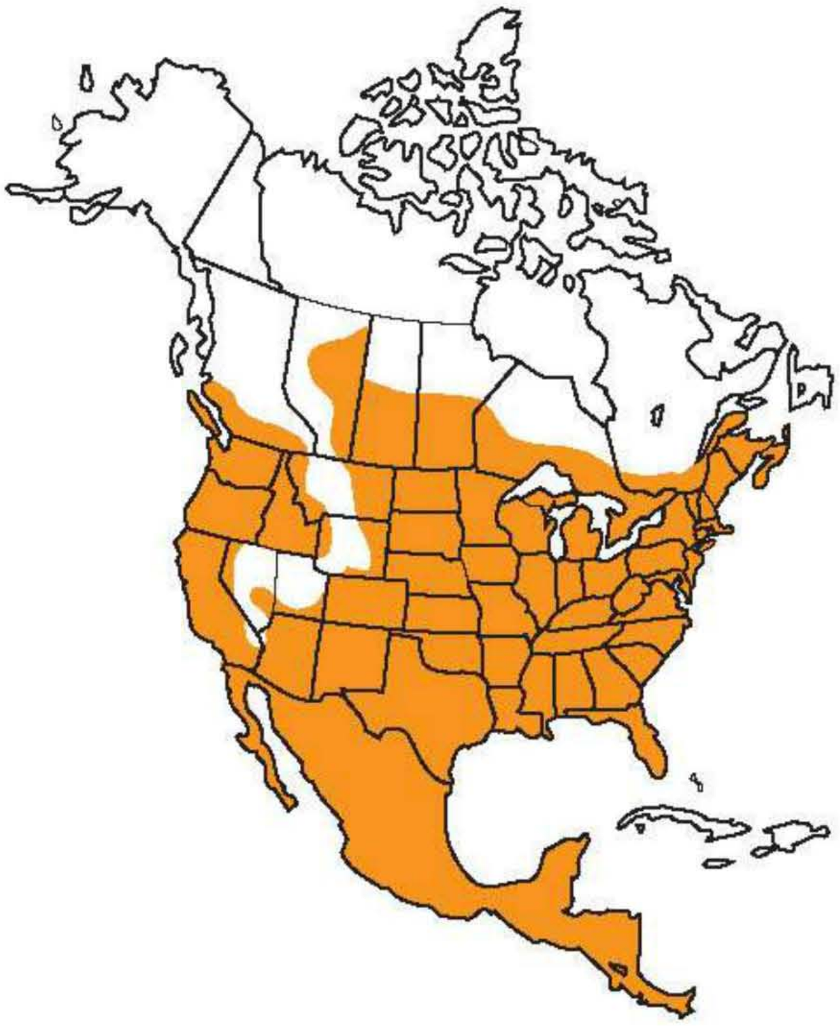

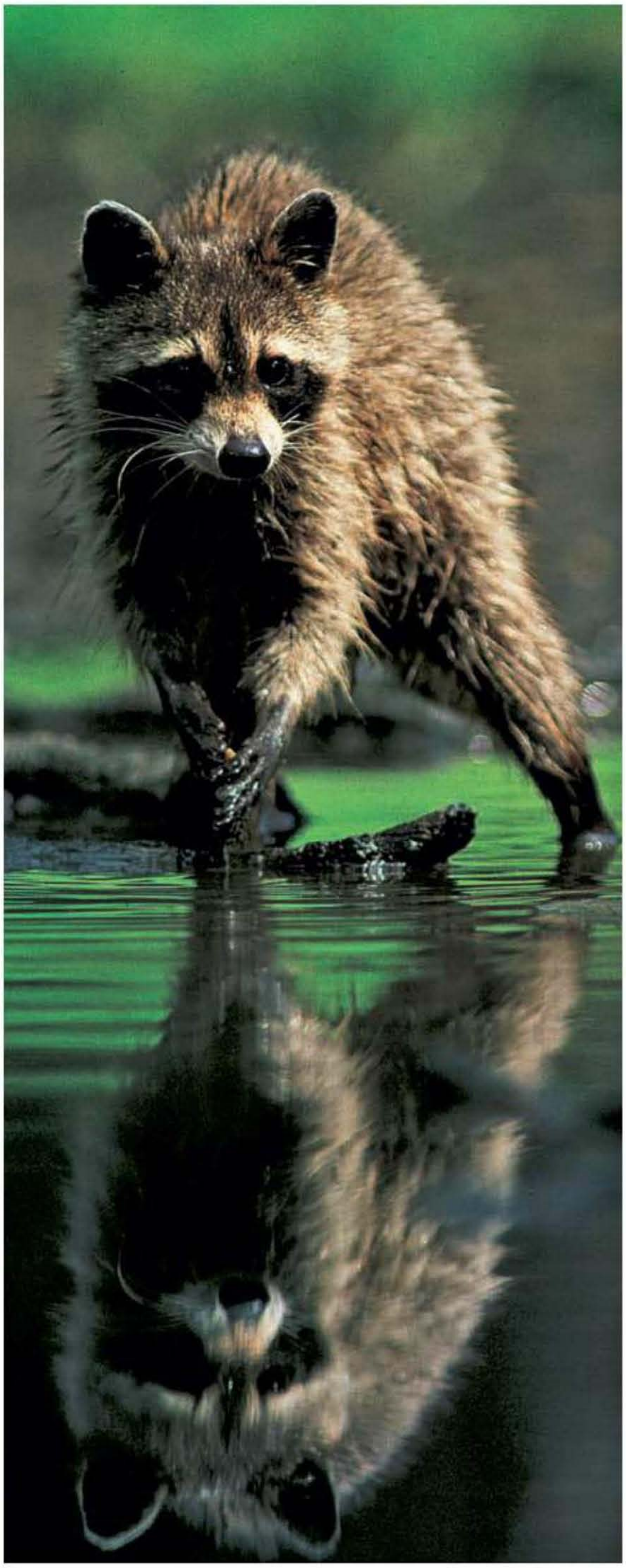

Raccoon have a black, mask-like band across the eyes and black rings on the tail. Most raccoon weigh between 15 and 20 pounds (6.75 and 9 kg). The largest on record, 62 pounds, 6 ounces, (28.1 kg) was shot in Wisconsin in 1950.

They can run up to 15 mph (24 km/h) and are good swimmers.

The vast majority of raccoon hunting is done with hounds, especially Walkers, black-and-tans and redbones. Tracking is easiest on damp, cool nights with a slight breeze. Some dogs can pick up a cold trail, making it possible to hunt in daylight.

When pursued by hounds, raccoon usually run in large circles, crawling in and out of holes and climbing up and down trees to lose the dogs. They may jump to the ground from as high as 50 feet (15 m) and scurry away unharmed. They stay in a tree only when the hounds get so close that other avenues of escape are impossible.

Early in the season you may be able to call raccoon. After dark, take a stand along a stream bank, lakeshore or cornfield. Use calls to imitate an injured bird or rabbit.

Most raccoon hunters use .22 rifles, although some prefer 20-gauge shotguns with No. 6 or 7½ shot. Where legal, a few hunters use small-caliber handguns.

WHERE TO FIND RACCOON

Look for holes in trees that could be den sites (left). You seldom see raccoon during the day, but an area with many dens would be a good spot to hunt after dark.

Check for tracks in the soft dirt around ponds, marshes or streams. Raccoon go to water to drink and to find foods like crayfish, frogs and small fish.

HOW TO HUNT WITH HOUNDS

Release the hounds and allow them to range ahead to pick up raccoon scent. To better control the hounds, some hunters keep them on leash until the dogs detect fresh scent, then turn them loose (left).

Listen for the hounds to start baying, then follow the sound. Some hunters use orange or red headlamps to find their way through the woods. The colored light is less noticeable to a raccoon than a white light.

Continue to follow the dogs until they either tree the raccoon or lose it. Sometimes the chase goes on for miles (km). Hunters can tell when their dogs tree a raccoon, because the baying becomes more intense.

COYOTE

The coyote, once an inhabitant of the Western plains, has spread across the continent. They are thriving now because of humans. In the late-1800s and early-1900s, ranchers and trappers declared war on wolves. For decades, the coyote's ancient enemy has been largely absent from the lower 48 states. With little competition, coyotes have taken over. And, they have become educated about how to live around people. Even within the city limits of big cities, coyotes prowl golf courses and fields in search of their prey. In the suburbs, residents complain about the coyotes that roam at will, eating from dog food dishes and garbage cans or picking up unsuspecting dogs and cats.

A good place to hunt coyotes is wherever they can find their principal foods. Fields, meadows and prairies are full of mice; brushy draws hold rabbits and upland birds. Wherever there are antelope or deer, there will be coyotes close by. In the spring when ungulates are giving birth to their young, a fawn-in-distress call can draw in predators.

When you find an area that has all the right coyote foods, look for sign. Coyotes use trails and waterholes just like the other animals. If there are coyotes in the area, you'll soon see their sign.

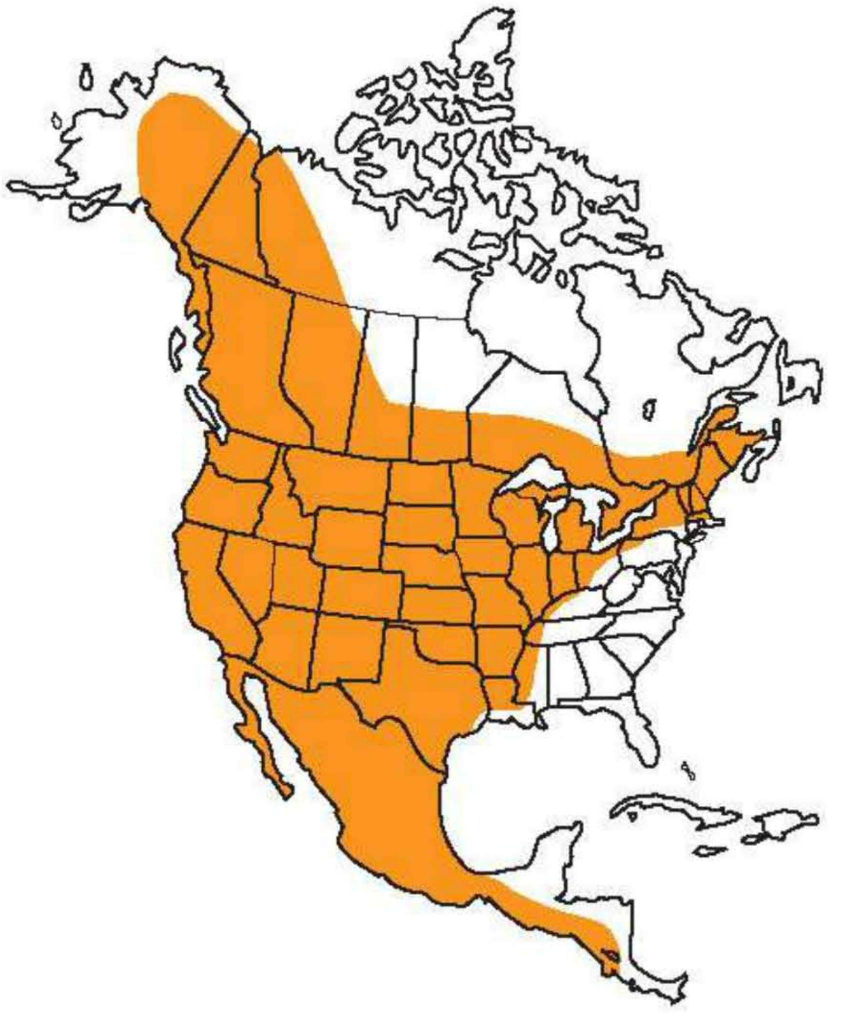

An adult coyote stands less than 24 inches (61 cm) tall and varies in color from whitish-gray to brown with sometimes a reddish cast to its pelt. Ears and nose appear long and pointed, in relation to the size of its head. It weighs between 20 and 50 pounds (9 and 22.5 kg) and can be identified by a thick, bushy tail, which it often holds low to the ground. He is an extremely lean, fast animal that may reach speeds up to 43 mph (69 km/h).

Coyote droppings are about the same size as a dog's of similar size but the scat is small and twisted on the end. It contains undigested bits of whatever the animal has been feeding on. Deer hair, bits of rabbit and mouse fur and even berries are found in coyote droppings.

Coyotes also need a den to raise their pups in spring and summer and for shelter in extreme conditions. A good coyote hunting area will be adjacent to food sources but also give ample cover to hide from their enemies.

In September and October the young of the year are out searching for food on their own. The population is strongest at this time of the year and many coyotes are taken during deer and elk seasons. But the coyote pelt is best after the first snows and a tanned pelt or a coyote rug makes a fine trophy to remember a successful hunt.

Calling Coyotes

A fawn-in-distress or a dying rabbit predator call costs only a few dollars and can be easily mastered. Electronic calls with remote controls and motorized decoys cost more, but provide a way for the hunter to watch the call from a distance.

The first two hours and last two hours of daylight are the best times for calling. A coyote may respond anytime of day but the likelihood is greater at first and last light.

The coyote knows that the sound of a rabbit or a rodent in distress means an easy meal, if it can take the prey away from whatever is killing it. That is why sometimes a dominant coyote comes right in while a younger dog might be hesitant.

Put fright and pain in the calling. The first sounds a rabbit makes when it is captured are a series of squalls

Coyotes normally have a territory of 10 to 40 square miles (26 to 104 sq km). They can sprint up to 35 mph (56 km/h), fast enough to overtake a deer. A coyote's track looks similar to that of a domestic dog.

which then become gasping cries. If the predator shifts its grip then the squalls might come again.

Study the land and determine likely approaches a coyote might make. The wind is of paramount importance. Situate yourself downwind from where the coyote is likely to be. Situate hunting partners up to 50 yards (45 m) away from the caller.

Wear full camouflage and grease paint on hands and face. Keep all movement to a minimum. As coyotes approach they zero-in on the location of the caller.

Call in one location for at least 20 minutes. Depending on the topography you might need only move a quarter of a mile before calling again.

For a dedicated coyote rifle, the best choice might be one of the flat-shooting .22 center-fires or a 6mm or .243. Beyond these, any rifle a big-game hunter is proficient with is the perfect choice for coyotes.

For calling situations where the coyote is likely to come in close, a shotgun is ideal. Use a load with sufficient knockdown power in the range between No. 2 and 00 buckshot.

Coyotes perform a valuable function, keeping rodent and rabbit populations in check. However, as coyote populations increase they can put a severe dent in the numbers of deer and antelope herds.

Coyotes are on the move when snow is on the ground. The cold weather speeds up metabolisms, making them vulnerable to the sound of a prey animal in distress.

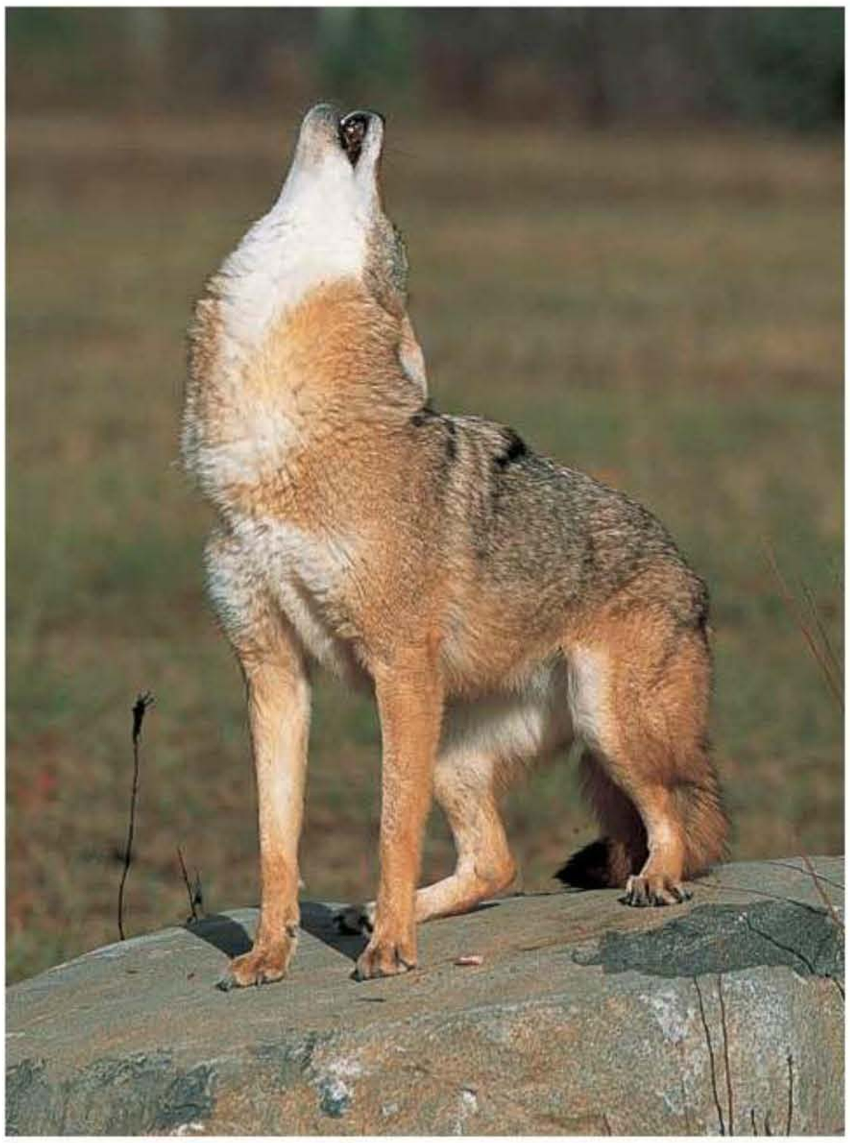

A shot at a coyote might be taken as close as 10 feet (3 m) or as far out as 300 yards (274 m). A good coyote gun should be able to accommodate either circumstance.

Some callers use confidence decoys to lure the coyotes in. Jackrabbit, cottontail and fawn decoys give the coyote something to look at as it approaches. Other hunters simply use an old stuffed animal that approximates the natural.

Chapter 6

HUNTING UPLAND GAME BIRDS

The memories that are made in the fall are the kind of images that stay with you for the rest of your life.

Dried stalks of corn rattling with the November breeze and the sudden rush of wings that is a gaudy ringneck rooster clawing for altitude. Chukar running the ridge top in front of a flushing dog, then a dozen birds briefly silhouetted against the sky. A German shorthair quivering on point, his eyes locked on a patch of color in the grass. The heft and grace of a fine double gun coming to your shoulder.

Upland bird hunting is a multi-faceted sport. Bird hunters cherish guns that they shoot well and spend months training their dogs for a few days afield each year. Some hunters focus on one or two species, but the continent holds a variety of game bird opportunities, from farmland pheasants to mountain quail, to the challenge of chukar.

Upland gunning is sport for hunters of any economic background and experience. Upland bird hunts are the province of anyone with a shotgun and a license, whether a nine-year-old novice or a well-traveled sportsman.

The term upland means high ground. But upland birds can also be found in and around lowlands. Pheasants frequent marsh edges and sharp-tailed grouse often live in bogs. Migratory upland birds also congregate in lowland areas. Woodcock feed along moist stream banks and mourning doves gather near water holes.

Upland bird hunters need little equipment other than waterproof boots and a thorn-proof jacket and pants. With the exception of the small-caliber rifles used for wild turkey, upland bird hunters almost always use shotguns. The choke and shot size varies with the size of the bird and the usual shooting distance.

RING-NECKED PHEASANT

Imported from Asia in the 1880s, the pheasant fast became a favorite. The bird is a walking, flying kaleidoscope of color. Flashing blue, purple, copper and iridescent green, it rises out of the amber grain in a blur of wings and sound.

The pheasant deserves its reputation as one of the most wily and elusive upland game birds. Its first instinct is to run rather than fly. And despite its large size and gaudy colors, a rooster can slink away unnoticed in ankle-high cover. Sometimes a bird sits tight, refusing to budge unless you actually step on it.

Pheasants rely on excellent eyesight and good hearing to elude hunters. They can detect ground vibrations not recognizable to humans. These vibrations often cause roosters to crow.

Ringnecks thrive in fertile agricultural areas with good nesting and wintering cover. In spring, they nest in moderately dense cover like hay meadows, clover fields and roadside ditches. They continue to use these areas as roosting sites into the fall, but also roost in weedy crop fields, short slough grass, willow patches and woodlots. They occasionally roost in trees. Pheasants need heavier cover to escape winter storms.

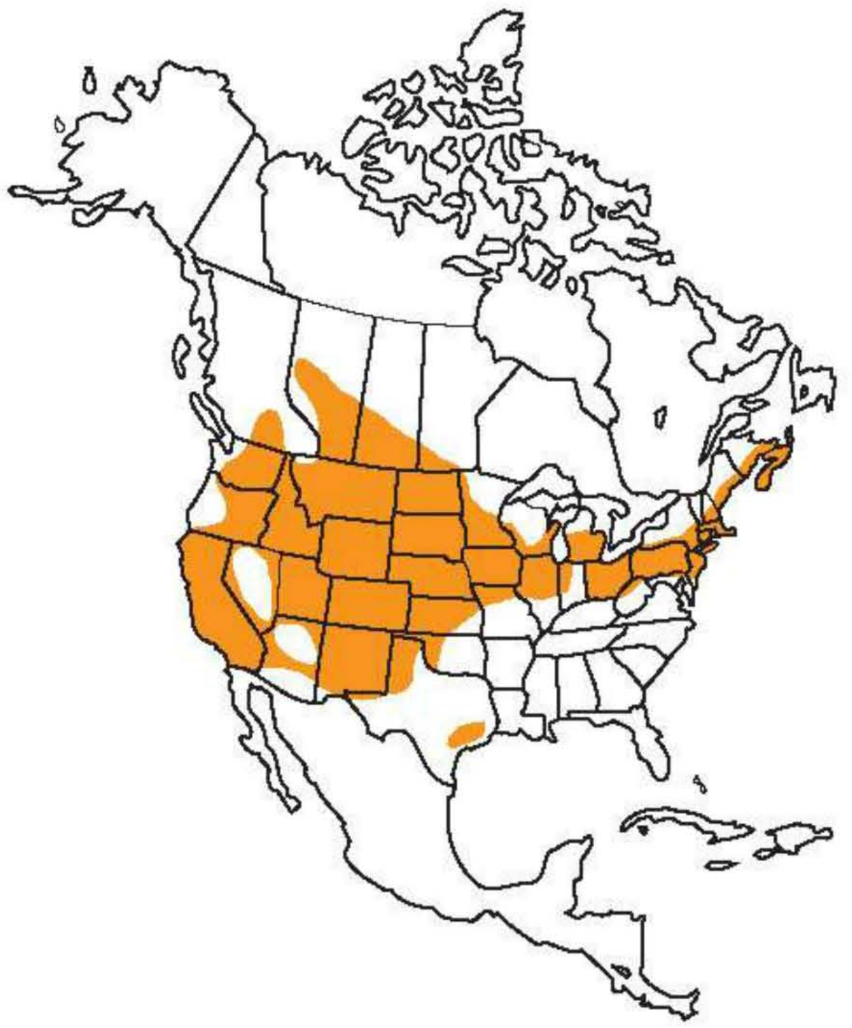

Primary foods include grain crops like corn, wheat, milo and soybeans. They also eat weed seeds and insects. The birds pick up grit in fields and along roads. This helps their gizzard to grind food.

From June through August, hen pheasants remain with their broods of four to eight chicks. By fall, the chicks begin to mature and the groups break up. As winter approaches, pheasants often flock together where they can find food and heavy cover. Flocks may contain hundreds of birds.

Despite the fact that pheasants are among the hardiest of game birds, their average life-span is only nine months. Normally, only 30 percent of the birds survive from one year to the next, even where there is no hunting season.

In most states and provinces, only roosters are legal game. Pheasant roosters can mate with many hens. Research has proven that hunters can harvest up to 90 percent of the roosters without affecting the next year's hatch.

Hunters can quickly distinguish the colorful rooster from the drab hen. In addition, roosters often cackle on take-off, removing any doubt about the bird's sex. The typical rooster measures 30 to 36 inches (76 to 91 cm) from head to tail and weighs 2½ to 3 pounds (1 to 1.3 kg). The hen has a much shorter tail and weighs about ½ pound less (0.25 kg).

Roosters have spurs on the lower part of the leg. The spurs grow longer and sharper as the bird gets older, reaching ¾-inch (1.9 cm) on three-year-old birds. They use their spurs in spring territorial battles. The sharp spurs on an old rooster can badly scratch a hunter or a dog.

When a rooster bursts from cover, it quickly reaches a speed of 35 to 40 mph (56 to 64 km/h). It may fly up to 1 mile (1.6 km), but usually only a few hundred yards (m). A bird generally spends its entire life in an area of ½ square mile (1.3 sq km) or less, leaving that area only if food or cover becomes inadequate.

Hunting pheasant is most glorious for what returns in memories. Good friends, a hard-working dog, long-tailed birds and sunlight on steel barrels.

WHERE TO FIND PHEASANTS

Wild ringnecks are often found in wheat, corn, barley and milo fields, feeding on grain, weed seeds and insects. Thick, brushy cover along fence rows, ditches, streams and marshes provide cover.

Pheasants feed in the morning and evening, then retreat to rest. Ideal habitat includes water, cover and feed within a short walk. Early morning and evening are good times to listen for calling.

When working brushy fence rows, one or two drivers should bust through the weeds with a dog, searching a ragged pattern back and forth. Post another hunter at the end of the row to jump skulking birds into the air. Expect pressured pheasants to seek out steep, brushy draws that defy all but the most determined.

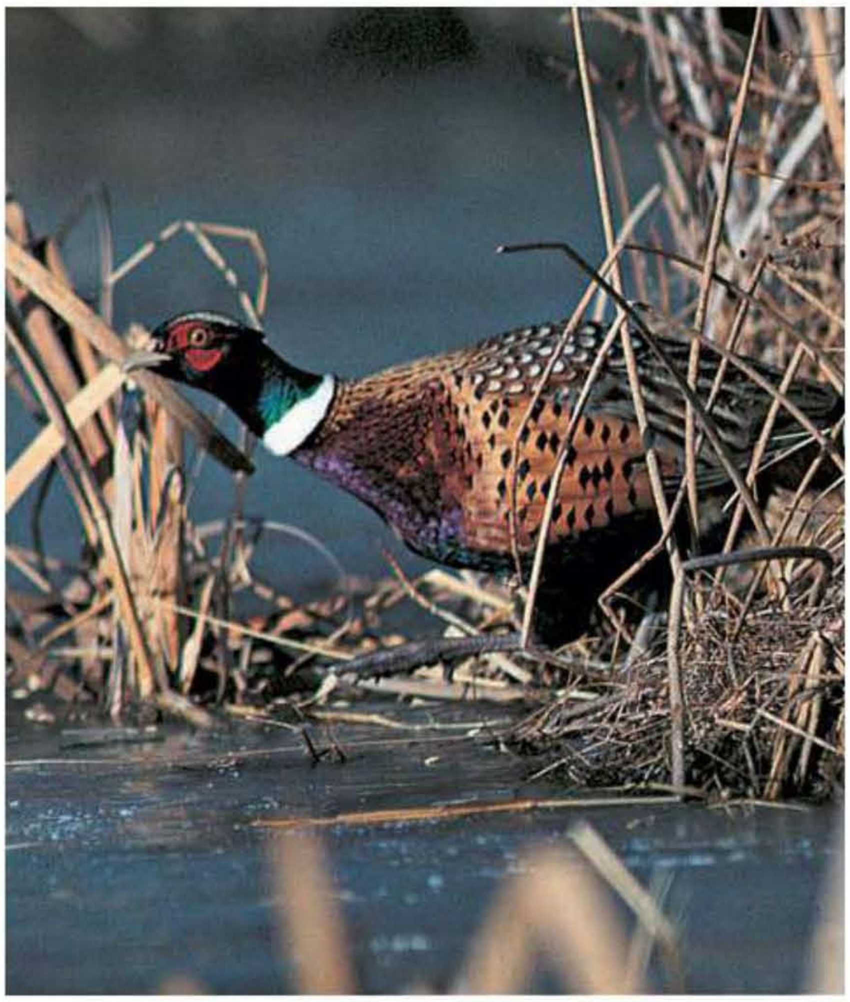

Ringnecks get their name from the white ring around the rooster's neck. Both sexes have brownish tails with black crossbars. The rooster has a reddish-copper breast and a powder-blue rump. Its head has shades of metallic blue, green and purple with a bright red eye patch.

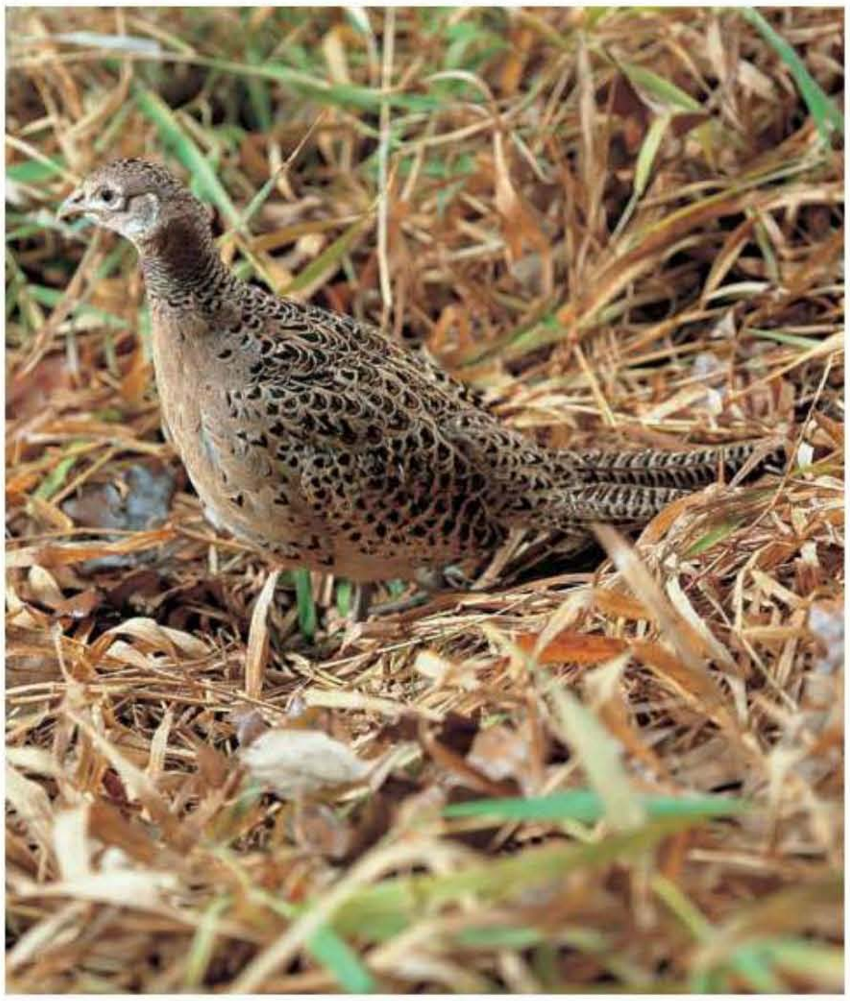

The hen is tan with dark flecks and creamy mottling.

TYPICAL PHEASANT HABITAT

Nesting cover is vital to pheasants (left). Hens need grassy cover at least 12 inches (30 cm) high which remains unmowed until after nesting.

Fertile croplands offer a good food supply. But expanses of corn and other row crops with little cover support few pheasants.

Winter cover includes cattail sloughs, thick brush or willows and woodlots. With good cover, pheasants can survive to –50°F (–45°C).

When hunting in standing corn, post blockers at the end of the rows to jump roosters into the air. Avoid shooting at birds until they are above the stalks to avoid endangering other hunters.

Pheasant Hunting Strategies

The pheasant hunting season can be split into two parts: the first few days and the rest of the season. Young pheasants lack the wariness of birds hatched the previous year, so the early days of the season usually offer the easiest hunting. Once the easier birds are gone, you will find it much more difficult to outwit the remaining roosters.

EARLY SEASON. You can locate a good area before the season opens by driving through the countryside and looking for pheasants. The best time to spot the birds is around sunrise on a clear, calm day with dew on the grass. You may also see pheasants in late afternoon. Another way to locate a good hunting area is to look for abundant nesting cover. Chances are there are birds in the vicinity. If you find a promising location, ask the farmer if you can return to hunt once the season begins.

Early-season hunters often find most of the land covered with crops. In this situation, pheasants may be almost anywhere. To flush birds from large crop or stubble fields, hunters conduct drives. With this much cover still standing, other techniques may not be as productive.

Ringnecks usually flush close in early season, so a shotgun with an improved cylinder or modified choke works best. Use shot no larger than No. 6.

LATE SEASON. Once farmers harvest their crops, the birds have fewer places to hide. But birds that survive to late season have learned to evade danger and often flush far ahead of approaching hunters. Or they hold extremely tight and let hunters walk past.

In late season, roosters hole up in much heavier cover than they did earlier in the year. They prefer areas with tall trees or other cover that breaks the wind. Look for them in brushy woodlots, thick fence lines and drainage ditches lined with slough grass. A favorite hiding spot is a fringe of cattails around the edge of a shallow wetland. Although it prefers heavy cover, a late-season rooster sometimes seeks refuge in a patch of grass not much larger than his body. After a snowfall, hunters often find pheasants under clumps of grass covered with snow. The birds evidently allow themselves to become snowed in.

To bag late-season roosters, one hunter blocks a possible escape route while another approaches from the opposite end of cover. If a rooster flushes too far ahead of the walking hunter, the blocker may get a shot.

For long-range shooting in late season, use a modified or full choke shotgun with No. 4 to 6 shot.

Hunting Pheasants with a Dog

Statistics show that hunters who use dogs bag twice as many ringnecks as those who do not. Most hunters agree that flushers and retrievers work best in heavy cover. Preferred breeds include springer spaniels and Labrador retrievers. Some hunters prefer pointing dogs for large expanses of light, grassy cover, but a wily rooster often runs rather than hold to a point.

A dog with good retrieving skills seldom loses a crippled ringneck. Pheasants are notoriously difficult to kill. A wounded rooster usually runs off and burrows under heavy grass or brush, where a hunter without a dog would have practically no chance of finding it. A good dog marks the bird down and relentlessly pursues it through even the thickest tangle of vegetation.

Skilled handlers refuse to work their dogs in standing cornfields or fields of other tall row crops. To a ringneck, an open row is an invitation to run. And most dogs follow, often disappearing into the field and flushing birds at the opposite end.

A novice pheasant hunter tends to over-command his dog, continually directing it to hunt likely-looking spots. Instead, let the dog use its nose to decide where to hunt. Allow it to range back and forth across cover until it finds fresh scent.

Hunt strip cover with your dog on the downwind side. When the dog detects fresh scent, it moves into the cover and flushes or points the bird. Late in the season, another hunter should post at the end of the strip.

Follow your dog through a large expanse of cover. A flushing dog may run when it picks up a scent. You must keep up so birds do not flush out of range. With pointing breeds, you do not have to stay as close.

Send your dog into a thick patch of cover to save you time and energy. Keep track of its location by watching and listening for moving grass or brush. Or, attach a bell to the dog's collar.

General Tactics

A ringneck's inclination to run rather than fly makes it one of the most difficult birds to hunt by yourself. But you can greatly improve your odds by choosing the right type of spot and by using proven techniques.

You are most likely to flush pheasants from small, isolated pieces of cover, such as a patch of short slough grass surrounded by a plowed field. In a large block of heavy cover, a rooster can easily give you the slip.

If you walk steadily in a straight line, pheasants probably sit tight and let you pass. But if you follow a zig-zag path, walking a few steps, then stopping for a moment, nearby pheasants generally become nervous and fly.

Watch and listen carefully for any indication of pheasant movement. If the cover is not too thick, you might catch a glimpse of a rooster running ahead. On a still day, you may hear the slight rustling of a rooster sneaking through the grass.

Driving for Pheasants

Driving takes advantage of the ringneck's habit of running at the sight or sound of humans. A row of hunters moves through all or part of a large block of cover, while posters wait quietly at the end.

Drivers may flush some pheasants, but more often the birds run to the end of the field. When the drivers approach the posters, the birds realize they are trapped and explode from cover.

TIPS FOR HUNTING PHEASANTS

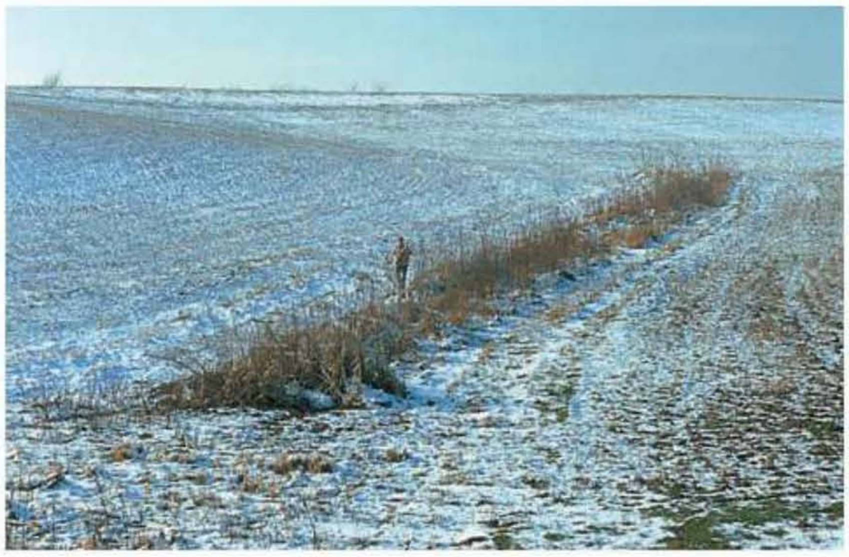

Select manageable-sized pieces of cover. Try to push the birds toward a spot where the cover suddenly ends. Pheasants are reluctant to run into an open field and usually flush near the edge.

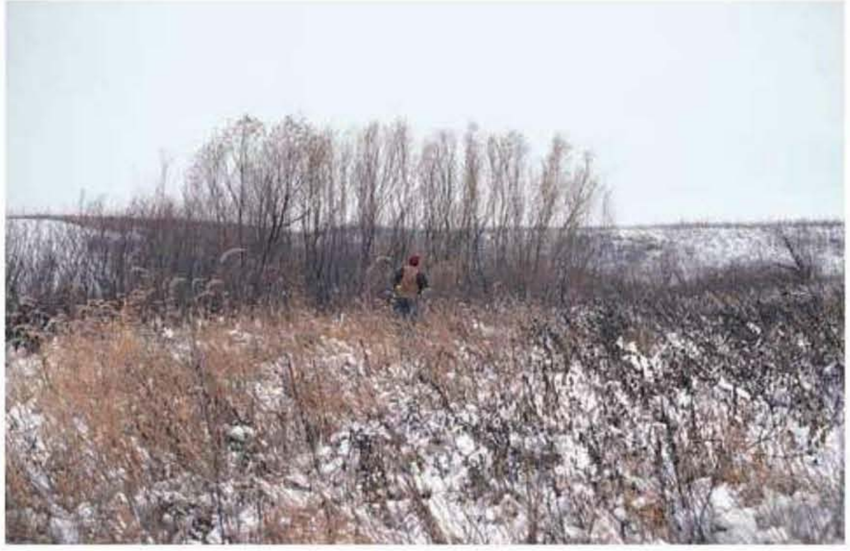

Look for the tallest, thickest cover, especially in cold or windy weather. Trees or tall weeds offer better shelter than the surrounding lower cover.

Hunt the sunny side of cover on frosty mornings. Pheasants move to the sunlit side to warm up and dry ant moisture on their feathers. Later in the day, the birds may retreat deeper into the cover.

Follow fresh tracks in the snow. Rooster tracks are slightly larger and farther apart than those of a hen. Often the tracks end at a snow-covered clump of grass. You may have to kick the clump to flush the bird.

Drive pheasants from the heaviest cover to the lightest. Pushing birds into sparse vegetation forces them to fly. If you drive from light to heavy cover, the birds are more likely to find a safe hiding spot.

Avoid driving into the sun, especially when it is low in the sky. The glare may blind you or prevent you from distinguishing a rooster from a hen. If you must hunt into the sun, listen for a cackle or look for a long tail.

Leapfrog stretches of cover. Drop off a partner, drive ahead, then start walking. He or she walks to the car, drives ahead of you, then resumes walking.

Choose dirty crop fields when hunting ringnecks. The weeds provide cover and food that is lacking in clean, well-manicured fields. Look for the tallest, thickest habitat, especially in cold or windy weather. Trees or tall weeds offer better shelter than the surrounding lower cover.

If the field is too wide to cover in one pass, the posters can take a vehicle to the end they are blocking. The drivers can take the vehicle back to the opposite end, then make another drive. Or the posters can return, while the drivers start another pass from the end where they finished.

When driving cover that narrows toward one end, start from the widest end. This way you will push the birds into a smaller area, increasing the likelihood of someone getting a shot.

BOBWHITE QUAIL

A covey of bobwhites exploding from cover can rattle even the veteran hunter. Often, each bird flies in a different direction. In the confusion, you may shoot hurriedly without touching a feather.

Despite their stubby wings, quail can fly up to 30 mph (48 km/h) and change direction instantly as they dodge through cover. But they usually fly less than 200 yards (183 m). Bobwhites are also excellent runners; in fact, they seldom fly unless threatened.

When alarmed, a member of the covey emits a barely audible signal which tells the others to freeze. They squat and remain motionless, relying on camouflage for concealment until the threat passes. If they feel too conspicuous, they scurry to another hiding spot.

Cocks have a white throat patch and a white line through the eye. On hens, the throat patch and line are buff-colored. Bobwhites measure about 10 inches (25 cm) long and weigh 6 to 8 ounces (170 to 226 g).

Unlike pheasants, bobwhites are monogamous, meaning that a cock breeds with only one hen. Quail eggs hatch from May to early July. The cock helps incubate the eggs and raise the chicks. By the start of hunting season, the chicks resemble the adults and the covey consists of 10 to 15 birds.

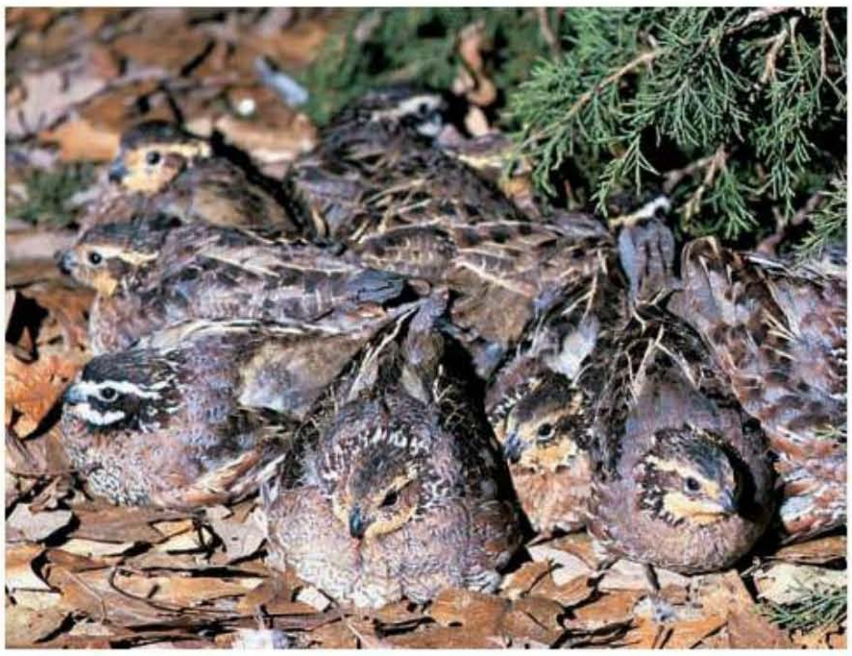

Bobwhite quail roost on the ground. During cool weather, they form a roosting ring, huddling in a plate-sized circle with their tails pointing toward the center. This tactic preserves body heat. Bobwhites prefer a roosting site with plenty of open space above them, enabling them to flush quickly should the need arise. Roosts generally have a south or west exposure, so the ground stays warm in late afternoon.

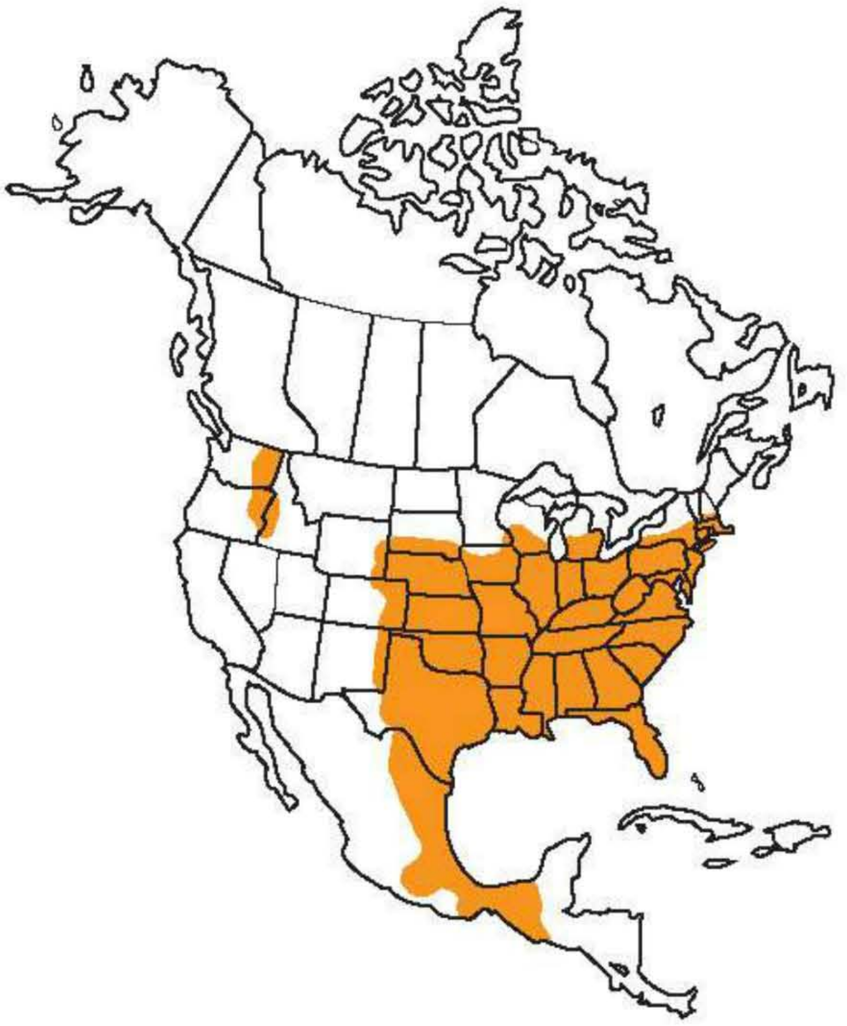

The birds thrive in areas with a mixture of grasslands, woodlands and brush adjacent to croplands. They use grasslands mainly for nesting; woodlands and brush provide roosting and escape cover. Bobwhites cannot survive in a climate with prolonged periods of deep snow or severe cold. They are unable to dig through deep snow to find food and their small bodies do not retain enough heat.

Nearly all bobwhites live within 50 to 100 feet (15 to 30 m) of field borders. They are seldom found in the middle of a crop field or woods.

The bobwhite's diet consists mainly of weed seeds, but they also eat insects, acorns and crops like corn, soybeans, wheat and milo. On hot, clear days, the birds feed in early morning and late afternoon. In midday, they take periodic dust baths along sunny field edges. But during cool, damp weather, bobwhites often stay in their roosts until mid-morning, feed intermittently until dark, then return to their roosts.

Each year, predators, severe weather and hunting take a heavy toll on the bobwhite population. On the average, a bird lives about 8½ months; only 15 to 20 percent survive to the next breeding season.

Hunting for Bobwhite Quail

To many bobwhite hunters, the biggest thrill comes from watching good pointing dogs in action.

A lone hunter who works the cover slowly and thoroughly can kick up some bobwhites by himself. But if the birds decide to hold tight, hunting can be extremely difficult without a dog.

Most hunters prefer English pointers or English setters for large expanses of cover. Brittanys and German shorthairs also work well, but do not cover quite as much ground. Handlers allow the dogs to range far ahead of the hunting party. When a dog detects fresh scent, it locks on point. The covey freezes, giving the hunters plenty of time to move into position.

Most of the birds burst from cover in unison. After the initial flush, work the area a little longer because a straggler or two may remain. Flushing the last birds is often more difficult.

Quail usually fly only a short distance, so you can flush them again. Sometimes the covey stays together, sometimes it breaks up. Watch carefully because the birds may veer off to the side just before they land.

Early season offers the best bobwhite hunting. Coveys consist mainly of young birds that have never been hunted. But as the season progresses, quail become much more unpredictable. Some birds flush when you slam your car door. Others run rather than hold. Flushed birds fly two or three times farther than they did in early season.

Finding downed quail can be difficult, even with a good dog. Some hunters maintain that the scent washes off as the birds fly. On the ground, bobwhites compress their feathers so little scent can escape. Bird dogs may walk within inches (cm) of wounded quail without finding them. Because of their small size and excellent camouflage, quail can hide in the lightest cover. Wear snakeproof chaps or leggings to protect yourself from snakebites. They also protect your legs from tough thorns and briers.

Many quail hunters prefer 20-gauge, double-barreled shotguns with improved cylinder and modified chokes. Small shot, usually No. 7½ or 8, works best.

WHERE TO FIND BOBWHITE QUAIL

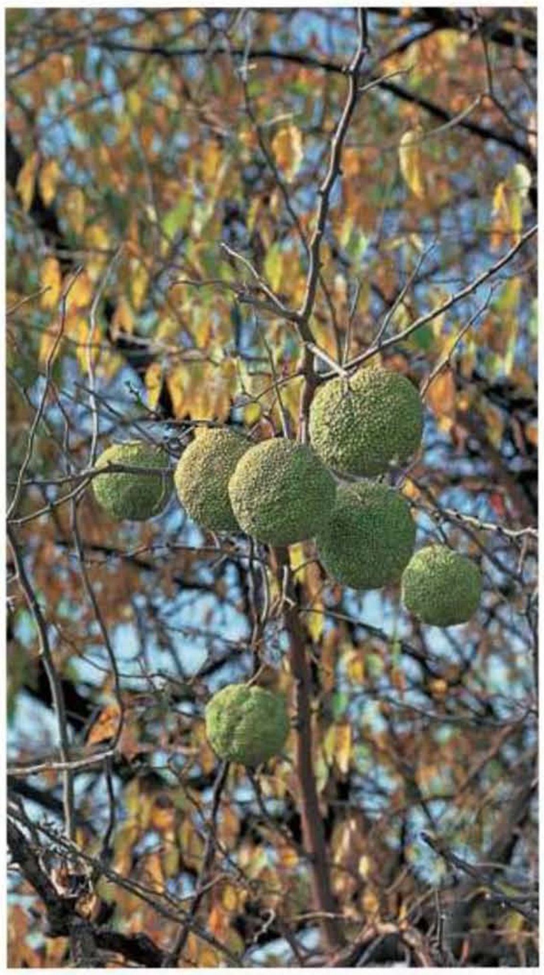

Osage orange means good quail cover. Its sharp spines prevent cattle from eating the underbrush. Look for the large, yellow fruits.

Giant ragweed provides roosting and escape cover. The plants stand up to 12 feet (3.6 m) tall. Ragweed seeds may be a major source of food in fall.

Lespedeza or Japan Clover, produces seeds which quail eat in winter. Found mainly in the southern United States, it is planted for erosion control.

BOBWHITE HUNTING BASICS

Pick a single bird and concentrate on that shot. Resist the tendency to flock shoot. A wild shot at the covey rarely brings down a bird.

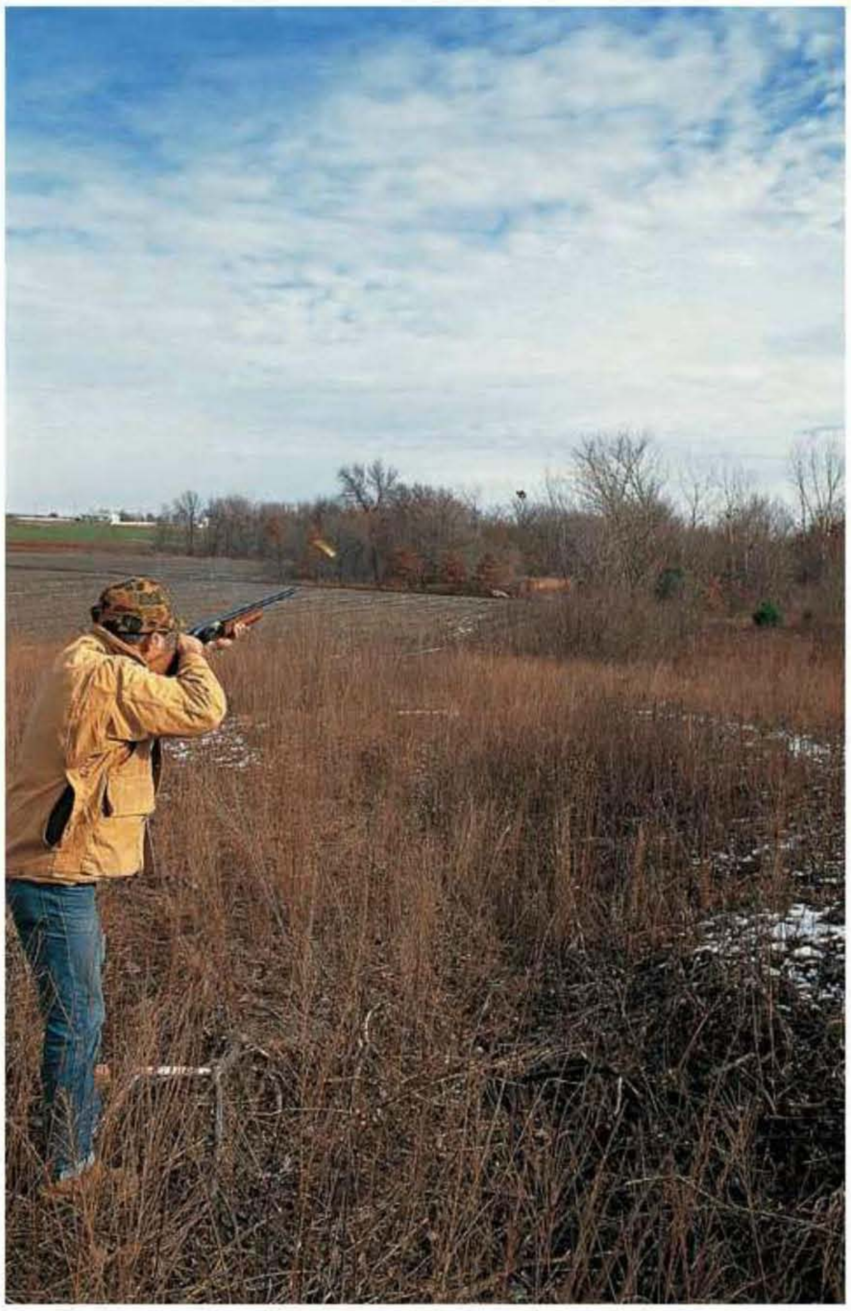

Learn to judge your effective shooting distance. Because bobwhites are so small, they are probably closer than you think.

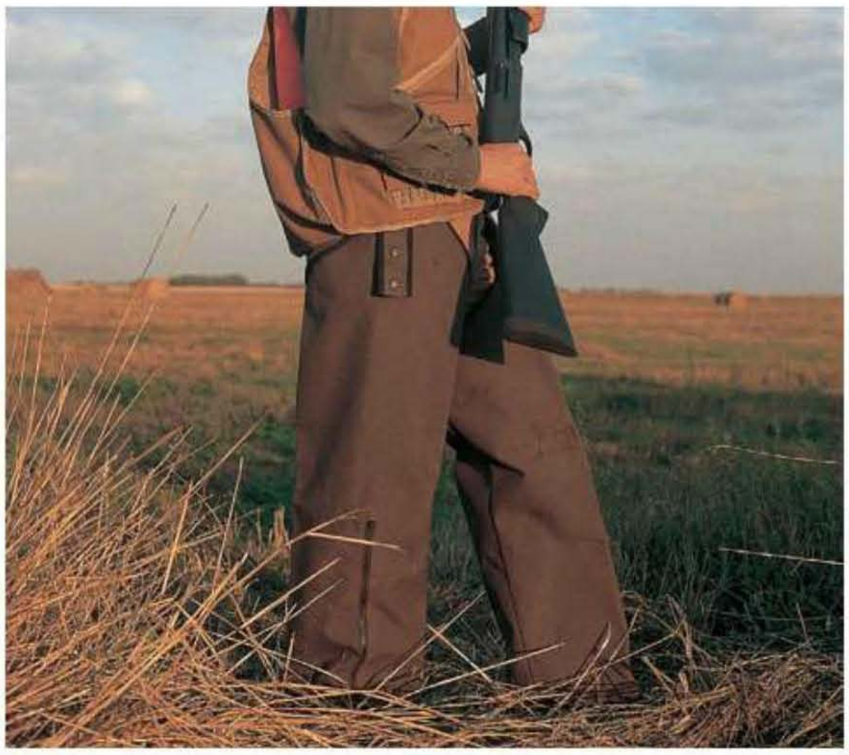

Wear snakeproof chaps or leggings to protect yourself from snakebites. They also protect from tough thorns and briers.

Listen for the typical bob-bob-white whistle during the mating period. You will find birds in the same area once the hunting season opens.

CALIFORNIA QUAIL

California quail are also called valley quail. They prefer semi-arid desert brushlands. They feed on the seeds of weeds and brush and roost in dense patches of tall shrubs or low trees. The birds need water each day and are seldom far from streams, springs or water holes.

In fall, they form large flocks numbering from 50 to over 100 birds. They prefer to run rather than fly. When flushed, they usually go only a short distance, then land in bushes or trees.

A good pointing dog can pin down an entire covey by circling the birds. If they fly, watch where they land. They hold tighter on the second flush.

Ideal shotguns for California quail are 12- or 20-gauge double barrels with modified and improved cylinder chokes. These guns give you an open pattern for birds that hold and a tight pattern for those that flush at long range. Most hunters use No. 7½ shot.

California quail, also called valley quail, can be found up and down the West Coast of North America, from southern British Columbia to the tip of the Baja Peninsula. They are a ground-dwelling and foraging bird that seldom moves far from the place where they were hatched. The coveys feed on seasonal seeds, leaves, and insects and search out sunny spots to take communal dust baths.

In fall, they form large packs numbering from 50 to over 100 birds. They prefer to run rather than fly. When flushed, they are loud flyers, bursting like popcorn from cover. They scatter to confuse predators and seldom fly very far, landing in bushes and trees.

These quail are native to the Southwest, but they have been introduced to other places around the world, including Australia, New Zealand, and Hawaii.

California quail prefer semi-arid desert brushlands. They feed on the seeds of weeds and brush, and roost in dense patches of tall shrubs or low trees. The birds need water each day and are seldom far from streams, springs, or water holes.

A good pointing dog can pin down an entire covey by circling the birds. If they fly, watch where they land. They hold tighter on the second flush.

Ideal shotguns for the quail on these pages are 12- or 20-gauge double barrels with modified and improved cylinder chokes. These guns give you an open pattern for birds that hold and a tight pattern for those that flush at long range. Most hunters use No. 7½ shot.

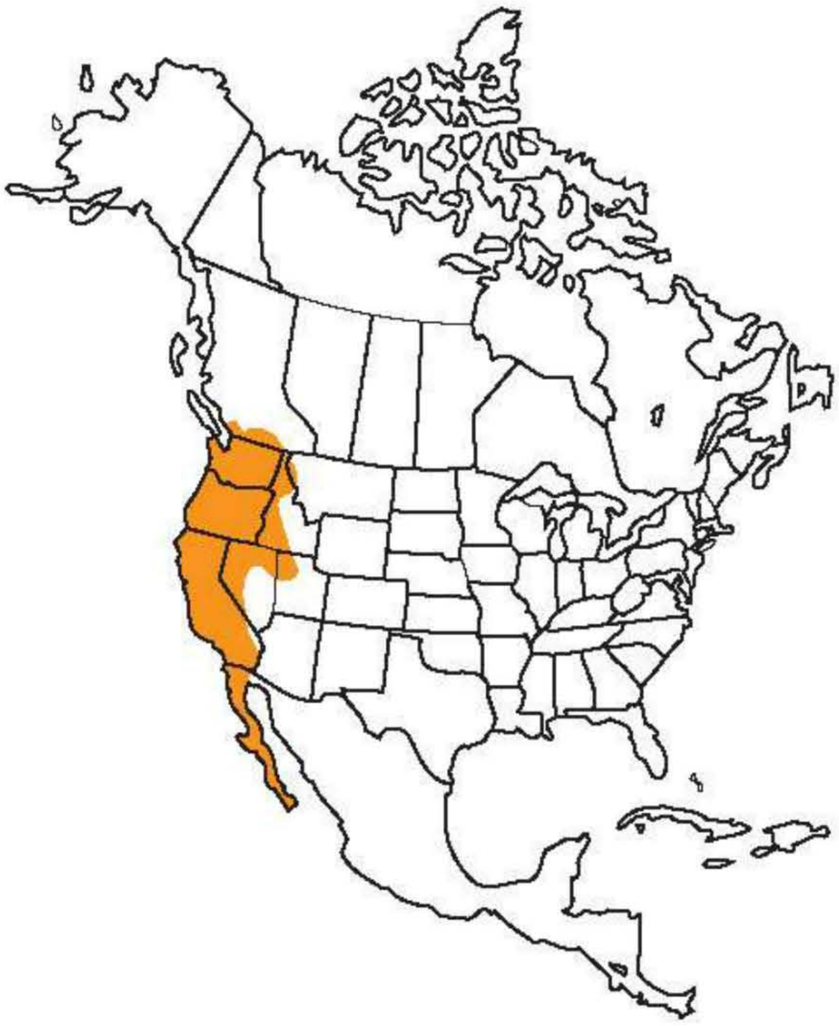

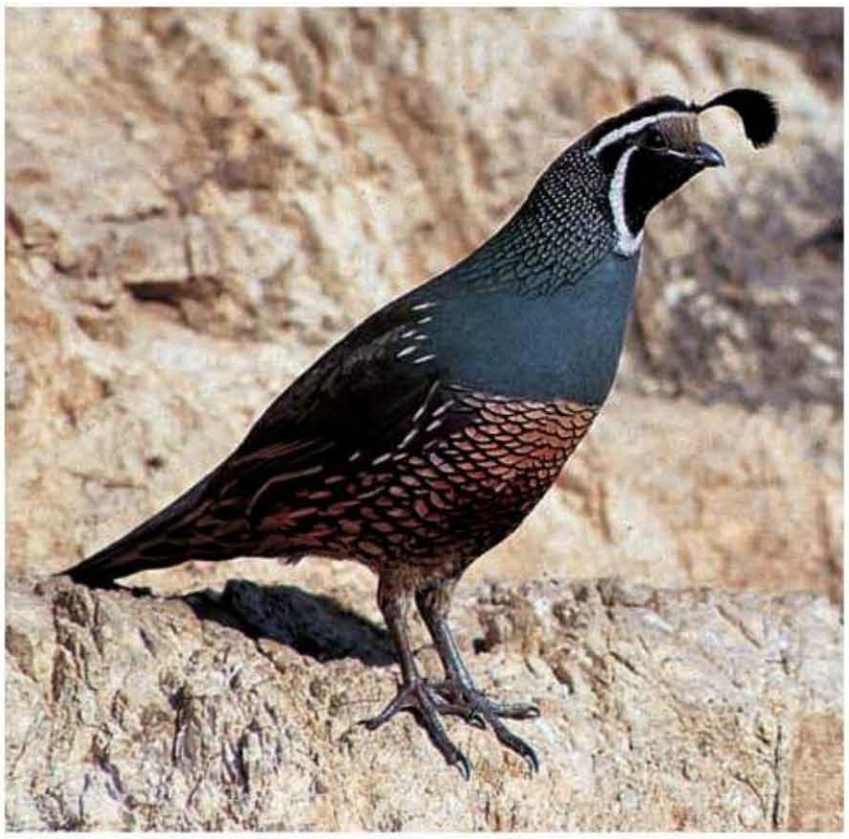

California quail have a scaled breast, grayish brown flanks with white streaks and a teardrop-shaped plume. Males have a black throat with a white border and a chestnut patch on the belly. Females are less boldly marked. California quail weigh 6 to 7 ounces (170 to 198 g).

GAMBEL'S QUAIL

Sometimes called desert quail, Gambel's quail are usually found along brushy slopes and in river valleys of arid and semi-arid deserts. They roost in dense thickets or trees and feed on the seeds of weeds and brush. Normally the birds get enough water from their foods and the dew. But during dry conditions, they rely on a watering site.

The birds come out to feed early and late. They spend the rest of the day in brushy cover to escape the heat. In cool weather, they leave their roosts earlier and may remain in the open all day.

If you approach Gambel's quail in the open, they will probably run before you can get a shot. But if you wait until they move to cover, the birds are more likely to hold. A good pointing dog improves your odds.

Poisonous snakes are rarely a problem, but some hunters wear leather or plastic leggings.

The Gambel's quail was named after William Gambel, a 19th century explorer of the southwestern United States. Sometimes called desert quail, Gambel's quail inhabit the southwestern desert country of California, Arizona, Colorado, New Mexico, Nevada, Texas, Utah, and Sonora, Mexico. Non-migratory, they are usually found along brushy slopes and in river valleys of arid and semi-arid deserts. They roost in dense thickets or trees, and feed on the seeds of weeds and brush. Normally the birds get enough water from their foods and the dew. But during dry conditions, they rely on a watering site.

The Gambel's average length is 11 inches (30 cm). It has a wingspan of 14 to 16 inches (35 to 40 cm). It moves through cover by walking and when pushed can run surprisingly fast. Its flight, like that of its cousin the California quail, is abrupt, explosive, and terminated by a long glide to the ground.

The birds come out to feed early and late. They spend the rest of the day in brushy cover to escape the heat. In cool weather, they leave their roosts earlier and may remain in the open all day.

If you approach Gambel's quail in the open, they will probably run before you can get a shot. But if you wait until they move to cover, the birds are more likely to hold. A good pointing dog will improve your odds.

In the cooler weather of hunting season, poisonous snakes are rarely a problem, but some hunters wear leather or plastic leggings.

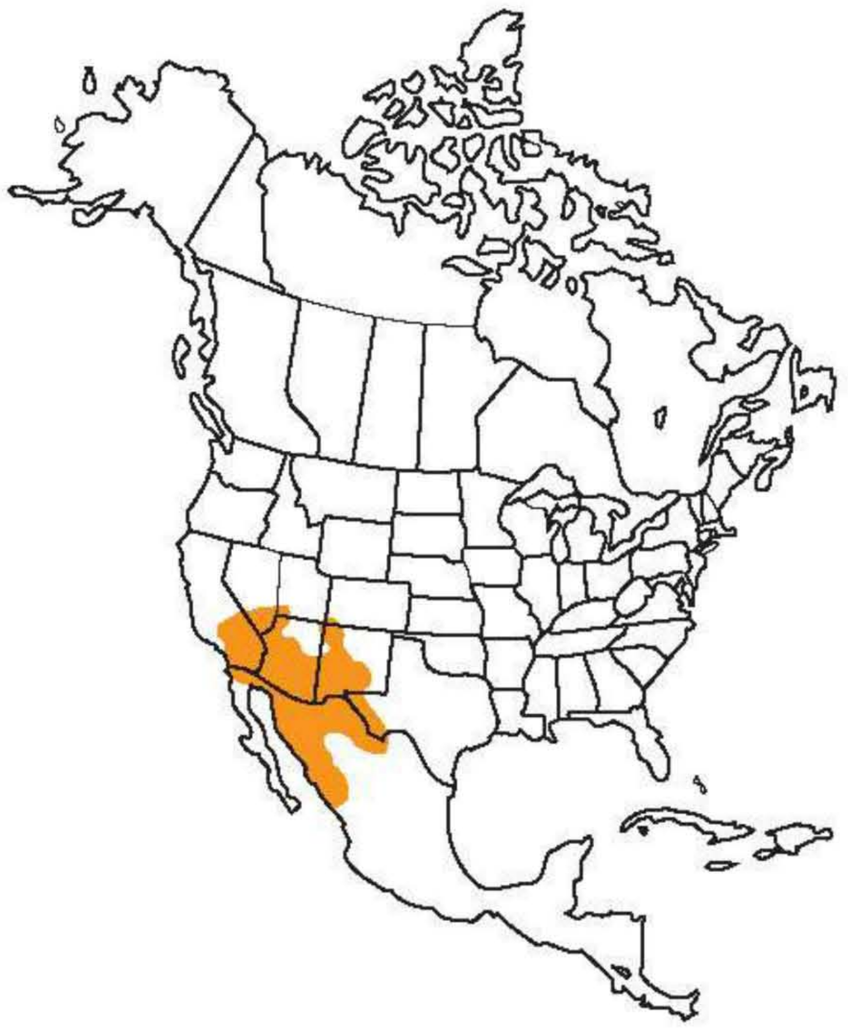

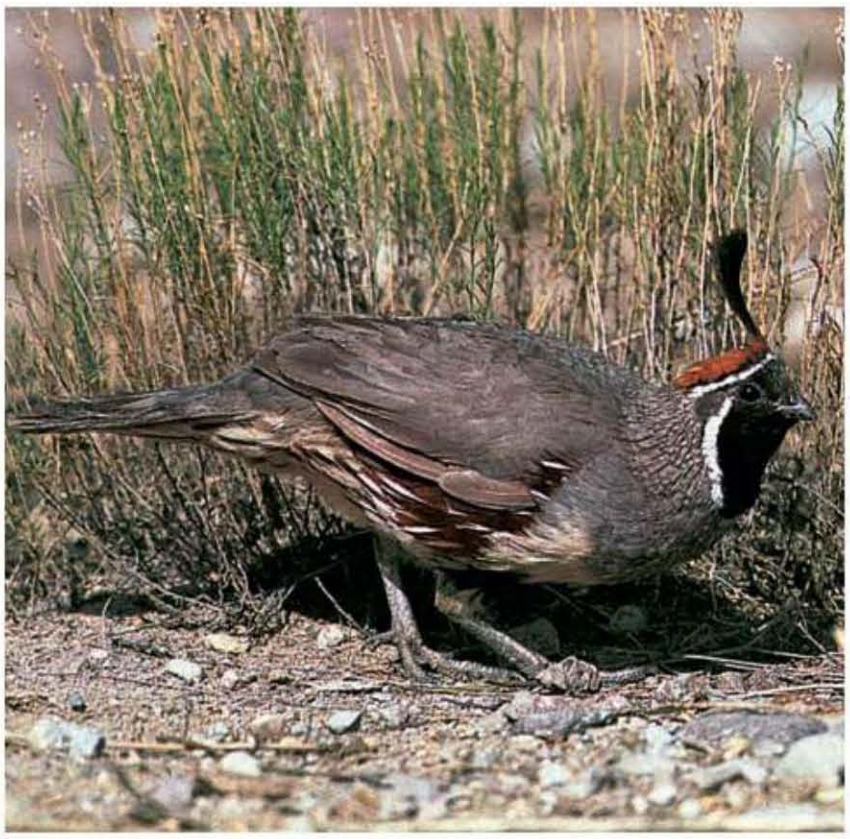

Gambel's quail resemble California quail, but their buff-white belly is not scaled and the cap and flanks on the male are reddish brown. Males also have a black patch on the lower part of the breast. Gambel's quail weigh 5½ to 6½ ounces (156 to 184 g).

MOUNTAIN QUAIL

Few upland game birds are as difficult to hunt as mountain quail. They run at the sight of a hunter and even a good pointing dog has difficulty pinning them down. A flusher or retriever is more likely to find single birds that sit tight.

Found at elevations of 2,000 to 10,000 feet (610 to 3,048 m), mountain quail live along brushy edges of conifer forests and streams. They eat berries, clover, wild oats and seeds of weeds and grasses; the birds roost under heavy brush or in small trees. Home territories take in large areas. They head to lower elevations in late fall to avoid severe weather.

Coveys generally number only seven to nine birds. They do not form large coveys, but hunters sometimes see loose groups of birds feeding in the same area.

Listen for the call any time of day, but pay close attention in the evening, as feeding birds reassemble near water prior to roosting. Food and cover are the keys to finding quail and to the lasting memory of a full game bag at the end of the day.

Few upland game birds are as difficult to hunt as mountain quail. They will run at the sight of a hunter, and even a good pointing dog has difficulty pinning them down.

Found at elevations of 2,000 to 10,000 feet (610 to 3048 m), the mountain quail inhabits mountainous chaparral in California, Oregon, and Washington. They live along brushy edges of conifer forests and streams. Mountain quail eat berries, clover, wild oats, and seeds of weeds and grasses; the birds roost under heavy brush or in small trees. They are not migratory, but their home territories take in large areas. They seek out seasonally available foods and head to lower elevations in late fall to avoid severe weather.

Like other quail, they move about primarily by walking. When they run, they can cover ground at an amazing rate of speed. When startled, they run through cover and flush when they hit an opening. The birds scatter when they fly then regroup by calling.

Coveys generally number only seven to nine birds. They do not form large coveys, but hunters sometimes see loose groups of birds feeding in the same area.

A flusher or retriever is more likely to find single birds that sit tight.

Listen for the call any time of day, but pay close attention in the evening, as feeding birds reassemble near water prior to roosting. Food and cover are the keys to quail, and to the lasting memory of a full game bag at the end of the day.

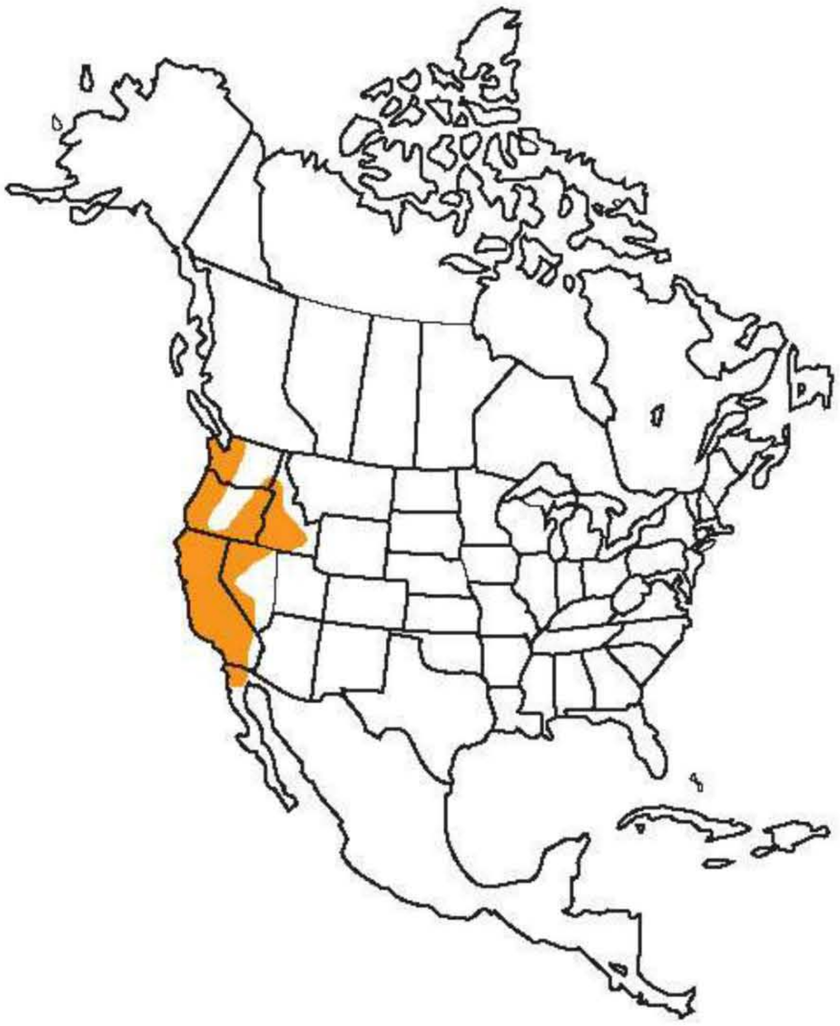

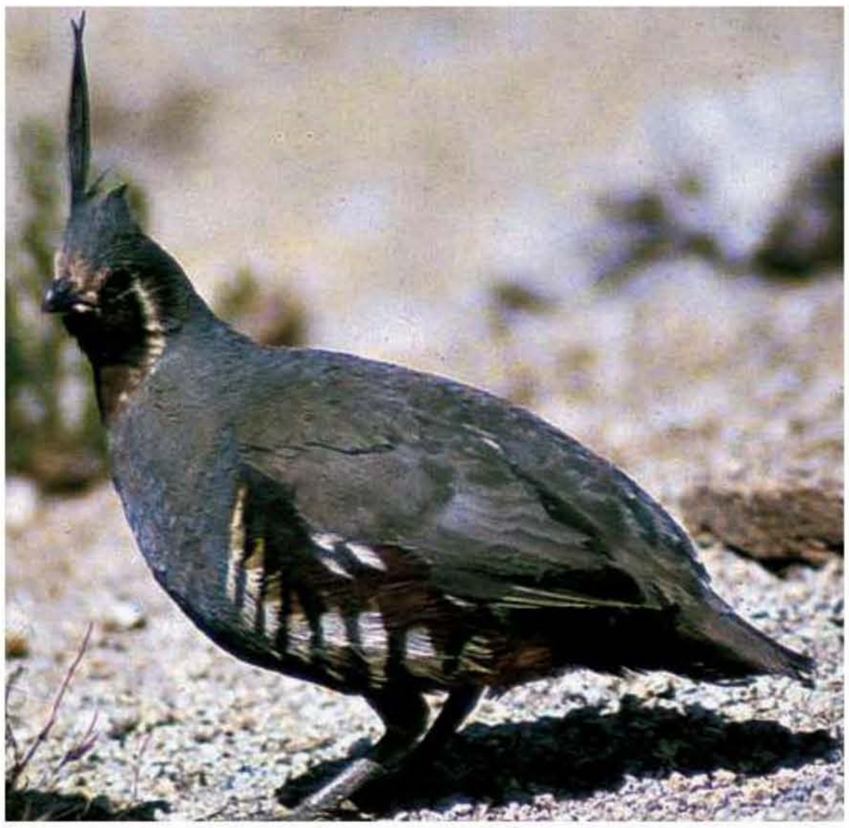

Mountain quail have a grayish-brown back, a chestnut throat and chestnut flank with heavy white bars. The head and breast are bluish-gray. Two long, dark feathers form the plume. The sexes look alike. Mountain quail weigh 8 to 9 ounces (227 to 255 g).

SCALED QUAIL

Some hunters refer to scaled quail as blue quail; others call them cottontops or blue racers. These birds are even more likely to run than mountain quail.

Scaled quail live in semi-arid desert grasslands and in grassy brushlands. They prefer weed and brush seeds for food. The birds roost in dense clumps of brush, grass or weeds. Their water requirements are similar to those of Gambel's quail. In some states and provinces, conservation agencies have installed watering devices, helping both quail species to survive during dry periods.

Scaled quail country can be hard on the dog. Snakes are a concern, but the most important thing is to carry enough water for both hunter and dog. Lava rock and cactus play havoc with a dog's paws. Bring boots and ointment to protect Fido's feet.

If you spot a covey, rush the birds to startle them into flying. Even if you are out of range, shoot to scatter the birds and scare up stragglers. Then hunt the singles, because they are more likely to hold.

Some hunters refer to scaled quail as blue quail; others call them cottontops or blue racers. They are named for their scaly appearance on back and breast feathers. They can be identified by their cotton-like white crest.

They are bluish-gray birds that live in semi-arid desert grasslands and in grassy brushlands. The scaled quail and its subspecies are found in parts of Colorado, New Mexico, Arizona, western Oklahoma, and Texas, as well as Sonora and Chihuahua in Mexico. The chestnut-bellied scaled quail is found in southern Texas down to northwestern Mexico.

They prefer weed and brush seeds for food. The birds roost in dense clumps of brush, grass, or weeds. Their water requirements are similar to those of Gambel's quail. In some states conservation agencies have installed watering devices, helping both quail species to survive during dry periods.

Scaled quail country can be hard on dogs. Snakes are a concern, but the cooler weather of hunting season is likely to keep them dormant. The most important thing is to carry enough water for both hunter and dog. Lava rock and cactus play havoc with a dog's paws. Bring boots and ointment to protect Fido's feet.

These birds are even more likely to run than mountain quail. If you spot a covey, rush the birds to startle them into flying. Even if you are out of range, shoot to scatter the birds and scare up stragglers. Then hunt the singles, because they will be more likely to hold.

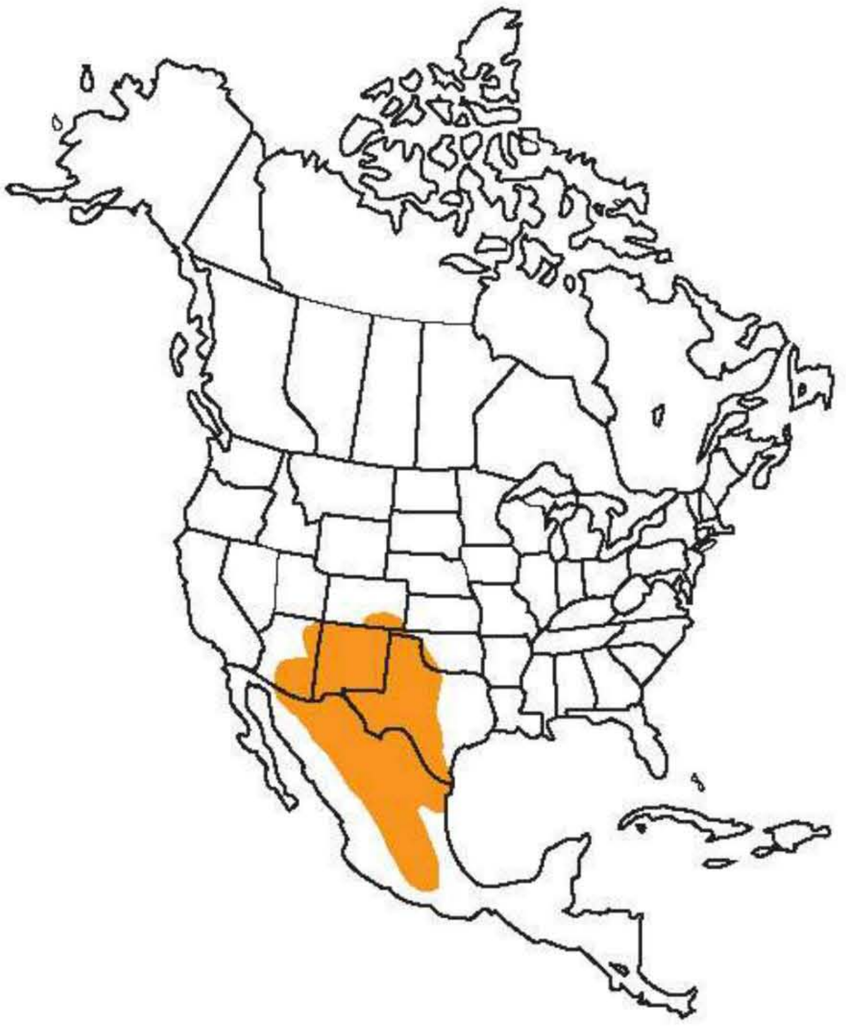

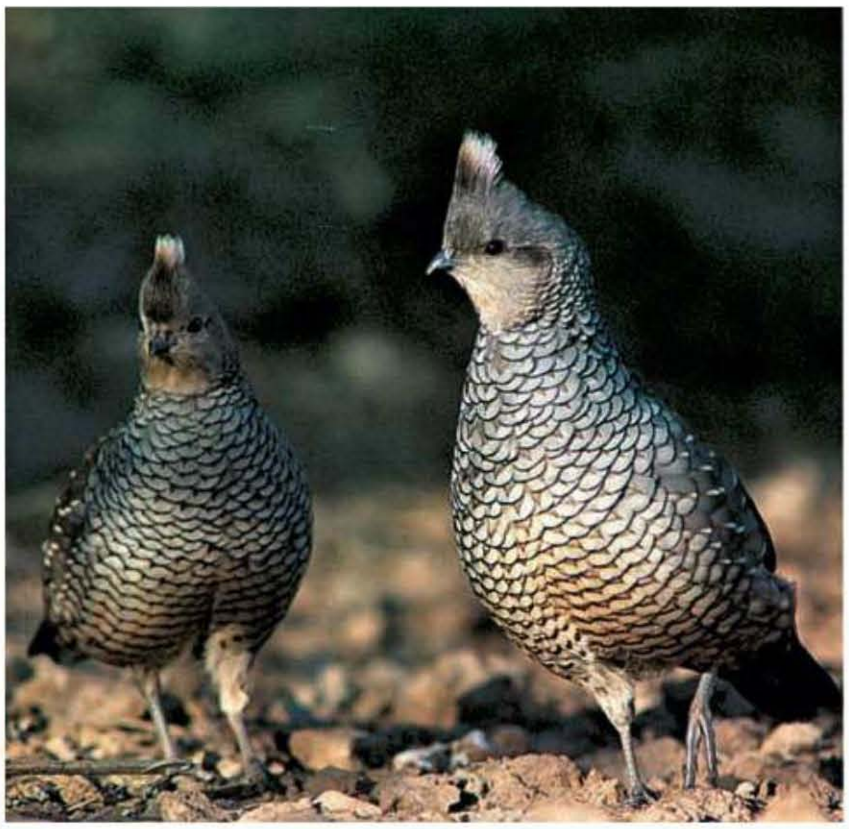

Scaled quail have breast feathers with dark margins, giving them a scaled appearance. The breast is bluish-gray; the head grayish-brown with a distinctive white-tipped crest. The crest is larger on males. Scaled quail weigh 6 to 7 ounces (170 to 184 g).

RUFFED GROUSE

Some hunters maintain that a ruffed grouse intentionally selects a flight path that places a tree directly in the line of fire. Whether intentional or not, the ruffed grouse is a difficult target for even the best wingshooters.

Grouse accelerate rapidly, reaching a speed of 40 mph (64 km/h) in seconds. But they seldom fly more than 200 yards (183 m). Grouse do not run as fast or as far as most other upland birds. When disturbed, birds of both sexes utter a call that sounds like pete-pete-pete.

Many hunters call the birds partridge, but the term is a misnomer. The only true partridge in North America are the chukar and Hungarian. The name ruffed grouse comes from the ruff, a tuft of dark feathers on the neck.

Hardwood forests with small clearings and mixed-age aspen trees make the best ruffed grouse habitat. Aspens provide good nesting cover and a year-round source of food. Grouse also favor birch trees, but also live in many other types of hardwood and conifer-hardwood forests.

Besides aspen buds, leaves and twigs, other foods include berries, fruits, clover, nuts, insects and occasionally corn. On warm, sunny days with little wind, grouse feed most of the morning and again in late afternoon. Cold, snowy or windy weather drives the birds into dense thickets or conifer stands. Grouse usually roost on the ground or in conifer trees, but when snow depth reaches 8 to 10 inches (20 to 25 cm), they often burrow into the snow.

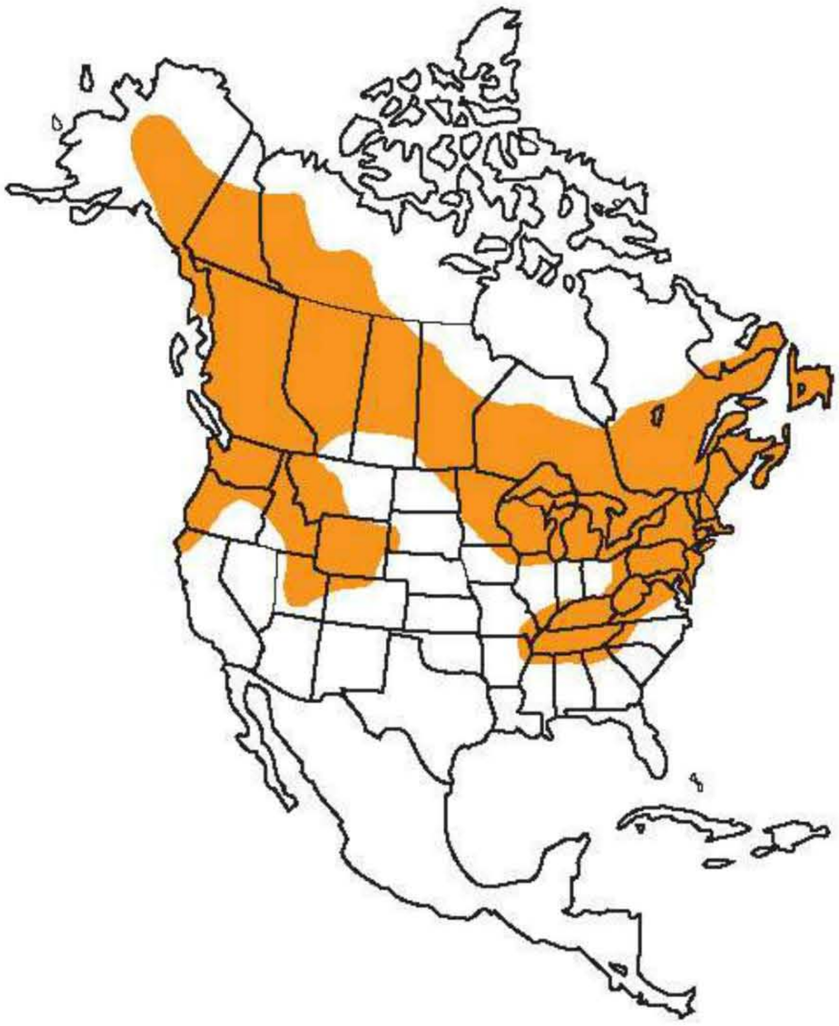

Before the spring breeding season, male grouse begin drumming to advertise themselves to females. A bird stands on a log, braces his body with his tail, then starts beating his wings. The sound resembles that of a one-cylinder gasoline engine starting slowly, then gradually speeding up. Some drumming continues through summer and into fall.

Grouse nest in April or May, producing a brood of eight to ten chicks. By fall, the birds are full grown and the brood scatters. Young grouse usually move at least 1 mile (1.6 km) and some relocate up to 10 miles (16 km) from the hatching site. Once they establish territories, they move very little the rest of their lives.

Young birds lack the wariness of those that have been pursued by hunters. Often a young grouse will sit on the ground or on a tree limb in full view of a hunter. But the birds gain experience quickly. They learn to flush well ahead or from behind a tree or patch of dense brush where the hunter has difficulty seeing them. In remote areas, adult grouse are no more wary than young birds.

Throughout much of the northern regions, ruffed grouse populations are cyclical, peaking about every ten years. Populations in peak years may exceed those in poor years by a ratio of 15 to 1.

Ruffed grouse are some of the most available but under-hunted upland birds. With a range from northern California to Alaska and east to the Atlantic, public land birds are available to anyone with a full tank of gas and a shotgun.

Hunting for Ruffed Grouse

Every grouse hunter has been frustrated upon hearing a whirr of wings, scanning the dense cover, then spotting a bird just as it disappears from sight.

In early season, you may scare up a dozen grouse for every one you see in time for a shot. Even though the birds generally flush close by, leaves severely limit your range of vision. By the time the leaves fall, grouse have become spookier and tend to get up farther away. But they are much easier to see and seldom flush out of shooting range.

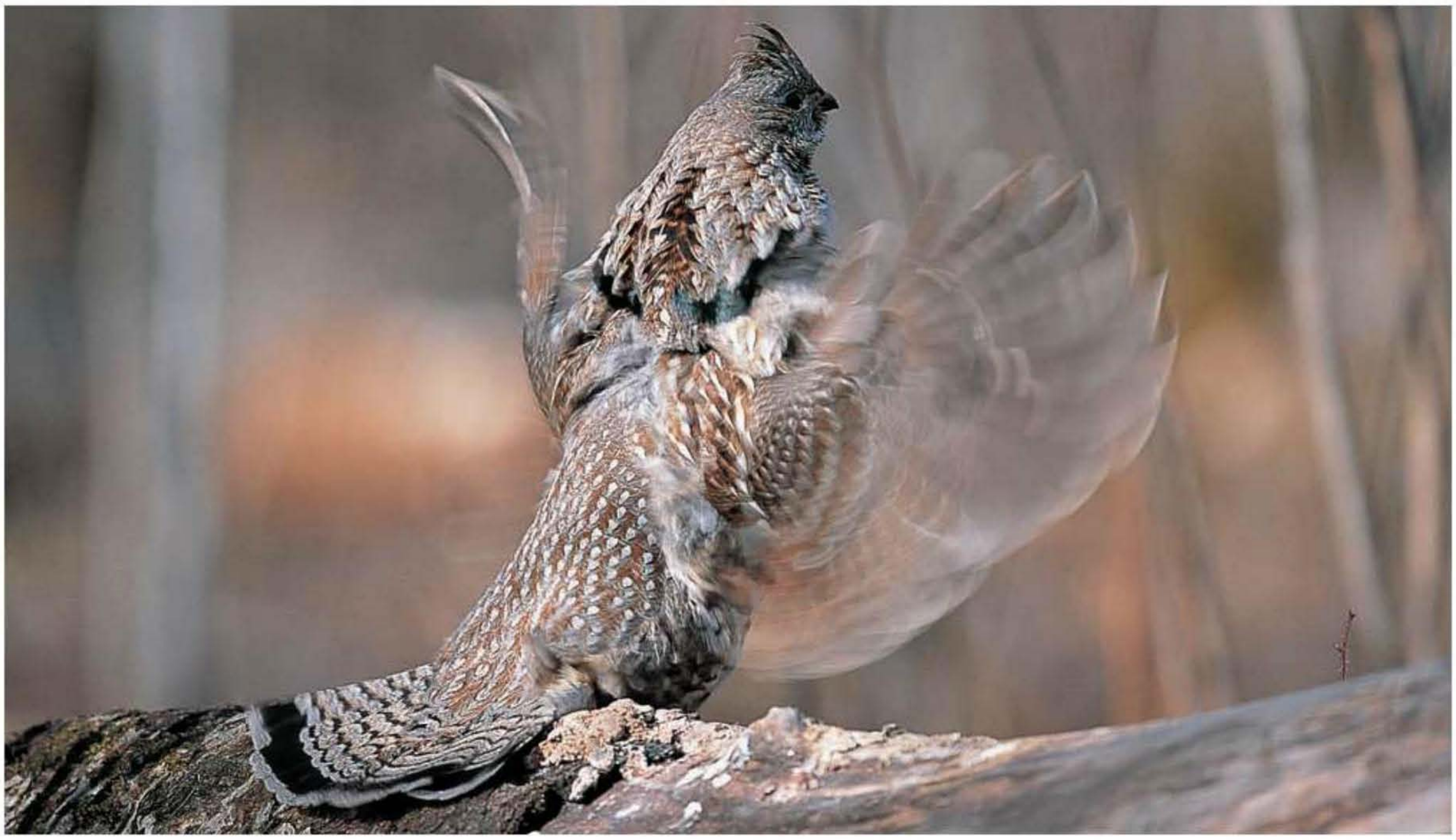

Ruffed grouse have a broad, black or dark brown band on the tail. The band is unbroken on most males, but broken on the center two feathers on females. There are two color phases, red and gray. Red-phase birds are more common in the South and at low altitudes; grays predominate in the North and at high altitudes. Grouse measure 17 to 20 inches (43 to 51 cm) long. They weigh 16 to 24 ounces (454 to 680 g), but occasionally reach 2 pounds (0.9 kg).

A good dog improves your grouse hunting success. But a dog is not essential because grouse are easier to flush than pheasants or quail. Most hunters prefer pointing breeds, but flushers and retrievers can also be effective if they work close to the handler. Because grouse are so well camouflaged, they can be difficult to find without a good dog.

When hunting without a dog, walk through likely cover and stop frequently. You can sometimes hear the birds nervously clucking before they fly. With two or more hunters, walk on opposite sides of thick patches. One hunter is likely to get a shot when a bird flushes.

If you miss or fail to get off a shot, watch where the bird lands. A flushed bird often flies off to a dense thicket or lands in a tree. You may be able to approach close enough for a second attempt.

You can improve your success by checking the crop of a freshly killed bird to find out what it was eating. Then, hunt where these foods are plentiful.

Grouse hunting is usually best on calm, sunny days. It seldom pays to hunt in early morning because the birds are still on their roosts. They begin moving after the woods have dried off.

A short-barreled, 20-gauge shotgun with an improved cylinder choke is an ideal grouse gun. The short barrel makes the gun easy to carry through heavy brush and enables you to swing quickly. Most grouse hunters prefer No. 7½ or 8 shot.

SIGNS OF GROUSE ACTIVITY

Tracks are closely spaced and about 2 inches (5 cm) long. A grouse places one foot directly in front of the other, so the tracks form a straight line.

Droppings on a drumming log mean that ruffed grouse are nearby. The brownish or greenish droppings are about ¼ inch (6 mm) in diameter.

Holes in the snow may be burrow-roosting sites. Burrows have an entrance hole and, if the bird has left, an exit hole.

Pointing dogs pin down grouse, enabling you to get close enough for an unobstructed shot. When you move in to flush the bird, avoid walking into dense cover where shooting would be difficult.

GROUSE HUNTING TECHNIQUES

Work the edges of cover. Grouse spend much of their time feeding on fruits and berries that grow along sunlit borders. If a bird flushes, your chances of getting a clear shot are better than in a thick woods.

Walk along a logging road, powerline cut or other trail through the woods. Trails make for easy walking and the edges produce food that attracts grouse. Try to avoid backtracking along the same path.

Listen for drumming ruffed grouse while hunting. Although most drumming activity takes place in the spring, some males continue to drum into the fall, revealing their locations to hunters.

TIPS FOR HUNTING RUFFED GROUSE

Plan your strategy using detailed maps of your hunting locale. State forestry departments may publish fire-control maps that show trails and logging roads.

Attach a bell to the collar of your pointing dog. This helps you keep track of its location. When the bell stops ringing, the dog is on point.

Scan the trees as you walk. Listen for a short flutter of wings that means a grouse has hopped up to a branch. Grouse in trees evidently feel safe, often allowing hunters to approach within gun range.

Other species are quicker to run or flush, but grouse hold tight while a hunter walks by. Its plumage blends so perfectly that you may never see the bird. That is until it flushes behind you.

SHARP-TAILED GROUSE

A novice hunter may mistake a sharptail for a hen pheasant, realizing the error only after the bird sails out of range. The two are about the same size and color, but a sharptail has a short, whitish tail and usually clucks when it flushes.

Sharptails prefer large ungrazed grasslands with numerous pockets of trees and brush for cover and fields of small grain for food. They favor wheat, oats and barley, but also eat berries, seeds, clover and buds. The birds feed in grain stubble or weed patches in early morning, loaf in wooded or brushy areas in midday, then feed again in late afternoon. Sharptails normally roost on the ground.

In early fall, a flock consists of six to eight birds. Later, many flocks combine to form a large pack which may have over 200 birds. When the birds flush, the flock stays together. They usually fly at least ½ mile (0.8 km) and occasionally a mile or more. But a bird or two may hold, so work the area thoroughly. Sharptails are strong runners and sometimes elude hunters by racing out the end of a field as the hunt begins.

Early-season hunters use 12- or 20-gauge shotguns with improved cylinder or modified chokes and No. 6 shot. After the birds form packs, they are harder to approach, so a full choke may work better.

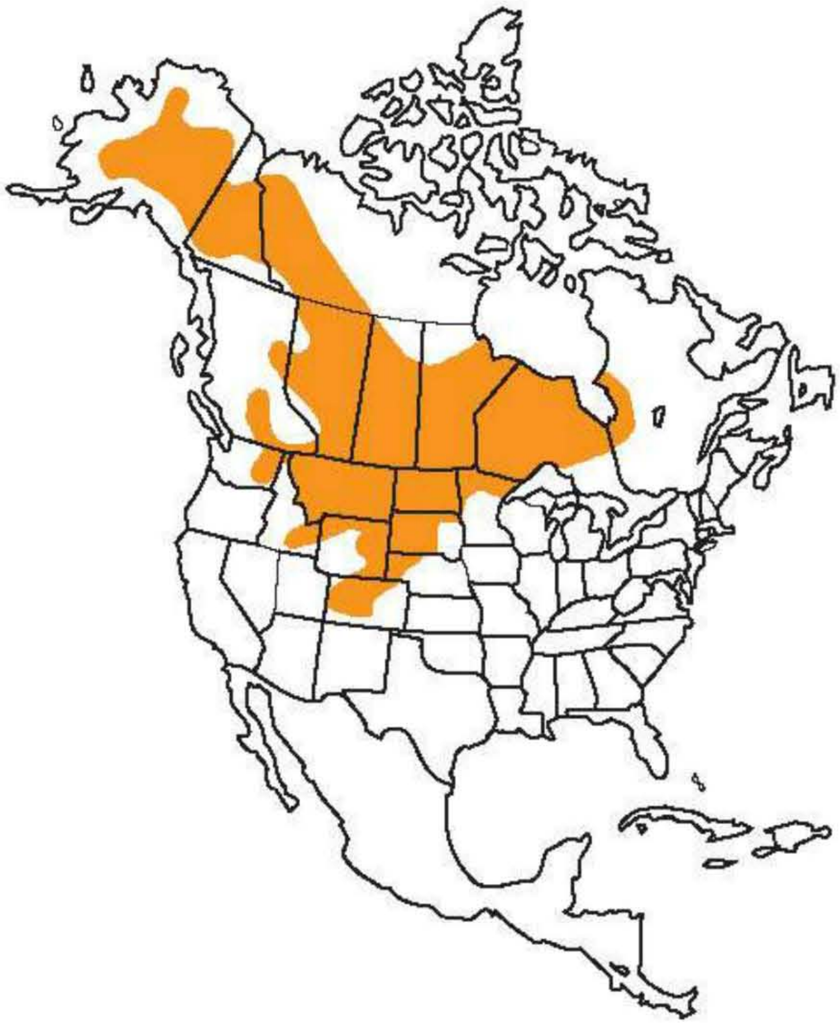

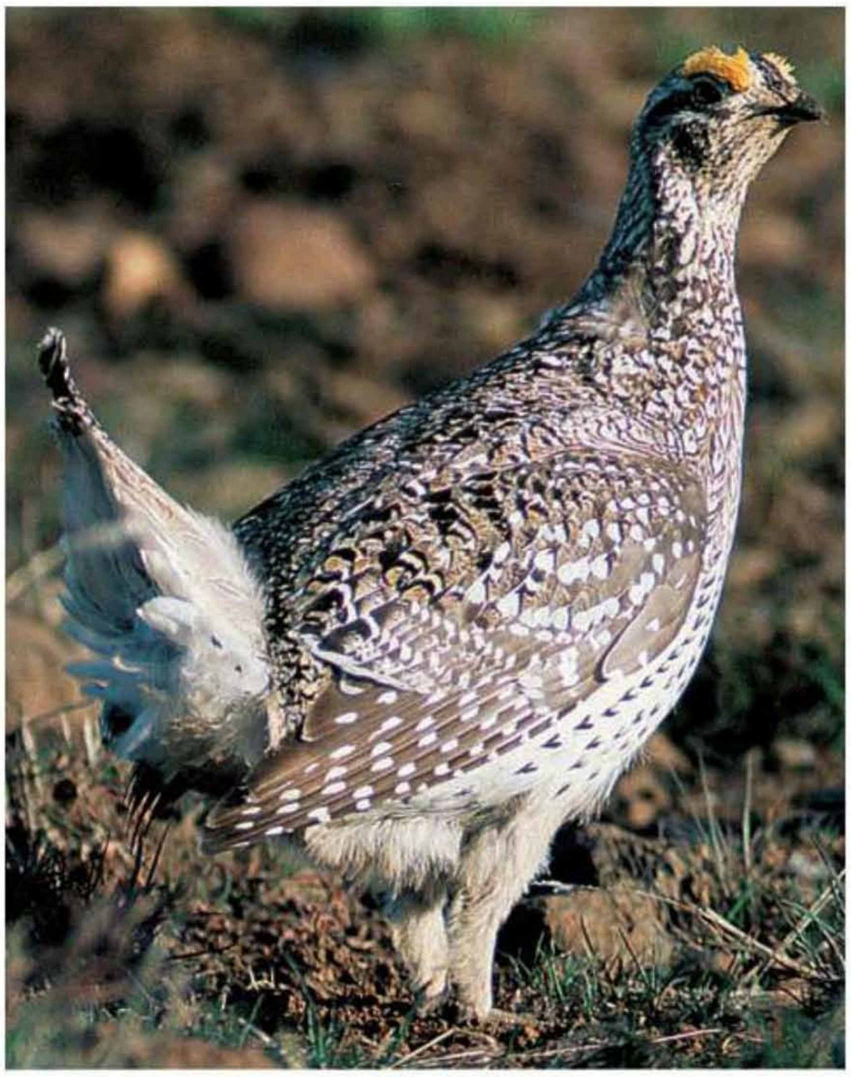

Sharptails are named for the two long feathers in the center of the tail. The brownish back and wings are covered with white spots. Males have a patch of yellow skin above each eye. These patches are less visible after the breeding season. The birds weigh up to 2 pounds (0.9 kg).

WHERE TO FIND SHARPTAILS

Look for sharptails in stubble fields early or late in the day. Carry binoculars to spot the birds and listen for clucking sounds.

Hunt along brushy ravines in windy or stormy weather. A ravine, grassy ditch or the lee side of a hill makes a good windbreak.

Check brush patches to find loafing sharptails. Other good loafing spots include shelterbelts and grassy areas around abandoned farms.

SAGE GROUSE

In the 1800s, up to 1.1 million sage grouse could be found in parts of what are now sixteen states and three provinces. Today, the sage grouse is a recluse, its population numbering around 160,000, found on sagebrush ground from Oregon through Wyoming and points north.

The birds inhabit open, semi-arid country with large sagebrush flats. Buds, leaves and shoots of sagebrush make up most of their diet. They also eat alfalfa, small grains, grasses, clover and berries. The birds feed along sagebrush edges in early morning. You can find them near water holes where they eat green vegetation. They rest in gullies and draws in midday, then feed again starting in mid-afternoon.

In early fall, birds may be found in groups of six to sixty, on open flats, plateaus, ditches and ridge tops. Hunt the edges of agricultural land, where the fields meet the flats. For best success, hunt close to stock tanks and ponds. In dry times, grouse may fly in from long distances to get the moisture they crave.

To find sage grouse, glass from a ridge in early morning. Check open areas and edges of cover. If you spot birds, attempt a stalk. If not, identify water holes, sagebrush edges and ravines that you could hunt later. A pointing dog works well in early season. But when the birds start to run, a flusher or retriever is more effective.

At the flush, the birds come up in ones and twos, giving you the time to pick a target, swing with it and shoot.

Use a 12-gauge with a modified or full choke and No. 6 shot to down sage grouse. Some hunters shoot only smaller birds. The meat of older, larger birds tends to be tough and gamey.

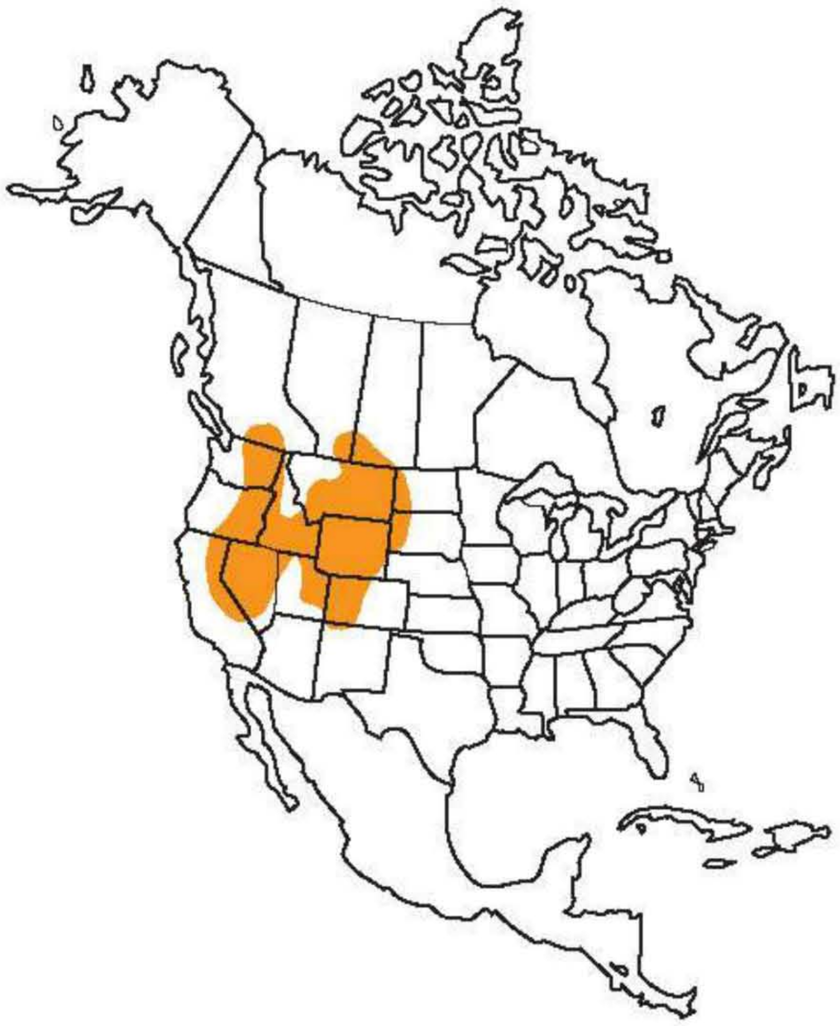

Sage grouse have a mottled, grayish-brown back, a pale breast and flanks and a black belly. The long, pointed tail feathers fan out in flight. Males have a black patch on the chin and upper neck and a white V across the throat. Females are much smaller than males. The mature male sage grouse can weigh up to 7 pounds (3 kg) and stretch the tape to 30 inches (76 cm) from his beak to the tip of his tail.

WHERE TO FIND SAGE GROUSE

Hunt near stock tanks, ponds, irrigation ditches or creeks, especially in dry weather. These spots are best in early morning and just before dark. Sage grouse are seldom more than ½ mile (0.8 km) from water.

Rush a flock of sage grouse once you stalk within shooting range. They cannot get off the ground as quickly as other upland birds, so this tactic enables you to gain yardage for a closer shot.

CHUKAR PARTRIDGE

Chukar habitat is steep, dry, almost inaccessible and generally unsuitable for anything but wildlife. Much of the good hunting is on public land and access is generally not a problem. Chukars use rocky slopes for roosting and escape cover.

Named for its call, the chukar was imported from India in the late 1800s. Populations became established in arid portions of the Northwest. Typical chukar habitat varies from bare rocks to sparse grasslands. Because of the open country, the coveys are easy to see, but difficult to approach. They run uphill at the sight of a hunter. When they flush, they usually fly downhill.

Seeds, grasses, leaves, insects and fruit make up most of the chukar's diet. The coveys, which average ten to twenty birds, begin feeding in mid-morning. On cool days they may feed through the afternoon. They often cover over 1 mile (1.6 km) on their feeding rounds. In hot weather, chukars spend the middle of the day loafing in the shade of rocky bluffs or near springs and water holes. In windy or stormy weather, they seek depressions or crevices in the rocks.

Early in the season, when the weather is dry and hot, chukars will not be far from water. After fall's first rains, they may be more widely dispersed. Cheatgrass is one of the chukar's principal food sources, though the birds may also be found in agricultural areas close to the river and canyon habitat they prefer.

In summer and early fall, chukars can be found at altitudes up to 11,000 feet (3,353 m). A heavy snowfall will drive them to lower elevations.

A 12- or 20-gauge modified-choke shotgun makes a good chukar gun. Most hunters use No. 6 or 7½ shot.

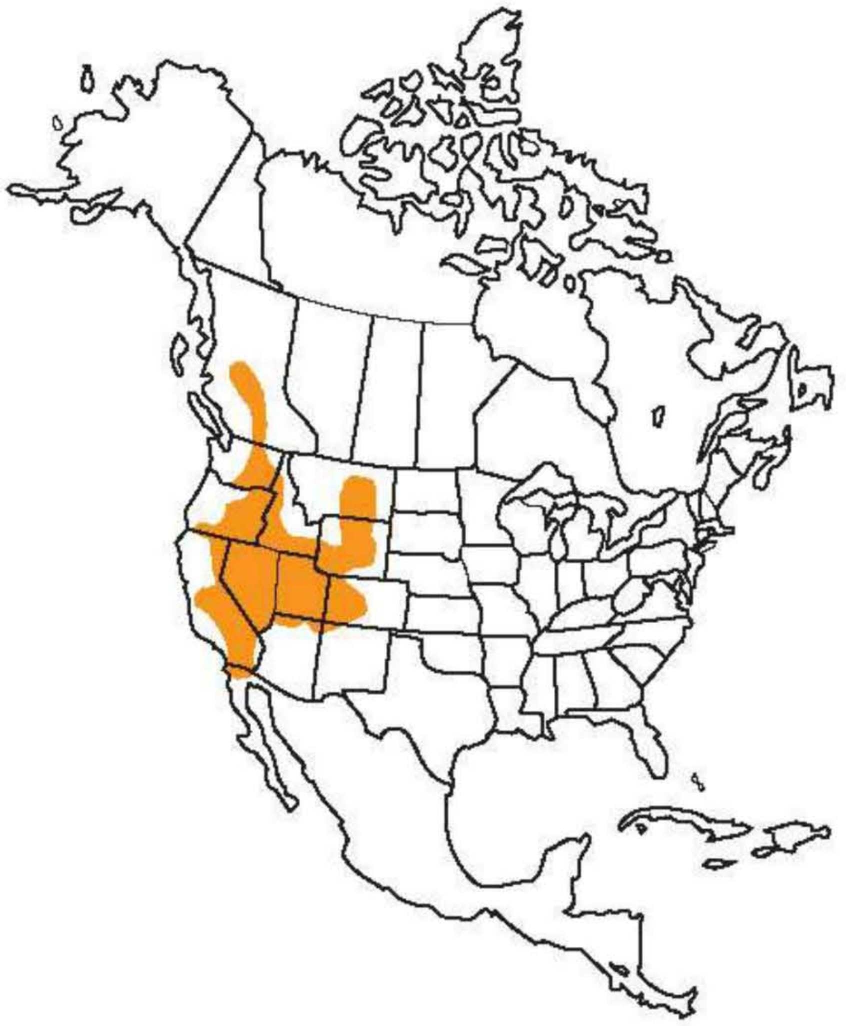

Chukars have a grayish-brown back, a reddish bill and legs and a black collar that passes through the eye. Black and chestnut bars cover the flanks. Chukars average about 15 inches (38 cm) long and weigh 18 to 24 ounces (510 to 680 g).

HOW TO HUNT CHUKARS

Hunt with pointing dogs to help pin down the birds and prevent them from running. When your dog detects fresh scent, move into shooting position quickly because chukars may not hold to a point for long.

Follow chukar tracks after a fresh snowfall. In addition to providing good tracking conditions, snow makes the birds hold tighter and prevents them from running as fast as they do on bare ground.

HUNGARIAN PARTRIDGE

Hungarian partridge are birds of the wide open spaces and can be found in cropland, sage and bunchgrass country in Washington, Oregon, Nevada, Idaho, Montana, Wyoming and points east. Look for foothill habitat close to irrigated ag lands for the best hunting.

Huns have an understated beauty, with a form that follows function. Wings of white and mottled brown allow the bird to hide from predators by holding still. A pastel-gray on the breast blends on the head with shades of brown. Their eyes are surrounded by a pencil-thin border of red. Flanks and breasts are splotched with markings of chocolate brown.

Hunters find the birds in short, light cover. Huns feed mainly on grains like corn, oats, wheat and barley, but also eat weed seeds and green leaves.

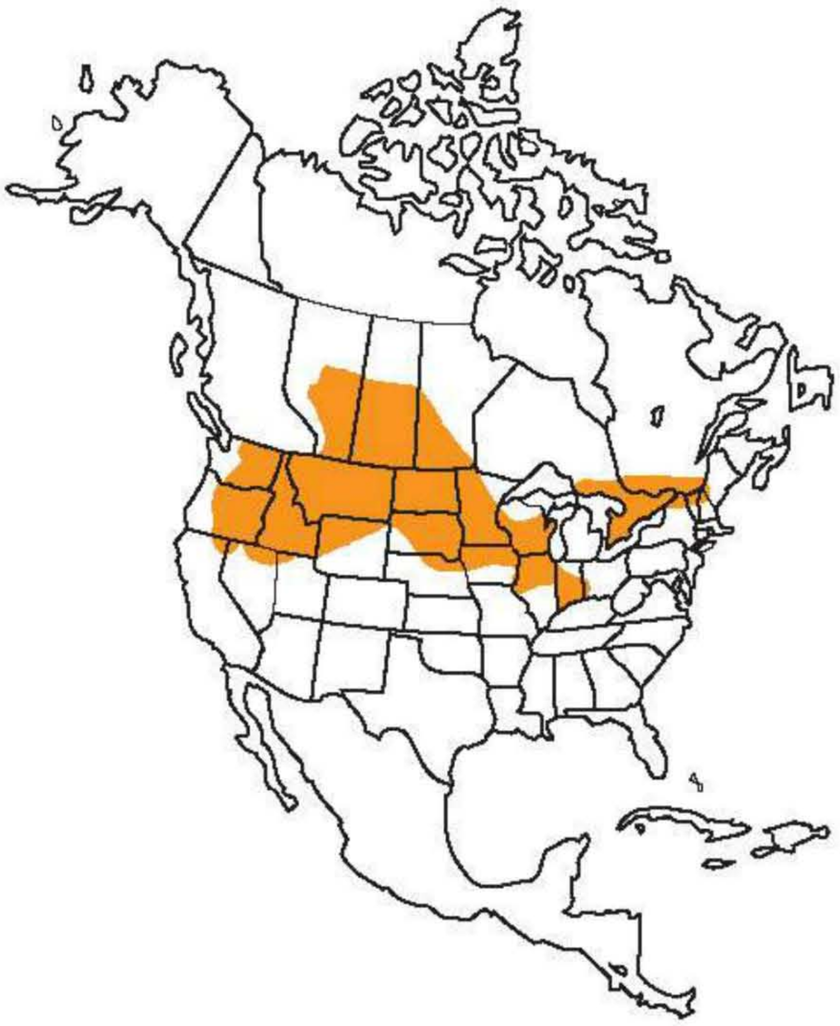

Huns begin feeding after most of the dew dries off the grass. They feed until late morning, loaf in a grassy area until mid-afternoon, then resume feeding until dusk. The birds roost in alfalfa fields, grain stubble, short grass or even on plowed ground. They often form roosting rings like those of bobwhite quail. In winter, huns may roost in a depression in the snow or burrow-roost under it.

Coveys normally consist of ten to twelve birds. Huns do not hold well when first approached. They tend to sneak ahead of a hunter or dog, then burst from cover in unison. The covey remains together. Early in the season, huns fly only a short distance. As the season progresses, they go farther, sometimes over ½ mile (0.6 km).

Shotguns and shells used for huns are similar to those used for chukars. Some hunters prefer 12-gauge guns with full chokes in late season.

Huns are as spooky as chukar, faster than quail and smart enough to run, circle and hide. Add to all that its taste on the table and this import from Hungary is a welcome and worthy quarry for the best dogs and hunters.

Hungarian partridge are technically named gray partridge. Native to Hungary, they have a cinnamon head and chestnut bars on the flanks. Males have a chestnut horseshoe on the breast. The birds weigh about 1 pound (0.45 kg) and measure 12 to 14 inches (31 to 35.5 cm) long.

HOW TO HUNT HUNGARIAN PARTRIDGE

Watch the covey land, then approach the spot from opposite sides. This prevents the birds from slipping away and assures someone of a shot. You can often flush the same covey again.

Hunt with a close-working retriever or flusher or use a pointing dog. Flushers and retrievers work best in high cover; wide-ranging pointers find more birds in sparse grasslands or other open terrain.

WILD TURKEY

Every year, hundreds of thousands of hunters head out in pursuit of a gobbler. Winter's weariness is receding with the snowline. Forests are coming to life. Fields of wildflowers are awash with the yellows, pinks and blues. There are spotted fawns in the thickets and new growth on the trees. To many hunters turkey hunting is a celebration of spring.

The wild turkey is the largest game bird in North America. It is found throughout North America (except for Alaska). Hawaii has an abundant (introduced) turkey population. There are five recognized subspecies, which vary slightly in color and size.

First-time turkey hunters may wonder if they could tell a male from a female. There are definite differences.

The male wild turkey, called a tom or gobbler, is a large, robust bird weighing up to 30 pounds (13.5 kg) and standing as high as 4 feet (1.2 m) tall. His body color is brownish black with a metallic, iridescent sheen. The head and neck, nearly bald, vary from white to blue to red. Bright red, fleshy bumps, called carbuncles, droop from the front and sides of the neck and a fleshy flap of skin, called a dewlap, is attached to the throat and neck. A fingerlike protrusion called a snood hangs over the front of the beak. When the tom is alert, the snood constricts and projects vertically as a fleshy bump at the top rear of the beak.

On the gobbler's breast there is a beard, a series of coarse hair-like fibers. On a jake, an immature gobbler, the beard may be only 3 inches (7.6 cm) long. A mature male has a beard approaching 8 inches (20 cm) in length and longer. This is not a feature totally exclusive to gobblers though, as some females may sport a beard.

Another feature that sets toms apart from hens is spurs. Hens have no spurs. Spurs are found on a gobbler's lower legs and average between ½ and 1½ inches (2.5 and 3.8 cm) long. These spurs are small and rounded on young birds; long, pointed and usually very sharp on mature birds.

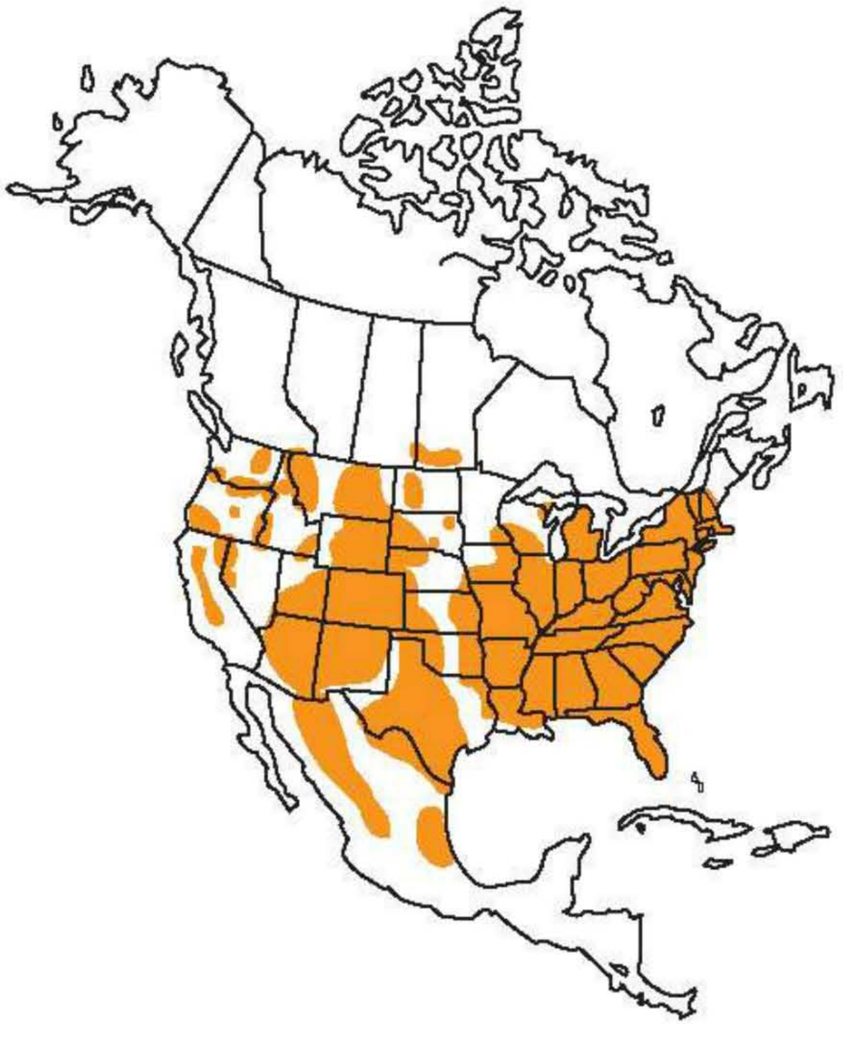

Toms begin their breeding displays in early spring. With tail fanned, feathers fluffed, and wingtips dragging, the tom struts boldly while emitting low-pitched hums.

In the spring season, big gobblers are looking for love. This makes them vulnerable to hunters who can call a tom into range by imitating the varied calls of the hen. Several styles of turkey calls are available, making it easy for a hunter to find one that suits his hunting style and abilities.

The male is called a gobbler for good reason: his rattling, deep-toned call is one of the most recognizable sounds in all of nature. At mating time, toms gobble with full-volume gusto, attempting to attract hens for breeding. Adult males display for hens by fanning their tail feathers, puffing up their body feathers and dragging their wings as they strut. Their heads and necks turn bright red during breeding season or when the tom is otherwise excited.

Adult females, or hens, are considerably smaller than toms, rarely weighing more than 10 to 12 pounds (4.5 to 5.4 kg). Their overall body color is duller than the male's and lacks his metallic, iridescent sheen. The hen's head and neck are usually blue-gray in color and sparsely covered with small, dark feathers. Carbuncles are sometimes present, but smaller than those on toms. Some hens grow small, rudimentary beards and spurs. Although they don't gobble, hens make a variety of cluck, purr, cutt and yelp sounds. Dominant hens may assert themselves with a display resembling that of the male, though they do not strut.

Juvenile birds mature quickly. By their fifth month, the juvenile male (jake) and juvenile female (jenny) closely resemble adult birds. However, juveniles have darker legs, which turn pink as the birds age. Jakes make feeble gobbles, higher in pitch than the calls of mature toms. Their beards are shorter in length and usually have amber-colored tips.

With its powerful legs, the wild turkey is an exceptional runner and has been clocked at speeds up to 12 mph (19 km/h). Although strong short-distance fliers, turkeys usually run when threatened. When necessary for escape, turkeys launch themselves with a standing leap or a running start and accelerate to 35 mph (56 km/h) in a matter of seconds. They cannot remain in the air for more than a few hundred yards (m), but can glide for ½ mile (0.8 km) or more when coasting down from a ridge top.

Subspecies

Subspecies can be difficult to distinguish from one another, since regional variations within the group can be more dramatic than physical differences between the subspecies. In addition, where the subspecies' ranges overlap, crossbreeding produces hybrid birds that show traits of both parents. A map showing the current range of all five subspecies is a good starting point in determining where subspecies live.

Turkey Calls and Decoys

Turkey calls have been an important part of the hunt ever since man discovered he could talk turkey and the birds would respond. Calls are deadly at drawing toms to hunters.

Calling is considered by most devoted turkey hunters to be the only sporting way to take a gobbler. Calling is the fun part of turkey hunting. Mastering a variety of different calls—learning when to yelp, cluck, purr or cutt, knowing when to call softly and when to crank up the volume—this is the art of spring turkey hunting.

Under the right circumstances, turkeys can be ridiculously easy to call. But don't count on it. When hens are abundant or hunting pressure is heavy, only a

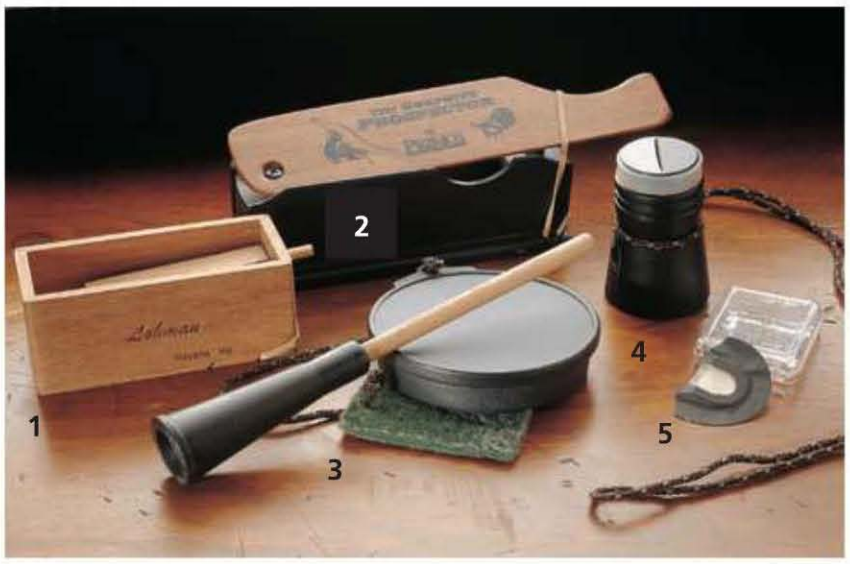

Turkey calls can be divided into two basic categories: Friction calls include the (1) push-pull call, the (2) box call and the (3) slate-and-peg call. Air-activated calls include the (4) tube call and the (5) diaphragm call.

PROPER DIAPHRAGM CALL PLACEMENT

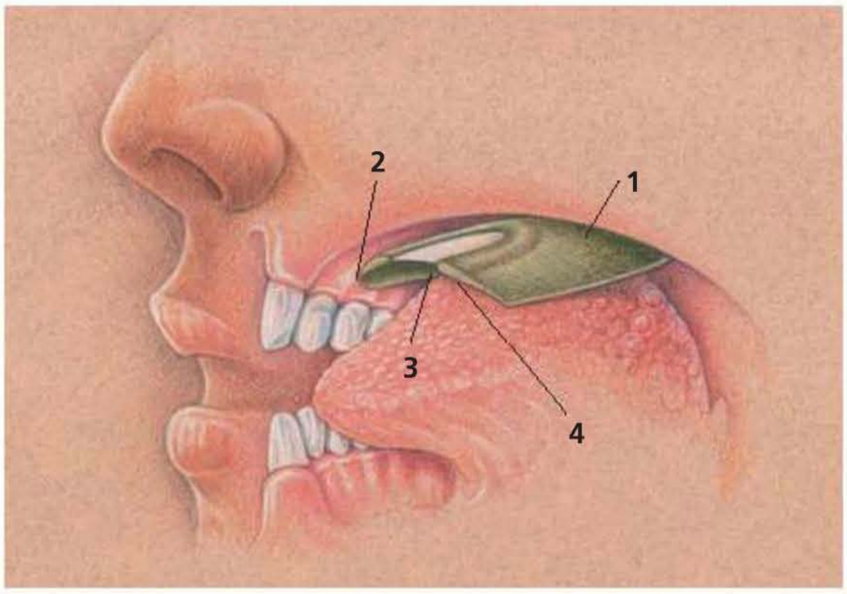

Proper placement of a diaphragm call in your mouth is crucial to success. (1) Use your tongue to position the call against the roof of your mouth with the straight edge facing forward. (2) Position the call so the forward edge nearly touches the back of your top front teeth. (3) Place the top of your tongue lightly against the latex reed and (4) expel short bursts of air between the top of your tongue and the reed, while saying the word "chirp." Stay at it and you will soon be able to produce acceptable yelps, the first call you should learn. Once you have mastered the yelp, begin practicing other calls.

CALLING TIPS

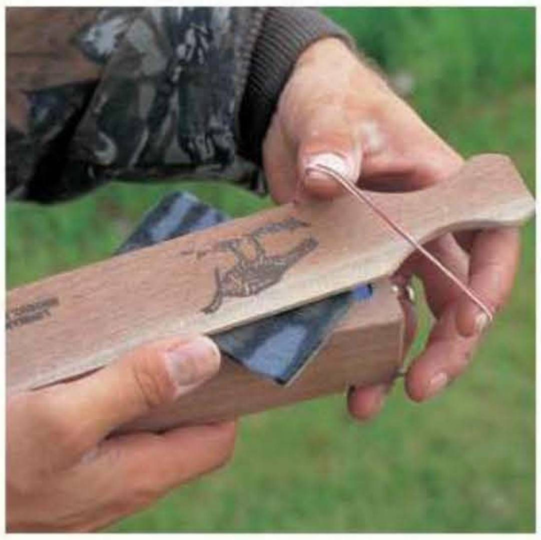

Insert a piece of cloth or fleece under the paddle, then wrap a rubber band around paddle and box. This keeps the call from squeaking as you walk.

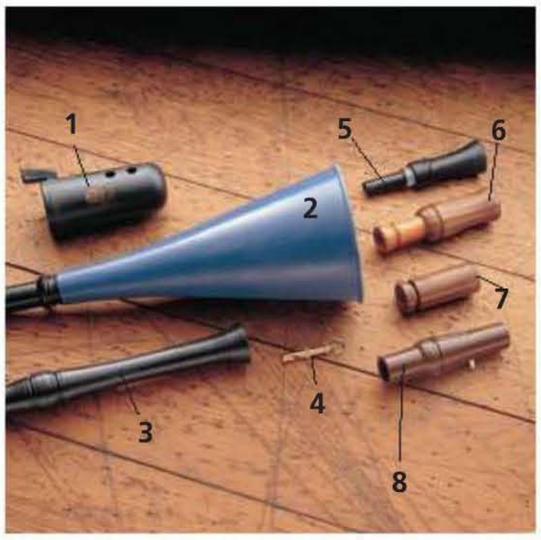

Dependable locator calls include: (1) owl hooter, (2) coyote howler with large bellows, (3) flute-style goose call, (4) silent whistle with adjustable frequency, (5) crow call, (6) pileated woodpecker call, (7) hawk screamer and (8) predator call.

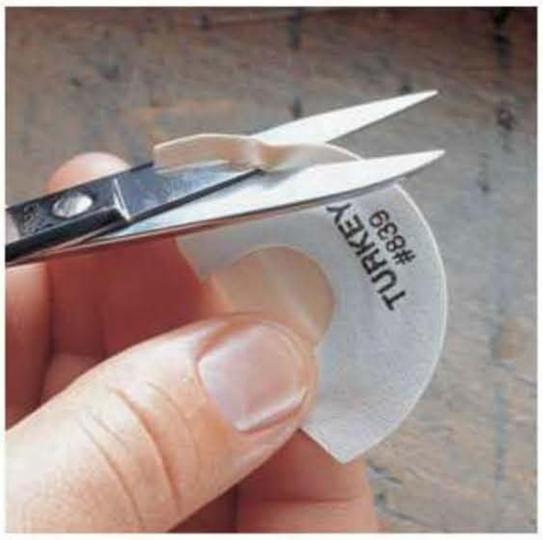

Trim the skirt of a diaphragm call with small scissors to make it fit comfortably against the roof of your mouth.

TYPES OF DECOYS

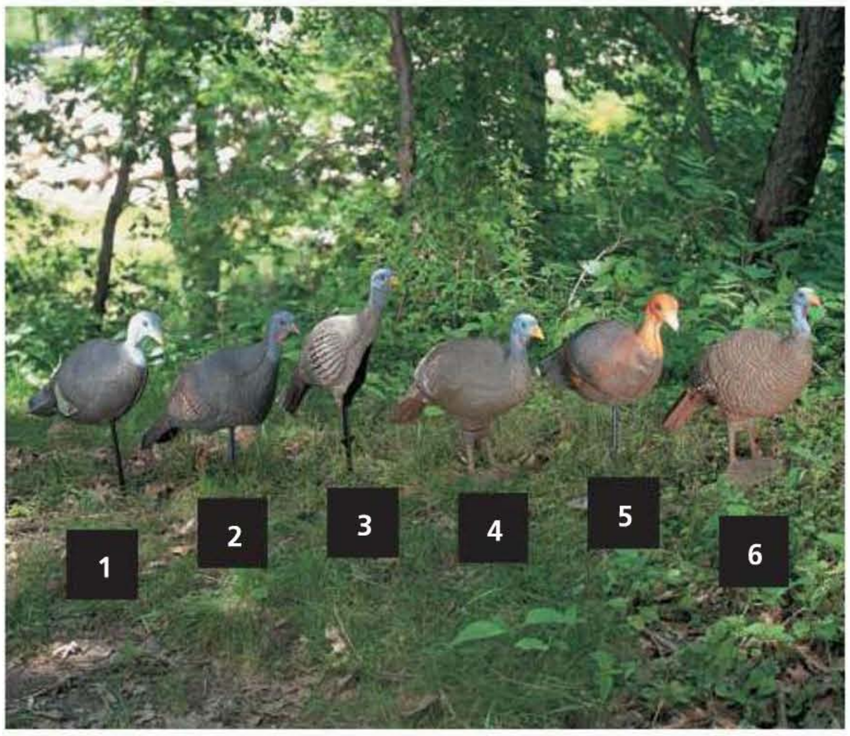

Hen decoy styles include: (1) collapsible foam; (2) folding shell; (3) silhouette; (4) full body; (5) hen with safety orange head, which is both safe and often aggravates toms; and (6) motion decoy, which allows the hunter to move the head of the decoy by pulling on a long string.

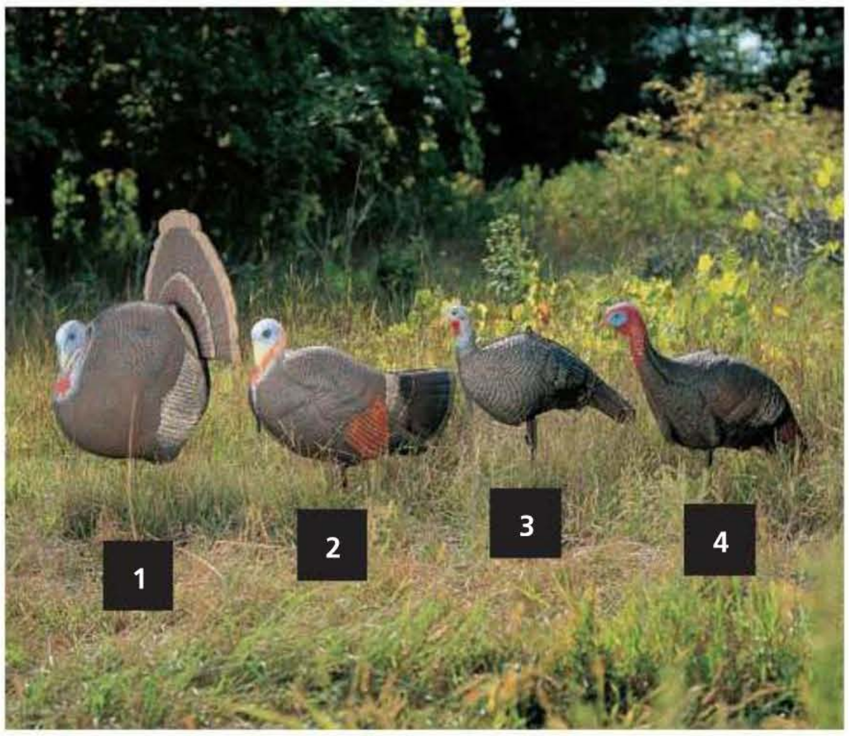

Jake decoy styles include: collapsible foam in (1) full strut; (2) half-strut; and (3) erect position; and (4) folding shell in erect position.

hunter who has mastered the art of calling walks out of the woods with a gobbler slung over a shoulder.

Calls designed to lure gobblers to the hunter can be divided into two broad categories: friction calls and air-activated calls. Although each has its advantages and disadvantages, the friction calls are generally easier to operate.

A third category of calls, locator calls, are used to elicit a response from a turkey without attracting the bird to your position. This response is most commonly referred to as a shock gobble. Turkeys have been known to shock gobble to the sound of a squeaking fence, slamming car door or a bellowing Holstein and a host of other sudden, loud sounds. Most locator calls imitate animals such as crows, owls and coyotes, but loud cutting on a turkey call can also prompt a shock gobble.

There are dozens of commercial calls on the market and while there is no need to buy and use all of them, a hunter is wise to use a variety of calls. Wild turkeys can be fickle, preferring a box call one morning, a diaphragm the next. A well-prepared hunter is ready to switch calls on those mornings when his old standbys just aren't making the grade.

However, as important as calling is, woodsmanship is equally important. In fact, an accomplished woodsman—a hunter intimately familiar with the land and the turkeys that live on it—can take gobblers with amazing consistency without ever touching a call.

A well-positioned turkey decoy is often the final touch needed to lure a shy tom into shooting range. Especially in open country, toms responding to a hen call may grow suspicious if they can't spot the source of the call. But if they see a decoy, the reassured gobblers often commit and waltz into range. A decoy can work wonders on hard-hunted birds, which tend to be leery of even the best calling.

Understanding Daily Patterns

At first light, as turkeys begin to awaken in their roosting trees, toms begin to gobble, while hens utter soft tree yelps. Usually the birds fly to the ground as soon as there is enough light for them to spot predators, but this fly-down can be as late as two hours after sunrise if the morning is rainy or foggy. Turkeys roosting in small woodlots in relatively open areas tend to fly down sooner than birds roosting in deep timber, where daylight is longer in arriving.

When a turkey gobbles on the roost in early morning, he is trying to draw a hen to him. If successful, the tom will fly down to join the hen. If a hen doesn't appear, the tom flies down and walks to his strutting area, while continuing to gobble.

To effectively hunt wild turkeys in spring, a hunter must understand the bird's daily movement patterns. Bagging a wild turkey is largely a matter of being in the right place at the right time.

Strutting zones are located in open areas where toms can be easily seen by hens, such as a field edge, logging road or pasture. In hill country, strutting areas are often located on elevated ridges, points or hilltops. In flat country, gobblers often strut in semi-open hardwood bottoms.

When the tom arrives at the strutting zone, he continues to gobble until he attracts a hen. Once a hen arrives, however, the gobbler begins to strut and his gobbling becomes infrequent. The tom may attract several hens simultaneously.

A gobbler follows hens as they wander to feed, mating with each several times during the morning, if they permit. Identifying feeding areas can be difficult, since wild turkeys eat such a wide variety of food. In big timber, turkeys often wander along, foraging on whatever they find. In agricultural areas, turkeys feed in crop fields a good deal of the time.

Only in the arid regions of the West and Southwest do turkeys habitually seek water during their daily route. In most regions they obtain sufficient water from dew or in the plants they consume.

By mid- to late morning, after satisfying their hunger, turkeys often treat themselves to a dust bath. Bred hens may then retire to the nest to lay or incubate. Gobblers may rest at this time or resume gobbling and continue to search for new hens.

Mid-afternoon finds turkeys on the move and feeding once again. During the last hour of day, the flock feeds toward roosting sites. They usually fly up at sunset or shortly thereafter. Birds often change limbs or even roosting trees a number of times before settling in for the night.

Turkey Hunting

One wrong move and the turkey will duck his head and run. Don't move a muscle, not yet. Let him close to 30 yards (27 m). He stops to strut, fanning out his tail feathers and puffing out his chest as if to say, "Look at me, ladies. Ain't I the pretty one, though?"

Ideally, you'll want to "roost" a gobbler on the eve of the hunt, which means you actually see or hear the bird in his roost tree. Return before first light, slip within 100 yards (91 m) of this roost, sound a couple of tree yelps at dawn and odds are good the gobbler will fly down practically in your lap.

To "put a bird to bed," set up near a roosting site in early evening. Turkeys feed toward a roost tree in late evening, usually fly into them around sunset and settle in before dark.

If you cannot spot birds, listen for them. Occasionally, toms gobble a time or two after roosting and hens sometimes yelp. More often, you'll have to make them gobble by using a locator call. Walk through likely habitat, pausing every ¼ mile (0.4 km) or so to call. If a gobbler responds, move toward him, eliciting additional shock gobbles until you've pinpointed his location. Mark the site and the trail to it with strips of toilet paper or surveyor's tape so you can find your way back before dawn.

As day breaks, give a few quiet tree yelps. Chances are good that the gobbler will reply, but if he doesn't, resist the temptation to call louder or more often. Overcalling while a tom is still on the roost is a common mistake.

If you failed to put a gobbler to bed the night before, plan to be at a good listening position well before first light. Hilltops and ridges are best. On most mornings, toms begin gobbling spontaneously, but if you hear no gobbling by the time there is enough light for you to see individual limbs on trees, then try to elicit a gobble with an owl or crow call. When you hear a gobbler within reasonable distance, try to get within 100 yards (91 m) before setting up to call him.

Turkeys usually fly from the roost as soon as there is enough light for them to see predators on the ground. However, on rainy, foggy or unseasonably cold mornings, they may stay on the roost for an hour or more after sunrise.

They use the same strutting grounds day after day. These are usually in relatively open areas, such as open timber, field edges, pastures or logging roads, where toms can watch for approaching hens and hens can clearly see their strutting displays. Once you find these sites, you can hunt them productively for years, because turkeys return to the same strutting zones season after season.

Hunters often make the mistake of yelping loudly and frequently. Call just enough to let a gobbler know you are there. Every few minutes, scratch the ground to mimic the sound of a foraging turkey. This assures the gobbler that a hen is still nearby, but since she ignores his calling, he sometimes goes against his own instinct and sashays toward her.

The prime early morning strutting/gobbling period may last a few minutes to an hour. Occasionally, a tom returns to his strutting zone in late morning after his original partners abandon him.

Two strategies can produce results after the prime early morning period: run-and-gun or sit-and-wait.

To use the run-and-gun approach, walk quickly through the area, stopping frequently to blow a locator call. Search for a bird that responds to every call you make. When you find such a bird, get within 200 yards (183 m) in open country, half that distance in timber; sit down and get ready before you make a single hen call. Call just often enough to keep the bird gobbling, so you know where he is. If the bird continues to gobble in response to your calls, but moves steadily away, determine his direction and attempt to get ahead of him. If he moves steadily in your direction, call sparingly and stay ready.

The sit-and-wait tactic is most successful if you know the area well enough to predict travel routes. Picking a good feeding spot and waiting patiently for a flock of hungry birds to arrive is also a good tactic before 10:00 a.m. A blind can be a real asset when playing this waiting game, as can a decoy or two. Patience is the key. Turkeys feed at midmorning and a feeding flock is in no hurry, covering only about 200 yards (183 m) per hour, on average.

Cover lots of territory after 10 a.m. Use locator calls and loud cutting to elicit gobbles. Locator calls are better than hen calls for prompting a tom to begin gobbling. Once you've elicited a shock gobble, set up and get ready to seduce the bird with your most authentic hen calls. If you begin by making hen yelps, a tom may arrive before you are prepared to deliver the proper greeting.

The first step in a successful hunt is to find good turkey habitat. Assuming you have scouted adequately, you should know several spots that hold birds.

If you prefer to sit and wait for your bird, late morning is a good time to stake out strutting zones. Gobblers that have been deserted by hens frequently meander back to their strutting zones and readily respond to calls.

Gobblers can still be taken in the afternoon, though it takes patience and a slightly different strategy. Find a turkey hangout—a dusting area or feeding site—set out a decoy or two, relax against a big tree and give a few yelps every 15 minutes or so. If a gobbler comes to your call at this hour, he usually won't announce himself until within shooting range. Many hunters have been jarred from midday naps by tiptoeing gobblers shouting at close range.

Action picks up again in late afternoon, when turkeys begin actively feeding and, later, working toward the roost. If you have located a roosting area, determine the most likely direction from which the birds will approach and set up in their path. If you haven't pinpointed a roosting area, spend the last hour of daylight covering ground and using a locator call. If you hear a gobbler, slip in close and work the bird with hen calls. Even if you don't get a shot, you'll know where to continue the hunt in the morning.

HUNTING STRATEGIES

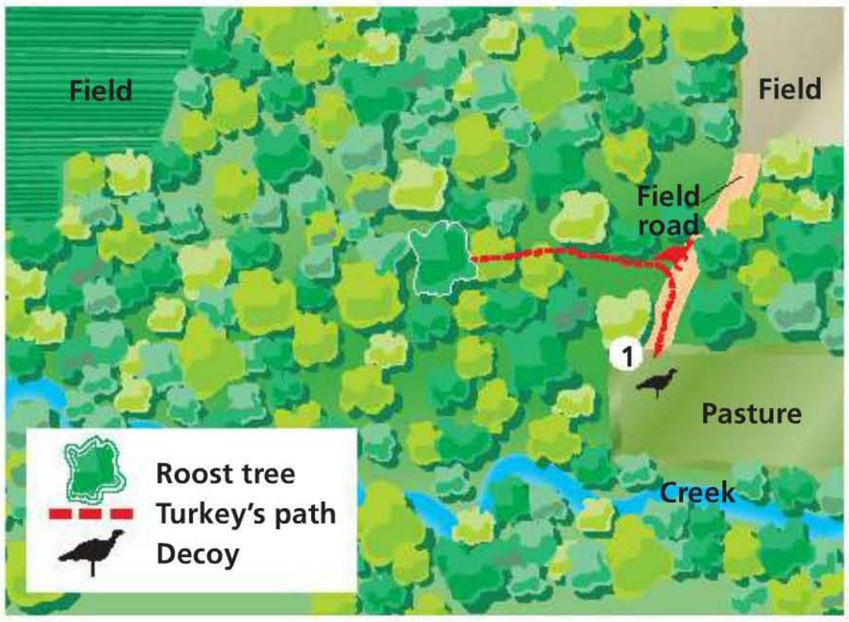

For typical early morning scenarios, set up in the nearest opening to the roost tree and put out a decoy. If the gobbler is alone, he will go to the field to gobble and strut. If he is with hens, he tags along as they go to feed in the field. The logging road is the obvious approach route.

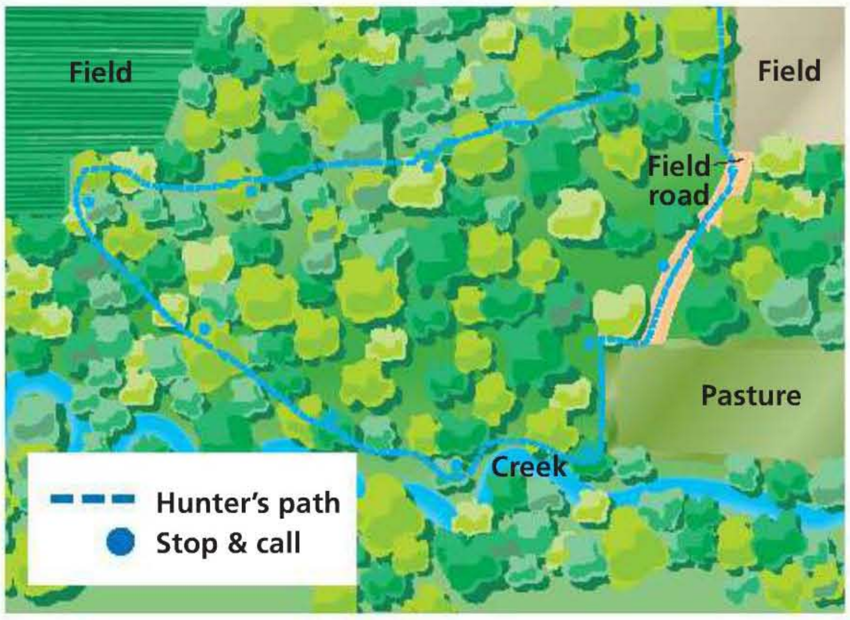

If you want to set up a typical run-and-gun scenario, follow the easiest route to cover the most ground in the least amount of time when running and gunning. Skirt the edges of all openings and call every 100 to 200 yards (91 to 183 m), until you get a response.

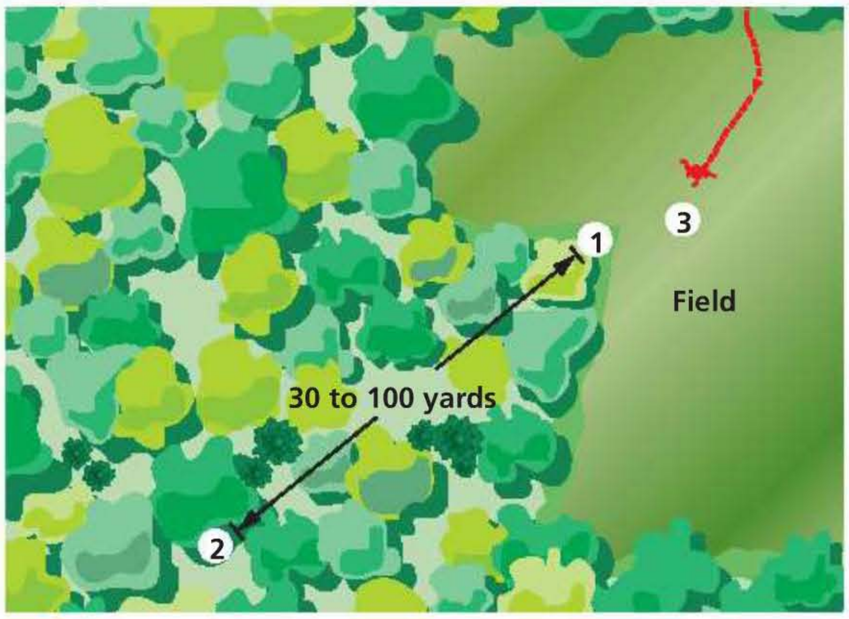

In a tag-team setup, position the shooter (1), 30 to 100 yards (27 to 91 m) in front of the caller (2). If a gobbler hangs up short of the caller, the bird will be within range of the shooter.

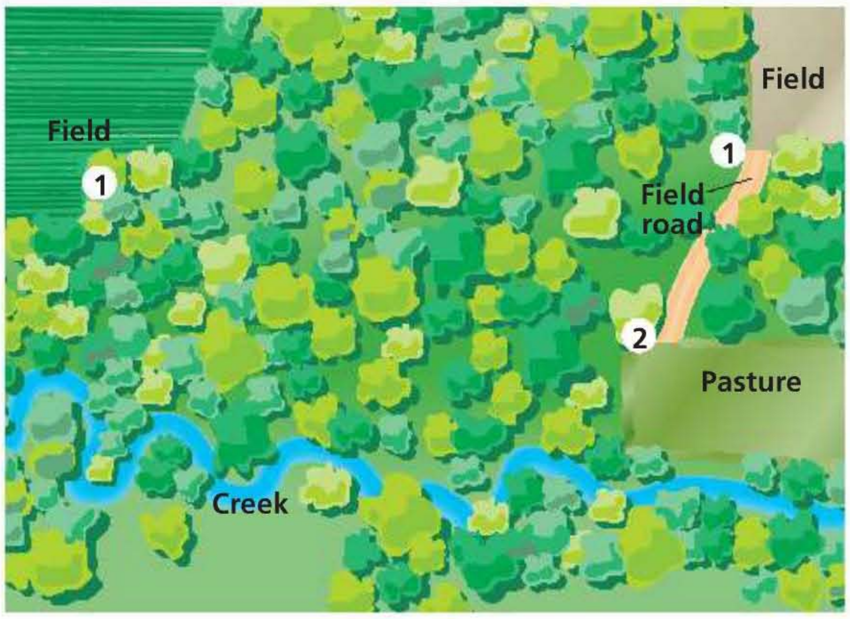

For a typical sit-and-wait scenario, sit and wait for birds where you can cover locations where birds enter or exit fields (1), pastures (2) or other openings. Be patient, call sparingly and stay alert.

MOURNING DOVE

It's hard to deny the thrill of seeing dozens of doves sweeping toward you and beyond, the sound of the wind in their wings. There's no more difficult wingshooting challenge than that you find on a dove hunt. These birds are hard to hit. You need an edge if you hope to bag enough for dinner.

North American hunters harvest about 50 million doves annually, more than all other migratory game birds combined. The continent's dove population is estimated at 500 million, making it the most common game bird.

Mourning doves don't like cold weather. In northern states, the birds begin flying south after the first frost. Prior to migration, they form flocks consisting of several hundred birds.

Doves prefer open fields with scattered trees and woodlots. The birds need water each day, so there must be a pond, stock tank, flooded gravel pit or river nearby. Doves nest in grass, shrubs, stubble fields or trees, especially evergreens.

Nesting begins in early spring and continues through early fall. Doves raise up to four broods, with both parents caring for the young. After predation and weather-related losses, broods average about one bird each. This trait of raising multiple broods is unique among game birds. It ensures a relatively stable population from year to year, with or without a hunting season.

Weed seeds form the bulk of the dove's diet. They favor foxtail, dove weed, ragweed and wild hemp. Other foods include corn, soybeans, sunflowers, oats and wheat.

Doves fly to their feeding areas at dawn, feed until mid-morning, then head for water. Through midday, they rest in dead or dying trees near a feeding area, a water hole or a small pond. Doves resume feeding in late afternoon, then return to a watering site late in the day.

At dusk, they fly to their roosts, often to the same trees they used during the day. If not disturbed, doves continue to use the same roosting and watering areas.

Mourning doves can fly up to 55 mph (88.5 km/h). Their normal flight is smooth, effortless and direct. But they flare quickly at the sight of hunters, darting and weaving erratically as they fly away.

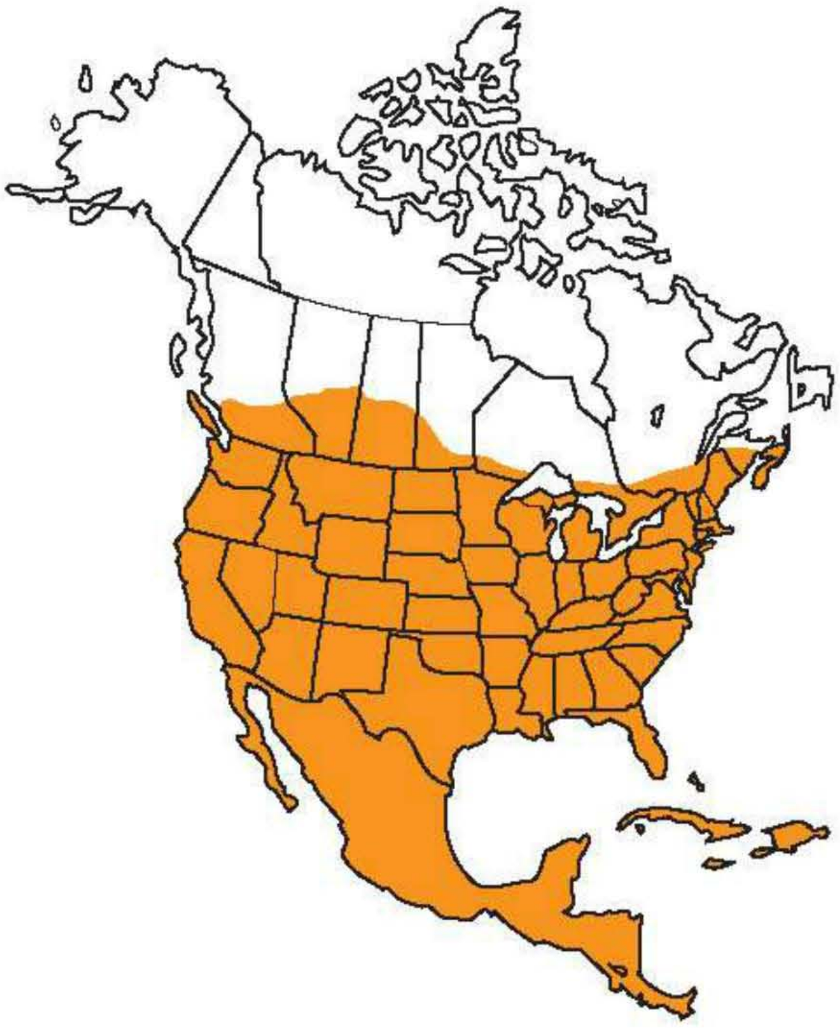

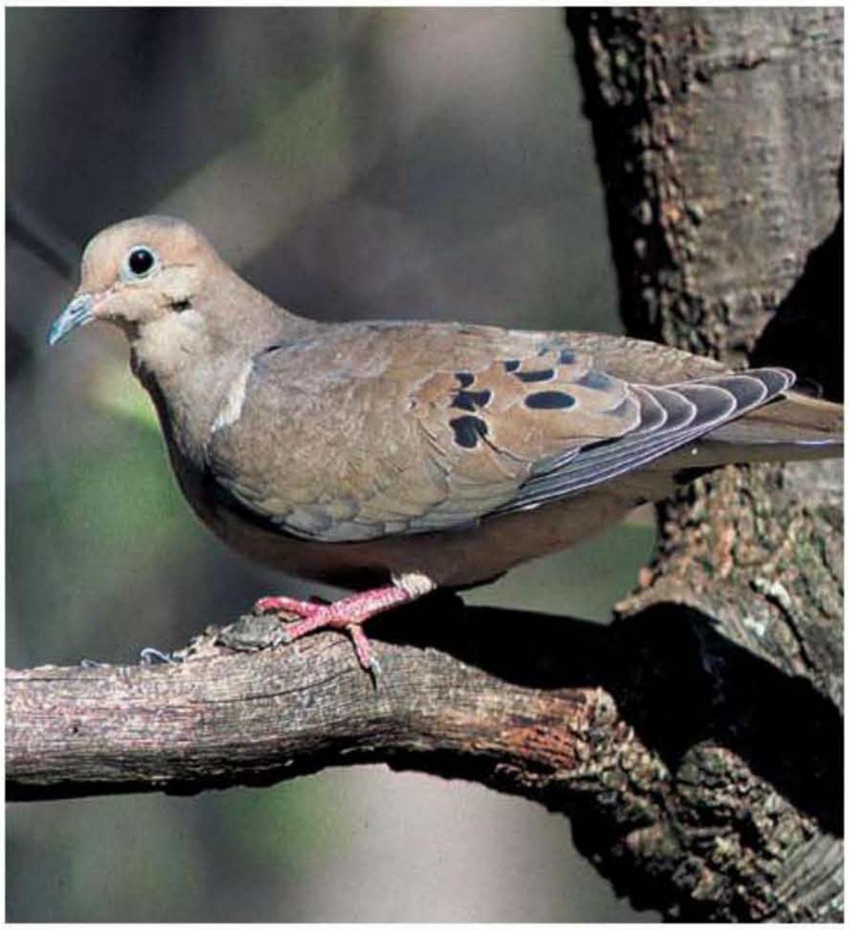

Mourning doves have a slate-blue back, a fawn-colored breast with a pinkish cast and a black spot behind the ear. They measure 11 to 13 inches (28 to 33 cm) from head to tail and weigh about 3½ to 5 ounces (99 to 142 g).

Roosting trees are usually within 2 miles (3.2 km) of water. A favorite roosting spot is a sand bar in a river where the birds find dead trees close to water.

Hunting for Doves

Most dove hunters find a stand along a flight path. It may be a route between a roosting and feeding site or between a feeding and watering area. Conceal yourself in a brush patch or near a fence line or tree, then pass-shoot as the birds fly through.

Doves often fly near a dead tree, telephone pole or any object taller than the surrounding cover. They frequently fly through a gap in a tree line to reach a feeding area or water hole. Often they skirt the end of a point of cover extending into a feeding field. If stands within gun range of these spots do not produce, watch where the birds are flying and change your location accordingly.

By stationing several hunters around the edge, you can keep the birds moving and improve shooting for everyone.

Ideal habitat includes an old dead tree and muddy water with a bank free of brush. Scout for set-ups where a fencepost or a dead tree will break up your outline.

The best way to put birds in the game bag is to improve your shooting. Train to focus on one part of one bird. Key on a dove's beak for a passing shot or the tail feathers as it goes straight away. The tighter the focus, the better your chances are of bringing birds home for dinner.

That's difficult to remember in the heat of the action, but it is easily practiced on any bird you see. In time it becomes automatic.

There are more techniques to help the bird hunter fill the game bag. The first is timing. Take it slow. With an unloaded gun, practice disengaging the safety and shouldering it. The cheekpiece should come against your cheek, with the butt anchored against your shoulder. The barrel should be in the same plane as your master eye, front bead on (or ahead of) the target.

Picture the shot leaving the barrel in a string, similar to water in a hose. Imagine swinging a spray nozzle across the sky. When you swing your barrel, following through as you squeeze the trigger, your shot will "wash" the target in the same way.

A stool can improve your shooting. Lunging from a crouched position to shoot is awkward. Rising easily from a chair promotes good stance and shooting position.

Another essential is the earplug. There may be more shooting on one dove hunt than in all the rest of your season's bird hunts combined. Earplugs keep you from flinching, but more important, they protect the auditory organs. You want to keep your hearing sharp so, on opening day of dove season twenty years from now, you can still hear the whistle of the wind in their wings.

Camouflage clothing is ideal, but any drab outerwear will do. Many hunters use semi-automatic or pump shotguns. A variable or screw-in choke enables you to vary the shot pattern, depending on the range of the birds. An improved cylinder or modified choke with No. 7½ or 8 shot usually works best.

WHERE TO FIND DOVES

Water holes used by doves often have muddy water and bare ground along the edge. The best water holes have trees with dead branches nearby.

Idle fields attract doves because they provide an ample supply of weed seeds. Doves prefer to pick seeds from bare ground. They seldom feed in dense vegetation.

Gravel roads provide grit. Doves pick up sand and small gravel along roads in early morning and late afternoon. They also find grit along streams and in fields.

WOODCOCK

Woodcock have a chunky body with a mottled brownish back and sides and black bars on top of the head. They measure 10 to 12 inches (24.5 to 31 cm), weigh 6 to 8 ounces (170 to 227 g) and have a bill about 2½ inches (5 cm) long.

Eyes on the side of the head give the birds excellent lateral vision Some researchers believe that woodcock rely on an acute sense of hearing to find worms in the ground. The birds can fly as fast as 30 mph (48 km/h). When flushed, they usually land within 100 yards (91 m).

Most woodcock breed in the northern states and Canada. They start their southerly migration in early to mid-October, leaving en masse once the ground freezes or after a heavy snow. Woodcock stop off in the same resting areas each year.

Young forests with trees 10 to 20 feet (3 to 6 m) tall make the best woodcock habitat. The soil should be damp with little grassy cover. Heavy ground cover makes it difficult for the birds to find earthworms, their favorite food.

Woodcock feed mainly around dawn and dusk. Besides earthworms, the birds eat insect larvae, seeds, berries and green leaves. During the day, they rest on the ground, feeding only occasionally. In cool weather, look for them on sunny hillsides or other sunlit areas. On a hot day, they sit in the shade, often below evergreens.

The best woodcock hunting is during the migration period. Northern hunters may get some shooting at resident birds, but when the migration peaks, they may flush up to thirty birds per hour.

Hunters who use dogs flush a least twice as many birds as those who do not. Because woodcock blend so well with the leaves, downed birds can be difficult to find without a dog.

A short-barreled 20-gauge with an improved cylinder choke is an excellent gun for woodcock. Most hunters use No. 7½ or 8 shot, but some prefer No. 9.

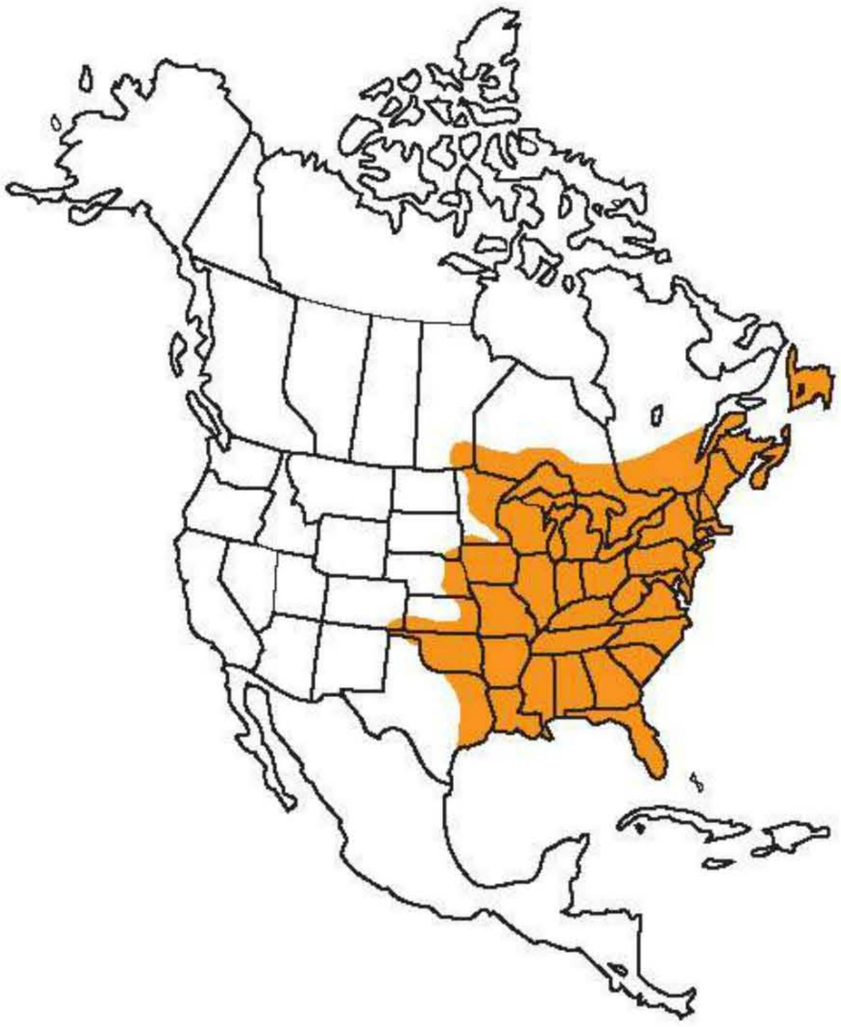

Many hunters who own pointing dogs consider woodcock or timberdoodle, the ideal quarry. Even with a dog holding a close point, a bird will hold tight, confident in its near-perfect camouflage.

Chapter 7

HUNTING WATERFOWL

Waterfowl are ideally suited for a life spent in and around water. Webbed feet give them extraordinary swimming ability and their long necks and broad, flattened bills allow them to feed on aquatic plant and animal life. Waterproof plumage and thick layers of insulating down keep these birds from losing body heat in cold water.

Geese are distinguished by their large size, with some approaching 15 pounds (6.75 kg). During the migration, geese may form enormous flocks numbering in the tens of thousands. Geese exhibit no coloration differences between the sexes. They rarely breed before their second year and may live for twenty-five years. Geese mate for life and maintain stronger family bonds than do ducks, with both parents caring for the young.

Ducks differ from geese in that sexes usually vary in coloration. Males have distinctive, colorful plumage, while females are camouflaged with mottled, drab browns. Pair bonds are temporary, with the drake deserting the female shortly after breeding. Most ducks are short lived, but a few have been known to live fifteen years.

Puddle ducks include some of the most widely hunted species. Puddle ducks are adept at walking and feeding on land. They can take flight almost instantly by jumping straight into the air.

Diving ducks have legs and feet positioned far back on their bodies. Rather awkward on land, they spend most of their time in the water.

Identification for species and sexes requires a great deal of practice, especially with ducks. But you can recognize the major types of waterfowl by their body size and wingbeat. Geese are larger than ducks and their wingbeat is slower. Puddle ducks have a faster wingbeat, but not as fast as diving ducks.

WATERFOWL EQUIPMENT

Waterfowling requires more equipment than most other types of hunting. In many cases, you will need decoys, camouflage clothing, waders, binoculars, calls and a duck boat.

Regulations require hunters to use nontoxic shot. Most hunters prefer a 12-gauge with a modified choke for shooting ducks and geese, increasing the size of shot when hunting larger waterfowl such as geese.

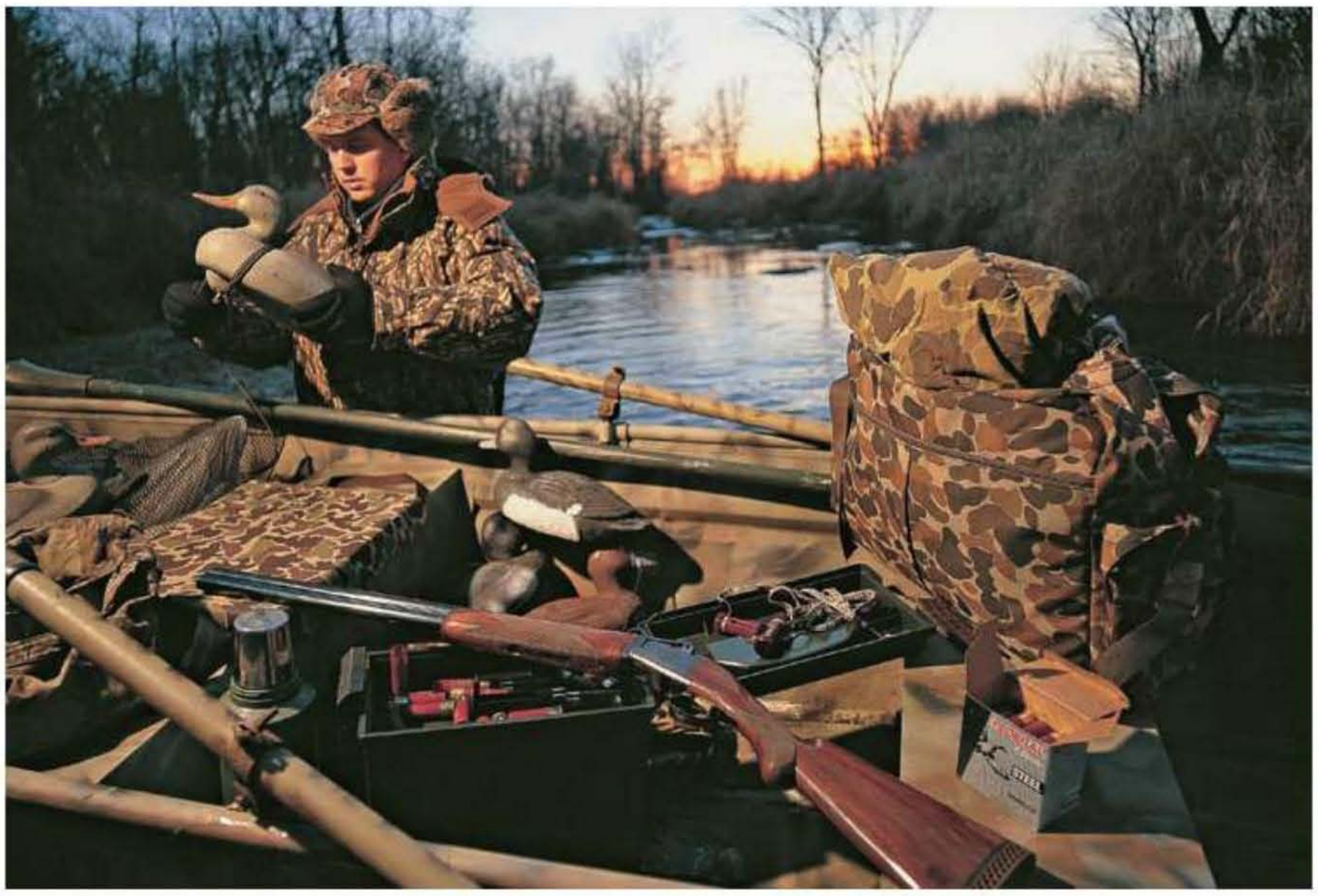

Waterfowl hunters must think total concealment when setting up a decoy spread. Camouflage the hunter and the boat, and then utilize the proper call to "sell" the spread to incoming birds.

BASIC DUCK AND GOOSE CALLS

Duck Calls. *There are two basic types of mallard calls: standard, or Reelfoot, style and Arkansas style. Reelfoot calls have a metal reed that is curved upward at the tip. Arkansas calls have a straight plastic reed. Reelfoot calls are generally harder to blow and have a narrower range of tones, but are much louder. Consult an experienced local duck hunter to determine what kind of call works best in your area. The sound of a particular call may be more effective in one region than in another.*

Calls for other species include those for wood ducks, diver ducks and pintail-widgeon-teal whistle calls.

Goose Calls. *The basic types of goose calls include the standard call and the flute call. Standard goose calls are easy to blow. They create two tones, one low and one high pitch tone, that imitate the basic sound made by Canada geese. Low grunting feeder sounds can also be imitated using standard calls. Flute goose calls are more difficult to master, but can imitate the many different communication sounds of Canada geese. Callers use both hands to operate flute calls; one to hold the call while the other is cupped over the end of the call to regulate air flow.*

Calls for other species include those for white-fronted geese and snow geese.

EQUIPMENT FOR DECOY HUNTING

Full-body floating decoys may be made of wood, cork, foam or molded plastic. A weighted or water-filled keel makes them float upright and a string and anchor keep them from drifting away. They come in standard and magnum sizes.

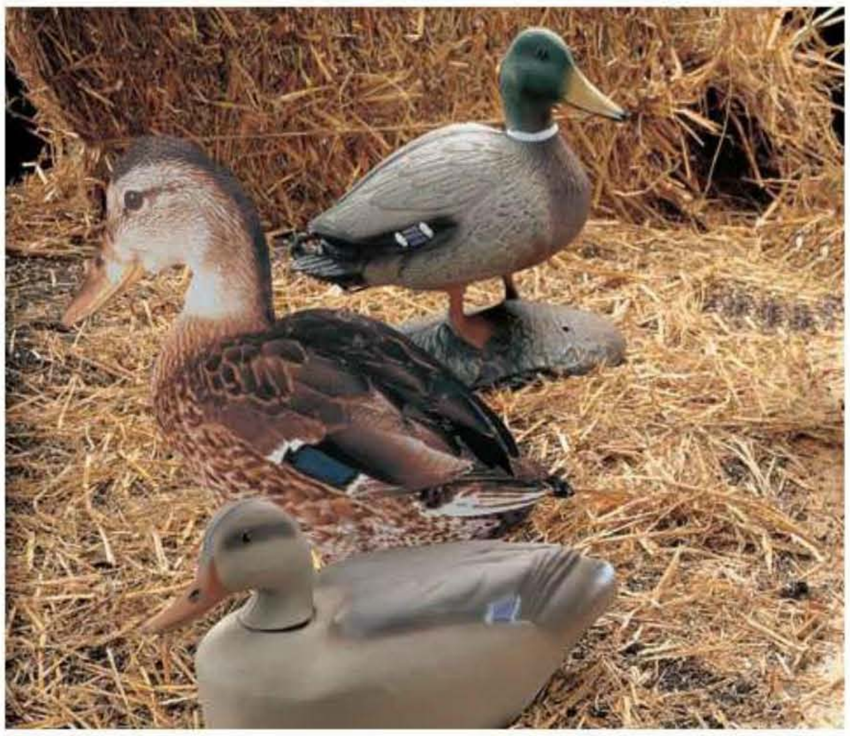

Field decoys: Shell decoys (bottom) are the most common among hunters because they are three-dimensional and can still be stacked for easy storage. Full-body decoys (top) add realism to your spread, but are probably the most bulky of all the styles. Silhouette decoys (middle) are very lightweight and easy to stack. Manufacturers now cut them in a wide variety of feeding, standing and resting positions, adding to the believability of the spread.

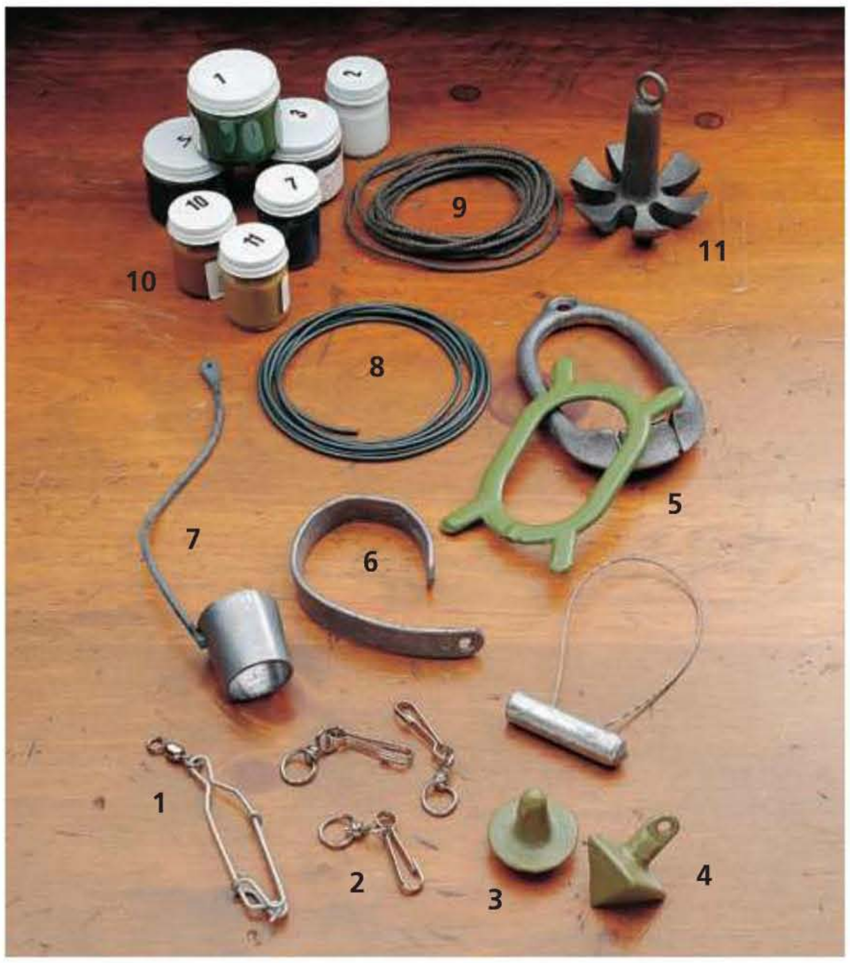

Decoy accessories include: (1) longline clip, for attaching decoy cords to longlines; (2) snap-swivels, which attach to decoy weights to prevent cord twist; decoy weights, such as (3) mushroom anchor, (4) pyramid anchor, (5) ring anchors, (6) strap anchor and (7) scoop anchor, which slips over the decoy's bill; (8) tangle-free decoy cord; (9) braided-nylon decoy cord; (10) decoy touch-up paint; and (11) grapple anchor for longlines.

Flagging devices attract the attention of waterfowl at a distance. Use a black or white flag for geese; a gray or black flag for ducks.

PUDDLE DUCKS

As their name suggests, these ducks commonly frequent small, shallow bodies of water, but they may also be found on big water. Also called dabblers, they feed on or just below the surface, mainly on aquatic vegetation. They may skim food off the water or they may tip up, submerging their upper body to feed while leaving their feet and rump pointing up. In fall, some puddle ducks also feed in grain fields. Their predominantly vegetable diet explains why they are considered better eating than most other ducks. Large wings give puddle ducks good maneuverability. They often circle a potential landing site several times, inspecting it closely before setting down. There are a total of twelve huntable puddle duck species in North America.

Puddle ducks are able to lift from the water almost instantly. They may be hunted by jump-shooting, pass-shooting, or by using decoys to draw them in.

HOW TO RECOGNIZE PUDDLE DUCKS

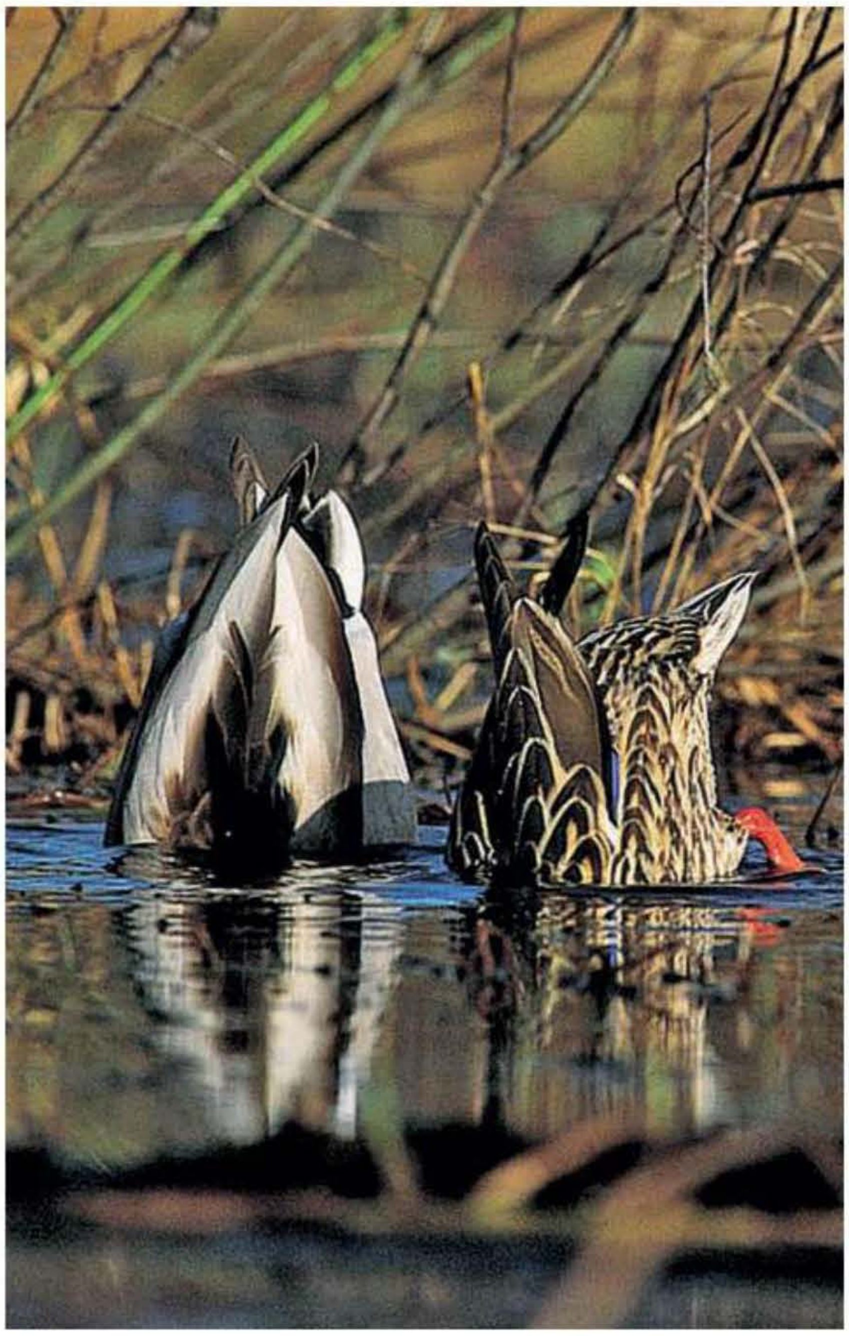

Tipping up or dabbling, is the usual puddle duck feeding method.

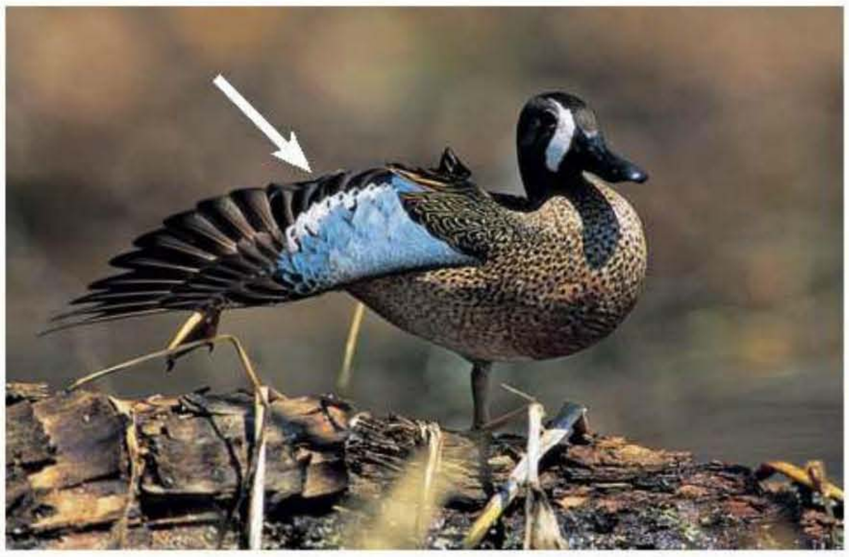

Colored wing patches are present on most puddlers. Most have a patch, called a speculum, on the wing's trailing edge. The speculum (arrow) is usually iridescent.

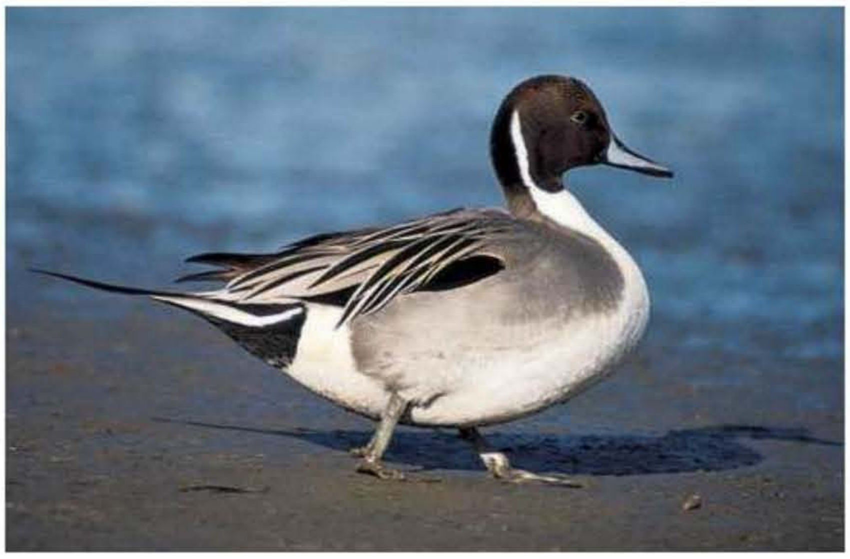

Legs positioned near the center of the body give a puddler good balance, so it can easily walk and feed on land.

Many waterfowl hunters prefer to use decoy spreads. Camouflage is important. Build a temporary blind from natural materials or bring in a portable blind. Permanent blinds can be built on private leases.

Puddle-duck Hunting

The widespread distribution of puddle ducks, combined with their excellent table quality, accounts for their tremendous popularity among waterfowl hunters.

As a rule, puddle ducks prefer smaller water than diving ducks; even the shallowest slough or tiniest creek may offer top-rate hunting. But puddlers are also found on some of the continent's largest waters, including the Great Lakes and Utah's Great Salt Lake. While divers are comfortable riding out big waves in open water, puddlers are usually found in calm water. In windy weather, they seek shelter along a lee shore.

You don't need a lot of expensive equipment to hunt puddle ducks. Jump-shooting and pass-shooting require nothing more than a shotgun and a few shells and you can decoy the birds into a small body of water with only a half-dozen blocks. Hunting on big water, however, is much more involved and requires considerably more equipment, including a good-sized boat and a minimum of several dozen decoys.

Mallards are the most numerous of the puddle ducks and most other puddler species feel comfortable in their company. Teal, for instance, are commonly seen flying with mallards and wood ducks often loaf in the same area as mallards. As a result, you can use mallard decoys and calls to attract most any kind of puddle duck.

In setting decoys for puddle ducks, remember that puddlers rest in looser groups than divers, so your decoys can be spread out much more. Some hunters leave as much as 10 feet (3 m) between individual blocks. Diving-duck hunters seldom leave more than 6 feet (1.8 m).

Calling is of utmost importance in puddle-duck hunting. The birds, by nature, are quite vocal and they look for reassurance from ducks on the water before making the decision to land. If you are not a competent caller, however, it's better not to call; you'll scare away more birds than you'll attract.

Puddlers are generally considered to be smarter than divers, but this "intelligence" may, in part, be a reflection of their migration and feeding habits. Because they migrate earlier than divers, they're exposed to a longer period of hunting pressure, so they're more likely to recognize the difference between decoys and the real thing.

PUDDLE-DUCK HUNTING TIPS

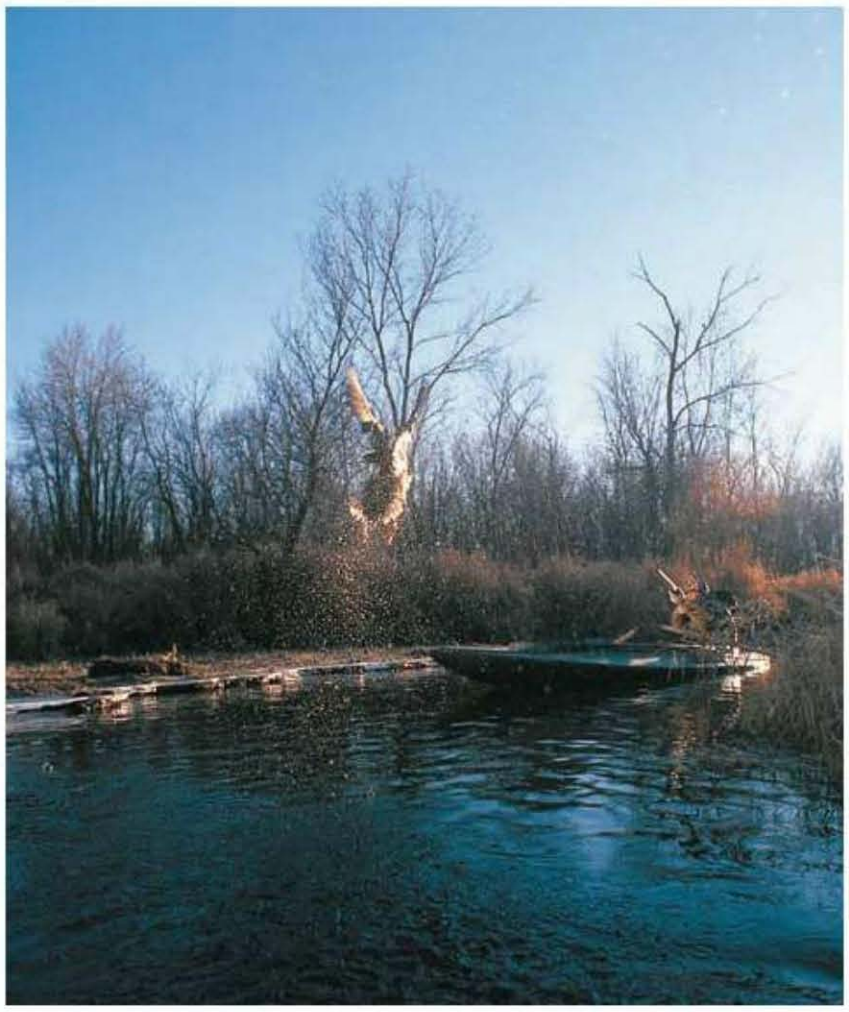

Float a small stream or river using a poke boat or a canoe. Hug the inside turns as long as possible. Ducks usually rest below the points and they won't see you coming until the last minute.

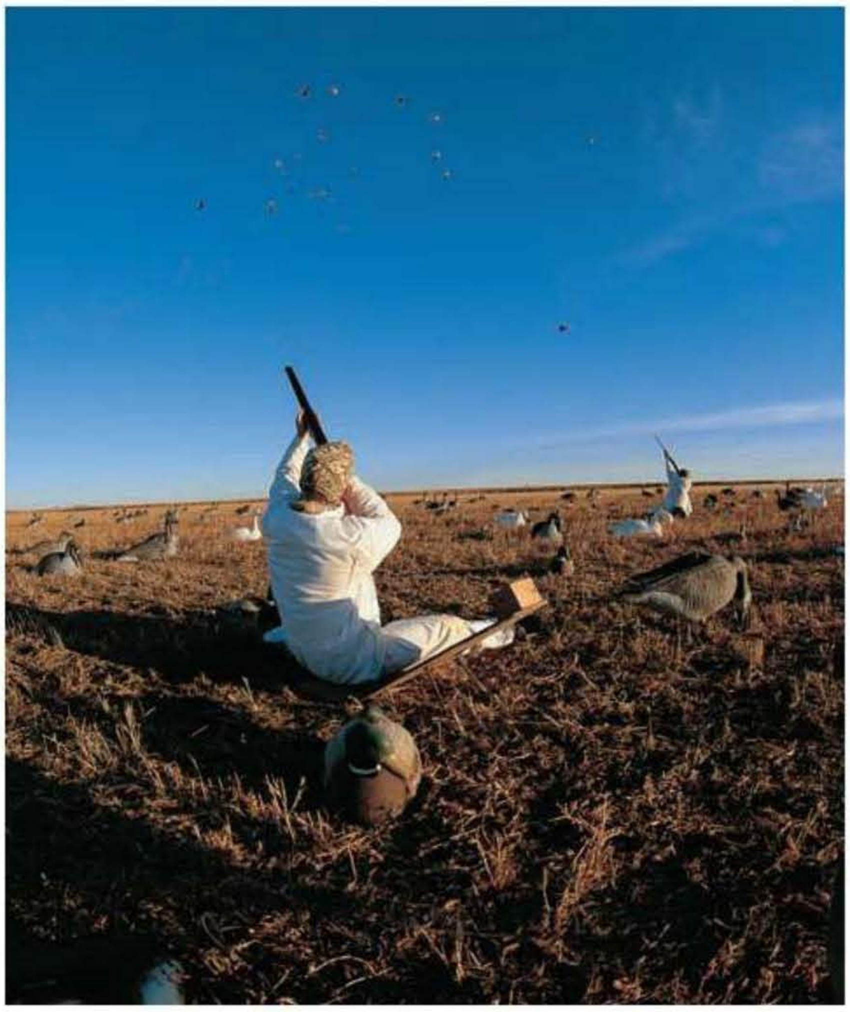

Use several goose decoys when hunting puddle ducks in the field. Ducks often feed with geese and the larger decoys add to the visibility of your spread.

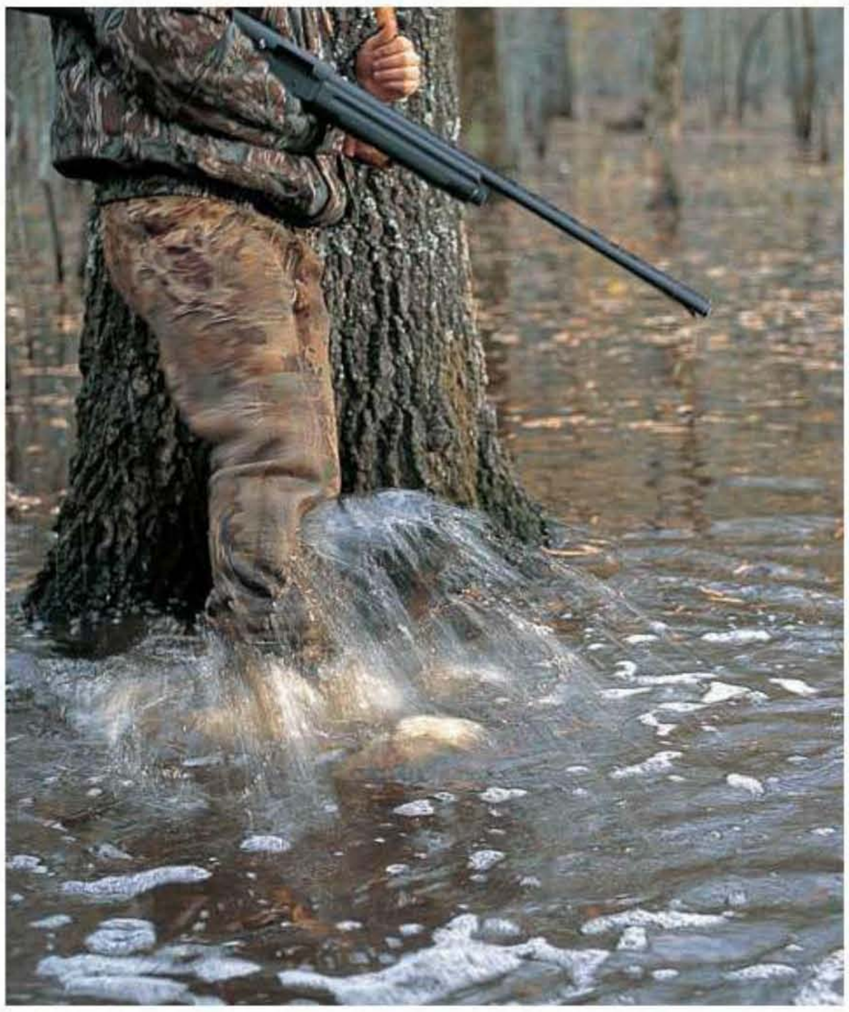

Kick water with your boot when hunting flooded timber. The ripples indicate to circling ducks that others are dabbling in the water.

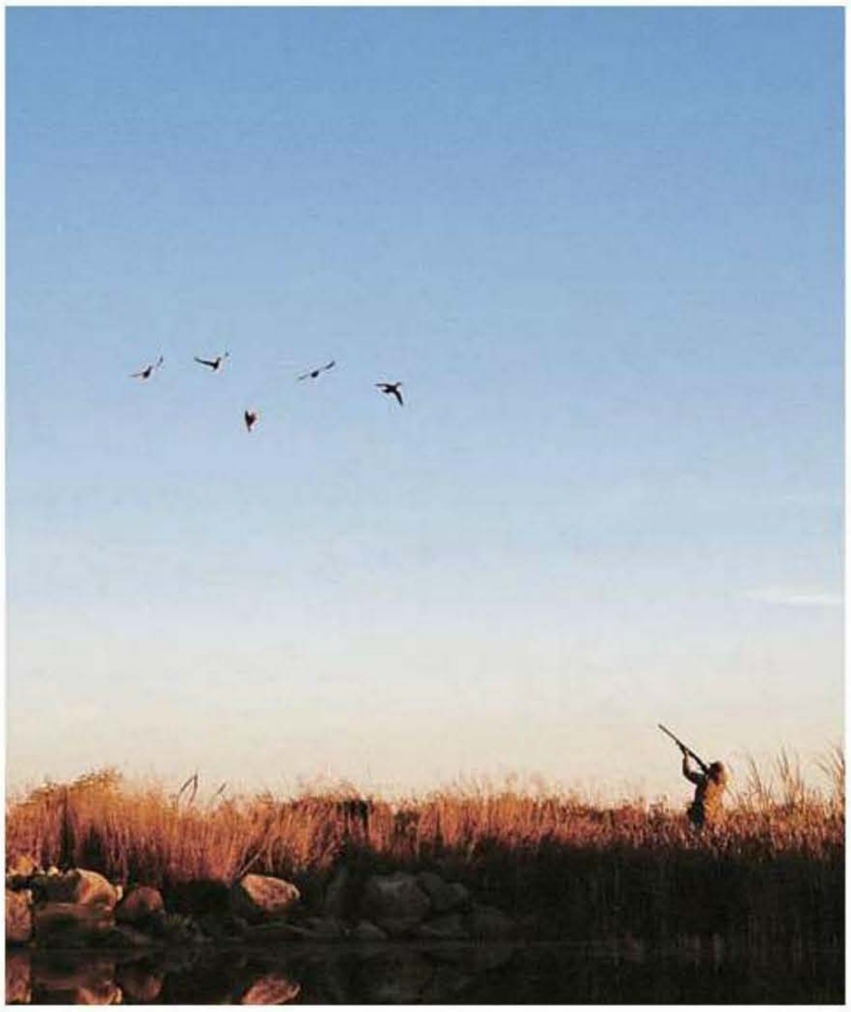

Pass-shoot from a strip of land between two lakes or between a lake and a grain field. Ducks choose the route that crosses the least land.

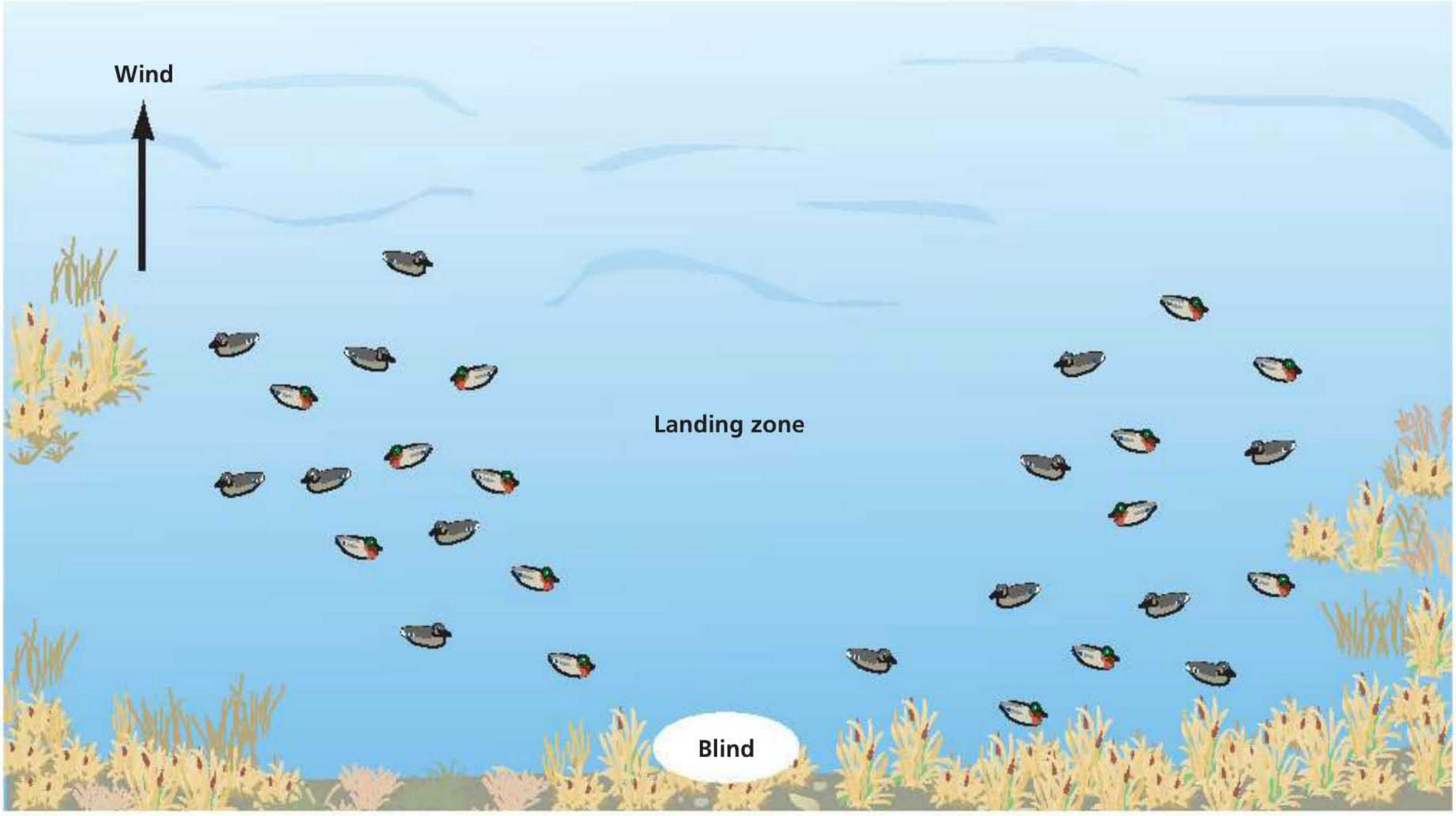

With the wind at hunters' backs, decoys should be set in two pods with a good-sized landing pocket between the pods. The outermost decoys should be no more than 30 yards (27 m) from the blind. The blind is positioned so the hunters are facing the pocket. Without a pocket, ducks may land beyond the decoys, out of shooting range.

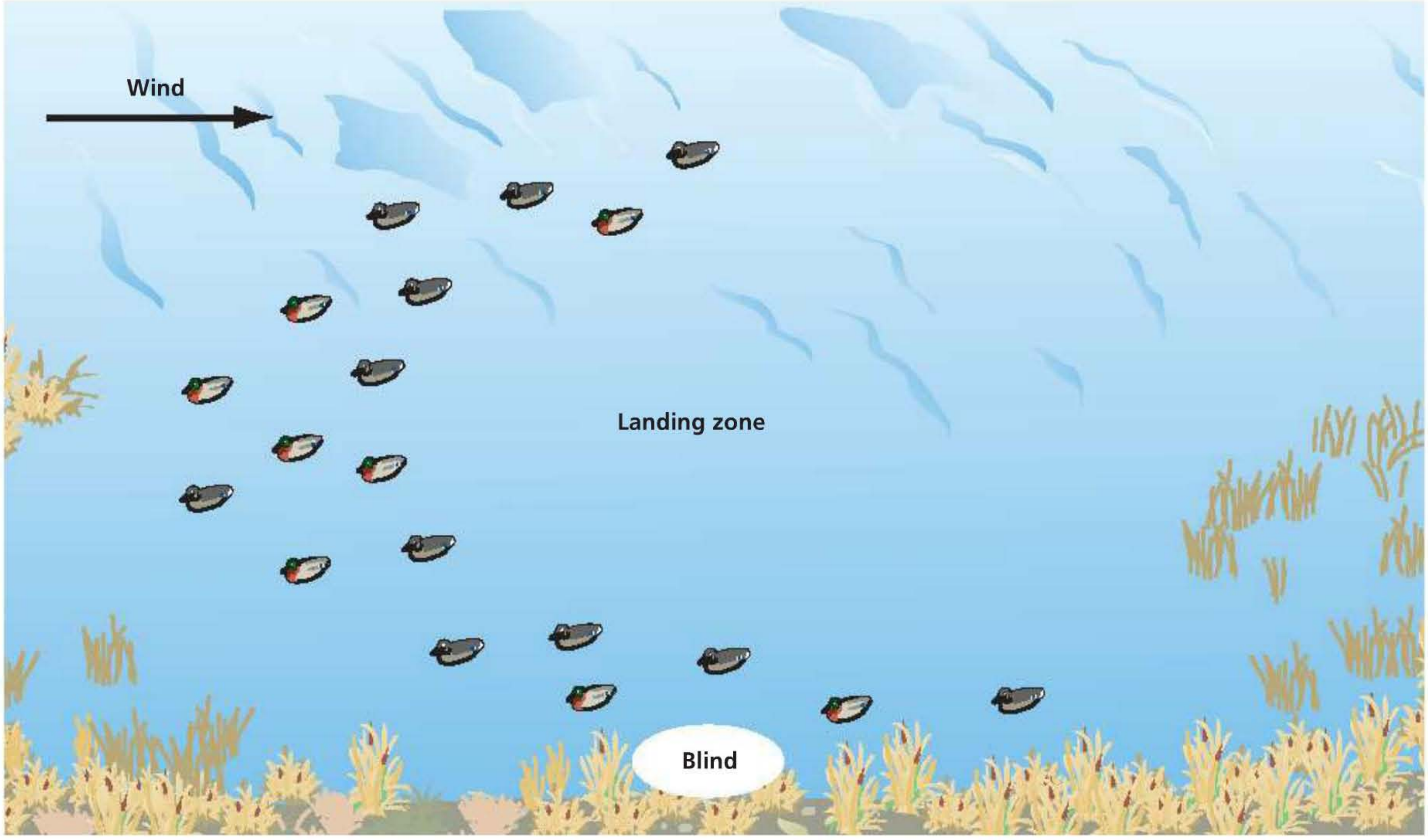

With the wind from the left, arrange the decoys in a "C" formation to the left of the blind, as shown. Ducks do not like flying over decoys to land; this way, they do not have to. When the wind is from the right, reverse the entire decoy setup.

DIVING DUCKS

With legs positioned farther back on the body, these ducks are more adept at diving beneath the surface than are puddle ducks. Their feet are larger for their size, so they are better underwater swimmers. The leg position also makes it more difficult for them to walk on land, explaining why they seldom feed in agricultural fields. The diet of most divers consists mainly of invertebrates and fish, explaining their strong taste. But some, such as canvasbacks and redheads, feed heavily on wild celery and other aquatic vegetation and are considered better eating. There are nineteen huntable diving duck species in North America.

Sea ducks, a subcategory of diving duck, differ from ordinary divers in that they spend most of their life in coastal areas. They have remarkable diving ability, with some species descending to depths of more than 200 feet (61 m) to feed on mollusks, crustaceans and fish.

Mergansers, often called fish ducks, are also classified as a subcategory of diving duck. They feed even more heavily on fish than other divers and their serrated bill is ideal for catching and holding small fish until they can be swallowed whole. Their crested head easily distinguishes them from other diving ducks.

Diving ducks plunge well beneath the surface to feed.

HOW TO RECOGNIZE DIVING DUCKS

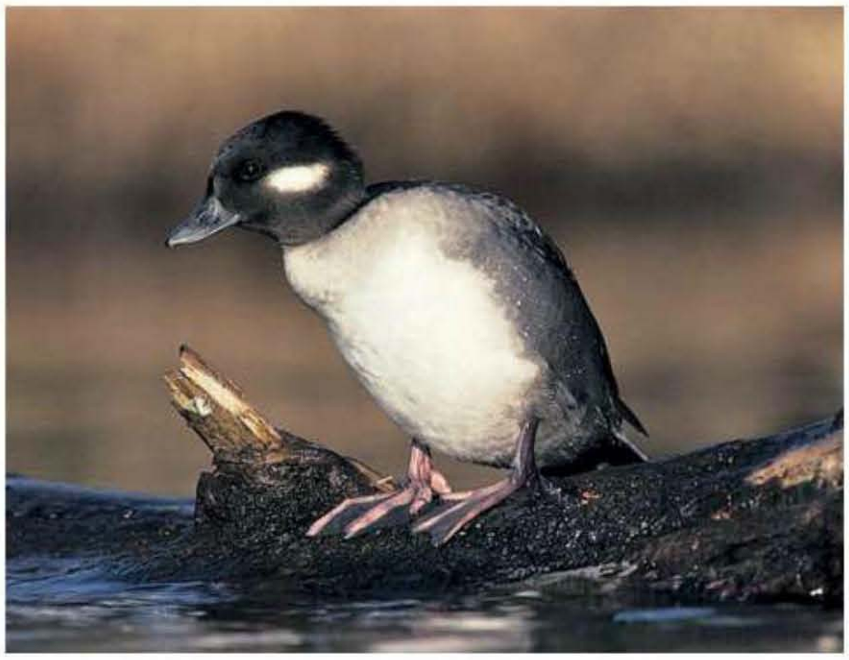

Large feet and legs positioned far back on the body account for the diving ability of these ducks. White to dark gray wing patches are found on most diving duck species.

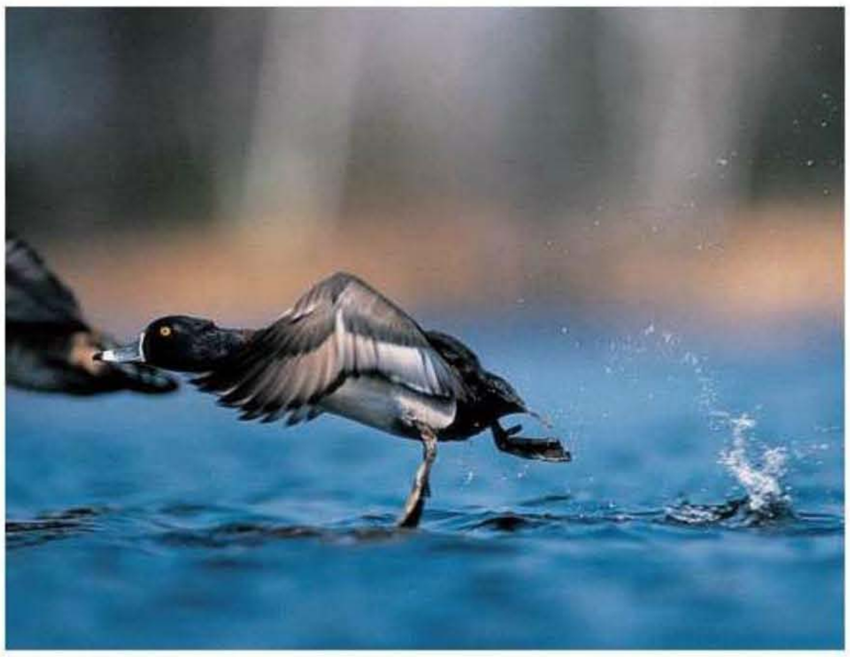

Flying low over the water in tight flocks is typical behavior among most diver species. In flight, divers can be distinguished from puddlers by their shorter, faster wingbeat. Running on the water helps divers gain enough speed for take-off. Their small wings provide less lift than those of puddlers.

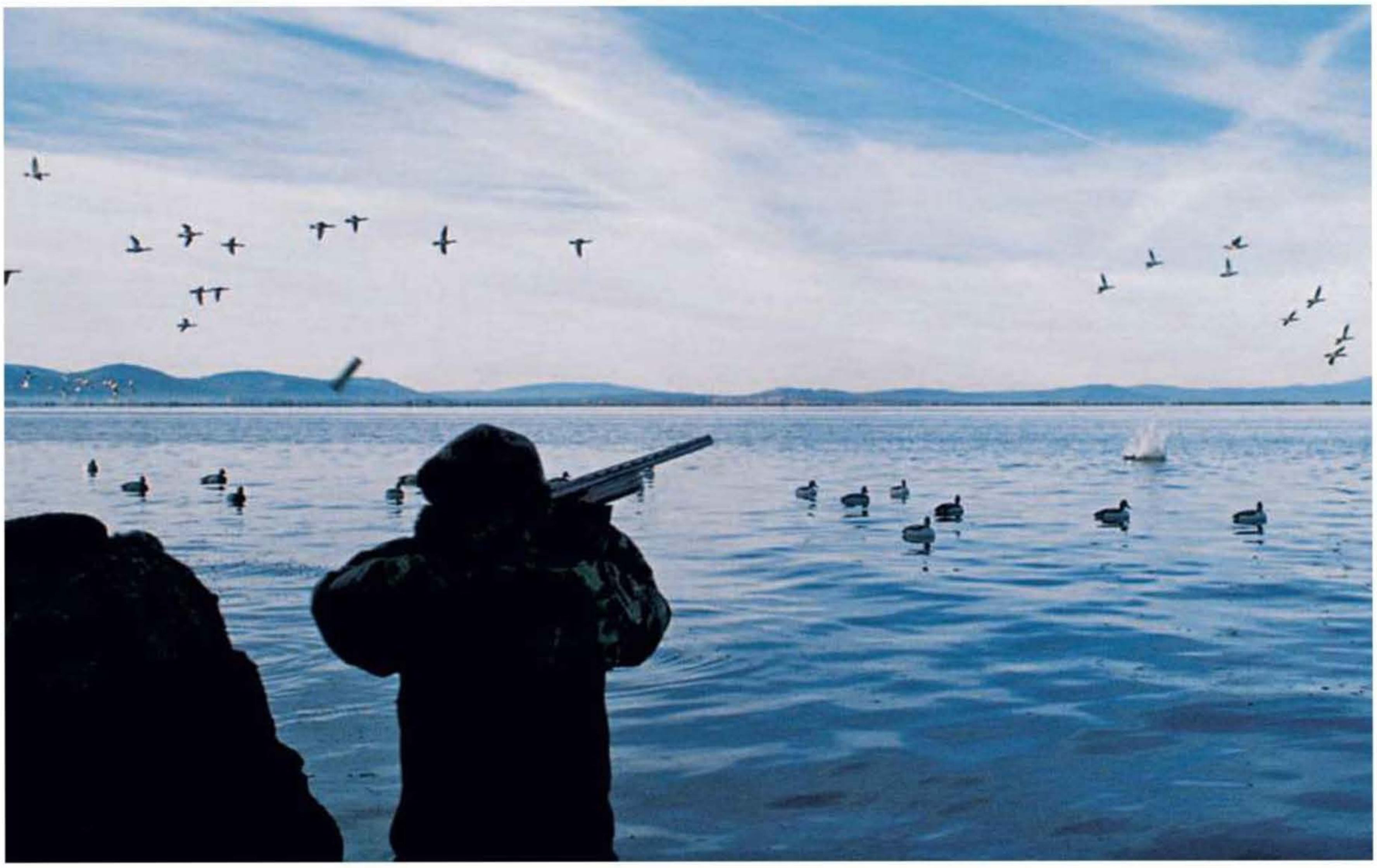

The pursuit of diving ducks takes hunters to broad rivers, estuaries, and in and around large saltwater bays. Strong fliers, they are one of waterfowling's greatest challenges.

Diving-duck Hunting

If you're looking for a wingshooting challenge, diving ducks are the perfect quarry. When a flock of divers streaks along a decoy line, it's not uncommon for a hunter to shoot at the lead bird and drop the third or fourth one back.

Due to the divers' preference for big water, hunters need more and different equipment than they ordinarily do for puddle-duck hunting. Instead of a 12- or 14-foot (3.7 to 4.3 m), shallow-draft duck boat, for instance, you'll need a 16- or 18-foot (5 to 5.5 m), deep-hulled semi-V to negotiate water that can turn rough in a hurry. You'll also need more decoys than you would for puddle-duck hunting and, possibly, a floating blind.

The divers' habit of forming large open-water rafts, particularly in late season, adds to the hunting challenge. Some hunters use low-profile sneak or sculling boats to approach birds in open water without alarming them.

Because divers migrate much later in the season than puddle ducks, you'll be hunting in considerably colder weather. Warm, waterproof clothing, including insulated boots and gloves, is a requirement.

Crippled divers often attempt to escape by swimming away underwater or with their head at water level. You'll need a determined, hardy retriever, preferably a Labrador or Chesapeake, that is capable of making long retrieves in cold, choppy water.

In puddle-duck hunting, the idea is to lure birds into the decoys and take them when they're "putting on the brakes," meaning that their wings are cupped and they are about to land. In diver hunting, the birds are less likely to slow down, so you'll probably have to take them as they're winging along a line of decoys that leads to your blind.

Instead of shooting at birds that are barely moving, you're attempting to hit birds barreling along at speeds up to 75 mph (120 km/h). To be successful, you must learn the swing-through shooting technique and do a lot of practicing to determine the proper lead.

Calling is seldom necessary in diver hunting. The birds spot the decoys from a distance and their curiosity draws them in without any vocal enticement. Nevertheless, many hunters try calling anyway, using a diver call or rolling their tongue against a mallard call to mimic the characteristic diver "purr."

Divers differ from puddlers in that they do not hesitate to land in rough water. As a result, there is no need to set your decoys off a lee shore. It's more important to set them where the birds are feeding, regardless of wind direction.

Yet another difference: divers do most all of their feeding on the water, so the field-hunting techniques that work so well for puddle ducks are not effective for divers. Practically all diver hunting is done on the water or on passes between two bodies of water.

Diver hunting is not for everyone. Not only does it require a high tolerance for inclement weather, it can be a lot of work. But if you're a hardy soul who loves the whistle of wings, there's no more exciting shooting.

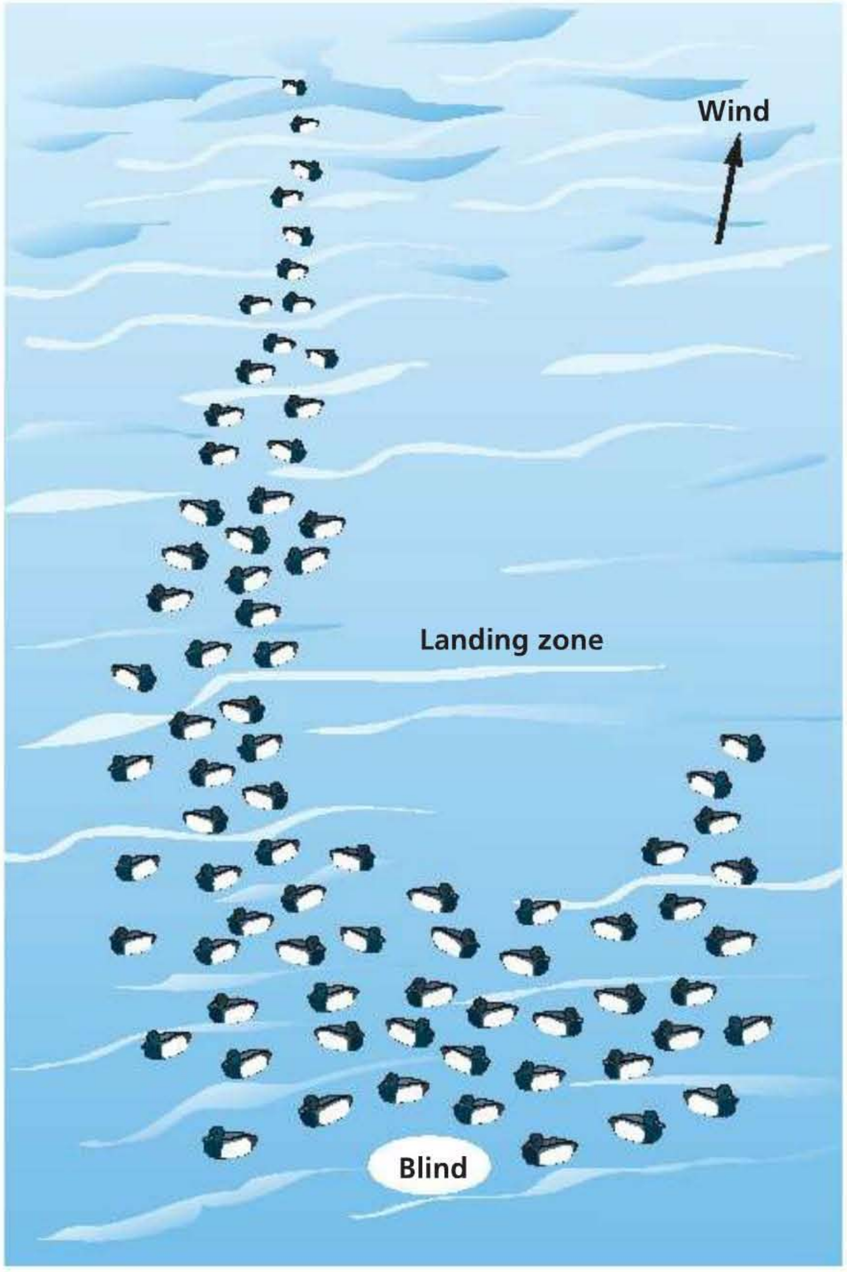

The J-hook. This setup works well off a point or in open water. The long tail extends well beyond shooting range. The largest clump of decoys is placed at the hook of the J, which forms a landing zone.

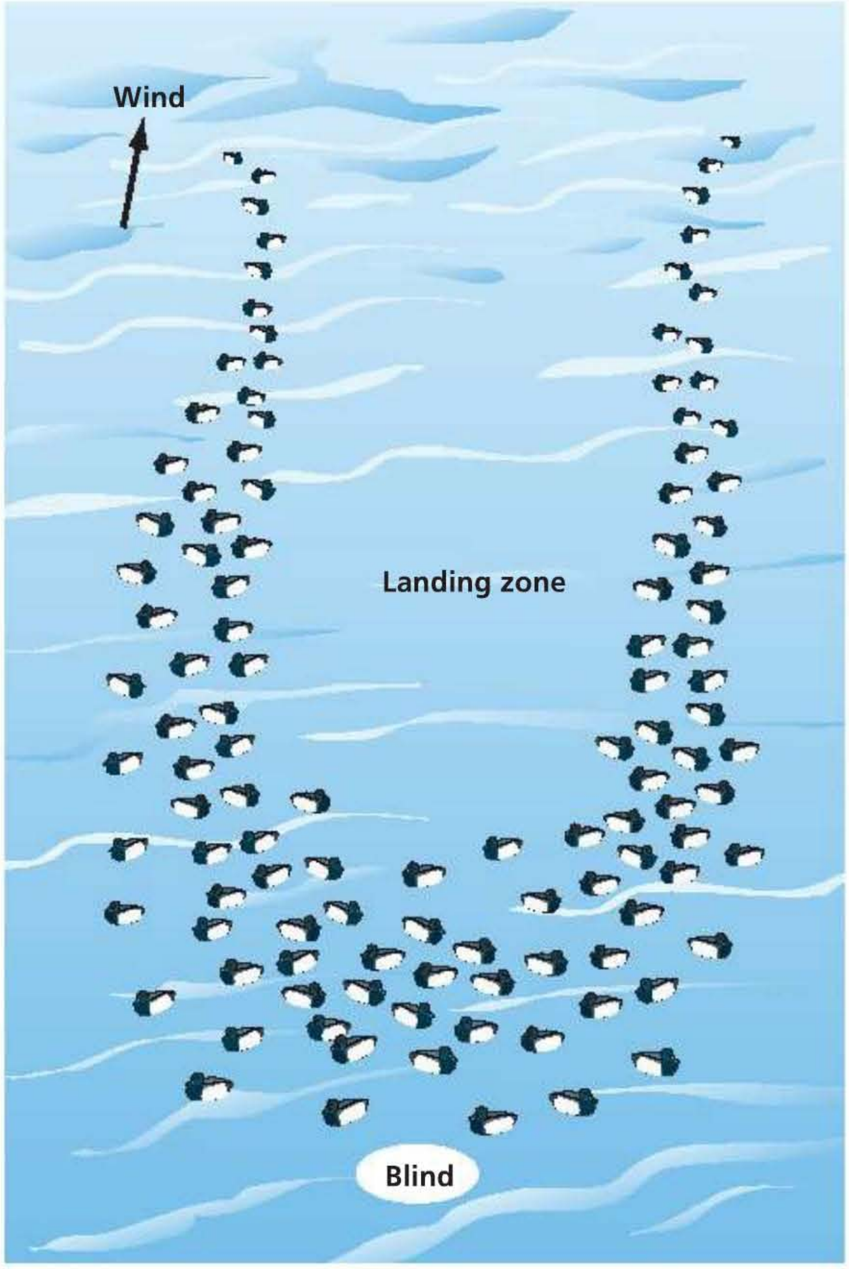

The V-formation. This setup is similar to the J-hook, but it has two tails extending out into the water, rather than one, giving the birds an additional option for approaching the blind.

DIVING-DUCK HUNTING TIPS

Use silhouette sleds, which can be neatly stacked so they take up less space than ordinary decoys. Attach the sleds to longlines; each should be rigged with about twelve sleds. Tie a full-body floating decoy to each end of the line to help keep the silhouettes afloat in rough water.

Make inexpensive diver decoys by painting black patches on white bleach jugs.

Low-profile boats allow you to get close enough for a good shot on diving ducks that have rafted up on the water.

Wave a black flag when you spot distant ducks. The motion resembles other ducks flying or landing. Stop flagging once the ducks start to head your way.

Look for openings in the trees or long points that extend out into the main body of water to pass-shoot divers.

GEESE

Geese often fly in a very distinct V-formation with a much slower wingbeat than ducks. Large, powerful wings enable geese to take flight very quickly, even though some Canada geese reach a weight of 15 pounds (6.75 kg). Once in flight, geese are easily recognized by their slower wingbeat when compared to the fast wingbeat of ducks.

Flocks consist of one or two family units, each numbering between four and six birds. As the migration progresses, many flocks join to form large concentrations, which may number in the thousands.

An exceptional homing instinct draws geese to the same waters every year. And they often return to the same feeding areas on consecutive days. If the food supply holds up, they come back to the exact spot in the field.

The fall diet of geese consists of waste grain, green plants such as clover, grasses and new shoots of barley, wheat and oats. Geese feed in open fields where they can see in all directions.

About three out of four bagged geese are juveniles, but some live twenty-five years. To survive that long, they must learn to be extremely cautious. Old birds scrutinize a landing site and veer off immediately if anything looks suspicious.

Dry land goose hunting is one of waterfowling's greatest thrills. The birds are big, with a wingspan of up to 71 inches (180 cm). The birds breed in Canada and the northern United States, then travel south in the fall in long Vs. They can be enticed with large decoy sets and calling.

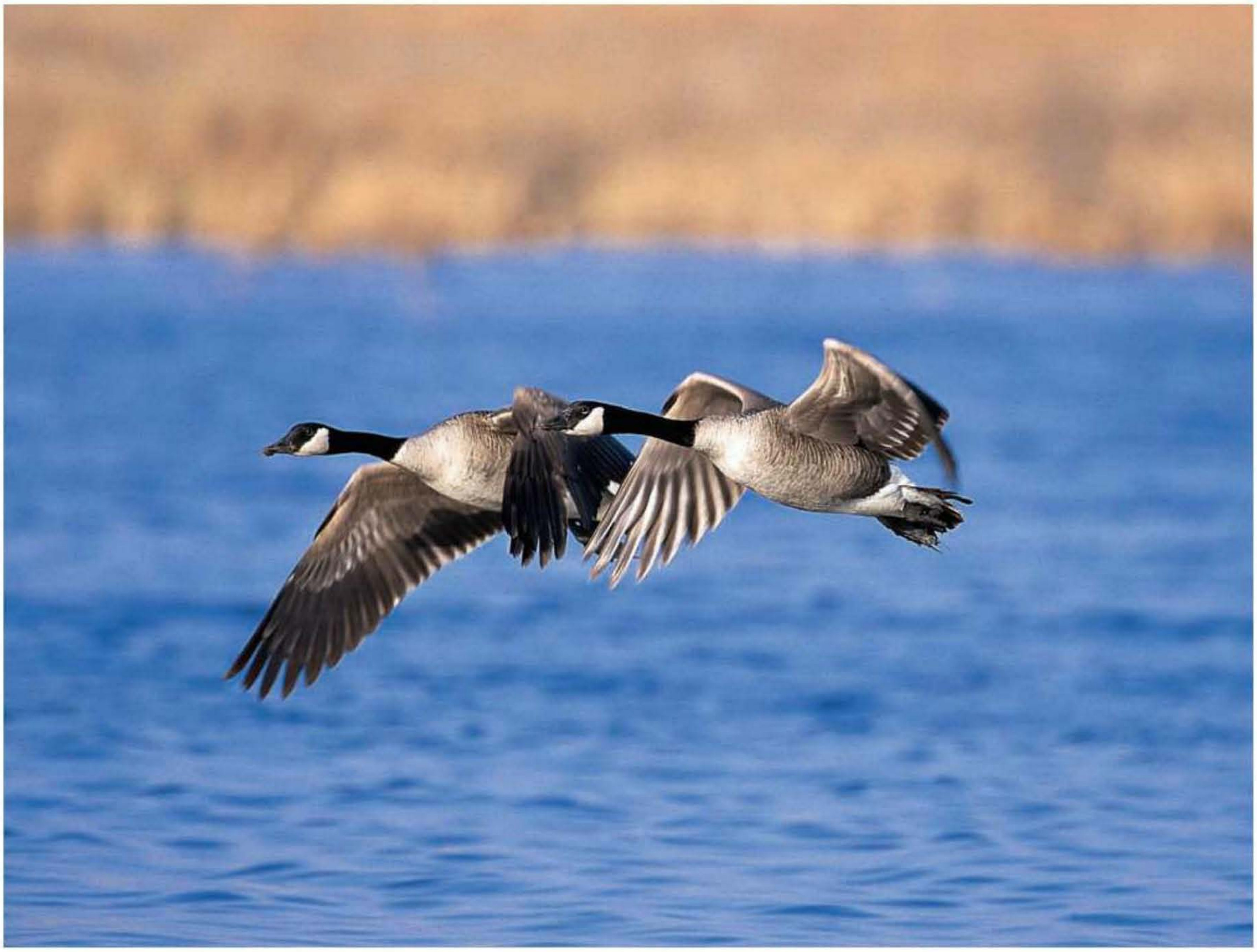

Canada geese have a black head with a white chin bar. The back is grayish or brownish and the belly is whitish. Biologists recognize more than 12 varieties of Canada geese, the largest being the giant Canada; the smallest, the cackling goose.

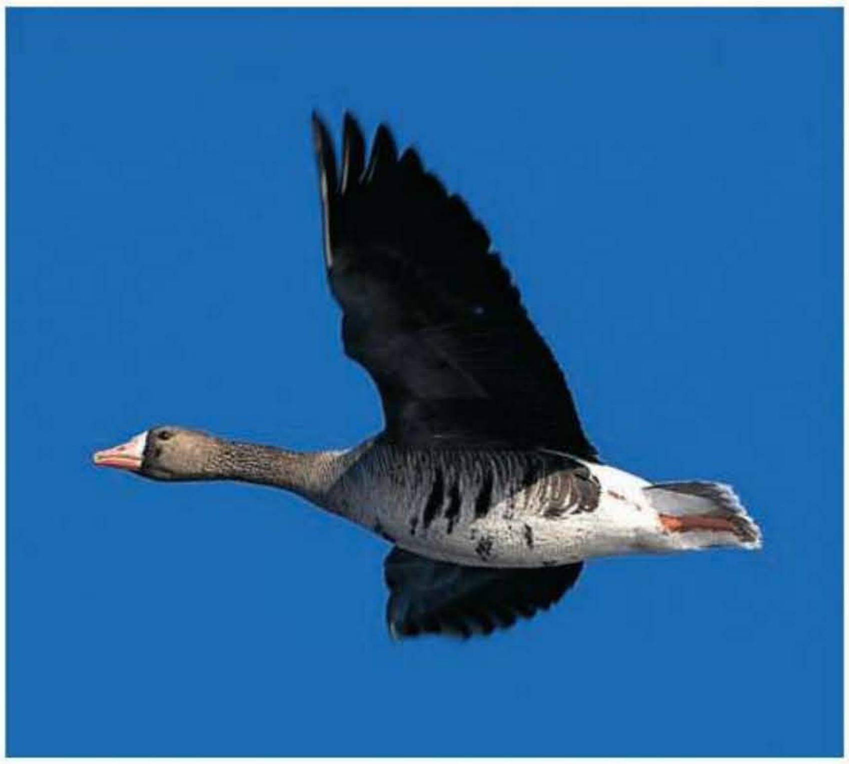

White-fronted geese, also called speckle bellies, have a brownish head, neck and back. The undersides are whitish with dark brown speckles. A distinctive white facial patch rings the base of the pinkish bill.

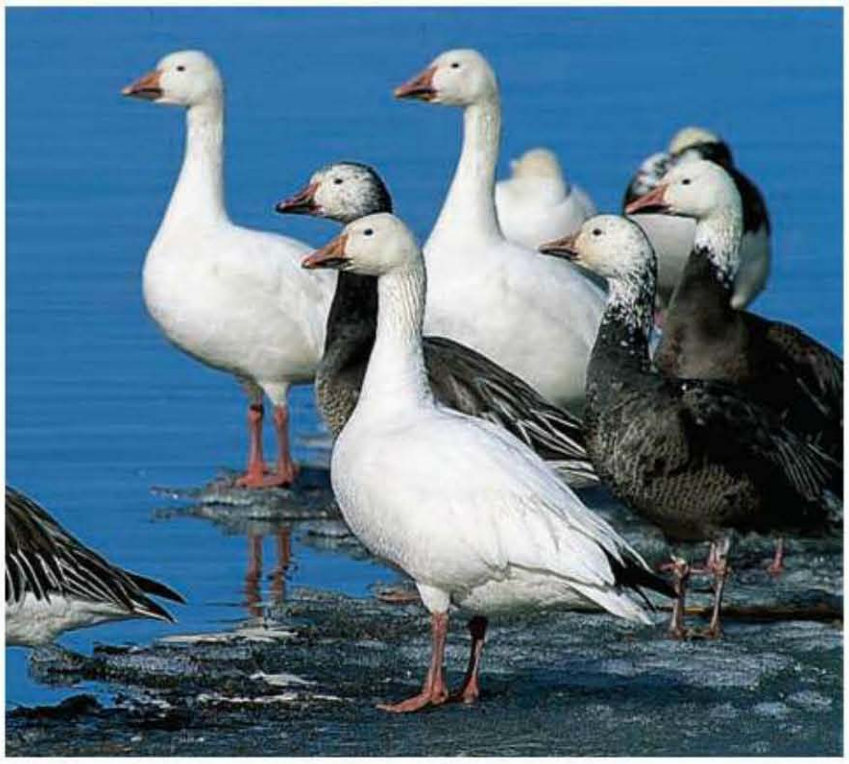

Snow geese are pure white with black wing tips. Two subspecies are recognized: the greater snow goose and the lesser snow goose. Lesser snow geese may sport either dark or white plumage.

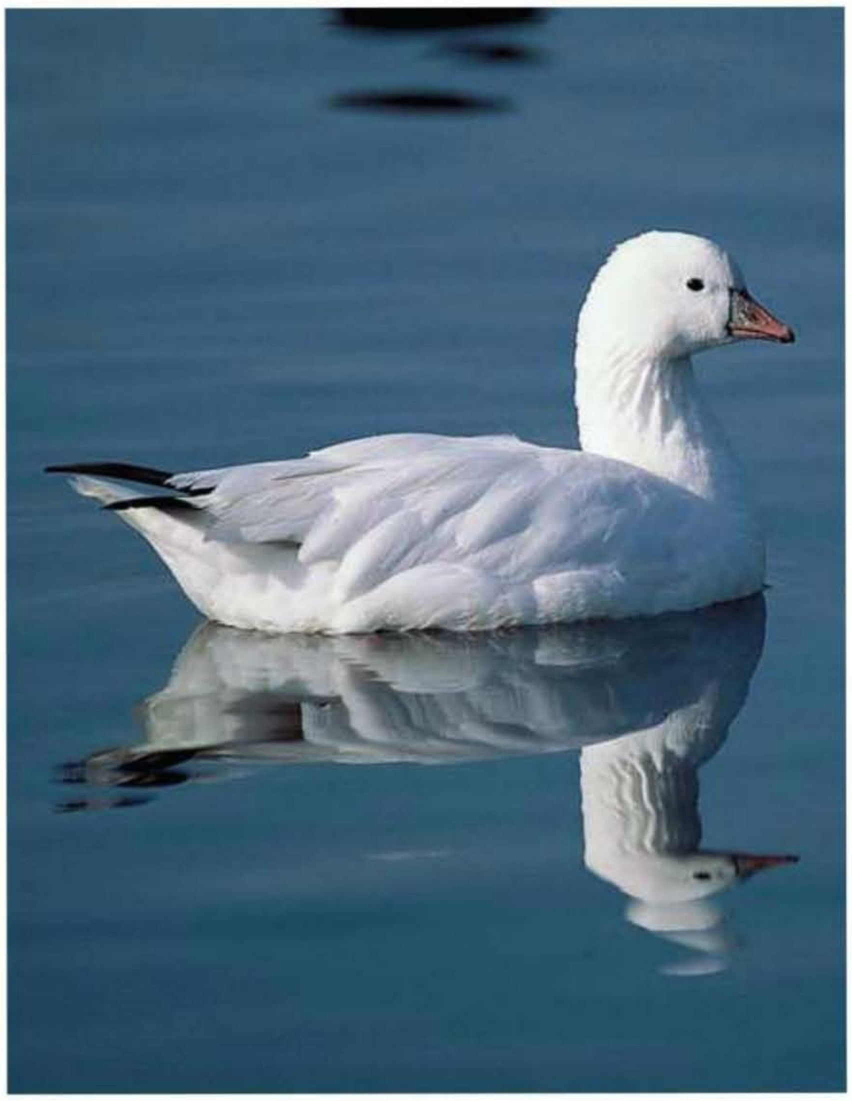

Ross geese resemble the lesser snow goose, but are smaller and have a stubbier bill. Ross geese often travel with snow geese and can easily be identified in flight by their faster wingbeat.

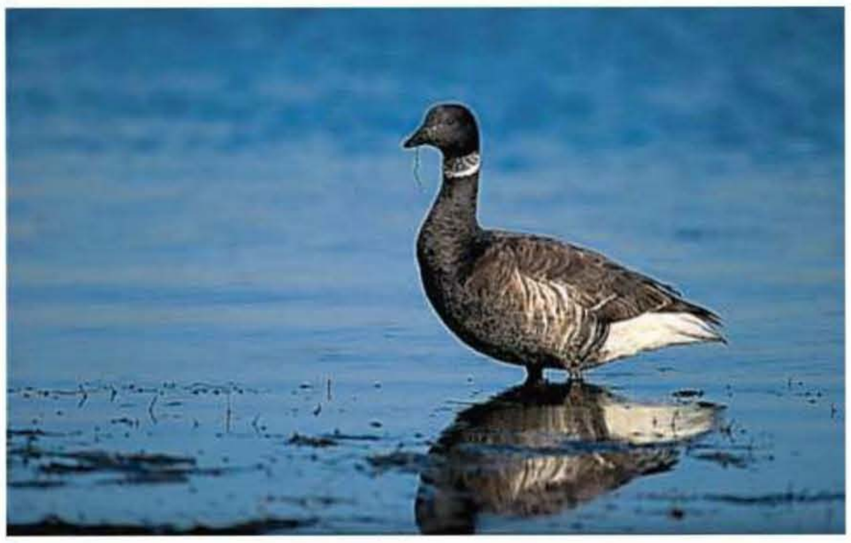

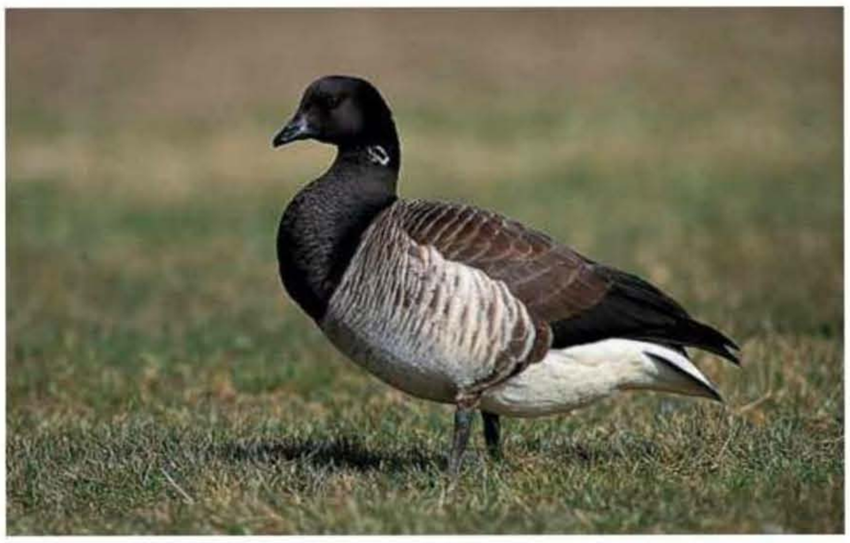

Brant geese, both Atlantic (bottom) and Pacific (top), can be recognized in flight by their rapid wingbeat. The head, neck and chest are black and the sides of the neck are marked with white streaks. The Pacific brant has a darker underside and more pronounced neck marks.

Hunting Geese

Those minutes before first light, the barley bending with the wind, the cool of the morning and the whispered anticipation of honking geese against a paling sky—this is what brings goose hunters to the field each year. That and the thrill of seeing the birds over the barrel of a gun, ten feet off the ground, their wings cupped to land.

Scouting, prior to the hunt, is essential. Note what time birds are seen in which fields and be there with your decoys the next day. Large flocks of strategically-placed plastic impostors, with good calling, will bring geese into shotgun range.

Hunting with decoys can be done both in feeding fields and on the water. When hunting in a field, you often need a large number of decoys to entice the birds.

When hunting for Canada or white-fronted geese in the early season, set your decoys in small clusters to imitate a small family group. Fewer decoys are often required early in the season because the geese haven't grouped up for the fall migration.

When hunting snow or Ross geese, some hunters use well over a thousand decoys in order to successfully decoy the most cautious waterfowl of North America. Combining large shell decoys and silhouette decoys, hunters mix in Texas Rag style decoys, white garbage bags and wind sock style decoys to increase the mass and visibility of the spread. Hunters rarely hunt snows on the water with decoys.

Brants can also be hunted with decoys, most commonly on the water. Setting two separate groups of decoys and leaving a place for the geese to land in the middle is the most common setup.

Most pass-shooting takes place in areas surrounding goose refuges or other waters with large concentrations of birds. Flight paths to and from a lake vary, depending on wind direction and the location of the feeding field being used. To select a good pass-shooting spot, spend a few hours watching the birds, taking note of the direction of the wind.

Hunters usually stalk geese after following or spotting a flock during their feeding period, which usually occurs twice daily, once in midmorning and again toward late afternoon.

Be prepared to use any available cover to sneak on a flock of geese. Sometimes they land in a field that is impossible to sneak. Better to find another flock where you are more likely to make a successful stalk.

When geese land in an open field, try team hunting. One hunter hides along an edge of the field, while another approaches from the opposite side. When the birds fly, they may pass over the hiding hunter.

Hunters accustomed to shooting ducks often have trouble hitting geese. Because the birds are so large, the tendency is to underestimate their distance and speed. Goose hunters use 12- or 10-gauge shotguns with modified or full chokes. Shot sizes range from No. 3 to T shot.

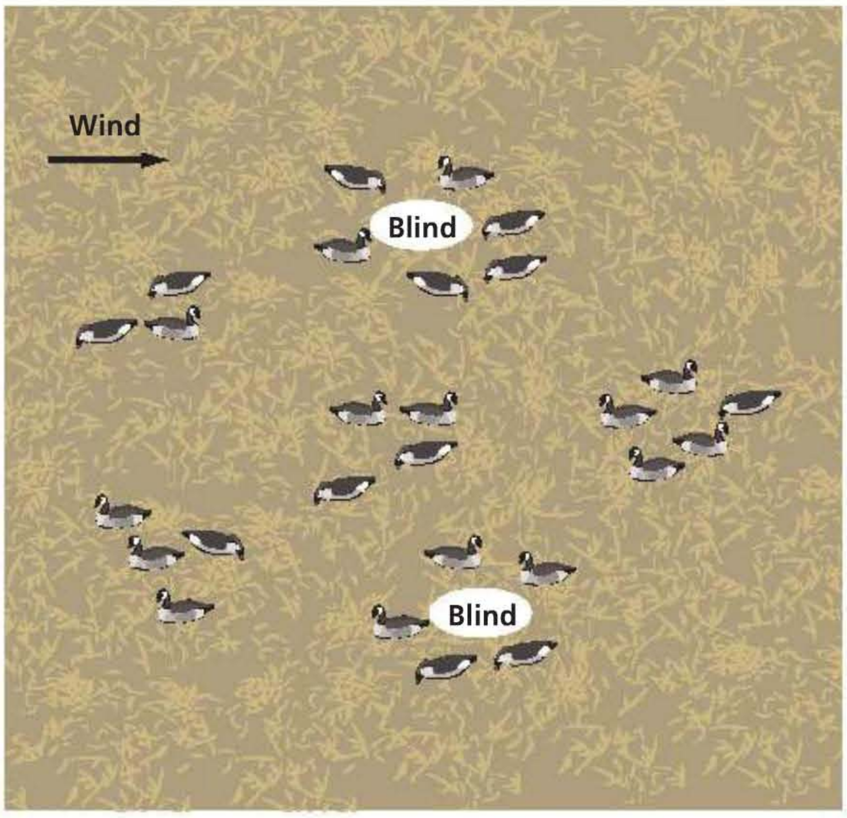

Family spread. Set Canada or white-fronted goose decoys in small family groups when hunting early in the season. Groups consist of four to ten decoys spaced about 5 to 10 yards (0.9 to 9 m) from each group. Set your blind(s) on the downwind side, shooting geese as they approach your spread.

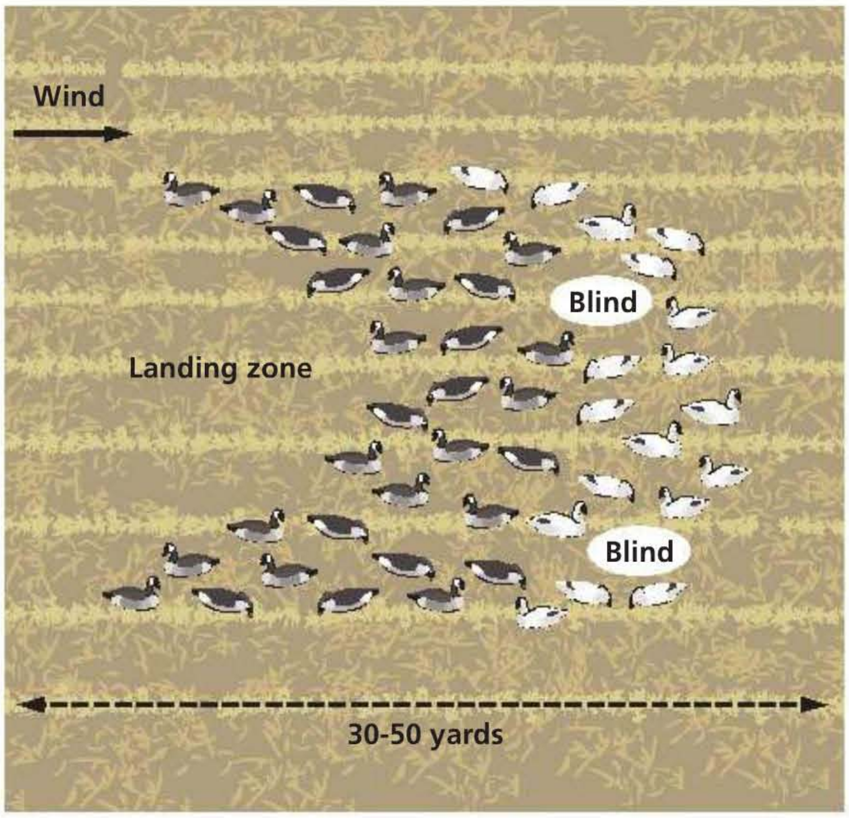

Combination field. Place snow goose decoys on the downwind side of the spread. On the upwind side, set your Canada decoys, leaving a large opening. Hide among the snow goose decoys or in a pit blind near the downwind side of the spread and shoot as the geese approach the landing zone from the downwind side.

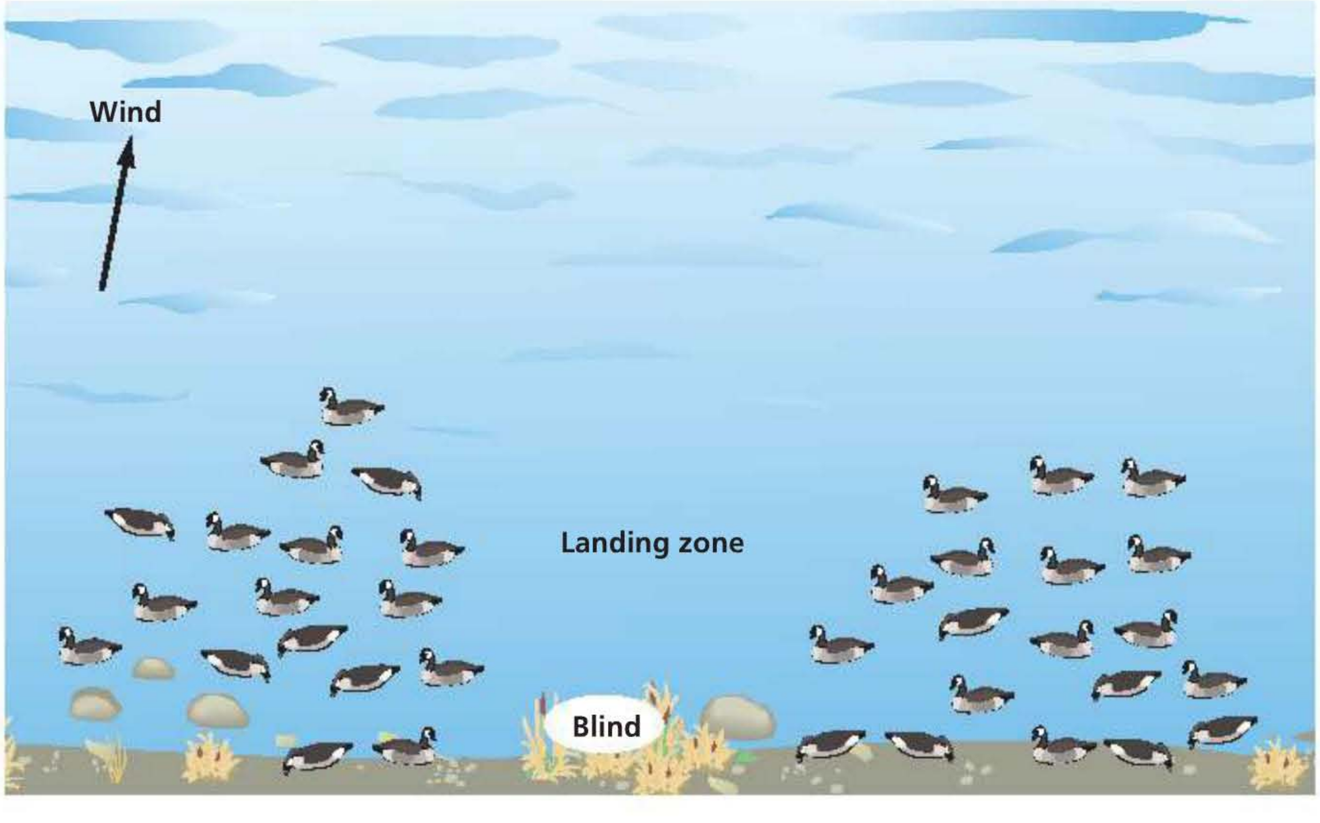

Water spread. Place your decoys in the most protected area of the bay, leaving an opening about 15 yards (13 m) wide for the birds to land in the middle of the spread. The opening should be directly in front of the blind. The outermost decoys should be no more than 35 yards (32 m) from the blind, giving you a distance guide to approaching geese. Depending on the size of the water you're hunting, you'll need from one to six dozen decoys to hunt geese on the water.

HOW TO USE A SHORT REED CALL

The Short Reed Goose Call provides the most versatile range of tone and pitch of any call on the market. Today, many variations are available, made from ABS, polycarbonate and acrylic. Calls made from ABS range are the least expensive. Polycarbonate calls are more expensive. Acrylics may be the most expensive, but they have the crispest tones.

The grip is critical. To create back pressure, put your index finger at the base of your thumb to form an O at the very end of the call. Then close your pinky finger and your ring finger against your palm as if in a fist. You use your middle finger to regulate the air pressure and flow to create all the sounds.

Play the call like you would play an instrument. Don't blow with your mouth. Air pressure should originate in your diaphragm.

Use your middle finger to change tones and your off-hand to direct the sound up and down and left and right. Begin to muffle the call as the birds get closer. The subtle sound keeps them coming.

Most goose hunters use a blend of several types of decoys. Here, the hunters are setting up shells. Flags and silhouettes are handy because they take up less storage space and can add depth to a spread.

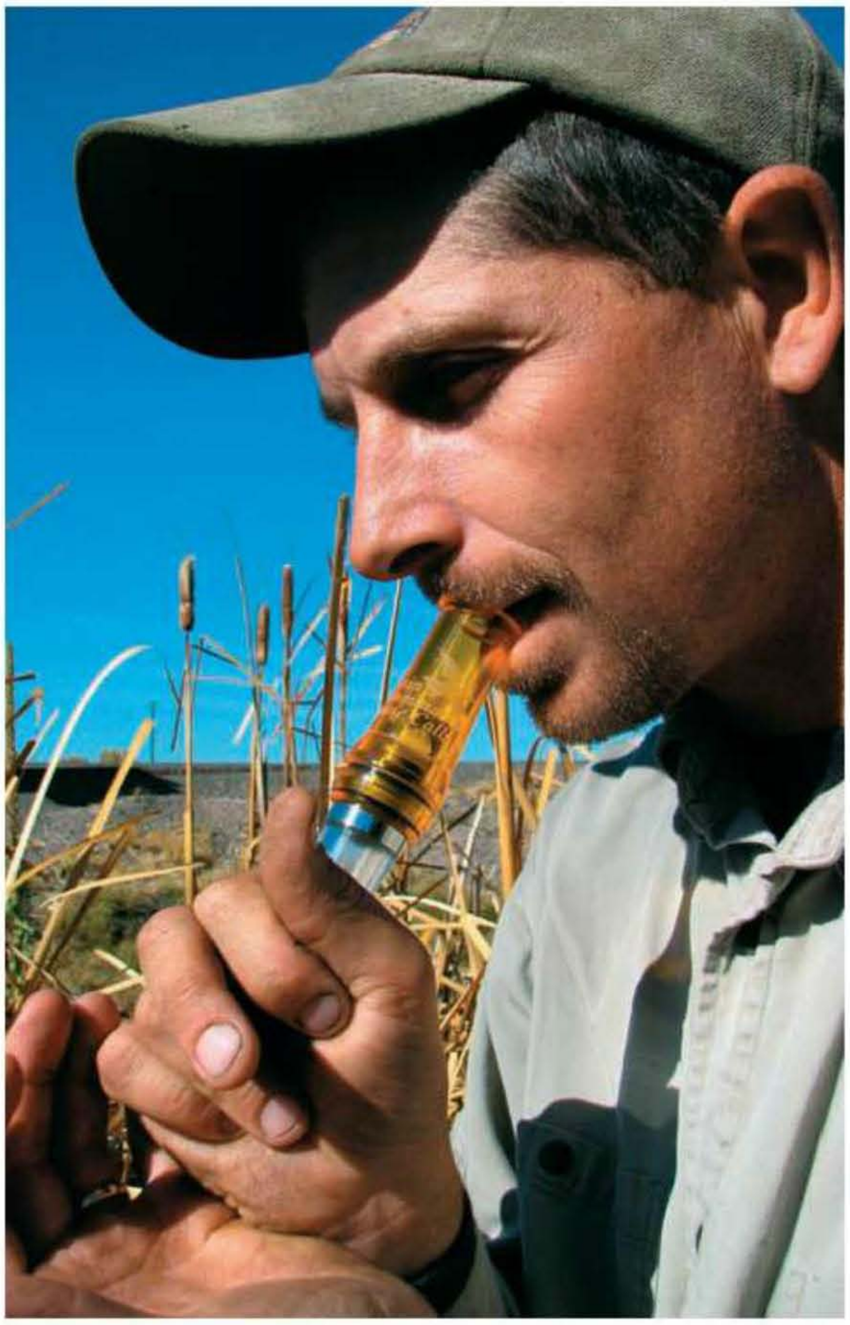

Outfitter Darren Roe demonstrates proper calling form (right), employing a Basin Abomination short reed goose call. Use your middle finger to change tones and your off-hand to direct the sound up and down, left and right. Begin to muffle the call as the birds get closer. The subtle sound keeps them coming.

GOOSE HUNTING TIPS

Locate geese midmorning or late afternoon by scouting grain fields near resting waters. Secure permission and hunt the field the next feeding period.

Use flags to gain the attention of distant geese. The waving flags imitate geese fluttering in to a flock of feeding geese.

Pass-shoot geese flying to and from feeding areas. Many refuges allow hunting along the boundaries.

INDEX

A

Action, 22, 23, 32
Ammunition, 24–25, 28
Antelope. See *Pronghorns*
Antler rattling, 64, 78
Antlers, 61
Arkansas style calls, 140
Arrows, 35, 36, 37

B

Ballistics tables, 24, 25
Bear. See *Black bear*
Beds and bedding, 3, 14, 47, 50
Bellows calls, 130
Big game, 8, 10–11, 14, 34, 40, 59–85
Binoculars, 46, 50, 70, 73, 140
Black bear, 82–83
  Breeding season, 82
  Calling for, 82
  Daily and seasonal movements, 82
  Description, 82
  Escape tactics, 82
  Food and feeding, 82
  Habitat, 12, 82
  Hibernation, 82
  Hunting techniques, 49, 82–83
  Range map, 82
  Record weight, 82
  Rifles, ammunition, 83
  Senses, 82
  Sign, 83
  Territories, 82
Black-tailed deer, 71–73
  Daily and seasonal movements, 73
  Description, 71
  Escape tactics, 73
  Food and feeding, 71, 73
  Habitat, 71, 72, 73
  Hunting techniques, 71, 72, 73
Blinds, 49, 65, 81, 133, 143
Boats, 46, 55
  Canoes, 55, 144
  Double-pointed duck, 55
  Duck, 140
  Jon, 55
  Low-profile, 147
  Sculling, 55, 147
  Semi-V, 55, 147
Bobwhite quail, 13, 114–16
  Daily and seasonal movement, 114
  Description, 114
  Escape tactics, 114–15
  Foods and feeding, 16, 114, 115
  Habitat, 114, 115
  Hunting techniques, 115, 116
  Nesting, 114
  Range map, 114
  Roosting, 114, 115
  Shotguns, chokes, shot sizes, 115
  Sounds, 114, 116
Bolt action rifles, 22–23
Bore, 23, 28, 32
Bore sighting, 28
Bowhunting, 35–39, 50, 76, 77, 81
  Energy, 35
  Equipment, 35, 36, 39
  Types of arrowheads, 36
  Types of bows, 35
Box calls, 130
Brant geese, 151
Bugling, 52, 85–87
Bullets, 23, 24–25, 28–31, 41
  Ballistics tables, 24
  Energy, 24, 25
  Factors that affect flight path, 24, 30
  Mushrooming, 24
  Slant range effect, 30
  Styles, 24, 25
  Trajectory, 24, 25, 41
  Velocity, 24, 25
  Weight, 24
  Wind drift, 30, 31

C

Caliber, 22–24, 40
California quail, 117
Calls and calling, 78, 103, 130, 140, 153
  Arkansas style, 140
  Bellows, 130
  Coyote howler, 130
  Crow call, 130
  Diaphragm, 76, 130
  Diver, 147
  Flute, 76, 130, 140
  Hawk screamer, 130
  Owl hooter, 130
  Pigtail, 76
  Pileated woodpecker, 130
  Predator, 130
  Pronghorn, 81
  Push-pull, 130
  Silent whistle, 130
  Slate-and-peg, 130
  Standard, 140
  Tube, 76, 130
Canada geese, 149, 150, 151
  Calls for, 140, 153
  Decoys, 149, 152, 153
Canoes, 55
Canvasbacks, 146
Carbines, 22
Cartridges, 24, 25
Chokes, 33
Chuckar partridge, 126
Clothing, 42, 48, 49, 50, 52, 66, 87, 104, 115, 116, 118, 136, 140, 143, 147
Color vision of game, 8, 10, 60
Compound bows, 35, 36, 38
Cottontail rabbits, 88–91
  Breeding season, 88
  Daily movements, 88
  Description, 89
  Diseases, 88
  Escape tactics, 88
  Foods and feeding, 88
  Habitat, 88, 89
  Home range, 88, 91
  Hunting techniques, 88, 90, 91
  Range map, 88
  Rifles, scopes, ammunition, 88
  Senses, 9, 88
  Shotguns, chokes, shot size, 33
  Sign, 90

Cover, 12, 13–15, 50, 51, 73
  Corridors, 52, 72
  Escape, 13
  For bedding, 13, 47
  From the elements, 14
  Islands, 52
  Midday, 14
  Types, 13–14
Coyotes, 103–5
  Calls for, 103, 104, 105
  Daily movement, 105
  Decoys, 105
  Dens and denning, 103
  Description, 103
  Foods and feeding, 103
  Habitat, 103
  Hunting techniques, 103, 104
  Range map, 103
  Rifles, ammunition, 104
  Shotguns, ammunition, 104
  Sign, 103
  Territories, 104

D

Deer, 16, 25, 49, 55, 60–73
Diaphragm calls, 76, 130
Diving ducks, 146–48
  Calls for, 147
  Decoys, 147, 148
  Description, 139, 146
  Foods and feeding, 146
  Hunting techniques, 147–48
  Migration, 147
Dogs, 53, 54, 56, 57, 88, 90, 91, 111, 115, 122–23, 126, 127, 137
  American water spaniel, 54
  Basset hound, 88
  Beagle, 54, 88
  Black-and-tan hound, 54, 82
  Bluetick hound, 54
  Boykin spaniel, 54
  Brittany spaniel, 54, 115
  Chesapeake Bay retriever, 54, 147
  Cocker spaniel, 54
  English pointer, 54, 115
  English setter, 54, 115
  German shorthair, 54, 115
  Golden retriever, 54
  Gordon setter, 54
  Labrador retriever, 54, 111, 147
  Plotts, 82
  Redbone hound, 54, 102
  Redtick hound, 54
  Springer spaniel, 54, 111
  Training, 54
  Walker hound, 54, 82, 102
  Weimaraner, 54
Doves. See *Mourning doves*
Driving, 52
  For elk, 76
  For jackrabbits, 92
  For mule deer, 70
  For pheasants, 110, 112–13
  For whitetails, 64, 67
  Safety, 52
  Where to drive, 52
Droppings, 47, 53, 63, 75, 79, 83, 91, 103, 122
Ducks, 8, 9, 142–48
  Calls for, 140
  Decoys, 141
  Description, 139
  Foods and feeding, 16,
  Hunting techniques, 49, 143–44
  Shotguns, choke, shot sizes, 33

E

Elk, 16, 74–77
  Beds and bedding, 74, 75
  Bows, 76
  Breeding season, 74, 76
  Calls for, 76–77
  Daily movements, 74, 75
  Description, 74
  Escape tactics, 74
  Foods and feeding, 16, 74, 75
  Habitat, 17, 74
  Hunting techniques, 76, 77
  Migration, 74–75
  Muzzleloading, 76
  Range map, 74
  Record antlers, 74
  Rifles, ammunition, 25, 76
  Senses, 74
  Sign, 46, 75, 76
  Sounds, 75, 76
  Territories, 74, 75
  Varieties, 74
  Wallows, 75, 77

F

Falling block action rifles, 23
Firing zone, 43
Flagging devices, 141, 153
Float-hunting, 55
  For ducks, 144
  For moose, 78, 79
  For squirrels, 98
  For whitetails, 64
Flushing, 53
  For bobwhite quail, 115
  For cottontails, 91
  For grouse, 121–23, 124, 125
  For partridges, 126
  For pheasants, 110, 111, 112–13
  For quail, 119
  For woodcock, 137
Flute call, 76, 130, 140
Fox squirrels, 95, 96–98
  Breeding season, 97
  Calls for, 98
  Daily movements, 95, 97
  Dens and denning, 96
  Description, 95
  Escape tactics, 95, 97
  Foods and feeding, 95, 97
  Habitat, 95, 97
  Hunting techniques, 97, 98
  Nests, 96
  Range map, 95
  Rifles, scopes, ammunition, 97
  Shotguns, shot size, 97
  Sign, 96
  Sounds, 95
Friction calls, 130

G
Gambel's quail, 16, 118
Game trails. See *Trails*
Geese, 149–53
  Calls for, 140, 153
  Daily movements, 152
  Decoys, 151, 152, 153
  Description, 139, 149, 150–51
  Escape tactics, 149
  Foods and feeding, 16, 149
  Habitat, 1551–52
  Hunting techniques, 49, 49, 151–53
  Migration, 149
  Shotguns, chokes, shot sizes, 33, 152
  Species, 150–51
Gray partridge. See *Hungarian partridge*
Gray squirrels, 94, 96–98
  Breeding season, 97
  Daily movements, 94, 97
  Dens and denning, 96
  Description, 94
  Escape tactics, 94, 97
  Foods and feeding, 94, 97
  Habitat, 94, 97
  Home range, 94
  Hunting techniques, 97, 98
  Nests, 96
  Range map, 94
  Rifles, scopes, ammunition, 97
  Senses, 94
  Shotguns, shot size, 97
  Sign, 96
  Sounds, 98
Ground squirrels, 100
Grouse, 8, 17, 33, 121–25

H
Habitat, 12–19, 46
Hares, 92, 93
Hearing. See *Senses*
Hounds, 54, 82, 83, 90, 93, 102. See also *Dogs*
Hungarian partridge, 127
Hunting equipment
  For big game, 59
  For small game, 87
  For upland game birds, 107
  For waterfowl, 140–41
Hunting techniques
  Antler rattling, 64
  Baiting, 82, 83
  Jump-shooting, 142
  Leapfrog technique, 113
  Pass-shooting, 136, 144, 153
  Running-and-gunning, 133, 134
  Sitting-and-waiting, 133, 134
  Sustained-lead technique, 34, 39
  Swing-through technique, 34, 47
  Tag-team, 134
  Walk-and-wait, 118, 122

J
Jackrabbits, 92
Jon boats, 55
Jump-shooting, 142

L
Leapfrog technique, 113
Lever action rifles, 23
Licenses, 45
Locator calls, 130, 131
Longbows, 35, 36, 38

M
Mallard ducks, 8, 143
Mergansers, 146
Metatarsal glands, 11
Moose, 5, 78–79
  Beds and bedding, 78
  Breeding season, 78
  Daily and seasonal movements, 78
  Description, 78, 79
  Escape tactics, 78, 79
  Foods and feeding, 78, 79
  Habitat, 17, 78, 79
  Home range, 78
  Hunting techniques, 15, 49, 78, 79
  Range map, 78
  Record antlers, 78
  Rifles, ammunition, 25, 78
  Senses, 78
  Sign, 71, 78, 79
  Wallows, 78
  Water, 49
Mountain quail, 119
Mourning doves, 49, 135–36
  Daily movements, 135
  Description, 135
  Escape tactics, 135
  Foods and feeding, 135–36
  Habitat, 135, 136
  Hunting techniques, 49, 136
  Migration, 135
  Nesting, 135
  Range map, 135
  Roosting, 135, 136
  Shotguns, chokes, shot sizes, 33, 136
  Water and watering sites, 49, 107, 135, 136
Mule deer, 11, 68–70
  Beds and bedding, 68, 70
  Breeding season, 69
  Daily movements, 68–69
  Description, 68
  Escape tactics, 11, 68, 70
  Food and feeding, 68–69, 70
  Habitat, 68, 70
  Home range, 68
  Hunting techniques, 69–70
  Migration, 69
  Range map, 68
  Record antlers, 68
  Record weight, 69
  Rifles, ammunition, 70
  Senses, 11, 68
  Sign, 71, 69
  Varieties, 69
Muzzleloaders, 40–41, 76

P
Partridge, 126–27
Pass-shooting,
  For doves, 136
  For geese, 151, 153
  For puddle ducks, 144
Pheasants. See *Ring-necked pheasants*
Pigtail call, 76
Prairie dogs, 99
Pronghorns, 11, 80–81
  Breeding season, 81
  Calls for, 81
  Daily movements, 80
  Decoys, 81
  Description, 80, 81
  Escape tactics, 11, 80
  Foods and feeding, 80
  Habitat, 80
  Home range, 80–81
  Hunting techniques, 49, 81
  Record horns, 80
  Rifles, scopes, ammunition, 81
  Senses, 80
  Territories, 81
  Water and watering sites, 16, 49
Puddle ducks, 139, 142–45
  Calls for, 143

  Decoys, 143, 144, 145
  Description, 142
  Food and feeding, 142
  Habitat, 142, 143
  Hunting techniques, 142–45
Pump action rifles, 23
Pump action shotguns, 32
Push-pull calls, 130

Q

Quail, 12, 33, 114–20

R

Rabbits, 8, 17, 54, 88–91
Raccoon, 10, 101–2
  Beds and bedding, 101
  Calls for, 102
  Daily movements, 101
  Dens and denning, 101
  Description, 101
  Escape tactics, 101, 102
  Foods and feeding, 101
  Habitat, 101, 102
  Handguns, 102
  Hibernation, 101
  Hunting techniques, 102
  Range map, 101
  Record weight, 101
  Rifles, 102
  Senses, 101
  Shotguns, shot sizes, 33, 102
  Sign, 102
Recurve bows, 35, 36, 38
Redheads, 146
Red squirrels, 95
Regulations, hunting, 7
Rifle sights, 26–29
Rifles, 5, 22–31
  Actions, 22
  Ammunition, 24–25
  Caliber, 22–23
  Combination guns, 35, 87
  Performance, 26, 28
  Shooting a rifle, 28, 30–31
  Sights, 26–29
  Weight, 22
Ring-necked pheasants, 5, 11, 15, 109–13
  Daily and seasonal movements, 108, 109
  Description, 108, 109
  Escape tactics, 11, 108
  Foods and feeding, 16, 108, 109, 113
  Habitat, 12, 13, 17, 107, 108, 109, 110, 112
  Home range, 108
  Hunting techniques, 109, 110, 111, 112
  Nesting, 108, 110
  Range map, 108
  Roosting, 108
  Senses, 9, 108
  Shotguns, chokes, shot sizes, 33, 110
  Sounds, 108, 109, 112
Ross geese, 151
Ruffed grouse, 121–23
  Breeding season, 121
  Daily and seasonal movements, 121
  Description, 122
  Drumming, 121, 122, 123
  Escape tactics, 121–23
  Foods and feeding, 121, 123
  Habitat, 12, 121
  Hunting techniques, 121–23
  Nesting, 121
  Range map, 121
  Roosting, 121, 122
  Shotguns, choke, shot sizes, 122
  Sign, 121, 122
  Sounds, 121, 122, 123
  Territories, 121
Running-and-gunning, 134

S

Sage grouse, 17, 125
Sea ducks, 146
Safety, 42–43, 55, 67, 92, 110, 136
Scaled quail, 120
Scent, 11
Scopes, 26, 29, 46, 50, 70, 73
Scouting for game, 17, 46–47, 49, 59, 62, 64, 65, 151
Sculling boats, 55
Semi-automatic action rifles, 23
Semi-automatic action shotguns, 32
Semi-Vs, 55
Senses, 8–11
Sharp-tailed grouse, 107, 124
Shotguns, 32–34
  Actions, 32
  Barrel type, 32
  Bore size, 32
  Chamber length, 32
  Chokes, 32, 33
  Shells, 32
  Shooting a shotgun, 34
  Weight, 32
Sitting-and-waiting technique, 134
Slate-and-peg calls, 130
Slugs, 34
Small game, 7, 14, 54, 87–105
Ammunition, 40, 87
Snow geese, 150
Snowshoe hares, 33 93
Squirrels, 6, 33, 46, 54, 55, 94–98, 100
Stalking, 50–51
  For black bears, 82
  For elk, 76
  For geese, 152
  For moose, 78
  For mule deer, 70
  For pronghorns, 81
  For sage grouse, 125
  For whitetails, 64
  For wild boars, 84
Stand-hunting, 49, 65
  For elk, 76, 77
  For deer, 65
  For moose, 78
  For mule deer, 70
  For pronghorns, 81
  For squirrels, 97, 98
  For whitetails, 64, 65
Still-hunting, 48, 66
  For bear, 82
  For elk, 76, 77
  For moose, 78
  For mule deer, 70
  For squirrels, 97
  For whitetails, 64, 66–67
Succession, 12
Swing-through technique, 34, 147

T

Tag-team hunting, 134
Teals, 143
Team-hunting, 98
Tracks, 47, 53, 63, 77, 83, 85, 90, 104, 112, 122
Trails, 47, 49, 85
Trajectory
  Of arrows, 37
  Of rifle bullets, 24, 25, 30
  Of muzzleloader balls, 41
Tube calls, 130
Turkey. See *Wild turkey*

U

Upland game birds, 7, 40, 54, 107–37

V

Varmints, 40
Vision. See *Senses*

W

Water, 16, 17, 49
Waterfowl, 7, 17, 21, 46, 49, 50, 54, 55, 139–53
  Calls for, 140
  Decoys, 141
  Equipment, 140–41
  Migration, 17
  Muzzleloaders, 40
  Population, 7
  Senses, 8
  Shotguns, ammunition, 32, 140
White-fronted geese, 150
White-tailed deer, 10, 60–67
  Beds and bedding, 62
  Breeding season, 62
  Daily and seasonal movements, 62
  Description, 60
  Escape tactics, 60, 61
  Foods and feeding, 16, 62
  Habitat, 12, 17, 62, 63
  Home range, 60, 62
  Hunting techniques, 61, 64–67
  Range map, 60
  Record antlers, 60
  Record weight, 60
  Rifles, ammunition, 65
  Senses, 10, 60,
  Sign, 62, 63
  Sounds, 60
  Varieties, 60
Wild boar, 84–85
  Beds and bedding, 84
  Daily and seasonal movements, 84, 85
  Description, 84
  Escape tactics, 84–85
  Foods and feeding, 84
  Hunting techniques, 84–85
  Rifles, ammunition, 85
  Senses, 84
  Sign, 85
  Wallows, 85
Wildlife management, 12, 16, 45
Wild turkey, 8, 128–34
  Breeding season, 129
  Calls for, 130–31, 132, 133
  Daily movement patterns, 131–33, 134
  Decoys, 130–31
  Description, 128–29
  Escape tactics, 129, 132
  Foods and feeding, 132, 133
  Habitat, 134
  Hunting techniques, 132–34
  Muzzleloaders, 40
  Nesting, 133
  Range map, 128
  Roosting, 131, 132, 134
  Senses, 8, 9, 129
  Shotguns, choke, shot sizes, 33
  Sounds, 128, 129, 132–33, 134
  Strutting zones, 131–33, 134
  Subspecies, 129
  Water and watering sites, 132
Wind drift, 30, 31
Woodcock, 11, 137

PHOTO CREDITS

Note: T = Top, B = Bottom, C = Center, L = left, R = Right, I = inset

© Grady Allen: pp. 101, 116BR
© Charles Alsheimer: p. 63TR
© Erwin Bauer: pp. 8R, 26, 61C, 61CR, 61BL, 76B, 83T, 123BR
© Craig Blacklock: p. 58
© Les Blacklock: p. 12
© Tom Brakefield: p. 61T, Back Cover TL
© The Browning Company: pp. 48C, 149
© Glenn D. Chambers: p. 114
© Tim Christie: p. 93I, Front Cover BC
© Herbert Clarke: pp. 117, 119
© Daniel J. Cox: p. 16, 61 far left
© John Ebeling: p. 11BC
© John Ford: Front Cover T
© Michael Francis: pp. 99B, 104, Front Cover TC
© Michael Furtman: pp. 138, 146T
© Eric J. Hansen: pp. 6, 20, 44, 62, 71
© istock photo: p. 100
© Larry D. Jones: pp. 75BC, 75BR
© Gary Kramer: pp. 72, 73R, 85TL, 86, 105BL, 122, 128, 133, 142BR
© Lon E. Lauber: pp. 35, 39T, 65T, 66, 73L, 84, 92, 105BR, 129, 132
© Gary Lewis: pp. 85TR, 85B, 153R
© Tom Mangelson: pp. 54, 109TR
© Bill Marchel: pp. 142BL, 142CR, 150T, 150BL, 150BR, Front Cover B
© Tom Martinson: p. 8L
© Jay Massey: p. 50
© Worth Mathewson: p. 94
© Bill McRae: pp. 77BR, 81TL, 109TL
© Wyman Meinzer: p. 120
© Arthur Morris/Birds as Art: pp. 151TR, 151BR
© William H. Mullins: p. 124
© Doug Murphy: p. 11TL
© Nebraska Game And Parks Commission, 13
© Scott Nielsen: p. 142T, 146BL, 146BR
© Jerome B. Robinson: p. 55
© Lynn Rogers: p. 93T
© Leonard Lee Rue III: pp. 47CR, 63TL, 89T, 102T, 135
© Dwight R. Schuh: p. 77T
© Ron Shade: pp. 69, 75T
© Jerry Smith: p. 61BR
© Dale C. Spartas: pp. 11TR, 123T, 123BL
© Ron Spomer: pp. 10TR, 11BL, 11BR, 19BL, 95, 151L
© Norm Strung: p. 65B
© Syl Strung: p. 79BR
© Ken Thommes: pp. 99T, 105TL, 105TR, Back Cover BR
© Charles Waterman: p. 125
© Chuck Wechsler: pp. 61CR, 61 far right
© Ron Winch: p. 79T
© James Zacks: p. 10TL
© Gary Zahm: pp. 10BR, 118, 136

Creative Publishing international

Your Complete Source of How-to Information for the Outdoors

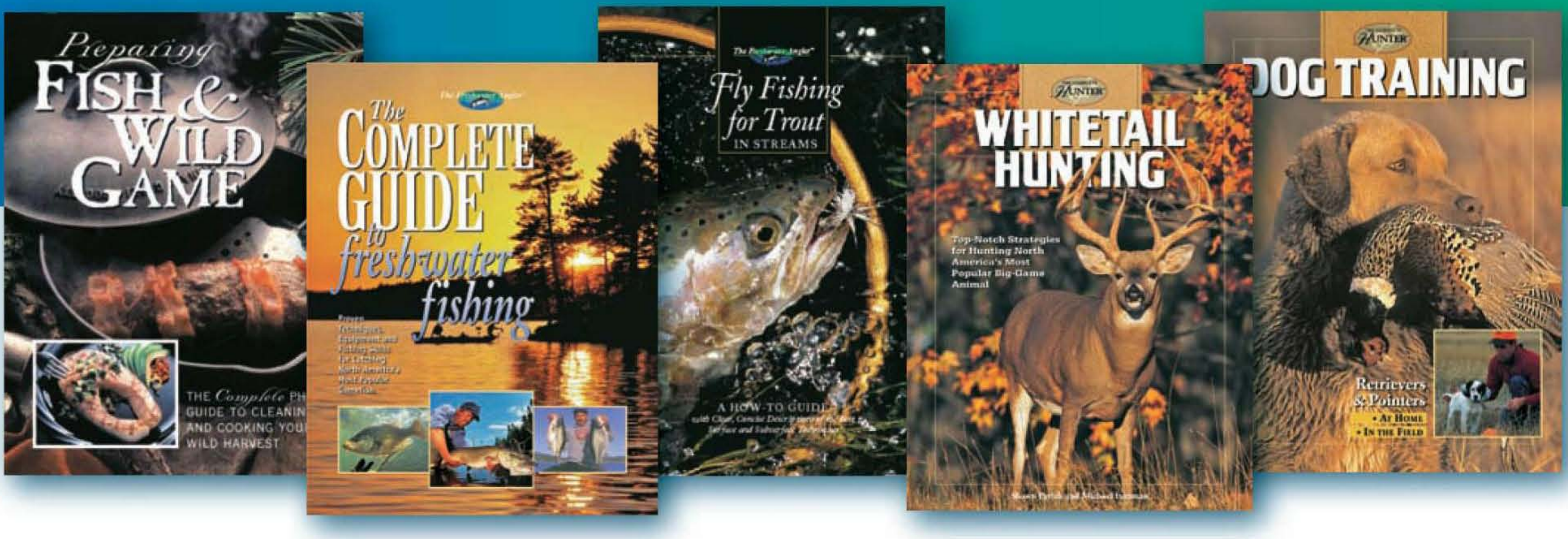
Preparing FISH & WILD GAME
THE Complete PHOTO GUIDE TO CLEANING AND COOKING YOUR WILD HARVEST
The COMPLETE GUIDE to freshwater fishing
Fly Fishing for Trout IN STREAMS
A HOW-TO GUIDE
WHITETAIL HUNTING
Top-Notch Strategies for Hunting North America's Most Popular Big-Game Animal
DOG TRAINING
Retrievers & Pointers
• AT HOME
• IN THE FIELD

Hunting Books
- Advanced Turkey Hunting
- Advanced Whitetail Hunting
- Beginner's Guide to Birdwatching
- Black Bear Hunting
- Bowhunting Equipment & Skills
- Bowhunter's Guide to Accurate Shooting
- The Complete Guide to Hunting
- Dog Training
- Elk Hunting
- How to Think Like a Survivor
- Hunting Record-Book Bucks
- Mule Deer Hunting
- Muzzleloading
- Outdoor Guide to Using Your GPS
- Waterfowl Hunting
- Whitetail Addicts Manual
- Whitetail Hunting
- Whitetail Techniques & Tactics
- Wild Turkey

Fishing Books
- Advanced Bass Fishing
- The Art of Freshwater Fishing
- The Complete Guide to Freshwater Fishing
- Fishing for Catfish
- Fishing Tips & Tricks
- Fishing with Artificial Lures
- Inshore Salt Water Fishing
- Kids Gone Campin'
- Kids Gone Fishin'
- Kids Gone Paddlin'
- Largemouth Bass
- Live Bait Fishing
- Modern Methods of Ice Fishing
- Northern Pike & Muskie
- Panfish
- Salt Water Fishing Tactics
- Smallmouth Bass
- Striped Bass Fishing: Salt Water Strategies
- Successful Walleye Fishing
- Ultralight Fishing

Fly Fishing Books
- The Art of Fly Tying + CD-ROM
- Complete Photo Guide to Fly Fishing
- Complete Photo Guide to Fly Tying
- Fishing Dry Flies
- Fly-Fishing Equipment & Skills
- Fly Fishing for Beginners
- Fly Fishing for Trout in Streams
- Fly-Tying Techniques & Patterns

Cookbooks
- All-Time Favorite Game Bird Recipes
- America's Favorite Fish Recipes
- America's Favorite Wild Game Recipes
- Backyard Grilling
- Cooking Wild in Kate's Kitchen
- Dressing & Cooking Wild Game
- The New Cleaning & Cooking Fish
- Preparing Fish & Wild Game
- The Saltwater Cookbook
- Venison Cookery
- The Wild Butcher
- The Wild Fish Cookbook
- The Wild Game Cookbook

To purchase these or other Creative Publishing international titles,
contact your local bookseller, or visit our website at
www.creativepub.com

The Freshwater Angler™ THE COMPLETE HUNTER™ The Complete FLY FISHERMAN™

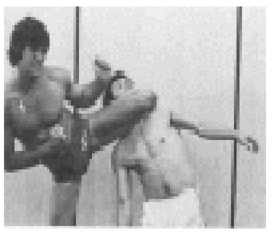

THE 100 DEADLIEST KARATE MOVES

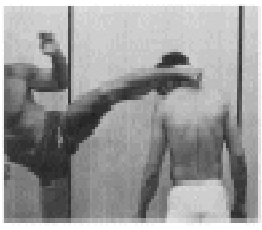

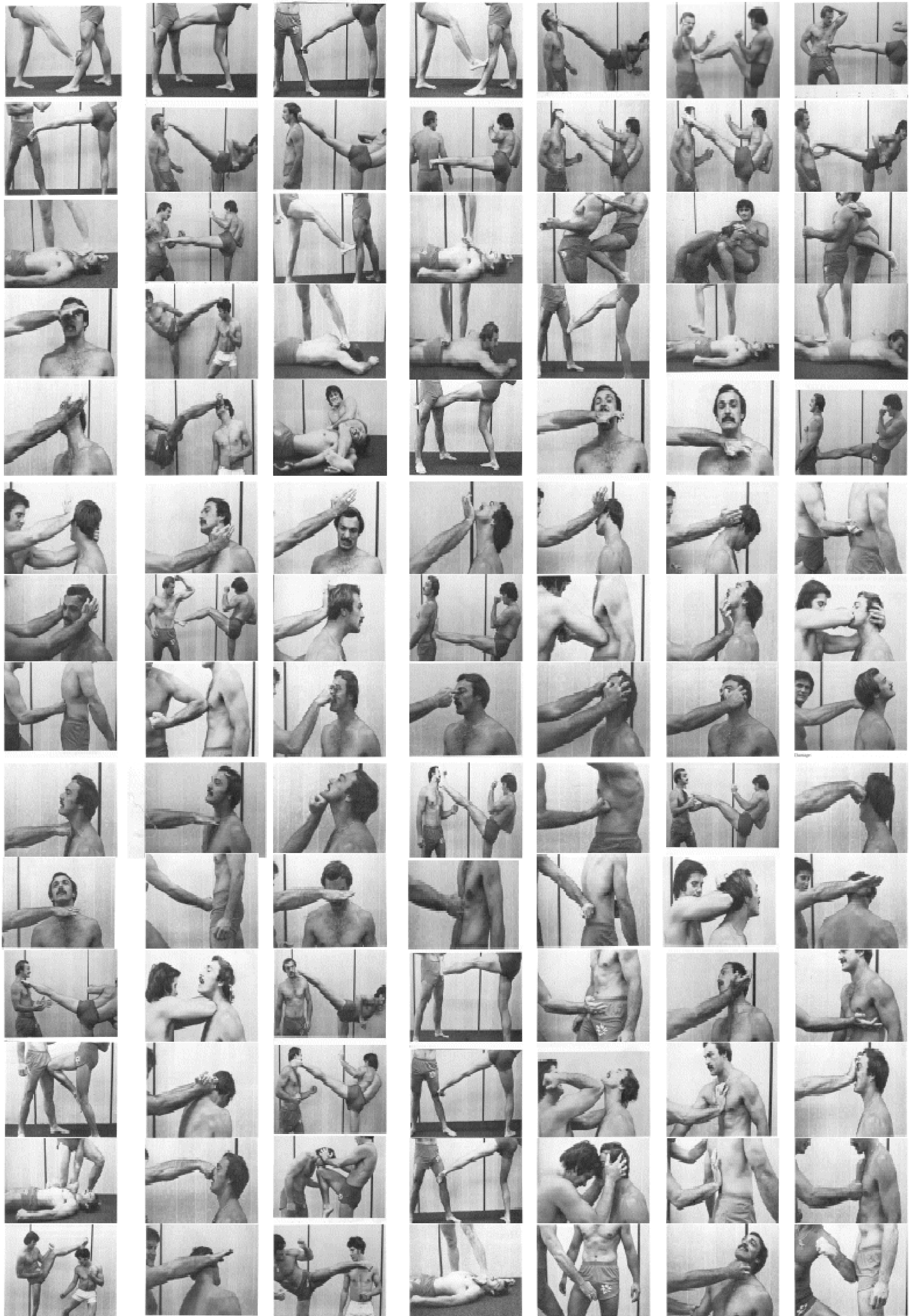

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 1

Fig. 1.

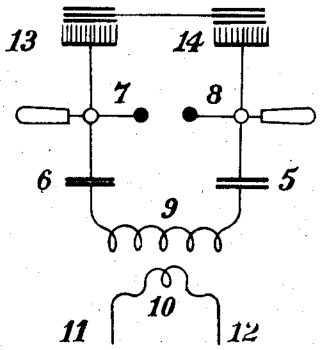
13
14
7
8
6
5
9
10
11
12

Fig. 2.

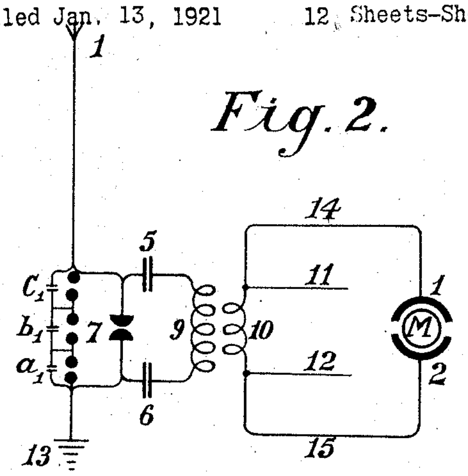
1
14
5
11
C1
b1
7
9
10
a1
12
1
2
6
15
13

Fig. 3.

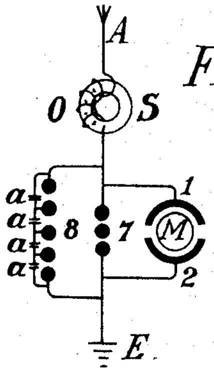
A
O
S
a
a
a
a
8
7
1
2
E.

Fig. 4.

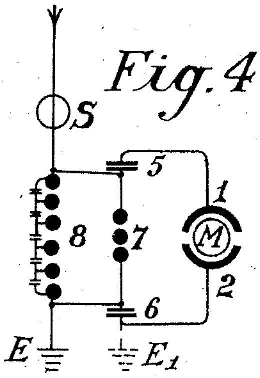
S
5
8
7
1
2
6
E
E1

Fig. 5.

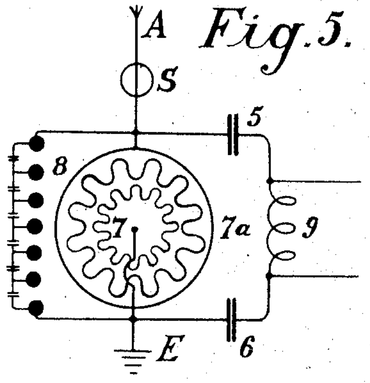
A
S
5
8
7
7a
9
E
6

Fig. 6

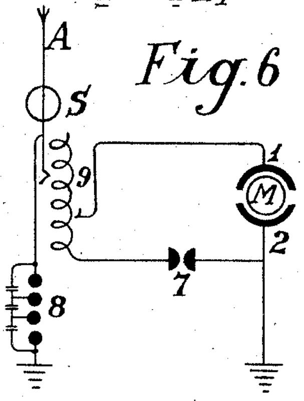

Inventor
Hermann Plauson

by Knight Bros
attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 2

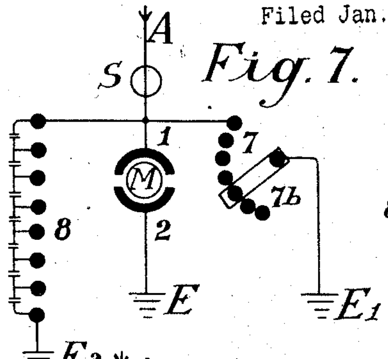
A
S
1
M
2
7
7b
8
E
E1
E2

Fig. 7.

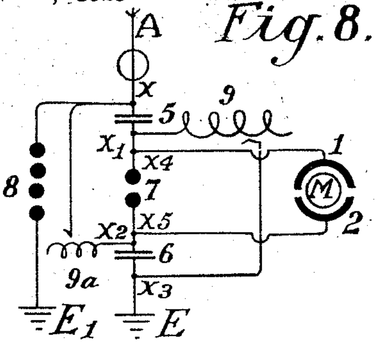
A
X
5
9
X1
X4
1
M
7
8
X5
2
X2
6
9a
X3
E1
E

Fig. 8.

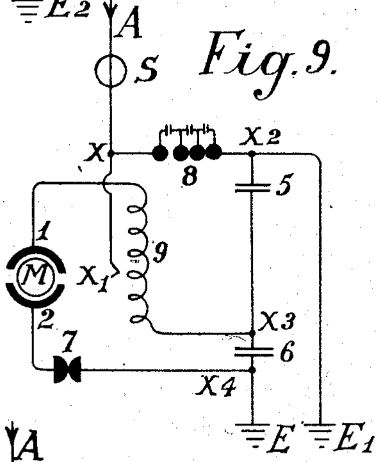
A
S
X
8
X2
5
1
M
X1
9
2
X3
7
6
X4
E
E1

Fig. 9.

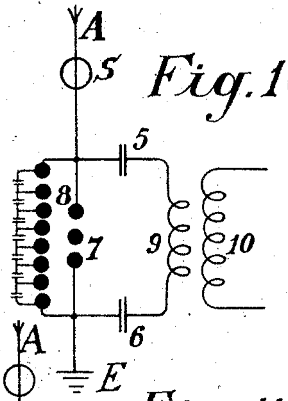
A
S
5
8
7
9
10
6
E

Fig. 10.

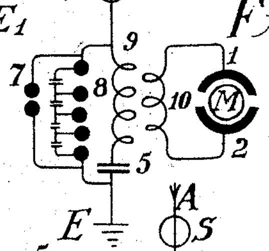
A
S
9
7
8
10
1
M
2
5
E

Fig. 11.

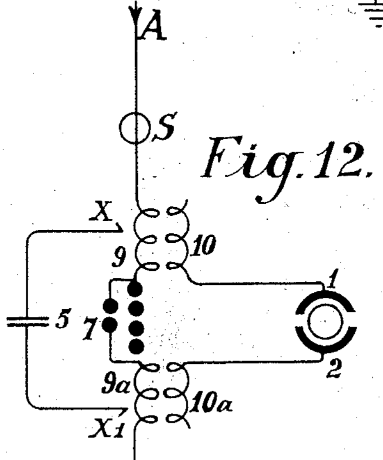
A
S
X
9
10
1
5
7
2
9a
10a
X1

Fig. 12.

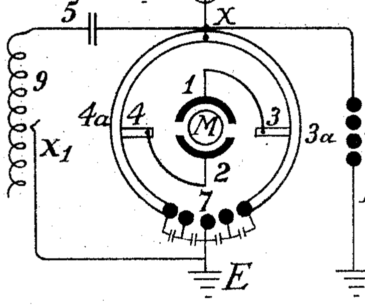
A
S
X
5
9
1
M
4a
4
3
3a
8
X1
2
7
E

Fig. 13.

Inventor

Normann Plauson

by Knight Bros attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 3

Fig. 14.

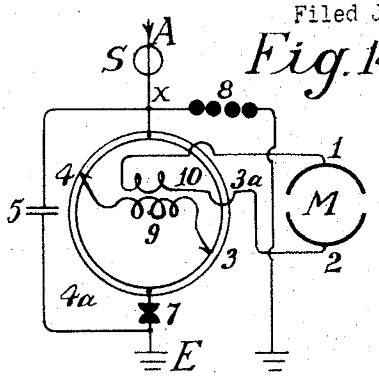
A
S
x
8
1
M
2
4
10
3a
5
9
3
4a
7
E

Fig. 15.

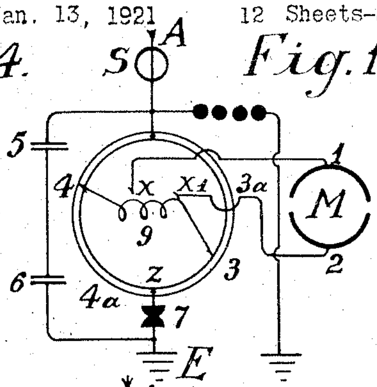
A
S
5
1
M
2
4
x
x1
3a
9
3
6
z
4a
7
E

Fig. 16.

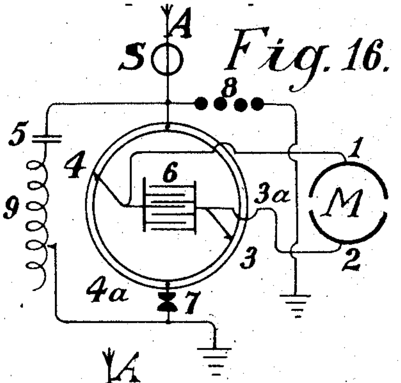
A
S
8
5
1
M
4
6
3a
9
3
2
4a
7

Fig. 17.

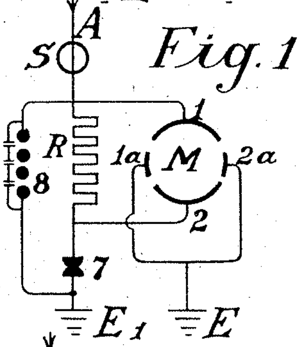
A
S
1
R
1a
M
2a
8
2
7
E1
E

Fig. 18.

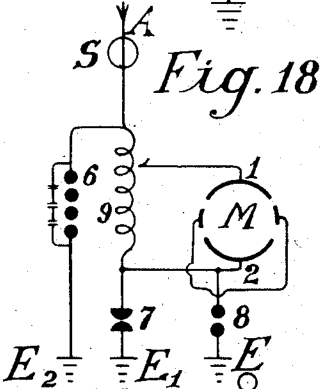
A
S
1
6
M
9
2
7
8
E2
E1
E

Fig. 19.

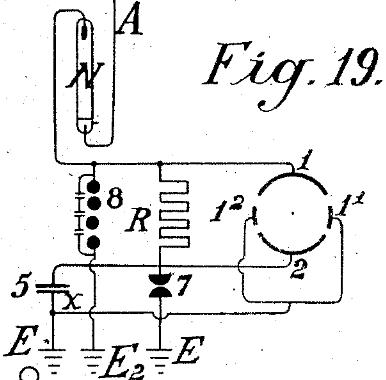
A
N
1
8
R
1²
1¹
2
5
7
x
E
E2
E

Fig. 20.

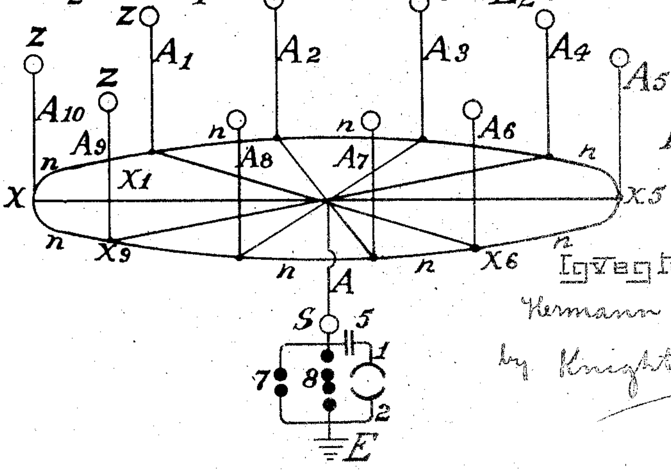
Z
A1
A2
A3
A4
A5
Z
A10
Z
n
n
A6
A9
n
A8
A7
n
X1
X
X5
n
X9
n
n
X6
n
A
S
5
1
7
8
2
E

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 4

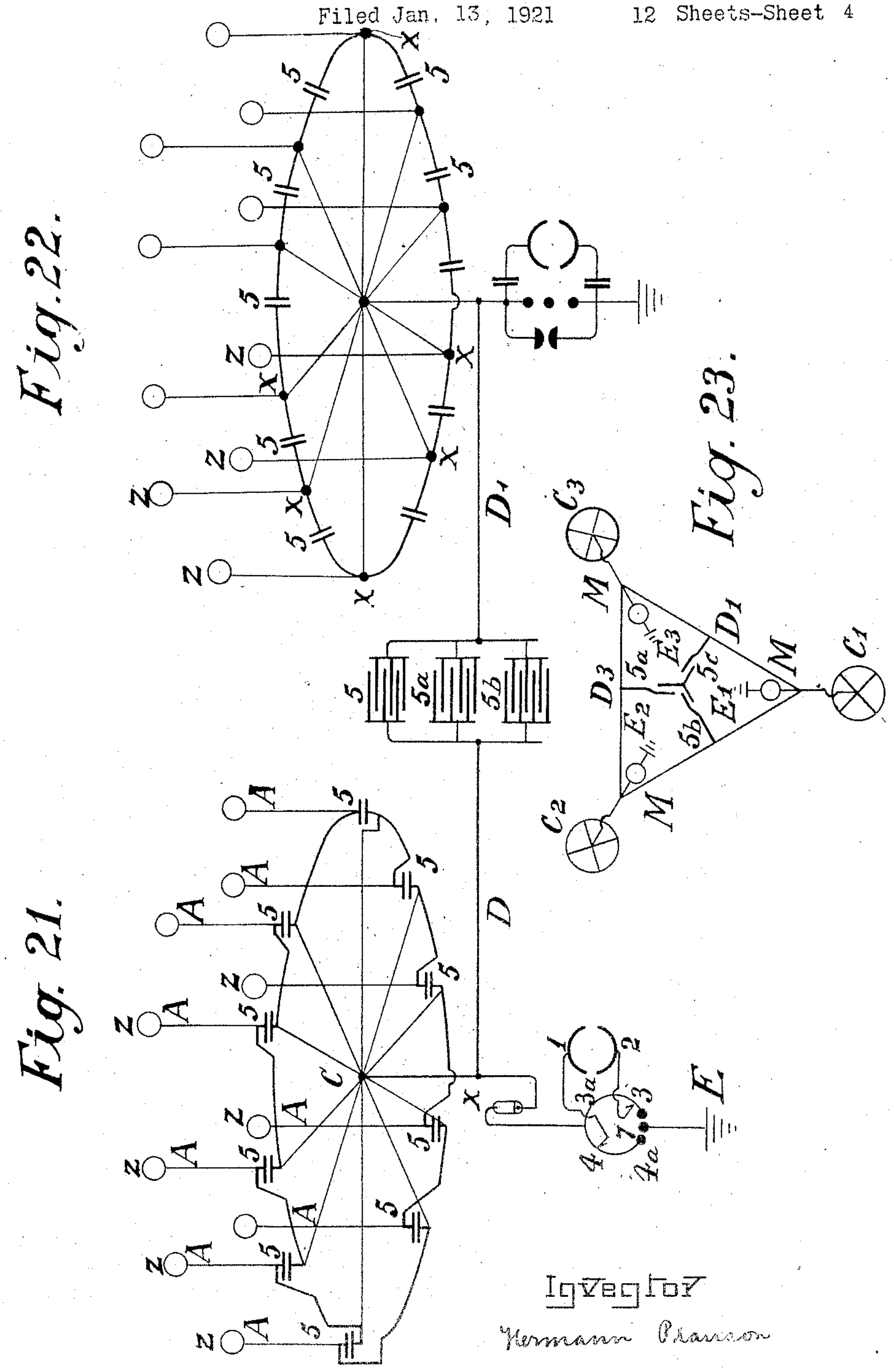
Fig. 22.

Fig. 21.

Fig. 23.
5
X
Z
D1
D
5a
5b
A
C
1
2
3
3a
4
4a
7
E
C1
C2
C3
M
D3
E1
E2
E3
5c

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925.

1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 5

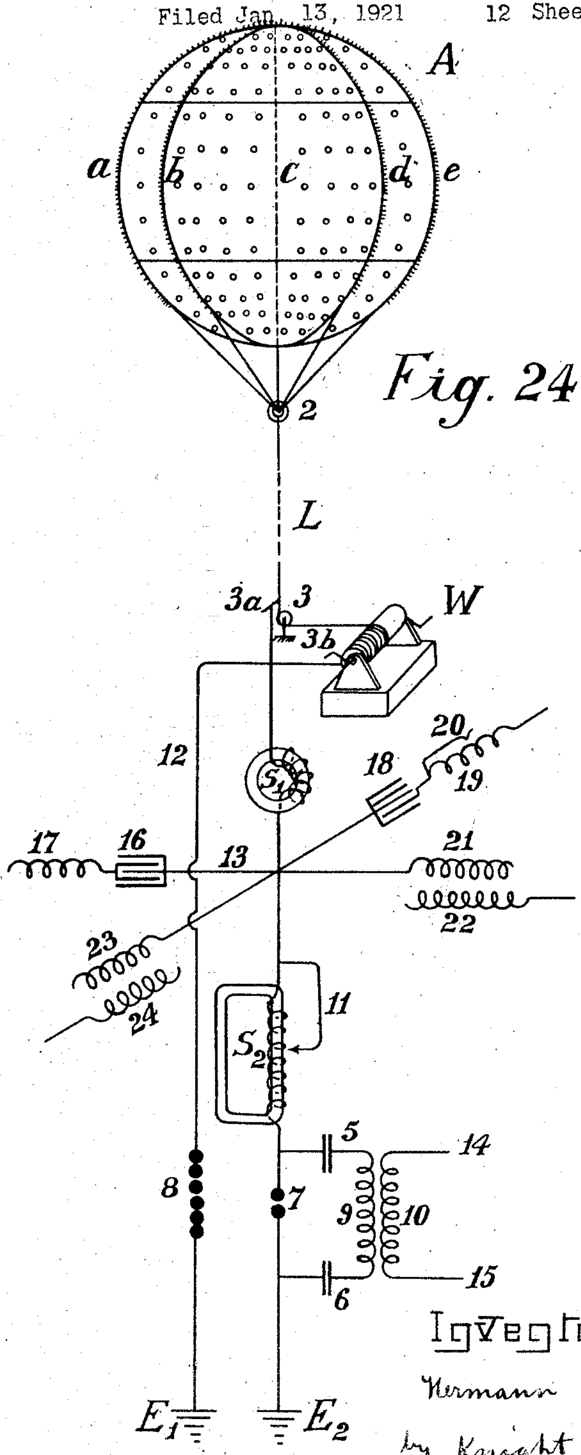
A
a
b
c
d
e
2
L
3a
3
W
3b
12
S_1
18
20
19
17
16
13
21
22
23
24
11
S_2
5
14
8
7
9
10
15
6
E_1
E_2

Fig. 24.

Inventor

Hermann Plauson

by Knight Bros

Attorneys

June 9, 1925. **1,540,998**

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 6

Fig. 25.

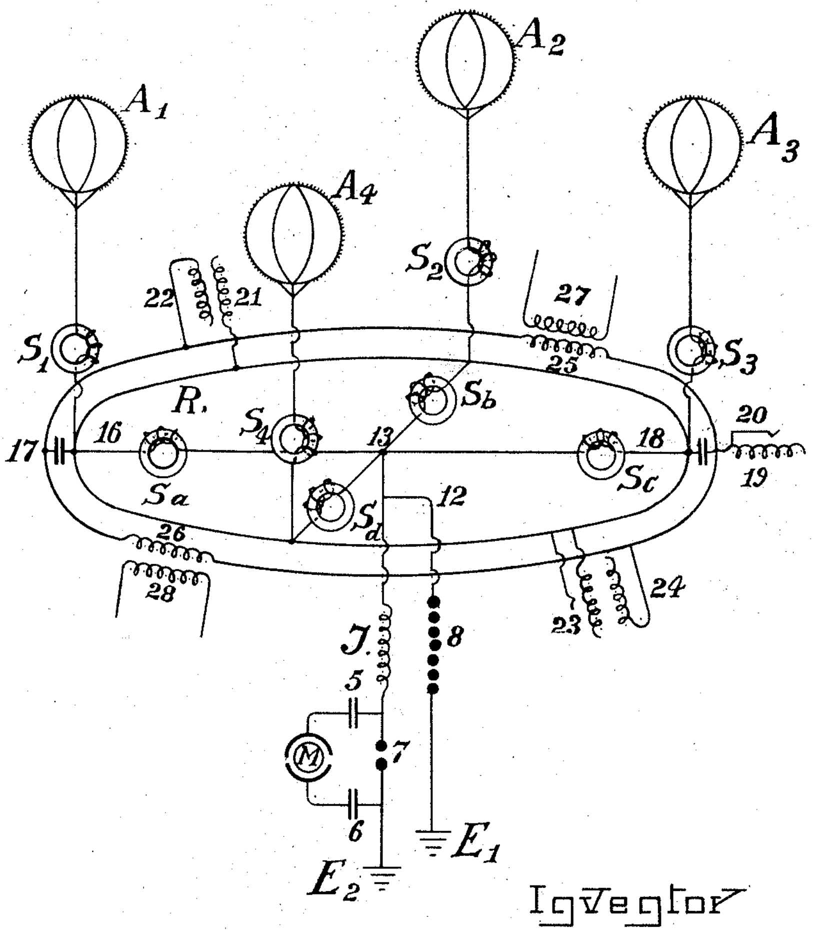

A_1
A_2
A_3
A_4
S_2
22
21
27
S_1
25
S_3
R.
S_b
S_4
13
16
18
20
17
S_a
S_c
19
S_d
12
26
28
24
23
J.
8
5
M
7
6
E_1
E_2

Inventor

Hermann Plauson

by

Knight Bros

attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 7

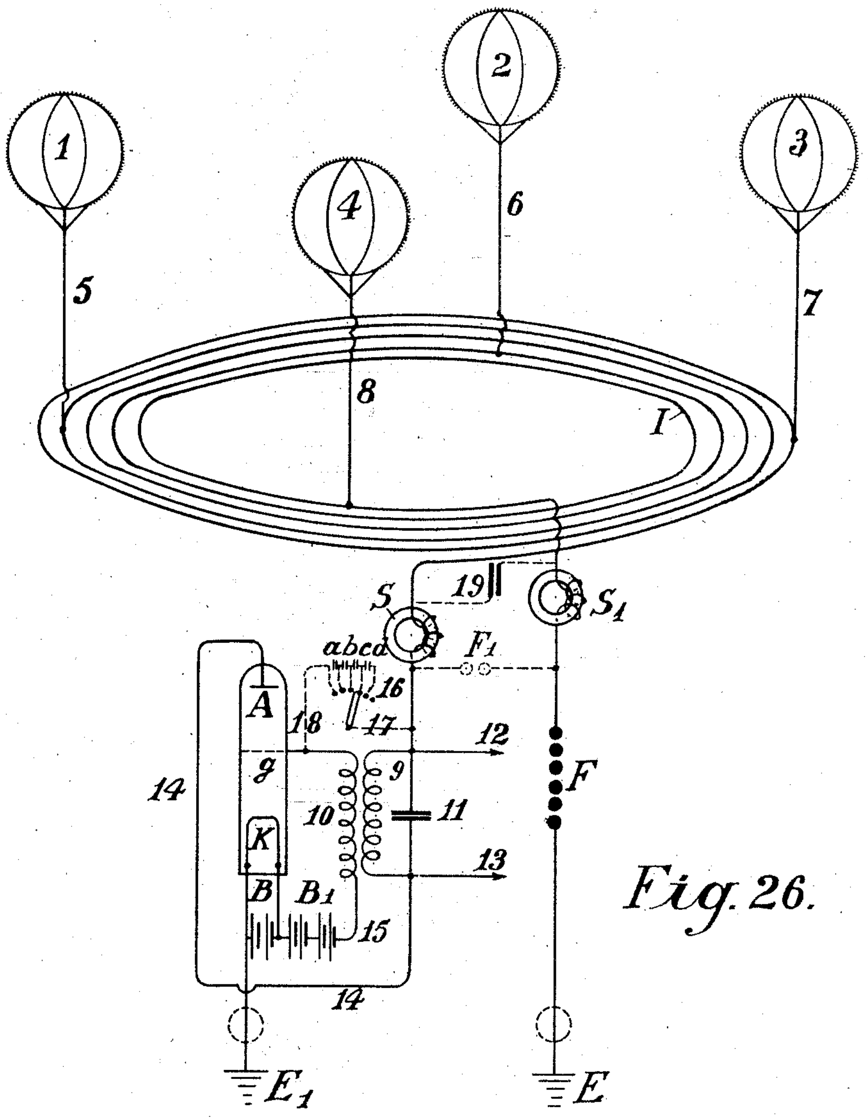
2
1
3
4
6
5
7
8
I
S
19
S_1
abcd
F_1
16
A
18
17
12
g
9
14
10
11
K
13
F
B
B_1
15
14
E_1
E

Fig. 26.

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921. 12 Sheets-Sheet 8

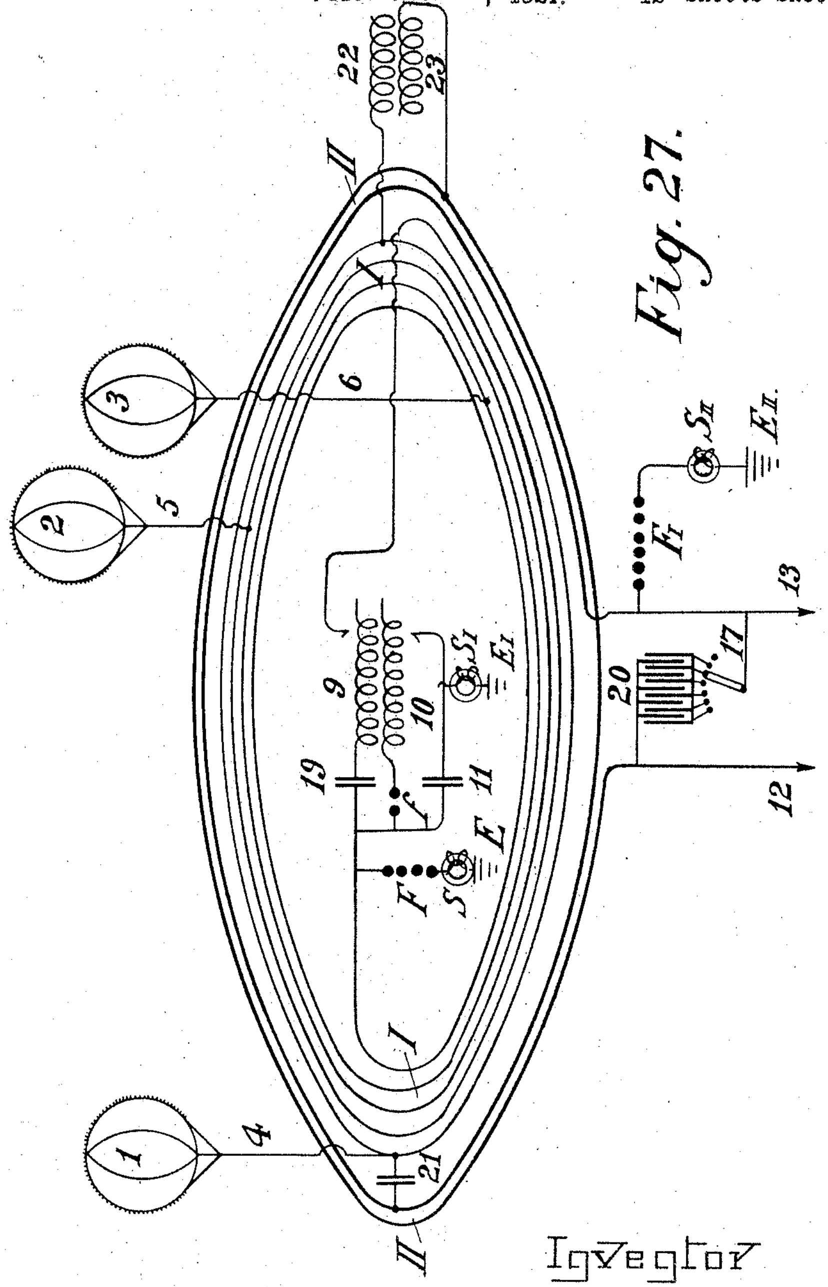
22
23
II
I
6
3
2
5
1
4
9
10
S_I
E_I
19
11
F
E
S
I
21
II
20
17
13
12
F_I
S_{II}
E_{II}

Fig. 27.

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 9

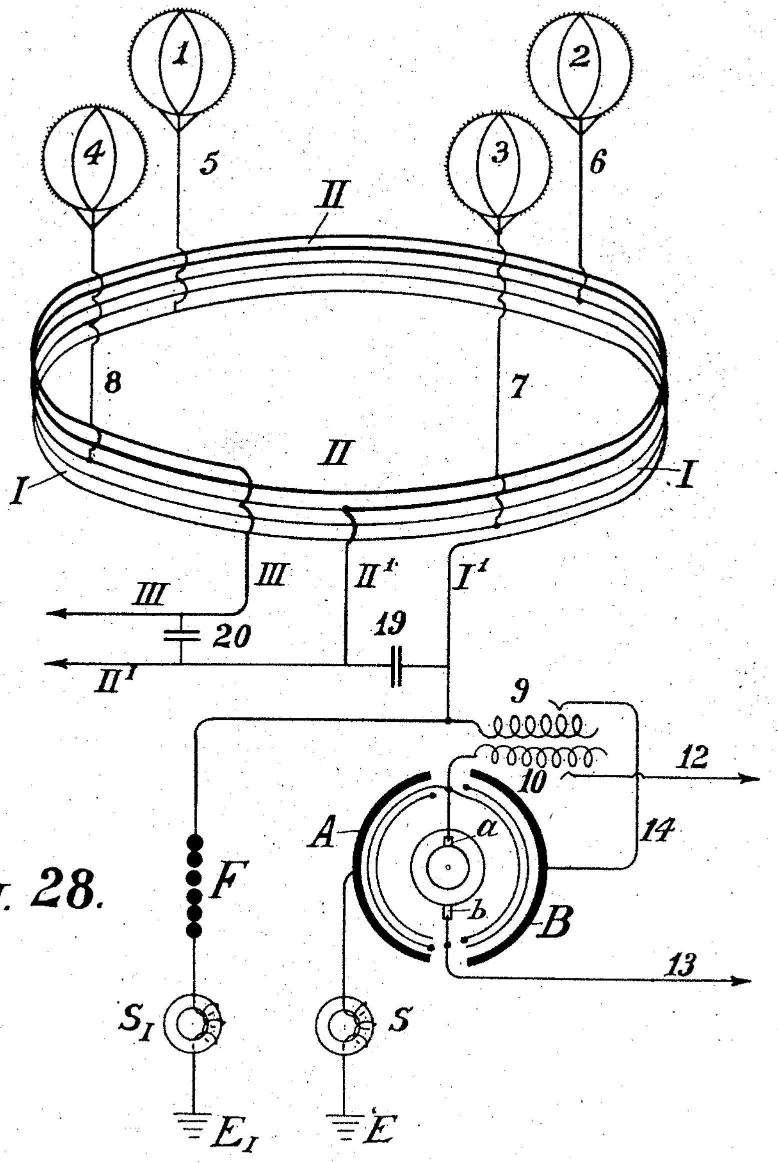
1
2
3
4
5
6
II
8
7
II
I
I
III
II¹
I¹
III
20
19
II¹
9
12
10
A
a
14
F
b
B
13
S₁
S
E₁
E

Fig. 28.

Inventor
Hermann Plauson
by
Knight Bros
attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921. 12 Sheets-Sheet 10

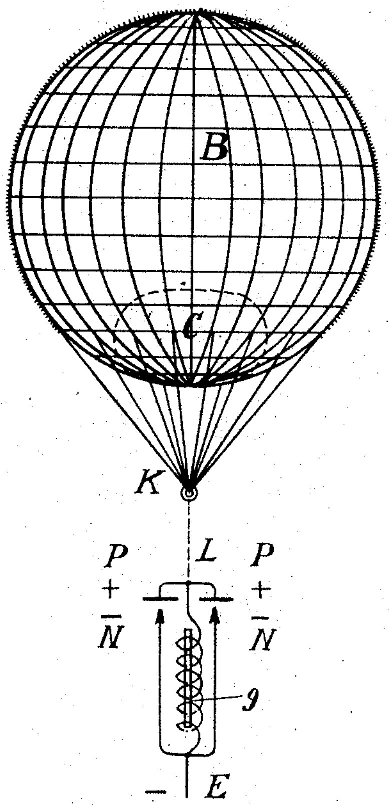
B
C
K
P
+
−
N
L
P
+
−
N
9
−
E

Fig. 29.

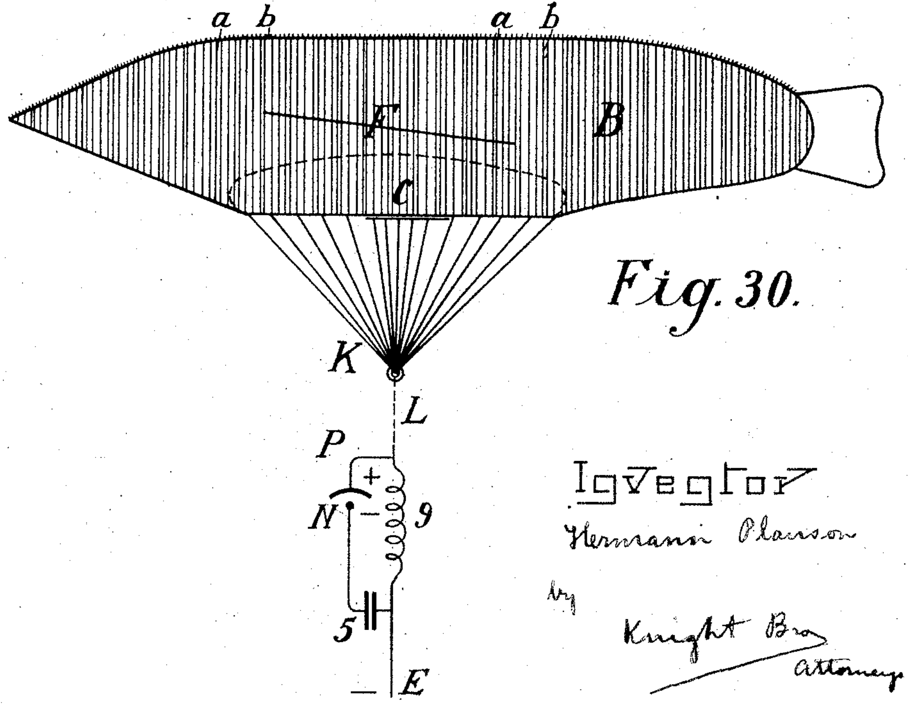
a
b
a
b
F
B
C
K
L
P
+
N
−
9
5
−
E

Fig. 30.

Inventor
Hermann Plauson
by
Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 11

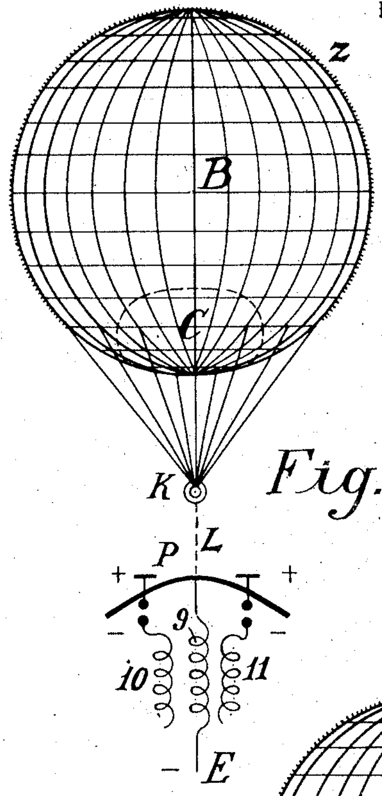
z
B
C
K
P
L
+
+
−
−
9
10
11
−
E

Fig. 31.

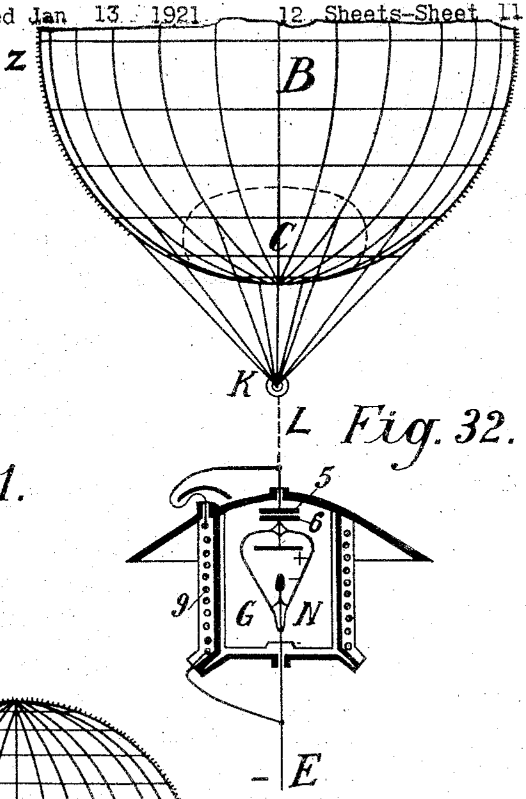
z
B
C
K
L
5
6
+
−
9
G
N
−
E

Fig. 32.

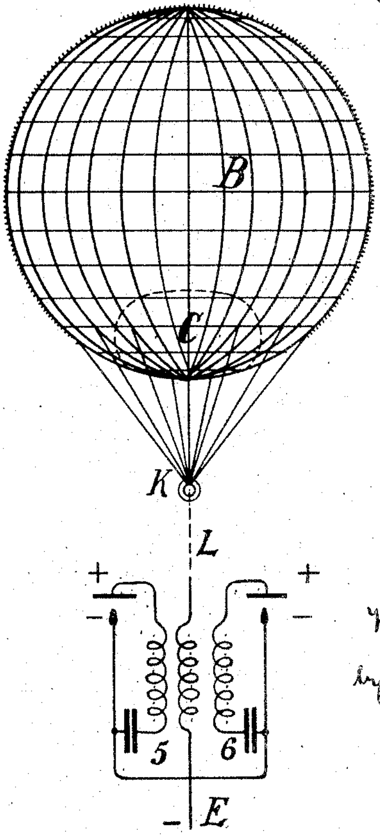

Fig. 33.

Inventor
Hermann Plauson
by Knight Bros
attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 12

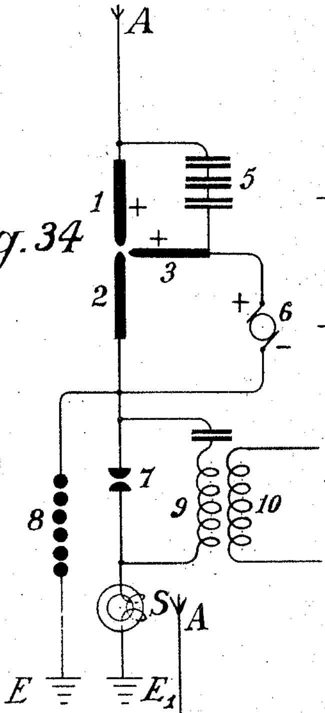
A
1
+
5
+
3
2
+
6
−
8
7
9
10
S
E
E1

Fig. 34

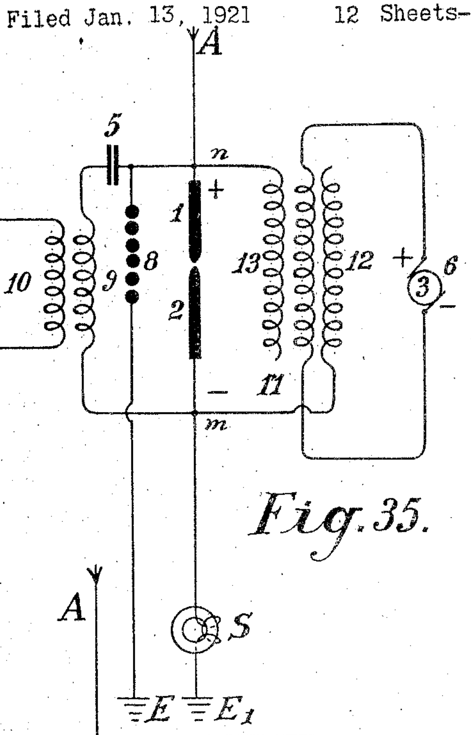
A
5
n
+
1
8
10
9
13
12
+
6
3
−
2
−
11
m
S
E
E1

Fig. 35.

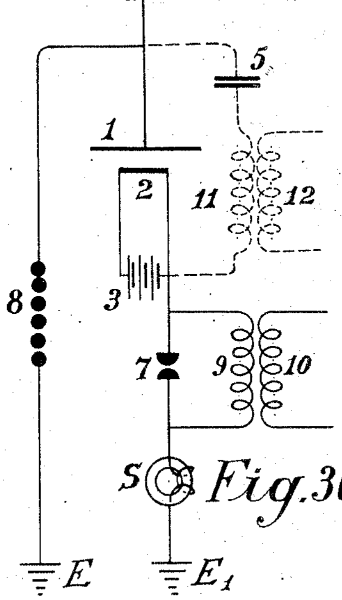
A
5
1
2
11
12
3
7
9
10
8
S
E
E1

Fig. 36.

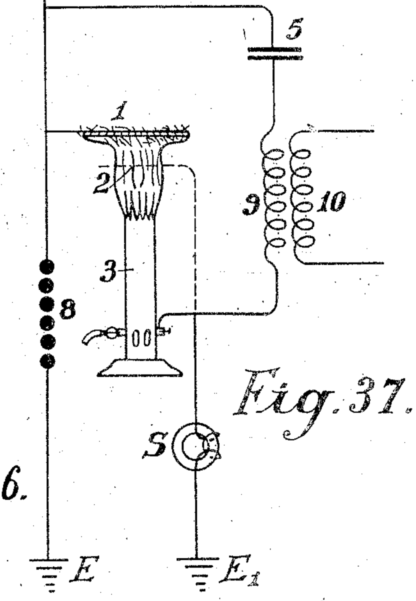
A
5
1
2
9
10
3
8
S
E
E1

Fig. 37.

Inventor
Hermann Plauson

by Knight Bros
attorneys

Patented June 9, 1925. **1,540,998**

UNITED STATES PATENT OFFICE.

HERMANN PLAUSON, OF HAMBURG, GERMANY.

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY.

Application filed January 13, 1921. Serial No. 437,107.

To all whom it may concern:

Be it known that I, HERMANN PLAUSON, Esthonian subject, residing in Hamburg, Germany, have invented certain new and useful Improvements in the Conversion of Atmospheric Electric Energy, of which the following is a specification.

Methods of obtaining atmospheric electricity by means of metallic nettings set with spikes which are held by means of ordinary or anchored kite balloons made of fabrics and filled with hydrogen, are in theory already known. Atmospheric electricity obtained in this way has been suggested to be used in the form of direct current for the charging of accumulators. This knowledge however is at present only theoretical as the conversion in practice has hitherto been a failure. No means are known of protecting the apparatus from destruction by lightning. The balloons used for collecting the charge must also be made of very large size in order to be able to support the weight of the metallic netting and the heavy cable connections.

Instead of using heavy metallic netting as collectors attached to single air balloons of non-conducting materials which are liable to be torn and are permeable to the gas, it is proposed to use metallic balloon collectors which have the following important advantages—

(*a*) The metallic cases are impenetrable to helium and hydrogen; they also represent large metallic weather-proof collecting surfaces.

(*b*) Radio active means and the like may be easily applied internally or externally; whereby the ionization is considerably increased and therewith also the quantity of atmospheric electricity capable of being collected.

(*c*) Such balloon collectors of light metal do not require to be of large size as they have to carry only their own moderate weight, and that of the conducting cable or wire.

(*d*) The entire system therefore offers little surface for the action of storm and wind and is resistant and stable.

(*e*) Each balloon can be easily raised and lowered by means of a winch so that all repairs, recharging and the like can be carried out without danger during the operation.

It is further proposed to use a collecting aerial network of several separate collectors spread out in the air above the earth, which collectors are interconnected by electrical conductors.

According to this invention charges of atmospheric electricity are not directly converted into mechanical energy, and this forms the main difference from previous inventions, but the static electricity which runs to earth through aerial conductors in the form of direct current of very high voltage and low current strength is converted into electro-dynamic energy in the form of high frequency vibrations. Many advantages are thereby obtained and all disadvantages avoided.

The very high voltage of static electricity of a low current strength can be converted by this invention to voltages more suitable for technical purposes and of greater current strength. By the use of closed oscillatory circuits it is possible to obtain electromagnetic waves of various amplitude and thereby to increase the degree of resonance of such current. Such resonance allows various values of inductance to be chosen whereby again the governing of the starting and stopping of machines driven thereby by simply tuning the resonance between coils of the machine and the transformer circuit forming the resonance can easily be obtained. Further, such currents have the property of being directly available for various uses, even without employing them for driving motors, of which there may be particularly mentioned, lighting, production of heat and use in electro-chemistry.

Further, with such currents a series of apparatus may be fed without direct current supply through conductors and also the electro-magnetic high frequency currents may be converted by means of special motors adapted for electro-magnetic oscillations into mechanical energy, or finally converted by special machines into alternating current of low frequency or even into direct current of high potential.

The invention is more particularly described with reference to the accompanying diagrams in which:—

Figure 1 is an explanatory figure.

Figure 2 is a diagrammatic view of the simplest form.

Figure 3 shows a method of converting atmospheric electrical energy for use with motors.

Figure 4 is a diagram showing the use of protective means.

Figure 5 is a diagram of an arrangement for converting large current strengths.

Figure 6 is a diagram of an arrangement including controlling means.

Figure 7 shows means whereby the spark gap length can be adjusted.

Figure 8 shows a unipolar connection for the motor.

Figure 9 shows a weak coupled system suitable for use with small power motors.

Figures 10, 11 and 12 show modified arrangements.

Figure 13 shows a form of inductive coupling for the motor circuit.

Figure 14 is a modified form of Figure 13 with inductive coupling.

Figure 15 is an arrangement with non-inductive motor.

Figure 16 is an arrangement with coupling by condenser.

Figures 17, 18 and 19 are diagrams of further modifications.

Figure 20 shows a simple form in which the aerial network is combined with special collectors.

Figure 21 shows diagrammatically an arrangement suitable for collecting large quantities of energy.

Figure 22 is a modified arrangement having two rings of collectors.

Figure 23 shows the connections for three rings of collectors.

Figure 24 shows a collecting balloon and diagram of its connection of condenser batteries.

Figures 25 and 26 show modified collector balloon arrangements.

Figure 27 shows a second method of connecting conductor for the balloon aerials.

Figure 28 shows an auto-transformer method of connection.

Figure 29 shows the simplest form of construction with incandescent cathode.

Figure 30 shows a form with cigar shaped balloon.

Figure 31 is a modified arrangement.

Figure 32 shows a form with cathode and electrode enclosed in a vacuum chamber.

Figure 33 is a modified form of Figure 32.

Figure 34 shows an arc light collector.

Figure 35 shows such an arrangement for alternating current.

Figure 36 shows an incandescent collector with Nernst lamp.

Figure 37 shows a form with a gas flame.

Figure 1 illustrates a simple diagram for converting static electricity into dynamic energy of a high number of oscillations. For the sake of clearness in the drawings an influence machine is assumed to be employed and not an aerial antenna. 13 and 14 are combs for collecting the static electricity of the influence machine. 7 and 8 are spark discharging electrodes, 6 and 5 condensers, 9 an inductive primary coil, 10 secondary coil, 11 and 12 ends of conductors of the secondary coil 10. When the disc of the static influence machine is rotated by mechanical means, the combs collect the electric charges one the positive and the other the negative, and charge the condensers 5 and 6 until such a high potential is formed across the spark gap 7—8, that the spark gap is jumped. As the spark gap 7—8 forms a closed circuit with condensers 6 and 5, and inductive resistance 9, as is well known, waves of high frequency electromagnetic oscillations will pass in this circuit.

The high frequency of the oscillations produced in the primary circuit induces waves of the same periodicity in the secondary circuit. Thus in the primary circuit electromagnetic oscillations are formed by the passage of the spark over the spark gap and these waves are maintained by fresh charges of static electricity.

By suitably selecting the ratio between the number of the coils in the primary and secondary circuits with regard to a correct application of the co-efficients of resonance (capacity, inductance, and resistance) the high voltage of the primary circuit may be suitably converted into low voltage and high current strength.

When the oscillatory discharges in the primary circuit becomes weaker or entirely cease, the condensers are charged again by the static electricity until the accumulated charge again breaks down the spark gap. All this is repeated as long as electricity is produced by the static machine by employing mechanical energy.

An elementary form of the invention is shown in Figure 2 in which two spark gaps in parallel are used one of which may be termed the working gap 7 in Figure 2, whilst the second serves as a safety device for excess voltage and consists of a larger number of spark gaps than the working section, which gaps are arranged in series and are bridged by very small capacities as is illustrated in a_1, b_1, c_1, Figure 2 which allow of uniform sparking in the safety section.

In Figure 2 A is the aerial antenna for collecting charges of atmospheric electricity, 13 is the earth connection of the second part of the spark gap, 5 and 6 are condensers, 9 a primary coil. Now when through the aerial A the positive atmospheric electricity seeks to combine with the negative charge to earth, this is prevented by (the air gap between) the spark gaps. The resistance of the spark gap 7 is, as shown in the drawings, lower than that of the other safety section which consists of three spark gaps connected in series, and consequently a three times greater air resistance is offered by the latter.

So long therefore, as the resistance of the spark gap 7 is not overloaded, so that the other spark gaps have an equal resistance with it the discharges take place only over spark gap 7. Should however the voltage be increased by any influences so that it might be dangerous for charging the condensers 5 and 6 or for the coil insulation 9 and 10 in consequence of break down, by a correct regulation of this spark gap the second spark gap can discharge free from inductive effects direct to earth without endangering the machine.

Without this second spark gap, arranged in parallel having a higher resistance than the working spark gap it is impossible to collect and render available large quantities of electrical energy.

The action of this closed oscillation circuit consisting of spark gap 7, two condensers 5 and 6, primary coil 9, and also secondary coil 10 is exactly the same as the one described in Figure 1 with the arrangement of the static induction machine with the only difference that here the second spark gap is provided. The electromagnetic high frequency alternating current obtained can be tapped off from the conductors 11 and 12 for lighting and heating purposes. Special kinds of motors adapted for working with these peculiar electrical charges may be connected at 14 and 15 which can work with static electricity charges or with high frequency oscillations.

In addition to the use of spark gaps in parallel a second measure of security is also necessary for taking off the current. This precaution consists according to this invention, in the introduction of and method of connecting certain protective electromagnets or choking coils in the aerial circuit as shown by S in Figure 3.

A single electromagnet only having a core of the thinnest possible separate laminations is connected with the aerial.

In the case of high voltages in the aerial network or at places where there are frequent thunder storms, several such magnets may however be connected in series.

In the case of large units or plants several electromagnets can be employed in parallel or in series parallel.

The windings of these electromagnets may be simply connected in series with the aerials. In this case the winding preferably consists of several thin parallel wires, which make up together, the necessary section.

The winding may be made of primary and secondary windings in the form of a transformer. The primary winding will be then connected in series with the aerial network, and the secondary winding more or less short-circuited over a regulating resistance or an induction coil. In the latter case it is possible to regulate to a certain extent the effect of the choking coils. In the further description of the connecting and constructional diagrams the aerial electromagnet choke coil is indicated by a simple ring S.

Figure 3 shows the simplest way of converting atmospheric electricity into electromagnetic wave energy by the use of special motors adapted for high oscillatory currents or static charges of electrical energy. Recent improvements in motors for working with static charges and motors working by resonance, that is to say, having groups of tuned electromagnetic cooperating circuits render this possible but such do not form part of the present invention.

A motor adapted to operate with static charges will for the sake of simplicity be diagrammatically indicated by two semicircles 1 and 2 and the rotor of the motor by a ring M. (Figure 3.) A is a vertical aerial or aerial network. S the safety choke or electromagnet with coil O as may be seen is connected with the aerial A. Adjacent the electromagnet S the aerial conductor is divided into three circuits, the circuit 8 giving the safety spark gap, the circuit 7 with the working spark gap, and then a circuit including the stator terminal 1, the rotor and stator terminal 2 at which a connection is made to the earth wire. The two spark gaps are also connected metallically with the earth wire. The method of working these diagrams is as follows:

The positive atmospheric electric charge collected tends to combine with the negative electricity (or earth electricity) connected with the earth wire. It travels along the aerial A through the electromagnet S without being checked as it flows in the same direction as the direct current. Further, its progress is arrested by two sparks gaps placed in the way and the stator condenser surfaces. The stator condenser surfaces are charged until the charge is greater than the resistance of the spark gap 7, whereupon a spark springs over the spark gap 7 and an oscillatory charge is obtained as by means of the motor M, stator surfaces 1 and 2, and spark gap 7, a closed oscillation circuit is obtained for producing the electromagnetic oscillations. The motor here forms the capacity and the necessary inductance and resistance, which, as is well known, are necessary for converting static electricity into electromagnetic wave energy.

The discharges formed are converted into mechanical energy in special motors and can not reach the aerial network by reason of the electromagnet or choke. If, however, when a spark springs over the spark gap 7 a greater quantity of atmospheric electricity tends to flow to earth, a counter voltage is induced in the electromagnet, which is greater the more rapidly and strongly the flow of current direct to the earth is. By the formation of this opposing voltage a sufficiently high resistance is offered to the flow of atmospheric electricity direct to earth to prevent a short circuit with the earth.

The circuit containing spark gap 8 having a different wave length which is not in resonance with the natural frequency of the motor, does not endanger the motor and serves as security against excess voltage, which, as practical experiments have shown, may still arise in certain cases, but can be conducted direct to earth through this spark gap.

In the diagram illustrated in Figure 4 the spark gap 7 is shunted across condensers 5 and 6 from the motor M. This construction affords mainly a better insulation of the motor against excess voltage and a uniform excitation through the spark gap 7.

In Figure 5 a diagram is illustrated for transforming large current strengths which may be employed direct without motors, for example, for lighting or heating purposes. The main difference is that here the spark gap consists of a star shaped disc 7 which can rotate on its own axis and is rotated by a motor opposite similarly fitted electrodes 7^a. When separate points of stars face one another, discharges take place, thus forming an oscillation circuit over condensers 5 and 6 and inductance 9 for oscillatory discharges. It is evident that a motor may also be directly connected to the ends of the spiral 9.

The construction of the diagram shown in Figure 6 permits of the oscillation circuit of the motor being connected with an induction coil. Here a regulating inductive resistance is introduced for counter-acting excess voltages in the motor. By cutting the separate coils 9 (coupled inductively to the aerial) in or out the inductive action on the motor may be more or less increased or variable aerial action may be exerted on the oscillation circuit.

In Figure 7 the oscillation circuit is closed through the earth (E and E_1). The spark gap 7 may be prolonged or shortened by more or fewer spark gaps being successively connected by means of a contact arm 7^b.

Diagram 8 shows a unipolar connection of the motor with the aerial network. Here two oscillation circuits are closed through the same motor. The first oscillation circuit passes from aerial A through electromagnet S, point x, inductance 9^a to the earth condenser 6 and further, over spark gap 7 to the aerial condenser 5 and back to x. The second oscillation circuit starts from the aerial condenser 5 at the point x^1 over the inductance 9 to the earth condenser 6 at the point x^3 and through the condenser 6 over the spark gap 7 back to x^1. The motor itself is inserted between the two points of the spark gap 7. From this arrangement slightly damped oscillation wave currents are produced.

In the diagram illustrated in Figure 9 a loosely coupled system of connections is illustrated which is assumed to be for small motors for measuring purposes. A indicates the aerial conductor, S the electromagnet in the aerial conductor, 9 the inductance, 7 the spark gap, 5 and 6 condensers, E the earth, M the motor, and 1 and 2 stator connections of the motor. The motor is directly metallically connected with the oscillation circuit.

In Figure 10 a purely inductive coupling is employed for the motor circuit. The motor is connected with the secondary wire 10 as may be seen in Figure 11 in a somewhat modified diagram connection. The same applies to the diagram of Figure 12.

The diagrams hitherto described preferably allow of motors of small and medium strength to be operated. For large aggregates, however, they are too inconvenient as the construction of two or more oscillation circuits for large amounts of energy is difficult; the governing is still more difficult and the danger in switching on or off is greater.

A means of overcoming such difficulties is shown in Figure 13. The oscillation circuit here runs starting from the point x over condenser 5, variable inductance 9, spark gap 7 and the two segments (3^a and 4^a) forming arms of a Wheatstone bridge, back to x. If the motor is connected by brushes 3 and 4 transversely to the two arms of the bridge as shown in the drawings, electromagnetic oscillations of equal sign are induced in the stator surfaces 1 and 2 and the motor does not revolve. If however, the brushes 3 and 4 are moved in common with the conducting wires 1 and 2 which connect the brushes with the stator poles a certain alteration or displacement of the polarity is obtained and the motor commences to revolve.

The maximum action will result if one brush 3 comes on the central sparking contact 7 and the other brush 4 on the part x. They are however, usually in practice not brought on to the central contact 7 but only held in the path of the bridge segments 4^a and 3^a in order not to connect the spark gaps with the motor oscillation circuit.

As however, the entire oscillation energy can thereby not act on the motor it is better to carry out the same system according to the diagram 14. The diagram 14 differs from the foregoing only by the motor not being directly metallically connected with the segments of the commutator, but only a primary coil 9 which induces in a secondary coil 10, current which feeds the motor M and takes the place of the rotor. By this arrangement a good transforming action is obtained, a loose coupling and also an oscillation circuit without a spark gap.

In Figure 15 the motor is not purely inductively as in 14, but directly metallically branched off from the primary coil (at x and x^1) after the principle of the auto-transformer.

In Figure 16 instead of an inductance a condenser 6 is in similar manner, and for the same object inserted between the segments 3^a and 4^a. This has the advantage that the segments 3^a and 4^a need not be made of solid metal but may consist of spiral coils whereby a more exact regulation is possible and further motors of high inductance may be employed.

The arrangements of Figures 17, 18 and 19 may be employed for use with resonance and particularly with induction condenser motors; between the large stator induction condenser surfaces, small reversing pole condensers are connected, which, as may be seen from Figures 17, 18 and 19 are led together to earth. Such reversing poles have the advantage that with large quantities of electrical energy the spark formation between the separate oscillation circuits ceases.

Figure 19 shows a further method which prevents electromagnetic oscillations of high number of alternations formed in the oscillation circuit striking back to the aerial conductor. It is based on the well known principle that a mercury lamp, one electrode of which is formed of mercury, the other of solid metal such as steel allows an electric charge to pass in only one direction from the mercury to the steel and not vice versa. The mercury electrode of the vacuum tube N is therefore connected with the aerial conductor and the steel electrode with the oscillation circuit. From this it results that charges can pass only from the aerial through the vacuum tube to the oscillation circuit, but not vice versa. Oscillations which are formed on being transformed in the oscillation circuit cannot pass to the aerial conductor.

In practice these vacuum tubes must be connected behind an electromagnet as the latter alone affords no protection against the danger of lightning.

As regards the use of spark gaps, all arrangements as used for wireless telegraphy may be used. Of course the spark gaps in large machines must have a sufficiently large surface. In very large stations they are cooled in liquid carbonic acid or better still in liquid nitrogen or hydrogen; in most cases the cooling may also take place by means of liquefied low homologues of the metal series or by means of hydrocarbons the freezing point of which lies at between −90° C. and −40° C. The spark gap casing must also be insulated and be of sufficient strength to be able to resist any pressure which may arise. Any undesirable excess super-pressure which may be formed must be automatically let off. I have employed with very good results mercury electrodes which were frozen in liquid carbonic acid, the cooling being maintained during the operation from the outside through the walls.

Figure 20 is one of the simplest forms of construction of an aerial network in combination with collectors, transformers and the like illustrated diagrammatically. E is here the earth wire, 8 the safety spark gap, 7 the working spark gap, 1 and 2 the stator surfaces of the motor, 5 a condenser battery, S the protective magnet which is connected with the coil in the aerial conductor, A^1 to A^{10} aerial antennæ with collecting balloons, N horizontal collecting or connecting wires from which, to the centre a number of connections run.

The actual collectors consist of metal sheaths preferably made of an aluminium magnesium alloy, and are filled with hydrogen or helium and are attached to copper plated steel wires. The size of the balloon is selected so that the actual weight of the balloon and the weight of the conducting wire is supported thereby. On the top of the balloon aluminium spikes, made and gilded in a special manner hereinafter described, are arranged in order to produce a conductor action. Small quantities of radium preparations, more particularly polonium-ionium or mesothorium preparations considerably increase the ionization, and therewith the action of these collectors.

In addition to metal balloons, fabric balloons which are superficially metal coated according to Schoop's metal spraying process, may however also be employed. A metallic surface may also be produced by lacquering with metallic bronzes, preferably according to Schoop's spraying process or lacquering with metallic bronze powders in two electrical series of widely different metals, because thereby the collecting effect is considerably increased.

Instead of the ordinary round balloons, elongated cigar shaped ones may be employed. In order also to utilize the frictional energy of the wind, patches or strips of non-conducting substances which produce electricity by friction, may be attached to the metallized balloon surfaces. The wind will impart a portion of its energy in the form of frictional electricity, to the balloon casing, and thereby the collecting effect is substantially increased.

In practice however, very high towers (up to 300 metres is fully admissible) may be employed as antennæ. In these towers copper tubes rise freely further above the top of the tower. A gas lamp secured against the wind is then lit at the point of the copper tube and a netting is secured to the copper tube over the flame of this lamp to form a collector. The gas is conveyed through the interior of the tube up to the summit. The copper tube must be absolutely protected from moisture at the place at which it enters the tower and also rain must be prevented running down the walls of the tower which might lead to a bad catastrophe. This is done by bell shaped enlargements which expand downwards, being arranged in the tower in the form of high voltage insulators of Siamese pagodas.

Special attention must be devoted to the foundations of such towers. They must be well insulated from the ground, which may be obtained by first embedding a layer of concrete in a box form to a sufficient depth in the ground and inserting in this an asphalt lining and then glass bricks cast about 1 or 2 metres in thickness. Over this in turn there is a ferro-concrete layer in which alone the metal foot of the tube is secured. This concrete block must be at least 2 metres from the ground and be fully protected at the sides by a wooden covering, from moisture. In the lower part of the tower a wood or glass house for the large condenser batteries or for the motors may be constructed. In order to lead the earth connection to the ground water, a well insulated pit constructed of vitreous bricks, must be provided. Several such towers are erected at equal distances apart and connected with a horizontal conductor. The horizontal connecting wires may either run directly from tower to tower or be carried on bell shaped insulators similar to those in use for high voltage conductors. The width of the network may be of any suitable size and the connection of the motors can take place at any suitable places.

In order to collect large quantities of electricity with few aerials it is well to provide the aerial conductor with batteries of condensers as shown in two methods of construction in Figures 21 and 22. In Figure 21 the batteries of condensers 5 are connected on the one hand with the aerial electricity collectors Z by the aerial conductor A, and on the other hand interconnected in series with an annular conductor from which horizontal conductors run to the connecting points C to which the earth wire is connected.

Figure 22 shows a similar arrangement. Should two such series of antennæ rings be shown by a voltmeter to have a large difference of potential (for example, one in the mountains and one in the plain) or even of different polarity these differences may be compensated for by connecting sufficiently large condenser batteries (5, 5^a, 5^b) by means of Maji star conductors D and D^1. In Figure 23 a connection of three such rings of collectors to form a triangle with a central condenser battery is illustrated.

The condenser batteries of such large installations must be embedded in liquefied gases or in liquids freezing at very low temperatures. In such cases a portion of the atmospheric energy must be employed for liquefying these gases. It is also preferable to employ pressure. By this means the condenser surfaces may be diminished, and still allow for large quantities of energy to be stored, secure against breakdown. For smaller installations the immersing of the condensers in well insulated oil or the like, suffices. Solid substances on the other hand cannot be employed as insulators.

The arrangement in the diagrams hitherto described was always such that the condenser batteries were connected with both poles directly to the aerial conductors. An improved diagram of the connections for obtaining atmospheric electricity for the condenser batteries has however, been found to be very advantageous, this arrangement consists in that they are connected by only one pole (unipolar) to the collecting network. Such a method of arrangement is very important, as by means of it a constant current and an increase of the normal working pressure or voltage is obtained. If for example a collecting balloon aerial which is allowed to rise to a height of 300 metres, shows 40,000 volts above earth voltage, in practice it has been found that the working voltage (with a withdrawal of the power according to the method hereinbefore described by means of oscillating spark gaps and the like) is only about 400 volts. If however, the capacity of the condenser surfaces be increased, which capacity in the above mentioned case was equal to that of the collecting surface of the balloon aerials, to double the amount, by connecting the condenser batteries with only one pole, the voltage rises under an equal withdrawal of current up to and beyond 500 volts. This can only be ascribed to the favourable action of the connecting method.

In addition to this substantial improvement it has also been found preferable to insert double inductances with electromagnets and to place the capacities preferably between two such electromagnets. It has also been found that the useful action of such condensers can be further increased if an induction coil be connected as inductive resistance to the unconnected pole of the condenser, or still better if the condenser itself be made as an induction condenser. Such a condenser may be compared with a spring which when compressed carries in itself accumulated force, which it again gives off when released. In charging, a charge with reversed sign is formed at the other free condenser pole, and if through the spark gap a short circuit results, the accumulated energy is again given back since now new quantities of energy are induced at the condenser pole connected with the conductor network, which in fact charges with opposite signs to that at the free condenser pole. The new induced charges have of course the same sign as the collector network. The whole voltage energy in the aerial is thereby however increased. In the same space of time larger quantities of energy are accumulated than is the case without such inserted condenser batteries.

In Figures 24 and 25 two different diagrams of connections are more exactly illustrated, Figure 24 shows a collecting balloon and the diagram of the connections to earth. Figure 25 four collecting balloons and the parallel connection of the condenser batteries belonging thereto.

A is the collecting balloon made of an aluminium magnesium alloy (electron metal, magnalium) of a specific gravity of 1.8 and a thickness of plate 0.1 to 0.2 mm. Inside there are eight strong vertical ribs of T shaped section about 10 to 20 mm. in height and about 3 mm. in thickness with the projecting part directed inwards (indicated by *a*, *b*, *c*, *d* and so forth); they are riveted together to form a firm skeleton and are stiffened in a horizontal direction by two cross ribs. The ribs are further connected with one another internally and transversely by means of thin steel wires, whereby the balloon obtains great power of resistance and elasticity. Rolled plates of 0.1 to 0.2 mm. in thickness made of magnalium alloy are then either soldered or riveted on this skeleton so that a fully metallic casing with smooth external surface is obtained. Well silvered or coppered aluminium plated steel wires run from each rib to the fastening ring 2. Further, the coppered steel hawser L preferably twisted out of separate thin wires (shown in dotted lines in Figure 24) and which must be long enough to allow the balloon to rise in the desired height, leads to a metal roller or pulley 3 and from thence to a winch W, well insulated from the earth. By means of this winch, the balloon, which is filled with hydrogen, or helium, can be allowed to rise to a suitable height (300 to 5,000 metres) and brought to the ground for recharging or repairs.

The actual current is taken directly through a friction contact from the metal roller 3 or from the wire, or even from the winch or simultaneously from all three by means of brushes (3, 3^a and 3^b). Beyond the brushes the conductor is divided, the paths being:—firstly over 12 to the safety spark gap 8, from thence to the earth conductor E^1, and secondly over electromagnet S^1, point 13, to a second loose electromagnet having an adjustable coil S^2, then to the spark gap 7 and to the second earth conductor E^2. The actual working circuit is formed through the spark gap 7, condensers 5 and 6, and through the primary coil 9; here the static electricity formed by oscillatory discharges is accumulated and converted into high frequency electromagnetic oscillations. Between the electromagnets S^1 and S^2 at the crossing point 13, four condenser batteries are introduced which are only indicated diagrammatically in the drawings each by one condenser. Two of these batteries (16 and 18) are made as plate condensers and prolonged by regulating induction coils or spirals 17 and 19 while the two others (21 and 23) are induction condensers. As may be seen from the drawings each of the four condenser batteries 16, 18, 21, 23 is connected only by one pole to the aerial or to the collector conductor. The second poles 17, 19, 22, 24 are open. In the case of plate condensers having no inductive resistance an induction coil is inserted. The object of such a spiral or coil is the displacement of phase of the induction current by ¼ periods, whilst the charging current of the condenser poles which lie free in the air, works back to the collector aerial. The consequence of this is that in discharges in the collector aerial the back inductive action of the free poles allows a higher voltage to be maintained in the aerial collecting conductor than would otherwise be the case. It has also been found that such a back action has an extremely favourable effect on the wear of the contacts. Of course the inductive effect may be regulated at will within the limits of the size of the induction coil, the length of the coil in action being adjustable by means of wire connection without induction (see Fig. 24, No. 20).

S^1 and S^2 may also be provided with such regulating devices in the case of S^2 (illustrated by 11). If excess voltage be formed it is conducted to earth through the wire 12 and spark gap 8 or through any other suitable apparatus, since this formation would be dangerous for the other apparatus.

The action of these condenser batteries has already been hereinbefore described.

The small circles on the collector balloon indicate places at which zinc amalgam or gold amalgam or other photoelectric acting metals in the form of small patches in extremely thin layers (.01 to .05 mm. in thickness) are applied to the balloon casing of light metal. Such metallic patches may also be applied to the entire balloon as well as in greater thickness to the conducting network. The capacity of the collector is thereby considerably strengthened at the surface. The greatest possible effect in collecting may be obtained by polonium amalgams and the like. On the surface of the collector balloon metal points or spikes are also fixed along the ribs, which spikes serve particularly for collecting the collector charge. Since it is well known that the resistance of the spikes is less the sharper the spike is, for this purpose it is therefore extremely important to employ as sharp spikes as possible. Experiments made as regards these have shown that the formation of the body of the spike or point also plays a large part. for example, spikes made of bars or rollers with smooth surfaces, have a many times greater point resistance as collector accumulator spikes than those with rough surfaces. Various kinds of spike bodies have been experimented with for the collector balloons hereinbefore mentioned. The best results were given by spikes which were made in the following way. Fine points made of steel, copper, nickel, or copper and nickel alloys, were fastened together in bundles and then placed as anode with the points in a suitable electrolyte (preferably in hydrochloric acid or muriate of iron solutions) and so treated with weak current at 2 to 3 volts pressure. After 2 to 3 hours according to the thickness of the spikes or pins the points become extremely sharp and the bodies of the spikes have a rough surface. The bundle can then be removed and the acid washed off with water. The spikes are then placed as cathode in a bath consisting of solution of gold, platinum, iridium, paladium or wolfram salts or their compounds and coated at the cathode galvanically with a thin layer of precious metal, which must however be sufficiently firm to protect them from atmospheric oxidation.

Such spikes act at a 20 fold lower voltage almost as well as the best and finest points made by mechanical means. Still better results are obtained if polonium or radium salts are added to the galvanic bath when forming the protective layer or coating. Such pins have a low resistance at their points and even at one volt and still lower pressures have an excellent collector action.

In Figure 24 the three unconnected poles are not connected with one another in parallel. That is quite possible in practice without altering the principle of the free pole. It is also preferable to interconnect in parallel to a common collector network, a series of collecting aerials.

Figure 25 shows a diagram for such an installation. A^1, A^2, A^3, A^4 are four metal collector balloons with gold or platinum coated spikes which are electrolytically made in the presence of polonium emanations or radium salts, which spikes or needles are connected over four electro-magnets S^1, S^2, S^3, S^4, through an annular conductor R. From this annular conductor four wires run over four further electromagnets S^a, S^b, S^c, S^d, to the connecting point 13. There the conductor is divided, one branch passing over 12 and the safety spark gap 8 to the earth at E^1, the other over inductive resistance J and working spark gap 7 to the earth at E^2. The working circuit, consisting of the condenser 5 and 6 and a resonance motor or a condenser motor M, such as hereinbefore described, is connected in proximity round the sparking gap section 7.

Instead of directly connecting the condenser motor of course the primary circuit for high frequency oscillatory current may also be inserted.

The condenser batteries are connected by one pole to the annular conductor R and can be either inductionless (16 and 18) or made as induction condensers as shown by 21 and 23. The free poles of the inductionless condensers are indicated by 17 and 19, those of the induction condensers by 22 and 24. As may be seen from the drawings all these poles 17, 22, 19, 24 may be interconnected in parallel through a second annular conductor without any fear that thereby the principle of the free pole connection will be injured. In addition to the advantages already set forth the parallel connection also allows of an equalization of the working pressure in the entire collector network. Suitably constructed and calculated induction coils 25 and 26 may also be inserted in the annular conductor of the free poles, by means of which a circuit may be formed in the secondary coils 27 and 28 which allows current produced in this annular conductor by fluctuations of the charges or the like appearances to be measured or otherwise utilized.

According to what has been hereinbefore stated separate collector balloons may be connected at equidistant stations distributed over the entire country, either connected directly with one another metallically or by means of intermediate suitably connected condenser batteries through high voltage conductors insulated from earth. The static electricity is converted through a spark gap into dynamic energy of a high number of oscillations and may in such form be coupled as a source of energy by means of a suitable method of connecting, various precautions being observed and with special regulations. The wires leading from the collector balloons have hitherto been connected through an annular conductor without this endless connection, which can be regarded as an endless induction coil, being able to exert any action on the whole conductor system.

It has now been found that if the network conductor connecting the aerial collector balloons with one another is not made as a simple annular conductor, but preferably short circuited in the form of coils over a condenser battery or spark gap or through thermionic tubes or valves or audions, then the total collecting network exhibits quite new properties. The collection of atmospheric electricity is thereby not only increased but an alternating field may be easily produced in the collector network. Further, the atmospheric electrical forces showing themselves in the higher regions may also be directly obtained by induction. In Figures 26 and 28 a form of construction is shown on the basis of which the further foundations of the method will be more particularly explained.

In Figure 26 1, 2, 3, 4 are metal collector balloons, 5, 6, 7, 8 their metallic aerial conductors and I the actual collector network. This consists of five coils and is mounted on high voltage insulators in the air, on high voltage masts (or with a suitable construction of cable embedded in the earth). One coil has a diameter of 1 to 100 km. or more. S and S^1 are two protective electromagnets, F the second safety section against excess voltage, E its earth conductor and E^1 the earth conductor of the working section. When an absorption of static atmospheric electricity is effected through the four balloon collectors, the current in order to reach the earth connection E^1 must flow spirally through the collector network over the electromagnet S, primary induction coil 9, conductor 14, anode A of the audion tube, incandescent cathode K, as the way over the electromagnet and safety spark gap F offers considerably greater resistance. Owing to the fact that the accumulated current flows in one direction, an electromagnetic alternating field is produced in the interior of the collector network coil, whereby the whole free electrons are directed more or less into the interior of the coil. An increased ionization of the atmosphere is therefore produced. In consequence of this the points mounted on the collector balloon show a considerably reduced resistance and therefore increased static charges between the points on the balloon and the surrounding atmosphere are produced. The result of this is a considerably increased collector effect.

A second effect which could not be obtained otherwise is obtained by the electromagnetic alternating field which running parallel to the earth surface, acts more or less with a diminishing or increasing effect on the earth magnetic field, whereby in the case of fluctuations in the current a return induction current of reversed sign is always produced in the collector coil by earth magnetism. Now if, however, a constantly pulsating continuous alternating field is produced as stated in the above collector network I, an alternating current of the same periodicity is produced also in the collecting network coil. As the same alternating field is further transmitted to the aerial balloon, the resistance of its points is thereby considerably reduced, whilst the collector action is considerably increased. A further advantage is that positive electrons which collect on the metal surfaces during the conversion into dynamic current produce a so-called drop of potential of the collector area. As an alternating field is present, the negative ions surrounding the collector surfaces, when discharge of the collector surfaces takes place produce by the law of induction, an induction of reversed sign on the collector surface and so forth (that is to say again a positive charge). In addition to the advantages hereinbefore set forth, the construction of connecting conductors in coil form when of sufficiently large diameter, allows of a utilization of energy arising in higher regions also in the simplest way. As is well known electric discharges frequently take place at very great elevations which may be observed, such as St. Elmo's fires or northern lights. These energy quantities have not been able to be utilized up to now. By this invention all these kinds of energy, as they are of an electromagnetic nature and the direction of the axis of the collector coils stands at right angles to the earth's surface, can be more or less absorbed in the same way as a receiver in wireless telegraphy absorbs waves coming from a far distance. With a large diameter of the spiral it is possible to connect large surfaces and thereby to take up also large quantities of energy.

It is well known that large wireless stations in the summer months, and also in the tropics are very frequently unable to receive the signals in consequence of interruptions which are caused by atmospheric electricity, and this takes place with vertical coils of only 40 to 100 m. diameter. If on the contrary horizontal coils of 1 to 100 km. diameter be employed very strong currents may be obtained through discharges which are constantly taking place in the atmosphere. Particularly in the tropics or still better in the polar regions where the northern lights are constantly present, large quantities of energy may probably be obtained in this way. A coil with several windings should act the best. In similar manner any alteration of the earth magnetism should act inductively on such a coil.

It is not at all unlikely that earthquakes and spots on the sun will also produce an induction in such collector coils of sufficient size. In similar manner this collector conductor will react on earth currents more particularly when they are near the surface of the earth or even embedded in the earth. By combining the previous kind of current collectors so far as they are adapted for the improved system with the improved possibilities of obtaining current the quantities of free natural energy which are to be obtained in the form of electricity are considerably increased.

In order to produce in the improved collector coil uniform current oscillations of an undamped nature so-called audion high vacuum or thermionic tubes of suitable connection are employed instead of the previously known spark gaps (Fig. 26, Nos. 9–18). The main aerial current flows through electromagnet S (which in the case of a high number of alternations is not connected here but in the earth conductor E^1) and may be conveyed over the primary coils in the induction winding through wire 14 to the anode A of the high vacuum grid tube. Parallel with the induction resistance 9 a regulating capacity of suitable size, such as condenser 11 is inserted. In the lower part of the vacuum grid tube is arranged the incandescent filament or the cathode K which is fed through a battery B. From the battery B two branches run, one to the earth conductor E^1 and the other through battery B^1 and secondary coil 10 to the grid anode *g* in the vacuum tube. By the method of connections shown in dotted lines, a desired voltage at the grid electrode *g* may also be produced through the wire 17 which is branched off from the main current conductor through switches 16 and some small condensers (*a*, *b*, *c*, *d*) connected in series, and conductor 18, without the battery B^1 being required.

The action of the entire system is somewhat as follows:—

On the connecting conductor of the aerial collector network being short circuited to earth, the condenser pole 11 is charged and slightly damped oscillations are formed in the short circuited existing oscillation circuit formed of the condenser 11 and self inductance 9. In consequence of the coupling through coil 10, fluctuations of voltage take place in the grid circuit 15 with the same frequency, which, fluctuations in turn influence the strength of the electrode current passing through the high vacuum amplyfying tube and thus produce current fluctuations of the same frequency in the anode circuit. A permanent supply of energy to the oscillation circuits 9 and 10 consequently takes place, until a condition of balance is set up, in which the consumed oscillation energy is equal to that absorbed. Thereby constant undamped oscillations are now produced in the oscillation circuits 9–11.

For regular working of such oscillation producers high vacuum strengthening tubes are necessary and it is also necessary that the grid and anode voltages shall have a phase difference of 180° so that if the grid is negatively charged, then the anode is positively charged and vice versa. This necessary difference of phase may be obtained by most varied connections, for example, by placing the oscillation circuit in the grid circuit or by separating the oscillation circuit and inductive coupling from the anodes and the grid circuit and so forth.

A second important factor in this way of converting static atmospheric electricity into undamped oscillations is that care must be taken that the grid and anode voltages have a certain relation to one another; the latter may be obtained by altering the coupling and a suitable selection of the self induction in the grid circuit, or as shown by dotted lines 18, 17, 16 by means of a larger or smaller number of condensers of suitable size connected in series; in this case the battery B^1 may be omitted. With a suitable selection of the grid potential a glow discharge takes place between the grid *g* and the anode A, and accordingly at the grid there is a cathode drop and a dark space is formed. The size of this cathode drop is influenced by the ions which are emitted in the lower space in consequence of shock ionization of the incandescent cathodes K and pass through the grid in the upper space. On the other hand the number of the ions passing through the grid is dependent on the voltage between the grid and the cathode. Thus if the grid voltage undergos periodic fluctuations (as in the present case) the amount of the cathode drop at the grid fluctuates and consequently the internal resistance of the tube correspondingly fluctuates, so that when a back coupling of the feed circuit with the grid circuit takes place, the necessary means are afforded for producing undamped oscillations and of taking current, according to requirements from the collecting conductor.

The frequency of the undamped oscillations produced is with a suitably loose coupling equal to the self frequency of the oscillation circuits 9 and 10. By a suitable selection of the self induction of the coil 9 and capacity 11 it is possible to extend from frequencies which produce electromagnetic oscillations of only a few metres wave length down to the lowest practical alternating current frequency. For large installations a suitable number of frequency producing tubes in the form of the well known high vacuum transmission tubes of .5 to 2 kw. in size may be connected in parallel so that in this respect no difficulty exists.

The use of such tubes for producing undamped oscillations, and also the construction and method of inserting such transmission tubes in an accumulator or dynamo circuit is known and also that such oscillation producing tubes only work well at voltages of 1,000 up to 4,000 volts, so that on the contrary their use at lower voltages is considerably more difficult. By the use of high voltage static electricity this method of producing undamped oscillations as compared with that through spark gaps must be regarded as an ideal solution particularly for small installations of outputs of from 1 to 100 kw.

By the application of safety spark gaps, with interpolation of electro-magnets, not only is short circuiting avoided but also the taking up of current is regulated. Oscillation producers inserted in the above way form a constantly acting electromagnetic alternating field in the collector coil, whereby as already stated, a considerable accumulating effect takes place. The withdrawal wire or working wire is connected at 12 and 13, but current may be taken by means of a secondary coil which is firmly or movably mounted in any suitable way inside the large collector coil, i. e. in its electromagnetic alternating field, so long as the direction of its axis runs parallel with that of the main current collecting coil.

In producing undamped oscillations of a high frequency (50,000 per second and more) in the oscillation circuits 9 and 11, electromagnets S and S^1 must be inserted if the high frequency oscillations are not to penetrate the collector coil, between the oscillation producers and the collector coil. In all other cases they are connected shortly before the earthing (as in Figs. 27 and 28).

In Figure 27 a second method of construction of the connecting conductor of the balloon aerials is illustrated in the form of a coil. The main difference consists in that in addition to the connecting conductor I another annular conductor II is inserted parallel to the former on the high voltage masts in the air (or embedded as a cable in the earth) but both in the form of a coil. The connecting wire of the balloon aerials is indicated as a primary conductor and also as a current producing network; the other is the consumption network and is not in unipolar connection with the current producing network.

In Figure 27 the current producing network I is shown with three balloon collectors 1, 2, 3 and aerial conductors 4, 5, 6; it is short circuited through condenser 19 and inductance 9. The oscillation forming circuit consists in this diagram of spark gap *f*, inductance 10, and condenser 11; the earth wire E, is connected to earth over electromagnet S^1. F is the safety spark gap which is also connected to earth through a second electromagnet S at E. On connecting up the condenser circuit 11 this is charged over the spark gap *f* whereby an oscillatory discharge is formed. This discharging current acts through inductance 10 on the inductively coupled secondary 9, whereby in the producing network a modification of the potential of the condenser 19 is produced. The consequence of this is that oscillations arise in the coil shaped producer network. These oscillations induce a current in the secondary circuit II, which has a smaller number of windings and a less resistance, the voltage of which, according to the proportion of the number of windings and of the ohmic resistance, is considerably lower whilst the current strength is greater.

In order to convert the current thus obtained into current of an undamped character, and to tune its wave lengths, a sufficiently large regulatable capacity 20 is inserted between the ends 12 and 13 of the secondary conductor II. Here also current may be taken without an earth conductor, but it is advisable to insert a safety spark gap E^1 and to connect this with the earth over an electromagnet S^2.

The producer network may be connected with the working network II over an inductionless condenser 21 or over an induction condenser 22, 23. In this case the secondary conductor is unipolarly connected with the energy conductor.

In Figure 28 the connecting conductor between the separate accumulator balloons is carried out according to the autotransformer principle. The collecting coil connects four aerial balloons 1, 2, 3, 4, the windings of which are not made side by side but one above the other. In Figure 28 the collector coil I is shown with a thin line, the metallically connected prolongation coils II with a thick line. Between the ends I^1 and II^1 of the energy network I a regulating capacity 19 is inserted. The wire I^1 is connected with the output wire and with the spark gap F.

As transformer of the atmospheric electricity an arrangement is employed which consists in using rotary pairs of condensers in which the one stator surface B is connected with the main current, whilst the other A is connected with the earth pole. Between these pairs of short circuited condensers are caused to rotate from which the converted current can be taken by means of two collector rings and brushes, in the form of an alternating current, the frequency of which is dependent on the number of balloons and the revolutions of the rotor. As the alternating current formed in the rotor can act, in this improved method of connection described in this invention, through coils 10 on the inductance 9, an increase or diminution of the feed current in I can be obtained according to the direction of the current by back induction. Current oscillations of uniform rhythm thereby result in the coil shaped windings of the producer net-work.

As the ends of this conductor are short circuited through the regulatable condenser 19 these rhythms produce short circuited undamped oscillations in the energy conductor, the periodicity and wave lengths of which oscillations can be adjusted according to desire by altering the capacity 19 to a given wave length and therewith also to a given frequency. These currents may also be employed in this form directly as working current through the conductors II^1 and III. By inserting the condenser 20 a connection between these conductors may also be made, whereby harmonic oscillations of desired wave length are formed. By this means quite new effects as regards current distribution are obtained. The withdrawal of current can even take place without direct wire connection if, at a suitable point in the interior of the producing network (quite immaterially whether this has a diameter of 1 or 100 km.) a coil tuned to these wave lengths and of the desired capacity is firmly or movably mounted in the aerial conductor in such a way that its axial direction is in parallel with that of the collector coil. In this case a current is induced in the producing network, the size of which is dependent on the total capacity and resistance and also on the periodicity employed. A possibility is thereby afforded in future, of taking energy from the producer network by wireless means. As thereby in addition to atmospheric electricity also magnetic earth currents and the energy from the higher atmosphere (at least partially) may be simultaneously obtained, this last system for collecting the atmospheric energy is of particular importance for the future.

Of course everywhere instead of spark gaps suitable grid vacuum tubes may be employed as producers for undamped oscillations. The separate coils of the producer net-work with large diameters may be connected with one another through separate conductors all in parallel or all in series or in groups in series. By regulating the number of oscillations and also the extent of the voltage more or less large collector coils of this kind may be employed. The coils may also be divided spirally over the entire section. The coils may be carried out in annular form or also in triangular, quadrangular, hexagonal or octagonal form.

Of course wires may be carried from a suitable place to the centre or also laterally which serve the current waves as guides. This is necessary when the currents have to be conducted over mountains and valleys and so forth. In all these cases the current must be converted into a current of suitable periodicity.

As already hereinbefore mentioned separate collecting balloons may be directly metallically interconnected at equidistant stations distributed over the entire country or may be connected by interpolation of suitable condenser batteries by means of high voltage conductors. The static electricity is converted through a spark gap into dynamic energy of a high number of oscillations, and could then in such form, with a suitable arrangement of the connections, observing various measures of precaution, be employed as source of energy after separate or special regulation.

According to this invention in order to increase the collecting effect of the balloon in the aerial collector conductor or in the earth wire, radiating collectors are employed. These consist either of incandescent metal or oxide electrodes in the form of vacuum grid tubes, or electric arcs (mercury and the like electrodes) Nernst lamps, or finally flames of various kinds may be simply connected with the respective conductor.

It is well known that energy can be drawn off from a cathode consisting of an incandescent body opposite an anode charged with positive electricity (vacuum grid tube). Hitherto however, a cathode was always first directly placed opposite an anode, and secondly the system always consisted of a closed circuit.

Now if we dispense with the ordinary ideas in forming light or flame arcs in which a cathode must always stand directly opposite an anode, and if we place an incandescent cathode opposite an anode charged to a high potential or another body freely floating in the air, or regard the incandescent cathode only as a source of unipolar discharge (which represent group and point discharges in electro-static machines similar to unipolar discharges), it may be ascertained that incandescent cathodes and less perfectly all incandescent radiators, flames and the like admit of relatively large current densities and allow large quantities of electric energy to radiate into the open space in the form of electron streams as transmitters.

The object of this invention is as described below, if such incandescent oxide electrodes or other incandescent radiators or flames are not freely suspended in space but connected metallically with the earth so that they can be charged with negative terrestrial electricity, these radiators possess the property of absorbing the free positive electrical charges contained in the air space surrounding them (that is to say of collecting them and conducting them to earth). They can therefore, serve as collectors and have, in comparison to the action of the spikes, or points, a very large radius of action R; the effective capacity of these collectors is much greater than the geometrical capacity (R_0-) calculated in an electro-static sense.

Now as our earth is surrounded as is well known, with an electro-static field and the difference of potential

$$\frac{\delta V}{\delta h}$$

of the earth field according to the latest investigations, is in summer about 60 to 100 volts and in winter 300 to 500 volts per metre of difference in height (δh), a simple calculation gives the result that when such a radiation collector or flame collector is arranged for example on the ground, and a second one is mounted vertically over it at a distance of 2,000 metres and both are connected by a conducting cable, there is a difference of potential in summer of about 2,000,000 volts and in winter even of 6,000,000 volts and more.

According to Stefan Boltzmann's law of radiation, the quantity of energy which an incandescent surface (temperature T) of 1 sq. cm. radiates in a unit of time into the open air (temperature T_o) is expressed by the following formula:

$$S=\sigma\,(T^4-T^4_0)\ \text{watt/sq. cm.}$$

and the universal radiation constant σ is according to the latest researches of Ferry (Annales de Chimie et de physique 17 page 267 (1909)) equal to 6.30×10^{-12} watt/sq. cm.

Now if an incandescent surface of 1 sq. cm. shows, as compared with the surrounding space a periodic fall of potential δV it radiates (independent of the current direction, that is to say of the sign) in accordance with the above formula, for example at a temperature of 3725° C. an energy of 1.6 kw. per sq. cm. per second. As for the radiation the same value can be calculated for the collection of energy, but reversed. Now as carbon electrodes at the temperature of the electric arc support on the current basis a current density up to from 60 to 65 amperes per sq. cm. no difficulties will result in this direction in employing radiating collectors as accumulators.

If the earth be regarded as a cosmically insulated condenser in the sense of geometrical electro-statics *x* there results from the geometric (compare Ewald Rasch, "das elektrische Bogenlicht" (The electric arc light) page 169) capacity of the earth according to Chwolson:

For negative charging 1.3×10^6 Coulomb

For negative potential $V=10\times10^8$ volts. From this there results however, $EJT\cong24.7\times10^{24}$ watt/Sec. Now if it is desired to make a theoretic short circuit through an earthed flame collector this would represent an electric total work of about 79,500 10^{10} kilowatt years. As the earth must be regarded as a rotating mechanism which is thermo-dynamically, electromagnetically, and also kinematically coupled with the sun and stars system by cosmic radiations and gravitation a diminution of the electric energy of the earth field is not to be feared. The energies which the incandescent collectors would withdraw from the earth field can only cause by the withdrawal of motor work a lowering of the earth temperature (temperature $T_E=300$) and reduce this to that of the world space ($T=O$) by using the entire energy. This is however not the case as the earth does not represent a cosmically entirely insulated system. On the contrary there is conveyed to the same according to the recent value corrected by Ferry for the solar constants through the radiation from the sun an energy of $18{,}500\times10^{10}$ kw. Accordingly any lowering of the earth temperature (T_E) without a simultaneous lowering of the sun's temperature (T_S) would contradict Stefan Boltzmann's law of radiation.

$$S=\sigma\,(T_S^4-T_E^4).$$

From this it must be concluded that if the earth temperature (T_E) sinks, the total radiation S absorbed by the earth increases, and further also that the secular speed of cooling of the earth is directly dependent on that of the sun and the other radiators cosmically coupled with the sun and is connected most closely with these.

The incandescent radiation collectors may, according to this invention, be employed for collecting atmospheric electricity if they (1) are charged with the negative earth electricity (that is to say when they are directly connected by means of a metallic conductor with the earth) and (2) if large capacities (metal surfaces) charged with electricity are mounted opposite them as positive poles in the air. This is regarded as the main feature of the present invention as without these inventive ideas it would not be possible to collect with an incandescent collector, sufficiently large quantities of the electrical charges contained in the atmosphere as technology requires; the radius of action of the flame collectors would also be too small, especially if it be considered that the very small surface density (energy density) (σ about$=2\times7\,.\,10^9$ St. E. per sq. cm.) does not allow of large quantities of charge being absorbed from the atmosphere.

x) Calculated according to Poisson's calculation;

$\Delta V=-4\pi\delta$; as here the alteration of the potential or potential gradients only takes place in the direction of the normal, this calculation assumes the simple form

$$\delta=\frac{1}{4\pi}\times\frac{\delta^2V}{\delta n^2}$$

It has indeed already been proposed to employ flame collectors for collecting atmospheric electricity and it is known that their collecting effect is substantially greater opposite the points. It is however, not known that the quantities of current which could hitherto be obtained are too small for technical purposes. According to my experiments the reason for this is to be found in the too small capacities of the collector conductor poles. If such flame or radiating collectors have no or only small positive surfaces, their radius of action for large technical purposes is too small. If the incandescent collectors be constantly kept in movement in the air they may collect more according to the speed of the movement, but this is again not capable of being carried out in practice.

By this invention the collector effect is considerably increased by a body charged with a positive potential and of the best possible capacity being also held floating (without direct earth connection) opposite such an incandescent collector which is held floating in the air at a desired height. If for example, a collecting balloon of sheet metal or of metalized balloon fabric be caused to mount to 300 up to 3,000 metres in the air and as positive pole it is brought opposite such a radiating collector connected by a conductor to earth, quite different results are obtained.

The metallic balloon shell (with a large surface) is charged to a high potential by the atmospheric electricity. This potential is greater the higher the collecting balloon is above the incandescent collector. The positive electricity acts concentratedly on the anode floating in the air as it is attracted through the radiation shock ionization, proceeding from the incandescent cathode. The consequence of this is that the radius of action of the incandescent cathode collector is considerably increased and thereby also the collecting effect of the collecting balloon surface. Further the large capacity of the anode floating in the air plays therefore an important part because it allows of the taking of large charges, and thereby a more uniform current is obtained even when there is a large consumption: this cannot be the case with small surfaces.

In the present case the metallic collecting balloon is a positive anode floating in the air and the end of the earth conductor of this balloon serves as positive pole surface opposite the surface of the radiating incandescent cathode, which in turn is charged with negative earth electricity being conductingly connected to earth.

The process may be carried out by two such contacts (negative incandescent cathode and anode end of a capacity floating in the air) a condenser and an inductive resistance being switched on in parallel, whereby simultaneously undamped oscillations may be formed.

In very large installations it is advisable to connect two such radiating collectors in series. Thus an arc light incandescent cathode may be placed below on the open ground and an incandescent cathode which is heated by special electro-magnetic currents be located high in the air. Of course for this the special vacuum Liebig tubes with or without grids may also be employed. An ordinary arc lamp with oxide electrodes may be introduced on the ground and the positive pole is not directly connected with the collecting balloon, but through the upper incandescent cathode or over a condenser. The method of connecting the incandescent cathode floating in the air may be seen in Figs. 29–33.

B is the air balloon, K a Cardan ring (connection with the hawser) C the balloon, L a good conducting cable, P a positive pole, N negative incandescent cathode, and E earth conductor.

Fig. 29 represents the simplest form of construction. If electric oscillations are produced below on the ground by means of a carbon arc lamp or in other suitable way a considerably greater electric resistance is opposed to that in the direct way by inserting an electrical inductive resistance 9. Consequently between P and N a voltage is formed, and as, over N and P only an inductionless ohmic resistance is present, a spark will spring over so long as the separate induction co-efficients and the like are correctly calculated. The consequence of this is that the oxide electrode (carbon or the like) is rendered incandescent and then shows as incandescent cathode an increased collecting effect. The positive poles must be substantially larger than the negative in order that they may not also become incandescent. As they are further connected with the large balloon area which has a large capacity and is charged at high voltage, an incandescent body which is held floating in the air and a positive pole which can collect large capacities is thereby obtained in the simplest way. The incandescent cathode is first caused to become incandescent by means of separate energy produced on the earth, and then maintained by the energy collected from the atmosphere.

Fig. 30 only shows the difference that instead of a round balloon a cigar shaped one (of metal or metalized fabric) may be employed and also a condenser 5 is inserted between the incandescent cathode and the earth conductor so that a short circuited oscillation circuit over P. N 5 and 9 is obtained. This has the advantage that quite small quantities of electricity cause the cathode to become incandescent and much larger cathode bodies may be rendered incandescent.

In this form of construction both the incandescent cathode and also the positive electrode may be enclosed in a vacuum chamber as may be seen in Fig. 32. A cable L is carried well insulated through the cover of a vessel and ends in a condenser disc 5. The cover is arched in order to keep off the rain. The vessel is entirely or partially made of magnetic metal and well insulated inside and outside. Opposite the disc 5 another disc 6 and on this again a metallic positive pole of the vacuum tube *g* with the incandescent cathode (oxide electrode) N is arranged. The negative electrode is on the one hand connected with the earth conductor E, and on the other hand with the inductive resistance 9 which is also connected with the cable L with the positive pole and wound round the vessel in coils. The action is exactly the same as that in Fig. 29 only instead of an open incandescent cathode one enclosed in vacuo is employed. As in such collectors only small bodies can be brought to incandescence in large installations a plurality of such vacuum tubes must be inserted in proximity to one another. According to the previous constructions Figs. 31 and 33 are quite self evident without further explanations.

Figs. 34–37 represent further diagrams of connections over radiating and flame collectors, and in fact, how they are to be arranged on the ground.

Fig. 34 shows an arc light collector with oxide electrodes for direct current and its connection; Fig. 35 a similar one for alternating current, Fig. 36 an incandescent collector with a Nernst lamp and Fig. 37 a similar one with a gas flame.

The positive pole 1 of the radiating collectors is always directly connected to the aerial collecting conductor A. In Fig. 34 this is further connected over the condenser battery 5 with a second positive electrode 3. The direct current dynamo *b* produces current which flows over between the electrodes 3 and 2 as an arc light. On the formation of an arc the negative incandescent electrode 2 absorbs electricity from the positive poles standing opposite it and highly charged with atmospheric electricity and conveys the same to the working circuit. The spark gap 7, inductive resistance 9 and induction coil 10 are like the ones previously described. The protective electromagnet S guards the installation against earth circuiting, the safety spark gap 8 from excess voltage or overcharging.

In Fig. 35 the connection is so far altered that the alternating current dynamo feeds the exciting coil 11 of the induction condenser. 12 is its negative and 13 its positive pole; if the coil 3 on the magnet core of the dynamo is correctly calculated and the periodicity of the alternating current is sufficiently high an arc light can be formed between the two poles 1 and 2. As the cathode 2 is connected with the negatively charged earth, and therefore always acts as a negative pole, a form of rectification of the alternating current produced by the dynamo 3 is obtained, the second half of the period is always suppressed. The working circuit may be carried out in the same way as in Fig. 34; the working spark gap 7 may however be dispensed with, and instead thereof between the points *n* and *m* a condenser 5 and an induction resistance 9 may be inserted from which the current is taken inductively.

Fig. 36 represents a form of construction similar to Fig. 34 only that here instead of an arc lamp a Nernst incandescent body is employed. The Nernst lamp is fed through the battery 3. The working section is connected with the negative pole, the safety spark gap with the + poles. The working spark gap 7 may also be dispensed with and the current for it taken at 12 over the oscillation circuit 5, 11 (shown in dotted lines).

Flame collectors (Fig. 37) may also be employed according to this invention. The wire network 1 is connected with the aerial collector conductor A and the burner with the earth. At the upper end of the latter, long points are provided which project into the flame. The positive electrode is connected with the negative over a condenser 5 and the induction coil 9 with the earth.

The novelty in this invention is firstly, the use of incandescent cathodes opposite positive poles which are connected with large metallic capacities as automatic collecting surfaces, (2) the connection of the incandescent cathodes with the earth whereby, in addition to the electricity conveyed to them from the battery or machine which causes the incandescing, also the negative charge of the earth potential is conveyed, and (3) the connection of the positive and negative poles of the radiating collectors over a condenser circuit alone or with the introduction of a suitable inductive resistance, whereby simultaneously an oscillatory oscillation circuit may be obtained. The collecting effect is by these methods quite considerably increased.

I declare that what I claim is:—

1. An electrical energy generating system, comprising a conducting surface for static charges, means to support same at a distance above the earth, a conductor leading to the earth level, a spark gap associated with said conductor to convert electrostatic charges into electromagnetic high frequency oscillations means to supply said electromagnetic energy to a net work, and a spark gap of greatly increased relative resistance in parallel therewith.

2. An electrical energy generating system

comprising a conductor, means to support same above the earth level, an inductance therein, a spark gap associated with said conductor, a second spark gap of much higher relative resistance in parallel therewith and an energy receiving circuit coupled with the spark gap of lesser resistance.

3. An electrical energy generating system comprising a collecting surface, means to support same above the earth level, a conductor connecting said collecting surface with the earth level, a choke in said conductor, an electromagnetic resistance converting electrostatic energy to electromagnetic energy, a safety higher resistance in parallel therewith and a net work coupled with the conversion resistance of lesser value.

4. An electrical energy generating system comprising electric conductors spaced above the earth to form electromagnetic oscillating circuits, conductors connecting to earth level, electrostatic to electromagnetic energy conversion means therein, a safety high electrostatic resistance in parallel therewith and means to alter the electromagnetic characteristic of the circuits.

5. An electrical energy generating system comprising in combination a static collecting surface arranged above the earth, conductors connecting to earth level, a pair of spark gaps in parallel of different electrostatic resistance, a utilization net work shunted across the spark gap of lesser resistance and an electromagnetic choke in said conductors.

6. An electrical energy generating system comprising an open circuit energy collecting aerial, a pair of sparking gaps in parallel of widely different resistance, connected thereto and a closed electric oscillation circuit in shunt across the gap of lesser resistance.

7. An electrical energy generating system comprising an open circuit energy collecting aerial, a pair of sparking gaps in parallel of widely different resistance connected thereto, a closed electric oscillation circuit in shunt across the gap of lesser resistance, a plurality of electrostatic collecting surfaces, means to connect said collecting surfaces in parallel in groups and means to connect said groups symmetrically with said aerial.

In witness whereof, I have hereunto signed my name this 30 day of Dec., 1920, in the presence of two subscribing witnesses.

HERMANN PLAUSON.

Witnesses:
H. F. Armstrong,
W. H. Beeston.

OLEG D. JEFIMENKO
West Virginia University

Electrostatic Motors

Their History, Types, and Principles of Operation —

With many illustrations, of which 57 are
by
David K. Walker
of
Waynesburg College

Electret Scientific Company
Star City

UNIVERSITY OF COLORADO
DENVER CAMPUS LIBRARY

Copyright © 1973 by
Electret Scientific Company
P.O. Box No. 4132
Star City, West Virginia 26505

All rights reserved

Printed in the United States of America

This book is dedicated to an accomplished
farmer, craftsman, musician,
educator, and lawyer
MY FATHER

PREFACE

The fascinating science of electrostatics was developed mainly in the 18th century at a time when the technology and industry were too primitive to put this science to practical uses. As a result, serious research in electrostatics soon lost its momentum and, except for a few isolated efforts, was practically nonexistent during the entire 19th century. Only very recently practical aspects of electrostatics began to make their impact on the industry and economy, and the once glamorous but long forgotten science has again appeared at the focal point of serious scientific investigations.

A peculiar obstacle stands, however, in the way of many such investigations: because of the one and one half century of neglect of electrostatic explorations there is a singular lack of easily available quantitative and qualitative information on even the most basic electrostatic phenomena, techniques, and devices. Whereas experimental and theoretical data pertaining to most modern sciences are well documented and are easily retrievable from numerous reference sources, many data in the field of electrostatics must be extracted through thorough and laborious firsthand studies of old books and magazines, the very existence of which is not generally known, and which are not readily available in any of the present-day libraries or repositories.

It is precisely this kind of obstacle that the author encountered, when several years ago he started his research on electrostatic motors. Luckily, however, the obstacle soon transformed itself into a highly rewarding experience of studying various "ancient" books and periodicals and searching through electrostatic inventories of various European and American science museums. In the course of these studies he found numerous forgotten publications on electrostatic motors and found that several old electrostatic motors can actually be seen in some of the museums. (It is interesting to note that the latter motors are usually not on public display, apparently because of a lack of adequate information about their purpose and mode of operation available to museum personnel. It is also interesting to note that certain types of electrostatic motors were frequently employed in various animated toys during the second half of the 18th century, and that, although some of these toys are shown in museums, they are usually shown without an explanation of what made them move or how they were supposed to function.)

The purpose of this book is to describe the various types of electrostatic motors reported in the scientific literature between 1700 and the present, and to discuss in general terms their various design features and principles of operation. The book is written for a wide circle of readers, and care has been taken to avoid technical details that would be useful only to a small group of specialists. The readers interested in additional scientific and technical data on the various motors are referred to appropriate original publications cited in this book.

The material is presented in several chapters, each chapter describing a particular class of motors. The sequence of the chapters corresponds to the chronological order in which the earliest motor belonging to each particular chapter was invented. The sequence of material in each chapter is normally arranged also in a chronological order. The last chapter describes some of the most recent research results on electrostatic motors and extrapolates these results into the immediate future.

The author is grateful to his wife Valentina for typing the manuscript and for otherwise helping him to make the book ready for publication. He is also grateful to Dr. David K. Walker for reading the manuscript and for drawing the illustrations for the book.

Oleg D. Jefimenko

CONTENTS

Preface vii

1. What are Electrostatic Motors? 1

Electromagnetic and Electrostatic Motors. Some Unique Properties of Electrostatic Motors. Classification of Electrostatic Motors.

2. Electric Pendulum Motors; Contact Motors. 6

Gordon Electric Bells. Electric Swing. Electric Seesaw. Electric Perpetuum Mobile. Dailey Reciprocating Motor. Electric Racing Ball.

3. Electric Wind Motors. 22

Gordon Electric Fly. Hamilton Electrometer. Langenbucher Electrometer. Electric Orrery. Kinnersley Electric Horse-Race. Bichat Electric Tourniquet. Corona-Kraft Levitation. Electric Wind Wheels and Turbines. King Electric Rabbit Hunt.

4. Franklin's Spark Motors and Their Descendants. 39

Franklin First Motor. Franklin Second Motor. Moore Motors. Bell Jar Motors. Another Electric Perpetuum Mobile. Dressel Nuclear Battery Motor.

5. Corona Motors and Motors Similar to Them. 56

Holtz Rotation Phenomenon. Poggendorff Motor. Christiansen Motors. Gruel Electric Tourbillion. Ruhmkorff Motor. Johnson Motors. McVay Motors.

6. Capacitor Motors. 73

Zipernowsky Motor. Van Huffel Motor. Synchronous and Asynchronous Motors. Bollée Capacitor Motors. Moore Interdigital Motor.

7. Induction Motors. 84

Arno Motor. Weiler Induction Motors. Bollée Induction Motors. Choi and Dunn Motor.

8. Liquid-Immersed Motors. 91

Weiler Liquid-Immersed Motor. Secker Motors. Quincke Motor.

9. Electret Motors. 97

Gubkin Motor. Slot-Effect Motors. Synchronous Electret Motor.

10. What to Expect from Electrostatic Motors. 106

Motors for Miniaturized Systems. New Corona Motors. Sources of Power for Electrostatic Motors. Operation of Electrostatic Motors from Atmospheric Electricity.

11. Index 120

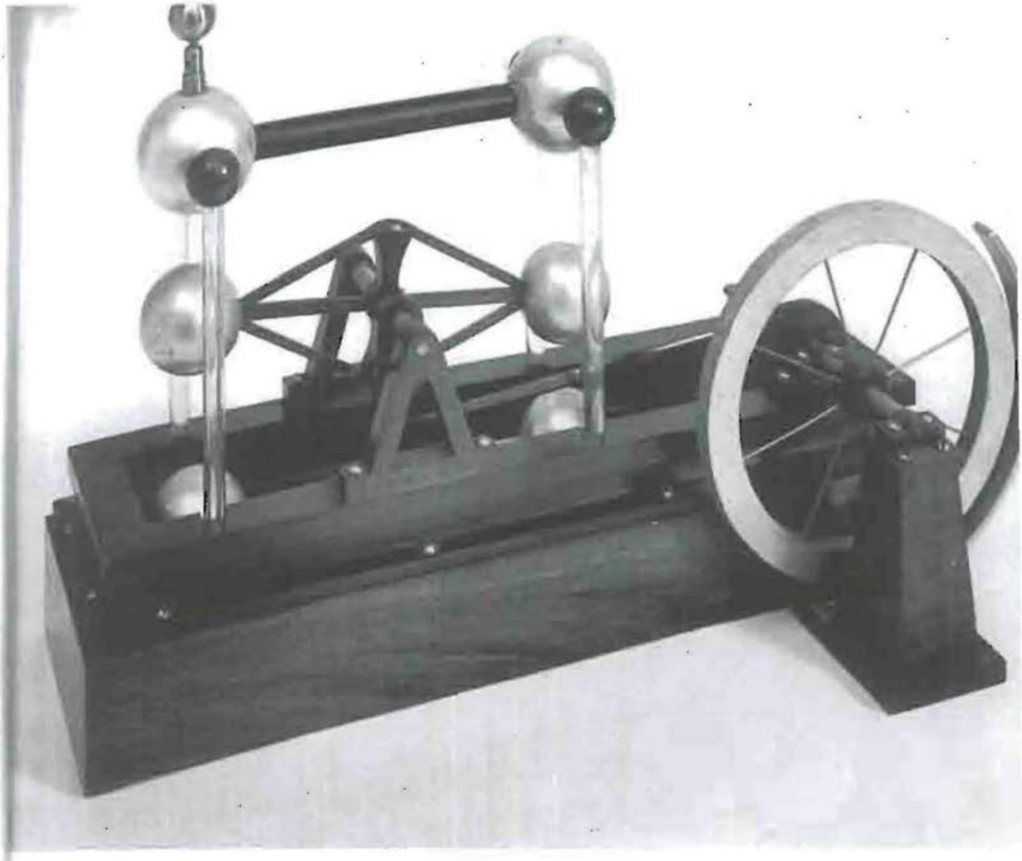

PLATE 1

Here is what H. B. Dailey wrote about this electrostatic motor which he and his father built in 1880: "Whether or not there is about the instrument and its history that which would ever give it a claim to any measure or serious interest, it is at least an electric motor most unique. A motor without magnetism, wiring, or any iron in its make up. A motor that runs by the action of the direct push and pull of the pure unconverted electricity itself".

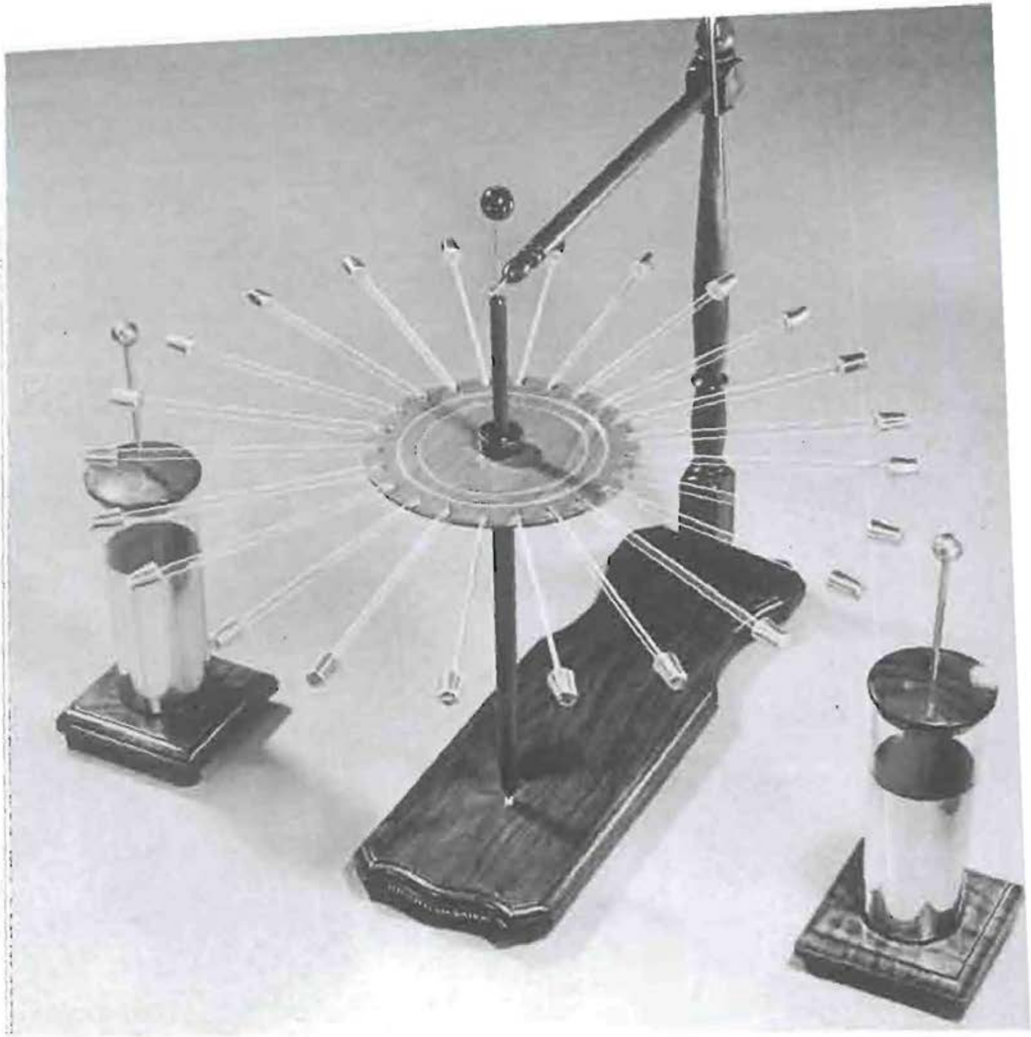

PLATE 2

It was Benjamin Franklin who in 1748 constructed the first electrostatic motors. No original drawings or models of his motors are known to exist, but his first motor must have looked very similar to this replica designed by the author for the Electret Scientific Company.

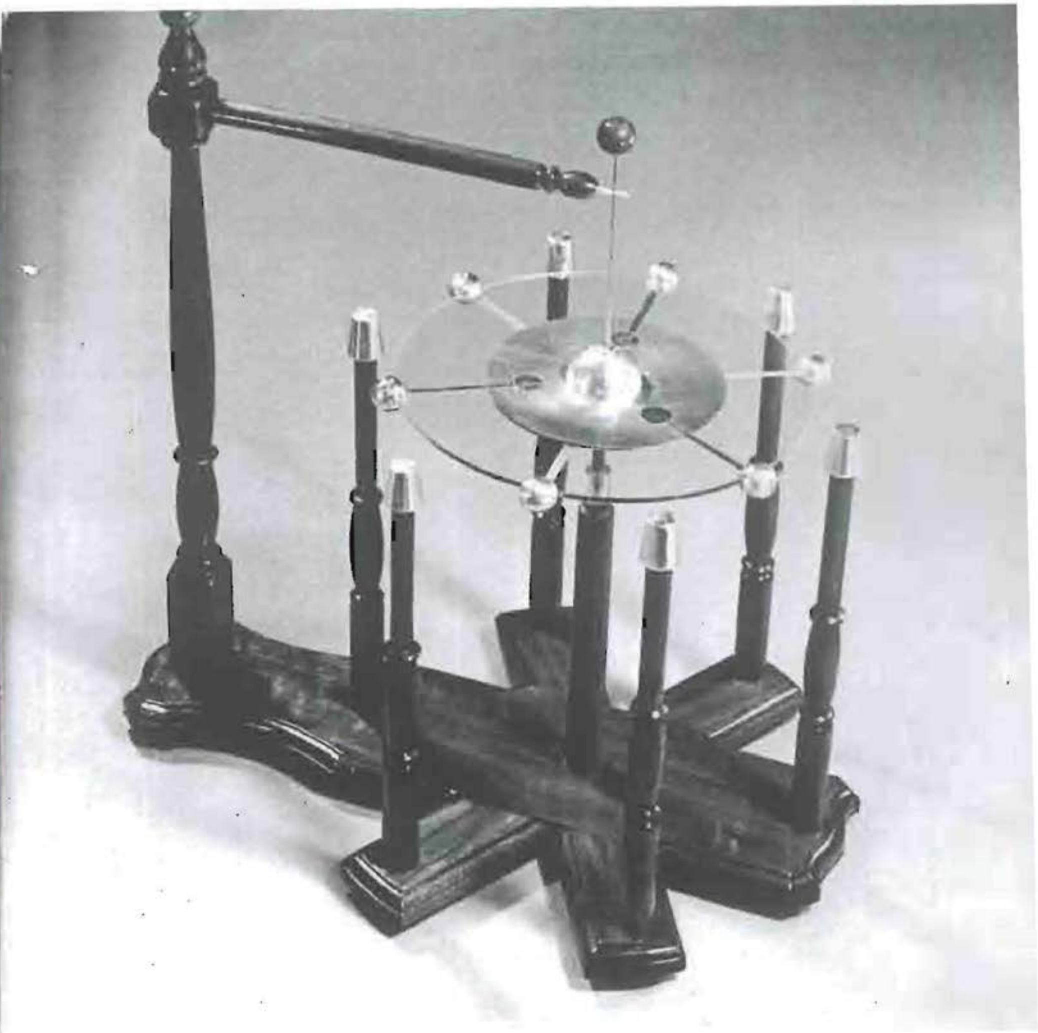

PLATE 3

Franklin's second motor probably looked very much like this replica also designed by the author for the Electret Scientific Company. Whereas the first motor operated from the electricity stored in Leyden jars, this motor operated from the electricity stored in the motor itself.

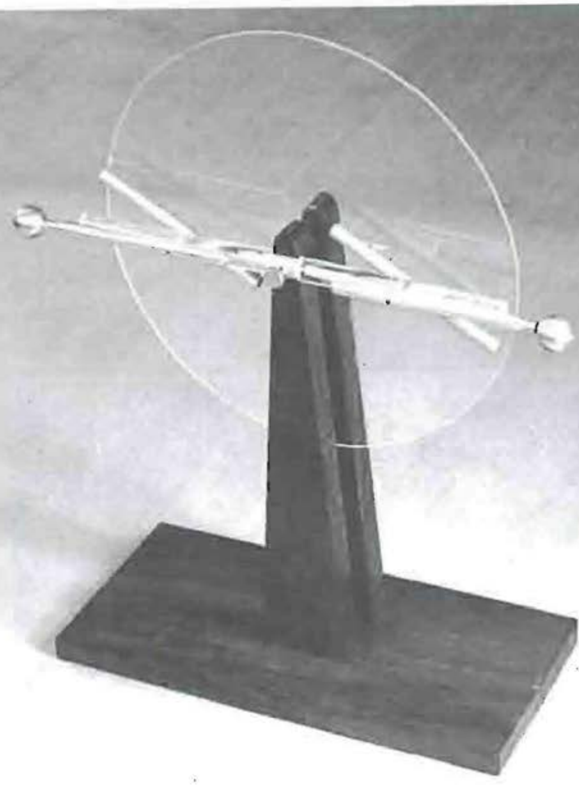

PLATE 4

The first corona motor was designed around 1869 by the German physicist Poggendorff. Poggendorff made a thorough study of the motor, but failed to appreciate its possibilities in the mistaken belief that no sources of electricity could supply enough power to any electric motor to make it do useful work. This simplified version of Poggendorff's motor was built by the author.

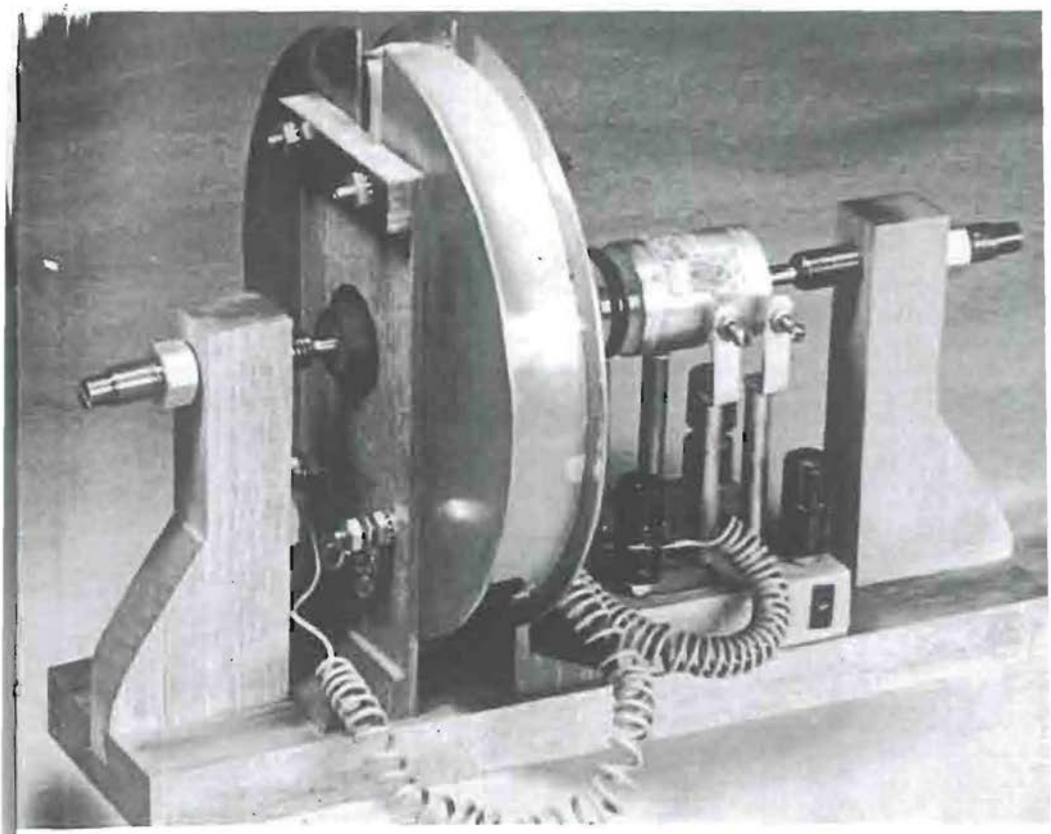

PLATE 5

The electret—a permanently electrized dielectric—is essentially a product of the 20th century. When an electret is placed between slotted electrodes to which a voltage is applied, the electret experiences a force. This motor designed by the author utilizes the slot effect. The rotor of the motor is a carnauba wax electret made of two oppositely polarized half-disks. The motor was built in 1966.

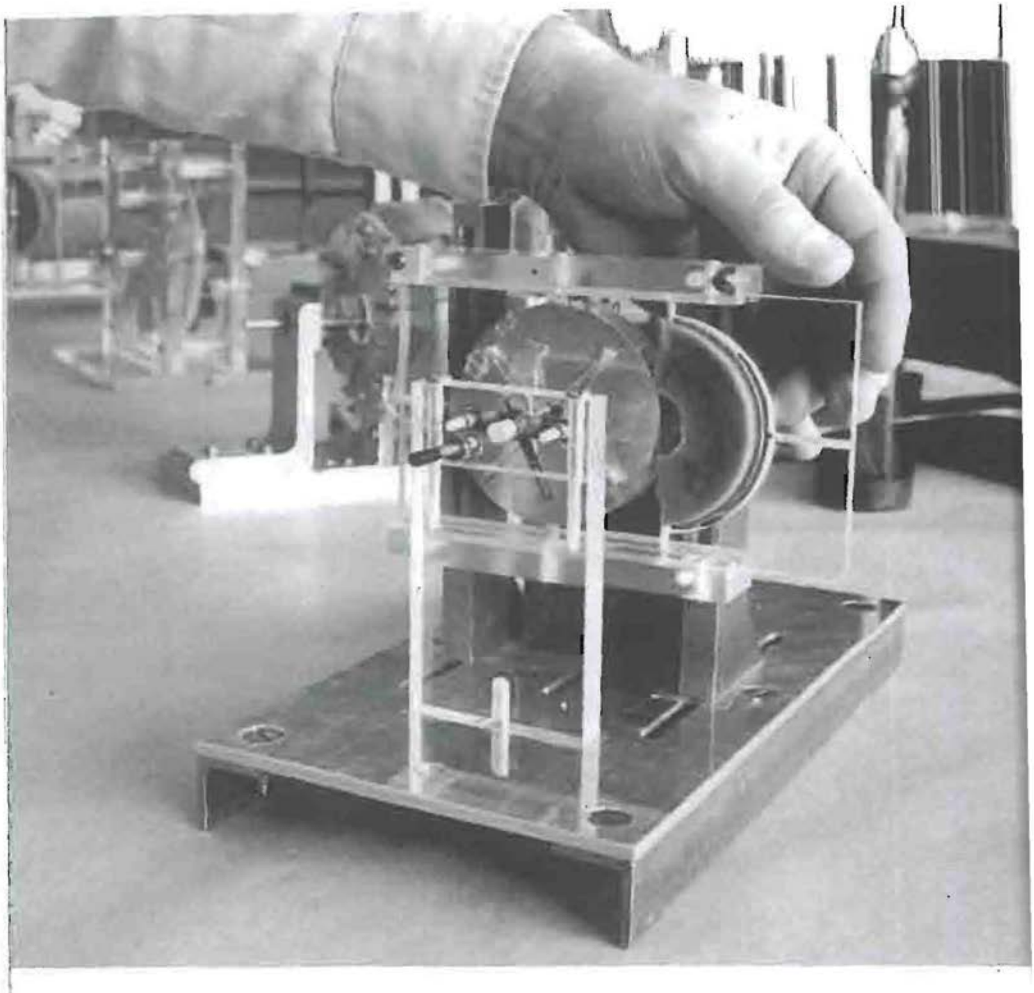

PLATE 6

A more practical design of an electret motor incorporates stationary electrets and rotating electrodes. The electrets can be easily withdrawn from the motor for servicing or replacement. The motor operates from approximately 60 volts dc. This particular motor was the first ever operated from atmospheric electricity.

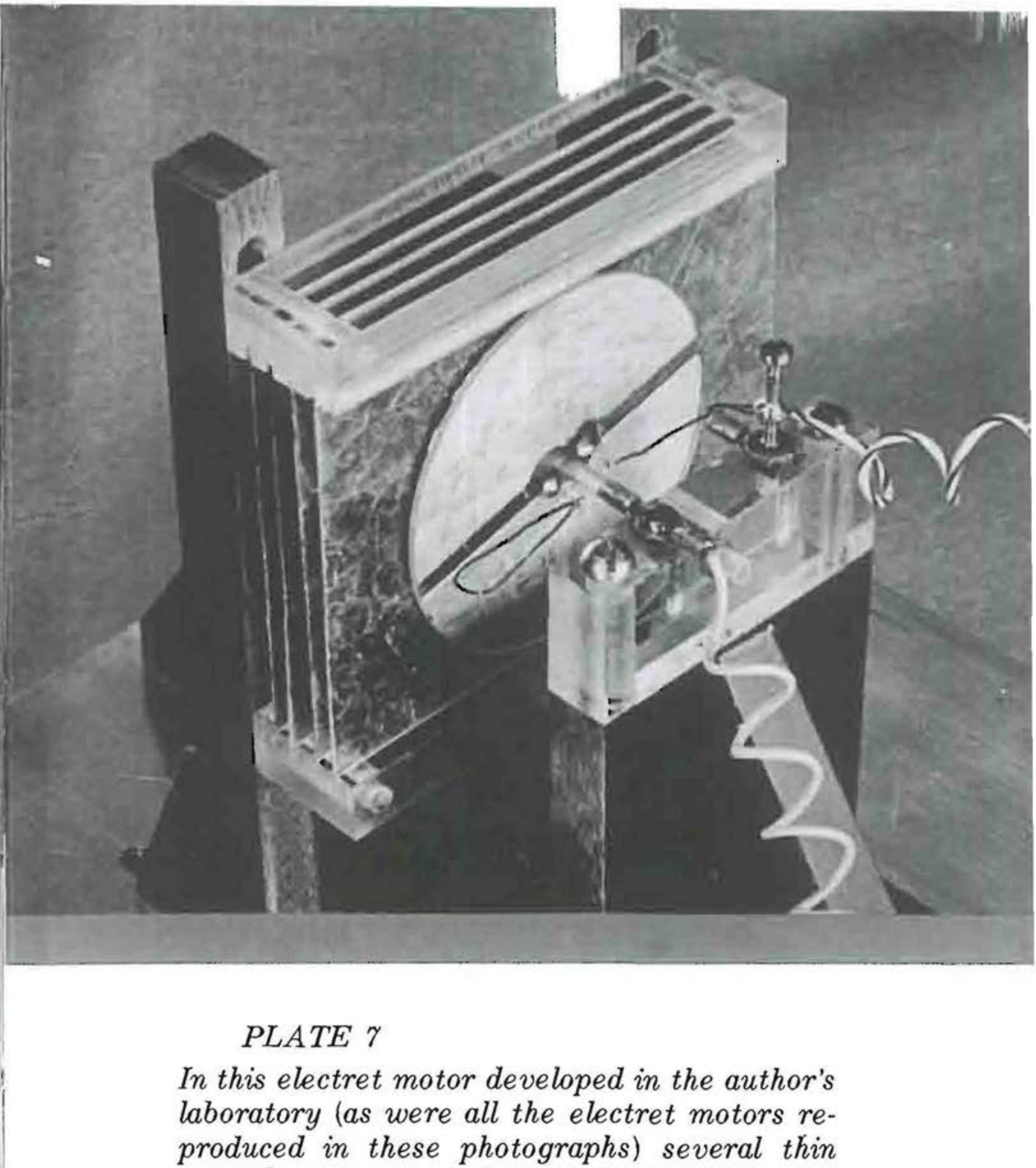

PLATE 7

In this electret motor developed in the author's laboratory (as were all the electret motors reproduced in these photographs) several thin mica electrets are used as active elements. This arrangement of electrets makes good use of the available space and allows one to construct relatively powerful electret motors.

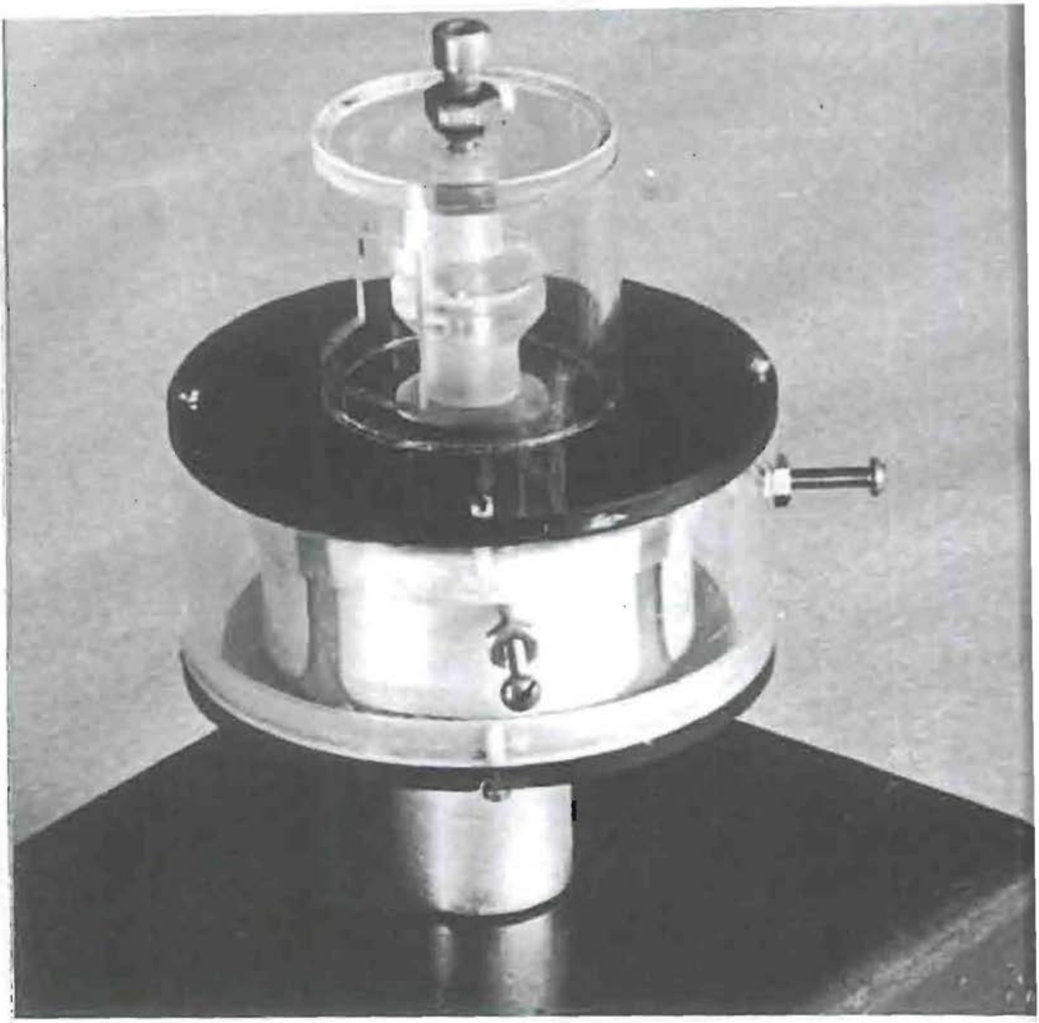

PLATE 8

This "hoop" electret motor built in 1967 by D. K. Walker has a stationary cylindrical electret. The rotor is made of four pairs of bent aluminum plates enclosing the electret. The motor requires no commutator. The motor has been operated for several years without any servicing.

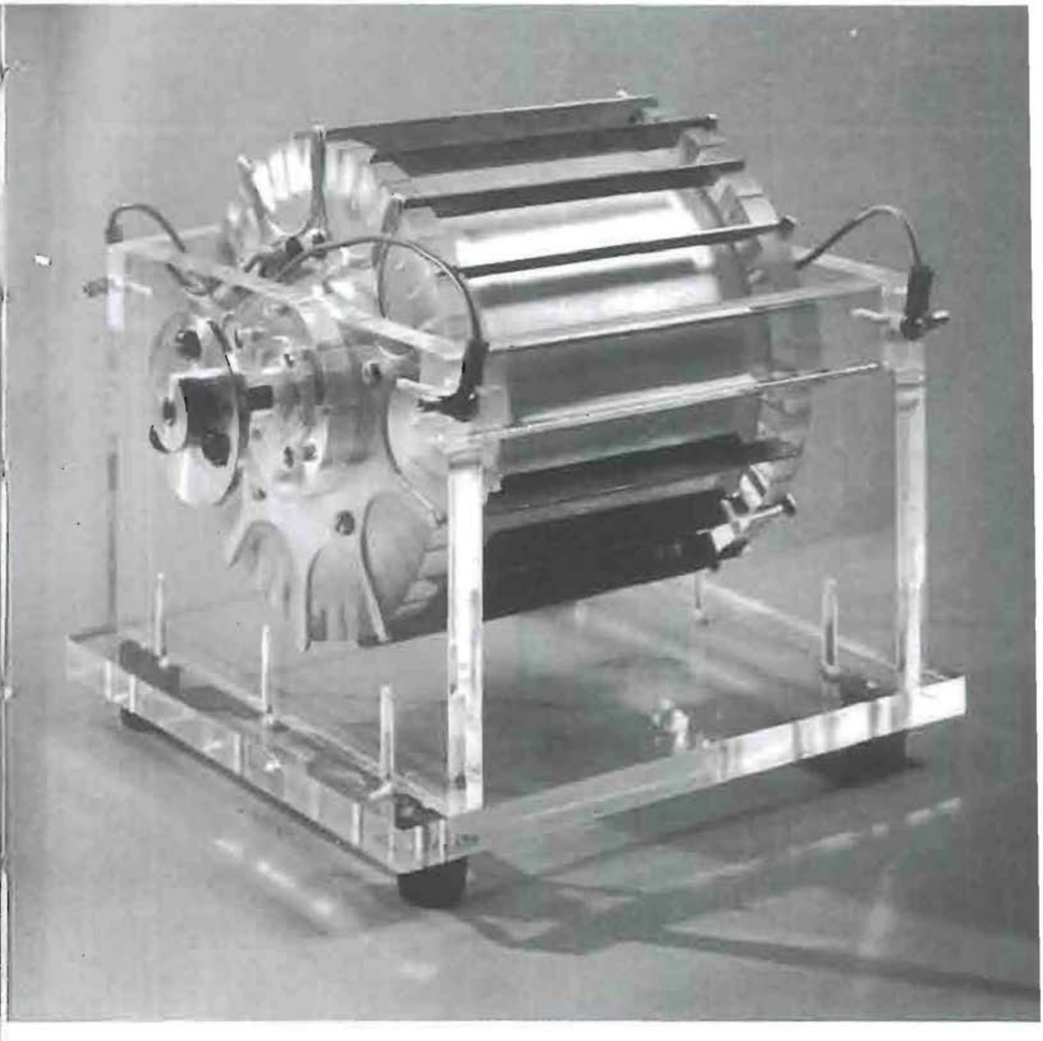

PLATE 9

Corona motors are probably the most promising electrostatic motors. This 0.1-hp modern version of the Poggendorff motor has a cylindrical rotor instead of a disk. The motor operates from a 6000-volt power supply as well as from an earth field antenna.

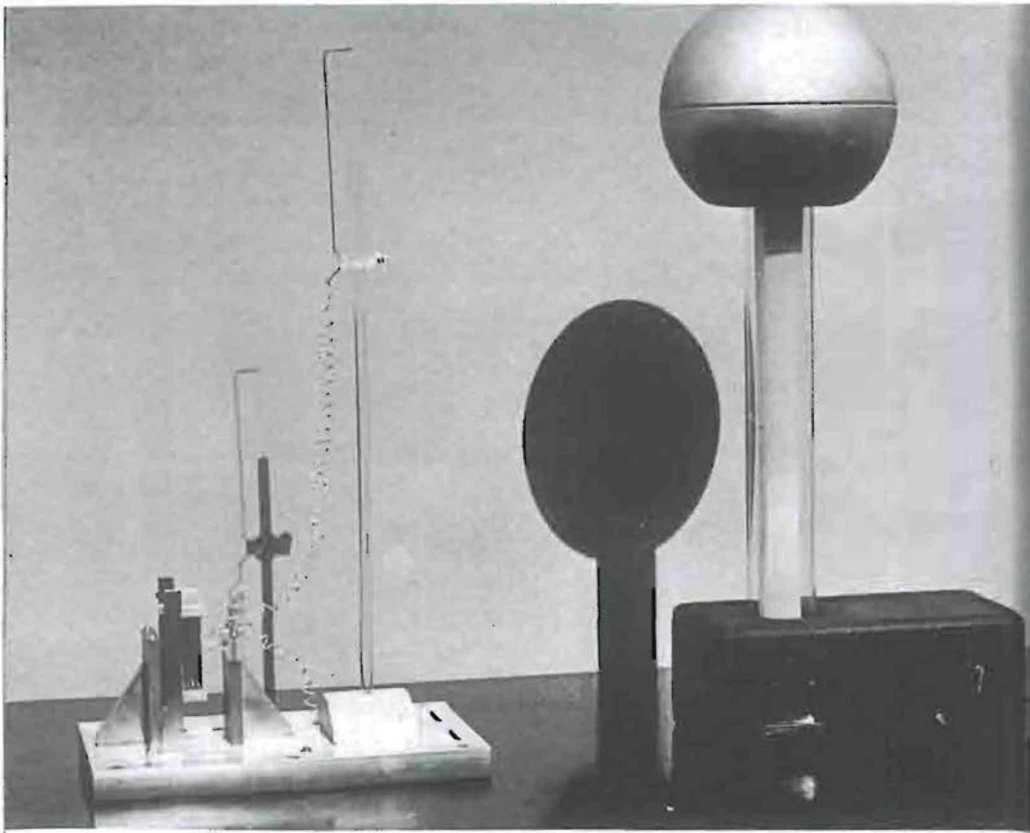

PLATE 10

The operation of electrostatic motors from antennas can be demonstrated with this lecture-room apparatus. The Van de Graaff generator produces electric charges in the air. The sharp-point antenna collects the charges from the air and delivers them to one terminal of an electrostatic motor whose other terminal is grounded.

PLATE 11

The author and D. K. Walker attempt to operate an electret motor from a 20-foot pole antenna in front of the Physics Building of the West Virginia University. The tall building screened the atmospheric field, and the motor did not run, although it operated very well from the same antenna in the nearby unobstructed parking lot.

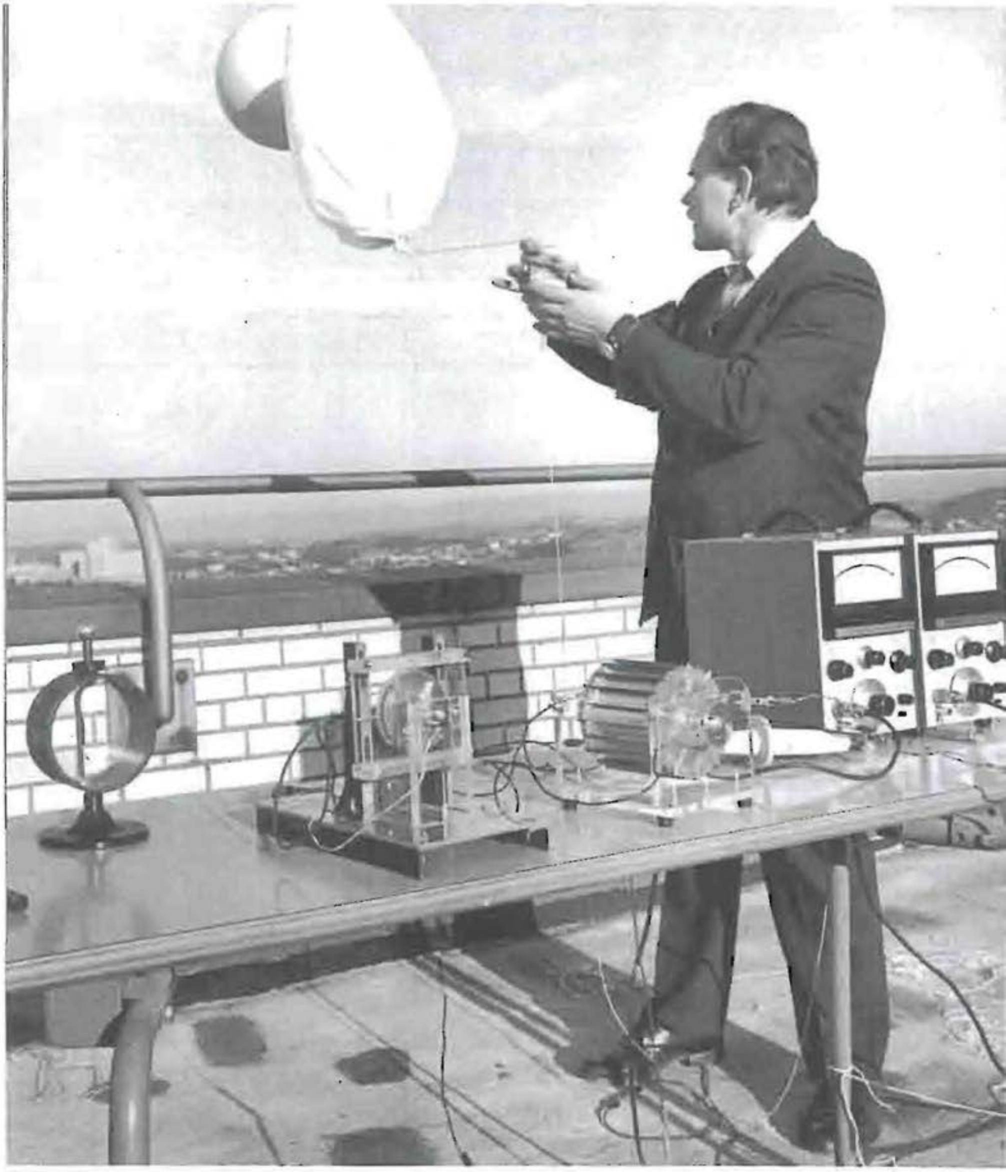

PLATE 12

The author operates electrostatic motors by means of an airborne antenna.

1

WHAT ARE ELECTROSTATIC MOTORS?

Conventional electric motors create mechanical motion as a result of magnetic forces acting upon electric currents. These motors are properly called *electromagnetic motors*. There is, however, another type of electric motor, in which the motion is created as a result of electric, or "electrostatic", forces acting between electric charges. Motors of this type are called *electrostatic motors*.

It is interesting to note that in nature the electrostatic forces are much stronger than the magnetic ones. There are many ways in which this can be demonstrated. For example, although a considerable effort may be needed to separate a magnet from an object attracted by it, a much greater effort is needed to break the magnet; this is because the magnet and the object are held together by magnetic forces, while the molecules of matter in the magnet (as well as in any other body) are held together by electrostatic forces.

Why, then, do we not use the electrostatic forces in our electric devices, and in motors in particular, at least on as wide a scale as we use the "inferior" magnetic forces? There are two main reasons for that. First, it is difficult to establish appreciable concentrations of electric charges without causing an electric breakdown in the medium surrounding or supporting the charges (although with modern insulating materials and techniques this difficulty becomes progressively less serious). Second, powerful electrostatic devices require voltages of many kilovolts for their operation, and until recently such voltages could not be produced conveniently and economically.

Motors of great power, however, are not the only ones needed. Equally important are low-power motors capable of performing various special tasks. In this respect electrostatic motors may compete successfully with their electromagnetic cousins even now.

Although electrostatic motors are not yet widely known or used, they already hold at least five very impressive records as compared with the electromagnetic motors:

1. The first electric motor ever invented was an electrostatic one. It was built about 100 years before the first electromagnetic motor was conceived.

2. The electric motor that operated without interruption longer than any other electric motor was an electrostatic one. This was a pendulum-type motor, known as the "electric perpetuum mobile", installed at the University of Insbruck, Austria, in 1823. Since then it operated continuously at least until 1909, powered by a Zamboni pile (an early high-voltage battery).
3. Electrostatic motors have been operated from voltages in excess of 10^5 volts, which is much higher than the voltages suitable for operating electromagnetic motors.
4. Electrostatic motors have been operated by using currents smaller than 10^{-9}amp, which is much less than the currents needed to operate electromagnetic motors.
5. The first electric motor that operated directly from the atmospheric electricity was an electrostatic motor. None of the presently available electromagnetic motors can operate directly from this source.

Even this short list reflecting some of the more obvious peculiarities of electrostatic motors shows quite clearly that electrostatic motors possess a number of unique properties. These properties undoubtedly will make electrostatic motors increasingly more important for the science, engineering, and technology of the future.

Many different types and designs of electrostatic motors are possible. It is customary to classify electrostatic motors in accordance with some prominent feature of their mode of operation or some prominent feature of their design. Thus, in reference to the techniques used for delivering electric charges to the active part of a motor one speaks of *contact motors*, *spark motors*, *corona motors*, *induction motors*, and *electret motors*. In reference to the medium in which the active part of a motor is located one speaks of *liquid-* or *gas-immersed motors*. In reference to the material and design of the active part of a motor one speaks of *dielectric motors* (the active part is made mainly of dielectric material) and *conducting-plate*, or *capacitor*, *motors* (the active part is made mainly of metal and resembles a variable capacitor). Finally, in reference to the rate of rotation of a motor relative to the period of the applied voltage (for ac operated motors) one speaks of *synchronous motors* and *asynchronous motors*.

This classification of electrostatic motors does not make it possible, however, to specify uniquely each individual motor. Also, two or more different operation modes are usually possible for most electrostatic motors (for example, certain motors can be operated at will as contact, spark, or corona motors). Therefore an assignment of a particular motor to one or another of the above types or categories is frequently more or less arbitrary. This applies, of course, also to the present book, where certain motors described in the chapters that follow could be equally well discussed under different chapter headings, and where the inclusion of certain motors in a particular chapter was occasionally dictated by the fact that such an inclusion resulted in a more coherent development of the subject matter.

2

ELECTRIC PENDULUM MOTORS; CONTACT MOTORS

If "electric motor" is understood to be a device converting electric energy into a continuous mechanical motion, then the first two electric motors were invented in the early 1740's by Andrew Gordon, a Scottish Benedictine monk and professor of philosophy at Erfurt, Germany.[1] Gordon's first motor was a device known as the "electric bells" (his second motor will be described in the next chapter). The device and its operation were as follows. A metallic clapper (pendulum) was suspended by a silk thread between two oppositely charged bells (Fig. 1). From an initial contact with one of the bells the clapper acquired a charge of the same polarity as that of the bell. Due to repulsion of like charges and attraction of opposite charges the clapper was then repelled by this bell and attracted by the second bell. As the clapper struck the second bell, it gave off its initial charge and acquired a charge of the same polarity as that of the second bell. Then the clapper was repelled by the latter and attracted by the first bell, which it struck again, and so on.

Numerous variations of Gordon's bells have been described by later authors, mostly as devices for lecture-room demonstrations of electrostatic forces. Such devices are widely used for this purpose even now.

An ingenious application of Gordon's bells was made by Benjamin Franklin in 1752. He connected the bells to an insulated lightning rod as a warning device "to give notice when the rod should be electrified".[2] (It appears that at that time neither Franklin nor any other scientist suspected how dangerous this "warning" device could be. The extreme danger of experiments with insulated lightning rods became clear a year later, when in 1753 the Russian physicist George Wilhelm Richmann was killed by lightning that entered his laboratory through such a rod, as he approached it in order to measure its electrification with a specially constructed electrometer.)

In a later modification of Gordon's invention one of the two bells was mounted on a Leyden jar (Fig. 2); the bells would then ring for as long as there was enough electric energy stored in the jar to move the clapper. (Franklin described the operation of an electric pendulum powered by a Leyden jar in 1747; see Ref. 2, p. 189.)

Sometimes Gordon's bells were made as a set

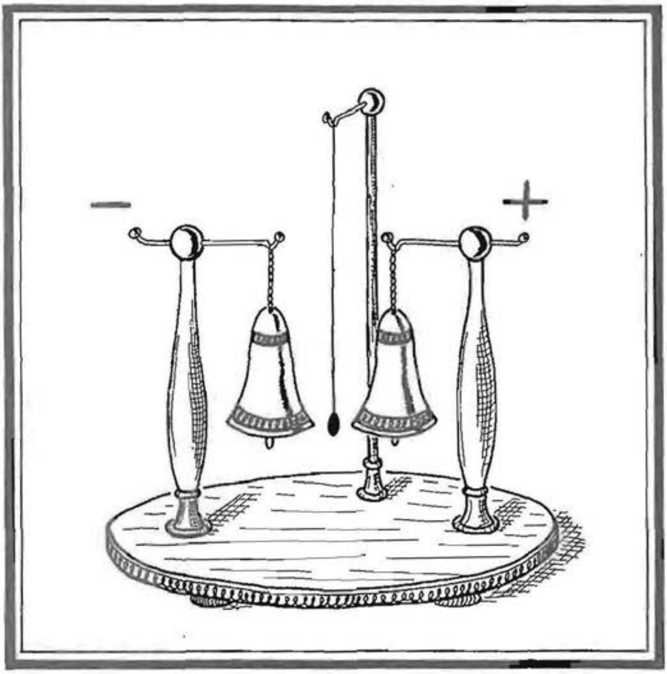

FIGURE 1

The so-called "electric bells" constituted the first device that converted electrical energy into a continuous mechanical motion. They were invented in about 1742 by Andrew Gordon, a professor of philosophy at Erfurt, Germany.

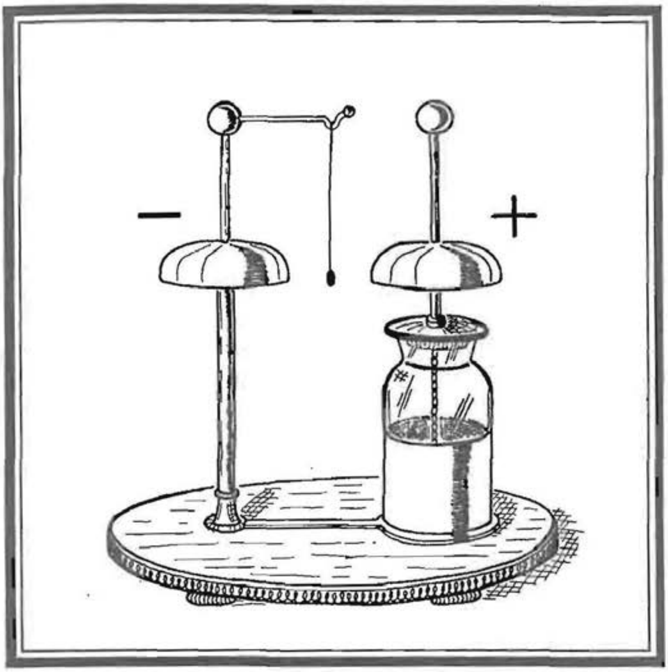

FIGURE 2

In a later modification of Gordon's invention one of the two bells was mounted on a Leyden jar. The bells would then ring for as long as there was enough electrical energy stored in the jar to move the clapper.

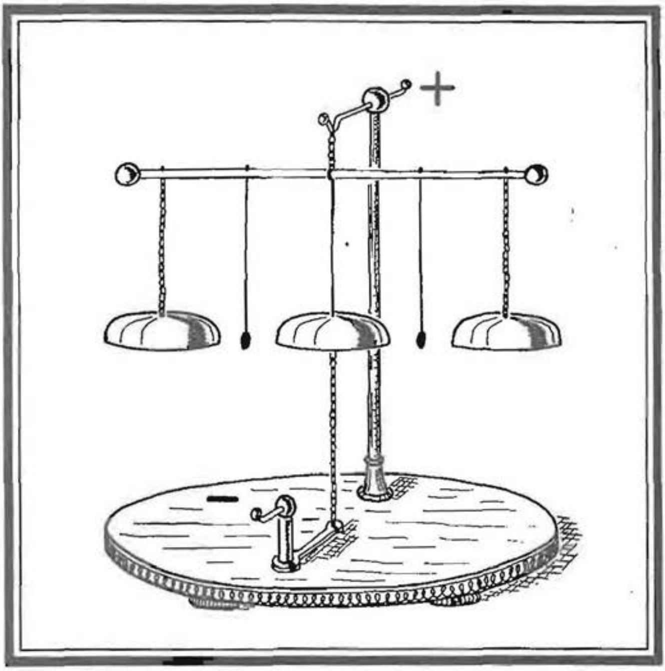

FIGURE 3

Sometimes Gordon's bells were made as a set of three bells and two clappers, all suspended from a horizontal bar, the central bell being insulated from the bar and grounded by means of a light chain.

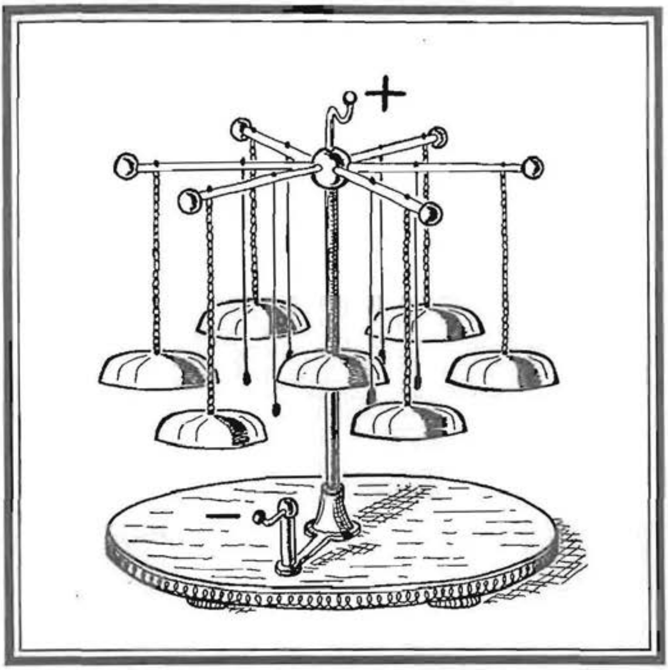

FIGURE 4

Another modification of Gordon's bells was a set of bell chimes arranged on a wooden base in a circular formation around the central bell. A separate clapper was then present for each of the outside bells.

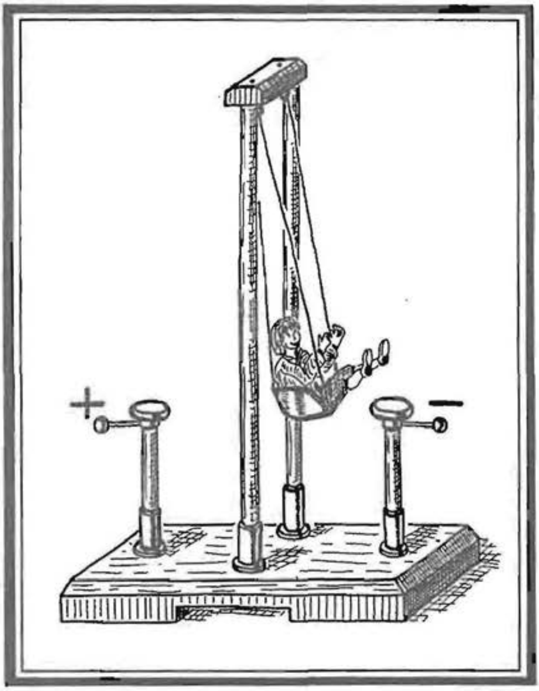

FIGURE 5

The electrical and mechanical principle of Gordon's invention could be used not only for ringing bells but for creating a mechanical motion in general. In this "electric swing" this principle was used to operate an animated toy.

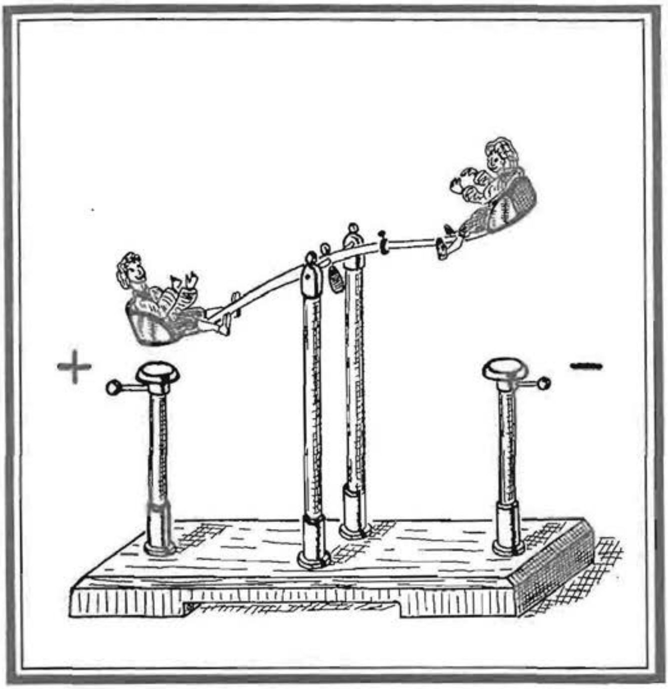

FIGURE 6

Various electrostatically operated toys were invented in the 18th and 19th centuries. This "electric seesaw" was basically a conducting bar oscillating about a horizontal axis on the same principle as the clapper of Gordon's bells.

of three bells and two clappers, all suspended from a horizontal bar, the central bell being insulated from the bar and grounded by means of a light chain (Fig. 3). Another modification of Gordon's bells was a set of bell chimes arranged in a circular formation around the central bell (Fig. 4); a separate clapper was then present for each of the outside bells, each clapper moving between one of the latter and the central bell (the central bell was usually grounded, the outside bells were connected to an electrostatic generator).

The electrical and mechanical principle of Gordon's invention could be used, of course, not only for ringing bells but also for a variety of other purposes. In particular, this principle was later used in two electrically operated toys: the "electric swing" and the "electric seesaw", which utilized, respectively, an insulated pendulum made as a swing (Fig. 5) and an insulated conducting swing bar capable of oscillating about a horizontal axis (Fig. 6).[3,4]

Except for Franklin's bells, all of the above devices were designed for operation from an electrostatic generator or from Leyden jars charged by such a generator. However, in 1806 a high-voltage chemical battery ("dry pile", later known as the "Zamboni pile") was invented by Georg Behrens, and in 1810 this pile was adapted by Giuseppe Zamboni to operate primitive electrostatic motors.[5] Inasmuch as a Zamboni pile could remain active for many years, the little motors that operated from it were known as the "electric perpetuum mobile". An example of such a perpetuum mobile is shown in Fig. 7 (see Chapter 4 for another design). The apparatus consisted of a light rigid insulated pendulum, pivoted just above the center of gravity, with a light conducting ring at the top. The ring was located between the two knobs of a Zamboni pile. When the ring contacted one of the knobs, it acquired a charge from the knob and was repelled from this knob and attracted to the second knob, and so on, just like the clapper in Gordon's bells. One such perpetuum mobile is reported to have operated without interruption for at least 86 years.[6]

The most sophisticated device derivable from Gordon's electric bells was probably the reciprocating electrostatic motor (Fig. 8 and Plate 1) built in 1880 by Howard B. Dailey and Elijah M. Dailey.[7] The motor, which is now at the Museum of History and Technology of the Smithsonian Institution, was described by one of its builders as follows:

"This machine operates by the direct action of static electric attractions and repulsions. It is constructed entirely of fine wood, glass and hard rubber, there being no magnetic materials used. The flywheel is of laminated, soft wood and runs in journal bearings of very small diameter. The moving balls, mounted on the walking beam of vulcanite, are made of wood, hollowed out so that the walls are about 2 millimeters thick.

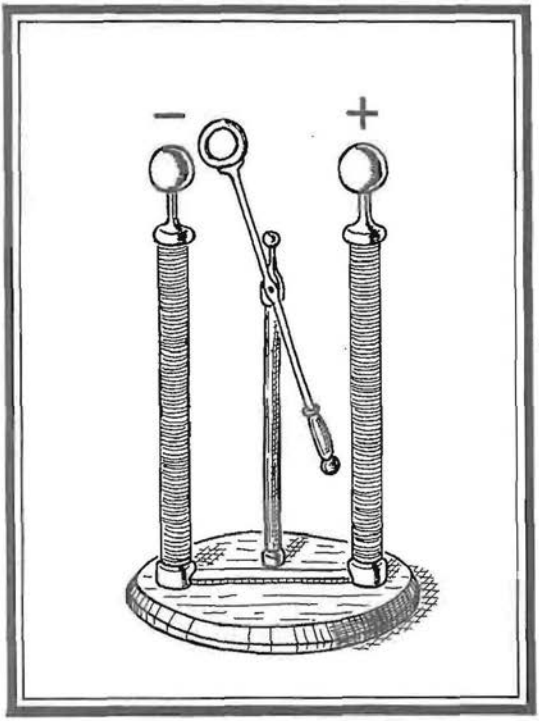

FIGURE 7

This electric pendulum operated from a Zamboni pile, an early high-voltage battery. Since the pile remained active for years, the pendulum was known as an "electric perpetuum mobile".

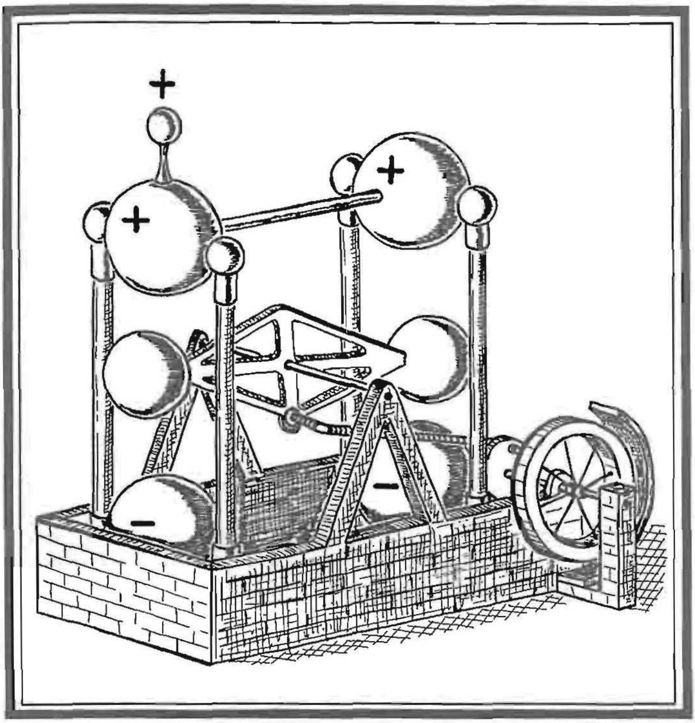

FIGURE 8

The most sophisticated device derivable from Gordon's bells was probably this reciprocating electrostatic motor built in 1880 by H. B. Dailey and E. M. Dailey.

They are covered with aluminum foil for static conductivity. The stationary balls are of solid wood.

"To operate the engine the stationary balls are charged with electricity from a static electric generator, such as a Holtz machine, the upper balls being connected through the brass ball to one pole of the machine while the lower stationary balls are connected through the binding post on the bed frame to the opposite pole of the machine. Under proper conditions, when charged, the engine will make about 375 revolutions per minute.

"The walnut base upon which the engine is mounted is 14" long, 4" wide and 1¾" thick. The movable balls are about 1½" in diameter; the upper stationary balls are 1¾" in diameter; and the lower stationary balls, 1½". The four glass rods, mounted vertically, are about 6" high and spaced 6" apart along the bed. The diameter of the flywheel is 5¾". It is gilded and has small wire spokes. The connecting rod is 7" in length."*

The electric principle of Gordon's bells could be used also for producing a continuous unidirectional motion. A fascinating device of this type was the "electric racing ball", or the "electric planetarium", an excellent description of

---

*The author is grateful to Mr. Elliot N. Sivowitch, Smithsonian Institution Museum Specialist, for kindly providing this description and the photograph of Dailey's motor.

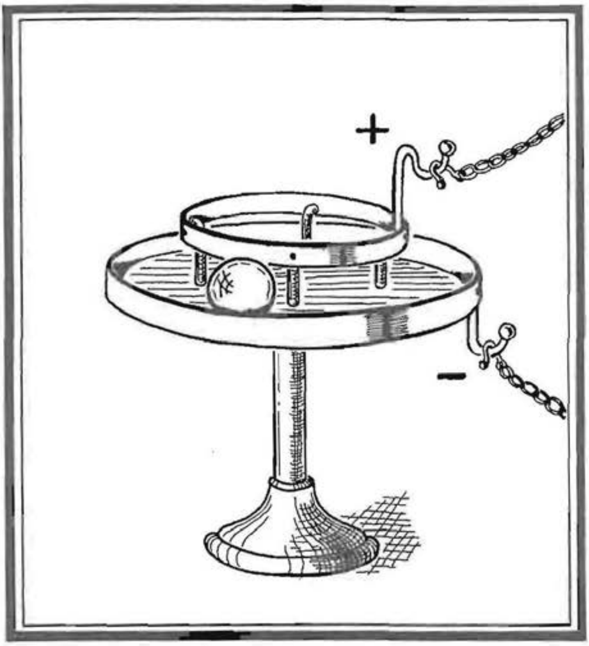

FIGURE 9

The electric principle of Gordon's bells could be used also for producing a continuous unidirectional motion. This "electric racing ball" was propelled along a conducting hoop by the same mechanism that moved the clapper in Gordon's bells.

which has been provided by M. Guyot.[8] The apparatus consisted of an insulated horizontal conducting hoop positioned above a conducting disk having a vertical rim concentric with the hoop (Fig. 9). The arrangement constituted a track for a light sphere of glass placed on the disk between the hoop and the rim. When the hoop and the disk were connected to an electrostatic generator, the sphere ran along this track due to attraction and repulsion exerted by the hoop and the disk upon the points of the sphere that acquired charges from contact with the disk and the hoop (the rolling of the sphere usually had to be initiated with a light push by hand, but once the sphere had started to run, it continued running for as long as the electrostatic generator was in operation).

Since the moving elements in the above described motors were charged by contact, the motors may be classified as "contact motors". This method of charging, however, is neither the only possible nor the most expedient one. Therefore contact motors will probably always be subordinate to most of the motors described in the following chapters. In this connection it may be useful to note that, upon a closer examination of the operation of Gordon's bells and other similar devices, one recognizes that the charging of the moving elements in these devices usually occurs by means of a spark that jumps from a stationary electrode (bell) to a moving element (clapper) as soon as the latter comes close enough to the former. Thus the actual physical contact between the moving and stationary components in these devices is not at all necessary for their operation. In fact, such a contact is harmful insofar as it results in energy losses from impact and friction.

REFERENCES

1. P. Benjamin, *A History of Electricity*, John Wiley and Sons, New York, 1898, pp. 506, 507.
2. J. Sparks, Ed., *The Works of Benjamin Franklin*, Whittemore, Niles, and Hall, Boston, 1856, Vol. 5, p. 301.
3. F. M. Avery, *Modern Electricity and Magnetism*, Sheldon and Company, New York, 1885, p. 241.
4. A. Neuburger, *Ergötzliches Experimentierbuch*, Ullstein & Co., Berlin, 1920, pp. 434-436.
5. Gilbert, *Annalen der Physik*, Ser. 1, Vol. 49, pp. 35-46, (1815).
6. Müller-Pouillet, *Lehrbuch der Physik und Meteorologie*, Friedrich Vieweg und Sohn, Braunschweig, 1909, 10th edn., Vol. 4, p. 338.
7. H. B. Dailey, *Modern Electrics*, Vol. 5, pp. 916-917 (1912).
8. M. Guyot, *Nouvelles Recreations Physiques et Mathematiques*, Gueffier, Paris, 1786, Vol. 1, pp. 272-274 and Plate 27.

3

ELECTRIC WIND MOTORS

As already mentioned in the preceding chapter, Andrew Gordon invented two electric motors. His second motor was the "electric fly",[1,2] also known as the "electric whirl", "electric pinwheel", or "electric reaction wheel" (Fig. 10). The electric fly consisted of one or more light metal arms with sharp-point ends bent at right angles to each arm and in the same circumferential direction. The fly was pivoted at its center on an insulated needle. When the needle was connected to an electrostatic generator, a corona discharge* took place from the sharp points of the fly. The air near these points became charged with charges of the same polarity as that of the fly and was then repelled from the points due to repulsion of like charges (the resulting motion of the air is known as the "electric wind"). Similarly, the points themselves were repelled from the charges in the air, and the fly rotated therefore in the opposite sense to that in which the points were directed.

---

*Corona discharge is a spontaneous electric conduction in a gas (air) originating from sharp conducting bodies whose electrostatic potential relative to the ground is 3000 volts or higher.

Also the electric fly was used for ringing bells or chimes. This was accomplished by suspending from the fly a clapper, which, as the fly rotated, struck in turn several bells positioned in a circular formation underneath the fly (Fig. 11).[3]

In 1760, Hamilton, professor of philosophy at Dublin, suggested to use a similar device as an electrometer.[4] In this electrometer small weights were suspended from the fly. As the fly turned, the weights were deflected from the vertical direction by centrifugal forces. This deflection was a measure of the fly's speed and, hence, of the strength of the source to which the fly was connected. A more sophisticated electrometer utilizing the electric fly was described by Jakob Langenbucher,[5] a silversmith at Augsburg, Germany, who spent a large part of his time and fortune on improving electrostatic generators and studying electrostatic phenomena. In this electrometer the fly was mounted on a thin vertical cylinder. A fine string was fastened to the cylinder. The string passed over a small pulley and had a light pan attached to its free end. A small weight was placed on the pan. As the fly turned, the string

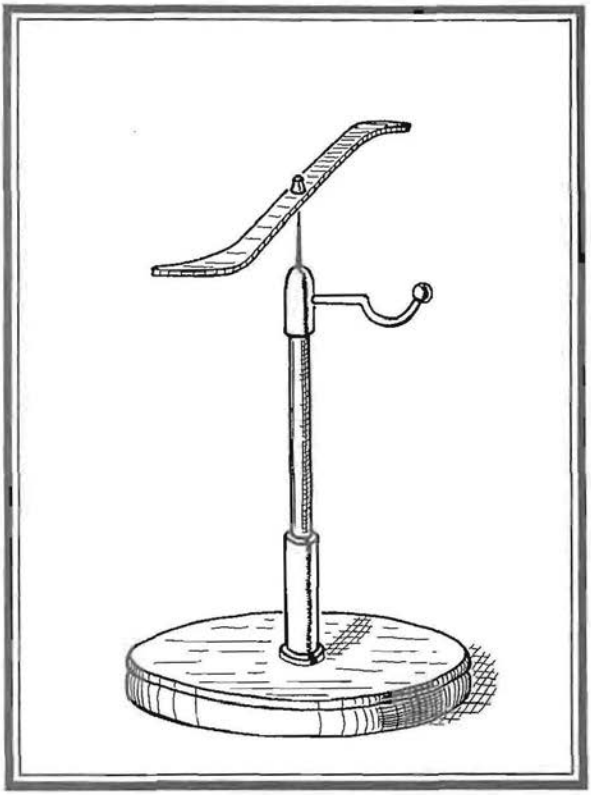

FIGURE 10

Andrew Gordon invented two electric motors. His second motor was the so-called "electric fly".

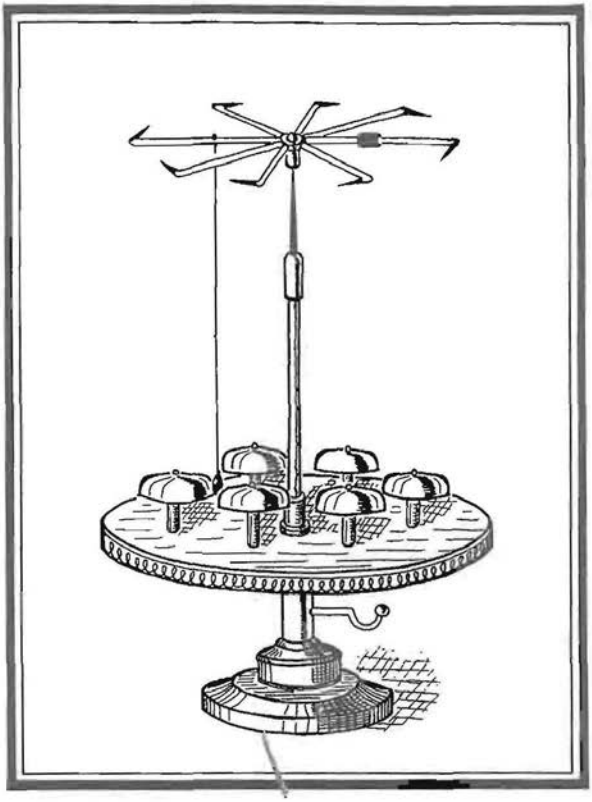

FIGURE 11
Also the electric fly was used for ringing bells with the aid of a clapper suspended from it.

wound itself on the cylinder lifting the pan and the weight. The maximum weight that could be lifted by this electrometer was a measure of the strength of the source from which the fly was operated.

The principle of the electric fly was later used to operate an orrery depicting the orbital motions of the Sun, the Moon, and the Earth (Fig. 12).[6] In such an orrery, also known as the "electrical tellurium", the longer bent arm carried a large sphere at one end, while at the other end it carried a similar but smaller bent arm, each end of which also carried a small sphere. The three spheres represented the Sun, the Moon, and the Earth. The arms were pivoted on sharp points, and each arm had a perpendicular sharp point in the horizontal plane. When the instrument was connected to an electrostatic generator, the two systems rotated about their points of support, the smaller system, being lighter, rotated more rapidly than the larger one. The relative masses of the spheres were adjusted to make the Earth-Moon system rotate twelve times for every single rotation of the Sun.

An interesting variation of the electric fly was constructed in 1761 by Ebenezer Kinnersley, a Philadelphia school teacher and close friend of Benjamin Franklin. Kinnersley called his device the "electrical horse-race".[7] This was a cross formed by two pieces of wood of equal length (probably about 18 inches long) supported horizontally on a central pin. A light figure of horse

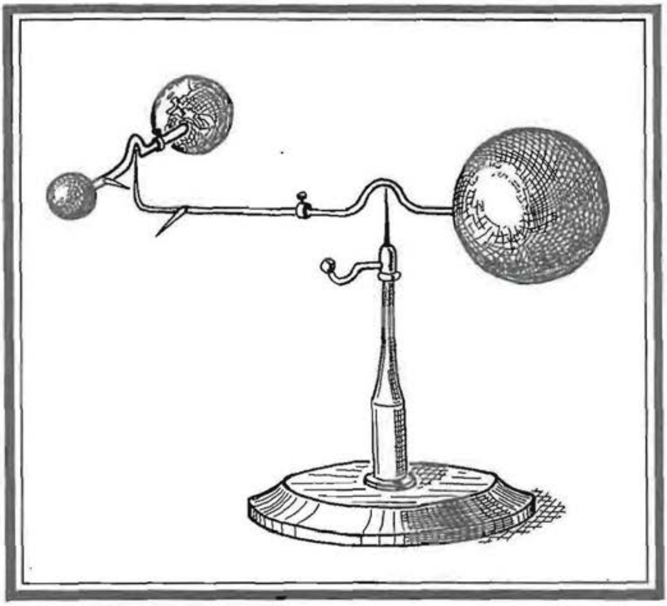

FIGURE 12

The principle of the electric fly was later used to operate an orrery depicting the orbital motions of the Sun, the Moon, and the Earth. The masses of the spheres were adjusted to make the Earth-Moon system rotate twelve times for every single rotation of the Sun.

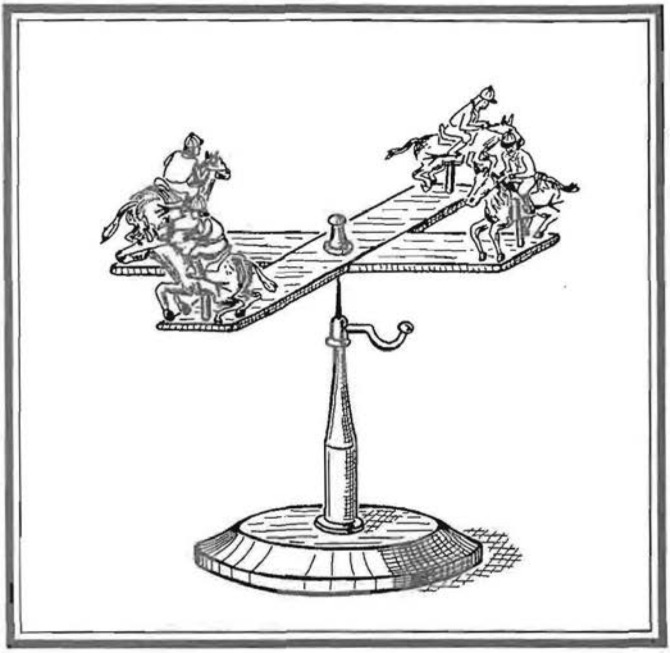

FIGURE 13

An interesting variation of the electric fly was constructed in 1761 by Ebenezer Kinnersley, a friend of Benjamin Franklin. Kinnersley called his device the "electrical horse-race". The electrical horse-race was a popular toy in the latter part of the 18th century.

with his rider was placed upon each extremity of the cross. The motion was produced by corona discharge from the spurs of the riders (Fig. 13). The electrical horse-race was a popular toy in the latter part of the 18th century. Interesting descriptions of this toy have been provided by Langenbucher (Ref. 5, p. 206 and Plate 8), and by Guyot.[8]

To show more convincingly that the electric fly converted electrical energy into mechanical energy, electric flies were placed on horizontal axes and were arranged to lift small weights (Fig. 14) or to climb inclined rails (Fig. 15).[9]

Operation of the electric fly in liquids was investigated by Aimé.[10] A historical account of electric fly studies was presented by Tomlinson.[11] The mechanism of the fly was discussed at some length by Mascart.[12] A quantitative study of the forces acting upon the electric fly was performed by Kämpfer.[13]

A more efficient modification of the electric fly was proposed in 1887 by Bichat.[14] Bichat's electric "tourniquet" consisted of a light vertical frame capable of turning about a vertical axis and supporting two thin vertical wires producing a corona discharge along their entire length. The vertical sides of the frame acted as "spoilers" preventing a corona discharge from the "back side" of the wires. A simple motor based on Bichat's device is shown in Fig. 16.

In recent years the principle of the electric fly (reaction force experienced by a sharp point as a

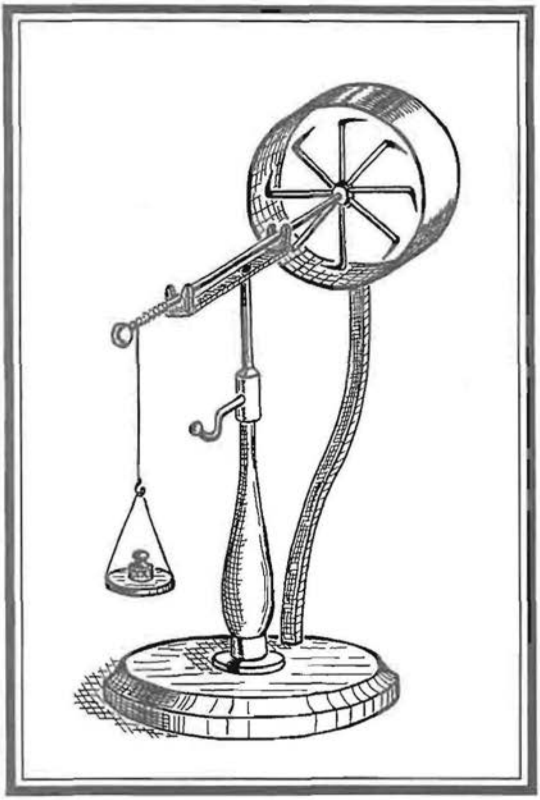

FIGURE 14

Electric fly converted electrical energy into mechanical energy and could lift small weights.

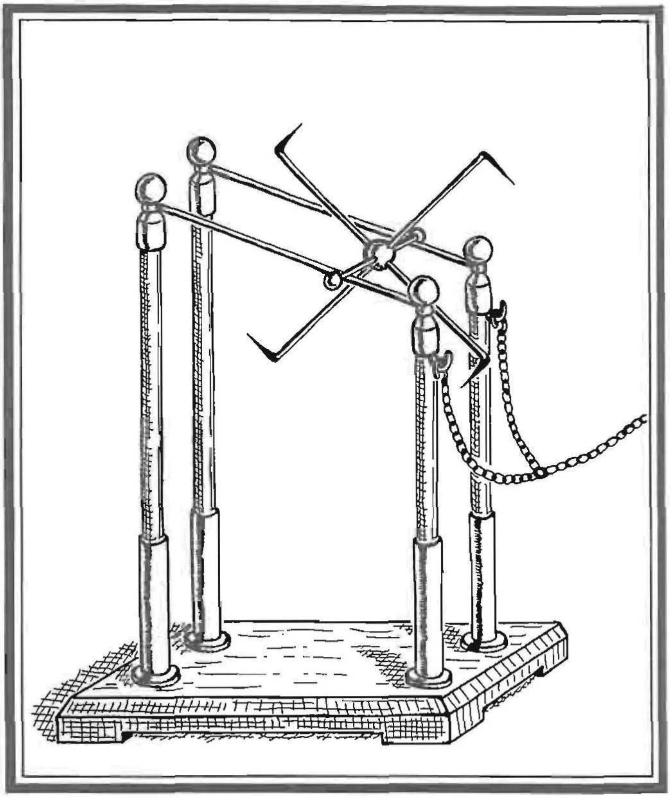

FIGURE 15

Electric fly developed enough mechanical energy for climbing inclined rails.

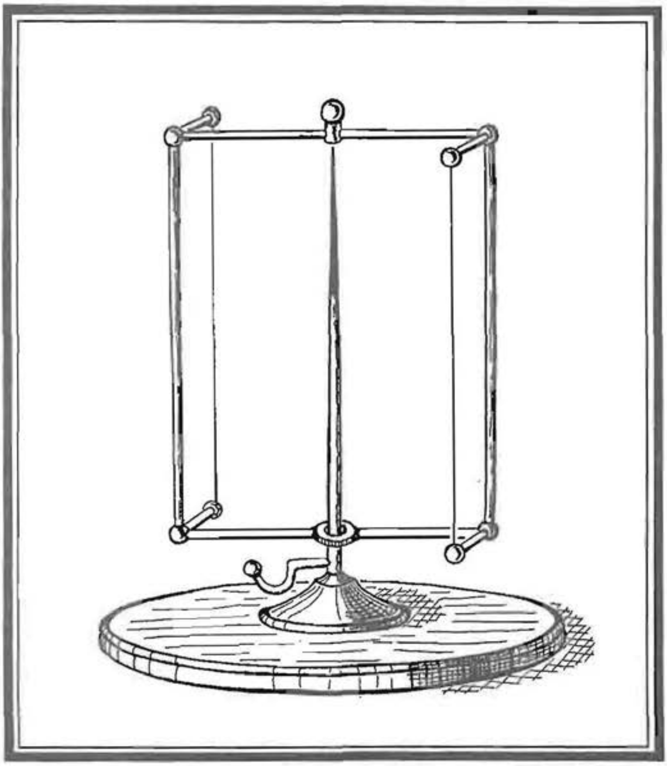

FIGURE 16

A more efficient modification of the electric fly was proposed in 1887 by Bichat, who called his device the "electric tourniquet".

result of a corona discharge from it) has been suggested for a flying machine ("corona-kraft levitation").[15] The lift realizable from such a machine does not, however, appear to be of practical significance.[16]

Closely related to the electric fly were the "electric wind wheels" and the "electric wind turbines". Like the electric fly, they were invented in the 1740's and utilized a corona discharge from sharp points for their operation. Benjamin Franklin studied these devices in 1747 (he first learned of them from his friend Philip Syng). According to Franklin (Ref. 7, p. 184) they were "light windmill-wheels made of stiff paper vanes; also. . .little wheels, of the same matter, but formed like water-wheels". Their probable appearance was as shown in Figs. 17 and 18. Later versions of these devices have been described by Guyot (Ref. 8, pp. 274, 275 and Plate 27), by Langenbucher (Ref. 5, pp. 75, 76 and Plate 3), by Gale,[17] and by Neuburger.[18]

A delightful electric toy using the electric wind for its propulsion was made by W. R. King, an associate of James Wimshurst (the designer of the electrostatic machine carrying his name). This toy, which may be called an "electrical rabbit hunt", now at the Science Museum in London, is shown in a simplified form in Fig. 19. It consists of a small electric wind turbine mounted on a vertical shaft carrying five light metal arms on its upper end; the arms support

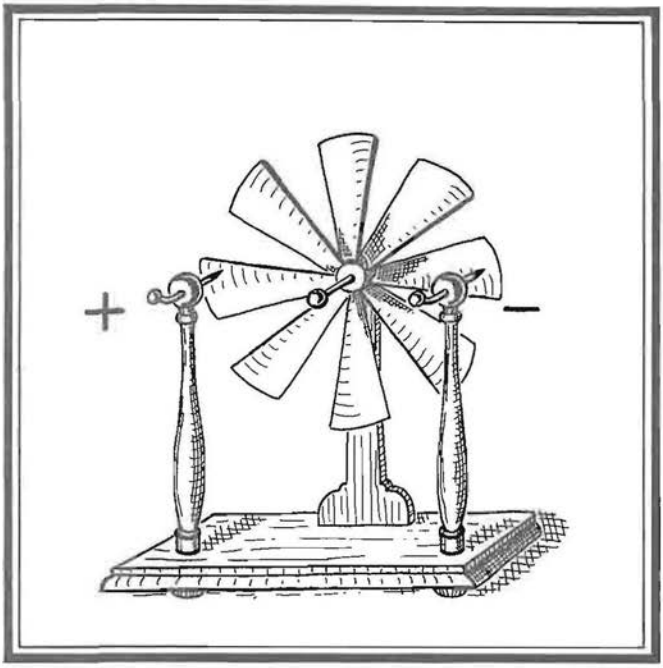

FIGURE 17

Closely related to Gordon's electric fly was the "electric wind wheel". Like the electric fly, it was invented in the 1740's and utilized a corona discharge from two sharp points for its operation.

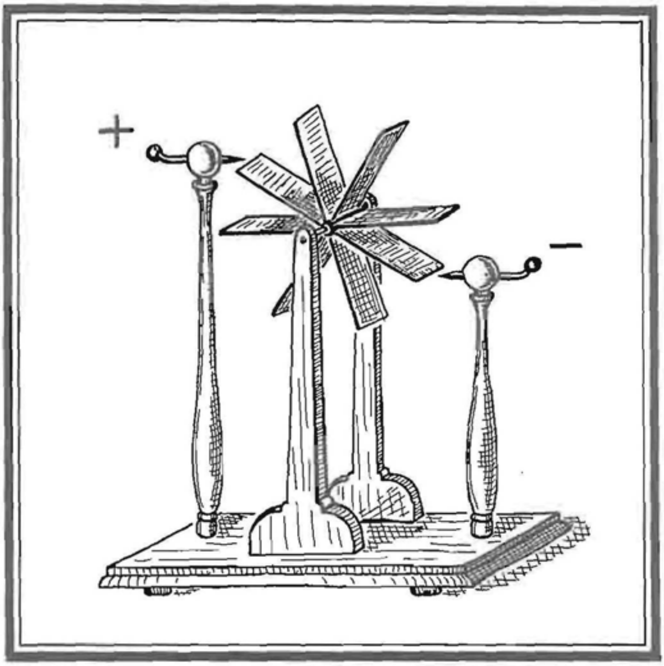

FIGURE 18

Benjamin Franklin studied this "electric wind turbine" in 1747. He concluded that its action depended not as much on the electric wind as on the forces between the charges present on its moving and stationary parts.

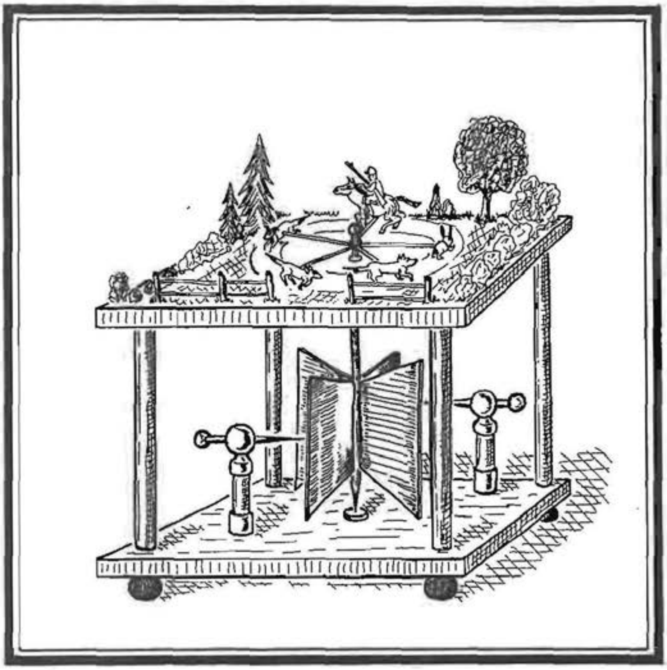

FIGURE 19

A delightful electric toy utilizing the electric wind for its propulsion was made by E. R. King, an associate of James Wimshurst. This toy, which may be called an "electrical rabbit hunt", is now at the Science Museum in London.

figurines of a rabbit, three dogs, and a horseman, who chase each other in a beautiful little garden, as the arms turn powered by the turbine below.

In studying the electric wind motors, Benjamin Franklin concluded that their action depended not as much on the electric wind as such, but on the repulsion and attraction of electric charges present on the moving and stationary parts of the motors (Ref. 7, p. 184). It is very probable that this conclusion led him to the invention of his big electric motors, the famous "electrical wheels", which are described in the next chapter.

REFERENCES

1. P. Benjamin, *A History of Electricity*, John Wiley and Sons, New York, 1898, pp. 506, 507.
2. J. C. Poggendorff, *Geschichte der Physik*, Johann A. Barth, Leipzig, 1879, p. 846.
3. V. K. Chew, *Physics for Princes*, Her Majesty's Stationery Office, London, 1968, p. 13.
4. B. Wilson, *Philosophical Transactions*, Vol. 51, p. 896 (1760).
5. J. Langenbucher, *Beschreibung einer beträchtlich verbesserten Elektrisiermaschine*, Mathäus Reigers, Augsburg, 1780, pp. 54-56 and Plate 3.
6. G. Adams, *An Essay on Electricity*, London, 1799, 5th edn., pp. 580-581 and Plate 4.

7. J. Sparks, Ed., *The Works of Benjamin Franklin*, Whittemore, Niles, and Hall, Boston, 1856, Vol. 5, p. 371.
8. M. Guyot, *Nouvelles Recreations Physiques et Mathematiques*, Gueffier, Paris, 1786, Vol. 1, pp. 249, 250 and Plate 25.
9. H. J. Oosting, *Zeitschrift für den Physikalischen und Chemischen Unterricht*, Vol. 9, pp. 84-85 (1896).
10. R. Aimé, *Annales de Chimie et de Physique*, Ser. 2, Vol. 62, p. 421 (1833).
11. C. Tomlinson, *Philosophical Magazine*, Ser. 4, Vol. 27, p. 202 (1864).
12. M. E. Mascart, *Traité d'Electricité Statique*, G. Masson, Paris, 1876, Vol. 1, pp. 175-179.
13. D. Kämpfer, *Annalen der Physik*, Ser. 3, Vol. 20, pp. 601-614 (1887)
14. E. Bichat, *Annales de Chimie et de Physique*, Vol. 12, pp. 64-79 (1887).
15. A. Christenson and S. Moller, *AIIA Journal*, Vol. 5, pp. 1768-1733, (1967).
16. L. B. Loeb, *Electrical Coronas*, University of California Press, Berkley, 1965, pp. 402-406.
17. L. D. Gale, *Elements of Natural Philosophy*, Mark H. Newman, New York, 1846, p. 164.
18. A. Neuburger, *Ergötzliches Experimentierbuch*, Ullstein & Co., Berlin, 1920, pp. 432, 433.

4

FRANKLIN'S SPARK MOTORS AND THEIR DESCENDANTS

If one defines the "electric motor" as a machine capable of converting substantial amounts of electrical energy into mechanical energy, then it was Benjamin Franklin who invented the electric motor, for he was the first person to design and to construct a machine capable of accomplishing such a conversion.

Franklin built two electric motors, which he called "electrical wheels". He described them in a letter to Peter Collinson, Fellow of the Royal Society of London, written in Philadelphia in 1748.[1]

His first motor was a big machine, about 40 inches in diameter. The main part of this motor was a wooden disk mounted horizontally on a vertical wooden axle and carrying 30 glass spokes with brass thimbles on their ends (Fig. 20 and Plate 2). The axle turned on a sharp point of

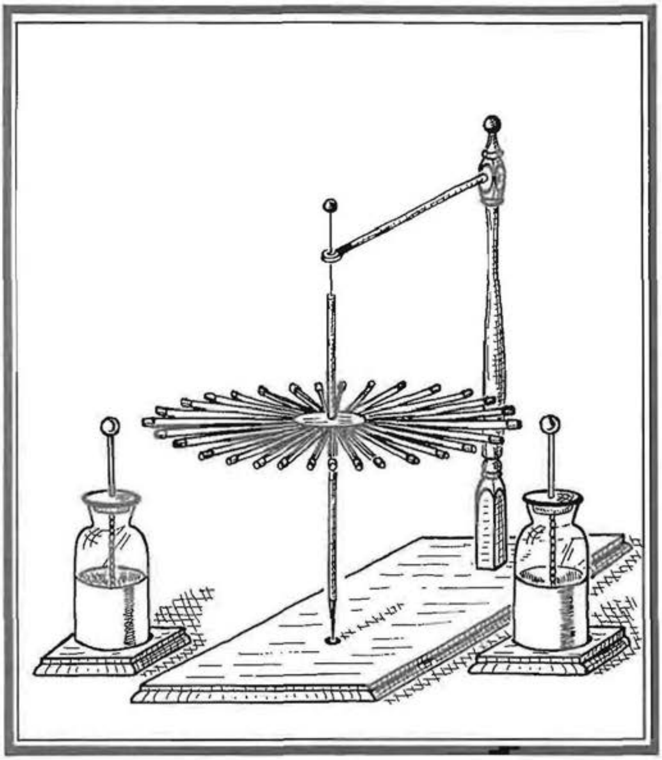

FIGURE 20

Franklin's first "electrical wheel" was the earliest device converting substantial amounts of electrical energy into mechanical energy.

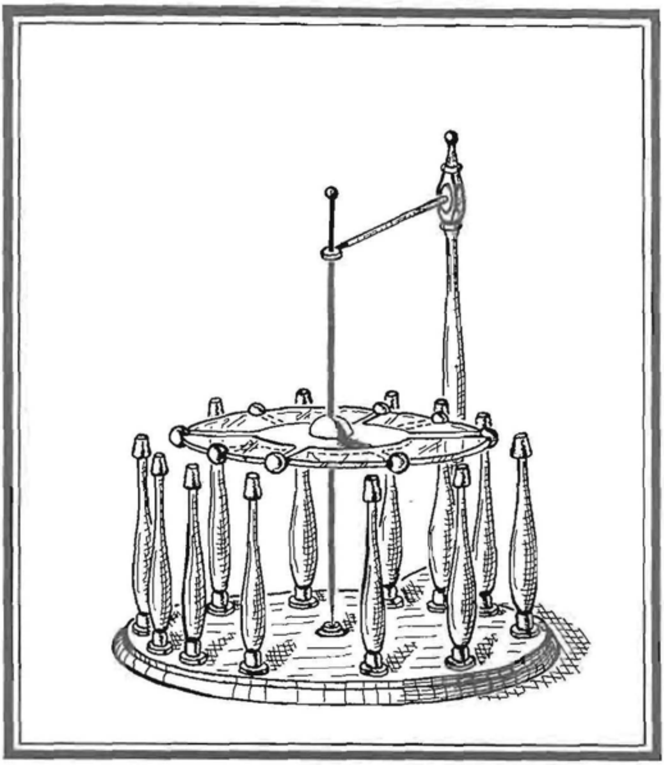

FIGURE 21

Franklin's second "electrical wheel" was more efficient than the first. It was powered by the electric charge stored in its capacitor-rotor.

iron fixed in its lower end and was kept in the vertical position by a strong wire mounted in its upper end and passing through a small hole in a thin stationary brass plate. Two oppositely charged Leyden jars (bottles) were placed in close proximity to the thimbles on the spokes, so that when the wheel turned, the thimbles barely missed the knobs of the jars.

As the thimbles moved past the Leyden jars, a spark jumped between a jar knob and the passing thimble, charging the latter with a charge of the same polarity as that of the knob. Therefore each knob attracted the oncoming thimbles ("opposite charges attract") and repelled the departing thimbles ("like charges repel"), causing the motor to turn. The motor, being perfectly symmetric, could turn in either direction, and usually required a starting push by hand to initiate the charging of the thimbles.

Here is how Franklin himself described the operation of his motor:[1]

"If now the wire of a bottle, electrified in the common way, be brought near the circumference of this wheel, it will attract the nearest thimble, and so put the wheel in motion; that thimble, in passing by, receives a spark, and thereby being electrified, is repelled, and so driven forwards; while a second being attracted, approaches the wire, receives a spark, and is driven after the first, and so on till the wheel has gone once round, when the thimbles before electrified approaching the wire, instead of being attracted as they were at first, are repelled, and the motion presently ceases. But if another bottle, which had been charged through the coating (e.i. charged oppositely to the first, O. J.), be placed near the same wheel, its wire will attract the thimble repelled by the first, and thereby double the force that carries the wheel round; and, not only taking out the fire (electric charge, O. J.) that had been communicated to the thimbles by the first bottle, but even robbing them of their natural quantity, instead of being repelled when they come again towards the first bottle, they are more strongly attracted, so that the wheel mends its pace, till it goes with a great rapidity, twelve or fifteen rounds in a minute, and with such strength, as that the weight of one hundred Spanish dollars, with which we once loaded it, did not seem in the least to retard its motion".

Franklin was not quite pleased with this motor, however, because it required "a foreign force, to wit, that of the bottles" for its operation. He constructed therefore a second, "self-moving", motor, which itself contained the electrical energy to be converted into the mechanical motion.

The second motor was made as follows (Fig. 21 and Plate 3). A disk of window glass, 17 inches in diameter, was gilded on each side, except for narrow areas next to the edge. Several lead spheres (bullets) were fixed to the edge of the disk and were connected in alternation to the two gold layers. The disk was mounted on a vertical axle consisting of two pieces of strong wire, each about 9 inches long; the upper wire being in contact with the top layer of gold, the lower wire being in contact with the bottom layer. The lower wire terminated in a sharp point resting on a piece of brass cemented within a glass salt-cellar. The upper wire passed through a hole in a thin brass plate cemented on a long strong piece of glass. This wire carried a small ball on its upper end for preventing a corona discharge from it. Several glass pillars with thimbles on their tops were positioned in a circle around the disk, so that the spheres on the edge of the disk barely missed the thimbles when the disk was in motion. The motor was powered by electricity stored in the disk itself, which, due to the two layers of gold, constituted a capacitor capable of storing an appreciable quantity of charge (to charge the rotor, the two layers of gold were temporarily connected to an electrostatic friction generator).

The motor operated on the same principle as Franklin's first motor, except that now the thimbles were stationary and received positive or negative charges from the passing spheres, rather than the other way around. As in the first motor, the rotation persisted as long as there was enough charge stored in the capacitor (rotor) to produce sparks between the spheres and the thimbles.

Franklin reported that his second motor

(which was more powerful than the first one) could make 50 turns in a minute and could run for up to 30 minutes on a single charging of the rotor.[1] Experiments with replicas of Franklin's motors performed by the author indicate that the original motors developed a power of the order of 0.1 watt,[2,3] which was a very impressive achievement for that very early period in the history of electric science.

Several authors have described simplified versions of Franklin's motors.[4,5] Miniature Franklin-type motors have been constructed and described by A. D. Moore[6] (Fig. 22). One such motor consists of a Plexiglas disk about 4 inches in diameter with small conducting balls embedded in the rim. The disk turns on a horizontal axis supported by a base behind the disk. The motor runs when placed between two conductors connected to an electrostatic generator. To make the motor self-starting and one-way running, the conductors may be provided with short auxiliary rods tangential to the disk; since the rods dispatch sparks predominantly in the direction of their free ends, the disk also turns in this direction. Another little motor of the same type designed by Moore is shown in Fig. 23.

An obvious improvement of Franklin's motors would be to use a cylindrical rotor with electrodes in the shape of long strips. Such electrodes would accept more charge than thimbles or spheres, and the power of the motor would be increased considerably. Motors of this type

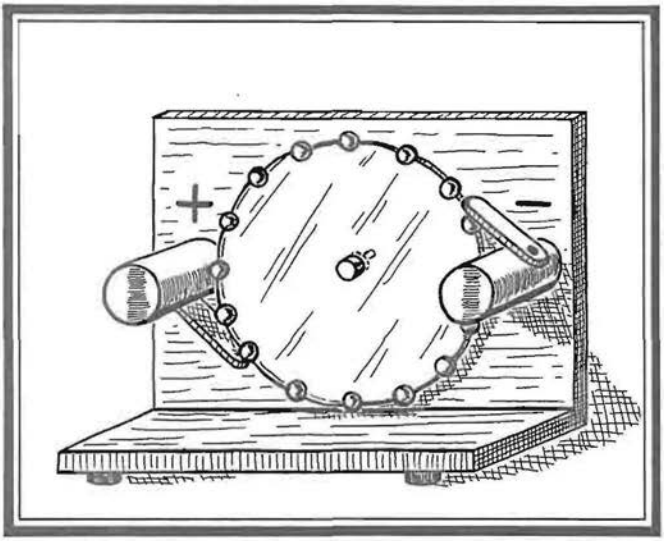

FIGURE 22

Several authors have described simplified versions of Franklin's motors. One such motor consists of a Plexiglas disk about 4 inches in diameter with small conducting balls embedded in the rim. The disk turns on a horizontal axis supported by a base behind the disk.

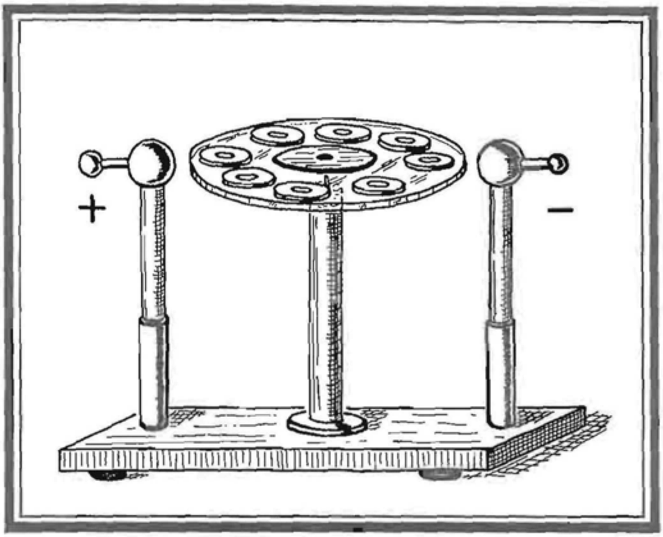

FIGURE 23

This inexpensive miniature Franklin-type motor of recent design can be used for lecture-room demonstrations of electrostatic forces. The rotor disk can be made from cardboard, wood, or plastic. For the rotor electrodes one can use thumbtacks or small aluminum foil disks.

have indeed been constructed. Here is a description of a "bell jar" motor taken from a book published in 1847 (Fig. 24):

"On a pivot, *a*, suspend a bell jar having four pieces of tinfoil pasted on its sides, *b*, *c*, *d*; connect the jar, by means of the insulated wire, *y*, with the prime conductor (high-voltage terminal of an electrostatic friction generator, O. J.) so that the pieces of tinfoil may receive sparks. On the opposite side arrange a conductor, *x*, in connection with the ground by a chain. On putting the machine into activity, the jar will commence rotating on its pivot."[7]

Small motors closely related to both Franklin's first motor and to pendulum motors were used in the rotary-type "perpetuum mobile" (see Chapter 2) operated from a Zamboni pile.[8] The main part of these motors (Fig. 25) was a light insulated arm pivoted on a sharp point and carrying a piece of tinsel on each end. The tinsels could just touch the knobs of the pile when the arm was moving on its pivot. The device functioned in the same manner as Franklin's first motor, the two tinsels corresponding to the thimbles of the latter.

An analogous design has recently been described as a "millifleapower motor" operating from a nuclear battery.[9] In this motor a light aluminum vane with pointed ends (Fig. 26) bent so as to form an inverted V is mounted at its midpoint on the tip of a sharp needle protruding from the positive terminal of the battery. As the

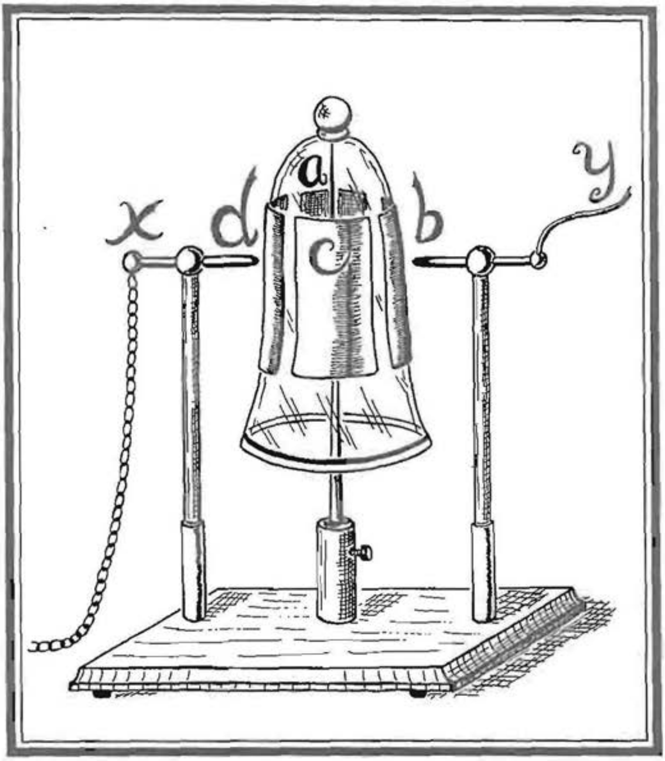

FIGURE 24

This "bell jar" Franklin-type motor had a cylindrical rotor with tinfoil electrodes in the shape of long strips.

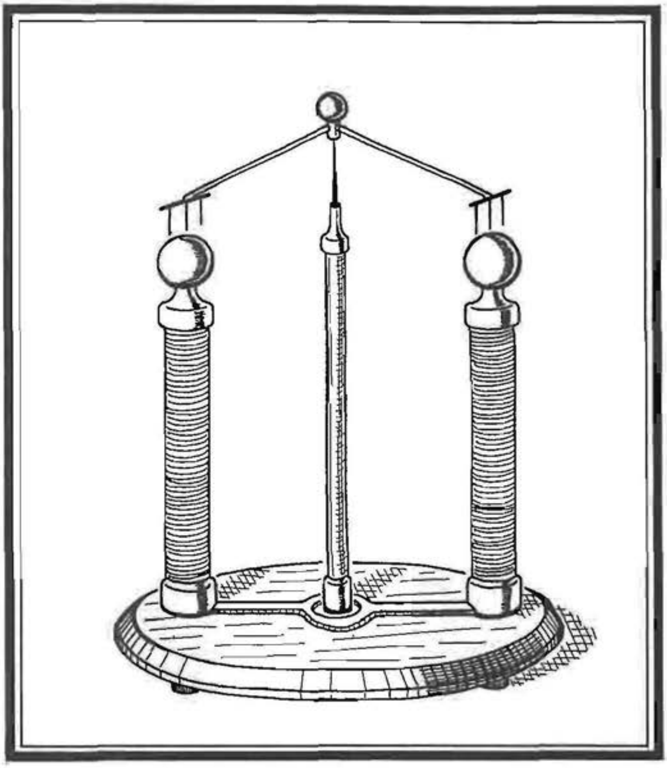

FIGURE 25

This small motor was used as a rotary-type "perpetuum mobile" operating from a Zamboni pile. Such motors turned for years without stopping.

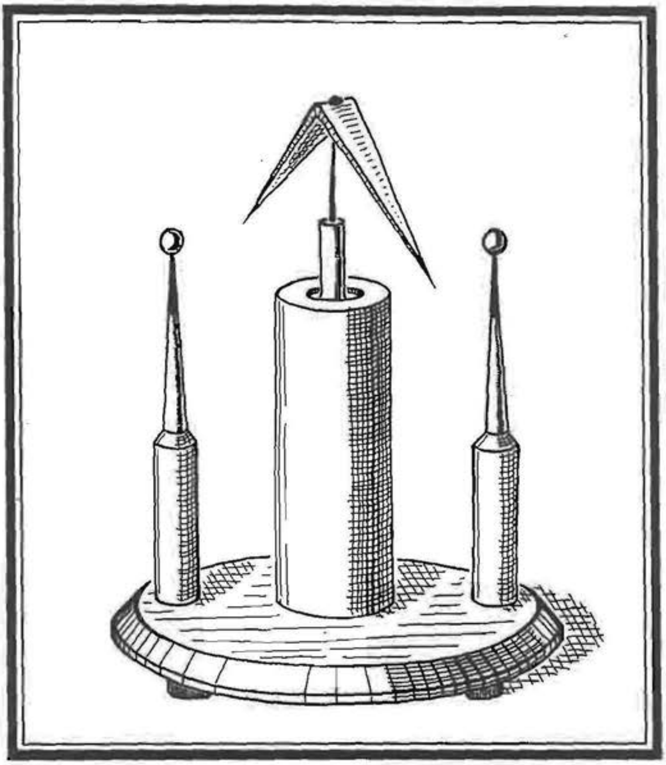

FIGURE 26

An analogous design has recently been described as a "millifleapower motor" operating from a nuclear battery.

vane turns, its ends come close to two spheres connected with the ground through a high resistance and become attracted by them. A discharge from the ends of the vane charges the two spheres, and the spheres then repel the ends of the vane. During the time it takes the vane to complete one half of its revolution, the charges of the spheres leak off to the ground, so that the spheres can again attract the ends of the vane, accept new charges from them, and then again repel them. The motor is expected to run for many years.

An interesting motor derivable from Franklin's first motor and from pendulum motors was suggested for operation from an electrophorus.[10] The main part of this motor (Fig. 27) was a light conducting ring supported in a vertical plane on an insulated needle. In front of the ring there were two insulated spheres, which were so positioned that the ring just missed them as it turned on its needle. The spheres were either charged oppositely by the electrophorus, or only one sphere was charged, in which case the second sphere was grounded. The device operated on the same principle as Franklin's first motor.

Of all the motors discussed above, Franklin's original motors and the bell jar motor were the most powerful ones. One can make such motors even more powerful by increasing the number of moving and stationary electrodes. But this can be done only up to a limit, because if the electrodes are too close to one another, sparks will

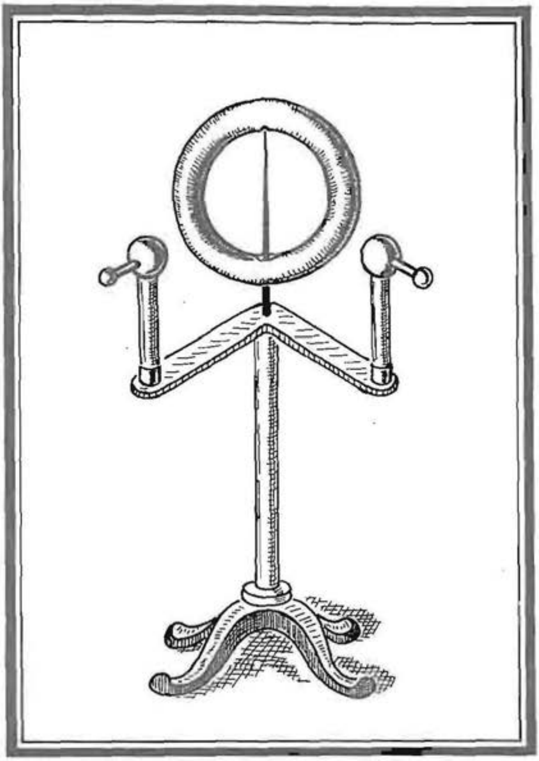

FIGURE 27

This motor derivable from Franklin's first motor and from pendulum motors was suggested for operation from an electrophorus.

jump all around from one electrode to the next, thus short-circuiting the motor. The motor will then not work at all, unless the operating voltage is lowered. But by lowering the voltage one also lowers the power, so that the gain obtained by increasing the number of electrodes becomes offset by the decrease in the operating voltage. Franklin type motors have therefore an inherent limitation in their power characteristics, and the fact that they depend for their operation on the presence of many metallic electrodes is one of their disadvantages.

Since in Franklin's motors and in other motors described in this chapter the rotor is charged by means of electric sparks, these motors may be classified as "spark motors". Toward the end of the 19th century a way was found to construct electrostatic motors based on the same mechanism of attraction and repulsion between electric charges as in Franklin's motors, but with rotors (or stators) charged more efficiently than it is possible to do by means of the sparks. These motors are described in the next chapter.

REFERENCES

1. J. Sparks, Ed., *The Works of Benjamin Franklin*, Whittemore, Niles, and Hall, Boston, 1856, Vol. 5, pp. 204-207.
2. O. Jefimenko and D. K. Walker, *The Physics Teacher*, Vol. 9, pp. 121-129 (1971).
3. O. Jefimenko, *American Journal of Physics*, Vol. 39, pp. 1139-1140 (1971).

4. M. Guyot, *Nouvelles Recreations Physiques et Mathematiques*, Gueffier, Paris, 1786, Vol. 1, pp. 272-274 and Plate 27.
5. Dayton C. Miller, *Sparks, Lightning, Cosmic Rays*, MacMillan, New York, 1939, pp. 63, 64.
6. A. D. Moore, *Electrostatics*, Doubleday and Company, Inc., New York, 1968, pp. 104-108.
7. J. W. Draper, *Text-Book on Chemistry*, Harper and Brothers, New York, 1847, p. 105.
8. M. J. Jamin, *Cours de Physique*, Gauthier-Villars, Paris, 1869, 2nd edn., Vol. 3, p. 39.
9. R. H. Dressel, *High School Science Bulletin of the New Mexico State University*, Vol. 6, pp. 3-4 (April 1, 1964).
10. H. Wommelsdorf, *Annalen der Physik*, Ser. 4, Vol. 70, pp. 135-138 (1923).

5

CORONA MOTORS AND MOTORS SIMILAR TO THEM

In 1867, W. Holtz, the inventor of the electrostatic "influence" machine (generator) carrying his name, discovered that this machine could function as a motor when powered by another similar machine.[1] In order to study this effect in some detail, the famous German physicist J. C. Poggendorff ordered the construction of a special device, analogous to Holtz's machine, but designed to operate only as a motor and provided with numerous adjustable components so as to operate under a wide range of conditions.[2]

The basic part of this motor was a vertical glass-disk rotor placed between two ebonite crosses carrying metallic sharp-needle "combs" (Fig. 28 and Plate 4). The sharp points of the combs almost touched the surface of the disk, so that the disk could be charged by means of a corona discharge from the combs when they were connected in alternation to an electrostatic generator. Each comb could be oriented so as to make an arbitrary angle with respect to a radial direction.

Just like the sparks in Franklin's motors, the corona discharge from a comb deposited on the rotor a charge of the same polarity as that of the comb. Each comb repelled therefore the segment of the rotor carrying charges which it sprayed onto the rotor and attracted the segment carrying charges sprayed onto the rotor by the preceding comb. Because of a continuous discharge between the combs and the rotor, almost the entire surface of the rotor was sprayed with electric charges. The repulsion and attraction between the combs and the rotor, were therefore not only much stronger than in Franklin's motors, (where only the electrodes were charged) but were also much steadier, since the distances between the combs and the charges on the rotor were always the same. As a result, the torque and power obtainable with this motor were far greater than those obtainable with Franklin's motors of similar dimensions.

Poggendorff made a thorough study of his motor, varying all its important parameters (his article comprises 32 pages). He used several rotors of different thickness and surface characteristics, whose weight ranged from 2.5 to 4.5 pounds (all rotors were 15 inches in diameter). He varied the number, the orientation, and the polarity of the combs.

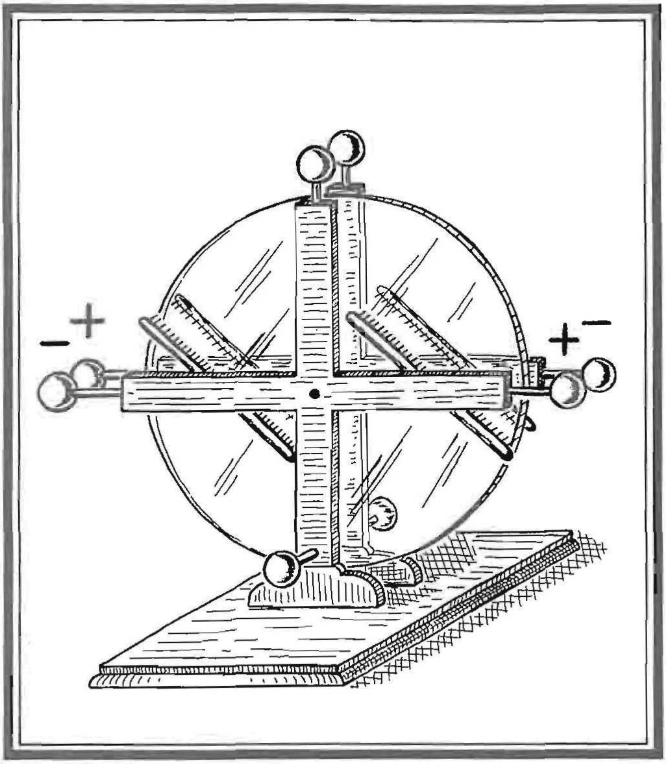

FIGURE 28

Poggendorff's corona motor had a glass-disk rotor placed between two ebonite crosses carrying metallic sharp-needle "combs".

He found that a motor with combs arranged along the radii of the rotor could rotate equally well in either direction but was not self-starting, since such combs sprayed charges onto the rotor in a perfectly symmetric fashion. However, if the combs were slanted relative to the radii, the motor became self-starting and unidirectional, because such combs sprayed charges predominantly in one direction. He also fould that the performance of the motor could be improved by placing pieces of glass, cardboard, or metal close to the surface of the rotor (this is the principle of the modern "backing plate" or "lining"; see Chapter 10). When properly adjusted, the motor could run at better than 300 rpm.

Poggendorff had fully appreciated the superiority of his motor in comparison fo Franklin's motors, and had stated so at the end of his article. In fact, he had indicated that his motor used from 1200 to 1800 times as much current as Franklin's motors, in which case, assuming that the operating voltage of his motor was the same as that of Franklin's motors (approximately 100 kilovolts), the power of his motor must have been close to 100 watts.

At the time when Poggendorff's article was written, however, it was considered unscientific to expect that electricity (*any* electricity) could ever become a significant source of motive power. This may be the reason why Poggendorff, speaking of the rotation phenomenon that he investigated, declared at the end of his article: "...it would be a sanguine hope if one wanted to believe that any useful mechanical effect could be achieved with it. That this is not possible, follows already from the consideration of how small is the quantity of electricity that is here put into play in comparison to that developed by a voltaic battery, with which one nevertheless, even with the help of the magnetismus produced by it, has so far achieved nothing substantial ("Erkleckliches")". The argument was obviously weak, but taking into account Poggendorff's reputation as an eminent scientist, it was probably sufficient to discourage further serious investigations of the corona-induced rotation for many years to come.

Since Poggendorff's motor depended on a corona discharge for its operation, the motor may be classified as a "corona motor".

Holtz's rotation phenomenon was also studied by the Danish physicist C. Christiansen (whose interest in the phenomenon had been aroused by a preliminary report of Poggendorff's investigations published in 1867).[3] Christiansen described two motors based on Holtz's discovery. His first motor consisted of a horizontal disk rotor and two stationary corona-producing combs. His second motor was similar to the first except that in it the glass disk was stationary and the combs constituted the rotor. The rotor combs in this second motor were located below the glass disk and were charged by means of two foil strips hanging from them and sliding along

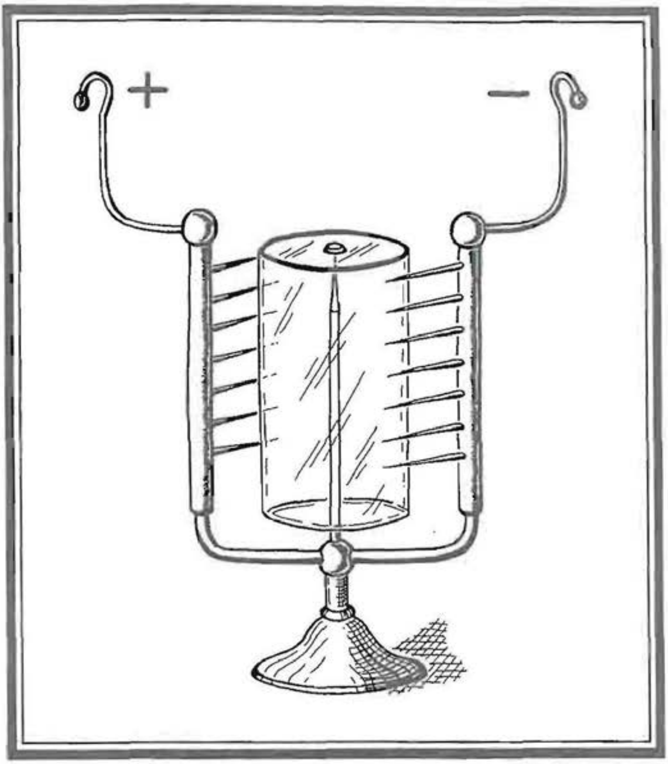

FIGURE 29

This self-starting corona motor with a cylindrical rotor was described in 1871 under the name of "electric tourbillion".

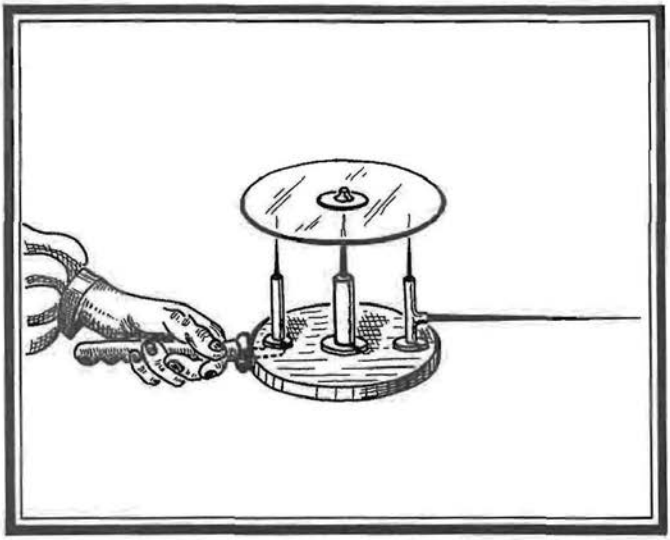

FIGURE 30

In this hand-held corona motor (attributed to Ruhmkorff) the horizontal wire, by a corona discharge, extracted charges from the power source, and the vertical needle connected to this wire, also by a corona discharge, sprayed these charges onto the rotor.

two horizontal concentric contact rings mounted on the base of the motor.

Cylindrical and spherical motors operating on the same principle as Poggendorff's motor were described in 1871 by W. Gruel (Fig. 29), who called them the "electric tourbillions".[4] Gruel pointed out that these motors could be made self-starting and unidirectional by inclining the combs in the same rotational direction with respect to the surface of the rotor. He also suggested that by providing the rotors with a series of holes, the motors could be used as sirens.

A disk motor not noticeably different from Poggendorff's motor was patented in the U.S.A. in 1891 by J. W. Davis and J. B. Farrington.[5]

An interesting hand-held corona motor (attributed to Ruhmkorff) was described in 1876 by M. E. Mascart.[6] The motor had a horizontal mica disk rotor (Fig. 30) supported on a vertical needle by a jewel bearing, such as are used for supporting compass needles. Below the disk there were two vertical corona-producing needles mounted on a hard rubber base. One of the latter needles was connected to a metal handle, the other was connected to a long, stiff, horizontal wire terminating in a sharp point. To set the motor in motion, the operator held the motor so that the sharp point of the horizontal wire was in the proximity of a charged conductor. By a corona discharge, the horizontal wire then extracted charges from the conductor, and the ver-

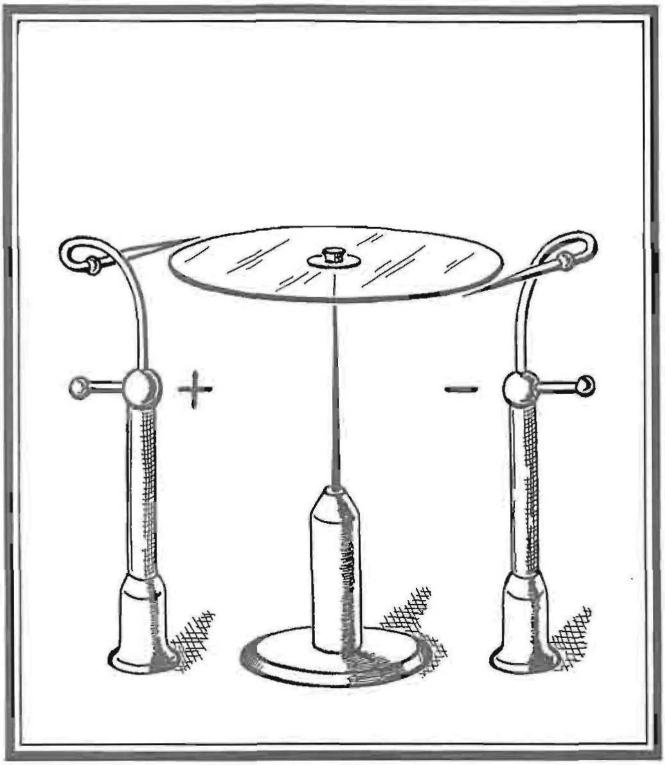

FIGURE 31

This motor had two insulated corona-producing points tangential to the rotor disk and oriented in the same rotational direction.

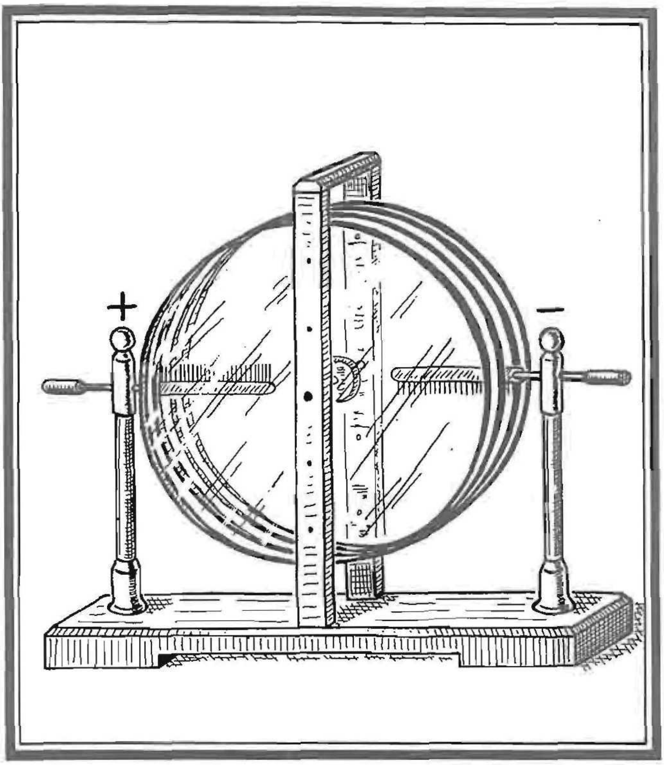

FIGURE 32

This corona motor with a compound glass rotor could run at about 2000 rpm and had a power of about 90 watts.

tical needle connected to this wire, also by a corona discharge, sprayed these charges onto the rotor. The second vertical needle discharged the rotor and conducted its charges to the ground through the hand and body of the operator.

A number of experiments with Poggendorff type corona motors were reported in 1921 by V. E. Johnson.[7] One of his motors was similar to Ruhmkorff's motor (Fig. 31). This was a horizontal mica disk, about 6 inches in diameter, supported on a vertical needle point. At the edge of the disk there were two insulated corona-producing points tangential to the disk and oriented in the same rotational direction. Another of Johnson's motors was a big machine consisting of two stationary glass disks and a three-piece glass rotor that could rotate on ball bearings between the stationary disks. The rotor was made of two large glass disks separated by a smaller disk so as to have a deep slot along its edge. Two flat sharp-point combs were inserted diametrically opposite to one another into the slot of the rotor (Fig. 32). Johnson estimated that his motor could run at about 2000 rpm and had a power of about 90 watts.

Just as Poggendorff's motor was derived from Holtz's electrostatic machine, a series of electrostatic motors were similarly derived from Wimshurst's electrostatic machine.

In 1891 five such motors were constructed by William McVay of New York City.[8] McVay's first motor (Fig. 33) consisted of two horizontal glass disks, about 12 inches in diameter, one stationary and the other rotating on the vertical axis just above the first. The lower disk had two quadrants of tinfoil, and the upper disk had 16 tinfoil sectors, as shown in the figure. The power (from a Wimshurst machine) was delivered to the motor by means of two insulated arms, each of which terminated in two brushes, one touching continually one of the lower quadrants, the other charging a sector on the upper disk just clear of the edge of the quadrant. Charges of the same polarity were thus deposited on the quadrant and on the sector, causing them to repel each other. An "equalizer" reduced the charge of the sectors before they passed over the further edge of an oppositely charged quadrant, thus reducing the back torque on the rotating disk. An important new feature of this motor was the simultaneous charging of the stationary quadrants and of the moving sectors, which assured a relatively strong starting torque and, together with the neutralizing system, assured a reliable unidirectional operation. McVay also constructed motors of cylindrical geometry, one of which is shown in Fig. 34. In this motor the quadrants were located on the inner cylinder, while the charging and neutralizing brushes were on the neck of the outside cylinder.

McVay's first motor was later modified for charging by combs, rather than by brushes (Fig. 35). The motor then operated essentially as

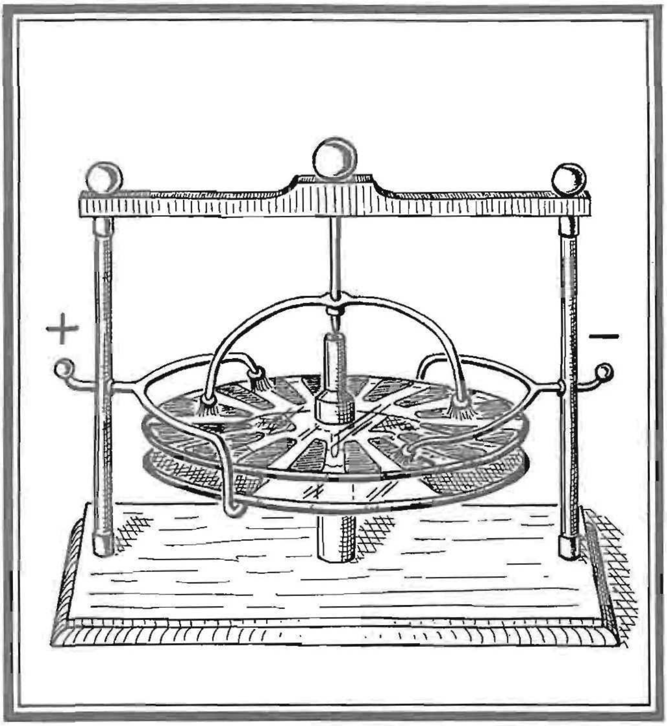

FIGURE 33

McVay's first motor consisted of two horizontal glass disks, about 12 inches in diameter, one stationary and the other rotating on the vertical axis just above the first.

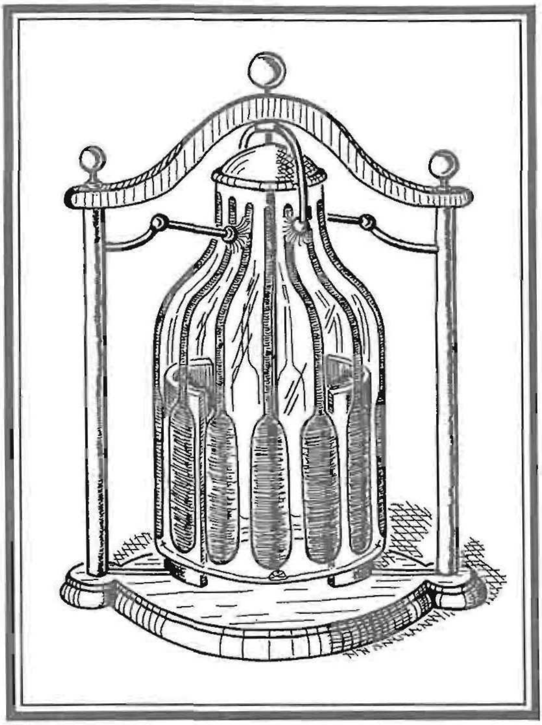

FIGURE 34

In this McVay's motor the quadrants were located on the inner cylinder (shown as two segments), while the charging and neutralizing brushes were on the neck of the outside cylinder.

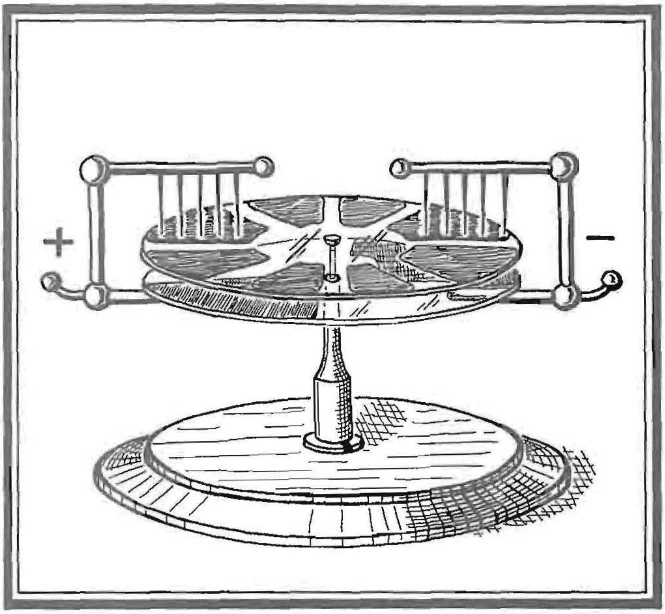

FIGURE 35

McVay's first motor was later modified for charging by combs, rather than by brushes. The motor then operated essentially as Poggendorff's corona motor, retaining, however, its self-starting and unidirectional qualities.

Poggendorff's corona motor, retaining, however, its self-starting and unidirectional qualities.

Instructions for building a Poggendorff motor may be found in the May 1971 issue of Popular Science.[9] Instructions for building a simple McVay motor may be found in the January 1914 issue of Electrical Experimenter[10] (the author is grateful to Mr. Thorn L. Mayes for this information).

Since a corona discharge in air at atmospheric pressure requires a minimum voltage of about 3000 volts, ordinary corona motors can operate only from sources capable of supplying a voltage of this magnitude. Such voltages are, of course, easily attainable at the present time, and since the corona motors are very simple and efficient devices, they have been further developed and studied in recent years. We shall return to them in the last chapter of this book.

REFERENCES

1. W. Holtz, *Annalen der Physik*, Ser. 2, Vol. 130, pp. 168-171 (1867).
2. J. C. Poggendorff, *Annalen der Physik*, Ser. 2, Vol. 139, pp. 513-546 (1870).
3. C. Christiansen, *Annalen der Physik*, Ser. 2 Vol. 137, p. 490 (1869).
4. "Elektrischer Tourbillion", *Annalen der Physik*, Ser. 2 Vol. 144, p. 644 (1871).
5. U. S. Patent No. 459678.

6. M. E. Mascart, *Traité d'Electricité Statique*, G. Masson, Paris, 1876, Vol. 1 pp. 179, 180.
7. V. E. Johnson, *Modern High Speed Influence Machines*, E. and F. N. Spon, London, 1921, pp. 175-205.
8. "A Group of Static Motors", *Electrical World*, Vol. 18, p. 418 (1891).
9. C. P. Gilmore and W. J. Hawkins, *Popular Science Monthly*, Vol. 198, pp. 95-97 and 114 (May 1971).
10. "A Static Electric Motor", *Electrical Experimenter*, January 1914, p. 137.

6

CAPACITOR MOTORS

In 1889, Karl Zipernowsky, a Hungarian engineer (co-inventor of practical electrical transformers), constructed a new type of electrostatic motor,[1] which was derived from Thomson's quadrant electrometer.[2] The rotor of this motor (Fig. 36) consisted of two pairs of aluminum sectors insulated from each other and from the rest of the apparatus. The stator consisted of four double (hollow) sectors of brass enclosing the rotor. The rotor was fitted with a commutator in four parts, by means of which the sectors of the rotor were charged oppositely to those sectors of the stator into which they were entering and identically to those sectors of the stator which they were leaving. An interesting property of this motor was that it could operate from high-voltage dc as well as from high-voltage ac sources.

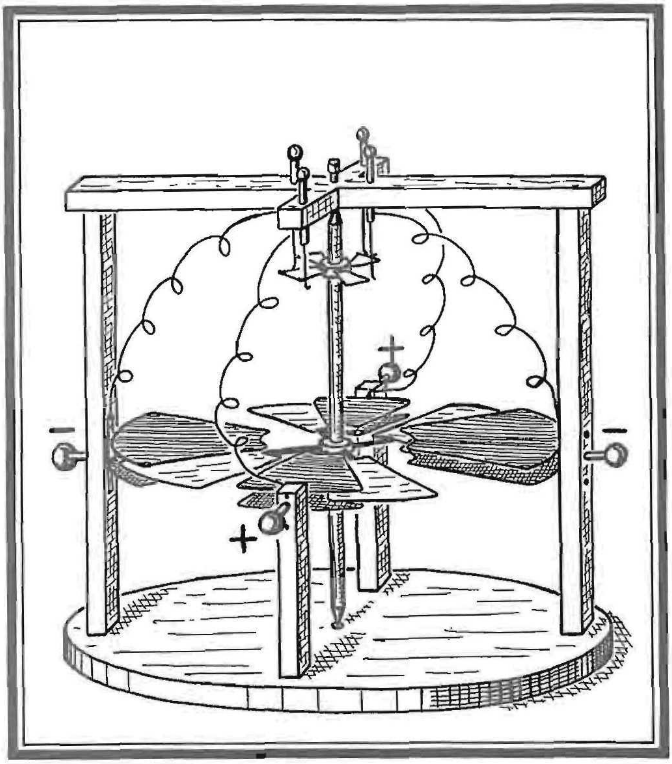

FIGURE 36

This motor was built in 1889 by Karl Zipernowsky, a Hungarian engineer. It operated from dc as well as from ac sources.

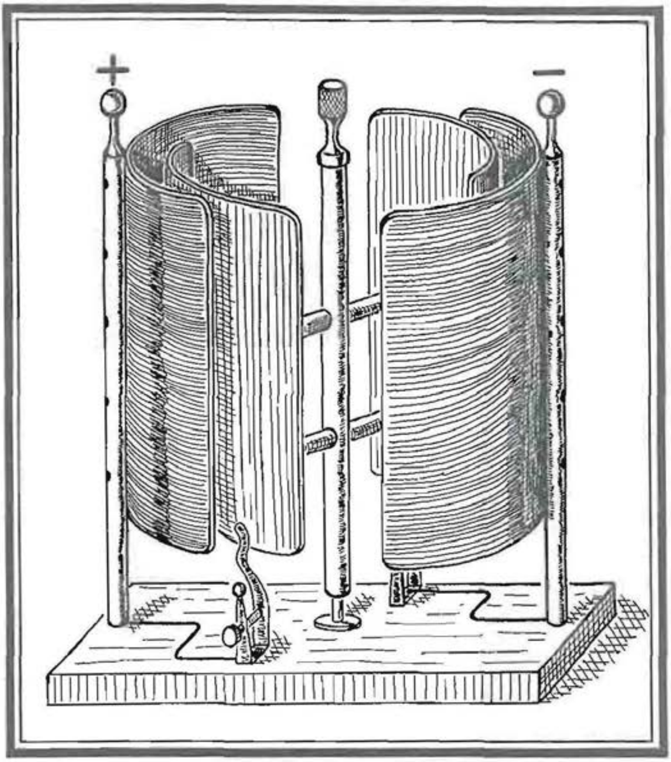

FIGURE 37

This capacitor motor was described in 1904 by van Huffel. Its rotor could be charged by contact or by sparks.

Inasmuch as Zipernowsky's motor operated as a result of the electric forces exerted by one charged conducting plate upon a second charged conducting plate (which are the same forces that act upon the two plates of a capacitor) it constituted what is now called an electrostatic "capacitor motor".

A simpler version of a capacitor motor was described in 1904 by van Huffel.[3] This motor was based on the so-called Thomson's replenisher.[4] The motor had an essentially cylindrical geometry (Fig. 37) and consisted of a rotor with two insulated bent brass plates mounted on a vertical axle and located between two similar, but slightly larger, plates of the stator. The rotor plates (which were not quite coaxial) were charged by means of two metallic tongues connected to the stator plates; the tongues touched the rotor plates just as the latter began to clear the stator plates. The charging of the rotor plates could occur by a direct contact as well as through a spark, depending on the adjustment of the tongues.

Since capacitor motors do not require sparks or a corona discharge for their operation, they can operate, at least in principle, from as low a voltage as one desires to use. This is one of their important advantages and is one of the reasons that such motors have been given considerable attention in recent years. Furthermore, as already indicated, capacitor motors can operate not only from dc sources, but also from ac sources. Finally, when powered by an ac source, they can operate both as synchronous and asynchronous motors (Zipernowsky's original motor operated from ac as an asynchronous motor).

A synchronous capacitor-type electrostatic motor is merely a multi-electrode capacitor motor without a commutator, the proper charging of the rotor being accomplished by continuously supplying an ac voltage of proper frequency between the stator and the rotor (Fig. 38). It is easy to see that if the rotor moves by one electrode in one period of the supply voltage, then the ac voltage accomplishes the same effect as that accomplished by a dc voltage with a commutator. The synchronous velocity is therefore $2\pi f/N$, where f is the frequency of the supply voltage and N is the number of the electrodes.

Several precision-made synchronous motors were described in 1969 by B. Bollée.[5] One of these motors ran on sapphire bearings and had a cylindrical aluminum rotor 1 centimeter long and 0.45 centimeters in diameter; the rotor and the stator had 60 electrodes each, with a 0.001 centimeter gap between stator and rotor electrodes. The maximum power of this motor was about 100 microwatts at 220 volts and 50 hertz. Another motor (for an electric clock) had a dielectric disk rotor, with four rows of 100 electrodes on each side, located between two bearing covers, each with five rows of 100 elec-

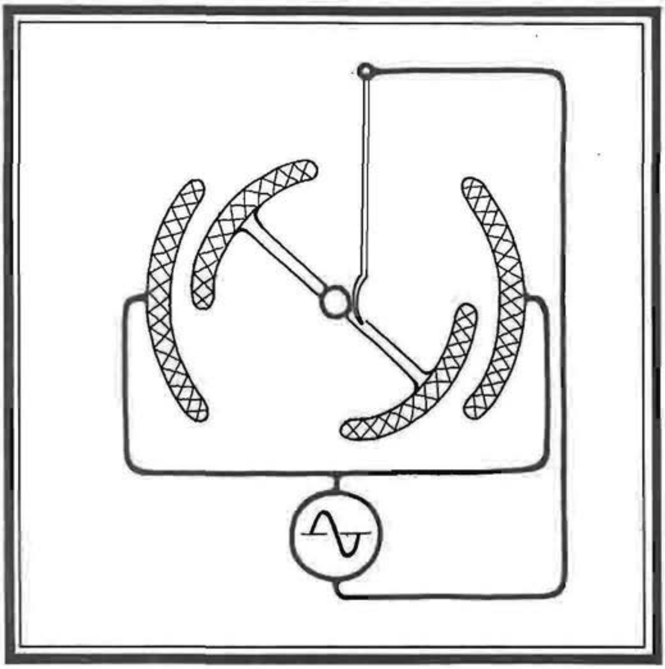

FIGURE 38

In a synchronous capacitor-type electrostatic motor the proper charging of the rotor is accomplished by continuously supplying an ac voltage of proper frequency between the stator and the rotor.

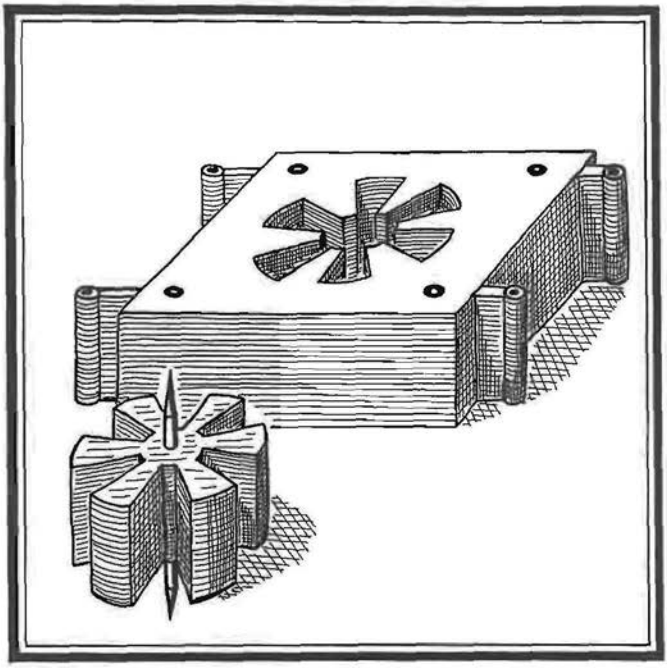

FIGURE 39

This synchronous electrostatic motor is essentially a variable capacitor with a rotatable shaft. The design of this motor is similar to the design of an electrostatic variable-capacitance generator.

trodes; the gap width in the axial direction was 0.01 centimeter. At 200 volts this motor had a maximum output of 600 microwatts. An interesting feature of this motor was that it had no slip ring on the rotor shaft, and that the rotor was therefore not connected to the power source. Instead, the stator was divided into two mutually insulated halves. This arrangement reduced the friction of the rotor and improved the performance of the motor. Still another motor was essentially a variable capacitor with a rotatable shaft. It had 15 plates in the stator and 14 in the rotor, each 0.03 centimeters thick; the gaps between stator and rotor plates were 0.015 centimeters (Fig. 39). The design of this motor is similar to the design of the electrostatic variable-capacitance generator described by N. J. Felici.[6]

According to Bollée, the maximum average torque on the rotor in a capacitor-type electrostatic motor is $T_{\max} = kNV^2$, where k is a geometrical constant, and V is the peak voltage applied. A motor with many electrodes is therefore slow but produces greater torque. The power of the motor is proportional to the square of the applied voltage and to the frequency but does not depend on the number of electrodes.

An important parameter in capacitor-type electrostatic motors is the capacitance difference $C_{\max} - C_{\min}$, where $C_{\max}$ is the greatest capacitance between the rotor and the stator and $C_{\min}$ is the smallest capacitance as they oc-

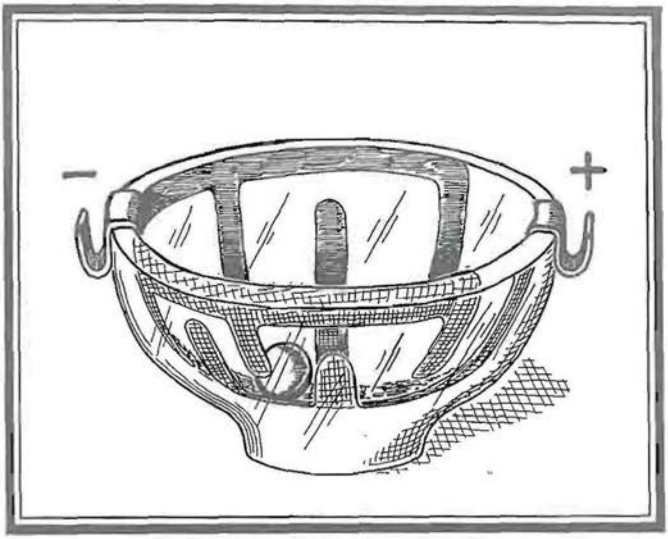

FIGURE 40

This interesting motor has been described by Moore, who calls it the "Interdigital Motor". The stator consists of a glass bowl with strips of aluminum foil glued to it. The rotor is a single conducting ball inside the bowl.

cur when the rotor turns. The larger this parameter is, the better are both the torque and the power. The main objective in the design of such motors is therefore to obtain the greatest possible variation of the capacitance.

An interesting motor which can be classified either as a modified Franklin's second motor or as a modified capacitor-type motor has been described by Moore,[7] who calls it the Interdigital Motor. The stator consists of a glass bowl with strips of aluminum foil glued to it (Fig. 40). The rotor is a single conducting ball inside the bowl; the ball runs along the sloping side when the strips are connected to an electrostatic generator (positive and negative strips are connected in alternation). Several balls can be placed into the bowl at the same time, and all of them will then run.

REFERENCES

1. "Zipernowsky Electrostatic Motor", *Electrical World*, Vol. 14, p. 260 (1889).
2. W. Thomson, *British Association Report*, 1855 (2), p. 22.
3. N. G. van Huffel, *Zeitschrift für den Physikalischen und Chemischen Unterricht*, Vol. 17, pp. 316-317 (1904).
4. W. Thomson, *British Association Report*, 1867, pp. 489-513.
5. B. Bollée, *Philips Technical Review*, Vol. 30, pp. 178-194 (1969).

6. N. J. Felici, *Electronics and Power*, Vol. 11, pp. 169-171 (1965).
7. A. D. Moore, *Electrostatics*, Doubleday and Company, Inc., New York, 1968, p. 109.

7

INDUCTION MOTORS

In 1892-1893, electrostatic motors of a fundametally new kind were described by Riccardo Arno[1,2] and by W. Weiler.[3] Their motors operated on the principle that the polarization of a dielectric in a variable electric field lags behind the field inducing the polarization. Therefore if a dielectric body is placed into a rotating electric field, the polarization charges* induced on the body will experience a force causing the body to follow the rotation of the field.

Arno's motor (Fig. 41) had a cylindrical stator formed by four insulated copper segments enclosing the space in which the rotating field was to be produced. The rotor was a hollow closed ebonite cylinder supported on two steel points turning in holes in glass; the cylinder weighed 40.33 grams, was 18 centimeters long and 8 centimeters in diameter. The rotating field was produced by means of a high-voltage transformer (3800 volts output), an RC-circuit, and a mercury commutator. The rotor attained a speed of about 250 rpm and developed a torque of 176 $\text{centimeter}^2 \times \text{gram} \times \text{second}^{-2}$.

---

*See, for example, O. Jefimenko "Electricity and Magnetism", Appleton-Century-Crofts, Inc. (1966), pp. 245-249.

Weiler used a hand-operated double commutator for his motors. The commutator delivered high voltage from an electrostatic generator in sequence to the four segments of a stator, thus producing the needed rotating electric field. Weiler described four different motors utilizing rotating fields. One of his motors was similar to Arno's motor. The operation of this motor (and that of Arno's motor) can be explained with the aid of Fig. 42, representing the top view of the motor. When segment A of the stator is positive, and segment B is negative, the electric field of the stator induces in the dielectric rotor (assumed to be at rest) polarization charges as shown in the figure. After the commutator has completed ¼ revolution, segment C is positive and segment D is negative. Thus the field has rotated by 90°, and now it polarizes the rotor in a new direction (at 90° with respect to the first). However, since it takes a certain time for the polarization charges to relax to zero, some of the initial polarization charges are still on the rotor, so that the distribution of the polarization charges on the rotor is not symmetrical relative

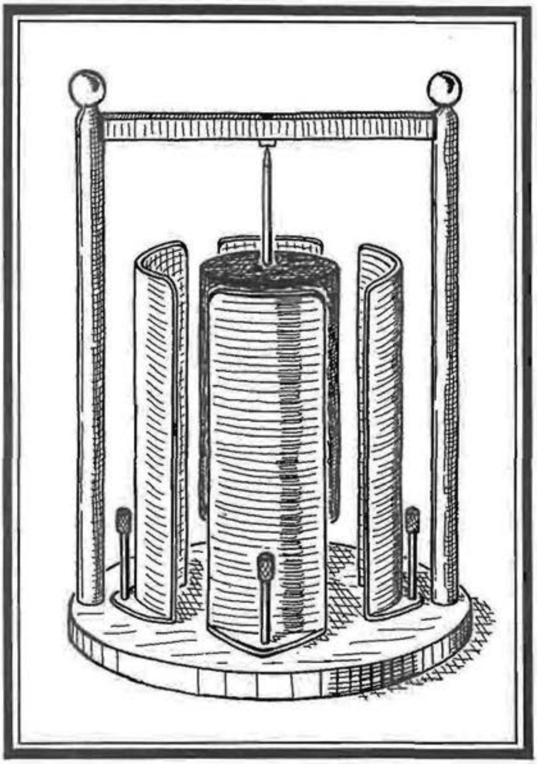

FIGURE 41

Arno's induction motor operated on the principle that the polarization of a dielectric in a variable electric field lags behind the field.

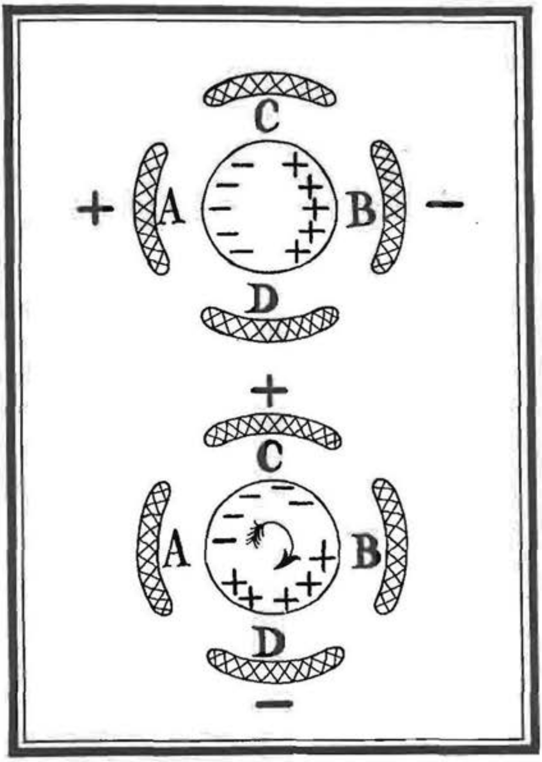

FIGURE 42

The operation of Arno's and Weiler's induction motors can be explained with the aid of this diagram.

to the charges on the stator. The rotor experiences therefore a torque due to attraction and repulsion between the rotor charges and stator charges, and the rotor turns.

Weiler also experimented with motors having noncylindrical rotors. One such motor is shown in Fig. 43. Rotating field electrostatic motors with disk-shaped rotors were built and studied between 1894 and 1901 in Japan by Hiderato Ho.[4]

Since Arno's and Weilers motors depend for their operation on induced polarization charges, they may be called "induction motors".

Induction motors, too, have a number of appealing features which attracted considerable interest to such motors in recent years: the motors can operate from both dc and ac; the motors can operate from low-voltage sources; since the motors require no brushes or slip rings, friction losses in these motors are very small.

Several carefully designed induction motors (both cylindrical and flat) were described in 1969 by Bollée.[5] One such motor of cylindrical geometry operated from a three-phase power supply at 220 volts and 50 hertz. The stator had 12 electrodes, which surrounded a rotor 15 centimeters in diameter. A disk-type induction motor operating from a three-phase power supply had a rotor consisting of 10 glass disks, each 6 centimeters in diameter, coated with a thin layer of slightly conductive material and mounted on a common axle. The rotor turned inside a ring

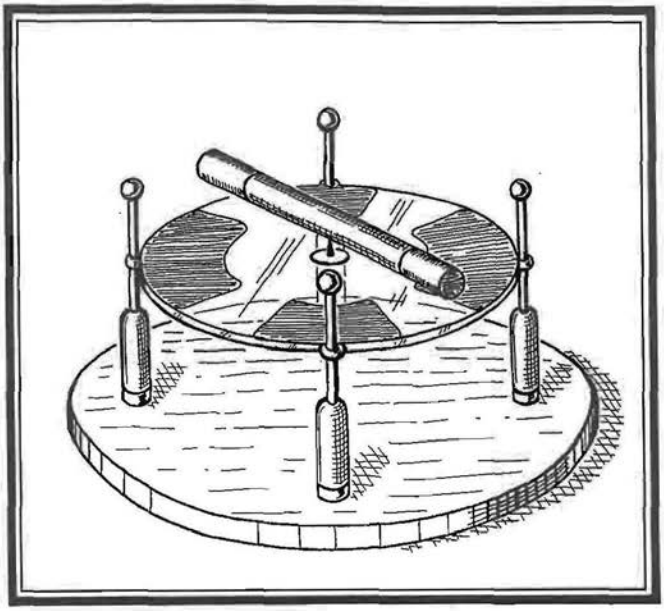

FIGURE 43

In 1892-1893, Weiler also experimented with induction motors having noncylindrical rotors, such as the motor shown in this drawing. Induction motors have a number of appealing features which attracted considerable interest to such motors in recent years.

formed by 60 comb-shaped stator electrodes. The maximum torque was 2000 centimeter^2 × gram × second $^{-2}$ at 220 volts and 50 hertz.

Another interesting motor described by Bollée was a linear induction motor, which "may be regarded as a segment of rotating-field motor bent straight".[5]

An extensive mathematical analysis of the induction motors was published in 1970 by Soon Dal Choi and D. A. Dunn.[6] They tested their theoretical results on a cylindrical motor with 72 electrodes. The motor was 14 inches wide, 14 inches high, and 19¼ inches long. It had a Plexiglas rotor that turned at a maximum rate of about 1000 rpm when powered by a 10-kilovolt source.

REFERENCES

1. R. Arno, *Rediconti della Atti Reale Accademia dei Lincei*, Vol. 1 (2), p. 284 (1892).
2. R. Arno, *The Electrician*, Vol. 29, pp. 516-518 (1892).
3. W. Weiler, *Zeitschrift für den Physikalischen und Chemischen Unterricht*, Vol. 7, pp. 1-4 (1893).
4. "A Rotary Field Electrostatic Induction Motor", *Electrical World*, Vol. 37, pp. 1012-1014 (1901).
5. B. Bollée, *Philips Technical Review*, Vol. 30, pp. 178-194 (1969).
6. S. D. Choi and D. A. Dunn, *Proceedings of the IDEE*, Vol. 59, pp. 737-738 (1971).

8

LIQUID-IMMERSED MOTORS

In 1893, W. Weiler discovered that a glass cylinder placed in a poorly conducting liquid between two spherical electrodes began to rotate when the electrodes were connected to an electrostatic generator.[1] He then constructed a small motor based on this discovery (Fig. 44). The principle of the operation of this motor was essentially the same as that of a cylindrical corona motor, except that the charges were now deposited on the rotor not by a corona discharge, but by the conduction current in the liquid.

In 1896, G. Quincke reported the same phenomenon and published a very comprehensive experimental study of it.[2] His device is shown in a simplified form in Fig. 45. It is interesting to note that Quincke, rather than Weiler, has been credited by subsequent investigators with the discovery of the rotation of dielectric bodies in

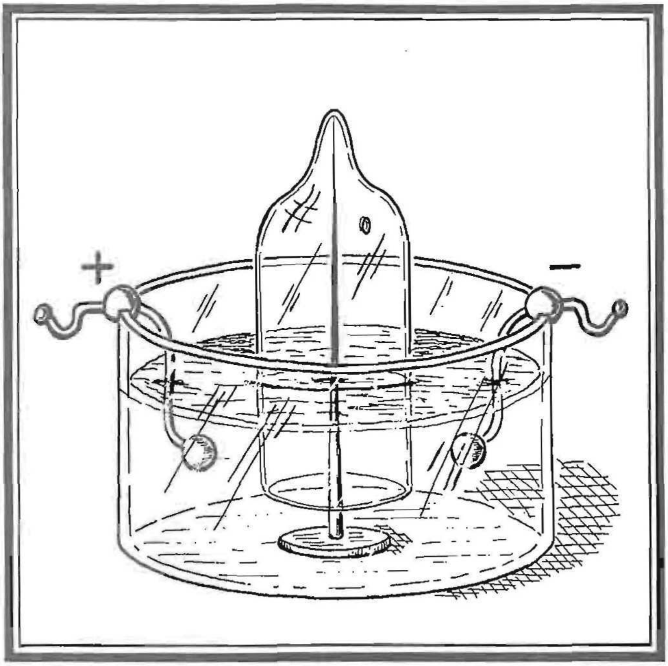

FIGURE 44

In 1893, W. Weiler discovered that a glass cylinder placed in poorly conducting liquid between two spherical electrodes began to rotate when the electrodes were connected to an electrostatic generator.

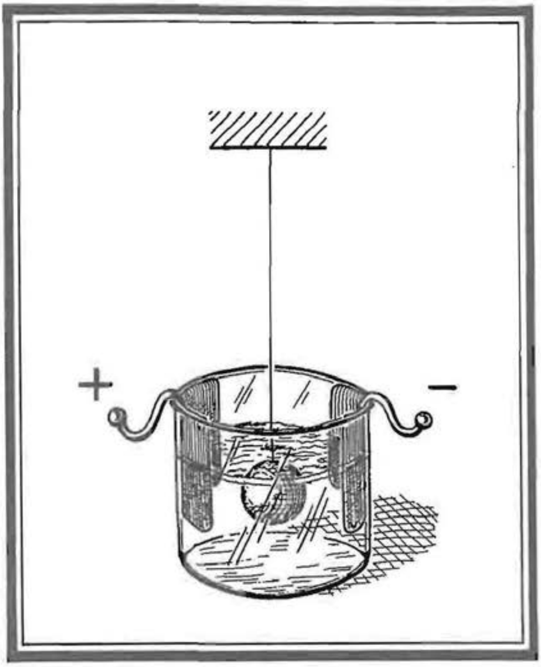

FIGURE 45

In 1896, G. Quincke reported the same phenomenon of rotation and published a very comprehensive experimental study of it. His liquid-immersed motor is shown in a simplified form in the above drawing.

poorly conducting liquids.

Weiler-type liquid-immersed dielectric motors were recently studied in considerable detail by P. E. Secker and his co-workers.[3,4] Their motors operated in various semi-insulating liquids such as hexane, hexane doped with amyl alcohol or ethyl alcohol, and isoamyl alcohol.

In one of their motors the electrodes were stainless steel squares 1.8 centimeters on a side. Their separation was varied from 1 to 2.5 centimeters. The rotor, of ebonite or Perspex, was 0.95 centimeters in diameter and 2.2 centimeters long. The operating voltage was from 3 to 30 kilovolts. The speed of the motor was found to be a linear function of the applied voltage, with the maximum recorded speed of about 2500 rpm.

In another motor, a stator with six electrodes was used (Fig. 46). The rotor was made of Perspex and was 7/8 inches in diameter and 2 inches long; it had a layer of high-permittivity material on its surface (titanium ceramic in polystyrene). Powered by a 30-kilovolt source, the motor turned at 1700 rpm, and the total power input amounted to 5.4 watts, of which 2.7 watts appeared at the rotor shaft.

A discussion of electrohydrodynamic effects that may take place during an operation of liquid-immersed motors has been presented by J. R. Melcher and G. I. Taylor.[5]

Because of considerable hydrodynamical losses, the motors of this type will apparently always have a relatively low output torque.

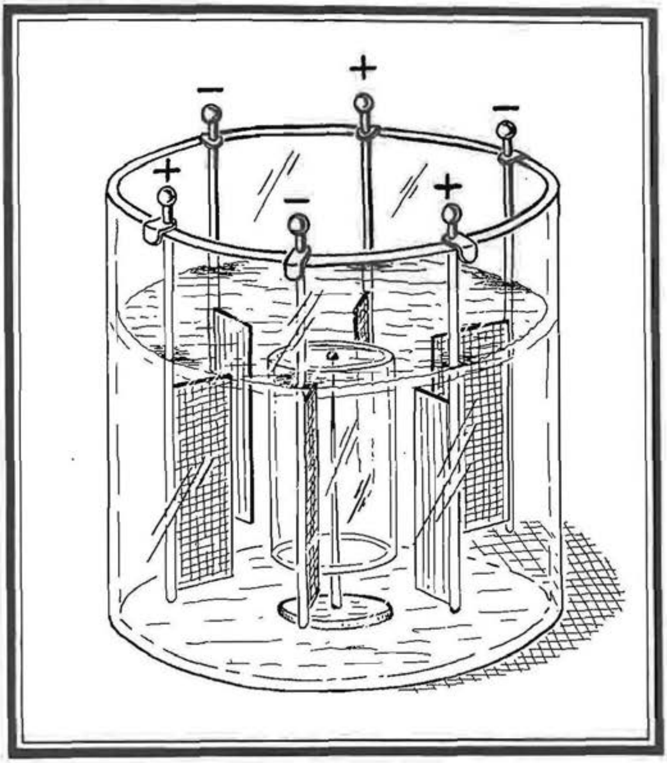

FIGURE 46

This liquid-immersed dielectric motor was recently studied in considerable detail by P. E. Secker and his co-workers.

REFERENCES

1. W. Weiler, *Zeitschrift für den Physikalischen und Chemischen Unterricht*, Vol. 6, pp. 194-195 (1893).
2. G. Quincke, *Annalen der Physik*, Ser. 3, Vol. 59, pp. 417-486 (1896).
3. P. E. Secker and I. N. Scialom, *Journal of Applied Physics*, Vol. 39, pp. 2957-2961 (1968).
4. P. E. Secker and M. R. Belmont, *Journal of Physics D: Applied Physics*, Vol. 3, pp. 216-220 (1970).
5. J. R. Melcher and G. I. Taylor, *Annual Review of Fluid Mechanics*, Vol. 1, pp. 111-146 (1969).

9

ELECTRET MOTORS

In general, an electret is a permanently charged dielectric body. In a more restricted sense of the word, an electret is a permanently polarized dielectric body and may be considered to constitute the electrical counterpart of a permanent magnet. It is this latter type of electret that has been used for constructing electrostatic motors.

In 1961, the Russian physicist A. N. Gubkin described an electret motor[1] shown schematically in Fig. 47. The motor consisted of a stator formed by two horizontal, axially symmetric, parallel-plate capacitors and a rotor formed by two flat, axially symmetric, oppositely polarized electrets mounted on a vertical axle and capable of passing between the plates of the capacitors. The motor operated as follows. When a voltage was applied to the two capacitors, so that

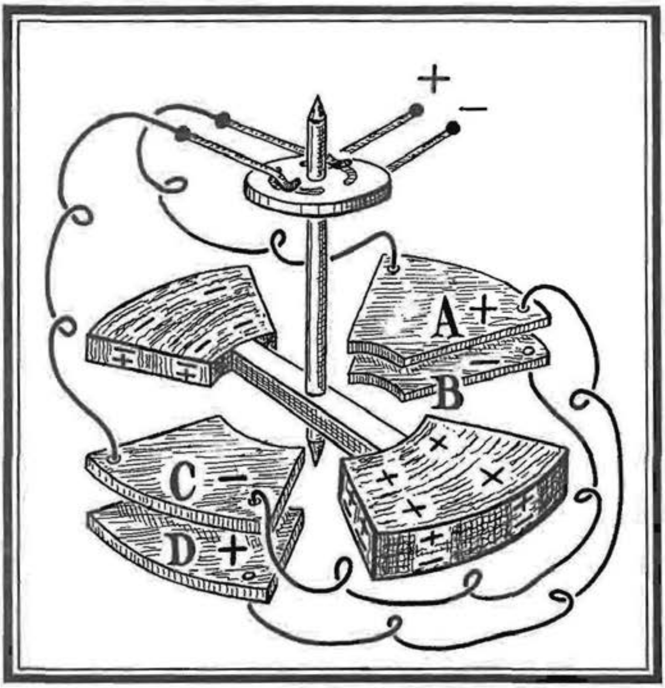

FIGURE 47

In 1961, the Russian physicist A. N. Gubkin described this motor with two electrets in the rotor and two parallel-plate capacitors in the stator. A commutator was used to change the polarity of the capacitors.

plates A and D were positive and plates B and C were negative, the capacitors attracted the electrets as soon as the latter came close to the plates, and the electrets were pulled into the capacitors. A commutator changed the polarity of the plates just as the electrets were coming out from the capacitors, and the electrets were then repelled from them, and so on. Thus the rotor was set in a continuous rotation.

A well-known property of electrets is that they lose their polarization in the absence of adequate shielding. Electret motors with almost perfect shielding were described by this author and by D. K. Walker in 1970.[2] These motors were based on the so-called "electret slot effect".[3] The slot effect works as follows (Fig. 48). An electret is placed between two pairs of electrodes, each pair forming a slot. If a voltage is applied to one or both electrode pairs, the electret (or the electrodes) experiences a force in a direction perpendicular to the slot and parallel to the plates. This arrangement insures both a nearly perfect shielding and a relatively large force.

A simple electret motor utilizing the slot effect is shown in Fig. 49 and Plate 5.* It uses a disk-shaped carnauba wax electret rotor consisting of two oppositely polarized half-disks. The stator has two pairs of electrodes connected to a

---

*The motor was designed by the author and built by one of his students, Charles Lynn Walls, in 1966.

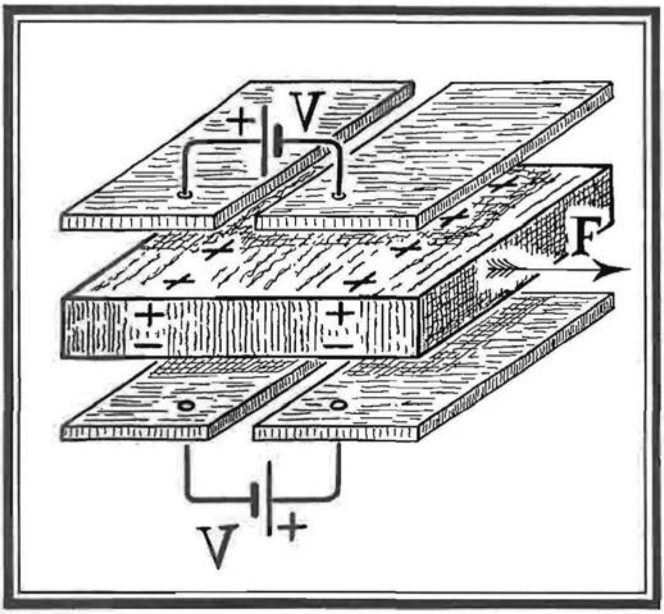

FIGURE 48

The slot effect works as follows. An electret is placed between two pairs of electrodes, each pair forming a slot. If a voltage is applied to one or both electrode pairs, the electret experiences a force in a direction perpendicular to the slot and parallel to the plates.

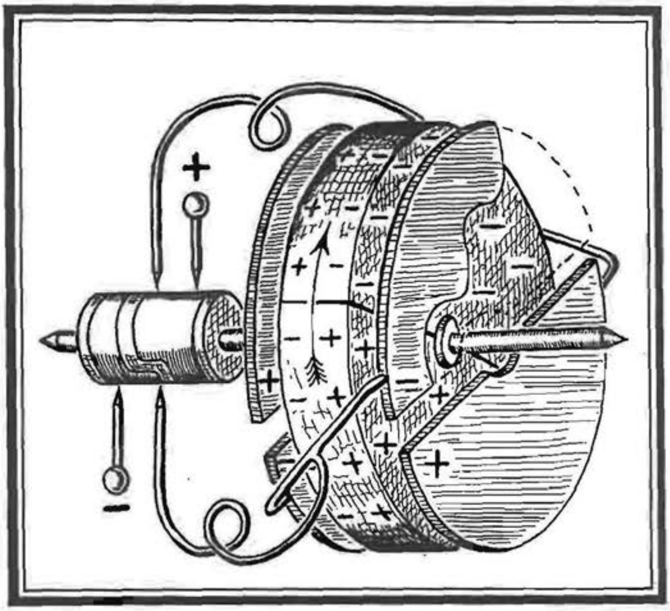

FIGURE 49

This simple electret motor utilizes the slot effect. The rotor is made of a disk-shaped electret consisting of two oppositely polarized half-disks. The stator has two pairs of slotted electrodes connected to a cylindrical commutator, which is charged by contact or by sparks.

cylindrical commutator. The thickness of the rotor is $1/2$ inch, the diameter is 5 inches. The motor operates from an 8-kilovolt power supply and rotates at 1500 rpm.

Similar motors with stationary electrets and rotating electrodes are shown in Plates 6 and 7.

A more sophisticated slot-effect motor[2] uses a cylindrical electret (Fig. 50 and Plate 8). The electret is stationary; it is shaped as a hollow cylinder and has four sections of opposite radial polarization. The rotor is made of four internal and four external aluminum electrodes forming two cylinders with four slots in each. The inner electrodes are cross-connected with the outer ones, and all electrodes are supported by a Plexiglas disk mounted on a steel axle. This motor uses no commutator. The power is delivered to two adjacent external electrodes by means of two sharp points which charge the rotor through a corona discharge. The overall diameter of this motor is 3 inches, the operating voltage is 6 kilovolts, the speed is up to 5000 rpm, the power is about 20 milliwatts.

Synchronous electret motors for electric clocks have been recently announced by the General Time Corporation. The motors are about $1 1/4$ inches in diameter and $1/4$ inch thick. In these motors (Fig. 51) a thin plastic electret disk with 15 active sectors and 15 equally large cut-outs is the rotor. The rotor is placed between two stator plates fabricated as printed circuit boards. Each stator plate has 30 elec-

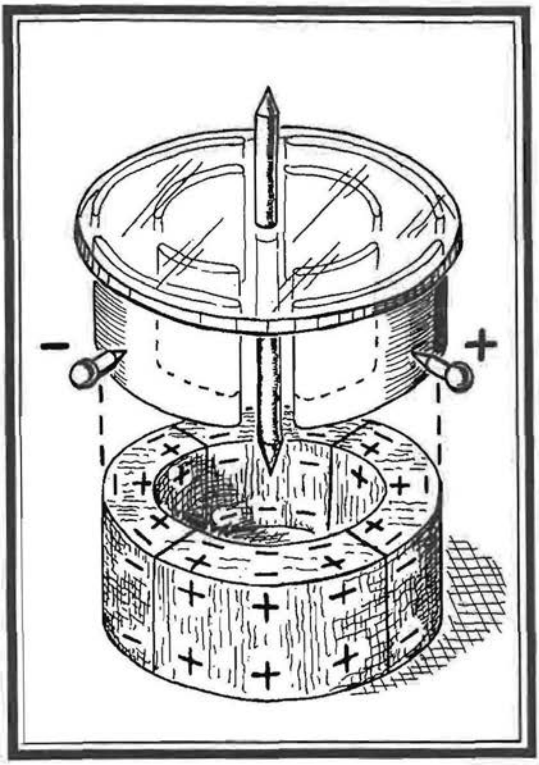

FIGURE 50

This more advanced slot-effect motor uses a stationary cylindrical electret and works without a commutator. The rotor has 8 slots.

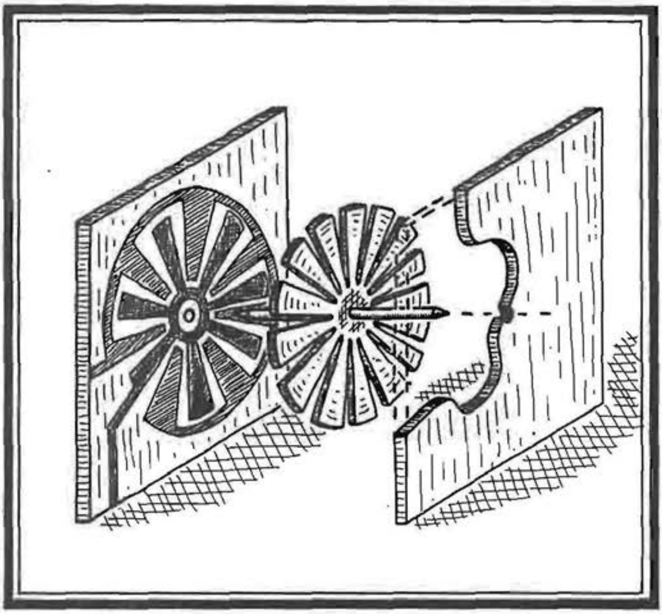

FIGURE 51

In this simplified version of a synchronous electret motor for electric clocks a thin plastic electret disk with several active sectors and several cut-outs is used as the rotor.

trodes connected in alternation to the two input terminals. The operation of this motor is similar to that of the first slot-effect motor described above (Fig. 49) except that the reversal of polarity of the electrodes is accomplished directly by the applied ac voltage rather than by a commutator.

Of all presently known types of electrostatic motors, electret motors are the newest and (together with corona motors) the most highly promising ones. We shall return to them once again in the next chapter.

REFERENCES

1. A. N. Gubkin, *Electrets*, Academy of Sciences, Moscow, 1961, pp. 130-133 (in Russian).
2. O. Jefimenko and D. K. Walker, *Conference on Dielectric Materials, Measurements and Applications*, The Institution of Electrical Engineers, London, 1970, pp. 146-149.
3. O. Jefimenko, *Proceedings of the West Virginia Academy of Sciences*, Vol. 40, pp. 345-348 (1968).

10

WHAT TO EXPECT FROM ELECTROSTATIC MOTORS

Having described the various types of electrostatic motors we shall now present a brief discussion of the current aims in the electrostatic motor research and a brief discussion of the most probable future uses and applications of these motors.

It has been pointed out by Bollée[1] that electromagnetic motors rapidly lose their efficiency in scaled-down versions (which is due to the relative increase in energy dissipation in the magnet coils) and that very small capacitor-type and induction-type electrostatic motors may be a better choice for miniaturized systems. Therefore one may expect that miniature electrostatic motors of these types will find applications in various sensor and control devices where only very small torques and power are needed.

Experiments with electret motors conducted in the author's laboratory[2] indicate that these motors may be very useful in systems where powers of up to 1 watt are needed.

The most promising electrostatic motors appear to be, however, the corona motors. These motors possess a number of highly desirable features. Here are some of them: the motors are extremely simple in design; they require no expensive materials; their maintenance is very simple; having only few metal parts they possess a very good power-to-weight ratio; they are fully capable of developing appreciable amounts of power; and they can attain very high speeds.

Our present-day awareness of the many attractive properties of corona motors is to a great extent due to the work of the Russian engineers Yu. Karpov, V. Krasnoperov, and Yu. Okunev published in 1958 and 1960.[3,4] They described a 6-watt cylindrical corona motor of improved design (Fig. 52) operating from a 7-kilovolt power supply and turning at a rate of 6000 rpm. This motor had a hollow Plexiglas rotor 10.5 centimeters in diameter and 17 centimeters long with a conducting lining on its inner surface. The stator supported 16 knife-like electrodes inclined relative to the surface of the rotor in the direction of the desired rotation. The lining of the rotor served to increase the electric field in the gap between the electrodes and the surface of the rotor and thus to enhance the corona discharge from the electrodes.

A high-speed corona motor with a disk rotor

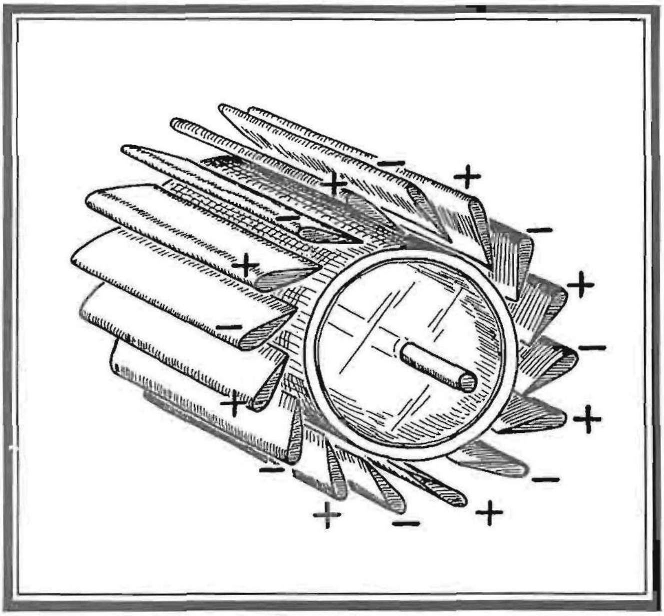

FIGURE 52

In 1958, the Russian engineers Yu. Karpov, V. Krasnoperov, and Yu. Okunev described a 6-watt cylindrical corona motor of an improved design operating from a 7-kilovolt power supply and turning at a rate of 6000 rpm. This motor had a Plexiglas rotor.

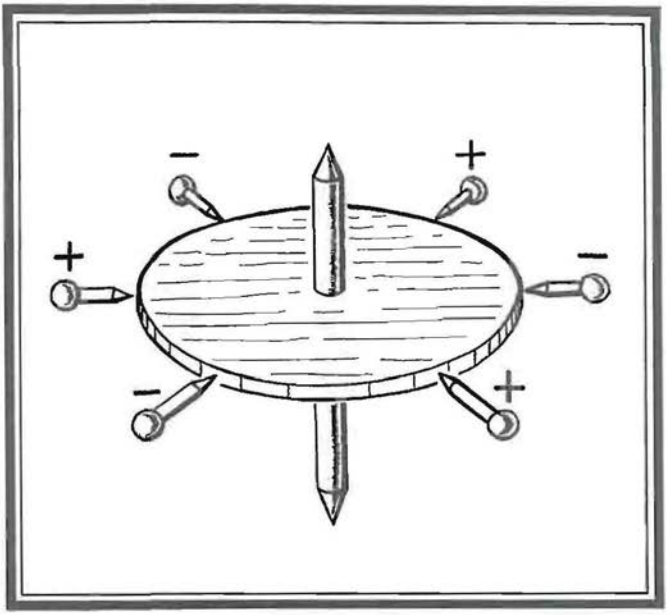

FIGURE 53

This corona motor with a disk rotor and circumferential electrodes was described by J. D. N. Van Wyck and G. J. Kühn of South Africa in 1961. It had a rotor 1.5 inches in diameter turning in jewel bearings. Operating from 13 kilovolts, the motor ran at 12000 rpm.

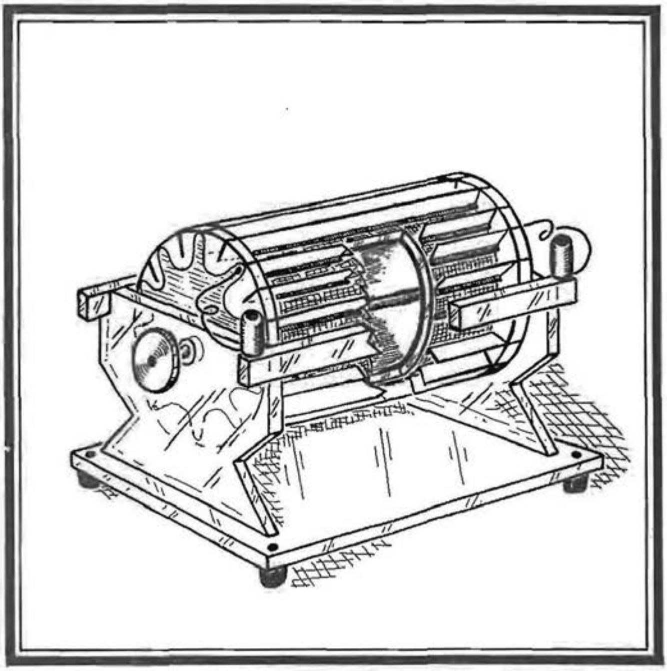

FIGURE 54

The performance of a corona motor depends on the shape and arrangement of electrodes and the structure of the rotor. The curved electrodes of this motor produce an especially large torque. The lining of the rotor increases the power.

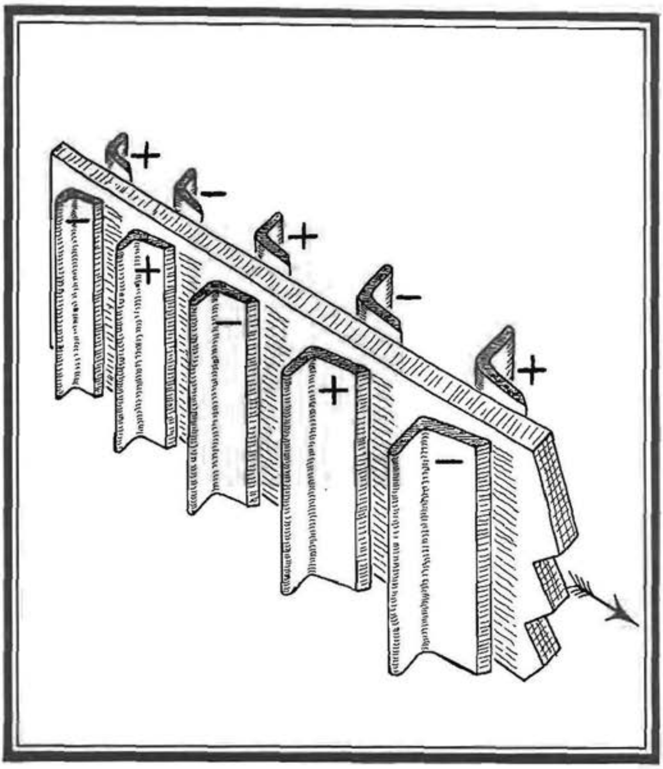

FIGURE 55

The mechanism of electric interaction producing rotational motion in corona motors can be used to achieve a translational motion as well. This drawing shows the principle of a linear corona motor with multiple electrodes.

and circumferential electrodes was described by J. D. N. Van Wyck and G. J. Kühn of South Africa in 1961.[5] The motor (Fig. 53) had a rotor 1.5 inches in diameter turning in jewel bearings. The stator supported 6 sharp-point electrodes. Operating from 8-13 kilovolts, the motor developed speeds of up to 12000 rpm.

A similar motor was studied in Poland by B. Sujak and W. Heffner in 1963.[6]

A number of advanced corona motors were studied in the author's laboratory;[7] an example of these motors is given in Fig. 54 and Plate 9. A diagram of a linear corona motor designed by the author is shown in Fig. 55. Instructions for building the corona motors of the author's design may be found in the May 1971 issue of Popular Science.[8]

It appears that corona motors with input power of 100 to 1000 watts and efficiency of at least 60% can be constructed without any difficulty. There seems to be no reason why even more powerful corona motors could not be built. It is likely that the motors can be further improved by using rotors immersed in a gas other than the air at atmospheric pressure.

Performance characteristics of corona motors strongly depend on the design and geometrical configuration of the electrodes spraying charges onto the rotor. Several typical electrode types and electrode configurations for corona motors are shown in Fig. 56. Especially interesting is the electrode arrangement shown in Fig. 56c.

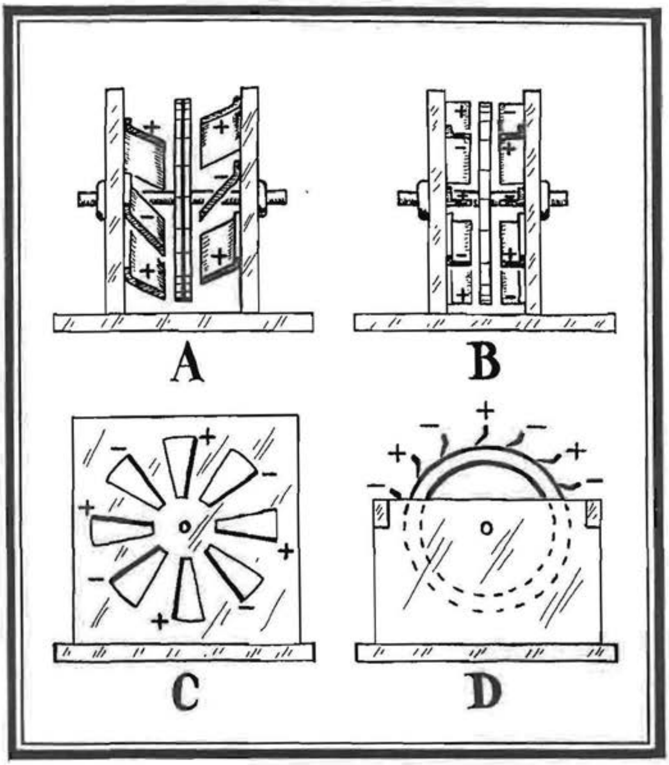

FIGURE 56

(a) Slanted electrodes and disk rotor with a conducting lining in the middle; (b) Straight electrodes; (c) Foil electrodes shielded on one side; (d) Curved electrodes and cylindrical rotor with a conducting lining.

This arrangement is used in conjunction with thin disk-shaped rotors "sandwiched" between two insulating stator plates. An even number of windows is cut in each stator plate, and metal foil electrodes are glued to the walls of the holes, as shown by heavy lines in Fig. 56c. In this arrangement there is a dielectric medium on one side of each electrode, and there is air on the other side. Therefore the electrodes spray charges predominantly in the direction of the side facing the air. Thus a motor with electrodes of this type is unidirectional and self-starting. Furthermore, the entire "sandwich" comprising the rotor and the two stator plates may be very thin. Therefore many such "sandwiches" can be placed on a single axle thus forming a compound motor in which good use is made of the entire volume occupied by the motor. There are indications that compound motors of this type can develop up to 1000 horsepower for each cubic meter of their volume.

An important property of electrostatic motors is that they can operate from a much greater variety of sources than the electromagnetic motors. It is, of course, obvious that any sources used for operating conventional electromagnetic motors can be used to operate the electrostatic motors. However, the electrostatic motors can operate also from sources from which no other motors can operate. The reason for this is that the electrostatic motors are extra-high impedence devices and thus require extremely small currents for their operation.

One very interesting source of electricity for electrostatic motors is the ordinary capacitor. It is true that capacitors, as we know them now, do not store appreciable amounts of electric energy. However, if high-permittivity low-loss dielectric materials were developed, then the capacitor would become a very useful device for storage of electric energy, and possibly could be used in place of the chemical batteries for operating electrostatic motors.

Another possible source of power for electrostatic motors are high-impedance high-voltage batteries of the type of Zamboni pile.

A potentially very important source of electricity for powering electrostatic motors are electrostatic generators. Considerable advances in the development of such generators have been made in recent years[9,10] and it is conceivable that electrostatic motor-generators will be used to convert the high-voltage dc produced by such generators into the conventional low-voltage dc or ac.

Finally, a very interesting source for powering electrostatic motors is the atmospheric electric field. In fact, it appears possible to extract the energy contained in this field by means of electrostatic motors. Experiments on such an energy extraction have been conducted by the author.[11,12] In these experiments an electret motor and a corona motor were powered by simple earth-field antennas (Fig. 57; Plates 10 to

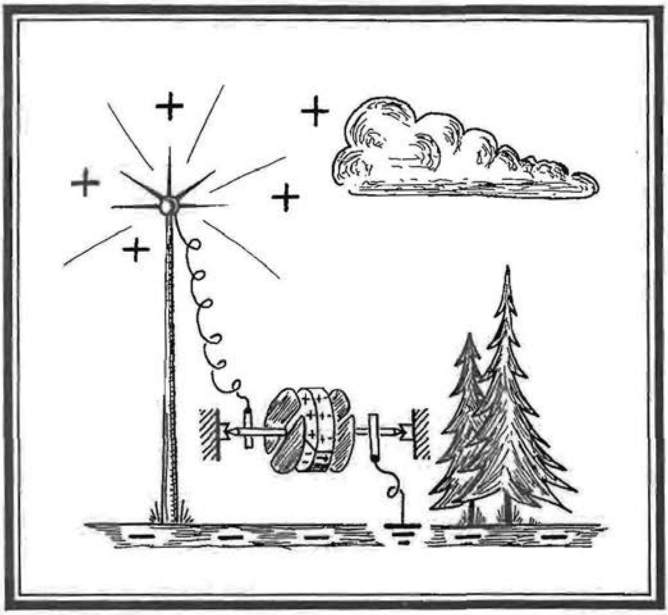

FIGURE 57

Small electrostatic motors can be powered from atmospheric electricity by means of simple earth-field antennas. Whether or not it will be possible to operate large motors in this manner will depend on how successful we are in designing antennas capable of extracting appreciable power from the earth electric field.

12). The corona motor was the one shown in Plate 9. The electret motor was similar to that shown in Fig. 49, but with the electret stationary and the electrodes rotating (Plate 6). These experiments indicate that it is entirely possible to operate small electrostatic motors from atmospheric electricity. Whether or not it will be possible to operate large motors in this manner will depend on how successful we will be in designing and building earth-field antennas capable of extracting appreciable power from the earth electric field.

In conclusion it may be useful to mention that a quantitative electrostatic motor research is now only in its very rudimentary stage. Almost all papers on electrostatic motors published thus far deal with the qualitative aspects of the performance of the various types of motors. Only very few papers deal with the optimization of design and with the quantitative aspects of the operation of the motors, likewise only very few papers present a theoretical analysis of the electrostatic motors. It is clear therefore that the electrostatic motors still constitute an essentially unexplored area of physics and engineering, and that the electrostatic motor research is presently one of the potentially most rewarding research fields in modern electrostatics.

REFERENCES

1. B. Bollée, *Philips Technical Review*, Vol. 30, pp. 178-194 (1969).

2. O. Jefimenko and D. K. Walker, *Conference on Dielectric Materials, Measurements and Applications*, the Institution of Electrical Engineers, London, 1970, pp. 146-149.
3. Yu. Karpov, V. Krasnoperov, and Yu. Okunev, *Tekhnika Molodezhi*, Vol. 26, pp. 36-37 (Sept. 1958) (in Russian).
4. Yu. Karpov, V. Krasnoperov, Yu. T. Okunev, and V. V. Pasynkov, "On the Motion of Dielectrics in an Electric Field", in *Dielectric Physics*, Academy of Sciences, Moscow, 1960, pp. 124-131 (in Russian).
5. J. D. N. Van Wyck and G. J. Kühn, *Nature*, Vol. 192, pp. 640-650 (1961).
6. B. Sujak and W. Heffner, *Acta Physica Polonica*, Vol. 23, pp. 715-726 (1963).
7. O. Jefimenko and H. Fischbach-Nazario, *Proceedings of the West Virginia Academy of Sciences*, Vol. 42, pp. 216-221 (1971).
8. C. P. Gilmore and W. J. Hawkins, *Popular Science Monthly*, Vol. 198, pp. 95-97, 114 (May 1971).
9. N. J. Felici, *Elektrostatische Hochspannungs Generatoren*, G. Braun, Karlsruhe, 1957.
10. J. Hughes and P. Secker, *New Scientist and Science Journal*, Vol. 49, pp. 468-470 (1971).
11. O. Jefimenko, *American Journal of Physics*, Vol. 39, pp. 776-778 (1971).
12. W. Aston, *West Virginia University Magazine*, Vol. 3, Number 4, Spring 1971, pp. 6-11.

INDEX

INDEX

Adams, G., 37
Aimé, R., 29,38
Arno, Q., 84,90
Arno Motor, 84,85,87
Aston, W., 118
Asynchronous motors, 4, 77
Atmospheric electricity, 3, 115-117
Avery, F. M., 21

Backing plate, 59, 107
Behrens, G., 14
Bell jar motor, 48, 49, 52
Belmont, M. R., 96
Benjamin, P., 21
Bichat, E., 29, 32, 38
Bollée, B., 77, 80, 82, 90, 106, 117
Bollée capacitor motors, 77, 79, 80
Bollée induction motors, 88, 90

Capacitor motors, 4, 73-77, 80, 82, 106
Chew, V. K., 37
Choi, S. D., 90
Christenson, A., 38
Christiansen, C., 60, 71
Christiansen motor, 60
Classification of electrostatic motors, 4
Collinson, P., 39
Contact motors, 4, 6, 20
Corona discharge, 22, 56, 71
Corona-kraft levitation, 33
Corona motors, 4, 56, 58, 60-66, 68-71, 107-112, 115, 116

Dailey, H. B. and E. M., 15, 21
Dailey motor, 15-18
Davis, J. W., 63
Draper, J. W., 55
Dressel, R. H., 51, 55
Dunn, D. A., 90

Earth-field antennas, 115, 116, 117
Electret, 97, 99
Electret motors, 4, 97-104, 106, 114, 115
Electret slot effect, 99, 100
Electric bells, 6, 8-11, 23, 25
Electric chimes, 11, 14, 23, 25
Electric fly, 22-26, 29-31
Electric horse-race, 26, 28, 29

Electric motor, 1, 6, 39
Electric orrery, 26, 27
Electric pendulum, 6
Electric pendulum motors, 6, 15, 16
Electric perpetuum mobile, 3, 15, 16, 48, 50
Electric pinwheel, 22
Electric planetarium, 18, 19
Electric rabbit hunt, 33, 36
Electric racing ball, 18, 19
Electric reaction wheel, 22
Electric seesaw, 13, 14
Electric swing, 12, 14
Electric tellurium, 26, 27
Electric tourbillion, 61, 63, 71
Electric tourniquet, 29, 32
Electric toys, 12, 13, 14, 28, 29, 33, 36
Electric wheels, 37, 39
Electric whirl, 22
Electric wind, 23
Electric wind motors, 22 ff, 33-35
Electric wind turbines, 33, 35
Electric wind wheels, 33, 34
Electromagnetic motors, 1-3, 106, 114
Electrophorus, 52
Electrostatic forces, 1
Electrostatic generators, 80, 115
Electrostatic motors, 1-4, 106, 107, 114-117

Farrington, J. B., 63
Felici, N. J., 80, 83, 118
Franklin, Benjamin, 7, 33, 37, 39, 42, 43
Franklin bells, 7, 14
Franklin first motor, 39-45, 48, 52
Franklin motors, 39-45, 54, 57, 59
Franklin second motor, 41, 43-45, 82

Gale, L. D., 33, 38
Gilmore, C. P., 72, 118
Gordon, A., 6, 22
Gordon bells, 6, 8, 15, 18, 20
Gruel, W., 63
Gruel motor, 61, 63
Gubkin, A. N., 97, 105
Gubkin motor, 97-99
Guyot, M., 20, 21, 29, 33, 38, 55

Hamilton, 23
Hamilton electrometer, 23
Hawkins, W. J., 72, 118
Heffner, W., 112, 118
Ho, H., 88
Holtz, W., 56, 71
Holtz rotation phenomenon, 56, 59, 60
Van Huffel, N. G., 76, 82
Van Huffel motor, 75, 76
Hughes, J., 118

Induction motors, 4, 84-90, 106
Interdigital motor, 81, 82

Jamin, M. J., 55
Jefimenko, O., 54, 84, 105, 117, 118
Johnson, R., 66
Johnson motors, 64, 65, 66

Kämpfer, D., 29, 38
Karpov, Yu., 107, 108, 118
King, W. R., 33
Kinnersley, E., 26
Krasnoperov, V., 107, 108, 118
Kühn, G. J. 110, 118

Langenbucher, J., 23, 29, 33, 37
Langenbucher electrometer, 23, 24
Levitation, electrostatic, 33
Leyden jar, 6, 42, 43
Linear corona motor, 111, 112
Linear induction motor, 90
Liquid-immersed motors, 4, 91-95
Loeb, L. B., 38

McVay, W., 66
McVay motors, 66-71
Magnetic forces, 1
Mascart, M. E., 29, 38, 63, 72
Mayes, T. L., 71
Melcher, J. R., 94, 96
Miller, D. C., 55
Millifleapower motor, 48, 51
Moller, S., 38
Moore, A. D., 45-47, 55, 81-83
Müller-Pouillet, 21

Neuburger, A., 21, 33, 38

Okunev, Yu., 107, 108, 118
Oosting, H. J., 38

Pendulum motors, 3, 6, 15, 16
Perpetuum mobile, electric, 3, 15, 16, 48, 50
Poggendorff, J. C., 37, 56, 59, 60, 71
Poggendorff motor, 56-60, 66, 71

Quincke, G., 91, 93, 96

Reciprocating motor, 15, 17, 18
Richmann, G. W., 7
Rotating field motors, 84-90
Ruhmkorff, 63
Ruhmkorff motor, 62, 63

Scialom, I. N., 96
Secker, P. E., 94, 96, 118
Secker motor, 94, 95
Sivowitch, E. N., 18
Slot effect, 99, 100
Slot-effect motors, 99-103
Spark motors, 4, 39-41, 46, 47, 49, 50, 53, 54
Sparks, J., 21, 38, 54
Sujak, B., 110, 118
Synchronous motors, 4, 77, 102
Syng, P., 33

Taylor, G. I., 94, 96
Thomson, W., 82
Thomson electrometer, 73
Thomson replenisher, 76
Tomlinson, C., 29, 38

Walker, D. K., 54, 99, 105
Weiler, W., 84, 90, 91, 96
Weiler induction motors, 87-89
Weiler liquid-immersed motor, 91, 92
Wilson, B., 37
Wimshurst, J., 33
Wommelsdorf, H., 55
Van Wyck, J. D. N., 110, 118

Zamboni, G., 14
Zamboni pile, 3, 14, 15, 48, 115
Zipernowsky, K., 73, 77
Zipernowsky motor, 73, 74, 82

This book was set in Century Textbook.
The format was designed by the author.
The drawings were done by
Dr. D. K. Walker from the author's sketches.
The printer and binder was
Morgantown Printing & Binding Company.

Picture Credits

Plate 1: courtesy Smithsonian Institution, Washington, D.C.
Plates 2 and 3: courtesy Electret Scientific Company, Star City, W. Va.
Plates 6, 9, 11, and 12: courtesy West Virginia University, Morgantown, W. Va.
Plates 4, 5, 7, 8, and 10: author's photographs.

FREE-POWER AM RECEIVER (P,E)

This project uses a strange way to get its power. It uses power "extracted" from the tuned station to give to the amplifier transistor. With a reasonably strong applied modulation envelope, this type of detector produces a strong demodulated output.

Naturally, the closer you are to a strong station, the more current your radio will be able to supply.

For best results, you must do everything possible to deliver a strong signal to the transistor detector. We recommend a good antenna and ground, the latter preferably being made to a water pipe or solid external ground composed of a pipe driven at least 5 feet into moist earth. This is important for ensuring maximum signal pickup.

If you have plenty of space available, the longer the antenna (up to about 100 feet), the better the results.

Figure 1 shows the schematic diagram of the Free-Power AM receiver. Observe that the heart of the circuit is a germanium transistor that works as a detector and audio amplifier.

Components placement on a terminal strip used as a chassis is shown in ***Figure 2***. The terminal strip can be fixed on an experiment board that can be constructed using some common tools and materials.

Q1 is any germanium transistor, such as GE-2 or 2N107. You can find germanium transistors in old nonworking AM transistor radio receivers. Nonworking AM radios can also supply the variable capacitor and the loopstick (ferrite rod).

Earphone must be a high-impedance crystal type. Low-impedance earphones don't function in this circuit.

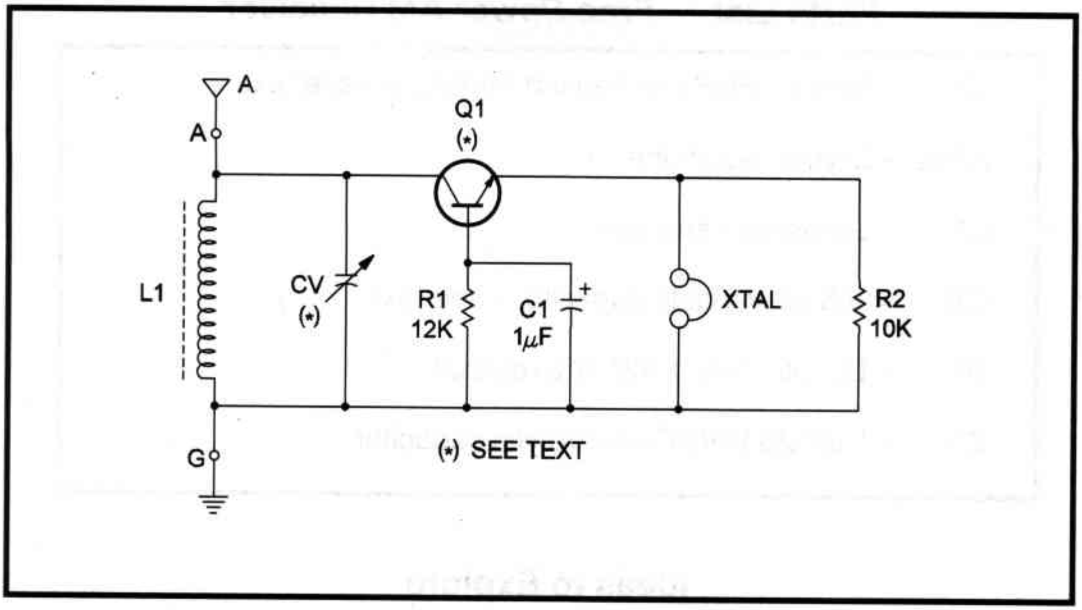
A
A
Q1
(*)
L1
CV
(*)
R1
12K
C1
1μF
+
XTAL
R2
10K
G
(*) SEE TEXT

Figure 1

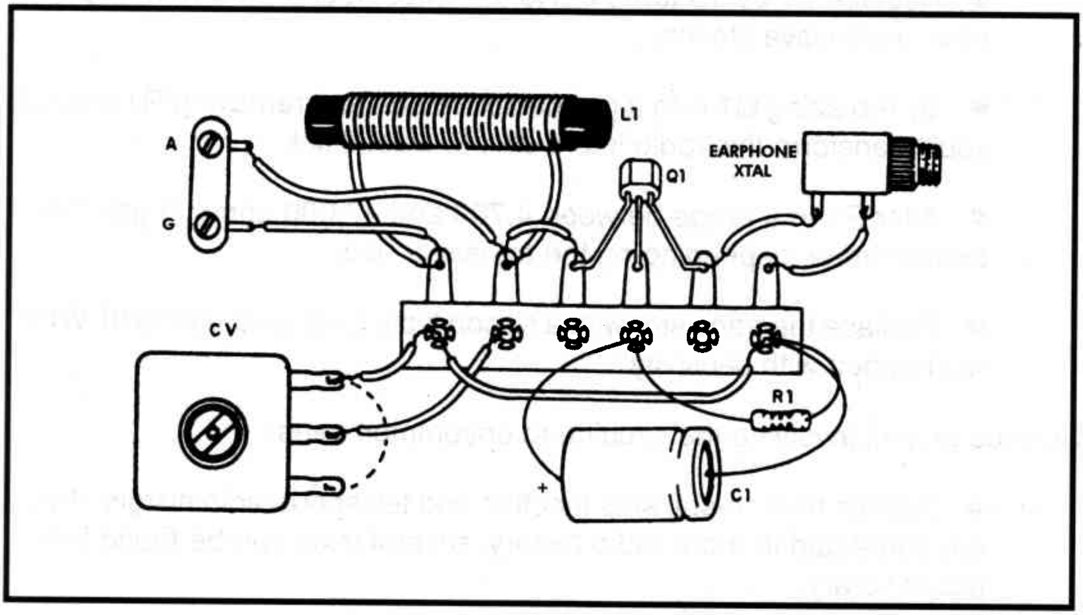
L1
A
G
EARPHONE
XTAL
Q1
CV
R1
+
C1

Figure 2

Parts List — Free-Power AM Receiver

Q1	- NPN or PNP germanium transistor - see text
XTAL	- Crystal earphone
L1	- Loopstick - see text
CV	- 365 pF variable capacitor - see text
R1	- 12,000 ohm, 1/4W, 5% resistor
C1	- 1 uF, 25 WVDC electrolytic capacitor

Ideas to Explore

To learn more about the circuit:

- Replace L1 by 15 turns of No. 28 wire, wound on a ferrite rod, to tune short-wave stations.
- By replacing Q1 with a germanium diode and removing R1 and C1 you'll transform this radio into a simple crystal set.
- Alter R1 in a range between 4,700 and 47,000 ohms to get better performance, depending on the transistor used.
- Replace the transistor with a silicon type, such as the BC548. What will happen with sensivity?

Science project involving the circuit and uncommon uses:

- Explain how crystal sets function and tell about radio history. If you are interested in more radio history, several texts can be found in the public library.
- If you want to learn more about radio receivers, try to find information about heterodyne and super-hereterodyne types.

US 20150102676A1

(19) **United States**

(12) **Patent Application Publication**

Dinwiddie et al.

(10) **Pub. No.: US 2015/0102676 A1**

(43) **Pub. Date: Apr. 16, 2015**

(54) **METHOD AND APPARATUS FOR EXTRACTING AND CONVEYING ELECTRICAL ENERGY FROM THE EARTH'S IONOSPHERE CAVITY**

(71) Applicant: **Earth Energies, Inc.**, Johns Creek, GA (US)

(72) Inventors: **John Dinwiddie**, Cary, NC (US); **Terry L. Wright**, Suwanee, GA (US)

(21) Appl. No.: **14/509,846**

(22) Filed: **Oct. 8, 2014**

Related U.S. Application Data

(60) Provisional application No. 61/889,894, filed on Oct. 11, 2013.

Publication Classification

(51) **Int. Cl.**
H01F 38/14 (2006.01)

(52) **U.S. Cl.**
CPC ***H01F 38/14*** (2013.01); *H01F 2038/146* (2013.01)

(57) **ABSTRACT**

The system and apparatus of one or more embodiments of the present invention extracts, conditions, and conveys electric power from the earth ionosphere cavity through the integrated and collaborative operation of the system and apparatus consisting of a capacitively-coupled insulated elevated terminal (coupled capacitor upper plate), an evacuated spark gap, an integrated step-down transformer and resonant capacitor, and a ground terminal (coupled capacitor lower plate). The architecture governing exemplary embodiments of the system and apparatus of the present invention emulates the natural architecture governing the interaction of living trees with the electrical energy resident in the earth ionosphere cavity. The implementation of such exemplary embodiments of the present invention utilizes standard and customized components appropriate for their function within the system and apparatus.

The extracted electric power from the earth ionosphere cavity manifests in the form of broadband electromagnetic waves in the 0 to 200 Hz frequency range. Electric energy from these waves produce magnetic fields inside the integrated step-down transformer which are added together by the primary coil of the integrated step-down transformer, and delivered as 60 Hz 120/240 VAC to an electric load via the secondary coil of the integrated step-down transformer resonating at 60 Hz due to the influence of the integrated resonant capacitor.

The system and apparatus of exemplary embodiments of the present invention extracts, conditions, and conveys ionosphere cavity-resident electric power in alternating current and direct current realms.

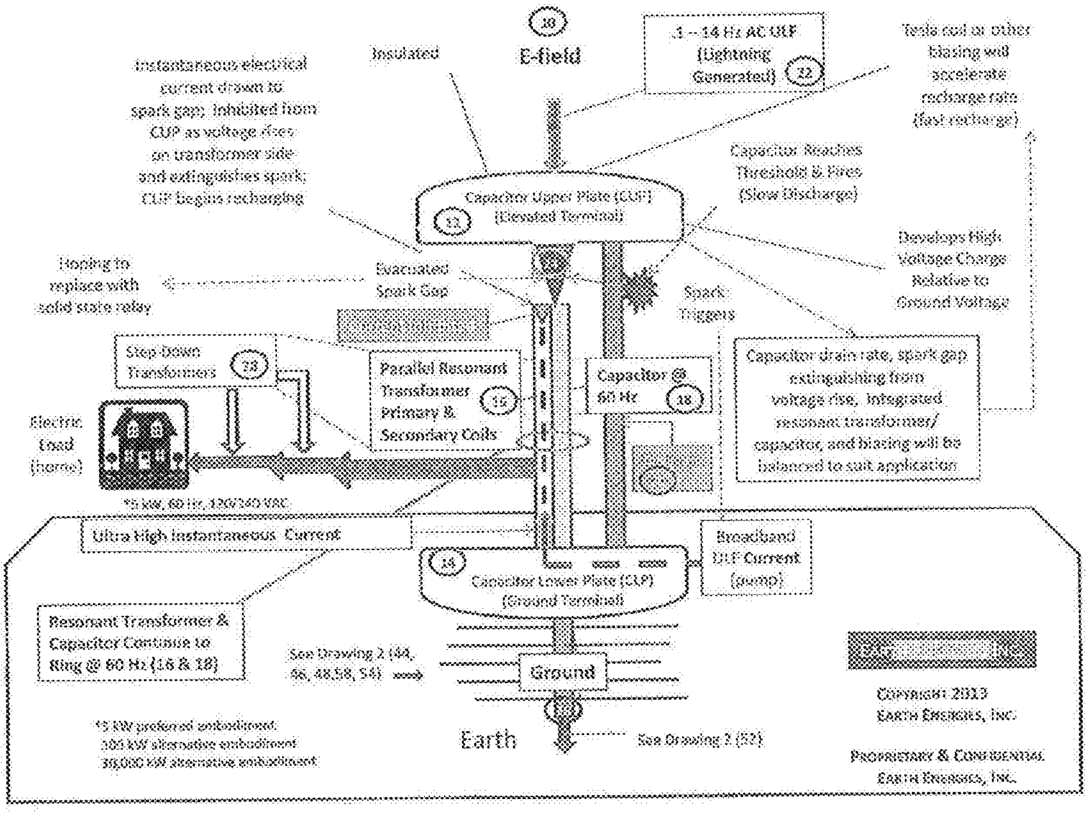
10
E-field
.1 – 14 Hz AC ULF
(Lightning
Generated) 22
Tesla coil or other
biasing will
accelerate
recharge rate
(fast recharge)
Insulated
Instantaneous electrical
current drawn to
spark gap; Inhibited from
CUP as voltage rises
on transformer side
and extinguishes spark;
CUP begins recharging
Capacitor Upper Plate (CUP)
(Elevated Terminal)
12
Capacitor Reaches
Threshold & Fires
(Slow Discharge)
Develops High
Voltage Charge
Relative to
Ground Voltage
Hoping to
replace with
solid state relay
Evacuated
Spark Gap
Spark
Triggers
Step-Down
Transformers
28
Parallel Resonant
Transformer
Primary &
Secondary Coils
15
Capacitor @
60 Hz
18
Capacitor drain rate, spark gap
extinguishing from
voltage rise, Integrated
resonant transformer/
capacitor, and biasing will be
balanced to suit application
Electric
Load
(home)
*5 kW, 60 Hz, 120/240 VAC
Ultra High Instantaneous Current
Broadband
ULF Current
(pump)
14
Capacitor Lower Plate (CLP)
(Ground Terminal)
Resonant Transformer &
Capacitor Continue to
Ring @ 60 Hz (16 & 18)
See Drawing 2 (44,
46, 48,58, 54)
Ground
Earth
See Drawing 2 (52)
*5 kW preferred embodiment
100 kW alternative embodiment
35,000 kW alternative embodiment
COPYRIGHT 2013
EARTH ENERGIES, INC.
PROPRIETARY & CONFIDENTIAL
EARTH ENERGIES, INC.

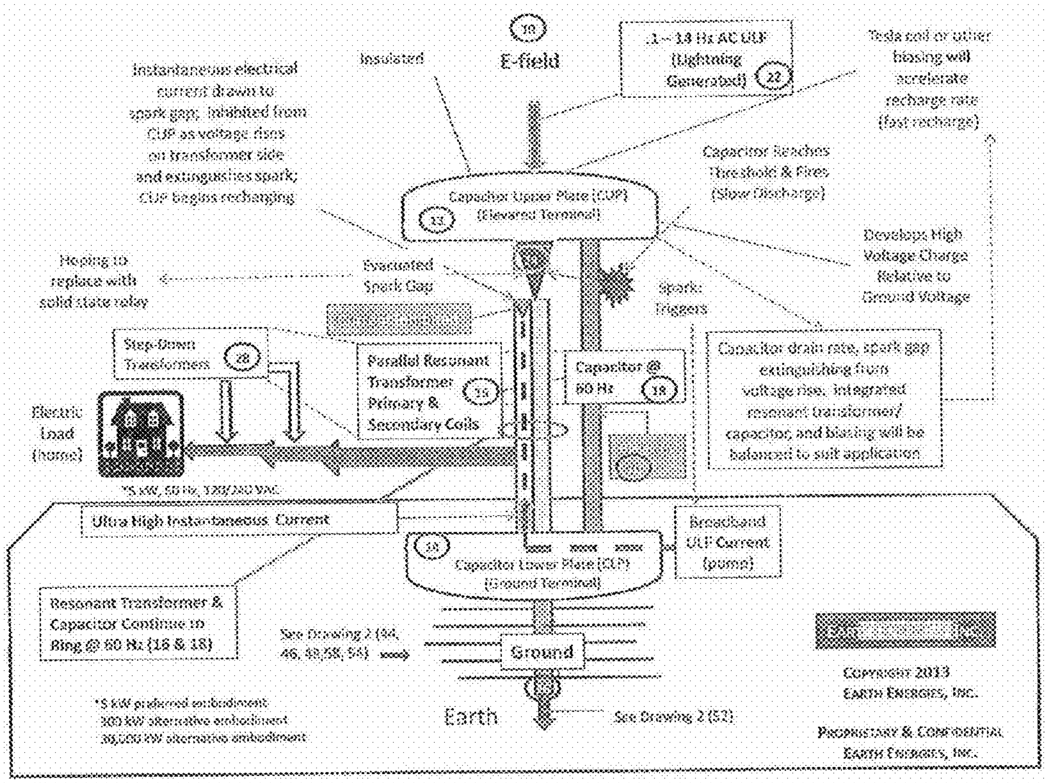
10
E-field
Insulated
.1 – 14 Hz AC ULF (Lightning Generated) 22
Tesla coil or other biasing will accelerate recharge rate (fast recharge)
Instantaneous electrical current drawn to spark gap; inhibited from CUP as voltage rises on transformer side and extinguishes spark; CUP begins recharging
Capacitor Upper Plate (CUP) (Elevated Terminal) 12
Capacitor Reaches Threshold & Fires (Slow Discharge)
Develops High Voltage Charge Relative to Ground Voltage
Hoping to replace with solid state relay
Evacuated Spark Gap
Spark Triggers
Step-Down Transformers 28
Parallel Resonant Transformer Primary & Secondary Coils 15
Capacitor @ 60 Hz 18
Capacitor drain rate, spark gap extinguishing from voltage rise, integrated resonant transformer/ capacitor, and biasing will be balanced to suit application
Electric Load (home)
*5 kW, 60 Hz, 120/240 VAC
Ultra High Instantaneous Current
Broadband ULF Current (pump)
18
Capacitor Lower Plate (CLP) (Ground Terminal)
Resonant Transformer & Capacitor Continue to Ring @ 60 Hz (16 & 18)
See Drawing 2 (44, 46, 48,58, 54)
Ground
Earth
See Drawing 2 (52)
*5 kW preferred embodiment
300 kW alternative embodiment
30,000 kW alternative embodiment
Copyright 2013
Earth Energies, Inc.
Proprietary & Confidential
Earth Energies, Inc.

FIG. 1

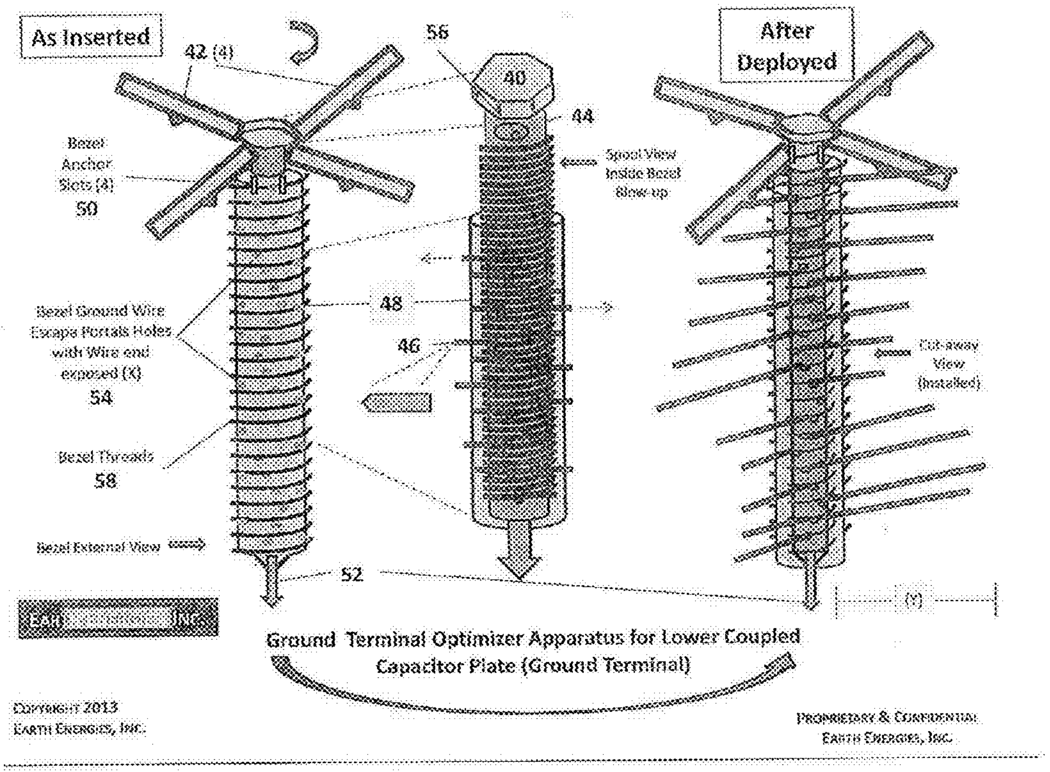
As Inserted
42 (4)
56
40
44
After Deployed
Bezel Anchor Slots (4)
50
Spool View Inside Bezel Blow-up
48
Bezel Ground Wire Escapa Portals Holes with Wire end exposed (X)
54
46
Cut-away View (Installed)
Bezel Threads
58
Bezel External View
52
(Y)
INC.
Ground Terminal Optimizer Apparatus for Lower Coupled Capacitor Plate (Ground Terminal)
COPYRIGHT 2013
EARTH ENERGIES, INC.
PROPRIETARY & CONFIDENTIAL
EARTH ENERGIES, INC.

FIG. 2

METHOD AND APPARATUS FOR EXTRACTING AND CONVEYING ELECTRICAL ENERGY FROM THE EARTH'S IONOSPHERE CAVITY

[0001] This application claims benefit of U.S. Provisional Application No. 61/889894, filed 11 Oct. 2013, titled, "Method and Apparatus for Extracting and Conveying Electrical Energy From the Earth's Ionosphere Cavity," the disclosure of which is incorporated herein by reference in its entirety.

BACKGROUND OF THE INVENTION

[0002] A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by any-one of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright whatsoever.

[0003] The invention is an apparatus and methodology for extracting electric power manifesting as a broadband collection of electric field oscillations in the 0 to 200 Hz frequency range within the Earth's ionosphere cavity, through the use of an elevated biased hemispherical or other curved surface capacitor coupling the ionosphere cavity and Earth, and then conveying this electrical energy at the proper frequency, amperage, and voltage to existing homes and buildings where it is delivered for consumption over existing home or building wiring.

DESCRIPTION OF THE RELATED ART

[0004] The current state of the art in this field is focused on the bending of radio frequency waves in the Earth's ionosphere cavity for the purpose of distorting (through reflection) the resident cavity's magnetic field to physically direct the stimulating electrical energy to targeted locations within the Earth's ionosphere cavity and/or coordinates on the Earth's surface at specific energy levels appropriate to specific applications. The regenerative feedback was done by Sutton's active antenna.

[0005] Nikola Tesla attempted to transmit wireless power in the "natural medium"; however, Tesla bad no description of a device to receive natural oscillations as described in the present disclosure.

[0006] There is no known prior art in the field of electric energy extraction from the ionosphere cavity and conveyance to other terminals within the cavity for delivery of usable electric power.

SUMMARY OF THE INVENTION

[0007] In one exemplary embodiment of the present invention, a properly-scaled electric power extraction/conveyance apparatus is installed at a typical consumer residential home to deliver 60 Hz 120/240 VAC electric power, sufficient to convey the equivalent of 5 kW of continuous electricity, to the home's existing electric service entry component to power electrical appliances in the home over existing home electric wiring.

[0008] In another exemplary embodiment of the present invention, a properly-scaled electric power extraction/conveyance apparatus is installed at an appropriate location within an industrial park or neighborhood to deliver 60 Hz 120/240 V AC electric power, sufficient to convey the equivalent of 300 kW of continuous electricity, to the service entry components of industrial buildings or neighborhood homes to power electrical devices and appliances over existing building or home electric wiring.

[0009] Alternative exemplary embodiments of the present invention will extract direct current electric power from the Earth ionosphere cavity, generated by the earth's rotating magnetosphere, and convey in the form of conditioned direct current to direct current load sources, or convert to alternating current and convey in the form of conditioned alternating current to alternating current load sources.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] A disclosure embodiments of the present invention, is set forth more particularly in the remainder of the specification, which makes reference to the accompanying figures, in which:

[0011] FIG. **1** provides an example of an electric power extraction and conveyance apparatus for one residential-scale embodiment of the present invention. Certain additional example embodiments of the present invention utilize similar fundamental technologies, but differ in the scale (size) of the components and characteristics of the electricity conveyed for consumption; and

[0012] FIG. **2** is a schematic illustration of ground terminals for use with embodiments of the present invention.

DETAILED DESCRIPTION

[0013] Applicant has discovered that trees extract and use electric power from the Earth ionosphere cavity. The present invention emulates the architecture of electric power extraction and usage methodology inherent in the natural operation of living trees, but uses standard and customized components, common in the field, for its implementation. The natural electric power extraction and use of electric power by trees leverages the shape and physical composition of trees as instances of large electrical capacitors coupled to the ionosphere cavity. The canopy of trees (upper limbs and leaf structures) emulates the upper plates of a coupled capacitor, the trunks of trees emulate a combined dielectric and resonant transformer/capacitor, and the root systems of trees emulate ground terminals (lower coupled capacitor plates). The extraction of electric power from the ionosphere cavity by trees is in the form of ultra-low frequency broadband waves (oscillations) in the frequency range of 0.1 to 14 Hertz (ULF). Inherent tree physical composition and control structure (i.e., DNA/RNA) govern the use of the extracted electric power to accomplish tree growth, seasonal metabolic adjustments, and reproductive processes. Any excess extracted electric power by trees is returned to the ionosphere cavity. The present invention is an exemplary instance of this same natural architecture using standard and customized components that facilitate extraction and conveyance of ionosphere cavity-resident broadband electric power at useful scales sufficient to provide continuous electrical power for electrical devices used by people.

[0014] One preferred exemplary embodiment of the present invention is in the form shown in Drawing **1** above and will deliver continuous 60 Hz 120/240 V AC electric power to consumer residences (and other structures with similar electricity conditioning and consumption requirements) equivalent to electric power capacity of 5 kilowatts.

[0015] A secondary exemplary embodiment of the present invention is also in the form shown in FIG. **1**, appropriately scaled for larger electrical power demands and related conveyance capacity, and will deliver continuous 60 Hz 120/240 V AC electric power to larger buildings or electrical distribution facilities for commercial office parks with electric power capacity of 300 kilowatts.

[0016] A third exemplary embodiment of the present invention is also in the form shown in FIG. **1**, appropriately scaled for larger electrical power capacity, and will deliver continuous 60 Hz 120/240 VAC electric power to utility substations with electric power capacity of 30,000 kilowatts (or higher).

[0017] The following system, power sources, and component descriptions provide additional details of the present invention in various embodiments.

[0018] Earth's ionosphere cavity (**10**) is a reservoir of continuously-generated electrical energy (E-Field) in both direct current (DC) and alternating current (AC) realms. The DC electric energy is created by the constant rotation of the Earth's magnetosphere exposed to the solar wind. Earth's perpetually moving magnetic fields create the DC electric power in the ionosphere cavity E-Field per the equations in James Clerk Maxwell's fundamental theory of electromagnetism. AC electrical energy in the ionosphere cavity is generated from the persistent E-Field disturbances (i. e., oscillations) caused by lightning discharges. The present invention includes the direct extraction of AC electric power from the ionosphere cavity, and the direct extraction of DC electric power from the ionosphere cavity which is converted to AC electric power before conveyance. Optional operational configurations of the present invention will also enable conveyance of extracted DC electric power for potential scenarios where the conveyance of conditioned DC electricity is required.

[0019] Electric-field oscillations (**22**) are induced by lightning and encapsulated in the earth-ionosphere cavity, creating a continuous source of renewable electrical energy. The frequency of lightning occurrences in the ionosphere cavity averages thirty (30) strikes per second, with each strike representing approximately one (1) terawatt (TW) of instantaneous energy. Assuming an average energy dissipation rate for each lightning strike of fifteen (15) microseconds (μsec), i. e., the time involved for one (1) terawatt of instantaneous energy to dissipate to zero (0) energy, these lightning-induced electric field oscillations are sufficient to provide 4,500 times the entire electric energy consumption on planet Earth at the projected consumption rate for the year 2100 (55.3 TWh annually). Lightning-induced E-Field oscillations vary between 0.1 and 14 Hertz with wavelengths exceeding the practicality of an electromagnetic antenna for collecting the energy. The present invention utilizes capacitive coupling to overcome this obstacle.

[0020] An insulated upper capacitor plate (terminal) (12) elevated above the influence of the ground floor collects E-Field oscillations from the ionosphere cavity. This insulated elevated terminal, capacitively-coupled to the E-Field bounded by the earth-ionosphere cavity, is arranged in a surface of large radii of curvature, supporting maximum surface connection to the E-Field without leaking the high voltage charge applied to the elevated terminal. The capacitance and resistance of the elevated terminal are responsive to the reception of broadband electric-field frequencies in the 0 to 200 hertz range. The elevated terminal is subject to high voltage alternating current biasing used to collect electric-field oscillations generated by lightning impulses in the earth-ionosphere cavity.

[0021] An evacuated spark gap (**24**) connected to the elevated terminal (**12**) prevents electrical discharge occurring between the elevated terminal and the ionosphere cavity. As the high voltage alternating current approaches peak voltage the spark gap energizes and a large voltage is applied to the step-down transformer primary coil (**16**). The high voltage pumps high current from the ground terminal (**14**, **26**) converting the electric-field to a magnetic field within the transformer (**16**). The step-down transformer is connected to a 60 hertz resonant capacitor (**18**), the secondary side of the transformer resonating at 60 hertz will supply a 60 hertz 120/240 VAC power to the load. Filter circuits will condition the power. The evacuated spark gap function may be accomplished by solid state circuitry in the implementation of the present invention.

[0022] The resonant transformer (**16**) is integrated with a parallel resonant capacitor across the transformer where the integrated combination resonates at 60 hertz supplying power to the electricity consuming structure (home or business). Various step-down transformers (**28**) are utilized to condition the conveyed electricity to accommodate the requirements of the receiving station for various embodiments of the present invention.

[0023] Ground (**26**) is the source of electric current for the present invention. A ground terminal, connected to the soil, is arranged in a surface of large radii of curvature to permit instantaneous sourcing of high currents from earth ground without temperature and voltage rise inhibiting the ground terminal collection of current. The capacitance and resistance of the ground terminal to earth is minimized to promote on demand current sourcing. Drawing **2** illustrates an exemplary apparatus of the present invention that optimizes ground connection quality by maximizing the penetration extent and resulting surface area of the ground terminal's extension into the soil.

[0024] Step-down transformers (**28**) are scaled to transform the electrical current and voltage into structured and conditioned electricity that is compatible with the electric power needs of the load sources for various embodiments of the present invention.

[0025] A mechanical nut (**40**) solidly attached to shaft **44** to facilitate extension, when turned, of large gauge copper wire extrusions (**46**).

[0026] A downward pointed anchor flat shaft (**42**) attached to the bezel **50** via a T-slot in bezel.

[0027] A copper (or other highly-conductive metal) spool base (**44**) sourcing "X" independently attached large gauge pointed copper wires (**46**) of length "Y", where "X" is equal to the number of escape portals in bezel (**50**), and "Y" is the length determined to yield optimum surface area of wires of type **46** for high quality connectivity to soil. The spool can freely rotate inside the bezel (**48**) and emulates the primary root structure of a living tree.

[0028] Sharpened large gauge (reinforced) copper wires (**46**), quantity "X", of length "Y" emulating the root structure of living trees.

[0029] Threaded bezel (spool-housing sleeve) (**48**) with drilled escape portal holes with end of copper wires (**46**) protruding through holes ½ inch.

[0030] Bezel anchor slots (**50**) of twice the height of anchor flat shafts (**42**) and T-slotted to accommodate anchor flat shafts.

[0031] Extended pointed end (**52**) of inner spool base shaft.

[0032] Escape portals (**54**) in bezel to allow for large gauge copper wire (**46**) extension into soil.

[0033] Connector port (**56**) for connection to ground terminal (**14**) in Drawing **1**.

[0034] Bezel threads (**58**).

[0035] Various embodiments of the present invention will include one or more of the following characteristics in various combinations. These descriptions are provided for purposes of example. Not all of the following characteristics will be required for any given exemplary embodiment.

[0036] Characteristic 1, The form (shape), physical composition, size, insulation, location altitude, discharge capacitive threshold and discharge rate, and the strength of AC or DC biasing applied to the (CUP) upper coupled capacitor plate (**12**: elevated terminal) all form an integrated whole, the purpose of which is to extract and temporarily store DC or AC electric power from the Earth's ionosphere cavity (**10**) at high voltage, and to trigger the energizing of the evacuated spark gap (**24**).

[0037] Characteristic 2, The elevated terminal (**12**) is capacitively coupled to the E-Field (**10**) bounded by the earth ionosphere cavity, and arranged in a surface of large radii of curvature, supporting maximum surface connection to the E-Field without leaking the high voltage charge applied to the elevated terminal (**12**).

[0038] Characteristic 3, The capacitance and resistance of the elevated terminal (**12**) are responsive to the reception of broadband electric-field frequencies in the 0 to 200 hertz range.

[0039] Characteristic 4, The elevated terminal (**12**) is subject to high voltage alternating current biasing used to collect E-Field oscillations generated by lightning impulses in the earth-ionosphere cavity.

[0040] Characteristic 5, An alternative instantiation of elevated terminal (**12**) is subject to high voltage direct current biasing used to collect E-Field direct current generated by the Earth's rotating magnetosphere.

[0041] Characteristic 6, The evacuated spark gap (**24**) may be in the form of a cohesive physical spark ignition or an operational emulation of this function through solid state circuitry.

[0042] Characteristic 7, The evacuated spark gap (**24**) connected to the elevated terminal (**12**) prevents electrical discharge occurring between the elevated terminal and the ionosphere cavity.

[0043] Characteristic 8, The high voltage alternating current from the E-Field, induced into the elevated terminal (**12**), approaches peak voltage and causes the evacuated spark gap (**24**) to energize and a large voltage is applied to the step-down transformer primary coil (**16**).

[0044] Characteristic 9, The high voltage alternating current impulse from the elevated terminal (**12**), which causes the evacuated spark gap (**24**) to energize, manifests as alternating current on broadband frequencies between 0.1 and 200 Hz.

[0045] Characteristic 10, In an alternative instantiation of the elevated terminal (**12**) the stored high voltage direct current, which causes the evacuated spark gap (**24**) to energize, manifests as direct current on broadband frequencies between 0 and 200 Hz.

[0046] Characteristic 11, The high voltage applied to the step-down transformer primary coil (**16**) pumps high current from the ground terminal (**14**), converting the electric-field to a magnetic field within the transformer (**16**).

[0047] Characteristic 12, The high current pumped from the ground terminal (**14**), caused by high voltage applied to the step-down transformer primary coil (**16**), will attempt to normalize the field and extinguish the evacuated spark gap (**24**).

[0048] Characteristic 13, The transformer's (**16**) magnetic field expands in response to the high current electric-field inrush from the ground terminal (**14**).

[0049] Characteristic 14, The primary winding inside the step-down transformer (**16**) will add the energy from each broadband frequency together into a single magnetic field, where it will become a 60 Hz frequency at the secondary winding of the transformer (**16**).

[0050] Characteristic 15, The step-down transformer (**16**) is connected to a 60 Hertz resonant capacitor (**18**).

[0051] Characteristic 16, The secondary (coil) side of the step-down transformer (**16**), resonating at 60 hertz, will supply a 60 hertz 120/240 VAC power to the electric load.

[0052] Characteristic 17, Filter circuits within the step-down transformer will condition the power according to the electric power needs of the electric load.

[0053] Characteristic 18, Electric-field lines exist (within the Earth's ionosphere cavity) that are horizontal to the surface of the Earth and extend to heights above 30,000 feet, with voltage gradients typically 100 Volts per vertical meter.

[0054] Characteristic 19, Lightning impulses create global broadband oscillations in the Earth ionosphere cavity electric-field between 0.1 and 14 Hertz.

[0055] Characteristic 20, Rotation of the Earth's magnetosphere creates global electric field oscillations in the Earth ionosphere cavity that manifest as electromagnetic waves in the 0 to 200 Hertz frequency range.

[0056] Characteristic 21, Global broadband oscillations in the Earth ionosphere cavity electric-field can be used for wireless energy transmission without harm to people, plants, or trees.

[0057] Characteristic 22, A maximum quality ground connection is achieved by expanding the surface area of the ground connection (**48** and **44**) through extension of high-conductor large gauge copper wires (**46**) horizontally through bezel (**48**) escape portals (**54**) along depth of source spool (**44**) and bezel (**48**).

[0058] Characteristic 23, The technique described in Characteristic 22 effectively expands the diameter of the ground source spool (**44**) with regard to quality of ground connection.

[0059] Characteristic 24, Bezel threads (**58**) expand surface area of bezel (**48**) touching earth, enhancing ground connection quality.

[0060] Characteristic 25, Bezel threads (**58**) enhance ground connection quality of bezel (**48**), source spool (**44**), and high conductor large gauge copper wires (**46**) through rotational insertion and resulting friction hold v. other (direct drive) insertion methods which can result in corruption of ground hole integrity caused by imprecise angle and/or vibration during insertion.

[0061] While one or more embodiments of the present invention have been described above, it should be understood that any and all equivalent realizations of the present invention are included within the scope and spirit thereof. Thus, the embodiments presented herein are by way of example only and are not intended as limitations of the present invention. Therefore, it is contemplated that any and all such embodiments are included in the present invention.

What is claimed is:

1. A power receiver for extracting electrical energy from the earth's electric field, said power receiver comprising:

a resonant transformer connected between a ground terminal disposed below the surface of the earth and an elevated terminal;

impulse generator for generating and applying a high voltage electrical impulse to a primary winding of the resonant transformer to induce current flow from the ground terminal through the primary winding of the transformer; and

a power conversion circuit connected to a secondary winding of the resonant transformer to convert electrical current flowing through the secondary winding to a desired form.

2. The power receiver of claim **1** wherein a resonant frequency of the resonant transformer is below 200 Hz.

3. The power receiver of claim **1** wherein the elevated terminal comprises an upper capacitive plate coupled to the earth's ionosphere cavity.

4. The power receiver of claim **3** wherein the impulse generator comprises:

the upper capacitive plate;

a pair of electrodes separated by a spark gap, said electrodes connected between the upper capacitive plate and the primary winding of the resonant transformer and configured to generate a spark when a voltage difference between the electrodes reaches a predetermined level.

5. The power receiver of claim **1** further comprising a lower capacitive plate coupled to the ground terminal.

6. The power receiver of claim **1** wherein the resonant transformer comprises a capacitor connected in parallel with a primary winding of the resonant transformer.

* * * * *

US008686575B2

(12) **United States Patent**
McCowen

(10) **Patent No.: US 8,686,575 B2**
(45) **Date of Patent: Apr. 1, 2014**

(54) **ENERGY COLLECTION**

(75) Inventor: **Clint McCowen**, Navarre, FL (US)

(73) Assignee: **Ion Power Group, LLC**, Navarre, FL (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **12/255,130**

(22) Filed: **Oct. 21, 2008**

(65) **Prior Publication Data**

US 2009/0040680 A1 Feb. 12, 2009

Related U.S. Application Data

(63) Continuation of application No. 11/358,264, filed on Feb. 21, 2006, now Pat. No. 7,439,712.

(51) **Int. Cl.**

F02B 63/04	(2006.01)
F03G 7/08	(2006.01)
H02K 7/18	(2006.01)
H02N 1/00	(2006.01)

(52) **U.S. Cl.**
USPC .... **290/1 R**; 322/2 A

(58) **Field of Classification Search**
USPC .... 290/1 R; 322/2 A; 126/660; 60/641.8, 60/641.11, 641.12, 659, 698; 244/158.2
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

674,427 A		5/1901	Palencsar	307/149
911,260 A		2/1909	Pennock	307/149
1,014,719 A		1/1912	Pennock	307/149
2,473,819 A	*	6/1949	Pittman	315/36
3,532,959 A	*	10/1970	Erickson	320/166
3,780,722 A	*	12/1973	Swet	126/680
3,946,227 A		3/1976	Bingham	250/281
4,079,225 A	*	3/1978	Warner	219/110
4,092,250 A		5/1978	Sano et al.	210/500.29
4,104,696 A		8/1978	Cochran, Jr.	361/229
4,146,800 A	*	3/1979	Gregory et al.	290/44
4,180,056 A	*	12/1979	Schnabel et al.	126/660
4,201,197 A	*	5/1980	Dismer	126/600
4,206,396 A	*	6/1980	Marks	322/2 A
4,224,496 A	*	9/1980	Riordan et al.	219/110
4,307,936 A	*	12/1981	Ochiai	385/120
4,314,192 A	*	2/1982	Caro	322/2 R
4,346,478 A	*	8/1982	Sichling	398/107
4,425,905 A	*	1/1984	Mori	126/578
4,433,248 A	*	2/1984	Marks	290/44
RE31,678 E	*	9/1984	Ochiai	385/115

(Continued)

FOREIGN PATENT DOCUMENTS

WO	2007098341	8/2007

OTHER PUBLICATIONS

"Emissivity Values for Common Materials & Non-Metals", Raytek—A Fluke Company, http://www.raytek.com/Raytek/en-r0/IREducation/EmissivityNonMetals.htm, 1999.*

Primary Examiner — Pedro J Cuevas
(74) *Attorney, Agent, or Firm* — Benjamin A. Balser; Balser & Grell IP Law

(57) **ABSTRACT**

An energy collection system may collect and use the energy generated by an electric field. Collection fibers are suspended from a support wire system supported by poles. The support wire system is electrically connected to a load by a connecting wire. The collection fibers may be made of any conducting material, but carbon and graphite are preferred. Diodes may be used to restrict the backflow or loss of energy.

25 Claims, 10 Drawing Sheets

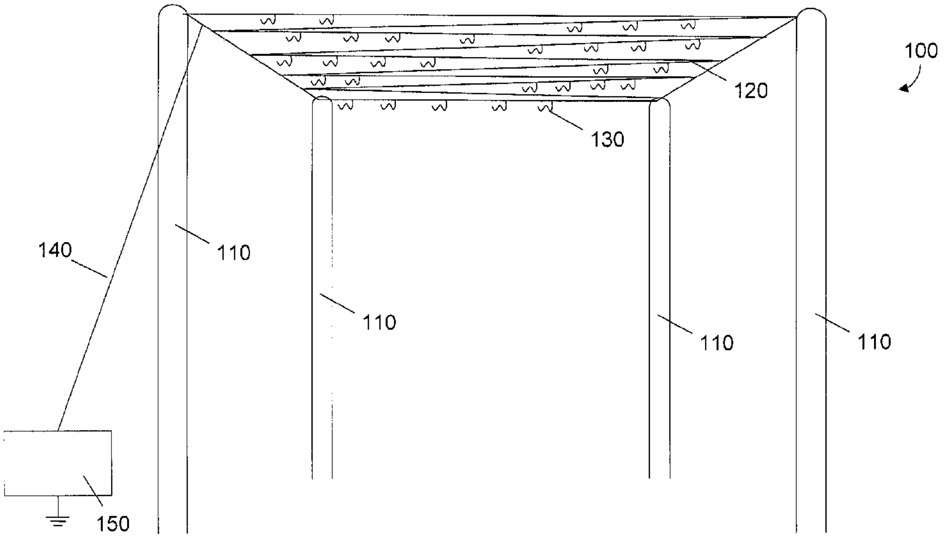

(56) References Cited

U.S. PATENT DOCUMENTS

4,483,311	A *	11/1984	Whitaker	126/602
4,489,269	A *	12/1984	Edling et al.	322/2 R
4,512,335	A *	4/1985	Mori	126/678
4,653,223	A *	3/1987	Mori	47/17
4,653,472	A *	3/1987	Mori	126/683
4,676,226	A *	6/1987	Mori	126/680
4,809,675	A *	3/1989	Mori	126/578
4,852,454	A *	8/1989	Batchelder	89/1.11
4,943,125	A *	7/1990	Laundre' et al.	385/35
4,988,159	A *	1/1991	Turner et al.	385/33
5,047,892	A	9/1991	Sakata et al.	361/231
5,114,101	A *	5/1992	Stern et al.	244/172.8
5,145,257	A *	9/1992	Bryant et al.	374/131
5,379,103	A *	1/1995	Zigler	356/73
5,851,309	A *	12/1998	Kousa	136/248
5,942,806	A *	8/1999	Veliadis	290/1 R
6,025,591	A	2/2000	Taylor et al.	250/292
6,038,363	A *	3/2000	Slater et al.	385/147
6,116,544	A *	9/2000	Forward et al.	244/158.2
6,173,922	B1 *	1/2001	Hoyt et al.	244/158.2
6,226,440	B1 *	5/2001	Lyons	385/119
6,419,191	B1 *	7/2002	Hoyt et al.	244/158.2
6,425,391	B1 *	7/2002	Davoren et al.	126/683
6,735,830	B1	5/2004	Merciel	250/324
6,765,212	B2 *	7/2004	Goetz et al.	250/339.11
6,846,447	B2 *	1/2005	Mbachu et al.	264/406
6,853,447	B2 *	2/2005	Goetz	356/237.1
6,894,772	B2 *	5/2005	Goetz et al.	356/237.1
6,925,852	B2 *	8/2005	Susko	73/23.2
6,974,110	B2	12/2005	Grandics	244/171.5
7,109,597	B1 *	9/2006	Bose	290/43
7,128,867	B2 *	10/2006	Mbachu et al.	264/406
7,231,809	B2 *	6/2007	Susko	73/31.05
7,259,475	B2 *	8/2007	Hong et al.	307/64
7,439,712	B2 *	10/2008	McCowen	322/2 A
7,449,668	B2 *	11/2008	Schutten et al.	250/214 SW
7,478,712	B2 *	1/2009	McCowen	191/45 R
7,532,819	B1 *	5/2009	Triebes et al.	398/121
2002/0026933	A1 *	3/2002	Gottlieb	126/628
2002/0108892	A1 *	8/2002	Goetz et al.	209/576
2002/0109094	A1 *	8/2002	Goetz et al.	250/339.11
2002/0109835	A1 *	8/2002	Goetz	356/237.1
2003/0107869	A1	6/2003	Pan et al.	361/306.2
2003/0125630	A1 *	7/2003	Furnish	600/461
2003/0193319	A1 *	10/2003	Wood et al.	322/2 A
2004/0011925	A1 *	1/2004	Grandics	244/62
2004/0083793	A1 *	5/2004	Susko	73/31.05
2004/0094853	A1 *	5/2004	Mbachu et al.	264/40.1
2004/0160711	A1	8/2004	Stumberger	361/20
2005/0101024	A1 *	5/2005	Mbachu et al.	436/85
2005/0136311	A1 *	6/2005	Ueda et al.	429/30
2005/0270525	A1 *	12/2005	Susko	356/326
2007/0107765	A1 *	5/2007	Schutten et al.	136/243
2007/0195481	A1 *	8/2007	McCowen	361/233
2007/0218323	A1 *	9/2007	Sudo et al.	429/13
2007/0223171	A1	9/2007	Guy et al.	361/126
2007/0273206	A1 *	11/2007	McCowen	307/9.1
2007/0297892	A1 *	12/2007	Kildegaard	415/1
2008/0145714	A1 *	6/2008	Kagami	429/13
2009/0107842	A1	4/2009	Park et al.	204/554
2009/0114495	A1 *	5/2009	McCowen	191/45 R
2009/0216292	A1 *	8/2009	Pless et al.	607/33
2010/0090562	A1 *	4/2010	McCowen	310/309
2010/0090563	A1 *	4/2010	McCowen	310/309
2012/0164555	A1 *	6/2012	Yamazaki et al.	429/482
2012/0299559	A1 *	11/2012	McCowen	322/2 A
2013/0276776	A1 *	10/2013	Te Kamp et al.	126/660

* cited by examiner

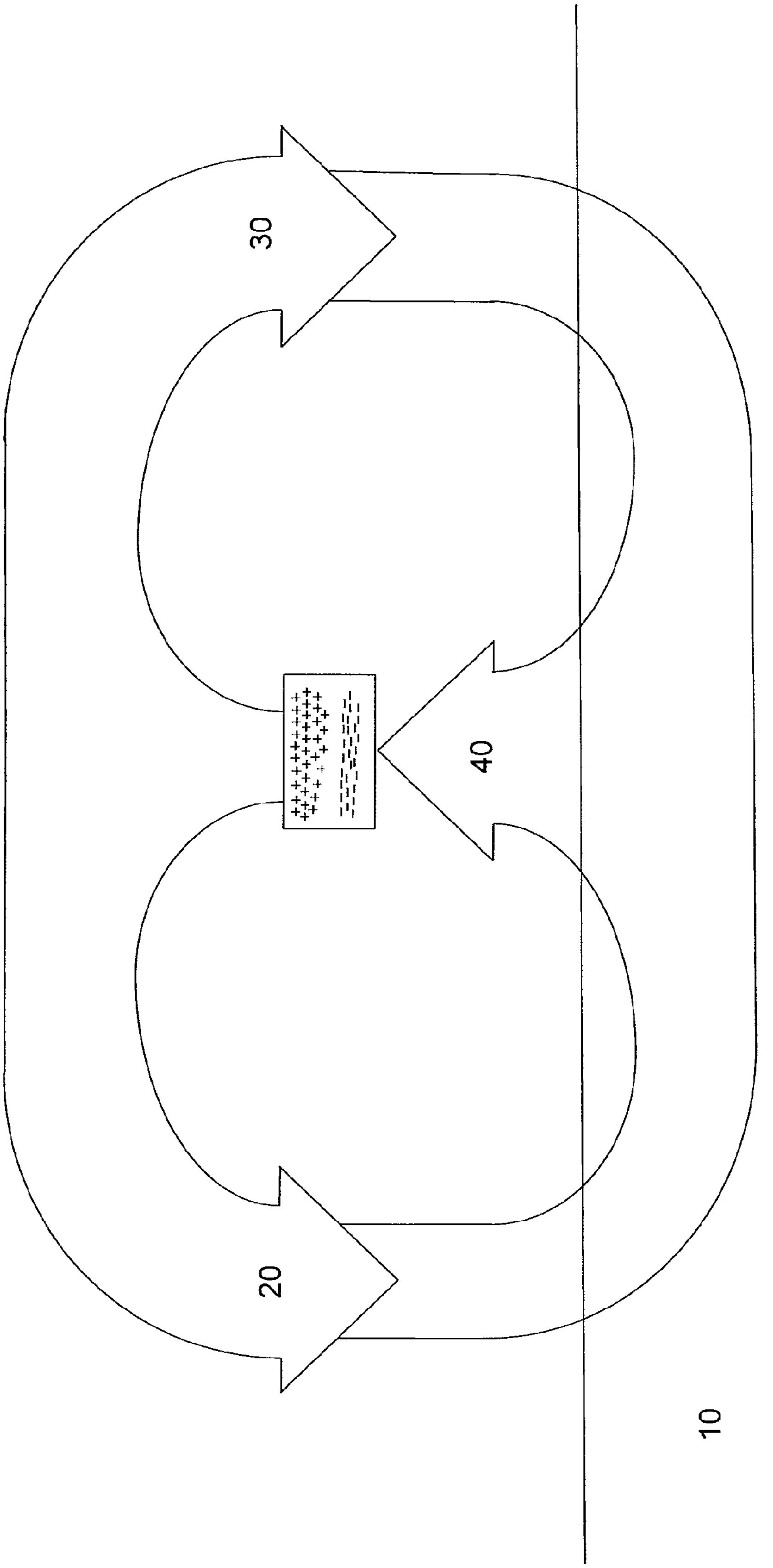

FIGURE 1

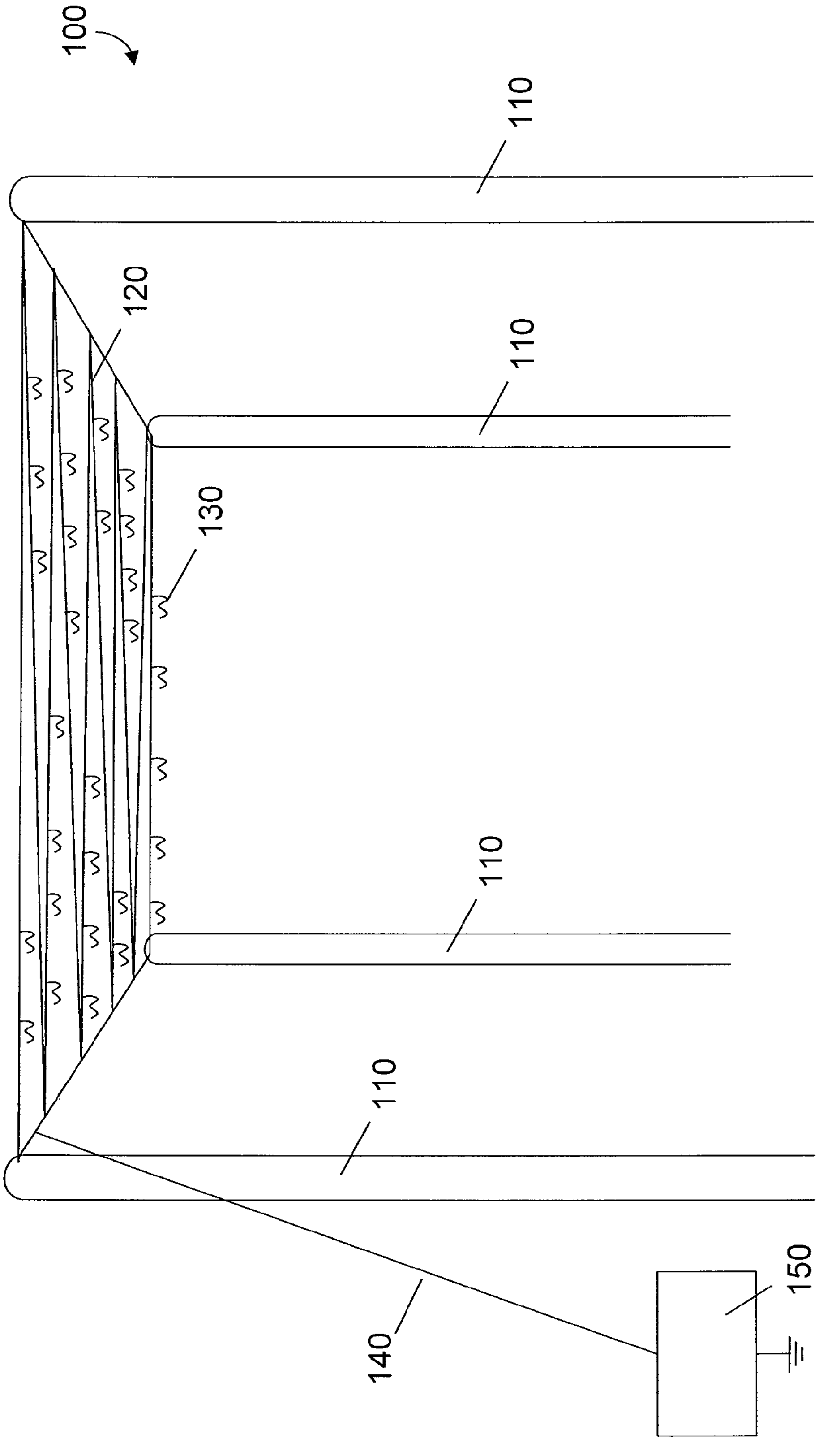
100
110
110
120
130
110
110
140
150

FIGURE 2

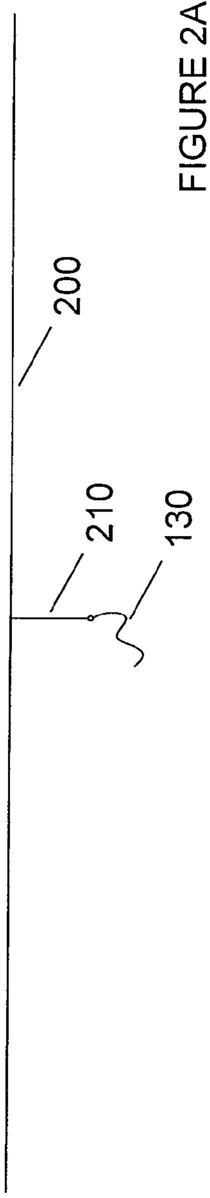

FIGURE 2A

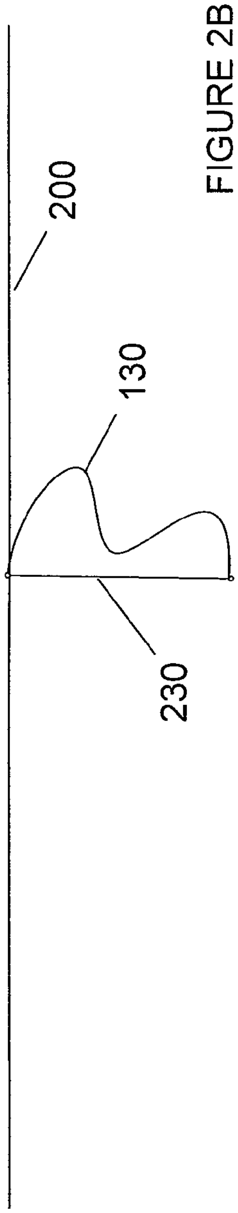

FIGURE 2B

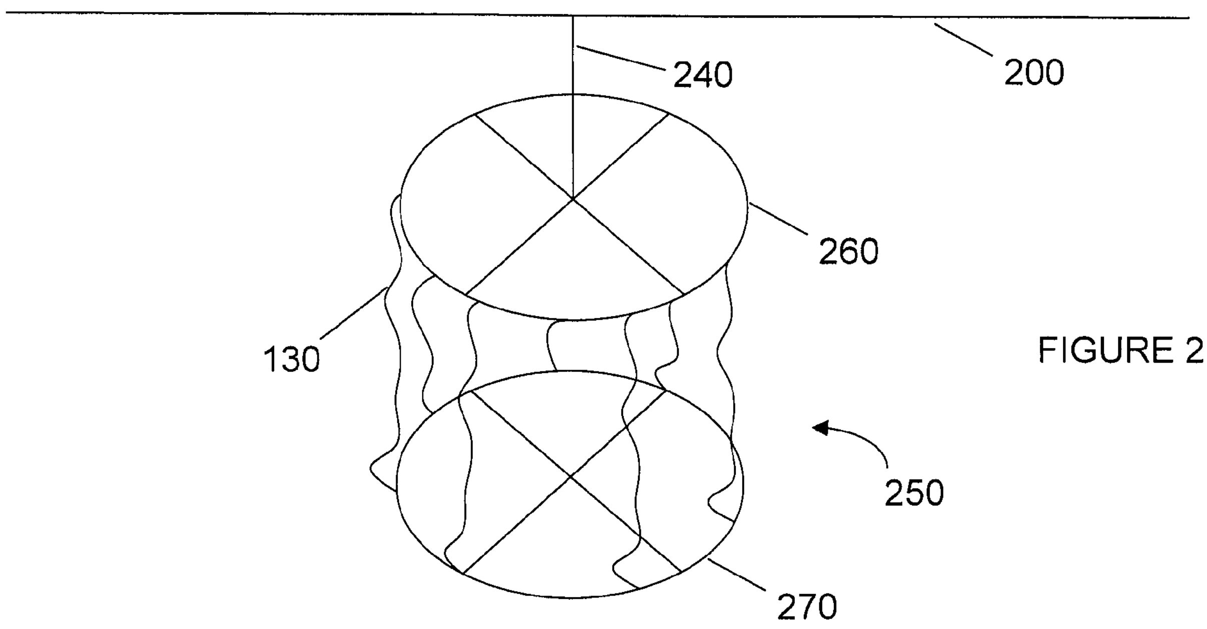
240
200
260
130
250
270

FIGURE 2C

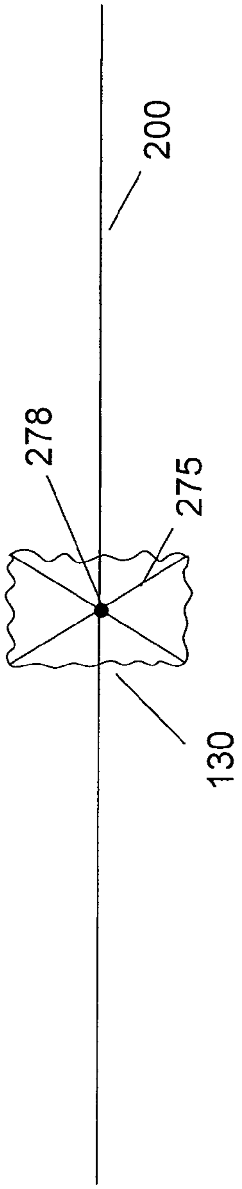
200
278
275
130

FIGURE 2D

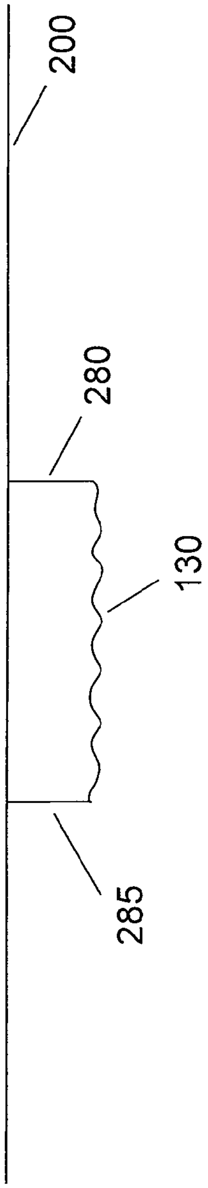
200
280
130
285

FIGURE 2E

FIGURE 2F

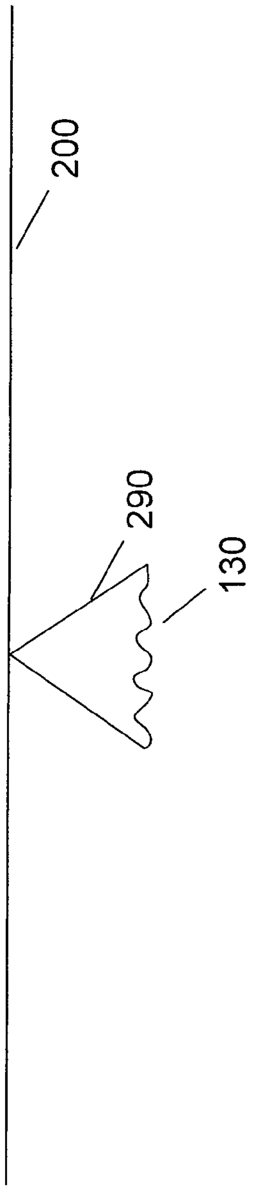

FIGURE 2G

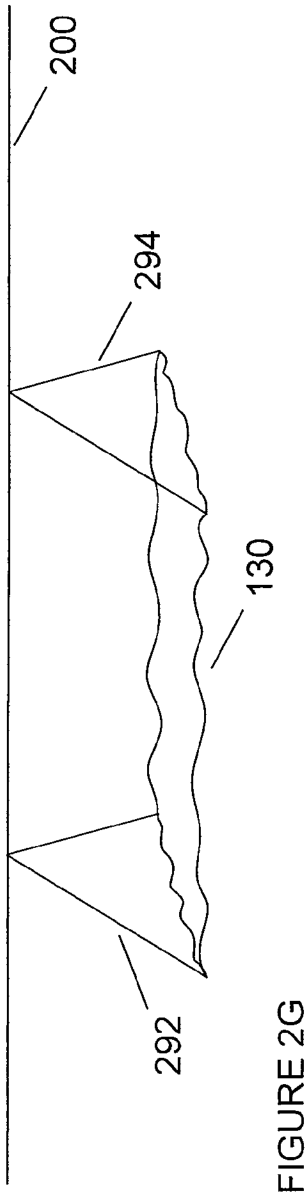
200
294
130
292

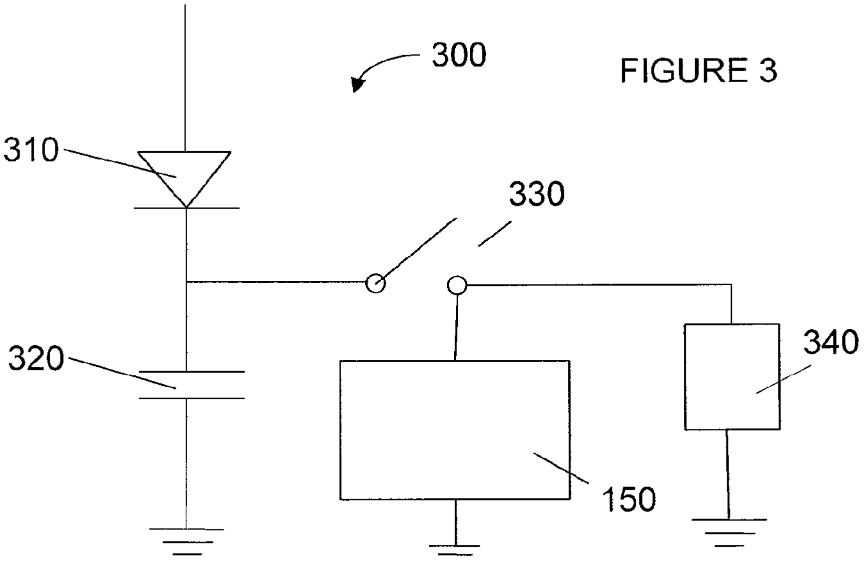
300
310
330
320
340
150

FIGURE 3

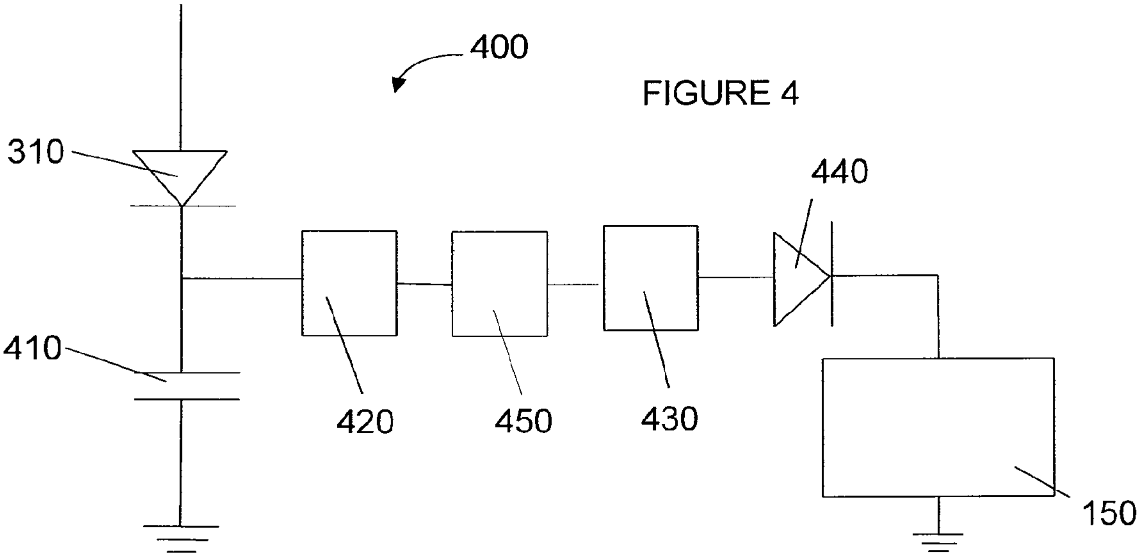
400
310
440
410
420
450
430
150

FIGURE 4

FIGURE 5

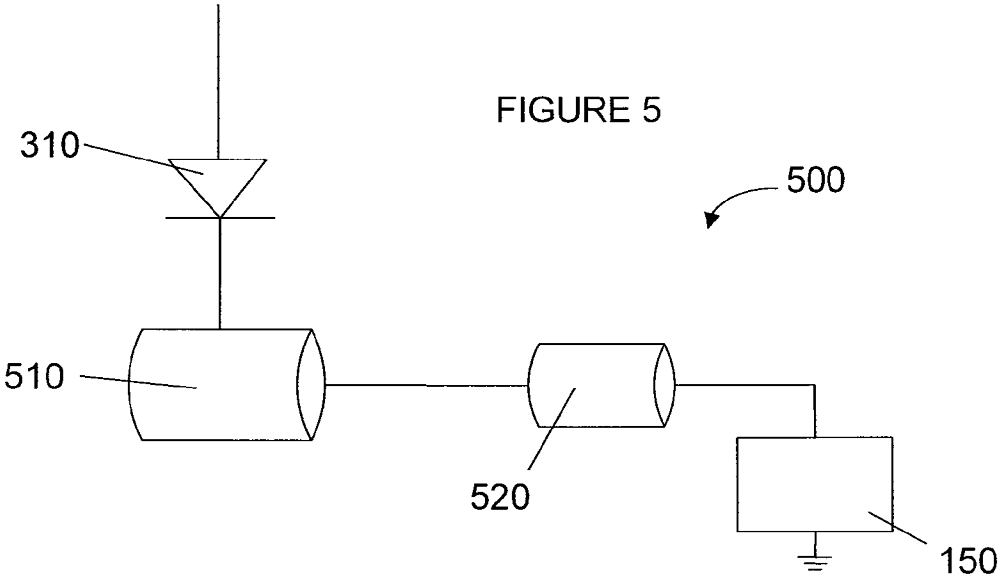
310
500
510
520
150

FIGURE 6

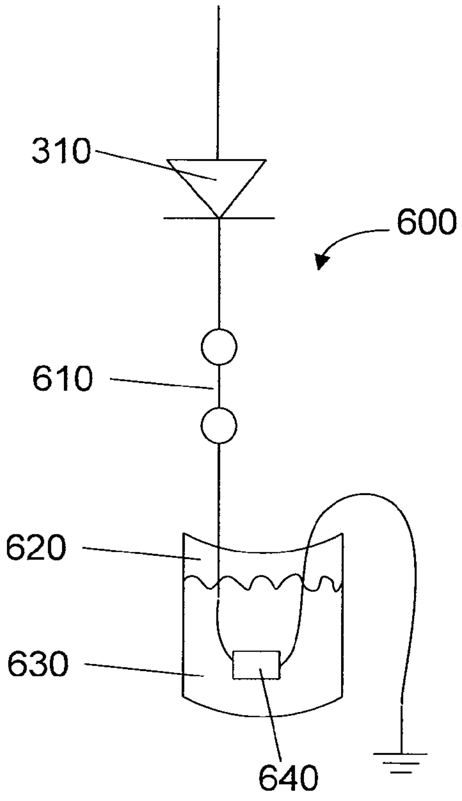
310
600
610
620
630
640

FIGURE 7

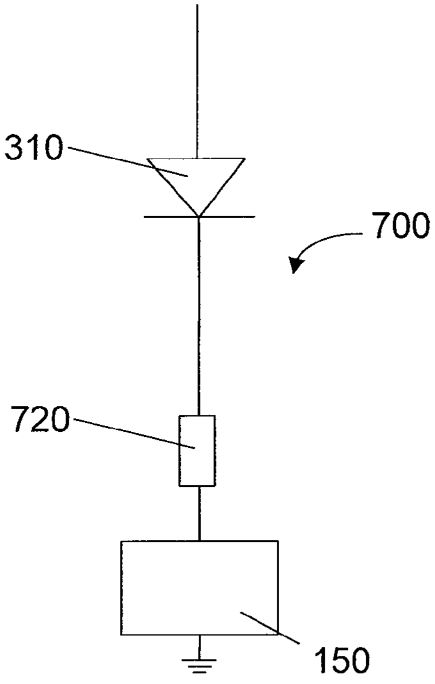
310
700
720
150

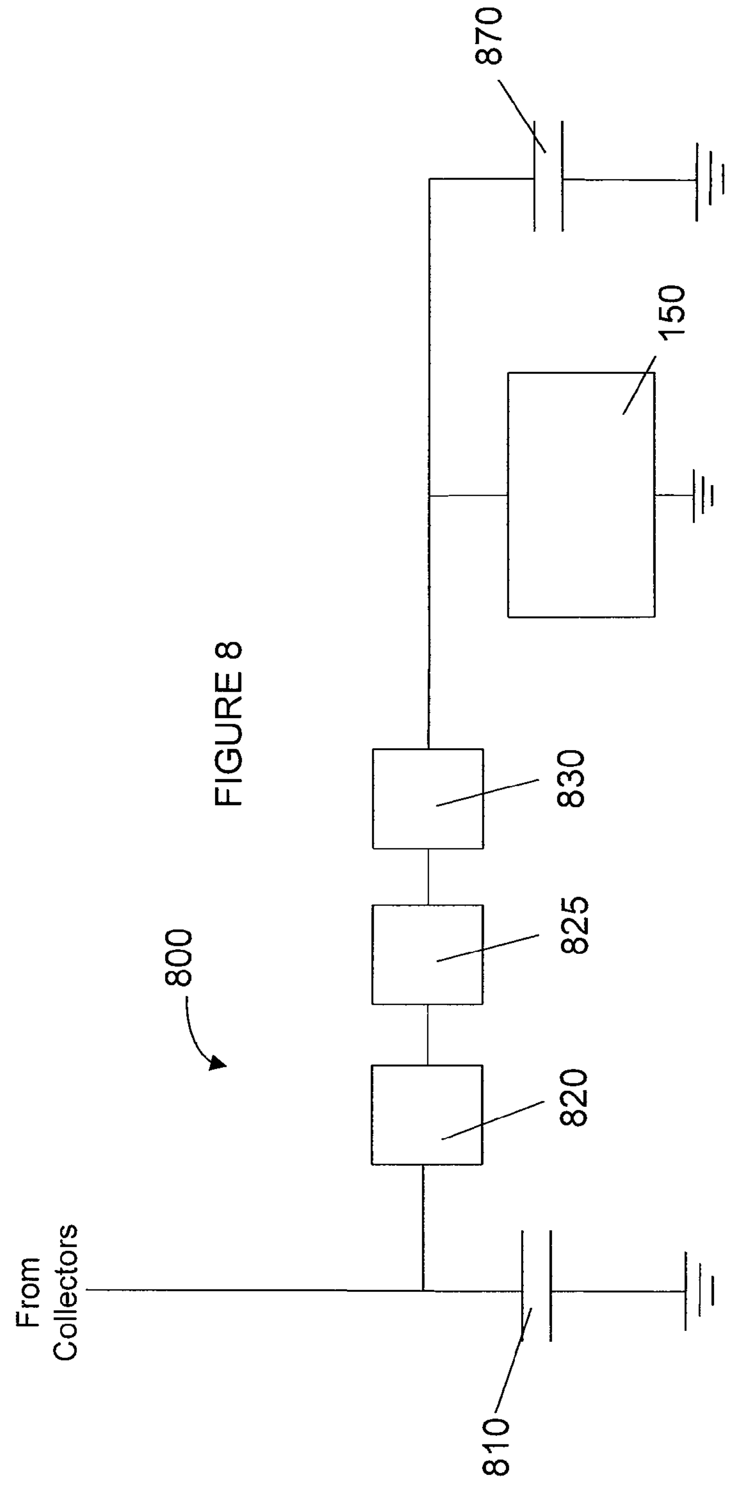
FIGURE 8

800
From Collectors
810
820
825
830
150
870

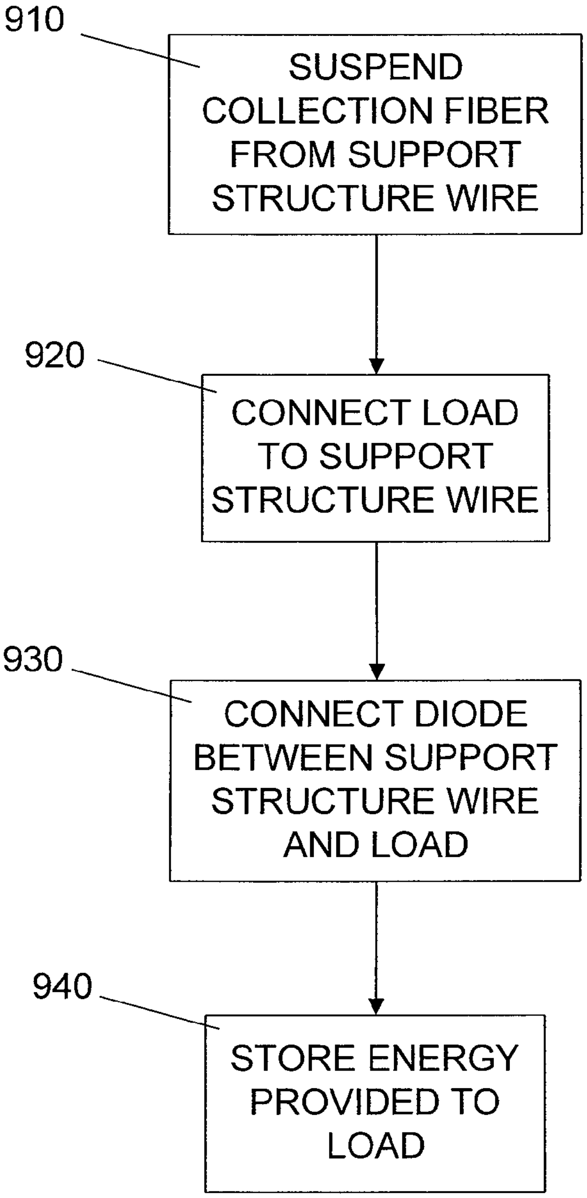
910
SUSPEND COLLECTION FIBER FROM SUPPORT STRUCTURE WIRE
920
CONNECT LOAD TO SUPPORT STRUCTURE WIRE
930
CONNECT DIODE BETWEEN SUPPORT STRUCTURE WIRE AND LOAD
940
STORE ENERGY PROVIDED TO LOAD

FIGURE 9

ENERGY COLLECTION

CROSS REFERENCE TO RELATED APPLICATION

This application is a continuation application of U.S. patent application Ser. No. 11/358,264, filed on Feb. 21, 2006, which is incorporated by reference herein.

TECHNICAL FIELD

The present disclosure is generally related to energy and, more particularly, is related to systems and methods for collecting energy.

BACKGROUND

The concept of fair weather electricity deals with the electric field and the electric current in the atmosphere propagated by the conductivity of the air. Clear, calm air carries an electrical current, which is the return path for thousands of lightening storms simultaneously occurring at any given moment around the earth. For simplicity, this energy may be referred to as static electricity or static energy. FIG. **1** illustrates a weather circuit for returning the current from lightning, for example, back to ground **10**. Weather currents **20**, **30** return the cloud to ground current **40**.

In a lightening storm, an electrical charge is built up, and electrons arc across a gas, ionizing it and producing the lightening flash. As one of ordinary skill in the art understands, the complete circuit requires a return path for the lightening flash. The atmosphere is the return path for the circuit. The electric field due to the atmospheric return path is relatively weak at any given point because the energy of thousands of electrical storms across the planet are diffused over the atmosphere of the entire Earth during both fair and stormy weather. Other contributing factors to electric current being present in the atmosphere may include cosmic rays penetrating and interacting with the earth's atmosphere, and also the migration of ions, as well as other effects yet to be fully studied.

Some of the ionization in the lower atmosphere is caused by airborne radioactive substances, primarily radon. In most places of the world, ions are formed at a rate of 5-10 pairs per cubic centimeter per second at sea level. With increasing altitude, cosmic radiation causes the ion production rate to increase. In areas with high radon exhalation from the soil (or building materials), the rate may be much higher.

Alpha-active materials are primarily responsible for the atmospheric ionization. Each alpha particle (for instance, from a decaying radon atom) will, over its range of some centimeters, create approximately 150,000-200,000 ion pairs.

While there is a large amount of usable energy available in the atmosphere, a method or apparatus for efficiently collecting that energy has not been forthcoming. Therefore, a heretofore unaddressed need exists in the industry to address the aforementioned deficiencies and inadequacies.

SUMMARY

Embodiments of the present disclosure provide systems and methods for collecting energy. Briefly described in architecture, one embodiment of the system, among others, can be implemented by a support structure wire elevated above a ground level, at least one collection fiber electrically connected to the support structure wire; a load electrically connected to the support structure wire; and a diode electrically connected between the load and at least one collection fiber.

Embodiments of the present disclosure can also be viewed as providing methods for collecting energy. In this regard, one embodiment of such a method, among others, can be broadly summarized by the following steps: suspending at least one collection fiber from a support structure wire elevated above ground level, the fiber electrically connected to the support structure wire; providing a load with an electrical connection to the support structure wire to draw current; and providing a diode electrically connected between the collection fiber and the load.

Other systems, methods, features, and advantages of the present disclosure will be or become apparent to one with skill in the art upon examination of the following drawings and detailed description. It is intended that all such additional systems, methods, features, and advantages be included within this description, be within the scope of the present disclosure, and be protected by the accompanying claims.

BRIEF DESCRIPTION OF THE DRAWINGS

Many aspects of the disclosure can be better understood with reference to the following drawings. The components in the drawings are not necessarily to scale, emphasis instead being placed upon clearly illustrating the principles of the present disclosure. Moreover, in the drawings, like reference numerals designate corresponding parts throughout the several views.

FIG. **1** is a circuit diagram of a weather energy circuit.

FIG. **2** is a perspective view of an exemplary embodiment of many energy collectors elevated above ground by a structure.

FIG. **2**A is a side view of an energy collection fiber suspended from a support wire.

FIG. **2**B is a side view of an exemplary embodiment of an energy collection fiber suspended from a support wire and with an additional support member.

FIG. **2**C is a perspective view of a support structure for multiple energy collection fibers.

FIG. **2**D is a side view of an exemplary embodiment of a support structure for multiple energy collection fibers.

FIG. **2**E is a side view of a support structure for an energy collection fiber.

FIG. **2**F is a side view of an exemplary embodiment of a support structure for an energy collection fiber.

FIG. **2**G is a side view of a support structure for multiple energy collection fibers.

FIG. **3** is a circuit diagram of an exemplary embodiment of a circuit for the collection of energy.

FIG. **4** is a circuit diagram of an exemplary embodiment of a circuit for the collection of energy.

FIG. **5** is a circuit diagram of an exemplary embodiment of an energy collection circuit for powering a generator and motor.

FIG. **6** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy and using it for the production of hydrogen and oxygen.

FIG. **7** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy, and using it for driving a fuel cell.

FIG. **8** is a circuit diagram of an exemplary embodiment of a circuit for collecting energy.

FIG. **9** is a flow diagram of an exemplary embodiment of collecting energy with a collection fiber.

DETAILED DESCRIPTION

Electric charges on conductors reside entirely on the external surface of the conductors, and tend to concentrate more around sharp points and edges than on flat surfaces. Therefore, an electric field received by a sharp conductive point may be much stronger than a field received by the same charge residing on a large smooth conductive shell. An exemplary embodiment of this disclosure takes advantage of this property, among others, to collect and use the energy generated by an electric field in the atmosphere. Referring to collection system **100** presented in FIG. **2**, at least one collection device **130** may be suspended from a support wire system **120** supported by poles **110**. Collection device **130** may comprise a diode or a collection fiber individually, or the combination of a diode and a collection fiber. Support wire system **120** may be electrically connected to load **150** by connecting wire **140**. Supporting wire system **120** may be any shape or pattern. Also, conducting wire **140** may be one wire or multiple wires. The collection device **130** in the form of a fiber may comprise any conducting or non-conducting material, including carbon, graphite, Teflon, and metal. An exemplary embodiment utilizes carbon or graphite fibers for static electricity collection. Support wire system **120** and connecting wire **140** can be made of any conducting material, including aluminum or steel, but most notably, copper. Teflon may be added to said conductor as well, such as non-limiting examples of a Teflon impregnated wire, a wire with a Teflon coating, or Teflon strips hanging from a wire. Conducting wire **120**, **140**, and **200** may be bare wire, or coated with insulation as a non-limiting example. Wires **120** and **140** are a means of transporting the energy collected by collection device **130**.

An exemplary embodiment of the collection fibers as collection device **130** includes graphite or carbon fibers. Graphite and carbon fibers, at a microscopic level, can have hundreds of thousands of points. Atmospheric electricity may be attracted to these points. If atmospheric electricity can follow two paths where one is a flat surface and the other is a pointy, conductive surface, the electrical charge will be attracted to the pointy, conductive surface. Generally, the more points that are present, the higher energy that can be gathered. Therefore, carbon, or graphite fibers are examples that demonstrate exemplary collection ability.

In at least one exemplary embodiment, the height of support wire **120** may be an important factor. The higher that collection device **130** is from ground, the larger the voltage potential between collection device **130** and electrical ground. The electric field may be more than 100 volts per meter under some conditions. When support wire **120** is suspended in the air at a particular altitude, wire **120** will itself collect a very small charge from ambient voltage. When collection device **130** is connected to support wire **120**, collection device **130** becomes energized and transfers the energy to support wire **120**.

A diode, not shown in FIG. **2**, may be connected in several positions in collection system **100**. A diode is a component that restricts the direction of movement of charge carriers. It allows an electric current to flow in one direction, but essentially blocks it in the opposite direction. A diode can be thought of as the electrical version of a check valve. The diode may be used to prevent the collected energy from discharging into the atmosphere through the collection fiber embodiment of collection device **130**. An exemplary embodiment of the collection device comprises the diode with no collection fiber. A preferred embodiment, however, includes a diode at the connection point of a collection fiber to support system **120** such that the diode is elevated above ground. Multiple diodes may be used between collection device **130** and load **150**. Additionally, in an embodiment with multiple fibers, the diodes restricts energy that may be collected through one fiber from escaping through another fiber.

Collection device **130** may be connected and arranged in relation to support wire system **120** by many means. Some non-limiting examples are provided in FIGS. **2A-2G** using a collection fiber embodiment. FIG. **2A** presents support wire **200** with connecting member **210** for collection device **130**. Connection member **210** may be any conducting material allowing for the flow of electricity from connection device **130** to support wire **200**. Then, as shown in FIG. **2**, the support wire **200** of support system **120** may be electrically connected through conducting wire **140** to load **150**. A plurality of diodes may be placed at any position on the support structure wire. A preferred embodiment places a diode at an elevated position at the connection point between a collection fiber embodiment of collection device **130** and connection member **210**.

Likewise, FIG. **2B** shows collection fiber **130** electrically connected to support wire **200** and also connected to support member **230**. Support member **230** may be connected to collection fiber **130** on either side. Support member **230** holds the fiber steady on both ends instead of letting it move freely. Support member **230** may be conducting or non-conducting. A plurality of diodes may be placed at any position on the support structure wire. A preferred embodiment places a diode at elevated position at the connection point between collection fiber **130** and support wire **200** or between fiber **130**, support member **230**, and support wire **200**.

FIG. **2C** presents multiple collection fibers in a squirrel cage arrangement with top and bottom support members. Support structure **250** may be connected to support structure wire **200** by support member **240**. Structure **250** has a top **260** and a bottom **270** and each of the multiple collection fibers **130** are connected on one end to top **260** and on the other end to bottom **270**. A plurality of diodes may be placed at any position on support structure **250**. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support structure wire **200**.

FIG. **2D** presents another exemplary embodiment of a support structure with support members **275** in an x-shape connected to support structure wire **200** at intersection **278** with collection fibers **130** connected between ends of support members **275**. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support wire **200**.

FIG. **2E** provides another exemplary embodiment for supporting collection fiber **130**. Collection fiber **130** may be connected on one side to support member **285**, which may be connected to support structure wire **200** in a first location and on the other side to support member **280**, which may be connected to support structure wire **200** in a second location on support structure wire **200**. The first and second locations may be the same location, or they may be different locations, even on different support wires. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places one or more diodes at elevated positions at the connection point(s) between collection fiber **130** and support wire **200**.

FIG. **2F** presents another exemplary embodiment of a support for a collection fiber. Two support members **290** may support either side of a collection fiber and are connected to support wire **200** in a single point. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places a diode at an elevated position at the connection point between collection fiber **130** and support wire **200**.

FIG. **2G** provides two supports as provided in FIG. **2F** such that at least two support members **292**, **294** may be connected to support structure wire **200** in multiple locations and collection fibers **130** may be connected between each end of the support structures. Collection fibers **130** may be connected between each end of a single support structure and between multiple support structures. A plurality of diodes may be placed at any position on the support structure. A preferred embodiment places one or more diodes at elevated positions at the connection point(s) between collection fiber **130** and support structure wire **200**.

FIG. **3** provides a schematic diagram of storing circuit **300** for storing energy collected by one or more collection devices (**130** from FIG. **2**). Load **150** induces current flow. Diode **310** may be electrically connected in series between one or more collection devices (**130** from FIG. **2**) and load **150**. A plurality of diodes may be placed at any position in the circuit. Switch **330** may be electrically connected between load **150** and at least one collection device (**130** from FIG. **2**) to connect and disconnect the load. Capacitor **320** may be connected in parallel to the switch **330** and load **150** to store energy when switch **330** is open for delivery to load **150** when switch **330** is closed. Rectifier **340** may be electrically connected in parallel to load **150**, between the receiving end of switch **330** and ground. Rectifier **340** may be a full-wave or a half-wave rectifier. Rectifier **340** may include a diode electrically connected in parallel to load **150**, between the receiving end of switch **330** and ground. The direction of the diode of rectifier **340** is optional.

In an exemplary embodiment provided in FIG. **4**, storage circuit **400** stores energy from one or more collection devices (**130** from FIG. **2**) by charging capacitor **410**. If charging capacitor **410** is not used, then the connection to ground shown at capacitor **410** is eliminated. A plurality of diodes may be placed at any position in the circuit. Diode **310** may be electrically connected in series between one or more collection devices (**130** from FIG. **2**) and load **150**. Diode **440** may be placed in series with load **150**. The voltage from capacitor **410** can be used to charge spark gap **420** when it reaches sufficient voltage. Spark gap **420** may comprise one or more spark gaps in parallel. Non-limiting examples of spark gap **420** include mercury-reed switches and mercury-wetted reed switches. When spark gap **420** arcs, energy will arc from one end of the spark gap **420** to the receiving end of the spark gap **420**. The output of spark gap **420** may be electrically connected in series to rectifier **450**. Rectifier **450** may be a full-wave or a half-wave rectifier. Rectifier **450** may include a diode electrically connected in parallel to transformer **430** and load **150**, between the receiving end of spark gap **420** and ground. The direction of the diode of rectifier **450** is optional. The output of rectifier **450** is connected to transformer **430** to drive load **150**.

FIG. **5** presents motor driver circuit **500**. One or more collection devices (**130** from FIG. **2**) are electrically connected to static electricity motor **510**, which powers generator **520** to drive load **150**. A plurality of diodes may be placed at any position in the circuit. Motor **510** may also be directly connected to load **150** to drive it directly.

FIG. **6** demonstrates a circuit **600** for producing hydrogen. A plurality of diodes may be placed at any position in the circuit. One or more collection devices (**130** from FIG. **2**) are electrically connected to primary spark gap **610**, which may be connected to secondary spark gap **640**. Non-limiting examples of spark gaps **610**, **640** include mercury-reed switches and mercury-wetted reed switches. Secondary spark gap **640** may be immersed in water **630** within container **620**. When secondary spark gap **640** immersed in water **630** is energized, spark gap **640** may produce bubbles of hydrogen and oxygen, which may be collected to be used as fuel.

FIG. **7** presents circuit **700** for driving a fuel cell. A plurality of diodes may be placed at any position in the circuit. Collection devices (**130** from FIG. **2**) provide energy to fuel cell **720** which drives load **150**. Fuel cell **720** may produce hydrogen and oxygen.

FIG. **8** presents exemplary circuit **800** for the collection of energy. Storage circuit **800** stores energy from one or more collection devices (**130** from FIG. **2**) by charging capacitor **810**. If charging capacitor **810** is not used, then the connection to ground shown at capacitor **810** is eliminated. A plurality of diodes may be placed at any position in the circuit. The voltage from capacitor **810** can be used to charge spark gap **820** when it reaches sufficient voltage. Spark gap **820** may comprise one or more spark gaps in parallel or in series. Non-limiting examples of spark gap **820** include mercury-reed switches and mercury-wetted reed switches. When spark gap **820** arcs, energy will arc from one end of spark gap **820** to the receiving end of spark gap **820**. The output of spark gap **820** may be electrically connected in series to rectifier **825**. Rectifier **825** may be a full-wave or a half-wave rectifier. Rectifier **825** may include a diode electrically connected in parallel to inductor **830** and load **150**, between the receiving end of spark gap **820** and ground. The direction of the diode of rectifier **825** is optional. The output of rectifier **825** is connected to inductor **830**. Inductor **830** may be a fixed value inductor or a variable inductor. Capacitor **870** may be placed in parallel with load **150**.

FIG. **9** presents a flow diagram of a method for collecting energy. In block **910**, one or more collection devices may be suspended from a support structure wire. In block **920**, a load may be electrically connected to the support structure wire to draw current. In block **930** a diode may be electrically connected between the support structure wire and the electrical connection to the load. In block **940**, energy provided to the load may be stored or otherwise utilized.

Any process descriptions or blocks in flow charts should be understood as representing modules, segments, or portions of code which include one or more executable instructions for implementing specific logical functions or steps in the process, and alternate implementations are included within the scope of the preferred embodiment of the present disclosure in which functions may be executed out of order from that shown or discussed, including substantially concurrently or in reverse order, depending on the functionality involved, as would be understood by those reasonably skilled in the art of the present disclosure.

It should be emphasized that the above-described embodiments of the present disclosure, particularly, any "preferred" embodiments, are merely possible examples of implementations, merely set forth for a clear understanding of the principles of the disclosure. Many variations and modifications may be made to the above-described embodiment(s) of the disclosure without departing substantially from the spirit and principles of the disclosure. All such modifications and variations are intended to be included herein within the scope of this disclosure and the present disclosure and protected by the following claims.

Therefore, at least the following is claimed:

1. A method of collecting energy comprising:

suspending at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the at least one collection device electrically connected to the support structure;

providing a load with an electrical connection to the at least one collection device to passively draw current; and powering a fuel cell between the support structure and the load, the powering performed with energy collected from the at least one collection device.

2. The method of claim **1**, wherein the collection device comprises a diode.

3. The method of claim **1**, wherein the collection device comprises a collection fiber.

4. The method of claim **1**, wherein the collection device comprises a diode and a collection fiber and the diode is electrically connected between the collection fiber and the load.

5. The method of claim **1**, further comprising storing energy provided to the load.

6. The method of claim **5**, wherein storing energy provided to the load comprises storing energy in a capacitor or an inductor.

7. The method of claim **3**, wherein the collection fiber comprises carbon fiber or graphite fiber.

8. A system of energy collection comprising:

a support structure;

at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the collection device electrically connected to and suspended from the support structure, the at least one collection device configured to passively draw current;

a load electrically connected to the at least one collection device; and

a diode electrically connected between the at least one collection device and the support structure.

9. The system of claim **8**, wherein the collection device comprises a diode.

10. The system of claim **8**, wherein the collection device comprises a collection fiber.

11. The system of claim **8**, wherein the collection device comprises a collection fiber and a diode electrically connected between the load and the collection fiber.

12. The system of claim **9**, wherein the diode is elevated relative to the ground level.

13. The system of claim **10**, wherein the collection fiber comprises a carbon fiber or a graphite fiber.

14. The system of claim **8**, further comprising:

a switch connected in series between the at least one collection device and the load; and

a capacitor connected in parallel with the switch and the load.

15. The system of claim **14**, wherein the switch comprises an interrupter connected between the at least one collection device and the load, and wherein the interrupter comprises at least one of a fluorescent tube, a neon bulb, an AC light, and a spark gap.

16. The system of claim **15**, further comprising a transformer connected between the interrupter and the load.

17. The system of claim **8**, further comprising:

a motor for providing power, the motor connected between the at least one collection device and the load; and

a generator powered by the motor.

18. The system of claim **8**, further comprising a fuel cell between the support structure and the load.

19. The system of claim **18**, wherein the fuel cell produces hydrogen and oxygen.

20. A system of collecting energy comprising:

means for suspending at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the at least one collection device electrically connected to the means for suspending;

means for passively inducing current flow, the means for inducing current flow electrically connected to the means for suspending; and

means for restricting the backflow of charge carriers, the means for restricting the backflow of charge carriers electrically connected between the at least one collection device and the means for inducing current flow.

21. A system of energy collection comprising:

a support structure;

at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the collection device electrically connected to and suspended from the support structure, the at least one collection device configured to passively draw current;

a load electrically connected to the at least one collection device; and

a switch connected in series between the at least one collection device and the load, the switch comprising an interrupter connected between the at least one collection device and the load, and wherein the interrupter comprises at least one of a fluorescent tube, a neon bulb, an AC light, and a spark gap.

22. The system of claim **21**, further comprising a transformer connected between the interrupter and the load.

23. A system of energy collection comprising:

a support structure;

at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the at least one collection device electrically connected to and suspended from the support structure, the collection device configured to passively draw current;

a load electrically connected to the at least one collection device;

a motor for providing power, the motor connected between the at least one collection device and the load; and

a generator powered by the motor.

24. A system of energy collection comprising:

a support structure;

at least one collection device with, in operation, microscopic points of a cross-section of the collection device exposed to the environment, the at least one collection device electrically connected to and suspended from the support structure the at least one collection device configured to passively draw current;

a load electrically connected to the at least one collection device; and

a fuel cell between the support structure and the load.

25. The system of claim **24**, wherein the fuel cell produces hydrogen and oxygen.

* * * * *

United States Patent [19]
Eastlund

[11] **Patent Number: 4,686,605**
[45] **Date of Patent: Aug. 11, 1987**

[54] METHOD AND APPARATUS FOR ALTERING A REGION IN THE EARTH'S ATMOSPHERE, IONOSPHERE, AND/OR MAGNETOSPHERE

[75] Inventor: **Bernard J. Eastlund**, Spring, Tex.
[73] Assignee: **APTI, Inc.**, Los Angeles, Calif.
[21] Appl. No.: **690,333**
[22] Filed: **Jan. 10, 1985**
[51] **Int. Cl.⁴** **H05B 6/64**; H05C 3/00; H05H 1/46
[52] **U.S. Cl.** .... **361/231**; 89/1.11; 380/59; 244/158 R
[58] **Field of Search** .... 361/230, 231; 244/158 R; 376/100; 89/1.11; 380/59

[56] **References Cited**

PUBLICATIONS

Liberty Magazine, (2/35) p. 7 N. Tesla.
New York Times (9/22/40) Section 2, p. 7 W. L. Laurence.
New York Times (12/8/15) p. 8 Col. 3.

Primary Examiner—Salvatore Cangialosi
Attorney, Agent, or Firm—Roderick W. MacDonald

[57] ABSTRACT

A method and apparatus for altering at least one selected region which normally exists above the earth's surface. The region is excited by electron cyclotron resonance heating to thereby increase its charged particle density. In one embodiment, circularly polarized electromagnetic radiation is transmitted upward in a direction substantially parallel to and along a field line which extends through the region of plasma to be altered. The radiation is transmitted at a frequency which excites electron cyclotron resonance to heat and accelerate the charged particles. This increase in energy can cause ionization of neutral particles which are then absorbed as part of the region thereby increasing the charged particle density of the region.

15 Claims, 5 Drawing Figures

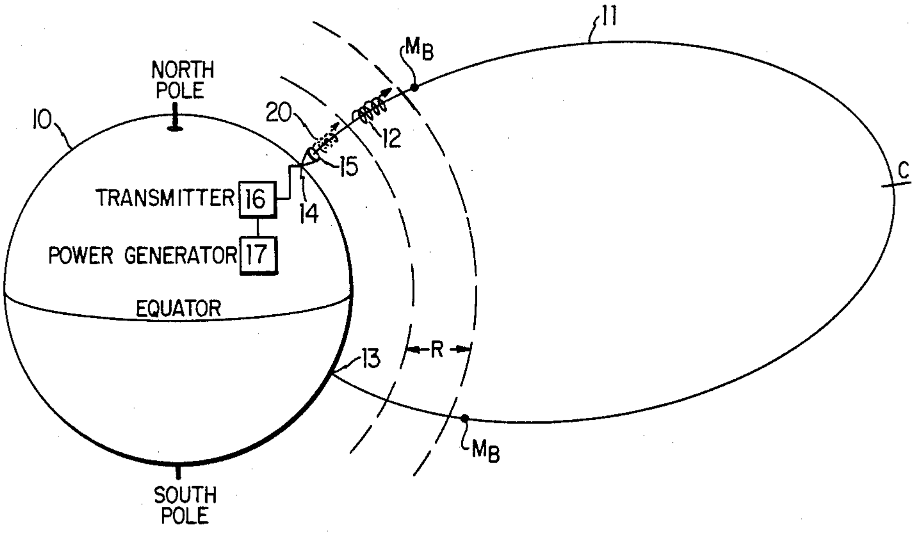

FIG. 1

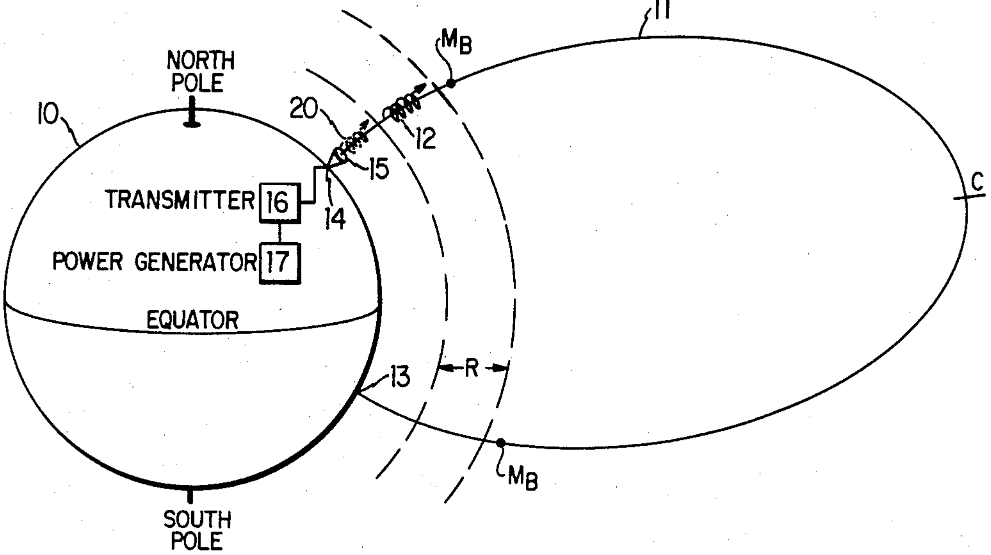
11
MB
NORTH POLE
10
20
12
15
C
TRANSMITTER 16
14
POWER GENERATOR 17
EQUATOR
13
R
MB
SOUTH POLE

FIG. 2

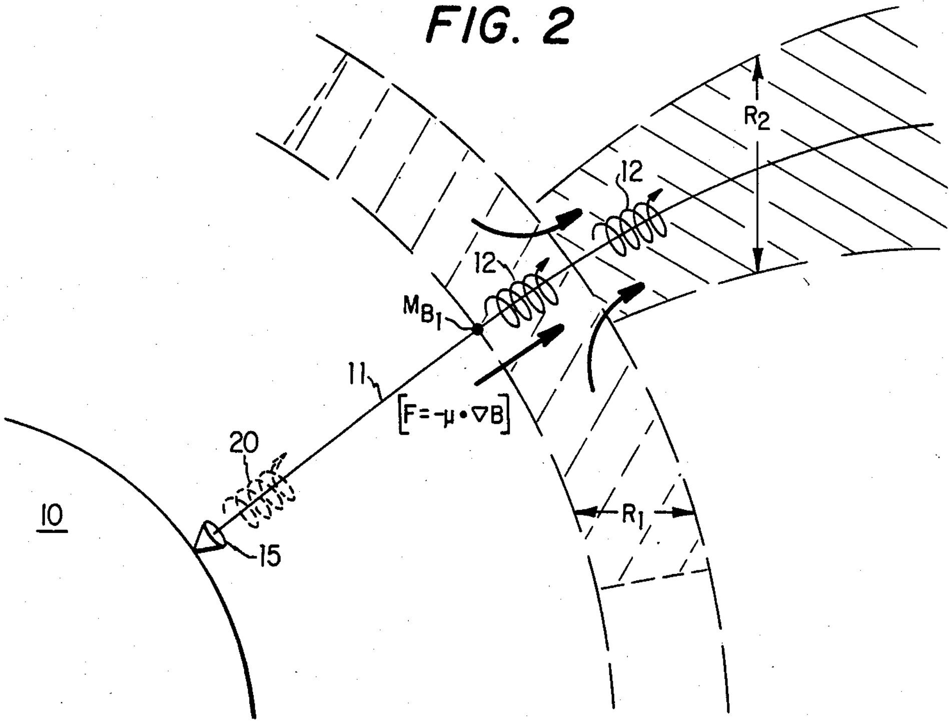
R2
MB1
[F=-μ•∇B]
R1

FIG. 3

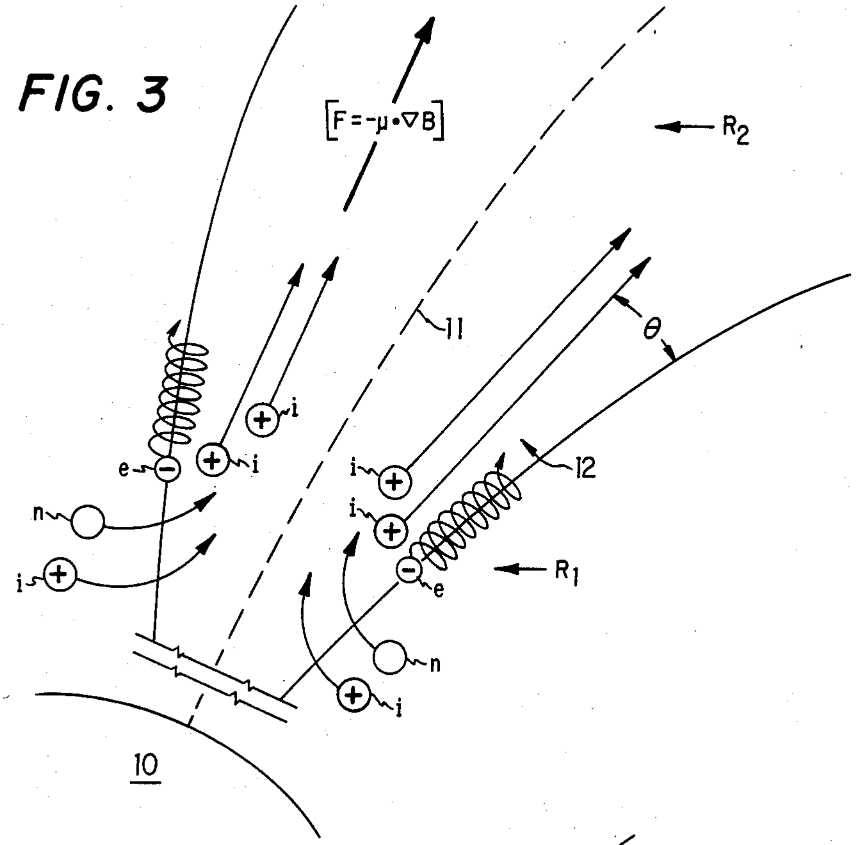
[F=-μ•∇B]
R2
11
θ
e
i
n
i
i
i
12
R1
e
n
i
10

FIG. 4

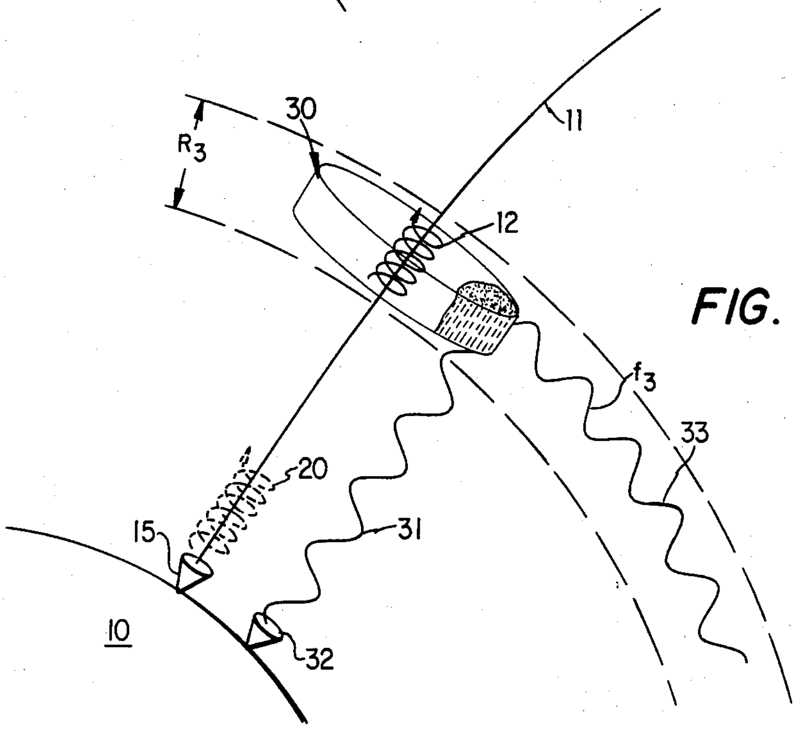
R3
30
11
12
f3
33
20
15
31
10
32

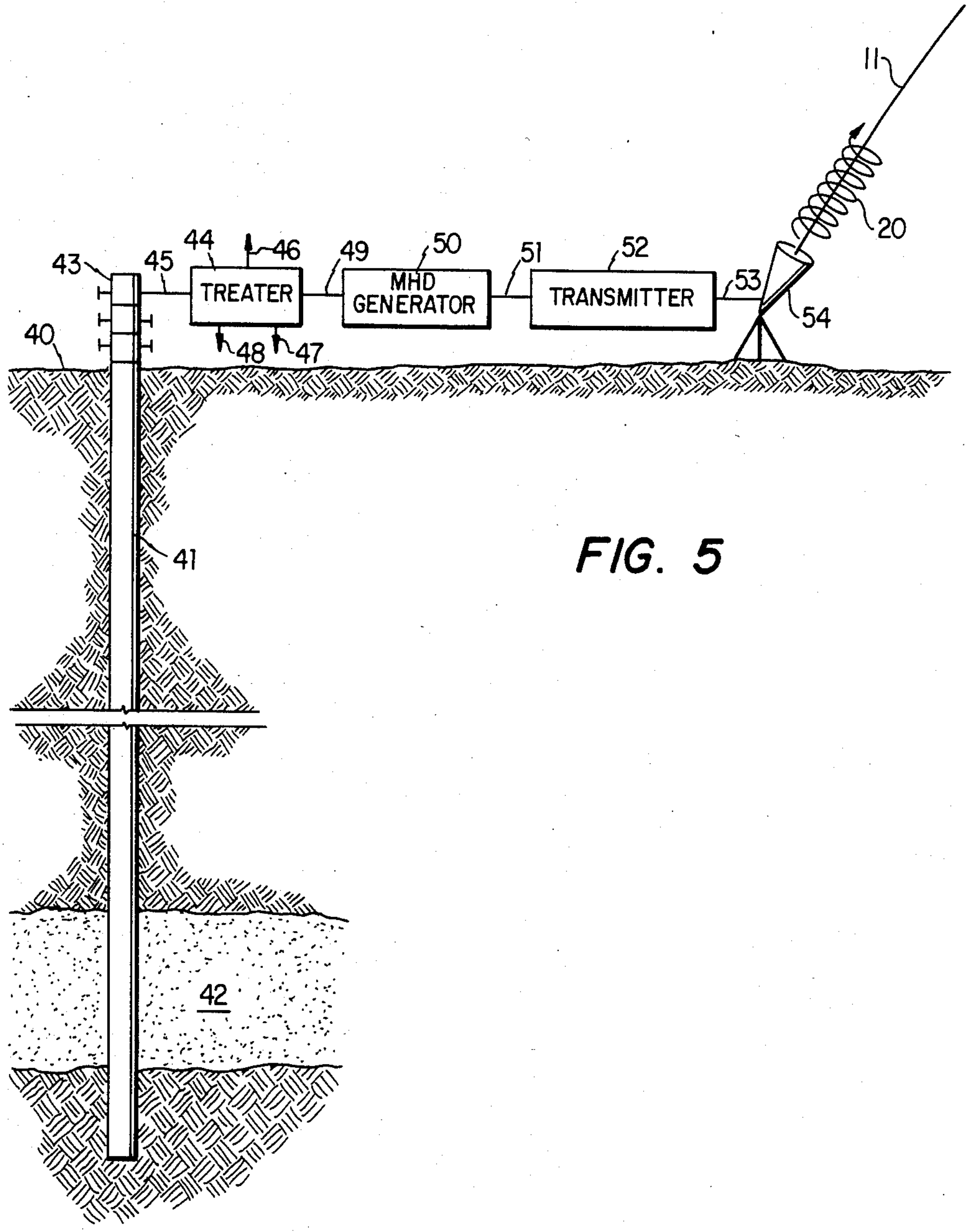
TREATER
MHD GENERATOR
TRANSMITTER

FIG. 5

METHOD AND APPARATUS FOR ALTERING A REGION IN THE EARTH'S ATMOSPHERE, IONOSPHERE, AND/OR MAGNETOSPHERE

DESCRIPTION

1. Technical Field

This invention relates to a method and apparatus for altering at least one selected region normally existing above the earth's surface and more particularly relates to a method and apparatus for altering said at least one region by initially transmitting electromagnetic radiation from the earth's surface essentially parallel to and along naturally-occurring, divergent magnetic field lines which extend from the earth's surface through the region or regions to be altered.

2. Background Art

In the late 1950's, it was discovered that naturally-occuring belts exist at high altitudes above the earth's surface, and it is now established that these belts result from charged electrons and ions becoming trapped along the magnetic lines of force (field lines) of the earth's essentially dipole magnetic field. The trapped electrons and ions are confined along the field lines between two magnetic mirrors which exist at spaced apart points along those field lines. The trapped electrons and ions move in helical paths around their particular field lines and "bounce" back and forth between the magnetic mirrors. These trapped electrons and ions can oscillate along the field lines for long periods of time.

In the past several years, substantial effort has been made to understand and explain the phenomena involved in belts of trapped electrons and ions, and to explore possible ways to control and use these phenomena for beneficial purposes. For example, in the late 1950's and early 1960's both the United States and U.S.S.R. detonated a series of nuclear devices of various yields to generate large numbers of charged particles at various altitudes, e.g., 200 kilometers (km) or greater. This was done in order to establish and study artifical belts of trapped electrons and ions. These experiments established that at least some of the extraneous electrons and ions from the detonated devices did become trapped along field lines in the earth's magnetosphere to form artificial belts which were stable for prolonged periods of time. For a discussion of these experiments see "The Radiation Belt and Magnetosphere", W. N. Hess, Blaisdell Publishing Co., 1968, pps. 155 et sec.

Other proposals which have been advanced for altering existing belts of trapped electrons and ions and/or establishing similar artificial belts include injecting charged particles from a satellite carrying a payload of radioactive beta-decay material or alpha emitters; and injecting charged particles from a satellite-borne electron accelerator. Still another approach is described in U.S. Pat. No. 4,042,196 wherein a low energy ionized gas, e.g., hydrogen, is released from a synchronous orbiting satellite near the apex of a radiation belt which is naturally-occurring in the earth's magnetosphere to produce a substantial increase in energetic particle precipitation and, under certain conditions, produce a limit in the number of particles that can be stably trapped. This precipitation effect arises from an enhancement of the whistler-mode and ion-cyclotron mode interactions that result from the ionized gas or "cold plasma" injection.

It has also been proposed to release large clouds of barium in the magnetosphere so that photoionization will increase the cold plasma density, thereby producing electron precipitation through enhanced whistler-mode interactions.

However, in all of the above-mentioned approaches, the mechanisms involved in triggering the change in the trapped particle phenomena must be actually positioned within the affected zone, e.g., the magnetosphere, before they can be actuated to effect the desired change.

The earth's ionosphere is not considered to be a "trapped" belt since there are few trapped particles therein. The term "trapped" herein refers to situations where the force of gravity on the trapped particles is balanced by magnetic forces rather than hydrostatic or collisional forces. The charged electrons and ions in the ionosphere also follow helical paths around magnetic field lines within the ionosphere but are not trapped between mirrors, as in the case of the trapped belts in the magnetosphere, since the gravitational force on the particles is balanced by collisional or hydrostatic forces.

In recent years, a number of experiments have actually been carried out to modify the ionosphere in some controlled manner to investigate the possibility of a beneficial result. For detailed discussions of these operations see the following papers: (1) Ionospheric Modification Theory; G. Meltz and F. W. Perkins; (2) The Platteville High Power Facility; Carrol et al.; (3) Arecibo Heating Experiments; W. E. Gordon and H. C. Carlson, Jr.; and (4) Ionospheric Heating by Powerful Radio Waves; Meltz et al., all published in Radio Science, Vol. 9, No. 11, November, 1974, at pages 885–888; 889–894; 1041–1047; and 1049–1063, respectively, all of which are incorporated herein by reference. In such experiments, certain regions of the ionosphere are heated to change the electron density and temperature within these regions. This is accomplished by transmitting from earth-based antennae high frequency electromagnetic radiation at a substantial angle to, not parallel to, the ionosphere's magnetic field to heat the ionospheric particles primarily by ohmic heating. The electron temperature of the ionosphere has been raised by hundreds of degrees in these experiments, and electrons with several electron volts of energy have been produced in numbers sufficient to enhance airglow. Electron concentrations have been reduced by a few percent, due to expansion of the plasma as a result of increased temperature.

In the Elmo Bumpy Torus (EBT), a controlled fusion device at the Oak Ridge National Laboratory, all heating is provided by microwaves at the electron cyclotron resonance interaction. A ring of hot electrons is formed at the earth's surface in the magnetic mirror by a combination of electron cyclotron resonance and stochastic heating. In the EBT, the ring electrons are produced with an average "temperature" of 250 kilo electron volts or kev (2.5×10^{9}K) and a plasma beta between 0.1 and 0.4; see, "A Theoretical Study of Electron—Cyclotron Absorption in Elmo Bumpy Torus", Batchelor and Goldfinger, Nuclear Fusion, Vol. 20, No. 4 (1980) pps. 403–418.

Electron cyclotron resonance heating has been used in experiments on the earth's surface to produce and accelerate plasmas in a diverging magnetic field. Kosmahl et al. showed that power was transferred from the electromagnetic waves and that a fully ionized plasma was accelerated with a divergence angle of roughly 13 degrees. Optimum neutral gas density was 1.7×10^{14} per cubic centimeter; see, "Plasma Acceleration with Microwaves Near Cyclotron Resonance", Kosmahl et al., Journal of Applied Physics, Vol. 38, No. 12, Nov., 1967, pps. 4576–4582.

DISCLOSURE OF THE INVENTION

The present invention provides a method and apparatus for altering at least one selected region which normally exists above the earth's surface. The region is excited by electron cyclotron resonance heating of electrons which are already present and/or artifically created in the region to thereby increase the charged particle energy and ultimately the density of the region.

In one embodiment this is done by transmitting circularly polarized electromagnetic radiation from the earth's surface at or near the location where a naturally-occurring dipole magnetic field (force) line intersects the earth's surface. Right hand circular polarization is used in the northern hemisphere and left hand circular polarization is used in the southern hemisphere. The radiation is deliberately transmitted at the outset in a direction substantially parallel to and along a field line which extends upwardly through the region to be altered. The radiation is transmitted at a frequency which is based on the gyrofrequency of the charged particles and which, when applied to the at least one region, excites electron cyclotron resonance within the region or regions to heat and accelerate the charged particles in their respective helical paths around and along the field line. Sufficient energy is employed to cause ionization of neutral particles (molecules of oxygen, nitrogen and the like, particulates, etc.) which then become a part of the region thereby increasing the charged particle density of the region. This effect can further be enhanced by providing artificial particles, e.g., electrons, ions, etc., directly into the region to be affected from a rocket, satellite, or the like to supplement the particles in the naturally-occurring plasma. These artificial particles are also ionized by the transmitted electromagnetic radiation thereby increasing charged particle density of the resulting plasma in the region.

In another embodiment of the invention, electron cyclotron resonance heating is carried out in the selected region or regions at sufficient power levels to allow a plasma present in the region to generate a mirror force which forces the charged electrons of the altered plasma upward along the force line to an altitude which is higher than the original altitude. In this case the relevant mirror points are at the base of the altered region or regions. The charged electrons drag ions with them as well as other particles that may be present. Sufficient power, e.g., 10^{15} joules, can be applied so that the altered plasma can be trapped on the field line between mirror points and will oscillate in space for prolonged periods of time. By this embodiment, a plume of altered plasma can be established at selected locations for communication modification or other purposes.

In another embodiment, this invention is used to alter at least one selected region of plasma in the ionosphere to establish a defined layer of plasma having an increased charged particle density. Once this layer is established, and while maintaining the transmission of the main beam of circularly polarized electromagnetic radiation, the main beam is modulated and/or at least one second different, modulated electromagnetic radiation beam is transmitted from at least one separate source at a different frequency which will be absorbed in the plasma layer. The amplitude of the frequency of the main beam and/or the second beam or beams is modulated in resonance with at least one known oscillation mode in the selected region or regions to excite the known oscillation mode to propagate a known frequency wave or waves throughout the ionosphere.

BRIEF DESCRIPTION OF THE DRAWINGS

The actual construction, operation, and apparent advantages of this invention will be better understood by referring to the drawings in which like numerals identify like parts and in which:

FIG. **1** is a simplified schematical view of the earth (not to scale) with a magnetic field (force) line along which the present invention is carried out;

FIG. **2** is one embodiment within the present invention in which a selected region of plasma is raised to a higher altitude;

FIG. **3** is a simplified, idealized representation of a physical phenomenon involved in the present invention; and

FIG. **4** is a schematic view of another embodiment within the present invention.

FIG. **5** is a schematic view of an apparatus embodiment within this invention.

BEST MODES FOR CARRYING OUT THE INVENTION

The earth's magnetic field is somewhat analogous to a dipole bar magnet. As such, the earth's magnetic field contains numerous divergent field or force lines, each line intersecting the earth's surface at points on opposite sides of the Equator. The field lines which intersect the earth's surface near the poles have apexes which lie at the furthest points in the earth's magnetosphere while those closest to the Equator have apexes which reach only the lower portion of the magnetosphere.

At various altitudes above the earth's surface, e.g., in both the ionosphere and the magnetosphere, plasma is naturally present along these field lines. This plasma consists of equal numbers of positively and negatively charged particles (i.e., electrons and ions) which are guided by the field line. It is well established that a charged particle in a magnetic field gyrates about field lines, the center of gyration at any instance being called the "guiding center" of the particle. As the gyrating particle moves along a field line in a uniform field, it will follow a helical path about its guiding center, hence linear motion, and will remain on the field line. Electrons and ions both follow helical paths around a field line but rotate in opposite directions. The frequencies at which the electrons and ions rotate about the field line are called gyromagnetic frequencies or cyclotron frequencies because they are identical with the expression for the angular frequencies of gyration of particles in a cyclotron. The cyclotron frequency of ions in a given magnetic field is less than that of electrons, in inverse proportion to their masses.

If the particles which form the plasma along the earth's field lines continued to move with a constant pitch angle, often designated "alpha", they would soon impact on the earth's surface. Pitch angle alpha is defined as the angle between the direction of the earth's magnetic field and the velocity (V) of the particle. However, in converging force fields, the pitch angle does change in such a way as to allow the particle to turn around and avoid impact. Consider a particle moving along a field line down toward the earth. It moves into a region of increasing magnetic field strength and therefore sine alpha increases. But sine alpha can only increase to 1.0, at which point, the particle turns around and starts moving up along the field line, and alpha decreases. The point at which the particle turns around is called the mirror point, and there alpha equals ninety degrees. This process is repeated at the other end of the field line where the same magnetic field strength value B, namely Bm, exists. The particle again turns around and this is called the "conjugate point" of the original mirror point. The particle is therefore trapped and bounces between the two magnetic mirrors. The particle can continue oscillating in space in this manner for long periods of time. The actual place where a particle will mirror can be calculated from the following:

$$\sin^2 \text{alpha}_o = B_o/B_m \quad (1)$$

wherein:

alpha_o = equatorial pitch angle of particle

B_o = equatorial field strength on a particular field line

B_m = field strength at the mirror point

Recent discoveries have established that there are substantial regions of naturally trapped particles in space which are commonly called "trapped radiation belts". These belts occur at altitudes greater than about 500 km and accordingly lie in the magnetosphere and mostly above the ionosphere.

The ionosphere, while it may overlap some of the trapped-particle belts, is a region in which hydrostatic forces govern its particle distribution in the gravitational field. Particle motion within the ionosphere is governed by both hydrodynamic and electrodynamic forces. While there are few trapped particles in the ionosphere, nevertheless, plasma is present along field lines in the ionosphere. The charged particles which form this plasma move between collisions with other particles along similar helical paths around the field lines and although a particular particle may diffuse downward into the earth's lower atmosphere or lose energy and diverge from its original field line due to collisions with other particles, these charged particles are normally replaced by other available charged particles or by particles that are ionized by collision with said particle. The electron density (N_e) of the plasma will vary with the actual conditions and locations involved. Also, neutral particles, ions, and electrons are present in proximity to the field lines.

The production of enhanced ionization will also alter the distribution of atomic and molecular constituents of the atmosphere, most notably through increased atomic nitrogen concentration. The upper atmosphere is normally rich in atomic oxygen (the dominant atmospheric constituent above 200 km altitude), but atomic nitrogen is normally relatively rare. This can be expected to manifest itself in increased airglow, among other effects.

As known in plasma physics, the characteristics of a plasma can be altered by adding energy to the charged particles or by ionizing or exciting additional particles to increase the density of the plasma. One way to do this is by heating the plasma which can be accomplished in different ways, e.g., ohmic, magnetic compression, shock waves, magnetic pumping, electron cyclotron resonance, and the like.

Since electron cyclotron resonance heating is involved in the present invention, a brief discussion of same is in order. Increasing the energy of electrons in a plasma by invoking electron cyclotron resonance heating, is based on a principle similar to that utilized to accelerate charged particles in a cyclotron. If a plasma is confined by a static axial magnetic field of strength B, the charged particles will gyrate about the lines of force with a frequency given, in hertz, as $f_g = 1.54 \times 10^3 B/A$, where: B = magnetic field strength in gauss, and A = mass number of the ion.

Suppose a time-varying field of this frequency is superimposed on the static field B confining the plasma, by passage of a radiofrequency current through a coil which is concentric with that producing the axial field, then in each half-cycle of their rotation about the field lines, the charged particles acquire energy from the oscillating electric field associated with the radio frequency. For example, if B is 10,000 gauss, the frequency of the field which is in resonance with protons in a plasma is 15.4 megahertz.

As applied to electrons, electron cyclotron resonance heating requires an oscillating field having a definite frequency determined by the strength of the confining field. The radio-frequency radiation produces time-varying fields (electric and magnetic), and the electric field accelerates the charged particle. The energized electrons share their energy with ions and neutrals by undergoing collisions with these particles, thereby effectively raising the temperature of the electrons, ions, and neutrals. The apportionment of energy among these species is determined by collision frequencies. For a more detailed understanding of the physics involved, see "Controlled Thermonuclear Reactions", Glasstone and Lovberg, D. Van Nostrand Company, Inc., Princeton, N.J., 1960 and "The Radiation Belt and Magnetosphere", Hess, Blaisdell Publishing Company, 1968, both of which are incorporated herein by reference.

Referring now to the drawings, the present invention provides a method and apparatus for altering at least one region of plasma which lies along a field line, particularly when it passes through the ionosphere and/or magnetosphere. FIG. **1** is a simplified illustration of the earth **10** and one of its dipole magnetic force or field lines **11**. As will be understood, line **11** may be any one of the numerous naturally existing field lines and the actual geographical locations **13** and **14** of line **11** will be chosen based on a particular operation to be carried out. The actual locations at which field lines intersect the earth's surface is documented and is readily ascertainable by those skilled in the art.

Line **11** passes through region R which lies at an altitude above the earth's surface. A wide range of altitudes are useful given the power that can be employed by the practice of this invention. The electron cyclotron resonance heating effect can be made to act on electrons anywhere above the surface of the earth. These electrons may be already present in the atmosphere, ionosphere, and/or magnetosphere of the earth, or can be artificially generated by a variety of means such as x-ray beams, charged particle beams, lasers, the plasma sheath surrounding an object such as a missile or meteor, and the like. Further, artificial particles, e.g., electrons, ions, etc., can be injected directly into region R from an earth-launched rocket or orbiting satellite carrying, for example, a payload of radioactive beta-decay material; alpha emitters; an electron accelerator; and/or ionized gases such as hydrogen; see U.S. Pat. No. 4,042,196. The altitude can be greater than about 50 km if desired, e.g., can be from about 50 km to about 800 km, and, accordingly may lie in either the ionosphere or the magnetosphere or both. As explained above, plasma will be present along line **11** within region R and is represented by the helical line **12**. Plasma **12** is comprised of charged particles (i.e., electrons and ions) which rotate about opposing helical paths along line **11**.

Antenna **15** is positioned as close as is practical to the location **14** where line **11** intersects the earth's surface. Antenna **15** may be of any known construction for high directionality, for example, a phased array, beam spread angle (θ) type. See "The MST Radar at Poker Flat, Alaska", Radio Science, Vol. 15, No. 2, Mar.-Apr. 1980, pps. 213–223, which is incorporated herein by reference. Antenna **15** is coupled to transmitter **16** which generates a beam of high frequency electromagnetic radiation at a wide range of discrete frequencies, e.g., from about 20 to about 1800 kilohertz (kHz).

Transmitter **16** is powered by power generator means **17** which is preferably comprised of one or more large, commercial electrical generators. Some embodiments of the present invention require large amounts of power, e.g., up to 10^9 to 10^{11} watts, in continuous wave or pulsed power. Generation of the needed power is within the state of the art. Although the electrical generators necessary for the practice of the invention can be powered in any known manner, for example, by nuclear reactors, hydroelectric facilities, hydrocarbon fuels, and the like, this invention, because of its very large power requirement in certain applications, is particularly adapted for use with certain types of fuel sources which naturally occur at strategic geographical locations around the earth. For example, large reserves of hydrocarbons (oil and natural gas) exist in Alaska and Canada. In northern Alaska, particularly the North Slope region, large reserves are currently readily available. Alaska and northern Canada also are ideally located geographically as to magnetic latitudes. Alaska provides easy access to magnetic field lines that are especially suited to the practice of this invention, since many field lines which extend to desirable altitudes for this invention intersect the earth in Alaska. Thus, in Alaska, there is a unique combination of large, accessible fuel sources at desirable field line intersections. Further, a particularly desirable fuel source for the generation of very large amounts of electricity is present in Alaska in abundance, this source being natural gas. The presence of very large amounts of clean-burning natural gas in Alaskan latitudes, particularly on the North Slope, and the availability of magnetohydrodynamic (MHD), gas turbine, fuel cell, electrogasdynamic (EGD) electric generators which operate very efficiently with natural gas provide an ideal power source for the unprecedented power requirements of certain of the applications of this invention. For a more detailed discussion of the various means for generating electricity from hydrocarbon fuels, see "Electrical Aspects of Combustion", Lawton and Weinberg, Clarendon Press, 1969. For example, it is possible to generate the electricity directly at the high frequency needed to drive the antenna system. To do this, typically the velocity of flow of the combustion gases (v), past magnetic field perturbation of dimension d (in the case of MHD), follow the rule:

$$v = df$$

where f is the frequency at which electricity is generated. Thus, if $v = 1.78 \times 10^6$ cm/sec and $d = 1$ cm then electricity would be generated at a frequency of 1.78 mHz.

Put another way, in Alaska, the right type of fuel (natural gas) is naturally present in large amounts and at just the right magnetic latitudes for the most efficient practice of this invention, a truly unique combination of circumstances. Desirable magnetic latitudes for the practice of this invention interest the earth's surface both northerly and southerly of the equator, particularly desirable latitudes being those, both northerly and southerly, which correspond in magnitude with the magnetic latitudes that encompass Alaska.

Referring now to FIG. **2** a first ambodiment is illustrated where a selected region R_1 of plasma **12** is altered by electron cyclotron resonance heating to accelerate the electrons of plasma **12**, which are following helical paths along field line **11**.

To accomplish this result, electromagnetic radiation is transmitted at the outset, essentially parallel to line **11** via antenna **15** as right hand circularly polarized radiation wave **20**. Wave **20** has a frequency which will excite electron cyclotron resonance with plasma **12** at its initial or original altitude. This frequency will vary depending on the electron cyclotron resonance of region R_1 which, in turn, can be determined from available data based on the altitudes of region R_1, the particular field line **11** being used, the strength of the earth's magnetic field, etc. Frequencies of from about 20 to about 7200 kHz, preferably from about 20 to about 1800 kHz can be employed. Also, for any given application, there will be a threshhold (minimum power level) which is needed to produce the desired result. The minimum power level is a function of the level of plasma production and movement required, taking into consideration any loss processes that may be dominant in a particular plasma or propagation path.

As electron cyclotron resonance is established in plasma **12**, energy is transferred from the electromagnetic radiation **20** into plasma **12** to heat and accelerate the electrons therein and, subsequently, ions and neutral particles. As this process continues, neutral particles which are present within R_1 are ionized and absorbed into plasma **12** and this increases the electron and ion densities of plasma **12**. As the electron energy is raised to values of about 1 kilo electron volt (kev), the generated mirror force (explained below) will direct the excited plasma **12** upward along line **11** to form a plume R_2 at an altitude higher than that of R_1.

Plasma acceleration results from the force on an electron produced by a nonuniform static magnetic field $(\overline{B})$. The force, called the mirror force, is given by

$$F = -\mu \nabla B \tag{2}$$

where μ is the electron magnetic moment and $\nabla \overline{B}$ is the gradient of the magnetic field, μ being further defined as:

$$W_\perp / B = mV_\perp^2 / 2B$$

where $W_\perp$ is the kinetic energy in the direction perpendicular to that of the magnetic field lines and B is the magnetic field strength at the line of force on which the guiding center of the particle is located. The force as represented by equation (2) is the force which is responsible for a particle obeying equation (1).

Since the magnetic field is divergent in region R_1, it can be shown that the plasma will move upwardly from the heating region as shown in FIG. **1** and further it can be shown that

$$\tfrac{1}{2}M_eV_{e\perp}^2(x) \approx \tfrac{1}{2}M_eV_{e\perp}^2(Y)+\tfrac{1}{2}M_iV_{i\parallel}^2(Y) \tag{3}$$

where the left hand side is the initial electron transverse kinetic energy; the first term on the right is the transverse electron kinetic energy at some point (Y) in the expanded field region, while the final term is the ion kinetic energy parallel to B at point (Y). This last term is what constitutes the desired ion flow. It is produced by an electrostatic field set up by electrons which are accelerated according to Equation (2) in the divergent field region and pulls ions along with them. Equation (3) ignores electron kinetic energy parallel to B because $V_{e\parallel} \approx V_{i\parallel}$, so the bulk of parallel kinetic energy resides in the ions because of their greater masses. For example, if an electromagnetic energy flux of from about 1 to about 10 watts per square centimeter is applied to region R, whose altitude is 115 km, a plasma having a density (N_e) of 10^{12} per cubic centimeter will be generated and moved upward to region R_2 which has an altitude of about 1000 km. The movement of electrons in the plasma is due to the mirror force while the ions are moved by ambipolar diffusion (which results from the electrostatic field). This effectively "lifts" a layer of plasma **12** from the ionosphere and/or magnetosphere to a higher elevation R_2. The total energy required to create a plasma with a base area of 3 square kilometers and a height of 1000 km is about 3×10^{13} joules.

FIG. **3** is an idealized representation of movement of plasma **12** upon excitation by electron cyclotron resonance within the earth's divergent force field. Electrons (e) are accelerated to velocities required to generate the necessary mirror force to cause their upward movement. At the same time neutral particles (n) which are present along line **11** in region $R_1$are ionized and become part of plasma **12**. As electrons (e) move upward along line **11**, they drag ions (i) and neutrals (n) with them but at an angle θ of about 13 degrees to field line **11**. Also, any particulates that may be present in region R_1, will be swept upwardly with the plasma. As the charged particles of plasma **12** move upward, other particles such as neutrals within or below R_1, move in to replace the upwardly moving particles. These neutrals, under some conditions, can drag with them charged particles.

For example, as a plasma moves upward, other particles at the same altitude as the plasma move horizontally into the region to replace the rising plasma and to form new plasma. The kinetic energy developed by said other particles as they move horizontally is, for example, on the same order of magnitude as the total zonal kinetic energy of stratospheric winds known to exist.

Referring again to FIG. **2**, plasma **12** in region R_1 is moved upward along field line **11**. The plasma **12** will then form a plume (cross-hatched area in FIG. **2**) which will be relatively stable for prolonged periods of time. The exact period of time will vary widely and be determined by gravitational forces and a combination of radiative and diffusive loss terms. In the previous detailed example, the calculations were based on forming a plume by producing 0^+energies of 2 ev/particle. About 10 ev per particle would be required to expand plasma **12** to apex point C (FIG. **1**). There at least some of the particles of plasma **12** will be trapped and will oscillate between mirror points along field line **11**. This oscillation will then allow additional heating of the trapped plasma **12** by stochastic heating which is associated with trapped and oscillating particles. See "A New Mechanism for Accelerating Electrons in the Outer Ionosphere" by R. A. Helliwell and T. F. Bell, Journal of Geophysical Research, Vol. 65, No. 6, June, 1960. This is preferably carried out at an altitude of at least 500 km.

The plasma of the typical example might be employed to modify or disrupt microwave transmissions of satellites. If less than total black-out of transmission is desired (e.g., scrambling by phase shifting digital signals), the density of the plasma (N_e) need only be at least about 10^6 per cubic centimeter for a plasma orginating at an altitude of from about 250 to about 400 km and accordingly less energy (i.e., electromagnetic radiation), e.g., 10^8 joules need be provided. Likewise, if the density $N_e$is on the order of 10^8, a properly positioned plume will provide a reflecting surface for VHF waves and can be used to enhance, interfere with, or otherwise modify communication transmissions. It can be seen from the foregoing that by appropriate application of various aspects of this invention at strategic locations and with adequate power sources, a means and method is provided to cause interference with or even total disruption of communications over a very large portion of the earth. This invention could be employed to disrupt not only land based communications, both civilian and military, but also airborne communications and sea communications (both surface and subsurface). This would have significant military implications, particularly as a barrier to or confusing factor for hostile missiles or airplanes. The belt or belts of enhanced ionization produced by the method and apparatus of this invention, particularly if set up over Northern Alaska and Canada, could be employed as an early warning device, as well as a communications disruption medium. Further, the simple ability to produce such a situation in a practical time period can by itself be a deterring force to hostile action. The ideal combination of suitable field lines intersecting the earth's surface at the point where substantial fuel sources are available for generation of very large quantitities of electromagnetic power, such as the North Slope of Alaska, provides the wherewithal to accomplish the foregoing in a practical time period, e.g., strategic requirements could necessitate achieving the desired altered regions in time periods of two minutes or less and this is achievable with this invention, especially when the combination of natural gas and magnetohydrodynamic, gas turbine, fuel cell and/or EGD electric generators are employed at the point where the useful field lines intersect the earth's surface. One feature of this invention which satisfies a basic requirement of a weapon system, i.e., continuous checking of operability, is that small amounts of power can be generated for operability checking purposes. Further, in the exploitation of this invention, since the main electromagnetic beam which generates the enhanced ionized belt of this invention can be modulated itself and/or one or more additional electromagnetic radiation waves can be impinged on the ionized region formed by this invention as will be described in greater detail herein after with respect to FIG. **4**, a substantial amount of randomly modulated signals of very large power magnitude can be generated in a highly nonlinear mode. This can cause confusion of or interference with or even complete disruption of guidance systems employed by even the most sophisticated of airplanes and missiles. The ability to employ and transmit over very wide areas of the earth a plurality of electromagnetic waves of varying frequencies and to change same at will in a random manner, provides a unique ability to interfere with all modes of communications, land, sea, and/or air, at the same time. Because of the unique juxtaposition of usable fuel source at the point where desirable field lines intersect the earth's surface, such wide ranging and complete communication interference can be achieved in a resonably short period of time. Because of the mirroring phenomenon discussed hereinabove, it can also be prolonged for substantial time periods so that it would not be a mere transient effect that could simply be waited out by an opposing force. Thus, this invention provides the ability to put unprecedented amounts of power in the earth's atmosphere at strategic locations and to maintain the power injection level, particularly if random pulsing is employed, in a manner far more precise and better controlled than heretofore accomplished by the prior art, particularly by the detonation of nuclear devices of various yeilds at various altitudes. Where the prior art approaches yielded merely transitory effects, the unique combination of fuel and desirable field lines at the point where the fuel occurs allows the establishment of, compared to prior art approaches, precisely controlled and long-lasting effects which cannot, practically speaking, simply be waited out. Further, by knowing the frequencies of the various electromagnetic beams employed in the practice of this invention, it is possible not only to interfere with third party communications but to take advantage of one or more such beams to carry out a communications network even though the rest of the world's communications are disrupted. Put another way, what is used to disrupt another's communications can be employed by one knowledgeable of this invention as a communications network at the same time. In addition, once one's own communication network is established, the far-reaching extent of the effects of this invention could be employed to pick up communication signals of other for intelligence purposes. Thus, it can be seen that the disrupting effects achievable by this invention can be employed to benefit by the party who is practicing this invention since knowledge of the various electromagnetic waves being employed and how they will vary in frequency and magnitude can be used to an advantage for positive communication and eavesdropping purposes at the same time. However, this invention is not limited to locations where the fuel source naturally exists or where desirable field lines naturally intersect the earth's surface. For example, fuel, particularly hydrocarbon fuel, can be transported by pipeline and the like to the location where the invention is to be practiced.

FIG. 4 illustrates another embodiment wherein a selected region of plasma R_3 which lies within the earth's ionosphere is altered to increase the density thereof whereby a relatively stable layer **30** of relatively dense plasma is maintained within region R_3. Electromagnetic radiation is transmitted at the outset essentially parallel to field line **11** via antenna **15** as a right hand circularly polarized wave and at a frequency (e.g., 1.78 megahertz when the magnetic field at the desired altitude is 0.66 gauss) capable of exciting electron cyclotron resonance in plasma **12** at the particular altitude of plasma **12**. This causes heating of the particles (electrons, ions, neutrals, and particulates) and ionization of the uncharged particles adjacent line **11**, all of which are absorbed into plasma **12** to increase the density thereof. The power transmitted, e.g., 2×10^6 watts for up to 2 minutes heating time, is less than that required to generate the mirror force F required to move plasma **12** upward as in the previous embodiment.

While continuing to transmit electromagnetic radiation **20** from antenna **15**, a second electromagnetic radiation beam **31**, which is at a defined frequency different from the radiation from antenna **15**, is transmitted from one or more second sources via antenna **32** into layer **30** and is absorbed into a portion of layer **30** (cross-hatched area in FIG. **4**). The electromagnetic radiation wave from antenna **32** is amplitude modulated to match a known mode of oscillation f_3 in layer **30**. This creates a resonance in layer **30** which excites a new plasma wave **33** which also has a frequency of f_3 and which then propogates through the ionosphere. Wave **33** can be used to improve or disrupt communications or both depending on what is desired in a particular application. Of course, more than one new wave **33** can be generated and the various new waves can be modulated at will and in a highly nonlinear fashion.

FIG. **5** shows apparatus useful in this invention, particularly when those applications of this invention are employed which require extremely large amounts of power. In FIG. **5** there is shown the earth's surface **40** with a well **41** extending downwardly thereinto until it penetrates hydrocarbon producing reservoir **42**. Hydrocarbon reservoir **42** produces natural gas alone or in combination with crude oil. Hydrocarbons are produced from reservoir **42** through well **41** and wellhead **43** to a treating system **44** by way of pipe **45**. In treater **44**, desirable liquids such as crude oil and gas condensates are separated and recovered by way of pipe **46** while undesirable gases and liquids such as water, H_2S, and the like are separated by way of pipe **47**. Desirable gases such as carbon dioxide are separated by way of pipe **48**, and the remaining natural gas stream is removed from treater **44** by way of pipe **49** for storage in conventional tankage means (not shown) for future use and/or use in an electrical generator such as a magnetohydrodynamic, gas turbine, fuel cell or EGD generator **50**. Any desired number and combination of different types of electric generators can be employed in the practice of this invention. The natural gas is burned in generator **50** to produce substantial quantities of electricity which is then stored and/or passed by way of wire **51** to a transmitter **52** which generates the electromagnetic radiation to be used in the method of this invention. The electromagnetic radiation is then passed by way of wire **53** to antenna **54** which is located at or near the end of field line **11**. Antenna **54** sends circularly polarized radiation wave **20** upwards along field line **11** to carry out the various methods of this invention as described hereinabove.

Of course, the fuel source need not be used in its naturally-occurring state but could first be converted to another second energy source form such as hydrogen, hydrazine and the like, and electricity then generated from said second energy source form.

It can be seen from the foregoing that when desirable field line **11** intersects earth's surface **40** at or near a large naturally-occurring hydrocarbon source **42**, exceedingly large amounts of power can be very efficiently produced and transmitted in the direction of field lines. This is particularly so when the fuel source is natural gas and magnetohydrodynamic generators are employed. Further, this can all be accomplished in a relatively small physical area when there is the unique coincidence of fuel source **42** and desirable field line **11**. Of course, only one set of equipment is shown in FIG. **5** for sake of simplicity. For a large hydrocarbon reservoir **42**, a plurality of wells **41** can be employed to feed one or more storage means and/or treaters and as large a number of generators **55** as needed to power one or more transmitters **52** and one or more antennas **54**. Since all of the apparatus **44** through **54** can be employed and used essentially at the sight where naturally-occurring fuel source **42** is located, all the necessary electromagnetic radiation **20** is generated essentially at the same location as fuel source **42**. This provides for a maximum amount of usable electromagnetic radiation **20** since there are no significant storage or transportation losses to be incurred. In other words, the apparatus is brought to the sight of the fuel source where desirable field line **11** intersects the earth's surface **40** on or near the geographical location of fuel source **42**, fuel source **42** being at a desirable magnetic latitude for the practice of this invention, for example, Alaska.

The generation of electricity by motion of a conducting fluid through a magnetic field, i.e., magnetohydrodynamics (MHD), provides a method of electric power generation without moving mechanical parts and when the conducting fluid is a plasma formed by combustion of a fuel such as natural gas, an idealized combination of apparatus is realized since the very clean-burning natural gas forms the conducting plasma in an efficient manner and the thus formed plasma, when passed through a magnetic field, generates electricity in a very efficient manner. Thus, the use of fuel source **42** to generate a plasma by combustion thereof for the generation of electricity essentially at the site of occurrence of the fuel source is unique and ideal when high power levels are required and desirable field lines **11** intersect the earth's surface **40** at or near the site of fuel source **42**. A particular advantage for MHD generators is that they can be made to generate large amounts of power with a small volume, light weight device. For example, a 1000 megawatt MHD generator can be construed using superconducting magnets to weigh roughly 42,000 pounds and can be readily air lifted.

This invention has a phenomenal variety of possible ramifications and potential future developments. As alluded to earlier, missile or aircraft destruction, deflection, or confusion could result, particularly when relativistic particles are employed. Also, large regions of the atmosphere could be lifted to an unexpectedly high altitude so that missiles encounter unexpected and unplanned drag forces with resultant destruction or deflection of same. Weather modification is possible by, for example, altering upper atmosphere wind patterns or altering solar absorption patterns by constructing one or more plumes of atmospheric particles which will act as a lens or focusing device. Also as alluded to earlier, molecular modifications of the atmosphere can take place so that positive environmental effects can be achieved. Besides actually changing the molecular composition of an atmospheric region, a particular molecule or molecules can be chosen for increased presence. For example, ozone, nitrogen, etc. concentrations in the atmosphere could be artificially increased. Similarly, environmental enhancement could be achieved by causing the breakup of various chemical entities such as carbon dioxide, carbon monoxide, nitrous oxides, and the like. Transportation of entities can also be realized when advantage is taken of the drag effects caused by regions of the atmosphere moving up along diverging field lines. Small micron sized particles can be then transported, and, under certain circumstances and with the availability of sufficient energy, larger particles or objects could be similarly affected. Particles with desired characteristics such as tackiness, reflectivity, absorptivity, etc., can be transported for specific purposes or effects. For example, a plume of tacky particles could be established to increase the drag on a missile or satellite passing therethrough. Even plumes of plasma having substantially less charged particle density than described above will produce drag effects on missiles which will affect a lightweight (dummy) missile in a manner substantially different than a heavy (live) missile and this affect can be used to distinguish between the two types of missiles. A moving plume could also serve as a means for supplying a space station or for focusing vast amount of sunlight on selected portions of the earth. Surveys of global scope could also be realized because the earth's natural magnetic field could be significantly altered in a controlled manner by plasma beta effects resulting in, for example, improved magnetotelluric surveys. Electromagnetic pulse defenses are also possible. The earth's magnetic field could be decreased or disrupted at appropriate altitudes to modify or eliminate the magnetic field in high Compton electron generation (e.g., from high altitude nuclear bursts) regions. High intensity, well controlled electrical fields can be provided in selected locations for various purposes. For example, the plasma sheath surrounding a missile or satellite could be used as a trigger for activating such a high intensity field to destroy the missile or satellite. Further, irregularities can be created in the ionosphere which will interfere with the normal operation of various types of radar, e.g., synthetic aperture radar. The present invention can also be used to create artificial belts of trapped particles which in turn can be studied to determine the stability of such parties. Still further, plumes in accordance with the present invention can be formed to simulate and/or perform the same functions as performed by the detonation of a "heave" type nuclear device without actually having to detonate such a device. Thus it can be seen that the ramifications are numerous, far-reaching, and exceedingly varied in usefulness.

I claim:

1. A method for altering at least one region normally existing above the earth's surface with electromagnetic radiation using naturally-occurring and diverging magnetic field lines of the earth comprising transmitting first electromagnetic radiation at a frequency between 20 and 7200 kHz from the earth's surface, said transmitting being conducted essentially at the outset of transmission substantially parallel to and along at least one of said field lines, adjusting the frequency of said first radiation to a value which will excite electron cyclotron resonance at an initial elevation at least 50 km above the earth's surface, whereby in the region in which said electron cyclotron resonance takes place heating, further ionization, and movement of both charged and neutral particles is effected, said cyclotron resonance excitation of said region is continued until the electron concentration of said region reaches a value of at least 10^6 per cubic centimeter and has an ion energy of at least 2 ev.

2. The method of claim **1** including the step of providing artificial particles in said at least one region which are excited by said electron cyclotron resonance.

3. The method of claim 2 wherein said artificial particles are provided by injecting same into said at least one region from an orbiting satellite.

4. The method of claim 1 wherein said threshold excitation of electron cyclotron resonance is about 1 watt per cubic centimeter and is sufficient to cause movement of a plasma region along said diverging magnetic field lines to an altitude higher than the altitude at which said excitation was initiated.

5. The method of claim 4 wherein said rising plasma region pulls with it a substantial portion of neutral particles of the atmosphere which exist in or near said plasma region.

6. The method of claim 1 wherein there is provided at least one separate source of second electromagnetic radiation, said second radiation having at least one frequency different from said first radiation, impinging said at least one second radiation on said region while said region is undergoing electron cyclotron resonance excitation caused by said first radiation.

7. The method of claim 6 wherein said second radiation has a frequency which is absorbed by said region.

8. The method of claim 6 wherein said region is plasma in the ionosphere and said second radiation excites plasma waves within said ionosphere.

9. The method of claim 8 wherein said electron concentration reaches a value of at least 10^{12} per cubic centimeter.

10. The method of claim 8 wherein said excitation of electron cyclotron resonance is initially carried out within the ionosphere and is continued for a time sufficient to allow said region to rise above said ionosphere.

11. The method of claim 1 wherein said excitation of electron cyclotron resonance is carried out above about 500 kilometers and for a time of from 0.1 to 1200 seconds such that multiple heating of said plasma region is achieved by means of stochastic heating in the magnetosphere.

12. The method of claim 1 wherein said first electromagnetic radiation is right hand circularly polarized in the northern hemisphere and left hand circularly polarized in the southern hemisphere.

13. The method of claim 1 wherein said electromagnetic radiation is generated at the site of a naturally-occurring hydrocarbon fuel source, said fuel source being located in at least one of northerly or southerly magnetic latitudes.

14. The method of claim 13 wherein said fuel source is natural gas and electricity for generating said electromagnetic radiation is obtained by burning said natural gas in at least one of magnetohydrodynamic, gas turbine, fuel cell, and EGD electric generators located at the site where said natural gas naturally occurs in the earth.

15. The method of claim 14 wherein said site of natural gas is within the magnetic latitudes that encompass Alaska.

* * * * *

N. TESLA.

APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.

APPLICATION FILED JAN. 18, 1902. RENEWED MAY 4, 1907.

1,119,732. Patented Dec. 1, 1914.

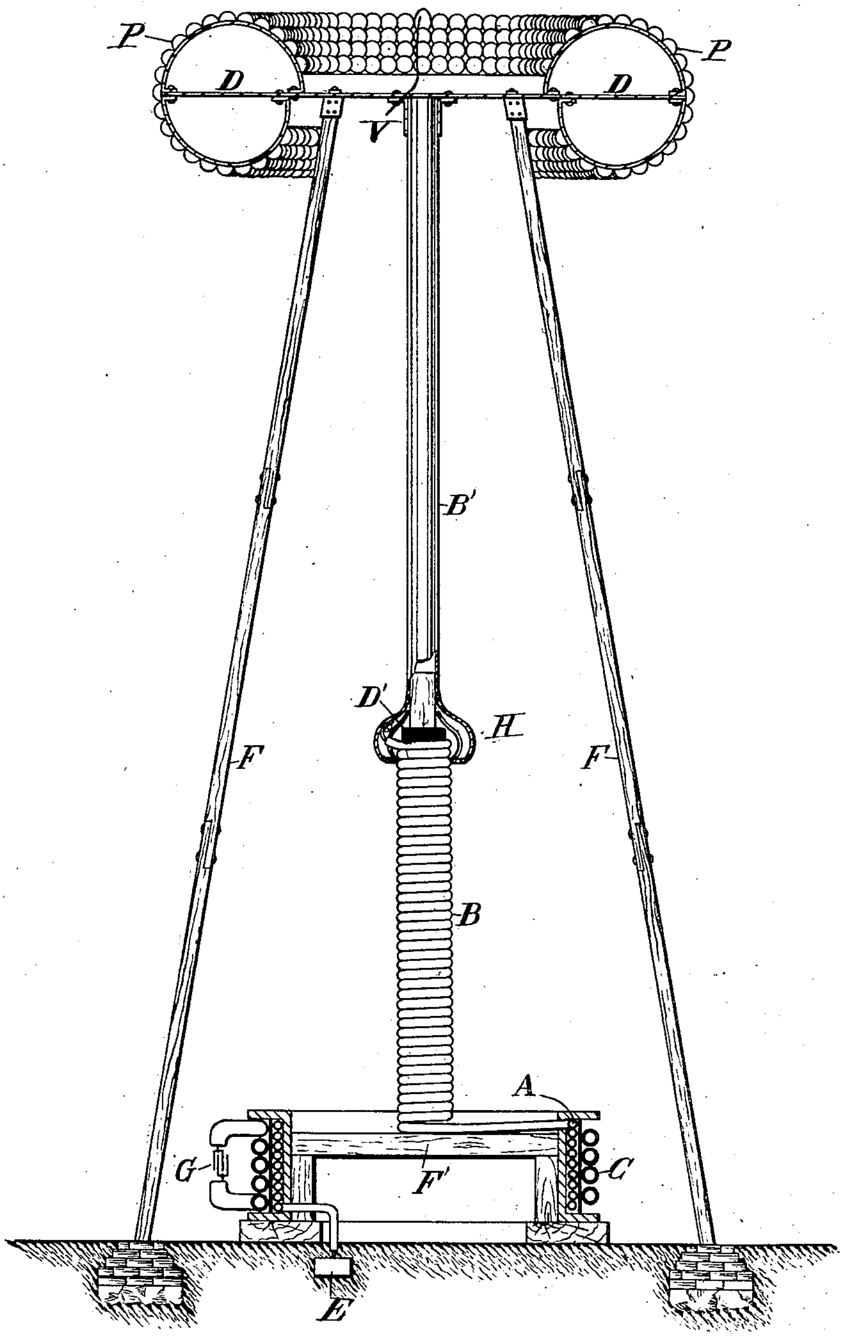

WITNESSES:

M. Lawson Dyer

Benjamin Miller.

Nikola Tesla, INVENTOR,

BY Kerr, Page & Cooper,

his ATTORNEYS.

UNITED STATES PATENT OFFICE.

NIKOLA TESLA, OF NEW YORK, N. Y.

APPARATUS FOR TRANSMITTING ELECTRICAL ENERGY.

1,119,732. Specification of Letters Patent. **Patented Dec. 1, 1914.**

Application filed January 18, 1902, Serial No. 90,245. Renewed May 4, 1907. Serial No. 371,817.

To all whom it may concern:

Be it known that I, Nikola Tesla, a citizen of the United States, residing in the borough of Manhattan, in the city, county, and State of New York, have invented certain new and useful Improvements in Apparatus for Transmitting Electrical Energy, of which the following is a specification, reference being had to the drawing accompanying and forming a part of the same.

In endeavoring to adapt currents or discharges of very high tension to various valuable uses, as the distribution of energy through wires from central plants to distant places of consumption, or the transmission of powerful disturbances to great distances, through the natural or non-artificial media, I have encountered difficulties in confining considerable amounts of electricity to the conductors and preventing its leakage over their supports, or its escape into the ambient air, which always takes place when the electric surface density reaches a certain value.

The intensity of the effect of a transmitting circuit with a free or elevated terminal is proportionate to the quantity of electricity displaced, which is determined by the product of the capacity of the circuit, the pressure, and the frequency of the currents employed. To produce an electrical movement of the required magnitude it is desirable to charge the terminal as highly as possible, for while a great quantity of electricity may also be displaced by a large capacity charged to low pressure, there are disadvantages met with in many cases when the former is made too large. The chief of these are due to the fact that an increase of the capacity entails a lowering of the frequency of the impulses or discharges and a diminution of the energy of vibration. This will be understood when it is borne in mind, that a circuit with a large capacity behaves as a slackspring, whereas one with a small capacity acts like a stiff spring, vibrating more vigorously. Therefore, in order to attain the highest possible frequency, which for certain purposes is advantageous and, apart from that, to develop the greatest energy in such a transmitting circuit, I employ a terminal of relatively small capacity, which I charge to as high a pressure as practicable. To accomplish this result I have found it imperative to so construct the elevated conductor, that its outer surface, on which the electrical charge chiefly accumulates, has itself a large radius of curvature, or is composed of separate elements which, irrespective of their own radius of curvature, are arranged in close proximity to each other and so, that the outside ideal surface enveloping them is of a large radius. Evidently, the smaller the radius of curvature the greater, for a given electric displacement, will be the surface-density and, consequently, the lower the limiting pressure to which the terminal may be charged without electricity escaping into the air. Such a terminal I secure to an insulating support entering more or less into its interior, and I likewise connect the circuit to it inside or, generally, at points where the electric density is small. This plan of constructing and supporting a highly charged conductor I have found to be of great practical importance, and it may be usefully applied in many ways.

Referring to the accompanying drawing, the figure is a view in elevation and part section of an improved free terminal and circuit of large surface with supporting structure and generating apparatus.

The terminal D consists of a suitably shaped metallic frame, in this case a ring of nearly circular cross section, which is covered with half spherical metal plates P P, thus constituting a very large conducting surface, smooth on all places where the electric charge principally accumulates. The frame is carried by a strong platform expressly provided for safety appliances, instruments of observation, etc., which in turn rests on insulating supports F F. These should penetrate far into the hollow space formed by the terminal, and if the electric density at the points where they are bolted to the frame is still considerable, they may be specially protected by conducting hoods as H.

A part of the improvements which form the subject of this specification, the transmitting circuit, in its general features, is identical with that described and claimed in my original Patents Nos. 645,576 and 649,621. The circuit comprises a coil A which is in close inductive relation with a primary C, and one end of which is connected to a ground-plate E, while its other end is led through a separate self-induction coil B and a metallic cylinder B′ to the terminal D.

The connection to the latter should always be made at, or near the center, in order to secure a symmetrical distribution of the current, as otherwise, when the frequency is very high and the flow of large volume, the performance of the apparatus might be impaired. The primary C may be excited in any desired manner, from a suitable source of currents G, which may be an alternator or condenser, the important requirement being that the resonant condition is established, that is to say, that the terminal D is charged to the maximum pressure developed in the circuit, as I have specified in my original patents before referred to. The adjustments should be made with particular care when the transmitter is one of great power, not only on account of economy, but also in order to avoid danger. I have shown that it is practicable to produce in a resonating circuit as E A B B′ D immense electrical activities, measured by tens and even hundreds of thousands of horse-power, and in such a case, if the points of maximum pressure should be shifted below the terminal D, along coil B, a ball of fire might break out and destroy the support F or anything else in the way. For the better appreciation of the nature of this danger it should be stated, that the destructive action may take place with inconceivable violence. This will cease to be surprising when it is borne in mind, that the entire energy accumulated in the excited circuit, instead of requiring, as under normal working conditions, one quarter of the period or more for its transformation from static to kinetic form, may spend itself in an incomparably smaller interval of time, at a rate of many millions of horse power. The accident is apt to occur when, the transmitting circuit being strongly excited, the impressed oscillations upon it are caused, in any manner more or less sudden, to be more rapid than the free oscillations. It is therefore advisable to begin the adjustments with feeble and somewhat slower impressed oscillations, strengthening and quickening them gradually, until the apparatus has been brought under perfect control. To increase the safety, I provide on a convenient place, preferably on terminal D, one or more elements or plates either of somewhat smaller radius of curvature or protruding more or less beyond the others (in which case they may be of larger radius of curvature) so that, should the pressure rise to a value, beyond which it is not desired to go, the powerful discharge may dart out there and lose itself harmlessly in the air. Such a plate, performing a function similar to that of a safety valve on a high pressure reservoir, is indicated at V.

Still further extending the principles underlying my invention, special reference is made to coil B and conductor B′. The latter is in the form of a cylinder with smooth or polished surface of a radius much larger than that of the half spherical elements P P, and widens out at the bottom into a hood H, which should be slotted to avoid loss by eddy currents and the purpose of which will be clear from the foregoing. The coil B is wound on a frame or drum D^1 of insulating material, with its turns close together. I have discovered that when so wound the effect of the small radius of curvature of the wire itself is overcome and the coil behaves as a conductor of large radius of curvature, corresponding to that of the drum. This feature is of considerable practical importance and is applicable not only in this special instance, but generally. For example, such plates at P P of terminal D, though preferably of large radius of curvature, need not be necessarily so, for provided only that the individual plates or elements of a high potential conductor or terminal are arranged in proximity to each other and with their outer boundaries along an ideal symmetrical enveloping surface of a large radius of curvature, the advantages of the invention will be more or less fully realized. The lower end of the coil B—which, if desired, may be extended up to the terminal D—should be somewhat below the uppermost turn of coil A. This, I find, lessens the tendency of the charge to break out from the wire connecting both and to pass along the support F′.

Having described my invention, I claim:

1. As a means for producing great electrical activities a resonant circuit having its outer conducting boundaries, which are charged to a high potential, arranged in surfaces of large radii of curvature so as to prevent leakage of the oscillating charge, substantially as set forth.

2. In apparatus for the transmission of electrical energy a circuit connected to ground and to an elevated terminal and having its outer conducting boundaries, which are subject to high tension, arranged in surfaces of large radii of curvature substantially as, and for the purpose described.

3. In a plant for the transmission of electrical energy without wires, in combination with a primary or exciting circuit a secondary connected to ground and to an elevated terminal and having its outer conducting boundaries, which are charged to a high potential, arranged in surfaces of large radii of curvature for the purpose of preventing leakage and loss of energy, substantially as set forth.

4. As a means for transmitting electrical energy to a distance through the natural media a grounded resonant circuit, comprising a part upon which oscillations are impressed and another for raising the ten-

sion, having its outer conducting boundaries on which a high tension charge accumulates arranged in surfaces of large radii of curvature, substantially as described.

5. The means for producing excessive electric potentials consisting of a primary exciting circuit and a resonant secondary having its outer conducting elements which are subject to high tension arranged in proximity to each other and in surfaces of large radii of curvature so as to prevent leakage of the charge and attendant lowering of potential, substantially as described.

6. A circuit comprising a part upon which oscillations are impressed and another part for raising the tension by resonance, the latter part being supported on places of low electric density and having its outermost conducting boundaries arranged in surfaces of large radii of curvature, as set forth.

7. In apparatus for the transmission of electrical energy without wires a grounded circuit the outer conducting elements of which have a great aggregate area and are arranged in surfaces of large radii of curvature so as to permit the storing of a high charge at a small electric density and prevent loss through leakage, substantially as described.

8. A wireless transmitter comprising in combination a source of oscillations as a condenser, a primary exciting circuit and a secondary grounded and elevated conductor the outer conducting boundaries of which are in proximity to each other and arranged in surfaces of large radii of curvature, substantially as described.

9. In apparatus for the transmission of electrical energy without wires an elevated conductor or antenna having its outer high potential conducting or capacity elements arranged in proximity to each other and in surfaces of large radii of curvature so as to overcome the effect of the small radius of curvature of the individual elements and leakage of the charge, as set forth.

10. A grounded resonant transmitting circuit having its outer conducting boundaries arranged in surfaces of large radii of curvature in combination with an elevated terminal of great surface supported at points of low electric density, substantially as described.

NIKOLA TESLA.

Witnesses:
M. LAMSON DYER,
RICHARD DONOVAN.

Nov. 12, 1957

L. R. CRUMP

POWERING ELECTRICAL DEVICES WITH ENERGY ABSTRACTED FROM THE ATMOSPHERE

Filed March 12, 1954

2,813,242

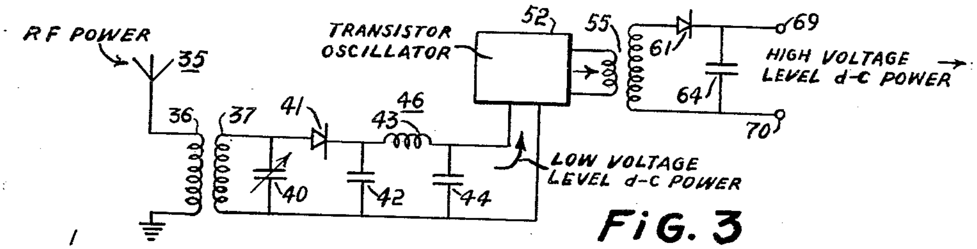
RF POWER
35
36
37
41
46
43
40
42
44
TRANSISTOR OSCILLATOR
52
55
61
64
69
70
HIGH VOLTAGE LEVEL d-c POWER
LOW VOLTAGE LEVEL d-c POWER

FiG. 3

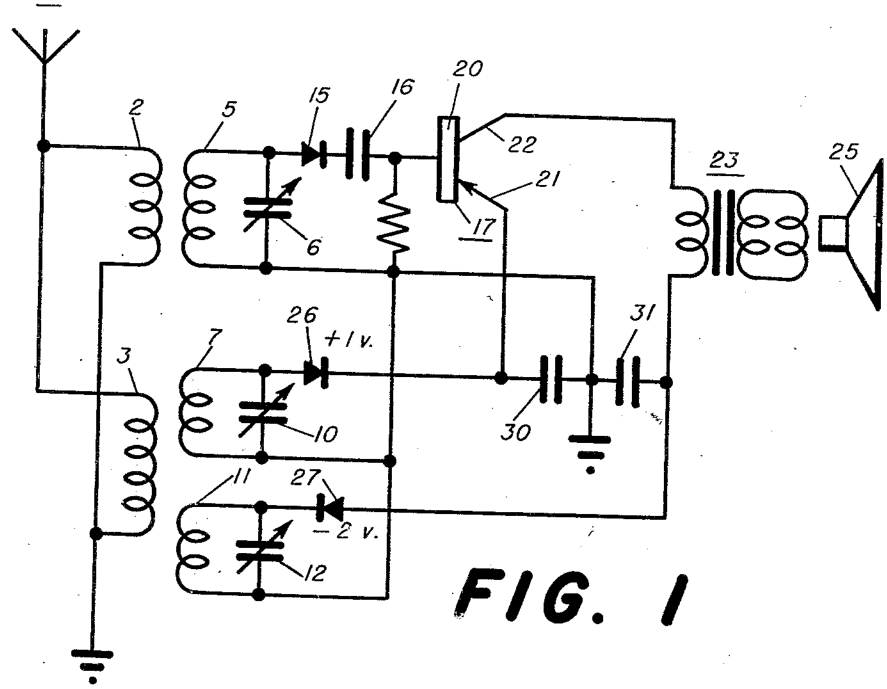
1
2
3
5
6
7
10
11
12
15
16
17
20
21
22
23
25
26
+1 v.
27
−2 v.
30
31

FIG. 1

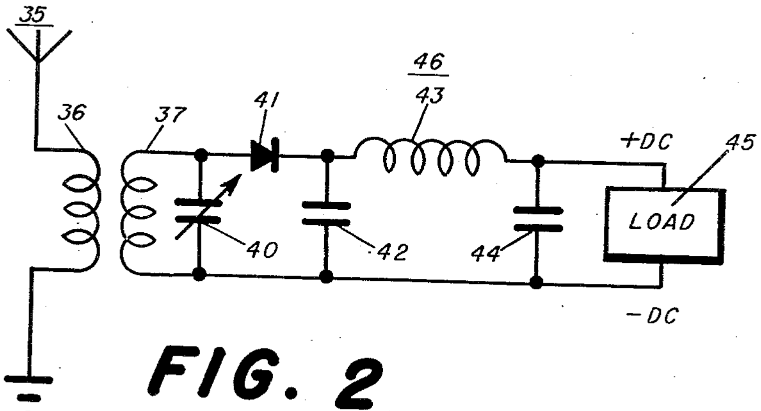
35
36
37
41
46
43
40
42
44
+DC
45
LOAD
−DC

FIG. 2

INVENTOR

Lloyd R. Crump

BY W. E. Thibodeau + A. W. Dew

ATTORNEYS

United States Patent Office

2,813,242

Patented Nov. 12, 1957

2,813,242

POWERING ELECTRICAL DEVICES WITH ENERGY ABSTRACTED FROM THE ATMOSPHERE

Lloyd R. Crump, Silver Spring, Md.

Application March 12, 1954, Serial No. 415,986

1 Claim. (Cl. 321—2)

(Granted under Title 35, U. S. Code (1952), sec. 266)

The invention described herein may be manufactured and used by or for the Government for governmental purposes without payment to me of any royalty thereon.

This invention relates to the convenient and economical provision of power for the operation of electronic circuits and devices using transistors, and of other electrical devices having modest power requirements.

A great advantage of transistors, and a major reason for their enthusiastic reception since their introduction a few years ago, is the fact that they will operate satisfactorily with very low supply voltages and currents. One milliwatt or even less is sufficient to power a transistor in many applications. Various batteries have been developed to provide, in a minimum of space, the relatively minute amounts of power needed by transistors.

My invention provides methods and means that permit transistor circuits, and also other low-powered electrical devices, to be economically and conveniently operated without any batteries whatever, and indeed without any power supply whatever as power supplies are ordinarily conceived.

The invention centers around my discovery that it is practicable to construct operative transistor circuits that are able to abstract from the atmosphere sufficient electromagnetic energy to provide all necessary supply voltages and currents for their own operation. Circuits and devices powered according to my invention will operate indefinitely without any local power source whatever.

I have successfully constructed and demonstrated such circuits. For example, I have constructed a batteryless transistor radio receiver on which I have listened to either nearby or distant broadcast stations as desired, using either headphones or a loudspeaker; this receiver has been powered entirely by electromagnetic energy abstracted from the atmosphere.

From the successful operation of this receiver, and from other experimental work, it becomes clear that, by the methods and means of the invention, a great variety of practical and useful transistor circuits can be powered entirely by energy abstracted from the atmosphere.

Furthermore, as will become apparent below, my invention is applicable to the powering of other electrical devices requiring relatively small amounts of power.

An object of the present invention is to provide methods and means for powering transistor circuits entirely from radiofrequency energy abstracted from the atmosphere.

Another object is to provide methods and means for powering remote radio receivers, low-powered radio transmitters, and other low-powered electrical devices, with energy received by radio from a master station, so that no local power supplies are needed by the devices and so that the powering or non-powering of the remote device is under the control of the master station.

A further object is to provide methods and means for powering transistor circuits and other low-powered electrical devices with radiofrequency energy received from one or more remote radio transmitters.

Other objects, aspects, uses, and advantages of the invention will become apparent from the following description and from the drawing.

Figure 1 is a schematic diagram of a transistor radio receiver in which all necessary power is supplied by energy abstracted from the atmosphere in accordance with the invention.

Figure 2 is a schematic diagram showing a general application of the invention to provide direct-current power to a load.

Figure 3 is a schematic diagram of a system for obtaining a high energy D.-C. source at a high voltage level using energy abstracted from the atmosphere.

Referring to Figure 1, a receiving antenna **1** is connected to antenna coupling coils **2** and **3**, the other ends of which are connected to ground. A parallel resonant circuit consisting of coil **5** and capacitor **6** is coupled to coil **2**. A second parallel resonant circuit consisting of coil **7** and capacitor **10** is coupled to coil **3**. A third parallel resonant circuit consisting of coil **11** and capacitor **12** is also coupled to coil **3**.

Coil **5** and capacitor **6** are tuned to the frequency of a radio transmitter from which it is desired to receive information—for instance, an amplitude-modulated standard broadcast station. The signal received from this transmitter need not be strong. The signal is detected by diode **15** to obtain an audio-frequency information signal. This audio signal is coupled through a capacitor **16** and is amplified by a circuit that includes a transistor **17** having a base **20**, an emitter **21**, and a collector **22**. The amplified audio output of the transistor is coupled through an audio transformer **23** to an electroacoustical transducer, preferably a permanent-magnet dynamic loudspeaker **25** as shown.

The novelty of the invention lies largely in the method and means by which the necessary direct-current power is supplied to the emitter and collector circuits of transistor **17**. This method and means will now be described.

Coil **7** and capacitor **10**, and also coil **11** and capacitor **12**, are tuned to receive radio signals of relatively high strength. It does not matter whether these signals contain information. These power signals are rectified by diodes **26** and **27** to provide direct-current power that is filtered by capacitors **30** and **31**. The D.-C. power thus obtained is utilized to power the transistor **17**.

In the circuit shown, two tuned circuits (coil **7** and capacitor **10**, and coil **11** and capacitor **12**) are tuned to power signals and the D.-C. voltages obtained from each are connected in series. The tuned power circuits may be tuned to the same or different power signals. Under certain circumstances it may be desirable to use more than two tuned power circuits and to tune them to more than two power signals; in this way power can be obtained from several signals and combined. On the other hand, if a strong power signal is available, a single tuned power circuit may suffice to give the needed D.-C. power.

Even weak information signals can be received successfully. A plurality of transistor amplifier stages can be used if desired, or other circuits such as superheterodyne circuits can be used. It is merely necessary that a sufficiently strong power signal or signals be available to provide the small amount of D.-C. needed to power the transistors.

If the information signal happens to be strong, it can be used as the power signal; all of the tuned circuits (coil **5** and capacitor **6**, coil **7** and capacitor **10**, coil **11** and capacitor **12**) are tuned to the information signal.

Engineers who have observed my invention in operation have been surprised at the unexpectedly good results obtained, even with readily available power signals of quite moderate strength. For instance, sufficient power for satisfactory operation of a loudspeaker at low volumes is readily obtained from a 5-kilowatt standard broadcast station 5 miles away, using only an indoor antenna to pick up the power signal as well as information signals. In typical operation under these conditions a D.-C. voltage of about 2.5 to 3 volts is obtained between the emitter and the collector, at a current of about 250 microamperes; D.-C. power input to the transistor is thus of the order of 0.5 to 1 milliwatt. So far as I am aware, no one has ever before discovered and demonstrated the practicability of this method of powering a radio receiver.

Because existing broadcast stations within a radius of a number of miles provide adequate power signals, the invention is readily practicable with existing power signals in almost any location in or near any city in the United States.

Although I have described a transistor radio receiver powered by my invention, it will be readily apparent that the invention is applicable to the powering of any transistor circuit using one or a number of transistors, and to the powering of other devices requiring relatively small amounts of power. For instance, sensitive electromechanical, electrochemical, or electrothermal devices can be operated by the method of the invention.

Referring to Figure 2, which shows a more general embodiment of my invention, an antenna **35** picks up radiofrequency energy from the atmosphere. This energy flows through coil **36**, which is coupled to a tuned circuit consisting of coil **37** and capacitor **40**. The radiofrequency voltage across capacitor **40** is rectified by diode **41** and filtered by a low-pass filter **46** consisting of capacitors **42** and **44** and choke coil **43**. The resulting D.-C. voltage is applied to a load **45**.

In the practice of my invention, larger amounts of power can be obtained for short periods of time by storing received energy in a suitable energy storage device. Stored energy may then be withdrawn at intervals at a more rapid rate than that at which it was received and put into the storage device. In this way the invention can be used to provide short pulses of relatively very high electrical energy. This result can be readily obtained by charging a relatively large capacitor with direct current and then discharging the capacitor rapidly into a load when desired. This rapid discharge can be initiated automatically when the voltage across the capacitor reaches a certain level, or it can be initiated when a transistor radio receiver receives a certain information signal.

Higher voltages can be obtained with the invention by means of well known devices for raising D.-C. voltages as shown in Figure 3. The D.-C. voltage output from the capacitor **44** can be used to power a low frequency transistor oscillator **52** whose A.-C. output is raised to a higher voltage level by the transformer **55**. This relatively high A.-C. voltage can then be rectified by a diode **61** and fed to a capacitor **64** to provide a high energy D.-C. source at a relatively high voltage level at the terminals **69** and **70**. If desired, energy can now be withdrawn from the capacitor **64** at intervals in short pulses of high energy at a high voltage level. Pulsed radio transmission is one of the possible uses for this form of the invention. Other uses would be to provide a single relatively powerful pulse needed to actuate an electrothermal or electromechanical device.

As has been indicated above, in many locations and particularly anywhere in or near most American cities, power signals normally present in the atmosphere are readily available for the easy and convenient practice of the invention. However, the invention also has important applications in systems in which the necessary power signal is generated and transmitted specifically for the operation of the particular system. Such systems can, for example, comprise a master station transmitting all the power that is needed for hundreds or thousands of fixed or mobile transistor receivers or other remote devices over a range of many miles. This eliminates the need for hundreds or thousands, as the case may be, of local power supplies. At the same time, such a system has the advantage that all of the remote devices can be simultaneously activated or deactivated at the will of the master station, simply by starting or stopping the transmission of the power signal. In such systems it will often be advantageous to use power signals of frequencies sufficiently high to permit the use of resonant receiving antennas of small physical dimensions for signal pickup at the remote devices. In addition to the power signal, the master station may transmit an information signal on the same or a different carrier.

Certain types of devices powered entirely by received radio waves are of course well known. The well-known "crystal set" of the early days of radio, which used a diode rectifier to demodulate an amplitude-modulated radiofrequency signal, is an outstanding example of such a device. My invention is readily distinguishable from such prior devices, however. In typical prior devices a modulated radiofrequency signal is applied to a diode to obtain unidirectional half-wave pulses whose amplitudes vary with modulation. These pulses are integrated by means of a capacitor to obtain a unidirectional signal the amplitude of which follows the audiofrequency modulation envelope. If the radiofrequency signal is received with sufficient strength the audio signal may have sufficient power to operate headphones or similar utilization device without power amplification; but the signal is utilized for its information content, rather than to supply non-information-containing power.

My invention, on the other hand, entails the utilization of received radiofrequency energy to supply power to at least one pair of circuit points (across capacitor **31** in Fig. 1, for example), such circuit points requiring power solely for its power content and not for any information or modulation it may contain. In other words, my invention entails the utilization of radiofrequency energy to supply power that would otherwise have to be supplied by batteries, generator, or other local power source.

It will be apparent that the embodiments shown are only exemplary and that various modifications can be made in construction and arrangement within the scope of the invention as defined in the appended claim.

I claim:

An electrical device for obtaining a high energy D.-C. source at a high voltage level using energy abstracted from the atmosphere, said device comprising in combination: resonant means for receiving radio waves, first rectifier means for converting said radio waves into first direct current energy, first capacitor means for storing said first direct current energy, an oscillator powered by said direct current energy, said oscillator producing an A.-C. output, transformer means for raising said A.-C. output to an increased voltage level, second rectifier means for converting the A.-C. output of increased voltage level from said transformer into second direct current energy, and second capacitor means for storing said second direct current energy.

References Cited in the file of this patent

UNITED STATES PATENTS

1,540,998	Plauson	June 9, 1925
2,056,824	Cawley	Oct. 6, 1936
2,553,734	Adler	May 22, 1951
2,569,687	Pan	Oct. 2, 1951
2,647,957	Mallinckrodt	Aug. 4, 1953

OTHER REFERENCES

Turner: "A crystal receiver with transistor amplifier," Radio and Television News, January 1950.

H. C. VION.

Electric Apparatus.

No. 28,793. Patented June 19, 1860.

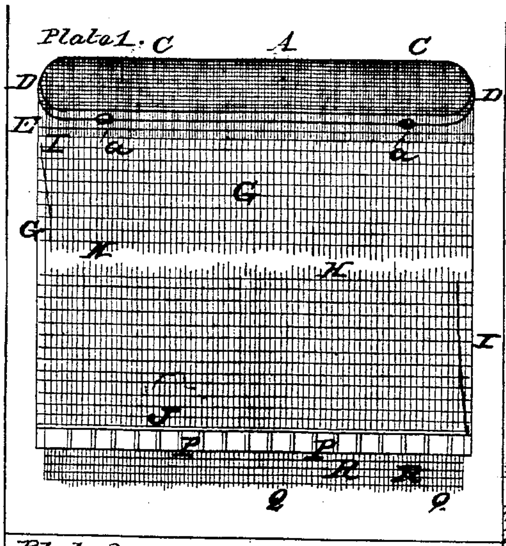

Plate 1.
C
A
C
D
D
E
L
a
a
G
G
K
K
I
J
P
P
R
R
Q
Q

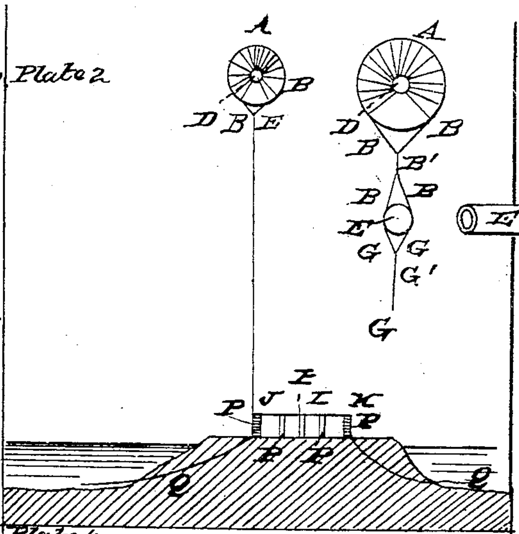

Plate 2
A
B
D
B
E
A
D
B
B
B'
B
B
E
G
G
G'
G
E
I
J
L
K
P
P
P
P
Q
Q

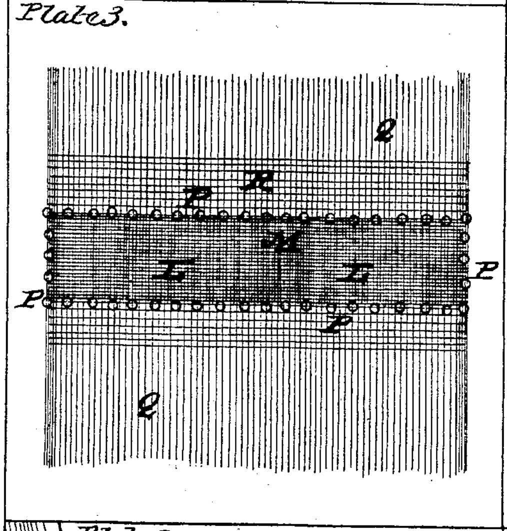

Plate 3.
Q
R
P
M
L
L
P
P
P
Q

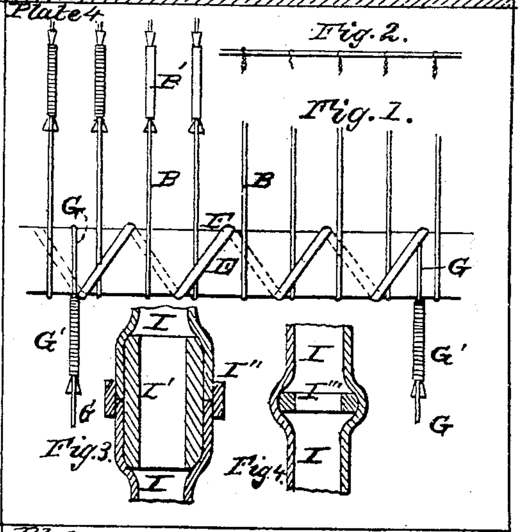

Plate 4.
Fig. 2.
Fig. 1.
E'
B
B
G
E
D
G
G'
G'
G
G
I
I'
I''
I
I'''
I
I
Fig. 3.
Fig. 4.

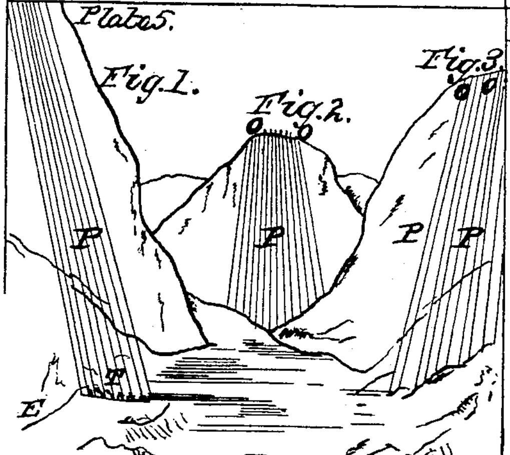

Plate 5.
Fig. 1.
Fig. 2.
Fig. 3.
O
O
O
O
P
P
P
P
P
E

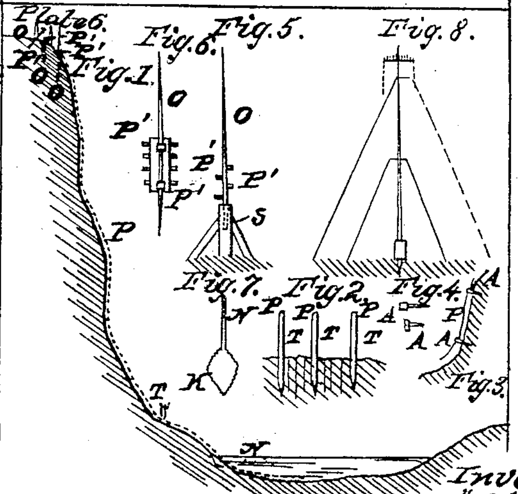

Plate 6.
Fig. 1.
Fig. 6.
Fig. 5.
Fig. 8.
O
O
P'
P'
P'
P'
S
P
Fig. 7.
Fig. 2.
Fig. 4.
N
P
P
T
T
T
A
A
A
P
A
Fig. 3.
T
K
N

Witnesses:
Arthur L. McIntire
Lewis Bosworth

Inventor:
H. C. Vion
by Munn & Co
Attorneys

N. PETERS, Photo-Lithographer, Washington, D. C.

UNITED STATES PATENT OFFICE.

H. CHARLES VION, OF PARIS, FRANCE.

IMPROVED METHOD OF UTILIZING ATMOSPHERIC ELECTRICITY.

Specification forming part of Letters Patent No. **28,793**, dated June 19, 1860.

To all whom it may concern:

Be it known that I, HIPPOLYTE CHARLES VION, of Paris, in the Empire of France, engineer, have invented a new Mode of Obtaining Atmospheric Electricity and Terrestrial Electricity and its Industrial Applications; and I do hereby declare that the following is a full, clear, and exact description of the same, reference being had to the accompanying drawings, making part of this specification, in which—

Plate 1 represents a front elevation; Plate 2, a side elevation of an aerostat, in combination with certain other devices for obtaining atmospheric and terrestrial electricity; Plate 3, a plan of the device for obtaining terrestrial electricity; Plate 4, detached portions of the apparatus. Plate 5 shows a front view, and Plate 6 a vertical section and various details of the apparatus when applied in mountainous regions, the use of an aerostat being dispensed with.

The object of my invention is to form an electric pile of great power by using the positive electricity contained in the atmosphere, and the negative electricity contained in the earth, so as to make the electricity therein contained available for industrial purposes.

To enable others skilled in the art to make and use my invention, I will proceed to describe its construction and operation.

A is aerostat of a tubular form closed at both ends. It is made of suitable material, so as to be impervious to air. Its dimensions are such as to give it great ascensional power when filled with gas.

a a are valves in the surface of the aerostat, to be opened when the inflation of the aerostat should become too great.

A long india-rubber tube, I, communicates with the interior of the aerostat, being attached to the latter, near one end of its ends D, as seen in Plate 1, while the lower end of the tube I is attached to a gasometer, (not shown in the drawings.) The gasometer is to be fed with hydrogen gas, produced by the action of the pile itself, the negative wires of which (afterward to be described) enter a body of water at or near the base of the apparatus and decompose the water so as to produce the hydrogen gas. As the aerostat is supposed to be at a considerable height the tube I must be of corresponding length, and is constructed of a number of tubes, short wooden tubes I′ being inserted where the tubes I are joined, and a fastening-ring, I″, being slipped over each of the joints, as seen in Fig. 3 of Plate 4. At certain distances the tube I is fastened to the net-work of the positive wires (afterward to be described) in order to secure the tube against the action of the wind, and at each of these fastening-places a washer, I‴, is inserted in the tube in order that the tube shall not be compressed by the wire or other means employed to fasten it to the net-work of positive wires. (See Fig. 4, Plate 4.)

The aerostat is surrounded with a net-work of wires, one layer of the wires, C, being parallel with the axis of the aerostat and fastened to rings D at both ends of the aerostat, and the other layer of wires, B, extending partially around the aerostat at right angles to the wires C. One end of each of the wires B extends around an iron tube, E, some distance below the aerostat and meets the other end of it between the tube E and the aerostat. The two ends are fastened together by a ligature, B′. (See Fig. 1, Plate 4.) Each end of the tube E terminates into a ball, *e*. The wires B are fastened to the surface of the tube E by means of a helical wire, F, wound around the tube and across the wires B, as seen in Fig. 1, Plate 4. The upper ends of long vertical wires G are also wound around cylinder E, each wire G between two of the wires B, and the ends secured by a ligature, G′, as seen in Fig. 1, Plate 4. The helical wire F is also wound across the wires G, so as to keep them in their places on tube E.

The vertical wires G, which are to be the conductors of the positive electricity of the atmosphere, must be of a length proportionate to the desired efficacy of the electric pile, and the size and ascensional power of the aerostat must, of course, be adequate to sustain the weight of and keep suspended the wires G, (a weight still further increased by the horizontal cross-wires H, with which the vertical wires G are interlaced, in order to form a net-work not liable to be deranged by the action of the winds or similar influences.) The two outside wires, G, are stronger than the rest of them, and their lower ends are fastened to dynamometers of any suitable construction. These dynamometers are attached to the ends or a massive iron cylinder, J, and they serve to indicate the tension in the outside wires, G, and the corresponding ascensional power of the aerostat. According to the reading of these dynamometers the aerostat has to be supplied (through tube I) with more or less gas. The lower end of each of the wires G is wound around the cylinder J, and secured by ligatures similar to those above described. The wires G are all insulated (by a coating of gutta-percha or similar substance,) except where they are in contact with tube E and with cylinder J, and a similar insulating coating is laid on cylinder J, after the lower ends of the wires G have been fastened to it.

Another cylinder, K, similar to J, is placed at some distance from and parallel to cylinder J. It is connected with cylinder J by wires L, wound around both cylinders and interlaced with cross-wires M. The wire-work L M and cylinders are insulated (in a manner already described) against outside influences, so that the only electric communication between the two cylinders will be through the wires L. The two cylinders are placed upon insulated columns P. The cylinder K may be used as a substitute for cylinder J, and vice versa, whenever repairs become necessary. Insulated branch wire or wires are attached to the cylinders J K and wires L, so as to conduct the positive electricity obtained from the atmosphere by means of the above-described apparatus to wherever it is desired for industrial purposes. The insulated wires Q (interlaced with cross-wires R) are placed on the ground underneath and parallel to the positive wires L. Both ends of each of the wires Q are sunk into the earth or submerged in water, and fastened to a metallic plate coated with a metal not subject to oxidation. These wires Q are the conductors for the negative electricity of the earth, and a branch wire or wires attached to the wires Q serve to transmit the negative terrestrial electricity to wherever it is wanted for industrial or other purposes.

By uniting to the ends of the positive and the negative branch wire or wires a powerful electric current will be obtained, one pole of which is the atmosphere and the other the earth, and may be applied to any suitable useful purpose.

I will now proceed to describe the modification of the above-described apparatus when to be applied in mountainous countries.

P represents the positive electric copper or other metal wires coated over with an insulating substance. The upper ends of each of the positive wires is soldered to a prompter, O, at P′, Figs. 1, 5, and 6, Plate 6. The lower portion of each of the positive-wires is secured to an insulator, T, Fig. 2, Sheet 6. The positive wires are held above the ground by joints A, Figs. 3 and 4, Sheet 6, projecting from the soil at suitable distances from each other. The wires P are intended to follow the inequalities of the ground on which they are laid.

The prompters O, Figs. 5 and 6, Sheet 6, are iron rods sharpened to a point and silvered or coppered at their upper ends. The lower part of the prompter is fastened into a pole, S, covered with tar, which isolates the prompter and holds it in a firm position. A large metallic plate may be soldered to each prompter, as shown in Fig. 6. The positive wires may be soldered to the rod of each prompter or to the plate which is fastened thereto.

One or more branch lines, E, are soldered up to the positive wires to transmit the positive atmospheric electricity for which the wires P are the conductors to any desirable point.

N are negative iron or other metal conductors coated with an insulating substance. The upper ends of these wires rest on the ground near the positive insulators. The lower ends of these wires are soldered to a metallic plate or plates, V, Fig. 7, Plate 6, coated with a metal not subject to oxidation. The negative wires are sunk into the ground at very great depth or into wells, rivers, or into the sea. The negative electric branch wires are attached to the negative conductors N in the same manner as the positive branch wires are to the positive conductors. The branch wires and the soldering are coated over with an insulating substance. They are intended to carry the negative terrestial electricity to any desired point.

What I claim as my invention, and desire to secure by Letters Patent, is—

1. The peculiar arrangement of means herein specified, whereby I am enabled to use the positive electricity contained in the atmosphere and the negative electricity contained in the earth, and thus form an electric pile of considerable power and make the electricity therein contained available for industrial purposes, as set forth.

2. The combination of an aerostat and vertical wire-work with a tube, I, for admitting gas into the aerostat, in the manner and for the purposes above set forth.

CHARLES VION.

Witnesses:
HARRY W. SPENCER,
A. GUION, Jr.

rexresearch.com

Roy J. MEYERS

Absorber

How To Patent Your Idea
3 Easy Steps To Patent & Make Money Get A Free Patent & Invention Kit
www.Patent.InventionHome.com

Superior X-Ray Solutions
Performance, Reliability & Speed from the world leader in SCD
www.Bruker-AXS.com

Do You Have an Idea?
We Can Help Turn Your Idea Into A Real Product. Get Free Information!
NewProductHelp.com

Buckyballs Magnetic Toy
Get the Gadget on Everyone's List. Best prices on this year's hot toy!
www.getbuckyballs.com

Car Door Magnets: 50% Off
Magnetic Car Signs $4.99. No Mins. Full-Color. Upload Photos & Logos!
www.Vistaprint.com

Power Your Own Home
Reduce Utility Bills by 50% Lower C02 Emissions by 1/3
www.ClearEdgePower.com

Ads by Google

Renewable Clean Energy
Compact Electricity & Heat Unit For Homes & Small Businesses
www.ClearEdgePower.com

Do You Have an Idea?
We Can Help Turn Your Idea Into A Real Product. Get Free Information!
NewProductHelp.com

Patent Attorney/Engineer
Over 36 years experience. Staff available 24/7 by phone.
www.invention.net

C. Edholm: "Picks Power from Air" (*Technology World Magazine*, 1912 ?)
Dr L. Hirshberg: "Electricity from Air New Great Discovery (*Modern Electrics*, 1914 ?)
Roy J. Meyers: British Patent # 1098 (1913) -- Improvements in and Relating to Apparatus for Producing Electricity (PDF Format, requires Adobe Reader)
R. J. Meyers: British Patent # 1098 (1913) --- Transcription w/ enlarged figures
Notes & Comments
R. J. Meyers: British Patent # 1098 (1913) --- JPG version

Technology World Magazine, p. 279-281 (Year unknown, apparently circa 1912; another article about Meyers appeared in the November 1912 issue of _Electronic World_)

"Picks Power from the Air"

by
Charlton Lawrence Edholm

A remarkable scene took place in the legislature of Arizona this spring when the lawmakers enthusiastically voted for the parole of a certain convict in the State penitentiary, granting him a leave of absence for 30 days and by means of private contributions raising a fund to defray his expenses to Washington DC and return.

The prisoner, Roy J. Meyers, is serving a 3-1/2 year sentence, but in spite of the fact that he bears the stigma of a convicted lawbreaker, he has demonstrated that a convict can be a useful member of society. During his imprisonment he perfected an electrical device of such original character as to arouse feelings of wonder and skepticism until experts had seen it in actual operation. It is a device to draw electricity from the atmosphere for light and power, and the 30-day parole was granted in order that the inventor might protect his rights through the patent office at Washington.

With the acquiescence of the legislature, Governor Hunt granted the parole and the prisoner was allowed to go free without any guard or any assurance but his word of honor that he would return. Two days before the period had elapsed, Meyer again presented himself before the governor, having accomplished his mission, and then returned to the penitentiary at Florence, where he continues to serve his sentence.

This, in brief, is the picturesque story which has called attention of the civilized world to a newly discovered electrical genius, and to another feature of the case which is of equal importance and human interest; namely, the enlightened policy pursued by our youngest State in its treatment of convicts...

Before entering the prison, Meyer had already applied for various patents, among them one for an improved trolley wheel head which prevents the trolley wheel from jumping the wire. Meyers had a conference with Superintendent Sims and Parole Clerk Sanders, and it was to these gentlemen that the inventor first explained the principles of his new device for securing electrical energy from the air. The officials were willing to give the man the opportunity to develop his plan and a little wooden building outside the walls was turned over to Meyers and was fitted up as a workshop and a laboratory. The first demonstration of the new apparatus was made shortly thereafter, the electricity drawn from the atmosphere being used to spark the gas engines of the pump house, and although the device was crude yet it did the work, and removed the doubts of his friends. Furtter development of the "absorber" followed, and his second model was constructed, and developed 8 volts. The machine came to the attention of the remarkable woman who brought his name before the legislature.

This was Miss Kate Barnard, State Commissioner of Charities and Corrections of Oklahoma, who was a guest of Mr. Sims, while studying prison conditions. She saw the machine at work, became familiar with the facts of Meyers' case, and was impressed by his rather blunt and unaffected personality, for Meyers has nothing of the polish or glibness of the poseur. He is a simple, earnest student of mechanical problems and not the sort of man to make a sentimental appeal for sympathy because of any grace of person or manner. Therefore it was the value of Meyers' invention, together with his essential integrity (in spite of his lapse) which so strongly impressed Miss Barnard that when she appeared before the Arizona legislature not long afterwards, addressing that body on the need of enlightened legislature along the line of her own work, she told the story of Roy Meyers and his epoch-making invention.

So, early in May, Meyers set out for Washington, unaccompanied.

In his own words: "When I arrived in Washington and laid my plans before the patent office experts, they merely smiled and told me that I would have to build a model and demonstrate my claims --- that it seemed strange that I, unknown as I am in the electrical world, should have accomplished the things for which Edison, Tesla and other experts have been striving for years.

"They could grasp the meaning of my drawings nor the explanation I tried to make to them. There was little time to spare, as I had only 20 days left of my leave, but I set to work in a few days was able to take a crude model around to the patent office to make a demonstration.

"Arriving at the patent office I telephoned to a friend who had been so kind as to introduce me and aid me in reaching the proper officials. The absorber was hoisted on two short poles and made to work. While they were as yet unable to understand the principles involved and hardly willing to believe their eyes, they were forced to admit that I had something new and different, and they told me that there would be no further objection; that I might file my application without further delay.

"I hope to construct my first large machine right here in Phoenix. I feel grateful; to Governor Hunt and others for what they have done for me and to the help they have given in securing protection I might not otherwise have had, and I am desirous of demonstrating this gratitude. I am going back to Florence today to resume the serving of my sentence, which will expire in 10 months. Then, here in Phoenix, I will begin the work of making my machines."

While there are some details of the device which the inventor refuses to make public, yet there are many general features that may be explained. It is planned that the machine, to be set up in Phoenix, will generate sufficient power to light the city, and will consist of a 200 foot tower upon which is placed the "absorber". The latter consists of a series of magnetized steel plates set in a circle (the manner of preparing them is kept secret) and this mechanism attracts the electricity from the atmosphere. This is carried by wires to a transformer in the engine house below and thence is applied to produce either power or light after the usual manner.

In an authorized statement Meyers says: "The flow of electricity is constant. When it emerges into the transformer it is in the form of a direct current. It will absorb the electricity day and night and will work whenever the wireless will work. I can put up a plant to supply such a building as the Adams Hotel for about $1500, and one of the principal items of the expense is the cost of the towers, the wires, the magnetizing of one set of plates, which is part of the secret of the treatment which makes it respond to the accumulations of the atmosphere.

"For use in the case of an electrical storm I have made what I call a modified form of circuit breaker, such as is commonly used as a lightning arrester on telegraph lines. In case of a storm the accumulator would suddenly become overcharged, possibly, and as the electricity would not of itself flow back into the air, the result might be disastrous. So I send it down into the ground, whenever the voltage rises above a certain amount."

Roy J. Meyers & the "Absorber"

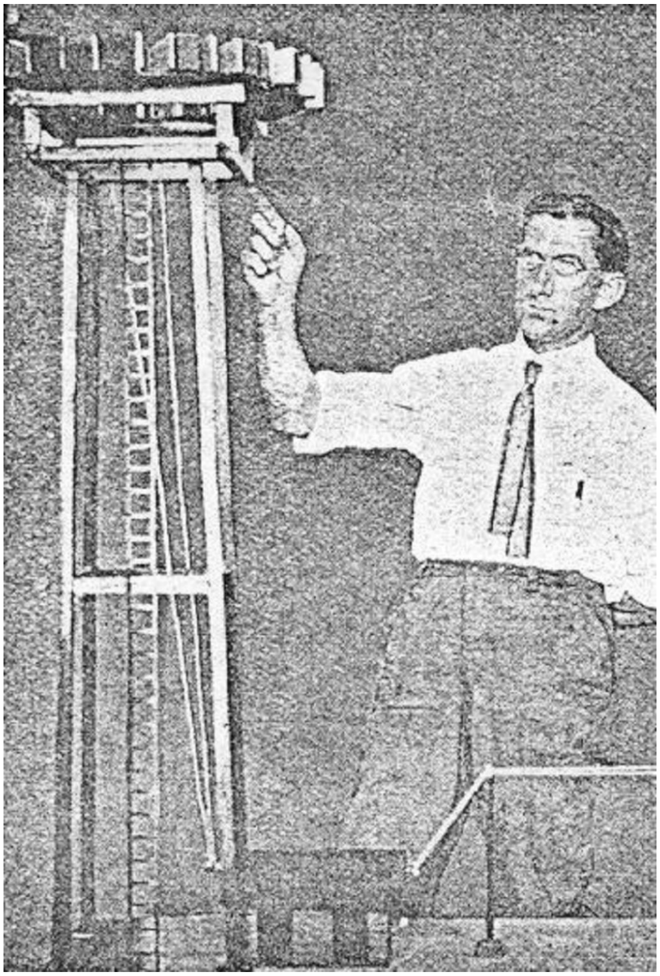

Modern Electronics **(1914 ?)**

Electricity From Air New Great Discovery

by

Dr Leonard Keene Hirshberg

Working quietly in the heart of Baltimore for weeks on an invention which some critics say will revolutionize the method of converting electricity to practical use has been Roy J. Meyers, who like Benjamin Franklin, extracts the electric current from the air.

Mr Meyers invention was made last summer while he was confined in the penitentiary at Florence, Arizona. His first finished apparatus was made in Baltimore.

A practical, unlettered electrician, Mr Meyers, while in Arizona, was arrested on a comparatively minor charge and sent to the penitentiary. There he was placed in charge of the prison electrical plant, and there he says he made his discovery that the current which the civilized world is beginning to use most extensively for light and power could be transformed from the atmosphere without the aid of moving machines or batteries.

Miss Kate Barnard, Commissioner of Charities and Corrections, of Oklahoma, hearing of Meyers' invention and of his desire to have it patented, appeared before the Arizona Legislature to make an appeal in behalf of the young convict. As a result a special bill was passed which granted Meyers a month's leave of absence on parole. He went unaccompanied to Washington, filed his patent applications and returned to the penitentiary. Since then he has been indefinitely paroled.

He came to Baltimore as the place where he could easily obtain the mechanical parts needed to make a more nearly perfect machine than the crude model he has fashioned in the penitentiary workshop, and is making his headquarters here while working on his invention. With him is W.E. Chenot, who has been his assistant in assembling and testing the machine and who says that he has bought Meyers' patent rights for Germany.

They have proved beyond doubt that the invention is practical and that when finally brought to a state of perfection it will introduce a new epoch in the industrial use of electricity. By Westinghouse meters they tested the strength of the current gathered from the air, and with the use of only two of the four rectifying transformers the voltmeter recorded four and one-half volts, and the ammeter, which had the capacity of recording 75 amperes, was broken by the force of the current.

The machine itself is simple. It is in reality a transformer, which is familiar to anyone knowing anything at all about electricity in its practical uses. On a high tripod, which resembles somewhat the framework of a windmill tower, is the transformer, which Mr Meyers calls his 'absorber'. It is made up of an iron core, wrapped with copper wire. The secret of the invention is the manner in which the disks composing this 'absorber' are magnetized, and this secret Meyers says he found by accident while at work in prison.

What the machine, when finally perfected, will do is yet to be seen. Its inventor claims that it will greatly reduce the cost of making electricity. No batteries of any kind are needed, he says, and not a part of the machine turns upon the other. It is as durable, apparently, as an electric light pole. One of these machines, says Meyers, when perfected may be placed on a vehicle and transform enough electricity to give motive power, be that vehicle a locomotive or an automobile. He declared it can be placed on a building to furnish electric lights or power, and that the only wear will be upon the machinery which its current runs.

Meyers is 34 years old and he gained his knowledge of electricity by working in shops along the Pacific Coast. The depths of the mysteries of electricity he has not explored, but he is certain that he has found the means of absorbing it from the air and of converting it to the use of mankind.

British (GB) Patent # 191301098

Improvements in and Relating to Apparatus for Producing Electricity.

1-14-1914

Roy Jerome Meyers

Classification: - international: H05F7/00; H05F7/00; - european: H05F7/00
Application number: GBD191301098 19130114
Priority number(s): GBT191301098 19130114

Abstract ~ Vapour apparatus, arrangements of. - A rectifier for use with apparatus for producing electricity from the earth consists of mercury- vapour lamps constructed and arranged as shown in Fig. 4. Each lamp comprises two wires 6<1>, 7<1> wound around a steel tube 15 surrounding a mercury tube 11 preferably of copper. The coil 6<1> is connected between the electrode 14 and the terminal 18, and the coil 7<1> between the terminals 19, 5. The coils 6<1>, 7<1> are preferably composed of soft iron. Reference has been directed by the Comp- troller to Specifications 16,709/87, 14,033/99, and 5457/11, [all in Class 53, Galvanic batteries], and 15,412/06.

British Patent # 1098

(January 14, 1913)

Improvements in and Relating to Apparatus for Producing Electricity

Roy Jerome Meyers

This invention relates to improvements in apparatus for the production of electrical currents, and the primary object in view is the production of a commercially serviceable electrical current without the employment of mechanical or chemical action. To this end the invention comprises means for producing what I believe to be dynamic electricity from the earth and its ambient elements.

I am, of course aware that it has been proposed to obtain static charges from upper strata of the atmosphere, but such charges are recognized as of widely variant potential and have thus far proved of no practical commercial value, and the present invention is distinguished from all such apparatus as has heretofore been employed for attracting static charges by the fact that this improved apparatus is not designed or employed to produce or generate irregular, fluctuating or other electrical charges which lack constancy, but on the other hand I have by actual test been able to produce from a very small apparatus at comparatively low elevation, say about 50 or 60 feet above the earth's surface, a substantially constant current at a commercially usable voltage and amperage. This current I ascertained by repeated tests is capable of being readily increased by additions of the unit elements in the apparatus hereinafter set forth, and I am convinced from the constancy of the current obtained and its comparatively low potential that the current is dynamic and not static, although, of course, it is not impossible that certain static discharges occur and, in fact, I have found occasion to provide against the damage which might result from such discharge by the provision of lightning arresters and cut-out apparatus which assist in rendering the obtained current stable by eliminating sudden fluctuations which sometimes occur during conditions of high humidity from what I consider static discharges. The nature of my invention is obviously such that I have been unable to establish authoritatively all of the principles involved, and some of the theories herein expressed may possibly prove erroneous, but I do know and am able to demonstrate that the apparatus which I have discovered does produce, generate, or otherwise acquire a difference of potential representing a current amperage above stated, or varied therefrom at the will of the operator according to the uses which the current is to be subjected.

The invention comprises generically means for producing electrical currents of serviceable potential substantially without the employment of mechanical or chemical action, and in this connection I have been able to observe no chemical action whatever on the parts utilized although deterioration may possibly occur in some of the parts, but so far as I am able to determine such deterioration does not add to the current supply but is merely incidental to the effect of climatic action.

The invention more specifically comprises the employment of a magnet or magnets and a co-operating element, such as zinc disposed adjacent to the magnet or magnets and connected in such manner and arranged relative to the earth so as to produce current, my observation being that current is produced only when such magnets have their poles facing substantially to the north and south and the zincs are disposed substantially along the magnets.

The invention also comprehends other details of construction, combinations and arrangements of parts as will hereinafter be fully set forth and claimed.

In the accompanying drawings:

Figure 1 is a top plan view of an apparatus embodying the features of the present invention, the arrow accompanying the figure indicating substantially the geographical north, parts of the figure being diagrammatic for condensing the showing.

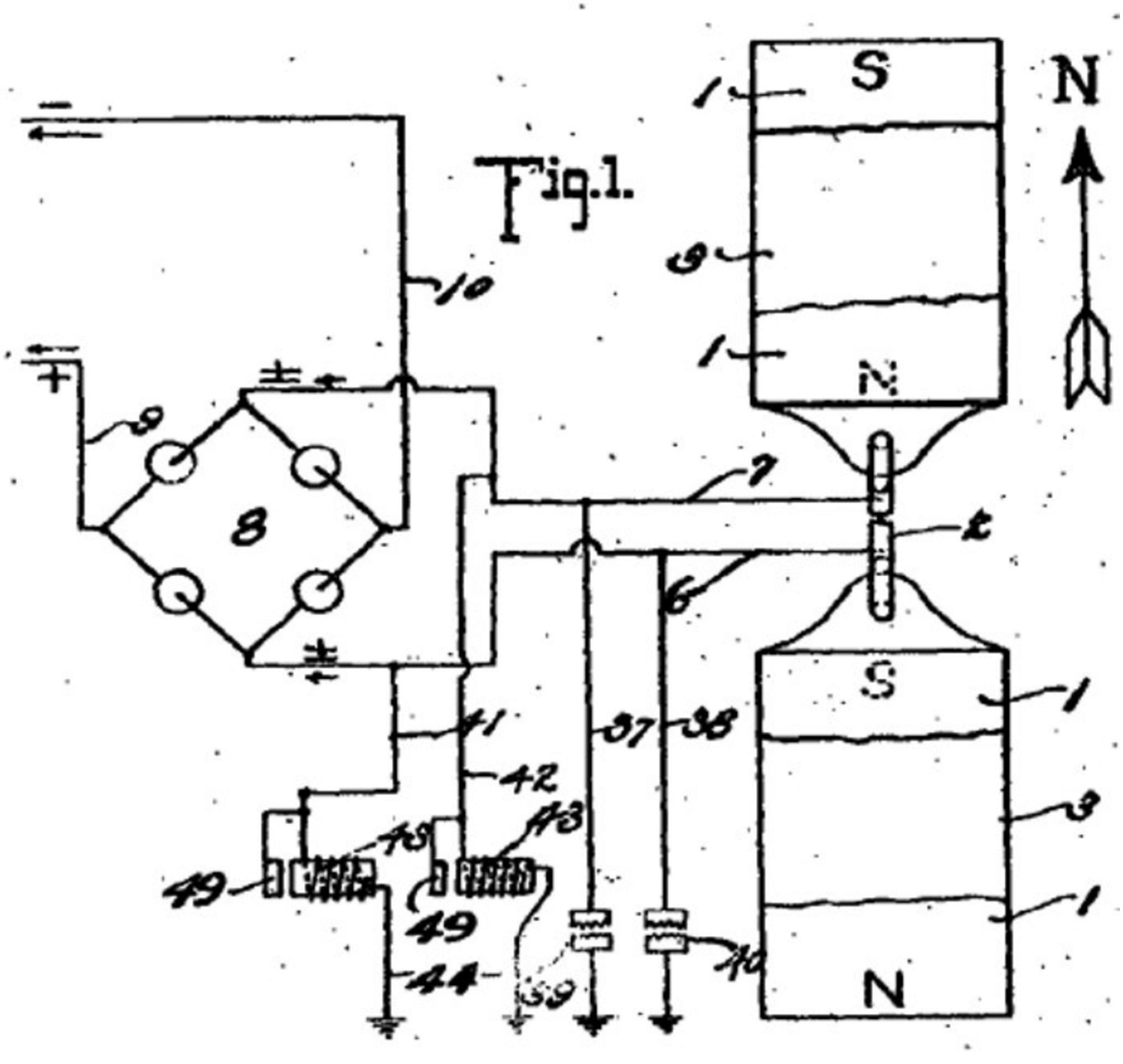
Fig. 1.

Figure 2 is a view is side elevation of the parts seen in plan in Figure 1.

Fig. 2.

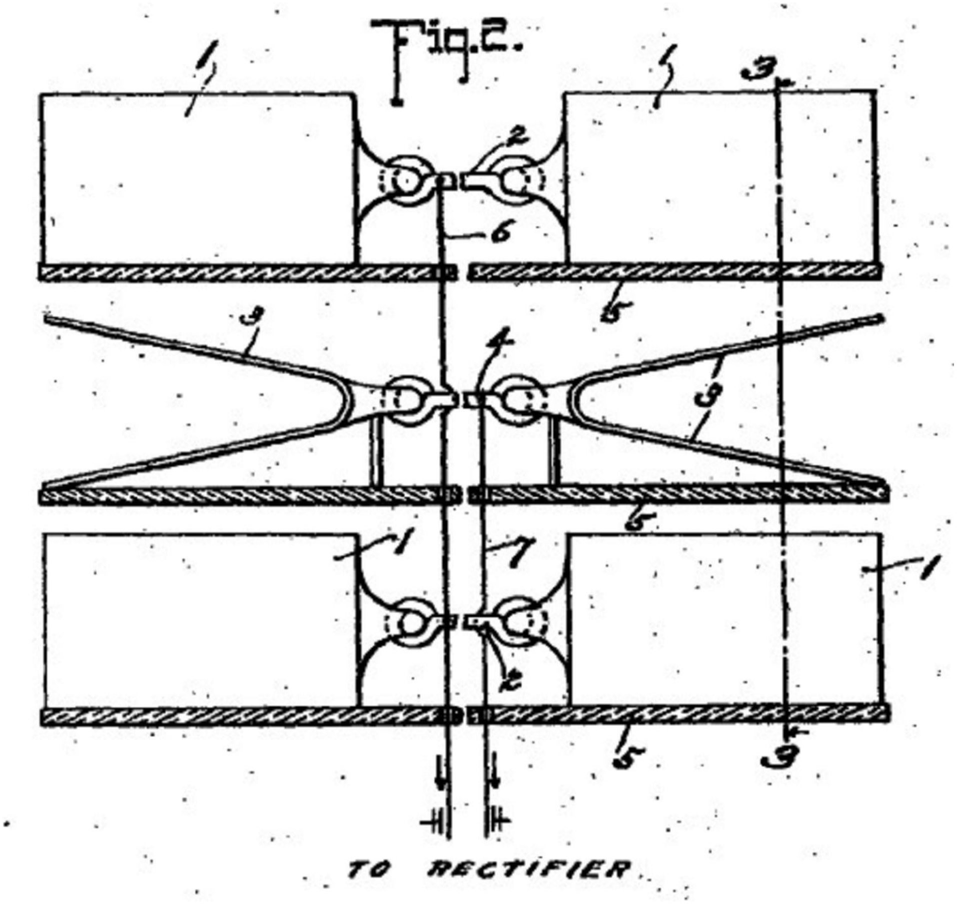
1
1
3
2
6
5
9
4
9
6
1
7
1
2
5
3
TO RECTIFIER

Figure 3 is a vertical section taken on the plane indicated by the line 3-3 of Figure 2 and looking in the direction indicated by the arrow.

Fig. 5.

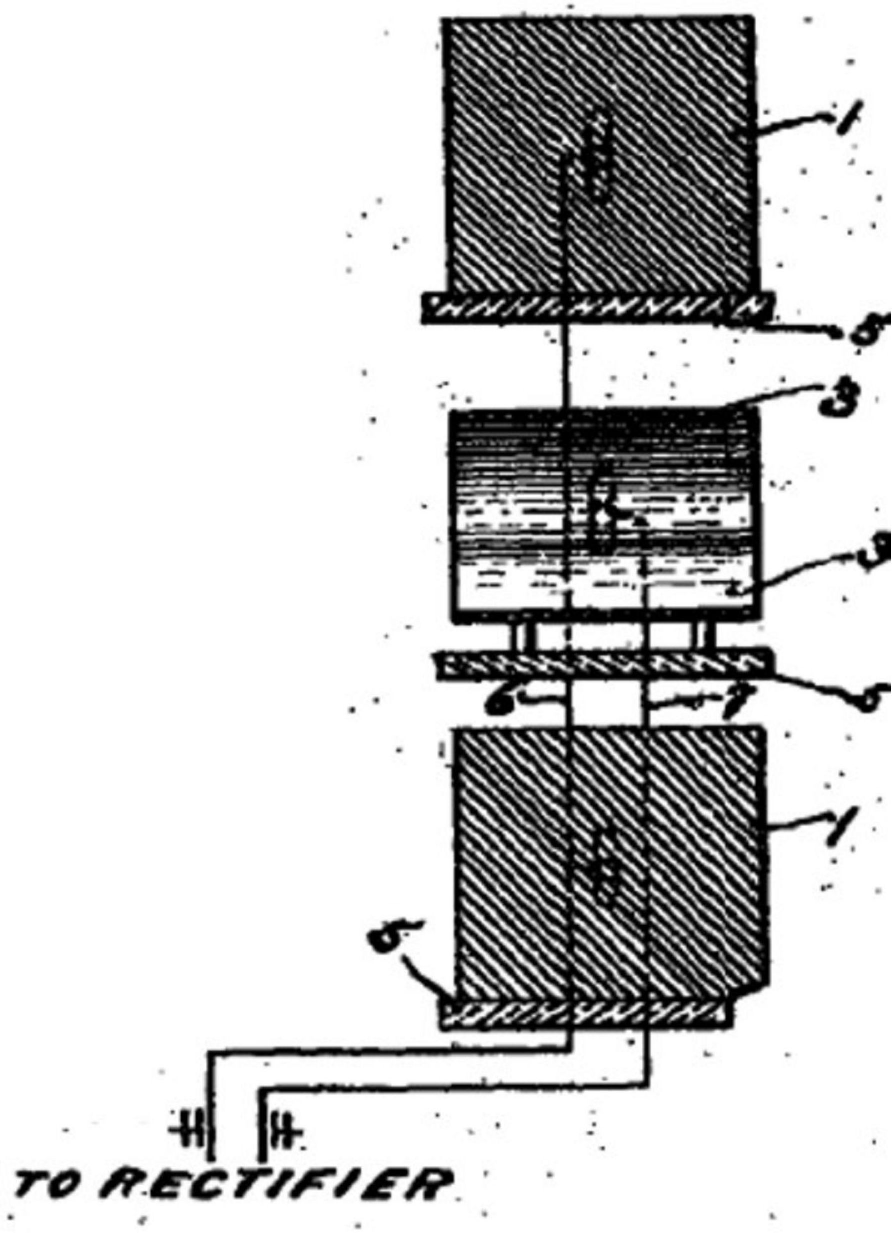

Figure 4 is a detail view partly in elevation and partly in section showing the detail connections of the converter and intensifier.

Fig4

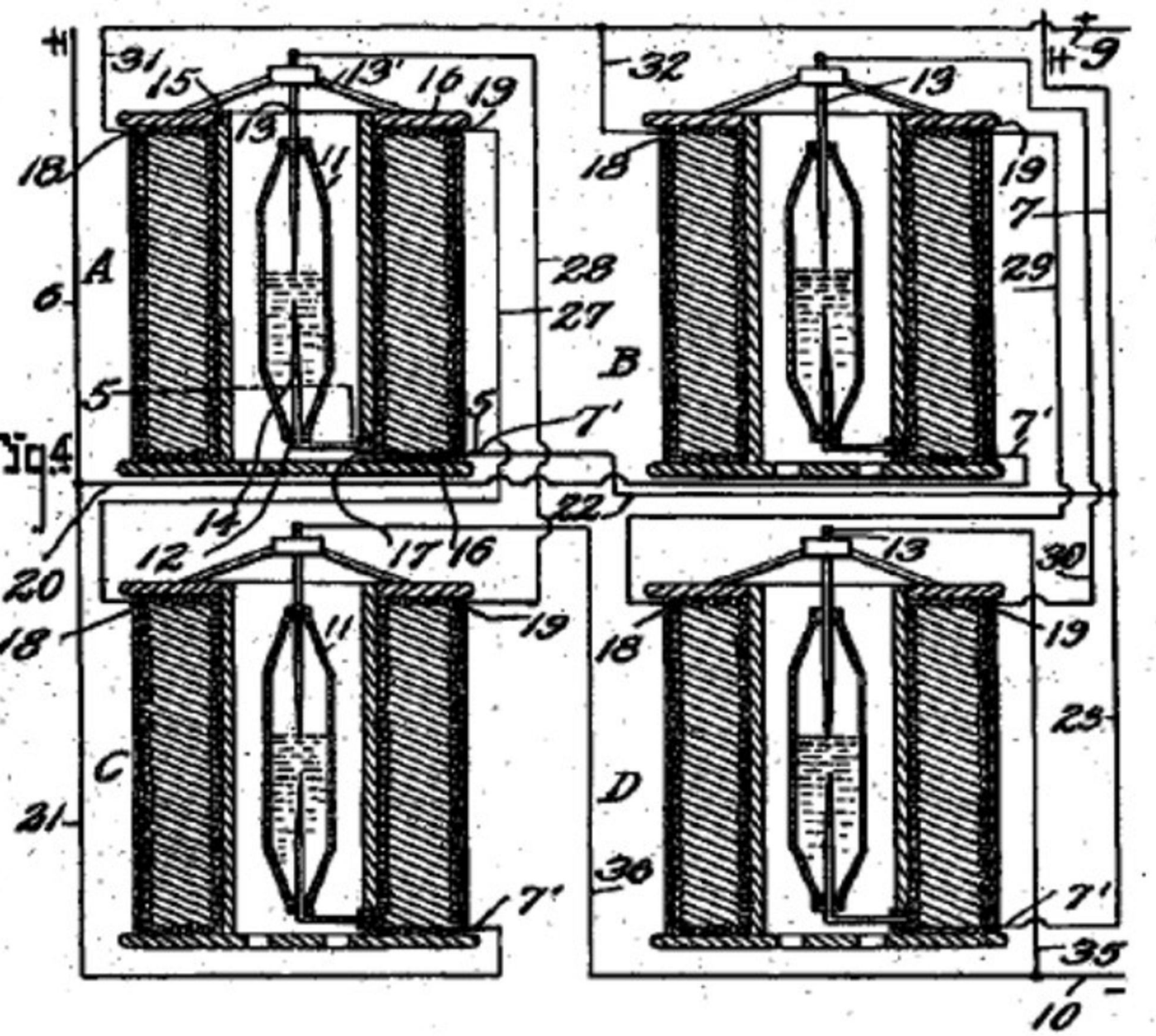
+
31
15
13'
16
19
13
11
18
6
A
5
5
28
27
B
7'
14
20
22
12
17
16
32
13
9
19
7
29
7'
13
30
18
19
23
C
21
D
36
7'
7'
35
10
−

Figure 5 is a transverse section taken on the planes indicated by line 5-5 of Figure 4 and looking downwardly.

Fig5.

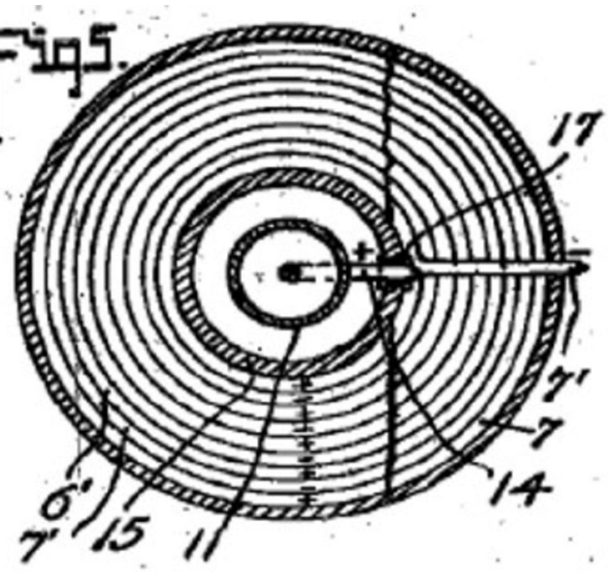
17
+
−
7'
7
14
6'
7'
15
11

Figure 6 is an enlarged detail fragmentary section illustrating the parts at the juncture of the conductors and one of the intensifiers.

Fig.6

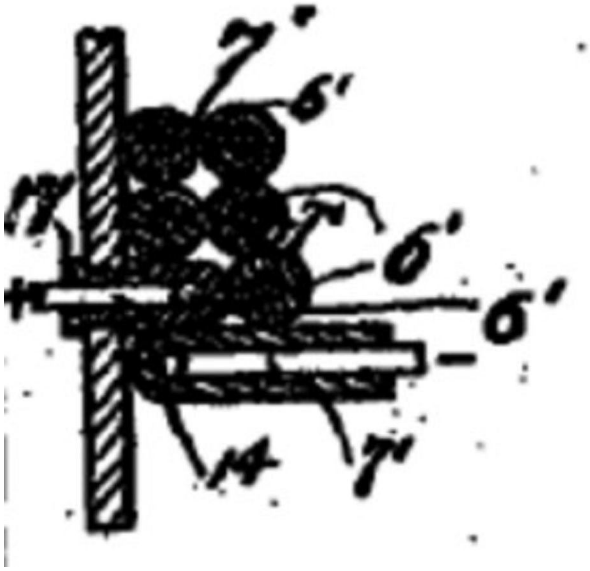

Figure 7 is an enlarged detail view partly in elevation and partly in section of one of the automatic cut-outs and

Fig.7.

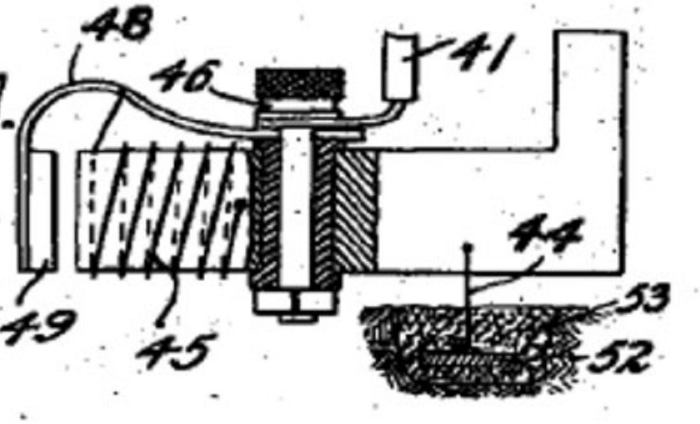

Figure 8 is a diagrammatic view of one of the simplest forms of embodiment of the invention.

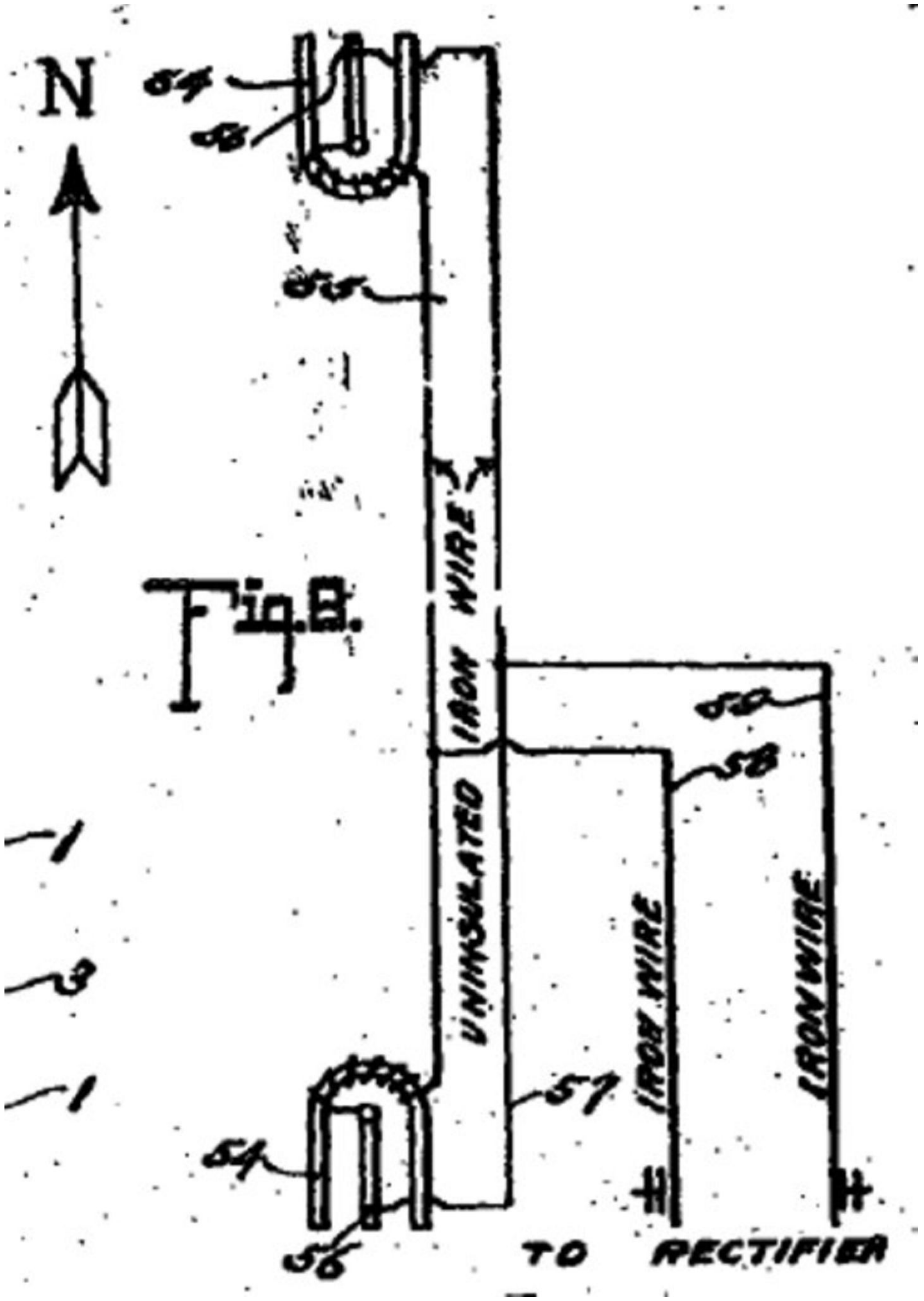

Referring to the drawing by numerals, 1,1 indicates magnets connected by a magnetic substance 2, preferably an iron wire. The magnets 1 are arranged in pairs, one pair being spaced beneath the other, and interposed between the magnets are zinc plates 3,3 connected by an iron wire conductor 4. Suitable insulating supports 5 are arranged for sustaining the respective magnets 1 and plates 3,3. Each plate 3 is preferably bent substantially into V form, as clearly seen in Figure 1, and the V1 of one of the plates opens or faces toward the north and the V of the other plate to the South. I have determined by experimentation that it is essential that the plates 3 be disposed substantially north and south with their flat faces approximately parallel to the adjacent faces of the co-operating magnets, although by experience I have not discovered any material difference in the current obtained when the plates are disposed slightly to one side of north and south, as for instance when the plates are disposed slightly to one side of north and south, as for instance when disposed in the line of the magnetic polarity of the earth. The same is true with respect to the magnets 1, the said magnets being disposed substantially north and south for operative purposes, although I find that it is immaterial whether the north pole of one of the magnets is disposed to the north and the south pole to the south, or vice versa, and it is my conviction from experience that it is essential to have the magnets of each pair connected by magnetic material so that the magnets substantially become one with a pole exposed to the north and a pole exposed to the north. In Figure 1, I have indicated in full lines by the letters 8 and N the respective polarities of the magnets 1, and have indicated in dotted lines the other pole of those magnets when the connection 2 is severed. I have found that the magnets and zinc plates operate to produce, whether by collection or generation I am not certain, electrical currents when disposed substantially north and south, but when disposed substantially east and west no such currents are produced. I also find that the question of elevation is by no means vital, but it is true that more efficient results are obtained by placing the zincs and magnets on elevated supports. I furthermore find from tests that it is possible to obtain currents from the apparatus with the zincs and magnets disposed in a building or otherwise enclosed, although more efficient results are obtained by having the said elements arranged in the open.

While in Figures 1, 2, and 3, I have shown the magnets and the zinc plates as superimposed, it will be apparent, as hereinafter fully set forth, that these elements may be juxtaposed in horizontal planes, and substantially the same results will be secured. Furthermore, the magnets 1 with the interposed zincs 3, as shown in Figures 1, 2 and 3 merely represent a unit which may be repeated either horizontally or vertically for increasing the current supply, and when the unit is repeated the zinc plates are arranged alternating with the magnets throughout the entire series as hereinafter indicated.

A conductor 6 is connected in multiple with the conductors 2 and a conductor 7 is connected with conductor 4, the conductor 6 extending to one terminal of a rectifier which I have indicated by the general reference character 8, and the conductor 7 extending to the other terminal of said rectifier. The rectifier as seen in diagram in Figure 1 may assume any of several well known embodiments of the electrical valve type and may consist of four asymmetric cells or Cooper-Hewitt mercury vapor lamps connected as indicated in Figure 1 for permitting communication of the positive impulses from the conductor 6 only to the line conductor 9 and the negative impulses from conductor 6 on only to the line conductor 10. The current from this rectifier may be delivered through the conductors 9 and 10 to any suitable source for consumption.

While the said rectifier 8 may consist of any of the known types, as above outlined, it preferably consists of a specially constructed rectifier which also has the capacity of intensifying the current and comprises specifically the elements shown in detail in Figures 4, 5, and 6 wherein I have disclosed the detail wiring of the rectifier when composed of four of the rectifying and intensify in elements instead of asymmetric cells or simple mercury vapor valves. As each of these structures is an exact embodiment of all the others, one only will be described, and the description will apply to all. The rectifying element of each construction consists of a mercury tube 11 which is preferably formed of glass or other suitable material, and comprises a cylinder having its end portions tapered and each terminating in an insulating plug or stopper 12. Through the upper stopper 12 is extended the electrode 13 which extends well into the tube and preferably substantially one-half the length thereof to a point adjacent the inner end of an opposing electrode 14 which latter electrode extends thence downwardly through the insulation 12 at the lower end of the tube. The tube 11 is supplied with mercury and is adapted to operate on the principle of the mercury vapor lamp, serving to rectify current by checking back impulses of one sign and permitting passage of impulses of the other. To avoid the necessity for utilizing a starter, as is common with the lamp type of electrical valve, the supply of mercury within the tube may be sufficient to contact with the lower end of the electrode 13 when current is not being supplied, so that as soon as current is passed from one electrode to the other sufficiently for volatilizing that portion of the mercury immediately adjacent the lower end of electrode 13, the structure begins its operation as a rectifier. The tube 11 is surrounded by a tube 15 which is preferably spaced from tube 11 sufficiently for allowing atmospheric or other cooling circulation to pass the tube 11. In some instances, it may be desirable to cool the tube 11 by a surrounding body of liquid, as hereinafter indicated. The tube 15 may be of insulating material but I find efficient results attained by the employment of a steel tube, and fixed to the ends of the of the tube are insulating disks 16, 16 forming a spool on which are wound twin wires 6^1 and 7^1, the wire 6^1 being connected at the inner helix of the coil with the outer end of the electrode 14, the lower portion of said electrode being extended to one side of the tube 11 and passed through an insulating sleeve 17 extending through the tube 15, and at its outer end merging into the adjacent end of the wire 6^1. The wire 7^1 extends directly from the outer portion of the spool through the several helices to a point adjacent the juncture of the electrode 14 with wire 6^1 and thence extends in mechanical parallelism with the wire throughout the coil, the wire 6^1 ending in a terminal 18 and the wire 7^1 ending in a terminal 19. For the sake of convenience of description and of tracing the circuits, each of the apparatus just above described and herein known as an intensifier and rectifier will be mentioned as A, B, C and D, respectively. Conductor 6 is formed with branches 20 and 21 and conductor 7 is formed with similar branches 22 and 23. Branch 20 from conductor 6 connects with conductor 7^1 of intensifier B and branch 21 of conductor 6 connects with the conductor 7^1 of intensifier C, while branch 22 of conductor 7 of intensifier C, while branch 22 of conductor 7 connects with conductor 7^1 of intensifier D. A conductor 27 is connected with terminal 19 of intensifier A and extends to and is connected with the terminal 18 of intensifier C, and a conductor 7 connects with conductor 7^1 of intensifier D. A conductor 27 is connected with terminal 19 of intensifier A. and extends to and is connected with terminal 18 of intensifier C, and a conductor 28 is connected with the terminal 19 of intensifier C and extends from the terminal 19 of intensifier B to the terminal 18 of intensifier D to electrode 13 of intensifier B. Each electrode 13 is supported on a spider 13^1 resting on the upper disk 16 of the respective intensifier. Conductors 31 and 32 are connected with the terminals 18 of intensifiers A and B and are united to form the positive line wire 9 which co-operates with the negative line wire 10 and extends to any suitable point of consumption. The line wire 10 is provided with branches 35 and 36 extending to the electrodes 13 of intensifiers C and D for completing the negative side of the circuit.

Thus it will be seen that alternating currents produced in the wires 6 and 7 will be rectified and delivered in the form of a direct current through the line wires 9 and 10, and I find by experiment that the wires 6 and 7 should be of iron, preferably soft, and may of course be insulated, the other wiring not specified as iron being of copper or other suitable material.

In carrying out the operation as stated, the circuits may be traced as follows: A positive impulse starting at the zincs 3 is directed along conductor 7 to branch 23 to conductor 7^1 and the winding of the rectifier of intensifier B through said rectifier to the conductor 6^1, through the winding thereof to the contact 18, conductor 32 and to the line wire 9. The next or negative impulse directed along conductor 7 cannot find its way along branch 23 and the circuit just above traced because it cannot pass across the rectifier of intensifier B but instead the negative impulse passes along conductor 22 to conductor 7 of intensifier A and the winding thereof to the contact 19 and to conductor 27 to contact 18 of intensifier C, to the winding of the wire 6^1 thereof to the electrode 14 through the rectifier to the of the electrode 13 and conductor of intensifier A, electrode 14 thereof and conductor 6^1 to contact 18 and wire 31 to line wire 9. Obviously the positive impulse cannot pass along the wire 20 because of its inverse approach to the rectifier of intensifier B. The next impulse or negative impulse delivered to conductor 6 cannot pass along conductor 21 because of its connection with electrode 13 of the rectifier of intensifier A, but instead passes along conductor 20 to the wire 7^1 and its winding forming part of intensifier B to the contact 19 and conductor 29 to contact 18 and the winding of wire 6^1 of intensifier D to the electrode 14 and through the rectifier to the electrode 13 and conductor 35 to line wire 10. Thus the current is rectified and all positive impulses directed along one line and all negative impulses along the other lie s that the potential difference between the two lines will be maximum for the given current of the alternating circuit. It is, of course, apparent that a less number of intensifiers with their accompanying rectifier elements may be employed with a sacrifice of the impulses which are checked back from a lack of ability to pass the respective rectifier elements, and in fact I have secured efficient results by the use of a single intensifier with its rectifier elements, as hereinafter set forth.

Grounding conductors 37 and 38 are connected respectively with the conductors 6 and 7 and are provided with the ordinary lightning arresters 39 and 40 respectively for protecting the circuit against high tension static charges.

Conductors 41 and 42 are connected respectively with the conductors 6 and 7 and each connects with an automatic cutout 43 which is grounded as at 4. Each of said automatic cutouts is exactly like the other and one of the same is shown in detail in Figure 7 and comprises the inductive resistance 45 provided with an insulated binding post 46 wit which the respective conductor 6 or 7 is connected, said post also supporting a spring 48 which sustains an armature 49 adjacent to the core of the resistance 45. The helix of resistance 45 is connected preferably through the spring to the binding post at one end and at the other end is grounded on the core of the resistance, the said core being grounded by ground conductor 44 which extends to the metallic plate 52 embedded in moist carbon or other inductive material buried in the earth. Each of the conductors 41, 42 and 44 is of iron, and in this connection I wish it understood that where I state the specific substance I am able to verify the accuracy of the statement by the results of tests which I have made, but of course I wish to include along with such substances al equivalents, as for instance, where iron is mentioned its byproducts, such as steel, and its equivalents such as nickel and other magnetic substances are intended to be comprehended. The cutout apparatus seen in detail in Figure 7 is employed particularly for insuring against high tension currents, it being obvious from the structure shown that when potential rises beyond the limit established by the tension of the spring sustaining the armature 40, the armature will be moved to a position contacting with the core of the cutout device and thereby directly close the ground connection for line wire 41 with conductor 44, eliminating the resistance of winding 45 and allowing the high tension current to be discharged to the ground. Immediately upon such discharge the winding 45 losing its current will allow the core to become demagnetized and release the armature 49 whereby the ground connection is substantially broken leaving only the connection through the winding 45 the resistance of which is sufficient for insuring against loss of low tension current.

In Figure 8 I have illustrated an apparatus which though apparently primitive in construction and arrangement comprehends the first successful embodiment which I produced in the course of discovery of the present invention, and it will be observed that the essential features of the invention are therein disclosed. The structure delineated in said figure consists of horseshoe magnets 54, 55, one facing north and the other south, that is, each opening in the respective directions indicated and the two being connected by an iron wire 55 which is uninsulated and wrapped about the respective magnets each end portion of the wire 55 being extended from the respective magnets to and connected with, as by being soldered to, a zinc plate 56, there being a plate 56 for each magnet and each plate being arranged longitudinally substantially parallel with the legs of the magnet and with the faces of the plate exposed toward the respective legs of the magnet, the plate being thus arranged endwise toward the north and south. An iron wire 57 connects the plates 56, the ends of the wire being preferably connected adjacent the outer ends of the plates but from experiment I find that the wire may be connected at practically any point to the plate. Lead wires 58 and 59 are connected respectively with the wires 55 and 57 and supply an alternating current at a comparatively low tension, and to control such current the wires 58 and 59 may be extended to a rectifier or combined rectifier and intensifier, as above set forth.

The tests which I have found successful with the apparatus seen in Figure 8 were carried out by the employment first of horseshoe magnets approximately 4 inches in length, the bar comprising the horseshoe being about one inch square, the zincs being dimensioned proportionately and from this apparatus with the employment of a single intensifier and rectifier, as above stated, I was able to obtain a constant current of 8 volts.

It should be obvious that the magnets forming one of the electrodes of this apparatus may be permanent or may be electromagnets, or a combination of the two.

While the magnets mentioned throughout the above may be formed of any magnetic substance, I find the best results obtained by the employment of the nickel chrome steel.

While the successful operation of the various devices which I have constructed embodying the present invention have not enabled me to arrive definitely and positively at fixed conclusion relative to the principles and theories of operation and the source from which current is supplied, I wish it to be understood that I consider myself as the first inventor of the general type hereinbefore described capable of producing commercially serviceable electricity, for which reason my claims hereinafter appended contemplate that I may utilize a wide range of equivalents so far as concerns details of construction suggested as preferably employed.

The current which I am able to obtain is dynamic in the sense that it is not static and its production is accomplished without chemical or mechanical action either incident to the actual chemical or mechanical motion or incident to changing caloric conditions so that the elimination of necessity for the use of chemical or mechanical action is to be considered as including the elimination of the necessity for the use of heat or varying degrees thereof.

Having now particularly described and ascertained the nature of my said invention, and in what manner the same is to be performed, I declare that what I claim is: --- [Claims not included here]

NOTES & COMMENTS

From the Article in *Tech. World Mag*.:

1. First demo model was powerful enough to spark a gas engine.
2. Second model developed 8 volts.
3. Demo model at Patent Office was elevated on short poles.
4. The model planned to power Phoenix AZ would be elevated 200 feet.
5. The Absorber "consists of a series of magnetized steel plates set in a circle (the manner of preparing them is kept secret)".
6. "[T]he magnetizing of one set of plates... is part of the secret of the treatment which makes it respond to the accumulations of the atmosphere".

From British Patent # 1098 (1913):

"I have been able... to produce from a very small apparatus at comparatively low elevation, say about 50 or 60 feet above the earth's surface, a substantially constant current at a commercially useable voltage and amperage".

"This current... is capable of being readily increased by additions of the unit elements in the apparatus".

Fig. 1 and Fig. 2 show the magnet poles are connected N-S by a thick iron rod (thick compared to the lines used for wires in the drawings).

No angle is specified for the V-shaped zinc plates. The article (but not the patent) states that the plates are magnetized (obviously not zinc). Zinc-galvanized steel? Will a thin film of Zn work? Or, powdered Zn in a binder (more surface area)? Or, zinc-galvanized iron wire in a coil?

The Palmer Craig device (www.rexresearch.com/craig/craig.htm) is powered by the terrestrial magnetic field, and employs a thin film of bismuth to capture the energy as diamagnetism. Perhaps this can be integrated with Meyers' device.

Figure 8 (the demonstration of principle) show uninsulated iron wire being used to connect the plates and magnets. The wire is wound around the Bloch wall area of the horseshoe magnets. Perhaps Coler-type windings around the poles could be used here (See: www.rexresearch.com/coler/coler.htm). Coler used copper plates as "condensers" in his device. Could copper plates be used for the Meyers device? Perhaps flat (Tesla non-inductive) coils could be integrated here.

"It is essential that the plates 3 be disposed substantially N and S with their flat faces approximately parallel to the adjacent faces of the co-operating magnets....

"I find that it is immaterial whether the N pole of one of the magnets is disposed to the N and the S pole to the S, or vice versa".

"[T]he magnets and zinc plates... produce electrical currents when disposed... N and S, but when disposed... E and W no such currents are produced".

"[E]levation is by no means vital, but... more efficient results are obtained by placing the zincs and magnets on elevated supports".

"The elements may be disposed in horizontal planes [or vertically]...".

The "zinc plate 56... [is] arranged longitudinally substantially parallel with the legs of the magnet and with the faces of the plate exposed toward the respective legs of the magnet, the plate being thus arranged endwise toward the north and south".

The first model used "horseshoe magnets approximately 4 inches in length, the bar comprising the horseshoe being about one inch square, the zincs being dimensioned proportionately and from this apparatus with the employment of a single intensifier and rectifier, as above stated, I was able to obtain a constant current of 8 volts... [T]he magnets... may be permanent or may be electromagnets, or a combination of the two... I find the best results obtained by... nickel chrome steel".

Comments & Questions:

The rectifier is described as a preferred embodiment, but other designs also work. The Ed Gray capacitor design comes to mind (www.rexresearch.com/evgray/1gray.htm). The Tate Ambient Power Module also might apply (www.rexresearch/tate/tate.htm).

Would non-ferrous magnets work? Is there a frequency involved (oscilloscope tests)? Coler found that ferromagnetism has a resonant frequency about 180 KHz. Can the components be made adjustable for RLC-resonance?

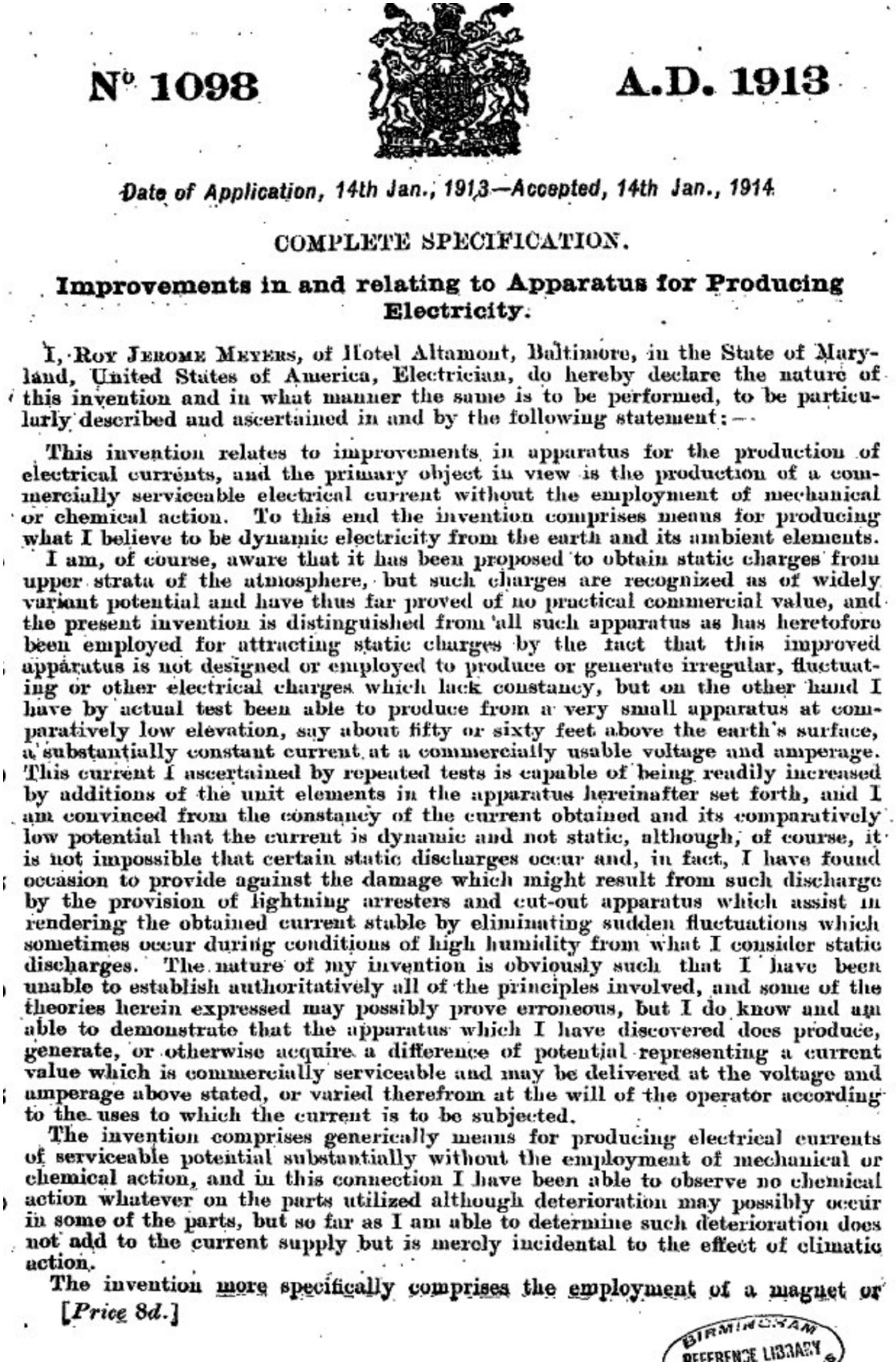

Nº 1098 A.D. 1913

Date of Application, 14th Jan., 1913—Accepted, 14th Jan., 1914

COMPLETE SPECIFICATION.

Improvements in and relating to Apparatus for Producing Electricity.

I, Roy Jerome Meyers, of Hotel Altamont, Baltimore, in the State of Maryland, United States of America, Electrician, do hereby declare the nature of this invention and in what manner the same is to be performed, to be particularly described and ascertained in and by the following statement:—

This invention relates to improvements in apparatus for the production of electrical currents, and the primary object in view is the production of a commercially serviceable electrical current without the employment of mechanical or chemical action. To this end the invention comprises means for producing what I believe to be dynamic electricity from the earth and its ambient elements. I am, of course, aware that it has been proposed to obtain static charges from upper strata of the atmosphere, but such charges are recognized as of widely variant potential and have thus far proved of no practical commercial value, and the present invention is distinguished from all such apparatus as has heretofore been employed for attracting static charges by the fact that this improved apparatus is not designed or employed to produce or generate irregular, fluctuating or other electrical charges which lack constancy, but on the other hand I have by actual test been able to produce from a very small apparatus at comparatively low elevation, say about fifty or sixty feet above the earth's surface, a substantially constant current at a commercially usable voltage and amperage. This current I ascertained by repeated tests is capable of being readily increased by additions of the unit elements in the apparatus hereinafter set forth, and I am convinced from the constancy of the current obtained and its comparatively low potential that the current is dynamic and not static, although, of course, it is not impossible that certain static discharges occur and, in fact, I have found occasion to provide against the damage which might result from such discharge by the provision of lightning arresters and cut-out apparatus which assist in rendering the obtained current stable by eliminating sudden fluctuations which sometimes occur during conditions of high humidity from what I consider static discharges. The nature of my invention is obviously such that I have been unable to establish authoritatively all of the principles involved, and some of the theories herein expressed may possibly prove erroneous, but I do know and am able to demonstrate that the apparatus which I have discovered does produce, generate, or otherwise acquire a difference of potential representing a current value which is commercially serviceable and may be delivered at the voltage and amperage above stated, or varied therefrom at the will of the operator according to the uses to which the current is to be subjected.

The invention comprises generically means for producing electrical currents of serviceable potential substantially without the employment of mechanical or chemical action, and in this connection I have been able to observe no chemical action whatever on the parts utilized although deterioration may possibly occur in some of the parts, but so far as I am able to determine such deterioration does not add to the current supply but is merely incidental to the effect of climatic action.

The invention more specifically comprises the employment of a magnet or

[*Price 8d.*]

BIRMINGHAM REFERENCE LIBRARY

magnets and a co-operating element, such as zinc disposed adjacent to the magnet or magnets and connected in such manner and arranged relative to the earth so as to produce current, my observation being that current is produced only when such magnets have their poles facing substantially to the north and south and the zincs are disposed substantially along the magnets.

The invention also comprehends other details of construction, combinations and arrangements of parts as will hereinafter be fully set forth and claimed.

In the accompanying drawings:

Figure 1 is a top plan view of an apparatus embodying the features of the present invention, the arrow accompanying the figure indicating substantially the geographical north, parts of the figure being diagrammatic for condensing the showing.

Figure 2 is a view in side elevation of the parts seen in plan in Figure 1.

Figure 3 is a vertical section taken on the plane indicated by the line 3—3 of Figure 2 and looking in the direction indicated by the arrow.

Figure 4 is a detail view partly in elevation and partly in section showing the detail connections of the converter and intensifier.

Figure 5 is a transverse section taken on the planes indicated by line 5—5 of Figure 4 and looking downwardly.

Figure 6 is an enlarged detail fragmentary section illustrating the parts at the juncture of the conductors and one of the intensifiers.

Figure 7 is an enlarged detail view partly in elevation and partly in section of one of the automatic cut-outs and

Figure 8 is a diagrammatic view of one of the simplest forms of embodiments of the invention.

Referring to the drawing by numerals, 1, 1 indicates magnets connected by a magnetic substance 2, preferably an iron wire. The magnets 1 are arranged in pairs, one pair being spaced beneath the other, and interposed between the magnets are zinc plates 3, 3 connected by an iron wire conductor 4. Suitable insulating supports 5 are arranged for sustaining the respective magnets 1 and plates 3, 3. Each plate 3 is preferably bent substantially into **V** form, as clearly seen in Figure 1, and the **V** of one of the plates opens or faces toward the north and the **V** of the other plate to the south. I have determined by experimentation that it is essential that the plates 3 be disposed substantially north and south with their flat faces approximately parallel to the adjacent faces of the co-operating magnets, although by experience I have not discovered any material difference in the current obtained when the plates are disposed slightly to one side of north and south, as for instance when disposed in the line of the magnetic polarity of the earth. The same is true with respect to the magnets 1, the said magnets being disposed substantially north and south for operative purposes, although I find that it is immaterial whether the north pole of one of the magnets is disposed to the north and the south pole to the south, or *vice versa*, and it is my conviction from experience that it is essential to have the magnets of each pair connected by magnetic material so that the magnets substantially become one with a pole exposed to the north and a pole exposed to the south. In Figure 1, I have indicated in full lines by the letters S and N the respective polarities of the magnets 1, and have indicated in dotted lines the other poles of those magnets when the connection 2 is severed. I have found that the magnets and zinc plates operate to produce, whether by collection or generation I am not certain, electrical currents when disposed substantially north and south, but when disposed substantially east and west no such currents are produced. I also find that the question of elevation is by no means vital, but it is true that more efficient results are obtained by placing the zincs and magnets on elevated supports. I furthermore find from tests that it is possible to obtain currents from the apparatus with the zincs and magnets disposed in a building or otherwise enclosed, although more efficient results are obtained by having the said elements arranged in the open.

While in Figures 1, 2 and 3, I have shown the magnets and the zinc plates as superimposed, it will be apparent, as hereinafter fully set forth, that these elements may be juxtaposed in horizontal planes, and substantially the same results will be secured. Furthermore, the magnets 1 with the interposed zincs 3, as shown in Figures 1, 2 and 3 merely represent a unit which may be repeated either horizontally or vertically for increasing the current supply, and when the unit is repeated the zinc plates are arranged alternating with the magnets throughout the entire series as hereinafter indicated.

A conductor 6 is connected in multiple with the conductors 2 and a conductor 7 is connected with conductor 4, the conductor 6 extending to one terminal of a rectifier which I have indicated by the general reference character 8, and the conductor 7 extending to the other terminal of said rectifier. The rectifier as seen in diagram in Figure 1 may assume any of several well known embodiments of the electrical valve type and may consist of four assymmetric cells or Cooper-Hewitt mercury vapor lamps connected as indicated in Figure 1 for permitting communication of the positive impulses from conductor 7 only to the line conductor 9 and the negative impulses from the conductor 7 only to the line conductor 10, and permitting the positive impulses from the conductor 6 only to the line conductor 9 and the negative impulses from conductor 6 only to the line conductor 10. The current from this rectifier may be delivered through the conductors 9 and 10 to any suitable source for consumption.

While the said rectifier 8 may consist of any of the known types, as above outlined, it preferably consists of a specially constructed rectifier which also has the capacity of intensifying the current and comprises specifically the elements shown in detail in Figures 4, 5 and 6 wherein I have disclosed the detail wiring of the rectifier when composed of four of the rectifying and intensifying elements instead of assymmetric cells or simple mercury vapor valves. As each of these structures is an exact embodiment of all the others, one only will be described, and the description will apply to all. The rectifying element of each construction consists of a mercury tube 11 which is preferably formed of copper but may be formed of glass or other suitable material, and comprises a cylinder having its end portions tapered and each terminating in an insulating plug or stopper 12. Through the upper stopper 12 is extended the electrode 13 which extends well into the tube and preferably substantially one-half the length thereof to a point adjacent the inner end of an opposing electrode 14 which latter electrode extends thence downwardly through the insulation 12 at the lower end of the tube. The tube 11 is supplied with mercury and is adapted to operate on the principle of the mercury vapor lamp, serving to rectify current by checking back impulses of one sign and permitting passage of impulses of the other. To avoid the necessity for utilizing a starter, as is common with the lamp type of electrical valve, the supply of mercury within the tube may be sufficient to contact with the lower end of the electrode 13 when current is not being supplied, so that as soon as current is passed from one electrode to the other sufficiently for volatilizing that portion of the mercury immediately adjacent the lower end of electrode 13, the structure begins its operation as a rectifier. The tube 11 is surrounded by a tube 15 which is preferably spaced from tube 11 sufficiently for allowing atmospheric or other cooling circulation to pass the tube 11. In some instances, it may be desirable to cool the tube 11 by a surrounding body of liquid, as hereinafter indicated. The tube 15 may be of insulating material but I find efficient results attained by the employment of a steel tube, and fixed to the ends of the tube are insulating disks 16, 16 forming a spool on which are wound twin wires 6^1 and 7^1, the wire 6^1 being connected at the inner helix of the coil with the outer end of the electrode 14, the lower portion of said electrode being extended to one side of the tube 11 and passed through an insulating sleeve 17 extending through the tube 15, and at its outer end merging into the adjacent end of the wire 6^1. The wire 7^1 extends directly from the outer portion of the spool through the several helices to a point adjacent the juncture of the electrode 14 with wire 6^1 and thence extends in mechanical parallelism with the wire throughout the coil, the wire 6^1 ending in a terminal 18 and the wire 7^1 ending in a terminal 19. For the sake of convenience of description and of tracing the circuits, each of the apparatus just above described and herein known as an intensifier and rectifier will be mentioned as A, B, C and D, respectively. Conductor 6 is formed with branches 20 and 21 and conductor 7 is formed with similar branches 22 and 23. Branch 20 from conductor 6 connects with conductor 7^1 of intensifier B and branch 21 of conductor 6 connects with the conductor 7^1 of intensifier C, while branch 22 of conductor 7 connects with conductor 7^1 of intensifier A and branch 23 of conductor 7 connects with conductor 7^1 of intensifier D. A conductor 27 is connected with terminal 19 of intensifier A and extends to and is connected with terminal 18 of intensifier C, and a conductor 28 is connected with the terminal 19 of intensifier C and extends to and is connected with electrode 13 of intensifier A. A conductor 29 extends from the terminal 19 of intensifier B to the terminal 18 of intensifier D and a conductor 30 extends from the terminal 19 of intensifier D to electrode 13 of intensifier B. Each electrode 13 is supported on a spider 13^1 resting on the upper disk 16 of the respective intensifier. Conductors 31 and 32 are connected with the terminals 18 of intensifiers A and B and are united to form the positive line wire 9 which co-operates with the negative line wire 10 and extends to any suitable point of consumption. The line wire 10 is provided with branches 35 and 36 extending to the electrodes 13 of intensifiers C and D for completing the negative side of the circuit.

Thus it will be seen that alternating currents produced in the wires 6 and 7 will be rectified and delivered in the form of a direct current through the line wires 9 and 10, and I find by experiment that the wires 6 and 7 should be of iron, preferably soft, and may of course be insulated, the other wiring not specified as iron being of copper or other suitable material.

In carrying out the operation as stated, the circuits may be traced as follows: A positive impulse starting at the zincs 3 is directed along conductor 7 to branch 23 to conductor 7^1 and the winding of intensifier D to terminal 19 through conductor 30 to electrode 13 of the rectifier of intensifier B through said rectifier to the conductor 6^1, through the winding thereof to the contact 18, conductor 32 and to the line wire 9. The next or negative impulse directed along conductor 7 cannot find its way along branch 23 and the circuit just above traced because it cannot pass across the rectifier of intensifier B but instead the negative impulse passes along conductor 22 to conductor 7 of intensifier A and the winding thereof to the contact 19 to conductor 27 to contact 18 of intensifier C, to the winding of the wire 6^1 thereof to the electrode 14 through the rectifier to the electrode 13 and conductor 36 to the line wire 10. A positive impulse delivered to wire 6 passes along the said wire to the branch 21 to the conductor 7^1 of intensifier C and the winding thereof to the contact 19, conductor 28, electrode 13 of the rectifier of intensifier A, electrode 14 thereof and conductor 6^1 to contact 18 and wire 31 to line wire 9. Obviously the positive impulse cannot pass along the wire 20 because of its inverse approach to the rectifier of intensifier B. The next impulse or negative impulse delivered to conductor 6 cannot pass along conductor 21 because of its connection with electrode 13 of the rectifier of intensifier A, but instead passes along conductor 20 to the wire 7^1 and its winding forming part of intensifier B to the contact 19 and conductor 29 to contact 18 and the winding of wire 6^1 of intensifier D to the electrode 14 and through the rectifier to the electrode 13 and conductor 35 to line wire 10. Thus the current is rectified and all positive impulses directed along one line and all negative impulses along the other line so that the potential difference between the two lines will be maximum for the given current of the alternating circuit. It is, of course, apparent that a less number of intensifiers with their accompanying rectifier elements may be employed with a sacrifice of the impulses which are checked back from a lack of ability to pass the respective rectifier elements, and in fact I have secured efficient results by the use of a single intensifier with its rectifier elements, as hereinafter set forth.

Grounding conductors 37 and 38 are connected respectively with the conductors 6 and 7 and are provided with the ordinary lightning arresters 39 and 40 respectively for protecting the circuit against high tension static charges.

Conductors 41 and 42 are connected respectively with the conductors 6 and 7 and each connects with an automatic cut-out 43 which is grounded as at 44. Each of said automatic cut-outs is exactly like the other and one of the same is shown in detail in Figure 7 and comprises the inductive resistance 45 provided with an insulated binding post 46 with which the respective conductor 6 or 7 is connected, said post also supporting a spring 48 which sustains an armature 49 adjacent the core of the resistance 45. The helix of resistance 45 is connected preferably through the spring to the binding post at one end and at the other end is grounded on the core of the resistance, the said core being grounded by ground conductor 44 which extends to the metallic plate 52 imbedded in moist carbon or other inductive material 53 buried in the earth. Each of the conductors 41, 42 and 44 is of iron, and in this connection I wish it understood that where I state the specific substance I am able to verify the accuracy of the statement by the results of tests which I have made, but of course I wish to include along with such substance all equivalents, as for instance, where iron is mentioned its by-products, such as steel, and its equivalents such as nickel and other magnetic substances are intended to be comprehended. The cut-out apparatus seen in detail in Figure 7 is employed particularly for insuring against high tension currents it being obvious from the structure shown that when potential rises beyond the limit established by the tension of the spring sustaining the armature 49, the armature will be moved to a position contacting with the core of the cut-out device and thereby directly close the ground connection for line wire 41 with conductor 44, eliminating the resistance of winding 45 and allowing the high tension current to be discharged to the ground. Immediately upon such discharge the winding 45 losing its current will allow the core to become demagnetized and release the armature 49 whereby the ground connection is substantially broken leaving only the connection through the winding 45 the resistance of which is sufficient for insuring against loss of low tension current.

In Figure 8 I have illustrated an apparatus which though apparently primitive in construction and arrangement comprehends the first successful embodiment which I produced in the course of the discovery of the present invention, and it will be observed that the essential features of the invention are therein disclosed. The structure delineated in said figure consists of horse shoe magnets 54, 54, one facing north and the other south, that is, each opening in the respective directions indicated and the two being connected by an iron wire 55 which is uninsulated and wrapped about the respective magnets at or adjacent the neutral zone thereof, and the wire 55 is preferably soldered to the respective magnets each end portion of the wire 55 being extended from the respective magnet to and connected with, as by being soldered to, a zinc plate 56, there being a plate 56 for each magnet and each plate being arranged longitudinally substantially parallel with the legs of the magnet and with the faces of the plate exposed toward the respective legs of the magnet, the plate being thus arranged endwise toward the north and south. An iron wire 57 connects the plates 56, the ends of the wire being preferably connected adjacent the outer ends of the plates but from experiment I find that the wire may be connected at practically any point to the plate. Lead wires 58 and 59 are connected respectively with the wires 55 and 57 and supply an alternating current ordinarily at a comparatively low tension, and to control such current the wires 58 and 59 may be extended to a rectifier or combined rectifier and intensifier, as above set forth.

The tests which I have found successful with the apparatus seen in Figure 8 were carried out by the employment first of horse shoe magnets approximately four inches in length, the bar comprising the horse shoe being about one inch square, the zincs being dimensioned proportionately and from this apparatus with the employment of a single intensifier and rectifier, as above stated, I was able to obtain a constant current of 8 volts.

It should be obvious that the magnets forming one of the electrodes of this apparatus may be permanent or may be electro-magnets, or a combination of the two.

While the magnets mentioned throughout the above may be formed of any magnetic substance, I find the best results obtained by the employment of the nickel chrome steel.

While the successful operation of the various devices which I have constructed embodying the present invention have not enabled me to arrive definitely and positively at fixed conclusions relative to the principles and theories of operation and the source from which current is supplied, I wish it to be understood that I consider myself as the first inventor of an apparatus of the general type hereinbefore described capable of producing commercially serviceable electricity, for which reason my claims hereinafter appended contemplate that I may utilize a wide range of equivalents so far as concerns details of construction suggested as preferably employed.

The current which I am able to obtain is dynamic in the sense that it is not static and its production is accomplished without chemical or mechanical action either incident to the actual chemical or mechanical motion or incident to changing calorific conditions so that the elimination of necessity for the use of chemical or mechanical action is to be considered as including the elimination of the necessity for the use of heat or varying degrees thereof.

Having now particularly described and ascertained the nature of my said invention, and in what manner the same is to be performed, I declare that what I claim is:—

1. Means for producing dynamic electricity without mechanical or chemical action.
2. Electricity producing means as claimed in Claim 1, characterized in that said means is capable of producing low tension current.
3. Electricity producing means as claimed in Claim 1, characterized in that said means comprises a magnet, and means co-operating with the latter.
4. Means as set forth in Claim 3, characterized in that the parts are stationary.
5. Means as claimed in Claim 3, characterized in that the magnet is disposed substantially north and south.
6. Means as claimed in Claims 1, 3 and 5, wherein the means coacting with the magnet is a metal and said parts are disposed substantially north and south.
7. Means as claimed in Claims 1, 3 and 6, in which the metal is zinc, and combined with conductors connected with the co-operating parts to deliver current therefrom.
8. Means as set forth in Claim 1, characterized in that said means comprises spaced magnets with an adjacent co-operating zinc, and an iron wire connecting the magnets, and means to deliver electrical current from said parts.
9. Means as set forth in Claim 8, characterized in that the current delivery means comprises an intensifier and rectifier.
10. Means as set forth in Claim 8, characterized in that the current delivery means comprises wires connected with the magnets and zinc, an electrical valve connected with one of said wires, and independent helices surrounding said valve, one of said helices being connected with the terminal of the valve at the opposite side from the contact of said wire and the other of said helices being connected with the other wire, the electrical valve comprising a mercury vapor rectifier.
11. The process of producing electricity comprising exposing a magnet and

co-operating means disposed substantially north and south and taking off current therefrom.

12. The process of producing electricity comprising disposing stationary elements in co-operative relation in respect to each other and with respect to the earth for giving off relatively low tension electrical current, substantially without chemical action.

13. Electricity producing means constructed and operating substantially as described with reference to the accompanying drawings.

Dated this 13th day of January, 1913.

W. P. THOMPSON & Co.,
6, Lord Street, Liverpool, and at
Bradford & London,
Agents for the Applicant.

Reference has been directed, in pursuance of Section 7, Sub-section 4, of the Patents and Designs Act, 1907, to Specifications No. 16,709 of 1887, No. 14,033 of 1899, No. 15,412 of 1906, and No. 5457 of 1911.

Redhill: Printed for His Majesty's Stationery Office, by Love & Malcomson, Ltd.—1914.

Fig. 1.

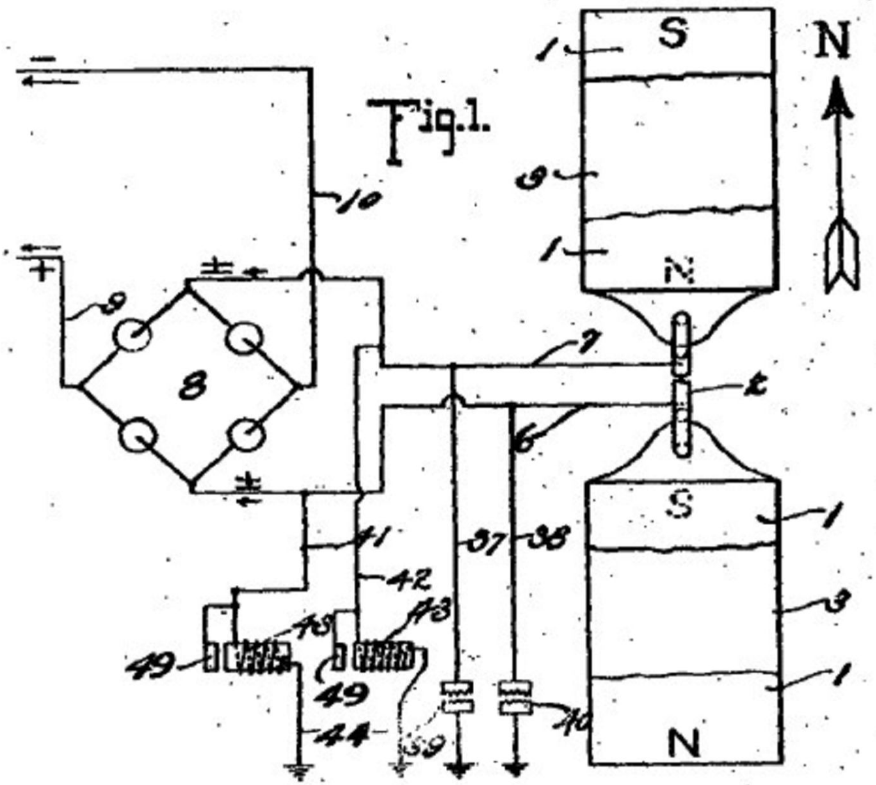
−
+
10
9
8
7
6
t
S
N
1
3
41
42
43
48
49
44
39
37
38
40

Fig. 8.

Fig. 6

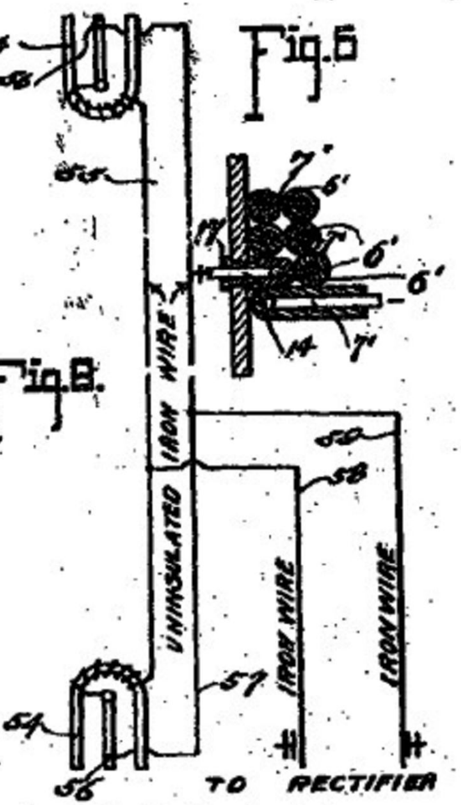
54
55
56
57
58
59
UNINSULATED IRON WIRE
IRON WIRE
IRON WIRE
TO RECTIFIER
7'
6'
6'
6'
14
17
7'

Fig. 2.

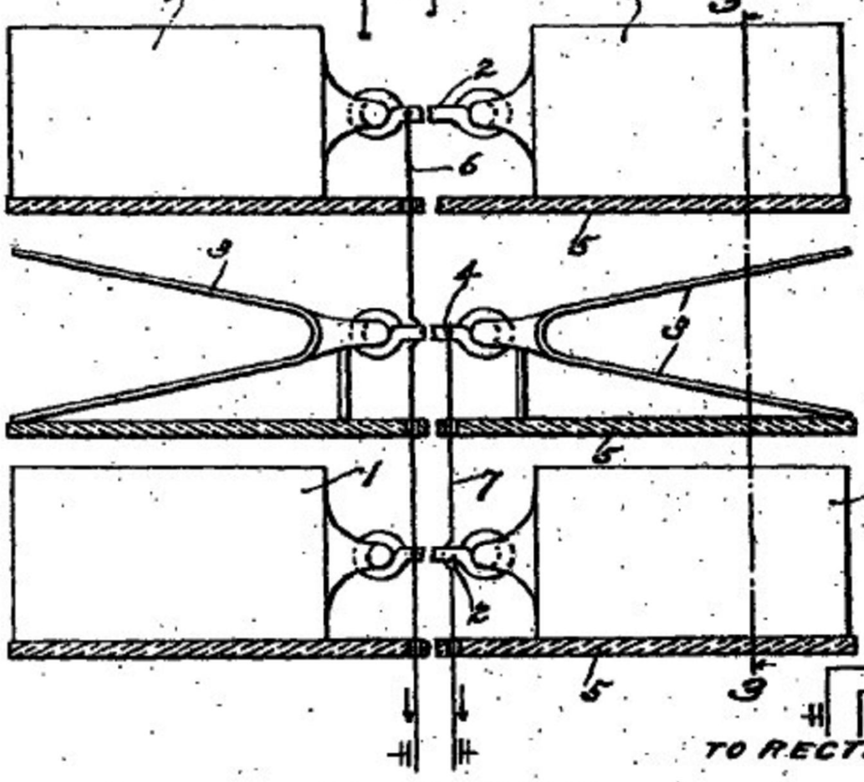
1
2
3
4
5
6
7
t
TO RECTIFIER

Fig. 5.

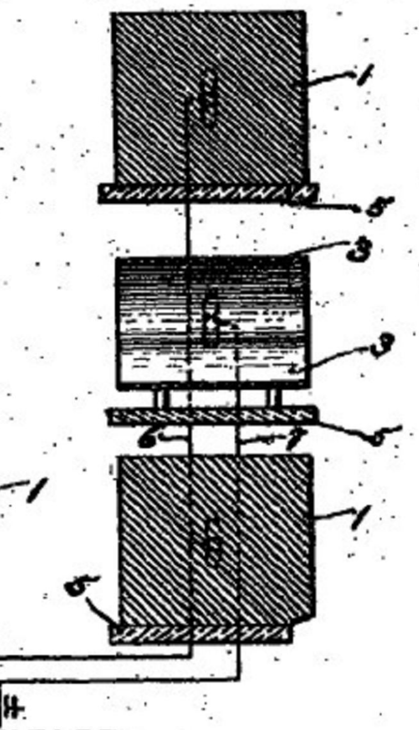
1
3
5
6
7
TO RECTIFIER

Fig 7.

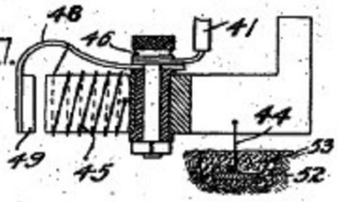
48
46
41
44
53
52
49
45

Fig 4.

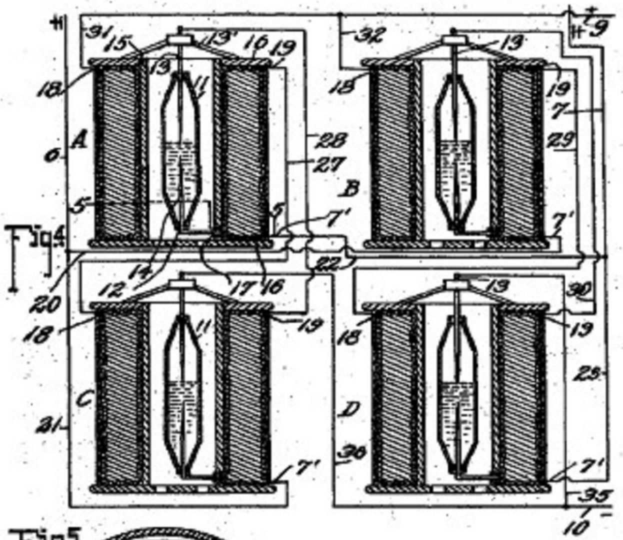
11
31
15
13'
16
19
13
18
A
6
5
28
27
B
7'
32
9
7
29
12
14
17
16
22
20
30
C
D
21
36
35
10
25

Fig 5.

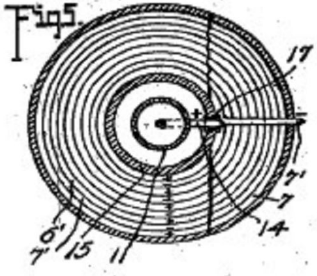
17
7'
7
14
6'
7'
15
11

Your Support Maintains this Service. Please Invest in the mind-boggling Rex Research Website CD -- Always The Best Investment ... Everything on this Website & More (the Bonus Files CD) -- Only $13, Postpaid Anywhere ! Secure Transaction via PayPal: All Major Credit Cards Accepted:

Buy Now

No Credit ? Don't like PayPal ? Send a Check/Money Order to:

Rex Research, PO Box 19250, Jean, NV 89019 USA

---

rexresearch.com

PHGN 300
Crystal Radio
(in 45 minutes)

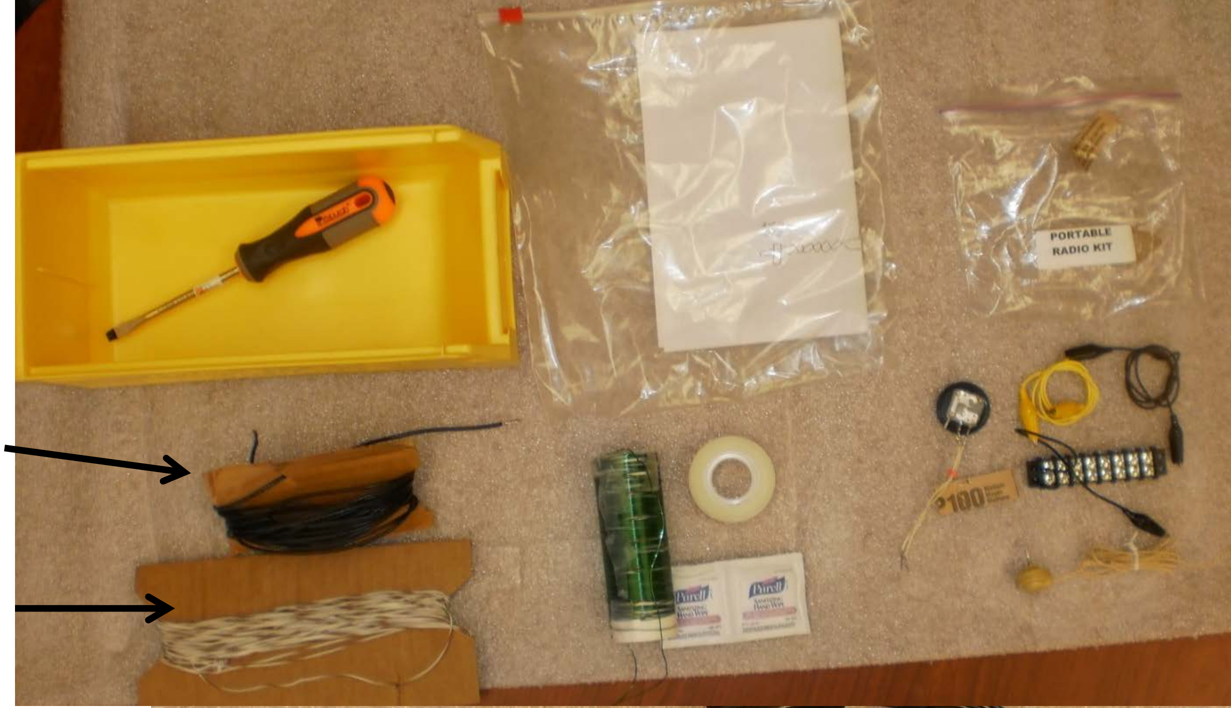
PORTABLE RADIO KIT
Purell
P100

Ground Wire

Antenna
Need double
length

PORTABLE RADIO KIT

Please put away like
this when done!

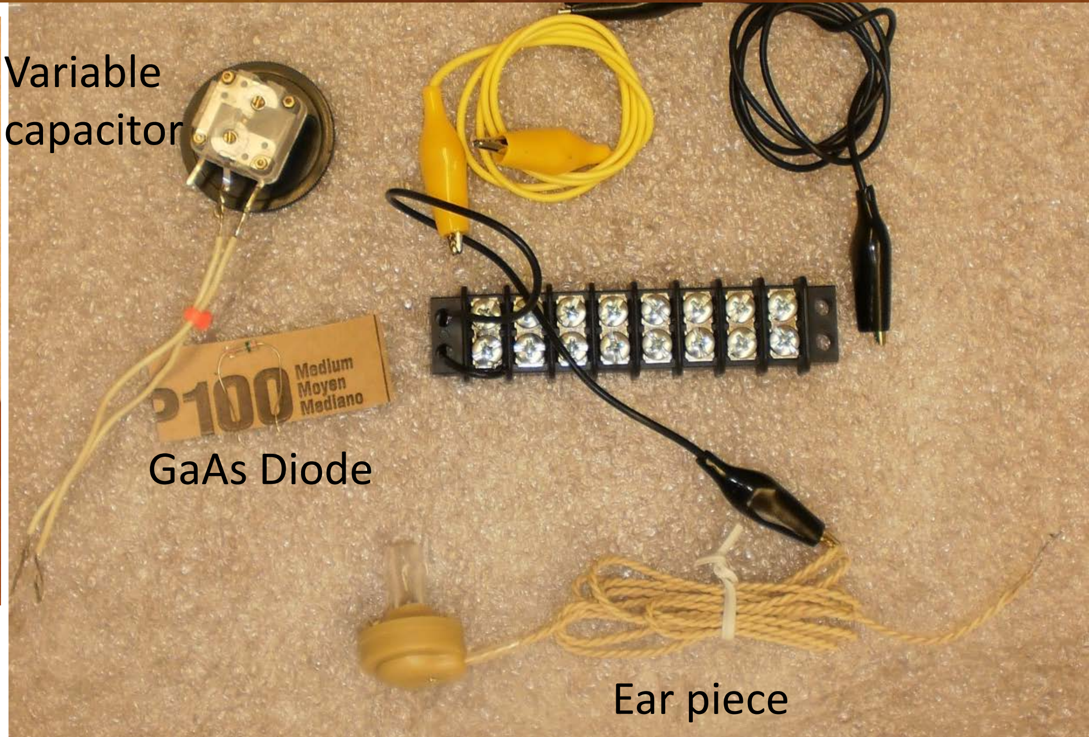
P100
Medium
Moyen
Mediano

Variable
capacitor

GaAs Diode

Ear piece

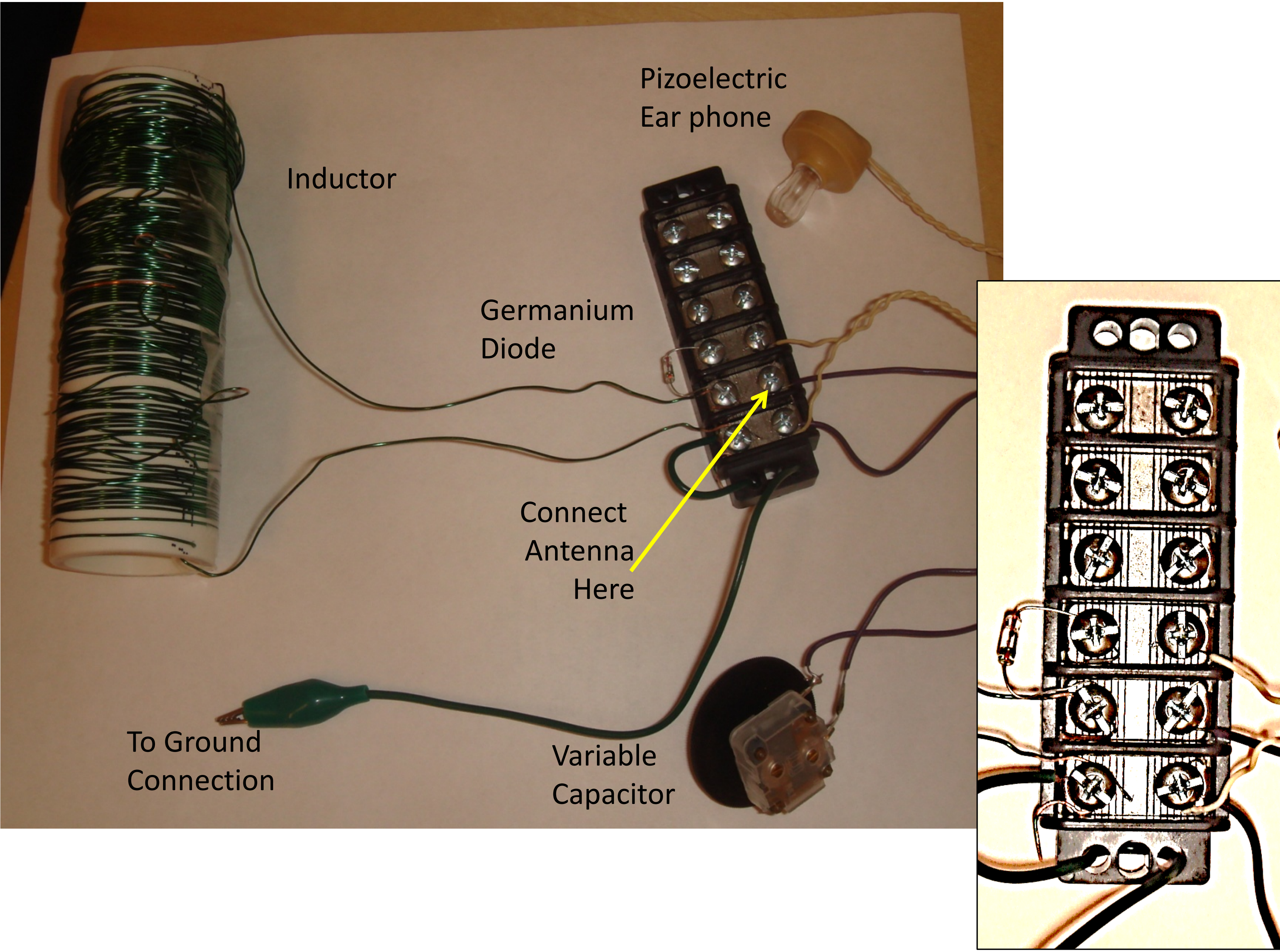
Pizoelectric
Ear phone
Inductor
Germanium
Diode
Connect
Antenna
Here
To Ground
Connection
Variable
Capacitor

Crystal Radio in 45 minutes.

Work in groups of 4.

2 people should assemble the radio and connect the ground wire.

2 people should string the antenna

If you get your radio to work using the GeAs diode,
Try the Gallinium + Cat Whisker

Return everything the way you found it before leaving

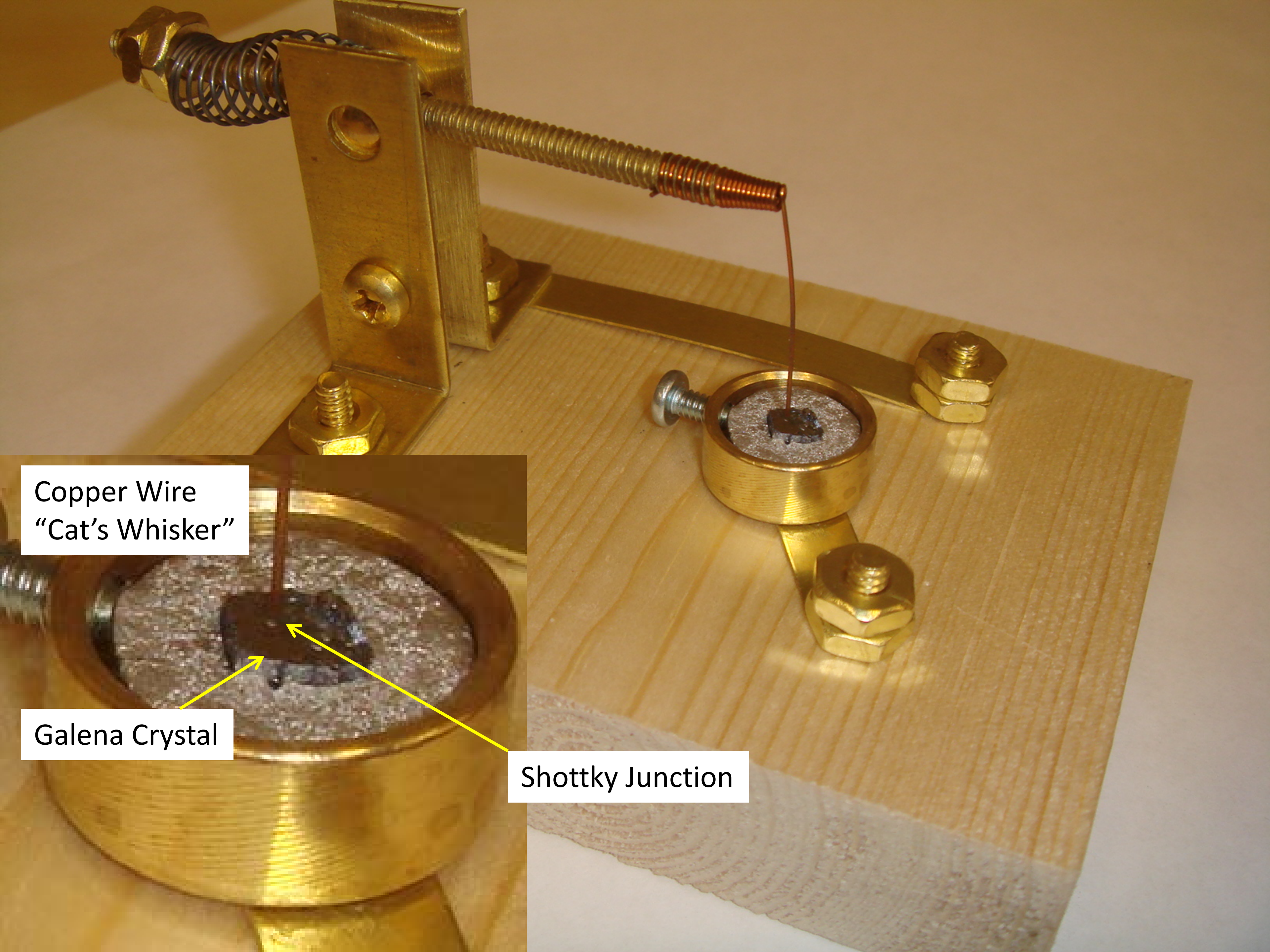
Copper Wire
"Cat's Whisker"
Galena Crystal
Shottky Junction

The radios **must** have a good ground connection and a long antenna. Long means 100’-200’ of antenna wire outside. 200’ is better

Two possible routes are shown here. Use one.

Attach to tree branches you can reach.

When crossing a sidewalk, make it high so pedestrians don’t have to duck (they won’t).

You will take antenna down when class ends.

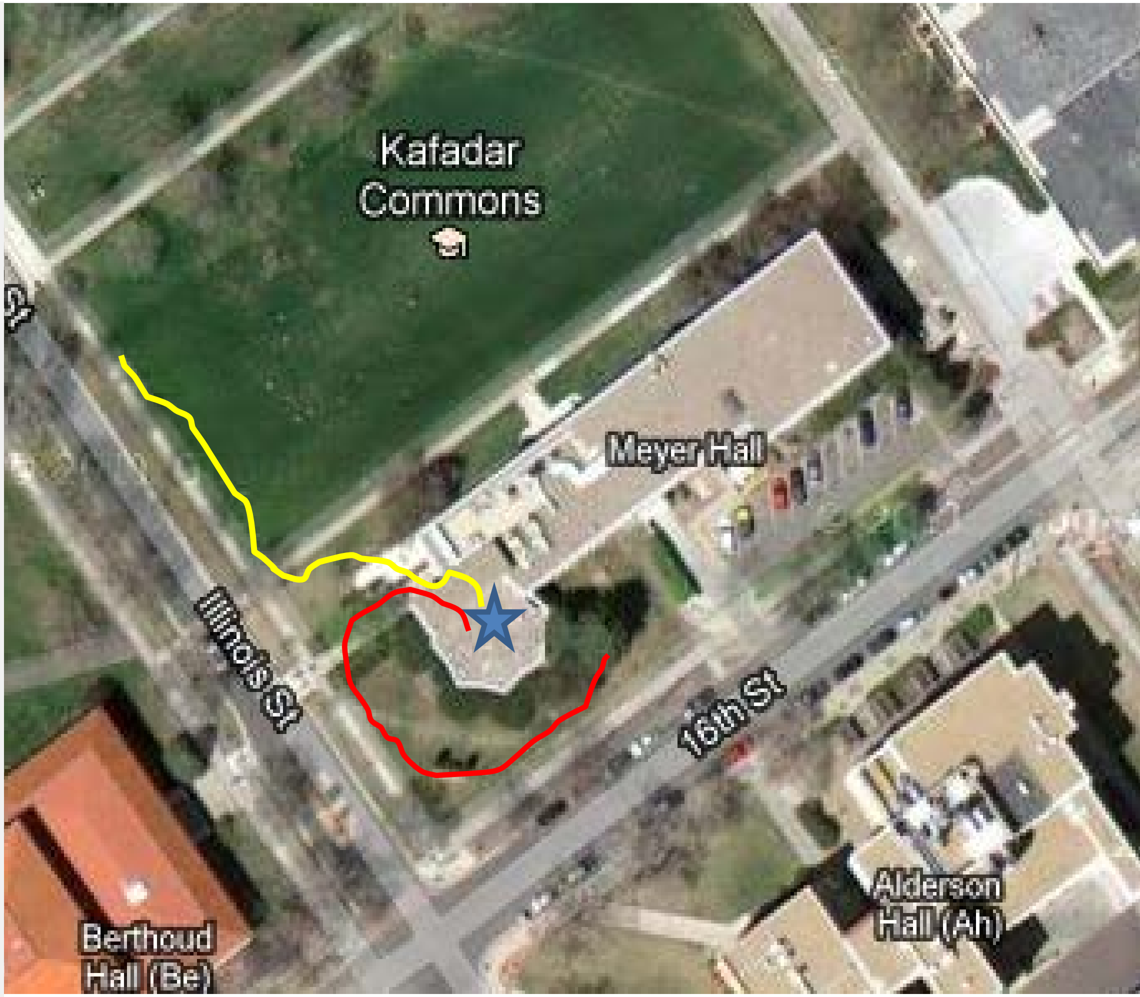
Kafadar Commons
Meyer Hall
St
Illinois St
16th St
Alderson Hall (Ah)
Berthoud Hall (Be)

4/6
45c
AG

BECK BOOK CO. LTD.
New and Secondhand Booksellers
53 PULTENEY STREET,
ADELAIDE, S.A.

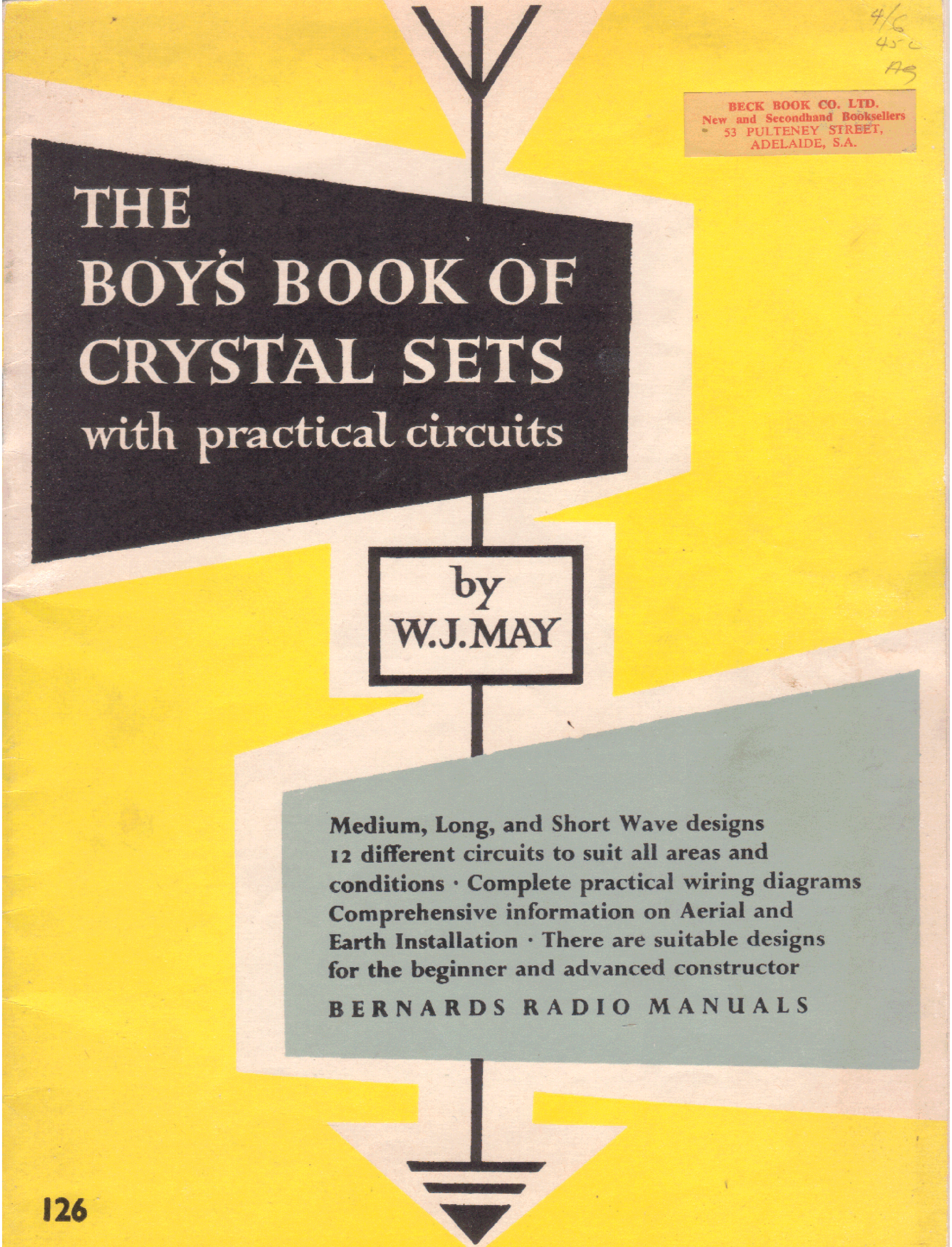

THE BOY'S BOOK OF CRYSTAL SETS

with practical circuits

by
W. J. MAY

**Medium, Long, and Short Wave designs
12 different circuits to suit all areas and conditions · Complete practical wiring diagrams
Comprehensive information on Aerial and Earth Installation · There are suitable designs for the beginner and advanced constructor**

BERNARDS RADIO MANUALS

126

THE BOY'S BOOK of CRYSTAL SETS

BY

W. J. MAY

⁂

BERNARDS (PUBLISHERS) LTD.
THE GRAMPIANS
WESTERN GATE
LONDON W.6

General Editor
WALTER J. MAY

First Published—March, 1954
Second Impression—June, 1954
Third Impression—April, 1955
Fourth Impression—January, 1956
Fifth Impression—February, 1957
Sixth Impression—August, 1957
Seventh Impression—March, 1958
Eighth Impression—October, 1958
Ninth Impression—April, 1959
Tenth Impression—December, 1959
Eleventh Impression—April, 1960
Twelfth Impression—October, 1960
Thirteenth Impression—March, 1961
Fourteenth Impression—August, 1961
Fifteenth Impression—January, 1962
Sixteenth Impression—November, 1962
Seventeenth Impression—August, 1963
Eighteenth Impression—September, 1964

⁂

Printed by V. Cooper & Partners, Ltd., Flitcroft St., W.C.2
for Bernards (Publishers) Ltd., The Grampians, Western Gate, London, W.6

CONTENTS

Page

Preface - - - - - 4

Introduction - - - - 5

Receiver No. 1 - - - - 12

Receiver No. 2 - - - - 13

Receiver No. 3 - - - - 14

Receiver No. 4 - - - - 16

Receiver No. 5 - - - - 19

Receiver No. 6 - - - - 20

Receiver No. 7 - - - - 23

Receiver No. 8 - - - - 24

Receiver No. 9 - - - - 26

Receiver No. 10 - - - - 27

Receiver No. 11 - - - - 30

Receiver No. 12 - - - - 31

PREFACE

The Publishers make no apology for reviving interest in the crystal set. For too long the technical press has immersed itself in an orgy of complex superheterodyne receivers and television equipment, all of which is necessarily expensive. In the 1920's, crystal set construction was a fascinating and inexpensive pastime indulged in by the majority of boys of that day. Today this pleasurable hobby can be even more attractive, the introduction of the germanium crystal diode and high performance coils, has opened up new paths which will capture the interest of boys of all ages. Perhaps most important of all. not only can a receiver capable of impeccable reproduction be produced without any technical knowledge; but the cost is no greater than that of a few visits to the cinema

❦

INTRODUCTION

Before any attempt is made to construct a receiver, it is necessary to examine the problems which surround the crystal set, so that the best can be obtained from any of the designs attempted.

Firstly, it must be understood that the crystal set as it is to-day does **not** provide any amplification. It relies entirely on what is fed into it via the aerial and earth system and gives a very faithful replica of the original transmission.

From this it will be obvious that the aerial and earth system must be as efficient as possible if the final results are to be in any way outstanding. This because these are the only means by which the signals are fed to the receiver.

Secondly it is necessary to understand the nature of the transmitted signal, then it will be easy to understand the working of the set and to appreciate the function of each of the components.

When crystal sets first became popular, very little information was generally available, at least, not in a form that school-boys could understand.

We all built sets of all shapes and sizes, with coil designs that had to be seen to be believed, but very few of us had much idea of how they worked.

I well remember the case of a cousin of mine who, in those days acquired a magnificent variable capacitor or condenser as it was then called, this instrument had a most impressively engraved dial of polished ebonite, brass vanes and nickel plated end plates. Having observed similar dials on several highly priced commercial receivers in the town, he at once ripped out the somewhat tattered coil from the family receiver and replaced it with this device. The profound silence which ensued caused considerable amazement and dismay until a better informed adult explained the mysteries of L and C to him. Readers of this manual, however, will be better informed and there is no risk of failure with any of the designs described provided the instructions are carefully followed.

The Signal

To commence, we will assume that an orchestra is playing in a broadcasting studio. Since the principle of radio transmission is electrical it is necessary to change the sound produced by the orchestra into an electrical equivalent. This is carried out by the microphone, which picks up the sound and changes it into minute electric currents.

As they appear at the output of the microphone they are too small to be of use and accordingly are passed through a high power amplifier. These amplified currents could now be transmitted, but unfortunately, as we shall learn, owing to the inherent nature of the signal in this state, transmission over any useful distance would be impractical.

When the music from the orchestra is transformed into electrical currents they are in the form of alternating currents, usually called A.C., that is they rise to a maximum in one direction, fall to a minimum, rise to a maximum in the opposite direction and then fall to minimum again. This process is repeated over and over again. One complete rise and fall in each direction is called a cycle and is drawn in Fig. 1. Every time a note is struck on a piano, vibrations are sent out which reach the ear enabling you to hear it. These vibrations are also spoken of as cycles, they rise and fall in intensity the same way as an alternating current. The number of cycles radiated by any given note over a period of one second are referred to as its frequency. Middle C on the piano sends out 261 cycles every second and is known as having a frequency of 261. The microphone also " hears " the note and in the case of middle C produces minute A.C. at 261 cycles. This can be drawn as in Fig. 2, the only difference between Fig. 1 and 2 is, that the time factor is given so that the frequency can be identified. The higher the pitch of a note the higher the frequency and the lower the pitch the lower the frequency. On a piano the frequency of the top note is 3515 cycles and that of the lower 27 cycles. Those of you who have listened to an organ in a concert hall will have noticed that when a very deep note was played, it sounded like a growl to the ear, but the vibrating fre-

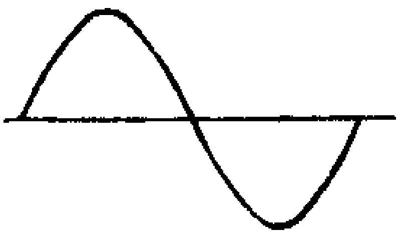

FIG 1

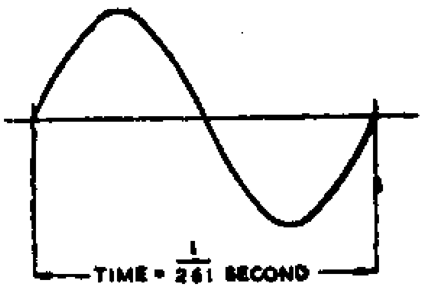

FIG 2

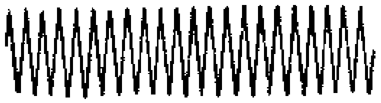

FIG 3

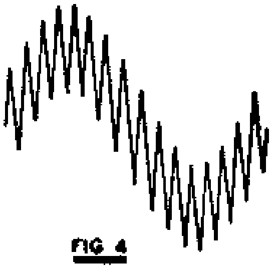

FIG 4

quency could be distinctly felt through the seat. Higher notes have too high a frequency to be observed in this manner.

The range of sounds which can be detected by the human ear are known as audio or low frequencies. From this you will understand that a low frequency amplifier is one which amplifies sound.

So that the transmitter will carry the programme over a useful distance it is necessary to radiate high frequencies. Now, as explained, the programme to be transmitted consists of low frequencies, and to overcome the difficulty, the transmitter generates A.C. of high frequency and combines it with the low frequencies. It will now be understood that the transmitted signal consists essentially of two different parts, a high and a low frequency content.

Fig. 3 gives a representation of the high frequency signal generated by the transmitter. In the case of the London Home Service, the frequency is 908000 cycles.

When referring to a high frequency signal on the medium or long wave-band it is usual to express the frequency in thousands of cycles, thus 908000 cycles becomes 908 kilo-cycles, which in turn may be abbreviated to 908 k/cs. It might at first be thought that by adding the low frequency or L.F. signal to the high frequency or H.F. carrier a form such as in Fig. 4 would result, such a combination is useless, and, so that the original L.F. content can be satisfactorily extracted by the receiver, the L.F. signal must vary the amplitude or output power of the H.F. signal as in Fig. 5.

It is in this form that the signal arrives at the receiving aerial. The aerial in itself is incapable of discriminating between one signal and another, and countless signals will be collected by the aerial at any one time. Many of these are too weak to be of use but the stronger ones must be sorted out since there is no point in receiving several programmes at once.

Fig. 6 shows the basic circuit of the input to a crystal. The coil L possesses a quality known as inductance, and the capacitor C, that of capacitance. If the coil had no capacitance whatever across it, all signals arriving at the aerial would be effectively short circuited to earth. As a matter of interest, it is impossible to obtain this state of affairs since even without any additional capa-

citor any coil **must contain a certain amount of self capacitance.**

By combining a coil and capacitor as in Fig. 6 a peculiar effect is observed, at one particular frequency, the signals are not short circuited to earth, but are developed across the coil. In other words the combined effect of L and C no longer provides a short circuit, **but only at one particular frequency.** If the value of C is altered the effect will be observed at a different frequency; likewise by altering L the frequency at which the effect will take place can be changed. There is a name for this phenomena, the frequency at which it occurs with any given L and C combination is known as the **resonant frequency.**

The values of the coils and capacitors shown in this manual have been carefully chosen so that resonance will be obtained at all frequencies where stations are broadcasting. Broadcasting stations work in bands of frequencies, those of major interest to crystal set constructors are the medium wave-band 1200 k/cs—600 k/cs and the long wave-band 300 kc/s—150 k/cs.

Usually a variable capacitor is used with a fixed inductance to cover one band and an additional coil switched in to increase the inductance to cover the other. In this way the L and C combination can be adjusted to provide resonance at the desired frequency of any given station. In other words you can **select** the station you want by varying C, that is turning the dial of the variable capacitor. This procedure is referred to as tuning. Having selected or tuned the required station it still remains necessary to change the form of the signal back to that of the original transmission. This process is called detection or de-modulation.

Examination of Fig. 5 will show that the signal has been duplicated, in other words, as it rises in one direction it also rises equally in the other. In this form the signal is useless since each half of the signal cancels the other, and if this signal is applied to a pair of headphones silence will result. Obviously some provision must be made to get rid of the unwanted half of the signal, and it is here that the crystal detector must be considered. This device will pass current in one direction only, ignoring any signal in the opposite direction, so that if the signal of Fig. 5 is passed through such a crystal, that of Fig. 7 will result. There is still the H.F. content to be reckoned

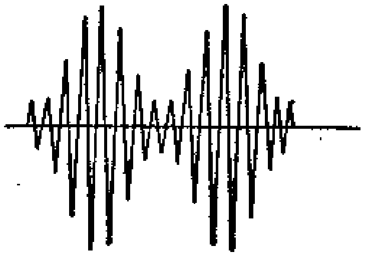

FIG 5

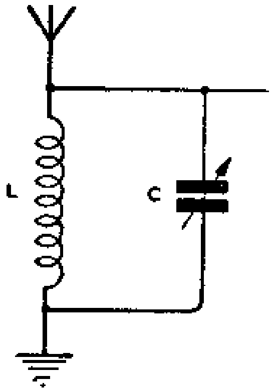

FIG 6

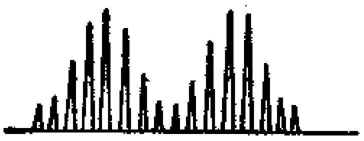

FIG 7

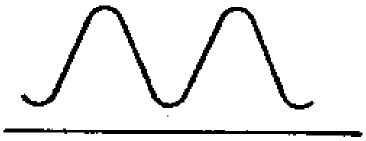

FIG 8

with, fortunately this is easily dealt with, a capacitor connected across the headphone terminals effectively disposes of this, leaving only the audio or L.F. content, as shown in Fig. 8. This audio content, which is a faithful replica of the original transmission is fed to the headphones. These in turn reverse the process of the microphone and transform the electrical currents into sound waves acceptable to the human ear.

Briefly then, your requirements are as follows:—

(1) A good aerial and earth installation, to make the most of the available signals.

(2) A receiver containing:—

   (a) Some form of coil and capacitor (L & C) combination to select or tune in the wanted station.

   (b) A crystal to get rid of the unwanted half of the signal (detection).

   (c) A fixed capacitor across the headphones to get rid of any remaining carrier.

(3) A pair of sensitive high resistance headphones.

Aerials

By this time the intending constructor will be able to appreciate the necessity of a good aerial. It is a point which cannot be over-emphasised. Assuming that you are in the fortunate position of being able to erect an outdoor aerial there are two main considerations, height and length.

One of the best that can be used is the inverted L shown diagramatically in Fig. 9.

It should be erected as high as is practical, every foot counts. The horizontal wire, that is the aerial proper, should have a minimum length of 60′ to which of course the length of the down lead is added. Where it is impossible to erect an aerial with an ideal horizontal length, a compromise must be effected.

Fig. 10 shows a three wire spreader aerial which gives quite a good effective length.

Suitable wire for a receiving aerial will not set any problems, stranded copper about 7/22 gauge is the best. 7/22 means that it consists of 7 strands of 22-gauge wire. This wire may be obtained covered, and for the present purpose is better than the plain or enamelled kind.

Note that insulators are used between the actual aerial wire and its anchoring supports, it is important that these are used, otherwise leakage will occur which will of course spoil its efficiency. Fig. 11 and 12 show how the wire may best be attached to two of the most common types of insulator available. Fig. 11 is of porcelain and is usually referred to as an egg insulator, whereas Fig. 12 shows a more modern (and more expensive) type in glass. If you are using the glass pattern, one is usually sufficient at each end of the aerial, but with the egg type two should be used. No doubt many readers will not be in a position to erect an out-door aerial, and must necessarily be content with an indoor installation.

The next best thing to a good out-door aerial is a replica constructed in a loft. If this form of construction is used, care must be taken when feeding the down lead to avoid close contact with the wall of the house. At the point where the lead feeds under the eaves, a length of rubber tubing can be used to cover the wire. The lead is fed through one of the small spaces left for ventilation purposes. A general idea is given by Fig. 13.

A less elaborate but quite effective aerial can be obtained by using a bed-spring. A length of 7/22 copper insulated wire is connected between the receiver and the spring. The spring should first be cleaned with emery cloth. Remove about 2″ of insulation from the wire and bind it tightly round the prepared spring. The joint may be covered with insulating tape which is obtainable from all electrical stores for a few pence.

Such aerials are quite popular since so many crystal sets are built for bedroom use.

When an aerial is required in the living room, the picture rail can be conveniently used. Insulated screw-eyes are fixed at intervals of about 3 feet along the rail, the wire is firmly anchored to one of them, and stretched right round the room until you arrive back at the starting point. By use of one of the insulated screw-eyes, the wire is secured and the down lead fed to the receiver. The idea is illustrated in Fig. 14.

Soldering

Before leaving the subject of aerials, a few words on soldering will not come amiss. Down leads on outdoor aerials should be soldered, and the same applies to loft types. Apart from aerial leads; earth leads and all the connections in the receiver will require soldering.

For the type of soldering necessary in radio construction an electric iron is the best solution. One of the small types marketed by Adcola or Henleys will be found admirable for the job. These are excellent for actual set construction but are hardly large enough for soldering the down lead to an outdoor type of aerial. Here the heat is dissipated much more quickly and a larger iron is required in order to get the solder to flow.

Ordinary irons which may be heated by a gas flame can be obtained very cheaply from most ironmongers' stores. Assuming the use of such an iron, first heat the iron until the copper bit is giving off a green coloured flame, the iron is now at the correct operating temperature. The bit is now discoloured or oxidised and its tip

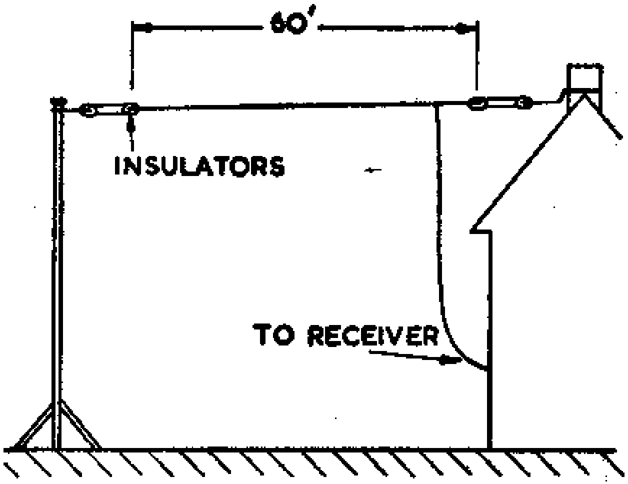
60'
INSULATORS
TO RECEIVER

INVERTED 'L' AERIAL

FIG. 9

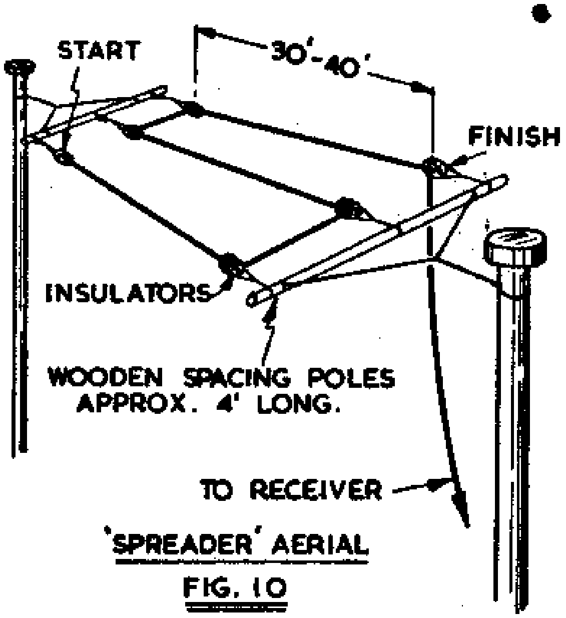
START
30'-40'
FINISH
INSULATORS
WOODEN SPACING POLES APPROX. 4' LONG.
TO RECEIVER

'SPREADER' AERIAL

FIG. 10

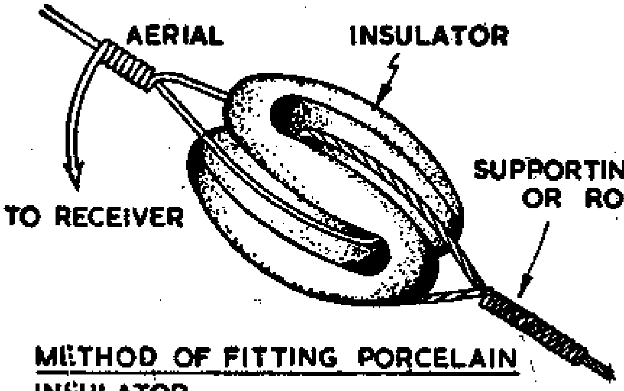
AERIAL
INSULATOR
TO RECEIVER
SUPPORTING WIRE OR ROPE.

METHOD OF FITTING PORCELAIN INSULATOR

FIG. 11

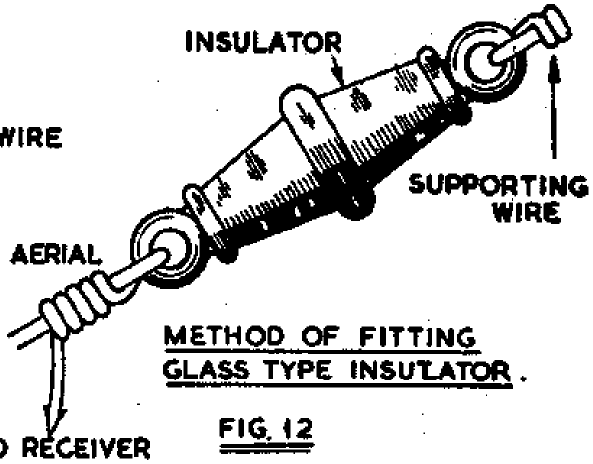
INSULATOR
SUPPORTING WIRE
AERIAL
TO RECEIVER

METHOD OF FITTING GLASS TYPE INSUTLATOR.

FIG. 12

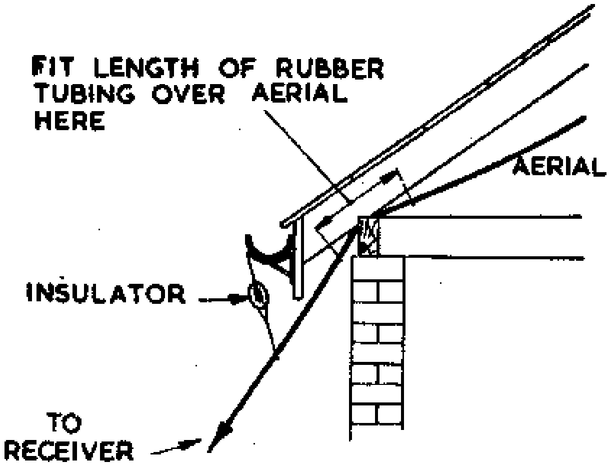
FIT LENGTH OF RUBBER TUBING OVER AERIAL HERE
AERIAL
INSULATOR
TO RECEIVER

LOFT AERIAL.

FIG. 13

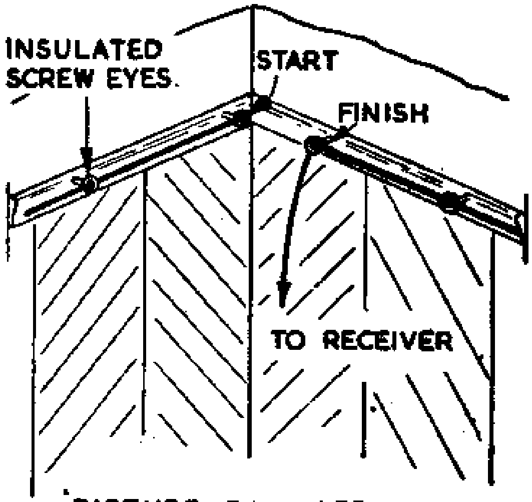
INSULATED SCREW EYES.
START
FINISH
TO RECEIVER

'PICTURE RAIL AERIAL.

FIG. 14

should be quickly cleaned with an old file which should be kept specially for the purpose. A better idea is to obtain a small block of sal-ammoniac and rub the tip of the hot iron on it. This will clean the tip of the bit perfectly. Next take a length of cored solder such as Ersin Multicore, and melt a little on to the prepared tip faces, now smooth it evenly over the surface with a piece of old rag (be very careful not to burn your fingers) the iron is now "tinned" and ready for use. When heating the iron, be very careful not to let the bit overheat or get red hot, otherwise the tinned surface will be destroyed and the whole process will have to be gone through again.

An electric iron will not overheat, and the tinning will last much longer than with ordinary types, and since the heat does not deteriorate there is less likelihood of making faulty joints.

Having obtained a tinned iron, the process of soldering joints is remarkably simple and anyone with a little patience can acquire the art in a very short time.

To solder two copper wires together: clean the wire with emery cloth, apply a prepared iron and some cored solder to it, the solder will flow evenly over the wire thus tinning it. Repeat the process with the remaining wire and then twist the two together. Now apply the iron to the joint from the underside and the cored solder to the joint on top. Solder will flow evenly over the joint. Remove both iron and solder and allow to set. The solder will harden or set in a few seconds but during this period the joint must not be touched or moved, as otherwise the joint will be "dry" and quite useless mechanically or electrically. On radio components, tags for soldering are already tinned though if they are old, or discoloured it is best to re-tin them. Normally, however, it is only necessary to twist the connecting wire to it and apply the iron and solder as explained.

Remember: never apply the solder to the iron and then the iron to the joint, always apply the iron and the solder to the joint. It is however a good thing to apply a little solder to the iron tip even when it is perfectly tinned just before making a new joint.

One final "DON'T," you will have noticed that the solder referred to is "cored," that is, it contains resin and other substances through its centre, plain solder as used by electricians and plumbers will not do, as it is the resin or flux

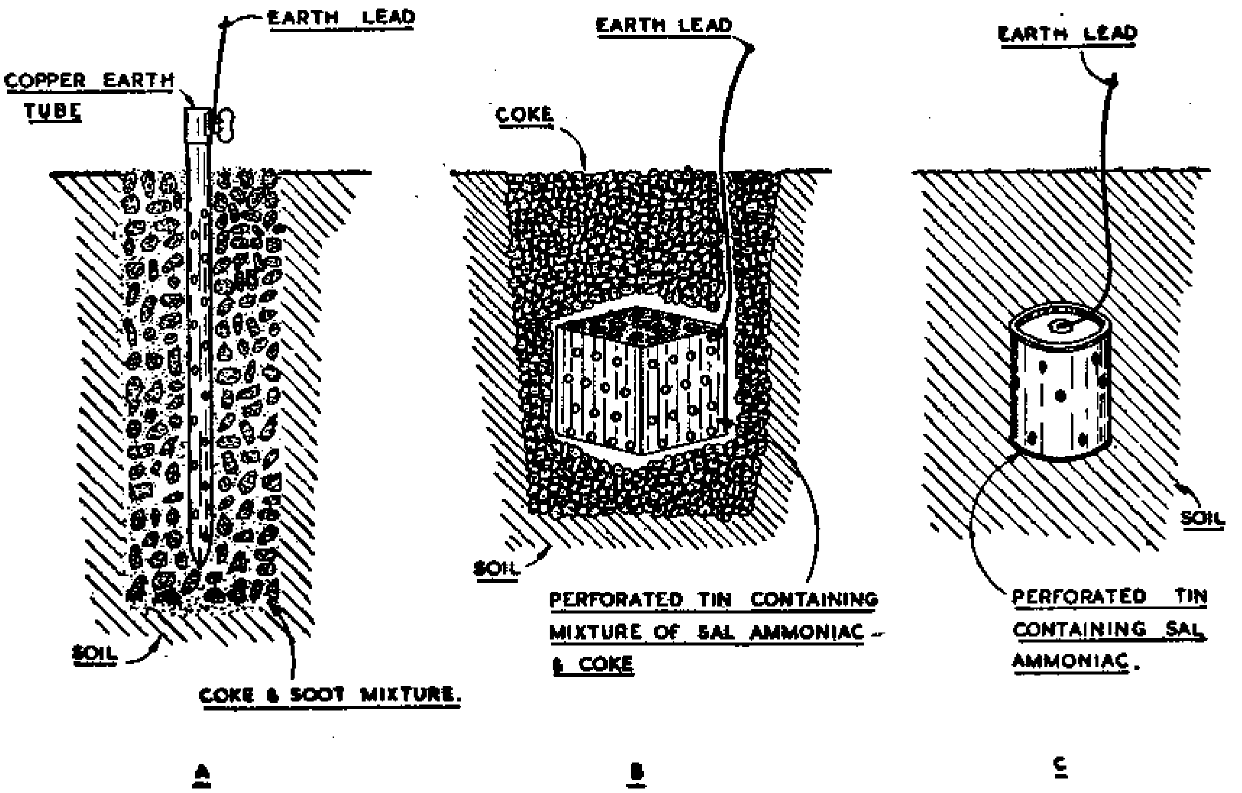
EARTH LEAD
COPPER EARTH TUBE
SOIL
COKE & SOOT MIXTURE.
A
EARTH LEAD
COKE
SOIL
PERFORATED TIN CONTAINING MIXTURE OF SAL AMMONIAC & COKE
B
EARTH LEAD
SOIL
PERFORATED TIN CONTAINING SAL AMMONIAC.
C

FIG. 15.

as it is called which makes the solder flow evenly and permits a good electrical joint.

Equally important do not attempt to use a separate flux or soldering fluid with plain solder or to "help" the cored solder, since these are almost certain to cause eventual corrosion and will destroy any components that have been contaminated.

Earths

The provision of a good earth is just as important as the rest of the installation.

Much disappointment would be avoided if this fact were not lost sight of, to avoid any slip up in this direction I am proposing to outline several well tried and efficient earth systems.

If it can be obtained, an earth rod, specially designed for the purpose provides the basis of a good earth connection. These were very popular some years ago when the majority of receivers were either crystal or battery operated, but in these days of modern mains driven sets they are not often used and consequently not always readily available.

Fig. 15 shows how it is used. First excavate a hole to a depth of three feet and fill with a mixture of soot and coke. Drive in the earth rod, which is a hollow copper tube, perforated. The lead to the receiver is connected at the top, and should be of covered 7/22 gauge, as used for the aerial system. Though not obvious it is important that this lead-in should be in covered wire, otherwise a number of indifferent earth contacts are likely to be made at various points along its length until it reaches the set. This is very undesirable and will spoil the efficiency of the system. It is essential to keep the soil surrounding the rod moist, which is one reason why the tube is hollow, and care must be taken to pour water into the tube at intervals.

A very efficient earth is the percolative type also popular at one time. Due to the chemicals used, it will extract moisture from its surroundings, thus maintaining a permanently moist earth. It should be installed as in Fig. 15b.

The container is of copper or zinc, anything else will quickly rust away, again a good quantity of coke is used. Fill the container with a mixture of sal-ammoniac and coke, and then bury in coke as illustrated. This earth will not require further attention. It is possible to use powdered calcium chloride instead of sal-ammoniac but unless a chemist can be persuaded to make some up it is better to stick to sal-ammoniac.

If a zinc container is not available and you do not use or cannot easily get coke, Fig. 15c should be used. This makes a better earth than many so-called "earths" that I have come across. Obtain as large a tin as possible, make a number of holes as shown. Solder the lead in the bottom and fill it with sal-ammoniac. Replace the lid and bury in the ground. The tin will eventually rust away, but the replacement cost is negligible.

If it is quite impossible to make direct contact with the ground, a water pipe must be pressed into service. This should be a main pipe feeding straight to ground and not a hot water pipe or one fed from a tank. Scrape the pipe clean and twist the lead in tightly around it, a copper clip is even better. Do not attempt to solder on to the water pipe; since cold water is flowing through the pipe, it is extremely unlikely that your soldering iron will heat up the water supply sufficiently to allow a sound electrical joint, though you may spring a leak.

No attempt should be made to utilise gas pipes; the possibility of causing a fire is certainly very remote but, they make incredibly bad earth connections due to a number of joints made before true ground is reached. These joints are at best only semi-conductors, at least from an electrical standpoint.

Headphones

Since the late war there have been a large number of head-phone sets available on the surplus market. These may be roughly divided into two types, high-impedance and low-impedance. For the crystal sets detailed in this book high-impedance 'phones are required and the low-impedance pattern will not be suitable unless a matching transformer is used. As this is likely to cost more than the rest of the installation including the set, it will be as well to avoid them.

High-impedance types have an impedance of 2000Ω to 4000Ω whereas the low-impedance types are usually 600Ω.

The remaining consideration is weight; often cheap headphones are very heavy and uncomfortable to wear, every endeavour should be made to obtain 'phones as light in weight as possible.

Crystals

All the sets shown in this book have been designed to work with modern germanium crystals rather than the older galena crystal.

These germanium crystals require no adjustment, which, in itself, removes the main objection to this class of receiver. Suitable crystals are available from the following manufacturers and on the surplus market: Mullard, G.E.C., Brimar, Westinghouse and B.T.H. They are of robust construction, some in glass, others in ceramic or plastic, but must not be subjected to heavy knocks, otherwise the contact point may become dislodged.

From the information given, you should now be able to install an excellent aerial and earth system, and have some idea of how the crystal set works, so it is time to pass on to the actual receiver construction.

CONSTRUCTION 1

If you examine the following circuits you will find that each one is different. In most cases the difference lies in the coil design and/or the method by which the crystal and aerial is tapped into it. Each of these circuits has its own particular advantage to suit different conditions and the ideal circuit in some localities is not necessarily the best in others. It is not just a matter of a given circuit giving louder results than another, if it were there would be no point in showing more than one.

The main problem is to obtain adequate selectivity without reducing the volume level.

A receiver is said to be selective when it tunes sharply, a set with poor selectivity allows the stations to spread over the dial and when used near a transmitter will receive the local stations mixed together, which of course is useless.

Consider Fig. 16a, this is a very simple receiver, with no special attempt to provide any great amount of selectivity. In areas where signal

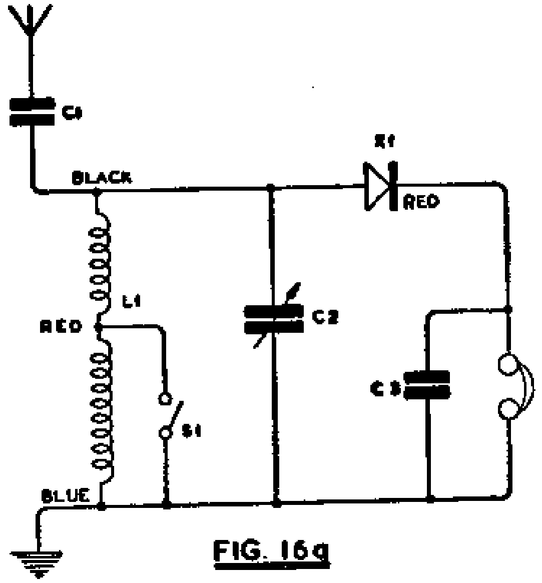

FIG. 16a

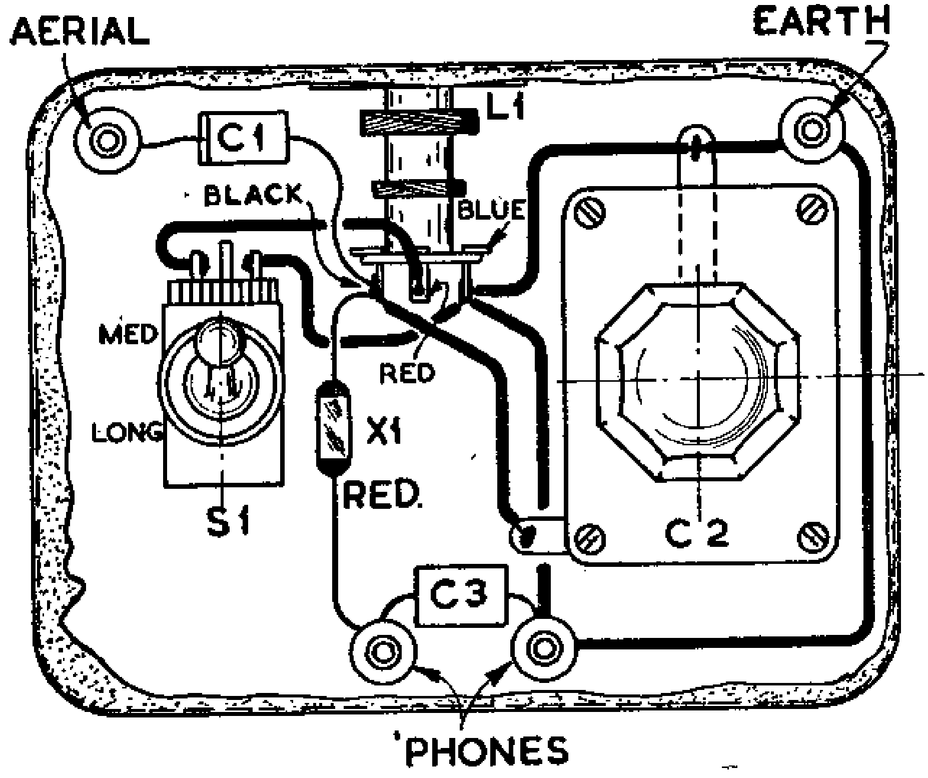

LID SHOWN BROKEN AWAY TO REVEAL POSITION OF COMPONENTS.

FIG 16b

strength is not high, or a short aerial is used, it will probably be ideal.

It would have been quite easy to increase the selectivity by providing a tap on the coil for the aerial as in Fig. 17a, but unfortunately, as the selectivity is increased overall volume is likely to decrease, so that unless you live close enough to the transmitter to have a large signal available and therefore need the selectivity, the circuit of Fig. 16a will be quite satisfactory.

Capacitor C1 is to prevent the aerial damping the circuit too heavily because this would flatten the tuning unnecessarily, however with small aerials it may be better to take the aerial direct to the black tag on L1. The signals are selected or tuned by L1 and C2, X1 is the crystal and C3 the capacitor across the 'phones to prevent unwanted carrier or R.F. reaching the 'phones.

A practical diagram (Fig. 16b) is provided showing the layout and all the wiring. You will require nuts and bolts to fix the coil, about ½", 4BA size will do, the other parts have locking nuts provided. An old 2-oz. tobacco tin makes a very good container and keeps the size down. Note that the metal box is connected to earth. With the exception of AC/DC receivers the metal work on any receiver or amplifier is connected to earth.

When S1 is open as in the diagram, the set will tune in long-wave stations, but when closed the medium wave-band will be received.

Components List, Fig. 16a

C1	100pF mica capacitor.
C2	500pF tuning capacitor, solid dielectric.
C3	1000pF mica capacitor.
X1	Germanium crystal.
L1	Crystal Set Coil. R.E.P.
S1	Single Pole toggle switch.
4	Insulated wander-plug sockets and plugs.
1	2-oz. tobacco tin (or similar container).

Make sure that the wander-plug sockets are of the insulated type, otherwise the metal case will join all the sockets together electrically.

Try to follow the theoretical diagram when wiring, a little practice will soon enable you to wire up a set without a practical diagram, which is a great advantage because often only the theoretical diagram is given when circuits are detailed in the technical press.

Fig 28 on page 34 gives a list of symbols used on the theoretical diagrams, so that you can readily identify the components.

2

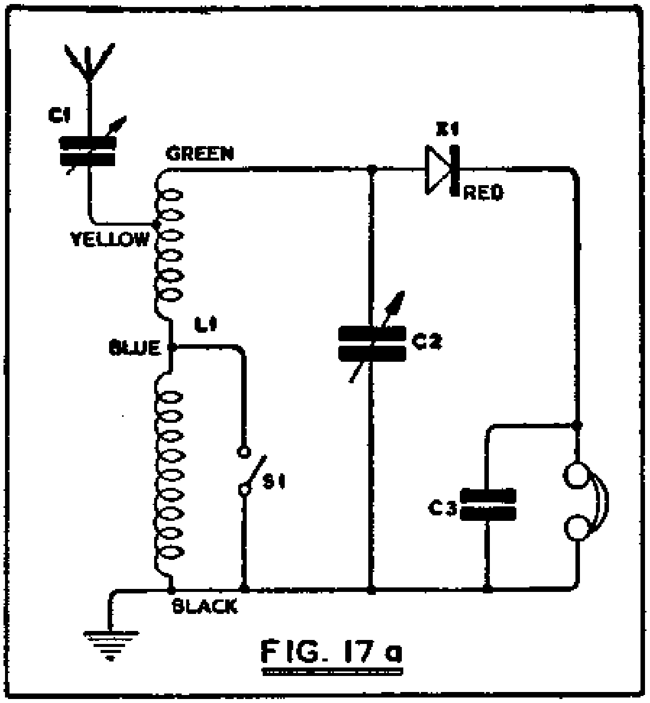

FIG. 17 a

Fig. 17a circuit is similar in many respects to that of Fig. 16a. The difference is purely one of selectivity, tuning will certainly be sharper, and even with comparatively inefficient aerials the design will put up a very good performance. There is no reason why the construction should not follow the same lines as the previous receiver but, by enlarging the set a little and using an air spaced tuning capacitor, the efficiency is improved. Observe that the colour coding on the coil, an R.E.P. Dual Range (Blue Box) is different from that of the Crystal set coil used on the previous design, and make sure it is correctly wired in.

C1 is a mica compression capacitor, often referred to as a "trimmer" and because it is adjustable, permits the set to be matched to aerials of varying lengths. Construction is carried out on a square panel of bakelite, perspex or wood, as shown in Fig. 17b. Perspex being clear like glass presents a most attractive finish, provided the set is neatly wired. It must be drilled slowly, however, otherwise the generated heat of the drill will make the hole wander, and spoil the panel.

After the set has been wired and tested a small wooden container can be made to house the completed set.

When carrying out reception tests, a little experiment is well worth while, to get the best out of the set. Try removing the germanium crystal from the green tag on L1 and connecting it to yellow, at the same time removing the lead from C1 to yellow and connecting it to green. Once the best arrangement has been found the wiring

can be left permanently in that position. Note that two tags on the coil are unused, this is intentional because on this circuit the extra winding connected to these tags is not required. With the switch S1 closed, the coil will cover the medium wave-band, and when open, long wave stations can be received. It is customary to abbreviate the expression medium wave-band to M.W. and long wave-band to L.W.

Components List, Fig. 17a

C1	100pF mica trimmer capacitor.
C2	500pF variable capacitor (air spaced).
C3	1000pF mica capacitor.
L1	Dual Range Coil R.E.P. (Blue Box).
S1	Single Pole toggle switch.
X1	Germanium Crystal.
4	Terminals or wander-plugs and sockets. Perspex, bakelite or wood for mounting panel.

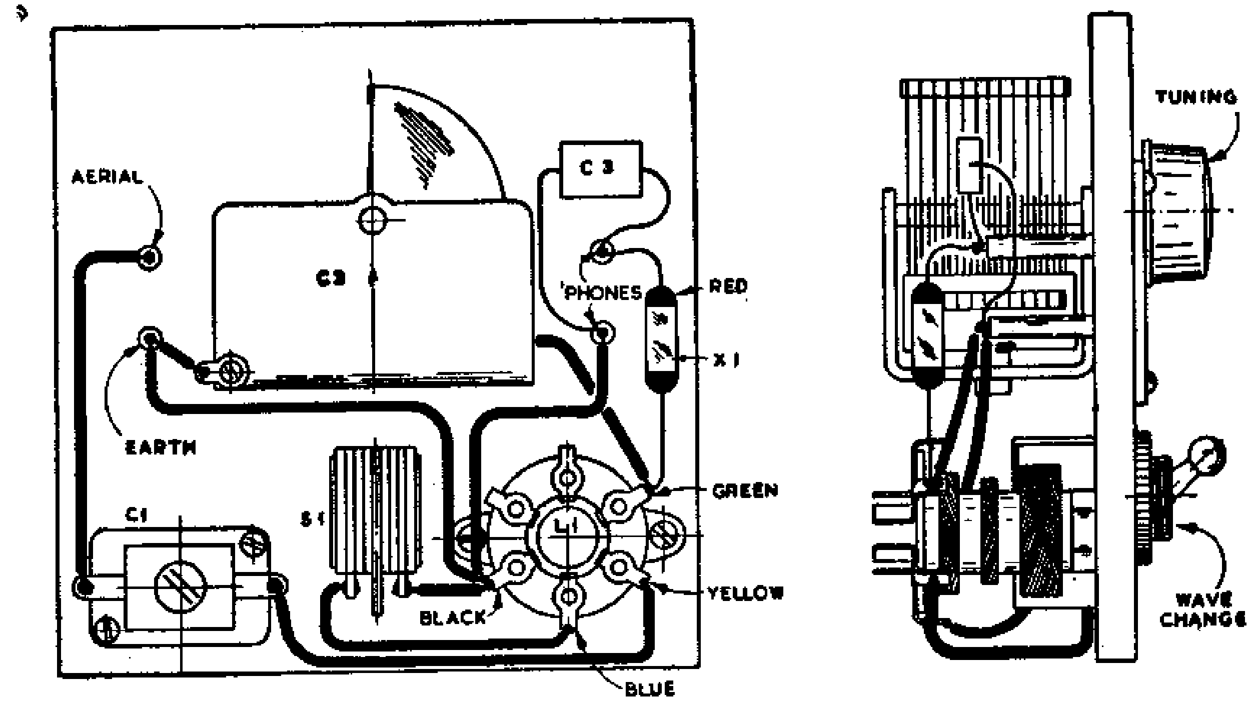

FIG. 17b

3

The design shown in Fig. 18a and 18b is more flexible than either of the preceding sets. That is, it can be varied to suit local conditions to a greater extent than the first two sets. This is made possible by the large number of taps provided on the coil.

You will notice that in this instance the coil is home-constructed on a cardboard former and is much larger than commercially produced coils. Coils wound in this manner are called **solenoids**, the turns are wound on side by side; in this case to a depth of some three inches. The diameter of the coil is made purposely large so as to obtain high efficiency. Modern commercial coils are invariably wave-wound and quite often are litzendraht which is the German for litz wire. This wire is made up of a number of strands of fine copper wire, each strand is enamelled to insulate it from the others, the whole is then silk covered. Litzendraht is more efficient than solid copper wire, and by using this and wave-winding, manufacturers can produce an efficient coil which is also small. Unfortunately wave-wound coils cannot be produced without a complex winding machine and home-constructed coils must take the solenoid form. It would be possible to use litzendraht but it is not easy to obtain by the reel, it is most expensive, and is difficult to handle. This last point is because at termination points, each strand must be cleaned of its enamel before a joint is made, and if one strand is broken, its advantage over plain copper wire is lost. How-

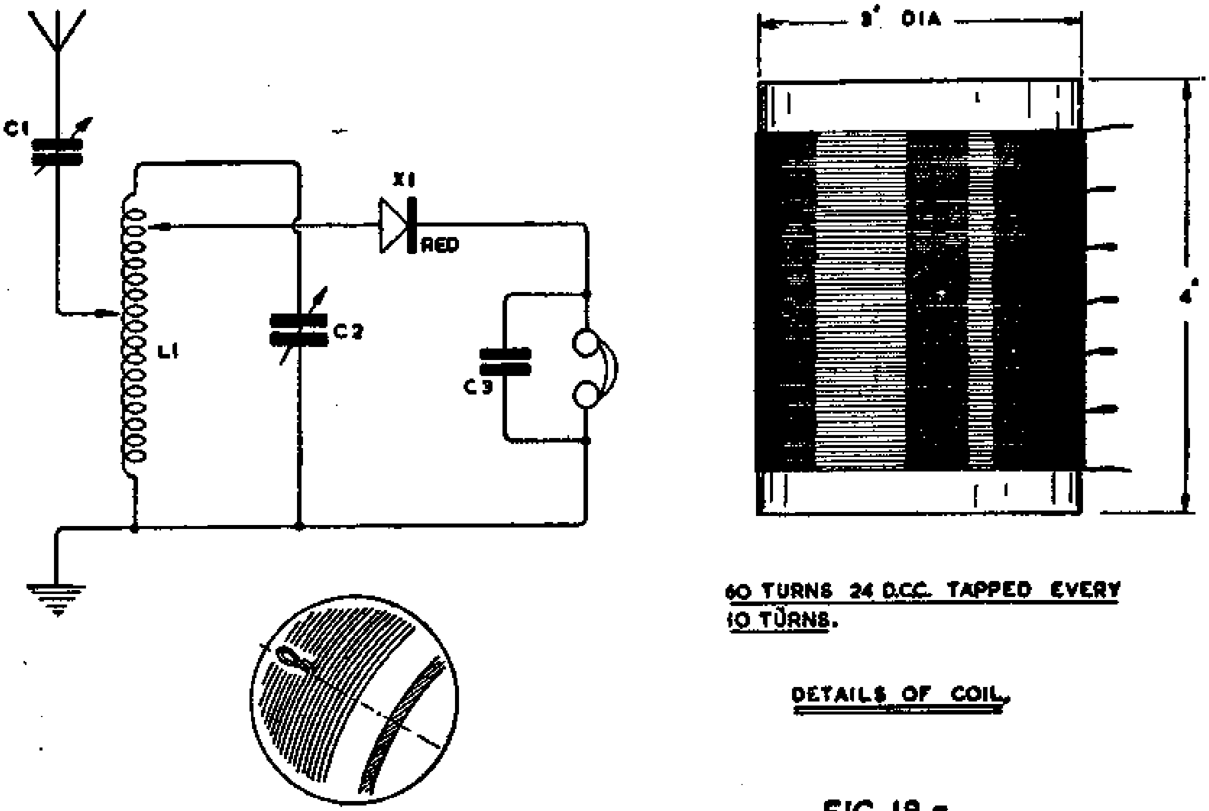

60 TURNS 24 D.C.C. TAPPED EVERY 10 TURNS.

DETAILS OF COIL.

INSET SHOWING METHOD OF TAPPING COIL.

FIG. 18 a

ever, by using a former of reasonably large dimensions an efficient coil can be produced cheaply using ordinary copper wire.

The wire must be firmly secured at the start and finish of the windings. Pierce three small holes about ¼″ apart ½″ from the end of the former. Pass the wire through the first from the outside, return it through the second and pass it back again through the third. Leave some 6″ of wire at the end to make off the connection. It will now be possible to wind the turns on tightly without wire slipping.

Count on ten turns and make a loop 1″ long. The method of preparing loops is shown in Fig. 18a, loops or taps should always be made in this manner, never by baring the wire and soldering a further length of wire to it. Carry on with the winding, making off the taps every ten turns as directed until the coil is complete.

Construction is carried out on a wooden base-board and front panel. Secure the coil to the base-board as shown on Fig. 18b, mount the remaining components and proceed with the wiring.

When testing out, it will be found that the further the aerial is tapped down the coil towards the earth end, the greater the selectivity. A position should be found which permits separation of local stations without excessive loss of sensitivity.

The crystal tap is adjusted for best results and different settings of C1 tried out, when choosing the best position for the aerial tap.

Components List, Fig. 18a

C1	450pF padder.
C2	500pF variable capacitor.
C3	1000pF mica capacitor.
L1	See text and Fig. 18a.
X1	Germanium Crystal.
2	Crocodile Clips (to connect leads to coil taps).
4	Wander-plugs and sockets.

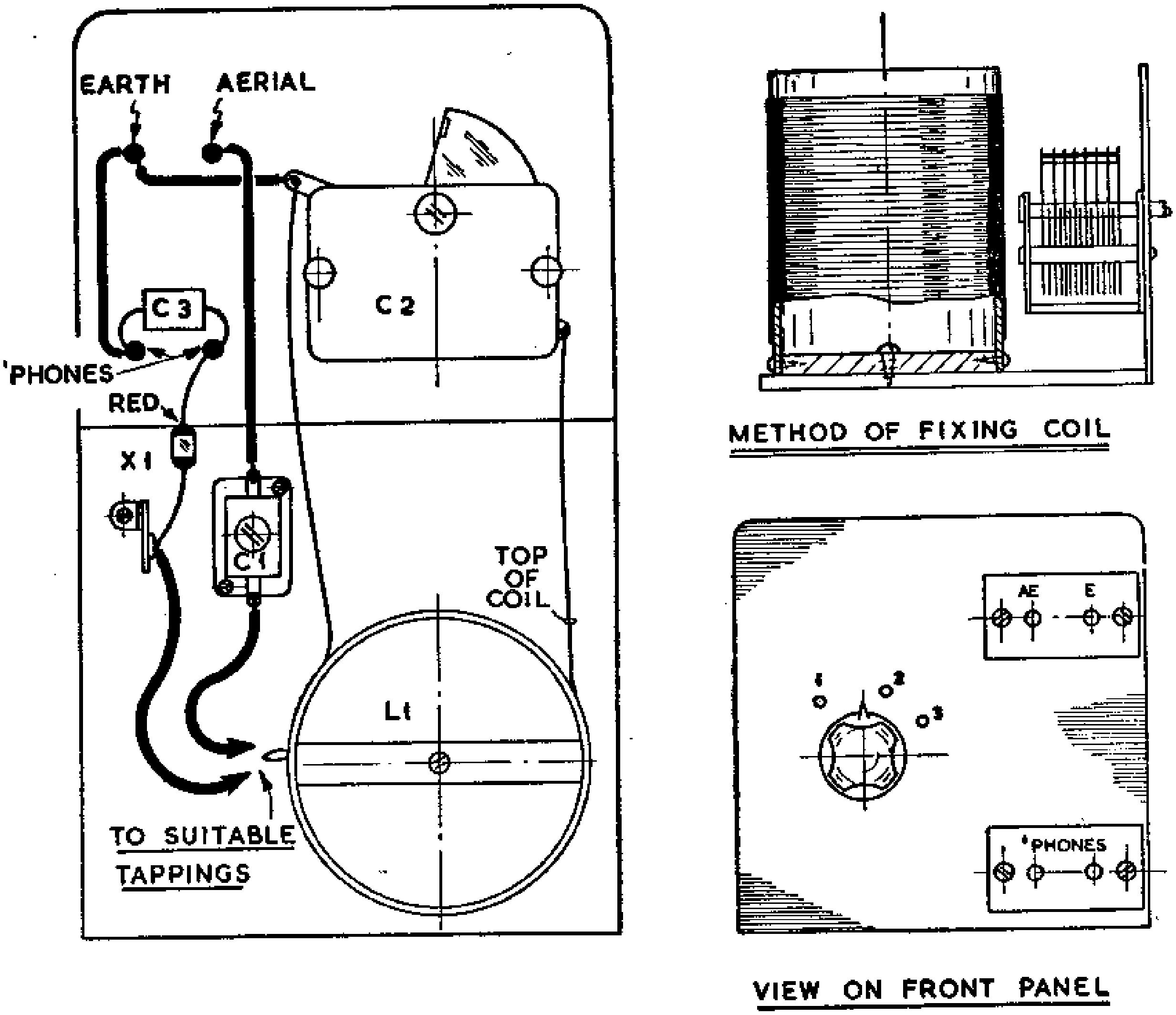

FIG. 18 b

4

A most unusual design is reproduced in Fig. 19a.

It first appeared in America about three years ago, and it certainly does offer some advantage over more conventional sets. The tuning circuits are duplicated and two crystals are used, so the circuit may be described as a full-wave receiver. Signals received will certainly be louder than with simple sets, but care must be taken with the coil winding, as with the other receivers using home-made coils, otherwise results will be disappointing. Be very careful to ensure that all three windings are in the same direction, this is very important.

The distance between each winding should be ⅛". After the coil has been wound it is a good plan to warm it before a fire and paint the windings with "Durafix." Heating the coils makes the "Durafix" run freely. This substance sets quite hard and there will be no risk of the windings loosening. A small quantity of enamelled wire will cover the requirements of this coil, a 2 oz. reel will provide more than sufficient.

Before wiring in the germanium crystals, examine them carefully, note that one end is coloured red or in some cases marked with a positive sign thus +. You will notice, the sign is the same as the addition symbol used in arithmetic.

It is essential that both the red or positive ends are connected together, note that this is clearly marked on Fig. 19b. The receiver cannot work if

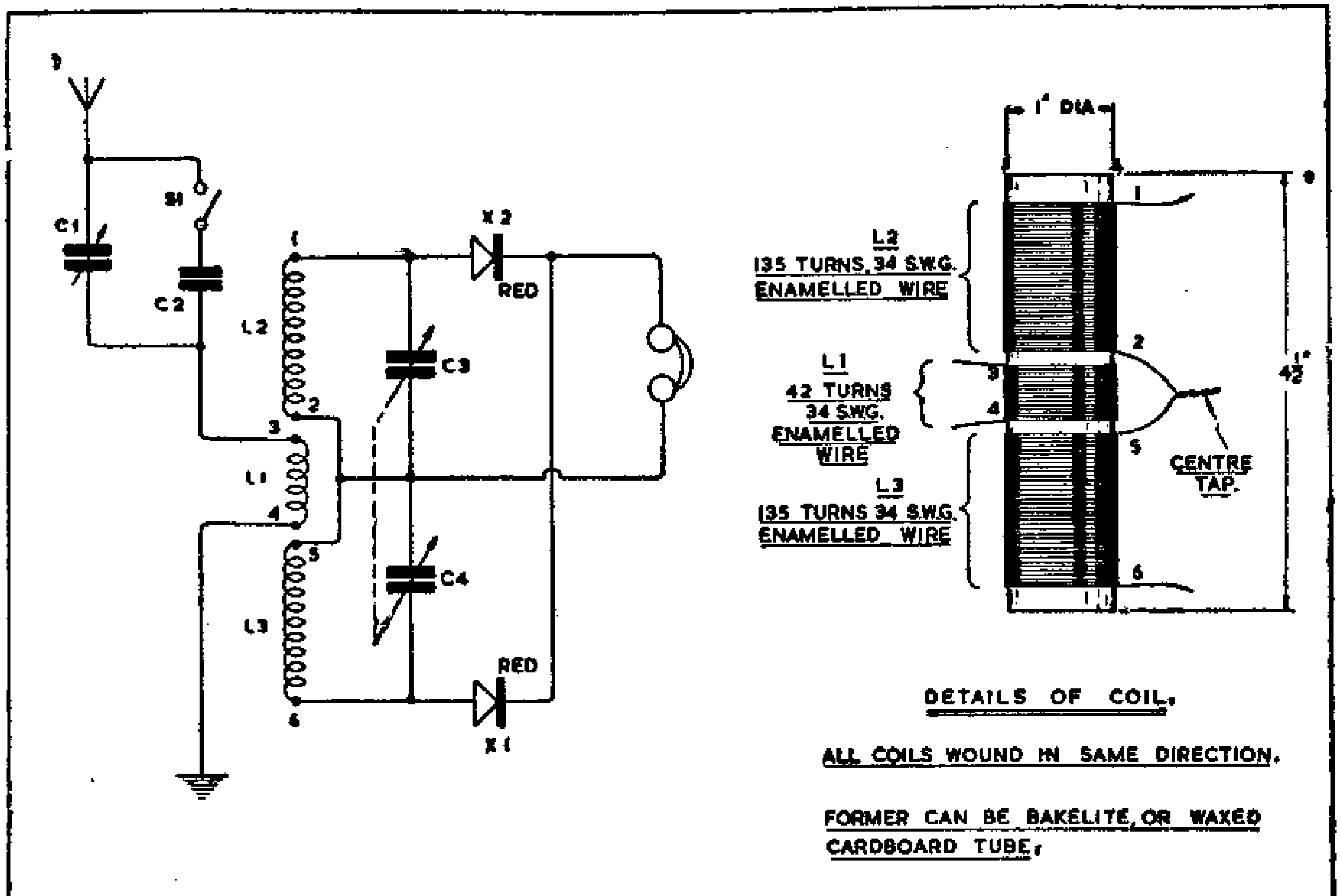
C1
S1
C2
L2
1
2
3
L1
4
5
L3
6
X2
RED
C3
C4
RED
X1
1" DIA.
L2
135 TURNS, 34 S.W.G. ENAMELLED WIRE
L1
42 TURNS 34 S.W.G. ENAMELLED WIRE
L3
135 TURNS 34 S.W.G. ENAMELLED WIRE
1
2
3
4
5
6
CENTRE TAP.
4½"

DETAILS OF COIL.

ALL COILS WOUND IN SAME DIRECTION.

FORMER CAN BE BAKELITE, OR WAXED CARDBOARD TUBE.

FIG. 19a

one of the crystals is connected the reverse way round.

A baseboard and panel form of assembly is used for this set which looks quite attractive if housed in a small polished cabinet. Dimensions are not at all critical but Fig. 19b gives a general idea of the layout that should be used. To avoid any unnecessary losses the coil should be mounted on perspex.

Obtain two strips of perspex 5½" long by ½" wide, drill fixing holes at both ends of each strip. It is best to clamp the two together when drilling these holes. Place one strip on the board in position, then put the coil and remaining strip over the first one and screw down. The coil will be clamped neatly and rigidly into position.

Twin socket bakelite strips are used to carry the aerial/earth and 'phone connections, if desired terminals mounted on pieces of bakelite could be used. These socket strips can be purchased from most shops selling components. If feet are not provided, drill two holes and screw into the edge of the base-board. Tag-strips are used to anchor some of the wires from the coil, and the germanium crystals. These also can be purchased for a few pence. At least two tags are necessary on one, and three on the other, though if the strips have more tags than required it is of no consequence. C3-C4, the ganged tuning capacitors have trimmers fitted, this should be stipulated when purchasing, and when first operating the set these small trimming capacitances are set about half-way. The idea of these trimmers is, that they compensate for any difference between the self-capacity of the tuning coils. With the switch S.1 open, that is in the OFF position, two

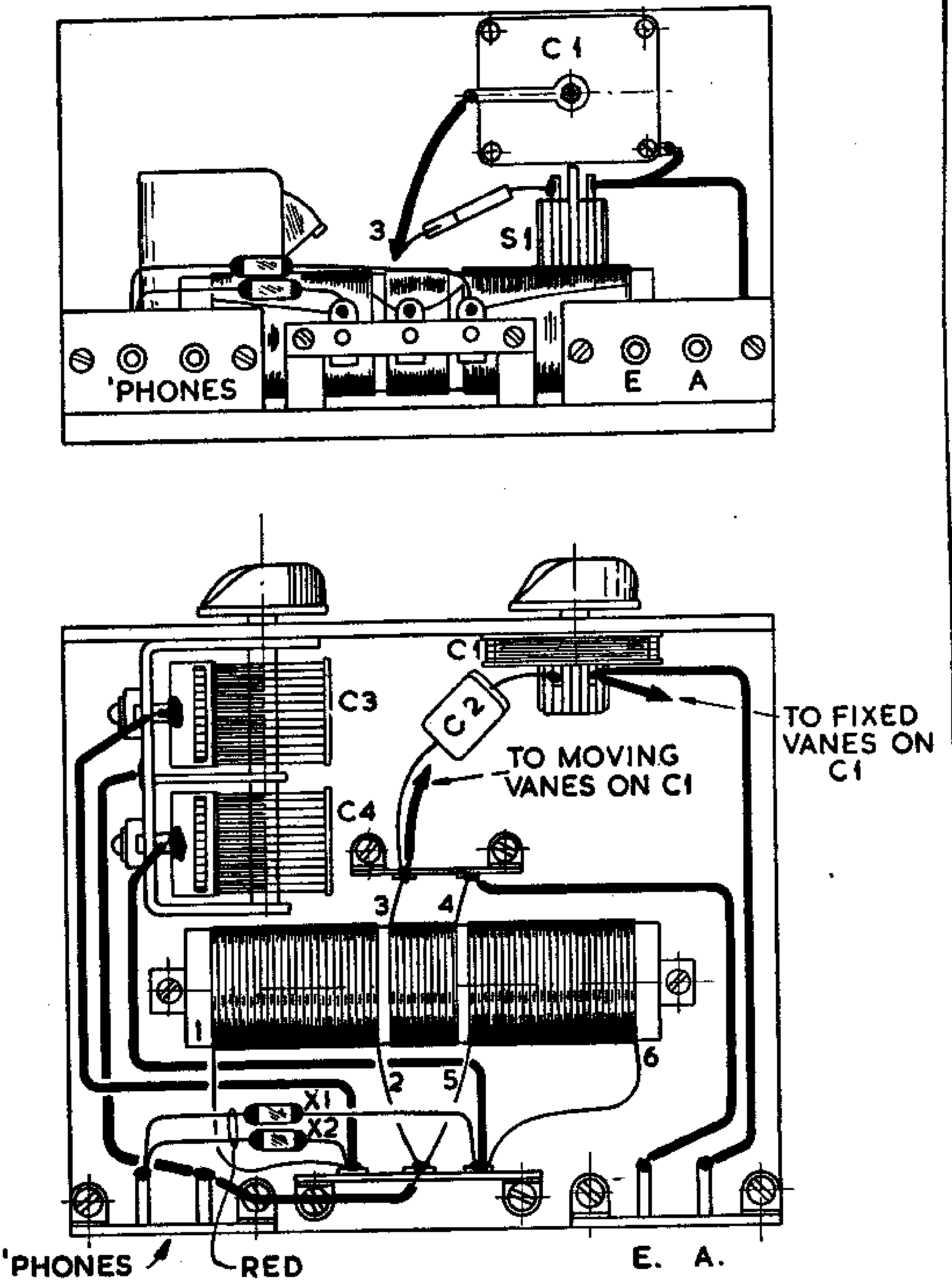
C1
3
S1
'PHONES
E
A
C1
C2
C3
C4
TO FIXED VANES ON C1
TO MOVING VANES ON C1
3
4
1
2
5
6
X1
X2
1
'PHONES
RED
E.
A.

FIG. 19 b

in a station at the high-frequency end of the band, that is with the vanes of C3-C4 towards the disengaged position. Adjust C1 for maximum volume, without interference from other stations. Now adjust the trimmers for loudest headphone strength. It may be necessary to run over the adjustments several times for best results. Once they are correctly set, the trimmers require no further adjustment. The tuning coil covers the M.W. band, 1500 kc/s. to 600 kc/s., as the capacitor C3-C4 is advanced and the vanes start to mesh, the frequency to which the receiver tunes decreases so that with the vanes disengaged the set is tuned to 1500 kc/s. approximately and at full mesh, 600 kc/s. Switch S1 is used as follows, for stations lower in frequency than 850 kc/s it should be in the ON position with the contact closed, but for stations higher in frequency, it is in the OFF position. As an example both the London Home Service and London Light Programme transmitters on the M.W. band operate at a higher frequency than 850 kc/s. If any doubt exists as to the operating frequency of a given station, reference should be made to the Radio Times which quotes both wave-lengths and frequency. If you know the wave-length in metres, it is simple to find the frequency. Divide 300,000 by the wave-length in metres, the dividend equals the frequency in kilocycles, i.e., 300,000 ÷ 300 metres=1,000 kc/s. In the same way, dividing 300,000 by the frequency in kilocycles will produce the wave-length in metres.

To use the receiver, tune in the signal by C3-C4, adjust C1 for maximum volume without allowing stations to overlap.

Components List, Fig. 19a

C1	300pF solid dielectric variable capacitor.
C2	1000pF mica capacitor.
C3-4	500pF twin gang variable capacitor.
L1-2-3	See text and Fig. 19a.
X1-2	Germanium Crystal.
2	Tag-strips.
2	Twin socket strips.
S1	Single Pole toggle switch.

5

Whenever crystal sets are discussed, constructors are apt to think in terms of medium-wave reception. When you consider that at least 98% of published circuits are designed for this band (sometimes with the long-wave band thrown in as an afterthought) it is understandable. However, if you can provide a good outdoor aerial (and an equally good earth) there is a lot of fun to be had listening to the short-wave bands. This receiver is designed specially for short-wave reception.

A metal chassis is used for construction, these can be obtained ready made in aluminium from most good supply houses. It can be quite small and on the original model a 6" x 4" was used. Low loss components are used so as to obtain greatest efficiency on the short wave bands. The coil is a commercial product which plugs into a 4-pin base. Best results were obtained on the 3mc/s to 7mc/s band though coils covering other bands are available.

Note that the tuning capacitor C1 is smaller than normally used and has a maximum capacitance of 140pF.

Coil type 706/R covers the 3mc/s to 7mc/s band, but the set is equally satisfactory on the M.W. band, and to cover this coil type 706P should be used. This coil has an iron dustcore which can be adjusted to make the coil cover the required band. The effect of the core is as if turns were being added or removed from an ordinary coil. When dealing with the short waves it is customary to refer to the frequencies in terms of mega-cycles (mc/s) rather than kilo-cycles. One mega-cycle is equivalent to 1000 kilo-cycles, that is 1,000,000 cycles.

Layout and wiring can be clearly followed from Fig. 20b and a theoretical diagram is shown in Fig. 20a. Care should be taken when arranging a mounting for the aerial terminal. The best plan is to cut a 1" hole in the chassis, mount the aerial terminal on a piece of perspex and screw into the chassis. This avoids any unnecessary losses between aerial and chassis. The earth terminal is screwed directly into the chassis.

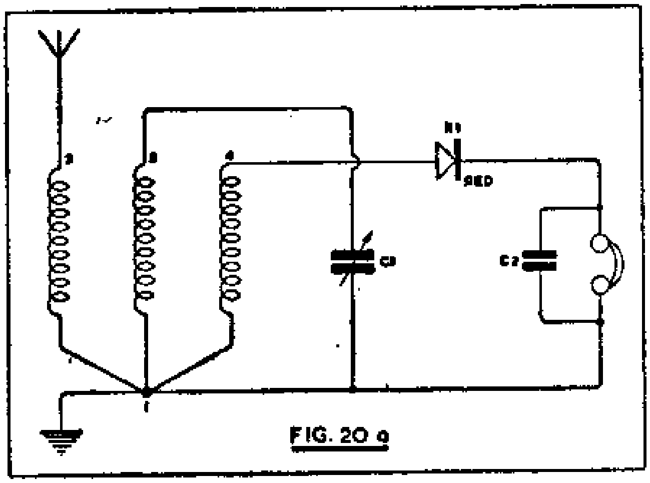

FIG. 20 a

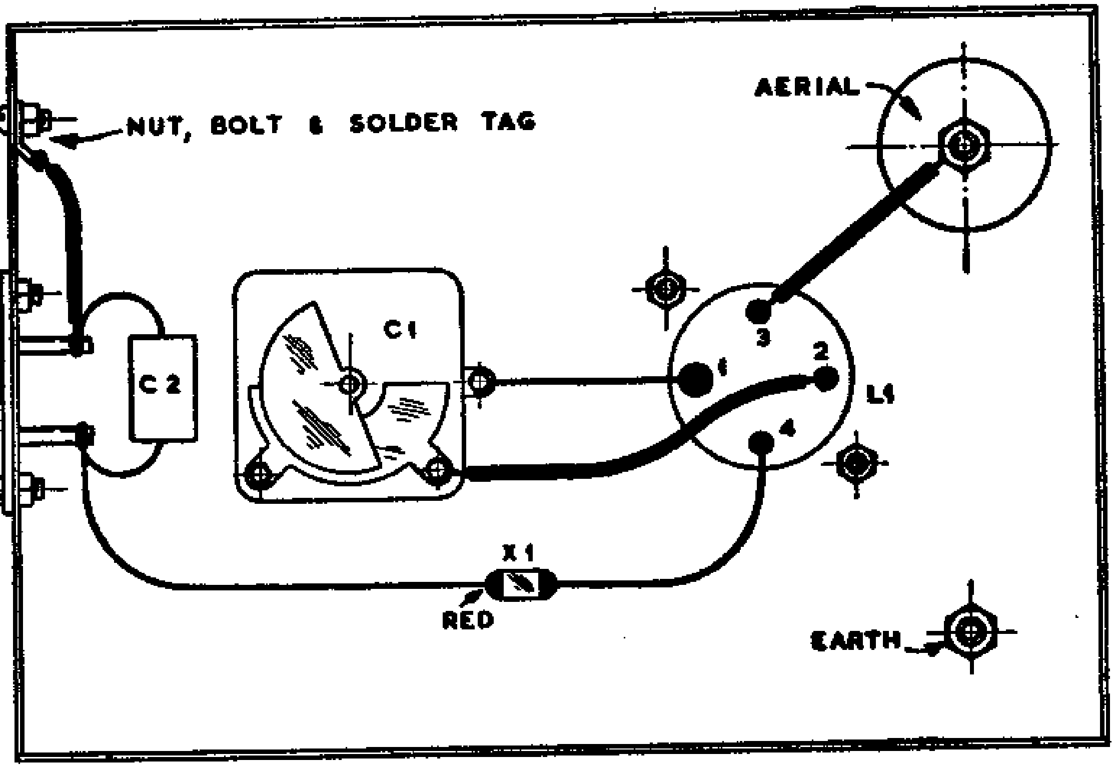

FIG. 20 b

Wiring as shown in Fig. 20a is likely to give the best results, but other arrangements are possible and to get the best out of the set they should be tried.

Refer to Fig. 20a. The aerial is connected to pin 3 on L1, C1 to pin 2 and X1 the crystal to pin 4. If you have a long aerial try this combination—Aerial to pin 4, C1 to pin 2, and the crystal to pin 3. Connections to pin 1 are not altered. In a few cases the following arrangement may prove best. Crystal and C1 to pin 2, aerial to pin 3, pin 4 left free and again pin 1 is unaltered.

Tuning on the short-wave bands is more critical than on the medium waves so tune very slowly over the band and remember signal strength is likely to vary from day to day.

Components List, Fig. 20a

C1	140pF variable capacitor Eddystone 586.
C2	1000pF mica capacitor.
1	Coil holder Eddystone 707.
L1	3m/c—7m/c coil Eddystone 706/R.
1	Engraved tuning dial.
1	6"x4"x2½" chassis.
2	Terminals (Aerial-Earth).
1	Twin socket strip ('phones).
X1	Germanium Crystal.

6

Fig. 21a-b is yet another design from America and appeared in Radio Craft some years ago, designed by Mr. W. J. Spain. The original used a silicon crystal which preceded the development of the germanium. Results are certainly very good though of course on the model built here, a germanium crystal was used. Selectivity can be adjusted to suit all conditions and even at short distances from local stations. Good volume can be obtained without the programmes overlapping. Home made coils are used and for best results a 75' outdoor aerial is desirable.

Baseboard and panel construction is used, which can conveniently be housed in a small wooden cabinet.

Operation is a little more complicated than some of the other designs, but the results certainly merit the extra trouble taken. As with the other receivers much of the success obtainable is due to the coils and these must be carefully made otherwise results will be disappointing.

Layout and wiring can be followed quite easily from Fig. 21b and requires no special comment.

First make the coils. L1 is a tapped coil, the taps should be made in the same way as shown in Fig. 18b. It consists of 90 turns of 22 D.C.C. (double cotton covered) copper wire tapped at the following number of turns, 5, 10, 15, 25, 30, 40, 50, 60, 70 and 80. The former used has a diameter of 2″. L2 is also wound on a 2″ diameter former, this coil however is not tapped, it consists of 110 turns of 38 D.C.C. or enamelled copper wire. Great care must be taken when handling this wire, it is quite fine and in the hands of the inexperienced will easily break.

After wiring has been carried out the set can be put into operation.

Selectivity is controlled by S2, in position B tuning is broad but in position A it can be quite sharp.

When searching for a station switch S2 to B. Tune the station and switch S2 to A. The frequency range is controlled by the position of the tap on L1, a good plan is to start with the 5 turn tap on L1.

Components List, Fig. 21a

C1	500pF variable capacitor.
C2	500pF variable capacitor.
C3	1000pF Mica Capacitor.
X1	Germanium Crystal.
L1 & L2	See text.
S1	Single Pole toggle switch.
S2	Single Pole 2-way toggle switch.
2	Twin socket strips (Aerial/Earth and 'Phones).

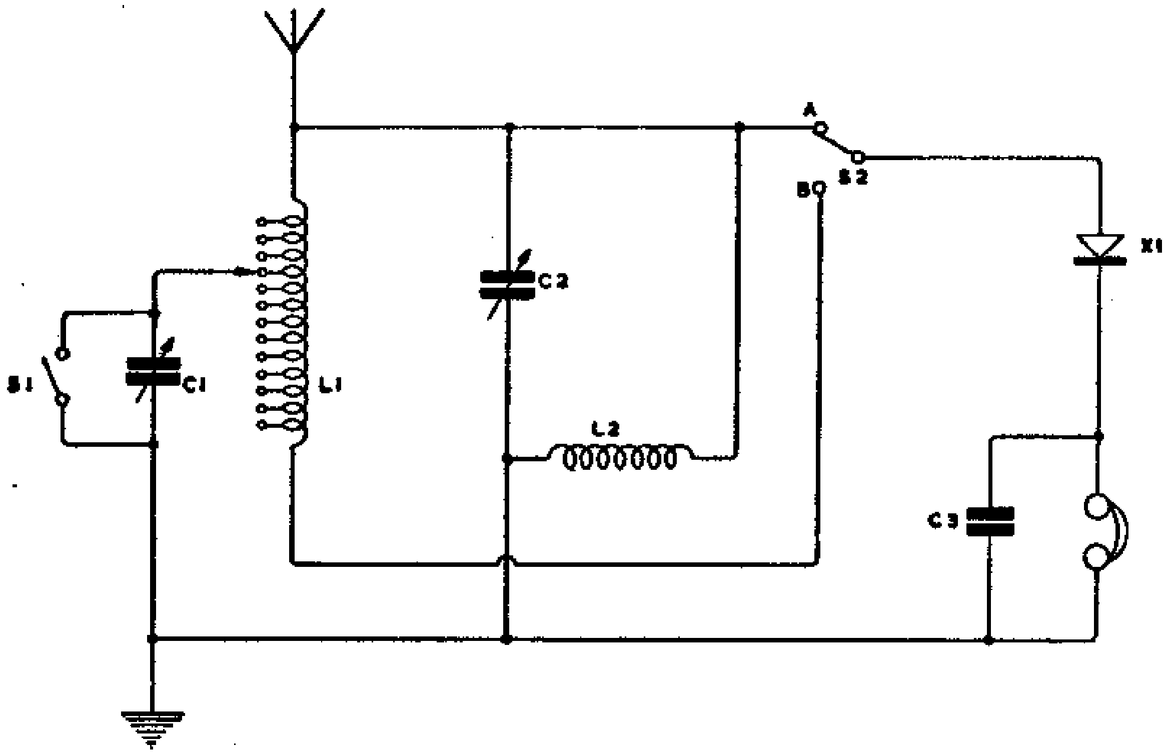

FIG. 21 a

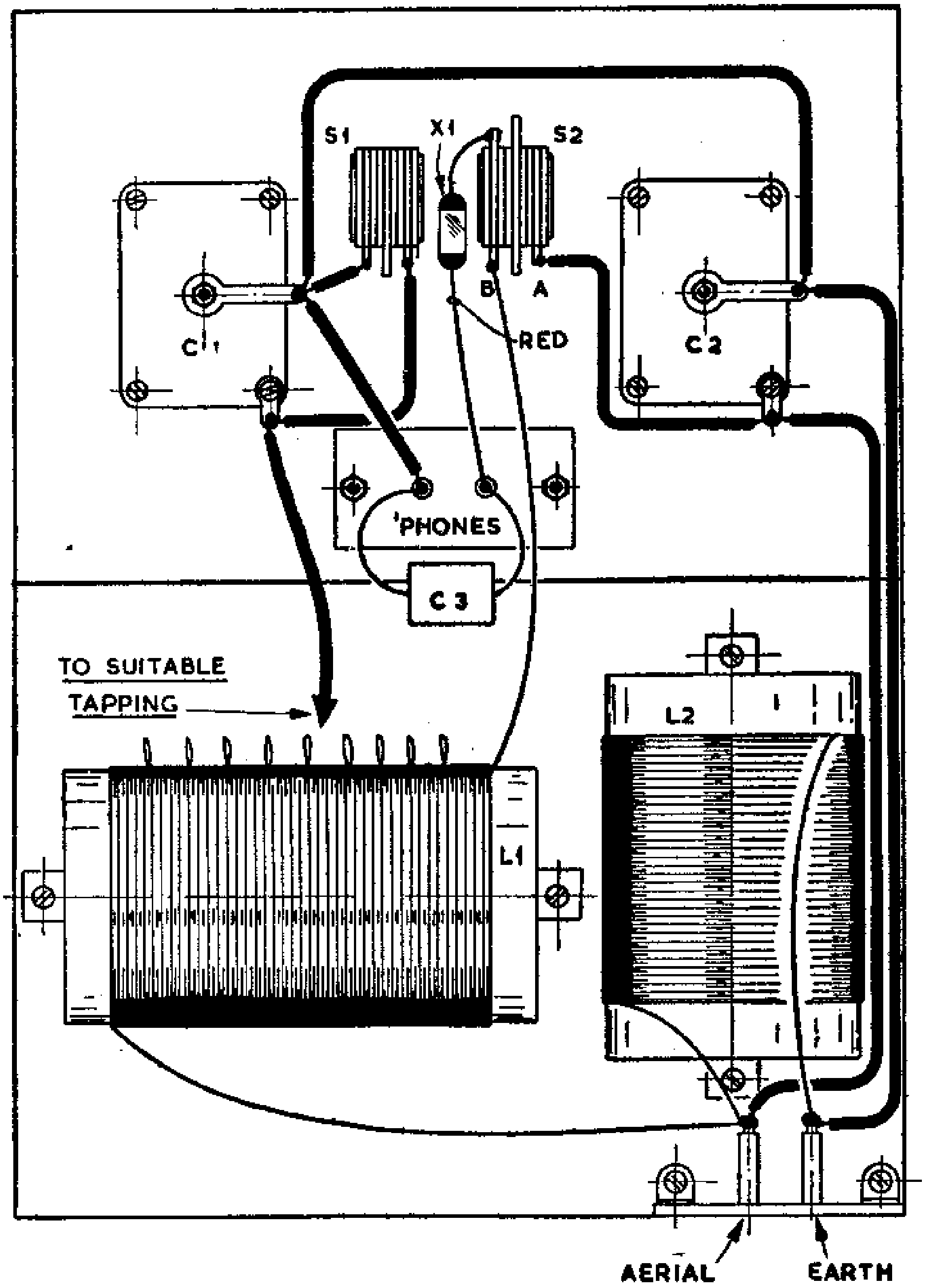
S1
X1
S2
B
A
RED
C 1
C 2
'PHONES
C 3
TO SUITABLE TAPPING
L2
L1
AERIAL
EARTH

FIG. 21 b

Keep S1 open and C1 for tuning, adjusting the tap on L1 for best results. Very powerful signals are best brought in with S1 closed. When using C1 and adjusting the taps on L1, switch S2 is kept in the A position.

It will be found that C2 acts as a fine control on C1. One division of the tuning dial on C1 is roughly equal to a movement of 20 divisions on C2.

Note that the coils are mounted at right angles to one another and that in this instance wood has been used as far as possible, an aluminium or other metal chassis is not suitable.

7

A very useful receiver for both medium and long-wave reception is possible with the circuit of Fig. 22a-b. A home-made coil is used. On the original a 1¼″ diameter former was used but in fact this is not very critical and a 1½″ former could be used. An interesting point is the construction of the long-wave coil section. Solenoid coils as used for the medium wave-band are not efficient if the length is too great relative to the diameter. Now as the long-wave section has 300 turns, if wound as a solenoid the winding length would be very great and a lot of efficiency would be lost. To overcome the difficulty the coil is pile wound. If the whole coil were to be wound in one pile it would still be inefficient because of the capacity formed in the winding, but by dividing it into five sections a coil of reasonable efficiency results. The same remarks concerning winding apply to this coil as to all the others in

Components List, Fig. 22a

C1	300pF Mica Capacitor.
C2	500pF Variable Capacitor.
C3	1000 Mica Capacitor.
X1	Germanium Crystal.
L1-2	See Text.
S1	Single Pole Toggle Switch.

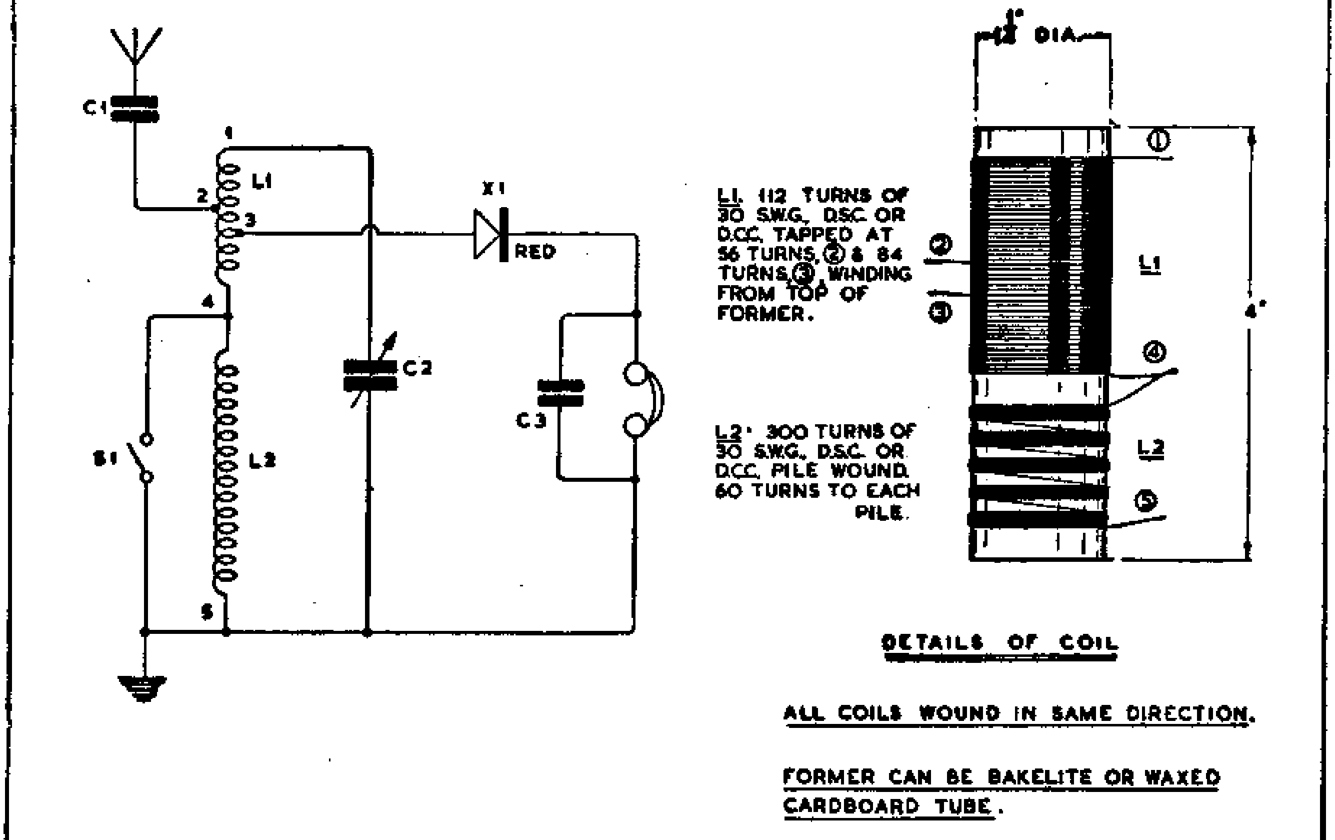
C1
1
2
L1
3
X1
RED
4
C2
C3
S1
L2
5
1½" DIA.
①
②
③
④
⑤
L1
L2
4"
L1. 112 TURNS OF 30 S.W.G. D.S.C. OR D.C.C. TAPPED AT 56 TURNS ② & 84 TURNS ③, WINDING FROM TOP OF FORMER.
L2· 300 TURNS OF 30 S.W.G. D.S.C. OR D.C.C. PILE WOUND. 60 TURNS TO EACH PILE.
DETAILS OF COIL
ALL COILS WOUND IN SAME DIRECTION.
FORMER CAN BE BAKELITE OR WAXED CARDBOARD TUBE.

FIG. 22a

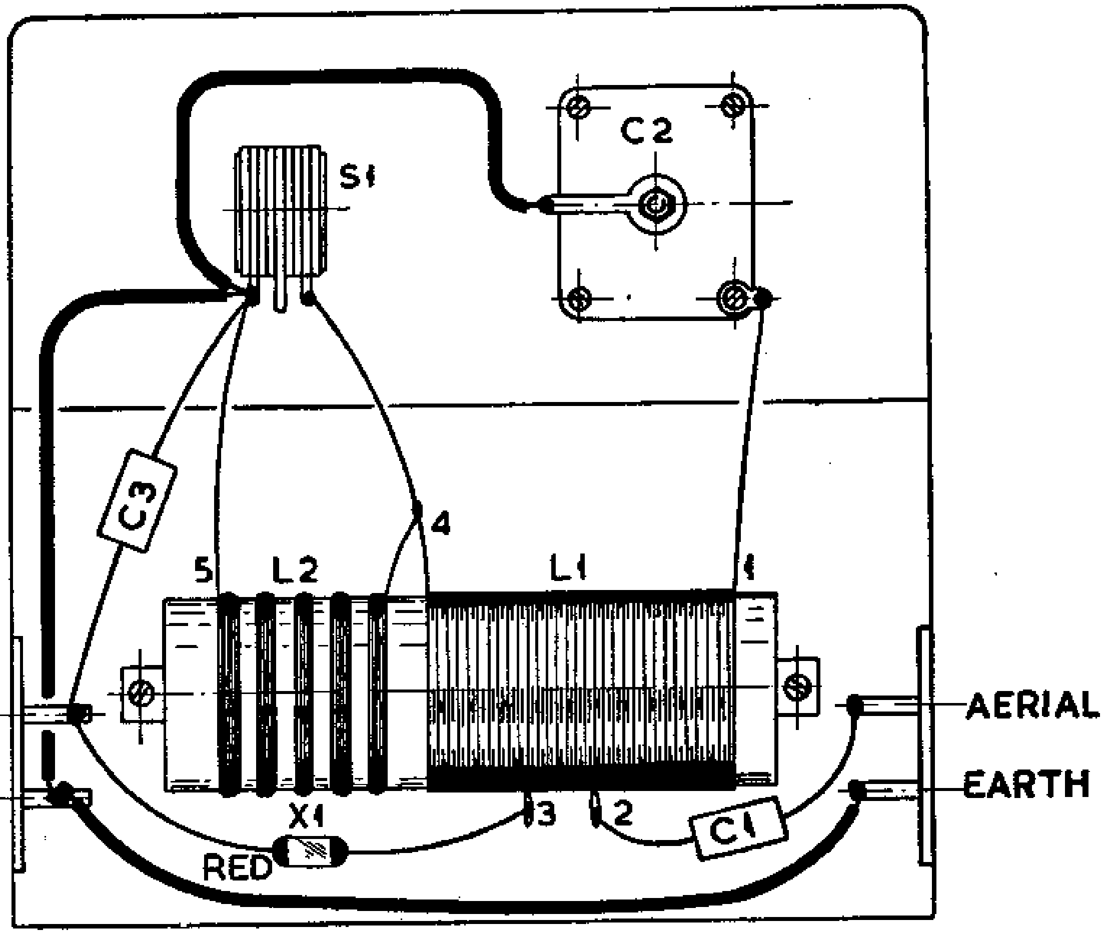

FIG. 22 b

this book, and readers should not, by this time, encounter any difficulties.

The design provides good selectivity with good sensitivity, and provided the set is not required to operate within a few miles of a powerful transmitter, no trouble will be encountered from interference between adjacent stations.

A wooden base-board and front panel is best for this design, and the wiring and layout can be clearly followed from Fig. 22b. Switch S1 is open for long-wave reception and closed for the medium-wave band.

Coil mounting can be on the same principle as for the full-wave design Fig. 19a-b.

8

If you look at all the other circuits shown in this book you will find that in each case the tuning capacitor and tuning coil are in parallel, this is by no means essential and the circuit of Fig. 23a shows a series arrangement that is the tuning capacitor is between the coil and earth instead of being wired up across it. This design also appeared in **Radio Craft** a few years ago.

An interesting point is that the coil is a conversion adapted from another type of component. To make this coil you must obtain an old I.F transformer as used in superheterodyne receivers.

You will require one designed to work anywhere between 450 and 470 kc/s. Some of the

very early ones worked at 110 kcs/ but these are unsuitable.

The I.F. transformer must also be of the type tuned by an iron dust core at each end.

Dismantle the transformer, remove the parallel capacitors and mounting wires so that you are left with two coil bobbins and their former. Be very careful when un-soldering the wires from the supports because the coils are usually wound in litzendraht and must remain intact. Remove 25% (approximately) of the turns from one coil which we will call the secondary, and completely remove the iron dust slug from this coil, now cut the former in half, and mount as in Fig. 23a. Warm the coils with heat from your soldering iron or from a fire and gently slide the coils into position at the ends of the former as in Fig. 23a. The wax will set again holding the coils in their new position.

Note that the wires from the coils are connected to tags on the mountings. These are pieces of bakelite, and usually the original end pieces used on the I.F. transformer can be used.

A word about soldering litzendraht. Do not attempt to use emery paper. Prepare the end of wire by removing the outer silk covering and dip the end in methylated spirits. Ignite with a match and after about five seconds wipe with a cloth held between finger and thumb. This will quench the flame and will remove all traces of enamel from the wire. It is a good plan to experiment with the odd length of wire removed from the secondary coil. If the coils prove to be constructed of ordinary copper wire, it can of course be prepared with fine emery cloth in the normal manner.

Components List, Fig. 23a

C1	500pF variable capacitor.
X1	Germanium Crystal.
L1	See Text.
4	Plugs and sockets or terminals.

Wiring is straight forward as can be seen from Fig. 23b and requires no explanation.

Operation is as follows, selectivity is controlled by varying the distance between coils with the wing nut, and to an extent by the position of the primary iron dust core. Remember that the aerial and earth installations are as important with this set as with any other crystal receiver. No parallel capacitance appears necessary across the 'phones, the set works quite happily relying on the self-capacitance of the head-phones.

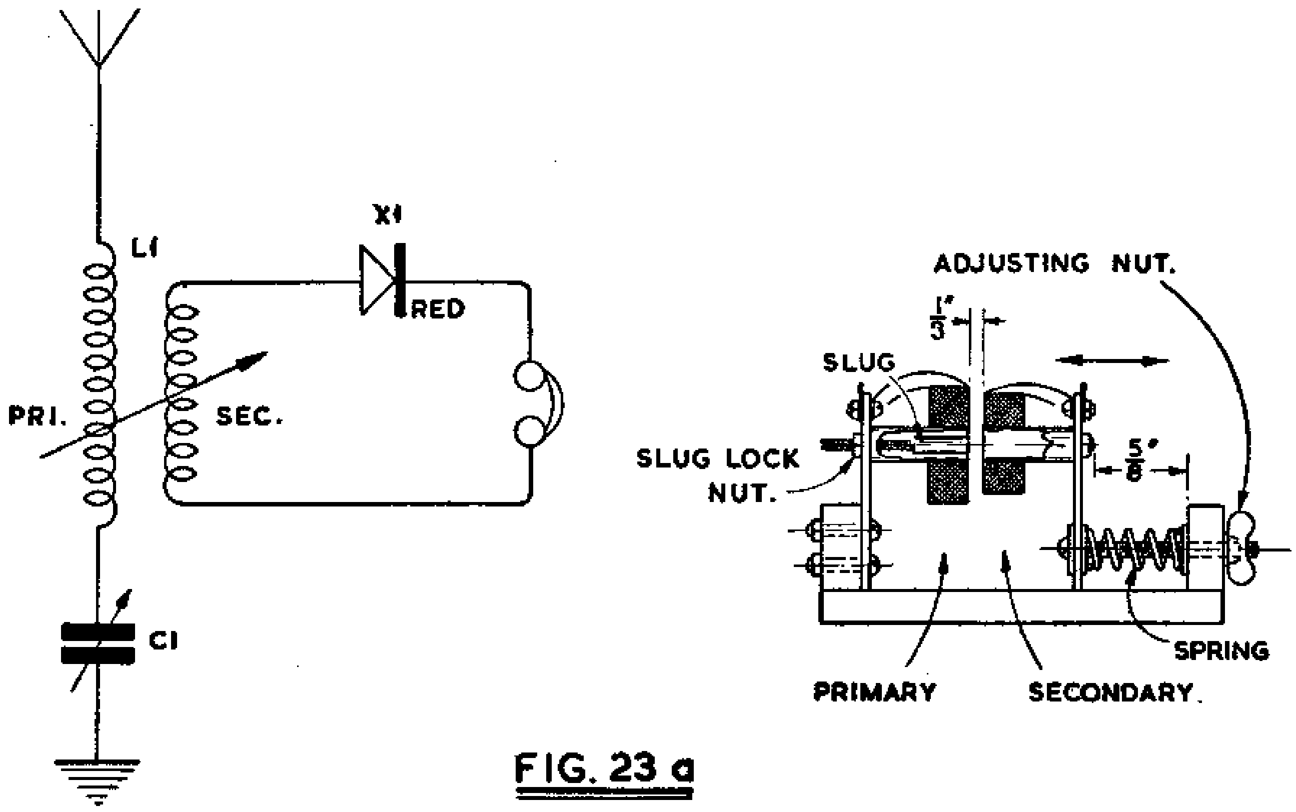

FIG. 23 a

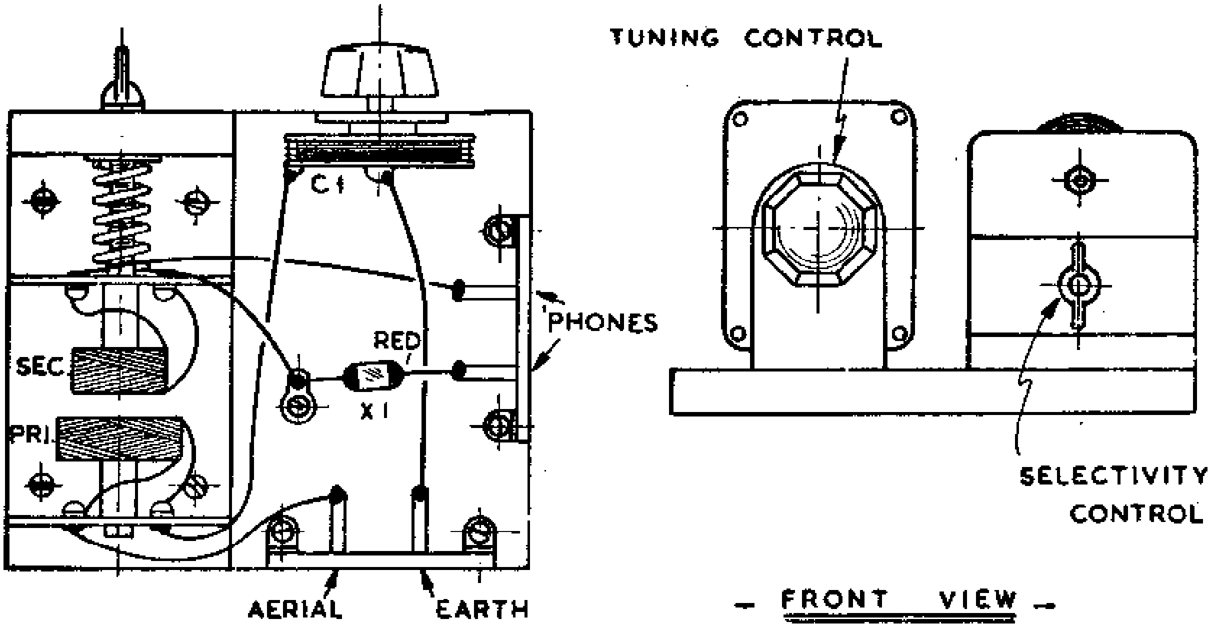
TUNING CONTROL
C1
RED
PHONES
SEC.
PRI.
X1
SELECTIVITY CONTROL
AERIAL
EARTH
- FRONT VIEW -

FIG. 23 b

9

So as to give a truly varied selection of circuits, some of the sets have been designed around commercial coils whilst others use coils which must be home-constructed.

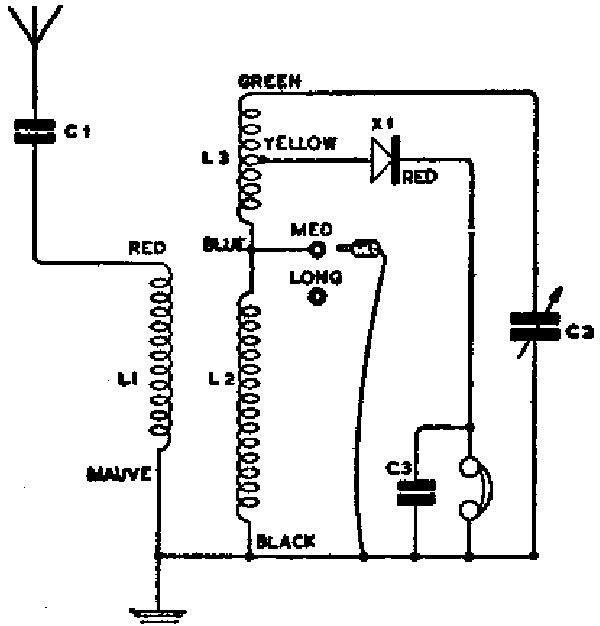
C1
GREEN
X1
L3
YELLOW
RED
MED
RED
BLUE
LONG
C2
L1
L2
MAUVE
C3
BLACK

FIG. 24 a

Most enthusiasts like to build their own coils but when it is a question of something really small combined with efficiency a commercial coil is the obvious choice.

Fig. 24a-b uses a tobacco tin or a similar container just as the design shown in Fig. 16a.

This receiver however is a little more selective and is more suitable than the earlier design, if you are near a transmitter or have a very long aerial. Take great care to connect the coil according to the colours shown on the diagrams. Sockets are used for wave-changing, this cheapens the cost a little and they are just as efficient.

Components List, Fig. 24a

C1	100pF Mica capacitor.
C2	500pF variable capacitor solid dielectric.
C3	1000pF Mica capacitor.
X1	Germanium crystal.
L1-2	R.E.P. dual range coil. Blue box.
6	Insulated wander-plug sockets.
5	Wander Plugs.
2oz.	Tobacco tin or similar container.

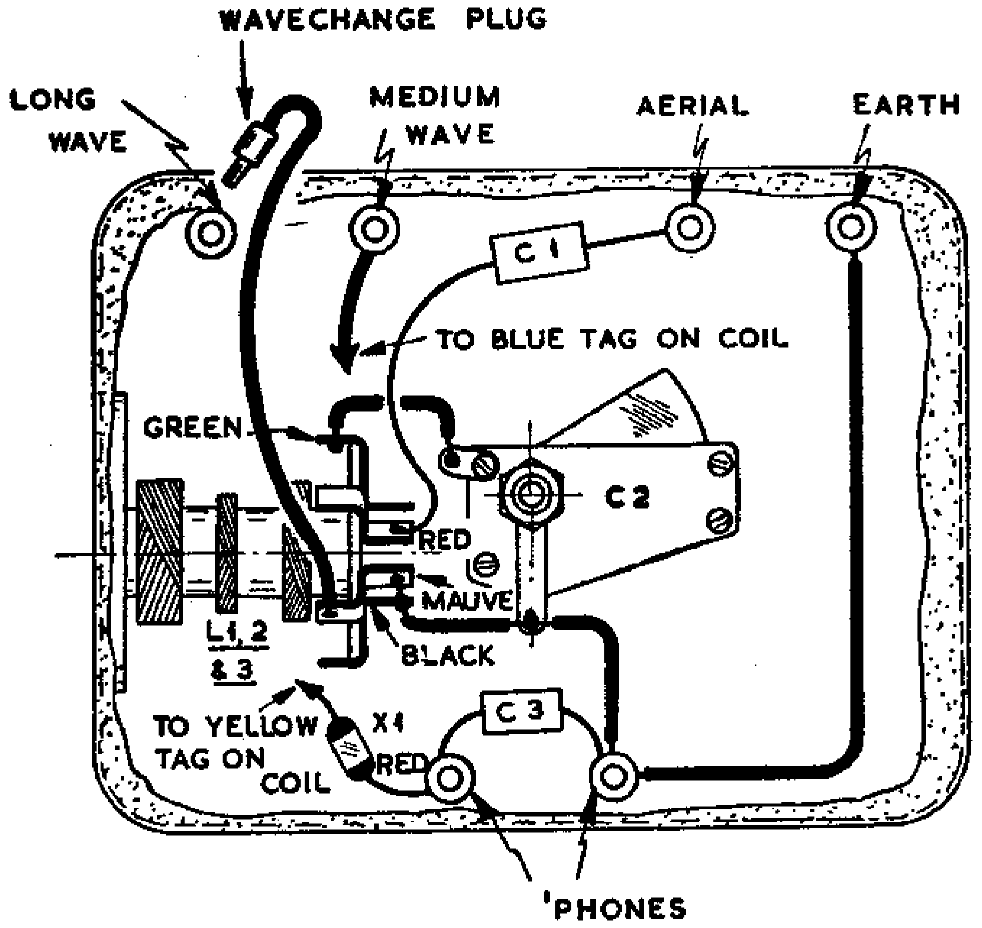

LID SHOWN BROKEN AWAY TO REVEAL POSITION OF COMPONENTS.

FIG. 24 b

10

In certain cases, signals from the local transmitter are too powerful to permit a crystal set with a normal coil to be of much use. I myself live in Hertfordshire a few miles from the London transmitters, and find a more elaborate circuit is necessary. Now in radio circles it is a well-known fact that the greater the number of tuned circuits, the greater the overall selectivity. Look at Fig. 25a, you will see that the usual L and C tuning arrangements have been duplicated. This type of circuit is known as a Band-Pass Filter. There are many kinds of such filters, this particular type is an inductively-coupled filter, note that energy from the first half (L2) is conveyed to L3 by way of two small inductors L1 and L4.

Tuning is quite sharp and the "feel" of the set is quite different from that of single coil designs. A small chassis is ideal for construction.

6"x4"x2½" was used on the original, though it could have been smaller.

Fig. 25b gives the layout and wiring.

Take particular notice that L1-2-3 is mounted horizontally and at right angles to L4-5-6, this is to prevent the coupling from being too "tight" which would ruin the idea of the filter. A good aerial and earth is of course necessary. It may be found on test that the selectivity is too great

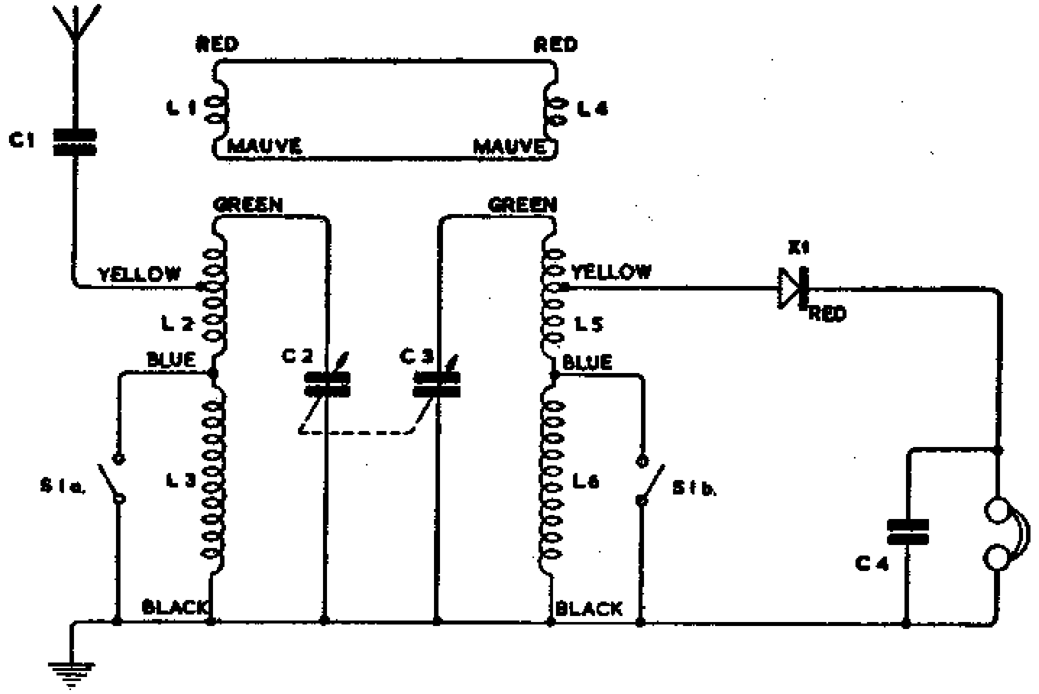

FIG. 25 a

and that volume has suffered. This can be ingeniously overcome by connecting a very small capacitance between the two Green tags on the coils. The simplest way is to connect it across the fixed vane connections on C2-C3. Values for this extra capacitance are a matter for experiment usually a value between 10pF and 47pF is sufficient. A small 50pF trimmer could be used and adjusted to suit. The ideal setting is to enable powerful adjacent stations to be received just short of overlap so that as much volume as possible is obtained.

There is one point to be observed with the ganged capacitor. It should be of the type fitted with trimmers though of course, there is no reason why you should not fit them yourself. These are to allow for differences between coils to be balanced out, and consist of a small variable capacitance across each main section.

When setting up the receiver, first set each trimmer at half way and tune in a station near the high-frequency end of the band (vanes nearly out) adjust the trimmers for maximum volume. If extra coupling capacity is added as described, try adjusting them further.

Once properly set up, no further adjustment is necessary.

Components List, Fig. 25a

C1	100pF Mica capacitor.
C2-3	2 x 500pF variable capacitor (see text).
C4	1000pF Mica capacitor.
L1-2-3	Dual Range Coil R.E.P. (Blue Box).
L4-5-6	Dual Range Coil R.E.P. (Blue Box).
X1	Germanium Crystal.
S1a-b	2 pole single throw toggle switch.
4	Terminals or plugs and sockets.

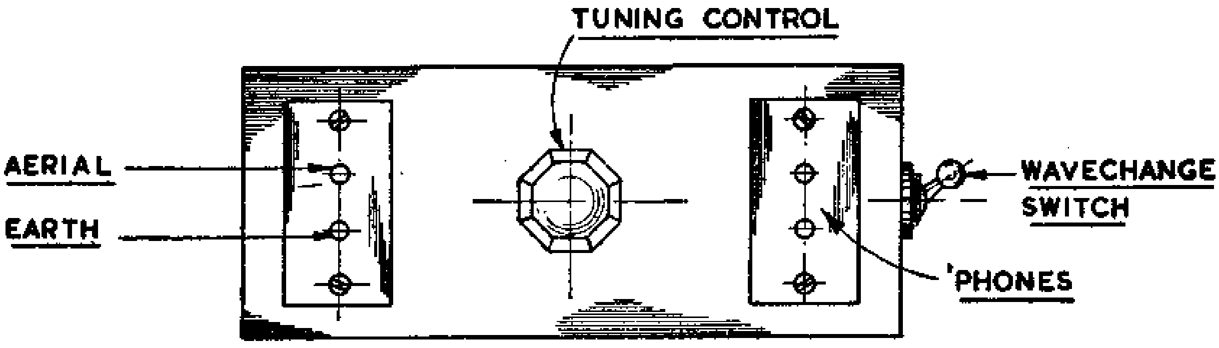
TUNING CONTROL
AERIAL
EARTH
WAVECHANGE SWITCH
'PHONES

FRONT OF COMPLETED RECEIVER.

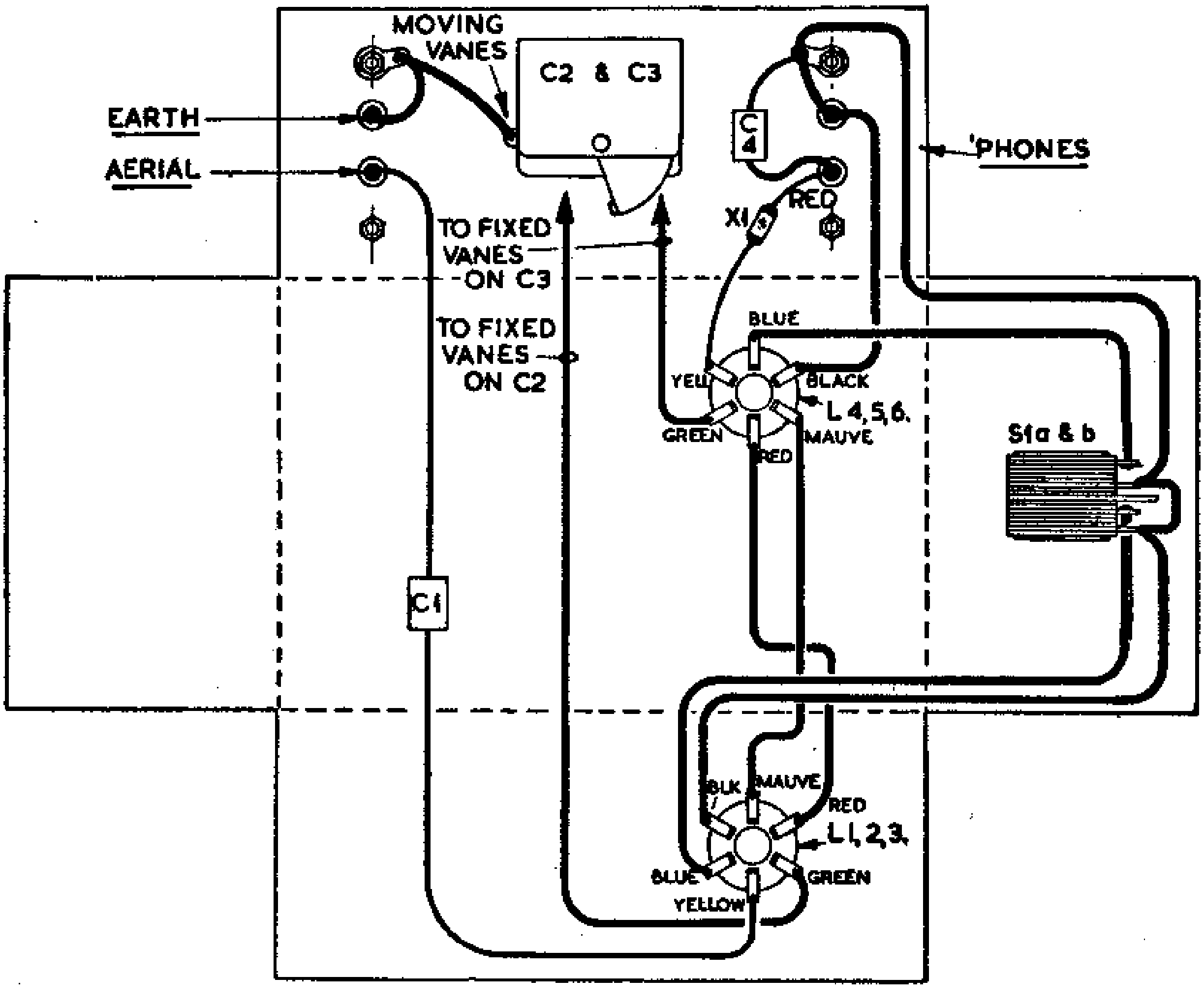
MOVING VANES
C2 & C3
EARTH
AERIAL
TO FIXED VANES ON C3
TO FIXED VANES ON C2
C4
X1
RED
'PHONES
BLUE
YELL
BLACK
L 4,5,6.
GREEN
MAUVE
RED
S1a & b
C1
BLK
MAUVE
RED
L1,2,3.
BLUE
GREEN
YELLOW

EXPLODED UNDER-CHASSIS VIEW

FIG. 25 b

11

Another band-pass circuit is shown in Fig. 26a. This time home-made coils are used. This filter is not inductively coupled but entirely capacitive, relying on C3. As in the previous circuit it must be adjusted to give just sufficient " broadness " to the tuning to provide adequate volume without station overlap. To avoid unwanted coupling a screen is mounted between the coils, this is necessary because due to the size of the coils they would be bound to couple quite tightly without it. Such precautions were not necessary with the circuit of Fig. 25a, as the commercial coils are much smaller physically and the risk of unwanted coupling reduced.

The coils for this receiver are for medium wave reception only, this keeps the size down. Coil winding procedure has already been fully explained, and covers all the requirements of these coils. Base-board and panel construction is used, and for more accurate tuning since the coils are home-made, **separate tuning capacitors are** recommended. Fig. 26b gives the layout and practical wiring.

Components List, Fig. 26a

C1	100pF Mica capacitor.
C2	500pF variable capacitor.
C3	50pF Mica trimmer.
C4	500pF variable capacitor.
C5	1000pF mica capacitor.
L1-2	See text.
X1	Germanium Crystal.
4	Terminals or plugs and sockets.

For those who do not wish to bother with these a ganged capacitor can be used, but tuning may not be as good as with separate ones. The screen mounted between the coils is of aluminium or copper and should be about 1" higher than the coils and the same length as the baseboard depth.

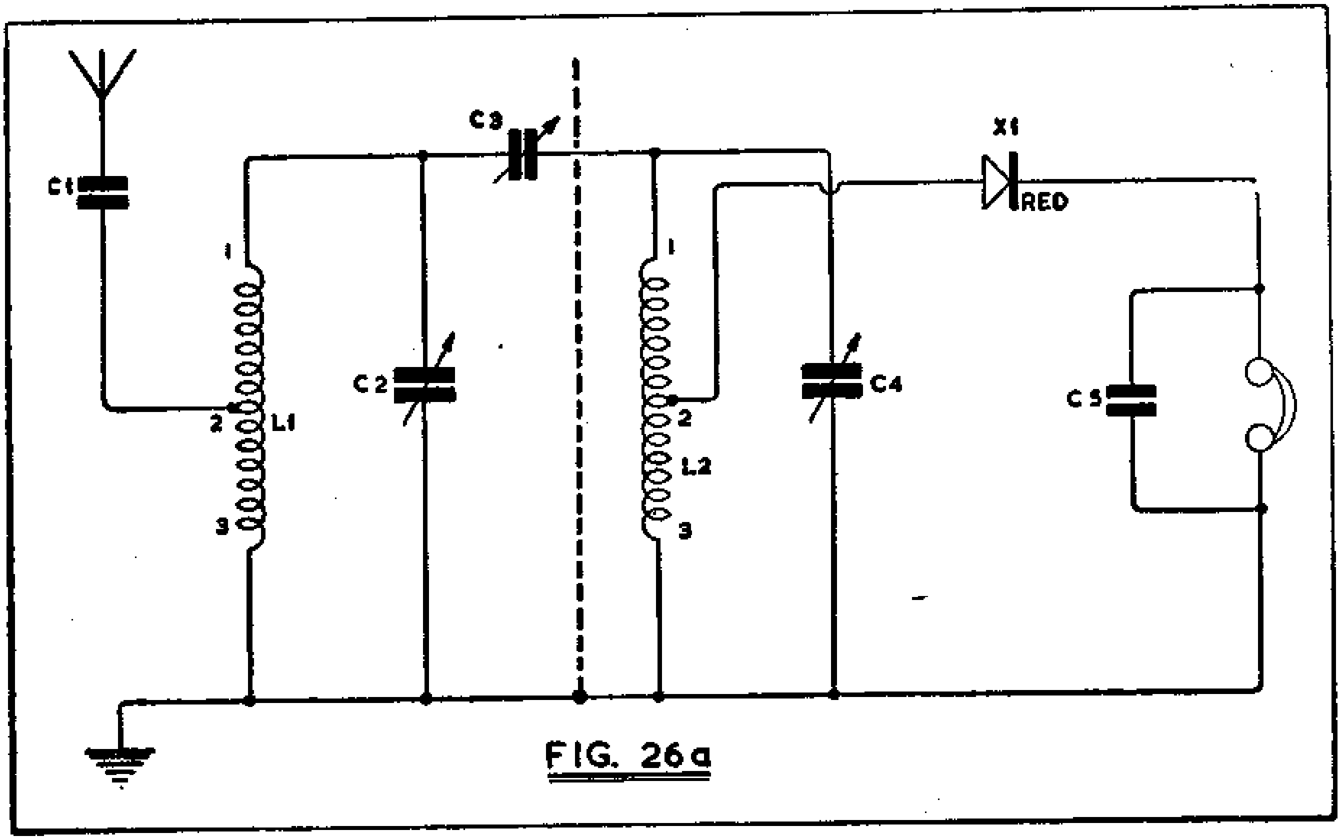

FIG. 26a

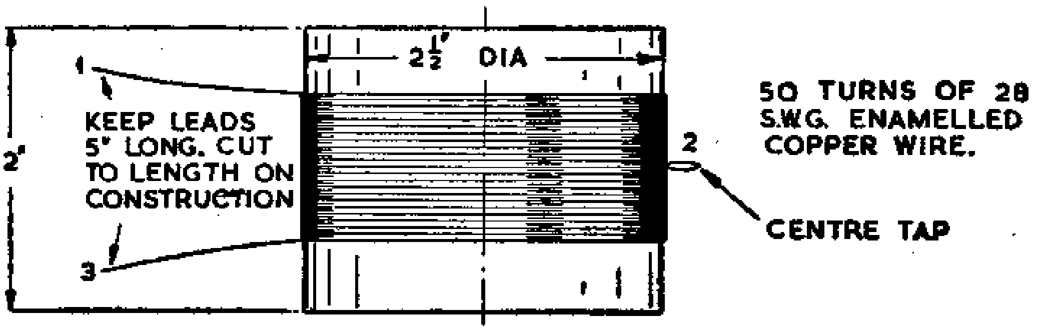
2½" DIA
1
KEEP LEADS 5" LONG. CUT TO LENGTH ON CONSTRUCTION
2"
3
2
50 TURNS OF 28 S.W.G. ENAMELLED COPPER WIRE.
CENTRE TAP

COIL DETAILS.

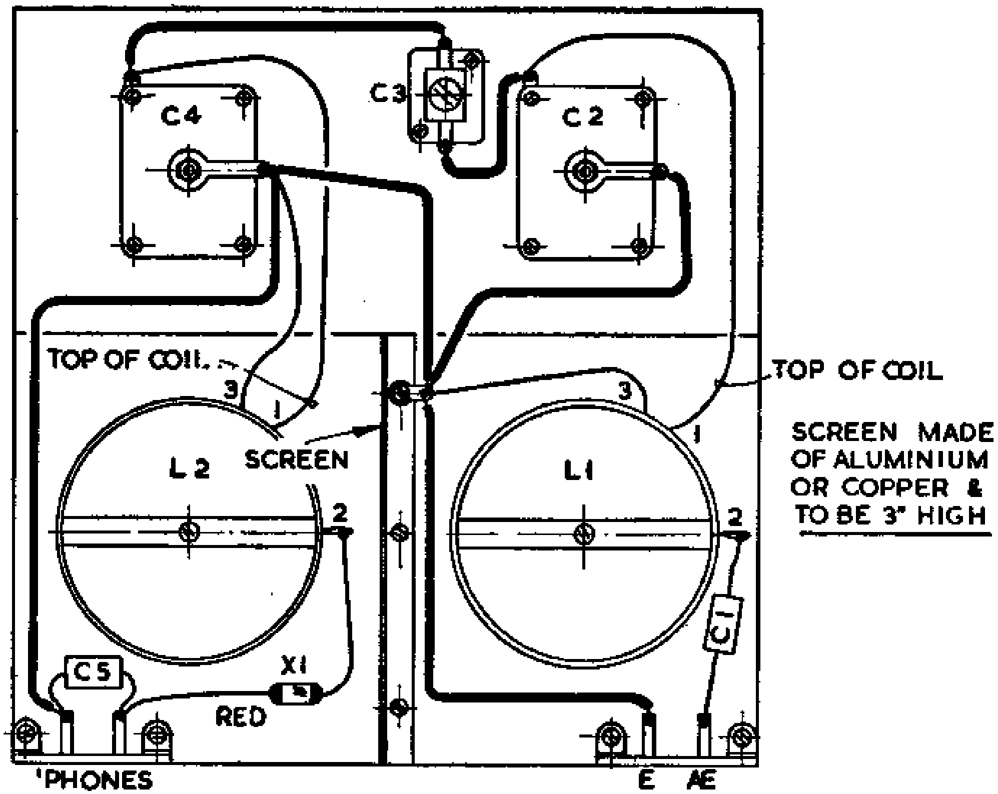
C4
C3
C2
TOP OF COIL
3
1
L2
SCREEN
2
TOP OF COIL
3
1
L1
2
SCREEN MADE OF ALUMINIUM OR COPPER & TO BE 3" HIGH
C1
C5
X1
RED
'PHONES
E
AE

FIG. 26 b

12

Fig. 27a-b uses a band-pass filter for tuning, and the coils are home-constructed.

With this design a ganged capacitor is quite suitable as the coils tune quite accurately. It is essential that the coupling condenser C3 is **non-inductive**, there will be no difficulty in obtaining a new component of this pattern but it is as well to avoid old components which may not employ this form of construction.

Receivers using this type of filter are used to the best advantage when situated a short distance from a powerful transmitter and if selectivity is not important this particular design is not the most suitable.

Both medium-waves and long-waves are catered for: consider L1-2-3, this is the first half of the filter.

L1 is the aerial coupling coil, with S1 closed,

the medium-waves are received. This coil is inductively coupled to L2 and L3. Again with S2 open L2 and L3 combined cover the long-waves and when S2 is closed the medium waves are covered.

Tuning for this half of the filter is by means of the capacitor C1, which is one half of the ganged capacitor. A screen is erected between the two sets of coils comprising the filter, and coupling is effected by C3, an earth return for the coils is provided by R1.

L4 and L5 operated in a similar manner to L2 and L3, tuning is by means of C2, the remaining section of the ganged capacitor. You will realise that the coils must be accurately wound otherwise tuning will not remain constant between the two sections over the band.

The ganged capacitor should be fitted with trimmer capacitors when purchased so that the two circuits can be balanced.

You will have noticed that this circuit uses three sets of switch contacts. A three-pole two-way wafer switch is the best choice otherwise either two or even three separate switches are necessary. All the necessary coil details are given in the diagrams, the tap on L4 is made in the same way as shown for other coils in the manual. One other component used only on this design, is the resistor R1, its size is not important because the current flowing is so small it can be ignored.

A ¼ watt rating is adequate, its precise value is not critical so that a 20% tolerance component is quite suitable.

Resistors are coded by colours and the 1000Ω type you require will bear the following sequence. At one end will be painted three coloured rings.

Components List, Fig. 27a

C1-2	500pF 2 gang, variable capacitor.
C3	0.05mfd. non-inductive paper capacitor.
C4	1000pF mica capacitor.
R1	1000 ohm resistor.
X1	Germanium Crystal.
S1-2-3	3-pole 2-way wafer switch.
L1-2-3 & L4-5	See text.
4	Terminals (Aerial, Earth and 'Phones).
	Sheet copper or aluminium for screen.

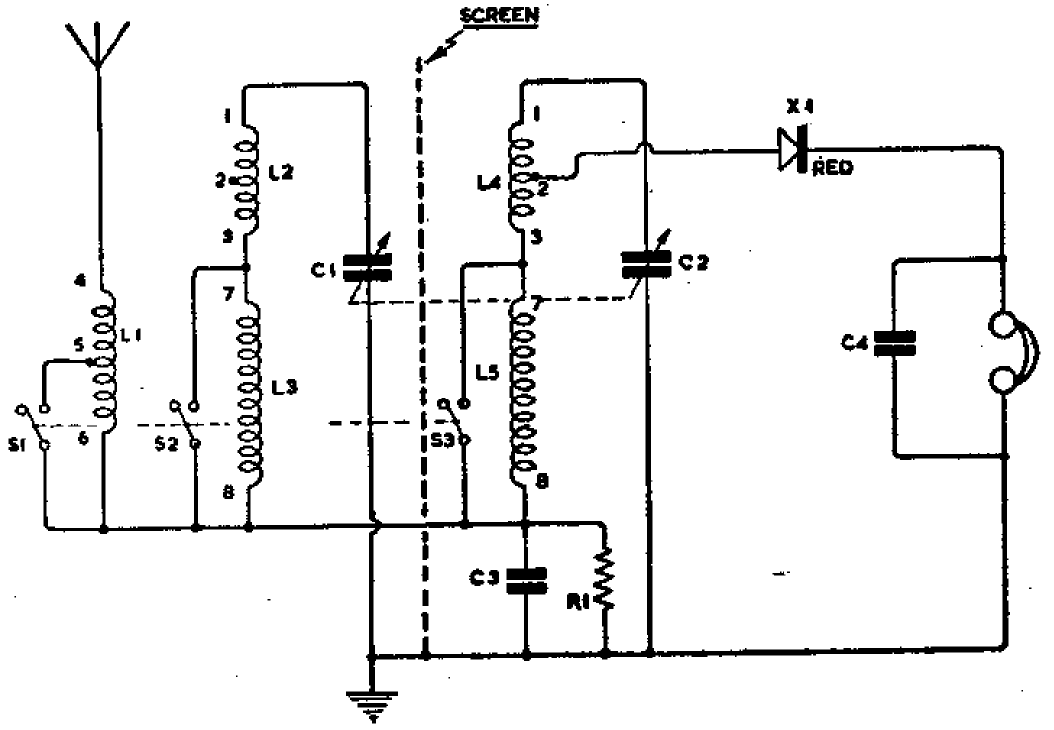

FIG. 27a

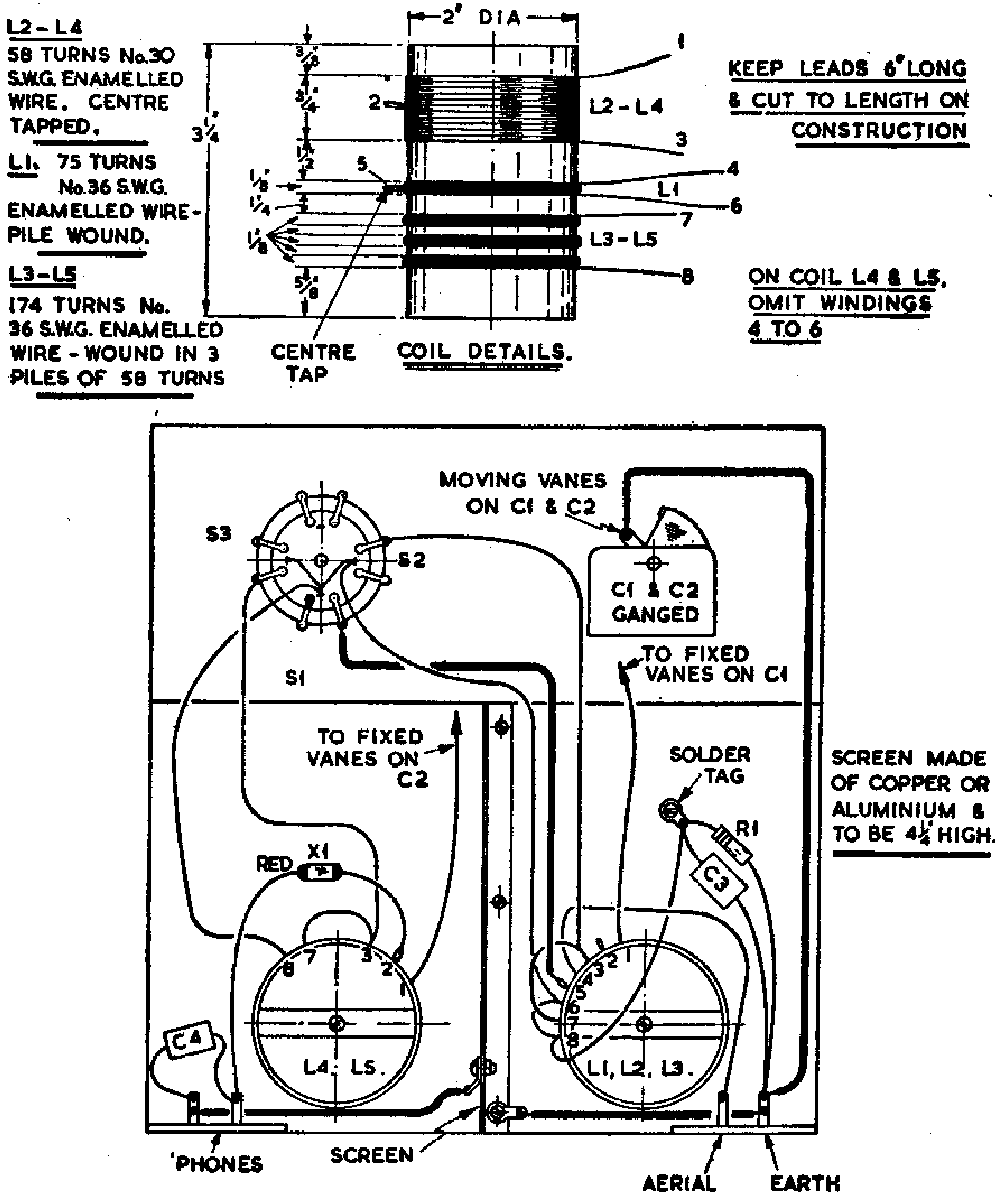

L2 - L4
58 TURNS No.30 S.W.G. ENAMELLED WIRE. CENTRE TAPPED.

L1. 75 TURNS No.36 S.W.G. ENAMELLED WIRE - PILE WOUND.

L3 - L5
174 TURNS No. 36 S.W.G. ENAMELLED WIRE - WOUND IN 3 PILES OF 58 TURNS

KEEP LEADS 6" LONG & CUT TO LENGTH ON CONSTRUCTION

ON COIL L4 & L5, OMIT WINDINGS 4 TO 6

SCREEN MADE OF COPPER OR ALUMINIUM & TO BE 4¼" HIGH.

2" DIA
3⅛
¾
½
⅛
¼
⅛
⅝
3¼"
1
2
3
4
5
6
7
8
L2 - L4
L1
L3 - L5
CENTRE TAP
COIL DETAILS.
MOVING VANES ON C1 & C2
S3
S2
S1
C1 & C2 GANGED
TO FIXED VANES ON C1
TO FIXED VANES ON C2
SOLDER TAG
R1
C3
RED
X1
C4
L4, L5.
L1, L2, L3.
'PHONES
SCREEN
AERIAL
EARTH

FIG. 27 b

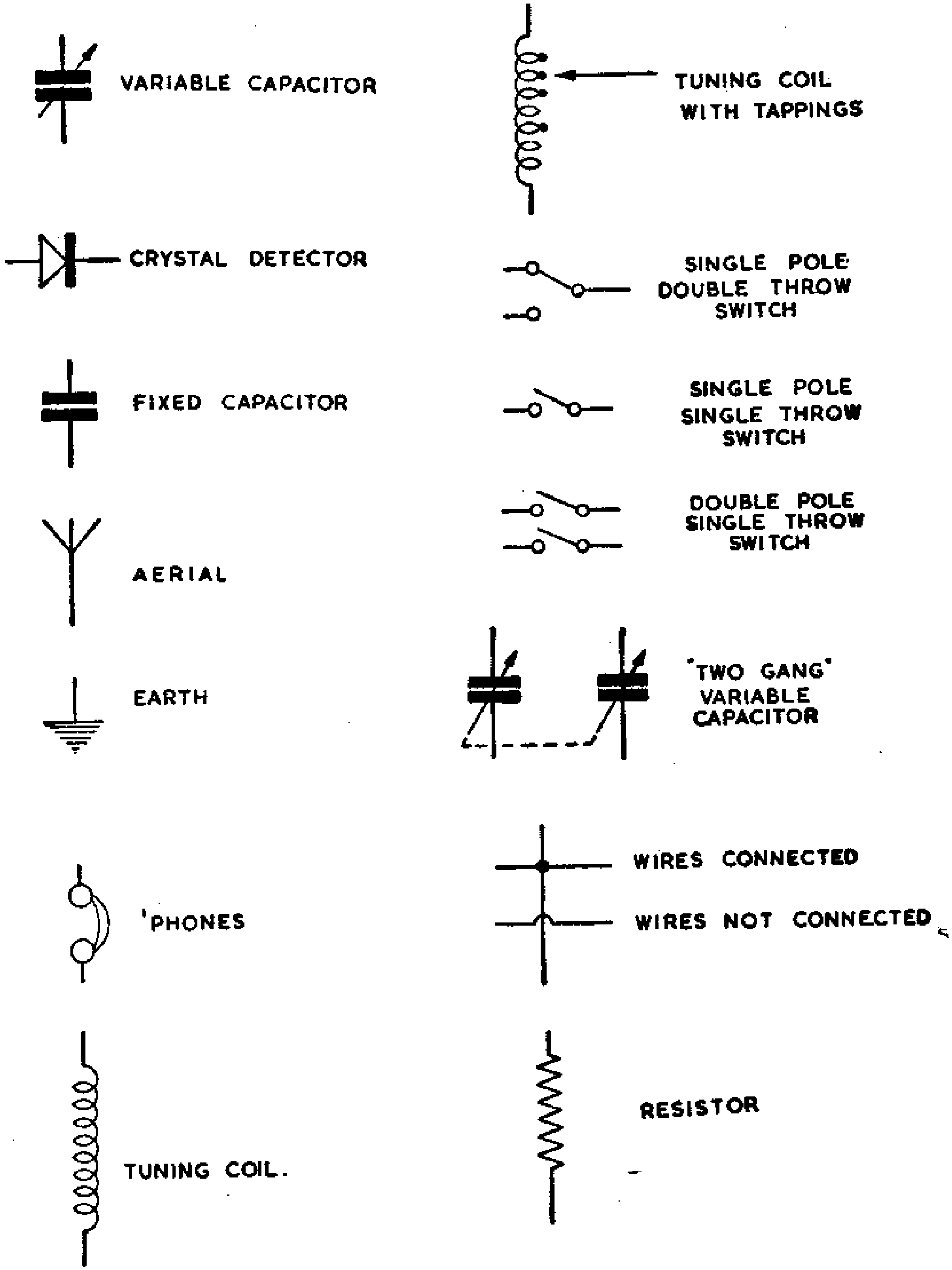
VARIABLE CAPACITOR
TUNING COIL WITH TAPPINGS
CRYSTAL DETECTOR
SINGLE POLE DOUBLE THROW SWITCH
FIXED CAPACITOR
SINGLE POLE SINGLE THROW SWITCH
DOUBLE POLE SINGLE THROW SWITCH
AERIAL
EARTH
"TWO GANG" VARIABLE CAPACITOR
WIRES CONNECTED
WIRES NOT CONNECTED
'PHONES
RESISTOR
TUNING COIL.

SYMBOLS USED IN THE CIRCUITS DESCRIBED.

FIG. 28

The first one is BROWN which represents 1, the first figure of the value, the second BLACK, this indicates the second numeral is nought, and the third colour is RED which tells us that two further ciphers or noughts are added to give the total value. From this you can see, we have 1 plus a nought equalling 10, plus two further noughts which total 1000. Some earlier types of resistor were painted differently, the whole body colour represented the first figure, one tip was coloured to represent the second figure, and a painted dot in the centre gave the number of ciphers to be added.

There are no other special points regarding construction except perhaps that if a long aerial is used, a small 100pF capacitor might be tried in the aerial lead to prevent damping the first uned circuit.

When setting up the receiver, tune to a station the high-frequency end of the M.W. band and the trimmers for loudest volume, no other stments are necessary.

CONCLUSION

The twelve receivers described are representative of the best crystal set designs available to-day. Their construction will provide many hours of useful enjoyment, and the results will give lasting pleasure.

A final word about components, if no actual maker is specified, any good class component can be used. Switches, fixed and variable capacitors and crystals, are available from numerous manufacturers all of which are invariably of excellent quality. Where a particular manufacturer's product is called for, the specification should be adhered to. The commercial coils specified are freely available from most supply houses specialising in components for constructors, but in case of difficulty write to Bernards (Publishers) Ltd., The Grampians, Western Gate, London, W.6, who will, on receipt of a S.A.E. be pleased to supply the address of your nearest stockist. Coil formers are not always easy to obtain and in this case advice should be sought from Post Radio Supplies, 395, Queensbridge Road, London, E.8, who will be pleased to help in any way possible.

* *

BERNARDS RADIO BOOKS

No.		Price
56.	Radio Aerial Handbook	2/6
57.	Ultra-Shortwave Handbook	2/6
58.	Radio Hints Manual	2/6
64.	Sound Equipment Manual	2/6
68.	Frequency Modulation Receivers' Manual	2/6
83.	Radio Instruments and their Construction	2/6
96.	Crystal Set Construction	1/-
99.	One Valve Receivers	1/6
100.	A Comprehensive Radio Valve Guide, Book 1	5/-
103.	"Radiofolder" A. The Master Colour Code Index for Radio and T.V.	1/6
104.	Three Valve Receivers	1/6
121.	A Comprehensive Radio Valve Guide, Book 2	5/-
123.	"Radiofolder" F. The Beginners' Push-Pull Amplifier	1/6
126.	Boys' Book of Crystal Sets and Simple Circuits	2/6
129.	Universal Gram Motor Speed Indicator	1/-
135.	All Dry Battery Portable Construction	2/6
138.	How to Make F.M. and T.V. Aerials, Bands 1, 2 and 3	2/6
141.	Radio Servicing for Amateurs	3/6
143.	A Comprehensive Radio Valve Guide, Book 3	5/-
145.	Handbook of AM/FM Circuits and Components	2/-
146.	High Fidelity Loudspeaker Enclosures	5/-
147.	Practical Tape Recording Handbook	5/-
148.	Practical Transistor Receivers, Book 1	5/-
149.	Practical Stereo Handbook	3/6
150.	Practical Radio Inside Out	3/6
151.	Transistor Superhet Receivers	7/6
156.	Transistor Circuits Manual, No. 1	2/6
157.	A Comprehensive Radio Valve Guide, Book 4	5/-
158.	Radio, Television, Industrial Tubes, Semiconductor and Diodes Equivalents Handbook (208 pages)	9/6
159.	Realistic High Fidelity	5/-
160.	Coil Design and Construction Manual	5/-
161.	Radio, Television and Electronics Data Manual	3/6
162.	High Fidelity Stereo Gramophone	5/-
163.	Transistor Circuits Manual, No. 2	2/6
165.	Hartley on Hi-Fi, Book 1. Radio Tuners	5/-
166.	Public Address Systems	2/6
168.	Transistor Circuits Manual, No. 4	2/6
170.	Transistor Circuits for Radio Controlled Models	7/6
171.	Super Sensitive Transistorised Pocket Set Construction	3/6
172.	International Radio Stations List	2/6
173.	Practical Transistor Audio Amplifiers	3/6
174.	Transistor Subminiature Receivers Handbook	5/-
175.	Transistorised Test Equipment and Servicing Manual	3/6
176.	Transistor Audio Amplifiers Manual	6/-
177.	Modern Transistor Circuits for Beginners	7/6
178.	Comprehensive Radio Valve Guide Book 5	6/-
179.	Transistor Circuits Manual No. 5	5/-
180.	British Semiconductor Survey	15/-
181.	22 Tested Circuits Using Micro Alloy Transistors	5/-
182.	"At A Glance" Radio Valve & T.V. Tube Equivalents	3/6
183.	How to Receive Foreign T.V. Programmes on your Television Set by simple Modifications	5/-
184.	Tested Transistor Circuits Handbook	2/6
185.	Tested Short-Wave Receiver Circuits using Micro Alloy Transistors (MAT's)	5/-
186.	Tested Superhet Circuits for Short-Wave and Communication Receivers using Micro Alloy Transistors (MAT's)	6/-
187.	The TSL Mark "4" Valved F.M. Tuner and its Construction	3/6
188.	Construction of the B.H. High Fidelity "14" watt Audio Amplifier	1/6
189.	How to build the world's smallest transistor radio. The Sinclair Micro 6	1/-
190.	How to build the world's smallest High Fidelity Amplifier	1/6
191.	Practical Car Radio Handbook	6/-
501.	ABC's of Magnetism	5/-
502.	ABC's of Missile Guidance	5/-
503.	ABC's of Ultrasonics	5/-
504.	ABC's of Transistors	5/-
	Resistor Colour Code Disc Calculator	1/6
	Engineers' Reference Tables	1/6
	International Radio Tube Encyclopaedia—3rd Edition	63/-

Jules GUILLOT

Atmospheric Electric Generator

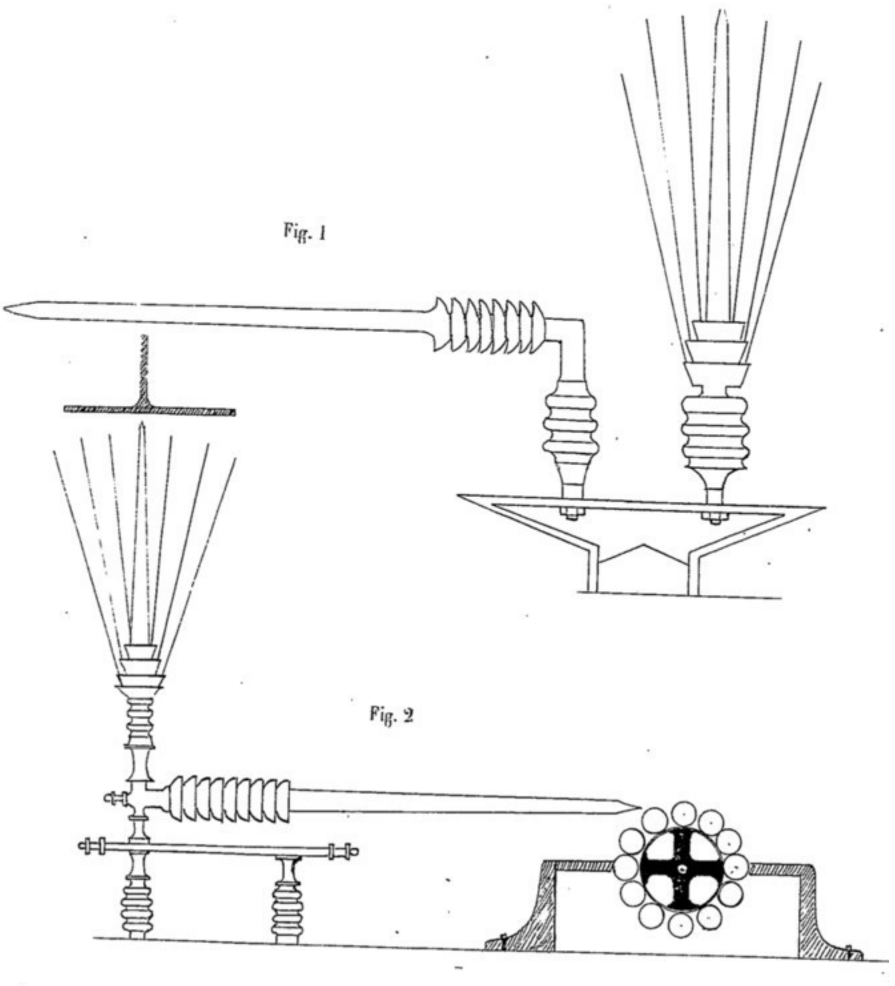

Fig. 1

Fig. 2

Shared by L. Cozma (Thanks !) --

The Guillot device generated about 2.5-3 Kilowatts with antenna height of ~ 20 meters. Power depends on the total collector surface and height of the vertical antenna. The apparatus in the photo produced ~300 watts with a collector 2 meters tall.

"Practical Utilization of Atmospheric Electricity"

by

Gillbert Darida

in *The Invention Encyclopedia,* pp. 204-207 (1930, Geo. Constantinescu, Ed.)

"The Earth has its own negative electricity, in the soil, and the atmosphere which surrounds the Earth is charged with positive electricity. The electric potential (the voltage) increases with the altitude, so we can say that the electric force is proportional to the atmospheric altitude (after Franklin, Quetelet, Lord Kelvin, Mascart, Joubert and other scientists).

The recent observations demonstrate that the air at 6000-7000 meters in altitude, is very highly charged with positive electricity, which could be explained by the friction between the external photosphere and the upper atmosphere of Earth, which rotates at a speed of more than 100,000 Kilometer per hour.

In that way, the Earth works like an electrostatic generator with electric charging by influence --- the upper atmosphere is positively charged by influence and the Earth crust obtains the negative polarity.

Between the two environments, the air and the soil, and inside the low atmosphere, in conditions of good weather, there are sbout 800 positive ions and also, 680 negative ion (and electrons) in just one square-centimeter of normal ionised air.

The Earth behaves like an huge electric armature negatively charged, which repels the electrons and attracts the positive ions. That positive ions' attraction determinates an electric current, also called "convective current". That's like an invisible continous bombardment, subject to daily and seasonal variations, which could be aproximated at **3 x 10 ^ -16 Amperes per square-centimeter,** and that is a total value of 1500 Amperes for the entire surface of the Earth.
The question is --- how does this current always maintains the same direction ? We can suppose that the natural radioactive emissions of the soil is responsable for this. We also know that radioactive emissions of the Earth works usually near the soil surface, and that explains the ionisation phenomenona inside caves.

The ionisation of the low atmosphere could also be the effect of the radioactive emissions of the Earth, especially when the X radiation works. Also, the Earth atmosphere is ionisated by the external radiation proceeding from the Sun and from the space environment, especially the action of ultraviolet radiation and the electron fascicles emitted by the Sun surface, at the temperature of 6000 deg. Celsius.

The capture of atmosphere electricity has been used in France, with aerial cables mounted on the Mont Blanc, and also in Germany --- with conductive cables carried by the captive baloons.

The atmosphere electricity collect system invented by eng. Jules Guillot is most ingenious and it relies on "the **electric siphon**" [*m.n. --- the discharging devices or spark-gap used today from the protection of aerial electric cables against the atmospheric electricity ; Jules Guillot has thought to recuperate and to use that electricity*].

His method consisted in the direct "pumping" of the atmosphere electricity using a collecting device which had two antennae and several collecting rods.

One antenna is **vertical** and it has a lot of **rods** scattered like an opened **fan**, with the tips against the **zenith**, for collecting the **negative** electricity which comes from the air ; the **horizontal** antenna is orientated against the **South** and its role is to collect the **positive** electricity.

The air electricity seems to have the double-polarity, as we can see at two electrised clouds when between them apears the electric discharge. The air could have different electric charging and the ionisation processus of atmosphere is very heterogenous. The inventor Guillot used **two separated and insulated armatures with the positive armature against the South** (more precisely, against the **Equator**...) and the **negative** armature against the **zenith**.

We can see in the scheme :

1,- the device P is the "**collector apparatus**" having **15-20 meters height** ; it also has a lot of **rods** which are scattered **fan**-like, all this on top of an **iron pillar** ; and also an **horizontal antenna** S oriented against the **South** ; the two antennae are disposed at **90 deg. angle** forming the two armatures where the electrical tension must apear. As the **height** of the pillar **increases**, the **voltage** increases more and more ;

2,- a system of **lightning-rods** (p1, p2, p3) for protection against discharges of atmospheric electricity during storms;

3,- the regulator system (**voltage regulator**, etc) R ;

4,- the electric **resistors** (Cn) for absorbing the parasite currents ;

5,- the "electric siphons" which have the role of making a magnetic field as a good environment for the transport of the ionized fluid, absorbing the electricity from the air ; the E1 and E2 are their regulator devices ;

6,- an **auxilliary continous current power supply** which send the current against an special **excitation coil**, used as starting device ; it works only when the system is starting.
This collector device works very good for lighting and heating. The device made by Guillot had
20 meters height on the total surface of **3 square-meters**. Also, J.Guillot used and electrical transformer for the industrial utilization of this "collector" --- as power supply for industrial electric engines."

French Patent # 551,882

Apparatus for Capture of Electric Currents in the Atmosphere

16 April 1923

Considering the terrestrial globe like the inductor of a dynamo, where the extraterrestrial ether is the inductor of electric currents circulating in the atmosphere; the apparatus of the invention described here selects through the device described below two perfectly distinct currents and eliminates others.

The device includes:

1, The air sensor; 2, a series of lightning rods; 3, vacuum; 4, an array of resistances; 5, vacuum; 6, witnesses.

1. An aerial sensor mounted on a pole is composed of **magnetic steel** fixed and sealed by a porcelain **insulator**, and surrounded a base by a **bronze ring** which are screwed **32 points of soft iron**, all forming antenna a pole. level beside it and isolated from the first, forming the 2nd pole, is fixed a sharp point of magnetic steel fixed in the horizontal position and directed towards the south and **movable to an angle of 45** is sealed in a porcelain insulator. This point is also circled a **ring of copper** notches figure. 1.

2. A series of 6 lightning surge arrestors at corners between each pole and the earth and the various gauges.

3. A first **regulator** form of a device similar absolutely similar to the aerial amtenna, but the two poles are superposed and opposed to the vertical point, and a **copper disk connects to ground**. Opposite the horizontal point, a **ring of tin** to which are welded tubes alternatively **16 tubes** composed each composed of **copper-lead** and **iron-lead** . The ring connects to the **ground** as shown in figure. 2.

A second regulator form of **automatic breakers** balanced on 3 poles and two poles also equally balanced.

4. A panel of **resistances** composed of **wire mesh** in bunches from different sections of **glass tubes** containing **copper** dust, **coal** and flowers of **sulfur**.

Vacuum cleaners consists of a **wooden box** on each pole, containing a **porcelain vase** in which **layers** isolated with **mica** is made up of **mercury, tin, coal, copper** and **sulfur**, all contaied in a **copper tube**.

6. The witnesses are comprised of ordinary incandescent lamps.

Resume

By the point at the zenith and the point to the south we channel two currents forming the two poles. We also protect from lightning. It regularizes the flow by a regulator and similar devices by a controlling each devise of adequate strength of the current harmful nature of these devices do not have loads. The refined current is conducted by ordinary copper wires**.**

French Patent # 565,395

Combined Apparatus for Capture of Atmospheric Electric Currents with Immediate Implementation

25 January 1924

We know that earth with its constitution, its rotation and movement in space, provides the electricity in the atmosphere. The electrical currents escape into space or largely accumulate towards the equator, as a result of the greater periphery of the globe.

In the atmosphere, there are two clearly distinct poles, i.e., clouds that can be electrically positive or negative; everything in nature shows this, otherwise it would be difficult to explain the lightning that occurs between the clouds and which are none other than contrary cloud electrical charges discharged by too close proximity.

The invention relates to a set of devices capable of capturing the atmospheric electricity.

So far, all searches made for this purpose has been to capture the atmosphere, i.e., that brought together into one system to capture two poles and opposes effect has been to destroy or cancel each other leaving the amount as the difference of the strongest to weakest.

It is therefore easy to understand why, considering the two poles of air as about equal, it is almost impossible to measure some potential with the land, the highest on the lowest remaining which can be positive or negative, copper being the strongest of one or other of these polarities.

It is on this basis that all research until now have been made, and that is why we must consider that all the time, between the two poles captured in the atmosphere and not in any mixture of systems uptake, it is possible to obtain tension and intensity, a considerable and measurable power captured separately between the two poles and remaining isolated from one another until their utilization .

In accordance with this invention, this collection is made by attraction of primary electrical currents in teh atmosphere, either positively or negatively charged, by an antenna that has two points absolutely isolated from one another, where one points to the zenith to attract negative electricity, and the other turned south, toward the equator to draws in the masses grouped in this area.

This shows and proves the existence of atmospheric electrical currents, and that it is possible to capture them with specially designed equipment primarily acting as regulators, as the persistent difficulty encountered so far in realizing their capture lies with large variations of tension in which the current atmosphere is present in space, for each of the poles, and an overload could inevitably fatal.

The patent in its present form of presentation does not cover the regulators or devices that are anticipated in the system, because they can be designed in different ways and give the same result, but covers the application and grouping of equipment operating as automatic relays with variable influence to limit current tension to that chosen for utilization.

These devices have a role associated with an overflow reservoir outside carrying excess liquid; iti these regulators will divert the excess voltage flows to ground.

We can still absorb these regulators having filters because they have the sole purpose of diverting the post being used , currents that are not yet known but likely that we will call for simplicity, abnormal currents. [sic...]

The patent also covers systems for regulating currents, applied to each of the poles because it is recognized that the shape and tension of the currents of the two poles are not equivalent.

The method of capture is by special antenna has directed two points in the atmosphere, as will be stated after this, with Boot ecoulement prior to the current atmosphere is also of great importance.

Finally, the current atmospheric reception is adequate even with the installation of a post, which can be expected at any location without the need for special altitude, as has been attempted in previous efforts.

We refer to the attached drawings:.

Figure 1 represents in elevation and side view an antenna that only has two points, to overcome an elevation ;

Figure 2 is the end of one of the peaks, the horizontal:

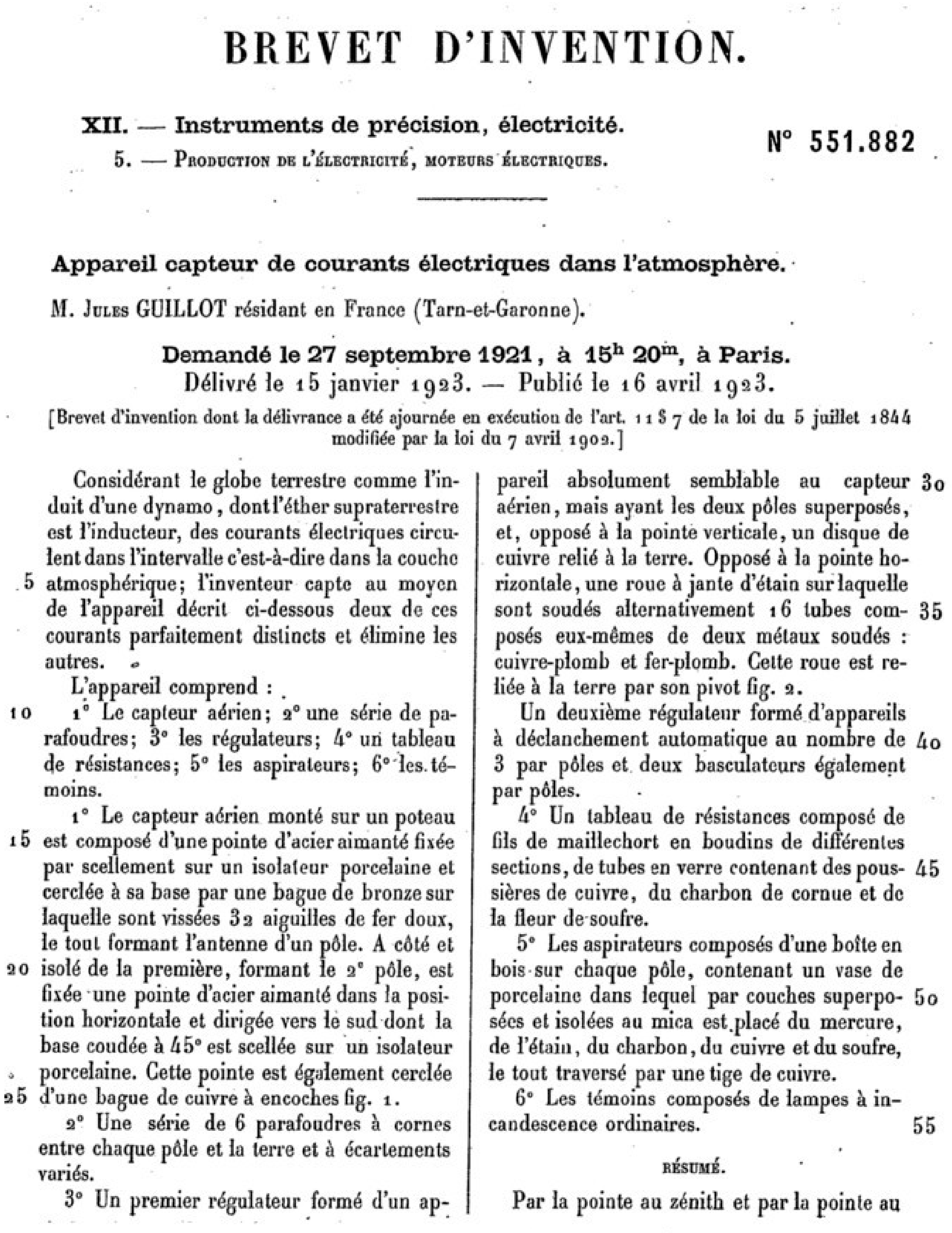

BREVET D'INVENTION.

XII. — Instruments de précision, électricité.

5. — Production de l'électricité, moteurs électriques.

N° 551.882

Appareil capteur de courants électriques dans l'atmosphère.

M. Jules GUILLOT résidant en France (Tarn-et-Garonne).

Demandé le 27 septembre 1921, à 15h 20m, à Paris.
Délivré le 15 janvier 1923. — Publié le 16 avril 1923.

[Brevet d'invention dont la délivrance a été ajournée en exécution de l'art. 11 § 7 de la loi du 5 juillet 1844 modifiée par la loi du 7 avril 1902.]

Considérant le globe terrestre comme l'induit d'une dynamo, dont l'éther supraterrestre est l'inducteur, des courants électriques circulent dans l'intervalle c'est-à-dire dans la couche atmosphérique; l'inventeur capte au moyen de l'appareil décrit ci-dessous deux de ces courants parfaitement distincts et élimine les autres.

L'appareil comprend :

1° Le capteur aérien; 2° une série de parafoudres; 3° les régulateurs; 4° un tableau de résistances; 5° les aspirateurs; 6° les témoins.

1° Le capteur aérien monté sur un poteau est composé d'une pointe d'acier aimanté fixée par scellement sur un isolateur porcelaine et cerclée à sa base par une bague de bronze sur laquelle sont vissées 32 aiguilles de fer doux, le tout formant l'antenne d'un pôle. A côté et isolé de la première, formant le 2° pôle, est fixée une pointe d'acier aimanté dans la position horizontale et dirigée vers le sud dont la base coudée à 45° est scellée sur un isolateur porcelaine. Cette pointe est également cerclée d'une bague de cuivre à encoches fig. 1.

2° Une série de 6 parafoudres à cornes entre chaque pôle et la terre et à écartements variés.

3° Un premier régulateur formé d'un appareil absolument semblable au capteur aérien, mais ayant les deux pôles superposés, et, opposé à la pointe verticale, un disque de cuivre relié à la terre. Opposé à la pointe horizontale, une roue à jante d'étain sur laquelle sont soudés alternativement 16 tubes composés eux-mêmes de deux métaux soudés : cuivre-plomb et fer-plomb. Cette roue est reliée à la terre par son pivot fig. 2.

Un deuxième régulateur formé d'appareils à déclanchement automatique au nombre de 3 par pôles et deux basculateurs également par pôles.

4° Un tableau de résistances composé de fils de maillechort en boudins de différentes sections, de tubes en verre contenant des poussières de cuivre, du charbon de cornue et de la fleur de soufre.

5° Les aspirateurs composés d'une boîte en bois sur chaque pôle, contenant un vase de porcelaine dans lequel par couches superposées et isolées au mica est placé du mercure, de l'étain, du charbon, du cuivre et du soufre, le tout traversé par une tige de cuivre.

6° Les témoins composés de lampes à incandescence ordinaires.

RÉSUMÉ.

Par la pointe au zénith et par la pointe au

Figure 3 shows schematically the table on which are mounted various devices or accessories that act as regulators of current tensions.

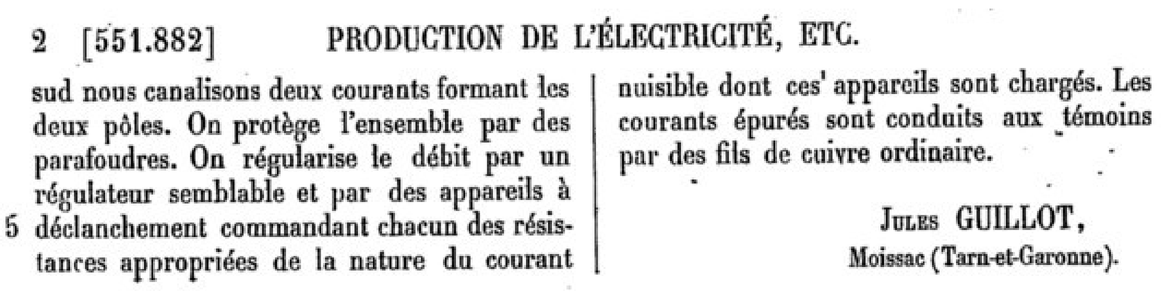

2 [551.882] PRODUCTION DE L'ÉLECTRICITÉ, ETC.

sud nous canalisons deux courants formant les deux pôles. On protège l'ensemble par des parafoudres. On régularise le débit par un régulateur semblable et par des appareils à déclanchement commandant chacun des résistances appropriées de la nature du courant nuisible dont ces appareils sont chargés. Les courants épurés sont conduits aux témoins par des fils de cuivre ordinaire.

Jules GUILLOT,
Moissac (Tarn-et-Garonne).

Figure 4 is a scheme of surge arrestors, and

Figure 5 a table of resistances.

Fig.3.

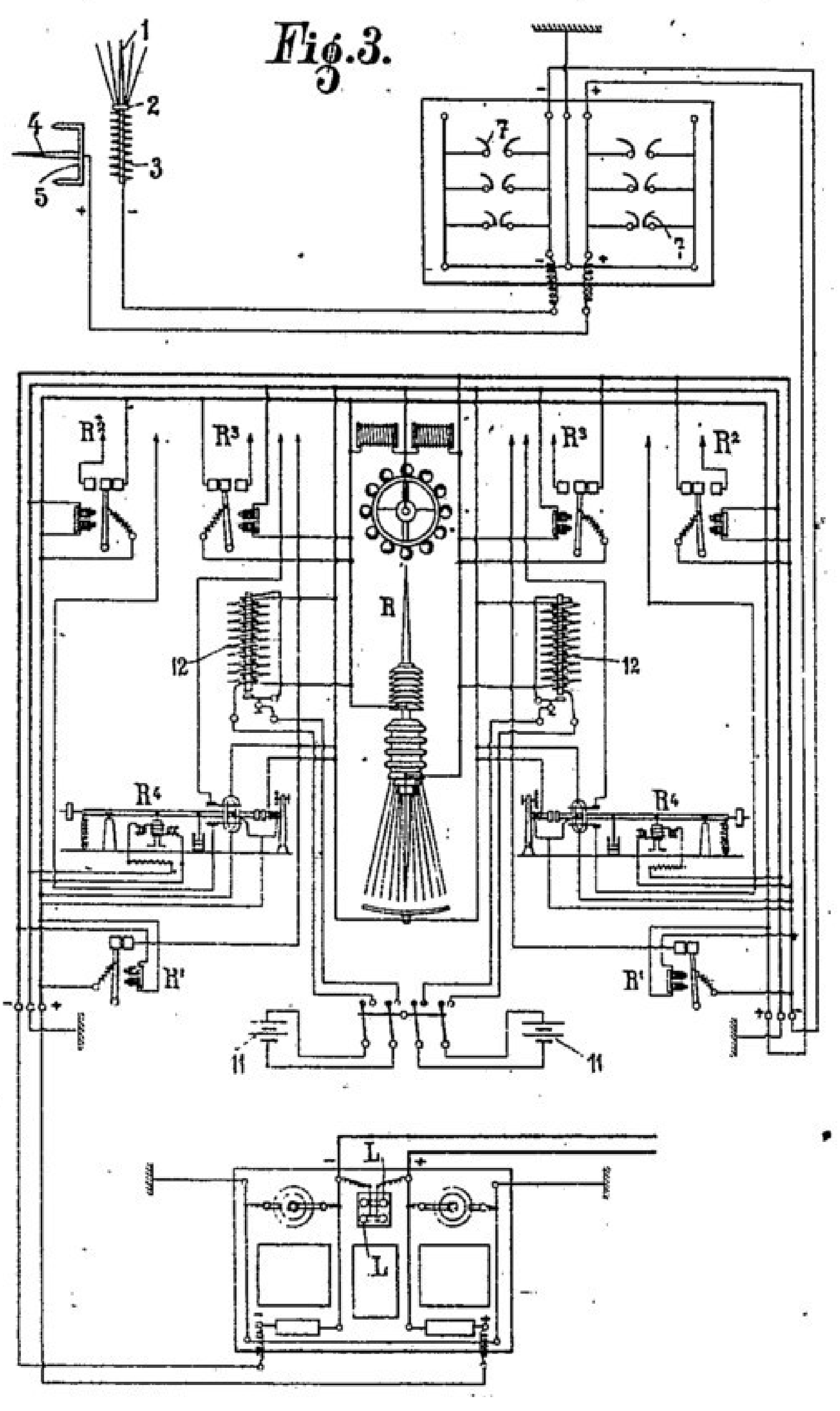

The necessary arrangements to capture atmospheric current form is as follows:

1. The **antenna** in figure 1 is formed with two peaks: the 1 is **steel**, supported by a **bronze ring** 2, surrounded by a **spiral form** 3; this point is oriented for attracting the **zenith** layers of **negative electricity** and the other edge 4 contains **copper coils** 6, and is oriented **south** to the **equator** to draw in the masses grouped in this area. To use this antenna it suffices to be raised on a mast or **pole** to a height a little higher than that of surrounding houses.

2. On the table **lightning arrestors**, contained 4, forms of 7 **copper horns** used to absorb large atmospheric discharges by certain time, as it would be dangerous to allow movement in the devices regulators.

3. On a table of **two** symmetrical systems **regulators**, Figure 3, one for each pole, to obtain and allow a regularization of large excesses in the current tensions, by automatically absorbing surges before passing over the

surfacea chosen for the desired use. These regulators are represented on the drawing by the references R, R1, R2, R3, R4.

4. In the **resistance** panel in communication with regulators, these resistances are formed as **spiral coils or screens** 8, **tubes 9 of glass filled with fine coal dust , aluminum powder, sulfur and fine copper dust**. Finally, **two** special 10 **processors** and constitutes **layers** of metal forming an **absorbant** for harmful or abnormal currents.

5. As a source of energy formed **auxiliary batteries** 11 and **Ruhmkorff coils** 12 to get into the antenna sufficient attraction by means of an adjustment of intensity, which once it begins, allows the constant flow of atmospheric current in the system.

In considering Figure 3, which represents a sort of an installation scheme of the device, the reception of atmospheric electrical current will, as set out above, producing a priming of these currents. This boot is made in launching the system which is double to answer each of the poles, the current batteries 11 through 12 and reels going to the antenna.

This boot will continue until the disposal of electrical current atmosphere is evident by the one or more table lamps L of resistance shown in Figure 5.

Once the lights illuminate it is advisable to charge the batteries to boot. Natural flow and constant currents will be captured by the regulators , Figure 3; they will regularize in tension and in amperage and then directed to use in table lamps.

It has been said above that regulators R, etc., are intended and designed to automatically eliminate earth surge currents capture and abnormal currents that cannot yet be level but that these regulators could be replaced by devices with the same function.

There is therefore no need to describe in detail and especially the claims. We only specify that these regulators must be proper protection for the purposes described above.

Fig. 4.

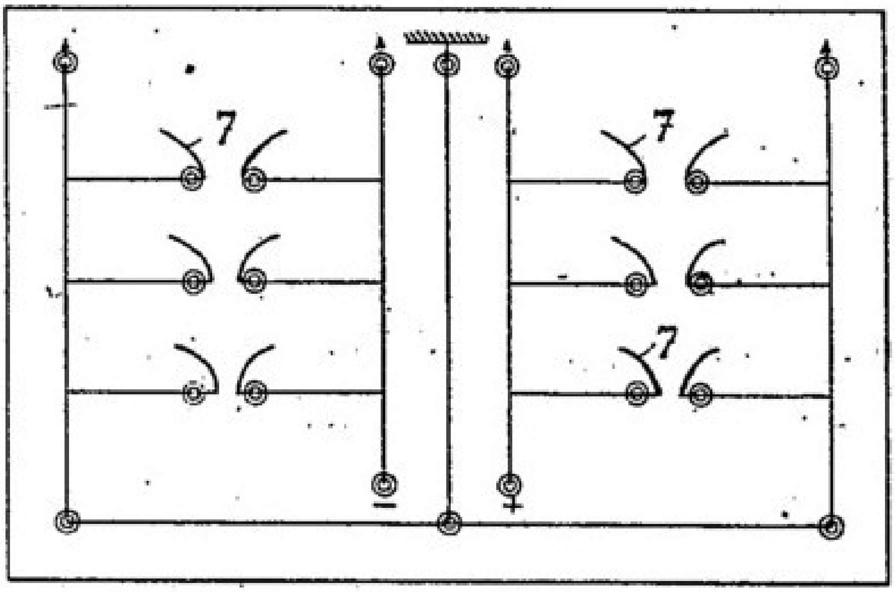

Fig. 5.

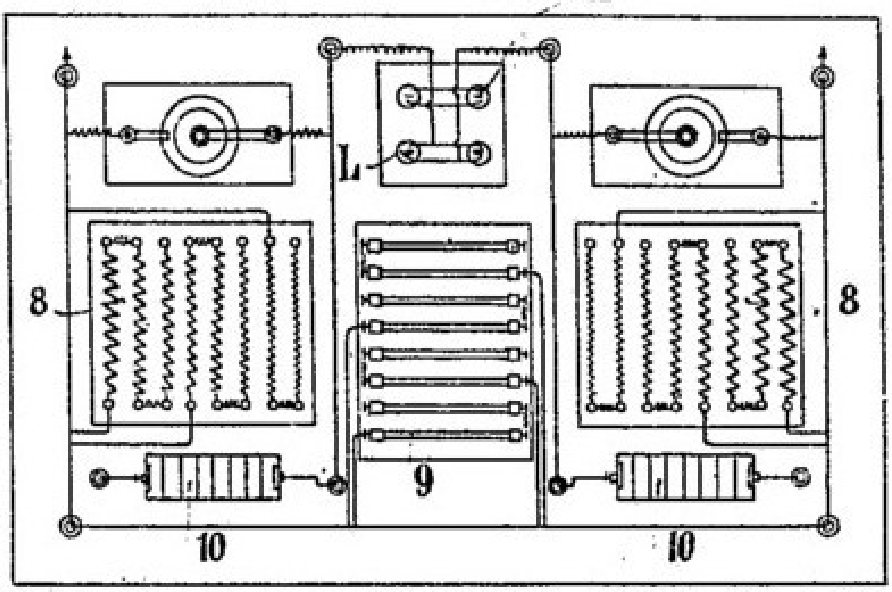

XII. — Instruments de précision, électricité.

5. — PRODUCTION DE L'ÉLECTRICITÉ, MOTEURS ÉLECTRIQUES.

N° 565.395

Dispositif combiné pour capter les courants électriques atmosphériques avec application immédiate.

M. JULES GUILLOT résidant en France (Allier).

Demandé le 6 avril 1923, à 11h 30m, à Paris.

Délivré le 7 novembre 1923. — Publié le 25 janvier 1924.

On sait que la terre, par sa propre constitution, sa rotation et son mouvement dans l'espace, fournit à elle seule l'électricité que l'on utilise et qui est en suspension dans l'atmosphère. Les courants électriques, qui s'échappent de la surface du globe, montent directement dans l'espace où les grandes masses sont accumulées vers l'équateur, par suite de la plus grande périphérie du globe.

Dans l'atmosphère, il y a deux pôles parfaitement distincts, c'est-à-dire que des nuages peuvent être électrisés, soit d'électricité positive ou négative; ceci, tout dans la nature le démontre, sans quoi, il serait difficile d'expliquer la provenance des éclairs qui se produisent entre les nuages et qui ne sont autres que des nuages chargés d'électricité contraire et qui s'amorcent entre eux par un trop grand rapprochement.

L'invention dont il s'agit concerne un dispositif formé d'un ensemble d'appareils aptes à la captation de ces courants atmosphériques.

Jusqu'à ce jour, toutes les recherches faites dans ce but ont été de capter dans l'atmosphère une seule catégorie de courants atmosphériques, c'est-à-dire, qu'on a réuni dans un même système de captation, deux pôles opposés et dont l'effet a été de se détruire ou de s'annuler l'un à l'autre laissant comme quantité que la différence du plus fort au plus faible.

Il sera donc facile de comprendre pourquoi, en considérant les deux pôles d'un courant atmosphérique comme à peu près égaux, il est presque impossible par certains temps de mesurer le potentiel avec la terre, du plus fort sur le plus faible subsistant et qui peut être positif ou négatif, suivant que le plus fort est de l'une ou de l'autre de ces polarités.

C'est sur cette base que toutes les recherches jusqu'à nos jours ont été faites et c'est pourquoi, il faut considérer que par tous les temps, entre les deux pôles captés dans l'atmosphère et non mélangés dans aucun des systèmes de captation, il est possible d'obtenir, en tension et en intensité, une puissance aussi considérable que l'on puisse désirer et qui sera mesurée entre les deux pôles captés séparément et toujours isolés l'un de l'autre jusqu'à leur utilisation.

Suivant l'invention, cette captation se fait par attraction dans l'atmosphère au moyen d'un amorçage des courants électriques, soit positif ou négatif, par une antenne à deux pointes absolument isolées l'une de l'autre, dont l'une tournée vers le zénith attire les couches d'électricité négative, et l'autre, tournée vers le sud, puise vers l'équateur, dans les masses groupées dans cette zone.

Ceci exposé et l'existence des courants électriques atmosphériques étant prouvée, il est

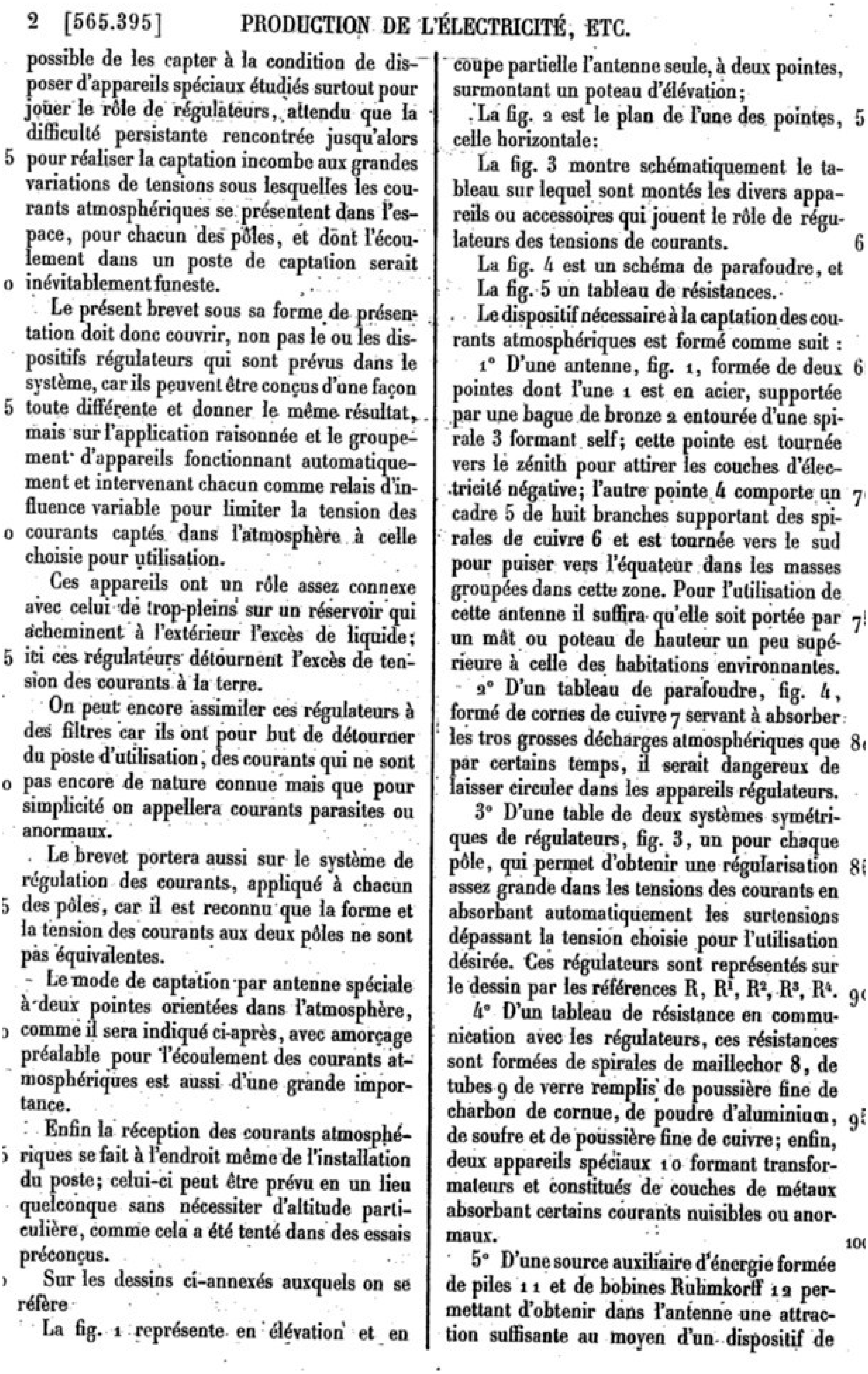

possible de les capter à la condition de disposer d'appareils spéciaux étudiés surtout pour jouer le rôle de régulateurs, attendu que la difficulté persistante rencontrée jusqu'alors pour réaliser la captation incombe aux grandes variations de tensions sous lesquelles les courants atmosphériques se présentent dans l'espace, pour chacun des pôles, et dont l'écoulement dans un poste de captation serait inévitablement funeste.

Le présent brevet sous sa forme de présentation doit donc couvrir, non pas le ou les dispositifs régulateurs qui sont prévus dans le système, car ils peuvent être conçus d'une façon toute différente et donner le même résultat, mais sur l'application raisonnée et le groupement d'appareils fonctionnant automatiquement et intervenant chacun comme relais d'influence variable pour limiter la tension des courants captés dans l'atmosphère à celle choisie pour utilisation.

Ces appareils ont un rôle assez connexe avec celui de trop-pleins sur un réservoir qui acheminent à l'extérieur l'excès de liquide; iti ces régulateurs détournent l'excès de tension des courants à la terre.

On peut encore assimiler ces régulateurs à des filtres car ils ont pour but de détourner du poste d'utilisation, des courants qui ne sont pas encore de nature connue mais que pour simplicité on appellera courants parasites ou anormaux.

Le brevet portera aussi sur le système de régulation des courants, appliqué à chacun des pôles, car il est reconnu que la forme et la tension des courants aux deux pôles ne sont pas équivalentes.

Le mode de captation par antenne spéciale à deux pointes orientées dans l'atmosphère, comme il sera indiqué ci-après, avec amorçage préalable pour l'écoulement des courants atmosphériques est aussi d'une grande importance.

Enfin la réception des courants atmosphériques se fait à l'endroit même de l'installation du poste; celui-ci peut être prévu en un lieu quelconque sans nécessiter d'altitude particulière, comme cela a été tenté dans des essais préconçus.

Sur les dessins ci-annexés auxquels on se réfère

La fig. 1 représente en élévation et en coupe partielle l'antenne seule, à deux pointes, surmontant un poteau d'élévation;

La fig. 2 est le plan de l'une des pointes, celle horizontale:

La fig. 3 montre schématiquement le tableau sur lequel sont montés les divers appareils ou accessoires qui jouent le rôle de régulateurs des tensions de courants.

La fig. 4 est un schéma de parafoudre, et

La fig. 5 un tableau de résistances.

Le dispositif nécessaire à la captation des courants atmosphériques est formé comme suit :

1° D'une antenne, fig. 1, formée de deux pointes dont l'une 1 est en acier, supportée par une bague de bronze 2 entourée d'une spirale 3 formant self; cette pointe est tournée vers le zénith pour attirer les couches d'électricité négative; l'autre pointe 4 comporte un cadre 5 de huit branches supportant des spirales de cuivre 6 et est tournée vers le sud pour puiser vers l'équateur dans les masses groupées dans cette zone. Pour l'utilisation de cette antenne il suffira qu'elle soit portée par un mât ou poteau de hauteur un peu supérieure à celle des habitations environnantes.

2° D'un tableau de parafoudre, fig. 4, formé de cornes de cuivre 7 servant à absorber les tros grosses décharges atmosphériques que par certains temps, il serait dangereux de laisser circuler dans les appareils régulateurs.

3° D'une table de deux systèmes symétriques de régulateurs, fig. 3, un pour chaque pôle, qui permet d'obtenir une régularisation assez grande dans les tensions des courants en absorbant automatiquement les surtensions dépassant la tension choisie pour l'utilisation désirée. Ces régulateurs sont représentés sur le dessin par les références R, R^1, R^2, R^3, R^4.

4° D'un tableau de résistance en communication avec les régulateurs, ces résistances sont formées de spirales de maillechor 8, de tubes 9 de verre remplis de poussière fine de charbon de cornue, de poudre d'aluminium, de soufre et de poussière fine de cuivre; enfin, deux appareils spéciaux 10 formant transformateurs et constitués de couches de métaux absorbant certains courants nuisibles ou anormaux.

5° D'une source auxiliaire d'énergie formée de piles 11 et de bobines Ruhmkorff 12 permettant d'obtenir dans l'antenne une attraction suffisante au moyen d'un dispositif de

DARIDA, Gillbert : "Practical Utilization of Atmospheric Electricity";
(in *The Invention Encyclopedia,* pp. 204-207 ; 1930, Geo. Constantinescu, Ed.) --

possible de les capter à la condition de disposer d'appareils spéciaux étudiés surtout pour jouer le rôle de régulateurs, attendu que la difficulté persistante rencontrée jusqu'alors pour réaliser la captation incombe aux grandes variations de tensions sous lesquelles les courants atmosphériques se présentent dans l'espace, pour chacun des pôles, et dont l'écoulement dans un poste de captation serait inévitablement funeste.

Le présent brevet sous sa forme de présentation doit donc couvrir, non pas le ou les dispositifs régulateurs qui sont prévus dans le système, car ils peuvent être conçus d'une façon toute différente et donner le même résultat, mais sur l'application raisonnée et le groupement d'appareils fonctionnant automatiquement et intervenant chacun comme relais d'influence variable pour limiter la tension des courants captés dans l'atmosphère à celle choisie pour utilisation.

Ces appareils ont un rôle assez connexe avec celui de trop-pleins sur un réservoir qui acheminent à l'extérieur l'excès de liquide; ici ces régulateurs détournent l'excès de tension des courants à la terre.

On peut encore assimiler ces régulateurs à des filtres car ils ont pour but de détourner du poste d'utilisation, des courants qui ne sont pas encore de nature connue mais que pour simplicité on appellera courants parasites ou anormaux.

Le brevet portera aussi sur le système de régulation des courants, appliqué à chacun des pôles, car il est reconnu que la forme et la tension des courants aux deux pôles ne sont pas équivalentes.

Le mode de captation par antenne spéciale à deux pointes orientées dans l'atmosphère, comme il sera indiqué ci-après, avec amorçage préalable pour l'écoulement des courants atmosphériques est aussi d'une grande importance.

Enfin la réception des courants atmosphériques se fait à l'endroit même de l'installation du poste; celui-ci peut être prévu en un lieu quelconque sans nécessiter d'altitude particulière, comme cela a été tenté dans des essais préconçus.

Sur les dessins ci-annexés auxquels on se réfère

La fig. 1 représente en élévation et en coupe partielle l'antenne seule, à deux pointes, surmontant un poteau d'élévation;

La fig. 2 est le plan de l'une des pointes, celle horizontale:

La fig. 3 montre schématiquement le tableau sur lequel sont montés les divers appareils ou accessoires qui jouent le rôle de régulateurs des tensions de courants.

La fig. 4 est un schéma de parafoudre, et

La fig. 5 un tableau de résistances.

Le dispositif nécessaire à la captation des courants atmosphériques est formé comme suit :

1° D'une antenne, fig. 1, formée de deux pointes dont l'une 1 est en acier, supportée par une bague de bronze 2 entourée d'une spirale 3 formant self; cette pointe est tournée vers le zénith pour attirer les couches d'électricité négative; l'autre pointe 4 comporte un cadre 5 de huit branches supportant des spirales de cuivre 6 et est tournée vers le sud pour puiser vers l'équateur dans les masses groupées dans cette zone. Pour l'utilisation de cette antenne il suffira qu'elle soit portée par un mât ou poteau de hauteur un peu supérieure à celle des habitations environnantes.

2° D'un tableau de parafoudre, fig. 4, formé de cornes de cuivre 7 servant à absorber les tros grosses décharges atmosphériques que par certains temps, il serait dangereux de laisser circuler dans les appareils régulateurs.

3° D'une table de deux systèmes symétriques de régulateurs, fig. 3, un pour chaque pôle, qui permet d'obtenir une régularisation assez grande dans les tensions des courants en absorbant automatiquement les surtensions dépassant la tension choisie pour l'utilisation désirée. Ces régulateurs sont représentés sur le dessin par les références R, R^1, R^2, R^3, R^4.

4° D'un tableau de résistance en communication avec les régulateurs, ces résistances sont formées de spirales de maillechor 8, de tubes 9 de verre remplis de poussière fine de charbon de cornue, de poudre d'aluminium, de soufre et de poussière fine de cuivre; enfin, deux appareils spéciaux 10 formant transformateurs et constitués de couches de métaux absorbant certains courants nuisibles ou anormaux.

5° D'une source auxiliaire d'énergie formée de piles 11 et de bobines Ruhmkorff 12 permettant d'obtenir dans l'antenne une attraction suffisante au moyen d'un dispositif de

2 [565.395] PRODUCTION DE L'ÉLECTRICITÉ, ETC.

possible de les capter à la condition de disposer d'appareils spéciaux étudiés surtout pour jouer le rôle de régulateurs, attendu que la difficulté persistante rencontrée jusqu'alors pour réaliser la captation incombe aux grandes variations de tensions sous lesquelles les courants atmosphériques se présentent dans l'espace, pour chacun des pôles, et dont l'écoulement dans un poste de captation serait inévitablement funeste.

Le présent brevet sous sa forme de présentation doit donc couvrir, non pas le ou les dispositifs régulateurs qui sont prévus dans le système, car ils peuvent être conçus d'une façon toute différente et donner le même résultat, mais sur l'application raisonnée et le groupement d'appareils fonctionnant automatiquement et intervenant chacun comme relais d'influence variable pour limiter la tension des courants captés dans l'atmosphère à celle choisie pour utilisation.

Ces appareils ont un rôle assez connexe avec celui de trop-pleins sur un réservoir qui acheminent à l'extérieur l'excès de liquide; ici ces régulateurs détournent l'excès de tension des courants à la terre.

On peut encore assimiler ces régulateurs à des filtres car ils ont pour but de détourner du poste d'utilisation, des courants qui ne sont pas encore de nature connue mais que pour simplicité on appellera courants parasites ou anormaux.

Le brevet portera aussi sur le système de régulation des courants, appliqué à chacun des pôles, car il est reconnu que la forme et la tension des courants aux deux pôles ne sont pas équivalentes.

Le mode de captation par antenne spéciale à deux pointes orientées dans l'atmosphère, comme il sera indiqué ci-après, avec amorçage préalable pour l'écoulement des courants atmosphériques est aussi d'une grande importance.

Enfin la réception des courants atmosphériques se fait à l'endroit même de l'installation du poste; celui-ci peut être prévu en un lieu quelconque sans nécessiter d'altitude particulière, comme cela a été tenté dans des essais préconçus.

Sur les dessins ci-annexés auxquels on se réfère

La fig. 1 représente en élévation et en coupe partielle l'antenne seule, à deux pointes, surmontant un poteau d'élévation;

La fig. 2 est le plan de l'une des pointes, celle horizontale:

La fig. 3 montre schématiquement le tableau sur lequel sont montés les divers appareils ou accessoires qui jouent le rôle de régulateurs des tensions de courants.

La fig. 4 est un schéma de parafoudre, et

La fig. 5 un tableau de résistances.

Le dispositif nécessaire à la captation des courants atmosphériques est formé comme suit :

1° D'une antenne, fig. 1, formée de deux pointes dont l'une 1 est en acier, supportée par une bague de bronze 2 entourée d'une spirale 3 formant self; cette pointe est tournée vers le zénith pour attirer les couches d'électricité négative; l'autre pointe 4 comporte un cadre 5 de huit branches supportant des spirales de cuivre 6 et est tournée vers le sud pour puiser vers l'équateur dans les masses groupées dans cette zone. Pour l'utilisation de cette antenne il suffira qu'elle soit portée par un mât ou poteau de hauteur un peu supérieure à celle des habitations environnantes.

2° D'un tableau de parafoudre, fig. 4, formé de cornes de cuivre 7 servant à absorber les tros grosses décharges atmosphériques que par certains temps, il serait dangereux de laisser circuler dans les appareils régulateurs.

3° D'une table de deux systèmes symétriques de régulateurs, fig. 3, un pour chaque pôle, qui permet d'obtenir une régularisation assez grande dans les tensions des courants en absorbant automatiquement les surtensions dépassant la tension choisie pour l'utilisation désirée. Ces régulateurs sont représentés sur le dessin par les références R, R^1, R^2, R^3, R^4.

4° D'un tableau de résistance en communication avec les régulateurs, ces résistances sont formées de spirales de maillechor 8, de tubes 9 de verre remplis de poussière fine de charbon de cornue, de poudre d'aluminium, de soufre et de poussière fine de cuivre; enfin, deux appareils spéciaux 10 formant transformateurs et constitués de couches de métaux absorbant certains courants nuisibles ou anormaux.

5° D'une source auxiliaire d'énergie formée de piles 11 et de bobines Ruhmkorff 12 permettant d'obtenir dans l'antenne une attraction suffisante au moyen d'un dispositif de

réglage d'intensité qui, une fois amorcé, permet l'écoulement constant des courants de l'atmosphère dans le système.

En considérant la fig. 3 qui représente en quelque sorte le schéma d'installation du dispositif, la réception des courants électriques atmosphériques se fera de la façon indiquée ci-dessus, en produisant un amorçage de ces courants. Cet amorçage est fait en lançant dans le système qui est double pour répondre à chacun des pôles, le courant des piles 11 passant par les bobines 12 et se rendant à l'antenne.

Cet amorçage sera maintenu et réglé jusqu'à ce que l'écoulement des courants électriques atmosphériques se soit manifesté par l'allumage d'une ou plusieurs lampes L du tableau des résistances représenté en fig. 5 portant la prise d'utilisation.

Dès que l'allumage des lampes sera constaté on sera averti de la captation des courants atmosphériques et on pourra couper le courant des piles d'amorçage. L'écoulement naturel et constant des courants captés se produira dans les appareils régulateurs de la table, fig. 3; ils seront régularisés en tension et en nature puis dirigés au tableau des lampes pour utilisation.

Il a été dit ci-dessus que les régulateurs R, etc., étaient prévus et conçus pour éliminer automatiquement à la terre les surtensions des courants captés ainsi que les courants anormaux qui ne peuvent encore être classés. mais que ces régulateurs pouvaient être remplacés par des appareils faisant le même office.

Il n'y a donc pas lieu de les décrire en détail et de les revendiquer spécialement. L'avenir seul précisera si ces régulateurs doivent faire l'objet d'une protection propre pour l'application décrite ci-dessus.

RÉSUMÉ.

Dispositif combiné pour capter les courants électriques atmosphériques avec application immédiate, caractérisé :

1° En ce que l'organe de captation est une antenne à deux branches orientées différemment, l'une vertical avec self est tournée vers le zénith pour puiser les courants négatifs, l'autre horizontale comporte un plateau central avec enroulement de fil et est orientée vers le sud pour puiser les courants positifs; l'orientation des pointes restant fonction du lieu terrestre où se fait la captation des courants considérés.

2° En ce que l'écoulement des courants atmosphériques dans le système récepteur est subordonné à un amorçage qui consiste à lancer dans l'antenne un courant provenant d'une source auxiliaire, telle que des piles reliées à des bobines de Ruhmkorff.

3° En ce qu'un tableau parafoudre est interposé entre l'antenne ou organe de captation et le tableau des régulateurs pour préserver l'installation en éliminant à la terre les décharges brusques atmosphériques.

4° En ce que les régulateurs sont disposés suivant deux systèmes symétriques correspondant à chacun des pôles des courants atmosphériques.

J. GUILLOT.

Par procuration :

V. Prévost.

French Patent # 551,882

Appareil Capteur de Courants Electriques dans L'Atmosphere

The classic spark-gap with magnetic blowing circuit

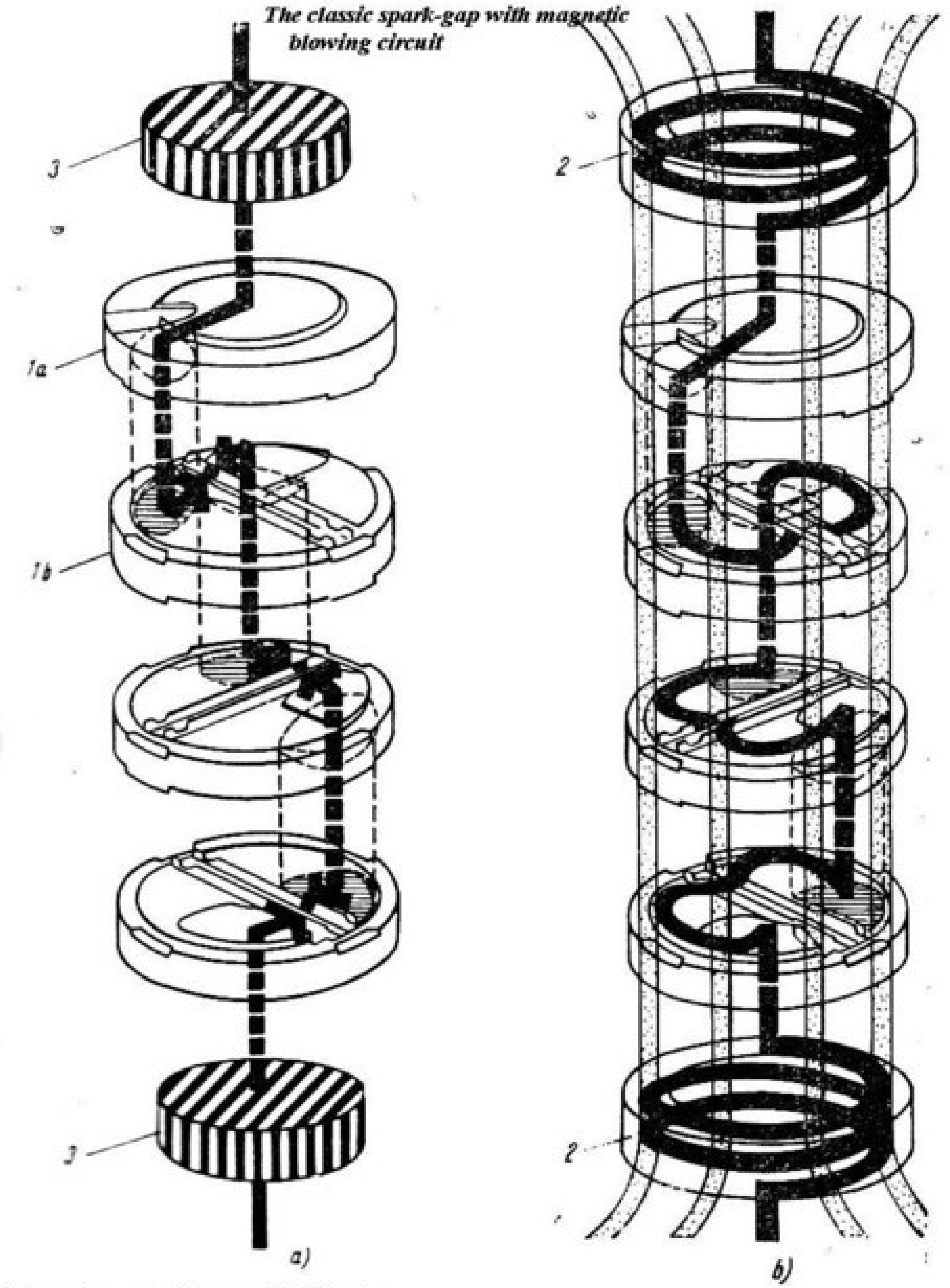

The spark-gap with magnetic blowing :
a,- the initial stage ; the impulse current send in soil through the shunt-resistences 3, because the coils 2 have an impedance too big for the impulse of 10/20 μs ;
the spark-gap 1 has two elements disc-shaped 1a and 1b, the inferior disc have one electrode and the upper disc the other electrode of the spark-gap circuit.
b,- the second stage ; the accompanying current discharge - after the impulse current ;
this accompanying current pass through the coils 2 which produce the magnetic field with the role to extend the electric arc into the slit between insulator discs 1a and 1b ; the thickness of this slit is smaller and smaller in accord with the extension of the electric arc, in this manner producing the magnifing of its electric resistance and finally, the electric arc extinction.

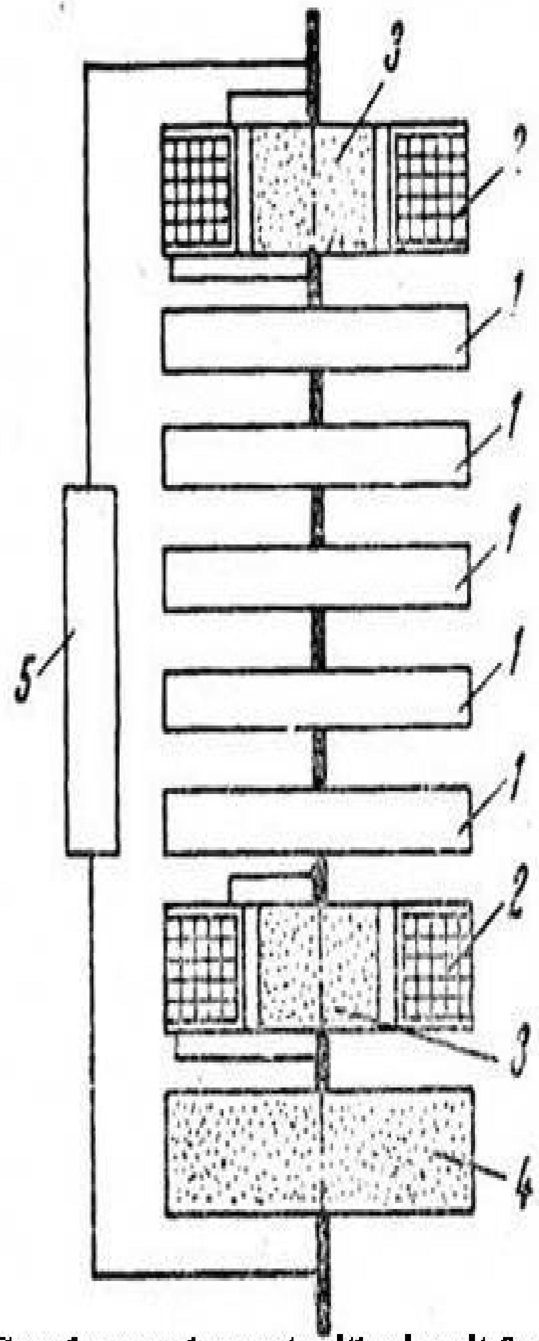

Spark-gap element with circuit for the magnetic blowing of the electric fascicle

1,- the spark-gap individual devices which work in the manner to extend the electric arc of the accompanying electric current ;

2,- the coils which produce the magnetic field for the magnetic blowing ;

3,- the electric shunt-resistances in parallel connection on the blowing coils, with the role of reducing the current inside this coils

4,- the disc with unlinear resistance ;

5,- electrical resistance for the homogenous distribution of the electric field in all the elements of the spark-gap.

In Jules Guillot scheme, that kind of device could be used for leading the current.

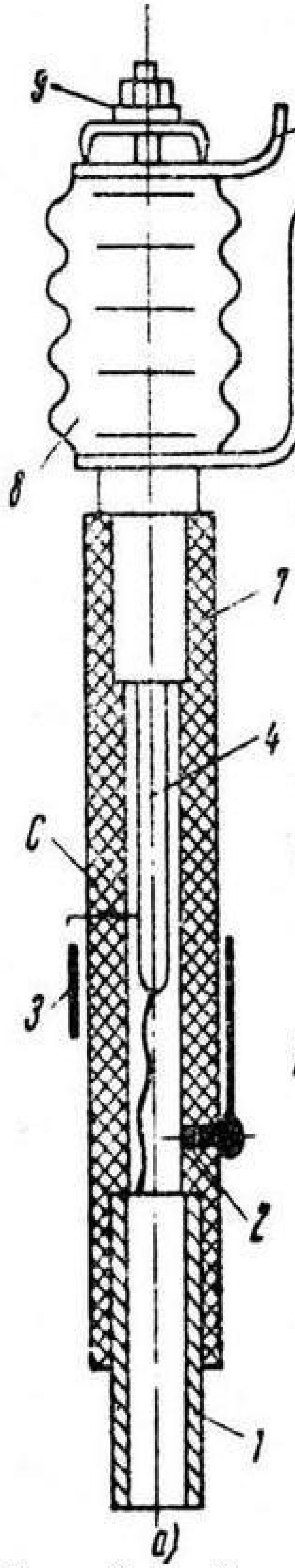

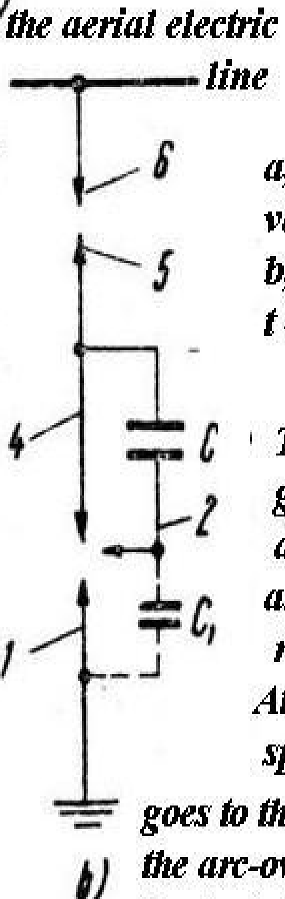

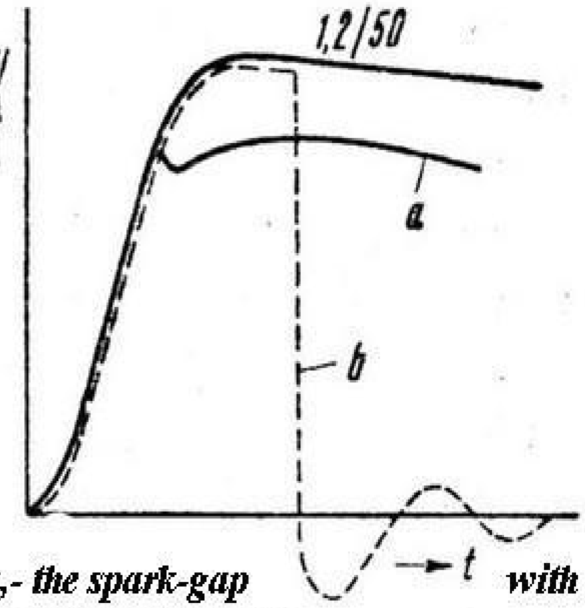

a,- the spark-gap with variable resistance ;
b,- the tube spak-gap ;
t - the time

The tube spark-gap has the aerial spark gap with the electrodes 5 and 6 disposed at the insulator 8 extremities ; and one arc-over spark-gap device between the rod 4 and the electrode 1.

At start, the first spark apear in the aerial spark-gap and then, the electric dicharge goes to the gas-producing tube 7, where it makes the arc-over. The local pressure in the gas allow the electric arc extinction, after managing the over electric tension into soil. The capacity C which is created between the external ring 3 and the rod 4 is bigger than the Capacity C_1 between 1 an 2 in that manner, when the overtension apears, the auxiliary spark-gap 1, 2 will start working and then, the main spark-gap 1, 4. In this circuit, the electrode 9 is connected at the aerial electric line and the electrode 1 at the Earth.

In the image above, it is an classic tube spark-gap used for the protection of aerial electric cables. The atmospheric electricity is discharged in soil and lost... The Guillot's inventions used that kind of devices for leading and absorbing of the atmospheric electricity.

Pământul ca și celelalte astre, posedă o electricitate proprie negativă în interior, iar aerul care-l înconjoară este pozitiv. Potențialul sau tensiunea electrică a aerului se mărește pe măsură ce ne ridicăm în înălțime, astfel că putem spune că intensitatea electrică este proporțională cu înălțimea (după lucrările remarcabile ale lui Franklin, Quetelet, Lord Kelvin, Mascart, Joubert și a altor mari savanți fizicieni).

Observațiile făcute recent, au arătat că aerul regiunilor înalte dela 6000—7000 m este foarte încărcat cu electricitate pozitivă, a cărei explicație se datorește frecării fotosferei sau atmosferei terestre, prin învârtirea pământului în jurul soarelui cu mai mult de o sută de mii de km pe oră. Această frecare directă în fluidul eteric ambiant, face ca aerul să se încarce cu electricitate pozitivă prin influență, iar pământul să se încarce negativ.

Intre cele două medii, fluidic (aerul) și (pământul) materie, știința ne spune, că în partea de jos a atmosferei, pe timp frumos un centimetru cub de aer ionizat, conține circa 800 particule cu ioni pozitivi și 680 ioni negativi (electroni).

Pământul se comportă ca un enorm conductor încărcat negativ, respingând electroni și atrăgând ionii pozitivi. Această atracție a ionilor pozitivi determină un curent electric, zis și curent de convențiune. E un fel de bombardament invizibil supus variațiunilor zilei și a sezoanelor, putându-l evalua aproximativ la o densitate mijlocie remarcabilă, dela 3×10^{16} sau a 30 quatrilionime de amper pe cm cub, ceea ce dă pentru întreagă suprafață a pământului un curent formidabil, de 1500 de amperi.

Se pune întrebarea cum un astfel de curent poate să-și întreție mereu același sens? Aceasta este enigma pe care ne-o punem, studiind curenții naturali ai atmosferei noastre, presupunem că poate să fie o legătură a acțiunii emanațiunilor radioactive. Aceste

Utilizarea practică a electricității atmosferice

Pământul ca și celelalte astre, posedă o electricitate proprie negativă în interior, iar aerul care-l înconjoară este pozitiv. Potențialul sau tensiunea electrică a aerului se mărește pe măsură ce ne ridicăm în înălțime, astfel că putem spune că intensitatea electrică este proporțională cu înălțimea (după lucrările remarcabile ale lui Franklin, Quetelet, Lord Kelvin, Mascart, Joubert și a altor mari savanți fizicieni).

Observațiile făcute recent, au arătat că aerul regiunilor înalte dela 6000—7000 m este foarte încărcat cu electricitate pozitivă, a cărei explicație se datorește frecării fotosferei sau atmosferei terestre, prin învârtirea pământului în jurul soarelui cu mai mult de o sută de mii de km pe oră. Această frecare directă în fluidul eteric ambiant, face ca aerul să se încarce cu electricitate pozitivă prin influență, iar pământul să se încarce negativ.

Intre cele două medii, fluidic (aerul) și (pământul) materie, știința ne spune, că în partea de jos a atmosferei, pe timp frumos un centimetru cub de aer ionizat, conține circa 800 particule cu ioni pozitivi și 680 ioni negativi (electroni).

Pământul se comportă ca un enorm conductor încărcat negativ, respingând electroni și atrăgând ionii pozitivi. Această atracție a ionilor pozitivi determină un curent electric, zis și curent de convențiune. E un fel de bombardament invizibil supus variațiunilor zilei și a sezoanelor, putându-l evalua aproximativ la o densitate mijlocie remarcabilă, dela 3×10^{16} sau a 30 quatrilionime de amper pe cm cub, ceea ce dă pentru întreagă suprafață a pământului un curent formidabil, de 1500 de amperi.

Se pune întrebarea cum un astfel de curent poate să-și întreție mereu același sens? Aceasta este enigma pe care ne-o punem, studiind curenții naturali ai atmosferei noastre, presupunem că poate să fie o legătură a acțiunii emanațiunilor radioactive. Aceste

prin ajutorul unui aparat de captare, format din 2 antene aeriene și o serie de bobine după cum se vede în fig. 108. Una din antene este verticală în formă de evantaiu dirijată cu vârfurile către zenit, pentru a capta electricitatea negativă din atmosferă; cea care este orizontală, este dirijată către Sud, pentru a capta electricitate pozitivă.

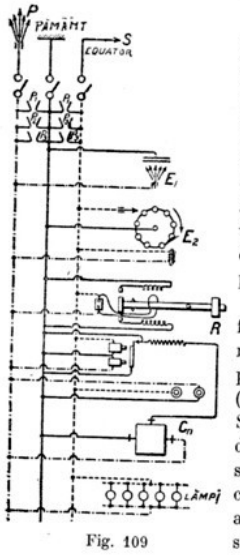

Fig. 109

Electricitatea în aer pare să fie cu dublă polaritate, după cum se vede la 2 nori electrizați, cari deplasându-se în sens contrar, dau naștere unei scântei (fulger) când distanța respectivă poate să fie străbătută de puternica lor diferență de potențial. Straturile de aer de densități diferite, nu sunt toate antrenate de aceeași viteză prin rotația pământului, ele se electrizează deci prin frecarea decalajului lor respectiv.

Aceste sunt considerațiunile ce l-au făcut pe inventator să așeze în atmosferă doi poli complet distincți și perfect izolați, al cărui pol pozitiv *S* (fig. 109) poate fi dirijat în direcțiunea Sudului, și a celuilalt negativ *P*, în direcțiunea zenitului. Fig. 109 arată schema montajului și a conexiunilor compuse după cum urmează: 1. dintr'o antenă dublă aeriană, montată pe stâlpi de 15—20 m înălțime. Această parte care se mai numește și aparat de captat, este format din vârfurile *P*, dispuse în evantai în vârful unei tije de oțel dirijate către zenit, pentru a capta curenții negativi ai atmosferei și dintr'o altă antenă orizontală a cărei săgeată este îndreptată către sud *S*, adică către equator, pentru a recepta curenții pozitivi. Aceste două părți dispuse în unghiu de 90°, formează cei doi poli al curentului aerian, a cărei diferență de potențial este suficientă pentru a permite utilizarea electricității atmosferice.

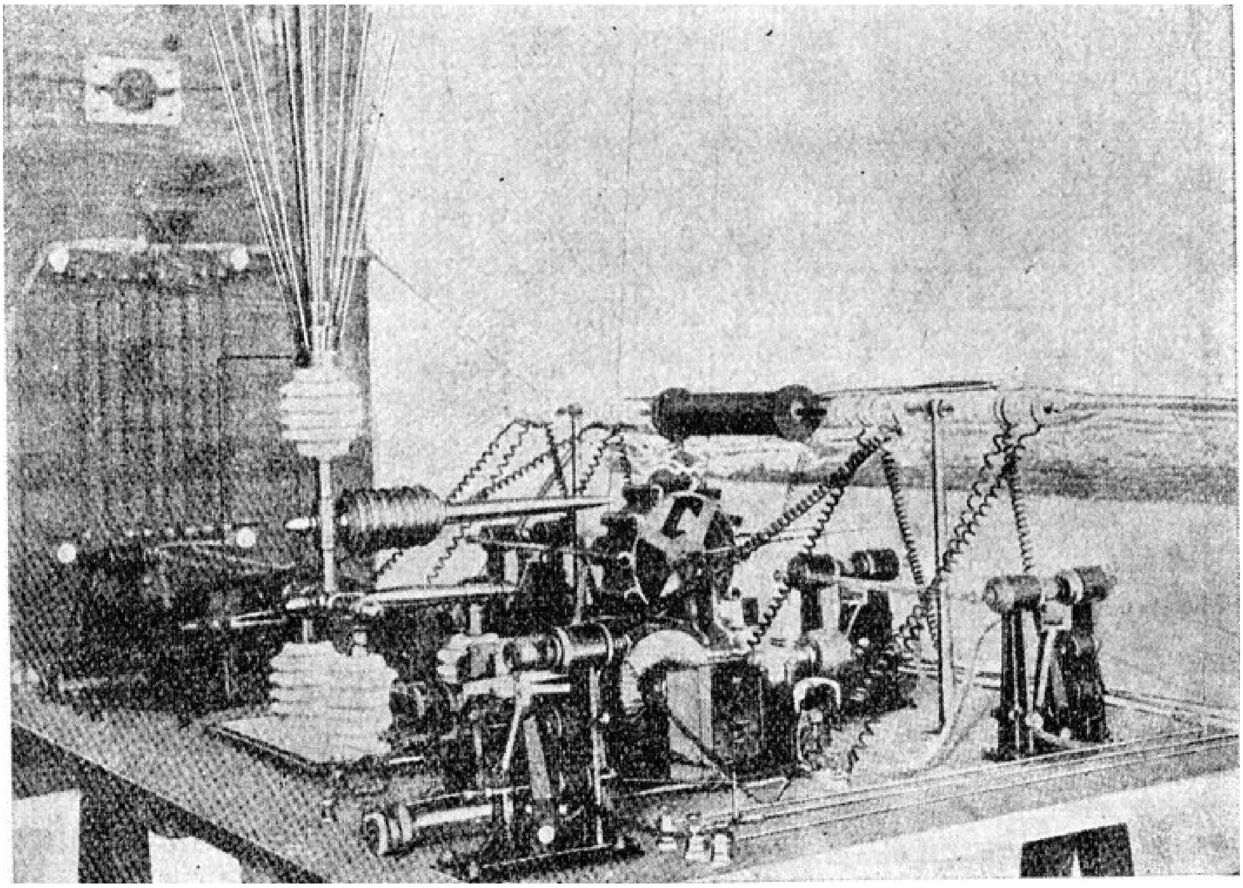

Parts List

Steel Rods, magnetic, 2
Insulator, porcelain
Ring, bronze
Rods, soft iron, 32
Ring, copper
Surge Protector
Disk, copper
Ring, tin
Tubes, steel
Rx Cu-Pb, Fe-Pb
Breakers, automatic
Resistances : Wire Mesh, Glass Tube. Cu, C, S powder
Vacuum Cleaners : Box, Jar, Mica &
Tube, copper;
Rx : Hg, Sn, C, Cu, S
Battery
Ignition Coil

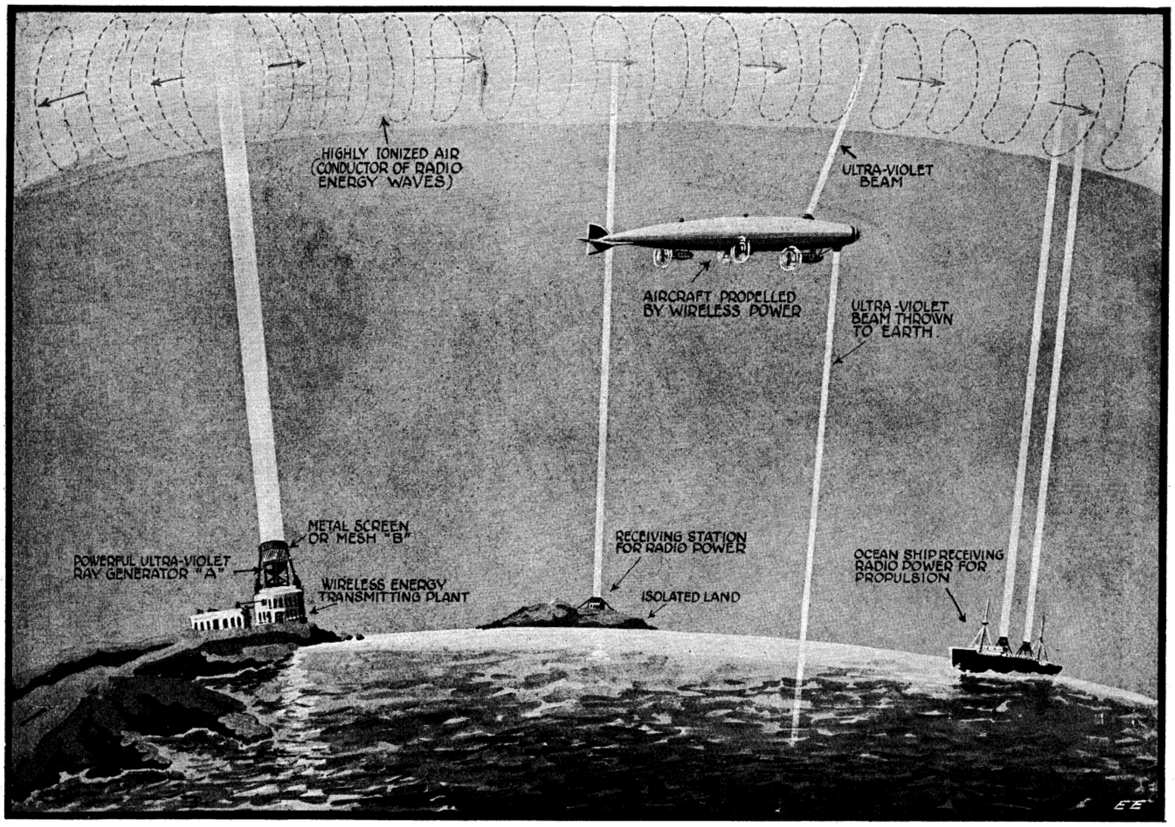
HIGHLY IONIZED AIR (CONDUCTOR OF RADIO ENERGY WAVES)
ULTRA-VIOLET BEAM
AIRCRAFT PROPELLED BY WIRELESS POWER
ULTRA-VIOLET BEAM THROWN TO EARTH
METAL SCREEN OR MESH "B"
POWERFUL ULTRA-VIOLET RAY GENERATOR "A"
WIRELESS ENERGY TRANSMITTING PLANT
RECEIVING STATION FOR RADIO POWER
ISOLATED LAND
OCEAN SHIP RECEIVING RADIO POWER FOR PROPULSION
EE

Showing the Wireless Transmission of Power by Means of Ionized Searchlight Beams. Powerful Tesla Currents of High Potential and Frequency Are Imprest on the Ionized Beams and Received from Similar Beams as Shown.

Wireless Transmission of Power Now Possible

By THOMAS W. BENSON

RECENT developments tend to fulfill the old, old dream of power transmission without wires. For years men have labored in vain to transmit power without wires and thus solve a myriad of transportation problems, such as propelling vehicles, ships, trains, aircraft, etc., without having to include a source of power in their construction. It makes one's head almost reel to think of the possibilities in such an achievement and now the dream promises to come true.

By one bold stroke an English genius, Mr. John Hettinger, has unthinkingly opened up the way for its realization. Like numerous other inventions, the way lay before our very eyes, but we were all too blind to see it.

While working on a means for transmitting radio waves without material aerials, Mr. Hettinger hit upon the idea of using ionized light beams as elevated conductors. Obviously simple when one thinks of it.

It is a well-known fact that an ionized gas is a conductor; in fact, the conductivity of a gas is a measure of its ionization. Furthermore, it is possible to ionize a gas by means of a stream of ultra violet rays. Working with these facts, Mr. Hettinger devised means to utilize them for radio transmission in the following manner.

As shown in Fig. 1, a source of ultra violet rays, an arc or mercury vapor lamp, A, is arranged to throw a beam of ionizing rays vertically.

These rays result in an ionized stream of air that acts as a conductor, the surrounding un-ionized air being practically a perfect insulator. By making connection to this conducting beam with a metallic screen or mesh at B, it can be utilized as an aerial for radio communication.

We are perfectly aware that this ionized beam rapidly loses its conducting power

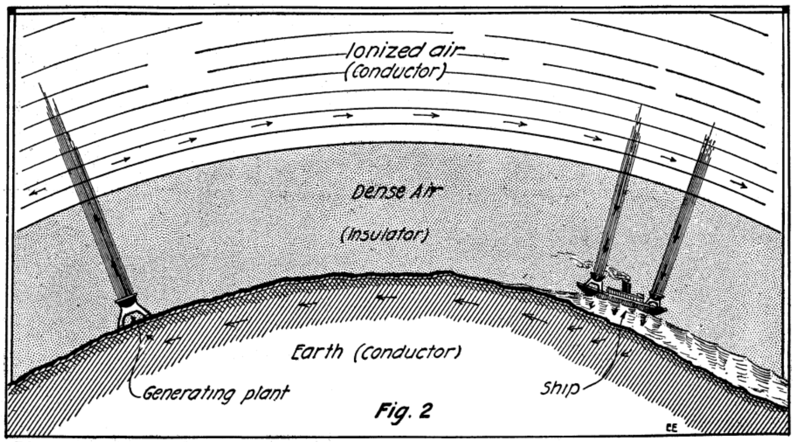
Ionized air (Conductor)
Dense Air (Insulator)
Earth (Conductor)
Generating plant
Ship
CE

Fig. 2

Diagram of the Author's Scheme for the Wireless Transmission of Power Thru Ionized Light Beams and the Upper Ionized Strata of the Atmosphere.

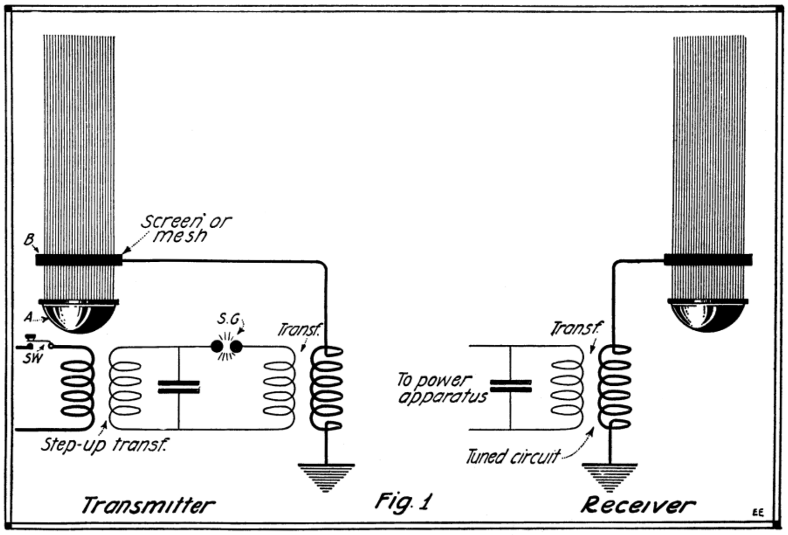

Fig. 1

Diagram Showing the Transmitting and Receiving Circuits for the Transmission and Reception of Electric Power by Wireless.

as the distance from the source increases, therefore the effective height of such an arrangement is determined by the intensity of the ray generator. So much for the idea as applied to radio communication.

But let us consider further. In studying the spectrum we find that the sunlight on reaching the earth contains few ultra violet rays of shorter wavelengths than 3,000 Angstrom units, yet we are able to generate much shorter rays with an arc light. It is unreasonable to assume that shorter waves do not leave the sun, and in view of the fact that it has been determined that the upper stratum of our atmosphere is ionized, it is accepted that the shorter rays are absorbed to cause this ionization.

It will now be apparent that a very good condition exists for the transmission of energy without wires. As shown in Fig. 2, we have the earth, a good conductor, surrounded with a blanket of insulating air about ten miles thick, beyond which, up to about a hundred miles, the air is highly rarefied and kept in a state of ionization by light from the sun and stars.

Hence, we have two good conductors separated by several miles of atmosphere, practically a perfect insulator, an ideal arrangement for our purposes. We have but to connect a source of current to the earth and to the ionized strata of air and energy can be transmitted entirely around the earth without wires. The energy could be utilized in any part of the earth by merely connecting to the pair of conductors.

This condition has been recognized for some time, but the difficulty lies in making connections to the upper layer of conducting atmosphere. It is hardly practical to construct towers six or seven miles high for the purpose, then again the energy could not be utilized without using a similar structure at the point of reception, which makes this scheme hardly feasible for aircraft or moving vehicles.

Return then to the ionized stream for radio aerials. We can easily construct arcs that will throw a beam ten or twenty miles. Then why not make connection to that upper stratum with an ionized stream of air? Wonderful in its simplicity.

In Fig. 2 the writer attempts to portray a generating station embodying these principles. In the main building will be housed the generating units for supplying the current to the arc and the primary of the high voltage apparatus for transmission of energy. Supported on the roof of this structure will be a monstrous arc lamp capable of throwing a stream of ultra violet rays for at least ten miles. In this beam is supported a conducting screen to make connection to the conducting stream of ionized air. The insulation of these parts present no great difficulties to modern engineering.

The building will also house a huge step-up transformer, a set of high voltage condensers, a spark gap either of the rotary or quenched type and a huge oscillation transformer or Tesla coil. These instruments are connected in the usual manner and tuned so that the inductance in the circuit and the capacity formed by the earth and ionized upper strata will have a well defined oscillation period.

By these means the whole earth will become an electrified body, energy to be drawn at any point by simply making metallic connection to the earth and pointing a beam of light vertically, a screen being used to take the energy that will flow down the beam. An inductance coil being necessary in the circuit to insure resonance with the transmitter at the receiving set. After once started, the receiving energy can be used to keep the arc going.

Ships will have an arc and reflector mounted on the masts, aeroplanes have two arcs, one on top, the other below, and so on.

This scheme will not interfere with present short range transmission, it being hardly practical for each home to have an arc on the roof when one arc would serve for a town and several for a large city.

Rather its benefit will lie in the utilization of water falls far from civilization, now unused on account of difficulties in transmitting the power hundreds or thousands of miles.

Aside from the transmission of power other advantages may accrue. Will such a stress between the earth and upper air strata reduce the presence of dust particles in the atmosphere? Or, again, cause the immediate condensation of fogs and moisture in the air to give us perfectly cloudless days? Perhaps it will become simply a matter of changing the frequency or voltage of the current to cause rain or to prevent it.

Then the question of effect on radio communication: will it make our present sets obsolete, a transmitter of the future being connected directly to the power set and acting by superimposing different frequencies on the power frequency? Or will radio-telephony work hand in hand with the power system, the voice currents being impress on the power circuit and filtered out at the receiver?

There are a thousand and one things such an arrangement might effect. Will trees and vegetation increase in growth as experiments in electrical culture would indicate? What effect would such stresses have on germs and bacilli? Their numbers will be decreased by the destructive effects of the ultra-violet light. Will man grow to an unsuspected height and become uniformly healthier due to the electrical treatment he would be constantly undergoing?

And so on—conjecture is without bounds, but the scheme is practical to all appearances; it requires but some financial genius to give it a trial.

And to think it all was within reach such a short time ago! Had Tesla but put a gigantic arc on top of the tower of his experimental station in Long Island accurate data would be at hand now—were it but even a promise it would mean another step towards the final mastery of all matter by man.

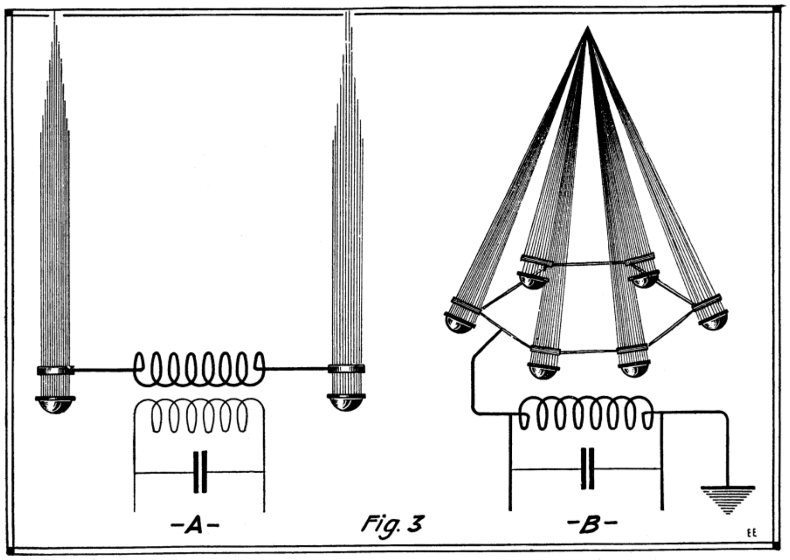

Fig. 3

Two Optional Forms of Wireless Antennae Formed of Searchlight Beams—Ionized Atmospheric Streams.

Creative SCIENCE & Research

Information Big Brother & the Oil Companies do not want you to know!

459

Copyright 2002 - 2004 Creative Science & Research

PO BOX 557 New Albany, IN. 47151 USA

Free Energy *From The* EARTH

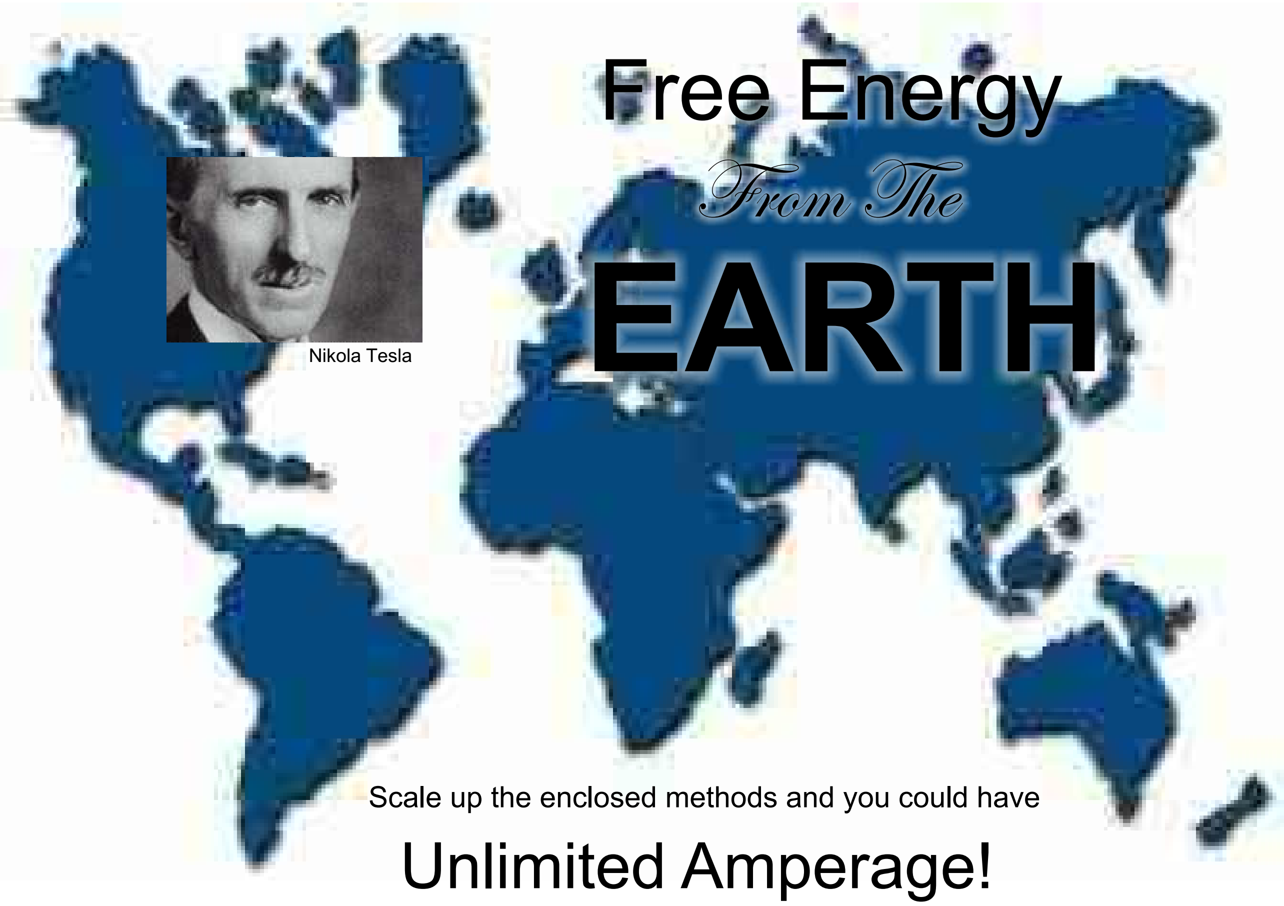

Nikola Tesla

Scale up the enclosed methods and you could have

Unlimited Amperage!

By Unlimited we mean you can keep adding more earth cells to your earth power station
as you can afford it year by year adding more and more amp power.

Aluminum

Copper

There are many ways to pick up on the free energy in the earth, Just a few are described in these plans. These devices if made right will pick up earth energy that many call ether or static currents as well as acting as high efficient earth battery's. Ether travels through plastic wood etc... Study and research these plans closely there is more to these plans than what you may think. You can learn to get enough power to run your home!

Copyright 2002 - 2004 Creative Science & Research

Learning The basic’s of our invention

Experiment #1 How to assemble a 12 vdc system

1. You will have 12- 4" 3/4" copper piping, or If you did not purchase the kit then cut your copper to this size. KIT: take each piece of copper and tape the tops with 1" masking tape.

2. Now using RED LACQUER spray paint (Which you can buy at any hardware store), Paint the outside of the copper all around. (DO NOT PAINT THE INSIDE.) let air dry for 24 hrs. Now remove 1" masking tape.

3. Take the copper pipe outside to the area that you are going to use. use a hammer and hammer each one Into the ground up to the 1" copper exposed top. do not let the exposed copper touch the earth in the final placement. Dirt should be moist. Not real dry.

4. Now remove each pipe and place a thick piece of plastic on bottom of copper pipe to cap it. LET IT BE LOSE, DO NOT TAPE IT. YOU WANT RAIN WATER TO BE ABLE TO ESCAPE FROM EACH PIPE, Do not let exposed outside copper touch the earth.

5. Now hammer each pipe back Into the same holes. place each 2" piece of zinc rod (or a long bolt, most bolts are zinc plated) into the center of each pipe. The zinc Is the NEGATIVE and the copper is the + POSITIVE.

6. Now connect them in series to get 12 volts, use clip wire or solder them. If you are going to leave It outside for a long time then make sure connections are soldered and then painted to protect them from the weather. The rain will replenish your earth batteries. Top get more amaperage with this type simply add more copper and zinc rod cells USE 3' PIECES OF COPPER PIPE INSTEAD OF 4" PIPE. THEN CONNECT EVERY ROW OF 12 PIPES IN PARALLEL.. ADD AS MANY ROWS AS YOU NEED UNTIL YOU GET THE DESIRED AMPERAGE YOU WANT. **(THINK BIG!)** This can be a very powerful free energy system for your home or other. Not only will you get an earth battery effect, but you will also be collecting earth currents from ether /static and under ground radio waves as well.

7. The deeper you go with the copper pipe, the more amperage and voltage you can get.

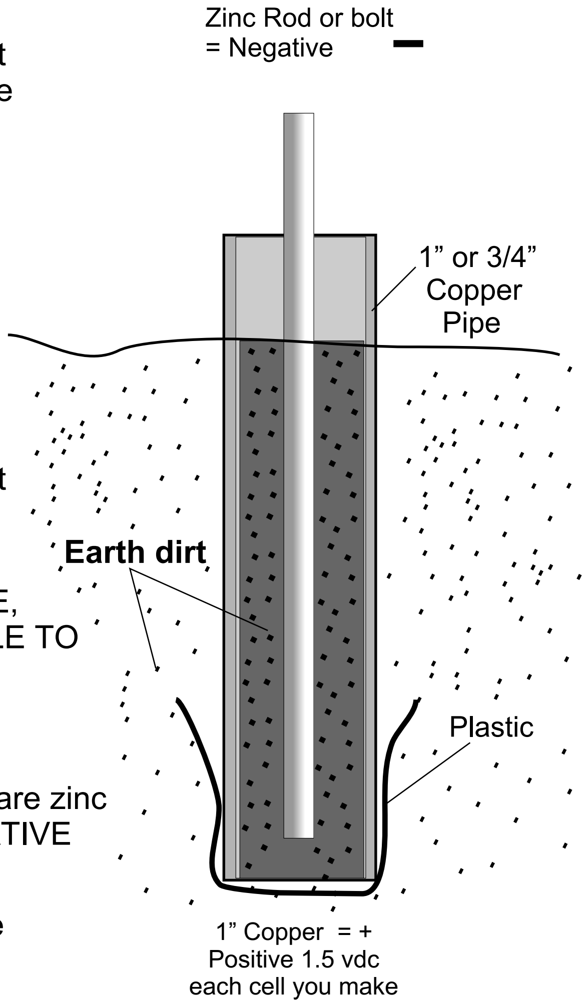

This represents one cell. A qty of 12 to 14 will give you 12 - 14 volts dc when connected in series.

Cut 12 - 4” L x 3/4” copper pipe

Copyright 2002 - 2004 Creative Science & Research

Learning The basic's of our invention

Experiment #2 How to assemble a 12 vdc system

This is just an example to go by and is not the best way but is low in efficiency! See our capacitor earth cells, coil type or flat sheet type. They produce far more amperage and voltages than this type.

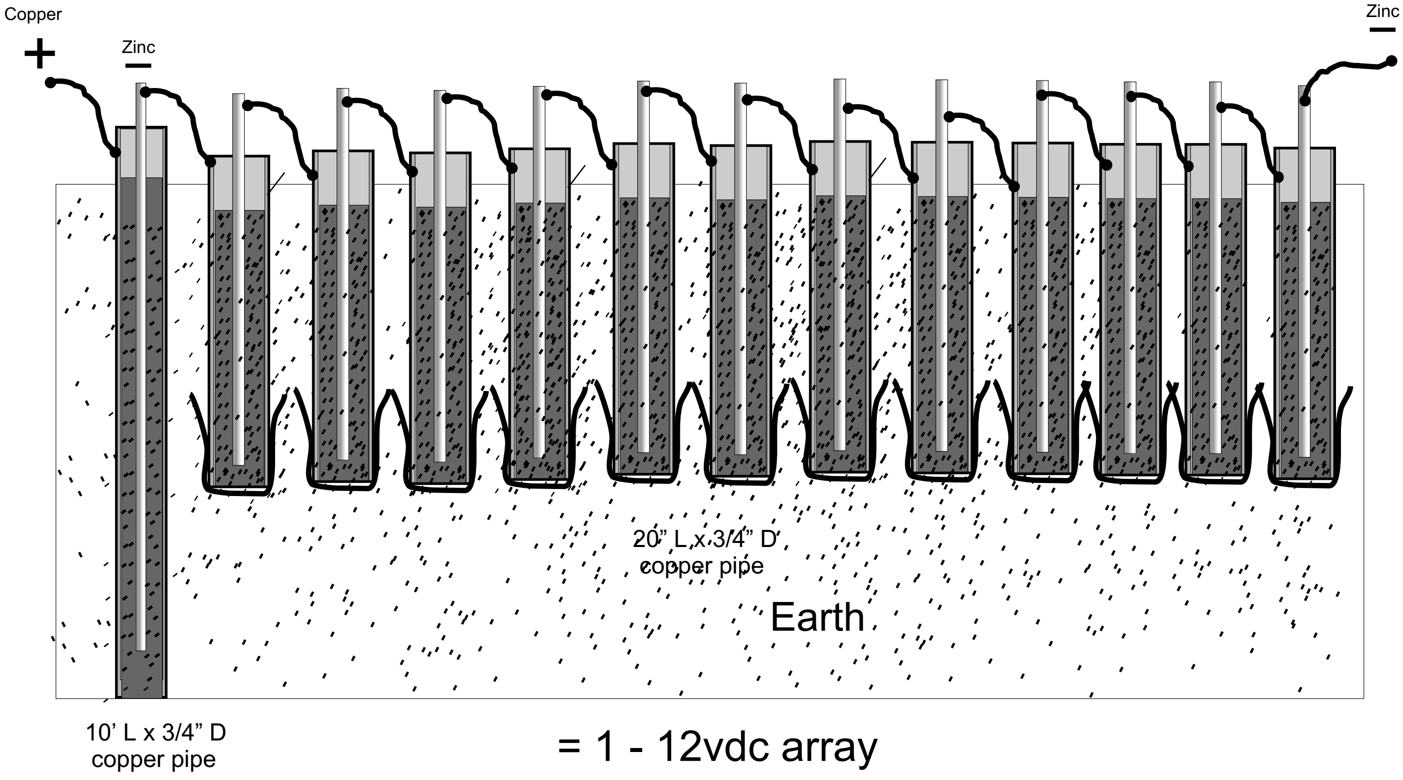

= 1 - 12vdc array

NOTICE: Do not paint the 10 foot copper cell. Keep all copper exposed to the earth. The more exposed copper the better. For higher amperage output, use a zinc or aluminum rod 10/16" Diameter which will leave a space of more than 1/16" to the copper.

Advantages of earth powered batteries

1. Free Energy
2. Long Life
3. The collection of aether energy.
4. Cells are also replenished by outside weather. Rain as well as lightning charges.

Copyright 2002 - 2004 Creative Science & Research

Learning The basic's of our invention

You can purchase Zinc rods or zinc plated bolts at any hardware store, most all nuts and bolts they sell are zinc coated and will work just fine. They also carry aluminum sheeting. Most hardware stores can order you copper sheeting as well.

Experiment #3 How to assemble a 12 vdc system

This is a simple way to produce more amperage but is not practical, we are only showing you this to educate you. The closer the zinc rod is to the copper the more amperage you will get as well as going deeper into the earth. If you do decide to build this basic set up. It is suggested that you solder all leads very well and dig deep enough into the earth to hide the tops of each cell. By covering the tops with dirt the cells can short out so what you want to do is to paint the exposed tops with lacquer or enamel paint then cover with plastic, cover with dirt and allow both the negative and the positive leads to stick out of the ground. But like I said this is not practical, what your goal should be is to make a very strong uf earth capacitor / battery, This will allow you to pick up and collect earth radiant waves of energy as well as lighting energy during storms, the energy you can collect in a strong lighting storm will amaze you! You must use an antenna on the zinc rod or copper rod. **WARNING! Be careful when collecting lighting it can kill you, be sure the positive and negative wires are not coming into your home. During a storm it is advised to have the lead wires going into a large uf cap bank via a one way diode on the positive lead to the positive of the cap or battery. The earth battery's can be dangerous at times and must be completely buried in the earth so a child, adult or animal can not touch or get to. Our earth capacitor batteries can store and hold thousands of volts. So be careful. Use rubber gloves and clothing to protect yourself. We are not responsible for anything in these plans you build at your own risk.**

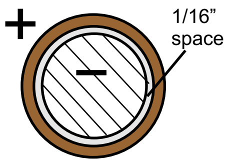

Example of a perfect Zinc rod copper pipe cell

Top view of a 12 volt dc array high amp system

Connect the series arrays in parallel to each other as shown.
Or for a high voltage system connect all in series! Pulse the dc into a 10x step up transformer, 15,000 v Neon transf.

Zinc rod can be replaced by using rolled up aluminum sheeting or Zinc sheeting, Use 100% cotton cloth wrapped around the rolled sheeting or rod to give a small space in between the aluminum sheet and copper.

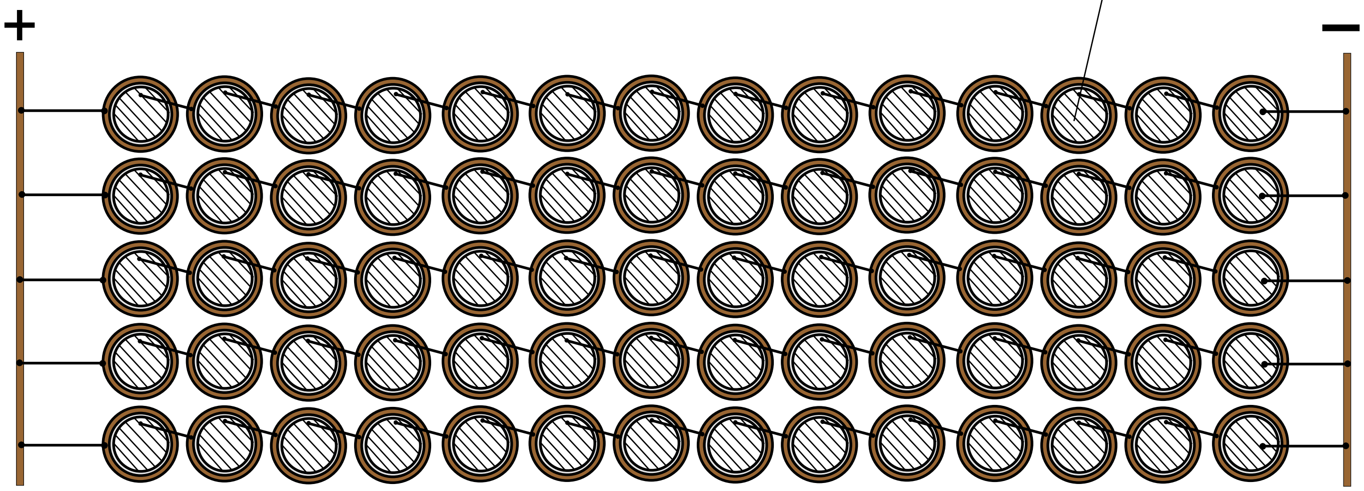

Copyright 2002 - 2004 Creative Science & Research

Learning The basic’s of our invention

Experiment #4 The 6 foot spacing method (old way)

There are many US Patents that were Issued back in the 1800's, one of them was issued to a Mr. Deckman. Deckman found that if you take several small pieces of Zinc and Carbon rod and stick them into the ground next to each other and connect them in series (much like a battery) you get very liittle gain at all.

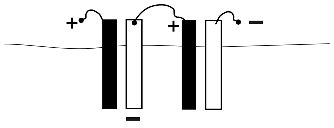

But if you put them 6 feet apart you will get a gain In voltage and they will not cancel one another out. so by doing this you can put them In series to increase your voltage and your volt amps. The theory is that there Is some sort of natural energy VORTEX that takes up about that much space for each unit or cell.

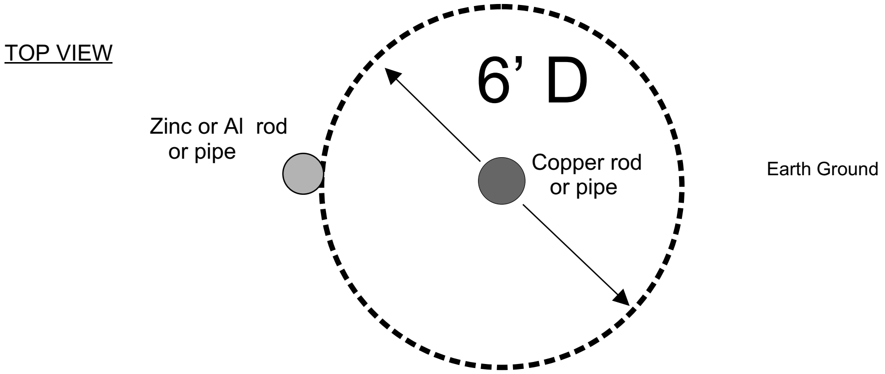

By using this method it will take a lot of ground space, ground space that many people just do not have with the exception of farmers. If you do have the space then think BIG! dig deep and have as many cells as you can get up to 2 to 4 acres. There are much better ways - vs - the old method as you will see. but by using Deckmans old way or our new methods, you can get as much free energy as you want with voltages or amperages as high as you want or desire. The higher the desired amperage the more costly it can get. We are always working to better our inventions to keep the cost down.

Copyright 2002 - 2004 Creative Science & Research

Learning The basic's of our invention

By using our new system you can get as much voltage or amperage as you want. Just size It to meet your needs or your budget.

How to get AC current for house hold use.

You can convert the incoming dc current to ac by using a home made inverter (which we sell the plans to, **5 kw inverter plans $40 order #579**) or you can purchase a 12 v to 115 vac inverter from a store or solar panel supply house.. If you are planing to make a 115 VAC system this Is the only way you can do it. You can not get AC current from the ground. You must use our system to charge a deep cycle 12 v marine battery, and then connect our 5,000 watt 120 vdc to ac inverter to the battery to run your home. (we do not sell kits or inverters assembled, we only sell the plans at this time.)

More on the basic pipe method

Please note: the dirt level inside each pipe should be up to 1/4" from the top. each pipe should be tightly fliied with dirt. Which will automatically happen when you hammer the pipe into the ground the first time.

Each copper pipe (or cell) should be spaced 1 Inch to 1 1/2 inch apart. Doing this will give you more amperage and voltage.

How to dig a simple deep hole for 8' pipe: There are many ways in which you can do this. you can rent a ground drill bit and rig it up with a electric 2 hp motor or you can use 1" copper pipe.

First: prepare the ground with water to make it moist but not muddy, 2nd: hammer a 1" x 5' copper pipe into the moist ground about 5 inches deep/ then remove it from the ground and take a stick or a water hose to remove the dirt from inside the pipe. Then just repeat the process over and over until you get the desired depth that you want. you can get about 10 feet or more using this method.

WARNINGI BE CAREFUL OF ELECTIC POWER LINES WHEN YOU PLACE OR PULL THE COPPER PIPE FROM THE GROUND, IT WILL KILL YOU IF YOU TOUCH A POWER LINE WITH THE END OF THE COPPER PIPE.

We are dealing with a new type of energy as well as conventional energy < Ether energy can go through rubber/ plastic. Iron, etc....

Copyright 2002 - 2004 Creative Science & Research

What is ether or Radiant energy?

The understanding comes from a very deep study of etheric or static forces, this energy that we are getting out of the ground is what some call LIGHT ETHER, when you separate the light ether from the warm ether you can get these superconductive events. It appears like electricity but it has no heat events related to It.

The point is that the best polarity's of things to get a differential in the light ether is to have a silvered coloured metal and then something that is totally absorbitive/ Like black Carbon.

So now you have something that's the perfect reflector and something that is the perfect absorber, and once you understand what your pulling out of the ground (and that it Is not classic electricity as we get out of a normal type of chemical battery), then your going to be farther down the road to understanding how to get much more power out of this stuff.....

There are researchers now that are getting 1 /2 volt to 2 volts dc on a "Quote" volt meter, but it will shock you through an inch of rubber shoes this is not normal electricity!

Q: Is this energy coming from the sun and being absorbed In the earth or from the center of the earth?

A: Yes, but the primary source of light ether is the sun light, yes the earth does have an economy of light ether which does follow the sun around. BUT IT IS ALSO IN THE GROUND! AND ITS IN THE AIR!

Q: Did NokolaTesla know about this in the 1900's?

A: Yes, if you read into his writings he was always saying things like, there is more energy moving than just the electricity in this stuff.

Example: If you take a large Tesla coil and fire it up, and you drive it with a van de graph generator and you put the output of that into a Tesla coil. you can throw sparks across the room! What these electrostatic machines are gathering is light ether.

Copyright 2002 - 2004 Creative Science & Research

A Very Strange and New Discovery!

While working outside on another experiment I discovered a very strange generator effect, much to my surprise, I found that by using 2 aluminum 8” disks connected to my DC volt meter and by placing both disks in a vertical position while touching the bare earth dirt, and moving one of the disks and allowing the other to be stationary I get a dc voltage positive output! This effect works with any size or shape of aluminum, we used 8” diameter disk x 1/8” thick, we found that size to work best. Also by stacking them like a capacitor you will find that the current output is far greater, For Example: see figure 1, Disk A would have 4 more disks connected in parallel of the same size separated by paper die electrics. The same goes for Disk B .

First find a spot any where on the ground where there is no grass (best when wet). Please note, it's not just creating an acid battery effect, you are also getting a good contact with the ground as a conductor to pick up ether energy.

Now place the disks side by side, attach your DC voltmeter probes onto the disks, one negative and one positive. DC meter should be on the lowest setting, now place both disks on the ground very softly and move **disk B** and **do not move disk A** (positive.) You will notice you are getting a small amount of voltage and amperage. Now Move **disk A** and do not move **disk B.** **THE POLARITY CHANGES**.

The rule here is; The disk that moves is always positive. since we can get a polarity change just by moving the disks at opposite times. You can make a very simple mechanical device to get AC current from on top of the earth. Again, this is not just a chemical reaction we have done many test to prove this. You can also get a small voltage reading in dry dirt. The reason for the smaller voltage in using dry dirt is because a low conductance resistance is taking place here, keeping the system from being in contact with the earth and collecting the earth **ether energy.**

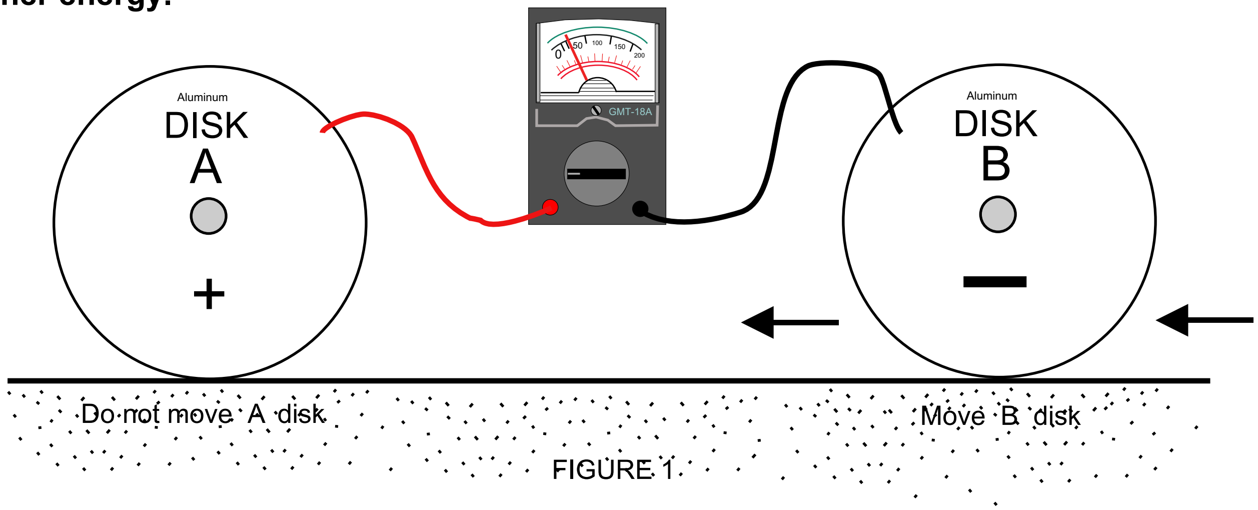

FIGURE 1

Copyright 2002 - 2004 Creative Science & Research

Capacitor Sheet Method

This method is far better than using pipes or rods. By using copper and zinc or aluminum foil sheets you will get much more amperage out of your system!

The amperage you will be collecting will be coming from 3 different sources,

1. **The acid in the ground and water**
2. **Energy that is being transmitted from the earth itself**
3. **Energy that is being transmitted from the sky and space.**

This may all seem unbelievable but it's true and is a fact! The more plates that you add the more power you will get! The copper sheet(s) is the positive, it must be facing down toward the earth. The aluminum or zinc sheet(s) is the negative and must be facing toward the sky! Between each sheet place 1- cotton or course screen printing mesh or other plastic course mesh cloth material. you can use a hole puncher to punch exact holes in each sheet, use 2 holes per sheet copper and aluminum. Use a 8 1/2" x 11" paper puncher one that is adjustable. next use a 3/4 wood base and drill hole for wood dowel rods to fit to the holes you punched in the sheets, you will have a total of 4 wood dowel rods, 2 on the right for the copper sheets and 2 on the left for the aluminum sheets to fit over, cloth die electric can be done the same. You then need to assemble dry and use another 3/4" wood board for top holding plate, which will fit over the wood dowel rods on top, sandwhich and press the assembly together and tape in place with duct tape, drill 2 long bolt holding holes on bottom wood base and top wood base then attach long bolts and nuts and tighten down. Remove tape. and attach wires, water the entire capacitor assembly and bury in the ground.

Copper sheet

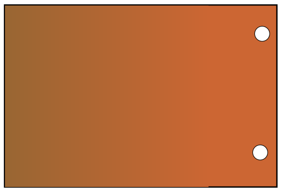

Top View

Aluminum sheet

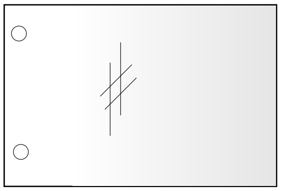

Top View

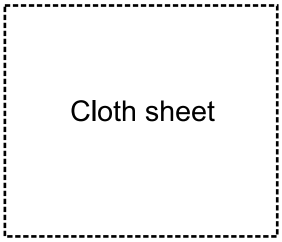

Top View

Place Aluminum sheet on top of copper sheet

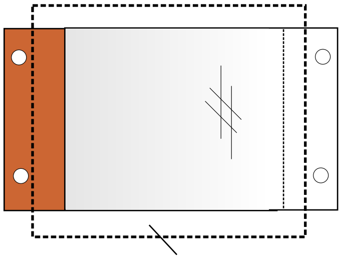

Dotted lines indicate where cloth goes

Top View

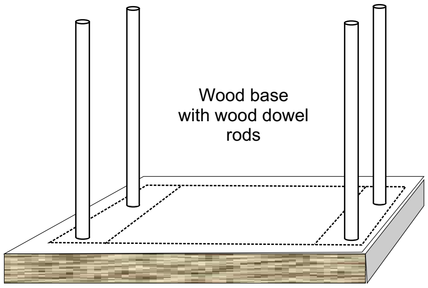

Side View

Copyright 2002 - 2004 Creative Science & Research

Capacitor Sheet Method

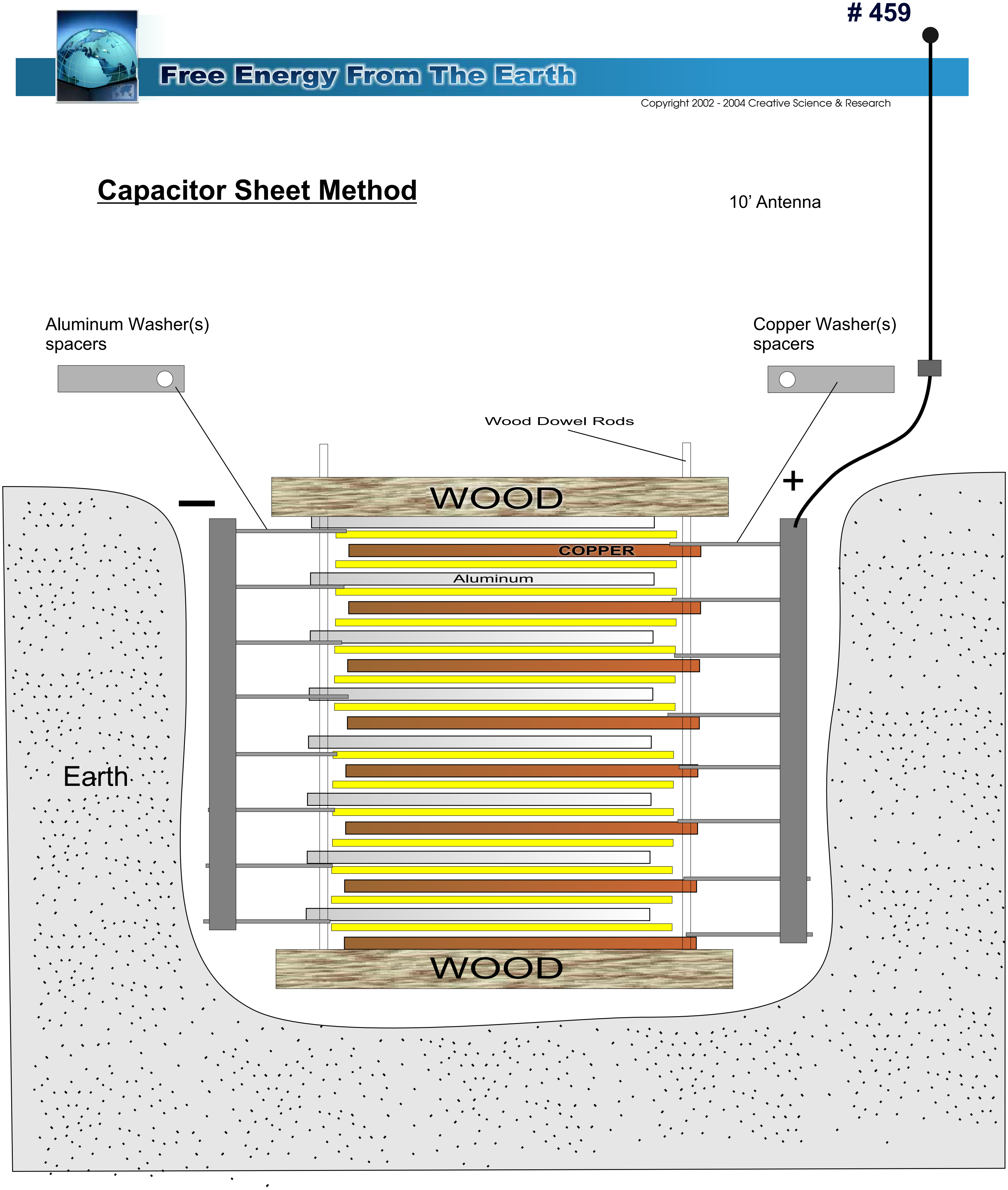
10' Antenna
Aluminum Washer(s)
spacers
Copper Washer(s)
spacers
Wood Dowel Rods
WOOD
−
+
COPPER
Aluminum
Earth
WOOD

Copyright 2002 - 2004 Creative Science & Research

Capacitor Sheet Method

Again, the more plate sheets of metal you add (or cells) the more power you will get! Remember it's all about capacitance. You are collecting more than just simple battery type power. The copper plate sheet is the Positive + and the Aluminum sheet is the negative. There are many ways in which one can build this type of earth energy collector, below is simply an example of another shape and form. In this type we use water holes. water holes must all be in exact spots.

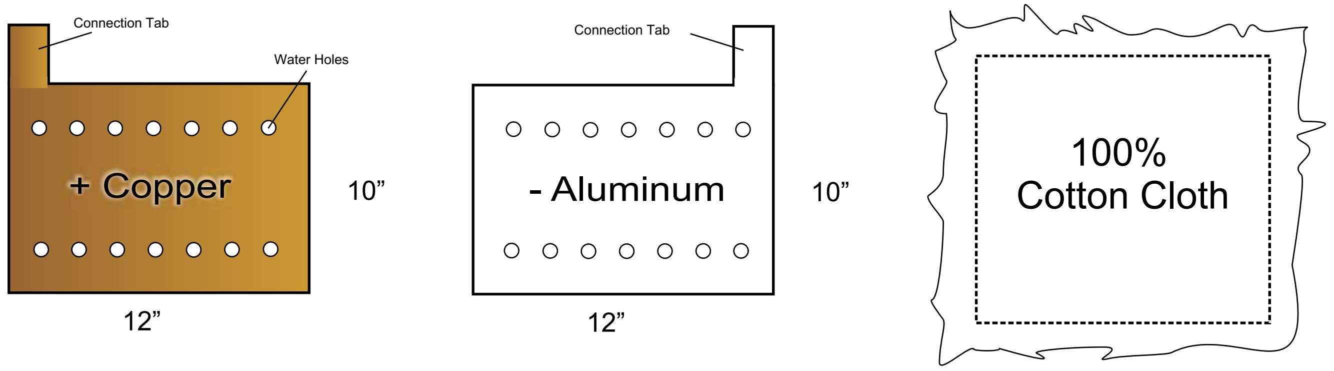
Connection Tab
Water Holes
+ Copper
10"
12"
Connection Tab
- Aluminum
10"
12"
100%
Cotton Cloth

Figure #1 Top View of separate pieces

Figure #1 Side View

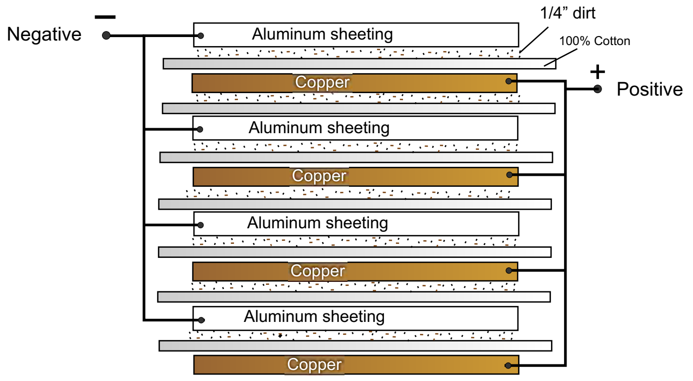
Negative
−
Aluminum sheeting
1/4" dirt
100% Cotton
Copper
+
Positive
Aluminum sheeting
Copper
Aluminum sheeting
Copper
Aluminum sheeting
Copper

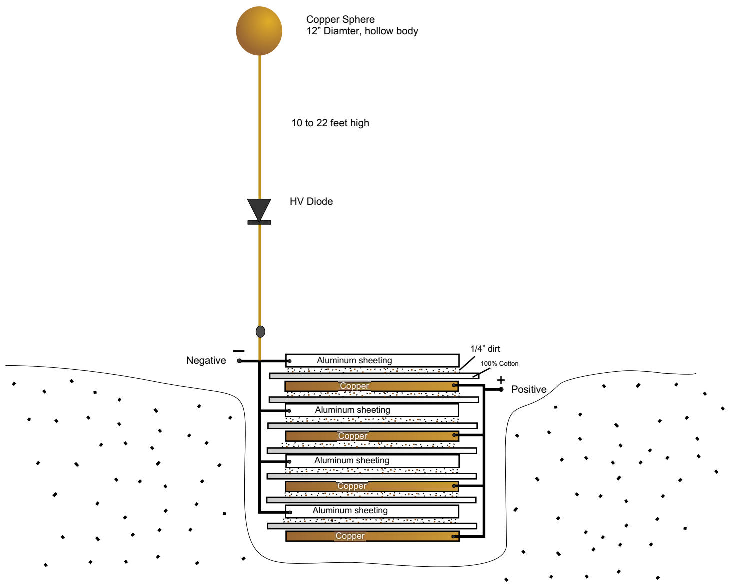
Copper Sphere
12" Diamter, hollow body
10 to 22 feet high
HV Diode
Negative
−
Aluminum sheeting
1/4" dirt
100% Cotton
Copper
+
Positive
Aluminum sheeting
Copper
Aluminum sheeting
Copper
Aluminum sheeting
Copper

Copyright 2002 - 2004 Creative Science & Research

OUR MORE ADVANCED EARTH CAPACITOR!

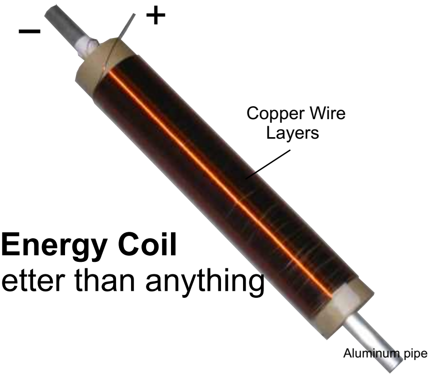

(EECC) Earth Energy Coil Capacitor!

To collect more ether earth energy atoms, build the following **Earth Energy Coil Capacitor**, **Designed & Created by Creative Science & Research**. This is far better than anything we have found yet! Much more power and is easier to build.

Start with our EECC Model #1, Take a 5/16” zinc rod or aluminum rod, cut to 7.5” long. Now use #27 bare copper wire, secure it in a drill press (which you will then turn by hand). Cover zinc rod with paper one layer, use very tiny bits of tape to secure. Now apply the copper wire over the paper, remember to use copper that is not coated! Tape one end of the copper wire to the end of the zinc or aluminum rod and begin to slowly wind. Each wind must be perfectly fitted side by side, paper should be 5.5” long, total length of winds should be 4” long centered on paper. Once you are done with the first layer, apply another piece of paper to the top layer of the first copper layer, do not use to much tape to secure paper in place! Now begin your 2nd layer, repeat this same process until you have about 10 layers, (the more layers the better!) Copper must not touch the zinc or aluminum. When done hot glue or epoxy a small end piece of the copper wire so it will not come unwound. This is one complete cell, keep in mind this is a small test version, to get more power you have to THINK BIG! Build larger cells, using larger size copper. Then place them in salt water to test, 2 cells can be connected in series, after that you must connect them by use of capacitors, diodes and electronic or hand wound spring switches. If you build 20 of these and try to connect them all in series without a cap/ diode switching method, the cells will cancel each other out. Keep in mind you want water to get to each layer as well as the zinc or aluminum. You can also use Aluminum or zinc pipe.

Now concerning the paper Die Electric, Paper will not last very long, test and find a water absorbing plastic die electric,(mesh type). We suggest that you use Plastic screen mesh that you can buy at any hardware store, if all you are interested in is collecting small voltages then use one layer in between each layer of copper, if you are interested in collecting higher voltages such as lightning charges then you must use more layers of plastic screen mesh to thicken the die electric so the higher voltages will not short out your system. Paper may not last longer than 1 year, Plastic screen mesh or polyester screen printing mesh will last for many years.

EECC Model #2 Same as #1, but different. In this design we use #27 copper COATED wire, instead of bare wire. since the copper wire is coated and not bare, each layer must be sanded, that is, the top portion of the copper coated layer must be sanded off so it will be exposed to the water and acid, a die electric should still be used in this case. This type makes for a very high efficient earth cap!

EECC Model #3 Same as #1 but reversed! You will use aluminum wire and a copper rod or pipe. Use a 3/4” copper pipe and wind the die electric paper or plastic mesh and Aluminum bare wire around the copper pipe as explained in Model #1

Copyright 2002 - 2004 Creative Science & Research

OUR MORE ADVANCED EARTH CAPACITOR!

(EECC) Earth Energy Coil Capacitor!

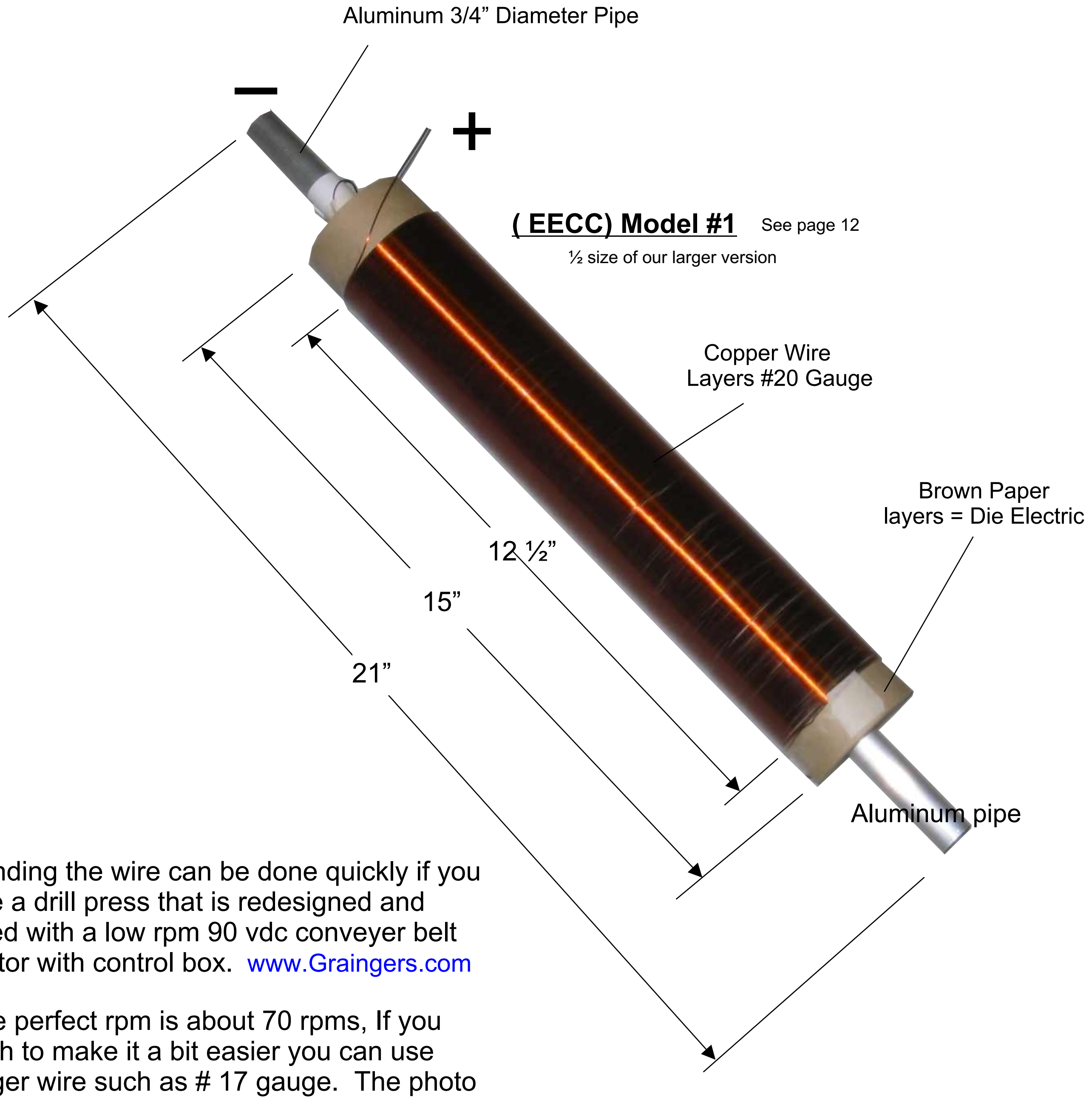
Aluminum 3/4” Diameter Pipe
–
+
(EECC) Model #1
See page 12
½ size of our larger version
Copper Wire Layers #20 Gauge
Brown Paper layers = Die Electric
12 ½”
15”
21”
Aluminum pipe

Winding the wire can be done quickly if you use a drill press that is redesigned and fitted with a low rpm 90 vdc conveyer belt motor with control box. www.Graingers.com

The perfect rpm is about 70 rpms, If you wish to make it a bit easier you can use larger wire such as # 17 gauge. The photo is an example of a mid size cell and can be very powerful! This is ½ size of a large cell.

Copyright 2002 - 2004 Creative Science & Research

OUR MORE ADVANCED EARTH CAPACITOR!

(EECC) Earth Energy Coil Capacitor!

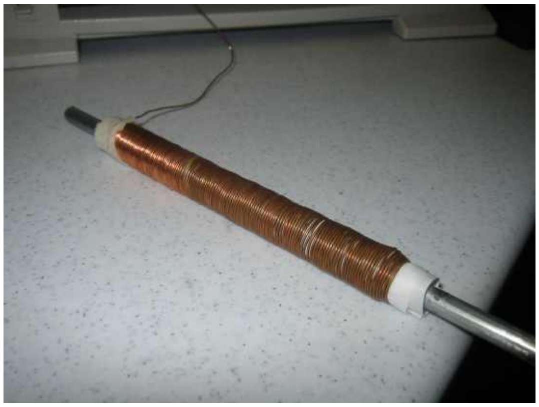

Our small test model #1, we used #22 bare copper wire wrapped around a zinc rod, each layer had a paper die electric, tested in tap water. 5 layers.
Output = 89 vdc x 3.5 ma, this is low.
more layers will = more capacitance which ='s more amperage and voltage.

Small Drill press turned into a Coil Winder

The more layers, the more free earth energy you can collect!

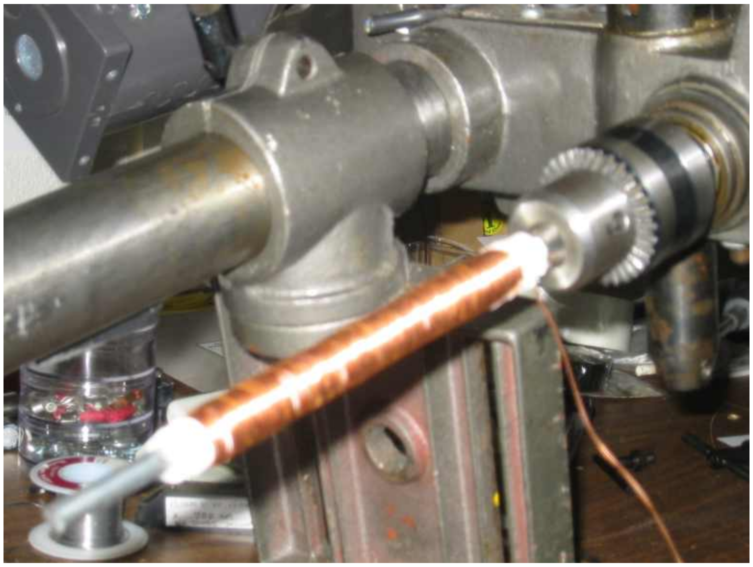

Small Drill press turned into a Coil Winder. We turned this by hand.
Larger coils is best to use low rpm motor at about 70 rpms, In this picture you see that we wound each wind side by side, this makes for a more high efficient earth cap.
You can wild wind if you like if you use smaller wire like #31 to #34.

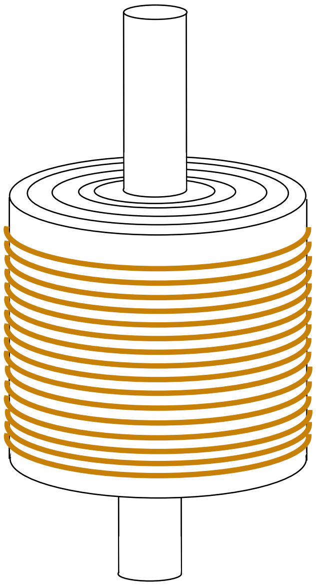

Side View
Ideal size for experimenting

Copyright 2002 - 2004 Creative Science & Research

(EECC) Earth Energy Coil Capacitor!

Rolled Earth Capacitor

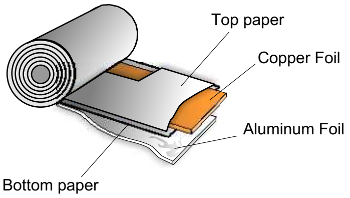
Top paper
Copper Foil
Aluminum Foil
Bottom paper

MODEL #4 Rolled Earth Capacitor

Copper foil is placed between 2 pieces of paper strips and Aluminum is placed on bottom. Use 400 foot strips 2" wide. This method is very hard to do. But is very high eff.

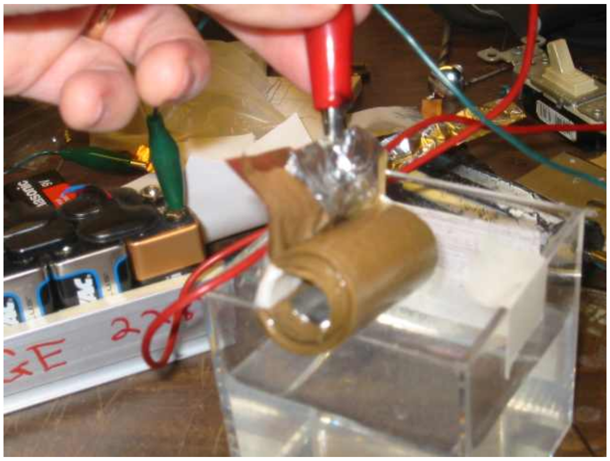

We used salt water to test rolled caps before placing them in the ground.

Copyright 2002 - 2004 Creative Science & Research

(EECC) Earth Energy Coil Capacitor!

MODEL #4 Rolled Earth Capacitor

Rolled Earth Capacitor

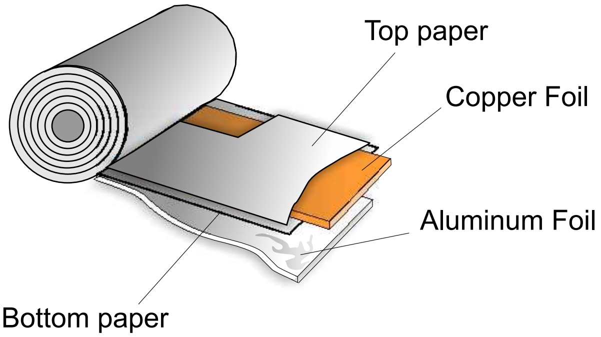

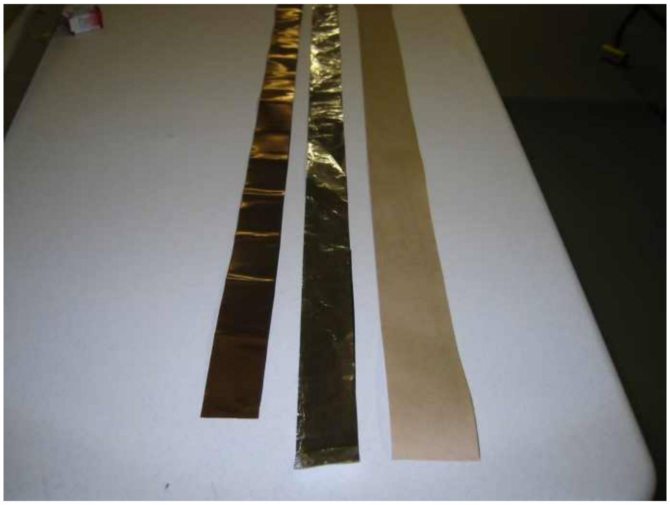

Again rolling this type of earth capacitor is very hard, Everything must be kept straight and inline with one another, Manufacturers of rolled electrolytic capacitors use computers and machinery. I am sure a simple home made roller with pin registration could be designed very easily. This type of earth battery is very high efficient. Paper is not a good die electric for high voltages! If you are wanting to collect earth and storm HV Currents you must use a thicker die electric. Metals must be spaced properly for storing high voltage.

Copyright 2002 - 2004 Creative Science & Research

(EECC) Earth Energy Coil Capacitor!

Rolling a capacitor is the best way but not the easiest,
Which makes this pipe coil model much more desirable to build

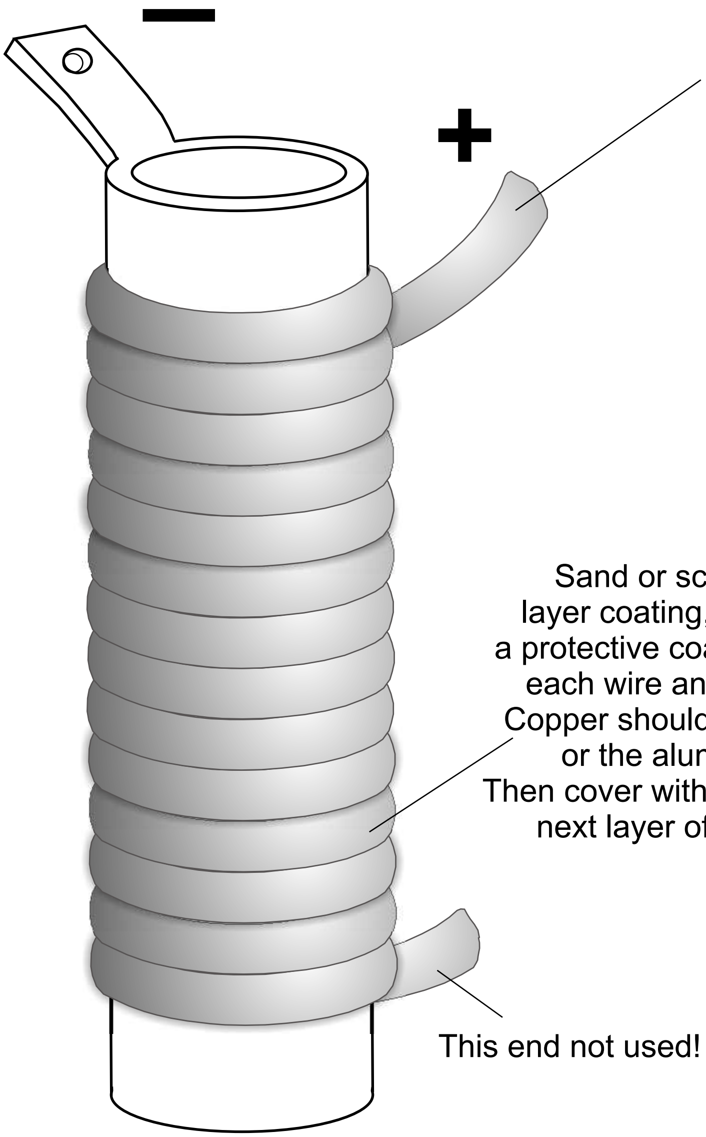

Sand or scratch off top
layer coating, this will leave
a protective coating in between
each wire and underneath.
Copper should not touch itself
or the aluminum pipe.
Then cover with paper and begin
next layer of copper wire.

Symbols

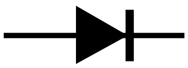

= One way DIODE

Photo of Diode

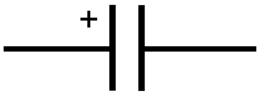

Capacitor (Electrolytic)

Photo of one type
of Capacitor

Copyright 2002 - 2004 Creative Science & Research

CONNECTING THE CELLS IN SERIES

How to connect the cells in series to increase the output voltage without allowing the cells to cancel each other out. Figure **A** The Copper pipe and zinc rod method without the use of plastic on bottom of pipe. **Figure B** is the coil earth cap method.

Figure A

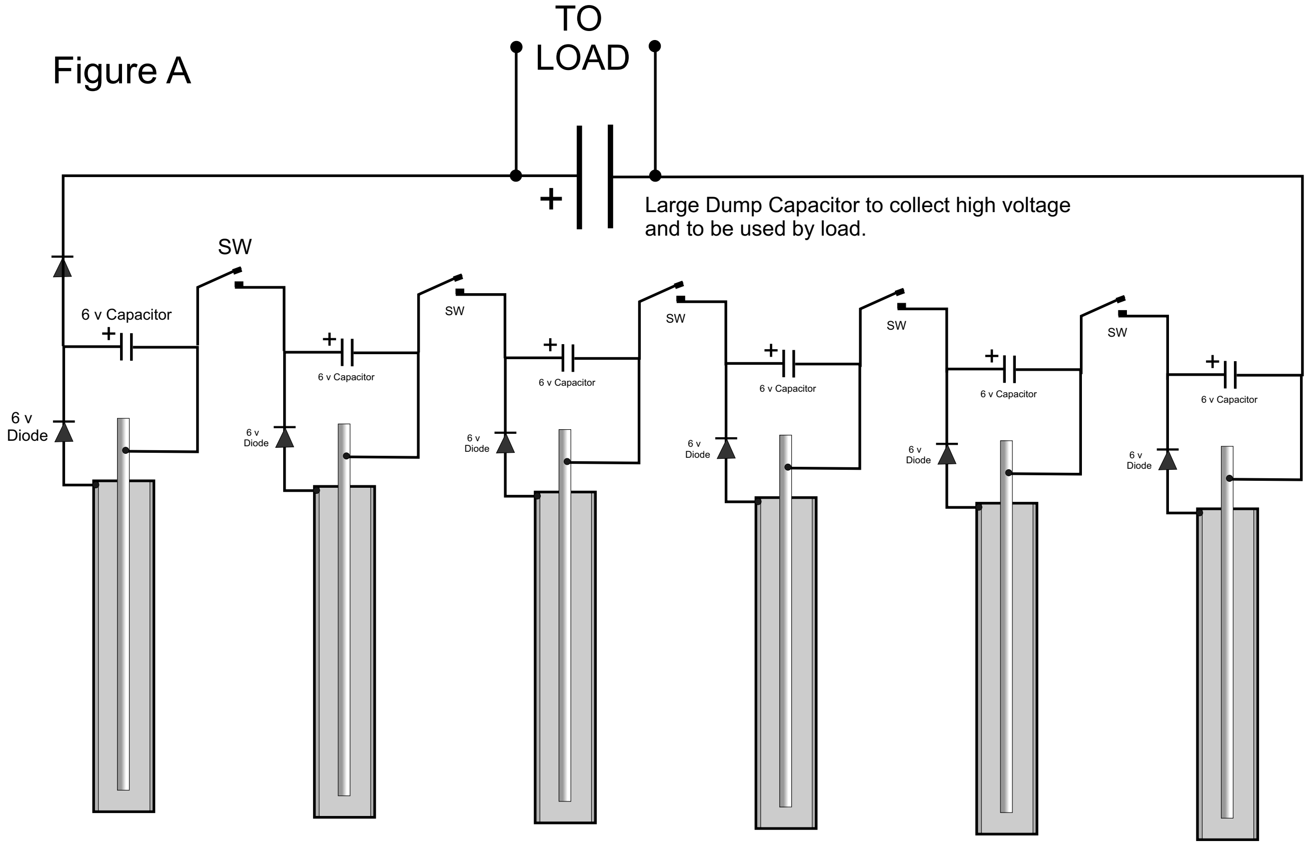
TO LOAD
+
Large Dump Capacitor to collect high voltage and to be used by load.
SW
6 v Capacitor
+
6 v Diode
SW
6 v Capacitor
6 v Diode
SW
6 v Capacitor
6 v Diode
SW
6 v Capacitor
6 v Diode
SW
6 v Capacitor
6 v Diode
SW
6 v Capacitor
6 v Diode

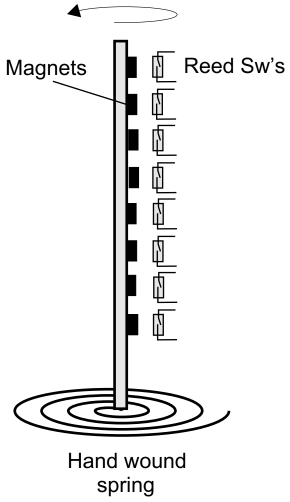
Magnets
Reed Sw's
Hand wound spring

Use electrolytic capacitors to store the charge coming from each cell, then connect them in series using Commutator Sw's connected to a small rotor shaft, connected to a milliamp solar cell hobby motor or build a small fuelless engine electric motor to run on and off switch's. You can also build a very low milliamp electronic switching circuit and use the earth energy to power it. Or another more simple way would be to use a hand spring type device (timer type) glue small magnets on a long rotating rotor arm connected to the inner spring. Then place small reed relay magnetic switches, on each timing magnet ON position. should have multi reed sw's, as the unit turns if gives time for the caps to charge and then comes around and connects caps all in series which in turn discharges into main cap dump.

Copyright 2002 - 2004 Creative Science & Research

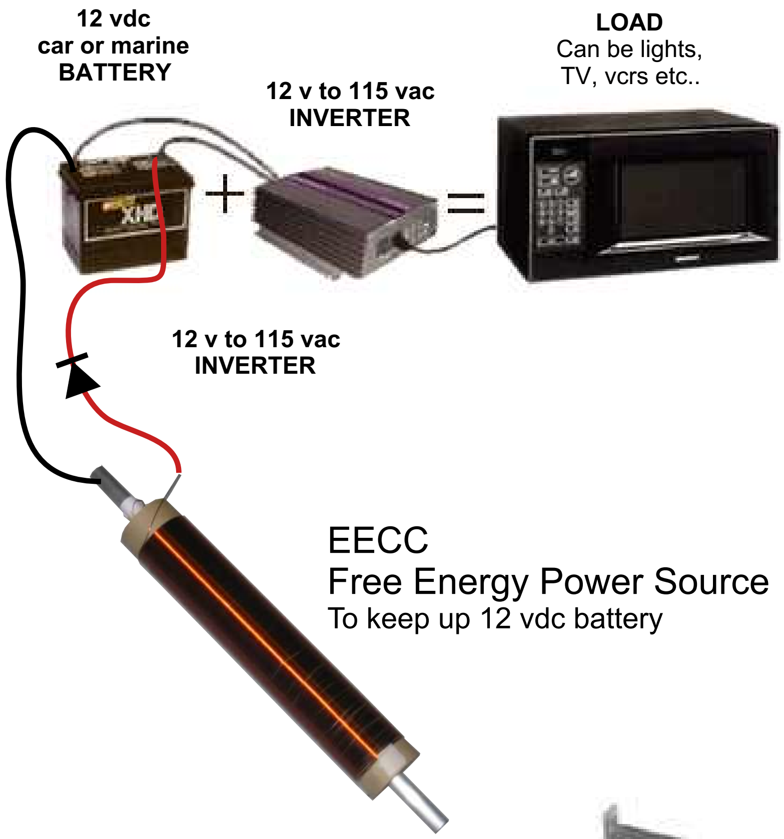
12 vdc
car or marine
BATTERY
12 v to 115 vac
INVERTER
LOAD
Can be lights,
TV, vcrs etc..
+
=
12 v to 115 vac
INVERTER
EECC
Free Energy Power Source
To keep up 12 vdc battery

Copper

Symbols

= One way DIODE

Photo of Diode

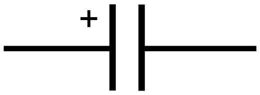

Capacitor (Electrolytic)

Photo of one type
of Capacitor

Copyright 2002 - 2004 Creative Science & Research

Suppliers List

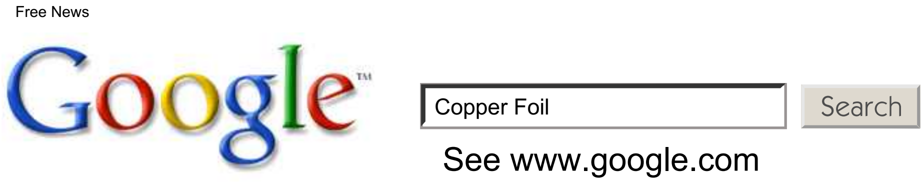

Free News

Copper Foil

Search

See www.google.com

1. Http://www.delphiglass.com/

   Use there search engine to find copper foil tape

2. http://www.glassmart.com/foil_menu.htm
3. http://www.globalsources.com

Search Google.com for:

1. Copper flashing
2. Inverters
3. Batteries
4. Copper coated wire
5. Copper wire
6. Electronic parts and supplies
7. Diodes
8. Capacitors
9. Free energy from the earth
10. Zinc rods

Copyright 2002 - 2004 Creative Science & Research

EECC update

Step One

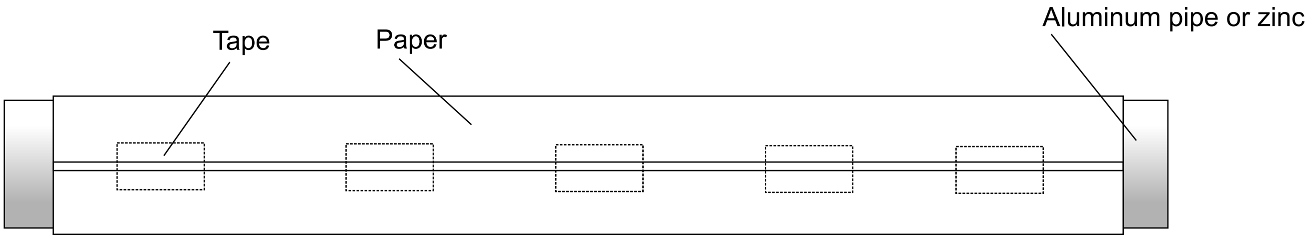
Tape
Paper
Aluminum pipe or zinc

Step two

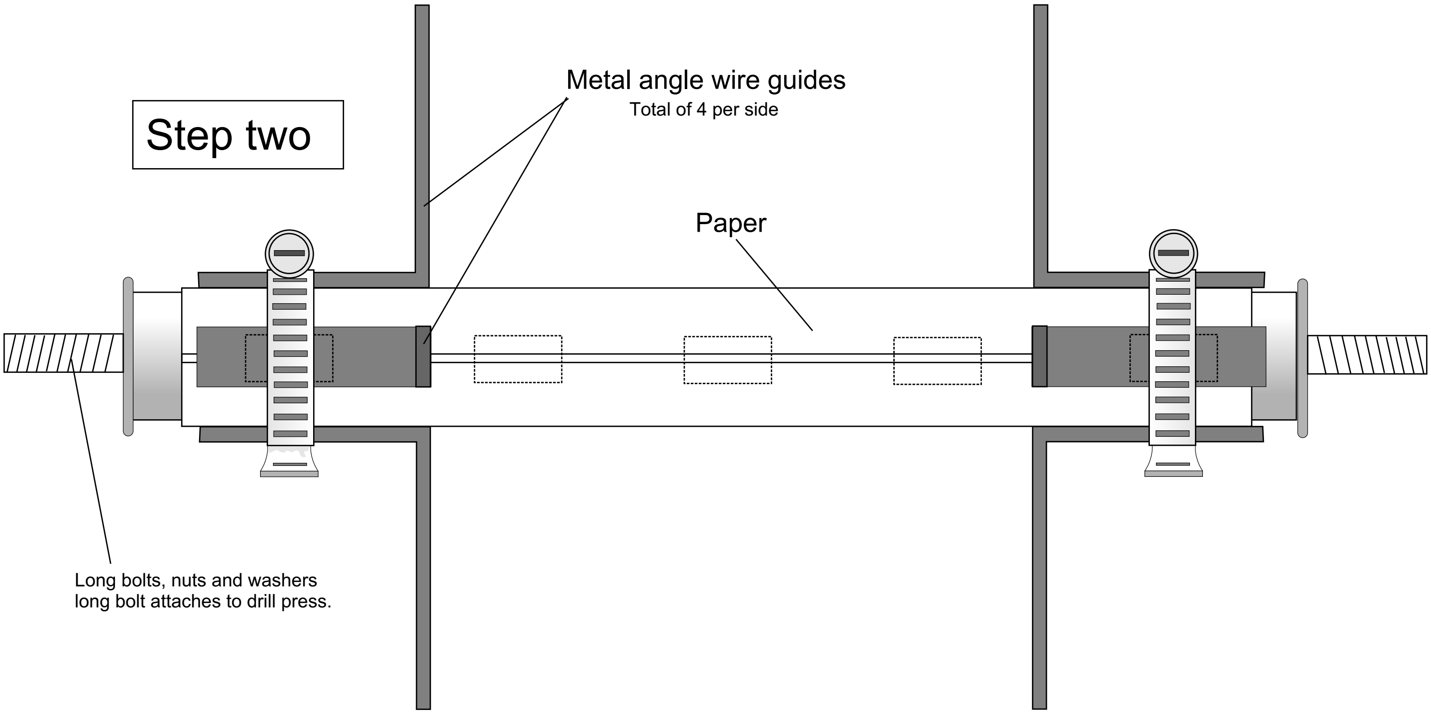
Metal angle wire guides
Total of 4 per side
Paper
Long bolts, nuts and washers
long bolt attaches to drill press.

Apply first layer of wire, remove, metal wire guides, apply more paper, then repeat over and over again until you have about a 5 to 6 inch diameter roll of wire and paper. Makes a very powerful earth capacitor!

Copyright 2002 - 2004 Creative Science & Research

EECC update

These are just a few extra idea's that we threw together. I hope they are helpful to you. Let us know how you like these plans. Send us photo's of what you have done we love to hear from our customers. If you got any new free energy idea's let us know we can research them for you. Many customers do not have the extra research time that we have.

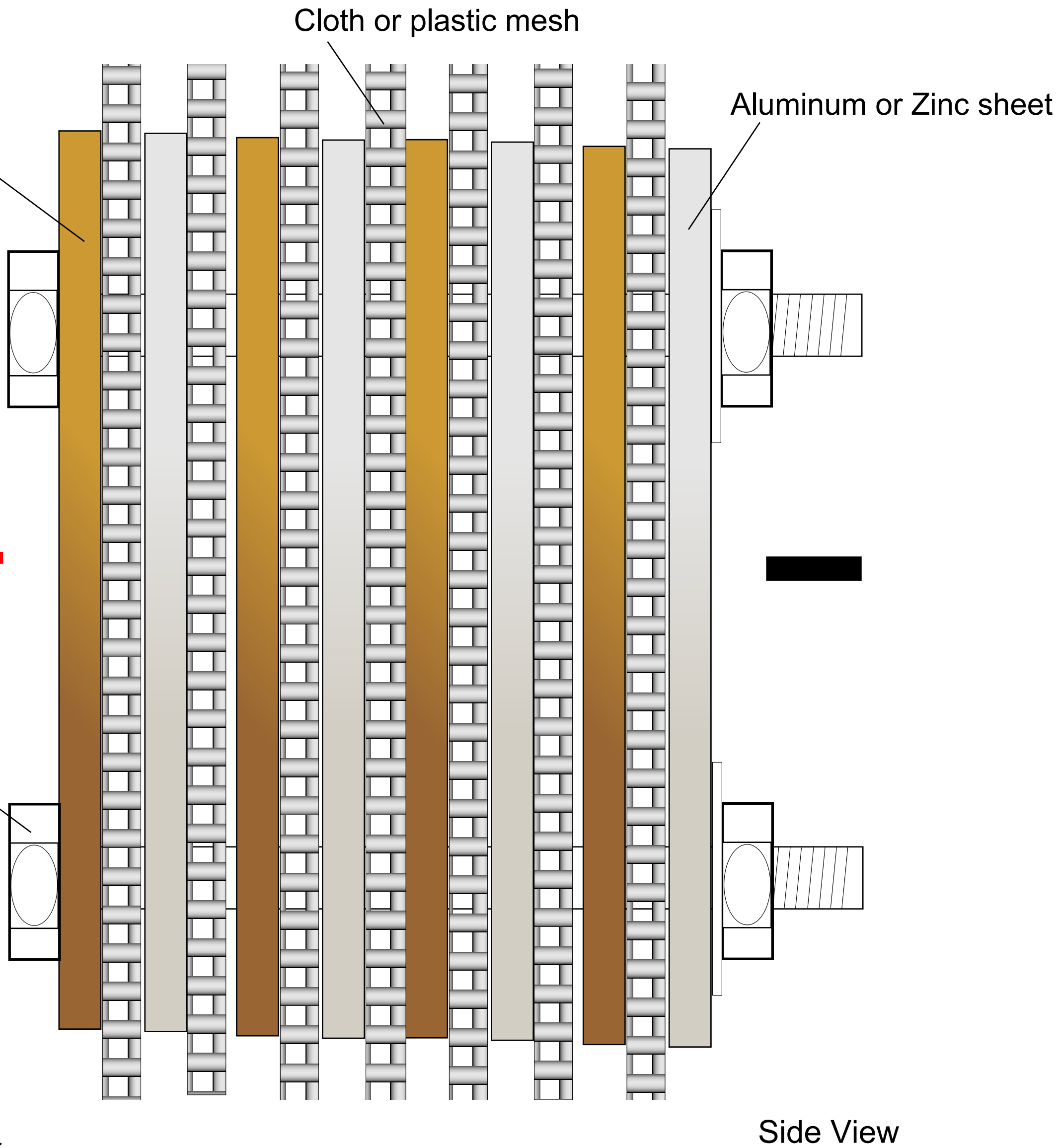
Cloth or plastic mesh
Copper
Aluminum or Zinc sheet
+
White Nylon Bolts
-
Side View

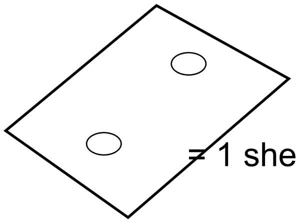

Metal sheet type earth battery

rexresearch.com

The Next Bigly Thing: Power from the Air

by Robert A. Nelson

1 | Man Cheats Credit Score
1 simple trick & my credit score jumped 217 pts. Banks hate this! thecreditsolutionprogram.com

2 | $4,400 in the first month
Looking for a job? Drive with Lyft instead & make $4,400 in the first month. Terms apply. lyft.com

1 | Man Cheats Credit Score
1 simple trick & my credit score jumped 217 pts. Banks hate this! thecreditsolutionprogram.com

2 | Tesla Generator - Build Your Own Generator
Eliminate Your Power Bill Easy. Do It Yourself. Great Discount Now! Just $49. tesla-energy.org/Generator

1 | Part-time Online Work - Up to $45 Per Survey
Make up to $45 per survey in 20 min surveyjunkie.com/Extra-Cash

2 | Tesla Generator - Build Your Own Generator
Eliminate Your Power Bill Easy. Do It Yourself. Great Discount Now! Just $49. tesla-energy.org/Generator

The Next Bigly Thing: Power from the Air

by Robert A. Nelson

The lightning bolts hurled at us by angry gods amply demonstrate the abundance of energy latent in the heavens. Lightning also introduced humans to fire.

Benjamin Franklin's legendary kite flight during a thunderstorm in 1752 marks the beginning of our Promethean attempts to draw power from atmospheric electrostatic (ES) energy. Considerable progress has been made since then, and the industrial scale production of electrical power from the sky has been an accomplished fact since the early 1900s. The promising technology fell dormant after World War One, however, and it has slept since then.

Atmospheric electricity offers several advantages over other energy sources; the technology is simple and robust, costs less than wind or solar power, and it is available anywhere anytime. Earth is an electrostatic generator -- the atmosphere is positively charged relative to the negatively polarized Earth. The system generates approximately 3 x 10 ^ -16 Amperes/sq. cm. (1500 Amperes for the entire Earth).

Andor Palenscar

Andor Palenscar's "Apparatus for Collecting Atmospheric Electricity" included a novel motor to utilize the ES energy, described in US patent US674427, granted in 1901:

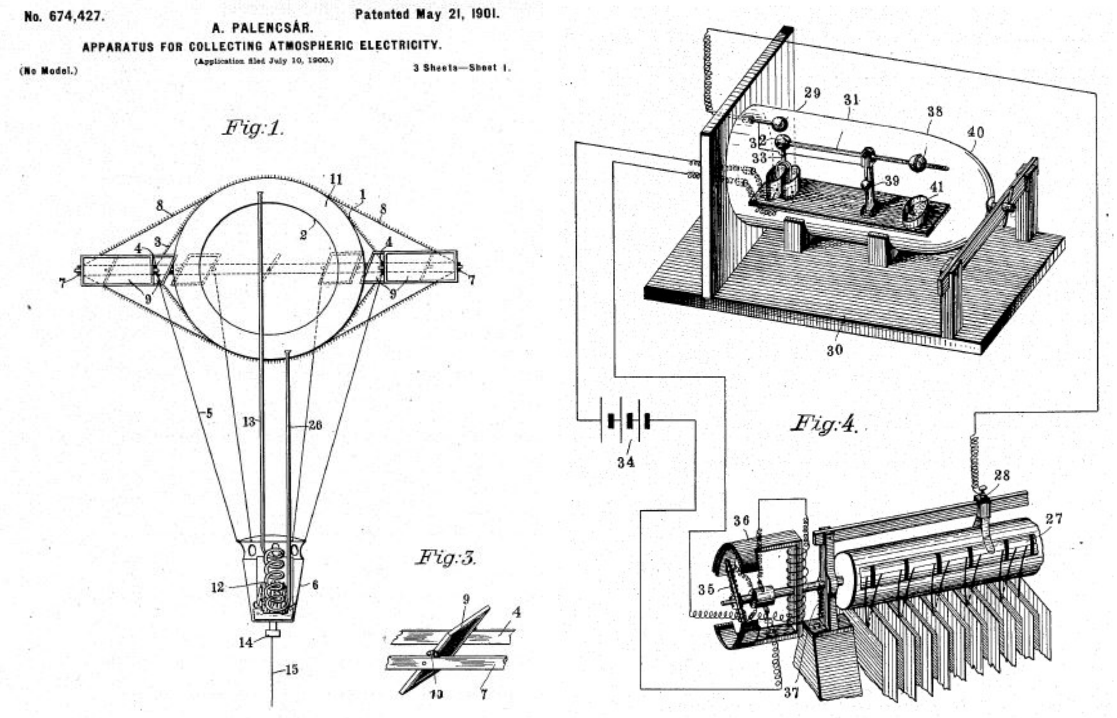
No. 674,427.
Patented May 21, 1901.
A. PALENCSÁR.
APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.
(Application filed July 10, 1900.)
(No Model.)
3 Sheets—Sheet 1.
Fig:1.
Fig:3.
Fig:4.

Walter Pennock

Walter Pennock developed an aerial energy collector, for which he was granted US patent US911260 in 1909 and US1014719 in 1912.

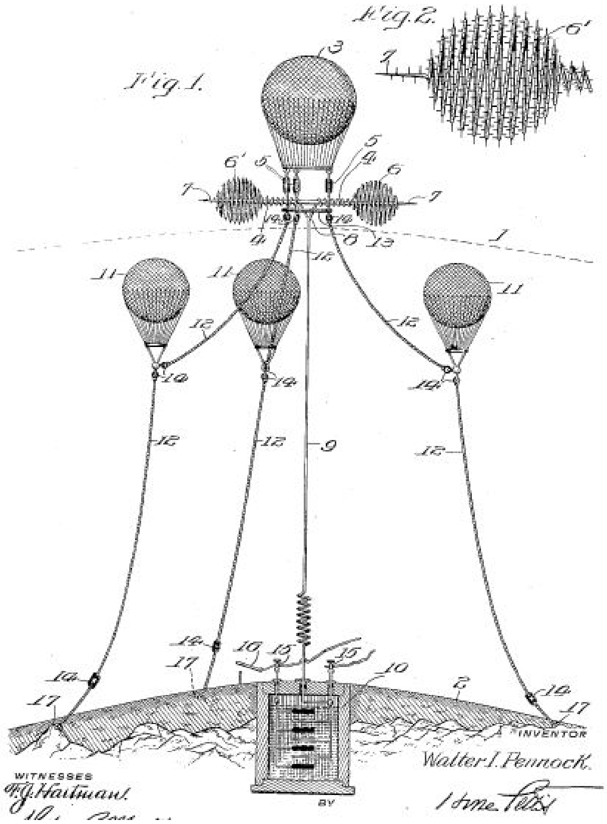
Fig.1.
Fig.2.
WITNESSES
INVENTOR
Walter I. Pennock.
BY

Jules Guillot

The Atmospheric Electric Siphon invented by Jules Guillot in the 1920s generated about 2.5-3 kilowatts with antenna height of ~ 20 meters. The power depends on the total collector surface and height of the vertical antenna. A tabletop apparatus with only a 2 meter tall collector produced ~300 watts.

The great promise of Guillot's device was reported in *The Invention Encyclopedia* (1930, edited by the eminent engineer George Constantinescu):

> "The capture of atmosphere electricity has been used in France, with aerial cables mounted on the Mont Blanc, and also in Germany --- with conductive cables carried by the captive balloons.
>
> "The atmosphere electricity collect system invented by eng. Jules Guillot is most ingenious and it relies on "the electric siphon". His method consisted in the direct "pumping" of the atmosphere electricity using a collecting device which had two antennae and several collecting rods.
>
> "One antenna is vertical and it has a lot of rods scattered like an opened fan, with the tips against the zenith, for collecting the negative electricity which comes from the air; the horizontal antenna is orientated against the South and its role is to collect the positive electricity:
>
> "Guillot used two separated and insulated armatures with the positive armature against the South (more precisely, against the Equator...) and the negative armature against the zenith... Also, J. Guillot used an electrical transformer for the industrial utilization of this collector as power supply for industrial electric engines."

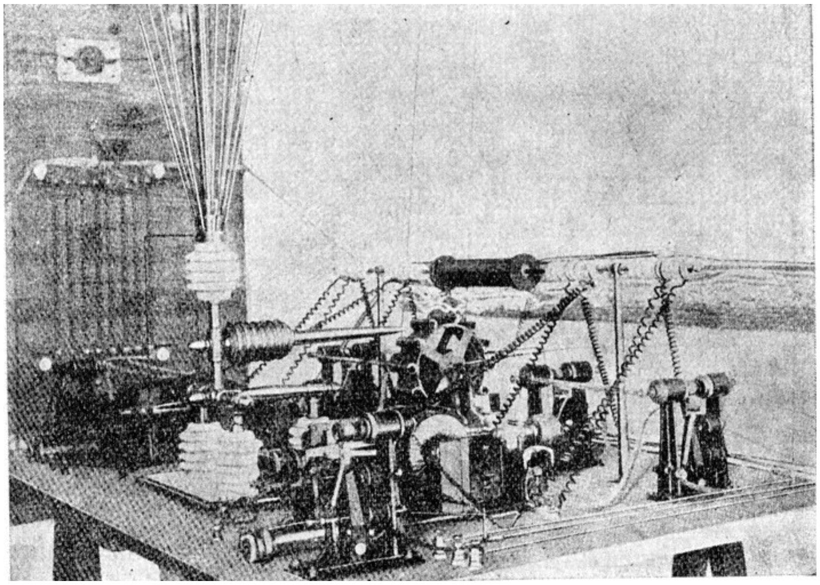

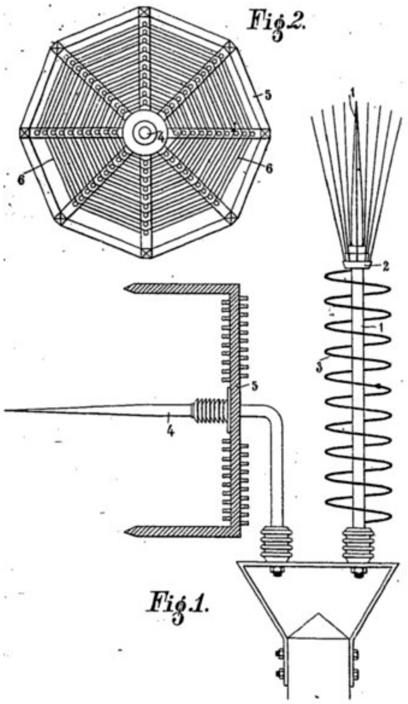

Fig.2.

Fig.1.

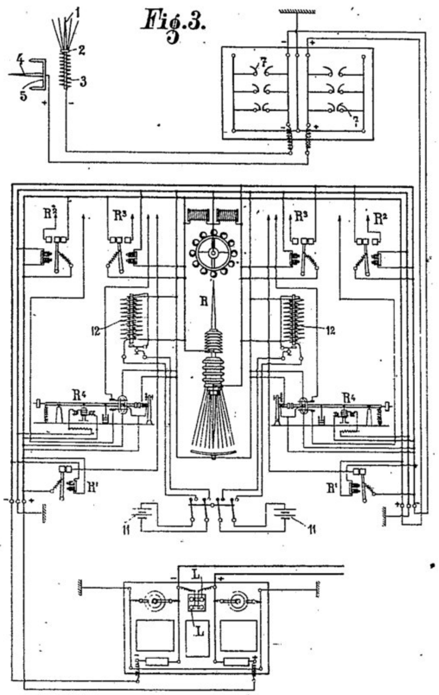

Fig.3.

The electric siphons produce a magnetic field and absorb the aerial electricity. A Ruhmkorff induction coil is used to jump-start the system. Guillot received three French Patents for his "Apparatus for Capture of Atmospheric Electric Currents with Immediate Implementation" (FR551882, FR565395, and FR551882)

Hermann Plauson

The Estonian Hermann Plauson was director of the Fischer-Tropsch Laboratories in Hamburg, Germany during the 1910s and 20s,. He thoroughly investigated atmospheric electricity and constructed practical apparatus to utilize it. His book "*Gewinnung und Verwertung der Atmosphärischen Elektrizität*" (1920) provides a detailed explanation of the technology.

Hugo Gernsbach presented the invnetion in *Science & Invention* magazine (March 1922, June 1928). He noted that the system was "actually in use small power plants, that generate electricity direct from the air, day and night, without interruption at practically no cost, once the plant is constructed."

> "Herr Plauson found in his experiments that a single balloon sent aloft to a height of 300 yards gave a constant current at 400 volts of 1.8 amperes, or in 24 hours over 17-1/4 kilowatts! By using two balloons in connection with a special condenser battery, the power obtained was 81-1/2 kilowatts in 24 hours. The actual current delivered was 6.8 amperes at 500 volts.
>
> "The best balloons... are made of thin aluminum leaf, filled with helium...The surface is dotted with electrolytically sharpened pins amalgamated with zinc, and a pinch of radium salt to further ionize the air. By dotting the balloon with photoelectric zinc or polonium amalgam, the amount of electricity can be greatly increased."

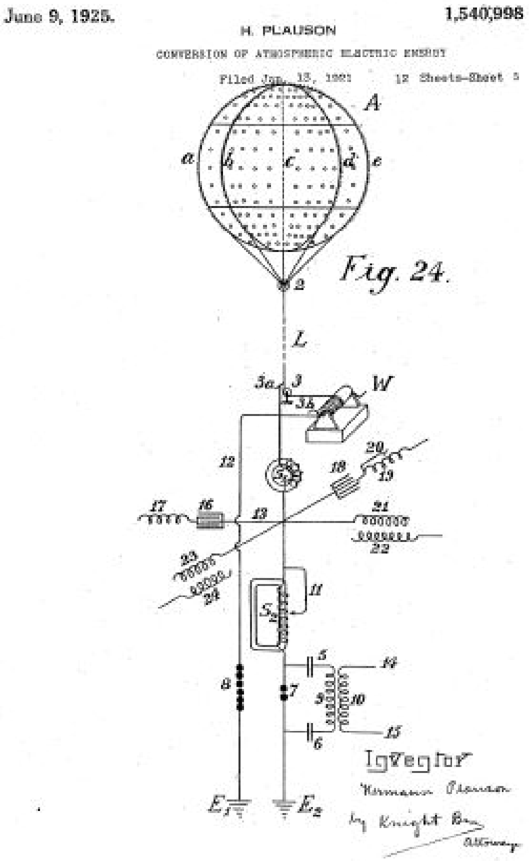

One hundred balloons, 100 yards apart, will generate at least 200 horsepower, up to 400 during winter. Plauson used batteries of condensers and high voltage transformers to light lamps, run motors and charge batteries, etc. He also invented an electrostatic rotary transformer to produce alternating current. The system literally sucks electricity from the collector balloons. The balloons also act as lightning arresters and quickly discharge thunder clouds.

> "There is no doubt that this invention will soon come into universal use all over the world. We will see the land dotted with captive balloons, particularly in the country and wherever water power does not abound. Indeed, the time is not distant when nearly all of our power will be derived from the atmosphere. So far it seems to be the cheapest form of power known... safely extracting several kilowatts of electrical power from the atmosphere with metallic surfaced balloons, elevated to a height of only 1000 feet.
>
> "German patents show the use of a kite balloon from which hangs a metal net to collect electricity. The tether-conductor leads to a windlass. The patents claim that at a height of one mile 225,000 volts will be available. Plauson proposed the construction of insulated towers about 1000 feet high to support the aerials... [H]e carried out experiments with a balloon made of aluminum leaf with collecting needles of amalgamated zinc with a radium preparation as an ionizer. The surface of the balloon was sprinkled over with zinc amalgam. It was sent up to a height of 300 meters, early 1,000 feet, and was held by a copper-plated steel wire. A constant current of 1.8 amperes at an average of 400 volts potential difference was obtained. This gave nearly three-quarters of a kilowatt, or close to one horsepower. The collector of the balloon insulated from the earth showed a tension of 42,000 volts. By sending up a second balloon with an antenna to the same height at a distance of 100 meters from the first balloon, a current of over 3 amperes was obtained. Then by putting into the circuit a large condenser, whose capacity was equal to the surface capacity of both balloons, and of the antenna connections, the current rose to 6.8 amperes with about 500 volts mean tension. By the use of these two balloons, he eventually ran up the power to 3.4 kilowatts."

No. 674,427. Patented May 21, 1901.

A. PALENCSÁR.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

(Application filed July 10, 1900.)

(No Model.) 3 Sheets—Sheet 1.

Fig: 1.

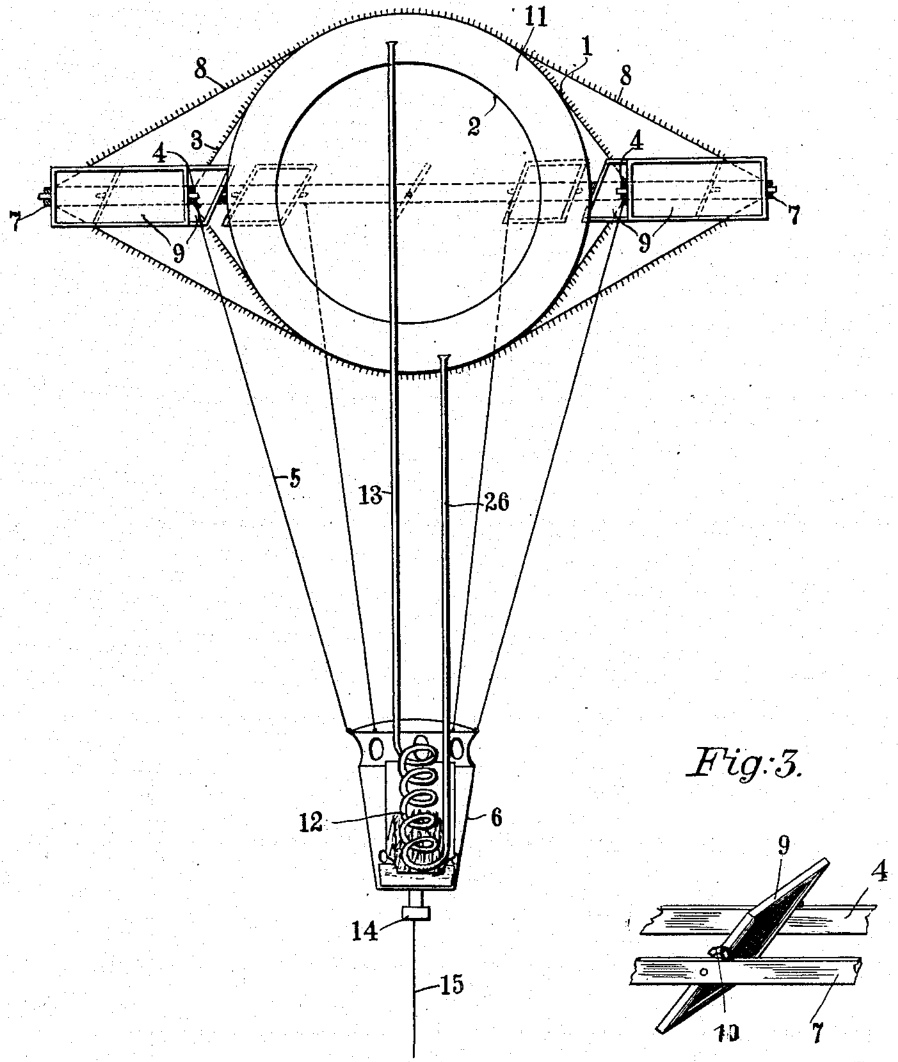
11
1
8
8
2
3
4
4
7
7
9
9
9
5
13
26
12
6
14
15

Fig: 3.

9
4
10
7

WITNESSES:
Ella L. Giles
Otto Munk

INVENTOR
Andor Palencsár
BY
Richards
ATTORNEYS

THE NORRIS PETERS CO., PHOTO-LITHO., WASHINGTON, D. C.

No. 674,427. Patented May 21, 1901.

A. PALENCSÁR.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

(Application filed July 10, 1900.)

(No Model.) 3 Sheets—Sheet 2.

Fig: 2.

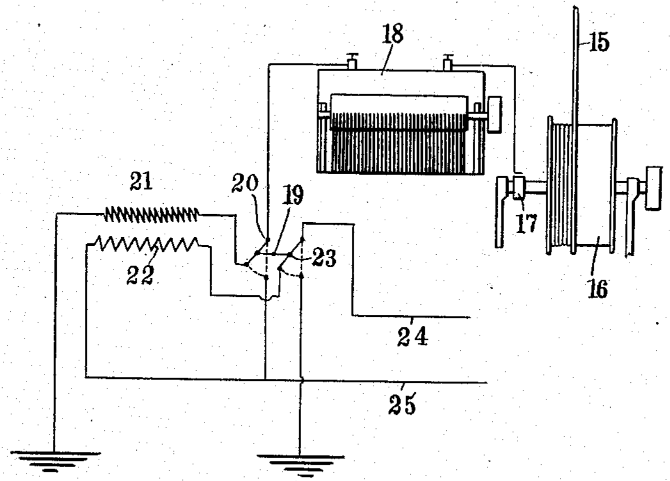

WITNESSES:
Ella L. Giles
Otto Munk

INVENTOR
Andor Palencsár
BY
Richards & Co
ATTORNEYS

THE NORRIS PETERS CO., PHOTO-LITHO., WASHINGTON, D. C.

No. 674,427. **Patented May 21, 1901.**

A. PALENCSÁR.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

(Application filed July 10, 1900.)

(No Model.) 3 Sheets—Sheet 3.

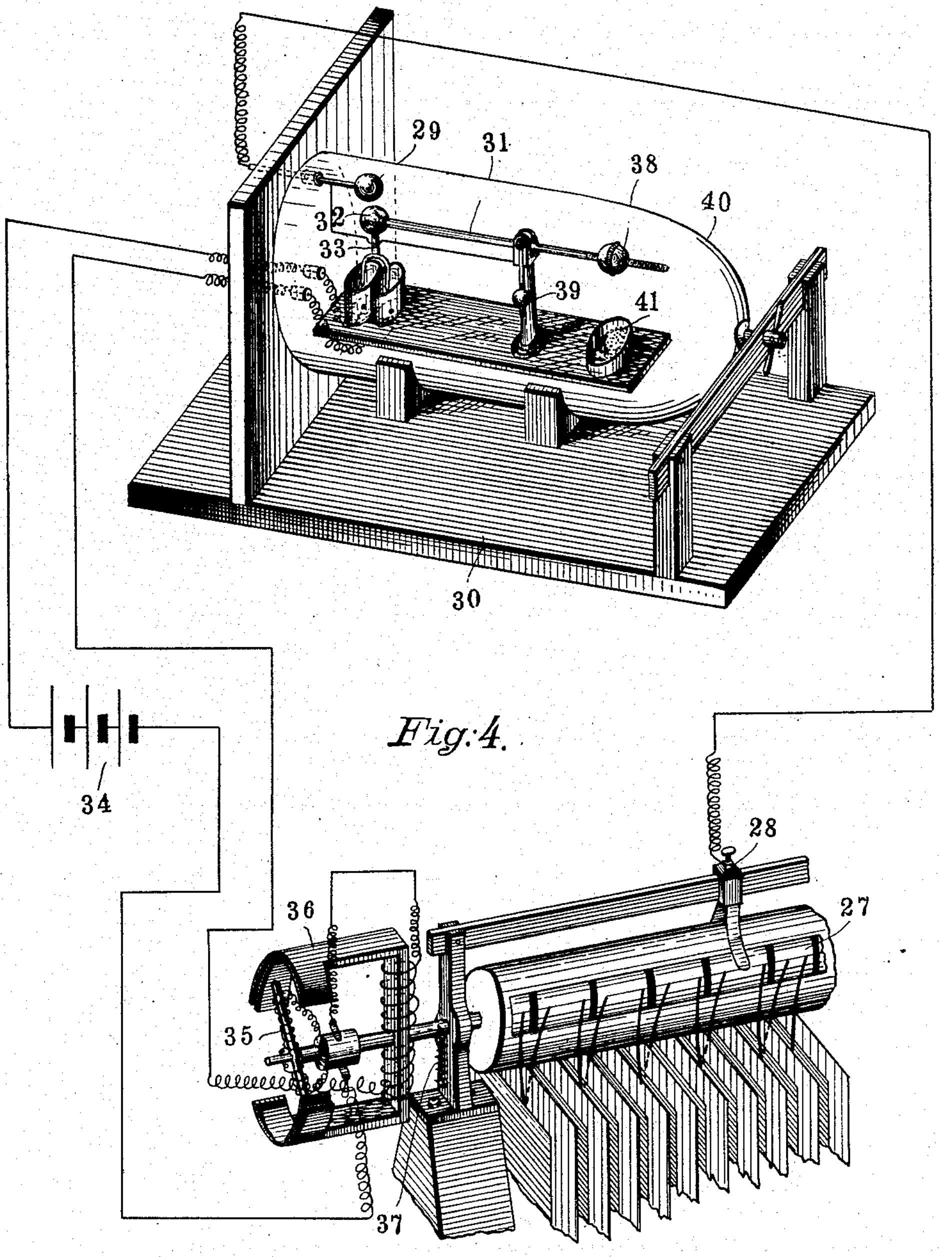
29
31
38
40
32
33
39
41
30
34
28
27
36
35
37

Fig. 4.

WITNESSES:

Ella L. Giles

Otto Munk

INVENTOR

Andor Palencsár

BY

Richards & Co

ATTORNEYS

THE NORRIS PETERS CO., PHOTO-LITHO., WASHINGTON, D. C.

UNITED STATES PATENT OFFICE.

ANDOR PALENCSÁR, OF BUDA-PESTH, AUSTRIA-HUNGARY.

APPARATUS FOR COLLECTING ATMOSPHERIC ELECTRICITY.

SPECIFICATION forming part of Letters Patent No. 674,427, dated May 21, 1901.

Application filed July 10, 1900. Serial No. 23,102. (No model.)

To all whom it may concern:

Be it known that I, ANDOR PALENCSÁR, a subject of the King of Hungary, residing at Buda-Pesth, Austria-Hungary, have invented a certain new and useful Apparatus for Collecting Atmospherical Electricity, of which the following is a full, clear, and exact specification.

The present invention concerns an apparatus for collecting and driving or conducting atmospherical electricity which renders it practicable to obtain material quantities of the same in serviceable form.

The experiments made hitherto have been limited to deviating or conducting the electricity by simple lightning-rods without making allowance for the progress of the theory of atmospherical electricity. It is obvious that only extremely small quantities of electricity can be collected in this manner, and, moreover, it was obtained in a form which entirely precluded its utilization. It is evident that a source of electricity of irregular yield can only be utilized by means of accumulators and for charging accumulators only a current of constant potential can be employed. The potential of electricity derived or deviated by means of a lightning-rod varies within wide limits, and, moreover, it is so high that it cannot be used at all for charging accumulators. All these drawbacks are remedied by the present invention, by which the atmospheric electricity is obtained in larger quantities with as low and constant a potential as may be desired.

The idea underlying the invention is based on the modern theory of atmospheric electricity, according to which it is produced by the condensation of steam or aqueous vapors, and that the increase of potential is effected by the concentration of the small drops of water into larger ones, as the proportion of the surface of the drops to the volume of same is materially reduced thereby. According to this theory the water-droplets floating in the layers of air are considered as vehicles or carriers of the electricity, and a rational system of the deviation of the atmospheric electricity must derive it from the water-drops. This is attained by the present invention in the following manner: A collecting-body of as large a surface as possible is provided with sharp needles. This is moved in the higher layers of air, while being continually heated by a suitable heating device. Owing to the heat the water-drops immediately ambient to the collecting-body will be evaporated, their capacity is gradually reduced, while the potential of the charge grows until it reaches an infinite height with infinite smallness of the drops. It is readily apparent that the whole charge of the drops as soon as they have been evaporated will have passed to the collector or collecting-body, from which it can be conducted. For the purpose of replacing the evaporated drops, which have been deprived of their charge by new-charged drops, the collector is moved in relation to the ambient air.

One way of carrying out the invention is represented in diagram in the accompanying drawings, in which—

Figure 1 is a longitudinal section of the collecting-balloon. Fig. 2 shows the deviating and conversion device, and Fig. 3 a detail. Fig. 4 is a diagrammatic perspective view of a form of automatic regulator for the rheostatic machine.

The apparatus consists of a balloon having two walls and covered with a light wire net, preferably of aluminium wire, said net being studded with needles. Besides, the balloon carries the net 3, on which the ring 4, made of a solid but light material, (wood, cane, &c.,) is fixed. This ring carries the basket 6 by means of the cords or ropes 5. On a level with the ring 4 is the ring 7, which is kept spread by the blades or wings 9, which are journaled in a manner to rotate easily. The ring 7 is kept fixedly in position by the cords or ropes 8. The blades or wings consist of a frame covered with light material, and their rotation in either direction is limited by the stops or ledges 10. The blades or wings form advantageously an angle of sixty to seventy degrees with the vertical line.

All wire nets, ropes, rings, and blades or wings may be covered with small metal needles which are electrically connected with each other.

From the space 11, formed between the double walls of the balloon, the pipe 26 leads from the lowest point of the same to the serpentine 12, arranged in the basket 6, which in turn is connected with the pipe 13, ending in the upper part of the space 11 between the two walls 12. The serpentine is heated by means of a suitable source of heat, whereby a warm current of gas or air circulates continually between the double walls of the balloon.

Under the basket the ball-bearing 14 is arranged in electrical connection with the wire nets, and its stud is electrically connected with the carefully-insulated light though sufficiently strong cable 15.

On the earth's surface is a winch 16, Fig. 2, by means of which the balloon can be made to ascend or descend as soon as the interior space of the balloon is filled with illuminating gas or hydrogen.

The end of the cable-core is soldered to a collector arranged upon and insulated from the axle of the winch, and the electricity is conducted from this collector by means of a sliding contact. The collecting of the electricity takes place by moving the balloon continuously up and down by means of the winch. In this movement the balloon is turned by means of the wings or blades 9, which are adapted to turn the balloon always in the same direction whether ascending or descending, as in the change from ascending into descending of the balloon, or vice versa, the blades are turned over by the aerial resistance, and thus impart the rotating motion to the balloon in the same direction. In order to avoid torsion of the cable, the ball-bearing 14 is provided. This up-and-down motion and rotating of the balloon accomplishes the purpose of bringing it into contact with as many water particles floating in the air as possible.

As the electricity conducted from the collector-wing 17 possesses a much too high and varying potential for making its direct application practicable, and as it is usual with an irregular source of electricity to first charge accumulators and to further utilize the easily-regulated current of the same only, it becomes necessary to seek to maintain the electricity conducted from the collector 17 for the charging of the accumulators at a constant potential and convert the potential to a much lower one; but as we deal in this case with a direct current ordinary converters cannot be used for this purpose. Moreover, the electricity possesses in this case a much too high potential, so that with the employment of ordinary converters the greatest part of the collected electricity would be lost again. The only practical method for this purpose is the converting by means of the Planté rheostatic machine, by which this high-voltage electricity can be transformed almost without any loss whatever. Thus the question of conversion would be solved, and only the question of maintenance of constant potential remains. This is obtained in the following manner: The rheostatic machine 18 or only a part of the plates is connected with an electrometer of any construction, the movable part of which closes a contact which eventually actuates an electromagnet which effects the switching or reversion of the rheostatic machine. After the switching of the rheostatic machine it is discharged, the potential falls to zero, (*nil*,) and the electrometer resumes its initial position, whereby the current of the electromagnet which effects the reversion is interrupted and the plates of the rheostatic machine are reconnected to potential. The machine is then ready for renewed charge and is again discharged when the determined potential is reached. This action is continually repeated as long as the apparatus is in operation.

A form of the automatic regulator for the rheostatic machine is represented in Fig. 4. 27 is the contact-cylinder of the rheostatic machine, on which, for clearness sake, only the contacts for the charging position of the condenser-plates are shown, while the contacts for the discharging position, which come into action after the contact-cylinder has been turned, are omitted. The coating of the condenser-plates is connected in electrical circuit with the stationary ball 29 and the movable ball 32 of the charge-meter 30. When the charge of the rheostatic machine rises, the ball 32, arranged on the one extremity of the lever 31, is repelled and at a certain stroke actuated by means of the fork 33, which is fastened on the ball 32, and by dipping into a mercury-cup it closes the circuit of the source of current 34. This current passes through the windings of the anchor or armature 35, fixed on the shaft of the contact-cylinder, and it also passes through the electromagnet 36. Thereby the contact-cylinder is turned by a certain angle and the rheostatic machine is thus reversed. Then if the tension decreases by discharge and the repulsion of the balls 29 32 declines so far that the contact is interrupted at 33 the spring 37 turns the contact-cylinder into its normal position and the rheostatic machine is again switched to tension. The regulation of the electrometer is effected by the adjustable weights 38 39. When larger quantities of electricity shall be derived or deviated, two rheostatic machines may operate alternately so, that while the one is being discharged the other can be charged. It is readily apparent that if the capacity of the rheostatic machine is not changed the switching or reversion by the electrometer will always take place at the same potential of the rheostatic machine, and as the number of plates, and hence the proportion of conversion, remains the same the current impulses derived or deviated from the rheostatic machine will also have the same potential.

The irregularities of the source of electricity change the interval of time in which the charges follow each other; but as long as the potential remains constant this has no injurious effect on the charge of the accumulators. The current derived, or deviated from the rheostatic machine can be further transformed by an ordinary converter 21 22, and whenever it is sufficiently constant it can be utilized without the intervention of the accumulators. The converters can be connected or disconnected by means of the double switch 19 20 23.

24 and 25 are the conducting-wires, which run either directly to the place of consumption or to an accumulator-battery.

Having now described my invention, what I claim as new, and desire to secure by Letters Patent, is—

1. An apparatus for collecting atmospheric electricity comprising a collecting-body adapted to be kept in motion, heating means for said body, and a rheostatic machine and a converter connected with the said body electrically, substantially as described.

2. An apparatus for collecting atmospheric electricity for storage comprising a collecting-body adapted to be kept in motion in the ambient air, heating means for said body, a conductor leading from said body, a rheostatic machine connected with the said conductor, an electrometer connected with said rheostatic machine electrically, a contact controlled by the electrometer, and an electromagnet controlled by said contact, said electromagnet controlling the reversing of the rheostatic machine, substantially as described.

3. In combination, the balloon-like collecting-body, having the collecting-points, means for heating the interior space of said body, a conductor leading from the balloon-like body, and electrical devices for receiving the current therefrom, substantially as described.

4. In combination, the balloon-like collecting-body, means for turning the same constantly in one direction in both ascending and descending, conducting means leading from the balloon and electrical devices for receiving the current therefrom, substantially as described.

In witness whereof I have hereunto set my hand in presence of two witnesses.

ANDOR PALENCSÁR.

Witnesses:
EUGÈNE HARRÄNJO,
PAUL BÖLESKEY.

US 20120286623A1

(19) **United States**

(12) **Patent Application Publication**
Ogram

(10) **Pub. No.: US 2012/0286623 A1**

(43) **Pub. Date: Nov. 15, 2012**

(54) **ATMOSPHERIC ENERGY COLLECTION**

(75) Inventor: **Mark E. Ogram**, Tucson, AZ (US)

(73) Assignee: **Sefe, Inc.**, Tempe, AZ (US)

(21) Appl. No.: **13/103,963**

(22) Filed: **May 9, 2011**

Publication Classification

(51) **Int. Cl.**
H02N 3/00 (2006.01)

(52) **U.S. Cl.** **310/309**

(57) **ABSTRACT**

The subject matter described herein is an atmospheric energy collector. The atmospheric energy collector includes of a windsock arrangement that has a large up-wind opening on one side and that tapers from the larger up-wind opening on the one side to a small down-wind opening on the other side. The up-wind side is secured to a tether such that an electrically conducting material (e.g. metal) included in construction of the atmospheric energy collector is connected to the tether. The windsock arrangement is extended outwards by wind and the like atmospheric conditions such that the electrically conducting material collects the atmospheric energy and transfers the collected energy to the tether.

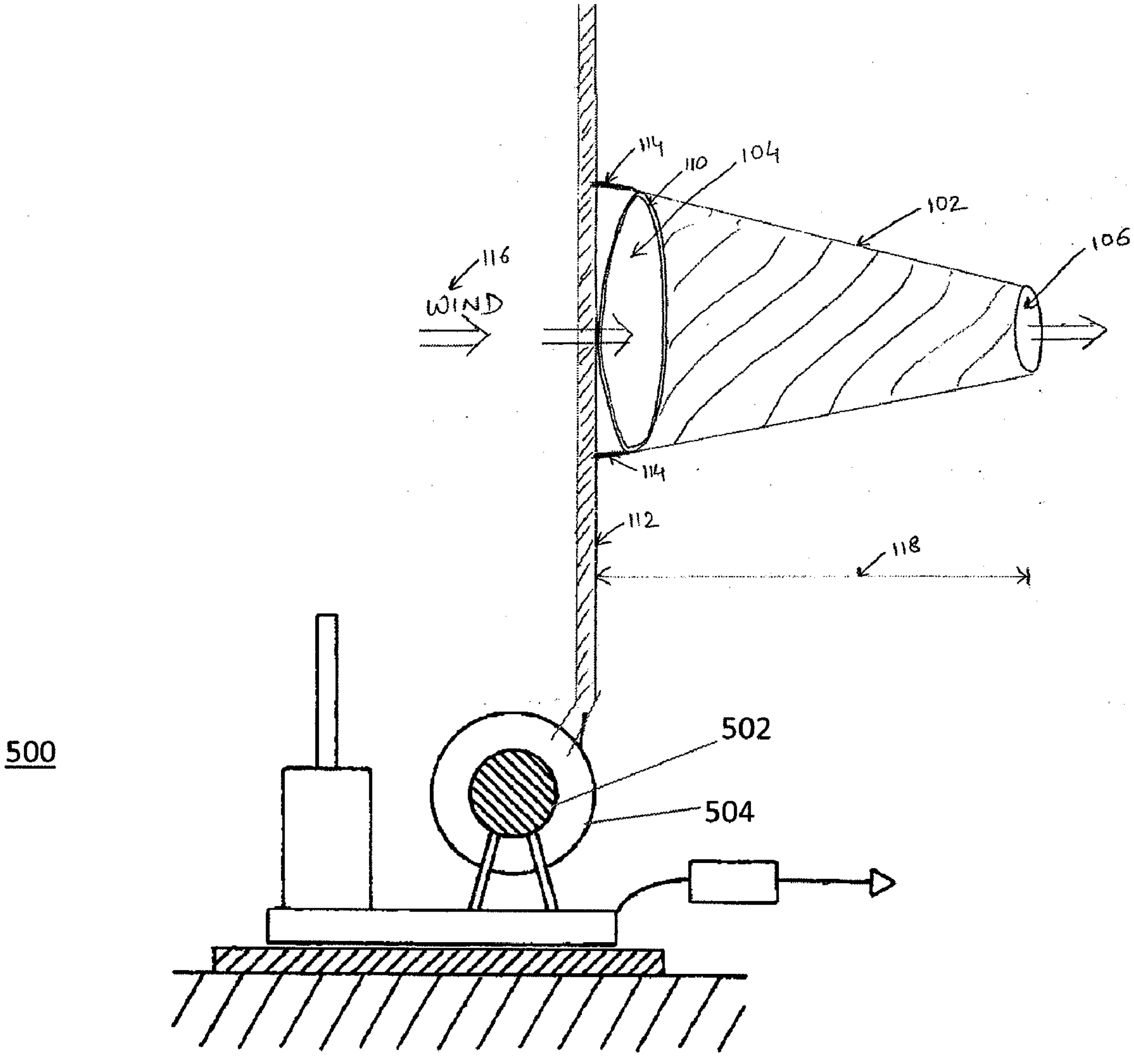

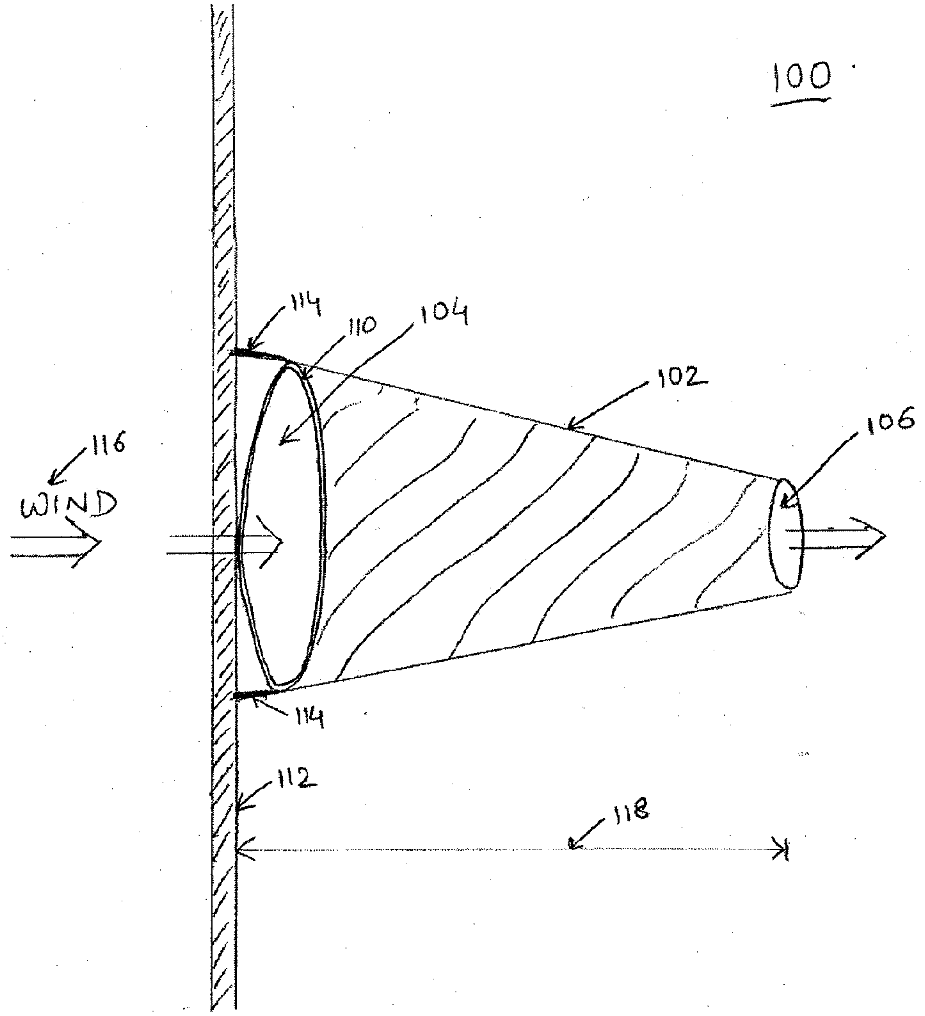
100
114
110
104
102
106
116
WIND
114
112
118

FIG. 1A

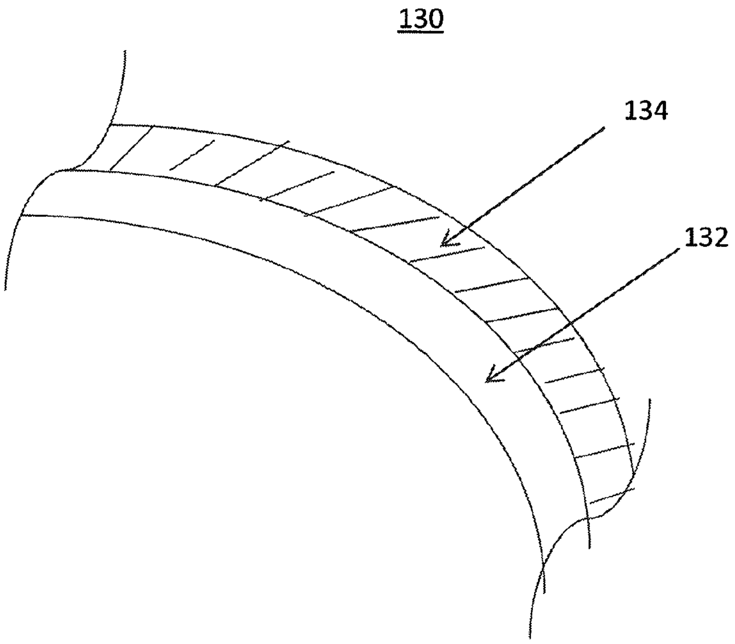

FIG. 1B

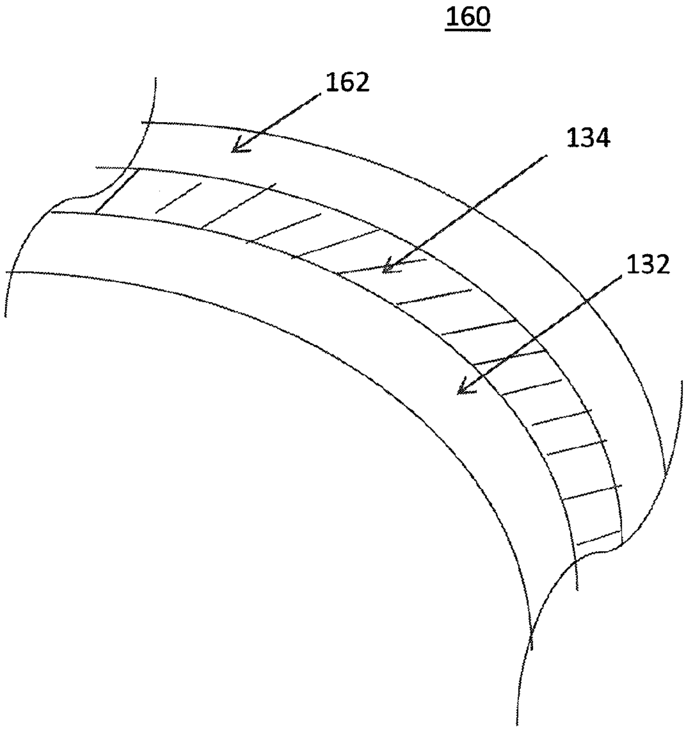
160
162
134
132

FIG. 1C

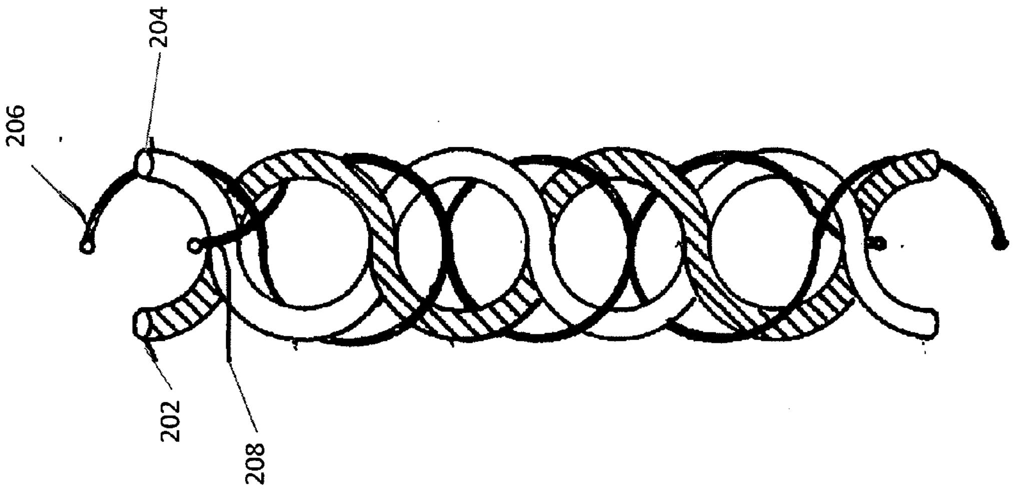
206
204
202
208

FIG. 2

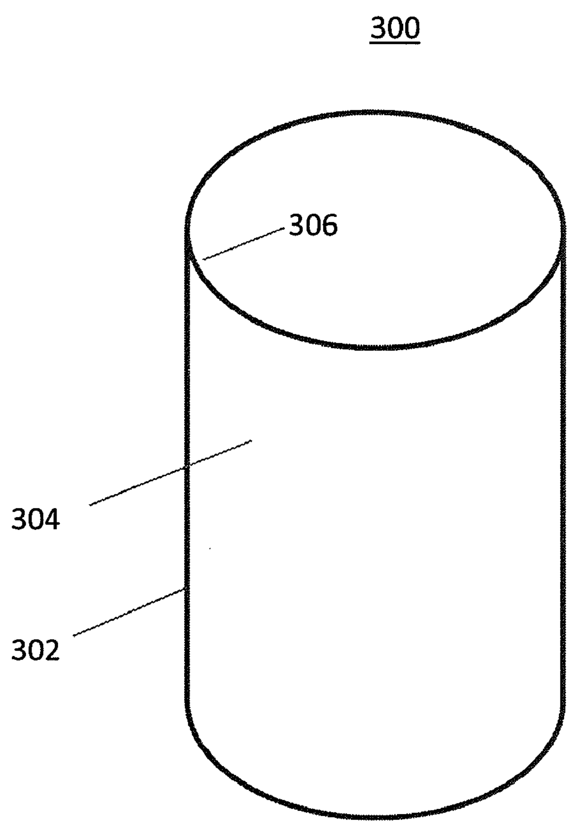
300
306
304
302

FIG. 3A

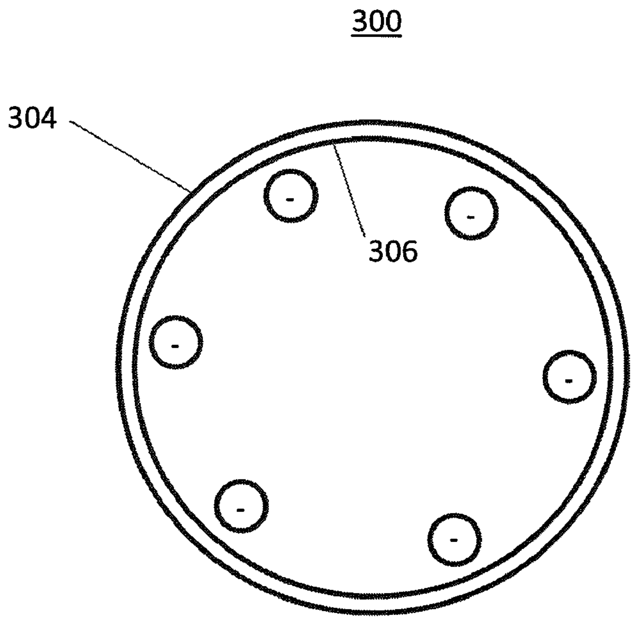

FIG. 3B

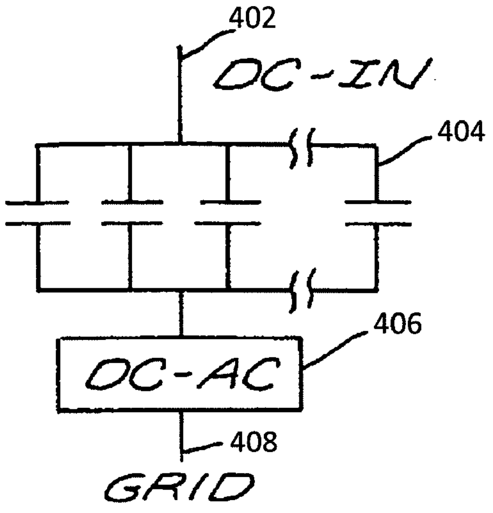
402
DC-IN
404
406
DC-AC
408
GRID

FIG. 4A

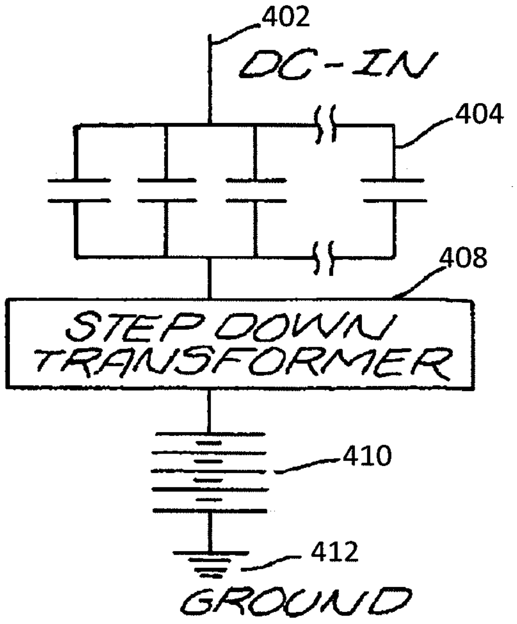
402
DC-IN
404
408
STEP DOWN TRANSFORMER
410
412
GROUND

FIG. 4B

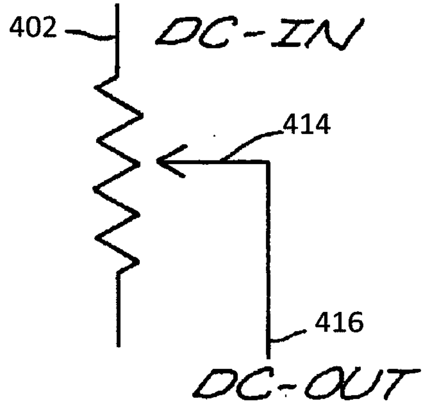
402
DC-IN
414
416
DC-OUT

FIG. 4C

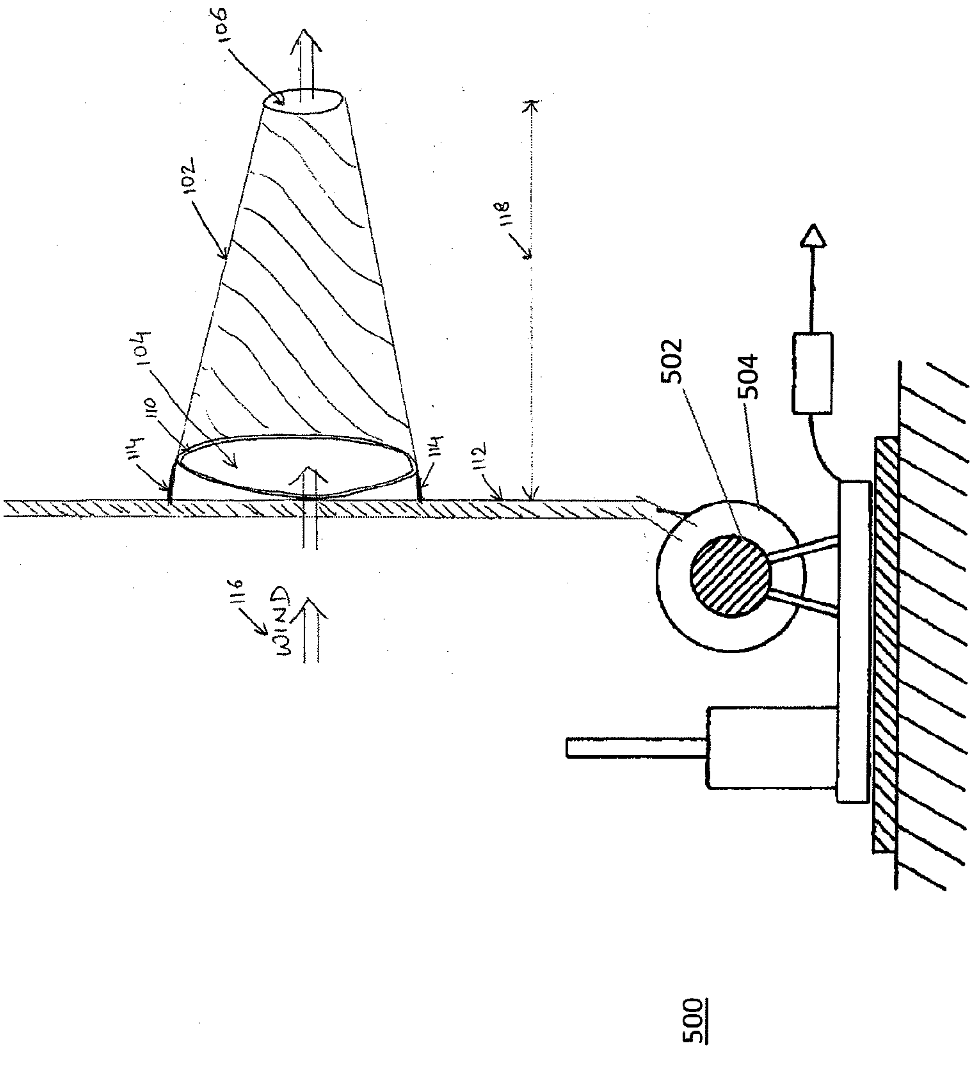
106
102
118
104
110
114
114
112
116
WIND
502
504
500

FIG. 5

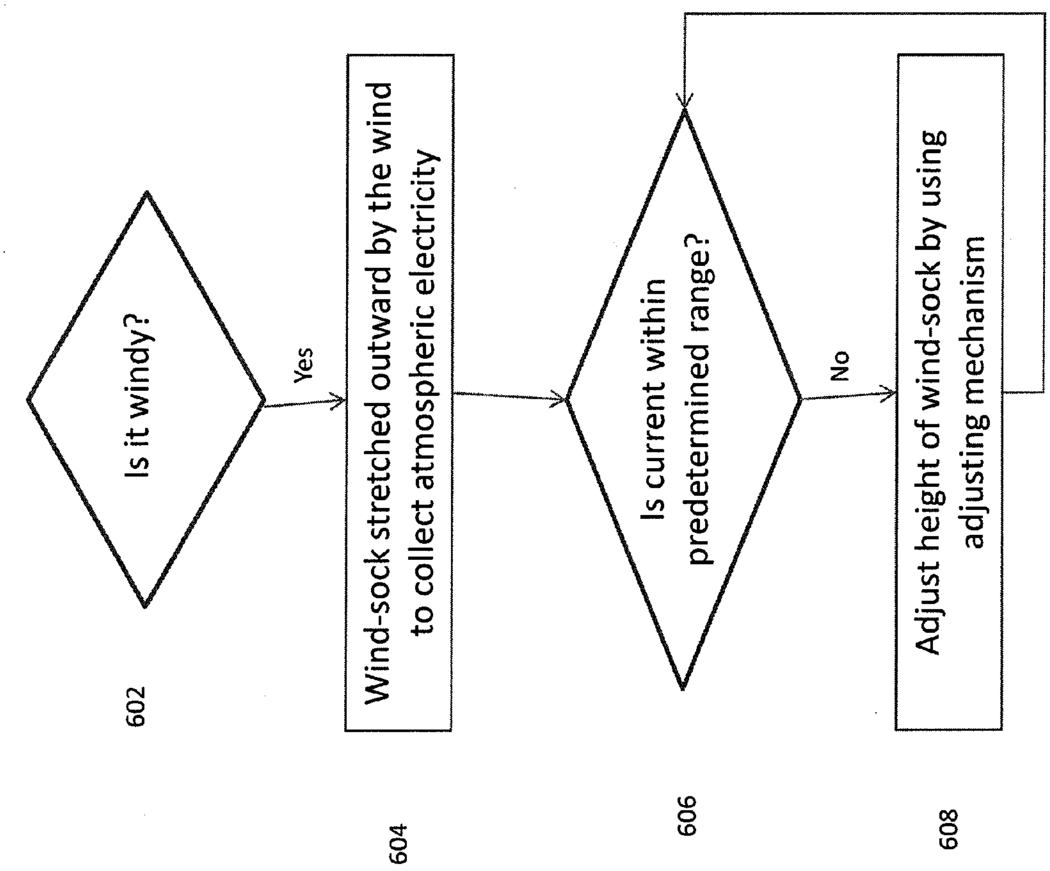
602
Is it windy?
Yes
604
Wind-sock stretched outward by the wind to collect atmospheric electricity
606
Is current within predetermined range?
No
608
Adjust height of wind-sock by using adjusting mechanism

FIG. 6

ATMOSPHERIC ENERGY COLLECTION

TECHNICAL FIELD

[0001] The subject matter described herein relates to a light-weight, extendible electricity collecting windsock arrangement comprising an enhanced collection surface for atmospheric electrical energy collection.

BACKGROUND

[0002] The atmosphere above the earth is known to include electrical charge. The earth's surface is negatively charged, while the air above it is positively charged. All atmospheric effects are a result of an interplay between these two huge areas of opposite electrical energy. The potential difference between the positively charged atmosphere and the negatively charged earth surface causes atmospheric electrical energy to be developed. Some of this atmospheric electrical energy may be experienced through thunderstorms.

[0003] Even though there is a huge amount of atmospheric electrical energy in the atmosphere, collection of the atmospheric electrical energy still remains a problem. Due to high voltage and low current conditions, there is a need for large collection surfaces for maximizing the collection of atmospheric electrical energy. Note that the terms atmospheric electricity, atmospheric energy, and atmospheric electrical energy have been used interchangeably in this specification.

[0004] The problem of low current is compounded by a heavy weight of the collectors of atmospheric electrical energy. The heavy weight of the collectors increases payload requirements of a lift mechanism, which may be used to lift the collectors to an appropriate height or altitude so that the collection of atmospheric electrical energy is maximized. Thus, a need exists for large yet light-weight collectors such that payload requirements for the lift mechanism remain reasonable.

SUMMARY

[0005] The subject matter described herein relates to an atmospheric energy collector. The atmospheric energy collector includes an electricity collecting windsock arrangement that has a large up-wind opening on one side and that tapers from the larger up-wind opening on the one side to a small down-wind opening on the other side. The up-wind side is secured to a tether such that an electrically conducting material (e.g. metal) included in construction of the atmospheric energy collector is connected to the tether. The electricity collecting windsock arrangement is extended outwards by wind and the like atmospheric conditions such that the electrically conducting material collects the atmospheric energy and transfers the collected energy to the tether. Further, the electricity collecting windsock includes light-weight collectors such that payload requirements for the lift mechanism remain reasonable (i.e. remain within a predetermined value).

[0006] In one aspect, an electricity collection apparatus includes a tether and a windsock. The windsock is formed of an insulation material with a metal deposited on at least a portion of the insulation material. The metal of the windsock is electrically connected to the tether. The windsock extends in a direction of ambient wind to provide a surface area based on the ambient wind, and the surface area provides the metal to collect electrical energy from the ambient wind, the collected electrical energy being transferred to the tether.

[0007] In another aspect, an electricity collection apparatus includes a windsock formed of an insulation material the windsock to extend in a direction of ambient wind to provide a surface area based on the ambient wind. The apparatus further includes one or more electrical conductors provided on at least a portion of the insulation material to collect electrical energy from the ambient wind. The apparatus further includes an electrically conductive tether connected with windsock to anchor the windsock in the direction of the ambient wind, the electrically conductive tether being electrically connected to the one or more electrical conductors provided on at least a portion of the insulation material to transfer the collected electrical energy from the one or more electrical conductors to an electrical storage.

[0008] The subject matter described herein provides many advantages. For example, large collection surfaces allow a maximized collection of atmospheric energy. Moreover, the light-weight collectors allow payload requirements for the lift mechanism to remain reasonable (i.e. remain within a predetermined value).

[0009] The details of one or more variations of the subject matter described herein are set forth in the accompanying drawings and the description below. Other features and advantages of the subject matter described herein will be apparent from the description and drawings, and from the claims.

DESCRIPTION OF DRAWINGS

[0010] The accompanying drawings, which are incorporated in and constitute a part of this specification, show certain aspects of the subject matter disclosed herein and, together with the description, help explain some of the principles associated with the disclosed implementations. In the drawings,

[0011] FIG. **1**A illustrates extendible electricity collecting windsock arrangement that is used for collection of atmospheric electrical energy in accordance with implementations of the current subject matter;

[0012] FIG. **1**B illustrates a cross-section of one implementation of the electrically conducting windsock in accordance with implementations of the current subject matter;

[0013] FIG. **1**C illustrates a cross-section of another implementation of the electricity collecting windsock in accordance with implementations of the current subject matter;

[0014] FIG. **2** illustrates an exemplary conductive line in accordance with implementations of the current subject matter;

[0015] FIG. **3**A illustrates a front view of a conductive line in accordance with implementations of the current subject matter;

[0016] FIG. **3**B illustrates a top view of the conductive line shown in FIG. **3** in accordance with implementations of the current subject matter;

[0017] FIGS. **4**A, **4**B, and **4**C illustrate electrical schematics for handling the static charge from the atmosphere in accordance with implementations of the current subject matter;

[0018] FIG. **5** illustrates an adjusting apparatus (winch motor and spool) that enables adjustment of location/height of the electricity collecting windsock arrangement in accordance with implementations of the current subject matter; and

[0019] FIG. **6** is a process flow diagram illustrating aspects of a method in accordance with implementations of the current subject matter.

[0020] When practical, similar reference numbers denote similar structures, features, or elements.

DETAILED DESCRIPTION

[0021] To address these and potentially other issues with currently available solutions, one or more implementations of the current subject matter provide methods, systems, articles or manufacture, and the like that can, among other possible advantages, provide an energy collector formed as a windsock and having an enhanced collection surface for atmospheric electrical energy collection. In preferred implementations, an electricity collecting windsock is formed as a flexible cylinder, a truncated cone, or a cone with a metal surface, and is mounted to a mast or other tether so as to capture ambient wind which extends the windsock outward from the mast, in turn forming the largest possible surface area for the wind conditions to enable the metal surface to collect atmospheric amperage. The collected amperage is transferred from the metal surface of the electricity collecting windsock to the mast or tether, and eventually to an electricity storage device. Although implementations in which windsock arrangements having one windsock each are discussed below, those or ordinary skill in the art understand that a windsock arrangement may include two or more windsocks for enhanced atmospheric electricity collection.

[0022] FIG. **1**A illustrates an implementation of electricity collecting windsock arrangement **100** that is used to collect atmospheric electrical energy. The electricity collecting windsock arrangement **100** includes an extendible electricity collecting windsock **102**. The extendible electricity collecting windsock **102** is made of a material that has a high tensile strength and provides electrical insulation. This electrically insulating material **134** (discussed below with respect to FIG. **1**B) may be a plastic or a flexible glass, such as polyester like polyethylene terephthalate (PET) or biaxially-oriented polyethylene terephthalate (BoPET). BoPET is known in the industry by different trade names, some of which are Mylar, Melinex and Hostaphan.

[0023] As is described later with respect to FIGS. **1**B and **1**C, which illustrate cross-sections of implementations of electricity collecting windsock **102**, a thin coat of an electrically conducting material **132**, **162** is deposited on the electrically insulating material **134**. The deposited electrically conducting material **132**, **162** is a good conductor of electricity, and helps electricity collecting windsock **102** to tap atmospheric electrical energy. The electrically conducting material may be any one of a metal or a suitable alloy, such as one of or a combination of one or more of gold, silver, copper, aluminum, and the like.

[0024] The electricity collecting windsock **102** has a large up-wind opening up-wind opening **104** on one side. The electricity collecting windsock **102** tapers from the large up-wind opening **104** on the one side to a small down-wind opening **106** on the other side of the electricity collecting windsock **102**. The up-wind opening **104** has a metal loop **110** on the circumference of the up-wind opening **104**. The metal loop **110** is attached to a tether **112** using an attachment mechanism **114**. The attachment mechanism **114** may include a soldering mechanism. In some implementations, other attachment mechanisms **114** are also known to be used, such as nut and bolt mechanism, threading mechanism, gluing mechanism, and the like. The electricity collecting windsock **102** is extended outwards by wind **116** and/or like atmospheric conditions such that the electrically conducting material (e.g. metal) collects the atmospheric electrical energy and transfers the collected energy to the tether **112**.

[0025] The tether **112** includes a conductive line that is isolated from the ground. The conductive line is used to transfer the collected atmospheric electrical energy from the electricity collecting windsock **102** to an electricity storage device (not shown) where the collected atmospheric charge may be gathered for later or simultaneous use.

[0026] When wind **116** strikes against the electricity collecting windsock arrangement **100** in the direction illustrated, the electricity collecting windsock **102** extends outward, thus allowing the electrically conducting material **132** (discussed below with respect to FIG. **1**B) to collect maximum possible atmospheric electricity. In one implementation, the outward extension **118** is fifty feet or more. When there is no wind, the electricity collecting windsock **102** may be in a relatively compressed state such that the length of the electricity collecting windsock **102** is less than the outward extension **118** caused by windy conditions or like atmospheric conditions. The collected atmospheric electrical energy is communicated to the tether **112**.

[0027] FIG. **1**B illustrates a cross-section **130** of one implementation of the electrically conducting windsock **102**. The cross-section **130** illustrates the electrically conducting material **132** deposited on the inner surface of the electrically insulating material **134**. The electricity collecting windsock **102** includes the electrically conducting material **132** and the electrically insulating material **134** such that the electrically conducting material **132** forms the inner surface of the electricity collecting windsock **102**. When the wind **116** strikes the electricity collecting windsock arrangement, the electricity collecting windsock **102** extends outward, thus allowing the electrically conducting material **132** to collect maximum possible atmospheric electricity. The electrically conducting material **132** may include multiple collectors that are spaced apart such that the collection surface is greatly enhanced. The multiple collectors may be spaced apart in patterns, such as spaced apart lines, squares, rectangles, circles, and the like, or any other pattern that may maximize the collection of the atmospheric electrical energy. As noted above, the electrically insulating material **134** may be a plastic or a flexible glass, such as polyester like polyethylene terephthalate (PET) or biaxially-oriented polyethylene terephthalate (BoPET). The electrically conducting material **132** may be a metal or alloy, such as one of or a combination of one or more of gold, silver, copper, aluminum, and the like.

[0028] FIG. **1**C illustrates a cross-section **160** of another implementation of the electricity collecting windsock **102**. The electrically conducting material (e.g. metal deposit) **132**, **162** may be placed on both sides of the electrically insulating material **134** such that the collection ability is increased even further. Multiple collectors are implemented and are spaced apart on electrically conducting material **132**, **162** such that the collection surface is greatly enhanced. The multiple collectors may form one or more patterns, as noted above with respect to FIG. **1**B. Furthermore, the electrically insulating material **134** (e.g. mylar) used in the electricity collecting windsock **102** is light, thereby minimizing the overall weight of the electricity collecting windsock **102** and reducing the weight for payload calculations. The electrically conducting materials **132**, **162** may be an electrically conducting metal or alloy, such as one of or a combination of at least one of gold, copper, aluminum, silver, and the like. The electrically conducting materials **132**, **162** may be made of the same metal/alloy or a different metal/alloy.

[0029] The electrically conducting materials **132**, **162** may be attached to the electrically insulating material **134** by a gluing mechanism. In some implementations, other attachment mechanisms are known to be implemented, such as paint coating mechanism, a threading mechanism, a nut and bolt mechanism, soldering mechanism, and the like. The attachment mechanism between the electrically conducting material **132** and the electrically insulating material **134** may or may not be the same as the attachment mechanism between the electrically conducting material **162** and the electrically insulating material **134**.

[0030] FIG. **2** illustrates an exemplary conductive line in accordance with one implementation. This type of conductive line is commonly referred to as poly-wire. The conductive line consists of multiple interwoven strands of plastic **202** and **204** woven into a cord or rope arrangement having intertwined therein exposed metal wires **206** and **208**. Although FIG. **2** illustrates two plastic strands and two metal wires, any number of possible combinations of plastic strands and metal wires is possible. The exposed metal wires **206** and **208** attract the atmospheric static charge and transmit the charge down to the electricity storage device (not shown).

[0031] FIG. **3**A illustrates a front view of a conductive line **300** in accordance with another implementation. FIG. **3**B illustrates a top view of the conductive line **300** illustrated in FIG. **3**. The conductive line **300** creates an ionized pathway for the flow of the static charges from the atmosphere to the electricity storage device via the electricity collecting windsock arrangement **100**. This conductive line utilizes a tube **302** having an outer layer **304** of PET Film (Biaxially-oriented polyethylene terephtalate polyester film). The tube **302** provides exceptionally high tensile strength and is chemically and dimensionally stable. In one implementation, the tube **302** may have an ideal diameter of between two and three inches. An interior metal coating **306** provides an initial conduit for the flow of static charge. The static charge through the metal may force the tube **302** to expand due to the repulsion experienced by like charges. Further, the flow of electricity causes the interior of the tube **302** to become ionized to provide an additional pathway for the atmospheric static charges to the electricity storage device (not shown).

[0032] FIGS. **4**A, **4**B, and **4**C illustrate electrical schematics for handling the static charge from the atmosphere. By maintaining the voltage being collected in a prescribed range, an electrical conversion system is easily designed. While FIGS. **4**A, **4**B, and **4**C illustrate some electrical configurations, those of ordinary skill in the art readily recognize a variety of other configurations that may serve the same function.

[0033] Referencing FIG. **4**A, Direct Current In (DC IN) **402** is buffered by a gang of capacitors **404** before being communicated to a DC/AC converter **406**. The DC/AC converter converts the direct current into alternating current suitable for placement over an existing electrical grid **408** such as normally found from a power-plant. Those of ordinary skill in the art readily recognize a variety of DC/AC converters that may work in this capacity.

[0034] FIG. **4**B illustrates an electrical arrangement suitable for use in charging a battery. DC IN **402** is buffered by capacitor bank **404** before entering into a step down transformer **408**. Step down transformer **408** reduces the voltage so that the voltage can safely be introduced into battery **410**, which is connected to ground **412** at the battery's other pole. Those of ordinary skill in the art readily recognize a variety of batteries that may work in this capacity.

[0035] In FIG. **4**C, DC IN **402** is fed into an adjustable rheostat **414**, which is controlled by the controller so that the DC OUT **416** falls within a specified range.

[0036] FIG. **5** illustrates an adjusting apparatus **500** that enables adjustment of height of the electricity collecting windsock arrangement **100** in accordance with one implementation. As per one implementation, the height of the electricity collecting windsock **102** can be adjusted using a winch motor **502** and a spool **504**. The winch motor **502** and the spool **504** can release or withdraw the tether **112** to adjust the height of the electricity collecting windsock **102** that is connected to the tether **112**. In one implementation, this release and the withdrawal may be performed manually. In another implementation, this release and the withdrawal may be performed automatically based on information obtained from a sensor system that measures atmospheric electrical energy being collected per unit time. If the sensor indicates that a current flow is diminishing, then the tether **112** is released/extended from the spool **504** to increase the altitude of the electricity collecting windsock **102** such that more static charge from the atmosphere is gathered. If the sensor indicates that collected atmospheric electrical energy exceeds a preset level per unit time, the tether **112** is withdrawn onto the spool **504** to decrease the static charge being collected from the atmosphere.

[0037] FIG. **6** is a process flow diagram illustrating aspects of a method consistent with implementations of the current subject matter. At step **602**, the functioning of the electricity collecting windsock arrangement **100** depends on whether atmospheric conditions are windy. If the atmospheric conditions are windy, the wind **116** stretches/extends outward the electricity collecting windsock **102**—Step **604**. Next, it is determined whether the atmospheric electrical energy (atmospheric current) that is collected is in a predetermined range—Step **606**. If the collected atmospheric electrical energy is in the predetermined range, the flow goes back to step **604**. If the current is not in the predetermined range, the height of the electricity collecting windsock **102** may be adjusted using adjusting apparatus **500**, such as the winch motor **502** and spool **504**. One skilled in the art understands that other implementations may include any other adjusting apparatus **500**, such as a pulley, a wheel mechanism, or the like.

[0038] The implementations set forth in the foregoing description do not represent all implementations consistent with the subject matter described herein. Instead, they are merely some examples consistent with aspects related to the described subject matter. Although a few variations have been described in detail herein, other modifications or additions are possible. In particular, further features and/or variations can be provided in addition to those set forth herein. For example, the implementations described above can be directed to various combinations and sub-combinations of the disclosed features and/or combinations and sub-combinations of one or more features further to those disclosed herein. In addition, the logic flows depicted in the accompanying figures and/or described herein do not necessarily require the particular order shown, or sequential order, to achieve desirable results. The scope of the following claims may include other implementations or embodiments.

What is claimed is:

1. An electricity collection apparatus comprising:

a tether; and

a windsock formed of an insulation material with a metal deposited on at least a portion of the insulation material, the metal of the windsock being electrically connected to the tether, the windsock to extend in a direction of ambient wind to provide a surface area based on the ambient wind, the surface area providing the metal to collect electrical energy from the ambient wind, the collected electrical energy being transferred to the tether.

2. The apparatus in accordance with claim **1**, wherein the metal is deposited on at least one of opposite sides of the insulation material.

3. The apparatus in accordance with claim **2**, wherein the metal is a metal selected from a group of metals comprising gold, silver, copper and aluminum.

4. The apparatus in accordance with claim **1**, wherein the insulation material is a polyester film.

5. The apparatus in accordance with claim **1**, wherein the transferred electrical energy at the tether is further transferred from the tether to an electricity storage device.

6. The apparatus in accordance with claim **1**, wherein the metal comprises conductors that are light-weight thereby allowing payload requirements for adjusting the windsock to an optimum location to be within a predetermined value, the optimum location comprising a height of the windsock from ground level.

7. The apparatus in accordance with claim **6**, wherein the deposited metal conductors form a pattern that maximizes the collection of the electrical energy at the optimum location.

8. An electricity collection apparatus comprising:

a windsock formed of an insulation material having a metal deposited on at least a portion of the insulation material, the windsock to extend in a direction of ambient wind to provide a surface area based on the ambient wind, the surface area providing the metal to collect electrical energy from the ambient wind; and

an electrically conductive tether connected with windsock to anchor the windsock in the direction of the ambient wind, the electrically conductive tether being electrically connected to the metal of the insulation material, the collected electrical energy being transferred to the tether.

9. The apparatus in accordance with claim **8**, wherein the metal is deposited on at least one of opposite sides of the insulation material.

10. The apparatus in accordance with claim **9**, wherein the metal is a metal selected from a group of metals comprising gold, silver, copper and aluminum.

11. The apparatus in accordance with claim **8**, wherein the insulation material is a polyester film.

12. The apparatus in accordance with claim **8**, wherein the transferred electrical energy at the tether is further transferred from the tether to an electricity storage device.

13. The apparatus in accordance with claim **8**, wherein the metal comprises conductors that are light-weight thereby allowing payload requirements for adjusting the windsock to an optimum location to be within a predetermined value, the optimum location comprising a height of the windsock from ground level.

14. The apparatus in accordance with claim **13**, wherein the metal conductors form a pattern that maximizes the collection of the electrical energy at the optimum location.

15. An electricity collection apparatus comprising:

a windsock formed of an insulation material the windsock to extend in a direction of ambient wind to provide a surface area based on the ambient wind;

one or more electrical conductors provided on at least a portion of the insulation material to collect electrical energy from the ambient wind; and

an electrically conductive tether connected with windsock to anchor the windsock in the direction of the ambient wind, the electrically conductive tether being electrically connected to the one or more electrical conductors provided on at least a portion of the insulation material to transfer the collected electrical energy from the one or more electrical conductors to an electrical storage.

* * * * *

Sept. 7, 1965 L. G. SMITH 3,205,381

IONOSPHERIC BATTERY

Filed March 9, 1962

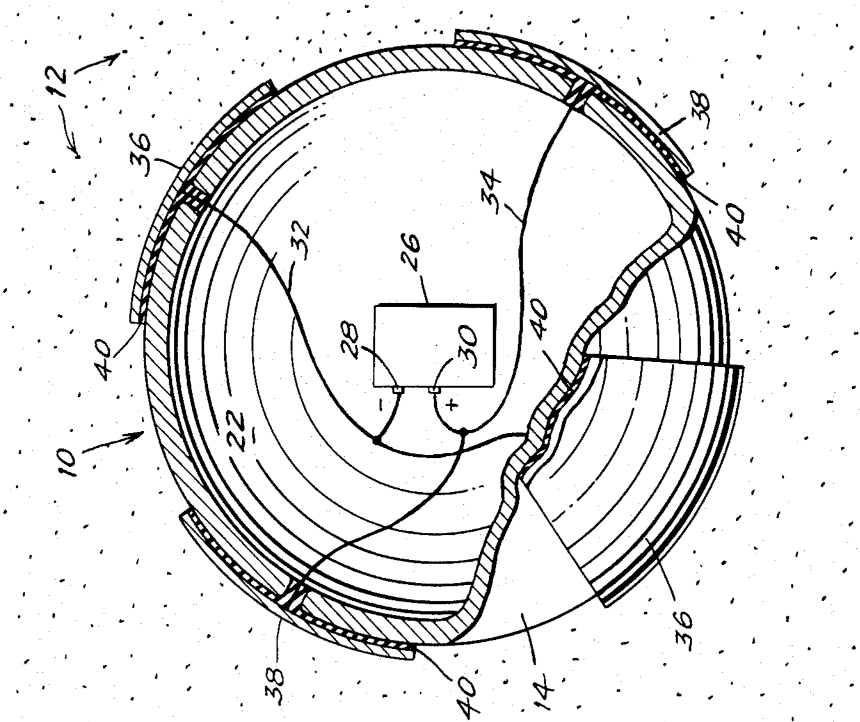
12
36
38
34
32
26
28
30
40
40
40
22
10
38
40
14
36
-
+

Fig. 2

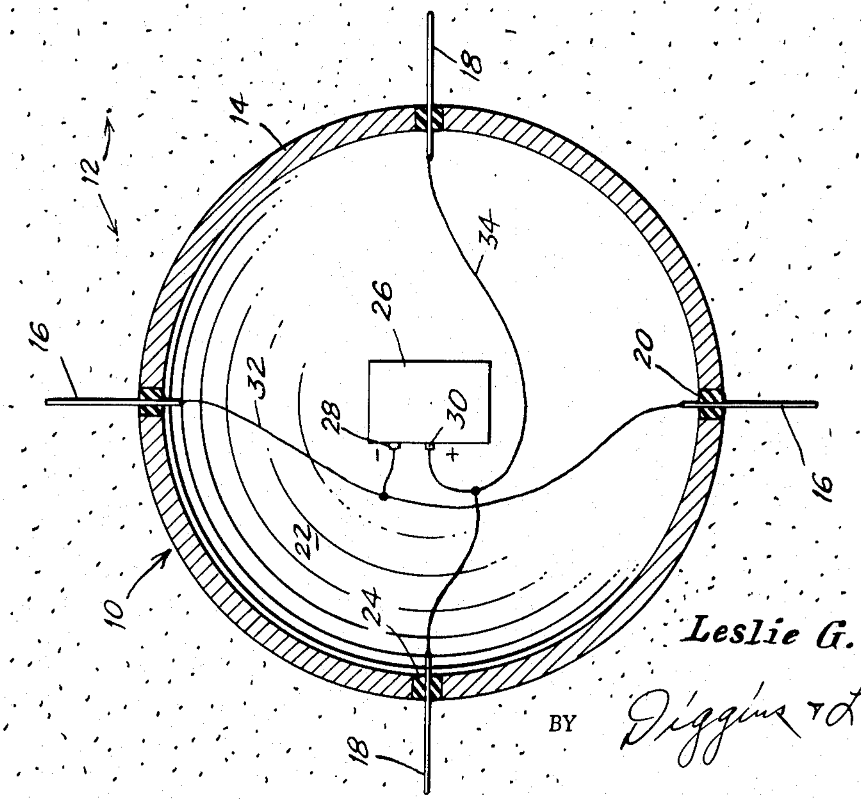
18
14
12
34
26
32
28
30
16
20
16
10
22
24
18
-
+

Fig. 1

INVENTOR

Leslie G. Smith,

BY Diggins & Le Blanc

ATTORNEYS

United States Patent Office

3,205,381

Patented Sept. 7, 1965

3,205,381

IONOSPHERIC BATTERY

Leslie G. Smith, Waltham, Mass., assignor, by mesne assignments, to the United States of America as represented by the Administrator of the National Aeronautics and Space Administration

Filed Mar. 9, 1962, Ser. No. 178,721

11 Claims. (Cl. 310—5)

This invention relates to a method and apparatus for tapping the enormous reservoir of energy existing in the ionosphere and more particularly for furnishing electrical power to space vehicles.

Within recent years various space probes and space vehicles such as satellites have been launched by man in an effort to learn more about the earth and the solar system within which he lives. Each of these various space craft has employed numerous electrical circuits which are utilized to detect and transmit back to earth certain information ascertained by the craft as it moves through space. In view of the presence of such electrical circuits, some source of electrical power must be made available whereby these circuits can be energized. In the past there have been three main sources of electrical power carried by space vehicles for energizing these circuits, namely, (1) conventional storage batteries, (2) solar batteries and (3) a combination of the conventional storage and solar batteries. Although these various electrical power sources have functioned with some degree of success each has certain serious limitations which cannot be readily overcome.

In the first instance, the conventional storage battery may be of either the wet or dry type but, due to its superior storage capacity, the wet cell type is used nearly exclusively in space vehicles. This additional storage capacity of the wet cell battery is of great importance since the useful life of the vehicle is usually dependent upon the proper operation of its electrical circuits which are, in turn, directly dependent upon the capacity and life of the available power source. Use in space vehicles of a storage battery that utilizes a liquid electrolyte presents several problems that have not been completely solved, however, nor is there any apparent solution to some of these problems. For example, to increase the output capacity of the battery and thereby extend the useful life of the satellite a greater amount of electrolyte must be carried by the battery. But to increase the quantity of electrolyte carried by the battery results in both an increase in size and weight of the power source and space vehicle which cannot, in most instances, be tolerated. Furthermore, the electrolyte may not function correctly if it is subjected to the extreme conditions of space such as, for example, the changes in temperature that are normally encountered. Thus, to insure the proper operation of the wet cell battery involved in powering a space vehicle and to prevent any leakage that might occur should the battery be damaged, some shielding or insulating of the battery is usually involved which, needless to say, adds materially to the cost and weight of the vehicle.

To overcome the space, weight and leakage problems encountered when storage batteries of the conventional wet cell type are used, space engineers have turned to the recently developed solar cell or battery which converts the light rays received from the sun directly into electrical energy. While these solar batteries do not require any type of electroylte and are, therefore, relatively light in comparison to the wet cell storage battery, they are expensive and their inherent principle of operation presents one problem which cannot be overcome, namely, the inability of the solar cell to produce power while the space vehicle is shaded from the sun such as occurs when a satellite is in the shadow cast by the earth.

In an effort to overcome this loss of power which occurs each time a satellite passes into the shadow cast by the earth or some other heavenly body, a storage battery of the wet cell type was added to the vehicle. While the satellite is in the sunlight the solar battery powers the various electrical circuits and charges the storage battery so that, once the sunlight is no longer available, the storage battery can be switched on to furnish the necessary electrical power to the vehicle. While this combined system functions satisfactorily, there is added to the already existing problems of weight, size and electrolyte leakage the problem of incorporating a battery charging and switch-over system between the solar and storage batteries. The addition of this type of system to the power source is undesirable since it adds unduly to the complexity, sensitivity and cost of the space vehicle.

According to the present invention, it has been found that some of the foregoing disadvantages of space vehicle power sources can be overcome by employing an ionospheric type battery which utilizes the ionosphere or other ionized medium existing in outer space as a supplemental source of energy. Since power is a scarce commodity essential for the operation of the space vehicle and for the proper functioning of its experiments, any means for supplying additional power while adding only slightly to the weight of the vehicle, increases the efficiency and life of the space vehicle. Although it has been known for some time that the ionosphere contains an enormous reservoir of energy in the form of excited and ionized atoms and molecules, no practical method has heretofore been found whereby this energy could be utilized as a source of power for earth satellites even though much thought has been given to this problem. The operation of the present invention is based upon the discovery that the ionosphere can be utilized as a part of supplemental power source for the load circuit in a space vehicle when said vehicle is equipped with electrodes having different work functions which, in turn, are electrically connected through the load circuit.

Accordingly, a primary object of this invention is to provide a method of producing electrical power from ionized media existing in nature.

Another object of this invention is to provide a simple, highly reliable supplemental power source for space vehicles.

Yet another object of this invention is to provide an ionospheric battery for use by satellites.

Still another object of this invention is to provide an ionospheric battery which utilizes the ionosphere as an electrolyte.

A further object of this invention is to provide a supplemental power source for space vehicles that will operate when shielded from the sun.

Yet a further object of this invention is to provide a power source for a satellite that is light and inexpensive to produce.

Another object of this invention is to provide a satellite with various shaped electrodes constructed of metals having different contact potentials.

These and further objects and advantages of this invention will become more apparent upon reference to the following description, claims and the appended drawings wherein:

FIGURE 1 is a cross sectional view of an orbiting satellite equipped with a power source constructed in accordance with one embodiment of the present invention; and

FIGURE 2 is a view in partial section of an orbiting satellite equipped with a power source constructed in accordance with a second embodiment of the present invention.

With continuing reference to the accompanying drawings wherein like reference numerals designate similar parts throughout the various views, and with initial attention directed to the embodiment of FIGURE 1, reference numeral **10** is used to generally designate a space satellite which is orbiting or otherwise passing through an ionized layer or medium such as, for example, the ionosphere which surrounds the earth. As is well known, the ionosphere contains an enormous number of excited and ionized atoms and molecules which are generated or produced when the gasses surrounding the earth are subjected to bombardment by high energy particles and rays such as, for example, ultra-violet or cosmic rays that originate throughout the solar system. Although these various ionized particles created by this bombardment are invisible to the eye they are represented pictorially by the various dots **12** for purposes of clarity.

The satellite **10** consists of an envelope or outer skin surface **14** which may be constructed of metal, ceramic or any other like material which is suitable for space vehicles of this type. As seen in FIGURE 1, a plurality of electrically conductive electrodes **16–18**, which are shown in the form of rods and are constructed of materials having different contact potentials (work functions), are mounted in a spaced relationship in the medium **12** through apertures **20** in the envelope **14** of the satellite. In order to prevent a leak from occurring between the normally pressurized interior **22** of the satellite and the ionosphere, which is a near vacuum, and to insulate the electrodes in case a metal is used for the envelope **14**, vacuum tight insulator feed-through elements **24** are fitted around the electrodes **16–18** and are securely mounted within the aperture **20**. These feed-through elements may be of any suitable type so long as they are constructed to withstand the changes in environmental conditions encountered in space without losing their insulative and sealing qualities.

Carried within the cavity or interior **22** of the satellite **10** is a load circuit **26** which may be, for example, a transmitter or any other electrical device that requires a source of electrical power for its proper operation. The load circuit **26** is mounted to the interior wall of the satellite by suitable means (not shown) and has its positive and negative power input terminals **28–30** connected to the various electrodes **16–18** respectively, by suitable electrical conductors **32–34**.

The overall principle and operation of the ionospheric battery in relationship to the satellite **10** can be explained substantially in the following manner. While the satellite remains on earth prior to its being hurled into or through an ionized medium, the electrodes **16–18** are at substantially the same potential since very few ionized particles are present. Thus, no current will flow through the load circuit **26**. However, once the satelllite is launched and it enters an ionized medium a potential difference will be established between the electrodes **16–18** and an electrical current will begin to flow through the load circuit **26**. This difference in electrical potential that exists between the electrodes **16–18** while the satellite is in an ionized medium is attributed to the fact that the electrodes **16** are constructed of a material having a different contact potential or work function from that of the electrodes **18**. This difference in contact potential between the electrodes **16–18** will result in the free electrons of the ionized medium being collected on one group of electrodes, for example the electrodes **16**, while the positive ions produced when the gas is ionized will be collected on the other electrodes **18**. With the electrodes **16** charged to a negative potential in relationship to electrodes **18**, a predetermined amount of current will flow through the load circuit **26** and can thus be utilized to power an electrical circuit such as the transmitter mentioned above.

The exact material from which the two types of electrodes **16** and **18** are constructed will depend upon several factors such as, for example, the amount of output power desired, the material readily available for constructing the electrodes, the susceptibility of the material to damage or destruction when exposed to the environmental conditions existing in space, and various other considerations that will be readily apparent to a worker in the battery field.

While the electrodes **16–18** used in the ionospheric battery of FIGURE 1 have been described as rods or whip-like elements it is to be realized that the electrodes may take other shapes. By way of example, a plurality of flat, plate-like electrodes **36–38** are shown mounted on a satellite in FIGURE 2. These electrodes are constructed of materials having different contact potentials and are insulated from the outer surface **14** of the satellite when this is necessary by an insulative layer **40** of any suitable material having the desired insulative properties under the severe conditions existing in space. The use of these plate electrodes **36–38** has the advantage of increasing the surface area exposed to the ionized medium thus permitting a larger output current to be produced without the necessity of using excessively long electrodes as might be the case if the rod electrodes of FIGURE 1 were used.

The output current of the battery is dependent upon the degree to which the medium is ionized and to the extent that the electrode surface area is exposed to the medium. An increase in either results in an increase in the efficiency of the battery. From an analysis of various experiments it is expected that a current of about two microamps per cm.^{2} of electrode surface at a potential of one volt would be produced if the satellite is orbiting in the ionosphere at an altitude of approximately 350 km. and electrodes of gold and platinum are used.

It will also be readily apparent that the envelope of the satellite itself could be constructed in sectional form and of materials having different contact potential thus eliminating the necessity of attaching different electrodes to the satellite or the slight increase in weight necessarily involved. Of course the various sections of material making up the envelope of the satellite would have to be insulated from one another to prevent electrical shorting from occurring.

It will be apparent from the foregoing that the power source of this invention is extremely light, rugged and involves a minimum of cost to construct. By using the vast natural reservoir of energy existing in the form of ionized gases for producing the electrical power necessary to energize the various electrical circuits of a space vehicle a practically indefinite source of electrical power is made available to the vehicle. An ionospheric battery constructed in accordance with this invention is also much more reliable in operation than other heretofore known power sources since no liquid electrolyte is involved which can leak, deteriorate or otherwise become spent, nor is the operation of the battery dependent upon its position in relationship to a source of light.

The invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. The present embodiments are therefore to be considered in all respect as illustrative and not restrictive, the scope of the invention being indicated by the appended claims rather than by the foregoing description, and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced therein.

What is claimed and desired to be secured by United States Letters Patent is:

1. A space vehicle equipped with a generator of electrical energy operable in the ionosphere as a supplemental source of potential for a load thereof, said generator comprising at least two electrode means on said vehicle and positioned to contact said ionosphere when said vehicle is in said ionosphere, said two electrode means each having different contact potentials, one of said two electrode means being gold and the other being

platinum, and electrical connector means connecting said two electrode means to said load.

2. Apparatus for generating electrical energy from the ionosphere comprising, a pair of spaced electrode means having different contact potentials positioned to contact said ionosphere so that a direct-current electric potential is established between said electrode means, load circuit means, and means electrically connecting said load circuit means to said electrode means whereby said load circuit means is powered by the direct-current electrical potential existing between said electrodes.

3. Apparatus for generating electrical energy from the ionosphere comprising, carrier means, a first electrode having a first contact potential mounted on said carrier means, a second electrode having a second contact potential spaced from said first electrode and mounted on said carrier means, said first and second electrodes being positioned on said carrier means so that they will be in contact with said ionosphere when said apparatus is placed in the ionosphere so that a direct-current electric potential is established between said first and second electrodes, and an electrical load circuit connected between said first and second electrodes, said load circuit being powered by the direct-current potential produced by said first and second electrodes.

4. Apparatus according to claim 3 wherein said first and second electrodes extend away from said carrier means.

5. Apparatus according to claim 3 wherein said first and second electrodes form a portion of said carrier means.

6. In a space vehicle adapted to pass through ionized areas in space, a power source for said vehicle while it is passing through said ionized areas comprising, a plurality of electrode means located on the outer surface of said vehicle, said electrode means being constructed of materials having different contact potentials, and means connecting said electrode means together through a load circuit whereby said circuit receives a source of direct-current as said space vehicle passes through said ionized areas.

7. In a space vehicle according to claim 6 wherein said electrode means are rod shaped.

8. In a space vehicle according to claim 6 wherein said electrode means are plate shaped.

9. A satellite comprising an envelope, electrical load means carried in said envelope, apertures extending through said envelope at spaced points, electrode means having different contact potentials, said electrode means passing through said apertures, means in said aperture for insulating said electrode means from said envelope, and electrical conductor means connected between said electrode means and said electrical load means whereby said electrical load means is energized by the direct-current electric potential established when said electrode means passes through an ionized medium.

10. A space satellite comprising an envelope, an electrical load circuit carried in said envelope, a power supply for said electrical load circuit comprising a first electrode having a first contact potential connected to said electrical load circuit and extending through said envelope and a second electrode having a second contact potential connected to said electrical load circuit and extending through said envelope, said first and second electrodes being insulated from one another so that said electrical load circuit is powered by the direct-current electric potential established by said first and second electrodes coming in contact with an ionized medium.

11. The method of powering electrical circuit means carried by a space vehicle comprising the steps of positioning first and second electrode means having different work functions on said vehicle, passing said vehicle through the ionosphere so that said first and second electrode means are in contact therewith and produce a direct-current potential between them, and connecting said first and second electrodes to said electrical circuit means by means of electrical connector means, whereby said direct-current potential is made available for operating said circuit.

References Cited by the Examiner

UNITED STATES PATENTS

414,943	11/89	Dewey	310—5 X
911,260	2/09	Pennock	310—5 X
2,303,321	12/42	Bennett	175—264
2,445,305	7/48	Hochgesaug	250—83.6
2,510,397	6/50	Hansell	171—330
2,802,113	8/57	Ohmart	250—83.6
2,864,012	12/58	Thomas	310—3
2,900,535	8/59	Thomas	310—3
3,021,472	2/62	Hernqvist	322—2

LLOYD McCOLLUM, *Primary Examiner.*

JOHN H. MACK, *Examiner.*

FIP8102 XR 3,273,066

X5264R

Sept. 13, 1966 L. H. RUHNKE 3,273,066

APPARATUS FOR DETECTING CHANGES IN THE ATMOSPHERIC ELECTRIC FIELD

Filed Dec. 20, 1963 2 Sheets-Sheet 1

Fig. 1

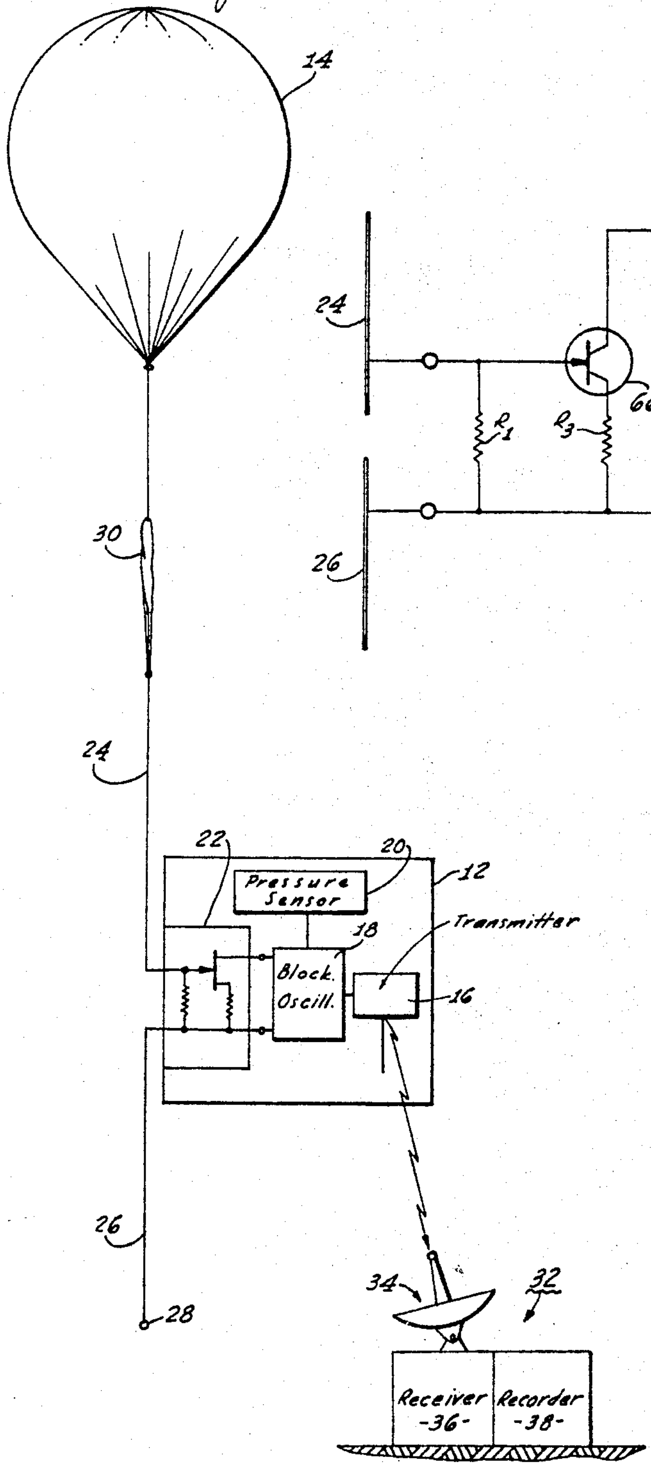
14
30
24
22
20
Pressure Sensor
12
18
Transmitter
Block. Oscill.
16
26
28
34
32
Receiver -36-
Recorder -38-

Fig. 3

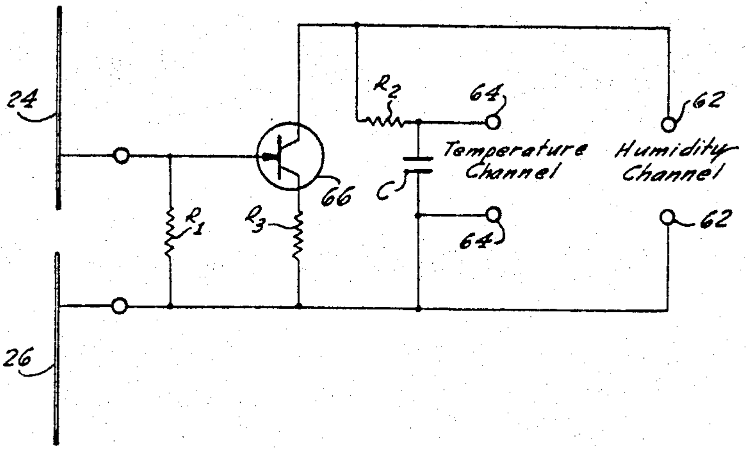
24
R2
64
62
Temperature Channel
Humidity Channel
C
66
R1
R3
64
62
26

INVENTOR:
Lothar H. Ruhnke

By Alan C. Rose
Attorney

Sept. 13, 1966

L. H. RUHNKE

3,273,066

APPARATUS FOR DETECTING CHANGES IN
THE ATMOSPHERIC ELECTRIC FIELD

Filed Dec. 20, 1963

2 Sheets-Sheet 2

Fig. 2

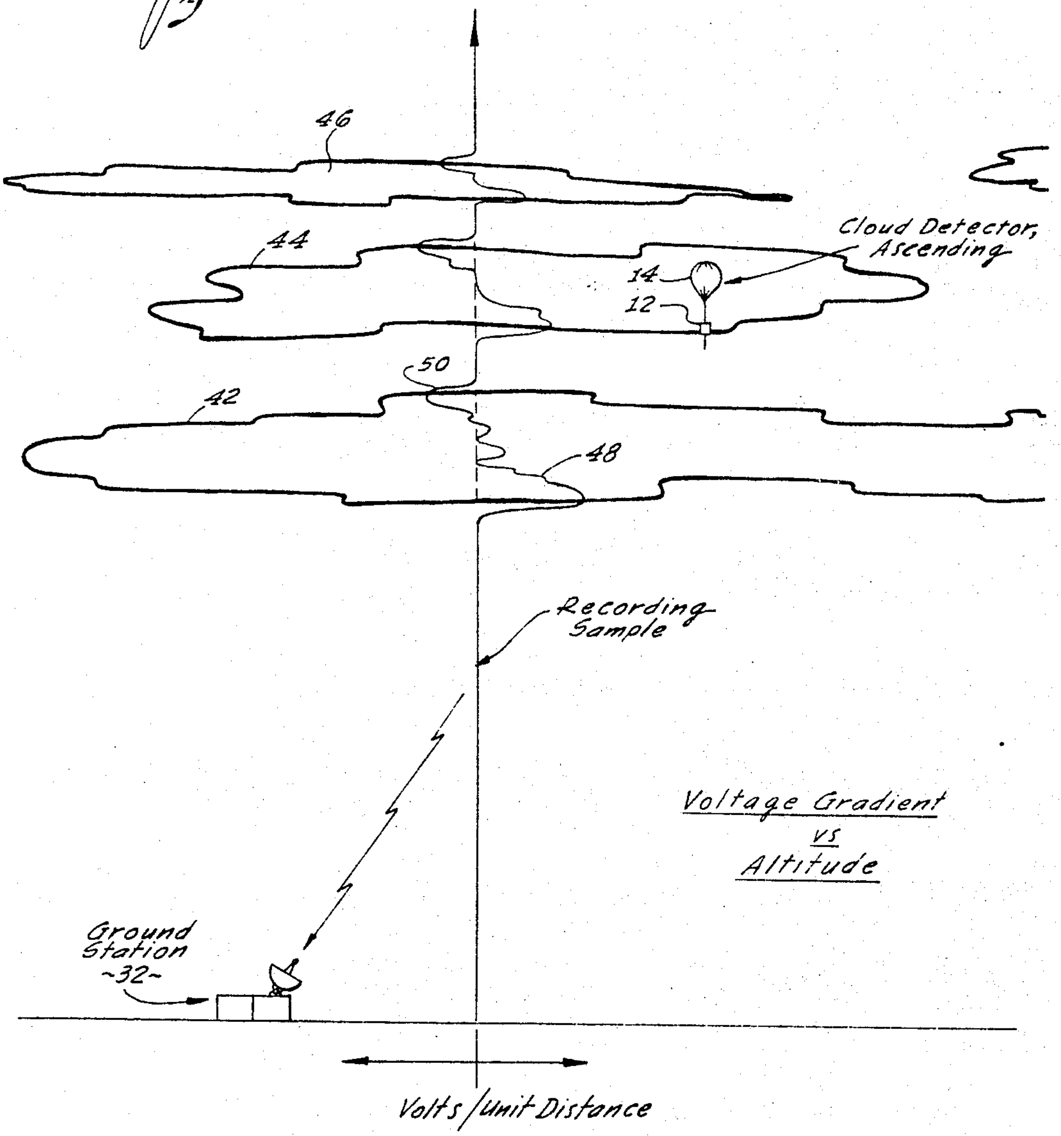
46
44
Cloud Detector, Ascending
14
12
50
42
48
Recording Sample
Voltage Gradient vs Altitude
Ground Station ~32~
Volts/Unit Distance

INVENTOR:
Lothar H. Ruhnke

By Alan C. Rose
Attorney

3,273,066
APPARATUS FOR DETECTING CHANGES IN THE ATMOSPHERIC ELECTRIC FIELD

Lothar H. Ruhnke, Minneapolis, Minn., assignor to Litton Systems, Inc., Beverly Hills, Calif.

Filed Dec. 20, 1963, Ser. No. 332,186
4 Claims. (Cl. 325—113)

This invention relates to a cloud detector or an apparatus for sensing electrostatic conditions in the earth's atmosphere.

Cloud detection apparatus which is presently employed is usually of the "ceilometer" type which merely measures the height of the lower surface of the clouds to indicate the ceiling conditions. One object of the present invention is to determine the location of both the top and the bottom of one or more cloud formations of the atmosphere.

While electrostatic sensors of one type or another have been proposed heretofore for use in the atmosphere, they have been so complex and expensive that they have not enjoyed widespread use. In addition, cloud radars have recently been developed, but their usefulness is limited by failures when different zones of the atmosphere contain droplets of different sizes, as discussed more fully below. Accordingly, it is another object of the present invention to simplify and reduce the cost of cloud detectors, while maintaining reliability, compatibility with existing equipment, and adequate sensitivity for all practical purposes.

In accordance with an illustrative embodiment of the present invention, a conventional U.S. Weather Bureau or U.S. Army meteoroligical radiosonde transmitter may be provided with a modulator which is controlled by a field effect transistor. The input to the field effect transistor is a dipole antenna. Each of the arms of the dipole antenna may be five to fifty feet in length, for example, and one of the dipole elements may trail the radiosonde package while the other element may extend forward in the direction of movement of the unit. When a balloon is employed to carry the cloud detector unit aloft, a weight may be attached to one of the antenna elements so it hangs below the instrument package, while the other antenna element can extend along, or even form the flexible connection between the radiosonde unit and the balloon. A receiving unit is provided to pick up the modulated signals from the radiosonde apparatus, demodulate these signals and record them. While it is contemplated that the receiver may normally be situated on the ground and receive signals from a rising radiosonde unit, the radiosonde unit could well be employed by an aircraft to determine conditions below a cloud layer. Under these conditions the receiver would be mounted on the plane and the radiosonde unit could be parachuted to the ground.

In accordance with a feature of the present invention, therefore, a system for detecting electrostatic field conditions in the atmosphere includes a transmitter, arrangements for moving the transmitter through the atmosphere, an antenna extending a substantial distance from the transmitter, and a field effect transistor connected to receive electrostatic input signals from said antenna and to modulate the output of the transmitter in accordance with voltages detected by the antenna.

In accordance with one aspect of the present invention, simplicity of circuitry and stability of operation are achieved by the use of dynamic circuitry which measures the instantaneous difference in potential between two antenna elements. Thus, a high input resistor may be connected between the input antenna elements. With this type of arrangement the instrument package must be continuously moved through the atmosphere in order to give meaningful results. Under these conditions, the recorded output indicates changes in electric fields. A constant electric field, corresponding to a linear change in potential produces no substantial output signal. It has been determined that a balloon velocity of about five meters per second will produce good results with a dipole antenna shunted by a 1,000 megohm resistor and connected to a field effect transistor as discussed in detail in the present specification. With lower velocities, a higher input impedance or greater amplification would be required, with higher velocities, such as rocket propulsion or the like, permitting even lower input impedances and simpler input circuits.

Accordingly, it is a feature of the invention that apparatus is provided for moving a radiosonde unit through the atmosphere at a predetermined rate, that two element input antenna preferably of the dipole type is provided, and that the antenna is connected to the radiosonde by simple dynamic amplification circuitry, which requires continuous movement of the apparatus to give useful output information.

The novel features which are believed to be characteristic of the invention both as to its organization and method of operation, together with further objects and advantages thereof, will be better understood from the following description considered in conjunction with the accompanying drawings in which a typical embodiment of the invention is illustrated by way of example. It is to be expressly understood, however, that the drawings are for the purpose of illustration and description only and are not intended as a definition of the limits of the invention.

In the drawing:

FIG. 1 is a schematic block diagram of a radiosonde equipment for sensing changes in the electric field of the atmosphere, and thus detecting clouds, in accordance with the present invention;

FIG. 2 is a diagrammatic showing of the record obtained under typical cloud conditions; and

FIG. 3 is a circuit diagram indicating a typical arrangement of connections of a dipole antenna, a field effect transistor, and the input terminals of a standard meteorological radiosonde apparatus.

With reference to the drawings, FIG. 1 shows a radiosonde package 12 supported by a balloon 14. The radiosonde unit includes a transmitter 16, blocking oscillator 18, pressure sensor 20, and a field effect transistor unit 22. A dipole antenna having an upper element 24 and a lower element 26 is connected to the input of the field effect transistor unit 22. A small weight 28 is secured to the lower end of the antenna element 26 to hold it in the extended position. Similarly, the upper antenna element 24 extends in the other direction, and it may either form the support for the radiosonde unit 12 or may be closely associated with the supporting cable. A parachute 30 is provided for safe return of the instrument package to ground.

A ground or receiving station 32 may include an antenna 34, receiver and demodulator 36, and a recording apparatus 38.

FIG. 2 shows a typical recording of the electrostatic field obtained as the ballon 14 and instrument package 12 rise through the successive layers of clouds 42, 44, and 46, along the center line of the drawing. It is known that the electrostatic potential in the atmosphere increases from ground potential to plus 300,000 volts at elevations of about 30 miles. It is also known that the atmospheric electric field, which is the gradient of the potential, is substantially increased within clouds. As the package enters the lower edge of a cloud, the sudden increase in electric field causes a pulse of current to flow through resistor R_1 in one direction, and upon leaving the cloud, the sudden reduction in electric field causes an oppositely directed current pulse. This produces the positive swing shown at **48** in FIG. 2 as the unit enters the cloud and the negative peak **50** as the apparatus is rising through the upper surface of the cloud stratus **42**. A similar pattern is observed as the unit goes through the higher cloud formations **44** and **46**. With a knowledge of the speed of ascent of the balloon and instrument package, the location of the upper and lower surfaces of successive cloud layers may be determined by the present apparatus. Accordingly, despite the presence of a relatively low cloud layer which completely obscures other cloud formations, ground personnel may readily determine the meterological conditions relating to clouds above this low ceiling.

Returning to FIG. 1, the radiosonde equipment represented by blocks **16**, **18** and **20** is well known. The meterological radiosonde system which will be taken as typical is designated by the military by the numbers AN/AMT 4B. The Army technical manual which describes this radiosonde system is identified as TM 11–2432A–TO 31M4–2AMT–11, dated June 26, 1958. The radiosonde transmitter operates at a frequency of 1680 megacycles per second. This transmitter is modulated by a blocking oscillator **18**, the frequency of the blocking oscillator being controlled by the meteorological conditions which are detected. Normally, a thermistor for measuring the temperature and a humidity sensor are alternately switched to the input terminals of the blocking oscillator **18**. In the present case, the cloud detector circuitry is connected to both of these terminals. The barometric pressure sensor **20** may also be retained in the unit, and it may be connected to control the frequency of the blocking oscillator **18** by a switching action as described below.

With reference to FIG. 3, the humidity channel input terminals **62** and the temperature channel input terminals **64** of the standard radiosonde unit are shown at the right-hand side of the circuit. Atmospheric electric signals received by the dipole antenna elements **24** and **26** are applied to the field effect transistor **66** and are then coupled by a suitable circuit to the input terminals **62** and **64** of the radiosonde. While other field effect transistors may be employed, type C 610 of Crystallonics, Inc., was found to be suitable from a sensitivity and reliability standpoint for the present purposes. The input to the field effect transistor is shunted by a 1,000 megohm resistor R_1 to stabilize the input characteristic. Suitable biasing of the transistor **66** is provided by resistor R_3. Because the impedance of the transistor **66** is substantially lower than that of the temperature and humidity sensors which are normally employed, a 110,000 ohm resistor in the blocking oscillator was removed and a capacitor C was connected in parallel to the input terminals. The capacitor was selected to produce a modulation frequency of approximately 100 cycles per second in the absence of an input signal. Positive and negative voltages will then raise or lower this frequency.

In the radiosonde apparatus, signals from the pressure sensor **20** are normally employed to switch from the temperature channel terminals **64** to the humidity channel terminals **62**. This switch is well known in the art as a baroswitch. The barometric pressure switching action provides an indication of altitude, so it is useful to preserve this information. Accordingly, the resistor R_2 is employed between the transistor **66** and one terminal **64** to reduce the output signal slightly, thus changing the modulation level and preserving the pressure and thus altitude information.

With regard to the length of the dipole elements **24** and **26**, the size of resistor R_1 and the amplification of the field effect transistor, these are fixed by several factors. First, the antenna elements should be large enough to provide a substantial input signal, but should be relatively small as compared with normal cloud thicknesses. Lengths of 5 to 50 feet for each dipole would be of the proper order of magnitude. The resistor R_1 was chosen to reduce the input impedance between dipoles **24** and **26** below the point where fog and dampness and the resultant shunt impedance will introduce instability into the transmitted signals. With these parameters, and a speed of ascent of the order of 5 meters per second, it was determined that a single field effect transistor provided the necessary amplification for operation with the standard radiosonde unit.

As mentioned generally above, cloud radars are subject to failure under certain atmospheric conditions. Thus, the radar return is proportional to the sixth power of the droplet or particle size, while the human eye and optical systems respond to the second power of the droplet size. With the radar return proportional to such a high power of the droplet diameter, if a radar is adjusted to pick up a cloud having large particles, other clouds having small size particles are not detected. The present electrostatic system, however, responds in a manner which is nearly the same as the optical response, and therefore does not fail under atmospheric conditions where different size droplets or particles are present in successive clouds.

It is to be understood that the above described arrangements are illustrative of the application of the principles of the invention. Numerous other arrangements within the scope of the invention may be devised by those skilled in the art. Thus, by way of example and not of limitation, a somewhat different shape of antenna could be employed to pick up electrostatic field variations and apply them to the field effect transistor, an inherently stabilized transistor may be used, and the instrument package may be carried by rocket or other suitable propulsion arrangements than the balloon shown in the drawings. Accordingly, it is evident that various changes may be made in the present invention without departing from the spirit of the invention as defined in the present claims.

What is claimed is:

1. An apparatus for sensing electrostatic conditions in the atmosphere comprising:

a radiosonde apparatus including means for modulating the output frequency thereof, said means including an R-C circuit;

a pair of probes positioned in respectively different locations in the atmosphere with respect to said radiosonde;

a field effect transistor having its drain and gate terminals connected across said pair of probes and having its source and drain terminals connected in series with the resistance of said R-C circuit to modulate said radiosonde; and

a stabilizing high resistance connected across the two probes and across said gate and drain terminals of said transistor.

2. In an apparatus for sensing atmospheric electrostatic field conditions;

a balloon;

an instrument package connected to and spaced from said balloon, said instrument package including a transmitter, means including an R-C circuit for modulating said transmitter, and a field effect transistor having source and drain terminals connected in series with the resistance in said R-C circuit to modulate said transmitter;

antenna means including an antenna and resistance means connected across said antenna, said resistance means also being connected across the gate and the drain of said field effect transistor, said antenna extending upward from said package toward said balloon and downwardly from said package for applying signals to said field effect transistor in response to said field conditions to render said modulating means responsive to said field conditions.

3. An apparatus for sensing changes of the atmospheric electric field at cloud boundaries, which comprises:

a blocking oscillator-type modulated radiosonde including fixed resistance means and capacitance means connected in parallel with said fixed resistance means, said capacitance means being effective to discharge through said resistance means to control the modulation of said radiosonde;

dipole antenna means responsive to changes of the atmospheric electric field for producing a signal indicative of said changes;

resistance means connected across said dipole antenna means; and

a field effect transistor having source and drain terminals connected in series with said fixed resistance means and a gate terminal connected to said dipole antenna means for reception of said signal, said field effect transistor being responsive to said signal for varying the rate of discharge of said capacitance means to cause the blocking oscillator to modulate said radiosonde in accordance with the second changes in the atmospheric electric field.

4. An apparatus for sensing changes of the atmospheric electric field at cloud boundaries, which comprises:

radiosonde means including a blocking oscillator having fixed resistance means and capacitance means connected in parallel with said fixed resistance means, said capacitance means being effective to discharge through said resistance means to control the frequency of oscillation of said blocking oscillator;

dipole antenna means responsive to changes of the atmospheric electric field for producing a signal indicative of said changes;

resistance means connected across said dipole antenna; and

a field effect transistor having source and drain terminals connected in series with said fixed resistance means and a gate terminal connected to said dipole antenna means for reception of said signal, said field effect transistor being responsive to said signal for varying the rate of discharge of said capacitance means to control the frequency of oscillation of said blocking oscillator in accordance with the sensed changes in the atmospheric electric field.

References Cited by the Examiner

UNITED STATES PATENTS

2,463,527	3/1949	Dunmore	325—113
3,038,154	6/1962	Zworykin et al.	325—113 XR
3,056,888	10/1962	Atalla	307—88
3,062,043	11/1962	Marsh et al.	340—193.2

DAVID G. REDINBAUGH, *Primary Examiner.*

JOHN W. CALDWELL, *Examiner.*

United States Patent [19]

Gundlach

[11] **Patent Number: 4,595,852**

[45] **Date of Patent: Jun. 17, 1986**

[54] **ELECTROSTATIC GENERATOR**

[76] Inventor: **Robert W. Gundlach,** 2434 Turk Hill Rd., Victor, N.Y. 14564

[21] Appl. No.: **599,199**

[22] Filed: **Apr. 11, 1984**

[51] **Int. Cl.**^{4} **H02N 1/00**

[52] **U.S. Cl.** **310/309**

[58] **Field of Search** 310/308–310

[56] **References Cited**

U.S. PATENT DOCUMENTS

2,610,994	9/1952	Bosch et al.	310/309
2,644,903	7/1953	Lehrer	310/307
2,739,248	3/1956	Meier	310/309
2,756,352	7/1956	Felici et al.	310/308 X

Primary Examiner—Donovan F. Duggan

[57] **ABSTRACT**

A novel induction-type electrostatic generator is disclosed in which moving electrodes are transported back and forth between two Faraday cages. Each moving electrode, once enclosed by a cage, is contacted to the cage, then separated by more than the sparking distance before making electrical connection to the opposite moving electrode or to ground. In this way, charge opposite in polarity but proportional in magnitude to the charge on each cage, is induced in each moving electrode and carried into and imparted onto the next cage, progressively elevating the cage potentials until the corona or arcing limits are reached. Initiation of charging is spontaneous, deriving from inevitable statistical imbalances in nature.

5 Claims, 4 Drawing Figures

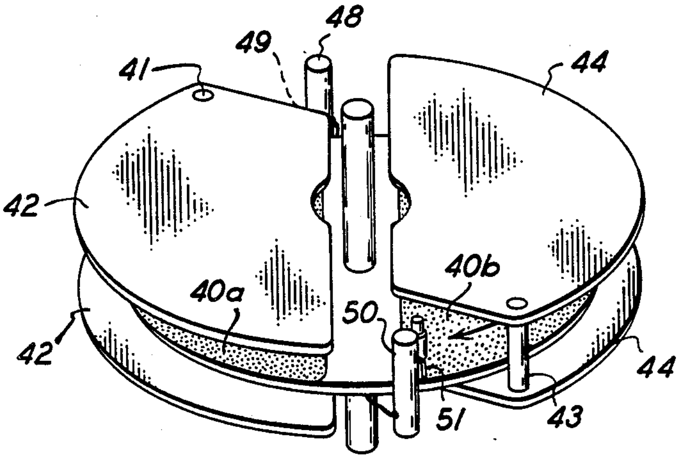

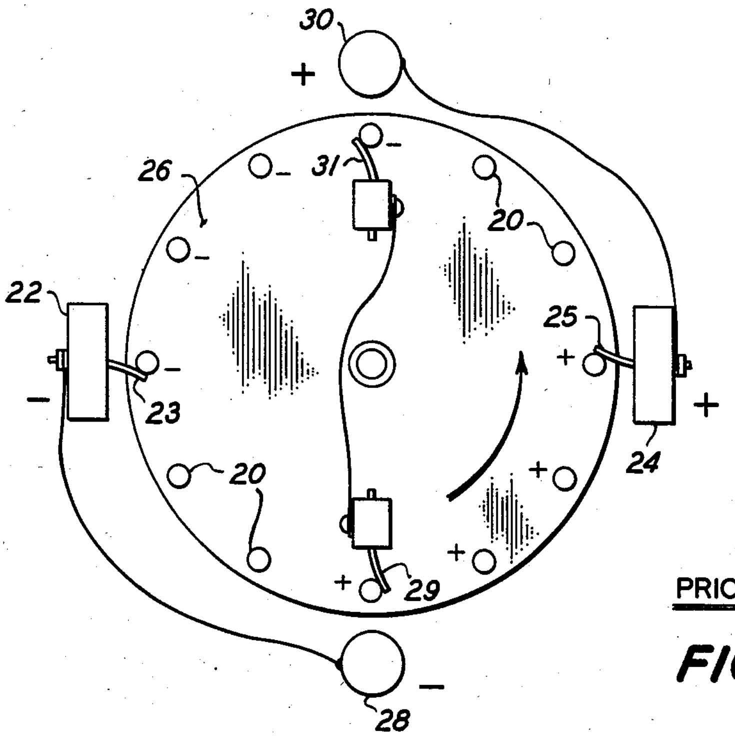
30
+
26
31
20
22
25
+
-
23
24
20
29
28
-

PRIOR ART

FIG. 1

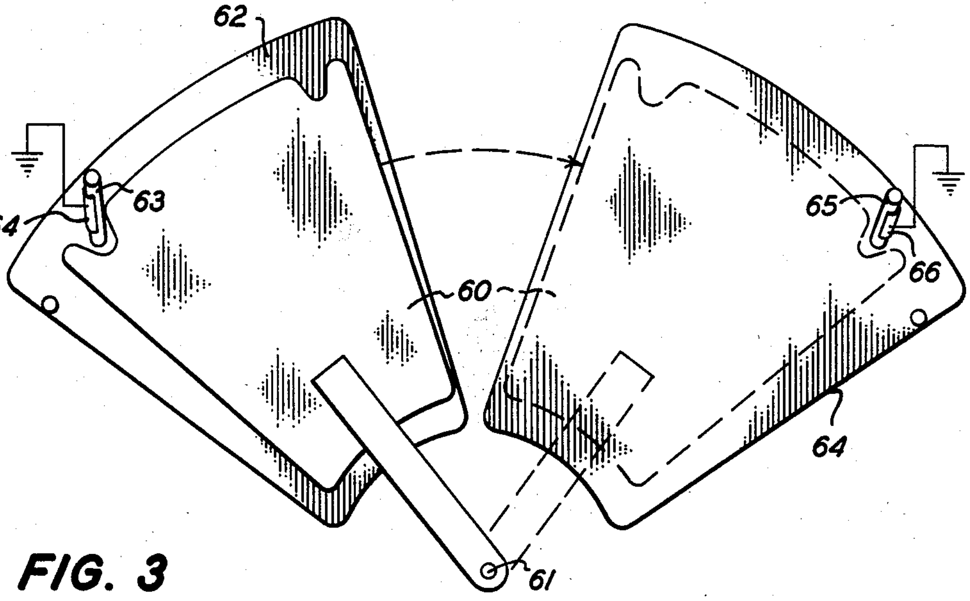
62
63
64
65
66
60
64
61

FIG. 3

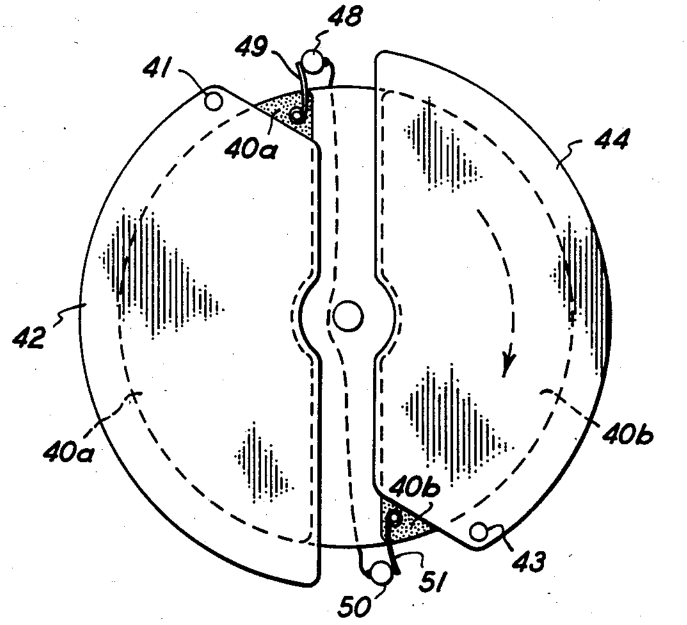

FIG. 2a

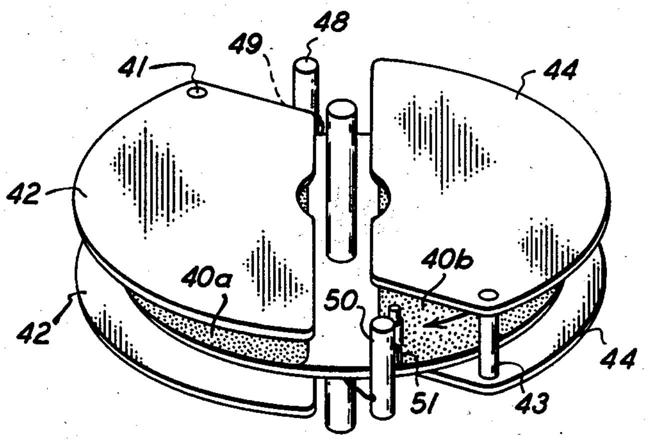

FIG. 2b

ELECTROSTATIC GENERATOR

BACKGROUND OF THE INVENTION

This invention relates to electrostatic generators, and more particularly to an improved induction-type electrostatic generator that is efficient, compact, and inexpensive, in which electrodes are transported back and forth between two Faraday cages.

The Greeks were the first to record observations of static electricity several thousand years ago. They discovered that amber, when rubbed vigorously with fur or fabrics under dry conditions, takes on electrostatic charge. In fact, the Greek word for amber, "electron", has become the root of many of our words relating to electricity.

It was not until about 1660 that the first electrostatic generator was invented. Otto von Guericke of Magdeburg, Germany showed electrostatic sparks could be produced by spinning and rubbing a dry sulfur globe. In this device, high potentials were developed by friction, or triboelectric charging. More than 200 years passed before James Wimshurst invented the first reliable induction-type generator, called the "influence machine" at the time (1878). This involved two insulating discs sliding against each other as they rotated in opposite directions on a common axis. Attached to these oppositely rotating plates were metal sectors into which charges are induced and carried to a capacitor for storage. The metal sectors continue inducing opposite charges in each other and carrying the charges to the appropriate ends of the capacitor, gradually building up the potentials of the capacitor until sparking voltages are reached. Wimshurst built many such machines, some with oppositely rotating discs more than two meters in diameter. The only substantial application of the Wimshurst generator appears to have been to operate the early X-ray tubes.

Then, in 1937, Robert Van de Graaff invented a belt-type generator in which an insulating belt, sometimes carrying conducting elements, transported charges into a giant conducting sphere, or "Faraday cage", to generate, in some cases, over a million volts. Van de Graaff charged the moving belt by spraying corona charge onto it or, alternatively, by frictionally charging it.

In the 1960's Prof. A. D. Moore designed an induction-type generator for instructional purposes. This is the simplest of the prior art generators to understand, and works purely by induction, requiring neither corona nor frictional charging means to produce the potential differences. FIG. **1** is a schematic diagram of Moore's "Dirod" generator, which he described in his book "Electrostatics", (Doubleday, Garden City, NJ, 1968; a Doubleday Anchor Book). In this generator, rods $\mathbf{20}_{(1\rightarrow N)}$ are rotated between two collector electrodes, **22** and **24**. Each rod makes momentary electrical contact to collectors **22** and **24** through flexible conducting brushes **23** and **25**, respectively, as insulating disk **26** is rotated counterclockwise about its center. Just above and below rotating disc **26** are supported conducting rods **28** and **30**, called "inductors", electrically connected to collector electrodes **22** and **24**, respectively. At the point of nearest approach to each inductor, each of the rods $\mathbf{20}_n$ is momentarily contacted, one at a time, to ground or to the opposite rod $\mathbf{20}_{(n+N/2)}$ through flexible conducting brushes **29** and**31**, respectively. Any incipient charge or potential difference between collector electrodes **22** and **24** will induce opposite charges in those rods $\mathbf{20}_n$ and $\mathbf{20}_{(n+N/2)}$ adjacent the inductors **28** and **30**. For example, if we assume that, of the many billions of electrons in each of the collector electrodes, there is a slight excess in collector **22**, then collector **24** will be slightly positive relative to collector **22**. Inductor **28** will therefore induce positive charges in each rod $\mathbf{20}_n$ as it is grounded in its proximity to inductor **28**. At the same time inductor **30** induces negative charges in each rod $\mathbf{20}_{(n+N/2)}$ when grounded next to inductor **30**. As the charged rods separate from inductor rods **30** and **28**, their potentials increase so that part of the induced charges flow off to the collector plates **22** and **24**, increasing their negative and positive potentials, respectively. As the potentials of the collector electrodes increase, so the charges induced in rods $\mathbf{20}_n$ by inductors **28** and **30** increase, quickly driving the collector electrodes to their maximum values. Once charging begins it will be seen how the potential difference grows at a progressively increasing rate. The statistical imbalance of nature insures that the initial imbalance can always be depended on. (To understand how the instability builds on itself, we need only understand why a sharpened pencil cannot be balanced long on its point; the restoring force is negative, and is proportional to the degree of imbalance.) Professor Moore's Dirod generator routinely produces 60 to 70 kV, depending on the spacing of the rods **20** from inductors **28** and **30**, since air breakdown (arcing) occurs at about 30 kV per cm for gaps larger than a few millimeters. This generator is not entirely satisfactory for manufacturing as an educational toy or for use for demonstrations because it is relatively complicated and expensive to produce and complex to understand.

SUMMARY OF THE INVENTION

Accordingly, the present invention provides a simplified electrostatic generator for an educational toy. Also, the present invention provides an electrostatic generator that is more efficient, more affordable, more rugged, and easier to understand for teaching purposes and experimental uses than prior art devices.

The foregoing and other features of the instant invention will be more apparent from a further reading of the specification and claims and from the drawings, in which:

FIG. **1** is a schematic plan view of a prior art generator.

FIG. *2a* is a schematic plan view of the preferred embodiment of the present invention.

FIG. *2b* is an isometric view of the preferred embodiment of the present invention.

FIG. **3** shows an alternative embodiment of the present invention that employs a single mobile electrode.

While the invention will be described hereinafter in connection with a preferred embodiment and an alternative embodiment, it will be understood that it is not intended to limit the invention to those embodiments. On the contrary, it is intended to cover all alternatives, modifications, and equivalents as may be included within the spirit and scope of the invention as defined by the appended claims.

DESCRIPTION OF THE PREFERRED EMBODIMENT

In the pesent invention the induction-type generator has been simplified and made much more efficient by combining each pair of collector and inductor electrodes into a single Faraday cage, as shown in FIGS. 2*a* and 2*b*. As Faraday taught, and Van de Graaff put to use, by enclosing a charged object in a conducting cage, the potential of the whole cage is increased as if all the charges were placed onto the cage, itself. In fact, if the enclosed charged object is conducting and makes conductive contact with the cage, all charges do, indeed, flow from the charged object onto the surrounding "Faraday cage".

In this example only two rotor electrodes **40***a* and **40***b* are shown. In fact, any even number can be used, within practical space limitations. An odd number could be used if external ground is provided during charge induction. With an even number of rotor electrodes, contacts **48** and **50** can be connected to each other, instead of to external ground, to induce electrons from one rotor electrode to the other. Rotor electrode **40***a* is transported by clockwise rotation from one cage **42**, to the other cage **44**, making sequential contact first with the surrounding, or enclosing case, through contact post **43**, then, after adequate separation from contact post **43**, making momentary contact with ground or the opposite rotor electrode **40***b* through contact posts **50** and **48**, whereby charges opposite to those on the enclosing cage are induced on each rotor plate and then transported to the opposite cage through brushes **49** and **51**. This results in a progressive build-up of opposite electrostatic potentials on collector cages **42** and **44**. Assuming thin, flat rotor plates **40**, the capacitance of each rotor plate **40** to the shell or cage **42** when the plate is totally enclosed by the cage is given by:

$$C = \frac{K\epsilon_0 A}{t}$$

where C is capacitance, K is the dielectric constant of the medium (K=1 for air), ϵ_o is the permittivity of space ($\epsilon_o = 9 \times 10^{-12}$), A is the area of the surface of the rotor plate, and t is the distance between the rotor plate **40** and the inner surface of the shell **42**. Each plate carries a charge q out of one shell and into the other shell every half rotation, equal to:

$$q = CV.$$

It is possible, then, to calculate the current that can be supplied by this generator for any specified values of rotor plate area A, distance t between the rotor plate **40** and the parallel surface of the cage **42**, and speed of rotation of the rotor plates.

$$i = \frac{dq}{dt} = 4\omega CV$$

where ω is the spin velocity in rotations per second. For example, an area of 100 cm^2 on each face of the rotor plate ($A = 2 \times 10^{-2} m^2$), and a plate-to-cage spacing of 1 cm ($t = 10^{-2}$m), and a spin velocity of ten rotations per second, gives a generator current of:

$$i = 40\ CV = 40 \times \left(\frac{9 \times 10^{-12} \times 2 \times 10^{-2}}{10^{-2}}\right) V = 7.2 \times 10^{-10}\ V \text{ amperes}$$

Clearly, then, current is very low for very low voltages, but builds exponentially with time, the initial voltage build-up can be expedited by placing a charged object on one of the cages. Once the maximum potential of, say, 20 kV is reached, the generator of this example of the present invention would produce a current of:

$$i = 1.4 \times 10^{-5} \text{ amperes, or about } 14\ \mu A$$

FIG. **3** shows another embodiment of the present invention. Here a single mobile electrode **60** is mounted on a hinge **61** and swings reciprocally between Faraday cages **62** and **64**. In this figure, the top plate of each cage **62** and **64** has been removed for clarity. Movable electrode **60** enters either cage **62** or **64**, makes conducting contact with said cage, separates from said conducting contact by more than the sparking distance (1 cm for every 30 kV), and then makes the electrical contact to ground while it is still enclosed by said cage. During grounding, charges opposite to the charge on the enclosing cage are induced in electrode **60**, and these are transported by movable electrode **60** into and deposited onto the opposite Faraday cage. Swinging contacts **63** and **65** prevent electrical contact between movable plate **60** and ground during entry of plate **60** to each cage, while they insure contact with conducting edges **64** and **66** during exit, while plate **60** is still fully enclosed within the associated Faraday cage. Since the charges induced on movable plate **60** are proportional to the charges on the enclosing cage, repeating the cyclic process produces progressively increasing potentials until corona or sparking limits are reached.

In summary, the advantages of the electrostatic generator of the present invention are its elegant simplicity for construction, for providing a simple understanding for the student, its ruggedness of design for portability without damage, affordability of cost, and its high efficiency derived by maximizing the rotor area and thus capacitance, for a given spacing, the spacing determining the inherent limitation of voltage imposed by dielectric breakdown of air or other fluid medium in which the generator is operated. If we wish to develop potentials of 30 kV operating in air, for example, the spacing of rotors **40** to cages **42** must be at least one cm, since the dielectric breakdown of air for such gaps is 30 kV/cm.

While the principles described herein have been described by the example of FIGS. 2*a*, 2*b*, and **3**, it should be understood that the concepts disclosed can be embodied in various mechanical designs without departing from the teachings of this invention and may come within the scope of the following claims.

What is claimed is:

1. A spontaneous non-frictional electrostatic induction-type generator, comprising:

(a) two Faraday cages:

(b) electrode means, substantially surrounded by and mounted for rotary movement relative to said Faraday cages in order to enable said Faraday cages to serve alternately as both the inductor and collector electrodes: and

(c) flexible contact means comprising a single set of brushes mounted to said rotary electrode means so as to extend by centrifugal force into electrical contact with said Faraday cages, then, after separating from each cage by more than the sparking distance, momentarily contacting each said flexible contact means to ground or to opposite ends of a conductor connecting said flexible contact means to each other to progressively charge said Faraday cages.

2. The spontaneous non-frictional electrostatic induction-type generator of claim **1**, wherein said electrode means comprises an even number of movable electrodes, and flexible contact means mounted for centrifugal extension to make momentary electrical contact separately to each cage; then, after separating from each cage by more than the sparking distance, momentarily connecting said movable elctrodes to each other.

3. The method of claim **1**, wherein an even number of rotary movable electrodes is provided, and means, mounted for extension by centrifugal force, is provided for contacting said rotary movable electrodes separately to each Faraday cage, then, after separating from each cage by more than the sparking distance, momentarily connecting said rotary movable electrodes to each other.

4. The spontaneous electrostatic induction-type generator of claim **1**, wherein said electrode means comprises a plurality of movable electrodes.

5. A method of generating high electrostatic potentials using a spontaneous, non-frictional electrostatic induction-type generator comprising the steps of:

(a) providing two Faraday cages;

(b) providing electrode means substantially surrounded by and mounted for rotary movement relative to said Faraday cages; and

(c) providing flexible contact means mounted to said rotary electrode means so as to extend by centrifugal force into electrical contact with each cage separately, then, after separating from each cage by more than the sparking distance, momentarily contacting to ground or to a neutral conductor, in order to progressively charge said Faraday cages.

* * * * *

US007855476B2

(12) **United States Patent**
Ogram

(10) **Patent No.: US 7,855,476 B2**
(45) **Date of Patent: Dec. 21, 2010**

(54) **ATMOSPHERIC ELECTRICAL GENERATOR**

(76) Inventor: **Mark Ellery Ogram**, 780 S. Freeman Rd., Tucson, AZ (US) 85748

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 254 days.

(21) Appl. No.: **12/218,297**

(22) Filed: **Jul. 14, 2008**

(65) **Prior Publication Data**

US 2010/0007218 A1 Jan. 14, 2010

(51) **Int. Cl.**
H02G 11/00 (2006.01)

(52) **U.S. Cl.** **307/145**; 307/149; 361/218; 361/230

(58) **Field of Classification Search** 307/149, 307/145; 244/30, 31; 174/2; 361/212, 215–218, 361/230, 231
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

674,427	A	*	5/1901	Palencsar	307/149
911,260	A	*	2/1909	Pennock	307/149
1,014,719	A	*	1/1912	Pennock	307/149
1,540,998	A	*	6/1925	Plauson	310/309
3,121,196	A	*	2/1964	Kasemir	324/464
4,234,167	A	*	11/1980	Lane	254/291
4,486,669	A	*	12/1984	Pugh	290/44
4,842,219	A	*	6/1989	Jakubowski et al.	244/31
5,203,542	A		4/1993	Coley et al.	
6,254,034	B1	*	7/2001	Carpenter	244/153 R
6,829,911	B2		12/2004	Davis et al.	
6,961,662	B2		11/2005	Murphy	
7,016,785	B2		3/2006	Makela et al.	
7,200,418	B2		4/2007	Kaikuranta et al.	
7,330,366	B2		2/2008	Lee et al.	
7,375,492	B2		5/2008	Calhoon et al.	
7,378,181	B2		5/2008	Skinlo	
7,388,350	B1		6/2008	Wright	
7,394,671	B2		7/2008	Fukumoto et al.	
7,397,220	B2		7/2008	Uchida et al.	

* cited by examiner

Primary Examiner—Fritz M Fleming
(74) *Attorney, Agent, or Firm*—Mark Ogram

(57) **ABSTRACT**

A mechanism to tap an electrical source which uses an aircraft (preferably a lighter than air balloon) tethered by a conductive line. The conductive line is extended/withdrawn by a winch motor to adjust the altitude of the aircraft. The conductive line is isolated from the ground and an electrical conductor is connected to the conductive line and to an electrical load. In this manner, static electricity generated in the atmosphere is gathered for use.

15 Claims, 4 Drawing Sheets

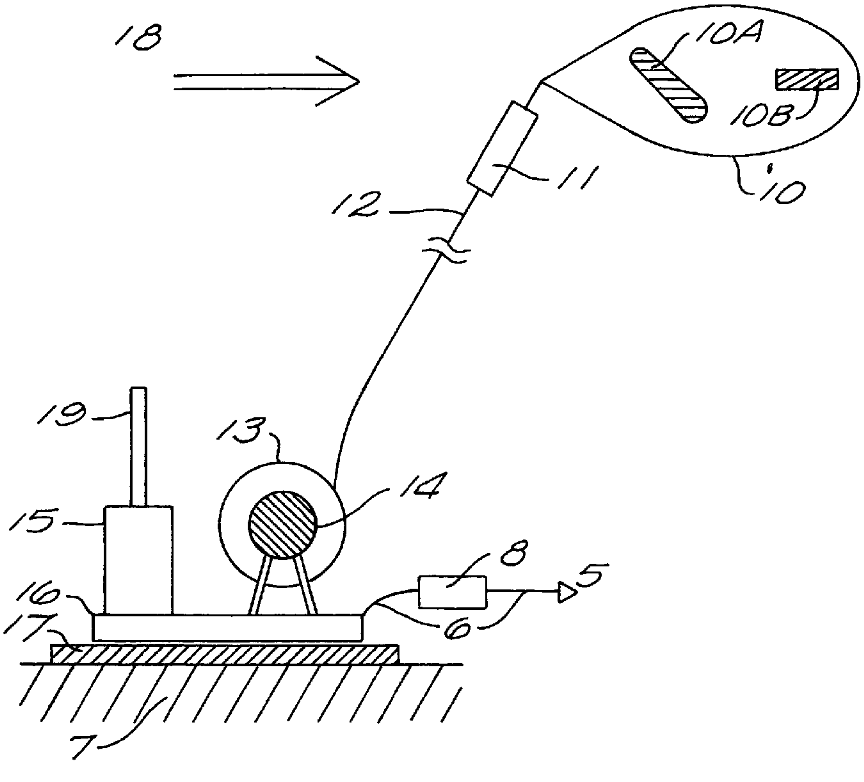

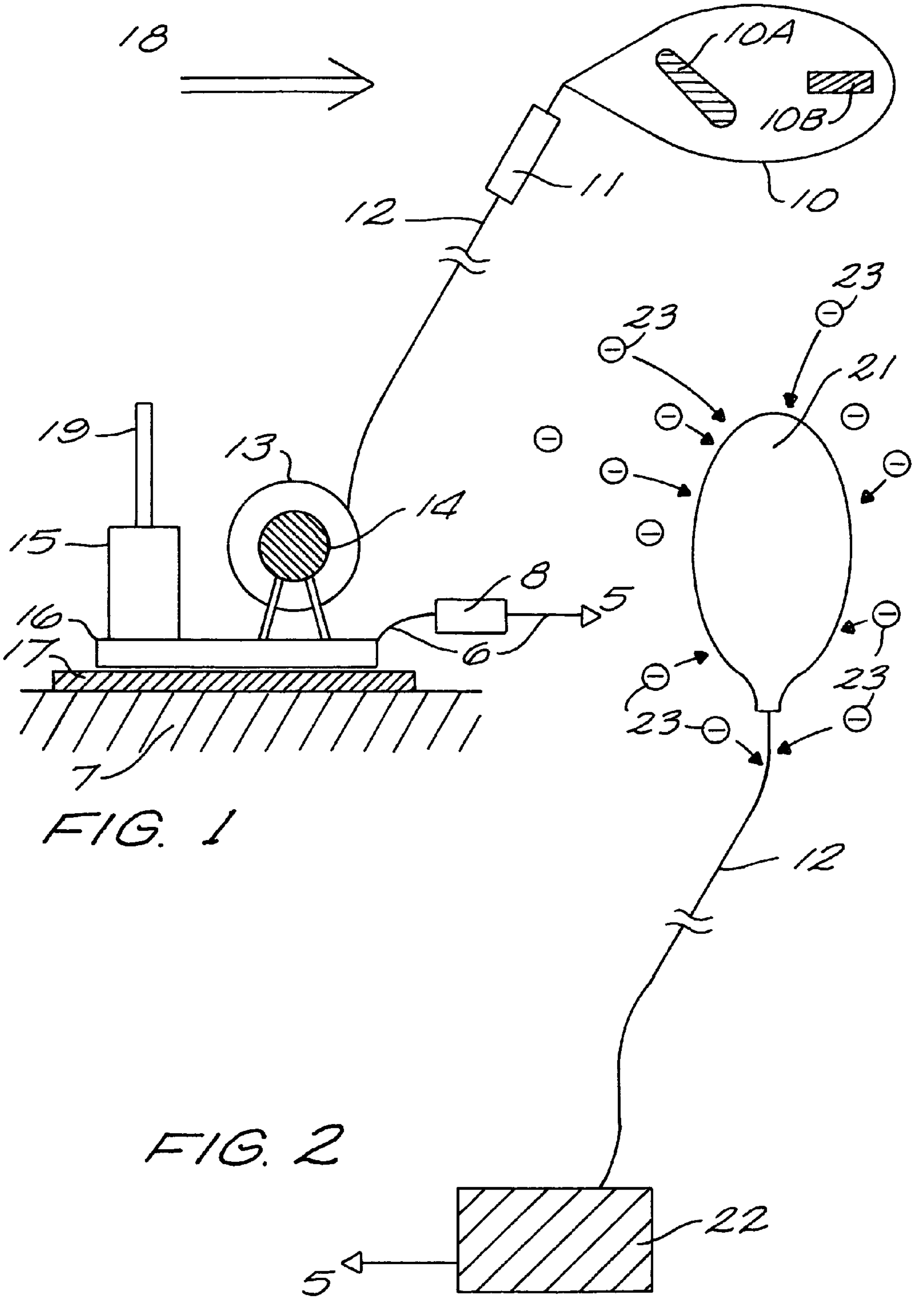
18
10A
10B
10
11
12
19
13
14
15
8
5
16
6
17
7
23
23
21
23
23
23
12
22
5

FIG. 1

FIG. 2

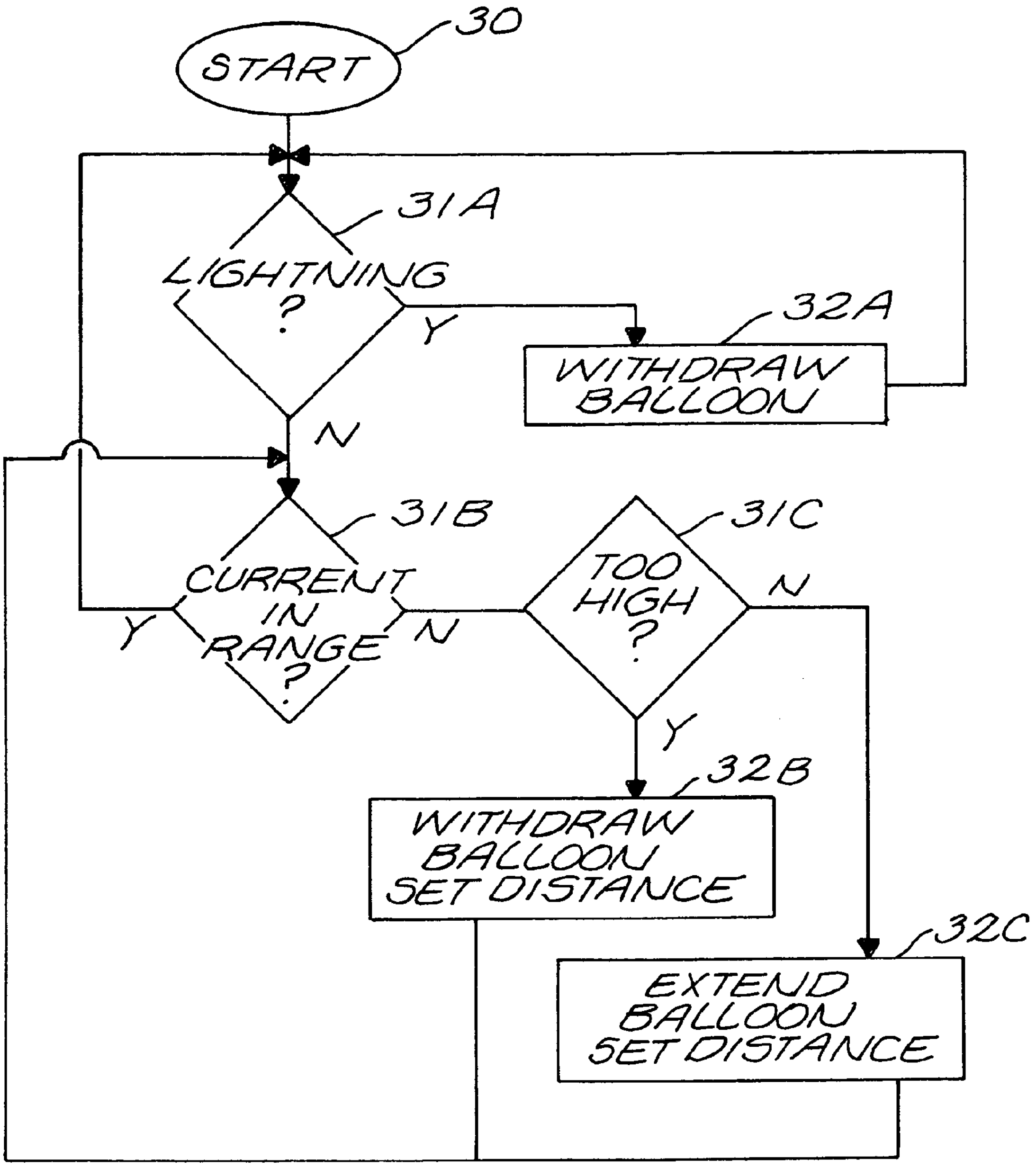
30
START
31A
LIGHTNING ?
Y
N
32A
WITHDRAW BALLOON
31B
CURRENT IN RANGE ?
Y
N
31C
TOO HIGH ?
N
Y
32B
WITHDRAW BALLOON SET DISTANCE
32C
EXTEND BALLOON SET DISTANCE

FIG. 3

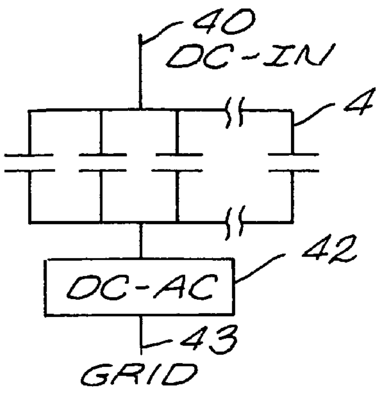
40
DC-IN
41
42
DC-AC
43
GRID

FIG. 4A

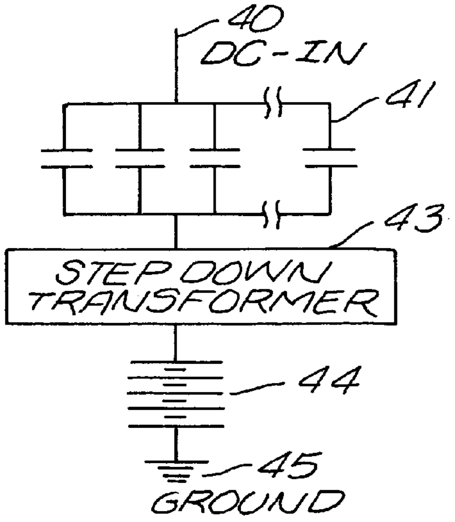
40
DC-IN
41
43
STEP DOWN TRANSFORMER
44
45
GROUND

FIG. 4B

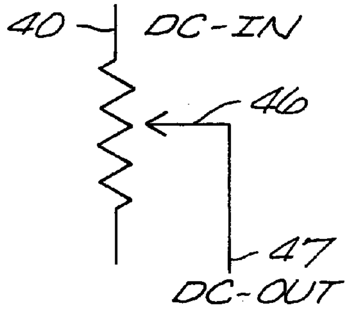
40
DC-IN
46
47
DC-OUT

FIG. 4C

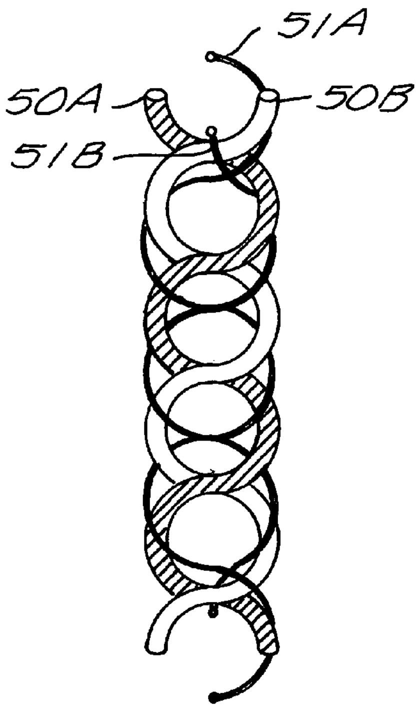
51A
50A
50B
51B

FIG. 5

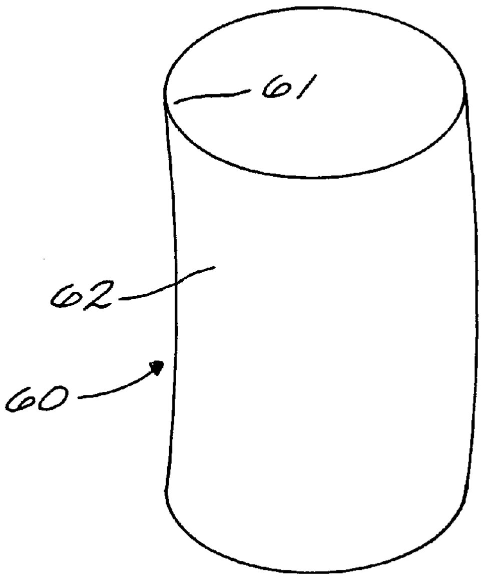

FIG. 6A

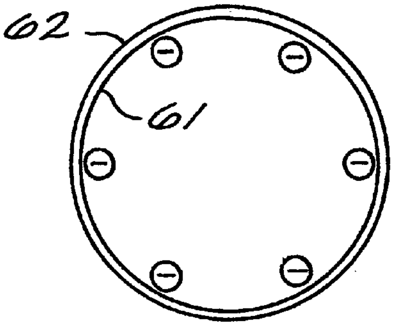

FIG. 6B
REPLACEMENT

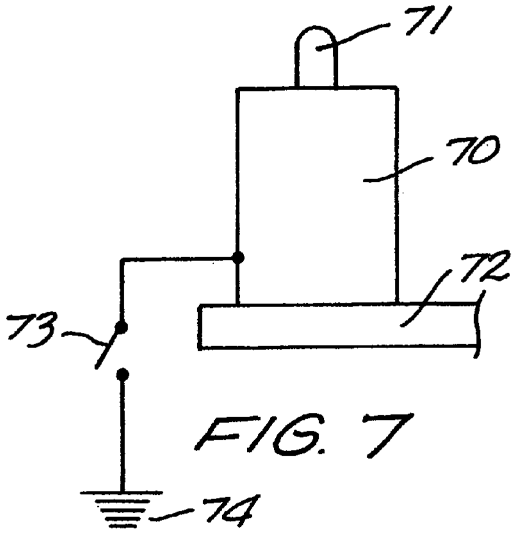
71
70
72
73
74

FIG. 7

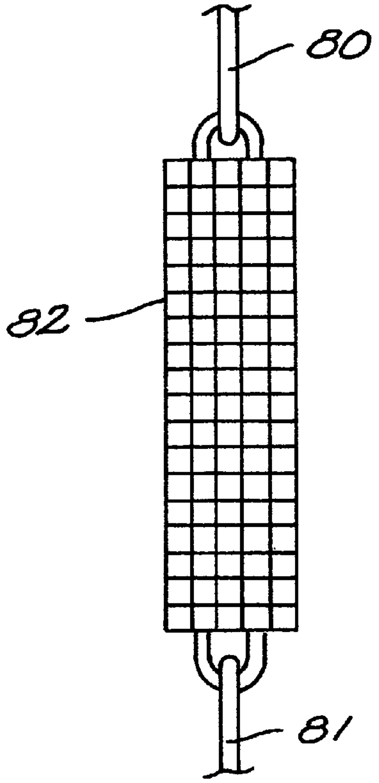

FIG. 8A

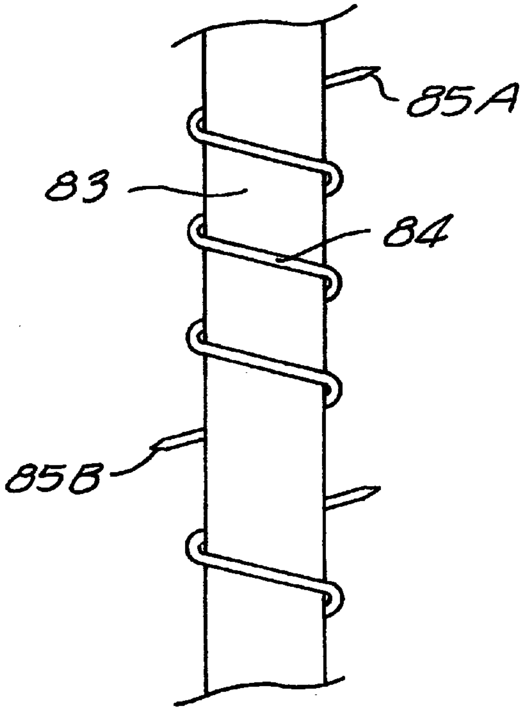
85A
83
84
85B

FIG. 8B

ATMOSPHERIC ELECTRICAL GENERATOR

BACKGROUND OF THE INVENTION

This invention relates to the production of electrical energy and more particularly to the production of electrical energy from the atmosphere.

Everyone is familiar with Benjamin Franklin's kite experiment of 1752. Using a kite whose string had become wet, negative charges from the passing clouds flowed into the string, down to the suspended key, and then into a Leyden jar via a thin metal wire. Franklin was protected by a dry silk string; but, when Franklin's knuckle came too close to the key, he received a strong shock. Fortunately, Benjamin Franklin was not killed, others who tried this same experiment were not so lucky.

Since then, the formation of lightning has remained something of a mystery. Lightning bolts are triggered when a negatively charged cloud base induces a positive charge from the ground, thereby forming a "pathway" for the discharge of the collected electrical energy.

Lightning travels up to 60,000 miles per hour with a flash that is brighter than ten million 100-watt lightbulbs. This wattage is as much power as is produced by all of the electricity plants in the United States and with a voltage of up to 300 million volts.

It is this very fact, the power within lightning is immense, that has prevented any successful collection of the electrical energy from lightning. The electricity in lightning is far too extreme for current technology to harness.

While lightning has attracted a energy starved industrial world, no one has developed any technique to harness this naturally occurring electrical source.

It is clear there is a continuing need for an electrical source other than carbon-based fuels and that the naturally occurring electricity in the atmosphere is being ignored.

SUMMARY OF THE INVENTION

The invention is a mechanism which taps into the naturally occurring static electricity in the atmosphere. Whereas heretofore, the attempt to garner electricity from the atmosphere has focused exclusively on capturing lightning, the present invention syphons off the static electricity which is generated from any agitated air and avoids lightning.

Lightning is only the final discharge of the static electricity, whether that lightning is intra-cloud lightning, cloud-to-ground lightning, or inter-cloud lightning. Other types of final discharges are known as heat lightning, summer lightning, sheet lightning, ribbon lightning, silent lightning, ball lightning, bead lightning, elves, jets, and sprites. Well before these discharges are observed, as the atmosphere becomes agitated by wind or thermal activity, static electricity is being generated.

The present invention recognizes that this static electricity is being formed and creates a mechanism to capture it.

The mechanism of this invention utilizes an aircraft such as a lighter than air balloon. While the preferred embodiment uses a foil balloon, a variety of other aircraft are obvious to those of ordinary skill in the art, including, but not limited to: gliders, rubber balloons (such as weather balloons), biaxially-oriented polyethylene terephthalate polyester film balloons, and latex balloons.

Within this discussion, the balloon is referenced, but, the invention is not intended to be limited solely to balloons.

The balloon is sent aloft and is tethered by a conductive line. In this context, the conductive line may be any obvious to those of ordinary skill in the art. For the preferred embodiment, the conductive line is a generically referred to as a "poly-rope" and is commercially available through a variety of sources. A suitable conductive line is described in U.S. Pat. No. 5,203,542, entitled "Apparatus for Improved Electric Fence Wire Construction for use with Intensive Grazing" issued Apr. 20, 1993, to Coley, et al. and incorporated hereinto by reference.

The conductive line is played out of a winch to control the altitude of the balloon. The motor controlling the winch is able to reverse direction to both extend and withdraw the conductive line which is wrapped around a spool on the winch. The winch/spool combination are part of a base unit.

In some embodiments of the invention, the spool is constructed of rubber so as to insulate the conductive line from the winch assembly. In this embodiment, only the conductive line is charged by the atmospheric static electricity while the winch remains neutral.

In yet another embodiment, the winch/spool are part of a base unit which is itself isolated from the ground by an insulator. In this embodiment, the entire base unit is charged by the atmospheric static electricity.

A conductor, such as an insulated wire, is electrically connected to the conductive line. In one embodiment, where the conductive line is electrically isolated from the spool and winch motor, the conductor is connected to the conductive line. In the embodiment where the conductive line is electrically connected to the base unit, then the conductor is connected anywhere on a metallic base unit.

The other end of the conductor is connected to a load. The load in this case can be any of a variety of electrical loads well known to those of ordinary skill in art, including, but not limited to a motor, a battery system, or the electrical grid for the system.

In the preferred embodiment, a sensor array is used to monitor the activities both at the base unit (such as electrical flow within the conductor) and in the surrounding locale.

A sensor monitoring the electrical flow (i.e. voltage and/or current) within the conductor is used to monitor the electrical activity within the conductor.

In the preferred embodiment, a lightning sensor monitors for lightning activity within the locale. As noted earlier, the electrical characteristic of lightning is so extreme that ideally this discharge is avoided as it might damage the mechanism of this invention.

The sensor array is utilized by a controller, such as microprocessor, programmed to operate the mechanism as outlined herein.

The controller operates the winch motor to extend or withdraw the conductive line and by extension the altitude of the balloon. The controller is programmed to operate the winch by monitoring the electrical characteristics of the conductor and adjusting the balloon's altitude to maintain these characteristics within the conductor within a preset range.

This preset range is established either in the base programming of the controller or is established by an operator of the system.

As example, by controlling the amount of current being withdrawn from the atmosphere, the mechanism operates within a safe range and also provides a relatively stable current flow from which a variety of activities can take place (such as DC-AC conversion).

The controller also utilizes the lightning sensor to protect the mechanism from a lightning strike. Should lightning be detected within a pre-determined range (as established by the software or defined by an operator), then the balloon is pulled down to minimize the risk of damage from a lightning strike.

The invention, together with various embodiments thereof will be more fully explained by the following description of the accompanying drawings.

DRAWINGS IN BRIEF

FIG. **1** diagrams the preferred embodiment of the invention.

FIG. **2** illustrates the collection of the negative charged particles in the atmosphere.

FIG. **3** is a flow-chart of the operation of the controller for the preferred embodiment of the invention.

FIGS. **4**A, **4**B, and **4**C are electrical schematics for handling the static charge from the atmosphere.

FIG. **5** illustrates a conductive line used in the preferred embodiment of the invention.

FIGS. **6**A and **6**B illustrate an alternative conductive line creating an ionized pathway for the flow of the static charges from the atmosphere.

FIG. **7** illustrates the controller of an alternative embodiment and the associated safety devices.

FIGS. **8**A and **8**B illustrate two embodiments of enhanced electrical collection leads.

DRAWINGS IN DETAIL

FIG. **1** diagrams the preferred embodiment of the invention.

Balloon **10** is an aircraft which, in this illustration, is a lighter than air balloon. Wings **10**A, extending from the body of balloon **10**, provide additional lift in air flow **18**. Tail **10**B helps to stabilize balloon **10**.

Balloon **10** is tethered to the ground via conductive line **12**. As noted earlier, a variety of configurations and materials are available to serve as conductive line **12**. In this illustration, a poly-wire is used Poly-wire is commercially available through a variety of vendors, including, but not limited to: Jeffers Livestock and Zareba Systems, Inc, of Ellendale, Minn.

In this embodiment, located proximate to balloon **10**, is an electrical collection enhancement lead **11** which assists in the collection of the static electrical charge in the atmosphere. Electrical collection enhancement lead **11** is configured to attract the static charge and conduct the charge into the conductive line **12**.

The electricity flows down the conductive line into spool **13**, where the conductive line **12** is collected and either withdrawn or dispensed through operation of winch motor **14**.

Winch motor **14** and spool **13** are mounted onto base unit **16** which is electrically isolated from ground **7** using insulator **17**. Note, in this embodiment of the invention, when electricity is being collected from the atmosphere, the entire base unit **16** becomes charged. In another embodiment of the invention, spool **13** is constructed of rubber, thereby preventing base unit **16** from becoming charged, thereby restricting the charging from the atmosphere to only conductive line **12**.

In this embodiment, conductor **6** is connected to base unit **16** (since the entire base unit **16** is charged and the base unit is metallic) to communicate the electrical current to load **5**. Conductor **6** is ideally an insulated wire.

The electrical current through conductor **6** is measured using sensor **8**.

In the alternative embodiment discussed above, where only the conductive line **12** is charged, then conductor **6** is connected to conductive line **12**.

Controller **15**, located in this embodiment on base unit **16**, operates winch motor **14** in response to signals from sensor **8** (measuring the current being discharged to load **5**) to maintain the current flow within a pre-defined range. As the current flow diminishes, then the conductive line **12** extended from spool **13** to increase the altitude of balloon **10** to that more static charge from the atmosphere is gathered; as the current flow falls exceeds a preset level, conductive line **12** is withdrawn onto spool **13** to decrease the static charge being collected from the atmosphere.

The range of current flow through conductor **6** is ideally set by the program, although some embodiments of the invention permit an operator to establish this range of operation.

In an alternative embodiment, the sensor monitoring conductor **6** monitors the voltage therein.

In the preferred embodiment of the invention, controller **15** is also equipped with a lightning sensor **19**. In this embodiment, when lightning is sensed within a preset range, then substantially all of conductive line **12** is wound onto spool **13** to pull balloon **10** near the ground and protect the entire mechanism from being damaged from a lightning discharge.

In the preferred embodiment, the "safe" distance form lightning is set in the programming of controller **15** and is ideally two miles; other embodiments permit the operator to "safe" distance.

There are a variety of lightning sensors well known to those of ordinary skill in the art, including, but not limited to those described in: U.S. Pat. No. 7,016,785, entitled "Lightning Detection" issued to Makela, et al. on Mar. 21, 2006; U.S. Pat. No. 6,829,911, entitled "Lightning Detection and Prediction Alarm Device" issued to Jones, et al. on Dec. 7, 2004; U.S. Pat. No. 7,200,418, entitled "Detection of Lightning" issued to Karikuranta, et al. on Apr. 3, 2007; and U.S. Pat. No. 6,961,662, entitled "Systems and Methods for Spectral Corrected Lightning Detection" issued to Murphy on Nov. 1, 2005; all of which are incorporated hereinto by reference.

In another embodiment of the invention, controller **15** is not located on base unit **16**, rather it is remote and communicates its control signals to winch motor **14** using radio waves.

FIG. **2** illustrates the collection of the negative charged particles in the atmosphere.

Static charges **23** are generated in the atmosphere by agitated air. These static charges are often collected at the bottom of clouds, but exist in other environments as well.

Balloon **21** is extended into this strata of static charges **23** which are then attracted to conductive line **12** to flow to base unit **22** and then onto load **5**.

By increasing or decreasing the altitude of balloon **21** (defined by the length of the extended conductive line **12**), conductive line **12** is selectively exposed to varying densities and levels of the static charge strata, and by extension, the current flow or voltage is increased or decreased.

FIG. **3** is a flow-chart of the operation of the controller for the preferred embodiment of the invention.

Once the program starts **30**, the lightning sensor is checked to determine if lightning has occurred within the unsafe range **31**A, if it has, then the balloon is lowered **32**A, and the program continues monitoring the status of lightning until no lightning is detected.

When the lightning status is acceptable, then the current within the conductor is checked to see if the current is within the prescribed range **31**B. If the current is acceptable (within range) the program returns to check the lightning status **31**A; otherwise a determination is made to see if the current is above the prescribed range **31**C.

If the current is above the prescribed range, then the altitude of the balloon is withdrawn a set amount **32**B (ideally twenty-five feet) and the program loops back to see if the current is within range **31**B.

If the current is below the prescribed range, then the altitude of the balloon is extended a set amount **32**C (ideally twenty-five feet) and the program loops back to see if the current is within range **31**B.

In this manner of feed-back and minor adjustments in the altitude of the balloon, the current is maintained within a prescribed range which can be handled by the downstream electrical system.

As noted earlier, some embodiments of the invention monitor the voltage instead of the current.

FIGS. **4**A, **4**B, and **4**C are electrical schematics for handling the static charge from the atmosphere.

By maintaining the voltage being collected in a prescribed range, an electrical conversion system is easily designed. While FIGS. **4**A, **4**B, and **4**C illustrate some electrical configurations, those of ordinary skill in the art readily recognize a variety of other configurations which will serve the same function.

Referencing FIG. **4**A, Direct Current In (DC IN) **40** is buffered by a gang of capacitors **41** before being communicated to a DC/AC converter **42**. The DC/AC converter converts the direct current into a an alternating current suitable for placement over an existing electrical grid **43** such as normally found from a power-plant.

Those of ordinary skill in the art readily recognize a variety of DC/AC converters, including, but not limited to: U.S. Pat. No. 7,394,671, entitled "Controller IC, DC-AC Conversion Apparatus, and parallel running system of DC-AC Conversion Apparatuses" issued to Fukumoto, et al. on Jul. 1, 2008; and, U.S. Pat. No. 7,330,366, entitled "DC-AC Converter" issued to Lee, et al. on Feb. 12, 2008; all of which are incorporated hereinto by reference.

FIG. **4**B illustrates an electrical arrangement suitable for use in charging a battery. DC IN **40** is buffered by capacitor bank **41** before entering into a step down transformer **43**. Step down transformer **43** reduces the voltage so that the voltage can safely be introduced into battery **44** which is connected to ground **45** at the battery's other pole.

Those of ordinary skill in the art readily recognize a variety of batteries which will work in this capacity, including, but not limited to those described in: U.S. Pat. No. 7,378,181, entitled "Electric Storage Battery Construction and Manufacture" issued to Skinlo on May 27, 2008; U.S. Pat. No. 7,388,350, entitled "Battery with Electronic Compartment" issued to Wright on Jun. 17, 2008; U.S. Pat. No. 7,397,220, entitled "Connection Member and Battery Pack" issued to Uchida, et al. on Jul. 8, 2008; and, U.S. Pat. No. 7,375,492, entitled "Inductively Charged Battery Pack" issued to Calhoon, et al. on May 20, 2008; all of which are incorporated hereinto by reference.

In FIG. **4**C, DC IN **40** is fed into an adjustable rheostat **46** which is controlled by the controller so that the DC OUT **47** falls within a specified range.

FIG. **5** illustrates a conductive line used in the preferred embodiment of the invention.

This type of conductive line is commonly called poly-wire and consists of multiple interwoven strands of plastic **50**A and **50**B woven into a cord or rope arrangement having intertwined therein exposed metal wires **51**A and **51**B. While this illustration shows two plastic strands and two metal wires, any number of possible combinations is possible.

The exposed metal wires **51**A and **51**B attract the atmospheric static charge and transmit the charge down to the base unit (not shown).

FIGS. **6**A and **6**B illustrate an alternative conductive line creating an ionized pathway for the flow of the static charges from the atmosphere.

This conductive line utilizes a tube **60** having an outer layer **62** of PET Film (Biaxially-oriented polyethylene terephtalate polyester film) which provides exceptionally high tensile strength and is chemically and dimensionally stable. The tube has an ideal diameter of between two and three inches.

An interior metal coating **61** provides an initial conduit for the flow of static charge. The static charge through the metal forces the tube to expand due to the repulsion experienced by like charges. Further, the flow of electricity causes the interior of the tube **60** to become ionized to provide an additional pathway for the atmospheric static charges to the base unit (not shown).

Because outer layer **62** provides a gas barrier, the resulting ionization is not dissipated by air currents, thereby providing a highly stable pathway.

FIG. **7** illustrates the controller of an alternative embodiment and the associated safety devices.

In this embodiment, controller box **70**, resting on insulating pad **72**, is in communication with the sensors as described above. Using the input from these sensors, when there is flow of electricity through the base unit, warning flashing light **71** is illuminated. To electrically neutralize the mechanism, switch **73** is activated to pass any existing current into the ground **74**.

FIGS. **8**A and **8**B illustrate two embodiments of enhanced electrical collection leads.

Referencing FIG. **8**A, enhanced electrical collection lead **82** is a wire mesh which is in electrical communication with conductive line **81** and balloon **80**. Because of the significant amount of metal exposed by enhanced electrical collection lead **82**, more static electricity from the atmosphere is drawn to the collection lead **82**, and then down conductive line **81** to the base unit (not shown).

Conductive lead **82** is positioned proximate to balloon **80**.

In FIG. **8**B, poly-wire **83** has enhanced electrical collection leads **84** wrapped therearound. Collection leads **84** have pointed ends **85**A and **85**B which have a propensity to attract more electricity than rounded ends do.

It is clear from the foregoing that the present invention captures an entirely new source of electrical energy.

What is claimed is:

1. A mechanism to tap an electrical source comprising:

a) a lighter than air balloon suspended in the atmosphere;

b) a base unit having a spool of conductive line on a winch motor, one end of said conductive line secured to said lighter than air balloon, a portion of said conductive line collecting electricity in the atmosphere, said winch motor capable of selectively extending or withdrawing said conductive line from said spool;

c) an insulator electrically isolating said conductive line from ground;

d) a conductor having a first end electrically connected to said conductive line and a second end electrically connected to a load being powered by collected electricity from said conductive line;

e) an electrical flow sensor monitoring electrical flow through said conductor and generating an electrical flow indicia indicative of said electrical flow in said conductor; and,

f) a controller receiving said electrical flow indicia and selectively operating said winch motor such that said electrical flow indicia remains within a selected operating range.

2. The mechanism to tap an electrical source according to claim **1**, wherein said selected operating range is established by an operator.

3. The mechanism to tap an electrical source according to claim **2**, further including a warning light activated when said electrical flow indicia is non-zero.

4. The mechanism to tap an electrical source according to claim **1**,

a) further including a lightning sensor generating a lightning presence indicia indicative of lightning within a prescribed range, said presence indicia being communicated to said controller; and,

b) wherein said control mechanism, in response to said lightning presence indicia, operates said winch motor to withdraw substantially all of said conductive line onto said spool.

5. The mechanism to tap an electrical source according to claim **4**, wherein said prescribed range is established by an operator.

6. The mechanism to tap an electrical source according to claim **1**, further including an electrical collection enhancement lead in electrical contact with a first end of said conductive material, said electrical collection enhancement lead configured to attract static electricity.

7. The mechanism to tap an electrical source according to claim **6**, wherein said electrical collection enhancement lead is positioned proximate to said lighter than air balloon.

8. The mechanism to tap an electrical source according to claim **7**, wherein said electrical collection enhancement lead includes at least two pointed electrical conductors.

9. The mechanism to tap an electrical source according to claim **1**, wherein said insulator electrically isolates said conductive line from said winch motor.

10. The mechanism to tap an electrical source according to claim **1**, wherein said insulator electrically isolates said base unit from ground.

11. A mechanism comprising:

a) an airborne aircraft having a conductive line secured to a winch capable of extending or withdrawing said conductive line from a spool, when said airborne aircraft is aloft, said conductive line collecting electricity from the atmosphere;

b) a conductor having a first end electrically connected to said conductive line and a second end electrically connected to a load such that collected electricity from said conductive line powers said load;

c) an electrical flow sensor monitoring electrical flow through said conductor and generating an electrical flow indicia indicative of said electrical flow in said conductor; and,

d) a controller receiving said electrical flow indicia adjusting an altitude of said airborne aircraft via said winch such that said electrical flow indicia remains within a selected operating range.

12. The mechanism to tap an electrical source according to claim **11**,

a) further including a lightning sensor communicating with said controller; and,

b) wherein said controller, in response to selected signals from said lightning sensor withdraws substantially all of said conductive line.

13. A mechanism to tap an electrical source comprising:

a) a lighter than air balloon suspended in the atmosphere;

b) a base unit having a spool of conductive line on a winch motor, one end of said conductive line suspended by said balloon, a portion of said conductive line collecting electricity from the atmosphere, said winch motor capable of adjusting an altitude of said lighter than air balloon by selectively extending or withdrawing said conductive line from said spool;

c) a conductor having a first end electrically connected to said conductive line and a second end electrically connected to a load being powered by said collected electricity from said conductive line;

d) a sensor array having,

1) an electrical flow sensor monitoring electrical flow through said conductor and generating an electrical flow indicia indicative of said electrical flow in said conductor, and,

2) a lightning sensor monitoring existence of proximate lightning; and,

e) a controller responsive to said electrical flow indicia from said sensor array to selectively operate said winch motor.

14. The mechanism to tap an electrical source according to claim **13**,

a) further including a warning light; and,

b) wherein said controller activates said warning light when said electrical flow in said conductor is non-zero.

15. The mechanism to tap an electrical source according to claim **14**, further including an electrical collection enhancement lead in electrical contact with said conductive material, said electrical collection enhancement lead configured to attract static electricity and positioned proximate to said lighter than air balloon.

* * * * *

US009160156B2

(12) United States Patent
Allen et al.

(10) **Patent No.: US 9,160,156 B2**
(45) **Date of Patent: Oct. 13, 2015**

(54) SYSTEM FOR HARVESTING ATMOSPHERIC ELECTRICITY

(75) Inventors: **Edward H. Allen**, Lancaster, CA (US); **Sergey Macheret**, Palmdale, CA (US)

(73) Assignee: **Lockheed Martin Corporation**, Bethesda, MD (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 357 days.

(21) Appl. No.: **13/276,259**

(22) Filed: **Oct. 18, 2011**

(65) **Prior Publication Data**

US 2013/0093261 A1 Apr. 18, 2013

(51) **Int. Cl.**

H05F 7/00	(2006.01)
H02G 13/00	(2006.01)
B64D 45/02	(2006.01)

(52) **U.S. Cl.**
CPC ***H02G 13/20*** (2013.01); ***B64D 45/02*** (2013.01); *H05F 7/00* (2013.01)

(58) **Field of Classification Search**
CPC H02G 13/20; B64D 45/02; H05F 7/00
USPC 361/231, 230
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

3,719,829	A *	3/1973	Vaill	307/149
4,017,767	A *	4/1977	Ball	361/117
4,950,978	A *	8/1990	Govaert	324/72
5,726,855	A *	3/1998	Mourou et al.	361/213
8,045,314	B2 *	10/2011	Ibok	361/231
2014/0231679	A1 *	8/2014	Kremeyer et al.	250/505.1

OTHER PUBLICATIONS

Tzzortzakis et al. "Concatenation of plasma filaments created in air by femtosecond laser infrared laser pulses", 2003, Applied Physics B, vol. 767, pp. 609-612.*

Zhou et al. , Revival of femtosecond laser plasma filaments in air by a nanosecond laser, Jun. 24, 2009, Optic Express, vol. 17, No. 14, pp. 1 -7.*

Assion et al., "Control of Chemical Reactions by Feedback-Optimized Phase-Shaped Femtosecond Laser Pulses", Science, Oct. 30, 1998, pp. 919-922, vol. 282.

* cited by examiner

Primary Examiner — Zeev V Kitov
(74) *Attorney, Agent, or Firm* — Baker Botts L.L.P.

(57) ABSTRACT

Systems, methods and apparatus for harvesting atmospheric electricity are provided. The system includes a laser configured to form a plasma filament and a collector configured to collect electricity flowing along the plasma filament. The plasma filament comprises an electrically conducting plasma filament. Atmospheric electricity may be collected by having the plasma filament form at least a part of a conducting path: (1) between ground and a cloud, (2) between differently charged regions of the same cloud, (3) between differently charged regions of different clouds, and (4) between different regions of atmosphere, where there is a vertical voltage gradient. When the plasma filament is not long enough to form the entire conducting path, a lightning may be triggered to complete the conducting path needed to collect atmospheric electricity.

21 Claims, 6 Drawing Sheets

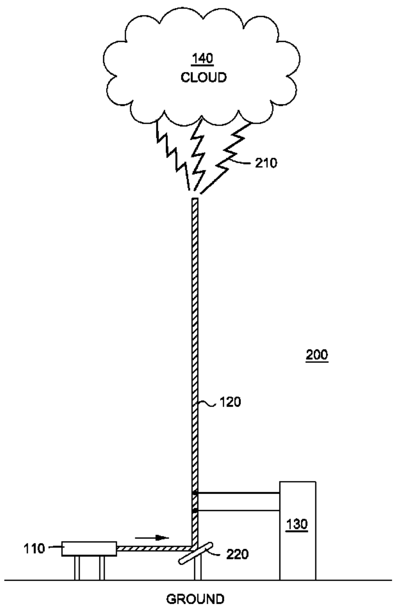

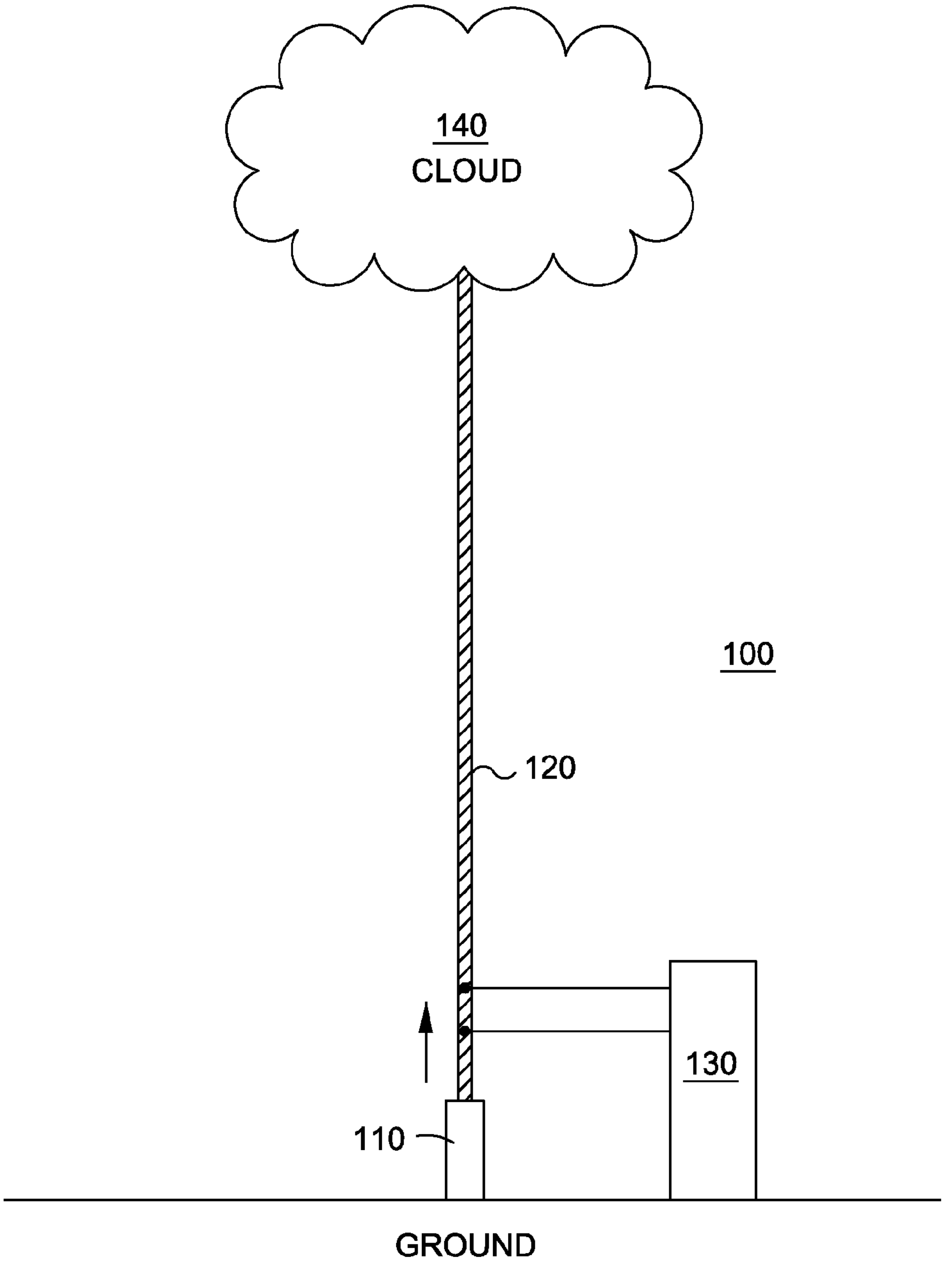
140
CLOUD
100
120
130
110
GROUND

FIG. 1

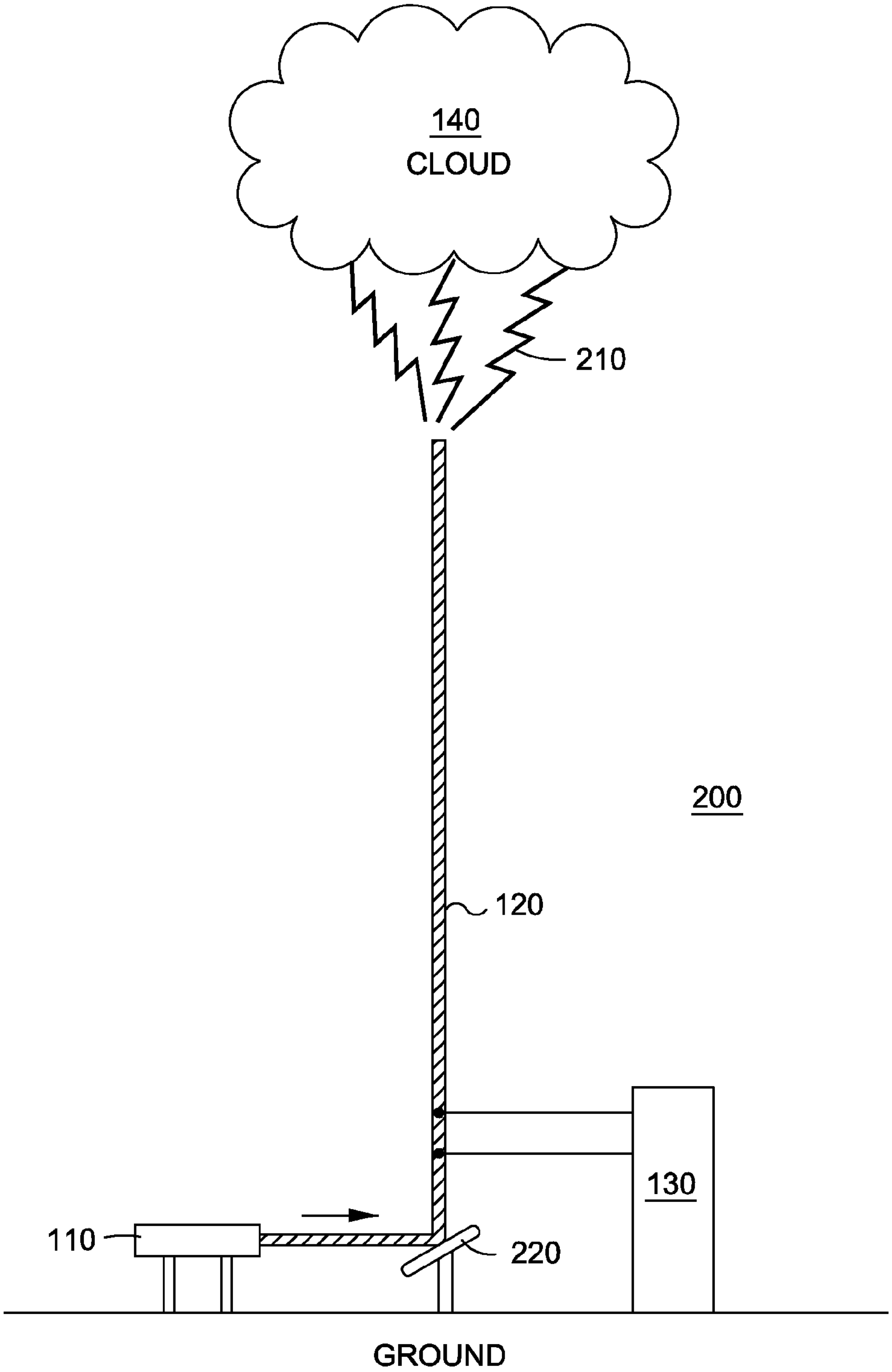
140
CLOUD
210
200
120
130
110
220
GROUND

FIG. 2

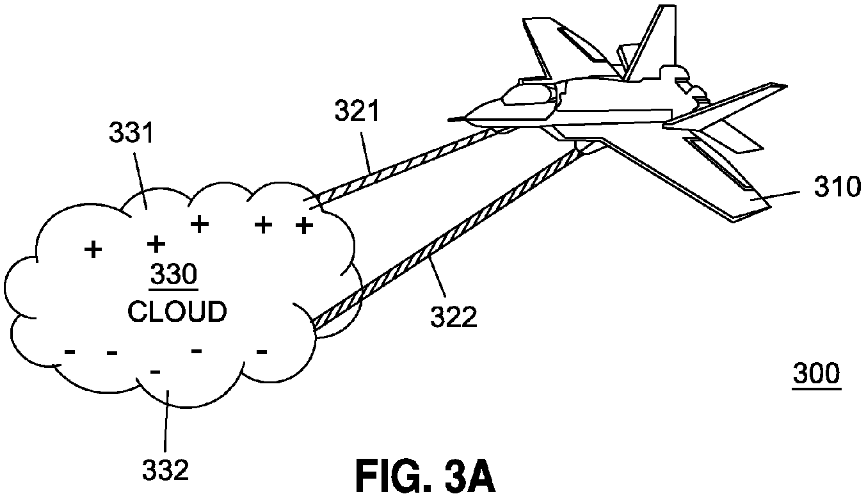
+ + + + + +
CLOUD
- - - - -
-

FIG. 3A

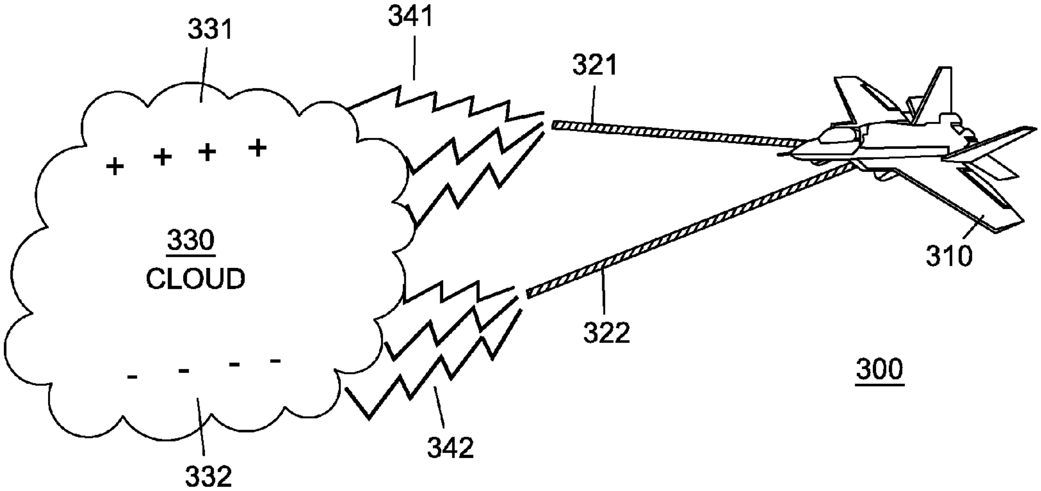
+ + + +
CLOUD
- - - -

FIG. 3B

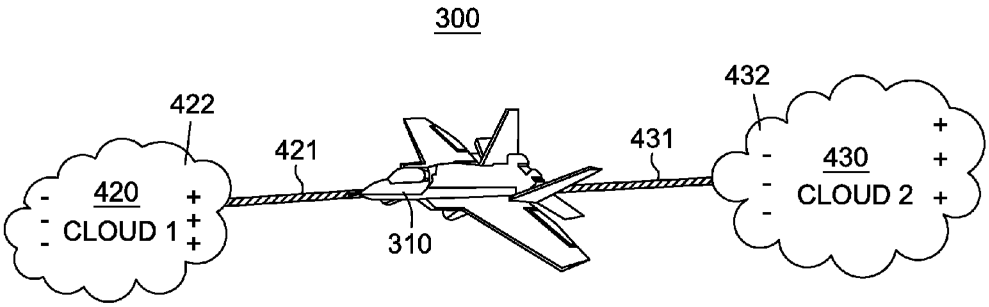
300
422
421
420
CLOUD 1
310
431
432
430
CLOUD 2

FIG. 4

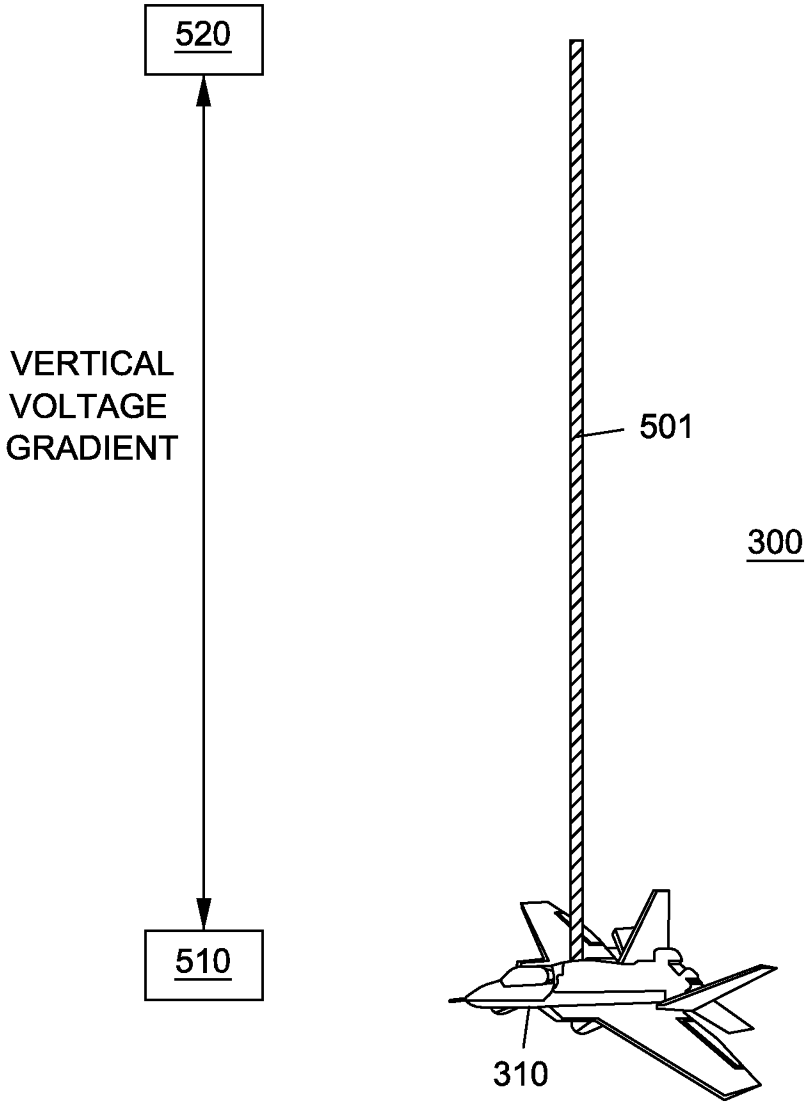
520
VERTICAL
VOLTAGE
GRADIENT
510
501
300
310

FIG. 5

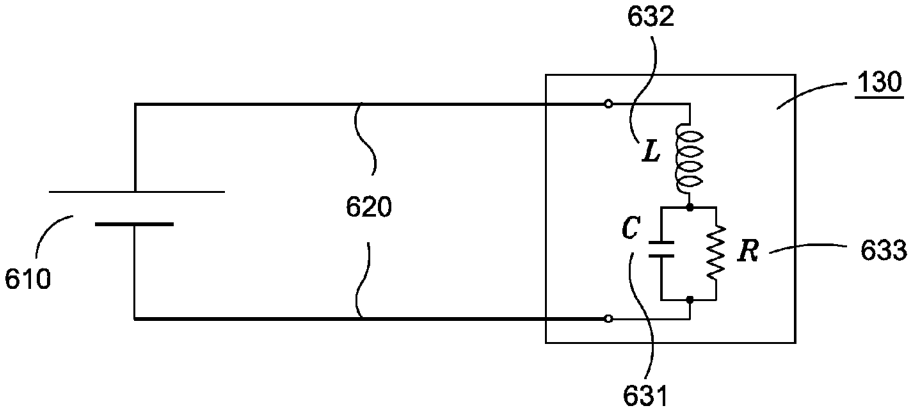
632
130
L
620
C
R
633
610
631

FIG. 6

SYSTEM FOR HARVESTING ATMOSPHERIC ELECTRICITY

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

Not applicable.

FIELD OF THE INVENTION

The subject technology relates in general to a system for harvesting atmospheric electricity, and more particularly to a system that uses a laser to form an electrically conducting plasma filament for harvesting atmospheric electricity.

BACKGROUND

It is well known that large quantities of electrical energy are present in the atmosphere and in lightning. Lightning was one of the first forms of electricity harnessed in the modern age by Benjamin Franklin in his famous kite experiment.

A lightning discharge typically contains on the order of 10^{10} Joules of energy. It has been estimated that the total electrical power of lightning across the Earth is of the order of 10^{12} watts. Collection of electrical energy from clouds associated with lightning is a problem that needs to be resolved.

Lightning is but a small part of the total electrical activity of the atmosphere. When a local build up of charge above the Earth exceeds the local breakdown potential of the atmosphere a lightning discharge occurs. However, there is a continual invisible flow of charge from Ionosphere to Earth occurring day and night over the entire surface of the globe, which exceeds the global lightning power output by many times. The source of this flow of charge from the atmosphere to Earth is the Sun. Radiation from the Sun helps form most of the ions found in the Ionosphere, a highly charged region above the atmosphere, through ionization of atmospheric molecules. Solar radiation, particularly in the UV and soft x-ray bands, consist of photons whose energies are well suited to ionization. The Sun also emits continuously a solar wind of positively charged particles. These are captured by the Earth's magnetic field and further contribute to the Ionosphere. The positively charged region in the Ionosphere in turn induces (by electrostatic induction) a negative charge on the surface of the Earth. The Earth becomes in effect an enormous spherical capacitor. A potential gradient or electric field is thus established between the two "plates" of this capacitor, the Ionosphere (or Electrosphere) and the surface of the Earth. While the upper strata of the atmosphere conduct electricity reasonably well, the lower levels act as an insulator or dielectric. Near the surface of the Earth, this electrostatic potential gradient is on the order of about 100 Vm^{-1} in summer, rising to 300 Vm^{-1} in winter. This flow of charge can be tapped and directed to provide useable electrical power. This source of atmospheric electricity has the following advantages: (1) Simple and robust technology; (2) Low cost technology—much cheaper than photovoltaics or wind turbines; (3) Available day and night in all weather conditions—in fact, more power is produced at night than during the day; and (4) Available at any point on the Earth's surface. Collection of this source of atmospheric electricity is another problem that needs to be resolved.

SUMMARY

According to various aspects of the subject technology, a system for harvesting atmospheric electricity is provided that solves some or all of the foregoing problems. In some aspects, the system may be used to harvest atmospheric electricity arising from an electrostatic potential gradient: (1) between ground and a cloud, (2) between differently charged regions of the same cloud, (3) between differently charged regions of different clouds, and (4) between different regions of atmosphere, where there is a vertical voltage gradient. In other aspects, the system may be used to harvest atmospheric electricity arising from an electrostatic potential gradient between the Ionosphere (or Electrosphere) and the surface of the Earth.

According to various aspects of the subject technology, a system for harvesting atmospheric electricity is provided. The system comprises a laser configured to form a plasma filament. The system also comprises a collector configured to collect electricity flowing along the plasma filament.

According to various aspects of the subject technology, a method for harvesting atmospheric electricity is provided. The method comprises forming a plasma filament with a laser. The method also comprises collecting electricity flowing along the plasma filament.

According to various aspects of the subject technology, an apparatus for harvesting atmospheric electricity is provided. The apparatus comprises means for forming a plasma filament with a laser. The system also comprises means for collecting electricity flowing along the plasma filament.

It is understood that other configurations of the subject technology will become readily apparent to those skilled in the art from the following detailed description, wherein various configurations of the subject technology are shown and described by way of illustration. As will be realized, the subject technology is capable of other and different configurations and its several details are capable of modification in various other respects, all without departing from the scope of the subject technology. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are included to provide further understanding of the subject technology and are incorporated in and constitute a part of this specification, illustrate aspects of the subject technology and together with the description serve to explain the principles of the subject technology.

FIG. **1** illustrates a first example of a system for harvesting atmospheric electricity. The system comprises a laser configured to form a plasma filament and a collector configured to collect electricity flowing along the plasma filament. The plasma filament forms a conducting path between ground and a cloud.

FIG. **2** illustrates a second example of a system for harvesting atmospheric electricity. A plasma filament and a lightning together form a conducting path between ground and a cloud. A condensing mirror directs the beam of laser towards the cloud.

FIG. **3**A illustrates a third example of a system for harvesting atmospheric electricity. A laser system situated on an airplane forms two plasma filaments that together complete a conducting path between a first region of a cloud and a second region of the cloud. The first region of the cloud and the second region of the cloud carry opposite electric charges. A collector configured to collect electricity flowing along the two plasma filaments is also situated on the same airplane as the laser system.

FIG. **3**B illustrates a fourth example of a system for harvesting atmospheric electricity. It is similar to FIG. **3**A, except two lightnings complete the conducting path between a first region of a cloud and a second region of the cloud (i.e., a first lightning electrically connects the first region of the cloud to one of the plasma filaments and a second lightning electrically connects the second region of the cloud to the other plasma filament).

FIG. **4** illustrates a fifth example of a system for harvesting atmospheric electricity. A laser system situated on an airplane forms two plasma filaments that together complete a conducting path between a first cloud and a second cloud. The region of the first cloud that is connected to the first plasma filament and the region of the second cloud that is connected to the second plasma filament carry opposite electric charges.

FIG. **5** illustrates a sixth example of a system for harvesting atmospheric electricity. A laser situated on an airplane forms a plasma filament that completes a conducting path between a first region of atmosphere and a second region of atmosphere. A collector situated on the same airplane as the laser can collect electricity flowing along the plasma filament, because a vertical voltage gradient exists between the first region of atmosphere and the second region of atmosphere.

FIG. **6** illustrates an example of a collector configured to collect electricity flowing along the plasma filament. The collector may be a charge collection circuit characterized by capacitance, inductance, and resistance. The electricity may be collected by charging a capacitor.

DETAILED DESCRIPTION

The detailed description set forth below is intended as a description of various configurations of the subject technology and is not intended to represent the only configurations in which the subject technology may be practiced. The appended drawings are incorporated herein and constitute a part of the detailed description. The detailed description includes specific details for the purpose of providing a thorough understanding of the subject technology. However, it will be apparent to those skilled in the art that the subject technology may be practiced without these specific details. In some instances, well-known structures and components are shown in block diagram form in order to avoid obscuring the concepts of the subject technology. Like components are labeled with identical element numbers for ease of understanding.

FIG. **1** illustrates a first example of a system **100** for harvesting atmospheric electricity. The system **100** comprises a laser **110** configured to form a plasma filament **120** and a collector **130** configured to collect electricity flowing along the plasma filament **120**. The plasma filament **120** forms a conducting path between ground and a cloud **140**.

The plasma filament **120** comprises an electrically conducting plasma filament. The plasma comprises of electrons and positive ions. The electrically conducting plasma filament may be formed in atmosphere using an ultra short pulse laser (USPL). The USPL forms a plasma filament by raising the conductivity of a very small diameter channel (~100 microns or less) for a very brief time. The propagation of high power (~10^{14} W/cm^2) pulses from the USPL is accompanied by filamentation—self-channeling of femtosecond laser pulses in stable high-intensity light filaments with ~100 microns diameters (so that the Kerr effect focusing balances defocusing due to plasma formation). This filamentation keeps the beam virtually free of diffraction divergence. The long plasma strings have high electron densities (10^{16}/cm^3) which support high conductivities. The ionization of air at these intensities results from multi-photon processes. Filamentation has been produced and detected at altitudes as high as 13-20 km. Almost no beam scattering occurs in filamentation phenomenon, and virtually all the pulse energy is invested in ionization of air molecules. With the typical electron density and channel diameter, a 160 mJ pulse is sufficient to generate a 1 km long channel.

Once initiated, plasma filaments cannot go on forever, and typically only propagate a kilometer or so before energy depletion and exhaustion. However, by using two coaxial USPL projection optics having different time-lensing parameters, a near and far filament could be initiated in a "daisy-chain" to extend the total length of the conductive channel. The onset of filamentation in the far filament is time-lensed to coincide with the exhaustion range of the near filament. Additionally, using negative chirp, femtosecond laser pulses can propagate almost without loss until different frequency components of the wave packet, propagating at different speeds due to dispersion in air, come together at a certain desired location so that the laser intensity there exceeds the filamentation threshold, resulting in long ionized filaments created at a distance of up to several miles from the laser source. In other words, negative chirp is used to essentially focus a beam of the laser at a desired distance to form the plasma filament there. Accordingly, this technique may be used to move the starting point and to extend the length of the plasma filament.

The size, electrical conductivity, and lifetime of the plasma filaments are enhanced by an auxiliary radio-frequency electromagnetic waves or microwaves. The plasma filament may act like an antenna to absorb the auxiliary radio-frequency (RF) electromagnetic waves or microwaves.

The pulsed collection of electric charge and energy is repeated with an optimal (possibly very high) repetition rate of the pulsed laser and of the auxiliary RF waves or microwaves. The energy collected per one pulse should exceed the energy spent on creating and enhancing the plasma filaments, so that the net energy collected is positive. The optimal pulse repetition rate (PRR) depends upon atmospheric conditions which can vary widely over even short time durations. As a general rule, one tries to tune the repetition rate to the plasma relaxation time (the ion-electron recombination time) so that the conductivity of the filament remain as high as is optimal given the atmospheric conditions at the time. A feedback control system that varies the PRR as well as the rest of the pulse-shaping parameters may be employed to optimize the net power production of the system. For more information regarding pulse shaping and tailoring, which is the central theme of the field called "quantum control", please see A. Assion, T. Baumert, M. Bergt, T. Brixner, B. Kiefer, V. Seyfried, M. Strehle, G. Gerber (30 October). "Control of Chemical Reactions by Feedback-Optimized Phase-Shaped Femtosecond Laser Pulses". Science 282 (5390): 919-922. doi: 10.1126/science.282.5390.919. PMID 9794756.

FIG. **2** illustrates a second example of a system **200** for harvesting atmospheric electricity. A plasma filament **120** and a lightning **210** together form a conducting path between ground and a cloud **140**. A condensing mirror **220** directs the beam of laser **110** towards the cloud **140**. A collector **130** is configured to collect electricity flowing along the plasma filament **120**.

In the case of electricity collection from the cloud **140**, the initial and RF-enhanced plasma filaments can be relatively short, not reaching all the way from ground to the cloud **140**. The relatively short plasma filament can be optimized so that it would trigger propagation of the so-called leader (as in natural lightning), which would effectively use the energy and potential of the cloud to eventually create an electrically conducting path between ground and the cloud **140**. The final conducting path would comprise of a plasma filament **120** and a lightning **210**.

It is not shown in the figures, but alternatively both the laser **110** and the collector **130** may be placed on a mobile device, which moves together with the cloud. One such mobile device may be a vehicle moving on the ground, but another such mobile device may be an airplane flying in the air. In the case of an airplane flying in the air, plasma filaments may be formed by the laser to reach both ground and the cloud at the same time. In the event the plasma filament directed at ground is not long enough to reach ground, a lightning may also be formed to complete the conducting path to ground.

FIG. **3**A illustrates a third example of a system **300** for harvesting atmospheric electricity. A laser system situated on an airplane **310** forms two plasma filaments (**321**, **322**) that together complete a conducting path between a first region **331** of a cloud **330** and a second region **332** of the cloud **330**. The first region **331** of the cloud **330** and the second region **332** of the cloud **330** carry opposite electric charges. In the example shown here, the first region **331** of the cloud **330** carries positive electric charges, while the second region **332** of the cloud **330** carries negative electric charges. A collector configured to collect electricity flowing along the two plasma filaments is also situated on the same airplane **310** as the laser. Positive and negative charges are naturally separated within the cloud, so this example makes use of the electrostatic potential gradient between different regions of the same cloud.

FIG. **3**B illustrates a fourth example of the system **300** for harvesting atmospheric electricity. It is similar to FIG. **3**A, except two lightnings (**341**, **342**) complete the conducting path between a first region **331** of a cloud **330** and a second region **332** of the cloud **330** (i.e., a first lightning **341** electrically connects the first region **331** of the cloud **330** to one of the plasma filaments **321** and a second lightning **342** electrically connects the second region **332** of the cloud **330** to the other plasma filaments **322**). The lightning formation is again because the initial and RF-enhanced plasma filaments can be relatively short, not reaching all the way from the collector situated on airplane **310** to the cloud **330**. These relatively short plasma filaments can be optimized so that they would trigger propagation of the so-called leader (as in natural lightning), which would eventually create an electrically conducting path between the two regions of cloud **330** via a connection through airplane **310**. The final conducting path would comprise of two plasma filaments (**321**, **322**) and two lightnings (**341**, **342**). Of course, one of the lightnings may not be formed, if one of the plasma filament can reach all the way from airplane **310** to cloud **330**.

FIG. **4** illustrates a fifth example of the system **300** for harvesting atmospheric electricity. A laser system situated on an airplane **310** forms two plasma filaments (**421**, **431**) that together complete a conducting path between a first cloud **420** and a second cloud **430**. The region **422** of the first cloud **420** that is connected to the first plasma filament **421** and the region **432** of the second cloud **430** that is connected to the second plasma filament **431** carry opposite electric charges. In the example shown here, region **422** of cloud **420** carries positive electric charges, while region **432** of cloud **430** carry negative electric charges. Positive and negative charges are naturally separated within the cloud, so this example makes use of the electrostatic potential gradient between different regions of different clouds. Once again, in case one of the plasma filament cannot reach all the way from airplane **310** to any one of the clouds, then lightning may be formed to complete an electrically conducting path from airplane **310** to any one of the clouds. But these embodiments are not shown in the figures.

FIG. **5** illustrates a sixth example of the system **300** for harvesting atmospheric electricity. A laser system situated on an airplane **310** forms a plasma filament **501** that completes a conducting path between a first region **510** of atmosphere and a second region **520** of atmosphere. A collector situated on the same airplane **310** as the laser system can collect electricity flowing along the plasma filament **501**, because a vertical voltage gradient exists between the first region **510** of atmosphere and the second region **520** of atmosphere. Because no clouds are needed for this embodiment, atmospheric electricity collection is possible under fair weather conditions (relying on the naturally existing vertical voltage gradient in the atmosphere) as well as under foul weather conditions.

FIG. **6** illustrates an example of a collector **130** configured to collect electricity flowing along a plasma filament **620**. The collector may be a charge collection circuit characterized by capacitance **631**, inductance **632**, and resistance **633**. The electricity may be collected by charging a capacitor associated with capacitance **631**. Battery **610** represents these naturally existing voltage sources: (1) between ground and a cloud, (2) between differently charged regions of the same cloud, (3) between differently charged regions of different clouds, and (4) between different regions of atmosphere, where there is a vertical voltage gradient. The plasma filament **620** can, for the purpose of analysis and optimization of the charge collection process, be viewed as electromagnetic transmission lines characterized by their effective capacitance, inductance, and resistance per unit length. Accordingly, the parameters of the charge collection circuit and of the effective transmission line representing the plasma filaments are optimized, so as to maximize the net charge and energy collected (minus the energy spent on generating and enhancing the plasma filaments).

The foregoing description is provided to enable a person skilled in the art to practice the various configurations described herein. While the subject technology has been particularly described with reference to the various figures and configurations, it should be understood that these are for illustration purposes only and should not be taken as limiting the scope of the subject technology.

There may be many other ways to implement the subject technology. Various functions and elements described herein may be partitioned differently from those shown without departing from the scope of the subject technology. Various modifications to these configurations will be readily apparent to those skilled in the art, and generic principles defined herein may be applied to other configurations. Thus, many changes and modifications may be made to the subject technology, by one having ordinary skill in the art, without departing from the scope of the subject technology.

It is understood that the specific order or hierarchy of steps in the processes disclosed is an illustration of exemplary approaches. Based upon design preferences, it is understood that the specific order or hierarchy of steps in the processes may be rearranged. Some of the steps may be performed simultaneously. The accompanying method claims present elements of the various steps in a sample order, and are not meant to be limited to the specific order or hierarchy presented.

A phrase such as "an aspect" does not imply that such aspect is essential to the subject technology or that such aspect applies to all configurations of the subject technology. A disclosure relating to an aspect may apply to all configurations, or one or more configurations. An aspect may provide one or more examples of the disclosure. A phrase such as an "aspect" may refer to one or more aspects and vice versa. A phrase such as an "embodiment" does not imply that such embodiment is essential to the subject technology or that such embodiment applies to all configurations of the subject technology. A disclosure relating to an embodiment may apply to all embodiments, or one or more embodiments. An embodiment may provide one or more examples of the disclosure. A phrase such an "embodiment" may refer to one or more embodiments and vice versa. A phrase such as a "configuration" does not imply that such configuration is essential to the subject technology or that such configuration applies to all configurations of the subject technology. A disclosure relating to a configuration may apply to all configurations, or one or more configurations. A configuration may provide one or more examples of the disclosure. A phrase such as a "configuration" may refer to one or more configurations and vice versa.

Furthermore, to the extent that the term "include," "have," or the like is used in the description or the claims, such term is intended to be inclusive in a manner similar to the term "comprise" as "comprise" is interpreted when employed as a transitional word in a claim.

The word "exemplary" is used herein to mean "serving as an example, instance, or illustration." Any embodiment described herein as "exemplary" is not necessarily to be construed as preferred or advantageous over other embodiments.

A reference to an element in the singular is not intended to mean "one and only one" unless specifically stated, but rather "one or more." The term "some" refers to one or more. Underlined and/or italicized headings and subheadings are used for convenience only, do not limit the subject technology, and are not referred to in connection with the interpretation of the description of the subject technology. All structural and functional equivalents to the elements of the various configurations described throughout this disclosure that are known or later come to be known to those of ordinary skill in the art are expressly incorporated herein by reference and intended to be encompassed by the subject technology. Moreover, nothing disclosed herein is intended to be dedicated to the public regardless of whether such disclosure is explicitly recited in the above description.

What is claimed is the following:

1. A system for harvesting atmospheric electricity, comprising:

a laser configured to form a plasma filament comprising at least a first filament and a second filament, wherein onset of formation of the second filament coincides with an exhaustion range of the first filament;

an auxiliary source of electromagnetic energy operable to enhance the plasma filament by applying electromagnetic energy to the plasma filament; and

a collector configured to collect electricity flowing along the plasma filament.

2. The system of claim **1**, wherein the plasma filament comprises an electrically conducting plasma filament.

3. The system of claim **1**, wherein the laser is a pulsed laser.

4. The system of claim **3**, wherein the pulsed laser is pulsed at an optimal repetition rate.

5. The system of claim **1**, wherein the electromagnetic energy comprises at least one of radio-frequency energy or microwave energy.

6. The system of claim **1**, wherein the plasma filament forms at least a part of a conducting path between ground and a cloud.

7. The system of claim **6**, wherein at least one lightning forms another part of the conducting path between ground and the cloud.

8. The system of claim **1**, wherein the plasma filament forms at least a part of a conducting path between a first region of a cloud and a second region of the cloud.

9. The system of claim **8**, wherein at least one lightning forms another part of the conducting path between the first region of the cloud and the second region of the cloud.

10. The system of claim **1**, wherein:

the laser is further configured to form a second plasma filament,

the plasma filament is connected to a region of a first cloud, and

the second plasma filament is connected to a region of a second cloud.

11. The system of claim **1**, wherein the plasma filament completes a conducting path between a first region of atmosphere and a second region of atmosphere.

12. The system of claim **11**, wherein a vertical voltage gradient exists between the first region of atmosphere and the second region of atmosphere.

13. The system of claim **1**, wherein the collector comprises a charge collection circuit.

14. The system of claim **1**, wherein the laser is situated on ground.

15. The system of claim **1**, wherein the laser is situated on a mobile device.

16. The system of claim **15**, wherein the mobile device is an airplane.

17. The system of claim **1**, wherein negative chirp is used to move the starting point of formation of the plasma filament to a desired location.

18. A method for harvesting atmospheric electricity, the method comprising:

forming a plasma filament with a laser, wherein the plasma filament comprises at least a first filament and a second filament, and wherein onset of formation of the second filament coincides with an exhaustion range of the first filament;

applying electromagnetic energy to the plasma filament to enhance the plasma filament; and

collecting electricity flowing along the plasma filament.

19. An apparatus for harvesting atmospheric electricity, the apparatus comprising:

means for forming a plasma filament with a laser, wherein the plasma filament comprises at least a first filament and a second filament, and wherein onset of formation of the second filament coincides with an exhaustion range of the first filament;

means for applying electromagnetic energy to the plasma filament to enhance the plasma filament; and

means for collecting electricity flowing along the plasma filament.

20. The system of claim **1**, wherein the laser is an ultra short pulse laser (USPL).

21. The system of claim **1**, wherein the laser comprises two coaxial USPL projection optics having different time-lensing parameters.

* * * * *

US 20100220424A1

(19) **United States**

(12) **Patent Application Publication**
IBOK

(10) **Pub. No.: US 2010/0220424 A1**

(43) **Pub. Date: Sep. 2, 2010**

(54) **METHOD OF ATMOSPHERIC DISCHARGE ENERGY CONVERSION, STORAGE AND DISTRIBUTION**

(75) Inventor: **EFFIONG ETUKUDO IBOK**, SUNNYVALE, CA (US)

Correspondence Address:
THE TRAVIS BUSINESS GROUP, INC.
1005 E. HOMESTEAD RD.
SUNNYVALE, CA 94087 (US)

(73) Assignees: **EFFIONG ETUKUDO IBOK**, SUNNYVALE, CA (US); **THE TRAVIS BUSINESS GROUP, INC**, SUNNYVALE, CA (US)

(21) Appl. No.: **12/534,115**

(22) Filed: **Aug. 1, 2009**

Publication Classification

(51) **Int. Cl.**
H02H 3/22 (2006.01)
H02J 7/00 (2006.01)

(52) **U.S. Cl.** **361/117**; 320/166

(57) **ABSTRACT**

A method of converting atmospheric electrical discharge to a useable form of energy by arresting, storage and retransmission of lightning induced electrical discharge is disclosed. The invention discloses methods of deploying this technology even in isolated locations where no electricity infrastructure exists. Additionally, the potential for achieving in excess of 1 GWe of electrical power supply at costs orders of magnitude lower than fossil fuel or solar is also disclosed. Isolated collection units are disclosed. A method of deployment of these devices on individual cars and recharge stations is disclosed which in effect unplugs these automotive and recharge stations from the grid. This capability significantly decreases their dependence on the electricity grid infrastructure, enabling a much more enhanced rollout capability for the industry. The concept of Energy Dams is disclosed. These are facilities with substantial electrical energy storage capacities with the capability of receiving energy feeds from various generation sources.

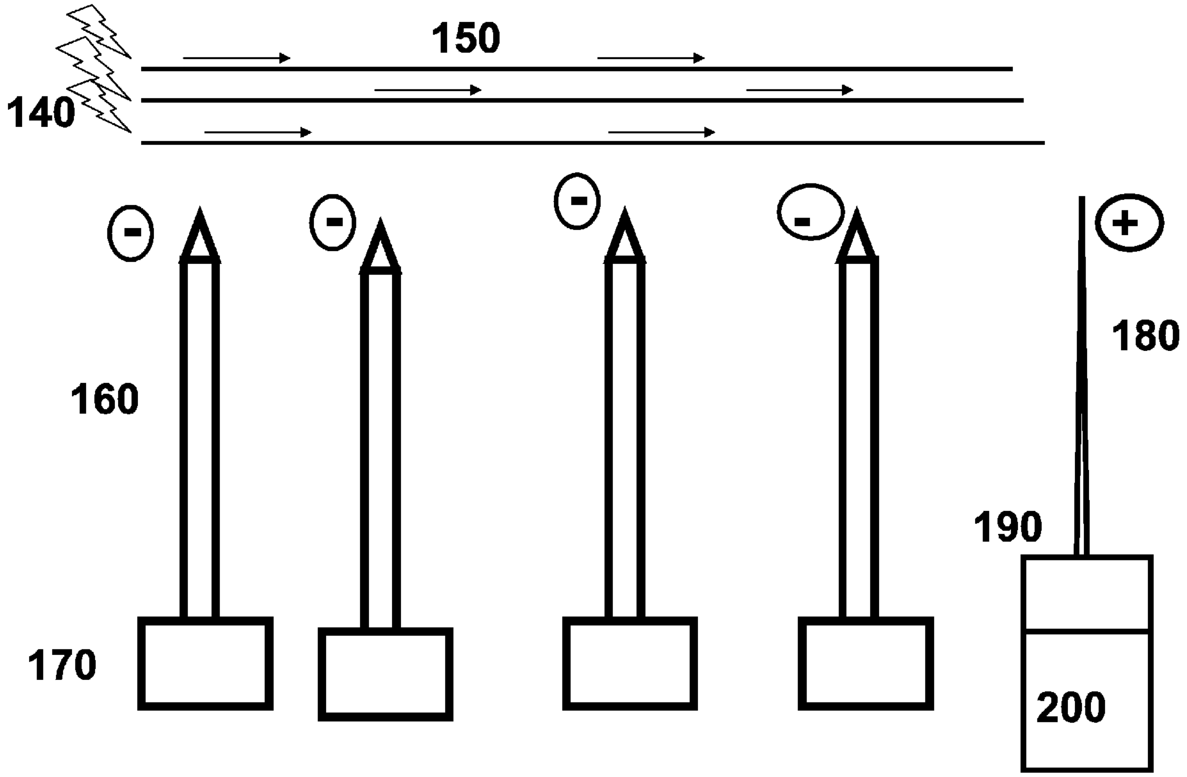

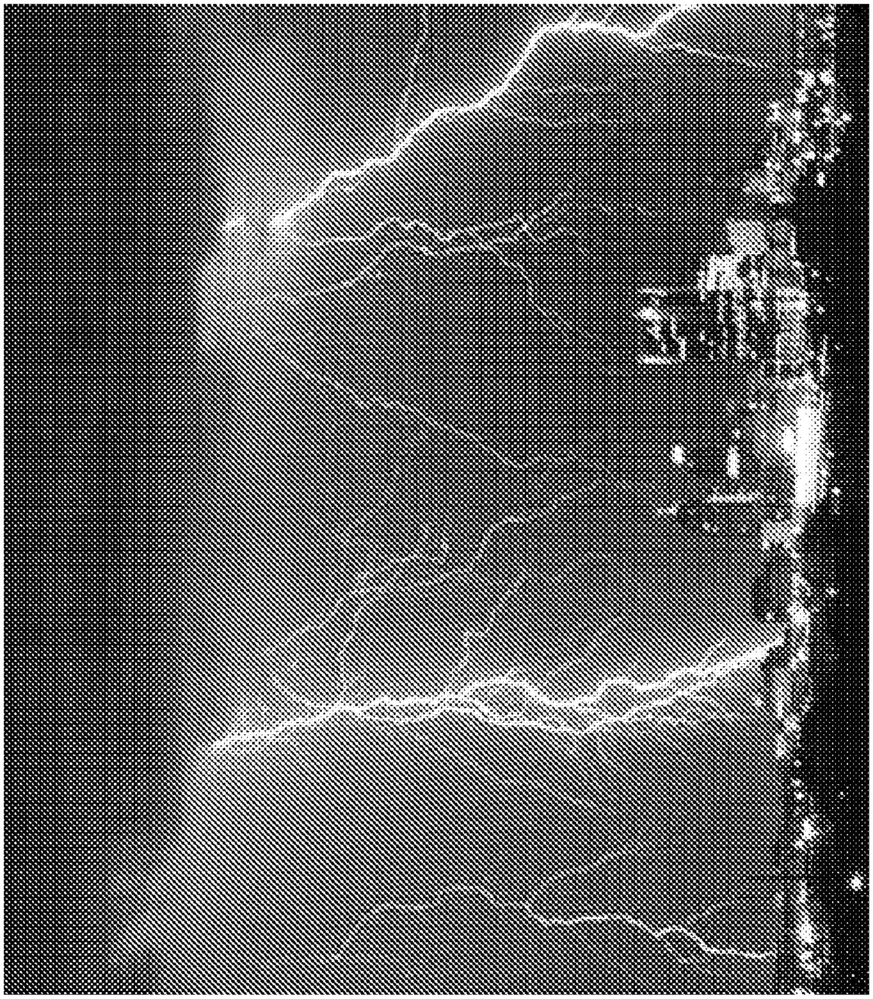

Fig. 1a

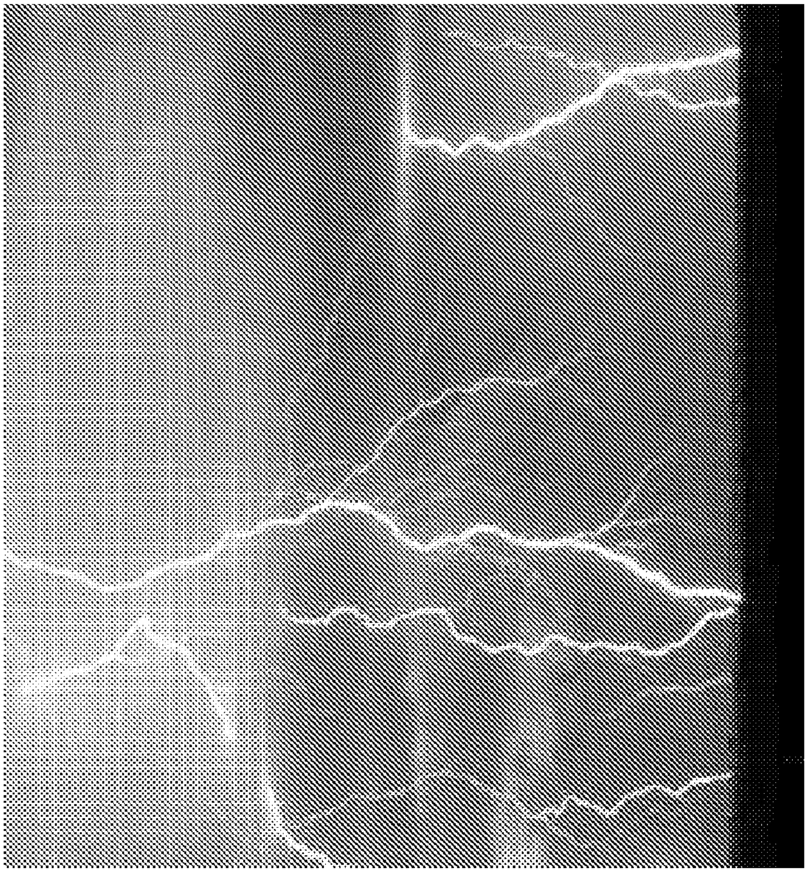

Fig. 1b

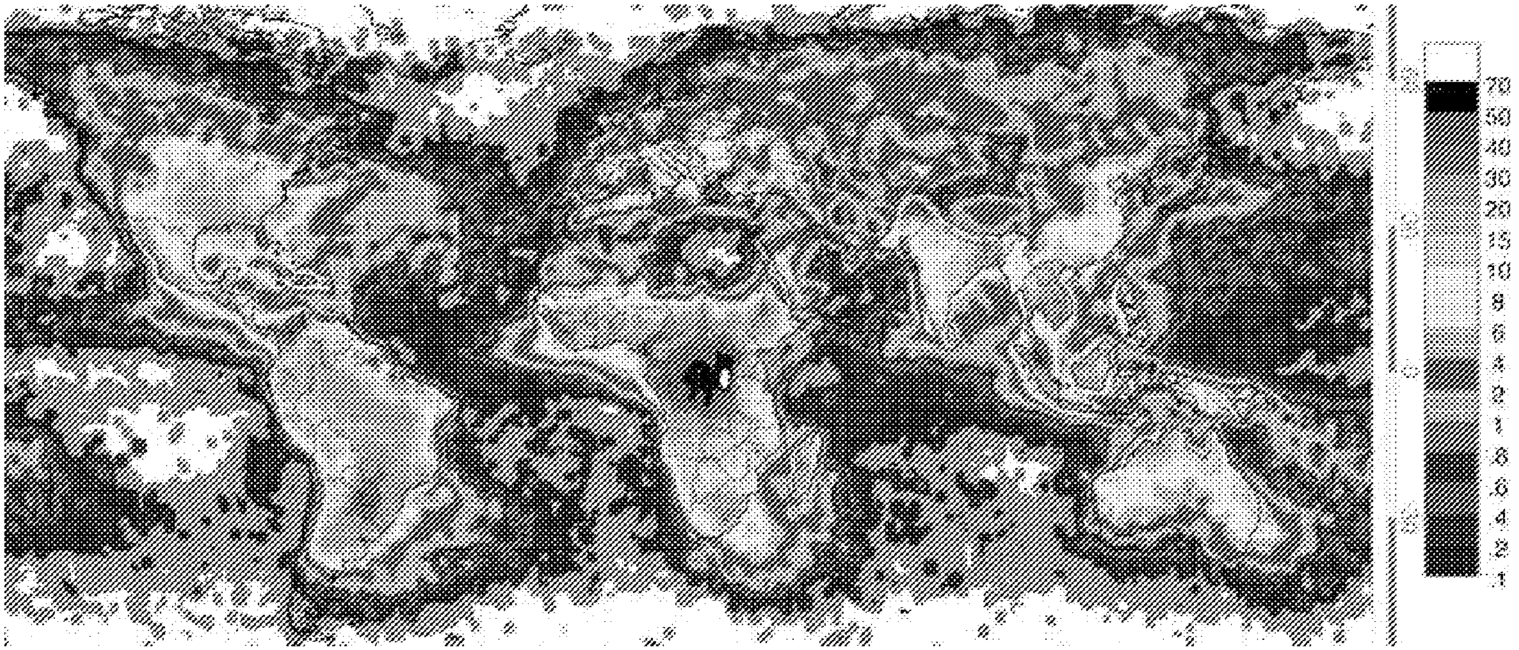
70
50
40
30
20
15
10
8
6
4
2
1
.8
.6
.4
.2
.1

Global Distribution of Lightning Strikes: April 1995 – February, 2003. (Source: NASA)

Fig. 1c

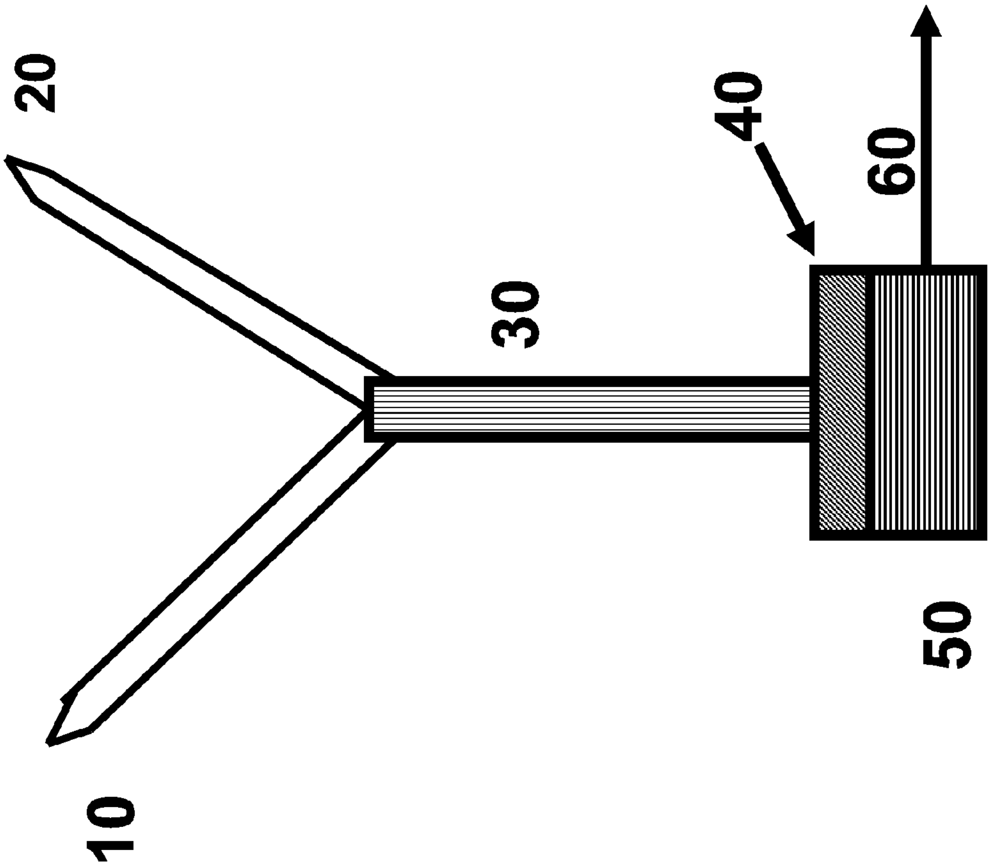
20
10
30
40
60
50

Fig. 2

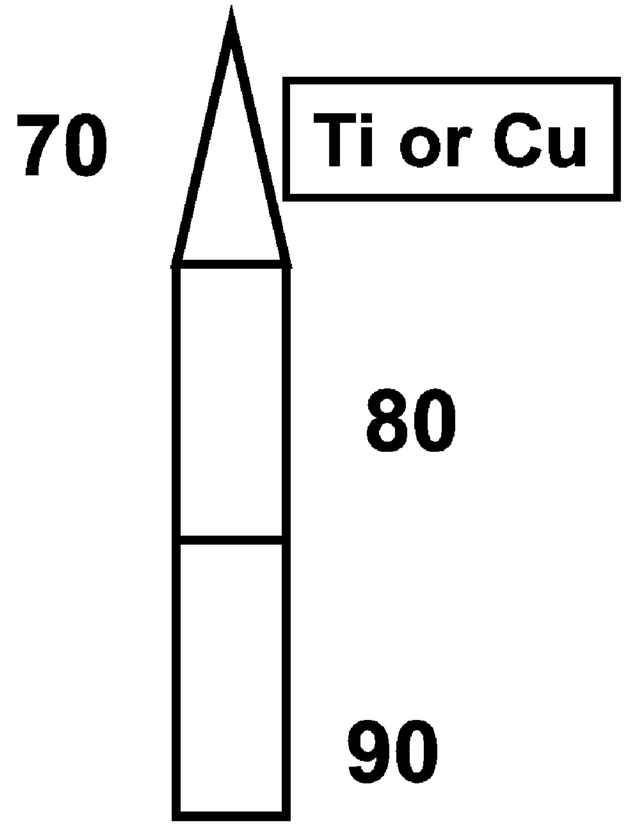
70
Ti or Cu
80
90

Fig. 3

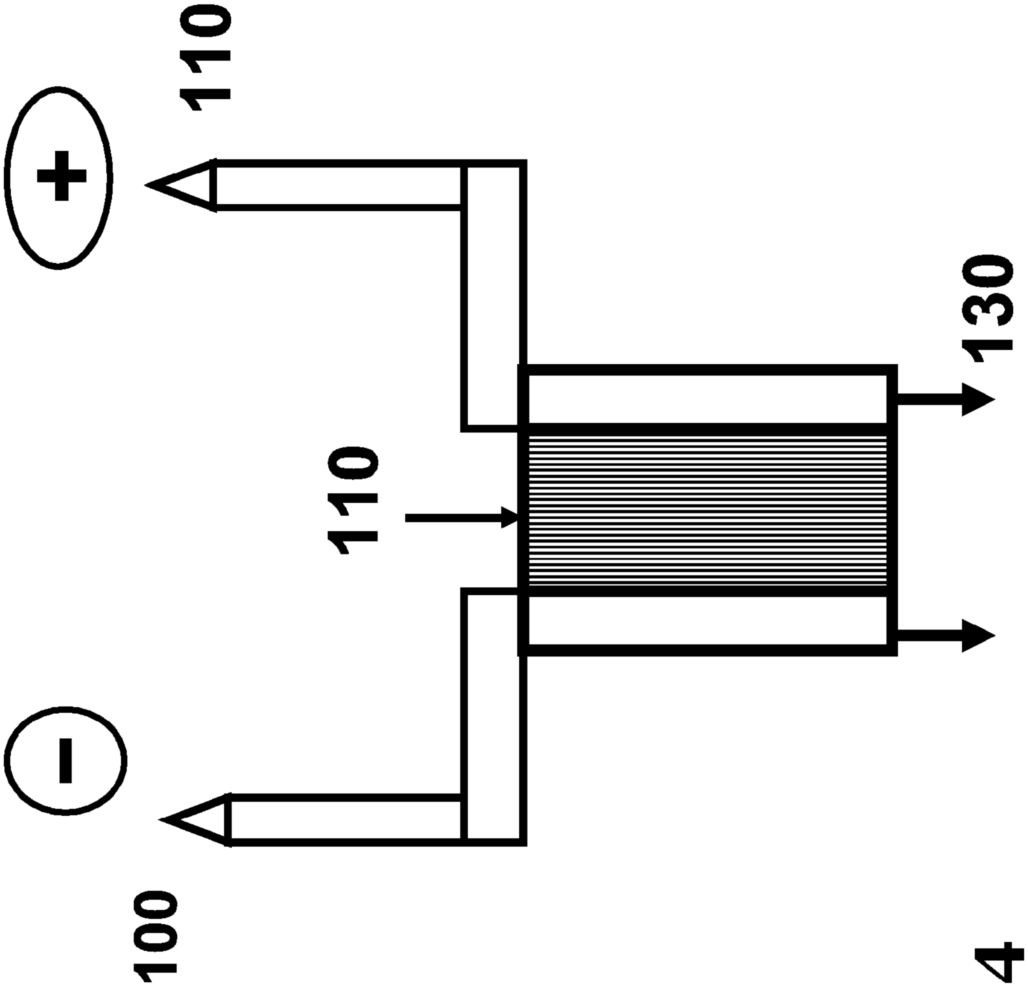
+
110
110
130
−
100

Fig. 4

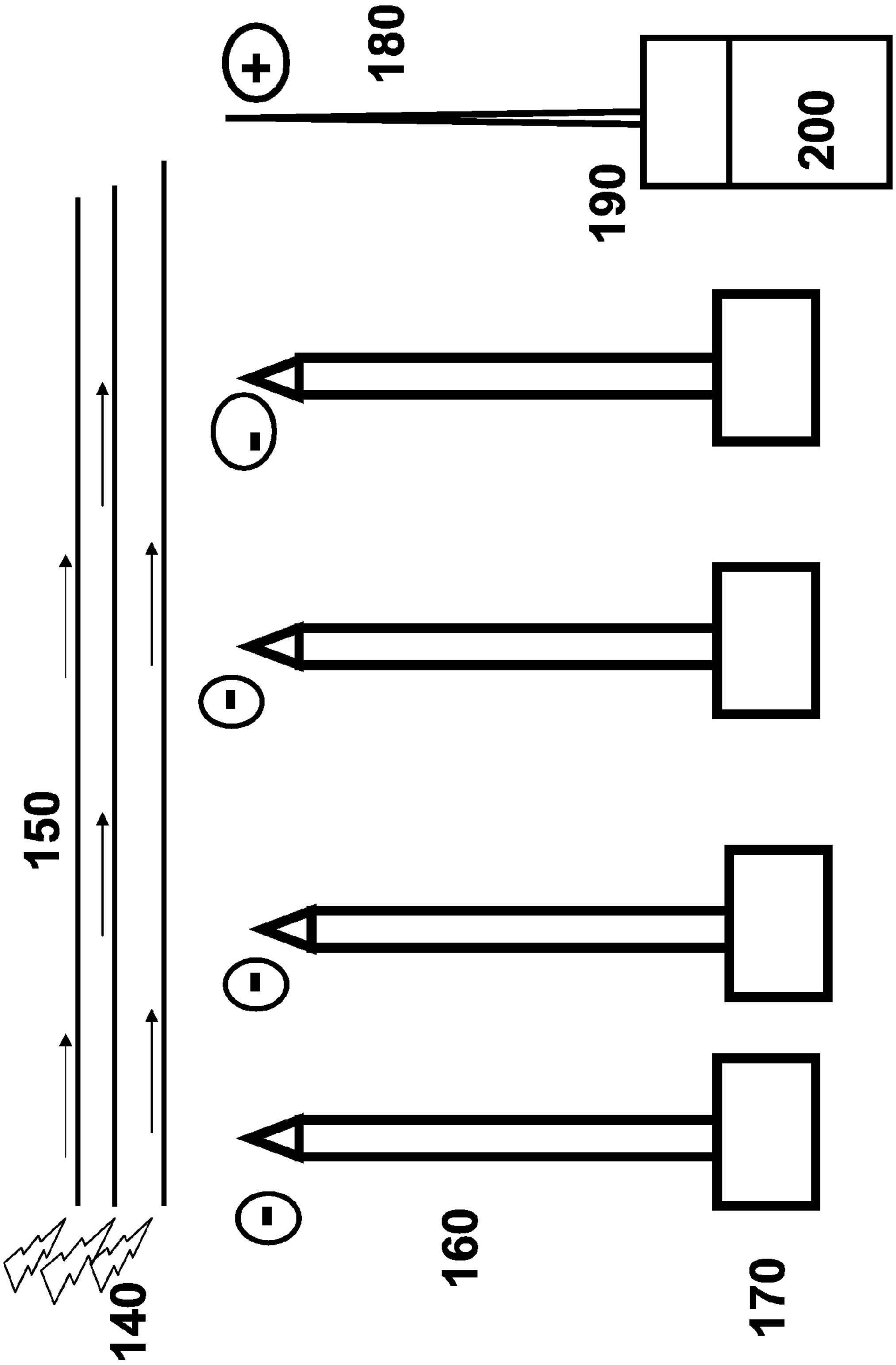
140
150
160
170
180
190
200
+
-
-
-
-

Fig. 5

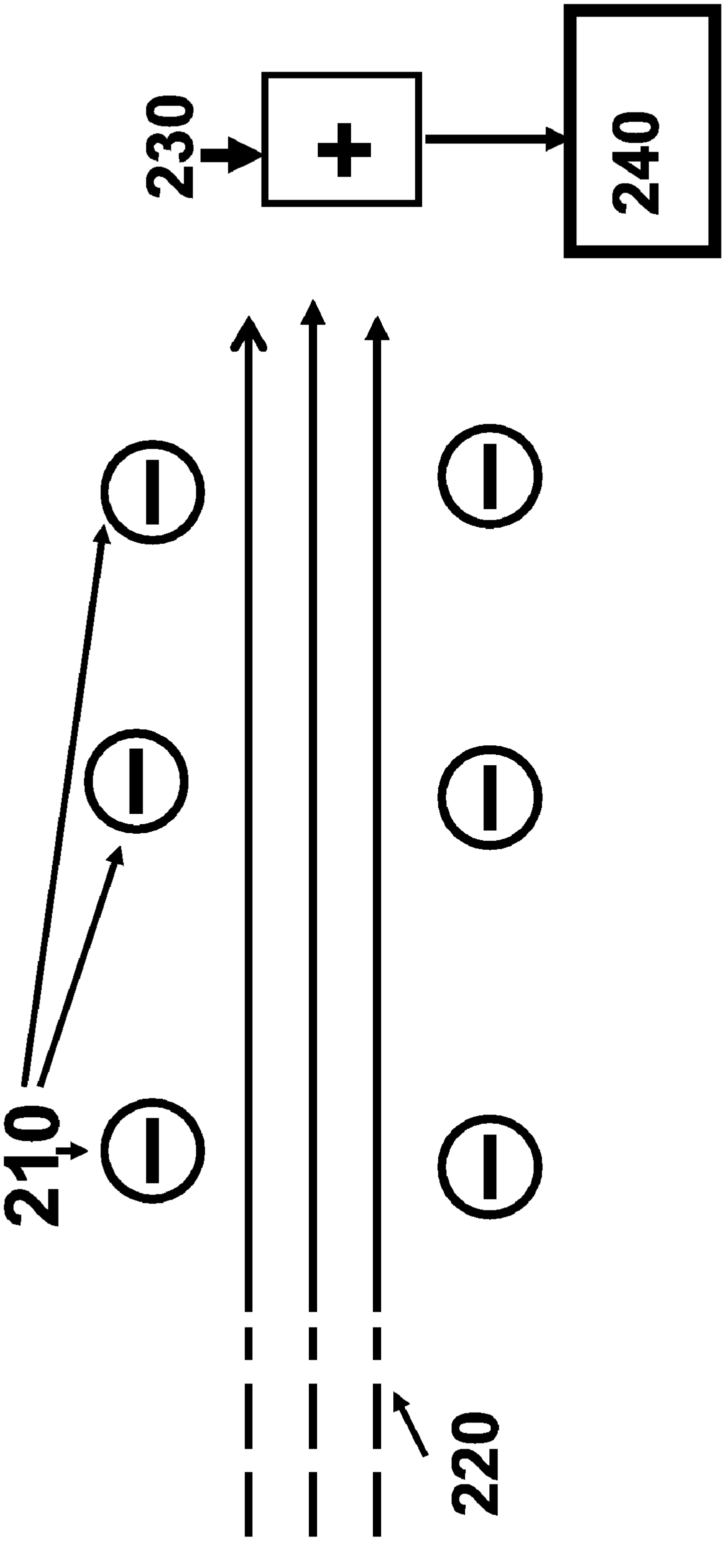
230
+
240
210
220

Fig. 6

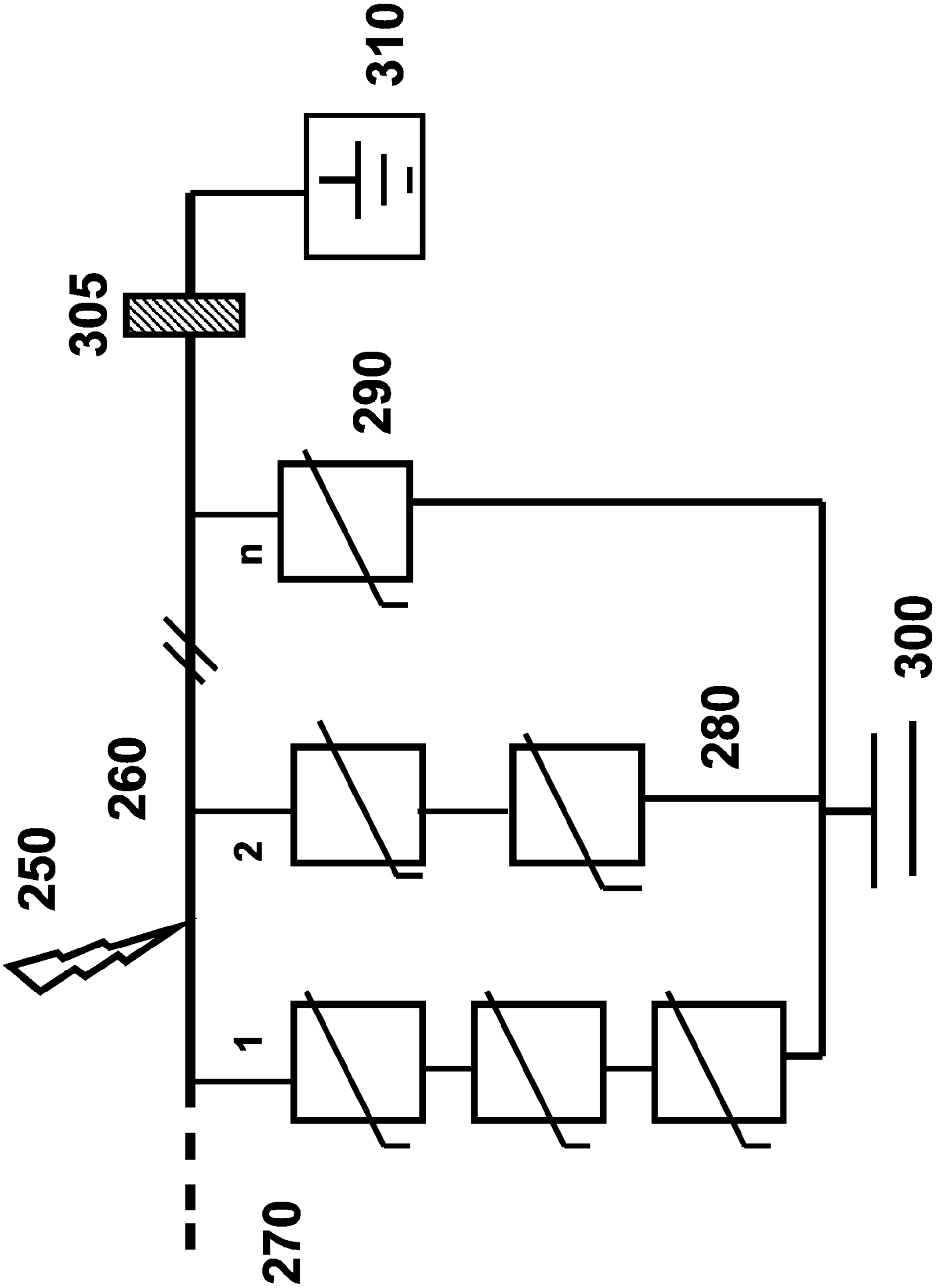
310
305
290
n
300
280
260
2
250
1
270

Fig. 7a

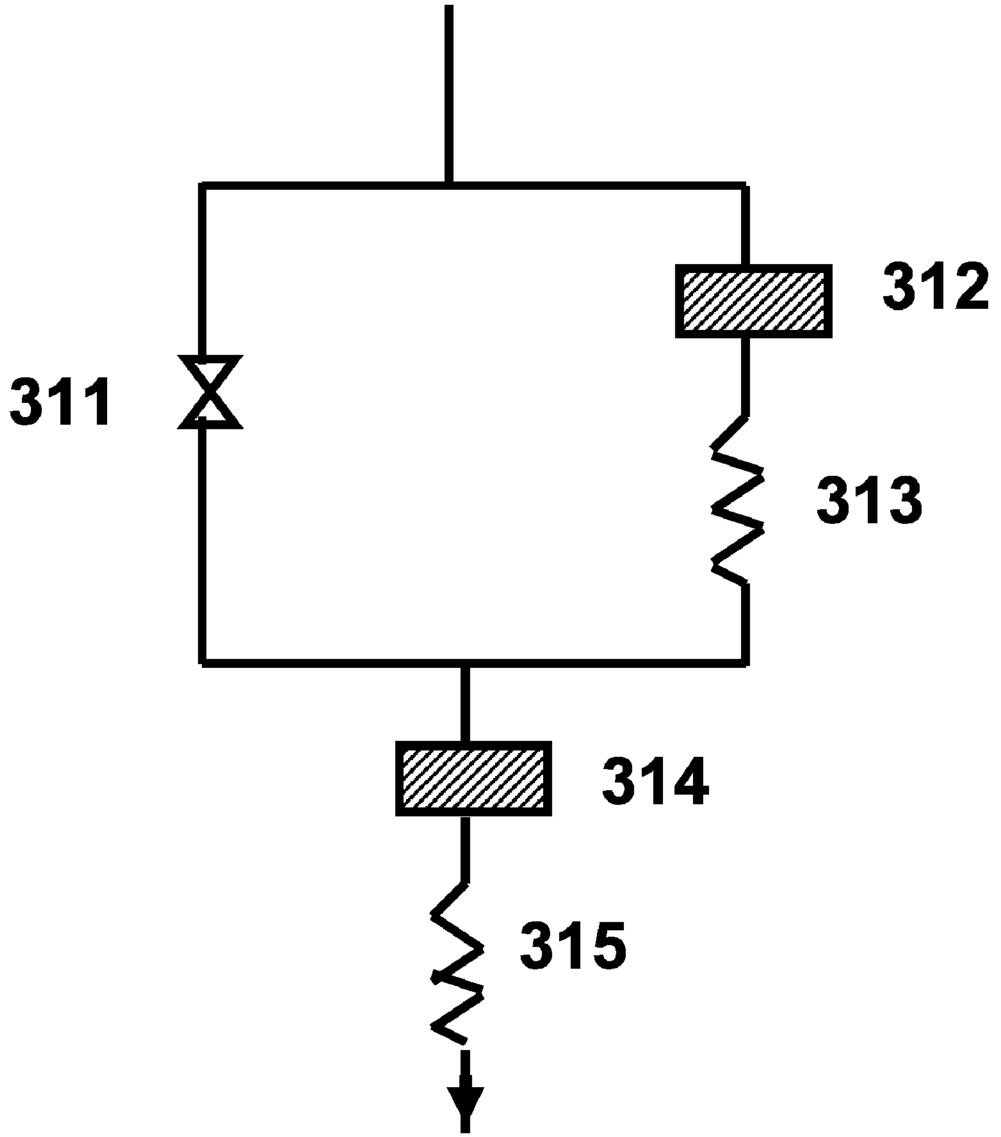
312
311
313
314
315

Fig. 7b

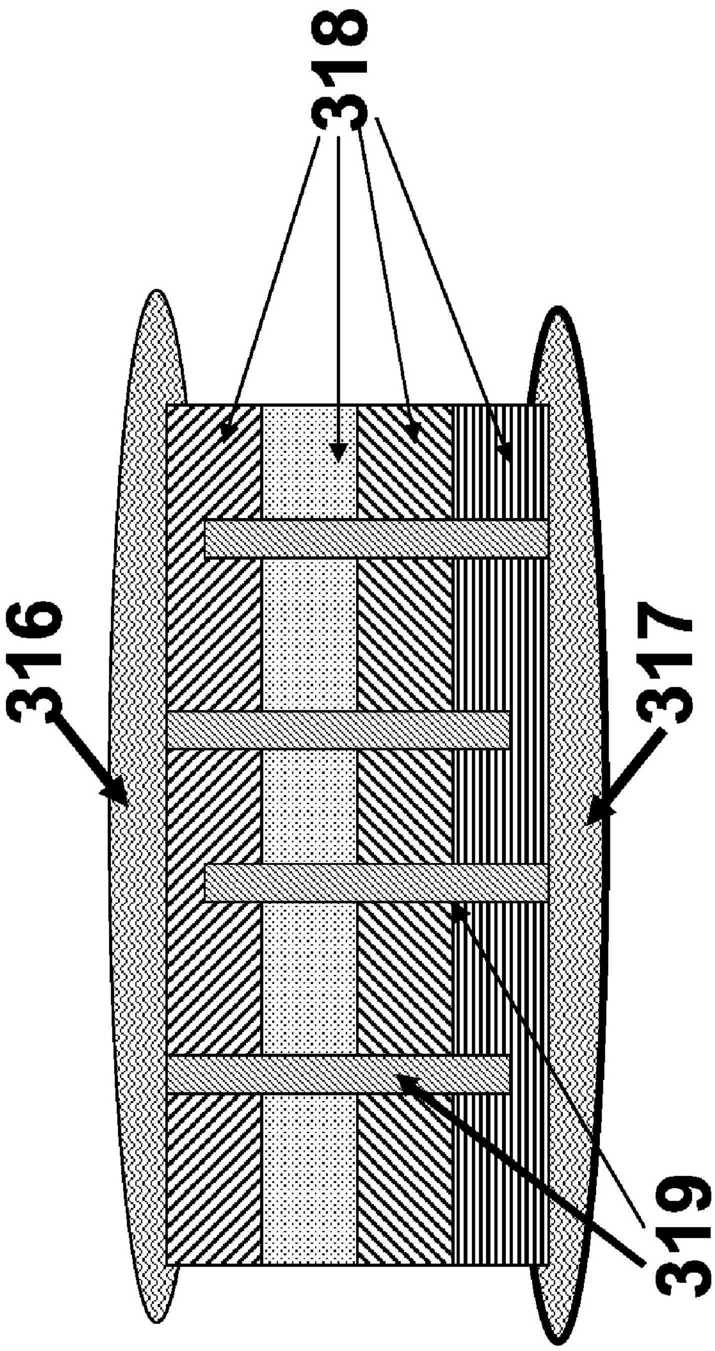
318
316
317
319

Fig 7c

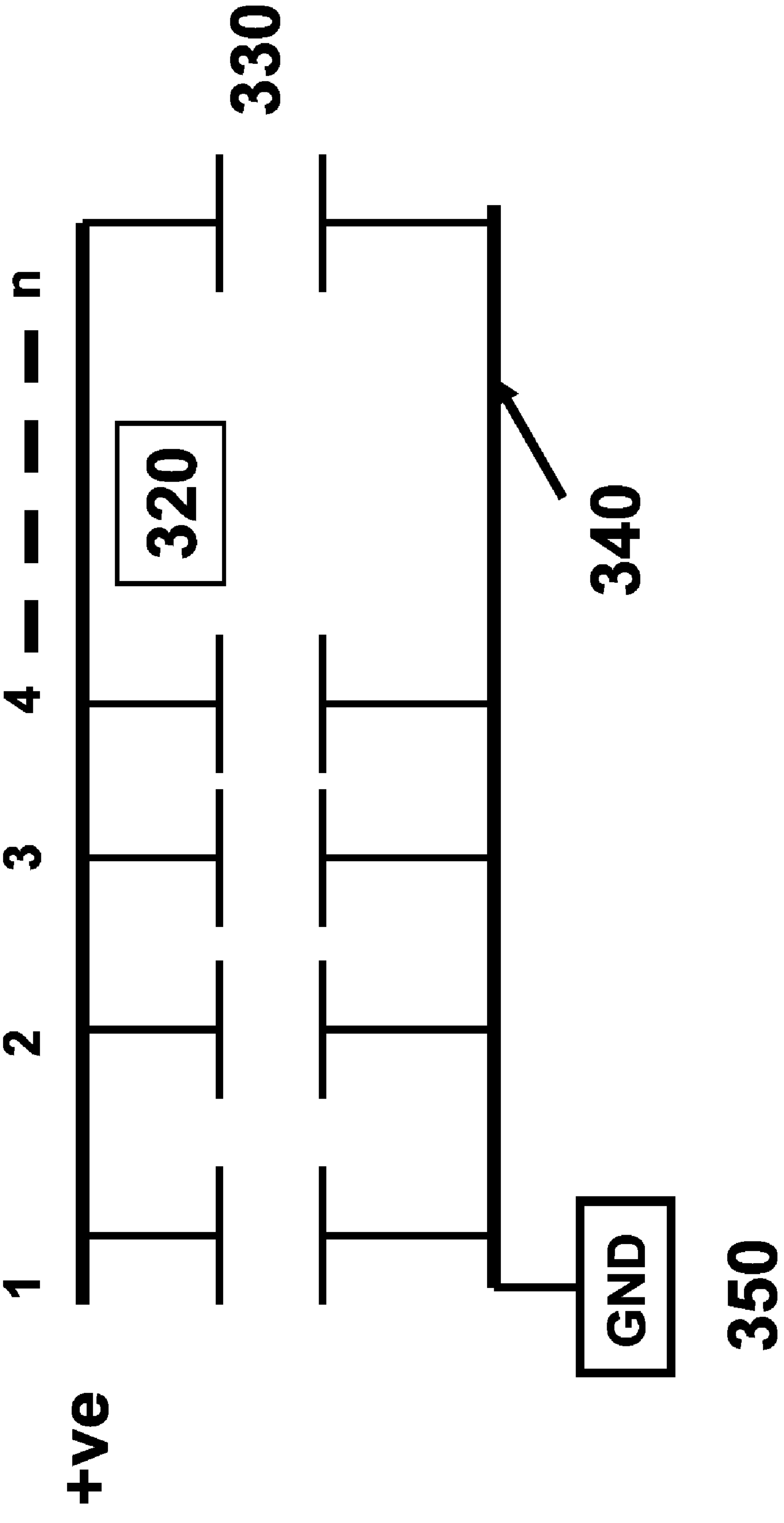
+ve
1
2
3
4
n
320
330
340
GND
350

Fig. 8a

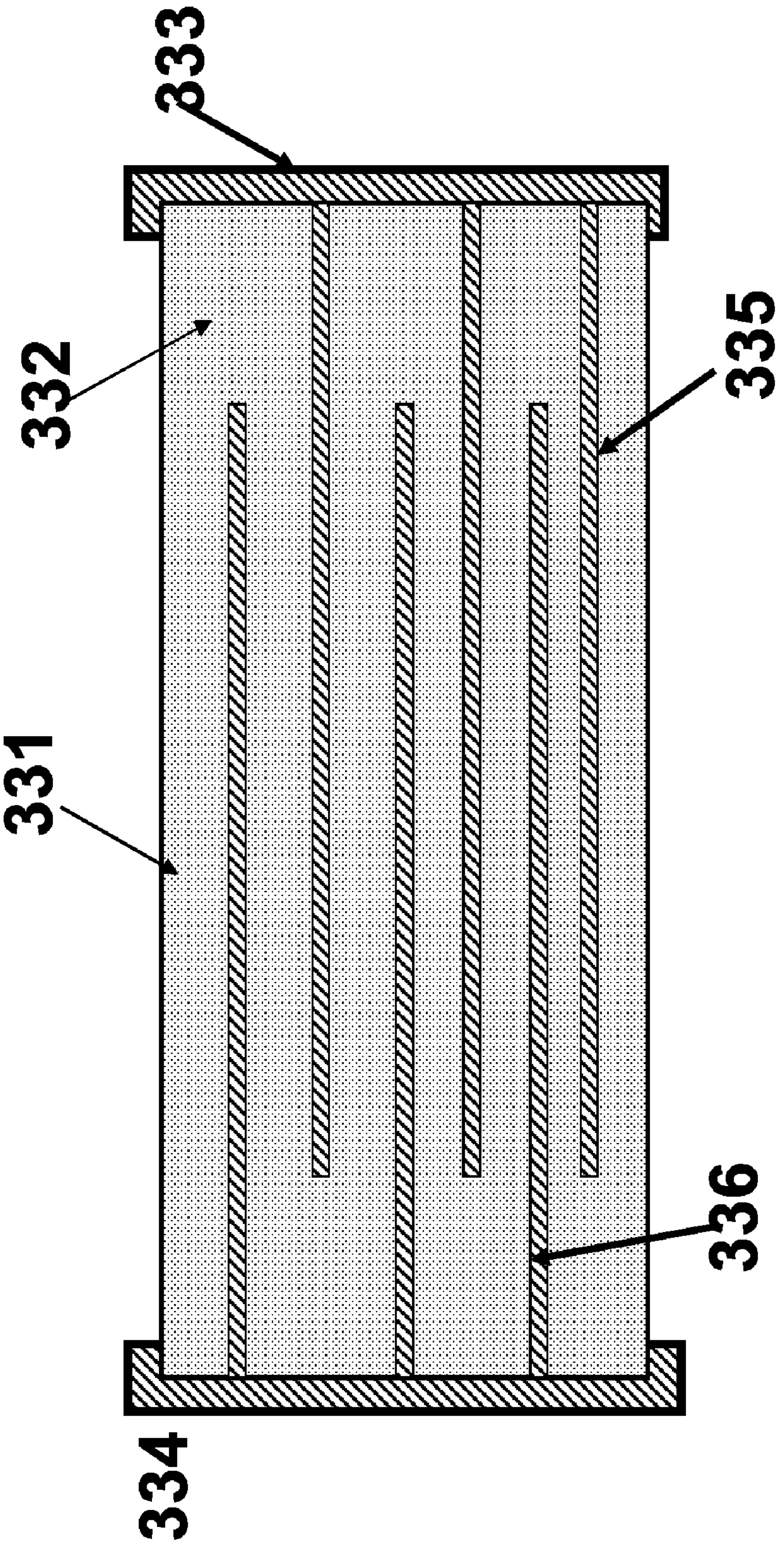
333
332
335
331
336
334

Fig. 8b

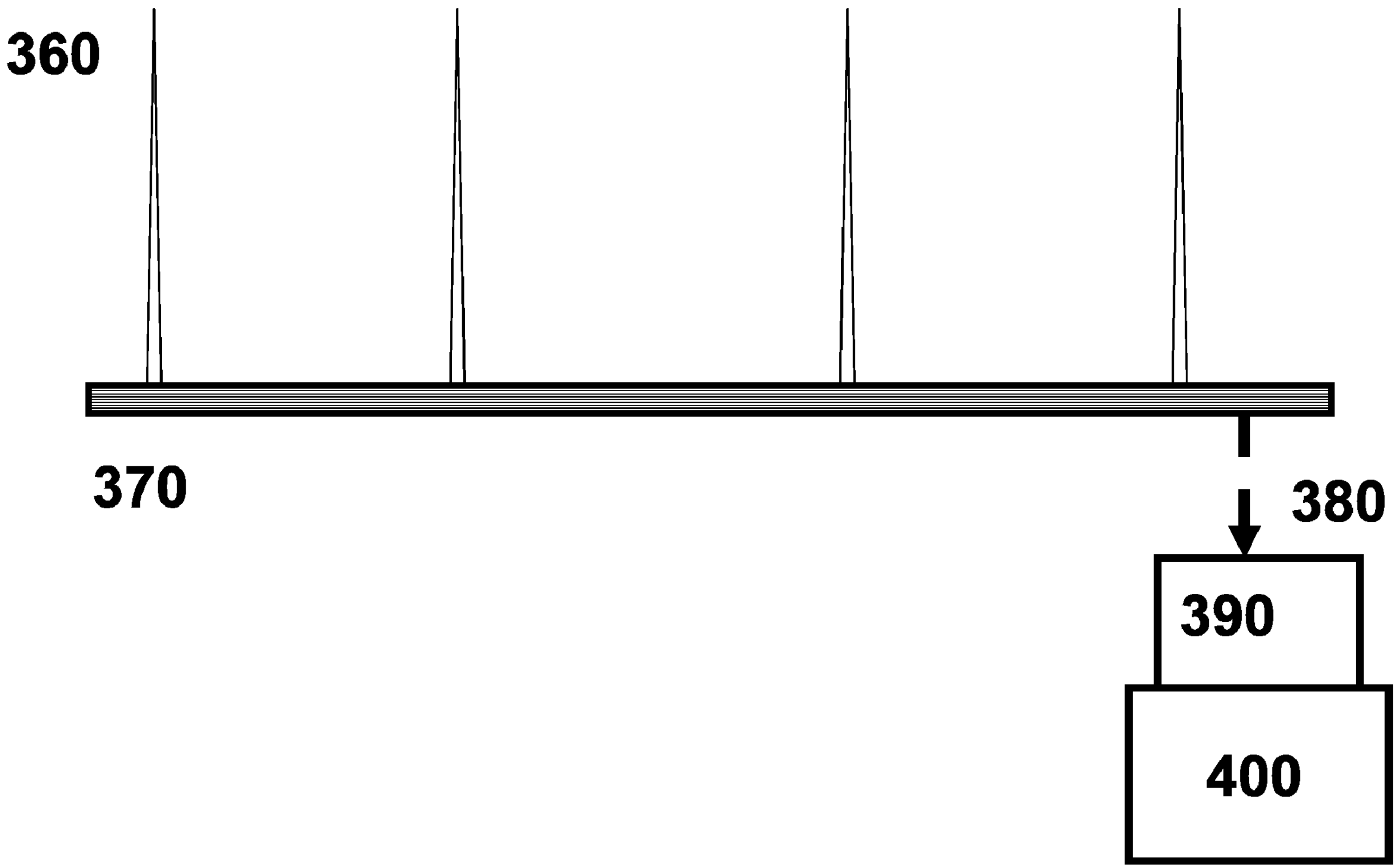
360
370
380
390
400

Fig. 9

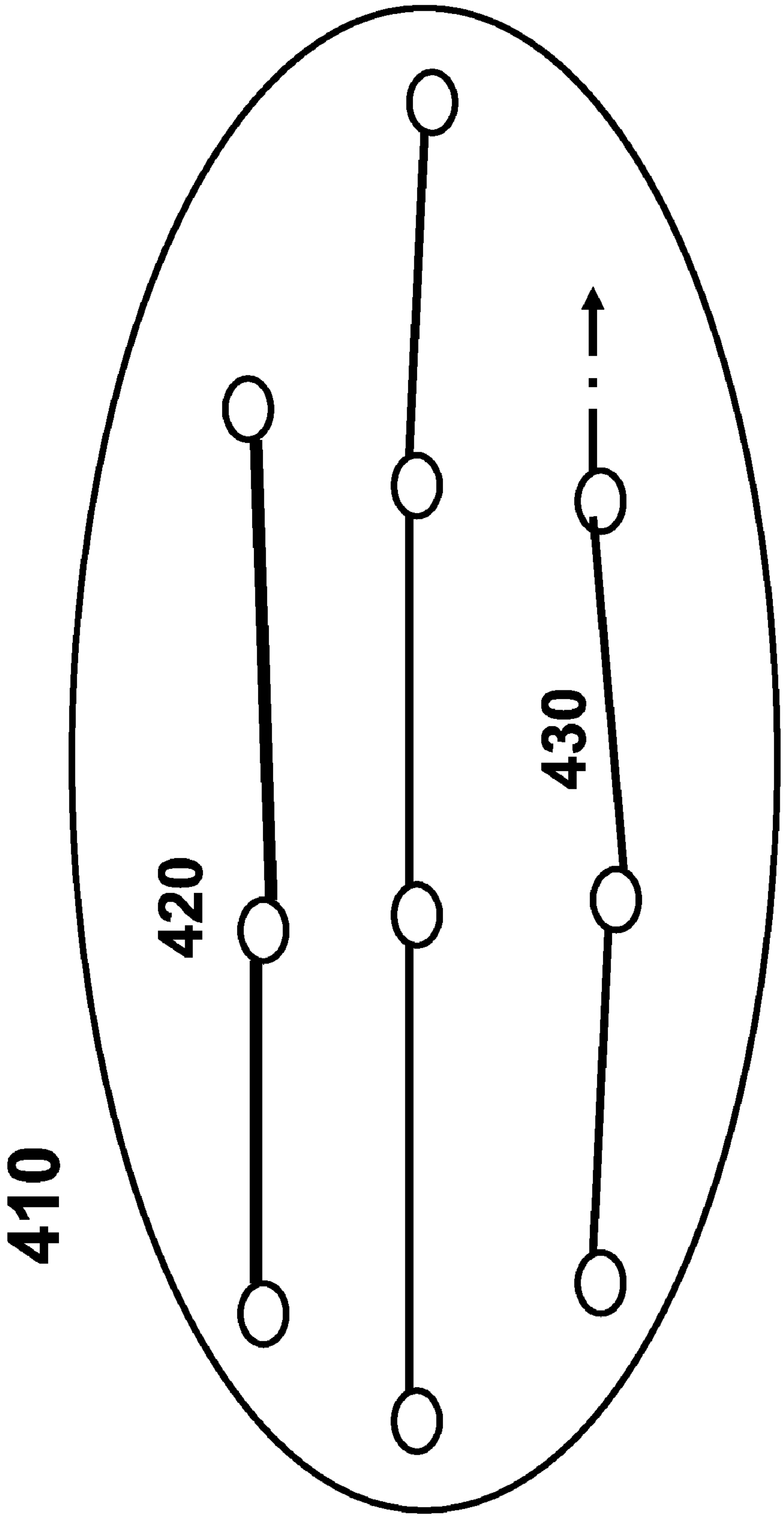

Fig. 10

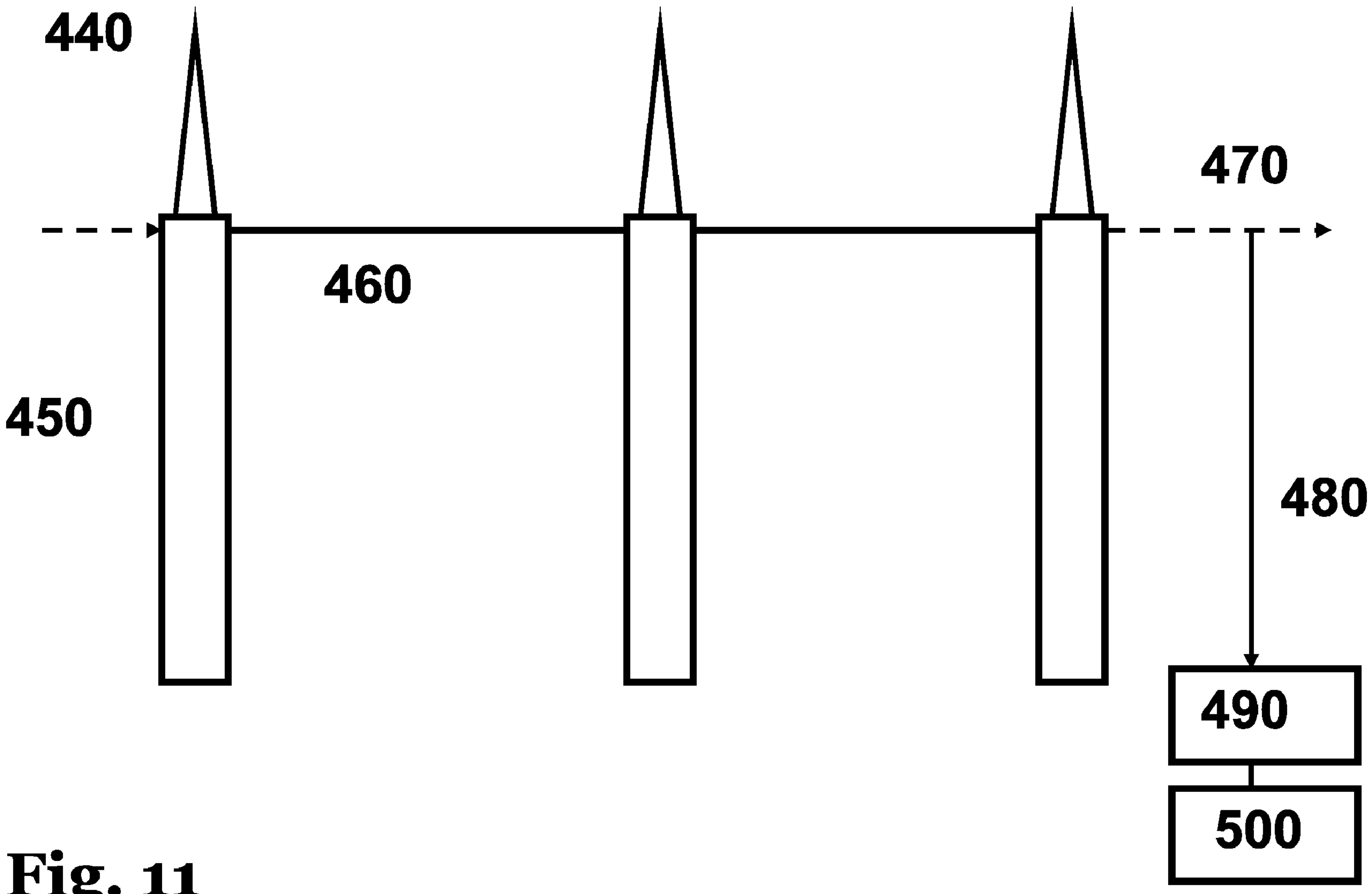
440
470
460
450
480
490
500

Fig. 11

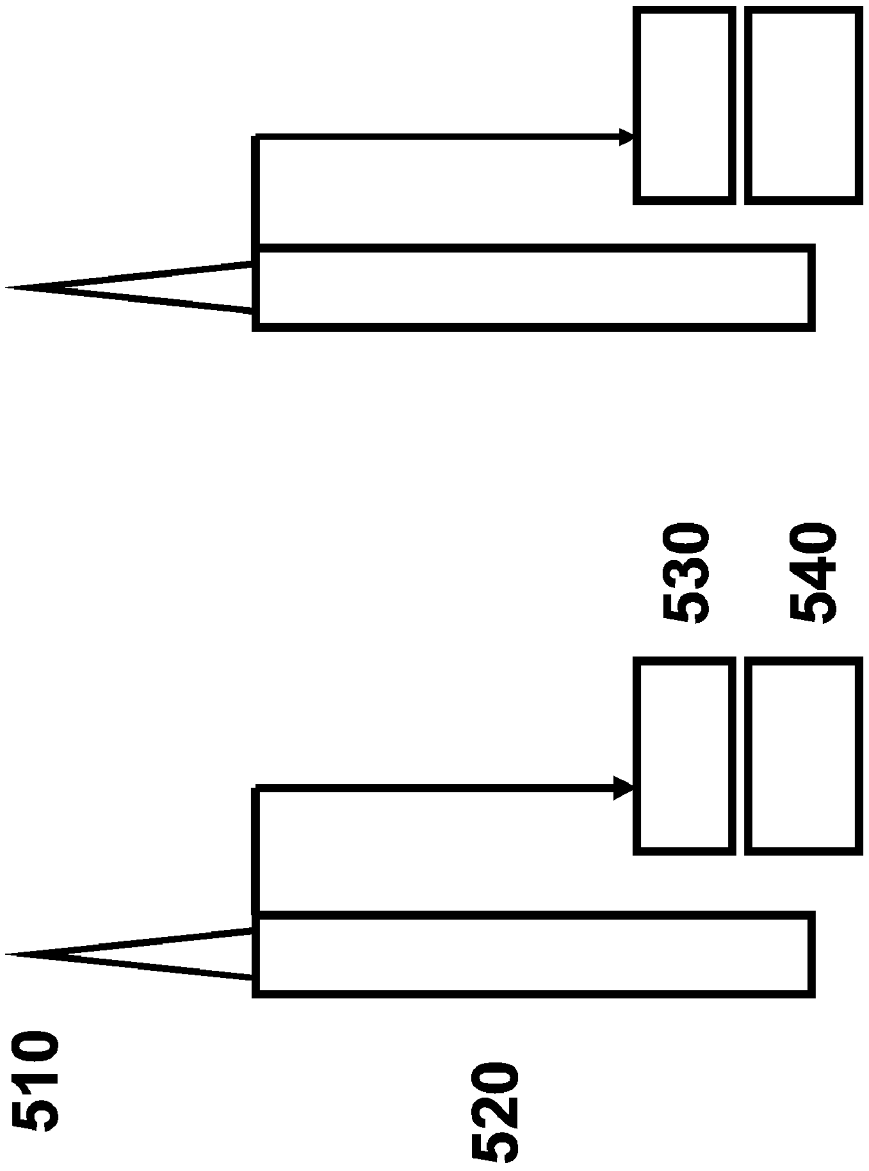
510
520
530
540

Fig. 12

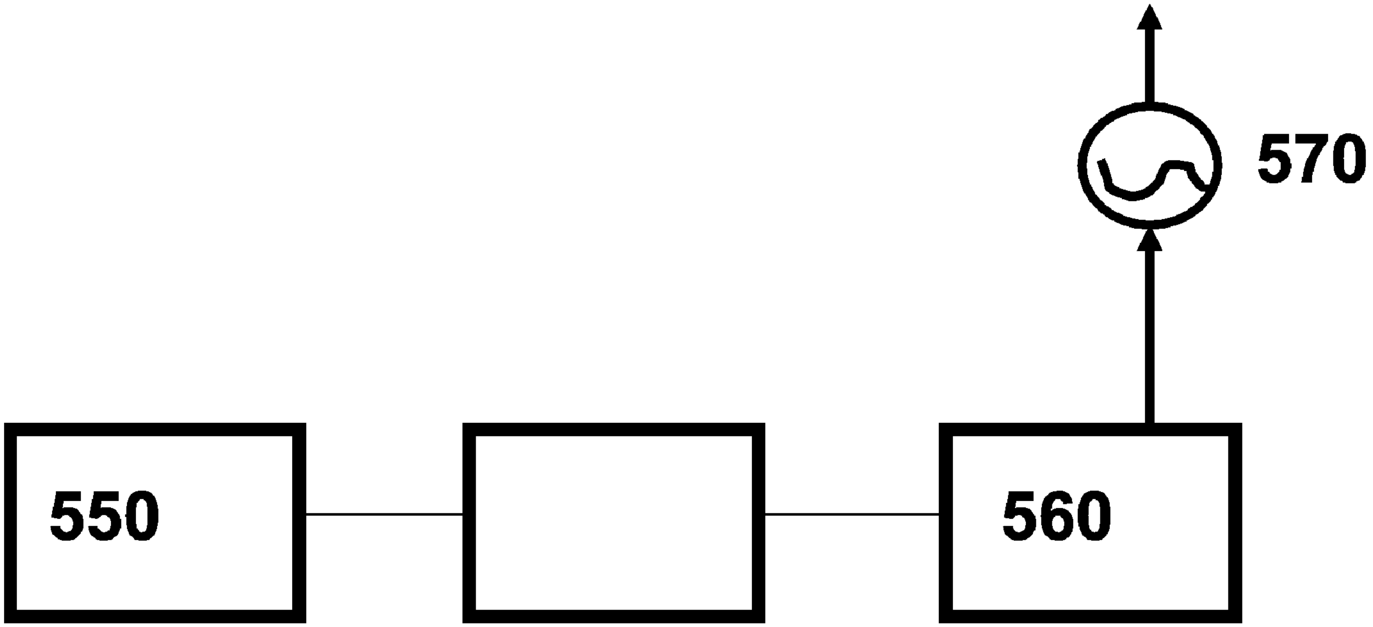

Fig. 13

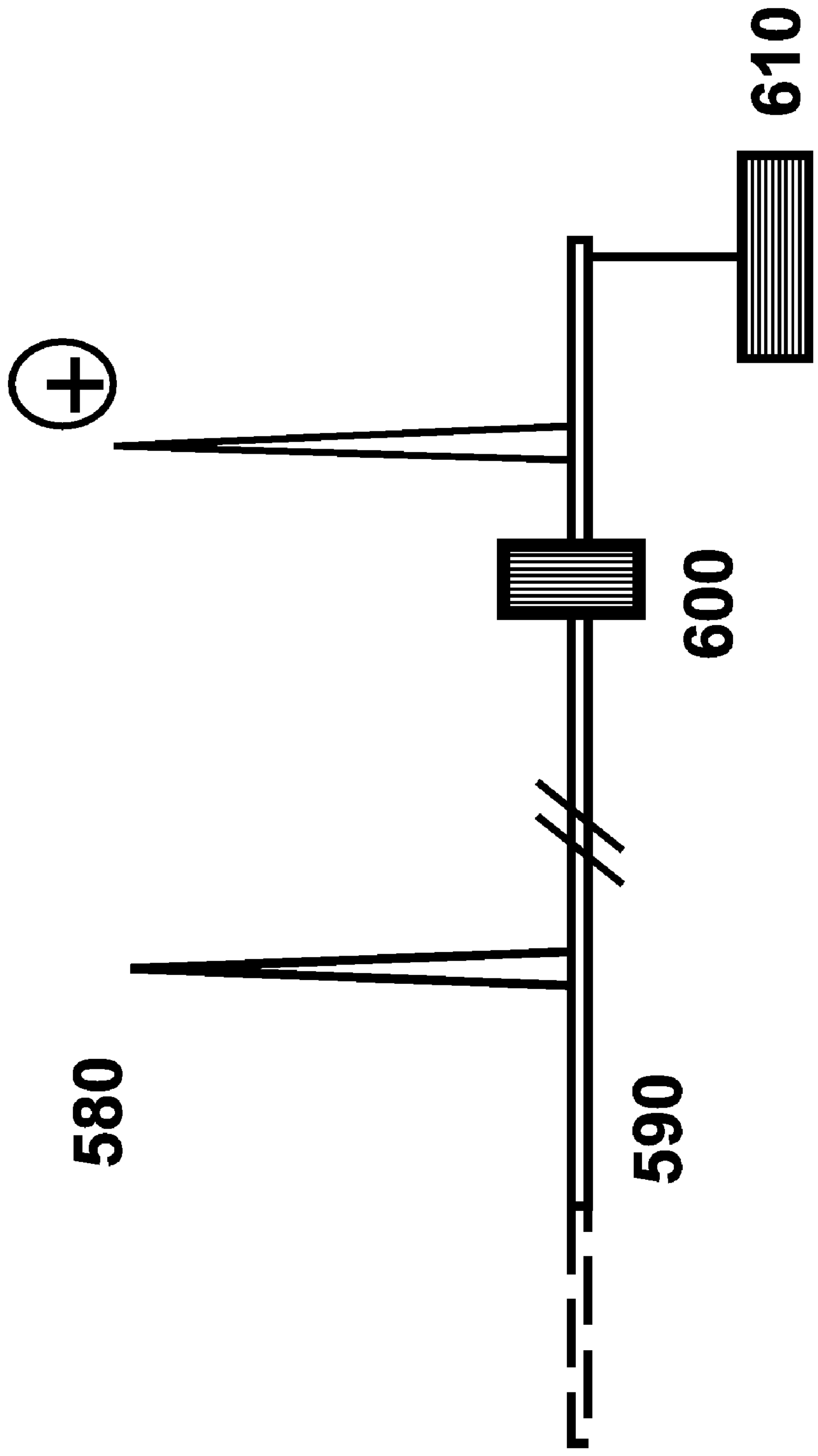
610
+
600
580
590

Fig. 14

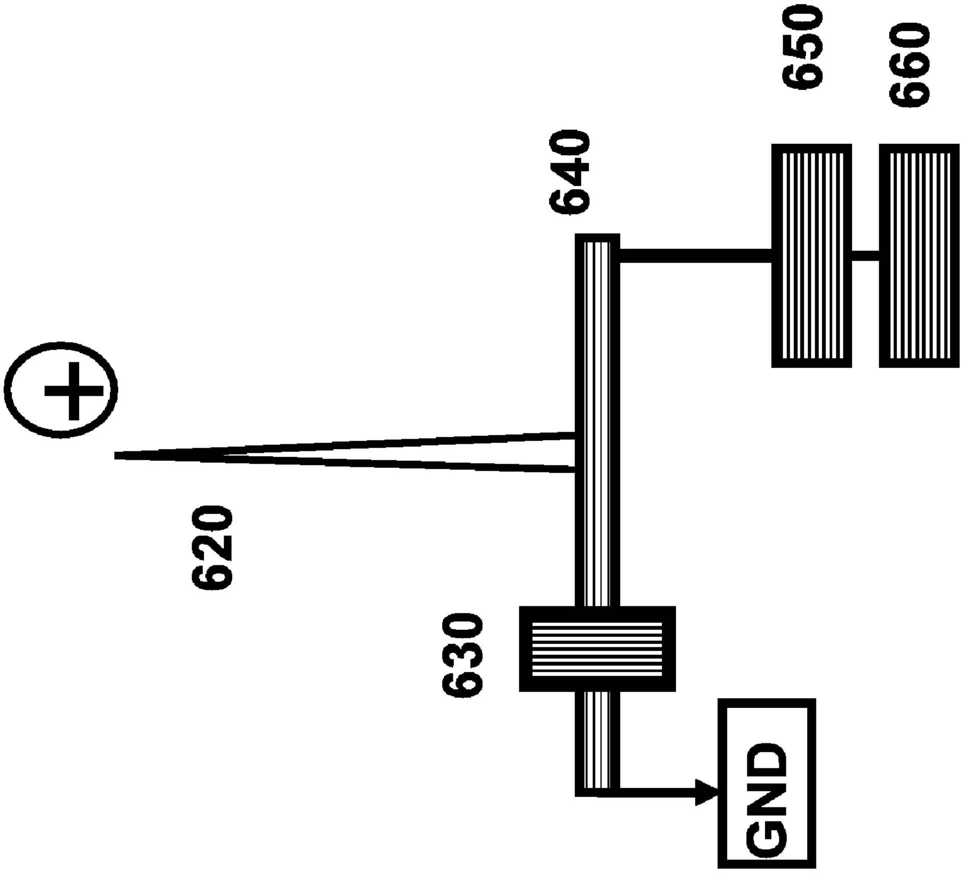
650
660
640
+
620
630
GND

Fig. 15

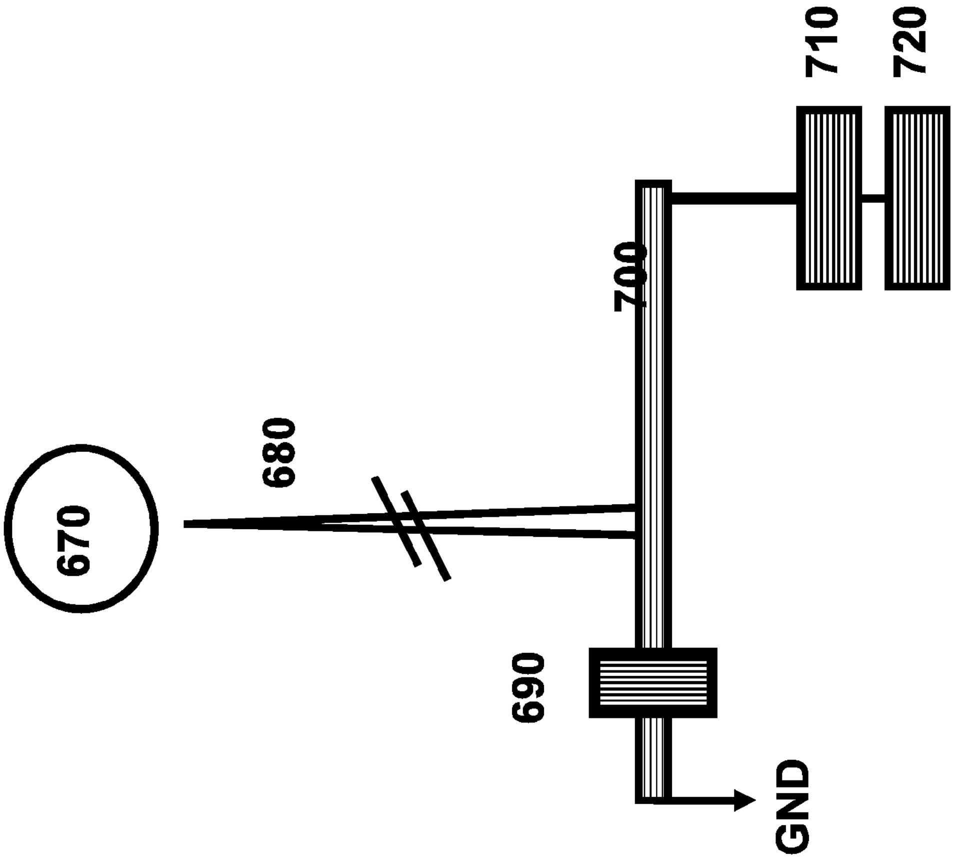
670
680
690
700
710
720
GND

Fig. 16

METHOD OF ATMOSPHERIC DISCHARGE ENERGY CONVERSION, STORAGE AND DISTRIBUTION

TECHNICAL FIELD

[0001] This invention relates to the field of electrical energy generation and storage in general and the conversion of atmospheric electrical discharge to a useable form of energy by arresting, storage and retransmission of lightning induced electrical discharge in particular.

BACKGROUND OF INVENTION

[0002] Lightning has always been a destructive force of nature. Causing death by electrocution and starting fires both in residential areas and in forests. In either case the loss of life and property is substantial. Lightning strikes have affected communication systems, and by uncontrolled arcing, electricity grids leading to loss of power with substantial economic loss

[0003] Lightning is the number one cause of forest and farm fires just in the United States and causes over 80 percent of all livestock losses due to accidents. Property losses run into billions of dollars annually and thousands of deaths and injury are reported yearly. Business losses from power and communications failure run into billions of dollars annually. Up to now, lightning strike damage has generally been controlled primarily by grounding.

[0004] A lightning discharge can have voltage of 10 MV to 100 MV and current of 1 kA to 300 kA. It packs so much voltage that it can leap a mile through the air and strike another object. Lightning can strike a building directly or leaping to it after striking a tree or other nearby object or by following a power line. It can strike any object that provides an easier path to the ground for it than the air. Because of the large network of lines, lightning can strike a utility pole at one point and be transmitted to any location in the network. Because of this, utility companies use multiple grounding and other protective devices per mile to ameliorate the lightning effects.

[0005] There has been a lot of activity in the alleviation of lightning damage but none in utilizing its energy potential.

[0006] U.S. Pat. No. 7,495,168 discloses a dipole lightning conductor which can gather electrical charges from the atmosphere during a thunderstorm, opposite to the polarity of the adjacent cloud on one side and induce a different polarity on the other end contacting the ground. When a sufficient potential difference is reached between the dipole and the earth, a dielectric breakdown occurs sending a large amount of earth charge to the thundercloud, inducing a thunderbolt.

[0007] In U.S. Pat. No. 6,012,330, Palmer discloses a method of inducing lightning strikes by shooting a stream of ionized water into a thundercloud and triggering electrical conduction through the ionized water column to the ground. In this invention, while ionized water is used, advantage is taken also of capillary wetting of conducting mesh cable at the tip of which is tethered a balloon of conductor coated Mylar. It is easier to float a balloon than to pump a column of water to any significant height.

[0008] In U.S. Pat. No. 6,320,119 Gumley discloses a capacitively coupled composite air terminal with sharp and rounded end components for voltage manipulation to minimize electric field ahead of streamers to minimize corona discharge.

[0009] Kato, in U.S. Pat. No. 5,280,335, discloses a compound lightning low voltage arrester with high energy tolerance for power and telephone lines comprising a spark gap is connected to a serial combination of Zinc Oxide arrestor and a non-inductive resistor. By cooperative execution between the spark gap and the Zinc Oxide arrester, the Zinc Oxide is chosen to be of a rating such that thermal breakdown does not occur before the spark gap ignition voltage is reached. The arrangement ensures reliable and predictable discharge and arrester survival even under multiple lightning strikes of close proximity.

[0010] Subbarao in U.S. Pat. No. 4,338,648, discloses a discharge counter and arrester current meter to determine arrester steady state current and the number of discharge events occurring during a thunder storm lightning arrester event.

[0011] Greenwald and Moses in U.S. Pat. No. 5,610,813 disclose a thunderstorm cell detection and mapping system for acquiring localized lightning strike information and identifying and locating active thunderstorm cells.

[0012] In U.S. Pat. No. 4,272,720, a method of differentiating between cloud-cloud and cloud-ground discharges is disclosed. Indicating discharge events with no accompanying lightning ground strike. This indicates that even in the absence of streamers, there is significant electrical discharge that can be converted and stored.

[0013] Weir and Nelson disclose in U.S. Pat. No. 7,033,406, a high energy density ceramic capacitor with energy capacity of 52 kW.h in 2005 cu.in, the equivalent of 1600 kW.h/m^3. Hansen in U.S. Pat. No. 6,078,494 discloses multilayer ceramic capacitors comprising doped barium-calcium-zirconium-titanate dielectric, the materials basis for U.S. Pat. No. 7,033,406. It is claimed to be characterized by high dielectric constant, high stability of capacitance value, long service life, low loss factor, high insulation resistance capacitance, low voltage dependence, and wide temperature range stability. Electrodes of base metal alloys from the group of Ni, Fe, Co or their alloys are claimed to be perhaps just as effective as noble metals containing gold, silver, platinum and palladium and may also contain Cr, Ti, Zr, V, Al, Zn, Cu, Sn, Pb, Mn, Mo and W.

[0014] Although the energy density capacity is not disclosed, it may be in the same range as U.S. Pat. No. 7,033,406 showing a path for storing the energy that can be extracted from lightning discharge.

[0015] U.S. Pat. No. 5,361,187, claims dielectric constant (k) of up to 19000 using similar group of dielectrics and processes. In U.S. Pat. No. 5,604,167 dielectric constant values of between 11,000 and 14,000 are disclosed by the same inventor with equivalent materials and processes.

[0016] Since these capacitors use standard powder processing technique, it can reasonably be expected that equivalent characteristics may be observed with naturally occurring doped silicates such as phyllosilicates for example, montmorillonite: $(Na,Ca)_{0.33}(Al,Mg)_2(Si_4O_{10})(OH)_2.nH_2O$, or cyclosilicates such as Benitoite —$BaTi(Si_3O_9)$ as base materials appropriately calcined and doped.

[0017] In U.S. Pat. No. 6,078,494 multilayer ceramic capacitors comprising doped barium-calcium-zirconium-titanate dielectric is disclosed. It is claimed to be characterized by high dielectric constant, high stability of capacitance value, long service life, low loss factor, high insulation resistance and low voltage dependence, and wide temperature range stability. Electrodes of base metal alloys from the group of Ni, Fe, Co or their alloys are claimed to be perhaps just as effective as noble metals containing gold, silver, platinum and palladium and may also contain Cr, Ti, Zr, V, Al, Zn, Cu, Sn, Pb, Mn, Mo and W. In U.S. Pat. No. 5,361,187, the inventors claim dielectric constants of up to 19000 using similar group of dielectrics and processes. In U.S. Pat. No. 5,604,167 dielectric constant values of between 11,000 and 14,000 is disclosed by the same inventors with equivalent materials and processes. Since these capacitors use standard powder processing technique, it can reasonably be expected that equivalent characteristics may be observed with naturally occurring doped silicates such as phyllosilicates for example, various grades of montmorillonite: $(Na,Ca)_{0.33}(Al,Mg)_2(Si_4O_{10})(OH)_2.nH_2O$, or cyclosilicates such as Benitoite—$BaTi(Si_3O_9)$ as base materials appropriately calcined and doped.

[0018] Although the technology to arrest lightning strikes has been around for at least a hundred years, this invention to harness and store that power has only become viable due to recent advances in super-capacitor technology which has made possible the fabrication of ultrahigh capacitance capacitors with very long lifetimes, high power, high energy densities and very fast charge and discharge rates.

[0019] Just as flood water, another destructive force of nature in times past was controlled by building dams, and from the dams eventually came hydroelectric power; and wind including thunderstorms has now been harnessed to generate electricity, this invention describes the technology to harness the power of atmospheric discharges including lightening storms to generate useable electrical energy.

[0020] This disclosure also introduces the concept of an Energy Dam. Facilities of substantial energy storage capacity capable of receiving energy feeds from various energy sources such as Solar, Wind, Thermal Power plants and Lightning. For example, during a thunderstorm, electricity generated by wind and lightning is collected and stored.

[0021] Lightning strikes every part of the globe but not uniformly. The regions with the highest historical concentration of lightning strikes are shown in FIG. **1***c*. These include Florida and the Gulf Coast in the Americas, the Equatorial Highlands of DRC, Rwanda and Burundi in Central Africa and the Monsoon Belt in Asia.

[0022] Except for the Americas, typically, these regions have very little electricity infrastructure. With the capability disclosed here, substantial reserves of electricity can be generated, stored and possibly traded.

[0023] In regions of epileptic electricity supply, disruptions in power do not necessarily translate into disruption in service. Since when power is generated, it is fed into these energy dams or local storage units for eventual use. Just as in water supply, the customer is not necessarily conscious of when the local water company is pumping water into storage tanks or local dams, in energy dams, the customer need not be conscious of whether or when electricity is being generated by the utility company because the power to his home is drawn not necessarily live from the grid but from the local energy storage units or energy dams.

BRIEF SUMMARY OF THE INVENTION

[0024] A method of harnessing atmospheric discharge and storing said energy in a useful form for subsequent use is presented.

[0025] Since the voltage of a lighting bolt can be as high as 100 MV, a methodology for safely channeling this discharge for storage is disclosed.

[0026] It is the also the object of this invention to present a storage medium in which an energy capacity of 1 MW.h in a volume of under 1 m^3 can be achieved.

[0027] It is also a further objective of this invention to show how lightening strikes on tall buildings can be harnessed to provide energy for the building either directly or fed into the grid.

[0028] It is also an objective of this invention to present a methodology for harnessing lightening, heretofore a source of power disruptions for electricity grids and converting it into a useable form of energy and feeding it back into the grid.

[0029] Since each electricity grid has a lightning suppression unit, a methodology for harnessing the impulse from these units from lightening strikes and feeding it back into the grid is disclosed.

[0030] a method of deploying a network of said storage units along a utility grid is also disclosed.

[0031] It is also the objective of this invention to disclose a method of modifying the storage units for single building deployment.

[0032] A method of deploying said storage units as vehicular charging centers even in locations not readily served by a utility grid is also disclosed.

[0033] A method of manipulating the direction and movement of the lightening bolt in open air locations and redirecting them to predetermined collection points is disclosed.

[0034] Lightning storms are some of the major causes of wild fires in the world. A method of redirecting the electrical discharges in any given space via a network of polarized and non-polarized collectors is presented. The discharges can be stored, fed into the grid or grounded.

[0035] A method for non-spontaneous discharge generation and storage is disclosed. And a method for adapting the methodology for electric automotive deployment is also disclosed.

BRIEF DESCRIPTION OF DRAWINGS

[0036] For understanding of the present invention reference is made of the accompanying drawings in the following DETAILED DESCRIPTION OF THE INVENTION. In the drawings:

[0037] FIG. **1***a*. Picture of lightning strike over a city.

[0038] FIG. **1***b*. Picture of lightning strike over open plain.

[0039] FIG. **1***c*: Global Distribution of Lightning Strikes: April 1995-February, 2003. (Source: NASA).

[0040] FIG. **2** shows a schematic of discharge collection and storage system.

[0041] FIG. **3** shows a schematic of collection unit showing metallic composition variation.

[0042] FIG. **4** Shows a Schematic of composite collector showing polarized ends.

[0043] FIG. **5** Schematic of collector network configuration with negatively polarized ends to manipulate and redirect lightning electrical discharge away from the earth for eventual collection at a predetermined location.

[0044] FIG. **6** Plan view of polarized collection network as shown in FIG. **5** acting to concentrate and modulate the electrical discharge and redirect such for collection at predetermined location.

[0045] FIG. **7***a*. Schematic showing a representation of a network of varistors arranged in rows and columns with the net effect of applying resistive arrestors in parallel and in series to effect voltage mitigation of atmospheric discharges prior to collection and storage.

[0046] FIG. **7***b:* Schematic of varistor of power mitigation unit showing cascade of ZnO arrestor in series with non-inductive resistor in parallel with spark gap.

[0047] FIG. **7***c:* Multilayer varistor with ZnO, ZnS and WO_3/WSi_2

[0048] FIG. **8***a*. Schematic showing cross section of a collection unit.

[0049] FIG. **8***b*. Schematic showing cross-section of multi-layer ceramic capacitor of **8***a*.

[0050] FIG. **9** Schematic of collection unit network mesh in forested area with air frames deployed on each tree as collection of trees linked in a mesh network.

[0051] FIG. **10** Plan-view schematic of collection unit network mesh of FIG. **9**

[0052] FIG. **11** Schematic showing collection unit deployment on electricity transmission poles linked in a network and fed into the electricity grid.

[0053] FIG. **12** Schematic showing collection unit deployment on light poles or wireless towers.

[0054] FIG. **13** Schematic showing network of linked collection units and storage units on transmission lines, light poles or wireless towers fed into the grid.

[0055] FIG. **14**. Schematic of a system for non-spontaneous atmospheric discharge generation and storage.

[0056] FIG. **15** Schematic of atmospheric discharge collection and storage unit deployable on automotive units.

[0057] FIG. **16** Schematic of balloon-assisted discharge collection and storage unit.

DETAILED DESCRIPTION OF THE INVENTION

[0058] This invention describes a method of harnessing the electrical energy in atmospheric discharges, including, but not limited to lightning, storing said electrical energy and transmitting said energy for use. To this end, as an illustration, the preferred embodiments are represented in the drawings shown in FIGS. **2** to **15**.

[0059] Since lightning discharge voltages range from 10 MV to 100 MV with current of 10 kA to 100 kA, the technology challenge will be in mitigating the ultrahigh power which range from 100 GW to in excess of 10,000 GW. Although individual strikes last less than 100 μsec, thunderstorms usually last longer than 30 min. in any cloud formation and location. Indicating that the energy that can be mined from any location in a thunderstorm can run in excess of 50 GW.H. Enough energy for 500,000 homes consuming 1000 kW.h per month. At $100 per 1000 kWh, the revenue equivalent is $5 million per thunderstorm per location.

[0060] In solar energy generation, the cost of a 1 MWe rating PV panel is about $1 million. To achieve 1 GWe, the cost implications will be about $1 billion. In fossil fuel plants, the construction cost is approximately $1.3 million per MWe, putting the cost for a 1 GWe at ~$1.3 billion excluding fuel, carbon footprint, and other operating cost. Though thunderstorms are not continuous, with storage capability, the cost basis for an atmospheric discharge energy source will be several orders of magnitude lower than fossil fuel or solar of any architecture.

[0061] One embodiment of this invention is depicted in FIG. **2**. Lightning is collected at the air terminals **10**, **20** and transmitted through the conductor **30** to an arrestor **40** and on to the capacitor **50** where the electricity is stored. From which it can be extracted **60** at point of use or transmitted into the grid.

[0062] The degree of atmospheric discharge charge collection can be influenced by the metallurgical composition of the collectors. FIG. **3** illustrates the fabrication of the collector with different metals of varying conductivities and atmospheric and electrical stability. In this embodiment, metallurgical compositions for the various segments comprise Ti, Cu, Al. Other metals or their alloys can be used depending on the required properties and the locations of deployment.

[0063] In yet another embodiment of the invention as shown in FIG. **4**, a composite collector with opposite polarity air terminals **100**, **110** is used. The air terminals are isolated from each other by an insulator **120**. The negative polarity terminal serves to deflect and manipulate the charge to the positive terminal from which it is conducted to the storage unit **130**.

[0064] Another embodiment of the invention is illustrated in FIGS. **5** and **6**. In this representation, negative polarity collectors, **160** deflect the atmospheric discharge so that they do not reach the ground but are deflected for collection at a predetermined terminal **180** of positive polarity and thereafter storage **200**. A plan view representation of this embodiment is shown in FIG. **6**.

[0065] It is essential to mitigate the power of the lightning before storage. In one embodiment of this invention, a collection of varistors of various power capacities, linked in series **270**,**280** and **290** are connected in parallel in lines **1** to n, to achieve a desired capability as shown in FIG. **7***a*. The resultant power is stored in the capacitor **300**. An insulator or fuse **305** is placed at the end of the varistor chain to protect the capacitor **300** against over-current. In the event of overstressing, the dielectric or varistor is deformed and the excess current is grounded.

[0066] Each component of the power mitigation unit is any of the following configurations as shown in FIG. **7***b:* A ZnO varistor **312** connected in series with a non-inductive resistor **313** and in parallel with a spark gap **311**. The resultant current is further modulated through another ZnO unit **314** and resistor **315**. A cascade of this configuration could be adopted to achieve the desired properties.

[0067] Zinc Oxide has a relatively large band gap of ~3.3 eV at room temperature. The consequence of this large band gap includes, higher breakdown voltages, ability to sustain large electric field and high power and temperature operation. Because of this, ZnO is the primary component in varistors and in voltage suppression devices. Zinc Sulfide has a band gap of 3.5 eV for the cubic form and as high as 3.91 eV for the hexagonal form at 300K. While ZnO band gap can be tuned from between 3-4 eV by alloying with magnesium oxide or cadmium oxide, the co-processing of ZnO with ZnS of the hexagonal wurtzite form should provide electrical field and other operational advantages over ZnO alone.

[0068] Multilayers of ZnO and ZnS or as sandwich layers with each other with ZnO achieving 50-99.9 mol. % of each layer can be adopted to improve on the suppression properties of the single layer ZnO varistor as is shown in FIG. **7***c*. Said varistor comprising a ceramic base body **318** of multiple layers, terminal electrodes **316**, **317** and internal electrodes **319**.

[0069] Alloying of ZnO with metal oxides and nitrides of transition elements including W, Zr, Mo, Mg, Cr, and Ti may improve its band gap characteristics and improve its voltage response non-linearity. Tungsten, in the form of tungsten silicide has very unique crystallization characteristics, particularly in the presence of dopants such as phosphorus or arsenic. Upon post deposition anneal, the Si in the non-stoichiometric tungsten silicide (WSix) segregates to the grain boundaries and the film interfaces. The dopant atoms diffuse preferentially along the grain boundaries. When oxidized, it is the segregated silicon that is oxidized selectively. If the W is exposed to the oxidizing ambient of O_2/Ar or O_2/N_2 at temperatures in excess of 700° C., tungsten oxide WOx (x=3, or 1.5) is formed. It has no corresponding crystallographic orientation to the underlying tungsten silicide or metal, being in the form of an amorphous powder. The fabrication of multilayer varistors with ZnO and WSi_2/WO_3, SiO_2 should provide the flexibility of modulation of the effective grain sizes of the host material to within 0.5-10μ as the case may be. This precise grain size control provides corresponding modulation of the electrical properties including overvoltage suppression.

[0070] A schematic of the storage unit **320** is shown in FIG. **8***a*. A collection of capacitors **330** are connected in parallel to achieve a desired capacitance. A schematic of the cross-section of each capacitor unit is shown in FIG. **8***b*. The unit **331** with dielectric layers **332** is a multilayer capacitor comprising alternately stacked internal electrodes **335**, **336** connected to external electrodes **333**, **334**. The dielectric layers may be of uniform composition, or layers of different compositions optimized for flexibility in the modulation of desired electrical, physical and reliability properties.

[0071] Capacitance in excess of $1\ kF/m^3$ with energy storage capacity of 1.5 MW.h is reported as being possible by existing art. With compositional and manufacturing improvements disclosed here these benchmarks are expected to be surpassed making possible energy storage capacities of up to 1 GW.h. Sufficient energy for 1000 households consuming 1000 kW.h of electricity per month for one month.

[0072] Lightning storms are a major cause of wildfire. To address this, in yet another embodiment of this invention as represented in FIG. **9**, air terminals **360** are set up on a selection of trees and connected by a conductor **370** in a mesh array as shown in FIG. **10**. The impinging charge is transmitted through a conductor **380**, through a suppressor unit **390** and on to the storage unit **400**.

[0073] Utility transmission lines usually have lightning collectors along the length of the network. In this invention as demonstrated in FIG. **11** air terminals **440** deployed transmission poles **450** are linked by a polarized conductor **460**. At predetermined locations the impinging charge is conducted through a suppression unit **490** and thereafter into the storage unit **500**. The stored energy is periodically released into the grid.

[0074] In yet another embodiment of the invention, air terminals **510** are deployed on light poles **520**. The collected charged are modified and stored in a storage unit **540**.

[0075] In another embodiment, the air terminal/collection units are deployed on radio transmission towers. The charge could be stored individually at each tower location or linked in a network as shown in FIG. **13**. In either case, the stored energy be used at each individual location or fed into the grid **570**.

[0076] In the absence of a thundercloud, lightening flash can be generated by the modification of the natural bias between the cloud and ground. This is demonstrated in FIG. **14**, where a positively biased conductor **580** of height greater than nearest ground elevation or nearest structure. The increased elevation and bias facilitates cloud to ground discharge even when there is no active lightning storm. The incoming charge is isolated by an insulator **600** and stored in the storage unit **610**.

[0077] In yet another embodiment of this invention, a balloon of conductive material is tethered to a terminal via a mesh line of conducting material. The balloon captures cloud to cloud discharges which are transmitted down to storage. In yet another embodiment this tether line may be wetted by a conductive solution such as but not limited to salt solution to facilitate the discharge conduction.

[0078] Another embodiment of this invention is shown in FIG. **15**. In an isolated collection system as shown here, the air terminal **620** which may be biased or unbiased is connected to a collection/storage unit **650**, **660** respectively by conductor **640** and isolated by a dielectric or insulator **670**. Should the discharge exceed the capacity of the isolation, the isolation is bridged and the charge grounded. This isolated collection system is particularly useful where there is no utility grid, particularly in very remote and isolated locations.

[0079] This isolated collection capability enables deployment on individual houses and isolated farms. In electric automotive recharge stations, this capability enables the location of these facilities in any location even those not served by regular electricity utilities lines.

[0080] In isolated rural areas, the deployment of these isolated collection units enables the provision of electricity where electricity grid infrastructure is not available or practical. The storage units can be deployed in banks as the estimated anticipated usage dictates.

[0081] In another embodiment of this invention, the isolated collection unit is modified to be portable for deployment on automobiles. The charge in the storage unit is linked with the automobile's electrical power or storage unit. Alternatively, it can replace the power storage unit of the automobile for pure EV or hybrid cars. Additionally, it can unplug these cars from the grid. With the ambitious industry plans for electric or hybrid cars the infrastructure for charging these cars either in garages or charging stations is somewhat limited. The capability to both 'charge as you go' and unplug from the grid would expand the infrastructure significantly.

[0082] In yet another embodiment of this invention, the isolated collection unit is deployed on trains. This could be in commuter trains that directly fed from the grid, or long-haul trains to augment or replace the fossil fuel used.

[0083] It is understood that the presentation of these steps in this disclosure is not exhaustive. Only the preferred embodiments of the invention and but a few of the examples of its versatility are shown and described in the present disclosure. It should be readily apparent to those of ordinary skill in the art that the invention is capable of use in various other combinations, environments and applications and is capable of changes or modifications within the scope of the inventive concept as expressed herein. These changes and modifications may be made without departing from the spirit and scope of the invention as set forth in the appended claims.

REFERENCES

[0084] U.S. Pat. No. 7,495,168,
[0085] U.S. Pat. No. 6,012,330
[0086] U.S. Pat. No. 6,078,494
[0087] U.S. Pat. No. 4,338,648
[0088] U.S. Pat. No. 6,407,900
[0089] U.S. Pat. No. 6,924,598
[0090] U.S. Pat. No. 5,569,972

[0091] U.S. Pat. No. 5,289,335
[0092] U.S. Pat. No. 4,272,720
[0093] U.S. Pat. No. 4,338,648
[0094] U.S. Pat. No. 5,610,813
[0095] U.S. Pat. No. 6,320,119
[0096] U.S. Pat. No. 5,604,167

1. A method of converting an atmospheric electrical discharge into a useable form of electrical energy comprising the following steps:

(a) Arresting the discharge via an air terminal,

(b) Said air terminal connected via a conducting unit to a power mitigating (step down) unit which can reduce the discharge voltage from approximately higher than 100 MV and 100 KA to below 20 kV, 10 kA.

(c) The step down unit connected to an energy storage unit which may comprise modules of high energy density capacitors

(d) Said storage units arranged such that the stored electrical energy can be discharged into the grid or to an appliance at point of use as desired.

2. A method of claim **1** wherein the arrestor and mitigating unit comprise an array of self-healing varistors in series and parallel chosen to ensure total conversion of the incident lightning discharge to a voltage/current level suitable for storage in said capacitors.

3. A method of claim **2**, where the power mitigation unit is any of the following configurations:

(a) Single ZnO varistor or

(b) Zinc Oxide arrester connected in series with a non-inductive resistor

(c) Zinc Oxide arrester in series with said non-inductive resistor and in parallel with a spark gap.

(d) A cascade of configuration as described in (c).

4. A method of claim **1** wherein the storage unit comprises modular array of high energy density doped ceramic capacitors with storage capacity of 10 kWe.h to 1 GWe.h and volume 0.1 to 1 m^3 and up to 1000 units as the desired capacity dictates.

5. A method of broad spatial lightning protection by the manipulation of the path of an electrical discharge via the manipulation of the air terminal bias magnitude and polarity Said arrestors being arranged in an array to achieve the redirection of the discharge path. The discharge path being directed to a collection unit at some location significantly removed from the initial discharge impact position.

6. A method of claim **5** wherein the discharge redirection is via a biased conductor to the said collection and storage unit.

7. A method of claim **5** wherein the area of protection is a forest whereby:

(a) A network array of air terminals is deployed on trees, in close proximity to said trees with tip approximating the treetops, or girding the trees to redirect the discharge along a biased conductor to a collection unit for the purpose of mitigating lightning induced fires.

(b) Said discharge electrical energy being stored or fed into the electricity grid.

8. A method of claim **1** wherein the discharge collection is on electrical power lines and the stored energy is fed back into the grid.

9. A method of claim **1** wherein the discharge collection and storage are at telecom towers, light poles or other free standing structures. Said storage being at individual collection sites or some location serving an array of collectors and said stored energy used for individual structure's operational energy need or fed back into the utility grid.

10. A method of non-spontaneous generation of atmospheric discharge by the following:

(a) Modulation of the magnitude and bias in a dipole air terminal relative to the ground

(b) Conductive material balloon tethered by a conductive mesh cable.

(c) Said cable being soaked by an ionic liquid, including, a salt solution to facilitate said atmospheric discharge conduction.

The discharge thus formed is stored for feeding into the greed or for operation as a stand-alone energy dam.

11. A method of claim **1** in which the discharge collection unit is configured as a mobile unit. Said unit being deployed on an automobile or any mobile platform. The said deployed unit extracting electrical charge from the atmosphere and storing such energy in a deployed storage unit on the said automobile. The energy thus stored, feeds the car battery which obviates the need for recharging at charging centers or from the grid.

12. A method of claim **1** wherein the electrical discharge energy collection unit is a stand-alone unit comprising modular capacitors configured as an electric vehicle charging station.

13. A method of claim **1** wherein the air terminals are deployed on buildings and the discharge energy collected thereby are:

(a) Stored and fed back into the grid.

(b) Stored for use by each building

(c) Stored in a central energy storage facility for use by the community in general.

14. A method of claim **1** wherein the said storage units are underground.

15. A method of claim **1** wherein the storage unit is of capacity and configuration to accept energy feeds from multiple electricity generation sources: Thermal, Hydro, Solar, Wind and Lightning.

16. A method of claim **4** wherein the storage unit is integrated into a Utility's Switch Farm.

17. A method of claim **4** wherein the multilayer ceramic capacitor unit comprises dielectric layers of varying compositions individually or as sandwich layers with each other containing Barium-Titanate silicates as host material doped with Zr, WSi_2/WO_3 to modulate grain growth to between 0.1 to 10 μm as the case may be; said modulation providing corresponding control of electrical and reliability properties.

18. A method of claim **1** wherein the varistor comprise multilayers of ZnO and ZnS individually layered or as sandwich layers with each other with ZnO achieving 50-99.9% of each layer; said layers being alloyed metal oxides of transition elements including W, Zr, Mo, Mg, Cr, Ti to improve voltage response non-linearity.

* * * * *

US008253639B2

(12) **United States Patent**
Cohen

(10) **Patent No.: US 8,253,639 B2**
(45) **Date of Patent: Aug. 28, 2012**

(54) **WIDEBAND ELECTROMAGNETIC CLOAKING SYSTEMS**

(76) Inventor: **Nathan Cohen**, Belmont, MA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 498 days.

(21) Appl. No.: **12/547,104**

(22) Filed: **Aug. 25, 2009**

(65) **Prior Publication Data**

US 2010/0156556 A1 Jun. 24, 2010

Related U.S. Application Data

(60) Provisional application No. 61/189,966, filed on Aug. 25, 2008.

(51) **Int. Cl.**
H01Q 15/02 (2006.01)
H01Q 19/06 (2006.01)

(52) **U.S. Cl.** **343/753**; 343/909

(58) **Field of Classification Search** 343/753, 343/909, 910; 333/135
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

5,471,224	A *	11/1995	Barkeshli	343/909
5,714,961	A *	2/1998	Kot et al.	343/769
6,473,048	B1 *	10/2002	Diaz	343/753

* cited by examiner

Primary Examiner — Michael C Wimer
(74) *Attorney, Agent, or Firm* — McDermott Will & Emery LLP

(57) **ABSTRACT**

Arrangement of resonators in an aperiodic configurations are described, which can be used for electromagnetic cloaking of objects. The overall assembly of resonators, as structures, do not all repeat periodically and at least some of the resonators are spaced such that their phase centers are separated by more than a wavelength. The arrangements can include resonators of several different sizes and/or geometries arranged so that each size or geometry corresponds to a moderate or high "Q" response that resonates within a specific frequency range, and that arrangement within that specific grouping of akin elements is periodic in the overall structure. The relative spacing and arrangement of groupings can be defined by self similarity and origin symmetry.

14 Claims, 4 Drawing Sheets

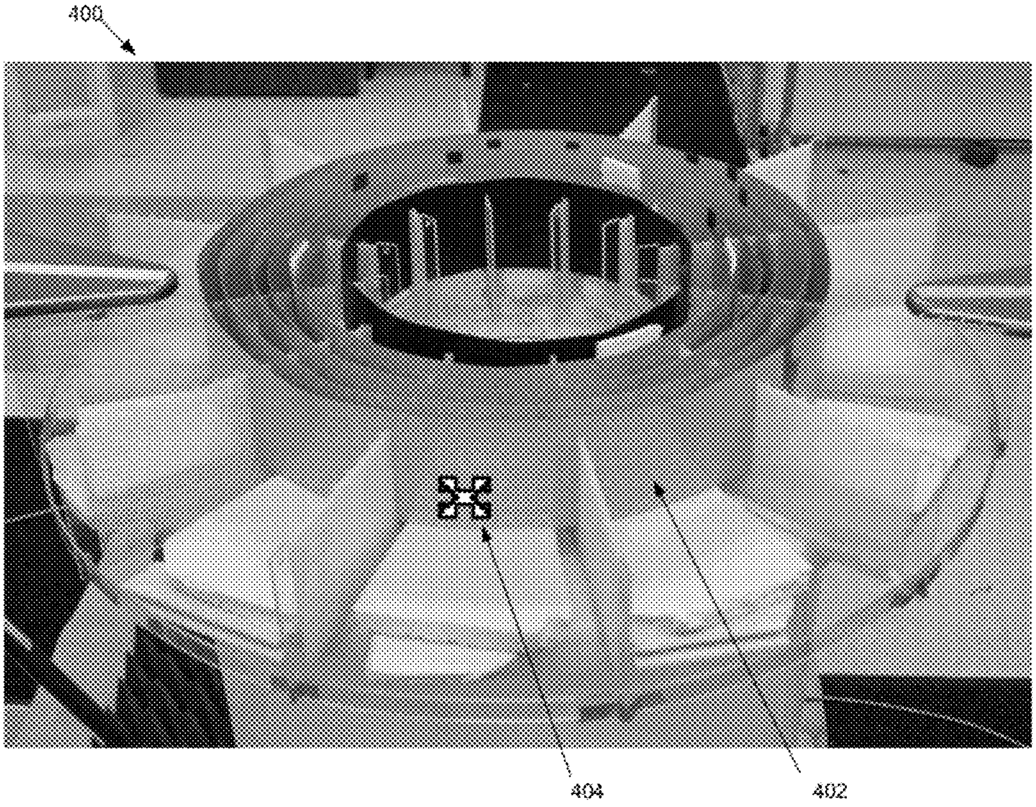

FIG. 1

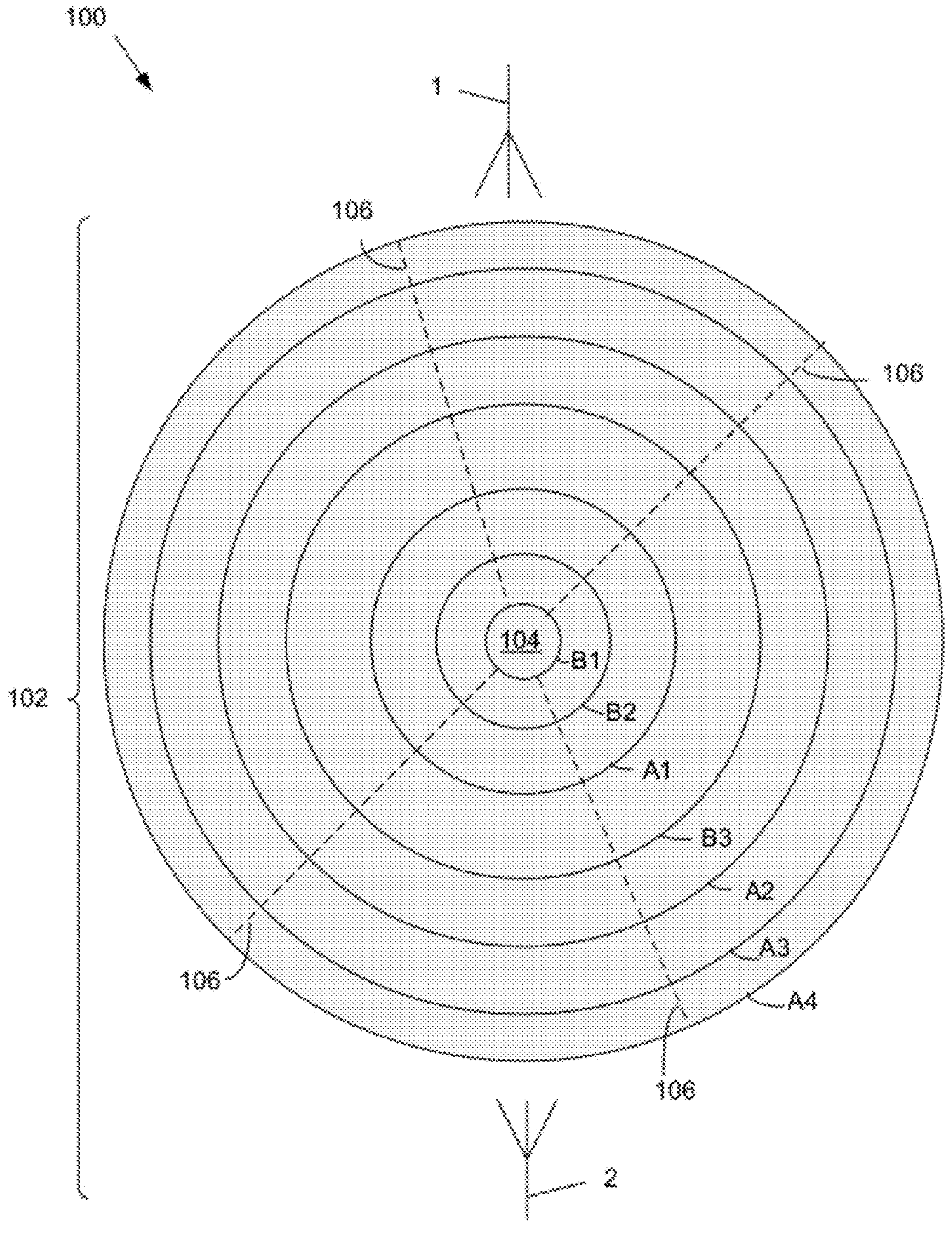
100
1
106
106
102
104
B1
B2
A1
B3
A2
A3
A4
106
106
2

FIG. 2

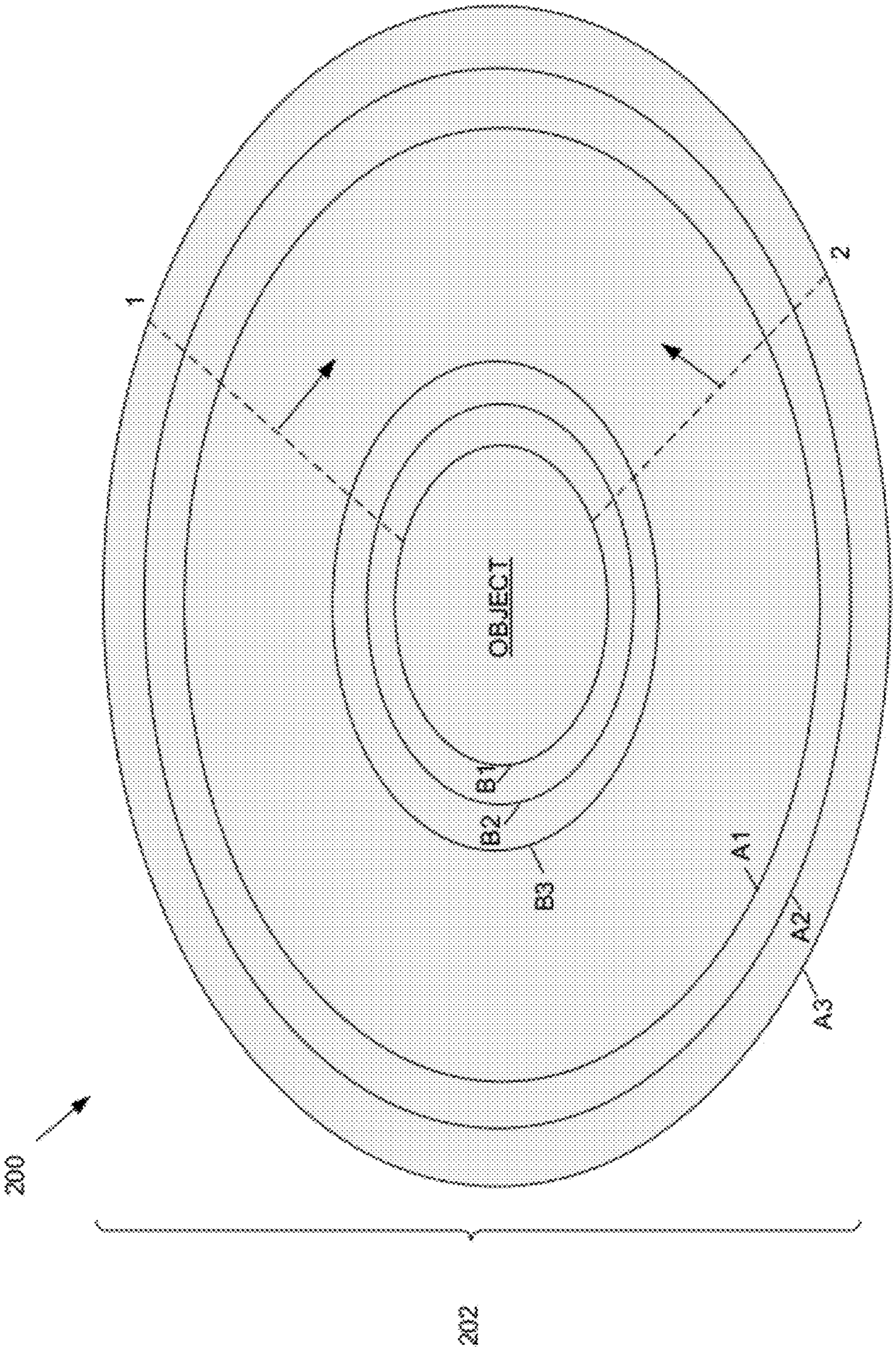
200
1
2
OBJECT
B1
B2
B3
A1
A2
A3
202

FIG. 3

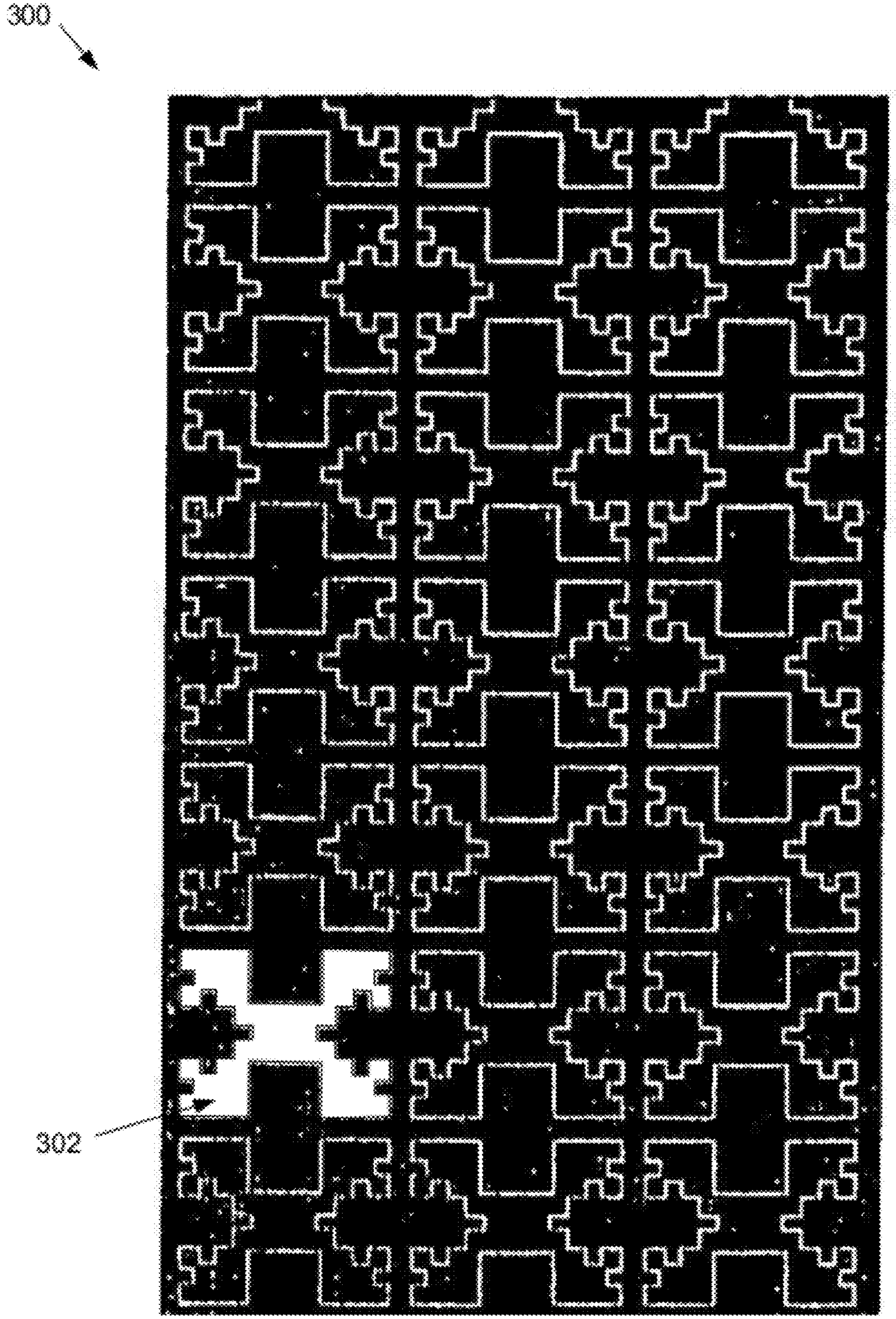

FIG. 4

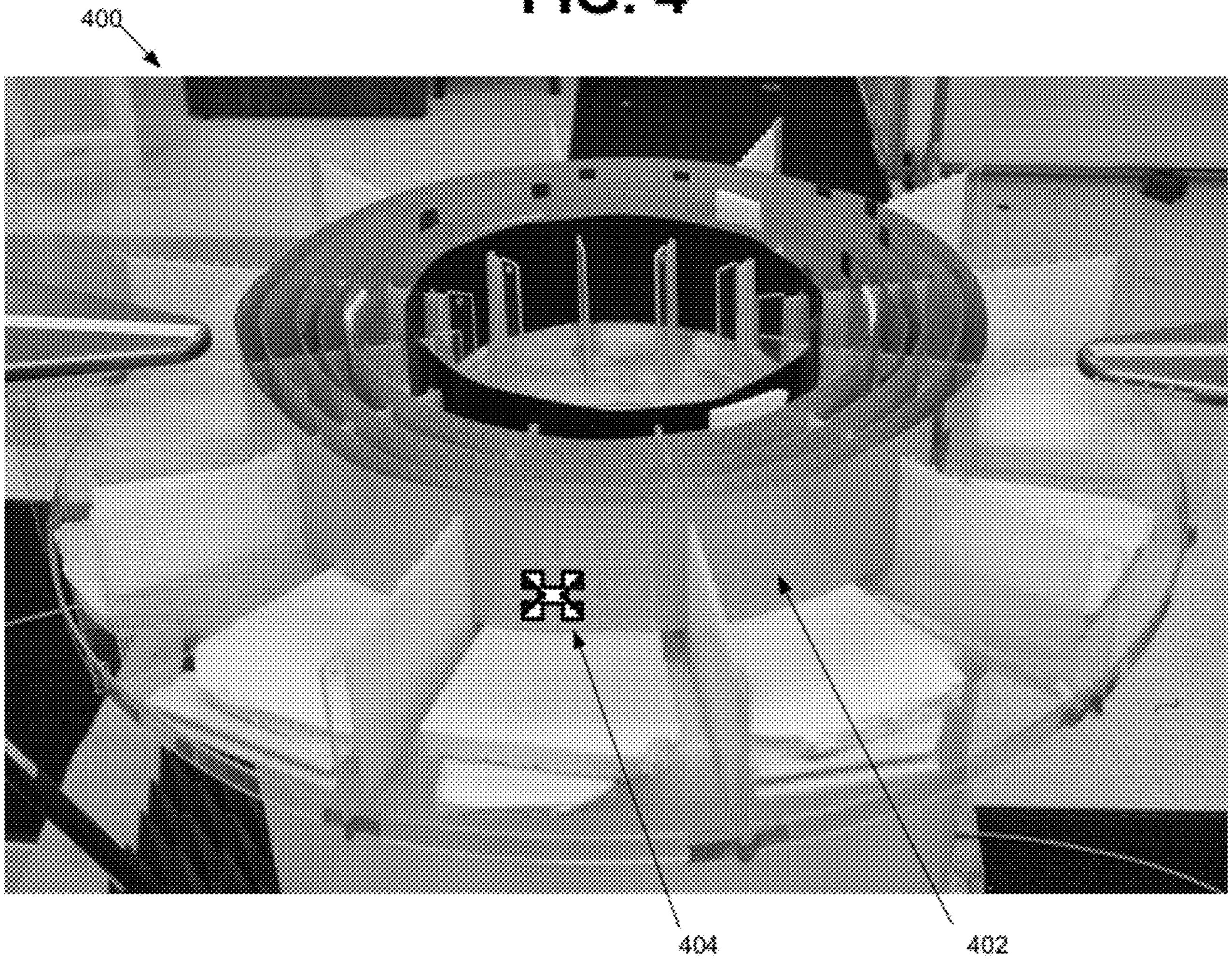

WIDEBAND ELECTROMAGNETIC CLOAKING SYSTEMS

RELATED APPLICATION

This application claims priority to U.S. Provisional Patent Application No. 61/189,966, filed 25 Aug. 2008 and entitled "Method and Apparatus for Wideband Electromagnetic Cloaking, Negative Refractive Index Lensing and Metamaterial Applications," the entire contents of which are incorporated herein by reference.

BACKGROUND

Much time and effort has been devoted to the quest for so-called invisibility machines. Beyond science fiction, however, there has been little if any real progress toward this goal.

Materials with negative permittivity and permeability leading to negative index of refraction were theorized by Russian noted physicist Victor Veselago in his seminal paper in *Soviet Physics* USPEKHI, 10, 509 (1968). Since that time, metamaterials have been developed that produce negative index of refraction, subject to various constraints. Such materials are artificially engineered micro/nanostructures that, at given frequencies, show negative permeability and permittivity. Metamaterials have been shown to produce narrow band, e.g., typically less than 5%, response such as bent-back lensing. Such metamaterials produce such a negative-index effect by utilizing a closely-spaced periodic lattice of resonators, such as split-ring resonators, that all resonate. Previous metamaterials provide a negative index of refraction when a sub-wavelength spacing is used for the resonators.

In the microwave regime, certain techniques have been developed to utilize radiation-absorbing materials or coatings to reduce the radar cross section of airborne missiles and vehicles. While such absorbing materials can provide an effective reduction in radar cross section, these results are largely limited to small ranges of electromagnetic radiation.

SUMMARY

Embodiments of the present disclosure can provide techniques, including systems and/or methods, for cloaking objects at certain wavelengths/frequencies or over certain wavelength/frequency ranges (bands). Such techniques can provide an effective electromagnetic lens and/or lensing effect for certain wavelengths/frequencies or over certain wavelength/frequency ranges (bands).

The effects produced by such techniques can include cloaking or so-called invisibility of the object(s) at the noted wavelengths or bands. Representative frequencies of operation can include, but are not limited to, those over a range of 500 MHz to 1.3 GHz, though others may of course be realized. Operation at other frequencies, including for example those of visible light, infrared, ultraviolet, and as well as microwave EM radiation, e.g., K, Ka, X-bands, etc. may be realized, e.g., by appropriate scaling of dimensions and selection of shape of the resonator elements.

Exemplary embodiments of the present disclosure can include a novel arrangement of resonators in an aperiodic configuration or lattice. The overall assembly of resonators, as structures, do not all repeat periodically and at least some of the resonators are spaced such that their phase centers are separated by more than a wavelength. The arrangements can include resonators of several different sizes and/or geometries arranged so that each size or geometry ("grouping") corresponds to a moderate or high "Q" (that is moderate or low bandwidth) response that resonates within a specific frequency range, and that arrangement within that specific grouping of akin elements is periodic in the overall structure—even though the structure as a whole is not an entirely periodic arrangement of resonators. The relative spacing and arrangement of groupings (at least one for each specific frequency range) can be defined by self similarity and origin symmetry, where the "origin" arises at the center of a structure (or part of the structure) individually designed to have the wideband metamaterial property.

For exemplary embodiments, fractal resonators can be used for the resonators in such structures because of their control of passbands, and smaller sizes compared to non-fractal based resonators. Their benefit arises from a size standpoint because they can be used to shrink the resonator(s), while control of passbands can reduce or eliminates issues of harmonic passbands that would resonate at frequencies not desired.

It should be understood that other embodiments of wideband electromagnetic resonator or cloaking systems and methods according to the present disclosure will become readily apparent to those skilled in the art from the following detailed description, wherein exemplary embodiments are shown and described by way of illustration. The systems and methods of the present disclosure are capable of other and different embodiments, and details of such are capable of modification in various other respects. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

BRIEF DESCRIPTION OF THE DRAWINGS

Aspects of the disclosure may be more fully understood from the following description when read together with the accompanying drawings, which are to be regarded as illustrative in nature, and not as limiting. The drawings are not necessarily to scale, emphasis instead being placed on the principles of the disclosure. In the drawings:

FIG. **1** depicts a diagrammatic plan view of a resonator cloaking system utilizing a number of cylindrical shells, in accordance with exemplary embodiments of the present disclosure;

FIG. **2** depicts a diagrammatic plan view of a resonator cloaking system utilizing a number of shells having an elliptical cross-section, in accordance with an alternate embodiment of the present disclosure;

FIG. **3** depicts an exemplary embodiment of a portion of shell that includes repeated conductive traces that are configured in a fractal-like shape; and

FIG. **4** depicts a perspective view (photograph) of an exemplary embodiment of the present disclosure.

While certain embodiments depicted in the drawings, one skilled in the art will appreciate that the embodiments depicted are illustrative and that variations of those shown, as well as other embodiments described herein, may be envisioned and practiced within the scope of the present disclosure.

DETAILED DESCRIPTION

The present disclosure is directed to novel arrangements of resonators useful for obscuring or hiding objects at given bands of electromagnetic radiation. Embodiments of the present disclosure can provide techniques, including systems and/or methods, for hiding or obscuring objects at certain wavelengths/frequencies or over certain wavelength/frequency ranges or bands. Such techniques can provide an effective electromagnetic lens and/or lensing effect for certain wavelengths/frequencies or over certain wavelength/frequency ranges or bands. The effects produced by such techniques can include cloaking or so-called invisibility of the object(s) at the noted wavelengths or bands.

Representative frequencies of operation can include, but are not limited to, those over a range of about 500 MHz to about 1.3 GHz, though others may of course be realized. Operation at other frequencies, including for example those of visible light, infrared, ultraviolet, and as well as microwave EM radiation, e.g., K, Ka, X-bands, etc. may be realized, e.g., by appropriate scaling of dimensions and selection of shape of the resonator elements.

Embodiments of the present disclosure include arrangement of resonators or resonant structures in aperiodic configurations or lattices. The overall assembly of resonator structures can include nested or concentric shells, that each include repeated patterns of resonant structures. The resonant structures can be configured as a close-packed arrangement of electrically conductive material. The resonant structures can be located on the surface of a circuit board.

The overall assemblies, as structures, do not all repeat periodically and at least some of the resonators are spaced such that their phase centers are separated by more than a wavelength. The arrangements can include resonators of several different sizes and/or geometries arranged so that each size or geometry ("grouping") corresponds to a moderate or high quality-factor "Q" response (that is, one allowing for a moderate or low bandwidth) that resonates within a specific frequency range, and that arrangement within that specific grouping of like elements is periodic in the overall structure—even though the structure as a whole is not an entirely periodic arrangement of resonators. The relative spacing and arrangement of groupings (at least one for each specific frequency range) can be defined by self similarity and origin symmetry, where the "origin" arises at the center of a structure (or part of the structure) individually designed to have the wideband metamaterial property.

For exemplary embodiments, fractal resonators can be used for the resonators because of their control of passbands, and smaller sizes. A main benefit of such resonators arises from a size standpoint because they can be used to shrink the resonator(s), while control of passbands can reduce/mitigate or eliminate issues of harmonic passbands that would resonate at frequencies not desired.

Exemplary embodiments of a resonator system for use at microwave (or nearby) frequencies can be built from belts of circuit boards festooned with resonators. These belts can function to slip the microwaves around an object located within the belts, so the object is effectively invisible and "see thru" at the microwave frequencies. Belts, or shells, having similar closed-packed arrangements for operation at a first passband can be positioned within a wavelength of one another, e.g., ⅒λ, ⅛λ, ¼λ, ½λ, etc.

An observer can observe an original image or signal, without it being blocked by the cloaked object. Using no power, the fractal cloak can replicates the original signal (that is, the signal before blocking) with great fidelity. Exemplary embodiments can function over a bandwidth from about 500 MHz to approximately 1500 MHz (1.5 GHz), providing 3:1 bandwidth; operation within or near such can frequencies can provide other bandwidths as well, such as 1:1 up to 2:1 and up to about 3:1.

FIG. **1** depicts a diagrammatic plan view of a cloaking system **100** and RF testing set up in accordance with exemplary embodiments of the present disclosure. As shown in FIG. **1**, a number of concentric shells (or bands) **102** are placed on a platform (parallel to the plane of the drawing). The shells include a flexible substrate (e.g., polyimide with or without composite reinforcement) with conductive traces (e.g., copper, silver, etc.) in fractal shapes or outlines. The shells **102** surround an object to be cloaked (shown as **104** in FIG. **1**). A transmitting antenna **1** and a receiving antenna **2** are configured at different sides of the system **100**, for verifying efficacy of the cloaking system **100** and recording results. The shells **102** can be held in place by radial supports **106** (while only four are shown, **12** were used in the exemplary embodiment indicated).

The shells indicated in FIG. **1** are of two types, one set (A**1**-A**4**) configured for optimal operation over a first wavelength/frequency range, and another set (B**1**-B**3**) configured for optimal operation over a second wavelength/frequency range. (The numbering of the shells is of course arbitrary and can be reordered, e.g., reversed.)

For an exemplary embodiment of system **100**, the outer set of shells (A**1**-A**4**, with A**1** being the innermost and A**4** the outmost) had a height of about 3 to 4 inches (e.g., 3.5 inches) and the inner set of shells had a height of about 1 inch less (e.g., about 2.5 to 3 inches). The spacing between the shells with a larger fractal shape (A**1**-A**4**) was about 2.4 cm while the spacing between shells of smaller fractal generator shapes (B**1**-B**3**) was about 2.15 cm (along a radial direction). In a preferred embodiment, shell A**4** was placed between shell B**2** and B**3** as shown. The resonators formed on each shell by the fractal shapes can be configured so as to be closely coupled (e.g., by capacitive coupling) and can serve to propagate a plasmonic wave.

It will be appreciated that while, two types of shells and a given number of shells per set are indicated in FIG. **1**, the number of shell types and number of shells for each set can be selected as desired, and may be optimized for different applications, e.g., wavelength/frequency bands.

FIG. **2** depicts a diagrammatic plan view of a cloaking system (or electrical resonator system) according to an alternate embodiment in which the individual shells have an elliptical cross section. As shown in FIG. **2**, a system **200** for cloaking can include a number of concentric shells (or bands) **202**. These shells can be held in place with respect to one another by suitable fixing means, e.g., they can be placed on a platform (parallel to the plane of the drawing) and/or held with a frame. The shells **202** can include a flexible substrate (e.g., polyimide with or without composite reinforcement) with a close-packed arrangement of electrically conductive material formed on the first surface. As stated previously for FIG. **1**, the closed-packed arrangement can include a number of self-similar electrical resonator shapes. The resonator shapes can be made from conductive traces (e.g., copper, silver, gold, silver-based ink, etc.) having a desired shape, e.g., fractal shape, split-ring shape, and the like. The shells **202** can surround an object to be cloaked, as indicated in FIG. **2**.

As indicated in FIG. **2** (by dashed lines **1** and **2** and arrows), the various shells themselves do not have to form closed surfaces. Rather, one or more shells can form open surfaces. This can allow for preferential cloaking of the object in one direction or over a given angle (solid angle). Moreover, while dashed lines **1** and **2** are shown intersecting shells B**1**-B**3** and A**1**-A**3** of system **200**, one or more shells of each group of shells (B**1**-B**3** and A**1**-A**3**) can be closed while others are open.

With further regard to FIGS. **1**-**2**, it should be appreciated that the cross-sections shown for each shell can represent closed geometric shapes, e.g., spherical and ellipsoidal shells.

As indicated previously, each shell of a cloaking system can include multiple resonators. The resonators can be repeated patterns of conductive traces. These conductive traces can be closed geometric shapes, e.g., rings, loops, closed fractals, etc. The resonator(s) can being self similar to at least second iteration. The resonators can include split-ring shapes, for some embodiments. The resonant structures are not required to be closed shapes, however, and open shapes can be used for such.

In exemplary embodiments, the closed loops can be configured as a fractals or fractal-based shapes, e.g., as depicted by **302** in FIG. **3** for an exemplary embodiment of a shell **300**, or **402** in FIG. **4**. The dimensions and type of fractal shape can be the same for each shell type but can vary between shell types. This variation (e.g., scaling of the same fractal shape) can afford increased bandwidth for the cloaking characteristics of the system (e.g., system **100** of FIG. **1**) This can lead to periodicity of the fractal shapes of common shell types but aperiodicity between the fractal shapes of different shell types.

Examples of suitable fractal shapes (for use for shells and/or a scatting object) can include, but are not limited to, fractal shapes described in one or more of the following patents, owned by the assignee of the present disclosure, the entire contents of all of which are incorporated herein by reference: U.S. Pat. Nos. 6,452,553; 6,104,349; 6,140,975; 7,145,513; 7,256,751; 6,127,977; 6,476,766; 7,019,695; 7,215,290; 6,445,352; 7,126,537; 7,190,318; 6,985,122; 7,345,642; and, U.S. Pat. No. 7,456,799.

Other suitable fractal shape for the resonant structures can include any of the following: a Koch fractal, a Minkowski fractal, a Cantor fractal, a torn square fractal, a Mandelbrot, a Caley tree fractal, a monkey's swing fractal, a Sierpinski gasket, and a Julia fractal, a contour set fractal, a Sierpinski triangle fractal, a Menger sponge fractal, a dragon curve fractal, a space-filling curve fractal, a Koch curve fractal, an Iypanov fractal, and a Kleinian group fractal.

FIG. **3** depicts an exemplary embodiment of a shell **300** (only a portion is shown) that includes repeated conductive traces that are configured in a fractal shape **302** (the individual closed traces). For the exemplary embodiment shown, each resonator shape **302** is about 1 cm on a side. Such resonator could, e.g., be used for the fractal shapes of shells B**1**-B**3** of FIG. **1**, in which case similar fractal shapes of larger size (e.g., about 1.5 cm on a side) could be used for shells A**1**-A**4**. The conductive trace is preferably made of copper. While exemplary fractal shapes are shown in FIG. **3**, the present disclosure is not limited to such and any other suitable fractal shapes (including generator motifs) may be used in accordance with the present disclosure.

It will be appreciated that the resonant structures of the shells may be formed or made by any suitable techniques and with any suitable materials. For example, semiconductors with desired doping levels and dopants may be used as conductive materials. Suitable metals or metal containing compounds may be used. Suitable techniques may be used to place conductors on/in a shell, including, but no limited to, printing techniques, photolithography techniques, etching techniques, and the like.

It will also be appreciated that the shells may be made of any suitable material(s). Printed circuit board materials may be used. Flexible circuit board materials are preferred. Other material may, however, be used for the shells and the shells themselves can be made of non-continuous elements, e.g., a frame or framework. For example, various plastics may be used.

FIG. **4** depicts a perspective view (photograph) of an exemplary embodiment of a cloak system **400** according to the present disclosure. As shown, the system includes a number of resonator shells **402** each having a close-packed arrangement of electrically conductive material (self-similar resonators **404**) formed on one surface. Two different shell configurations are shown, four larger shells and two smaller shells. The smaller shells included close-packed arrangements of resonator structures in which each resonator shape (as shown by **302** in FIG. **3**) was about 1 cm on a side. Similar fractal shapes of larger size (e.g., about 1.5 cm on a side) were used for the larger shells.

In FIG. **4**, a transmitting (source) antenna and a receiving antenna are shown as triangular shapes on the left and right, respectively (though functionally of each could of course be interchanged for the other). Twelve radially arrayed spacers are shown in FIG. **4**. The system **400** is shown supported on a Nalgene tank and Delrin platform and Delrin supports (radial supports) RF absorbers were placed in the immediate vicinity of the set up; further RF tripods (e.g., available from ETS) were used; all such materials were substantially transparent at the RF frequencies investigated/used. The cloak system **400** consists of six belts of fractal metamaterial (i.e., fractal-resonant structures shown in FIG. **3**) on flexible Taconic EF35 (low loss) circuit board. The belts are shown surround a scattering ring (object). The arrangement is supported by RF transparent plastics in a comb support. The entire system **400** was shown to be easily built up and broken down within a minute or two. The scale in FIG. **4** is about 0.7 meters across. The height of each shell can of course be selected as desired depending on the situation/application.

While embodiments are shown and described herein as having shells in the shape of concentric rings (circular cylinders), shells can take other shapes in other embodiments. For example, one or more shells could have a generally spherical shape (with minor deviations for structural support). In an exemplary embodiment, the shells could form a nested arrangement of such spherical shapes, around an object to be shielded (at the targeted/selected frequencies/wavelengths). Shell cross-sections of angular shapes, e.g., triangular, hexagonal, while not preferred, may be used.

One skilled in the art will appreciate that embodiments and/or portions of embodiments of the present disclosure can be implemented in/with computer-readable storage media (e.g., hardware, software, firmware, or any combinations of such), and can be distributed and/or practiced over one or more networks. Steps or operations (or portions of such) as described herein, including processing functions to derive, learn, or calculate formula and/or mathematical models utilized and/or produced by the embodiments of the present disclosure, can be processed by one or more suitable processors, e.g., central processing units ("CPUs") implementing suitable code/instructions in any suitable language (machine dependent on machine independent).

While certain embodiments and/or aspects have been described herein, it will be understood by one skilled in the art that the methods, systems, and apparatus of the present disclosure may be embodied in other specific forms without departing from the spirit thereof.

For example, while certain wavelengths/frequencies of operation have been described, these are merely representative and other wavelength/frequencies may be utilized or achieved within the scope of the present disclosure.

Furthermore, while certain preferred fractal generator shapes have been described others may be used within the scope of the present disclosure. Accordingly, the embodiments described herein are to be considered in all respects as illustrative of the present disclosure and not restrictive.

What is claimed is:

1. An electrical resonator system, comprising:

a plurality of concentric electrical resonator shells, each shell including a substrate having first and second surfaces and a close-packed arrangement of electrically conductive material formed on the first surface, wherein the closed-packed arrangement comprises a plurality of self-similar electrical resonator shapes and is configured to operate at a desired passband of electromagnetic radiation;

wherein the close-packed arrangements of at least two of the electrical resonator shells are different in size and/or shape; and

wherein a resonator in the close-packed arrangement comprises a second order or higher fractal.

2. The system of claim **1**, wherein said passband is about 2:1.

3. The system of claim **2**, wherein said passband is about 3:1.

4. The system of claim **1**, wherein the electrical system is configured and arranged so that radiation incident on the system from a given direction has an intensity on a point-by-point basis such at each respective antipodal point, relative to an object placed at the center of the system, the radiation has the same or similar intensity.

5. The system of claim **1**, wherein the system is configured and arranged so that radiation incident on the system from a direction in cylindrical coordinates has the same or similar intensity at the antipodal point after having traversed the system.

6. The system of claim **1**, wherein the plurality of shells comprises a first pair of shells having similar closed-packed arrangements for operation at a first passband, wherein the two shells are positioned within ⅛λ of one another.

7. The system of claim **6**, wherein the plurality of shells comprises a second pair of shells having similar closed-packed arrangements for operation at a second frequency band, wherein the two shells are positioned within ⅛λ of one another.

8. The system of claim **1**, wherein the plurality of shells are hemispherical.

9. The system of claim **1**, wherein the plurality of shells are cylindrical.

10. The system of claim **1**, wherein the plurality of shells are spherical.

11. The system of claim **1**, wherein said fractal is selected from the group consisting of a Koch fractal, a Minkowski fractal, a Cantor fractal, a torn square fractal, a Mandelbrot, a Caley tree fractal, a monkey's swing fractal, a Sierpinski gasket, and a Julia fractal.

12. The system of claim **1**, wherein the fractal is selected from the group consisting of a contour set fractal, a Sierpinski triangle fractal, a Menger sponge fractal, a dragon curve fractal, a space-filling curve fractal, a Koch curve fractal, a Lypanov fractal, and a Kleinian group fractal.

13. The system of claim **1**, wherein the plurality of concentric electrical resonator shells are configured and arranged for operation at K band, Ka band, or X-band.

14. The system of claim **1**, wherein the system is operational over a bandwidth from about 500 MHz to about 1500 MHz.

* * * * *

NU ENERGY TECHNOLOGIES

Your Energy Source for the New Millennium

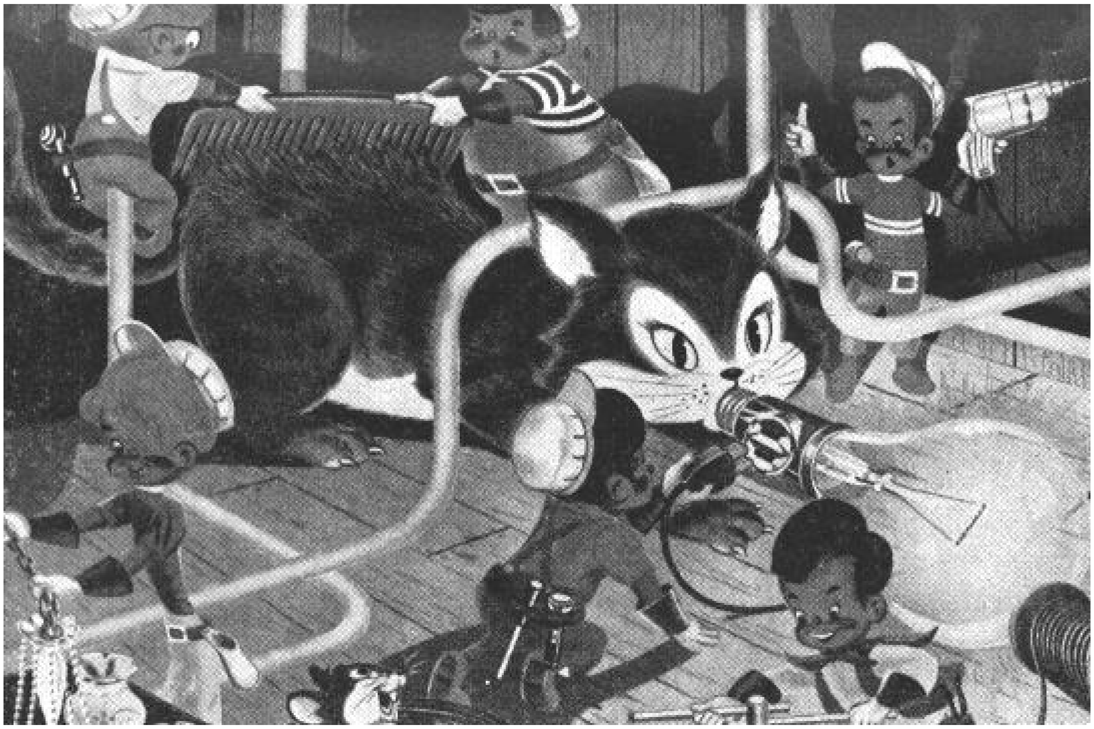

Radiant Energy
Power Generation

revised millennium edition

Information in this document is subject to change without notice and does not represent a commitment on the part of Bruce A. Perreault. The information described in this document is furnished under a license agreement or nondisclosure agreement. The information may be used or copied only in accordance with the terms of the agreement. It is against the law to copy this information on magnetic tape, disk, or any other medium for any purpose other than the purchaser's personal use.

LICENSE AGREEMENT

IMPORTANT – By reading the pages to follow and/or by using the enclosed information, you agree to be bound by the terms set-forth herein by Bruce A. Perreault (BAP). BAP grants you a limited, non-exclusive, non-transferable, royalty-free license for the use of one copy of the enclosed information. BAP retains all title and all ownership rights to this information. All rights not expressly granted are reserved to BAP. You shall not rent, lease, sell, sub-license, assign, or otherwise transfer the information. You shall not decompose or disassemble the information except to the extent that this restriction is expressly prohibited by law. Your license for the use of the information shall commence upon your receipt of it. The license shall continue until terminated in writing by the author or his estate. The information is deemed accepted by you. In no event shall Bruce A. Perreault be liable for any consequential, incidental, direct, indirect, special, punitive, or other damages whatsoever arising out of this agreement or the use of or inability to use the information, except to the extent that this limitation is prohibited by law.

Copyright Information

Unless otherwise noted, all materials (including without limitation all text, graphics, instructions, and demonstrations) are copyrighted, ©1999, 2000, July 2001 & November 2001 by Bruce A. Perreault. The material available through this publication may be freely used for attributed **noncommercial educational** purposes only. Due credit is hereby requested and notification is to be given to the author.

All materials may not be reproduced, stored in any retrieval system, or used in any way for commercial purposes without the express prior written permission of the copyright holder.

ISBN 1-930216-01-7

Dedication

Dr. Thomas Henry Moray, of Salt Lake City, Utah, was probably the first to use Radiant Energy as a source of power. His tireless effort, to gain recognition for this work, has been an inspiration to me.

The original concepts presented by the late Dr. Thomas Henry Moray, of Salt Lake City, between the years 1909 to 1974, have paved a road to our energy independence. He made history by drawing the attention of the scientific community to certain discoveries of far-reaching importance. His research involved the direct application of Radiant Energy, to the ultimate electric particles of the atoms and molecules of the Universe. Dr. Moray's discoveries, when demonstrated, convinced large numbers of scientific men who had witnessed them in action. However, there were a few scientists in high places who prevented the commercial success of this technology.

If these eminent men, the supposed leaders of the scientific community, had the privilege of studying under Dr. Moray; they would have understood his discoveries and applauded them. But in their ignorance of Radiant Energy, they refused to believe or even listen politely to the facts that he often demonstrated.

One of his prototypes measured about 42 x 26 x 22 inches and weighed no more than sixty pounds. It generated around four kilowatts of electrical power and was demonstrated in front of scores of witnesses. Unfortunately, these few people, knowing nothing of the truth, have done incalculable damage to Moray's discoveries over the years.

In its' widest form, Radiant Energy can be said to be the "Holy Grail of Free Energy." Those who wish to study this exciting subject will do well to give serious attention to Dr. Moray's original discoveries and to my recent research.

It should be as easy to accept the fact that a device can be built to receive radiation from its surroundings as it is to accept a radio that receives broadcast waves. One device transforms radio energy coming from a transmitter to sound while the other transposes electrical particles into heat, light, or power. The fundamentals are the same in radio and in the Radiant Energy (R. E.) device. The radio receiver receives transmitted energy waves out of the air; the R. E. device receives energy waves from atomic ions and converts them into electricity. As in the reception of radio waves, the Radiant Energy circuit is tuned to respond to the natural energy that is emitted from certain intermolecular reactions.

Table of Contents

BOOK DEDICATION ii
LIST OF FIGURES iv
INTRODUCTION v

SECTION 1
ATOMOSPHERIC ELECTRICITY 1

SECTION 2
THE WIRELESS TELEGRAPHIC CONNECTION 3

SECTION 3
MORAY'S CIRCUITRY 7

SECTION 4
EARLY DEMONSTRATIONS 10

SECTION 5
MY RADIANT ENERGY RESEARCH 31

SECTION 6
ION VALVE TECHNOLOGY EXPLAINED 35

SECTION 7
EDWIN GRAY'S ENERGY SYSTEM 39

SECTION 8
TREATISE ON MATTER & ENERGY 41

SECTION 9
ALPHA FUSION 47

SECTION 10
THE TRUTH ABOUT NUCLEAR SCIENCE 53

REFERENCE PATENT
HERMANN PLAUSON - JUNE 9, 1925 - 1,540,998 56

List of figures

Number **Page**

Figure 1 - EARLY RADIANT ENERGY CIRCUIT 7

Figure 2 - SOURCES OF RADIANT ENERGY 8

Figure 3 - PROOF OF CONCEPT 34

Figure 4 - ION VALVE 35

Figure 5 - RADIANT ENERGY GENERATOR CONCEPT DEVICE 37

Figure 6 - EDWIN GRAY ENERGY SYSTEM 40

Radiant Energy is emitted as charged particles from all bodies at a constant rate. The mystery surrounding it will unlock untold amounts of energy. Under the proper conditions, it can be rendered susceptible to the most incomprehensible changes caused by the oscillations, pulsation and surging throughout the Universe.

It was discovered by the Curie’s that all matter is radioactive. It can be shown that radium, barium, uranium, thorium, polonium, vanadium, cerium, molybdenum, zinc, aluminum, etc... will imprint radiographs on sensitive photographic paper.

The Sun’s explosions hurl intense surges of Radiant Energy at the Earth. The atmosphere absorbs these particles, shielding all life on the planet from deadly harm. Electrical charge is induced into the clouds that form in the lower atmosphere. This is due to the fact that clouds are in constant motion and travel parallel to the highly charged upper atmosphere. Once enough charge is built up in a cloud, it discharges to the ground as lightening. Static charge is thus transformed into kinetic oscillations. The Earth’s magnetic field quivers in response to these surgings. Is it too far fetched to reason that we can duplicate this natural process to obtain electrical power?

If an oscillating tank circuit has the correct impedance, reactance and inductance it will absorb energy from an external oscillating electrical source. Energy is captured. The tank oscillations can be kept alive by establishing resonance with the external source. Therefore, energy is not drawn from the transformer that powers the tank circuit. In the case of the Radiant Energy Receiver, it becomes possible to harness the atomic electrons that are generated within specially constructed energized plasma tubes.

What I have just revealed here is the "Holy Grail of Energy." The implications of this discovery are far and wide. Is humankind ready for such a revelation? What will be explained in the pages to follow is nothing more than a glorified radio receiver, one that is designed to oscillate with the oscillations of the Universe. It locks onto the very wheelwork of nature. This Radiant Energy receiver should last for many years with very little maintenance, no more than for a good radio.

"Electric power is everywhere present in unlimited quantities and can drive the world's machinery without the need of coal, oil, gas, or any other of the common fuels."

-Nikola Tesla

Atmospheric Electricity

We have succeeded in penetrating many secrets of Nature. Because of this we have an obligation to honor her abundant miracles with care and respect.

Since the time of Benjamin Franklin and his famous kite experiments not much has been learned about the nature of atmospheric electricity and how to convert it into useful power. Certain people have discussed this form of energy to some extent. Some theoretical papers in the so-called "free energy" field have been written. Many of these theorists deny that our ionized atmosphere can be utilized to produce usable amounts of electrical power. They are determined to cling to their wild unproven theories.

Nevertheless, I have taken it upon myself to find a solution to converting ionized particles into useful electrical currents. The properties of radioactive matter have shown me that large quantities of energy can be collected and converted. I have found that with the right circuits and components, rivers of high-voltage static electricity can be efficiently transformed into a lower practical voltage. It is my desire to be able to furnish irrefutable proof of this process, in this new millennium. However, there are still more questions that need to be answered. The utilization of the atmospheric electricity on completely new ground has to first take place.

The production and utilization of atmospheric electricity for the well being of humankind has been the goal of scholars and researchers for countless years now. However, the implementation of a real working device has remained a pious desire until today. This is because well-intentioned researchers turn away from the simple fact that there is free electricity in the atmospheric air just waiting to power our appliances. Only in the last year have I realized through new experiences and investigations that atmospheric electricity may be drawn into an electronic device in usable quantities to be converted into useful power.

Many researchers after investigating atmospheric electricity become astonished when they realize that there is nothing new in the World. They find out that the features of atmospheric electricity have been with us for a very long time, before the knowledge of electricity. The wisdom of the effects and utilization of atmospheric electricity dates back many thousands of years, as we can infer from the biblical texts. We read in the first book of Moses, that the all-holiest temple of Jehovah housed the Ark of the Covenant. With the exception of the high priests, this Ark would kill anyone getting to close to it. Furthermore, we read that 40 priests dared to approach the Ark, in the absence of Moses and Aaron. Lightning issued out from it, killing all 40 priests.

It is written in the second book of Moses, Chapter 25 verses 10-11:

10. Frame an ark of acacia-wood; the length whereof shall be of two cubits and a half (45 inches): the width, a cubit and a half (27 inches): the height, likewise a cubit and a half.
11. And thou shall overlay it with the purest gold within and without (exactly like a capacitor).

The biblical text sited here indicates that the Ark was made out of a noble wood that acted as a good electrical insulator and that gold was to be laid inside and out. Thereby, all conditions were just right for creating a good electrical capacitor. From the description we can tell that Moses and Aaron knew how to collect electrostatic energy from the atmosphere. This Ark of the Covenant could only have been a very large capacitor with a huge capacity. It was charged by atmospheric electricity from the accounts given. The fact that it remained fully charged with this electricity must be attributed to the dry climate in Palestine. Knowledge of this energy was a coveted secret kept by Moses and the high priests of Egypt. They were the first connoisseurs of atmospheric electricity.

When the study of meteorology first began there was a big interest in atmospheric electricity. Many in this field thought to use it. This is revealed through the patents of the time. The first relevant patent was granted to inventor Dr. Heinrich Rudolph of St. Goarshausen according to the German Patent Office; D.R.P. No. 98180 January 19, 1897.

Benjamin Franklin in America made a few attempts to utilize atmospheric electricity. He once built a motor that implied a commercially significant amount of energy might be derived from the Earth's atmosphere. It has been observed that a potential gradient of 150 volts per meter of height is always available on average.

There are three major causes of atmospheric ionization shown in **Figure 2**. The first one is due to the cosmic rays and the second cause is due to the solar winds. The solar winds consist of 96% protons and the remaining percentage is atoms of tritium. There are other particles but they are relatively unimportant. Almost all of these high-energy particles are absorbed by the Earth's upper atmosphere causing it to become ionized. This is how the ionosphere is created. However, the area where the reaction between the solar wind and the magnetosphere takes place is far outside of the Earth's atmosphere. The charged solar wind particles rattle our planet's magnetic cage long before its gas envelope neutralizes them. The third cause of atmospheric ionization is due to the content of radioactive gasses contained in the air we breathe. In the ground radioactive substances such as radium, thorium and actinium are spread throughout the soil. These elements generate gaseous radioactive emanations that spread in the air and ionize its molecules. The actinium-emanation generated from the decay of U235 plays a major role because of its brief radioactive half-life of 3.9 seconds.

Radio Ionics

There are not enough charged ions in the air to be converted into electrical power. This on the surface appears to be a correct calculation and assumption. There simply are not enough of them. However, what has been overlooked is the fact that these ions are in a state of constant surging motion. This is the mechanism to obtain power from this sea of energy! It is not the charge that should be looked at. These ions are in motion like magnets are in motion in a generator. Can you see the correlation here? It is the motion of the ions that **induce** electrical current. By means of properly tuned radio ionic circuit rests a possible solution to harnessing this source of energy. Hermann Plauson was successful in converting atmospheric energy into useful electrical currents. He was granted a patent titled "Conversion of Atmospheric Electric Energy" – U.S. Patent No. 1,540,998. In this patent he states that "St. Elmo's fire and the northern lights can be more or less absorbed in the same way as a receiver in wireless telegraphy absorbs waves coming from a far distance." This patent has been added to the end of this book for your study. It begins on page 59.

The Wireless Telegraphic Connection

Radio in the early years was called "wireless telegraphy." Beginning around the year 1913, amateur wireless telegraph stations were set up in trees and housetops. Aerials and masts dotted the countryside. It is estimated that there was almost "a quarter of a million stations." They were used to send and receive wireless telegraph signals, not voice. Vacuum tubes were not yet widely known. These stations had to rely on spark gaps.

It was the energy spikes, or surges put out by the spark-gap transmitters that first caught T. H. Moray's attention back in 1903. In 1909 Moray devised a circuit that would draw electricity from these surges of energy that traveled through the ground. By the fall of 1910 he obtained enough electrical energy from the ground to power a miniature arc lamp. During the Christmas holidays of 1911 he was able to power an old type 16-candle carbon arc lamp at about half of its brightness.

In 1912 while Henry Moray was on mission with the Mormon Church in Uppsala, Sweden, his passion for crystal radios got him started in his research in the science of radiant energy. Every spare moment he searched for a mineral that could possibly work as a radio detector. Moray had found two specimens that worked well as radio detectors. The material found in the hillside could have very well been a type of **argenti-zinciferrous-galena.** This type of galena is highly sensitive to radio waves that allowed his receiver to function without a battery. I make this assumption because a galena type synthetic formula can be found in Moray's Electrotherapeutic Apparatus – U.S. Patent No. 2,460,707. This material using only the power transmitted from a local wireless station could drive a small horn speaker as reported by Moray.

The other detector material was a white, powdery, stone-like material that he found in a railway car, located in Abisko, Sweden. From military records we know for certain that this white material was "**fused silica.**" Silica is the chemical name for the simple oxide of silicon, silicon dioxide (SiO_4). Mineralogists call this compound **quartz.** This is normally found in nature in its crystalline form. Finding this mineral in a fused state can only mean one thing. That, what Moray found was **metamict quartz.** A metamict is a crystalline mineral that loses its crystal structure due to radioactive destruction. For this to occur the quartz had to have contained trace amounts of uranium and/or thorium, which is why it was found in a "fused" state, in a more or less amorphous state, the so-called metamict state, owing to radiation damage from α-decay of these impurities. "Over the course of hundreds of millions of years, α-decay doses as high as 10^{19} decays/g can occur, which may lead to the complete amorphization"**(1)** of the quartz structure.

Moray's rare mineral find became known as the "Swedish stone." He was able to light a standard 100-watt General Electric light bulb in June of 1925 with this Swedish stone. By August 1925 he was able power an electric flat iron along with the 100-watt bulb. Therefore, bringing the total power consumed to 655 watts. People would quite often demand that he would draw too much power from his device and the white, stone-like material would overheat and burn up.

A problem of drawing power from telegraph stations appeared when the Federal Radio Commission on Nov. 11, 1928 limited the amount that could be transmitted. The transmitted frequency also had to be well defined. Spark-gap transmitters simply could not meet these requirements. Scores of spark gap driven telegraph stations had to be dismantled and in effect had been given death sentences. This spelled disaster for Moray. He could still draw energy from the background radio noise produced by the great cosmos but the amount of power that he could obtain was severely limited. Originally, Moray’s source of power was drawn from the local spark gap radio transmitting stations through the ground. Moray could draw power from the radio frequencies generated by nature but unfortunately only very low power could be had from this source. This forced Moray to develop a more favorable detector material. It was not too long before he got the inspiration to add radioactive dopants to his Swedish stone material and to other types of semiconductive materials. This occurred shortly after the Radio Allocation Act of 1928 when Moray's detectors were no longer able to receive steady incoming power. His Radiant Energy receiver could produce power but it only came in regular pulses. It was not constant without the added radioactive material. The addition of radioactive dopants to semiconductors allowed his receiver to produce a steady flow of energy.

Radiant Energy Early Definition

Professor Langley showed that out of the total amount of radiation coming from the Sun, only the visible portion of the spectrum represents only 19%. He presented the theory that there is only one kind of energy radiated from sunlight. That, heat, and chemical effects depend entirely upon the state or condition which radiant energy may happen to fall on matter.

Professor Langley wrote that up until 1872 it was almost universally believed that there were three different kinds of **entities**–actinic, luminous, and thermal, represented in the spectrum. There is only one radiant energy that appears to us as 'actinic,' or 'luminous,' or 'thermal' radiation, according to the way we observe it. Heat and light, therefore, can not be things in themselves, but separate sensations to our bodily sensors. They are merely effects of this mysterious thing called "radiant energy."

The Italian physicist Melloni stated that, "light is merely a series of caloric indications sensible to the organs of sight, or vice versa, the radiation of obscure heat are veritable invisible radiation of light." Melloni wrote this in 1843, but it was not adopted until Langley by his elaborate researches, more refined and complex, proved it. The great physicists of society then adopted the doctrine of one radiant energy.

Only a few researchers know the immediate effect of radiant energy: electromagnetic. The results of two centuries of observation all point to this conclusion. When a mighty tongue of white-hot matter darts across the abyss of a large spot or cavern on the Sun, the equilibrium of the Earth's magnetic field is disturbed and the effect is a magnetic storm. The needles of magnetographs throughout the world quiver and oscillate. The vibrations take place on opposite sides of our planet, the impulse passes from the Sun to the side nearest of the Earth and then to the most distant side, whether through or around the Planet's surface.

After centuries of investigation from Gilbert to Tesla, this most wonderful research still holds admiration and mystery for all that study this vast science; electrodynamics. Power is cut out of the seemingly emptiness of space, and the hurrying waves are caught and chained to servitude in artificial light and electric appliances. The Sun, being electromagnetic, emits waves that carry power, which beat and surge against the Earth. A magnetic field is space that surrounds a magnet. This space might be filled with air, wood, stone, glass, or might be a vacuum. But the waves are not quenched. They flow through all of these things. A freely suspended magnet in a magnetic field will move, and the Earth's surface is surrounded by a magnetic field, that is acted upon by the Sun's magnetic field. Suspend a sewing needle by a silk fiber in the Earth's magnetic field and it will come to rest parallel to the field's north and south poles. Now, if this field becomes disturbed, that is; if it becomes stronger or weaker, the needle will move. This is a magnetic storm. When a gas jet is hurled across a spot on the Sun the disturbance reaches the Earth in the same time that light does. So, radiance traveling at 186,000 miles per second is reaching the Earth in eight minutes and nineteen seconds.

The most memorable magnetic storm occurred on November 17, 1882. This was one of the most violent recorded. The daily press was burdened with accounts of widespread magnetic disturbance. In some places telegraphic communication was suspended. The turbulence filled that great quadrilateral from New York to Yankton, Nashville and Winnipeg. In Milwaukee, the carbons in the electric lamps were lighted, rendered incandescent by currents of electricity flowing on the wires. At other locations, switchboards in telegraph offices were set on fire and sending keys were melted, while electric balls were seen hovering on the telegraph lines in Nebraska.

The Earth's aurora holds the key to harnessing the Sun's daily pulsation. An aurora is the visible effect of obscure undulations from the Sun, as they come dashing on the Earth with a speed of 186,000 miles per second. In a six-month winter, say at the North Pole of the Earth, the Sun is far south of the equator, and none of its rays can shine on the Earth's northern pole. However, the aurora is very bright. It displays many colors, and these flash and glow with rapid variations. The light, although caused by the Sun, does not come direct. It is caused by the turbulence set up in the Earth's magnetic field by electromagnetic upheaval on the Sun. The field of the Earth is "tuned" with the Sun's field, as was the coherer in the days of wireless telegraphy and telephony. The aurora is known to be electrical, for magnets and compass needles on ships are always affected. Could the coherer of the days of old teach us something new?

No magnet can be placed near a "current" of electricity, or a static charge without making an oscillation. Every oscillation sends out a wave, like a stone falling in water. An electromagnetic wave from the Sun disturbs the Earth's magnetic field in the same way that an induction coil used in wireless telegraphy does. The Earth and coherer are both doing the same thing with the same kind of electromagnetic waves. The coherer in a distant receiving station will react to natural electromagnetic wave fronts from the Sun. It will likewise react to the artificial wave fronts coming from the induction coil. The electromagnetic wave makes the loose particles of metal between the knobs in the glass tube coherer generate dots and dashes.

Making a device "oscillate with the oscillations of the Universe" may not be too far off the mark in obtaining energy. Perhaps a radon-powered diode (ion valve) will act as a one way gate to this energy? This type of specialized diode offers zero resistance to electrical current. It is also an ion gate that can function in the million-volt range. Today's state of the art high voltage diodes will not meet our requirement. This is because as more are added in series the greater the electrical resistance becomes. I believe that Moray's Swedish stone met our requirements utilizing only a trace amount of radon generating material.

Moray found that semiconductive materials doped with radioactive impurities have an increased sensitivity. One such satisfactory formula included a triboluminescent zinc mixture consisting of pure zinc sulfide, radioactive impurities and pure germanium metal. He also found that "artificial radiation" could be used in place of the radioactive impurities. For example, by exposing bismuth to wavelengths of energy emitted from one of his unique vacuum tubes, he was able to create "active bismuth," today known as Polonium–210. This is what Moray called "artificial radiation." When added to the zinc sulfide, it would glow a bluish-green. When this glowing material was then doped into ultra-pure germanium it made an excellent emitter of secondary electrons. This synthetic substance was called "fission material" by Moray. This transistor type material is said to be superior even to what is available today.

"The Moray germanium mixture gives certain unique results in functioning as in a valve and or booster (amplifier). Made in form of rounded stones or pellets compressed under high pressure and fused. Combination contains **radium chloride, thorium, uranium, and pure germanium metal-triboluminescent zinc.** Some pellets fastened to envelop with pure tin in place of solder. Bismuth pellets fused to side. Germanium mixture pellets float between other pellets but making firm, needlepoint-like contact. Have used silicon too, which has some of the properties of germanium. Germanium works best when impurities are introduced. Care must be taken when alloyed with other substances, as too much other mixture added worsens conductivity and germanium loses its properties." * T.H. Moray, 5th edition, The Sea of Energy in Which the Earth Floats (1978), p. 71.

I had often asked myself: what ever became of Moray's radiant energy tubes? Wouldn't it be nice to run some tests on them? These questions have been put to rest because these very tubes are now in my possession. It was communicated to me that the widow of T. H. Moray gave these tubes and his notebook to a man named Stan Mahurin. This man knew Moray and his wife quite well.

Mr. Mahurin passed the tubes onto a close friend because he had no clue as to their function but insisted on keeping the notebook. The next owner of the tubes had no clue either. These tubes remained in storage for over twenty-five years. I was contacted because of my passion for Radiant Energy research and was given the opportunity to purchase them. I promptly replied and have had the honor to inspect them first hand. Upon inspection of these tubes there does not appear to be any radioactive material in them. However, this does not mean that it was never in the tubes. Moray stated that he was using what he called "artificial radiation." From his notes we know that he created this artificial radioactive material by attracting radon gas to a metal by giving it a negative charge. We also know that radon decays to lead-210, a beta emitter with a half-life of around 22 years. This would mean that Moray's source of induced radioactive material would eventually burn itself out. This would explain why Moray did not have a running prototype in his later years.

My own research is now nearly complete. My next project is to build a radiant energy generator that will demonstrate a few hundred watts of excess power above what the circuits require to be self-sustaining. The final step will be for me to develop a commercial unit for general use. Through this millennium edition of my book, the shroud of mystery surrounding this old source of energy may now be lifted.

1) Miami University, Herbert Jaeger Physics

Moray’s Circuitry

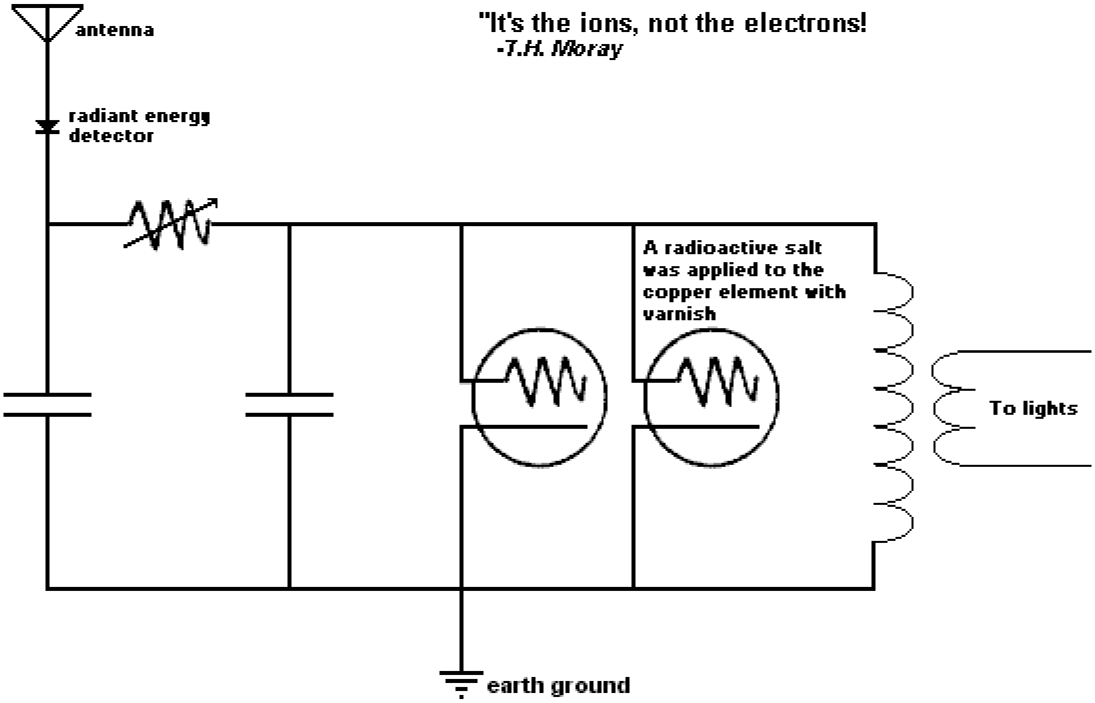
antenna
"It's the ions, not the electrons!
-T.H. Moray
radiant energy detector
A radioactive salt was applied to the copper element with varnish
To lights
earth ground

Early Radiant Energy Circuit - Figure 1

It is assumed that the average reader of this publication is familiar with the elementary principles of radio transmission. Granting this, it is of course reasonable to believe that such readers will understand how an oscillating current is set up in a circuit comprising an inductance or coil of wire, a capacitor and a spark gap. The capacitor is charged with a high tension current from any convenient source such as a transformer or induction coil. When the potential stored in the capacitor reaches a critical value, the air in the gap between the spark gap electrodes can no longer stand the strain. The capacitor discharges across the gap in a succession of crashing sparks. As the current from the capacitor crosses the gap in one direction, it literally over-reaches itself just as a pendulum will swing past the neutral point when given a push by hand. After the first rush of current passes in one direction, a reversal of the cycle occurs and a second rush in the opposite direction occurs. This operation is repeated many thousands of times per second. The discharge gradually dies down as all the energy is either completely radiated or used up in heat or performing work. The oscillations cease once the potential across the capacitor has been lowered so that the spark can no longer jump the air gap. The capacitor immediately takes a fresh charge from the transformer and the entire cycle of operations is repeated. It will be understood that all of this passes in an infinitesimal fraction of a second. The charge and discharge of the capacitor takes place so rapidly that the observer can detect no change in the solid spark that appears continuously to fill the gap. It should also be understood that if the oscillating tank circuit has the correct impedance, reactance and inductance, it would absorb energy from an external oscillating electrical source. Energy is therefore **captured.** The tank oscillations can be kept alive by establishing resonance with the external source. In the case of Moray's early radiant energy receiver, the capacitor was charged through a material that acted like a high voltage diode. This material he called "Swedish stone." Electrons (electricity) were actually generated in the discharge tubes and fed back to the antenna and ground circuitry. This mode of energy generation is explained in more detail on pages 37 to 38.

What has been explained here is the Holy Grail of energy. The implications of this discovery are far and wide. Is humankind ready for such a revolutionary design? This receiver locks onto the very wheelwork of Nature. This device should last for many years with very little maintenance, no more than for a good radio.

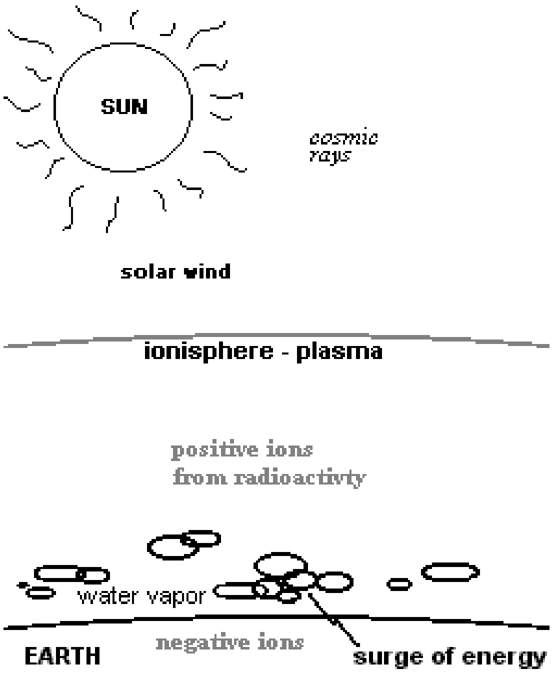

Sources of Radiant Energy - Figure 2

SOME FACTS TO THINK ABOUT

1. Cosmic particles are hurled at the Earth from outer space.
2. The Sun emits ionized **protons** and **tritium** atoms that are carried to the Earth by the solar winds 24 hours each and every day.
3. The Earth's upper atmosphere absorbs the charge carried by the particles in the solar winds causing it to become ionized.
4. The Earth's ionized atmospheric region is called the "ionosphere." It is sort of a half gas and half plasma.
5. Water vapor (clouds) become charged by induction as they travel parallel to the ionosphere.
6. When clouds become sufficiently charged they will discharge according to Pupin’s theory of capacitive discharge. The energy is released in the form of light and electrical oscillations.
7. There is a cloud discharge about one hundred times each second that is randomly located somewhere on the planet, generating a tremendous amount of oscillating electrical energy. The total energy per discharge is approximately "**two million-million watts.**" Now multiply this by one hundred and you have an unbelievable amount of unleashed energy.
8. The facts presented are Nature’s process of generating oscillating electrical energy.

Moray’s Radiant Energy Detector Tube

T.H. Moray invented a very unique vacuum tube. He called this discovery, the "Moray-valve." It functioned as a high-voltage, high frequency, minute capacitance detector tube. It included solid-state semiconductors mounted "under ideal conditions."

This Radiant Energy valve contained Moray’s fission material. It replaced the original Swedish Stone. It’s anode acted as a collector made with a molybdenum disulfide rod. The metal case served as a parabolic reflector to concentrate and focus the secondary ions that were expelled from the fission material.

Radiant Energy in Moray's detector comes to us in waves of separate and distinct units called photons. There are definite relationships between the energies of these particles and the zinc sulfide (triboluminescent zinc) germanium mixture, contained in Moray’s fission material. The zinc sulfide in the mixture is highly sensitive to the photons generated from the fission material. The sulfide transforms the high frequency photons into lower frequency photons. These photons then **impinge** upon the germanium. The germanium being sensitive to these lower frequency photons emit secondary photo ions. The hemisphere that contains the active materials was filled with a low-pressure water vapor and was comparatively simple in construction.

Thomas H. Moray successfully demonstrated an electrical device, weighing less than sixty pounds, that extracted enormous amounts of energy: 4,000 watts. This unconventional source of energy is abundant day or night throughout the year. To understand how this device generates energy, you must carefully read the material contained in this book. I have compiled this information to assist you in understanding radiant energy power generation without giving away the entire contents of my discoveries.

On the introduction page in the "The Sea of Energy in Which the Earth Floats," 4th edition, Moray explains that radiant energy is kinetic in nature. It is rendered sensible by conversion of its energy. He states that "this is the phenomenon of the transducer fueled by a fission reaction. In the final analysis, radiant energy is a means of using the energy released by the fissionable reactions taking place in the stellar crucibles of the universe." Moray in his 2nd edition, on page 61, describes the "universe" as being both stellar and atomic. The final clue to the inter-most secret of Moray’s energy device can be found on page 33 in this same edition; "It is an accepted fact that when various substances are bombarded with alpha particles they are found to give off **electrons.**" How much clearer could he have made this? He further states; "just as sodium, potassium, cesium, rubidium, barium and strontium react to visible light or wavelengths within a certain range, might not certain other substances react to oscillations from the Cosmos or artificially produced radiation?" Moray’s **fission material** clearly fits this description. How could we not know from this description how Moray’s "detector" functioned? He also states; "the universe is analogous to a radio transmitting station. It is continually emitting energy, only of a greater range of wavelengths." Is it this type of statement that has confused people seeking answers? I say that this is a certainty.

What is revealed within this publication should clear a path to the understanding of Moray’s radiant energy device for the reader. It will be understood that ionic recombination can be directed to produce intense surges of electromagnetic energy. More clearly, ionic energy is converted into electromagnetic energy. The conversion from ionic to electromagnetic energy can then be transformed into useful electrical power by conventional radio circuitry.

Electronics vs. Ionics

The Earth’s electric field that exists between the surface and the ionosphere is charged to about 360,000 volts. It is estimated that the stored energy ranges from about one million to a billion kilowatts. This field provides a low current at high voltage and will not power today’s appliances because they use high current, low voltage electromagnetic power. In 1748, Benjamin Franklin was the first to invent an electrostatic motor that could run on the earth’s energy field. Did this scientist of yesteryear point us in the right direction?

At the Paris Electrical Exhibition of 1881, Professor Amos Dolbear successfully demonstrated his electrostatic transducer for the telephone. He obtained US Patents No. 239,742 and No. 240,578 for this invention.

Thomas Edison’s 1880 discovery of thermionic emission, the current of electron flow in a vacuum tube, created the first electronic device. Sir William H. Preece called this the "Edison Effect." Edison had discovered the fundamentals of the electron tube and was granted US Patent No.307, 031. Its’ basic principle was to be used in radio communication for many years to follow.

In 1887 Dolbear was forced to go to court against Alexander Graham Bell to establish that he was the inventor of the telephone. Dolbear lost this infringement case. The judge consequently ruled that Bell was the inventor of the telephone. This decision was the pivotal point of our electrical development. Because of this one decision by a court, inventors moved in the direction of electromagnetic driven components.

Conventional radio tubes to this day are based on electron emission. Electron flow through a vacuum tube is where we get the word "electronic." If science had gone the way of ionic emission instead of electron emission in radio tubes I believe that our technology would be drastically different today.

I have designed radio tubes that carry an ion current instead of electrons. Vacuum tubes that carry electrons are called "electron valves." My tubes carry ions and therefore are called "ion valves." This changes everything about inductance and capacitance in any circuitry. Capacitance then becomes the real key element in electron transfer through associated wires and circuitry. It may be of interest to you that "ohm’s law" only applies to conventional electron tubes. In a conventional electron tube the mean velocity of an electron is about 1 cm per second. In my ion tube charge transfer is much faster. When using an ion tube as a rectifier or oscillator this changes the whole mathematical equation on the electrodynamics of the associated electrical circuits.

I believe that T. H. Moray's tube oscillators were actually ion valves. This is what allowed them to run cold. This is the Holy Grail of energy that can be used to tap into the sea of energy that is all around us. We are bathed in this sea of ions at all times. This is not just a wild dream. The circuits given in this publication are proof positive that a new technology is just on the horizon.

With the release of this publication, there can now be no doubt that we are at the dawn of a revived energy revolution. I have given a great gift to science. All I ask for in return is that you shout from your rooftops the information gleaned and to please give me the credit for what I have earned.

Early Demonstrations

What follows are accounts given by observers of the T. H. Moray radiant energy device that utilizes the vast stores of energy in the Universe without mechanical force, but through **ionic oscillations**. The statements have been edited for clarity without changing the meaning being conveyed.

A brief description of the device

A brief description of what the device has done will be given. Also, successful demonstrations will be described like those that have been made in the presence of hundreds of reliable witnesses, many of whom are highly trained and hold degrees in physics and electrical engineering. There will be a brief description of why the device works based on the oscillations of condensers. I do not believe in perpetual motion. I have no quarrel with those that may; the operation of the Moray device is **not** perpetual motion. It utilizes energy that exists and transforms it into useful forms. An electrical generator is not really a generator, as it creates nothing. The generator does not make electricity. It is merely pumped. From that standpoint, an electric generator might be referred to as an electron pump. The Moray energy device might be called an **electric siphon**, or high speed-oscillating turbine.

T. H. Moray's device was enclosed in a box 10 x 10.5 x 26 inches, one wire leading from the device to a special balanced antenna. Another wire was connected to a balanced ground connection.

In my version, a special tube ion is used. The switch of the device is simply turned on. The device is adjusted so that a synchronized resonance is established. Once it is operating, the device "siphons" energy. A specially constructed high frequency transformer is used to control the output so that any desired voltage may be obtained.

On page 17 of the "Nature of the World and of Men" (compiled by the technical staff, Chicago Institute of Technology), we learn that our planet is receiving energy from the Sun continuously at the rate of 160,000 horsepower per inhabitant.

In the "Physical Review," Dr. Gunn of the U. S. Navy Research Laboratories states that the Earth itself is a huge dynamo, producing 200,000,000 amperes of electrical current. For more detail, read pages 334 to page 344, "Physical Review," July 15, 1939.

The Aurora Borealis is definitely considered to be an electrical phenomenon produced by the passage of electric charges through the rarefied gases of the higher atmosphere.

The conversion of matter to energy in the stars is now generally accepted as fact. During radioactive disintegration energy waves are radiated. Thus, we may conclude that energy waves of very high frequency are sent out from the stars, one of which is our Sun.

Similar to the reception of radio waves, the Moray circuit is resonance tuned by the right arrangement of inductance, capacities and oscillators. These components respond to the particular wave frequency "oscillations" of energy from the Cosmos. Moray used a special valve to prevent return of the power to the outer circuit and forced it to go through the power application circuit.

All that is necessary to put the device in operation is to start electric oscillations in the circuit due to a difference of potential between ground and atmosphere.

The frequency of the current is extremely high, as is shown by the brush discharge when either the antenna lead or the ground wire is disconnected from the machine. Certain difficulties of insulation inherent in high frequency currents are avoided by the ingenious operations of the oscillator tubes that reduce the frequency on the output side.

Standard electric light globes become exceedingly hot in one spot about the size of a dime when operated on this current. This is due to the fact that the gas in the globes becomes incandescent under the influence of the very high frequency current. Also, the luminosity of the incandescent gas is much higher and whiter than with ordinary current. A variation of the Tesla, high frequency, button type globe would be ideal for use with this device.

Ordinarily, when a potential is applied to the terminals of a condenser, a full charge is acquired almost instantaneously. The filling up of Moray's oscillator tubes is similar to water being poured into a bucket. That is, the longer the voltage is applied the greater the charge that is taken, up to the maximum capacity for a given potential and frequency.

R.E. experiments have been made at different places, many miles from all power lines. One experiment occurred more than 50 miles from any power lines and 26 miles from even a farmer's telephone. The locations for these experiments were selected by those making the test and not by the inventor. These locations were not prearranged but selected as they drove along in their own car, not the auto of the inventor.

In an endurance test, the device was operated under seals for a total of 157 hours and 55 minutes. Then the seals were broken and the device was put under severe strain and tests for about another hour and then shut off.

Close examination of the device, showed that all parts were in perfect shape and could have been run indefinitely. During the entire test the light burned evenly and brightly without flickering. There was no change in the brilliance from day to day.

The quantity of current passing through the secondary of the transformer, experts have asserted, is sufficient to burn up similar wire if ordinary current was used. Yet, there is no heating of the transformer, even though cooling air was not present, as it is completely enclosed. All parts of the machine run absolutely cool regardless of the length of time operated.

It makes no difference whether one 50-watt lamp is used or whether fifty of them are connected to the machine; the current is adapted to the load.

This current has another physical characteristic as seen in photographs of light bulbs operated by the radiant energy (R.E.) device. The photographic images of such bulbs are extremely bright. So, the emulsion density in the area of the bulb image is extremely low, virtually zero. If one takes a photograph of the original print the resulting film has the same characteristics. Prints made from this second film have these same characteristics as well. The light bulb images have such extreme low density that one has to reduce the exposure of other parts of the image in order to get a satisfactory picture. Yet the light from the R. E. device is whiter and less harsh, even to stressed eyes. Here is further evidence that this current is different from any coming from conventional sources in use today.

There is no sound coming from the machine when in operation, there being no moving parts, electrical oscillations or prime mover operations.

The current cannot be derived from batteries, as it can be seen. Doctors of Physics have pronounced this current to be high frequency. Only an alternating current has these properties. Batteries provide only direct current.

It cannot be induction from adjacent power lines, as an induced current is always the same frequency as the inducing current. There is no power line on Earth carrying current of these frequencies. This current cannot be caused by radio signals. These are much too weak. To operate a loud speaker, it is necessary to amplify a radio signal many times, to say nothing of lighting a lamp or heating a flat iron. Also, the frequency of the incoming current is higher than radio signals.

Moray gave a demonstration for a nationally known physicist, a professor in one of the greatest universities of this country. The physicist made these observations:

Moray states that when the oscillators are connected to the circuit the condensers fill slowly. The longer the current is applied the greater the charge they take, up to their maximum for the applied voltage. This is similar to filling a bucket by pouring the water into it. The condensers do not take the charge instantaneously, as is ordinarily the case.

Moray has also stated that the size of wire in the transformer could not carry the amperage passing through it without burning up if ordinary current were used. Yet the wires remained absolutely cool no matter how long the machine operated.

The above points show these currents to be entirely out of the ordinary. They cannot be accounted for by induction from existing power lines or current from batteries.

Another short account follows

Today Mr. Judd, Mr. Adams and attorney Nebeker visited the inventor's laboratory. After the "radiant energy" device was packed into the attorneys' auto, we drove away. The three above-mentioned gentlemen began to discuss where they should go to make the test experiment. The inventor did not want to have any say in where the test was made. He wanted the test to be made at a place selected by them. At last, the three men decided to go up Emigration Canyon, as there are no power lines there. After driving about four miles up the canyon they selected a place. Then they changed their minds, selecting another place a few hundred feet further up the canyon.

Mr. Judd stayed in the car because of an injured foot, while Mr. Nebeker and Mr. Adams put up the antenna and ground. The inventor then took the device out of the car and connected it to the "antenna and ground." The switch on the device was opened and closed many times as in all former experiments, but no light appeared. The device was then "tuned in" as Mr. Judd had witnessed in previous tests. Then the switch was closed and the lights came on. The "antenna wire" was momentarily disconnected. The lights went out but came back on when the "antenna" was again connected to the device. The same thing happened when the "ground wire" was disconnected and then reconnected. All this with Mr. Judd hopping around on one foot, he having gotten out of the car when the tuning process was started. All three gentlemen were very well satisfied and pleased with what they saw. It was dusk when they left the canyon.

The following is a letter describing a similar test performed at a different location

Dear Mr. Cooley,

This letter is being written for your information. It is a record of the electrical demonstration made October 29th, 1926, by inventor T. H. Moray for Attorney Judd, Mr. Knight and myself.

As prearranged, I met Moray and Judd at the Moray laboratory about 7:10 a.m. on October 29th, 1926. We carried the electrical equipment in my car and left Mr. Judd's car in the Moray parking lot.

I remember that the odometer registered 19 miles at Charleston and 26 miles as we left the last electric power line near the mouth of Daniel's Canyon. It registered 52 miles when we stopped to do the demonstration. Thus, we were 26 miles from the nearest power line and also 26 miles from the nearest one wire rural telephone line.

Moray requested that we select a place near a stream of water so that the ground pipe could be sunk in its bed to be more effective. The ground in the mountains, at this time of year, was frozen. We stopped at a place about 10 miles southeast of the Daniel's Strawberry summit and about 200 yards west of the main road to Duchesne. This location was almost due east from what Mr. Knight called Haystack Mountain. It was perhaps ¾ of a mile east of Strawberry Lake on a little stream that made a zigzag course through a gently sloping grassy flat.

The antenna wire was put up without any aid or instructions whatever from Moray, as it had been "balanced." Moray did suggest that the wire be stretched tighter to prevent so much sag at its' center. This was done and the wire then appeared to clear the ground by 7 or 8 feet at its lowest point.

Operation of the R.E. device;

The balanced ground rod was pointed at the end to make it easier to drive into the ground.

The antenna wire was insulated from the poles with two quartz glass insulators about six inches long. A piece of wire about two feet long connected each insulator to the poles. The lead-in wire was fastened to the antenna at a point about 10 or 15 feet from the east pole. I helped Moray solder the connection where the lead-in wire was fastened onto the antenna wire. I also helped him solder the ground wire to the rod. I stepped the distance between the two antenna poles and estimated it to be 87 feet as I took 29 steps intended to be three feet each.

Moray took his electrical equipment out of the automobile and placed it on the running board of the car. Two dry boards were laid on the ground and a rubber mat from my office was placed on the boards for Moray to stand on. This was as a precaution against electric shocks. The running board was hardly large enough for the equipment so we took the seat cushion out of the front seat and placed it on the mat. Moray transferred the equipment to the seat cushion and connected it up there.

Very light snowflakes fell occasionally and a tarpaulin was hung over the top of the auto doors to protect the equipment from getting wet. When all of the wire connections were made and the device synchronized in resonance by Moray. It was just 1:05 PM by my watch. Before "tuning in" he closed the switch but no light appeared. After "tuning in" for slightly more than 10 minutes the switch was closed and the light appeared immediately. It was slightly after 1:15 PM by my watch. Moray closed the switch two or three times before and during the tuning operation but no light appeared until perfect "balance" was established.

While the lights were burning the antenna lead-in wire was disconnected from the apparatus. The lights went out. When reconnected again, the lights reappeared. Moray disconnected the "ground wire" and the lights went out. He then reconnected it and the lights appeared again.

E. C. Johnson of Salt Lake City, Utah signed the above letter.

Mr. Nelton Welling wrote this letter, describing yet another test

Your interest in the Moray invention to take electrical energy from the "air" and make it usable for light, heat and power purposes is the reason for this letter. It describes a demonstration that I saw of his device on Saturday, February 21.

A dozen people were present, including Paul Harsh, Mark Yuri and Mr. Ferguson.

I first witnessed a demonstration of this device three months ago. Since then, the cabinet containing the machine has been simplified and improved. It was quite apparent that there was no possible faking of the power produced.

The tuning device was improved. The time required to bring in the energy was shortened from five minutes to less than a minute.

The operation was as simple as tuning in a well-equipped radio set. A lady who witnessed the demonstration for the first time also performed this tuning. She operated the device as easily as Dr. Moray himself did. This was after she had seen him "tune in" the energy.

A pilot light on the cabinet first became illuminated. The switch was then closed, connecting a light rack with the current. Instantly thirty 50-watt lamps and five 100-watt lamps were brilliantly lit.

A regular, Hot Point flat iron was then connected without dimming the lights in the least. The inventor stated that the result would have been the same had one hundred lamps been used in place of the 35 on the light rack.

The lights and iron together were consuming more than four-horse power of electric energy.

On account of the brilliancy of the lights it was apparent that much more than ordinary voltage was going into them. The excessive heat, which developed in one spot, made me feel that they would soon burn out, but they did not.

I confidently believe Dr. Moray is on the threshold of perfecting one of the most amazing fundamental inventions of history.

Very sincerely yours,

Nilton H. Welling

The next document is a lengthy report written by T. J. Yates to the Secretary of Utah about Morays work

Dear Secretary of the State of Utah,

Many men of science have come from foreign countries and from the East and the West and have been shown demonstrations. Not one of them has been able to find any fault with what they have seen or heard. Names of these men will be furnished upon request. These men have had the device opened for their inspection. They have pronounced the experiments to be wonderful, that the current is high frequency, the color of the light different, that the device carries many times as much current without even getting slightly warm. Any other electrical device of like construction, known to man, would burst into flame if it were carrying that much current. They state that the "tubes used are far more powerful than anything known to science today." The drawings, circuits and theory have been pronounced by leading men of science as scientifically, electrically, mechanically sound and correct.

It is generally accepted now by science. Moray's device proves there is energy coming from somewhere, that such a field of energy surrounds the Earth.

As Moray explains it, the oscillator tubes pick up electrical oscillations through the circuit of the device itself. As stated, the Moray device picks up these surgings or oscillations of energy coming and returning to the Universe. It is tuned to oscillate in harmony (sympathetically) with the oscillations of the Universe just as musical instruments can be made to vibrate together. Every oscillation, whether large or small, is completed during the same interval of time, the heartbeats of life. The oscillations of the Universe are governed by the same eyelet of time and are completed during the same interval of time. As Moray stated years ago, these waves of energy have a regular beat note of time, coming and going like the waves of the sea. They are in a very definite mathematical order of time, coming to the Earth from every direction. They are stronger in the day time than at night, but always coming with a regular beat note that might be referred to as the Father of Time; the Sire of Gravitation.

This energy has a definite elastic rigidity and density, which is subject to displacement and strains. When the strain is removed, this medium will spring back to its old position and beyond, surging back and forth as the waves of the sea. It will continue to oscillate until the original pressure is used up. If the internal impedance is too great, there will be no oscillations. It will merely slide back in a dead beat to its unrestrained state. By cutting down the resistance to the minimum and by synchronous resonance (sympathy) of the device with that of the Universe, recovery will be quicker and quicker. Finally inertia will assert itself and lengthen the time of final recovery by carrying the recoil beyond the natural oscillation. Thus, the vibrations are prolonged by oscillation. When the recovery is distinctly oscillatory, resonance sets in. The oscillations will go on forever because they come from the Universe. These electric oscillations are not simple ones, but surges with a definite beat note.

One will ask, how can you get steady energy from such surging? Since there is a great amount of energy at such a terrific potential could not a steady flow of water be obtained from the surging of the sea?

The operation of the condensers need not be repeated here as they have been fully explained in the enclosed account.

It is not claimed that all of Moray's theory is proven. But it is claimed that the device works. The results are certain. In the absence of better explanations, Moray's theories are as good as any.

Moray explained his theory to a well-known and noted American scientist, who said, "You go in your theory back to the law of gravitation."

In this day and age, anything is possible. What was done with radiant energy has been called "radical." Nevertheless, of the great number of learned men who have seen and heard of Moray's work, not one has been able to disprove his claims, theories or discoveries. Some of these men are amongst America's foremost in science. They have spent from five hours to many days on the theory and claims and made test experiments. Hence, these further explanations of the Moray theory, on which he has spent his time from boyhood.

Fully realizing that whatever the difficulties in discovering new truths, there are still greater obstacles in getting them recognized. As Mark Twain once said, "It takes many years to get a new idea fixed in the human mind." One should not be long surprised at the attacks of some or at the exasperation of a certain number of worthy people. Not surprising is the silence of a greater number of the scholars who have heard of these experiments. It is hard for the average human mind to rid itself of inherited ideas, which so completely and unconsciously control our line of thought. Franklin, Faraday, Cavendish and others were unable to always express themselves so that their peers understood their inner meaning. They gave to the World their ideas in a form unintelligible to others of their day. Yet their ideas have since become known facts.

The theory described here is not new in the main, nor is it contrary to ideas that science has accepted today. It is however original with Moray in application. When Moray first advanced them years ago they were "killed" by those who heard of them before they got very far. Be that as it may, Moray has studied "radiant energy" and found a means of using it. Heat, light, power and energy are not things in themselves. They are sensations, or effects produced by this "cosmic power," directly or indirectly.

"ASTRONOMY," by Robert H. Baker, P.H. D. Professor of Astronomy, University of Illinois

Page 303:

Another problem relates to the apparent lavish expenditure of this radiation. Of all the energy that pours forth from the Sun, the planets and their satellites intercept less than one part in 200 million. The remainder spreads through interstellar space with little chance, so far as we know, of being recovered. The suggestion that the Sun shines only in the direction of material that can intercept it, may appeal from the point of view of economy, but appears to have little else to recommend it. It would seem that Nature is squandering its resources of energy so prodigally that it must end in bankruptcy. But we doubtless have at present, an imperfect account of the situation.

"FOUNDATIONS OF THE UNIVERSE," by M. Luckiesh, D. SS. Director of Lighting Research Laboratory, General Electric Company printed in 1925

Page 5:

In the far-off stellar crucibles, we see the same laws being obeyed as in our laboratories. Tracing down, to the almost infinitesimal constituents of the extremely minute atom, we find that apparently it does not exist at all. It is not even as the realistic matter that we have supposed it to be. There at its very foundation, it seems to consist of electric charges that probably simulate the motions of celestial bodies. It is becoming more and more certain that the apparent complexity of Nature is due to our lack of knowledge. As the picture unfolds, it promises a marvelous simplicity.

Pages 41-43:

The great success of the atomistic principle relating to the kinetic theory of matter, is one of the wonders of the modern scientific age. As expected, it has found other applications equally fascinating and promising. It is now being pressed further into the service of explaining the structure of matter.

Maxwell's Theory

When Maxwell (1873) propounded the electromagnetic theory of light (radiation), his achievement was epochal. The exact manner in which the radiant energy traversed space was not known. The next epochal event was the founding by Planck (1900) of the quantum theory. Here we have the atomistic principle applied to energy instead of being confined to the material of the Universe as it had been. In other words, in the quantum theory, we have the atomistic idea applied to physical processes. We now have the atom of matter, the atom (electron) of electricity and the atom (quantum) of action (a product of energy and time). Planck assumed the emission of radiation (from the Sun, a lamp filament, etc.), to occur discontinuously. He conceived elements of energy of equal magnitude; analogous to the equality of electrons, or atoms of a given element. Radiant energy is emitted at various wavelengths or frequencies that must be taken into account in the laws of radiation. Now, the physicist uses quanta as commonly as he does electrons and atoms and molecules. Bodies are built of molecules, the molecules of atoms, and the atoms of electrons (and protons). Here we see the atomistic principle applied to "material" (matter) and then to electricity (what shall we call it)? Finally, a physical process, the radiation emitted by the electrons, is divided into quanta. With such pictures of the Universe being constructed, we may cease to be surprised at anything, but our interest and admiration will grow.

One of the most marvelous relationships that has ever been revealed in the entire science of physics is that between light and electricity. Knowing now how to view the structure of atoms, this relationship is not quite so surprising. Half century ago, there was a total absence of knowledge, pertaining to the existence of electrons in atoms of matter. So, the sudden revelation, that light (and radiation in general) is an electrical phenomenon, was very startling and revolutionary. Even today those persons, who are unfamiliar with fundamental physics, find it difficult to believe that energy traveling from yonder star to the eyes, is electromagnetic in nature. But that has been amply proved. It is the atoms in those distant stellar crucibles that possess moving electrons that are emitting electromagnetic waves of many wavelengths or frequencies. Here on Earth, we have many 'receiving stations' that are tuned to certain ranges of wavelengths.

"RADIANT ENERGY," by Edgar Lucien Larkin (1903)

Pages 17-18:

Radiant here means proceeding from a center in straight lines in every direction. Energy is internal and inherent.

Professor Barker, "Physics," page 4, says: "Energy is defined as a condition of matter in which any definite portion may effect changes in any other definite portion." This was written in 1892 and discoveries since confirm it. Energy then, is a state of matter. Or rather, is it the result of a particular state in which matter may be when any observed phase of energy appears?

These two notions, matter and energy, or possibly one, are the sum total of all that has been found during three centuries of incessant research? This search has been in that portion of the Universe visible in a forty-inch telescope, armed with the most powerful spectroscope ever made.

It is the belief of the writer that all this space is saturated with inconceivably minute corpuscles. J. J. Thomson recently discovered these. These are doubtless either electricity in its ultimate refinement, or very closely allied to it, or its immediate carriers. The smallest particle of hydrogen has long been thought to be the smallest mass of any known particle of matter. But the corpuscles detected by Thomson have only one-thousandth the mass of the hydrogen atom. The Earth and Sun, all suns and dark bodies in space, all granular matter, moves through the primordial cosmic mass of electrical corpuscles as would a wire screen through water. The wide spaces in diamond, glass, steel, flint, or anything else, allow these "bodies smaller than atoms," as Thomson calls them, to pass through.

Larkin's Theory on the transmission of energy

From the definition of energy, it is the potential of the Universe. When matter is in a phase allowing it to be active, it affects other quantities of matter at a distance. The method of transfer is known to be by means of wave motion. Each impulse moves from the emitting to the receiving mass on a rigorously straight line. One continuous set of oscillations in this straight line is called a ray. Each negative or Thomsonian corpuscle makes a double vibration to and fro like a pendulum straight across the direction of the ray – i.e., at right angles to it. The corpuscle moves over and returns to the original position it had before the excursion. The corpuscles are negative and can be drawn out of their original straight path by the action of magnetism. So, the entire wave motion of the Universe is electromagnetic. This is what Maxwell prophesied forty years ago. Thomson fulfilled the prophecy.

After one corpuscle makes an oscillation across the direction of the ray and returns, the next does likewise and the next and so on. After the first corpuscle makes a swing, another distant from it 186,000 miles in the same straight line, will also make a vibration at the end of the first second of time.

"ATOMS AND RAYS," by Sir Oliver Lodge, Fellow of the Royal Society of Science and holder of five or six Doctors degrees from colleges, with honors from a score more. He has held offices of President in a dozen or more Scientific Societies. Printed in 1924

The term 'light' strictly speaking, means that kind of ethereal radiation which is able to affect the eye. But it is common knowledge that there are many other wavelengths of radiation besides those that effects the eye. It is not clearly known why the eye is sensitive to some kinds of ethereal radiation and not to others. That, no doubt, is a question for physicists and physiologists in collaboration. But the eyes of animals, insects and man, all appear to be sensitive to a limited range of ethereal radiation, which is therefore called light. Some kinds of radiation can affect a photographic plate. Other kinds can stimulate the chemical actions going on in the leaves of plants and thereby supply the energy needed for vegetable growth. Another kind, a rather deeper harmony as it were, supplies everything on Earth with warmth and by evaporating water contributes to most of the phenomena of weather. Other kinds, again, are omitted when individual electrons, traveling at a high speed in a vacuum, encounter the obstruction of a target. These kinds of invisible radiation are called x-rays. And, at the opposite end of the scale, great antennas emit another kind of radiation. These are the Hertzian waves employed in radio.

If describing these kinds of radiation as different, we are not speaking quite accurately. They differ only as treble notes differ from bass notes. They differ in rapidity or rate, or vibration or wavelength. They do not differ in any other essential particular. The longest waves are telegraphic ones, which may be a mile long. Much, much shorter are the x-rays, whose wavelength is actually smaller than atoms, and only expressible in billionths of an inch. This whole range of waves, travel at precisely the same speed. This is the only speed at which the ether is able to transmit energy. They are all of the same electromagnetic character. They are all subject to the same optical laws of interference, of reflection and polarization, which have long been studied in departments of physics.

Moray would qualify the above by saying there are speeds greater than the speed of light

Whatever an electric charge is, or is not, it is certainly a focus of energy. Imagine an ether vortex, containing the known mass of the electron and circulating with the velocity of light. The energy of this vortex would be equal to that of the electric field in the space surrounding the electron. This coincidence, if it were a coincidence, can hardly fail to have some meaning. Some thinkers are beginning to view the whole material Universe as being built up of ether in various states of self contained or intrinsic motion. These adjectives are intended to discriminate between rotary motion, like that of a top or a whirlpool and ordinary locomotion, or whiffling from place to place. Locomotion has not been attributed to the ether, which is the most stationary thing we know. It is perhaps the only stationary thing that exists, but it may be full of what is sometimes called "stationary motion." This is a paradoxical term, appropriate to the condition of a sleeping top.

Those who hold this view of the Universe are strengthened in their position by Einstein's expression of energy in general. It is well known that all ordinary energy, such as the motion of railway trains, or croquet balls, etc, is merely relative to the Earth, or to some other piece or matter. There is nothing absolute about it. But Einstein gives an expression for what could be called absolute energy, in which the only relevant velocity is that of light. All the phenomena of Nature, at any rate in inorganic Nature may be due to, this great ethereal velocity. It must appear in a form that enables it to appear to our animal derived senses. For the spinning motion itself is impalpable and beyond the ken of our instruments. It can be detected only when it partially exhibits itself as transmitted waves in the form of radiation.

All the light that we experience can be resolved into vibrations or tremors in the ether. That is how we first knew about the ether. But all electric and magnetic phenomena and therefore, all chemical activity are likewise known to be modes of manifestation of the ether of space. The complete manner and meaning of which have still to be worked out.

So the question arises, What is matter? Is that too a manifestation of some peculiar properties in the ether? We know now that matter is built up of protons and electrons. But analyzing these into their fundamentals, we find far more than a hint that they are but special modifications in the all-pervading ether. They are essentially resolvable into ethereal energy of a specific kind. Hence, we are beginning to think that matter itself is a form of energy.

Energy is one thing in the physical universe that directly appeals to us. We apprehend it under a great variety of forms. And it is becoming provable that what we call matter is one of those forms. Most of the known forms of energy are convertible one into another. The energy of motion turns into heat, as does the energy of electric currents, unless it is converted into the energy of chemical separation or electric charge. Conversion from one form to another, without loss, is the sign-natural of energy. The proof that matter is a form of energy will not be clinched until it can be demonstrated that matter too is convertible into other forms of energy.

Such a process has not yet been performed in our laboratories, or has it? Though it is believed to be occurring in the giant stars. The interior of stars is at an altogether exceptional temperature and pressure.

This constitutes a laboratory where results can be beyond the scope of our present manipulation. In the light from those stars we see some small residual outcomes of this production of energy at the expense of matter. In their motions, we probably see the same thing. That, which we ordinarily recognize as the locomotive energy of bodies, seems now to be overflow or surplus of the violent constitutional energy within. This energy, at present seems inaccessible to us, but is possessed in enormous amount by the very constitution of the atoms of matter. Fortunately, a few of those atoms have given us the hint. They have spontaneously emitted their energy. We call it radioactivity. It is only the heavy atoms of fission material and other substances at that end of the series, which still retain the property of spontaneous disintegration. The other more familiar atoms seem to have lost that power, and settled down into apparent stability and quiescence. These atoms show no obvious sign of possessing any such power. But to the eye of science, it is there. Means have been sought for getting at it.

Combination of atoms into molecules and the interaction of molecules generally have long been known to give rise to various forms of energy. Witness ordinary combustion and the power of explosives. **Simple atoms like those of hydrogen, can be packed together so as to form the more complex atoms of higher elements. This process will liberate vast stores of energy, much greater than could be obtained from ordinary kinds of chemical combination.** It is highly unlikely that this will go on spontaneously or uncontrollably or dangerously, under such conditions, as we are familiar with on Earth. They may be violent enough under the conditions in the interior of stars, including perhaps our Sun. But here, on Earth, they are traceable, guided and controllable. They will not run rampant and do damage, except by reason of bad or malevolent arrangements and then only on a relative small scale.

How soon energy of this kind may become commercial, no one can say. The practical way is by using energy from the cosmos; i.e., not by so-called fission material.

Another Testimonial

TO WHOM IT MAY CONCERN,

This is to certify that on the evening of March 16, 1929, along with Dr. Wilkinson, of Cedar City, I witnessed a demonstration at the laboratory of T. Henry Moray, Salt Lake City, Utah.

Dr. Moray claims to have invented an apparatus that will produce electrical energy without the use of a prime mover. This apparatus was demonstrated on this occasion.

1. Description of the Apparatus

The apparatus uses an antenna specially balanced or an aerial capacitor and a special ground wire. These were connected to the terminals of a switch. Two wooden boxes were placed on a table. On one of these boxes was a high-frequency transformer. In the other box were two sets of condensers, ten large ones in one set and ten small ones in the other set. There were two composition cylinders, each about 1 – 1/8 inches in diameter and four inches long. Each of these cylinders weighed about three or four ounces. In addition, there was another box approximately hemispherical in shape, about two inches in diameter and weighing about two ounces. There were also coils of wires and other equipment. These pieces of apparatus were connected by a number of wires.

Two of these wires were led out to the switch. One was attached to the blade of the switch and the other its' jaw. When the switch was closed, the antenna, lead-in, apparatus in the boxes and ground wire, were all in series.

The other wires leading out of the box were connected to six 100-watt lamps connected in multiple during part of the demonstration. A flat iron was connected during part of the demonstration, as well.

2. Demonstration

During the demonstration the apparatus was connected in series as above described, except that a small switch connected in series with the coil was left open.

Dr. Moray energized and synchronized the device in three or four minutes. The lamps were lighted and remained bright as long as the switch was left closed, which was about 60 minutes. He then connected an electric flat iron. In a short time, the iron was hot. When the "ground wire" was disconnected and then the lead-in was disconnected, the lights went out.

3. Objections people have made

I. That there is a hidden wire from the electric lighting current obtains the power.

II. That the power is obtained from batteries.

4. Tests

Before and after the demonstration I closed the big switch that connects or shorts the antenna and ground. I also made other tests. If the antenna or lead-in were connected to the lighting circuit this would have produced a short circuit. I further tested the device by closing and opening the switch several times to see if any sparks appeared. But there were no sparks. I placed my wet finger between the blade and the jaws of the switch and could not feel any electricity. I touched my hand to both sides of the switch and the wall to check for ground but could not feel anything. We turned the table over and examined it carefully for hidden wires but found none. With the apparatus all connected as when operating the lights the contacts with the switch were moved but produced no arcing. This indicates that the circuit was dead.

While the demonstration was being conducted and the lamps were receiving energy through the apparatus, the main switch, that controls the lights in the building, was opened. All lights on the house circuit went out but the lights on the Radiant Energy circuit were not altered. They were neither brighter nor dimmer at that time. Thus the lights could not have received their power from that source.

The condensers were thoroughly tested. The terminals were shorted, the positive to the negative. If they had been batteries they would have showed a spark. But no sign of spark appeared. I connected them to the electric terminals then tested them. After being charged, the large condensers gave a vigorous discharge, showing a brilliant strong arc and a loud sound. They showed a sudden discharge as condensers are supposed to do and batteries never do.

The small condensers were less vigorous in discharging. They displayed the same snappy discharge of a condenser and not how a battery discharges. These tests proved positively that condensers and not batteries were in the cases. Besides no batteries of such size, could produce such power. The boxes were completely emptied thus leaving no possible place for batteries to be stored. Besides the boxes were not large enough in order to hide batteries.

During the time that the lights were burning the connections with the big switch were moved along the switch and vigorous arcing occurred. Thus, proving that electrical energy was passing through this apparatus.

5. Conclusions

The electric lamps received energy from some source. During the demonstration, which lasted for more than an hour, the lights were brilliant at all times, just as bright at the end as at the beginning of the demonstration.

The lights were a different color and brighter and whiter than those on the house circuit. The electric energy that lighted the lamps and heated the flat iron was not received from the house circuit.

One is therefore forced to the conclusion that the electric energy was received from some other source. As difficult as it is to understand, with our present knowledge, the only conclusion that can be drawn from the demonstration is that the energy was received by and through the apparatus as claimed by Dr. Moray.

T. J. YATES, KE N.E.

This next letter is from Murray O. Hayes Ph.D.

To all whom it may concern,

It is now more than two years since I first became acquainted with Dr. T. H. Moray and the work he is carrying on. In that time he has demonstrated inventive ability of an exceptional order.

Perhaps the most wonderful of his inventions is a device whereby he is able to draw electric power from the Cosmos. This energy is not derived by induction from power lines, as has been suggested by some, nor is it derived from radio stations. This has been demonstrated by taking the apparatus more than 50 miles from the nearest power line and over a hundred miles from the nearest radio station. At this location it operated just as well as anywhere else. This device was subjected to an endurance test in which it was operated continuously for a week. During this test a 100-watt lamp was lighted simultaneously along with the heating of a 575-watt standard Hot Point flat iron, making a total of 675-watts. It is very evident that no batteries could sustain such a current drain as this. Besides the current is high frequency.

He has also invented a very sensitive sound detector whereby it is possible to hear conversations carried on in an ordinary tone of voice, at a distance of several blocks.

He has also worked out numerous radio designs that eliminate many of the parts now considered necessary for good reception. Yet, there was no apparent diminution in quality or volume. In fact, there was a notable elimination of interference from static when some of these designs were used.

He has devised a means for measuring with some degree of accuracy the energy evolved during mental activity. His device shows variable deflections of a sensitive galvanometer needle that appeared to be related to the rigor of mental activity.

There are a great many other equally remarkable things that he has done. For example, reducing old rubber from truck tires to the state of a viscous fluid. This fluid is readily vulcanized without the addition of a smoke sheet as is necessary with other processes. He also created a high frequency therapeutic device and numerous other devices, which show great ingenuity.

Murray O. Hayes Ph.D.

This letter is addressed to Mr. Lovesy is from Murray O. Hayes, Ph.D.

Dear Mr. Lovesy,

As promised during our recent discussion, I am writing about my acquaintance with the construction and operating principles of the Moray device for utilizing cosmic energy.

You are already aware that I have seen many demonstrations of what this mechanism will do. Also, I have seen the parts of which it is built. Recently Dr. Moray has shown to me the wiring diagram of the device. I can find no inconsistencies in it, not anything that does not appear to be logical and sound. While this device appears to be very complicated, when looking at the machine, it is in reality very simple in essence. It is based on reorganized laws of electricity, when all is explained. There are many features which appear to be incidental, but they are in reality of basic importance.

He has also shown to me and explained the detector that his device uses. In this he has applied a fundamental principle of electric circuits which, I believe, would not have been noticed unless pointed out by him. This detector also has numerous features which appear incidental but are the heart of the matter and of first importance.

In my own home he connected his detector into a crystal set for receiving radio in place of the crystal. The reception was better than with the Erla crystal, though the antenna was merely one of the bell type. He also took a lump of lead treated according to a process of his own. He used in it place of the crystal and got wonderful reception of radio signals. They were loud enough, in fact, to operate an old fashion horn speaker of the type put out of RCA about 1923.

I was recently present when an electrical engineer representing the Russian government was given a demonstration of the energy machine. He at first said that the amplification of radio waves accounted for the output energy. However, it would be a real achievement to amplify such waves sufficiently to light six one hundred-watt lights at one time. This is in addition to heating a flat iron of the standard five hundred seventy-five-watt type. When engineer saw the inside of the device, he admitted that it could not have been radio waves. He remarked many times, "It is very interesting."

This machine has been operated in my presence so many times, under so many different conditions of weather and of Season that I am positively convinced that it is what its' inventor claims it to be. I also believe that its' commercial adaptation is feasible. I believe that Dr. Moray has explained everything to me without reservation. I am sure that this is a revolutionary and epoch making invention.

Sincerely yours,

Dr. Murray O. Hayes

Patent Correspondence

Dear Henry,

For your information, I wish to report on the matter of your radiant energy machine since coming here.

We arrived here Friday, July 10. On Monday, July 13, I spent the whole day in the search room of the Patent Office. I examined every patent issued by the U. S. on devices for the reception of radiant energy. Most of them were so obviously dissimilar to your invention that they did not consume much time. Not one is there which shows, on reading, even the remotest resemblance to your radiant energy machine. Not one mentions such an idea.

Neither is there any patent that suggests anything like your oscillator tubes. I feel perfectly safe in saying that there is not one chance in a hundred thousand that the examiner will find any disclosures that will restrict your claims. I have always felt certain that you would obtain broad protection on your radiant energy invention. But now I know it for certain.

You may use this letter in any way that you see fit.

Very truly yours,

Murray O. Hayes

Thomas J. Yates, E.E., M.E, wrote this notarized letter

TO WHOM IT MAY CONCERN,

I witnessed a demonstration of the apparatus invented by Dr. T. H. Moray by which he produces electrical energy. A report of that visit has already been made.

I now have, about two years later, witnessed a second demonstration of the same device; some changes in the apparatus were noted. A great improvement was made in the performance. The condensers, coils and other parts, noted at the previous test, were now in one small box about 10 x 10 x 30 inches. A safer and better means of starting the device has been devised. Once the apparatus is put in operation, it will continue to charge itself.

The lamp rack that formerly held six 100-watt lamps has been replaced. The new rack is larger and holds thirty-two 50-watt lamps, three 100-watt lamps and a receptacle to receive an extension plug and cord. On the other end of this cord a 575-watt electric iron was connected. During this test, all of the electric lamps were lighted at one time. Their light was whiter and more brilliant than other lights in the room, which were supplied from the house circuit. When the lights were on bright, the electric iron was plugged in. There was no noticeable blinking or sudden diminution of the lights as in the usual house lighting systems when an electrical iron is turned on. The lights were as bright as before the iron was plugged in. The iron heated quite rapidly. In about 2½ minutes it was hot enough to be used for ironing.

I made a thorough and careful examination of the entire system to see if electricity could have been introduced from some other source. I inspected every part of the system, all wires were checked; the switch was removed and examined. Leads from the switch to the box containing the condensers and coils were checked. The box itself was lifted and examined. The leads from the box to the lamp rack were inspected. The lamp rack was lifted and examined, as was the ground wire from the switch to the water pipe on which it was grounded. Not an inch of space was overlooked where there might be any metal that would carry the power. I am prepared to say that no such connection existed from which power could be had from any other system.

A crystal radio set was connected to the lead-in wire from the antenna on one side and to the ground wire on the other. Perfect reception was had. If there had been power from other sources on the antenna or lead-in wire radio reception would not have been possible. Instead a loud humming sound would be produced. This was demonstrated that by connecting the radio set to the house lighting system.

When Moray's system was working, the lights were burning bright. The terminal attached to the antenna was disconnected. A vigorous arcing occurred and the sparks jumped over an air gap as long as eighteen inches. This would indicate that the power was high voltage. The nature of the sparks indicated a high frequency current.

I do not understand the principle by which Dr. Moray's device produces the electric energy. The condensers and the coils of wire are common. The cylinders called "oscillators" and the small conical shell called the "detector" are the only things not commonly known, but the system works. It produces electric power in abundance and does all that Dr. Moray claims for it.

I do not own any stock or interest in Moray's work. My only purpose in issuing this statement is in the interest of the advancement of science. I consider this development a great advance in the science of producing electrical energy.

Very truly yours,

In witness to the above I hereunto sign my name.

(Signed) Thomas J. Yates, E.E., M.E.

STATE OF UTAH COUNTY OF SALT LAKE

Thomas J. Yates, being first duly sworn, deposes and says that he has read the foregoing statement and acknowledges that he wrote and signed the same as above set up.

(SEAL)

Thomas J. Yates

Subscribed and sworn to before me this 18th day of December 1930.

My Commission expires R. J. Chapman May 7, 1934 Notary Public – residing at Salt Lake City, Utah.

This is a letter addressed to Moray from W. H. Lovesy

Dear Henry,

I have handed you the original letter from Murray O. Hayes, in which he described being shown and having explained to him the wiring diagram of your R.E. device. Also the fundamentals of the detector were explained to him. In that letter he stated plainly these fundamentals would not have been noticed unless pointed out by you. At the same time, he had become familiar with all of the details of the construction of your device for utilizing cosmic energy. He added that you had applied only fundamental principles of electric circuits.

There is attached a second letter containing details of Mr. Hayes educational record.

I made the memorandums on the face of these two original documents at the time I was consulting with Murray O. Hayes in my office. At the time he definitely advised me of being shown every detail of your invention. Mr. Hayes stated that he could make one of the machines himself from the information that had been given him.

Murray O. Hayes made similar statements to me several times. I have made repeated requests that he makes these declarations. Namely that he had been shown every detail of the invention. Perhaps these repeated requests gave him the impression that I had a doubt about the invention.

Personally, I have never had a doubt. But I thought that only you held the real detailed secrets of the invention. I had a conference in New York with Dr. Harvey Fletcher and Dr. Carl Eyring at the Bell laboratories. I agreed with them to ask you to give the details of your invention to Murray O. Hayes. The purpose was to determine if he could duplicate your machine. In every instance his answer was emphatic and to the effect that he could do so.

Yours truly, W. H. Lovesy

This letter is from Geo. R. Pyper

To Whom It May Concern,

I have worked in electricity all my life. I was with the Utah Power and Light Company for thirteen years and worked in all departments including sub-stations. I have been with the Kearns Corporation for over seventeen years. I have been in charge of all the electrical work for the Tribune and Telegram Publishing Company and in the Kearns and Tribune Buildings.

In last December I witnessed a demonstration at Dr. Moray's laboratory of his electrical marvel box. He let me see inside this box. There was a H. F. transformer, some of his cold tubes and some condensers. Dr. Moray connected this box to a special balanced aerial and ground wire from outside the building. Two of us held a counter poise antenna attached to glass insulators in the room. When he connected the box to this counter poise antenna, I saw the same results as with the external aerial.

During this demonstration and while the lamp and appliances were operating, I shorted the aerial and ground wires. There was no spark; it just turned off the power from the box. I then took hold of both of these wires and felt nothing, they were both cold.

He then took a larger box, about 18 x 30 x 15 inches and connected it to the outside aerial and ground wire. From this box he lighted about fifty 100-watts, 120-volt lamps, ran a small motor specially wound at great speed, an original electric iron and a 500-watt glow heater.

Dr. Moray then disconnected the appliances and lamps and attached two long wires to the box. We pulled the main line, Utah Power and Light Company service switch for his building. We attached the wire from the box to the building side of the switch. He lighted his building and heated iron heaters. In fact, he did everything the Utah Power and Light Company service would do except run motors that I understand have to be specially wound.

Standard globes were used. They seemed to give a softer whiter light that was more like daylight.

I am satisfied from my experience in electricity that there was no fake of any kind such as concealed batteries or wires. Everything was in the open so I could see every operation. I did not see the inside of the larger box.

It was a very remarkable demonstration and one that I will always remember.

Geo. R. Pyper

SECTION SUMMARY

Whenever a fairly large amount of oppositely charged ions coexist within a small volume of space extensive recombination occurs. The resulting accumulative radiation can add up to an intense, constant wide-band frequency. This phenomenon prevails around many stars in our own galaxy and is concentrated around nebulas and star clusters. Indeed, this is a very old source of energy. All we have to do is attach our machines to this "wheelwork of nature."

SECTION REFERENCES

"The Sea of Energy In Which The Earth Floats" by T. H. Moray, 5th Edition 1978

The Eyring Research Institute, before Henry Moray died, was working on "Direct Energy Conversion Systems." This involves using radioactive material in conjunction with the quartz junction. p. 89

In 1942, shortly after World War II began for the United States, Henry Moray attempted to rebuild a radiant energy device. He used the remaining bit of what was known as the "Swedish Stone." This material limited the amount of power that his device could draw. Consequently, in the larger unit, he developed a second detector that forced him into extensive research involving nuclear materials and radioactive reactions. He became deeply involved in the study of synthetic radioactivity as described by Gustave LeBon in his book, "The Evolution of Matter." p. 186

Others may discover Direct Energy Conversion Systems using synthetic radioactive materials. Bell laboratories have made millions of dollars from semiconductors, though no credit has been given to Henry Moray. p. 190

This is from a speech given by T. H. Moray on Jan. 23, 1962 at the Valley State College Northridge, California

It is now estimated, because of present advances in dielectrics that a 100-lb. unit can be made to deliver 300KW. That is gross weight, not net weight. p. 11

"The Sea of Energy In Which The Earth Floats" 4th Edition 1960, by T. H. Moray

Nikola Tesla was not referring to so-called atomic energy or nuclear energy but to the energy that is continually bombarding the Earth from outer space. Call it cosmic, or what one will. p. 1

Enough energy is coming to the Earth to light 1,193,600 one hundred-watt lamps for every human being alive today. No fuel of any kind will be taken. Energy can be "picked-up" directly by great ocean liners, railroads, airplanes, automobiles or any form of transportation. In addition, heat, light and power can be available for use in all kinds of buildings. p. 1

Electrons are spontaneously being emitted from the nuclei found in Nature. Every new discovery on the subject bears out the claim that all "space" is filled with energy containing millions of amperes at very high voltages. p. 89

In the case of R. E. devices, a different effect is established. The oscillations from the Universe are trapped by the R. E. valve and continue to enter the circuit as the waves of the sea beat upon the shore. p. 92

Pellets were made of a mixture of the **Moray lead** that withstands heat over 1800°F. Using pure germanium mixed with bismuth, iron sulfide, triboluminescent zinc and certain other impurities **including the Moray fission material,** he obtained a substance that had wonderful properties as a detector or valve for radio signals. p. 128

With this germanium combination alloy used as a detector, it was found that radios would operate without batteries or any source of power other than that obtained from the **radio transmitting station.** p. 128

To make this special R. E. energy valve, he used a lump of lead (Moray lead). It was treated according to the process that he has discovered and got wonderful reception on a radio. p. 129

"Radiant Energy" 3rd edition by T. H. Moray, 1945

This energy, or as Dr. Moray explains it, these oscillations of energy, are picked up by the device through the oscillators, or neutron bombardment. p. 8

In reference to electrons, neutrons, protons and ions: It is my theory, in using these terms, that they are the energy of the Universe. p. 17

My device oscillates because of the oscillations of the Universe caused by the disintegration of matter. p. 18

Radiant Energy 2nd edition by T. H. Moray, 1931

As far as atomic energy is concerned for heat, light and power, it is and always will be nothing more or less than an expensive, dangerous, glorified steam plant or an equally dangerous thermoelectric device. The process involves breaking into the lines of force of the radiation field that surrounds the reactor to capture energy. Such a plant, no matter how efficient, never is the complete answer to the World's energy problems. Costs and weight will always be a great problem.

Any nuclear fuel, even the most "super nuclear pile reactor," or what have you, is, and always will be, just another way to operate an energy plant from any heat, light and power viewpoint. p. 1

When Moray used the word "Universe" he was often referring to the atom. Moray refers to cosmic energy that relates to both "stellar space" as it does to "atomic space." p. 61

The facts involved are the same in both the radio receiving set and this device. The nature of the mechanism, however, in one differs greatly from the other. The radio receiving set receives transmitted energy waves out of the air and transposes them into sound waves. The R.E. device receives oscillations from the Universe, and transposes them into electricity. p. 13

If the frequency is raised high enough, molecules will separate and the atoms are freed. Raise the vibration of the atoms still higher and higher and they will resolve themselves into the original elements of which all matter is constituted. p. 15

Cosmic rays are constantly creating radioactive carbon. The fusion of small atoms together to make larger ones gives off more energy than so-called "splitting" of the larger uranium or plutonium atoms. The uranium energy release ratio is only 1 to 1000 – that is only 1/1000 of the heavy atoms are changed in the uranium "atom splitting" process of the atomic pile. p. 60

11 April 1990

Arthur Glenn Foster
Principal Engineer wrote:

There are some interesting side details. "We" (a group of us in government and industry, loosely coupled) were pursuing "Project X". We had a contract to process some of the Moray "valve," fine-ground, "Swedish Stone" into pressed "detector" pellets. The pellet material was doped with a tiny amount of unpurified radium reduced from U238 yellow cake.

The pellets were subjected to a huge number of millibars of pressure by a tetrahedral press used for manufacturing artificial diamonds. This technique was developed and patented by Dr Tracy Hall after he left GE. When a certain critical pressure was reached "something" in the material "flashed." The electric pulse produced was so violent that it burned out the 2000-amp carbon-die heater power supply, blew the main breakers in the building and blew the power line transformer on the outside power pole!

My Radiant Energy Research

Whenever radiation from a radioactive substance ionizes atoms, electrons are emitted. The electrons almost immediately recombine with the positively charged atoms that expelled them in the first place. Charge equilibrium occurs at the exact moment of recombination, along with electromagnetic oscillation. A by-product of this phenomenon is called **ionic oscillation**. Radioactive particles will again separate the newly recombined atom. In fact, any electrically neutral atom will lose electrons when exposed to radioactive matter. The atom that is left behind becomes a positive ion. This process of ionization and recombination is a continuous cycle as long as the atom is exposed to a source of ionizing radiation.

Whenever a fairly large amount of oppositely charged ions coexist within a small volume of space, the accumulative radiation from recombination can add up to an intense, constant wide-band frequency. This phenomenon prevails around many stars in our own galaxy and is concentrated around nebulas and star clusters. Indeed, this is a very old source of energy. All we have to do is attach our machines to this "wheelwork of nature."

What has just been revealed to you here is the final secret to harnessing the very wheelwork of nature. Ionic recombination of the decay products of radioactive isotopes will produce intense surges of electromagnetic energy. More often the wavelengths created are in the infrared (thermal) range. However, the wavelengths generated can also manifest within the radio frequency range or even far into the **gamma** end of the electromagnetic spectrum. They can even be made to manifest as free electrons. What does the decay of radioactive matter teach us about the nature of energy?

Under the right conditions, radioactive particles can be made to generate intense surges of energy in the radio frequency range. Tune into this energy with a tank circuit that will resonate with it and you have yourself a powerful source of electrical power.

To the best of my knowledge no person has ever harnessed radioisotopic energy using the process that I have just explained, with the exception of T. H. Moray and myself. There is not a patent to be found that even closely resembles it. Radio astronomy only gives us hints about these **"ionic oscillations."**

Many radioactive isotopes could be used to generate electrical power. However, the real power behind these isotopes is the radon gas that they release. I have pointed this out many times in my writings, lectures and articles. Polonium can be cheaply manufactured by bombarding the metal bismuth with a source of neutrons.

In United States Patent 2,728,867, it is stated that each gram of nuclear fuel that is undergoing fission gives off around twenty watts of power. In comparison, one gram of **polonium** will generate about one hundred and forty watts of power. This is **seven times** more energy than what is available from a conventional nuclear reactor!

If polonium is made to generate surges of intense electron oscillations then it becomes a powerful source of radio energy. A simple **ion valve**, or call it what you will, can be used to generate these oscillations. They in turn can be directly converted to electrical power. This unique valve is shown in **Figure 4.** Before it can function it must be connected to a source of high-voltage. In order for electrical energy to be extracted, it must also be connected to a proper circuit. One such circuit is detailed in **Figure 5.** The electronic circuit shown is basically what is required for this conversion. Many additional stages can be added for more power. Many other circuit and component configurations can also be used but the fundamental conversion principle will remain the same. The circuit and components shown in this book will give you the general idea of the principles involved.

The actual mechanics and electronics of building and validating a radiant energy prototype are relatively simple. It is fun and easy. Moray's devices used a detector that was ionized by radioactive impurities. However, radioactive material need not be used, from what we now know.

Those involved in building prototypes, must first confront some serious issues before beginning. I will try to enumerate a few pitfalls and make descriptions as clear as I can. This will fortunately and blessedly push some of you into action, but for others it will be viewed as an impossible challenge.

What are the most important assets for beginning my research?

1. Dedication
2. Motivation
3. Passion
4. Do instead of just theorizing.

What is the second most important thing?

The lone wolf researcher will need a lot of skills if their project is to be successful. Among the most valuable skills are those involving electronics technology. Plus, some familiarities with machine shop practice and materials. A good sense of organization is important. Being able to stand-back and observe the big picture, along with a lot of plain, good old common sense; are necessary skills to have. The "hands-on-imperative" must be at the very core of your efforts.

What if I have few tools and skills but have the dedication?

Then you need to have some cash, the green stuff. You will have to buy the skills, knowledge and time of others. You can farm out electronics and machining chores but it will be expensive. The more that you can do for yourself the better off it will be for you in the long run.

A fully equipped, skilled, advanced amateur radiant energy experimenter might slip by for about $300.00 for everything. If you have only a few tools hanging around then a proof of concept prototype might cost you around $500.00 because of the extra tools that you will have to purchase. Later you can build a more powerful prototype to replace the proof of concept design. Your basic working design will not have to be replaced. You can add more components as your research progresses. Simple dedication goes a long way. Watching, paying attention to details and doing it right the first time; are all very important. With this borne in mind, you could complete a powerful prototype for around $5,000.00 that could provide power for the average home. However, it is not now my intention to instruct you on how to build a home power unit. This book was written so that you may validate the technology; refinements still have to be made.

Are there any dangers involved?

Yes, of course there are! There are a number of ways to wind up in an early grave! Most of the lethal dangers can be avoided with **skills**. Read first and do second. Reference books and a solid, broad-based scientific library are your true assets. Access to the Internet would also be very helpful. Know the danger zones in the various disciplines that you are going to come across in your research endeavors.

The radiant energy circuitry is connected to an antenna and ground. A lightning strike could fry the device and burn down your house to boot. I take no responsibility for this. You build these circuits at your own risk. Take the time to install lightning protection on the antenna and ground leads. It could save you a lot of grief in the long run.

Some specific dangers include:

1. Crushed and mangled hands of the unwary or metal splinters in the eyes or skin.

2. There are burn and eye damage hazards.

3. Imploding glass bell jars, power valves, can scatter glass splinters at high velocity into face and body. Always use glass enclosures that can withstand the vacuum that is being applied.

Do you have what it takes to go the distance?

If you have what it takes and there is a burning passion for this research then you are probably ready to begin. This is not a place for a wimpy mentality. If you become frustrated or fail in your experiments, have the courage and good grace to realize how this came about. Make the effort to find out where you made your mistakes.

More than likely, you are going into this research with learning from books that have been written to hide certain facts from you. Do not expect results over night. Learn all that you can by careful observation, educate yourself, search out the truth first and let Nature be your guide.

There are political powers in this World that would not like to see us succeed in getting radiant energy power implemented. They know quite well that this energy source will liberate the global community. This mode of power generates not only power, but it will also generate hope for our decaying World.

Fundamental Validation Test Circuit

First of all, let me say, there is no fringe science involved with radiant energy power generation. This is a very old source of energy being extracted and harnessed in a unique, new way. I do not claim to have discovered anything new in nature. What I do claim is an improved method to convert radiant energy (atomic ions) into useful electrical power. Think of the radiant energy receiver as a type of electronic heat pump, only it is pumping ions, not heat. No laws of physics are being violated. No new laws are being implied.

Energy from the Dissociation of Matter

It is no secret that electrical current can be obtained by utilizing a couple of dissimilar metals. You can plunge a dissimilar couple into an electrolyte to obtain electrical current. This is called electrochemical energy. You can heat the junction of a pair of dissimilar wires that have been twisted together at their ends to generate electrical potential. This is called "thermoelectricity." However, what is not generally known is that you can also generate electrical current with plasma or a glow discharge as your medium between two dissimilar electrodes. This effect is the result of "triboelectricity." All these forms of electrical energy are caused by the dissociation of matter. This dissociation of matter is called "radioactivity" by the scientific world. All electrical phenomenon is the result of this radioactive decay.

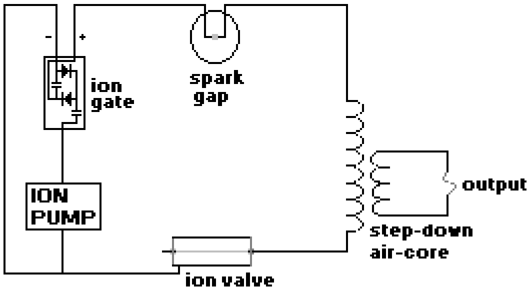

Proof of Concept - Figure 3

- The "ion gate" in this circuit uses two .001mfd/20kvdc capacitors and two 20kv, 100-nanosecond diodes.
- There are varieties of electronic circuits that can be built to serve as ion pumps. The circuit used was designed by Robert E. Iannini and can be found on my website at: http://www.nuenergy.org/plasma.htm
- The ion valve in the above proof of concept prototype consists of a standard aluminum electrical conduit that is six inches in length. Its center wire is a 60 µm diameter tungsten wire secured with two half-inch PVC end caps.

Preliminary Ion Valve Validation

1. The output at the ion gate reads around 5KVDC using an analog high voltage test probe without the ion-regulator valve in the circuit. With the ion valve in circuit a 20KVDC reading is obtained with at least a 25% reduction in current being drawn to power the circuit.
2. The apparent energy gain in the above ion valve appears from the fact that when negative ions produced at the surface of the inner cylinder bombard the center wire secondary negative ions are generated. The negative ions, when accelerated with sufficient voltage are capable of dislodging (dissociation) other negatively charged ions by from the target against which they impinge upon; and a significant result is that the number of secondary ions may be several times greater than the number of primary ions. This has the effect of reducing power drain from the power supply. A.W. Hull observed this effect in a thermionic tube – HULL, Proc. I.R.E., 6, 5, 1918.
3. The ion valve can be used as a quenching (regulator) device that can provide clean sinusoidal currents to the output current.

Ion Valve Technology Explained

An ion valve that has been preconditioned with radon gas will uniquely add electrons to an applied high-voltage source. Inputted energy can be converted to a higher wattage electrical output. Using an ion valve in this embodiment the basic design has an axial, negatively charged cathode wire that extends the length of the cylinder. The anode is made out of an aluminum cylinder that is positively charged. Its cathode is a tungsten wire or rod that is negatively charged.

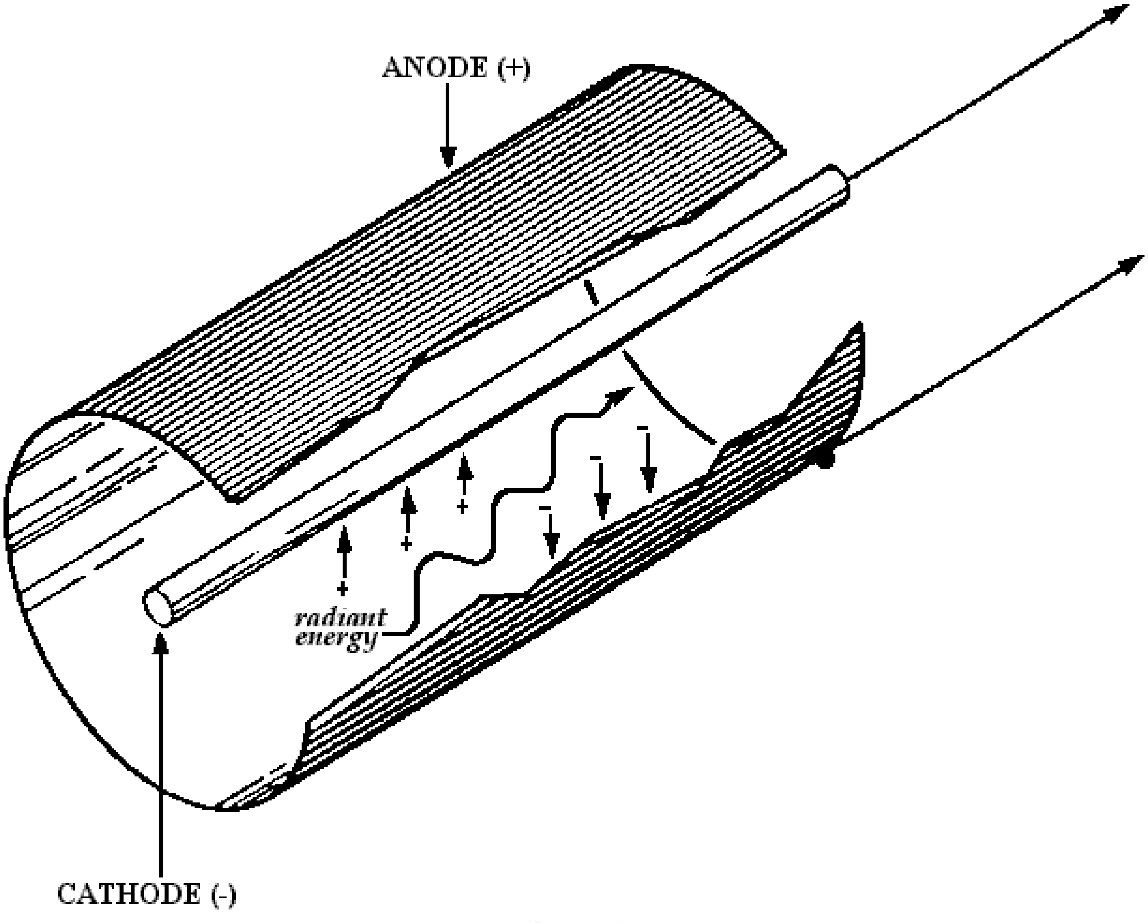

Ion Valve - Figure 4

In the diagram shown here negatively charged ions are formed at the surface of the center cathode wire. Positively charged ions are formed on the inner surface of the anode cylinder.

When a surge of ions flows through the valve they are added to the negative ions that sit on the cathode. They are now attracted to the positive anode cylinder. A few milliseconds later the positive ions on the surface of the inner cylinder are attracted to the negatively charged onrushing ions. When these negative and positive charges collide they neutralize each other generating electromagnetic oscillation. This characteristic has the effect of suppressing any **inductive kick back** seen from the loads that would normally flow back towards the power supply. Feedback from the loads to the power supply will result in wasteful heat losses.

I am fully convinced that my ion valve is a refined version of Edwin Gray's conversion tube described in his US Patents No. 4,595,975 and No. 4,661,747. **Chauncey J. Britten** also describes a similar component in his US Patent No. 1,826,727. Mr. Britten's component appears to be a prior variation of my ion-regulator valve design. Britten's valve was said to have lit up his home in 1930 according to a local newspaper article of that time period. Russian scientist **Alexander Chernetski** experimented with what appears to have been a type of ion-regulator valve. It is said that he got up to five times more energy out of his device than what he put into it. Plans were made to commercialize the invention but like the Moray, Gray and Britten devices it never became a commercial product.

The ion valve is not to be confused with a current rectifier. The ion valve prevents ions from flowing back to the power source. It is an 'energy dam' for the lack of a better description. Both negative and positive ions can flow through this component. The valve has a charge blocking effect. This is why Moray stated that "the valves are not rectifiers in the sense that they operate as radio valves in changing AC or HF to DC. Ion valves have an actual valve action in stopping the 'flow' of energy in one direction. Its function may be thought of as an oscillatory action similar to the waves of the sea, without rectification, from returning to the outer circuit." p. 209, The Sea of Energy, 5th edition. This rather cryptic language now makes more sense seeing first hand how my ion valve works.

The metal atoms in the cathode wire are bonded to each other through electron sharing. In another embodiment, to break these bonds the wire is given a positive charge and the cylinder is given a negative charge. The negative charge on the cylinder negatively ionizes any gas that contacts it. These ions rush towards the positive cathode. When an ion that carries an excess electron hits head on with a positively charged metal atom the electron combines with the metal atom. The absorbed electron passes along its kinetic energy onto the metal atom causing it to break its bonds with the atoms that were joined to it. This generates an intense burst of free electrons. The "sea electron" model helps to explain this effect. It is where metal atoms are bathed in a "sea" of valence electrons. If this model is taken one step further it can be shown that when metallic atoms are dissociated from each other excess electrons are released as pure electrical energy. This occurs because the electrons no longer take part in the intermolecular binding force that existed before the dissociation took place. The freed electrons will add amperage to the output circuit to which it is connected. Henceforth, the equation **I x E = P** holds true in this system. Where, "I" represents electrons (amperage), "E" electromotive force (ionic voltage) and "P" is the power generated.

The ion-regulator valve is an important component to the radiant energy converter. Add a grid to this valve and you then have an ion driven oscillator component. The ion-regulator valve can be used as a rectifier by heating its cathode. Electric motors can be made to draw less current from the power supply. Communication systems that take advantage of the ion valve as an oscillator could make present day radio transmission obsolete. Electrical power transmission and its generation could also receive a fresh new start. The possibilities are refreshing. It is an honor to offer humankind this technology. It is my goal to build and bring devices to market that will introduce new products that will better the quality of our lives.

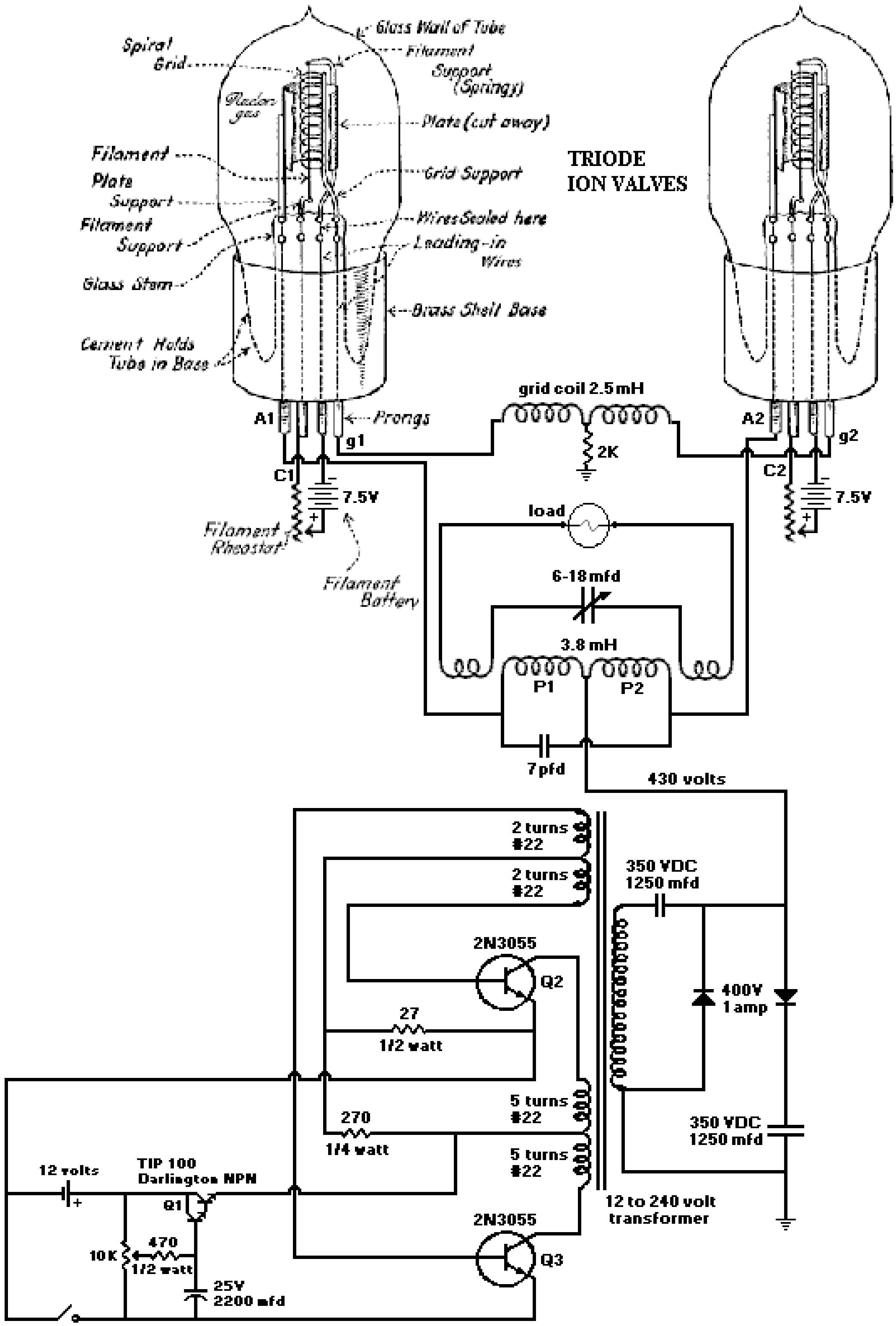
Spiral Grid
Glass Wall of Tube
Filament Support (Springs)
Radon gas
Plate (cut away)
Filament
Grid Support
Plate Support
Filament Support
Wires Sealed here
Leading-in Wires
Glass Stem
Brass Shell Base
Cement Holds Tube in Base
TRIODE ION VALVES
A1
Prongs
g1
C1
7.5V
Filament Rheostat
Filament Battery
grid coil 2.5 mH
2K
A2
g2
C2
7.5V
load
6-18 mfd
3.8 mH
P1
P2
7 pfd
430 volts
2 turns #22
2 turns #22
350 VDC 1250 mfd
2N3055
Q2
400V 1 amp
27
1/2 watt
5 turns #22
270
1/4 watt
350 VDC 1250 mfd
5 turns #22
12 volts
TIP 100 Darlington NPN
Q1
12 to 240 volt transformer
2N3055
Q3
10K
470
1/2 watt
25V 2200 mfd

Radiant Energy Generator Concept Device - Figure 5

To activate an ion valve a vacuum is pulled, radon gas is introduced and a negative charge is applied to the cathode wire. The radon decays and deposits lead–210. This is a beta emitter and will supply electrons without the need for filament current.

T55 Taylor electron tubes were used in the above public demonstration device due to nuclear regulatory restrictions. Other radio vacuum tubes may be substituted. The energy output of this device does not give the slightest shock when the hands are in direct contact with it. Voltage meters will not respond to this type of energy. A 100-watt light bulb placed across the output will be brighter than if powered by conventional current. The filament within the light bulb can be seen to glow a cherry red and the gas is excited throughout. It is not a form of electricity that we are accustomed to experiencing. This type of electrical current should not be confused with what powers the circuitry of **Figure 5**.

The source of energy that drives the circuit is a wavelength of energy that T. H. Moray loosely called "radiant energy." To obtain power from this energy source a receiver must be built that will oscillate in harmony with it. This particular type of radiant energy possesses great powers of penetration, exceeding, in fact, the penetrating powers of ordinary x-rays. These rays have been proven to be harmless and may be used, if desired, to produce clear radiographs. They are antiseptic, germicidal, and are good tissue builders possessing wonderful healing powers. They will also effect photographic plates, produce phosphorescence in certain substances, render the surrounding air conductive to electricity and can produce heat in some materials.

WARNING!

Electricity is EXTREMELY DANGEROUS and caution should always be used when working around it. Please take the time to educate yourself sufficiently. If you don't understand something, investigate further before attempting it. Take liberal advantage of encyclopedias, science books, and other reference materials. This book was compiled for information purposes only. I do not control the materials used in construction, the methods of construction, nor the applications of the construction, therefore; I AM NOT RESPONSIBLE FOR DANGEROUS OR UNDESIRABLE RESULTS. Always use rubber gloves when working with high-voltages.

ALL SAFETY PROCEDURES MUST BE CAREFULLY OBSERVED
MEASURING HIGH VOLTAGE IS NOT RECOMMENDED

Disclaimer: *The author assumes no liability for any incidental, consequential or other liability from the use of this information. All risks and damages, incidental or otherwise, arising from the use or misuse of the information contained herein are entirely the responsibility of the user. Although careful precaution has been taken in the preparation of this material, no responsibility is assumed for omissions or errors.*

Edwin Gray's Energy System

Thomas H. Moray invented a high efficiency 6,000-hertz power supply. It turns out that Edwin Gray's power supply also produced this same frequency. Gray claimed that electrical equipment such as solid state radio; television and light bulbs would operate on this type of current.

The overall efficiency of the Gray device was said to be extremely high. From one of his company press releases we are told that it will power loads four times longer in comparison to a standard DC to AC inverter power supply. This particular press statement indicates that he was planning to market a device that put out a 6Khz frequency, as compared to the standard 60Hz frequency coming from the public utility power grid.

Both the Moray and Gray energy systems harnessed a very old source of energy in very unique ways. They could be the solution to our power needs by using an inexpensive renewable energy. The advantages over other systems are many. The World is indeed moving at a suicidal pace, we need this technology more than ever. It will light lights, run motors and electrical appliances.

When you put a high potential current across two separate elements, ionization will occur between them. The high voltage elements **12** & **34** develop a pre-discharge plasma glow between them. The charge polarity given to the tube **14** in the Gray circuit puts it into an **impulse mode.** When a glow pre-discharge is present in this mode no current flows. The tube **14** functions pretty much like a Geiger radiation detection tube waiting for an electromagnetic surge to trigger it into conduction.

Discharge occurs across the spark gap **62** once the capacitor **16** charges to the gaps' breakdown voltage. Once the arc is broken the capacitor **16** recharges, the cycle repeats and the cycle repeats so long as the batteries remain charged. This discharge current flows along the red wire shown in **Figure 6**.

Consider an electron moving through a magnetic field, whose lines of force are at right angles to its direction of motion. It is known that the electron is constrained to move in a circular path. In fact, if the field is made strong enough the electron may be forced to revolve in a circular orbit whose diameter is smaller than the field. If a magnetic field is applied transversely to the motion of the electrons, they may be deflected to such an extent as to prevent them from reaching the plate. Thus they are constrained to return to the cathode. A.W. Hall of the General Electric Co. was the first to use this principle in a highly efficient vacuum tube converter that he called a magnetron. The sudden discharge through the element **12** generates an intense electromagnetic field that collapses very abruptly. This discharges the energy stored in capacitor **38** in series with the load **36** and capacitor **16**. During this discharge cycle capacitors **16** & **38** recharge and in the same moment a powerful surge of electrical oscillations flows through the load **36.** The collapsing magnetic field in conjunction with the reappearance of the pre-glow discharge instantly generates a pulse train of electrical oscillations. This surge of power runs Gray's unique high frequency motor. The motor functions with extreme efficiency because DC resistance is reduced to a minimum and wasteful heat loss is practically non-existent in this system.

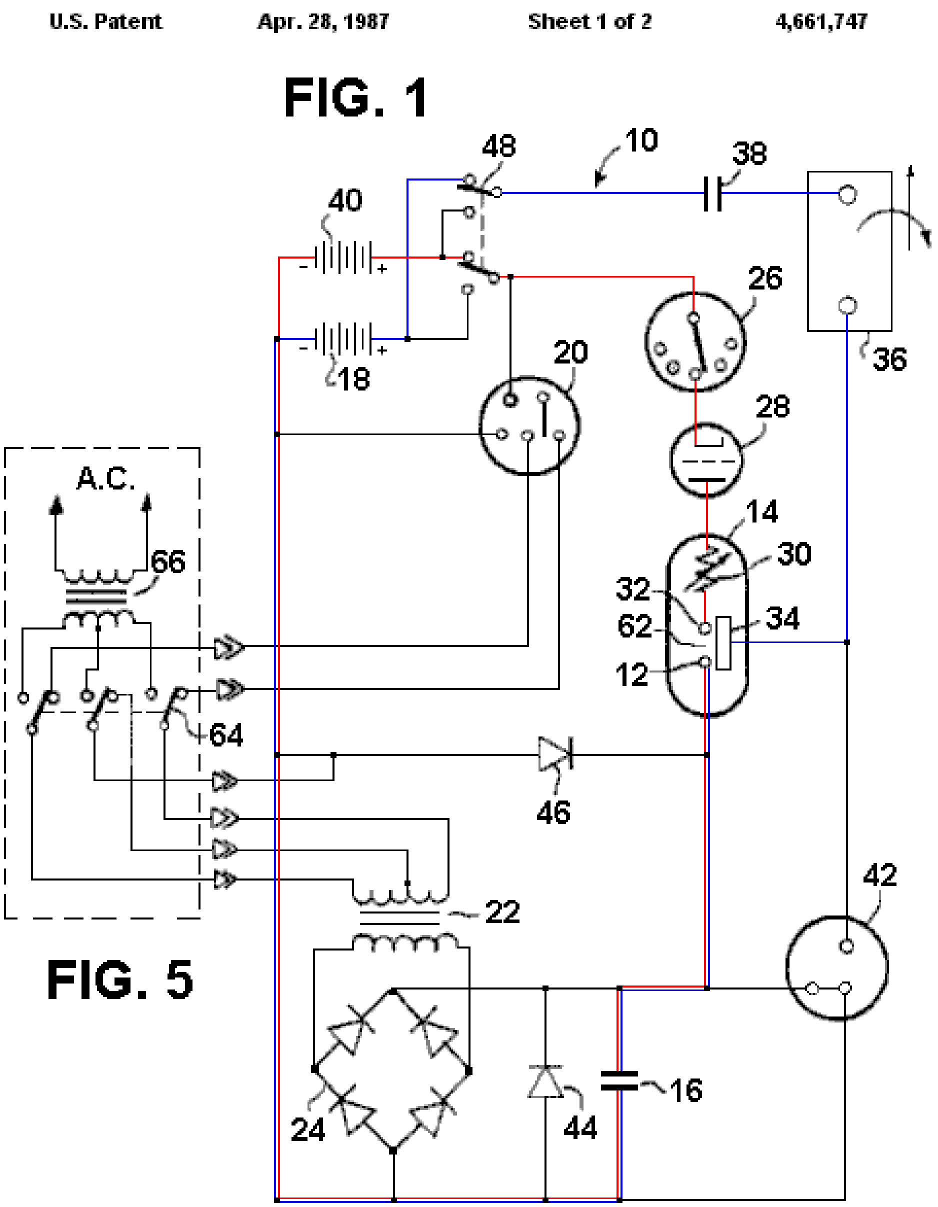
U.S. Patent
Apr. 28, 1987
Sheet 1 of 2
4,661,747
FIG. 1
48
10
38
40
26
36
20
18
28
A.C.
66
14
30
32
34
62
12
64
46
42
22
FIG. 5
16
24
44

Edwin Gray Energy System - Figure 6

1. The blue wires represent the high frequency current path that powers the load **36**.

2. The red wires represent the current discharge path that occurs when the spark gap in the Gray conversion tube **14** fires. Storage batteries **40** and **18** alternately recover the discharge energy through a double-poled switch **48**.

3. When the capacitor **16** discharges an intense oscillating high frequency electromagnetic field is generated through the "switching tube" and load circuit.

Treatise On Matter & Energy

Matter and Energy are one and the same…

Three Primary Particles

Atomic theory does not predict every reaction created by the various states of matter and of energy. There are principles in Nature yet to be revealed. Something has been missing from the over-all atomic picture of matter and energy. There are three primary particles that bring the elements into existence. Just as the colors red, green and blue combine to give us the entire rainbow, the same holds true of the three primary particles: protons, neutrons and electrons. The combination of these three basic primary particles is what gives physical mass to an element and its' unique characteristics. Its physical size, its mass, gives it a signature. This signature is given an "elemental" name.

If a given element is not in a kinetic state then it is said to be at complete rest. It is cold. The external forces acting upon it determine a particle's energy level. When struck by another particle, it becomes kinetic. It is this kinetic activity that we call heat. This heat or kinetic energy level may be raised by particles which impinge upon it that have higher kinetic energy levels. The "heat" of an element may be lowered by other elements that impinge upon it that have lower kinetic energy levels. This will cause the element to lose some of its heat. It is therefore "cooled." However, the elements fundamental make up is not changed. An element may acquire the kinetic energy level of visible light and still retain its fundamental properties. Its' signature is not changed.

Gamma Radiation

An atom that absorbs or expels neutron particles will ring like a bell. It emits gamma energy. Whenever an atom gains or loses neutron energy it generates gamma energy. Gamma energy is a disturbance phenomenon.

The only difference between natural and artificial gamma energy levels is their origins. Natural gamma rays are emitted from the nucleus of a radioisotope. Artificial gamma energy is generated outside the atom. Gamma rays and x-rays are often called photons. They have no mass and no charge and may be considered to be energy disturbances in space.

Wavelength

The kinetic energy level of the radiating element determines the wavelength of energy. A low kinetic level is represented by a long wavelength on the chart. A high kinetic level represents a short wavelength on the chart. The "speed of light" is therefore a relative term. Energy is contingent upon the matter that it impinges upon.

The x-rays begin at about the middle of the ultraviolet portion of the spectrum. These are called soft x-rays. Those that extend into the lower portion of the gamma range are called hard x-rays. Beyond this range, they are cosmic rays. These energy disturbances are called "photons." They are not particles. They are disturbances caused by the kinetic states of particles in motion. This energy spectrum proceeds in order of decreasing wavelength: sound, thermal, visible light, ultraviolet, gamma, and the cosmic rays.

Pair Production

If a photon has sufficient energy, its' energy may be absorbed by atoms. The photon interacts with electrons. This gives rise to a process called "pair production," in which the energy of the photon is absorbed by the atom.

Photo Disintegration – Gamma Energy

Atoms can absorb high-energy x-ray photons, those with energies above 10 MeV (Million electron Volts). Here, the electrons within atoms are raised to an excited state and instantaneously emit high-energy gamma radiation. This makes the atom receptive to nuclear transformation. This process is called "photo-disintegration."

1.02 MeV Energy Barrier

It has been calculated that when one electron pairs with another, 0.51 MeV of photon energy results. Since two electrons are involved in a pair production event, a photon must have at least 1.02 MeV of energy to separate them. A photon with less than 1.02 MeV will not produce a pair production event. Any energy in excess of 1.02 MeV is distributed equally between the two electrons.

Electrical Energy – Electron Cloud

An atom is negatively ionized when it gains electrons. The atom is positively ionized when it loses electrons. How can it be then that a helium particle expelled from radioactive decay is charged to millions of volts positive? This observation can mean only one thing, that the electron shells are really an abstract thought for mathematical calculations, when concretely the atom is a mixture of electrons, neutrons and protons. The electron cloud consists of thousands of actual electrons. It must be noted that an electron is approximately 1,830 times smaller than the mass of a proton. The scientific community has ignored this little fact. As an example it will take around 1,830 electrons to equal the mass of one proton in the hydrogen atom. So, it is said that there is one proton and one electron in the hydrogen atom. What really should be said is that there is one proton and approximately 1,830 electrons. However, it is much easier to state that there is one electron in the hydrogen atom for simplicity. The mathematical electron shell theory need not be modified to fit this reality. It is convenient to use the square root of the actual number of electrons that surround each nucleus, for day to day chemical formulations.

When electrons or negative ions rush towards a positively charged atom in order to reach equilibrium, it is said that an electric current is generated. This current is created by a particular energy state called "voltage." If a metallic conductor has electrical current flowing in one direction one end of the conductor has an excess of electrons. The opposite end has a deficit of electrons. All "electric oscillations" are a result of this phenomenon. Battery power, direct current generators, alternating current alternators, radio, microwaves, radar, electronic oscillators, metallic coils and condensers, electron-tubes, ion-tubes and transistors, magnetrons, klystrons, lasers, masers and molecular transitions; all carry these oscillations.

So then, what is electricity? This question has eluded the most able intellects throughout human history. Could this be because science ignored the fact that the electron in an atom is actually representing more than one thousand electrons, not just one? As observation shows, two states of energy exist, one "negative" and the other "positive." How could it be that there are two states, identical in their properties, opposite in character, both clinging to matter, both attracting and yet completely neutralizing each other? From this analogy, it is safe to say that electron differences create the effects of electricity.

Chemical Cohesion

Atoms are attracted by virtue of their differing kinetic phase relationships. Matter stays together because of this principle. Atoms are continually expanding and contracting in unison with their kinetic frequencies. Expanding particles are attracted to contracting ones and visa-versa. "Gravitation" might turn out to be a grand display of this phenomenon. Atoms will repel atoms only if they are in the state of coherent kinetic vibration. This occurs when atoms are vibrating in phase with each other. Coherence will raise energy levels. Differing elements will form "compounds" when their fundamental frequencies are in accord with each other. They must ring harmoniously.

Ultra Chemistry

The first law of thermodynamics describes the principle of the conservation of energy. It states that "energy is not created or destroyed; it merely changes form." The fact is that the creation or destruction of energy is a result of matter being broken down or built up. They both go hand in hand.

I strongly believe that the three primary particles were created at the beginning of the Universe and that they can not be destroyed. However, the good news is that to serve our needs we can artificially alter the elements, (cosmic soups) of these three particles, protons, neutrons and electrons).

I will dare to further state that neutrons not only reside in the "nucleus," but they also exist throughout the atom as a homogenous mix with the other primary particles. The most simplistic expression of the Perreault Atom involves the bold statement that neutrons reside in balance with the protons and electrons in the atoms. This concept applies throughout the entire Periodic Chart of the Elements.

Isotopes

Three hydrogen elements (isotopes) are known to exist that are chemically similar but have different weights. The three weigh differently due to neutron content. These are all hydrogen atoms. However, it is the neutron count in the atomic shell that determines the isotope.

It is clearly established that triple weight hydrogen (tritium) decays to helium–3. In this particular transformation a charged normal weight hydrogen element (protium) carries an excess electron in its shell. It can react with tritium and emit an electron as a by-product of the reaction. At this moment the electron is emitted at an accelerated rate and is seen as beta energy. This model will not alter the way we view chemistry, however, it should change how nuclear changes are viewed.

Gravitation

Each individual atom is at the center of it's own universe and has a unique resonant frequency that is representative of it's specific mass. Lithium for an example has a greater mass than the hydrogen atom and will vibrate at a lower frequency in comparison. Therefore, we may say that it is caught by earth's gravitational pull more readily and has more weight. This may help to explain how physical objects are attracted to the Earth.

Stored Nuclear Energy

Have you ever wondered about spontaneous radioactive decay? Is there a super charged state of matter? If an atom becomes ionized by either gaining or by losing electrons then why can there not be a nuclear ionization too? I hypothesize that atoms do become ionized on the nuclear level by gaining or losing neutrons. This may appear incidental but could be at the very heart of an ultra-chemistry. I say that it is highly likely that this very phenomenon is behind natural spontaneous radioactive decay and explains many unanswered questions in nuclear science.

Atoms can gain or expel neutron particles, some to a greater or lesser degree. Thorium can readily be transformed into a fissile isotope of uranium by absorbing neutrons. It is a known fact that thorium stores more energy than uranium, coal, oils and all other fuels combined. This represents a tremendous amount of stored energy. Transformed thorium could be our most economic source of energy. It is almost as abundant as the element lead.

It is only by passing from one state of equilibrium to another that matter can lose or gain energy, consequently emitting radiation. The notion that radioactivity originated in a peculiar chemical process was adopted and defended by Rutherford. "Radioactivity," says he, "is due to a succession of chemical changes." Research has found that this "succession" is due to neutron transformation.

Nuclear Transformation

Elements that absorb neutrons and become unstable are called "radioisotopes." When an atom absorbs neutrons it is seen as a "fusion" reaction. When an atom loses neutrons, it is seen as a "fission" reaction. All elements can "transform" into a radioisotope and visa-versa. Radioactive isotopes are unstable and will transform into stable elements. A radioactive isotope will eventually revert back to the original element on it's own accord. Radioisotopes are nature's storehouses of energy. We need not look any farther for our energy needs. Nature holds an inexhaustible supply.

Induced Radioactivity

When bismuth is exposed to radium, initially it becomes several hundred times more active than the radium. This is due to emanation that clings to the surface of the bismuth. Emanation is widely known as radon gas that will eventually decay to polonium. It releases more energy than any single natural element ever discovered.

If beryllium is exposed to radium, neutrons are generated. These neutrons will have penetrating power several times that of the most energetic gamma energies. This activity is due, not to an alteration of the inactive matter itself, but to an admixture of a very small quantity of intensely active matter. Is this the neutron? Could it be that a neutron is actually an inert monatomic gas? This would certainly explain a few gaps in accepted atomic theory.

I have found that "radioactive" matter can be electronically synthesized. I find that radioactive matter is simply in an atomically charged state. It has been recognized that when an atom gains or loses electrons, it becomes "charged." However, I have taken this one step further by suggesting that matter will also become super charged when it gains or loses neutrons. This super charge that the atom acquires depends largely on its energy state and the type of radioactive gas that it reacts with at the moment of contact.

Strange Nuclear Facts

Primary neutron particles and gamma energy levels are generated in nuclear transformations. This results in the atom becoming super charged.

A "chain-reaction" in radioactive matter is due to the concentration of excess neutrons. The greater the concentration of excess neutrons, the more accelerated the decay process becomes.

For example, if beryllium metal is added to radium metal, neutrons are expelled. These neutrons can be used to cause secondary nuclear reactions in fissile material. The beryllium becomes the nuclear trigger.

The energy released by the transformation of beryllium is estimated to be at around 50 MeV and that of boron nearly as large. The excess neutrons generated can be used to super charge atoms thus transforming them into radioisotopes.

Cold Fusion

Here is a "cold fusion" reaction to consider. Radon gas (a pure alpha emitter) generated from the decay of radium is deposited onto a beryllium wire. Beryllium will react with an excited alpha particle that will generate a "cold fusion" reaction without the need for excessive heat, as is the case for "hot fusion." The reaction that results is: the ${}_4Be^9 + {}_2He^4 \rightarrow {}_6C^{12} + {}_0n^1$ (cold fusion) mentioned above. This secondary neutron energy now can effect a reaction in nuclear fuels. Atoms need not be fragmented in this reaction. Therefore, this reaction is "clean." There are other examples but this formula will give you a general idea.

Alpha fusion is actually cold fusion. The end result generates neutrons from the production of fusion by alpha particle collision. A simple equation where mass squared by it's speed equates to temperature. This equation can be applied to a kinetically energetic alpha particle in my model. Simply stated, a fusion reaction occurs when there is a temperature differential between the alpha particle and the target that it impinges upon. It must also be noted that no new neutrons will result because the reaction consumes what it releases. This is cold fusion.

The over all reaction is 'cold' in the general sense; in contrast to reactions that produce excess residual heat. To site a cross correlation; a thermal heat pump will absorb hot particles right out of the cold air of winter. The average temperature is cold. This analogy can be applied to the alpha fusion reaction just sited. The particles involved are in the midst of relatively cool ones. The average temperature is cold. This is my justification for dubbing it a "cold fusion" reaction. You may still consider the reaction to be hot. In fact, it is in every real sense of the word. Even so, the term 'cold fusion' is a non-mathematical one. We can use it to denote the difference between reactions that result from a CERN particle reactor and those that are naturally directed by ultra chemical reactions. As you can see, I do not buy into what the textbooks say about what constitutes the nuclear verses chemical reaction. All reactions are chemical in my point of view. The sited alpha reaction is an ultra extension of chemistry. The textbooks have over complicated the "nuclear reaction" in my assessment and in doing so have fogged the truth in many instances.

So-called "cold fusion" researchers have overlooked the fact that the electrodes that they are using contain traces of naturally occurring radioisotopes. Alpha particle emitters such as Pt190 & Pt192 are two examples. What is an alpha particle anyway? It is a super charged helium atom is it not? Couldn’t you also say that it is the atomic ash of nuclear decay? That is exactly what researchers are finding in their experimental cold fusion setups. Neutrons may also be generated as a result of a cold fusion reaction but we might not always detect them. If the neutrons generated equal the neutrons absorbed in the reaction there is nothing left to detect. In my opinion this is the signpost of a clean nuclear reaction.

Some Ignored Uranium Facts

1. U235 emits ultraviolet light. Galen Winsor[1]

2. U238 expels electrons most readily when exposed to a wavelength of 3000 Angstroms in the ultraviolet spectrum. H. C. Renschler[2]

3. The UV light wavelength of 254 nanometers (nm) is a catalyst to oxidation processes.

4. The 254 nm frequency of light kills viruses, molds and other pathogens by rendering their DNA inactive. It therefore disinfects drinking water.

5. The human skin is reddened (erythema) and is most sensitive to the wavelength of 3000 Angstroms (254 nm range). Hauser and Vahle[3]

SECTION SUMMARY

Bombardment by alpha or beta particles does not "disintegrate" atoms. Rather, this bombardment goes beyond normal chemical changes. Energy is transformed by neutron absorption or by neutron emission.

What other discoveries await us? The future is indeed a glorious jewel to behold. Will we continue to follow the path that leads us further away from this great storehouse of energy?

Alpha Fusion

The Department of Energy (DOE) has hindered the development of this technology by holding fast to its' bogus restrictions on radioactive materials. I allege that they fear a miniature, clean atomic power source that will deny the Department of Defense its supply of weapons-grade plutonium.

The DOE’s 45-year old H-bomb fusion program sucks up all government fusion funding. The DOE’s fusion projects are located at the Princeton University Plasma Physics Lab. These projects are top-secret, cold-war research projects using fusion for massive production of plutonium and tritium for nuclear weapons.

The Princeton fusion reactor burns tritium, a radioactive fuel that produces deadly **neutron** radiation, 80% of its total output of energy. This reactor, if used in conjunction with a breeder reactor, is capable of "breeding" 100 times more plutonium than can be produced by a conventional utility reactor. This is a fuel machine for nuclear weapons of destruction.

Congressman George Brown, Chairman of the Science, Space and Technology Committee, had introduced a new bill, in early May of 1994. This bill would have allowed for clean, compact, "neutron-less" fusion energy technologies that would compete with the DOE’s weapons program. This bill was never passed!

In spite of resistance from the DOE against clean atomic reactions, the "Fusion Energy Research and Development Act" (Brown's bill) would have created an Office for Alternative Fusion Energy and a separate budget for Alternative Fusion Technologies.

A neutron free, atomic device could generate electricity at minimal cost. It would compete with any petroleum-based fuels. This would mean no more oil spills, among many other benefits that I have already outlined.

Dr. Bogdan Maglich was the driving force behind the neutron free technology and the alternative fusion bill. He discovered the **omega meson** and other particles. In 1961, his omega meson discovery won him and three associates, a citation from President John F. Kennedy.

Dr. Maglich invented the "Migma," a miniature reactor that demonstrated the principle of a non-radioactive nuclear reaction in a 6-cubic-inch volume. While most fusion reactions are radioactive, Dr. Maglich’s reactor utilizes a design that is virtually non-radioactive. In 1984, Dr. Maglich planned to demonstrate the magmas' energy capabilities. The tests were to be done in cooperation with nuclear labs in Russia, under the State Department's "Conversion of Peaceful Nuclear Research" program for Russian scientists.

Dr. Maglich had flown supporters with him to Washington to persuade the Congress to give the project an affirmative vote. The DOE and the nuclear-industry lobbyist defeated the vote. Humanity was on the brink of an energy technology that would have been compact, safe and almost zero residues except for the electricity that it produced.

A space shuttle full of moon dust would contain enough helium–3 to meet all the energy needs of the U.S. for an entire year. Dr. Maglich pointed out that nine grams of helium–3 are the equivalent of 1,000 barrels of crude oil. A Migma reactor could breed helium–3 from seawater. You do not need to go to the moon.

The Bechtel Corporation in 1989 subcontracted with Dr. Maglich to produce an engineering design of a clean nuclear reactor slated for a space power station. Dr. Maglich at the time was about ready to publish his completed design. The Defense Department (DOD) at this point stepped in and classified the plans before he could utter one word. DOD called his reactor "export controlled," a classification forbidding any written or verbal presentation of this invention to anyone except for government officials. This clearly violated an international agreement that forbids withholding information on controlled nuclear reactions. The agency justified this classification by saying that this type of research contains "ground-breaking work."

The bottom line is that **neutron free** atomic reactions do not generate weapons fuel. The DOE would stand to lose its funding for nuclear weapons fuel if non-radioactive reactors were to be commercialized. These groundbreaking designs would threaten all nuclear power plants.

In the final analysis, it is relatively simple. The alpha particle proves to be an effective energy source in inducing fusion when combined with the right material. We can safely come to the conclusion that through the proper conditions that the alpha particle can provide incredible amounts of power for our use.

Energy Released by Alpha Fusion vs. Uranium Fission

"Alpha fusion" is an amazing process. Here is a comparison of the energy released per atom of uranium that undergoes fission verses alpha fusion via neutron absorption.

U235 fission=161,000 kWh/kg (2x10^8 BTU/lb.)

vs.

Alpha fusion=1.68x10^{11} kWh/kg (2.6x10^{14} BTU/lb.)

The potential implication of this alpha fusion formula staggers the imagination. There is room for unlimited innovation and creativity. Imagine having a device that could safely provide power for several decades without recharging. Sounds almost like science fiction you say? It is a scientific fact! Plutonium–238 is a non-fissile isotope that decays by **alpha particle** emission with essentially zero associated gamma emissions. This type of plutonium spontaneously produces about one kilowatt of energy for two kilograms of mass. This reaction will release alpha particles that can only travel a short distance before they are stopped. This process decreases slowly over time. After ten years the energy production is about 92 percent of the initial value. Even after 87 years the material produces half as much energy as it did when was first started.

When positively charged helium atoms (alpha particles) collide with the **electron cloud** that envelops another atom the positive charge is eventually canceled by these collisions. This is called "absorption." Atoms ring like a bell when impinged upon by alpha particles. This ringing state generates **x-rays**. The wavelength of this generated x-ray depends on the charge of the alpha particle and what type of atom it hits. If the x-ray generated has an energy level that exceeds 10MeV, atoms absorb it. In this case, the neutrons are raised to an excited state. They may be broken loose from the shells of an element or a neutron may be added to a shell. Gaining or losing neutrons via this process may create isotopes. We can efficiently do this by exposing atoms to alpha particles. This model clearly demonstrates that energy is transformed from one state to another. It is neither created nor is it destroyed.

Generally, the formation of a chemical compound is accompanied by the release of a definite amount of heat per unit mass of the substance formed. The burning of a fuel is a chemical reaction in which oxygen is consumed. The quantity of heat released when a unit of fuel is burned with oxygen is called the "heat of combustion." For the average coal, this value is from 7,000 to 8,000 calories per gram or 12,000 to 14,000 BTUs per pound.

In contrast, **radioisotopic** transformation is accompanied by the release of enormous energy when compared to conventional oxygen – electron oxidation. However, this does not justify the theory advanced that neutrons occupy the center of atoms. It is my hypothesis that neutrons reside in the orbital shells of elements and not within the nucleus. To envision this atomic structure one only has to look at our own solar system to see this on a grand scale. Isotopic oxidation (neutron oxidation) simply involves the neutrons. However, the energy involved is much greater because the neutrons represent neutral charge and therefore the binding energy of the elemental shells is of a greater magnitude.

When the unstable isotope of an element reverts back to its original stable state energy is released in the transformation. Now, consider that incident electrons are impinging upon, for example, the naturally occurring element U235. Assuming its absorption wavelength is correct, according to the resonant frequency of this element, then excess energy can be released. This may be done with any isotope. However, U235 will "self-fission" and under the right conditions, will not require any source of energy other than what it emits. Excess energy is released only after enough neutrons are captured to cause the transformation of the U235 atom. This is called the "break even" point, beyond which excess energy is released to effect a "chain-reaction."

We are not limited to the U235 isotope to use as a fuel source. This element is used in a commercial nuclear reactor because it undergoes self-fission and requires no additional catalyst in order to "burn." The public does not generally know that the U238 in a commercial reactor does not undergo fission. The only reason for its presence is to capture neutrons that transform it into weapons grade plutonium (Pu^{239}).

U238 will "fission" via fast neutrons generated by the alpha fusion process. However, there is one small hitch to the alpha fusion process. A nuclear fusion fuel material is needed along with the U238 alpha emitter. This material generates the required fast neutrons for this type of fission. All that is needed is the right type of spark to speed up the decay rate. Radium or radiothorium could also be used as the alpha source. Beryllium, boron, or carbon can be used as the fusion catalyst. Any of the elements below periodic number fifteen feasibly could be utilized. However, these three elements sited pack the most wallop. My process uses alpha fusion formulas that show the way to a simple and economical method. No excess neutrons result to create a toxic waste problem. Fully depleted U238 could be used that does not contain U235. Instead of making nuclear bullets, like the ones used in the Gulf War, we could be using our reserves of depleted uranium as a beneficial fuel source.

Plutonium is not feasible at this time, so plain old, natural, untreated uranium will also work. One pound of "yellow-cake" (U_3O_8) is equivalent to 31 barrels of oil or 10 tons of coal. At the moment yellow cake costs only $12.50 per pound.

Radon gas from naturally occurring radio emissive uranium or thorium ores can be directly converted into electricity. The case surrounding these materials will provide adequate shielding. These devices can be designed to be compact so that a lightweight unit could be built. It has been estimated that for every pound of fuel, that, one kilowatt of power can be generated.

For political reasons' people, organizations and government agencies have strongly opposed the use of all forms of nuclear energy. These people have made the general population paranoid of any and all radioactive materials. Despite this strong opposition and threats against it's development, work continues with the dream of harnessing transformations safely. The atom makes energy available to us. I have proposed a radical atomic theory to explain the nuclear energy process. Conventional atomic theory has many flaws. It does not completely explain the energy created by a nuclear reaction. Does this mean that the nuclear reactor does not completely create energy? No matter what theory you wish to use in explaining atomic power, the fact still remains that energy is transformed.

James Chadwick discovered the neutron in February 1932. Later, further studies were made to discover what effects would be produced by bombarding different materials with this new particle. Researchers for the next several years bombarded every known element with neutrons and recorded hundreds of new radioactive isotopes. My own research in this area has revealed many secrets of the vast reservoirs of energy contained within the atom.

The First Neutron Initiator

The initiator for the first plutonium weapon was composed of polonium coated onto beryllium. The polonium came from the Radium Chemical Company.

Electronically Activated Neutron Source

In the winter of 1996, I applied a 90,000-volt negative charge to a carbon rod. This carbon rod was surrounded with radioactive ore in a closed container. Under these conditions, the radon gas emitted from the ore was attracted and absorbed into the charged rod. This began my research into electronically assisted atomic transformations.

"It must be first remarked that helium is a gas which accompanies all radioactive minerals. It was from these bodies that it was first obtained – a catalyst with an interesting effect."

"This derivation from radium is a **special helium** since it appears to possess the property of spontaneously vanishing. Its' only resemblance to ordinary helium seems to consist of the momentary presence of some spectral rays." * The Sea of Energy, 5th edition, pages 229-230.

Breeding U234

Mix a dilute solution of uranium nitrate and finely divided carbon together. The solution is boiled for about half an hour. Radon is united with the carbon. After which the carbon is filtered out. To unite the radon with the carbon, the two are exposed to a strong source of **gamma radiation.** This is the **catalyst** that accelerates nuclear (ultra-chemical) reactions.

The reaction is as follows: $C^{12} + Rn^{222} \rightarrow U^{234}$ (uranium "X") + $\mathbf{e}\uparrow$

To concentrate the U234 the excess carbon can be removed by burning.

The results of my life-long research have opened a doorway that could lead to electronically induced nuclear fusion. I originally named this process "alpha fission" to connote the electronically induced atomic reaction. The term was chosen so as to cause the least adverse social reaction to this discovery, due to the fact that the public has been duped into a state of paranoia by its global governments. Later, this term was changed to alpha fusion, as this is a more accurate term than the original chosen term.

Conventional nuclear reactors generate deadly toxic radioactive wastes as by-products. These reactors are even more deadly than anyone realizes. Fuel rods in the reaction process are extremely hot and are subjected to enormous internal pressure. You can not tell me that untold amounts of deadly radon gas are not leaking into our environment. The fact of the matter is that a special detector would be required to detect this leakage. Scintillation, Geiger and Alpha counters will not detect this gas because the mass of the radon gas can not penetrate the sensors on these meters. We must take a close look at clean nuclear processes before it is too late (if it isn't already).

With the advent of electronically triggered neutrons, a self-sustaining chain reaction can be obtained and may be highly controlled. I find this especially true if **thorium ore** is used. Thorium emits **thoron** (Rn^{220}). Owing to its extremely short half-life, thoron by itself can not be used as a source of strong radiation. However, in conjunction with its parent substance radiothorium, it can indeed be used as a strong source of radioactivity.

Thoron has a short half-life of 54.5 seconds. On the other hand Radon–222 has a considerably long half-life of 3.5 days making it a hazard. Thoron is safe and convenient to work with because there are no long-lived products in its disintegration chain. Therefore, there is minimal danger of permanent contamination. The parent substance, radiothorium, may be obtained from thorium solutions in the manner described next.

Radiothorium is generally used as a parent source of thoron. Radiothorium has a conveniently long half-life of nearly two years. Extractions of radiothorium can be obtained in a highly emanating state. A small quantity of iron chloride is added to a thorium solution and precipitated as hydroxide by ammonia or as oxalate by oxalic acid. The precipitate, the hydroxide or oxalate of iron, is washed very carefully and dried at room temperature. This precipitate should have a high activity rate. It will not lose more than 20% of its initial activity rate for many months. This radiothorium preparation may be spread into a thin layer in a shallow brass tray placed at the bottom of a small lead-shielded container. The active deposit thorium can be collected on a negatively charged wire, insulated with respect to the container. The emanated **thoron** gas may also be absorbed into various materials.

Many radioactive isotopes could be used to generate electrical power. However, the power behind these isotopes is really with the radon gas. I have pointed this out many times in my writings, lectures and articles. Simply by bombarding bismuth, we can create polonium. In United States Patent 2,728,867, it is stated that the available energy of a commercial nuclear fuel rod is around twenty watts of power per gram. In comparison, **polonium** will generate one hundred and forty watts of power per gram. This is **seven times** more energy than is available from the conventional fission process!

The atomic ions emitted from radioactive isotopes can be directly converted to electrical power. My conversion method shown in Section 5, **Figure 4** can be used to convert this source of radiant energy. Additional stages can be added for more power. The circuit shown gives a general idea of how such an energy conversion device can function. By no means is this technology limited to this one circuit or configuration.

Radon Fuel

My research has revealed that the energy devices of T. H. Moray, Alfred Hubbard, Reverend Antonio d'Angelo, Joseph Papp, and others all contained a source of **radon gas.** They may or may have not realized that this was what they were using as their fuel. We will never know this for sure. However, what we do know for certain is that they all used radium in their energy devices. When my prototype is ready to be demonstrated I too will show how radon can be used to generate electricity. Demonstrations will have to be performed in private until my device obtains an exemption from government restrictions.

The facts show beyond doubt that radon gas was the fuel source that has powdered many so-called "free energy" devices. You can say that these devices ran on the energy of the cosmic rays. It is these rays that created radioactive thorium and uranium that we find in the ground. High-energy cosmic rays are really neutrons as Nikola Tesla had discovered. When a cosmic particle reacts with either thorium or uranium a radioisotope is created. This is why you see a consistent percentage of U235 all over the planet. This percentage represents an accumulation of comic ray bombardment over the lifetime of our world. Nature offers us this fuel that can be used to obtain electrical power with no moving parts and it is out there free for the taking. We do not have to split atoms to harness this storehouse of energy. Nature has already safely done this for us.

In order for clean nuclear power generation to enter into public use, high-grade radioactive ores must be secured without restrictions. Recent attempts to obtain these ores have been blocked by the United States Government. Once these ores are obtained I will build a high wattage radiant energy converter. Through my writings, lectures and published articles I have attempted to explain my research. Over the years I have made my research available to the public and have asked for little in return. I am asking you at this to help me to get this technology implemented, within the constraints of the law. As a former United States President once stated: "Sometimes when we reach for the stars, we fall short but we must press on."

President Ronald Reagan

SECTION 10

The Truth about Nuclear Science

Present day alternative energy researchers find hope in that, one day, power will be obtained from the energy that runs the Cosmos. In reality, this dream has only proved to be a lure to its followers, beyond the limits of credibility. A quagmire of unsubstantiated hypotheses exists that are bottomless and lack reality. The facts of many newly conceived energy concepts are wholly inadequate, lacking justification, and incapable of throwing any light on the promise of inexpensive, abundant sources of energy.

The Earth is indeed bathed in a sea of cosmic and solar energies. Can a device be built that can transform these energies into usable power? I believe that T. H. Moray did build such a device. This has inspired me to conduct my own research and to build my own prototypes. How the radiant energy device really worked might shock a few people once the facts become realized.

Nature is my best teacher. Within one of her classrooms there are mineral samples that give out pure electrical energy. They are the radioactive rocks. If radioactive matter from these rocks derives its energy elsewhere, this must be of a type entirely different from any presently known. It is capable of penetrating, without loss, hundreds of feet of solid rock.

Nikola Tesla’s investigations brought him to the conclusion that the Earth was being showered with "tiny particles, each carrying so small a charge that we are justified in calling them neutrons." He stated that "…they move with great velocity, exceeding that of light."

Dr. Thomas Moray called into existence the concept of unlimited radiant energy (the cosmic ray), permeating all space and all matter, unimpeded. Such a view, at first glimpse, merges perfectly with the continuous and permanent activity of radium extracted from certain rocks. This activity seems to last for an indefinite time. There appears to be no rational reason why radioactivity, however intense and powerful, should decay or diminish with the passage of time. Yet, radiant matter (radioactivity) does decay. The demonstrated "half-life" values have proven that the stored energy within radiant matter will eventually be depleted.

At some point in my own research, I had to ask myself three important questions:

1) Is there anything opposed, either to reason or to probability, to the view that the energy evolved from radium is actually derived from an existing previously unsuspected internal process within the atom?

2) Within the decay process does the element experience a transformation into other elements?

3) How is it that such enormous stores of energy in matter have remained so long unknown?

In my experience, a specimen of radium chloride is dissolved in water and the liquid is then evaporated. Bringing this compound back to its dry state, as the result of this very simple operation, the radium has lost the greater part of its radioactivity. The penetrating beta and gamma rays completely disappear. The remaining non-penetrating alpha rays will be only one quarter as powerful as they were initially. Then a strange thing happens. Left alone, the radium will spontaneously recover its lost activity. Little by little, day by day, until at the end of a month, it will be no less active than it was before it went into solution. This appears to be in direct conflict with the zealous statement that "radioactivity of radium cannot be affected by any known process." On studying this experiment carefully, we will find that when radium chloride is dissolved in water something escapes into the air. This "something" is intensely radioactive. It diffuses in the air, but will remain if contained within a gas-tight, closed vessel. In short, this "something" is a gas possessing the property of radioactivity to a very intense degree. This gas was called "emanation" in days of old. Today it is called **RADON-222**. It has a half-life of around 3.8 days. This radon gas is highly radioactive. Thorium releases **RADON-220** that is intensely radioactive and has a half-life of around 55.6 seconds. Uranium emits **RADON-219**, the most intensely radioactive substance, having a half-life of only 3.96 seconds.

Radon that has been separated from its parent source will rapidly decay. It is easy enough to conclude from this experiment that energy is stored within the radium atom. This accounts for the comparatively rapid decay of the activity of the generated radon gas. This gas is depleting its internal store of energy so rapidly that it is soon exhausted. On the other hand how is the gradual recovery of the radioactivity of the parent radium explained? This question is the golden key to unlocking the secret of its power.

If it is true that energy comes from within, as large as the store of energy within the atom must be to explain its radioactivity, yet it cannot be infinite. Therefore, it is expected that the activity will slowly decay with the passing of time. Two radioactive bodies, one much more radioactive than the other, are compared together. It is expected that the activity of the more powerful body will decay faster than that of the other. A time will come for both of these substances, when the internal stores of energy are exhausted, that the radioactivity will come to an end. Therefore, radium, a substance containing the store of energy, can no longer be radium when the energy is lost. Coal is not coal after it is burnt. When energy is obtained from matter, the matter changes. Before it can be restored to its former state, the energy evolved must be put back. In no case, is it possible for matter to part with its store of chemical energy and remain the same, otherwise perpetual motion could be easily constructed.

We conclude that, if the energy is stored in the radium, it must be within the atom. Therefore, if radium changes, it must be a change of the atom and of the element itself. This change would be its ultra-chemical reaction. This is a more fundamental and deep-seated change than the normal chemistry or any known kind of material change. Until the discovery of radioactivity, such changes certainly had never been observed. If the energy of radium comes from within, then it must be subjected to this ultra-chemistry that has eluded the chemist. This process of ultra-chemistry must be understood as a special extension of chemistry. It is not some vaguely possible, transcendental nuclear condition in the Sun and stars. Once this is understood, the answers will become as sure as the nose on your face.

Imagine that a month has passed. The radium chloride is once again dissolved in water and evaporated down to dryness exactly as before. Once more you will find in the process, that the radium has lost the same large proportion of its radioactivity. Again it gains a new amount of radon gas no less than before it went into solution. Repeat this experiment as often as you like. You will find the result always the same. The radon that you separate from the radium is decaying away from day to day. But the radium will also generate a spontaneous fresh new batch of radon. In order to make sense of this, we have to answer these nagging questions; "from where does the energy of radium come?" Does it come from nowhere, or is it being newly created out of nothing? This is a view that is not acceptable for a plausible twenty-first-century physical science.

If our doctrines of energy are true, then there are only two possible answers to consider. Either the energy must be derived from within the radium, or it must be supplied from elsewhere. This simple narrowing down of all the possible answers may appear to you somewhat simplistic, but in reality, it carries with it far more than appears on the surface. Being a minute property of the element, radioactivity is therefore a property of the atom. If we say that the energy comes from within, this means that there must exist an enormous and not previously suspected store of energy in matter, or at least in radioactive matter.

Dr. Moray often advanced the theory that radium acts merely as a transforming mechanism, that it converts the cosmic rays into electrical particles. He states that it retransmits energy in a form that we recognize. Tesla also entertained this view. There are electrical transformers dotted all over our country receiving transmitted but dangerous high-tension currents from the central power station. These are "transformed" into comparatively safe, low-tension currents for your house. Are the atoms of radium acting as the transformers of a mysterious and unknown source of external energy, first receiving it and then delivering it up again in a form that can be recognized? It may be said that so vague a view, postulating the existence of limitless and mysterious supplies of omnipresent energy, cannot be directly disproved. One thing is for certain; this view provides a way of escape from some of the more distasteful "logical" assertions, namely that all forms of energy are stored within the atoms and molecules.

If radium owes its activity to an energy source outside itself, one has only to isolate it. We have seen that to quench radioactivity, or to modify it in any way is one of the things mainstream science cannot do.

An experiment has proven that even in its natural state in the mine, hundreds of feet deep down in the Earth, pitchblende exhibits its normal radioactivity. Over 100 years ago, scientists concluded that some sort of radiation must exist in the atmosphere. This conclusion arises because under all circumstances, air is found to be have a slight positive concentration of ionized particles. The atoms are minus an electron. This radiation is constantly much stronger at high altitudes than at the Earth’s surface. This leads us to conclude that such radiation does not originate on this planet but rather from somewhere in the outer regions of space. The fact that the intensity of this radiation does not increase during daylight hours eliminates the sun as its source. Also, the intensity does not decrease during the day or increase by night, therefore it cannot be a result of any particular group of stars.

Spontaneous Decay

Does not the "free emanating" state of a salt of radium teach anything about its decay mechanism? Results obtained by Dr. Alois Gaschler and Otto Hahn in the 1920's proved conclusively that when radium becomes finely divided, it most readily emits radon. This teaches us that radioactive change is effected by confinement. How many scientists have paid attention to this all-important lesson? Is it actually the true mechanism behind what is being called "fission?"

It is in my humble educated opinion that the seat of radioactivity is due to radon gas trapped within the lattice of certain heavy metals. The best comparison that I can think of is coal and the stored hydrogen gas within its lattice. Its stored hydrogen reserve reacts with oxygen transforming these two gasses into H_2O with the release of heat. The carbon contained in the coal also reacts with oxygen when a high enough heat is generated from the hydrogen reaction and combustion occurs, transforming the carbon and oxygen into CO_2. Pretty much the same fundamental principle must occur when radon gas is released from its heavy metal prison. The reactant in this later case is a neutron from a previous radon gas emission. Radon reacts with a metal, losing a neutron to it. When this occurs, the radon becomes highly unstable, emitting alpha particles and rapidly transforming into stable lead. From this reaction it is clear that radon is not as inert as we were lead to believe. In fact, I say that radon is at the very foundation of nuclear reactions.

Since the discovery of radium by Madame Curie, the fundamental method of its extraction has not changed. The process utilizes the same methods that were instituted by Madame Curie and M. Debierne. Radioactive ores are first treated with a warm solution of hydrochloric acid that dissolves the metals. The addition of sulfuric acid precipitates the insoluble barite's that are always present, carrying along with them all the radium as barium sulfate. It is necessary that barium is present in the process. If the solution does not contain a sufficient quantity, then some barium has to be added. It is said that the barium acts as a "carrier" to the radium. Can there be any other explanation? It never made much sense to me why the radium would not precipitate out without the barium. I simply could not accept the notion that it acted as the carrier. The answer to this problem becomes relatively simple if you let Nature be your teacher. Forget what the textbooks tell you. We know from her lessons that radiation can be induced into non-radioactive elements, do we not? Could it be then, that something in the separated hydrochloric solution is transforming the barium into what we call radium? Ala-ka-zam, low and behold, there sure is, it is called **URANIUM–235.**

Nu Radium Formula

Radium may be synthetically manufactured by fusing dry uranium nitrate mixed with about 0.5% of dry barium nitrate. This is done in the heat of an electric arc. The mass is then treated with nitric acid, water and sulfuric acid, successively. Radioactive barium sulfate, possessing all the physical properties of the "element" discovered by Madame Curie, is obtained by this formula. This material will emit ultra-violet rays, x-rays, excite phosphor screens and cause air to conduct electricity. The chloride and the carbonate of radium may also be created by using uranium chloride or uranium carbonate respectively and will have the same properties.

Uranium nitrate typically contains about 0.72 per cent **U235**. The barium reacts with the U235 creating pure radium. Take note here that this is an ultra-chemical reaction and that the barium is not a carrier as has been believed. Best results are obtained when 0.72 per cent barium is added, making it equal or less than the amount of U235 contained in the nitrate. If sulfuric acid is added after the reaction takes place, then radium sulfate is precipitated. This material can now be filtered for removal from the solution. The synthetically created radium radioisotope results when barium reacts with **RADON–219**. It can not come into being otherwise. **All other radioisotopes may be called into existence by using U235 as the source of energy transformation.** There are other processes involved but this will be discussed in a future paper. Radioisotopes created by ultra-chemical processes will have far reaching ramifications, if and when it is taken under the wing of mainstream science. Radioisotopes could be made to order literally in a jar. These findings indicate that all radioactive substances originate from the U235 atom. Now the question arises; where did the U235 atom originate? I can only speculate at this point that it is created when a cosmic ray hits a stable U238 atom head-on. If my conclusions turn out to be correct, then the energy stored in radiant matter is a product of an external cosmic force and an internal force without contradiction.

Clean Nuclear Reactors

Mother Nature offers us a clean nuclear process to supply the World with energy. My reasoning is based upon her process that leaves behind a clean, non-toxic ash. You have seen it in children's balloons at birthday parties. The inert gas that I am referring to here is called "helium." When I shine ultraviolet light onto a radioactive ore, the trapped helium gas will fluoresce a lime-green. The only ash, left behind by Nature’s process, is what is seen under the ultraviolet lamp. There are no deadly radio toxins like the ones found in the commercial reactor. It makes all the difference in creation how the reaction is carried out and for what purpose. I will also tell you that eating radium, like the women did who painted the glow-in-the-dark watch and clock dials will create serious health problems. A little common sense goes a long way.

The commercial nuclear reactor releases excessive amounts of neutrons. This is a direct result of the reaction of **RADON–219** with the zircalloy fuel rod casing. It is highly likely that we can design reactors that will release only pure electrical particles in place of the toxic radioisotopes created by stray neutrons. Why not build clean reactors that do not generate excessive neutron release? Maybe it is because this is not presently perceived by mainstream science? Knowing the true facts makes it a relatively simple matter to acquire a cleaner burning reaction. The only major modification in reactor design is to find a more suitable fuel rod casing. Eliminate the excess neutrons and you will have a clean burn.

Are Neutrons What Really Cause Fission?

1. Alpha bombardment of any element with atomic numbers from **1** through **15** will generate a new and stable element that will release one excess neutron from the reaction.
2. Neutron bombardment of any stable element from **16** and above will generate a new and unstable element by capturing neutrons. One or more neutrons are released from the reaction. The targeted element will gain the weight of the absorbed neutrons.
3. Neutron bombardment of an unstable element generates one helium–4 atom from each reaction. The atomic weight of the targeted element looses the weight of one helium–4 atom as a result of the reaction.

From these three facts, it is not too hard to figure out that the "moderator" in a commercial nuclear reactor is actually what initially supplies the required neutrons for a "fission reaction." The excess neutrons created are what react with a "fissile fuel" that creates a secondary reaction. The release of thermal energy is a by-product. Scientists of yesteryear had these facts straight. Why are present day nuclear theories so off course? It will only be when these facts resurface that advances will be made in the way we obtain energy.

CONCLUSION

The devices that I have presented at public tests do not contain radioactive materials for obvious reasons. However, the underlying principles have been adequately demonstrated. I have used star-mode electrodes in the past to slowly discharge the capacitors in a prototype device. Advanced discharge systems are now being utilized. In 1995 I discovered ion valve circuits that are now being used in my most recent prototype. Even though these public tests have not contained radioactive material, my research has been validated.

Funding has not permitted me to build a device that can power a home or electric car. However, my prototypes have proven beyond doubt that with the proper funding this can be accomplished. It is my hope, through the publication of my findings, that funding will be forthcoming. At this point in time, it is most important to me that I receive credit for my for work and that my research discoveries do not fall on deaf ears.

Public tests are performed as necessary to validate the technologies. These demonstrations are deemed as research and development tests only. No power claims are expressed or implied.

For almost fifteen years now I have carried out many experiments. The fruits of this research are ripening. Many fruits are within grasp but they are not yet quite ripe for the picking. The intent of revealing my progress to the World is to insure that this work is not done in vain. Through my writings I have attempted to explain the many facets of "radiant energy" and how it is possible to draw useable power from it. If radiant energy can be given to the World then humankind will have the potential to reach heights not yet dreamed or conceived. Today, I freely give to you the results of my life's work. Together we can reclaim our independence!

REFERENCES

Winsor, "Nuclear Scare Scam" – lecture

"Radiations from Radioactive Substances" by
Sir E. Rutherford, J. Chadwick, and **C. D. Ellis** 1930

Rentschler, A. I. E. E. 1930, p. 49 & p. 576

Hauser and **Vahle, Strahlentherapie,** 1921, p. 13 & p. 41

"The Interpretation of Radium and the Structure of the Atom" **by Frederick Soddy, M.A., F.R.S.** 1920

Annual Report of the Smithsonian Institution (1899) "An Experimental Study of Radio-active Substances" p.162

United States Patent No. 1,644,370 **A. Gaschler** – Oct. 3, 1927

United States Patent No. 1,655,184 **Otto Haun** – Jan 3, 1928

HISTORICAL COMMENTARY

Tesla's last ten years of his life, 1933 - 1943, contrary to popular belief, were his most productive years in his private laboratory located on Canadian soil. This was communicated to me through his last known living assistant, Arthur H. Matthews, E.E., and B.Sc. before he passed away in 1986. He also revealed to me that special primary batteries with replaceable zinc electrodes powered Tesla's electric car. This car also had a "fluid transmission" and could travel at 90 miles per hour. The electric motor could run on the batteries for three hundred miles before you had to pull over and attend to the batteries and then you were on your way for another three hundred miles. Matthews claimed that there were enough spare parts in the trunk to run the car for one year! The performance of this vehicle seems to be impossible. However, the characteristics of Tesla's unidirectional electrical current can generate short duration high-energy that will generate high frequency oscillations when dumped through an inductive coil. This unique principle makes possible many seemingly impossible innovations. It was this principle that was at the very heart of Tesla's Magnifying Transmitter, also known as his "DC transformer" to his assistant Arthur H. Matthews.

June 9, 1925. **1,540,998**

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 1

Fig. 1.

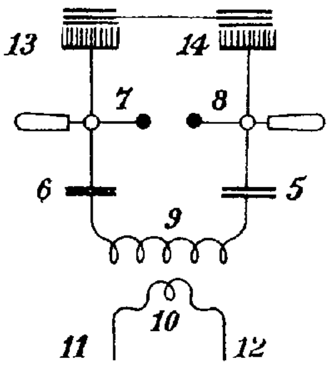
13
14
7
8
6
5
9
10
11
12

Fig. 2.

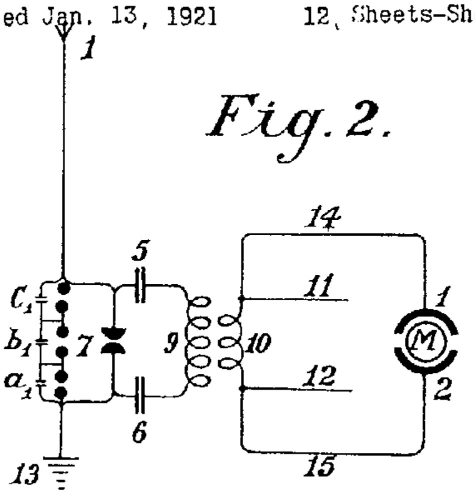
1
C1
b1
a1
7
5
6
9
10
14
11
12
15
1
M
2
13

Fig. 3.

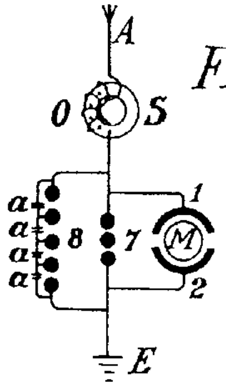
A
0
S
a
a
a
a
8
7
1
M
2
E

Fig. 4.

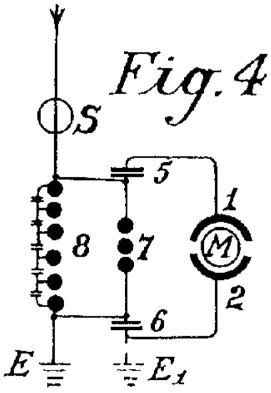
S
5
8
7
1
M
2
6
E
E1

Fig. 5.

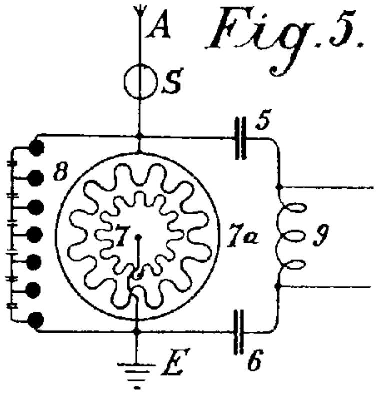
A
S
5
8
7
7a
9
E
6

Fig 6

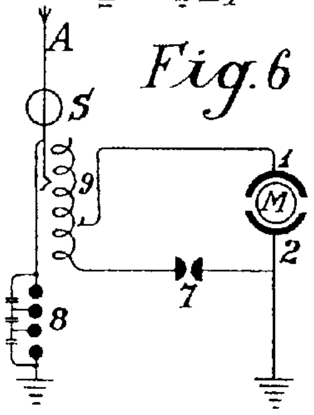

Inventor
Hermann Plauson

by Knight Bros
attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 2

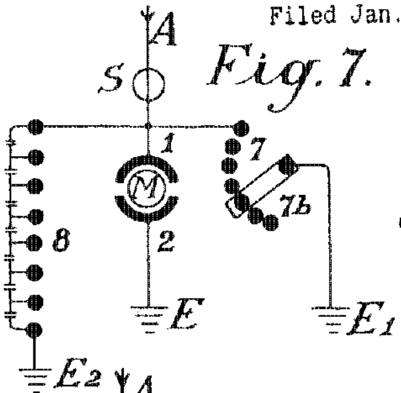
A
S
1
M
2
7
7b
8
E
E1
E2

Fig. 7.

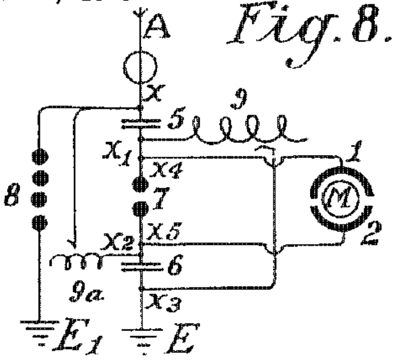
A
X
5
9
X1
X4
7
1
M
2
8
X5
X2
6
9a
X3
E1
E

Fig. 8.

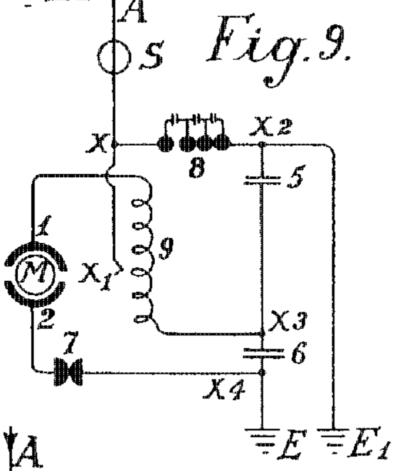
A
S
X
8
X2
5
1
M
X1
9
2
X3
7
6
X4
E
E1

Fig. 9.

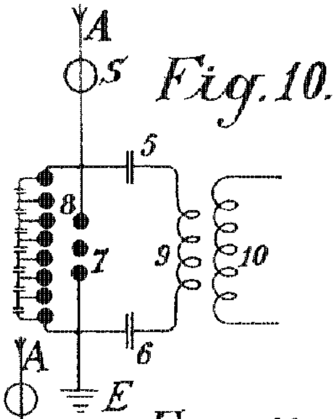
A
S
5
8
7
9
10
6
E

Fig. 10.

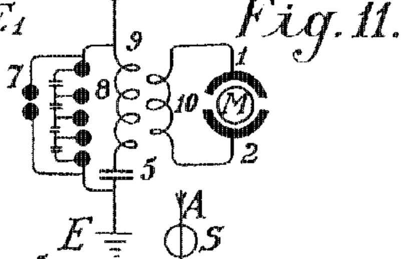
A
S
9
7
8
10
1
M
2
5
E

Fig. 11.

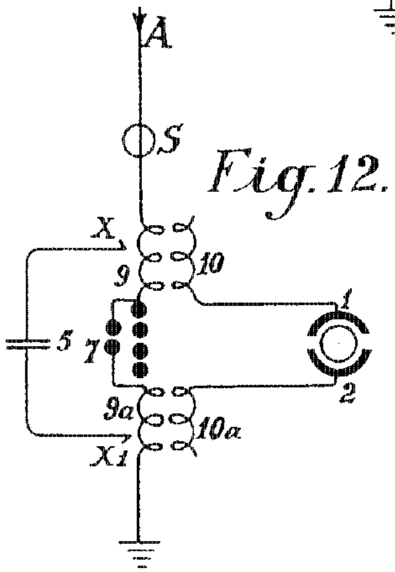
A
S
X
9
10
1
5
7
2
9a
10a
X1

Fig. 12.

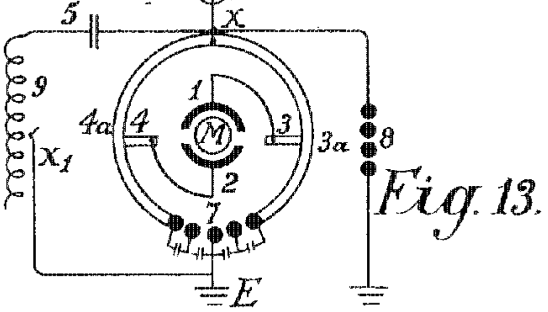
A
S
5
X
9
1
4a
4
M
3
3a
8
X1
2
7
E

Fig. 13.

Inventor

Normann Plauson

by Knight Bros attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 3

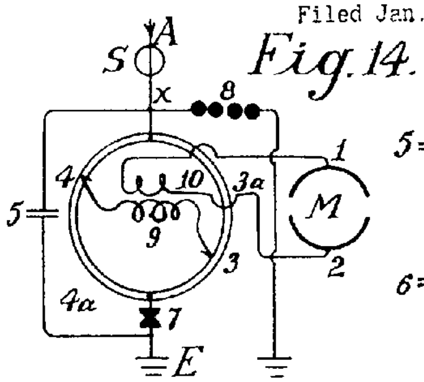
Fig. 14.
A
S
X
8
1
M
2
10
3a
4
5
9
3
4a
7
E

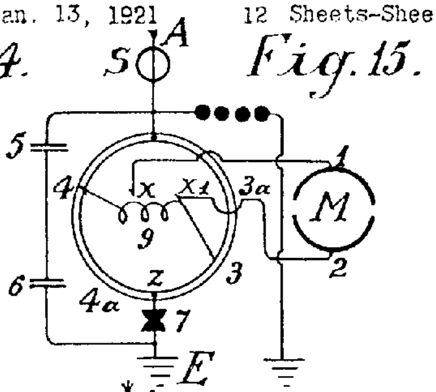
Fig. 15.
A
S
5
4
X
X1
3a
1
M
9
3
2
6
Z
4a
7
E

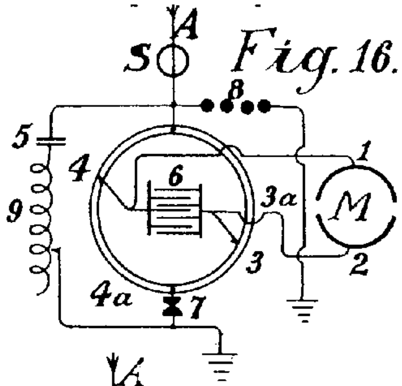
Fig. 16.
A
S
8
5
4
6
1
3a
M
9
3
2
4a
7

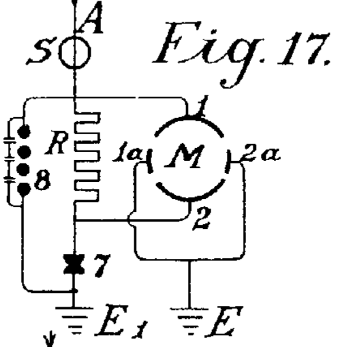
Fig. 17.
A
S
1
R
1a
M
2a
8
2
7
E1
E

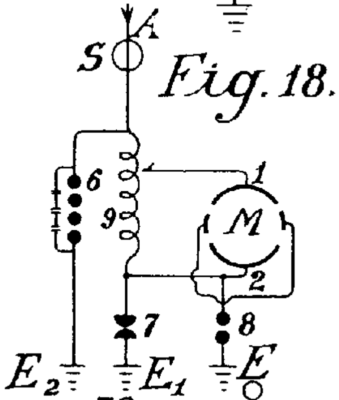
Fig. 18.
A
S
6
9
1
M
2
7
8
E2
E1
E

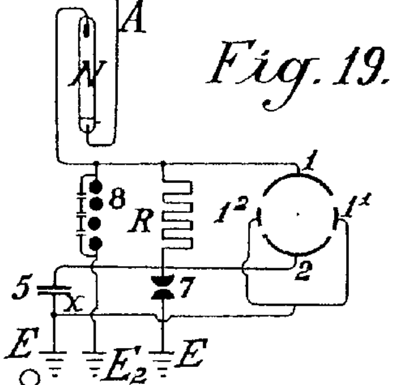
Fig. 19.
A
N
8
R
1
1²
1¹
2
5
X
7
E
E2
E

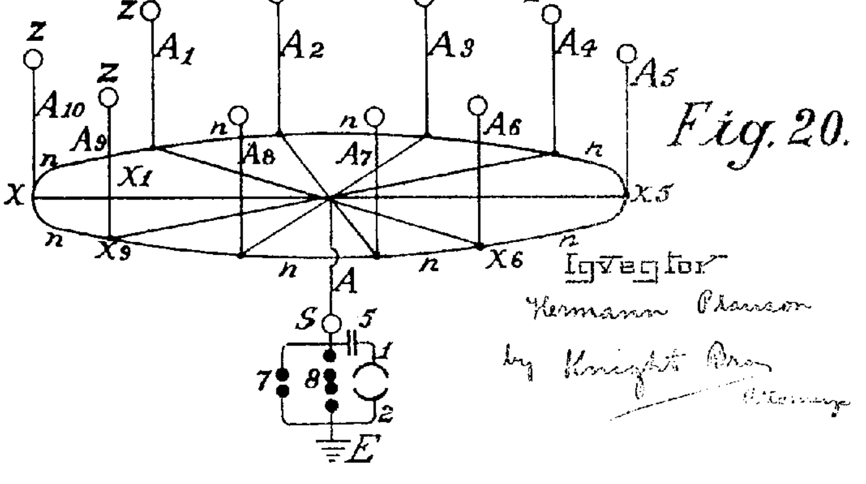
Fig. 20.
Z
A1
A2
A3
A4
A5
Z
A10
Z
A9
n
A8
n
A7
A6
n
X
X1
X5
n
X9
n
n
X6
n
A
S
5
1
7
8
2
E

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. **1,540,998**

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 4

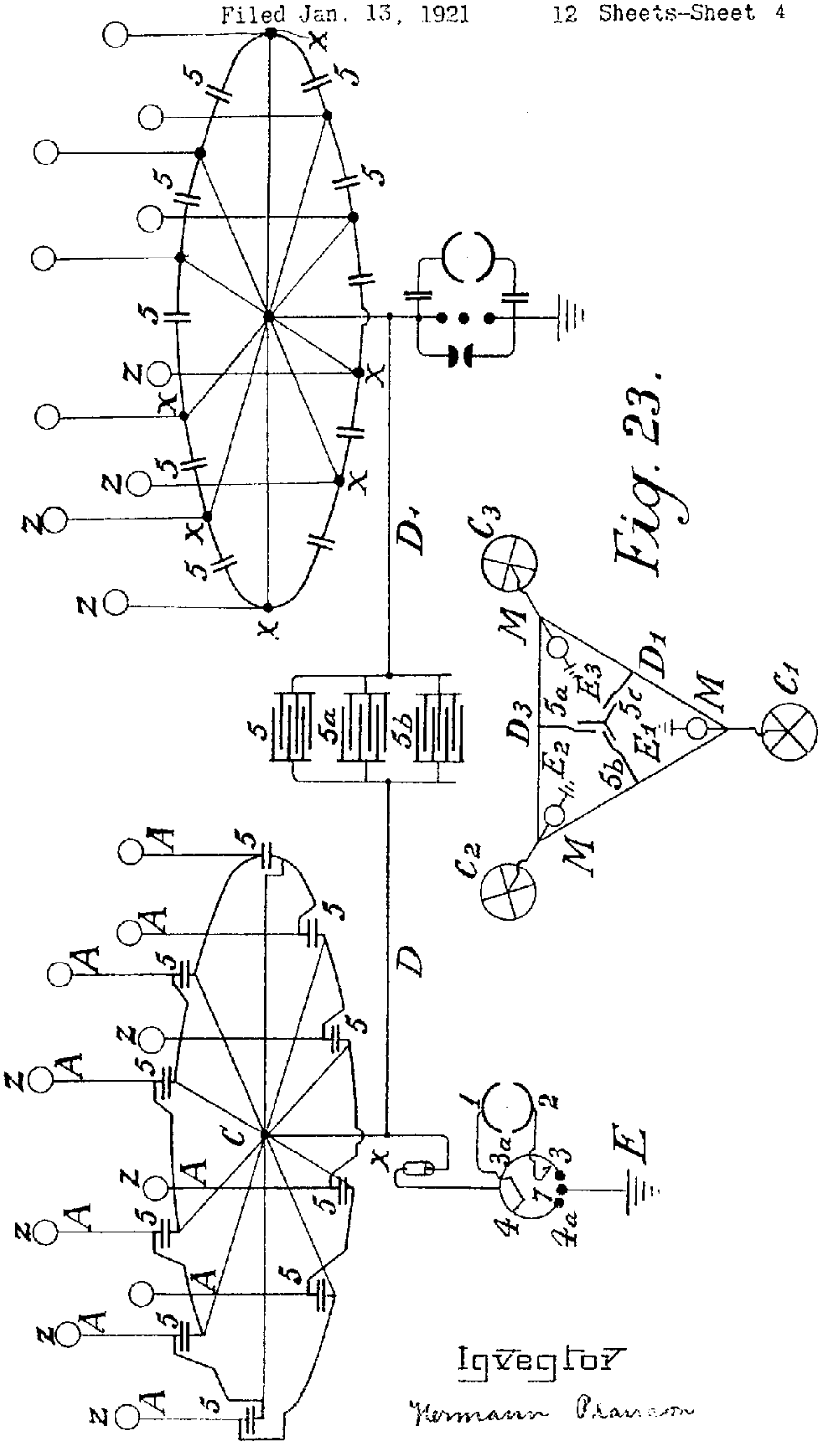

Fig. 22.

Fig. 21.

Fig. 23.

X
5
Z
D'
D
5a
5b
A
C
1
2
3
3a
4
4a
7
E
C1
C2
C3
M
D1
D3
E1
E2
E3
5c

Inventor
Hermann Plauson

by Knight Bros
Attorneys

June 9, 1925.

1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 5

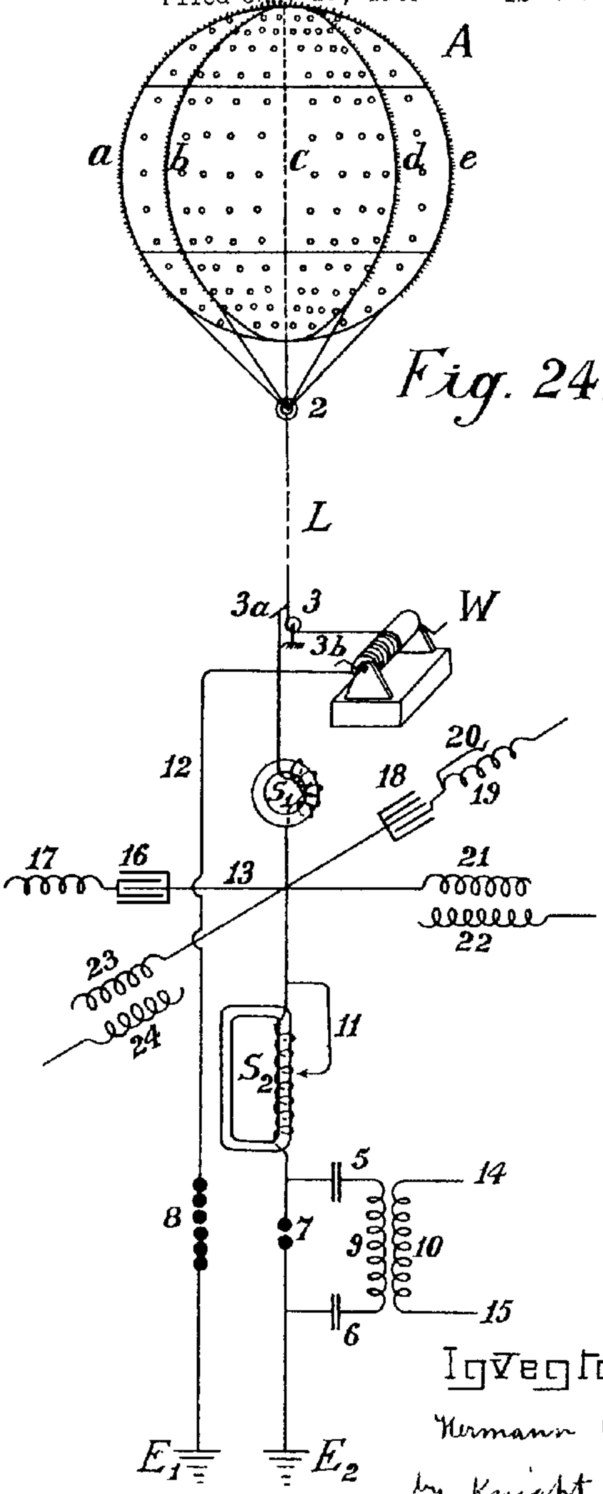
A
a
b
c
d
e
2
L
3a
3
3b
W
12
20
19
18
17
16
13
21
22
23
24
11
S1
S2
5
14
8
7
9
10
6
15
E1
E2

Fig. 24.

Inventor

Hermann Plauson

by Knight Bros

Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 6

Fig. 25.

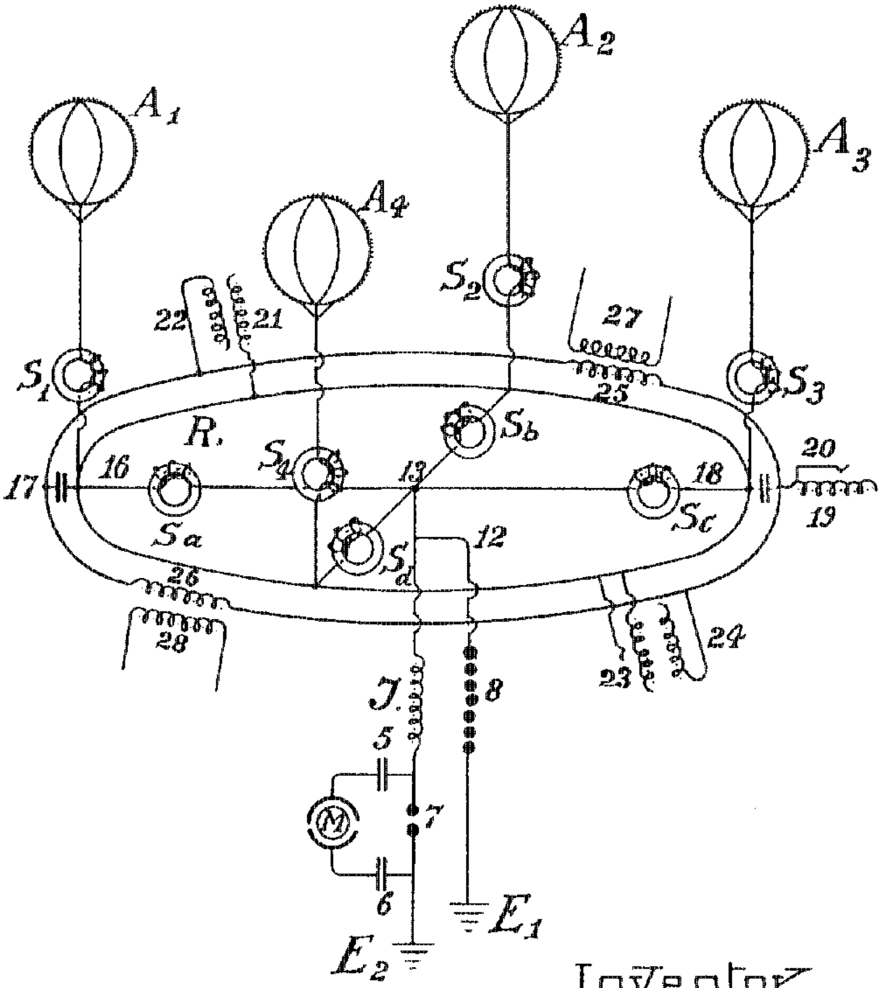

A_1
A_2
A_3
A_4
S_1
S_2
S_3
S_4
S_a
S_b
S_c
S_d
R.
12
13
16
17
18
19
20
21
22
23
24
25
26
27
28
J.
5
6
7
8
M
E_1
E_2

Inventor

Hermann Plauson

by

Knight Bros

attorneys

June 9, 1925. **1,540,998**

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 7

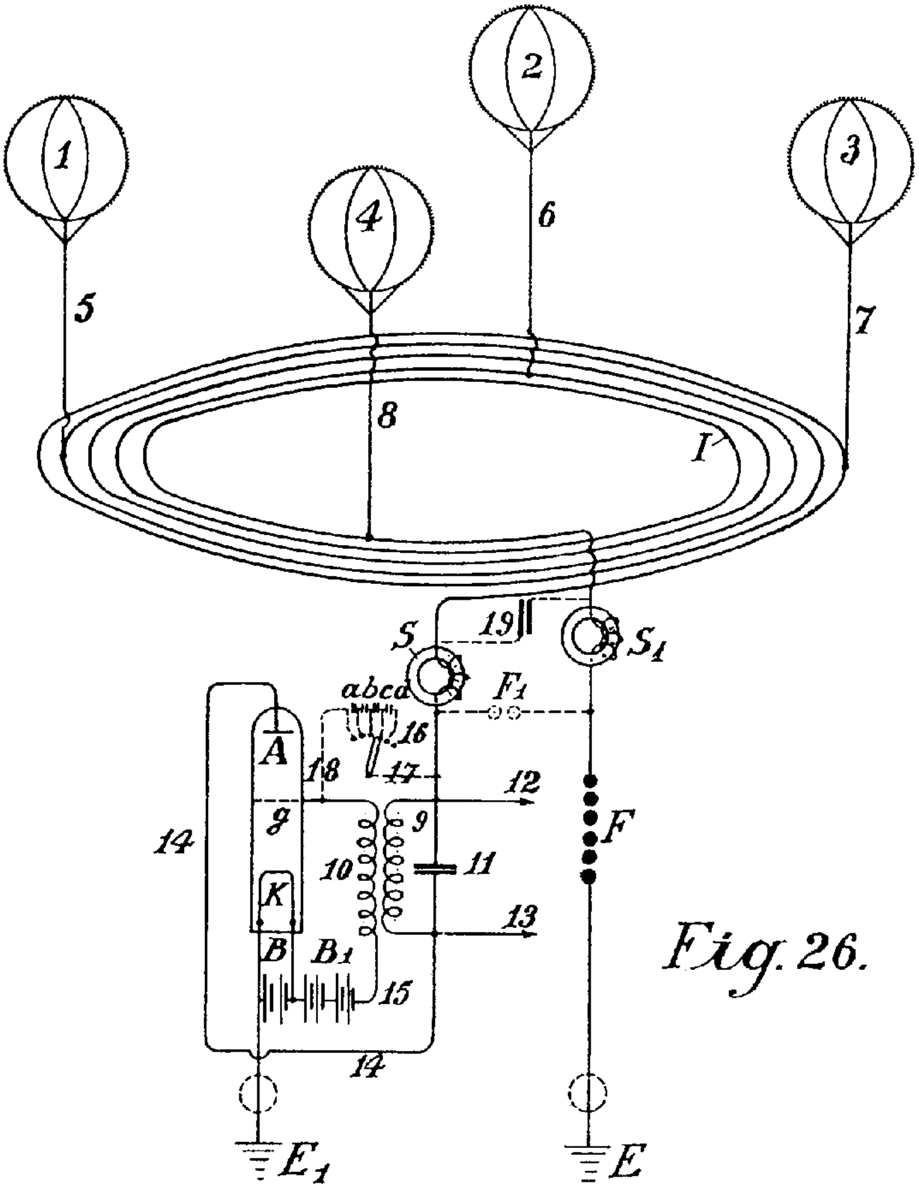
1
2
3
4
5
6
7
8
I
19
S
S$_1$
abcd
F$_1$
16
A
18
17
12
g
9
14
10
11
K
13
B
B$_1$
15
F
14
E$_1$
E

Fig. 26.

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. **1,540,998**

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921. 12 Sheets-Sheet 8

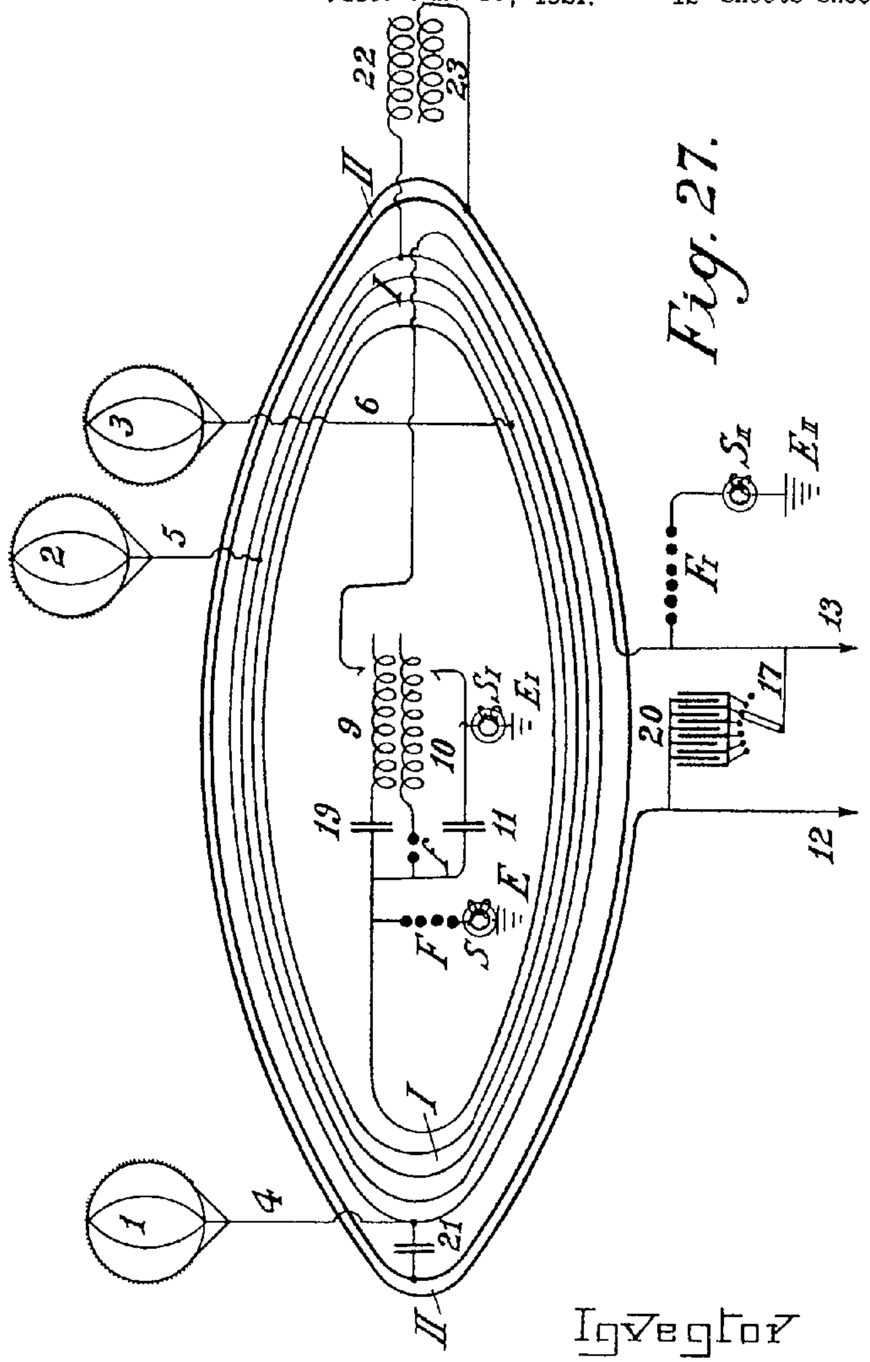
22
23
II
I
6
3
2
5
1
4
21
II
9
10
19
11
f
F
S
E
S_I
E_I
20
17
13
12
F_II
S_II
E_II
I

Fig. 27.

Inventor

Hermann Plauson

by Knight Bros

Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 9

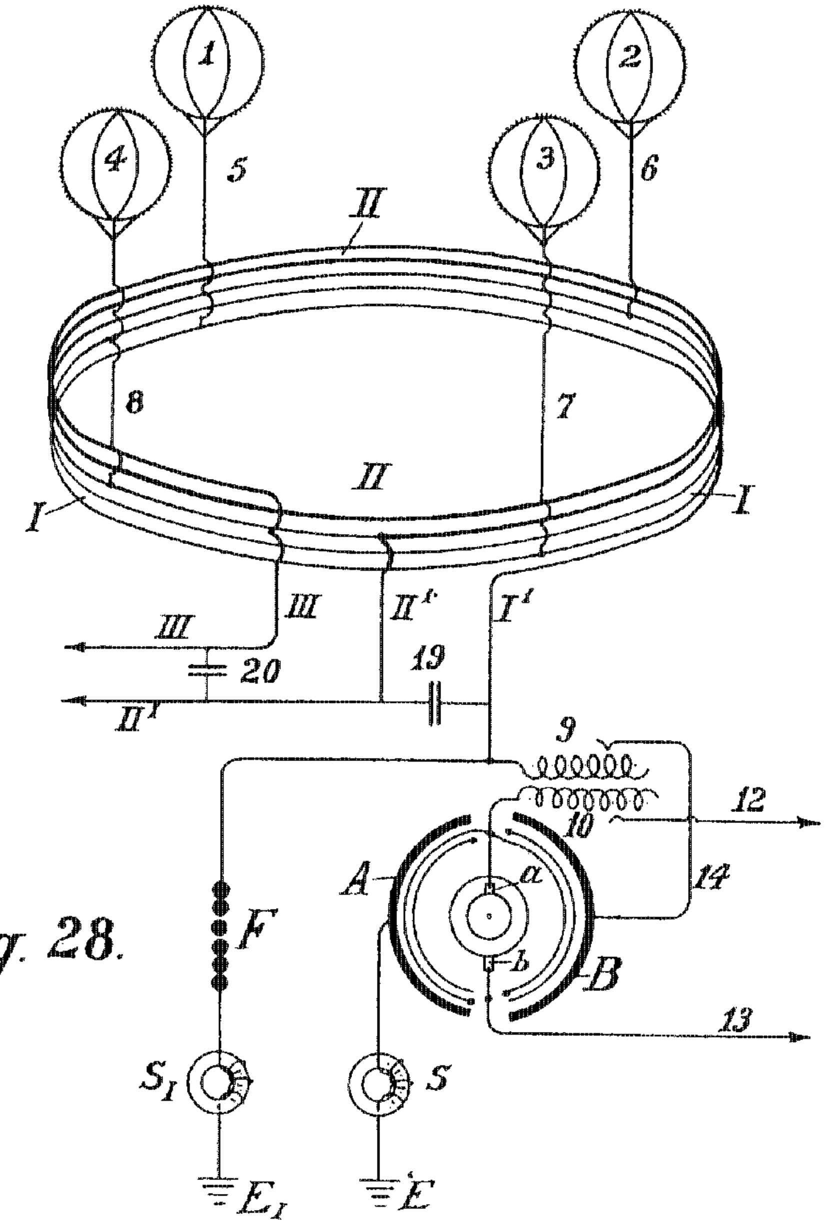
1
2
4
5
3
6
II
8
7
II
I
I
III
III
II¹
I¹
20
19
II¹
9
12
10
A
a
14
F
b
B
13
S₁
S
E₁
E

Fig. 28.

Inventor
Hermann Plauson
by
Knight Bros
attorneys

June 9, 1925. **1,540,998**

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921. 12 Sheets-Sheet 10

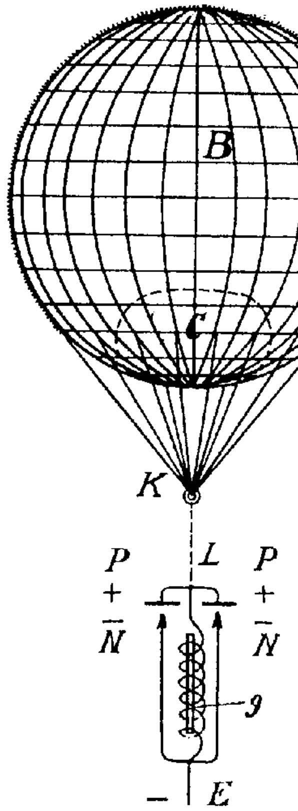

B
C
K
L
P
+
N
P
+
N
9
−
E

Fig. 29.

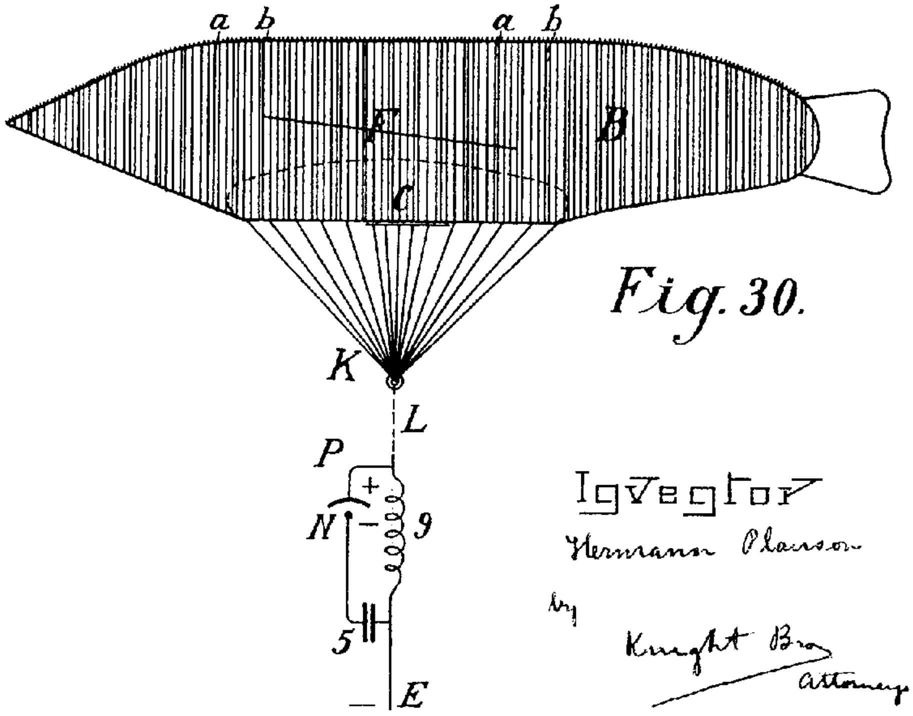

a
b
a
b
F
B
C
K
L
P
+
N
−
9
5
−
E

Fig. 30.

Inventor
Hermann Plauson
by
Knight Bros
Attorneys

June 9, 1925. **1,540,998**

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan 13 1921 12 Sheets-Sheet 11

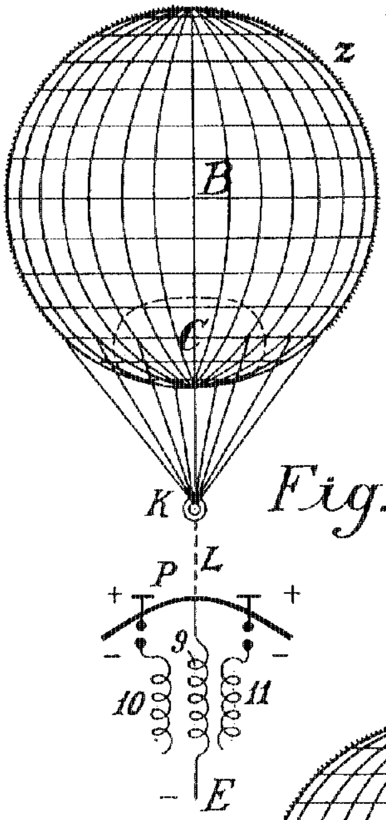
Z
B
C
K
L
P
+
+
−
−
9
10
11
−
E

Fig. 31.

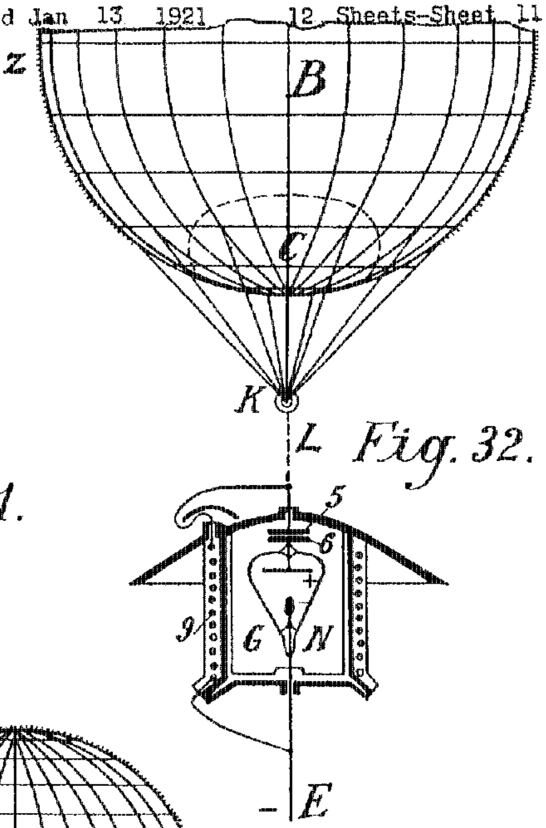
Z
B
C
K
L
5
6
+
9
G
N
−
E

Fig. 32.

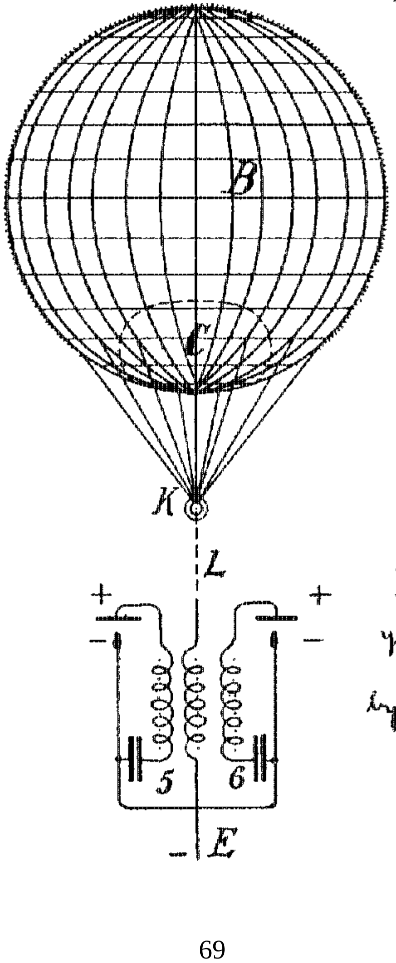

Fig. 33.

Inventor
Hermann Plauson
by Knight Bros
Attorneys

June 9, 1925. 1,540,998

H. PLAUSON

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY

Filed Jan. 13, 1921 12 Sheets-Sheet 12

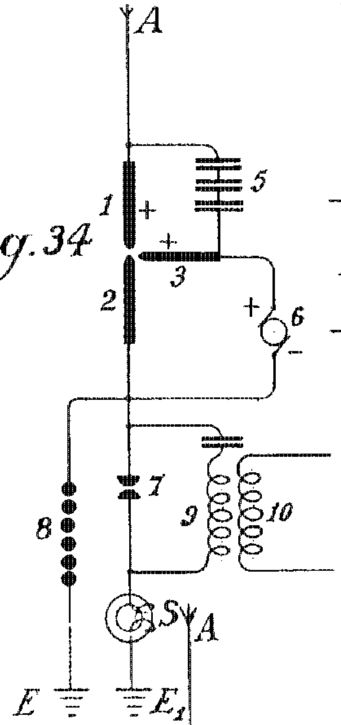
A
1
+
5
+
3
2
+
6
−
7
8
9
10
S
E
E1

Fig. 34

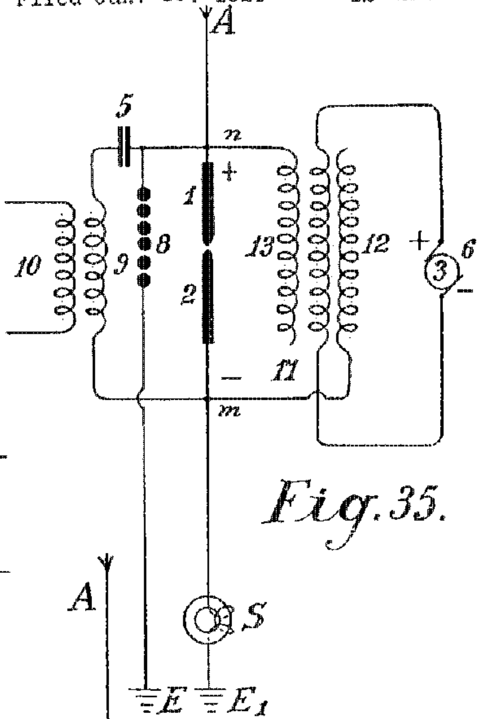
A
5
n
+
1
8
2
10
9
13
12
+
3
6
−
−
11
m
S
E
E1

Fig. 35.

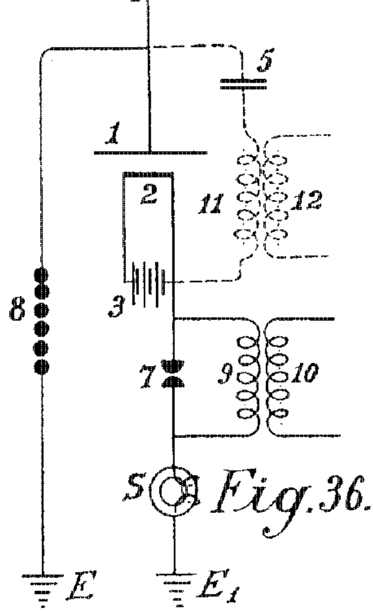
A
5
1
2
11
12
8
3
7
9
10
S
E
E1

Fig. 36.

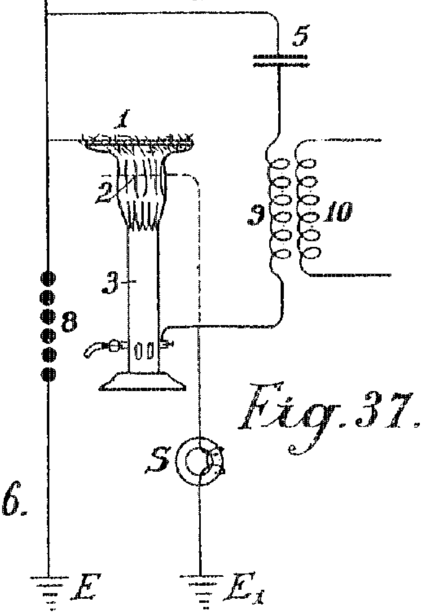
A
5
1
2
9
10
3
8
S
E
E1

Fig. 37.

Inventor

Hermann Plauson

by Knight Bros

Attorneys

Patented June 9, 1925. **1,540,998**

UNITED STATES PATENT OFFICE.

HERMANN PLAUSON, OF HAMBURG, GERMANY.

CONVERSION OF ATMOSPHERIC ELECTRIC ENERGY.

Application filed January 13, 1921. Serial No. 437,107.

To all whom it may concern:

Be it known that I, HERMANN PLAUSON, Esthonian subject, residing in Hamburg, Germany, have invented certain new and useful Improvements in the Conversion of Atmospheric Electric Energy, of which the following is a specification.

Methods of obtaining atmospheric electricity by means of metallic nettings set with spikes which are held by means of ordinary or anchored kite balloons made of fabrics and filled with hydrogen, are in theory already known. Atmospheric electricity obtained in this way has been suggested to be used in the form of direct current for the charging of accumulators. This knowledge however is at present only theoretical as the conversion in practice has hitherto been a failure. No means are known of protecting the apparatus from destruction by lightning. The balloons used for collecting the charge must also be made of very large size in order to be able to support the weight of the metallic netting and the heavy cable connections.

Instead of using heavy metallic netting as collectors attached to single air balloons of non-conducting materials which are liable to be torn and are permeable to the gas, it is proposed to use metallic balloon collectors which have the following important advantages—

(*a*) The metallic cases are impenetrable to helium and hydrogen; they also represent large metallic weather-proof collecting surfaces.

(*b*) Radio active means and the like may be easily applied internally or externally; whereby the ionization is considerably increased and therewith also the quantity of atmospheric electricity capable of being collected.

(*c*) Such balloon collectors of light metal do not require to be of large size as they have to carry only their own moderate weight, and that of the conducting cable or wire.

(*d*) The entire system therefore offers little surface for the action of storm and wind and is resistant and stable.

(*e*) Each balloon can be easily raised and lowered by means of a winch so that all repairs, recharging and the like can be carried out without danger during the operation.

It is further proposed to use a collecting aerial network of several separate collectors spread out in the air above the earth, which collectors are interconnected by electrical conductors.

According to this invention charges of atmospheric electricity are not directly converted into mechanical energy, and this forms the main difference from previous inventions, but the static electricity which runs to earth through aerial conductors in the form of direct current of very high voltage and low current strength is converted into electro-dynamic energy in the form of high frequency vibrations. Many advantages are thereby obtained and all disadvantages avoided.

The very high voltage of static electricity of a low current strength can be converted by this invention to voltages more suitable for technical purposes and of greater current strength. By the use of closed oscillatory circuits it is possible to obtain electromagnetic waves of various amplitude and thereby to increase the degree of resonance of such current. Such resonance allows various values of inductance to be chosen whereby again the governing of the starting and stopping of machines driven thereby by simply tuning the resonance between coils of the machine and the transformer circuit forming the resonance can easily be obtained. Further, such currents have the property of being directly available for various uses, even without employing them for driving motors, of which there may be particularly mentioned, lighting, production of heat and use in electro-chemistry.

Further, with such currents a series of apparatus may be fed without direct current supply through conductors and also the electro-magnetic high frequency currents may be converted by means of special motors adapted for electro-magnetic oscillations into mechanical energy, or finally converted by special machines into alternating current of low frequency or even into direct current of high potential.

The invention is more particularly described with reference to the accompanying diagrams in which:—

Figure 1 is an explanatory figure.

Figure 2 is a diagrammatic view of the simplest form.

Figure 3 shows a method of converting atmospheric electrical energy for use with motors.

Figure 4 is a diagram showing the use of protective means.

Figure 5 is a diagram of an arrangement for converting large current strengths.

Figure 6 is a diagram of an arrangement including controlling means.

Figure 7 shows means whereby the spark gap length can be adjusted.

Figure 8 shows a unipolar connection for the motor.

Figure 9 shows a weak coupled system suitable for use with small power motors.

Figures 10, 11 and 12 show modified arrangements.

Figure 13 shows a form of inductive coupling for the motor circuit.

Figure 14 is a modified form of Figure 13 with inductive coupling.

Figure 15 is an arrangement with non-inductive motor.

Figure 16 is an arrangement with coupling by condenser.

Figures 17, 18 and 19 are diagrams of further modifications.

Figure 20 shows a simple form in which the aerial network is combined with special collectors.

Figure 21 shows diagrammatically an arrangement suitable for collecting large quantities of energy.

Figure 22 is a modified arrangement having two rings of collectors.

Figure 23 shows the connections for three rings of collectors.

Figure 24 shows a collecting balloon and diagram of its connection of condenser batteries.

Figures 25 and 26 show modified collector balloon arrangements.

Figure 27 shows a second method of connecting conductor for the balloon aerials.

Figure 28 shows an auto-transformer method of connection.

Figure 29 shows the simplest form of construction with incandescent cathode.

Figure 30 shows a form with cigar shaped balloon.

Figure 31 is a modified arrangement.

Figure 32 shows a form with cathode and electrode enclosed in a vacuum chamber.

Figure 33 is a modified form of Figure 32.

Figure 34 shows an arc light collector.

Figure 35 shows such an arrangement for alternating current.

Figure 36 shows an incandescent collector with Nernst lamp.

Figure 37 shows a form with a gas flame.

Figure 1 illustrates a simple diagram for converting static electricity into dynamic energy of a high number of oscillations. For the sake of clearness in the drawings an influence machine is assumed to be employed and not an aerial antenna. 13 and 14 are combs for collecting the static electricity of the influence machine. 7 and 8 are spark discharging electrodes, 6 and 5 condensers, 9 an inductive primary coil, 10 secondary coil, 11 and 12 ends of conductors of the secondary coil 10. When the disc of the static influence machine is rotated by mechanical means, the combs collect the electric charges one the positive and the other the negative, and charge the condensers 5 and 6 until such a high potential is formed across the spark gap 7—8, that the spark gap is jumped. As the spark gap 7—8 forms a closed circuit with condensers 6 and 5, and inductive resistance 9, as is well known, waves of high frequency electromagnetic oscillations will pass in this circuit.

The high frequency of the oscillations produced in the primary circuit induces waves of the same periodicity in the secondary circuit. Thus in the primary circuit electromagnetic oscillations are formed by the passage of the spark over the spark gap and these waves are maintained by fresh charges of static electricity.

By suitably selecting the ratio between the number of the coils in the primary and secondary circuits with regard to a correct application of the co-efficients of resonance (capacity, inductance, and resistance) the high voltage of the primary circuit may be suitably converted into low voltage and high current strength.

When the oscillatory discharges in the primary circuit becomes weaker or entirely cease, the condensers are charged again by the static electricity until the accumulated charge again breaks down the spark gap. All this is repeated as long as electricity is produced by the static machine by employing mechanical energy.

An elementary form of the invention is shown in Figure 2 in which two spark gaps in parallel are used one of which may be termed the working gap 7 in Figure 2, whilst the second serves as a safety device for excess voltage and consists of a larger number of spark gaps than the working section, which gaps are arranged in series and are bridged by very small capacities as is illustrated in a_1, b_1, c_1, Figure 2 which allow of uniform sparking in the safety section.

In Figure 2 A is the aerial antenna for collecting charges of atmospheric electricity. 13 is the earth connection of the second part of the spark gap, 5 and 6 are condensers, 9 a primary coil. Now when through the aerial A the positive atmospheric electricity seeks to combine with the negative charge to earth, this is prevented by (the air gap between) the spark gaps. The resistance of the spark gap 7 is, as shown in the drawings, lower than that of the other safety section which consists of three spark gaps connected in series, and consequently a three times greater air resistance is offered by the latter.

So long therefore, as the resistance of the spark gap 7 is not overloaded, so that the other spark gaps have an equal resistance with it the discharges take place only over spark gap 7. Should however the voltage be increased by any influences so that it might be dangerous for charging the condensers 5 and 6 or for the coil insulation 9 and 10 in consequence of break down, by a correct regulation of this spark gap the second spark gap can discharge free from inductive effects direct to earth without endangering the machine.

Without this second spark gap, arranged in parallel having a higher resistance than the working spark gap it is impossible to collect and render available large quantities of electrical energy.

The action of this closed oscillation circuit consisting of spark gap 7, two condensers 5 and 6, primary coil 9, and also secondary coil 10 is exactly the same as the one described in Figure 1 with the arrangement of the static induction machine with the only difference that here the second spark gap is provided. The electromagnetic high frequency alternating current obtained can be tapped off from the conductors 11 and 12 for lighting and heating purposes. Special kinds of motors adapted for working with these peculiar electrical charges may be connected at 14 and 15 which can work with static electricity charges or with high frequency oscillations.

In addition to the use of spark gaps in parallel a second measure of security is also necessary for taking off the current. This precaution consists according to this invention, in the introduction of and method of connecting certain protective electromagnets or choking coils in the aerial circuit as shown by S in Figure 3.

A single electromagnet only having a core of the thinnest possible separate laminations is connected with the aerial.

In the case of high voltages in the aerial network or at places where there are frequent thunder storms, several such magnets may however be connected in series.

In the case of large units or plants several electromagnets can be employed in parallel or in series parallel.

The windings of these electromagnets may be simply connected in series with the aerials. In this case the winding preferably consists of several thin parallel wires, which make up together, the necessary section.

The winding may be made of primary and secondary windings in the form of a transformer. The primary winding will be then connected in series with the aerial network, and the secondary winding more or less short-circuited over a regulating resistance or an induction coil. In the latter case it is possible to regulate to a certain extent the effect of the choking coils. In the further description of the connecting and constructional diagrams the aerial electromagnet choke coil is indicated by a simple ring S.

Figure 3 shows the simplest way of converting atmospheric electricity into electromagnetic wave energy by the use of special motors adapted for high oscillatory currents or static charges of electrical energy. Recent improvements in motors for working with static charges and motors working by resonance, that is to say, having groups of tuned electromagnetic cooperating circuits render this possible but such do not form part of the present invention.

A motor adapted to operate with static charges will for the sake of simplicity be diagrammatically indicated by two semicircles 1 and 2 and the rotor of the motor by a ring M. (Figure 3.) A is a vertical aerial or aerial network. S the safety choke or electromagnet with coil O as may be seen is connected with the aerial A. Adjacent the electromagnet S the aerial conductor is divided into three circuits, the circuit 8 giving the safety spark gap, the circuit 7 with the working spark gap, and then a circuit including the stator terminal 1, the rotor and stator terminal 2 at which a connection is made to the earth wire. The two spark gaps are also connected metallically with the earth wire. The method of working these diagrams is as follows:

The positive atmospheric electric charge collected tends to combine with the negative electricity (or earth electricity) connected with the earth wire. It travels along the aerial A through the electromagnet S without being checked as it flows in the same direction as the direct current. Further, its progress is arrested by two sparks gaps placed in the way and the stator condenser surfaces. The stator condenser surfaces are charged until the charge is greater than the resistance of the spark gap 7, whereupon a spark springs over the spark gap 7 and an oscillatory charge is obtained as by means of the motor M, stator surfaces 1 and 2, and spark gap 7, a closed oscillation circuit is obtained for producing the electromagnetic oscillations. The motor here forms the capacity and the necessary inductance and resistance, which, as is well known, are necessary for converting static electricity into electromagnetic wave energy.

The discharges formed are converted into mechanical energy in special motors and can not reach the aerial network by reason of the electromagnet or choke. If, however, when a spark springs over the spark gap 7 a greater quantity of atmospheric electricity tends to flow to earth, a counter voltage is induced in the electromagnet, which is greater the more rapidly and strongly the flow of current direct to the earth is. By the formation of this opposing voltage a sufficiently high resistance is offered to the flow of atmospheric electricity direct to earth to prevent a short circuit with the earth.

The circuit containing spark gap 8 having a different wave length which is not in resonance with the natural frequency of the motor, does not endanger the motor and serves as security against excess voltage, which, as practical experiments have shown, may still arise in certain cases, but can be conducted direct to earth through this spark gap.

In the diagram illustrated in Figure 4 the spark gap 7 is shunted across condensers 5 and 6 from the motor M. This construction affords mainly a better insulation of the motor against excess voltage and a uniform excitation through the spark gap 7.

In Figure 5 a diagram is illustrated for transforming large current strengths which may be employed direct without motors, for example, for lighting or heating purposes. The main difference is that here the spark gap consists of a star shaped disc 7 which can rotate on its own axis and is rotated by a motor opposite similarly fitted electrodes 7^a. When separate points of stars face one another, discharges take place, thus forming an oscillation circuit over condensers 5 and 6 and inductance 9 for oscillatory discharges. It is evident that a motor may also be directly connected to the ends of the spiral 9.

The construction of the diagram shown in Figure 6 permits of the oscillation circuit of the motor being connected with an induction coil. Here a regulating inductive resistance is introduced for counter-acting excess voltages in the motor. By cutting the separate coils 9 (coupled inductively to the aerial) in or out the inductive action on the motor may be more or less increased or variable aerial action may be exerted on the oscillation circuit.

In Figure 7 the oscillation circuit is closed through the earth (E and E_1). The spark gap 7 may be prolonged or shortened by more or fewer spark gaps being successively connected by means of a contact arm 7^b.

Diagram 8 shows a unipolar connection of the motor with the aerial network. Here two oscillation circuits are closed through the same motor. The first oscillation circuit passes from aerial A through electromagnet S, point x, inductance 9^a to the earth condenser 6 and further, over spark gap 7 to the aerial condenser 5 and back to x. The second oscillation circuit starts from the aerial condenser 5 at the point x^1 over the inductance 9 to the earth condenser 6 at the point x^3 and through the condenser 6 over the spark gap 7 back to x^1. The motor itself is inserted between the two points of the spark gap 7. From this arrangement slightly damped oscillation wave currents are produced.

In the diagram illustrated in Figure 9 a loosely coupled system of connections is illustrated which is assumed to be for small motors for measuring purposes. A indicates the aerial conductor, S the electromagnet in the aerial conductor, 9 the inductance, 7 the spark gap, 5 and 6 condensers, E the earth, M the motor, and 1 and 2 stator connections of the motor. The motor is directly metallically connected with the oscillation circuit.

In Figure 10 a purely inductive coupling is employed for the motor circuit. The motor is connected with the secondary wire 10 as may be seen in Figure 11 in a somewhat modified diagram connection. The same applies to the diagram of Figure 12.

The diagrams hitherto described preferably allow of motors of small and medium strength to be operated. For large aggregates, however, they are too inconvenient as the construction of two or more oscillation circuits for large amounts of energy is difficult; the governing is still more difficult and the danger in switching on or off is greater.

A means of overcoming such difficulties is shown in Figure 13. The oscillation circuit here runs starting from the point x over condenser 5, variable inductance 9, spark gap 7 and the two segments (3^a and 4^a) forming arms of a Wheatstone bridge, back to x. If the motor is connected by brushes 3 and 4 transversely to the two arms of the bridge as shown in the drawings, electromagnetic oscillations of equal sign are induced in the stator surfaces 1 and 2 and the motor does not revolve. If however, the brushes 3 and 4 are moved in common with the conducting wires 1 and 2 which connect the brushes with the stator poles a certain alteration or displacement of the polarity is obtained and the motor commences to revolve.

The maximum action will result if one brush 3 comes on the central sparking contact 7 and the other brush 4 on the part x. They are however, usually in practice not brought on to the central contact 7 but only held in the path of the bridge segments 4^a and 3^a in order not to connect the spark gaps with the motor oscillation circuit.

As however, the entire oscillation energy can thereby not act on the motor it is better to carry out the same system according to the diagram 14. The diagram 14 differs from the foregoing only by the motor not being directly metallically connected with the segments of the commutator, but only a primary coil 9 which induces in a secondary coil 10, current which feeds the motor M and takes the place of the rotor. By this arrangement a good transforming action is obtained, a loose coupling and also an oscillation circuit without a spark gap.

In Figure 15 the motor is not purely inductively as in 14, but directly metallically branched off from the primary coil (at x and x^1) after the principle of the auto-transformer.

In Figure 16 instead of an inductance a condenser 6 is in similar manner, and for the same object inserted between the segments 3^a and 4^a. This has the advantage that the segments 3^a and 4^a need not be made of solid metal but may consist of spiral coils whereby a more exact regulation is possible and further motors of high inductance may be employed.

The arrangements of Figures 17, 18 and 19 may be employed for use with resonance and particularly with induction condenser motors; between the large stator induction condenser surfaces, small reversing pole condensers are connected, which, as may be seen from Figures 17, 18 and 19 are led together to earth. Such reversing poles have the advantage that with large quantities of electrical energy the spark formation between the separate oscillation circuits ceases.

Figure 19 shows a further method which prevents electromagnetic oscillations of high number of alternations formed in the oscillation circuit striking back to the aerial conductor. It is based on the well known principle that a mercury lamp, one electrode of which is formed of mercury, the other of solid metal such as steel allows an electric charge to pass in only one direction from the mercury to the steel and not vice versa. The mercury electrode of the vacuum tube N is therefore connected with the aerial conductor and the steel electrode with the oscillation circuit. From this it results that charges can pass only from the aerial through the vacuum tube to the oscillation circuit, but not vice versa. Oscillations which are formed on being transformed in the oscillation circuit cannot pass to the aerial conductor.

In practice these vacuum tubes must be connected behind an electromagnet as the latter alone affords no protection against the danger of lightning.

As regards the use of spark gaps, all arrangements as used for wireless telegraphy may be used. Of course the spark gaps in large machines must have a sufficiently large surface. In very large stations they are cooled in liquid carbonic acid or better still in liquid nitrogen or hydrogen; in most cases the cooling may also take place by means of liquefied low homologues of the metal series or by means of hydrocarbons the freezing point of which lies at between −90° C. and −40° C. The spark gap casing must also be insulated and be of sufficient strength to be able to resist any pressure which may arise. Any undesirable excess super-pressure which may be formed must be automatically let off. I have employed with very good results mercury electrodes which were frozen in liquid carbonic acid, the cooling being maintained during the operation from the outside through the walls.

Figure 20 is one of the simplest forms of construction of an aerial network in combination with collectors, transformers and the like illustrated diagrammatically. E is here the earth wire, 8 the safety spark gap, 7 the working spark gap, 1 and 2 the stator surfaces of the motor, 5 a condenser battery, S the protective magnet which is connected with the coil in the aerial conductor, A^1 to A^{10} aerial antennæ with collecting balloons, N horizontal collecting or connecting wires from which, to the centre a number of connections run.

The actual collectors consist of metal sheaths preferably made of an aluminium magnesium alloy, and are filled with hydrogen or helium and are attached to copper plated steel wires. The size of the balloon is selected so that the actual weight of the balloon and the weight of the conducting wire is supported thereby. On the top of the balloon aluminium spikes, made and gilded in a special manner hereinafter described, are arranged in order to produce a conductor action. Small quantities of radium preparations, more particularly polonium-ionium or mesothorium preparations considerably increase the ionization, and therewith the action of these collectors.

In addition to metal balloons, fabric balloons which are superficially metal coated according to Schoop's metal spraying process, may however also be employed. A metallic surface may also be produced by lacquering with metallic bronzes, preferably according to Schoop's spraying process or lacquering with metallic bronze powders in two electrical series of widely different metals, because thereby the collecting effect is considerably increased.

Instead of the ordinary round balloons, elongated cigar shaped ones may be employed. In order also to utilize the frictional energy of the wind, patches or strips of non-conducting substances which produce electricity by friction, may be attached to the metallized balloon surfaces. The wind will impart a portion of its energy in the form of frictional electricity, to the balloon casing, and thereby the collecting effect is substantially increased.

In practice however, very high towers (up to 300 metres is fully admissible) may be employed as antennæ. In these towers copper tubes rise freely further above the top of the tower. A gas lamp secured against the wind is then lit at the point of the copper tube and a netting is secured to the copper tube over the flame of this lamp to form a collector. The gas is conveyed through the interior of the tube up to the summit. The copper tube must be absolutely protected from moisture at the place at which it enters the tower and also rain must be prevented running down the walls of the tower which might lead to a bad catastrophe. This is done by bell shaped enlargements which expand downwards, being arranged in the tower in the form of high voltage insulators of Siamese pagodas.

Special attention must be devoted to the foundations of such towers. They must be well insulated from the ground, which may be obtained by first embedding a layer of concrete in a box form to a sufficient depth in the ground and inserting in this an asphalt lining and then glass bricks cast about 1 or 2 metres in thickness. Over this in turn there is a ferro-concrete layer in which alone the metal foot of the tube is secured. This concrete block must be at least 2 metres from the ground and be fully protected at the sides by a wooden covering, from moisture. In the lower part of the tower a wood or glass house for the large condenser batteries or for the motors may be constructed. In order to lead the earth connection to the ground water, a well insulated pit constructed of vitreous bricks, must be provided. Several such towers are erected at equal distances apart and connected with a horizontal conductor. The horizontal connecting wires may either run directly from tower to tower or be carried on bell shaped insulators similar to those in use for high voltage conductors. The width of the network may be of any suitable size and the connection of the motors can take place at any suitable places.

In order to collect large quantities of electricity with few aerials it is well to provide the aerial conductor with batteries of condensers as shown in two methods of construction in Figures 21 and 22. In Figure 21 the batteries of condensers 5 are connected on the one hand with the aerial electricity collectors Z by the aerial conductor A, and on the other hand interconnected in series with an annular conductor from which horizontal conductors run to the connecting points C to which the earth wire is connected.

Figure 22 shows a similar arrangement. Should two such series of antennæ rings be shown by a voltmeter to have a large difference of potential (for example, one in the mountains and one in the plain) or even of different polarity these differences may be compensated for by connecting sufficiently large condenser batteries ($5, 5^a, 5^b$) by means of Maji star conductors D and D^1. In Figure 23 a connection of three such rings of collectors to form a triangle with a central condenser battery is illustrated.

The condenser batteries of such large installations must be embedded in liquefied gases or in liquids freezing at very low temperatures. In such cases a portion of the atmospheric energy must be employed for liquefying these gases. It is also preferable to employ pressure. By this means the condenser surfaces may be diminished, and still allow for large quantities of energy to be stored, secure against breakdown. For smaller installations the immersing of the condensers in well insulated oil or the like, suffices. Solid substances on the other hand cannot be employed as insulators.

The arrangement in the diagrams hitherto described was always such that the condenser batteries were connected with both poles directly to the aerial conductors. An improved diagram of the connections for obtaining atmospheric electricity for the condenser batteries has however, been found to be very advantageous, this arrangement consists in that they are connected by only one pole (unipolar) to the collecting net-work. Such a method of arrangement is very important, as by means of it a constant current and an increase of the normal working pressure or voltage is obtained. If for example a collecting balloon aerial which is allowed to rise to a height of 300 metres, shows 40,000 volts above earth voltage, in practice it has been found that the working voltage (with a withdrawal of the power according to the method hereinbefore described by means of oscillating spark gaps and the like) is only about 400 volts. If however, the capacity of the condenser surfaces be increased, which capacity in the above mentioned case was equal to that of the collecting surface of the balloon aerials, to double the amount, by connecting the condenser batteries with only one pole, the voltage rises under an equal withdrawal of current up to and beyond 500 volts. This can only be ascribed to the favourable action of the connecting method.

In addition to this substantial improvement it has also been found preferable to insert double inductances with electromagnets and to place the capacities preferably between two such electromagnets. It has also been found that the useful action of such condensers can be further increased if an induction coil be connected as inductive resistance to the unconnected pole of the condenser, or still better if the condenser itself be made as an induction condenser. Such a condenser may be compared with a spring which when compressed carries in itself accumulated force, which it again gives off when released. In charging, a charge with reversed sign is formed at the other free condenser pole, and if through the spark gap a short circuit results, the accumulated energy is again given back since now new quantities of energy are induced at the condenser pole connected with the conductor network, which in fact charges with opposite signs to that at the free condenser pole. The new induced charges have of course the same sign as the collector network. The whole voltage energy in the aerial is thereby however increased. In the same space of time larger quantities of energy are accumulated than is the case without such inserted condenser batteries.

In Figures 24 and 25 two different diagrams of connections are more exactly illustrated, Figure 24 shows a collecting balloon and the diagram of the connections to earth. Figure 25 four collecting balloons and the parallel connection of the condenser batteries belonging thereto.

A is the collecting balloon made of an aluminium magnesium alloy (electron metal, magnalium) of a specific gravity of 1.8 and a thickness of plate 0.1 to 0.2 mm. Inside there are eight strong vertical ribs of T shaped section about 10 to 20 mm. in height and about 3 mm. in thickness with the projecting part directed inwards (indicated by *a*, *b*, *c*, *d* and so forth); they are riveted together to form a firm skeleton and are stiffened in a horizontal direction by two cross ribs. The ribs are further connected with one another internally and transversely by means of thin steel wires, whereby the balloon obtains great power of resistance and elasticity. Rolled plates of 0.1 to 0.2 mm. in thickness made of magnalium alloy are then either soldered or riveted on this skeleton so that a fully metallic casing with smooth external surface is obtained. Well silvered or coppered aluminium plated steel wires run from each rib to the fastening ring 2. Further, the coppered steel hawser L preferably twisted out of separate thin wires (shown in dotted lines in Figure 24) and which must be long enough to allow the balloon to rise in the desired height, leads to a metal roller or pulley 3 and from thence to a winch W, well insulated from the earth. By means of this winch, the balloon, which is filled with hydrogen, or helium, can be allowed to rise to a suitable height (300 to 5,000 metres) and brought to the ground for recharging or repairs.

The actual current is taken directly through a friction contact from the metal roller 3 or from the wire, or even from the winch or simultaneously from all three by means of brushes (3, 3^a and 3^b). Beyond the brushes the conductor is divided, the paths being:—firstly over 12 to the safety spark gap 8, from thence to the earth conductor E^1, and secondly over electromagnet S^1, point 13, to a second loose electromagnet having an adjustable coil S^2, then to the spark gap 7 and to the second earth conductor E^2. The actual working circuit is formed through the spark gap 7, condensers 5 and 6, and through the primary coil 9; here the static electricity formed by oscillatory discharges is accumulated and converted into high frequency electromagnetic oscillations. Between the electromagnets S^1 and S^2 at the crossing point 13, four condenser batteries are introduced which are only indicated diagrammatically in the drawings each by one condenser. Two of these batteries (16 and 18) are made as plate condensers and prolonged by regulating induction coils or spirals 17 and 19 while the two others (21 and 23) are induction condensers. As may be seen from the drawings each of the four condenser batteries 16, 18, 21. 23 is connected only by one pole to the aerial or to the collector conductor. The second poles 17, 19, 22, 24 are open. In the case of plate condensers having no inductive resistance an induction coil is inserted. The object of such a spiral or coil is the displacement of phase of the induction current by ¼ periods, whilst the charging current of the condenser poles which lie free in the air, works back to the collector aerial. The consequence of this is that in discharges in the collector aerial the back inductive action of the free poles allows a higher voltage to be maintained in the aerial collecting conductor than would otherwise be the case. It has also been found that such a back action has an extremely favourable effect on the wear of the contacts. Of course the inductive effect may be regulated at will within the limits of the size of the induction coil, the length of the coil in action being adjustable by means of wire connection without induction (see Fig. 24, No. 20).

S^1 and S^2 may also be provided with such regulating devices in the case of S^2 (illustrated by 11). If excess voltage be formed it is conducted to earth through the wire 12 and spark gap 8 or through any other suitable apparatus, since this formation would be dangerous for the other apparatus.

The action of these condenser batteries has already been hereinbefore described.

The small circles on the collector balloon indicate places at which zinc amalgam or gold amalgam or other photoelectric acting metals in the form of small patches in extremely thin layers (.01 to .05 mm. in thickness) are applied to the balloon casing of light metal. Such metallic patches may also be applied to the entire balloon as well as in greater thickness to the conducting network. The capacity of the collector is thereby considerably strengthened at the surface. The greatest possible effect in collecting may be obtained by polonium amalgams and the like. On the surface of the collector balloon metal points or spikes are also fixed along the ribs, which spikes serve particularly for collecting the collector charge. Since it is well known that the resistance of the spikes is less the sharper the spike is, for this purpose it is therefore extremely important to employ as sharp spikes as possible. Experiments made as regards these have shown that the formation of the body of the spike or point also plays a large part. for example. spikes made of bars or rollers with smooth surfaces, have a many times greater point resistance as collector accumulator spikes than those with rough surfaces. Various kinds of spike bodies have been experimented with for the collector balloons hereinbefore mentioned. The best results were given by spikes which were made in the following way. Fine points made of steel, copper, nickel, or copper and nickel alloys, were fastened together in bundles and then placed as anode with the points in a suitable electrolyte (preferably in hydrochloric acid or muriate of iron solutions) and so treated with weak current at 2 to 3 volts pressure. After 2 to 3 hours according to the thickness of the spikes or pins the points become extremely sharp and the bodies of the spikes have a rough surface. The bundle can then be removed and the acid washed off with water. The spikes are then placed as cathode in a bath consisting of solution of gold, platinum, iridium, paladium or wolfram salts or their compounds and coated at the cathode galvanically with a thin layer of precious metal, which must however be sufficiently firm to protect them from atmospheric oxidation.

Such spikes act at a 20 fold lower voltage almost as well as the best and finest points made by mechanical means. Still better results are obtained if polonium or radium salts are added to the galvanic bath when forming the protective layer or coating. Such pins have a low resistance at their points and even at one volt and still lower pressures have an excellent collector action.

In Figure 24 the three unconnected poles are not connected with one another in parallel. That is quite possible in practice without altering the principle of the free pole. It is also preferable to interconnect in parallel to a common collector network, a series of collecting aerials.

Figure 25 shows a diagram for such an installation. A^1. A^2, A^3, A^4 are four metal collector balloons with gold or platinum coated spikes which are electrolytically made in the presence of polonium emanations or radium salts, which spikes or needles are connected over four electro-magnets S^1, S^2, S^3, S^4, through an annular conductor R. From this annular conductor four wires run over four further electromagnets S^a, S^b, S^c, S^d, to the connecting point 13. There the conductor is divided, one branch passing over 12 and the safety spark gap 8 to the earth at E^1, the other over inductive resistance J and working spark gap 7 to the earth at E^2. The working circuit, consisting of the condenser 5 and 6 and a resonance motor or a condenser motor M, such as hereinbefore described, is connected in proximity round the sparking gap section 7.

Instead of directly connecting the condenser motor of course the primary circuit for high frequency oscillatory current may also be inserted.

The condenser batteries are connected by one pole to the annular conductor R and can be either inductionless (16 and 18) or made as induction condensers as shown by 21 and 23. The free poles of the inductionless condensers are indicated by 17 and 19, those of the induction condensers by 22 and 24. As may be seen from the drawings all these poles 17, 22, 19, 24 may be interconnected in parallel through a second annular conductor without any fear that thereby the principle of the free pole connection will be injured. In addition to the advantages already set forth the parallel connection also allows of an equalization of the working pressure in the entire collector network. Suitably constructed and calculated induction coils 25 and 26 may also be inserted in the annular conductor of the free poles, by means of which a circuit may be formed in the secondary coils 27 and 28 which allows current produced in this annular conductor by fluctuations of the charges or the like appearances to be measured or otherwise utilized.

According to what has been hereinbefore stated separate collector balloons may be connected at equidistant stations distributed over the entire country, either connected directly with one another metallically or by means of intermediate suitably connected condenser batteries through high voltage conductors insulated from earth. The static electricity is converted through a spark gap into dynamic energy of a high number of oscillations and may in such form be coupled as a source of energy by means of a suitable method of connecting, various precautions being observed, and with special regulations. The wires leading from the collector balloons have hitherto been connected through an annular conductor without this endless connection, which can be regarded as an endless induction coil, being able to exert any action on the whole conductor system.

It has now been found that if the network conductor connecting the aerial collector balloons with one another is not made as a simple annular conductor, but preferably short circuited in the form of coils over a condenser battery or spark gap or through thermionic tubes or valves or audions, then the total collecting network exhibits quite new properties. The collection of atmospheric electricity is thereby not only increased but an alternating field may be easily produced in the collector network. Further, the atmospheric electrical forces showing themselves in the higher regions may also be directly obtained by induction. In Figures 26 and 28 a form of construction is shown on the basis of which the further foundations of the method will be more particularly explained.

In Figure 26 1, 2, 3, 4 are metal collector balloons, 5, 6, 7, 8 their metallic aerial conductors and I the actual collector network. This consists of five coils and is mounted on high voltage insulators in the air, on high voltage masts (or with a suitable construction of cable embedded in the earth). One coil has a diameter of 1 to 100 km. or more. S and S^1 are two protective electromagnets, F the second safety section against excess voltage, E its earth conductor and E^1 the earth conductor of the working section. When an absorption of static atmospheric electricity is effected through the four balloon collectors, the current in order to reach the earth connection E^1 must flow spirally through the collector network over the electromagnet S, primary induction coil 9, conductor 14, anode A of the audion tube, incandescent cathode K, as the way over the electromagnet and safety spark gap F offers considerably greater resistance. Owing to the fact that the accumulated current flows in one direction, an electromagnetic alternating field is produced in the interior of the collector network coil, whereby the whole free electrons are directed more or less into the interior of the coil. An increased ionization of the atmosphere is therefore produced. In consequence of this the points mounted on the collector balloon show a considerably reduced resistance and therefore increased static charges between the points on the balloon and the surrounding atmosphere are produced. The result of this is a considerably increased collector effect.

A second effect which could not be obtained otherwise is obtained by the electromagnetic alternating field which running parallel to the earth surface, acts more or less with a diminishing or increasing effect on the earth magnetic field, whereby in the case of fluctuations in the current a return induction current of reversed sign is always produced in the collector coil by earth magnetism. Now if, however, a constantly pulsating continuous alternating field is produced as stated in the above collector network I, an alternating current of the same periodicity is produced also in the collecting network coil. As the same alternating field is further transmitted to the aerial balloon, the resistance of its points is thereby considerably reduced, whilst the collector action is considerably increased. A further advantage is that positive electrons which collect on the metal surfaces during the conversion into dynamic current produce a so-called drop of potential of the collector area. As an alternating field is present, the negative ions surrounding the collector surfaces, when discharge of the collector surfaces takes place produce by the law of induction, an induction of reversed sign on the collector surface and so forth (that is to say again a positive charge). In addition to the advantages hereinbefore set forth, the construction of connecting conductors in coil form when of sufficiently large diameter, allows of a utilization of energy arising in higher regions also in the simplest way. As is well known electric discharges frequently take place at very great elevations which may be observed, such as St. Elmo's fires or northern lights. These energy quantities have not been able to be utilized up to now. By this invention all these kinds of energy, as they are of an electromagnetic nature and the direction of the axis of the collector coils stands at right angles to the earth's surface, can be more or less absorbed in the same way as a receiver in wireless telegraphy absorbs waves coming from a far distance. With a large diameter of the spiral it is possible to connect large surfaces and thereby to take up also large quantities of energy.

It is well known that large wireless stations in the summer months, and also in the tropics are very frequently unable to receive the signals in consequence of interruptions which are caused by atmospheric electricity, and this takes place with vertical coils of only 40 to 100 m. diameter. If on the contrary horizontal coils of 1 to 100 km. diameter be employed very strong currents may be obtained through discharges which are constantly taking place in the atmosphere. Particularly in the tropics or still better in the polar regions where the northern lights are constantly present, large quantities of energy may probably be obtained in this way. A coil with several windings should act the best. In similar manner any alteration of the earth magnetism should act inductively on such a coil.

It is not at all unlikely that earthquakes and spots on the sun will also produce an induction in such collector coils of sufficient size. In similar manner this collector conductor will react on earth currents more particularly when they are near the surface of the earth or even embedded in the earth. By combining the previous kind of current collectors so far as they are adapted for the improved system with the improved possibilities of obtaining current the quantities of free natural energy which are to be obtained in the form of electricity are considerably increased.

In order to produce in the improved collector coil uniform current oscillations of an undamped nature so-called audion high vacuum or thermionic tubes of suitable connection are employed instead of the previously known spark gaps (Fig. 26, Nos. 9–18). The main aerial current flows through electromagnet S (which in the case of a high number of alternations is not connected here but in the earth conductor E^1) and may be conveyed over the primary coils in the induction winding through wire 14 to the anode A of the high vacuum grid tube. Parallel with the induction resistance 9 a regulating capacity of suitable size, such as condenser 11 is inserted. In the lower part of the vacuum grid tube is arranged the incandescent filament or the cathode K which is fed through a battery B. From the battery B two branches run, one to the earth conductor E^1 and the other through battery B^1 and secondary coil 10 to the grid anode *g* in the vacuum tube. By the method of connections shown in dotted lines, a desired voltage at the grid electrode *g* may also be produced through the wire 17 which is branched off from the main current conductor through switches 16 and some small condensers (*a*, *b*, *c*, *d*) connected in series, and conductor 18, without the battery B^1 being required.

The action of the entire system is somewhat as follows:—

On the connecting conductor of the aerial collector network being short circuited to earth, the condenser pole 11 is charged and slightly damped oscillations are formed in the short circuited existing oscillation circuit formed of the condenser 11 and self inductance 9. In consequence of the coupling through coil 10, fluctuations of voltage take place in the grid circuit 15 with the same frequency, which, fluctuations in turn influence the strength of the electrode current passing through the high vacuum amplyfying tube and thus produce current fluctuations of the same frequency in the anode circuit. A permanent supply of energy to the oscillation circuits 9 and 10 consequently takes place, until a condition of balance is set up, in which the consumed oscillation energy is equal to that absorbed. Thereby constant undamped oscillations are now produced in the oscillation circuits 9–11.

For regular working of such oscillation producers high vacuum strengthening tubes are necessary and it is also necessary that the grid and anode voltages shall have a phase difference of 180° so that if the grid is negatively charged, then the anode is positively charged and vice versa. This necessary difference of phase may be obtained by most varied connections, for example, by placing the oscillation circuit in the grid circuit or by separating the oscillation circuit and inductive coupling from the anodes and the grid circuit and so forth.

A second important factor in this way of converting static atmospheric electricity into undamped oscillations is that care must be taken that the grid and anode voltages have a certain relation to one another; the latter may be obtained by altering the coupling and a suitable selection of the self induction in the grid circuit, or as shown by dotted lines 18, 17, 16 by means of a larger or smaller number of condensers of suitable size connected in series; in this case the battery B^1 may be omitted. With a suitable selection of the grid potential a glow discharge takes place between the grid *g* and the anode A, and accordingly at the grid there is a cathode drop and a dark space is formed. The size of this cathode drop is influenced by the ions which are emitted in the lower space in consequence of shock ionization of the incandescent cathodes K and pass through the grid in the upper space. On the other hand the number of the ions passing through the grid is dependent on the voltage between the grid and the cathode. Thus if the grid voltage undergos periodic fluctuations (as in the present case) the amount of the cathode drop at the grid fluctuates and consequently the internal resistance of the tube correspondingly fluctuates, so that when a back coupling of the feed circuit with the grid circuit takes place, the necessary means are afforded for producing undamped oscillations and of taking current, according to requirements from the collecting conductor.

The frequency of the undamped oscillations produced is with a suitably loose coupling equal to the self frequency of the oscillation circuits 9 and 10. By a suitable selection of the self induction of the coil 9 and capacity 11 it is possible to extend from frequencies which produce electromagnetic oscillations of only a few metres wave length down to the lowest practical alternating current frequency. For large installations a suitable number of frequency producing tubes in the form of the well known high vacuum transmission tubes of .5 to 2 kw. in size may be connected in parallel so that in this respect no difficulty exists.

The use of such tubes for producing undamped oscillations, and also the construction and method of inserting such transmission tubes in an accumulator or dynamo circuit is known and also that such oscillation producing tubes only work well at voltages of 1,000 up to 4,000 volts, so that on the contrary their use at lower voltages is considerably more difficult. By the use of high voltage static electricity this method of producing undamped oscillations as compared with that through spark gaps must be regarded as an ideal solution particularly for small installations of outputs of from 1 to 100 kw.

By the application of safety spark gaps, with interpolation of electro-magnets, not only is short circuiting avoided but also the taking up of current is regulated. Oscillation producers inserted in the above way form a constantly acting electromagnetic alternating field in the collector coil, whereby as already stated, a considerable accumulating effect takes place. The withdrawal wire or working wire is connected at 12 and 13, but current may be taken by means of a secondary coil which is firmly or movably mounted in any suitable way inside the large collector coil, i. e. in its electromagnetic alternating field, so long as the direction of its axis runs parallel with that of the main current collecting coil.

In producing undamped oscillations of a high frequency (50,000 per second and more) in the oscillation circuits 9 and 11, electromagnets S and S^1 must be inserted if the high frequency oscillations are not to penetrate the collector coil, between the oscillation producers and the collector coil. In all other cases they are connected shortly before the earthing (as in Figs. 27 and 28).

In Figure 27 a second method of construction of the connecting conductor of the balloon aerials is illustrated in the form of a coil. The main difference consists in that in addition to the connecting conductor I another annular conductor II is inserted parallel to the former on the high voltage masts in the air (or embedded as a cable in the earth) but both in the form of a coil. The connecting wire of the balloon aerials is indicated as a primary conductor and also as a current producing network; the other is the consumption network and is not in unipolar connection with the current producing network.

In Figure 27 the current producing network I is shown with three balloon collectors 1, 2, 3 and aerial conductors 4, 5, 6; it is short circuited through condenser 19 and inductance 9. The oscillation forming circuit consists in this diagram of spark gap *f*, inductance 10, and condenser 11; the earth wire E, is connected to earth over electromagnet S^1. F is the safety spark gap which is also connected to earth through a second electromagnet S at E. On connecting up the condenser circuit 11 this is charged over the spark gap *f* whereby an oscillatory discharge is formed. This discharging current acts through inductance 10 on the inductively coupled secondary 9, whereby in the producing network a modification of the potential of the condenser 19 is produced. The consequence of this is that oscillations arise in the coil shaped producer network. These oscillations induce a current in the secondary circuit II, which has a smaller number of windings and a less resistance, the voltage of which, according to the proportion of the number of windings and of the ohmic resistance, is considerably lower whilst the current strength is greater.

In order to convert the current thus obtained into current of an undamped character, and to tune its wave lengths, a sufficiently large regulatable capacity 20 is inserted between the ends 12 and 13 of the secondary conductor II. Here also current may be taken without an earth conductor, but it is advisable to insert a safety spark gap E^1 and to connect this with the earth over an electromagnet S^2.

The producer network may be connected with the working network II over an inductionless condenser 21 or over an induction condenser 22, 23. In this case the secondary conductor is unipolarly connected with the energy conductor.

In Figure 28 the connecting conductor between the separate accumulator balloons is carried out according to the autotransformer principle. The collecting coil connects four aerial balloons 1, 2, 3, 4, the windings of which are not made side by side but one above the other. In Figure 28 the collector coil I is shown with a thin line, the metallically connected prolongation coils II with a thick line. Between the ends I^1 and II^1 of the energy network I a regulating capacity 19 is inserted. The wire I^1 is connected with the output wire and with the spark gap F.

As transformer of the atmospheric electricity an arrangement is employed which consists in using rotary pairs of condensers in which the one stator surface B is connected with the main current, whilst the other A is connected with the earth pole. Between these pairs of short circuited condensers are caused to rotate from which the converted current can be taken by means of two collector rings and brushes, in the form of an alternating current, the frequency of which is dependent on the number of balloons and the revolutions of the rotor. As the alternating current formed in the rotor can act, in this improved method of connection described in this invention, through coils 10 on the inductance 9, an increase or diminution of the feed current in I can be obtained according to the direction of the current by back induction. Current oscillations of uniform rhythm thereby result in the coil shaped windings of the producer net-work.

As the ends of this conductor are short circuited through the regulatable condenser 19 these rhythms produce short circuited undamped oscillations in the energy conductor, the periodicity and wave lengths of which oscillations can be adjusted according to desire by altering the capacity 19 to a given wave length and therewith also to a given frequency. These currents may also be employed in this form directly as working current through the conductors II^1 and III. By inserting the condenser 20 a connection between these conductors may also be made, whereby harmonic oscillations of desired wave length are formed. By this means quite new effects as regards current distribution are obtained. The withdrawal of current can even take place without direct wire connection if, at a suitable point in the interior of the producing network (quite immaterially whether this has a diameter of 1 or 100 km.) a coil tuned to these wave lengths and of the desired capacity is firmly or movably mounted in the aerial conductor in such a way that its axial direction is in parallel with that of the collector coil. In this case a current is induced in the producing network, the size of which is dependent on the total capacity and resistance and also on the periodicity employed. A possibility is thereby afforded in future, of taking energy from the producer network by wireless means. As thereby in addition to atmospheric electricity also magnetic earth currents and the energy from the higher atmosphere (at least partially) may be simultaneously obtained, this last system for collecting the atmospheric energy is of particular importance for the future.

Of course everywhere instead of spark gaps suitable grid vacuum tubes may be employed as producers for undamped oscillations. The separate coils of the producer net-work with large diameters may be connected with one another through separate conductors all in parallel or all in series or in groups in series. By regulating the number of oscillations and also the extent of the voltage more or less large collector coils of this kind may be employed. The coils may also be divided spirally over the entire section. The coils may be carried out in annular form or also in triangular, quadrangular, hexagonal or octagonal form.

Of course wires may be carried from a suitable place to the centre or also laterally which serve the current waves as guides. This is necessary when the currents have to be conducted over mountains and valleys and so forth. In all these cases the current must be converted into a current of suitable periodicity.

As already hereinbefore mentioned separate collecting balloons may be directly metallically interconnected at equidistant stations distributed over the entire country or may be connected by interpolation of suitable condenser batteries by means of high voltage conductors. The static electricity is converted through a spark gap into dynamic energy of a high number of oscillations, and could then in such form, with a suitable arrangement of the connections, observing various measures of precaution, be employed as source of energy after separate or special regulation.

According to this invention in order to increase the collecting effect of the balloon in the aerial collector conductor or in the earth wire, radiating collectors are employed. These consist either of incandescent metal or oxide electrodes in the form of vacuum grid tubes, or electric arcs (mercury and the like electrodes) Nernst lamps, or finally flames of various kinds may be simply connected with the respective conductor.

It is well known that energy can be drawn off from a cathode consisting of an incandescent body opposite an anode charged with positive electricity (vacuum grid tube). Hitherto however, a cathode was always first directly placed opposite an anode, and secondly the system always consisted of a closed circuit.

Now if we dispense with the ordinary ideas in forming light or flame arcs in which a cathode must always stand directly opposite an anode, and if we place an incandescent cathode opposite an anode charged to a high potential or another body freely floating in the air, or regard the incandescent cathode only as a source of unipolar discharge (which represent group and point discharges in electro-static machines similar to unipolar discharges), it may be ascertained that incandescent cathodes and less perfectly all incandescent radiators, flames and the like admit of relatively large current densities and allow large quantities of electric energy to radiate into the open space in the form of electron streams as transmitters.

The object of this invention is as described below, if such incandescent oxide electrodes or other incandescent radiators or flames are not freely suspended in space but connected metallically with the earth so that they can be charged with negative terrestrial electricity, these radiators possess the property of absorbing the free positive electrical charges contained in the air space surrounding them (that is to say of collecting them and conducting them to earth). They can therefore, serve as collectors and have, in comparison to the action of the spikes, or points, a very large radius of action R; the effective capacity of these collector is much greater than the geometrical capacity (R_0–) calculated in an electro-static sense.

Now as our earth is surrounded as is well known, with an electro-static field and the difference of potential

$$\frac{\delta V}{\delta h}$$

of the earth field according to the latest investigations, is in summer about 60 to 100 volts and in winter 300 to 500 volts per metre of difference in height (δh), a simple calculation gives the result that when such a radiation collector or flame collector is arranged for example on the ground, and a second one is mounted vertically over it at a distance of 2,000 metres and both are connected by a conducting cable, there is a difference of potential in summer of about 2,000,000 volts and in winter even of 6,000,000 volts and more.

According to Stefan Boltzmann's law of radiation, the quantity of energy which an incandescent surface (temperature T) of 1 sq. cm. radiates in a unit of time into the open air (temperature T_o) is expressed by the following formula:

$$S=\sigma\ (T^4-T^4_O)\ \text{watt/sq. cm.}$$

and the universal radiation constant σ is according to the latest researches of Ferry (Annales de Chimie et de physique 17 page 267 (1909)) equal to 6.30×10^{-12} watt/sq. cm.

Now if an incandescent surface of 1 sq. cm. shows, as compared with the surrounding space a periodic fall of potential δV it radiates (independent of the current direction, that is to say of the sign) in accordance with the above formula, for example at a temperature of 3725° C. an energy of 1.6 kw. per sq. cm. per second. As for the radiation the same value can be calculated for the collection of energy, but reversed. Now as carbon electrodes at the temperature of the electric arc support on the current basis a current density up to from 60 to 65 amperes per sq. cm. no difficulties will result in this direction in employing radiating collectors as accumulators.

If the earth be regarded as a cosmically insulated condenser in the sense of geometrical electro-statics *x* there results from the geometric (compare Ewald Rasch, "das elektrische Bogenlicht" (The electric arc light) page 169) capacity of the earth according to Chwolson:

For negative charging 1.3×10^{6} Coulomb

For negative potential $V=10\times10^{8}$ volts.

From this there results however, $EJT\cong24.7\times10^{24}$ watt/Sec. Now if it is desired to make a theoretic short circuit through an earthed flame collector this would represent an electric total work of about 79,500 10^{10} kilowatt years. As the earth must be regarded as a rotating mechanism which is thermo-dynamically, electromagnetically, and also kinematically coupled with the sun and stars system by cosmic radiations and gravitation a diminution of the electric energy of the earth field is not to be feared. The energies which the incandescent collectors would withdraw from the earth field can only cause by the withdrawal of motor work a lowering of the earth temperature (temperature $T_E=300$) and reduce this to that of the world space ($T=O$) by using the entire energy. This is however not the case as the earth does not represent a cosmically entirely insulated system. On the contrary there is conveyed to the same according to the recent value corrected by Ferry for the solar constants through the radiation from the sun an energy of $18,500\times10^{10}$ kw. Accordingly any lowering of the earth temperature (T_E) without a simultaneous lowering of the sun's temperature (T_S) would contradict Stefan Boltzmann's law of radiation.

$$S=\sigma\,(T_S^4-T_E^4).$$

From this it must be concluded that if the earth temperature (T_E) sinks, the total radiation S absorbed by the earth increases, and further also that the secular speed of cooling of the earth is directly dependent on that of the sun and the other radiators cosmically coupled with the sun and is connected most closely with these.

The incandescent radiation collectors may, according to this invention, be employed for collecting atmospheric electricity if they (1) are charged with the negative earth electricity (that is to say when they are directly connected by means of a metallic conductor with the earth) and (2) if large capacities (metal surfaces) charged with electricity are mounted opposite them as positive poles in the air. This is regarded as the main feature of the present invention as without these inventive ideas it would not be possible to collect with an incandescent collector, sufficiently large quantities of the electrical charges contained in the atmosphere as technology requires; the radius of action of the flame collectors would also be too small, especially if it be considered that the very small surface density (energy density) (σ about$=2\times7\ .\ 10^{9}$ St. E. per sq. cm.) does not allow of large quantities of charge being absorbed from the atmosphere.

x) Calculated according to Poisson's calculation;

$\Delta V=-4\pi\delta$; as here the alteration of the potential or potential gradients only takes place in the direction of the normal, this calculation assumes the simple form

$$\delta=\frac{1}{4\pi}\times\frac{\delta^2V}{\delta n^2}$$

It has indeed already been proposed to employ flame collectors for collecting atmospheric electricity and it is known that their collecting effect is substantially greater opposite the points. It is however, not known that the quantities of current which could hitherto be obtained are too small for technical purposes. According to my experiments the reason for this is to be found in the too small capacities of the collector conductor poles. If such flame or radiating collectors have no or only small positive surfaces, their radius of action for large technical purposes is too small. If the incandescent collectors be constantly kept in movement in the air they may collect more according to the speed of the movement, but this is again not capable of being carried out in practice.

By this invention the collector effect is considerably increased by a body charged with a positive potential and of the best possible capacity being also held floating (without direct earth connection) opposite such an incandescent collector which is held floating in the air at a desired height. If for example, a collecting balloon of sheet metal or of metalized balloon fabric be caused to mount to 300 up to 3,000 metres in the air and as positive pole it is brought opposite such a radiating collector connected by a conductor to earth, quite different results are obtained.

The metallic balloon shell (with a large surface) is charged to a high potential by the atmospheric electricity. This potential is greater the higher the collecting balloon is above the incandescent collector. The positive electricity acts concentratedly on the anode floating in the air as it is attracted through the radiation shock ionization, proceeding from the incandescent cathode. The consequence of this is that the radius of action of the incandescent cathode collector is considerably increased and thereby also the collecting effect of the collecting balloon surface. Further the large capacity of the anode floating in the air plays therefore an important part because it allows of the taking of large charges, and thereby a more uniform current is obtained even when there is a large consumption: this cannot be the case with small surfaces.

In the present case the metallic collecting balloon is a positive anode floating in the air and the end of the earth conductor of this balloon serves as positive pole surface opposite the surface of the radiating incandescent cathode, which in turn is charged with negative earth electricity being conductingly connected to earth.

The process may be carried out by two such contacts (negative incandescent cathode and anode end of a capacity floating in the air) a condenser and an inductive resistance being switched on in parallel, whereby simultaneously undamped oscillations may be formed.

In very large installations it is advisable to connect two such radiating collectors in series. Thus an arc light incandescent cathode may be placed below on the open ground and an incandescent cathode which is heated by special electro-magnetic currents be located high in the air. Of course for this the special vacuum Liebig tubes with or without grids may also be employed. An ordinary arc lamp with oxide electrodes may be introduced on the ground and the positive pole is not directly connected with the collecting balloon, but through the upper incandescent cathode or over a condenser. The method of connecting the incandescent cathode floating in the air may be seen in Figs. 29–33.

B is the air balloon, K a Cardan ring (connection with the hawser) C the balloon, L a good conducting cable, P a positive pole, N negative incandescent cathode, and E earth conductor.

Fig. 29 represents the simplest form of construction. If electric oscillations are produced below on the ground by means of a carbon arc lamp or in other suitable way a considerably greater electric resistance is opposed to that in the direct way by inserting an electrical inductive resistance 9. Consequently between P and N, a voltage is formed, and as, over N and P only an inductionless ohmic resistance is present, a spark will spring over so long as the separate induction co-efficients and the like are correctly calculated. The consequence of this is that the oxide electrode (carbon or the like) is rendered incandescent and then shows as incandescent cathode an increased collecting effect. The positive poles must be substantially larger than the negative in order that they may not also become incandescent. As they are further connected with the large balloon area which has a large capacity and is charged at high voltage, an incandescent body which is held floating in the air and a positive pole which can collect large capacities is thereby obtained in the simplest way. The incandescent cathode is first caused to become incandescent by means of separate energy produced on the earth, and then maintained by the energy collected from the atmosphere.

Fig. 30 only shows the difference that instead of a round balloon a cigar shaped one (of metal or metalized fabric) may be employed and also a condenser 5 is inserted between the incandescent cathode and the earth conductor so that a short circuited oscillation circuit over P. N 5 and 9 is obtained. This has the advantage that quite small quantities of electricity cause the cathode to become incandescent and much larger cathode bodies may be rendered incandescent.

In this form of construction both the incandescent cathode and also the positive electrode may be enclosed in a vacuum chamber as may be seen in Fig. 32. A cable L is carried well insulated through the cover of a vessel and ends in a condenser disc 5. The cover is arched in order to keep off the rain. The vessel is entirely or partially made of magnetic metal and well insulated inside and outside. Opposite the disc 5 another disc 6 and on this again a metallic positive pole of the vacuum tube *g* with the incandescent cathode (oxide electrode) N is arranged. The negative electrode is on the one hand connected with the earth conductor E, and on the other hand with the inductive resistance 9 which is also connected with the cable L with the positive pole and wound round the vessel in coils. The action is exactly the same as that in Fig. 29 only instead of an open incandescent cathode one enclosed in vacuo is employed. As in such collectors only small bodies can be brought to incandescence in large installations a plurality of such vacuum tubes must be inserted in proximity to one another. According to the previous constructions Figs. 31 and 33 are quite self evident without further explanations.

Figs. 34–37 represent further diagrams of connections over radiating and flame collectors, and in fact, how they are to be arranged on the ground.

Fig. 34 shows an arc light collector with oxide electrodes for direct current and its connection; Fig. 35 a similar one for alternating current, Fig. 36 an incandescent collector with a Nernst lamp and Fig. 37 a similar one with a gas flame.

The positive pole 1 of the radiating collectors is always directly connected to the aerial collecting conductor A. In Fig. 34 this is further connected over the condenser battery 5 with a second positive electrode 3. The direct current dynamo *b* produces current which flows over between the electrodes 3 and 2 as an arc light. On the formation of an arc the negative incandescent electrode 2 absorbs electricity from the positive poles standing opposite it and highly charged with atmospheric electricity and conveys the same to the working circuit. The spark gap 7, inductive resistance 9 and induction coil 10 are like the ones previously described. The protective electromagnet S guards the installation against earth circuiting, the safety spark gap 8 from excess voltage or overcharging.

In Fig. 35 the connection is so far altered that the alternating current dynamo feeds the exciting coil 11 of the induction condenser. 12 is its negative and 13 its positive pole; if the coil 3 on the magnet core of the dynamo is correctly calculated and the periodicity of the alternating current is sufficiently high an arc light can be formed between the two poles 1 and 2. As the cathode 2 is connected with the negatively charged earth, and therefore always acts as a negative pole, a form of rectification of the alternating current produced by the dynamo 3 is obtained, the second half of the period is always suppressed. The working circuit may be carried out in the same way as in Fig. 34; the working spark gap 7 may however be dispensed with, and instead thereof between the points *n* and *m* a condenser 5 and an induction resistance 9 may be inserted from which the current is taken inductively.

Fig. 36 represents a form of construction similar to Fig. 34 only that here instead of an arc lamp a Nernst incandescent body is employed. The Nernst lamp is fed through the battery 3. The working section is connected with the negative pole, the safety spark gap with the + poles. The working spark gap 7 may also be dispensed with and the current for it taken at 12 over the oscillation circuit 5, 11 (shown in dotted lines).

Flame collectors (Fig. 37) may also be employed according to this invention. The wire network 1 is connected with the aerial collector conductor A and the burner with the earth. At the upper end of the latter, long points are provided which project into the flame. The positive electrode is connected with the negative over a condenser 5 and the induction coil 9 with the earth.

The novelty in this invention is firstly, the use of incandescent cathodes opposite positive poles which are connected with large metallic capacities as automatic collecting surfaces, (2) the connection of the incandescent cathodes with the earth whereby, in addition to the electricity conveyed to them from the battery or machine which causes the incandescing, also the negative charge of the earth potential is conveyed, and (3) the connection of the positive and negative poles of the radiating collectors over a condenser circuit alone or with the introduction of a suitable inductive resistance, whereby simultaneously an oscillatory oscillation circuit may be obtained. The collecting effect is by these methods quite considerably increased.

I declare that what I claim is:—

1. An electrical energy generating system, comprising a conducting surface for static charges, means to support same at a distance above the earth, a conductor leading to the earth level, a spark gap associated with said conductor to convert electrostatic charges into electromagnetic high frequency oscillations means to supply said electromagnetic energy to a net work, and a spark gap of greatly increased relative resistance in parallel therewith.

2. An electrical energy generating system

comprising a conductor, means to support same above the earth level, an inductance therein, a spark gap associated with said conductor, a second spark gap of much higher relative resistance in parallel therewith and an energy receiving circuit coupled with the spark gap of lesser resistance.

3. An electrical energy generating system comprising a collecting surface, means to support same above the earth level, a conductor connecting said collecting surface with the earth level, a choke in said conductor, an electromagnetic resistance converting electrostatic energy to electromagnetic energy, a safety higher resistance in parallel therewith and a net work coupled with the conversion resistance of lesser value.

4. An electrical energy generating system comprising electric conductors spaced above the earth to form electromagnetic oscillating circuits, conductors connecting to earth level, electrostatic to electromagnetic energy conversion means therein, a safety high electrostatic resistance in parallel therewith and means to alter the electromagnetic characteristic of the circuits.

5. An electrical energy generating system comprising in combination a static collecting surface arranged above the earth, conductors connecting to earth level, a pair of spark gaps in parallel of different electrostatic resistance, a utilization net work shunted across the spark gap of lesser resistance and an electromagnetic choke in said conductors.

6. An electrical energy generating system comprising an open circuit energy collecting aerial, a pair of sparking gaps in parallel of widely different resistance, connected thereto and a closed electric oscillation circuit in shunt across the gap of lesser resistance

7. An electrical energy generating system comprising an open circuit energy collecting aerial, a pair of sparking gaps in parallel of widely different resistance connected thereto, a closed electric oscillation circuit in shunt across the gap of lesser resistance, a plurality of electrostatic collecting surfaces, means to connect said collecting surfaces in parallel in groups and means to connect said groups symmetrically with said aerial.

In witness whereof, I have hereunto signed my name this 30 day of Dec., 1920, in the presence of two subscribing witnesses.

HERMANN PLAUSON.

Witnesses:
H. F. Armstrong,
W. H. Beeston.

Nu Energy Technologies

A Leading Source for Clean & Economical Energy...
powering our future with revolutionary radio ionic energy

INVENTION AND DISCOVERY

Radiant Energy Power Generation

Nuclear Science has been incorrect since the advent of the Atomic "chain reaction"

It would be logical to expect that the discovery of fission was the authoritative answer to nuclear science. It would be normal also to expect that no single individual could really claim to be its "inventor," since it is a natural physical process. The efforts of hundreds of dedicated scientists were required to bring it into existence. Countless descriptions on the origin of atomic decay have been theorized. Since the advent of the atomic chain reaction nuclear science has been written into scientific law. However, this was based on atomic-bomb era mentality and is not correct according to what the author of this book has unveiled.

Mr. Bruce A. Perreault knows the truth about nuclear reactions. He is a catalyst who will ultimately lead humankind to the successful development of radiant energy power generators.

By any normal standard this man is the discoverer of a revised nuclear standard that promises to open new roads to invention and discovery.

Bruce A. Perreault, P. O. Box 22, Rumney, New Hampshire 03266-0022 USA
Copyrighted © 1999, 2000 & 2001 All Rights Reserved

Website: http://www.nuenergy.org

ISBN 1-930216-01-7
52495
9 781930 216013

$24.95 USD Retail

advertisement

TAX REFORM
SHOULDN'T HURT
CALIFORNIANS
CALL CONGRESSM
JEFF DENHAM

EVERYTHING
A NEW
MOM NEEDS
Baby Depot
at Burlington
SHOP NOW

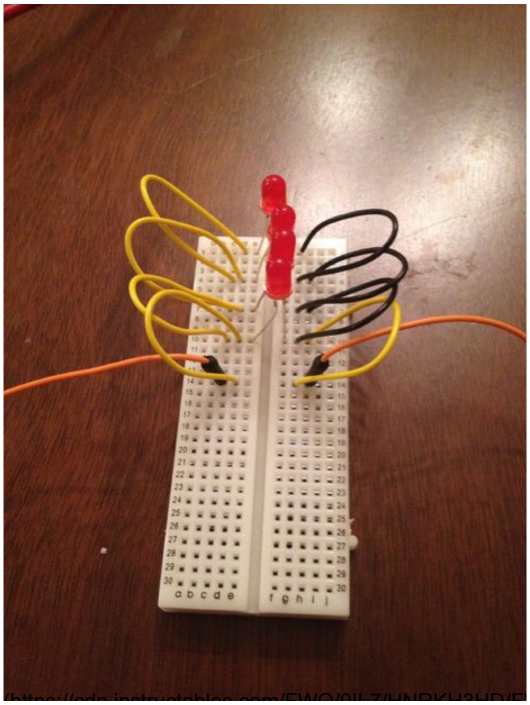

(https://cdn.instructables.com/FWQ/0IL7/HNRKH3HD/F

advertisement

ALLIED ELECTRONICS
Make a
Stronger
Connection
Avoid Failures with
Terminations That last
PHOENIX CONTACT
LEARN MORE

Step 1: Round Up Some LEDs!

About This Instructable

8,885 views

59 favorites

License:

BY-NC-SA

Free Energy From LEDs

by Throttle78 (/member/Throttle78/) in electronics (/technology/electronics/)

Download (/id/Free-Energy-From-LEDs/) 3 Steps + Collection I Made it! Favorite Share

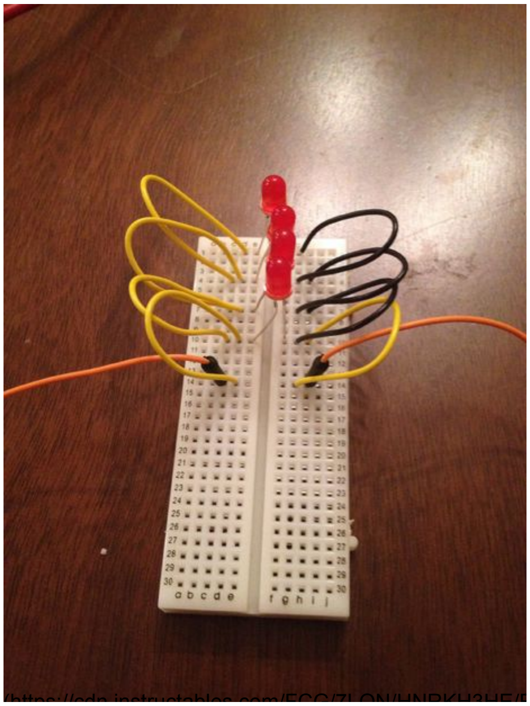
(https://cdn.instructables.com/FCG/7LON/HNRKH3HF/

First you'll need to get however many LEDs you want. You'll then want to connect the LEDs in a series, leaving a positive and negative wire.

NOTE: This works with any LEDs, but I have found that it works best with Red LEDs

Step 2: Go Towards the Light!

(https://cdn.instructables.com/FAY/6EWM/HNRKH3HC/

Then you'll want to get a light source (lamp or daylight) and put the circuit right in the strongest part of the light.

Step 3: It's As Simple As That!

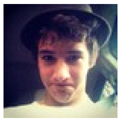

Throttle78 (/member/Throttle78/)

Follow 4

(/member/Throttle78/)

advertisement

EVERYTHING A NEW MOM NEEDS
Baby Depot
at Burlington
SHOP NOW

Just hook up the wires to a voltmeter and you'll see how much power they are generating. I was able to produce 1 volt just by using 3 LEDs and a bright lamp.

(As if you couldn't tell, this is my first ible, so some advice would be greatly appreciated)

Inspiration for this project and more information on how this works can be found at https://www.instructables.com/id/Better-LED-as-light-sensor

EVERYTHING A NEW MOM NEEDS
Baby Depot at Burlington
SHOP NOW

advertisement

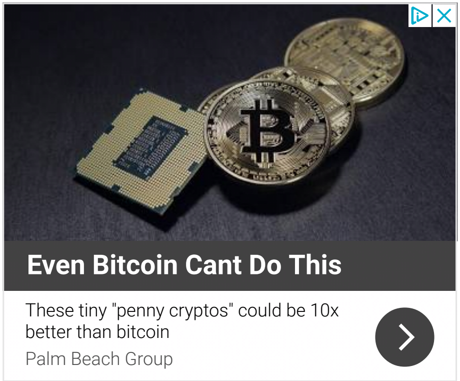
Even Bitcoin Cant Do This
These tiny "penny cryptos" could be 10x better than bitcoin
Palm Beach Group

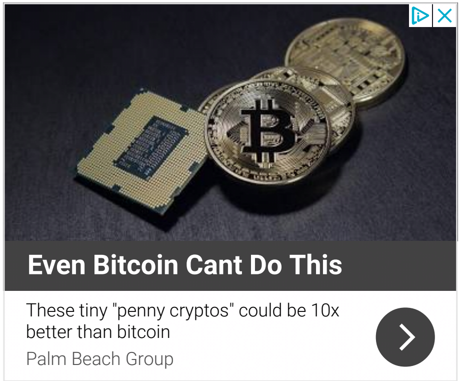
Even Bitcoin Cant Do This
These tiny "penny cryptos" could be 10x better than bitcoin
Palm Beach Group

Comments

be nice

We have a be nice comment policy.
Please be positive and constructive.

I Made it!
Add Images
Post Comment

---

uwezi (/member/uwezi/) 2015-05-20 Reply

An LED is a very poor solar cell in the same way that a solar cell is a very poor LED. Both are optimized for their primary function: to emit light OR to convert light ot electricity. Both work in the opposite way as well, but apart from studying the underlying physics there is nothing to be gained from using LEDs as solar cells.

Get a cheap solar-powered garden light for EUR 0.50 or an equivalent amount of money from your nearest hardware store and you get an LED and a useful solar cell module for your projects.

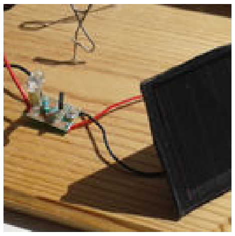

(https://cdn.instructables.com/FTO/W3XL/I9UE2ST3/FTOW3XLI9UE2ST3.LARGE.jpg)

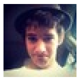

Throttle78 (/member/Throttle78/) 2013-11-10 Reply

I only have 4 LEDs, so I have not been able to generate enough current to power something.

assasinsareus (/member/assasinsareus/) ▸ Throttle78 (/member/Throttle78/) 2014-12-17 Reply

I'm afraid even with 4,000 LEDs you would struggle to actually power something. Voltage and Current are interdependent. A large voltage does not mean a large current. Power in terms of Watts is represented in terms of Voltage x Current. You would be much better off with actual solar cells. Using the LED as a light sensor is a good use for this phenomonon.

zeldekins (/member/zeldekins/) 2014-01-13 Reply

I think you can use powe LED (1W) to achive more energy!

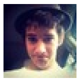

Throttle78 (/member/Throttle78/) 2013-11-11 Reply

Lean04, you do realize a simple schematic of this would be an LED?

lean04 (/member/lean04/) ▸ Throttle78 (/member/Throttle78/) 2013-12-04 Reply

since you´ve asked for advice i was suggesting a schematic so you have a more complete and friendly instructable, and it wouldn´t be a simple LED, because you´d be showing a lots of simple LEDs in series (in this particular case) for those DIYers that are new to electronics...

Bastl3r (/member/Bastl3r/) 2013-11-19 Reply

The LEDs are acually in reverse (anode on negative and cathode on positive connector). The current is there, but it is somewhere between nano- and microamperes.

The current that is generated could be used to indicate sunlight (or the absence of it) as shown in this schematic (http://www.dieelektronikerseite.de/Pics/Tricks/Dioden-Sperrstrom%20S01.GIF) , the D1 is the diode in the way you use it.

Your method of producing energy is barely enough to keep a microcontroller in low-power-mode, but not using it to make any program run.

With an array of 1000 LEDs in parallel you could produce about 1-10mA which would be nothing fancy, if you think of the costs compared to a little solar power module.

Most of the energy you put into the LEDs (the 100W lamp) is wasted, because of the less energy you get out
P = 1V * (ca. 1-10µA) = 1-10µW.

EVERYTHING
A NEW
MOM NEEDS
Baby Depot
at Burlington
SHOP NOW

Even if it would be more current (for example 10µA per LED) the output of power would still be around 30-50µA with 3-5 LEDs

I would suggest a net of lemons to build a lemon battery or some other chemical ways (vinegar and copper) which produce more power than the LEDs in reverse.

EVERYTHING A NEW MOM NEEDS

Baby Depot at Burlington

SHOP NOW

---

lean04 (/member/lean04/) 2013-11-10 Reply

Maybe placing some schematics would help and measuring the current too

---

andrea biffi (/member/andrea+biffi/) 2013-11-10 Reply

that's interesting, but maybe a picture of the multimeter with output voltage will explain better the goal..

Kiteman (/member/Kiteman/) 2013-11-10 Reply

So, you're using LEDs as tiny solar cells?

Have you powered anything with them yet?

Kimerserd (/member/Kimerserd/) 2013-11-09 Reply

U go my robot bro!

Throttle78 (/member/Throttle78/) 2013-11-09 Reply

I was unable to detect any amps with only 3 LEDs.

---

mman1506 (/member/mman1506/) 2013-11-09 Reply

now how much current can they supply?

---

mettaurlover (/member/mettaurlover/) ▸ mman1506 (/member/mman1506/) Reply
2013-11-09

Nowhere near enough to be useful, unfortunately, since they aren't designed to do this. A similarly-priced solar panel typically outputs far, far more voltage and current than a bunch of LEDs can, and isn't likely to decide that it wants to emit light instead of generating power if there's any surplus power in the system, such as a backup battery.

More Comments

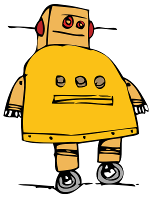

Newsletter

Let your inbox help you discover our best projects, classes, and contests. Instructables will help you learn how to make anything!

enter email | I'm in!

About Us

Who We Are (/about/)

Advertise (/advertise/)

Contact (/about/contact.jsp)

Jobs (/community/Positions-available-at-Instructables/)

Help (/id/how-to-write-a-great-instructable/)

Find Us

Facebook (http://www.facebook.com/instructables)

Youtube (http://www.youtube.com/user/instructablestv)

Twitter (http://www.twitter.com/instructables)

Pinterest (http://www.pinterest.com/instructables)

Google+ (https://plus.google.com/+instructables)

Resources

For Teachers (/teachers/)

Residency Program (/pier9residency)

Gift Premium Account (/account/give?sourcea=footer)

Forums (/community/?categoryGroup=all&category=all)

Answers (/tag/type-question/?sort=RECENT)

Sitemap (/sitemap/)

EVERYTHING A NEW MOM NEEDS
Baby Depot
at Burlington
SHOP NOW

Terms of Service (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21959721) |
Privacy Statement (http://usa.autodesk.com/adsk/servlet/item?siteID=123112&id=21292079) |
Legal Notices & Trademarks (http://usa.autodesk.com/legal-notices-trademarks/) | Mobile Site (https://www.instructables.com)
(http://usa.autodesk.com/adsk/servlet/pc/index?id=20781545&siteID=123112)

© 2017 Autodesk, Inc.

Visite 109 675
Pagine 245 310

MAREA SISTEMI

HOME

YOU ALSO PARTICIPATE IN THE SOCIAL NETWORK OF MAREASISTEMI

GOOGLE +

FACEBOOK CONTACT

facebook
Giuseppe Mazzone
SISTEMI

FACEBOOK PAGE

facebook
mareasistemi.com

YOUTUBE

Electricity from the earth's atmosphere. An old forgotten idea

Source: http://scienzaearte.blogspot.it/2011/03/energia-elettrica-dallatmosfera.html

Solar, wind, hydroelectric, geothermal, biogas. These are the main renewable and non-polluting energy sources. But
sure not to leave something out?
It is known that the Earth's atmosphere has a potential difference that increases with height. For non-professionals it
the atmosphere was a charged battery. This is evident during thunderstorms, when this charge finds its outlet in th
frightening and fascinating phenomena that are lightning. Many have imagined and studied methods for storing ligh
energy, but this presents numerous problems:

1) it is not known in advance where a lightning will strike or in any case where a thunderstorm will take place; 2) ev
lightning strike could be intercepted, for example by attracting it with a kind of lightning rod, it would be extreme
difficult to store all its energy, since the discharge is very short; 3) moreover, most of the lightning energy dissipate
light, heat and sound. In short, a "power plant" based on the interception of lightning would not be very profitable. **W**
less known, however, is that the atmosphere has this charge even when there is not even a cloud in the sk

. The idea of deriving electricity directly from the earth's atmosphere dates back more than a century ago. To m
knowledge - I urge readers to post comments if they are aware of other studies - the latest and most promising study
field was carried out by the Estonian engineer and inventor **Hermann Plauson** . His studies took place in the 1920
almost a century ago. Is it possible that since then there has been no progress in this field? Why did the idea of extra
electricity from the Earth's atmosphere go into oblivion, even though it seemed promising at the time of Plauson's stu

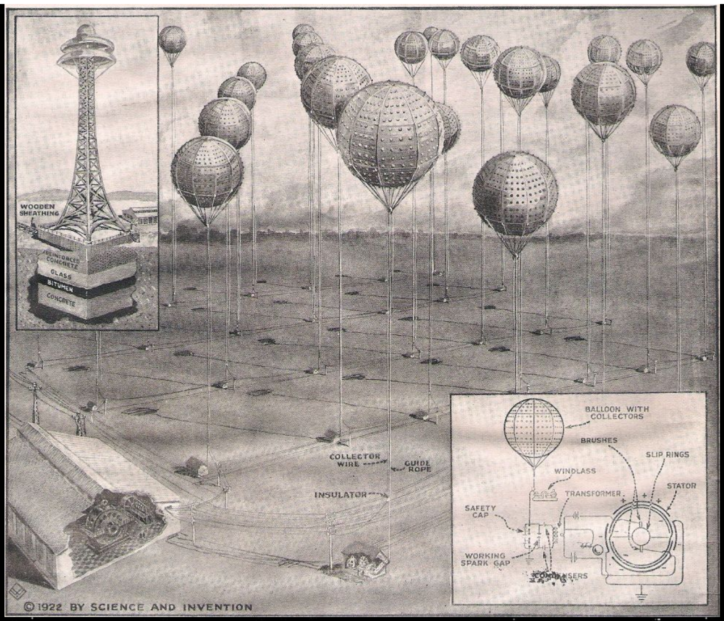

Image taken from the magazine "Science and Invention" (1922), where we describe the invention of Hermann Plau
Let's see before giving a brief description of the method used by Plauson. An electricity collector, consisting for exan
a balloon filled with helium, was raised to an altitude of several hundred meters, connected to the ground by a condu
cable. The balloon itself was made of conductive material and was covered with innumerable little needles, capabl
collecting atmospheric electricity through a phenomenon called field effect emission . High-voltage electricity w
conveyed to the ground by the conductor cable, and was converted into low-voltage electricity to be used by comn
electric users.

In the description of his invention , Plauson reports that he managed to get about 3.4 kW (kilowatt) in a "pilot" exper
using two of these balloons.
Returning to the first question: why such a promising technology was abandoned?
Several hypotheses can be made. The Plauson method provided - even if it was not bound - the use of radioactive ma
(radio and / or polonium), whose ionizing properties were used as a "catalyst" to increase the current inflow. It is ob
that today any alternative energy proposal based on the use of radioactive material would not be frowned upon,
considering the possible environmental impact that could result from accidental contamination. As a note
it is interesting *to mention* that until the eighties a type of lightning conductor covered with a thin layer of americiu
radioactive material), similar in concept to the Plauson collectors, was used in Italy: it is useless to say that these de
are currently become outlaws.
Another possible reason why the idea was abandoned may be related to the fact that the energy needs in the 20s wer
lower than today's ones: a typical "central" "at Plauson", composed of 100 separate balls the one from the other of a
100 meters (so 1 km of side), would have generated some hundreds of kilowatt: a power of all respect for those time
that pales a bit 'when compared to the energy produced by a single wind generator - about 1 Megawatt.
This reason, however, would justify only part of the abandonment of research in this field, since a system that allows
obtain even a few kilowatts using just a couple of balls could be competitive - compared to wind energy - in areas
windy and not served by the electricity network. **Let's** add at this point **an advantage that this system would hav**
wind and solar: uninterrupted production 24 hours a day and 365 days a year .
Apparently, therefore, the only major obstacle would seem to be represented by the use of radioactive material. And
we are meeting the recent technological development. In recent years, in fact, techniques have been developed to o
nanomaterials that are very efficient from the point of view of field effect emission: in practice they are super-pins
could function as effective collectors of atmospheric electricity without resorting to use of radioactive material
Some might argue, and rightly so, "But are these particular nanomaterials safe from the health point of view?"
Unfortunately, since these are new technologies, it is still difficult to estimate the impact in this sense, but some studi
under way. If, as is hoped, these materials will be considered harmless to health, then it can not be excluded that t
research on electric energy obtained directly from the earth's atmosphere will find a new vigor and this source not
exploited can go alongside the best known renewable and non-polluting sources.
As a final note it should be added that even without using radioactive material and the nanomaterials mentioned abo

system similar to that of Plauson could provide modest but useful quantities of electricity, finding application especi
the areas of developing countries far from large centers inhabited.

Hermann PLAUSON

Conversion of Atmospheric Electricity

Source: Source: rexresearch.com

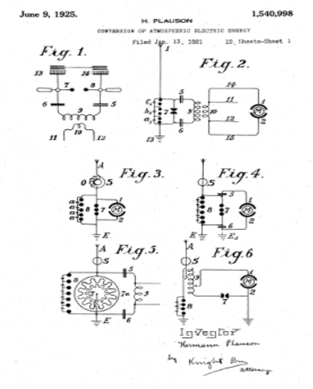

In 1920 the German engineer Hermann Plauson published a text "Gew und Verwertung der Atmospharischen Elektrizitat" (Acquisition and atmospheric electricity).

Where he concluded that atmospheric electricity could be captur exploited by men.

In this regard he wrote:

"Humanity will be free from worry about the threat of freezing death, knowing that in a pred time it will be possible to exploit the natural resources to the fullest ... Humanity will no longer suffer from the cold, because Mother Nature will give her abundant gift power."

Plauson designed a series of helium-filled balloons, able to collect atmospheric electricity in t through the use of thermionic rectifiers, Leida bottle condensers and induction coils, his idea supply energy to whole Germany.

Biography (Wikipedia)

Meridian International Research: Atmospheric Electricity Research

***Science & Invention* (Feb. 1922) : "Power from the Air" (I)**

***Science & Invention* (March 1922) : "Power from the Air" (II)**

Plauson's Patents (List)

H. Plauson: USP # 1,540,998 - Conversion of Atmospheric Electricity

H. Plauson: British Patent # 157,262 - Improvements in Electric Motors

H. Plauson: British Patent # 157,263 - Process & Apparatus for Converting Static Atmospheric Electrical Energy into Dynamic Electrical Energ

***Science & Invention* (June 1928) - "Harnessing Nature's Electricity"**

H. Plauson: British Patent # 299735 - Rapidly Moving Electron Process for Producing [PDF]

H. Plauson: *Gewinnung und Verwertung der Atmosphatischen Elecktrizitat* (1922) [PDF]

---

http://en.wikipedia.org/wiki/Hermann_Plauson

Biography

Hermann Plauson was an Estonian engineer and inventor. Plauson investigated the production of energy and power via atmospheric electricity.

Plauson was the director of the Fischer-Tropsch "Otto Traun Research Laboratories" in Hamburg, Germany during the Weimar Republic of the 1920s.
Nikola Tesla's idea for connecting machinery to the "wheelwork of nature". Plauson's US Patent # 1,540,998 methods of converting alternating radiar
electricity into rectified continuous current pulses. He developed the Plauson's converter, an electrostatic generator. In 1920, Plauson published a b
Production and Utilization of the Atmospheric Electricity " (*Gr., Gewinnung und Verwertung der Atmospharischen Elektrizitat*). A copy of this boo
British Library.

It is believed to be Gertrud Plauson (the exact relationship is unknown;

"Power from the Air". *Science and Invention* , Feb. 1922, no. 10. Vol IX, Whole No. 106. New York. (nuenergy.org)
"Power from the Air". *Science and Invention* , March 1922.

***Science and Invention*, Vol. IX (106) # 10 (February 1922)**

Power from the Air (I)

by

Hugo Gernsback

In the war there was developed in Germany in the new art --- or science --- that bids fair to revolutionize our present means of obtaining power.

This art, which is as new now as wireless, is 25 years ago, will be attainments during the next 25 years that may appear fantastic today. Hermann Plauso
of the new science, has devoted years of labor to his researches and he has now been using small power plants, generating electricity direct from the a
night, without interruption at practically no cost, once the plant is constructed.
We have occasion, to describe the system, roughly, from cabled dispatches, but complete information is available now. The amount of electrical power
in our atmosphere is astounding. Herr Plauson found in his experiments a single balloon sent to a height of 300 yards gave a constant current at 400
amperes, or in 24 hours over 17-1 / 4 kilowatts! By using two balloons in connection with a special condenser battery, the power obtained was 81-1 / 2
24 hours. The actual current delivered was 6.8 amperes at 500 volts.

The best balloons used by the inventor are made of thin aluminum leaf. No fabric was used. A simple internal system of ribs, stays and wires, gives
rigidity as well as a certain amount of elasticity. The balloon, when made airtight, is filled with hydrogen or better, with helium. It will then stay aloft fo
time. The outer surface is dotted with extremely sharp pins, made sharp electrolytically. Ordinary pins did not prove good current collectors, as they lack
sharpness. The pins themselves were made from amalgamated zinc, containing a radium preparation, in order to ionize the air. It was also found that by
outer surface of the balloon with zinc-amalgam more current could be collected. Even better results were obtained with polonium amalgam. Plauson sta
function of these amalgams is purely photoelectric.

One hundred of such captive balloons, separated one hundred yards from each other, will give a steady yield of 200 horsepower. This is the minimum,
the winter this figure increases up to 400 horsepower, due to the higher electrification of the atmosphere.

We need not go into the technic of how the current is finally made useable for industrial purposes, suffice it to say that the problem has been entirely sol
Plauson. By using batteries of condensers, high tension transformers, etc., the current can be transformed to any form desires. Such as for lighting lam
motors, charging storage batteries, etc.

Plauson also invented a sort of electrostatic rotary transformer which gives alternating current without the use of condensers and transformers. Indeed, i
very great, as it actually 'sucks' the current down rapidly from the collector balloons. There is no doubt that this invention will soon come into universal
the world. We will see the land dotted with captive balloons, particularly in the country and wherever water power does not abound. Indeed, the time is
when nearly all of our power will be derived from the atmosphere. So far it seems to be the cheapest form of power known, it being much cheaper even
power --- the cheapest form of power known today. Not only that, but as the inventor points out, no devastating thunder storms occur near such aerial p
because the balloons act not only as lightning arresters, but they quickly discharge the biggest thunder cloud, safely and noiselessly through their grou
gaps.

***Science & Invention* (March 1922), page 1006, 1007**

Power from the Air (II)

by

Hugo Gernsback

[For many years electrical engineers have endeavored to devise some means whereby it would become possible to utilize the free electrical energy eve
the atmosphere, but they were not successful, as every now and then an extra heavy surge of static current would rush down the elevated conductor a
the lives of the experimenters, or else destroy the apparatus connected with it. A German engineer has, however, devised the somewhat elaborate s
shown in brief, and he has succeeded, at least so his report states, in safely extracting several kilowatts of electrical power from the atmosphere w
surfaced balloons, elevated to a height of only 1000 feet.]

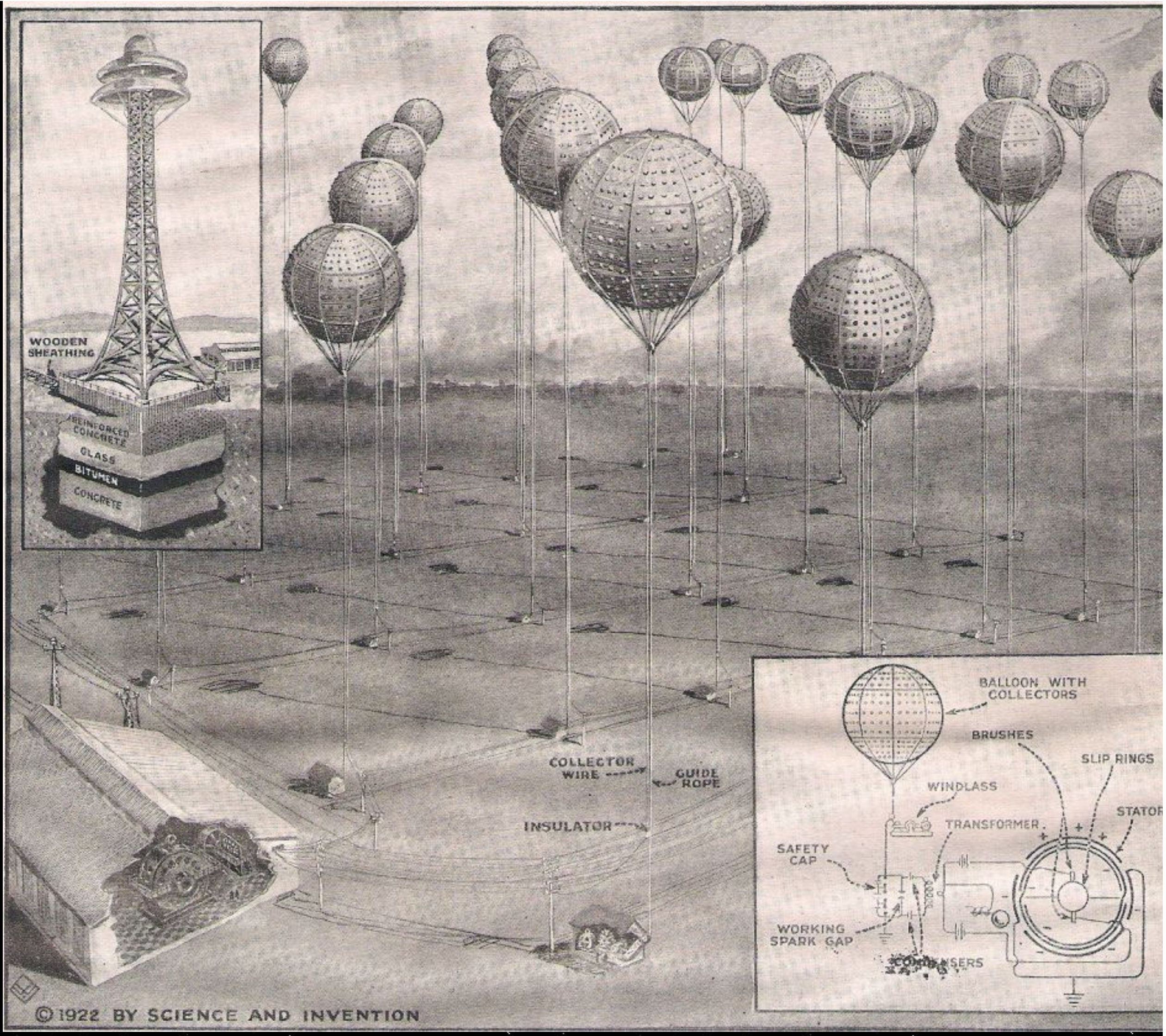

© 1922 BY SCIENCE AND INVENTION

We have previously treated of the extraction of electrical energy from the atmosphere. The difference of electric potential in different parts of the atmo
the difference between the upper air and earth make it a tempting proposition to obtain power from atmospheric electricity. The power would take the fo
potential difference with a discharge almost of a static nature. It has long appeared rather doubtful to conservative engineers, if such a source of power sh
be available. Yet when we see the lightning flash, it certainly suggests very high power, even though the total of its energy may be small, on account o
duration of the discharge It is not to the thunder storm that we look for getting power from the atmosphere, as the subject is now being seriously inve
German scientist, Hermann Plauson, has published a very elaborate work on this subject, and has investigated the use of kites, balloons and tow
utilization of the high potentials existing in the air at different altitudes, and has studied out the construction of motors to be operated by the pecu
discharge which will be obtained, if the projects are successfully carried out.

We will first speak of the methods used for collecting electricity from the upper air. The author cites several German patents. One of them shows the u
balloon. The balloon is shown floating in the air, kite fashion, and from it hangs a great net or aerial for the collection of electricity. The conductor fro
leads to the ground station; quite an elaborate description is given of the net-work which the patentee proposes to have covered with needle points. A wi
in or pays out cable for the balloon, and the patentee claims that by sending the apparatus to a height of about one mile he will have 225,000 volts to dra
then speaks o a battery of 20,000 cells in series, which will use up to 40,000 to 50,000 volts in the charging. This certainly provides for a reasonable
potential.

But our author discards this idea and first suggests something more permanent. He proposes the erection of towers to be in the neighborhood of 1,000 f
about the height of the Eiffel Tower. At the summit he has his collecting aerial. The appliance consists of a number of copper tubes; within each one he
burn gas lamps, whose products of combustion will reach the aerial, a collecting net-work covering the tops of the tubes. One of his apprehensions is
should wet his connections trouble might ensue, so he proposes a protection at the top in the shape of a great bell-like shield, resembling in his terms
pagoda". He also compares the form of the protection to that of a great petticoat insulator. Another of his difficulties is that he must have his tower ins
the earth. He, therefore describes a complicated foundation for his structure. He proposes first to pour in at the bottom of the excavation a foundatio
concrete. On this he places a layer of asphalt, and then a layer of cast glass, three to ten feet thick, and then comes a reinforced concrete foundation, t
metallic foot of the tower is to be anchored. This foundation must rise at least seven feet above the ground level, and is to be boarded in on all sides
from moisture. The author's idea s to erect a number of these towers connected by a horizontal cable, to which the aerials for collection of potentials are

The author strongly advocates balloons as collectors of the electric power of the air. These he depicts covered with spots. These spots indicate areas to b
coated and prepared to collect potential from the atmosphere.

In the first place he describes the balloon as made of thin metallic leaf supported by internal ribs. Steel wires silver-plated, copper-plated, or aluminum-
from the balloon to the pendant or junction ring. To this ring the tether cable is attached and runs to an insulated windlass on the surface of the earth. Th
to rise to an altitude varying from 300 feet to three miles.

The coating of the spots is to be of the thinnest amalgam, of mercury and gold, or zinc, or even polonium, perhaps only 1/2500 inch thick. All over the u the balloon are numberless metal points. To prepare the needle-like wires, they are collected into bundles and are treated electrolytically in a bath, dissolved in part. This gives a sharp point and roughened surface, all adapted for collecting the electric energy. The points may be of copper, steel, or metallic alloy. After this corrosion. As it may be termed, the wires are plated with gold or other of the so-called noble metals. It is advised that poloniu salts be added to the plating bath.

Dr Plauson devotes many pages of his book to describing his motor. This is a rotary motor including a stator and rotor and its peculiarity is that it contai develops no electromagnetic field properly speaking, but works by static excitation. One typical arrangement is shown in our illustration. The stator plat plates are concentric with each other, representing segments of cylinders. The alternation of negative and positive charged plates produces the rota connections there is included a safety spark gap to take care of dangerous potentials. Inductances and capacities are also used and indicated. It was fo plates heated, owing to the Foucalt currents, and to overcome this, several methods of subdividing the stator and rotor plates, are described by the author.

The whole subject is quite captivating, and it really seems as if the utilization of the electricity of the air may be almost in sight. It would seem possible experiments in this direction by means of the Eiffel Tower, but of course, the trouble here is that the tower is grounded, and perfect insulation of th surface is absolutely essential.

And now our author gives us some practical details. He says that on the Finland plains he carried out experiments with a balloon made of aluminu collecting needles of amalgamated zinc with a radium preparation as an ionizer. The surface of the balloon was sprinkled over with zinc amalgam. It wa a height of 300 meters, early 1,000 feet, and was held by a copper-plated steel wire. A constant current of 1.8 amperes at an average of 400 volts potentia was obtained. This gave nearly three-quarters of a kilowatt, or close to one horsepower. The collector of the balloon insulated from the earth showed 42,000 volts. By sending up a second balloon with an antenna to the same height at a distance of 100 meters from the first balloon, a current of over 3 a obtained. Then by putting into the circuit a large condenser, whose capacity was equal to the surface capacity of both balloons, and of the antenna conn current rose to 6.8 amperes with about 500 volts mean tension. By the use of these two balloons, he eventually ran up the power to 3.4 kilowatts.

Science & Invention (June 1928)

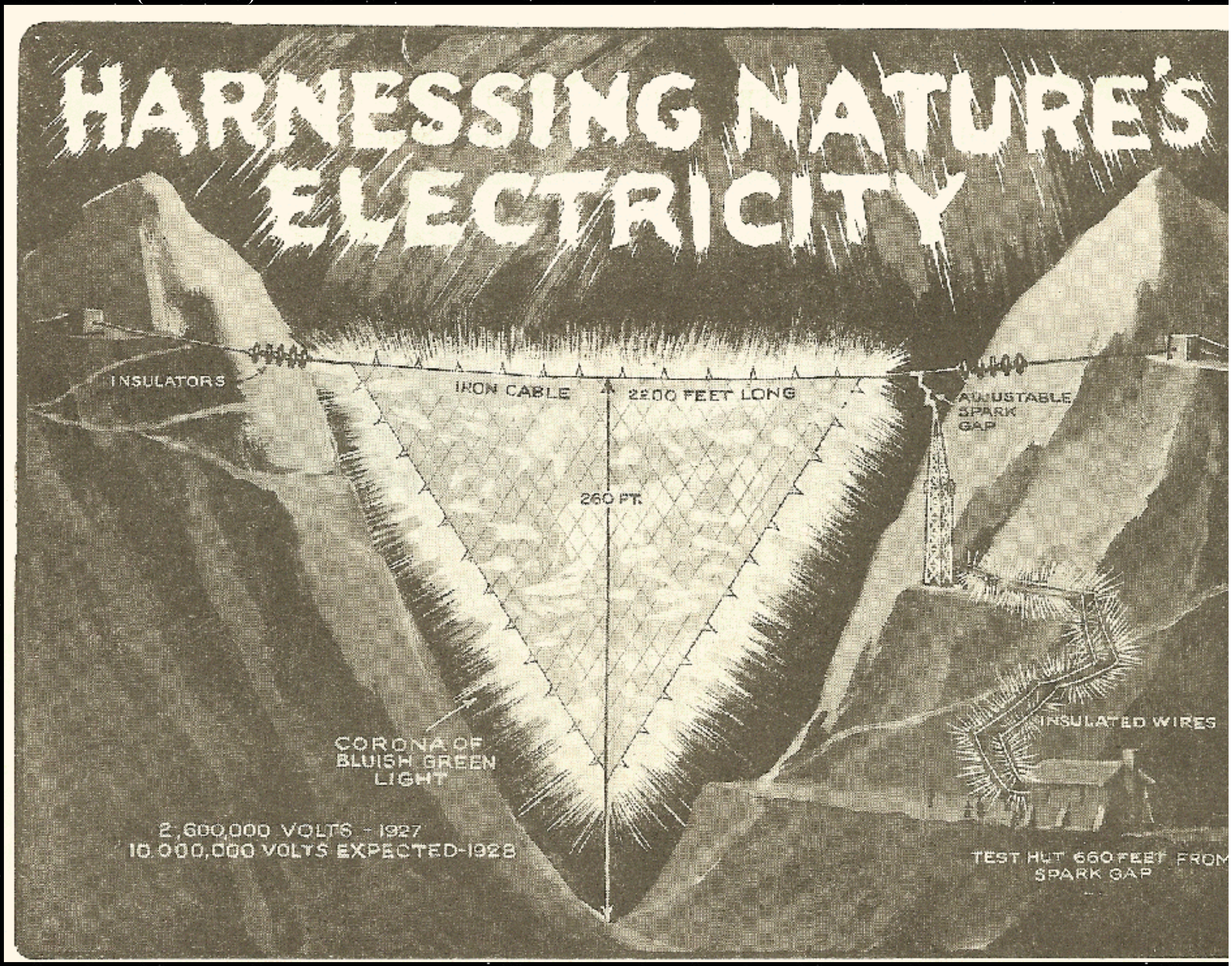
HARNESSING NATURE'S ELECTRICITY
INSULATORS
IRON CABLE
2200 FEET LONG
ADJUSTABLE SPARK GAP
260 FT.
INSULATED WIRES
CORONA OF BLUISH GREEN LIGHT
2,600,000 VOLTS - 1927
10,000,000 VOLTS EXPECTED-1928
TEST HUT 660 FEET FROM SPARK GAP

Remarkable European Experiments with Atmospheric Electrical Discharges with Potentials as High as 3,000,000 Volts

By HENRY TOWNSEND

elevation of 350 ft., and these students of natural electrical phenomena have found a very desirable location in the Alps, where they can suspend between one mountain and an adjacent one, a strong iron cable having a length of about 2,000 ft. This cable is about 250 feet above the intervening valley, and from it these daring engineers have suspended a coarsely woven wire net, which serves as an electrical capacity to gather the electricity from the atmosphere. As shown in the pictures, the wire net is supplied with numerous sharp points to aid in collecting the current from the air.

As the accompanying photographs of the actual apparatus and wire cable used last year clearly show, an adjustable spark gap of considerable length is provided. By adjusting this spark gap to various lengths, it is possible to judge the voltage of the discharge which leaps the gap at any moment. Mr. F. W. Peek, Jr., the well-known American worker in the realm of high voltage measurements, together with other engineers, have provided tabulated data and curves for various lengths of both needle and sphere type spark gaps. As one of the accompanying diagrams shows, it is a simple matter to calculate the voltage when a certain length of gap is used. The engineer first checks the length of the gap on the chart; he then follows a line horizontally from the gap length, to where it intersects with the angular line on the chart; and from the point of intersection he looks in a visual line downward to a place where the voltage is given. For needle spark gap measurements, the characteristic curve on the chart is practically a straight line, while for sphere gaps the characteristic curve on the voltage versus gap length, is a curved line. Those interested in high voltage measurements by means of the spark gap method can find the voltage-gap tables and charts in the *Standardization Rules* of the American Institute of Electrical Engineers. According to Mr. Peek's researches, the voltage per foot of atmospheric electrical discharges is about 100,000, while in laboratory measurements with A.C., transformer high potential discharges, the average voltage per foot of spark was found to be about 150,000 volts. The voltage of a lightning flash may

(*Continued on page* 156)

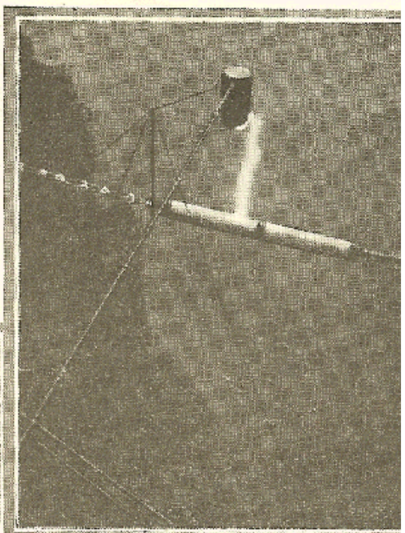

↑ Actual photo at
shows 13 ft., he
spark obtained f
the collecting net
the Alps by the
man scientists.
voltage is about 2,
000. The spark occu
once per second fo
minutes.

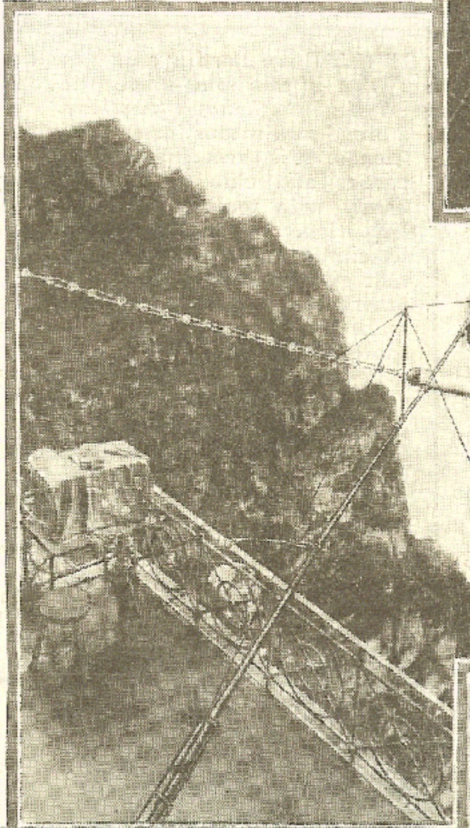

Photo, left, shows
adjustable spark
used in the Alps.
tice the heavy e
trode on the end of
adjustable arm to wh
the spark jumps

←

Below we see 3,000
volt artificial light
stroke produced in
E. Laboratory at P
field, Mass. Note n ↓

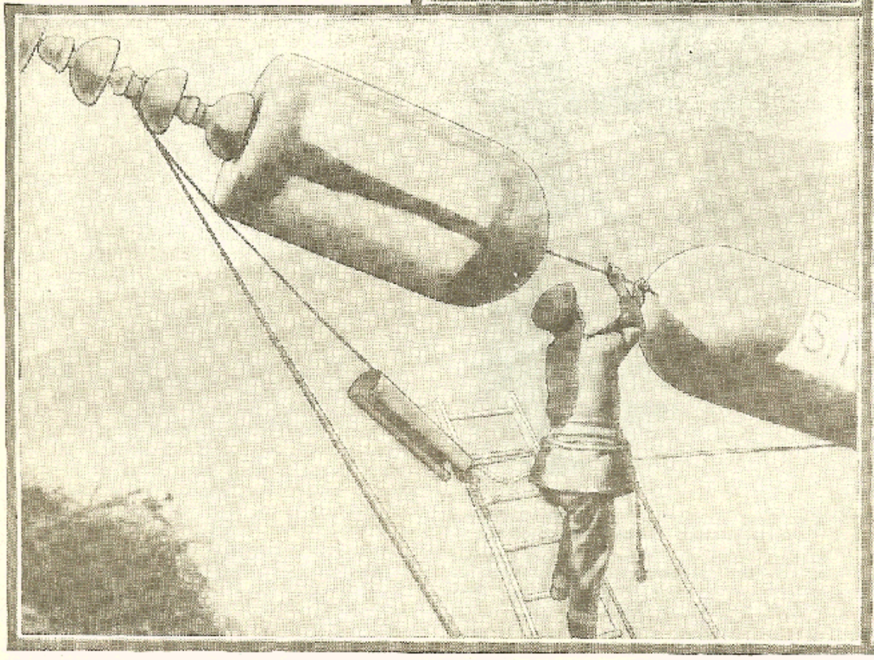

Actual photograph of the experimental "kite" used by the German experimenters in the Alps Mountains, for the purpose of accumulating high potential electrical discharges from the atmosphere. Note the size of the insulators.

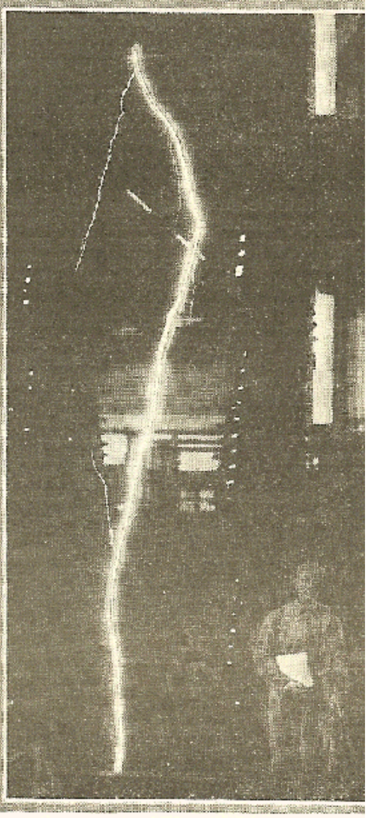

if-

you pledge
yourself to secrecy we will teach you the secrets of
Real Professional
Magic

Dr. Harlan Tarbell, internationally known magician and teacher of some of the world's greatest magicians.

Learn to be a Magician

For the first time in the history of Magic the age-old, sacredly guarded secrets of the Mystic Art are being revealed. Now at last you can learn to be a Real Professional Magician. Yes, and you can learn this wonderful art *easily and quickly* AT HOME. You are taught the mysterious *Principles* by which the Professional Magician works. You are taught everything from wonderful impromptu effects to massive stage illusions. But before you can study, you must sign the Magician's Solemn Pledge of Secrecy. This means that you are getting the *guarded secrets* of the Magic Profession. Think of that!

Earn $250 to $1000 a Month
There's a big demand for Magic entertainment. Magicians get big money. Dr. Tarbell gets as high as $250 for a half hour's work. Unlimited opportunities for you!

Low Prices Easy Terms

Write Today Send today for free Magic book, low prices and easy terms. Don't delay. Mail the coupon NOW!

Tarbell System, Inc., Studio A-142
1920 Sunnyside Avenue, Chicago, Ill.

Gentlemen: Without any obligation send me your free literature and information all about your wonderful "Tarbell Course in Magic." Also tell me about your Low Price Offer and Easy Payment Plan.

Name..

Address..

.... *Age*..........

HARNESSING NATURE'S ELECTRICITY

By HENRY TOWNSEND

(*Continued from page* 111)

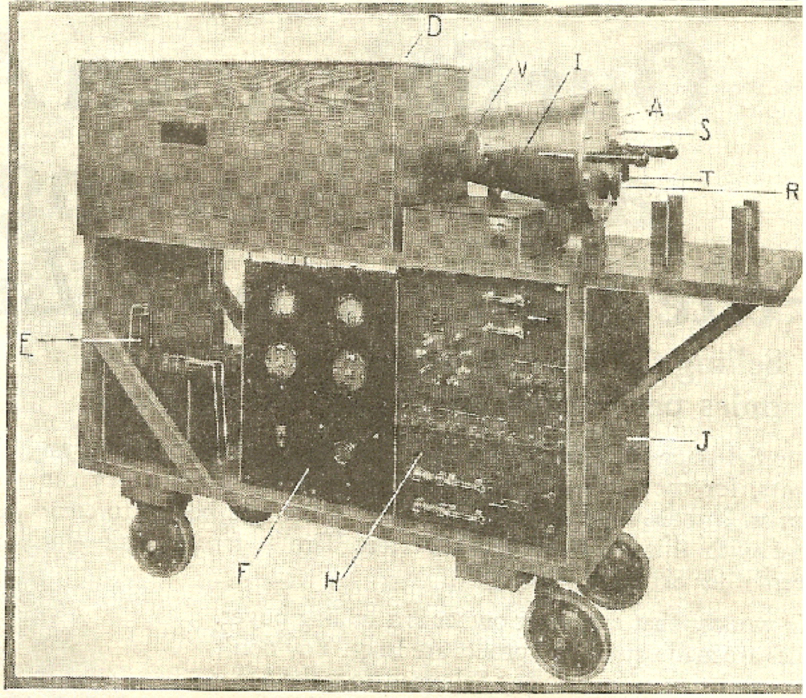

Remarkable cathode ray oscillograph used by Mr. F. W. Peek, Jr., in causing lightning to write its autograph with a pencil of electrons on a photographic film.

easily be 100,000,000 volts, as Mr. Peek has pointed out in one of his scientific papers.

WHY THE EXPERIMENTS ARE BEING MADE

ONE of the main reasons why these dangerous experiments are being carried on by the three young German scientists, whose names we have already learned, is because science believes that with a sufficiently high voltage, it will be possible for man to disintegrate the atom, and in this way make available a tremendous source of power as yet untapped. These experts have calculated that they will be able to obtain electrical energy in sufficient quantity from one of these powerful atmospheric discharges, to equal the Alpha rays obtained from 220 pounds of radium. As we have mentioned before, these experiments are of course fraught with great danger, and for that reason the experimenters seek refuge in a special lightning-proof hut, which is located about 600 ft., from the spark gap. When electrical storms are in the vicinity, it is especially important that the scientists keep within their protected fortress, for otherwise they would very probably be killed by a stray electrical discharge.

One of the peculiar things about this whole line of experiment is that the average layman does not realize perhaps, that there is a high electrical stress in the atmosphere on clear days, as well as when thunder storms are overhead. This fact has been known for a hundred years and more, and many years ago measurements of the various electrical potentials at increasing altitudes, were observed and measured by scientific investigators. There are a number of different ways in which these high electrical potentials found in the atmosphere can be measured; one of these methods involves the use of a calibrated spark gap. In this case the gap is set to a predetermined length, and when a discharge jumps this gap, the engineers know of course from previous experience and measurements, just what voltage is present. Another method of measuring extra high potentials, such as here
countered, requires the use of a static v
meter, which involves the use of a station
and of a movable or rotary set of m
plates, forming a condenser, to which
indicating needle is attached. For volta
above 2,000, static voltmeters have been
in a great many American central stati
and they have many desirable and us
characteristics. Of course as the vol
to be measured increases, the space betw
the quadrant shaped stationary and mov
plates is increased and vice versa.
electrostatic field from voltages below 2
is not sufficient to warrant the use o
static voltmeter. Another method of m
uring high potentials involves the use
the so-called vacuum tube voltmeter.

The general characteristics of the
mospheric electrical discharges, inclu
lightning, have been measured and reco
by one of the newest scientific instrum
known as the cathode ray oscillograph.
means of this quite remarkable, high v
age scientific apparatus, Mr. Peek, one
the well-known General Electric Comp
staff of research engineers, has made s
very interesting and remarkable discove
concerning the nature of natural electr
discharges, particularly of lightning
charges. Many people will probably wo
why Mr. Peek and some of his colleag
in the engineering profession, including
three daring German students, Mes
Brasch, Lange and Urban, play with
dangerous electrical discharges, and
they are at all interested in them. We
plained previously why the German sav
are intent on finding out all they can a
these tremendous voltages obtained
the atmosphere, while Mr. Peek, we
say, also has a very practical reason
carrying on experiments with these de
dealing bolts of Thor. Mr. Peek has
for many years intent on finding out
causes the huge insulators on long dista
high potential transmission lines to b
down when electrical storms break loose
these regions.

Plauson's Electrical Patents

USP # 1,540,998
Conversion of Atmospheric Electric Energy
6-09-1925

GB157262
Improvements in Electric Motors
1922-07-10

GB157263
Process and Apparatus for Converting Static Atmospheric Electrical Energy into Dynamic Electrical Energy of any Suitable High Periodicity
7-10-1922

British Patent # 299,735
Apparatus for Producing Rapidly Moving Electrons
7-15-1930

FI21227
Elektrisk uppvärmningsanordning
4-25-1946

Varmelegeme med elektriske varmemodstande
DK67691C
9-27-1948

FR877362
Dispositif de chauffage électrique
12-04-1942

DE734794
Elektrisches Heizsystem
4-24-1943

CH222509
Elektrischer Heizkörper zur Erwärmung von Flüssigkeiten
7-31-1942

DE738107
Elektrolyt fuer unmittelbare elektrische Warmwasser-Radiatorenheizung mit Elektroden
8-03-1943

DE433476
Verfahren zur Herstellung von Elektroden und Schleifkontakten fuer Dynamomaschinen
8-31-1926

CH94021
Elektrode und Verfahren zu deren Herstellung
4-01-1922

CA226423
Electrode for Electrolytic Apparatuses
11-21-1922

---

http://www.meridian-int-res.com/Energy/Atmospheric.htm

Atmospheric Electricity Research
[Excerpts]

In the nineteenth and early twentieth centuries, a large number of researchers investigated ways to extract electrical power from the Earth's ambient electr

The leader in this field was Dr Hermann Plauson who in the 1920s succeeded in generating significant quantities of electrical power comparable with m photovoltaic systems of a similar scale...

The leader in this field before the Second World War appears to have been Dr Hermann Plauson. Dr Plauson was an Estonian citizen who lived in Ha Switzerland. He carried out experiments in Finland with aerostats manufactured from magnesium-aluminium alloy, covered with electrolytically deposit The needles were further doped with a radium compound to increase local ionisation of the air. (This was the era in which the hands of watches were h with radium to make them luminous in the dark). Zinc amalgam patches were also painted onto the aerostats. Plauson obtained a power output of betwe and 3.4kW from one and two aerostats 300m above ground level. Dr Plauson filed patents in the USA, Great Britain and Germany in the 1920s "Gewinnung und Verwertung der Atmosphärischen Elektrizität" is the most detailed known account of the technology.

Other atmospheric electricity researchers contemporary to Dr Plauson included Walter Pennock and MW Dewey in the USA, Andor Palencsar in Hung Heinrich Rudolph in Germany. Hippolyte Charles Vion in Paris predated them all, putting forward proposals in the 1850s and 1860s.

Heinrich Rudolph made an interesting contribution to the design of the aerostat collectors. In 1898 he designed an elliptical aerostat made up of faceted minimise the effect of wind. The design bears a strong resemblance to Northrop's 2003 UCARS unamnned helicopter UAV project. The design uses Effect to help keep the aerostat on station and minimise wind effects.

In recent times, the only person who seems to have been active in this field is Dr Oleg Jefimenko. Dr Jefimenko carried out experiments on driving e motors from the Earth's electric field in the 1970s and has recently called for research into the neglected field of electrostatic motors to be renewed.

MIR's Research Programme

Since 1997 we have been carrying out theoretical research into conversion of atmospheric electricity into useable electrical power.

From a low level (5m high) simple zinc antenna we are able to obtain sufficient charge to light a number of white power LEDs. Further experimental in with metallic aerostat collectors and cavity resonant slow wave antennae concepts are ongoing...

Advantages of Atmospheric Electricity

Simple and robust technology
Low Cost technology - much cheaper than photovoltaics or wind turbines
Available day and night in all weather conditions - in fact, more power is produced at night than during the day
Available at any point on the Earth's surface

1. *Gewinnung und Verwertung der Atmosphärischen Elektrizität*, Dr Hermann Plauson, Hamburg, (1920)
2. Conversion of Atmospheric Electric Energy, USP 1,540,998, Dr Hermann Plauson, (1925)
3. Assembly for the Induction of Lightning into a Superconducting Magnetic Energy Storage System, USP 5,367,245 Goven Mims, (1994)
4. Electrostatic Motors are Powered by Electric Field of the Earth; CL Stong, *Scientific American*, (October 1974)
5. Operation of Electric Motors from the Atmospheric Electric Field; Dr Oleg Jefimenko, *American Journal of Physics*, vol. 39, July 1971.
6. *Electrostatic Motors: Their Principles, Types and Theory of Operation*; Dr Oleg Jefimenko, Electret Scientific, (1972).
7. Parametric Electric Machine, USP 4,622,510, Ferdinand Cap, (1986).

US Patent # 1,540,998

Conversion of Atmospheric Electric Energy

(9 June 1925)

Hermann PLAUSON

Be it known that I, Hermann Plauson, Estonian subject, residing in Hamburg, Germany, have invented certain new and useful improvements in the Conversio Electric Energy, of which the following is a specification.

Methods of obtaining atmospheric electricity by means of metallic nettings set with spikes which are held by means of ordinary or anchored kite balloons m filled with hydrogen, are in theory already known. Atmospheric electricity obtained in this way has been suggested to be used in the form of direct current fo accumulators. This knowledge however is at present only theoretical as the conversion in practice has hitherto been a failure. No means are known of protect from destruction by lightning. The balloons used for collecting the charge must also me be made of very large size in order to be able to support the weig netting and the heavy cable connections.

Instead of using heavy metallic netting as collectors attached to single air ballons of non-conducting materials which are liable to be torn and are permeabl proposed to use metallic balloon collectors which have the following important advantages ---

(a) The metallic cases are impenetrable to helium and hydrogen; they also represent large metallic weather-proof collecting surfaces.

(b) Radio active means the like may be easily applied internally or externally; whereby the ionization is considerable increased and therewith also the quanti electricity capable of being collected.

(c) Such balloon collectors of light metal do not require to be of large size as they have to carry only their own moderate weight, and that of the conducting cab

(d) The entire system therefore offers little surface for the action of storm and wind and is resistant and stable.

(e) Each balloon can be easily raised and lowered by means of a winch so that all repairs, recharging and the like can be carried out without danger during the o

It is further proposed to use a collecting aerial network of several separate collectors spread out in the air above the earth, which collectors are interconne conductors.

According to this invention charges of atmospheric electricity are not directly converted into mechanical energy, and this forms the main difference from pre but the static electricity which runs to earth through aerial conductors in the form of direct current of very high voltage and low current strength is conve dynamic energy in the form of high frequency vibrations. Many advantages are thereby obtained and all disadvantages avoided.

The very high voltage of static electricity of a low current strength can be converted by this invention to voltages more suitable for technical purposes and of gr the use of closed oscillatory circuits it is possible to obtain electromagnetic waves of various amplitude and thereby to increase the degree of resonance of s resonance allows various values of inductance to be chosen whereby again the governing of the starting and stopping of machines driven thereby by simply tun between coils of the machine and the transformer circuit forming the resonance can easily be obtained. Further, such currents have the property of being dire various uses, even without employing them for driving motors, of which there may be particularly mentioned, lighting, production of heat and use in electro-ch

Further, with such currents a series of apparatus may be fed without direct current supply through conductors and also the electro-magnetic high frequency converted by means of special motors adapted for electro-magnetic oscillations into mechanical energy, or finally converted by special machines into alternati frequency or even into direct current of high potential.

The invention is more particularly described with reference to the accompanying diagrams in which: ---

Figure 1 is and explanatory figure.

Figure 2 is a diagrammatic view of the simplest form.

Figure 3 shows a method of converting atmospheric electrical energy for use with motors.

Figure 4 is a diagram showing the use of protective means.

Figure 5 is a diagram of an arrangement for converting large current strengths.

Figure 6 is a diagram of an arrangement including controlling means.

Figure 7 shows means whereby the spark gap length can be adjusted.

Figure 8 shows a unipolar connection for the motor.

Figure 9 shows a weak coupled system suitable for use with small power motors.

Figures 10, 11, and 12 show modified arrangements.

Figure 13 shows a form of inductive coupling for the motor circuit.

Figure 14 is a modified form of Figure 13 with inductive coupling.

Figure 15 is an arrangement with non-inductive motor.

Figure 16 is an arrangement with coupling by condenser.

Figure 17, 18, and 19 are diagrams of further modifications.

Figure 20 shows a simple form in which the serial network is combined with special collectors.

Figure 21 shows diagrammatically an arrangement suitable for collecting large quantities of energy.

Figure 22 is a modified arrangement having two rings of collectors.

Figure 23 shows the connection for three rings of collectors.

Figure 24 shows a collecting balloon and diagram of its connection of condenser batteries.

Figure 25 and 26 show modified collector balloon arrangements.

Figure 27 shows a second method of connecting conductor for the balloon aerials.

Figure 28 shows an auto-transformer method of connection.

Figure 29 shows the simplest form of construction with incandescent cathode.

Figure 30 shows a form with cigar shaped balloon.

Figure 31 is a modified arrangement.

Figure 32 shows a form with cathode and electrode enclosed in a vacuum chamber.

Figure 33 is a modified form of Figure 32.

Figure 34 shows an arc light collector.

Figure 35 shows such an arrangement for alternating current.

Figure 36 shows an incandescent collector with Nernst lamp.

Figure 37 shows a form with a gas flame.

Fig. 1.

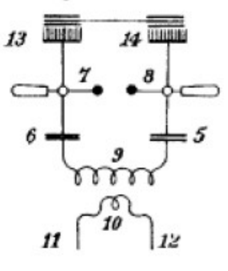

Fig. 2.

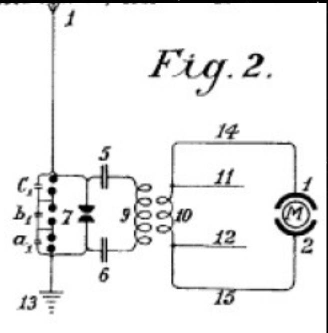

Figure 1 illustrates a simple diagram for converting static electricity into dynamic power of a high number of oscillations. For the sake of clearness in the draw machine is assumed to be employed and not an aerial antenna. 13 and 14 are combs for collecting the static electricity of the influence machine. 7 and 8 are electrodes. 5 and 6 are condensers, 9 an inductive primary coil, 10 secondary coil, 11 and 12 ends of conductors of the secondary coil 10. When the disc of th machine is rotated by mechanical means, the combs collect the electric charges one the positive and the other the negative, and charge the condensers 5 and 6 potential is formed across the spark gap 7-8, that the spark gap is jumped. As the spark gap 7-8 forms a closed circuit with condensers 5 and 6, and inductive well known, waves of high frequency electromagnetic oscillations will pass in this circuit.

The high frequency of the oscillations produced in the primary circuit induces waves of the same periodicity in the secondary circuit. Thus in the primary circui oscillations are formed by the passage of the spark over the spark gap and these waves are maintained by fresh charges of static electricity.

By suitably selecting the ratio between the number of the coils in the primary and secondary circuits with regard to a correct application of the co-effici (especially, inductance and resistance) the high voltage of the primary circuit may be suitably converted into low voltage and high current strength.

When the oscillatory discharges in the primary circuit becomes weaker or entirely cease, the condensers are charged again by the static electricity until the ac again breaks down the spark gap. All this is repeated as long as electricity is produced by the static machine employing mechanical energy.

An elementary form of the invention is shown in Figure 2 in which two spark gaps in parallel are used one of which may be termed the working gap 7 in Fi second serves as a safety device for excess voltage and consists of a larger number of spark gaps than the working section, which gaps are arranged in series a very small capacities as is illustrated in a, b, c, Figure 2 which allow of uniform sparking in the safety section.

In Figure 2 A is the aerial antenna for collecting charges of atmospheric electricity. 13 is the earth connection of the second part of the spark gap, 5 and 6 ar primary coil. Now when through the aerial A the positive atmospheric electricity seeks to combine with the negative charge to earth, this is prevented by (the the spark gaps. The resistance of the spark gap 7 is, as shown in the drawings, lower than that of the other safety section which consists of three spark gaps co and consequently a three times greater air resistance is offered by the latter.

So long, therefore, as the resistance of the spark gap 7 is not overloaded, so that the other spark gaps have an equal resistance with it the discharges take plac gap 7. Should however the voltage be increased by and influences so that it might be dangerous for charging the condensers 5 and 6 or for the coil insul consequence of break down, by a correct regulation of this spark gap the second spark gap can discharge free from inductive effects direct to earth withou machine.

Without this second spark gap, arranged in parallel having a higher resistance than the working spark gap it is impossible to collect and render available l electrical energy.

The actions of this closed oscillation circuit consisting of spark gap 7, two condensers 5 and 6, primary coil 9, and also secondary coil 10 is exactly the described in Figure 1 with the arrangement of the static induction machine with the only difference that here the second spark gap is provided. The elec frequency alternating current obtained can be tapped off from the conductors 11 and 12 for lighting and heating purposes. Special kinds of motors adapted these peculiar electrical charges may be connected at 14 and 15 which can work with static electricity charges or with high frequency oscillations.

In addition to the use of spark gaps in parallel a second measure of security is also necessary for taking off the current. This precaution consists according to the introduction of and method of connecting certain protective electromagnets or choking coils in the aerial circuit as shown by S in Figure 3.

A single electromagnet only having a core of the thinnest possible separate laminations is connected with the aerial.

In the case of high voltages in the aerial network or at places where there are frequent thunder storms, several such magnets may however be connected in serie

In the case of large units or plants several electromagnets can be employed in parallel or in series parallel.

The windings of these electromagnets may be simply connected in series with the aerials. In this case the winding preferably consists of several thin parallel w up together, the necessary section.

The winding may be made of primary and secondary windings in the form of a transformer. The primary windings will be then connected in series with the ae the secondary winding more or less short-circuited over a regulating resistance or an induction coil. In the latter case it is possible to regulate to a certain exten choking coils. In the further description of the connecting and constructional diagrams the aerial electromagnet choke coil is indicated by a simple ring S.

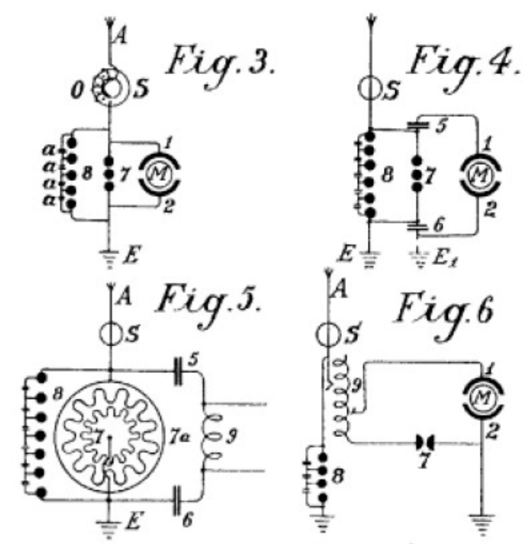

Figure 3 shows the simplest way of converting atmospheric electricity into electromagnetic wave energy by the use of special motors adapted for high oscil static charges of electrical energy. Recent improvements in motors for working with static charges and motors working by resonance, that is to say, having electromagnetic cooperating circuits render this possible but such do not form part of the present invention.

A motor adapted to operate with static charges will for the sake of simplicity be diagrammatically indicated by the two semicircles 1 and 2 and the rotor of th M (Figure 3). A is a vertical aerial or aerial network. S the safety choke or electromagnet with coil O as may be seen is connected with the aerial A. Adjacent S the aerial conductor is divided into three circuits, the circuit 8 giving the safety spark gap, the circuit 7 with the working spark gap, and then a circuit in terminal 1, the rotor and stator terminal 2 at which a connection is made to the earth wire. The two spark gaps are also connected metallically with the earth wi working these diagrams is as follows:

The positive atmospheric electric charge collected tends to combine with the negative electricity (or earth electricity) connected with the earth wire. It travels through the electromagnet S without begin checked as it flows in the same direction as the direct current. Further, its progress is arrested by two sparks gaps and the stator condenser surfaces. The stator condenser surfaces are charged until the charge is greater than the resistance of the spark gap 7, whereupon a spar spark gap 7 and an oscillatory charge is obtained as by means of the motor M, stator surfaces 1 and 2, and spark gap 7, a closed oscillation circuit is obtained electromagnetic oscillations. The motor here forms the capacity and the necessary inductance and resistance, which, as is well known, are necessary for electricity into electromagnetic wave energy.

The discharge formed are converted into mechanical energy in special motors and cannot reach the aerial network by reason of the electromagnet or choke. If, spark springs over the spark gap 7 a greater quantity of atmospheric electricity tends to flow to earth, a counter voltage is induced in the electromagnet, wh more rapidly and strongly the flow of current direct to the earth is. By the formation of this opposing voltage a sufficiently high resistance is offered to the flo electricity direct to earth to prevent a short circuit with the earth.

The circuit containing spark gap 8 having a different wave length which is not in resonance with the natural frequency of the motor, does not endanger the mo security against excess voltage, which, as practical experiments have shown, may still arise in certain cases, but can be conducted direct to earth through this sp

In the diagram illustrated in Figure 4 the spark gap 7 is shunted across condensers 5 and 6 from the motor M. This construction affords mainly a better insula against excess voltage and a uniform excitation through the spark gap 7.

In Figure 5 a diagram is illustrated for transforming large current strengths which may be employed direct without motors, for example, for lighting or heati main difference is that here the spark gap consists of as star shaped disk 7 which can rotate on its own axis and is rotated by a motor opposite similarly fit When separate points of stars face one another, discharges take place, thus forming an oscillation circuit over condensers 5 and 6, and inductance 9 for oscillat is evident that a motor may also be directly connected to the ends of the spiral 9.

The construction of the diagram shown in Figure 6 permits of the oscillation circuit of the motor being connected with an induction coil/ Here a regulating indu introduced for counter-acting excess voltages in the motor. By cutting the separate coils 9 (coupled inductively to the aerial) in or out the inductive action on more or less increased or variable aerial action may be exerted on the oscillation circuit.

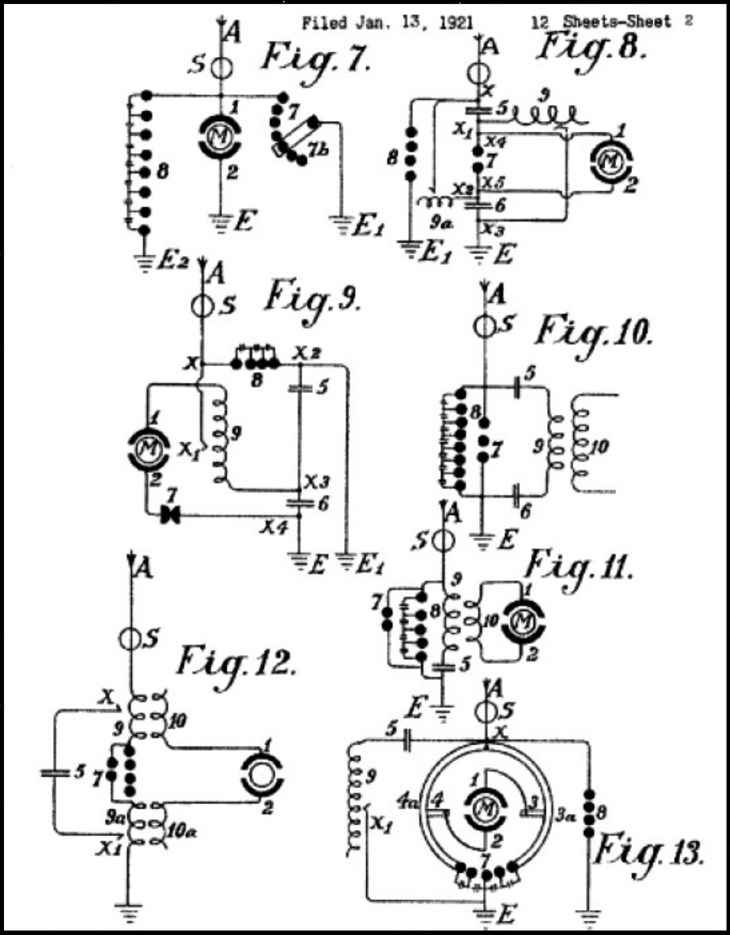

In Figure 7 the oscillation circuit is closed through the earth (E and E1). The spark gap 7 may be prolonged or shortened by more or fewer spark gaps b connected by means of a contact arm 7b.

Diagram 8 shows a unipolar connection of the motor with the aerial network. Here two oscillation circuits are closed through the same motor. The first oscillat from aerial A through electromagnet S, point *x*, inductance 9a to the earth condenser 6 and further, over spark gap 7 to the aerial condenser 5 and back oscillation circuit starts from the aerial condenser 5 at the point x^1 over the inductance 9 to the earth condenser 6 at the point x^3 and through the condenser 6 c 7 back to x^1. The motor itself is inserted between the two points of the spark gap 7. From this arrangement slightly damped oscillation wave currents are produc

In the diagram illustrated in Figure 9 a loosely coupled system of connections is illustrated which is assumed to be for small motors for measuring purpose aerial conductor, S the electromagnet in the aerial conductor, 9 the inductance, 7 the spark gap, 5 and 6 condensers, E the earth, M the motor, and 1 and 2 stat the motor. The motor is directly metallically connected with the oscillation circuit.

In Figure 10 a purely inductive coupling is employed for the motor circuit. The motor is connected with the secondary wire 10 as may be seen in Figure modified diagram connection. The same applies to the diagram of Figure 12.

The diagrams hitherto described preferably allow of motors of small and medium strength to be operated. For large aggregates, however, they are too in construction of two or more oscillation circuits for large amounts of energy is difficult; the governing is still more difficult and the danger in switching on or of

A means of overcoming such difficulties is shown in Figure 13. The oscillation circuit here runs starting from the point x over condenser 5, variable inductan and the two segments (3a and 4a) forming arms of a Wheatstone bridge, back to x, If the motor is connected by brushes 3 and 4 transversely to the two arm shown in the drawings, electromagnetic oscillations of equal sign are induced in the stator surfaces 1 and 2 and the motor does not revolve. If however the br moved in common with the conducting wires 1 and 2 which connect the brushes with the stator poles a certain alteration or displacement of the polarity is motor commences to revolve.

The maximum action will result if one brush 3 comes on the central sparking contact 7 and the other brush 4 on the part *x*. They are however, usually in practi the central contact 7 but only held in the path of the bridge segments 4a and 3a in order not to connect the spark gaps with the motor oscillation circuit.

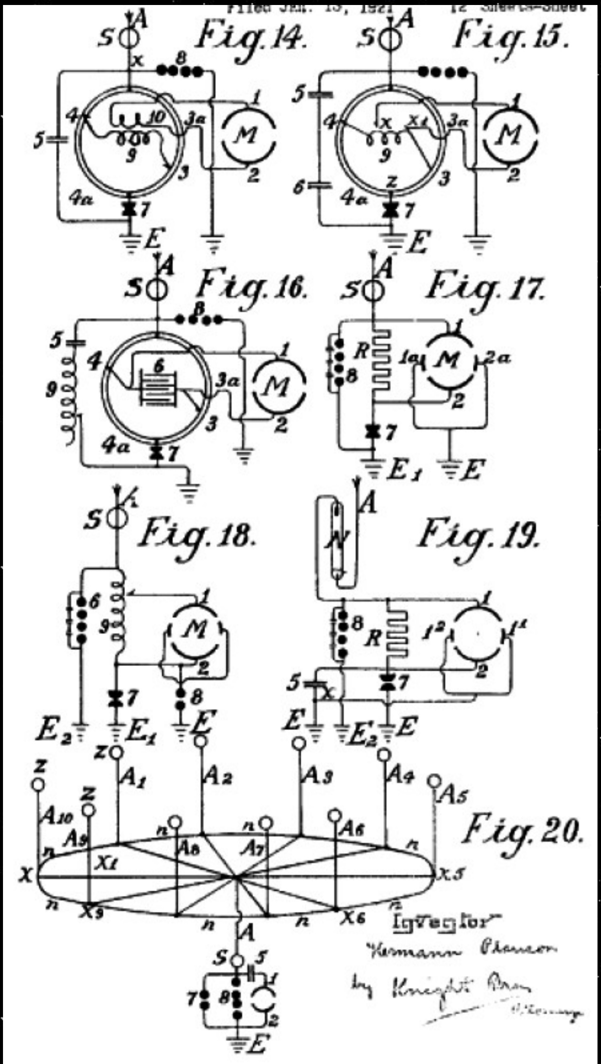
Filed Jan. 13, 1921
12 Sheets-Sheet
Fig. 14.
S A x 8 4 10 3a 5 9 3 1 M 2 4a 7 E
Fig. 15.
S A 5 4 x x1 3a 9 3 1 M 2 6 4a z 7 E
Fig. 16.
S A 8 5 4 6 3a 9 3 1 M 2 4a 7
Fig. 17.
S A R 8 1a M 2a 1 2 7 E1 E
Fig. 18.
S A 6 9 1 M 2 7 8 E2 E1 E
Fig. 19.
N A 8 R 1² 1¹ 2 5 x 7 E E2 E
Fig. 20.
z A1 A2 A3 A4 A5 A10 A9 z n A8 A7 A6 X X1 X5 X9 X6 n A S 5 7 8 1 2 E
Inventor
Hermann Plauson
by Knight Bros
Attorneys

As however, the entire oscillation energy can thereby not act on the motor it is better to carry out the same system according to the diagram 14. The diagram 1 foregoing only by the motor not being directly metallically connected with the segments of the commutator, but only a primary coil 9 which induces in a s current which feeds the motor M and takes the place of the rotor. By this arrangement a good transforming action is obtained, a loose coupling and also an without a spark gap.

In Figure 15 the motor is not purely inductive as in 14, but directly metallically branched off from the primary coil (at *x* and x^1) after the principle of the auto-tr

In Figure 16 instead of an inductance a condenser 6 is in similar manner, and for the same object inserted between the segments 3a and 4a. This has the a segments 3a and 4a need not be made of solid metal but may consist of spiral coils whereby a more exact regulation is possible and further motors of high in employed.

The arrangements of Figures 17, 18 and 19 may be employed for use with resonance and particularly with induction condenser motors; between the larg condenser surfaces, small reversing pole condenser surfaces, mall reversing pole condensers are connected, which, as may be seen from Figures 17, 18 and 19 earth. Such reversing poles have the advantage that with large quantities of electrical energy the spark formation between the separate oscillation circuits ceases

Figure 19 shows a further method which prevents electromagnetic oscillations of high number of alternations formed in the oscillation circuit striking b conductor. It is based on the well known principle that a mercury lamp, one electrode of which is formed of mercury, the other of solid metal such as steel charge to pass in only one direction from the mercury to the steel and not vice versa. The mercury electrode of the vacuum tube N is therefore connecte conductor and the steel electrode with the oscillation circuit. From this it results that charges can pass only from the aerial through the vacuum tube to the osci not vice versa. Oscillations which are formed on being transformed in the oscillation circuit cannot pass to the aerial conductor.

In practice these vacuum tubes must be connected behind an electromagnet as the latter alone affords no protection against the danger of lightning.

As regards the use of spark gaps, all arrangements as used for wireless telegraphy may be used. Of course the spark gaps in large machines must have a surface. In very large stations they are cooled in liquid carbonic acid or better still in liquid nitrogen or hydrogen; in most cases the cooling may also take p liquefied low homologues of the metal series or by means of hydrocarbons the freezing point of which lies at between –90° C and –40° C. The spark gap ca insulated and be of sufficient strength to be able to resist any pressure which may arise. Any undesirable excess super-pressure which may be formed must be off. I have employed wit very good results mercury electrodes which were frozen in liquid carbonic acid, the cooling being maintained during the operation through the walls.

Figure 20 is one of the simplest forms of construction of an aerial network in combination with collectors, transformers and the like illustrated diagrammatic earth wire, 8 the safety spark gap, 7 the working spark gap, 1 and 2 the stator surfaces of the motor, 5 a condenser battery, S the protective magnet which is c coil in aerial conductor, A^1 to A^{10} aerial antennae with collecting balloons, N horizontal collecting or connecting wire from which, to the center a number of co

The actual collectors consist of metal sheaths preferably made of an aluminum magnesium alloy, and are filled with hydrogen or helium and are attached t c wires. The size of the balloon is selected so that the actual weight of the balloon and the weight of the conducting wire is supported thereby. On top of the b spikes, made and gilded in a special manner hereinafter described, are arranged in order to produce a conductor action. Small quantities of radium pr particularly polonium-ionium or meso-thorium preparations considerably increase the ionization, and therewith the action of these collectors.

In addition to metal balloons, fabric balloons which are superficially metal coated according to Schoop's metal spraying process, may also be employed. A me also me produced by lacquering with metallic bronzes, preferably according to Schoop's spraying process or lacquering with metallic bronze powders in two e widely different metals, because thereby the collecting effect is considerably increased.

Instead of the ordinary round balloons, elongated cigar shaped ones may be employed. In order also to utilize the frictional energy of the wind, patches conducting substances which produce electricity by friction, may be attached to the metalized balloon surfaces. The wind will impart a portion of its ener frictional electricity, to the balloon casing, and thereby the collecting effect is substantially increased.

In practice however, very high towers (up to 300 meters is fully admissible) may be employed as antennae. In these towers copper tubes rise freely further ab tower. A gas lamp secured against the wind is then lit at the point of the copper tube and a netting is secured to the copper tube over the flame of this lamp to The gas is conveyed through the interior of the tube up to the summit. The copper tube must be absolutely protected from moisture at the place at which it ent also rain must be prevented running down the walls of the tower which might lead to a bad catastrophe. This is done by bell shaped enlargements which ex being arranged in the tower in the form of high voltage insulators of Siamese pagodas.

Special attention must be devoted to the foundations of such towers. They must be well insulated from the ground, which may be obtained by first embedding a in a box form to a sufficient depth in the ground and inserting in this an asphalt lining and then glass bricks cast about 1 or 2 meters in thickness. Over this ferro-concrete layer in which alone the metal foot of the tube is secured. This concrete block must be at least 2 meters from the ground and be fully protecte wooden covering, from moisture. In the lower part of the tower a wood or glass house for the large condenser batteries or for the motors may be constructed. In earth connection to the ground water, a well insulated pit constructed of vitreous brick, must be provided. Several such towers are erected at equal distances ap with a horizontal conductor. The horizontal connecting wires may either run directly from tower to tower or be carried on bell shaped insulators similar to tho voltage conductors. The width of the network may be of any suitable size and the connection of the motors can take place at any suitable places.

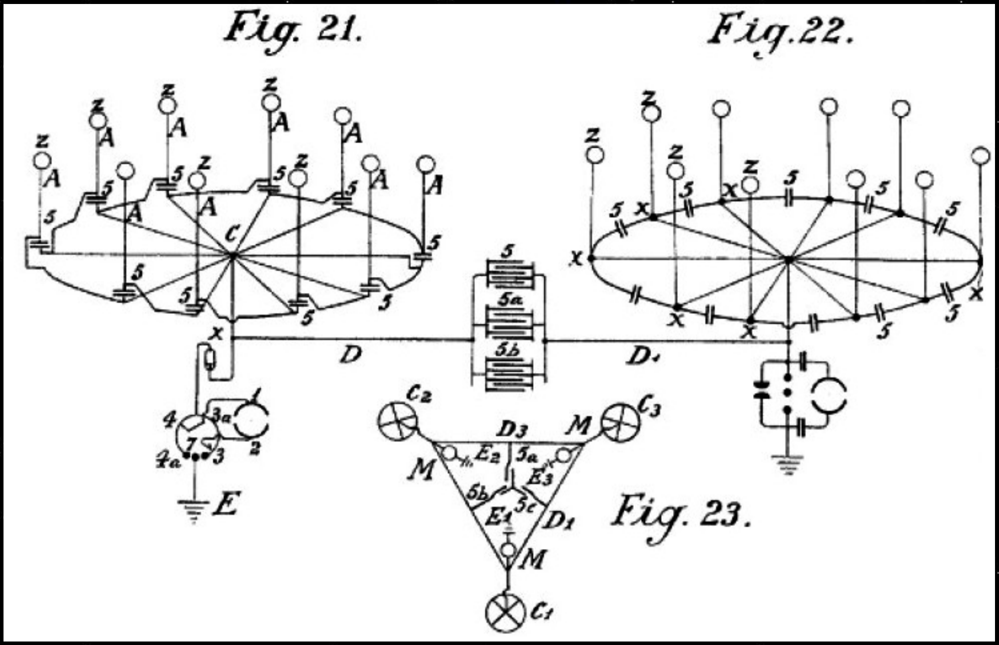
Fig. 21. Fig.22. Fig. 23.

In order to collect large quantities of electricity with few aerials it is well to provide the aerial conductor with batteries of condensers as shown in Figures 21 21 the batteries of condensers 5 are connected on the one hand with the aerial electricity collectors Z by the aerial conductor A, and on the other hand interc with an annular conductor from which horizontal conductors run to the connecting points C to which the earth wire is connected.

Figure 22 shows a similar arrangement, Should two such series of antennae rings be shown by a voltmeter to have a large difference of potential (for ex mountains and one in the plain) or even of different polarity these differences may be compensated for by connecting sufficiently large condenser batteries (5, of Maji star conductors D and D^1. In Figure 23 a connection of three such rings of collectors to form a triangle with a central condenser battery is illustrated.

The condenser batteries of such large installations must be embedded in liquid gases or in liquids freezing at very low temperatures. In such cases a portion o energy must be employed for liquefying these gases. It is also preferable to employ pressure. By this means the condenser surfaces may be diminished, and st quantities of energy to be stored, secure against breakdown. For smaller installation the immersing of the condensers in well-insulated oil or the like, suffices. on the other hand cannot be employed as insulators.

The arrangement in the diagrams hitherto described was always such that the condenser batteries were connected with both poles directly to the aerial condens diagram of the connections for obtaining atmospheric electricity for the condenser batteries has however, been found to be very advantageous. This arrangeme they are connected only by one pole (unipolar) to the collecting network. Such a method of arrangement is very important, as by means of it a constant curren of the normal working pressure or voltage is obtained. If for example a collecting balloon aerial which is allowed to rise to a height of 300 meters, shows 4 earth voltage, in practice it has been found that the working voltage (with a withdrawal of the power according to the method hereinbefore described by me spark gaps and the like) is only about 400 volts. If however, the capacity of the condenser surfaces be increased, which capacity in the above mentioned case w the collecting surface of the balloon aerials, to double the amount, by connecting the condenser batteries with only one pole, the voltage rises under an eq current up to and beyond 500 volts. This can only be ascribed to the favorable action of the connecting method.

In addition to this substantial improvement it has also been found preferable to insert double inductances with electromagnets and to place the capacities prefer
such electromagnets. It has also been found that the useful action of such condensers can be further increased if an induction coil be connected as inductive
unconnected pole of the condenser, or still better if the condenser itself be made as an induction condenser. Such a condenser may be compared with a sp
compressed carries in itself accumulated force, which it again gives off when released. In charging, a charge with reversed sign is formed at the other free cond
through the spark gap a short circuit results, the accumulated energy is again given back since now new quantities of energy are induced at the condenser po
the conductor network, which in fact charges with opposite signs to that at the free condenser pole. The new induced charges have of course the same sig
network. The whole voltage energy in the aerial is thereby however increased. In the same space of time larger quantities of energy are accumulated than is
such inserted condenser batteries.

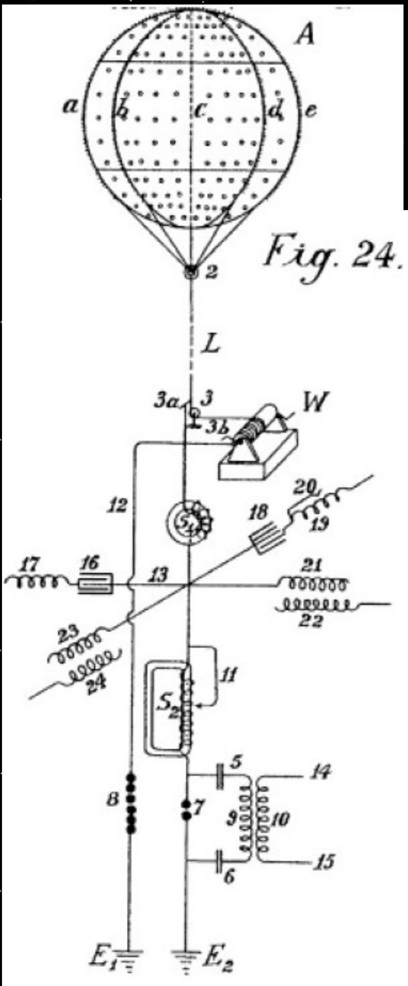

Fig. 24.

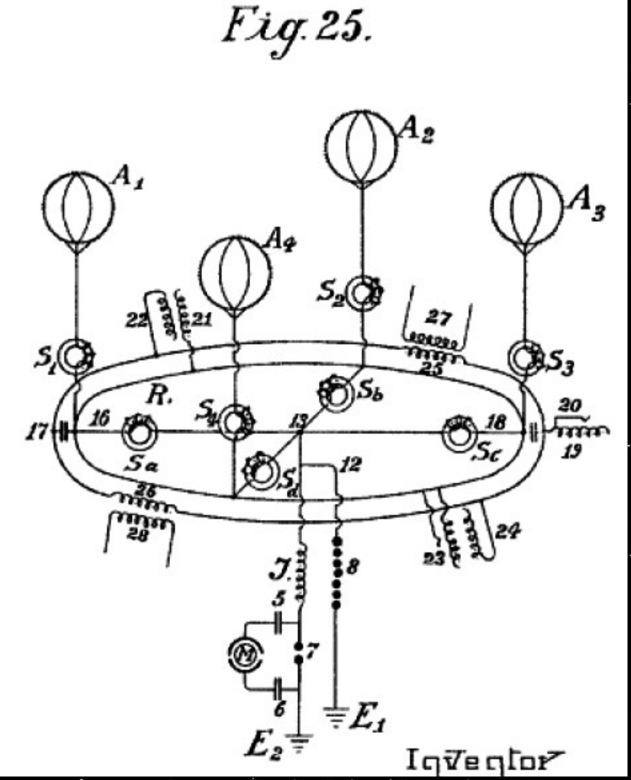

Fig. 25.

In Figures 24 and 25 two different diagrams of connections are more exactly illustrated, Figure 24 shows a collecting balloon and the diagram of the connectio
25 shows four collecting balloons and the parallel connection of the condenser batteries belonging thereto.

A is the collecting balloon made of an aluminum magnesium alloy (electron metal, magnalium) of a specific gravity of 1.8 and a thickness of plate 0.1 to 0.2
are eight strong vertical ribs of T-shaped section about 10 to 20 mm in height and about 3 mm in thickness with the projecting part directed inwards (indicated
so forth); they are riveted together to form a firm skeleton and are stiffened in a horizontal direction by two cross ribs. The ribs are further connected with one
and transversely by means of thin steel wires, whereby the balloon obtains great power of resistance and elasticity. Rolled plates of 0.1 to 0.2 mm in th
magnalium alloy are then either soldered or riveted on this skeleton so that a fully metallic casing with smooth external surface is obtained Well silvered or co
plated steel wires run from each rib to the fastening ring 2. Further, the coppered steel hawser L preferably twisted out of separate thin wires (shown in dotted l
and which must be long enough to allow the balloon to rise in the desired height, leads to a metal roller or pulley 3 and from thence to a winch W, well insulat
By means of this winch, the balloon, which is filled with hydrogen, or helium, can be allowed to rise to a suitable height (300 to 5000 meters) and brought
recharging or repairs.

The actual current is taken directly through a friction contact from the metal roller 3 or from the wire, or even from the winch or simultaneously from all t
brushes (3, 3a and 3b). Beyond the brushes the conductor is divided, the paths being: --- firstly over 12 to the safety spark gap 8, from thence to the earth c
secondly over electromagnet S^1, point 13, to a second loose electromagnet having an adjustable coil S^2, then to the spark gap 7 and to the second earth conduc
working circuit is formed through the spark gap 7, condensers 5 and 6, and through the primary coil 9; here the static electricity formed by oscillatory discharg
and converted into high frequency electromagnetic oscillations. Between the electromagnets S^1 and S^2 at the crossing point 13, four condenser batteries are
are only indicated diagrammatically in the drawings each by one condenser. Two of these batteries (16 and 18) are made as plate condensers and prolon
induction coils or spirals 17 and 19 while the two others (21 and 23) are induction condensers. As may be seen from the drawings each of the four condenser b
and 23 is connected by only one pole to the aerial or to the collector conductor. The second poles 17, 19, 22 and 24 are open. In the case of plate condensers ha
resistance an induction coil is inserted. The object of such a spiral or coil is the displacement of phase by the induction current by 1/4 periods, whilst that of th
of the condenser poles which lie free in the air, works back to the collector aerial. The consequence of this is that in discharges in the collector aerial the back i
the free poles allows a higher voltage to be maintained in the aerial collecting conductor than would otherwise be the case. It has also been found that such a b
extremely favorable effect on the wear of the contacts. Of course the inductive effect may be regulated at will within the limits of the size of the induction coil
coil in action being adjustable by means of wire connection without induction (see Figure 24, No. 20).

S^1 and S^2 may also be provided with such regulating devices in the case of S^2 (illustrated by 11). If excess voltage be formed it is conducted to earth throug
spark gap 8 or through any other suitable apparatus, since this formation would be dangerous for the other apparatus.

The small circles on the collector balloon indicate places at which zinc amalgam or gold amalgam or other photoelectric acting metals in the form of small pat thin layers (0.01 to 0.05 mm in thickness) are applied to the entire balloon as well as in greater thickness to the conducting network. The capacity of the co considerably strengthened at the surface. The greatest possible effect in collecting may be obtained by polonium amalgams and the like. On the surface of the metal points or spikes are also fixed along the ribs, which spikes serve particularly for collecting the collector charge. Since it is well known that the resistanc less the sharper the spike is, for this purpose it is therefore extremely important to employ as sharp spikes as possible. Experiments made as regards these ha formation of the body of the spike or point also plays a large part, for example, spikes made of bars or rollers with smooth surface, have a many times greater p collector accumulator spikes than those with rough surfaces. Various kinds of spike bodies have been experimented with for the collector balloons hereinbefo best results were given by spikes which were made in the following way. Fine points made of steel, copper, nickel, or copper and nickel alloys, were fastened to and then placed as anode with the points in a suitable electrolyte (preferably in hydrochloric acid or muriate of iron solutions) and so treated with weak curr pressure. After 2 to 3 hours according to the thickness of the spikes or pins the points become extremely sharp and the bodies of the spikes have a rough surfac then be removed and the acid washed off with water. The spikes are then placed as cathode in a batch consisting of solution of gold, platinum, iridium, palla salts or their compounds and coated a the cathode galvanically with a thin layer of precious metal, which must however be sufficiently firm to protect them oxidation.

Such spikes act at a 20-fold lower voltage almost as well as the best and finest points made by mechanical means. Still better results are obtained if polonium c added to the galvanic bat when forming he protective layer or coating. Such pins have a low resistance at their points and even at one volt and still lower p excellent collector action.

In Figure 24 the three unconnected poles are not connected with one another in parallel. That is quite possible in practice without altering the principle of the f preferable to interconnect in parallel to a common collector network, a series of collecting aerials:

Figure 25 shows a diagram for such an installation. A^1, A^2, A^3, A^4 are four metal collector balloons with gold or platinum coated spikes which are electrolyti presence of polonium emanations or radium salts, which spikes or needles are connected over four electromagnets S^1, S^2, S^3, S^4, through an annular condu annular conductor four wires run over four further electromagnets Sa, Sb, Sc, Sd, to the connecting point 13. There the conductor is divided, one branch passin safety spark gap 8 to the earth at E^1, the other over inductive resistance J and working spark gap 7 to the earth at E^2. The working circuit, consisting of the c and a resonance motor M, such as hereinbefore described, is connected in proximity round the sparking gap section 7.

Instead of directly connecting the condenser motor of course the primary circuit for high frequency oscillatory current may also be inserted.

The condenser batteries are connected by one pole to the annular conductor R and can be either inductionless (16 and 18) or made as induction condensers as 23. The free poles of the inductionless condensers are indicated by 17 and 19, those of the induction condensers by 22 and 24. As may be seen from the drawin 17, 22, 19, 24 may be interconnected in parallel through a second annular conductor without any fear that thereby the principle of the free pole connection v addition to the advantages already set forth the parallel connection also allows of an equalization of the working pressure in the entire collector network. Su and calculated induction coils 25 and 26 may also be inserted in the annular conductor of the free poles, by means of which a circuit may be formed in the s and 28 which allows current produced in this annular conductor by fluctuations of the charges of the like appearances to be measured or otherwise utilized.

According to what has been hereinbefore stated separate collector balloons may be connected at equidistant stations distributed over the entire country, either c with one another metallically or by means of intermediate suitably connected condenser batteries through high voltage conductors insulated from earth. The s converted through a spark gap into dynamic energy of a high number of oscillations and may in such form be coupled as a source of energy b y means of a connecting, various precautions being observed and with special regulations. The wires leading from the collector balloons have hitherto been connected th conductor without this endless connection, which can be regarded as an endless induction coil, being able to exert any action on the whole conductor system.

It has now been found that if the network conductor connecting the aerial collector balloons with one another is not made as a simple annular conductor, bu circuited in the form of coils over a condenser battery or spark gap or through thermionic tubes or valves or audions, then the total collecting network e properties. The collection of atmospheric electricity is thereby not only increased but an alternating field may be easily produced in the collector netw atmospheric electrical forces showing themselves in the higher regions may also be directly obtained by induction. In Figures 26 and 28 a form of constructio basis of which the further foundations of the method will be more particularly explained.

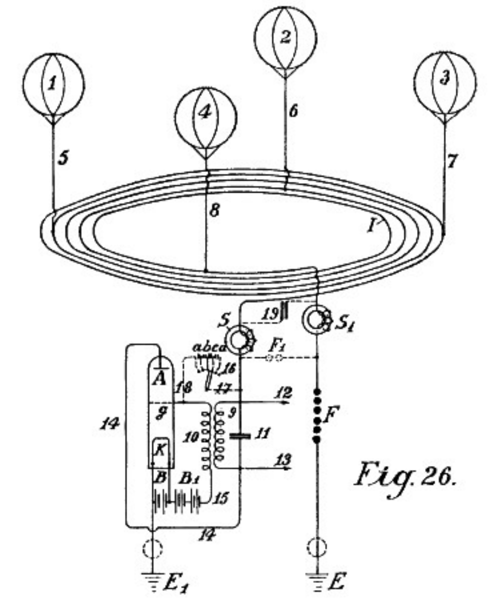

Fig. 26.

In Figure 26, 1, 2, 3, 4 are metal collector balloons, 5, 6, 7, 8 their metallic aerial conductors and I the actual collector network. This consists of five coils a high voltage insulators in the air, in high voltage masts (or with a suitable construction of cable embedded in the earth). One coil has a diameter of 1 or 100 k S^1 are two protective electromagnets, F the second safety section against excess voltage, E its earth conductor and E1 the earth conductor of the working absorption of static atmospheric electricity is effected through the four balloon collectors, the current in order to reach the earth connection E1 must flow sp collector network over the electromagnet S, primary induction coil 9, conductor 14, anode A of the audion tube, incandescent cathode K, as the way over the e safety spark gap F offers considerably greater resistance. Owing to the fact that the accumulated current flows in one direction, an electromagnetic alternating in the interior of the collector network coil, whereby the whole free electrons are directed more or less into the interior of the coil. An increased ionization of thereby produced. In consequence of this the points mounted on the collector balloon show a considerably reduced resistance and therefore increased static ch points on the balloon and the surrounding atmosphere are produced. The result of this is a considerably increased collector effect.

A second effect which could not be obtained otherwise is obtained by the electromagnetic alternating field which running parallel to the earth surface, acts m diminishing or increasing effect on the earth magnetic field, whereby in the case of fluctuations in the current a return induction current of reversed sign is al the collector coil by earth magnetism. Now if, however, a constantly pulsating continuous alternating field is produced as stated in the above collector networ current of the same periodicity is produced also in the collecting network coil. As the same alternating field is further transmitted to the aerial balloon, the resis is thereby considerably reduced, whilst the collector action is considerably increased. A further advantage is that positive electrons which collect on the meta the conversion into dynamic current produce a so-called drop of potential of the collector area. As an alternating field is present, the negative ions surround surfaces, when discharge of the collector surfaces takes place produce by the law of induction, an induction of reversed sign on the collector surface and so fo again a positive charge). In addition to the advantages hereinbefore set forth, the construction of connecting conductors in coil form when of sufficiently large of a utilization of energy arising in higher regions also in the simplest way. As is well known electric discharges frequently take place at very great elevatio observed as St Elmo's fire or northern lights. These energy quantities have no been available to be utilized up to now. By this invention all these kinds of ener an electromagnetic nature and the direction of axis of the collector coils stands at right angles to the earth's surface, can be more or less absorbed in the same in wireless telegraphy absorbs waves coming from a far distance. With a large diameter of the spiral it is possible to connect large surfaces and thereby to t quantities of energy.

It is well known that large wireless stations in the summer months, and also in the tropics are very frequently unable to receive the signals in consequence of in are caused by atmospheric electricity, and this takes place with vertical coils of only 40 to 100 meters diameter. If on the contrary horizontal coils of 1 to 10 employed very strong currents may be obtained through discharges which are constantly taking place in the atmosphere. Particularly in the tropics or still b regions where the northern lights are constantly present, large quantities of energy may probably be obtained in this way. A coil with several windings shou similar manner any alteration of the earth magnetism should act inductively on such a coil.

It is not at all unlikely that earthquakes and spots on the sun will also produce an induction in such collector coils of sufficient size. In similar manner this co will react on earth current more particularly when they are near the surface of the earth or even embedded in the earth. By combining the previous kind of cur far as they are adapted for the improved system with the improved possibilities of obtaining current the quantities of free natural electricity which are to be obt of electricity are considerably increased.

In order to produce in the improved collector coil uniform current oscillations of an undamped nature so-called audion high vacuum or thermionic tubes of su are employed instead of the previously known spark gaps (Figure 26, Nos. 9-18). The main aerial current flows through electromagnet S (which in the case of alternations is not connected here but in the earth conductor E^1) and may be conveyed over the primary coils in the induction winding through wire 14 to th high vacuum grid tube. Parallel with the induction resistance 9 a regulating capacity of suitable size, such as condenser 11 is inserted. In the lower part of the is arranged the incandescent filament or the cathode K which is fed through a battery B. From the battery B two branches run, one to the earth conductor through battery B^1 and secondary coil 10 to the grid above *g* in the vacuum tube. By the method of connections shown in dotted lines, a desired voltage at th may also be produced through the wire 17 which is branched off from the main current conductor through switches 16 and some small condensers (*a, b, c* series, and conductor 18, without the battery B^1 being required.

The action of the entire system is somewhat as follows: --

On the connecting conductor of the aerial collector network being short circuited to earth, the condenser pole 11 is charged and slightly damped oscillations short circuited existing oscillation circuit formed of the condenser 11 and self inductance 9. In consequence of the coupling through coil 10, fluctuations of vol the grid circuit 15 with the same frequency, which, fluctuations in turn influence the strength of the electrode current passing through the high vacuum amplify produce current fluctuations of the same frequency in the anode circuit. A permanent supply of energy to the oscillation circuits 9 and 10 consequently ta condition of balance is set up, in which the consumed oscillation energy is equal to that absorbed. Thereby constant undamped oscillations are now produced circuits 9-11.

For regular working of such oscillation producers high vacuum strengthening tubes are necessary and it is also necessary that the grid and anode voltages s difference of 180° so that if the grid is negatively charged, then the anode is positively charged and vice versa. This necessary difference of phase may be varied connections, for example, by placing the oscillation circuit in the grid circuit or by separating the oscillation circuit and inductive coupling from the an circuit and so forth.

A second important factor in this way of converting static atmospheric electricity into undamped oscillations is that care must be taken hat the grid and anod certain relation to one another; the latter may be obtained by altering the coupling and a suitable selection of the self-induction in the grid circuit, or as shown b 17, 18 by means of a larger or smaller number of condensers of suitable size connected in series; in this case the battery B^1 may be omitted. With a suitable se potential a glow discharge takes place between the grid *g* and the anode A, and accordingly at the grid there is a cathode drop and a dark space is formed cathode drop is influenced by the ions which are emitted in the lower space in consequence of shock ionization of the incandescent cathodes K and pass throu upper space. On the other hand the number of ions passing through the grid is dependent on the voltage between the grid and the cathode. Thus is the grid v periodic fluctuations (as in the present case) the amount of the cathode drop at the grid fluctuates and consequently the internal resistance of the tube correspon so that when a back coupling of the feed circuit with the grid circuit takes place, the necessary means are afforded for producing undamped oscillations and according to requirements from the collecting conductor.

The frequency of the undamped oscillations produced is with a suitably loose coupling equal to the self-frequency of the oscillation circuits 9 and 10. By a su the self induction of the coil 9 and capacity 11 it is possible to extend from frequencies which produce electromagnetic oscillation of only a few meters wavel lowest practical alternating current frequency. For large installations a suitable number of frequency-producing tubes of the well-known high vacuum transmi to 2 kw in size may be connected in parallel so that in this respect no difficulty exists.

The use of such tubes for producing undamped oscillations, and also the construction and method of inserting such transmission tubes in an accumulator or known and also that such oscillation producing tubes only work well at voltages of 1000 up to 4000 volts, so that on the contrary their use at lower voltages is c difficult. By the use of high voltage static electricity this method of producing undamped oscillations as compared with that through spark gaps must be reg solution particularly for small installations of outputs of from 1 to 100 kw.

By the application of safety spark gaps, with interpolation of electromagnets, not only is short-circuiting avoided but also the taking up of current is regu producers inserted in the above way form a constantly acting electromagnetic alternating field in the collector coil, whereby as already stated, a considerable ac takes place. The withdrawal wire or working wire is connected at 12 and 13, but current may be taken by means of a secondary coil which is firmly or movabl suitable way inside the large collector coil, i.e., in its electromagnetic alternating field, so long as the direction of its axis runs parallel with that of the main coil.

In producing undamped oscillations of a high frequency (50,000 per second or more) in the oscillation circuits 9 and 11, electromagnets S and S^1 must be in frequency oscillations are not to penetrate the collector oil, between the oscillation producers and the collector coil. In all other cases they are connected s earthing (as in Figures 27 and 28).

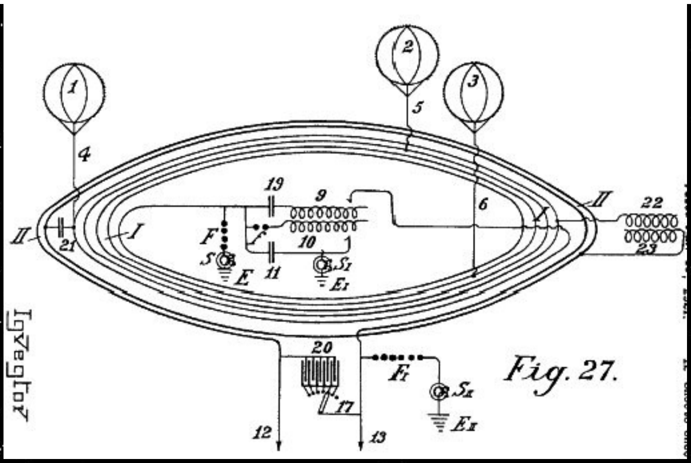

Fig. 27.

In Figure 27 a second method of construction of the connecting conductor of the balloon aerials is illustrated in the form of a coil. The main difference c addition to the connecting conductor I another annular conductor II is inserted parallel to the former on the high voltage masts in the air (or embedded as a cabl both in te form of a coil. The connecting wire of the balloon aerials is indicated as a primary conductor and also as a current producing network; the other is network and is not in unipolar connection with the current producing network.

In Figure 27 the current producing network I is shown with three balloon collectors 1, 2, 3 and aerial conductors 4, 5, 6; it is short-circuited through c inductance 9. The oscillation forming circuit consists in this diagram of spark gap f, inductance 10, and condenser 11; the earth wire E, is connected to earth o S^1. F s the safety spark gap which is also connected to earth through a second electromagnet S at E. On connecting up the condenser circuit 11 this is charged c f whereby an oscillatory discharge is formed. This discharging current acts through inductance 10 on the inductively coupled secondary 9, whereby in the pro modification of the potential of the condenser 19 is produced. The consequence of this is that oscillations arise in the coil shaped producer network. These osc current in the secondary circuit II, which has a smaller number of windings and a less resistance, the voltage of which, according to the proportion of the nu and of the ohmic resistance, is considerably lower whilst the current strength is greater.

In order to convert the current thus obtained into current of an undamped character, and to tune its wavelengths, a sufficiently large regulatable capacity 20 in the ends 12 and 13 of the secondary conductor II. Here also current may be taken without an earthy conductor, but it is advisable to insert a safety spark gap l this with the earth over an electromagnet S2.

The producer network may be connected with the working network II over an inductionless condenser 21 or over an induction condenser 22, 23. In this ca conductor is unipolarly connected with the energy conductor.

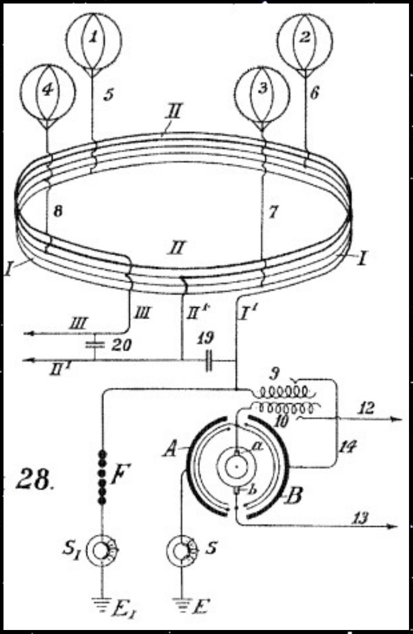

28.

In Figure 28 the connecting conductor between the separate accumulator balloons is carried out according to the autotransformer principle. The collecting c aerial balloons 1, 2, 3, 4, the windings of which are not made side by side but one above the other. In Figure 28 the collector coil I is shown with a thin lin connected prolongation coils II with a thick line. Between the ends I^1 and II^1 of the energy network I a regulating capacity 19 is inserted. The wire I^1 is co output wire and with the spark gap F.

As transformer of the atmospheric electricity an arrangement is employed which consists in using rotary pairs of condensers in which the one stator surface B the main current, whilst the other A is connected with the earth pole. Between these pairs of short-circuited condensers are caused to rotate from which the con be taken by means of two collector rings and brushes, in the form of an alternating current, the frequency of which is dependent on the number of balloons and the rotor. As the alternating current formed in the rotor can act, in this improved method of connection described in this invention, through coils 1 on the increase or diminution of the feed current in I can be obtained according to the direction of the current by back induction. Current oscillations of uniform rhyt in the coil shaped windings of the produce network.

As the ends of this conductor are short-circuited through the regulatable condenser 19 these rhythms produce short-circuited undamped oscillations in the ener periodicity and wave lengths of which oscillations can be adjusted according to desire by altering the capacity 19 to a given wavelength and therewith also to a

These currents may also be employed in this form directly as working current through the conductors II[1] and III. By inserting the condenser 20 a connecti conductors may also be made, whereby harmonic oscillations of desired wavelengths are formed. By this means quite new effects as regards current distribu The withdrawal of current can even take place without direct wire connection if, at a suitable point in the interior of the producing network (quite immaterially a diameter of 1 or 100 km) a coil tuned to these wavelengths and of the desired capacity is firmly or movably mounted in the aerial conductor in such a direction is in parallel with that of the collector coil. In this case a current is induced in the producing network, the size of which is dependent on the total capac and also on the periodicity employed. A possibility is thereby afforded in future, of taking energy from the producer network by wireless means. As there atmospheric electricity also magnetic earth currents and energy from the higher atmosphere (at leas partially may be simultaneously obtained, this last system atmospheric energy is of particular importance for the future.

Of course everywhere instead of spark gaps grid vacuum tubes may be employed as producers for undamped oscillations. The separate coils of the producer n diameters may be connected with one another through separate conductors all in parallel or all in series or in groups in series. By regulating the number of osc the extent of the voltage more or less large collector coils of this kind may be employed. The coils may also be divided spirally over the entire section. The co out in angular form or also in triangular, quadrangular, hexagonal or octagonal form.

Of course wires may be carried from a suitable place to the center or also laterally which serve the current waves as guides. This is necessary when the cu conducted over mountains and valley and so forth. In all these cases the current must be converted into a current of suitable periodicity.

As already hereinbefore mentioned separate collecting balloons may be directly metallically interconnected at equidistant stations distributed over the entire c connected by interpolation of suitable condenser batteries by means of high voltage conductors. The static electricity is converted through a spark gap into dy high number of oscillations, and could then in such forms, wit a suitable arrangement of the connections, observing various measures of protection, be empl energy after separate or special regulation.

According to this invention in order to increase the collecting effect of the balloon in the aerial collector conductor or in the earth wire, radiating collectors are consist either of incandescent metal or oxide electrodes in the form of vacuum grid tubes, or electric arcs (mercury and the like electrodes), Nernst lamps, or various kinds may be simply connected with the respective conductor.

It is well known that energy can be drawn of from a cathode consisting of an incandescent body opposite an anode charged with positive electricity (vacuum gr however, a cathode was always first directly placed opposite an anode, and secondly the system always consisted of a closed circuit.

Now if we dispense with the ordinary ideas in forming light or flame arcs in which a cathode must always stand directly opposite an anode, charged to a another body freely floating in the air, or regard the incandescent cathode only as a source of unipolar discharge (which represent group and point discharg machines similar to unipolar discharges), it may be ascertained that incandescent cathodes and less perfectly all incandescent radiators, flames and the like a large current densities and allow large quantities of electric energy to radiate into the open space in the form of electron streams as transmitters.

The object of this invention is as described below, if such incandescent oxide electrodes or other incandescent radiators or flames are not freely suspen connected metallically with the earth so that they can be charged with negative terrestrial electricity, these radiators possess the property of absorbing the free charges contained in the air space surrounding them (that is to say of collecting them and conducting them to earth). They can therefore serve as collec comparison to the action of the spikes, or points, a very large radius of action R; the effective capacity of these collectors is much greater than the geometri calculated in an electrostatic sense.

Now as our earth is surrounded as is well known with an electrostatic field and the difference of potential

$$\frac{\delta V}{\delta h}$$

of the earth field according to the latest investigations, is in summer about 60 to 100 volts and in winter 300 to 500 volts per meter of difference in heigh calculation gives the result that when such a radiation collector or flame collector is arranged for example on the ground, and a second one is mounted vert distance of 2000 meters and both are connected by a conducting cable, there is a difference in potential in summer of about 2,000,000 volts and in winter e volts and more.

According to Stefan Boltzmann's law of radiation, the quantity of energy which an incandescent surface (temperature T) of 1 sq cm radiates in a unit of time (temperature T_o) is expressed by the following formula:

$S = \delta (T^4 - T^4_o)$ watt./sq cm.

And the universal radiation constant δ is according to the latest researches of Ferry (*Annales de Chimie et de Physique*, 17: 267 [1909]) equal to 6.30 x 10^{-12} w

Now if an in incandescent surface of 1 sq cm shows, as compared with the surrounding space a periodic fall of potential δV it radiates (independent of the curr is to say of the sign) in accordance with the above formula, for example at a temperature of 3725° C an energy of 1.6 kw/sq cm/second. As for the radiation th be calculated for the collection of energy, but reversed. Now, as carbon electrodes at the temperature of the electric arc support on the current basis a current d 60 to 65 amperes per sq cm no difficulties will result in this direction in employing radiating collectors as accumulators.

If the earth be regarded as a cosmically insulated condenser in the sense of geometrical electrostatics *x* there results from the geometric (compare Ed *Elektrische Bogenlicht* [The Electric Arc Light], page 169) capacity of the earth according to Chwolson:

For negative charging 1.3 x 10^6 Coulomb

For negative potential V = 10 x 10^8 volts.

From this there results however, EJT = 24.7 x 10^{24} watt/sec. Now if it is desired to make a theoretic short circuit through an earthed flame collector this w electric total work of about 79,500 x 10^{10} kilowatt years. As the earth must be regarded as a rotating mechanism which thermodynamically, electromagr kinematically coupled with the sun and star system by cosmic radiations and gravitation a diminution of the electric energy of the earth field is not to be fea which the incandescent collectors would withdraw from the earth field can only case by the withdrawal of motor work a lowering of the earth temperature (300) and reduce this to that of the world space (T = O) by using the entire energy. This is however not the case as the earth does not represent a cosmically system. On the contrary there is conveyed to the same according to the recent value corrected by Perry for the solar constants through the radiation from the 18,500 x 10^{10} kw. Accordingly, any lowering of the earth temperature (T_E) without a simultaneous lowering of the sun's temperature (T_s) would contradict St law of radiation.

$S = \delta (s^4 - T^4)$.

From this it must be concluded that if the earth temperature (T_E) sinks the total radiation S absorbed by the earth increases, and further also that the secular sp
the earth is directly dependent on that of the sun and the other radiators cosmically coupled with the sun and is connected most closely with these.

The incandescent radiation collectors may, according to this invention, be employed for collecting atmospheric electricity if they (1) are charged with t
electricity (that is to say when they are directly connected by means of a metallic conductor with the earth) and (2) if large capacities (metal surfaces) charge
are mounted opposite them as positive poles in the air. This is regarded as the main feature of the present invention as without these inventive ideas it would
collect with an incandescent collector, sufficiently large quantities of the electrical charges contained in the atmosphere as the technology requires; the radiu
flame collectors would also be too small, especially if it be considered that the very small surface density (energy density) (δ about = 2 x 7 . 10^9 St. E. per
allow of large quantities of charge being absorbed from the atmosphere.

x) Calculated according to Poisson's calculation:

$\Delta V = -4\pi\delta$; as here the alteration of the potential or potential gradients only takes place in the direction o the normal, this calculation assumes the simple fo

$$\delta = \frac{1}{4\pi} \times \frac{\delta^2 V}{\delta n^2}$$

It has indeed already been proposed to employ flame collectors for collecting atmospheric electricity and it is known that their collecting effect is substantially
the points. It is however, not known that the quantities of current which could hitherto be obtained are too small for technical purposes. According to my exper
for this is to be found in the too small capacities of the collector conductor poles. If such flame or radiating collectors have no or only small positive surfac
action for large technical purposes is too small. If the incandescent collectors be constantly kept in movement in the air they may collect more according to
movement, but this again is not capable of being carried out in practice.

By this invention the collector effect is considerably increased by a body charged with a positive potential and of the best possible capacity being also held
direct earth connection) opposite such an incandescent collector which is held floating in the air at a desired height. If for example, a collecting balloon of
metalized balloon fabric be caused to mount to 300 up to 3000 meters in the air and as positive pole it is brought opposite such a radiating collector connected
the earth, quite different results are obtained.

The metallic balloon shell (with a large surface) is charged to a high potential by the atmospheric electricity. This potential is greater the higher the collecting
the incandescent collector. The positive electricity acts concentratedly on the anode floating in the air as it is attracted through the radiation shock ionization
the incandescent cathode. The consequence of this is that the radius of action of the incandescent cathode collector is considerably increased and thereby a
effect of the collecting balloon surface. Further the large capacity of the anode floating in the air plays therefore an important part because it allows of the takin
and thereby a more uniform current is obtained even when there is a large consumption: this cannot be the case with small surfaces.

In the present case the metallic collecting balloon is a positive anode floating in the air and the end of the earth conductor of this balloon serves as positive pol
the surface of the radiating incandescent cathode, which in turn is charged with negative earth electricity being conductively connected to earth.

The process may be carried out by two such contacts (negative incandescent cathode and anode end of a capacity floating in the air) a condenser and an in
being switched on in parallel, whereby simultaneously undamped oscillations may be formed.

In very large installations it is advisable to connect two such radiating collectors in series. Thus an arc light incandescent cathode may be placed below on the
an incandescent cathode which is heated by special electromagnetic currents be located high in the air. Of course for this the special vacuum Liebig tubes wi
may also be employed. An ordinary arc lamp with oxide electrodes may be introduced on the ground and the positive pole is not directly connected with the c
but through the upper incandescent cathode or over a condenser. The method of connecting the incandescent cathode floating in the air may be seen in Figures 2

B is the air balloon, K a Cardan ring (connection with the hawser), C the balloon, L a good connecting cable, P a positive pole, N negative incandescent cat
conductor.

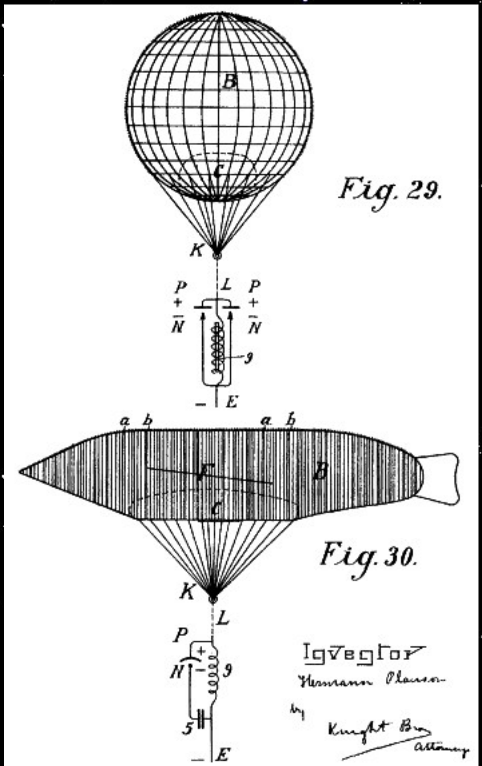

Figure 29 represents the simplest form of construction. If electric oscillations are produced below on the ground by means of a carbon arc lamp or in oth considerably greater electric resistance is opposed to that in the direct way by inserting an electrical inductive resistance 9. Consequently between P and N, a and as, over N and P only an inductionless ohmic resistance is present, a spark will spring over so long as the separate induction co-efficients and the calculated. The consequence of this is that the oxide electrode (carbon or the like) is rendered incandescent and then shows as incandescent cathode an in effect. The positive poles must be substantially larger than the negative in order that they may not also become incandescent. As they are further connect balloon area which has a large capacity and is charged at high voltage, an incandescent body which is held floating in the air and a positive pole which capacities is thereby obtained in the simplest way. The incandescent cathode is first caused to become incandescent by means of separate energy produced on t maintained by the energy collected from the atmosphere.

Figure 30 only shows the difference that instead of a round balloon a cigar-shaped one (of metal or metalized fabric) may be employed and also a conde between the incandescent cathode and the earth conductor so that a short circuited oscillation circuit over P.N. 5 and 9 is obtained. This has the advantag quantities of electricity cause the cathode to become incandescent and much larger cathode bodies may be rendered incandescent.

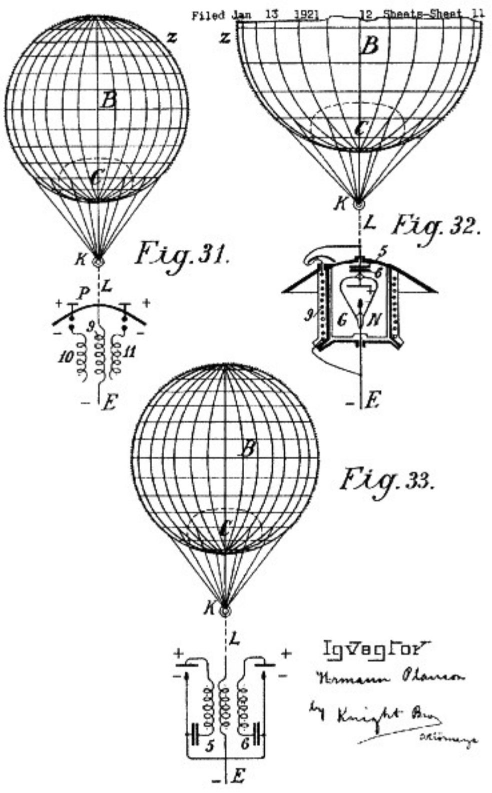

In this form of construction both the incandescent cathode and also the positive electrode may be enclosed in a vacuum chamber as may be seen in Figure carried well insulated through the cover of a vessel and ends in a condenser disc 5. The cover is arched in order to keep off the rain. The vessel is entirely or magnetic metal and well-insulated inside and outside. Opposite the disc 5 another disc 6 and on this again a positive pile of the vacuum tube g with the inca (oxide electrode) N is arranged. The negative electrode is on the one hand comected with the earth conductor E; and on the other hand with the inductive resi also connected with the cable L with the positive pole and wound round the vessel in coils. The action is exactly the same as that in Figure 29 only in incandescent cathode one enclosed in vacuo is employed. As in such collectors only small bodies can be brought to incandescence in large installations a vacuum tubes must be inserted in proximity to one another. According to the previous constructions Figures 31 and 33 are quite self evident without further exp

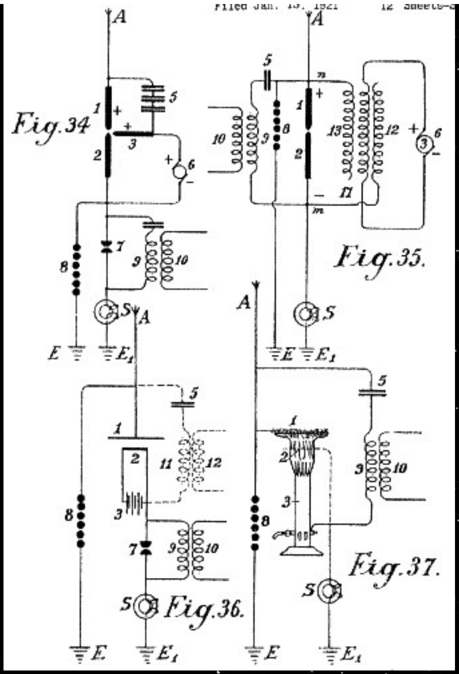

Figures 34-37 represent further diagrams of connections over radiating and flame collectors, and in fact, how they are to be arranged on the ground.

Figure 34 shows an arc light collector with oxide electrodes for direct current and its connection; Figure 35 a similar one for alternating current, Figure 3 collector with a Nernst lamp and Figure 37 a similar one with a gas flame.

The positive pole 1 of the radiating collectors is always directly connected to the aerial collecting conductor A. In Figure 34 this is further connected over the 5 with a second positive electrode 3. The direct current dynamo *b* produces current which flows over between the electrodes 3 and 2 as an arc light. On the fo the negative incandescent electrode 2 absorbs electricity from the positive poles standing opposite it and highly charged with atmospheric electricity and conve working circuit. The spark gap 7, inductive resistance 9 and induction coil 10 are like the ones previously described. The protective electromagnet S guar against earth circuiting, the safety spark gap 8 from excess voltage or overcharging.

In Figure 35 the connection is so far altered that the alternating current dynamo feeds the exciting coil 11 of the induction condenser. 12 is its negative and 13 i the coil 3 on the magnet core of the dynamo is correctly calculated and the periodicity of the alternating current is sufficiently high an arc light can be forme poles 1 and 2. As the cathode 2 is connected with the negatively charged earth, and therefore always acts as a negative pole, a form of rectification of the a produced by the dynamo 3 is obtained, the second half of the period is always suppressed. The working circuit may be carried out in the same way as in Figur gap 7 may however be dispensed with, and instead thereof between the points m and n a condenser 5 and an induction resistance 9 may be inserted from wh taken inductively.

Figure 36 represents a form of construction similar to Figure 34 only that here instead of an arc lamp a Nernst incandescent body is employed. The Nernst la the battery 3. The working section is connected with the negative pole, the safety spark gap with the + poles. The working spark gap 7 may also be dispe current for it taken at 12 over the oscillation circuit 5, 11 (shown in dotted lines).

Flame collectors (Figure 37) may also be employed according to this invention. The wire network 1 is connected with the aerial collector conductor A and th earth. At the upper end of the latter, long points are provided which project into the flame. The positive electrode is connected with the negative over a co induction coil 9 with the earth.

The novelty in this invention is firstly, the use of incandescent cathodes opposite positive poles which are connected with large metallic capacities as aut surfaces, (2) the connection of the incandescent cathodes with the earth whereby, in addition to the electricity conveyed to them from the battery or machine incandescing, also the negative charge of the earth potential is conveyed, and (3) the connection of the positive and negative poles of the radiating collectors circuit alone or with the introduction of a suitable inductive resistance, whereby simultaneously an oscillatory oscillation circuit may be obtained. The colle these methods quite considerably increased.

I declare that what I claim is: --- [Claims not included here]

British Patent # 157,262
(10 July 1922)

Improvements in Electric Motors

Hermann Plauson // Otto Traun's Forchungs-Laboratorium GmbH

This invention relates to that type of motor in which rotation is produced by means of the attraction and repulsion of surfaces carrying charges of electricity.

According to this invention a stator and rotor are formed of condensor surfaces and charges of electricity thereon imposed in the form of alternating currents of

The invention is more particularly described with reference to the accompanying drawing in which: ---

Figure 1 shows a simple form of motor and feed.

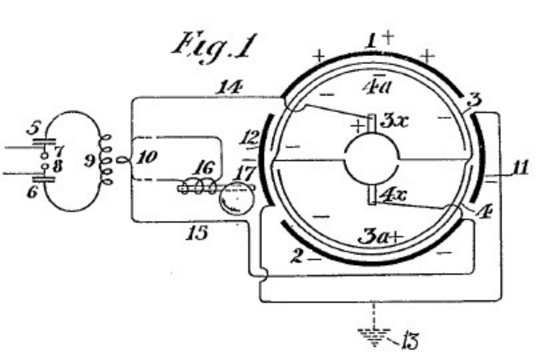
Fig. 1
1+
+
+
14
4a
3
3x
5
7
8
6
9
10
12
16
17
11
4x
4
15
3a+
2
13

Figure 2 is a modification of Figure 1.

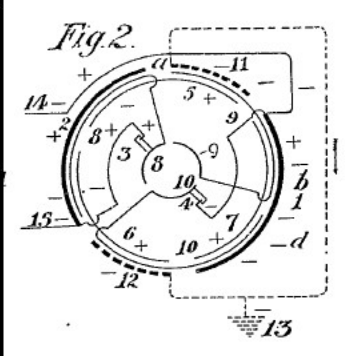
Fig. 2.
a
11
14
2
5+
9
8+
3
8
9
10
4
7
b
1
d
15
6+
10+
12
13

Figure 3 shows one form of a spiral condenser surface.

Figure 4 shows a wire wound condenser surface.

Figure 5 is a diagram of one type of rotor.

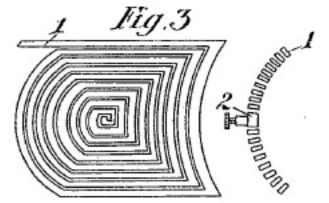
Fig. 3

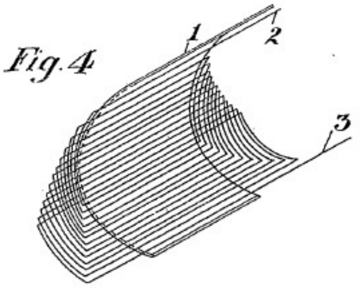
Fig. 4

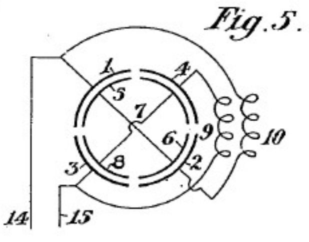
Fig. 5.
1
4
5
7
6
9
10
3
8
2
14
15

The inner plates of the condenser 5 and 6 are charged from a spark gap 7, 8 connected to a source of energy of sufficiently high pressure (alternating or direct
potential has risen so far that a spark springs over.

The spark gap 7, 8 forms with the condenser 5 and self-inductance 9 and condenser 6 a closed oscillatory circuit and alternating currents of high frequency w
this circuit. The high frequency current produced in the primary circuit 9 excite by induction in the secondary circuit 10 currents of the same periodicity.

The improved type of motor is fed by the discharges produced by the induction in the secondary circuit.

Hitherto only Tesla's motor system (shown diagrammatically in Figure 1, 16 and 17) was known for this purpose. The above-mentioned diagram is only shov
the fundamental principle. It has however no practical interest for carrying out large machines by reason of the impossibility of the regulation and the low effici

Now according to this process, all these defects are overcome by the construction of a machine which is applicable for high frequency currents and of a more
nature. The difference between the principle of construction of these motors as compared with those hitherto customary consists in that the motor is not based c
magnetic induction only (as have been all motors hitherto and also Tesla's motors).

It has been fund that the machine constructed according to Figure 1 cannot only be fed directly with static electricity but if it is connected to a source of high fr
will operate.

The applicants call this new type of motors ‘condenser motors’ to differentiate them from hitherto existing types.

The simplest form of construction of such condenser motors is shown in Figure 1, and this motor may be fed with high frequency alternating currents.

At a given moment positive electricity is charged by means of the lead 14 to the stator surface 1 and to the brush 3x (Figure 1). The brush 3x is connect
condenser surface 3, so that both the stator surface 1 and also the rotor surface 3 is charged with positive electricity. The stator surface 1 and the rotor sur
charged with positive electricity and the second rotor surfaces 4 and 4a by brushes 4x with negative electricity, such motors can then be started by providing i
surfaces 11, 12, the earth connection 13 of one of which is broken by a switch (not shown) according to the direction of rotation desires, or alternatively t
started by a separate source of alternating current in a manner similar to the starting of synchronous motors of known construction. After a half revolution of t
3x comes in contact with the second collector surface 4 so that now this surface is connected by the brush 3x with the stator surface 1 and the brush 4x with the
3. Consequently with a reverse direction of current through the second half of the oscillation period all the hereinbefore mentioned effects take place in the
which, however, produces no alteration in the direction of rotation because the dead points between two directions of oscillation are overcome by inertia.

Although this motor is easy to start it can only be employed for small experimental and measuring purposes because the stator and rotor surfaces are made of s
heated by Foucalt (eddy) currents. In spite of its simplicity and its unsuitability for use in practice it must however be regarded as a basic type for technical calc

The condenser motor shown in Figure 2 differs from Figure 1 by the rotor surfaces consisting of six condenser surfaces connected one behind the other in s
connected with three collector surfaces, so that at any one moment only two adjacent collector surfaces come under the two brushes (3 and 4). In its other actic
to Figure 1. The leads 14 and 15 may be connected either to the ends of the secondary coil 10 or directly with the source of energy. The outer thicker line ir
surfaces 1 and 2 (that is to say the unmoving part of the motor), 11 and 12 shown by thick dotted lines means earthed additional poles of the stator, 8, 9, anc
parts of the rotor condenser surfaces which in turn are connected with the collector surfaces 8, 9, and 10. 5, 6, and 7 are the inner parts of the condenser surface
3 and 4 are brushes.

Hitherto stator and rotor surfaces of compact metal have been spoken of. These however become highly heated with eddy currents and hardly yield 10-15% c
examining into such small useful effects it was found that certain forms of metal sections in the stator and condenser surfaces highly increase these. It was t
that if slots or notches be cut in the metal surfaces of the stator and rotor in the form of a spiral, not only was a higher useful effect possible, but also an easier
a regulation could be obtained.

Experiments have shown that by such a form of construction it is possible to build a very useful motor for high frequency alternating currents more partic
undamped nature.

If for example the system of construction of a stator shown in Figure 1, but four polar, be taken and the system of rotor construction shown diagrammaticall
with the form of construction of the condensers of the stator as well as of the rotor according to Figure 3, a condenser motor is obtained which works well in e
high frequency alternating current. It was also observed that the motors in such forms of construction were found to be more sensitive to resonance effects. S
works the best if stator and rotor surfaces have equal capacity and self-inductance so that the windings both in the stator and also in the rotor are in resonance.

A motor constructed according to the foregoing kind is already fully technically applicable. But even these motors have a series of faults, more particularly
construction. For example, the attachment of the spiral condenser surfaces both of the stator and of the rotor sown in Figure 3 are in practice difficult to carry
practice the condenser and stator surfaces are simply wound of wire or bands in the form shown in Figure 4. Such stator and rotor surfaces may, without fur
regarded as electromagnetic poles, although they are not made of iron as is the case in electromagnets. Such machines may be spoken of directly as motors f
alternating currents in which the separate pole surfaces consists of wound induction condenser surfaces of which one is sound on the stator and the other on the

If the coil as shown in Figure 4 be made of well insulated wires the coil can be embedded in insulating material either for the stator or motor surfaces as has a
in the case of ordinary single and multiphase motors. At the same time the possibility is afforded by increasing the number of turns to produce a greater or sm
the self induction co-efficients.

In Figure 5 is shown a modified construction of a rotor for a four pole motor consisting of four condenser surfaces 1, 2, 3, 4, of which 1 and 2 are connected thc
9 coupled with the coil 10. Four inner surfaces 5, 6, 7, 8 are provided of which 5 and 6 are directly connected also 7 and 8 similarly connected.

The pairs of like poles are connected by wires 14 and 15 to the source of energy. By a suitable selection of the values of the reactance and capacity in these c
circuits may be formed.

Having now particularly described and ascertain the nature of our said invention and in what manner the same is to be performed, we declare that what we cla
not included here]

British Patent # 157,263

Process & Apparatus for Converting Static Atmospheric Electrical Energy into Dynamic Electrical Energy of any Suitable High

Hermann Plauson // H.O. Traun's Forschungs-Laboratorium GmbH

Static aerial electricity in the form of direct current can be converted by using spark gaps and with the assistance of oscillatory circuits into dynamic electrical
high number of alternations of a more or less undamped nature and in such form --- either direct or by means of a special kind of resonance or 'condenser moto
utilized for technical purposes as mechanical energy.

For small installations this system may be very well employed; about 100 horsepower may be stated as practical limit. In constructions of larger aggregate diff
the spark gaps however increase considerably. Further it is desirable to convert the accumulated currents of from 100 to 1000 periods which may then be use
types of alternating current machines instead of into electromagnetic waves of a high number of alternations.

In experimenting with condenser motors the construction of which forms the object of British Patent # 157,262 it was observed that the rotor, if one pole of the
connected with the aerials collecting aerial electricity and the other pole with the earth, not only could act as a motor, but if vice versa the rotor connection
interrupted and the rotor caused to rotate by means of another motor, that when the brushes supply an alternating current the periodicity of which is dependent
poles and the revolutions of the rotor. Such an apparatus may therefore be regarded as a transformer of static into dynamic electrical energy.

The invention is more particularly described with reference to the accompanying diagrams in which: ---

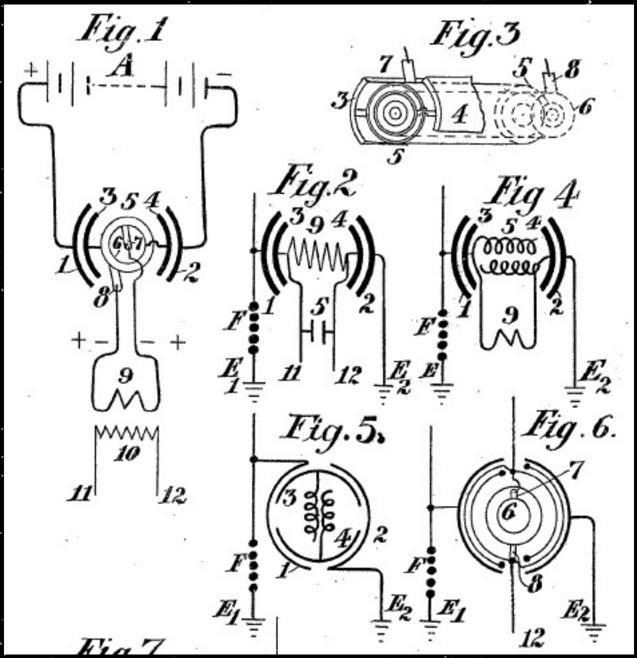

In Figure 1A is a strong accumulator battery, 1 and 2 are the outer poles of the transformer, consisting of simple metallic plates or are as shown in Figures 8-
coils without an electromagnet being present. Between these poles an armature is revolubly mounted on a shaft, which armature also consists of two similar cy
plates 3 and 4. These are metallically connected with two collector rings 5 and 6 on which two brushes 7 and 8 freely run which again are short-circuited with c
primary coil 9. 10 is the secondary coil with the free ends 11 and 12. If through the accumulator battery the stator plate 1 is charged with positive electricity, it
of reverse sign on the rotor surface 3 which is connected by the brushes 7 and 8 over the primary coil 9 with the second rotor surface 4. This latter is there
positive electricity, which in turn induces negative electricity on the stator surface 2/ Up to this moment everything takes place in the same way as if two
connected one behind the other in the current circuit A. If however, by means of mechanical power, this rotor be caused to rotate, the surface conditions ar
quarter revolution the rotor plates are between the stator plates and therefore no condenser surface faces another. By this means however, the capacity of th
reduced to a minimum and a change of current will also result in the main 9. Now if the rotor be turned further through 90 degrees by mechanical energy the ro
opposite the stator plate 2 and the rotor plate 4 opposite the stator plate 1, so that then the rotor pates are in a field of reverse sign. A fresh charge of curr
direction now runs through the primary coil 9. After a further half revolution the same action is repeated so that after a full revolution the initial condition is ag
result of such a revolution is an alternating current the periodicity of which is equal to the number of revolution. In practice of course not two poles but
possible would be employed because thereby the number of alternations would e considerably increased. The primary alternating current thus obtained induces
circuit an alternating current the potential of which is dependent on the winding of the coil. Figure 7 shows a multipolar machine.

If the stator surface 1, instead of being connected with the battery be connected with a collecting aerial network and the other stator surface 2 be directly eart
wich is otherwise constructed as hereinbefore, be rotated by a separate motor a much stronger alternating current results which is to be ascribed to the circums
higher potential can be charged on the pole surfaces of the stator by reason of the higher pressure of the static electricity than where accumulators are employ
the transformer has of course much larger quantities of energy supplied to it.

Figure 2 shows a mode of connections. The stator surface 1 is connected with the aerial antennae which is connected through the safety spark gap F to earth
surface 2 is directly earthed at E2. The inner revoluble rotor surfaces 3 and 4 are interconnected by means of an induction coil which is constructed directly
current is taken as in Figure 1 up to collector rings by means of brushes, which are not shown for the sake of clearness, and further conveyed through the cond
Between these a condenser 5 may be inserted. There is thereby formed a short oscillatory oscillation circuit free from spark gaps, which circuit consists of th
and condenser 5 and is fed by the periodic charging current impulse. By this means the possibility is afforded of obtaining a kind of current which is charac
periods and is undamped and oscillatory. Of course a simple alternating current may be obtained by cutting out the condenser.

Instead of the induction coil the condenser may also b constructed in the rotor. His can be carried out in such a way that its ends serve directly as collecto
current through the brushes. In Figure 3 such a motor is sketched in perspective, 3 and 4 are the rotor surfaces, 5 and 6 are the condenser surfaces constructed t
rotor consisting of two co-axial cylinders fitting one in the other in such a way that free room is left for the brushes 7 on one end of the condenser cylinder 6.

The condenser may be made in the form of a cylindrically wound spiral forming the capacity and reactance as shown in Figure 4. A further type of transfo
Figure 5. The difference consists in the stator and rotor surfaces not only each assuming a quarter of the circuit but almost the half. By this means the space
condenser surface is better utilized. Charge is produced only when the rotor surfaces face the full scope of the stator surfaces.

In addition a condition is obtained in which the stator surfaces are inductively connected by the rotor surfaces. The consequence of this is that an a
simultaneously results which is produced without sparking otherwise the connection is as before.

Figure 6 shows the alternation of the rotor surfaces; the rotor here consists of two cylindrical condenser plates arranged concentrically, each divided int
connected so that half the inner cylinder is connected to half the outer. Such a machine shows the more complete transformer action.

Fig.7.

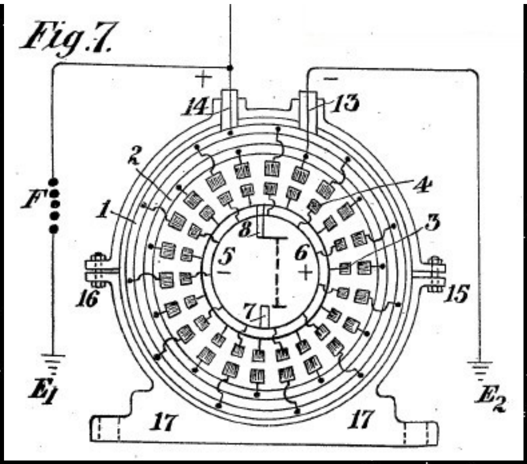

Figure 7 shows a four polar transformer. It consists of a metal casing, the lower half of which is fastened with the foundation plates 17 to the support or foun
half, the cover, is connected by bolts 15 and 16 firmly with the under part. This upper sleeve or casing is insulated from the under part. Two rings 1 and 2
constructed in the casing. The ring 1 is metallically connected with the collector aerial and the ring 2 with the earth. On both rings an equal number of s
mounted side by side but well insulated from one another and thus form an electrostatic field similar to the electromagnetic in many alternating current ma
consists in similar manner of two rings 5 and 6 on which an equal number of rotor surfaces are fixed so that each stator surface faces a rotor surface. By the br
alternating current formed is removed from the collector. The charge is conveyed by the conductor 14 to and by 13 away. If this rotor be then rotated by mea
positive and negative fields precisely as in the case of magnetizing will alter and thereby an alternating current is formed in the rotor, the periods of which are
number of the poles and the revolutions per second.

At the commencement it was thought that this apparatus could only be regarded as alternating current converters, but it was soon found that much more energy
rotate the rotor than might be necessary to overcome the friction. It was then found that the considerable expenditure of energy for rotating the rotor was caus
being moved through strongly electrostatic fields since the electrostatic lines of force must be cut at right angles and that further in the conductors a stronger
was otherwise to be expected. This apparatus must therefore not only be regarded as a transformer, but also as an energy producer, with the difference that the
obtained instead of by means of electromagnets, by static fields of high pressure. The entire system may, to some extent, be compared with a dynamo in wh
takes place by means of a fixed constant magnet. It was further ascertained that this way of using the atmospheric electricity produced a sort of suction on the
and that thus suitable greater quantities of current could be obtained.

The effects which in this apparatus became evident are extremely interesting and open a prospect of being able to obtain here a great deal more. Merely that
made it possible to transform suitable quantities of atmospheric electricity into alternating current of high or lower frequency (without the use of spark gaps)
extreme utility of these apparatus. Should in future, the construction of larger aggregates be necessary the transformer installation may be constructed in such
which are fed by a current obtained from an installation with spark gaps produce a certain quantity of energy which may then be employed for producing cu
the last described system.

The results of the examinations made for this may be construed as follows.

(1) If solid electrodes (condenser surfaces, rotor and stator surfaces) are employed they become hot. This effect may be considerably reduced by cutting the el
form Figure 8, but not entirely removed. This form allows the surface of the condenser plate to be enlarged or increased; the electrodes may be fastened in a
the under frame by perforation 1, 2, 3, 4, 5.

Fig.8

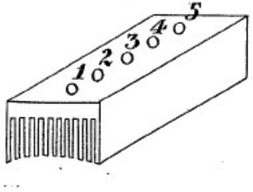

Fig. 9.

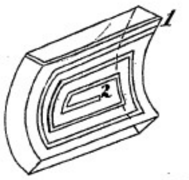

(2) If nicks or notches in spiral form as shown in Figure 9 seen from the side end in Figure 10 in section are employed, not only is the transformer effect gre
yields also more current, but require greater quantities of energy for their movement than a simple commutator action would require.

Fig. 10

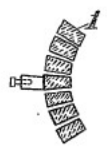

Fig. 11

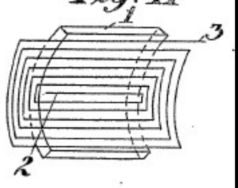

(3) The greatest effect is obtained if the rotor and stator surfaces are wound in flat spiral form of suitably thick wire, and in such a way that the inductive effec
the capacity is calculated in suitable proportion and this result is adapted to a suitable periodicity. In practice this is preferably done by the wire bent in spiral fo
in a separate vulcanite or hard rubber mass (see Figure 11) so that a smooth pole surface is formed similar to that in phase motors.

Fig. 12

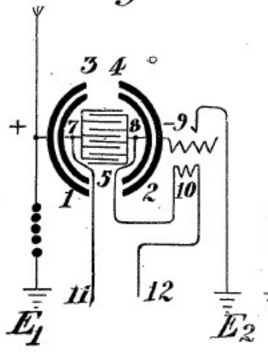

Fig. 13.

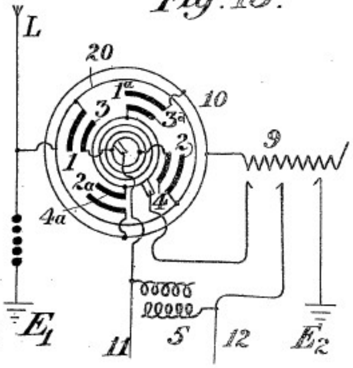

Regular undamped oscillations of a high frequency may however be produced if the converter be carried out in the manner shown in Figure 13. The aerial wir
connected with the ring 20. To this two pole surfaces 1 and 2 are connected. The inductive earth pole is also connected with a second ring 10 from which again
2a are branched off. Of course in similar manner any suitable number of poles may be branched of. In similar manner there are in the rotor two poles fastened
and 4 and 3a and 4a) connected with separate collector rings. From these two rings the current is collected by means of two brushes. The induced alternating c
directly metallically connected with an inductive earth stator conductor over an induction coil 9. Further a combined inductance and capacity 5 is inserted betw
11 and 12 in parallel with the converter. By this means a sparkless oscillatory circuit is obtained which can act on the exciting current in the stator. This pro
periodic alteration of the charging quantities according to the oscillation curves of the rotor currents in consequence of which the stator charge also commenc
oscillations and if the stator and rotor surfaces are calculated to one another in such a way that they are adapted to form oscillations of waves of similar
converter is caused to oscillate and furnishes undamped oscillations of a high number of alternations, but of periodically changed amplitude, the form of whic
the amplitude of the main alternating current and is caused by the number of the poles and revolutions per second. Thus an alternating current of, for examp
formed, the separate periods of which are formed by undamped oscillations of a higher number of alternations. In Figures 14-16 four other diagrams of convert
the object of which is not to produce usual alternating current, but oscillations of high frequency.

Fig. 14

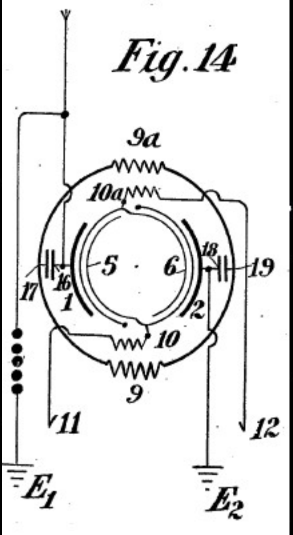

The main difference of these systems from those previously described is that from the connection of the collecting aerials is made between the stator pole 1 (F
pole 16 of the condenser 17 and the earth connection between a second stator pole 2 and the pole 18 of the condenser 19. The other poles of these condensers 1
circuited through a ring over two inductive primary coils 9 and 9a with one another. The secondary coils form the rotor conductors 10 and 10a. The rotor itsel

the manner shown in Figure 6 of two short-circuited plate condensers which may be wound as shown in Figure 11. In similar manner of course the stator sur
formed. The collector rings of the rotor with the two brushes for collecting current are here not shown in order to simplify the drawings. By the conn
condensers in the exiting circuit of the converter and also by the action of the alternating current produced in the rotor on the stator circuit, with a correct
capacity and the self-induction co-efficients a maximum action may be obtained. The kind of current produced will be similar to that described for Figure 12.

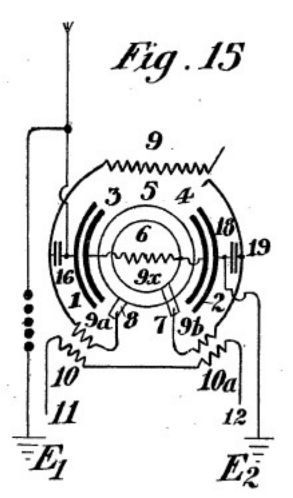
Fig. 15

The novelty of the converter illustrated in Figure 15 consists mainly in that the current resulting in the rotor is not directly employed, but only serves a exci
coils 9a and 9b. The working current is produced in the secondary coils 10 and 10a and further conducted through the conductors 11 and 12. The stator curren
by the regulatable inductive resistance 9 to the same resonance as the rotor current.

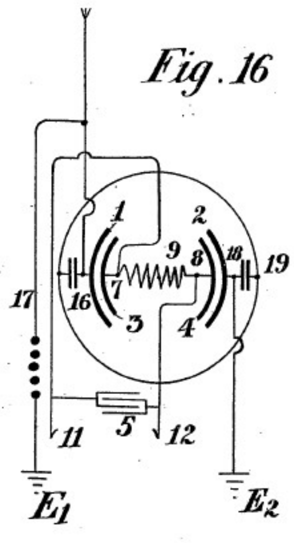
Fig. 16

In Figure 16 a very similar system is shown to Figure 14. The condenser 5 is however connected in parallel with the converter; and by the inductive resistance
rotor a short circuited oscillatory circuit is formed which gives extraordinarily good results and is simple in construction.

The inductive resistance 9 may also instead of being constructed in the rotor be constructed as primary coil employed outside the rotor and short circuit the
over the stator surfaces (see Figure 17).

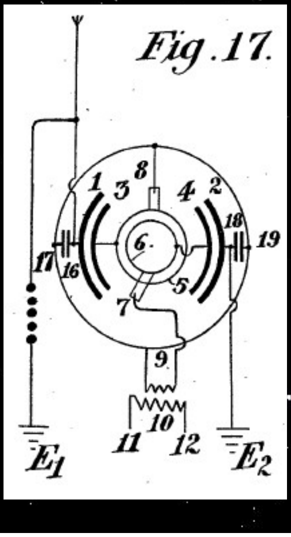
Fig. 17.

The last six types serve only for producing oscillations of a high number of alternations. If it be desired to obtain ordinary alternating current there complica
arrangements are not required as the types illustrated in Figures 1 to 11 suffice. It is self-evident that these arrangements may be altered in various ways by r
condenser surfaces in practice.

Having now particularly described and ascertained the nature of our said invention and in what manner the same is to be performed, we declare that what
Claims not included here]

**If you want to support our research, make an optional donation.
Thanks**

If you want to support our research make a donation Thanks

Donazione

AUTODESK. Make anything. (http://www.autodesk.com)

advertisement

FREE ENERGY MICROWAVE

Technology (/technology/) > Electronics | by GEORGE CHANIOTAKIS (/member/GEORGE+CHANIOTAKIS/)

Follow

4,604 56 5

Posted Dec. 22, 2015 | (cc) BY-NC-SA

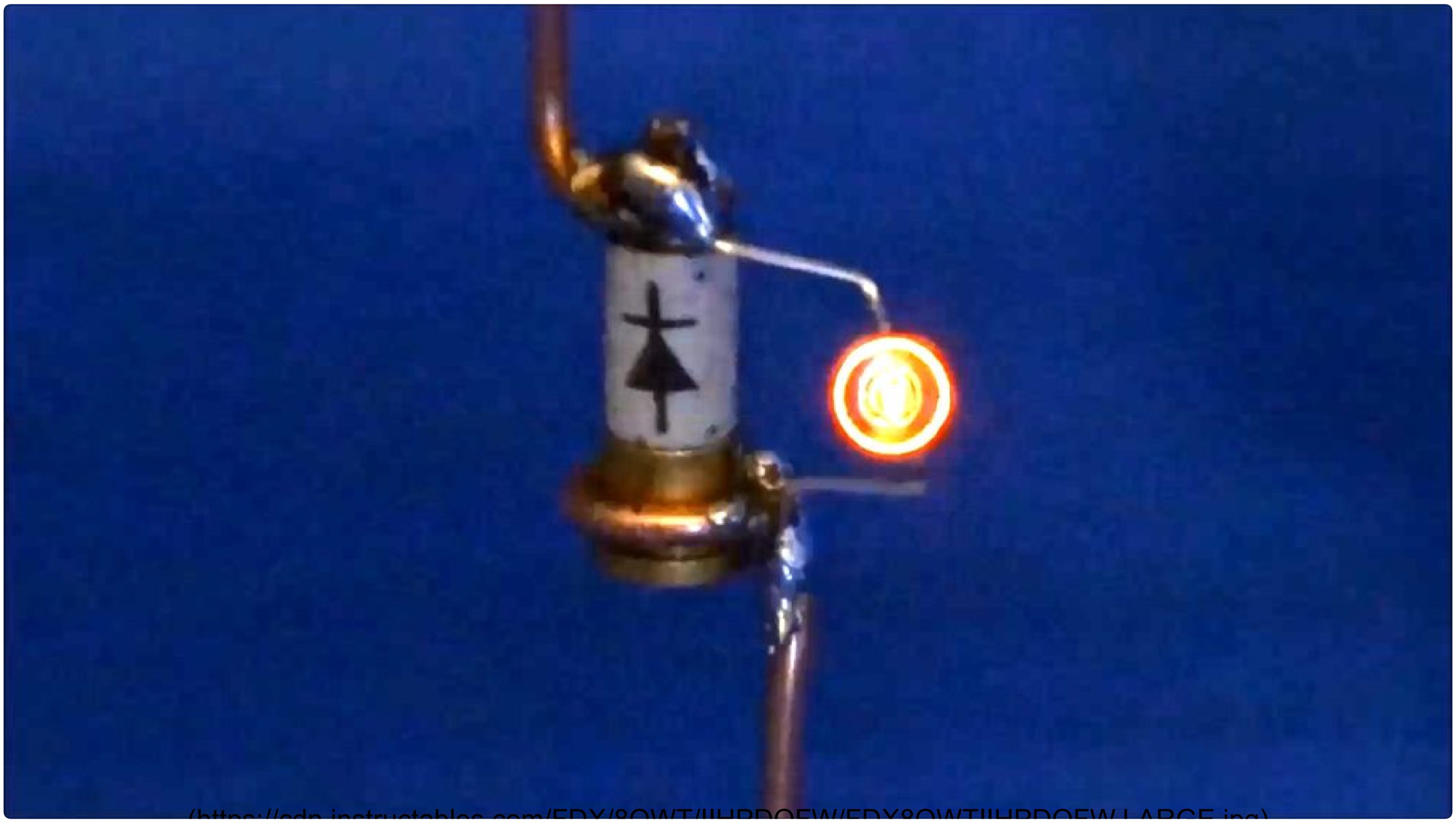

(https://cdn.instructables.com/FDX/8OWT/IIHRDOFW/FDX8OWTIIHRDOFW.LARGE.jpg)

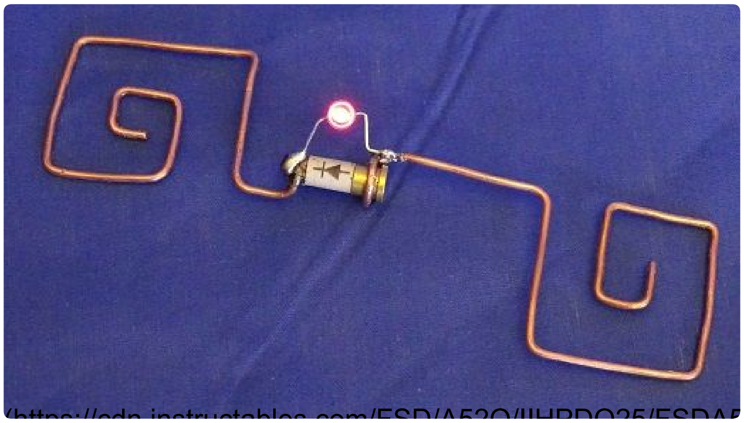

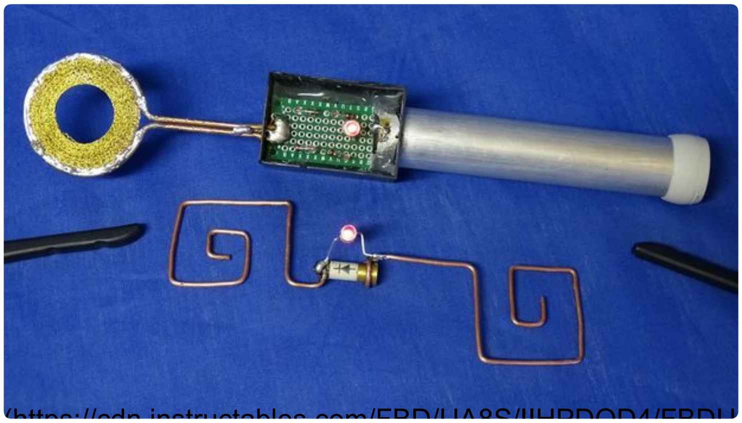

In electronics, a diode is a two-terminal electronic component that conducts primarily in one direction; it has low resistance to the flow of current in one direction, and high resistance in the other. A semiconductor diode, the most common type today, is a crystalline piece of semiconductor material with a p–n junction connected to two electrical terminals. A vacuum tube diode has two electrodes, a plate and a heated cathode. Semiconductor diodes were the first semiconductor electronic devices. The discovery of crystals' rectifying abilities was made by German physicist Ferdinand Braun in 1874. The first semiconductor diodes, called cat's whisker diodes, developed around 1906, were made of mineral crystals such as galena. Today, most diodes are made of silicon, but other semiconductors such as selenium or germanium are sometimes used.

This diode is able to detect RF signals and converts it to DC electricity to power a Led the amazing thing about this is that its completely passive and doesn't require any batteries... Gold Plated Slug 1N23B is Vintage U.S. ARMY Military for General Purpose UHF-MW Silicon Mixer. Primarily used in the Allies’ radar systems during World War II.

Add Tip Ask Question

Step 1:

PRELIMINARY PLANS TO RUN YOUR CAR ON TAP WATER!

IT ALSO WORKS ON YOUR
TRUCK / RV / MOTORCYCLE / AIRPLANE (ETC)

Will This Work?

These plans were sent to the Spirit of Ma'at anonymously, from someone who does not want his or her name printed (for obvious reasons).
We have had them checked by an expert who believes that they are real.
We also have talked with another individual who has patented a similar device, and we know by personal experience that the technology is sound.
So although we cannot guarantee it, we believe these plans will enable you to build a car that runs on water. If you test it out, though, do as the writer suggests and use an old car that doesn't represent a loss of value if you can't make it work. And leave everything intact so that you can always reconnect back to gas if you have to.

But if you do get it working, please send us your experience for our readers. You could be a national hero and help save our country and our world.

We know for certain that an automobile will run on water. So this could be an interesting project for you mechanical types, with a great reward of never having to purchase gasoline for the rest of your life - and helping humanity at the same time.

Re The Need to Rustproof Your Exhaust System

It is possible to make a hybrid of both gas and water (a system that is being tested now in Mexico), which would eliminate the need to open the head and remove the exhaust system. Just a thought. It takes only a small amount of gas to keep the system dry.

The text sent from the anonymous individual was edited slightly for better reading. The following is his/her words and drawings, which has been given into the public domain.
- Spirit of Ma’at LLC

Introduction

It is suggested you try this out to begin with on a second vehicle you own, one that you don't need to live with everyday, until you perfect this technology.

Do-it-yourself plans allow the individual (that's you and me, folks) to make a difference. This is the easiest and lowest-cost way to convert your car to run on (relatively) free energy.

Now, with existing technology, anyone can stand up and make a difference by reducing the local automotive pollution, eliminate gasoline expenses, help restore our atmosphere, and breathe a little easier.

In putting these plans into operation, you will be making use of your entire existing system except for the fuel tank and the catalytic converter.

The Plan

Build and install a low-cost alternative method for running your vehicle (internal combustion engine) on tap water, using off-the-shelf components.

This is simply an efficient way to convert ordinary tap water into gaseous hydrogen and oxygen, and then burn these vapors in the engine, instead of gasoline.

This "minisystem" runs easily from your existing battery and electrical system, and it plugs into your carburetor with simple off-the-shelf fittings.

You will be installing a plastic water tank, a control circuit, a reaction chamber, a hi-pressure carb/FI fitting, and 3 gauges, and then hooking into your existing carb/FI.

The simplicity comes from its being an "on-demand" system requiring no fancy storage or plumbing. You crank the gas pedal or throttle, and you electrically create more vapor for immediate consumption, on demand; low-high flow rate as needed, from idle to maximum power. The only real change is that you are using tap water as fuel, instead of the traditional petroleum-based fuel.

Given a choice, which way would you choose?

Frequently Asked Questions

Q: Does it really work ?
A: Yes; this is well-established technology dating back to stainless steel. But be sure to follow these instructions using the proper mechanical and electrical assembly techniques, as this plan incorporates the best qualities of several techniques.

Q: How does it qualify as "free energy"?
A: If you're paying someone for the water you use, then it is not strictly free. But the alternative is to keep buying into expen$ive ga$oline and its resultant hydrocarbon pollution.

Q: Is it safe?
A: Technically, it is safer than running on fossil fuel because you are no longer choking on your own emissions (health-wise). In general, it is practically as safe as your current gasoline arrangement. You will be installing a few simple safety devices, using current automotive standards.

Q: What kind of performance can I expect?
A: Properly adjusted, your modified vapor-only fuel system will run cooler, and at a modestly higher power level. The mileage performance expected from this design ranges from 50-300 mpg (of water), depending on your adjusting skills.

Q: Can I do the modification myself?
A: Why not? If you don't have any mechanical skills, and you know someone with basic mechanical and/or electrical skills, you can even delegate some of the construction. If you are using a fuel-injected engine, you may have to get a mechanic's opinion. [There will have to be an adapter inserted into the fuel-injection system, just as you would have to do if you were going to run on propane, hydrogen, or natural gas. Ed.)

Q: What is the environmental impact that my vehicle will have?
A: It will be producing H20 steam (water vapor) and unburnt O2 (Oxygen). Hence, it will be cleaning the environment, rather than dumping nauseous toxins into it. Plus you will be helping to save our dwindling supply of atmospheric oxygen. Any excess vapor in the reaction becomes either steam or oxygen. You can also expect to be receiving more than casual interest from those around you.

Q: Isn’t this really a steam engine?
A: No. Really. Exceedingly high temperature and pressure are not used. This is strictly an internal-combustion engine (burning orthohydrogen) with residual steam in the exhaust as a by-product.

There are a few things you should know about gasoline:

Gasoline as a fuel is not necessary; it is optional.

Gasoline versus Water

There is a lot of thermochemical energy in gasoline, but there is even more energy in water. The DOE (Department of Energy) has quoted about 40%, so it is probably much more than that.

Most people are unaware that "internal combustion" is defined as "a thermo-vapor process" - as in "no liquid in the reaction." Most of the gasoline in a standard internal combustion engine is actually consumed, (cooked, and finally, broken down) in the catalytic converter after the fuel has been not-so-burnt in the engine. Sadly, this means that most of the fuel we use in this way is used only to cool down the combustion process, a pollution-ridden and inefficient means of doing that.

How It Works

Exceedingly simple. Water is pumped as needed to replenish and maintain the liquid level in the chamber. The electrodes are vibrated with a 0.5-5A electrical pulse which breaks $2(H_2O) \Rightarrow 2H_2 + O_2$. When the pressure reaches say 30-60 psi, you turn the key and go. You step on the pedal, you send more energy to the electrodes, and thus more vapor to the cylinders; i.e. fuel vapor on demand.

You set the idle max-flow rate to get the most efficient use of power, and you're off to the races.

In the big picture, your free energy is coming from the tap water in an open system, as the latent energy in the water is enough to power the engine and hence drive the alternator and whatever belt-driven accessories. And the alternator is efficient enough to run the various electrical loads (10 - 20 amps), including the additional low current to run this vapor reaction. No extra batteries are required.

STEP BY STEP CONSTRUCTION (Please refer to diagrams at the end of this document)

OVERVIEW - Here is the suggested sequence of steps:

1. Install the CHT (or EGT) gauge and measure your current operating temp range (gasoline), for comparison.

2. Build and test the controller to verify the correct pulse output.

3. Build the reaction chamber and test it with the controller (i.e pressure out).

4. Install the tank, controller, chamber, and pressure fittings.

5. Run engine and adjust the control circuit as necessary for best performance.

6. Install the stainless steel valves and get the pistons/cylinders coated with ceramic.

7. Coat the exhaust system with ceramic without the catalytic converter (or let it rust out and then replace the whole dang thang with stainless steel pipe sections).

YOU WILL NEED

plastic water tank with pump and level sensor.
control circuit, wiring, connectors, and epoxy.
reaction chamber with electrodes and fittings.
3/8" stainless steel flex-tubing, fittings and clamps.
carb/FI vapor-pressure fitting kit. - pressure, CHT (or EGT), & level gauges.
stainless steel valves.
copper mesh junction.
ceramic surface treatment for cylinders & pistons.
stainless steel or ceramic treated exhaust assembly.

BASIC TOOLS

drill, screwdriver and pliers
hole cutter
wire-wrap, solder-iron and clippers
DVM and oscilloscope.

REACTION CHAMBER

Construct as shown in the diagrams. Use a section of 4" PVC waste pipe with a threaded screw-cap fitting on one end and a standard end-cap at the other. Make sure to drill-and-epoxy or tap threads thru the PVC components for all fittings. Set and control the water level in the chamber so that it well submerses the pipe electrodes; yet leave some headroom to build up the hydrogen/oxygen vapor pressure. Use stainless steel wires inside the chamber or otherwise use a protective coating; use insulated wires outside. Ensure that the epoxy perfects the seal, or otherwise lay down a bead of water-proof silicone that can hold pressure.

The screw fitting may require soft silicone sealant, or a gasket; its purpose is to hold pressure and allow periodic inspection of the electrodes. No leaks, no problems. Make sure you get a symmetric 1-5mm gap between the 2 stainless steel pipes. The referenced literature suggests that the closer to 1mm you get, the better. You will want to get your chamber level sensor verified before you epoxy the cap on.

Make your solder connections at the wire/electrode junctions nice, smooth, and solid; then apply a water-proof coating, e.g. the epoxy you use for joining the pipes to the screw cap. This epoxy must be waterproof and be capable of holding metal to plastic under pressure. You will want to get your chamber level sensor verified before you epoxy the cap on.

CONTROL CIRCUIT

The diagrams show a simple circuit to control and drive this mini-system. You are going to make a 'square-pulse' signal that 'plays' the electrodes like a tuning fork; which you can watch on an oscilloscope. The premise given by the literature is: the faster you want do go down the road, the 'fatter' you make the pulses going into the reaction chamber. Duty cycle will vary with the throttle in the vicinity of 90%MARK 10%SPACE (OFF/ON).

There is nothing sacred about how the pulse waveform is generated; there are many ways to generate pulses, and the attached diagrams show a few. The diagram shows the NE555-circuit approach from the referenced patent. The output switching transistor must be rated for 1-5 amps @ 12VDC (in saturation).

Go with a plan that works for you or your friendly neighborhood technoid or mechanic, and go get all the circuit elements from your local electronics store, such as Radio-Shack or Circuits-R-Us, including the circuit board, IC sockets, and enclosure/box.

DigiKey has better selection, service, and knowledge; plus they have no minimum order. Be sure to use a circuit board with a built-in ground plane, and to accommodate room for mounting 2 or 3 of the gauges. Mounting the reaction chamber in the engine compartment will require running a stub to your pressure gauge where you can watch it.

You can easily make 30-gauge wire-wrap connections between the socket pins and thru-hole discrete components having wire leads. Also make sure to get spec sheets on any IC you use. More details of the best circuits to use will be announced pending prototype testing. You will want to get your chamber level sensor verified before you epoxy the cap on.

Throttle Control

If you have a throttle position sensor, you should be able to access the signal from the sensor itself OR from the computer connector. This signal is input to the circuit as the primary control (i.e. throttle level = pulse width = vapor rate).

If you don't have such a signal available, you will have to rig a rotary POT (variable resistor) to the gas linkage (i.e. coupled to something at the gas pedal or throttle cable running to the carb or FI. If you make the attachment at the carb/FI, be sure to use a POT that can handle the engine temp cycles. Don't use a cheezy-cheapy POT; get one rated for long life and mechanical wear; mount it securely to something sturdy and stationary that will not fall apart when you step on the gas.

Control Range. The full throttle RANGE (idle-max) MUST control the vapor rate, i.e. pulse-width (duty). The resistor values at the throttle signal must allow the throttle signal voltage, say 1-4 Volt swing, to drive the VAPOR RATE. You will be using this voltage swing to generate a 10% ON 'square' pulse. The patent implies using a 'resonant' pulse in the 10-250 KHz frequency range; but it is not explicitly stated so.

In this circuit, you will simply tune to whatever frequency makes the most efficient vapor conversion. You will have to get into the specs for each IC you use, to insure you connect the right pins to the right wires, to control the frequency and pulse width. You can use spare sockets to try out different discrete component values. Just keep the ones that are spec-compatible in the circuit, and get the job done.

You crank up the throttle signal and put more electrical energy (fatter pulses) into the electrodes; verify you can get 10% duty on the scope (2 - 100 usec on the horizontal time-base). Your averaging DVM will display the 90%-10% DC voltage across the output transistor (Vce or Vds or Output to Ground). Set and connect DVM in the supply current and measure .5 - 5 amps, without blowing the DVM fuse. Now verify that you got everything you wanted.

Verify your wiring connections using your DVM as a continuity detector. Check your wiring 1 at a time and yellow line your final schematic as you go. You can best use board-mount miniature POTs for anything you want to set-and-forget. The LEDs are there to give you a quick visual check of normal vs abnormal operation of your new creation. You will want to get your chamber level sensor verified before you epoxy the cap on.

CARB/FI CONNECTION

The diagram also shows that fittings are required to the carb/FI l. There are ready-made kits (such as by Impco) available for making your pressure fittings to the carburetor or fuel-injector as the case may be. You will necessarily be sealing the built-in vents and making a 1-way air-intake.

The copper mesh comprises the inadvertent backfire' protection for the reaction chamber. Make sure that all vapor/duct junctions are air-tight and holding full pressure without leakage. Your new 'system' is considered successful and properly adjusted when you get the full power range at lower temp and minimum vapor flow without blowing the pressure safety valve.

CHT (or EGT)

Monitor your engine temp with the CHT (cylinder head temp) or EGT (exhaust gas temp) instead of your original engine temp indicator (if any). Your existing gauge is too slow for this application and will not warn you against overheating until after you have burnt something. Make sure that your engine runs no hotter than in the gasoline arrangement. VDO makes a CHT gauge with a platinum sensor that fits under your spark plug against the cylinder head (make sure it is really clean before you re-install your spark plug (as this is also an electrical ground).

ENGINE/EXHAUST TREATMENT

Get the valves replaced with stainless steel ones and get the pistons/cylinders ceramic-treated ASAP when you have successfully converted and run your new creation. Do not delay as these items will rust, either by sheer use or by neglect (i.e. letting it sit). You could make max use of your current exhaust system by using it with your new deal until it rusts through, then have your mechanic or welder friend to fit a stainless steel exhaust pipe (no catalytic converter is required). But it could be easier and cheaper to send your existing exhaust system out for the ceramic treatment, and then simply re-attach it to the exhaust ports.

GENERAL

1. Do not discard or remove any of the old gasoline setup components, e.g. tank, carb/FI, catalytic converter, unless necessary. Better to always leave an easy way to revert back to something that at least runs, just in case. Some people are leaving their gasoline setup completely intact, and switching back and forth at will, just to have a backup plan.

2. Set your throttle circuit so that you get minimum vapor flow at idle, and maximum vapor flow at full power without blowing the pressure relief valve. In this way, you control how 'lean' your mixture is by the strength of the pulse (i.e. “fatness” at the optimum pulse frequency).

3. If you just don't get enough power (at any throttle setting), it means that you need to (1) change the pulse frequency, (2) change the gap between the electrodes, (3) change the size (bigger) electrodes, or (4) make a higher output pulse voltage (last resort). Always use an output transistor, such as a MOSFET, that is rated for the voltage and current you need to get the job done. OK so you might have to play around with it some. Isn't that where all the Fun is anyhow?

4. If you get any engine knock our loud combustions (not compensated by adjusting the timing), it means that you need to install an additional coil in the chamber, and drive the coil with an additional pulse signal (about 19 Hz on the .1sec time base (see diagram). Here,

you will be slowing down the burn rate just enough so that the vapors burn thru out the power stroke of the piston. Be sure to include a board-mount POT to set the correct strength of this 2nd pulse signal into the coil. This is a stainless steel coil of about 1500 turns (thin wire) that you can arrange like a donut around the center pipe (but NOT touching either electrode), directly over the circular 1-5mm gap. You want no knocking at any power/throttle setting; smooth power only, but also no excess hydrogen leftover from the combustion.

5. Build the canister(s) as tall as you can without compromising your ability to mount them conveniently near the dash panel, or in the engine compartment, as the case may be. This way, you can always make the electrodes bigger, if necessary without undue hardship. Remember that anything in the engine compartment should be mounted in a bullet-proof, vibration and temperature tolerant fashion.

6. If you have to drill a thru-hole for wiring or plumbing thru metal, make sure to also install a grommet for protection against chafing. Always watch your chamber pressure range from IDLE (15-25 psi) - FULL POWER (30-60 psi). Set your safety-pressure relief-valve to 75 psi and make sure it's rated for much higher.

7. Shut OFF the power switch and pull over if there is any malfunction of the system. Your engine will last longest when it still develops FULL POWER+ at some minimum temperature that we are sure you can find, by leaning back the Royal Vapor Flow and/or by making use of the water-vapor cooling technique (see diagram). Keep good mpg performance records, and periodic maintenance/inspection. Keep it clean; save some money; clean the air; heal the planet; happy motoring; tell a friend; enjoy your freedom and self-empowerment.

8. There lacks documented material for perfecting this vapor system thru a fuel injector; there may be some details you will discover on your own as working prototypes progress. For example, you may be restricted to inject the hydrogen/oxygen vapor without any water vapor, as it may rust the injectors. If engine temp and CHT is a problem, then you will want to re-think your plan, e.g. ceramic-coating the injectors. There is always “replacing the FI system with a Carb.”

9. If you install the water-vapor system (for lower operating temp/stress), you will want to lean the mixture (vapor/air) for minimum vapor flow rate to achieve any given throttle position (idle - max). Make sure that you get a minimum flow for IDLE and a modestly sufficient flow for MAX, that does the cooling job without killing the combustion.

10. If you cannot find stainless steel pipe combinations that yield the 1-5mm gap, you can always regress back to alternating plates of +/- electrodes.

11. If you are concerned about the water freezing in your system, you can (a) add some

98% isopropyl alcohol and re-adjust the pulse frequency accordingly; or (b) install some electric heating coils.

12. Do not let ANYONE ever compromise your dream, your freedom, your independence or your truth.

REFERENCES

Stephen Chambers 'Apparatus for Producing Orthohydrogen and/or
Parahydrogen' US Patent 6126794, uspto.gov
Stanley Meyer 'Method for the Production of a Fuel Gas' US Patent 4936961,
uspto.gov
Creative Science & Research, 'Fuel From Water', fuelless.com
Carl Cella “A Water-Fuelled Car” Nexus Magazine Oct-Nov 1996
Peter Lindemann “Where in the World is All the Free Energy”, free-energy.cc <http://www.free-energy.cc/>
George Wiseman “The Gas-Saver and HyCO Series” eagle-research.com <http://www.eagle-research.com/>
C. Michael Holler “The Dromedary Newsletter” and “SuperCarb Techniques”
Stephen Chambers “Prototype Vapor Fuel System” xogen.com <http://www.xogen.com/>

COMMON LAW COPYRIGHT #285714: All rights to the use and duplication of these plans are hereby reserved for the People, in their efforts to heal and restore the environment. Dare to express your uniqueness and environmental ideals. This technology is an exercise in responsible self-determination.

DISCLAIMER: The Spirit of Ma'at LLC and the Spirit of Ma'at ezine and the author of this document assumes no liability for the use or misuse of this information; which is made available as public-domain information and free of charge, for the purposes of education, ecology, health, well-being, freedom, liberty, and pursuit of happiness.

CONVERT YOUR ENGINE TO BURN HYDROGEN AND OXYGEN VAPOR FROM WATER....... ON DEMAND + POLLUTION-FREE

FIGURE 1

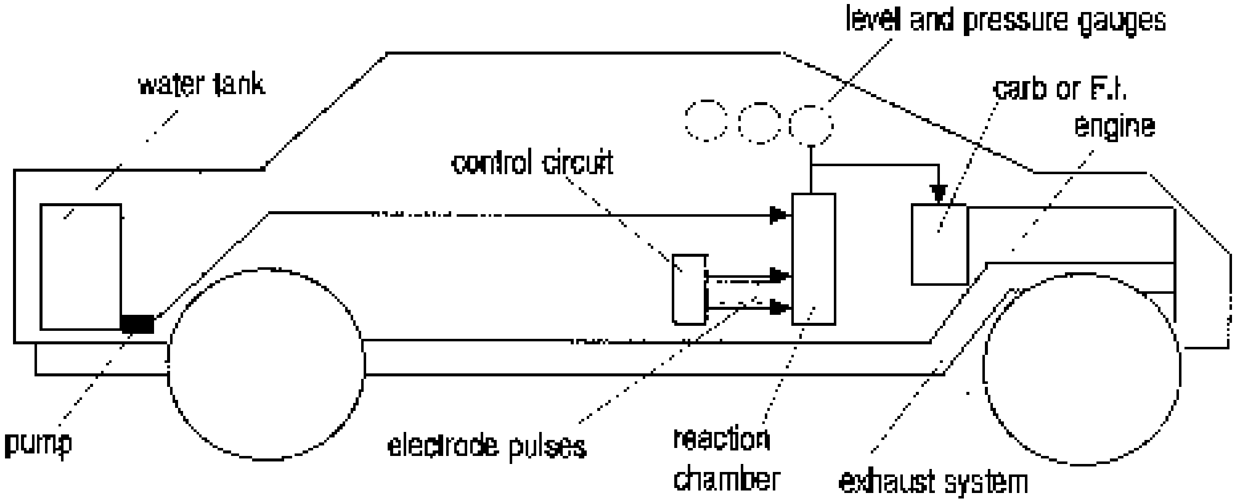
level and pressure gauges
water tank
carb or F.I.
engine
control circuit
pump
electrode pulses
reaction
chamber
exhaust system

BUILD YOUR OWN WATER VAPOR SYSTEM
CONTROLLER FROM THE-SHELF MATERIALS

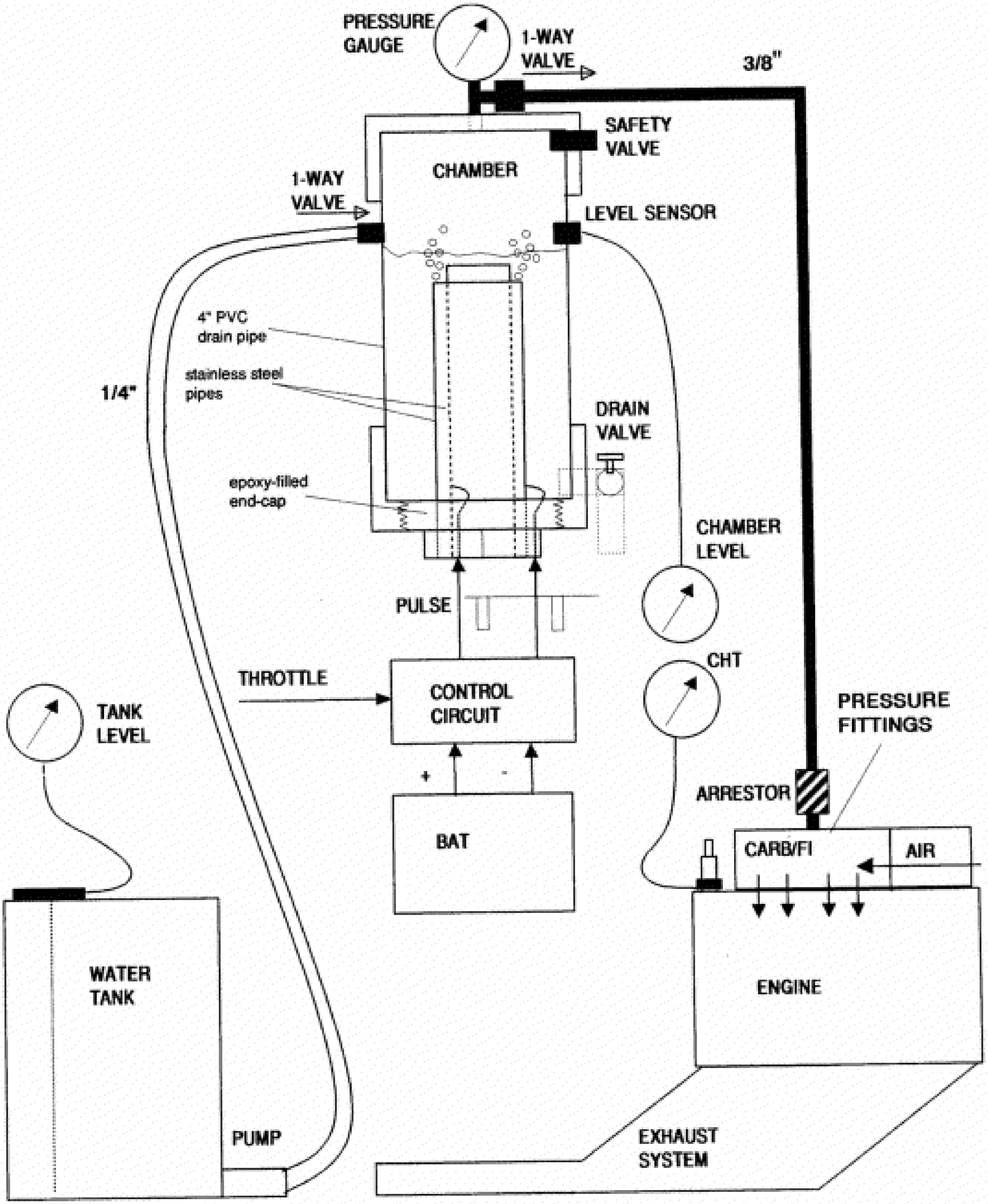
PRESSURE GAUGE
1-WAY VALVE
3/8"
SAFETY VALVE
CHAMBER
1-WAY VALVE
LEVEL SENSOR
4" PVC drain pipe
stainless steel pipes
1/4"
DRAIN VALVE
epoxy-filled end-cap
CHAMBER LEVEL
PULSE
CHT
THROTTLE
CONTROL CIRCUIT
PRESSURE FITTINGS
TANK LEVEL
+
-
ARRESTOR
BAT
CARB/FI
AIR
WATER TANK
ENGINE
PUMP
EXHAUST SYSTEM

FIGURE 2 BLOC DIAGRAM

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

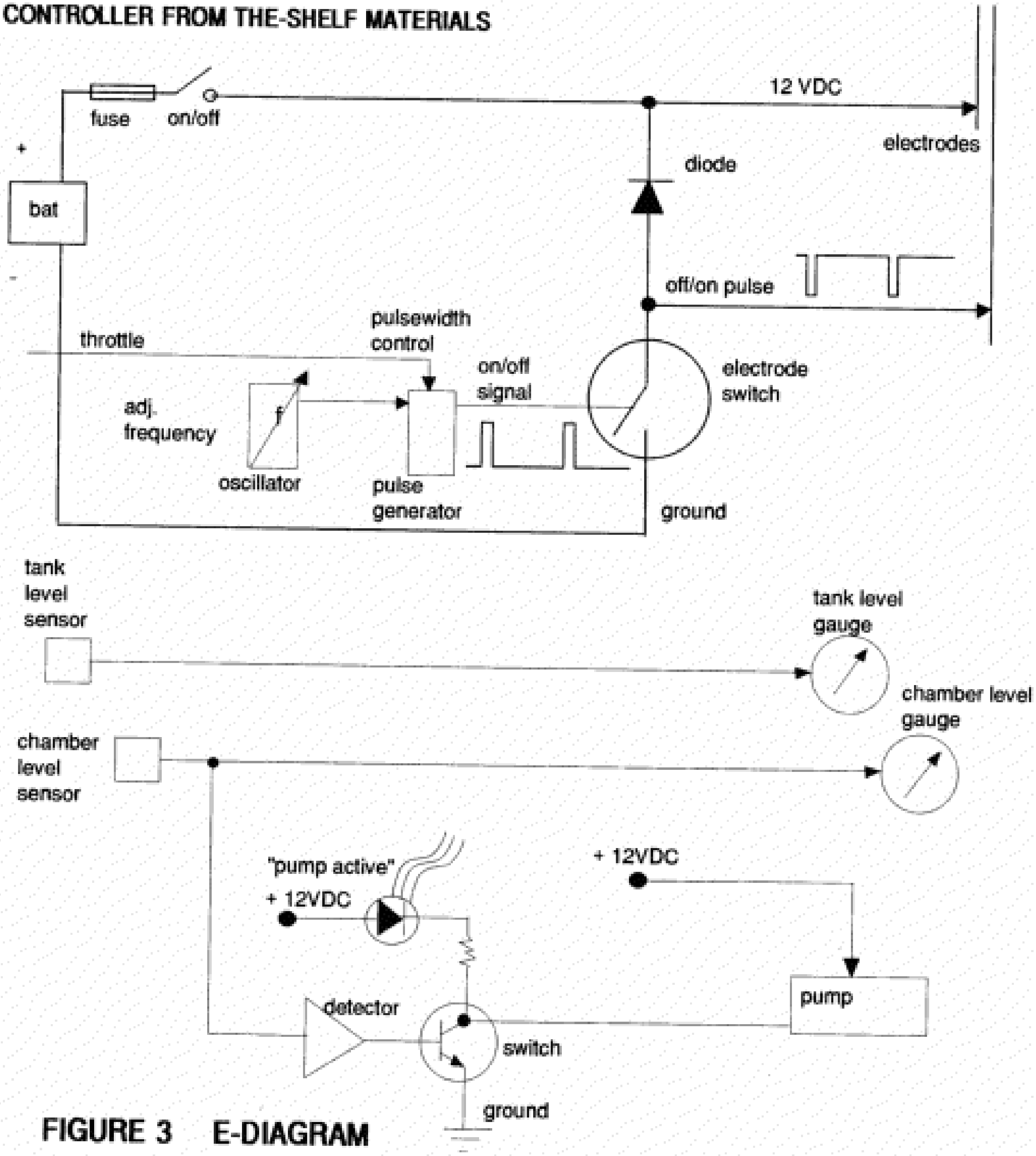
12 VDC
fuse
on/off
+
bat
-
electrodes
diode
off/on pulse
throttle
pulsewidth
control
on/off
signal
electrode
switch
adj.
frequency
f
oscillator
pulse
generator
ground
tank
level
sensor
tank level
gauge
chamber level
gauge
chamber
level
sensor
"pump active"
+ 12VDC
+ 12VDC
detector
switch
pump
ground

FIGURE 3 E-DIAGRAM

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

REV. A

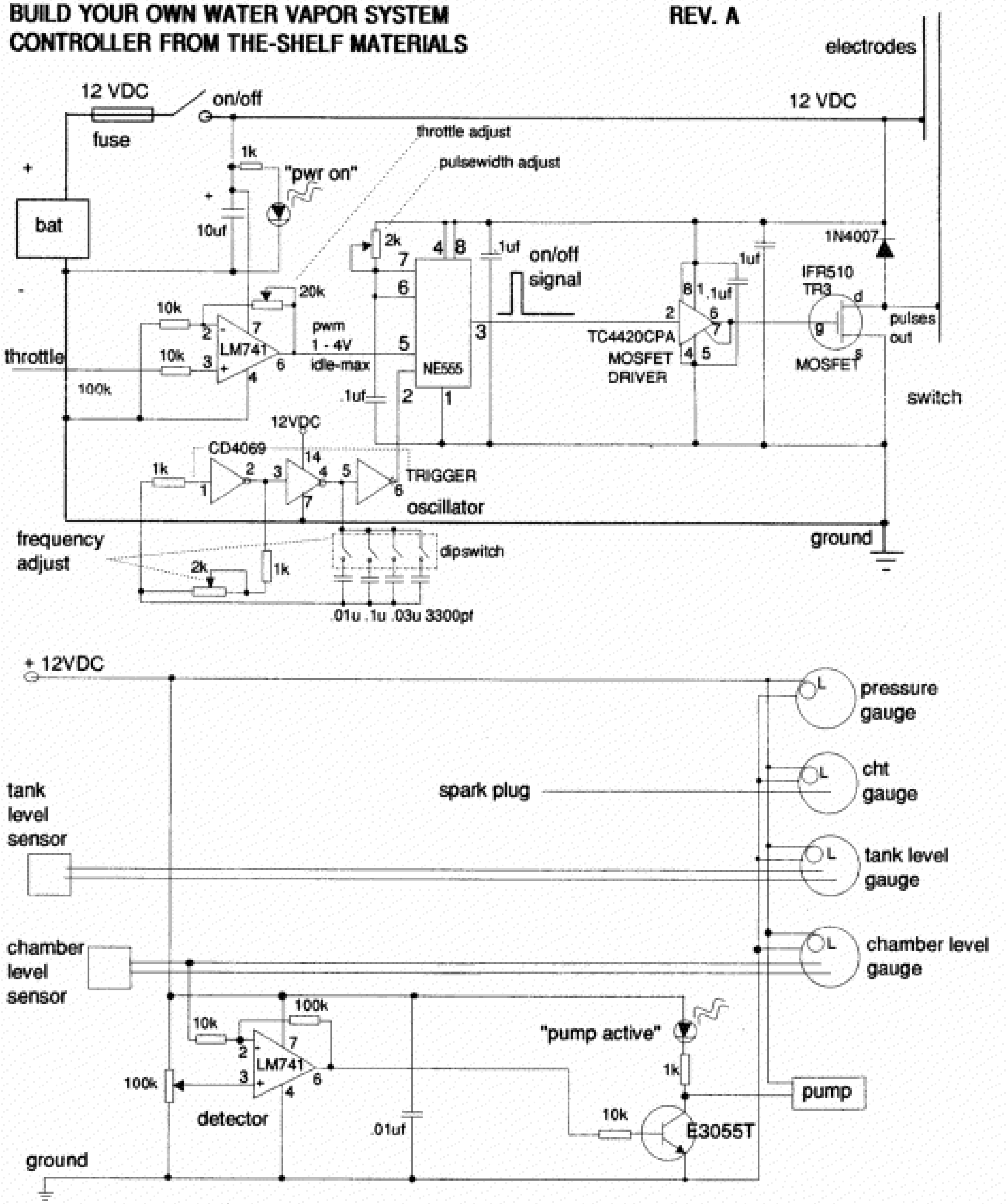
electrodes
12 VDC
on/off
fuse
12 VDC
+
bat
-
1k
"pwr on"
+
10uf
throttle adjust
pulsewidth adjust
2k
7
4
8
.1uf
on/off
signal
6
20k
10k
2
-
7
LM741
pwm
1 - 4V
idle-max
5
3
throttle
10k
3
+
6
4
100k
NE555
.1uf
2
1
1N4007
.1uf
8
1
.1uf
IFR510
TR3
d
2
6
7
TC4420CPA
MOSFET
DRIVER
4
5
g
pulses
out
MOSFET
s
switch
12VDC
CD4069
14
1k
2
3
4
5
1
7
6
TRIGGER
oscillator
ground
frequency
adjust
2k
1k
dipswitch
.01u
.1u
.03u
3300pf
+ 12VDC
L
pressure
gauge
L
cht
gauge
spark plug
tank
level
sensor
L
tank level
gauge
chamber
level
sensor
L
chamber level
gauge
100k
10k
2
7
LM741
"pump active"
100k
3
+
6
4
1k
pump
detector
.01uf
10k
E3055T
ground

FIGURE 4 E-SCHEMATIC

BUILD YOUR OWN WATER VAPOR SYSTEM
CONTROLLER FROM THE-SHELF MATERIALS

REV. A

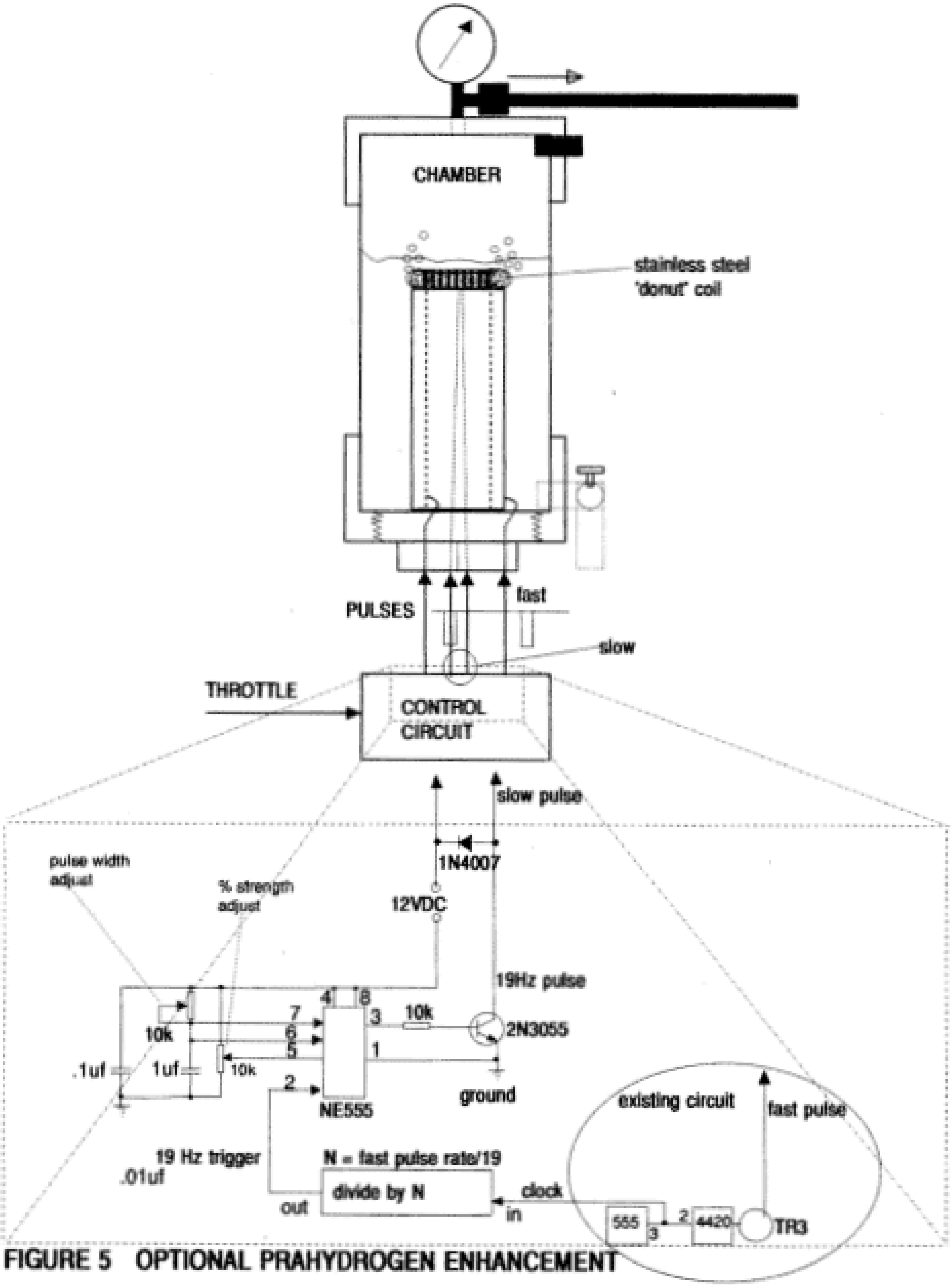
CHAMBER
stainless steel
'donut' coil
PULSES
fast
slow
THROTTLE
CONTROL
CIRCUIT
slow pulse
1N4007
12VDC
pulse width
adjust
% strength
adjust
19Hz pulse
4
8
7
6
5
3
1
2
10k
10k
10k
2N3055
.1uf
1uf
ground
NE555
19 Hz trigger
.01uf
N = fast pulse rate/19
divide by N
out
in
clock
existing circuit
fast pulse
555
3
2
4420
TR3

FIGURE 5 OPTIONAL PRAHYDROGEN ENHANCEMENT

BUILD YOUR OWN WATER VAPOR SYSTEM CONTROLLER FROM THE-SHELF MATERIALS

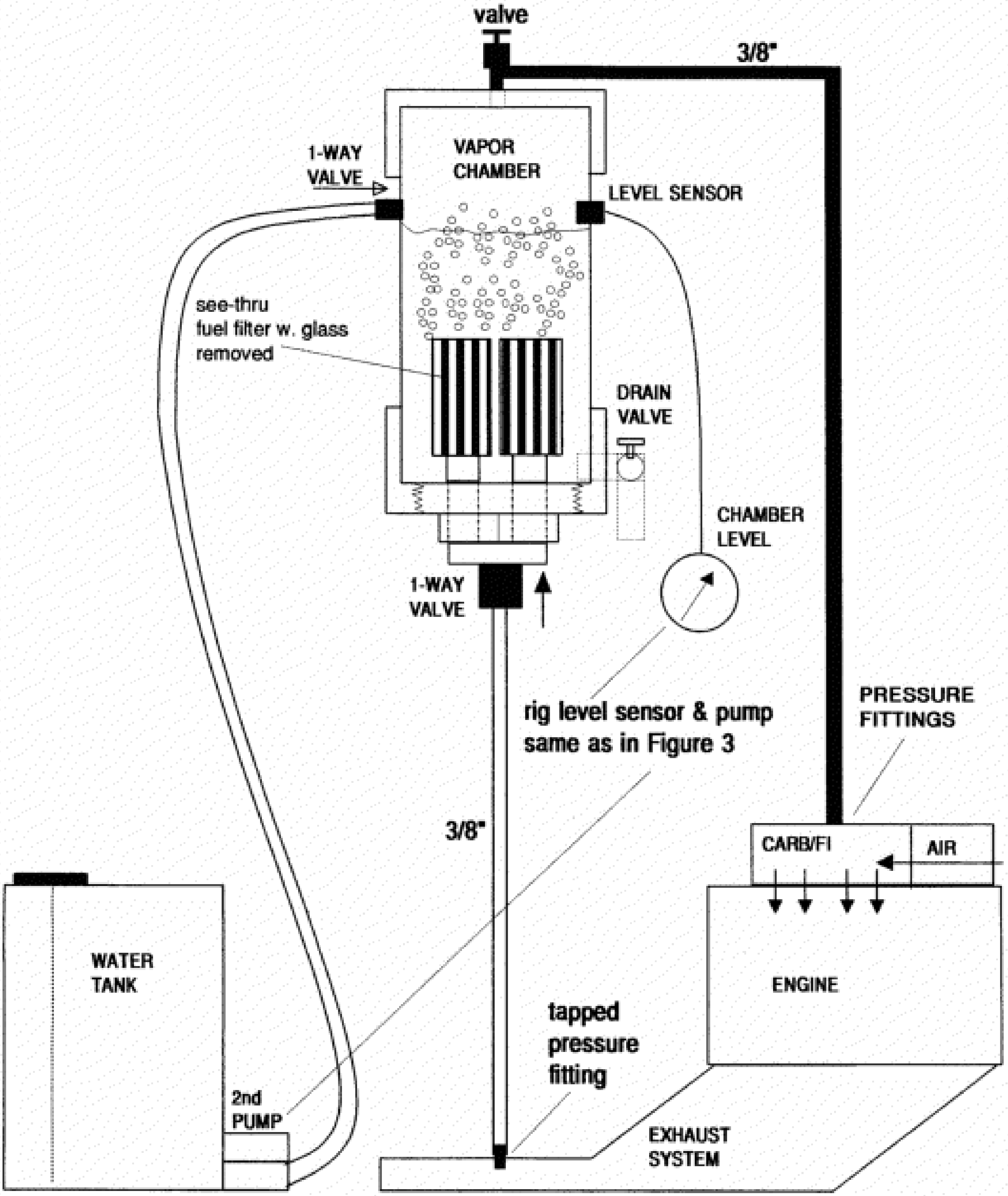
valve
3/8"
VAPOR CHAMBER
1-WAY VALVE
LEVEL SENSOR
see-thru fuel filter w. glass removed
DRAIN VALVE
CHAMBER LEVEL
1-WAY VALVE
rig level sensor & pump same as in Figure 3
PRESSURE FITTINGS
3/8"
CARB/FI
AIR
WATER TANK
ENGINE
tapped pressure fitting
2nd PUMP
EXHAUST SYSTEM

FIGURE 6 OPTIONAL WATER COOLING

Capacitors we used are electrolytic rated at 400 volts x 47 uF put in series to equal 6,000 volts, the diodes we used were silicon 1000 volt 2 amp placed in series to equal 6,000 volts......... Ground was connected to laboratory wall out let ground.

If you decide to try Tesla's experiment by pumping DC into the ground be careful, I tried this and it does work but is very dangerous to you or your neighbors. If someone is taking a shower or using water they can get killed or shocked. do this experiment far away from humans and animals. you can get far more energy out than you put in. I will not tell you much more because it is such a dangerous experiment.

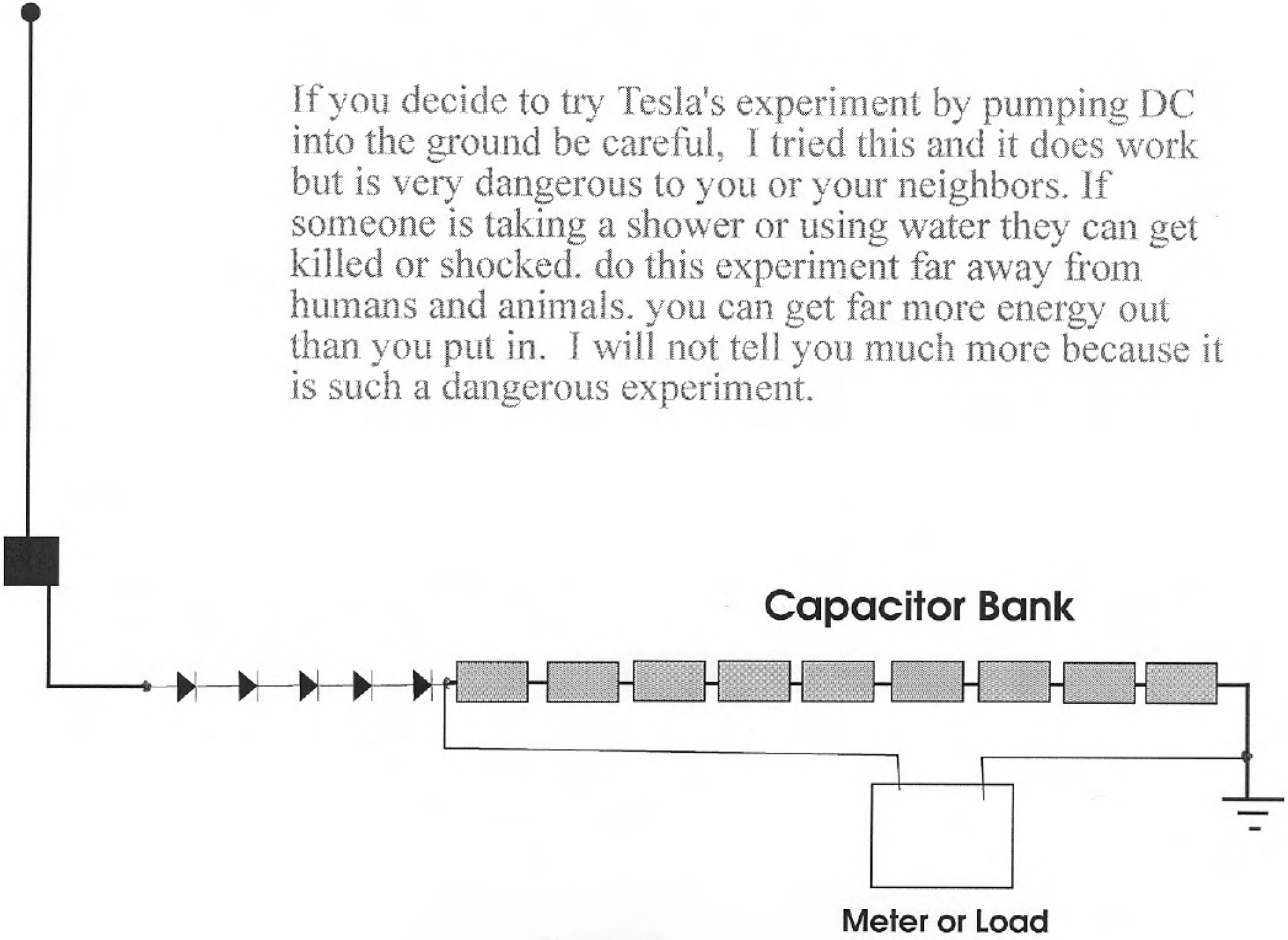

RADIO SCIENCE Journal of Research NBS/USNC–URSI
Vol. 69D, No. 8, August 1965

Schumann Resonances[1]

Janis Galejs

Applied Research Laboratory, Sylvania Electronic Systems, a Division of Sylvania Electric Products Inc.,
40 Sylvan Road, Waltham, Mass. 02154

(Received February 3, 1965)

This paper reviews the current work in the field of Schumann resonances and discusses the topics of waveforms and frequency estimates, resonance frequencies and Q factors, source distributions and noise spectra, diurnal and seasonal changes of power spectra, and variations of resonance frequencies. Observations of the above phenomena are correlated with theoretical considerations.

1. Introduction

The resonances in the spherical shell between the earth and ionosphere that are due to azimuthal waves have been considered in two recent survey papers [Galejs, 1964a; Wait, 1964a]. This work contains experimental data as well as the required theory. The basic theory of the resonances has been reviewed by Wait [1964b]. In the present summary the emphasis will be on correlation between measurements and available theory. There will be no extensive analytical developments, but frequent references will be made to literature.

Based on theoretical considerations, Schumann [1952a, b, 1957] postulated resonances of the earth-to-ionosphere cavity. Koenig [1958, 1959, 1961] obtained the first experimental indication of Schumann resonances by observing noise waveforms in the output of a narrow band amplifier. Detailed frequency spectra of this noise were first obtained by Balser and Wagner [1960]. Other measurements have been reported by Fournier [1960]; Benoit and Houri [1961, 1962]; Lokken et al. [1961, 1962]; Polk and Fitchen [1962]; Gendrin and Stefant [1962a]; Balser and Wagner [1962a, b, 1963]; Rycroft [1963]; Chapman and Jones [1964].

The fundamental theory of Schumann resonances [Schumann 1952a, b, 1957] is discussed in a book by Wait [1962]. Raemer [1961a, b] considers the observable noise spectra as the response of the earth-to-ionosphere cavity due to lightning flashes all over the world, but the homogeneous sharply bounded ionosphere model of Raemer introduces high losses, and he does not succeed in reproducing the spectral measurements of Balser and Wagner [1960]. Galejs [1961a, b] uses an isotropic ionosphere model of exponentially increasing conductivity which is based on measured or calculated characteristics of the lower ionosphere. This model permits a close reproduction of the noise spectra measurements in the resonance region, and it also provides an agreement with measured ELF attenuation rates [Chapman and Macario, 1956; Jean et al., 1961].

Models of an anisotropic ionosphere are difficult to apply to the problem of earth-to-ionosphere cavity resonances because of the variations of the magnetic field vector along the surface of the earth. The propagation parameters can be estimated in the presence of a horizontal magnetic field component from the work of Galejs and Row [1964] and of Galejs [1964b]. The earth-to-ionosphere cavity resonances have been analyzed for a nonhomogeneous ionosphere in the presence of a radial (vertical) magnetic field by Thompson [1963] and Galejs [1965]. A thin-shell approximation of the ionosphere with a superimposed radial magnetic field has been considered by Wait [1964c]. Some other work of Wait [1963a, b], although intended for a different frequency range, can also be applied to the earth-to-ionosphere resonance problem in the presence of a radial magnetic field.

Additional references are available in the comprehensive bibliography by Brock-Nannestad [1962].

2. Waveforms and Frequency Estimates

In wide band recordings of atmospheric signals an ELF component or slow tail follows the initial VLF component of the transient [Hepburn and Pierce, 1953; Liebermann, 1956a, b; Tepley, 1959; Pierce, 1960]. An oscillatory structure of the signal becomes noticeable by passing it through a band-pass filter. An example of a noise recording in the output of a 2 to 30 c/s band-pass filter [Lokken, 1964, private communication] is shown in figure 1. The two horizontal magnetic field components have been observed simultaneously in Canada and in Antarctica. There is a correlation between the high intensity ELF noise bursts observed at the various sites. A recording of one of the waveforms is shown in figure 2. The darkened areas indicate resonance frequencies around 8, 14, and 20 c/s. Noise waveforms in the 5 to 20 c/s band have been reported by Polk and Fitchen [1962] and Polk [1962]. Figure 3 shows sample recordings of the outputs of two coils the axes of which are oriented in the north to south (N–S) and east to west (E–W) directions. Frequency of these waveforms has been estimated by counting the cycles between the second markers shown on the records. The average frequency computed in this procedure should be higher than the resonance frequencies of a spectral analysis [Galejs, 1962a, 1964a]. Noise recordings made simultaneously in Germany and in Rhode Island have been reported by Keefe, Polk, and Koenig [1964].

---

[1] Paper presented at the ULF Symposium, Boulder, Colo., Aug. 17–20, 1964.

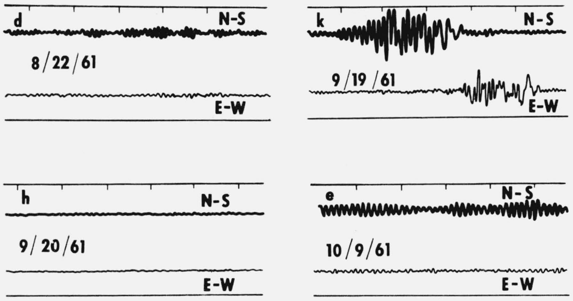

FIGURE 3. *Noise recording in the 5 to 20 c/s band.*

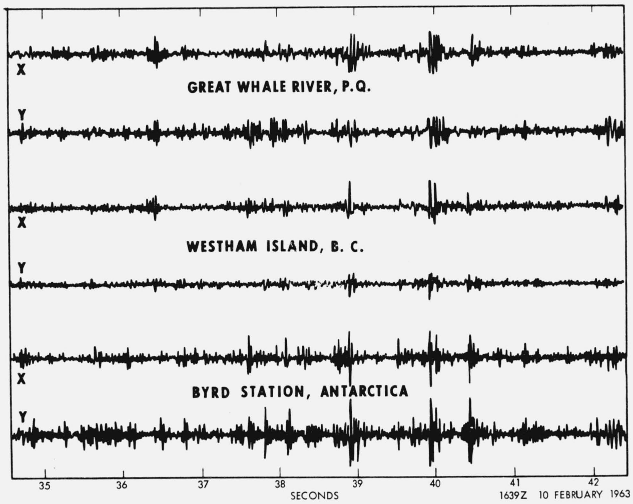

FIGURE 1. *Noise recording in the 2 to 30 c/s band.*

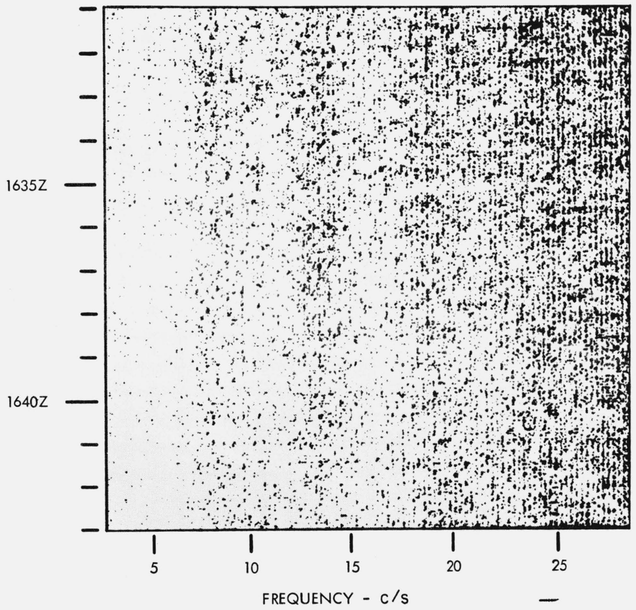

FIGURE 2. *Recording of the noise waveform X of Westham Island, B.C.*

3. Resonance Frequencies and Q Factors

3.1. Definitions

A resonance frequency and Q factor may be determined from spectral measurements. Analytically it is also possible to compute the power spectrum of the noise waveform. However, it is more convenient to relate the resonance frequency and the Q factor to the analytically determined propagation parameters. The radial electrical field due to the vertical dipole excitation can be represented by

$$E_r = -\frac{iIds}{4\omega\epsilon ha^2}\frac{\nu(\nu+1)}{\sin\nu\pi}P_\nu(-\cos\theta) \tag{1}$$

where

$$\nu(\nu+1) = (k_0aS)^2. \tag{2}$$

For a wave that propagates in the spherical shell between the earth and the ionosphere S is defined as the ratio of the complex wave number k to free space wave number $k_0(S = k/k_0)$. The real part of S can be seen to be inversely proportional to the phase velocity, and the imaginary part of S is proportional to the attenuation constant. This parameter S is determined from solving the appropriate boundary value problem of the earth, the air space and the chosen ionosphere model [Wait, 1962; Galejs, 1964a]. The resonance frequencies of the nth mode are determined from the minima of $\sin\nu\pi$ in (1) [Galejs, 1962a, 1964a] which gives

$$f_n \approx \frac{7.5\sqrt{n(n+1)}}{\mathrm{Re}S}. \tag{3}$$

Typical values of ReS are 1.4 to 1.2 for frequencies in the resonance region.

The cavity Q may be determined as a ratio between the stored energy and energy loss per cycle or simply by the width of the resonance curves. As shown in the appendix, in the present problem the electrically stored energy W_e differs from the magnetically stored energy W_h, and the Q definition should consider the sum of the average stored energies W_e and W_h. This expression for Q is

$$Q=\frac{(\mathrm{Re}S)^2+1}{4\mathrm{Re}S\,\mathrm{Im}S}. \tag{4}$$

When considering only the magnetically stored energy Q is given by

$$Q=\frac{1}{2\,\mathrm{Re}S\,\mathrm{Im}S}. \tag{5}$$

For a given set of propagation parameters this will give a lower Q figure than (4). This expression has been used in the past work of Galejs [1961a, 1962a, b, 1964a]. Considering only the electrically stored energy gives

$$Q=\frac{\mathrm{Re}S}{2\,\mathrm{Im}S}. \tag{6}$$

The last expression can also be derived from the concept of complex resonance frequencies [Wait, 1964a] or by considering the half-power band width of the resonance curve. All three Q definitions yield the same results as ReS approaches unity, but they will differ near the lower resonance frequency of the earth ionosphere cavity. Equation (4) will be used in the present calculation.

Experimentally determined Q values that are estimated from the half-power bandwidths of the resonance curves neglect the effects of adjacent resonances and of near-field noise that will tend to add to the background level. Hence, the half-power level is estimated too low, the apparent half-power bandwidth is larger and the estimated Q factor may be too low.

3.2. Characteristics of the Boundary

In the theoretical models that are advanced for explaining the resonance frequencies and the Q-factor measurements [Wait, 1962; Galejs, 1964a; Wait, 1964a] it is necessary to consider the boundary properties of the spherical shell.

In the frequency range of the resonances the displacement currents of ground are negligible and the ground conductivity $\sigma_g=10^{-3}$ to 1 mho/m is much higher than the effective ionospheric conductivity σ_i that is estimated to be of the order of 10^{-7} to 10^{-6} mho/m. The ground conductivity σ_g may be assumed as infinite in the first approximation and only the ionospheric properties require a detailed analysis.

The ionosphere acts as an anisotropic conductor and for a z-directed magnetic field its conductivity tensor is of the form

$$[\sigma]=\begin{bmatrix}\sigma_1 & -\sigma_2 & 0\\ \sigma_2 & \sigma_1 & 0\\ 0 & 0 & \sigma_0\end{bmatrix}. \tag{7}$$

Equations for computing the conductivity components are available in literature [Galejs, 1964a and b] and typical conductivity profiles are shown in figure 4 but some uncertainties still apply to such models [VanZandt, 1964]. The above computations consider the ion effects, but assume the operating frequency to be zero. These profiles can be further approximated by straight lines through the lower ionosphere in the semilogarithmic plots of figure 4, which corresponds to an expoential height variation of the conductivity components. In a further simplification the anisotropy may be ignored ($\sigma_1=\sigma_0$, $\sigma_2=0$). Even a simple one or two layer model can be assumed, such as shown in figure 5. These particular models have been used in past work by Raemer [1961a and b], Galejs [1962a and b], and by Chapman and Jones [1964]. In the isotropic exponential model the conductivity is given by

$$\sigma(z)=\sigma(z_0)\exp\left[\beta(z-z_0)\right] \tag{8}$$

where $z_0=60$ km and the day and night models are characterized by $\sigma(z_0)=4.63\times10^{-8}$ mho/m and $\beta=0.308$ km^{-1} and by $\sigma(z_0)=6.5\times10^{-10}$ mho/m and $\beta=0.44$ km^{-1}, respectively.

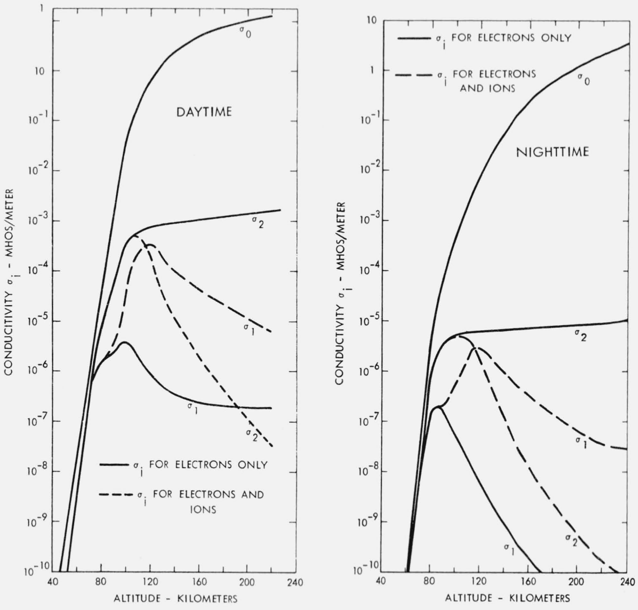

FIGURE 4. *Components of the tensor conductivity* F=0.

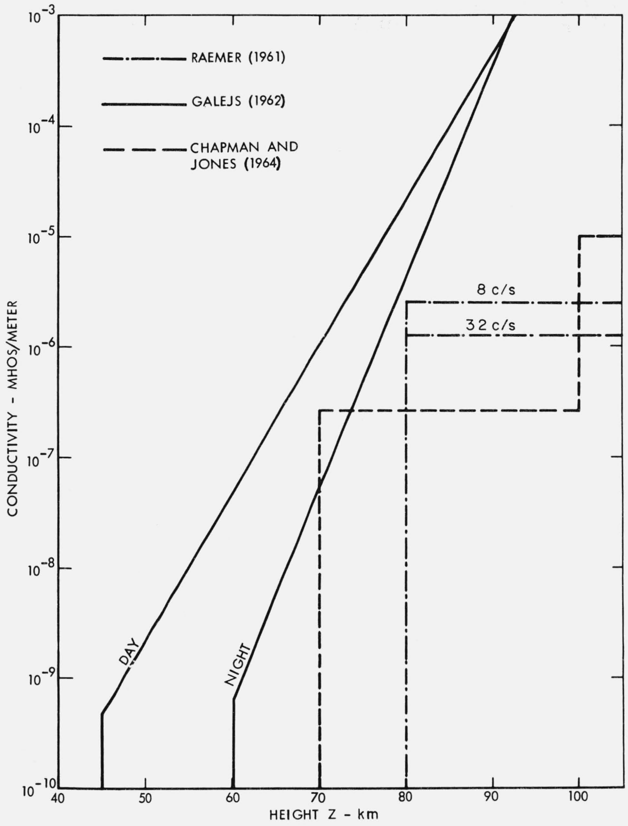

FIGURE 5. *Models of ionospheric conductivity.*

3.3. Single Layer Ionosphere

The single layer ionosphere model of Raemer [1961a and b] assumes the boundary height of the ionosphere to be constant, but the conductivity to change with frequency so as to obtain the correct resonant frequencies. However, the resulting Q factors are approximately two times lower than in measurements. The same difficulties are experienced with an ionosphere model of constant conductivity, but frequency dependent boundary height. Similar results are obtained by Madden [1961] with a conductivity of $\sigma = 1.8 \times 10^{-6}$ mho/m and the ionospheric boundary at $h = 80$ km. The deficiencies of a single layer ionosphere model are less pronounced if it is attempted to achieve only a partial match of the resonant frequencies, as seen from Wait [1964a]. However, the Q factors will be too low in particular at the higher resonance frequencies ($f \approx 30$ c/s).

Compressibility effects of the ionosphere have been considered by Seshadri [1965b] using the model of a perfectly conducting flat earth and a sharply bounded homogeneous isotropic and lossless ionosphere. The compressibility of the ionosphere does not affect the wave propagation below it for frequencies in the range of Schumann resonances and in the ELF band. However, the compressibility should be considered for waves that propagate inside the ionosphere [Seshadri, 1965a].

3.4. Two-Layer Ionosphere Models

Two-layer models give more accurate values of the resonance frequencies and of cavity Q or of the complex propagation parameter S. Two-layer and multilayer models have been discussed by Wait [1958, 1962]; Madden [1961]; Jean et al. [1961], and a particularly successful two-layer model has been constructed by Chapman and Jones [1964a, b]. This model, which is indicated in figure 5, has been applied in the frequency range of earth-to-ionosphere cavity resonances and also for frequencies up to 1 kc/s.

Such models are usually derived by trial and error procedures and they are difficult to correlate with continuous ionosphere profiles, such as shown in figure 4.

3.5. Models of Continuously Varying Ionosphere

The isotropic exponential model gives a reasonable approximation to the ionospheric conductivity versus height profile through the lower ionosphere layers, and the propagation parameters of the model can be determined from closed form expressions of the surface impedance [Galejs, 1961a, 1962a, b, 1964a]. Calculated resonance frequencies and Q factors that are based on (3), (4), and available values of S [Galejs, 1962a] are listed in tables I and II. Calculations are based on daytime or day and nighttime ionosphere models. The Q factors considered also the effects of cosmic ray ionization [Bourdeau et al., 1959] that give a nearly exponential conductivity versus height variation as indicated by Galejs [1962a]. Cosmic ray ionization decreases the effective Q figures, in particular for the lower frequencies. The calculated Q values are still somewhat higher than those estimated from experimental data.

TABLE I

RESONANCE FREQUENCIES OF THE SPHERICAL SHELL BETWEEN THE EARTH AND IONOSPHERE

	n	1	2	3	4
RESONANCE FREQUENCY f_n IN C/S	LOSSLESS CAVITY	10.6	18.3	25.9	33.5
	DAYTIME WITH h = 45 KM	7.6	13.5	19.4	25.4
	AVERAGE DATA OF DAY WITH h = 45 KM AND OF NIGHT WITH h = 60 KM	8.05	14.25	20.5	26.8
	MEASUREMENTS	8	14.1	20.3	26.4

TABLE II

Q FACTORS OF EARTH-IONOSPHERE CAVITY RESONANCES

	f IN C/S	10	30	100
Q - FACTORS	DAY WITH h = 45 KM, NIGHT WITH h = 60 KM	8.5	7	6.7
	DAYTIME ONLY WITH h = 45 KM	6.5	5.4	5.2
	DAY WITH h = 45 KM, NIGHT WITH h = 60 KM, COSMIC RAY IONIZATION	5.5	6	6.4
	MEASUREMENTS	4	6	

3.6. Magnetic Field Effects

The magnetic field effects for a homogeneous sharply bounded ionosphere model have been discussed by Wait [1962] in a quasi-longitudinal approximation for a superimposed radial magnetic field and also for a transverse magnetic field (east-west or west-east propagation along the magnetic equator).

The propagation parameters have been determined by Galejs and Row [1964] and Galejs [1964a, 1964b] in the presence of a horizontal magnetic field. These data provide an indication of propagation characteristics of the waves near the magnetic equator, but they do not apply directly to the earth-ionosphere resonance problem, where the **H** vector is nearly horizontal only over a small surface portion of the earth.

The earth-ionosphere cavity resonance frequencies have been considered in the thesis of Thompson [1963] for a radial magnetic field. He uses a multilayer ionosphere model that approximates the continuous altitude variation of the conductivity tensor. The azimuthal field variation is expressed in terms of the Legendre polynomials $P_n(\cos\theta)$ for the nth mode. The wave equation has four different solutions for each of the ionosphere layers, but it has only two solutions for the space between the ionosphere and the perfectly conducting ground surface. The solution of the problem and the resonance condition for the mode n are obtained after multiplying a sequence of 4×4 matrices. In these computations the displacement currents of the ionosphere are neglected. The computed resonance frequencies are in agreement with measurements, but the Q values are in the range of 15 to 20, with the exception of nighttime anisotropic solutions, where $Q=7$ to 10. The lowering of the nighttime Q figures by the magnetic field has been attributed to wave penetration through the ionosphere and to energy thus escaping from the earth-ionosphere cavity.

Another solution to the problem of earth's ionosphere cavity resonance in the presence of a radial magnetic field is given by Galejs [1965]. The azimuthal field variation is expressed in terms of the Legendre functions $P_\nu(-\cos\theta)$, and the solution of the wave equation is set up so that the four solutions of the ionospheric layers where $\sigma > \omega\epsilon_0$ degenerate into two solutions for the lower ionospheric layers where $\sigma << \omega\epsilon_0$ and for free space. After computing the surface impedance at the lower boundary of the ionosphere, the propagation parameters are determined from the solutions of the modal equation for TM modes, althought the coupling between TE and TM modes is considered in the processes of computing the surface impedance. Only the vertically polarized TM modes are assumed to propagate below the ionosphere. Such assumptions have been justified for a single layer ionosphere model at higher frequencies [Wait 1962, p. 269] and for a thin ionospheric shell also for frequencies in the resonance range [Wait, 1964c].

In the numerical calculations the anisotropic ionosphere models are those of figure 4 and further calculations are made for ionosphere models which consider electrons and ions at an operating frequency of 20 c/s, as indicated in figures 3 and 4 of Galejs [1964b]. In the isotropic ionosphere models σ_0 remains unaltered, $\sigma_1=\sigma_0$, and $\sigma_2\rightarrow 0$. The calculated resonance frequencies f_n are listed in table III. The isotropic day or night models exhibit too high resonance frequencies, but the average for the anisotropic day or night models gives nearly correct results, although the first resonance occurs near 7.7 c/s. At night the various anisotropic models give different results, while their results are nearly the same for daytime or for the isotropic models. The calculated Q factors are listed in table IV. At daytime the Q factors are lowered slightly due to anisotropy. The losses are principally due to absorption in the lower D region where the collision frequencies are high and the magnetic field has only small effects. At nighttime the anisotropy reduces the Q values drastically, and the ions have a considerable effect. For the isotropic models the energy remains below the ionosphere, the losses in the lower ionosphere are small and the Q values are high. For the anisotropic models some energy penetrates the ionosphere and escapes from the earth to ionosphere cavity. Hence the Q figure is low.

TABLE III

RESONANCE FREQUENCIES OF THE SPHERICAL SHELL BETWEEN THE EARTH AND IONOSPHERE IN THE PRESENCE OF A RADIAL MAGNETIC FIELD

	IONOSPHERE MODEL	n	1	2	3	4
RESONANCE FREQUENCY f_n IN C/S	DAY, ANISOTROPIC		7.8	13.8	19.8	26
	DAY, ISOTROPIC		8.1	14.2	20.4	26.6
	NIGHT, ANISOTROPIC	EL.+ IONS F = 20	7.6	15.2	23.2	29.5
		EL.+ IONS F = 0	7.7	13.8	20.1	26.9
		EL. ONLY F = 0	7.8	14.1	20.7	27.8
	NIGHT, ISOTROPIC		8.6	15.3	21.3	27.4
	MEASUREMENTS		8	14.1	20.3	26.4

TABLE IV

Q-FACTORS OF EARTH-IONOSPHERE CAVITY RESONANCES IN THE PRESENCE OF A RADIAL MAGNETIC FIELD

	IONOSPHERE MODEL	F	10	30
Q-FACTORS	DAY, ANISOTROPIC		6.8	5.6
	DAY, ISOTROPIC		8.3	7
	NIGHT, ANISOTROPIC	EL. + IONS F = 20	1.7	6.6
		EL. + IONS F = 0	2.0	2.7
		EL. ONLY F = 0	2.8	4.5
	NIGHT, ISOTROPIC		8.8	9.3
	MEASUREMENTS		4	6

The previous calculations have assumed a constant radial magnetic field (S pole) all over the globe. A reversal of the static magnetic field (N pole) causes negligible differences in the propagation parameters or in the values of f_n and Q. Hence a model using the average of the radial magnetic field may provide a first order estimate of the anisotropy effects. The present calculations assume the radial magnetic of the polar regions to be extended over the whole surface of the globe and the anisotropy effects should be too pronounced. The analytical results are highly dependent on the detailed structure of the nighttime anisotropic ionosphere, and further studies of it may be in order.

4. Source Distributions and Noise Spectra

Most of the ELF energy is of terrestrial origin and is caused by thunderstorm activity, although there are some high latitude events of an apparent extraterrestrial origin. Such emissions of approximately 800 c/s have been detected near Kiruna, Sweden [Egeland, 1964] and of 14 to 17 c/s on Kerguelen Island in the Southern part of the Indian Ocean [Gendrin and Stephant, 1964, private communication]. The Spectran recording of the signals received at Kerguelen is shown in figure 6. The upper trace shows a normal signal with resonances near 8, 14, 20, and 25 c/s. There is an additional signal of 14 to 17 c/s on the two lower traces. When neglecting the presence of such extraneous signals the terrestrial noise spectra can be calculated after estimating the dipole moments of the sources and using propagation parameters that are appropriate to the selected ionosphere model.

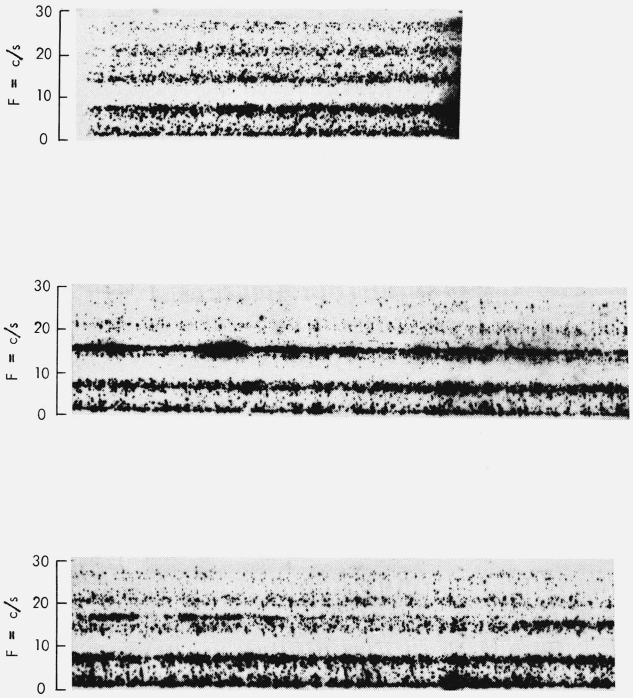

FIGURE 6. *Spectran recordings of noise waveforms at Kerguelen Island.*

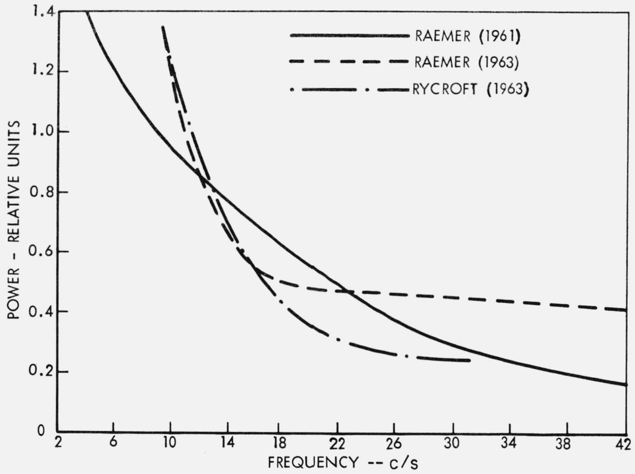

FIGURE 7. *Estimated lighting flash spectrum.*

Raemer [1961a, b] has estimated the power spectrum of the dipole moment in a median lightning flash Ids (or the first derivative of the electrostatic moment dM/dt) based on statistics of lightning strokes compiled by Williams [1959]. Later work of Raemer [1963 private communication] has lead to a revised estimate. These spectra are shown in figure 7. Pierce [1963] has computed the magnitude of the frequency spectra of the second derivative d^2M/dt^2, and Rycroft [1963] has calculated the frequency spectra of dM/dt using a similar representation of the lightning moments. The square of the frequency spectrum computed by Rycroft [1963] is also indicated in figure 7.

An equivalent power spectrum for the dipole moment of terrestrial noise sources has been deduced by Harris and Tanner [1962] from the measurements of Balser and Wagner [1960] by estimating the complex propagation constant within this frequency range in a "trial and error" approach. Harris and Tanner [1962] do not require an ionosphere model or any knowledge of lightning waveforms and their statistics for obtaining this equivalent power spectrum.

After determining the source spectra and the propagation parameters the power spectrum of the received noise may be computed from

$$G(i\omega)=\frac{g(i\omega)\,|\nu(\nu+1)|^2}{8(ak_0h)^2\,|\nu+\frac{1}{2}|\,[\cosh(2\pi\,\mathrm{Im}\,\nu)-\cos(2\pi\,\mathrm{Re}\,\nu)]}\cdot\left\{\frac{\sinh[2\,\mathrm{Im}\,\nu\,(\pi-\theta)]}{(-\mathrm{Im}\,\nu)}-\frac{\sin[(2\,\mathrm{Re}\,\nu+1)(\pi-\theta)-(\pi/2)]}{\mathrm{Re}\,\nu+\frac{1}{2}}\right\}_{\theta_1}^{\theta_2}, \quad (9)$$

where the ν is given by (2) and where the sources are assumed to be uniformly distributed in the θ interval of $\theta_1<\theta<\theta_2$. The derivation of (9) starts from (1) and Legendre functions $P_\nu(-\cos\theta)$ are replaced by an asymptotic expansion. There are also equivalent spectral representations in terms of Legendre polynomials [Raemer, 1961a, b].

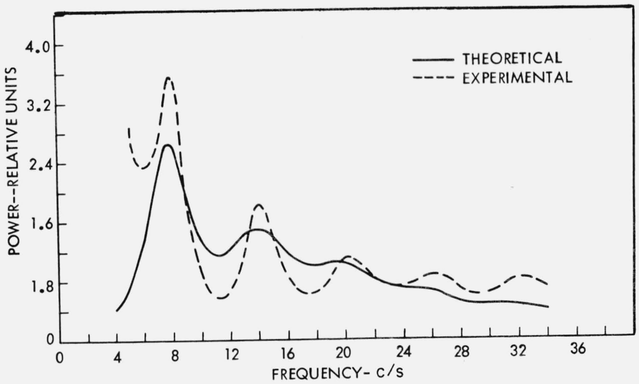

FIGURE 8. *ELF Noise spectrum—effective ionospheric height assumed constant.*

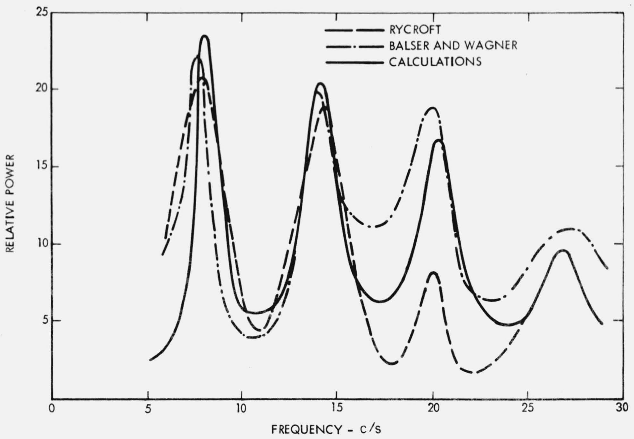

FIGURE 10. *Noise spectrum for 47° ≤ θ ≤ 128°.*

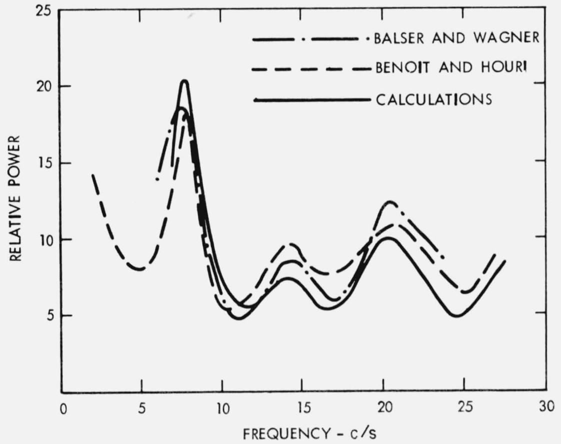

FIGURE 9. *Noise spectrum for 38° ≤ θ ≤ 72°.*

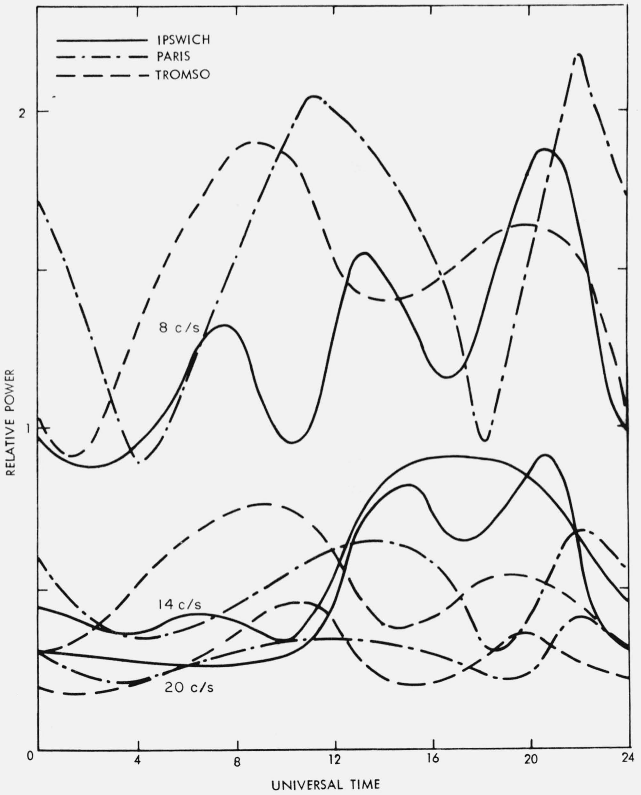

FIGURE 11. *Diurnal variation of the resonance modes at Ipswich, Mass., February 1961 (vertical electric field), Chambon-la-Foret (near Paris) July 1962 and at Tromso (Norway) April 1962 (horizontal magnetic field).*

The noise spectrum of Raemer [1961b] has been used for computing the observed noise spectra in conjunction with a sharply bounded ionosphere model. This spectrum is shown in figure 8 and is seen to exhibit excessively damped higher resonance peaks.

The exponential ionosphere model [Galejs, 1961b, 1962a, 1964a] produces a better agreement with measurements as seen in figures 9 and 10. In this work propagation is assumed to be uniform around the surface of the earth and the propagation parameters represent the average of day and night conditions. The sources are assumed to be uniformly distributed over the θ interval indicated in the figures (θ is defined with respect to the observer).

5. Diurnal and Seasonal Changes of Power Spectra

Balser and Wagner [1962a] have recorded diurnal variations of measured noise spectra. Their measurements of the diurnal variations of the power level at the first three resonance frequencies are shown in figure 11 together with similar measurements by Gendrin and Stefant [1962a]. Diurnal variations of the fields near the two lower resonance frequencies have been measured at Kingston, R.I., by Polk, Huck, and Yu [1963, private communication]. The maximum signal intensity occurs between 1500 and 1800 UT. In an amplitude versus time plot the signal peak is broader during summer than winter. Seasonal and diurnal variations of the noise power in the ELF frequency band below 35 c/s have been described by Wright [1963] without singling out the individual resonance modes. The measurements are made at Byrd Station, Antarctica (80°S, 120°W) and show peak intensities near 1000 UT during spring and summer and near 2300 UT during most of the year.

The gross features of the spectrum, such as the number of peaks and their approximate time of occurance, may be explained qualitatively by noting areas of globe with particularly high thunderstorm activity and by measuring the polar angle θ from these source regions to the observation point. The vertical electric field E_r and the horizontal magnetic field H_ϕ of the mode n exhibit the θ dependence of

$$E_r \sim P_n(\cos\theta) \tag{10}$$

$$H_\phi \sim \frac{d}{d\theta} P_n(\cos\theta), \tag{11}$$

where $P_n(\cos\theta)$ is a Legendre polynomial of the order n. A particular geographical region will contribute to the measured values of E_r or H_ϕ in the mode n if the right hand sides of (10) or (11) are sizeable for those values of θ. Such considerations have been reported by Balser and Wagner [1962a] and Gendrin and Stefant [1962a].

More detailed calculations have been carried out by Abraham [1964, private communication; Galejs 1964a] who used available estimates of geographical distributions of the noise sources and who calculated the day and night propagation effects after expressing the fields as a sum of recirculating azimuthal waves.

In the expression for the electric field component E_r of (1), the Legendre function is replaced by the leading term of its asymptotic expansion. The function $\sin \nu\pi$ is expanded into

$$(\sin \nu\pi)^{-1} = -2i \sum_{n=0}^{\infty} e^{i(2n+1)\nu\pi} \tag{12}$$

which is valid for $\operatorname{Im} \nu > 0$. This results in [Abraham, 1964, private communication]

$$E_r = \frac{Ids}{2\omega\epsilon h a^2} \frac{\nu(\nu+1)}{\sqrt{2\pi(\nu+\frac{1}{2})\sin\theta}} e^{i\frac{3\pi}{4}} \sum_{n=0}^{\infty} \left\{ e^{i(\nu+\frac{1}{2})(\theta+2n\pi)-i2n\frac{\pi}{2}} + e^{i(\nu+\frac{1}{2})[2\pi(n+1)-\theta]-i(2n+1)\frac{\pi}{2}} \right\}. \tag{13}$$

The first term of the summation represents a wave which has traveled the direct distance from the source to the observation point (θ) plus n full circles around the sphere ($2n\pi$), while experiencing a phase shift of ($-\pi/2$) radians at each traversal of the source point or its antipode. The other term of the summation represents a wave which has traveled toward the observation point across the antipode of the source ($2\pi-\theta$) and has made furthermore n circles around the sphere ($2n\pi$) and has also experienced a phase shift of ($-\pi/2$) radians at each traversal of the source point or its antipode. This physical interpretation of (13), which is equivalent to (1) for $\nu=$ const., permits a heuristic consideration of the ν differences between the day and night hemispheres. The multiplicative ν factors in front of the n summation are assumed to be constant and equal to $\bar{\nu}$ which is intermediate between the day and night ν values. The θ factors of the exponentials are replaced by integrals in order to account for the ν variation seen by the circular wave front as it propagates in the θ-direction with respect to the source. The amplitude changes at the day and night boundary can be also considered for each traveling wave by assuming that the energy of the wave remains constant after passing the day and night boundary. This results in decreased amplitudes for the larger ionospheric height of the night hemisphere. In the numerical work the surface of the earth was subdivided into 10° cells. Each cell had a constant surface density of noise sources but a variation was allowed from cell to cell. The plots of thunderstorm days over the world were used to set relative values of lightning activity for each of the cells, by assuming that the distribution of squared dipole moment per unit area is directly proportional to the thunderstorm day plots according to the Handbook of Geophysics [1960]. The diurnal variations were accounted for by a multiplicative factor in terms of local time at the source [Williams, 1959]. The final equation used in the spectrum calculations is similar to (9), although it is algebraically more involved [Abraham, 1964, private communication].

The diurnal variations of the noise power at the first three Schumann resonances have been computed using the exponential ionosphere model of figure 5 and are shown in figure 12. The calculations have disregarded the noise sources near the antipode where the usual asymptotic approximation of the Legendre functions is inaccurate. Figure 12 shows a qualitative agreement with the measurements of Balser and Wagner [1962a], that are shown in figure 11.

An example of seasonal power variations is shown in figure 13. It indicates a decreased noise power near the higher frequencies and/or during winter months.

Based on measurements at one location of the earth it is also possible to predict the noise levels at other locations using this procedure of noise spectrum calculations. The noise that is due to a worldwide distribution of thunderstorms depends on the frequency variation and also on the absolute value of the power spectrum of the mean dipole moment of the source $g(i\omega, \theta, \phi)$. The θ and ϕ dependence may be estimated from the thunderstorm frequency maps as was indicated before, which leaves the absolute value and the frequency dependence of $g(i\omega)$ as an unknown. Assuming that $g(i\omega)$ is the same for all noise sources, it may be determined by making the calculations to agree with measurements at one particular location. This gives a uniquely determined noise distribution over the surface of the earth and the accuracy of this model can be checked by com-

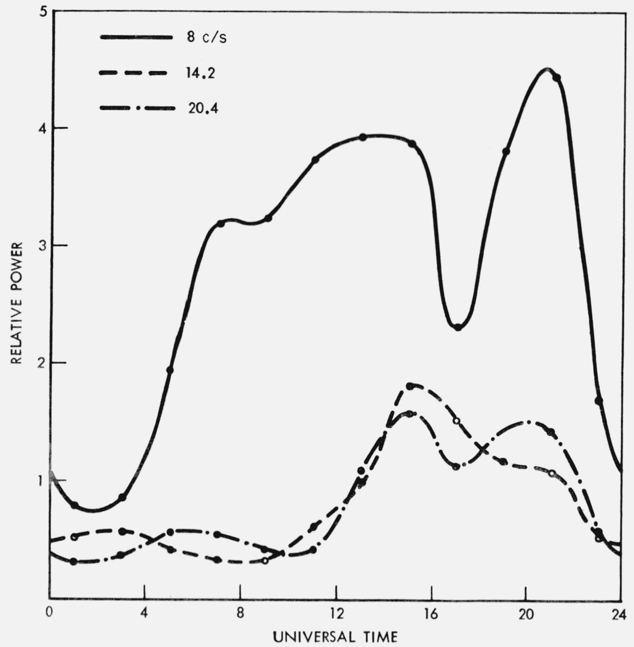

FIGURE 12. *Calculated diurnal variation of noise power at the peak of the resonance modes—wintertime—Boston, Mass.*

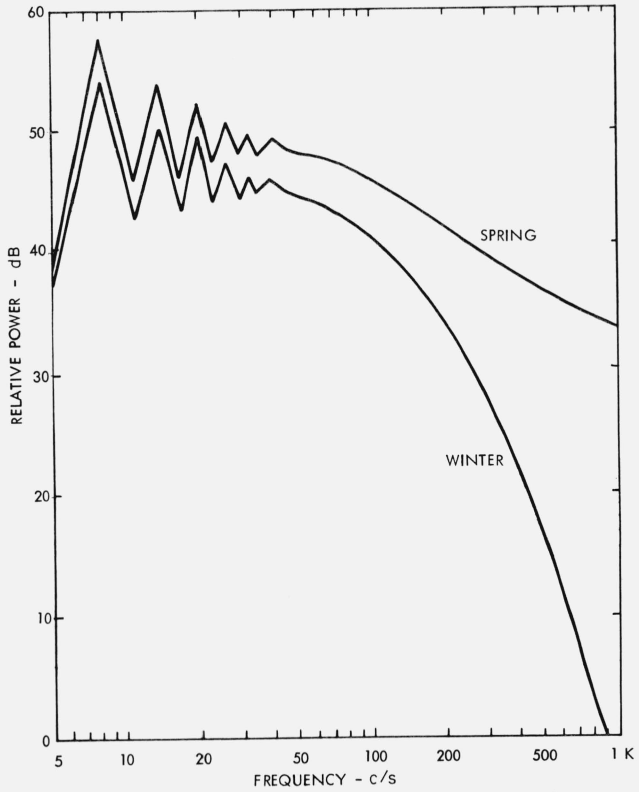

FIGURE 13. *Calculated noise spectrum for local noon in northern hemisphere.*

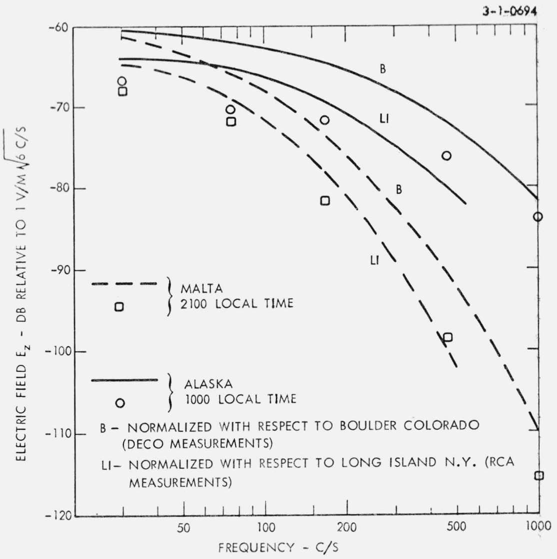

FIGURE 14. *Comparison of calculated field strength's data with July 1963 measurements by DECO and RCA.*

paring the calculations with measurements at other locations. In an example, measurements by DECO at Boulder, Colo., [Maxwell, 1963, private communication] and by RCA at Long Island, N.Y. [Powers, 1964, private communication], were used to determine $g(i\omega)$. The calculations by Abraham [1964, private communication] based on these values of $g(i\omega)$ have been compared with simultaneous measurements made in Malta and Alaska [Maxwell, 1963, private communication]. As seen from figure 14 the analytic curves agree in shape with the measurements, but the normalization with respect to Boulder gives somewhat higher predictions of the noise level. This may possibly be caused by summertime thunderstorm activity in the mountains near Boulder, Colo., that would give significant near-fields and a high estimate for the function $g(i\omega)$. The calculations are extended above the region of the earth-ionosphere cavity resonances to show that more significant noise level changes occur at higher frequencies.

In the calculations leading to figures 12 through 14 the day and night effects are considered most accurately if the day and night boundary is located symmetrically with respect to source (i.e., source at local noon or midnight). For nonsymmetrical day and night boundaries it is difficult to estimate the effects of wave front distortion. When neglecting such distortions (or assuming propagation strictly in the θ-direction), this method can account for $\nu(\theta)$ variations, which may be caused by ionosphere parameter changes or by changes of the static magnetic field along the surface of the earth. The assymptotic expansions of $P_\nu(-\cos\theta)$ used in (13) or (9) can be modified to increase their accuracy for observations near the source or its antipode [Abraham, 1965].

6. Variations of Resonance Frequencies

Diurnal frequency variations of the earth-ionosphere cavity modes have been reported by Balser and Wagner

[1962b], Gendrin and Stefant [1963], Rycroft [1963] and by Chapman and Jones [1964b]. The resonance frequencies vary by about ±0.2 to 0.3 c/s, but the changes are not simultaneous for the different modes; a low resonance frequency of one mode occurs simultaneously with a high frequency of a different mode.

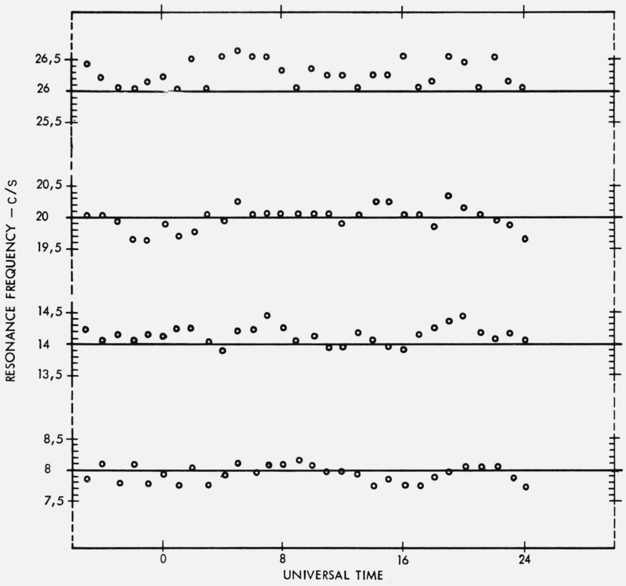

FIGURE 15. *Diurnal frequency variations measured by Gendrin and Stefant.*

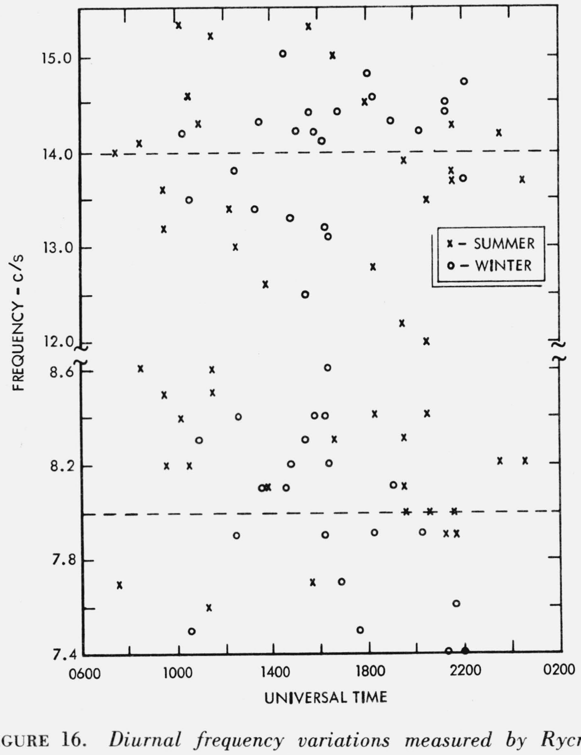

FIGURE 16. *Diurnal frequency variations measured by Rycroft.*

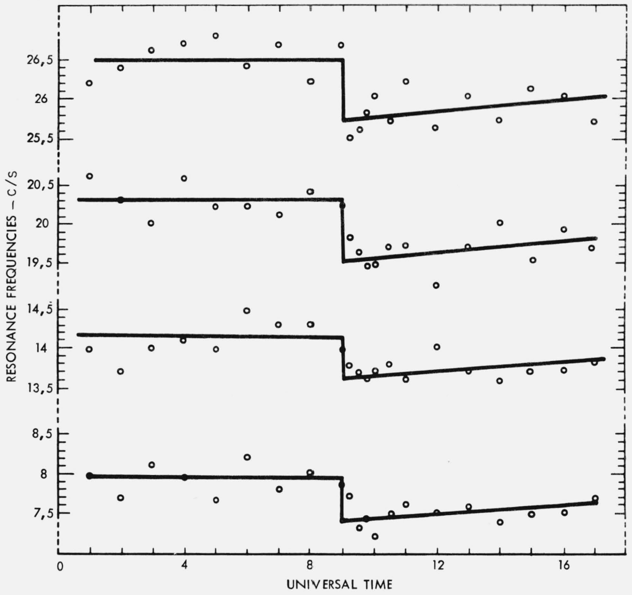

FIGURE 17. *Nuclear effects on earth-ionosphere cavity resonances.* Measurements by Gendrin and Stefant.

A high altitude nuclear explosion has been shown to affect the earth-to-ionosphere cavity resonances by simultaneously lowering all the resonance frequencies [Gendrin and Stefant, 1962a and b; Balser and Wagner, 1963]. Tepley et al., [1963] noted also an increase of signal strength for observations near the conjugate point of the explosion. The lowering of resonance frequency shown in figure 17 has been attributed to worldwide lowering of the effective ionosphere in a model of a single layer isotropic ionosphere [Gendrin and Stefant, 1962b]. The depression of the resonance frequency can be also explained by assuming an exponential daytime ionosphere to be effective around the globe [Row, 1964, private communication]. As an illustration the daytime resonance frequencies of table 1 compare with the measured resonance frequencies of figure 17. There are VLF observations [Zmuda et al, 1963] that can be explained by a lowering (or increasing electron density) of the *D*-region boundary. Brady et al., [1964] interpreted VLF diurnal phase shifts to show that some enhanced *D*-region ionization persists for several days following the event.

The lowering of the resonance frequency following a nuclear explosion has been attributed by Chapman and Jones [1964a] in their two-layer ionosphere model solely to a decreasing conductivity of the *E*-region.

After measuring a change of the resonance frequencies following a nuclear explosion, it is possible to advance simple ionosphere models for explaining a particular change. However, no attempt has been reported to predict resonance shifts for different nuclear yields and detonation altitudes. This rather complex problem requires a quantitative evaluation of the various nuclear mechanisms that may lead to ionosphere perturbations. If such information is available, the computation of the resonance shifts is relatively simple. It would be also of interest to investigate the effects of sudden ionosphere disturbances (SID) and to discriminate between SID's and nuclear effects.

The resonant frequency changes can be considered more accurately with nonuniform models of the propagation geometry. A perturbation method [Wait, 1964b] or numerical techniques [Madden and Thompson, 1964] have been used for solving such problems.

Madden and Thompson [1964] compute the local surface impedance and propagation constant for a uniform cylindrical shell between the anisotropic ionosphere and the earth using matrix multiplication techniques. There is a smooth variation of the parameters between polar and equatorial regions and also between day and night hemispheres. These local surface impedances and propagation constants are assumed to characterize a spherical transmission line of nonuniform parameters. The resonance frequencies and Q factors are computed numerically using a lumped parameter approximation of the two-dimensional transmission line. This model has been applied to explain the diurnal frequency variations and also the frequency changes due to a nuclear explosion.

7. Areas of Future Work

There are numerous observations of earth-ionosphere cavity resonances and existing ionosphere and source models account for most observed characteristics like the resonance frequency, Q factors, shape of the noise spectra, and diurnal variations of the received power.

Future work appears to be required in a systematic study of diurnal and seasonal amplitude and frequency variations of resonances. These changes should also be correlated with ionospheric disturbances. The parameters of the nighttime ionosphere require also a future clarification. The variations of the ionosphere parameters and the variations of the magnetic field vector along the surface of the earth should be incorporated in future calculations [Wait, 1964b; Madden and Thompson, 1964]. Much work remains to be done for assessing nuclear effects on the earth ionosphere resonance and for discriminating between lower level nuclear and sudden ionospheric disturbances.

The work of Sylvania reported in this paper has been supported principally by the Office of Naval Research under Contract Nonr 3185(00).

8. Appendix. Q Factors

The basic definition of the Q factor of a cavity with lossy walls is

$$\frac{1}{Q_i}=\frac{\int\int \mathbf{E}\times\mathbf{H}^*\cdot\mathbf{ds}}{\frac{\omega\epsilon_0}{2}\int\int\int \mathbf{E}\cdot\mathbf{E}^*dv+\frac{\omega\mu_0}{2}\int\int\int \mathbf{H}\cdot\mathbf{H}^*dv} \tag{14}$$

where the surface integral is extended over the cavity walls and the volume integral covers the inside of the cavity. The ground surface is assumed to be a perfect conductor and E_θ is related to H_ϕ at the ionospheric boundary by the surface impedance Z_s as $E_\theta=Z_sH_\phi$. However E_θ is assumed to be small relative to E_r for vertically polarized waves. Equation (14) is rewritten as

$$\frac{1}{Q_i}=\frac{2\int\int \mathrm{Re}Z_s|H_\phi|^2(a+h)^2\sin\theta d\theta d\phi}{\omega\int\int\int[\epsilon_0|E_r|^2+\mu_0|H_\phi|^2]r^2dr\sin\theta d\theta d\phi} \tag{15}$$

$$\approx\frac{2\mathrm{Re}Z_s}{\omega h}\,\frac{\int|H_\phi|^2\sin\theta d\theta}{\int[\epsilon_0|E_r|^2+\mu_0|H_\phi|^2]\sin\theta d\theta} \tag{16}$$

where a is the radius of earth, h is the ionospheric height and where the cavity is assumed to be uniform. For excitation by a radial electric dipole of moment Idr the field component E_r and H_ϕ follow from (21), (61), and (62) of Galejs [1964a] as

$$E_r=\frac{iIdr}{4\pi h\omega\epsilon a^2}\,\frac{n(n+1)(2n+1)}{D_n}\,P_n(\cos\theta) \tag{17}$$

$$H_\phi=-\frac{Idr}{4\pi ha}\,\frac{(2n+1)}{D_n}\,\frac{\partial}{\partial\theta}\,P_n(\cos\theta) \tag{18}$$

with $D_n=n(n+1)-\nu(\nu+1)$. Substituting (17) and (18) in (16) and evaluating the integrals gives

$$\frac{1}{Q_i}=\frac{2\mathrm{Re}Z_s}{\omega\mu_o h\left[\frac{n(n+1)}{(k_0a)^2}+1\right]} \tag{19}$$

where $k_0=\omega\sqrt{\mu_0\epsilon_0}$. From the approximate modal equation

$$S=\sqrt{1+\frac{iZ_s}{\omega\mu_0h}} \tag{20}$$

it follows that

$$2\ \mathrm{Re}S\ \mathrm{Im}S=\frac{\mathrm{Re}Z_s}{\omega\mu_0h}\,. \tag{21}$$

Applying (21) and the resonance condition

$$n(n+1)\approx(k_0a\ \mathrm{Re}S)^2 \tag{22}$$

(19) simplifies to

$$\frac{1}{Q_i}=\frac{4\ \mathrm{Re}S\ \mathrm{Im}S}{(\mathrm{Re}S)^2+1} \tag{23}$$

which was listed as (4). When considering the differences of E_r, H_ϕ, Z_s, and h between the day and night hemispheres it follows that

$$\frac{1}{Q_i} \approx \frac{2[\mathrm{Re}S\ \mathrm{Im}S|_{\mathrm{day}} + \mathrm{Re}S\ \mathrm{Im}S|_{\mathrm{night}}]}{(\mathrm{Re}S|_{\mathrm{avg}})^2 + 1}. \qquad (24)$$

The form (5) or (6) is obtained by neglecting the E integral or H integral, and by doubling the value of the H integral or E integral in (14).

9. References

Abraham, L. G. Jr., Aspects of the terrestrial ELF noise spectrum when near the source or its antipode, Radio Sci., J. Res. NBS **69D** (July 1965).

Balser, M., and C. A. Wagner (Nov. 1960), Observations of earth-ionosphere cavity resonances, Nature **188,** 638–641.

Balser, M., and C. A. Wagner (Feb. 1962a), Diurnal power variations of the earth-ionosphere cavity modes and their relationship to worldwide thunderstorm activity, J. Geophys. Res. **67,** No. 2, 619–625.

Balser, M., and C. A. Wagner (Sept. 1962b), On frequency variations of the earth-ionosphere cavity modes, J. Geophys. Res. **67,** No. 10, 4081–4083.

Balser, M., and C. A. Wagner (July 1, 1963), Effect of a high-altitude nuclear detonation on the earth-ionosphere cavity, J. Geophys. Res. **68,** No. 13, 4115–4118.

Benoit, R., and A. Houri (1961), Propagation of very-low-frequencies in the earth-ionosphere system, Ann. Geophys. (France) **17,** No. 4, 370–373.

Benoit, R., and A. Houri (Nov. 5, 1962), Power spectrum measurements of geophysical noise. Application to earth-to-ionosphere cavity, C. R. Acad. Sci. (France) **255,** 2496–2498.

Bourdeau, R. E., E. C. Whipple, and J. F. Clark (Oct. 1959), Analytic and Experimental Conductivity between the stratosphere and the ionosphere, J. Geophys. Res. **64,** 1363–1370.

Brady, A. H., D. D. Crombie, A. G. Jean, A. C. Murphy, and F. K. Stede (May 1, 1964), Long-lived effects in the D region after the high altitude nuclear explosion of 9 July 1962, J. Geophys. Res. **69,** 1921.

Brock-Nannestad, L. (Sept. 1962), Bibliography electromagnetic phenomena with special reference to ELF (1–3000 c/s) TR 10 Saclant, ASW Research Center, La Spezia, Italia.

Chapman, F. W., and D. L. Jones (May 16, 1964a), Earth-ionosphere cavity resonances and the propagation of extremely low frequency radio waves, Nature **202,** 654–657.

Chapman, F. W. and D. L. Jones, Observations of earth-ionosphere cavity resonances and their interpretation in terms of a two-layer ionosphere model, Radio Sci., J. Res. NBS **68D,** 1177–1185 (Nov. 1964b).

Chapman, F. W., and R. C. Macario (May 1956), Propagation of audio frequency radio waves to great distances, Nature **177,** 930–933.

Egeland, A. (1964a) ELF (500–1000 c/s) emissions at high latitude, Proc. NATO Advanced Study Institute on Natural Electromagnetic Phenomena Below 30 kc/s, Bad Homburg, West Germany, July–Aug. 1963, pp. 157–166 (Plenum Press, New York, N.Y.).

Fournier, H. (Aug. 17, 1960), Some aspects of the first high-frequency geomagnetic recordings obtained at Garchy, C. R. Acad. Sci. (France) **251,** No. 7, 962–964.

Galejs, J. (1961a), Terrestrial extremely-low-frequency noise spectrum in the presence of exponential ionospheric conductivity profiles, J. Geophys. Res. **66,** No. 9, 2787–2792.

Galejs, J. (Nov. 1961b), ELF waves in presence of exponential ionospheric conductivity profiles, IRE Trans. Ant. Prop. **AP-9,** 554–562.

Galejs, J. (July 1962a), A further note on terrestrial extremely-low frequency propagation in the presence of an isotropic ionosphere with an exponential conductivity-height profile, J. Geophys. Res. **67,** No. 7, 2715–2728.

Galejs, J. (1962b), Terrestrial extremely-low frequency propagation in the presence of an isotropic ionosphere with an exponential conductivity height profile, Proc. International Conference on the Ionosphere, London, pp. 467–474 (Chapman and Hall).

Galejs, J. (1964a), Terrestrial extremely-low frequency propagation, Proc. NATO Advanced Study Institute on Natural Electromagnetic Phenomena Below 30 kc/s, Bad Homburg, West Germany, July–Aug. 1963, pp. 205–258 (Plenum Press, New York, N.Y.).

Galejs, J. (June 1964b), ELF and VLF waves below an inhomogeneous anisotropic ionosphere, Radio Sci. J. Res. NBS **68D,** No. 6, 693–707.

Galejs, J., and R. V. Row (Jan. 1964), Propagation of ELF waves below an inhomogeneous anisotropic ionosphere, IEEE Trans. Ant. Prop. **AP-12,** No. 1, 74–83.

Galejs, J., On the terrestrial propagation of ELF and VLF waves in the presence of a radial magnetic field, Radio Sci., J. Res. NBS **69D,** No. 5 (May 1965).

Gendrin, R., and R. Stefant (Sept. 1962a), Magnetic records between 0.2 and 30 c/s, AGARD Conference on Propagation of Radio Frequencies Below 300 kc/s, Munich, Germany.

Gendrin, R., and R. Stefant (1962b), Effects of the very high altitude nuclear explosion of the July 9, 1962 on the earth to ionosphere cavity resonances, experimental results and interpretation, C. R. Acad. Sci. **255,** 2273–2275 and 2493–2495.

Handbook of Geophysics (1960), USAF Air Research and Development Command, Airforce Cambridge Research Center (Macmillan Co., New York, N.Y.).

Harris, F. B., Jr., and R. L. Tanner (1962), A method for determination of lower ionosphere properties by means of field measurements on sferics, J. Res. NBS **66D** (Radio Prop.), No. 4, 463–478.

Hepburn, F., and E. T. Pierce (May 9, 1953), Atmospherics with very-low-frequency components, Nature (GB) **171,** 837–838.

Jean, A. G., Jr., and A. C. Murphys, J. R. Wait, and D. F. Wasmundt (1961), Propagation attenuation rates at ELF, J. Res NBS **65D** (Radio Prop.), No. 5, 475–479.

Keefe, T. J., C. Polk, and H. L. Koenig (1964), Results of simultaneous ELF measurements at Brannenburg (Germany) and Kingston, R.I., ULF Symposium, Boulder, Colo. Aug. 17–20, 1964.

Koenig, H. (Dec. 1958), Atmospherics geringster frequenzen, Dissertation an der Technischen Hochschule München.

Koenig, H. (July 1959), Atmospherics geringster frequenzen, Z. Angew. Phys. **11,** No. 7, 264–274.

Koenig, H., E. Haine, and C. H. Antoniadis (Aug. 1961), Messung von 'atmospherics' geringster frequenzen in Bonn, Z. Angew. Phys. **13,** No. 8, 364–367.

Liebermann, L. (Dec. 1956a), Extremely-low-frequency electromagnetic waves: I. Reception from lightning, J. Appl. Phys. **27,** No. 12, 1473–1476.

Liebermann, L. (Dec. 1956b), Extremely-low-frequency electromagnetic waves: II. Propagation properties, J. Appl. Phys. **27,** No. 12, 1477–1483.

Lokken, J. E., J. A. Schand, and C. S. Wright (1962), A note on the classification of geomagnetic signals below 30 c/s, Can. J. Phys. **40,** 1000–1009.

Madden, T. R. (1961), An analysis of the earth-ionosphere electromagnetic resonances, unpublished class notes for MIT course 12.88, April 24.

Madden, T. Rand, W. Thompson, Low-frequency electromagnetic oscillations of the earth-ionosphere cavity, Report on Project NRL 371–401, Geophysics Laboratory, MIT, Cambridge, Mass. (Sept. 22, 1964). Also Rev. Geophys. **3,** No. 2 (May 1965).

Pierce, E. T. (July–Aug. 1960), Some ELF phenomena, J. Res. NBS **64D** (Radio Prop.), No. 4, 383–386.

Pierce, E. T. (July 1, 1963), Excitation of earth-ionosphere cavity resonances by lightning flashes, J. Geophys. Res. **68,** No. 13, 4125–4127.

Polk, C., and F. Fitchen (May–June 1962), Schumann resonances of the earth-ionosphere cavity—extremely low frequency reception at Kingston, R.I., J. Res. NBS **66D** (Radio Prop.), No. 3, 313–318.

Polk, C. (1962), ELF propagation and the earth-ionosphere resonant cavity, paper presented at the meeting of the Ionospheric Research Committee of AGARD in Munich, West Germany, Sept. 17–21, 1962.

Raemer, H. R. (1961a), On the extremely-low-frequency spectrum of earth-ionosphere cavity response to electric storms, J. Geophys. Res. **66,** No. 5, 1580–1583.

Raemer, H. R. (Nov.–Dec. 1961b), On the spectrum of terrestrial radio noise at extremely low frequencies, J. Res. NBS **65D** (Radio Prop.), No. 6, 581–594.

Rycroft, M. J. (Dec. 1963), Low frequency disturbances of natural origin of the electric and magnetic fields of the earth, Ph.D. Thesis, University of Cambridge.

Schumann, W. O. (1952a), Über die strahlungslosen Eigenschwingungen einer leitenden Kugel, die von einer Luftschicht und einer Ionosphärenhülle umgeben ist, Z. Naturforsch, **72,** 149–154.

Schumann, W. O. (1952b), Über die Dämpfung der elektromagnetischen Eigenschwingungen des Systems Erde-Luft-Ionosphäre, Z. Naturforsch **72,** 250–252.

Schumann, W. O. (Aug. 1957), Elektrische Eigenschwingungen des Hohlraumes Erde-Luft-Ionosphäre, Z. Angew. Phys. **9,** 373–378.

Seshadri, S. R., Radiation from electromagnetic sources in a plasma, IEEE Transactions on Antennas and Prop. **AP–13,** No. 1, 79–88 (Jan. 1965a).

Seshadri, S. R. (March 1965b), VLF propagation in a sharply bounded, compressible isotropic ionosphere, IEEE Trans. Ant. Prop. **AP–13,** No. 2, 311–317.

Tepley, L. R. (Dec. 1959), A comparison of sferics as observed in the very-low-frequency and extremely-low-frequency bands, J. Geophys. Res. **64,** No. 12, 2315–2329.

Tepley, L., R. C. Wentworth, K. G. Lambert, H. V. Prentiss, K. D. Amundsen and D. R. Hillendahl (26 Dec. 1963), Sub ELF geomagnetic fluctuations, III final report Contract AF 19(628)–462, Lockheed Missile and Space Co., Palo Alto, Calif.

Thompson, W. B. (Feb. 1963), A layered model approach to the earth-ionosphere cavity resonance problem, Ph.D. Thesis, Department of Geology and Geophysics, MIT.

VanZandt, T. E. (1964), Ionosphere, ULF Symposium, Boulder, Colo., Aug. 17–20.

Wait, J. R. (March 1958), Extension to mode theory of VLF propagation, J. Geophys. Res. **63,** 125–135.

Wait, J. R. (1962), Electromagnetic waves in stratified media (Macmillan Co., New York, N.Y.).

Wait, J. R. (Feb. 1963a), The mode theory of VLF radio propagation for a spherical earth and a concentric anisotropic ionosphere, Can. J. Phys. **41,** 299–315.

Wait, J. R. (May–June 1963b), Concerning solutions of the VLF mode problem for an anisotropic curved ionosphere, J. Res. NBS **67D** (Radio Prop.), No. 3, 297–302.

Wait, J. R. (April 1964a), On the theory of Schumann resonances in the earth-ionosphere cavity, Can. J. Phys. **42,** 575–582.

Wait, J. R. (1964b), Earth-ionosphere cavity resonances and the propagation of ELF radio waves, ULF Symposium, Boulder, Colo., Aug. 17–20. Also Radio Sci. J. NBS **69D,** No. 8.

Wait, J. R. (1964c), Cavity resonances for a spherical earth with a concentric anisotropic shell (private communication).

Wait, J. R. (1964d), A note on the Schumann problem for nonuniform boundaries (private communication).

Williams, J. C. (May 1959), Thunderstorms and VLF radio noise, Ph.D. Thesis, Harvard University, Division of Engineering and Applied Physics.

C. S. Wright (July–Aug. 1963), Conjugate relations of micropulsations and other ionospheric phenomena, NATO Advanced Study Institute on Natural Electromagnetic Phenomena Below 30 kc/s, Bad Homburg, Germany.

Zmuda, A. J., B. W. Shaw, and C. R. Haave (Feb. 1, 1963), VLF disturbances and the high-altitude nuclear explosion of July 9, 1962, J. Geophys. Res. **68,** No. 3, 745–758.

(Paper 69D8–538)

US006960975B1

(12) United States Patent
Volfson

(10) **Patent No.: US 6,960,975 B1**
(45) **Date of Patent: Nov. 1, 2005**

---

(54) **SPACE VEHICLE PROPELLED BY THE PRESSURE OF INFLATIONARY VACUUM STATE**

(76) Inventor: **Boris Volfson**, 5707 W. Maple Grove Rd., Apt. 3046, Huntington, IN (US) 46750

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 8 days.

(21) Appl. No.: **11/079,670**

(22) Filed: **Mar. 14, 2005**

Related U.S. Application Data

(63) Continuation of application No. 10/633,778, filed on Aug. 4, 2003, now abandoned.

(51) **Int. Cl.**^{7} **H01F 6/00**; F03H 5/00
(52) **U.S. Cl.** **335/216**; 60/200.1
(58) **Field of Search** 335/216; 60/200.1

(56) References Cited

U.S. PATENT DOCUMENTS

3,626,605	A	12/1971	Wallace
3,626,606	A	12/1971	Wallace
3,823,570	A	7/1974	Wallace
5,197,279	A	3/1993	Taylor
6,353,311	B1	3/2002	Brainard et al.

OTHER PUBLICATIONS

M.T. French, "To the Stars by Electromagnetic Propulsion", http://www.mtjf.demon.co.uk/antigravp2.htm#cforce.

Evgeny Podkletnov, "Weak Gravitational Shielding Properties of Composite Bulk $YBa_2Cu_33O_{(7-x)}$ Superconductor Below 70K Under E.M. Field", LANL database number cond-mat/9701074, v. 3, 10 pages, Sep. 16, 1997.

N. LI & D.G. Torr, "Effects of a Gravitomagnetic Field on Pure Superconductors", Physical Review, vol. 43, p. 457, 3 pages, Jan. 15, 1991.

Evgeny Podkletnov, Giovanni Modanese "Impulse Gravity Generator Based on Charged $YBa_2Cu_33O_{7-y}$ Superconductor with Composite Crystal Structure", arXiv.org/physics database, #0108005 vol. 2, 32 pages, 8 figures, Aug. 30, 2001.

S. Kopeikin & E. Fomalont, "General Relativistic Model for Experimental Measurement of the Speed of Propagation of Gravity by VLBI", Proceedings of the 6th European VLBI Network Symposium Jun. 25-28, 2002, Bonn, Germany, 4 pages.

Sean M. Carroll, "The Cosmological Constant", http://pancake.uchicago.edu/~carroll/encyc/, 6 pages.

Chris Y. Taylor and Giovanni Modanese, "Evaluation of an Impulse Gravity Generator Based Beamed Propulsion Concept", American Institute of Aeronautics and Astronautics, Inc., 2002.

Peter L. Skeggs, "Engineering Analysis of the Podkletnov Gravity Shielding Experiment", Quantum Forum, Nov. 7, 1997, http://www.inetarena'.com/~noetic/pls/podlev.html).

Primary Examiner—Ramon M. Barrera

(57) ABSTRACT

A space vehicle propelled by the pressure of inflationary vacuum state is provided comprising a hollow superconductive shield, an inner shield, a power source, a support structure, upper and lower means for generating an electromagnetic field, and a flux modulation controller.

A cooled hollow superconductive shield is energized by an electromagnetic field resulting in the quantized vortices of lattice ions projecting a gravitomagnetic field that forms a spacetime curvature anomaly outside the space vehicle. The spacetime curvature imbalance, the spacetime curvature being the same as gravity, provides for the space vehicle's propulsion. The space vehicle, surrounded by the spacetime anomaly, may move at a speed approaching the light-speed characteristic for the modified locale.

13 Claims, 6 Drawing Sheets

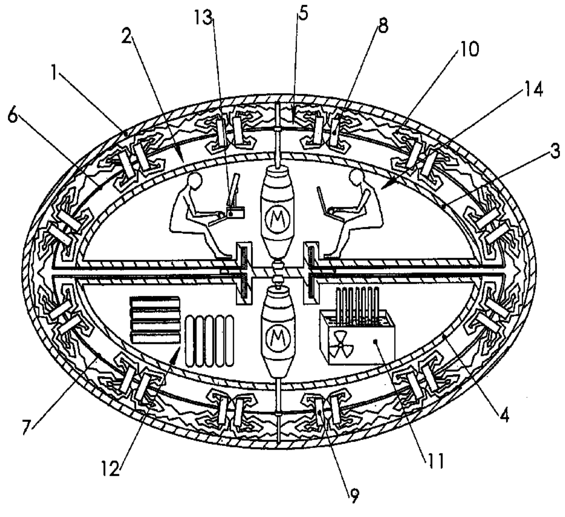

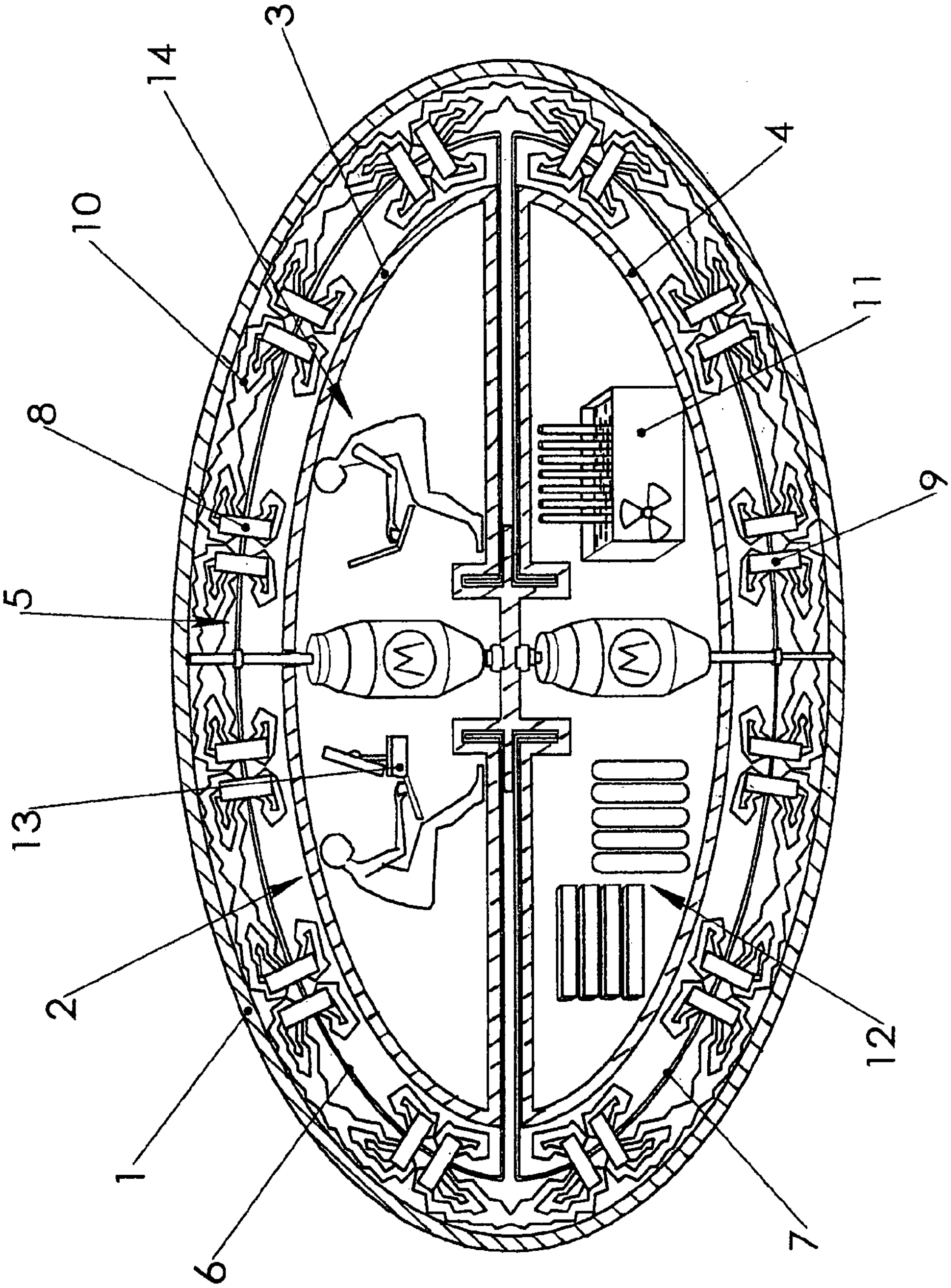
3
14
10
8
5
13
2
1
6
4
11
9
12
7

FIG. 1

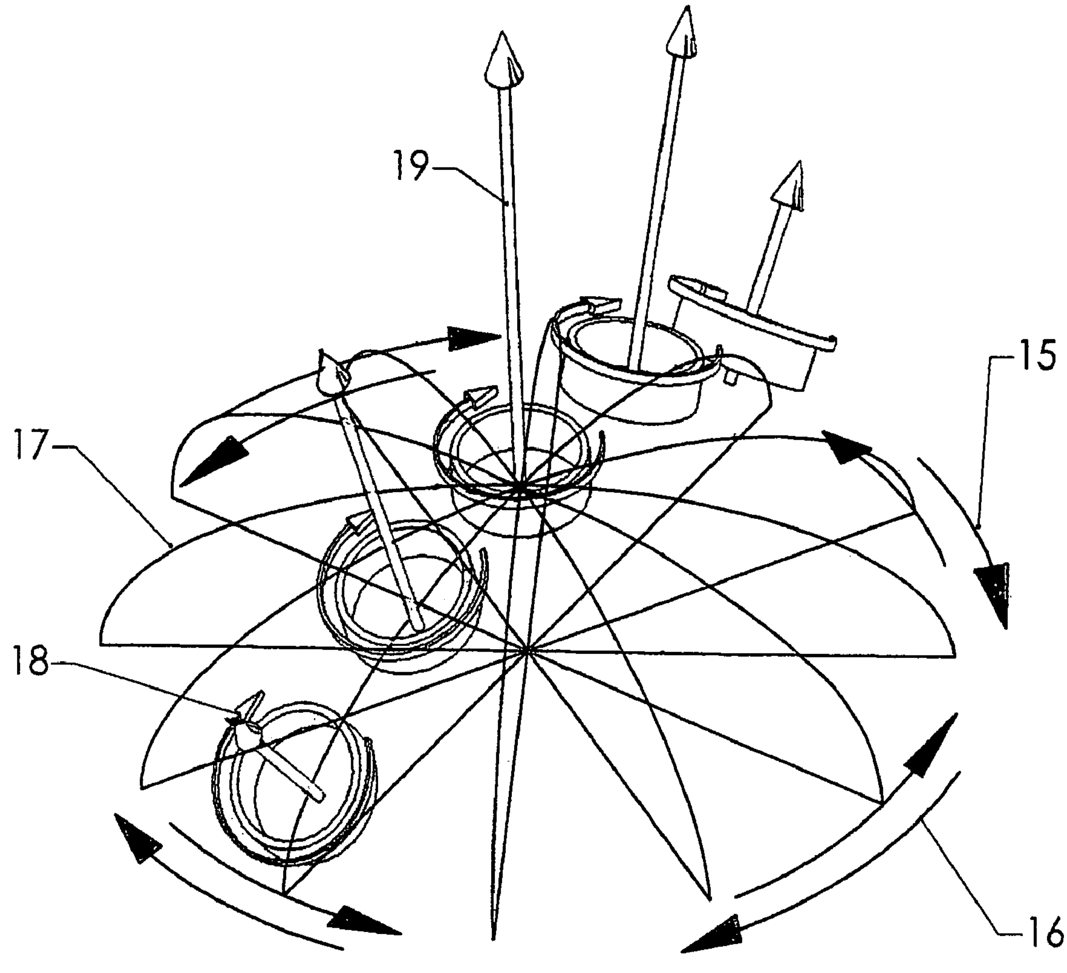
19
15
17
18
16

FIG. 2A

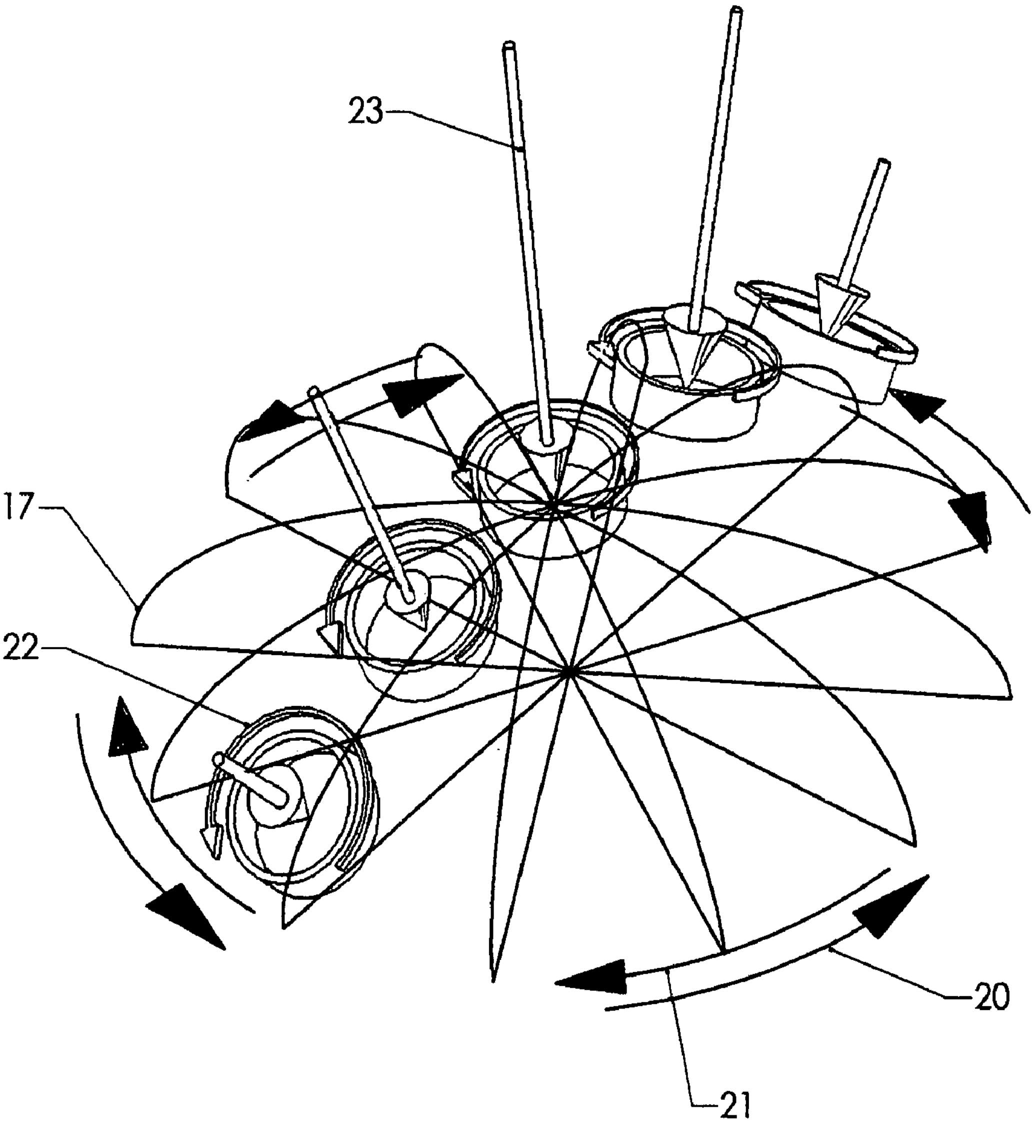
23
17
22
20
21

FIG. 2B

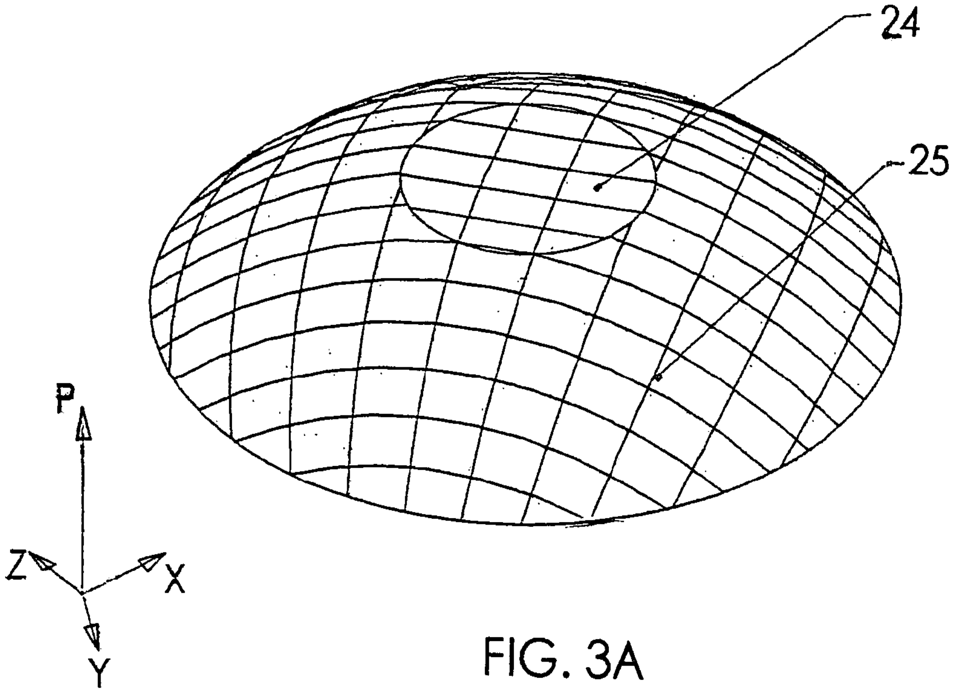

FIG. 3A

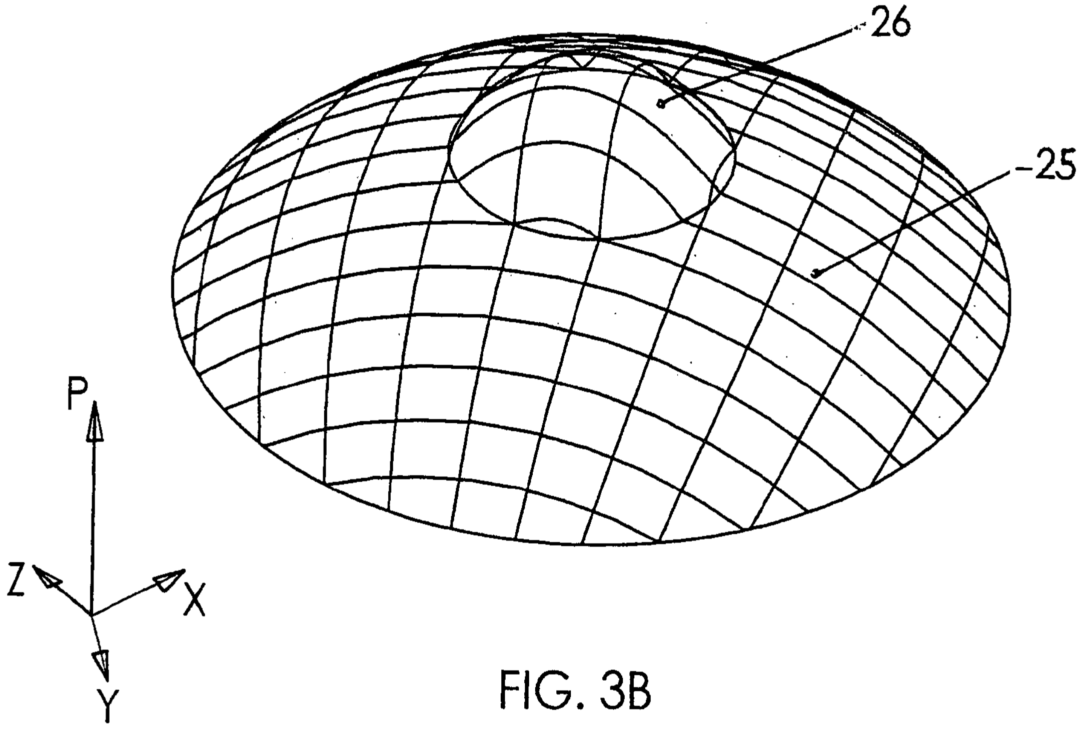

FIG. 3B

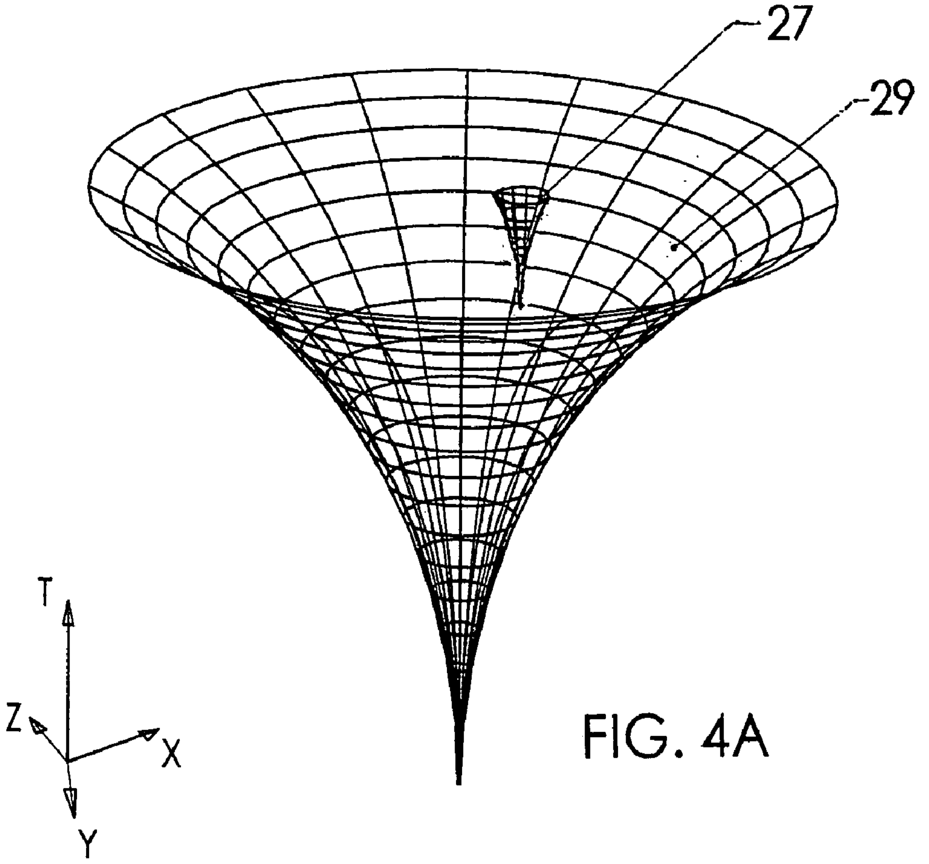

FIG. 4A

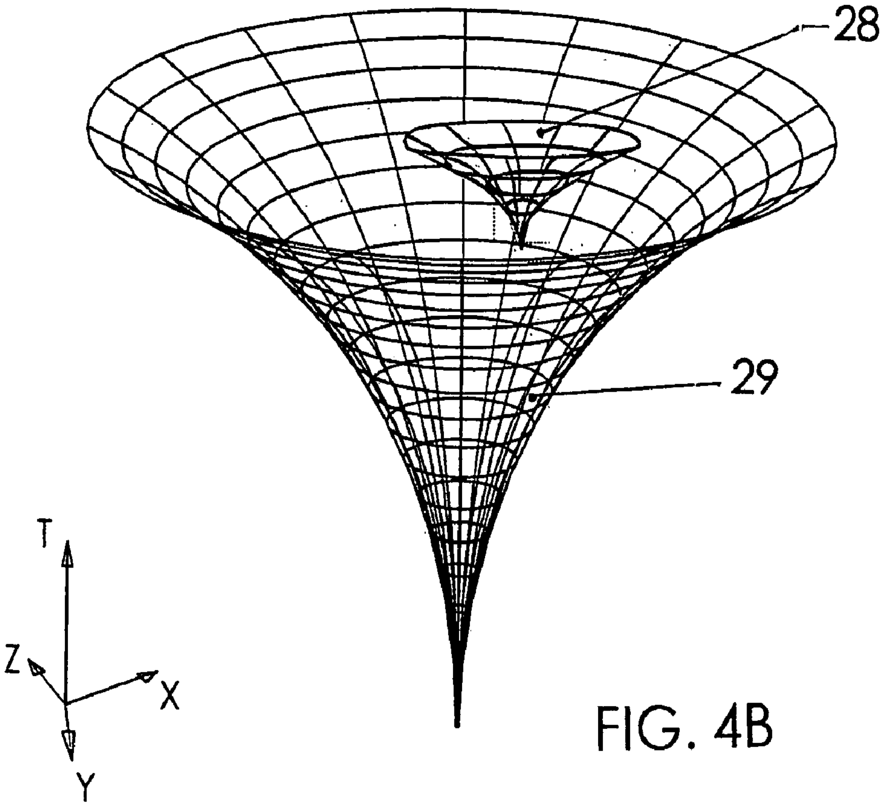

FIG. 4B

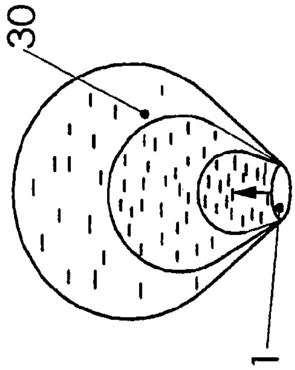

FIG.5A

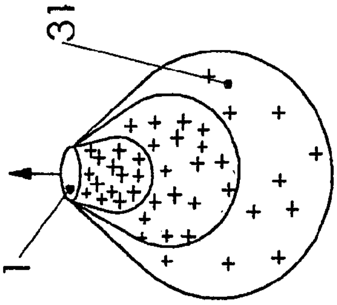

FIG.5B

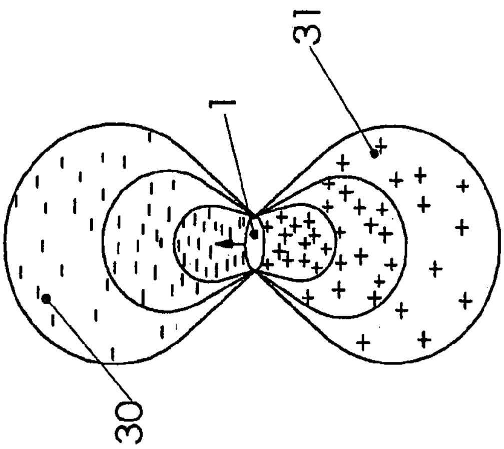

FIG.6

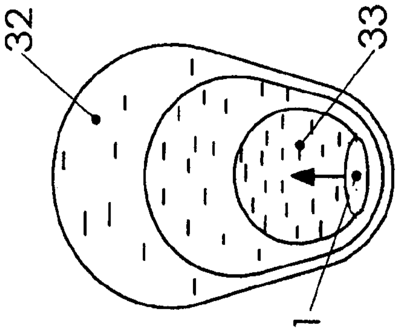

FIG.7A

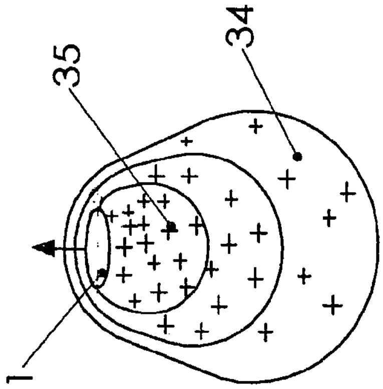

FIG.7B

SPACE VEHICLE PROPELLED BY THE PRESSURE OF INFLATIONARY VACUUM STATE

CROSS REFERENSE TO RELATED APPLICATIONS

This is a continuation of application Ser. No. 10/633,778 filed on Aug. 4, 2003, now abandoned.

TECHNICAL FIELD

The technical field of this disclosure represents space vehicles propelled by the electro-magnetically changed properties of vacuum, particularly the pressure of inflationary vacuum state.

BACKGROUND OF THE INVENTION

The existence of a magnetic-like gravitational field has been well established by physicists for general relativity, gravitational theories, and cosmology. The consequences of the effect of electromagnetically-affected gravity could be substantial and have many practical applications, particularly in aviation and space exploration.

There are methods known for converting electromagnetism into a propulsive force that potentially generates a large propulsive thrust. According to these methods, the machine thrust is produced by rotating, reciprocating masses in the following ways: centrifugal thrust, momentum thrust, and impulse thrust. ("To the Stars by Electromagnetic Propulsion", M. T. French, http://www.mtjf.demon.co.uk/antigravp2.htm#cforce).

However, the electromagnetic propulsion in an ambient space, or space that is not artificially modified, is not practical for interstellar travel because of the great distances involved. No interstellar travel is feasible without some form of distortion of space. In turn, no alteration of space is possible without the corresponding deformation of time. Gravitomagnetic alteration of space, resulting in the space-time curvature anomaly that could propel the space vehicle, could be a feasible approach to future space travel.

In the late 1940s, H. B. G. Casimir proved that the vacuum is neither particle nor field-free. It is a source of zero-point-fluctuation (ZPF) of fields such as the vacuum gravitomagnetic field. ZPF fields lead to real, measurable physical consequences such as the Casimir force. The quantized hand-made electromagnetic processes, such as those occurring in superconductors, affect the similarly quantized ZPFs. The most likely reason is the electron-positron creation and annihilation, in part corresponding to the "polarization effect" sited by Evgeny Podkletnov in explaining the gravitomagnetic effect reportedly observed by him in 1992. ("Weak Gravitational Shielding Properties of Composite Bulk $YBa_2Cu_33O_{(7-x)}$ Superconductor Below 70 K Under E.M. Field", Evgeny Podkletnov, LANL database number cond-mat/9701074, v. 3, 10 pages, 16 Sep. 1997).

The investigation of gravitomagnetism, however, started well before Podkletnov. In the U.S. Pat. No. 3,626,605, Henry Wm. Wallace describes an experimental apparatus for generating and detecting a secondary gravitational field. He also shows how a time-varying gravitomagnetic field can be used to shield the primary background of a gravitoelectric field.

In the U.S. Pat. No. 3,626,606, Henry Wm. Wallace provides a variation of his earlier experiment. A type III-V semiconductor material, of which both components have unpaired nuclear spin, is used as an electronic detector for the gravitomagnetic field. The experiment demonstrates that the material in his gravitomagnetic field circuit has hysterisis and remanence effects analogous to magnetic materials.

In the U.S. Pat. No. 3,823,570, Henry Wm. Wallace provides an additional variation of his experiment. Wallace demonstrates that, by aligning the nuclear spin of materials having an odd number of nucleons, a change in specific heat occurs.

In the U.S. Pat. No. 5,197,279, James R. Taylor discloses Electromagnetic Propulsion Engine where solenoidal windings generate an electromagnetic field that, without the conversion into a gravitomagnetic field, generates the thrust necessary for the propulsion.

In the U.S. Pat. No. 6,353,311 B1, John P. Brainard et al. offer a controversial theory of Universal Particle Flux Field, and in order to prove it empirically, provide a shaded motor-type device. This device is also intended for extracting energy from this hypothetical Field.

In the early 1980s, Sidney Coleman and F. de Luca noted that the Einsteinean postulate of a homogeneous Universe, while correct in general, ignores quantized local fluctuation of the pressure of inflationary vacuum state, this fluctuation causing local cosmic calamities. While the massless particles propagate through large portions of Universe at light speed, these anomaly bubbles, depending on their low or high relative vacuum density, cause a local increase or decrease of the propagation values for these particles. Scientists disagree about the possibility, and possible ways, to artificially create models of such anomalies.

In the early 1990s, Ning Li and D. G Torr described a method and means for converting an electromagnetic field into a gravitomagnetic field. Li and Torr suggested that, under the proper conditions, the minuscule force fields of superconducting atoms can "couple", compounding in strength to the point where they can produce a repulsion force ("Effects of a Gravitomagnetic Field on Pure Superconductors", N. Li and D. G. Torr, Physical Review, Volume 43, Page 457, 3 pages, 15 Jan. 1991).

A series of experiments, performed in the early 1990s by Podkletnov and R. Nieminen, reportedly resulted in a reduction of the weights of objects placed above a levitating, rotating superconductive disk subjected to high frequency magnetic fields. These results substantially support the expansion of Einstainean physics offered by Li & Torr. Podkletnov and Giovanni Modanese have provided a number of interesting theories as to why the weight reduction effect could have occurred, citing quantum gravitational effects, specifically, a local change in the cosmological constant. The cosmological constant, under ordinary circumstances, is the same everywhere. But, according to Podkletnov and Modanese, above a levitating, rotating superconductive disk exposed to high frequency magnetic fields, it is modified. ("Impulse Gravity Generator Based on Charged $YBa_2Cu_3O_{7-y}$ Superconductor with Composite Crystal Structure", Evgeny Podkletnov, Giovanni Modanese, arXiv.org/physics database, #0108005 volume 2, 32 pages, 8 figures, Aug. 30, 2001).

In the July 2004 paper, Ning Wu hypothesized that exponential decay of the gravitation gauge field, characteristic for the unstable vacuum such as that created by Podkletnov and Nieminen, is at the root of the gravitational shielding effects (Gravitational Shielding Effects in Gauge Theory of Gravity, Ning Wu, arXiv:hep-th/ 0307225 v 1 23 Jul. 2003, 38 pages incl. 3 figures, July 2004).

In 2002, Edward Fomalont and Sergei Kopeikin measured the speed of propagation of gravity. They confirmed that the speed of propagation of gravity matches the speed of light. ("General Relativistic Model for Experimental Measurement of the Speed of Propagation of Gravity by VLBI", S. Kopeikin and E. Fomalont, Proceedings of the 6^{th} European VLBI Network Symposium Jun. 25–28 2002, Bonn, Germany, 4 pages).

String theory unifies gravity with all other known forces. According to String theory, all interactions are carried by fundamental particles, and all particles are just tiny loops of space itself forming the spacetime curvature. Gravity and bent space are the same thing, propagating with the speed of light characteristic of the particular curvature. In light of the Fomalont and Kopeikin discovery, one can conclude that if there is a change in the speed of propagation of gravity within the spacetime curvature, then the speed of light within the locality would also be affected.

In general relativity, any form of energy affects the gravitational field, so the vacuum energy density becomes a potentially crucial ingredient. Traditionally, the vacuum is assumed to be the same everywhere in the Universe, so the vacuum energy density is a universal number. The cosmological constant Λ is proportional to the vacuum pressure ρ_Λ: $\Lambda=(8\pi G/3c^2)\rho_\Lambda$, where G is Newton's constant of gravitation and c is the speed of light ("The Cosmological Constant", Sean M. Carroll, http://pancake.uchicago.edu/~carroll/encyc/, 6 pages). Newer theories, however, permit local vacuum fluctuations where even the "universal" constants are affected: $\Lambda_1=(8\pi G_1/3c_1^2)\rho_{\Lambda_1}$.

Analyzing physics laws defining the cosmological constant, a conclusion can be drawn that, if a levitating, rotating superconductive disk subjected to high frequency magnetic fields affects the cosmological constant within a locality, it would also affect the vacuum energy density. According to the general relativity theory, the gravitational attraction is explained as the result of the curvature of spacetime being proportional to the cosmological constant. Thus, the change in the gravitational attraction of the vacuum's subatomic particles would cause a local anomaly in the curvature of the Einsteinean spacetime.

Time is the fourth dimension. Lorentz and Einstein showed that space and time are intrinsically related. Later in his life, Einstein hypothesized that time fluctuates both locally and universally. Ruggero Santilli, recognized for expanding relativity theory, has developed the isocosmology theory, which allows for variable rates of time. Time is also a force field only detected at speeds above light speed. The energy of this force field grows as its propagation speed declines when approaching light-speed. Not just any light-speed: the light-speed of a locale. If the conditions of the locale were modified, this change would affect the local time rate relative to the rate outside the affected locale, or ambient rate. The electromagnetically-generated gravitomagnetic field could be one such locale modifier.

Analyzing the expansion of Einstainean physics offered by Li & Torr, one could conclude that gravity, time, and light speed could be altered by the application of electromagnetic force to a superconductor.

By creating a spacetime curvature anomaly associated with lowered pressure of inflationary vacuum state around a space vehicle, with the lowest vacuum pressure density located directly in front of the vehicle, a condition could be created where gravity associated with lowered vacuum pressure density pulls the vehicle forward in modified spacetime.

By creating a spacetime curvature anomaly associated with elevated pressure of inflationary vacuum state around the space vehicle, with the point of highest vacuum pressure density located directly behind the vehicle, a condition could be created where a repulsion force associated with elevated vacuum pressure density pushes the space vehicle forward in modified spacetime. From the above-mentioned cosmological constant equation, re-written as:

$$\rho_\Lambda = \frac{3c^2}{8\pi G}\Lambda$$

it is clear that the increase in the vacuum pressure density could lead to a substantial increase in the light-speed. If the space vehicle is moving in the anomaly where the local light-speed is higher than the light-speed of the ambient vacuum, and if this vehicle approaches this local light-speed, the space vehicle would then possibly exceed the light-speed characteristic for the ambient area.

The levitating and rotating superconductor disk, which Podkletnov used to protect the object of experiment from the attraction produced by the energy of the vacuum, was externally energized by the externally-powered solenoid coils. Thus, Podkletnov's system is stationary by definition and not suitable for travel in air or space. Even if the superconductive disk is made part of the craft, and if it is energized by the energy available on the craft, the resulting anomaly is one-sided, not enveloping, and not providing the variable speed of light (VSL) environment for the craft.

In a recent (2002) article, Chris Y. Tailor and Modanese propose to employ an impulse gravity generator directing, from an outside location, an anomalous beam toward a spacecraft, this beam acting as a repulsion force field producing propulsion for the spacecraft. ("Evaluation of an Impulse Gravity Generator Based Beamed Propulsion Concept", Chris Y. Taylor and Giovanni Modanese, American Institute of Aeronautics and Astronautics, Inc., 2002, 21 pages, 10 figures). The authors of the article, however, didn't take into account the powerful quantized processes of field dispersion, which would greatly limit the distance of propagation of the repulsive force. At best, the implementation of this concept could assist in acceleration and deceleration at short distances from the impulse gravity generator, and only along a straight line of travel. If the travel goal is a space exploration mission rather than the shuttle-like commute, the proposed system is of little use.

Only a self-sufficient craft, equipped with the internal gravity generator and the internal energy source powering this generator, would have the flexibility needed to explore new frontiers of space. The modification of the spacetime curvature all around the spacecraft would allow the spacecraft to approach the light-speed characteristic for the modified locale, this light-speed, when observed from a location in the ambient space, being potentially many times higher than the ambient light-speed. Then, under sufficient local energies, that is, energies available on the spacecraft, very large intergalactic distances could be reduced to conventional planetary distances.

In "The First Men in the Moon" (1903), H. G. Wells anticipates gravitational propulsion methods when he describes gravity repelling "cavorite." Discovered by Professor Cavor, the material acts as a "gravity shield" allowing Cavor's vehicle to reach the Moon. Prof. Cavor built a large spherical gondola surrounded on all sides by cavorite shutters that could be closed or opened. When Prof. Cavor closed all the shutters facing the ground and opened the shutters facing the moon, the gondola took off for the Moon.

Until today, no cavorite has been discovered. However, recent research in the area of superconductivity, nanomaterials and quantum state of vacuum, including that of Li, Torr, Podkletnov, and Modanese, has resulted in important new information about the interaction between a gravitational field and special states of matter at a quantum level. This new research opens the possibility of using new electromagnetically-energized superconductive materials allowing stable states of energy, the materials useful not only in controlling the local gravitational fields, but also in creating new gravitomagnetic fields.

BACKGROUND OF INVENTION: OBJECTS AND ADVANTAGES

There are four objects of this invention:

The first object is to provide a method for generating a pressure anomaly of inflationary vacuum state that leads to electromagnetic propulsion.

The second object is to provide a space vehicle capable of electromagnetically-generated propulsion. The implementation of these two objects leads to the development of the space vehicle propelled by gravitational imbalance with gravity pulling, and/or antigravity pushing, the space vehicle forward.

The third object is to provide a method for generating a pressure anomaly of inflationary vacuum state, specifically, the local increase in the level of vacuum pressure density associated with the greater curvature of spacetime. The speed of light in such an anomaly would be higher than the speed of light in the ambient space.

The fourth object is to provide the space vehicle capable of generating an unequally-distributed external anomaly all around this vehicle, specifically the anomaly with the elevated level of vacuum pressure density. The anomaly is formed in such a way that gravity pulls the space vehicle forward in the modified spacetime at a speed possibly approaching the light-speed specific for this modified locale. If the vacuum pressure density of the locale is modified to be substantially higher than that of the ambient vacuum, the speed of the vehicle could conceivably be higher than the ambient light-speed.

SUMMARY OF THE INVENTION

This invention concerns devices self-propelled by the artificially changed properties of the pressure of inflationary vacuum state to speeds possibly approaching the light-speed specific for this modified locale. Furthermore, this invention concerns devices capable of generating the spacetime anomaly characterized by the elevated vacuum pressure density. The devices combining these capabilities may be able to move at speeds substantially higher than the light-speed in the ambient space.

The device of this invention is a space vehicle. The outside shell of the space vehicle is formed by a hollow disk, sphere, or the like hollowed 3-dimensional shape made of a superconductor material, hereinafter a hollow superconductive shield. An inner shield is disposed inside the hollow superconductive shield. The inner shield is provided to protect crew and life-support equipment inside.

A support structure, upper means for generating an electromagnetic field and lower means for generating an electromagnetic field are disposed between the hollow superconductive shield and the inner shield. A flux modulation controller is disposed inside the inner shield to be accessible to the crew.

Electrical energy is generated in a power source disposed inside the hollow superconductive shield. The electrical energy is converted into an electromagnetic field in the upper means for generating an electromagnetic field and the lower means for generating an electromagnetic field.

Electrical motors, also disposed inside the hollow superconductive shield, convert the electrical energy into mechanical energy.

The mechanical energy and the electromagnetic field rotate the hollow superconductive shield, and the upper and the lower means for generating an electromagnetic field, against each other.

The electromagnetic field is converted into a gravitomagnetic field in the hollow superconductive shield.

The gravitomagnetic field, propagated outward, orthogonally to the walls of the hollow superconductive shield, forms a pressure anomaly of inflationary vacuum state in the area of propagation. The pressure anomaly of inflationary vacuum state is comprised of an area of relatively lower vacuum pressure density in front of the space vehicle and an area of relatively higher vacuum pressure density behind the vehicle.

The difference in the vacuum pressure density propels the space vehicle of this invention forward.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** is a cross-sectional view through the front plane taken along the central axis of a space vehicle provided by the method and device of this invention.

FIGS. **2**A & **2**B are diagrams, presented as perspective views, showing some of the physical processes resulting from a dynamic application of an electromagnetic field to a hollow superconductive shield. Only one line of quantized vortices, shown out of scale, is presented for illustration purposes.

FIGS. **3**A & **3**B are diagrams, presented as perspective views, showing a vacuum pressure density anomaly associated with lowered pressure of inflationary vacuum state and a vacuum pressure density anomaly associated with elevated pressure of inflationary vacuum state, respectively. Both anomalies are shown on the background of Universal curvature of inflationary vacuum state.

FIGS. **4**A & **4**B are diagrams, presented as perspective views, showing a spacetime anomaly associated with lowered pressure of inflationary vacuum state and a spacetime anomaly associated with elevated pressure of inflationary vacuum state, respectively. Both anomalies are shown on the background of Universal spacetime.

FIGS. **5**A, **5**B, **6**, **7**A, & **7**B are diagrams of spacetime curvature anomalies generated by the space vehicle of the current invention, these anomalies providing for the propulsion of the space vehicle.

DRAWINGS—REFERENCE NUMERALS

1 hollow superconductive shield
2 inner shield
3 upper shell
4 lower shell
5 support structure
6 upper rotating element
7 lower rotating element
8 upper means for generating an electromagnetic field
9 lower means for generating an electromagnetic field
10 flux lines
11 power source

12 life-support equipment
13 flux modulation controller
14 crew
15 clockwise shield motion vector
16 counterclockwise EMF motion vector
17 wire grid
18 clockwise quantized vortices of lattice ions
19 outward gravitomagnetic field vector
20 counterclockwise shield motion vector
21 clockwise EMF motion vector
22 counterclockwise quantized vortices of lattice ions
23 inward gravitomagnetic field vector
24 vacuum pressure density anomaly associated with lowered pressure of inflationary vacuum state
25 Universal curvature of inflationary vacuum state
26 vacuum pressure density anomaly associated with elevated pressure of inflationary vacuum state
27 spacetime anomaly associated with lowered pressure of inflationary vacuum state
28 spacetime anomaly associated with elevated pressure of inflationary vacuum state
29 Universal spacetime
30 substantially droplet-shaped spacetime curvature anomaly associated with lowered pressure of inflationary vacuum state
31 substantially droplet-shaped spacetime anomaly associated with elevated pressure of inflationary vacuum state
32 substantially egg-shaped spacetime anomaly associated with lowered pressure of inflationary vacuum state
33 area of the lowest vacuum pressure density
34 substantially egg-shaped spacetime anomaly associated with elevated pressure of inflationary vacuum state
35 area of the highest vacuum pressure density

DESCRIPTION OF THE PRESENTLY PREFERRED EMBODIMENT

FIG. **1** is a cross-sectional view through the front plane taken along the central axis of a space vehicle provided by the method and device of this invention. A hollow superconductive shield **1** forms a protective outer shell of the space vehicle. The hollow superconductive shield **1** may be shaped as a hollow disk, sphere, or the like 3-dimensional geometrical figure formed by the 2-dimensional rotation of a curve around the central axis.

In the preferred embodiment, the hollow superconductive shield **1** is made of a superconductor such as YBa2Cu3O7-y, or a like high-temperature superconductor with a composite crystal structure cooled to the temperature of about 40 degrees K. The skilled in the art may envision the use of many other low- and high-temperature superconductors, all within the scope of this invention.

An inner shield **2** is disposed inside the hollow superconductive shield **1**. The inner shield **2** is comprised of an upper shell **3** and a lower shell **4**, the shells **3** and **4** adjoined with each other. Executed from insulation materials such as foamed ceramics, the inner shield **2** protects the environment within the shield from the electromagnetic field and severe temperatures.

A support structure **5** is disposed between the hollow superconductive shield **1** and the inner shield **2**, concentric to the hollow superconductive shield. The support structure **5** is comprised of an upper rotating element **6** and a lower rotating element **7**.

The upper rotating element **6** is pivotably disposed inside the hollow superconductive shield **1** and may envelope the upper shell **3**. The lower rotating element **7** is pivotably disposed inside the hollow superconductive shield **1** and may envelope the lower shell **4**. Even though the preferred embodiment has two rotating elements, a skilled in the art may envision only one rotating element, or three or more rotation elements, all within the scope of this invention.

Upper means for generating an electromagnetic field **8** are disposed between the hollow superconductive shield **1** and the upper shell **3**. The upper means for generating an electromagnetic field **8** are affixed to the upper rotating element **6** at an electromagnetic field-penetrable distance to the hollow superconductive shield **1**.

Lower means for generating an electromagnetic field **9** are disposed between the hollow superconductive shield **1** and the lower shell **4**. The lower means for generating an electromagnetic field **9** are affixed to the lower rotating element **7** at an electromagnetic field-penetrable distance to the hollow superconductive shield **1**.

The upper means for generating an electromagnetic field **8** and the lower means for generating an electromagnetic field **9** could be solenoid coils or electromagnets. In the process of operation of the space vehicle, the electromagnetic field identified by flux lines **10**, is controllably and variably applied to the hollow superconductive shield **1**.

Electric motors are disposed inside the hollow superconductive shield along its central axis.

A power source **11** is disposed inside the hollow superconductive shield **1** and may be disposed inside the lower shell **4**. The power source **11** is electrically connected with the upper means for generating an electromagnetic field **8**, the lower means for generating an electromagnetic field **9**, and the electric motors. In-turn, the upper means for generating an electromagnetic field **8**, the lower means for generating an electromagnetic field **9**, and the electric motors provide for the rotation of the upper rotating element **6** and the lower rotating element **7**. The power source **11** may be executed as a nuclear power generator.

Life-support equipment **12** is disposed inside the inner shield **2**, and may be disposed inside the lower shell **4**. The life-support equipment **12** may include oxygen, water, and food.

A flux modulation controller **13** is disposed inside the inner shield **2**, and may be disposed inside the upper shell **3**. The flux modulation controller **13** is in communication with the upper means for generating an electromagnetic field **8**, the lower means for generating an electromagnetic field **9**, the power source **11**, and the electric motors.

The flux modulation controller **8** may be executed as a computer or a microprocessor. The flux modulation controller **8** is provided with a capability of modulating the performance parameters of the upper means for generating an electromagnetic field **8**, the lower means for generating an electromagnetic field **9**, the power source **11**, and the electric motors.

A crew **14** may be located inside the upper shell **3** of the inner shield **2** and may consist of one or more astronauts. The crew has a free access to the life-support equipment **12** and the flux modulation controller **8**. A skilled in the art may envision a fully-automated, pilotless craft, which is also within the scope of this invention.

A skilled in the art may also envision the embodiment (not shown), also within the scope of this invention, where the hollow superconductive shield is pivotable, and the support structure with the means for generating an electromagnetic field is affixed on the outside of the inner shield.

FIGS. **2A** & **2B** are diagrams showing the results of the quantized electromagnetic turbulence within the superconductive shell of the hollow superconductive shield provided by the relative rotational motion of the hollow superconductive shield against the upper means for generating an electromagnetic field.

FIG. **2A** shows the clockwise relative rotational motion of the hollow superconductive shield, this motion identified by a clockwise shield motion vector **15**, and the counterclockwise relative rotational motion of upper means for generating an electromagnetic field, this motion identified by a counterclockwise EMF motion vector **16**.

The electromagnetic field, controllably and variably applied by the upper means for generating an electromagnetic field, whose various positions are identified by a wire grid **17**, to the hollow superconductive shield (not shown), causes quantized electromagnetic turbulence within the hollow superconductive shield. This turbulence is represented by a plurality of clockwise quantized vortices of lattice ions **18**. Only one line of the clockwise quantized vortices of lattice ions **18**, out of scale, is shown for illustration purposes only. Each of the clockwise quantized vortices of lattice ions **18** generates a gravitomagnetic field identified by an outward gravitomagnetic field vector **19** directed orthogonally away from the hollow superconductive shield.

FIG. **2B** shows the counterclockwise relative rotational motion of the hollow superconductive shield, this motion identified by a counterclockwise shield motion vector **20**, and the clockwise relative rotational motion of upper means for generating an electromagnetic field, this motion identified by a clockwise EMF motion vector **21**.

The electromagnetic field, controllably and variably applied by the upper means for generating an electromagnetic field identified by the wire grid **17**, to the hollow superconductive shield (not shown), causes quantized electromagnetic turbulence within the hollow superconductive shield, this turbulence represented by a plurality of counterclockwise quantized vortices of lattice ions **22**. Only one line of the counterclockwise quantized vortices of lattice ions **22**, out of scale, is shown for illustration purposes only. Each of the counterclockwise quantized vortices of lattice ions **22** generates a gravitomagnetic field identified by an inward gravitomagnetic field vector **23** directed orthogonally toward the hollow superconductive shield.

The electrical requirements for providing the Li-Torr effect are as follows:

Podkletnov has reported using the high frequency current of 10^5 Hz. He also used 6 solenoid coils @ 850 Gauss each. The reported system's efficiency reached 100% and the total field in the Podkletnov's disk was about 0.5 Tesla. The maximum weight loss reported by Podkletnov was 2.1%.

The preferred embodiment of the device of current invention is capable of housing 2–3 astronauts and therefore is envisioned to be about 5 meters in diameter at the widest point. The preferred space vehicle's acceleration is set at 9.8 m/s/s providing that gravity on board is similar to that on the surface of Earth.

The means for generating an electromagnetic field may be comprised of 124 solenoid coils. At the same 100% efficiency reported by Podkletnov, the total field required providing the acceleration of 9.8 m/s/s is 5,000 Tesla, or about 40 Tesla per coil. Skeggs suggests that on the Podkletnov device, out of 850 Gauss developed on the coil surface, the field affecting the superconductor and causing the gravitomagnetism is only 400 Gauss ("Engineering Analysis of the Podkletnov Gravity Shielding Experiment, Peter L. Skeggs, Quantum Forum, Nov. 7, 1997, http://www.inetarena.com/~noetic/pls/podlev.html, 7 pages). This translates into 47% device efficiency.

In this 47%-efficient space vehicle, the total field required achieving the 9.8 m/s/s acceleration is about 10,600 Tesla, or 85.5 Tesla per each of 124 solenoid coils. It must be noted that at this acceleration rate, it would take nearly a year for the space vehicle to reach the speed of light.

It also must be noted that Skeggs has detected a discrepancy between the Li-Torr estimates and Podkletnov's practical results. If Podkletnov's experimental results are erroneous while the Li-Torr estimates are indeed applicable to the space vehicle of this invention, then the energy requirements for achieving the sought speed would be substantially higher than the above estimate of 10,600 Tesla.

Podkletnov has concluded that, in order for the vacuum pressure density anomaly to take place, the Earth-bound device must be in the condition of Meissner levitation. As are all space bodies, the space vehicle is a subject to the pressure inflationary vacuum state and the gravitational force, which, within the migrating locality of the expanding Universe, in any single linear direction, are substantially in equilibrium. Thus, for the space vehicle, the requirement of Meissner levitation is waved.

The propagation of the gravitomagnetic field identified by the outward gravitomagnetic field vector **19** and the inward gravitomagnetic field vector **23** would cause exotic quantized processes in the vacuum's subatomic particles that include particle polarization, ZPF field defects, and the matter-energy transformation per $E=mc^2$. The combination of these processes would result in the gravitational anomaly. According to the general relativity theory, gravitational attraction is explained as the result of the curvature of spacetime being proportional to the gravitational constant. Thus, the change in the gravitational attraction of the vacuum's subatomic particles would cause a local anomaly in the curvature of the Einsteinean spacetime.

Gravity is the same thing as bent space, propagating with the speed of light characteristic for the particular spacetime curvature. When bent space is affected, there is a change in the speed of propagation of gravity within the spacetime curvature anomaly. The local speed of light, according to Fomalont and Kopeikin always equal to the local speed of propagation of gravity, is also affected within the locality of spacetime curvature anomaly.

Creation of spacetime curvature anomalies adjacent to, or around, the space vehicle, these anomalies characterized by the local gravity and light-speed change, has been the main object of this invention.

FIG. **3A** shows a diagram of a vacuum pressure density anomaly associated with lowered pressure of inflationary vacuum state **24** on the background of Universal curvature of inflationary vacuum state **25**. The vacuum pressure density anomaly associated with lowered pressure of inflationary vacuum state **24** is formed by a multitude of the inward gravitomagnetic field vectors. According to the cosmological constant equation, $\Lambda=(8\pi G/3c^2)\rho_\Lambda$, where the cosmological constant Λ is proportional to the vacuum energy pressure ρ_Λ, G is Newton's constant of gravitation, and c is the speed of light, the curvature of spacetime is proportional to the gravitational constant. According to the general relativity theory, the change in the vacuum pressure density is proportional to the change in the spacetime curvature anomaly. By replacing ρ_Λ with the vacuum pressure density, P times the vacuum energy coefficient κ, and replacing c with Δdistance/Δtime, we derive to $\Lambda=[8\pi G/3(\Delta\text{distance}/\Delta\text{time})^2]P_\kappa$, and could now construct a vacuum pressure density curvature diagram.

The vacuum pressure density curvature anomaly associated with lowered pressure of inflationary vacuum state **24** is shown here as a flattened surface representing the lowered pressure of the inflationary vacuum state. This anomaly is the result of the exotic quantized processes in the subatomic particles caused by the quantized turbulence occurring in the hollow superconductive shield. The XYZ axes represent three dimensions of space and the P axis represents the vacuum pressure density.

FIG. **3**B shows a diagram of a vacuum pressure density anomaly associated with elevated pressure of inflationary vacuum state **26** on the background of the Universal curvature of inflationary vacuum state **25**. The vacuum pressure density anomaly associated with elevated pressure of inflationary vacuum state **26** is formed by a multitude of the outward gravitomagnetic field vectors. The anomaly is shown here as a convex surface representing the elevated pressure of inflationary vacuum state. The diagrams of FIGS. **3**A & **3**B are not to scale: the anomaly sizes are exaggerated for clarity.

FIGS. **4**A & **4**B show diagrams of a spacetime anomaly associated with lowered pressure of inflationary vacuum state **27**, and a spacetime anomaly associated with elevated pressure of inflationary vacuum state **28**, respectively, each on the background a diagram of Universal spacetime **29**.

The quaterized Julia set $Q_{n+1}=Q_n^2+C_0$ is assumed to be an accurate mathematical representation of the Universal spacetime. The generic quaternion Q_0 belongs to the Julia set associated with the quaternion C, and n tends to infinity. If we assume that the quaternion value C_0 is associated with the Universal spacetime **29**, C_1 is the value of quaternion C for the spacetime anomaly associated with lowered pressure of inflationary vacuum state **27**, and C_2 is the value of quaternion C for the spacetime anomaly associated with elevated pressure of inflationary vacuum state **28**, then we can construct two diagrams.

The diagram of FIG. **4**A shows the spacetime anomaly associated with lowered pressure of inflationary vacuum state **27** as a quaterized Julia set contained in a 4-dimensional space: $Q_{n+1}=Q_n^2+C_1$ on the background of the Universal spacetime **29** represented by $Q_{n+1}=Q_n^2+C_0$.

The diagram of FIG. **4**B shows the spacetime anomaly associated with elevated pressure of inflationary vacuum state **28** as a quaterized Julia set $Q_{n+1}=Q_n^2+C_2$, also on the background of the Universal spacetime **29** represented by $Q_{n+1}=Q_n^2+C_0$. On both diagrams, the XYZ axes represent three dimensions of space, and the T axis represents time. The diagrams are not to scale: the anomaly sizes are exaggerated for clarity, and the halves of quaterized Julia sets, conventionally associated with the hypothetical Anti-Universe, are omitted.

FIGS. **5**A, **5**B, **6**, **7**A, & **7**B show simplified diagrams of spacetime curvature anomalies generated by the space vehicle of the current invention, these anomalies providing for the propulsion of the space vehicle. In each case, the pressure anomaly of inflationary vacuum state is comprised of an area of relatively lower vacuum pressure density in front of the space vehicle and an area of relatively higher vacuum pressure density behind the space vehicle. Because the lower pressure of inflationary vacuum state is associated with greater gravity and the higher pressure is associated with the higher repulsive force, the space vehicle is urged to move from the area of relatively higher vacuum pressure density toward the area of relatively lower vacuum pressure density.

FIG. **5**A illustrates the first example of spacetime curvature modification. This example shows a substantially droplet-shaped spacetime curvature anomaly associated with lowered pressure of inflationary vacuum state **30** adjacent to the hollow superconductive shield **1** of the space vehicle. The anomaly **30** is provided by the propagation of a gravitomagnetic field radiating orthogonally away from the front of the hollow superconductive shield **1**. This gravitomagnetic field may be provided by the relative clockwise motion of the upper means for generating an electromagnetic field, and relative counterclockwise motion of the hollow superconductive field, as observed from above the space vehicle.

In this example, the difference between the spacetime curvature within the substantially droplet-shaped spacetime anomaly associated with lowered pressure of inflationary vacuum state, and the ambient spacetime curvature, the spacetime curvature being the same as gravity, results in the gravitational imbalance, with gravity pulling the space vehicle forward.

FIG. **5**B illustrates the second example of spacetime curvature modification. This example shows a substantially droplet-shaped spacetime anomaly associated with elevated pressure of inflationary vacuum state **31** adjacent to the hollow superconductive shield **1** of the space vehicle. The anomaly **31** is provided by the propagation of a gravitomagnetic field radiating orthogonally away from the back of the hollow superconductive shield. This gravitomagnetic field may be provided by the relative counterclockwise motion of the lower means for generating an electromagnetic field, and relative clockwise motion of the hollow superconductive field, as observed from below the space vehicle.

In this example, the difference between the spacetime curvature within the substantially droplet-shaped spacetime anomaly associated with elevated pressure of inflationary vacuum state, and the ambient spacetime curvature, the spacetime curvature being the same as gravity, results in the gravitational imbalance, with the repulsion force pushing the space vehicle forward.

FIG. **6** illustrates the third example of spacetime curvature modification. This example shows the formation of the substantially droplet-shaped spacetime anomaly associated with lowered pressure of inflationary vacuum state **30** combined with the substantially droplet-shaped spacetime anomaly associated with elevated pressure of inflationary vacuum state **31**. This combination of anomalies may be provided by the relative clockwise motion of the upper means for generating an electromagnetic field and relative clockwise motion of the hollow superconductive field, combined with the relative clockwise motion of the lower means for generating an electromagnetic field, as observed from above the space vehicle.

In this example, the difference between the spacetime curvature within the substantially droplet-shaped spacetime anomaly associated with lowered pressure of inflationary vacuum state, and the spacetime curvature of the substantially droplet-shaped spacetime anomaly associated with elevated pressure of inflationary vacuum state, the spacetime curvature being the same as gravity, results in the gravitational imbalance, with gravity pulling, and the repulsion force pushing, the space vehicle forward.

FIG. **7**A illustrates the fourth example of spacetime curvature modification. This example shows the formation of a substantially egg-shaped spacetime anomaly associated with lowered pressure of inflationary vacuum state **32** around the hollow superconductive shield **1** of the space vehicle. The anomaly **32** is provided by the propagation of gravitomagnetic field of unequally-distributed density, this gravitomagnetic field radiating in all directions orthogonally away from the hollow superconductive shield. The propagation of the unequally-distributed gravitomagnetic field leads to the similarly unequally-distributed spacetime curvature anomaly. This unequally-distributed gravitomagnetic field may be provided by the relatively faster clockwise motion of the upper means for generating an electromagnetic field relative to the hollow superconductive field, combined with the relatively slower counterclockwise motion of the lower means for generating an electromagnetic field, as observed from above the space vehicle.

An area of the lowest vacuum pressure density **33** of the substantially egg-shaped spacetime anomaly associated with lowered pressure of inflationary vacuum state **32** is located directly in front of the space vehicle.

In this example, the variation in the spacetime curvature within the substantially egg-shaped spacetime anomaly associated with lowered pressure of inflationary vacuum state, the spacetime curvature being the same as gravity, results in a gravitational imbalance, with gravity pulling the space vehicle forward in modified spacetime.

FIG. **7**B illustrates the fifth example of spacetime curvature modification, also with the purpose of providing for a propulsion in modified spacetime. This example shows the formation of a substantially egg-shaped spacetime anomaly associated with elevated pressure of inflationary vacuum state **34** around the hollow superconductive shield **1** of the space vehicle. The anomaly **34** is provided by the propagation of gravitomagnetic field of unequally-distributed density, this gravitomagnetic field radiating in all directions orthogonally away from the hollow superconductive shield. The propagation of the unequally-distributed gravitomagnetic field leads to the similarly unequally-distributed spacetime curvature anomaly. This unequally-distributed gravitomagnetic field may be provided by the relatively slower counterclockwise motion of the upper means for generating an electromagnetic field relative to the hollow superconductive field, combined with the relatively faster clockwise motion of the lower means for generating an electromagnetic field, as observed from above the space vehicle.

An area of the highest vacuum pressure density **35** of the substantially egg-shaped spacetime anomaly associated with elevated pressure of inflationary vacuum state **34** is located directly behind the space vehicle.

In this example, the variation in the spacetime curvature within the substantially egg-shaped spacetime anomaly associated with elevated pressure of inflationary vacuum state, the spacetime curvature being same as gravity, results in a gravitational imbalance, with the repulsion force pushing the space vehicle forward in modified spacetime at speeds approaching the light-speed characteristic for this modified area. This light-speed might be much higher than the light-speed in the ambient space.

By creating alternative anomalies and modulating their parameters, the space vehicle's crew would dilate and contract time and space on demand. The space vehicle, emitting a vacuum pressure modifying, controllably-modulated gravitomagnetic field in all directions, would rapidly move in the uneven spacetime anomaly it created, pulled forward by gravity or pushed by the repulsion force. The time rate zone of the anomaly is expected to have multiple quantized boundaries rather than a single sudden boundary affecting space and time in the immediate proximity of the vehicle. Speed, rate of time, and direction in space could be shifted on demand and in a rapid manner. The modulated light-speed could make the space vehicle suitable for interstellar travel. Because of the time rate control in the newly created isospace, the accelerations would be gradual and the angles of deviation would be relatively smooth. The gravity shielding would further protect pilots from the ill-effects of gravity during rapid accelerations, directional changes, and sudden stops.

What is claimed is:

1. A space vehicle propelled by the pressure of inflationary vacuum state is claimed comprising:

a hollow superconductive shield,

an inner shield, the inner shield disposed inside said hollow superconductive shield, said inner shield comprising an upper shell and a lower shell,

a support structure, the support structure disposed between said hollow superconductive shield and said inner shield concentrically to said hollow superconductive shield, said support structure comprised of an upper rotating element and a lower rotating element,

upper means for generating an electromagnetic field, the upper means for generating an electromagnetic field disposed between said hollow superconductive shield and said upper shell, affixed to said upper rotating element at an electromagnetic field-penetrable distance to said hollow superconductive shield,

lower means for generating an electromagnetic field, the lower means for generating an electromagnetic field disposed between said hollow superconductive shield and said lower shell, affixed to said lower rotating element at an electromagnetic field-penetrable distance to said hollow superconductive shield,

electric motors, the electric motors disposed inside said hollow superconductive shield along the central axis of said hollow superconductive shield,

a power source, the power source disposed inside said hollow superconductive shield, said power source electrically connected with said upper means for generating an electromagnetic field, said lower means for generating an electromagnetic field, and said electric motors,

life-support equipment, the life-support equipment disposed inside said inner shield,

a flux modulation controller, the flux modulation controller disposed inside said inner shield, said flux modulation controller in communication with said upper means for generating an electromagnetic field, said lower means for generating an electromagnetic field, said power source, and said electric motors, and

a crew, the crew disposed inside said inner shield accessibly to said life-support and said flux modulation controller.

2. The space vehicle of claim **1** wherein said hollow superconductive shield is composed of a high-temperature ceramic superconductor with a composite crystal structure, and is cooled to a temperature of about 40 degrees K.

3. The space vehicle of claim **1** wherein said inner shield is composed of insulating materials such as foamed ceramics.

4. The space vehicle of claim **1** wherein said power source is a nuclear generator.

5. The space vehicle of claim **1** wherein said flux modulation controller is a computer.

6. The space vehicle of claim **1** wherein said upper rotating element and said lower rotating element are pivoted, urged to rotate by said upper means for generating an electromagnetic field, said lower means for generating an electromagnetic field, and said electric motors.

7. The space vehicle of claim **1** wherein said hollow superconductive shield is pivoted and urged to rotate by said upper means for generating an electromagnetic field, said lower means for generating an electromagnetic field, and said electric motors.

8. A method for generating pressure anomaly of inflationary vacuum state is claimed comprising the following steps, some occurring simultaneously:

providing a space vehicle comprising a hollow superconductive shield, an inner shield, a support structure, upper means for generating an electromagnetic field, lower means for generating an electromagnetic field, electric motors, a power source, life-support equipment, a flux modulation controller, and a crew,

generating electrical energy in said power source,

converting said electrical energy into an electromagnetic field in said upper means for generating an electromagnetic field and said lower means for generating an electromagnetic field,

converting said electrical energy into mechanical energy in said electrical motors,

rotating said upper means for generating an electromagnetic field and said lower means for generating an electromagnetic field relative to said hollow superconductive shield,

converting said electromagnetic field into a gravitomagnetic field in said hollow superconductive shield,

propagating said gravitomagnetic field outward, orthogonal to the walls of said hollow superconductive shield, and

forming a pressure anomaly of inflationary vacuum state in the area of propagation of said gravitomagnetic field, the pressure anomaly of inflationary vacuum state comprised of an area of relatively lower vacuum pressure density in front of said space vehicle and an area of relatively higher vacuum pressure density behind said vehicle,

whereby said space vehicle is urged to move from said area of relatively higher vacuum pressure density toward said area of relatively lower vacuum pressure density.

9. The method of claim **8** wherein said pressure anomaly of inflationary vacuum state is a substantially droplet-shaped anomaly associated with lowered pressure of inflationary vacuum state provided by the propagation of a gravitomagnetic field radiating orthogonally away from the front of said hollow superconductive shield,

whereby providing for the gravitational imbalance where the lowered pressure of inflationary vacuum state is pulling said space vehicle forward.

10. The method of claim **8** wherein said pressure anomaly of inflationary vacuum state is a substantially droplet-shaped anomaly associated with elevated pressure of inflationary vacuum state provided by the propagation of a gravitomagnetic field radiating orthogonally away from the back of said hollow superconductive shield,

whereby providing for the gravitational imbalance where the elevated pressure of inflationary vacuum state is pushing said space vehicle forward.

11. The method of claim **8** wherein said pressure anomaly of inflationary vacuum state is comprised of a substantially droplet-shaped anomaly associated with lowered pressure of inflationary vacuum state and a substantially droplet-shaped anomaly associated with elevated pressure of inflationary vacuum state,

whereby providing for the gravitational imbalance where the lowered pressure of inflationary vacuum state is pulling, and the elevated pressure of inflationary vacuum state is pushing, said space vehicle forward.

12. The method of claim **8** wherein said pressure anomaly of inflationary vacuum state is a substantially egg-shaped anomaly associated with lowered pressure of inflationary vacuum state provided by the unequally-distributed gravitomagnetic field propagated in all directions orthogonally away from said hollow superconductive shield with an area of lowest vacuum pressure density disposed directly in front said space vehicle,

whereby providing for the gravitational imbalance such that the lowered pressure of inflationary vacuum state is pulling said space vehicle forward in modified spacetime.

13. The method of claim **8** wherein said pressure anomaly of inflationary vacuum state is a substantially egg-shaped anomaly associated with elevated pressure of inflationary vacuum state provided by the unequally-distributed gravitomagnetic field propagated in all directions orthogonally away from said hollow superconductive shield with an area of highest vacuum pressure density disposed directly behind said space vehicle,

whereby providing for the gravitational imbalance such that the highest pressure of inflationary vacuum state is pushing said space vehicle forward in modified spacetime at a speed possibly approaching a local light-speed, the local light-speed which may be substantially higher than the light-speed in the ambient space.

* * * * *

Don't Let The Big Energy Corporations Control You – Free Energy Is HERE

(*+6 rating*, *6 votes*, *rated*)

You need to be a registered member to rate this post.

May 26, 2012 *in* *Activism*, *Preparedness*, *Resistance*, *Science*, *Technology*

by Kellie

The below added text I have been looking into has been an easily made design. Our government goes to no end to stop men and women who help human kind with their awesome God given inventions. I am not afraid to share this as Alex does with his knowledge and research. It is to open the minds of those that are dumbed down by corporate America and the global elite to do as they bid…No…stand up and fight against their system.

Free energy is NOT a hoax as they want you to believe.

If you have any engineer in you or technical expertise – Read…It works !

Jes Ascanius' Version of Nikola Tesla's Aerial System

This sort of information may seem confusing and maybe a little too technical for you, so let me tell you about the practical and useful applications used by Jes Ascanius, a Danish developer, to whom thanks is due for sharing his design. Initially, he set up a system to charge his mobile phone battery overnight from an aerial. Then he went on to produce a full-size Tesla Aerial System as described at the start of this chapter. Let's start with the very simple system and progress from that to the more powerful arrangements.

The initial circuit uses one strand of solid wire which rises vertically to a 700 mm diameter drum where there are some twenty turns. The arrangement is like this:

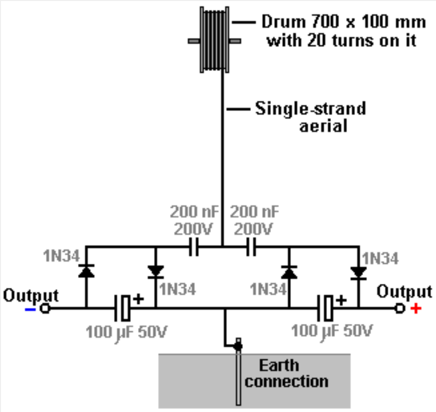

The aerial wire is several metres long, and in the prototype, was supported by (and insulated from) the eaves of a house. The aerial should be vertical or near vertical and a proper earth connection provided by driving a metal rod into the ground or connecting a wire to a metal plate and burying the plate in the ground as a good electrical connection is needed here. The earth connection used here is a 12 mm copper pipe 3 metres long, driven into the ground and the ground around it saturated with water:

The wire used to connect with the earthing rod is very important and should not be less than 8 swg copper wire, that is, 4 mm diameter and 13 sq. mm. cross-sectional area. As with all free-energy devices, the exact constructional details are vital.

The diodes used are germanium 1N34 or 1N34a as germanium diodes drop far less voltage than do silicon diodes and the 1N34 types are low-loss diodes, particularly suited to this application. Ceramic disc types are recommended for the 200 nF capacitors. The prototype build looked like this:

REFEED 24/7
LISTEN NOW

UPCOMING MISSIONS

- **Complete Your 'Base, Loc Interests' Sections Of You**
  Complete by:Thursday Dec 31, 2
- **Flood Alex With Demands Does And Does So Very T**
  Complete by:Friday Jan 1, 2016 /
- **#InfowarsMusic**
  Complete by:Saturday Apr 30, 20
- **Prevent Tyranny In All Nat**
  Complete by:Sunday Jan 1, 2017
- **Collaborators Remembere**
  Complete by:Sunday Dec 31, 201
- **Mission #4: Earth Psychic**
  Complete by:Thursday Dec 31, 2
- **N.D.A.A (National Defense**
  Complete by:Sunday Jul 31, 2022
- **Be The Best You Can Be**
  Complete by:Wednesday Nov 30
- **Agenda 21**
  Complete by:Saturday Dec 31, 2
- **U.N. Gun Arms Trade Trea**
  Complete by:Saturday Dec 31, 2

INSPIRATION SITES

Infowars.com
PrisonPlanet.tv
Infowars Shop
PrisonPlanet.com

Asked by Infowars

How do you fee players kneeling anthem?

I stand with

I stand with

(https://www.squareoffs

INFOWARS.COM NE

UPCOMING EVENTS

There are no upcoming group e

SEARCH:

Search

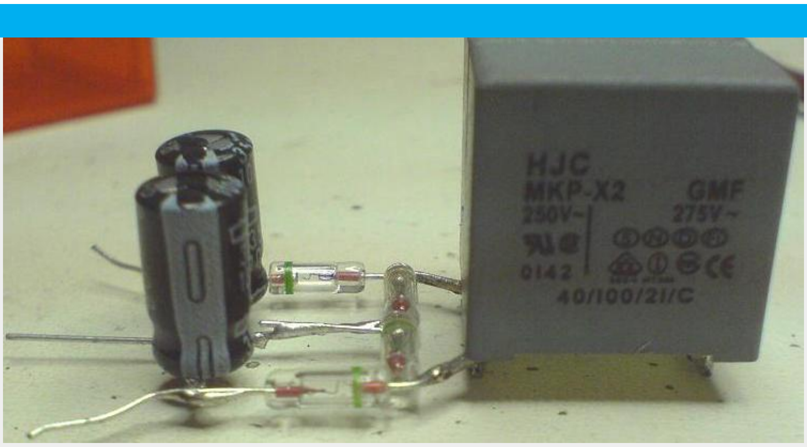

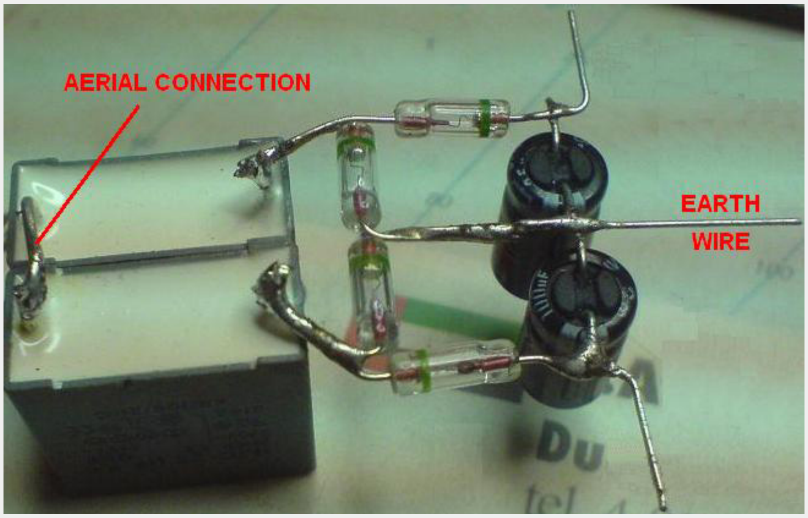

Now, consider this circuit as described, to be one modular building block which can lead to unlimited power from an aerial. I will represent the circuit shown above as a rectangle, showing the above circuit as:

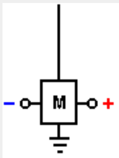

While it is possible to use more than one module with the aerial to get more power, the Danish developer then switched to the full-blown Tesla arrangement by attaching a 800 x 600 x 2 mm aluminium plate inside the sloping roof of his house:

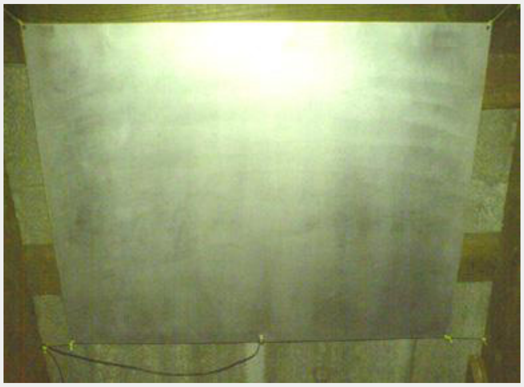

The plate being suspended using nylon cord to prevent it touching the roof or anything else:

The plate is positioned between 3 and 3.5 metres (10 to 12 feet) above the ground and the attachment to the plate is also heavy-duty 8 swg cable:

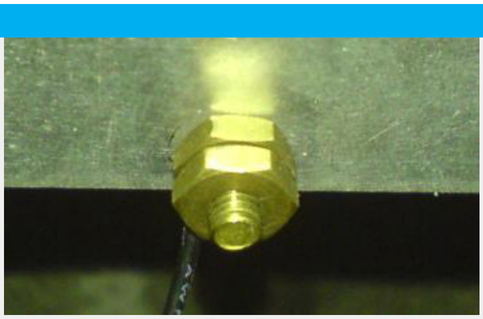

The cable is connected to the aluminium plate using a brass bolt and nuts which the builder thinks may be significant, quite apart from avoiding any galvanitic connection to the circuit. The cable is then run vertically downwards to the circuit. For this arrangement a second earthing point is also used. This is a galvanised iron pipe 3 metres long, driven vertically into the ground which is saturated with water. The second earth is 2 metres away from the first earth.

This arrangement provides serious power, enough to cause injury to, or kill a careless human. With two modules, it will light an LED very brightly, driving it to 2.6 volts. If the LED is removed, then the voltage climbs to about twenty volts and is easily sufficient to charge a 12V battery or battery bank although that takes time. With twenty modules as 12V battery can be charged over night. It is estimated that with two hundred modules, the power would be sufficient to power a household although that has not yet been done. It should be borne in mind that each module is easy and cheap to make, so arranging for a stack of them where additional modules can be added at a later date for more power, is an ideal arrangement. The circuit is like this:

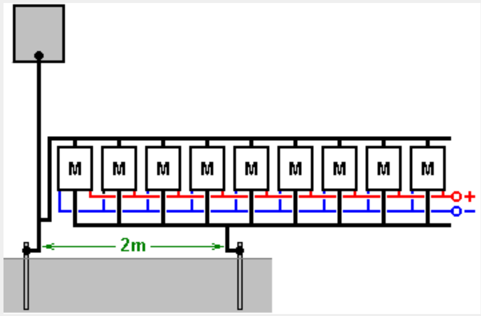

This circuit looks completely mad as the aerial input to the circuit appears to be directly short-circuited by the second earth connection. In spite of this, the circuit works very well when connected this way. Additional modules can be added without any known limit. Increased power can be had by either raising the aluminium plate higher above the ground, to say, 10 metres (33 feet), or by adding one or more additional aerial plates. As you have a good aerial connected through to a very good earth, there has to be the possibility of the equipment being hit by lightning, and so it is suggested that a protective spark-gap is installed between the aerial and the earth, close to the circuit, so that if high-voltage is suddenly applied to the aerial, the spark gap will fire and shunt the excess power through to the earth. Alternatively, possibly a better solution is to install a standard lightning rod system a few metres away from the aerial and a metre or two higher up, so that it forms a more attractive point for a lightning strike.

Further experimentation has shown that altering the connection point for the aerial has a significant effect on the results. If the connection is made at the mid point between the aerial plate and the earth connection, it produces a greater output:

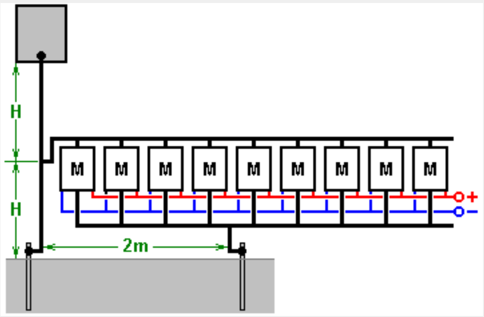

With this arrangement a single module produces around 30 volts while the original method of connecting near the earth was giving about 26 volts with two modules.

Dragan Kljajic has been experimenting with this circuit and has started by building many of these modules on a printed circuit board like this:

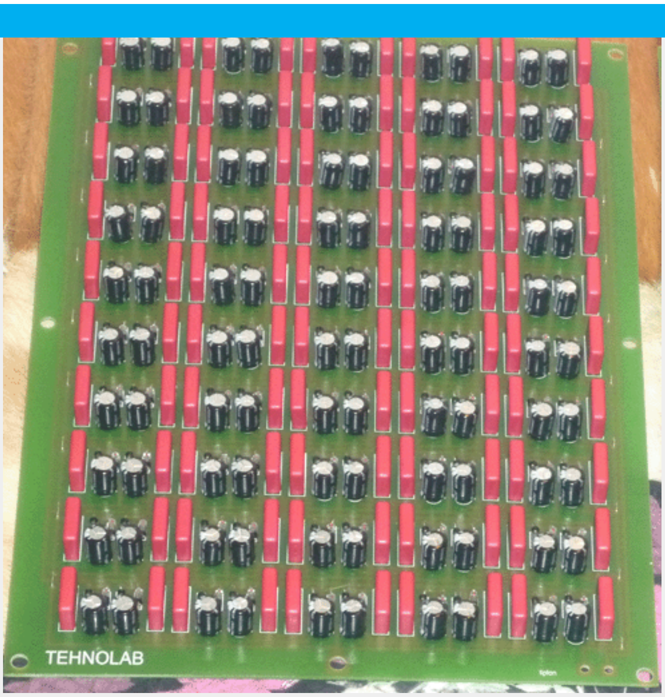

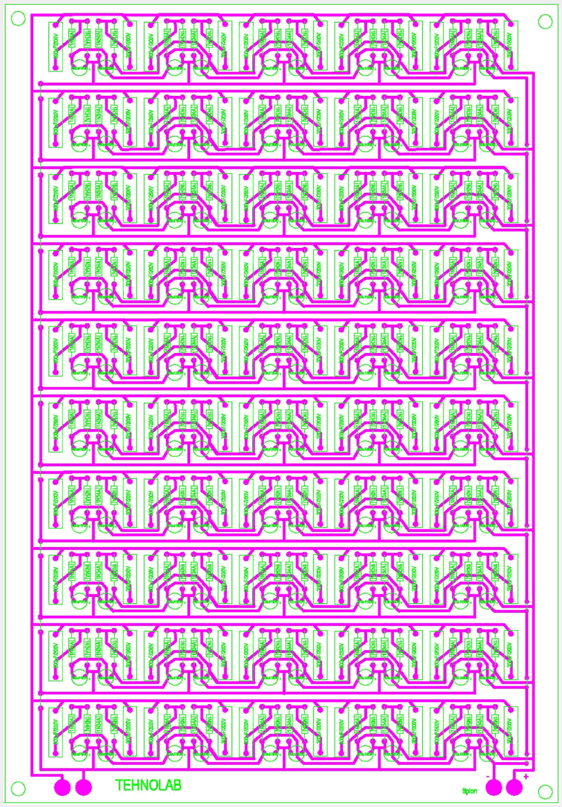

Using two of these boards, Dragan is pulling 96 watts continuously from his aerial plate. He intends to extend this arrangement much further, but is being hindered at present by a local civil war.

There is a forum where some builders of this system share comments. One comment is that there is an increased risk of a lightning strike where you have an earthed aerial, and so it is advisable not to place the aerial plate inside a house, but perhaps suspended between two trees. Also, using a car spark plug connected across the module set can protect against lightning strikes damaging the circuitry.

← Truth is a commodity… Sad →

INFOWARS.COM VIDEOS:

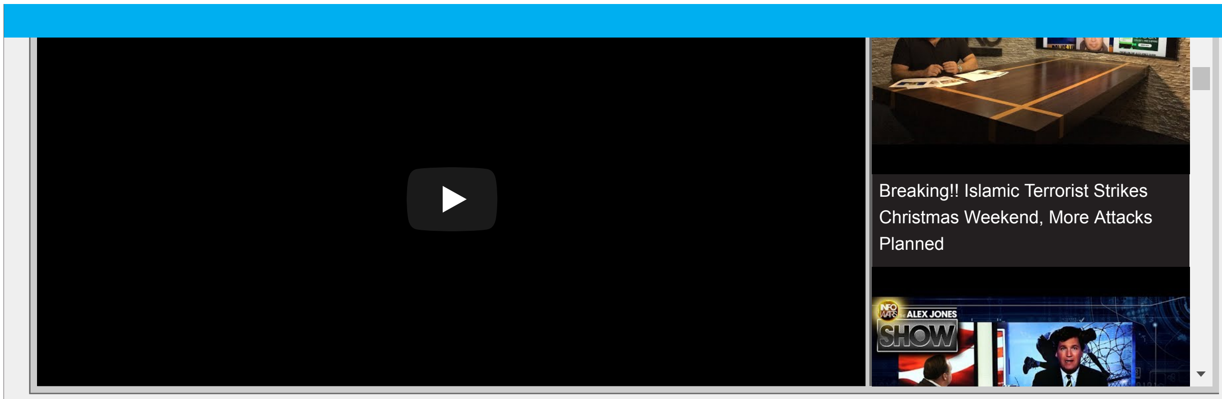
Breaking!! Islamic Terrorist Strikes Christmas Weekend, More Attacks Planned
ALEX JONES
SHOW

COMMENT ON THIS ARTICLE:

Comment rules

9 responses to *Don’t let the big energy corporations control you – Free energy is HERE*

Hendra said on April 26, 2016

Hello, I am new to this website and I am trying to build this things right now with the pcb. Has anyone successfully built this things with the PCB that Dragan make ?

THank you

Hendra

Log in to Reply

ifree said on June 1, 2012

cool. I remember reading an article about a farmer who was able to light a bunch of bulbs from a really long cable that somehow just inducted energy… this was after that idiotic show “Mythbusters” said it was impossible.. so, yeah. I don’t recommend that show.

Log in to Reply

Shaun said on June 1, 2012

I have made two of these boards but in A4 size with 32 modules on each board im in the process of drilling holes.
The alexkor circuit in Chap 7 is also worth looking at I will let you know when Im done with the Tesla /Jes Ascanius
circuit and what results I get I know its important to polish the plate and to have a very good earthing plate.

Log in to Reply

United said on May 28, 2012

Great post! thanks

Planet Infowars Log In Visit

Kakashi said on May 28, 2012

Is this transduction?

Log in to Reply

Kellie said on May 28, 2012

The whole article is here http://www.free-energy-info.com/Chapt7.html

Log in to Reply

WizarDave said on May 26, 2012

Gotta love Tesla!

Kellie said on July 6, 2012

Yes, and look how our great government treated him later in life up unto his death, and still to this day…Corrupticians !

Log in to Reply

ZerMaster said on May 26, 2012

Don’t forget about Tesla’s Turbine at 1PSI Steam hooked up to a generator (1PSI! think about it …) cheap to build and use. http://books.google.ro/books?id=BOMDAAAAMBAJ&pg=PA188&dq=popular+mechanics+Tesla&hl=en&sa=X&ei=BeC_T4G1FdDn-gbS-rCcCg&sqi=2&ved=0CCwQ6AEwAA#v=onepage&q=popular%20mechanics%20Tesla&f=false

http://www.youtube.com/watch?v=Oqyh-JyCmQc

Log in to Reply

Leave a reply

You must be logged in to post a comment.

BE ACTIVE

LOGIN

GROUPS

MEMBERS

F.A.Q.

TUTORIALS

YOUR ACCOUNT

LOGIN

MORE

CONTACT

WRITING TIPS

TERMS OF SERVICE

DIGITAL MILLENNIUM COPYRIGHT ACT

©2012 Planet.Infowars.com is a Planet Infowars, LLC company. All rights reserved - Digital Millennium Copyright Act Notice

rexresearch.com

Paul BAUMANN

Testatika Generator

---

life.augmented
INNOVATIVE SEMICONDUCTOR SOLUTIONS
Digi-Key ELECTRONICS
Learn More

Methernitha Group: CH-3517 Linden, Switzerland

Photo Gallery & Videos

Links:

Methernitha ~ http://www.methernitha.com
http://utenti.tripod.it/testatikmachine/
Paul E. Potter ~ http://website.lineone.net/~aarekhu/index.html

---

Hans Holzerr: Testatika Demonstration Notes

Paul Potter: Back-Engineered Testatika

P. Potter: Principle Experiment

Anonymous: Believe It or Not --- Here It is!

Cyril Smith: Testatika Generator and Over-Unity

Methernitha: The Research Work of Methernitha in the Field of the so-called Free Energy"

P. Potter: Transcript of Methernitha Testatika Video

Testatika Demonstration (4 August 1999)

Translation by Stefan Hartmann < harti@harti.com > & Hans Holzherr

From: Hans Holzherr
To: Stefan Hartmann

Recently, over 30 technicians and engineers (most of them retired) were allowed to visit the Methernitha group in Linden, Switzerland, where they witnessed a demo of the different Testatika machines.

Here is a report from Hans Holzherr from Switzerland who was present:

Hello Mr. Hartmann,

To your questions: > Have you seen live a machine with a load? If so, what load?

I am referring in the following to the model with the 50-cm diameter disks. This machine was already running when the visitors stepped into the room, and was not halted during the whole time --- we were there for about 1.5 hrs. As a first load a 1000-Watt lamp was connected for approximately 10 seconds whose brightness did NOT diminish --- the corresponding sequence on the Testatika film is just an effect of the camera aperture's automatic adjusting to the sudden brightness! The second load was a U-shaped heating element, that Mr.Baumann handed to me. It became so hot within one second that I had to put it down immediately! What was particularly impressive was that while he pulled back one of the contact wires (that was with the lamp, I believe), a 1-cm long arc appeared between the output electrode and the connecting wire for approximately one second. The apparatus was under a plexiglass hood. Near the base it had two holes which Baumann used to insert the contact wires to touch the output electrodes.

How do you think the high wattage is produced ?

Good question! I'd love to know the answer, too!

Did the disks slow down when a load was placed across the output electrodes ?

I did not notice that (nobody else did), but of course you tend to turn your look to 'where the action is' (the lamp etc.) The disks turned with 15 rpm, which is quite slow. The spin rate was regulated magnetically.

What general impression did you have ?

It was really impressive! One can hardly believe it, with this slow rotation. In any case, this cannot be explained in terms of bare electrostatics in the sense of the Wimshurst machine. The perforated sheets seem to have a key function... Beside the pick-up and the drive electrodes there are a number of small plexiglass blocks with glued-on perforated sheets, whose function is unknown.

As Adolf Schneider already mentioned, my colleague Bernhard XXX and I want to try to copy the principle experiment shown by Baumann --- without much hope to find anything extraordinary, though.

The device consists of a horizontal swiveling plexiglass arm with a small rectangular plexiglass plate at both ends glued to the lower side of the arm. The lower side of the arm is covered with perforated aluminum sheets (square holes), while the bottom of the plates is covered with brass wire mesh. Beneath each plate five additional plates are glued onto the base plate. There is also wire mesh between each pair of plates in the two blocks. From the mesh layer between the lowest plate and the base a wire goes to the two capacitors, which are connected in parallel . Baumann seized the arm with both hands and turned it about ten times back and forth (a full rotation was not possible, because the capacitors were in the way), then measured the DC voltage with a digital measuring instrument: 60 Volts. Then, as he short-circuited the condensers a loud crack could be heard. I don´t know if that already is an abnormal result...

On my question Baumann replied that with metal foil (instead of wire mesh) the device would not produce that effect.

For more information, Contact: leo@zelator.in-berlin.de

"Back-Engineered Testatika"

by

Paul E. Potter

http://website.lineone.net/~aarekhu/index.html

(Copyright 2000)

That the Swiss Methernitha group's Testatika machine is thought to be based on a Wimshurst electrostatic generator, is only a sparse approximation of the truth --- of the great multitude of electrostatic influence machines developed around the 1900's it more closely follows the charge-separation-and-collection system used by the 1898 Pidgeon machine [note 1] for its electrical circuit.

Its 50-per-disc steel grilles or 'gitter-grilles' are plainly unique to the Methernitha (see fig.1) but in principle follow on from previous research and patents for corrugated sectors which were found to be more efficient charge carriers [note 2] than flat ones, and from a similar example in more recent times of aluminium rods extending out like wheel spokes from an insulating hub of perspex [note 3].

Another unique function of these perforated grilles attached to the discs is how they induce charge from the rotating discs onto the special collecting pads, or 'tasten' antennae keys (which are also perforated --- so as to more readily pick up charge); for in a Wimshurst you had conductive brushes or rails of sharp points which actually touched the discs or were placed very close to them, but in the Methernitha the charge has to be made to traverse a parallel air-gap to the pads and for this purpose the metal gitter-grilles are so designed to create miniature eddy-currents of charged air which circulate in and out of the perforated metal's surface charges, and are more easily bounced out to the collecting pads. This process is categorised as VARIABLE CAPACITANCE electrostatic generation.

Careful note needs to be made of how the Methernitha uses its basically Pidgeon setup with regard to its neutralising rods (that equalise and stabilise the opposite charges --- see fig.2), and how charges are picked up from one area and accumulated at others, so that the polarities of charge are distributed correctly to specific areas on both discs [note 4].

And although there have been some fanciful claims, or misinformation, that it uses all sorts of radioactive materials to achieve its pulsed output I most strongly believe that the auxiliary electromagnetic circuit, that wraps itself around the rotating discs, portrays a simple electronic approach; afterall, who would use radium radioactive emission alongside leyden jar capacitors ! Indeed, the more you look into certain elements of its construction the more they point to three main eras of electronics development, the 1900's, the 1920's and the 1950/60's. The authentic Methernitha was designed and developed by purists who believed they had discovered a previously unknown electronic phenomenon, but they wanted to keep an integrity to the early pioneering days of the Pidgeon, Wimshurst and Holtz electrostatic machinery; they would not use such modern devices as transistors or IC chips (more's the pity) --- but they do use some pretty uncommon electronic engineering in their circuit [note 5].

Obviously, the electronics are in two parts; one --- the electrostatic generator and its particular technologies of how to direct what charge where, and two --- the very unique auxiliary electromagnetic circuit of inductances, capacitances and rectification that mobilises that 'static' electricity. To understand how they convert static energy into an electromotive force you would do well to go back to the earliest years of radio. From the pages of spark radio you soon appreciate just how important oscillation circuits and their valve rectifiers were, and moreover, how difficult it proved to engineer them. For although radio transmitters and receivers from the 1900's used resonating circuits their oscillations were controlled by sparks between two contacts and, of course, they were relatively inefficient. Not until the 1920's did the first electric current oscillations become an observable controlled phenomenon when someone coupled a rectifier valve, a capacitor, and a resistor together [note 6]. The early 1920's also saw the best era of experimentation and invention for novel devices that turned static energy into useable electromagnetic energy; it was in a 1921 patent that we see a German physicist Hermann Plauson describe in great detail his methods to convert static power, not only from rotary influence machines but also from balloons collecting atmospheric electricity up in the sky; and by using thermionic rectifiers, leyden jar capacitors and inductor coils he proposed a free-energy network that was to power the whole of Germany [note 7] ! The thermionic rectifier valve heralded a new era for radio and high voltage physics, and as it was then subjected to such a broad array of experiments and modifications to improve its efficiency so it paved the way for all sorts of new avenues in electronics. Indeed, with such a technical catalogue of similarities with what we see in the available photographs of Testatika it can be assumed without doubt that the horizontal glass tube which sits on top of the Methernitha machines is exactly what a home-made vacuum thermionic rectifying valve would look like; with its internal anode mesh-plate, surrounded by a coiled copper grid, fed by a glowing (heated) cathode wire running horizontally across its centre and capped by two black end-pieces, which are too big and bulbous to be mere end-caps and must surely be black rubber vacuum seals to seal the glass tube and the input/output wires [note 8].

With such a rectifier, some induction coils, and some leyden jar capacitors you have a circuit that oscillates, and that's what has to happen with a Methernitha, the electromagnetic circuit has to oscillate for it to work, and then the oscillations have to be rectified (or even modulated) so that the resulting single-pole pulses can be channeled through the big cans, which are basically high-efficiency transformers, and outputted as reduced voltage higher current DC pulses (see fig.3).

The precise components used to oscillate the primary oscillating circuit are, I believe, not to be seen in any of the available photographs, but there are various hints for their approximate whereabouts on the machine. Firstly, according to electronic design there should be a capacitor and coil configuration in close proximity to the rectifier. Well, from the picture "3KWREAR" can be seen the two long upright tubes which, according to those who have seen them first-hand, comprise a spirally turned aluminium strip (which indicates they are chokes [note 9]) inside a glass tube, inside the same sort of outer shielding that the big cans have (which indicates they are electrostatic shields), inside yet another glass tube, and are terminated at the top with a brass connecting rod which does a right-angled turn and passes into the side of the tower --- but only two-thirds up the height of the tower. These two assemblies must form a connection to the rectifier, because the rectifier is at the top of the tower, so why don't these electrostatically sensitive tubes extend all the way up to it ? Again, from the photographs of the rear and front of the Methernithas there is a wire that comes out of the tower's side wall at about 4 inches above the upright's brass terminals and this wire then passes through a short black tube and on to the rectifier valve. This, of course, would happen on both sides of the tower, enabling a connection to both ends of the rectifier. But why have this 4 inch gap of connections at the top of the tower ? Something is placed inside the top of the tower in this intermediate space which is very necessary to the circuit, and I think it must be the location of the capacitor/inductor configuration to oscillate the circuit. This (fig.4) is how I would see the inside of the top of the towers [note 10].

I've seen some of the patented inventions that rotate discs --- by using magnets (i.e., H.Rosenberg's permanent magnet excited rotational machine, US Patent # 3,411,027), and by utilising inscribed metalised discs (US Patent # 3,239,705 for instance), but there simply isn't enough room for these to be located in the Methernitha disc setup --- also, you don't want to interfere with the ES fields that zip around the revolving discs: From the reports of those who have seen the small machines working it appears their discs were rotated by small DC electric motors after they were hand-started, some re-wound with thinner wire (to presumably increase their torque) and powered directly from the discs' generated electricity --- but I have also seen how two discs can continue to rotate simply by careful placement of curved electrodes [note 11] which would act on the charges on the discs --- like the 3kw Testatica Distatica generators.

After reading through the many early accounts of electrostatic rotary machines, and some of the more recent ones, you can't help but be puzzled by the Methernitha's incredibly low rotational speed of just 60 rpm (and in the 1999 engineers report as low 15 rpm !). Most other early experimenters boasted up to 3000 rpm. J.G.Trump in his work on high voltage generation in space [note 12] spun his rotary machine at 10,000 rpm (to produce 433 Watts at 24 KV no less). One reason for this low speed might be to do with the close proximity of the 50 lamellas (gitter-grilles) on the discs at their inner ends, they are very close together, I think too close. Air, normally an insulator, breaks down and conducts at around 25-35 KV (this figure has been fairly constant from day-one of electrostatic machine experiments right through to the present day --- because air has a breakdown field strength of 3x106 volts/metre) and short-circuits the circuit. I feel that because this design of grilles is prone to short-circuiting at high voltages the Methernitha people have limited their rotational speed so as to ensure a low operating voltage --- of what I'd guesstimate to be only 12 to 24KV.

But, is this a waste of extra potential ? Not necessarily, for I don't think that the main power output comes solely from what the two counter-rotating discs supply.

There is, I believe, a far more important power generator --- the electron cascade generator, and the Methernitha has two of them, held inside the two horseshoe magnets, and providing the circuits to the magnets are made to oscillate at the right frequency at a high enough voltage then these metalised-perspex laminated blocks can enmass A MUCH LARGER AMOUNT OF ELECTRICITY THAN WHAT IS PUT INTO THEM.

This, perhaps, is the previously unknown electronic phenomenon that the Methernitha group have so zealously been trying to protect against unscrupulous entrepreneurs. But I would say that this copious supply of free energy is already known to the world --- it is not readily available - and its principles are not fully understood, as yet, but it is known about.

As the descriptions say (on the Testatika website), between the horseshoe magnet legs are four blocks of transparent 'plexiglass' type material alternated with copper and aluminium plates (that may or may not be perforated), in the sequence c-p-a-c-p-a-c-p-a-c-p-a (also see fig.6). And according to the Linden Experiment, where Paul Baumann induces a resonance of about 80-140 MHz in a coiled horseshoe and then has an aluminium-insulator-copper block moved between the horseshoe legs, a voltage could be taken off the plates of the block which measured 700 volts (DC presumably) [note 13]. This incredible phenomenon has never been replicated by any 'outside researcher', and is said to be the basis by which the Methernitha machine could be understood how to work [the clue, possibly, to this Principle Experiment may be variable-capacitance and dielectric-absorpsion].

But what, I hear you say, is an electron cascade... Well, it was only by chance, very recently, that I happened to listen to an audio tape by a Dr. Flanagan about crystal water; when I switched the tape over after the end of side one Dr. Flanagan then began talking about an electronic configuration that applied a high frequency, high voltage alternating field across an insulator --- that created what he called an electron cascade effect --- Yes, I thought, here is the answer to the Methernitha Machine.

The electron cascade or avalanche effect is where air molecules are accelerated to the device at such a high velocity that they collide with other molecules and atoms in the air to liberate new electrons which in turn also collide and liberate even more 'free electrons' from other air molecules (see fig.5), all of which become accelerated by the electric field, and an avalanche of electron-multiplications progresses throughout the whole immediate environment [note 14]. It's a chain reaction, and an entirely safe one, it happens in a more ferocious way in lightning strikes, and is a natural phenomenon. And, as in this case, the environment actually becomes part of the circuit [note 15] because the process is actually negatively-ionising the air surrounding the Methernitha machines, and that is why those who have been near these generators when working say the air around about them is cool and fresh [note 16].

In view of the fact that it's designers have chosen to wind insulated wire (which may or may not be bifilar) around the horseshoe metal it is likely that the horseshoes are used for some form of induction [note 17], it would also be very possible to draw directly from this part of the circuit the extra electric current produced from the electron cascade blocks, with suitable connections that might lead downward into the wooden base (where it is believed that an alternate layering of perforated metal plates and insulating plates - making up a large high-voltage storage capacitor - is located). This power could then be discharged as a pulsed output of high wattage, especially if the final output part of the electronic circuit is configured as a Pulse Forming Network of multiple sections of inductor / capacitor combinations [note 18].

The two big cans at the side, the big capacitors, are probably not highly technical (see fig.7), once the fundamental formula has been decided upon all models of a Testatika generator would follow a similar construction process. The written descriptions are a little contradictory but they seem to suggest a central input rod, or tube, connecting at the bottom of the cans to a stack of inter-linked pancake coils, that are wound secondary-outside primary-inside, fitted around a core of 6 hollow donut-ring magnets stacked in such a way with plastic spacers as to allow air gaps between them, and then finally the output of each can is a connection from the top coil of the secondaries of the pancake coils to a brass ring around the centre of the black plastic top lid --- and from the photographs can be seen a large diameter wire or tube [note 19] connecting that polarity's output terminal to the top lid's brass ring via a brass screw terminal. I would suggest that the ring magnets (of anistropic ferrite perhaps) are gapped in this way to prevent the magnetic flux fields of the pancake primaries co-joining as one sprawling field, because it would be more advantageous, and safer, to have each separate pancake's magnetic flux cut it's own adjoining secondary coil, and divide the secondary output voltage into smaller amounts of potential, thus depending less on complicated insulating procedures that accompany high voltage single primary / single secondary transformers.

The use of aluminium mesh and solid copper sheeting is commonly used in electronic construction; the outer aluminium mesh cylinder would be used to shield stray electrostatic charges, and the solid copper cylinder is to shield the large amount of stray electromagnetic fields produced by the transforming process from high voltage/low current to lower voltage/higher current [note 20]. Obviously they don't want field contamination taking place between the sensitive electrostatic generator and the transformers.

Within these two outside shielding-cylinders are 'grid condensers' which, according to the 1999 report by the 30 engineers, can be as many as 20 layers of perforated sheet (presumably as concentric cylinders) - which I have indicated (in fig.7 for instance) as being electrically connected

BETWEEN each separate secondary winding - in the fashion of an old discovery from the early days of wireless telegraphy and based on the 'disruptive discharge coil' devised by Nikola Tesla, that such a condenser connected in the center of a secondary coil collects the maximum amount of voltage created by that secondary. This configuration of one condenser inside another inside another etc etc, has a striking similarity to the layout of a pulse forming network [see note 18].

In the red wired can the transformer is wired to output negative, and the blue wired can's transformer is wired to output positive polarity. Special note should be made of a similar arrangement for divided primary / secondary windings devised by Van de Graaff in his 'High Voltage Electromagnetic Charged-Particle Accelerator Apparatus Having an Insulating Magnetic Core' [note 21] with respect to magnetic reluctance gaps.

Whilst it has been said that the clear perspex disc was designated the 'cloud' disc, and the (rear) dark disc the 'ground' disc I would think this relates to different types of acrylics or plastics that might become charged to different polarities, as in the triboelectric series, where frictional charging of different plastics --- and then bringing them close together, might cause donation or acceptance from one to the other; I would think from the above that cloud represents a donator (positive charge) and that ground must mean an acceptor (negative charge). Has anyone tried the combination of a teflon disc (extremely negative charge) with a glass disc (highly positive charge) ?

Or discs doped with paramagnetic particles perhaps [note 22] ?

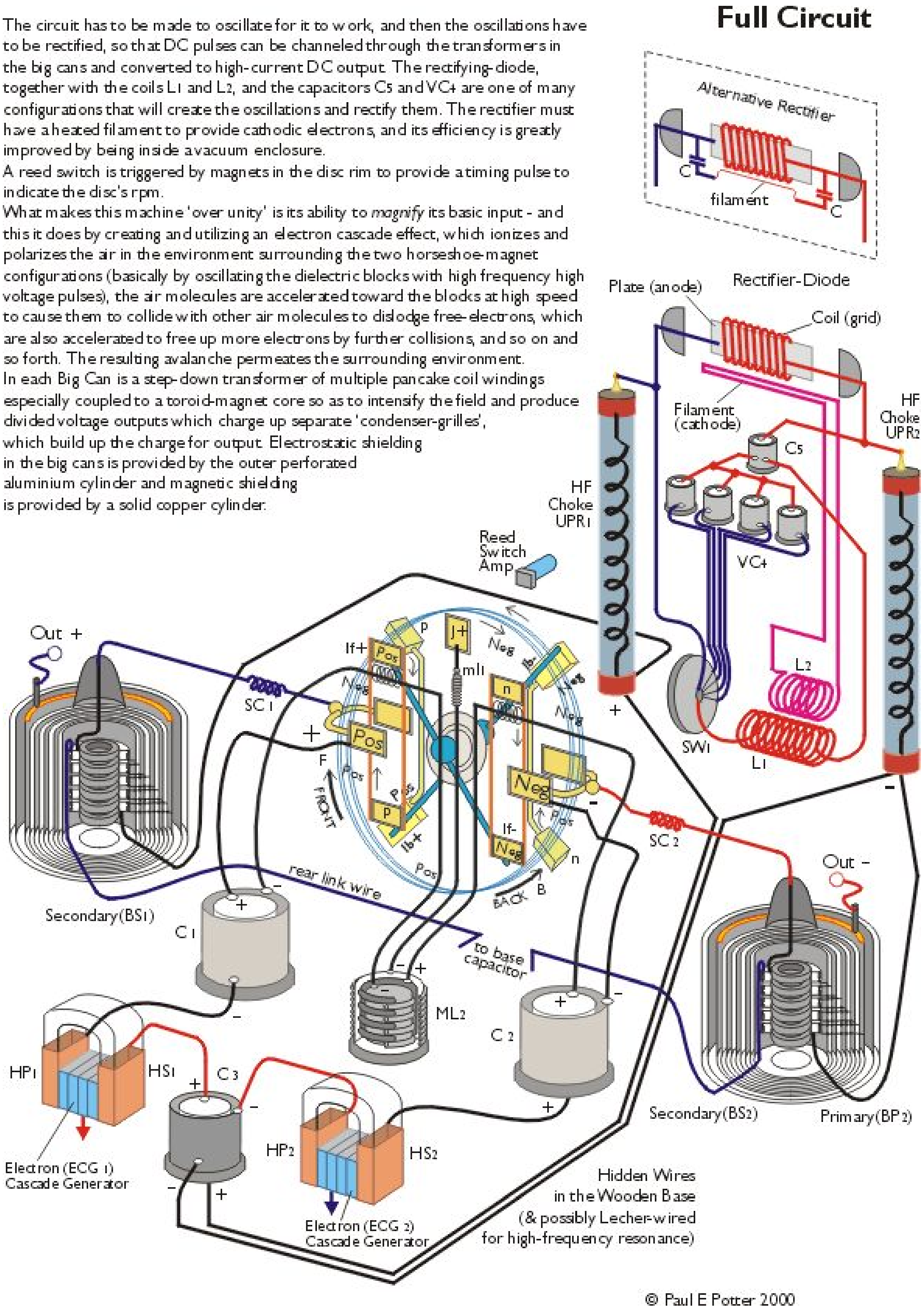

The circuit has to be made to oscillate for it to work, and then the oscillations have to be rectified, so that DC pulses can be channeled through the transformers in the big cans and converted to high-current DC output. The rectifying-diode, together with the coils L1 and L2, and the capacitors C5 and VC4 are one of many configurations that will create the oscillations and rectify them. The rectifier must have a heated filament to provide cathodic electrons, and its efficiency is greatly improved by being inside a vacuum enclosure.

A reed switch is triggered by magnets in the disc rim to provide a timing pulse to indicate the disc's rpm.

What makes this machine 'over unity' is its ability to *magnify* its basic input - and this it does by creating and utilizing an electron cascade effect, which ionizes and polarizes the air in the environment surrounding the two horseshoe-magnet configurations (basically by oscillating the dielectric blocks with high frequency high voltage pulses), the air molecules are accelerated toward the blocks at high speed to cause them to collide with other air molecules to dislodge free-electrons, which are also accelerated to free up more electrons by further collisions, and so on and so forth. The resulting avalanche permeates the surrounding environment.

In each Big Can is a step-down transformer of multiple pancake coil windings especially coupled to a toroid-magnet core so as to intensify the field and produce divided voltage outputs which charge up separate 'condenser-grilles', which build up the charge for output. Electrostatic shielding in the big cans is provided by the outer perforated aluminium cylinder and magnetic shielding is provided by a solid copper cylinder.

Full Circuit
Alternative Rectifier
C
filament
C
Plate (anode)
Rectifier-Diode
Coil (grid)
Filament (cathode)
HF Choke UPR2
C5
VC4
HF Choke UPR1
Reed Switch Amp
Out +
SC1
If+
Pos
P
J+
Neg
ml1
n
Neg
Ib-
Neg
+
Pos
F
Pos
Pos
FRONT
P
Ib+
Neg
If-
Neg
Pos
n
Pos
B
BACK
-
SC2
SW1
L2
L1
Out -
rear link wire
to base capacitor
Secondary(BS1)
C1
+
-
ML2
+
-
C2
+
-
HP1
HS1
+
C3
-
-
+
HP2
HS2
Electron (ECG 1) Cascade Generator
Electron (ECG 2) Cascade Generator
Secondary(BS2)
Primary(BP2)
Hidden Wires in the Wooden Base (& possibly Lecher-wired for high-frequency resonance)

© Paul E Potter 2000

Back-Engineered Methernitha --- Notes

Note 1 --- For more information on the Pidgeon machine see "Electrical Influence Machines" by John Gray, 1903 p. 206 & "Philosophical Magazine" Dec 1898, p. 564, and of course the Pidgeon patents.

Note 2 --- See "Modern High Speed Influence Machines" by V.E. Johnson, 1921, p. 76. Johnson was not only a researcher of electrostatic machines but was also an innovative constructor of them, and as such was keen to try any technique that made his generators more powerful than even the specialised Wommelsdorf multi-disc condenser machines. This book is an absolute must for those who wish to work in this field. Another 'must' is the website of Antonio Carlos M. de Queiroz (http://www.coe.ufrj.br/~acmq/electrostatic.html) which is absolutely full of information about (and with links to) present-day developments in electrostatic machines.

Note 3 --- See "Self-Excited, Alternating, High-Voltage Generation Using A Modified Electrostatic Influence Machine" by M.Zahn, et al., American Journal of Physics, Vol 42 (1974) p. 289.

Note 4 The Methernitha designers have taken a basic Pidgeon electric field system and added a few modifications of their own, partly to lock a certain polarity of charge to a certain area so as to stabilise it, and also to boost certain areas with charge. As in their use, for example, of an extra field plate located at the top-centre in front of the front disc (just under the rectifying valve), note also that this plate, or antennae key, is indirectly coupled to the rest of the circuit, via a coil setup. Much the same occurs with the two plates slightly below it, these plates are connected to a brass terminal which connects to a copper wire that goes down and winds in a coil shape around a hollow plastic tube, and inside the tube will be another wire or small coil that draws off the electric charge. So these three plates are using not direct connection but induction to get their charge.

Note 5 ------- By looking at how each of the photographed machines have been constructed you can see that these are high quality crafted structures. I would think each would start off as sub-assemblies fitted together by pairs or small groups of members, those sub-assemblies of wooden base, big cans, perspex framework, discs with bearings and axles, when completed would be passed on to the electrical engineers of the community who would then fit the wiring connections, vacuum tube rectifier and make sure that not only did they work but that they looked like a work of art.

Note 6 --- The Fleming valve had been around since 1905 and while it progressed to the thermionic valve and audion, by 1922 the ‘Pearson and Anson Effect’ was discovered whereby oscillating currents could be produced with a resistor, capacitor and thermionic valve coupled together.

Note 7 --- See US Patent 1,540,998 (9 June 1925) Conversion of Atmospheric Electric Energy by Hermann Plauson. He also wrote a book of the subject titled "Gewinnung und Verwertung der Atmospharischen Elektrizitat" in 1920 in German (which is currently held in the British Library).

Note 8 --- Whilst some have seen the smaller 300 Watt machine’s discharger/rectifier quite open and not encased in a vacuum tube the vacuum tube models would be much more efficient and would waste less current. Also, the rectifier tube must have a heated filament (which on the 3KW machines can be seen as a glowing line running the whole length of the grid and coil assembly between the two black end caps, and in the films you can see faint flashes coming from behind the rectifier so possibly the filament is wrapped around the other side of the grid/coil assembly as well). Coolridge, back in the 1900’s, discovered that no discharge from the cathode to the anode would occur, even at 100,000 volts, unless the filament was heated (Physics Review, Vol. 2, Dec 1913, p. 418). Aluminium mesh will give off electrons quite readily and can be used as a cold cathode --- but a heated cathode offers the advantage of being able to control the oscillations.

Note 9 --- The two long upright tubes are without doubt choke coil assemblies in precisely the right place to slow down the current where it gets oscillated and rectified. In a choke the higher the flow of current the greater will be its resistance to that current flow. An even better form of choke will have some form of iron core inside it.

Note 10 --- I have come up with 6 different circuits for this oscillation section, some of which include small quartz crystals. (See notes 13 and 16 on frequency of oscillation). The black dial at the rear of the 3kw machine is most likely to select a variety of capacitances so as to control the oscillations of the circuit, which in turn control disc rotational speed.

Note 11 --- The phenomenon of electrostatic motors has been well researched over the years (see "Electrostatic Motors" O. Jefimenko in "Physics Teacher" Vol. 9, March 1971, p. 121-9, and in "Electrostatics And Its Applications" by A.D. Moore (1973) p. 131-147; "Electrostatic Motors" by B.Bollee in "Philips Tech. Review" Vol 30 (1969), p. 178-194). The Methernitha Testatika generators (see a recent report by 30 engineers) auto-rotate, after they have been started by hand, by the same principles of these ES motors.

Note 12 --- J.G.Trump worked for the US Air Force and pioneered some highly efficient electrostatic machines around the 1960's (see"Electrostatic Sources of Electric Power" in "Elec. Eng." 66:525, June 1947; and "High Voltage Generation in Space:The Parametric Electrostatic Machine" in "Progr. Astronaut. Rocketry" (vol 3 --- Energy Conversion for Space Power) 1961 p745).

Note 13 --- Although the 'Linden Experiment' was thought to register a frequency of 80-140 MHz this does not necessarily mean that the Methernitha generators would oscillate at that rate also. Such a frequency seems unnecessarily high.

Note 14 --- See "Plasma --- The Fourth State of Matter" by D.A. Frank-Kamenetskii (1972) pp10, and Dr.Patrick Flanagan's US patents 4,743,275 (May 10, 1988) and 4,391,773 (Jul 5, 1983).

Note 15 --- The effect is very similar to the converging forces in a non-uniform field, the oscillating perspex blocks become one 'electrode' and the surrounding air in the room becomes the opposite 'electrode', and by the processes of electrophoresis and dielectrophoresis the electrically charged particles in the air (the electrons and negative ions) are drawn toward the central electrode, which in this case is the perspex block assembly (see "Nonuniform Electric Fields" by Herbert A. Pohl in "Scientific American" (Dec 1960) p. 107-8). I am much more inclined to believe that the ingenuity of the design of these types of machines comes from physicists and not electronic engineers.

Note 16 --- Dr. Flanagan actually uses the electron field generator in his own special ionizer (see Method of Purifying Air and Negative Field Generator US Patent 4,391,773).

How does an electron cascade generator work, I would think that while you have an alternating electron movement (and Dr. Flanagan reckons this effect occurs with a high voltage field alternating at above 20 KHz) at the metal electrodes, the perspex blocks sandwiched between them would transfer the electricity not through their mass but around it, as surface charge --- actually in the layer of air right next to the insulator's surface. Its the same principle as dielectric absorption - the perspex blocks don't discharge themselves fast enough to keep up with the alternating voltage and so they accumulate more and more charge --- until it forms as a layer of charge on the insulator's surface. This means that at a high enough frequency the surface-air molecules polarize, with the more mobile electrons separating from the slower bulks of those molecules and while the electrons get thrusted back and forth a secondary layer of (slower) positive air ions develops, and so on, and the process of high voltage high frequency polarization triggers the electron avalanche effect.

In the event that the perspex blocks are indeed ELECTRETS (as free-energy researcher Geoff Egel and others suggest) I would think that they would work in a similar fashion to the above process, of dielectric absorption that charges up the blocks before they produce the cascade-effect. Because in the electret the electrons charged into the perspex/plastic, and the positive ions, would still be manipulated by the reversing electromagnetic field in such a way as to orientate (as with dipoles) back and forth, to eventually reach the point (if the whole circuit is tuned properly) where they would attain resonance with the immediate air surrounding them. And if this effect is similar to an inductance then possibly a back-emf will result also, to increase the output voltage. Either way I believe the effect will still be an electron cascade through the environment and the product of this oscillating output (at the blocks) could similarly be drawn off and accumulated in the multi-layered base capacitor network.

I would suggest that a test program to find the best type of blocks would be; One --- try different types of plastic/acrylic/ceramic materials for the blocks.
Two --- try different methods of electrifying the plastics (as with electrets). Three --- try plastics doped with semiconducting particles. Four - try plastics doped with paramagnetic particles. Five - try hollow plastic blocks containing an electrolytic fluid. More information on plastics will be found on the Electret vs Dielectric Absorption Page.

Note 17 --- There are several definitions of Bifilar, one where the wires cancel out their magnetic fields, and one where the wires are wound to ensure a tight low-loss magnetic flux coupling, in this case you need all the magnetic flux you can get so it must be the latter --- See "Transformers For Electronic Circuits" by Nathan R. Grossner (1967), p. 224 etc.

Most commonly used magnetic metal is Mumetal, which is an easily saturable magnetic material, routing magnetic flux through it rather than in the surrounding air, so as to enhance the mutual induction between the two coilings of red wire around the horseshoe legs.

Note 18 --- So that the machine's output voltage doesn't drain away when it is connected up to a large load, what is needed here is a Pulse Forming Network (or artificial delay line). "Such a network is an improvement on simple capacitor storage because of the cascading action from one capacitor to the next along the chain. At the beginning, all capacitors are charged to the same voltage but as soon as the first one starts to loose voltage, the one behind it is then free to discharge into it. This topping-up action, which trickles down the network from capacitor to capacitor, is the mechanism by which the voltage across the output terminals tends to hold onto its original level." (see "High Energy Discharge Systems" A.P.Stephenson, Electronics Today International, March 1992, pp. 24-26).

Note 19 --- When voltage of a high potential and high frequency flows along a wire it does so on the outer surface (called the ‘skin effect’) and so the Methernitha would use thick wiring or even 1/8" tubing to connect its circuit.

Note 20 --- Two references for shielding are: "A Shielded Loop" by S. Goldman in "Electronics" Vol 11 (1938), pp. 20-22; and "Measurements in Radio Engineering" by F.E. Terman (1935), p. 218 & p. 341.

Note 21 --- For information on maximum voltage in the center of a secondary coil see "A Handbook of Wireless Telegraphy" by J. Erskine-Murray (1913) p. 42; and an article called "Dielectric Hysteresis at Radio Frequencies" by E.F.W. Alexanderson in "Proc. I.R.E., Vol. 2 (June 1914) p137-157. For Van de Graaff's transformer see US patents 3,323,069 (May 30, 1967) and 3,187,208 (June 1, 1965). These patents were not just for a Van de Graaff high voltage generator, they were for a special system devised by Van de Graaff long after his generator had been in use to convert static electricity into current electricity. This system may be a little too complicated for the Methernitha but, nevertheless, the principles he used for multiple primary / secondary windings may be of some interest.

Note 22 --- Dr. Flanagan modified his insulator blocks, made of resin, by doping them with paramagnetic granules (such as silicon carbide) to enhance even more the electron cascade effect; which is an idea that the physicist Thomas Townsend Brown first experimented with (by using lead oxide granules) in his US Patent 3,187,206 (June 1, 1965) to good effect. The surrounding air could also be ‘enhanced’ in similar fashion to polarise it's electric charge and improve its side of the performance (for those interested in the ‘physics’ of this see an article by W.A.Douglas Rudge "On Some Sources of Disturbance of the Normal Atmospheric Potential Gradient" in Proc. Royal Soc. A, Vol. 90 (1914) pp. 571, etc).

Note 23 --- Some other generators with similarities to the Testatika machine are the "Electrostatic Energy Field Power Generating System" invented by William W. Hyde (US Patent 4,897,592 of Jan 30, 1990) is a rotor/stator variable capacitance machine capable of producing 300 KV. Other such generators are; "Parametric Electric Machine" invented by Ferdinand Cap (US Patent 4,622,510 of Nov 11, 1986) which has a series resonant (LCR) circuit structured into it so that it oscillates --- and indeed operates AT RESONANCE to ensure a high output; "Electrostatic Generator" invented by Dan B. Le May, et al. (US Patent 3,094,653 of Jun 18 1963) is a very ingenious system of variable capacitance; the "Electrostatic Machine" by Noel Felici (US Patent 2,522,106 of Sep 12, 1950) is a good standard which utilizes a valve rectifier; and the "Electrostatic Generator" by William S. Spencer (US Patent 1,415,779 of May 9, 1922) is an early rotor/stator generator which transferred its electric impulses through a transformer to produce a higher current output.

Fig.2

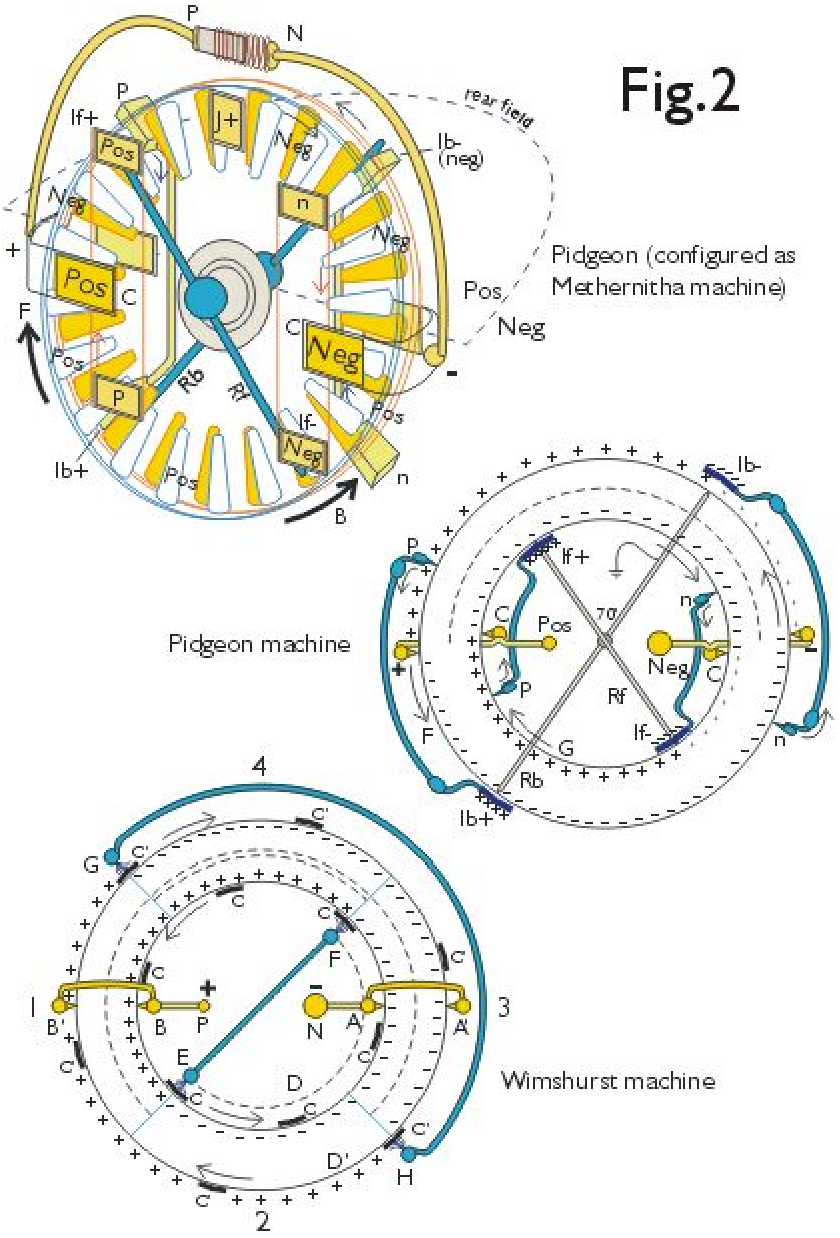
P
N
P
If+
J+
Neg
rear field
Ib-
(neg)
Pos
n
Neg
Neg
+
Pos
C
F
Pos
P
Rb
Rf
C
Neg
Pos
Neg
Pos
If-
Neg
Ib+
Pos
n
B
-
Pidgeon (configured as
Methernitha machine)
Ib-
P
If+
C
Pos
70
n
Neg
C
Pidgeon machine
+
P
F
Rf
If-
G
Rb
n
Ib+
4
C'
G
C'
C
C
F
C
I
B'
B
P
N
A
A
3
C'
E
D
C
C'
D'
H
C
2
Wimshurst machine

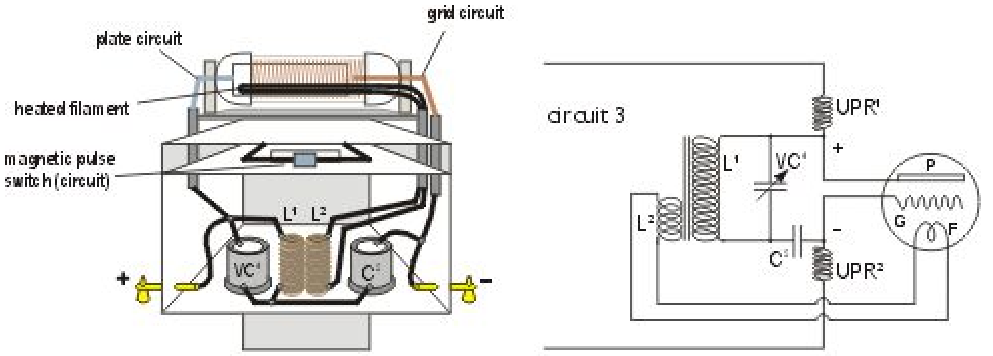
grid circuit
plate circuit
heated filament
magnetic pulse
switch (circuit)
L¹ L²
VC¹
C¹
+
-
circuit 3
UPR¹
L¹
VC¹
+
P
L²
G
F
C¹
-
UPR²

Fig. 4 - Tower Circuits

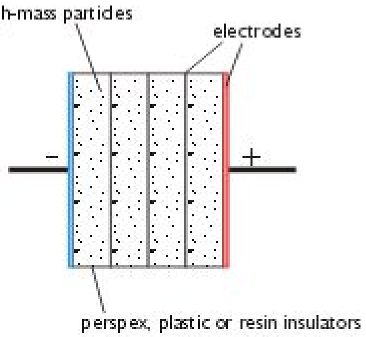
high-mass particles
electrodes
-
+
perspex, plastic or resin insulators

Fig. 6 - ECG Doped

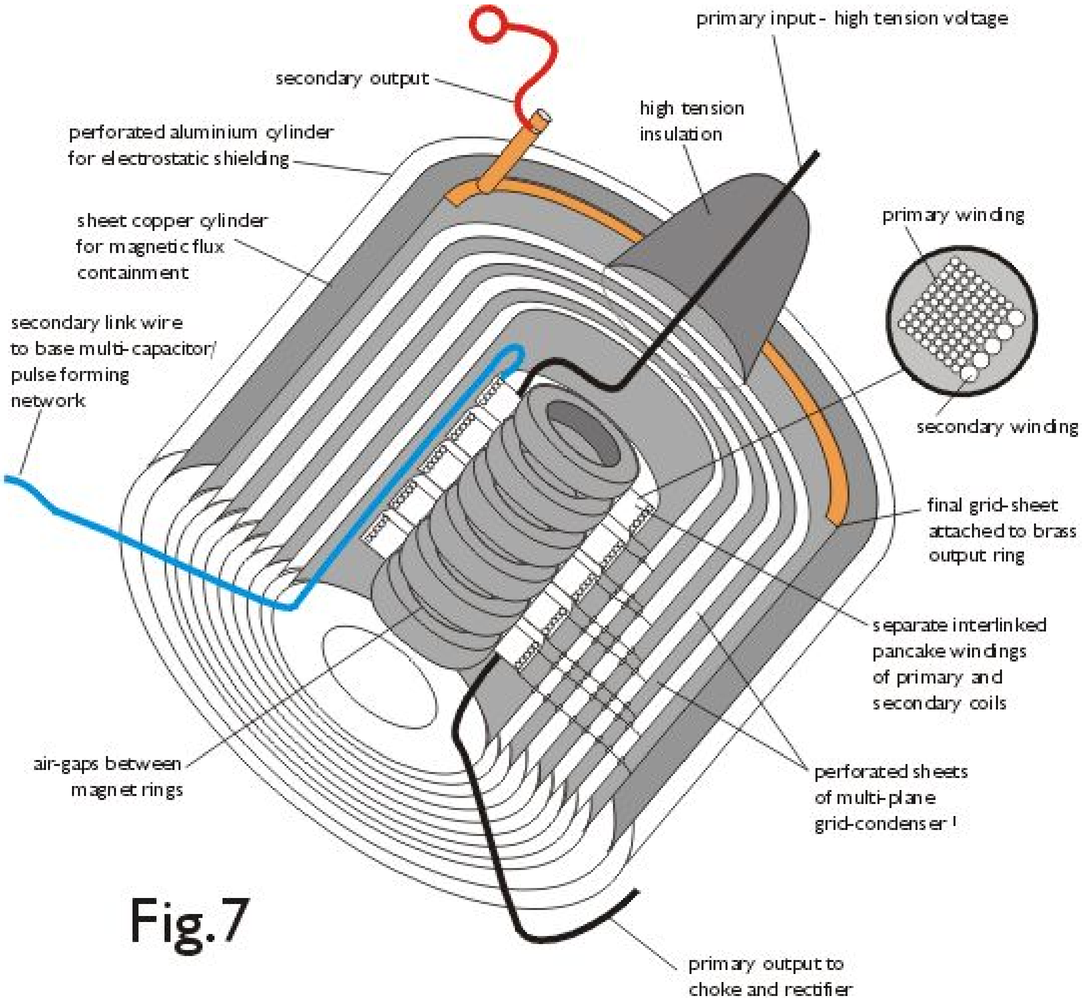
primary input - high tension voltage
secondary output
high tension insulation
perforated aluminium cylinder for electrostatic shielding
sheet copper cylinder for magnetic flux containment
primary winding
secondary link wire to base multi-capacitor/ pulse forming network
secondary winding
final grid-sheet attached to brass output ring
separate interlinked pancake windings of primary and secondary coils
air-gaps between magnet rings
perforated sheets of multi-plane grid-condenser [1]
primary output to choke and rectifier

Fig.7

[1] In the Testatika-Distatica Video it has... "By means of **grid-condensers** the energy is stored and it is then uniformly discharged, at the same time reducing the high-voltage and building up power with additional devices."

"Principle Experiment"

by

Paul E. Potter

With respect to the 30 engineer's report (of 1999) the 'principle experiment' was set up in such a way (see diag 1) that by swivelling the cross-piece over two other plexiglas blocks a current was transferred into two capacitors. Possibly when someone grabs hold of the swivelling cross-piece they transfer some static electricity from their body to the grilles of the cross-piece which, when vigorously turned back and forth will initiate a transfer of electric charge to the two stacks of blocks below them attached to the base. This may either be because anyone walking over a carpet or wearing man-made fibres will already hold thousands of volts of electrostatic charge - which will be enough to 'prime' this apparatus after touching it, to get it to work, much the same as when a rotating influence machine has to be primed with an electric charge before it will work; or, as many researchers have surmised already, that the plexiglas blocks glued to the swivelling cross-piece had been charged up with electricity as an ELECTRET.

The use of mesh around the cross-piece guarantees a variable capacitance to the moving arms --- and as these are moved quickly over the base blocks (again, alternated with mesh and Plexiglas) an amount of electric charge will be transferred to them EACH TIME the arms go backward-and-forward over them. Electronically, you have one variable capacitance acting upon another variable capacitance - which is why Baumann said that if metal foil were used instead of the mesh it would not produce the same effect.

What happens next is called "dielectric absorption", and the following extract, from "Dielectrics" by P.J.Harrop (1972) pp71, will explain;

"When a given dielectric [i.e., plexiglas] has a field put across it, and this field is then removed, the electrodes are briefly shorted out (not enough for all the ions and particles to relapse back) and the device left on open circuit one finds subsequently that it has partially charged up again.

This is due to the slow-moving charge carriers remaining in position. It can be a frightening phenomenon with practical dielectrics since a person who has briefly discharged a charged insulator may subsequently touch it and be severely shocked." Some dielectric materials being more prone to this effect than others, and so obviously some experimentation will be needed to choose the best perspex, plexiglas, plastic, or acrylic materials so as to take the most advantage from this dielectric absorption effect.

Once the blocks are being charged up the electric charge simply flows into the two interconnected capacitors and accumulates.

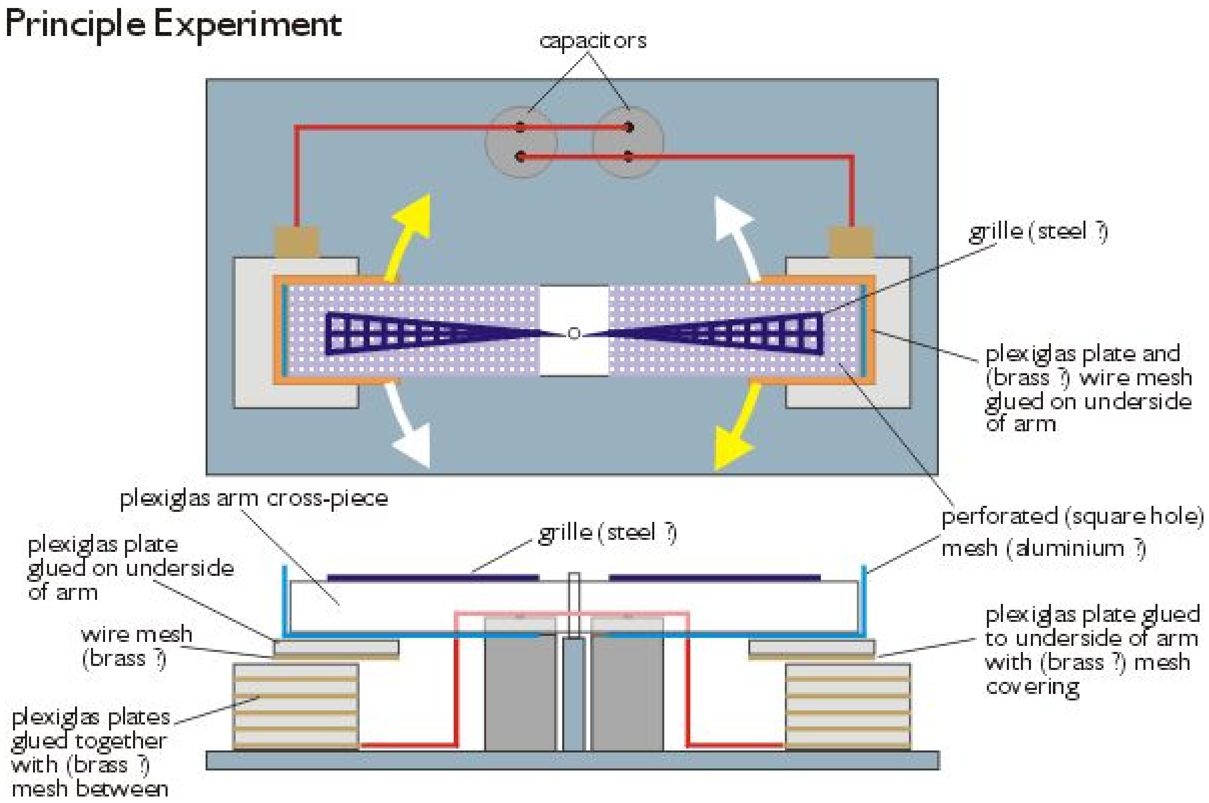
Principle Experiment
capacitors
grille (steel ?)
plexiglas plate and (brass ?) wire mesh glued on underside of arm
plexiglas arm cross-piece
grille (steel ?)
perforated (square hole) mesh (aluminium ?)
plexiglas plate glued on underside of arm
wire mesh (brass ?)
plexiglas plate glued to underside of arm with (brass ?) mesh covering
plexiglas plates glued together with (brass ?) mesh between

Linden Experiment ~

A look at the only available diagram of the Linden experiment (see diag 2) shows a horseshoe magnet coupled to what seems to be a closed-ended wire. It can also be seen in this diagram that the wire coils around the magnet are sufficiently spaced apart so as to provide capacitance (between each coil of the wire) in the circuit. So, with capacitance, coils (for induction), and magnetic flux, you have all the ingredients needed to make an oscillating circuit --- provided there is a supply of voltage and some sort of interruption mechanism - then you have a circuit that will resonate. And this leads us back to that oft repeated question --- is the block used in the Linden Experiment an electret. The first person to PUBLISH how they have duplicated this experiment will be the one to answer that question...

But there is something wrong with this diagram --- the resonant frequency of this simple circuit --- depending only on the values of inductance and capacitance included in the circuit would consequently enable it to oscillate at only a few kilohertz at best. No circuit of capacitance/inductance can be made to resonate at such a high frequency as claimed, not even if it included any form of quartz crystal. If, as the diagram says, it resonates at 140 MHz then there must be a Lecher wire system (the running of two conducting wires in a parallel line, about a foot apart or less, for several feet) to make it resonate at such a high frequency. It would seem to me also that a required modification of the Linden Experiment diagram would be to put a different metal into this circuit where the wire is said to be "closed-ended", say for instance a zinc plate, then there will be a contact voltage established between the copper wire and the zinc plate. Such a voltage would start the circuit resonating, for the reasons stated above.

And as with the Tini setup if a block comprising two metal plates separated by a suitable dielectric material (such as plexiglas) is placed in an oscillating electric field the electric charge on the metal plates will permeate into the dielectric, and for the same reasons as detailed above in the Principle Experiment, will temporarily accumulate voltage because of dielectric absorption. As I have already said elsewhere there is a great similarity between the phenomenon of the Testatika horseshoe blocks (and the Tini setup) and the principle behind Patrick Flanagan's Electron Field Generator which energizes an electron cascade mechanism throughout the immediate environment.

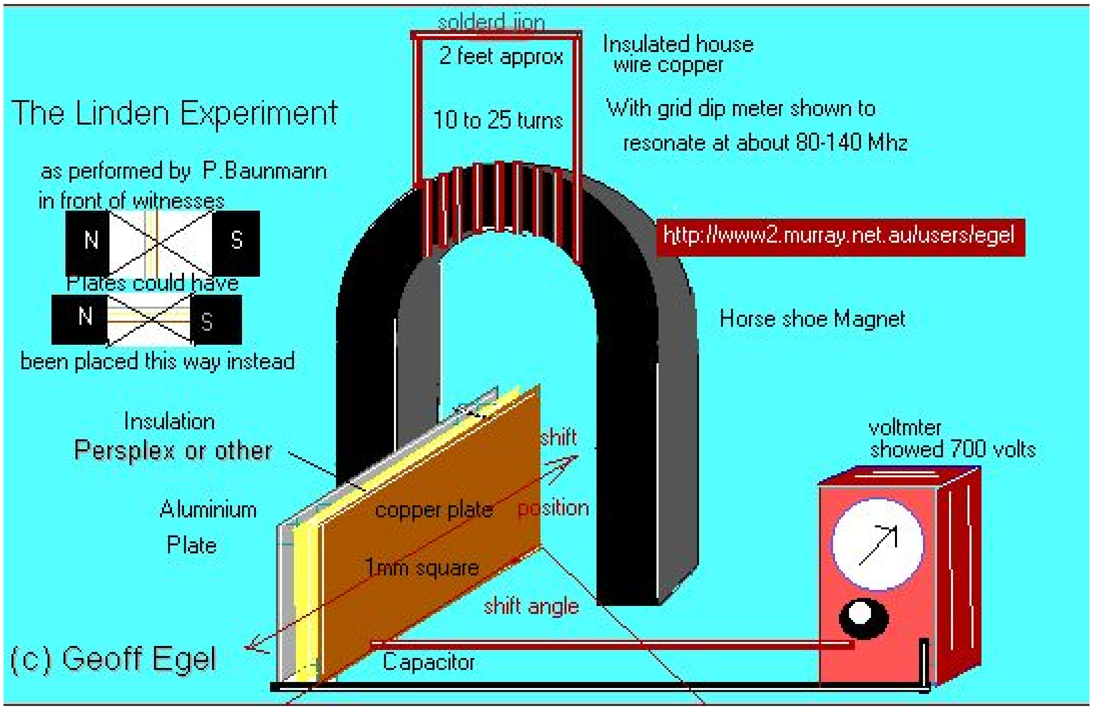
solderd jion
2 feet approx
10 to 25 turns
Insulated house wire copper
With grid dip meter shown to resonate at about 80-140 Mhz
The Linden Experiment
as performed by P.Baunmann in front of witnesses
N
S
Plates could have
N
S
been placed this way instead
http://www2.murray.net.au/users/egel
Horse shoe Magnet
Insulation
Persplex or other
shift
Aluminium Plate
copper plate
position
1mm square
shift angle
voltmter showed 700 volts
(c) Geoff Egel
Capacitor

Geoff Egel (1997) --- http://www2.murray.net.au/users/egel ~

Rectifier Circuit ~

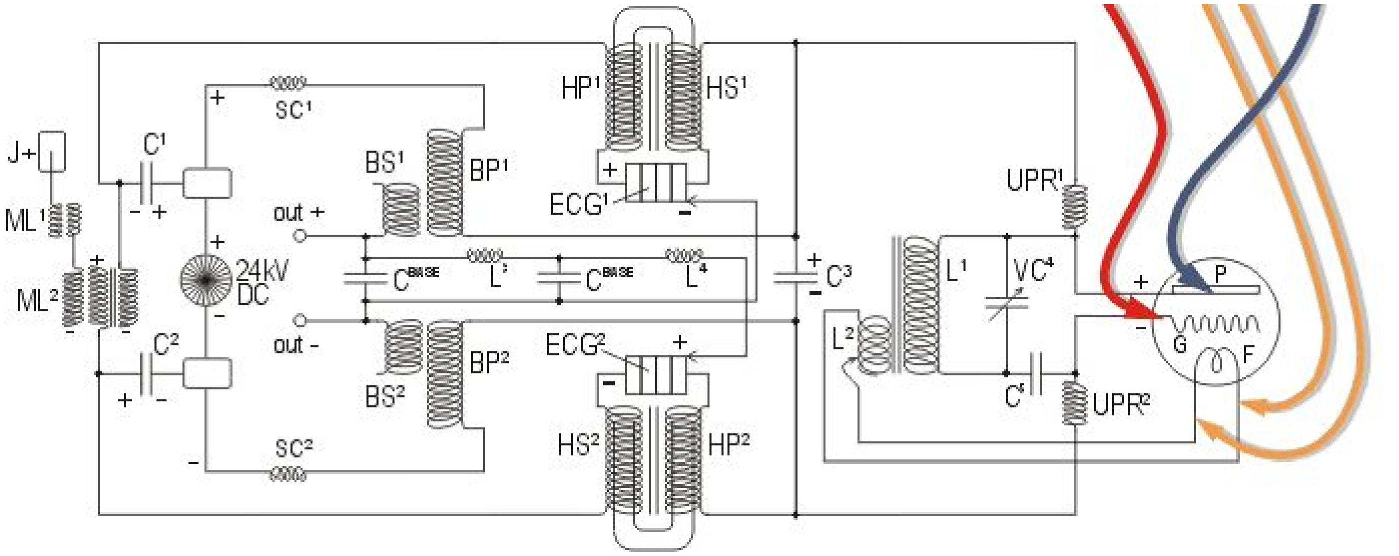
J+
ML¹
ML²
C¹
C²
SC¹
SC²
24kV DC
out +
out -
BS¹
BP¹
BS²
BP²
C^{BASE}
L³
L⁴
HP¹
HS¹
HS²
HP²
ECG¹
ECG²
C³
L¹
L²
VC⁴
UPR¹
UPR²
C⁵
P
G
F

The rectifier forms the basis for Testatika's operation
For whilst the LC circuits L^1, VC^4, UPR^1, UPR^2 and C^5 set up oscillations from the generator's high voltage output the rectifier controls the oscillations to provide positive dc pulses (which are then directed through a step-down transformer BP^1/BS^1 and BP^2/BS^2), to supply power at a higher current rate and lower voltage.

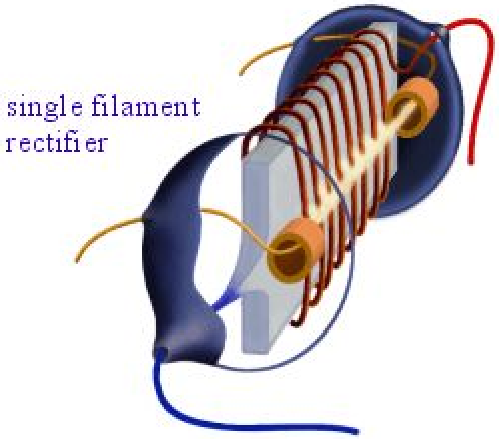
single filament rectifier

Unidentified European Report ~

"Believe or Not, Here It Is!"

The first really running "free energy" machine coming from Switzerland, Europe. It was developed over a 20 years Research period by a religious group called: METHERNITHA. This group lives in CH-3517 Linden, Switzerland. The inventor of this superb machine, Mr.Paul Baumann claims, its running principle was found by studying the lightning effects from nature.

And here are the facts :

* Testatika is an influence-type Wimhurst machine which runs on it's own energy, once started by counterclock revolving it's 2 dics by hand !!!

* Testatika not only runs on it' own energy, but produces also a huge amount of excess power, at least 3 KWATTS of power! That is enough to supply a one-person apartment with one machine!

* The machine you see during this demo-animation is only about 70 cm wide, 40 cm deep and about 60 cm tall !

* It delivers DC-Voltage ranging from 270 til 320 Volts, only depending on the dryness of the air! At this voltage it can supply at least 10 Amperes of DC-current !

* Testatika is not a perpetuum mobile, but an energy machine that collects it's huge amount of "FREE" energy from the charged and ionized air particles. But there are still some technological tricks implemented to overcome the normal drag-resistance of a conventional Wimhurst machine, which is still the secret of the Methernitha group !!!

When will this technology be available for everyone?

The Testatika machines, (there already exists also some smaller units which only deliver about 200-300 Watts) are not yet mass-production type models! They are still laboratory prototype units, although they are build with a very good craftmansship !

For more info contact directly:

METHERNITHA, CH-3517 LINDEN, Switzerland
TEL.-Nr.: ++ 41 31 97 11 24

The pictures are from some photos and from a video-tape Methernitha sells. It shows the machines (also the smaler ones) and explains also Methernitha's spiritual aims. The tape has a running length of about 40 minutes and is really worth looking at it! It is also available in VHS-NTSC and American language. It also shows Methernitha's earlier research developments like huge wind generators.

With the Testatika there is no fraud, hidden batteries or any other hidden energy source. It is really running! They already have the technology of ten years ahead! No more oil, no more pollution, no more atomic waste products, no more hunger in this world! This is the machine the mankind has dreamed about for centuries. Now it has come true!

More information about the Methernitha machine is also available in the German magazine *RAUM & ZEIT*, issue Nr.40 (8).

Jean L. Naudin Labs (4 February 1998) ~

"Testatika Generator and Over-Unity"

by

Cyril Smith

Having read Nelson Camus' article purporting to "explain" the Testatika over-unity generator (http://members.aol.com/overunity2/nelson/testatic.htm), I am of the opinion that the real explanation is hidden amongst the 19th century electrical mumbo jumbo which makes up much of the machine (Leyden Jars, Horseshoe Magnets, Wimshurst Generator and so on). From the description of spiral-wound foil capacitors containing radioactive material and placed within current carrying coils it strikes me that perhaps the real energy comes from Beta particles (fast electrons) absorbed into the electrical circuit.

To extract energy from Beta particles it is not just enough to capture them. Clearly this will result in a current flow according to the capture rate, but current by itself is not power. The kinetic energy of the particle must be captured and this will result in potential difference or voltage. If a Beta particle enters a thin foil conductor normal to its surface and is captured, then the potential difference associated with giving up its kinetic energy can be expected to occur across the opposite faces of the foil. If however the Beta particle were turned so as to enter at a shallow grazing angle to the surface, not only would the probability of capture be increased (since the particle will travel further within the material) but the potential difference will occur along the length of the foil. Beta particles travelling even at relavistic velocities can be turned within short distances by relatively weak magnetic fields, so the concept shown in the following Figure suggests itself.

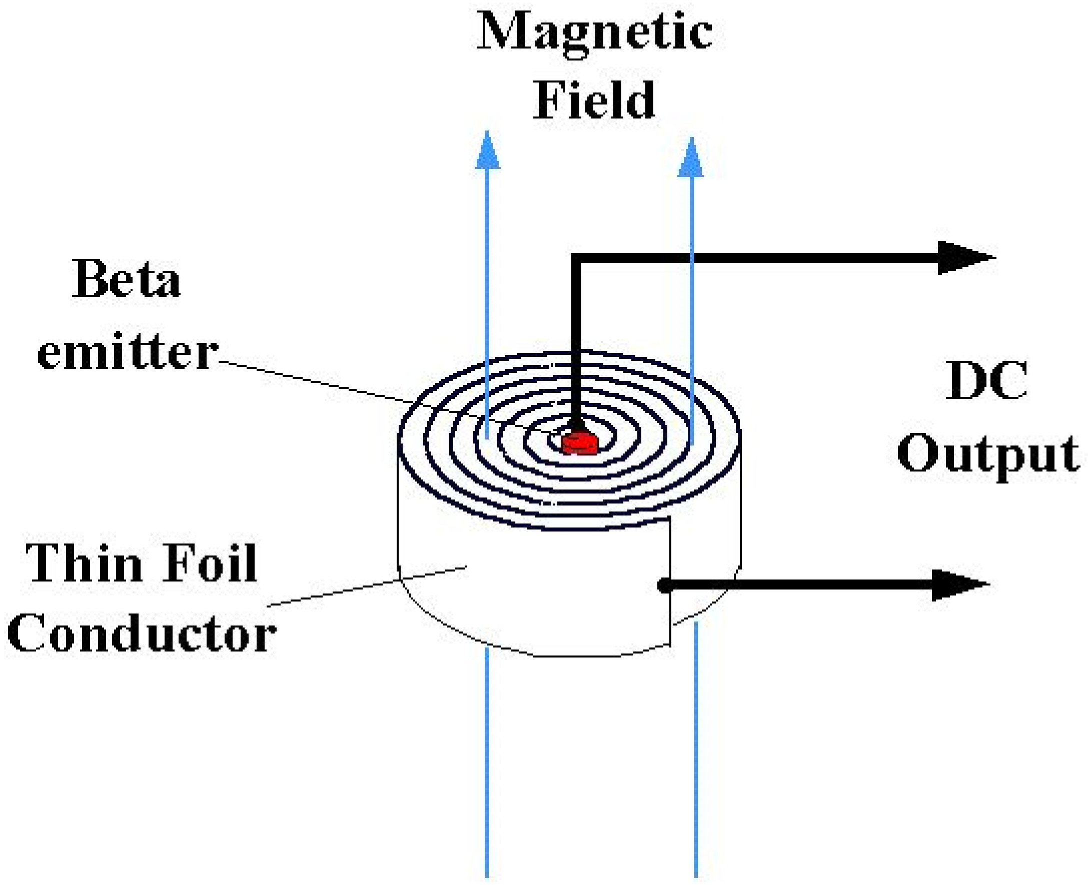

Methernitha Statement of Purpose ~

"The Research Work of Methernitha in the Field of the so-called Free Energy"

Ever since the foundation of Methernitha there existed a department for research, development and electronics, which was concerned with the problem of alternative energy sources, namely with technologies, which were suited to exploit the inherent forces of nature and thus to unlock sources of energy without disturbing natures ecological balance in any negative way. Any technology man invents should serve him in short as well as in long terms, and this condition is not fulfilled as soon as it opposes nature in any way.

This team of research within Methernitha works completely autonomous and is financed out of the co-operatives own resources, without any outside support.

The efficient utilisation of wind energy was one of the first objectives of Methernitha's research program. At the beginning generators with special excitation were developed, which allowed to load the cells of accumulators even at low RPM at times when there are only moderate air movements.

Utilising the kinetic energy of water currents was another field of interest of the development team, but it was pursued more as hobby. The key problem was here to transfer the slow revolutions of the waterwheel to an extent that the excitation threshold of the generator could be surpassed by a minimal loss of energy.

Also solar cells and solar heat collectors attracted the attention of our researchers since a long time. But since in these fields other institutions have attained outstanding results, Methernitha started, and this was already more than 30 years ago, to concentrate its efforts on lesser known and even generally unknown sources of energy. The result of this scientific work is the Thestatica machine.

The question arises: How it comes, that Methernitha, nothing more than a private organisation, could invest so much time, engagement, perseverance and financial resources in this kind of research.

Research and development are integrated parts of the general idealistic concept of Methernitha. To make you understand these ideals and goals, we will now introduce you into the practical sides of this spiritual community Methernitha, as it functions as a co-operative of people, living and working together.

Linden is a calm village of farmers just as dozens of others in the region of the valley called Emmental. However Linden is also the home of a special form of human life, which is capable of attracting more and more people from all over the globe, and which has repeatedly caused astonishment through its excellent functioning, even in extremely difficult situations.

There is no doubt that the formation of Methernitha is also a good part of the life story of Paul Baumann. Through his extraordinary technical capabilities, but also through his astonishing practicable wisdom he opened relations to all kind of people with ideals and in many places, and pretty soon, the idea to form a co-operative enterprise took shape.

"In the name of the God Almighty" people of equal thinking gathered and founded a co-operative, which could serve as the economic basis of this spiritual community.

"We want to be a united group of brothers, and never separate however severe the burden may be", this was the solemn vow.

Renunciation of alcohol and smoking and the will to realise a harmonious community life without dispute and discord like in the original Christian communities were the prerequisites to become a member. For the members Methernitha is an ideal opportunity to lead a fulfilled life by practising charity.

A workshop was erected and soon, one house after the other grew on these premises. Also all construction was done out of own resources and solely with the savings from the common work.

Today Methernitha is a social model proving the quality of its ideals through its well functioning just more than forty years. The people working here are almost without exception members of Methernitha.

Not even the prospering development that took place over the following decades could impair or even suppress the global idealistic aims. On the contrary, everybody works in his own interest with diligence and great joy towards the erection and the support of their new homeland and according to the principle:

One for all and all for one!

With this word one can conquer anything. Also the ancient truths appear again and attain respect.

The people living here feel themselves as members of a family, like a group sitting in the same boat and proud, but also grateful proprietors of their own beautiful homeland, which they may shape exactly as they wish to. Evidently, this form of human social life can function only on the base of idealistic principles.

The question arises, how the realisation of a sincere religious philosophy of life may be brought in harmony with a successful economic management. It is not obvious at all, that this is attainable.

Nowadays there are many amongst us that are caught in the world-wide and dense network of social and economic dependencies and obligations and many also feel the confrontation with its problems.

One member of the economic directory has recently phrased this in the following way:

"The fact that all essential functions of Methernitha are fulfilled without any external force, driven solely through inner conviction, which causes everybody to help and take care of the other, this is for me the most astonishing effect which is produced by this form of living together. It seems to be a miracle."

Another miracle within Methernitha is the Thestatica, which is the result of more than 20 years of research.

This wonder machine is lurked from nature, nothing else. Nature is the greatest source of power as well as knowledge which man has, and it still conceals many secrets, which are only revealed to those, who approach and tie in with them with highest respect and responsibility.

To understand nature and to perceive its voice, man is obliged to experience silence and solitude, and it was there, where the knowledge about this technology was obtained.

For these reasons it was always a great concern of Methernitha to acquire properties, untouched by man as far as possible, be it in valleys or forests, in the mountains or ashore of lakes, where one could study nature, ones own being and the creator of all this universe in silence and concentration and without being disturbed.

The public never understood this properly, rather interpreted it wrongly as an act of seclusion, assuming we had to hide something unclean. We actually had --- and still have --- to take great troubles to realise undisturbed all the things we intended to accomplish.

Such undertaking in research and development necessitate considerable financial expenditures. Therefore we are quite often obliged to construct things with most primitive means and materials. What was thrown away by our affluent society, we collect and possibly set up with it a cornerstone in the discovery of new forces and truth.

We are fortunate in gaining the experience that paradoxically the most beautiful and useful results can be achieved by just using the most simple means. Never did we use any borrowed capital because we want to stay free Swiss citizens and do not want to be hindered or even bound in any way in the pursuance of our aims.

The two conterrotating discs generate an electrostatic charge. One disc represents the earth, the other the cloud. Using grid electrodes the charges are bound. After that they are collected by non-contacting so-called antenna keys and then sorted.

After being initially turned on by hand, the discs rotate by themselves according to the electrostatic laws about attraction and repulsion. A rectifying diode keeps the cycles in steady state. Otherwise the impulses of attraction and repulsion would accumulate and cause the discs to run faster and faster. The correct speed is of great importance and for optimal power generation the discs have to run quite steady and slow.

By means of grid condensers the energy is stored and then uniformly discharged, at the same time reducing the high voltage and building up power with additional devices. Finally the machine supplies a uniform direct current, which varies according to the size of the model. The machine furnishes about 3-4kW permanent output, depending on humidity, whereby the electric potential ranges from 270 to 320Volt. High humidity of the atmosphere prevents the build-up of electric potential. The drier the air, the better.

No doubt, through the so far achieved results one main objective has been reached, namely to prove that it is possible to use Free Energy. Nevertheless the research work is not yet completed.

To the educated physicist many a thing of this machine may seem impossible, maybe even crazy. Maybe he is also offended by the conceptions used to explain the whole. Only partly we could use the concepts of conventional physical terminology to explain and define only approximately the functions and properties of the various parts of the machine.

After all it will be necessary to create some more new concepts like the one we have already used before, when we termed the non-contacting collectors of electric charges as antenna keys.

This machine puts experts, which are just trained in conventional physics to a very hard test, because its mode of action is not explainable with the state of the art of officially accepted physical knowledge, or at the most only partially explainable. However also a trained specialist should remain free and independent in his thinking, and should avoid to be limited by the temporal framework of publicly admitted knowledge in any science.

It has to be remembered that the established science was already many times forced to change or give up some of its very fundamental concepts. Think about Galilee, to name only one example. Our human society almost condemned this man as a sorcerer and magician, just because he investigated and discovered a truth that seemed unacceptable by the established science of the days.

The book knowledge of any times is not wrong, but it is incomplete, and therefore allows to draw wrong conclusions.

We are part of a new area which brings to light many new facts and new knowledge. The clothes of nowadays science have become too tight and should be stripped off, just as the larva of an insect does with its skin. Only this will allow a true metamorphosis to take its course, and finally, at the limits and at the destination of all worldly knowledge, an universal and unlimited spiritual science, radiating and beautiful, just as the completed imago of an insect, may give its blessings and benediction to a renovated humanity.

To ever reach there, a universal development of man has to take place. This however is only possible when man becomes aware of his true role within the whole creation, and again learns to recognise his true tasks, because the whole universe functions within a strict and precisely structured order according to the will and word of the creator. Therefore also man should recognise and realise these universal laws that are valid within the whole and within every part of this creation.

The hard facts rather show how far man has left the divine order through his self willed and authoritarian way of action and that he has become the actual cause of all discord and evil on this planet.

Unfortunately the ruling bodies that should be responsible for the well-being of the people work too often with the target to make life more and more difficult and to render impossible every free spiritual development.

Instead of utilising the achievements of science and technology for the benefit and preservation of all form of life, they are abused carelessly and irresponsibly to destroy and to kill, and thus turn them into a curse upon mankind.

To change all this, the evolution of a new technology is not enough, even if it were the most ecological and ingenious. To change this present status one has to go much deeper down, to the root-cause of all this evil, and this is mans way of thinking, his state of mind.

The ancient divine commandments are still valid today and also show today the way and direction humanity should go, as clearly spoken by the prophet Mica:

God has told you what is good. And what is it that the Lord asks of you? Only to act justly, to love loyalty, to walk wisely before your God.

"Testatika-Distatica Machine"

(Transcribed from the Methernitha Testatika video by Paul E Potter)

"... The efficient utilization of wind energy was one of the first objectives of Methernitha's research program.

At the beginning generators with special excitation were developed which allowed to load the cells of accumulators even at low rpm, at times when there are only moderate movements.

Utilizing the kinetic energy of water currents was another field of interest of the development team, but it was pursued more as hobby. The key problem was here to transfer the slow revolutions of the water-wheel to an extent that the excitation threshold of the generator could be surpassed by a minimal loss of energy. Also solar cells and solar heat collectors attracted the attention of our researchers since a long time. But since in these fields other institutions have obtained outstanding results Methernitha started, and this was already more than 20 years ago [since 1960], to concentrate its efforts on lesser-known and even generally unknown sources of energy, the result of this scientific work is the Testatika machine of which most of you may have heard already.

The two counter-rotating discs generate an electrostatic charge. One disc represents the earth, the other the cloud. Using grid-electrodes* the charges are bound. After that they are collected by non-contacted, so-called, antennae keys** and then sorted [shorted ?].

After being initially turned on by hand the discs rotate by themselves according to the electrostatic laws about attraction and repulsion.

A rectifying-diode keeps the cycles in steady state, otherwise the impulses of attraction and repulsion would accumulate and cause the disc to run faster and faster. The correct speed is of great importance, and for optimum power generation the discs have to run quite steady and slow.

By means of grid-condensers* the energy is stored and it is then uniformly discharged, at the same time reducing the high-voltage and building up power with additional devices.

Finally, the machine supplies a uniform direct current which varies according to size of the model.

The machine furnishes about 3 to 4 Kilo-Watt permanent output, depending on humidity, whereby the electrical potential ranges from 270 to 320 volt. High humidity of the atmosphere prevents the build up of electric potential. The dryer the air the better.

No doubt, through the so-far achieved results, one main objective has been reached, namely, to prove that it is possible to use free energy. Nevertheless, the research work is not yet completed: To finish a model which can be handed out more or less to anybody and without any haphazards, also to non-specialists, much work and also time will still be needed.

Only partly we could use the concepts of conventional physical terminology in order to explain and define only approximately the functions and properties of the various parts of the machine. After all, it will be necessary to create some more new concepts, like the one we have already used before, when we termed the non-contacting collectors of electric charge as antennae keys**. This machine puts experts who are just trained in conventional physics to a very hard test..."

* The German original has gitter-condensers, with gitter defined as wire lattice, grille, grate.

** the original word used is tasten or taster, which means antenna, probe, key.

Your Support Maintains this Service --

BUY

The Rex Research Civilization Kit

... It's Your Best Bet & Investment in Sustainable Humanity on Earth ...
Ensure & Enhance Your Survival & Genome Transmission ...
Everything @ rexresearch.com on a Data DVD !

ORDER PAGE

Energy Projects

Electrodynamic Tether Photos

Deployment of the Tether Satellite

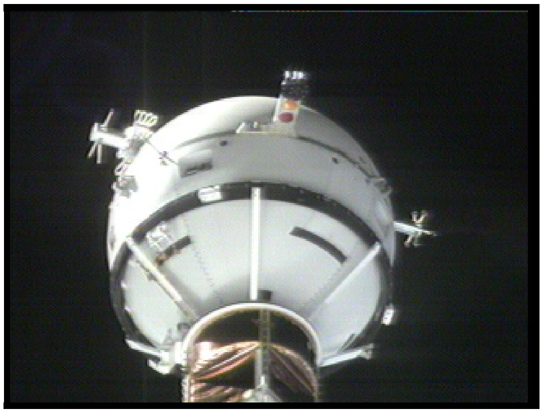

Boom Deploy

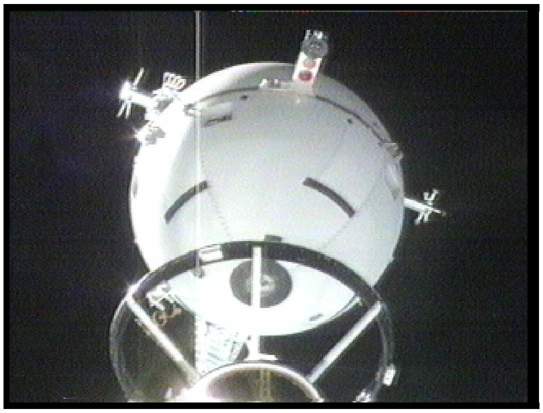

Initial Deploy

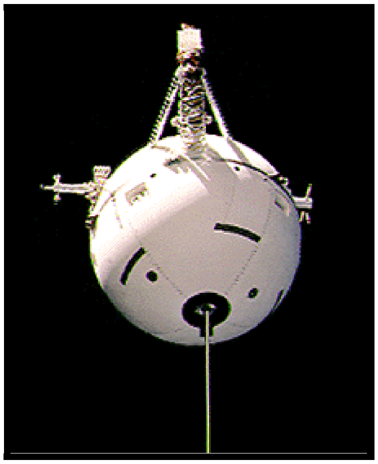

Satellite Deployed

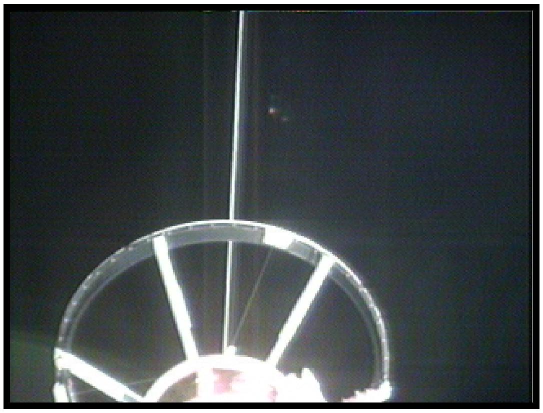

Closeup of boom showing tether

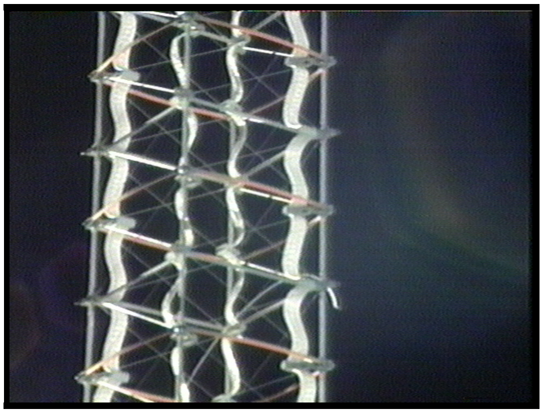

Close up of end of boom

Tether Break

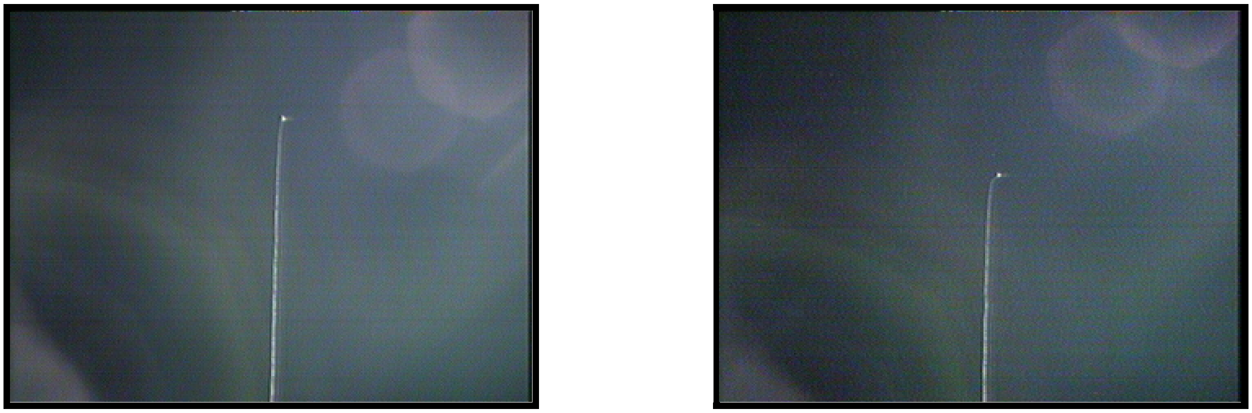

The tether deployed to one mile

Shortly after the breakreak

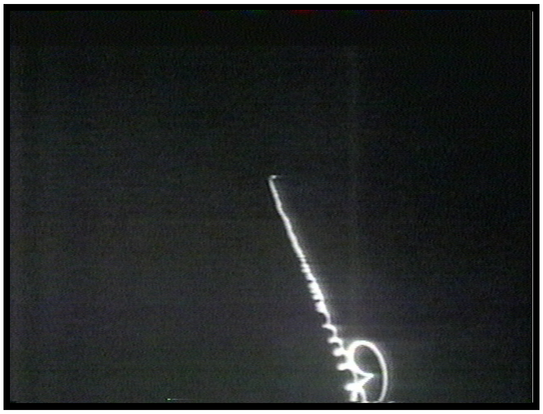

The tether coils as it moves away from the Shuttle

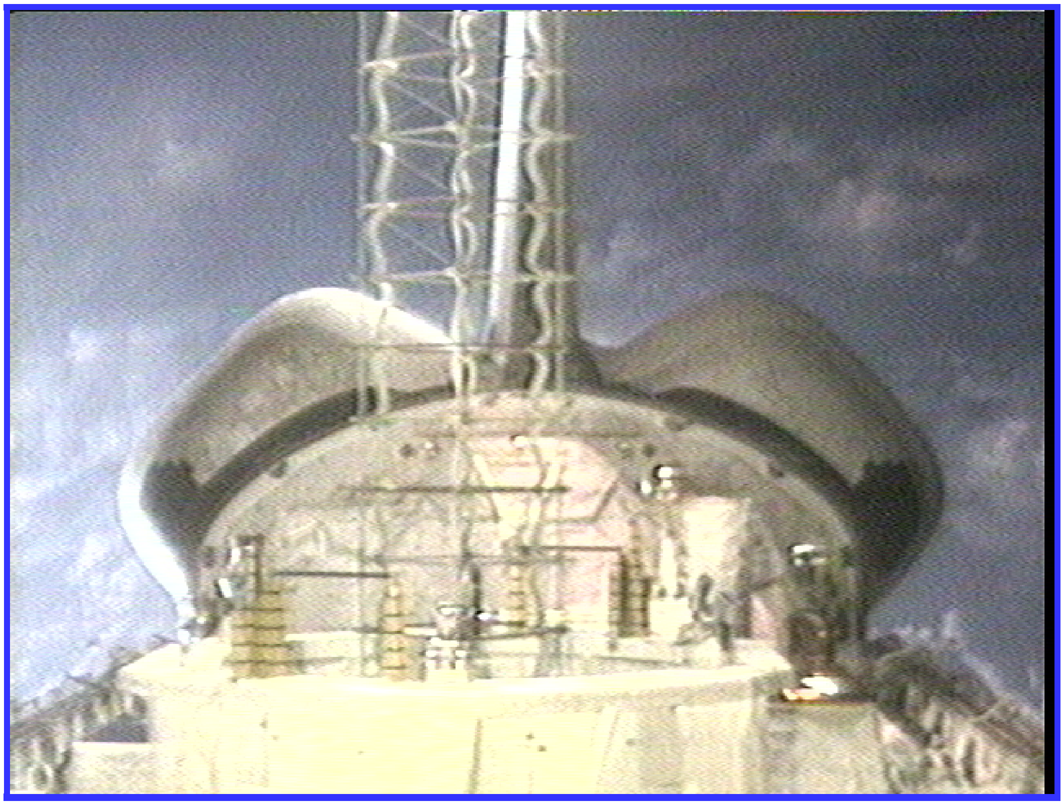

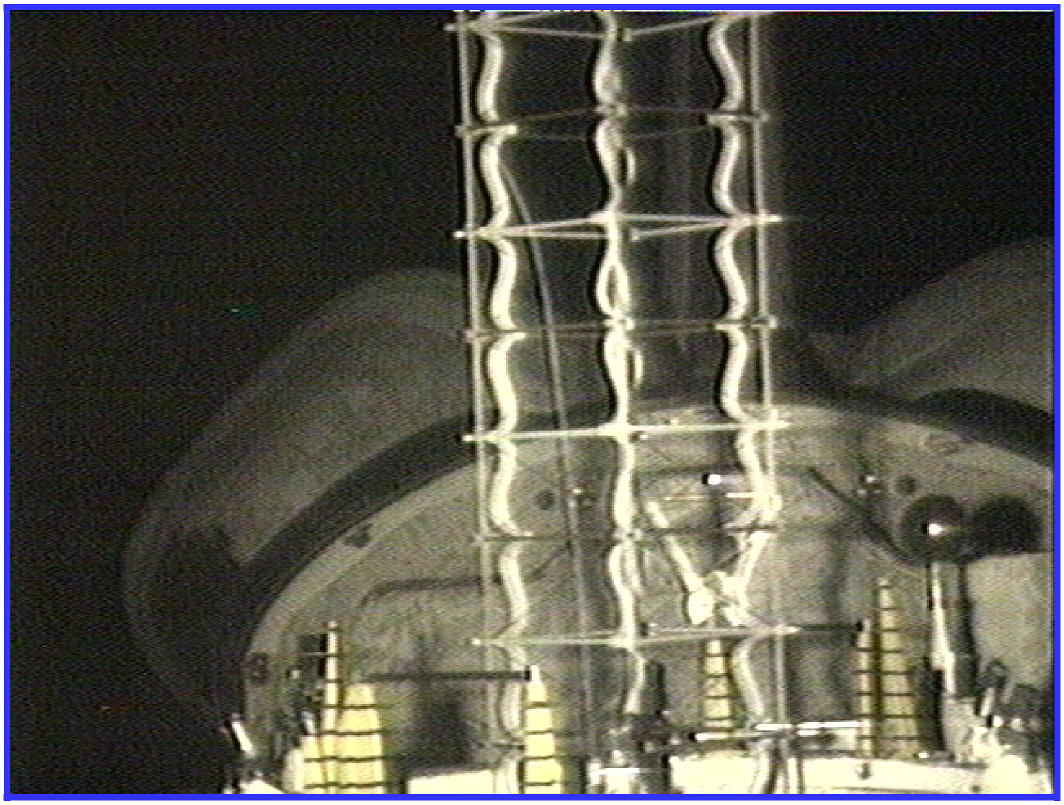

The base of the tether boom. The area looks more like a Tesla Laboratory than the space shuttle hanger bay.

After the Break

The following three images are those displayed for the public showing the white specks that appearall around the tether after it has broken free. NASA calls these "dust and debris" particles, though Mission control in the audio portion of the videos state that they see a "lot of stuff swimming around" and that the tether is wider than expected.

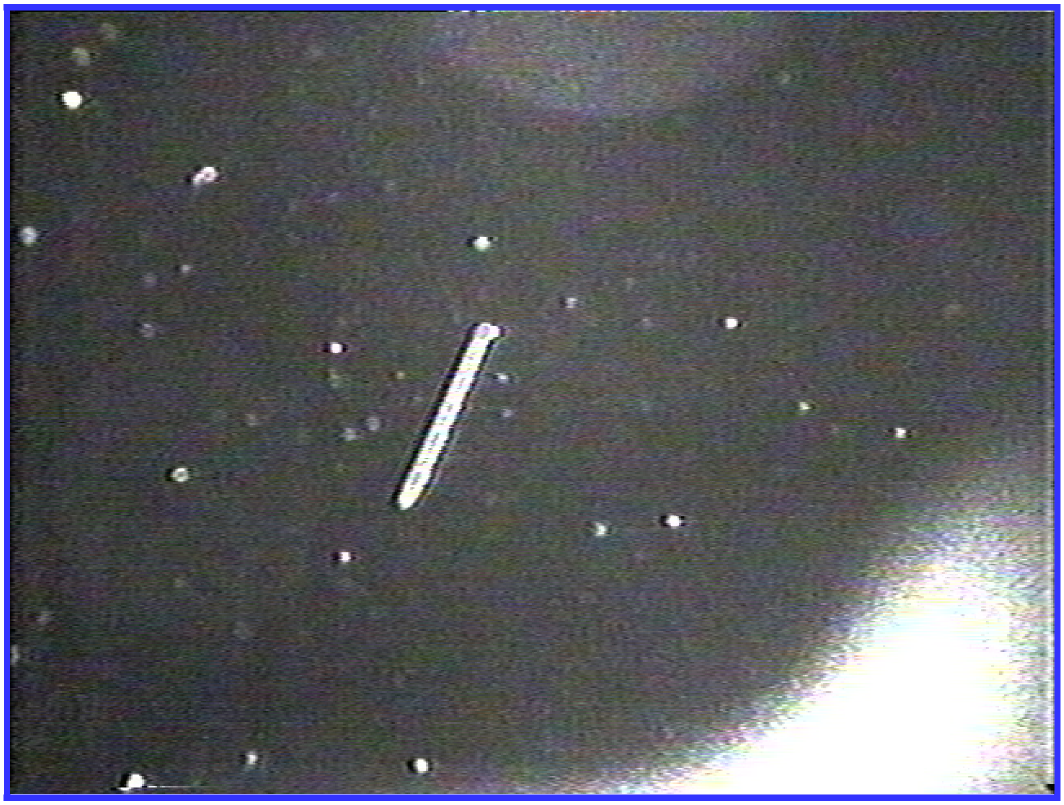

If you compare these images to the ones we captured from a hard to find high resolution copy of the video on this page you will see a remarkale difference. The video is a must see.

Video Clip
STS-75 'The Tether Incident' Feb 1996
Editors Note: This video has been removed several times by Youtube. Currently it is available again. If it does get removed again, email me and I can send you a copy as we have it on disk now. Contact me at webmaster
This one is from the hand held Infrared Camera and is not as clear as the other one was STS-75 Alternate Copy

Comments:

What is most interesting to note is that most of the video clips above are taken from this video, yet NASA does not show the video itself. The copy on Youtube is a transmission intercept copy. However the fact that the NASA info site DOES use those clips gives credibility to the video that it is indeed a NASA film - Zorgon

Tethered Satellite System (TSS-1R)
Source of Clips: NASA STS-75 Mission Pictures

Kennedy Space Center STS-75 Mission Fact Sheet

"Critters" from STS-75

The Report After the Fact

NASA/TP—2003-21228
Low Earth Orbit Spacecraft Charging Design Guidelines

Below are a few keynote excerpts from the PDF file. The entire 367 page document is available below for those with technically savy who want to read the entire paper...

Excerpt 1:
In the case of the TSS-1R tether, its 20 km length produced a maximum of about 3500 V potential between its most positive and negative ends, since it wasn?t oriented perfectly perpendicular to the velocity vector and the Earth?s magnetic field. A satellite at its upper end collected electrons, and an electron gun at the lower end emitted electrons to complete the circuit. When the electron gun was not in operation, a large resistance prevented the Shuttle from being biased thousands of volts negative of its surrounding plasma. However, there remained a large voltage between the tether lower end and the Shuttle orbiter. This enormous bias eventually led to a continuous arc on the tether (see The Continuous Arc, section 4.2.3.1 below), which broke, freeing the satellite and ending the experiment. During the arc, the satellite collected over 1 Amp of electron current to keep the arc going. Probe theory (Cohen et al, 19870010625 N) is usually used to calculate the total current collected by a wire with distributed potentials. However, before the break, TSS-1R demonstrated that a satellite at a high positive potential could collect an anomalously large electron current. See Zhang, et al (20000110580), Stone and Raitt (19990084046 and 20000025437), and Stone, et al (19980202347)

Excerpt 2:
Sustained arcs (continuous arcs) - These are the events that have been attributed with the destruction of on-orbit solar arrays. Generally, the process begins with a fast transient (a so-called trigger arc). Under some conditions, the transient develops into an arc that is fed directly by the entire array, effectively becoming a short-circuit. Such events invariably involve large quantities of energy and can be severely damaging to cells, interconnects or power traces.

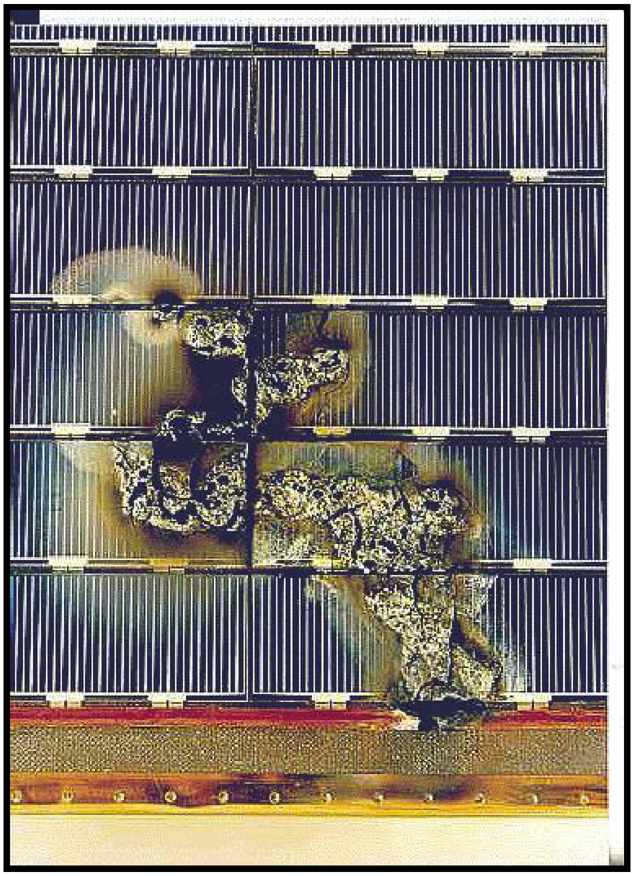

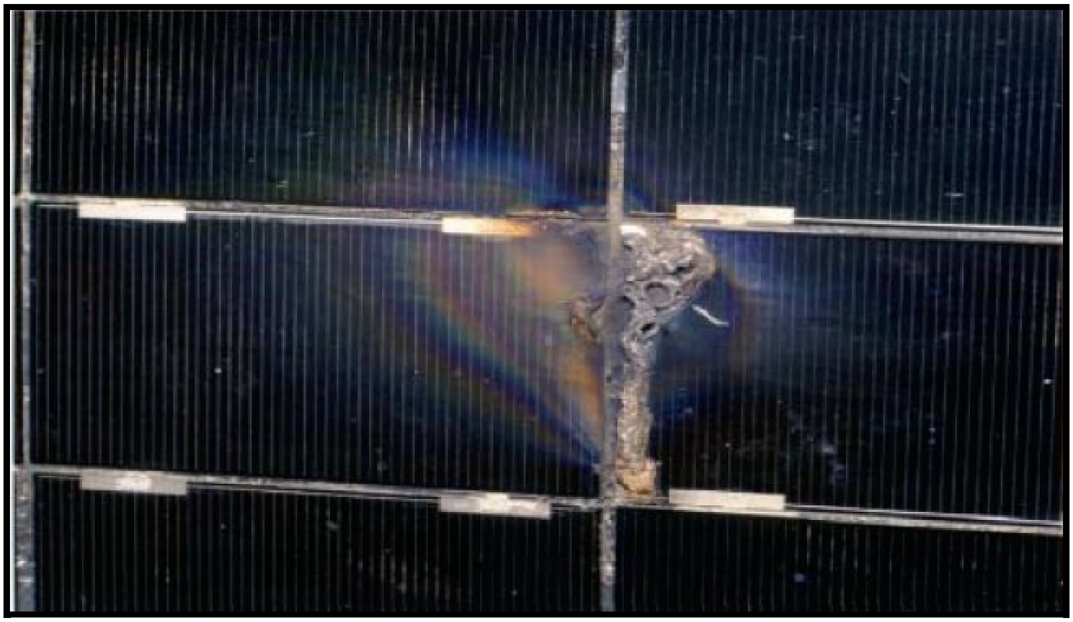

Excerpt 3:

When the structure or array capacitance electrically connected to the arc site is sufficiently large, the initial transient arcs themselves can be large enough to produce significant damage. In Figure 9, we see an anodized aluminum plate that has undergone repeated arcing in the laboratory with the ISS structure capacitance attached. Its thermal properties have been completely destroyed, along with most of the insulating surface layer of aluminum oxide.

Excerpt 4:

The most famous sustained arc event of all led to the breakage of the TSS-1R electrodynamic tether, and the loss of the attached satellite. The image below shows the burned, frayed and broken tether end still attached to the Shuttle after the break. **Incidentally, the tether continued arcing long after it and its satellite were drifting free, until finally it went into night conditions where the electron density was insufficient to sustain the arc.** Noel Sargent (2002) has investigated whether the TSS-1R arc was seen to disrupt Shuttle communications. Although he has found no record of disturbed communications during the event, for most of the time the arc was shielded by metallic structures from the communications antennas, and when the tether broke, the arc was many meters from the receiving antennas. It remains to be seen whether sustained arcs produce radio noise severe enough to be a communications problem.

Comments:

This is official confirmation that the tether continued to produce plasma energy long after it broke free, accounting for the "fluorescent bulb" glowing effect viewed after. We believe this concentrated collection of plasma energy is what attracted the "swarm" of "critters" to a "feeding frenzy"

To get the full PDF file you can download LEO_Charging_Guidelines_v1.3.1.zip

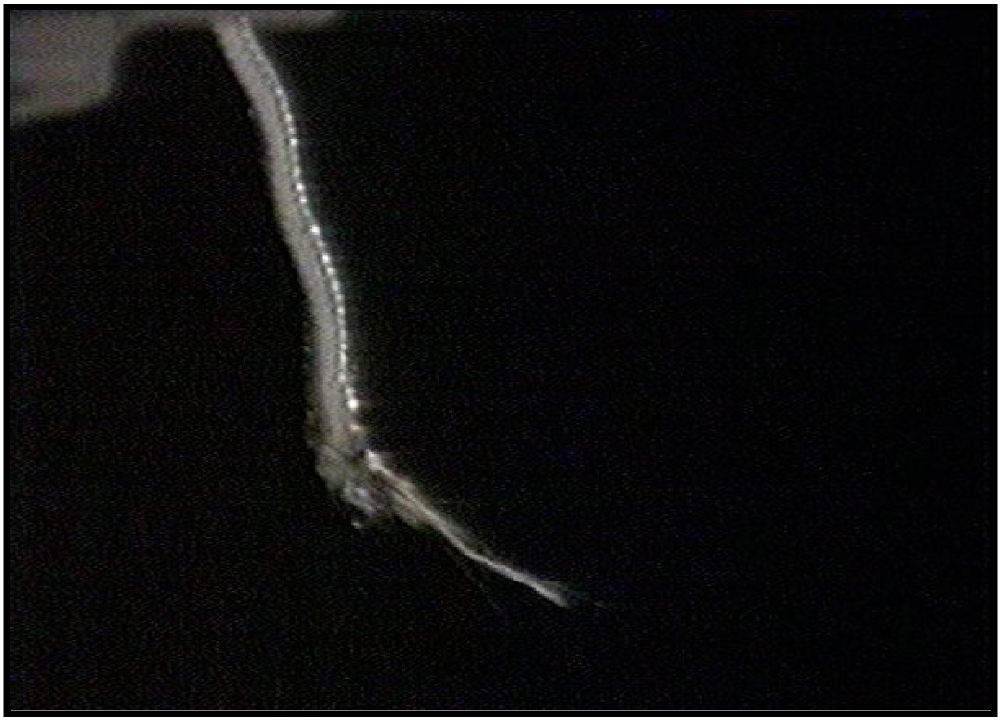

Tether Optical Phenomena Experiment (TOP)

Using a hand-held camera system with image intensifiers and special filters, the TOP investigation will provide visual data that may allow scientists to answer a variety of questions concerning tether dynamics and optical effects generated by TSS-1R. In particular, this experiment will examine the high-voltage plasma sheath surrounding the satellite...

In one mode of operation, the current developed in the Tethered Satellite System is closed by using electron accelerators to return electrons to the plasma surrounding the orbiter. The interaction between these electron beams and the plasma is not well understood...

Associate Investigator: Stephen Mende, Lockheed Martin

SOURCE

The Report After the Fact

...

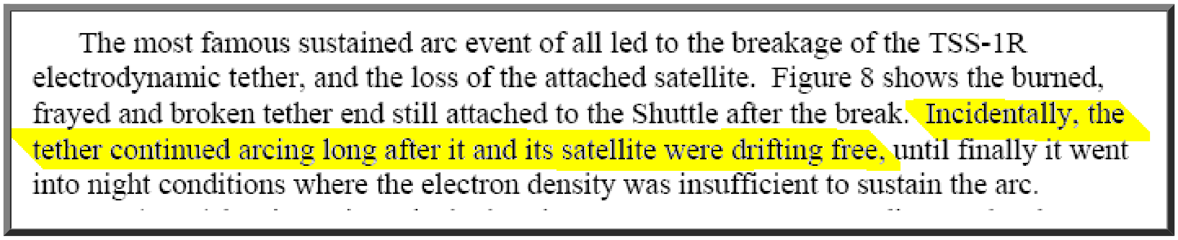

The most famous sustained arc event of all led to the breakage of the TSS-1R electrodynamic tether, and the loss of the attached satellite. Figure 8 shows the burned, frayed and broken tether end still attached to the Shuttle after the break. Incidentally, the tether continued arcing long after it and its satellite were drifting free, until finally it went into night conditions where the electron density was insufficient to sustain the arc.

.....

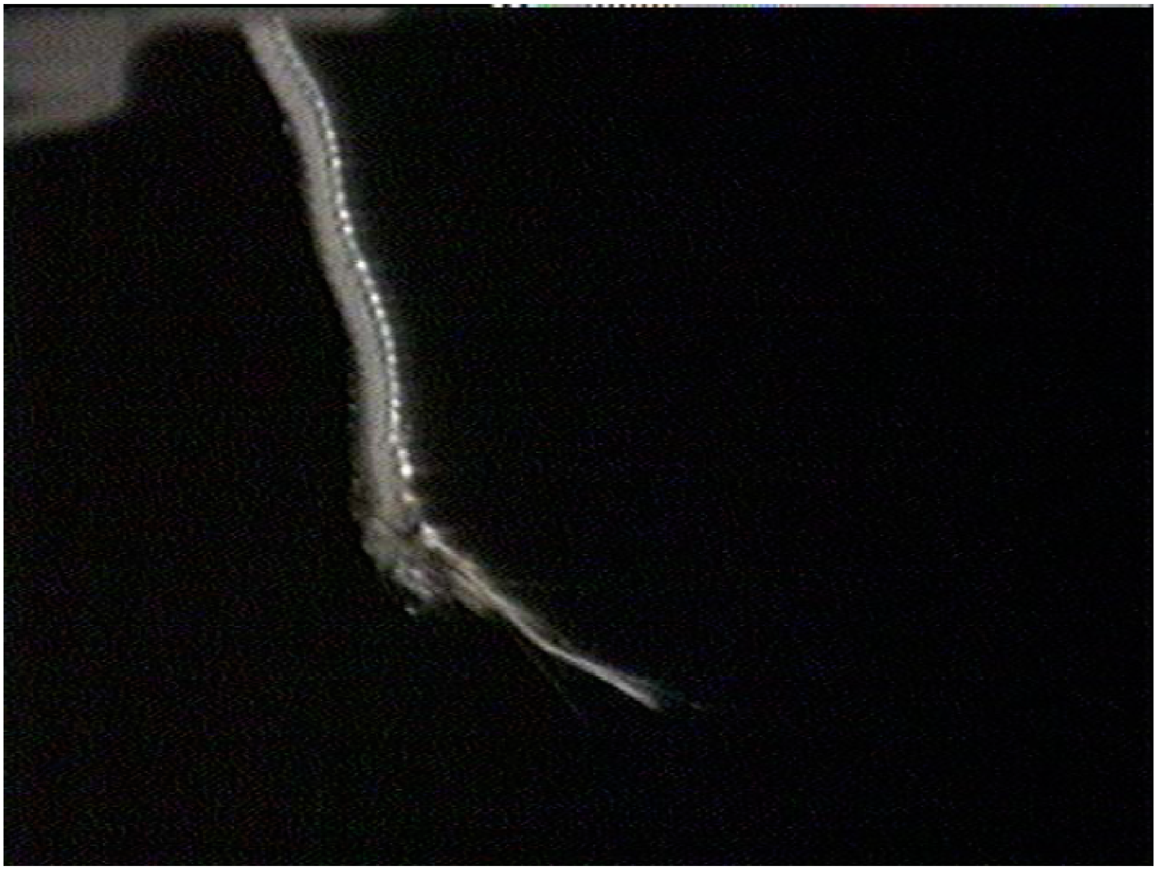

Figure 8 – The end of the remaining TSS-1R tether

…

It is thus possible that an astronaut, grounded to ISS by his tether or conductive tools, could undergo an arc at only –50 V. A sneak circuit analysis showed that such arcs could put 1 Amp of current through an astronaut's heart. Since 0.1 Amp is enough to cause heart stoppage, it is imperative that if the ISS plasma contactors are inoperable during astronaut EVAs, a method be used to prevent ISS astronaut workplaces from floating more than 50 V negative.

…

For many ISS surfaces, peak arc strengths of hundreds of Amps have been calculated. Arcs this strong will melt the arc site and spew molten metal through space. Plasma chamber tests of this kind of arcing are spectacular indeed! Arcs on one anodized surface have been seen to trigger arcs on nearby line-of-sight surfaces.

…

Finally, an arc on an electrodynamic tether may become continuous. The infamous arc on the TSS-1R tether that led to its break and the loss of the satellite was a continuous (sustained) arc with its power supplied by the tether.

…

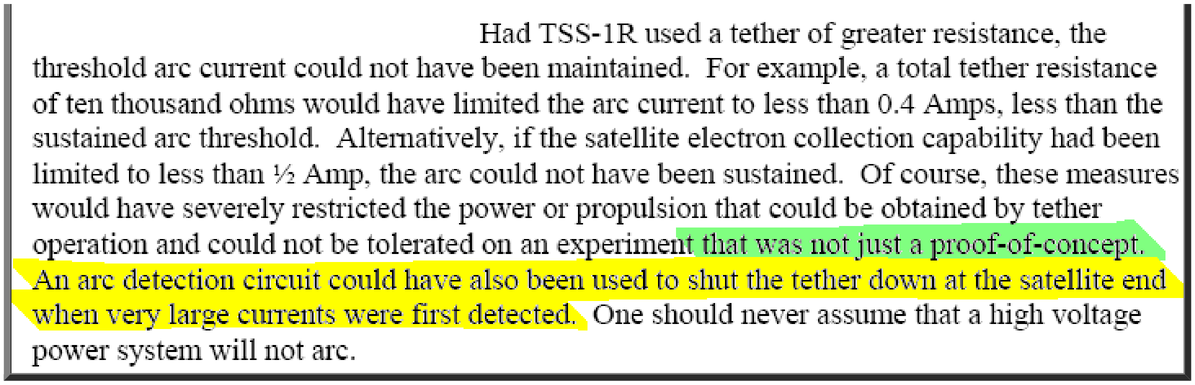

Had TSS-1R used a tether of greater resistance, the threshold arc current could not have been maintained. For example, a total tether resistance of ten thousand ohms would have limited the arc current to less than 0.4 Amps, less than the sustained arc threshold. Alternatively, if the satellite electron collection capability had been limited to less than ½ Amp, the arc could not have been sustained. Of course, these measures would have severely restricted the power or propulsion that could be obtained by tether operation and could not be tolerated on an experiment that was not just a proof-of-concept. An arc detection circuit could have also been used to shut the tether down at the satellite end when very large currents were first detected. One should never assume that a high voltage power system will not arc.

...

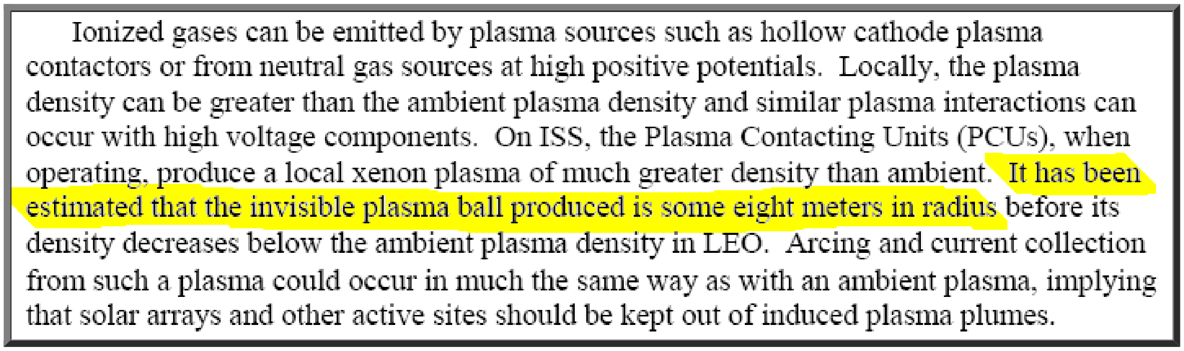

Ionized gases can be emitted by plasma sources such as hollow cathode plasma contactors or from neutral gas sources at high positive potentials. Locally, the plasma density can be greater than the ambient plasma density and similar plasma interactions can occur with high voltage components. On ISS, the Plasma Contacting Units (PCUs), when operating, produce a local xenon plasma of much greater density than ambient. It has been estimated that the invisible plasma ball produced is some eight meters in radius before its density decreases below the ambient plasma density in LEO. Arcing and current collection from such a plasma could occur in much the same way as with an ambient plasma, implying that solar arrays and other active sites should be kept out of induced plasma plumes.

...

Electrodynamic Tethers ~ Menu
Critters ~ Critters 01 ~ Critters 02 ~ Critters 03

Copyright Notice

All pictures and photos on this page, unless otherwise noted, have been gathered from pubic domain sources or are available under GNU License. Further documentation is available on our Copyright Page and our Legal Department.

Article Quotes all have links to their appropriate source and are noted. All other text not so noted is available under the terms of the GNU Free Documentation License

All information on this page is used only for non profit educational purposes. All reasonable attempts have been made that no credits are missed, but with a work of this size omissions may occur. If you see anything of yours that has not been properly credited or wish to have removed please contact the Webmaster

Menu

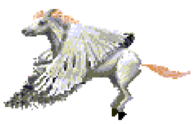

Webpages © 2001-2006
Blue Knight Productions

STS 75 TETHER INCIDENT

•

Tether Deployed

Tether Satellite being deployed from the Shuttle Bay of STS-75

••

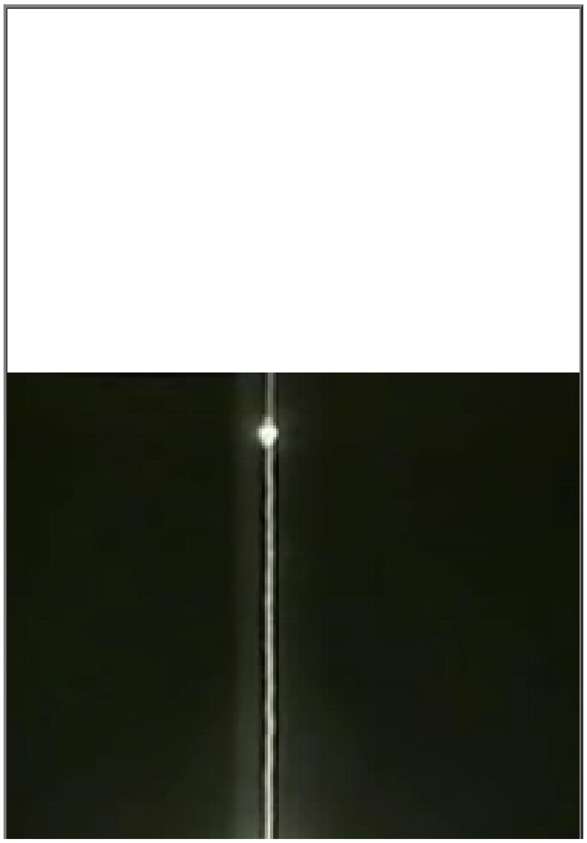

213506148 DAY:056

Tether Activated

Tether Satellite begins to glow as it gathers electricity from the Ionosphere

..

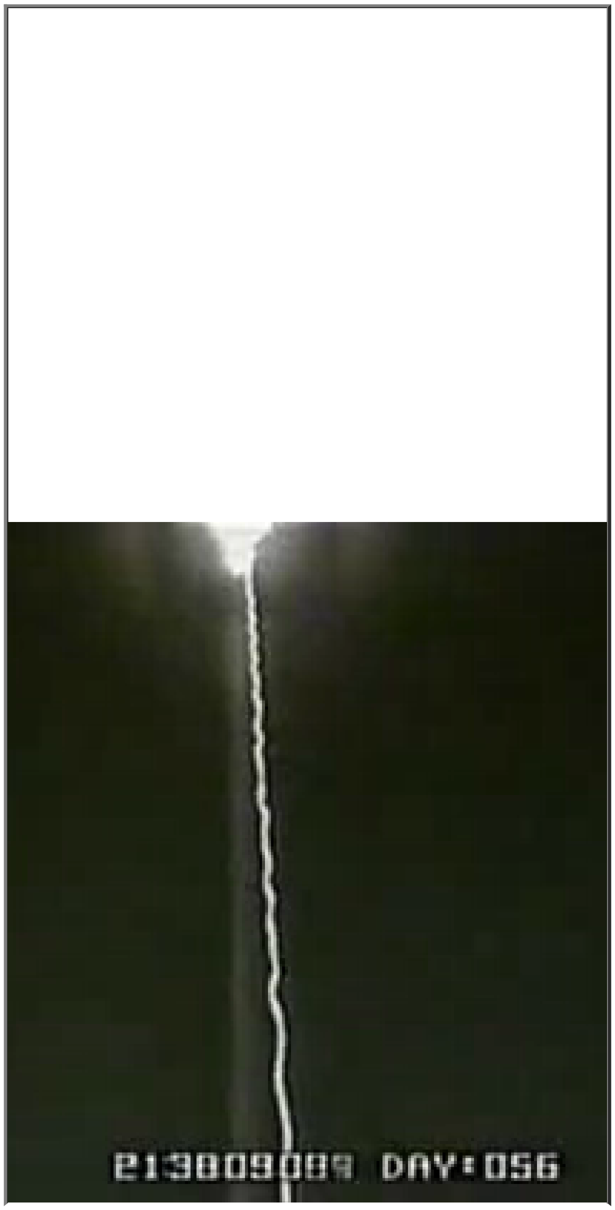
213809089 DAY:056

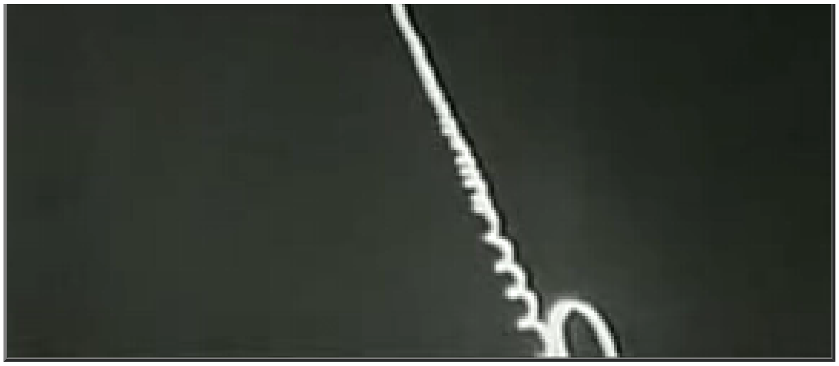

Tether Breaks Free

**Tether Satellite cable fries the connection and breaks free... still producing energy.
The glow increases and the cable gets wider. (More on this later)**

-

Tether Broke Free

Tether Satellite worth 100 million has overloaded the circuits within minutes of being activated because the NASA Scientists under estimated the potential. The tether in just a few moments of operation produced 10 times the anticipated power and they forgot to install a circuit breaker.

-

NOTES:
Add "Critter" Portion
Add Video
Add FLV Viewer

TETHER BROKEN FREE AND DRIFTING AWAY

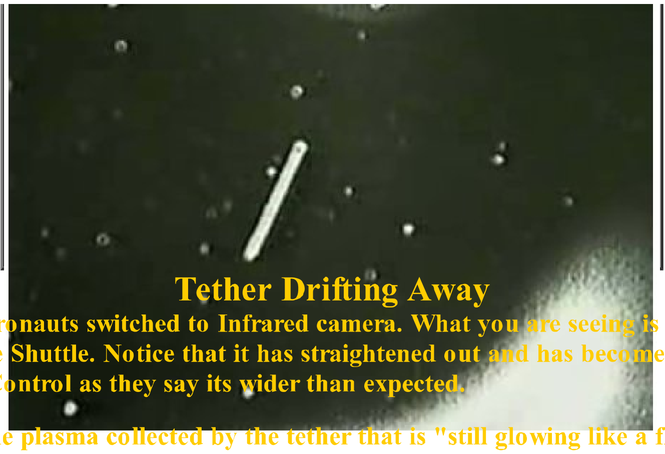

Tether Drifting Away

At this point the STS 75 Astronauts switched to Infrared camera. What you are seeing is the Tether at about 80 nautical miles away from the Shuttle. Notice that it has straightened out and has become much wider. This is commented on by Mission Control as they say its wider than expected.

It is also at this point that the plasma collected by the tether that is "still glowing like a fluorescent bulb" (according to a NASA report issued after the event (See Below)) attracts the "Critters" or UFO's as most people call them. To see more about the "Critters" VISIT HERE In the image include below you can see the Tether Satellite at the tip. By this time the Tether is almost as wide as the Satellite itself.

Going behind the tether you see one of the "Critters" It is CLEARLY behind the tether. This is what NASA calls "dust and debris" The purpose of this paper is to show the tether however, not the "Critters". If you want to know more about them either email us or check out the pages linked to above.

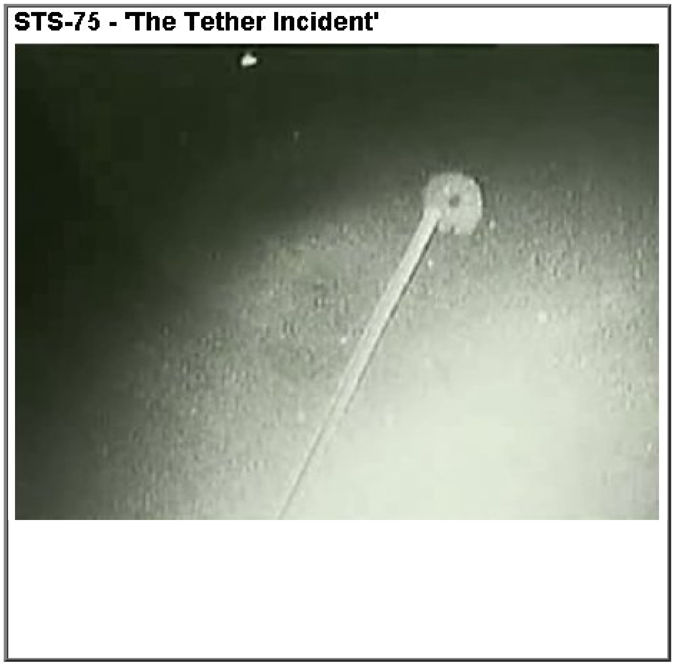

Tether Swarmed by "Critters"

Matt (Matyas) had written me that he has found documents hinting at a Tether in operation near Jupiter. He is still following that up, but it seems the Cassini Spacecraft has found one or more in the vicinity of Saturn

Cassini Spacecraft Image

Is this the lost Tether Satellite? Or is there another one in operation near Saturn?

-

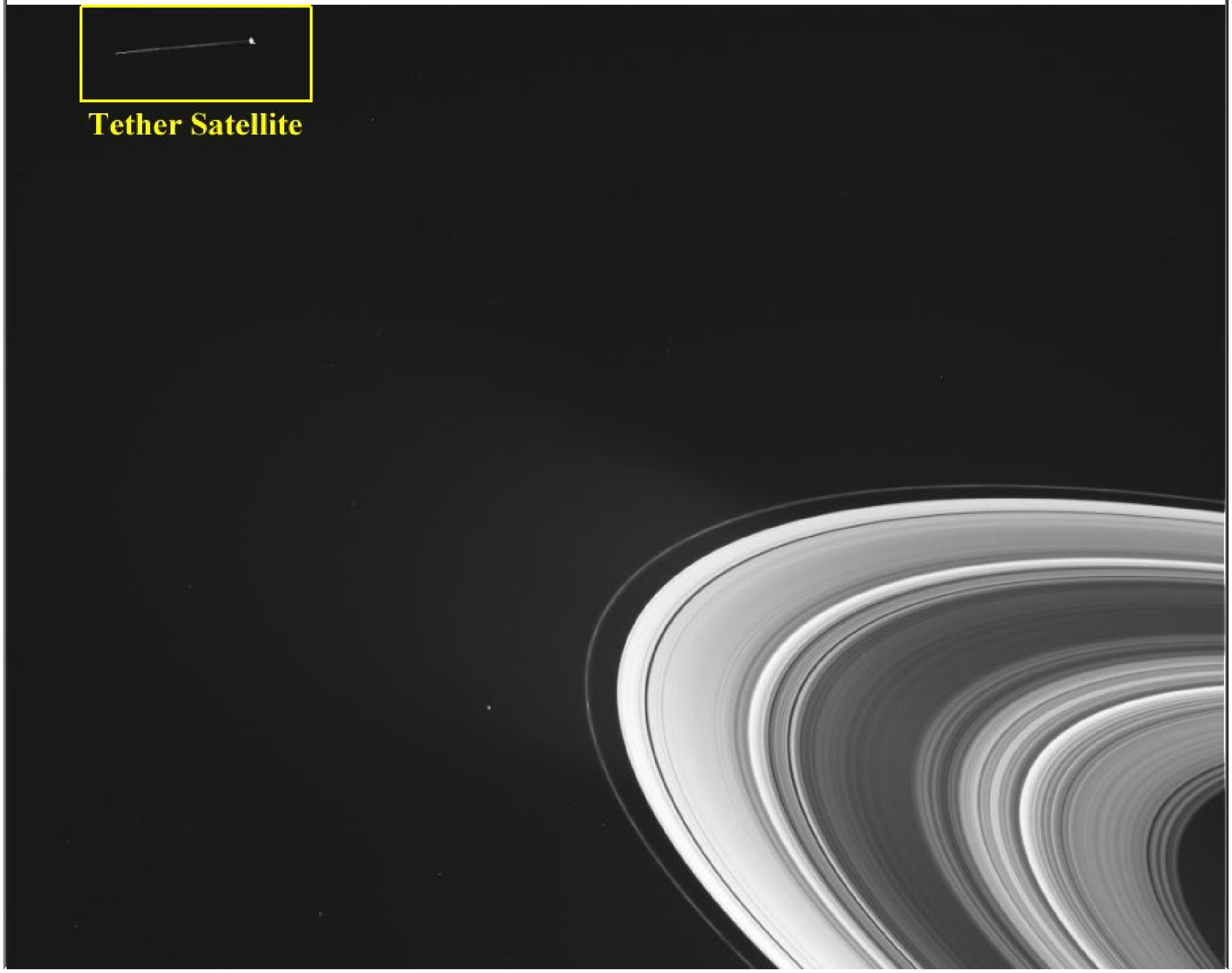
Tether Satellite

Tether Near Saturn

Judging by the scale of Saturn, it would seem this is NOT the lost Satellite, but a much larger one. The STS 75 Tether was 12 miles long... This one is huge

(Note Dates and Size are being calculated)

[Source: Original Image NASA Cassini/ Huygens W00028414.jpg]

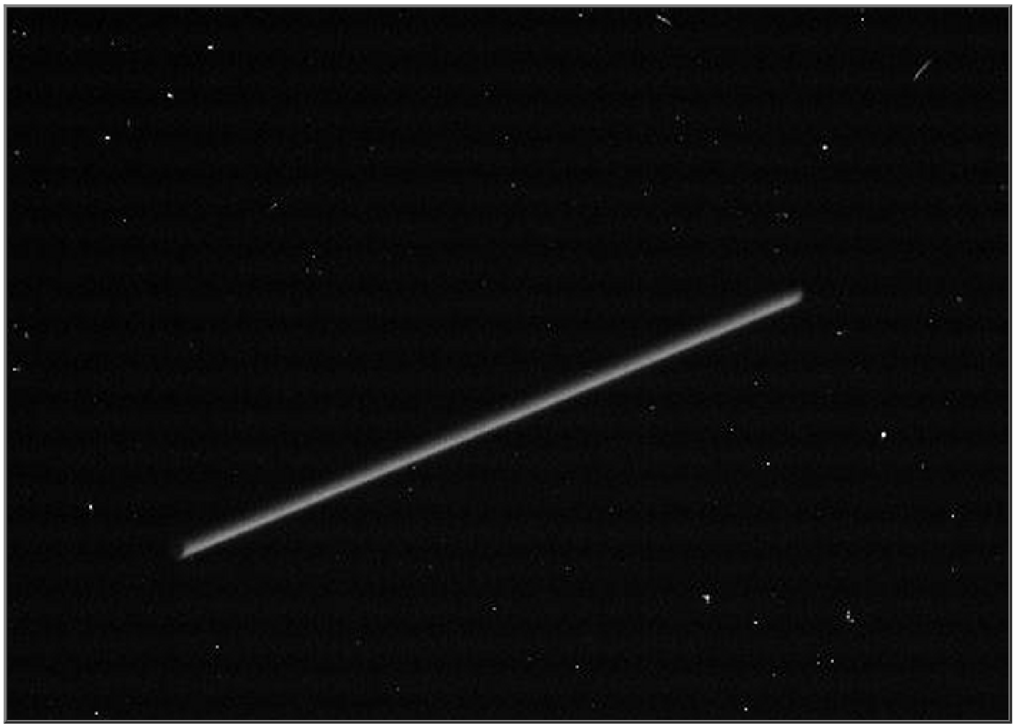

Cassini Spacecraft Image
Tether Near Saturn

..

Cassini Spacecraft Image
Tether Near Saturn

Source: Original Image NASA Cassini/ Huygens W00026015.jpg

Cassini Spacecraft Image

Source: Original Image NASA Cassini/ Huygens N00008771.jpg

~ MENU ~

US006116544A

United States Patent [19]

Forward et al.

[11] **Patent Number: 6,116,544**

[45] **Date of Patent: Sep. 12, 2000**

[54] **ELECTRODYNAMIC TETHER AND METHOD OF USE**

[75] Inventors: **Robert L. Forward**, Houston, Tex.; **Robert P. Hoyt**, Seattle, Wash.

[73] Assignee: **Tethers Unlimited, Inc.**, Seattle, Wash.

[21] Appl. No.: **08/929,271**

[22] Filed: **Sep. 12, 1997**

[51] **Int. Cl.**^{7} **B64G 1/00**

[52] **U.S. Cl.** .... **244/158 R**; 244/166; 244/172

[58] **Field of Search** .... 244/158 R, 164, 244/166, 172

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,097,010	6/1978	Colombo et al.	244/158 R
4,824,051	4/1989	Engelking	244/158 R
4,923,151	5/1990	Roberts	244/158 R

OTHER PUBLICATIONS

"Tethers Open New Space Options" by Ivan Bekey, *Astronautics & Aeronautics*, 1983.

Primary Examiner—J. Woodrow Eldred
Attorney, Agent, or Firm—Arthur M. Dula

[57] **ABSTRACT**

The present invention comprises an electrodynamic tether structure and a method of use. The structure of the tether taught by the present invention is a short, wide, interconnected multiwire (compared to the long, narrow single wires of the prior art) conductive tether whose area maximizes electrodynamic drag while simultaneously minimizing the Area-Time-Product swept by the tether during its operating life. The preferred tether length is two kilometers to five kilometers. The preferred tether mass is one percent to five percent of the spacecraft mass. The method of operation comprises orienting the tether structure at an angle to the local vertical to maximize electrodynamic drag on the host spacecraft and minimize tether instability. The angle of 35.26 degrees is preferred.

26 Claims, 12 Drawing Sheets

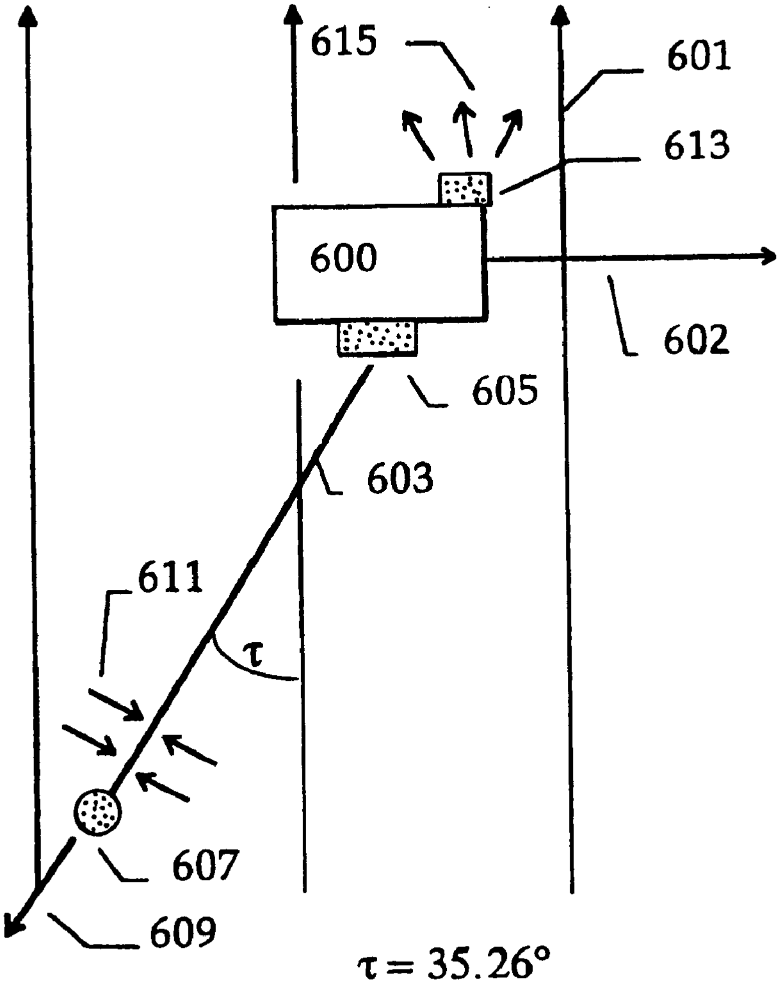

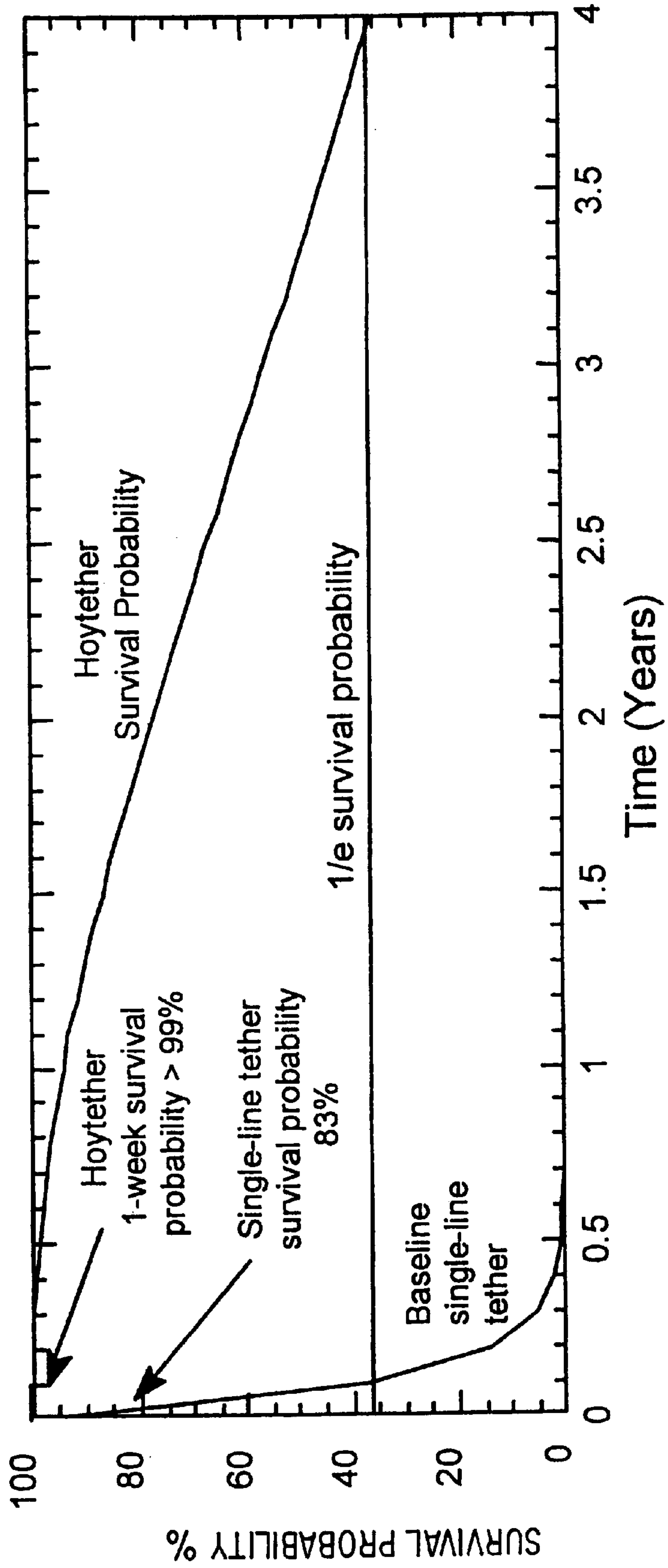
Hoytether
Survival Probability
1/e survival probability
Hoytether
1-week survival
probability > 99%
Single-line tether
survival probability
83%
Baseline
single-line
tether
SURVIVAL PROBABILITY %
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Time (Years)

FIG. 1a

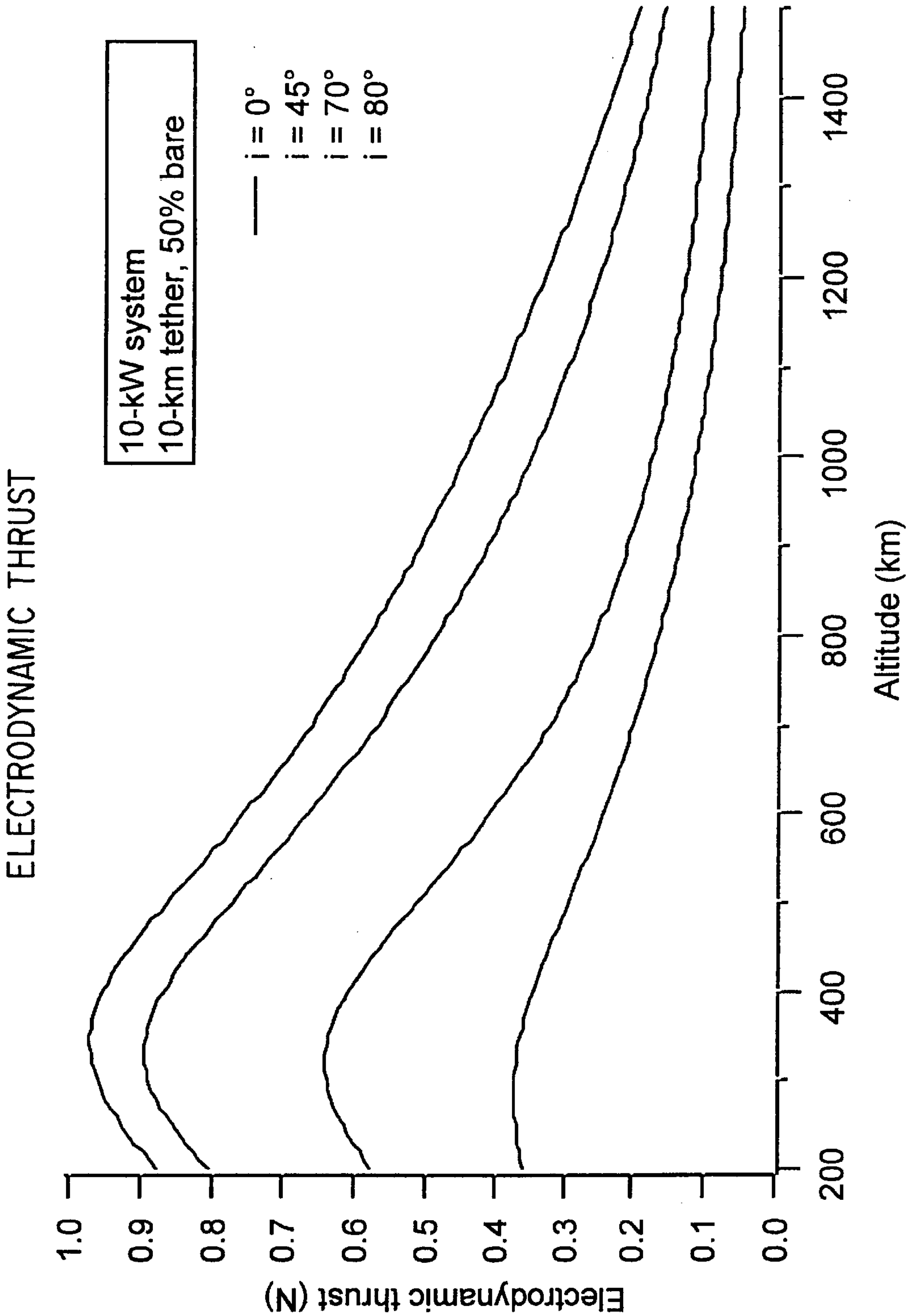
ELECTRODYNAMIC THRUST
10-kW system
10-km tether, 50% bare
i = 0°
i = 45°
i = 70°
i = 80°
Electrodynamic thrust (N)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
200
400
600
800
1000
1200
1400
Altitude (km)

FIG. 1b

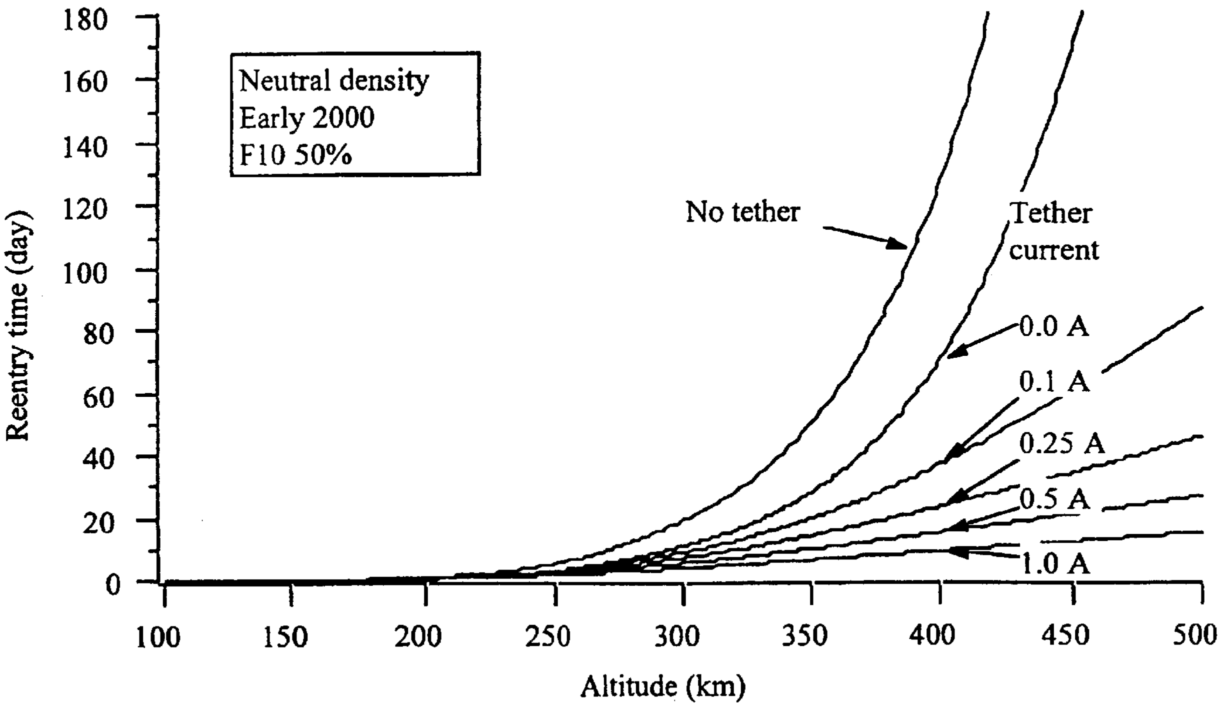

Reentry Time Sensitivity
Unpowered 5 km ProSEDS Tether

Enrico Lorenzini, SAO, NASA ProSEDS Mission (July 1997), NASA/MSFC Grant NAG8-1303

FIG. 1c

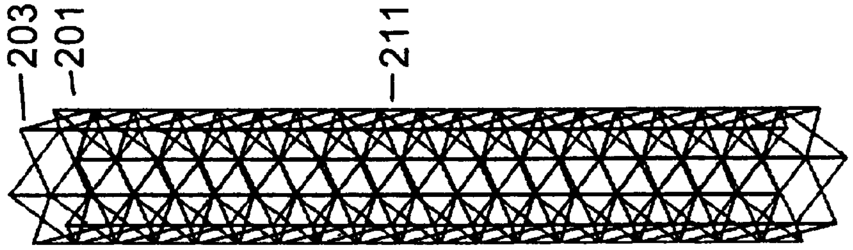

FIG. 2a

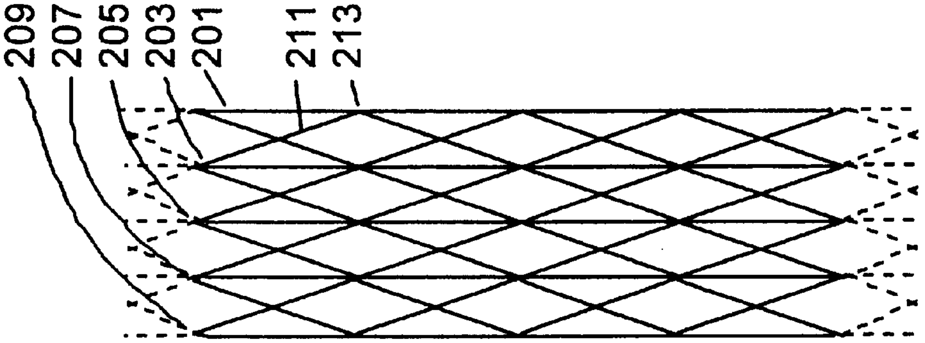
209
207
205
203
201
211
213

FIG. 2b

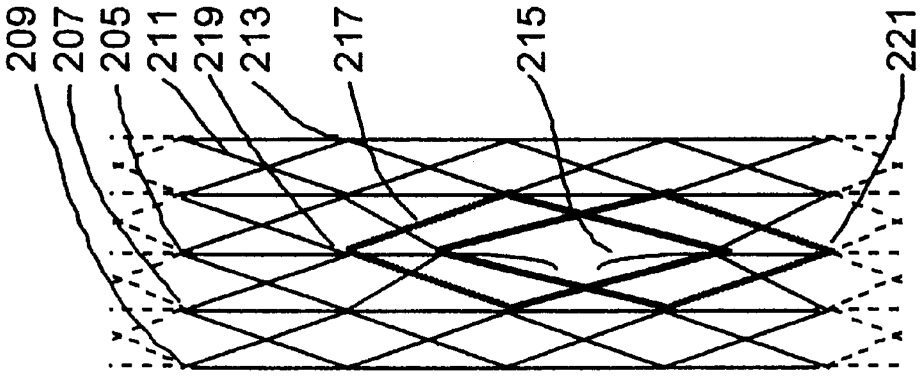
209
207
205
211
219
213
217
215
221

FIG. 2c

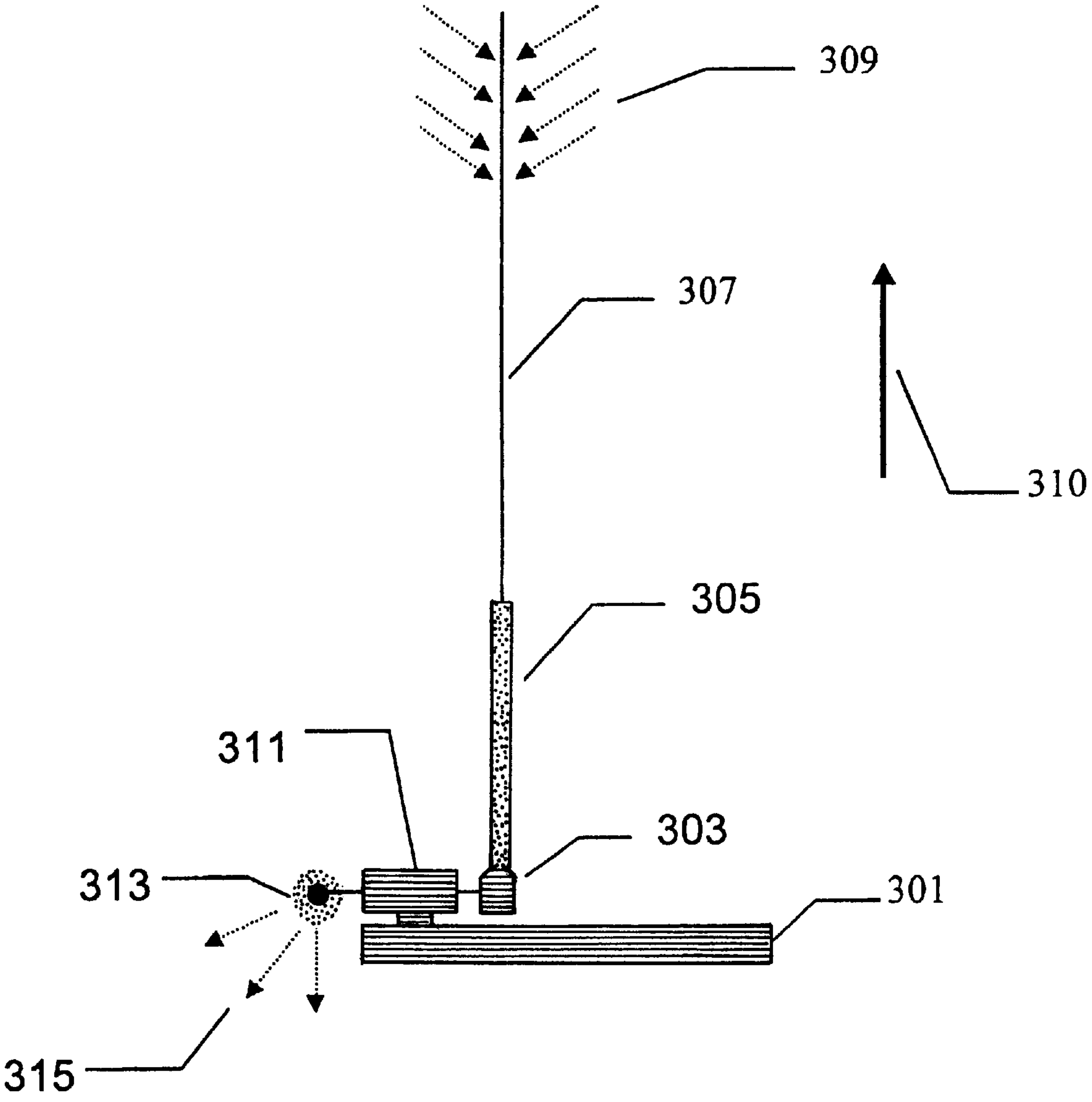
309
307
310
305
311
303
313
301
315

FIG. 3

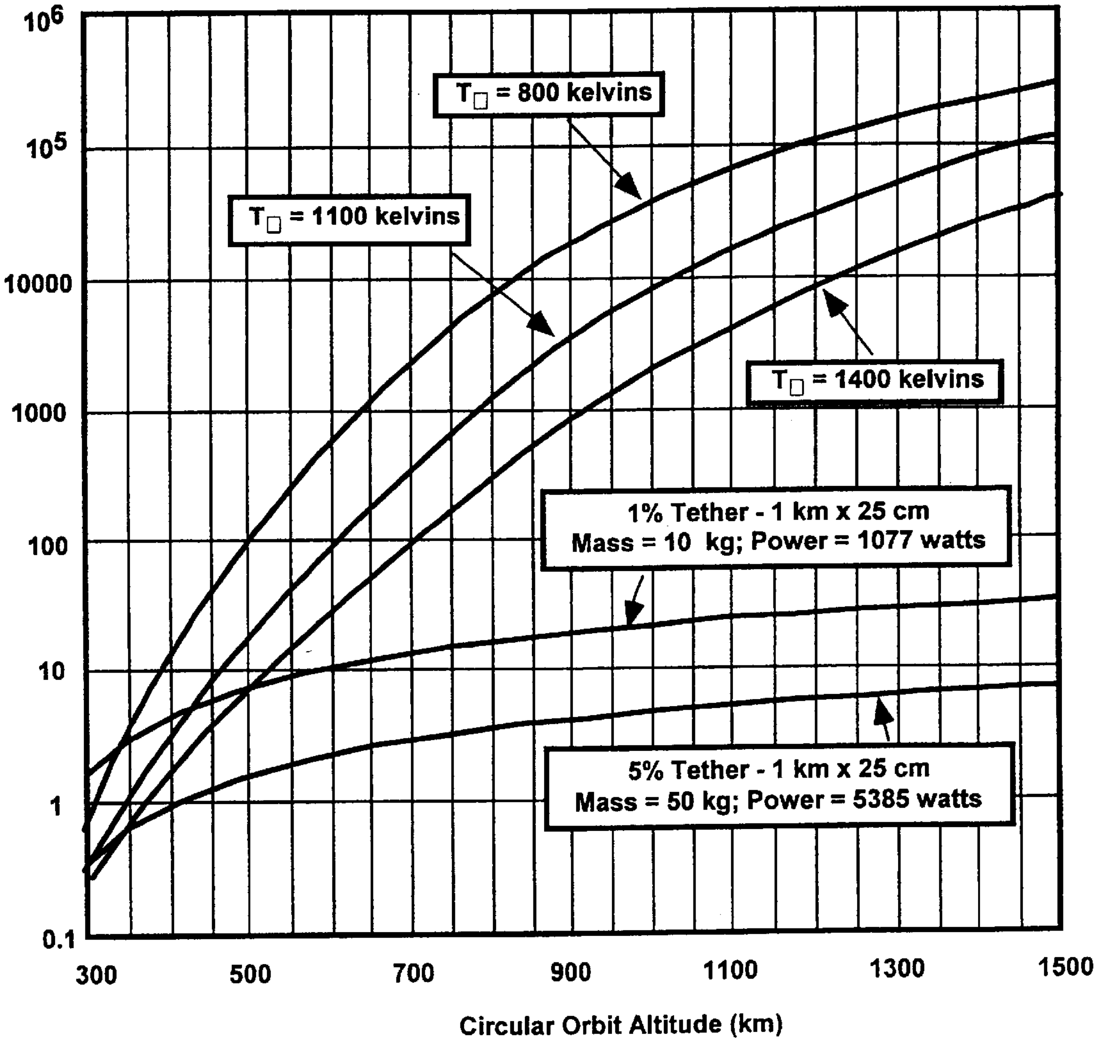

Area-Time Product for Mean and Extremes of Exospheric Temperature

(m = 1000 kg, Cd = 2.0 ; 1% Tether, 1 km x 25 cm; power = 1077 watts)

(5% Tether, 1 km x 25 cm; power =5385 watts)

FIG. 4

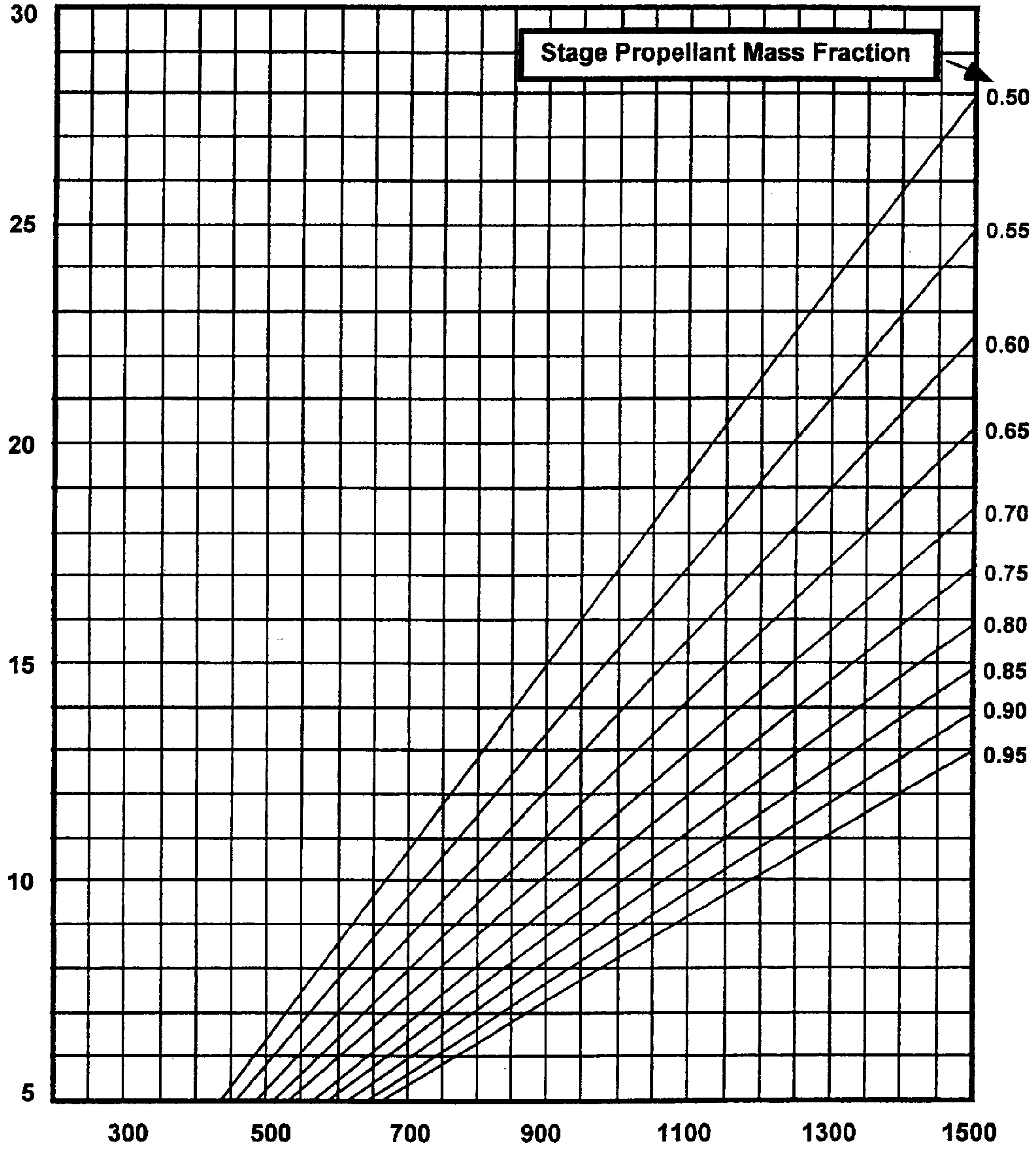

Conventional Solid Rocket Motor System

Percent Increase in Orbital Mass vs. Altitude and I' (Isp = 288 sec)

FIG. 5

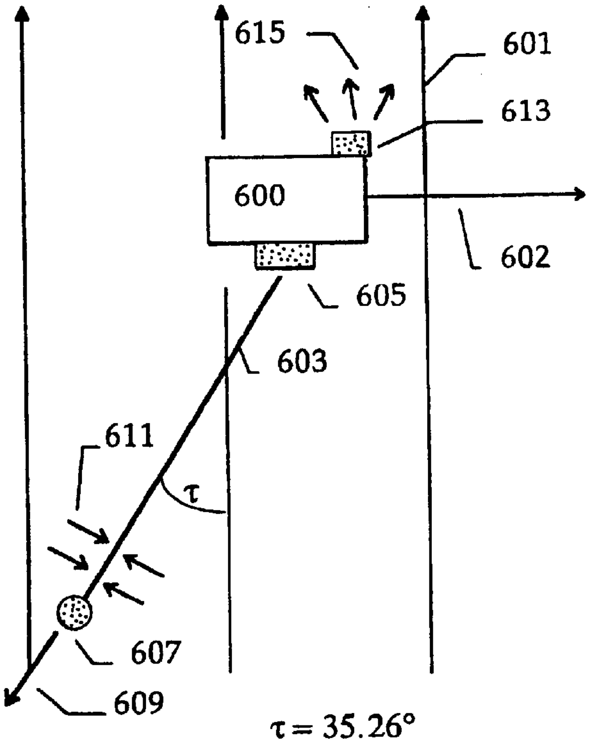
τ
τ = 35.26°

FIG. 6

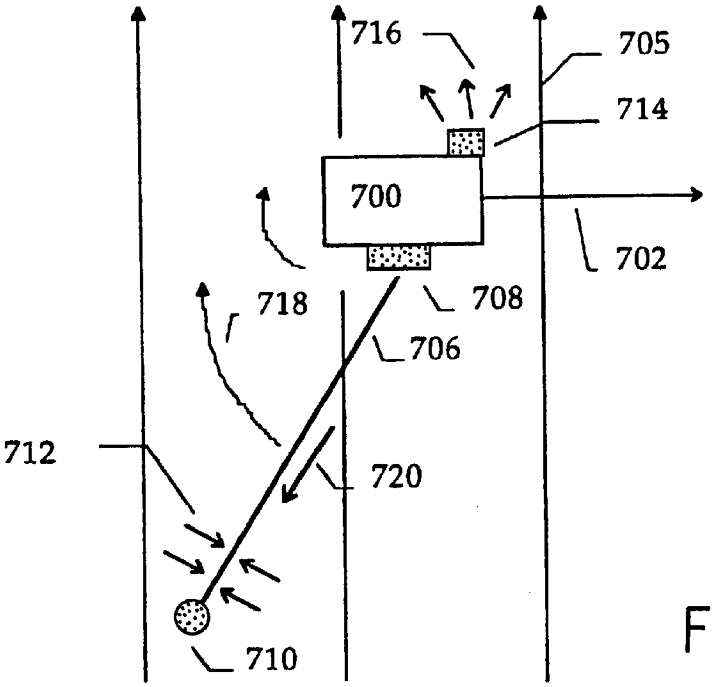

FIG. 7

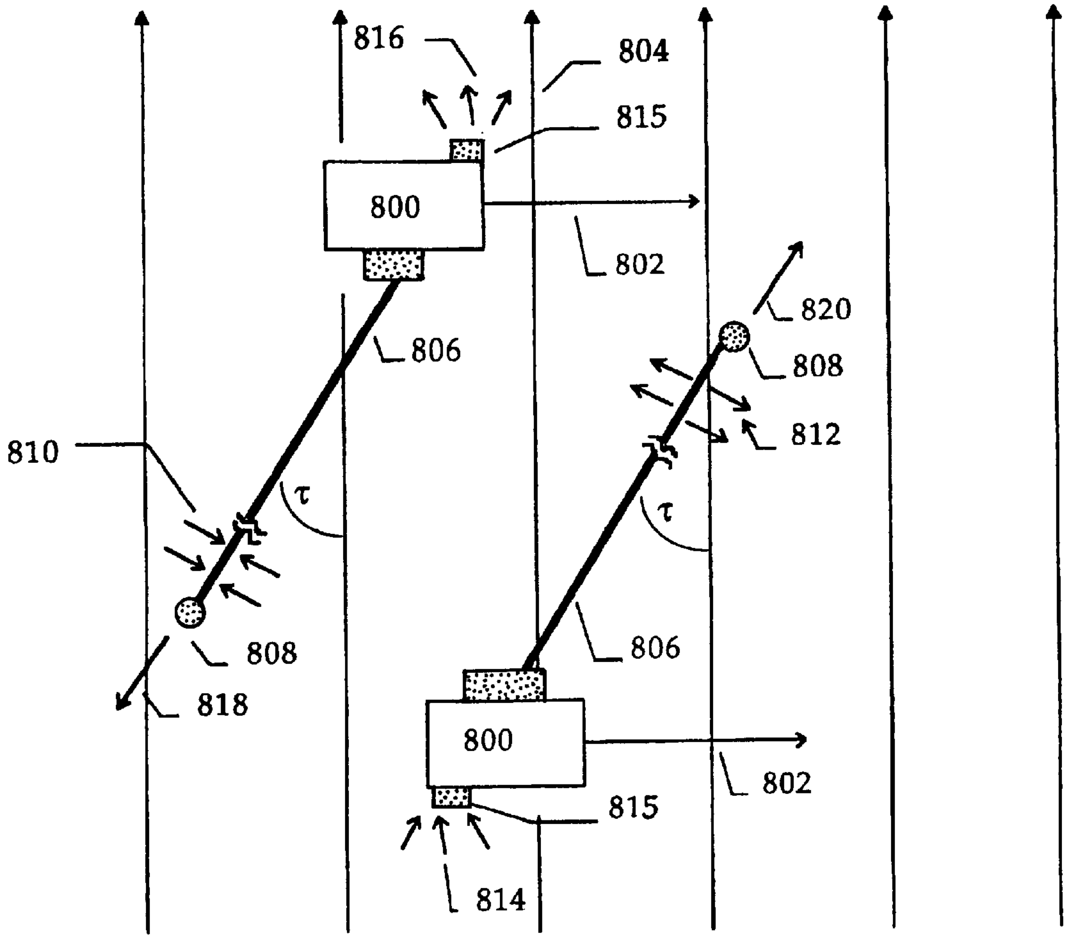
τ
τ

FIG.8

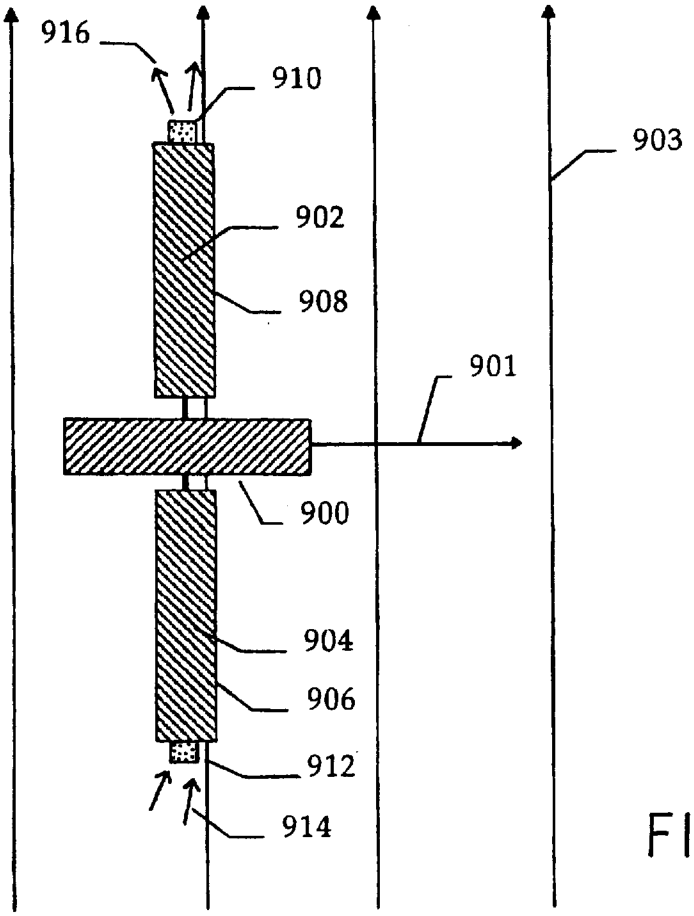

FIG. 9

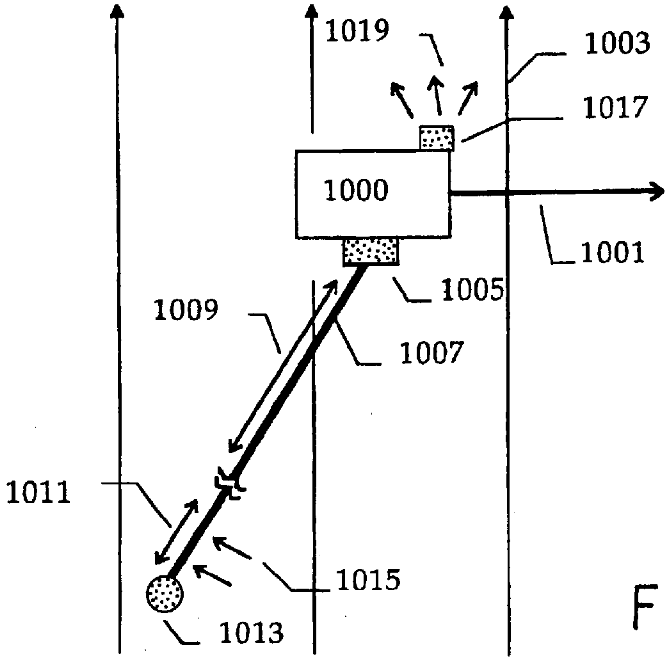

FIG. 10

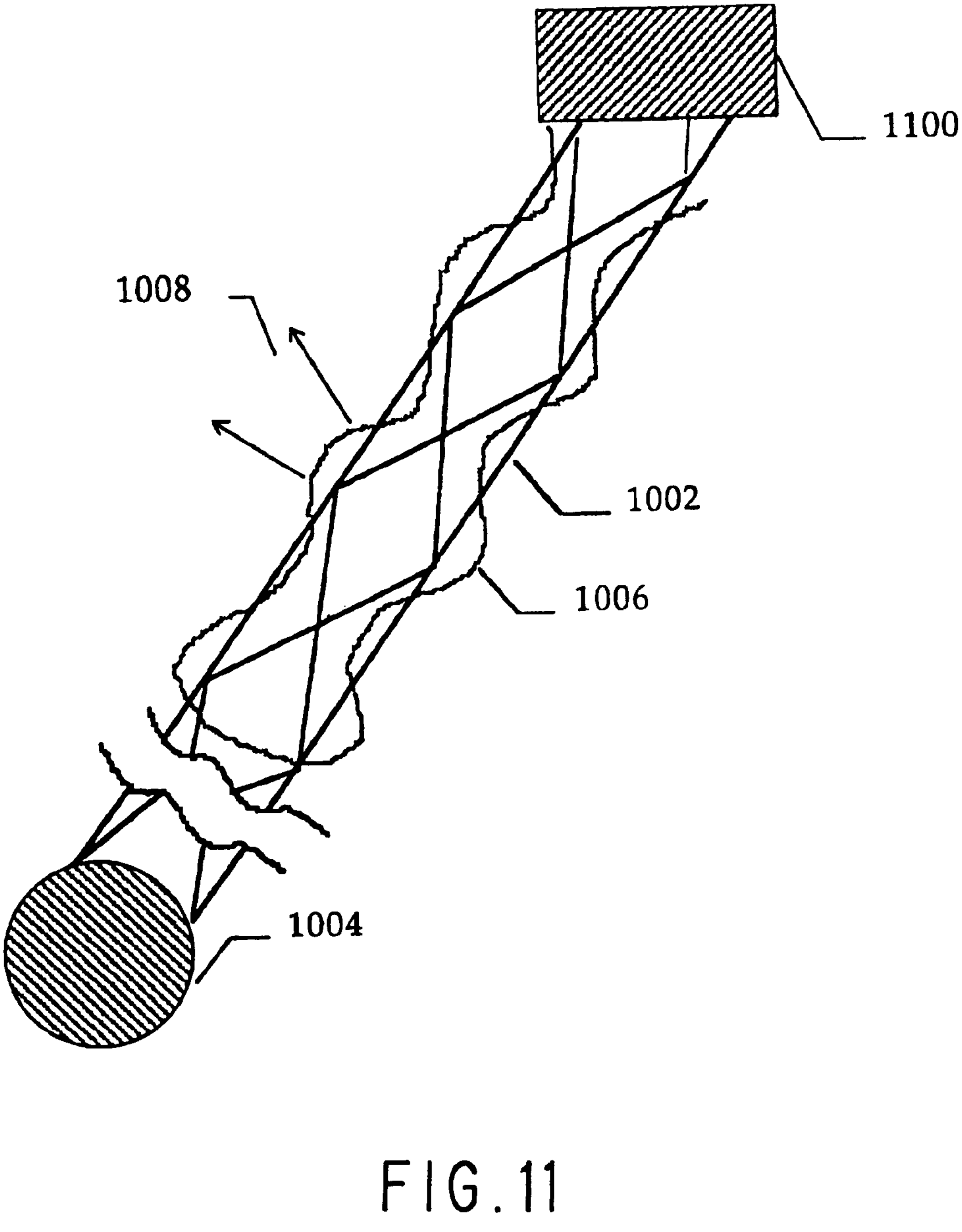
1100
1008
1002
1006
1004

FIG. 11

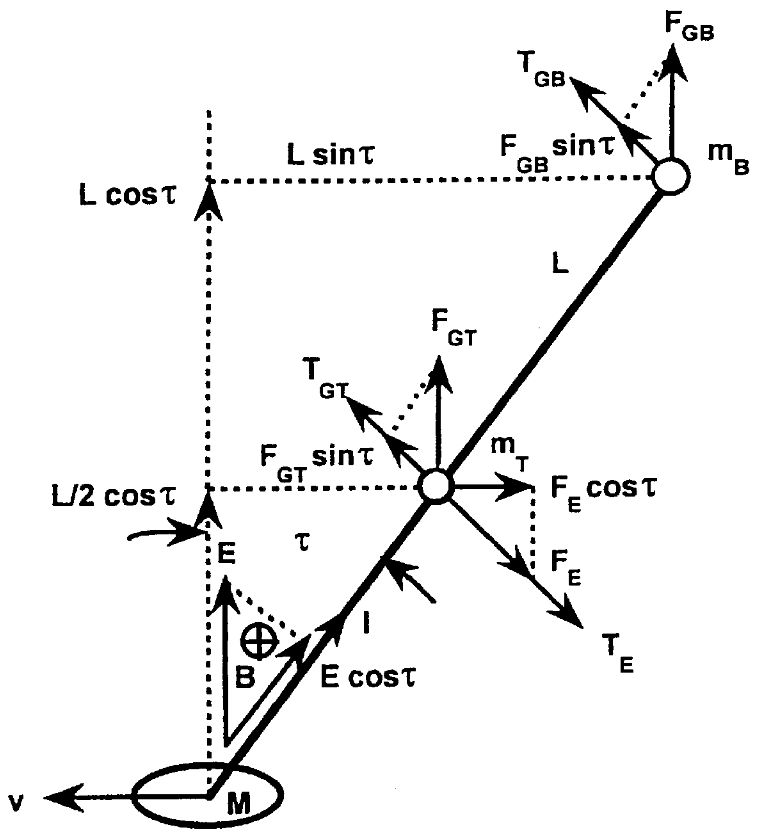
F_{GB}
T_{GB}
$F_{GB}\sin\tau$
m_B
$L\sin\tau$
$L\cos\tau$
L
F_{GT}
T_{GT}
$F_{GT}\sin\tau$
m_T
$F_E\cos\tau$
$L/2\cos\tau$
E
τ
F_E
I
T_E
B
$E\cos\tau$
v
M

FORCES AND TORQUES ON AN UPWARD DEPLOYED CONDUCTING TETHER DUE TO THE MOTION OF THE HOST SPACECRAFT THROUGH THE MAGNETIC FIELD OF THE EARTH.

FIG. 12

ELECTRODYNAMIC TETHER AND METHOD OF USE

TECHNICAL FIELD

This invention relates generally to apparatus and methods useful for changing the state vector of a space object when the space object is moving relative to a magnetic field. More specifically, the present invention relates to an apparatus and method of using a conducting tether to produce an electrodynamic force to deorbit a satellite from its orbit around a celestial body, such as the Earth, which has an associated magnetic field.

RESERVATION OF RIGHTS

This application is subject to certain rights of the U.S. government as a result of contracts between the U.S. government and the inventors.

BACKGROUND ART

The present invention has its principal utility in outer space, primarily for deorbiting satellites at the end of their useful life to mitigate the harm and reduce the liability created by the proliferation of space debris. In order to obtain a better understanding of the present invention it is helpful to understand the prior art of space tethers, especially tether dynamics and tether electrodynamics. The present invention may be more readily understood through a review of the experimental prior art and a mathematical analysis of electrodynamic space tethers.

Prior Art Tethers:

A tether was originally a rope or chain used to fasten an animal so that it grazed only within certain limits. Tethers have been used for decades in space to attach astronauts to their spacecraft.

In 1974 Professor Guiseppe Colombo, holder of the Galileo chair of astronomy at the University of Padua in Italy, proposed using a long tether to support a satellite from an orbiting platform. U.S. Pat. No. 4,097,010, which issued to Professor Colombo and Mario Grossi on Jun. 27, 1978, teaches a satellite connected by means of a long tether to a powered spacecraft. Colombo actively pursued the design of a tethered satellite system.

Several NASA experiments, such as the two Small Expendable Deployer System (SEDS 1 & 2) experiments and the Plasma Motor Generator (PMG) experiment used tethers in space. SEDS used a nonconducting tether. The PMG used a 500-meter conducting tether. The Tethered Satellite System flights in 1992 and 1996 (TSS-1 & 1R) used a 20,000-meter conducting tether.

On the TSS-1 mission the tether deployed only 260 meters (853 feet) before the deployer failed. On the TSS-1R the tether was deployed 19,500 meters. In the SEDS-2 flight, a 0.8-mm diameter, 20,000-meter long braided single-line tether was deployed to study tether dynamics and lifetime. Orbital debris or a meteoroid severed this tether in less than four days.

In the TSS-1R flight, the conducting single-line tether was severed after five hours of deployment. This failure was caused by an electric arc produced by the 3,500 volts of electric potential generated by the conductive tether's movement through the Earth's magnetic field.

The Tether Physics and Survivability (TiPS) satellite consists of two end masses connected by a 4,000-meter long non-conducting tether. This satellite was deployed on Jun. 20, 1996 at an altitude of 1,022 kilometers (552 nautical miles). Its tether is an outer layer of Spectra™ 1000 braid over a core of acrylic yarn. The yarn will "puff" its outer braid to two millimeters to "give it a larger cross section to improve its resistance to debris and small micrometeoroids", according to the National Reconnaissance Office (NRO), which is a sponsor of the TiPS mission. As of Jun. 21, 1997 the TiPS tether had survived one year.

References:

1. Joseph A. Carroll, "SEDS Deployer Design and Flight Performance", paper WSEDSA-1 at the 4th International Conference on Tethers in Space, Washington, D.C., April 1995.
2. James E. McCoy, et. al. "Plasma Motor-Generator (PMG) Flight Experiment Results", pp.57–84, Proceedings of the 4th International conference on Tethers in Space, Washington, D.C., April 1995.
3. W. John Raitt, et. al. "The NASA.ASI-TSS-1 Mission, Summary of Results and Reflight Plans, pp. 107–118, Proceedings of the 4th International conference on Tethers in Space, Washington, D.C., April 1995.
4. Joseph C. Anselmo, "NRO Orbiting Spacecraft Studies Tether Survivability", Aviation Week, page 24, Jul. 1, 1996.

These experiments all used single line tethers.

The following reference is illustrative of the current state of the art in space tethers: Paul A. Penzo and Paul W. Ammann. Tethers in Space Handbook—Second Edition. NASA Office of Space Flight, NASA Headquarters, Washington, D.C. 20546. See also the hundreds of references in the 33 page bibliography at the end of the handbook.

The "Hoytether"™, an Improved, High-Reliability Tether:

In 1991, one of the present inventors, Robert Hoyt, invented a lightweight net-like structure that provides many redundant load-bearing paths. A number of primary load bearing lines running the length of the structure are connected periodically by diagonal secondary lines. The disclosed embodiment of this invention has the secondary lines firmly connected by knots to the primary lines. The secondary lines are connected only to the primary lines. At either end of the disclosed structure, a support ring enforces the cylindrical spacing between the primary lines. The secondary lines are designed with a small amount of slack. These secondary lines are only put under load if a primary line fails. This specific tether structure was disclosed to the public in 1992 (Forward, R. L., "Failsafe Multistrand Tether Structures for Space Propulsion", AIAA paper 92–3214, 28th Joint Propulsion Conference, Nashville, Tenn., 1992 (hereinafter "1992 Hoytether structure"). This structure was named a "Hoytether". The term "Hoytether" is used throughout the remainder of this specification for this type of tether structure.

The present invention uses an improved Hoytether, which was invented by the same inventors as the present invention. This improved Hoytether is the subject of a copending PCT application. The Hoytether is discussed briefly in this specification to aid understanding of the present invention.

The 1992 Hoytether design teaches that the normally slack secondary lines have half the cross-section (0.707 the diameter) of the primary lines. There are twice as many secondary lines as primary lines, thus the mass of the secondary lines is equal to the mass of the primary lines. In an undamaged Hoytether, the primary lines carry the entire load, while none of the secondary lines are under load.

While the survival probability of a single-line tether decreases exponentially with time, the Hoytether can maintain a high, i.e. greater than 99 percent, survival probability for periods of months or years (forward and Hoyt, "Failsafe Multiline Hoythther Lifetimes", Paper AIAA 95–2890, 31st Joint Propulsion Conference, July 1995).

References:

1. Robert L. Forward, Failsafe Multistrand Tethers for Space Propulsion, Forward Unlimited, P.O. Box 2783, Malibu, Calif. 90265, July 1992, Final Report on NASA Contract NAS8-39318 SBIR 91-1 Phase I.
2. Robert L. Forward and Robert P. Hoyt, Failsafe Multistrand Tether SEDS Technology Demonstration, Final Report on NAS8-40545 with NASA/MSFC (Jun. 14, 1995).
3. Robert L. Forward and Robert P. Hoyt, "High Strength-to-Weight Tapered Hoytether for LEO to GEO Payload Transfer" Final Report on contract number NAS8-40690 with NASA/MSFC (Jul. 10, 1996).

The Hoytether is essentially a tri-axial net structure, with 'primary' lines running along the length of the tether and two sets of 'secondary' lines connecting these primaries diagonally. They can be made by hand and connected with knots as is taught by the 1992 Hoytether structure. Because knotted connections severely limit the strength of a structure, it is desirable to use a knotless fabrication technique to achieve interconnections that have strengths approaching the limits of the constituent material. As these tethers may be many kilometers long; fast and inexpensive mechanical methods are required for their practical fabrication.

Hoytethers may be made by mechanical braiding, i.e. three-dimensional braiding, such as 3-D rotation braiding using braiding machines such as those developed by the Herzog Company in Germany (August Herzog Maschinenfabrik GmbH & Co., Postfach 2260.26012, Oldenburg, Germany. The specialized loom developed by the Nichimo Company of Japan (Nichimo Company Ltd., 2-6-2 Ohtemachi, Chiyoda-Ku, Tokyo, Japan) is used to produce "Ultracross" knotless fishing nets in which the individual strands are braided as a 4-braid line, and the strands are interbraided where they cross. This produces netting that has slipless interconnections that are very strong, approaching the maximum capability of the fiber. Such a loom could, with some modifications, produce the present invention's structure. Only two such machines exist, one in Japan, the other in Washington State. Unfortunately neither can work with the small line diameters needed to practice the preferred embodiment of the present invention. See generally, Ko, F. K., "Braiding", in *Engineered Materials Handbook, Vol. 1., Composites*. ASM International, Metals Park, Ohio, 1957. Pp. 519–528.

The most common 3-dimensional braiding machines are 4-step braiders based upon the designs of Maistre (German Patent P230-16986, issued 1973) and Forentine (U.S. Pat. No. 4,312,261, issued 1982). Braiding is accomplished by using pneumatics or solenoids to push the parts of the braiding machine to the proper positions. This is a slow process and making a Hoytether kilometers long with these machines would be very time consuming and expensive. The composites division of Albany International (Albany International Research Company, 777 West Street, Mansfield, Mass.) also produces a 3-D braiding machine. This machine uses modular braiding components that arc assembled breadboard fashion on a large wall.

Although braiding is the preferred technique, alternate fabrication methods such as Raschel knitting and crocheting can be used successfully. Multikilometer long Hoyththers are presently being produced for the inventors by the vendors Culzean Fabrics and Flemings Textiles using an electronically controlled crochet machine produced by Comez in Italy.

Space Tether Systems:

The prior art teaches the use of tethers in space applications. U.S. Pat. No. 5,163,641, issued on Apr. 9, 1990 to Yasaka, teaches the use of a powered spacecraft connected by a tether to a satellite. This tether is disconnected to change the state vector of the satellite. The state of the art of energy and momentum transfer using space tethers is discussed in Ivan Beckey's article "Tethering, a new Technique for Payload Deployment", Aerospace America, March 1997, at pages 36–40. Beckey concludes, "Tethers can perform the same functions as propulsive upper stages of direct payload injection, but at lower weight and cost per pound." U.S. Pat. No. 4,923,151, issued Mar. 1, 1988 to Roberts, Wilknson and Webster, teaches a tether power generator for earth orbiting satellites. U.S. Pat. No. 4,580,747, issued Mar. 15, 1983 to Pearson, teaches use of a long tether extending downward into the atmosphere from a satellite. The state vector of the satellite is changed by forces acting on a lifting body connected to the end of the tether. U.S. Pat. No. 4,824,051, issued Jan. 12, 1987 to Engelking, teaches passing an electric current through a conductive tether attached to a satellite to provide propulsive force to alter the orbit of the satellite. U.S. Pat. No. 5,082,211, issued Jan. 21, 1992 to Werka, teaches use of a tether to deorbit space debris. U.S. Pat. No. 4,727,373, issued Mar. 31, 1986 to Hoover, teaches an orbiting stereo imaging radar system having two spacecraft in synchronous parallel orbits connected by a tether.

Tether Dynamics:

In order to understand the forces that cause a tethered satellite to move upward and away from an orbiting satellite, for example, it is first necessary to explain briefly how a satellite remains in orbit. An orbiting satellite is acted on by the force of gravity which pulls it toward Earth, and by a centrifugal force, which pushes it away from Earth. The centrifugal force" (actually inertia) results from the motion of the satellite around its circular orbit. This is the same force that one can experience by swinging a ball around on the end of a string. A satellite is maintained in its orbit when it travels at the natural speed for its altitude and, as a result, the centrifugal force is equal to the gravitational force.

At the typical orbital altitude of 250 kilometers for a low-Earth orbit satellite, for example, a speed of approximately 7.6-km per second is required to create sufficient centrifugal force to balance gravitational attraction on the satellite. If the altitude is changed, the two opposing forces will no longer be in balance unless the satellite also changes its speed. A higher orbital altitude requires a slightly lower speed so the satellite will take longer to complete an orbit. Because of this, if two free-flying satellites are in orbits at different altitudes, the lower satellite will circle the Earth in less time than the satellite in the higher orbit.

If two satellites, at different altitudes, are connected to each other by a tether, they are forced to travel around their orbits together—in the same period of time, which is longer than the natural period of the lower satellite but shorter than that of the upper satellite. The lower satellite will, therefore, slow down below the natural speed for its orbit and will tend to fall to a lower orbit because the centrifugal force will now be less than the gravitational attraction of the Earth. An upward force in the tether that makes up the difference between centrifugal and gravitational forces holds it in place, however.

Correspondingly, the upper satellite will be accelerated above its natural orbiting speed (increasing its centrifugal force above the gravitational attraction) and will tend to move to a higher orbit. It, too, is held in place by an additional force (downward) in the tether. In other words, the net force downward on the lower satellite is balanced, through the tether, by the net force upward on the upper satellite. The effect of unbalanced forces on the two satellites is, therefore, to create tension in the tether. During the TSS-1 & 1R experiments, the inertia of the tethered satellite causes the satellite to rise above the orbiter as the tether is reeled out. Very close to the orbiter, there is little difference in the two orbits, and the tension force is insufficient to overcome friction in the deployer mechanism; therefore, until the satellite reaches a separation of approximately 1000-meters, the tension is augmented by small tether-aligned thrusters on the satellite. Beyond this point, the tension in the tether is the only force required.

By experimenting with a ball hung on a piece of elastic cord (a paddleball, for example) it is possible to simulate all the different types of oscillations that are possible on a space-based tether system. The elastic cord, representing the tether, may compress and stretch, causing the ball to bounce up and down (longitudinal oscillation). It also may move in a circular (skip-rope) motion or may develop wave-like motions (transverse oscillations). Even if the string itself remains straight, it is possible to get the ball swinging back and forth about its attachment point on the paddle like a child on a swing rope (pendulous motion).

Each type of motion occurs with a particular frequency, which depends on the length and tension of the tether. When the frequencies are different, the motions do not interact; however, at some tether lengths, the frequencies of two or more types of oscillation can become very close. At this point, energy can be transferred from one typc of motion to another, a phenomenon known as resonance. For instance, the transverse oscillations in the tether may cause the satellite to swing back and forth in pendulous motion.

Many different factors may cause oscillations; the movements of the satellite or Shuttle are but two of these. For an electrodynamic tether, the skip-rope and pendulous oscillations are of particular interest. If a current is passed through a tether, the current will interact with Earth's magnetic field, resulting in a force that may produce skip-rope and pendulous oscillations. Because it is necessary to maintain control of the satellite, much study has gone into identifying the different types of possible motions and the methods used to control them.

One way to control the magnitude of those motions that cause a change in tension or transverse motion at the end of the tether is to have an end mass connected to the Hoytether that maintains a controlled tension on the tether, working much like a spring-loaded 'dog leash'. This may be as simple as a coiled spring, or as complex as an active control system that measures the tension and transverse forces on the tether and adjusts the applied tension according to a local or remotely operating algorithm.

Electrodynamic Effects of Conductive Tethers:

Electric potential is generated across a conductive tether by its motion through the Earth's magnetic field. Electromagnetic forces acting on a conductive tether in orbit can make the tether system behave like an electric motor or generator, thereby exerting a useful force to alter the state vector of the tether and any mass attached to it.

Electrodynamic tether propulsion is unlike most other types of space propulsion in use or being developed for space application today—there is no hot gas expelled to provide thrust. Instead, the environment of near-Earth space is being utilized to propel a spacecraft or upper stage via electrodynamic interactions.

A charged particle moving in a magnetic field experiences a force that is perpendicular to its direction of motion and the direction of the field. When a current flows through a long, conducting tether the electrons flowing through the tether experience this force due to the fact that they are moving along the wire in the presence of Earth's magnetic field. This force is transferred to the tether and to whatever the tether is attached (like a spacecraft, satellite, space station or upper stage). It can be an orbit-raising thrust force or orbit-lowering drag force, depending upon the direction of current flow. Operation in one mode allows boost from LEO to higher orbit while reversing the current flow provides negative thrust for deboost. The principle is much the same for an electric motor; reverse its operation and it acts as a generator. The current passing through the tether is returned through the ionosphere to complete the current loop with collection and emission occurring on opposite ends of the tether.

The PMG experiment demonstrated that a conducting tether can be used as both a motor and a generator. The TSS experiments, especially TSS-1R showed that very large voltages (about 3500 volts) can be generated by a sufficiently long tether.

Uses of an electrodynamic tether as an orbit raising and lower propulsion system has many advantages over competing systems:

a. It is nearly propellantless. Most other systems expel hot gases and require extensive resupply. To emit current, the electrodynamic tether propulsion system may use plasma contactors developed as a part of the International Space Station Program. These contactors consume less than 20 kg of xenon gas per year with a 50% duty cycle. The electrodynamic tether propulsion system can also use field emitter arrays, which emit electrons without the use of any gas.

b. It can change both altitude and inclination. The Earth's magnetic field is non-uniform and can therefore provide both in- and out-of-plane forces for inclination changes as well as altitude changes. This is of particular interest to payloads requiring polar orbits in that they can be launched on a small launch vehicle into a lower inclination orbit and have it raised in space by the proper phasing of current through the tether.

A demonstration of the propulsive capabilities of electrodynamic tethers was recently approved for a flight test in 1999. The Propulsive SEDS or ProSEDS mission, will fly as a secondary payload on a Delta II launch vehicle and deploy a 5-km conducting tether using the existing SEDS deployer concept. The ProSEDS experiment will be followed by the Electrodynamic Tether Upper Stage (EDTUS) experiment that will demonstrated the use of electrodynamic forces to change both the altitude and inclination of the experimental spacecraft. FIGS. **1**B and **1**C show the calculated electrodynamic thrust at several inclinations and the reentry time sensitivity of the ProSEDS tether, respectively.

One application for long-life conducting electrodynamic tethers is as a "Terminator Tether™" for removing from orbit unwanted Earth orbiting spacecraft at the end of their useful lives. When the mission of the satellite is completed, the Terminator Tether™, weighing a small fraction of the mass of the satellite, would be deployed. The electrodynamic interaction of the conducting tether with the Earth's magnetic field will induce current flow in the tether conductor. The resulting energy loss from the heat generated by the current flowing through the ohmic resistance in the conducting tether will remove energy from the spacecraft, eventually causing it to deorbit, thus reducing the amount of orbital space debris that must be coped with in the future.

In the following analysis, it is shown that the amount of energy loss generated by an electrodynamic tether is essentially independent of its length or area, and instead is primarily proportional to the tether mass and the physical properties of the conductor metal chosen. In the typical example calculated, a 1000-kg spacecraft can be deorbited from a 1000-km high Earth orbit by a 10-kg mass tether in a month, while a 1-kg tether can deorbit a 1000-kg spacecraft in less than a year.

To the knowledge of the inventors, Joseph P. Loftus of NASA/JSC first proposed the general concept of using an electrodynamic tether to deorbit spent satellites. (Joseph P. Loftus <JLoftus@ems.jsc.nasa.gov>, personal communication via email to Robert Forward, Monday Jun. 10, 1996 15:50:10.) In order to show that the Loftus deorbit concept was not obvious to those skilled in the art of electrodynamic tethers, Forward contacted the leading expert, Joseph Carroll, of Chula Vista, Calif., who built and participated in the flight test of the PMG. After being told of the Loftus concept in a telephone conversation, his reply in an Email message dated Aug. 5, 1996, was "such a system would be feasible . . . by it is still not obvious to me that it would be useful . . . "

Loftus was considering the use of electrodynamic drag from a conducting tether to achieve this goal of bringing the unwanted spacecraft down from its high orbit (where atmospheric drag is negligible) to a 200-km orbit, where atmospheric drag would rapidly finish off the task of removing the unwanted spacecraft from orbit. The tether Loftus was considering was a single-line, conducting tether, typically 1-mm in diameter, 1-km long, and, if made of aluminum, 2-kg in mass. He would include means at the ends of the tether to contact the ambient space plasma around the Earth to complete the current loop.

Unfortunately it is probable that space impactors would sever the 1-mm diameter, 1-km long single-line tether proposed by loftus within a 1/e lifetime of four months. This would produce orbital debris rather than removing it. The motivation for this work is the NASA Safety Standard NSS 1740.14 "(Guidelines and Assessment Procedures for Limiting Orbital Debris." The relevant portion of the Standard starts on page 6-3: General Policy Objective-Postmission Disposal of Space Structures. Item 6-1: "Disposal for final mission orbits passing through LEO: A spacecraft or upper stage with perigee altitude below 2000 km in its final orbit will be disposed of by one of three methods." The method of interest is the atmospheric reentry option, Option a: "Leave the structure in an orbit in which, using conservative projections for solar activity, atmospheric drag will limit the lifetime to no longer than 25 years after completion of mission. If drag enhancement devices are to be used to reduce the orbit lifetime, it should be demonstrated that such devices will significantly reduce the area-time product of the system or will not cause the spacecraft or large debris to fragment if a collision occurs while the system is decaying from orbit."

The NASA standard applies only to NASA spacecraft and even then only to completely new spacecraft designs. New versions of existing designs are to make a "best effort" to meet the standard, but will not be required to change their design to do so. The Department of Defense has adopted the NASA standard with the same provisos. An Interagency Group report has recommended that the NASA standard be taken as a starting point for a national standard. It is NASA's recommendation to the Interagency Group that the safety requirement be phased in only as spacefaring nations reach consensus internationally, which is being done through the International Debris Coordination Working Group whose members are Russia, China, Japan, ESA, UK, India, France, Italy, and the US.

Thus, although the NASA Safety Standard in its present form is not the "Law", the existence of the standard means that some time in the future a similar requirement may be imposed on all spacecraft. This could result in major growth in future space tether business, with a sale to every non-geostationary spacecraft being "mandated" by government safety regulations, somewhat as the sale of seat belts and airbags for every car are mandated.

In fact, three of the companies planning to set up "constellations" of low to medium orbit communications: Teledesic, Iridium and Odessey have committed their companies to abide by the spirit of NASA Safety Standard 1740.14 by using one means or another to deorbit their spacecraft before they reach end of life.

Problems with Prior Art Tethers:

All electrodynamic tether designs proposed by the prior art teach that the tether should be operated at a right angle to the magnetic field through which the tether is moving. This is a problem because the electrodynamic force acting on the tether causes the tether to align itself with the magnetic field force lines. To overcome this problem the prior art teaches the use of a large ballast mass attached to the end of the tether and/or use of a very long (tens to hundreds of kilometers) tether. The large ballast mass is expensive to take to orbit because it replaces useful payload. The long tether sweeps a larger Area-Time-Product during its useful life and thus is more likely to impact other space objects, either debris or another spacecraft.

Another problem common to all proposed prior art tethers is tether instability. If the tether produces a large electrodynamic drag force, which is desirable because a large drag force will cause the satellite to deorbit quickly, then the tether will be dynamically unstable. This instability can cause the tether to lose its effectiveness, act uncontrollably and even wrap around the satellite or otherwise malfunction. Experts skilled in the art of tether design have opined that this dynamic instability is inherently unavoidable in any electrodynamic tether system. The prior art solution, such as that presently being used in the ProSEDS experiment, has been to use a large ballast mass to increase the stabilizing gravity-gradient force and/or to limit the electrodynamic drag of the tether to less that the maximum that could be produced. In the ProSEDS experiment, the conducting electrodynamic tether is five kilometers long. To insure stability, it will be augmented by a 20–35 kilometer long non-conducting tether, which to further have stability will have a 40 kilogram ballast end mass.

Yet another problem of all proposed prior art electrodynamic tether systems is how to radiate away the energy produced by the tether's operation. A satellite moving at an orbital velocity of 18,000 miles per hour has a kinetic energy of over 10,000 calories per gram. To put this amount of energy in an understandable perspective, it may be noted that when nitroglycerine explodes it produces about 1,500 calories per gram. Prior art designs of electrodynamic drag tethers teach the use of the electrical energy generated by the tether to charge batteries or operate electronics, with the excess energy being converted into heat by a resistive load. This excess heat must be radiated to the space environment or it will melt the resistive load. Thus the resistive load, and/or its associated radiator structures, must be massive and replace useful payload.

DISCLOSURE OF THE INVENTION

The present invention comprises an electrodynamic tether structure and a method of use. The principal industrial utility of the present invention is to deorbit satellites in Earth orbit at the end of their useful life. This embodiment of the present invention is sometimes referred to in this specification as a "Terminator Tether™" because it terminates the orbital lifetime of the host spacecraft. The structure of the tether taught by the present invention is a short, wide, interconnected-multiwire (compared to the long single wires of the prior art) conductive Hoytether whose area maximizes electrodynamic drag while simultaneously minimizing the Area-Time-Product swept by the tether during its operating life. The preferred tether length is two to five kilometers. The preferred tether mass is one to five percent (1%–5%) of the spacecraft mass. The method of operation comprises orienting the tether structure at a 35.26-degree trailing angle to the local vertical to maximize electrodynamic force on the tether while avoiding tether instability and allowing use of a small tether end mass.

The present invention also teaches that the satellite-tether system may be rotated around its common center of mass to centrifugally produce tension force in the tether structure to oppose forces causing tether instability. The angle of the conductive tether structure of the present invention with respect to the velocity vector of the host spacecraft may be controlled by the method of the present invention so it interacts with the encountered magnetic field to induce a maximum current flow in the tether. This produces maximum electrodynamic drag. All or a portion of this electric power may be stored and then controllably applied to the conductive tether to produce an induced electrodynamic force. This induced electrodynamic force may by used to enhance the drag force, to rotate the tether-satellite system and/or to provide satellite propulsion, i.e. to change the state vector of the satellite for any useful purpose, e.g. to avoid collision or to change the host spacecraft's orbit to an orbit more favorable for more rapid deorbiting.

The present invention also teaches a tether structure that also functions as a thermal radiator and/or plasma contactor. An embodiment of the present invention using conducting elements of the satellite, e.g. the solar arrays, as electrodynamic tether structures is also disclosed.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention may be better understood with reference to the following drawings:

FIG. **1***a* is a graph comparing the survival probability of a single line tether as taught by the prior art to the Hoytether used by the present invention.

FIG. **1***b* is a graph showing the level of electrodynamic thrust for a 10 kilometer, 10 kilowatt tether at various altitudes and inclinations.

FIG. **1***c* is a graph showing reentry time sensitivity of a five kilometer ProSEDS tether for several conditions of current flow in the electrodynamic tether.

FIGS. **2***a*, **2***b* and **2***c* show the Hoytether structure used by the preferred embodiment of the present invention.

FIG. **2***d* shows the Hoytape structure used by the preferred embodiment of the present invention.

FIG. **3** is a block electrical diagram of an electrodynamic tether system.

FIG. **4** is a graph showing a graph of the Area-Time product for three separate levels of exospheric temperature (for neutral drag calculations) and for two separate assumptions regarding tether mass and power drawn from the ambient plasma (for the Terminator Tether™ calculations).

FIG. **5** is the percent additional mass required in orbit to drop the perigee of a circular orbit at altitude, at, to a value of 200 kilometers using standard chemical rocket as taught by the prior art for deorbiting unwanted spacecraft.

FIG. **6** is a deorbit tether system shown operating with its electrodynamic tether at a trailing 35.26-degree angle to the encountered magnetic field, as is taught by the preferred embodiment of the present invention.

FIG. **7** shows an embodiment of the present invention in which the electrodynamic tether—satellite system is rotated about its center of mass to avoid tether instability.

FIG. **8** shows an embodiment of the present invention wherein the tether is powered to provide induced electrodynamic force.

FIG. **9** shows an embodiment of the present invention wherein the solar power system structure of the satellite is used as an electrodynamic tether.

FIG. **10** shows the use of a portion of the tether structure as a thermal radiator and plasma contactor.

FIG. **11** shows the braiding of an ohmic resistive load into the tether structure.

FIG. **12** is a force diagram showing the forces and torques on an upward deployed conducting tether due to the motion of the host spacecraft through the magnetic field of the Earth.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The principal industrial utility of the present invention is the deorbiting of satellites in Earth orbit, which necessarily can happen only in outer space. Until the present invention is reduced to actual practice by use with satellites in outer space and a body of practical experience is thereby obtained, the present invention can only be constructively reduced to practice, such as by this written specification and its associated drawings, diagrams and graphs and by reference to electric measurements made on conducting tethers in space by the PMG and TSS experiments. The inventors provide herein a detailed discussion of the theory of the present invention to help those skilled in the art of aerospace and tether engineering to understand the present invention and to make and use the best embodiment of the present invention known to the inventors at the time this specification was prepared, without undue experimentation.

Although the present invention is discussed in this specification in its preferred embodiment as a means of deorbiting satellites, it must be emphasized that discussion of this specific use in no way limits the broad scope of the present invention. The present invention can be used to change the direction and speed, i.e. the state vector, of any space object providing that object is in motion relative to any magnetic field. This magnetic field may be that of the Earth or of any other celestial body, for example Jupiter or the sun.

In this specification the material forming the structure of the electrodynamic tether is specified as being a conductor. For example, a metal, such as copper or aluminum wire could be used. Likewise, nonmetallic conductor, such as carbon nanotubes, or a conductive polymer could form the conductive structure of the tether.

FIG. **1***a* shows the survival curve for a Hoytether as used by the present invention vs. the survival curve of a prior art single line tether. The survival probability curve of the Hoytether as a function of time does not conform to the well known "1/e decay" shape of a single-line tether. The Hoytether can maintain a high level of survival probability, about 99 percent, until it nears its 'lifetime'. Its survival probability then drops rapidly to zero. A detailed mathematical analysis of the difference between Hoytether and single tether survival probabilities is given in Appendix E "Small Impactor Survival Probabilities of Hoytethers" and Appendix F "Large Orbital Debris Survival Probabilities of Hoytethers", both contained in the Final Report of NASA Contract NAS8-40545. The resulting cut probability with time for the Hoytether has a "bingo curve" shapes. In a bingo game, at least five numbers must be called before anyone can win, and usually many numbers have to be called before one of the bingo cards gets five in a row. In the Hoytether at least four cuts must happen at the same level before any failure occurs, and many cuts have to be made before any one of the levels has all four lines cut. The bingo curve has the property that the probability of survival stays very high for periods short compared to the lifetime. The probability of survival is greater than 99% for periods shorter than 10% of the lifetime. This is much better performance than the 1/e curve of a single line tether, where the probability of survival is only 90% at 10% of the 1/e lifetime.

FIG. **1***b* shows the calculated level of electrodynamic thrust, in Newtons, produced by a ten kilowatt, ten kilometer single conductor tether at altitudes between 200 and 1400 kilometers and at orbital inclinations between zero degrees and eighty degrees.

FIG. **1***c* is a graph showing the reentry time sensitivity calculated for a five kilometer ProSEDS single conductor tether at altitudes between 200 and 500 kilometers with current flow in the tether between on flow and one ampere.

FIGS. **1***b* and **1***c* are the work of Enrico Lorenzini of the Smithsonian Astrophysical Observatory from his study "Performance Evaluation of the Electrodynamic Tether Tug", NASA/MSFC Grant NAG8-1303 (Nov. 12, 1966). They show that experts in the field believe that the present invention will have industrial utility, as the current flow of one ampere in FIG. **1***c* decreases the deorbit time for the calculated case from more than 180 days to less than 20 days.

FIG. **2***a* is an isometric drawing showing the generally cylindrical structure of the Hoytether. In FIG. **2***a*, primary lines **201** and **203** are shown connected via secondary lines **211**. In FIG. **2***b*, which is a schematic illustration of the structure shown in FIG. **2***a*, primary lines **201, 203, 205, 207** and **209** are connected, each to their respective adjacent primary lines, by a plurality of secondary lines **221**. These connections, for example as shown at interconnection **213**, are made by knotless, slipless interconnections, such as Soutache braiding of twisted connections.

FIG. **2***c* shows the Hoytether structure of FIG. **2***b* but with a severed primary line **205** at breakpoint **215**. The second level of secondary lines **217**, shown as bold lines in FIG. **2***c*, redistribute the load from the severed primary line **215** at point **217** back to line **215** at points **219** and **221**, above and below the break, respectively.

For the electrodynamic tether, a preferred embodiment would be to unroll the tubular Hoytetherr™ into a flat Hoytape™ as shown in FIG. **2***d*.

FIG. **3** generally shows the arrangement of functional elements of an electrodynamic tether system. Of course such a system must have a means of attachment to the spacecraft, a deployer, the tether itself, some form of space plasma cathode contactor, some form of space plasma anode contactor, a power dissipation system, and a communication and control system. Each of these individual elements are known in some form to those skilled in the art of aerospace engineering. Thus the present inventors do not feel it is necessary to describe them in detail in this specification.

In FIG. **3**, a space object **301**, i.e. a satellite in Earth orbit, or any other space object either natural or man made, is physically connected to the tether system. The tether system comprises a deployer **303** from which a conductive Hoytether **305** having a bare segment **307** extends upward from space object **301**. The positively biased anode end **309** of tether **305** collects electrons from the ionosphere as space object **301** moves in direction **310** across the Earth's magnetic field. These electrons flow through the conductive structure of the Hoytether to the power system interface **311**, where it supplies power to an associated load, not shown. The electrons then flow to the negatively biased cathode **313** where electrons are ejected into the space plasma **315**, thus completing the electric circuit.

In order to allow those skilled in the art to better appreciate the broad scope of the present invention, the inventors will now provide an example using specific tether and system values to compare the Terminator Tether™ to prior art means, ie a rocket, for deorbiting a satellite.

Terminator Tethers™:

When an Earth orbiting conducting space tether moves through the magnetic field of the Earth, an electric field is generated in the tether that is proportional to the velocity of the conductor, the magnetic field strength of the Earth, and the angle between the conducting tether and the magnetic field lines. From data obtained during the various electrodynamic experiments that have been conducted in space to date, such as the PMG, TSS-1, and TSS-1R experiments, a typical value of the generated emf per unit length of tether of E=100–200 Volts/kilometer can be assumed. The electric potential V developed at the ends of a tether of length L is then V=EL. For a tether of length L=10 km, the electric potential developed is V=1000 V. For calibration, the 20-km long TSS-1R tether, at the moment of failure, was developing a potential of 3500 volts, which is 175 volts/km.

The mass of a conducting tether of length L, cross-sectional area A, and densin, d is given by m=dLA. Typical values for the density are d=2700 kg per cubic meter for aluminum and 8900-kg per cubic meter for copper. For a typical aluminum tether of mass m=10 kg and length L=10 km, the cross-sectional area will be A=0.37 mm squared. If this were a solid-wire single-line tether, the diameter of the wire would be D=0.69 mm (21.5 gauge). If this were an 18 line tubular Hoytether, the diameter of the lines in the Hoytether would be D=0.16 mm (34 gauge).

The resistance of a conducting tether of length L and cross-sectional area A is given by R=rL/A, where r is the resistivity of the conductor in nano-ohm-meters (nΩ-m). Typical values for the resistivity are r=27.4nΩ-m for aluminum and 17.0 -nΩ-m for copper. For the m=10 kg aluminum tether of length L=10 km and cross-sectional area A=0.37 mm squared at 20 degrees C, the end-to-end resistance is 750 Ω. This value of resistance is the essentially the same whether the conductor area is concentrated into a single-line tether or divided up into many lines as in a multiline Hoytether.

The current I generated in the conducting tether by the electric potential V between the ends of the tether applied across the tether resistance R is I=V/R. For the 10 kg mass aluminum tether of length L=10 km, electric potential V=EL=1000 volts, and resistance R=750 Ω the current is I=V/R=1.33 Amps. Currents near these values were measured in the TSS-1R experiment at the time of failure. At the time of failure of the TSS-1R tether, none of the plasma contactors on the Space Shuttle or the Italian Satellite were operating. Thus the current was being collected by the conductive surface area of the Space Shuttle and the Italian

Satellite at the two ends of the tether. The Space Shuttle area is quite large, so it was not the limiting factor in current collection. The diameter of the Italian Satellite was 1.6 m, which would give it an effective plasma contact area of about 8 square meters. This shows that, if the plasma contact area of the ends of a conductive space tether can be make large enough, then ampere level currents can be extracted from the ambient space plasma without the use of space plasma contactors.

The power dissipated as ohmic heating in the tether is given by P=IE. For an aluminum tether of mass m=10 kg, resistivity r=27 nΩ-m, and density d=2700 kg/cubic meter, subjected to an electric field of E=100 V/m, the power dissipated in the ohmic losses of the conductor is P=IV=1330 Watts. For a 1 kg mass tether, it would be still a considerable 133 Watts of dissipation.

There will no doubt be additional dissipation of energy in plasma ohmic losses, plasma wave generation, and plasma ion acceleration, but the ohmic losses in the conducting tether alone are sufficient for the task of deorbiting an unwanted spacecraft massing 100 to 1000 times more than the tether.

The decay time of a metric ton spacecraft moving from a 1000 km altitude orbit to a 200 km altitude orbit with an energy difference of dU=3.3GJ, when its energy is being dissipated at a power of P=1330 W by an aluminum tether massing just 10 kg, or 1% the mass of the spacecraft, is about one month. This is a remarkably short time, and indicates that the concept of using a conductive tether to deorbit a spacecraft is indeed feasible. If the aluminum tether massed only 1 kg, or 1/1000th the mass of the spacecraft it was deorbiting, then the decay time would rise to 10 months, still a reasonable value.

In reality, of course, the actual decay time will be longer than this. If the electrodynamic drag force is very large, and becomes larger than the gravity gradient forces pulling on the ends of the tether (which force is proportional to the mass of the tether), then the tether will tend to align itself along the magnetic field lines instead of across them, and the drag force will decrease because of the small angle between the conductor length and the magnetic field lines. The tether will then settle into an angle determined by the balance between these two forces.

Is the Terminator Tether™ theoretically a better means than atmospheric drag or a rocket engine for deorbiting satellites at the end of their useful life?

In the formula for da/dt (change in altitude per change in time), using the assumptions of near-circular spiral trajectories, the Area-Time-Product (Z), the criteria by which NASA judges compliance with Safety Standard 1740.14, is given by:

$$Z = A\int dt = -\frac{m}{C_D}\int \frac{da}{\rho(a)\sqrt{\mu a}},$$

where ρ(a) is simply the density as a function of semi-major axis. Thus, for a static atmosphere, the problem of area-time product is reduced to quadrature. Note that the area-time product depends only on the density profile and the ratio of spacecraft mass to drag-coefficient and is linearly related to that ratio.

The basic altitude-dependence of the atmospheric density, represented here by ρ(a) can be expressed in a low-order power series of ln(density)=f(ln(altitude)), so that it will be possible to develop a representation of the log—log relationship between altitude and density. The three static levels of atmospheric density are modeled as 5th order polynomials in the natural log of the altitude and are representative of well known atmosphere models, for example, Jacchia, L. G., “Thermospheric Temperature, Density, and Composition: New Model,” SAO Special Report 375, March 1977. 2. Anon., Marshall Space Flight Center, “Long-Range Statistical Solar Activity Estimation,” Atmospheric Sciences Division, 1989.) The values of exospheric temperature are reasonable (about 2-sigma) high and low solar activity values of 1400 and 800 kelvins respectively. The mean value is taken to be 1100 kevins.

FIG. **4** shows a graph of the Area-Time product for three separate levels of exospheric temperature (for neutral drag calculations) and for two separate assumptions regarding tether mass and power drawn from the ambient plasma (for the Terminator Tether™ calculations). The density profiles for the three values of exospheric temperature represent the extremes and mean values to be expected during the next several decades. The exospheric temperature goes through a cycle of about an 11 year period with maximum about 1400 kelvins and minimum about 800 kelvins.

The tether calculations include an inherent assumption that the Terminator Tether™ transfers all the energy extracted from the ambient space plasma into drag that decreases the orbital energy of the spacecraft at a rate given by the power drawn by the tether. Thus, the change in energy for a spacecraft of mass m at an initial altitude h is given by the expression:

$$\Delta E = -m\left\{\frac{\mu}{2(R_e+h)} - \frac{\mu}{2(R_e+250)}\right\},$$

and the time required to effect this change in energy of the mass, m, is just

$$\Delta t=\Delta E/P,$$

where P is the power drawn by the tether from the ambient plasma charge.

The curves for 1% and 5% tethers do not include the effects of atmospheric drag because the electrodynamic drag is orders of magnitude greater than the neutral atmospheric drag. At altitudes greater than about 700 km, the electrodynamic drag is 200 to 3000 times greater than the neutral drag forces.

It is clear, from FIG. **4** that the Terminator Tether™ concept is far superior to neutral drag in removing spacecraft from orbit, no matter how much additional area is added to the passive spacecraft to increase the atmospheric drag. The analyst should note that FIG. **4** is proportional to mass; that is, the Area-Time-Product values should be multiplied by the mass of the spacecraft and divided by 1000 kg. This is true for the neutral drag and Terminator Tether™ curves alike. The power levels assumed for the 1% and 5% tethers are only 80% of their theoretical values. This is done to provide a 20% margin on the power available from the ambient plasma and electrodynamic gradient.

Conventional rocket mechanisms can remove spacecraft from orbit, but this mechanism is apparently not viable from mass considerations when these factors are compared with the capabilities of the Terminator Tether™. A satellite owner may decide to satisfy the NASA safety requirements by adding a small solid motor and the associated hardware, software, sensors and structure to make the package independent, to be used to deorbit the spacecraft in case the main spacecraft power, attitude, or propulsion system fails.

The requirements of such a rocket motor system are more stringent than those attributed to ordinary spacecraft. The rocket motor deorbit system must operate when some or all other systems of the spacecraft have failed. These more stringent requirements are balanced by lesser requirements of performance. The backup system must simply deorbit the spacecraft; it does not have to perform all the other duties of the spacecraft. But the backup system must know when to fire under all kinds of anomalous situations, including tumbling, offset of center of mass (because of loss of parts due to collisions), and lack of knowledge of the orbital position.

FIG. **5** shows the percent additional mass required by the rocket motor system to drop the perigee of a circular orbit at some initial perigee altitude to a value of 200 km. The atmospheric drag at this 200 km perigee altitude will remove any spacecraft (in the range considered) from orbit in a few revolutions. The contours of constant stage propellant mass fraction, λ, range from low values of 0.5 through reasonable values of 0.65 to 0.7, up to the values associated with the best solid motors ($\lambda \nabla 0.93$) without adding any extra hardware to the emergency stage. An effective, independent stage to provide a retro deltaV of from 50 to 325 m/s will almost certainly have a λ of the order of 0.6 to 0.75. If the emergency stage is required to perform its own attitude determination, the stage propellant mass fraction may be as low as 0.55 or 0.50. Note that the additional mass, shown in FIG. **5**, must also be lofted to orbit in the first place, to provide the mass on orbit for the originally intended service. The stage propellant mass fraction, λ, is a key stage performance parameter that describes the ratio of the mass of propellant to the mass of the stage. That is $\lambda=m_p/(m_p+m_I)$, where m_p is the mass of the propellant and m_I is the mass of everything else in the stage (not including the payload or any stages above or below the stage being considered). Typical independent small rocket stages have values of λ from 0.60 to 0.75. The curves of stage propellant mass fraction are shown as straight lines (a minor approximation) and are included only from 0.5 to 0.95, the extremes of reasonable design practice.

These calculations that show, beyond any reasonable doubt, that the Terminator Tether™ concept is far superior to conventional mechanisms such as drag enhancement devices or small rocket deorbit propulsion systems. The superiority is measured in terms of Area-Time-Product, NASA's measure of the likelihood of collision with other spacecraft in the path of the descending spent member of a constellation. Tether calculations were made using conservative assumptions that the power extractable from the ambient plasma and electrodynamic gradient is only 80% of the theoretical power available to a perfect tether crossing the magnetic field lines at a right angle, i.e. normally.

The following analysis is presented by the inventors to help those skilled in the art to better understand the present invention when it is used as a Terminator Tether™ for removing from orbit unwanted non-geostationary Earth-orbiting spacecraft at the end their useful lives. The primary result of that analysis is that the electrical power P in the tether that is converted into heat by the resistance of the tether and radiated away into space will remove energy from the spacecraft, causing it to rapidly deorbit, thus reducing the amount of orbital space debris that must be coped with in outer space. Tether electric power P is given by:

$$P=(vB)^2m/2rd$$

where m is the mass of the conducting tether, r and d are the resistivity and density of the conducting material, and v is the velocity of the spacecraft's motion through the Earth's magnetic field B. For a m=10 kg tether of aluminum with resistivity of r=27.4 nΩ-m and density d=2700 kg/m^3, moving at a velocity v=7037 m/s relative to the Earth's horizontal magnetic field B=26.5 μT, the power dissipated is P=2350 W! This energy loss in the form of heat must necessarily come out of the kinetic energy of the host spacecraft. For a typical example, a 1000 kg spacecraft in a 1000 km high orbit subjected to an energy loss of 2350 J/s from a 10 kg tether (1% the mass of the host spacecraft) will be deorbited in a few weeks.

Power levels of the magnitude estimated in the previous paragraph have been measured in a real orbital space experiment, the TSS-1R mission carried out on the Shuttle Orbiter in 1995. In that experiment, a large Italian spacecraft, 1.6 m in diameter, was deployed upward from the Shuttle Orbiter at the end of a conducting copper wire tether covered with electrical insulation. As the tether was slowly deployed upwards, a series of measurements were made of the open circuit voltage induced in the tether by its motion through the Earth's magnetic field. The voltage between the end of the tether and the Orbiter ground varied from zero volts at the start to 3500 V when the amount of tether deployed approached its maximum length of 20 km. Periodically, the end of the tether was connected either to one of two different electron guns, which supplied contact to the surrounding space plasma, or to the Orbiter ground. The bare surfaces of the Shuttle Orbiter proved to be a surprisingly good plasma contactor via a combination of ton collection and secondary electron emission. The current flow through the tether was deliberately limited by control circuits and the current capacity of the electron guns to about 0.5 amperes, but power levels of 1800 Watts were reached.

The tether was intended to have a fully deployed length of 20 km, but at a deployed length of 19.5 km, when about 3500 V was being induced at the end of the tether inside the Orbiter reel mechanism, a flaw in the insulation allowed an electrical spark to jump in an uncontrolled manner from the tether to the Orbiter ground. With no control circuits to keep the current level down to 0.5 amperes, the current flow jumped to 1.1 amperes, and the total power generated was P=3850 Watts Most of this energy went into the electrical arc, which burned through the tether, causing it to break and halting the experiment. This experiment showed that large areas of bare conducting material, such as that provided by the 8 square meter area of the Italian spacecraft at one end of the tether and the very large surface area of the Shuttle Orbiter spacecraft at the other end of the tether, can collect and emit amperes of current, while thousands of volts of potential can be generated by sufficiently long tethers moving at orbital speeds.

Thus, both theory and experimental data collected in the space environment indicate that significant amounts of electrodynamic drag force can be obtained from a low mass conducting tether attached to a host spacecraft, provided the ends of the conductor can exchange sufficient numbers of electrons with the surrounding space plasma.

Experimental data from the TSS-1R data also produced the amazing result that the efficiency of a bare metal surface in "contacting" the space plasma is many times better than the standard theory would predict. The 8 square meters of bare surface area of the Italian spacecraft were sufficient to collect the 1.1 A of electron current. This amount of area is easily replicated by a few hundred meters of bare wire, considering that the effective collection diameter around the wire is the Debye length, which is a few centimeters at the typical values for space plasma density and temperature.

Because of this result, that a bare wire can easily collect electrons, Les Johnson, Nobie Stone, Chris Rupp, and others at NASA Marshall Space Flight Center have formed a team, which includes the present inventors, which is embarked on a new flight experiment. The experiment is scheduled for a piggy-back flight on a Delta II launch of an AF Global Positioning Satellite in early 2000. The goal of the experiment is to demonstrate that electrodynamic drag from a wire moving at orbital speeds through the Earth's magnetic field will create a large enough electrodynamic drag force to deorbit the Delta II second stage, whose mass is greater than 1000 kilograms, in a few weeks. This is essentially a demonstration of the Loftus electrodynamic drag deorbit concept and the first step in the development of a Terminator Tether™.

The ProSEDS (Propulsion Small Expendable-tether Deployer System) mission will use a 5 km long copper wire conductive tether massing 18 kg connected to a 20–35 km long nonconducting tether, which is in turn connected to a 25–40 kg ballast mass. The total of 25–40 km of tether length and the 25–40 kg ballast mass on the end will provide enough gravity gradient force to keep the tether aligned near the zenith, so that the direction of the current in the tether is at right angles to both the direction of the spacecraft motion in the nominal EW direction and the Earth's near-equatorial magnetic field in the nominal NS direction.

An important feature of the ProSEDS experiment is that it is designed to be completely self-powered. It uses a battery to initiate deployment and to power up the plasma contactor, but once current is flowing through the tether, some of the power is tapped off and used to recharge the battery. The battery, in turn, powers the current control electronics, the telemetry, system, and the plasma contactor. The ProSEDS mission will not be designed to allow ground control changes in operation, primarily because of the increase in complexity and cost associated with that option.

The present invention is the use of a small, low-mass deployer/controller package containing a large collecting area, short length, multiline space tether, such as a Hoytape mesh made of aluminum wire, as a "Terminator Tether™" for a constellation spacecraft. The Terminator Tether™ would be deployed when the host spacecraft is no longer working or no longer wanted. The electrodynamic drag from the Terminator Tether™ would rapidly remove the unwanted spacecraft from the constellation and a few weeks later complete the deorbit of the host spacecraft from space by burnup in the upper atmosphere of the Earth. For a Terminator Tether™ to be of maximum usefulness for constellation spacecraft, it would be desirable to minmize the mass and the length of the tether. A lower added mass means more mass for revenue producing transponders, while a shorter tether length means a lower collision cross-section Area-Time Product during deorbit.

FIG. **6** shows a spacecraft **600** having a state vector **602** that causes the spacecraft to move across magnetic field lines **601**. Spacecraft **600** has a conductive tether **603** attached to it by tether deployer and control system **605**. The outer end of tether **603** is attached to tether end mass **607**. Spacecraft **600** also has an attached electron emitter means **613**.

In FIG. **6**, spacecraft **600** has a velocity and direction defined by state vector **602**. As spacecraft **600** moves it causes conductive tether **603** to cut magnetic field lines **601**. As this happens the distal end of the conductive tether collects electrons, shown as arrows **611**, from the space plasma. Electrons **611** move through the conductive tether **603**, are passed through a resistive load, not shown, in control system **605**, and are emitted back into the space plasma as electrons **615** from electron emitter **613**. This produces an electrodynamic drag in the direction shown by arrow **609**.

As is discussed in detail in the electrodynamic drag analysis below, especially in that sections of the analysis titled "Optimization of Tether Angle", the maximum electrodynamic drag of the tether is achieved when the angle, shown as the Greek letter τ in FIG. **6**, is 35.26 degrees, trailing, to the direction of the magnetic field lines. This angle may be maintained by a feedback control circuit that maximizes the electrodynamic drag of the tether system, as is discussed in detail below. The tether structure **603** is a Hoytether having a length of about 2 to 5 kilometers. The use of the 35.26 degree angle to the magnetic field reduces the tether instability, allowing the use of a short tether which will have a small Area-Time-Product. This will minmize the possibility that the tether will impact another space object during its use.

In FIG. **7**, spacecraft **700** has a state vector **702** across magnetic field lines **704**. Conductive tether **706** is attached to spacecraft **700** by tether deployer and control system **708**. The distal end of tether **706** is attached to a tether end mass **710**. Spacecraft **700** has an electron emitter means **714**.

Functionally, electrons **712** flow from the space plasma into conductive tether **706**, are passed through a resistive load in control unit **708** and are emitted into the space plasma by electron emitter **714** as free electrons **716**. As is discussed in detail in the analysis below, this produces an electrodynamic drag on satellite **700**, which causes it to deorbit.

In FIG. **7**, spacecraft **700** is rotating with an angular velocity indicated by arrow **718**. This rotation causes a centrifugal force **720** to place tension on conductive tether **706**. This force **720** places tension on the tether **706** in a direction that counters tether instability. The Tether **706** may be at any angle to the magnetic field while the rotation of the tether-satellite system is producing this useful tension. FIG. **7** shows the angle to be 35.26 degrees, which is optimum, but any angle will work.

FIGS. **8***a* and **8***b* show a powered Terminator Tether™. In FIG. **8***a* the spacecraft **800** has a state vector **802** that causes it and its associated conductive tether **806** to move across magnetic field **804**. Tether **806** has an end mass **808** that is proximate a plasma contactor **810**. Just as was discussed in the embodiment of the present invention shown in FIG. **6** and FIG. **7**, an electric current moves through the circuit formed by the space plasma, the plasma contactor **810**, the conductive tether **806**, and the electron emitter **815**, thence back into the space plasma as electrons **816**. This creates an electrodynamic drag on the tether in the direction shown by arrow **818**.

In FIG. **8***b*, similar numbers indicate similar structures. In FIG. **8***b*, however, an electric power source, not shown, in spacecraft **800** provides electrons to conductive tether **806**. The electrons are emitted from the end **812** of the tether **806**. Electrons are collected to from the space plasma **814** by contactor **815**, thus completing the electric circuit. The result is an electrodynamic force on tether **806**, and therefore on spacecraft **800**, in the direction show by arrow **820**.

Spacecraft **800** in FIG. **8***b* may be rotated as is taught in the embodiment of the invention shown in FIG. **7**, above, to allow the control system to time the application of electric current to the tether **806** to cause force **820** to be exert force in any desired direction along the circle made by the rotating tether. This allows the state vector of the spacecraft to be modified to raise or lower the spacecraft's orbital altitude, or to otherwise usefully change its orbital elements.

FIG. **9** show an embodiment of the present invention wherein the conductive structure of a satellite is used as the tether. In FIG. **9**, a spacecraft **900** which has a state vector

901 and is moving across magnetic field **903**, has two attached solar panels, **902** and **904**. Solar panels **902** and **904** have conductive aluminum frames **908** and **906**, respectively. Frame **906** is electrically and mechanically connected to plasma contactor **912**. Frame **908** is electrically and mechanically connected to electron emitter **910**. As spacecraft **900** moves across magnetic field **903**, electrons from the space plasma **914** flow through contactor **912** and frame **906** to the body of spacecraft **900**, where the current flows through a resistive load, not shown, and then into frame **908** and electron emitter **910** to the space plasma **916**, thus completing the circuit. This creates an electrodynamic drag on spacecraft **900**. If the power output from solar panels **902** and **904** was directed to flow into the frames, **906** and **908**, then the electrodynamic interaction with the space plasma would cause a useful force to be impressed on spacecraft **900**, as was discussed in connection with FIG. **8** above. The spacecraft may be rotated and/or oriented to a desired angle with the magnetic field lines, as was also discussed above.

FIG. **10** shows an embodiment of the present invention wherein the plasma contactor and the resistive load are structurally part of the conductive tether. In FIG. **10** spacecraft **1000** has a state vector **1001** that causes it to pass through magnetic field **1003**. Spacecraft **1000** has a tether control and deployment system **1005** that is connected to a conductive tether **1007** having an end mass **1013**. A contactor portion **1015** of the conductive tether **1007** near end mass **1013** is adapted to contact the space plasma so as to receive or emit electrons. This may be the bare wire strands of a wide Hoytether, or it may be any other plasma contactor means that can be conveniently made electrically part of the Hoytether structure, such as points or metal fuzz. A second resistive portion **1009** of tether structure **1007** is adapted to be a resistive load. This may be done by making a portion of the tether structure from a material, such as nichrome steel wire, that has suitable electrical resistance The large surface area of the Hoytether provides an excellent radiator structure into the three degree Kelvin radiation sink of outer space.

FIG. **11** shows an alternative embodiment of the load resistance/radiator structure of the present invention. In FIG. **11**, a biline Hoytether **1002** is attached to its control and deployer system **1100** on a spacecraft, not shown, and to a tether end mass **1004**. A resistive load, **1006** is woven into the Hoytether structure, whereby this load **1006** and the section of the Hoytether **1002** into which it is woven act together as a thermal radiator allowing photons **1008** to radiate into the thermal sink of outer space.

The following detailed analysis of the present invention will help those skilled in the art to better understand and use the present invention.

Electrodynamic Drag Analysis:

The following is an optimization analysis of the electrodynamic drag produced by a conducting tether deployed from a host spacecraft that it is to deorbit. The force and torque diagram used in the analysis is shown in FIG. **12**, which shows the forces and torques on an upward deployed conducting tether due to the motion of the host spacecraft through the magnetic field of the Earth.

Physical Constants and Assumptions:

The analysis that follows will use the following physical constants and assumptions:

Newton's gravitational constant $G=6.67\times10^{-11}$ $m^3/kg\text{-}s^2$.
Mass of Earth $M_e=5.976\times10^{24}$ kg.
Radius of Earth near equator $R_e=6378$ km.
Assumed host spacecraft altitude h=622 km.
Assumed host spacecraft orbital radius a=7000 km.
Host spacecraft orbital velocity $w=\omega a=[GM_e/a\ ]^{1/2}=7546$ m/s.
Vertical gravity gradient at spacecraft $2\Gamma=2GM_e/a^3=2.32\times10^{-6}\ s^{-2}$.
Centrifugal gradient at spacecraft $\omega^2=GM_e/a^3=1.16\times10^{-6}\ s^{-2}$.
Combined gradient at spacecraft $3\Gamma=3GM_e/a^3=3\omega^2=3.49\times10^{-6}\ s^{-2}$.
Magnetic field of Earth (tilted dipole approximation):
  Magnitude $B_o=35\ \mu T=0.35$ gauss.
  Angle between magnetic pole and spin pole 11.5°.
  Field components at spacecraft with altitude a and angle β between radius vector of spacecraft and magnetic equatorial plane.
  Horizontal $B_H=B_o(R_e/a)^3\cos\beta=26.5\ \mu T\cos\beta$.
  Vertical $B_V=2B_o(R_e/a)^3\sin\beta=52.9\ \mu T\sin\beta$.

Electrodynamic Tether Constants and Assumptions:

The electrodynamic tether is assumed to be made of a conducting metal, and have a length L, density d, resistivity r, and cross-sectional area A that is constant along the length of the tether. If the tether is a single round wire of diameter D, then the cross-sectional area is $A=\pi D^2/4$. Because of the micrometeorite and space debris hazard, however, it is likely the tether will be made up of redundantly interconnected multiple lines whose individual cross-sectional areas add up to A. Given these assumptions, the tether mass is then $m_T=dLA$, while the end-to-end tether resistance is $R_T=rL/A=rdL^2/m_T$.

Specific Conductivity Parameter:

The choice of the metal conductor to be used in a space tether is determined by a combination of low resistivity (high conductivity) and low density, with cost, strength, and melting point as secondary considerations for certain applications. Copper has a resistivity r=17.0 nΩ-m, a density d=8933 kg/m, and a "specific conductivity" of 1/rd=6,585 $m^2/\Omega\text{-}kg$. Aluminum has a resistivity r=27.4 nΩ-m, which is significantly greater than that of copper, but it has a much lower density of d=2700 kg/m^3. As a result, aluminum's "specific conductivity" of 1/rd=13,500 $m^2/\Omega\text{-}kg$ is twice the conductivity per unit mass of copper. Silver, because of its higher density and higher cost, is not competitive as an electrodynamic space tether even though its resistivity of 16.1 nΩ-m is slightly better than that of copper. An alternate candidate material would be beryllium, with a resistivity r=32.5 nΩ-m, density d=1850 kg/m^3, and a "specific conductivity" of 1/rd=16,630 $m^2/\Omega\text{-}kg$, slightly better than that of the much cheaper aluminum. Beryllium also has a higher melting point at 1551 K than aluminum at 933 K, so some of its alloys may be a preferred material for some electrodynamic applications despite its higher materials cost. Unfortunately, despite decades of metallurgical research by the nuclear power industry, highly ductile alloys of beryllium have not been found, so it is difficult to pull beryllium into wire. As a result, because of its high specific conductivity, low cost, and ready availability in ductile wire form, it will be assumed for this analysis that the electrodynamic tether will be made of aluminum wire.

Typical Resistance Values:

To be economically competitive, the mass of the tether needs to be a small fraction of the mass of the host spacecraft it is required to deorbit. Since a typical constellation spacecraft has a mass of about 1000 kg, a typical Terminator Tether™ with a mass that is 2% of the host spacecraft mass would consist of a deployer/controller package with a mass $m_D=10$ kg, containing an aluminum tether with a mass $m_T=10$ kg with a volume of $LA=m_T/d=3.70\times10^{-3}\ m^3$. If this 10 kg of aluminum were formed into a tether with a length of L=2 km and a cross-sectional area of A=1.85 mm^2, then the end-to-end resistance of the tether would be $R=rL/A=rdL^2/m_T=29.6\ \Omega$ A longer tether would have a proportionately smaller cross-sectional area and a higher resistance; for example, a 5 km long tether with the same mass would have a resistance of 185 Ω.

Orbit Inclination Assumption:

In order to make the remainder of the analysis mathematically tractable, it will be assumed that the orbit of the host spacecraft is circular and above the magnetic equator, so that the angle between the radius vector to the spacecraft and the magnetic equatorial plane is β=0 degrees. In this orbit, the velocity v of the spacecraft with respect to the magnetic field is the orbital velocity w less the rotational velocity of the magnetic field at the orbital radius a due to the once per day rotation of the Earth, or:

$$v=W-2\pi a/1\text{day}=(7546-509)\text{m/s}=7037\text{m/s}$$

This equation also shows why electrodynamic drag will not be useful for removing geostationary spacecraft from orbit. At the geostationary orbital radius of a=42,200 km, the relative velocity of the spacecraft and the rotating magnetic field of the Earth is zero.

With this assumption of an orbit above the magnetic equator, the vertical component of the Earth's magnetic field is zero and the horizontal component of the Earth's magnetic field is at right angles to both the local vertical and the direction of motion of the spacecraft. The total magnetic field seen by the host spacecraft and its Terminator Tether™ is then orthogonally horizontal and has the magnitude:

$$B=B_H=B_o[R_e/a]^3 \cos \beta=26.5\mu T.$$

Where for simplicity the subscript H will be dropped for the remainder of this analysis.

This orbit, with its inclination of 11.5° with respect to the spin equator, will not stay in the plane of the magnetic equator, but will vary ±11.5° above and below it as the Earth rotates, causing a variation in magnetic field strength of ±0.5 μT, as well as a slight variation in angle. This 2% variation is negligible in terms of the other uncertainties in this analysis. It is not until the orbital inclination of the host spacecraft orbit reaches 60°, where cos 60°=0.5, that there is a significant drop in the expected magnitude of the electrodynamic drag forces calculated in this analysis.

Electromagnetic Drag Effects in Polar Orbit:

In many medium Earth orbit communication satellite constellations, there are a significant number of spacecraft at high inclinations and in nearly polar orbits. The high inclination spacecraft, with inclinations between 60° and 78.5°, will all have orbits that stay between the magnetic poles. Although the amount of electrodynamic drag will be significantly less than that experienced by spacecraft with orbits at lower inclinations, the direction of the induced electric fields in the tether will always be in the proper direction. If the tether is vertically upward, the outer tip of the tether will be positively charged and the bare wire in the tether will pull electrons out of the surrounding space plasma, while the electron emitter at the host spacecraft end will eject the electrons back into the space plasma to complete the circuit.

For spacecraft in near polar orbits with inclinations between 78.5° and 90°, however, there will be much more drastic variations. First of all, for a spacecraft in a 90° orbit that happens to pass directly over the magnetic poles, there is no horizontal component of the magnetic field when it is passing over one of the magnetic poles, so no voltage is generated in the tether during that part of the orbit, while the horizontal component of the magnetic field near the magnetic equator, while strong in magnitude, is along the spacecraft velocity vector, so no voltage is generated in the tether in that part of the orbit either. As a result, there will be negligible electrodynamic drag experienced by the host spacecraft during that particular orbit. The Earth is rotating, however, and the magnetic pole is rotating with it. A few orbits later, the host spacecraft will be passing over the Earth's spin pole at a point where the horizontal component of the magnetic field is exactly at right angles to the direction of motion of the spacecraft, so the full voltage is generated in the tether. When passing over the spin pole, the spacecraft radius vector is at an angle of 11.5° from the magnetic pole and at an angle β=78.5° away from the magnetic equatorial plane. With these assumptions, the magnitude of the horizontal component of the magnetic field at the Earth's spin pole is a respectable:

$$B=B_o[R_e/a]^3 \cos 78.5°=5.28\mu T$$

or 20% of the maximum value experiences by spacecraft orbiting above the magnetic equator. This value will drop slightly as the orbit continues, then build back up as the spacecraft passes over the opposite pole. As the Earth continues to rotate, bringing the magnetic pole again under the orbit, and the interaction of the tether with the horizontal component of the magnetic field again drops to zero. The average coupling of a tether to the Earth's magnetic field over all polar trajectories has been calculated to be 12.3 percent.

A problem experienced only by spacecraft with orbit inclinations greater than 78.5°:

The spacecraft will no longer be traveling from west to east with respect to the magnetic field axis, but will have a retrograde motion as it moves through the magnetic field. As a result, the voltage generated by the motion of the tether through the Earth's magnetic field will switch direction. The outer tip will be negatively charged and will attempt to collect ions, which is a much less efficient process than collecting electrons.

There are a number of solutions to this problem. The first is to increase the mass and length of the tether supplied to a spacecraft assigned to a polar orbit, so that higher voltages, currents, and drag are generated during the limited times the spacecraft is passing over the poles in the right direction. The second is to supply a tether with plasma contactors at both ends that can emit electrons from either end, allowing the current to flow either way, depending upon which direction the spacecraft is passing around the magnetic pole. The third is to utilize the first number of passes to torque the orbit of the spacecraft until the orbit inclination has been shifted below 60°, then turn on the electrodynamic drag full time to deorbit the spacecraft from this more favorable orbit inclination. This orbit torqueing maneuver is accomplished by activating the electrodynamic drag mechanism only when the magnetic field orientation is such that a strong out-of-plane component of force is created. This orbit torqueing maneuver can be augmented by switching to a propulsion mode, where power saved in the batteries during the drag force mode is pumped back into the tether when the magnetic field is in the opposite direction, applying electrodynamic propulsion to torque the orbit even further. With the tether at a large and stable trailing orientation, the coupling to the magnetic field can be significant to a vertical tether.

In summary, spacecraft in near-polar orbits might take longer to bring down, and might have to utilize specially designed Terminator Tethers™ that might cost and mass more than the simpler Terminator Tethers™ usable in lower inclination orbits, but a Terminator Tether™ can still remove a spacecraft from a polar orbit when desired.

The inventors now present a detailed analysis and optimization of a typical Terminator Tether™ attached to a typical host spacecraft in a typical low inclination orbit.

Electromotive Generation of Voltage and Current in the Tether:

When an object is moved at a velocity v through a magnetic field B, an electric field is generated in the frame of reference of the moving object given by:

$$E=v\times B=vB$$

where the magnetic field B of the Earth, being mostly tangent to the Earth's surface in the north-south direction, is at right angles to the velocity vector v of the spacecraft, assumed to be orbiting in a generally west-east direction. The direction of the electric field E will be at right angles to both v and B, or along the local vertical. It should be noted that this electric field exists in the moving frame of reference because a moving magnetic field creates an electric field. No object actually has to be there, but if it is, then the relative motion of the magnetic field of the Earth will not only apply magnetic forces to whatever material the object is made out of, but electric forces too.

Note also that the velocity used in this equation is the relative velocity between the object and the magnetic field. Because the Earth's magnetic field rotates with the Earth, the motion of the magnetic field must be subtracted from the orbital velocity of the object to obtain the relative velocity.

Voltage generation in a Conducting Tether:

If the moving object is a long conducting wire of length L, the electric field E, generated in the wire produces a voltage V between the opposite ends of the wire given by:

$$V=E\bullet L=EL\cos\tau=vBL\cos\tau$$

where τ is the angle between the length vector L of the tether and the electric field vector E, assumed to be in the vertical direction at right angles to the velocity vector v in the plane of FIG. **12** and the magnetic field vector B out of the plane of FIG. **12**. A typical value for the voltage level built up in a vertically oriented tether with length L=5 km and $\tau=0$, moving at a velocity of v=7037 m/s through the Earth's horizontal magnetic field of strength 26.5 μT, is 932 V, or 0.186 V/m. Spacecraft in higher inclination orbits would experience somewhat smaller electric fields. For calibration, the voltage measured between the ends of the 19.5 km long TSS-1R tether was 3500 V, or 0.175 V/m.

Contacting the Space Plasma:

Although a voltage will build up between the ends of the conducting tether, no current will flow unless the circuit is completed. The circuit cannot be completed with another wire, for it too will have a similar voltage generated in it by the moving magnetic field. Fortunately, empty space is not empty, and in near-Earth regions not too distant from the Earth's atmosphere there exists highly electrically conductive space plasma, kept partially ionized by radiation from the Sun. The electron and ion density varies from the dark to light side, with altitude, with season, with sunspot cycle, with contamination level, etc. but it typically varies from 10^{11} to 10^{13} electrons/cc.

Fortunately, it has been found that if a bare conductive surface such as the spherical Italian spacecraft in the TSS-1R experiment, or the long bare wire to be used in the ProSEDS experiment, or a multiline conducting wire Hoytape mesh, is charged to a few hundred volts, the bare conductor will readily pull electrons out of the space plasma. So all that is needed to complete a connection to the space plasma at the positively charged (upper) end of the tether is a sufficient large area of uninsulated conductor.

It is more difficult to eject electrons from a wire or to collect positive ions from the space plasma. Although it is conceivable that a very large area at the other end of the tether could collect enough ions to complete the circuit as was demonstrated during the TSS-1R mission when the Shuttle orbiter was found to be an adequate plasma contactor for over an amp of current, the present method chosen is to use an electron emitter of some sort, either a hot cathode, a plasma cathode or contactor, a field-emission device, or something similar. Once provisions have been made at both ends of the tether to allow the flow of electrons out of one end of the tether and into the other end, and the altitude of the host spacecraft is not too high, then there will be sufficient conductivity in the space plasma surrounding the host spacecraft to allow current to flow through the tether.

Current Flow In The Tether:

The amount of current I flowing through the tether depends upon the total resistance R in the circuit. This resistance will consist of three components, the effective resistance of the plasma, the resistance of the tether, calculated earlier as $R_T=rL/A=rdL^2/m_T$, and a control resistor R_C, which will be varied as needed to optimize the Terminator Tether™ performance. There will also need to be a parasitic load on the current in the form of a charging device to charge some batteries. The batteries in turn will be used to power the control and communication circuits, and drive the electron emitting devices at the negative end of the tether. A well-designed Terminator Tether™ will thus be completely self-powered, except for an initial charge in the batteries to provide electrical power for the deployment and startup procedure. For simplicity of analysis it will be assumed that this battery charging load, which absorbs power like a resistor, but which stores it and uses it later instead of dissipating it immediately as heat, is included in the control resistor R_c. Normally, this load would act as a "base resistance" below which the control resistor could not be lowered, but since the charging circuit can be turned off when desired, and operations continued without interruption using the power stored in the batteries, it will be assumed that the control resistor can be taken to near zero value in those circumstances where the space plasma conductivity is low, or the magnetic field is in the wrong orientation and the voltage being generated in the tether is not large. Under these circumstances, lowering the control resistor to near zero allows a much higher current to flow for a given generated voltage, thus increasing the power being dissipated in the tether and maintaining a high level of electrodynamic drag on the host spacecraft.

A properly designed Terminator Tether™ will have plenty of bare metal area for electron collection at the positive end, while the electron emitters at the other end are efficient in terms of emitting large electron currents at low voltages and therefore low powers, while at the same time the mass, length, and area of the conducting tether have been made such that the resistance of the tether is moderately high. Under these conditions, the effective resistance of the space plasma will be much less than the design resistance of the tether. To make the mathematics more tractable it will be assumed that the plasma resistance is can be neglected and that the voltage available to drive the tether is the full voltage generated by the moving magnetic field. Although voltage will be needed to power the electron emitter, whatever its form, it will be assumed that the required voltage and power will be supplied by the batteries. The batteries in turn will be powered by the battery charging circuit, whose effective resistance is included in the control resistor (except for those short periods when the plasma or magnetic field interactions are weak).

Given all these assumptions concerning the total resistance in the circuit, the current I flowing through the tether is then given by:

$$I=V/R=vBL \cos \tau/(R_C+rdL^2/m_T)$$

A typical current level can be estimated by assuming that the control resistance R_C should be about the same size as the tether resistance R_T, which for an aluminum tether with resistivity r=27.4 nΩ-m, density d=2700 kg/m^3, length L=5 km, and mass m_T=10 kg, is R_T=185 Ω, giving a total resistance, including the control resistance R_C, of R=370 Ω. Thus the current flowing though a tether of length L=5 km, generating a voltage of V=932 V, would be I=V/R=2.52 A.

Power Dissipation Analysis:

The induced current I flowing through the tether resistance R_T will generate heat in the wire, which will be radiated into space and lost. That radiated energy must come from somewhere. There is also power taken out of the current flow by the battery charging circuit and ultimately dissipated. In order to conform to energy conservation laws, this power and energy must come from a decrease in the total kinetic and potential energy of the host spacecraft, causing it to deorbit. Set forth below is a detailed, optimized force analysis which will calculate the drag force on the host spacecraft. In this section, by using some simplifying assumptions, it is possible to produce a general argument based on energy conservation laws, which will illustrate the broad scope of the present invention.

For this energy conservation analysis it will be assumed that the plasma resistance is small compared with the tether resistance, and that the control resistance is equal to the tether resistance:

$$R_C=R_T=rL/A=rdL^2/m_T$$

Assume that the ballast mass at the end of the tether is a large piece of the defunct host spacecraft, such as a solar panel, antenna, or battery pack, so that the gradient force is large and the tether is always oriented along the vertical so that the angle τ=0°. Also assume the spacecraft is in an orbit above the magnetic equator and thus is moving at right angles to the Earth's magnetic field, so that the angle between the orbital velocity vector and the magnetic field vector is 90°, and the electric field vector is in the vertical direction, aligned perfectly with the tether length vector. Under these ideal conditions, the voltage between the ends of the tether is given by:

$$V=E\bullet L=(v\times B)\bullet L=vBL$$

The current in the tether is then just:

$$I=V/(R_C+R_T)=vBm_T/2rdL$$

The power dissipated as ohmic heating in the tether is then given by:

$$P=IV=(vB)^2m_T/2rd$$

This equation shows the interesting result that the power dissipated does not depend upon the length or the area of the tether (within reason), but only the mass of the tether m_T, the resistivity r and density d of the tether material, and the velocity v of the conducting tether through the Earth's magnetic field B.

For a typical aluminum tether of mass m_T=10 kg, the power dissipated is an impressive P=2350 W. Even if only a small fraction of this power is dissipated in a real system, the kinetic energy of the host spacecraft would suffer a significant energy loss with time.

Orbital Energy Extraction Analysis:

The total energy U of a spacecraft of mass M in orbit around the Earth consists of two components, its positive orbital kinetic energy and its negative gravitational potential energy. If a circular orbit of altitude h and radius a=R_e+h is assumed, then the total energy of the spacecraft moving at a velocity w=$(GM_e/a)^{1/2}$ is:

$$U=+Mw^2/2-GMM_e/a=-GMM_e/2a$$

The system will be dissipating energy in the tether to decrease the energy U of the unwanted spacecraft from its relatively low negative value in a high Earth orbit to a greater negative value in a lower Earth orbit. To give a specific example, an unwanted spacecraft with a mass of M=1000 kg (a metric ton) in an orbit with an altitude of h=1000 km and a high orbital radius of a(H)=R_e+h=7371 km, then its (negative) total energy is:

$$U(H)=-GMM_e/2a(H)=-27.1\text{ GJ or }-27.1\text{ MJ/kg}$$

It is desired to lower the spacecraft to an orbit with an altitude just outside the atmosphere of h=200 km or a lower orbital radius of a(L)=R_e+h=6571 km, where atmospheric drag will finish the job. The total (negative) energy of a metric ton spacecraft at 200 km altitude is:

$$U(L)=-GMM_e/2a(L)=-30.4\text{ GJ}$$

Thus, the amount of energy needed to be dissipated in order to move the 1000 kg spacecraft from a 1000 km orbit with a total energy U(H)=−27.1 GJ down to a 200 km orbit with energy U(L)=−30.4 CJ is:

$$dU=U(H)-U(L)=3.3\text{ GJ.}$$

The decay time T of a metric ton spacecraft moving from a 1000 km altitude orbit to a 200 km altitude orbit with an energy difference of dU=3.3 GJ, when its energy is being dissipated at a power of P=2350 W by a typical aluminum tether massing just 10 kg or 1% the mass of the spacecraft, is found to be T=dU/P=16 days.

The actual decay time will be longer than this. If the electrodynamic drag force is very large, and becomes larger than the gravity gradient forces pulling on the ends of the tether, then the tether will tend to align itself along the magnetic field lines instead of across them, and the drag force will decrease because of the small angle between the conductor length and the magnetic field lines. The tether current will need to be controlled until the angle of the tether settles into an angle determined by the balance between these two forces. These electrodynamic torque vs. gradient torque dynamic interaction effects are discussed in the next section.

Force and Torque Balance Analysis:

It is now possible to calculate the forces and torques on the tether. It is also possible to use the fact that the electrodynamic and gravity forces and torques on the tether must balance each other out to calculate and teach some optimum values for some of the Terminator Tether™ parameters.

Electrodynamic Force and Torque:

As discussed above, both theory and experiment show that: provided the conducting tether is moved rapidly through the Earth's magnetic field in order to generate a voltage across it, and provided good contact is made with the space plasma, the conducting tether will have a current flowing through it. When a wire (moving or not) carrying a current I is embedded in a magnetic field B, there will be an electrodynatnic force F_E generated on each element of the wire. The electrodynamic force will be at right angles to the magnetic field vector and the length vector of the wire element, with a magnitude given by:

$$F_E=J\times B=ILB$$

where B is the horizontal component of the magnetic field, which is perpendicular to the assumed magnetic equator orbit plane, while the tether length vector L is assumed to lie in the orbit plane. If, as will be the case most of the time, the electron current is leaving the space plasma and entering the tether along the length of the tether, then IL needs to be replaced with an integral of the current along the length of the tether.

Note that the electrodynamic force amplitude or direction is not directly dependent on the motion of the tether through the Earth's magnetic field. The electrodynamic force would be the same if the tether were not moving and the current was being supplied by a battery. Since the current I is a function of the orbital velocity, however, there is a secondary dependence of the electrodynamic force magnitude on the orbital velocity, but not its direction. The electrodynamic force is always at right angles to the conductor, and stays at right angles to the conductor as the angle τ varies, as shown in FIG. **12**.

Assuming that the electrodynamic drag force is applied uniformly along the length of the tether, it is possible to make the simplifying assumption that the integrated force is effectively applied at right angles to the center of mass of the tether at the point L/2 as shown in FIG. **12**. The electrodynamic torque on the tether is then:

$$T_E=F_EL/2=IBL^2/2=vB^2L^3\cos\tau/2(R_C+rdL^2/m_T)=vB^2L^3\cos\tau/2R$$

Gravity Gradient Forces and Torques:

When a tether and its ballast end mass are deployed from a host spacecraft, the gravity gradient force field of the Earth, combined with the orbital centrifugal gradient force field, will cause the tether to deploy either up or down from the host spacecraft. The direction desired depends on which end of the tether is connected to the electron emitter. Normally, the electron emitter will be on the end attached to the host spacecraft, in which case the desired direction of deployment will be upward so that the induced voltage in the tether will produce an excess of electrons at the electron emitter end of the tether. The desired upward direction is chosen by having the deployer eject the ballast mass in the upward direction. Once the ballast mass has been started in that direction, the centrifugal force due to the orbital motion of the ballast mass will cause the ballast mass to continue to accelerate in the upward direction until it is brought to a halt by the full deployment of the tether.

If there were no electrodynamic or atmospheric drag, the equilibrium direction of the tether would be exactly along the vertical, since the combined gradient field is a maximum in that direction. Because a significant amount of electromagnetic drag is expected, the actual angle of the tether with respect to the local vertical will be at some angle τ, lagging behind the spacecraft motion in the plane of the orbit, as shown in FIG. **12**. In the following analysis shows that there is an optimum angle for τ that produces the largest electrodynamic drag force on the host spacecraft without producing tether instability.

The combined vertical gravity gradient and centrifugal gradient field 3Γ acting on the ballast mass m_B at the end of the tether of length L will produce a gradient force F_{GB} given by:

$$F_{GB}=-3_{\Gamma}m_BL\cos\tau$$

The strength of the force depends not only on the ballast mass m_B and the strength of the gradient field 3Γ, but also the component along the radial direction of the distance of the ballast mass from the center of mass, which is L cos τ. This force acts in the vertical direction along the radius vector leading from the ballast mass away from the center of the Earth. The amount of gradient force is not large. For a ballast mass of m_B=10 kg and a tether length of L=5 km, the gradient force is about 0.175 N.

As can be seen in FIG. **12**, the component of this gradient force that is at right angles to the tether, given by $F_{GB}\sin\tau$, will produce a torque T_{GB} on the tether that tends to restore the tether toward the vertical, lessening the angle τ.

$$T_{GB}=LF_{GB}\sin\tau=-3_{\Gamma}m_BL^2\sin\tau\cos\tau$$

The tether mass m_T also contributes to the gradient force and torque. If it is assumed that the tether has a uniform cross section, then it is possible to replace the distributed mass of the tether with an equivalent point mass m_T placed at the center of mass of the tether, which is the point L/2 along the tether, and a distance L/2 cos τ in the radial direction. The gradient force due to the tether mass is then:

$$F_{GT}=-3/2_{\Gamma}m_TL\cos\tau$$

While the gradient torque is:

$$T_{GT}=-L/2F_{GT}\sin\tau=-3/4_{\Gamma}m_BL^2\sin\tau\cos\tau$$

The total gradient torque attempting to restore the tether to its vertical orientation is then:

$$T_G=T_{GB}+T_{GT}=3_{\Gamma}(m_B+m_T/4)L^2\sin\tau\cos\tau$$

It is important to notice the variation of the total gravity gradient torque as the tether angle τ is varied. Since the gradient force is always in the radial or vertical direction, there is no torque on the tether when the tether is vertical, as is the case when there are no aerodynamic or electromagnetic drag forces. Once the drag forces become important and start to apply torque to the tether, increasing the tether angle τ, those drag torques causing an increase in tether angle τ will be opposed by a rising gradient torque which will attempt to decrease the tether angle. For small tether angles, the gradient torque increases first linearly with τ, then as sin τ, since cos τ is near unity for small τ, then at τ=45° the gradient torque reaches its maximum, where sin τ=cos τ=0.707 and sin τ cos τ=0.50. When this angle is reached, the tether is a point of catastrophic instability, for if there is a further increase in the electrodynamic drag force, causing the angle τ to become greater than 45°, the gradient torque, instead of growing stronger to counteract the increased drag torque, will become weaker and the tether angle will go rapidly to 90°.

To restore control to the tether angle if the instability occurs, it will be necessary to turn off the electrodynamic drag forces by shutting off the current flow through the tether. The τ=90° position for the tether and ballast mass is a gravitationally unstable orientation. After a time, slight fluctuations in the gravity field will allow the gradient force to slowly take over and restore the tether to the vertical orientation, which, unless it can be controlled in some way, is equally likely to be up or down. It would therefore be desirable to maintain control of the tether angle so as to avoid the tether angle getting into the region of instability.

Torque Balance on the Tether:

The angle τ of the tether is determined by a balance between the electrodynamic torque T_E attempting to increase the angle τ and the gradient torque T_G attempting to decrease the angle τ. Balance is achieved when the two torques are equal:

$$T_E=T_G=T_{GB}+T_{GT}$$

or:

$$vB^2L^3 \cos \tau/2R=3_{\Gamma}(m_B+m_T/4)L^2 \sin \tau \cos \tau$$

Simplifing and rearranging yields an equation giving us the angle τ at which torque balance occurs:

$$\tau=\arcsin[vB^2L/6_{\Gamma}(m_B+m_T/4)R]$$

The physical interpretation of this equation is that the maximum electrodynamic force that can be sustained on the tether is limited by the gradient force on the tether. It is possible to increase the electrodynamic force by decreasing the total resistance R of the tether (for example, by using a tether with a larger mass and a lower tether resistance), but if the tether resistance is too low, then the quantity in brackets becomes greater than unity and this equation has no solution, indicating that the tether has reached an angle where instability sets in.

To maintain control of the tether angle, it will be necessary to vary the control resistance of the tether to compensate for variations in magnetic field strength and direction, plasma density (which affects the plasma resistance), and other factors, and thereby maintain the tether at an intermediate angle where both the electrodynamic and gradient forces are at an appreciable level and balance each other.

Optimization of Tether Angle:

At first glance, it might seem that the optimum angle for the tether would be 45°, since at that angle the gradient torque is largest and therefore can counteract a larger electrodynamic drag force. However, since the 45° angle is the point where instability sets in, it is desirable to set the tether angle at some value below 45°. The optimum angle is that which maximizes the horizontal or drag component of the electrodynamic force. This optimum angle τ is derived from the equation for the horizontal component of the electrodynamic force, or the electrodynamic drag force, since it opposes the host spacecraft motion:

$$F_{ED}=F_E \cos \tau=6[m_B+m_T/4]\Gamma L \sin \tau \cos^2\tau$$

By setting the partial derivative$\square F_{ED}/\square\tau=0$ and solving, it is possible to calculate that the optimum angle for the tether that gives the maximum electrodynamic drag force F_{ED}, while still keeping the tether torques balanced and under control, is $\tau=\arctan(0.707)=35.26°$. With this angle selected and maintained by varying the control resistor R_C to compensate for variations in plasma contact resistance and variatins in the strength and direction of the Earth's magnetic field B seen at the spacecraft, the tether experience the maximum stable value for the electromagnetic drag force of:

$$F_{ED}(\max,\tau=35.26°)=2.31[m_B+m_T/4]\Gamma L$$

There are many ways to provide the sensing information needed to provide the feedback signals to the control resistor, but the simplest is to merely measure the drag acceleration on the host spacecraft with a set of accelerometers, and maximiize the force in the direction opposite to the host spacecraft motion. Another method would be to measure the current in the tether, and knowing the tether resistance and the amount of control resistance, calculate the power being extracted and maximize that value. Alternate methods would be to use an optical position sensor or GPS receivers at both ends of the tether to measure the angle of the tether or the position of the ballast mass with respect to the host spacecraft.

To make an estimate of the magnitude of drag force attainable, assume that a typical aluminum tether with a length of L=5 km, a tether mass of m_T=10 kg and a ballast mass of m_B=10 kg is being used. Then the maximum gradient force limited electrodynamic drag force is F_{ED}=0.167 N. The power being dissipated in the canoical tether at the angle τ, with $R_C=R_T=rdL^2/m_T=185\ \Omega$ is given by:

$$P=IR=[vBL \cos \tau]^2(R_C+rdL^2m_T)$$

which for an angle τ=35.26° results in a power dissipation of 1570 W. Thus, the thrust level per unit power obtained by the typical Terminator Tether™ would be about 0.106 N/kW. This number is comparable to the value of 0.148 N/kW estimated for the much heavier and longer TSS-1R tether.

Reconciliation of Energy and Force Analysis:

By a force and torque balance analysis an optimum angle for the tether has been discovered at which it is possible to obtain a maximum in the drag component of the electrodynamic force. There is also an additional component of the electrodynamic force, the component in the vertical direction, $F_{EV=FE} \sin(\tau)$, which is downward for an upwardly deployed tether. This component of force combines with the gravity force of the Earth to effectively allow the host spacecraft to orbit a little faster than normal for that orbital altitude. It does not contribute to the deorbiting of the host spacecraft. But since this vertical component of force is created by current running through the tether, and that current is creating heat and dissipating energy as it passes through the tether, there might be some concern that the force and torque balance analysis above does not conform to the law of conservation of energy.

The inventors will now show, in a very general manner, that no matter what the tether angle, the electrical power being dissipated in the tether is exactly equal to the power being lost by the slowing of the host spacecraft.

The "deorbit power" P_D that must be removed from a spacecraft moving at a velocity v when that motion is opposed by a drag force F_{ED} is:

$$P_D=v\bullet F_{ED}=vF_E \cos \tau=vILB \cos \tau=I[vBL \cos \tau]$$

But, since the voltage V induced across the tether of length L and tilt angle τ moving at a velocity v through a horizontal magnetic field B is

$$V=E\bullet L=EL \cos \tau=vBL \cos \tau$$

Therefore:

$$P_D=IV=P_E$$

And the deorbit power P_D extracted from the slowing of the spacecraft by the drag component of the electrodynamic force is always exactly equal to the electrical power P_E being dissipated as heat in the tether circuits, independent of the tether angle.

Optimization of Tether Mass Distribution:

If it is not possible to use a piece of the host spacecraft as ballast mass, then the mass of the ballast must be included in the Terminator Tether™ mass. It would be desirable to minimize the total Terminator Tether™ mass, since every kilogram saved means that another kilogram's worth of revenue-producing transponders can be added to the working payload of the host spacecraft. Given a total mass for the Terminator Tether™ and the mass of the deployer/controller unit, it is possible to optimize the mass distribution between the ballast mass and the tether mass to obtain a minimum total Terminator Tether™ mass. A well-designed Terminator Tether™ will also have most, if not all, of the deployer mass incorporated into the ballast mass.

Assume that the total Terminator Tether™ mass consists of the deployer/controller mass m_D, the tether mass m_T, and the ballast mass m_B, with $m_B > m_T > m_D$. Of the three mass components in the Terminator Tether™, two of them affect the electrodynamic drag performance. If it is assumed that the ballast mass is a factor X larger than the tether mass, or $m_B = Xm_T$, the maxinum drag force that can be obtained is now:

$$F_{ED}(\max)=2.31m_T(X+0.25)\Gamma L$$

If it is assumed that the control resistor has been adjusted so that this maximum value for the electrodynamic drag force is maintained as the motion of the spacecraft along its orbit moves the spacecraft into regions with different magnetic field strengths and plasma densities, then for the angle $\tau=35.26°$, $\cos(35.26°)=0.817$, and the maximum electrodynamic drag force in terms of the electrodynamic parameters will be:

$$0.817vB^2L^2/(R_C+rdL^2/m_T)=F_{ED}(\max)=2.31m_T(X+0.25)\Gamma L$$

or canceling out terms and rearranging:

$$(X+0.25)=0.353vB^2L/\Gamma m_T(R_C+rdL^2/m_T)$$

It is now possible to make the further simplifying assumption that to maintain control of the tether, the control resistor needs to be roughly the same size as the tether resistance or $R_C \approx rdL^2/m_T$. Using this assumption produces an equation for the optimum value for the ratio X of the ballast mass to the tether mass:

$$X = vB^2/\Gamma rdL - 0.25 = -0.25 + \frac{vB^2}{12\sin\gamma\cos\gamma\Gamma rdL}$$
$$= -0.25\frac{\Lambda}{L}$$

Where Λ is the "effective electrodynamic length".

If the tether is short, so that L is small, this equation indicates that the ballast mass must be increased to increase the gradient force, since it is the gradient force which determines the upper limit to the amount of electrodynamic drag force that can be generated without losing control of the tether.

Assuming the usual values for the Earth's magnetic and gradient fields, the velocity of the host spacecraft through the magnetic field, and the resistivity and density of aluminum, this equation becomes:

$$X=10.9/L-0.25$$

where L is in kilometers.

For a typical tether of length L=5 km, X=1.93. If the total mass for the Terminator Tether™ is 20 kg, or 2% of the host spacecraft mass, and the mass of the deployer is 10 kg, then the remaining 10 kg should be distributed so that 6.59 kg is in the ballast mass while 3.41 kg is in tether mass.

Optimization of Tether Length:

If the ballast mass consists of a piece of the host spacecraft, then the mass of the ballast does not have to be considered in the optimization of the Tether system. In this case, the optimum distribution of the Tether mass is to put as much mass as possible into the tether, with a minimum in the deployer/controller package. Under this assumption, the optimum length of the tether is determined by the desire to keep the Area-Time Product of the host spacecraft plus Tether to a minimum.

In a long tether of length L, the effective "collision" cross-sectional area is not the area of the tether, but the larger area produced by multiplying the length of the tether by sum of the width of the tether plus the width of the "target" spacecraft that would be damaged by the tether hitting it. Thus, to decrease the "area" portion of the Area-Time Product, it would be desirable to shorten the tether, making it thicker, and perhaps slightly wider, at the same time, thus keeping the same tether mass and electron collection area, and thereby maintaining the drag force and keeping the decay time at the same level. The result will be a shortening of the Area-Time Product.

If the tether is shortened too much, however, the smaller voltage generated across the shorter tether will leave less voltage margin for the operation of the plasma contactors and the battery recharge system. Since the maximum voltage that can be generated is about 200 V/km, and typical plasma contactors and battery, chargers take 25–100 V to operate, a minimum length for a Tether would be roughly 2 km. Host spacecraft operating in polar orbits, where the conditions for voltage generation are not as good, may require a tether length of 5 km or more.

Since many watts of heat power will be dissipated in the control resistor, means must usually be provided to radiate the heat away into space. One low-mass method of accomplishing this is to make the control resistor (or resistors) in the form of a long, electrically insulated, high resistivity wire similar to those used in electric blankets, but designed to operate at a higher temperature, and during the fabrication of the electrodynamic tether, incorporate the high resistivity wire (or wires) into the end of the tether closest to the control circuit by weaving or braiding the insulated high resistivity wire in with the uninsulated aluminum wire of the electrodynamic tether. The surface area of a long wire is very large, so in this extended wire configuration, exposed to the space all around it, the hot control resistor wire can self-radiate its power into cold space without incurring the mass penalty of a separate radiating surface.

Implementation:

The basic optimum structure for a Terminator Tether™ would be one of the many types of Hoytethers. A multiline (6–12 primary line) Hoytape™ will provide the largest contact area with the plasma, since both sides of the tape would be able to pass current to the plasma. If the spacing between the primary lines is chosen to be larger than twice the Debye length of the plasma, then the effective current collection area per unit length of the Hoytape is proportional to the width of the Hoytape mesh, not the diameter of the wires in the mesh. Thus, a Hoytape not only provides an assured longer life for the Terminator Tether™, but very short lengths will also provide a very large current collection areas.

The deployer for the tether can deploy the Terminator Tether™ either down or up or both. The deployer can stay attached to the spacecraft as was done in the SEDS missions, or perhaps a better alternative would have the deployer ejected from the spacecraft, with one end of the tether still attached to the spacecraft, reeling out tether as it leaves. The empty deployer would then act as a ballast mass at the end of the Terminator Tether™, which would improve the performance.

The "Remora Remover"

In addition to attaching Terminator Tethers to spacecraft before launch, it is possible to consider a missile-like application. This "Remora Remover" missile would use a Terminator Tether™ carried by a seeker missile similar to the small "hit-to-kill" missiles developed by the Space Defense Initiative Office that deployed a net loaded with oriented metal rods. The Remora Remover missile would hunt down a spacecraft that needs to be removed from space, but instead of hitting the spacecraft, the missile would be programmed to rendezvous with the spacecraft and attach itself to the host spacecraft using a hooked net, harpoon, or adhesive "sucker". The Remora Remover missile would then deploy the Terminator Tether™, which would bring down both the spacecraft and the missile.

The inventors have presented two analyses, backed up with data obtained from space flight experiments. One analysis was based on generalized energy conservation laws while the other analysis use force and torque balance arguments. Both analyses show that optimized aluminum wire tethers 2 to 5 km in length and massing just 1% to 5% of the mass of the host spacecraft can deorbit the host spacecraft in the order of a few weeks, thus mitigating the long-lived orbital debris hazard created by a constellation spacecraft after their end-of-life.

Power Augmented Operation of the Tether:

It is well known in the aerospace literature that if a spacecraft has a power supply and a conducting tether hanging from it, and the spacecraft is in an orbit or trajectory that takes it near a body, such as the Earth or Jupiter, which has both a magnetic field and a region of moderate density electron-ion plasma, and electrical current is pumped from the power supply through the tether and back through the plasma, that the current flowing through the tether will interact with the magnetic field of the body, producing forces on the tether and thence on the spacecraft, changing its orbit or trajectory. These forces can be used to increase or decrease the spacecraft altitude and/or inclination. The amount of altitude or inclination change is proportional to the ratio of the power available from the power supply divided by the mass of the spacecraft and varies with altitude. The unpowered Terminator Tether™ is a simple example of such a system, in which only altitude decrease is possible, although small amounts of both inclination increase or decrease are possible at the same time. Some typical references to this prior art are: Les Johnson, "Propulsive Small Expendable Deployer System Mission (ProSEDS)", Proceedings of the OAST 8th Advanced Propulsion Workshop, JPL, Pasadena, Calif., 20–22 May 1997; and Les Johnson, Joe Carroll, Robert D. Estes, Enrico Lorenzini, Brian Gilchrist, Manuel Martinez-Sanchez, Juan Sanmartin, and Irwin Vas, "Electrodynamic Tethers for Reboost of the International Space Station and Spacecraft Propulsion, AIAA Paper 1996.

Since the Terminator Tether™ of the invention is normally associated with a satellite that has a power supply in the form of a solar panel combined with a storage battery, the Terminator Tether™ can be operated in the "powered propulsion" mode if desired. Such a mode would be useful when attempting to avoid a collision between the Terminator Tether™ and another spacecraft. In such cases, the power from the solar panel, augmented by the power stored in the battery, can be used to temporarily operate the Terminator Tether™ in the propulsion mode instead of the normal drag mode. This would enable to Terminator Tether™ to raise its altitude and/or increase or decrease its inclination to avoid a collision.

In addition, the Terminator Tether™ could be operated in the "power-augmented drag" mode. In this mode, there will be a current induced in the tether by the motion of the spacecraft through the magnetic field, which will cause a drag force on the spacecraft. This induced current would then be augmented by additional current in the same direction generated by the solar panel and the storage batteries in the power supply. The increased current will cause an increased drag force. If there is a significant vertical component of the magnetic field there will also be an increased force normal to the orbital plane, which will cause an increased rate of change of inclination.

When the Terminator Tether™ is operated in the "powered propulsion" or "power-augmented drag" mode, the conditions for stable optimum operation will be essentially the same as in the "unpowered drag" mode. The reason the conditions stay the same is that the maximum tether force than can be effectively utilized is limited by the strength of the gradient force, which has not changed, since the length and mass of the tether and the ballast mass has not changed for these different modes of operation. As a result the optimum operational angle for the tether is still 35.26 degrees, and the optimum ratio of ballast mass to tether mass for different tether lengths and tether material is still determined by the analysis of the invention as is set forth above.

In the analysis given above, where the optimum tether angle was found, the tether was assumed to be in the plane of the orbit and lagging behind the spacecraft motion, since in the "unpowered drag" mode, that would be its normal position, although if the component of the magnetic field along the orbit were high, there would be some tilt of the tether out of the plane of the orbit, producing a force tending to change the inclination of the orbit. In either of the "powered" modes of operation of the Terminator Tether™, however, the tilt of the tether could be forward toward the motion of the spacecraft or strongly tilted to one side or the other of the orbital plane. In all these cases, if the electrodynamic force is allowed to become too large, and the tether angle exceeds 45 degrees, the restoring force of the gradient field will drop off and the tether angle will go unstable. As a result, the optimum angle for the tether that will give the maximum stable force, whether it is a maximum drag force, and maximum propulsion force, or a maximum inclination change force, will be 35.26 degrees, and the optimum ratio for the mass of the tether versus the mass of the ballast will be the same as for the unpowered drag case.

The analysis that gave the optimum angle as 35.26 degrees assumed that the tether would be oriented like a rigid rotor. In reality, variations in forces along the tether will probably cause the tether to hang in a slightly curved shape, where the optimum angle may not be exactly 35.26 degrees from the local vertical. The optimum angle of the tether as it leaves the host spacecraft in such a case is where the drag force is largest.

The present invention is discussed in this disclosure in terms of its space applications as a Terminator Tether™ useful for the removal of unwanted satellites. It should be understood, however, that the present invention is useful in any application where a space object can use a conductive tether to produce electrodynamic force through interaction with a magnetic field. The invention, therefore, should be limited not by this description, but only by the appended claims and their equivalents.

What is claimed is:

1. A method of operating a conductive tether attached to a space object, said conductive tether being oriented at an angle to magnetic field lines of an external magnetic field and said conductive tether moving across the magnetic field lines of said external magnetic field at said angle to producing an electric current in said conducting tether and a resulting electrodynamic force acting on said tether and said attached space object comprising,

Controlling the amount of electric current flowing in said conducting tether to vary the angle between the conductive tether the field lines of the external magnetic field and the electrodynamic force acting on said tether and attached space object.

2. A method as in claim **1** wherein said angle is less than 45 degrees.

3. A method as in claim **1** wherein said angle is 35.26 degrees.

4. A method as in claim **1** wherein said angle is controlled to average 35.26 degrees over time.

5. A method as in claim **1** wherein said angle is controlled to average 35.26 degrees over the length of the tether.

6. A method as in claim **1** wherein the space object and the tether connected to it are rotated about their center of mass while the conductive tether interacts with the external magnetic field.

7. A method as in claim **1** wherein the angle of the conductive tether to the external magnetic field is controlled by sensing a measurable parameter of the space object/conductive tether system.

8. A method as in claim **7** wherein the parameter is current flow in the conductive tether.

9. A method as in claim **7** wherein the parameter is the electrical power dissipated in the conductive tether and the current control system.

10. A method as in claim **7** wherein the parameter is the relative position of the space object and the position of the far end tip of the tether.

11. A method as in claim **7** wherein the parameter is the acceleration of the space object.

12. A method as in claim **6** wherein the parameter is the state vector of the space object as measured by the Global Positioning System.

13. A method as in claim **1** including the step of applying electric power to the conductive tether to change the state vector of the space object.

14. A method of operating a conductive tether attached to a space object, said conductive tether being oriented at an angle to the local vertical between said space object and a celestial body having a magnetic field having magnetic field lines external to said conductive tether; said conductive tether moving across said magnetic field lines of said external magnetic field at an angle to the local vertical to produce an electric current in said conducting tether and a resulting electrodynamic force acting on said tether and said attached space object comprising,

Controlling the amount of electric current flowing in said conducting tether to vary the angle between the conductive tether the local vertical between said space object and said celestial body to vary the electrodynamic force acting on said tether and attached space object.

15. A method as in claim **14** wherein said angle is less than 45 degrees.

16. A method as in claim **14** wherein said angle is controlled to average 35.26 degrees over time.

17. A method as in claim **14** wherein the space object and the tether connected to it are rotated about their center of mass while the conductive tether interacts with the external magnetic field.

18. A method as in claim **14** wherein the angle of the conductive tether to the external magnetic field is controlled by sensing a measurable parameter of the space object/conductive tether system.

19. A method as in claim **18** wherein the parameter is current flow in the conductive tether.

20. A method as in claim **18** wherein the parameter is the acceleration of the space object caused by the electrodynamic force acting on the space object.

21. A method as in claim **18** wherein the parameter is the change in the state vector of the space object as measured by the Global Positioning System.

22. A method as in claim **14** including the step of applying electric power to the conductive tether to change the state vector of the space object.

23. A method as in claim **14** wherein the celestial body is the Earth and the magnetic field is the Earth's magnetic field.

24. A method as in claim **14** wherein the celestial body is the sun and the magnetic field is the Sun's magnetic field.

25. A method as in claim **14** wherein the magnetic field is the vector sum of the magnetic fields from all celestial bodies in the solar system.

26. A method as in claim **14** wherein the conductive tether is a Hoytether.

* * * * *

United States Patent [19] [11] **4,292,804**
Rogers, Sr. [45] **Oct. 6, 1981**

[54] **METHOD AND APPARATUS FOR OPERATING AN ENGINE ON COMPRESSED GAS**

[76] Inventor: **Leroy K. Rogers, Sr.,** #5 Capistrano Ct., Ft. Myers, Fla. 33908

[21] Appl. No.: **158,303**

[22] Filed: **Jun. 10, 1980**

[51] **Int. Cl.**³ **F15B 11/06**
[52] **U.S. Cl.** **60/407;** 91/187; 91/275
[58] **Field of Search** 60/407, 412; 91/187, 91/275, 364

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,881,399	5/1975	Sagi et al.	91/187 X
3,885,387	5/1975	Simington	60/407 X
4,018,050	4/1977	Murphy	60/412 X

Primary Examiner—Allen M. Ostrager
Attorney, Agent, or Firm—Burns, Doane, Swecker & Mathis

[57] **ABSTRACT**

The present invention relates to a method and apparatus for operating an engine having a cylinder and a piston reciprocable therein on compressed gas. The apparatus comprises a source of compressed gas connected to a distributor which distributes the compressed gas to the cylinder. A valve is provided to selectively admit compressed gas to the cylinder when the piston is in an approximately top dead center position. In one embodiment of the present invention the timing of the opening of the valve is advanced such that the compressed gas is admitted to the cylinder progressively further before the top dead center position of the piston as the speed of the engine increases. In a further embodiment of the present invention a valve actuator is provided which increases the length of time over which the valve remains open to admit compressed gas to the cylinder as the speed of the engine increases. A still further embodiment of the present invention relates to an apparatus for adapting a conventional internal combustion engine for operation on compressed gas.

22 Claims, 8 Drawing Figures

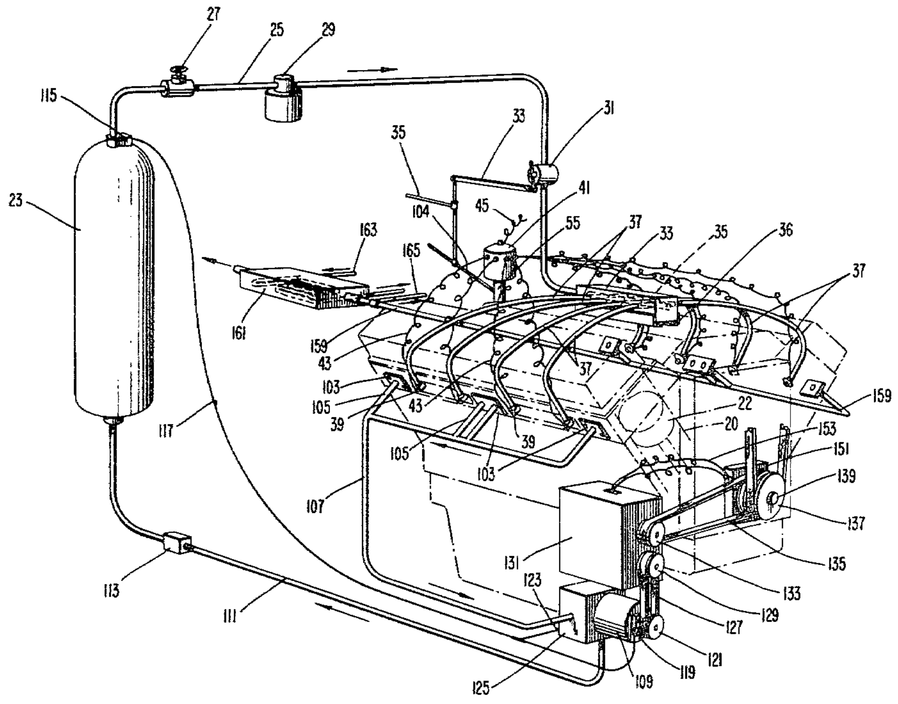

FIG. 1

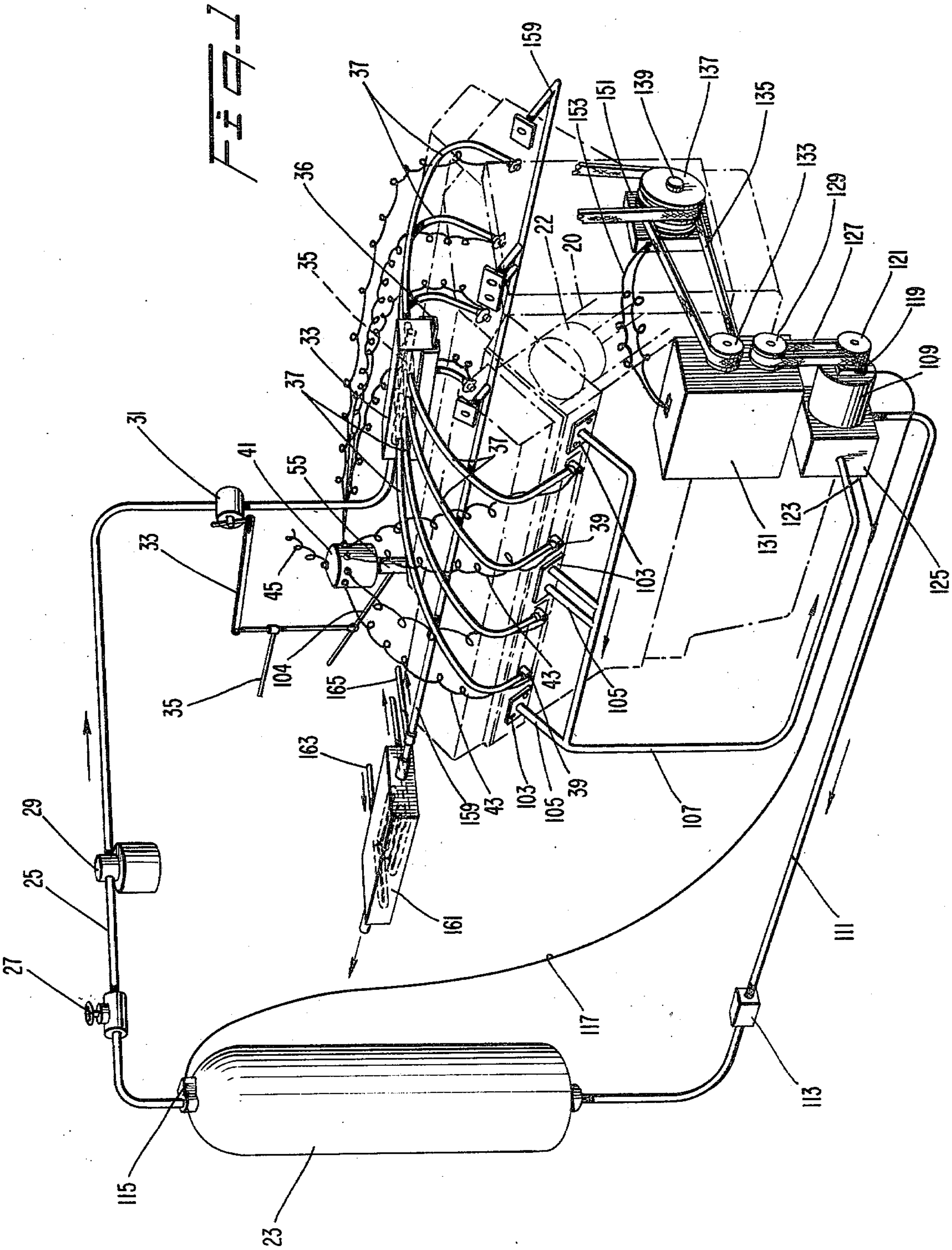

FIG. 6

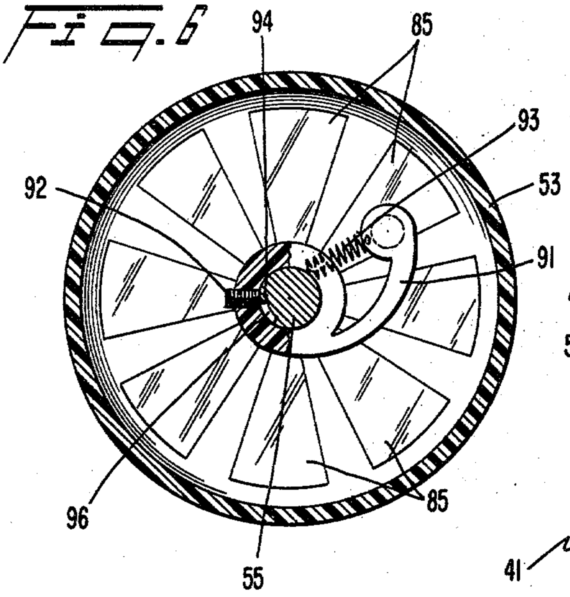
94
85
93
92
53
91
96
55
85

FIG. 7

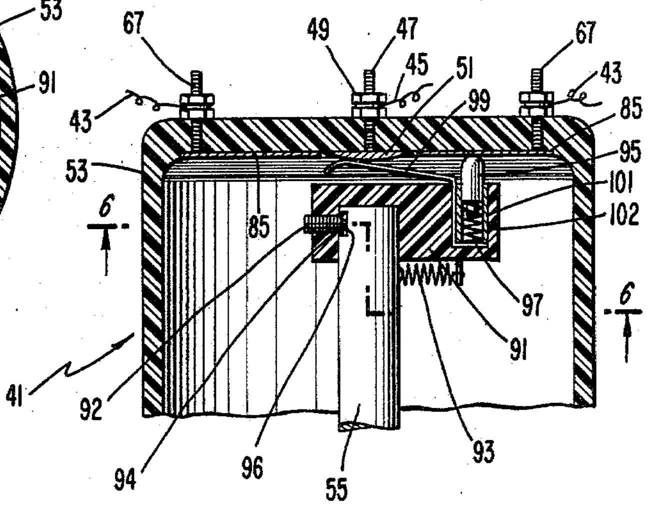
67
49
47
45
51
67
43
43
99
85
53
95
101
102
6
85
97
6
91
41
92
93
94
96
55

FIG. 3

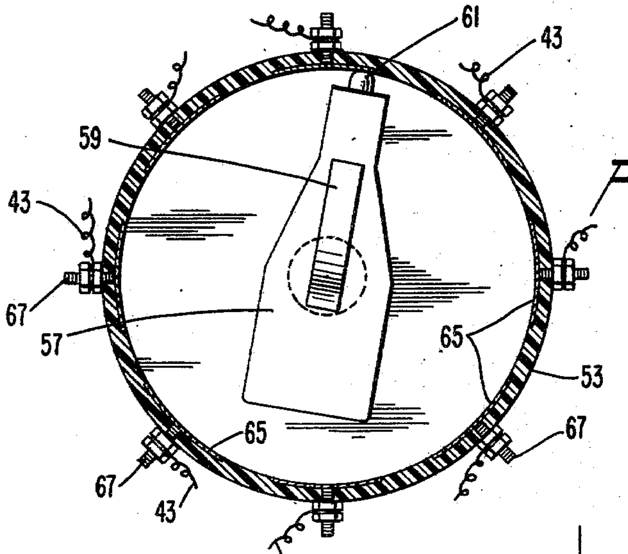
61
43
59
43
67
57
65
53
65
67
67
43
43

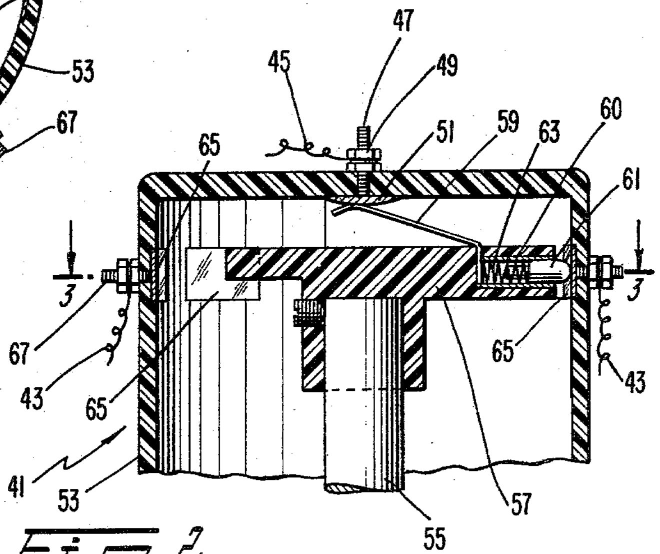
47
45
49
51
59
63
60
65
61
3
3
67
65
43
43
65
65
41
53
57
55

FIG. 2

Fig. 5

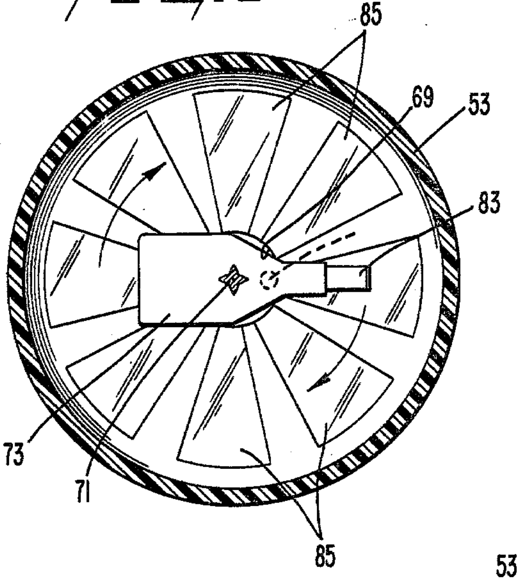
85
69
53
83
73
71
85

Fig. 4

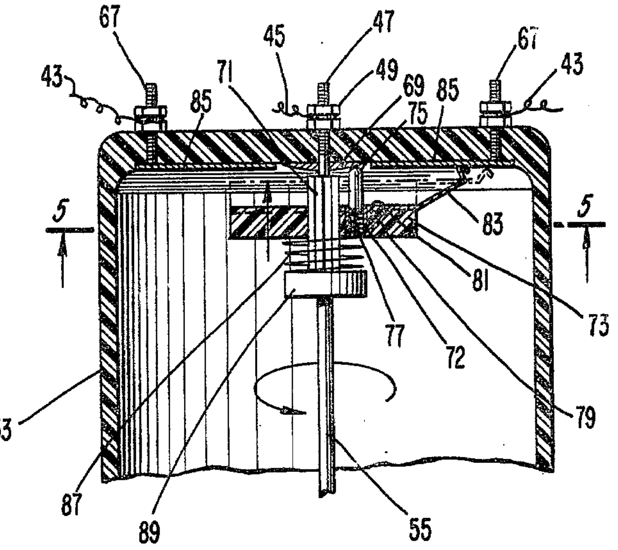
67
45
47
49
69
67
43
43
71
85
75
85
5
5
83
81
73
77
72
79
53
87
89
55

Fig. 6

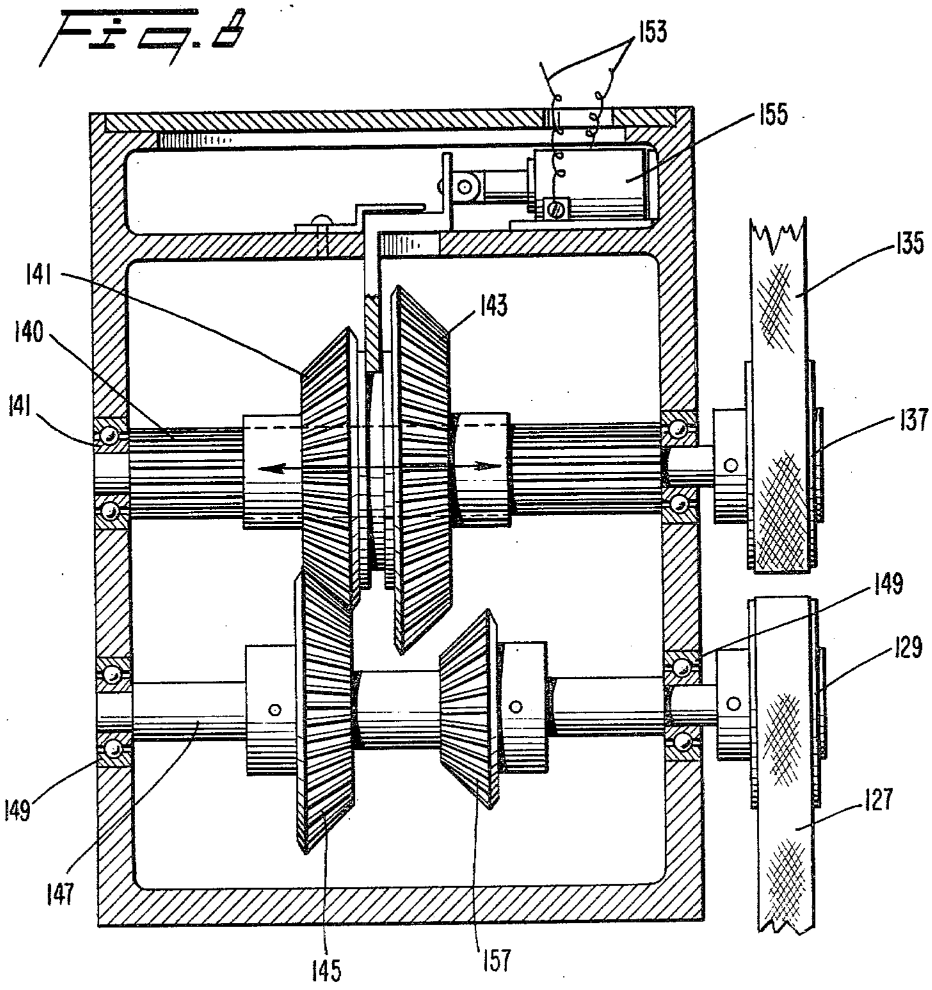
153
155
141
143
135
140
141
137
149
129
149
127
147
145
157

METHOD AND APPARATUS FOR OPERATING AN ENGINE ON COMPRESSED GAS

BACKGROUND AND SUMMARY OF THE PRESENT INVENTION

The present invention relates to a method and apparatus for operating an engine using a compressed gas as the motive fluid. More particularly, the present invention relates to a apparatus for adapting a pre-existing internal combustion engine for operation on a compressed gas.

Air pollution is one of the most serious problems facing the world today. One of the major contributors to air pollution is ordinary internal combustion engine which are used in most motor vehicles today. Various devices, including many items mandated by legislation, have been proposed in an attempt to limit the pollutants which an internal combustion engine exhausts to the air. However, most of these devices have met with limited success and are often both prohibitively expensive and complex. A clean alternative to the internal combustion engine is needed to power vehicles and other machinery.

A compressed gas, preferably air, would provide an ideal motive fluid for a engine since it would eliminate the usual pollutants exhausted from an internal combustion engine. An apparatus for converting an internal combustion engine for operation on compressed air is disclosed in U.S. Pat. No. 3,885,387 issued May 27, 1975 to Simington. The Simington patent discloses an apparatus including a source of compressed air and a rotating valve actuator which opens and closes a plurality of mechanical poppet valves. The valves deliver compressed air in timed sequence to the cylinders of an engine through adapters located in the spark plug holes. However, the output speed of an engine of this type is limited by the speed of the mechanical valves and the fact that the length of time over which each of the valves remains open cannot be varied as the speed of the engine increases.

Another apparatus for converting an internal combustion engine for operation on steam or compressed air is disclosed in U.S. Pat. No. 4,102,130 issued July 25, 1978 to Stricklin. The Stricklin patent discloses a device which changes the valve timing of a conventional four stroke engine such that the intake and exhaust valves open once for every revolution of the engine instead of once every other revolution of the engine. A reversing valve is provided which delivers live steam or compressed air to the intake valves and is subsequently reversed to allow the exhaust valves to deliver the expanded steam or air to the atmosphere. A reversing valve of this type however does not provide a reliable apparatus for varying the amount of motive fluid injected into the cylinders when it is desired to increase the speed of the engine. Further, a device of the type disclosed in the Stricklin patent requires the use of multiple reversing valves if the cylinders in a multi-cylinder engine were to be fired sequentially.

Therefore, it is an object of the present invention to provide a reliable method and apparatus for operating an engine or converting an engine for operation with a compressed gas.

A further object of the present invention is to provide a method and apparatus which is effective to deliver a constantly increasing amount of compressed gas to an engine as the speed of the engine increases.

A still further object of the present invention is to provide a method and apparatus which will operate an engine using compressed gas at a speed sufficient to drive a conventional automobile at highway speeds.

It is still a further object of the present invention to provide a method and apparatus which is readily adaptable to a standard internal combustion engine to convert the internal combustion engine for operation with a compressed gas.

Another object of the invention is to provide a method and apparatus which utilizes cool expanded gas, exhausted from a compressed gas engine, to operate an air conditioning unit and/or an oil cooler.

These and other objects are realized by a method and apparatus according to the present invention for operating an engine having at least one cylinder and a reciprricating piston therein using compressed gas as a motive fluid. The apparatus includes a source of compressed gas and a distributor connected with the source of the compressed gas for distributing the compressed gas to the at least one cylinder. A valve is provided for admitting the compressed gas to the cylinder when the piston is in approximately a top dead center position within the cylinder. An exhaust is provided for exhausting the expanded gas from the cylinder as the piston returns to approximately the top dead center position.

In a preferred embodiment of the present invention a device is provided for varying the duration of each engine cycle over which the valve remains open to admit compressed gas to the cylinder dependent upon the speed of the engine. In a further preferred embodiment of the present invention, an apparatus for advancing the timing of the opening of the valve is arranged to admit the compressed gas to the cylinder progressively further before the top dead center position of the piston as the speed of the engine increases.

Further features of the present invention include a valve for controlling the amount of compressed gas admitted to the distributor. Also, a portion of the gas which has been expanded in the cylinder and exhausted through the exhaust valve is delivered to a compressor to be recompressed and returned to the source of compressed gas. A gear train is selectively engagable to drive the compressor at different operating speed depending upon the pressure maintained at the source of compressed air and/or the speed of the engine. Still further, a second portion of the exhaust gas is used to cool a lubricating fluid for the engine or to operate an air conditioning unit.

In a preferred embodiment of the present invention, the valve for admitting compressed gas to the cylinder is electrically actuated. The device for varying the duration of each engine cycle over which the intake valve remains open as the speed of the engine increase comprises a rotating element whose effective length increases as the speed of the engine increases such that a first contact on the rotating element is electrically connected to a second contact for a longer period of each engine cycle. The second contact actuates the valve whereby the valve remains in an open position for a longer period of each engine cycle as the speed of the engine increases.

Still further features of the present invention include an adaptor plate for supporting the distributor above an intake manifold of a conventional internal combustion engine after a carburetor has been removed to allow air to enter the cylinders of the engine through the intake manifold and conventional intake valves. Another adaptor plate is arranged over an exhaust passageway of the internal combustion engine to reduce the cross-sectional area of the exhaust passageway.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred embodiments of a method and apparatus for operating an engine according to the present invention will be described with reference to the accompanying drawings wherein like members bear like reference numerals and wherein:

FIG. 1 is a schematic representation of an apparatus according to the present invention arranged on an engine;

FIG. 2 is a side view of one embodiment of a valve actuator according to the present invention;

FIG. 3 is a cross-sectional view taken along the line 3—3 in FIG. 2;

FIG. 4 is a cross-sectional view of a second embodiment of a valve actuator according to the present invention;

FIG. 5 is a view taken along the line 5—5 in FIG. 4;

FIG. 6 is a cross-sectional view of a third embodiment of a valve actuator according to the present invention;

FIG. 7 is a view taken along the line 7—7 in FIG. 6;

FIG. 8 is a cross-sectional view of a gearing unit to drive a compressor according to the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

With reference to FIG. 1, an engine block **21** (shown in phantom) having two banks of cylinders with each bank including cylinders **20** having pistons **22** reciprocable therein (only one of which is shown in phantom) in a conventional manner. While the illustrated engine is a V-8 engine, it will be apparent that the present invention is applicable to an engine having any number of pistons and cylinders with the V-8 engine being utilized for illustration purposes only. A compressed gas tank **23** is provided to store a compressed gas at high pressure. It may also be desirable to include a small electric or gas compressor to provide compressed gas to supplement the compressed gas held in the tank **23**. In a preferred embodiment, the compressed gas is air which can be obtained from any suitable source.

A line **25** transports the gas withdrawn from the tank **23** when a conventional shut off valve **27** is open. In addition, a solenoid valve **29** preferably operated by a suitable key operated switch (not shown) for the engine is also arranged in the line **25**. In normal operation, the valve **27** is maintained open at all times with the solenoid valve **29** operating as a selective shut off valve to start and stop the engine **21** of the present invention.

A suitable regulating valve **31** is arranged downstream from the solenoid valve **29** and is connected by a linkage **33** to a throttle linkage **35** which is operator actuated by any suitable apparatus such as a foot pedal (not shown). The line **25** enters an end of a distributor **33** and is connected to an end of a pipe **35** which is closed at the other end. A plurality of holes, which are equal to the number of cylinders in the engine **21**, are provided on either side of the pipe **35** along the length of the pipe **35**.

When the present invention is used to adapt a conventional internal combustion engine for operation on compressed gas, an adaptor plate **36** is provided to support the distributor **33** in spaced relation from the usual intake opening in the intake manifold of the engine after a conventional carburetor has been removed. In this way, air is permitted to enter the internal combustion engine through the usual passageways and to be admitted to the cylinders through suitable intake valves (not shown). The adaptor plate **36** is secured to the engine block **21** and the distributor **33** by any suitable apparatus, e.g., bolts.

Each of the holes in the pipe **35** is connected in fluid-tight manner to a single line **37**. Each line **37** carries the compressed gas to a single cylinder **20**. In a preferred embodiment, each of the lines **37** is $\frac{1}{2}$ inch high pressure plastic tubing attached through suitable connectors to the distributor **33** and the pipe **35**. Each of the lines **37** is connected to a valve **39** which is secured in an opening provided near the top of each of the cylinders **20**. In the case of a conversion of a standard internal combustion engine, the valves **39** can be conveniently screwed into a tapped hole in the cylinder **20** typically provided for a spark plug of the internal combustion engine. In a preferred embodiment, the valves **39** are solenoid actuated valves in order to provide a fast and reliable opening and closing of the valves **39**.

Each of the valves **39** is energized by a valve actuator **41** through one of a plurality of wires **43**. The valve actuator **41** is driven by a shaft of the engine similar to the drive for a conventional distributor of an internal combustion engine. That is, a shaft **55** of the valve actuator **41** is driven in synchronism with the engine **21** at one half the speed of the engine **21**.

A first embodiment of the valve actuator **41** (FIGS. 2 and 3) receives electrical power through a wire **45** which is energized in a suitable manner by a battery, and a coil if necessary (not shown) as is conventional in an internal combustion engine. The wire **45** is attached to a central post **47** by a nut **49**. The post **47** is connected to a conducting plate **51** arranged within a housing **53** for the valve actuator **41**. Within the housing **53**, the shaft **55** has an insulating element **57** secured to an end of the shaft **55** for co-rotation therewith when the shaft **55** is driven by the engine **21**. A first end of a flexible contact **59** is continuously biased against the conducting plate **51** to receive electricity from the battery or another suitable source. A second end of the contact **59** is connected to a conducting sleeve **60** which is in constant contact with a spring biased contact **61** which is arranged within the sleeve **60**. The contact **61** is biased by a spring **63** which urges the contact **61** towards a side wall of the housing **53**.

With reference to FIG. 3, a plurality of contacts **65** are spaced from one another and are arranged around the periphery of the housing **53** at the same level as the spring biased contact **61**. Each contact **65** is electrically connected to a post **67** which extends outside of the housing **53**. The number of contacts **65** is equal to the number of cylinders in the engine **21**. One of the wires **43**, which actuate the valves **39**, is secured to each of the posts **67**.

In operation, as the shaft **55** rotates in synchronism with the engine **21**, the insulating element **57** rotates and electricity is ultimately delivered to successive ones of the contacts **65** and wires **43** through the spring biased contact **61** and the flexible contact **59**. In this way, each of the electrical valves **39** is actuated and opened in the proper timed sequence to admit compressed gas to each of the cylinders **20** to drive the pistons **22** therein on a downward stroke.

The embodiment illustrated in FIGS. 2 and 3 is effective to actuate each of the valves 39 to remain open for a long enough period of time to admit sufficient compressed gas to each of the cylinders 20 of the engine 21 to drive the engine 21. The length of each of the contacts 65 around the periphery of the housing 53 is sufficient to permit the speed of the engine to be increased when desired by the operator by moving the throttle linkage 35 which actuates the linkage 33 to further open the regulating valve 31 to admit more compressed gas from the tank 23 to the distributor 33. However, it has been found that the amount of air admitted by the valves 39 when using the first embodiment of the valve actuator 41 (FIGS. 2 and 3) is substantially more than required to operate the engine 21 at an idling speed. Therefore, it may be desirable to provide a valve actuator 41 which is capable of varying the duration of each engine cycle over which the solenoid valves 39 are actuated, i.e., remain open to admit compressed gas, as the speed of the engine 21 is varied.

A second embodiment of a valve actuator 41 which is capable of varying the duration of each engine cycle over which each of the valves 39 remains open to admit compressed gas to the cylinders 20 dependent upon the speed of the engine 21 will be described with reference to FIGS. 4 and 5 wherein members corresponding to those of FIGS. 2 and 3 bear like reference numerals. The wire 45 from the electrical source is secured to the post 47 by the nut 49. The post 47 has a annular contact ring 69 electrically connected to an end of the post 47 and arranged within the housing 53. The shaft 55 rotates at one half the speed of the engine as in the embodiment of FIGS. 2 and 3.

At an upper end of the shaft 55, a splined section 71 slidably receives an insulating member 73. The splined section 71 of the shaft 55 positively holds the insulating member 73 for co-rotation therewith but permits the insulating member 73 to slide axially along the length of the splined section 71. Near the shaft 55, a conductive sleeve 72 is arranged in a bore 81 in an upper surface of the insulating element 73 generally parallel to the splined section 71. A contact 75, biased towards the annular contact ring 69 by a spring 77, is arranged within the conductive sleeve 72 in contact therewith. The conductive sleeve 72 also contacts a conductor 79 at a base of the bore 81.

The conductor 79 extends to the upper surface of the insulating element 73 near an outer periphery of the insulating element 73 where the conductor 79 is electrically connected to a flexible contact 83. The flexible contact 83 selectively engages a plurality of radial contacts 85 arranged on an upper inside surface of the housing 53. A weak spring 87 arranged around the splined section 71 engages a stop member 89 secured on the shaft 55 and the insulating element 73 to slightly bias the insulating element 73 towards the upper inside surface of the housing 53 to ensure contact between the flexible contact 83 and the upper inside surface of the housing 53. As best seen in FIG. 5, the radial contacts 85 on the upper inside surface of the housing 53 are arranged generally in the form of radial spokes extending from the center of the housing 53 with the number of contacts being equal to the number of cylinders 20 in the engine 21. The number of degrees covered by each of the radial contacts 85 gradually increases as the distance from the center of the upper inside surface of the housing 53 increases.

In operation of the device of FIGS. 4 and 5, as the shaft 55 rotates, electricity flows along a path through the wire 45 down through post 47 to the annular contact member 69 which is in constant contact with the spring biased contact 75. The electrical current passes through the conductive sleeve 72 to the conductor 79 and then to the flexible contact 83. As the flexible contact 83 rotates along with the insulating member 73 and the shaft 55, the tip of the flexible contact 83 successively engages each of the radial contacts 85 on the upper inside of the housing 53. As the speed of the shaft 55 increases, the insulating member 73 and the flexible contact 83 attached thereto move upwardly along the splined section 71 of the shaft 55 due to the radial component of the splines in the direction of rotation under the influence of centrifugal force. As the insulating member 73 moves upwardly, the flexible contact 83 is bent such that the tip of the contact 83 extends further radially outwardly from the center of the housing 53 (as seen in phantom lines in FIG. 4). In other words, the effective length of the flexible contact 83 increases as the speed of the engine 21 increases.

As the flexible contact 83 is bent and the tip of the contact 83 moves outwardly, the tip remains in contact with each of the radial contacts 85 for a longer period of each engine cycle due to the increased angular width of the radial contacts with increasing distance from the center of the housing 53. In this way, the length of time over which each of the valves 39 remains open is increased as the speed of the engine is increased. Thus, a larger quantity of compressed gas or air is injected into the cylinders as the speed increases. Conversely, as the speed decreases and the insulating member 73 moves downwardly along the splined section 71, a minimum quantity of air is injected into the cylinder due to the shorter length of the individual radial contact 85 which is in contact with the flexible contact 83. In this way, the amount of compressed gas that is used during idling of the engine 21 is at a minimum whereas the amount of compressed gas which is required to increase the speed of the engine 21 to a level suitable to drive a vehicle on a highway is readily available.

With reference to FIGS. 6 and 7, a third embodiment of a valve actuator 41 according to the present invention includes an arcuate insulating element 91 having a first end pivotally secured by any suitable device such as screw 92 to the shaft 55 for co-rotation with the shaft 55. The screw 92 is screwed into a tapped hole in the insulating element 91 such that a tab 94 at an end of the screw 92 engages a groove 96 provided in the shaft 55. In this way, the insulating element 91 positively rotates with the shaft 55. However, as the shaft 55 rotates faster, a second end 98 of the insulating element 91 is permitted to pivot outwardly under the influence of centrifugal force because of the groove 96 provided in the shaft 55. A spring 93 connected between the second end 98 of the element 91 and the shaft 55 urges the second end of the element 91 towards the center of the housing 53.

A contact 99 similar to the contact 59 (FIG. 2) is arranged such that one end of the contact 99 is in constant contact with the conducting plate 51 located centrally within the housing 53. The other end of the contact 99 engages a conductive sleeve 101 arranged in bore 102. A contact element 95 is arranged in the conductive sleeve 101 in constant contact with the sleeve 101. The bore 102 is arranged generally parallel to the shaft 55 near the second end of the arcuate insulating element **91**. The contact **95** is biased by a spring **97** towards the upper inside surface of the housing **53** for selective contact with each of the plurality of radial contacts **85** which increase in arc length towards the outer peripheral surface of the housing **53** (FIG. **6**).

In operation of the device of FIGS. **6** and **7**, as the shaft **55** rotates the arcuate insulating element **91** rotates with the shaft **55** and the second end **98** of the insulating element **91** tends to pivot about the shaft **55** due to centrifugal force. Thus, as the effective length of the contact **95** increases, i.e., as the arcuate insulating element **91** pivots further outwardly, the number of degrees of rotation over which the contact **95** is in contact with each of the radial contacts **85** on the upper inside surface of the housing **53** increases thereby permitting each of the valves **39** to remain open for a longer period of each engine cycle to admit more compressed gas to the respective cylinder **20** to further increase the speed of the engine **21**.

With reference to FIG. **1**, a mechanical advance linkage **104** which is connected to the throttle linkage **35**, advances the initiation of the opening of each valve **39** such that compressed gas is injected into the respective cylinder further before the piston **22** in the respective cylinder **20** reaches a top dead center position as the speed of the engine is increased by moving the throttle linkage **35**. The advance linkage **104** is similar to a conventional standard mechanical advance employed on an internal combustion engine. In other words, the linkage **104** varies the relationship between the angular positions of a point on the shaft **55** and a point on the housing **53** containing the contacts. Alternatively, a conventional vacuum advance could also be employed. By advancing the timing of the opening of the valves **39**, the speed of the engine can more easily be increased.

The operation of the engine cycle according to the present invention will now be described. The compressed gas injected into each cylinder of the engine **21** drives the respective piston **22** downward to drive a conventional crankshaft (not shown). The movement of the piston downwardly causes the compressed gas to expand rapidly and cool. As the piston **22** begins to move upwardly in the cylinder **20** a suitable exhaust valve (not shown) arranged to close an exhaust passageway is opened by any suitable apparatus. The expanded gas is then expelled through the exhaust passageway. As the piston **22** again begins to move downwardly a suitable intake valve opens to admit ambient air to the cylinder. The intake valve closes and the ambient air is compressed on the subsequent upward movement of the piston until the piston reaches approximately the top dead center position at which time the compressed gas is again injected into the cylinder **20** to drive the piston **22** downward and the cycle begins anew.

In the case of adapting a conventional internal combustion engine for operation on compressed gas, a plurality of plates **103** are preferably arranged over an end of the exhaust passageways in order to reduce the outlet size of the exhaust passageways of the conventional internal combustion engine. In the illustrated embodiment, a single plate having an opening in the center is bolted to the outside exhaust passageway on each bank of the V-8 engine while another single plate having two openings therein is arranged with one opening over each of the interior exhaust passageways on each bank of the V-8 engine. A line **105** is suitably attached to each of the adaptor places to carry the exhaust to an appropriate location. In a preferred embodiment, the exhaust lines **105** are 1½" plastic tubing.

In a preferred embodiment, the exhaust lines **105** of one bank of the V-8 engine are collected in a line **107** and fed to an inlet of a compressor **109**. The pressure of the exhaust gas emmanating from the engine **21** according to the present invention is approximately 25 p.s.i. In this way, the compressor **109** does not have to pull the exhaust into the compressor since the gas exhausted from the engine **21** is at a positive pressure. The positive pressure of the incoming fluid increases the efficiency and reduces wear on the compressor **109**. The exhaust gas is compressed in the compressor **109** and returned through a line **111** and a check valve **113** to the compressed gas storage tank **23**. The check valve **113** prevents the flow of compressed gas stored in the tank **23** back towards the compressor **109**.

A suitable pressure sensor **115** is arranged at an upper end of the tank **23** and sends a signal along a line **117** when the pressure exceeds a predetermined level and when the pressure drops below a predetermined level. The line **117** controls an electrically actuated clutch **119** disposed at a front end of the compressor **109**. The clutch **119** is operative to engage and disengage the compressor **109** from a drive pulley **121**. Also, the signal carried by the line **117** actuates a suitable valve **123** arranged on a compressor housing **125** to exhaust the air entering the compressor housing **125** from the line **107** when the clutch **119** has disengaged the compressor **109** from the drive pully **121**.

In a preferred embodiment, when the pressure is the tank **23** reaches approximately 600 p.s.i., the clutch **119** is disengaged and the compressor **109** is deactivated and the valve **123** is opened to exhaust the expanded gas delivered to the compressor **109** from the line **107** to the atmosphere. When the pressure within the tank **23** drops below approximately 500 p.s.i., the sensor **115** sends a signal to engage the clutch **119** and close the valve **123**, thereby operating the compressor **109** for supplying the tank **23** with compressed gas.

The pulley **121** which drives the compressor **109** through the clutch **119** is driven by a belt **127** which is driven by a pulley **129** which operates through a gear box **131**. With reference to FIGS. **1** and **8**, a second pulley **133** on the gear box is driven by a belt **135** from a pulley **137** arranged on a drive shaft **139** of the engine **21**. The pulley **137** drives a splined shaft **140** which has a first gear **141** and a second larger gear **143** arranged thereon for rotation with the splined shaft **140**. The splined shaft **140** permits axial movement of the gears **141** and **143** along the shaft **140**.

In normal operation (as seen in FIG. **8**), the first gear **141** engages a third gear **145** arranged on a shaft **147** which drives the pulley **129**. The shafts **140** and **147** are arranged in suitable bearings **149** arranged at each end thereof. When the speed of the engine **21** drops below a predetermined level, a suitable sensor **151** responsive to the speed of the drive shaft **139** of the engine **21** generates a signal which is transmitted through a line **153** to a solenoid actuator **155** arranged within the gear box **131**. The solenoid actuator **155** moves the first and second gears **141**, **143** axially along the splined shaft **140** to the right as seen in FIG. **8** such that the second, larger gear **143** engages a fourth smaller gear **157** which is arranged on the shaft **147**. The ratio of the second gear **143** to the fourth gear **157** is preferably approximately 3 to 1.

In this way, when the speed of the engine **21** drops below the predetermined level as sensed by the sensor **151** (which predetermined level is insufficient to drive the compressor **109** at a speed sufficient to generate the 500-600 pounds of pressure which is preferably in the tank **23**), the solenoid actuator **155** is energized to slide the gears **143**, **141** axially along the splined shaft **140** so that the second, larger gear **143** engages the fourth, smaller gear **157** to drive the pulley **129** and hence the compressor **109** at a higher rate of speed to generate the desired pressure. When the speed of the engine increases above the predetermined level, in a preferred embodiment approximately 1500 rpm, the solenoid actuator **155** is deactivated by the sensor **151** thereby moving the gears **143** and **141** to the left as seen in FIG. **8** such that the first gear **141** re-engages with the third gear **145** to effectuate a 1 to 1 ratio between the output shaft **139** of the engine **21** and the pulley **129**.

The other bank of the V-8 engine has its exhaust ports arranged with adapter plates **103** similar to those on the first bank. However, the exhaust from this bank of the engine **21** is not collected and circulated through the compressor **109**. In a preferred embodiment, a portion of the exhaust is collected in a line **159** and fed to an enlarged chamber **161**. A second fluid is fed through a line **163** into the chamber **161** to be cooled by the cool exhaust emmanating from the engine **21** in the line **159**. The second fluid in the line **163** may be either transmission fluid contained in a transmission associated with the engine **21** or a portion of the oil used to lubricate the engine **21**. A second portion of the exhaust from the second bank of the V-8 engine is removed from the line **159** in a line **165** and used as a working fluid in an air conditioning system or for any other suitable use.

It should be noted that the particular arrangement utilized for collecting and distributing the gas exhausted from the engine **21** would be determined by the use for which the engine is employed. In other words, it may be advantageous to rearrange the exhaust tubing such that a larger or smaller percentage of the exhaust is routed through the compressor **109**. It should also be noted that since the exhaust lines **105** are plastic tubing, a rearrangement of the lines for a different purpose is both simple and inexpensive.

In operation of the engine of the present invention, the engine **21** is started by energizing the solenoid valve **29** and any suitable starting device (not shown), e.g., a conventional electric starter as used on an internal combustion engine. Compressed gas from the full tank **23** flows through the line **25** and a variable amount of the compressed gas is admitted to the distributor **33** by controlling the regulator valve **31** through the linkage **33** and the operator actuated throttle linkage **35**. The compressed gas is distributed to each of the lines **37** which lead to the individual cylinders **20**. The compressed gas is admitted to each of the cylinders **20** in timed relationship to the position of the pistons within the cylinders by opening the valves **39** with the valve actuator **41**.

When it is desired to increase the speed of the engine, the operator moves the throttle linkage **35** which simultaneously admits a larger quantity of compressed gas to the distributor **33** from the tank **23** by further opening the regulator valve **31**. The timing of the valve actuator **41** is also advanced through the linkage **104**. Still further, as the speed of the engine **21** increases, the effective length of the rotating contact **83** (FIG. **4**) or **95** (FIG. **6**) increases thereby electrically contacting a

wider portion of one of the stationary radial contacts **85** to cause each of the valves **39** to remain open for a longer period of each engine cycle to admit a larger quantity of compressed gas to each of the cylinders **20**.

As can be seen, the combination of the regulating valve **31**, the mechanical advance **104**, and the valve actuator **41**, combine to produce a compressed gas engine which is quickly and efficiently adaptable to various operating speeds. However, all three of the controls need not be employed simultaneously. For example, the mechanical advance **104** could be utilized without the benefit of one of the varying valve actuators **41** but the high speed operation of the engine may not be as efficient. By increasing the duration of each engine cycle over which each of the valves **39** remains open to admit compressed gas to each of the cylinders **20** as the speed increases, conservation of compressed gas during low speed operation and efficient high speed operation are both possible.

After the compressed gas admitted to the cylinder **20** has forced the piston **22** downwardly within the cylinder to drive the shaft **139** of the engine, the piston **22** moves upwardly within the cylinder **20** and forces the expanded gas out through a suitable exhaust valve (not shown) through the adapter plate **103** (if employed) and into the exhaust line **105**. The cool exhaust can then be collected in any suitable arrangement to be compressed and returned to the tank **23** or used for any desired purpose including use as a working fluid in an air conditioning system or as a coolant for oil.

When using the apparatus and method of the present invention to adapt a ordinary internal combustion engine for operation with compressed gas it can be seen that considerable savings in weight are achieved. For example, the ordinary cooling system including a radiator, fan, hoses, etc. can be eliminated since the compressed gas is cooled as it expands in the cylinder. In addition, there are no explosions within the cylinder to generate heat. Further reductions in weight are obtained by employing plastic tubing for the lines which carry the compressed gas between the distributor and the cylinders and for the exhaust lines. Once again, heavy tubing is not required since there is little or no heat generated by the engine of the present invention. In addition, the noise generated by an engine according to the present invention is considerably less than that generated by an ordinary internal combustion engine since there are no explosions taking place within the cylinders.

The principles of preferred embodiments of the present invention have been described in the foregoing specification. However, the invention which is intended to be protected is not to be construed as limited to the particular embodiments disclosed. The embodiments are to be regarded as illustrative rather than restrictive. Variations and changes may be made by others without departing from the spirit of the invention. Accordingly, it is expressly intended that all such variations and changes which fall within the spirit and the scope of the present invention as defined in the appended claims be embraced thereby.

What is claimed is:

1. An apparatus for operating an engine having at least one cylinder and a reciprocating piston therein comprising:

a source of compressed gas;

- distributor means connected with the source of compressed gas for distributing the compressed gas to the at least one cylinder;
- valve means for admitting the compressed gas to the at least one cylinder when the piston is in approximately a top dead center position within the cylinder;
- altering means for increasing the duration of each engine cycle over which the valve means admits compressed gas to the at least one cylinder as the speed of the engine increases; and
- exhaust means for exhausting gas as the piston subsequently approaches approximately the top dead center position.

2. The apparatus of claim **1** further comprising control means for controlling the amount of compressed gas admitted to the distributor means.

3. The apparatus of claim **1** wherein the valve means is a solenoid valve secured in an opening in the cylinder above the level of the piston at the top dead center position.

4. The apparatus of claims **1** or **2** further comprising means for advancing the timing of the valve means as the speed of the engine increases such that compressed gas is admitted progressively further before the top dead center position as the speed of the engine increases.

5. The apparatus of claim **4** wherein the means for advancing the timing comprises a mechanical linkage connected to an operator actuated accelerator linkage.

6. The apparatus of claim **1** wherein a portion of the gas exhausted through the exhaust means is compressed in a compressor driven by an output shaft of the engine and is returned to the source of compressed gas.

7. The apparatus of claim **1** wherein a portion of the gas exhausted through the exhaust means is used to cool transmission fluid for a transmission associated with the engine.

8. The apparatus of claim **1** wherein a portion of the gas exhausted through the exhaust means is used as a working fluid in an air conditioning system.

9. The apparatus of claim **6** further comprising first gearing means interposed between the output shaft of the engine and the compressor for increasing the speed at which the compressor is driven.

10. The apparatus of claim **6** further comprising clutch means attached to the compressor both for disengaging the compressor from the output shaft of the engine when a first predetermined pressure at the source of compressed gas is exceeded and for engaging the compressor with the output shaft of the engine when the pressure at the source of compressed gas drops below a second predetermined pressure.

11. The apparatus of claim **9** further comprising means for both disengaging the first gearing means when a predetermined speed of the engine is exceeded and engaging a second gearing means for driving the compressor at a speed slower than the first gearing means when the predetermined speed of the engine is exceeded.

12. The apparatus of claim **1** wherein the valve means is electrically actuated and wherein the altering means comprises:

- a rotating member timed with the at least one cylinder and arranged within a housing;
- first and second contacts arranged on a first end of the rotating member and on an inside surface of the housing, respectively;
- means for increasing the distance of the first contact from the rotational axis of the rotating member as the speed of the engine increases such that the first contact moves radially outwardly within the housing; and
- said second contact presenting a longer arc length to the first contact as the distance of the first contact from the rotational axis of the rotating member increases.

13. The apparatus of claim **12** wherein the rotating member comprises an arcuate arm and wherein the means for increasing the distance of the first contact comprises pivotally mounting a second end of the arcuate arm about the axis of rotation of the rotating member and spring means for biasing the first end of the arcuate arm towards a radially inward position whereby the first end of the arcuate arm pivots radially outwardly as the speed of the engine increases.

14. The apparatus of claim **12** wherein the rotating member is axially slidably received on a rotating shaft for co-rotation therewith, said shaft having splines with a radial component in the direction of rotation, and wherein the first contact comprises a flexible contact located on an upper surface of the rotating member, said flexible contact being biased against the inside surface of the housing which carries the second contacts whereby as the speed of the engine increases the rotating member is urged axially along the splined shaft towards the inside surface of the housing such that the flexible contact is forced radially outwardly along the inside surface.

15. The apparatus of claim **12** wherein the second contact comprises of radially extending conductor arranged on an upper inside surface of the housing, said conductor increasing in arc length as the conductor extends radially outwardly from a central portion of the housing.

16. An apparatus for adapting an internal combustion engine for operation with compressed gas, the internal combustion engine having at least one cylinder, a piston reciprocable within the at least one cylinder, intake and exhaust means disposed in the at least one cylinder, and a tapped hole in the at least one cylinder adapted to receive a spark plug, the apparatus comprising:

- a source of compressed gas;
- distributor means connected with the source of compressed gas for distributing the compressed gas to the at least one cylinder;
- valve means arranged in the tapped hole for admitting the compressed gas to the at least one cylinder when the piston is in approximately a top dead center position within the cylinder; and
- altering means for increasing the duration of each engine cycle over which the valve means remains open to admit the compressed gas as the speed of the engine increases.

17. An apparatus as in claim **16** further comprising first adapter plate means for supporting the distributor means above an intake manifold of the engine, which adaptor plate means allows ambient air to enter through the intake manifold.

18. The apparatus of claim **16** further comprising second adapter plate means for reducing the exit area of the exhaust means.

19. A method of operating an engine on compressed gas, said engine having at least one cylinder and a piston reciprocable therein comprising the steps of:

delivering compressed gas from a source to a distributor;

distributing the compressed gas to the at least one cylinder;

admitting compressed gas to the at least one cylinder through an intake valve when the piston is at approximately a top dad center position;

increasing the duration of each engine cycle over which compressed gas is admitted to the at least one cylinder as the engine speed increases; and

exhausting the remaining gas when the piston subsequently reaches approximately the top dead center position.

20. The method of claim **19** further comprising the step of controlling the amount of compressed gas which is delivered to the distributor.

21. The method of claim **19** further comprising the step of advancing the timing of the opening of the intake valve as the speed of the engine increases.

22. An apparatus for operating an engine having at least one cylinder and a piston reciprocable therein on compressed gas comprising:

a source of compressed gas;

distributor means connected with the source of compressed gas for distributing the compressed gas to the at least one cylinder;

electrically actuated valve means secured in an opening in the at least one cylinder for selectively admitting compressed gas to the at least one cylinder when the piston is in approximately a top dead center position; and

means for advancing the timing of the valve means as the speed of the engine increases whereby compressed gas is admitted progressively further before the top dead center position as the speed of the engine increases.

* * * * *

FortuneCity Get your free website today!
Web hosting Custom Email SiteBuilder

H&R BLOCK AT HOME™.
Get It Right.℠
File for FREE ▸
H&R BLOCK At Home

Ads by Google Pipe Fittings Pipe Break Pipe Welder Acrylic Pipe

---

Goodbye Chemtrails, Hello Blue Skies!
The Do-It-Yourself Kit for Sky Repair

This web page reproduced with permission from
Ken Adachi of Educate_yourself organization
original page
Jan. 10. 2002

A couple of years ago, I was standing outside a restaurant on a sunny day in southern California showing Joanna Beck, wife of Dr. Bob Beck, just how simple it was to disperse clouds using The Visual Ray (http://educate-yourself.org/visualray21may0.html), a technique I learned from reading Trevor James Constable's book, *The Cosmic Pulse of Life*. After demonstrating the technique to her on a couple of smaller clouds, we worked together on evaporating a particular large cumulus cloud and noticed how the blueness of the sky was slightly darker where the large cloud had existed only a few minutes earlier, a result of intensified orgone energy accumulating in that area. Almost simultaneously, we wondered aloud whether a group of people could use this technique to break up chemtrails! I tried it a number of times on my own, but I couldn't get chemtrail 'clouds' to break up and disperse the same way normal clouds would and I wondered silently whether a *Cloudbuster*, a simple invention of Wilhelm Reich.to direct orgone energy in the atmosphere, could tackle the job.

Over the course of 35 years, Trevor Constable had modified Reich's original long pipe arrayed Cloudbuster down to a much smaller, compact, and simpler design that could 'make' or 'break' weather in any direction he chose: He could make rain or stop it; he could relieve drought and break up smog or create a dry spell and intensify smog conditions. It all depended on how you manipulated the atmospheric ether in a given area. Trevor discovered that it all depends on how you apply his modified Cloudbuster design to direct the *flow* of etheric energy. If you bottle up the etheric flow, or stymie its normal expansive and contractile pulse, you will get drought (and smog). If you release the etheric flow after being 'bottled up' or stuck during a dry spell, you will get torrential rains. There is likely no civilian on the face of this earth who has accumulated more expertise about weather engineering than Trevor James Constable

Don Croft

However, the hero of this article is **Don Croft** (zapper16@earthlink.net), a man who was led to Reich's work with orgone energy and Cloudbusters by another route entirely. Someone had directed Don to a web site (http://www.orgone.net) that discussed Reich's observations about orgone energy and how you could utilize Reich's inventions of the Orgone Accumulator (abbreviated "oracs") involving a much smaller design than the conventional oracs built by Reich using metal shavings, magnets, crystals, and epoxy or polyester resin. Don started to make cone shaped oracs using the above mentioned materials (and plastic funnels for the mold) and called them "Holy Hand Grenades" (HHG's) (http://educate-yourself.org/HHGconstruction10jan02.html). He also incorporated these ideas into his "terminator" zappers (contact me if interested: Editor@educate-yourself.org) .

"Holy Hand Grenades" Batman!

These things had the ability to focus life positive energy and break up areas of noxious energies or DOR, a word coined by Reich which means Deadly Orgone Energy. Don calls them orgone *generators* since they neutralize DOR and seem to only generate positive, life affirming energies. Don began planting these HHG's in outdoor locations near nuclear power plants and other (classified) areas of noxious, life depleting energies and started to have a definite effect on improving the energy fields radiating from these zones. Since he already was thinking about using the cloudbuster to try and break up chemtrails, Don took it a step further and combined a large version of his HHG with a cloudbuster utilizing rigid copper pipes and a 2 gallon plastic bucket to hold the whole enchilada together. This modification of Reich's Cloudbuster is a unique and original design idea of Don's and produces some rather special effects.

He set up his combo orac/cloudbuster outside on a chemtrail overcast day with the pipes pointing straight up and sure enough, after a while, the chemtrails started to break up revealing blue sky above. The longer he left the orac/CB in place, the clearer the sky got , eventually covering an area out to 90 miles. He later improved the orac design and got the coverage to extend out to 120 miles.

Don had contacted me by E mail last year and was sending me reports of his success with the orac/CB both in this country and abroad. He had sent the instructions to an African man named Gert Botha living in Namibia who built a modified version of the device using locally scavenged materials (read "Africa's First Cloudsbuster" {http://educate-yourself.org/africasfirstcloudbusterjun01.html}and then see Gert's Report {http://educate-yourself.org/gertbothareportoct01.html}). Initially, Gert found that his Cloudbuster was having an effect on the incessant sand storms causing them to constantly change direction, but eventually he found that aiming the Cloudbuster south at an angle of 30 degrees to the horizon forced the sandstorms to stop. After being parked in this position for two months, his orac/CB brought on torrential rains for the first time in years in one of the driest regions of Africa.

Psychic Dames

Don gains many of his "insights" about his oracs and other devices through dowsing with his wife, Carol Derry, and his daughter, Nora, who both possess psychic ability, big time. His two psychic 'dames' also keep him one step ahead of the Bad Guys. Don is on the road most of the time and whenever the black op types try to move in on him, he learns about it before they do! The Bad Guys have their high tech advantages, but it seems the Universe takes care of its own when you're trying to do the Right Thing. In Don's case, he's been provided with exactly the sort of 'assistance' he needs (three cheers for Carol and Nora!).

Your Weekend Project

Complete instructions to make your own orac/CB or "ChemBuster" are included in this article. It will cost you about $150 US dollars if you buy all of the materials new, but if you're a scrounger, you can probably do it for the cost of the plastic resin. The instructions were written by Don Croft with a little bit of editing by another man who wished to remain anonymous. The pictures were mostly supplied by Don, but Mr. Anonymous threw in one photo of his own orac/CB breaking up chemtrails in the midwest, his neck of the woods. Anyone with modest do-it-yourself skills can build these devices, but be aware that an orac/CloudBuster is not a toy. It directs orgone energy, both the good type and the harmful DOR type.

If you don't educate yourself about a Cloudbusters and handle this thing carelessly, you could experience some discomfort, but not the sort of DOR that Reich's Cloudbuster design could accumulate. Trevor Constable told me first hand a story about the late Robert McCullough, Reich's friend and assistant in the Arizona desert in the 1950's. Dr Reich and Bob were running pioneering experiments with Reich's Cloudbuster when Bob had allowed a big (functioning) Cloudbuster to accidentally touch the side of his bare leg-which instantly paralyzed his leg and caused him excruciating pain. It took weeks to recover and Bob walked with a limp and had to use a cane for the rest of his life. You need to remember to discharge an active Cloudbuster before touching the pipes. Don recommends touching the base of the unit after touching the pipes to discharge any DOR energy. You could also lay the unit on its side on the ground and remove the pipes with a rubber glove. Laying the pipes on the bare earth should discharge them of any DOR.

So here's the edited reprint of Don's papers on Chemtrail Busting and the complete instructions for making your own "Chembuster" .

You can also review an earlier version of Don's Cloudbuster directions (http://educate-yourself.org/CBconstructioninfo1feb02.html) which he had sent to me in August, 2001. It includes some interesting uses for the CB not included in the directions given below. At the time, Don indicated that you could get 5 double pointed quartz crystals from "David" at **davdart@texas.net** for $7 including shipping, but I don't know if that contact is still valid. Give it a try.

How to Build A Cloudbuster to Disperse Chemtrails and Improve the Weather

By Don Croft

Definition of a contrail: A contrail occurs when a plane travels at a high altitude (about 30,000 ft. or more) and compresses the air into a water vapor or ice crystals through jet engines or the wing tips pushing through the air. This trail disappears after anywhere from a few seconds to 15 minutes, as the water vapor evaporates.

Definition of a chemtrail: The term "chemtrail" is a relatively new word, which appeared in the last few years along with the appearance of chemtrails. A chemtrail is very different from a contrail. At first a chemtrail might look a bit like a contrail. However, instead of disappearing like a contrail does, a chemtrail just keeps spreading out and forming a hazy cloud bank. These trails traverse the whole sky and stay for up to around five or even eight hours. They have been known to turn what was originally a clear blue sky into a grey haze. (Note: As of Nov 2001 there are already several dozen cloudbusters in seven countries, so if you see chemtrails that don't last anywhere near this long, there may already be a cloudbuster in the area).

It is believed chemtrails are composed at least in part by an oil based substance of unknown content that is actually being sprayed from jets on populated areas. It has also been found that some chemtrails contain red blood cells, aluminum, mycoplasma, viruses and other unknown biological components. There are many reports of increased diseases, especially respiratory ailments, in areas where there are chemtrails.

Chemtrails are often seen to be sprayed by white planes with no markings. Sometimes you will see these planes give off no contrail at all, and immediately after the same kind of plane in the same area will start to leave a thick chemtrail. At other times, distinct short gaps in chemtrails have been photographed, as though the plane ran out of chemicals and then had to start spraying with a new supply.

For information on Chemtrails see the sites listed at http://www.carnicom.com/contrails.htm

History of the Cloudbuster

The Wilhelm Reich cloudbuster (CB) was well known for its ability to create rain. But a problem existed with its continued use, in that the deadly orgone energy that it absorbed could in fact hurt the operator if it was not correctly "drained off". This "draining off" of deadly orgone energy was accomplished by 1. Connecting the cloudbuster to a body of running water (difficult) or 2. Connecting the cloudbuster output to an orgone accumulator. However, the orgone accumulator would soon become saturated with the deadly orgone energy which in turn caused more problems for the operator. So the cloudbuster, as good a technology as it is, has lain dormant for many years.

The Cloudbuster Re-Invented

By utilizing the research of Wilhelm Reich, we have created a cheap, portable and easy to build device that consistently destroys Chemtrails and heals the atmosphere. The "chembuster" is the answer to these ongoing attacks. Unlike the original cloudbuster, the CB changes the deadly orgone energy to good orgone energy and does NOT become saturated or dangerous to the operator. You can build one for about $150.

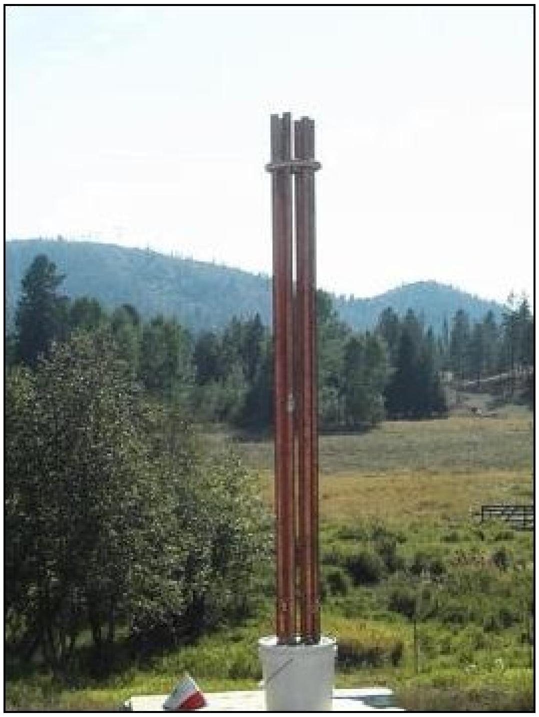

Join The Battle - Now!

It is human nature to sit on the sidelines and think that one person cannot change the things that are happening to our country. You now have no excuse for inaction. For a few dollars and a few hours of your time YOU can make a difference.

A properly constructed CB unit should be able to greatly reduce the effects of chemtrails for about 45 miles in all directions (90 mile diameter) from the unit. If used with Slim Spurling's environmental tools, and a few extra crystals, the diameter can be increased to about 120 miles in all directions.

With this device a mere few hundred concerned citizens can SHUT DOWN the multi-billion dollar chemtrail program. The chembuster cannot do any damage. It can only help and heal. Its up to you to join the fight to save America.

Chembuster Construction Details

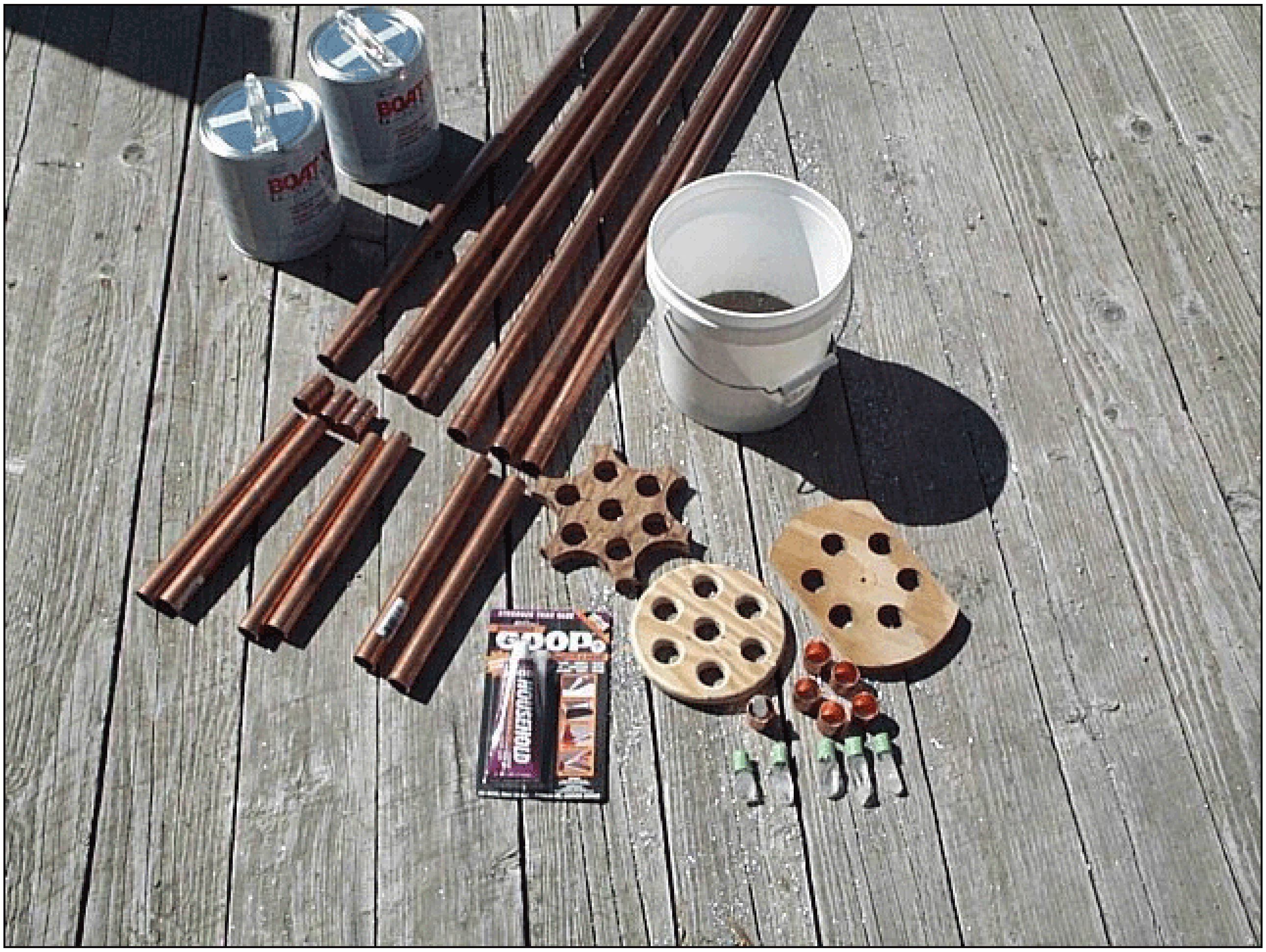

Cloudbuster Materials

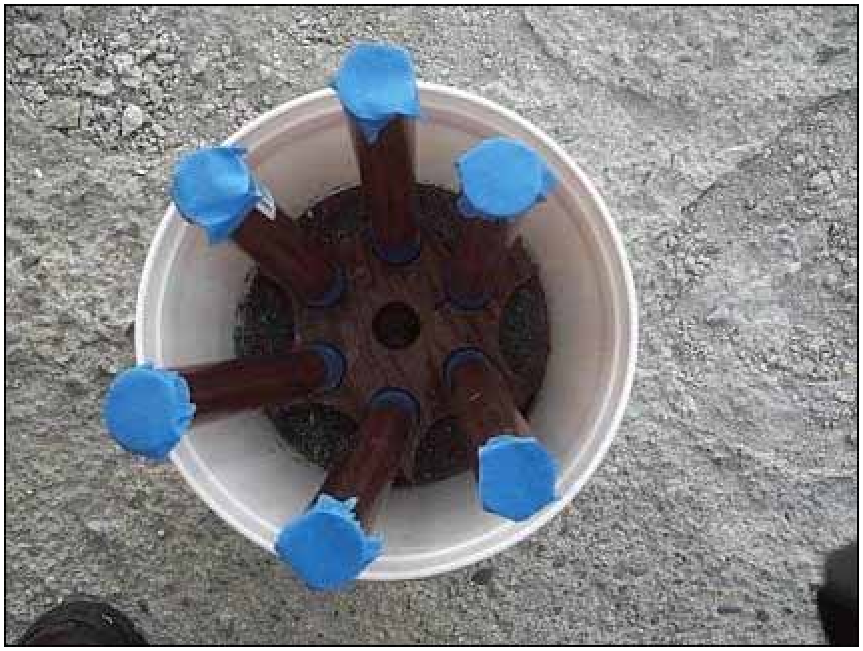

Set pipes in bottom spacer

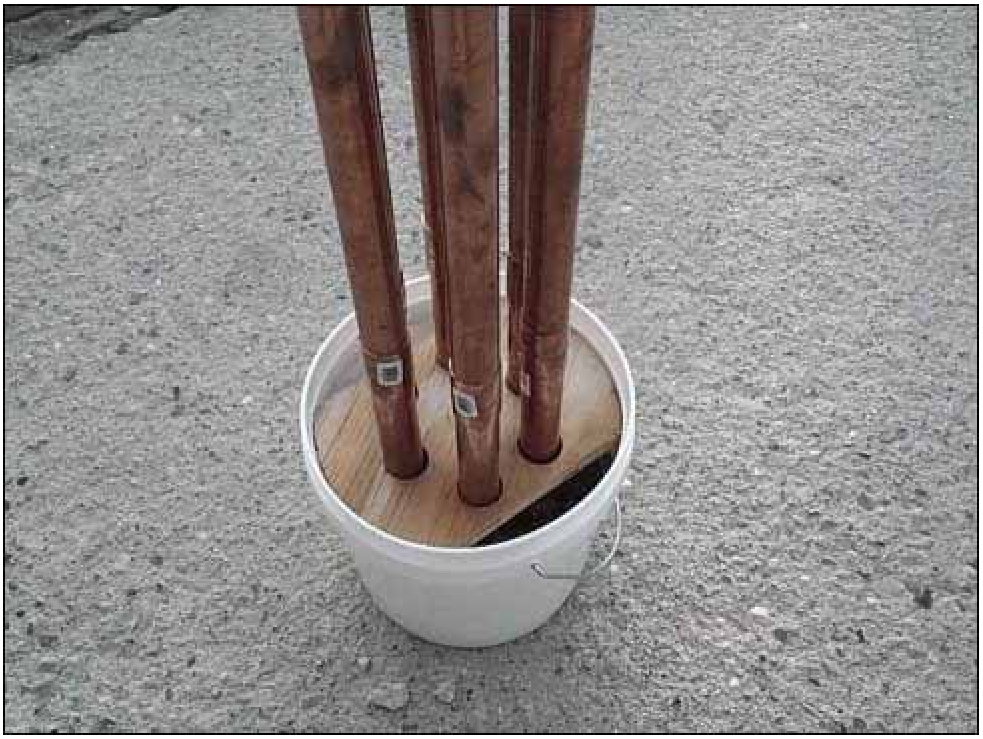

Middle Plywood Spacer

CLOUDBUSTER DIRECTIONS

BUCKET: Base: 9" in dia., 9" deep. I use two-gallon plastic buckets from the Home Depot paint department, but two-gallon food buckets, available from restaurants, or two-gallon spackle buckets from drywall contractors are fine. Leave the buckets on, as the handles make it easier to move the finished cloudbuster.

COPPER PIPES: Six standard Type M, 1" copper pipes, 6' long, open at one end, the capped ends submerged to within an inch or so of the bottom of the bucket before casting. For portability and shipping, you can use 12” sections of pipe in the base, adding five-foot lengths by putting couplers between the 12” and five-foot pieces. You won’t need to buy as much copper pipe this way, too.

CRYSTALS: I glued a double-terminated (two pointed ends) quartz crystal, about 2" long, into a .75” section of garden hose, (or use electrical tape) then, glue to the inside of each of the six copper end caps, pointing in the same direction as the pipe will be after the cap is glued on. I use 'Goop' glue. Tape, or solder around the cap where it joins the pipe to prevent resin from leaking into the cap during casting. (Note) Another option is to add one citrine gemstone to each pipe before adding the crystal, this smoothes the CB’s energy and helps to transmute negative energies. However, citrine does not seem to make the CB any more effective at eliminating chemtrails.

METAL PARTICLES: Get some metal particles from a machine shop, recycling yard or a place where they saw a lot of aluminum (sign shops and aluminum fabrication shops are good places to look). I use aluminum for weight considerations, but ANY metal will do. Very fine particles (such as filings from a key-making machine) aren’t quite as good as ones that will easily go through a ½” screen are okay. There really is a wide margin of suitability! .

PLYWOOD: Template 1 - 3/4” exterior plywood is used to make the base for the copper endcaps/pipes. I arrange the six pipes evenly around a 2 1/2” radius circle, which makes the pipes 2 1/2” apart on centers. A flat, 1 1/4” auger bit drills a suitable hole for each end cap.

Template 2 - Make another piece to fit inside the rim of the bucket. Remember to cut a section away so you can fit your fingers under it to remove it from the bucket. This piece is not part of the finished CB, just used to get the spacing for the pipes right during construction. Cut the holes in this with a 1 1/8” flat bit so that it will fit well over the six pipes. This piece will hold the pipes in position while the second batch of resin hardens.

Template 3 - I make a circular plywood spacer for the top ends of the pipes, 4” radius with 1 1/8” holes on the same 2 1/2” radius circles for the pipe ends. This keeps the pipes parallel with each other and fairly rigid.

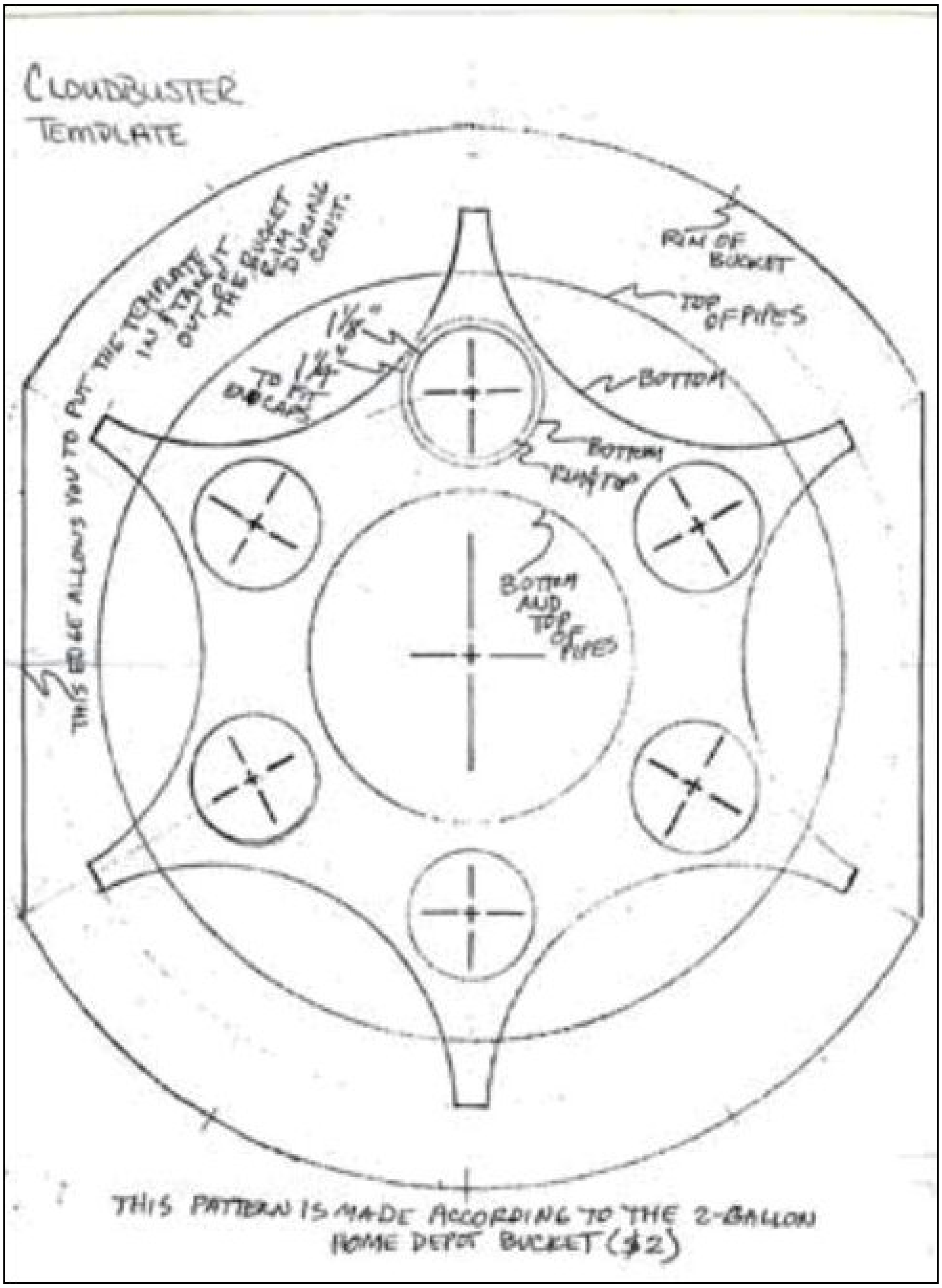
CLOUDBUSTER
TEMPLATE
THIS EDGE ALLOWS YOU TO PUT THE TEMPLATE IN STRAIGHT OUT OF THE BUCKET RIM DURING CONST.
RIM OF BUCKET
TOP OF PIPES
1 1/8"
TO FIT 1 1/4" ENDCAPS
BOTTOM
BOTTOM
RIM TOP
BOTTOM AND TOP OF PIPES
THIS PATTERN IS MADE ACCORDING TO THE 2-GALLON HOME DEPOT BUCKET ($2)

Not to scale Widest circle is 9 1/8”

RESIN: I use polyester resin (used to build fiberglass boats) but epoxy also works, as does envirotech resin from www.eti-usa.com (this sets up very quickly, so you need to be proficient with it). (If you are going to construct many of these units, try to purchase your resin in 5 gallon buckets over the internet, possibly from a surfboard or boat manufacturing supply company. Home depot gets $24 per gallon for resin)

On a level surface, pour an inch or so of catalyzed resin into the bottom of the bucket and stir in metal particles until the surface is uniform and flat. CAUTION: Resin fumes are flammable; use in a warm, but well ventilated area.

The ratio that seems to work is approx. one part resin to one part metal particles.
Put the base Template 1 onto that before it hardens, so that it just touches the surface and doesn't

sink in much. Pour a half gallon of catalyzed resin into the bucket after the six pipes are set into the holes in the base template. Use the dowel to stir in handfuls of metal particles in the same ratio as before. Note that the wood Template 1 should not be pushed down to the bottom of the bucket with the resin squishing over the top.

Carefully move the Template 2 down on to the six pipes until it is snugly within the rim of the bucket. Put the top Template 3 onto the top of the pipes and move them until the pipes are parallel. After the resin has hardened, remove the second template and set it aside. This is only used during construction, as I mentioned.

You can finish off putting the resin/metal mixture in now in two stages if you want, up to within a half inch of the rim of the bucket. Now it’s finished (wait until it hardens before you move it. You will actually use about one gallon of metal shavings and 1 ½ gal. Of resin.

A MONEY SAVING SUGGESTION, ALSO FOR BETTER PORTABILITY:

Cut six pipes, 12" long, and put the crystals/endcaps on them for the base. Buy six coupler joints which allow a pipe to come into them from both directions and stop at the middle. This allows you to make the rest of the pipe assembly from three pipes, 10' long, cut in half. The finished product leaves you with only one piece of pipe, four feet long leftover.

When positioning the 12" pipes for pouring the second layer of mix, temporarily tape the top ends of the 12" pipes so no metal particles are dropped in them by mistake.

After the mix is made, take off the tape and push the rim plywood piece down onto the pipes so that it fits into the rim of the bucket.

Put the spacers (soldered if desired), then the 5' pipes onto the 12" base pipes, then put the third (top) plywood piece onto the 5’ copper pipes.

The pipes will look skewed, most likely, so just twist the top plywood piece around until the pipes line up parallel. As you do that, the rim plywood piece will move around. You need to do this expeditiously, though not in a hurry, so that the resin doesn't harden before you're done.

We experimented with 5' extensions to the pipes to increase the range. It had a strong effect on the moon and on the upper atmosphere of the planet for hundreds of miles. We also apparently shut down the weather warfare facility under Homestead Air Force Base by leaving the extended cloudbuster pointing at the dead orgone field above it for a week.

We found that we don't need to point it in any particular direction in order to clear the atmosphere of bad orgone, and the longer it's left standing, the nicer it feels for a longer distance from the cloudbuster (many miles). We just stand it on its bottom, pointing straight up. We did once eliminate a belt of smog which was just off the coast by aiming at it for a few minutes, so direction can be a factor.

Cloudbusters destroy chemtrails consistently. It may take a few days for your cloudbuster to activate the environment enough to do this, but after that point the spew will disappear within seconds of leaving the UN jets and dissipated trails that drift within range of the cloudbuster will also disperse, but more slowly. The ones that are made up of solids take longer to dissipate, but we’ve found that these only make up about 10% or so of the spewing episodes.

Dr. Reich's cloudbusters were extremely dangerous when in operation since he had no orgone generator attached to them. Bad orgone does carry a charge which can quickly drain one's life force away when it's concentrated in one spot. It's similar to a strong positively charges static electrical field. When cloudcover is unproductive or unbalanced (too much lightning, for instance, or too much rain) the cloudbuster busts it up.

It also may help stop drought. Overall, it just balances the weather, I think it sort of mediates between the ground and the atmosphere somehow. The one in Namibia, Africa finally brought heavy rain to the desert there. It happened 9/5/01, and continued for several days. Gert, the fellow who made the cloudbuster and reported the occurrence, told me that this had simply never happened in the Namib desert before to the best of anyone's knowledge. It's arguably the driest region on the planet.

WARNING: If you touch the pipes when they're drawing in strong unbalanced or dead orgone and don't put your hand on the base after that, you may experience headaches, even into the next day, as a friend of ours did recently.

Mind you that our efforts, though somewhat pioneering, should be seen mainly as a springboard for further developments rather than the last word.

SUGGESTION: If you want even more rain, try putting water on top of the material in the bucket and/or hosing down the whole apparatus daily. In freezing climates, cover the top of the six pipes so water can't freeze and shatter the crystals. The CB appears to effect chemtrails equally well whether it's placed indoors or out. Up to seven crystals can be placed in each pipe to improve performance.

CB plowing through chemtrails in the midwest

Additional Information from Ken Adachi

Concering discharge of a CB. The origanl CB as built by Reich could channel both positive orgone and negative DOR and DOR is the problem when touching the pipes. Don Croft's modification for all practical purposes eliminates 99% of the potential dangers of DOR , but he does report that some people have gotten a headache by touching the pipes of an active unit (meaning it's been set up outdoors for a few days and pointing up) without grounding the energy by touching the base of the unit where the orac is located.

Concerning rain making. You don't have to do anything special. Simply set the unit up and give it a few days or a couple of weeks. One technique found to work was to pour water onto the base of the pipes (not into the pipes) and allow it to puddle on the top of the orac (encased in resin within the bucket). Orgone has a tremendous affinity for water and the presence of water on the orac seems to stimulate the production of rain clouds.

Below is Don Croft's chat list for the chembuster CB, HHgs info, etc. Very active site with many fascinating postings. You can read all of the previously posted messages about the success others have been having with Don's CB ref. breaking chemtrails and producing rain, etc.. Maybe you could post it to the Energy 2000 chat list and gain their attention faster than I.

cloudbuster newsgroup yahoo website

Anyone can read the previously posted messages. If you want to join and contribute to the chat, you have to agree to either be in the process of building a CB or have already built one. He doesn't want armchair 'experimenters' spouting off without hands-on experience. I can understand that.

Free Newsletter:

All information posted on this web site is the opinion of the author and is provided for educational purposes only. It is not to be construed as medical advice. Only a licensed medical doctor can legally offer medical advice in the United States. Consult the healer of your choice for medical care and advice.

Educate-Yourself

The Freedom of Knowledge, The Power of Thought

Further URL Links for further and related subjects at Educate yourself organization

Current News | Introduction | Colloidal Silver | Chemtrails | Emerging Diseases | Forbidden Cures | Immunity Boosting
Nutrition | Mind-Body Connection | Ozone | Bioelectrification | Story on Drugs | Vaccine Dangers | Newsletter | Links
Coming Attractions | New World Order | NWO News | Pam Schuffert | Reports From Peru | Phil Schneider | Al Bielek
Trevor James Constable | Mind Control | Brice Taylor | Ted Gunderson | Free Energy | Membership | Help | Discussion
Light Therapy | 2003 | Anna Hayes | Gary Wade | Tone Generator | Home | Contact Us | Products | Search | Letters

© Copyright 2002 Educate-Yourself.org All Rights Reserved.

Send any questions or comments to Educate-Yourself

New Information is being added to this website constantly. You may wish to bookmark it now.

Some additional notes and info from Graham Snook another cloudbuster builder

graham snook

Here are some photos of cloud busters The following photograph of a cloudbuster and is made of wood and plastic pipes, this is not my best subject i can make things but putting it down on paper is a problem for me, so here we go

They are 16 tubes 1"1/2 inch wide set in a wooden box with an accumulator section at the bottom where the end has two pipes just fit in the end, one pipe has a water pipe fitted the comes from the water mains and runs when in use it works well and the pond is full of life and i site around the pond for hours gives me a nice feeling as well as busting clouds and moving weather fronts away, its not a good idea to keep it on for long periods you can rewrite this if you wish i hope you get the idea of this type of cloud buster.

This cloud buster is of metal construction and has three metal pipes with the end sealed up and four slots cut in the pipe about four inch from the end or so about 1 inch wide by 8 inch then you make the end into a capacitor Cut some thick plastic the stuff used under concrete you can get it from the builders merchants, and then get some aluminium foil at the end wrap a piece of plastic round to cover the slots that you have cut in the ends then aluminium foil then plastic and keep going until you do this about 20 time then wrap it in duct tape to keep it all together make sure its tight the one in the photo can turn 360 degrees and move up or down its also can change a weather front over 1000 miles the tubes are insulated from the mount as well

A bit of a close up of the cloud busters if you need and more info let me know

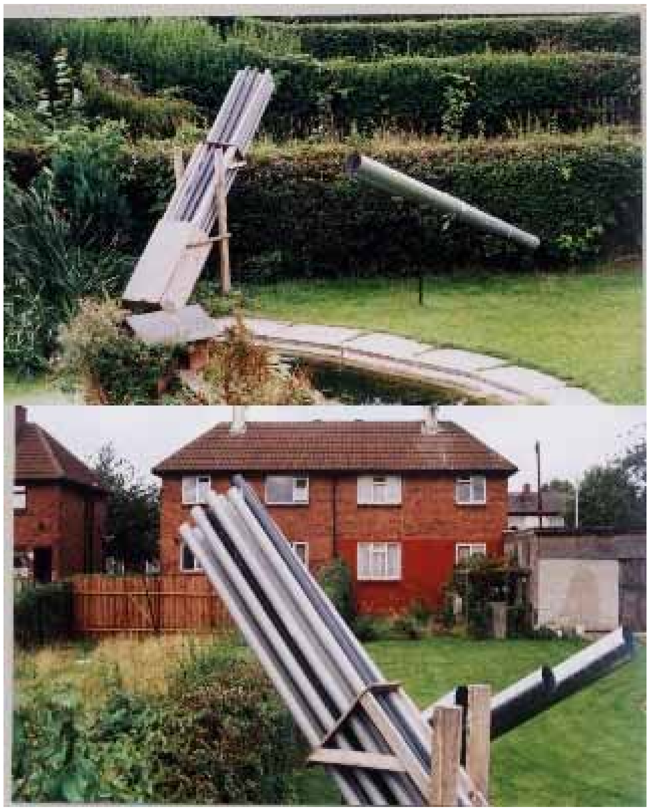

all the best graham snook

☐

Subscribe to energy2000

enter email address YAHOO! Groups Join Now!

Powered by groups.yahoo.com

LINKS

Experiments in non conventional energy CD
click for ordering details here

ENERGY 2000

Powered by FortuneCity,
Get your free website today!

FortuneCity

web hosting • domain names
web design • online games

Metaphysical Power Tools
Advanced Radionics Orgone Magick Life Force Energy Technologies
www.Metaphysical-Tools.com/

Don Croft Zapper
Don's Best Selling Zapper is Now On Sale.
YourHealthByDesign.com

Energy You Can FEEL
Reiki Orgone Generators make energy that you can FEEL + EMF Protection
ReikiOrgone.com

Electronic Projects for Beginners

F. G. RAYER, T.Eng. (CEI), Assoc. IERE

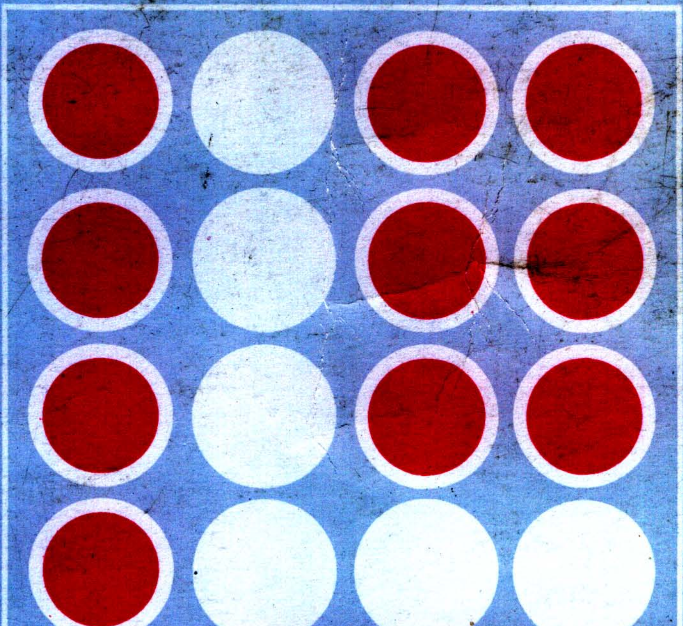

ELECTRONIC PROJECTS
FOR BEGINNERS

ALSO BY THE SAME AUTHOR

No. BP32	How to Build Your Own Metal and Treasure Locators
No. BP37	50 Projects Using Relays, SCRs & Triacs
No. BP39	50 (FET) Field Effect Transistor Projects
No. BP43	How to Make Walkie-Talkies
No. BP57	How to Build Your Own Solid State Oscilloscope
No. BP67	Counter Driver and Numeral Display Projects
No. BP75	Electronic Test Equipment Construction
No. BP79	Radio Control for Beginners
No. BP84	Digital IC Projects
No. BP90	Audio Projects
No. BP93	Electronic Timer Projects
No. BP97	IC Projects for Beginners

ELECTRONIC PROJECTS FOR BEGINNERS

by

F.G. RAYER, T.Eng.(CEI), Assoc. IERE

BERNARD BABANI (publishing) LTD
THE GRAMPIANS
SHEPHERDS BUSH ROAD
LONDON W6 7NF
ENGLAND

PLEASE NOTE

Although every care has been taken with the production of this book to ensure that any projects, designs, modifications and/or programs etc. contained herein, operate in a correct and safe manner and also that any components specified are normally available in Great Britain, the Publishers do not accept responsibility in any way for the failure, including fault in design, of any project, design, modification or program to work correctly or to cause damage to any other equipment that it may be connected to or used in conjunction with, or in respect of any other damage or injury that may be so caused, nor do the Publishers accept responsibility in any way for the failure to obtain specified components.

Notice is also given that if equipment that is still under warranty is modified in any way or used or connected with home-built equipment then that warranty may be void.

©1978 BERNARD BABANI (publishing) LTD

ISBN 0 85934 054 6

First Published – August 1978
Reprinted – September 1980
Reprinted – June 1982
Reprinted – January 1984
Reprinted – April 1986
Reprinted – February 1988
Reprinted – April 1989
Reprinted – August 1991

Printed and bound in Great Britain by Mayhew McCrimmon Printers Ltd.

CONTENTS

Page

INTRODUCTION 7
  Conversion Table 8

SECTION 1: "NO SOLDERING" PROJECTS 9
  Theoretical Circuits 11
  Signal Tracer 13
  Multivibrator 14
  "C" Substitution Box 16
  "R" Substitution Box 18
  SCR Controlled Light 18
  One Transistor Intercom 21
  Tuned Ferrite Aerial 23
  Diode Radio 25
  One Transistor Amplifier 31
  Two Transistor Amplifier 32
  Push-Pull Output Stage 34
  Boost Speaker 37

SECTION 2: MISCELLANEOUS DEVICES 39
  Soldering 39
  One Transistor Audio Oscillator 40
  Multivibrator 41
  Glissade 43
  Morse Oscillator for Phones 45
  Two Door Siren 46
  UJT Oscillator 48
  UJT Continuity Tester 49
  UJT High Voltage Producer 50
  Signal Blinker 51
  Mains Wiring 52
  Dimmer/Heat Control 53
  Two Power 55
  Sensitive Touch Switch 56
  Water Level Indicator 58
  Optical Isolator 60
  LED Indicator 61
  "Magic Candle" 62
  "Nerve Tester" 63

Page
Burglar Alarm 64
Transistor Tester 65
Treasure Locator 67

SECTION 3: RADIO AND AUDIO FREQUENCY 71
MW Booster 71
Midget MW Diode Radio/Tuner 74
One Transistor VHF Receiver/Converter 75
Morse and Sideband Resolver 78
Easy Preamplifier 81
741 Preamplifier 83
Tone Control for Your Discs Etc. 85
Two Channel Mixer 86
¼ Watt IC Amplifier 88
Class A Amplifier 90
Two Stage Amplifier 92

SECTION 4: POWER SUPPLIES 95
Easy PSU 95
Easy Regulated PSU 97
Multi-Voltage Regulated PSU 98
One Transistor Adjustable PSU 99
Zener Regulated Supply 102
Motor PSU with Reverse 103
Wrong Polarity Protector 104
Transistor Lead-Out Positions 105

EQUIVALENT SEMICONDUCTORS 107

Page 7 & 8 missing -------->

SECTION I
"NO SOLDERING" PROJECTS

With these, connections are generally made with small terminals, or nuts and bolts. By this means components are connected together, and fixed to an insulated board or panel.

A project can be built using one size of bolt and nut throughout, but it will be found convenient to have two or three sizes. Small nuts and bolts can then be used for connections between components, while larger bolts with extra nuts, or terminals, can be used to connect external items, such as headphones.

For one size throughout, 6ba screws or bolts will be satisfactory, about ½ inch long. Only a few are needed for each project, but they are quite inexpensive in 50 or 100 packs. Some extra nuts will prove useful, and also a quantity of 6ba washers. If the insulated board or panel is not too thick, the ½ inch bolts with an extra nut each will do as terminals for connecting external items.

A 4ba bolt is *larger* than 6ba, and a few of these, ½ inch or more long, also with washers and nuts, can be used for terminals, if preferred. It is also possible to obtain terminal heads, so that bolts with nuts can be used as terminals.

The 8ba size is *smaller* than 6ba. These can be of advantage in a project where size is to be kept down, or for positions where several leads come close together (as at a transistor). Bolts, nuts and washers (8ba throughout) can be obtained together.

Naturally all sorts of nuts and bolts, to hand or obtained from constructional toys or old electronic equipment, can be used. Results are not going to depend on the sizes used, which are more a matter of convenience and appearance

There is also considerable latitude in the insulated panels or boards used to carry the components. Some projects need only a few holes in a small board. Here, it is convenient to use 1/16 inch thick or similar paxolin sheet, cutting a piece of

the wanted size and drilling holes for the bolts or terminals. A small hand drill is satisfactory for this purpose. The drills used can be 3/32 inch for 8ba bolts, 1/8 inch for 6ba, and 5/32 inch for 4ba, or any similar sizes giving clearance.

Where drilling is inconvenient, a ready-perforated board must be used. For projects where size is not important, perforated 1/8 inch thick hardboard is suitable – holes are generally at 1 inch intervals, but this is not important. For this type of board, the size of the holes will generally mean that 4ba nuts and bolts will be beeded.

Plain perforated board (that is, without any conductor foil) can be obtained with holes at 0.1 inch or 0.15 inch spacing but these holes are too small and too close together. However, such boards can be used instead of plain unperforated paxolin sheet by drilling the holes larger in just those places where bolts must be placed.

There is also a type of eyelet board, with holes at 5mm spacing and just large enough to take 8ba bolts. This is handy for small projects.

It is also relatively easy to prepare a fully perforated board or panel, which can be used for a number of projects in turn.

For a board up to about 4 x 6 inch paxolin 1/16 inch thick is adequate, with 1/8 inch sheet for larger boards. It is also possible to use 1/8 inch or similar PVC or other transparent sheet, which allows connections to be seen on both sides. Plain hardboard, or thin plywood, is also satisfactory. Hardboard or wood should be given a coat of shellac or similar varnish, before use, to keep out damp.

Where size is not important, the board can be drilled with holes spaced 1 inch from each other throughout. Do *not* mark off a grid with pencil lines, as this may cause leakage between adjacent terminals. An easy method is to mark such a grid on paper, secure this over the board or panel with tape, and drill all the holes. Make these of an appropriate size for the terminals or bolts.

For small items, the board will need to be prepared to suit.

But where a board is going to be a general purpose support for several projects in turn, it can be much larger. It is also handy to support it on two strips of wood, about 1 inch high, so that there is clearance underneath for bolts and connections.

It will be found that neat and efficient projects can be assembled on such boards, and that it is easy to check connections or components, or to make changes for experimental or other purposes.

Theoretical Circuits

A theoretical circuits shows an electronic device in simple form. Capacitors, resistors, transistors and other items are shown by their own symbols. Lines representing connections are drawn from symbol to symbol. The layout of actual components in the device may not be the same as the positions of the symbols in the theoretical circuit.

It will be found that it soon becomes easy to follow a theoretical circuit. An understanding of such circuits develops when they are compared with an actual layout which shows components and the connections.

The Signal Tracer in Fig. 1 is an example of this. The main item here is the transistor, which has emitter (E), base (B) and collector (C) wires. For the transistors shown, these wires come from the base as in the diagram, which is of the underside of the transistor. (Emitter, base and collector wires may be in different positions in other transistors.) It will be seen that the emitter E is connected to negative at the battery, and to a clip on a lentth of flexible wire.

The headphones – or single earpiece – go between collector C and battery positive. A headphone jack, marked P, is used to take a plug for the phones. Generally, an on-off switch would be present in one battery connection, but this is not necessary here as the circuit can be interrupted (to switch "off") by unplugging the phones. Exactly the same results would be obtained if two separate sockets were used, to take two plugs from the headphones.

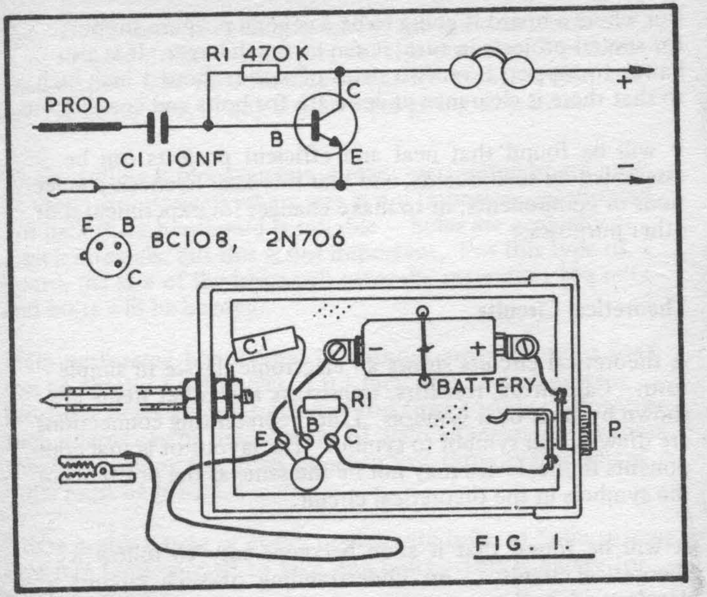

FIG. 1.

Resistor R1 is connected from collector C to base B, and it supplies a base current for the transistor, so that the latter can operate as an amplifier.

The signal to be amplified is obtained through capacitor C1. A capacitor has plates or foils separated from each other by an insulator. There is no direct current circuit through the capacitor, so placing the test prod on various points of a circuit being tested does not upset direct current working conditions.

It is clear, from Fig. 1, how the theoretical circuit is made up as a practical piece of equipment, with C1, R1, transistor, phones and battery connected together in the way required.

With later projects, the theoretical circuit and practical wiring diagram may be shown together in this way, or may appear as separate diagrams when they are larger. If the theoretical circuit and wiring diagram are compared, so that capacitors, resistors and other items on one are also found on the other, it will soon be found that a theoretical circuit is readily followed and understood. It will be noted that the theoretical circuit could be built up in numerous ways, all equally satisfactory.

Signal Tracer

This locates and amplifies audio (sound) signals which may be present in a circuit. Construction is most easily carried out in a small plastic case (Fig. 1). An insulated board about 1¼ x 2 inch (or to suit the case) has three small bolts to secure emitter, base and collector leads of the transistor. Bolts also hold two small brackets, which press on the negative and positive ends of the 9v battery. String or elastic through holes in the board helps to secure the battery.

The prod is a long 6ba or 4ba bolt, or length of screwed rod. Lock one end of C1 lead with nuts, or loop it under the bolt head. Thread the clip flex through a hole in the case, and pass the prod through another hole, as shown. A washer and nut then fix the prod, with the board inside the case. The case lid can then be fitted, and the securing nut can be screwed on the phone outlet. Connections here can be tightly twisted round the outlet tags, and compressed with pliers. A piece of insulating sleeving is put on the test prod, and a clip with fixing screw is attached to the flexible lead.

Making Connections

For all the projects in this book, any thin insulated connecting wire can be used; or bare wire with insulated sleeving. It is handy to have two or more colours – such as red for positive, and black for negative circuits.

A loop is made at the bare end of the wire, and is put under the screw head, or under a washer. Where two or more leads go to the same bolt, it is convenient in some cases to have a washer or two extra, to go between loops. Each connection must be reasonably tight, and short circuits must not be allowed between wires or loops which are shown separated from each other.

Battery and external leads are generally of flexible wire. Positive and negative battery clips, with leads already attached, are available for transistor radio and similar batteries. Elsewhere it may be convenient to arrange contacts as in Fig. 1, or to twist wires to the battery tags.

The battery polarity (positive and negative) *must* always be correct, as shown.

When dealing with transistors, it will be found that some types have long wire leads, more than adequate for easy screwed connections. Others, such as the BC108 and similar types, have rather short wires, but can be held satisfactorily with 8ba bolts fairly near eath other. (With other transistor types not used in this section, soldering may be essential for proper connections to be made.)

Circuit Tracing

To use the signal tracer, connect the clip to the earth return line of the receiver or amplifier. The prod is then touched one by one on points where the audio signal should be present, working systematically through the audio amplifier, or from the first point where the signal can be heard. Take in each capacitor, joint, or stage in turn. When signals cease to be heard, the fault lies between this point and the previous point, where signals were heard.

Do not test any mains operated or high voltage equipment, as this is a job for an engineer who will follow the necessary and essential precautions for safety.

Medium or high impedance phones are most suitable for this tracer. A 600 ohm DC headset will be ideal. The tracer can be utilised as an amplifier for a crystal diode radio receiver

Multivibrator

You will be able to use this as an audio oscillator for numerous test purposes, or for Morse Code practice, or as a signal device in place of a bell or buzzer.

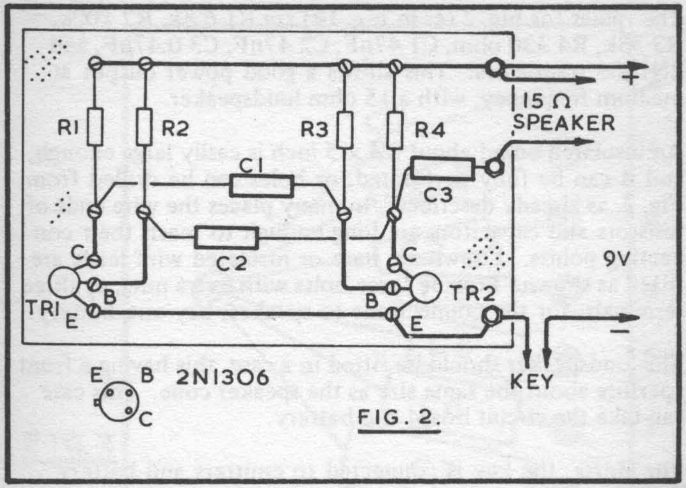

FIG. 2.

Fig. 2 shows the practical assembly of the device on an insulated board. A theoretical circuit is given in Fig. 14, and is not repeated here, so Figs. 2 and 14 should be looked at together.

R1 is the collector load resistor for transistor TR1. Collector current passes through this resistor, and an audio signal voltage is developed across it. Resistor R2 supplies base current for TR1.

TR2 has a similar collector resistor R4, and base resistor R3. Capacitor C1 couples the signal at the collector C of TR1 to the base B of TR2, while C2 similarly couples TR2 collector to TR1 base. Amplification in the transistors and feedback by the capacitors in this way result in the circuit oscillating. That is, each transistor goes rapidly in and out of conduction in turn, so that a voltage at audio frequency is produced across the collector load resistors. C3 is a coupling capacitor, allowing the signal voltage at the collector of TR2 to be taken to a loudspeaker or headphones.

With a circuit of this type, the component values, transistors, and battery voltage will not be critical. However, resistor values, capacitor values, battery voltage, transistors, and the loudspeaker or other sound-reproducing device will all have an influence on the frequency of oscillation, or tone produced.

The values for Fig. 2 (as in Fig. 14) are R1 6.8k, R2 100k, R3 33k, R4 330 ohm, C1 47nF, C2 47nF, C3 0.47μF, and 2N1306 transistors. This allows a good power output at medium frequency, with a 15 ohm loudspeaker.

An insulated board about 2½ x 5 inch is easily large enough, and it can be fully perforated, or holes can be drilled from Fig. 2, as already described. In many places the wire ends of resistors and capacitors are long enough to reach their connecting points. Elsewhere, bare or insulated wire leads are fitted as shown. Provide three bolts with extra nuts, or three terminals, for the connections to speaker, key and battery.

The loudspeaker should be fitted in a case, this having a front aperture about the same size as the speaker cone. This case can take the circuit board and battery.

For Morse, the key is connected to emitters and battery negative as shown, with flexible leads of convenient length. If the multivibrator is to be used as a warning device then these leads can instead go to a door bell push, or to a relay or SCR, used to complete the circuit, as with some later projects. In this way the multivibrator, with its speaker, can replace a buzzer or bell.

If the frequency of the multivibrator is to be changed, this can readily be done by altering the values of capacitors C1 and C2 in particular. The two capacitors need not be of the same value. Larger values reduce the frequency (lower the tone) while smaller capacitors increase the frequency. C3 also has some effect, due to loading from the speaker.

Such a circuit can be used with a battery voltage as low as 1.5v, for headphones. Transistor types are very uncritical. But if PNP transistors are fitted, the battery polarity must be reversed. The transistors shown are NPN.

"C" Substitution Box

When experimenting with circuits or checking a suspected capacitor, a box bringing into circuit a range of capacitor values may be used. In many circuit positions an exactly correct value is not required for testing by substitute, so that a very large number of capacitors need not be provided.

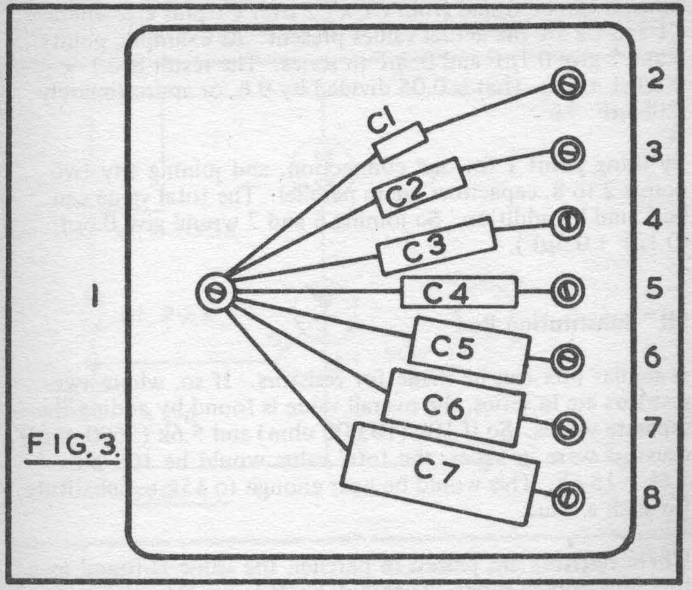

FIG. 3.

Such a box is easily made as in Fig. 3, using a plastic or other insulated case. Here, point 1 is common to all the capacitors, and any capacitor can be selected by connecting to points 2 to 8. Connections can be taken to 6ba bolts, using extra nuts. An alternative is to have eight sockets, with a pair of test leads equipped with plugs.

Suggested values are C1 100pF, C2 470pF, C2 1nF (1000pF or 0.001μF), C4 10nF (0.01μF), C5 0.1μF, C6 0.5μF, and C7 8μF. As the last capacitor will be polarised, mark its terminal to agree.

These values will be found sufficiently near for use in many coupling, decoupling, by-pass and similar circuit positions. The substitution box can be used in place of C1 or C2 in Fig. 2. But where a defective capacitor is found in a tuned circuit, it should be replaced by the exact value.

By ignoring terminal 1 and connecting leads to any pair of the terminals 2 to 8, two capacitors can be used in series. For those who like to work out the value of such combina-

tions, it can be found from C1 x C2 over C1 plus C2, where C1 and C2 are the actual values present. As example, points 6 and 7 give 0.1μF and 0.5μF in series. The result is 0.1 x 0.5/0.1 + 0.5. That is 0.05 divided by 0.6, or approximately 0.083μF.

By using point 1 for one connection, and joining any two points 2 to 8, capacitors are in parallel. The total value can be found by addition. So joining 6 and 7 would give 0.6μF (0.1μF + 0.5μF).

"R" Substitution Box

A similar box can be made for resistors. If so, where two resistors are in series, the overall value is found by adding the separate values. So if 10k (10,000 ohm) and 5.6k (5600 ohm) resistors were in series, the total value would be 10k plus 5.6k = 15.6k. This would be near enough to 15k to substitute for such a value.

Where resistors are placed in parallel, the value is found as for capacitors in series. That is, R1 x R2 over R1 + R2. So for the 10k and 5.6k resistors, this would be approximately 3.6k.

Naturally each resistor may be selected alone, as for the capacitors. If sufficient terminals are provided, a large number of resistors could be fitted. For most transistor circuits, values from about 150 ohm to 270k will fill most needs.

SCR Controlled Light

This device will switch on a light, when the level of general illumination falls below that for which triggering is set. It can thus be used for an automatic night light. Operation can be from a battery; or from a mains power supply, as described later.

Fig. 4 is the circuit, and introduces two new devices. One is the light-dependent resistor LDR. The resistance of this component depends on the strength of the light falling on it. The

resistance is low (some hundreds of ohms) with bright light, but rises to many thousands of ohms in dim light, or darkness.

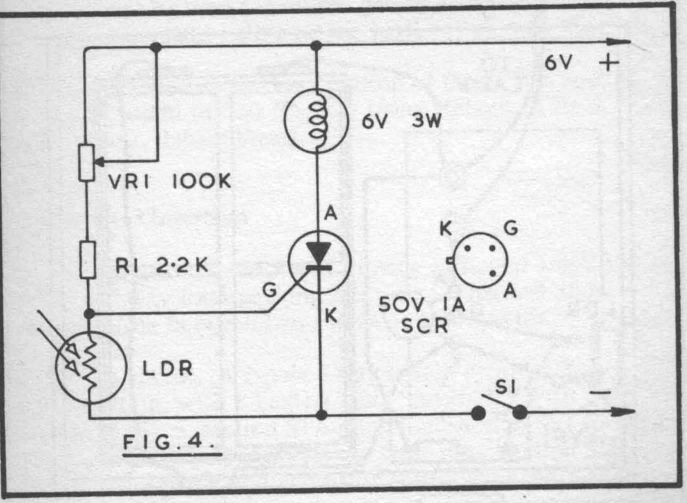

FIG. 4.

The second device is the silicon controlled rectifier SCR. Normally, this does not allow current to pass from its cathode K to its anode A. However, a small current applied to its gate G will trigger it so that it conducts from cathode to anode. It remains in this state even when gate current is removed. It is restored to the non-conducting state by removing the voltage from its cathode, by opening switch S1.

In this device the SCR is controlled by the LDR. With illumination present the resistance from gate to cathode is low, so that when S1 is closed, current does not flow from the anode circuit through the 6 volt 3 watt lamp. As illumination falls, the resistance of the LDR rises. This continues until a sufficient gate potential is present to operate the SCR, which moves into its avalanche or conducting state, lighting the lamp. The potentiometer VR1 enables the resistance from positive to gate to be adjusted, to set the illumination level at which the circuit operates.

A 50v 1 amp SCR is more than adequate for this circuit, and it resembles a large transistor externally, and can be mounted by its wire ends.

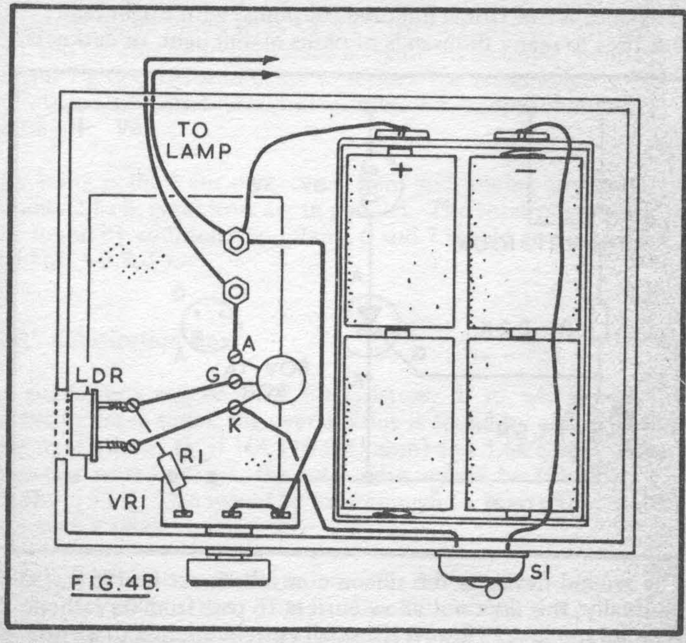

FIG.4B.

Fig. 4B shows constructional and wiring details. The SCR is mounted on a small insulated board. Clear bared wires are twisted tightly to the LDR, so that this also can be fitted by small bolts and nuts, as shown. The connections to VR1 are also tightly twisted and pressed together with pliers.

A battery holder for four 1.5v cells is shown, and connections to this can be by means of clips which have leads attached. The switch S1 is the small, low-voltage type having small screws, as used for toy lighting circuits and similar purposes.

A hole is cut in the box level with the LDR, which may be directed towards the window of the room, or positioned as found to give the results wanted. The bulb, in holder, can be positioned on top of the case. If light from it reaches the LDR this will not stop the circuit working, as the SCR can only be returned to the non-conducting state by opening the switch S1,

For an automatic night light or similar device, a case can be made from thin wood, with opening back or bottom. The device can also be used for various games, where momentarily interrupting the light falling on the LDR switches on the bulb.

If a detailed reference to the operation of the SCR is wanted, this can be found in "50 Projects Using Relays, SCRs & Triacs" (BP37, Babani Press).

One Transistor Intercom

A single high gain transistor will provide sufficient amplification for a 2-way loudspeaking intercom. This will allow communication between two rooms or other points.

Fig. 5 is the circuit. A 2-pole 3-way switch provides 2-way communication, with a central "Off" position. With the switch as in Fig. 5, section S1A takes loudspeaker SP1 to the base coupling capacitor C1. At the same time, section S1B connects the battery and transistor collector circuit to the

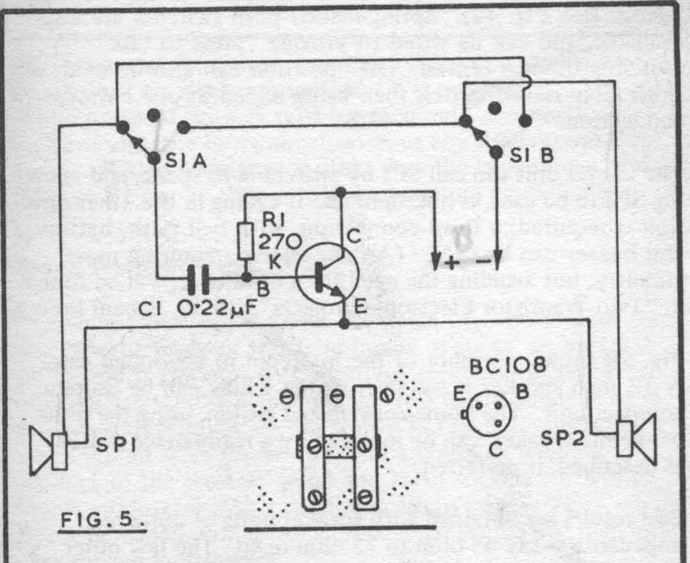

FIG. 5.

speaker SP2. SP1 then acts as a microphone, and sounds are reproduced in SP2.

With the switch in the opposite position, S1A selects SP2 as microphone, and amplified signals pass via S1B to SP1, so that speech can be in the other direction.

The switch, amplifier, battery and SP2 are assembled as the master control unit, and twin leads run to SP1. The person at SP1 should say "Over" so that the other person can operate the switch, and he in turn should say "Over" to signify that SP1 is again acting as microphone.

A simplified home-made change over switch is shown in Fig. 5. This is two strips of metal, pivoted so that they can be swung from side to side, and bearing on two of the three round-headed screws. A strip of insulating material is fitted so that both strips move together. In Fig. 5, this switch is in its "Off" position.

Ready made 2-pole 2-way switches can be obtained and used instead (see Fig. 49). Spring-loaded push switches are also available, and can be wired to provide "press to talk". A switch without a central "Off" position can also be used, a single pole on-off switch then being added in one battery connection.

The master unit can call SP1 by switching to speak, and allowing SP2 to be used as microphone. If calling in the other direction is required, a third connection, with bell push, battery and buzzer can be used. (An alternative, requiring more circuitry, but avoiding the need for a third wire, will be found in "Two Transistor Electronic Projects", (BP30, Babani Press.)

Fig. 5B shows assembly of the intercom in a wooden case. A 3½ inch speaker is suitable, so the whole can be quite a compact unit. The home-constructed switch, using the bolts below the speaker, can be replaced by a ready-made switch, as described, if preferred.

Best results are obtained with speaker units of quite high impedance – say 45 ohm to 75 ohm or so. The few other components are assembled on a small insulated board, as for earlier circuits.

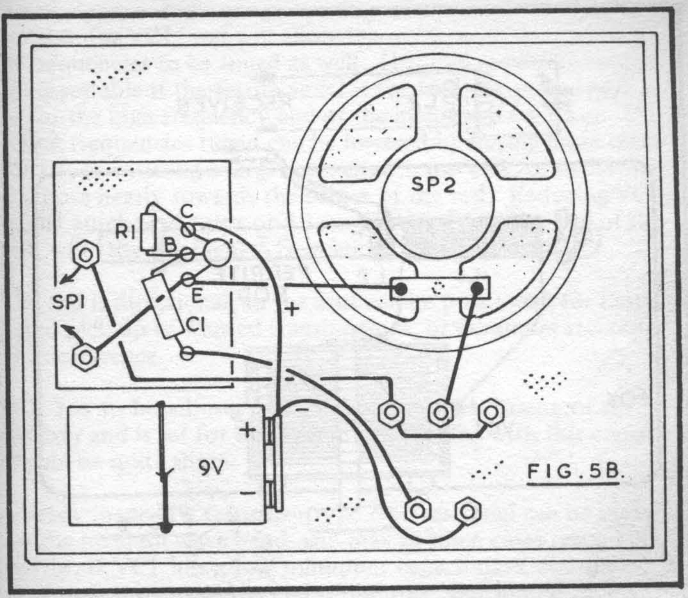

FIG.5B.

The speaker at the remote point also requires a small case. Dimensions are not important, and it is possible to utilise plastic boxes of suitable type for each unit. Connections between units are by twin bell wire, or any similar twin conductor. With a transistor of other than the type shown, the value of R1 can be chosen to give maximum amplification.

Tuned Ferrite Aerial

This simple device will allow "portable" reception with the old type of receiver which normally requires an external aerial, and it can also give improved reception in terms of freedom from interference on the 160 metre amateur band and adjacent frequencies. As shown in Fig. 6, it consists of a tuned circuit with ferrite rod, and a coupling winding to connect to the receiver aerial and earth sockets. It will be found very useful with a communications type receiver, where a long extended outdoor or similar aerial results in a high level of noise and interference.

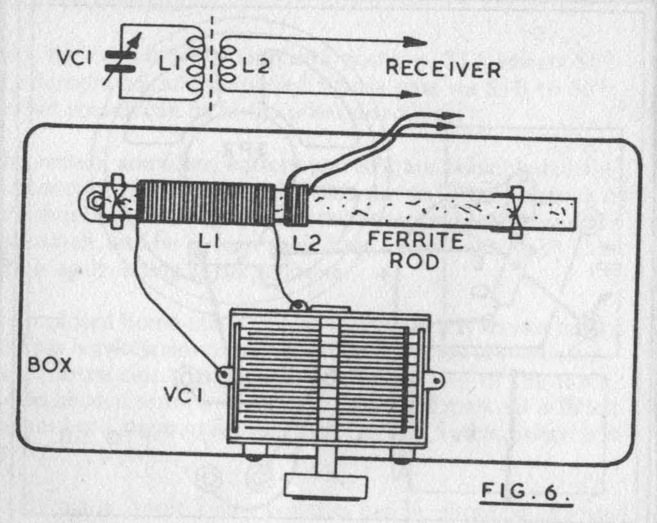

FIG. 6.

The device is constructed in a plastic lunch box about 6 x 4 x 4 inch or other *insulated* case. Some boxes of this kind, though strong, are of brittle material. These must be drilled without too much pressure, or they may crack. Drill points should also be sharp.

The rod is 6 inch long and 3/8 inch in diameter, and it is supported by two strips of insulating material, which are fixed with brackets. Each strip has a V-shaped notch, so that the rod can be held with thin string. L1 is the tuned winding, and it had sixty turns of 24 swg enamelled wire, side by side. Connections can be made to the frame (moving plates) and fixed plates of the capacitor VC1 by means of small bolts, or may be lightly twisted.

L2 is the coupling winding. Many communications type receivers have a low impedance input circuit, and twelve turns will be adequate for L2. The same wire can be used as for L1. With other receivers it may be worth while to experiment with the number of turns on L2, for best reception.

An air-spaced single gang capacitor of about 315pF is most suitable for VC1, and will allow the usual medium wave band of frequencies to be tuned as well. However, a smaller capacitor is suitable if the ferrite aerial is for 160m reception only, or for the high frequency end of the medium wave band. The actual frequencies tuned can be lowered by having more turns for L1, or by using a larger capacitance for VC1, or by having L1 more nearly towards the centre of the rod. Reducing VC1, or the number of turns on L1, or moving L1 to the end of the rod, will raise the highest frequency tuned.

The rod is directional, so the unit can be positioned for best signal pick up of wanted transmissions, or minimum reception of interference.

VC1 has to be adjusted in conjunction with tuning of the receiver and is set for best reception. Tuning with this control should be quite sharp.

A ready made MW transistor type receiver aerial can be useful for the medium wave band, and may in some cases reach 160 metres, if VC1 has a low minimum capacitance, and the winding is quite near the end of the rod. But should tuning not reach a high enough frequency, some turns will have to be removed.

Diode Radio

A lot of interest can be obtained from a diode radio receiver. It requires few components, no battery or other supply, and can give good headphone reception of a few of the more powerful transmitters. It can also be used as a tuner, as shown later, and will provide excellent audio quality for local station reception.

The circuit in Fig. 7 is intended for medium wave reception. That is, from about 200 metres to 500 metres, or 1500 kilohertz to 600 kilohertz.

The coil L1 is tuned by the variable capacitor VC1. At resonance with the wanted transmission, a signal voltage is developed across the coil. Detection or demodulation by the diode D1 allows the audio component of the signal to be

recovered, and taken to the phones or to an amplifier. Such "crystal sets" were used in very many homes, in the early days of radio.

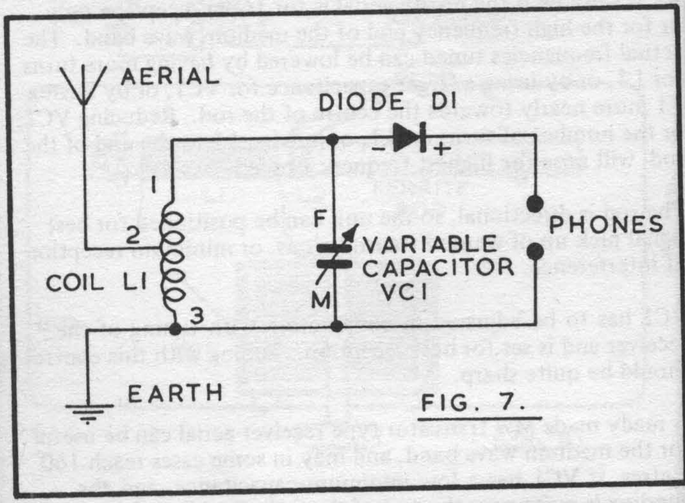

FIG. 7.

Fig. 7B shows the practical layout. It is convenient to use an insulated board larger than necessary, so that a transistor amplifier can be added later. A board about 5 x 8 inch is suggested, and it can be supported on 1 inch high runners (strips of wood) so that there is clearance for screws or terminals.

Coil L1 is wound with 26swg or similar wire, and has seventy turns, side by side. Secure the wire by passing it through two small holes, wind on about thirty five turns, and form a loop. Continue to wind in the same direction for a further thirty five turns, cut off the wire, and secure the end through two holes. The tube used can be 1½ inch to 2 inch in diameter, and can be paxolin, or can be made by winding thin card round a suitable object, smearing a little adhesive on to form a tube.

Point 1, the beginning of the coil, is taken to the fixed plates tag F of the variable capacitor, which is also connected to diode D1. The tapping 2 goes to the aerial terminal. End 3 is taken to the moving plates connection or metal frame of the capacitor. Ends 1 and 2, and loop 3, are left long enough to reach these points, and the insulation is carefully scraped off the wire where connections are to be made. The capacitor VC1 can best be about 500pF, but this exact value is not essential.

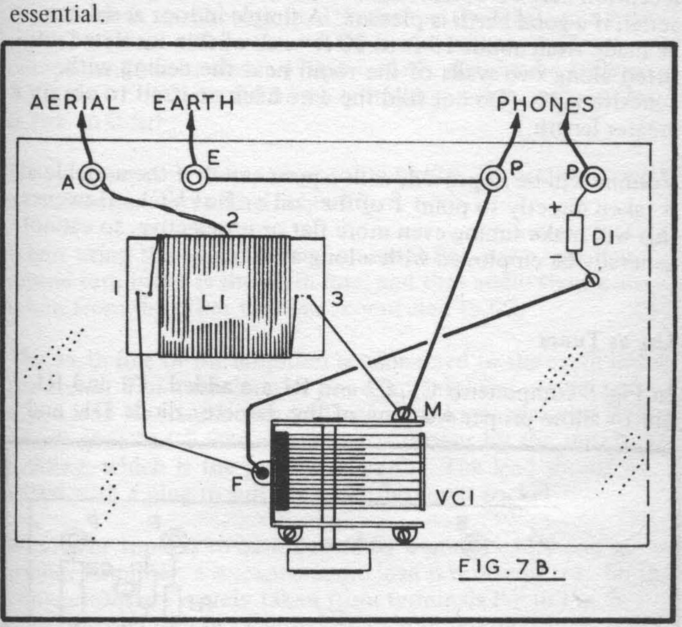

FIG.7B.

Four terminals, for Aerial, Earth, and Phones, will complete the receiver. Best results will be had by connecting a good headset, of about 500 ohm to 2,000 ohm. Very low impedance earpieces, such as are employed with some transistor receivers, are not suitable here.

With a simple receiver of this kind it is necessary to employ a reasonably effective aerial and earth. Only in some circumstances will volume be great enough with no earth, or with a short aerial. The earth connection can be taken to a metal spike in the ground, or to a metal object buried in soil, or to any other available earth other than a mains circuit earth. (No connection must be made to any mains plug or other mains circuit.) A cold water pipe may provide a good earth. Gas pipes should not be used.

For the aerial, a length of aerial wire some 30 ft to 75 ft or so in all may form the aerial itself and lead-in. If this can be suspended quite high, and clear of buildings or earthed objects, signal strength will be improved. However, satisfactory reception may be obtained from an indoor aerial, or shorter aerial, if a good earth is present. A simple indoor aerial can be made from about 15 ft to 20 ft or so of thin insulated wire, fitted along two walls of the room near the ceiling with adhesive tape. (Do not fold the wire back on itself to obtain a greater length.)

Volume will be improved, with a poor aerial, if the aerial lead is taken directly to point 1 of the coil or F at VC1. However, this will make tuning even more flat or unselective, so cannot generally be employed with a long aerial.

Use as Tuner

In Fig. 8 components C1, C2 and R1 are added. C1 and R1 are to allow proper working of the detector diode D1, and may be considered as replacing the headphones. C2 is a coupling capacitor passing audio signals to the following amplifier.

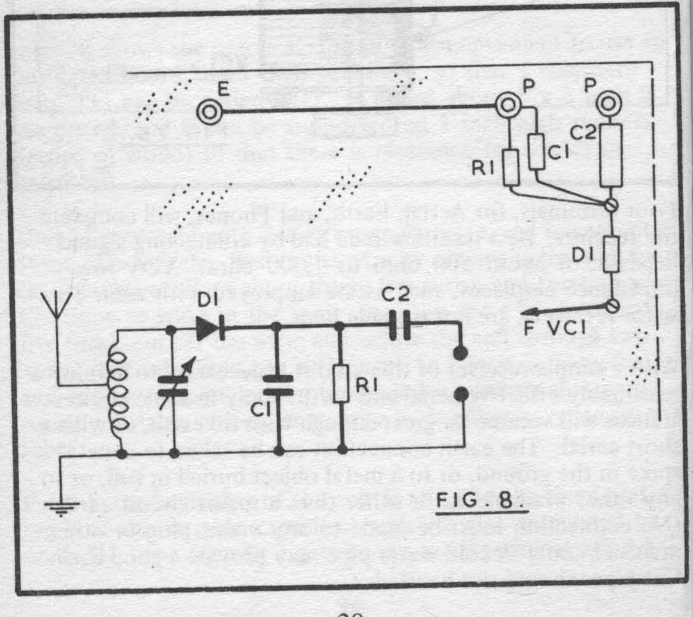

FIG. 8.

As with the previous circuit, any general purpose detection type diode is satisfactory for D1, such as the OA91 and many similar devices. The value of C1 is not critical, and this component can be about 470pF or 1nF. R1 can be 100k to 220k or so. C2 should be of quite large value, and can be 0.1μF to 0.5μF.

D1 can be moved from the position in Fig. 7B, and the few extra components added, as in Fig. 8.

When using the tuner to feed an amplifier, note that one phone terminal P is the earth line, and that audio signals are taken from the other terminal, connected to C2.

The earth line of the amplifier is connected to the earth line of the tuner. For a sensitive or powerful amplifier, a screened audio lead is recommended. This has an inner screened lead, which goes to C2, and a conductor former by the outer brading, which is the ground or earth. The lead should be fitted with a plug to suit the amplifier input socket.

Where the tuner is to be followed by a small headphone or similar amplifier, a screened audio lead is not required. So then connections are merely taken from terminals P-P in Fig. 8, remembering to join amplifier earth and tuner earth circuits as explained.

Medium and Long Wave

The addition of long wave coverage is worthwhile in some areas, and the coil can be modified as in Fig. 9 so that both MW and LW can be tuned.

A tube about 3 inch long and approximately 1½ inch in diameter will be necessary. The section L1 is wound as before. L2 consists of about two hundred turns of 34swg or 36swg wire, in two compact piles of one hundred turns each. All turns throughout are in the same direction. The junction of L1 and L2 goes to the waveband switch S1. The second

switch terminal goes to 4 and earth line. MW tuning is obtained with S1 closed, and LW tuning with S1 open.

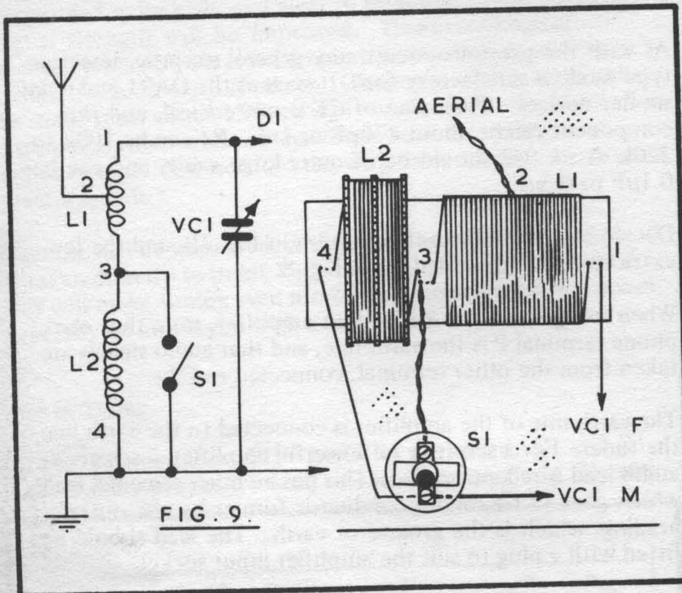

FIG. 9.

For accurate coverage of the usual MW and LW bands, and using a 1¼ inch diameter tube, L1 can be eighty-four turns of 32swg enamelled wire, with tapping 2 sixty turns from point 1. Leave a space of 5/8 inch and glue on a card disc cut to fit the tube. Also glue on two further discs, leaving spaces of ¼ inch between them. L2 is wound with 32swg double cotton covered wire, with one hundred and twenty turns in each slot (making 240 in all for L2). The coil can be fixed with two screws or small bolts, with spacers or extra nuts to hold it a little clear of the board. The windings should not be varnished, covered with adhesive, or wax.

With circuits of this kind there is no ganged tuning, or need for exact tuning of a particular frequency band. This means that coil winding details (wire gauge and covering, and diameter, or number of turns) can be varied to suit materials to hand. Ready-made coils can also be tried. Some commercially made coils with cores can give a slightly better performance than the air-cored home made type of coil. Coils can also be wound in a similar way to that in Fig. 6.

One Transistor Amplifier

Headphone volume obtained from the crystal diode receiver can be considerably boosted by adding a single transistor audio amplifier. Space for this can be found on the circuit board.

In Fig. 10, resistor R2 and the transistor have been added, with a switch S2 for on-off switching of the battery. R1, C1, D1 and C2 are already present, as shown in Fig. 8.

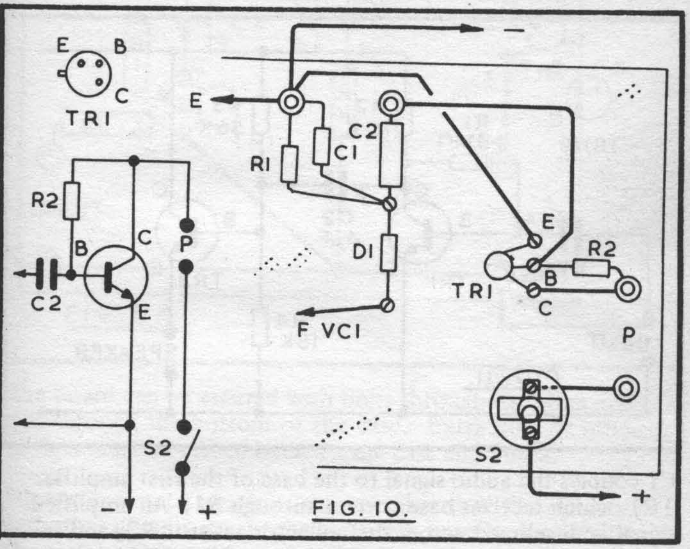

FIG. 10.

TR1, R2 and S2 are connected as shown. For TR1, a BC108, 2N706, or similar general purpose NPN transistor will be satisfactory. R2 can generally be 1 megohm with satisfactory results, but it can be worth while trying values from about 100k to 2.2 megohm here, as performance depends on the transistor, phones, and battery voltage.

Medium and high impedance phones, such as a 600 ohm DC headset, will be most suitable. The battery can be 4.5v to 9v.

If the phones are first connected to terminals P-P in Fig. 8, then taken to terminals P in Fig. 10, a very considerable increase in volume should be obtained.

If a PNP transistor is to be tried in this circuit, the polarity of the battery *must* be reversed.

Two Transistor Amplifier

Fig. 11 is the circuit for an easily made 2-transistor amplifier which can provide reasonable loudspeaker volume. The audio signal is taken to the input terminals, battery negative being the grounded side of this circuit.

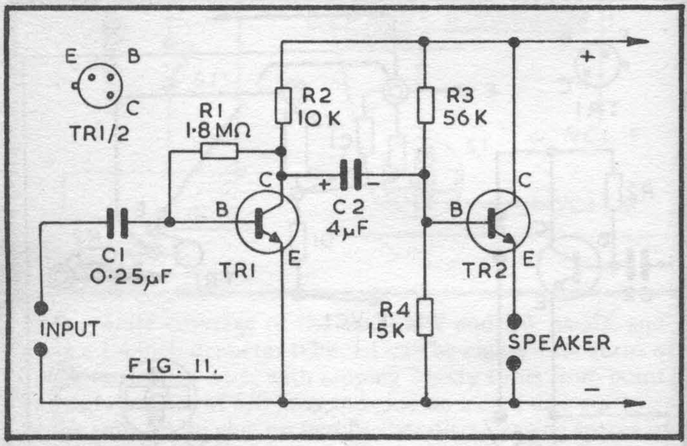

FIG. 11.

C1 couples the audio signal to the base of the first amplifier TR1, which receives base current through R1. An amplified signal is developed across the collector resistor R2, and is coupled to the base of the following amplifier TR2 by means of C2.

The two resistors R3 and R4 set the base operating conditions for TR2, and the output is taken from TR2 emitter.

Fig. 11B shows assembly of the components on a small insulated board. If preferred, connections between components can be on the underside of the board, in the way shown for later circuits. Fit red and black flexible leads for battery connections. The amplifier, with its battery, can be fixed in an insulated box or metal case. Sockets can then be used for input and output circuits, and an on-off switch can be placed in one battery lead.

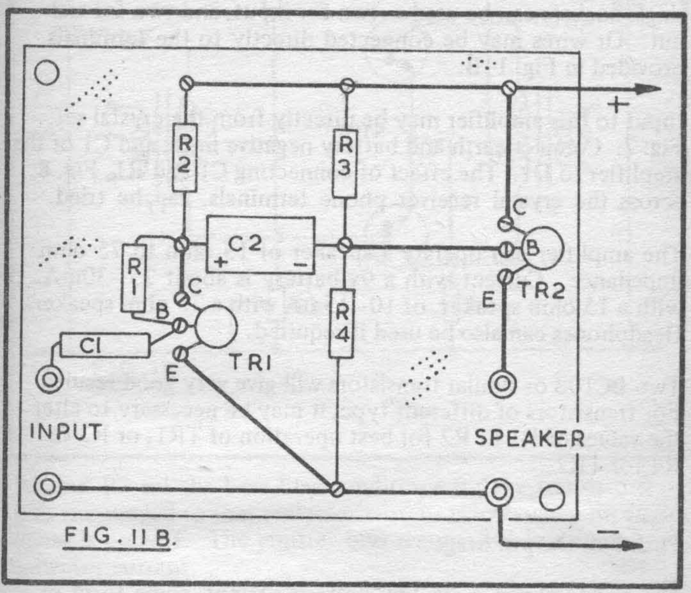

FIG. 11 B.

The board can be secured with bolts through its fixing holes, and through the bottom of the case. Extra nuts or other spacers must be placed between the case and board.

Whenever a metal box is used, remember that all connections and other circuits except the earth line must be insulated from the metal. The metal case itself is normally connected to the earth line, and a lead can be put under one of the fixing screws for this purpose. In Fig. 11, both input and speaker are returned to the earth line. So 3.5mm jack sockets of the type in which the outer or sleeve contact is common to the fixing bush could be used, as these points will be connected to battery negative and the earth line and metal box. But with some circuits it is necessary to use an insulated jack socket, or to insulate this from the metal. An example of this would

arise if the speaker were placed in the collector circuit of TR2, as the outer or sleeve circuit would then need to go to battery positive.

Instead of the type of socket which takes a jack plug, ordinary single sockets can be used – two for input, and two for output. Or wires may be connected directly to the terminals provided in Fig. 11B.

Input to this amplifier may be directly from the crystal set, Fig. 7. Connect earth and battery negative lines; and C1 of the amplifier to D1. The effect of connecting C1 and R1, Fig. 8, across the crystal receiver phone terminals, can be tried.

The amplifier can operate a speaker of 15 ohm to 75 ohm impedance. Current with a 9v battery is about 25–30mA with a 15 ohm speaker, or 10–15mA with a 75 ohm speaker. Headphones can also be used if required.

Two BC108 or similar transistors will give very good results. For transistors of different type, it may be necessary to alter the values of R1 or R2 for best operation of TR1, or R3 or R4 for TR2.

Push-Pull Output Stage

For good volume, with low battery current, some form of push-pull output stage is generally used. This has two transistors, and each amplifies one half of the audio signal. Base bias conditions of the two transistors are so arranged that only a low current flows during silent intervals. Overall battery drain is thus much lower than if a single transistor were used to obtain a similar level of volume.

Fig. 12 is a typical push-pull output stage. T1 is the input or driver transformer, and an audio signal is taken to its primary P. The secondary S is centre tapped. So for one half of the audio cycle one transistor base is driven positive and the other transistor base is driven negative, this being reversed for the remaining half cycle. Thus each transistor conducts alternately, and the centre-tapped output transformer T2 combines the collector currents, to operate the speaker connected to the secondary S.

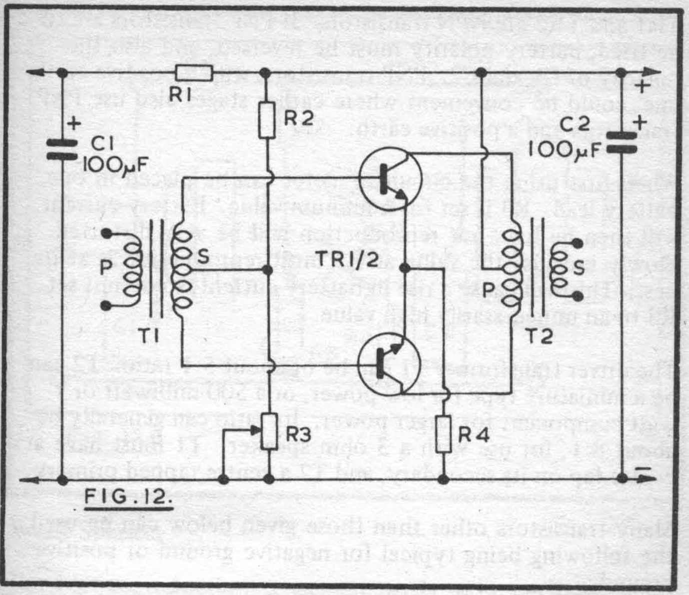

FIG. 12.

R2 and R3 set the base bias conditions for the transistors. Bias is arranged so that collector current is low, when no audio signal is present. The emitter bias resistor R4 limits peak collector current.

If R3 is a pre-set resistor, as in Fig. 12, operating conditions can be arranged to suit the transistors fitted. For best results TR1 and TR2 should be a matched pair. This means that they will have similar gain and current characteristics. However, two transistors of the same type number will generally be satisfactory.

It is assumed that the same battery will supply earlier stages of the amplifier, and current for this purpose is taken from C1 and R1. R1 and C1 prevent audio signals present in the positive line passing back to earlier amplifier stages. The by-pass capacitor C2 also helps in this respect. With simple circuits not giving much overall amplification, or small power outputs, R1, C1 and C2 might all be omitted.

TR1 and TR2 are NPN transistors. If PNP transistors are to be tried, battery polarity must be reversed, and also the polarity of C1 and C2. PNP transistors, with a positive earth line, could be convenient where earlier stages also use PNP transistors and a positive earth.

When first using the circuit, a meter can be placed in one battery lead. R3 is set for minimum value. Battery current will then be low, but reproduction will be very distorted. Slowly increase the value at R3 until reproduction is at its best. This will cause a rise in battery current, so do not set R3 to an unnecessarily high value.

The driver transformer T1 can be of about 5:1 ratio. T2 can be a miniature type for low power, or a 500 milliwatt or 1 watt component for larger power. Its ratio can generally be about 8:1, for use with a 3 ohm speaker. T1 must have a centre tap on its secondary, and T2 a centre tapped primary.

Many transistors other than those given below can be used, the following being typical for negative ground or positive ground.

Negative earth (NPN transistors). 2 x BC108. R1 1.2k. R3 100 ohm fixed or 250 ohm pre-set. R4 4.7 ohm.

Positive earth (PNP transistors). 2 x AC128 or 2 x AC142. R2 4.7k. R3 82 ohm or 250 ohm pre-set. R4 4.7 ohm.

R1 will generally be 1k to 2.2k, depending on the earlier amplifier. The primary of T1 can be connected to terminals P in Fig. 10; or alternatively to any audio source where a reasonably strong signal is present.

Where signals are too strong, an audio gain control needs to be fitted. In Fig. 8 this can be done by omitting R1, and disconnecting C2 from D1. Connect the outer tags of a 100k volume control or log pot from D1 to earth line. Take C2 to the centre tag or slider of the potentiometer. Any required audio level can then be taken off by adjusting the control knob.

Fig. 12B shows the board layout.

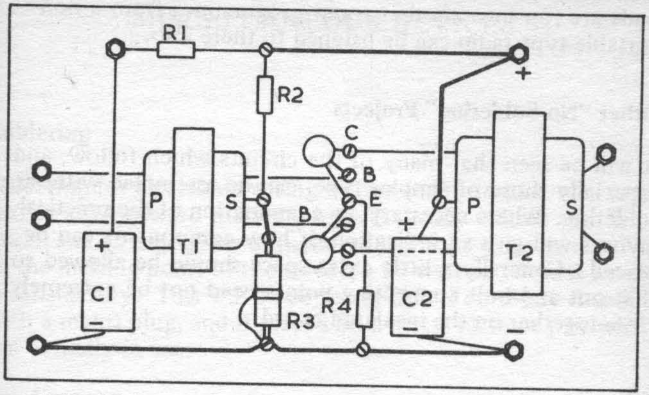

FIG. 12 B.

Boost Speaker

For best reproduction, a speaker needs to be enclosed in a cabinet or fitted to a baffle board. A cabinet can be made from wood, and can be about 6 x 10 inch and 4 inch deep, or as required for the speaker unit. Cut an aperture in the cabinet front to match the cone diameter, and glue thin fabric behind this opening. The speaker can be fixed with nuts and bolts, or short woodscrews driven in from behind.

Connect flexible leads, with separate plugs or a jack plug, to suit the equipment.

Small pocket receivers, which have a miniature speaker, can give much improved volume and reproduction when used with a larger speaker. If a headphone socket, operated from the receiver output stage, is present, it is only necessary to plug the speaker into this. A fairly large speaker, of sensitive type, will give best results. A unit 5 inch to 6 inch in diameter, or about 4 x 7 inch, will be convenient. The speaker has to be of approximately the correct impedance, which should be found in the receiver operating instructions. Where the speaker impedance is wrong, a matching transformer has to be employed between receiver and speaker.

Extension speakers are less frequently used today, but if two leads are run into another room programmes from a non-portable type radio can be listened to there also.

Other "No Soldering" Projects

It will be seen that many of the circuits which follow, and especially those of simpler type, can be assembled without soldering. Where necessary, an examination of the practical layouts will give an indication of how components can be placed. Generally a little extra space should be allowed so that nut and bolt connecting points need not be extremely close together on the insulated board.

SECTION 2
MISCELLANEOUS DEVICES

Soldering

Soldering for electronic wiring will be found to be a very easy and straightforward matter. An electrically heated iron is usual, and as large areas of metal do not have to be raised to the melting point of the solder, a 20 watt to 25 watt iron is satisfactory. This will be perfectly safe if correctly fitted with a mains plug, and if placed on a metal plate or tray when not actually in use.

It is best to use cored solder, as made for electronic wiring.

Leads, tags and other items to be joined should be clean and bright. They have to be raised in temperature to the melting point of the solder at the actual points where a joint is to be made, and the active core of the solder, or flux, must be present here when the joint is being made.

The iron should be tinned – that is, have melted solder on its tip – and have reached its correct temperature. Leads can be hooked or looped into position. The joint is then heated with the iron, and the solder is simultaneously applied to the joint, so that it melts and flows over the surfaces to be soldered. This usually takes place in a second or two. The solder and iron should be removed at once, as excessive heating may damage some components, and especially transistors or other semiconductors. The joint cools in a few seconds, and should be rigid, and solder having flown onto the items to be joined.

A probable cause of poor joints is insufficient heating of items to be soldered, or carrying solder to the joint on the iron, so that the activity of the flux is wasted.

Electronic items such as tags and the wire ends of components are generally designed for easy soldering. They will be copper, tinned copper, and similar metals and alloys which solder readily. For this reason, difficulties are unusual.

Some metals do not solder readily, or need special methods. Aluminium is one of these, and it is not used where popular cored solders intended for copper or general wiring are employed.

Make sure that heat from the iron does not damage furniture or other items, and keep the flexible cord away from the heated tip. Always switch the iron off at once when the work in progress is to be left.

One Transistor Audio Oscillator

This can be used instead of a bell or buzzer warning, or for code practice and similar purposes. It needs very few components, Fig. 13, and can be made to work satisfactorily with a wide range of transistors.

In Fig. 13, an NPN transistor is shown in the theoretical circuit, so battery polarity is as indicated, with negative to emitter. If a PNP transistor is fitted, battery polarity must be reversed.

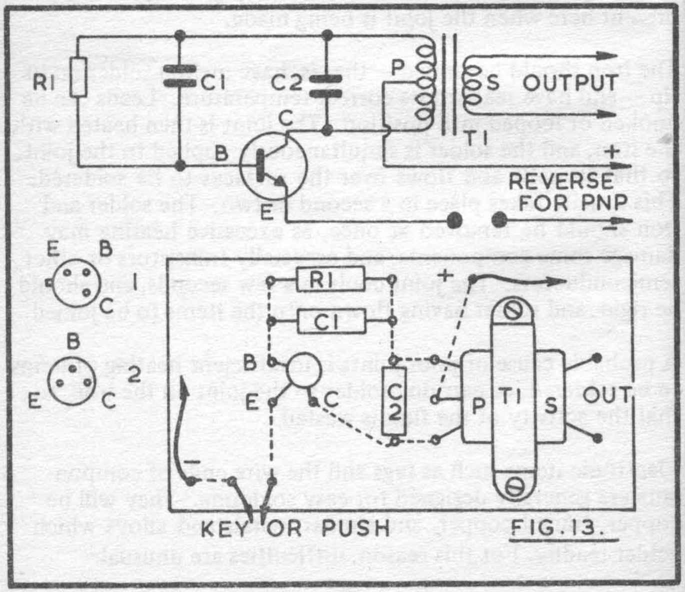

FIG.13.

(1) shows leads from a typical NPN transistor, such as the BC108, 2N706, and similar types. (2) is the base of the PNP AC128 transistor.

T1 can be a small speaker matching transformer, as used for push-pull stages. A centre tapped primary is necessary. The secondary S can be taken to a loudspeaker or headphones.

The values of R1, C1 and C2 can be varied considerably. C1 can be 0.1μF or so. R1 determines to some extent the power obtained, and can be about 22k, increased to 47k or 100k for lower power. The value of C2 has considerable influence on the note, or frequency of oscillation. Increasing the value of C2 lowers the pitch. C2 can be about 47nF to 4.7nF or thereabouts. A 9v battery can be used. Changes to the supply voltage will influence the note produced.

Fig. 13 also shows a layout for components, using either a fully perforated board, or a panel drilled to suit. Transformer T1 is fixed by screws. Small transformers of this type are usually wire ended, or have pins which pass down through holes in the board.

The wire ends of the resistors and capacitors are passed down through suitable holes, and bent over. The board is then turned so that the underside wiring can be carried out. In most places the wire ends of the components will be long enough to reach connecting points. Elsewhere, 24swg tinned copper wire, or similar connecting wire, will be used. Make the soldered joints as described, and snip off unwanted ends of wire. It is as well to leave the transistor to last, as this avoids heating it unnecessarily when soldering other items. Lengths of flex, anchored at holes in the board, are used for battery, key (or push) and speaker connections.

Multivibrator

Fig. 14 is the circuit of the multivibrator assembled in Fig. 2. Reference to Fig. 14B will show how this can also be prepared on a tag board.

The tag board has two rows of tags fitted to an insulated

board, and it is five tags long in Fig. 14B. Boards of this type can be obtained in many sizes. It is also easy to cut a section from a long tag board, to suit the needs of the item to be assembled. Such boards are very handy indeed for electronic assemblies, as the tags support resistors, capacitors and other items, and a neat layout is readily achieved.

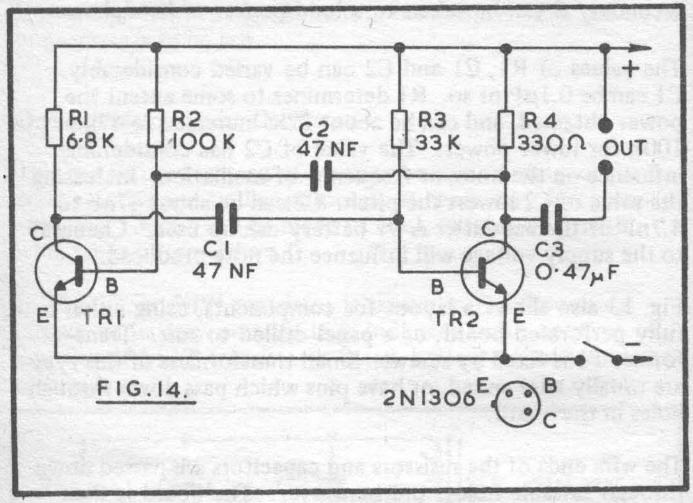

FIG. 14.

Tag strips are also useful for projects, and may have any number of tags required. Long tag strips can be cut down as necessary. Short tag strips have one tag for mounting the strip, longer strips having two tags formed as brackets for this purpose.

Place the components approximately as in Fig. 14B, and solder the wires to the tags. If any component values may be changed, do not twist the leads tightly round the tags before soldering, or removal will be difficult.

Wires which cross can be kept clear of each other, and then no insulated sleeving will be necessary. The transistors can be soldered on last, and can readily be changed without disturbing other parts.

The same layout can be used for the next project, though the component values are different.

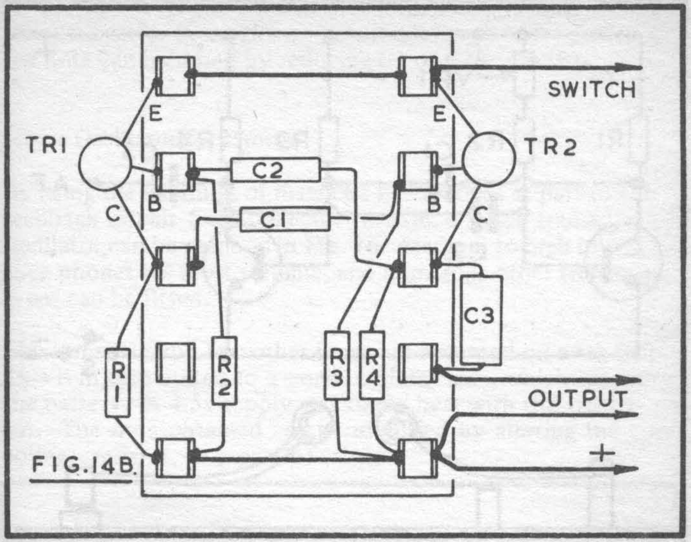

FIG. 14B.

Glissade

This device will produce swooping, rising and falling tones, for "effects" purposes, or can with additional skill be used to obtain musical notes through several octaves. The circuit is shown in Fig. 15, and reference can be made to Fig. 14B for a practical layout on a tag board.

In Fig. 15, suitable component values are:

R1	10k	C1	22nF
R2	12k	C2	22nF
R3	100k	C3	22nF
R4	10k	VR1	100k linear pot.

Transistors: 2 x BC108 or similar types.

The circuit is a wide frequency audio oscillator, in which the pitch can be controlled by VR1, and "on" switching is obtained by S. The oscillator is placed in a box, so that switch S and VR1 can be mounted on the top. The switch is a strip of springy metal, fitted with a spacer and bolt, so that it is just clear of a further bolt. A small downwards movement of the arm or knob completes the circuit.

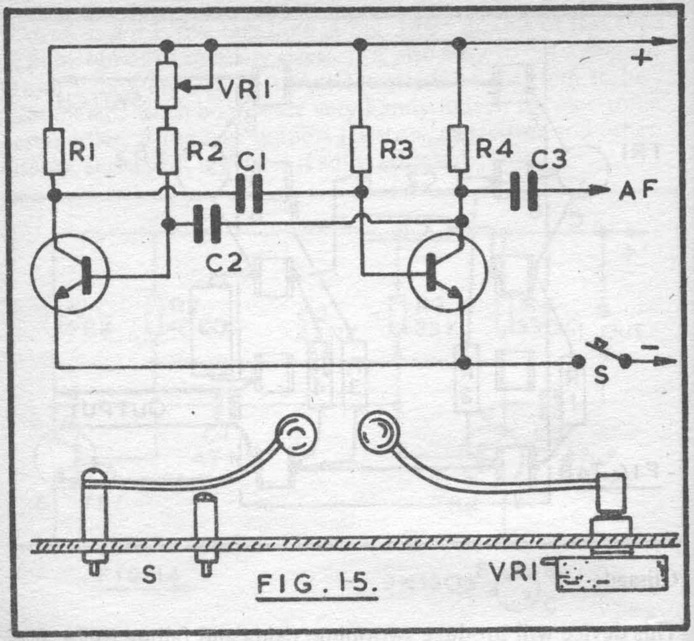

FIG. 15.

VR1 has a ¼ inch shaft coupler, into which is screwed a length of 4ba threaded rod. A knob is also fitted to the end of this rod. It can be swung backwards or forwards, to change the resistance of VR1. Connect VR1 so that movement of the knob towards the right raises the pitch.

Output from C3 can go to high resistance headphones, for amusement, but should be taken to an audio amplifier. Gain of the amplifier will have to be turned back somewhat, as the output from the glissade is at a quite high level.

Swooping tones are produced by holding down the switch S and swinging the control knob attached to VR1 from side to side. For notes which do not swoop to frequency, switch S must be released while VR1 is rotated. A slight quivering motion of the hand controlling VR1 will produce a vibrate effect. It is useful to have an ordinary switch in parallel with the knob control S, and to have the amplifier audio gain control near the left hand, so that fading effects can be produced as well.

About 4.5v to 9v can be used, this influencing volume. Should other transistor types give an unsuitable range of frequencies, the note can be raised by reducing C1 or C2 (or both).

Morse Oscillator for Phones

By using the windings of magnetic headphones as part of the feedback circuit from collector to base, a single transistor oscillator can be made as in Fig. 16. Medium to high impedance phones are most suitable, and numerous other transistor types can be fitted.

The transistor and few other items are mounted on a tag strip. This is in turn bolted to a small insulated box, which can carry the battery. A 4.5v supply was found best with this oscillator. The note obtained can be modified by altering the voltage, resistor, or capacitor values.

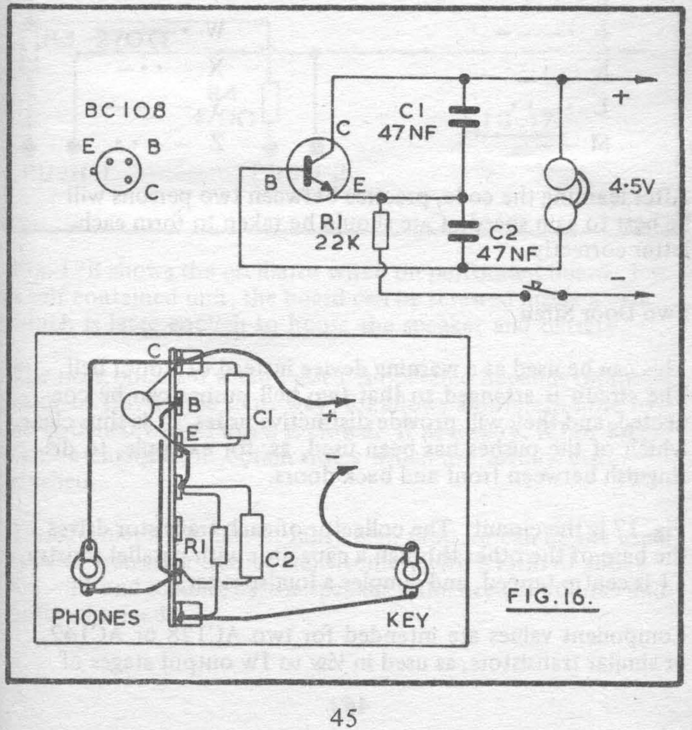

FIG. 16.

For serious code practice a good Morse key is required. It is adjusted so that only a small movement of the wrist is needed to operate it. Each letter consists of short and long sounds, shown as (•) and (–). The (–) has three times the length of the (•)

A •–
B –•••
C –•–•
D –••
E •
F ••–•
G ––•
H ••••
I ••
J •–––
K –•–
L •–••
M ––
N –•
O –––
P •––•
Q ––•–
R •–•
S •••
T –
U ••–
V •••–
W •––
X –••–
Y –•––
Z ––••

After learning the code, practice between two persons will be best to gain speed. Care should be taken to form each letter correctly.

Two Door Siren

This can be used as a warning device instead of a door bell. The circuit is arranged so that two bell pushes can be connected, and they will provide distinctive notes. It is thus clear which of the pushes has been used, as, for example, to distinguish between front and back doors.

Fig. 17 is the circuit. The collector of each transistor drives the base of the other through a capacitor with parallel resistor. T1 is centre tapped, and couples a loudspeaker.

Component values are intended for two AC128 or AC142, or similar transistors, as used in ½w to 1w output stages of receivers and small amplifiers. Other transistors will of course operate in this circuit. The pitch obtained can be modified by changing C3. The circuit shown is for PNP transistors, and battery polarity has to be reversed for NPN transistors.

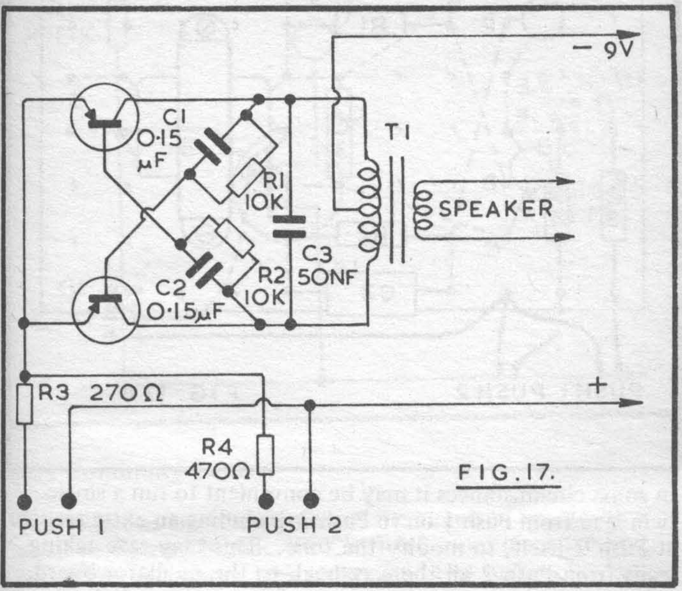

FIG. 17.

Fig. 17B shows the oscillator wired on perforated board. For a self contained unit, the board can be screwed inside a case which is large enough to house the speaker and battery.

The note obtained when Push 1 is operated depends on the value of R3. In the same way, R4 allows a different note to be obtained when Push 2 is closed. If necessary, R3 or R4 can be changed, to obtain notes which are readily distinguished.

Transformer T1 is best of fairly high ratio, or can be the usual type of transistor output transformer, with a ratio of about 8:1. T1 and loading by the speaker influence the volume and notes obtained.

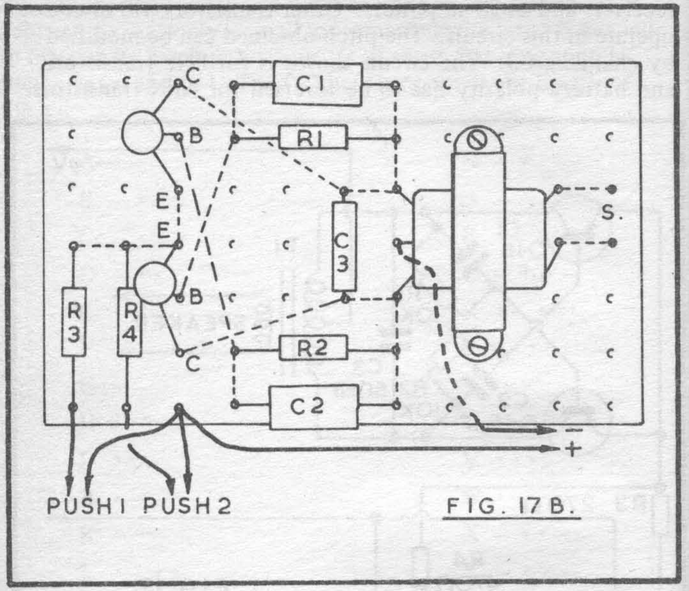

FIG. 17 B.

In some circumstances it may be convenient to run a single twin lead from Push 1 on to Push 2, including an extra resistor at Push 2 itself, to modify the tone. This may save taking leads from Push 2 all the way back to the oscillator board.

UJT Oscillator

The uni-junction transistor is a device having emitter, Base 1, and Base 2, Fig. 18. B2 receives positive bias, in this case through the headphones. When forward bias on the emitter E is reached the UJT conducts.

In Fig. 18, the emitter potential rises as C1 charges through the resistor R1. When the UJT conducts, C1 is discharged, and the process is repeated. This can take place over a very wide range of frequencies, depending on the values used for R1 and C1. Here, these are chosen to produce an audio tone.

A battery supply of 3v to 9v will be convenient. For use of the UJT oscillator during code practice, place the key in one battery lead.

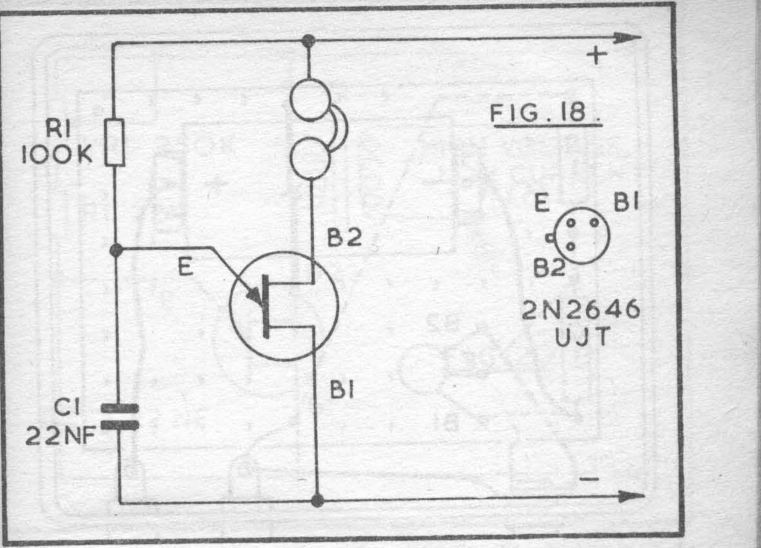

FIG. 18.

UJT Continuity Tester

Fig. 19 shows actual assembly of the circuit in Fig. 18 on a small perforated board. This is fitted in an insulated box, which carries a phone jack socket, and two sockets for Morse key or test leads. A 9v battery is used, secured with string or elastic.

For continuity tests, take prods or clips on the test leads to the suspected item. If the circuit is in order, the UJT oscillator will produce an audio tone. Oscillation will be obtained with a series resistance between the prods of up to 20k or more, so various components can also be tested, and some idea of the resistance in circuit can be gained from the sound obtained. The tester can also be used to check bulbs, lamps, fuses, extension and other leads, and similar items – these must of course be removed or disconnected from the equipment, before being tested.

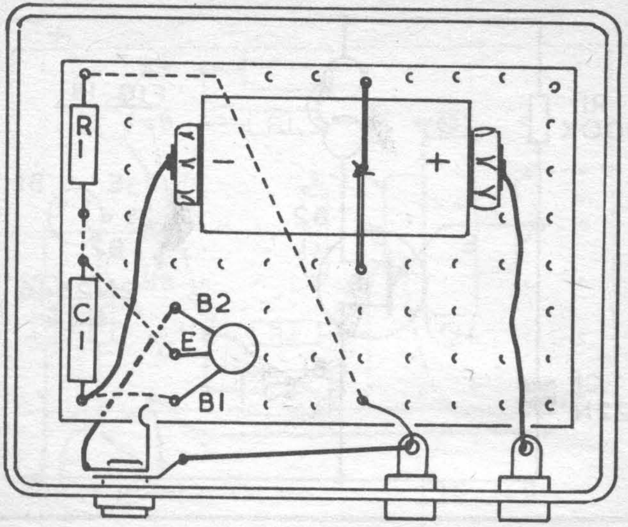

PHONES FIG. 19. KEY OR TEST LEADS

UJT High Voltage Producer

The UJT oscillator can be used to produce a high voltage, at very low current, by connecting a transformer as in Fig. 20. A loudspeaker transformer, as would be used in valve equipment, is suitable, and will have a ratio of about 45:1 or so. The usual secondary, or low-resistance winding, is connected to B2 and positive, so that a step-up of voltage is obtained. This should be found to light a small neon lamp.

VR1 is fitted to allow some adjustment, to find the most suitable frequency of oscillation. For maximum possible high voltage output, if required, operating voltage and current should be raised. No rating of the UJT must be exceeded. For the 2N2646, the supply should not exceed 30v, emitter current should not be over 50mA, and dissipation of power in the device should not be more than 300 milliwatts. (Dissipation can be found from Voltage x Current.)

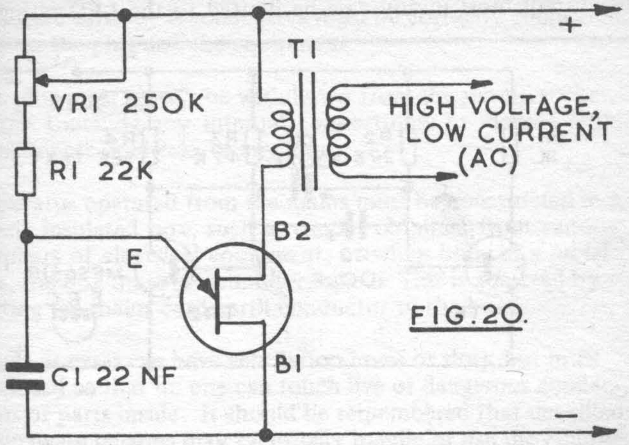

FIG.20.

A small mains transformer can be used for T1, with a low voltage winding, such as that for a 6.3v heater circuit, as primary, and the 200/250v winding as secondary. The high voltage output is alternating current.

Signal Blinker

Fig. 21 is a multivibrator in which the feedback capacitors C1 and C2 are of high value, so that conduction alternates between TR1 and TR2 at a low rate. A light emitting diode is placed in the collector circuit of TR2, and so flashes on and off.

This device can be used for various instruments and other equipment, and a flashing indicator draws attention to a greater extent than does one continuously illuminated. The LED can be located at any required point, with leads running to it.

Transistors other than the types shown can be utilised. R4 can usually be about 150 ohm for a 3v supply, 470 ohm to 1.2k for 6v, or 680 ohm to 1.5k for 9v. This depends to some

extent on the amount of light required when the indicator is operating. Lower resistor values at R4 will increase light, but current must in any case be limited to the LED rating.

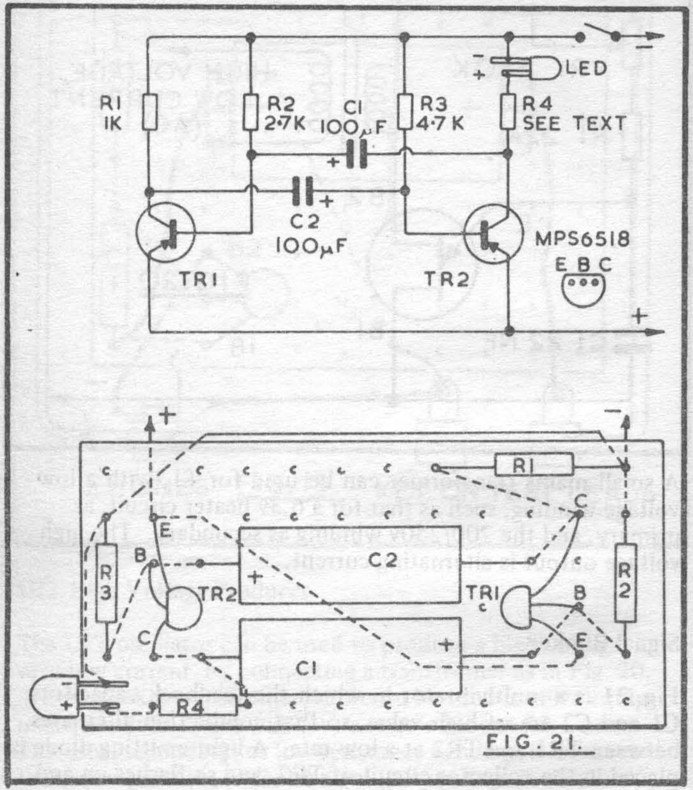

FIG. 21.

C1 and C2 need only be 6.4v working, and the device can be assembled on a very small perforated board. See p.61 for LED polarity.

Mains Wiring

All the devices described so far have operated from low voltage batteries, so than no shock hazard exists with them. However, it is necessary to employ mains power for some units, and proper care must then be taken to avoid any danger.

Power should be drawn from a 3-pin plug, and this should have a 3A or other low rating fuse installed. The live or L, neutral or N, and earth or E conductors must be correctly connected, both at the plug and the equipment.

The plug must always be withdrawn from the supply socket, before touching any internal connections, or making any adjustments or checks of any kind.

Apparatus operated from the mains must be constructed in a strong insulated box, such as can be obtained from various suppliers of electrical equipment; or when built in a metal box, the box must be soundly earthed. This is achieved by bolting the mains cord earth conductor to the box.

Boxes or cases can have ventilation holes or slots, but must be closed so that no one can touch live or dangerous connections or parts inside. It should be remembered that unskilled or unaware persons may eventually handle or use the equipment, so it must be arranged to present no unexpected hazards such as live terminals or parts, or openings for the fingers.

If these points are remembered, mains operated equipment can be safe to use and handle. No switch or other means of interrupting the circuit is put in the earth conductor, and any switch or equivalent device is placed in the live or L conductor. This leaves both earth and neutral permanently connected, for maximum safety.

Dimmer/Heat Control

It is often convenient to dim a filament lamp, or reduce the heat of a soldering iron to be left on for long periods, or the power of a low wattage heater. The silicon controlled rectifier circuit in Fig. 22 can be used for these purposes.

When the switch marked "Full" is closed, current for the load passes through this, so that full power is obtained.

When this switch is opened, current passes through the SCR. This device conducts only when the gate has been triggered, and the instantaneous point in the alternating current cycle

where this arises is determined by the setting of the potentiometer VR1. Triggering is advanced as the wiper of VR1 is moved towards R1. The average power supplied to the load can thus be adjusted between wide limits. Since the device is operating as a half wave controller, full power is not obtained with VR1 at maximum, so the switch is closed as described.

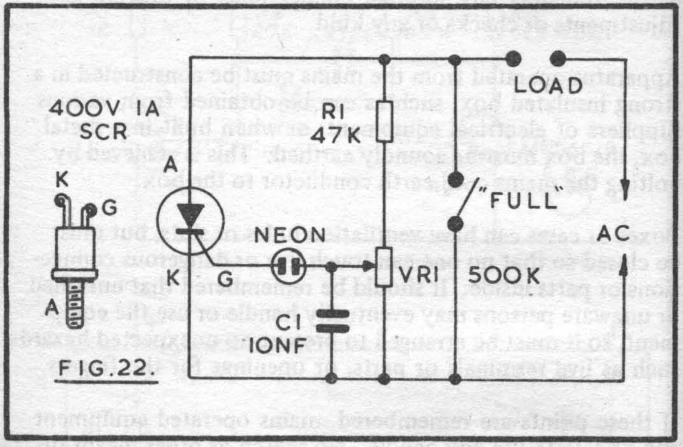

FIG.22.

Fig. 22B shows construction of the SCR controller in a metal box. The front of this carries a socket outlet, and the mains cord has a 3-pin plug, wired as explained earlier. It is thus easy to plug the controller into an outlet, and insert the plug of the lamp or other item to be used in the new outlet.

The mains cord is secured with a clamp, and passes through an insulated grommet. The tag strip supporting the small items is mounted by an unused tag. A strip of metal is cut to clamp round the body of the SCR, and is soldered to one tag, as shown. Run earth and neutral leads from the cord up through a hole, and connect them to the E and N terminals of the socket outlet. Also earth the case, at E. A well insulated wire runs from the L terminal of the outlet, to the "Full" switch. The outlet can be secured with bolts and nuts.

R1 should be 1 watt, and C1 is a 500v capacitor. The neon is a small type as used for indicators, etc.

After checking that the controller operates correctly, the back of the box should be screwed in place. A suitable size is about 4 x 5 x 2 inch deep.

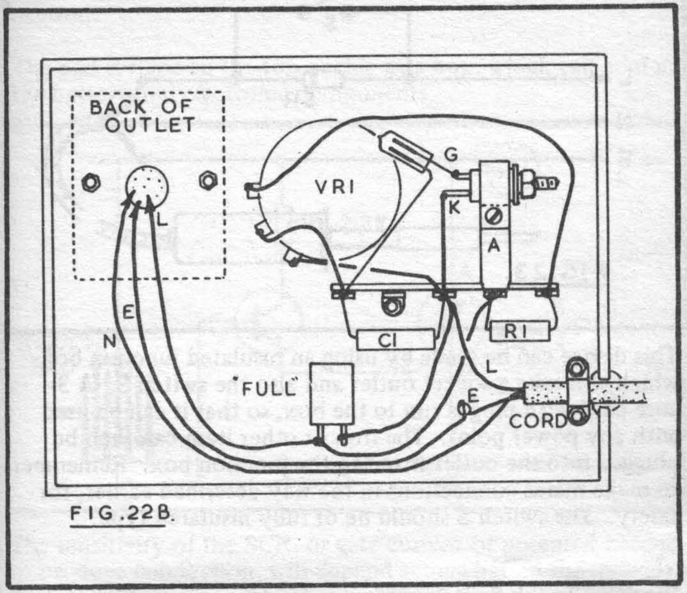

FIG.22B.

If a fully insulated box is used instead, then the bush of VR1 and of the switch should both be earthed by wires, except when these items are constructed wholly with insulated material. (This precaution is necessary to avoid any chance of metal parts outside the box becoming "live".)

Two Power

A circuit which will allow half or full power can be useful to reduce the heat of a soldering iron, or for dimming a filament lamp.

Fig. 23 is an easy circuit of this type. Full power is obtained when the switch S is closed. With S open, the silicon rectifier allows current to flow in one direction only, so that power is approximately halved. A 600v 1A or 1000v 1A rectifier is suitable.

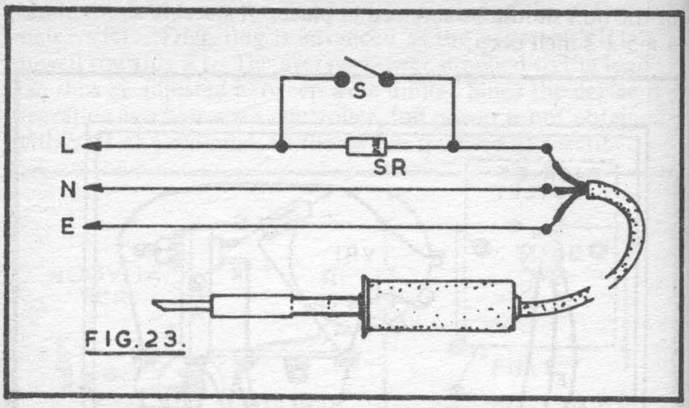

FIG.23.

This device can be made by using an insulated junction box which will take a socket outlet and also the switch S. A 3-core cord with plug is run to the box, so that it can be used with any power point. The iron or other item can then be plugged into the outlet fitted to the junction box. Remember to make mains connections in the way described earlier, for safety. The switch S should be of fully insulated type.

Sensitive Touch Switch

In Fig. 24, finger contact with the touch pad switches on a low voltage bulb. This circuit is suitable for a battery operated portable bedside light, and for some other devices.

Normally the silicon controlled rectifier is not conducting, so the lamp is extinguished. Finger contact with the pad provides positive base bias for the transistor so that it conducts. This moves the transistor emitter and SCR gate positive, triggering the SCR. Current then flows from cathode K to anode A, and through the bulb. The SCR remains in the conducting condition until the battery circuit is interrupted by opening the switch. A push switch, normally closed, is most suitable here. Pressing it then switches off the bulb.

This touch pad is easily made from a piece of circuit board with foil strips. It can be about 2 x 1 inch, or of any convenient size. Join together alternate foils – e.g., 1, 3, 5 and 7 – by soldering a connecting wire through the holes. This will provide one connection to the pad. The other foils – 2, 4 and 6 – are similarly joined together for the other pad connection.

The pad is fixed to the top or side of a box, which can contain the battery and electronic components.

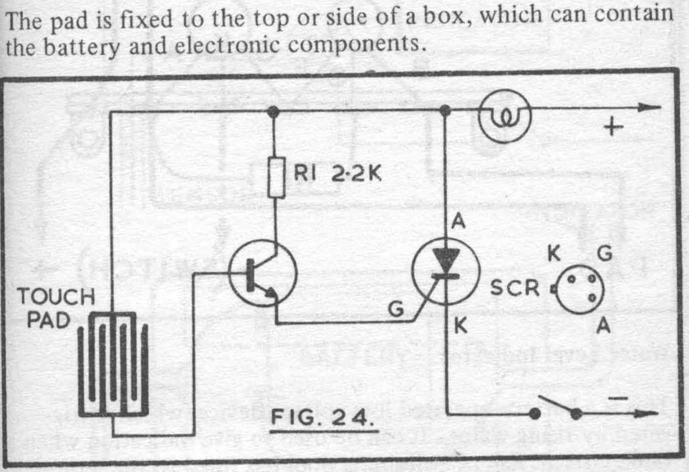

FIG. 24.

The sensitivity of the SCR, or gate current or potential needed to produce conduction, will depend somewhat on the individual device, while the gain of individual transistors will also be different. Numerous NPN types are suitable. If it is found that the switch is too sensitive, this can be corrected by placing a resistor from transistor base to SCR cathode or negative line. This resistor can be about 470k to 1 megohm or so. Excessive sensitivity will be apparent if the circuit is switched on by dust, damp air, or similar effects. The pad should be kept clean by wiping it with a dry cloth.

Fig. 24B shows assembly of the components on a tag strip. A small 50v 1A SCR is more than adequate. The bulb can be a screw in type, and can be 6.3v 0.15A, or 6.3v 0.3A. A 6v 3W (0.5A) bulb is also suitable, provided the battery is fairly large.

It is convenient to make a wooden box which will take the battery, and leave a little space for the tag strip assembly.

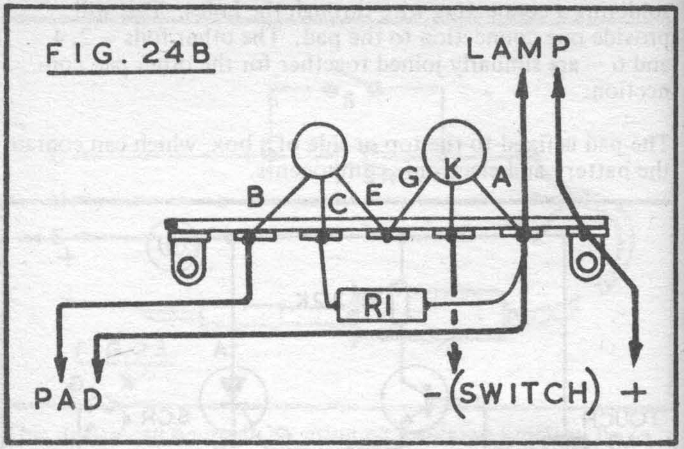

FIG. 24B

Water Level Indicator

This is a battery operated low voltage device, which is triggered by rising water. It can be used to give indication when a tank, cistern, fish or swimming pool has filled to the required level.

Fig. 25 shows the circuit, and also assembly on an insulated board. This also carries the battery, and it is fitted in a metal or plastic box.

In use, the base normally has no positive bias, as the sensor passes no current. When water reaches the sensor, it conducts, and the transistor passes collector current, to operate the indicator. R1 limits base current.

This circuit will operate satisfactorily with a wide range of PNP or NPN transistors, but battery polarity must be reversed for the PNP types.

The indicating or warning device can be one of the multivibrator or other audio oscillators shown earlier. These will give an audible warning, and will draw only a small current, so that a large transistor is not required.

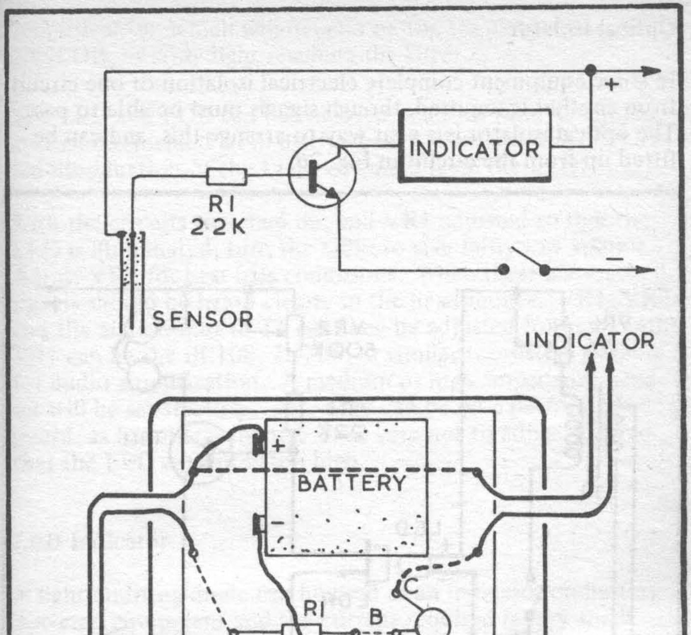

FIG. 25.

For indoor use or in a location protected from rain, the sensor can be two bare wires, about ¼ inch apart, bolted to a strip of insulated material. Where rain could operate the device, it needs to be covered. Alternatively, it can be made from a piece of metal tube, with a central wire, as in Fig. 25. The wire is bound with insulating tape for about half its length, to fit tightly in the tube. Flexible leads are soldered on, and the joints covered with tape. The sensor can be suspended inside the tank or pool, with its bottom at the required water level.

If the transistor is a high gain one, so that damp operates the indicator, sensitivity can be reduced by connecting a 100k resistor from base to emitter.

Optical Isolator

In some equipment complete electrical isolation of one circuit from another is required, though signals must be able to pass. The optical isolator is a neat way to arrange this, and can be fitted up from the circuit in Fig. 26.

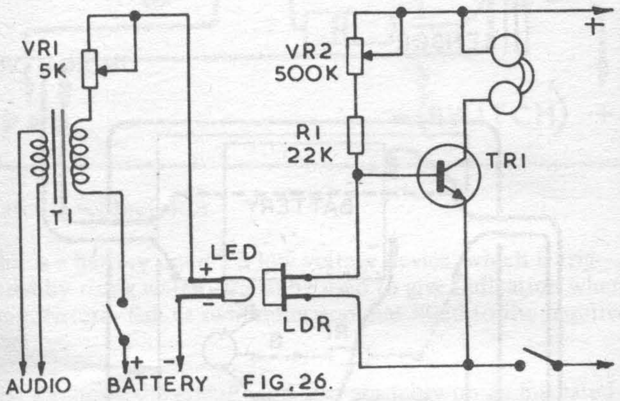

FIG. 26.

Electrically, the device is in two sections. The first has the light emitting diode LED, VR1, and transformer T1. When the battery is switched on, current flowing through the LED is adjusted by VR1, so that the LED glows dimly. A signal voltage in the secondary of T1 then modulates the LED light intensity.

To operate this circuit from a radio having a 2/3 ohm or similar speaker, a speaker matching transformer with a ratio of about 45:1 can be used. (The actual ratio is not very important.) Connect the usual transformer secondary to the radio speaker, so that T1 provides a step up effect.

The second part of the equipment has the light dependent resistor LDR, forming part of the network providing TR1 with base bias. VR2 allows adjustment of conditions for best amplification, which will depend on the transistor, and also the LDR, or stray light reaching the latter.

The LED and LDR can be enclosed in a card tube, or should at least be shielded from light, which will swamp the modulated illumination of the LDR, obtained from the LED.

With the circuits switched on, and VR1 adjusted so that the LED is illuminated, turn the radio to give fairly low volume. Adjust VR2 for best bias conditions. When these are reached, signals should be heard clearly in the headphones. VR1, VR2 and the audio input to T1 can then be adjusted for best results. TR1 can be the BC108, 2N706, or similar transistors suitable for audio amplification. A medium or high impedance headset will be satisfactory. Assembly can be on a perforated board, as for other circuits. Take care not to adjust VR1 so that the LED current is too high.

LED Indicator

A light emitting diode can be used as an indicator on battery powered equipment and the current required is very small. Fig. 27 is a LED indicator circuit. Here, the long wire of the LED is negative, but with some LEDs this is not so. A flat side may, or may not, be present. So polarity may have to be

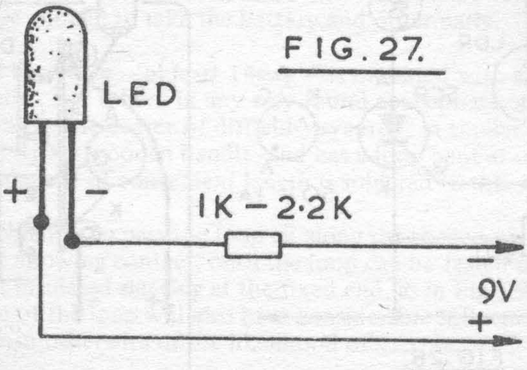

FIG. 27.

found by trial – the correct connections are those allowing the LED to light. The series resistor may be about 150 ohm to 1k or more, depending on the actual voltage, and brightness required. Current drawn varies with different LEDs, and with voltage and series resistance, but can be about 5mA or less, without reducing the brightness too far.

The LED indicator with resistor is connected to the circuit where voltage is present when the equipment is on. LEDs are available in red, yellow, green and other colours, with grommets or clips for panel mounting. Different circuits may be identified by a range of colours.

"Magic Candle"

The triggering of a silicon controlled rectifier can be used for slave photoflash and similar purposes. In Fig. 28 it is employed to switch on a light when a match or lighter is brought near. Sensitivity is best in somewhat subdued light, and it will then be found that striking a match at some distance can bring on the light.

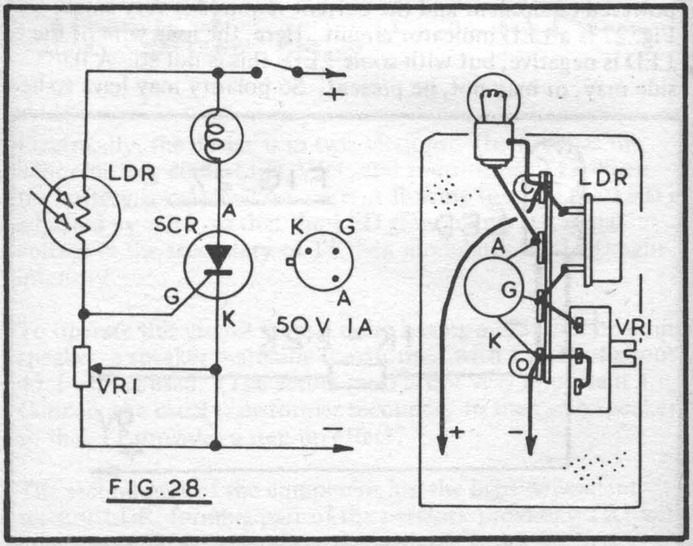

FIG. 28.

The LDR has a high resistance when in darkness or subdued light, so that the gate potential of the SCR is not sufficiently high to trigger this device. When light reaches the LDR, its resistance falls, and the SCR gate voltage rises. The SCR then moves into its conducting condition, so that current flows through the bulb. The sensitivity of the circuit, or level of illumination needed to trigger the SCR, is set by VR1.

Components can be assembled on a small tag strip. This, with a battery, can occupy a tubular case, with the lamp on top. An opening is necessary opposite the LDR, and a hole is also made so that VR1 can be adjusted with a screwdriver. Here, a miniature pre-set 50k or 100k linear potentiometer is most suitable. It is necessary to have a small push switch to interrupt the battery current, as the SCR remains in the conducting condition until the circuit is momentarily broken. A 6V battery and bulb can be used. SCRs may fail to operate at all if attempts are made to use them with extremely low voltages, such as a single 1.5v cell.

"Nerve Tester"

This device tests the nerve, concentration, and steadiness of hand, and always provides amusement.

Connections and construction will be seen from Fig. 29. A wooden box at least 10 inch or 12 inch long will be used, and large enough to take the battery and other parts.

A stout bare wire – at least 14swg – is mounted with a bolt and spacer, and is bent in any way found appropriate, and according to the degree of difficulty wanted. A similar wire is fitted with a wooden handle, and has a loop bent at one end. A flexible lead of convenient length is soldered to this wire.

The problem is to pass the loop all along the shaped wire, without allowing contact, until the loop can be rested on a piece of insulated sleeving at the fixed end, as in Fig. 29. The size of the loop will also have considerable influence on the ease or otherwise of the likelihood of success.

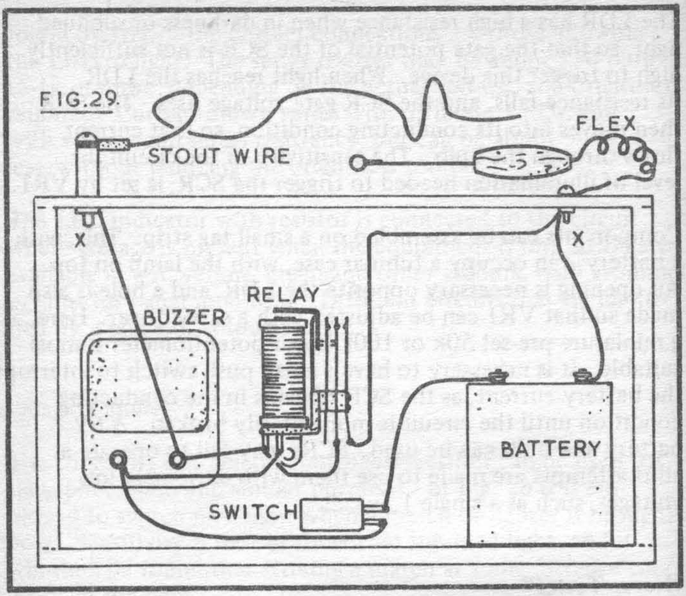
FIG.29.

If loop and wire touch, current flows from the battery to the buzzer, which sounds. Current also flows to the relay winding, so that the relay contacts close. Even a momentary contact between loop and wire will thus cause the buzzer to sound until the switch is used to break the circuit.

Any 4.5v, 6v or similar battery and buzzer will be satisfactory, with a relay which operates on a similar voltage. Connections are to the relay contacts which remain open until the relay is energised.

In use, the device can stand upright. A back is attached with small screws.

Burglar Alarm

The same circuit arrangement as in Fig. 29 will provide a burglar alarm operated by pressure mats. The latter are available in various sizes, and consist of a thin pad which can be placed under a carpet or ordinary mat.

Leads from the mat or mats will run to the terminals at X–X. The buzzer should be changed for a bell, which need not be fitted in the box, but can be situated where it is best heard. Momentary weight on any of the mats will then switch on the bell, which will be locked on by the relay as described, and will ring until the unit is switched off.

Transistor Tester

A very useful check of transistors can be made with the aid of a multi-range test meter, with a selection of base resistors to provide base current, as in Fig. 30. This method does not give an exact indication of transistor gain, but it does allow defective transistors to be eliminated, and provides a good idea of the performance to be obtained from surplus or other unknown transistors.

A 6-way switch allows the base to be open circuit, or to receive current through one of the resistors R1 to R5. These values are R1 330k, R2 110k, R3 33k, R4 11k and R5 3.3k. After allowing for the base-emitter potential, switch positions and corresponding base currents are near the following: 1, zero; 2, 10μA; 3, 30μA; 4, 100μA; 5, 0.3mA, and 6, 1mA.

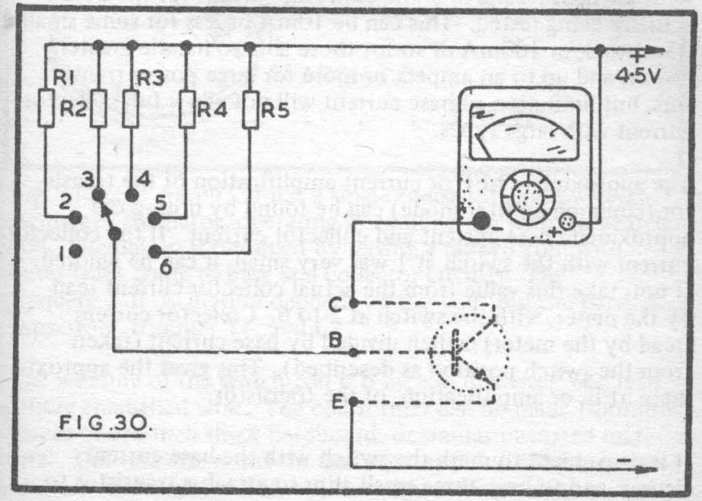
FIG.30.

Polarity is shown for NPN transistors. Reverse polarity of both meter and battery for PNP transistors.

For all tests, initially set the switch to 1, and the meter to its highest range (say 1 ampere). Afterwards, switch the meter to a lower range, after adjustment of the 6-way switch, as may prove to be necessary to obtain an easy reading.

With the switch at 1, a high reading on the meter shows that the transistor has a short circuit from emitter to collector, and is defective.

With germanium transistors, collector current (with the switch at 1) may be 0.1mA or so up to perhaps 0.5mA for small types, and up to some milliamperes for larger types. This small leakage current does not show that the transistor is defective.

In the case of silicon transistors, the leakage current (with switch at 1) will be zero or extremely small, if the device is in good condition.

The switch is then turned to one of the positions 2 to 6. If collector current does not rise, the transistor is defective. If current rises at each step, do not go beyond the switch setting which results in a safe collector current for the device actually being tested. This can be 10mA or less for some small transistors, or 100mA or so for those able to handle greater power, and up to an ampere or more for large power transistors, but limitation of base current will not allow full collector current with large types.

The approximate hFE or current amplification of the transistor (common emitter mode) can be found by noting the approximate base current and collector current. If the collector current with the switch at 1 was very small, it can be ignored. If not, take this value from the actual collector current read by the meter, with the switch at 2 to 6. Collector current (read by the meter) is then divided by base current (taken from the switch position as described). This gives the approximate hFE, or amplification, of the transistor.

It is convenient to mark the switch with the base current figures, and to have three small clips to attach a transistor to the circuit. Switch, clips, and battery can all be fitted to a small box. A development of the circuit would be to include a current meter, with switch and shunts, to replace the external general purpose test instrument.

Treasure Locator

This locator is among the simplest that can be constructed, as it uses an ordinary small portable radio as part of the detection circuitry. No modifications or connections to the radio are needed.

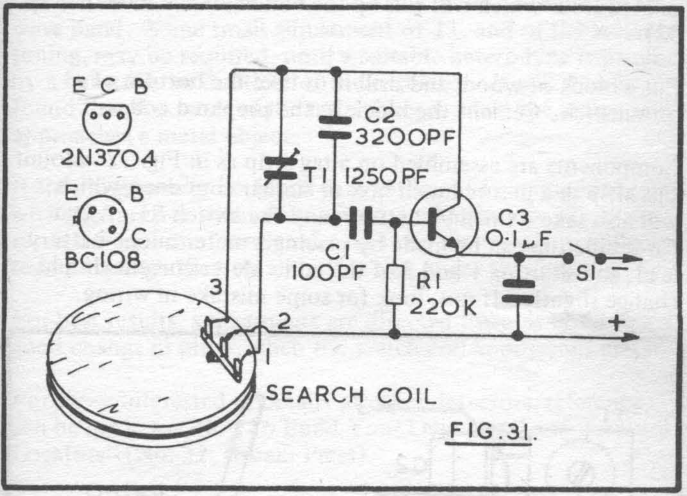

FIG.31.

Fig. 31 is the circuit. It has a single transistor as a low radio-frequency oscillator, operating in conjunction with the search coil. Oscillation is obtained by having a tapped coil, with outer leads going to base and collector circuits. The frequency of oscillation depends on the coil, and the parallel capacitors, trimmer T1, and C2.

The winding of the search coil is 6 inch in diameter, and is of 34swg enamelled wire. The coil former can be made from three pieces of 1/8 inch thick hardboard, or similar insulated material. Cut one disc 6 inch in diameter, and two others about 6½ inch in diameter. Smooth the edges with glasspaper, and

cement them securely together, with the small disc between the larger discs. This provides a former with a channel to take the winding. The adhesive should be allowed to dry with the discs under pressure, so that there is no gap into which turns can fall. The former should also be varnished, and allowed to dry, before winding.

The wire is passed up through a small hole, this being point 1. Wind on fifteen turns. Make a loop and pass this up through a hole, for point 2. Wind a further twenty-five turns in the same direction, finishing off at point 3. These leads are scraped and soldered to a small tag strip. This is an anchor point for insulated leads which will run up the handle of the locator (Fig. 32).

Cut a block of wood, and drill it to take the bottom of a broomstick. Cement the block to the prepared coil.

Components are assembled on a tag strip as in Fig. 32. Mount this strip in a plastic lunch box or similar container, which will also take a small 9v battery, and the switch S1. A test for oscillation can be made by placing a meter in one battery lead, and shorting 1 and 3 of the coil. Meter current should change slightly. If not, look for some mistake in wiring.

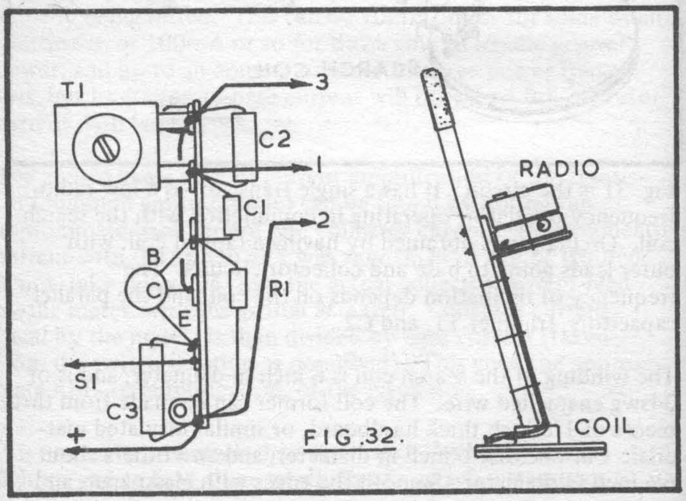

FIG.32.

A shelf is bolted to the handle, and the detector rests on this, with connections passing down to the search coil. A few loops of tape or string will secure these to the handle.

Operation in Great Britain can be on frequencies under 100kHz. Initially tune the radio receiver to 200kHz on long wave, and place it near the coil. Adjust T1 until a strong whistle is heard in the receiver speaker. This is the 2nd harmonic of the oscillator. T1 is then screwed down very slightly from this position.

The small portable radio is placed on top of the locator box, and can be held here with elastic. It is tuned to the medium wave band. Some small adjustment of T1, and of the receiver tuning, may be required, until a suitable heterodyne is produced by a harmonic or multiple of the oscillator. It will then be found that the pitch of the note changes when the search coil approaches a metal object.

It should generally be found that sufficient coupling is obtained with the receiver when they are situated as in Fig. 32. If not, run a few inches of insulated wire from the transistor collector tag up through the box and place this under the receiver.

For best results, adjustments are directed towards obtaining a good change in pitch, when the search coil approaches metal.

For those interested in details of other detectors, reference can be made to "How to Build Your Own Metal and Treasure Locators" (No. 32, Babani Press).

SECTION 3
RADIO AND AUDIO FREQUENCY

MW Booster

This unit can be employed in conjunction with any ordinary portable having a ferrite rod aerial, and will provide a considerable increase in range and sensitivity to weak signals. It has two functions – to allow the use of an external aerial and earth, and to amplify and couple signals to the portable receiver.

Tuning coverage is for medium waves. That is, about 200–550 metres, or 1500–550kHz. Many overseas and other stations can be picked up on this band, and distant reception can be particularly common on the higher frequencies, during the hours of darkness.

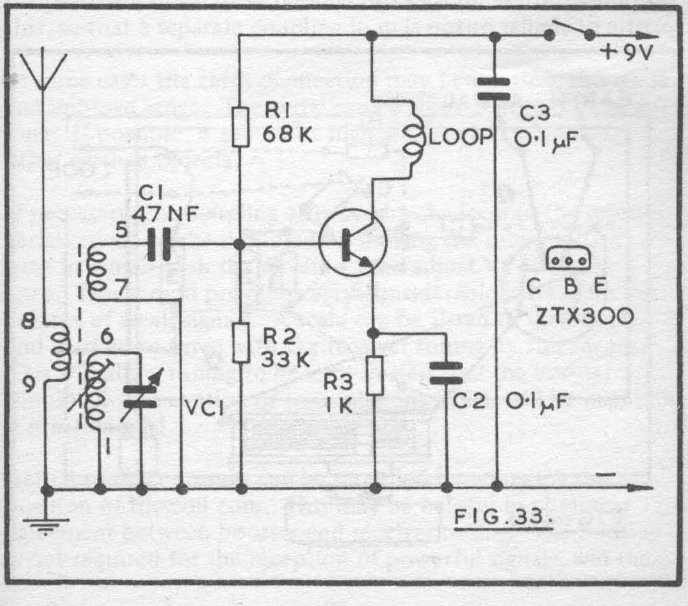

FIG.33.

Fig. 33 is the circuit, and it uses a cored tuning coil with aerial coupling winding 8–9, tuned winding 6–1, and base coupling coil 5–7. These numbers are for the pins of the Denco (Clacton) "Blue" Range 2 component. It will be apparent that other coils, or surplus or home-wound inductors, can be substituted.

For the coil mentioned, VC1 is 315pF. However, this value is not too critical, provided the capacitor fitted allows tuning the wanted frequencies. A solid dielectric capacitor can be used, though the air spaced type is likely to be a little more efficient.

Coupling to the base is by C1, and resistors R1 and R2 set the base bias conditions, with R3 for emitter bias. These values will suit numerous RF transistors.

Fig. 33B shows assembly of the components on an insulated board. The coil is mounted by its threaded bush. When the board has been wired, it can be mounted by means of bolts with extra nuts. The case is preferably metal, for screening purposes, but the booster can be used satisfactorily with a plastic or similar box.

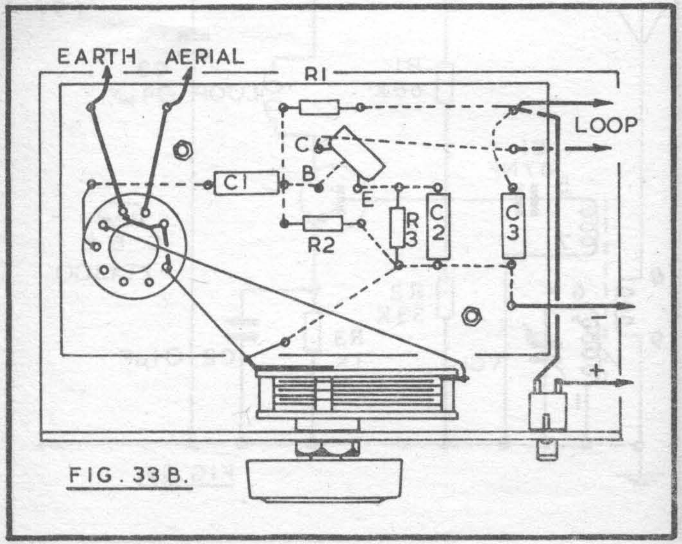

FIG. 33 B.

Fit a small slide switch to the box front, and aerial and earth sockets to the back, near the coil. A 9v battery is accommodated in the case.

Coupling to the transistor receiver is by means of a loop which is placed on the receiver ferrite rod, or near the rod. The loop can be about twenty turns of thin insulated wire, about ½ inch in diameter, and bound with cotton to keep the turns together The ends can be left long enough to reach the two connecting points in Fig. 33B.

If the case is metal, the booster and receiver can be near each other, so the loop leads will only need to be a few inches long. But when an insulated case is used, coupling from the receiver ferrite aerial or other circuits, to the booster coil, may cause instability. This will become apparent as whistles accompanying most signals. If this arises, the booster and receiver should be moved some inches apart. The aerial lead must not run near the loop connections.

Some receivers have an internal winding which is intended to couple an external aerial, and it may be possible to connect to this, so that a separate coupling loop is not required.

In some cases the earth connection may be omitted, though it can improve range. The aerial can be some 20ft to 40ft or so long. If possible, it should be high and clear of walls and other earthed objects.

If necessary, the coupling afforded by the loop to the receiver aerial circuit can be modified by moving the loop. Initially, tune in signals with the receiver, then adjust VC1 for resonance. This should provide a very considerable increase in the volume of weak signals. A scale can be fitted to the booster, and marked to agree with the receiver tuning by this means. This will allow tuning to be kept in step, and the booster should allow reception of transmissions which will be inaudible if it is removed.

Actual tuning coverage can be modified by adjusting the position of the coil core. This may be helpful in obtaining agreement between booster and receiver tuning. The booster is not required for the reception of powerful signals, and the

receiver automatic gain control circuit will tend to mask any increase in signal strength obtained with such transmissions.

Midget MW Diode Radio/Tuner

A small, compact tuner can be of utility in allowing radio reception through an amplifier, and a diode receiver can give excellent quality of reception. Such a unit can also be employed alone, with headphones, for personal listening.

Fig. 34 is a receiver/tuner which can be built in a very small plastic or other insulated box. VC1 is a compression type trimmer, fitted with a small knob for adjustment by hand. To do this, it is necessary to remove the short screw (generally 6ba) and replace this with a longer one, which projects to take the knob. The latter can be a terminal head, or the cap from an empty tube of adhesive, secured with lock nuts.

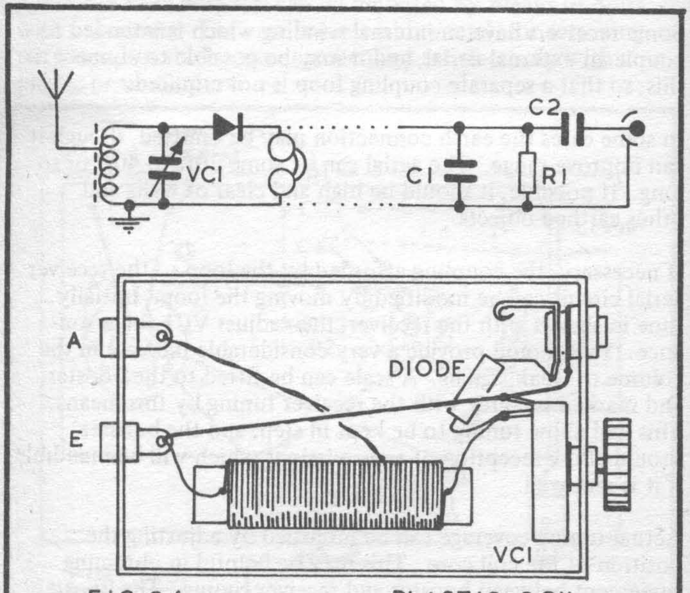

FIG.34.

For best efficiency and small size, the tuning coil is wound on a piece of ferrite rod. It consists of seventy turns of 24swg or similar enamelled wire, on a rod about 1½ inch long and 3/8 inch in diameter. Actual details will not be critical. A piece can be obtained from a longer rod by cutting a notch round the rod with a file and snapping at this point with the fingers. A few touches of adhesive will hold the ends of the winding secure. The loop for the aerial connection is about one-third the total number of turns from the diode end of the coil. (Moving this tapping towards the earthed end of the coil will increase selectivity or sharpness of tuning, but reduce volume.)

Aerial and earth sockets, and a 3.5mm jack outlet, complete the receiver. Headphones can be plugged directly in here.

The aerial and earth can be provided as explained for the diode radio in Fig. 7.

To use the radio as a tuner, C1, R1 and C2 may be added. Suitable values are C1 470pF, R1 220k, and C2 0.25μF. A screened lead with jack plug should be made up to connect the tuner to the amplifier input socket.

One Transistor VHF Receiver/Converter

The super-regenerative receiver is well known for its great sensitivity, which equals that of a multi-stage superhet. It has no alignment difficulties, and few components, so is an interesting project. Its main disadvantages are its background hiss, but this ceases when a signal is tuned in, and its lack of selectivity. The latter is however relatively unimportant on the VHF bands. The circuit here is particularly intended for 2m reception, but is easily modified. It is not intended for FM signals.

In Fig. 35 the detector is a VHF FET, and the field effect transistor will be found very satisfactory in circuits of this kind. T1 is the pre-set tuning capacitor, which allows adjustment over a fairly wide band of frequencies. Tuning is by VC1, which does not require a reduction drive.

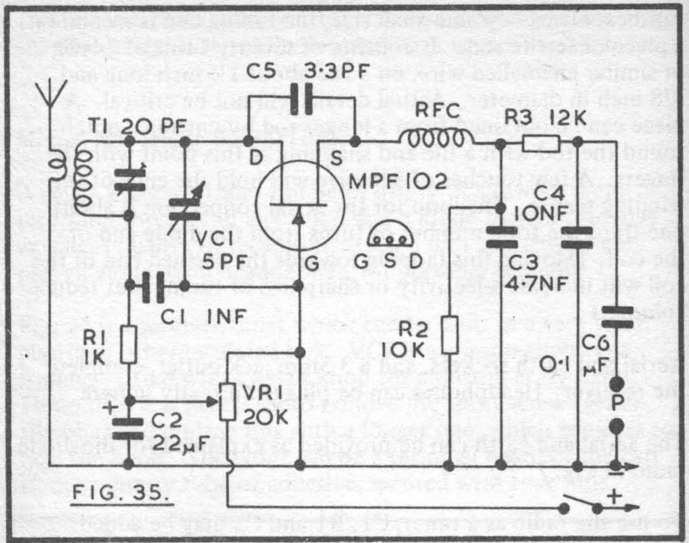

FIG. 35.

Control of super-regeneration is by means of the potentiometer VR1, which adjusts the FET drain potential. The value of VR1 is not very important, but a linear potentiometer is necessary.

C5 provides feedback from source to drain, for regeneration. R3 and the associated components form the quench network, with C6 feeding headphones, or an audio amplifier.

For best results, VR1 should enable the transistor to go smoothly into super-regeneration, which is shown by the hiss mentioned. Achieving this depends somewhat on the transistor, value of C5, adjustment of T1, and aerial loading. However, matters are not too critical.

Transistors unsuitable for VHF cannot be used. Individual FETs of the same type will also differ somewhat, so that in some cases it might be necessary to omit C5, or change the value. Alternatively, a small pre-set (not over 10pF) may be fitted. It will be found that a very small capacitance is required.

Another cause of failing to obtain super-regeneration is too tight aerial coupling, and this is easily corrected by moving the aerial coupling winding a little further from the larger winding.

The RF choke can be a VHF type, or can consist of seventy five turns of 36swg enamelled wire, side by side on a ¼ inch diameter insulated rod. Touches of adhesive may hold the end turns only.

Short leads are essential, and assembly on a tag board is shown in Fig. 35B. The variable capacitor is fitted first, and it can later attach the finished board to a panel, by means of an extra nut.

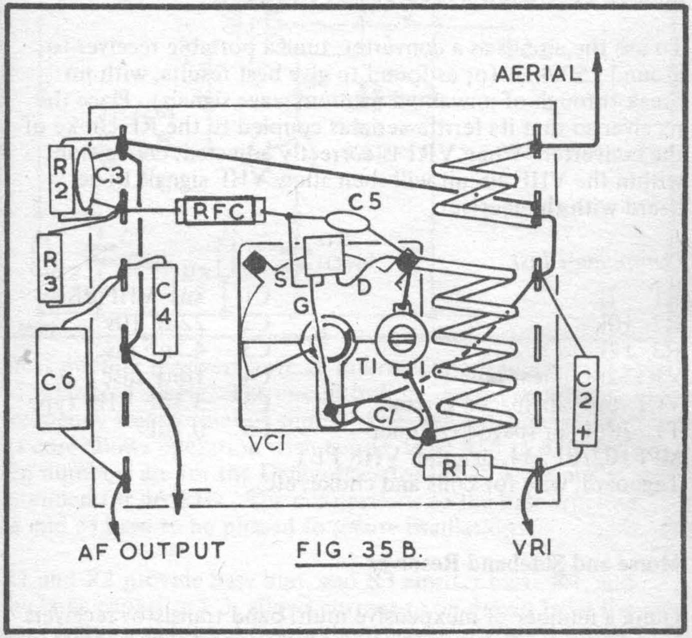

FIG. 35B.

The larger coil has five turns of 18swg wire, and is ½ inch in diameter and ¾ inch long. The smaller coil has two turns. Wiring to VC1, T1, C5, the coil and C1 must be as short as can be arranged. Connections elsewhere are reasonably direct. A ceramic disc capacitor may be used at C1. T1 should be a low loss trimmer.

The board can be fitted in a case which will carry VR1, and large enough for the battery and on-off switch. The case can be used to support a telescopic aerial, some 30 inch or so when extended.

Frequencies tuned by VC1 can be raised by unscrewing T1, or by stretching the larger coil to separate the turns more, or by using a coil with fewer turns, or of smaller diameter. The receiver is most suitable for use over the 70MHz to 200MHz range.

For the single transistor alone, a sensitive pair of medium or high resistance phones will be needed. Audio ouptut from C6 can be taken to an amplifier, for more powerful reception with phones, or to allow a loudspeaker to be used.

To use the circuit as a converter, tune a portable receiver to around 1500kHz (or as found to give best results, with no break through of unwanted medium wave signals). Place the receiver so that its ferrite aerial is coupled to the RF choke of the converter. When VR1 is correctly adjusted, conversion within the VHF circuit will then allow VHF signals to be heard with the receiver.

Component List

R1	1k	C1	1nF VHF disc
R2	10k	C2	22μF 10v
R3	12k	C3	4.7nF disc
VR1	20k linear pot.	C4	10nF disc
VC1	5pF air-spaced variable	C5	3.3pF VHF type
T1	20pF air spaced trimmer	C5	0.1μF

MPF102/BF244, or other VHF FET
Tagboard, wire for coils and choke, etc.

Morse and Sideband Resolver

Quite a number of inexpensive multi-band transistor receivers have frequency coverage which includes one or more Amateur bands. That is, 1.8, 3.5, 7, 14, 21, or 28MHz. On these bands transmissions are often in Morse code or CW, or single sideband. CW or SSB can be resolved by the specialised Amateur receivers and transceivers, but not with the general purpose home receiver. To receive CW or SSB with the latter type of radio, it is necessary to add an oscillator. This either generates a heterodyne or beat note for code reception, or replaces the carrier which was eliminated from the single sideband transmission.

The circuit in Fig. 36 may be used for these purposes. It generates a carrier of about the same frequency as that used in the intermediate frequency stages of the receiver, exact adjustment of frequency being by VC1,

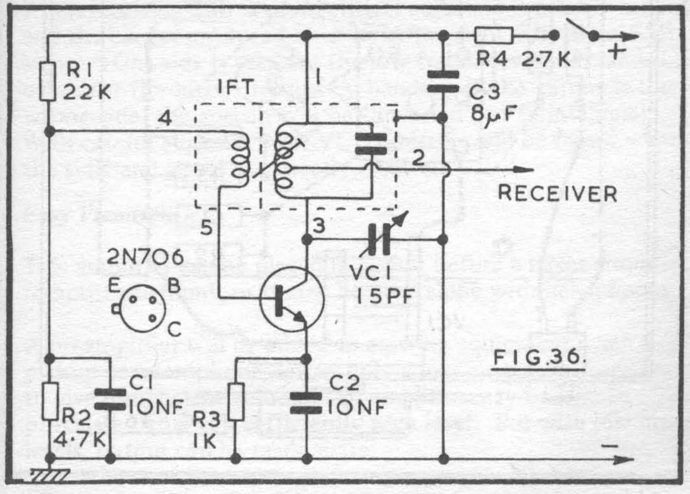

FIG.36.

Most popular receivers have an intermediate frequency of about 455kHz to 470kHz. The circuit in Fig. 36 uses an intermediate frequency transformer of similar type, and adjustment of its core allows operation with any IF from 455 to 470kHz. Pin numbers are for the Denco (Clacton) IFT13, normally intended for 465kHz. The connections to the base winding (4 and 5) have to be phased to secure oscillation.

R1 and R2 provide base bias, and R3 emitter bias. R4, and the large capacitor C3, allows current to be drawn from the same battery as is used to run the receiver, if wished. Alternatively, a separate 9v battery may be employed here.

Fig. 36B shows the practical layout of components, and wiring to them, using a perforated board. Connections and joints are kept close against the underside of the board. When wiring is finished, the board is mounted in a metal box, using ½ inch bolts with extra nuts for clearance.

VC1 can best have a dial with a central zero mark, this showing when the plates are half meshed.

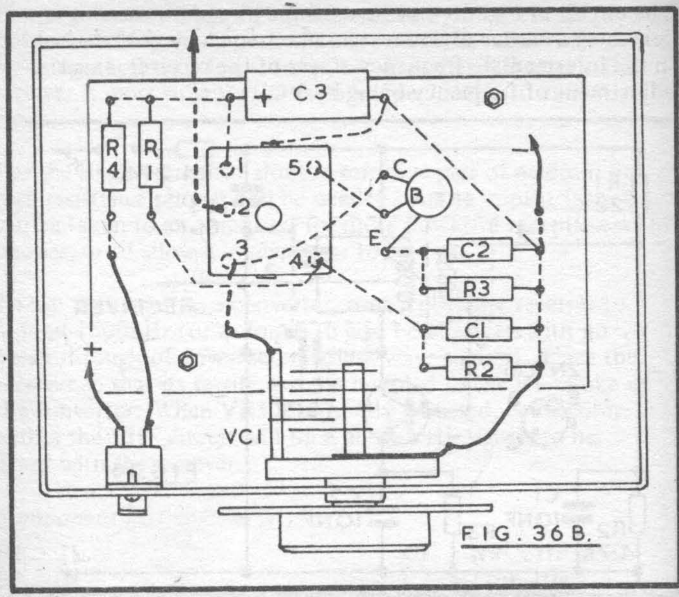

FIG. 36B.

An insulated lead is run from 2 on the IFT, and is placed near the receiver. This lead may be so positioned as to give best coupling into the IF stages of the receiver. If coupling is too weak, the signal from the CW/SSB Resolver will be inadequate. On the other hand, tight coupling will allow the carrier to swamp weak signals tuned in on the receiver. Fortunately the degree of coupling is not at all critical.

To set up the oscillator, first of all tune in any normal broadcast with the receiver. Set VC1 half closed and switch on. The core of the IFT should now be slowly adjusted until a loud heterodyne or beat note is heard accompanying the signal received. Set the core so that this note falls to about zero, or to a low pitch. It will then be possible to rotate VC1 either way from this central zero beat position, in each case producing a note which rises in frequency as VC1 is rotated.

To receive CW or Morse code, it is only necessary to tune in the code signal with the receiver, and then to adjust VC1 to obtain a note of the most suitable pitch. VC1 can be rotated either way from its central or zero beat position, to obtain best freedom from interference.

When receiving SSB, a more critical adjustment of VC1 is needed and the carrier produced must be at the right side of the SSB signal. (One side is used for the low frequency bands, and the other for the higher frequency bands.) If the carrier is the wrong side, the speech will be "inverted" or unintelligible. With careful adjustment of VC1, a setting will be found where the sideband signal is correctly resolved.

Easy Preamplifier

This amplifier can be placed in circuit before a larger amplifier, to boost the input, or it may be used alone with headphones.

A preamplifier will be added to existing equipment when a pickup or microphone having only a low output has been found to give insufficient volume. An amplifier may be able to provide a signal at a sufficiently high level. But with low input levels, output can be inadequate.

The stage in Fig. 37 can provide considerable gain. Audio input is to C1, and R1 provides base current for the transistor. Amplified signals are obtained across R2, and pass to the main amplifier (or phones) via the isolating capacitor C2.

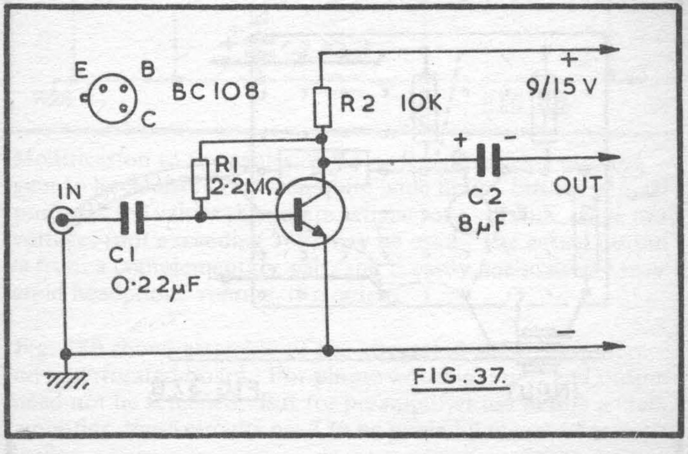

FIG. 37.

With a BC108 or similar transistor, the supply can be about 9v to 15v. Numerous NPN transistors, intended for preamplifier use and of low noise type, will perform well. Current may be from a separate battery. Or it may in many cases be drawn from the main amplifier. To avoid instability or similar troubles, the positive supply should be from a well decoupled and smoothed point in the main amplifier. If this is not available, connect a 220μF capacitor from positive to negative in Fig. 37, and draw the positive supply through a 1k resistor. Negative is returned to the negative line of the main amplifier.

Should the main amplifier have a positive earth line, then the transistor can be a PNP type, R1 and R2 being changed to suit.

Fig. 37B shows assembly on a perforated board. When the circuit provides input to a main amplifier, it should generally be in a metal case. Input is by a screened lead, and a screened lead is also fitted for the audio circuit to the main amplifier. The case should be earthed to the negative line. This can be by means of one of the bolts securing the board.

Screening to avoid hum or instability will generally prove to be unnecessary when using the stage to operate headphones. So two sockets, or a jack socket, will be more convenient for the phone circuit.

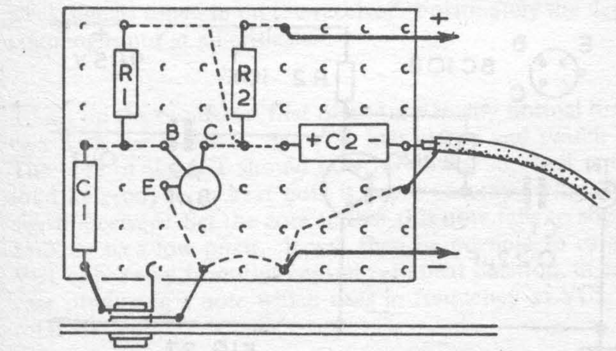

FIG.37B.

741 Preamplifier

Integrated circuits have several stages of amplification, formed as a single chip and enclosed in a small case. The 741 is an easily obtained audio IC which can be used in many circuits. The uA741CE or metal can 741 is convenient, as its leads may be spaced to suit any ready perforated board. The dual in line 741 is electrically similar.

Fig. 38 shows a preamplifier using the metal can 741. Input is via C1 to 3. As shown, 7 and 4 are positive and negative, while output is from 6. Negative feedback is applied to 2, via R4. C3 is an isolating capacitor.

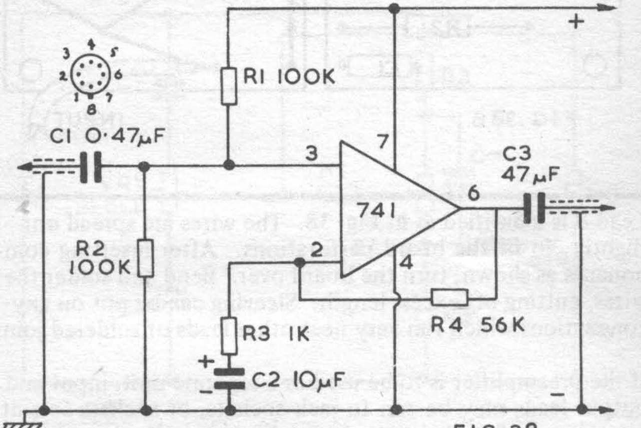

FIG.38.

Modification to the values of R3 or R4 allow feedback and gain to be adjusted between quite wide limits, but for general purposes the values shown are satisfactory. A wide range of voltages (not exceeding 36v) may be used. The actual output is from a complementary pair, and is easily adequate for very good headphone volume, if required.

Fig. 38B shows assembly of the integrated circuit amplifier on a perforated board. For phone working, input and output need not be screened. But for preamplifier use before a main amplifier, these circuits need to be made by means of screened leads.

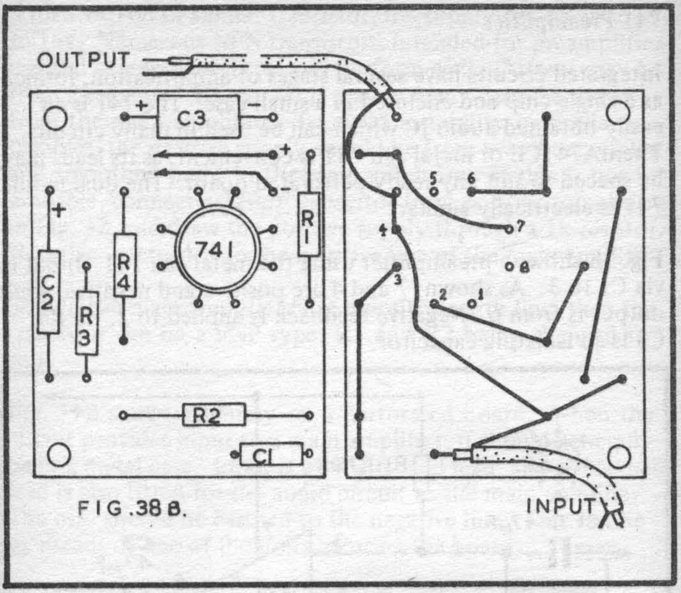

FIG.38B.

Lead 8 is identified as in Fig. 38. The wires are spread out slightly, to fit the board perforations. After inserting components as shown, turn the board over. Bend and solder the wires, cutting off excess length. Sleeving can be put on any connections which run very near other leads or soldered joints.

If the preamplifier is to be used as a separate unit, input and output leads may be run to jack sockets, or sockets to suit other items of equipment. A metal case is advised, as this will help to avoid any pick up of hum, or other troubles due to coupling to the board wiring. The board can be secured to the metal box or case with ½ inch 6ba bolts, and extra nuts to give a little clearance.

The amplifier may be run from its own battery; or from a supply used with a main amplifier. For the latter, it is wise to provide a decoupled point, in a similar manner to that described for the transistor preamplifier. This consists of a quite large capacitor (say 220μF to 470μF) from positive to negative across the preamplifier supply circuit, and a resistor of 1k to 2.2k or so from preamplifier positive to positive on the main amplifier.

Tone Control for Your Discs Etc.

The preamplifiers shown have a substantially flat response. That is, they do not purposely emphasize bass or treble, or have similar effects in modifying the audio frequencies present. In some circumstances it is necessary to boost treble or bass, or reduce the emphasis of these frequencies, in order to obtain the most satisfactory reproduction.

This may be achieved by making a fully adjustable tone control, and Fig. 39 is a suitable circuit. A volume control is also incorporated, as this has not been present in the preamplifiers.

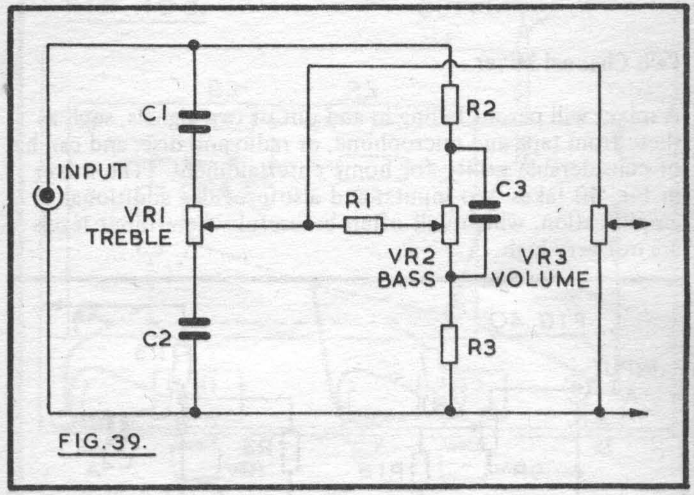

FIG.39.

The tone control is a passive circuit, needing no battery, and it can be placed between the source of audio signals (generally a microphone, pick-up, or radio tuner) and the amplifier or preamplifier. It may also be located between preamplifier and main amplifier, if wished.

Results are substantially flat with the treble control VR1 and base control VR2 in their central positions. Adjustment of VR1 and VR2 will provide separate boost or cut of treble, or boost or cut of bass frequencies. VR3 is the volume control.

The degree of control afforded by circuits of this type may be adjusted by alterations to the component values. In practice, it will be found that there is little point in having extreme levels of adjustment, which will never be required. The following values may be used: C1 and C2, 270pF; R1, 390k; R2 and R3, 100k; C3 680pF; VR1 and VR2, 2 megohm linear potentiometers; VR3, 2 megohm log potentiometer.

Components should be assembled in a metal case, for screening. The case is grounded to the earth line. Input and output sockets can be near the ends of the case, with VR1, VR2 and VR3 between. Components can be soldered directly to the potentiometer tags.

Two Channel Mixer

A mixer will permit fading in and out of two signals, such as these from tape and microphone, or radio and disc, and can be of considerable utility for home entertainment. The mixer in Fig. 40 takes two inputs, and also provides additional amplification, which will often be useful where input levels are not very high.

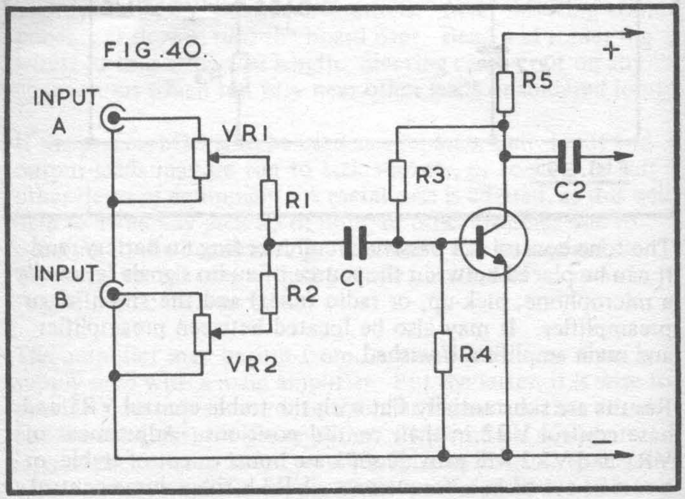
FIG.40.

Input A has its individual volume control VR1, and Input B is similarly provided with VR2. Resistors R1 and R2 prevent short circuiting one input when the other potentiometer is set at zero. Coupling to the transistor base is by C1. R3 and R4 set base bias conditions, with R5 as the collector load and audio output from C2.

Component values are not very critical, though R3, R4 and R5 need to be selected to suit the transistor fitted. The mixer is connected and used in a similar way to that explained for the preamplifiers.

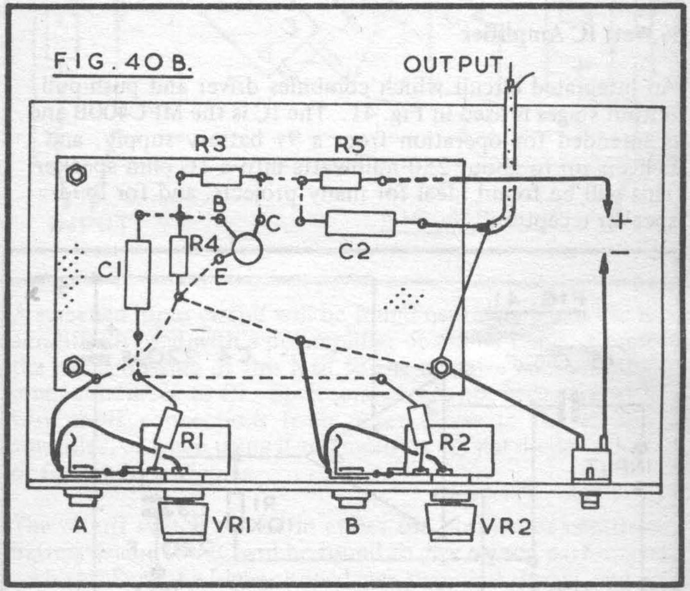
FIG.40B.

Fig. 40B is the layout, using a metal case or box. A screened output lead is provided. Input sockets A and B may be chosen to suit the type of plugs on existing equipment. The on-off switch is separate from the potentiometers, so that these can be left at pre-set levels if wished.

Component values may be chosen to suit low or high impedance inputs, if wished. Or the following may be used for general purposes:

Components List:

VR1, VR2, 1 megohm log potentiometers.
R1, R2, 200k
C1 0.5μF
R3 330k
R4 22k
R5 10k
C2 0.5μF
BC109

This is when using a 9v supply.

¼ Watt IC Amplifier

An integrated circuit which combines driver and push-pull output stages is used in Fig. 41. The IC is the MFC400B and is intended for operation from a 9v battery supply, and delivers up to about 250 milliwatts into a 16 ohm speaker. This will be found ideal for many projects, and for loudspeaker reception.

FIG. 41.

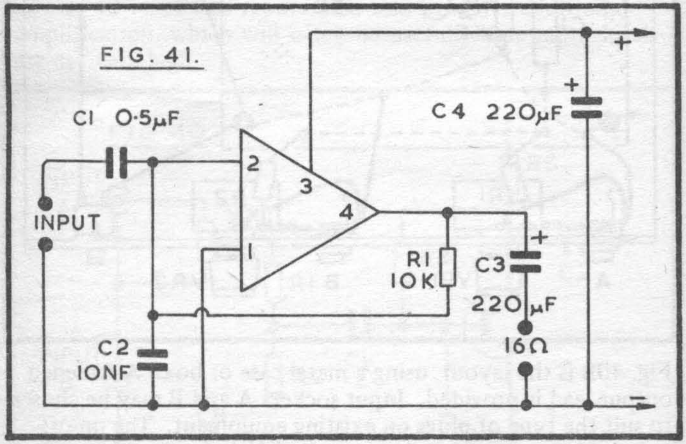

Audio input is to 2 of the IC, from the isolating capacitor C1. C2 is a by-pass capacitor to maintain stability. The IC has only four connections, 1 going to the negative line, and 3 to the positive supply. Audio output is from 4, and to the loudspeaker via C3. R1 provides negative feedback and bias to point 2.

The assembly of this amplifier on 0.15 inch matrix board is shown in Fig. 41B. Pin numbering is when looking at the top of the IC and the component side of the board.

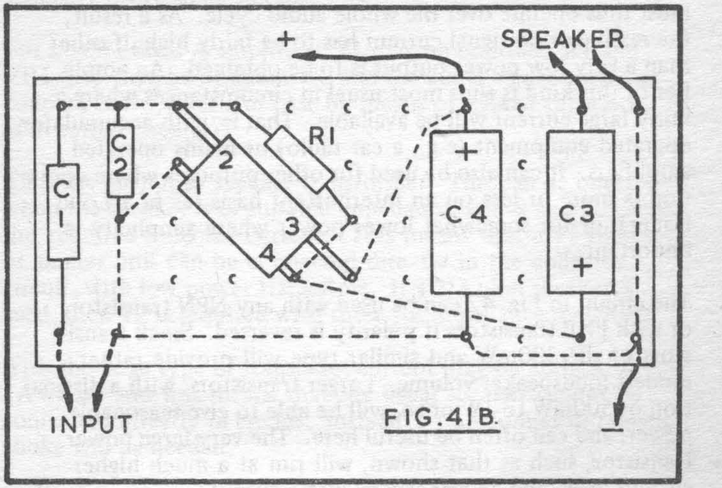

FIG. 41B.

A screened input circuit will be found necessary when the IC amplifier is used with a preamplifier or similar items. Connect the outer brading of this lead to the negative line, and the inner conductor to C1. Such screening is not necessary with very short connections from other stages to the IC amplifier, or when using it to amplify a crystal diode radio, or for similar purposes.

The on-off switch can be in either the positive or negative battery lead. The IC will be found to give a good performance with speakers of a higher impedance than that shown, and up to 75 ohm may be used (with some small loss in output). Units of lower impedance than 16 ohms should not be fitted.

The small size of this IC and the few extra components needed make is very suitable for miniature apparatus, probably with a 2½ inch speaker. It will also operate a large speaker well (see "Boost Speaker") and at a volume level adequate for many purposes.

Class A Amplifier

A Class A amplifier uses a single transistor, Fig. 42, which must thus operate over the whole audio cycle. As a result, the resting or no-signal current has to be fairly high, if other than a very low power output is to be obtained. An amplifier of this kind is thus most usual in circumstances where a fairly large current will be available. That is, with accumulator operated equipment (e.g., a car radio) or mains operated amplifiers. It can also be used for other purposes where operation is more or less on an intermittent basis (as in a Loud Hailer), or for somewhat lower power where simplicity is important.

The circuit in Fig. 42 can be used with any NPN transistor, or with PNP transistors if polarity is reversed. Small transistors of the 300mW and similar type will provide rather modest loudspeaker volume. Larger transistors, with a dissipation of 500mW to 1W or so, will be able to give reasonable power, and can often be useful here. The very large power transistor, such as that shown, will run at a much higher current level, and deliver several watts output.

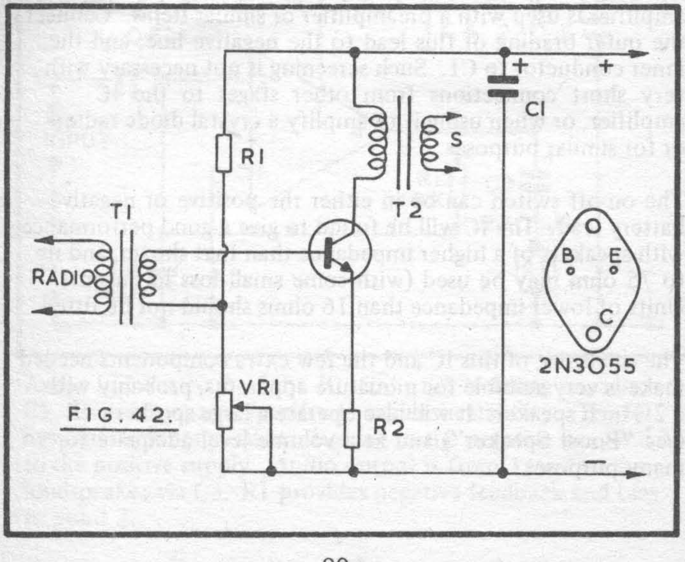

FIG. 42.

T1 is the coupling or driver transformer, which may receive audio signals from a radio receiver, smaller amplifier, or other equipment. A step down ratio of about 5:1 to 10:1 will generally prove to be suitable.

For smaller Class A amplifiers, T1 may be omitted. R1 and VR1 may then be of about 1k, and the audio signal can be coupled to the transistor base by means of a fixed capacitor, as in earlier circuits.

T2 is the output or speaker transformer. This can be omitted when the speaker is of suitable impedance, and is able to carry the required collector current. This means that a 75 ohm or similar unit can be connected directly in the collector circuit, with low power transistors. If a 2/3 ohm speaker is used, then T2 is required with these for impedance matching.

With the large type of transistor, current may be 500mA to 1A or so, and this means that the usual speaker cannot be connected directly in circuit. Instead, a transformer or tapped choke will be needed.

The circuit is often operated from 12v (a vehicle supply) but can be used with 9v, or more than 12v.

R1 and VR1 set bias conditions, and R2 limits peak collector current. For the 2N3055, R2 may be 1 ohm. R1 can be 100 ohm, and VR1 25 ohm. T2 can be 2:1 for a 3.5 ohm speaker. C1 is 3200μF. Begin with VR1 at minimum value, slowly increasing this until reproduction is satisfactory. (This will set collector current at about ½ ampere.)

With other transistors, adjustment can be made in the same way, by slowly increasing the base bias until best results are secured. Raising the value of VR1 in Fig. 42, or reducing the value of R1, will increase collector current, and thus raise the power which can be handled. However, current is limited by the power handling capacity of the transistor, so a meter should be put in one battery lead, and a check made that the power dissipation of the transistor is not exceeded. With small transistors, current can be limited to 50mA or so.

Fig. 42B shows assembly of the circuit in Fig. 42. The power

transistor is mounted on an aluminium plate, which acts as a heat sink to carry away heat from the transistor itself. Note that this plate is common to the collector. Base and Emitter pins pass through clearance holes. Smooth off any burr round the fixing and clearance holes, so that the transistor can rest flat on the metal. Where it is awkward to have the plate or case common to the collector, the usual insulation set can be fitted. This consists of bushes for the fixing bolts, and a thin mica washer to go under the transistor, so that electrical isolation is obtained.

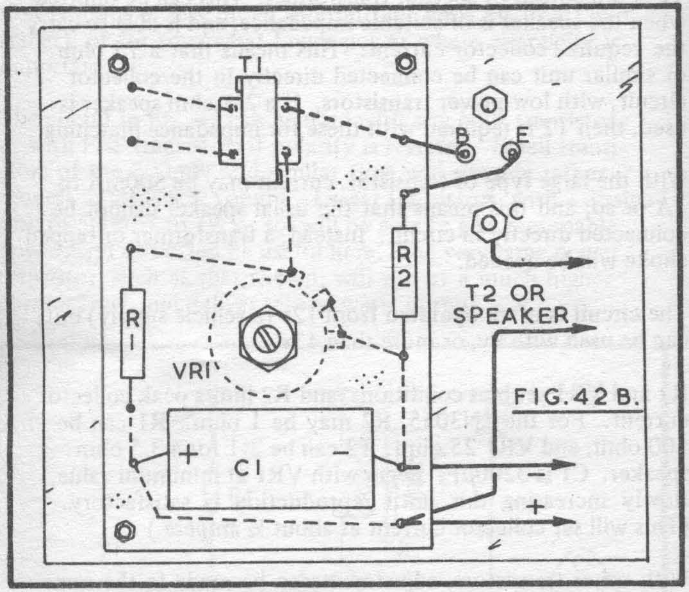

FIG. 42 B.

The insulated board carrying the other components is mounted by long bolts with extra nuts. A large Class A stage transistor of this type will grow very warm to the touch, if run at anything like maximum power rating. This does not indicate any defect.

Two Stage Amplifier

The circuit of a 2-stage amplifier is shown in Fig. 43, and this can be used for loudspeaker reproduction. Input to the driver stage is by the isolating capacitor C1. Operating conditions in this stage depend largely on the values of R2, R3, and the emitter bias resistor R5.

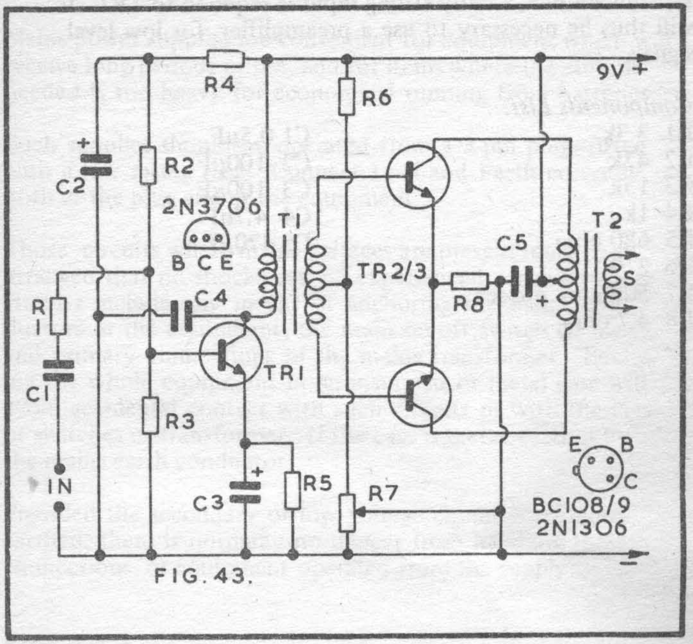

FIG. 43.

T1 is the driver transformer, with centre tapped secondary, so that TR2/3 may be operated in push-pull (see Fig. 12). T2 couples output to a 2/3 ohm speaker.

With the values shown, a BC109 will give very good results in the TR1 position, with 2 x BC108s for the output stage. R7 is set so that battery current is around 10mA, with no signal present. Current rises from 20mA to 60mA or so, according to volume, when the amplifier is driven.

A pair of 2N1306 transistors may be used for the output stage. Other transistors are also possible here, as operating conditions can be adjusted by R7.

The 2N3706 will operate in the TR1 position, though with

lower gain than the BC109. Numerous other small audio transistors can also be used here.

For full volume, a fairly strong input is required to TR1. It will thus be necessary to use a preamplifier, for low level signals.

Components List:

R1 3.3k
R2 47k
R3 15k
R4 1k
R5 680 ohm
R6 2.7k
R7 500 ohm pre-set
R8 4.7 ohm

C1 0.5μF
C2 100μF
C3 100μF
C4 4.7nF
C5 470μF

SECTION 4
POWER SUPPLIES

Mains power supplies are convenient for equipment which will receive long periods of use, and for items where the current needed is too heavy for economical running from batteries.

Such supplies should be operated from a 3-pin plug, fitted with a low rating fuse. Connect L, N and Earth correctly, both at the plug and at the equipment.

Those circuits where mains voltages are present must be so arranged that no shocks can be experienced. Mains voltage circuits include any means of anchoring the mains conductors in the equipment, the main on-off switch (if used), and primary connections of the mains transformer. Enclosing the whole equipment in an insulated or metal case will avoid accidental contact with such circuits or with the tags of switches or transformer. If the case is metal, earth it by the mains earth conductor.

Provided the secondary or low voltage circuit is correctly earthed, there is normally no danger from handling these connections, or equipment operated from the supply.

Easy PSU

A useful power supply suitable for many purposes can use the circuit in Fig. 44. Mains L (live) and N (neutral) conductors go to the primary P of the transforer. A switch is placed in the L circuit.

The secondary S can delivery any wanted low voltage, and also give isolation from the mains.

Almost all equipment will need direct current. The four separate rectifiers, or an equivalent single device, produce pulsating DC. The reservoir capacitor C1 smoothes this. Resistor R1, with the second capacitor C2, gives additional smoothing. Output from the supply is then substantially smooth DC, as would be obtained from a battery.

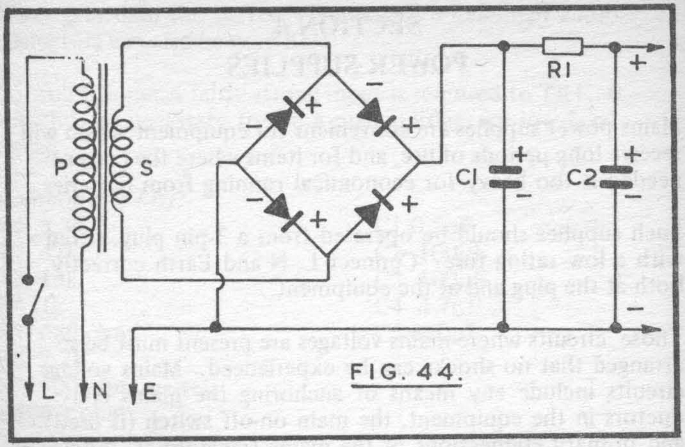

FIG. 44.

Sometimes R1 and C2 are omitted. Should hum accompany reproduction from an amplifier, this shows that smoothing is probably inadequate, and if so the cure is to add R1 and C2.

Any wanted voltage can be obtained by using a transformer with an appropriate secondary, with rectifiers and capacitors to suit.

C1 may charge up to approximately 1.4 times the RMS or usual voltage rating of the secondary S. Thus, if S were a 6.3v winding, about 8.8v would be expected across C1. But should R1 be present, a voltage drop will arise in it, the figure depending on the resistance value of R1, and current flowing.

For these reasons, the actual output may not be exactly the figure expected. This will often be unimportant, and such power supplies are largely used for all sorts of equipment.

For apparatus generally operated from a 9v or similar battery, a bell transformer with 3v, 5v and 8v output tappings is suitable. If R1 is 22 ohm, this will give around 9v at 100mA, with the 8v tap in use.

It is convenient to use 50v or 100v rectifiers, and the rating can be 1 ampere, with a 1 ampere transformer secondary. Four individual wire ended rectifiers can be mounted on a tag strip, and connected as in Fig. 44. A combined rectifier will have AC, positive, and negative tags (see Fig. 45) and can be fitted instead.

The capacitors need to be of large value, and have a voltage rating higher than the actual voltage which will be present. The 12v capacitors could be fitted in the circuit described. C1 can be 2500µF or larger, and C2 can be of similar value.

Easy Regulated PSU

With the circuit in Fig. 44, the exact voltage will depend on the current drawn, rising when current is low. To avoid this, some means of regulation may be employed.

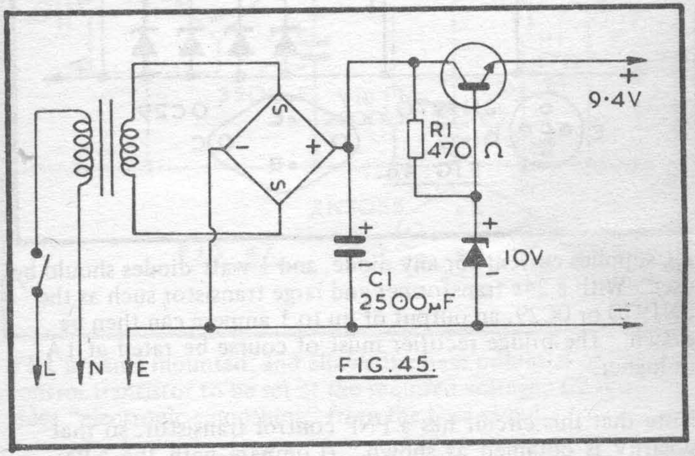

FIG.45.

Fig. 45 shows a typical regulation circuit, giving an output of approximately 9.4v. The output from the rectifier needs to be of somewhat higher voltage than required at the output, and a transformer with a 10v secondary is suitable. Current through R1 supplies the Zener diode, which holds the base of the transistor at virtually the same voltage for all current loads. Some voltage is lost in the transistor, so the output is nearly the same, for any current load. For up to about 200mA, a BFY51 or equivalent transistor is adequate, so this will do well for many purposes.

Multi-Voltage Regulated PSU

Fig. 46 is a somewhat similar circuit, but it has a switch, to select any one of the four Zener diodes. These may be chosen so that 6v, 9v, 12v and 15v will be available.

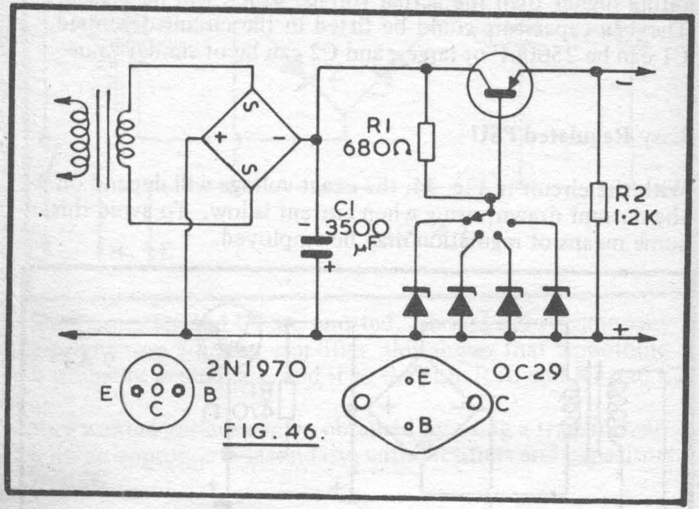

FIG. 46.

R1 supplies current for any diode, and 1 watt diodes should be used. With a 24v transformer and large transistor such as the 2N1970 or 0C29, an output of up to 1 ampere can then be drawn. The bridge rectifier must of course be rated at 1A or higher.

Note that this circuit has a PNP control transistor, so that polarity is obtained as shown. (Compare with the NPN transistor in the previous circuit.) However, either positive or negative can be grounded, and a separate earth terminal can be provided to allow this. A flying lead is placed between earth terminal and positive, or earth and negative, as required. The permanent negative earth connection shown in Fig. 45 must of course then be omitted.

More details of actual construction will be found later.

One Transistor Adjustable PSU

Fig. 47 is the circuit of a simple power supply unit which can be set to provide any output from zero to 12 volts. The transformer has a 17v secondary, rated at 1 ampere or higher. After rectification, DC is obtained from the reservoir capacitor C1.

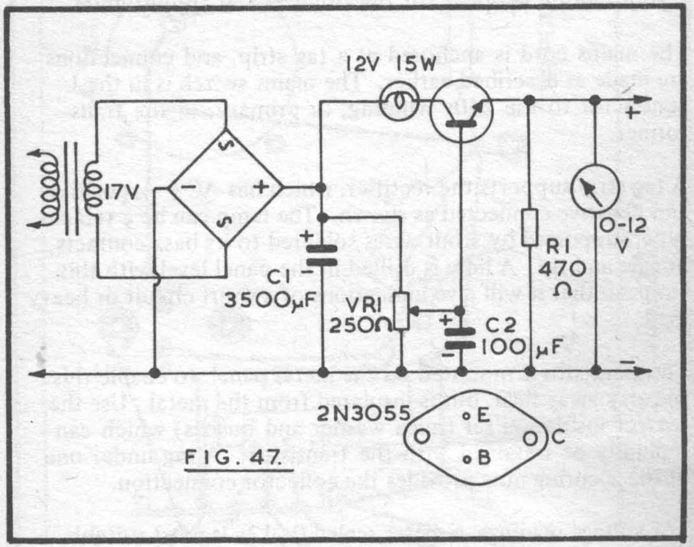

FIG. 47.

VR1 is panel mounted, and allows the base potential of the control transistor to be set at the required voltage. C2 provides "electronic smoothing" from the base circuit.

The 12 volt 15 watt bulb is for short circuit protection. With relatively low currents, little voltage is lost here. But if a short circuit arises in the output circuit, current is limited to a little over 1A by the bulb. This reduces the chances of any damage due to temporary shorts in leads or equipment.

R1 is a bleeder resistor, drawing some current even when no external load is connected. The meter is the usual high resistance type, scaled from zero to 12v. VR1 is adjusted until this instrument shows the wanted voltage.

The regulation of this supply is not so good as that of supplies using a number of transistors. But it will be found suitable for all ordinary purposes, and to supply power for circuits in this book, as only a small change in voltage arises, with changes to the load, or current drawn.

Fig. 47B shows practical construction of this PSU. A similar layout can be adopted for the other power supply units.

The mains cord is anchored at a tag strip, and connections are made as described earlier. The mains switch is in the L conductor to the 240v winding, or primary, of the transformer.

A tag strip supports the rectifier, which has AC (~), positive and negative connected as shown. The lamp can be a vehicle type, supported by stout wires soldered to its base contacts, or side and tip. A hole is drilled in the panel level with this lamp, so that it will give indications of a short circuit or heavy load.

The transistor is mounted on the metal panel, to enable this to carry away heat, but is insulated from the metal. Use the correct insulation set (mica washer and bushes) which can generally be obtained with the transistor. A tag under one of the securing nuts provides the collector connection.

For voltage readings, a meter scaled 0–12v is most suitable. Or a 1mA meter, with 12k 1% series resistor, may be used. If so, a new scale, marked 0–12, can be drawn for it.

A 3 watt linear potentiometer is used. Earth and negative are connected to the metal panel, for equipment with a negative ground line. The positive output terminal or socket is insulated from the metal.

The case can be metal, or wood. It should have a few ventilation holes. The transistor can be expected to run quite warm, at maximum current. Heating here rises as the output voltage is reduced by VR1, as in these circumstances more voltage is dropped in the transistor.

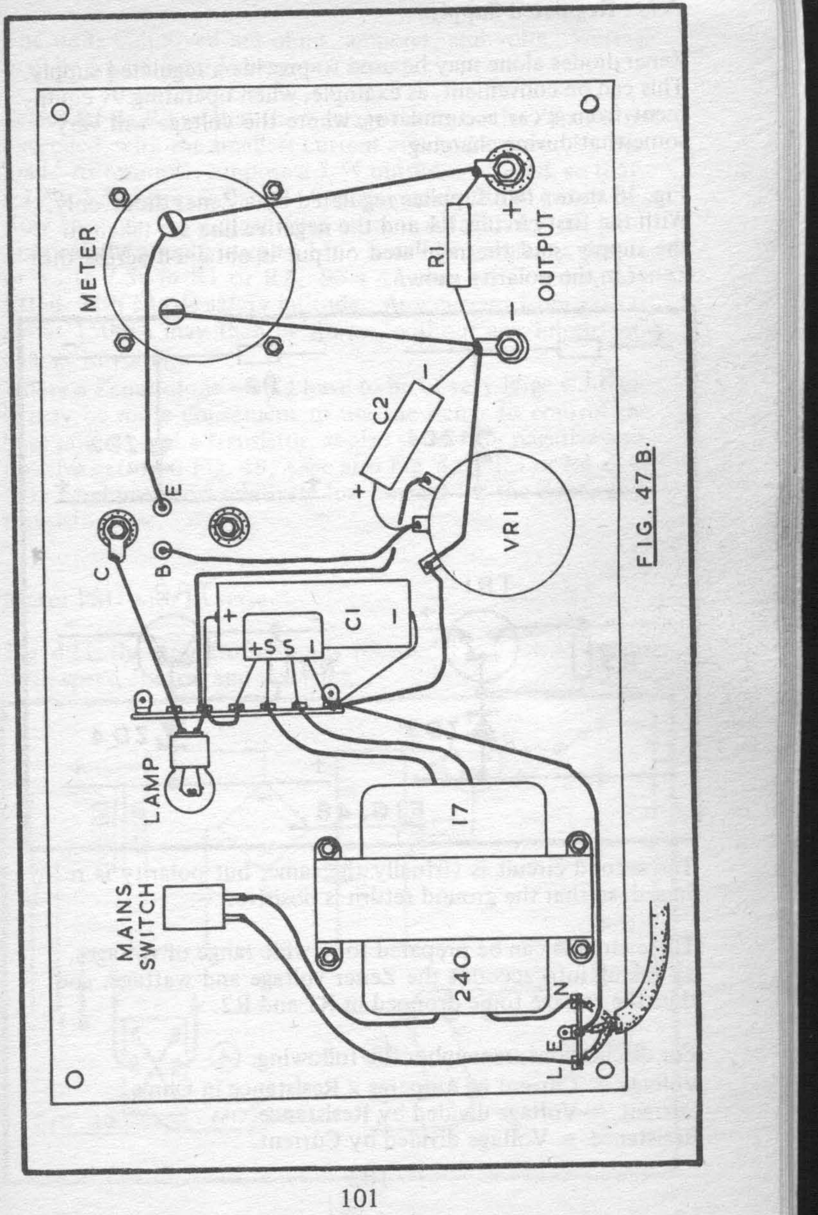

FIG. 47B.

Zener Regulated Supply

Zener diodes alone may be used to provide a regulated supply. This can be convenient, as example, when operating 9v equipment from a car accumulator, where the voltage will vary somewhat during charging.

Fig. 48 shows two supplies regulated by a Zener diode only. With the first circuit, R1 and the negative line are taken to the supply, and the regulated output is obtained across the Zener in the polarity shown.

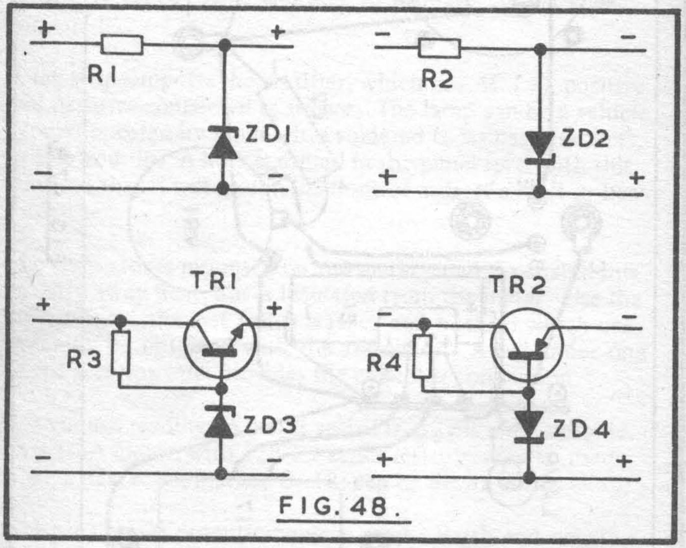

FIG. 48.

The second circuit is virtually the same, but polarity is reversed, so that the ground return is positive.

These circuits can be prepared for a wide range of voltages, by taking into account the Zener voltage and wattage, and thus the voltage to be dropped in R1 and R2.

For calculations, remember the following:
Voltage = Current in Amperes x Resistance in Ohms.
Current = Voltage divided by Resistance.
Resistance = Voltage divided by Current.

The units employed are ohms, amperes, and volts. Wattage may be found from Voltage x Current.

Select R1 or R2 so that the wattage rating of the Zener is not exceeded, with the smallest current ever to be drawn by the load. As example, suppose a 7.5v output is wanted, so that a 1½w 7.5v Zener is fitted. Approximately 200mA may thus flow through the Zener, if no current is taken by the load. Suppose the supply is up to 15v. It will be necessary to drop or waste 7.5v in R1 or R2. So a 42 ohm resistor may be fitted, with a little safety latitude. Any current from zero to about 150mA may then be drawn, without any important change in voltage.
Where a Zener diode would have to be of very large wattage, it may be more convenient to use the Zener to control the base potential of a transistor, as also shown for negative and positive earths in Fig. 48. (See also Fig. 45.) R3 or R4 then only need provide a relatively low current for the Zener and transistor base.

Motor PSU with Reverse

Fig. 49 is the circuit of a supply for model motors, and giving both speed control and reversing.

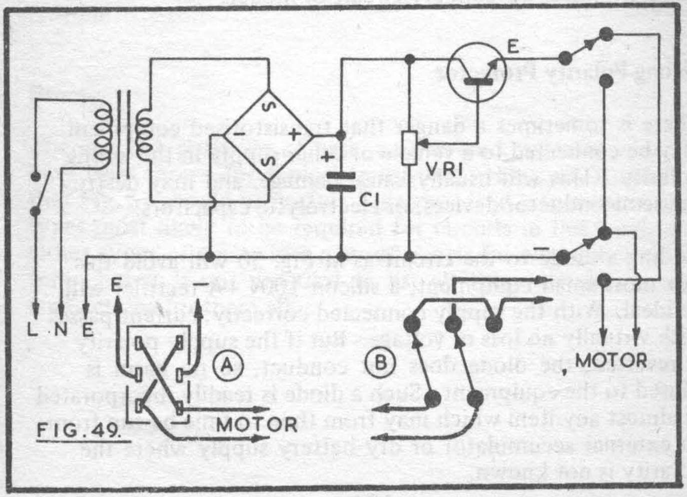

FIG. 49.

A transformer with a 12v secondary will be suitable for the usual model requiring up to 12v. The circuit is similar to that of the supply in Fig. 47. The potentiometer VR1 gives speed control. C1 can be around 2,000μF.

"A" is one type of ready-made switch, and it can be obtained with a central off position. "B" shows a home-built switch, similar to that described in Fig. 5. The switch must not short circuit the supply when changing over polarity, and one with central off is recommended.

Motors reversed in this way must be of the permanent magnet type, in which a change of polarity causes the armature to run in the opposite direction. Train and many other model motors are of this construction. Motors with a wound field will not be reversed by changing the polarity of supply.

A controller of this kind should be constructed as shown earlier, so that it is safe to use, and so that no joints in the mains circuits can be touched.

Earthing can only be omitted when a double-insulated transformer, of the type specifically made to avoid any need for earthing, has been used. The whole controller should be so constructed that, even if a fault develops, mains voltages cannot reach the secondary circuit or model.

Wrong Polarity Protector

There is sometimes a danger that transistorised equipment may be connected to a vehicle or other supply in the wrong polarity. This will usually cause damage, and may destroy the semiconductor devices, or electrolytic capacitors.

Adding a diode to the circuit as in Fig. 50 will avoid this. For most small equipment, a silicon 100v 1A rectifier will be ideal. With the supply connected correctly, current passes with virtually no loss of voltage. But if the supply polarity is reversed, the diode does not conduct, so no harm is caused to the equipment. Such a diode is readily incorporated in almost any item which may from time to time be run from an external accumulator or dry battery supply where the polarity is not known.

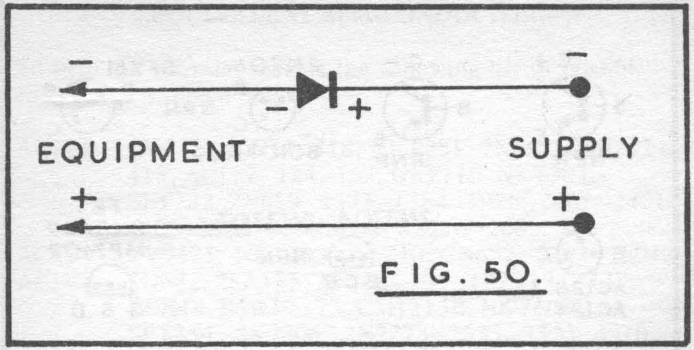

FIG. 50.

Component Tolerances

For circuits in this book, exact component values are generally not too important. Capacitor values such as 47nF (0.047μF) and 0.05μF may be regarded as interchangeable. In the same way, a 125μF capacitor may replace 100μF, or 200μF can be used instead of 220μF, and so on.

Resistors will generally be the standard 5% tolerance types. In some circuits exact resistor values are not at all critical. In others, and especially base, emitter and collector circuits, the values which provide best results can have much less latitude.

Bases

The lead-out positions of base, emitter and collector wires have been shown for typical transistors. However, it is felt that Fig. 51 should be useful. This gives lead information for types most likely to be required for circuits in this book. Other types of transistors may of course be used successfully, but it is not practical to list all these, or show connections for them all.

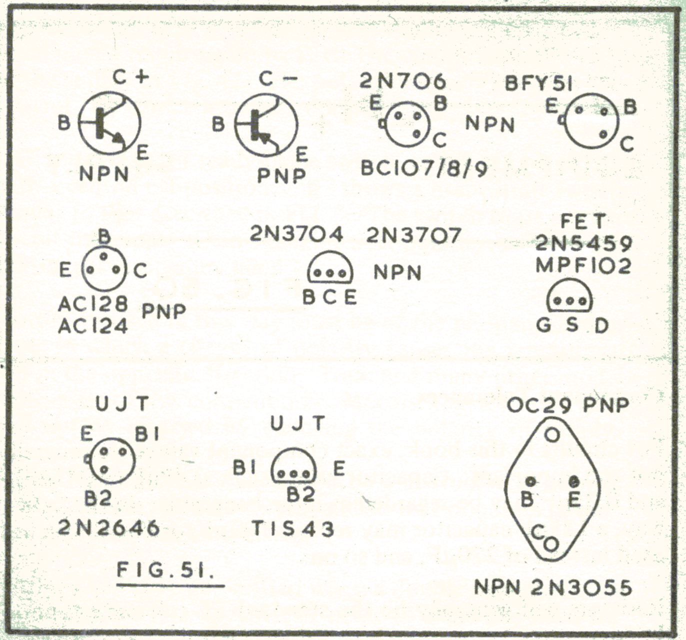
C +
B
E
NPN
C –
B
E
PNP
2N706
E B
C
BC107/8/9
NPN
BFY51
E B
C
B
E C
AC128
AC124
PNP
2N3704 2N3707
B C E
NPN
FET
2N5459
MPF102
G S D
UJT
E B1
B2
2N2646
UJT
B1 E
B2
TIS43
OC29 PNP
B E
C
NPN 2N3055

FIG. 51.

EQUIVALENT SEMICONDUCTORS

To help our overseas readers the following list of possible equivalents is shown below: –

AC128	CV9778, OC74–318, NKT281, 2N467, 2SB222–415, AC117–124–153, GET110, NKT302, SFT232, 2N659–1373–1384–1926–2001–2431–4106, 2SB370, SK3004, RS276–2005.
ACY39	CV9475, NKT243, ACZ10, 2N652–2303–2801.
AF117	2N642, 2SA155, AF102–105–127–133–137, BFX48, BSW19–73, GM1213B, NKT603F, SFT354, SK3006, 2N2273–2635–3324–4916–5354, RS276–2003.
BC107	CV9780, AM251, BC147–167–207–317, MPS6566, SK3020–3122, ZTX107, TT107, RS276–2009.
BC108	CV10541, AM252, BC148–168–208–318, MPS6520, SK3020, ZTX108, TT108, RS276–2009.
BC109	CV10769–10806, AM253, BC149–169–209–319, MPS6521, SK3020, ZTX109, TT109, RS276–2009.
BC147	BC171–182/L–207–317, MPS6566, SK3020, RS276–2013,
BC148	BC172–183/L–208–318, MPS6520, SK3020, BS276–2009.
BC149	BC184/L–209–173–319, MPS6521, SK3020, RS276–2009.
BF200	BF162–173–225–271–311–314, SE5020/1/2/3/4, SK3018, 2N616–2708–3288–3693, RS276–2011.
BFY51	CV7723–7726–10827–11238, BFX51–68A, BFR19, BSX45/X, BSY46–85, SK3024, 40347, 2SC708, ME6102 2N2193/4–2218–2410–3053–3252–4046, RS276–2009.
BSY95A	CV7648–9379, BC108A, BSX51–73–91, BFX96, BSY20–58–62, ME9022, SK3122, 2SC321, 2N744–914–1708–2218–2369–5187, RS276–2011.
MPF102	E305, SK3116.
OA81	AA117–118–132–144, AAZ15, OA91–95–161, SD38, 1N38–56–476–618, 1S33.
OC26	CV8560, NKT452, 2N257–297–458–553–1021, 2SB3, AD130–131–138–140–149, OC28, SK3009, TF80, T1156–3027, 2N174–1073–2870–RS276–2006.

OC29	CV7083–8356, NKT402, 2N457, 2SB86–425, AD150, ASZ16, AUY21, OC28–35–36, SK3009, T13028, 2N174–1073–2870, RS276–2006.
OC44	CV5710–7003, OC170–171–410–613, 2N1303, 2SA15, AF101–117–126, ASY55, AC191, BC126, NKT211, SKT358, SK3005, 2N36–1191/2–1352–1373–1384, RS276–2003.
OC70	CV5457, OC303–602, NKT214, 2N279–1305, 2SB75, AC122–125–151–163, ACY27–34, ASY27, BC213–206A, BSX36, SK3003, 2N1193–1352–1375–2429, RS276–2004.
OC71	CV5712–7005–8344, OC70–304–604, NKT214, 2N280–1305, 2SB77, AC122–125–151–163, ACY35, BSX36, OC3041, SK3003, 2N1193–1352–1375–1384–2429, RS276–2004.
OC72	AC122/3–125/6–131–151 IV–152–162–191, BC126–213–206A, NKT211, SFT353, 2N282–1190–1352–1371–1384–1991–2431, CV5713–7006–8440, RS276–2005.
OC76	CV8314, OC307–6025, 2N284–1305, NKT212, 2SB89, AC128–131–151–152, ACY28, ASY58–76, BC126–213–260A, SK3004, 2N394–1191–1352–1371–1384–1991, RS276–2004.
ORP12	RS276–116, LDR03, RPY25.
1N4001	BY127, EM502, RS276–1101/1135, 1S100.
1N4002	BY127, EM502, RS276–1102/1136, 1S100.
2N1306	CV7350–10686, NKT736, 2N634–635–636–1891–1993–1994–1995, ASY26–29, BSX19–20, RS276–2001.
2N1613	CV7440–8843, BFY33–34–67, BSY44, 2N698–699–717–1711–1889–1890–1893, BFY41, BSX45/6, BSY53, MPS6530, 2SC708, 2N2101.

N.B.: Although equivalent semiconductors may have similar electrical properties, physical dimensions may be different and this must be born in mind if space is tight and for mounting details.

Remember polarities if replacing PNP with NPN types and vice versa. If in doubt always be advised by your dealer for suitable equivalent semiconductors.

Notes

Please note overleaf is a list of other titles that are available in our range of Radio and Electronic Books.

These should be available from most good Booksellers, Radio Component Dealers and Mail Order Companies.

However, should you experience difficulty in obtaining any title in your area, then please write directly to the publishers enclosing payment to cover the cost of the book plus adequate postage.

If you would like a copy of our latest catalogue of Radio & Electronics Books then please send a Stamped Addressed Envelope.

BERNARD BABANI (publishing) LTD
The Grampians
Shepherds Bush Road
London W6 7NF
England

160	Coil Design and Construction Manual	£1.95
202	Handbook of Integrated Circuits (IC's) Equivalents & Substitutes	£1.95
205	First Book of Hi-Fi Loudspeaker Enclosures	£0.95
208	Practical Stereo and Quadrophony Handbook	£0.75
214	Audio Enthusiasts Handbook	£0.85
219	Solid State Novelty Projects	£0.85
220	Build Your Own Solid State Hi-Fi and Audio Accessories	£0.85
221	28 Tested Transistor Projects	£1.25
222	Solid State Short Wave Receivers for Beginners	£1.95
223	50 Projects Using IC CA3130	£1.25
224	50 CMOS IC Projects	£1.35
225	A Practical Introduction to Digital IC's	£1.75
226	How to Build Advanced Short Wave Receivers	£1.95
227	Beginners Guide to Building Electronic Projects	£1.95
228	Essential Theory for the Electronics Hobbyist	£1.95
RCC	Resistor Colour Code Disc	£0.20
BP1	First Book of Transistor Equivalents and Substitutes	£1.50
BP2	Handbook of Radio, TV, Ind & Transmitting Tube & Valve Equivalents	£0.60
BP6	Engineers and Machinists Reference Tables	£0.75
BP7	Radio and Electronic Colour Codes and Data Chart	£0.40
BP14	Second Book of Transistor Equivalents & Substitutes	£1.75
BP24	52 Projects Using IC741	£1.75
BP27	Chart of Radio, Electronic, Semi-conductor and Logic Symbols	£0.50
BP28	Resistor Selection Handbook	£0.60
BP29	Major Solid State Audio Hi-Fi Construction Projects	£0.85
BP32	How to Build Your Own Metal and Treasure Locators	£1.95
BP33	Electronic Calculator Users Handbook	£1.50
BP34	Practical Repair and Renovation of Colour TVs	£1.95
BP36	50 Circuits Using Germanium, Silicon and Zener Diodes	£1.50
BP37	50 Projects Using Relays, SCR's and TRIACS	£1.95
BP39	50 (FET) Field Effect Transistor Projects	£1.75
BP42	50 Simple L.E.D. Circuits	£1.50
BP43	How to Make Walkie-Talkies	£1.95
BP44	IC 555 Projects	£1.95
BP45	Projects in Opto-Electronics	£1.95
BP47	Mobile Discotheque Handbook	£1.35
BP48	Electronic Projects for Beginners	£1.95
BP49	Popular Electronic Projects	£1.95
BP50	IC LM3900 Projects	£1.35
BP51	Electronic Music and Creative Tape Recording	£1.95
BP52	Long Distance Television Reception (TV-DX) for the Enthusiast	£1.95
BP53	Practical Electronics Calculations and Formulae	£2.95
BP54	Your Electronic Calculator and Your Money	£1.35
BP55	Radio Stations Guide	£1.75
BP56	Electronic Security Devices	£1.95
BP57	How to Build Your Own Solid State Oscilloscope	£1.95
BP58	50 Circuits Using 7400 Series IC's	£1.75
BP59	Second Book of CMOS IC Projects	£1.95
BP60	Practical Construction of Pre-amps, Tone Controls, Filters & Attenuators	£1.95
BP61	Beginners Guide To Digital Techniques	£0.95
BP62	Elements of Electronics – Book 1	£2.95
BP63	Elements of Electronics – Book 2	£2.25
BP64	Elements of Electronics – Book 3	£2.25
BP65	Single IC Projects	£1.50
BP66	Beginners Guide to Microprocessors and Computing	£1.95
BP67	Counter, Driver and Numeral Display Projects	£1.75
BP68	Choosing and Using Your Hi-Fi	£1.65
BP69	Electronic Games	£1.75
BP70	Transistor Radio Fault-Finding Chart	£0.50
BP71	Electronic Household Projects	£1.75
BP72	A Microprocessor Primer	£1.75
BP73	Remote Control Projects	£1.95
BP74	Electronic Music Projects	£1.75
BP75	Electronic Test Equipment Construction	£1.75
BP76	Power Supply Projects	£1.95
BP77	Elements of Electronics – Book 4	£2.95

BP78	Practical Computer Experiments	£1.75
BP79	Radio Control for Beginners	£1.75
BP80	Popular Electronic Circuits – Book 1	£1.95
BP81	Electronic Synthesiser Projects	£1.75
BP82	Electronic Projects Using Solar Cells	£1.95
BP83	VMOS Projects	£1.95
BP84	Digital IC Projects	£1.95
BP85	International Transistor Equivalents Guide	£2.95
BP86	An Introduction to BASIC Programming Techniques	£1.95
BP87	Simple L.E.D. Circuits – Book 2	£1.35
BP88	How to Use Op-Amps	£2.25
BP89	Elements of Electronics – Book 5	£2.95
BP90	Audio Projects	£1.95
BP91	An Introduction to Radio DXing	£1.95
BP92	Easy Electronics – Crystal Set Construction	£1.75
BP93	Electronic Timer Projects	£1.95
BP94	Electronic Projects for Cars and Boats	£1.95
BP95	Model Railway Projects	£1.95
BP96	C B Projects	£1.95
BP97	IC Projects for Beginners	£1.95
BP98	Popular Electronic Circuits – Book 2	£2.25
BP99	Mini-Matrix Board Projects	£1.95
BP100	An Introduction to Video	£1.95
BP101	How to Identify Unmarked IC's	£0.65
BP102	The 6809 Companion	£1.95
BP103	Multi-Circuit Board Projects	£1.95
BP104	Electronic Science Projects	£2.25
BP105	Aerial Projects	£1.95
BP106	Modern Op-Amp Projects	£1.95
BP107	30 Solderless Breadboard Projects – Book 1	£2.25
BP108	International Diode Equivalents Guide	£2.25
BP109	The Art of Programming the 1K ZX81	£1.95
BP110	How to Get Your Electronic Projects Working	£1.95
BP111	Elements of Electronics – Book 6	£3.50
BP112	A Z-80 Workshop Manual	£2.75
BP113	30 Solderless Breadboard Projects – Book 2	£2.25
BP114	The Art of Programming the 16K ZX81	£2.50
BP115	The Pre-Computer Book	£1.95
BP116	Electronic Toys Games & Puzzles	£2.25
BP117	Practical Electronic Building Blocks – Book 1	£1.95
BP118	Practical Electronic Building Blocks – Book 2	£1.95
BP119	The Art of Programming the ZX Spectrum	£2.50
BP120	Audio Amplifier Fault-Finding Chart	£0.65
BP121	How to Design and Make your Own P.C.B.s	£1.95
BP122	Audio Amplifier Construction	£2.25
BP123	A Practical Introduction to Microprocessors	£2.25
BP124	Easy Add-on Projects for Spectrum ZX81 & Ace	£2.75
BP125	25 Simple Amateur Band Aerials	£1.95
BP126	BASIC & PASCAL in Parallel	£1.50
BP127	How to Design Electronic Projects	£2.25
BP128	20 Programs for the ZX Spectrum & 16K ZX81	£1.95
BP129	An Introduction to Programming the ORIC-1	£1.95
BP130	Micro Interfacing Circuits – Book 1	£2.25
BP131	Micro Interfacing Circuits – Book 2	£2.25
BP132	25 Simple Shortwave Broadcast Band Aerials	£1.95
BP133	An Introduction to Programming the Dragon 32	£1.95
BP134	Easy Add-on Projects for Commodore 64 & Vic-20	£2.50
BP135	The Secrets of the Commodore 64	£2.50
BP136	25 Simple Indoor and Window Aerials	£1.95
BP137	BASIC & FORTRAN in Parallel	£1.95
BP138	BASIC & FORTH in Parallel	£1.95
BP139	An Introduction to Programming the BBC Model B Micro	£2.50
BP140	Digital IC Equivalents and Pin Connections	£3.95
BP141	Linear IC Equivalents and Pin Connections	£3.95
BP142	An Introduction to Programming the Acorn Electron	£1.95
BP143	An Introduction to Programming the Atari 600XL and 800XL	£2.50
BP144	Further Practical Electronics Calculations and Formulae	£3.75
BP145	25 Simple Tropical and M.W. Band Aerials	£1.95

BERNARD BABANI BP48

Electronic Projects for Beginners

- In this book the newcomer to electronics will find a wide range of easily made projects, many complete with actual component and wiring layouts. Furthermore, a number of projects have been arranged so that they can be constructed without any need for soldering and, thus, avoid the need for a soldering iron.
- This book which is written by the very experienced author Mr. F.G. Rayer is divided into four sections:–
  1. "No Soldering" Projects
  2. Miscellaneous Devices
  3. Radio and Audio Frequency
  4. Power Supplies
- An absolute 'must' for all beginners in electronics.

ISBN 0-85934-054-6

00195

£1.95

9 780859 340540

frequency, in *Hertz* (named after a German physicist and abbreviated to Hz).

Often in electronic engineering we need to draw a *graph* or picture of such a vibration and we can imagine this being done as in Fig.1(i) for the pendulum. With an inked brush fixed as shown and with the paper moving, a wavy line is drawn and if for example the paper moves at 1 centimetre (cm) each second and the pendulum swings at 1 Hz (that is one to and fro movement per second) then the graph drawn on paper would appear as in Fig.1(ii) where the centre line is marked with a time scale.

This last graph is the form which we use to show a vibration pictorially. It is generally called a *frequency*, in this case of 1 Hz.

Audio frequencies (that is, those which can be heard) range upwards from about 20 to 20,000 Hz, from the lowest notes of a piano or organ to the highest as of a piccolo. Something vibrating at twenty thousand times in one second may be difficult to appreciate but it is nothing compared with the frequencies at which electrons can be made to vibrate, a million million times in one second is quite commonplace. So we are going to be confronted with very large numbers which are fortunately simplified by the metric system for just as 1 kilometre is the same as 1000 metres so 1 kilohertz is the same as 1000 Hz and the upper audio frequency of 20,000 Hz becomes 20 kHz. (kilo = 1000).

Now we all know how quickly sound waves are lost as they travel through the atmosphere so radio broadcasting uses much higher frequencies which peculiarly enough are not lost to the same extent. These *radio waves* are made to "carry" the audio frequencies of the program as we shall see later but first we should ourselves get rid of the confusion people experience with radio transmissions labelled both in "frequency" and "wavelength".

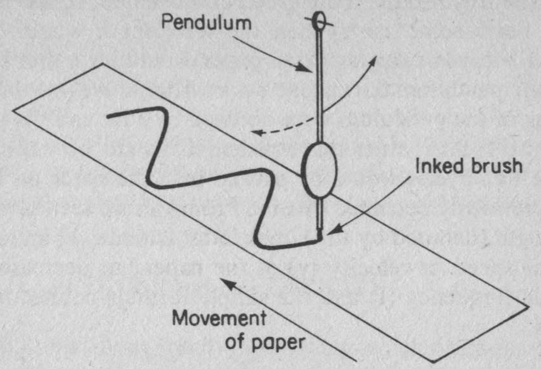

(i) Pendulum drawing its own graph

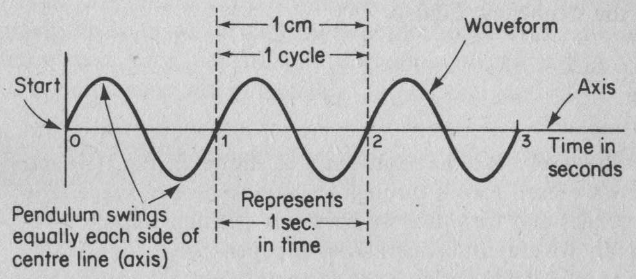
1 cm
1 cycle
Waveform
Start
Axis
0
1
2
3
Time in seconds
Pendulum swings equally each side of centre line (axis)
Represents 1 sec. in time

(ii) Graph when pendulum swings at 1 Hz and paper moves at 1 cm per second

Fig. 1 Pictorial representation of vibration

1.2 FREQUENCY AND WAVELENGTH

The dual labelling arises simply because the swing of the *waveform* of Fig.1(ii) can be described in two different ways, (i) as a frequency of 1 Hz or (ii) as the length along the axis of one cycle which is known appropriately as the wavelength. However the wavelength on the paper depends on how fast the paper moves and if we doubled the speed of movement to 2 centimetres per second (cm/s) then the wavelength would be *doubled* because twice as much paper would have moved under the pendulum during one cycle. Alternatively if the rate of swing of the pendulum were doubled to 2 Hz and the paper moved at 1 cm/s, then the wavelength would be *halved* because two cycles would be packed into the space on the graph previously occupied by one. From this we see that the wavelength (denoted by the Greek letter lambda, λ) increases with the speed or velocity (v) of the paper but decreases with the frequency (f) and the simple formula connecting them is

$$\lambda = \frac{v}{f} ,$$

that is, the wavelength is calculated by dividing the velocity by the frequency. Equally

$$f = \frac{v}{\lambda} .$$

Radio waves have the same form as shown in Fig.1(ii) and the wave itself travels through the atmosphere or space. The principles and formulae we have just examined apply but happily we are not concerned with paper speeds for all radio waves (and light) travel at the same unbelievable speed, three hundred million metres in one second (300,000,000 m/s – more than seven times round the Earth in one second!) So to find the wavelength of any radio transmission we simply divide 300,000,000 by the frequency (in Hz, not kHz), nasty numbers to deal with so the table in Appendix 1 is added as a guide, it covers the transmissions we will be tuning in on our crystal sets later. These range over the long and medium wavebands as follows:

Long Waves:	1000 – 1875m	(160 – 300 kHz)
Medium Waves:	190 – 550m	(545 – 1580 kHz).

As an example, a station broadcasting on 909 kHz has a wavelength of

$$\frac{v}{f} = \frac{300{,}000{,}000}{909{,}000} = 330 \text{ metres (330m).}$$

As radio constructors we shall often find it more convenient to work in frequency rather than wavelength even though many radio set tuning dials are only marked in the latter.

1.3 ELECTRONS and ELECTRICITY

We all know about *gravity*, something which brings us down to earth with a bump. We attract the Earth according to our size and the Earth attracts us according to its own size and we know which one wins. But we do not know how gravity works, only that it does. It is indeed fortunate that Nature put gravity there or we would disappear into space. It is a powerful but unseen *force*, the strength of which is around us always, we are said to be within its *field.*

These few remarks will perhaps help in appreciating another arrangement of unseen forces due to something we call *charges*. Whereas there is only one gravity which always attracts, there are two different charges which either attract or do exactly the opposite, that is, repel each other and this is according to which of the two types meet. They are labelled positive (+) and negative (–), and there is a simple law which governs their behaviour, "like charges repel, unlike attract".

Electricity is not new, it has been around since the world began. It is the movement of *electrons* which are tiny *negative* charges, within anything, the air if there is a flash of lightning, along mains wires in the home or even through us if we touch something "live". The stuff of which any substance is made (even us) is built up of *atoms*. Each atom has some electrons all to itself just like the Sun has its planets (the Earth is one) revolving round it. An atom with its full complement of electrons is neither positive nor negative because it has a *positively* charged centre which just balances the *negative* charges on its electrons. However some of these little ball-shaped electrons can escape from their parent atoms, a neighbouring positive charge attracts them towards itself or a negative charge pushes them away. When this happens a free electron moves on its own in the spaces between the atoms and because it has lost a negative charge, the atom becomes positive. When electrons are freed in this way and pushed in one direction, this constitutes an *electric current*. This is shown pictorially in Fig.2, where a battery is sending a current round a loop of wire. Chemical action within the battery has the effect of stripping electrons from their atoms and forcing them to crowd at the negative terminal, leaving positive atoms at the positive terminal. The more this happens, the greater the build-up of charge at each terminal and the difference in charge between the two terminals is called the *voltage*, in Fig.2 we have labelled it 3 volts. Thus higher voltage means greater charges which can therefore free and move more electrons in the wire, that is, cause a greater current.

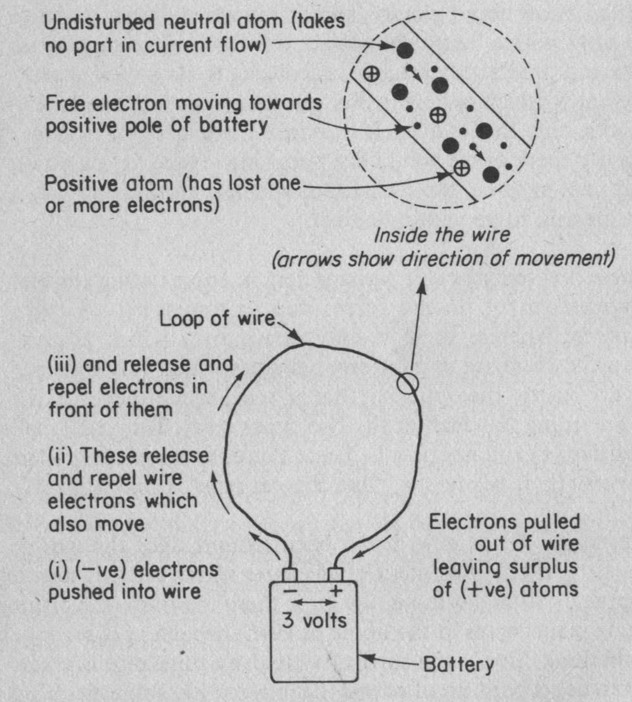
Undisturbed neutral atom (takes no part in current flow)
Free electron moving towards positive pole of battery
Positive atom (has lost one or more electrons)
Inside the wire (arrows show direction of movement)
Loop of wire
(iii) and release and repel electrons in front of them
(ii) These release and repel wire electrons which also move
(i) (−ve) electrons pushed into wire
Electrons pulled out of wire leaving surplus of (+ve) atoms
−
+
3 volts
Battery

Fig. 2 Electric current flow

The electron is tiny, so tiny in fact that we get lost in a sea of noughts when considering how many are involved in doing a job. Just as a single example, about 18 million million million (18 followed by 18 noughts) flow through a motor car headlamp bulb every second to light it. No wonder we can never see them.

So we can sum up by saying that an electron is a small charge of electricity normally living with an atom but when lots of them break away and move together in one direction, an electric current flows. When more exist at one point relative to another there is said to be a voltage between the points.

In Fig.2 the electron current is in one direction only and is known as a *direct current* (d.c.) If the battery connexions could be changed over at some very fast rate, electron flow in the wire would reverse its direction at the same rate, it is then known as an *alternating current* (a.c.) and audio and radio frequencies are such, that is, current flowing one way then the other at some frequency. As mentioned earlier we must get used to the idea that these tiny electrons can change direction millions of times in one second.

1.4 CONDUCTORS and INSULATORS

The fact that atoms give up electrons to make an electric current is the basis of electronic engineering, making currents of some predetermined size flow when and where we want. However not all atoms release electrons freely, in fact those of some materials hang on to them so tightly that releasing an electron is almost impossible. This divides materials broadly into two categories, *conductors* in which the atoms willingly give up electrons so that a current can flow, and *insulators* in which they do not. We therefore use a conductor to carry a current and wrap it with an insulator to stop the current going where it should not.

We rate the *ease* with which a material passes a current as its *conductance* but in practice we usually talk in terms of the opposite, its *resistance* (to current flow). Conductors have a low resistance, insulators high. Resistance is measured in *ohms*, (shorthand Ω, the Greek letter omega), practical values run from fractions of one ohm for a short length of wire to millions of ohms, expressed as megohms (MΩ) for insulators. In between there are kilohms (kΩ, 1kΩ = 1000Ω). Everything has a value of resistance, even we have.

We might see the relationship between voltage, current and resistance by considering a crowd of football fans at the stadium gates. The game is soon to commence and, anxious to get in, they create pressure on the gates. Each fan may be likened to an electron and those at the gates are being pushed by the agitation and anxiety of those behind. The pressure on the gates is the *voltage* but being closed, they present an infinitely high *resistance*. Nobody gets in (zero current). Now if the gates are opened slightly the gate resistance has fallen and some fans enter the ground (lower resistance, increased current). With the gates flung wide open everybody pours in (even lower resistance, high current). It is clear that the number of fans getting into the ground per minute (current) increases both with the pressure built up behind them (the driving force, voltage) and as the gates are opened more (lower resistance).

Copper is a good conductor and is used extensively in electronics for wires and components, silver is slightly better but

expensive. For insulators we can use for example, rubber, mica, most plastics, porcelain, glass or air. Thus wires usually consist of a copper conductor either bare or more usually coated with plastic (for example, electric light flex), current flows along the wire and it is confined to it by the insulation. Enamel is also an insulator so for many electronic purposes the insulation on a copper wire may consist of a thin coating of enamel only. This saves space but it must be treated with care for the enamel is easily damaged. In electronics diagrams when a wire is indicated for conducting current from one part or component to another we simply draw a single line and it is understood that the wire is either insulated or if not, supported in such a way that it touches nothing along its length. We will see how wires are drawn on circuit diagrams later.

There are many thicknesses or *gauges* of wire available from the hair-line to pencil thick. The world classifies them in different ways although generally the metric system is taking over in which a wire is described by its diameter in millimetres. Two other systems are still much in use, the *Standard Wire Gauge* (SWG) and the *American Wire Gauge* (AWG). Naturally we will use metric but where appropriate the nearest other gauges are also quoted.

Semiconductors are materials used in transistors etc. and they are neither good conductors nor good insulators, somewhere about half-way (semi = half).

1.5 RESONANCE

The pendulum mentioned in Section 1.1 is capable of keeping a clock accurate because it has a *natural* period of oscillation which depends only on its length (the shorter the pendulum, the faster it goes). It does not matter how hard it is pushed, eventually a pendulum settles down to the same time per cycle. A piano or any other musical string similarly has a natural period of oscillation but at a higher frequency, for example, for middle C the string vibrates at 262 Hz. The frequency in this case is determined by the thickness of the string and the tension in it and the piano tuner simply adjusts the tension to get the note right. When the note is played the wire is struck and set in vibration but it vibrates only at 262 Hz and this is called the *resonant* frequency. Similarly for all the other piano notes each wire has its resonant frequency. Resonance is a feature of all musical instruments by which they generate the delightful sounds we hear.

An electrical circuit can also resonate, that is, respond to one frequency only and two special components are required, an *inductor* and a *capacitor*. The inductor is a coil of wire and it has a distinctive property of *reactance* which affects a.c. only. The capacitor is a very different component from the inductor, it comprises metal plates which can hold a charge, it too has the property of reactance. We will not go into detail about these components yet because we will be looking at them more closely later. What is important here is that the inductor has one type of reactance which we label positive and the capacitor has the opposite type, so it is labelled negative. As with charges, a positive reactance can be cancelled out by an equal amount of negative reactance. The reactance of an inductor *increases* as frequency goes up but that of a capacitor falls and

> at low frequencies the reactance of an inductor is low, that of a capacitor is high
>
> at high frequencies the reactance of an inductor is high, that of a capacitor is low

so as frequency rises from low to high, with the reactance of the inductor rising and that of the capacitor falling, at some value they are both the same and cancel out.

Reactance, either positive or negative, is rather like resistance, it tries to stop current flowing so a circuit containing the two components can pass a high current at the frequency at which the reactances cancel but a lower current at all frequencies above and below this value because the reactances do not

cancel, there being more positive than negative or the other way round. The two components constitute a resonant or *tuned* circuit and we can begin to see how a radio receiver separates the wanted radio frequency from the others.

Let us now put this on a *circuit diagram* as in Fig.3. In diagrams the symbols are as shown and the shorthand for an inductor is L and for a capacitor, C. In Fig. 3(i) because both L and C have fixed values the circuit resonates at one frequency only. To make it resonate over a *range* of frequencies, for example the long or medium wavebands, we could vary either L or C and it will soon be evident that to vary C is much easier so Fig.3(ii) shows the complete tuned circuit we will construct. Variable capacitors are only available with certain maximum values so the inductors are wound to suit.

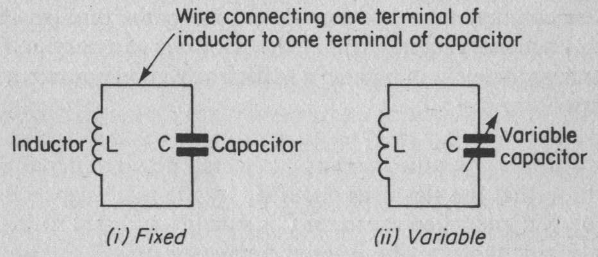
Wire connecting one terminal of inductor to one terminal of capacitor
Inductor
L
C
Capacitor
L
C
Variable capacitor
(i) Fixed
(ii) Variable

Fig. 3 Tuning circuits

1.6 AERIAL and EARTH

The drawing of Fig.3(ii) is therefore that of a tuned circuit capable of selecting one radio frequency out of the many offered to it. The radio waves are collected and delivered to this circuit by an *aerial* (also known as an *antenna*). Away at the radio transmitter we might imagine that the transmitting aerial is shaking a charge of electrons to and fro as we might shake a rope to make it take up a waveshape as in Fig.1(ii). This projects an alternating field outwards and just like the field of gravity, invisible but always there waiting to move anything it can, the radio field or wave causes the electrons in an aerial conductor to flow one way or the other according to the direction of the field at that instant. Technically we say that the field *induces* a voltage in the conductor. The degree to which the electrons flow depends on the strength of the field, the height of the aerial and how good a conductor it is.

Thus we erect an aerial which is simply a long copper conductor suitably insulated where necessary to prevent electrons leaking away. Because we are using no transistors to boost the aerial signals, as large a signal as possible must be collected and for this the earth itself is helpful as a return conductor back to the radio station. It may be surprising that the earth can be used in this way because some earthly materials such as sand and rock are actually insulators but it is obvious that there are countless other paths through the earth each carrying its share of the electrons of a current. One of our problems is in making a sufficiently good contact with "earth".

Fig.4 shows the tuned circuit connected between aerial and earth so that the electrons disturbed by the radio signals flow through it and when capacitor C is suitably adjusted so that the circuit tunes to the wanted frequency, this is the one which appears across the output terminals. Note the symbols used in diagrams for aerial and earth and how a connexion is indicated. A *connexion* implies two metals in such close contact that electrons flow unimpeded from one to the other, the metals are then said to be "in contact".

1.7 MODULATION

We have seen that a radio program consists of audio frequencies of a few kilohertz, yet the radio transmission is at

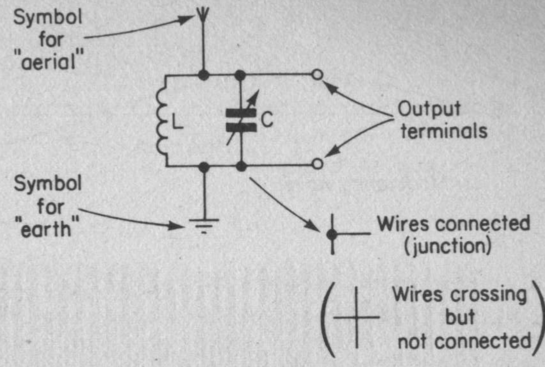

Fig. 4 Aerial feeding a tuned circuit

a much higher frequency, up to and over 1000 kHz. The radio frequency is called a *carrier* because in effect this is what it does, it carries the audio frequency on its back, so to speak. The process of getting it to do this is known as *modulation* and what happens is that the strength or *amplitude* of the radio-frequency (r.f.) carrier is caused to vary according to the amplitude of the audio frequency. This becomes more evident when shown pictorially as in Fig.5. The three parts of the figure are all graphs, on a basis of time, of how the amplitude of the modulating (audio), carrier and modulated waves vary. An *oscilloscope*, which is a measuring instrument with a small television-type screen, is the electronic counterpart of our rather crude method of making an oscillating device draw its own graph as in Fig.1(i) and this is how Fig.5 would be produced when the appropriate waves are applied to the input terminals. We take a practical example, that of the highest note of the piano (about 4 kHz) being transmitted on a carrier of 200 kHz (1500m) and consider a time period of one thousandth of a second (1/1000s), seemingly a very short time indeed but in which there are 200 complete cycles of the carrier wave.

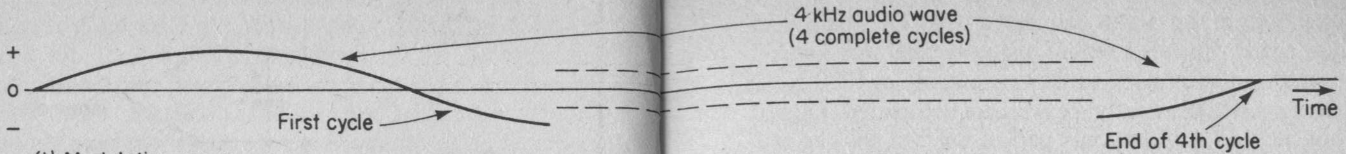
4 kHz audio wave
(4 complete cycles)
+
0
−
First cycle
Time
End of 4th cycle

(i) Modulating wave

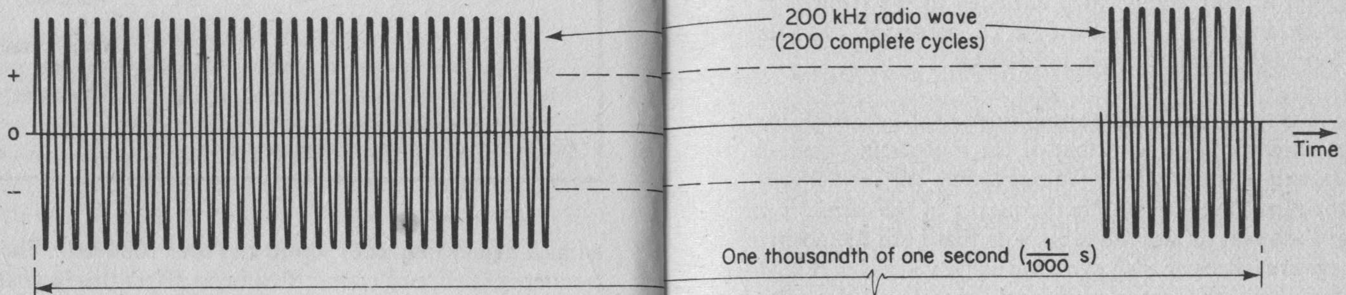
200 kHz radio wave
(200 complete cycles)
+
0
−
Time
One thousandth of one second ($\frac{1}{1000}$ s)

(ii) Carrier wave

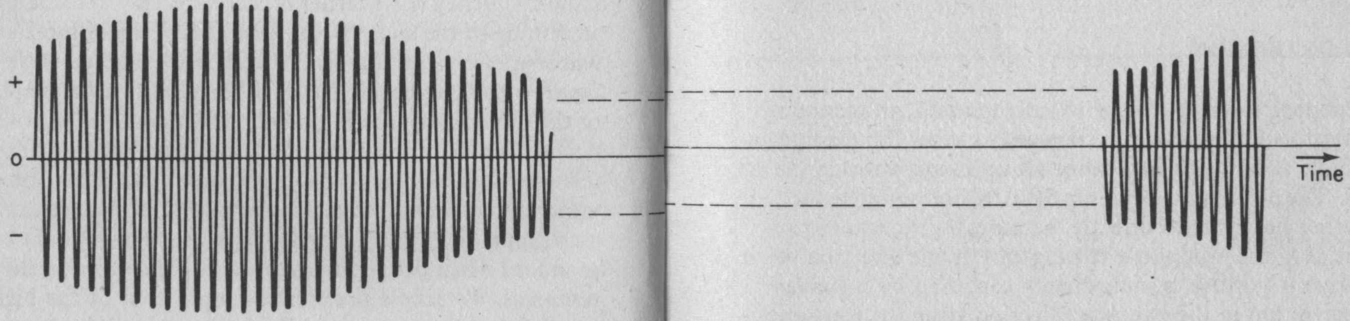

(iii) Modulated wave

Fig. 5 A modulated wave

A *microphone* listens to the sound waves generated by the piano wire. Its *diaphragm* vibrates at the frequency of the sound wave hitting it and causes electrons to move in alternate directions in the circuit connected to it. The result is an electrical signal which is a replica of the original piano wire vibration. In one second the audio wave completes 4000 cycles, therefore in 1/1000s exactly 4 cycles are completed. Fig.5(i) does not show all 4 cycles, only part of the first and the end of the last but this is sufficient to demonstrate the principle. For convenience it also shows the wave "cleaned up", the true graph is more complicated although of the same basic frequency. In Fig.5(ii) is shown the r.f. carrier at 200 kHz therefore 200 complete cycles in 1/1000s.

In the modulation process the amplitude of the carrier is made to vary in accordance with that of the modulating wave so from (i) and (ii) in the figure we can deduce (iii), and this is how the program is carried, the frequency of the carrier is not changed whereas its amplitude is. Note that both the positive and negative halves of each cycle are equally affected. Fig.5(iii) therefore shows the sort of waveform appearing at the output terminals of Fig.4 and what is needed next is a device to extract the audio frequency from the carrier.

1.8 DETECTION

Headphones work in reverse to microphones, an incoming electrical variation at audio frequency causes the diaphragm to vibrate accordingly and hence set up sound waves in the ear canal. The form shown in Fig.5(iii) is not suitable for operating headphones directly because the movement of electrons in the headphone driving unit in one direction when the wave is positive is immediately cancelled by a similar movement but in the opposite direction when the wave goes negative. Firstly one of these movements must be suppressed, it is then possible to find the average of all the unidirectional (all positive or all negative) half-cycles left. This brings us to the subject of rectification and a *rectifier* or *diode* is a device

which allows electrons to flow through it in one direction only, that is, it can behave either as a conductor or insulator according to which of its two terminals is presented with electrons. Hence if the wave of Fig.5(iii) is applied to a diode either all the positive half-cycles or the negative ones are blocked according to which way round it is connected. Effectively therefore in our example there would be 200,000 unidirectional *pulses* of current each second. But these pulses are not all equal in strength, they vary in amplitude according to the modulating frequency so by adding them together and somehow finding the average value as we go, the original modulating frequency is restored. A capacitor is suitable for this because it can store electrons on its plates and so add the various pulses together. Looking at Fig.5 again the *detection* process firstly removes either the top or bottom half of (iii), then by an averaging process regains (i). We can now complete our basic crystal receiver circuit diagram as in Fig.6. D is the diode and C_2 the capacitor just mentioned and except for the latter which we will find we can manage without, this is the circuit we will first build for experience.

All very well, but we are building a *crystal* set, where is the crystal? This in fact has gone out of fashion and its place is now taken by the diode, cheaper, smaller, more reliable and

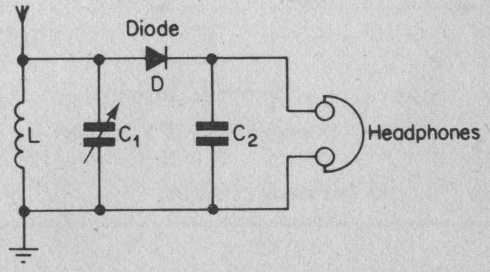

Fig. 6 Basic crystal (or diode) receiver circuit

needing no adjustment. Detection was accomplished in early receivers by means of a contact made between a fine-pointed wire and a crystal of galena, molybdenite or carborundum, each had its particular qualities. To obtain a suitable crystal/metal junction it was necessary to search around the crystal for a responsive spot using the end of the wire until the headphones sprang to life. This could take several minutes so woe to any member of the family thereafter causing sufficient rumpus for the contact to be lost. The wire became known as the *cat's whisker* and just to see what we have missed, Fig.7 shows a sketch of a crystal detector (minus its glass case) with a modern diode for comparison.

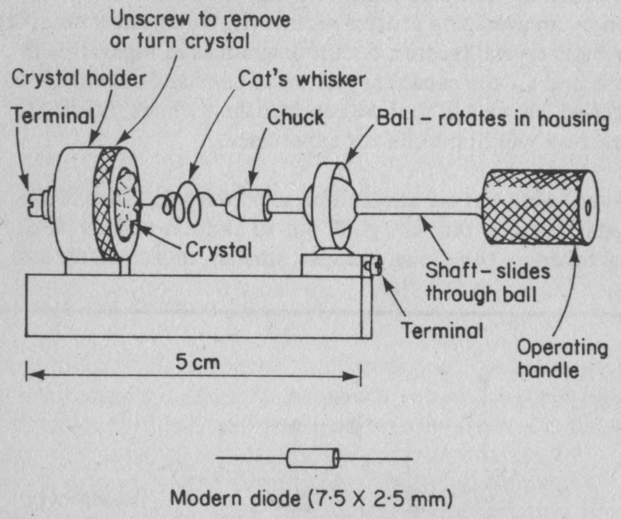

Fig. 7 Crystal and diode detectors

CHAPTER 2
COMPONENTS

Fig.6 shows the various components required for a basic receiver. Apart from the aerial and earth there is an inductor (L), variable capacitor (C_1), fixed capacitor (C_2), diode (D) and headphones. The inductor we will make ourselves although it could be bought (more of this later), the other components can be purchased from a radio component shop or mail-order service. The names and addresses are to be found in popular electronic or radio constructors' magazines, usually published monthly. Catalogues are so large nowadays that a charge may be made but to have one is an asset and note that the larger the catalogue, the more likely it is that the particular firm will be able to supply all the items required, this saves shopping around and postage. Nevertheless we must never forget that the defective radio set discarded by the family or friends, too expensive to repair yet somehow too good to throw away, can provide the experimenter with a plentiful supply of useful parts, especially the variable capacitor which we next consider.

So that we understand the components a little better, some notes follow.

2.1 CAPACITORS

To the newcomer a capacitor is a mysterious piece of equipment. It has no through connexion as do most other components, there is actually an insulator between its two sets of plates. Imagine two plates of metal placed close together as in Fig.8 and connected to a battery. As seen in Fig.2, a battery is a device which has a high pressure of electrons at its negative terminal and an equal lack of them at the positive terminal. When a battery is connected to the plates therefore there is a rush of electrons into the l.h. plate and equally out

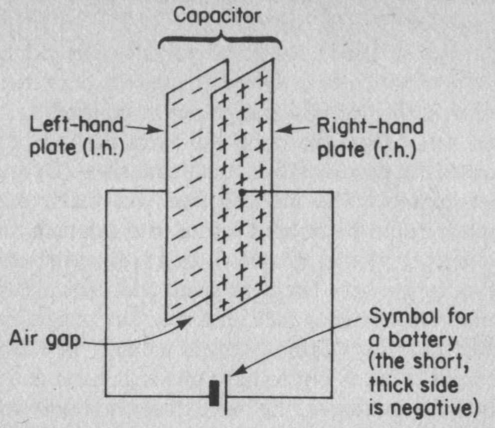

Fig. 8 Charging the plates of a capacitor

of the r.h. plate, in the latter case attracted by the surplus of positive atoms at the battery + terminal. The l.h. plate is therefore negatively charged and the r.h. positive, all in a very short time. There is no flow of electrons across the air-gap between the plates because air is an insulator so the flow ends when, for example, the l.h. plate has reached the same "electron pressure" as that at the battery negative terminal for then there are two equal negative charges opposing each other. The plates are now said to be *charged* and if the battery is disconnected they remain so. If subsequently a circuit or wire is connected across the plates then for the tiniest fraction of a second they act as a battery and drive electrons round until they are *discharged*.

Little imagination is required to see that *capacitance*, which is the capability of storing a charge, depends on the area of the plates. It also varies with the separation between them, reduced separation giving greater capacitance. The insulation between the plates (the *dielectric*) need not be air, other

insulating materials are used and the capacitance increased. The unit of capacitance is the *Farad* but generally in radio we work in *microfarads* (μF – μ is the Greek "mew") which are millionths of a farad or in *picofarads* (pF) each of which is one millionth of a microfarad. Just to add to our confusion some capacitors are now so small that there is no room to print the value in pF on them so the *nanofarad* is used which is one thousandth of 1 μF or equally one thousand pF, thus:

$$\text{nF} \times 1000 = \text{pF} \quad \mu\text{F} \times 1000 = \text{nF} \quad \mu\text{F} \times 1{,}000{,}000 = \text{pF}$$

so a 0.0005μF capacitor could equally be quoted as (0.0005 x 1000)nF, i.e. 0.5nF or (0.0005 x 1,000,000)pF, i.e. 500pF.

If required, C_2 in Fig.6 would be a *fixed* capacitor, with one value of capacitance only, say about 1000pF (or 1nF). By using thin dielectrics and plates, such a capacitor is no more than a bead a few mm across in the middle of its two connecting wires.

Variable capacitors such as C_1 are more complicated because they are mechanically operated. Those for tuning radio receivers have maximum values ranging from about 300 to 500pF. The capacitor works by enmeshing a set of movable plates (many small ones are connected together as the equivalent of a single large plate) with a second set of plates which are fixed as shown in Fig.9. The dielectric is usually air but solid dielectric types are also available which are more compact. A (tuning) knob is fixed to the shaft, by turning clockwise the moving plates interleave more with the fixed ones, so increasing the capacity. The moving plates are in contact with the frame and are therefore connected to the *earthy* side of the tuned circuit. The fixed plates are insulated from the frame.

If an old radio receiver is being *cannibalized* (dismantled to provide spare parts), the variable capacitor is easily recognized. Its spindle will be coupled to the tuning dial and it will look something like Fig.9 but duplicated or triplicated, that is two or three complete capacitors on one spindle. This is quite usable, we need only connect one section and improved designs shown later in the book need two sections anyway.

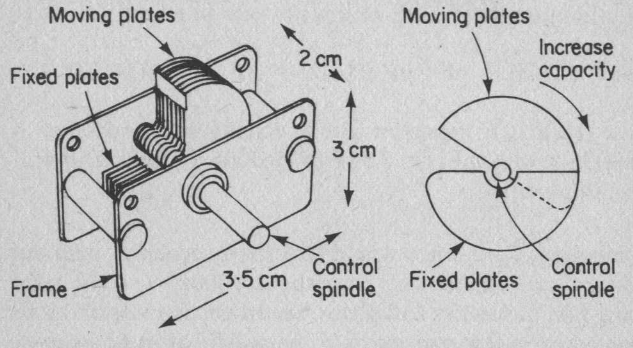

Fig. 9 Variable capacitor

2.2 INDUCTORS

These take many forms but at this stage we consider the *air-cored* type only. Ours will consist of enamelled wire close-wound on a cardboard or plastic tube, the number of turns of the wire being chosen to produce the correct *inductance*. This is measured in henrys, again a large unit for radio constructors so we work in *microhenrys* (μH) one of these being one millionth of 1 henry.

When electrons flow in one direction in a coil of wire, they create a magnetic field running through the centre which can turn a compass needle just as the Earth's magnetic field or that of a permanent magnet does. Again the field is invisible but the force it can exert is there just the same. If the electrons reverse direction, so does the field and if it attracted

one end of a compass needle in the first instance, it would now repel it. Of equal importance is the fact that when the current ceases, support for the magnetic field is removed and it collapses but in so doing, itself creates electron flow in the windings. In a resonant circuit the rise and fall of the magnetic field of the inductor is complementary to the discharging and charging of the capacitor. There is a to and fro movement of electrons between them which is greatest at the resonant frequency.

To cover the medium waveband (Section 1.2) an inductor of about 200μH is needed.

Enamelled wire is obtainable in over 60 different sizes, a good one to start with is 0.56mm diameter. Constructional details of coils follow in Chapter 4. Note that the general term "coil" is frequently used instead of "inductor" in radio tuning applications.

2.3 DIODES

Study of the diode, so called because it has two *electrodes* (di = 2) brings us into touch with the semiconductors on which modern transistor technology is based. Our explanation must of necessity be very simplified and it starts with the two fundamental types of semiconductor materials. These are *silicon* compounds which either have an excess of positive atoms or of negative electrons. The first is obtained by mixing silicon (a grey, brittle material) with a substance which "steals" electrons from some of the normally neutral atoms thereby leaving them positive, the final material is described as *p-type*. Alternatively a substance can be added which gives extra electrons to the silicon which is then classed as *n-type*. It is when p-type silicon is placed in contact with n-type silicon that a diode with its rectifying properties is formed. Fig.10 shows a battery connected across a diode with its junction between the two semiconductors shown diagrammatically, it of course does not have clean boundary lines as suggested.

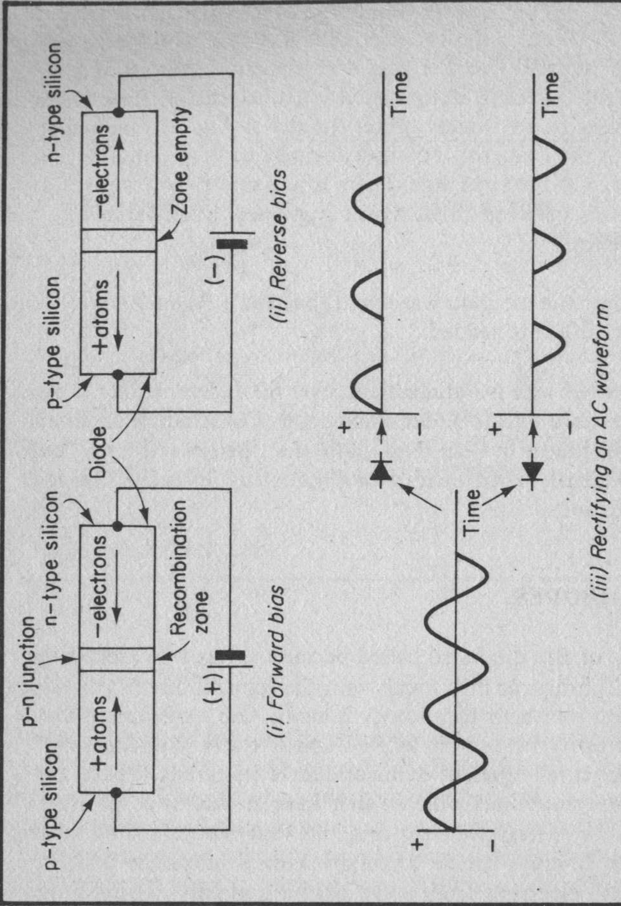

(i) Forward bias

(ii) Reverse bias

(iii) Rectifying an AC waveform

Fig. 10 Diode rectifier

In (i) the +atoms in the p-type and the −electrons in the n-type are forced straight across the junction by the similar charges of the battery. Some recombine to form neutral atoms in the junction area but the battery supplies plenty to make up for this. Thus current flows and the battery is then said to be connected in the *forward* direction.

(ii) When the battery connexions are reversed as shown, the *current carriers* (+ atoms and − electrons) are swept out of the junction whereupon no current can flow, that is, the device effectively becomes an insulator. This is the *reverse* direction.

Accordingly, if the battery potential is replaced by an a.c. waveform which in fact swings in both forward and reverse directions, the result is as shown in Fig.10(iii) where one half of each complete cycle is cut off.

We will see from the diode itself just how tiny the actual junction must be so if too many current carriers are pushed through the junction at once by applying too high a voltage, the heat created will burn it out. There are no such difficulties with a crystal set however because the voltages picked up by the aerial are minute. But we could be in trouble when soldering a diode wires onto tags because heat from the soldering iron travels along the wires to the junction, a golden rule therefore is not to cut the wires, to solder right at the ends and if possible use a *heat shunt* (see Appendix 3).

2.4 HEADPHONES

Resistance is mentioned in Section 1.4 and a crystal receiver needs headphones of 2000Ω resistance or more for most efficient working. Such headphones can be purchased but it is far better to make use of a type which is more available generally. Nowadays an increasing number of homes possess a pair of good quality stereo headphones, usually supplied with hi-fi systems, music centres, electronic organs and the like. This naturally saves purchasing them but if they do have to be bought, the low *impedance* types have many more uses for other projects. We can adapt them for ours.

We seem to have switched from the term "resistance" to "impedance", there is in fact a technical difference between the two even though they are both quoted in ohms, nevertheless as far as headphones are concerned we can look at both terms as meaning the same.

Stereo headphones are usually connected as shown in Fig.11 to a 6.5mm (¼ inch) jack plug. The wiring arranges for two separate *channels*, one to each earpiece. One wire is common to both so only three wires are needed. The plug has *tip*, *ring* and *sleeve* connexions insulated from each other as shown and we need to make our connexions to tip and ring only for the signal current then flows through the two earphones in *series* (that is, one after the other). This also happens to double the net headphone impedance to 16Ω since two 8Ω earphones in series gives 16Ω as shown in Fig.11. However we should not put 16Ω headphones directly on a circuit requiring a high impedance for some of the signal is wasted in so doing. What is necessary is to *match* the headphones to the receiver circuit by making them behave as though they have high impedance and this is effected by a *matching transformer*, miniature ones which are used between some transistor circuits and loudspeakers are suitable and inexpensive.

When electrons flow in a winding a magnetic field is set up and if another separate winding is within this field its electrons experience a force trying to move them. With a circuit connecting the two ends of the winding they can and do move and as we say, a current flows. This only happens while the current in the first winding is changing but this is the condition throughout crystal sets because all currents are alternating. A *transformer* consists of two or more separate windings which are in magnetic contact with each other, that is, they are close enough for the magnetic field of one to affect the other. Sometimes, as in the case of matching transformers, one winding is wound round the other with both on a metal core, alternatively as we shall find later (Section 5.5) many r.f. transformers have the two windings completely separated. The symbol for a transformer is shown in Fig.11.

Matching transformers convert impedances from one value to another and are sold quoting the impedances between

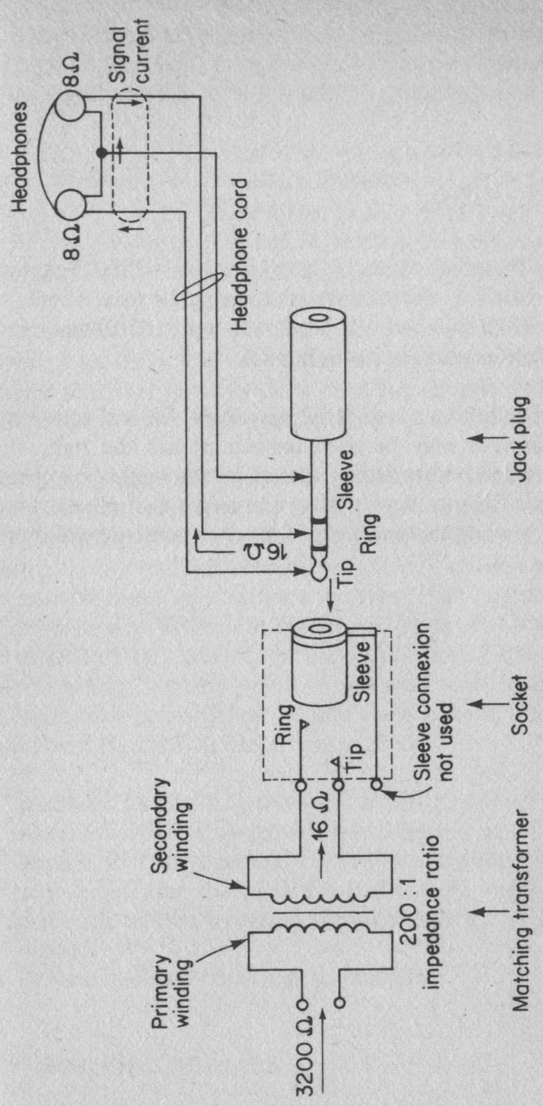

Fig. 11 Matching a high impedance receiver to low impedance headphones

which they should be used. We need not worry if we cannot get exact values as long as the impedance *ratio* is about right. Thus a transformer quoted as *primary* 1.0kΩ, *secondary* 5Ω (primary and secondary are the names of the windings) has a ratio of

$$\frac{1000\Omega}{5\Omega} = 200.$$

This ratio therefore raises 16Ω to 16 x 200 = 3200Ω as shown in Fig.11, the sort of value we are looking for (say, 1500 – 4000Ω). Thus the stereo headphones with transformer can replace high impedance headphones.

If stereo headphones cannot be borrowed, for real economy a crystal earpiece may be used instead, it has the right impedance and is surprisingly cheap, but as might be expected, much less efficient. Nevertheless one might usefully start with this type, leaving the purchase of more expensive headphones until later.

CHAPTER 3
AERIAL and EARTH

It is perhaps the two items at the input and output of the crystal set which affect the overall results most, the aerial/earth system and the headphones. In this chapter we discuss the former and how to install it, remembering that we are obtaining all the power for driving the headphones from the radio wave itself. We must therefore aim for the most efficient system possible but one which also suits our particular circumstances or location. The use of tiny aerials such as are found in portable receivers can only come by using circuits which are fed with extra power from batteries.

For our purpose aerial installations can be divided into three classes, in order of merit they are (i) outdoor, (ii) indoor loft and (iii) indoor room and the first decision must be as to whether we are building circuits simply for experience in the process of learning or whether a receiver is to be retained for use permanently. This naturally affects the types of wire used and the effort put into the installation, for example for a temporary installation a length of insulated wire slung around the room may be sufficient but we must bear in mind its limitations. Firstly two important reminders:

(i) recalling Section 1.6, the radio signal picked up by the aerial must flow through the tuned circuit, all of it. This implies that there should be no *leakage* paths to earth anywhere along the aerial for if there are, some signal electrons will be bypassed through these paths and be wasted.
(ii) the aerial should be as high as possible.

3.1 OUTDOOR AERIALS

Undoubtedly the best type although the most difficult to provide. Point (i) above reminds us that the aerial must be well insulated from earth all along its length. Air is an insulator so where an aerial runs through open space it can be of bare copper wire but insulated at its two supports. It so happens however that bare wire particularly suitable for outdoor aerials, for example 7 strand, of about 0.72mm (SWG 22, AWG 21 – stranding improves flexibility) may be difficult to obtain except from specialist wire firms so we will think in terms of insulated wire. This also happens to be easier to lead into the house and several kinds are available in most catalogues. Even twin-core plastic-coated mains flex or bell wire can be used quite cheaply as long as the wire itself is not likely to fracture under the tension of a long span. A typical outdoor installation is shown in Fig.12(i). With a little ingenuity and care that pipes or gutters are not pulled off the wall, fixings on both house and distant pole, tree or building can be made. Terminating the aerial wire on its insulators by twisting a loop round the span wire as shown in (ii) of the figure may be difficult with springy plastic-coated wires and here a binding of adhesive tape will help. A temporary insulated wire aerial needs no insulators and the wire can simply be tied at both ends.

The lead into the house must be insulated so if the aerial wire is bare an insulated lead must be jointed to it, preferably by soldering (see Appendix 3). Sometimes a ventilator or ill-fitting window or door will allow a thin plastic-insulated wire to be brought in without drilling through the woodwork (double-glazing makes life really difficult!), so much depends on the house or shed or wherever the receiver is to be used.

3.2 LOFT AERIALS

These are similar to the outdoor type but suspended within a loft or attic. Similar principles apply but in this case the down-lead is already in the house and may be brought down through the trap door or a tiny hole in the ceiling. Alternatively it may be run through the eaves and down the outside wall to be led

in as for the down-lead of an outdoor aerial. Preferably it should not be cabled down the wall but spaced from it.

3.3 ROOM AERIALS

Most likely to be used owing to their simplicity, these are unfortunately the least efficient. Nevertheless many readers will prefer to forego the rigours of outdoor or loft aerials and use this type. An insulated wire is simply run round the walls of the room (preferably an upstairs one) at picture rail or ceiling height back to the starting point at which the down-lead is taken. The wire is fixed by adhesive tape or preferably run through insulated screw hooks to keep it a short distance from the wall. The general arrangement is shown in Fig.12(iii).

3.4 EARTH

A good connexion to earth is essential and is most easily accomplished via a cold-water pipe (not gas, some of the joints may not conduct well) and a check should be made that the pipe does actually disappear into the ground and not find its way upwards to a plastic cold water tank in the attic. The pipe must be scraped or sandpapered clean and a connexion made to it by use of a special *earth clip* which is most likely to be available from an electrical (mains wiring) supplier. Alternatively a good connexion can be obtained by wrapping the earth wire round the pipe several times and then tightly twisting the free end over the main lead as shown in Fig.12(iv). For added security the loops of wire round the pipe may be pressed more tightly into contact by enclosing within a hose-clip of suitable dimensions. Soldering is preferable but not easily carried out because firstly the pipe must be emptied of water and secondly a very large iron or gas torch would be required to heat the mass of metal sufficiently for solder to flow.

Instead of a water-pipe connexion, a home-made earth spike

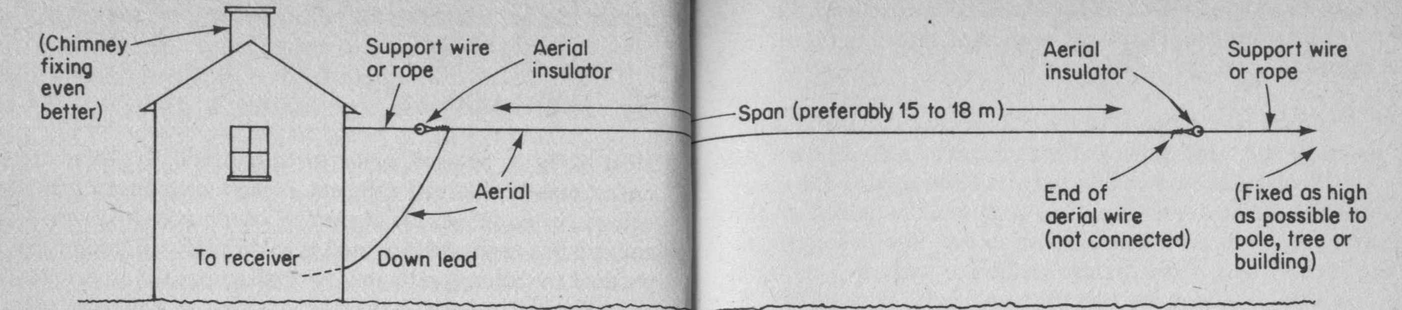
(Chimney fixing even better)
Support wire or rope
Aerial insulator
Span (preferably 15 to 18 m)
Aerial insulator
Support wire or rope
Aerial
To receiver
Down lead
End of aerial wire (not connected)
(Fixed as high as possible to pole, tree or building)

(i) Outdoor Aerial

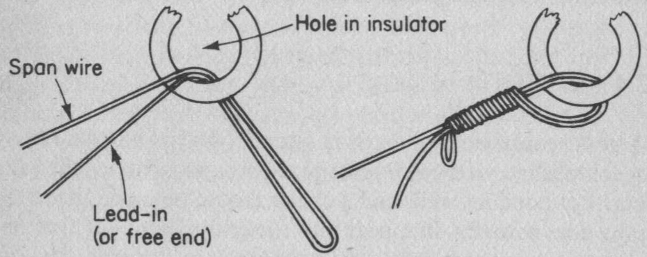
Hole in insulator
Span wire
Lead-in
(or free end)

(ii) Attaching aerial to insulator

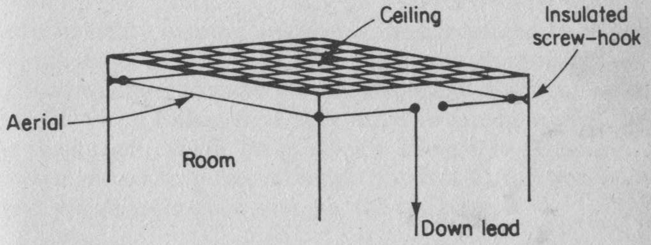
Ceiling
Insulated screw-hook
Aerial
Room
Down lead

(iii) Room aerial

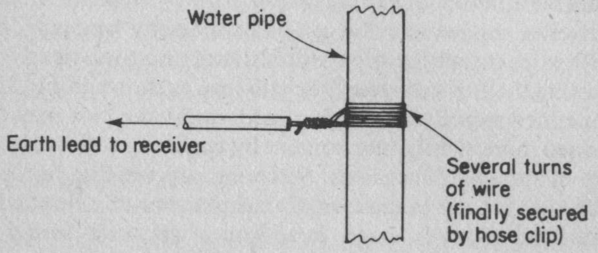
Water pipe
Earth lead to receiver
Several turns of wire (finally secured by hose clip)

(iv) Earth connection to water pipe

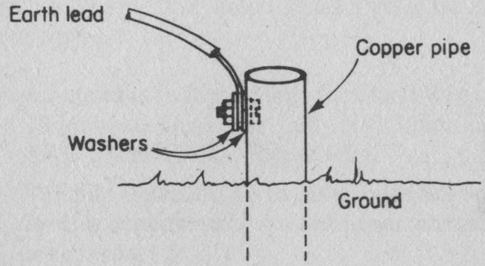
Earth lead
Copper pipe
Washers
Ground

(v) Lead bolted to earth spike

Fig. 12 Aerial/earth systems

consisting of a metre or so of copper pipe can be driven into previously soaked ground and connected to the earth wire as above or by using a bolt, nut and washers as shown in (v) of the Figure.

CHAPTER 4
BUILDING AN ELEMENTARY RECEIVER

There is nothing to stop us building an advanced receiver first, for example, one of the circuits following later. But since our aim is not only to build sets but also to have at least a little understanding of them it is more sensible to start with the simplest. This will give us practice in wiring and even more important, produce something against which improved models can be assessed. Thus assuming that we now have an aerial and earth at our disposal and are capable of soldering (Appendix 3), we build the circuit of Fig.13(i) which is in fact that of Fig.6 minus C_2 and with the aerial connected just over half-way down the inductor. C_2 may be omitted because the inertia (inability to move quickly) of the headphone diaphragm does the job as well. Each r.f. pulse (as in, say, the top half of Fig.5(iii) will tend to move the diaphragm and it will be helped by those which follow since they arrive in rapid succession. The diaphragm cannot move quickly enough to restore between successive pulses although it is able to follow the slower variations in pulse strength, the audio signal.

This is to be a medium wave receiver and for this, using a variable tuning capacitor of about the average value of those available, i.e. 365pF, the inductor has a value of 230μH so our list of parts is:

L	Air-cored inductor, 230μH (details follow). Requires 18 grammes (just over ½oz) of 0.56mm (SWG 24, AWG 23) enamelled copper wire.
C	Variable capacitor, up to 500pF (see Section 2.1 – for this experimental receiver lower maximum values are in order).
D	Signal diode, AA119, OA91 or equivalent.
H	*Either* crystal earpiece and appropriate chassis or line socket (most likely 3.5mm)

or Pair of 8ohm stereo headphones with appropriate stereo chassis or line socket [most likely 6.5mm (¼inch)].

T Miniature audio output transformer (see Section 2.4 and Fig.11).

Also: tuning knob, connexion or tag strip, "hook-up" wire.

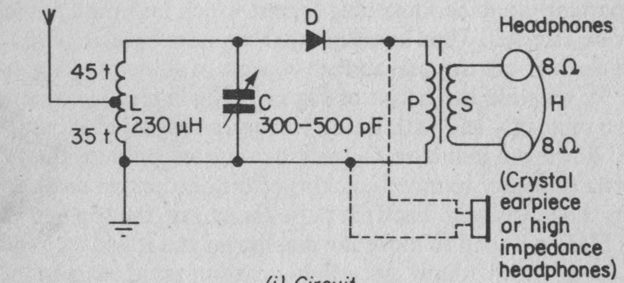

(i) Circuit

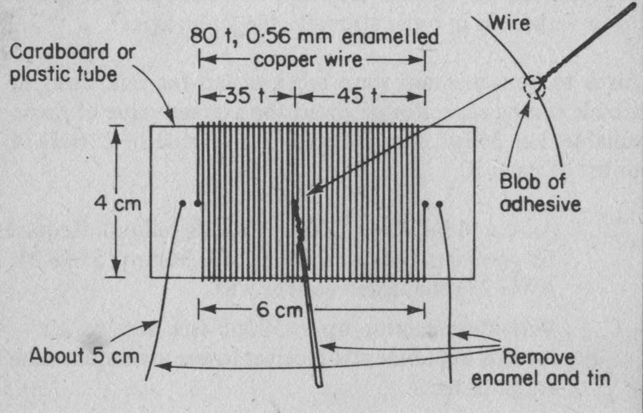

(ii) Inductor

Fig. 13 Elementary crystal set circuit and inductor

For the inductor a small quantity of enamelled wire, say 50-60 grammes (2oz) can usually be purchased, this is ample. 80 turns are wound on a 4cm (1½inch) diameter cardboard tube, we should not need to make one because they abound in plenty as the centre tubes of toilet or kitchen paper rolls. Failing this, any cylindrical plastic container having approximately this diameter and a length of at least 8cm will be suitable. Fig.13(ii) shows details of the inductor construction. Near one end of the tube two small holes about 1cm apart are made for anchoring the wire (a little clear adhesive helps) and some 5cm of wire is left for terminating. 35 turns are carefully wound along the tube, each turn just touching the previous one. Remember the enamel insulation is thin and easily scratched off. At this point a *tapping* is made as shown and again a little blob of adhesive holds the wire. The remaining 45 turns are then wound on, terminating through two holes as before. Scraping off the enamel and tinning the ends of the terminating wires and tapping loop completes the inductor. Should the end turns tend to loosen and ride along the tube, hold in position by pushing a pin through the tube while some adhesive over the wire and tube sets. We shall find after having built this set that modifications to improve it need additional tappings on the same type of coil. Readers who feel competent to tackle the hazards of seven tappings on one winding may prefer to wind the improved coil instead, simply leaving the tappings not required at this stage unused. See Section 5.1 and Fig.16. Precision is hardly needed so wires of diameter within about 10% of 0.56mm (1 gauge SWG or AWG either way) are quite suitable.

The capacitor is discussed in Section 2.1. If one can be obtained from an old radio set, so much the better. We need only a single-gang for this experimental model but some later ones will require 2-gang, leave unused gangs unconnected and check that no moving vane touches a fixed one. If a capacitor has to be purchased it will probably be found that the solid dielectric type is cheaper. Do not be too concerned about the maximum value, Section 5.3 shows that the coil can be adjusted to suit if required.

The tuning knob can be of any type, a large one (4-5cm diameter) makes fine tuning a little easier.

4.2 CONSTRUCTION

The component lay-out and wiring can to a great extent be to individual choice so what follows is one suggestion only, any other arrangement is equally suitable as long as the connexions are the same, that is, Fig.13(i) is followed. In true *breadboard* fashion a piece of wood at least 15 x 11 cm forms the base as shown in Fig.14, its thickness being such as to accommodate the screws holding the capacitor bracket and the terminal block. It is most likely that the capacitor bracket will have to be made, alternatively a small piece of plywood or hardboard nailed or screwed to the edge of the board and drilled for the capacitor to be fitted will suffice. As a general aid in connecting components together a 6-way mains terminal block is shown, these are usually purchased in 12-way flexible units which can be cut down as required. If desired a *tag-strip* may be used instead, all connexions on it being soldered instead of being held by screws. "Hook-up" wire can be purchased but it is possible to make do with many other types such as bare tinned copper, spare enamel-covered or even pieces of single core mains flex or bell wire, ends tinned where necessary. Remember to refer to Appendix 3 if there are still problems with soldering.

Fig.13(i) shows two alternative output arrangements, the crystal earpiece may be used in preference in which case the wires going to the primary winding of the audio output transformer are connected directly to terminals 4 and 5 of the terminal block instead, the transformer being omitted. The diode can be connected either way round, it does not matter whether we end up with the top or bottom half of Fig.5(iii), the result is the same. The miniature audio output transformer will probably fit in as shown, the tags of the secondary winding being connected directly into the terminal block. It is the secondary winding which is connected to the headphones, this is the winding marked with the lower impedance of the two to be matched. A centre tap on the primary is left unconnected.

Should a chassis rather than a line socket be required for plugging in headphones or a crystal earpiece, the appropriate chassis-type socket may be fitted on the baseboard or front panel.

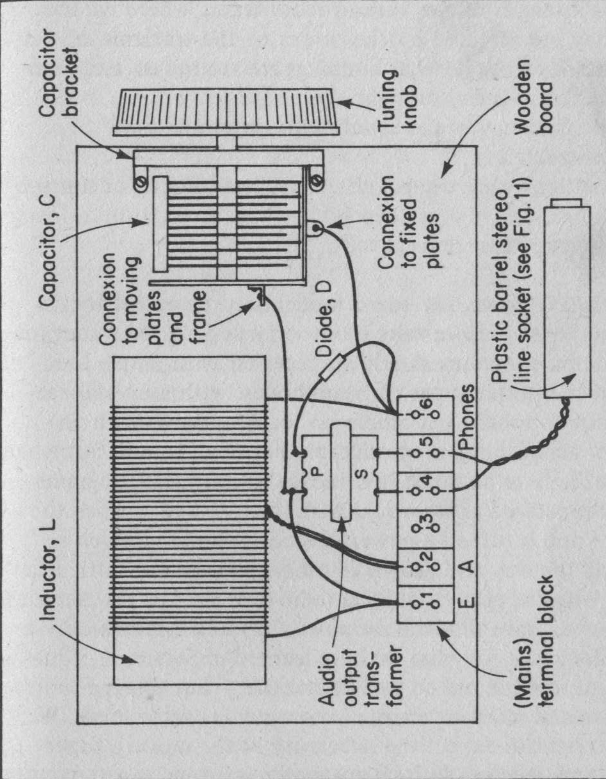
Capacitor bracket
Tuning knob
Wooden board
Capacitor, C
Connexion to fixed plates
Plastic barrel stereo line socket (see Fig.11)
Connexion to moving plates and frame
Diode, D
Phones
1
2
3
4
5
6
P
S
E
A
Inductor, L
Audio output trans-former
(Mains) terminal block

Fig. 14 Suggested component arrangement for elementary crystal set

A few hours work and our first set is completed. We cannot switch on because no on/off switch is required, all the power is free from the air so there is never a need to switch off.

4.3 LISTENING IN

When all is connected as in Fig.14 and with aerial and earth, medium wave radio stations should be heard and now we can assess what we have done. So far it is evident that individual circumstances will differ widely in

(i) the strength of the various radio waves where we live. They are affected by the power of the transmitter, the distance from it, what buildings are around us, and even whether it is day or night,
(ii) the efficiency of the aerial/earth system we have constructed,
(iii) whether we are using high quality headphones or simply an inexpensive crystal earpiece which apart from its lower efficiency uses one ear only.

Nevertheless given, say some moderately powerful local station(s) even a downstairs room aerial and a good waterpipe earth, radio programs should be received with ample loudness on the better types of headphones, with a crystal earpiece not so loudly yet ample on local stations. With an outdoor aerial things are noticeably better all round. But what is most likely to be much in evidence is that radio programs are not separated as we would wish them to be for when the tuning knob is turned a powerful local station may even be heard all the way and drown all others. One of our difficulties is that with the growth of local radio over the last few decades, many of us have moderately powerful radio stations only a few miles away. So what we have learned in Section 1.5 does not seem to work out so well in practice – but this is a simple experimental receiver, many improvements can be made. We have no right to expect the selectivity of the modern superheterodyne receiver with its hosts of transistors, tuned circuits

and special techniques. Nevertheless, by starting this way we have gained experience and confidence and will have little trouble in understanding each of the modifications which follow especially if we take them one at a time.

CHAPTER 5
MORE ADVANCED RECEIVERS

Such an uncomplicated receiver as we have constructed cannot be expected to perform as well as a more intricate one so we next look at some refinements, especially those from which can be discovered more about the basic principles of radio.

Readers with no local radio stations may find signals rather weak yet easy to separate while on the other hand those with one or two local stations will probably have more than adequate volume but little choice of station, the local ones swamping everything else and perhaps each other.

Selectivity is the ability of a receiver to respond to a particular radio frequency without interference from others and this may certainly seem to be lacking. However Fig.3(ii) in Section 1.5 considers only the ideal case, that of *pure* inductance and capacitance, that is, no electrons getting lost from the LC circuit in the process. When this does happen then the sharpness of tuning diminishes and the circuit is said to be *damped* – rather like reducing the brilliance of tone of a piano wire by holding a felt pad on it.

The damping arises from other circuits linked with the tuned circuit, in this case the aerial and earth, detector (diode) and headphones. The aerial/earth system hardly looks like a capacitor as in Fig.8, but in fact it is, the aerial wire is one plate, the earth the other. Thus the purity of the inductor is spoilt by having a tiny capacitor (C_a) connected across is as in Fig.15(i). The detector and headphones are connected directly across the tuned circuit and create even more damping, in a simple way we might consider that some electrons flow out of the tuned circuit instead of helping to build up the peak of activity which is required. Effectively this is rather like connecting a resistance (R) as also shown in Fig.15(i).

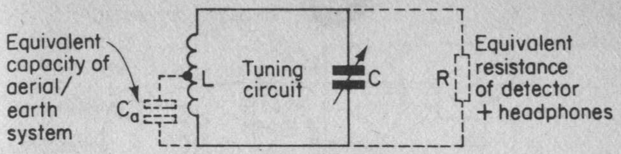
Equivalent capacity of aerial/ earth system
C_a
L
Tuning circuit
C
R
Equivalent resistance of detector + headphones

(i) Damping of LC circuit

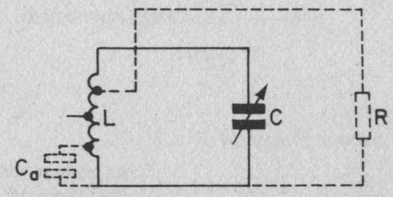

(ii) Damping reduced

Fig. 15 Reduction of damping to improve selectivity

One way of reducing the damping with consequent improvement in selectivity is by tapping the aerial and/or diode lower down the inductor (i.e. towards the earthy end) so that they are effective over less turns but therefore unfortunately providing a weaker headphone signal. Thus the sharper the tuning, the lower the loudness and obviously we will have to find the best compromise between the two.

5.1 IMPROVING SELECTIVITY

Fig.16(i) shows the modified circuit for finding the best compromise between selectivity and loudness for each of our own particular circumstances. To avoid the possibility of mistakes when wiring, a reminder has been added about how wires are shown connected or not connected together on a diagram. The inductor now has 7 taps instead of one but still has a total of 80 turns. There is a tap at every 5 turns from the earthy end

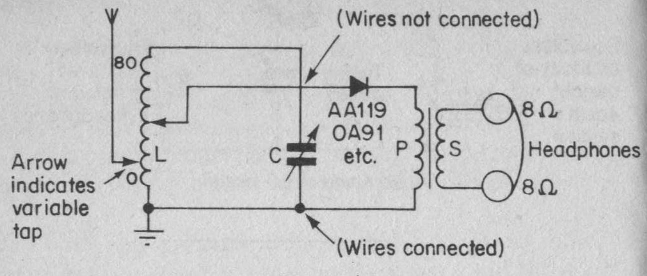
(Wires not connected)
80
L
O
Arrow indicates variable tap
C
AA119
OA91
etc.
P
S
8 Ω
Headphones
8 Ω
(Wires connected)

(i) Circuit

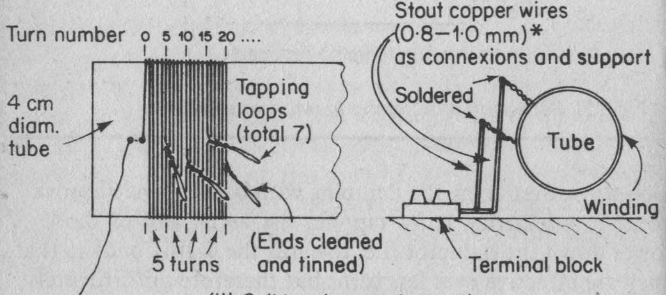
Turn number 0 5 10 15 20 ....
4 cm diam. tube
Tapping loops (total 7)
5 turns
(Ends cleaned and tinned)
Stout copper wires (0·8–1·0 mm)* as connexions and support
Soldered
Tube
Winding
Terminal block

(ii) Coil tappings and mounting

Fig. 16 Circuit with tapped inductor

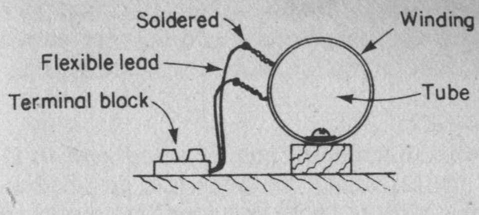
Soldered
Winding
Flexible lead
Tube
Terminal block

(iii) Alternative mounting

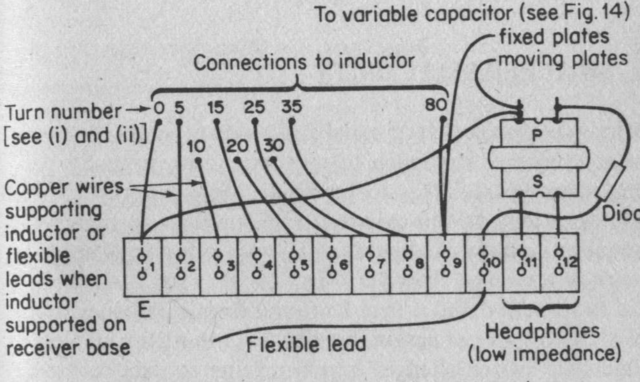
To variable capacitor (see Fig. 14)
fixed plates
moving plates
Connections to inductor
Turn number → 0 5 15 25 35 80
[see (i) and (ii)]
10 20 30
Copper wires supporting inductor or flexible leads when inductor supported on receiver base
P
S
Diode
1 2 3 4 5 6 7 8 9 10 11 12
E
Flexible lead
Headphones (low impedance)

(iv) Suggested component arrangement

* 21–19 SWG, 20–18 AWG

for the first 35 turns, i.e. 7 taps in all. We can either use the earlier method of a long (5 cm) loop twisted where it leaves the winding or perhaps a shorter loop onto which another wire is soldered. One suggestion is to have loops about 1.5cm long, staggered along the inductor with the latter supported on a 12-way terminal block using stout tinned copper wire supports as shown in Fig.16(ii). Do not be too concerned about support wire gauge, many junk boxes can produce something suitable, even the individual strands of scrap pieces of household mains cable. Equally the inductor might be fixed on two small blocks of wood with screws or adhesive as shown in Fig.16(iii) with flexible leads to the terminal block or to a tag strip instead.

The inductor requires 9 terminals of the block, the remaining 3 conveniently accommodate the headphone transformer (if

required) and a *flying lead* for the diode, these components are shown in Fig.16(iv). This is doing the job properly, there is no reason however (except for unreliability of contact) why both aerial and diode flying leads should not terminate on *crocodile clips* and be clipped onto the chosen taps as required.

Earth is connected to terminal 1 and the headphones to 11 and 12. Both the aerial and the diode lead are tried at terminals 2 – 9, a game of "poke and hope" but one which can be rewarding. Remember that generally but not always, the tappings with the lower numbers reduce loudness but increase selectivity. Aerial to terminal 8, diode to 9 brings us back to Fig.13(i) of course.

5.2 SWITCHED SELECTIVITY

A technique of selection of aerial and diode tappings by means of rotary switches has much to recommend it especially if setting changes are likely to be needed for reception of different stations. In this case two single-pole 8-way switches are required, usually obtainable as 12-way so leaving 4 spare. The drawing symbol shown in Fig.17 speaks for itself with regard to its action and it may be found from catalogues that there are two types of *action*, by which is meant the manner in which the switch changes over from one contact to the adjoining one. *Break before make* switches disconnect one circuit before the next is connected and *make before break* have a change-over period during which both circuits are connected at once for a short time. Either type is suitable for us. A knob to rotate each switch is also required, a small one, say 2 cm diameter or less is ample, if with an indicator line or pointer, so much the better. The modifications to Fig.16 for switching are given in Fig.17.

With such rotary switches a search for the optimum arrangement is quickly made. To tidy up such a receiver a front panel of plywood might be made and drilled to accommodate the

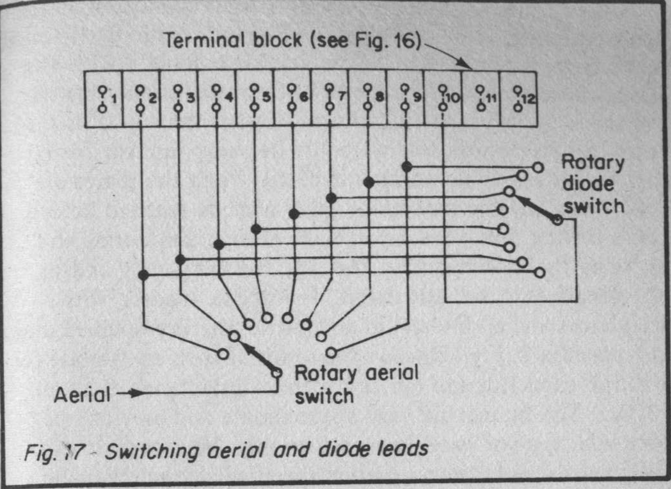

Fig. 17 - Switching aerial and diode leads

variable capacitor, the two switches and the headphone socket, the latter must then be of the chassis type. A pair of terminals for aerial and earth completes the receiver. On the panel the two rotary switches are easily labelled 1 – 8 but we are still uncertain about a tuning dial because we may not have the right variable capacitor for the inductor, this is the subject of the next section.

5.3 GETTING THE RANGE RIGHT

This is where we pause in receiver construction to remove some of the confusion which may still exist with regard to the tuning ingredients, wavelength, frequency, inductance and capacity, each of which has some bearing on the others, to fit together these four pieces of the puzzle, so to speak. For most crystal sets the inductance is fixed according to the variable capacitor available and we tune the receiver over the range by rotating the variable capacitor knob over half a circle. This is normally divided up into 180 *degrees* (180°) as shown on any *protractor* and as a reminder as the inner row of figures in Fig.18. The problem is to ascertain where on the dial a particular radio station should be found. Firstly we must relate degrees of rotation of the capacitor spindle with the actual capacitance and we might be forgiven for guessing that for, say, a 365pF capacitor the capacitance value at 0° is zero at at 180° is 365pF, but we could be sadly in error for (i) there is always some *residual* capacity when the plates are fully open, (ii) the maximum value may be reached before 180°. Having found the capacitance at each dial setting and knowing the inductance of the coil, the frequency and/or wavelength can be calculated. For those readers with calculators and an interest in arithmetic, this is explained more in Appendix 2. Fig.18 shows the result of such an exercise for a 365pF capacitor and our coil with its inductance of about 230μH. The figures are very approximate and are for one particular type of variable capacitor only. Most capacitors are designed to conform to a certain specified characteristic but there are inevitable differences in manufacture, furthermore there is no guarantee that our coils have exactly the inductance required. Accordingly it might pay to put on a few more turns than stated and then if wanted stations seem to be missing at the high-frequency end of the range, take some off to lower the inductance and hence increase the resonant frequency.

There is also the problem of having a capacitor of some other value than that used for Fig.18 or perhaps one has been taken from a derelict receiver with no idea whatsoever of the value. Measurement of capacitance is certainly beyond us at this stage, hence the technique above of adjusting the number of turns on the coil is the only one.

As an example of coil adjustment when the capacitance is known, say, 500pF, reduce the coil turns on Fig.13 to 65 – 70 to give an inductance of about 170μH. But if we are happy with a few local stations and none is too close to an end of the band, there is no problem.

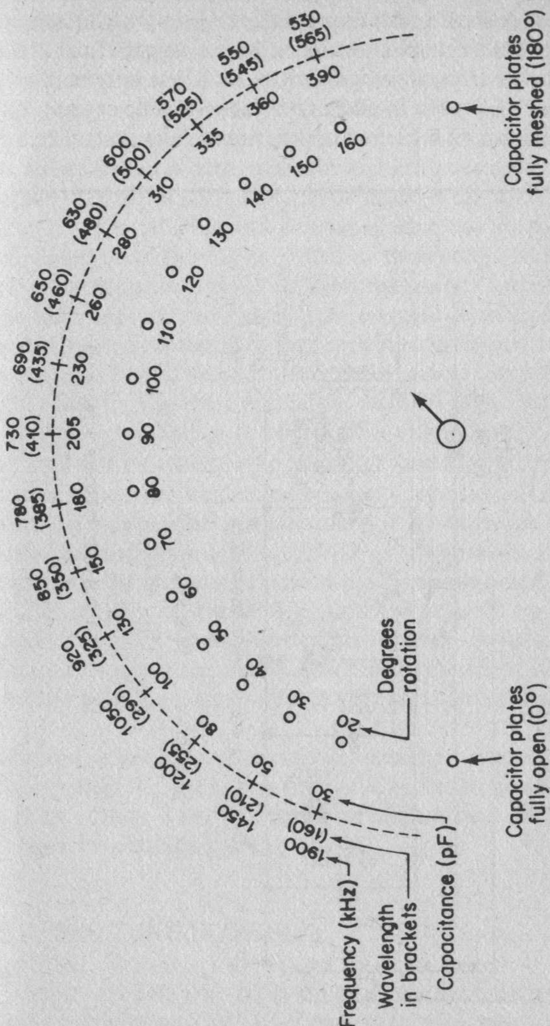
Frequency (kHz)
Wavelength in brackets
Capacitance (pF)
Degrees rotation
Capacitor plates fully open (0°)
Capacitor plates fully meshed (180°)
1900 (160) 30
1450 (210) 50
1200 (255) 80
1050 (290) 100
920 (325) 130
850 (350) 150
780 (385) 180
730 (410) 205
690 (435) 230
650 (460) 260
630 (480) 280
600 (500) 310
570 (525) 335
550 (545) 360
530 (565) 390
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160

Fig. 18 Approximate capacitance and frequency values for 365 pF variable capacitor and 230 μH coil

5.4 SWITCHED TUNING

This is a feature, simple to provide yet very useful in operation. A switch selects any one of a group of capacitors, each of the correct value for reception of its particular station. Fig.19 shows an arrangement for two stations only applied to the circuit of Fig.16. In addition a fully variable capacitor and tuning dial can be switched in if required. The system is not expensive and in fact for 2 or 3 fixed stations only, probably costs less than for a single variable. Suppose we are designing for two stations say on 909 and 1150 kHz. We could use a miniature single-pole change-over toggle switch or just two positions of a single-pole 8 or 12-way rotary switch for S. Both cost about the same so the latter might be preferable because it can accommodate more fixed stations if needed later. For the coil of Fig.16 the outer ring of Fig.18 is available for estimating the capacitance required in both cases, for example for 909kHz we might guess 135pF and for 1150kHz, 85pF (note that capacitance and frequency increase in opposite directions round the scale). We have considerable freedom of choice in setting up these two capacitances but a suggestion is to use 22pF *trimmers* with fixed ceramic capacitors as shown for $C_1 - C_4$. Trimmers are tiny variable capacitors, adjusted by a screw. When two capacitors are connected in *parallel* as C_1/C_2 and C_3/C_4 are, the total capacitance is the sum of the individual ones, thus a fixed 120pF in parallel with a 22pF trimmer can be varied over the range just over 120pF (all variable capacitors have some residual capacity) to 120 + 22 = 142pF. Hence 909 kHz can be tuned in by adjusting C_2 to make the combination about 135pF (more accurately, 133pF), this is done by adjusting C_2 until the station is heard at its best. Thereafter whenever S is switched to position 1, the station on 909 kHz should be heard. Similarly for as many other positions on the switch as we wish. Switched tuning and switched selectivity can be used together if required.

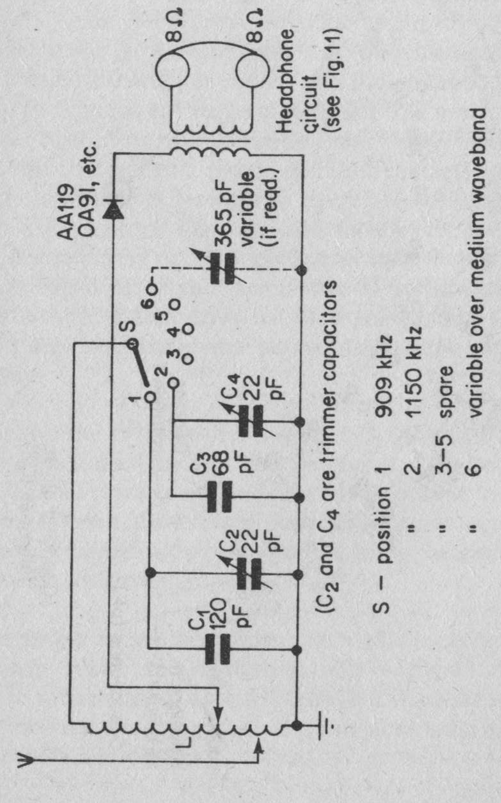
AA119
OA91, etc.
8Ω
8Ω
Headphone circuit (see Fig.11)
365 pF variable (if reqd.)
S
1
2
3
4
5
6
L
C1 120 pF
C2 22 pF
C3 68 pF
C4 22 pF
(C2 and C4 are trimmer capacitors)
S – position 1 909 kHz
" 2 1150 kHz
" 3–5 spare
" 6 variable over medium waveband

Fig. 19 Fig. 16(i) adapted for switched tuning

Capacitors may have a *maximum working voltage* quoted, above which the component is liable to break down. Any working voltage is suitable for these circuits because all signal voltages are extremely low.

5.5 TWO TUNED CIRCUITS

For many readers who have constructed the basic receiver, good selectivity may still be elusive even with variable tappings on the inductor as in Sections 5.1 and 5.2. Denied as we are in this book of overcoming the problem by the expedient of using transistors and batteries we must continue by exploring other ways of obtaining it. If a single tuned circuit is insufficient, why not use two, one after the other, rather like filtering wine once then a second time to further enhance its clarity? Technically this is a sound idea and the sharpness of tuning increases but one circuit must not damp the other too much for this is the very thing which itself impairs selectivity. Fig.20(i) shows such a circuit. Both inductors are the same as that in Fig.16 but because they are in magnetic contact they

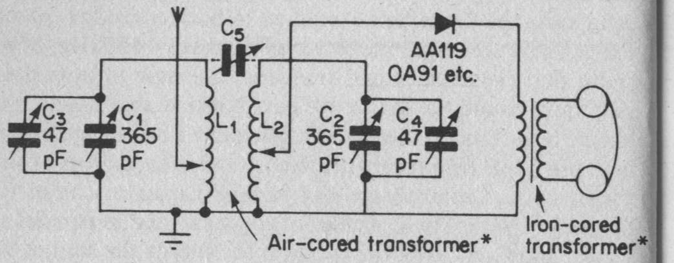

C5
AA119
OA91 etc.
C3
47
pF
C1
365
pF
L1
L2
C2
365
pF
C4
47
pF
Air-cored transformer*
Iron-cored
transformer*

(i) Circuit

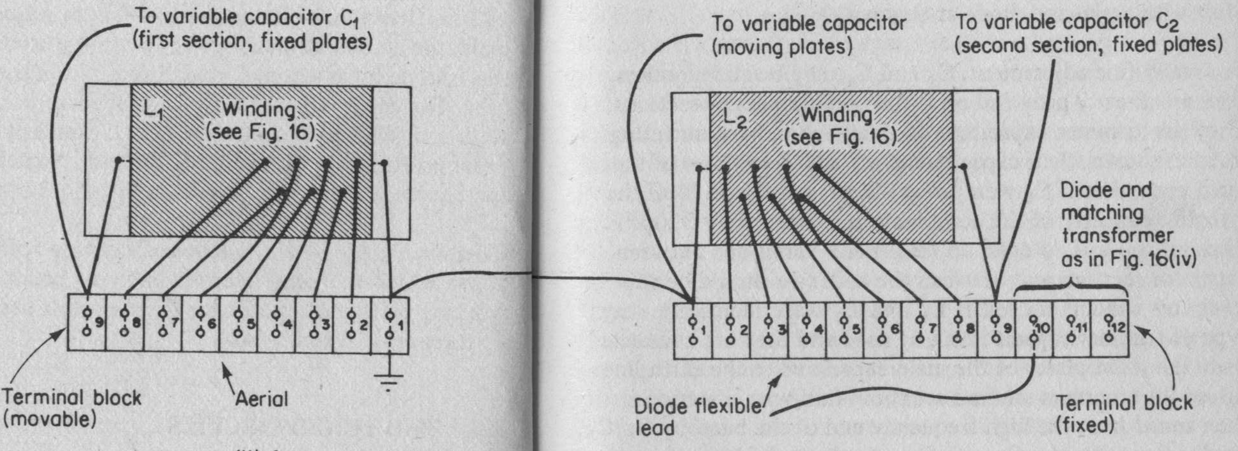

To variable capacitor C1
(first section, fixed plates)
L1
Winding
(see Fig. 16)
9
8
7
6
5
4
3
2
1
Terminal block
(movable)
Aerial
To variable capacitor
(moving plates)
To variable capacitor C2
(second section, fixed plates)
L2
Winding
(see Fig. 16)
Diode and
matching
transformer
as in Fig. 16(iv)
1
2
3
4
5
6
7
8
9
10
11
12
Diode flexible
lead
Terminal block
(fixed)

(ii) Arrangement of inductors

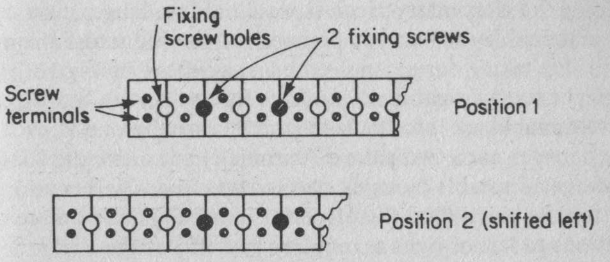

Fixing
screw holes
2 fixing screws
Screw
terminals
Position 1
Position 2 (shifted left)

(iii) Moving inductor L1

Fig. 21 A double-tuned crystal receiver

* See Section 2.4

form an r.f. transformer as mentioned in Section 2.4. The closer they are to each other, the more energy passes from the aerial to the detector but unfortunately each reduces the sharpness of tuning of the other. The practical circuit therefore requires that the separation between the two inductors is made variable to obtain the best compromise in any particular case. A suggestion is given in Fig.20(ii). Having some idea by now of the principles involved and the experience of building the elementary receiver we should no longer need a full practical layout except perhaps for the inductors themselves. By taking for example the arrangement of Fig.16(ii) we might fix the right-hand inductor by two screws through the terminal block into the base (there is usually one screw hole between each two pairs of terminals) but make the left-hand one adjustable by using two screws close together and shifting the block (and therefore the inductor) one or more positions to left or right as required [see (iii) of the Figure]. Adjustment should range from the tubes just touching end-on to some 4 – 5 cm apart. This adjustment is in addition to the aerial and diode tappings so we have plenty to play with. Start with aerial and diode at terminal 9.

For really fine adjustment, C_3 and C_4 may be added unless they are already provided on the variable capacitor sections. They are trimmer capacitors of small value, by connecting them as shown, their capacitances are added to those of the main ones, thus if C_1 were set at 300pF and C_3 to 20pF the effective capacity of the combination is 300 + 20 = 320pF. Their purpose is to even up (trim) any variations between capacitor sections and between the coils (we should be top-level coil winders indeed if L_1 and L_2 were identical). Any type of trimmer capacitor can be used and they are connected from the fixed plates of the main capacitors to the earth line. In use the trimmers are first set about half-way, a station is then tuned in at the high frequency end of the band by C_1/C_2, finally the trimmers C_3 and C_4 are adjusted for maximum loudness. This technique ensures that both tuned circuits resonate at the same frequency. Just in case when under weak signal conditions even more coupling between the inductors

is required than is obtained when they are at their closest, another trimmer capacitor C_5 of maximum value 4.7 or 5.5pF may be added and adjusted as required. Some of the signal then flows directly via C_5 in addition to that through the transformer $L_1 L_2$.

A circuit with double tuning which has already been published twice in this series is still especially worthy of our interest because again we learn more about the principles of radio. From Section 1.8, in looking at the basic idea of detection one could be forgiven for wondering whether blocking off and discarding every alternate half-cycle of the r.f. carrier wave was not somehow inefficient and certainly a power engineer would know it to be. What we have used so far is appropriately known as *half-wave* rectification, there is a more efficient system known as *full-wave*. Let us look at this graphically as in Fig.21(i), from which it is immediately evident that for a given input the full-wave circuit must have a greater output.

The crystal set circuit using full-wave detection is shown in Fig.21(ii), it is basically an adaptation of a standard full-wave rectification circuit used to produce direct current from the alternating current mains. L_1 and L_2 are so wound that when a signal flows in L_3 equal voltages are induced by transformer action in them in such directions that only one of the diodes passes current at a time. When the signal reverses direction the other diode is operative hence full-wave rectification as in (i) is obtained. Winding the coil needs a little more skill than hitherto but the result is a professional looking job as there are no untidy taps, the aerial is controlled by capacitors instead. Details of the coil are at (iii). A 2.5cm (1 inch) diameter tube is required, about 15 cm long. An odd piece of plastic water piping of this outside diameter might be found but failing this the tube can be made from a piece of thin cardboard 15 cm x 8 cm wrapped over a broom handle so that the long edges just meet, and lapped with gummed paper tape. Some 11 gm (less than ½oz) of 0.25mm (SWG 33, AWG 30) enamelled copper wire is required for the three windings and it is essential that L_1 and L_2 are wound in the same direction. The

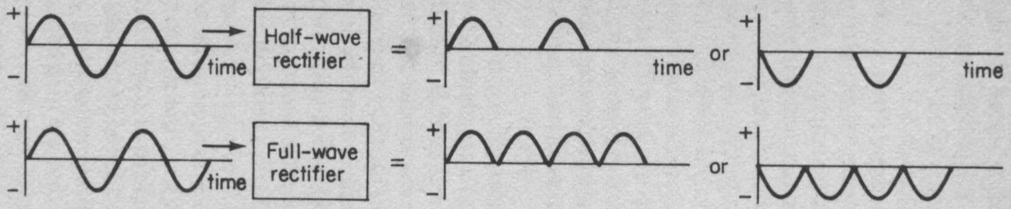

+
−
time
Half-wave rectifier
=
time
or
time
Full-wave rectifier

(i) (i) Rectifier outputs

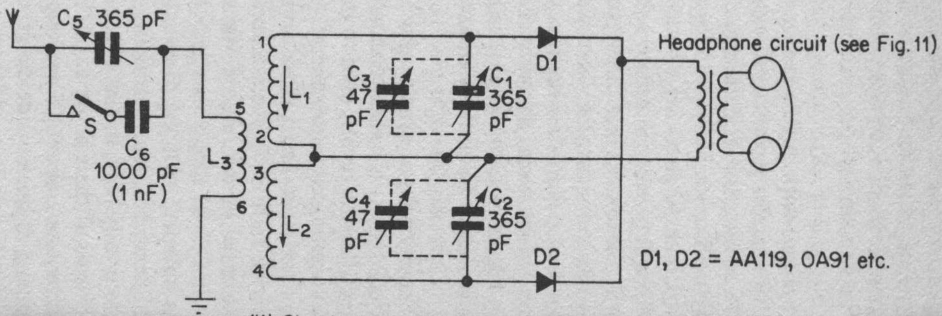

C₅ 365 pF
S
C₆
1000 pF
(1 nF)
L₃
1
2
3
4
5
6
L₁
L₂
C₃
47
pF
C₁
365
pF
C₄
47
pF
C₂
365
pF
D1
D2
Headphone circuit (see Fig. 11)
D1, D2 = AA119, OA91 etc.

(ii) Circuit of complete receiver

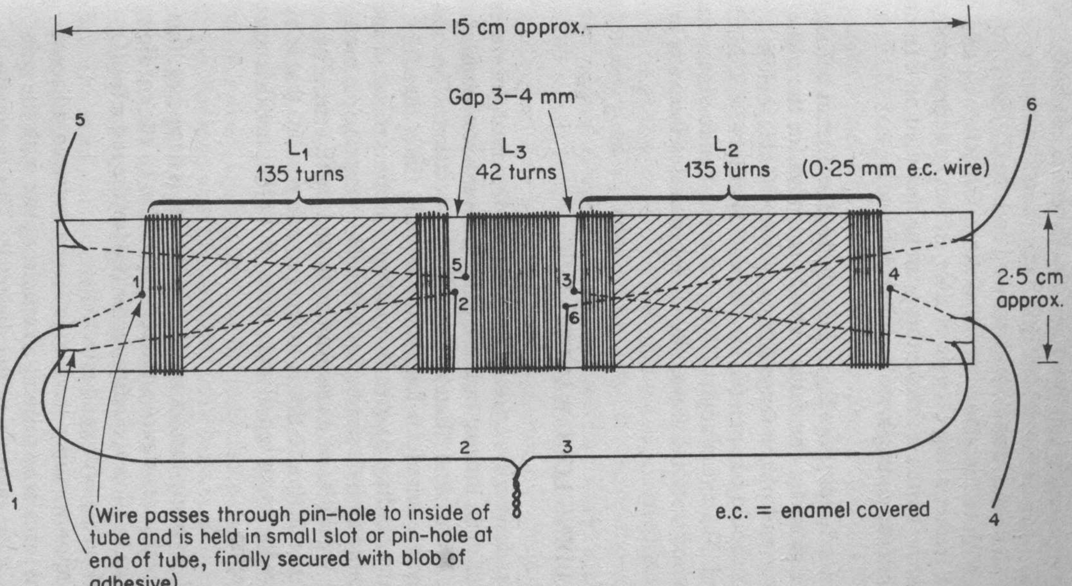

15 cm approx.
Gap 3–4 mm
L₁
135 turns
L₃
42 turns
L₂
135 turns
(0·25 mm e.c. wire)
2·5 cm approx.
1
2
3
4
5
6
(Wire passes through pin-hole to inside of tube and is held in small slot or pin-hole at end of tube, finally secured with blob of adhesive)
e.c. = enamel covered

(iii) Details of R.F. transformer windings

Fig. 21 Full-wave detection

complete inductor is not self-supporting, it must therefore be screwed or glued to small wooden blocks on the baseboard. The diodes can be connected either way round but they must both be the same, that is, both marked ends to C_1 and C_2 or equally both to the matching transformer. Capacitor C_5 and switch S (either rotary, toggle or slide, in fact any *single pole, single throw* switch is satisfactory) can be mounted on a front panel with the tuning knob and dial (if fitted).

The trimmer capacitors C_3 and C_4 if added, are set as before. The rule given for the aerial capacitors is that for the high-frequency end of the range (above about 850 kHz, below 350m) S should be open (C_6 disconnected) otherwise S should be closed. Thereafter adjust C_1/C_2 for maximum loudness on the required station followed by C_5 for increased loudness or better selectivity.

5.6 ADDING LONG WAVES

Most broadcast radio transmitters are found on medium waves, hence our main interest in this band. However a few transmit on long waves and therefore need consideration. The inductance required for long waves is many times that for medium, accordingly a greater number of turns is required and if we wound an inductor using the single-layer principle used so far it would be of excessive length and inefficient. The immediate reaction to this is to *layer* or *pile* wind it, that is by successive layers on top of each other. But other inefficiencies now creep in.

(i) with so many turns a smaller gauge wire is necessary, this has higher resistance and such opposition to current flow results in a lower *quality factor* (Q) for the inductor (Q must be high for good selectivity).

(ii) inductor windings have capacitance between adjacent turns, each wire forming a capacitor plate with the wire insulation plus air as the dielectric. This is greatly increased with pile winding because wires not only have

adjacent ones in the same layer but also in the next inner and outer layers. Special winding techniques using machines can minimize this effect but we cannot manage by hand. The net effect of the unwanted capacitance is similar to increasing the winding resistance, hence further lowering the Q.

We have to accept (i) or we end up with an enormous inductor therefore a finer wire is used, for example 0.315mm (SWG 30, AWG 28). For (ii) we can split the winding up to reduce the capacity and reduce it even further by changing to a different type of wire insulation with an extra covering or two of cotton, silk or nylon to increase the wire spacing. Unfortunately such covered wires are not so easily obtainable from the usual component firms and also are rather more expensive. It might be at this stage therefore that we should consider purchasing a ready-made coil. However, for those who wish to persist against all odds, a long wave/medium wave crystal receiver circuit which was published earlier is reproduced here [Fig.22(i)], it uses a home wound inductor for which DSC (double silk covered) or DCC (double cotton covered) wire can probably be obtained from a specialist wire firm. Details of the inductor are in (ii) of the Figure, some 50 – 60 gm (2 oz) of 0.315mm (SWG 30, AWG 28) wire are required and it is again wound on a 4cm diameter, 11cm long cardboard tube (e.g. toilet roll centre). Switch S is for wave band changing, when closed L_1 only is effective and this is the correct inductance with medium waves for a 500pF variable capacitor (using a lower value for C_1 requires an even larger inductor). When S is open, L_2 is added and the inductance rises to that required for long waves (about 2000μH). Note how L_2 is wound in separate piles to reduce the capacity, each pile winding can be built up without side supports but if preferred, cardboard discs can be fitted first [see (iii)] to make the job easier. The aerial and diode tappings are set for a compromise between loudness and selectivity but being a circuit without frills it is unlikely to reject fully local transmitters.

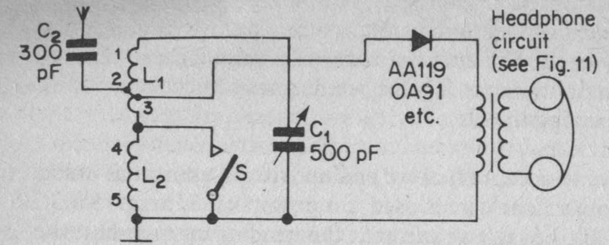
C₂
300 pF
1
2
L₁
3
4
L₂
5
S
C₁
500 pF
AA119
OA91
etc.
Headphone
circuit
(see Fig. 11)

(i) Circuit

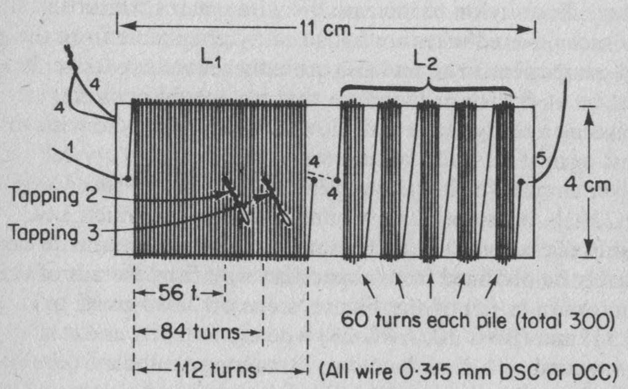
11 cm
L₁
L₂
4
4
1
Tapping 2
Tapping 3
4
4
5
4 cm
56 t.
84 turns
112 turns
60 turns each pile (total 300)
(All wire 0·315 mm DSC or DCC)

(ii) Details of coil

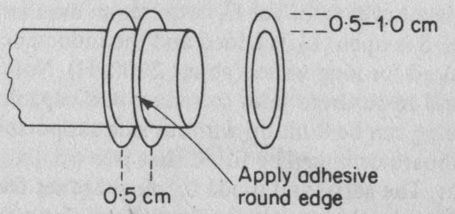
0·5–1·0 cm
0·5 cm
Apply adhesive
round edge

(iii) Adding discs to coil former

Fig. 22 A LW/MW receiver

5.7 USING READY-MADE COILS

In the previous section we begin to get suspicious about the success of winding our own coils when selectivity is of major concern and when this is the case it may well be preferable to purchase ready-made ones instead. These are likely to be more efficient than the home-made types because generally they use metal instead of air cores. By filling the centre of the tube or *former* with certain metallic compounds the inductance is raised, thus fewer turns are required so resulting in less winding resistance and a higher Q which as we have discovered, is the secret of selectivity. Such coils may be available especially for crystal sets or are the aerial coils of a range of different types for home construction of complete transistor receivers. Diagrams for use are usually included with the coil. Be sure to order for the correct wave range (long or medium) because many series of coils include those for short-wave ranges also.

Having read so far and perhaps also having built one or more of the designs, our understanding and constructional abilities should certainly be at the level at which a circuit diagram included with a manufactured coil creates no misgivings.

APPENDIX I
WAVELENGTH – FREQUENCY CONVERSION

LONG WAVEBAND

λ metres	f kHz
1000	300
1050	286
1100	273
1150	261
1200	250
1250	240
1300	231
1350	222
1400	214
1450	207
1500	200
1550	194
1600	188
1650	182
1700	176
1750	171
1800	167
1850	162
1900	158

MEDIUM WAVEBAND

λ metres	f kHz
190	1579
200	1500
210	1429
220	1364
230	1304
240	1250
250	1200
260	1154
270	1111
280	1071
290	1034
300	1000
310	968
320	938
330	909
340	882
350	857
360	833
370	811
380	789
390	769
400	750
410	732
420	714
430	698
440	682
450	667
460	652
470	638
480	625
490	612
500	600
510	588
520	577
530	566
540	556
550	545

APPENDIX 2
SIMPLIFIED RESONANT CIRCUIT CALCULATIONS

There is a standard, well known formula relating the resonant frequency (f_o) of a tuned circuit with the values of inductance and capacitance therein, it is

$$f_o = \frac{1}{2\pi\sqrt{LC}}$$

where f_o is the resonant frequency in Hz
L is the inductance in henrys
C is the capacitance in farads
π is the ratio of circumference to diameter of a circle and is equal to 3.14159.......

For convenience throughout this book we consider frequency only in kilohertz (f_k), inductance in microhenrys (L_μ) and capacitance in picofarads (C_p), therefore

$$f_o = f_k \times 10^3, \quad L = L_\mu \times 10^{-6}, \quad C = C_p \times 10^{-12}$$

and substituting in the above formula

$$f_k \times 10^3 = \frac{1}{2\pi\sqrt{L\mu \times 10^{-6} \times C_p \times 10^{-12}}}$$

$$= \frac{1}{2\pi \times 10^{-9}\sqrt{L_\mu \times C_p}}$$

$$\therefore f_k = \frac{10^{-3} \times 10^9}{2\pi\sqrt{L\mu C_p}} = \frac{10^6}{2\pi\sqrt{L\mu C_p}}$$

$$\therefore f_k^2 = \frac{10^{12}}{4\pi^2 L_\mu C_p} = \frac{2533 \times 10^7}{L_\mu C_p}$$

giving us three useful formulae:

$$L_\mu = \frac{2533 \times 10^7}{f_k^2 C_p},$$

$$C_p = \frac{2533 \times 10^7}{f_k^2 L_\mu},$$

$$f_k = \sqrt{\frac{2533 \times 10^7}{L_\mu C_p}}.$$

With scientific calculators there is little difficulty because scientific notation is available but with a home or pocket calculator which has no such facility and displays only 8 digits, it is still straightforward as long as it has a $\sqrt{x}$ facility.

We cannot get 2533 followed by 7 0's into the display hence work with 4 and multiply by 1000 later. As an example, when $L_\mu = 230$ and $C_p = 365$

$$f_k = \sqrt{\frac{25330000}{230 \times 365} \times 1000} = \sqrt{301.72722 \times 1000}$$

$$= \sqrt{301727.22} = 549.3 \text{ (kHz)}.$$

Similarly for an inductor of 2600μH, the capacitance required to tune to 210 kHz is

$$C_p = \frac{25330000}{210 \times 210 \times 2600} \times 1000 = 220.9 \text{ (pF)}.$$

APPENDIX 3
SOLDERING

To be accomplished at soldering is an asset indeed for so many jobs around the home can be undertaken once the techniques and principles are understood, even plumbing. But that is another story, we are concerned with electrical joints only and happily success with these is really quite easy to achieve.

A3.1 SOLDER

As we see in Section 1.4, electrons flow easily within a conductor it is therefore imperative that should the latter be connected or jointed to another conductor, tag or terminal, the joint itself should not resist this flow. This is accomplished by the two metals being in such intimate contact that electrons flow across the joint as though the metal were continuous. A screw connection such as used in the terminal block of Fig.14 achieves this by the screw being forced into the metal to give a very tight and clean contact. Soldering is another method of joining two conductors together and it provides an excellent and secure electrical contact. Electrical solder is an alloy of tin and lead usually in a 60/40 ratio, this melts and flows at 190°C. Water boils at 100°C so as far as we personally are concerned, molten solder and soldering irons must be treated with respect and fingers kept off. Tin is a metal which readily alloys with copper, that is, it actually mixes with the copper at its surface, thus solder joining two pieces of copper acts as a cement and because of its intimate contact with the surface, electrons flow across without hindrance. Electrical solder is called *soft*, naturally it has some of the characteristic softness of lead, in fact wires can be peeled out of a soldered joint. *Hard* solder on the other hand has a higher melting point and results in a much stronger joint but it is more difficult to manipulate.

It is easy to realize that for solder to form an alloy with copper the surface of the latter must be immaculate, that is with absolutely no dirt, corrosion or *oxide* in the way. Copper readily forms its oxide when it is heated. The oxide is not copper although it is derived from it and therefore any present must be removed. Once copper is coated with solder, formation of oxide is prevented, the process is aptly called *tinning*.

Solder flowing over clean metal and forming an alloy correctly is like water flowing, it has the resemblance of *wetness*, conversely if it tries to flow over a contaminated surface with which it cannot alloy it does not give the appearance of wetting the metal. To assist in prevention of oxide and to aid in cleanliness a *flux* (= to make flow) is employed and to make our life easier modern solders are available looking rather like a thick, soft grey wire with a resin flux running down the centre. For us a solder about 0.7mm diameter (SWG 22, AWG 21) is most suitable. It is available in small coils or larger reels.

A3.2 SOLDERING TOOLS

Undoubtedly the most useful and most used type of soldering "iron" (the working part is actually copper) is the electric one, a heating element within the barrel of the iron heats one end of a copper *bit*, heat travels along the bit to the working end to set up a temperature above the melting point of solder so that solder in contact with the bit melts. A 20 or 25 watt iron has a bit small enough for fine electrical work but it cannot handle bigger jobs which extract heat from the bit faster than the element generates it. A non-electric iron has a copper bit on an iron rod with a handle at the end remote from the bit, the bit is heated separately, for example in a gas flame, when the bit has cooled it must be reheated.

The two almost essential tools for wiring are (i) a pair of long-nosed pliers, with a little practice fine wires can be manipulated quite easily, and (ii) a pair of wire or side-cutters. These two items may seem expensive for in fact they may cost more than the whole crystal set (excluding headphones) but they have many uses around the home so if the family toolbox does not already possess them, the cost will soon be repaid. However it is just possible to manage without them, whether to purchase depends so much on what future intentions are with regard to electronics as a hobby. A small file will also be occasionally required for re-shaping the iron bit.

A3.3 SOLDERING

For a soldered joint, not only the parts to be joined but also the tip of the soldering iron must be tinned. Assuming an electric iron, plug in and after a few minutes it will have reached working temperature. Do not rest the bit on anything metal otherwise heat is conducted away. Most small irons have a hook attached so that they can be hung in an upright position with the bit uppermost. Wipe across the tip of the iron with a small piece of clean rag, if the tip surface is silvery bright it is suitably tinned, failing this, unplug and allow to cool then file until clean copper is exposed all over the tip, then apply the end of a piece of cored solder. When the solder has received sufficient heat from the iron it will melt and flow over the tip, forming a bead of molten solder. If this is wiped away the tip will appear bright and tinned. Remember that little beads of molten solder can raise larger blisters! Filing and re-tinning the tip is only needed occasionally when it has become pitted.

Most component wires and tags are tinned for us during manufacture but if they are dirty or corroded, re-tinning may be necessary. However for practice let us tin and then joint two pieces of copper wire (if no untinned wire is available then tinning can be removed with emery cloth). Wire insulation can first be removed where necessary by cutting round with a knife (carefully, so as not to cut the wire itself) and pulling off. The ends of the wire must be bright copper by rubbing

with sandpaper, emery cloth or by scraping. For solder to flow over the surface the wire temperature must be above that of the melting point of solder. Firstly add a little solder to the tip of the iron as shown in Fig.23(i), next push the end of the wire into the bead to heat the wire (ii), then feed solder (and flux) onto the wire and molten bead,it will flow over the copper and in fact the solder can be "painted" on (iii). Then move the iron and solder along the wire slowly continuing the painting process as far as is required (some surplus solder may fall off, this is inevitable when we start). Finally remove the wire and immediately wipe with the clean rag for a bright, silvery surface to appear, only the surplus solder wipes off, not the alloy.

Tin two pieces of wire and then twist the tinned ends together for about 1cm. Apply solder to the twist as we did in (iii) for the single wire and it will flow easily into the joint, cementing the two wires together, they can subsequently be untwisted only with considerable difficulty by tearing out of the solder.

Fig.23(iv) illustrates the soldering of a capacitor lead to a tag (for clarity we assume that the tag is already connected to something). First check that all surfaces are tinned then loop the capacitor wire through the tag. Press the tip of the iron against the tag and feed solder between the tag and the tip as shown. As the joint heats up the solder will flow, sufficient being fed in to ensure that the wire is fully cemented to the tag. Remove the iron. It looks as though we need three hands for iron, solder and capacitor so the latter must be made self-supporting, perhaps by squeezing the loop of wire onto the tag. Once soldered it certainly must not be moved until the solder has set, quite a short time, especially if we blow on the joint to cool it. Later on when a little skill has developed it will be found possible to solder wires without making a loop through the tag, this is especially useful for experimental work because it facilitates later removal of components.

Provided that all parts are brightly tinned and solder is seen to flow easily into all the crevices and not form globules which

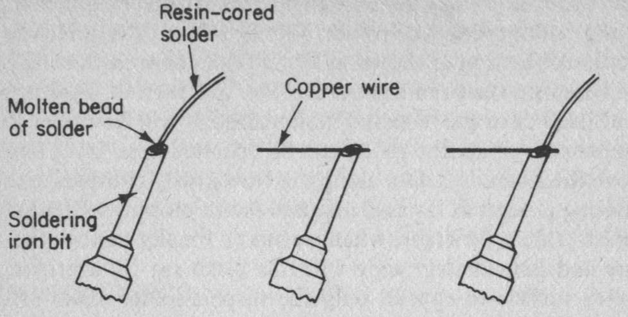
Resin-cored solder
Molten bead of solder
Copper wire
Soldering iron bit

(i) Preparing the iron (ii) Heating the wire (iii) Applying the solder

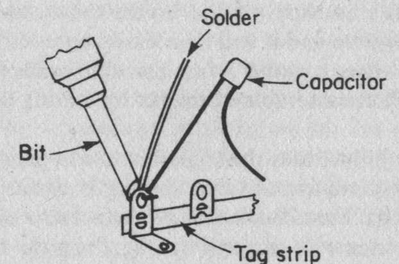
Solder
Capacitor
Bit
Tag strip

(iv) Soldering a component to a tag

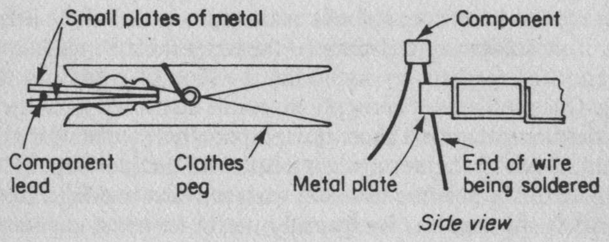
Small plates of metal
Component
Component lead
Clothes peg
Metal plate
End of wire being soldered
Side view

(v) A home-made heat shunt

Fig. 23 Soldering

refuse to flow, a perfect joint is obtained. *Dry* joints occur when dirt or oxide or perhaps too much flux prevent the solder from alloying – a little experience will take care of this.

Finally, remember that if used carelessly a soldering iron and molten solder can cause burns. Also even when unplugged the bit may still remain hot for sometime.

A3.4 HEAT SHUNTS

Some components, especially diodes and transistors are damaged by excessive heat arriving along their connecting wires. Thus soldering operations should not be prolonged nor should the leads be shortened. A further protection is provided by the use of a *heat shunt*, just a small chunk of metal clipped to the lead near the body of the component. Heat flowing along the lead is absorbed by the shunt and is therefore prevented from continuing its passage along the wire into the heart of the device. Heat shunts can be purchased quite cheaply or made for example by clipping strips of metal (copper, aluminium etc) onto the wire using a clothes peg [see Fig.23(v)].

A long appendix because it is easier to solder than to explain how to do it. Good luck!

PUBLISHER'S NOTE

Readers who feel they would like, or have a need to understand electronics more fully may be interested in a series of books by the same author entitled "Elements of Electronics". The series starts at the very beginning and assumes that the reader has no mathematical experience whatsoever and therefore teaches all that is required as each book progresses. It is an educational step-by-step series in which the reader can advance his or her knowledge book by book. The publication numbers are BP62, 63, 64, 77 and 89 and full details are contained within the current Catalogue of Babani Books.

Please note overleaf is a list of other titles that are available in our range of Radio, Electronics and Computer Books.

These should be available from all good Booksellers, Radio Component Dealers and Mail Order Companies.

However, should you experience difficulty in obtaining any title in your area, then please write directly to the publisher enclosing payment to cover the cost of the book plus adequate postage.

If you would like a complete catalogue of our entire range of Radio, Electronics and Computer Books then please send a Stamped Addressed Envelope to:

BERNARD BABANI (publishing) LTD
THE GRAMPIANS
SHEPHERDS BUSH ROAD
LONDON W6 7NF
ENGLAND

160	Coil Design and Construction Manual	1.50p
202	Handbook of Integrated Circuits (IC's) Equivalents & Substitutes	1.75p
205	First Book of Hi-Fi Loudspeaker Enclosures	95p
221	28 Tested Transistor Projects	1.25p
222	Solid State Short Wave Receivers for Beginners	1.25p
223	50 Projects Using IC CA3130	1.25p
224	50 CMOS IC Projects	1.25p
225	A Practical Introduction to Digital IC's	1.25p
226	How to Build Advanced Short Wave Receivers	1.20p
227	Beginners Guide to Building Electronic Projects	1.50p
228	Essential Theory for the Electronics Hobbyist	1.25p
RCC	Resistor Colour Code Disc	20p
BP1	First Book of Transistor Equivalents and Substitutes	95p
BP2	Handbook of Radio, TV & Ind. & Transmitting Tube & Valve Equiv.	60p
BP6	Engineers and Machinists Reference Tables	75p
BP7	Radio and Electronic Colour Codes and Data Chart	40p
BP14	Second Book of Transistor Equivalents and Substitutes	1.75p
BP24	52 Projects Using IC741	1.25p
BP27	Chart of Radio Electronic Semiconductor and Logic Symbols	50p
BP32	How to Build Your Own Metal and Treasure Locators	1.75p
BP33	Electronic Calculator Users Handbook	1.50p
BP34	Practical Repair and Renovation of Colour TVs	1.25p
BP35	Handbook of IC Audio Preamplifier & Power Amplifier Construction	1.75p
BP36	50 Circuits Using Germanium, Silicon and Zener Diodes	95p
BP37	50 Projects Using Relays, SCR's and TRIACs	1.75p
BP39	50 (FET) Field Effect Transistor Projects	1.75p
BP40	Digital IC Equivalents and Pin Connections	2.50p
BP41	Linear IC Equivalents and Pin Connections	3.75p
BP42	50 Simple L.E.D. Circuits	95p
BP43	How to Make Walkie-Talkies	1.50p
BP44	IC555 Projects	1.95p
BP45	Projects in Opto-Electronics	1.75p
BP46	Radio Circuits Using IC's	1.35p
BP47	Mobile Discotheque Handbook	1.35p
BP48	Electronic Projects for Beginners	1.35p
BP49	Popular Electronic Projects	1.45p
BP50	IC LM3900 Projects	1.35p
BP51	Electronic Music and Creative Tape Recording	1.75p
BP52	Long Distance Television Reception (TV–DX) for the Enthusiast	1.95p
BP53	Practical Electronic Calculations and Formulae	2.95p
BP55	Radio Stations Guide	1.75p
BP56	Electronic Security Devices	1.75p
BP57	How to Build Your Own Solid State Oscilloscope	1.50p
BP58	50 Circuits Using 7400 Series IC's	1.50p
BP59	Second Book of CMOS IC Projects	1.50p
BP60	Practical Construction of Pre-amps, Tone Controls, Filters & Attn	1.45p
BP61	Beginners Guide to Digital Techniques	95p
BP62	Elements of Electronics – Book 1	2.25p
BP63	Elements of Electronics – Book 2	2.25p
BP64	Elements of Electronics – Book 3	2.25p
BP65	Single IC Projects	1.50p
BP66	Beginners Guide to Microprocessors and Computing	1.75p
BP67	Counter Driver and Numeral Display Projects	1.75p
BP68	Choosing and Using Your Hi-Fi	1.65p
BP69	Electronic Games	1.75p
BP70	Transistor Radio Fault-Finding Chart	50p
BP71	Electronic Household Projects	1.75p
BP72	A Microprocessor Primer	1.75p
BP73	Remote Control Projects	1.95p
BP74	Electronic Music Projects	1.75p
BP75	Electronic Test Equipment Construction	1.75p
BP76	Power Supply Projects	1.75p
BP77	Elements of Electronics – Book 4	2.95p
BP78	Practical Computer Experiments	1.75p
BP79	Radio Control for Beginners	1.75p
BP80	Popular Electronic Circuits – Book 1	1.95p
BP81	Electronic Synthesiser Projects	1.75p
BP82	Electronic Projects Using Solar-Cells	1.95p
BP83	VMOS Projects	1.95p
BP84	Digital IC Projects	1.95p
BP85	International Transistor Equivalents Guide	2.95p
BP86	An Introduction to Basic Programming Techniques	1.95p
BP87	Simple L.E.D. Circuits – Book 2	1.35p
BP88	How to Use Op-Amps	2.25p
BP89	Elements of Electronics – Book 5	2.95p
BP90	Audio Projects	1.95p
BP91	An Introduction to Radio DX-ing	1.95p
BP92	Electronics Simplified – Crystal Set Construction	1.95p
BP93	Electronic Timer Projects	1.95p
BP94	Electronic Projects for Cars and Boats	1.95p
BP95	Model Railway Projects	1.95p
BP96	C B Projects	1.95p
BP97	IC Projects for Beginners	1.95p
BP98	Popular Electronic Circuits – Book 2	2.25p
BP99	Mini-Matrix Board Projects	1.95p
BP100	An Introduction to Video	1.95p
BP101	How to Identify Unmarked IC's	65p
BP102	The 6809 Companion	1.95p
BP103	Multi-Circuit Board Projects	1.95p
BP104	Electronic Science Projects	2.25p
BP105	Aerial Projects	1.95p
BP106	Modern Op-Amp Projects	1.95p

BERNARD BABANI BP92

Electronics Simplified - Crystal Set Construction

- This is a book written especially for those who wish to participate in the intricacies of electronics more through practical construction than by theoretical study. It is designed for all ages upwards from the day when one can read intelligently and handle simple tools.

- This book replaces several now out-of-date ones on crystal set construction that we previously published, especially the very successful book No. 126 ''Boys Book of Crystal Sets''.

- So many crystal set circuits exist that it is doubtful whether anything novel could be designed, hence the ones here are based on those from earlier publications, but completely modified to use modern inexpensive components and home-wound coils.

- A book highly recommended for all newcomers to the fascinating hobby of electronics.

ISBN 0 85934 067 8

BERNARD BABANI (publishing) LTD
The Grampians
Shepherds Bush Road
London W6 7NF
England

£1.75

www.ZapperLab.com

Order A Hulda Clark Zapper - **Order Dr. Clark's Books** - **Order Electrical Supplies**

Welcome to the Zapper Lab. Here we have the step-by-step instructions for building your very own Hulda Clark Zapper, reprinted with permission from the books by Dr. Hulda Clark. Dr. Clark has improved her original zapper circuit and published this update in 2003. Knowing her readers have varying experience with building electronics, this recent update released several different sets of instructions on how to build a zapper, depending on one's electronic expertise.

- The original published Zapper instructions require soldering or twisting the parts together. It is best when soldered, which does require some special tools and skill. These instructions are found reprinted below from Dr. Clark's 2003 update.
- A newer, easier to build design employs the use of an electronic *breadboard*. A breadboard does not require soldering, as one can push the electronic components into place onto the breadboard to easily build a circuit. Dr. Clark published designs for the 30 kHz Zapper, and the 1000 Hz Zappicator to be built on a breadboard. The instructions for building both breadboard zapper circuits are at ClarkZapper.net along with a nifty on-line video to help guide you through the process.
- And last, for advanced electricians, Dr. Hulda Clark published a schematic for building the updated Zapper (see photo below).

If you do not know anyone who could put one together for you, you can order a pre-built Zapper at the web site, www.ClarkZapper.com. Remember, the Hulda Clark Zapper is not a licensed medical device. It has not been tested by the FDA or AMA for use on humans or animals. The Clark Zapper can be sold and used only as an experimental device, and no claim can be made as to a diagnosis, cure or treatment for any medical condition or disease. Dr. Clark's book states to not experiment with a zapper if you are either pregnant or wearing an electronic pacemaker. Please be sure to read one of Dr. Clark's books before using a Hulda Clark Zapper.

Order A Hulda Clark Zapper - **Order Dr. Clark's Books** - **Order Electrical Supplies**

How To Build A Hulda Clark Zapper

click here to print the 2003 schematic for building a zapper

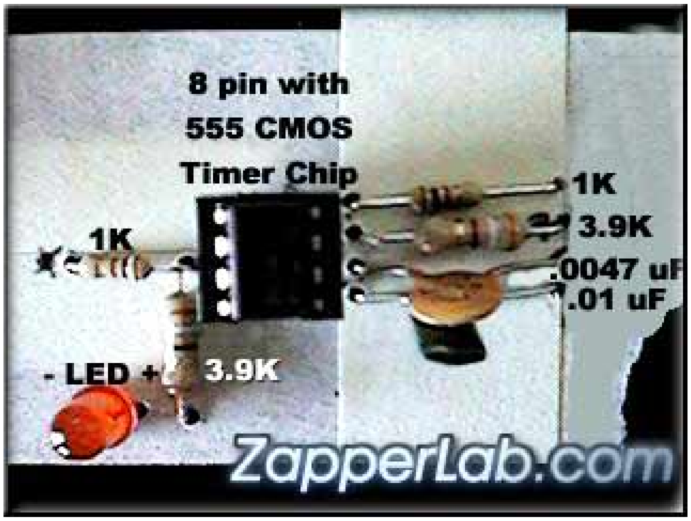

Order A Hulda Clark Zapper

Instructions for building your very own zapper can be found reprinted below from the books by Dr. Hulda Clark. You can get all the parts you need at online electrical supply stores such as: PositiveOffset.com. If you would rather order a pre-built zapper, just click: Clark Zapper.

Dr. Clark has released new devices called the *Food Zappicator* and the *Toothbrush Zappicator.* For the instructions on how to build this new type of zapper, visit: FoodZappicator.com and ToothZappicator.com

NEW! Build A Hulda Clark Zapper Video Instructions
Nine short videos that take you step-by-step through the process of building a Hulda Clark Zapper.

How To Build A Hulda Clark Zapper

***The Cure For All Cancers* © Copyright 1993 by Hulda Regehr Clark, Ph.D., N.D.**
Zapper instructions have been updated for 2003 by Dr. Hulda Clark

Hints for absolute novices: Don't let unusual vocabulary deter you. A "lead" is just a piece of wire used to make connections. When you remove a component from its package, label it with a piece of tape. A serrated kitchen knife works best, as does a large safety pin. Practice using the micro clips. If the metal ends are L-shaped bend them into a U with the long-nose pliers so they grab better. Chips and chip holders (wire wrap sockets) are very fragile. It is wise to purchase an extra one of each in case you break the connections. The "555" timer is a widely used component; if you can't locate this one, try another electronics shop.

The Shoebox Way

This circuit has been improved since the one given in earlier books. A resistor has been added that gives every pulse an added Positive Offset of 1/4 volt. You no longer need to operate your zapper so daringly close to a Negative voltage.

Get from Radio Shack, (serial number included to make it easy). Order these parts from their web site at: www.RadioShack.com. If any electrical parts are unavailable at Radio Shack, visit another on-line electrical supplier of these items, such as: www.PositiveOffset.com to buy your electrical parts.

Parts List for the 30,000 Hz Zapper Circuit	
Item	**Radio Shack Catalog Number**
Use Shoebox or Buy a Plastic Project Box	270-1809
9 volt battery	
9 volt battery snap connector	270-324 (set of 5 you need 1)
on-off toggle switch	275-624A micro mini toggle switch
If not available, choose any toggle switch with holes in the contact points or Radio Shack 275-612	
1 K Ohm resistor, brown-black-red-gold	271-312 (500 piece assortment) use 2
3.9 K Ohm resistor, orange-white-red-gold	Use 2 from 500 piece assortment
39 K Ohm resistor, orange-white-orange-gold	From 500 piece assortment
low-current red LED	276-044
.0047 uF capacitor	272-130 (set of 2, you need 1)
.01 uF capacitor	272-131 (set of 2, you need 1)
555 CMOS timer chip (TLC 555)	276-1718 (you may wish to buy a spare)
8 pin wire-wrapping socket for the chip	900-7242
If only 16 pin sockets are available, cut one in half OR leave half empty. Editor's Note: These parts have been discontinued at most Radio Shacks. 8 pin wire wrap sockets are available at many other electrical supply stores such as: www.PositiveOffset.com	
short (12") alligator clip leads	any electronics shop, get 10
If not available, use 14" length from Radio Shack, 278-1156	
Micro clip jumper wires	278-017 (you need 2 packages of 2)
If not available, use mini-clip jumper wires 278-016	
2 bolts, about 1/8" diameter, 2" long, with 4 nuts and 4 washers	hardware store
2 copper pipes, 3/4" diameter, 4" long	hardware store
Sharp knife, pin, long-nose pliers, tape, 4 twist ties or rubber bands.	

Editor's Note:

The 8 pin wire-wrapping socket has been discontinued at many Radio Shacks and readers have asked what to do. We know a great electrical supply store where 8 pin wire wrap sockets are available, as well as easy-to-build science kits for making a 30,000 Hz Frequency Generator with all the parts listed above in one simple package.

A Hulda Clark zapper is basically just a common Frequency Generator set at 30,000 Hz, with a 100% positive offset square wave. The cost for the electrical parts to build one is about $30-$40 for them all. Visit the electrical suppliers at PositiveOffset.com for all your electrical needs. They have everything you'd need to build a basic 30,000 Hz Positive Offset Square Wave Frequency Generator, and their kit is much easier to build than the instructions given below, because it uses a breadboard so there is no need to twist or solder wires, you simply push the electrical parts into place on the breadboard. **-editor.**

Assembling The Zapper

If you have tools such as a drill, needle nose pliers, and small drill bits, buy one of the plastic project boxes on the list, otherwise build your zapper in a shoe box, or a box half the size of a shoe box.

1. You will be using the lid of the shoe box or plastic lid of the project box to mount the components. Save the box to enclose the finished project.

2. Pierce two holes near the ends of the lid. Enlarge the holes with a pen or pencil until the bolts would fit through. Mount the bolts on the outside about half way through the holes so there is a washer and nut holding it in place on both sides. Tighten. Label one hole "grounding bolt" on the inside and outside.

3. Mount the 555 chip in the wire wrap socket. Find the "top end" of the chip by searching the outside surface carefully for a cookie-shaped bite or hole taken out of it. Align the chip with the socket and very gently squeeze the pins of the chip into the socket until they click in place.

4. Make 8 pinholes to fit the wire wrap socket. Enlarge them slightly with a sharp pencil. Mount it on the outside. Write in the numbers of the pins (connections) on both the outside and inside, starting with number one to the left of the "cookie bite" as seen from outside. After number 4, cross over to number 5 and continue. Number 8 will be across from number 1. The pins are numbered like this:

1	8
2	7
3	6
4	5

5. Pierce two holes ½ inch apart very near to pins 5,6,7, and 8. They should be less than 1/8 inch away. (Or, one end of each component can share a hole with the 555 chip.) Mount the .01 uF capacitor near pin 5 on the outside. On the inside connect pin 5 to one end of this capacitor by simply twisting them together. Loop the capacitor wire around the

pin first; then twist with the long-nose pliers until you have made a tight connection. Bend the other wire from the capacitor flat against the inside of the shoe box lid. Label it .01 on the outside and inside. Mount the .0047 uF capacitor near pin 6. On the inside twist the capacitor wire around the pin. Flatten the wire from the other end and label it .0047. Mount the 3.9 K Ohm resistor near pin 7, connecting it on the inside to the pin. Flatten the wire on the other end and label it 3.9. Mount the 1 K Ohm resistor and connect it similarly to pin 8 and label it 1K.

6. Pierce two holes ½ inch apart next to pin 3 (again, you can share the hole for pin 3 if you wish), in the direction of the bolt. Mount the other 1 K Ohm resistor and label inside and outside. Twist the connections together and flatten the remaining wire. This resistor protects the circuit if you should accidentally short the terminals. Mount the 3.9 K Ohm resistor downward. One end can go in the same hole as the 1K resistor near pin 3. Twist that end around pin 3 which already has the 1K resistor attached to it. Flatten the far end. Label.

7. Next to the 3.9 K Ohm resistor pierce two holes ¼ inch apart for the LED. Notice that the LED has a positive and a negative connection. The longer wire is the positive (anode). Mount the LED on the outside and bend back the wires, labeling them + and - on the inside.

8. Near the top pierce a hole for the toggle switch. Enlarge it until the shaft fits through from the inside. Remove nut and washer from switch before mounting. You may need to trim away some paper with a serrated knife before replacing washer and nut on the outside. Tighten.

9. Next to the switch pierce two holes for the wires from the battery holder and poke them through. Do not attach the battery yet.

10. An inch away from the switch pierce two holes 1/4 inch apart. Mount the 39 K Ohm resistor on the outside and label it inside and outside as "39 K, Positive offset." Flatten the wires on the inside.

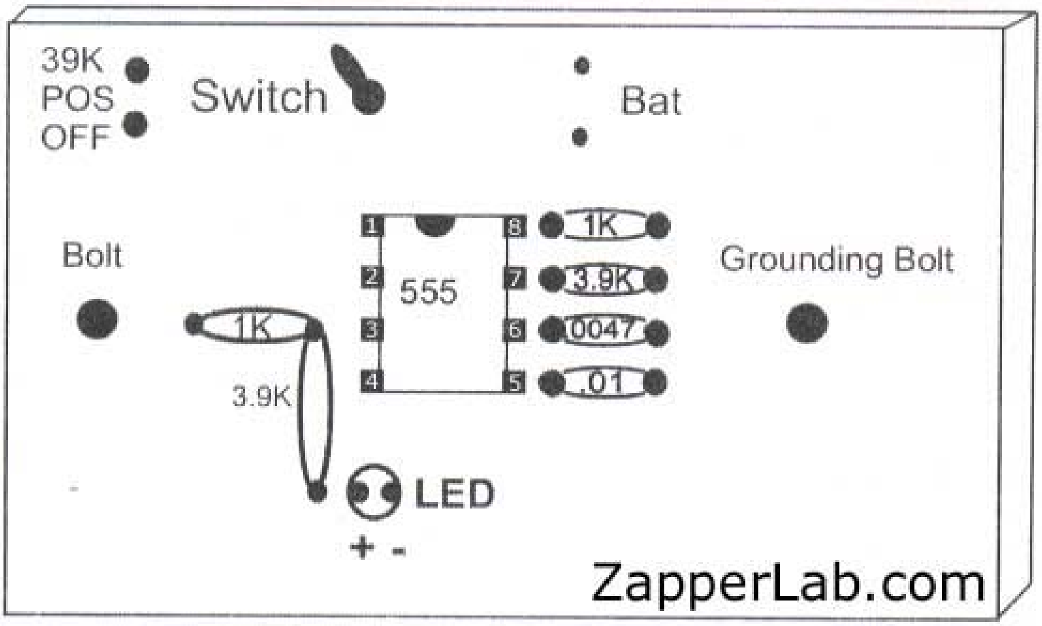

Now To Connect Everything

1. Twist free ends of the two capacitors .01 and .0047 together. Connect this to the Grounding Bolt using an alligator clip.

2. Bend the top ends of pin 2 and pin 6 (which already has a connection) inward towards each other in an L shape. Catch them both with an alligator clip and attach the other end of the alligator clip to the free end of the 3.9 K Ohm resistor by pin 7.

3. Using an alligator clip connect pin 7 to the free end of the 1 K Ohm resistor attached to pin 8.

4. Using three micro clips connect pin 8 to one end of the switch, pin 4 to the same end of the switch, and one end of the offset resistor to the same end of the switch. (Put one hook inside the hole and the other hooks around the whole connection. Check to make sure they are securely connected.) Connect the free end of the offset resistor to the bolt using an alligator clip.

5. Use an alligator clip to connect the free end of the 1 K Ohm resistor (by pin 3) to the bolt. It is the **output resistor**.

6. Twist the free end of the 3.9 K Ohm resistor by pin 3 around the plus end of the LED. Connect the minus end of the LED to the grounding bolt using an alligator clip.

7. Connect pin number 1 on the chip to the grounding bolt with an alligator clip.

8. Attach an alligator clip to the outside of one of the bolts. Attach the other end to a handhold (copper pipe). Do the same for the other bolt and handhold.

9. Connect the minus end of the battery snap connector (black wire) to the grounding bolt with an alligator clip.

10. Connect the plus end of the battery snap connector (red wire) to the free end of the switch using a micro clip lead. Attach the battery very carefully. Before attaching the battery to its snap connector, cover one terminal with tape. After snapping in one terminal, remove the tape to attach the other terminal. This is to prevent accidental touching of terminals in a backwards direction. If the LED lights up you know the switch is ON. If it does not, flip the switch and see if the LED lights. Label the switch clearly. If you cannot get the LED to light in either switch position, double-check all of your connections, and make sure you have a fresh battery. Even if it does light up, check every connection again.

11. Finally tie up the bunches of wire with twist-ties or rubber bands

and replace the lid on the box. Slip a couple of rubber bands around the box to keep it securely shut. For safer storage, place it inside a larger box.

Note: Having gained this much experience, you may prefer to build your next zapper on a piece of cardboard folded and able to fit inside a shoebox or plastic electronic project box for more protection.

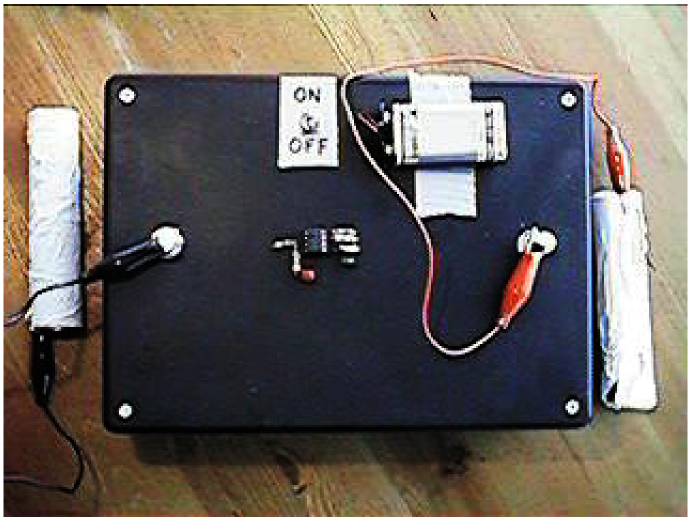

- Optional: measure the frequency of your zapper by connecting an oscilloscope or frequency counter to the handholds. Any electronics shop can do this. It should read between 20 and 40 kHz. The shop can also read the voltage (peak to peak) and the amount of Positive offset (on the .5 volt-per-division scale). The voltage output should be about 8 volts.
- Note: a voltage meter will only read 4 to 5 volts because it displays an average voltage.

Using The Zapper

1. Wrap handholds in one layer of wet paper towel before using. More will reduce the current. Grasp securely and turn the switch on to zap. Keep a bottle of water handy to keep them wet.

2. Zap for seven minutes, let go of the handholds, turn off the zapper, and rest for 20 minutes. Then seven minutes on, 20 minutes rest, and a final seven minutes on. This is the routine for **regular** zapping.

Tell us about your zapper experience at the zapper message board.

To Visit The Zapper Museum - Click Here!

Order A Hulda Clark Zapper - Click Here!

See The *'Build A Zapper'* Video - Click Here!

Learn about Dr Clark's latest zapping technique called: Plate Zapping

Learn about Dr Clark's latest inventions:
The Food Zappicator and The Toothbrush Zappicator

Order Electronic Parts at RadioShack.com and Frequency Generator Kits at PositiveOffset.com

All information is for educational purposes only. These statements have not been evaluated by the FDA. Not intended to diagnose, treat, prevent, mitigate or cure disease. Visit www.HuldaClark.com for more information on Dr. Clark's Syncrometer™ Science, and the Zapper. Legal Disclaimer Privacy Policy

www.HuldaClark.com

The Cure For All Cancers © Copyright 1993 by Hulda Regehr Clark, Ph.D., N.D.

KEELYNET

Zapper Schematics
No claims are made by KeelyNet for any of these plans
They are provided for informational purposes only!
KeelyNet 01/10/02

Simple Zapper Schematic

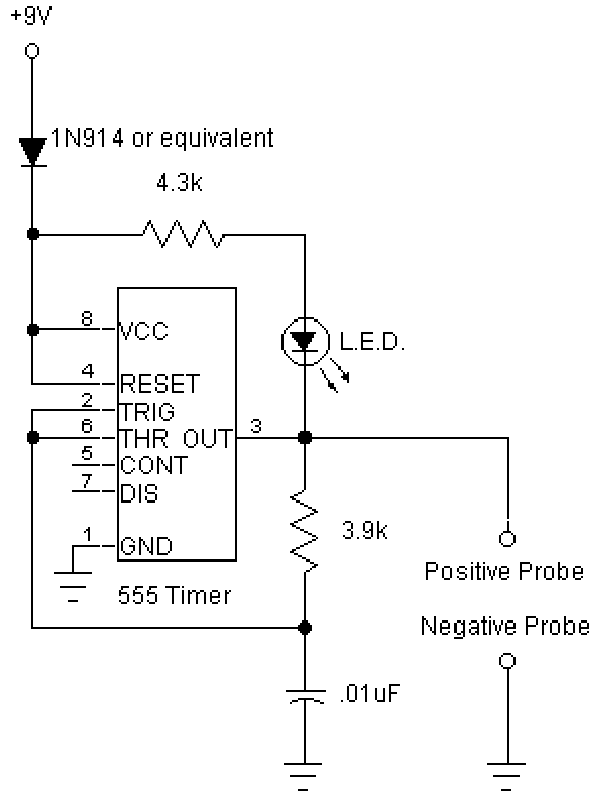
+9V
1N914 or equivalent
4.3k
8 VCC
4 RESET
2 TRIG
6 THR
OUT 3
5 CONT
7 DIS
1 GND
555 Timer
L.E.D.
3.9k
Positive Probe
Negative Probe
.01uF

The Research of Dr. Hulda Regehr Clark, Ph. D., N.D.

I have recently (9/15/2000) come across the research of Dr. Hulda Regehr Clark, Ph. D., N.D. In her books, she states that many diseases, among which are cancer and HIV/AIDS, can be cured by the use of, among other things, certain herbs and a frequency generator called a "Zapper" that was invented by her and her son.

Speaking of mothers :), my mother, 62, had fatigue and arthritic symptoms in her hands before I found out about and told her about Dr. Hulda Clark. She is a seamstress, so the arthritic symptoms hindered her work. After doing the Parasite Cleanses and using the Zapper, she doesn't have to take naps like before and her hands are flexible again. She has done the complete Parasite Cleanse only twice but she Zaps practically every day.

Simple Zapper Information

Schematic updated on 12/21/2001

Simple Zapper Information text updated on 12/21/2001.

I forgot to mention that the Simple Zapper that I'm using has a 1 Megohm potentiometer instead of the 2.2k ohm resistor in order to vary the output frequency. If you use a potentiometer with the circuit put it in series with the 2.2k ohm resistor because the chip gets hot when there is too little resistance.

The duty-cycle is pretty consistent with a range of 51% to 57%. With the oscilloscope's frequency analyzer in storage mode, I hold the Zapper probes and put the oscilloscope probes somewhere on my body and then I vary the frequency of the Zapper using the 1 Megohm potentiometer.

When testing my body on a spectrum analyzer while using the Simple Zapper with an output of around 5 Volts (tested on a DC voltmeter), I noticed that the frequencies around 4,000 hz to around 18,000 hz had larger peaks telling me that my body at the time was more conductive (less reactive?) and/or less resistive to frequencies in this range at the given voltage.

With my body being more conductive (less reactive?) at these frequencies, I believe that there is a better dispersal of the Simple Zapper's output within my body. I did this recent test with one Zapper probe in each hand (as opposed to both in one hand) and the oscilloscope probes (RCA to 1/4" phone plug adaptor) in my mouth near my wisdom teeth with my lips around the RCA ground (like a lollipop).

The oscilloscope's frequency analyzer was in storage mode. With the Simple Zapper circuit on this page, change the 2.2k ohm resistor to 4.3k ohm if you want a ~15,000 hz frequency. I have recently built a couple of Zappers with optimized output frequencies of around 11,000 hz. I haven't tested anyone using the Simple Zapper other than myself, so there is the possibility that your body's response may differ.

For more information, check out my Simple Zapper webpage at:

www.members.aol.com/mas1911/index.html

or for really interesting free circuit schematics:

www.members.tripod.com/~mas1911/zappers20011221tripod.html

You can write the Author of Simple Zapper if you have comments or questions.

The following document, written by Luke Parrish, provides instructions on building an astable multivibrator in DC pulse mode, identical in output to the Zapper, described in The Cure For All Diseases by Dr. Hulda Clark.

Plans For Zapper

This page tells how to make a Hulda Clark-style "Parasite Zapper". It's a simpler and more efficient version, but works just as well. We make no theraputic claims for this device. It is for experimental purposes only. We provide a ready made version for those who prefer not to spend so much time building it.

Materials

- A CD4069 hex inverter (Radio Shack link)
- 1 1M resistor, 1 10k resistor, and 1 1k resistor (Radio Shack link) (Resistor color code)
- 1 1000 pF capacitor (Radio Shack link)

NOTE: The size of the resistors and capacitors can vary, because they determine the frequency of the zapper, which doesn't really matter much according to Hulda Clark.

The following parts may be substituted according to preference, though something similar will be needed:
- Solderless breadboard (Radio Shack link)
- 9-volt battery
- 4 alligator clip-leads sets Radio Shack link
- 2 copper pipes (hand-holds)

Instructions

Plug the components into the solderless breadboard according to the following schematic. Make sure none of the wires touch each other. To hook up the power, attach an alligator clip-lead to each side of the capacitor that stretches from hot to ground. Clip the one that goes to hot to the positive terminal on the battery and the one that goes to ground to the negative terminal. The two hand-holds are attached with the other two clip-leads. (One goes to ground, the other to the end of a resistor that goes to pin 12.)

Schematic

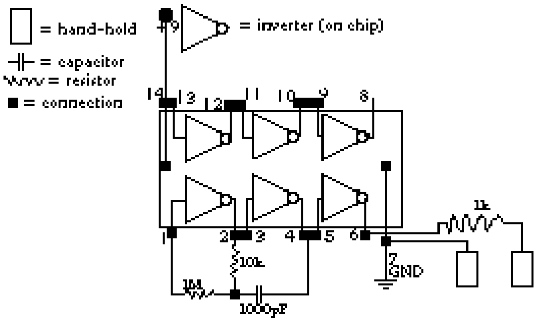

Schematic most recently updated 5-17-2001

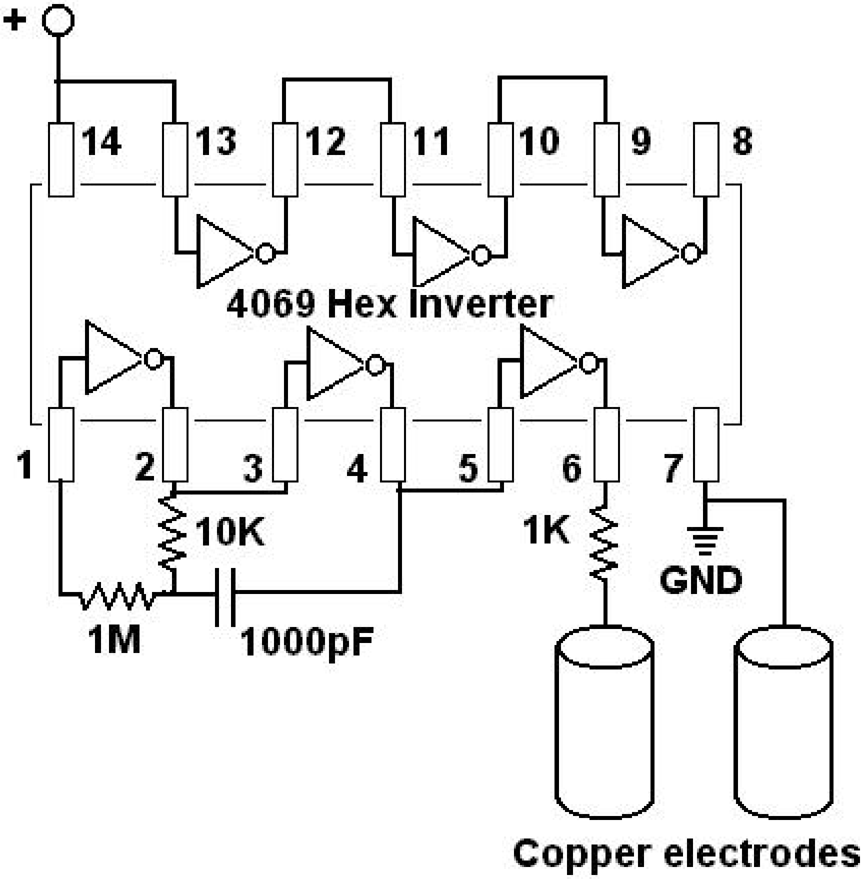

The above instructions are not the only way to do it. You can, for example, solder all the components directly to a 14-pin socket, like we do in our ready made version. This is a lot more trouble and more time-consuming for most people, which is why we recommend a breadboard. The resistors and capacitor don't all have to be the exact sizes mentioned here to make a working Zapper. However, any variation in capacitance or resistivity will change the frequency. (There are certain limits on how high of frequencies the chip can handle, so don't stray too far from the specified values.) If you have an oscilliscope, that's handy to check if it's working.

Ready made Zappers

Free plans from Zapperlab.com

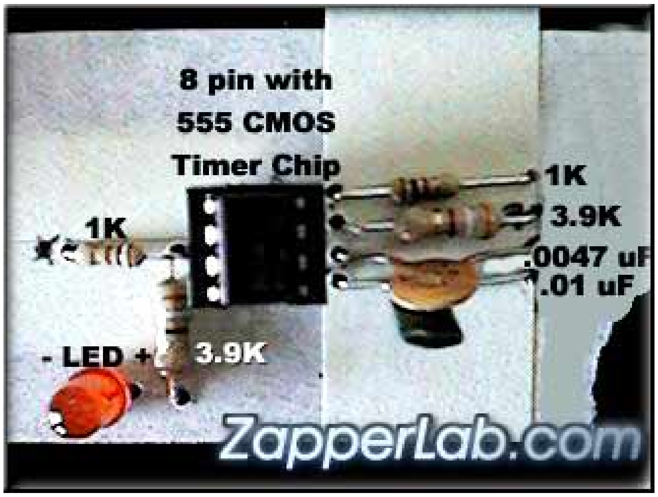

Here we have the easy step-by-step instructions for building your very own Hulda Clark Zapper reprinted with permission from her books (provided below). For advanced electricians there is a schematic for building a Hulda Clark Zapper as well. If you do not know anyone that could put one together for you, you can now order the zapper Hulda Clark uses as described in "The Cure For All Cancers" and "The Cure For All Diseases" at: www.SyncroZap.com. The Dr. Clark Zapper can be sold only as an experimental device, and no claim can be made by us legally as to a diagnosis, cure or treatment for any medical condition or disease.

Build A Dr. Clark Zapper

Order A Dr. Clark Zapper

Instructions for building your very own zapper can be found reprinted below from the books by Dr. Hulda Clark. If you would rather order a pre-built zapper, just click here.

How To Build A Zapper from *The Cure For All Cancers*:

The Cure For All Cancers © Copyright 1993 by Hulda Regehr Clark, Ph.D., N.D.

Get from Radio Shack, (serial number included to make it easy).
Order these parts from their web site at:
www.RadioShack.com

___ Black plastic Project Box Radio Shack # 270-1809

___ 9 volt battery

___ 9 volt battery clips: Radio Shack # 270-325

___ On off switch: Radio Shack # 275-624A

___ 1K Ohm resistor: RS# 271-1321

___ 3.9K Ohm resistor: RS# 271-1123

___ low current red LED: RS# 276-044

___ .0047 uF capacitor: RS# 272-130

___ .01 uF capacitor: RS# 272-1065

___ 555 CMOS timer chip: RS# 276-1723

___ 8 pin wire wrapping socket for CMOS chip: RS# 276-1988

___ two packs of Microclip test jumpers: RS# 278-017

___ one pack of 14" alligator clip leads RS# 278-1156C

Get from any Hardware Store:

2 Bolts 1/8" diameter, 2" long with 4 nuts and washers.

2 pieces of ¾" copper pipe, cut to 4 inches long each.

Editor's Note:

Remove 8 Pins From A 16 Pin Wire Wrap Socket

The **8 pin wire-wrapping socket** has been discontinued at many Radio Shacks and readers have asked what to do. Some say it is still available on their website at www.RadioShack.com, if not:

There are a couple options. One can still get the **16 pin wire-**

wrapping socket. Pull out the bottom grouping of 8 pins, and you now have an 8 pin socket. Just make sure when you plug the 555 Timer into it, you line it up with the remaining 8 pins. Very simple.

There is also a **Low Profile 8 pin socket**. The pins are so small one must solder to them.

The wire-wrap socket has long pins for easy twisting onto other wires. The low profile is for those that solder electronics.

If you are soldering your zapper together get the low profile 8 pin socket. Since the socket is just a saddle for the 555 Timer, it is not a vital part of the circuit. It's just a way to put long or short pins on the 555 Timer for easy twisting. If you cannot find one at your local Radio Shack, simply get the 16 pin wire-wrap socket, and pull out 8 pins. -editor.

Assembling The Zapper

If you have tools such as a drill, needle nose pliers, and small drill bits, buy one of the plastic project boxes on the list, otherwise build your zapper in a shoe box, or a box half the size of a shoe box.

1. You will be using the lid of the shoe box or plastic lid of the project box to mount the components. Save the box to enclose the finished project.

2. Pierce two holes near the ends of the lid. Enlarge the holes with a pen or pencil until the bolts would fit through. Mount the bolts on the outside about half way through the holes so there is a washer and nut holding it in place on both sides. Tighten. Label one hole "grounding bolt" on the inside and outside.

3. Mount the 555 chip in the wire wrap socket. Find the "top end" of the chip by searching the outside surface carefully for a cookie-shaped bite or hole taken out of it. Align the chip with the socket and very gently squeeze the pins of the chip into the socket until they click in place.

4. Make 8 pinholes to fit the wire wrap socket. Enlarge them slightly with a sharp pencil. Mount it on the outside. Write in the numbers of the pins (connections) on both the outside and inside, starting with number one to the left of the "cookie bite" as seen from outside. After number 4, cross over to number 5 and continue. Number 8 will be across from number 1. The pins are numbered like this:

1	8
2	7

3 6

4 5

5. Pierce two holes ½ inch apart very near to pins 5,6,7, and 8. They should be less than 1/8 inch away. (Or, one end of each component can share a hole with the 555 chip.) Mount the .01 uF capacitor near pin 5 on the outside. On the inside connect pin 5 to one end of this capacitor by simply twisting them together. Loop the capacitor wire around the pin first; then twist with the long-nose pliers until you have made a tight connection. Bend the other wire from the capacitor flat against the inside of the shoe box lid. Label it .01 on the outside and inside. Mount the .0047 uF capacitor near pin 6. On the inside twist the capacitor wire around the pin. Flatten the wire from the other end and label it .0047. Mount the 3.9 K Ohm resistor near pin 7, connecting it on the inside to the pin. Flatten the wire on the other end and label it 3.9. Mount the 1 K Ohm resistor and connect it similarly to pin 8 and label it 1K.

6. Pierce two holes ½ inch apart next to pin 3 (again, you can share the hole for pin 3 if you wish), in the direction of the bolt. Mount the other 1 K Ohm resistor and label inside and outside. Twist the connections together and flatten the remaining wire. This resistor protects the circuit if you should accidentally short the terminals. Mount the 3.9 K Ohm resistor downward. One end can go in the same hole as the 1K resistor near pin 3. Twist that end around pin 3 which already has the 1K resistor attached to it. Flatten the far end. Label.

7. Next to the 3.9 K Ohm resistor pierce two holes ¼ inch apart for the LED. Notice that the LED has a positive and a negative connection. The longer wire is the positive (anode). Mount the LED on the outside and bend back the wires, labeling them + and - on the inside.

8. Near the top pierce a hole for the toggle switch. Enlarge it until the shaft fits through from the inside. Remove nut and washer from switch before mounting. You may need to trim away some paper with a serrated knife before replacing washer and nut on the outside. Tighten.

9. Next to the switch pierce two holes for the wires from the battery holder and poke them through. Attach the battery and tape it to the outside.

Now To Connect Everything

Twist free ends of the two capacitors .01 and .0047 together. Connect to Grounding Bolt using an alligator clip.

Bend pin 2 and pin 6 together inward, using an alligator clip, catch them and connect to free end of 3.9K Ohm (by pin 7).

Alligator clip Pin 7 to free end 1K Ohm (near pin 8)

Using two microclips Pin 8 and pin 4 to one end of switch (use hole to attach both microclips).

Free end of 1K Ohm (by pin 3) to the Bolt using an alligator clip.

Alligator clip the free end of 3.9K Ohm (by pin 3) to plus end (long) of LED.

Minus end (short) of LED to Grounding Bolt with an alligator clip.

Alligator clip pin 1 to the Grounding Bolt.

Alligator clip Black battery wire to Grounding Bolt.

Micro clip the Red battery wire to free end of switch.

Hook up the battery, and the light should turn on, click the switch if it does not, check connections if LED does not light up.

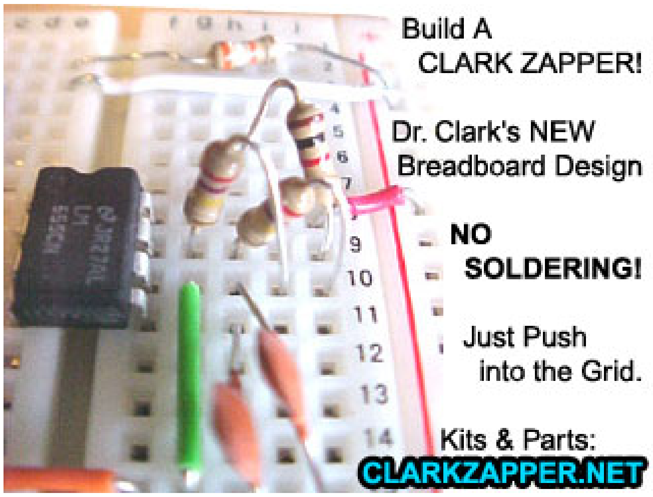

Attach a long lead wire to the grounding bolt & bolt, and then to the copper handles. Wrap handles in paper towel, clip with lead wire, wet handle when you are ready to zap.)

How To Zap

Wrap handles in wet paper towel, or wet cotton sleeves, and hold in each hand. If you have wrist cuffs, run them under water to wet them, attach to each wrist.

Attach handles or wrist cuffs to zapper.

Turn zapper on, zapping for 7 minutes.

At end of 7 minute zapping rest 20 minutes.

Do this a minimum of three 7 minute zapping sessions, with a 20 minute break in between each session. Some people zap five times in extreme cases.

Order A Dr. Hulda Clark Zapper - Click Here!

Learn about Dr Clark's latest zapping technique at: www.PlateZapping.com

All information is for educational purposes only. These statements have not been evaluated by the FDA. Not intended to diagnose, treat, prevent, mitigate or cure disease.

This Circuit designed by Geoff Clark and approved by Dr. Hulda Clark

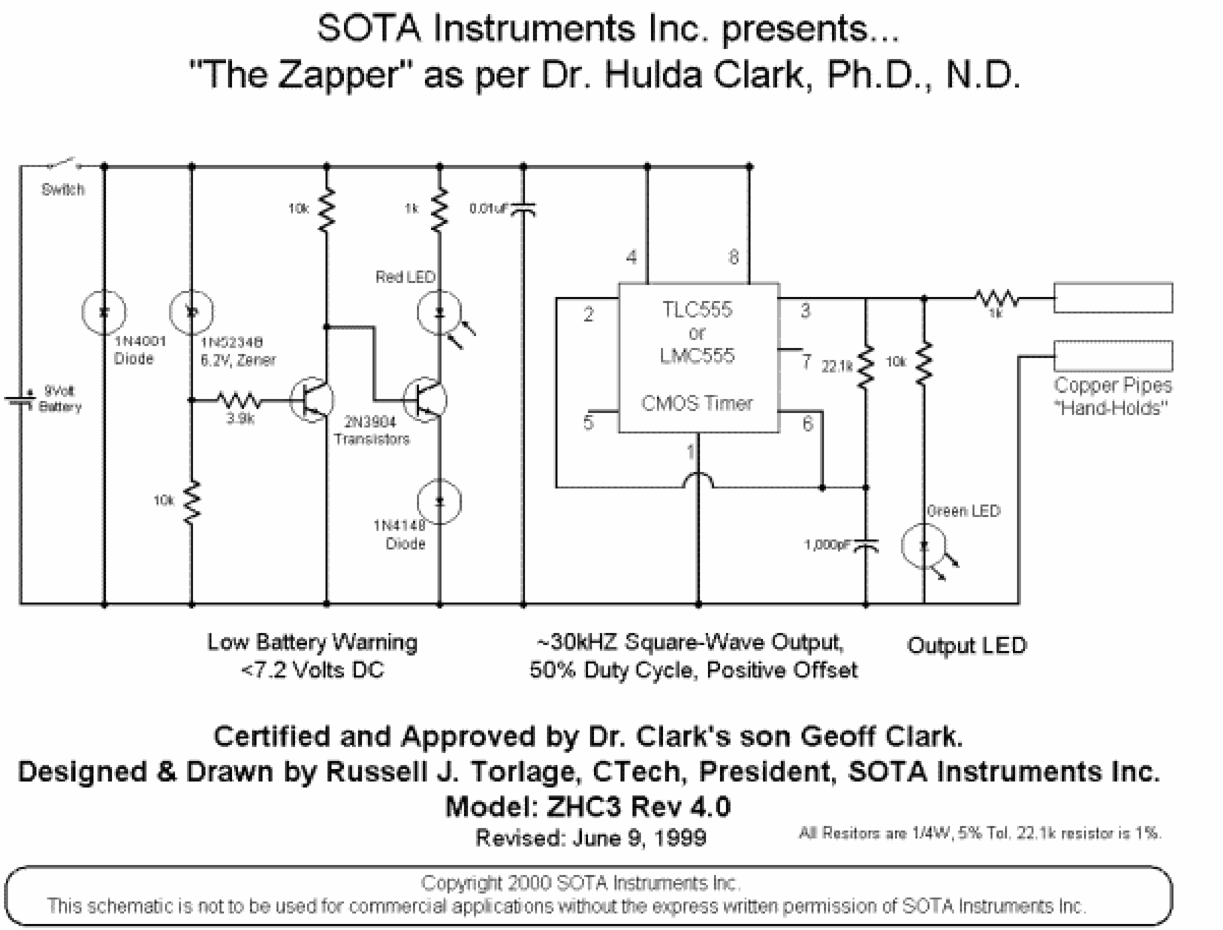

SOTA Instruments Inc. presents...
"The Zapper" as per Dr. Hulda Clark, Ph.D., N.D.

Switch
9Volt Battery
1N4001 Diode
1N5234B 6.2V, Zener
10k
3.9k
10k
1k
Red LED
2N3904 Transistors
1N4148 Diode
0.01uF
4
8
2
3
7
5
6
1
TLC555 or LMC555
CMOS Timer
22.1k
10k
1k
1,000pF
Green LED
Copper Pipes "Hand-Holds"
Low Battery Warning <7.2 Volts DC
~30kHZ Square-Wave Output, 50% Duty Cycle, Positive Offset
Output LED

Certified and Approved by Dr. Clark's son Geoff Clark.
Designed & Drawn by Russell J. Torlage, CTech, President, SOTA Instruments Inc.
Model: ZHC3 Rev 4.0
Revised: June 9, 1999

All Resitors are 1/4W, 5% Tol. 22.1k resistor is 1%.

Copyright 2000 SOTA Instruments Inc.
This schematic is not to be used for commercial applications without the express written permission of SOTA Instruments Inc.

If you found this file useful or interesting, please consider a donation or a purchase to help keep KeelyNet online and providing free information. Even a dollar will help. Others sell it, we prefer to share it, thanks!

PayPal DONATE

www.PositiveOffset.com

Electronic Kit & Parts for Educators, Researchers & Hobbyists

30 kHz / 1 kHz Zapper Kit
with Color Illustrated Instructions

$50.00 Add to Cart

Our Zapper Kit contains all the parts necessary to build either a 30,000 Hz or a 1,000 Hz Positive Offset Square Wave Frequency Generator Unit. This breadboard kit is based on Dr. Hulda Clark's zapper circuit design that was recently updated and published in 2003. Copper Pipes and 9 Volt Battery are not included in kit. For copper pipes and individual electronic parts, click the links in the right column.

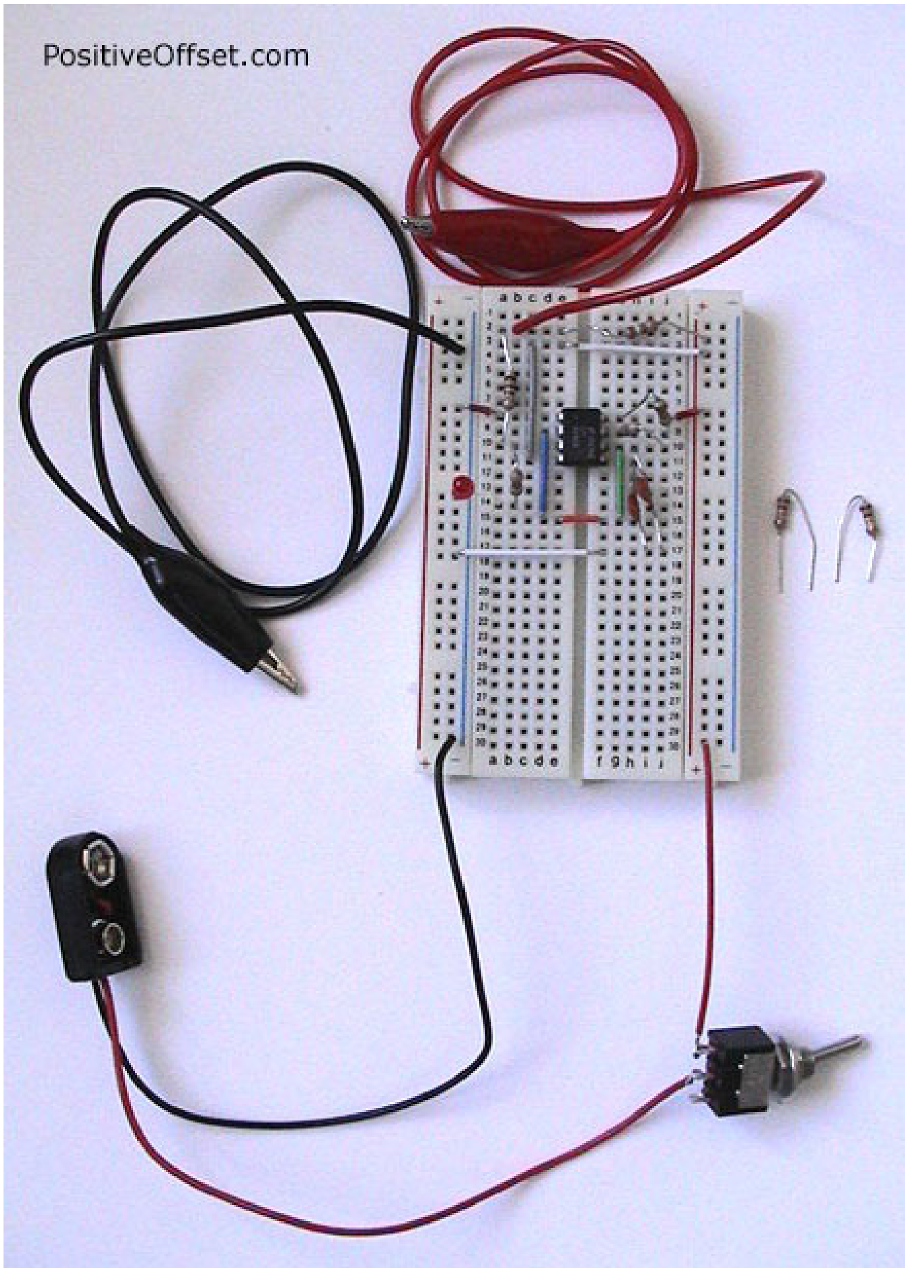

Hulda Clark Zapper Kit with Color Illustrated Instructions

$50.00 Add to Cart

Zapper Kit includes: Breadboard, 555 CMOS Timer, (2) 1K Ohm resistors, 2.2K Ohm resistor, 4.7K Ohm resistor, 39K Ohm resistor, (2) 270K Ohm

www.PositiveOffset.com

- 30 kHz/1 kHz Zapper Kit
- .0047 uF Capacitors
- 1 uH Inductors
- 1K Ohm Resistors
- 2.2K Ohm Resistors
- 3.9K Ohm Resistors
- 4.7K Ohm Resistors
- 39K Ohm Resistors
- 270K Ohm Resistors
- 555 CMOS Timer
- 8 Pin Wire Wrap Socket
- 9 Volt Battery Snap
- Alligator Clip Converter
- Alligator Leads
- Aluminum Square
- Banana to Alligator Leads
- Breadboard
- Color Coded Jump Wires
- Copper Pipes
- Nuts and Bolts
- On/Off Toggle Switch
- Red Mini LED
- Switching diode 1N914
- Video Instructions

Dual Frequency
1KHz 30KHz
www.FrequencyGeneration.com

www.PositiveOffset.com

VIEW CART

resistors, (2) .0047 uF Capacitors, Mini Red LED, (9) Color Coded Jump Wires, 9-Volt Battery Snap, (2) Alligator leads, and an On/Off Switch. All this comes in a sturdy little cardboard box you can use to house your finished circuit. The full-color instructions will help you build either a 30,000 Hz or 1,000 Hz circuit.

It's so easy to build. No soldering required! Watch the video instructions right here (36 minutes).

www.PositiveOffset.com
Breadboard
555 CMOS Timer
Jump Wires (9)
1 long red, 2 white, 1 gray, 2 red,
1 blue, 1 orange, 1 green
On/Off Switch
9-volt Battery Clip
.0047 uF Capacitors
Two 6" Wires 1 red, 1 black
1/4 Watt Resistors (7)
1K
2.2K
4.7K
39K
270K
Red LED
Alligator Leads (2)
Instructions
23:54
HD
vimeo

- YouTube video instructions: 9 easy steps to build this circuit
- Build a zapper Frequently Asked Questions and Answers
- Already have electronic parts? Download PDF Instructions

Remember the Hulda Clark Zapper has not been licensed as a medical device by the US FDA. Zapper Kit is sold as an educational tool only, no medical claims are intended by the sale of these electronic parts. Our Zapper Kit is a great project for a school science fair, or to learn the basics of building simple electronics.

Please note that when ordering electronic parts and kits we can assume no liability on the customer's ability to use them. Safety, craftsmanship, experience, and ability are the user's responsibility.

Electronic parts are small and fragile. We can not offer refunds or replacements on electronic parts. Buy electronic parts and kits only if you believe you can accomplish the work. If you experience difficulty with your build, or have little experience in electronics, please order replacement parts along with your kit.

The 555 CMOS Timer is the most likely part to need replacement, so it may be smart to order extras when building a kit.

The 555 CMOS Timer is vulnerable to static electricity. If you walk on a carpeted floor and the 555 receives a static shock it may burn out and need replacement.

555 CMOS Timers are sold separately for those who want to have several at hand during a kit build, or to troubleshoot a circuit that is not working.

If you are new to electronics, or just want to be cost effective, you may want to purchase extra 555 CMOS Timers when building your first kit. We can not offer refunds or replacements on electronic parts.

View Shopping Cart / Checkout

View Shopping Cart / Checkout

Customer Service Information

We accept all major credit cards via our secure shopping cart. We also accept checks and money orders through the mail. All packages are shipped United States Postal Service Priority Mail with Delivery Confirmation. If you have an existing order and would like to track your package's Delivery Confirmation, click: **USPS Delivery Confirmation**. If you have any other question or concern about your order, please send an email to Customer Service.

Natural Health Supply .com

www.PositiveOffset.com
Electronic Kit & Parts for Educators, Researchers & Hobbyists

Hulda Clark 30 kHz / 1 kHz Zapper Kit
with Color Illustrated Instructions
Order Zapper Kit Here

Hulda Clark Zapper Kit includes: Breadboard, 555 CMOS Timer, 7 resistors: (2) 1K Ohm, 2.2K Ohm, 4.7K Ohm, 39K, (2) 270K Ohm, (2) .0047 uF Capacitors, Red LED, (8) Color Jump Wires, (2) Alligator Leads, and Soldered Battery Snap/On-Off Switch. All this comes in a sturdy little cardboard box you can use to house your finished circuit. The full-color instructions will help you build either a 30,000 Hz or 1,000 Hz circuit. Kit does not include copper pipes. Click here to order copper pipes.

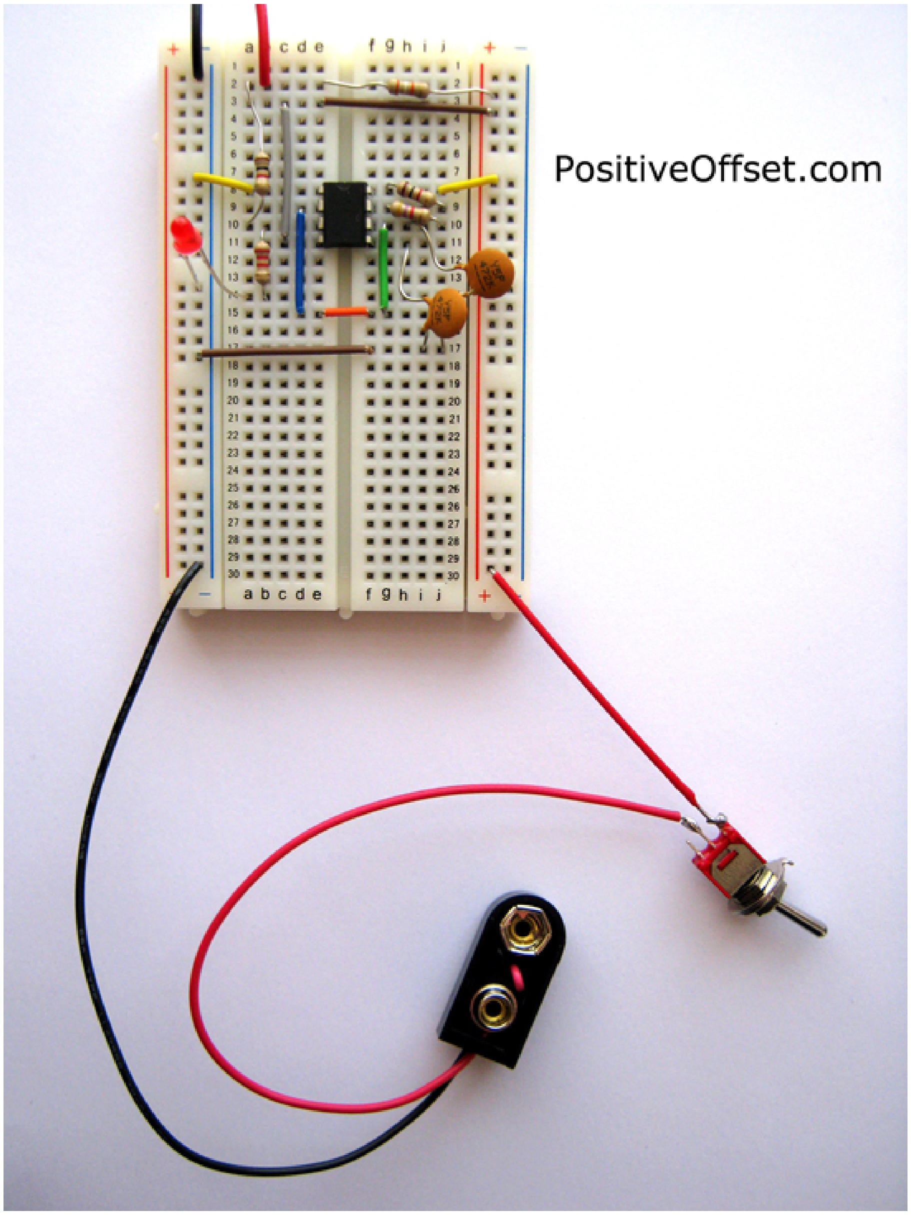

Hulda Clark Zapper Kit with Color Illustrated Instructions
Order Zapper Kit Here

If you are new to building electronics, you may also want to watch the instruction video. The 'Build A Zapper Video' shows how to build this Clark Zapper breadboard kit. Watch the video below. (Note: There are minor updates to the "Build Your Own Zapper" Kit. We have adjusted the jump wires for a

www.PositiveOffset.com

- 30 kHz/1 kHz Zapper Kit
- 555 CMOS Timer
- Switching diode 1N914
- Mini Alligator Leads
- Copper Pipes
- Dial-A-Zap
- Chair Zappicator
- Pet Zappicator
- Video Instructions

www.ClarkZapper.com

VIEW CART

better fit. We have also pre-soldered the 9-volt battery snap to the on/off switch. These updates are not shown in this video, but are clearly described in the kit instructions.)

It's so easy to build. No soldering required! Watch the video instructions right here (36 minutes).

More Video Instructions

- Build a Zapper, Plate Zapper, Food Zappicator, and Toothbrush Zappicator
- Zapper Kit Frequently Asked Questions and Answers
- Already have electronic parts? Download PDF Instructions

The Hulda Clark Zapper has not been licensed as a medical device by the US FDA. Zapper Kit is sold as an educational tool only, no medical claims are intended by the sale of these electronic parts. Our Zapper Kit is a great project for a school science fair, or to learn the basics of building simple electronics.

Please note that when ordering electronic parts and kits we can assume no liability on the customer's ability to use them. Safety, craftsmanship, experience, and ability are the user's responsibility.

Electronic parts are small and fragile. We can not offer refunds or replacements on electronic parts. Buy electronic parts and kits only if you believe you can accomplish the work.

The 555 CMOS Timer is the most likely part to need replacement, so it may be smart to order extras when building a kit.

The 555 CMOS Timer is vulnerable to static electricity. If you walk on a carpeted floor and the 555 receives a static shock it may burn out and need replacement. If you ever touch the 9 volt battery backwards to the battery snap, it may burn out the 555. You may want to protect your 555 by adding a 1N914 Switching Diode to your kit (diode sold separately).

555 CMOS Timers are sold separately for those who want to have several at hand during a kit build, or to troubleshoot a circuit that is not working. If you are new to electronics, or just want to be cost effective, you may want to purchase extra 555 CMOS Timers when building your first kit. We can not offer refunds or replacements on electronic parts.

View Shopping Cart / Checkout

Customer Service Information

We accept all major credit cards via our secure shopping cart. We also accept checks and money orders through the mail. All packages are shipped United States Postal Service Priority Mail with Delivery Confirmation. If you have an existing order and would like to track your package's Delivery Confirmation, click: **USPS Delivery Confirmation**. If you have any other question or concern about your order, please send an email to Customer Service.

2,100 ASANAS

THE COMPLETE YOGA POSES

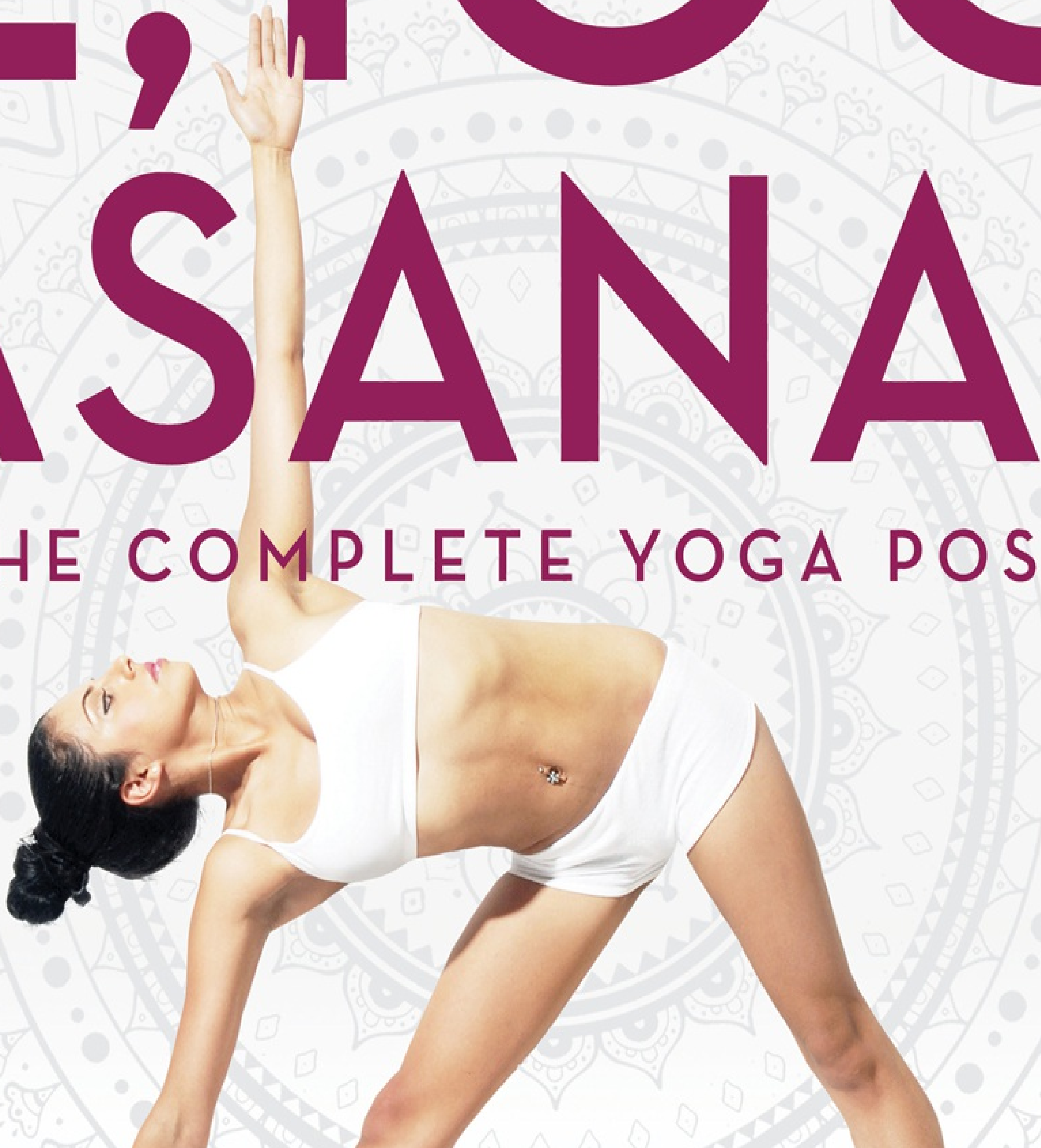

DANIEL LACERDA

FOUNDER OF MR. YOGA, INC.

2,100 ASANAS

The Complete Yoga Poses

Daniel Lacerda

FOUNDER OF MR. YOGA, INC.

BLACK DOG
& LEVENTHAL
PUBLISHERS
NEW YORK

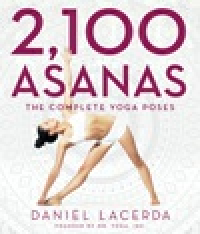

Begin Reading

Table of Contents

Newsletters

Copyright Page

In accordance with the U.S. Copyright Act of 1976, the scanning, uploading, and electronic sharing of any part of this book without the permission of the publisher constitute unlawful piracy and theft of the author’s intellectual property. If you would like to use material from the book (other than for review purposes), prior written permission must be obtained by contacting the publisher at permissions@hbgusa.com. Thank you for your support of the author’s rights.

Opening Prayer

Om

For the peaceful resolution
From the illusionary nature of dualistic existence,
I ground myself before the lotus feet of the gurus,
Who remind me that the light I search for is within me
Bringing stillness to the whirling of the cascading ego mind
I behold the awakened joy of my own Soul
Realizing the truth of pure radiance
That we are in fact the same
To the self-awakened gurus of the past, present, and future,
I salute

Om

Daniel Lacerda, Mr. Yoga

A MESSAGE FROM MR. YOGA

The ultimate goal of yoga is self-realization. You do not need to go to the mountaintop to find it or pay a teacher to show you the way. There are currencies that we exchange with one another that are much more valuable than money: kindness, selflessness, being one part of the greater good. Nor do you need to look outside yourself. If you have an open mind, a sincere desire to learn and to apply that knowledge on a daily basis, and the commitment to follow through on what you've begun, you can achieve self-realization.

Self-realization is the knowledge that we sentient beings are interconnected and that what we think, say, and do affects those around us. Burdened by the pressures and demands that exist outside of ourselves—of our jobs, bills, desire for status and for material possessions—we forget this. Self-realization is the ability to achieve freedom from these demands and to know that true happiness comes from fulfilling our own potential and from lifting up those around us without the thought of self-gain. Dedicating yourself to the regular practice of yoga can help bring you back to this place.

Yoga is, indeed, an excellent form of exercise that carries with it many immediate and long-term physical benefits from improved flexibility to stronger muscles and bones. However, yoga is not just about moving through the poses. Mindfulness plays an essential part in any dedicated yoga practice. If performed properly, yoga quiets the mind of all distracting thoughts from the outside world (*chittavritti*, meaning mind chatter), bringing you to a place of peace within. In turn, being mindful of your thoughts will allow you to be mindful of, and truly connected with, your body, thus completing the cycle of mental and physical health that will allow you to enjoy all the wonderful things that life has to offer.

For the past eleven years, I have dedicated my life to yoga, teaching an average of twenty-five classes seven days a week. I have done this to make a difference in the lives of my students. Now it's my pleasure to share this passion and dedication with you.

Namaste!

A HISTORY OF YOGA

Most of us know yoga as a set of poses performed in a gym or yoga studio setting. The majority of yoga styles practiced today were invented in the last quarter of the 20th century and are either a far cry from yoga's roots or have no authentic lineage.

If we really want to examine the roots of yoga, we need to go back to the Harrapan culture, dating back 3,500 years, when yoga was a meditative practice. According to some, around 1500 BCE, Harrapan culture was diminished due to Aryan invasion. Barbarians from Normandy introduced the caste system and enforced a set of religious rituals that involved blood sacrifice practices. Along with these religious practices came sacred scriptures called the Vedas, a large body of spiritual texts originating in India. The word "yoga" was first mentioned in the oldest of the Vedas, *Rig Veda*. It referred to the concept of discipline.

Fast forward to 800 BCE. The Upanishads, a collection of texts that contain some of the earliest concepts of Hinduism, prescribed the method of achieving enlightenment by studying under a teacher and dedicating one's life to a yoga practice. The Upanishads outlined two paths to enlightenment: Karma Yoga (selfless dedication to the service of others) and Jnana Yoga (intense study of spiritual writings). Around the 3rd century BCE, the *Maitrayaniya Upanishad* prescribed a six-step process to enlightenment, which included mastering *pranayama* (breath control), *pratyaharia* (sense withdrawal), *dhyana* (meditation), *dharana* (one-pointed concentration), *tarka* (self-reflection), and *samadhi* (absolute absorption) in order to unite the Atman (individual's spirit) and Brahman (universal spirit or source of creation). The sacred syllable *om* appeared in this particular Upanishad as a symbol of union between mind and breath.

At around the same time that *Maitrayaniya Upanishad* was introduced, *Bhagavad Gita* gained prominence. This scripture combined and mythological tales that later made their way into a celebrated collection of tales, *Mahabharata*. Three methods of devotion were outlined in *Bhagavad Gita*: Karma Yoga, Jnana Yoga, and Bhakti Yoga (devotion).

Compiled around 400 CE by Patanjali, *The Yoga Sutras* introduced the eight-fold path to yoga practice, which is considered to be the classical yoga manual and the foundation of many of today's yoga practices, particularly Ashtanga Yoga. We will hear more about this eight-fold path in The Eight Limbs of Yoga (here), which include *yama* (self-restraint) *niyama* (self-purification by self-restraint and discipline), *asana* (seat or posture), *pranayama* (control of breath), *pratyahara* (sense withdrawal), *dharana* (one-pointed concentration), *dhyana* (meditation), and *samadhi* (total absorption).

Around the 4th century CE, Tantra Yoga emerged. This new form of yoga celebrated the physical body as a vehicle to enlightenment. The philosophy behind Tantra Yoga can be summarized by the idea of uniting all the dualities within a human body (e.g., male and female; good and evil), which gave

Tantra a very sexual reputation. This is, however, a common misunderstanding, since Tantra practices extend far beyond sexuality.

Hatha Yoga was introduced in the 10th century CE. It combined the physicality and conscious intent of using bodily postures, or *asana* practice, and *pranayama* breath control for the goal of self-realization.

In 14th century CE, the Yoga Upanishads were introduced. One of these sacred texts, *Tejo Bindu Upanishad*, added seven more important parts of yoga practice on top of Patanjali's eight. They were as follows: *mula bandha* (root lock), balance, undisturbed vision, *tyaga* (abandonment), *mauua* (quiet), *desha* (space), and *kala* (time).

It was not until the 20th century that yoga gained any kind of popularity in Western Europe and North America. Swami Sivananda Saraswati was one of the first yogis to travel outside of India to spread the teachings of yoga to the West. He established yoga centers in North America at the time Swami Satchidananda also delivered an opening speech at the Woodstock Festival in 1969. However, T. Krishnamacharya is arguably the father of the yoga practice with which Westerners are familiar today. In the 1930s, he began teaching his students the Mysore vigorous sequences of yoga poses that emphasize strength and athletic ability. Students were only allowed to learn the next and more challenging pose after they had grasped the previous one. His three most prominent and influential students are Pattabhi Jois, Iyengar, and Indra Devi. Pattabhi Jois established Ashtanga yoga. It is one of the most popular types of yoga practiced in the West. Iyengar became successful by creating his own sequences of yoga poses, which were characterized by a focus on the alignment of the body and the use of various props. Indra Devi is considered the first famous yogini (female yoga master). Krishnamacharya also educated his son Desikachar in yoga. An engineer by training, Desikachar saw great value in studying yoga only when he was already a college graduate. Desikachar developed Vinyoga, which is a more therapeutic and less intense approach to physical practice, as compared to Ashtanga.

The 21st century presents us with an endless variety of yoga "styles" or "brands," such as Bikram Yoga, Power Yoga, Kundalini Yoga, and countless more. It is important to be open-minded, try as many styles and approaches as possible, and figure out what gives you the best results in terms of achieving both your physical and spiritual goals. There is no wrong way to achieve self-realization. Just make sure you are mindful, patient, practical, and consistent in your practice.

A Note on the Naming of Poses:

One of the ways that the distance from yoga's roots expresses itself in Western culture is in the naming of the poses. "Seated forward bend," "eagle pose," and "dolphin pose," for example, are all imprecise translations of the original Sanskrit name. *Garudasana*, for example, is widely known as eagle pose, but traditionally this pose was named in dedication to Garuda, who is a Hindu deity, portrayed as half-man and half-eagle. He is the charioteer of Lord Vishnu, who is part of the Holy Trinity in Hinduism. Knowing this history adds a whole new dimension to our understanding of the significance, philosophical depth, and essence of the pose, and can in turn enrich our practice.

The poses throughout this book are identified by both their English and Sanskrit names. The English name is a direct translation of the Sanskrit, which sometimes differs from the more common Western name, also provided in the notes. For a literal translation of each part of the Sanskrit name, you may consult the glossary at the back of the book. The intention is to provide you with as much information about the name of the pose and its history as possible, so no matter what style of yoga you practice, you will have the most complete understanding of the names of the poses.

THE EIGHT LIMBS OF YOGA

The Yoga Sutras, also known as *The Eight Limbs (Ashtanga) of Raja (King) Yoga*, was the first fully developed and recorded system of yoga. Created by Patanjali around 400 CE, this system influences much of the yoga that is practiced today. Although most of the sutras were originally focused on mindfulness, the yoga practiced in the West today seems to focus more on the body. Somewhere along the way, it seems, we began to practice the movement of yoga in isolation from its original philosophies.

For those interested in truly integrating the mindfulness of yoga with its movement, I recommend that you read *The Eight Limbs of Yoga* in its entirety and digest it very slowly. Take time to reflect on it piece by piece so you can implement it into both your practice and your daily life. Wisdom is in the doing. The following, however, is a useful summary of *The Eight Limbs of Yoga*, which will introduce you to the basic concepts of the philosophy. A deep understanding of yoga philosophy and history will greatly enhance the benefits of your practice and put you on the path to mindfulness and self-realization.

There is a wonderful lesson in Buddhism that applies here:

Once, a very old king went to see an old hermit who lived in a bird's nest in the top of a tree. He asked the hermit, "What is the most important Buddhist teaching?" The hermit answered, "Do no evil do only good. Purify your heart." The king expected to hear a long and detailed explanation. He protested, "Even a five-year-old child can understand that!" "Yes," replied the wise sage, "but even an eighty-year-old man cannot do it."

Your biggest obstacle to self-realization is you. As it says in the *Bhagavad Gita*, "The mind is restless and hard to control, but it can be trained by constant practice (*abhyasa*) and freedom from desire (*vairagya*). A man who cannot control his mind will find it difficult to attain this divine communion; but the self-controlled man can attain it if he tries hard and directs his energy by the right means."

Pantanjali's *Eight Limbs of Yoga* will help you form the necessary groundwork to get on the right track, but you must decide to confront your problems at their roots. Reading and intellectualizing is not enough. If you want to reap the full benefits of the yoga experience, implement the Eight Limbs into every aspect of your life. You must live it, breathe it, and engage this planet and its inhabitants with the lessons below.

The first and second limbs, ***Yama*** and ***Niyama***, form your foundation. Here, awareness and realization is established. *Yama* and *Niyama* lay the footing for everything to come. A serious student should be mindful of every limb, as each of these limbs need constant reflection. As you commit yourself to their study and practice, your depth of understanding for each limb will get deeper over time. In our world that perpetuates instant gratification, many people will take shortcuts and go straight to the yoga poses. Others will go straight to meditation and neglect physical health. I highly recommend starting with Pantanjali's first two limbs. Your practice will be at its deepest and most fulfilling if the first two limbs are practiced at a high level. If the first two limbs are not practiced at a proficient level, the rest of the limbs will be performed at a more superficial and less effective level.

FIRST LIMB **Yama** (Self-Restraint)

The focus of the first limb is on being an ethical and moral person, and on improving your relationship with the outer world. These values are as important today as they were centuries ago. The *Yamas*, as they are referred to, are not meant to be a moral straitjacket, but instead are meant to help develop a greater awareness of one's place in the world. It is not a coincidence that this is the first limb of the practice. When taking steps to transform our inner world, our outer world becomes a total reflection of this effort. There are five *Yamas*:

1. *Ahimsa:* **Non-violence**
   Replace harmful thoughts, speech, and actions with that of loving kindness toward yourself and others.
2. *Satya:* **Truth to be expressed in thought, word, and action**
   Be honest in your thoughts, words, and actions toward yourself and others.
3. *Asteya:* **Non-stealing and non-covetousness**
   Curb desires for things that are not your own. Share the beauty of your thoughts, speech, actions, and material belongings to uplift others instead of stealing and hoarding them for yourself.
4. *Brahmacharya:* **Abstinence from sexual intercourse when not married, practicing monogamy and not having sexual thoughts about another person who is not your spouse**
   It is believed that a life built on celibacy and spiritual studies done by free will increases energy and zest for life. Celibacy may sound like an unrealistic goal today, but it may help to remember that brahmacharya is also about monogamy. When brahmacharya is fully realized in marriage, the sex lives of both partners improve because the level of trust and devotion deepens their connection. It is important that the sexual activity is an expression based on the highest level of mutual respect, love, selflessness, and wisdom.
5. *Aparigraha:* **Non-possessiveness or non-greediness**
   Replace the habit of hoarding with sharing. Do not take without giving back. If you want something, work for it. This builds appreciation for what you have. This will help minimize the insatiable desire to constantly consume. An appetite that is not wisely disciplined leads to personal ill health, financial debt or poor credit, and destruction of the planet's natural resources. The Greek god Apollo's motto, "Nothing in excess. All things in moderation," is a great way to describe *aparigraha*.

SECOND LIMB **Niyama** (Self-Purification by Self-Restraint and Discipline)

The second limb helps refine your spiritual path. Discipline and self-restraint lead to a more orderly and productive life. From the perspective of ancient yoga texts, life is extremely short and we need to make the most of it while we can. This limb gives us guidance. There are five Niyamas:

1. *Shaucha:* **Purity of body and mind**
   When you develop *shaucha* (cleanliness), unwholesome thoughts that lead to foul speech and a sick body are cleared. Purity starts with your mind. Speech and action follow. So, the second limb directs you to make a habit of consuming both food and mental stimuli that support well-being for yourself and the environment (humanity and the planet). This will allow destructive habits (hatred, greed, and delusion) to dissolve.
2. *Santosha:* **Contentment with what one has**
   When you achieve *santosha* (contentment), bonds to the material world are broken and authentic peace and happiness are established within. A lack of contentment is often based on a distorted perception of what one has versus what others have. You advance on the path to self-realization when you can be content with your lot, whether you sit on a throne of dirt or gold.
3. *Tapas:* **Self-discipline, sometimes associated with austerity, and being able to conquer the body and mind through mental control**
   Tapas literally means “heat” or “glow.” This refers to a burning desire to accomplish one’s goal despite what obstacles may appear. The commitment to achieving a goal, no matter how challenging it becomes, builds character. However, note that the highest level of *tapas* is to complete one’s goal without a selfish motivation. When *tapas* is attained, laziness is overcome and willpower is developed for future use.
4. *Svadhyaya:* **Self-study that leads to introspection and a greater awakening of the soul and the source of creation; traditionally studied through Vedic scriptures**
   Svadhyaya (self-study) leads to a greater awakening of your true potential, the root of one’s place in this world and how to live in harmony with the Earth and all its inhabitants.
5. *Ishvara pranidhana:* **The surrender to God**
   When you accept that all things come from a higher power, pride and egocentric behavior are turned into humility and devotion. This strengthens your practice of all the limbs leading up to *samadhi* (the eighth limb).

Asana and *Pranayama* are the third and fourth limbs, and they relate to health and longevity, which allow us more time to achieve the ultimate goal of yoga, Self-Realization or Enlightenment. The third and fourth limbs are important, as they prepare the body for meditation, which will be the key to calming your mind and discovering your true potential.

THIRD LIMB **Asana** (Seat or Posture)

Here is a question: If Gandhi is one of the greatest yogis of our time, does that mean he can touch his toes or bring his foot behind his head? The answer is that it doesn’t matter. Gandhi’s ability to perform the *asanas* had very little to do with what he contributed to the world as a great yogi. The same applies to you. The practice of *asanas* is as much about training the mind as it is the body. How you approach your *asana* practice is often a reflection of how you approach life. Do you keep a sense of peace and calm when a challenge presents itself? Do you break down the impossible into smaller tasks, making the whole possible through commitment to and reflection on each of the parts? Do you overcome self-perceived limitations on your own or do you accept support from others?

Your practice of yoga poses should be characterized by two components: steadiness (*sthira*) and ease (*sukha*). Concentrating on the sound of your breath (*ujjayi*, the most commonly practiced breathing technique in yoga, see here), can provide the steadiness. If you lose your breath, it is most likely because you are pushing too hard; ease off the pose and let the pose cater to the breath.

There is no such thing as a perfect pose; let the poses come like the steps of a dance. Just like in dance, when we focus too much on the mechanics, we let go of the ability to enjoy the music. While the mechanics of alignment are important to prevent injury, never forget the final goal. Feel the music of life flow through you as you do each pose and your body will learn the moves naturally. There are more than enough postures to keep you busy for the rest of your life, so allow yourself to let go of ambition and enjoy the journey. Incorporating a combination of forward bends, backbends, twists, and inversions in your yoga session is optimal for health.

Remember, too, that *asanas* help prepare the mind and body for meditation, relieving tension and protecting the body from disturbances by purifying the nervous system.

FOURTH LIMB **Pranayama** (Control of Breath)

The English word “spirit” comes from the Latin *spiritus*, meaning “breath.”

The breath and the mind are interconnected. Deep, rhythmic, and fluid breathing will energize yet calm the mind and body. Rapid, irregular, and strained breathing produces a chaotic and disturbed mind. A calm mind will give you the mental space to make better decisions and a life in which you take control instead of feeling like a victim of circumstances.

Breathing properly is fundamental to our very existence. Your brain feeds on oxygenated blood, which is supplied with every inhalation. If you are unable to draw oxygen into your body, you will become brain dead after a few minutes. On the other hand, proper exhaling helps expel carbon dioxide. If your ability to exhale were impaired, you would most likely die due to the toxic buildup of carbon dioxide and poison. Stress tends to negatively affect breathing patterns, which contributes to a chain of effects that cause wear and tear on both your body’s nervous and immune systems. In fact, 90 percent of illness is stress-related and, for this reason, attention to breathing properly is, indeed, a matter of life and death.

FIFTH LIMB **Pratyahara** (Sense Withdrawal)

Our perception of reality is predominantly influenced by our sensory experience—what we see, feel, hear, touch, and taste. *Pratyahara* refers to the withdrawal of the senses from external objects and our modern-day need for constant gratification from sensory stimuli. Our minds are constantly being pulled outward to evaluate all the information the senses bring in. Evaluation involves categorizing what has been perceived; often, we hold on to what we believe is desirable, push away what we believe is undesirable, and ignore what we believe to be neutral. *Pratyahara* gives our minds a moment to rest and teaches us to be free of the grasping and clinging to the things we enjoy and avoiding the undesirable.

When you throw a pebble into a pond, your reflection becomes distorted by the resulting ripples. Your mind works in very much the same way: Every thought creates a ripple that distorts the ability to see your true self clearly. Constantly disrupted by these ripples, you begin to believe that the distorted reflection is who you really are. Practicing *pratyahar*a calms the mind, allowing you to see yourself clearly.

SIXTH LIMB **Dharana** (One-Pointed Concentration)

Asanas, *pranayamas*, and *pratyahara* help prepare us for meditation.

When the mind moves from experiencing random scattered thoughts to single one-pointed concentration, it can then find complete absorption in the present moment. By practicing one-pointed concentration, we clear the mind of all distracting thoughts. This can be achieved by focusing on your breath, counting, reciting mantras, or observing a candle flame or an image. Because we are constantly entangled in reliving past memories or living in anticipation of what is to come, it is very seldom that we live in the present moment. It is even less common to be mindful of the present moment with a calm and focused mind. However, this is crucial when trying to achieve self-realization. The power is in the now!

SEVENTH LIMB **Dhyana** (Meditation)

Just as there are many different types of yoga poses, there are many ways of meditating. Meditation is a form of inner contemplation that allows you to access a state of mind that has transcended the ego. This is a state of pure awareness of the present moment that is free of judgment. All meditation leads to a state of full awareness that does not discriminate or categorize things in a dualistic manner, which is to say the perception of what is good versus what is bad, beautiful versus ugly, pleasant versus unpleasant, etc. When we examine reasons behind such judgments, we find many of these beliefs are based on learned behavior, may vary from one culture to another, and have no fixed or concrete reality. With consistent reflection and an open mind, we can correct our biased perceptions. You will develop that part of you called “the Observer.” Once grounded in a regular seated mediation, it is important to take it into a moving mediation throughout your daily life.

EIGHTH LIMB **Samadhi** (Total Absorption)

Samadhi occurs when the analytical mind becomes absent and at one with the object of meditation. The object of meditation can be whatever you are focusing on in your meditation that is used to achieve one-pointed concentration. The word *om*, a deity, or a candle flame are all examples of objects of meditation. Total absorption involves the feeling of oneness with all creation, dissolving all lines between the act of meditation and the object being meditated upon.

It is the absorption in the present moment (*amanaska*) where dualistic thinking is transcended. Many are mistaken in believing *samadhi* is the final goal of yoga. It is but a temporary state of mind that we enter based on the conditions that we have nurtured to support it.

It’s useful to remember that every moment in your life gives you an opportunity to practice the eight limbs. Learn at your own pace, but stay focused, be consistent, and enjoy the journey!

THE *UJJAYI* BREATH

Breath is an essential staple of every yoga practice. By focusing on the breath, a yogi is able to stay in the present moment. Since the breath is neutral, a yogi neither seeks to avoid it, nor is she eager to chase it. Proper and continuous breath helps clear the mind of distracting thoughts and to remain in the present moment with one-pointed concentration.

Deep, conscious breathing also slows down the heartbeat and activates the parasympathetic nervous response, which soothes the nervous system, and allows the muscles to relax into the stretches and stay strong in the strength-based yoga poses. Maintaining deep, fluid breathing will help transform your yoga practice into a moving meditation.

Ujjayi pranayama is one of the most commonly used breathing techniques in every yoga practice. *Ujjayi* means “victorious” in Sanskrit. When practicing this breathing technique, a yogi creates what can be described as the “ocean sound” in the back of the throat by gently squeezing the glottis (the opening between the vocal cords in the throat). You’re aiming to create a breathy and whispery sound at the back of the throat. The sound is often compared to the sound Darth Vader makes when breathing! In order to practice *ujjayi* breath, first sit in a comfortable position. Inhale through the nose and, as you exhale, imagine you are trying to fog up a glass. Try to inhale with the same sound. Once you feel like you have an understanding of the *ujjayi* breath technique, seal your lips in order to prevent the throat from drying out. (A dry throat usually leads to coughing and drinking water, which distracts you from your yoga practice.) Inhale and exhale through the nose, while still maintaining the sensation in the throat of trying to fog up a glass.

Using breath to facilitate poses:

- Keep your breath deep and rhythmic.
- If your breathing becomes restricted or choppy, there is a good chance that you are pushing yourself unnecessarily. In this case, you should ease off the pose and return to a place that promotes smooth and fluid breathing.
- The body lifts up and lengthens on the inhale. For example, lift up from a forward bend into a mountain pose on the inhale, using the inhale to lengthen the limbs of the body and the spine. Use the exhales to go deeper into the pose.

THE *BANDHAS*

Bandha means “lock.” *Bandhas* were traditionally believed to regulate the flow of life energy (*prana*) throughout the body. In contemporary yoga practice, *bandhas* serve a more practical purpose. They are contractions, or “body locks,” that you can implement to help correct your posture or aid you in proper alignment.

There are three major *bandhas*: *mula bandha*, *uddiyana bandha,* and *jalandhara bandha*. The combination of all three *bandhas* is called *maha bandha*, or “the great lock.”

Mula bandha refers to the triggering of the perineum muscle that is located between the genitals and the anus. *Mula* means “root,” therefore *mula bandha* translates as “root lock.” When this *bandha* is engaged, you will feel a slight pull on the inside of the thighs, similar to what you feel when trying to stop the flow of urine.

Uddiyana bandha means “flying/moving up.” To engage this *bandha*, place three fingers below the belly button and pull your lower abdominal muscles slightly in and up. This will cause your pelvis to tilt forward slightly with an upward action, protecting the lower back and strengthening the lower abdominals.

Mula and *uddiyana bandhas* should be engaged throughout the yoga practice. Together they help correct the posture and create proper alignment, which will reduce the chance of injury.

Jalandara bandha is a chin lock. To practice this lock, bring the chin toward the clavicle bone while keeping your spine upright and moving your shoulder blades down the back. This *bandha* is rarely used, but can be found when engaged in *Dandasana*, Staff Pose.

THE *DRISHTIS*

Drishtis are the meditation gazing points to focus on while performing the poses. They are designed to aid with proper alignment, as well as to strengthen the focus on the present moment. While practicing we tend to look around, compare ourselves to others in the room, or look at the clock. This takes away from the focus on the internal aspects of the practice. *Drishtis* are meant help you look inward.

They are as follows:

1. *Nasagrai* or *Nasagre* (nose)
2. *Bhrumadhye* or *Ajna Chakra* (third eye, between the eyebrows)
3. *Nabhi, Nabhicakre,* or *Nabi Chakra* (belly button)
4. *Hastagrai* or *Hastagre* (hands)
5. *Padayoragrai* or *Padayoragre* (toes/feet)
6. *Parshva Drishti* (to the right)
7. *Parshva Drishti* (to the left)
8. *Angushtamadhye* or *Angushta Ma Dyai* (thumbs)
9. *Urdhva* or *Antara Drishti* (up to the sky)

Drishtis can be complicated to grasp at first. However, there are general guidelines for the gaze. It comes down to letting your eyes follow the direction of the stretch. For example, in backbends we look at our third eye in order to let the head roll back and deepen the backbend. Similarly, in seated forward bends, such as *Paschimottanasana* (Western Intense Stretch Pose), we gaze at the toes to lengthen the spine. The purpose of *drishtis* is not for you to become cross-eyed; they are a way to softly focus without intensely staring.

HOW TO APPROACH YOGA POSES

Yoga poses provide much more than a physical workout. Performing them builds character. Facing your fears and challenges that extend beyond your comfort zone with a sense of peace, calmness, and psychological equanimity will help you overcome your self-perceived limitations.

I find it helpful to think of every yoga pose as a prayer you do with your body. While you perform the pose, focus on what is good in your life with a feeling of gratitude. Being at one with your mind and body in this place of grace helps you transcend the ego, which helps you get closer to the ultimate goal of yoga, which is enlightenment.

Universal Alignment Cues

- Engage *mula* and *uddhiyana bandhas*.
- Engage the *ujjayi* breath and maintain deep conscious breathing all throughout the yoga session. If your breath is suffering, then back off a bit.
- Keep the chest open and shoulder blades down the back.
- Lengthen the body and limbs on the inhale, deepen the pose on the exhale.
- Avoid jerky and uncontrolled movements during the poses that are based on flexibility.
- Square your hips.
- Don’t let your knee go past your ankle when doing any kind of lunge.
- In Plank and the majority of arm balances, shoulders should be just over the fingertips.
- Even if you are intermediate or advanced, start with the beginner modifications in order to ensure proper form and get the blood into the targeted muscle as a warm-up.

Universal Flexibility Cues

- Hold flexibility for a minimum of 30 to 90 seconds.
- Stretch to the edge of comfort and stimulation without straining.
- Don’t overstretch to the point of pain; your muscles will tighten up in order to protect themselves and your flexibility will decrease.

RELAXATION AND MEDITATION

Relaxation

Corpse Pose (Shavasana)

(Shava in Sanskrit means "corpse")

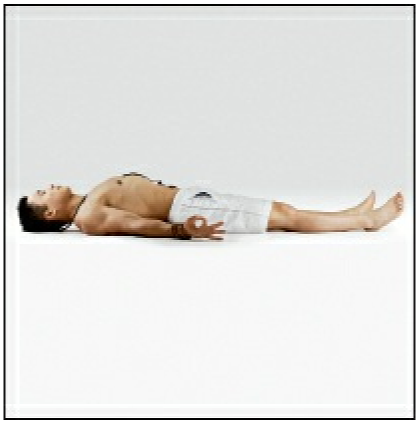

After an intensive *asana* practice, the final resting yoga pose, Corpse Pose or *Shavasana,* helps deepen the connection between our physical body and mind and helps prepare both for meditation. *Shavasana* can be thought of as an awakening, giving us the time to contemplate the question "If I died today, would I be fully satisfied and content with what I accomplished in this lifetime?" Have you lived up to your full potential? Have you fully acknowledged the people in your life that are of great importance to you? Would you be able to pass on with no regrets?

1. Lie down on your back, shoulder blades tucked in, legs apart. Relax your arms and let your palms face the ceiling. Let your fingers naturally curl. Relax your body into a neutral, comfortable position.
2. Close your eyes. Let your jaw naturally separate as you relax your whole body as if it were sinking down to the floor. Release all tension from your body.
3. Without letting the mind wander, concentrate on your breathing to reach a deep state of conscious relaxation, both physically and mentally.

Meditation

Easy Pose (Sukhasana), Seated Lotus Pose (Padmasana)

1. Start in Staff Pose (*Dandasana*), both legs extended in front of you. Grab onto your sitting bones and pull back the flesh to lengthen your legs and spine.
2. Sit in a cross-legged position that is comfortable. You can sit on a chair, in Easy Pose (*Sukhasana*), in Half Lotus (*Ardha Padmasana*), or Full Lotus (*Padmasana*). Press your sitting bones into the ground as you extend up through the spine. Lift the crown of your head up to the sky. Place one hand on top of the other and have the thumbs lightly touch. Gently close your eyes. Perform the 1:4:2 Healing Breath Zen Meditation as follows:

   Inhale for a count of 4 and feel your lower abdomen expand as you push it out. Hold and retain the air in your lungs for the count of 16. Exhale for a count of 8 as you squeeze the belly button back to the spine. Visualize the numbers as you count throughout the set. This will help develop what is called "one-pointed concentration." Your breathing must be so fine that it would not ruffle a feather. Each count should be one second long.

   After 10 breath cycles are performed, the exercise is finished. In an authentic yoga practice, the ego may not be permitted to intrude into the process. Don't "perform" the exercises as if you had an appreciative audience. This is your personal journey. Explore and express yourself while you gain the wonderful physical and mental benefits. This breathing technique is called Abdominal Breathing and should be maintained during the entire yoga practice.

A BRIEF SUMMARY OF *CHAKRAS*

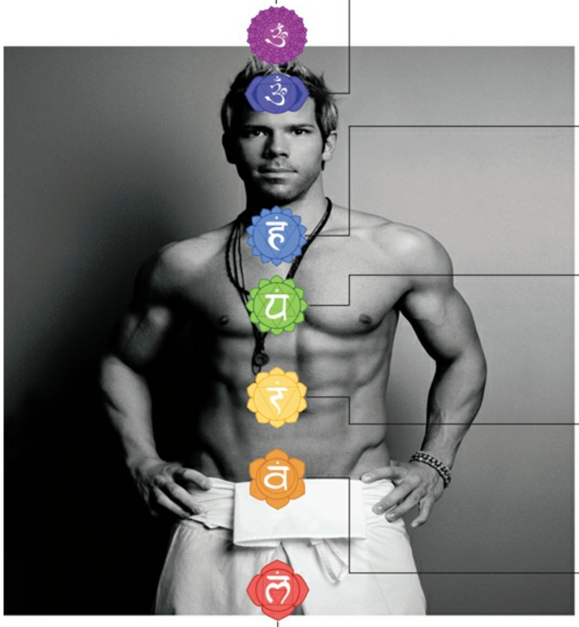

7th	**Crown Chakra**/Thousand Petal or Spoked Wheel/*Sahasrara Chakra* **Location:** just above the crown of the head **Color:** encompasses all colors **Mantra Seed Syllable:** encompasses all sounds	**Element:** the ultimate reality of all truth **Number of Petals:** 1000 (symbolic for unlimited) **Focus:** detachment from ego and illusory nature of the material world, attaining the goal of yoga (self-realization).
6th	**Third Eye Chakra/**Rule or Command Wheel/*Ajna Chakra* **Location:** in the middle of the forehead, between the eyebrows **Color:** indigo **Mantra Seed Syllable:** *om*	**Element:** the universal mind **Number of Petals:** 12 **Focus:** intuition, decision-making, and the surrendering of egocentric intellect in favor of attaining nondualistic wisdom
5th	**Throat Chakra**/Especially Pure Wheel/*Vishuddha Chakra* **Location:** throat **Color:** turquoise or blue **Mantra Seed Syllable:** *ham*	**Element:** ether/space **Number of Petals:** 16 **Focus:** self-expression and communication
4th	**Heart Chakra**/Wheel of the Unstruck or Singular Sound/*Anahata Chakra* **Location:** within the center of the chest **Color:** green **Mantra Seed Syllable:** *yam*	**Element:** air **Number of Petals:** 12 **Focus:** peace, love, and empathy
3rd	**Solar Plexus or Navel Chakra/**Wheel of the Jewel City/*Manipura Chakra* **Location:** solar plexus **Color:** yellow **Mantra Seed Syllable:** *ram*	**Element:** fire **Number of Petals:** 10 **Focus:** power, will, and self-esteem
2nd	**Sacral Chakra/**One's Own Self-Based Wheel/*Svadhisthana Chakra* **Location:** sacrum **Color:** orange **Mantra Seed Syllable:** *vam*	**Element:** water **Number of Petals:** 6 **Focus:** emotions, desires, and creativity
1st	**Root Chakra or Root Wheel/***Muladhara Chakra* **Location:** base of the spine **Color:** crimson **Mantra Seed Syllable:** *lam*	**Element:** earth **Number of Petals:** 4 **Focus:** physical survival, self-preservation, and security

In Sanskrit the word *chakra* can be translated into "wheel" or "turning." In the yogic interpretation, the *chakras* are based on the concept of a vortex and are visually portrayed as a lotus flower.

According to various Eastern yogic spiritual practices such as Hinduism and Tantric Buddhism, *chakras* are described as wheels or rings of energy found in the subtle (non-physical) body; the culmination of the mind, intelligence, and ego, which influences the gross physical body. Within this subtle body there are energy channels called *nadis* that carry the life force or vital energy (*prana*). The main *nadi* that runs through the *chakras* is called the *sushumna* (*brahma*) *nadi*. The *sushumna* joins two other important *nadis* (*ida* and *pingala*) together at the first and seventh *chakras*. The diameter of a singular *nadi* is believed to be no thicker than a thousandth of a hair's width and is located along the spine.

There are various opinions on how many *chakras* there are, but it is generally agreed that *chakras* spin in a "wheel-like" motion to draw in vital energy that creates a balance between the spiritual and physical body.

The earliest known recording of *chakras* dates back to the ancient Vedas (1700 BCE). The most popular *chakra* model used today is based on two Indian texts: *Shat-Cakra-Nirupana*, written by a Bengali yogi named Purnananda Swami in 1577, and the *Padaka-Panchaka*, written in the 10th century.

Chakras are activated in the following ways:
By stretching open the area where the *chakra* is located. For example, the throat *chakra* can be activated in Camel Pose (*Ushtrasana*). The head is rolled back so there is a stretch in the front of the throat.

By applying physical pressure on the area where the *chakra* is located. Throat *chakra*, for example, can be activated in Staff Pose (*Dandasana*). The Chin Lock (*Jaladhara Bandha*) is engaged by bringing the chin to the clavicle bone, therefore applying pressure to the throat area.

By combining the two methods above, throat *chakra* is activated in full version of Inverted Locust Pose (*Viparita Shalabhasana*). The head is rolling back, creating a stretch in the front of the throat. At the same time, the throat area is pressed to the floor, so there is physical pressure on that area.

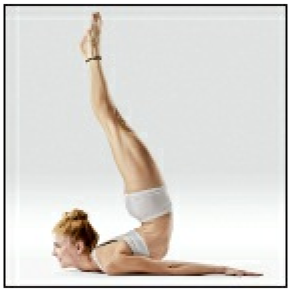

BENEFITS AND CAUTIONS FOR 8 CONDITIONS

CONDITION

1. Headache & Migraine

Many headaches are caused by tension and stress. In yoga we breathe deeply and relax. The yoga practice stretches the tight muscles in the upper body, releases endorphins (a “feel good” hormone) and relaxes the mind. It helps release tension by increasing blood flow to the muscles, making the nervous system less agitated and reducing a chance of a headache or migraine.

Poses that put weight or pressure on the head and neck should be avoided. If you suffer from migraines, avoid poses that dramatically increase blood flow to the head. If your migraines are severe, avoid practicing yoga poses and lie down in a dark room.

BENEFICIAL POSES

Seated forward bends–Ex. Both Hands to Ankle Head to Knee Pose (*Dwi Hasta Kulpa Janu Shirshasana*) here. Seated forward bends release tension in the hamstrings and lower back and help prevent headaches caused by tension in the legs and lower back.

Seated twists–Ex. Half Root Lord of the Fishes Pose (*Ardha Mula Matsyendrasana*) here. Seated twisting poses can help prevent headaches caused by tension in the upper and lower back.

Hand position of the pose dedicated to Garuda–Ex. Hand Position of the Pose Dedicated to Garuda in Child’s Pose (*Hasta Garudasana in Balasana*) here. Any pose with this hand position helps stretch the upper back and back shoulder heads and can help prevent headaches caused by tension in the muscles of the upper back.

Hand position of Cow Face Pose–Ex. Hand Position of Cow Face Pose in Bound Angle Pose (*Hasta Garudasana in Baddha Konasana*) here. Any pose with this hand position helps stretch the triceps, front shoulder heads, and rotator cuffs and can help prevent headaches caused by tension in the arms and shoulder muscles.

POSES TO APPROACH WITH CAUTION

Inversions–Ex. Peacock Feather Pose (*Picha Mayurasana*) here, Leg Position of One-Legged King Pigeon 1 Version B in Headstand 5 (*Pada Eka Pada Raja Kapotasana 1 B in Shirshasana 5*) here, and Leg Position of the Pose Dedicated to Garuda in Hands Bound Supported Whole Bod Pose (*Pada Garudasana in Baddha Hasta Salamba Sarvangasana*), also known as Shoulderstand here. Avoid any intense inversions that require a lot of strength, as they increase the heart rate and blood flow to the head and may trigger headaches or migraines.

Backbends with feet and head on the floor–Ex. Bridge Pose (*Setu Bandhasana*) here. Avoid poses that put pressure on the head and neck, as they may trigger headaches or migraines.

CONDITION

2. Carpal Tunnel

To help prevent or alleviate carpal tunnel syndrome with yoga, you'll need to practice poses that strengthen and stretch the flexor muscles of the forearm, which are the muscles on the palm side of the forearm. Depending on the severity of your condition, you may want to start with poses that bear less weight on the wrist joint.

BENEFICIAL POSES

Poses that strengthen the wrist without straining it–Ex. Staff Pose (*Dandasana*) here. Stretching and gently strengthening the wrist and muscles of the forearm can help prevent or reduce carpal tunnel syndrome.

Poses with hands in the reverse prayer position–Ex. Hidden Lotus Pose (*Gupta Padmasana*) here and Reverse Prayer Mountain Pose (*Viparita Namaskar Tadasana*) here. These poses help stretch out the wrists, forearms, front shoulder heads, chest, and rotator cuffs. Releasing tension from these areas is helpful because muscles in the upper body are interconnected and carpal tunnel can be caused or worsened by a chain reaction of tense muscles.

Poses that have hands in prayer (Anjali Mudra)–Ex. Revolved Prayer Standing Rising Wind Relieving Pose (*Parivritta Namaskar Stiti Utthita Vayu Muktyasana*) here. Stretching the wrists and forearm muscles can promote blood flow and lessen tension in the area.

Poses with bound hands, palms facing out–Ex. Mountain Pose—Raised Bound Hands (*Tadasana Urdhva Baddha Hastasana*) here These poses help stretch the forearm muscles that are strained and tight in most people who have carpal tunnel syndrome.

POSES TO APPROACH WITH CAUTION

Arm balances with both feet off the floor–Ex. Crane Pose (*Bakasana*) here. These poses have the entire body weight resting on the wrists, which puts a great amount of strain on the carpal tunnel and may drastically worsen its symptoms.

Backbends with hands and feet on the floor–Ex. Upward Bow Pose (*Urdhva Dhanurasana*) here and Wild Thing Pose (*Chamatkarasana*) here. These poses are very hard on the wrists.

Backbends with straight arms–Ex. Upward Facing Dog (*Urdhva Mukha Shvanasana*) here. These poses put strain on the wrist joints because most of the upper body weight rests on the hands.

ADDITIONAL NOTES: There are some options to avoid pressure on the wrist while practicin Upward Facing Dog Pose (*Urdhva Mukha Shvanasana*) here or Downward Facing Dog (*Adho Mukha Shvanasana*) here. If you feel that these poses irritate the carpal tunnel, drop the knees to the floor and rotate the hands 45 degrees to the outside to take the pressure off the nerve. You may also experiment with placing props (rolled-up yoga mat, thin book, or slant board) under the heels of the palms to shift the weight to the knuckles and fingers to reduce the compression of the wrist.

CONDITION

3. Asthma

Yoga can help bring awareness to your breathing patterns and release tension from the neck, upper back, chest, and shoulders. Focus on developing full and complete breaths through seated mediation sitting in Easy Pose (*Sukhasana*) here. Since a symptom of asthma is short shallow breaths, developing control of the breath will help the body obtain the oxygen needed and help calm the body, preventing future attacks.

Some yoga poses may be strenuous to the respiratory system and could cause asthma attacks. It's recommended that you pace yourself, gradually raising your body temperature and gradually cooling down. Cold air can cause bronchi to contract and cause an asthma attack. Hot and humid air can cause dehydration and can also cause an asthma attack. Find a room with a comfortable temperature.

BENEFICIAL POSES

Poses on hands and knees–Ex. Going from Tiger Pose (*Vyaghrasana*) here to Unsupported Tiger Pose (*Niralamba Vyaghrasana*) here. Many people who suffer from asthma have tension in the upper back and chest from coughing during asthma attacks. Combining mild backbends with mild forward bends gently stretches the chest, upper back, and the neck, which can help reduce the symptoms of asthma worsened by tension in those areas.

Mild backbends–Ex. Fish Pose (*Matsyasana*) here. Mild backbends help open the chest and front shoulder heads and improve the quality of breathing.

Lion's Pose variations–Ex. Lion Pose Dedicated to an Avatar of Lord Vishnu in Garland Pose (*Narasimhasana in Malasana*) here. Lion's Pose variations can help release tension in the throat, neck, and jaw because you "roar" like a lion in these poses. They also can help push the stale air out of your lungs.

Seated meditation focusing on breathing–Ex. Lotus Pose (*Padmasana*) here. Bringing awareness to your breath and developing control can be useful during an asthma attack. It can also help prevent an attack from happening.

Inversions–Ex. Headstand 5 (*Shirshasana 5*), also known as Tripod Headstand here. Inversions help promote proper movement of the diaphragm during an exhalation. Since the majority of the body is upside down, gravity works with the exhalation, not against it.

POSES TO APPROACH WITH CAUTION

Inversions on the shoulders–Ex. Ear Pressure Pose (*Karnapidasana*) here. Inversions on the shoulders compress the neck and chest, especially when the knees are bent toward the head. This compression restricts your breathing and may cause an asthma attack.

Backbends on the chin and chest–Ex. Inverted Locust Pose (*Viparita Shalabhasana*) here. These poses compress the throat and restrict your breathing, and may cause an asthma attack.

Seated forward bends–Ex. Western Intense Stretch Pose (*Paschimottanasana*) here. These poses compress the lungs and restrict your breathing, and may cause an asthma attack.

Cardio-intense poses–Ex. Crocodile Pose (*Nakrasana*) here. Poses that require a lot of strength and that are demanding on the cardiovascular system can cause shortness and shallowness of breath, and may cause an asthma attack.

Intense backbends–Ex. Little Thunderbolt Pose (*Laghuvajrasana*) here. These poses can be stimulating and cause shortness of breath if you have asthma. It is recommended to start with mild backbends and progress into deeper ones slowly, according to how your body feels.

Arm balances with both feet off the floor–Ex. Uneven Half Repose Pose Dedicated to Ashtavakra (*Vishama Ardha Shayana Ashtavakrasana*) here. These types of poses require a lot of strength and endurance and may cause shortness of breath.

CONDITION

4. Neck Pain

Yoga practice can help prevent and relieve neck pain. The combination of gentle stretches and strengthening movements can open up tight muscles in the body, increasing neck flexibility and rebalancing postural muscles. Simple and slow movements will lubricate the neck and increase its range of movement. Hold each pose for 30 to 90 seconds.

While it's important to strengthen and stretch the neck muscles in order to help prevent neck injury, if you already have a neck concern it's best not to aggravate it. Poses that are most strenuous for the neck are the ones that bear the majority of the body's weight on the head or neck.

BENEFICIAL POSES

Seated neck stretches–Ex. Easy Pose with Neck Stretch (*Sukhasana*) here. Neck pain can be caused by tension in the neck muscles. Stretching these muscles can help prevent or reduce neck pain.

Seated twists–Ex. Pose Dedicated to Bharadvaja 1 (*Bharadvajasana 1*) here. Muscles in the neck are connected to the muscles in the upper back. Seated twists increase the range of motion in the upper back and neck and can help prevent or reduce neck pain caused by tension in these areas.

Poses on hands and knees–Ex. Going from Tiger Pose (*Vyaghrasana*), modification knee to the forehead, here to Tiger Pose (*Vyaghrasana*), also known as Cat Tilt here, top. Rounding the back and then going into a mild backbend can help strengthen the neck muscles and stretch the front of the neck (in Dog Tilt) and the back of the neck (in Cat Tilt).

Hand position of the pose dedicated to Garuda–Ex. Hand Position of the Pose Dedicated to Garuda in Hero Pose (*Hasta Garudasana in Virasana*) here. Any pose with this hand position helps stretch the upper back and back shoulder heads and can help prevent neck pain caused by tension in the muscles of the upper back.

POSES TO APPROACH WITH CAUTION

Full inversions–Ex. Revolved Leg Position of One-Legged King Pigeon 1 Version B in Headstan 5 (*Parivritta Pada Eka Pada Raja Kapotasana 1 B in Shirshasana 5*) here, One-Legged Unsupported Whole Body Pose (*Eka Pada Niralamba Sarvangasana*), also known as Shoulderstand here Full inversions with the head on the floor put strain on the neck since the majority of the body weight tends to rest on the head or neck. If you have a neck injury, it’s best to avoid these.

Backbends with feet and head on the floor–Ex. Inverted Tip Toe Bow Pose (*Viparita Prapada Dhanurasana*), also known as Headstand Bow Pose (*Shirsha Dhanurasana*) here. Backbends with head and feet on the floor require a lot of neck strength and should be avoided if you have a neck concern.

CONDITION

5. High Blood Pressure

If you think you may be at risk or have already been diagnosed with high blood pressure, it is advisable to speak with your health care provider. Practicing yoga may help with blood pressure management since it combines the benefits of meditation, muscle relaxation, and strength training exercise. When practicing yoga poses, ensure that you are able to breathe comfortably and deeply. If you have any difficulty breathing, come out of the pose and rest or perform an easier version of the pose. If difficulty breathing persists, then consult your health care provider immediately.

BENEFICIAL POSES

Seated backbends–Ex. Hands Bound Lotus Pose (*Baddha Hasta Padmasana*) here. Seated backbends gently open the chest and improve the flow of oxygen to the lungs. They release tension in the chest and front shoulder heads often caused by stress and hunching over a computer on a daily basis. This can help lower high blood pressure that is a result of stress.

Seated twists–Ex. Revolved Easy Pose (*Parivritta Sukhasana*) here. Seated twists help release tension from the upper back and detoxify the body. This can help lower high blood pressure that is a result of tension in the upper back.

Supine forward bends–Ex. Reclining Both Hands to the Leg Pose (*Supta Dwi Hasta Padasana*) here. Supine forward bends stretch the hamstrings without increasing blood pressure, as opposed to standing forward bends where the head is below the heart. This can help lower high blood pressure that is a result of muscle tension of the lower back and legs.

POSES TO APPROACH WITH CAUTION

Inversions–Ex. Feet Spread Out Intense Stretch Pose 2 (*Prasarita Padottanasana 2*) here and Crane Pose in Headstand 5 Prep. (*Bakasana in Shirshasana 5 Prep.*) here. Inversions should be avoided if you have high blood pressure that is not controlled. They are very stimulating postures since the head is below the heart, which causes an increased demand for oxygen. These increase blood flow and heart rate, generating pressure to the blood vessels of the brain, and may cause blood pressure to rise dramatically.

Lunges with back knee off the floor–Ex. Revolved Son of Anjani (Lord Hanuman) Lunge Pose with Hands in Prayer (*Parivritta Anjaneyasana Namaskar*) here. These poses can require a lot of lower body strength if you are not used to practicing them. They can raise the heart rate and increase blood pressure.

Arm balances with both feet off the floor–Ex. Leg Position of Cow Face Pose in Pendant Pos (*Pada Gomukhasana in Lolasana*) here. These poses are demanding on the upper body and may increase heart rate, which can increase blood pressure.

Backbends with hands and feet on the floor–Ex. Elevated Both Legs Inverted Staff Pose (*Utthita Dwi Pada Viparita Dandasana*) here. Poses such as these can be demanding on the upper body and can elevate blood pressure by increasing heart rate.

CONDITION

6. Menstruation

During a period, uterine contractions can cause painful cramps in the lower abdomen and lower back. Yoga can help release endorphins to relax the body. Stretch out the lower body and back to help release the pain. Forward bends, inside hip openers, and gentle twists can help relieve the symptoms of menstruation.

Some are of the opinion that inversions can cause engorgement in the blood vessels of the uterus, which may increase blood flow, and should be avoided during a period. On the other hand, B.K.S. Iyengar's book *The Path to Holistic Health* recommends inversions during a period to reduce blood flow. You should listen to your own body and judge accordingly. If you are not feeling strong, engage in a slower-paced yoga pose practice.

BENEFICIAL POSES

Supine hip openers–Ex. Universal All-Encompassing Diamond Pose (*Vishvavajrasana*) here. Poses such as this open up the hips and groin and allow the lumbar spine to rest, which can help relieve the menstrual discomfort.

Low squats–Ex. One Leg Bound Garland Pose (*Eka Pada Baddha Malasana*) here. Practice poses such as this to stretch out the groin, the chest, and front shoulders. Releasing tension from those areas can help relieve menstrual discomfort.

Seated hip openers–Ex. One Legged King Pigeon Pose 1 Prep. (*Eka Pada Raja Kapotasana 1 Prep.*) here. Poses such as this help open up the hip and stretch the lower abdomen, which can help relieve menstrual discomfort due to tension in that area.

Backbends on the knees–Ex. Camel Pose (*Ushtrasana*) here. Poses such as this can help stretch out the lower abdomen and release tension from that region, which can help relieve menstrual discomfort.

Gentle seated twists–Ex. Easy Lord of The Fishes Pose Prep. (*Sukha Matsyendrasana Prep.*) here. Seated twists stimulate internal organs and can help relieve the symptoms of menstruation by gently encouraging the natural blood flow during menstruation.

POSES TO APPROACH WITH CAUTION

Arm balances with hands and feet on the floor–Ex. Four Limbed Staff Pose (*Chaturanga Dandasana*) here. Poses such as this are very demanding on the upper body and core, and may worsen the symptoms of menstruation due to overstraining.

Arm balances with both feet off the floor–Ex. Pose Dedicated to Galava, One-Legged Variation (*Eka Pada Galavasana*) here. These poses demand a lot of strength and may worsen the symptoms of menstruation due to overstraining.

High squats–Ex. Tip Toe Fierce Pose (*Prapada Utkatasana*) here. Poses such as this are demanding on the legs and core, and may worsen the symptoms of menstruation due to overstraining.

Backbends on hands and feet–Ex. Partridge Pose (*Kapinjalasana*) here. Intense backbends put a lot of strain on the body, because they engage the entire body. They may worsen the symptoms of menstruation due to overstraining.

CONDITION

7. Pregnancy

Don’t engage in a vigorous yoga practice with jump-through and jump-back *Vinyasas*. Jumping is dangerous during pregnancy. Avoid hot yoga classes that can dangerously elevate your core temperature or cause dehydration. After giving birth, be mindful when going into deep stretches. The levels of relaxin (the hormone that loosens the muscles and joints to accommodate birth) in the body may still be high, increasing the danger of injury due to overstretching. If you had a C-section, make sure the wound heals properly. Avoid doing any intense twists or backbends, as they may interfere with healing of the wound.

BENEFICIAL POSES

Seated inner hip openers–Ex. Knees Spread Wide Hero Pose (*Prasarita Janu Virasana*) here. Poses such as this stretch out the inner hips without compressing the abdomen.

Wide-legged squats–Ex. Lotus Hand Seal in Upward Hands Pose Dedicated to Goddess Kal (*Padma Mudra Urdhva Hasta Kalyasana*) here. Wide-legged squats strengthen quadriceps (front of the thighs), hamstrings (back of the thighs), and glutes (buttocks) without putting pressure onto the abdomen.

Standing side bends–Ex. Extended Side Angle Pose (*Utthita Parshva Konasana*) here. Standing side bends stretch out the side of the torso and lower back while strengthening the legs without putting pressure onto the abdomen.

Poses on hands and knees–Ex. Tiger Pose (*Vyaghrasana*) here. Poses on hands and knees can be done during pregnancy because they do not compress the abdomen.

Mild backbends on the knees–Ex. Half Camel Pose (*Ardha Ushtrasana*) here. Mild backbends can be practiced during pregnancy since they don’t compress the abdomen.

POSES TO APPROACH WITH CAUTION

Inversions–Ex. Repose Pose (*Shayanasana*) here. Don’t do any inversions where your heart is above your head beyond the first trimester. Turning your body upside down creates pressure on your internal organs and may be damaging to the developing fetus.

Forward bend twists–Ex. Two Hands Revolved Western Intense Stretch Pose (*Dwi Hasta Parivritta Paschimottanasana*) here. Don't do any forward bends or twists that compress the abdomen and squeeze the fetus and placenta.

Core poses–Ex. Revolved Boat Pose (*Parivritta Navasana*) here. Don't do strenuous core poses that compress the abdomen.

Prone poses–Ex. One-Legged Pose Dedicated to Siddhar Konganar (*Eka Pada Konganarasana*) here. Avoid prone poses since they put a lot of pressure on the abdomen.

Supine poses where the back is flat on the floor–Ex. Reclined Leg Position of Cow Face Pos (*Supta Pada Gomukhasana*) here. Supine poses with the back flat on the floor take out the natural curve in the lumbar spine that is present during pregnancy and can compress the fetus and placenta.

Intense backbends–Ex. One-Legged Pigeon Pose (*Eka Pada Kapotasana*) here. Intense backbends create too much stretch in the abdomen and should be avoided during pregnancy.

CONDITION

8. Menopause

Common symptoms of menopause such as hot flashes and mood swings can be alleviated with regular yoga pose practice. Focus on poses that open up the pelvic area as well as mediation to help control stress.

Avoid practicing hot yoga and avoid overexertion. Both can trigger menopause symptoms. It is recommended to avoid vigorous Sun Salutes, as they can increase body temperature and cause hot flashes.

BENEFICIAL POSES

Mild supine backbends–Ex. Bridge Whole Body Pose (*Setu Bandha Sarvangasana*) here. Mild backbends open up the chest and heart area. They can help balance blood pressure and hormonal secretions as well as help relieve mood swings and hot flashes.

Supine hip openers–Ex. Reclined Bound Angle Pose *(Supta Baddha Konasana*) here. These poses open up the chest, heart and pelvic areas. Blood flow is increased into the pelvic area and reproductive organs and that can help balance hormonal functions. These poses can help relieve high blood pressure, headaches, and breathing problems.

Supine thigh openers–Ex. Reclined Hero Pose (*Supta Virasana*) here. Poses such as this can help improve blood circulation in the ovarian region and stimulate the pelvic organs, which can help balance hormonal functions and relieve the symptoms of menopause.

POSES TO APPROACH WITH CAUTION

Backbends on hands and feet–Ex. Tip Toe One Hand Upward Bow Pose (*Prapada Eka Hasta Urdhva Dhanurasana*) here. Poses such as these can be too strenuous on the upper body (arms and shoulders). They can raise the body temperature and cause hot flashes.

Core poses–Ex. Boat Pose (*Navasana*) here. Avoid any vigorous core yoga poses, as it creates too much tension around the abdominal organs and may worsen the menopause symptoms.

Standing twists–Ex. Revolved Side Angle (*Parivritta Parshva konasana*) here. Avoid any intense twists, as it creates too much compression around the internal organs of the torso, which may worsen the menopause symptoms.

Inversions–Ex. Headstand 1 (*Shirshasana 1*) here. Avoid any full inversions, as they increase blood flow to the internal organs of the torso and raise heart rate which may cause hot flashes.

Standing Poses

MOUNTAIN POSE
SIDE BENDS
STANDING BACKBENDS
ONE LEG STANDING BALANCE
STANDING FORWARD BENDS
LEGS WIDE
DOWNWARD DOG
INTENSE SIDE STRETCH
TRIANGLE POSE
WARRIOR AND SIDE ANGLE POSES
LUNGES AND KNEELING POSES
SQUATS

Mountain Pose

Tadasana

(tuh-DAHS-uh-nuh)

Also Known As: Equal Steady Standing, State of Balance (Samasthiti)
Modification: palms rotated forward
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

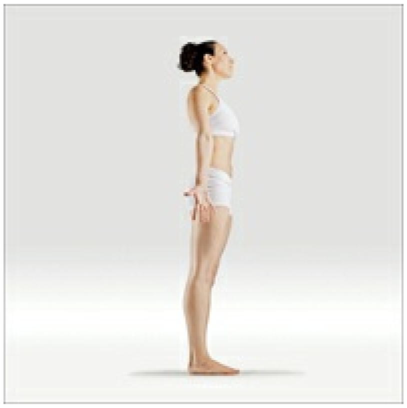

Mountain Pose with Hands in Prayer

Tadasana Namaskar

(tuh-DAHS-uh-nuh nuh-muhs-KAHR)

Modification: hands in Anjali Mudra (Hands in Prayer); feet to the front, toes lifted
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

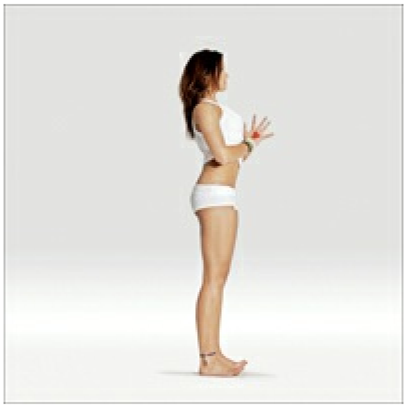

Reverse Prayer Mountain Pose

Viparita Namaskar Tadasana

(vi-puh-REE-tuh nuh-muhs-KAR tuh-DAHS-uh-nuh)

Also Known As: Penguin Pose, Back of the Body Prayer Mountain Pose (Paschima Namaskara Tadasana)

Modification: feet rotated out

Pose Type: standing

Drishti Point: Nasagrai or Nasagre (nose)

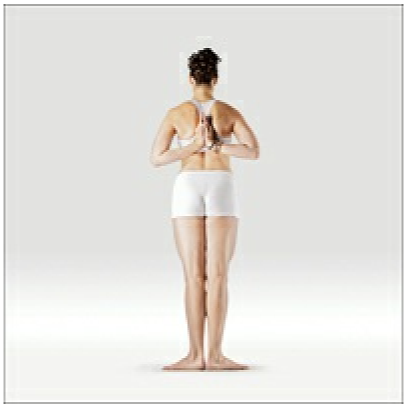

Hands Bound Mountain Pose

Baddha Hasta Tadasana

(BUH-duh HUH-stuh tuh-DAHS-uh-nuh)

Pose Type: standing

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

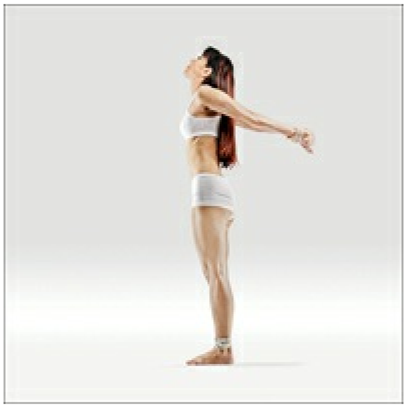

MOUNTAIN POSE: HANDS ABOVE THE HEAD & TANDAVA AND LASYA DANCE MODIFICATIONS

Upward Salute Pose

Urdhva Hastasana

(OORD-vuh huh-STAHS-uh-nuh)

Also Known As: Volcano Pose
Modification: arms shoulder width apart
Pose Type: standing, mild backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

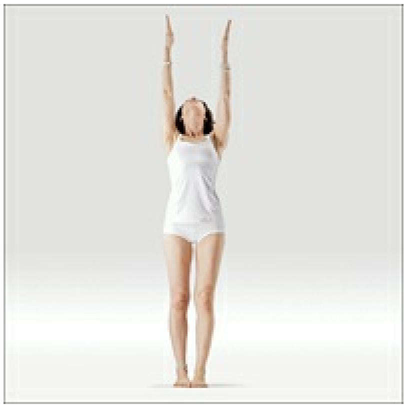

Upward Salute Pose

Urdhva Hastasana

(OORD-vuh huh-STAHS-uh-nuh)

Also Known As: Volcano Pose
Modification: palms pressed together
Pose Type: standing, mild backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

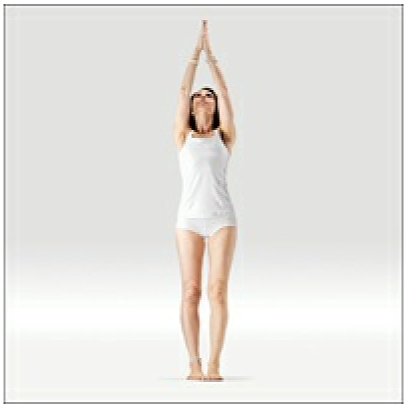

Mountain Pose—Raised Bound Hands

Tadasana Urdhva Baddha Hastasana

(tuh-DAHS-uh-nuh OORD-vuh BUH-duh huh-STAHS-uh-nuh)

Also Known As: Mountain Pose (Parvatasana)
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)
ल य

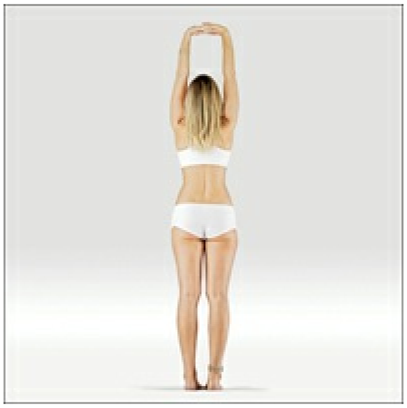

Mountain Pose

Tadasana

(tuh-DAHS-uh-nuh)

Modification: shoulder opener, intense version
Pose Type: standing
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल

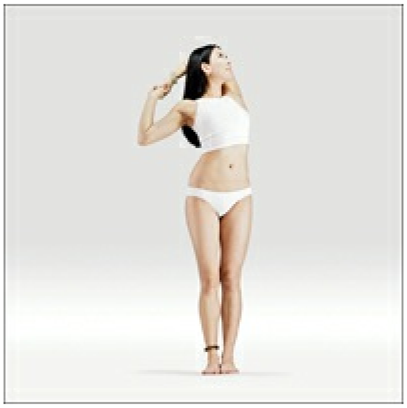

Pose Inspired by Parvati’s Graceful Dance

Lasyasana

(lahs-YAHS-uh-nuh)

Modification: both legs straight; one leg extended to the front, heel up; one arm up over the head reaching to the floor, other arm reaching up to the sky; deep backbend
Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

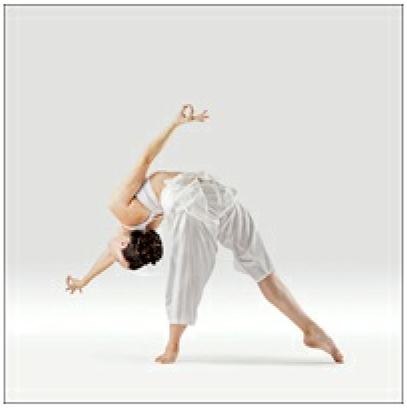

SIDE BENDS: FEET TOGETHER, ARMS UP OVER THE HEAD & TANDAVA AND LASYA DANCE MODIFICATIONS

Standing Crescent Pose

Indudalasana

(in-doo-duh-LAHS-uh-nuh)
Modification: grabbing onto the wrist of the top hand
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल

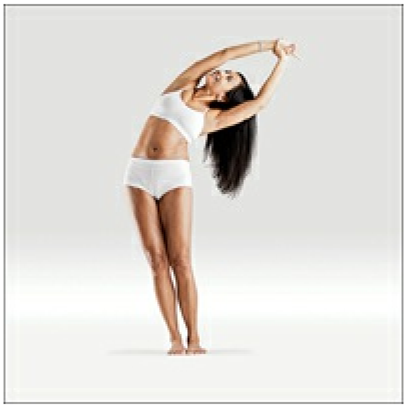

Sideways Mountain Pose—Raised Bound Hands

Parshva Tadasana Urdhva Baddha Hastasana

(PAHRSH-vuh tuh-DAHS-uh-nuh OORD-vuh BUH-duh HUH-STAHS-uh-nuh)
Also Known As: Side Bending Pose (Parshva Bhangi)
Pose Type: standing, side bend

Drishti Point: Nasagrai or Nasagre (nose)
लं

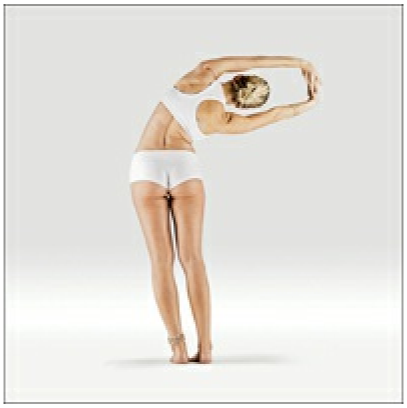

Revolved Upward Mountain Pose—Raised Bound Hands

Parivritta Urdhva Tadasana Urdhva Baddha Hastasana

(puh-ri-VRIT-tuh OORD-vuh tuh-DAHS-uh-nuh OORD-vuh BUH-duh HUH-STAHS-uh-nuh)

Also Known As: Side Bending Pose (Parshva Bhangi)
Pose Type: standing, side bend, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
लं रं यं हं ॐ

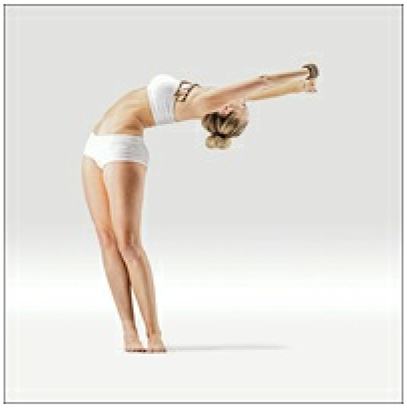

One-Legged Standing Crescent Pose

Eka Pada Indudalasana

(EY-kuh PUH-duh in-doo-LAHS-uh-nuh)

Modification: grabbing onto the wrist
Pose Type: standing one-legged balance, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)
लं वं

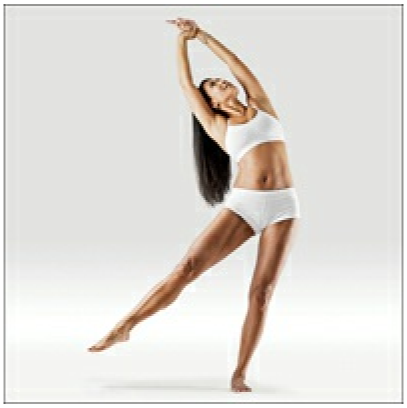

Side Stretch Pose

Parshvasana

(pahrsh-VAH-suh-nuh)

Modification: fingers interlocked, palms pressed together
Pose Type: standing, side bend
Drishti Point: Nasagrai or Nasagre (nose), Urdhva or Antara Drishti (up to the sky)
ल

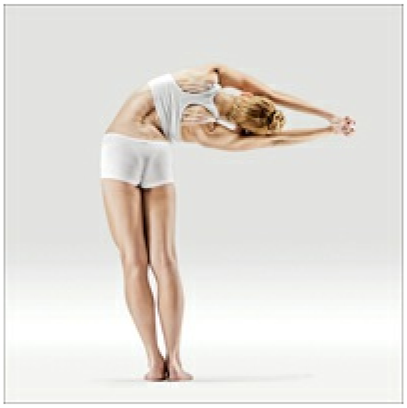

Pose Inspired by Shiva’s Vigorous Cycle of Life Dance

Tandavasana

(tahn-duh-VAHS-suh-nuh)

Modification: heels down, legs crossed, arms up over the head
Pose Type: standing, side bend, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व र य ह ॐ

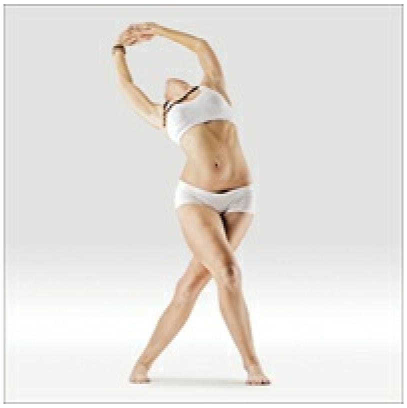

TANDAVA AND LASYA DANCE MODIFICATIONS

Pose Inspired by Parvati’s Graceful Dance

Lasyasana

(lahs-YAHS-uh-nuh)

Modification: legs crossed, knees bent; one hand on the hip, one arm over the head
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

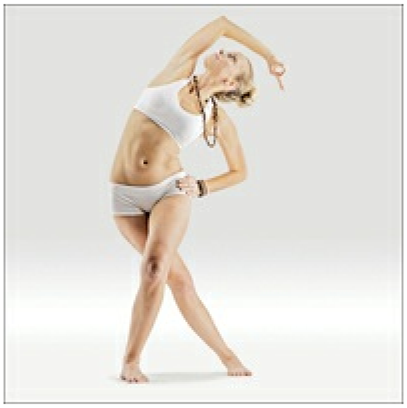

Pose Inspired by Parvati’s Graceful Dance

Lasyasana

(lahs-YAHS-uh-nuh)

Modification: front leg ankle stretch, both knees bent, legs crossed, one arm toward the opposite knee, other arm up over the head
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

SIDE BENDS: ONE ARM UP OVER THE HEAD, OTHER ARM ALONG THE SIDE OF TORSO

One Hand Side Stretch Pose

Eka Hasta Parshvasana

(EY-kuh HUH-stuh pahrsh-VAH-suh-nuh)

Modification: one arm up over the head, other hand sliding down the leg
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

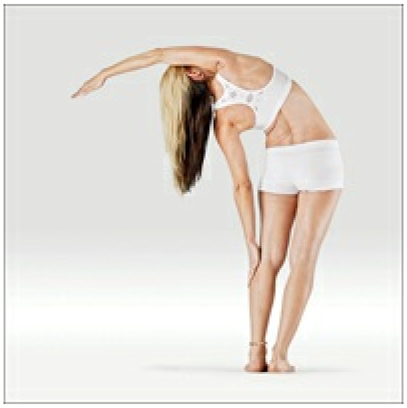

Swaying Palm Tree Pose

Tiryak Tala-Vrikshasana

(TIR-yuhk TAHL-uh-vrik-SHAHS-uh-nuh)

Pose Type: standing, side bend

Drishti Point: Hastagrai or Hastagre (hands)

How to Perform the Pose:

1. Begin by standing in Mountain Pose (*Tadasana*). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Inhale; step your feet slightly wider than shoulder-distance apart with toes facing forward and feet parallel to each other. Expand your chest and hold your arms straight out to the sides, parallel to the floor.
3. Exhale as you side bend to the left, dropping your left hand either to the side of the left thigh or the left shin. (Avoid putting pressure on the knee joint.) Bring your right arm over head and bend it at the elbow.
4. Inhale as you rotate your chest up to the sky; do not collapse it forward. Feel the deep stretch on the right side of your torso as you look toward your right hand (Pose #1).
5. On your next exhale, try to reach your right arm over your head with fingertips pointing toward the floor (Pose #2).
6. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch. Exhale as you release the pose. Inhale as you press strongly into both feet to come up.
7. Exhale, come back to Mountain Pose (*Tadasana*), and repeat on the left side.

Modification: elbow of the top arm bent
1. mild version
2. intense version

tiryak = horizontally, sideways, obliquely, across

tala-vrikshasana = palm tree

1.

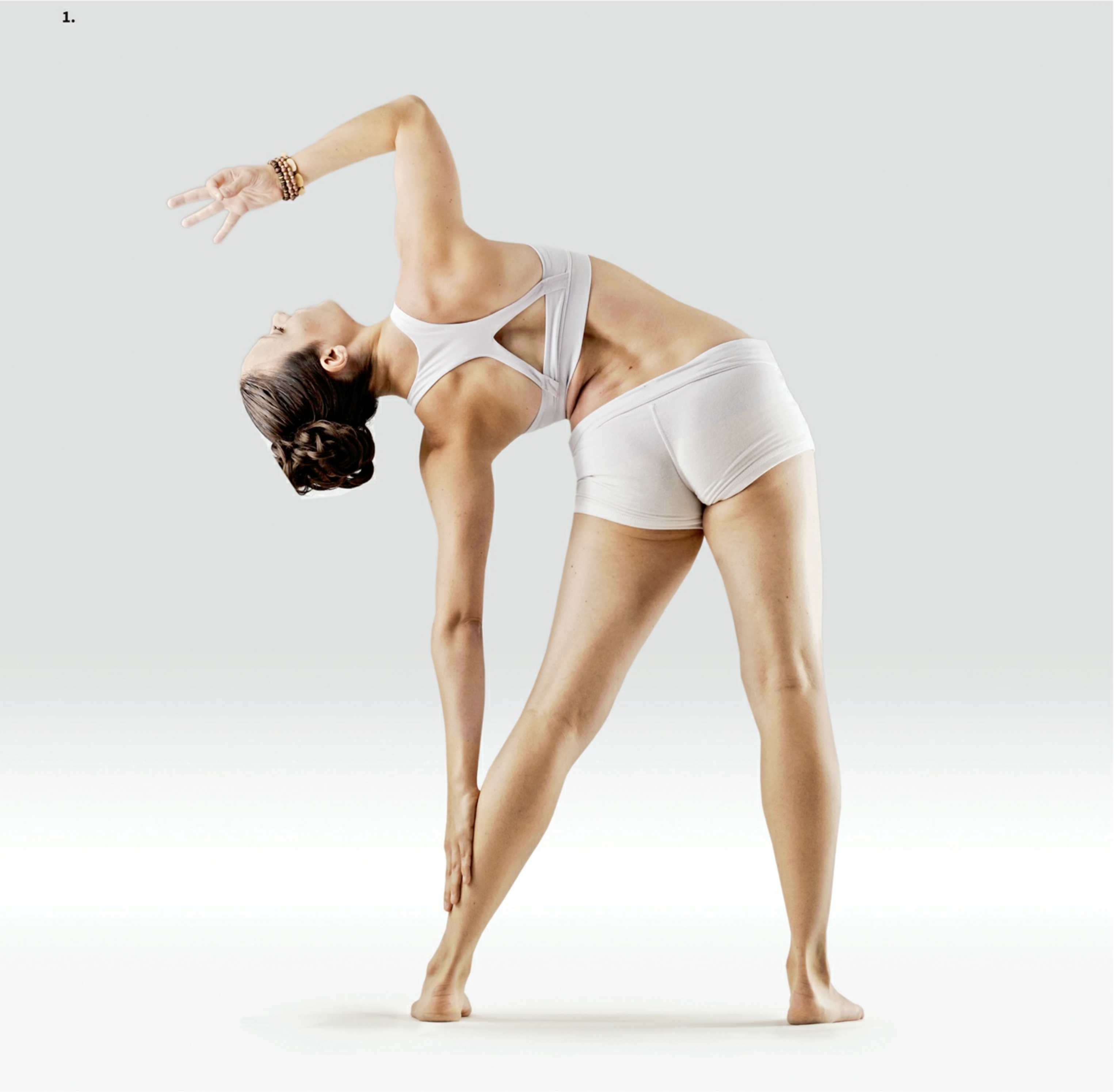

2.

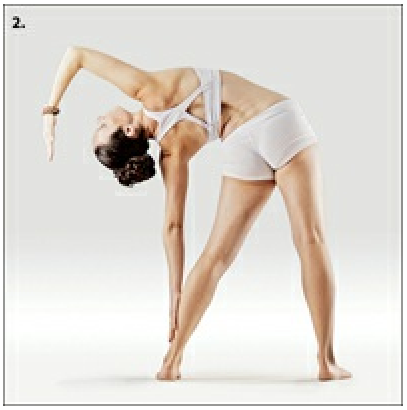

Hands Bound Rising Standing Locust Pose

Baddha Hasta Utthita Stiti Shalabhasana

(BUH-duh HUH-stuh UT-ti-tuh STI-ti shuh-luh-BAHS-uh-nuh)

Also Known As: Baddha Hasta Utthita Nindra Shalabhasana
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

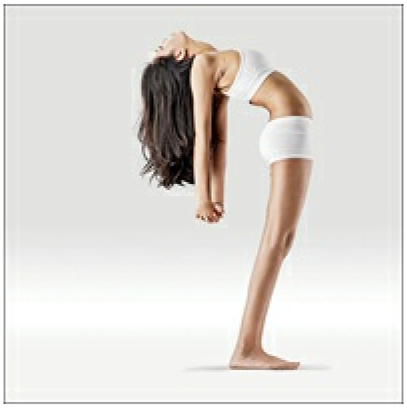

Sun Salutation Pose—Raised Bound Hands

Surya Namaskarasana Urdhva Baddha Hastasana

SOOR-yuh nuh-muhs-kahr-AHS-uh-nuh OORD-vuh BUH-duh huh-STAHS-uh-nuh)

Pose Type: standing, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल र य ह ॐ

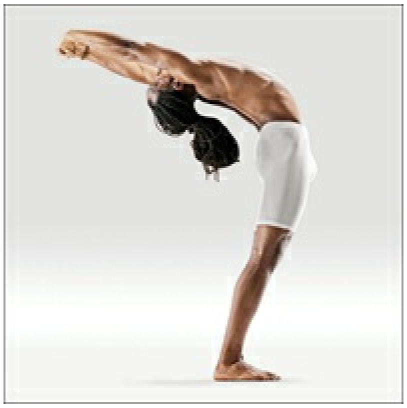

Rising Standing Cobra Pose

Utthita Stiti Bhujangasana

(UT-ti-tuh STI-ti bu-juhng-AHS-uh-nuh)

Also Known As: Utthita Nindra Bhujangasana
Modification: toes down, hands on the front of the knees
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

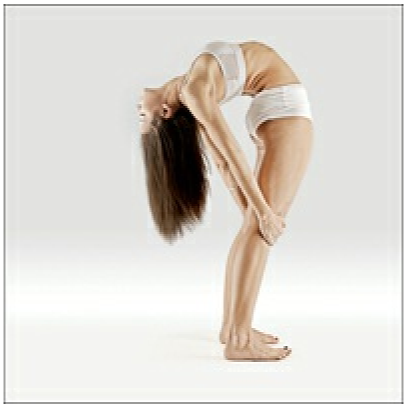

Sun Salutation Pose

Surya Namaskarasana

(SOOR-yuh nuh-muhs-kahr-AHS-uh-nuh)

Modification: palms facing up
Pose Type: standing, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

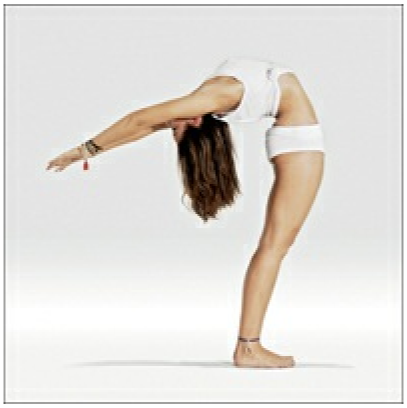

Sun Salutation Pose

Surya Namaskarasana

(SOOR-yuh nuh-muhs-kahr-AHS-uh-nuh-uh-nuh)

Modification: fingers interlocked, pointer fingers out
Pose Type: standing, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल र य ह ॐ

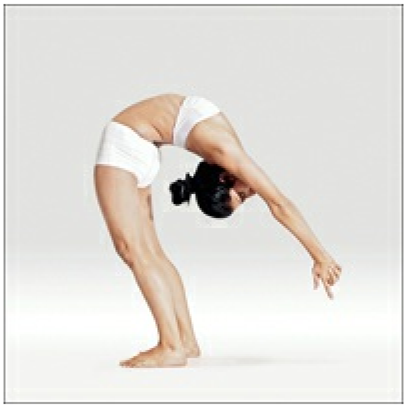

Reverse Facing Intense Stretch Pose

Tiryang Mukhottanasana

(TEER-yuhng mu-ko-tahn-AHS-uh-nuh)
Also Known As: Full Wheel Pose (Purna Chakrasana)
Modification: grabbing onto the shins
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

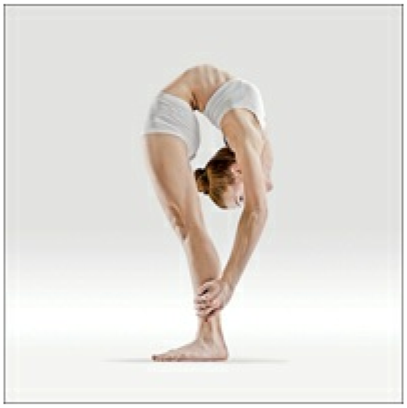

Reverse Facing Intense Stretch Pose

Tiryang Mukhottanasana

(TEER-yuhng mu-ko-tahn-AHS-uh-nuh)
Modification: grabbing onto the front of the knees
Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

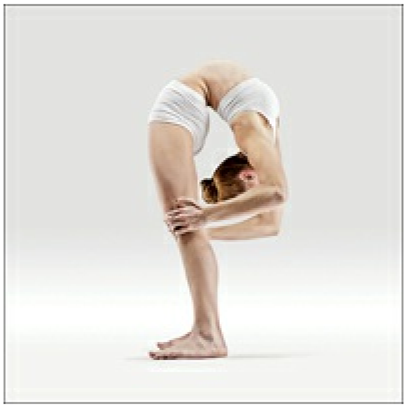

STANDING BACKBENDS: KNEES BENT

Standing Upward Facing Intense Ankle Stretch Bow Pose

Stiti Urdhva Mukgattana Kulpa Dhanurasana

(STI-ti OORD-vuh mu-ko-TAH-nuh KUL-puh duh-nur-AHS-uh-nuh)

Also Known As: Nindra Urdhva Mukgattana Kulpa Dhanurasana
Modification: hands to the heels
Pose Type: standing, backbend, balance
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

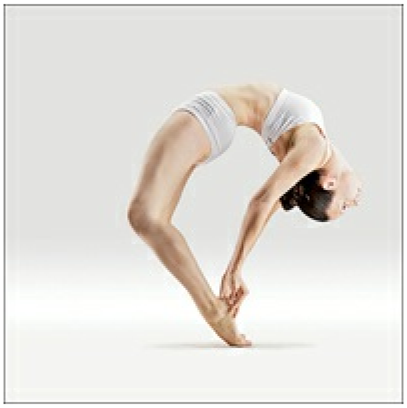

ONE LEG STANDING BALANCE: KNEE BENT—HEEL TO THE SITTING BONES

One Hand to Foot Rising Standing One-Legged Frog Pose

Eka Hasta Pada Utthita Stiti Eka Pada Bhekasana

(EY-kuh HUH-stuh PUH-duh UT-ti-tuh STI-ti EY-kuh PUH-duh bey-KAHS-uh-nuh)

Also Known As: Eka Hasta Pada Utthita Nindra Eka Pada Bhekasana

Modification: grabbing onto the foot, arm crossed in front; foot toward the glutes

Pose Type: standing one-legged balance, binding

Drishti Point: Hastagrai or Hastagre (hands)

🔴

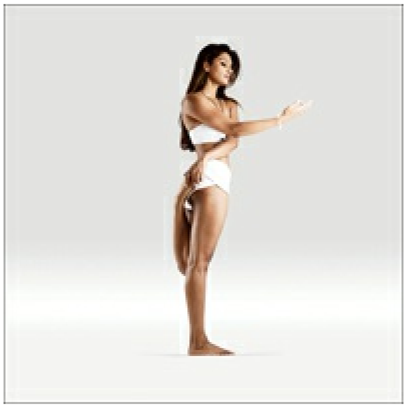

Lord of the Dance Pose

Natarajasana

(NAH-tuh-rahj-AHS-uh-nuh)

Modification: one hand to foot, foot toward the glutes

Pose Type: standing one-legged balance, backbend

Drishti Point: Hastagrai or Hastagre (hands)

🔴 🟢 🔵

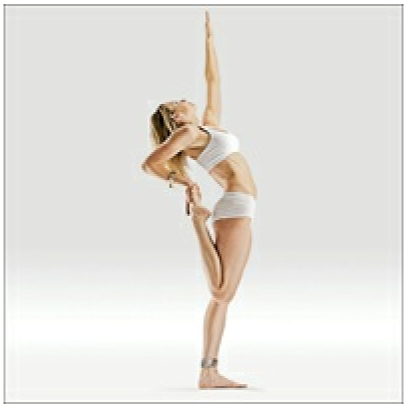

Both Hands to Foot Rising Standing One-Legged Frog Pose

Dwi Hasta Pada Utthita Stiti Eka Pada Bhekasana

(dwi-huh-stuh PUH-duh UT-ti-tuh STI-ti EY-kuh PUH-duh bey-KAHS uh-nuh)

Also Known As: Dwi Hasta Pada Utthita Nindra Eka Pada Bhekasana

Pose Type: standing one-legged balance

Drishti Point: Nasagrai or Nasagre (nose)

🔴 🟢

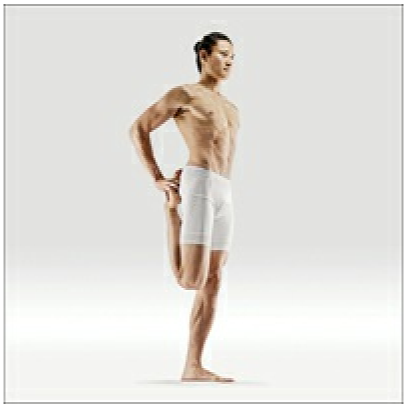

ONE LEG STANDING BALANCE: KNEE BENT—HEEL TO THE SITTING BONES—HALF FORWARD BEND

Bowing with Respect Lord of the Dance Pose

Nantum Natarajasana

(NUHN-tum NAH-tuh-rahj-AHS-uh-nuh)

Modification: half forward bend, both hands to the back foot, foot toward the glutes
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

🔴 🟠 🟢

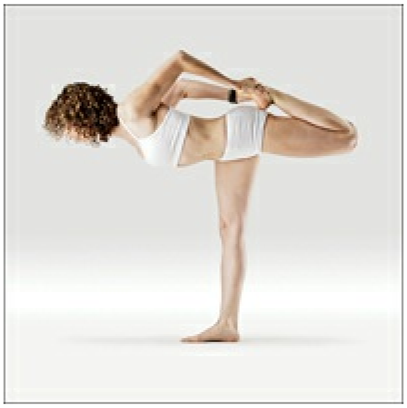

Half Bound Bowing with Respect Lord of the Dance Pose

Ardha Baddha Nantum Natarajasana

(UHR-duh BUH-duh NUHN-tum NAH-tuh-rahj-AHS-uh-nuh)

Modification: grabbing onto the back foot with opposite hand
Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

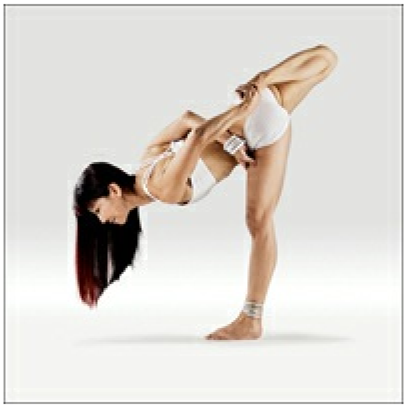

Half Moon Bow Pose Prep.

Ardha Chandrachapasana Prep.

(UHR-duh CHUHN-druh-chahp-AHS-uh-nuh)

Modification: grabbing onto the back foot with the hand on the same side
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

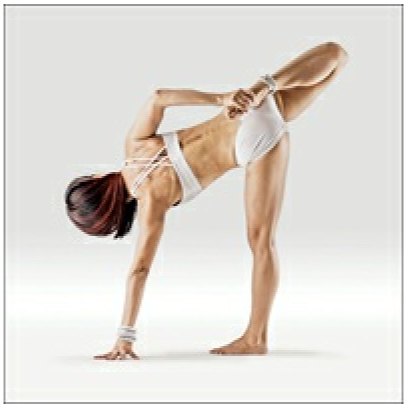

Hand to Foot Revolved Half Moon Bow Pose

Hasta Pada Parivritta Ardha Chandrachapasana

(HUH-stuh PUH-duh puh-ri-VRIT-tuh UHR-duh CHUHN-druh-chahp-AHS-uh-nuh)

Also Known As: Lord of the Dance Pose Modification (Natarajasana)
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

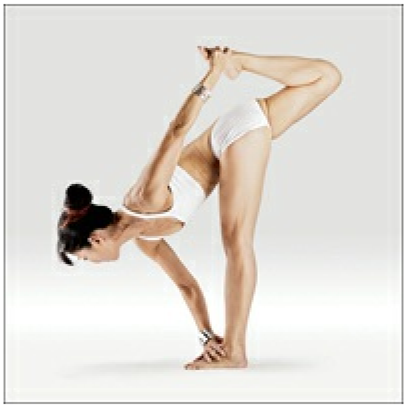

Half Bound Unsupported Bound One-Legged Intense Stretch Pose Prep.

Ardha Baddha Niralamba Baddha Eka Pada Uttanasana Prep.

(UHR-duh BUH-duh nir-ah-LUHM-buh BUH-duh EY-kuh PUH-duh ut-tahn-AHS-uh-nuh)

Also Known As: Lord of the Dance Pose Prep. Modification (Natarajasana Prep.) and Half Bound Unsupported Bound One-Legged Forward Bend Prep.
Modification: one hand to the floor
Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Hastagrai or Hastagre (hands), Nasagrai or Nasagre (nose)
ल व य

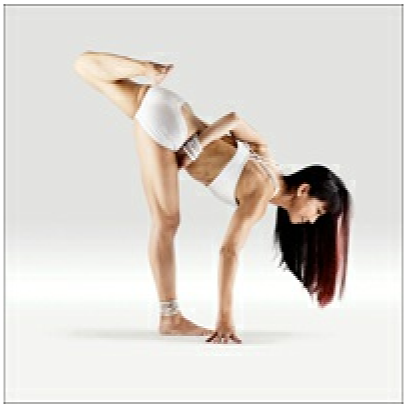

ONE LEG STANDING BALANCE: KNEE BENT—HEEL TO THE SITTING BONES—FORWARD BEND

One Leg Stretched Upward Pose Prep.

Urdhva Prasarita Ekapadasana Prep.

(OORD-vuh pruh-SAH-ri-tuh EY-kuh-puh-DAHS-uh-nuh)

Modification: top leg bent, heel toward the glutes; both palms to the floor
Pose Type: standing one-legged balance, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

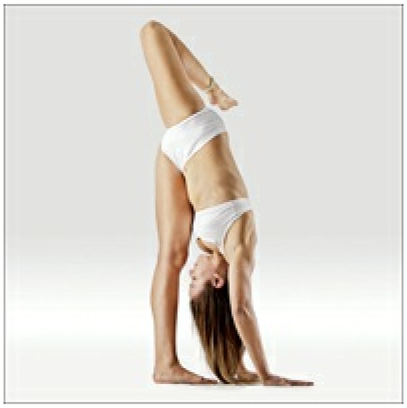

Unsupported Bound One Foot Intense Stretch Pose

Niralamba Baddha Eka Pada Uttanasana

(nir-ah-LUHM-buh BUH-duh EY-kuh PUH-duh ut-TAHS-uh-nuh)

Also Known As: Unsupported Bound One Foot Full Forward Bend
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

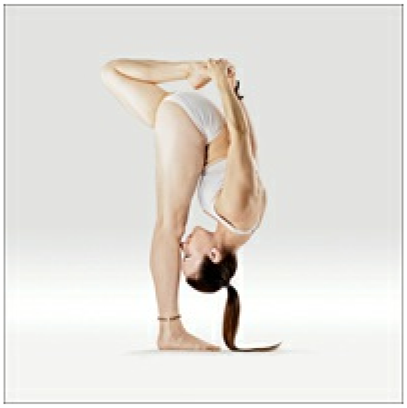

ONE LEG STANDING BALANCE: KNEE BENT—HALF FORWARD BEND

Revolved Half Moon Bow Pose

Parivritta Ardha Chandrachapasana

(puh-ri-VRIT-tuh UHR-duh CHUHN-druh-chahp-AHS-uh-nuh)

Pose Type: standing one-legged balance, backbend, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky), Padhayoragrai or Padayoragre (toes/feet), Hastagrai or Hastagre (hands)

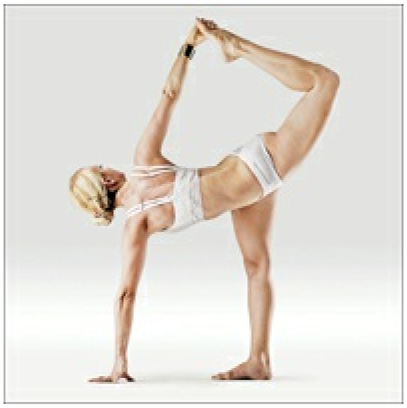

Pose Dedicated to Yogi Gitananda

Gitanandasana

(geet-ah-nuhn-DAHS-uh-nuh)

Also Known As: Lord of the Dance Pose Modification (Natarajasana)
Modification: foot away from the head
Pose Type: standing one-legged balance, backbend, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

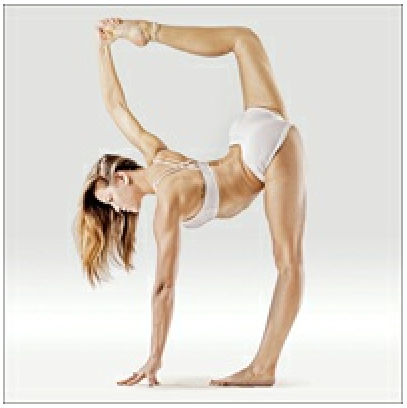

Pose Dedicated to Yogi Gitananda

Gitanandasana

(geet-ah-nuhn-DAHS-uh-nuh)

Also Known As: Lord of the Dance Pose Modification (Natarajasana)
Modification: foot to the head
Pose Type: standing one-legged balance, backbend, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ONE LEG STANDING BALANCE: HANDS BOUND BEHIND THE BACK

Toppling Tree Pose

Patan Vrikshasana

(PUH-tuhn vrik-SHAHS-uh-nuh)

Modification: both knees bent
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

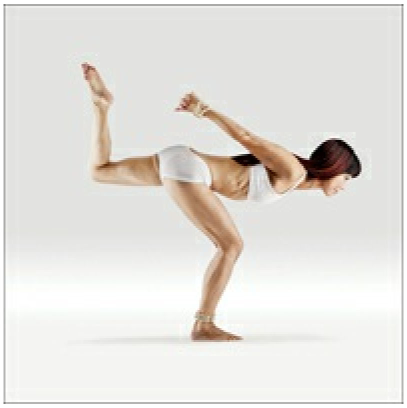

Toppling Tree Pose

Patan Vrikshasana

(PUH-tuhn vrik-SHAHS-uh-nuh)

Modification: both legs straight
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ONE LEG STANDING BALANCE: KNEE BENT—FOOT TO THE ELBOW CREASE

Lord of the Dance Pose Prep.

Natarajasana Prep.

(NAH-tuh-rahj-AHS-uh-nuh)

Modification: foot to the elbow crease, arms open to the sides, elbows bent
Pose Type: standing one-legged balance, backbend
Drishti Point: Hastagrai or Hastagre (hands)

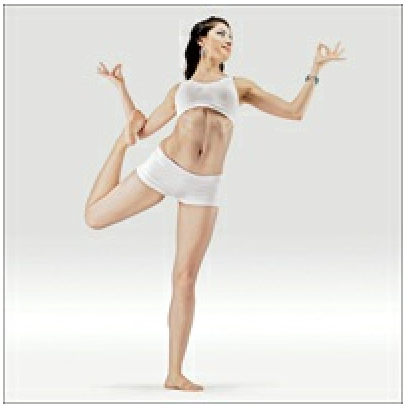

Prayer Lord of the Dance Pose

Namaskar Natarajasana

(nuh-muhs-KAHR NAH-tuh-rahj-AHS-uh-nuh)

Modification: foot to the elbow crease
Pose Type: standing one-legged balance, backbend
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

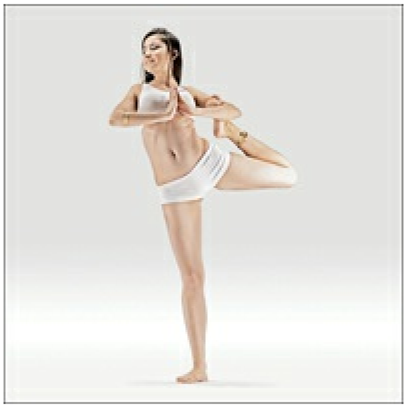

Half Moon Bow Pose Prep.

Ardha Chandrachapasana Prep.

(URH-duh chuhn-druh-chahp-AHS-uh-nuh)

Modification: foot to the elbow crease
Pose Type: standing one-legged balance, backbend, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

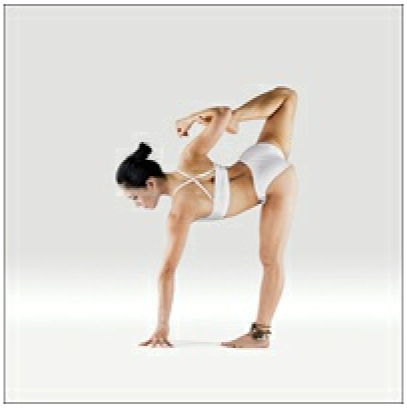

ONE LEG STANDING BALANCE: KNEE BENT BEHIND

Lord of the Dance Pose

Natarajasana

(NAH-tuh-rahj-AHS-uh-nuh)

Also Known As: Baby Dancer’s Pose (Bala Natarajasana)
Modification: hand to the foot on the same side using an under-head grip, back knee bent
1. thigh parallel to the floor
2. thigh 45 degrees to the floor, half forward bend

Pose Type: standing one-legged balance, backbend

Drishti Point: Hastagrai or Hastagre (hands)

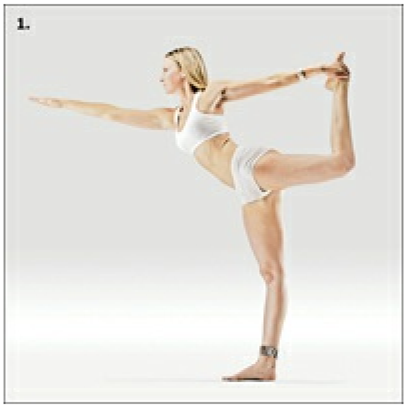
1.

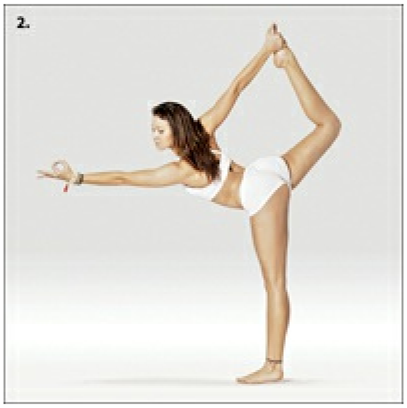
2.

Hand Position of Mermaid Pose in Bowing With Respect

Lord of the Dance Pose

Hasta Naginyasana in Nantum Natarajasana

(HUH-stuh nuh-gin-YAHS-uh-nuh in nan-toom NAH-tuh-rahj-AHS-uh-nuh)

Pose Type: standing one legged balance, backbend, forward bend

Drishti Point: Nasagrai or Nasagre (nose) or Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

How to Perform the Pose:

1. Begin by standing in Mountain Pose (*Tadasana*). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Inhale and bring your weight onto the left foot. Exhale as you bend your right knee, grab onto the right foot with your right hand, and bring your right heel to your right sitting bone. Keep your standing leg strong and straight by pulling the knee cap up and engaging the front thigh muscles (quadriceps). Keep your knees together.
3. On the next exhale, start pushing your right foot back and up to the sky, while tipping your torso forward.
4. Exhale as you bring the right foot to the right elbow crease and bend your right arm to keep the foot in place.
5. Inhale as you reach your left arm out in front of you. Exhale, bend the left elbow, and grab onto the right hand with your left hand (Pose #2). This is the arm position of Mermaid Pose (*Hasta Naginyasana*).
6. To deepen the shoulder stretch, on the next exhale start walking your left hand down your left forearm. Grab first onto the left elbow and then to your left tricep (Pose #1).
7. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch. Exhale as you release the pose, come back to Mountain Pose (*Tadasana*), and repeat on the right side.

hasta = hand

naga = great mythological snake

nantum = to bow with respect

Nataraj = name of Shiva as a cosmic dancer

1.

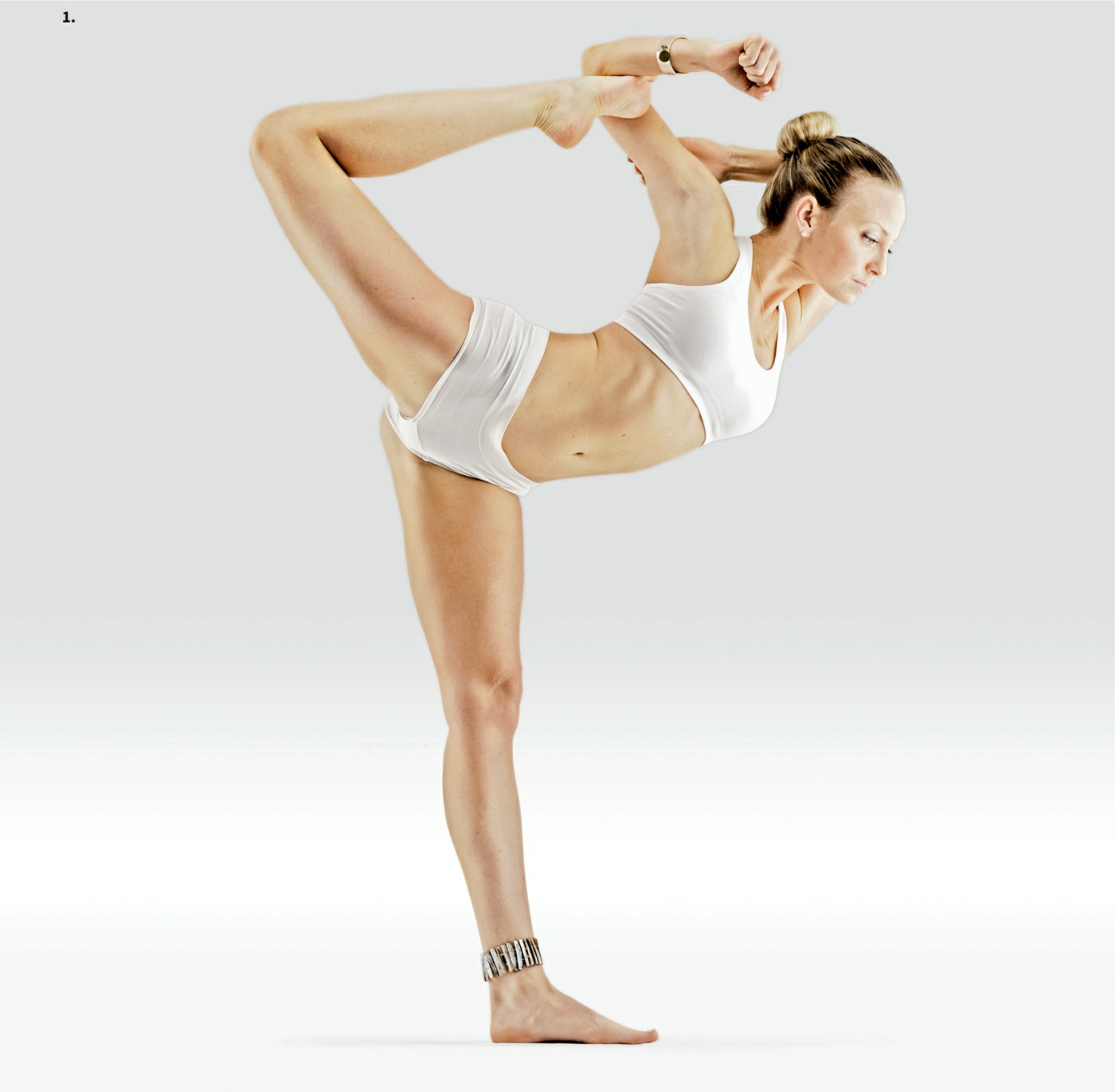

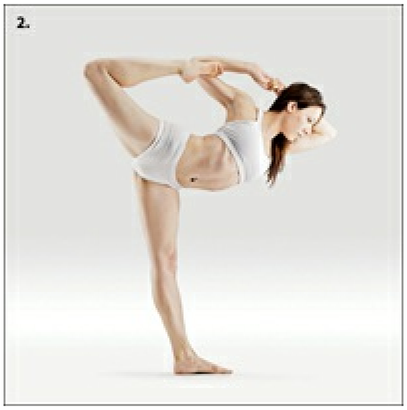

2.

Open Heart Lord of the Dance Pose

Anahata Chakra Natarajasana

(uhn-AH-huh-tuh CHUHK-ruh NAH-tuh-rahj-AHS-uh-nuh)

Modification: back knee bent, arms open to the sides, foot working toward the head
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

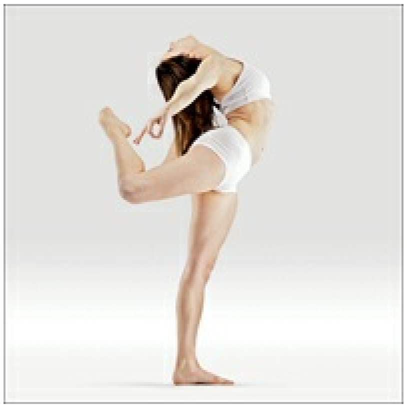

Lord of the Dance Pose

Natarajasana

(NAH-tuh-rahj-AHS-uh-nuh)

Modification: hand to opposite foot using an under-head grip, back knee bent
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

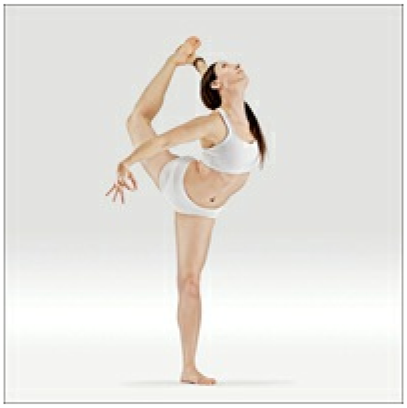

Lord of the Dance Pose

Natarajasana

(NAH-tuh-rahj-AHS-uh-nuh)

Modification: standing leg bent, heel to the floor; hand to the foot on the same side using an overhead grip, foot to the shoulder
Pose Type: standing one-legged balance, backbend
Drishti Point: Hastagrai or Hastagre (hands)
ल व र य

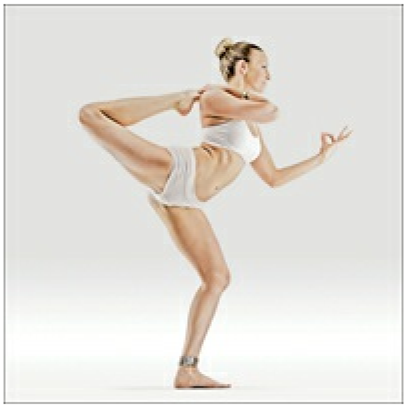

Lord of the Dance Pose

Natarajasana

(NAH-tuh-rahj-AHS-uh-nuh)

Also Known As: Lord of the Dance Pose 1 (Natarajasana 1)
Modification: both hands grabbing onto the back foot with overhead grip, foot away from the head
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

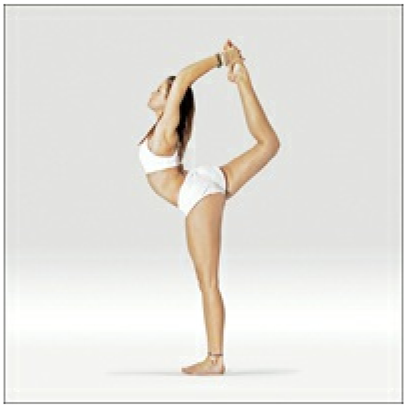

ONE LEG STANDING BALANCE: GRABBING ONTO THE BACK KNEE

Bound Lord of the Dance Pose

Baddha Natarajasana

(BUH-duh NAH-tuh-rahj-AHS-uh-nuh)

Also Known As: Yogi Yogananda Pose (Yoganandasana)
Modification: hand binding to the knee on the same side of the body; back foot away from the head
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
🔴 🟠 🟡 🟢 🔵

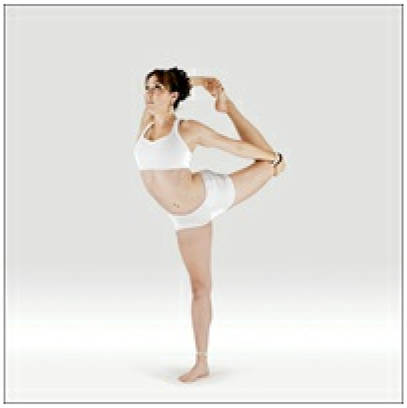

Bound Lord of the Dance Pose

Baddha Natarajasana

(BUH-duh NAH-tuh-rahj-AHS-uh-nuh)

Modification: one hand binding to the knee on the same side of the body, other arm crossed in front of the neck
1. left side view
2. right side view
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)
🔴 🟠 🟡 🟢 🔵

1.

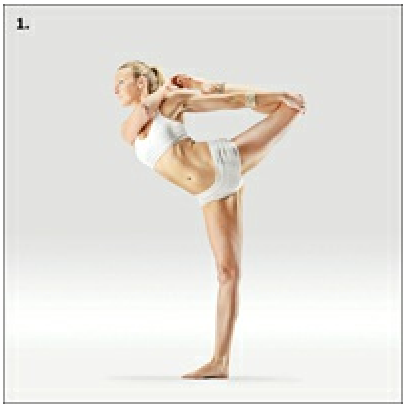

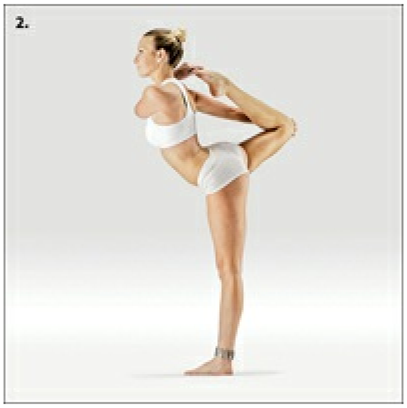

Bound Lord of the Dance Pose

Baddha Natarajasana

(BUH-duh NAH-tuh-rahj-AHS-uh-nuh)

Also Known As: Yogi Yogananda Pose (Yoganandasana)

Modification: hand binding to the inside of the thigh on the same side of the body, back foot away from the head

Pose Type: standing one-legged balance, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

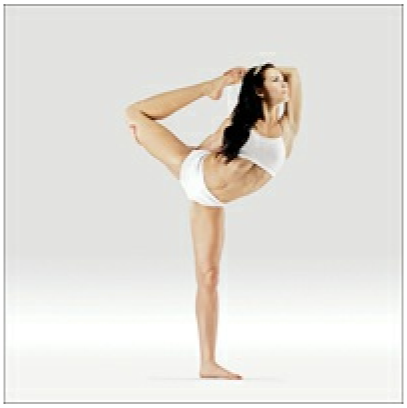

Supported Sideways Hand to Knee Lord of the Dance Pose

Salamba Parshva Hasta Janu Natarajasana

(sah-LUHM--buh PAHRSH-vuh HUH-stuh JAH-nu NAH-tuh-rahj-AHS-uh-nuh)

Pose Type: standing one-legged balance, forward bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

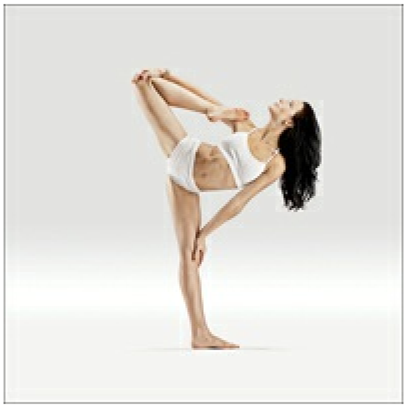

Unsupported Sideways Hand to Knee Lord of the Dance Pose

Niralamba Parshva Hasta Janu Natarajasana

(nir-ah-LUHM-buh-nuh PAHRSH-vuh HUH-stuh JAH-nu NAH-tuh-rahj-AHS-uh-nuh)

Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व र य

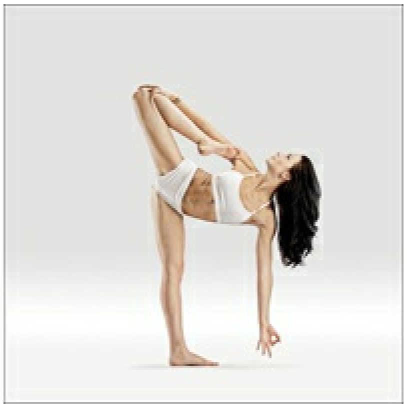

ONE LEG STANDING BALANCE: KNEE BENT IN FRONT

Hand Bound Rising Standing Wind Relieving Pose Prep.

Baddha Hasta Utthita Stiti Vayu Muktyasana Prep.

(BUH-duh HUH-stuh UT-ti-tuh STI-ti VAH-yu muk-TYAHS-uh-nuh)

Also Known As: Baddha Hasta Utthita Nindra Vayu Muktyasana Prep.
Modification: knee of the standing leg bent; fingertips to elbows, arms parallel to the floor, knee of the lifted leg to the forearm
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)
ल व र

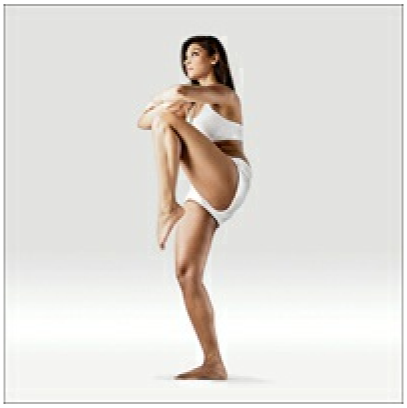

Standing Rising One-Legged Chin to Knee Pose Prep.

Stiti Utthita Eka Pada Chibi Janu Shirshasana Prep.

(STI-ti UT-ti-tuh EY-kuh PUH-duh CHI-bi JAH-nu sheer-SHAHS-uh-nuh)

Also Known As: Nindra Utthita Eka Pada Chibi Janu Shirshasana Prep.
Modification: 1. arms wrapped around the shin 2. fingers interlocked over the shin
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र ह

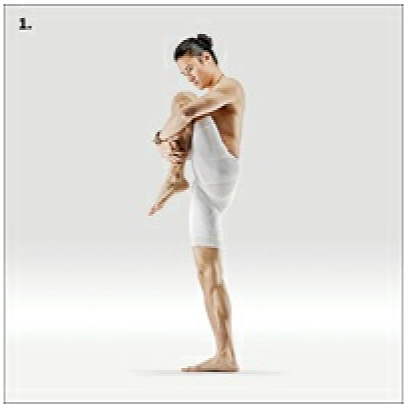

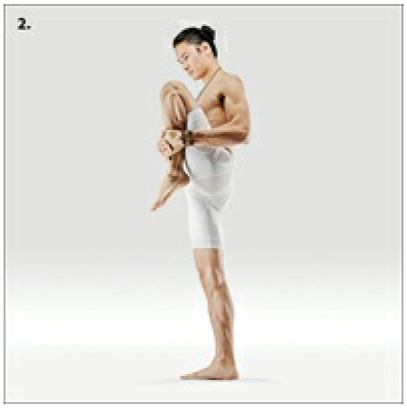

Standing Rising Wind Relieving Pose

Stiti Utthita Vayu Muktyasana

(STI-ti UT-ti-tuh VAH-yu muk-TYAHS-uh-nuh)

Also Known As: Nindra Utthita Vayu Muktyasana; Standing Pose Dedicated to Marichi (Nindra Marichyasana)
Modification: standing up straight
Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

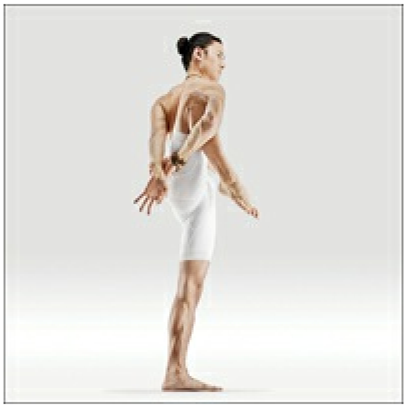

Revolved Prayer Standing Rising Wind Relieving Pose

Parivritta Namaskar Stiti Utthita Vayu Muktyasana

(puh-ri-VRIT-tuh nuh-muhs-KAHR STI-ti UT-ti-tuh VAH-yu mukTYAHS-uh-nuh)

Also Known As: Parivritta Namaskar Nindra Utthita Vayu Muktyasana
Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व र

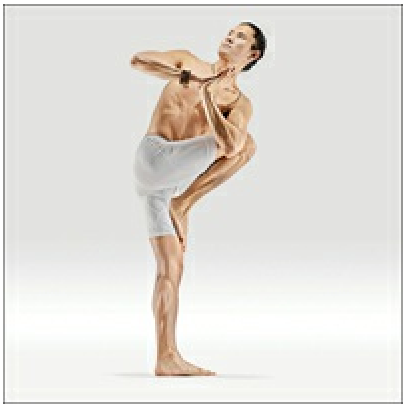

Revolved Bound Standing Rising Wind Relieving Pose

Parivritta Baddha Stiti Utthita Vayu Muktyasana

(puh-ri-VRIT-tuh BUH-duh STI-ti UT-ti-tuh VAH-yu muk-TYAHS-uh-nuh)

Also Known As: Parivritta Baddha Nindra Utthita Vayu Muktyasana
Modification: binding around the back of the thigh
1. foot on the knee
2. foot away from the knee
Pose Type: standing one-legged balance, forward bend, twist, binding
Drishti Point: Nasagrai or Nasagre (nose)

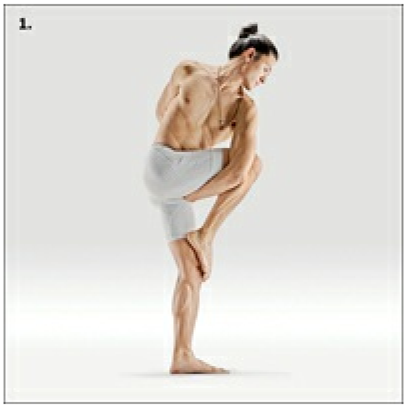

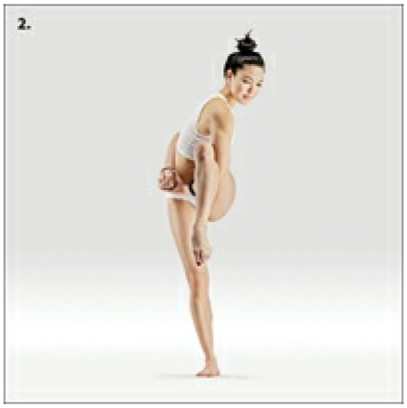

Revolved Bound Standing Rising Wind Relieving Pose

Parivritta Baddha Stiti Utthita Vayu Muktyasana

(puh-ri-VRIT-tuh BUH-duh STI-ti UT-ti-tuh VAH-yu muk-TYAHS-uh-nuh)

Also Known As: Parivritta Baddha Nindra Utthita Vayu Muktyasana
Modification: binding around the shin
Pose Type: standing one-legged balance, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

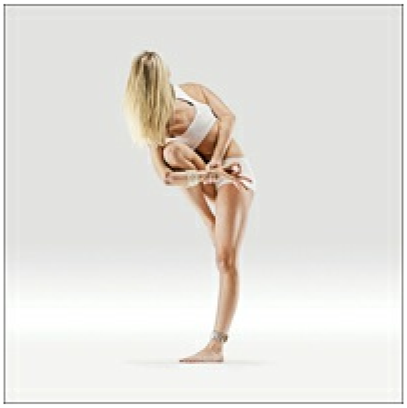

ONE LEG STANDING BALANCE: BOTH KNEES BENT—BINDING & TWISTS

Revolved Bound Standing Rising Wind Relieving Pose Prep.

Parivritta Baddha Stiti Utthita Vayu Muktyasana Prep.

(puh-ri-VRIT-tuh BUH-duh STI-ti UT-ti-tuh VAH-yu muk-TYAHS-uh-nuh)

Also Known As: Parivritta Baddha Nindra Utthita Vayu Muktyasana Prep.
Modification: foot resting on the knee, standing leg bent
Pose Type: standing one-legged balance, forward bend, twist, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

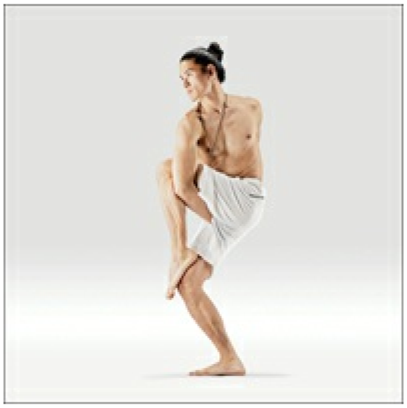

Pose Inspired by Shiva’s Vigorous Cycle of Life Dance

Tandavasana

(tahn-duh-VAHS-uh-nuh)

Modification: knee to the opposite elbow, foot resting on the knee, standing leg bent, heel down
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

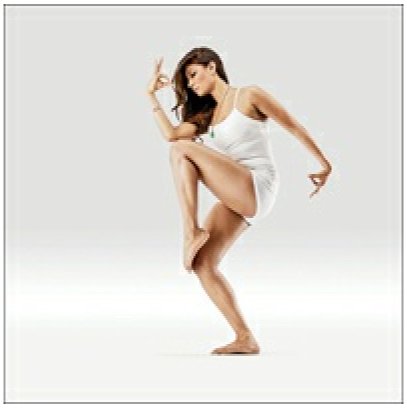

Pose Inspired by Shiva’s Vigorous Cycle of Life Dance

Tandavasana

(tahn-duh-VAHS-uh-nuh)

Modification: both arms straight, one forearm to the opposite knee, other arm straight up to the sky, both knees bent
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)
ल व र य

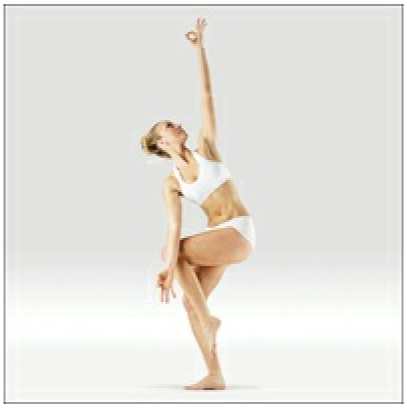

ONE LEG STANDING BALANCE: HEEL UP

Tip Toe Pose Inspired by Shiva’s Vigorous Cycle of Life Dance

Prapada Tandavasana

(PRUH-puh-duh tahn-duh-VAHS-uh-nuh)

Modification: knee to the opposite elbow, foot resting on the knee, standing leg bent
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)
ल व र

Tip Toe Pose Inspired by Parvati’s Graceful Dance

Prapada Lasyasana

(PRUH-puh-duh lahs-YAHS-uh-nuh)

Modification: deep backbend, one arm reaching up to the sky, other arm reaching down to the floor; standing leg straight, other knee bent; foot to the inside of the knee of the opposite leg
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)

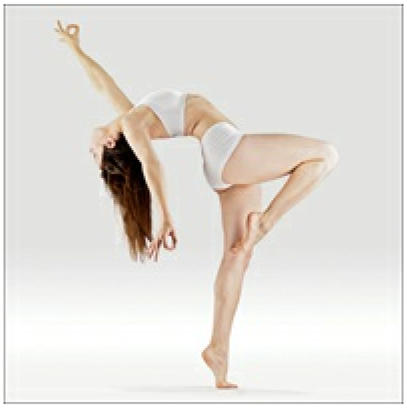

Intense Ankle Stretch Tip Toe Lord of the Dance Pose

Uttana Kulpa Prapada Natarajasana

(ut-TAHN-uh KUL-puh PRUH-puh-duh NAH-tuh-rahj-AHS-uh-nuh)

Also Known As: Tip Toe Pose Inspired by Parvati’s Graceful Dance (Prapada Lasyasana)
Modification: standing leg bent, toes curled under, overhead grip, foot to the shoulder
Pose Type: standing one-legged balance, backbend
Drishti Point: Hastagrai or Hastagre (hands)

ONE LEG STANDING BALANCE: KNEE BENT IN FRONT, BINDING, FORWARD BENDS

Half Standing Wind Relieving Intense Stretch Pose 1 & 2 Prep.

Ardha Stiti Vayu Muktyuttonasana 1 & 2 Prep.

(UHR-duh STI-ti VAH-yu muk-tew-ton-AHS-uh-nuh)

Also Known As: Ardha Nindra Vayu Muktyuttonasana; Standing Wind Relieving Half Forward Bend
Modification: fingers interlocked on the knee, half forward bend
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

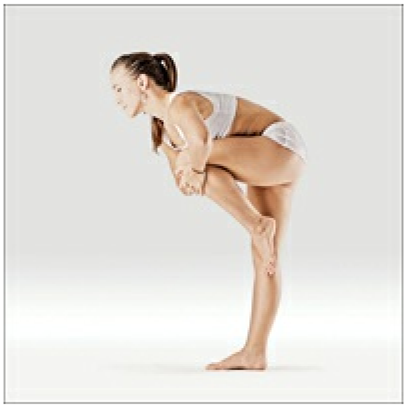

One Hand Bound Half Standing Wind Relieving Intense Stretch Pose

Eka Hasta Baddha Ardha Stiti Vayu Muktyuttonasana

(EY-kuh HUH-stuh BUH-duh UHR-duh STI-ti VAH-yu muk-tew-ton-AHS-uh-nuh)

Also Known As: Eka Hasta Baddha Ardha Nindra Vayu Muktyuttonasana; One Hand Bound Standing Wind Relieving Half Forward Bend
Modification: binding around the shin, palm to the rib cage, foot resting on the thigh of the standing leg
Pose Type: standing one-legged balance, forward bend, binding

Drishti Point: Hastagrai or Hastagre (hands)

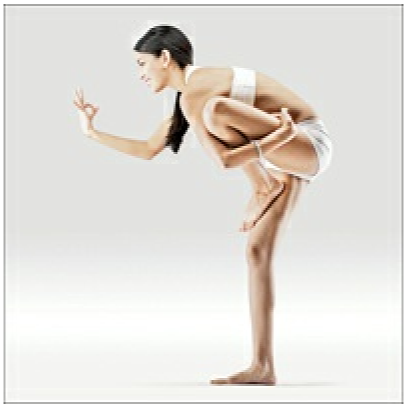

Standing Wind Relieving Intense Stretch Pose 1

Stiti Vayu Muktyuttonasana 1

(STI-ti VAH-yu muk-tew-ton-AHS-uh-nuh)

Also Known As: Nindra Vayu Muktyuttonasana; Standing Wind Relieving Half Forward Bend 1
Modification: chin away from the shin, binding on the inside of the thigh
Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

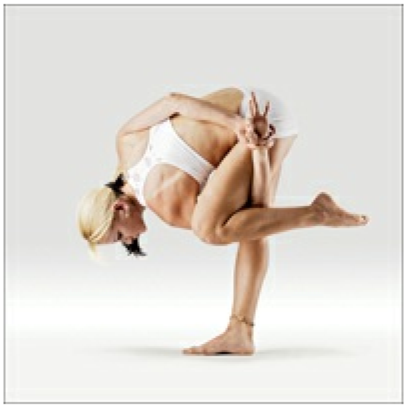

ONE LEG STANDING BALANCE: KNEE BENT IN FRONT, BINDING, TWIST FORWARD BENDS

Standing Wind Relieving Intense Stretch Pose 2

Stiti Vayu Muktyuttonasana 2

(STI-ti VAH-yu muk-tew-ton-AHS-uh-nuh)

Also Known As: Nindra Vayu Muktyuttonasana; Standing Pose Dedicated to Sage Marichi (Nindra Marichyasana) and Standing Wind Relieving Full Forward Bend

Modification: chin away from the shin, binding around the shin
Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)
ल व र य

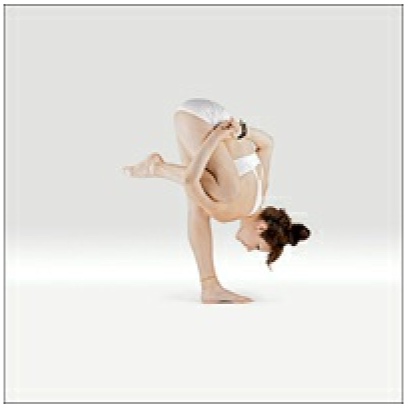

Revolved Half Standing Wind Relieving Intense Stretch Pose

Parivritta Ardha Stiti Vayu Muktyuttonasana

(puh-ri-VRIT-tuh UHR-duh STI-ti VAH-yu muk-tew-ton-AHS-uh-nuh)

Also Known As: Parivritta Ardha Nindra Vayu Muktyuttonasana; Revolved Standing Wind Relieving Half Forward Bend
Pose Type: standing one-legged balance, forward bend, twist, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व र य

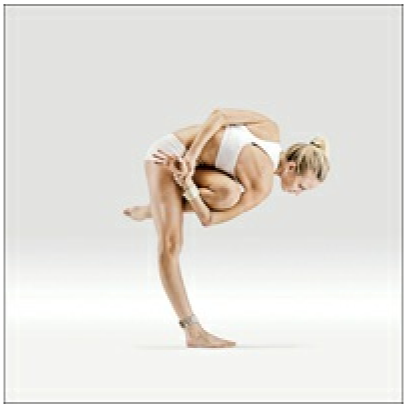

ONE LEG STANDING BALANCE: KNEE BENT IN FRONT—BACKBENDS

Hand to Knee Pose Inspired by Parvati’s Graceful Dance

Hasta Janu Lasyasana

(HUH-stuh JAH-nu lahs-YAHS-uh-nuh)

Modification: arm up over the head and parallel to the floor
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)
ल व र य ह ॐ

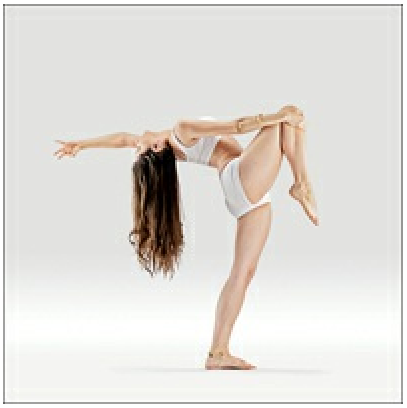

ONE LEG STANDING BALANCE: KNEE BENT IN FRONT—BACKBENDS—GRABBING ONTO THE ANKLE

Standing One-Legged Hero Pose

Stiti Eka Pada Virasana

(STI-ti EY-kuh PUH-duh veer-AHS-uh-nuh)

Also Known As: Nindra Eka Pada Virasana; Standing One-Legged Thunderbolt Pose (Nindra Eka Pada Vajrasana)
Modification: hand grabbing onto the foot, heel toward the glutes, shin parallel to the floor, other arm up to the sky
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)
ल व

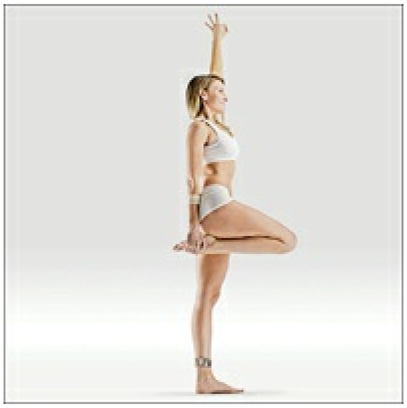

Hand to Foot Pose Inspired by Parvati’s Graceful Dance

Hasta Pada Lasyasana

(HUH-stuh PUH-duh lahs-YAHS-uh-nuh)

Modification: arm up over the head reaching to the sky
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)

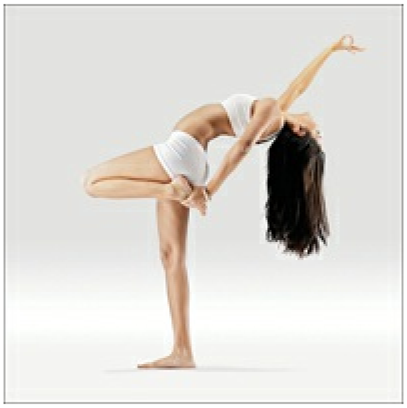

ONE LEG STANDING BALANCE: KNEE BENT, KNEES IN LINE, FORWARD BENDS

Hand to Ankle Half Standing Wind Relieving Intense Stretch Pose

Hasta Kulpha Ardha Stiti Vayu Muktyuttonasana

(HUH-stuh kul-puh UHR-duh STI-ti VAH-yu muk-tew-ton-AHS-uh-nuh)

Also Known As: Hasta Kulpha Ardha Nindra Vayu Muktyuttonasana; Knee to Ear Half Standing Wind Relieving Intense Stretch Pose (Janu Karna Ardha Nindra Vayu Muktyuttonasana) and Hand to Ankle Standing Wind Relieving Half For-ward Bend
Modification: one hand grabbing onto the ankle, knee to the temple; other arm up to the sky, elbow bent
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

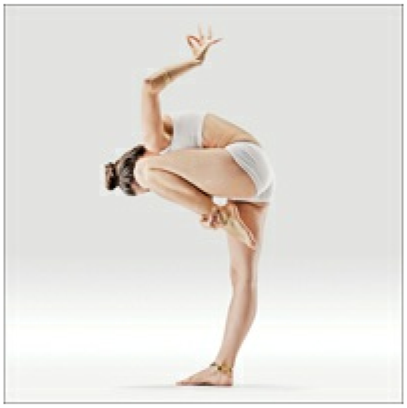

Hand to Foot One-Legged Half Intense Stretch Pose

Hasta Pada Eka Pada Ardha Uttanasana

(HUH-stuh PUH-duh EY-kuh PUH-duh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Hand to Foot One-Legged Half Forward Bend
Modification: hand to the foot from the same side, heel to the sitting bone, other palm to the floor
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल

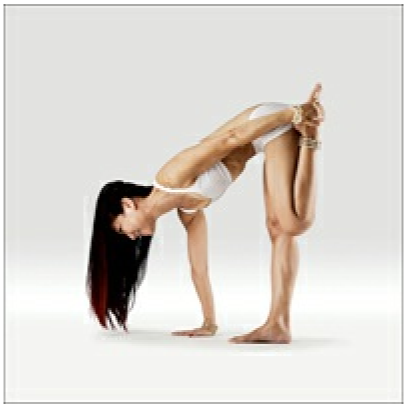

Hand to Foot One-Legged Intense Stretch Pose

Hasta Pada Eka Pada Uttanasana

(HUH-stuh PUH-duh EY-kuh PUH-duh ut-tahn-AHS-uh-nuh)

Also Known As: Hand to Foot One-Legged Full Forward Bend
Modification: hand to the foot on the same side, heel to the sitting bone; other palm to the floor
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल र य ह ॐ

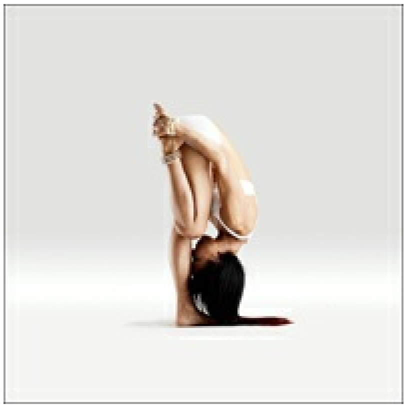

One-Legged Intense Stretch Pose 1

Eka Pada Uttanasana 1

(EY-kuh PUH-du ut-tahn-AHS-uh-nuh)

Also Known As: One-Legged Full Forward Bend 1
Modification: leg bent, heel to the sitting bone, knees together; fingertips to the floor, arms extended behind the foot
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

ल र य ह

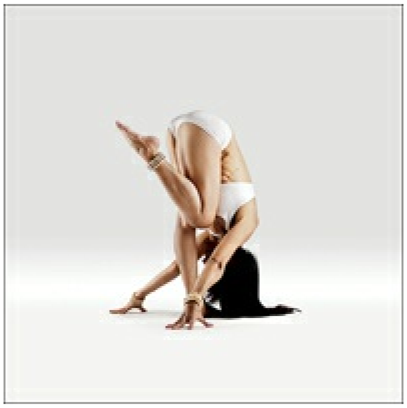

One-Legged Intense Stretch Pose 2

Eka Pada Uttanasana 2

(EY-kuh PUH-du ut-tahn-AHS-uh-nuh)

Also Known As: One-Legged Full Forward Bend 2
Modification: leg bent, heel to the sitting bone, knees together; fingertips to the floor, arms extended behind the foot
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

ल र य ह ॐ

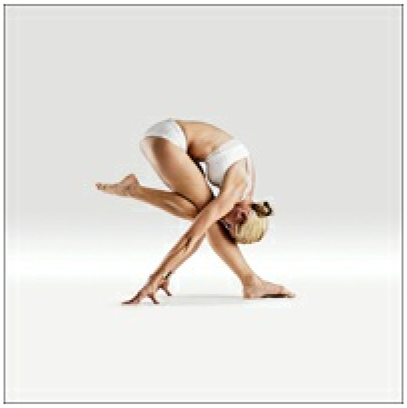

Pose Inspired by Shiva’s Vigorous Cycle of Life Dance

Tandavasana

(tahn-duh-VAHS-uh-nuh)

Modification: both knees bent, both elbows bent
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)

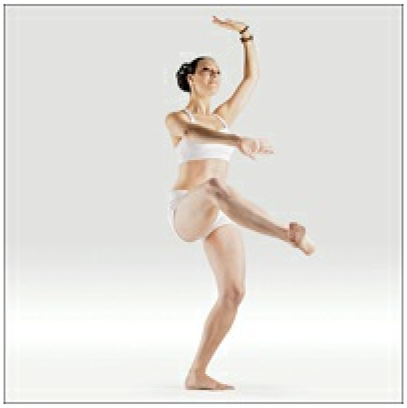

Pose Inspired by Shiva’s Vigorous Cycle of Life Dance

Tandavasana

(tahn-duh-VAHS-uh-nuh)

Modification: standing leg straight; other knee bent, foot to the inside of the knee; one forearm to the opposite knee, other arm extended up to the sky
Pose Type: standing one-legged balance, twist
Drishti Point: Hastagrai or Hastagre (hands)

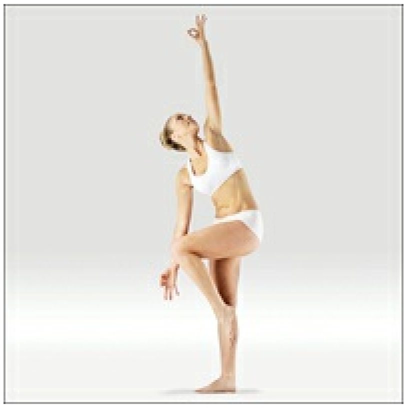

Pose Inspired by Shiva’s Vigorous Cycle of Life Dance

Tandavasana

(tahn-duh-VAHS-uh-nuh)

Modification: standing leg straight, other knee bent at 90 degrees; both elbows bent, fingertips reaching to the sky; twisting to the outside of the body
Pose Type: standing one-legged balance, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व

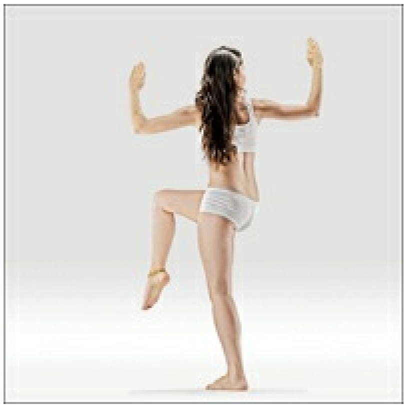

ONE LEG STANDING BALANCE: KNEE BENT TO THE SIDE

Pose Inspired by Shiva’s Vigorous Cycle of Life Dance

Tandavasana

(tahn-duh-VAHS-uh-nuh)

Modification: knee to the tricep on the same side; both elbows bent, one arm up over the head
Pose Type: standing one-legged balance, side bend
Drishti Point: Hastagrai or Hastagre (hands)
ल व

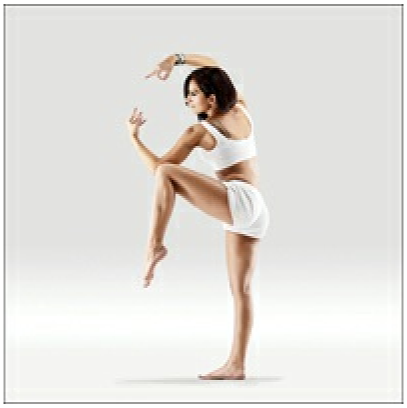

Pose Inspired by Shiva’s Vigorous Cycle of Life Dance

Tandavasana

(tahn-duh-VAHS-uh-nuh)

Modification: standing leg straight, other knee bent, toes pointing away from the head; fingertips to the temples
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)

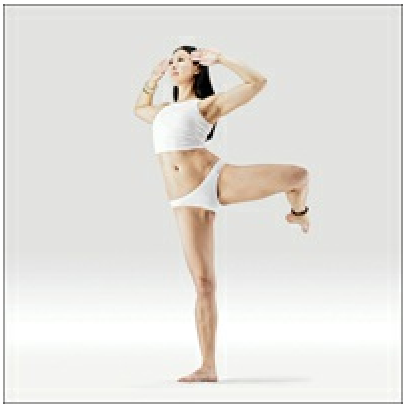

Extended Hand to Big Toe Pose 2 Prep.

Utthita Hasta Padangushtasana 2 Prep.

(UT-ti-tuh HUH-stuh puh-DAHNG-goosh-tahn-AHS-uh-nuh)

Also Known As: Extended Hand to Big Toe Pose B Prep. (Utthita Hasta Padangushtasana B Prep.)
Modification: knee bent
Pose Type: standing one-legged balance
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

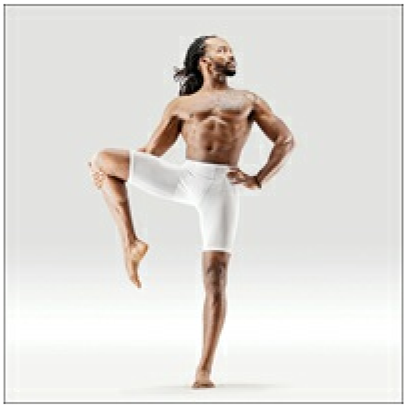

Easy Tree Pose Modification

Sukha Vrikshasana

(SOOK-uh vrik-SHAHS-uh-nuh)

Modification: hands on the hips
1. toes to the floor
2. foot to the calf muscle

Pose Type: standing one-legged balance
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

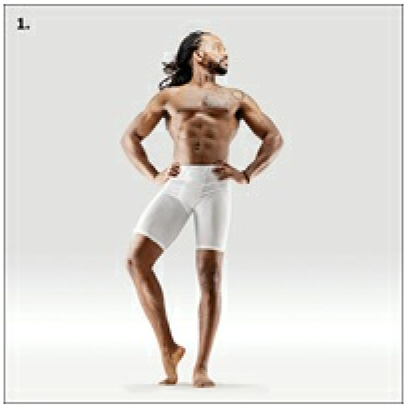

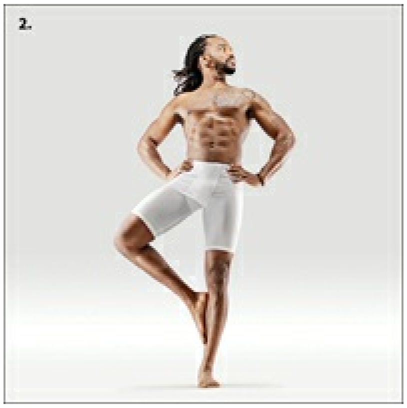

Tree Pose with Hands in Prayer

Vrikshasana Namaskar

(vrik-SHAHS-uh-nuh nuh-muhs-KAHR)

Pose Type: standing one-legged balance
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

Reverse Prayer Tree Pose

Viparita Namaskar Vrikshasana

(vi-puh-REE-tuh nuh-muhs-KAHR vrik-SHAHS-uh-nuh)

Also Known As: Back of the Body Prayer Tree Pose (Paschima Namaskara Vrikshasana)
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)

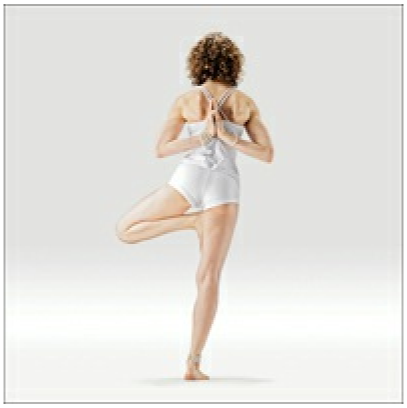

Revolved Half Bound Tree Pose

Parivritta Ardha Baddha Vrikshasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh vrik-SHAHS-uh-nuh)

Pose Type: standing one-legged balance, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

Tree Pose with Hands in Prayer

Vrikshasana Namaskar

(vrik-SHAHS-uh-nuh nuh-muhs-KAHR)

Modification: backbend
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

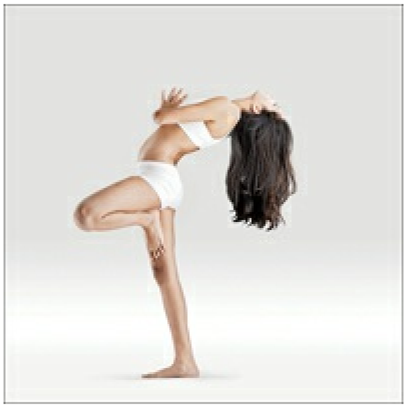

Tree Pose

Vrikshasana

(vrik-SHAHS-uh-nuh)

Also Known As: Upward Hands Tree Pose (Urdhva Hasta Vrikshasana) or Pose Dedicated to Royal Sage Bhagiratha (Bhagirathasana)

Pose Type: standing one-legged balance, mild backbend

Drishti Point: Angusthamadhye or Angustha Ma Dyai (thumbs)

How to Perform the Pose:

1. Begin by standing in Mountain Pose (*Tadasana*). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing. Find a still point on the floor to keep your gaze on. This will help you find and keep your balance.
2. Inhale and bring your weight onto the right foot. Exhale as you bend your left knee, and bring it out to the left side, opening the inside of your left hip. Keep your hips leveled and parallel to the floor (don't let your left hip go higher than your right). Keep the lower abdomen engaged to take out the compression (the arch) in the lower back.
3. Exhale as you place the sole of your left foot to the left calf muscle. Make sure the toes of your left foot are pointing to the floor.
4. On the next exhale, grab onto the left ankle with your left hand and slide the left foot up to the inside of your left thigh, keeping the toes of the left foot pointing to the floor. Avoid putting pressure on your right knee. Make sure to keep the left knee out to the side as you open the inside of your left hip.
5. Inhale as you reach both arms up over your head, fingertips pointing up to the sky. Keep them straight and shoulder-width apart. Make sure to lengthen your neck and keep your shoulder blades down your back.
6. Exhale and bring your gaze to your thumbs (Pose #1). You can experiment with rolling your head all the way back, feeling the stretch in the front of your neck (Pose #2).
7. If your shoulders are open, the neck is long, and your breathing is not constricted, you can bring your palms together on the exhale and look at your thumbs (Pose #3). With your palms pressed together, you can also experiment with rolling your head all the way back, feeling the stretch in the front of your neck (Pose #4).
8. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch. Exhale as you release the pose, come back to Mountain Pose (*Tadasana*), and repeat on the other side.

Modification: arms extended up over the head

1. arms shoulder width
2. arms shoulder width, head rolling back
3. palms together
4. palms together, head rolling back

vriksha = tree

1.

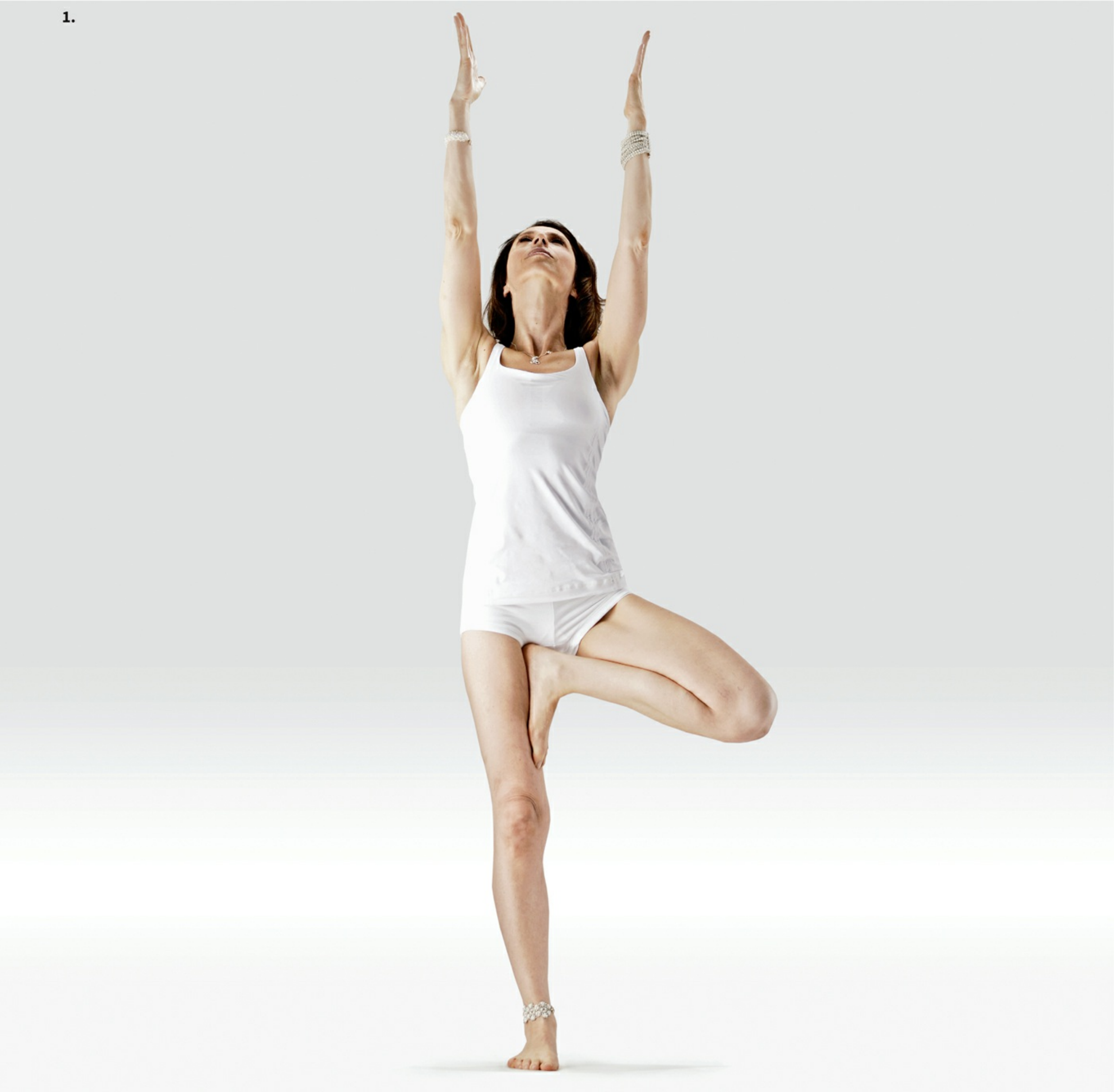

2.

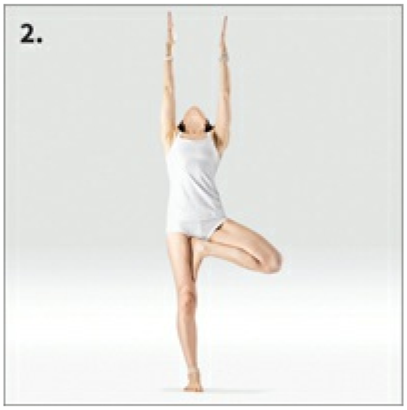

3.

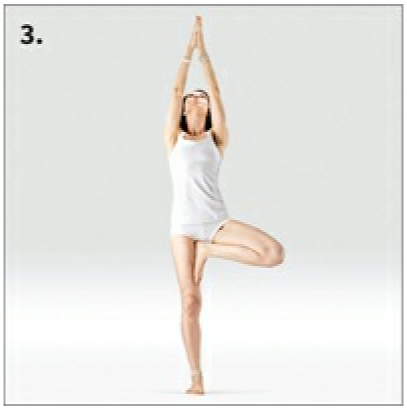

4.

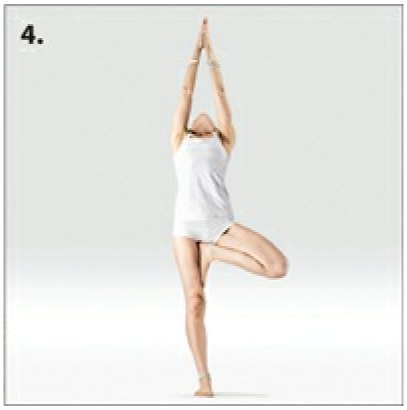

Hand Position of Cow Face Pose in Tree Pose

Hasta Gomukhasana in Vrikshasana

(HUH-stuh go-muk-AHS-uh-nuh in vrik-SHAHS-uh-nuh)

Modification: Gomukhasana arms

Pose Type: standing one-legged balance

Drishti Point: Nasagrai or Nasagre (nose)

ल व य

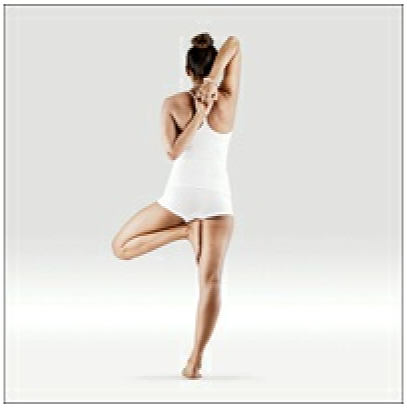

Hand Position of the Pose Dedicated to Garuda in Tree Pose

Hasta Garudasana in Vrikshasana

(HUH-stuh GUH-ru-duh-AHS-uh-nuh in vrik-SHAHS-uh-nuh)

Pose Type: standing one-legged balance

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

ल व

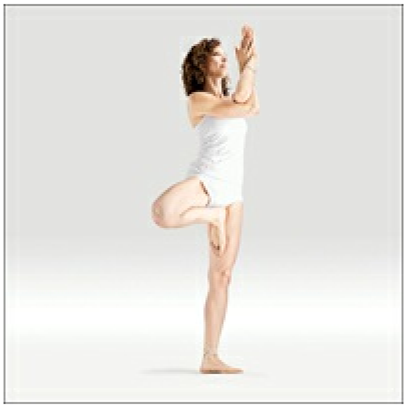

Sideways Tree Pose

Parshva Vrikshasana

(PAHRSH-vuh vrik-SHAHS-uh-nuh)

Modification: side bend toward the bent knee
Pose Type: standing one-legged balance, side bend
Drishti Point: Hastagrai or Hastagre (hands)

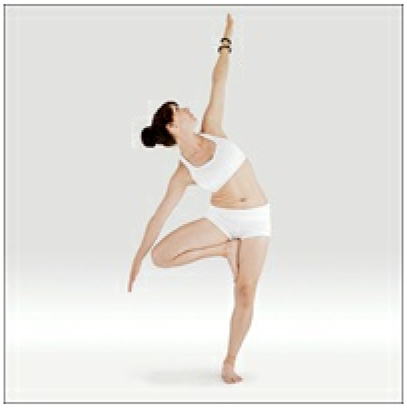

Upward One Hand Half Bound Lotus Tree Pose

Urdhva Eka Hasta Ardha Baddha Padma Vrikshasana

(OORD-vuh EY-kuh HUH-stuh UHR-duh BUH-duh PUHD-muh vrik-SHAHS-uh-nuh)

Pose Type: standing one-legged balance, binding
Drishti Point: Nasagrai or Nasagre (nose)

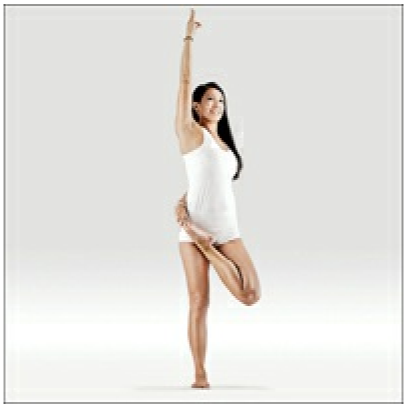

Revolved Half Bound Lotus Tree Pose

Parivritta Ardha Baddha Padma Vrikshasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh PUHD-muh vrik-SHAHS-uh-nuh)

Pose Type: standing one-legged balance, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

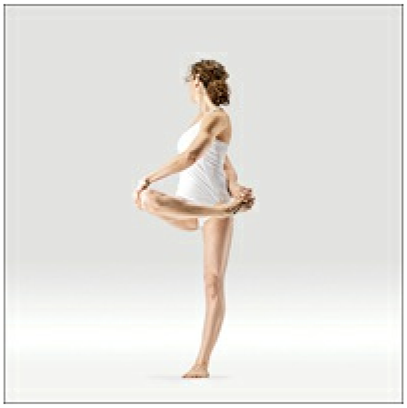

Hand to Foot Hand to Knee Tree Pose

Hasta Pada Hasta Janu Vrikshasana

(HUH-stuh PUH-duh HUH-stuh JAH-nu vrik-SHAHS-uh-nuh)

Modification: grabbing onto the foot and the knee, knee to the outside
Pose Type: standing one-legged balance
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

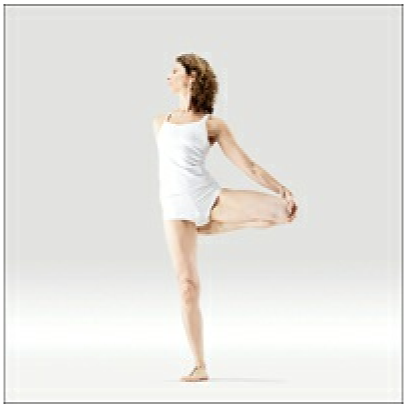

Hand to Foot Hand to Knee Tree Pose

Hasta Pada Hasta Janu Vrikshasana

(HUH-stuh PUH-duh HUH-stuh JAH-nu vrik-SHAHS-uh-nuh)

Modification: backbend, grabbing onto the foot and the knee, knee to the outside, looking straight ahead
Pose Type: standing one-legged balance, backbend
Drishti Point: Nasagrai or Nasagre (nose)

Hand to Foot Hand to Knee Toppling Tree Pose

Hasta Pada Hasta Janu Patan Vrikshasana

(HUH-stuh PUH-duh HUH-stuh JAH-nu PUH-tuhn vrik-SHAHS-uh-nuh)

Modification: grabbing onto the foot and the shin, knee to the outside
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य ह

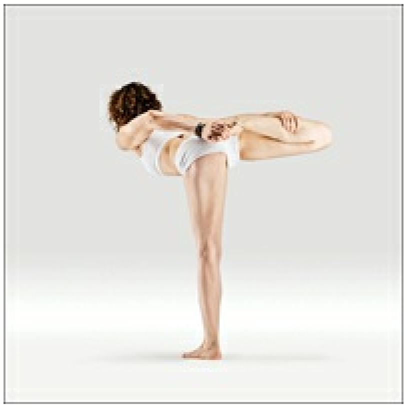

ONE LEG STANDING BALANCE: LEG STRAIGHT IN FRONT

Both Hands to Foot Pose

Dwi Hasta Padasana

(dwi-huh-stuh puh-DAHS-uh-nuh)

Also Known As: Raised Up Leg and Back Stretch Pose (Utthita Eka Pada Paschimottanasana), Standing Head to Knee Pose (Dandayamana-Janushirasana)
Modification: chin to the shin
Pose Type: standing one-legged balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

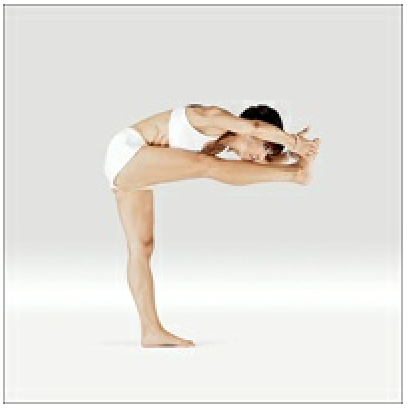

Extended One Foot Pose

Utthita Ekapadasana

(UT-ti-tuh EY-kuh-puh-DAHS-uh-nuh)

Also Known As: Extended Hand to Big Toe Pose D (Utthita Hasta Padangushtasana D; found in Ashtanga Yoga System)
Modification: hands on the hips
Pose Type: standing one-legged balance
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

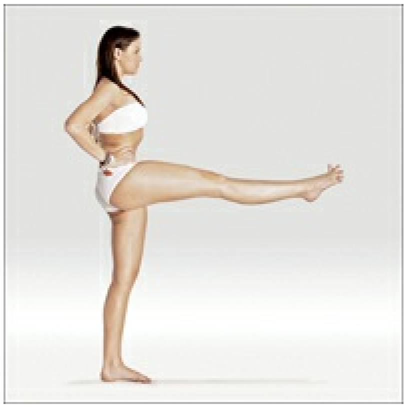

Extended Hand to Big Toe Pose 1

Utthita Hasta Padangushtasana 1

(UT-ti-tuh HUH-stuh puhd-ahng-goosh-TAHS-uh-nuh)

Also Known As: Extended Hand to Big Toe Pose A (Utthita Hasta Padangushtasana A)
Modification: lifted leg in front of the body
Pose Type: standing one-legged balance
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

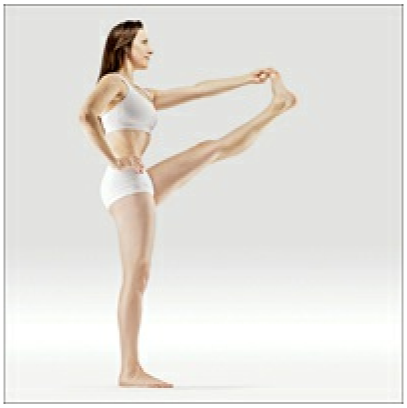

Extended Pose Dedicated to Trivikrama—Conqueror of the Three Worlds (Vishnu)

Utthita Trivikramasana

(UT-ti-tuh tri-vi-kruh-MAHS-uh-nuh)

Also Known As: Extended Hand to Foot Stretch Pose (Utthita Hastha Pada Uttanasana)
Modification: knee away from the shoulder
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)

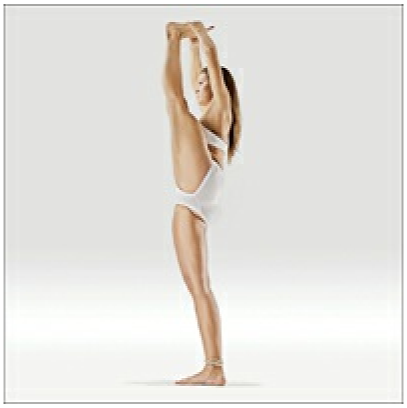

ONE LEG STANDING BALANCE: LEG STRAIGHT TO THE SIDE

Pose Dedicated to Trivikrama—Conqueror of the Three Worlds (Vishnu)

Trivikramasana

(tri-vi-kruh-MAHS-uh-nuh)

Also Known As: Leg to the Side Pose Dedicated to Trivikrama—Conqueror of the Three Worlds (Vishnu) (Parshva Pada Trivikramasana)
Modification: arm wrapping around the leg on the same side
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)
ल व

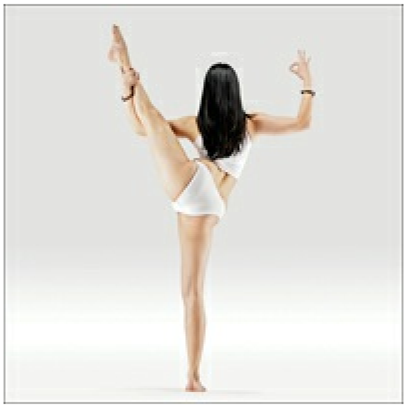

Rising Standing Sundial Pose

Utthita Stiti Surya Yantrasana

(UT-ti-tuh STI-ti SOOR-yuh yuhn-TRAHS-uh-nuh)
Also Known As: Utthita Nindra Surya Yantrasana
Modification: side bend
Pose Type: standing one-legged balance, side bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व र य

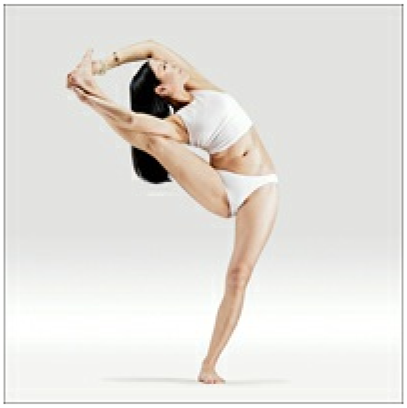

Pose Dedicated to Trivikrama—Conqueror of the Three Worlds (Vishnu)

Trivikramasana

(tri-vi-kruh-MAHS-uh-nuh)

Also Known As: Leg to the Side Pose Dedicated to Trivikrama—Conqueror of the Three Worlds (Vishnu) (Parshva Pada Trivikramasana)
Modification: both hands grabbing onto the leg, shoulder to the back of the knee, looking to the side
Pose Type: standing one-legged balance, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

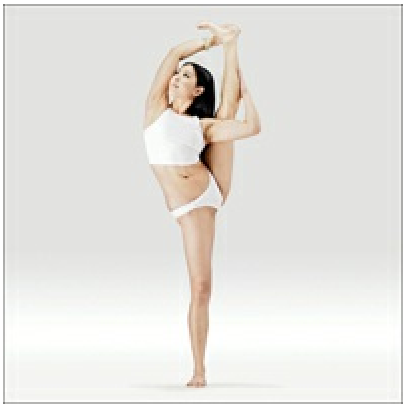

Pose Dedicated to Trivikrama—Conqueror of the Three Worlds (Vishnu)

Trivikramasana

(tri-vi-kruh-MAHS-uh-nuh)

Also Known As: Leg to the Side Pose Dedicated to Trivikrama—Conqueror of the Three Worlds (Vishnu) (Parshva Pada Trivikramasana)
Modification: one hand grabbing onto the opposite foot, shoulder to the front of the knee, other arm straight out to the side, looking straight ahead
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)

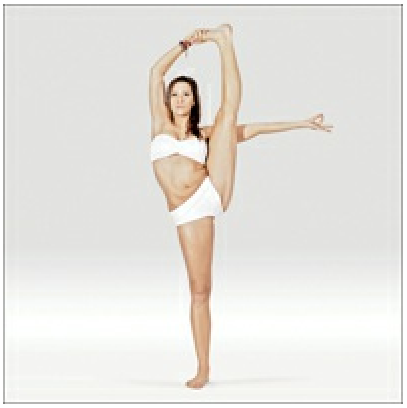

Extended Hand to Big Toe Pose 2

Utthita Hasta Padangushtasana 2

(UT-ti-tuh HUH-stuh puhd-ahng-goosh-TAHS-uh-nuh)

Also Known As: Extended Hand to Big Toe Pose B (Utthita Hasta Padangushtasana B), Standing Leg Going to the Side Pose Prep. (Utthita Parshvasahita Prep.)
Pose Type: standing one-legged balance
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व

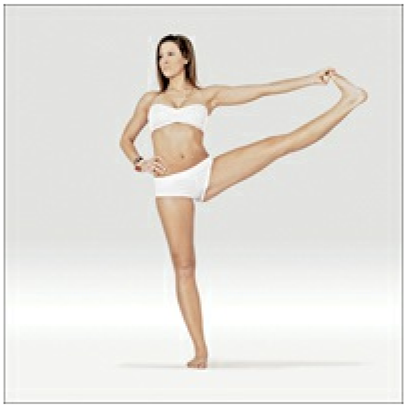

Half Bound Pose Dedicated to Trivikrama—Conqueror of the Three Worlds (Vishnu)

Ardha Baddha Trivikramasana *(UHR-duh BUH-duh tri-vi-kruh-MAHS-uh-nuh)*
Also Known As: Leg to the Side Half Bound Pose Dedicated to Trivikrama—Conqueror of the Three Worlds (Vishnu) (Parshva Pada Ardha Baddha Trivikramasana)
Modification: one arm wrapping around the leg on the same side, other arm behind the back, hand to the inside of the thigh
Pose Type: standing one-legged balance, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व य

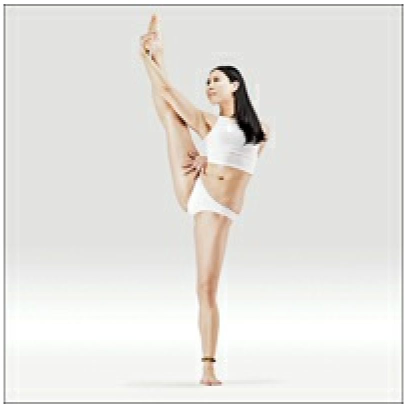

Bird of Paradise Pose

Svarga Dvijasana

(SVUHR-guh dwij-AHS-uh-nuh)

Pose Type: standing one-legged balance, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व य

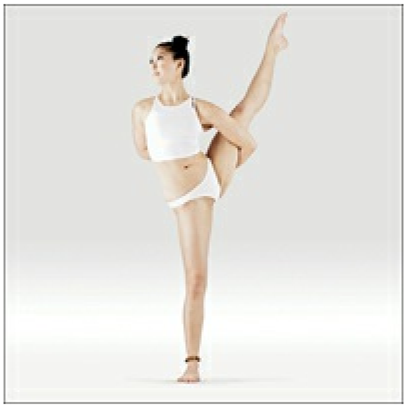

Revolved Extended Hand to Foot Pose

Parivritta Utthita Pada Hastasana

(puh-ri-VRIT-tuh UT-ti-tuh PUH-duh huh-STAHS-uh-nuh)

Also Known As: Revolved Hand to Big Toe Pose (Parivritta Hasta Padangushtasana)
Pose Type: standing one-legged balance, twist
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व

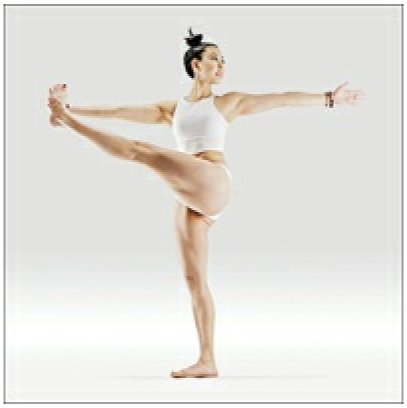

Revolved Bird of Paradise Pose

Parivritta Svarga Dvijasana

(puh-ri-VRIT-tuh SVUHR-guh dwij-AHS-uh-nuh)

Also Known As: Raised and Revolved Bound Leg Pose (Utthita Parivritta Baddha Padasana)
Pose Type: standing one-legged balance, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

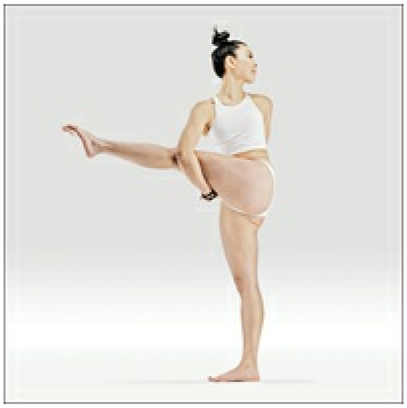

ONE LEG STANDING BALANCE: BOTH LEGS STRAIGHT—BACKBEND

Lord of the Dance Pose

Natarajasana

(NAH-tuh-rahj-AHS-uh-nuh)

Modification: under-head grip with one hand on the same side; both legs straight
Pose Type: standing one-legged balance, backbend
Drishti Point: Hastagrai or Hastagre (hands)

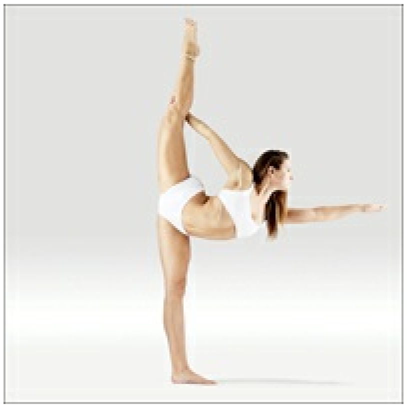

Pose Dedicated to Yogi Yogananda

Yoganandasana

(yo-gah-nuhn-DAHS-uh-nuh)

Also Known As: Pose Dedicated to Vishnu Devananda (Vishnu Devanandasana)
Modification: both legs straight, grabbing onto the back leg with an under-head grip
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

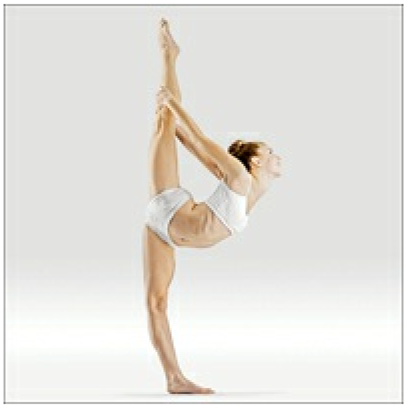

Yogi Sivananda’s Pose

Sivanandasana

(shiv-ah-nuhn-DAHS-uh-nuh)

Modification: grabbing onto the shin
Pose Type: standing one-legged balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

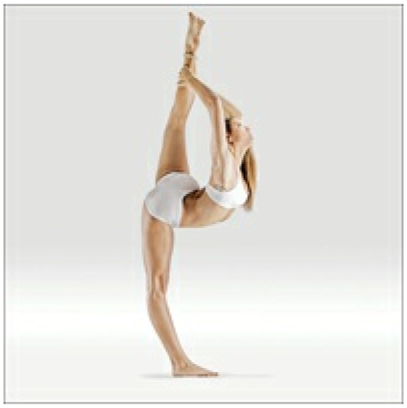

ONE LEG STANDING BALANCE: BOTH LEGS STRAIGHT—FORWARD BEND

Pose Dedicated to Sage Sundaranandar

Sundaranandarasana

(sun-duh-RAH-nuhn-duh-RAHS-uh-nuh)

Modification: palms to the floor, forehead to the shin
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

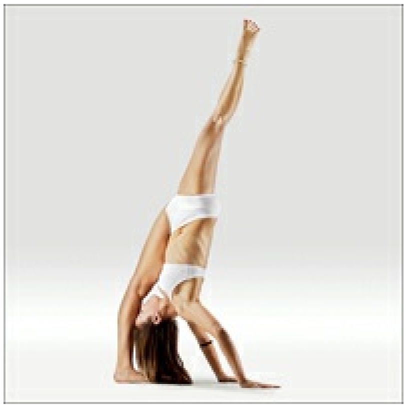

One Leg Stretched Upward Pose

Urdhva Prasarita Ekapadasana

(OORD-vuh pruh-SAH-ri-tuh ey-kuh-puhd-AHS-uh-nuh)

Modification: one hand grabbing onto the ankle on the same side, palm of the other hand to the floor, head away from the shin
Pose Type: standing one-legged balance, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet) or Nasagrai or Nasagre (nose)

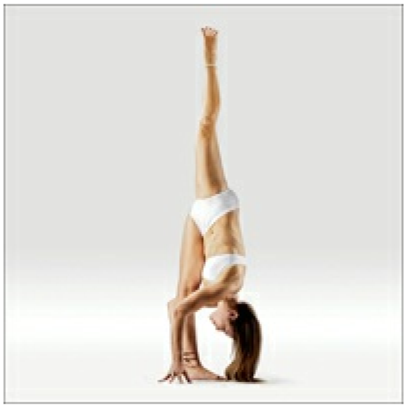

One Leg Stretched Upward Pose

Urdhva Prasarita Ekapadasana

(OORD-vuh pruh-SAH-ri-tuh ey-kuh-puhd-AHS-uh-nuh)

Modification: both hands grabbing onto the ankle, forehead to the shin
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose) if chin touches the shin
लं वं रं यं हं ॐ

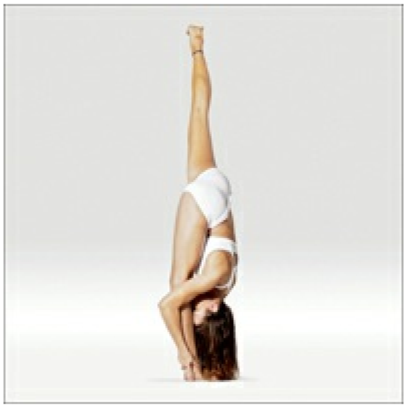

Unsupported One Leg Stretched Upward Pose

Niralamba Urdhva Prasarita Ekapadasana

(nir-AH-luhm-buh OORD-vuh pruh-SAH-ri-tuh ey-kuh-puhd-AHS-uh-nuh)

Modification: arms straight along the sides of the torso, fingertips to the sky
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Padayoragrai or Padayoragre (toes/feet), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं हं ॐ

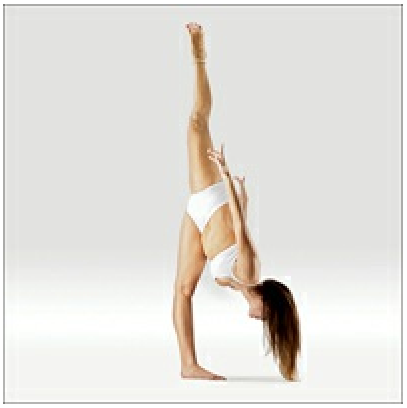

Bound Unsupported One Leg Stretched Upward Pose

Baddha Niralamba Urdhva Prasarita Ekapadasana

(BUH-duh nir-AH-luhm-buh OORD-vuh pruh-SAH-ri-tuh ey-kuh-puhd-AHS-uh-nuh-uh-nuh)

Pose Type: standing one-legged balance, forward bend, binding

Drishti Point: Nasagrai or Nasagre (nose), Padayoragrai or Padayoragre (toes/feet), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र ह ॐ

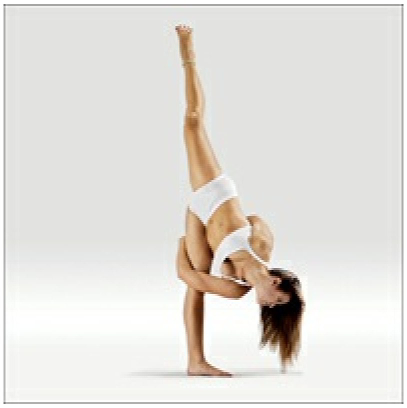

ONE LEG STANDING BALANCE: LEG STRAIGHT TO THE SIDE—HALF FORWARD BEND

Bowing with Respect Extended Hand to Big Toe Pose 2

Nantum Utthita Hasta Padangushtasana 2

(NUHN-tum UT-ti-tuh HUH-stuh puhd-ahng-goosh-TAHS-uh-nuh)

Modification: forward bend, grabbing onto the big toe of the lifted leg; other arm extended to the side parallel to the floor

Pose Type: standing one-legged balance, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

ल व ह

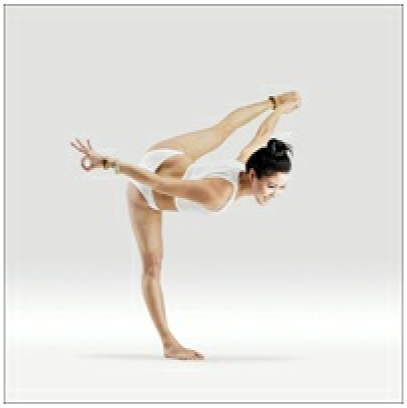

Bowing with Respect Both Hands Extended to Big Toes Pose 2

Nantum Utthita Dwi Hasta Padangushtasana 2

(NUHN-tum UT-ti-tuh DWI-huh-stuh puhd-ahng-goosh-TAHS-uh-nuh-uh-nuh)

Modification: forward bend, grabbing onto the big toes with both hands, leg lifted higher than the hip
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

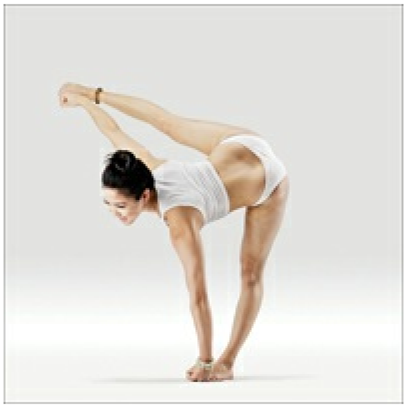

Bowing with Respect Half Bound Hand to Leg Pose

Nantum Ardha Baddha Hasta Padasana

(NUHN-tum UHR-duh BUH-duh HUH-stuh puhd-AHS-uh-nuh)

Modification: half forward bend
Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

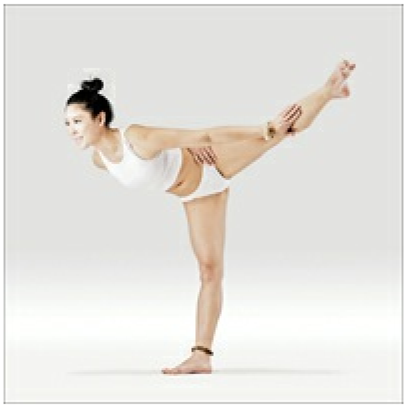

ONE LEG STANDING BALANCE: STANDING KNEE BENT & STRAIGHT—FORWARD BEND

Bowing with Respect Bird of Paradise Pose Prep.

Nantum Svarga Dvijasana Prep.

(NUHN-tum SVUHR-guh dwij-AHS-uh-nuh)

Modification: 1. standing leg bent; other knee to the back of the shoulder, leg bent
2. standing leg bent; other knee to the back of the shoulder, leg straight
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

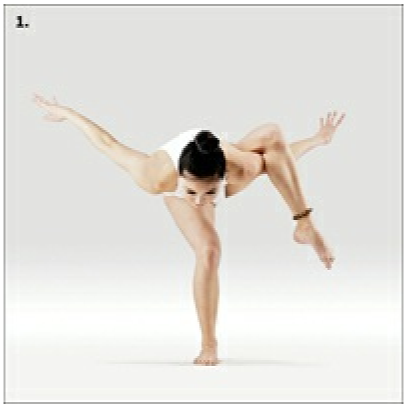

1.

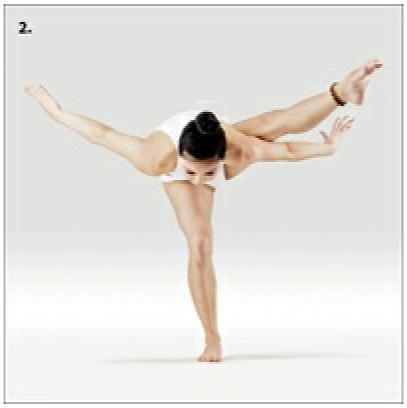

2.

Bowing with Respect Bird of Paradise Pose

Nantum Svarga Dvijasana

(NUHN-tum SVUHR-guh dwij-AHS-uh-nuh)

Modification: half forward bend

Pose Type: standing one-legged balance, forward bend, binding

Drishti Point: Nasagrai or Nasagre (nose)

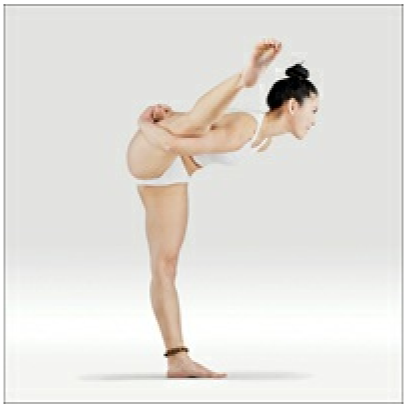

Benu Bird Pose 2

Benvasana 2

(ben-VAHS-uh-nuh)

Modification: both knees bent, chest toward the quadriceps, arms open to the sides

Pose Type: standing one-legged balance, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

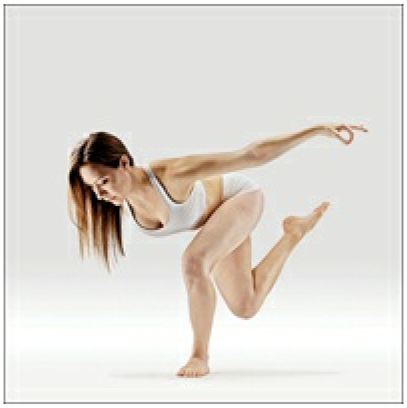

ONE LEG STANDING BALANCE: ONE LEG STRAIGHT TO THE BACK, HANDS TO THE FLOOR

Half One Leg Stretched Upward Pose

Ardha Urdhva Prasarita Ekapadasana

(UHR-duh OORD-vuh pruh-SAH-ri-tuh ey-kuh-puh-DAHS-uh-nuh)

Also Known As: One Leg Stretched Upwards Pose Prep. (Urdhva Prasarita Ekapadasana Prep.), Warrior 3 Prep. (Virabhadrasana 3 Prep.)
Modification: fingertips to the floor
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

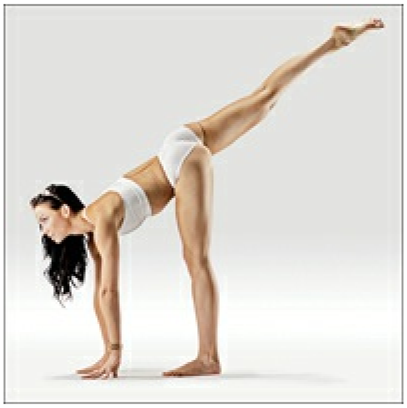

One Leg Stretched Upward Pose Prep.

Urdhva Prasarita Ekapadasana Prep.

(OORD-vuh pruh-SAH-ri-tuh ey-kuh-puh-DAHS-uh-nuh)

Modification: standing leg bent, shoulder to the back of the knee, arms extended to the back
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

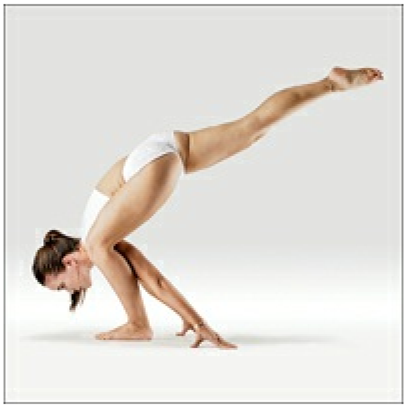

Half Moon Pose Prep.

Ardha Chandrasana Prep.

(UHR-duh chuhn-DRAHS-uh-nuh)

Modification: with yoga block prop, hand on the hip
Pose Type: standing one-legged balance, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

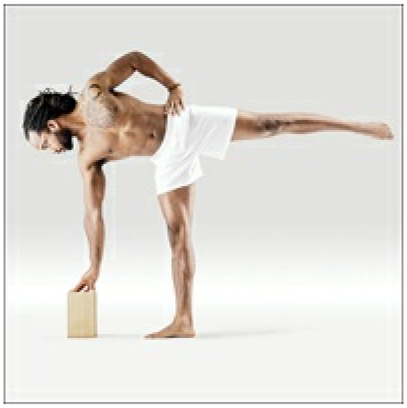

Half Moon Pose

Ardha Chandrasana

(UHR-duh chuhn-DRAHS-uh-nuh)

Pose Type: standing one-legged balance, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

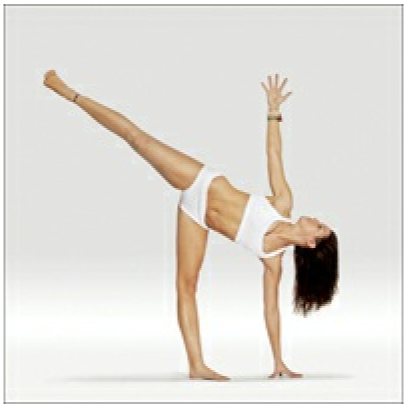

ONE LEG STANDING BALANCE: ONE LEG STRAIGHT TO THE BACK—HANDS OFF THE FLOOR

Warrior 3

Virabhadrasana 3

(VEER-uh buh-DRAHS-uh-nuh)

Also Known As: Bird Pose A (Dikasana A)
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

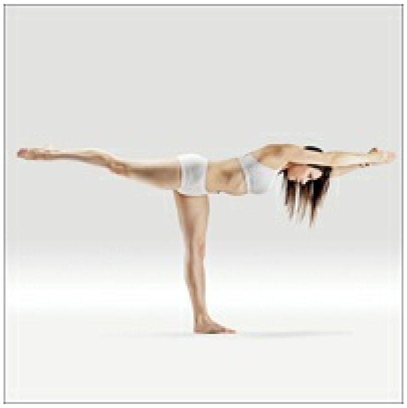

Warrior 3

Virabhadrasana 3

(VEER-uh buh-DRAHS-uh-nuh)

Also Known As: Bird Pose B (Dikasana B)
Modification: arms extended to the back and parallel to the floor
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

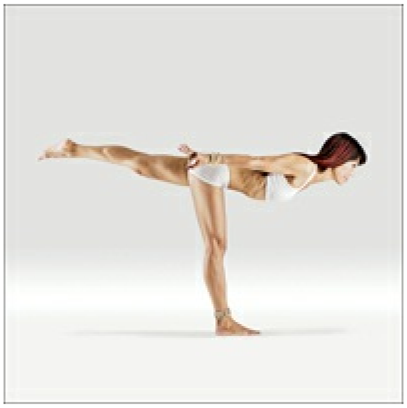

Reverse Prayer Warrior 3

Viparita Namaskar Virabhadrasana 3

(vi-puh-REE-tuh nuh-muhs-KAHR VEER-uh buh-DRAHS-uh-nuh)

Also Known As: Back of the Body Prayer Warrior 3 (Paschima Namaskara Virabhadrasana 3)
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

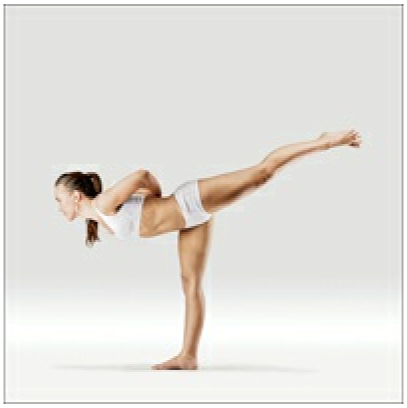

Hand Position of the Pose Dedicated to Garuda in Warrior 3

Hasta Garudasana in Virabhadrasana 3

(HUH-stuh guh-ru-DAHS-uh-nuh in VEER-uh-buh-DRAHS-uh-nuh)

Pose Type: standing one-legged balance, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

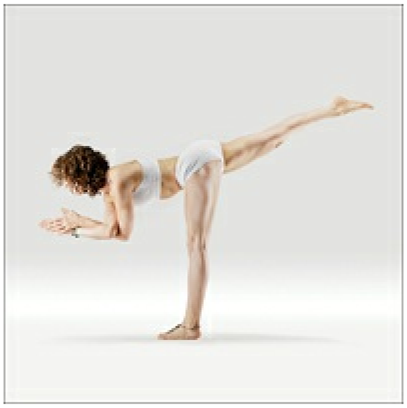

Half Lotus Warrior 3

Ardha Padma Virabhadrasana 3

(UHR-duh PUHD-muh VEER-uh-buh-DRAHS-uh-nuh-uh-nuh)

Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

Half Moon Pose

with Hands in Prayer

Ardha Chandrasana Namaskar

(UHR-duh chuhn-DRAHS-uh-nuh nuh-muhs-KAHR)

Pose Type: standing one-legged balance, forward bend

Drishti Point: 1. Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) 2. Urdhva or Antara Drishti (up to the sky)

How to Perform the Pose:

1. Begin by doing Extended Triangle Pose (*Utthita Trikonasana*) with the left foot in front and your left hand to the floor outside the left foot (either on the fingertips or with the palm flat on the floor). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale as you slide your left palm out to the front while bending your left knee. Experiment with the distance to find your balance.
3. Inhale and lift your right foot off the floor until your right leg is parallel to the floor. Keep it straight and reaching away from your head. Straighten your standing left leg.
4. Exhale and rotate your chest to the side so that your shoulders stack one on top of the other.
5. On your next exhale, lift your left arm up off the floor and bring your hands into prayer at the center of your chest.
6. You can either find a gazing point on the floor that is not moving to help keep your balance (Pose #1) or challenge yourself by looking up to the sky (Pose #2).
7. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch. Exhale, lower your right foot to the floor, coming back into Extended Triangle Pose (*Utthita Trikonasana*), and repeat on the other side.

Modification:

1. looking down
2. looking up to the sky

ardha = half

chandra = moon

namaskar = greeting with hands in Anjali Mudra (hands in prayer)

1.

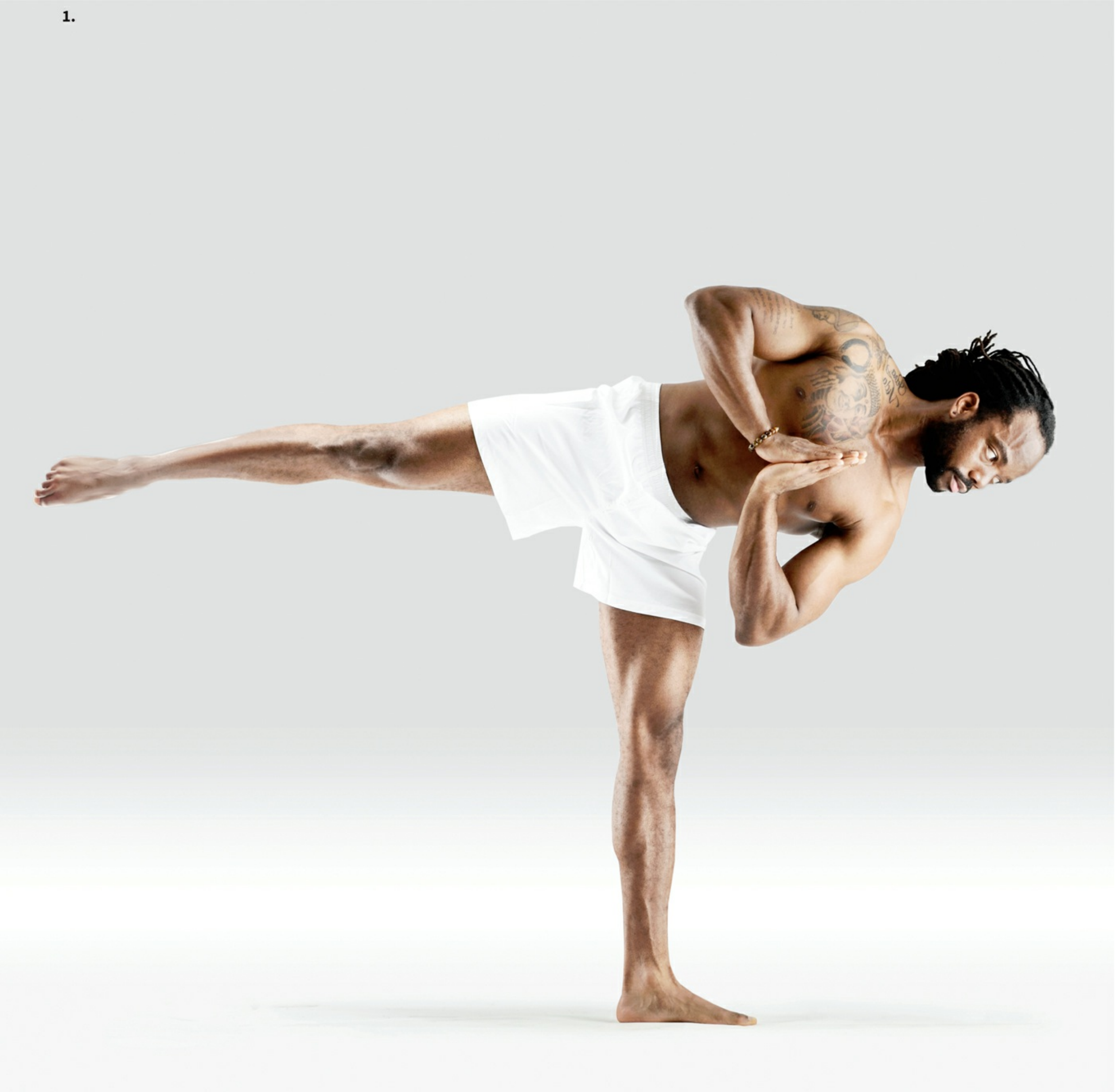

2.

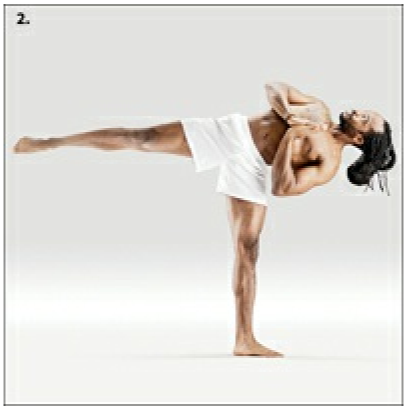

Reverse Prayer Half Moon Pose

Viparita Namaskar Ardha Chandrasana

(vi-puh-REE-tuh nuh-muhs-KAHR UHR-duh chuhn-DRAHS-uh-nuh)

Also Known As: Back of the Body Prayer Half Moon Pose (Paschima Namaskara Ardha Chandrasana)
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

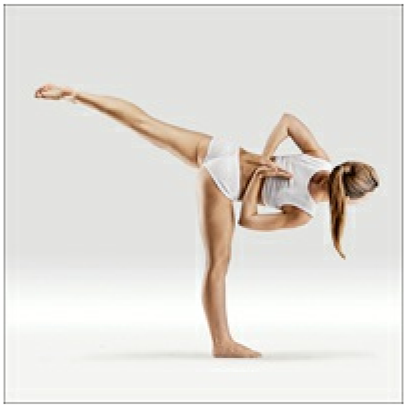

Bound Half Moon Pose

Baddha Ardha Chandrasana

(BUH-duh UHR-duh chuhn-DRAHS-uh-nuh)

Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose) or Padayoragrai or Padayoragre (toes/feet)

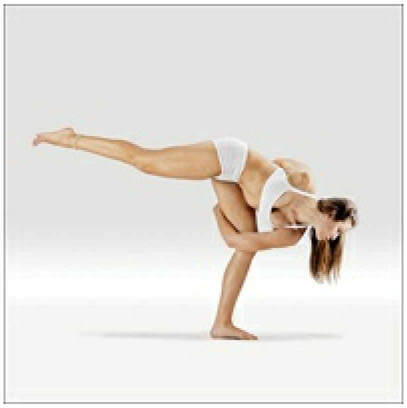

Revolved Bound Half Moon Pose

Parivritta Baddha Ardha Chandrasana

(vi-puh-REE-tuh BUH-duh UHR-duh chuhn-DRAHS-uh-nuh)

Modification: looking up to the sky
Pose Type: standing one-legged balance, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

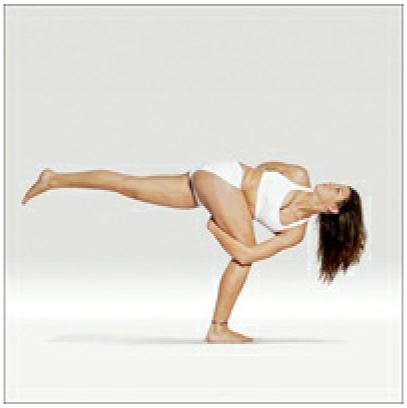

Half Firelog Pose in Half Moon Pose Prep.

Ardha Agnistambhasana in Ardha Chandrasana Prep.

(UHR-duh uhg-ni-stuhm-BAHS-uh-nuh in UHR-duh chuhn-DRAHS-uh-nuh)

Modification: foot on top of the knee, palm to the lower back
Pose Type: standing one-legged balance, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

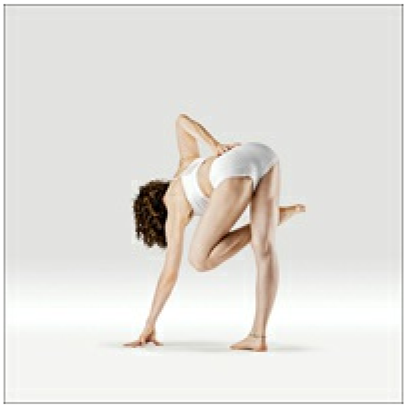

Half Bound Lotus Half Moon Pose Prep.

Ardha Baddha Padma Ardha Chandrasana Prep.

(UHR-duh BUH-duh PUHD-muh UHR-duh chuhn-DRAHS-uh-nuh)

Modification: arm behind the back, binding to the inside of the thigh of the leg in Half Lotus

Pose Type: standing one-legged balance, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

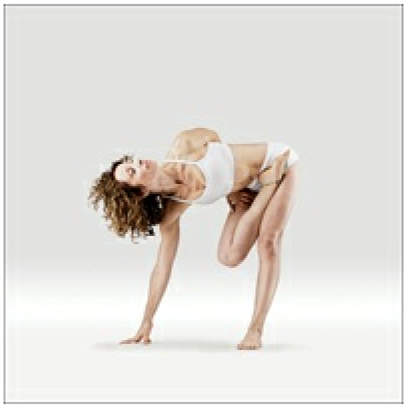

Leg Position of Half Cow Face Pose in Half Moon Pose Prep.

Pada Ardha Gomukhasana in Ardha Chandrasana Prep.

(PUH-duh UHR-duh go-muk-AHS-uh-nuh in UHR-duh chuhn-DRAHS-uh-nuh)
Also Known As: One-Legged Cow Face Pose in Half Moon Pose Prep. (Eka Pada Gomukhasana in Ardha Chandrasana Prep.)
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

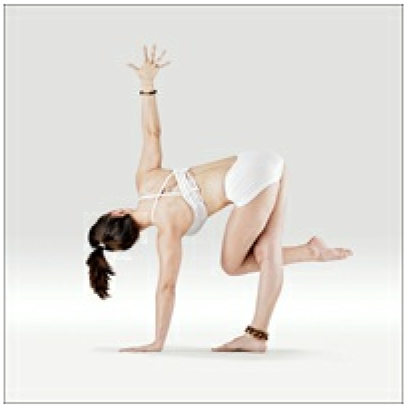

ONE-LEGGED STANDING BALANCE: HALF LOTUS

Half Bound Lotus Intense Stretch Pose Prep.

Ardha Baddha Padmottanasana Prep.

(UHR-duh BUH-duh puhd-mo-tahn-AHS-uh-nuh)
Also Known As: Half Bound Lotus Half Forward Bend Prep., Half Lotus Half Intense Stretch Pose

(Ardha Padma Ardha Uttanasana), Half Lotus Half Forward Bend
Modification: half forward bend, fingertips of both hands to the floor
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

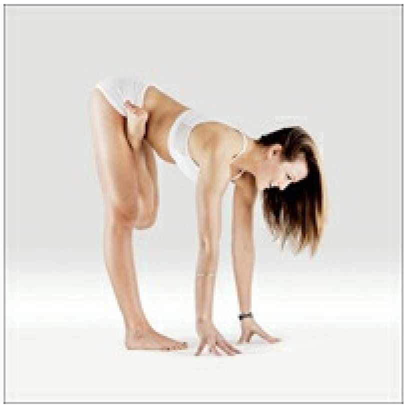

Half Bound Lotus Intense Stretch Pose Prep.

Ardha Baddha Padmottanasana Prep.

(UHR-duh BUH-duh puhd-mo-tahn-AHS-uh-nuh)
Also Known As: Half Bound Lotus Half Forward Bend
Modification: half forward bend
Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

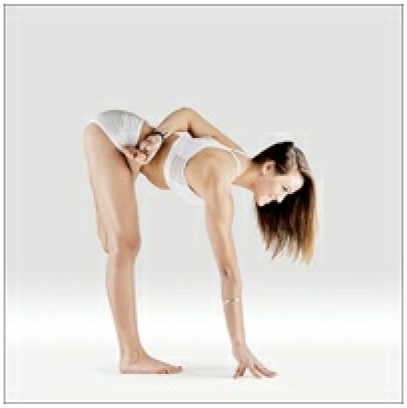

Half Bound Lotus Intense Stretch Pose

Ardha Baddha Padmottanasana

(UHR-duh BUH-duh puhd-mo-tahn-AHS-uh-nuh)

Also Known As: Half Bound Lotus Full Forward Bend
Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

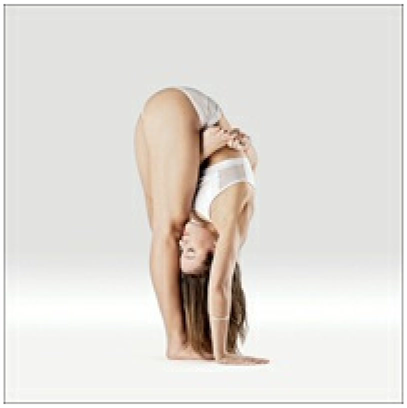

GARUDA LEGS

Pose Dedicated to Garuda

Garudasana

(guh-ru-DAHS-uh-nuh)
Also Known As: Eagle Pose
Modification: hands intertwined in front, low stance
Pose Type: standing one-legged balance
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व

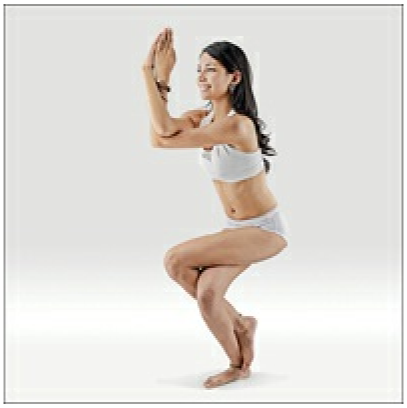

Reverse Heart Chakra Seal in Leg Position of the Pose Dedicated to Garuda

Viparita Anahata Chakra Mudra in Pada Garudasana

(vi-puh-REE-tuh un-AH-huh-tuh chuh-kruh MU-druh in PUH-duh guh-ru-DAHS-uh-nuh)

Also Known As: Reverse Heart Chakra Seal in Leg Position of the Eagle Pose
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)
ल व

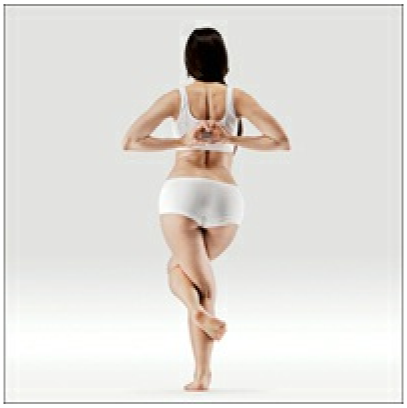

Pose Dedicated to Garuda

Garudasana

(guh-ru-DAHS-uh-nuh)

Also Known As: Eagle Pose
Modification: elbows to the knees
Pose Type: standing one-legged balance, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व र य

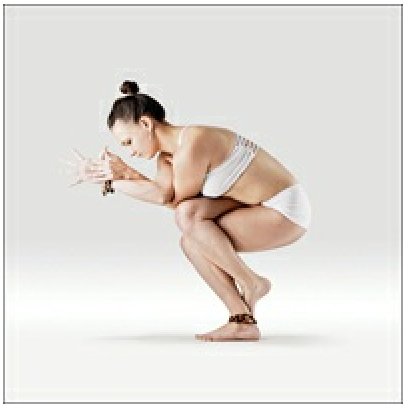

Leg Position of the Pose Dedicated to Garuda

Pada Garudasana

(PUH-duh guh-ru-DAHS-uh-nuh)

Also Known As: Leg Position of the Eagle Pose
Modification: arms straight out to the sides behind the back
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

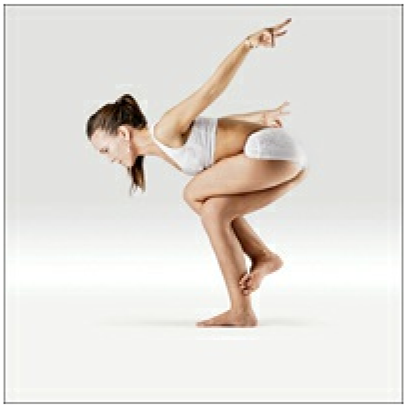

Revolved Bound Leg Position of the Pose Dedicated to Garuda

Parivritta Baddha Pada Garudasana

(puh-ri-VRIT-tuh BUH-duh PUH-duh guh-ru-DAHS-uh-nuh)
Also Known As: Revolved Bound Leg Position of the Eagle Pose
Pose Type: standing one-legged balance, forward bend, binding, twist
Drishti Point: Nasagrai or Nasagre (nose)
ल व र

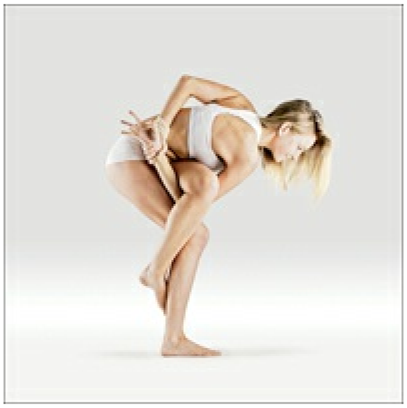

Hands Bound Leg Position of the Pose Dedicated to Garuda

Baddha Hasta Pada Garudasana

(BUH-duh HUH-stuh PUH-duh guh-ru-DAHS-uh-nuh)
Also Known As: Hands Bound Leg Position of the Eagle Pose
Pose Type: standing one-legged balance, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

ल व र य

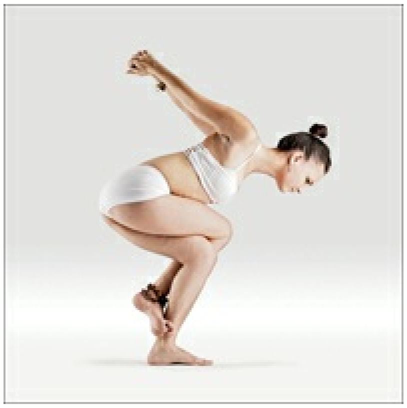

ONE-LEGGED SQUATS: TWISTS

Revolved Half Standing Wind Relieving Intense Stretch Pose

Parivritta Ardha Stiti Vayu Muktyuttonasana

(puh-ri-VRIT-tuh UHR-duh STI-ti VAH-yu muk-tew-to-NAHS-uh-nuh)

Also Known As: Parivritta Ardha Nindra Vayu Muktyuttonasana; Revolved Standing Wind Relieving Half Forward Bend
Modification: knees together; elbow to the opposite knee, other arm extended to the back
Pose Type: standing one-legged balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

ल ह ॐ

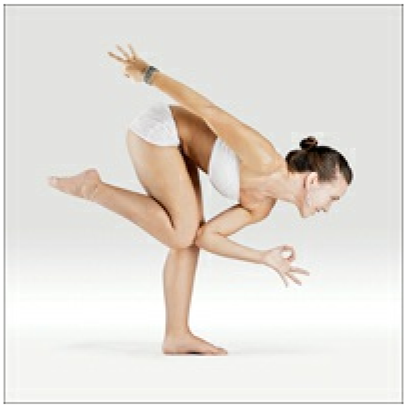

Revolved Half Standing Wind Relieving Intense Stretch Pose with Hands in Prayer

Parivritta Ardha Stiti Vayu Muktyuttonasana Namaskar

(puh-ri-VRIT-tuh UHR-duh STI-ti VAH-yu muk-tew-to-NAHS-uh-nuh nuh-muhs-KAHR)

Also Known As: Parivritta Ardha Nindra Vayu Muktyuttonasana Namaskar; Revolved Standing Wind Relieving Half Forward Bend with Hands in Prayer
Modification: back leg crossed under the front leg
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose)

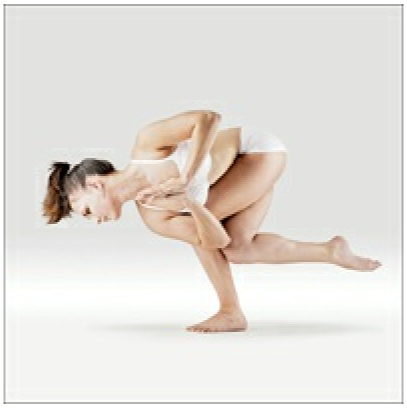

Bound Revolved Half Standing Wind Relieving Intense Stretch Pose

Baddha Parivritta Ardha Stiti Vayu Muktyuttonasana

(BUH-duh puh-ri-VRIT-tuh UHR-duh STI-ti VAH-yu muk-tew-to-NAHS-uh-nuh)

Also Known As: Baddha Parivritta Ardha Nindra Vayu Muktyuttonasana; Bound Revolved Standing Wind Relieving Half Forward Bend
Pose Type: standing one-legged balance, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

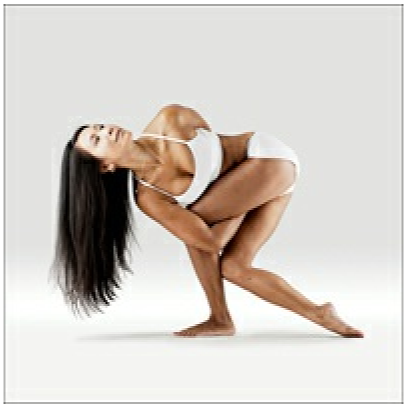

ONE-LEGGED SQUATS: KNEES IN LINE—HEEL TO THE SITTING BONES—GRABBING ONTO THE ANKLE

Revolved One Hand to Foot Half Standing Wind Relieving Intense Stretch Pose

Parivritta Eka Hasta Pada Ardha Stiti Vayu Muktyuttonasana

(puh-ri-VRIT-tuh EY-kuh HUH-stuh PUH-duh UHR-duh STI-ti VAH-yu muk-tew-to-NAHS-uh-nuh)

Also Known As: Parivritta Eka Hasta Pada Ardha Nindra Vayu Muktyuttonasana; Revolved One Hand to Foot Standing Wind Relieving Half Forward Bend
Modification: knees together; elbow to the opposite knee, other hand grabbing onto the ankle
Pose Type: standing one-legged balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

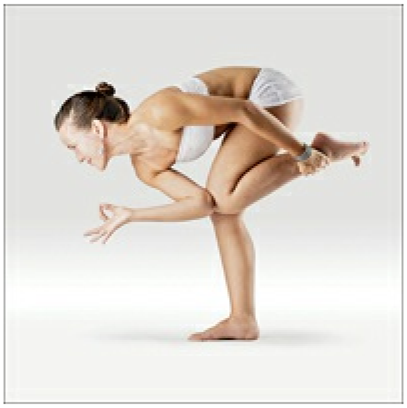

Both Hands to Foot Half Standing Wind Relieving Intense Stretch Pose

Dwi Hasta Pada Ardha Stiti Vayu Muktyuttonasana

(DWI-huh-stuh PUH-duh UHR-duh STI-ti VAH-yu muk-to-NAHS-uh-nuh)

Also Known As: Dwi Hasta Pada Ardha Nindra Vayu Muktyuttonasana
Pose Type: standing one-legged balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

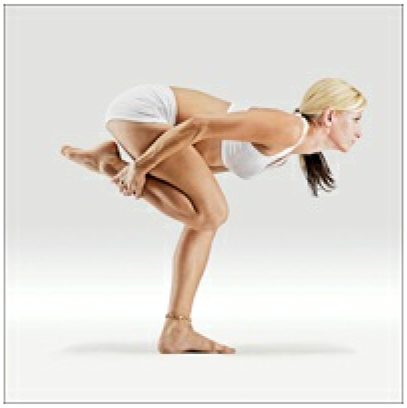

One Hand to Foot Tip Toe Half Standing Wind Relieving Intense Stretch Pose

Eka Hasta Pada Prapada Ardha Stiti Vayu Muktyuttonasana

(EY-kuh HUH-stuh PUH-duh PRUH-puh-duh UHR-duh STI-ti VAH-yu muk-to-NAHS-uh-nuh)

Also Known As: Eka Hasta Pada Prapada Ardha Nindra Vayu Muktyuttonasana
Pose Type: standing one-legged balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं हं ॐ

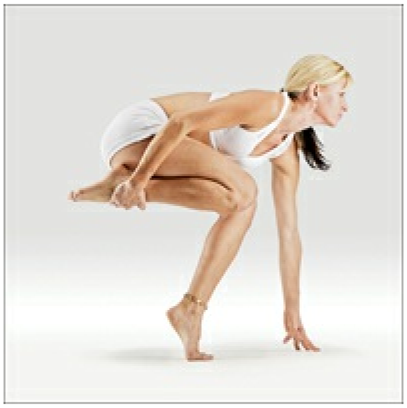

Both Hands to Foot Tip Toe Half Standing Wind Relieving Intense Stretch Pose

Dwi Hasta Pada Prapada Ardha Stiti Vayu Muktyuttonasana

(DWI-huh-stuh PUH-duh PRUH-puh-duh UHR-duh STI-ti VAH-yu muk-to-NAHS-uh-nuh)

Also Known As: Dwi Hasta Pada Prapada Ardha Nindra Vayu Muktyuttonasana
Pose Type: standing one-legged balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं यं हं ॐ

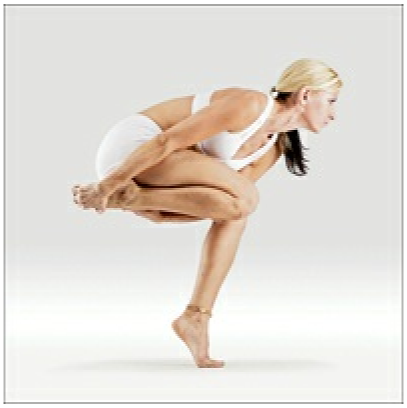

ONE-LEGGED SQUATS: TWISTS—GRABBING ONTO THE FOOT, LEG STRAIGHT IN FRONT

One-Legged Pose Dedicated to Yogi Shankara

Eka Pada Shankarasana

(EY-kuh PUH-duh shunk-uhr-AHS-uh-nuh)

Modification: forehead to the shin

Pose Type: standing one-legged balance, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

● ● ● ● ●

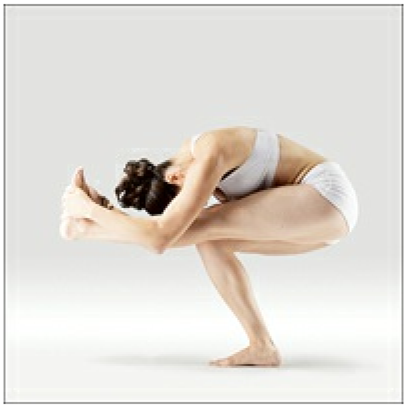

Revolved One-Legged Pose Dedicated to Yogi Shankara

Parivritta Eka Pada Shankarasana

(puh-ri-VRIT-tuh EY-kuh PUH-duh shunk-uhr-AHS-uh-nuh)

Modification: twisting to the inside of the leg

Pose Type: standing one-legged balance, forward bend, twist, side bend

Drishti Point: Urdhva or Antara Drishti (up to the sky)

● ●

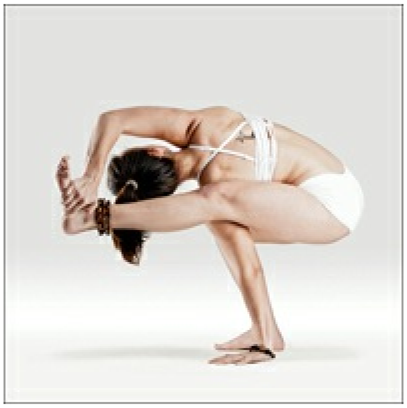

Revolved One-Legged Pose Dedicated to Yogi Shankara

Parivritta Eka Pada Shankarasana

(puh-ri-VRIT-tuh EY-kuh PUH-duh shunk-uhr-AHS-uh-nuh)

Modification: twisting to the outside of the leg
Pose Type: standing one-legged balance, forward bend, twist, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

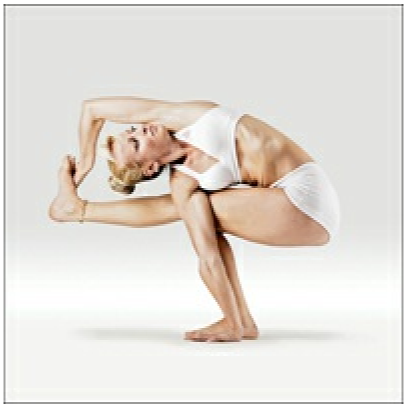

ONE-LEGGED SQUATS: FOOT ON TOP OF THE KNEE

One-Legged Fierce Pose 1

Eka Pada Utkatasana 1

(EY-kuh PUH-duh ut-kuh-TAHS-uh-nuh)

Also Known As: One-Legged Prayer Fierce Pose (Eka Pada Namaskar Utkatasana)
Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing one-legged balance
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

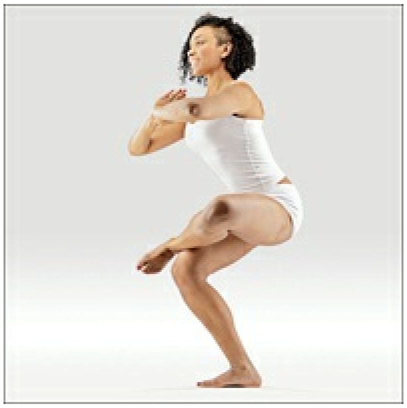

One-Legged Fierce Pose 2

Eka Pada Utkatasana 2

(EY-kuh PUH-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Both Hands to Ankle One-Legged Fierce Pose (Dwi Hasta Kulpa Eka Pada Utkatasana)
Modification: both hands to the ankle
Pose Type: standing one-legged balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

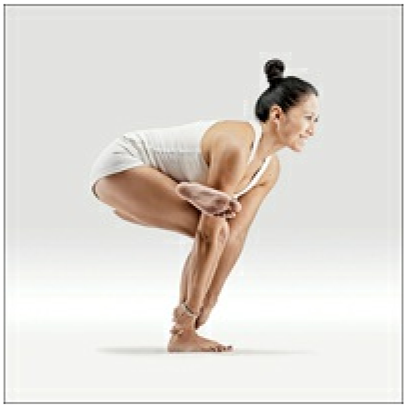

One-Legged Fierce Pose 3

Eka Pada Utkatasana 3

(EY-kuh PUH-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Arms Spread One-Legged Fierce Pose (Prasarita Hasta Eka Pada Utkatasana)
Modification: arms straight out to the sides, forward bend
Pose Type: standing one-legged balance, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

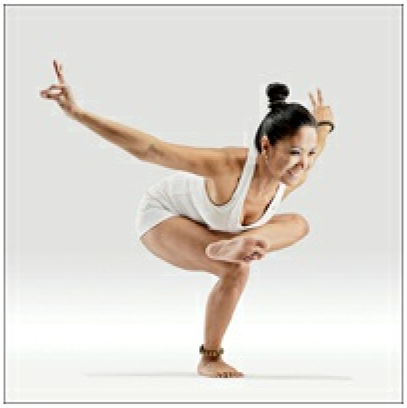

One-Legged Fierce Pose 4

Eka Pada Utkatasana 4

(EY-kuh PUH-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Hands Bound One-Legged Fierce Pose (Baddha Hasta Eka Pada Utkatasana)
Modification: hands bound behind the back, forward bend
Pose Type: standing one-legged balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

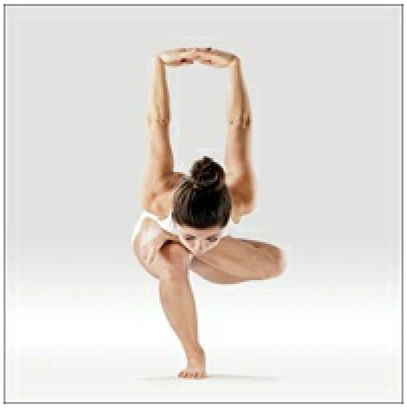

ONE-LEGGED SQUATS: HALF LOTUS

Half Bound Lotus Fierce Pose

Ardha Baddha Padma Utkatasana

(UHR-duh BUH-duh PUHD-muh ut-kuh-TAHS-uh-nuh)

Modification: hips low
Pose Type: standing one-legged balance, binding
Drishti Point: Hastagrai or Hastagre (hands)
ल व

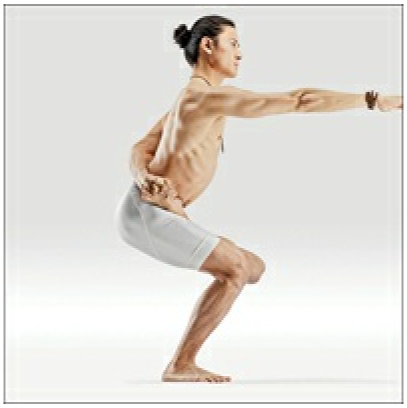

STANDING: FORWARD BEND—FOOT TO KNEE & FOOT TO ARMPIT

Baby Cradle Pose in Intense Stretch Pose

Hindolasana in Uttanasana

(hin-do-LAHS-uh-nuh in ut-tahn-AHS-uh-nuh)
Also Known As: Baby Cradle Pose in Full Forward Bend
Modification: foot behind the knee of the standing leg
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य ह ॐ

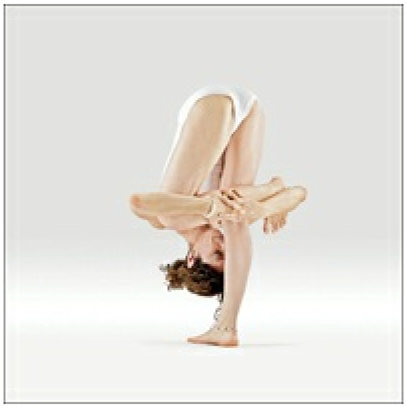

Standing Bound Yogic Staff Pose

Stiti Baddha Yoganandasana

(STI-ti BUH-duh yo-gah-nuhn-DAHS-uh-nuh)

Also Known As: Nindra Baddha Yoganandasana
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व ह

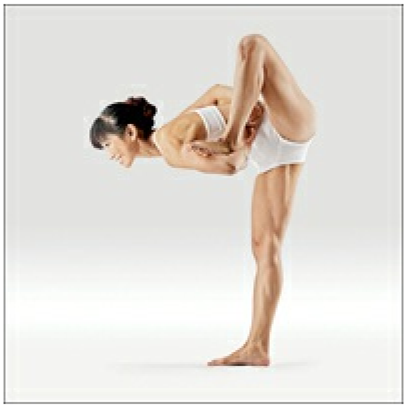

Yogic Staff Intense Stretch Pose

Yogadananda Uttanasana

(yo-gah-nuhn-duh ut-tahn-AHS-uh-nuh)
Also Known As: Yogic Staff Full Forward Bend
Pose Type: standing one-legged balance, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)
ल व र य ह

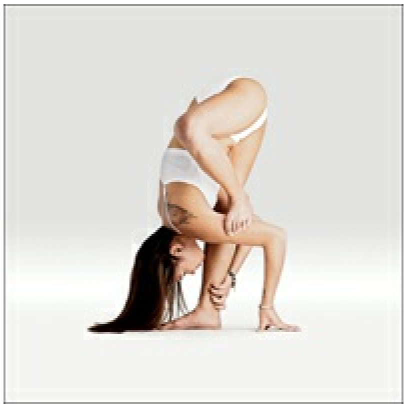

STANDING HALF FORWARD BEND: LEGS CROSSED

Hand Position of the Pose Dedicated to Garuda in Sideways Half Intense Stretch Pose

Hasta Garudasana in Parshva Ardha Uttanasana

(HUH-stuh guh-ru-DAHS-uh-nuh in PAHRSH-vuh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Hand Position of the Pose Dedicated to Garuda in Sideways Legs Crossed Half Forward Bend

Modification: legs crossed

Pose Type: standing, forward bend, twist

Drishti Point: Hastagrai or Hastagre (hands)

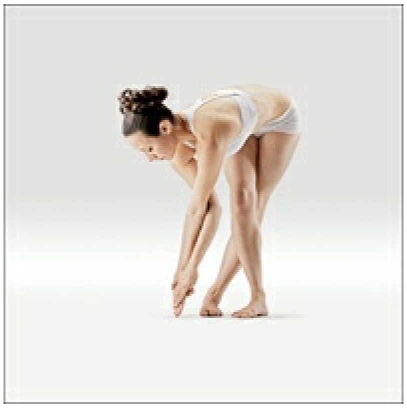

STANDING HALF FORWARD BEND: ONE KNEE BENT

Reverse Prayer Uneven Legs Half Intense Stretch Pose

Viparita Namaskar Vishama Pada Ardha Uttanasana

(vi-puh-REE-tuh nuh-muhs-KAHR VISH-uh-muh PUH-duh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Back of the Body Prayer Uneven Legs Half Intense Stretch Pose (Paschima Namaskara Vishama Pada Ardha Uttanasana), Reverse Prayer Uneven Legs Half Forward Bend

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

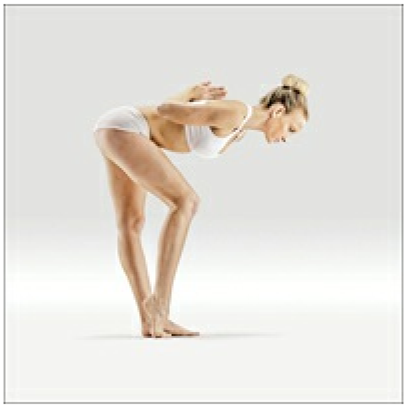

Bound Uneven Legs Half Intense Stretch Pose

Baddha Vishama Pada Ardha Uttanasana

(BUH-duh VISH-uh-muh PUH-duh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Bound Uneven Legs Half Forward Bend
Pose Type: standing, forward bend, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

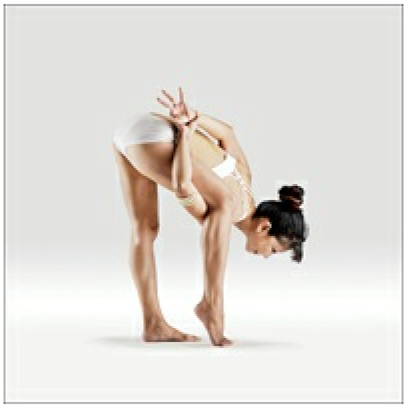

STANDING HALF FORWARD BEND: ONE KNEE BENT—TWISTS & BINDING

Revolved Uneven Legs Half Intense Stretch Pose

Parivritta Vishama Pada Ardha Uttanasana

(puh-ri-VRIT-tuh VISH-uh-muh PUH-duh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Revolved Uneven Legs Half Forward Bend
Modification: one hand to the floor, other hand up to the sky; looking down
Pose Type: standing, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

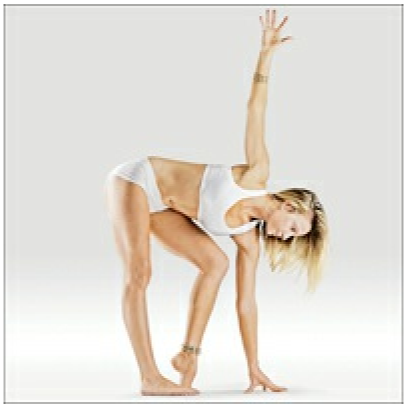

Reverse Prayer Revolved Uneven Legs Half Intense Stretch Pose

Viparita Namaskar Parivritta Vishama Pada Ardha Uttanasana

(vi-puh-REE-tuh nuh-muhs-KAHR puh-ri-VRIT-tuh VISH-uh-muh PUH-duh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Back of the Body Prayer Revolved Uneven Legs Half Intense Stretch Pose (Paschima Namaskara Ardha Parivritta Uttanasana), Reverse Prayer Revolved Uneven Legs Half Forward Bend

Pose Type: standing, forward bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

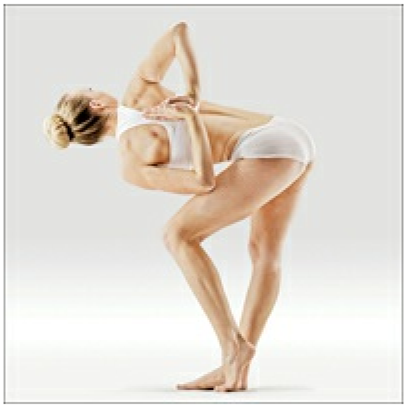

Hands Bound Revolved Uneven Legs Half Intense Stretch Pose

Baddha Hasta Parivritta Vishama Pada Ardha Uttanasana

(BUH-duh HUH-stuh puh-ri-VRIT-tuh VISH-uh-muh PUH-duh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Hands Bound Revolved Uneven Legs Half Forward Bend

Pose Type: standing, forward bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

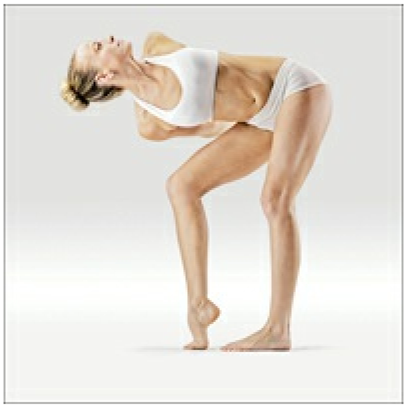

Bound Revolved Half Intense Stretch Pose

Baddha Parivritta Ardha Uttanasana

(BUH-duh puh-ri-VRIT-tuh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Bound Revolved Half Forward Bend

Pose Type: standing, forward bend, twist, binding

Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल य

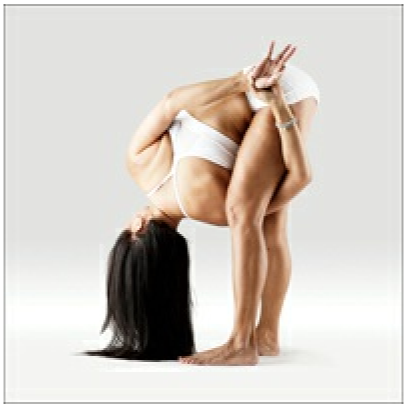

STANDING HALF FORWARD BEND: HANDS OFF THE FLOOR

Arms Extended Half Intense Stretch Pose

Utthita Hasta Ardha Uttanasana

(UT-ti-tuh HUH-stuh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Arms Extended Half Forward Bend

Modification: palms pressed together

Pose Type: standing, forward bend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

ल

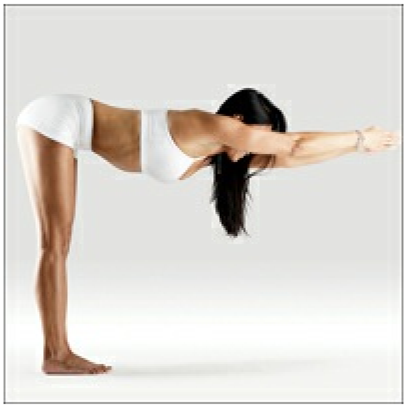

STANDING HALF FORWARD BEND: HANDS ON THE FLOOR

Half Foot Big Toe Pose

Ardha Padangushtasana

(UHR-duh puhd-ahng-goosh-TAHS-uh-nuh)

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

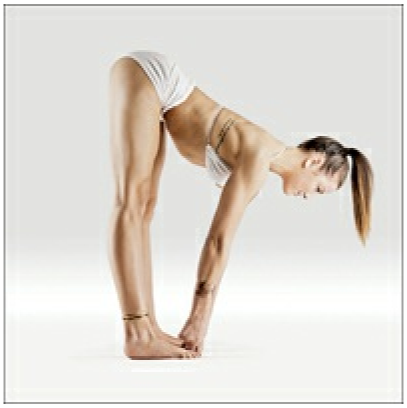

Sideways Half Intense Stretch Pose

Parshva Ardha Uttanasana

(PAHRSH-vuh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Sideways Half Forward Bend

Modification: fingers interlocked, palms to the floor on the outside edge of the foot

Pose Type: standing, forward bend, side bend, twist

Drishti Point: Hastagrai or Hastagre (hands)

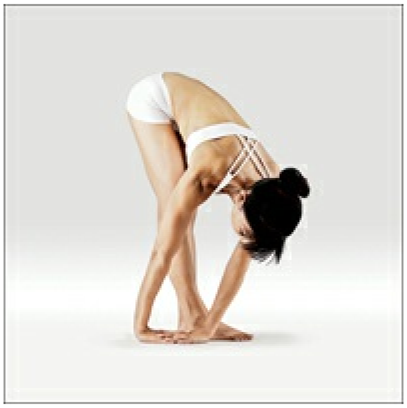

STANDING HALF FORWARD BEND: HANDS ON THE FLOOR—HEELS UP

Tip Toe Half Intense Stretch Pose

Prapada Ardha Uttanasana

(PRUH-puh-duh UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Tip Toe Half Forward Bend and Downward Facing Tree Pose Prep. (Adho Mukha Vrikshasana Prep.)
Pose Type: standing, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

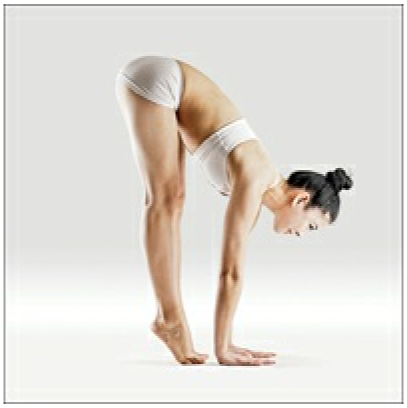

Tip Toe Intense Stretch Pose

Prapada Uttanasana

(PRUH-puh-duh ut-tahn-AHS-uh-nuh)

Also Known As: Tip Toe Full Forward Bend
Modification: arms straight and pointing to the back, both arms on the inside of the legs
1. head up
2. forehead toward the shins
Pose Type: standing, forward bend
Drishti Point: 1. Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) 2. Nasagrai or Nasagre (nose)

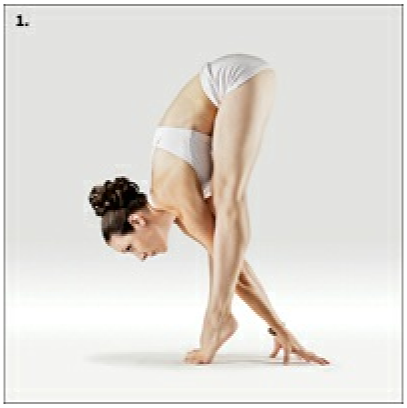

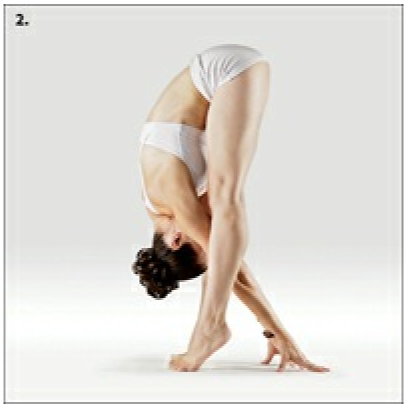

STANDING FULL FORWARD BEND: BALLET TOES

Intense Ankle Stretch Intense Stretch Pose 1

Uttana Kulpa Uttanasana 1

(ut-TAHN-uh KUL-puh ut-tahn-AHS-uh-nuh)

Also Known As: Intense Ankle Stretch Forward Bend
Modification: legs together
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

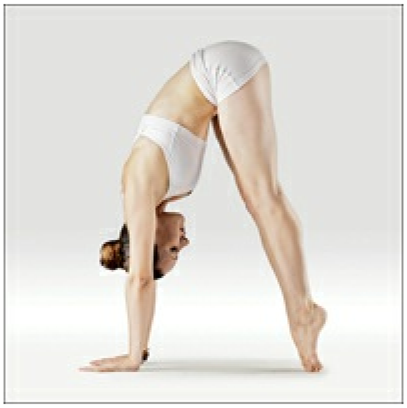

Intense Ankle Stretch Intense Stretch Pose 2

Uttana Kulpa Uttanasana 2

(ut-TAHN-uh KUL-puh ut-tahn-AHS-uh-nuh)

Also Known As: Intense Ankle Stretch Forward Bend
Modification: scissor legs
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

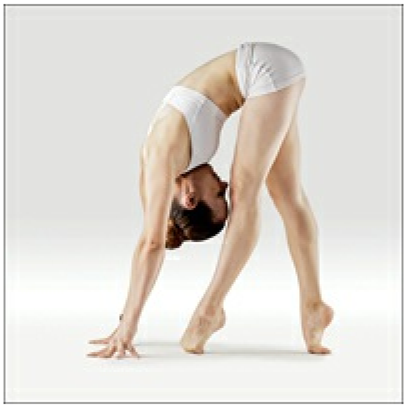

STANDING FULL FORWARD BEND: ONE-LEGGED

Extended One Hand to Foot Intense Stretch Pose 1

Utthita Eka Hasta Pada Uttanasana 1

(UT-ti-tuh EY-kuh HUH-stuh PUH-duh ut-tahn-AHS-uh-nuh)

Also Known As: Extended One Hand to Foot Full Forward Bend 1
Modification: hand grabbing the foot on the same side
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

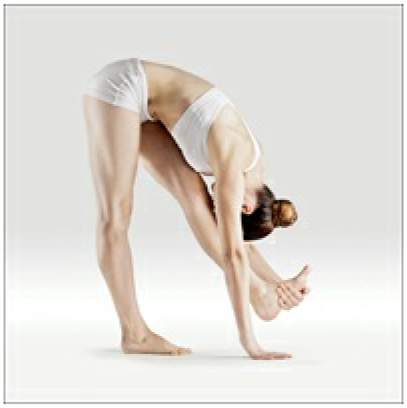

Extended One Hand to Foot Intense Stretch Pose 2

Utthita Eka Hasta Pada Uttanasana 2

(UT-ti-tuh EY-kuh HUH-stuh PUH-duh ut-tahn-AHS-uh-nuh)

Also Known As: Extended One Hand to Foot Full Forward Bend 2
Modification: hand grabbing the foot on the opposite side
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

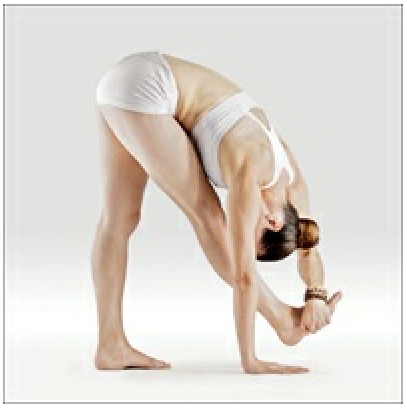

STANDING FULL FORWARD BEND: HANDS TO THE FLOOR & HANDS TO THE BACK OF THE LEGS

Intense Stretch Pose

Uttanasana

(ut-tahn-AHS-uh-nuh)

Also Known As: Full Forward Bend
Modification: palms to the floor, fingers pointing to the back
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

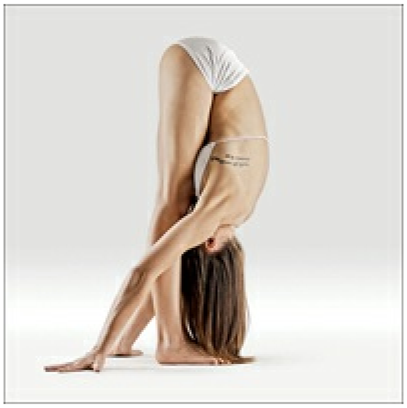

Intense Stretch Pose

Uttanasana

(ut-tahn-AHS-uh-nuh)

Also Known As: Full Forward Bend and Standing Turtle Pose (Nindra Kurmasana)
Modification: feet hip width apart, grabbing onto the calves, elbows bent
Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)
लं

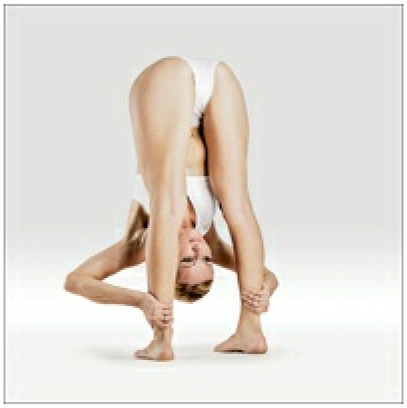

Intense Stretch Pose

Uttanasana

(ut-tahn-AHS-uh-nuh)

Also Known As: Full Forward Bend and Locked Elbows Standing Intense Stretch Pose (Baddha Padahastasana)
Modification: grabbing onto triceps behind the calves
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं यं हं

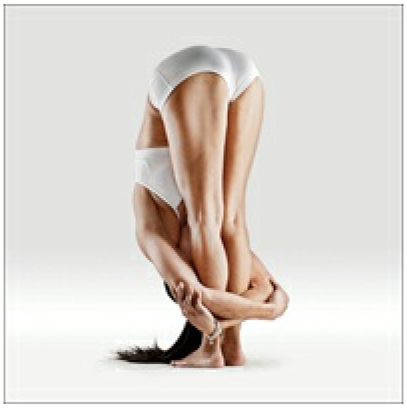

STANDING FULL FORWARD BEND: GRABBING ONTO THE FEET

Foot Big Toe Pose

Padangushtasana

(puh-ahng-goosh-TAHS-uh-nuh)

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)
लं

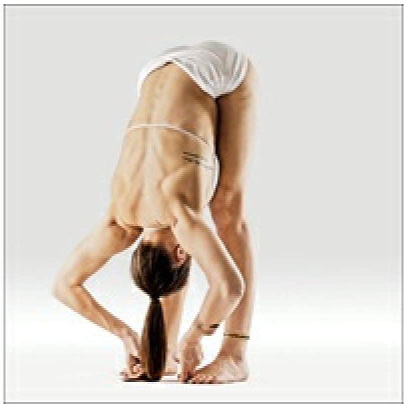

Hand Under Foot Pose

Pada Hastasana

(PUH-duh huh-STAHS-uh-nuh)
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं हं

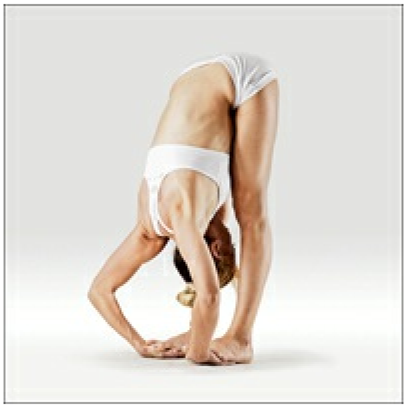

Intense Stretch Pose

Uttanasana *(ut-tahn-AHS-uh-nuh)*
Also Known As: Full Forward Bend
Modification: arms crossed, grabbing onto the outside edges of the feet, half forward bend
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet) or Nasagrai or Nasagre (nose)
लं यं

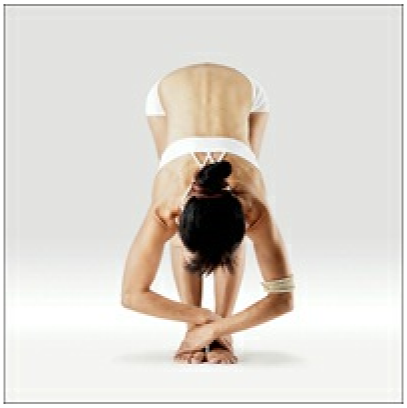

Forward Bend

Uttanasana

(ut-tahn-AHS-uh-nuh)

Also Known As: Full Forward Bend and Hands to Feet Pose Modification (Pada Hastasana)

Modification: feet hip width apart, arms crossed, grabbing onto the outside edges of the feet, elbows moving toward the feet

1. half forward bend
2. full forward bend

Pose Type: standing, forward bend

Drishti Point: 1. Padayoragrai or Padayoragre (toes/feet) 2. Nasagrai or Nasagre (nose)

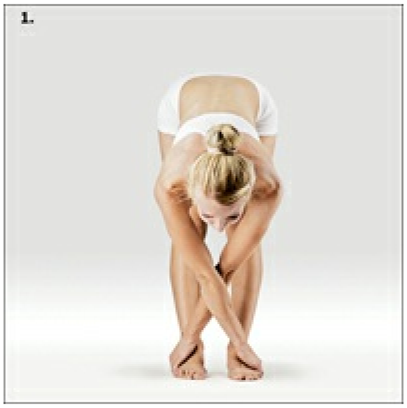

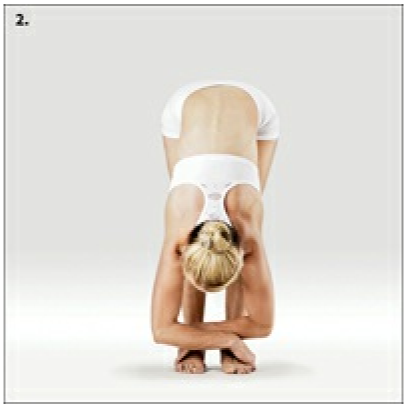

STANDING FULL FORWARD BEND: UNSUPPORTED

Intense Stretch Pose

Uttanasana

(ut-tahn-AHS-uh-nuh)

Also Known As: Full Forward Bend

Modification: grabbing onto the triceps, swaying from side to side

Pose Type: standing, forward bend, side bend

Drishti Point: Nasagrai or Nasagre (nose)

लं

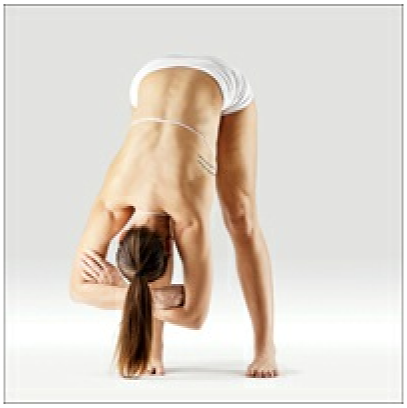

Intense Stretch Pose

Uttanasana

(ut-tahn-AHS-uh-nuh)

Also Known As: Full Forward Bend and Unsupported Forward Stretch Pose (Niralamba Uttanasana)

Modification: arms straight out to the sides behind the back, forehead to the shins

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

लं यं हं

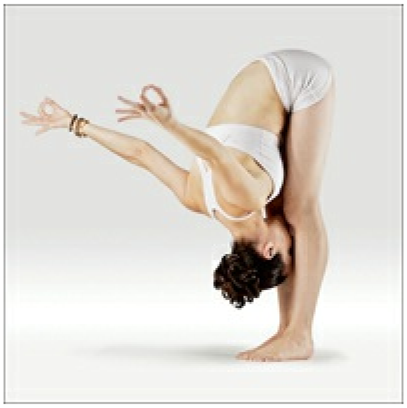

Hands Bound Forward Bend Modification

Baddha Hasta Uttanasana

(BUH-duh HUH-stuh ut-tahn-AHS-uh-nuh)

Also Known As: Hands Bound Full Forward Bend and Unsupported Forward Stretch Pose (Niralamba Uttanasana)
Modification: knees bent, chest to the quadriceps
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल य ह

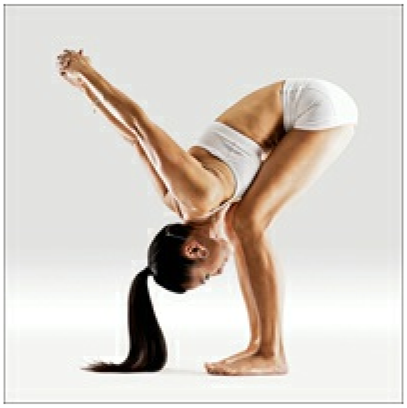

STANDING FULL FORWARD BEND: HANDS TO THE FLOOR—ARMS OPEN WIDE

Intense Stretch Pose

Uttanasana

(ut-tahn-AHS-uh-nuh)

Also Known As: Legs Crossed Full Forward Bend
Modification: legs crossed, forehead to the shins, hands to the floor in line with the feet, elbows bent
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व य ॐ

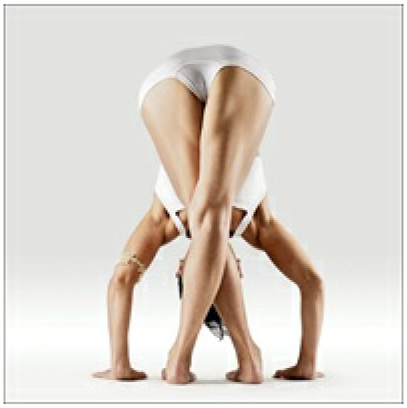

Sideways Intense Stretch

Parshva Uttanasana

(PAHRSH-vuh ut-tahn-AHS-uh-nuh)

Also Known As: Sideways Full Forward Bend

Modification: one hand grabbing onto the opposite ankle, fingertips of the other hand to the floor

Pose Type: standing, forward bend, twist

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

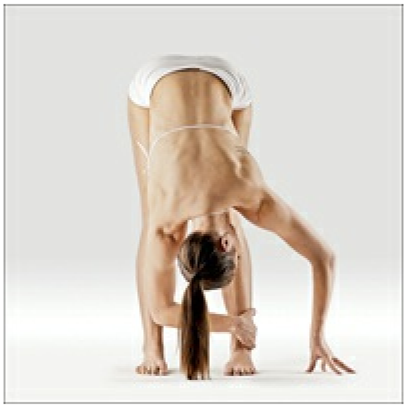

Sideways Intense Stretch

Parshva Uttanasana

(PAHRSH-vuhut-tahn-AHS-uh-nuh)

Also Known As: Sideways Full Forward Bend and Sideward Forward Stretch (Parshva Bhaga Uttanasana)

Modification: fingers interlocked, palms to the floor on the outside edge of the foot

Pose Type: standing, forward bend, twist

Drishti Point: Nasagrai or Nasagre (nose)

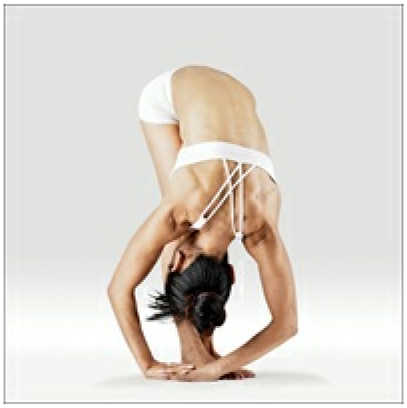

Revolved Intense Stretch Prayer Pose

Parivritta Uttana Anjalyiasana

(puh-ri-VRIT-uh ut-TAHN-uh uhn-juhl-YAHS-uh-nuh)

Also Known As: Revolved Full Forward Bend Prayer Pose
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

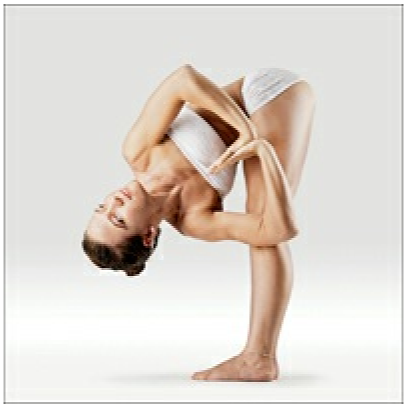

Revolved Intense Stretch

Parivritta Uttanasana

(puh-ri-VRIT-tuh ut-tahn-AHS-uh-nuh)

Also Known As: Revolved Full Forward Bend
Modification: grabbing the outside edge of the feet, feet hip width apart
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

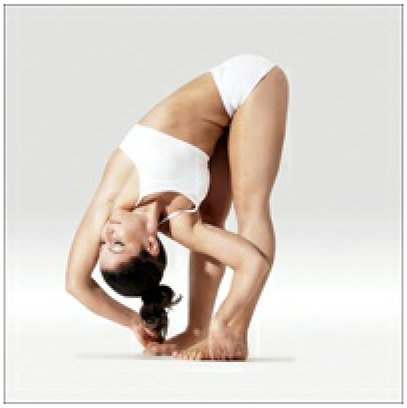

Revolved Intense Stretch

Parivritta Uttanasana

(puh-ri-VRIT-tuh ut-tahn-AHS-uh-nuh)

Also Known As: Revolved Full Forward Bend
Modification: feet wide; one hand grabbing onto the ankle, back of the other hand to the floor in front of the opposite foot
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

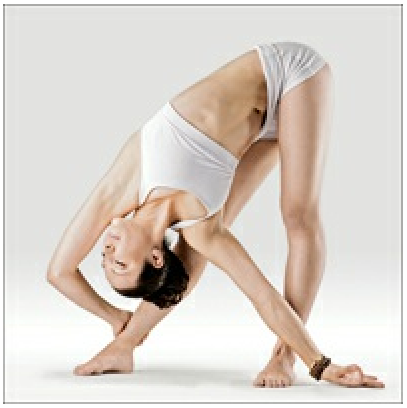

Revolved Intense Stretch

Parivritta Uttanasana

(puh-ri-VRIT-tuh ut-tahn-AHS-uh-nuh)

Also Known As: Revolved Legs Crossed Full Forward Bend
Modification: feet wide, legs crossed, feet flat on the floor
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

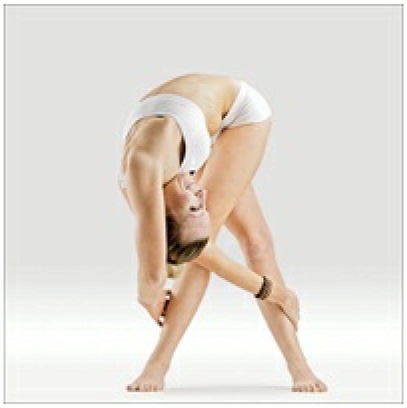

Firefly Pose 2 A

Tittibhasana 2 A

(ti-ti-BAHS-uh-nuh)

Also Known As: Intense Stretch Pose Stork Modification (Uttanasana), Standing Bound Arms Head Between Knees Pose (Utthita Baddha Hasta Janu Shirshasana)
Modification: arms wrapped around the legs, fingers interlocked behind the head
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

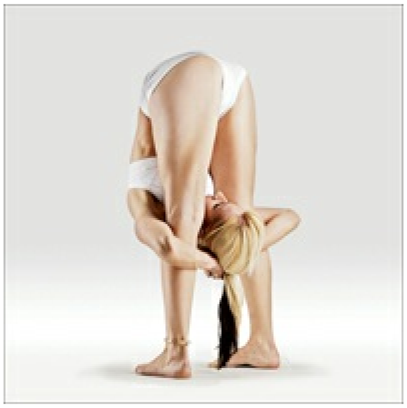

Firefly Pose 2 B

Tittibhasana 2 B

(ti-ti-BAHS-uh-nuh)

Also Known As: Inverted Firefly Prayer Pose (Viparita Tittibha Anjali Asana)
Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं

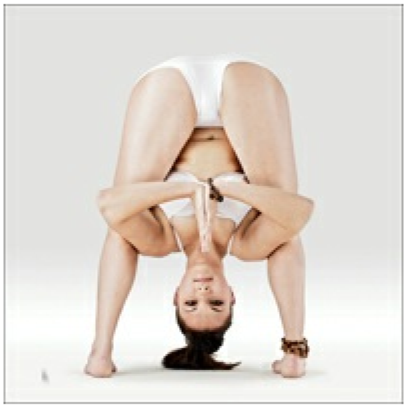

Complete Firefly Pose 2

Paripurna Tittibhasana 2

(puh-ri-POOR-nuh ti-ti-BAHS-uh-nuh)

Also Known As: Firefly Pose B (Tittibhasana B), Inverted Firefly Pose (Viparita Tittibhasana), Standing Firefly Pose Modification (Utthita Tittibhasana), Bound Firefly Pose, Standing Tortoise Pose (Nindra Kurmasana)
Modification: hands bound
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

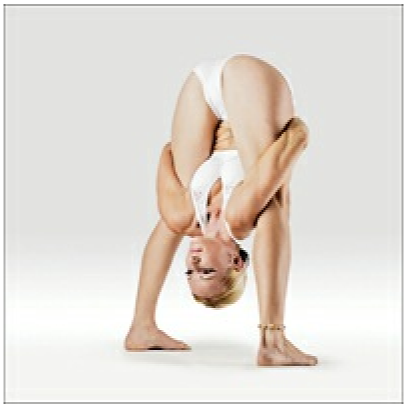

Firefly Pose 3 A Prep.

Tittibhasana 3 A Prep.

(ti-ti-BAHS-uh-nuh)

Modification: hands on the ankles, heels up
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

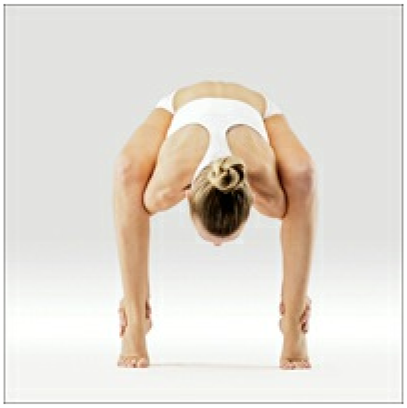

Firefly Pose 3 A

Tittibhasana 3 A

(ti-ti-BAHS-uh-nuh)

Also Known As: Firefly Pose C Prep. (Tittibhasana C Prep.)
Modification: hands on the ankles, one foot lifted
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व

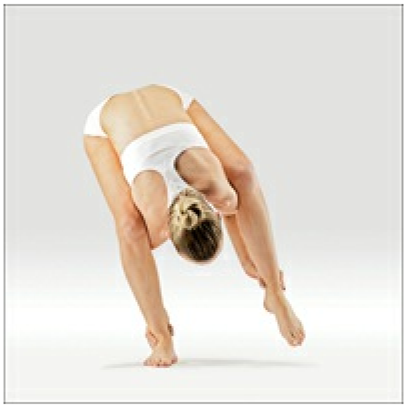

Firefly Pose 4 A

Tittibhasana 4 A

(ti-ti-BAHS-uh-nuh)

Also Known As: Firefly Pose D (Tittibhasana D), Standing Firefly Pose (Utthita Tittibhasana)
Modification: fingers interlocked in front of the ankles
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)
ल व ह

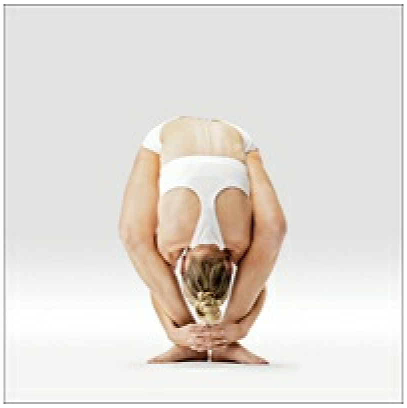

Firefly Pose 4 B

Tittibhasana 4 B

(ti-ti-BAHS-uh-nuh)

Also Known As: Inverted Both Legs Prayer Pose (Viparita DwiPada Anjaliasana)
Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

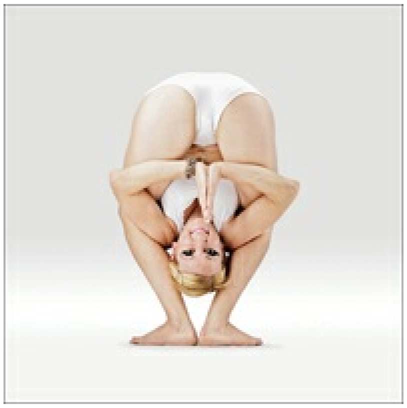

Complete Firefly Pose 4

Paripurna Tittibhasana 4

(puh-ri-POOR-nuh ti-ti-BAHS-uh-nuh)

Also Known As: Inverted Bound Legs Pose (Viparita DwiPada Baddhasana)
Modification: hands bound
Pose Type: standing, forward bend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

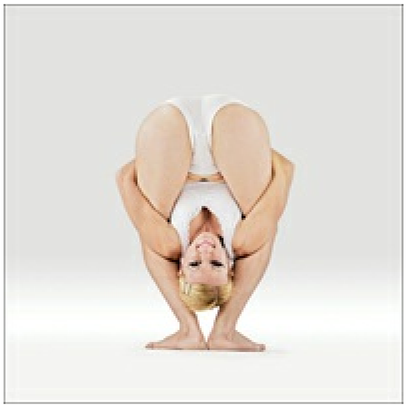

Book Stand Pose

Grantadara

(gruhn-tah-DAH-ruh)

Also Known As: Chikkyasana
Modification: hands to the head, palms facing up, looking to the side
Pose Type: standing, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

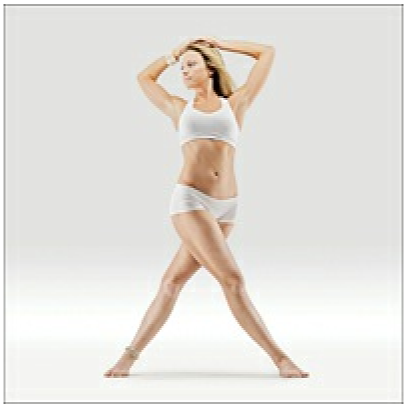

Revolved Feet Spread Mountain Pose

Parivritta Prasarita Pada Tadasana

(puh-ri-VRIT-tuh pruh-SAH-ri-tuh PUH-duh tuh-DAHS-uh-nuh)

Modification: arms straight out to the sides
Pose Type: standing, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

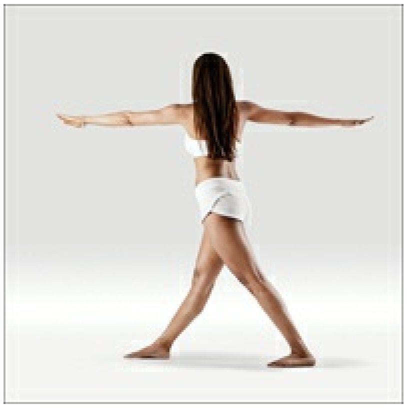

Feet Spread Mountain Pose in Hero Succession Series

Prasarita Pada Tadasana in Vira Parampara

(pruh-SAH-ri-tuh PUH-duh tuh-DAHS-uh-nuh in VEER-uh puh-ruhm-puh-RAH)

Modification: backbend, hands on the calf muscles

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

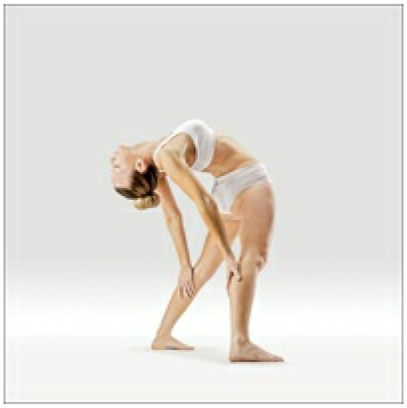

LEGS WIDE: ARMS UP OVER THE HEAD

Easy Feet Spread Upward Hands Pose

Sukha Prasarita Pada Urdhva Hastasana

(SUK-kuh pruh-SAH-ri-tuh PUH-duh OORD-vuh huh-STAHS-uh-nuh)

Also Known As: Part of Hero Succession Series (Vira Parampara)

Modification: palms together, looking up

Pose Type: standing, mild backbend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

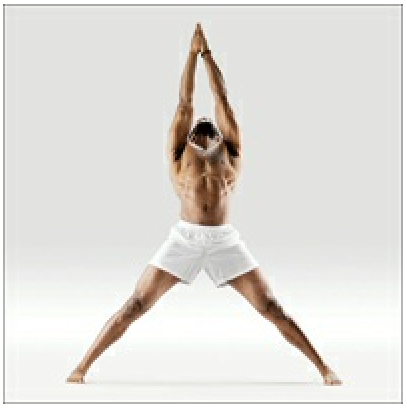

Feet Spread Upward Hands Pose

Prasarita Pada Urdhva Hastasana

(pruh-SAH-ri-tuh PUH-duh OORD-vuh huh-STAHS-uh-nuh)

Also Known As: Equal Angle Pose (Sama Konasana)
Modification: legs open extremely wide; palms together, looking straight ahead
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

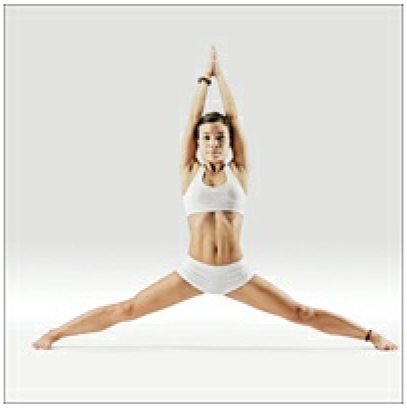

Angle Pose

Konasana

(ko-NAHS-uh-nuh)

Also Known As: Triangle Pose (Konasana)
Modification: fingers interlocked, palms together; side bend
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky), Nasagrai or Nasagre (nose)

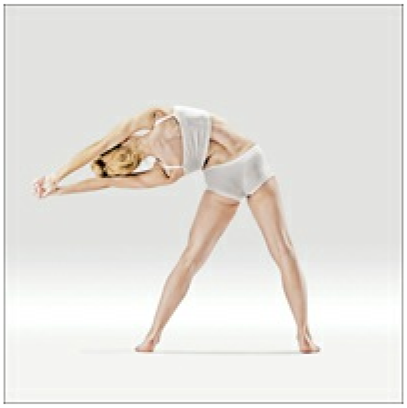

Book Stand Pose with Upward Bound Hands

Grantadara Urdhva Baddha Hastasana

(gruhn-tah-DAH-ruh OORD-vuh BUH-duh HUH-STAHS-uh-nuh)

Also Known As: Chikkiasana Urdhva Baddha Hastasana
Modification: fingers interlocked, palms facing up, looking up
Pose Type: standing
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

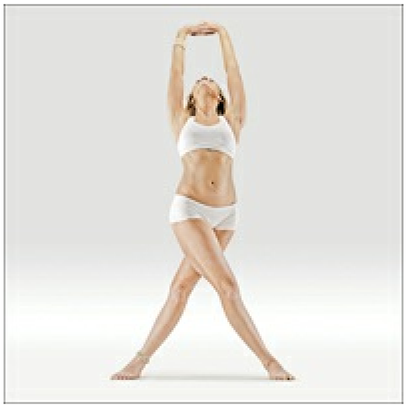

KALI SQUAT: HEELS DOWN—ARMS UP OVER THE HEAD

Lotus Hand Seal in Upward Hands Pose Dedicated to Goddess Kali

Padma Mudra Urdhva Hasta Kalyasana

(PUHD-muh MU-druh OORD-vuh HUH-stuh kahl-YAHS-uh-nuh)

Modification: heels down; arms extended to the sky, fingertips open
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

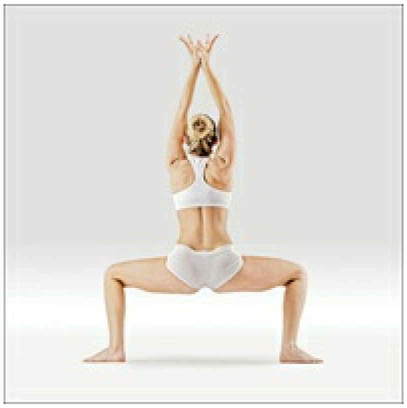

Sideways Hands Bound Pose Dedicated to Goddess Kali

Parshva Baddha Hasta Kalyasana

(PAHRSH-vuh BUH-duh HUH-stuh kahl-YAHS-uh-nuh)

Modification: side bend, fingers interlocked
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व य

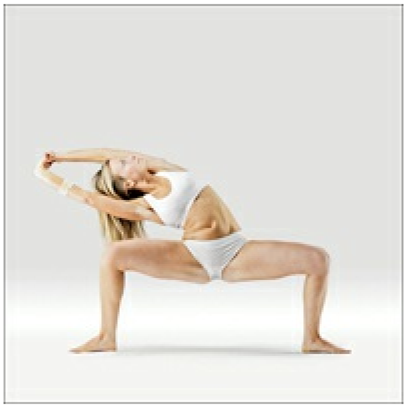

KALI SQUAT: HEELS DOWN—ARMS AT HEAD HEIGHT

Pose Dedicated to Goddess Kali

Kalyasana

(kahl-YAHS-uh-nuh)

Modification: elbows bent, palms facing the sky, heels down
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)
ल व

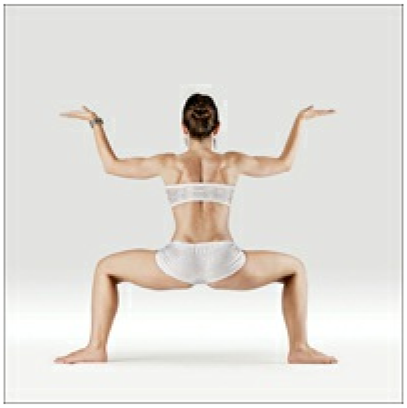

Hand Position of the Pose Dedicated to Garuda in Pose Dedicated to Goddess Kali

Hasta Garudasana in Kalyasana

(HUH-stuh guh-ru-DAHS-uh-nuh in kahl-YAHS-uh-nuh)

Pose Type: standing

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

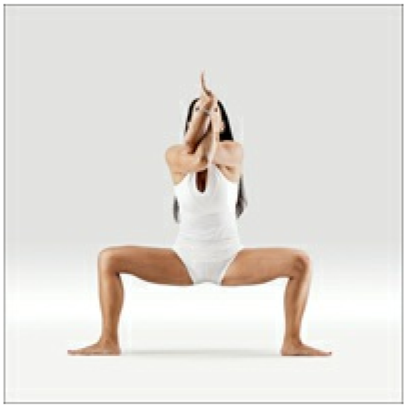

KALI SQUAT: HEELS DOWN—ARMS BEHIND

Hands Bound Pose Dedicated to Goddess Kali

Baddha Hasta Kalyasana

(BUH-duh HUH-stuh kahl-YAHS-uh-nuh)

Modification: head rolling back, heels down

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

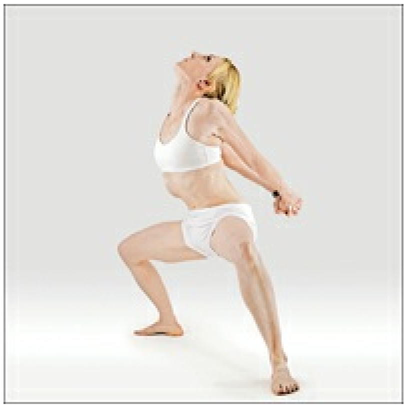

Bound Pose Dedicated to Goddess Kali

Baddha Kalyasana

(BUH-duh kahl-YAHS-uh-nuh)

Pose Type: standing, binding

Drishti Point: Nasagrai or Nasagre (nose)

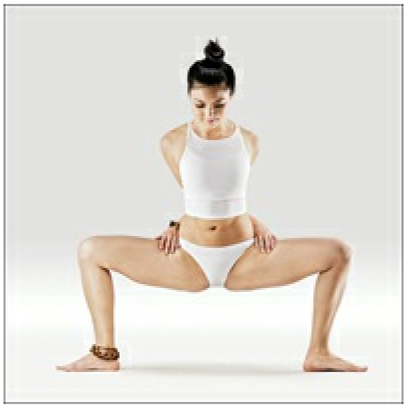

KALI SQUAT: HEELS DOWN—ONE ARM SIDE BENDS

Pose Dedicated to Goddess Kali

Kalyasana

(kahl-YAHS-uh-nuh)

Modification: one hand to the forehead, arm to the knee on the same side

Pose Type: standing

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

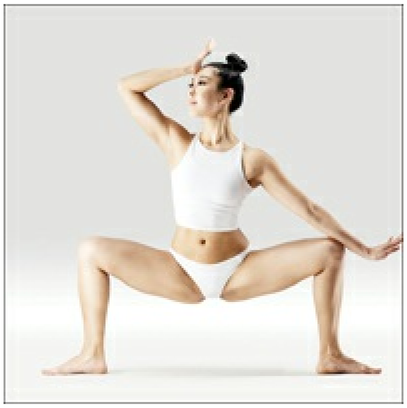

Pose Dedicated to Goddess Kali

Kalyasana

(kahl-YAHS-uh-nuh)

Modification: one arm straight and parallel to the floor, other arm bent, palm facing up
Pose Type: standing
Drishti Point: Hastagrai or Hastagre (hands)

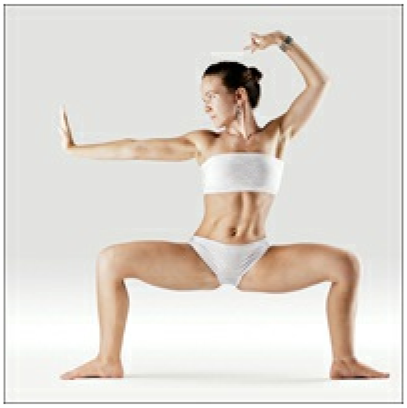

Sideways Pose Dedicated to Goddess Kali

Parshva Kalyasana

(PAHRSH-vuh kahl-YAHS-uh-nuh)

Modification: leaning to one side, one arm out to the side—elbow slightly bent, palm facing up, other hand to the face—palm down
Pose Type: standing, side bend
Drishti Point: Hastagrai or Hastagre (hands)

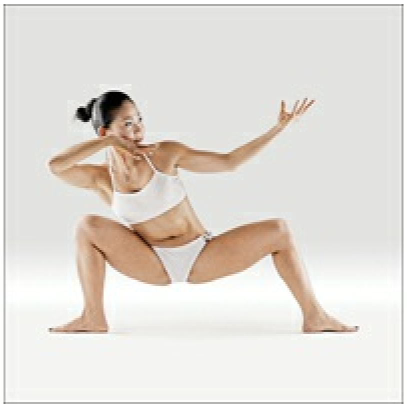

Sideways Half Bound Pose Dedicated to Goddess Kali

Parshva Ardha Baddha Kalyasana

(PAHRSH-vuh UHR-duh BUH-duh kahl-YAHS-uh-nuh)

Modification: forearm to the knee on the same side
Pose Type: standing, side bend, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

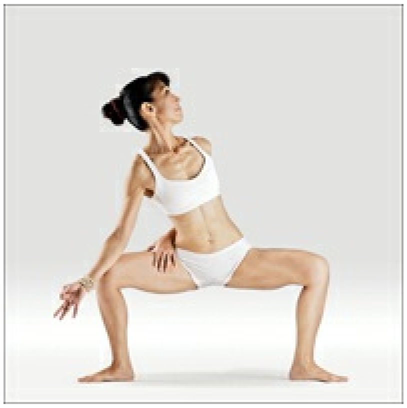

Sideways Pose Dedicated to Goddess Kali

Parshva Kalyasana

(PAHRSH-vuh kahl-YAHS-uh-nuh)

Modification: one hand to the floor, other arm up to the sky
Pose Type: standing, side bend
Drishti Point: Hastagrai or Hastagre (hands)

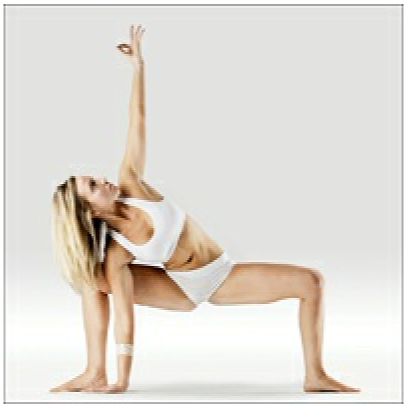

KALI SQUAT: HEELS DOWN—FORWARD BEND

Bound Pose Dedicated to Goddess Kali

Baddha Kalyasana

(BUH-duh kahl-YAHS-uh-nuh)

Modification: forward bend

Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

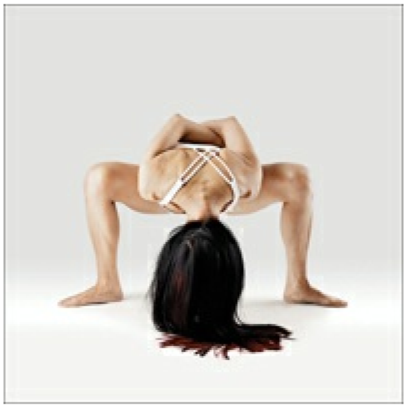

Pose Dedicated to Goddess Kali

Kalyasana

(kahl-YAHS-uh-nuh)
Modification: forward bend, arms straight and parallel to the floor, palms together
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

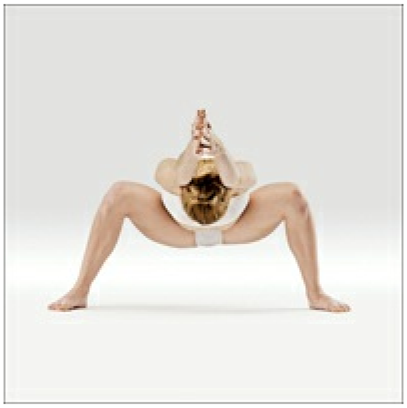

Pose Dedicated to Goddess Kali

Kalyasana

(kahl-YAHS-uh-nuh)
Modification: forward bend, arms crossed, hands to the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

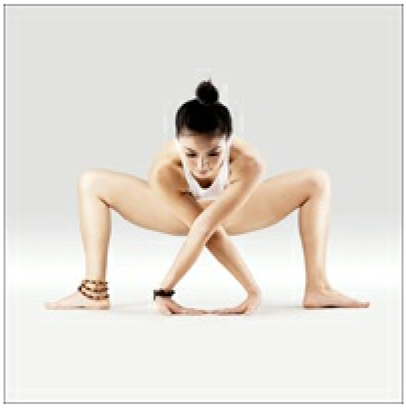

KALI SQUAT: HEELS DOWN—FORWARD BEND & FORWARD BEND TWISTS

Revolved Half Bound Pose Dedicated to Goddess Kali

Parivritta Ardha Baddha Kalyasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh kahl-YAHS-uh-nuh)

Modification: arm under the opposite leg
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

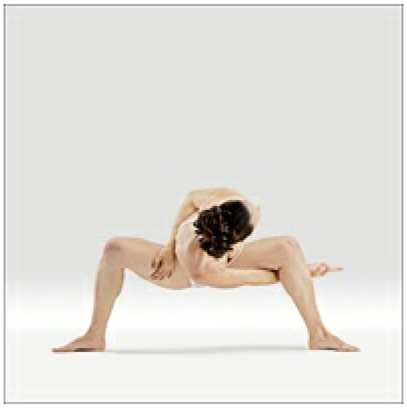

Revolved Pose Dedicated to Goddess Kali

Parivritta Kalyasana

(puh-ri-VRIT-tuh kahl-YAHS-uh-nuh)

Modification: hands to the knees
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

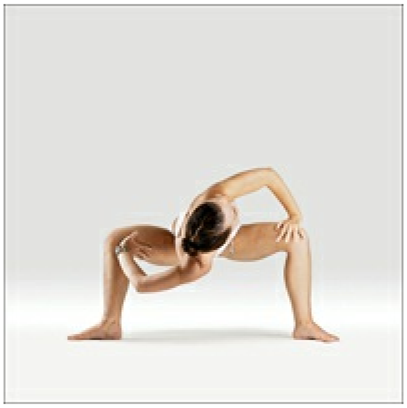

KALI SQUAT: HEELS UP—ARM POSITIONS

Tip Toe Pose Dedicated to Goddess Kali

Prapada Kalyasana

(PRUH-puh-duh kahl-YAHS-uh-nuh)

Modification: head down, elbows bent, back of the hands on the knees
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

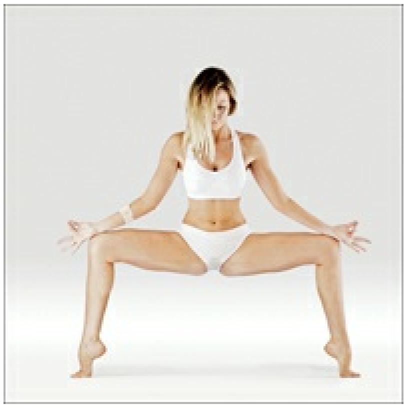

Lotus Hand Seal in Upward Hands Tip Toe Pose Dedicated to Goddess Kali

Padma Mudra in Urdhva Hasta Prapada Kalyasana

(PUHD-muh MU-druh in OORD-vuh HUH-stuh PRUH-puh-duh kahl-YAHS-uh-nuh)

Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)
लं वं

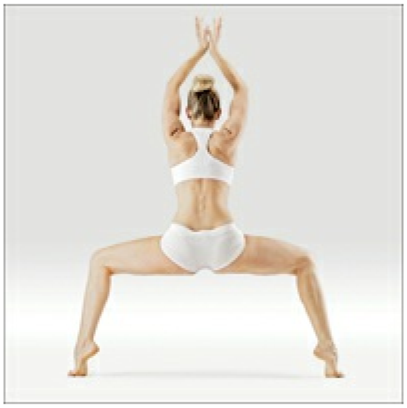

Tip Toe Pose Dedicated to Goddess Kali

Prapada Kalyasana

(PRUH-puh-duh kahl-YAHS-uh-nuh)

Modification: one hand behind the back in reverse prayer, one elbow on the knee
Pose Type: standing
Drishti Point: Hastagrai or Hastagre (hands)

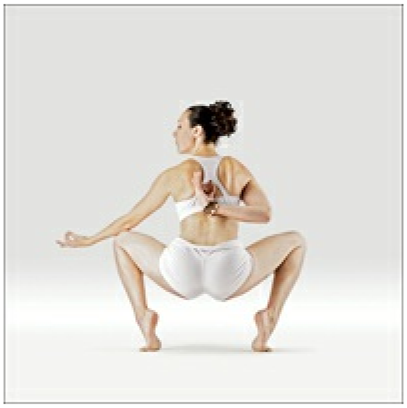

Reverse Prayer Tip Toe Pose Dedicated to Goddess Kali

Viparita Namaskar Prapada Kalyasana

(vi-puh-REE-tuh nuh-muhs-KAHR PRUH-puh-duh kahl-YAHS-uh-nuh)

Also Known As: Back of the Body Prayer Tip Toe Pose Dedicated to Goddess Kali (Paschima Namaskar Prapada Kalyasana)
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

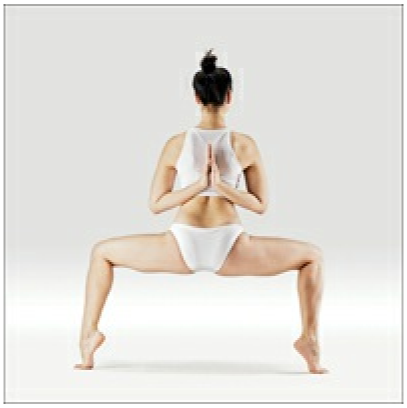

KALI SQUAT: HEELS UP—SIDE BENDS

Sideways Tip Toe Pose Dedicated to Goddess Kali

Parshva Prapada Kalyasana

(PAHRSH-vuh PRUH-puh-duh kahl-YAHS-uh-nuh)

Modification: one palm to the forehead, fingertips of the other hand to the floor
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

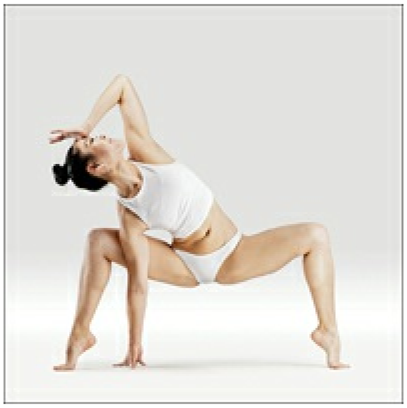

KALI SQUAT: ONE HEEL UP, ONE HEEL DOWN

Sideways Uneven Tip Toe Pose Dedicated to Goddess Kali

Parshva Vishama Prapada Kalyasana

(PAHRSH-vuh VISH-uh-muh PRUH-puh-duh kahl-YAHS-uh-nuh)

Modification: side bend
Pose Type: standing, side bend
Dristhi Point: Padhayoragrai or Padayoragre (toes/feet)

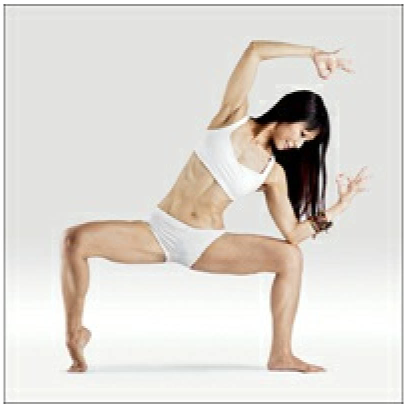

KALI SQUAT: HEELS UP—FORWARD BENDS

Tip Toe Pose Dedicated to Goddess Kali

Prapada Kalyasana

(PRUH-puh-duh kahl-YAHS-uh-nuh)

Modification: forward bend, arms crossed, hands to the floor, fingertips pointing toward each other
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

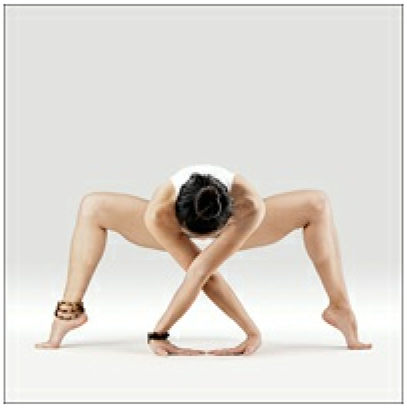

Tip Toe Pose Dedicated to Goddess Kali

Prapada Kalyasana

(PRUH-puh-duh kahl-YAHS-uh-nuh)

Modification: forward bend, arms crossed, hands to the floor, fingertips pointing away from each other
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

ल व ह

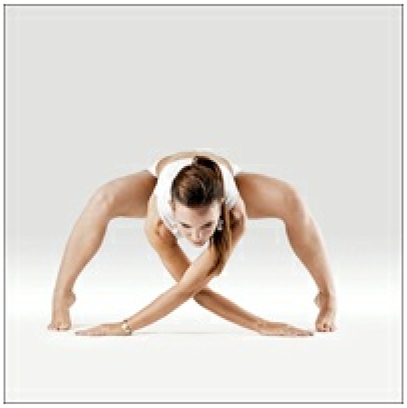

Tip Toe Pose Dedicated to Goddess Kali

Prapada Kalyasana

(PRUH-puh-duh kahl-YAHS-uh-nuh)

Modification: forward bend, hands to the floor, fingertips to the floor facing away from each other
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

ल व

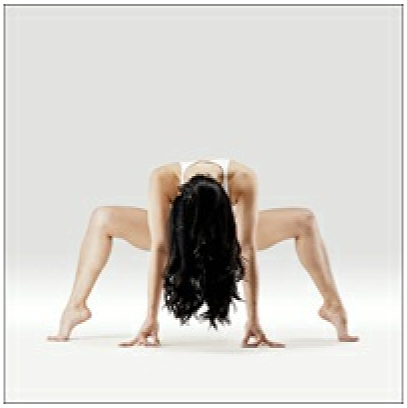

Tip Toe Pose Dedicated to Goddess Kali

Prapada Kalyasana

(PRUH-puh-duh kahl-YAHS-uh-nuh)

Modification: forward bend, arms open wide
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

ल व य

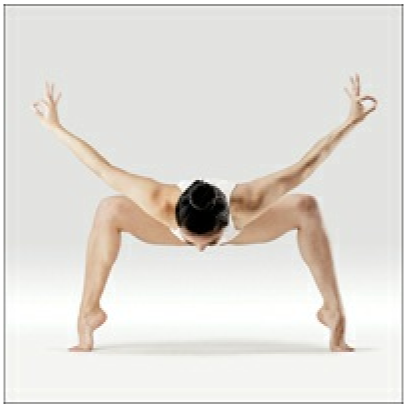

KALI SQUAT: HEELS UP—BINDING

Extended Side Revolved Hands Bound Half Pose Dedicated to Goddess Kali

Utthita Parshva Parivritta Baddha Hasta Ardha Kalyasana

(UT-ti-tuh PAHRSH-vuh puh-ri-VRIT-tuh BUH-duh HUH-stuh UHR-duh kahl-YAHS-uh-nuh)

Pose Type: standing, forward bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व य

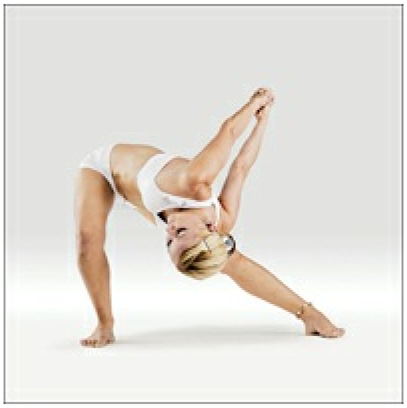

Hands Bound Pose Dedicated to Goddess Kali

Baddha Hasta Kalyasana

(BUH-duh HUH-stuh kahl-YAHS-uh-nuh)

Modification: hands bound under the leg, heels down

Pose Type: standing, forward bend, binding

Drishti Point: Nasagrai or Nasagre (nose)

ल व य

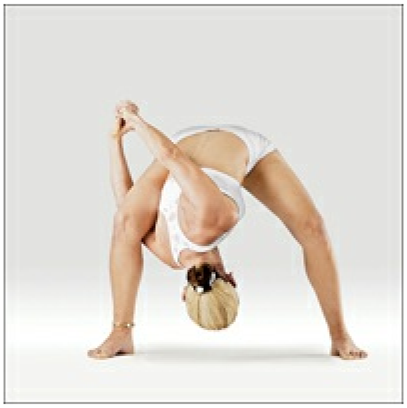

KALI SQUAT: HEELS UP—BINDING

Tip Toe Hands Bound Pose Dedicated to Goddess Kali

Prapada Baddha Hasta Kalyasana

(PRUH-puh-duh kahl-YAHS-uh-nuh)

Modification: hands bound under the leg
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

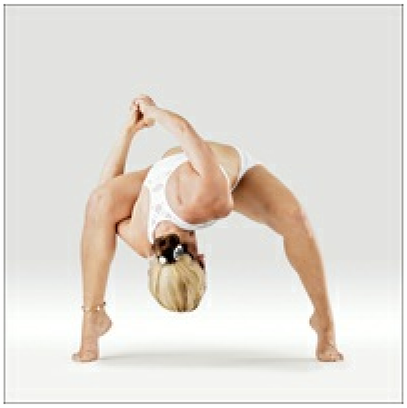

Tip Toe Half Bound Pose Dedicated to Goddess Kali

Prapada Ardha Baddha Kalyasana

(PRUH-puh-duh UHR-duh BUH-duh kahl-YAHS-uh-nuh)

Modification: one arm bound under the leg, palm to the rib cage, other forearm to the floor
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose), Angushtamadhye or Angushta Ma Dyai (thumbs)

LEGS WIDE: HALF FORWARD BEND

Half Feet Spread Out Intense Stretch Pose

Ardha Prasarita Padottanasana

(UHR-duh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Feet Spread Out Half Forward Bend
Modification: legs straight, hands on the shins
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

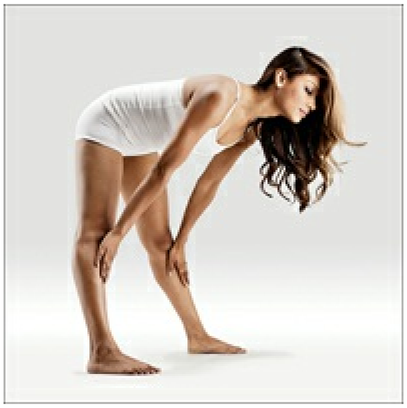

LEGS WIDE: HALF FORWARD BEND, HANDS TO THE FLOOR—HEELS DOWN & HEELS UP

Half Feet Spread Out Intense Stretch Pose

Ardha Prasarita Padottanasana

(UHR-duh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Feet Spread Out Half Forward Bend
Modification: 1. toes pointing straight ahead, palms flat on the floor 2. toes pointing slightly outward,

back of the hands to the floor, fingertips pointing toward each other

Pose Type: standing, forward bend

Drishti Point: Hastagrai or Hastagre (hands)

ल व ह

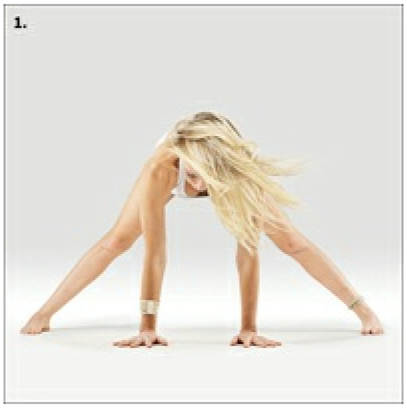

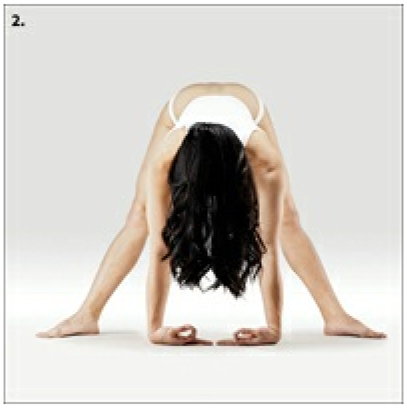

Tip Toe Half Feet Spread Out Intense Stretch Pose

Prapada Ardha Prasarita Padottanasana

(PRUH-puh-duh UHR-duh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Tip Toe Feet Spread Out Half Forward Bend

Modification: toes pointing straight ahead, heels up, palms flat on the floor

Pose Type: standing, forward bend

Drishti Point: Hastagrai or Hastagre (hands)

ल व ह

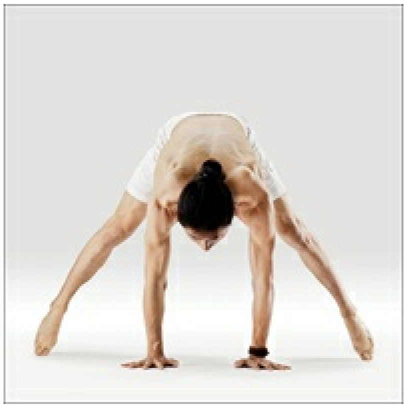

LEGS WIDE: FULL FORWARD BEND—HANDS TO THE FLOOR

Feet Spread Out Intense Stretch Pose 1

Prasarita Padottanasana 1

(pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Feet Spread Out Intense Stretch Pose A (Prasarita Padottanasana A), Feet Spread Out Full Forward Bend A
Modification: palms to the floor, elbows bent at 90 degrees
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
● ● ●

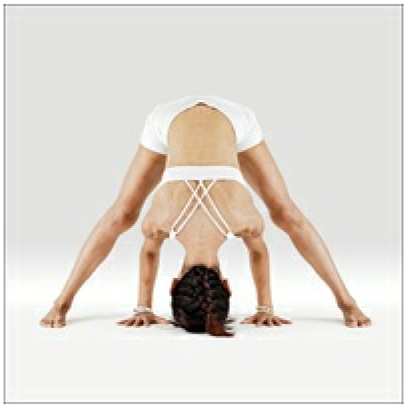

Feet Spread Out Intense Stretch Pose 1

Prasarita Padottanasana 1

(pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Feet Spread Out Intense Stretch Pose A (Prasarita Padottanasana A), Feet Spread Out Full Forward Bend A
Modification: arms straight, palms to the floor, fingertips pointing away from the head
Pose Type: standing, forward bend

Drishti Point: Hastagrai or Hastagre (hands)

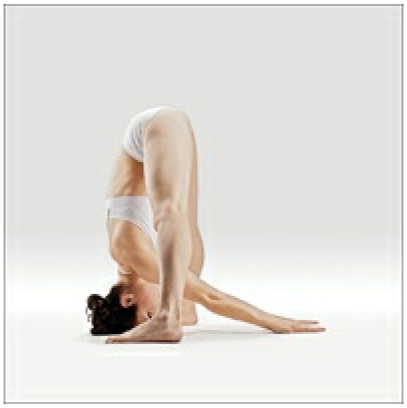

Feet Spread Out Intense Stretch Pose 4

Prasarita Padottanasana 4

(pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Feet Spread Out Intense Stretch Pose D (Prasarita Padottanasana D), Hands to Big Toes Feet Spread Out Full Forward Bend (Hasta Padangushta Prasarita Padottanasana)
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

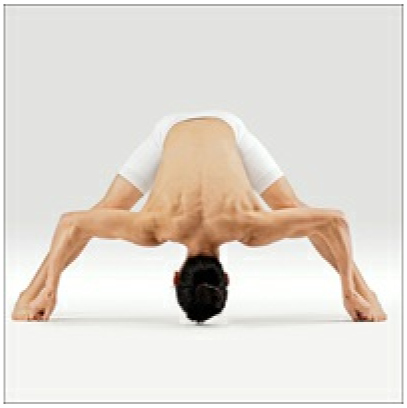

LEGS WIDE: FULL FORWARD BEND—ARMS BEHIND THE BACK

Feet Spread Out Intense Stretch Pose 2

Prasarita Padottanasana 2

(pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Feet Spread Out Intense Stretch Pose B (Prasarita Padottanasana B), Feet Spread Out Full Forward Bend

Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व य

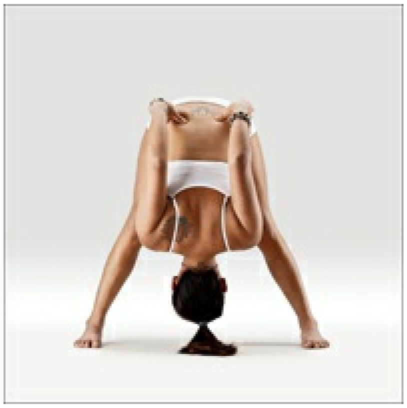

Feet Spread Out Intense Stretch Pose 5

Prasarita Padottanasana 5

(pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)
Also Known As: Reverse Prayer Feet Spread Out Full Forward Bend (Viparita Namaskar Prasarita Padottanasana)
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व

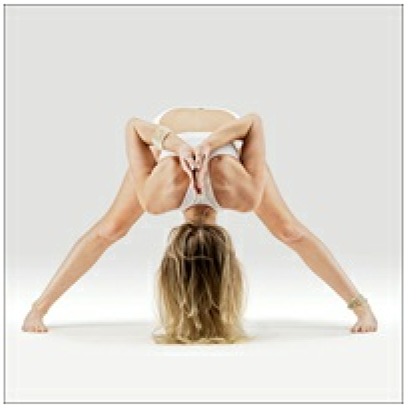

Feet Spread Out Intense Stretch Pose 3

Prasarita Padottanasana 3

(pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)
Also Known As: Feet Spread Out Intense Stretch Pose C (Prasarita Padottanasana C), Feet Spread Out Full Forward Bend, Hands Bound Wide Legs Forward Fold (Baddha Hasta Prasarita Padottanasana)

Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व य

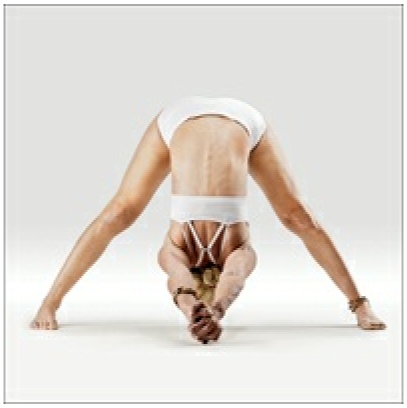

LEGS WIDE: FULL FORWARD BEND—ARMS OPEN WIDE

Feet Spread Out Intense Stretch Pose 6

Prasarita Padottanasana 6

(pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)
Also Known As: Feet Spread Out Full Forward Bend 6
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व य

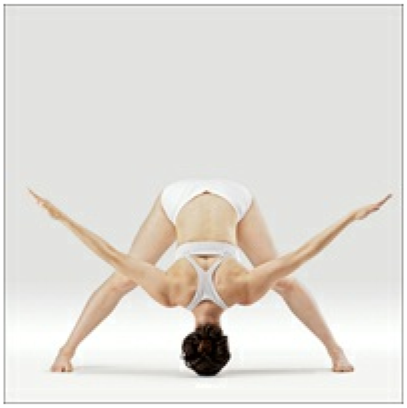

Feet Spread Out Intense Stretch Pose 7

Prasarita Padottanasana 7

(pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)
Also Known As: Feet Spread Out Full Forward Bend 7
Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)
ल व

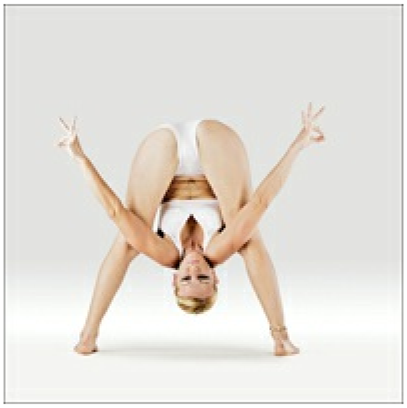

LEGS WIDE: FULL FORWARD BEND—TWISTS

Revolved Half Feet Spread Out Intense Stretch Pose

Parivritta Ardha Prasarita Padottanasana

(puh-ri-VRIT-tuh UHR-duh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Revolved Feet Spread Out Half Forward Bend
Modification: palm to the floor
Pose Type: standing, forward bend, twist
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व य

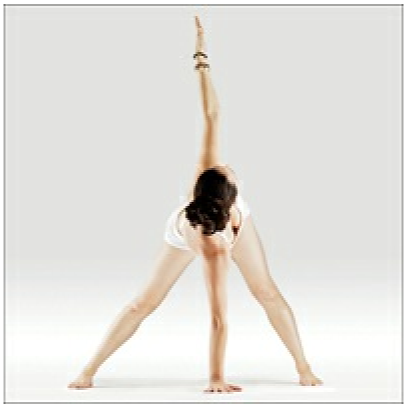

Revolved Half Feet Spread Out Intense Stretch Pose

Parivritta Ardha Prasarita Padottanasana

(puh-ri-VRIT-tuh UHR-duh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Revolved Feet Spread Out Half Forward Bend
Modification: forearm to the floor

Pose Type: standing, forward bend, twist
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

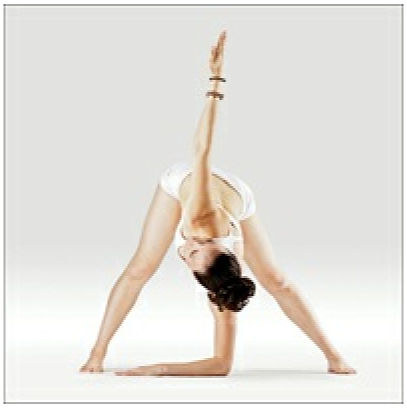

Revolved Paying Homage Feet Spread Out Intense Stretch Pose

Parivritta Namasya Prasarita Padottanasana

(puh-ri-VRIT-tuh nuh-MUHS-ya pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)
Also Known As: Revolved Paying Homage Feet Spread Out Forward Bend
Modification: legs crossed, hand to the back foot, other arm extended to the sky
Pose Type: standing, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

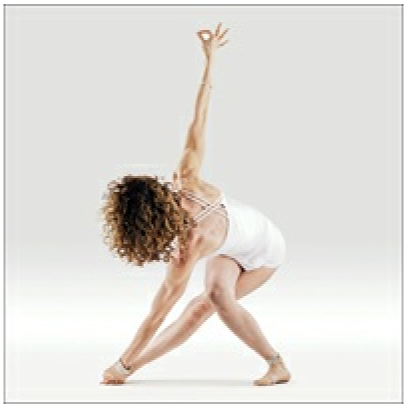

Bound Revolved Feet Spread Out Intense Stretch Pose

Baddha Parivritta Prasarita Padottanasana

(BUH-duh puh-ri-VRIT-tuh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)
Also Known As: Bound Revolved Feet Spread Out Full Forward Bend
Modification: arm straight, looking up
Pose Type: standing, forward bend, twist, binding

Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व य

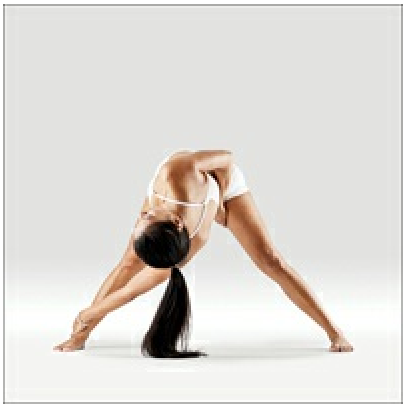

Hands Bound Revolved Feet Spread Out Intense Stretch Pose

Baddha Hasta Parivritta Prasarita Padottanasana

(BUH-duh HUH-stuh puh-ri-VRIT-tuh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Hands Bound Revolved Feet Spread Out Full Forward Bend
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व य

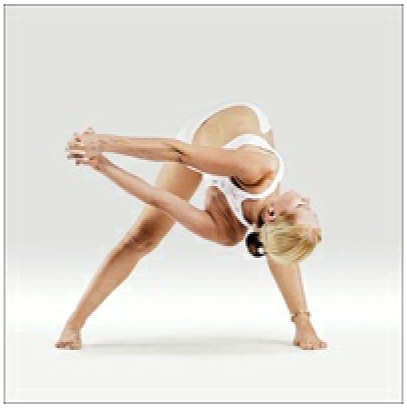

Revolved Paying Homage Feet Spread Out Intense Stretch Pose

Parivritta Namasya Prasarita Padottanasana

(puh-ri-VRIT-tuh nuh-MUHS-ya pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Modification: outside edges of the feet to the floor, legs crossed, forward bend, both hands on the back foot
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व

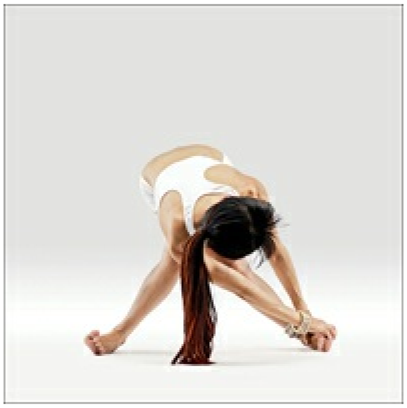

Revolved Feet Spread Out Intense Stretch Pose

Parivritta Prasarita Padottanasana

(puh-ri-VRIT-tuh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Revolved Feet Spread Out Full Forward Bend
Modification: grabbing onto the knee and the foot
Pose Type: standing, forward bend, side bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

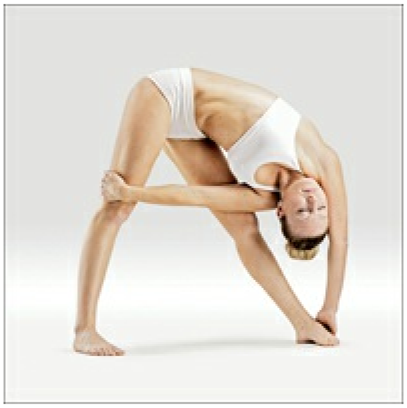

Revolved Feet Spread Out Intense Stretch Pose

Parivritta Prasarita Padottanasana

(puh-ri-VRIT-tuh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Revolved Feet Spread Out Full Forward Bend
Modification: grabbing onto the ankles
Pose Type: standing, forward bend, side bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

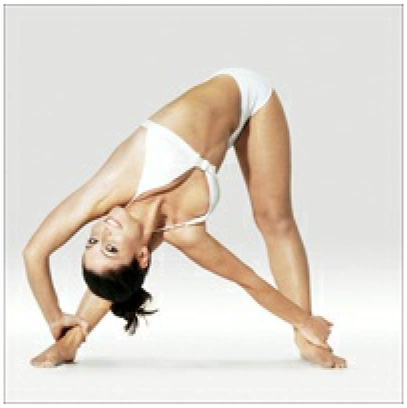

Revolved to the Side Feet Spread Out Intense Stretch Pose

Parivritta Parshva Prasarita Padottanasana

(puh-ri-VRIT-tuh PAHRSH-vuh pruh-SAH-ri-tuh puhd-o-tahn-AHS-uh-nuh)

Also Known As: Revolved to the Side Feet Spread Out Full Forward Bend)
Modification: grabbing onto the foot with both hands
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व

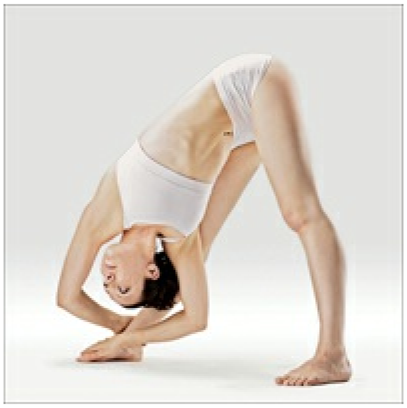

Downward Facing Dog Pose

Adho Mukha Shvanasana

(uh-DO MUK-uh shwa-NAHS-uh-nuh)

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

How to Perform the Pose:

1. Start by lying on your stomach with your whole body flat on the floor. Place your palms on the floor at the bottom of your ribs with fingertips facing forward and elbows tucked in. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale as you engage your core and lift your torso off the floor, and straighten your arms, coming into a distinct upside-down V shape with your sitting bones up to the sky. Your feet should be in line with your sitting bones and your hands should be shoulder width apart.
3. Try to keep your legs as straight as possible while pressing the heels to the floor. You should feel a deep stretch in the back of your thighs (the hamstrings) and your calf muscles.
4. Inhale as you lengthen your spine, keeping your neck long and your shoulder blades down your back. Try not to round your lower back, keeping it long. Exhale as you press through your chest, moving it toward your thighs. You should feel a deep stretch in the back of your shoulders.
5. You can either keep your head off the floor (Pose #1) or bring your forehead to the floor (Pose #2) if your shoulders are open.
6. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
7. To come out of the pose, inhale as you rock into Plank Pose, and exhale as you bend your elbows into Four Limbed Staff Pose (*Chaturanga Dandasana*) and lower all the way down to the floor into the starting position.

Modification:

1. head off the floor
2. forehead to the floor

adho = downward

mukha = face

shvana = dog

1.

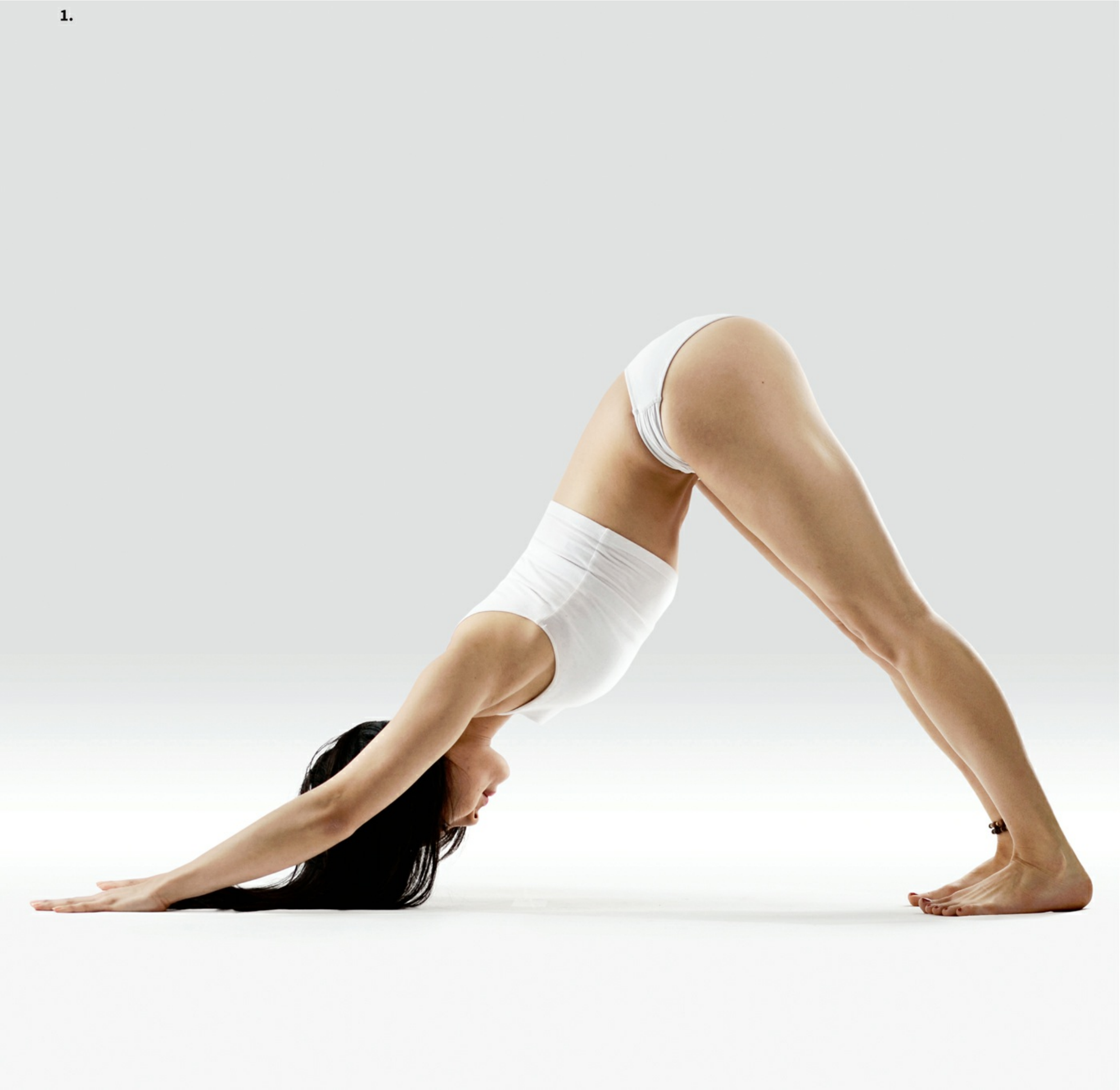

2.

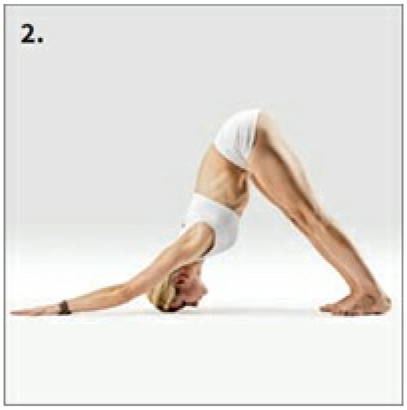

DOWNWARD DOG: FOREARMS AND HEAD ON THE FLOOR

Downward Facing Pose Dedicated to Makara on the Head

Adho Mukha Shirsha Makarasana

(uh-DO MUK-uh SHEER-shuh muh-kuh-RAHS-uh-nuh)

Pose Type: forward bend, inversion, core

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

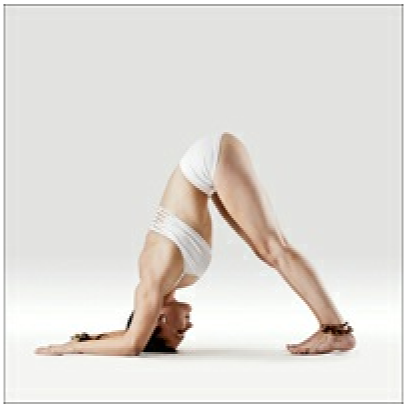

DOWNWARD DOG: KNEES BENT

Downward Facing Dog Pose

Adho Mukha Shvanasana

(uh-DO MUK-uh shwa-NAHS-uh-nuh)

Modification: knees bent, heels up

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

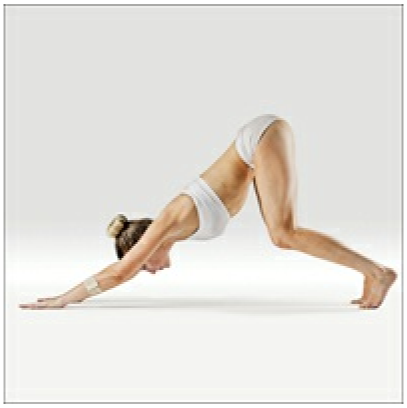

Downward Facing Dog Pose

Adho Mukha Shvanasana

(uh-DO MUK-uh shwa-NAHS-uh-nuh)

Modification: knees bent, heels up, spine in Cat Tilt, forehead toward the knees
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल ह

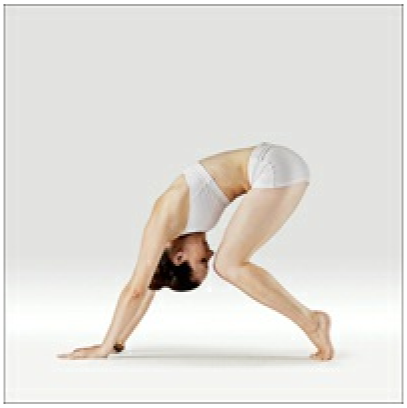

DOWNWARD DOG: WIDE-LEGGED

Feet Spread Downward Facing Dog Pose

Prasarita Pada Adho Mukha Shvanasana

(pruh-SAH-ri-tuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Modification: heels down, palms down
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

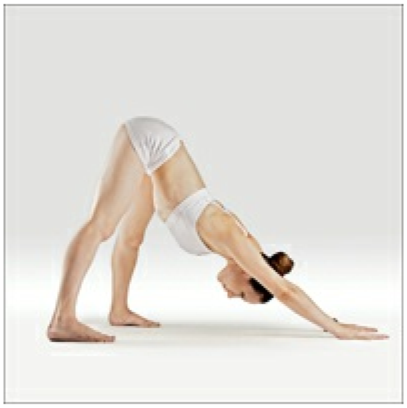

Tip Toe Feet Spread Downward Facing Dog Pose

Prapada Prasarita Pada Adho Mukha Shvanasana

(PRUH-puh-duh pruh-SAH-ri-tuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Modification: one palm on top of the other, fingertips to the floor, head above the arms
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands), Nasagrai or Nasagre (nose)

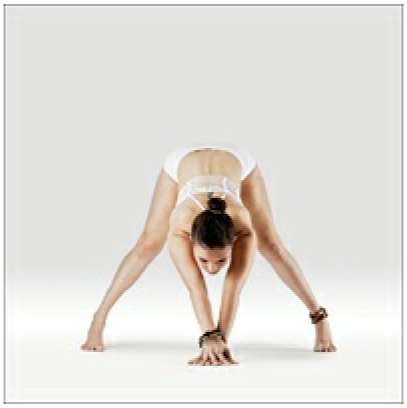

DOWNWARD DOG: ONE-LEGGED—KNEES TOGETHER

One-Legged Downward Facing Dog Pose

Eka Pada Adho Mukha Shvanasana

(EY-kuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Modification: back leg bent, knees together
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

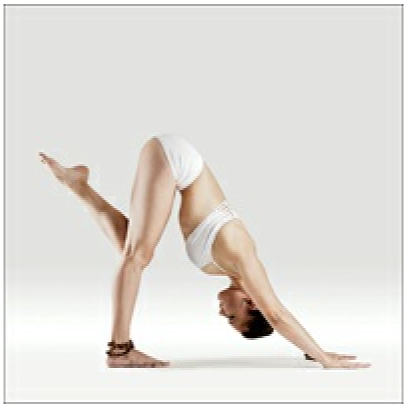

Leg Position of the Half Cow Face Pose in Downward Facing Dog Pose

Pada Ardha Gomukhasana in Adho Mukha Shvanasana

(PUH-duh URH-duh go-muk-AHS-uh-nuh in uh-DO MUK-uh shwa-NAHS-uh-nuh)

Also Known As: Leg Position of the One-Legged Cow Face Pose in Downward Facing Dog Pose (Eka Pada Gomukhasana in Adho Mukha Shvanasana)
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

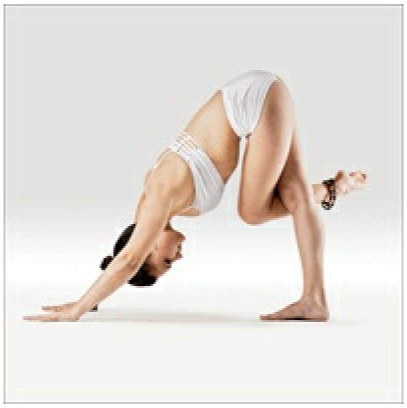

DOWNWARD DOG: ONE-ARMED

Revolved Downward Facing Dog Pose

Parivritta Adho Mukha Shvanasana

(puh-ri-VRIT-tuh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

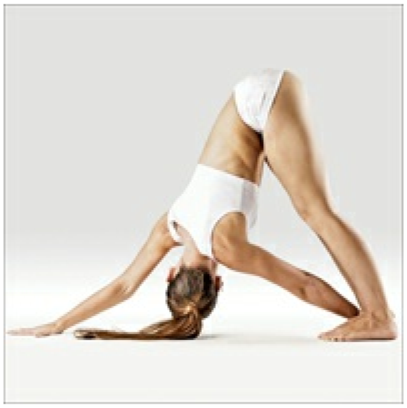

Hand to Knee One-Legged Downward Facing Dog Pose

Hasta Janu Eka Pada Adho Mukha Shvanasana

(HUH-stuh JAH-nu EY-kuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Modification: hand grabbing onto the opposite knee

Pose Type: standing one-legged balance, forward bend

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

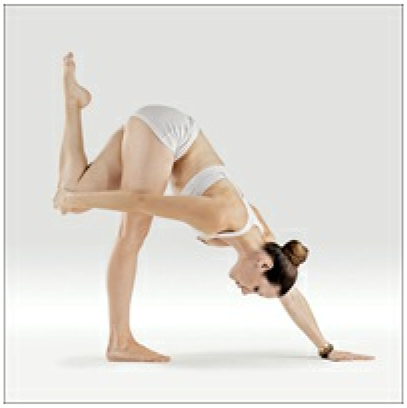

Hand to Ankle One-Legged Downward Facing Dog Pose 1

Hasta Kulpa Eka Pada Adho Mukha Shvanasana 1

(HUH-stuh KUL-puh EY-kuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Modification: grabbing onto the ankle on the same side; back leg bent, knees together

Pose Type: standing one-legged balance, forward bend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

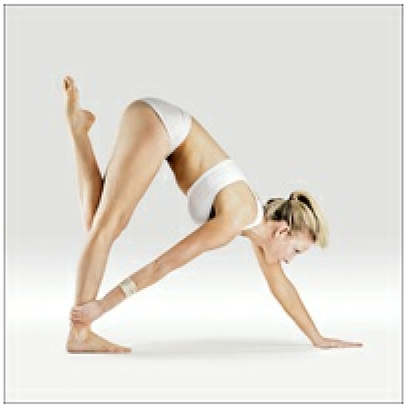

Hand to Ankle One-Legged Downward Facing Dog Pose 2

Hasta Kulpa Eka Pada Adho Mukha Shvanasana 2

(HUH-stuh KUL-puh EY-kuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Modification: grabbing onto the ankle on the opposite side; back leg bent, knees together
Pose Type: standing one-legged balance, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल

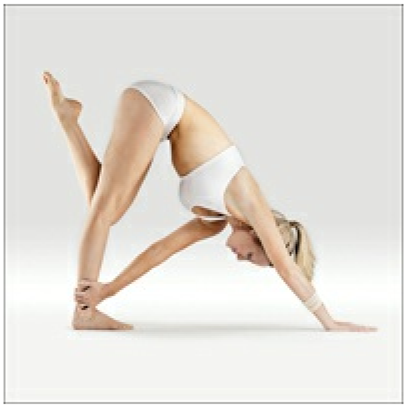

DOWNWARD DOG: ONE-LEGGED

One-Legged Downward Facing Dog Pose

Eka Pada Adho Mukha Shvanasana

(EY-kuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Modification: knee bent
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

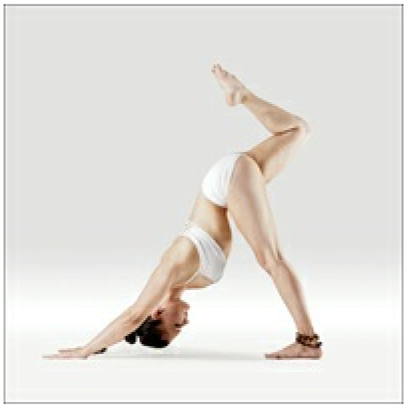

One Leg to the Side Downward Facing Dog Pose

Parshva Eka Pada Adho Mukha Shvanasana

(PAHRSH-vuh EY-kuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

ल व र य

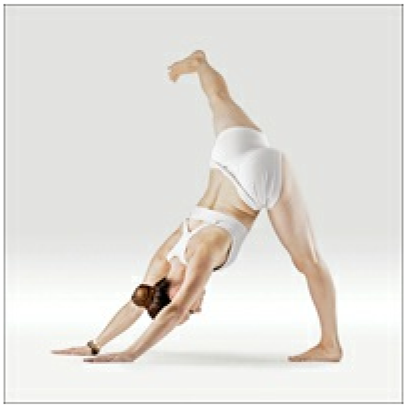

One-Legged Downward Facing Dog Pose

Eka Pada Adho Mukha Shvanasana

(EY-kuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Also Known As: One-Legged Downward Facing Dog Shoulder Press

Modification: bottom knee bent; elbows bent, fingertips to the floor

Pose Type: standing, forward bend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

ल व

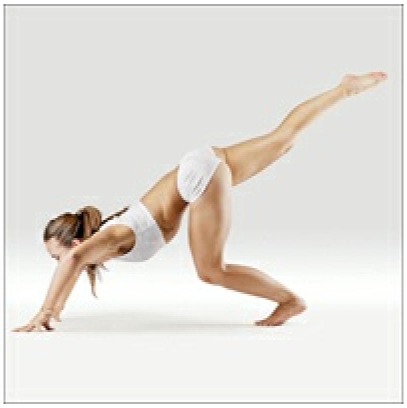

One-Legged Downward Facing Dog Pose

Eka Pada Adho Mukha Shvanasana

(EY-kuh PUH-duh uh-DO MUK-uh shwa-NAHS-uh-nuh)

Also Known As: One Leg Raised Up Downward Facing Dog Pose (Utthita Eka Pada Adho Mukha Shvanasana)

Modification: back leg straight and extended up to the sky
Pose Type: standing, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
लं वं रं यं हं

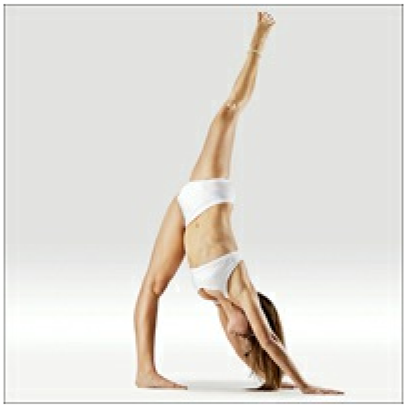

DOWNWARD DOG: TWISTS

Uneven Legs Revolved Downward Facing Dog Pose 1

Vishama Pada Parivritta Adho Mukha Shvanasana 1

(VISH-uh-muh PUH-duh puh-ri-VRIT-tuh uh-DO MUK-uh shwa-NAHS-uh-nuh)
Modification: one knee bent, foot crossed over
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
लं वं हं

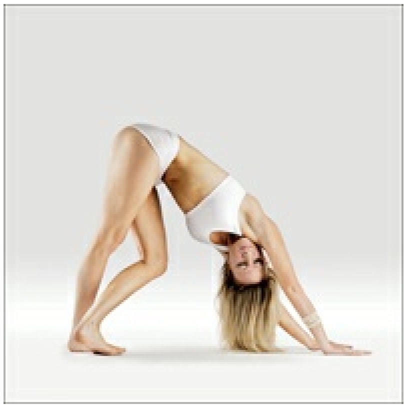

Uneven Legs Revolved Downward Facing Dog Pose 2

Vishama Pada Parivritta Adho Mukha Shvanasana 2

(VISH-uh-muh PUH-duh puh-ri-VRIT-tuh uh-DO MUK-uh shwa-NAHS-uh-nuh)
Modification: both legs straight, one leg crossed over

Pose Type: standing, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

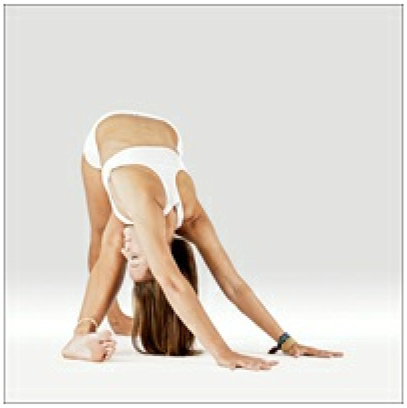

INTENSE SIDE STRETCH: STRAIGHT SPINE TO BACKBEND

Upward Bound Hands Pose in Hero Succession Series

Urdhva Baddha Hastasana in Vira Parampara

(OORD-vuh BUH-duh huh-STAHS-uh-nuh in VEER-uh puh-ruhm-puh-RAH)
Modification: arms up to the sky, fingers interlocked, spine straight
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

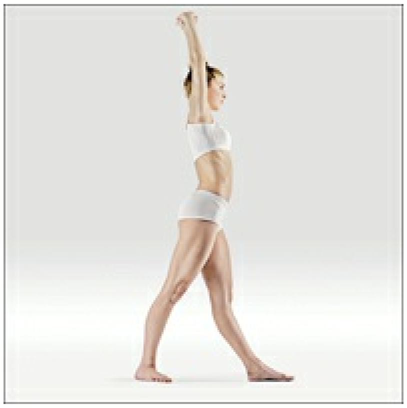

Hands Bound Pose in Hero Succession Series

Baddha Hastasana in Vira Parampara

(BUH-duh huh-STAHS-uh-nuh in VEER-uh puh-ruhm-puh-RAH)
Modification: grabbing onto the triceps behind the back
Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

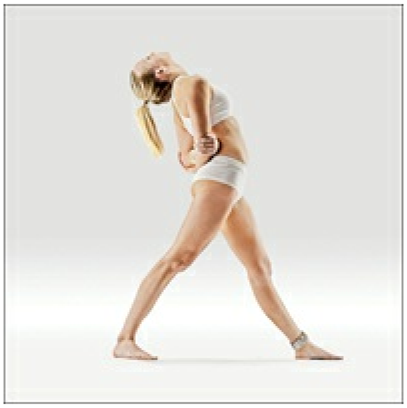

Reverse Prayer Pose in Hero Succession Series

Viparita Namaskarasana in Vira Parampara

(vi-puh-REET-tuh nuh-muhs-kahr-AHS-uh-nuh in VEER-uh puh-ruhm-puh-RAH)

Also Known As: Back of the Body Prayer in Hero Succession Series (Paschima Namaskara in Vira Parampara)
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

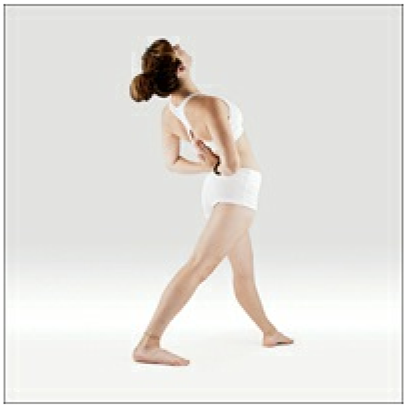

Upward Bound Hands Pose in Hero Succession Series

Urdhva Baddha Hastasana in Vira Parampara

(OORD-vuh BUH-duh huh-STAHS-uh-nuh in VEER-uh puh-ruhm-puh-RAH)

Modification: deep backbend
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angusthamadhye or Angustha Ma Dyai (thumbs)

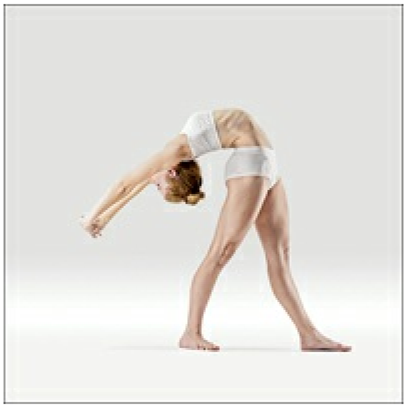

INTENSE SIDE STRETCH: HALF FORWARD BEND

Intense Side Stretch Pose Prep.

Parshvottanasana Prep.

(pahrsh-vo-tahn-AHS-uh-nuh)

Modification: arms straight out in front, palms pressed together
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

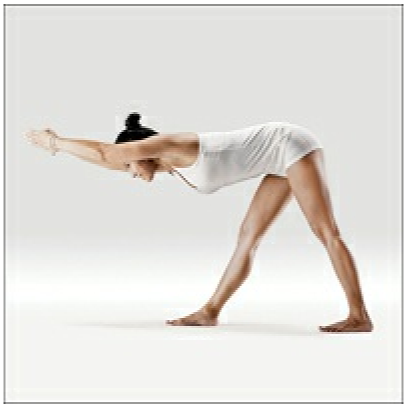

Intense Side Stretch Pose Prep.

Parshvottanasana Prep.

(pahrsh-vo-tahn-AHS-uh-nuh)

Modification: one hand to the lower back, one arm extended out in front and parallel to the floor
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)

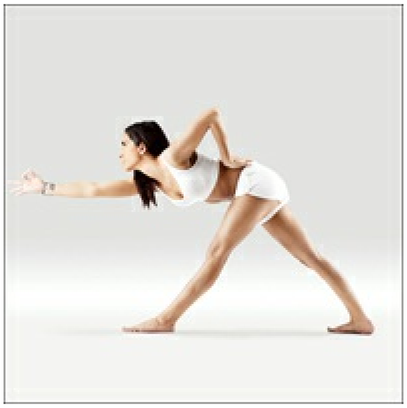

Intense Side Stretch Pose Prep.

Parshvottanasana Prep.

(pahrsh-vo-tahn-AHS-uh-nuh)

Modification: grabbing onto the triceps behind the back
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

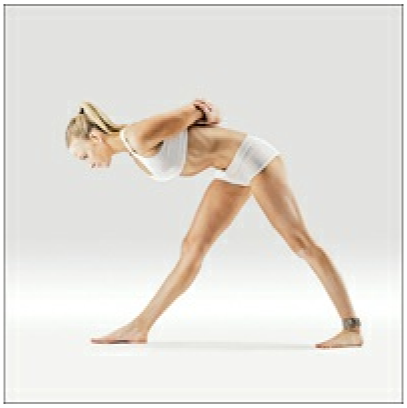

Intense Side Stretch Pose Prep.

Parshvottanasana Prep.

(pahrsh-vo-tahn-AHS-uh-nuh)

Modification: arms open to the sides
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

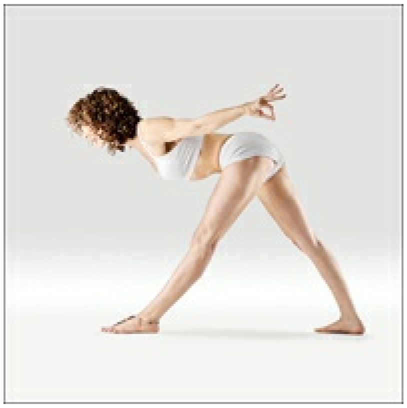

INTENSE SIDE STRETCH: HALF FORWARD BEND—GRABBING ONTO ONE LEG

Intense Side Stretch Pose Prep.

Parshvottanasana Prep.

(pahrsh-vo-tahn-AHS-uh-nuh)

Modification: both hands on the shin on the front leg
Pose Type: standing, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

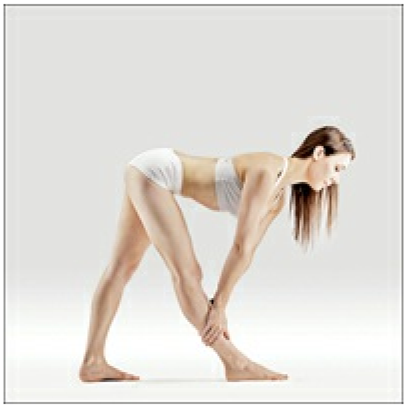

Extended Both Hands to Foot Intense Side Stretch Pose

Utthita Dwi Hasta Pada Parshvottanasana

(UT-ti-tuh DWI-huh-stuh PUH-duh pahrsh-vo-tahn-AHS-uh-nuh)

Modification: toes of the front foot flexed in and lifted off the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

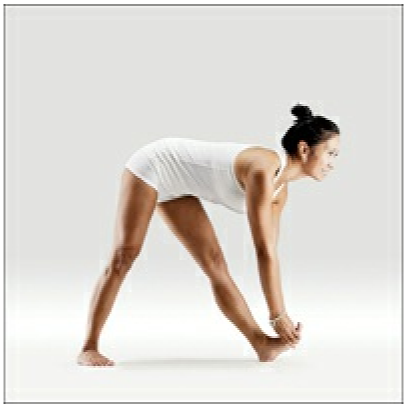

Intense Side Stretch Pose

Parshvottanasana

(pahrsh-vo-tahn-AHS-uh-nuh)

Modification: both hands grabbing onto the calf of the back leg
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
लं वं हं

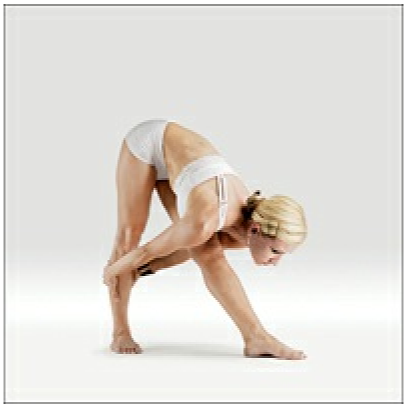

INTENSE SIDE STRETCH: FULL FORWARD BEND—ARMS IN FRONT AND ARMS POINTING TO THE BACK

Intense Side Stretch Pose Prep.

Parshvottanasana Prep.

(pahrsh-vo-tahn-AHS-uh-nuh)

Modification: palms to the floor on either side of the front foot
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं रं यं

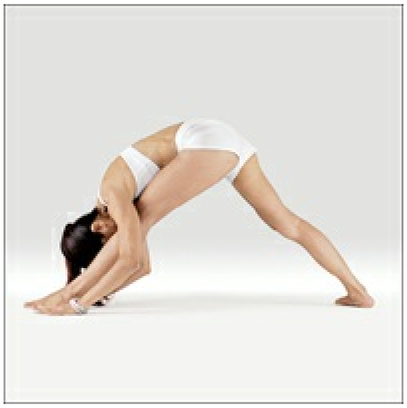

Intense Side Stretch Pose Prep.

Parshvottanasana Prep.

(pahrsh-vo-tahn-AHS-uh-nuh)

Also Known As: Hero Succession Series (Vira Parampara)
Modification: arms reaching to the back, fingertips to the floor
1. chin to the shin
2. forehead to the shin
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

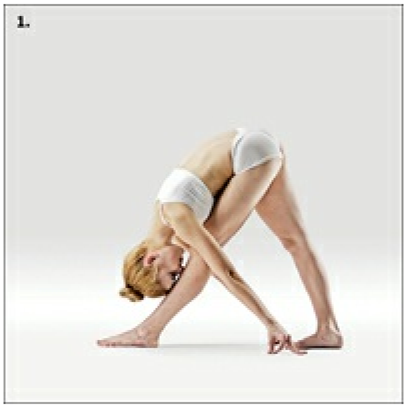

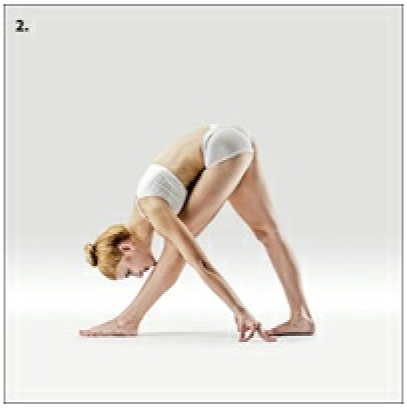

Intense Side Stretch Pose Prep.

Parshvottanasana Prep.

(pahrsh-vo-tahn-AHS-uh-nuh)

Modification: hands in Anjali Mudra (Hands in Prayer), arms wrapped around the front leg
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

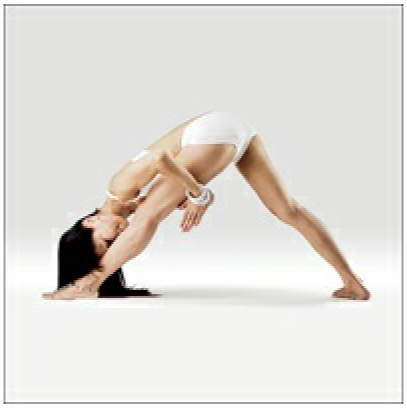

Revolved Intense Side Stretch Pose

Parivritta Parshvottanasana

(puh-ri-VRIT-tuh pahrsh-vo-tahn-AHS-uh-nuh)

Modification: hands in Anjali Mudra (Hands in Prayer), arms wrapped around the front leg, twist
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

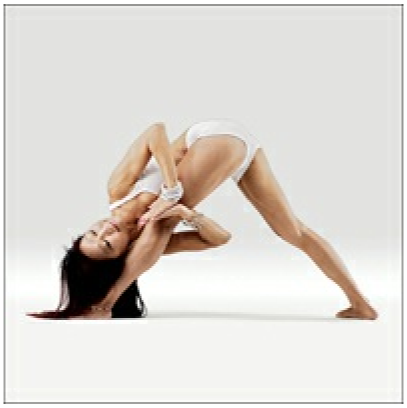

Intense Side Stretch Pose

Parshvottanasana

(pahrsh-vo-tahn-AHS-uh-nuh)

Also Known As: Reverse Prayer Intense Side Stretch Pose (Viparita Namaskar Parsvottanasana), Back of the Body Prayer Intense Side Stretch Pose (Paschima Namaskara Parshvottanasana)
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

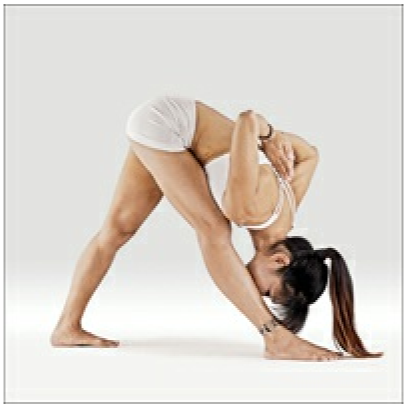

Hands Bound Intense Side Stretch Pose

Baddha Hasta Parshvottanasana

(BUH-duh HUH-stuh pahrsh-vo-tahn-AHS-uh-nuh)

Modification: both legs straight
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)
ल व र य ह

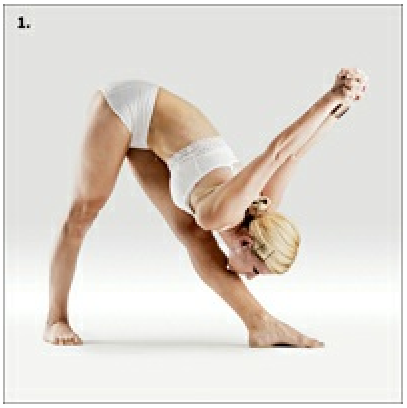

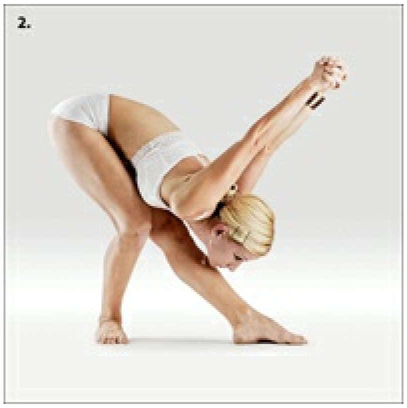

INTENSE SIDE STRETCH: FULL FORWARD BEND—ONE ARM IN FRONT—ONE ARM BEHIND

Uneven Legs Tip Toe Intense Side Stretch Pose

Vishama Pada Prapada Parshvottanasana

(VISH-uh-muh PUH-duh PRUH-puh-duh pahrsh-vo-tahn-AHS-uh-nuh)

Modification: 1. one arm stretched out to the back and up to the sky; other arm stretched out to the front, fingertips to the floor 2. both arms stretched out in front, fingertips to the floor
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व र य ह

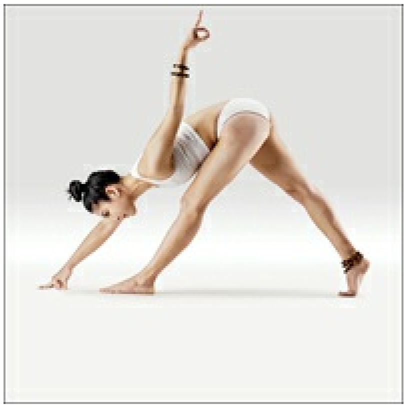

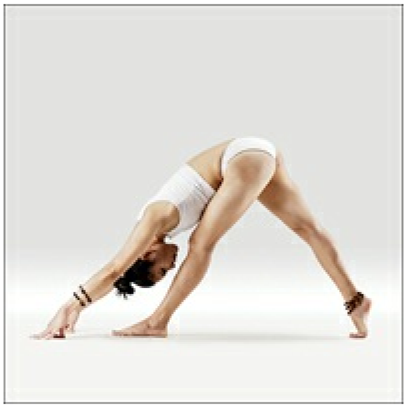

Intense Side Stretch Pose Prep.

Parshvottanasana Prep.

(pahrsh-vo-tahn-AHS-uh-nuh)

Modification: opposite arm wrapped around the front calf, chin to the shin
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व र य ह

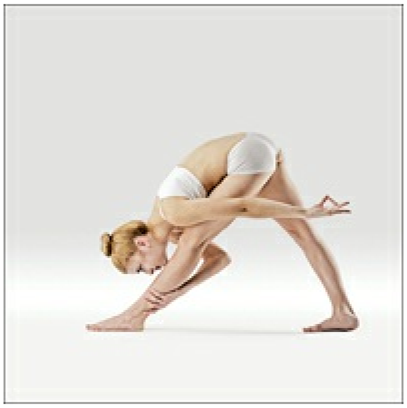

TRIANGLE POSE

Extended Side Triangle Pose

Utthita Parshva Trikonasana

(UT-ti-tuh PAHRSH-vuh tri-ko-NAHS-uh-nuh)

Modification: top arm parallel to the floor
Pose Type: standing, forward bend, side bend
Drishti Point: Hastagrai or Hastagre (hands)
ल व र य

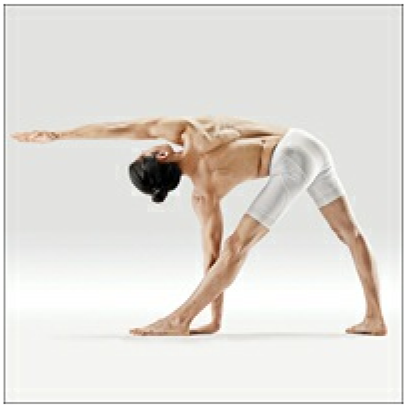

Extended Triangle Pose

Utthita Trikonasana

(UT-ti-tuh tri-ko-NAHS-uh-nuh)

Pose Type: standing, forward bend, side bend

Drishti Point: Angusthamadhye or Angustha Ma Dyai (thumbs)

How to Perform the Pose:

1. Begin by standing in Mountain Pose (*Tadasana*). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Inhale and step your feet about elbow distance apart with toes facing forward and feet parallel to each other. Expand your chest and hold your arms straight out to the sides, parallel to the floor.
3. Keep your legs strong and straight by engaging your quadriceps and pulling the kneecaps up. Turn your left foot 90 degrees to the left. Turn your right foot as close to 45 degrees to the left as possible.
4. On your next exhale, reach over to the left as far as you can with good form and drop your left hand toward the left shin (or left foot or to the floor either on the inside or the outside of the left foot).
5. Reach your right arm up to the sky and look at your right thumb (the *drishti* point). Make sure that your chest is not collapsing down toward the floor by lengthening both arms, grabbing onto the big toe of your left foot with pointer finger and middle finger and thumb locked on top, and leaning back far enough so that your shoulders are stacked one on top of the other.
6. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch. Exhale as you release the pose. Inhale as you press strongly into both feet to come up.
7. Exhale, come back to Mountain Pose (*Tadasana*) and repeat on the other side.

Modification:

1. front view
2. back view

utittha = extended

tri = three

kona = angle

1.

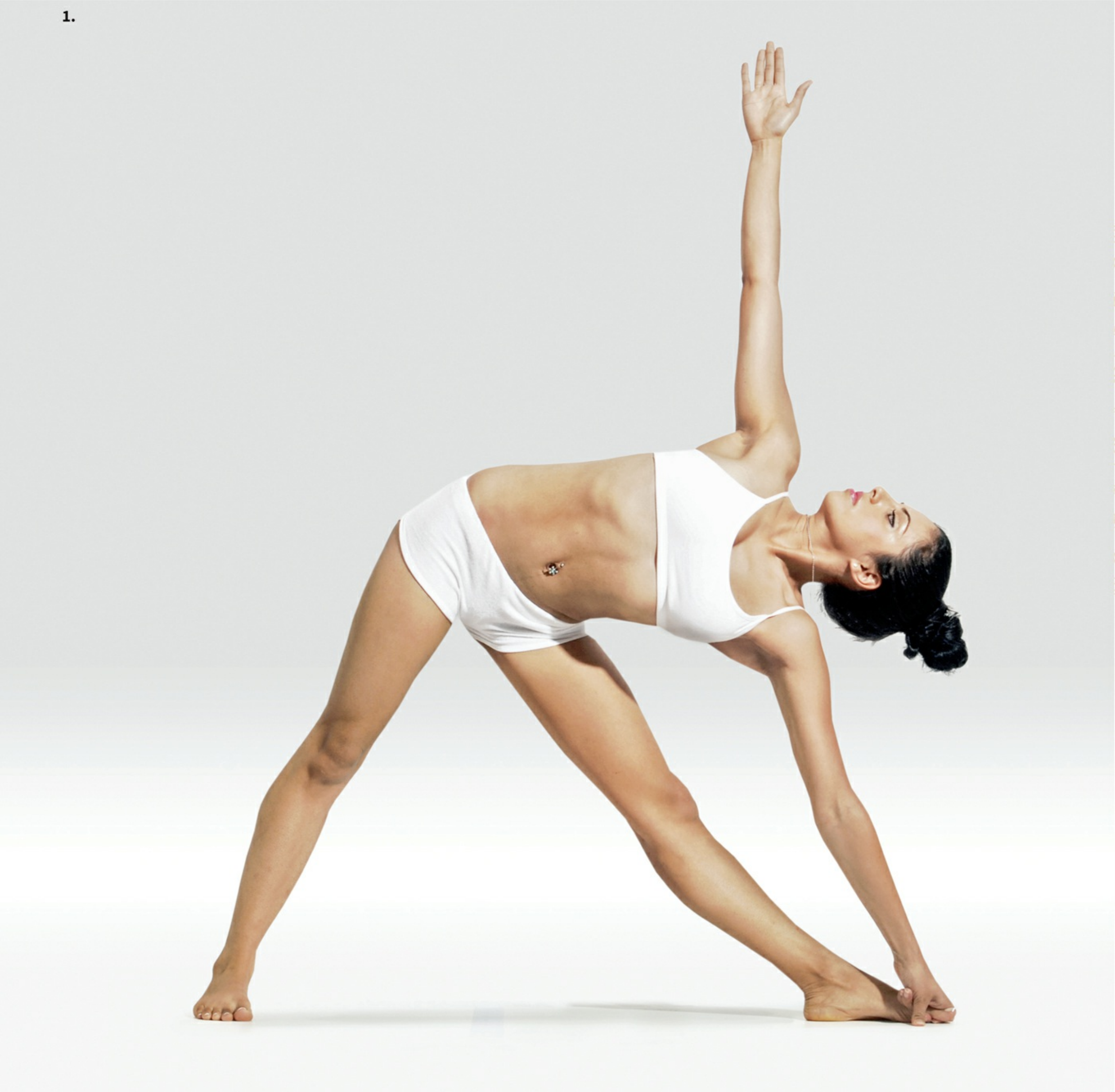

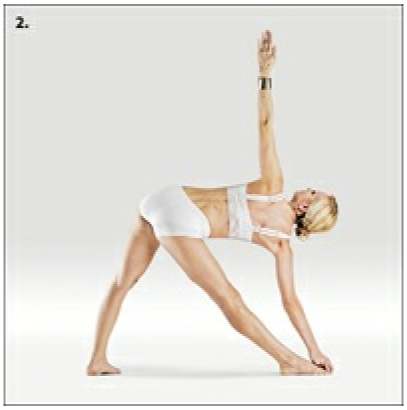

2.

Half Bound Extended Triangle Pose

Ardha Baddha Utthita Trikonasana

(UHR-duh BUH-duh UT-ti-tuh tri-ko-NAHS-uh-nuh)

Pose Type: standing, forward bend, side bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

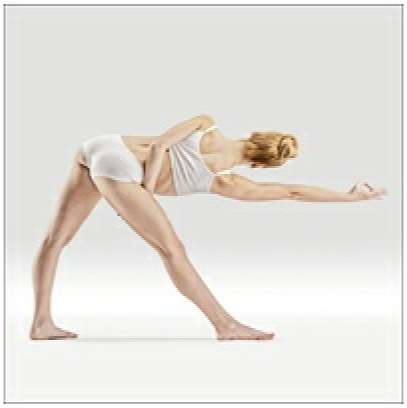

Bound Extended Triangle Pose

Baddha Utthita Trikonasana

(BUH-duh UT-ti-tuh tri-ko-NAHS-uh-nuh)

Modification: fingers interlocked on the inside of the thigh
Pose Type: standing, forward bend, side bend, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

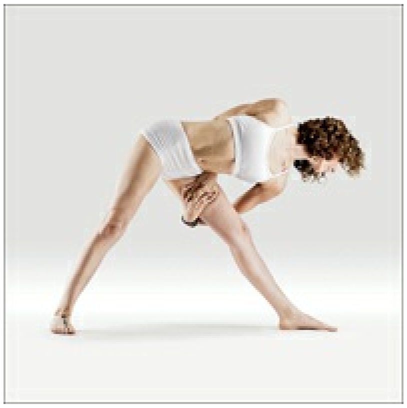

Bound Extended Triangle Pose

Baddha Utthita Trikonasana

(BUH-duh UT-ti-tuh tri-ko-NAHS-uh-nuh)

Pose Type: standing, forward bend, side bend, binding, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व र य

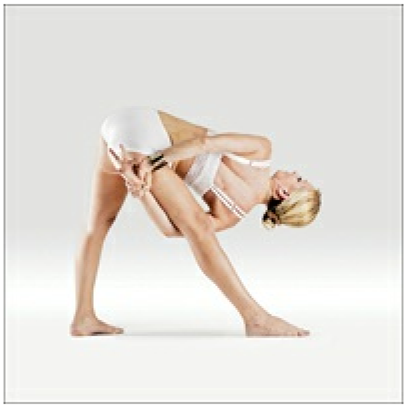

Revolved Sideways Extended Triangle Pose

Parivritta Parshva Utthita Trikonasana

(puh-ri-VRIT-tuh PAHRSH-vuh UT-ti-tuh tri-ko-NAHS-uh-nuh)

Modification: grabbing onto both shins

Pose Type: standing, forward bend, side bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व र य

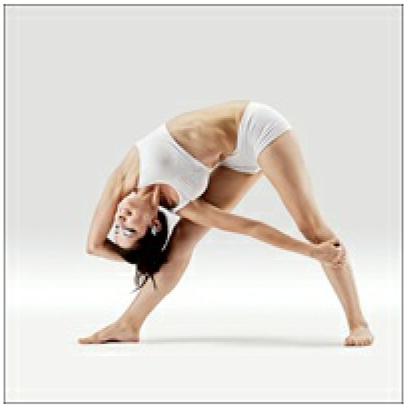

TRIANGLE POSE: REVOLVED & BINDING

Revolved Triangle Pose

Parivritta Trikonasana

(puh-ri-VRIT-tuh tri-ko-NAHS-uh-nuh)

Also Known As: Revolved Intense Side Stretch Pose (Parivritta Parshvottanasana)
Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

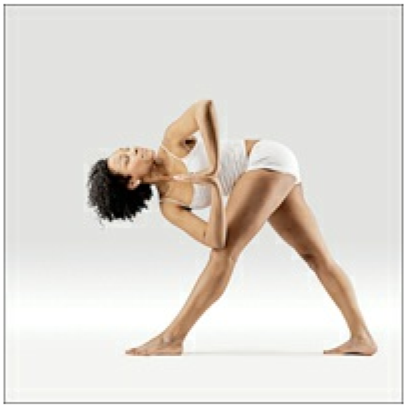

Revolved Triangle Pose

Parivritta Trikonasana

(puh-ri-VRIT-tuh tri-ko-NAHS-uh-nuh)

Pose Type: standing, forward bend, twist
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

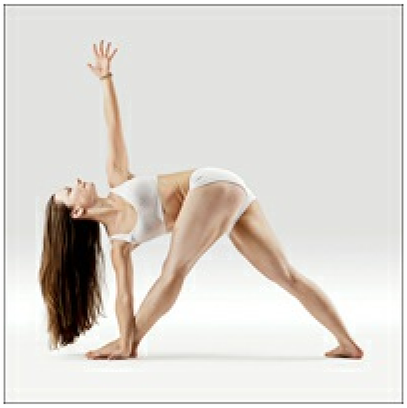

Revolved Bound Triangle Pose

Parivritta Baddha Trikonasana

(puh-ri-VRIT-tuh BUH-duh tri-ko-NAHS-uh-nuh)

Also Known As: Bound Leg Twisted Angle Pose (Baddha Pada Parivritta Konasana)
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

Extended Side Angle Pose

Utthita Parshva Konasana

(UT-ti-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Pose Type: standing, side bend

Drishti Point: Hastagrai or Hastagrahe (hands)

How to Perform the Pose:

1. Begin by standing in Mountain Pose (*Tadasana*). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Inhale and step your feet about wrist-distance apart with toes facing forward and feet parallel to each other. Expand your chest and hold your arms straight out to the sides parallel to the floor.
3. Keep your legs strong and straight by engaging your quadriceps and pulling the kneecaps up. Turn your left foot 90 degrees to the left. Turn your right foot 45 degrees to the left as much as possible.
4. Exhale and bend your left knee until your left thigh is parallel to the floor, coming into Warrior 2 Pose (Virabhadrasana 2).
5. On the exhale, bend to the left and drop your left elbow to the left thigh. Reach your right arm over your head and rotate your chest to the side until your right shoulder is on top of your left shoulder.
6. You can experiment with bringing your left palm to the floor on the inside of your left thigh or the outside of your thigh.
7. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch. Exhale as you release the pose. Inhale as you press strongly into both feet to come up.
8. Exhale, come back to Mountain Pose (*Tadasana*) and repeat on the right side.

Modification:

1. arm to the inside of the thigh
2. elbow on the thigh
3. arm to the outside of the thigh

utthita = extended, rising, risen

parsva = side

kona = angle

1.

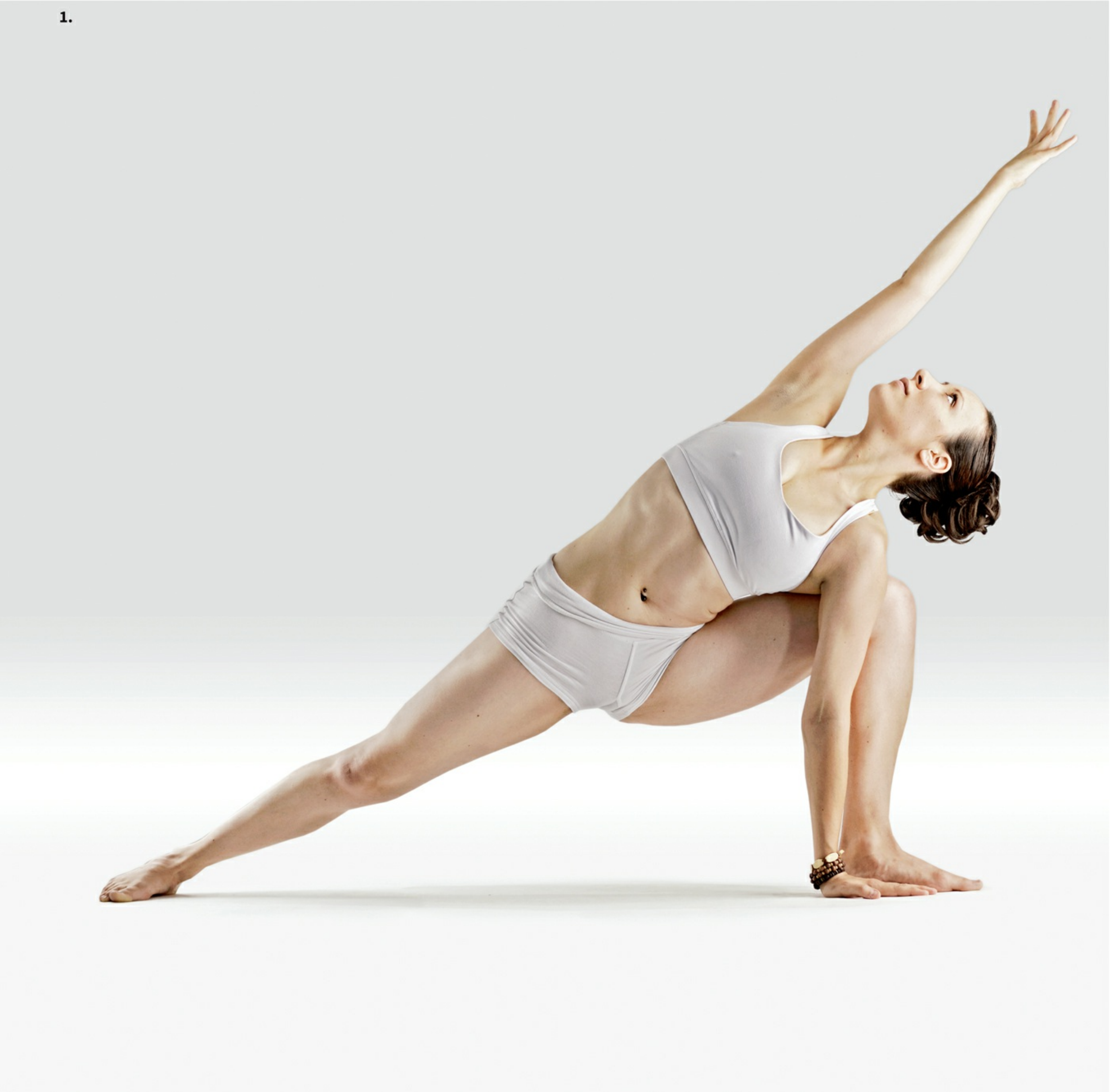

2.

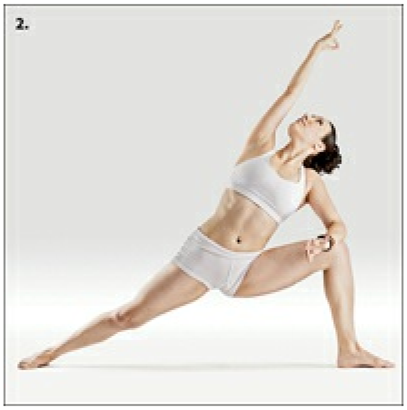

3.

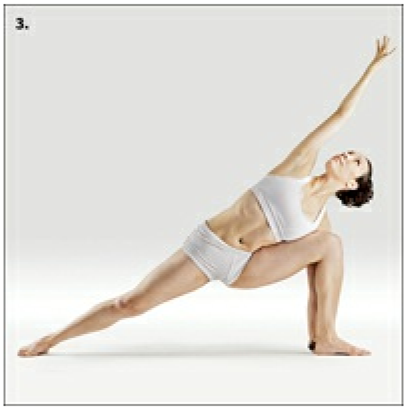

Warrior 2

Virabhadrasana 2

(veer-uh-buh-DRAHS-uh-nuh)

Also Known As: Half Side Angle Pose (Ardha Parshva Konasana)
Pose Type: standing
Drishti Point: Hastagrai or Hastagre (hands)

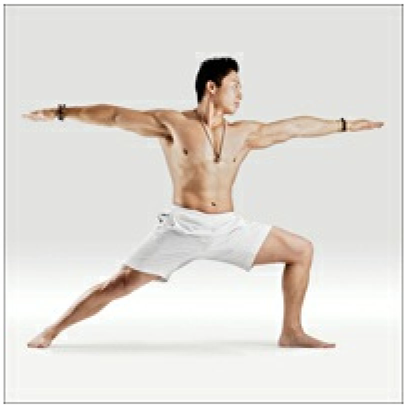

SIDE ANGLE POSE

Extended Side Angle Pose Prep.

Utthita Parshva Konasana Prep.

(UT-ti-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: back knee on the floor, palm to the floor on the inside of the front leg
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

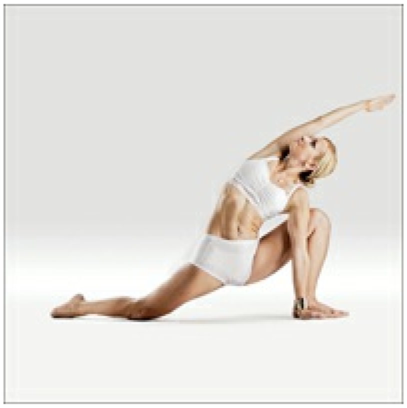

Both Arms Extended Side Angle Pose Prep.

Dwi Hasta Utthita Parshva Konasana Prep.

(DWI-huh-stuh UT-ti-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: back knee on the floor
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

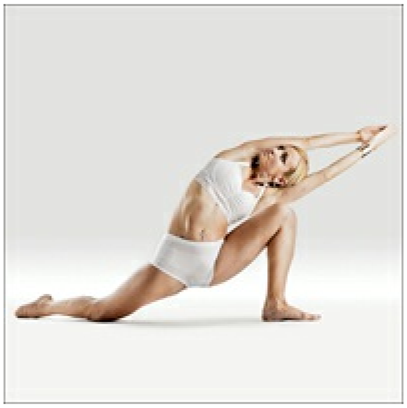

Both Arms Extended Side Angle Pose

Dwi Hasta Utthita Parshva Konasana

(DWI-huh-stuh UT-ti-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: back knee off the floor
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

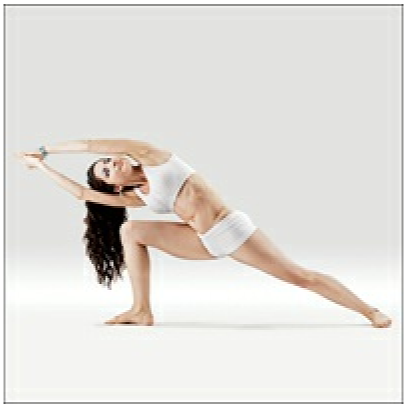

Tip Toe Extended Side Angle Pose Prep.

Prapada Utthita Parshva Konasana Prep.

(PRUH-puh-duh UT-ti-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: front heel lifted, fingertips to the floor on the inside of the leg, other arm resting on the side of the torso

Pose Type: standing, side bend

Drishti Point: Urdhva or Antara Drishti (up to the sky)

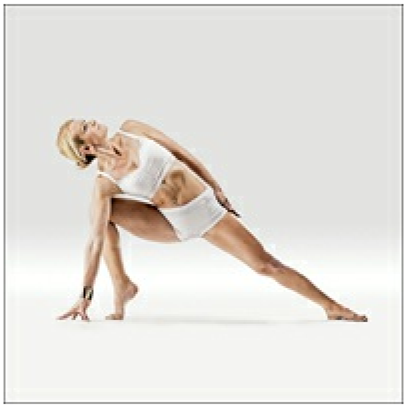

Tip Toe Both Arms Extended Side Angle Pose

Prapada Dwi Hasta Utthita Parshva Konasana

(PRUH-puh-duh DWI-huh-stuh UT-ti-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: back knee off the floor

Pose Type: standing, side bend

Drishti Point: Urdhva or Antara Drishti (up to the sky)

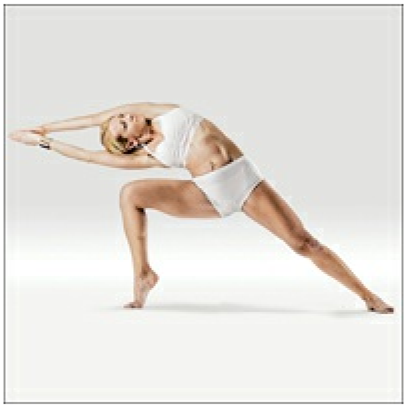

SIDE ANGLE POSE: BINDING

Half Bound Extended Side Angle Pose

Ardha Baddha Utthita Parshva Konasana

(UHR-duh BUH-duh UT-ti-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: hand to the floor on the inside of the front leg

Pose Type: standing, side bend, binding

Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व

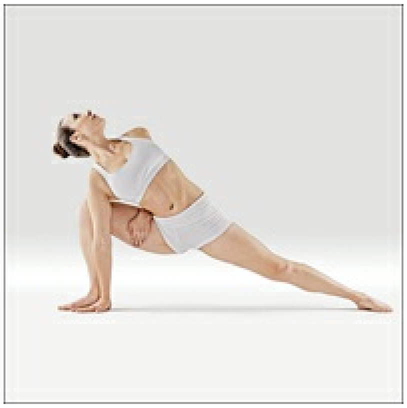

Bound Extended Side Angle Pose

Baddha Utthita Parshva Konasana

(BUH-duh UT-ti-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: fingers interlocked on the inside of the front thigh

Pose Type: standing, side bend, binding

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

ल व

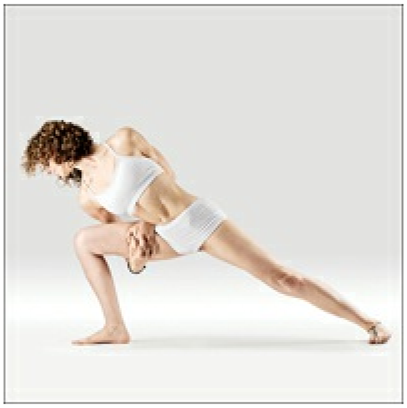

Bound Extended Side Angle Pose

Baddha Utthita Parshva Konasana

(BUH-duh UT-ti-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Also Known As: Bound Arms Side Angle Pose (Baddha Hasta Parshva Konasana)

Pose Type: standing, side bend, binding, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

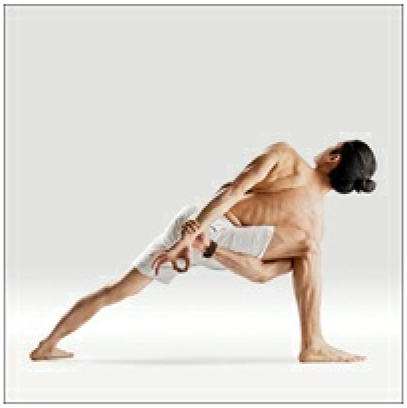

Pose Dedicated to Vishvamitra Prep.

Vishvamitrasana Prep.

(VISH-vah-mi-TRAHS-uh-nuh)
Modification: knee bent wrapped around the tricep, grabbing onto the foot
Pose Type: standing, arm balance
Drishti Point: Urdhva or Antara Drishti (up to the sky)

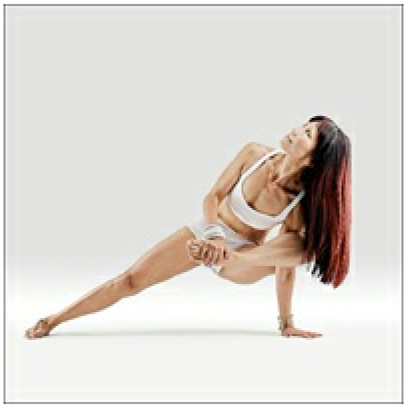

Pose Dedicated to Vishvamitra Prep.

Vishvamitrasana Prep.

(vish-vah-mi-TRAHS-uh-nuh)
Modification: knee bent, shin to the tricep
Pose Type: standing, arm balance
Drishti Point: Hastagrai or Hastagre (hands), Padayoragrai or Padayoragre (toes/feet)

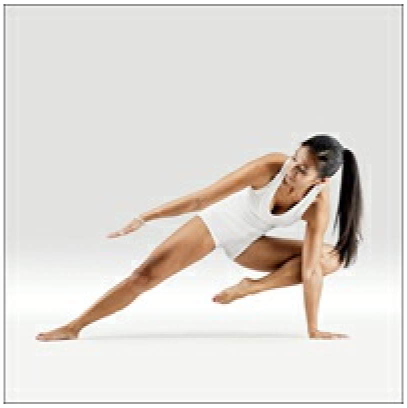

SIDE ANGLE POSE: REVOLVED

Revolved Side Angle Pose Prep.

Parivritta Parshva Konasana Prep.

(puh-ri-VRIT-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: one arm under the leg, hands in Anjali Mudra (Hands in Prayer),
1. back knee on the floor, looking up to the sky 2. back knee off the floor, looking down
Pose Type: standing, twist
Drishti Point: 1. Urdhva or Antara Drishti (up to the sky)
2. Padayoragrai or Padayoragre (toes/feet)
ल व र य

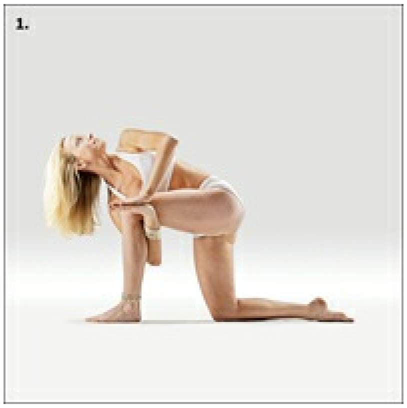

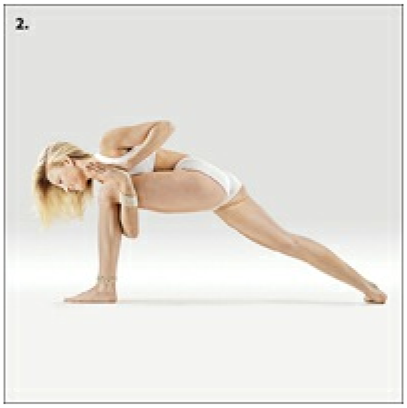

Revolved Side Angle Pose Prep.

Parivritta Parshva Konasana Prep.

(puh-ri-VRIT-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: hands in Anjali Mudra (Hands in Prayer)
1. back knee on the floor, looking up to the sky
2. back knee off the floor, looking up to the sky

Pose Type: standing, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व र य

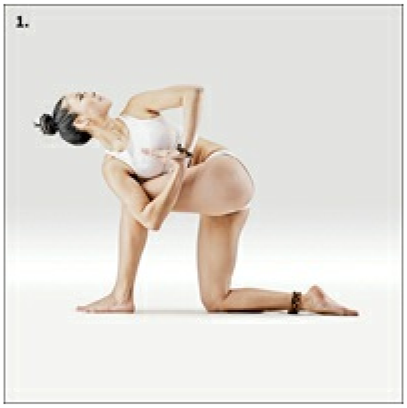

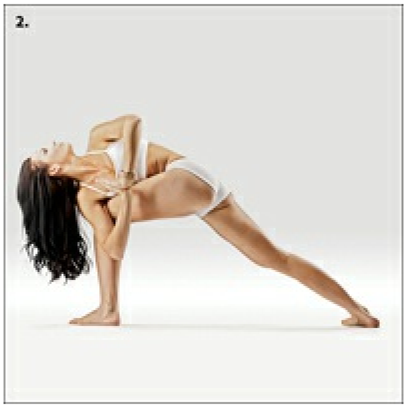

Both Hands Free Revolved Side Angle Pose

Dwi Mukta Hasta Parivritta Parshva Konasana

(DWI muk-tuh HUH-stuh puh-ri-VRIT-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: bottom hand off the floor, fingertips pointing in opposite directions
1. back knee on the floor
2. back knee off the floor

Pose Type: standing, twist, side bend
Drishti Point: Hastagrai or Hastagre (hands)
ल व र य

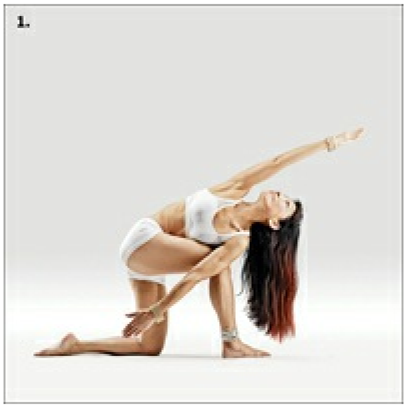

1.

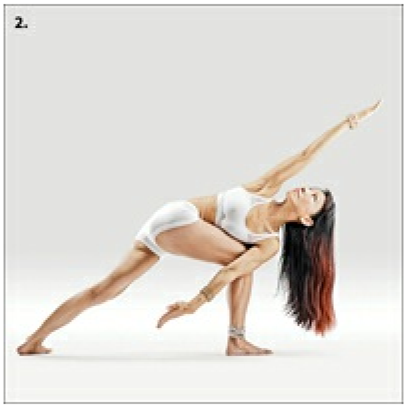

2.

Revolved Side Angle Pose Prep.

Parivritta Parshva Konasana Prep.

(puh-ri-VRIT-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: palm to the floor, other palm to lower back
Pose Type: standing, twist, side bend
Drishti Point: Hastagrai or Hastagre (hands) or Padhayoragrai or Padayoragre (toes/feet)

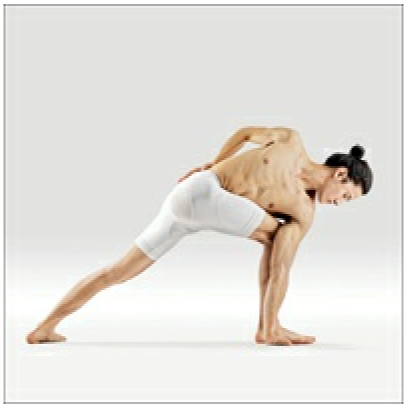

SIDE ANGLE POSE: REVOLVED BOUND

Revolved Side Angle Pose

Parivritta Parshva Konasana

(puh-ri-VRIT-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: 1. palm to the floor, back knee on the floor 2. palm to the floor, back knee off the floor
Pose Type: standing, twist, side bend
Drishti Point: Hastagrai or Hastagre (hands)
ल व र य

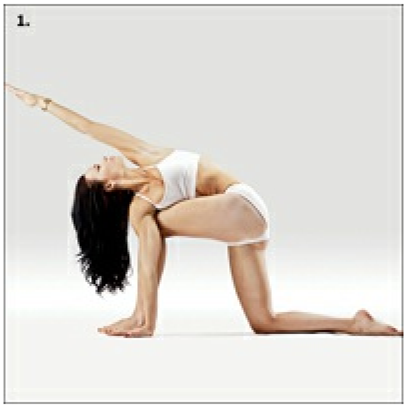

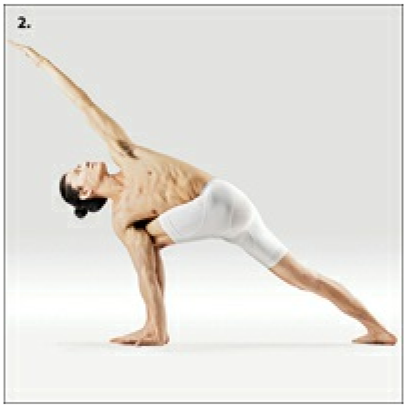

Bound Revolved Side Angle Pose

Baddha Parivritta Parshva Konasana

(BUH-duh puh-ri-VRIT-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Also Known As: Hands Bound Revolved Side Angle Pose Prep. (Baddha Hasta Parivritta Parshva Konasana Prep.)
Modification: 1. back knee on the floor 2. back knee off the floor
Pose Type: standing, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व र

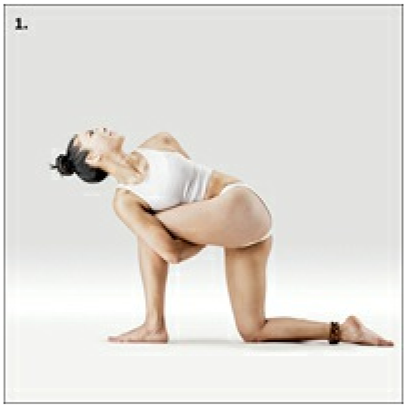

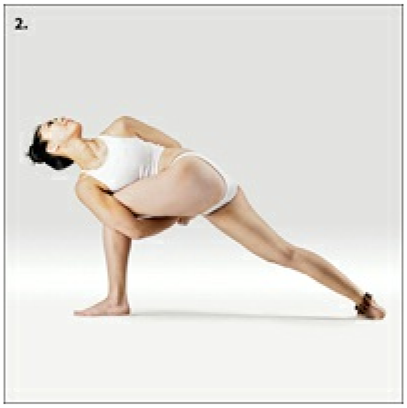

REVERSE WARRIOR

Both Arms Extended Reverse Warrior Pose Prep.

Dwi Hasta Utthita Viparita Virabhadrasana Prep.

(DWI-huh-stuh UT-ti-tuh vi-puh-REE-tuh vee-ruh-buh-DRAHS-uh-nuh)

Modification: back knee on the floor, arms straight, palms together
Pose Type: standing, backbend, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व र य ह

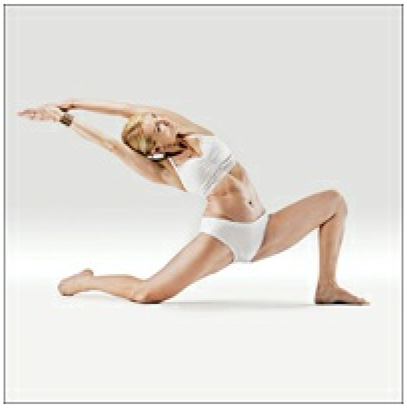

Tip Toe Both Arms Extended Reverse Warrior Pose

Prapada Dwi Hasta Utthita Viparita Virabhadrasana

(PRUH-puh-duh DWI-huh-stuh UT-ti-tuh vi-puh-REE-tuh vee-ruh-buh-DRAHS-uh-nuh)

Modification: arms straight, palms together
Pose Type: standing, backbend, side bend
Drishti Point: Bhrumadhye (third eye, between the eyebrows) or Hastagrai or Hastagre (hands)

ल व ह

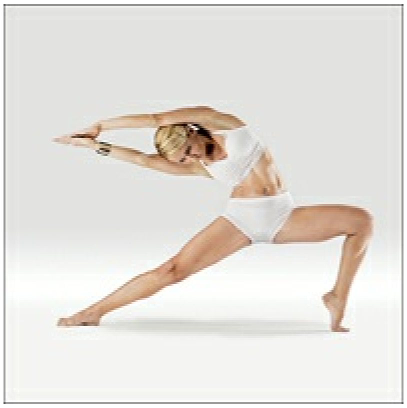

Revolved Sideways Tip Toe Reverse Warrior Pose

Parivritta Parshva Prapada Viparita Virabhadrasana

(puh-ri-VRIT-tuh PAHRSH-vuh PRUH-puh-duh vi-puh-REE-tuh vee-ruh-buh-DRAHS-uh-nuh)

Modification: front heel up, one arm in front and parallel to the floor; other elbow bent, palm facing up, hand to the crown of the head
Pose Type: standing, backbend, twist, side bend
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

ल व य

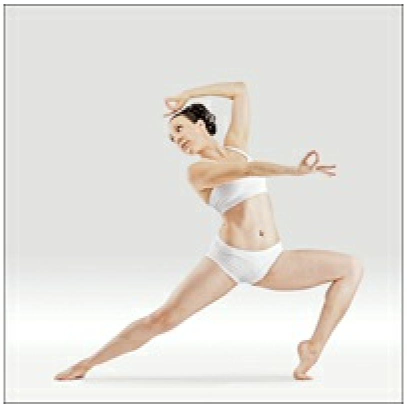

Tip Toe Reverse Warrior Pose

Prapada Viparita Virabhadrasana

(PRUH-puh-duh vi-puh-REE-tuh vee-ruh-buh-DRAHS-uh-nuh)

Modification: front heel up, both elbows bent, one arm up over the head, other arm in front of the torso

Pose Type: standing, backbend

Drishti Point: Hastagrai or Hastagre (hands)

ल व र य ह

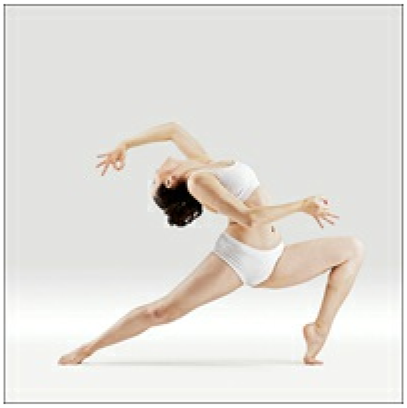

Half Bound Reverse Warrior Pose

Ardha Baddha Viparita Virabhadrasana

(UHR-duh BUH-duh vi-puh-REE-tuh vee-ruh-buh-DRAHS-uh-nuh)

Pose Type: standing, backbend, binding

Drishti Point: Hastagrai or Hastagre (hands)

ल व र य ह

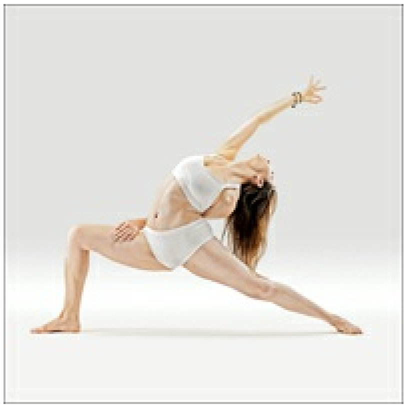

Reverse Warrior Pose

Viparita Virabhadrasana

(vi-puh-REE-tuh vee-ruh-buh-DRAHS-uh-nuh)

Also Known As: Reverse Warrior Pose 2 (Viparita Virabhadrasana 2)

Modification: hand to the calf of the back leg; low lunge, low stance
Pose Type: standing, backbend
Drishti Point: Hastagrai or Hastagre (hands)
ल व र य ह ॐ

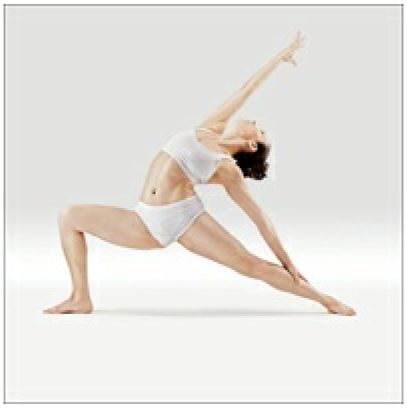

WARRIOR 1

Warrior 1 Prep.

Virabhadrasana 1 Prep.

(vee-ruh-buh-DRAHS-uh-nuh)
Modification: both legs straight, palms pressed together
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)
ल व

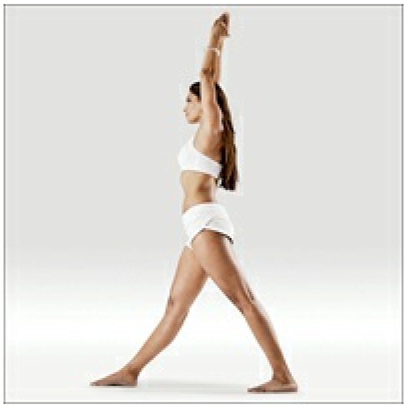

Warrior 1

Virabhadrasana 1

(vee-ruh-buh-DRAHS-uh-nuh)
Also Known As: Raised Arms Warrior Pose (Urdhva Hasta Veerasana), Hero Succession Series

(Vira Parampara)
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)
ल व

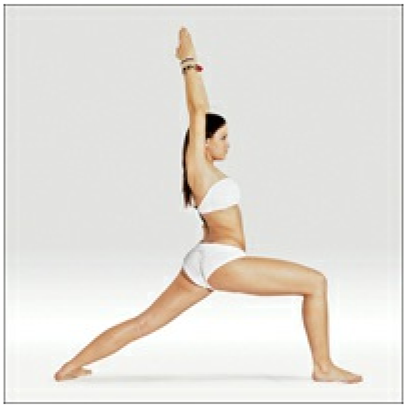

Warrior 1

Virabhadrasana 1

(vee-ruh-buh-DRAHS-uh-nuh)
Modification: arms out in front and parallel to the floor, palms facing the sky
Pose Type: standing
Drishti Point: Hastagrai or Hastagre (hands)
ल व

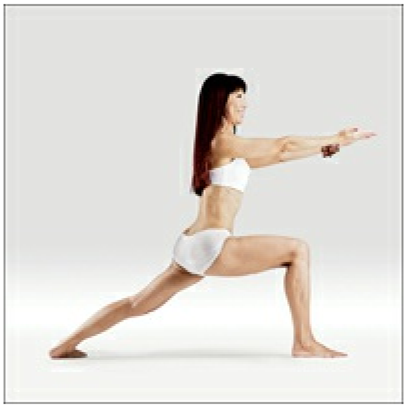

Bound Warrior 1

Baddha Virabhadrasana 1

(BUH-duh vee-ruh-buh-DRAHS-uh-nuh)
Modification: fingers interlocked on the inside of the front thigh
Pose Type: standing, binding
Drishti Point: Nasagrai or Nasagre (nose)

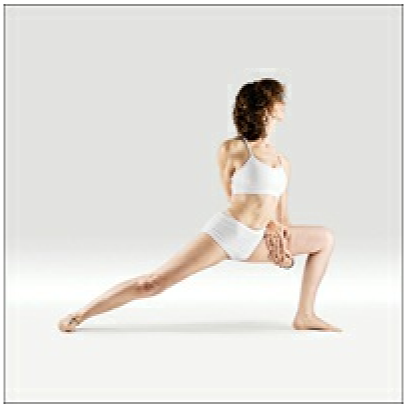

Warrior 1

Virabhadrasana 1

(vee-ruh-buh-DRAHS-uh-nuh)

Modification: arms up over the head and open to the sides; front leg, toes rotated out; deep backbend
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

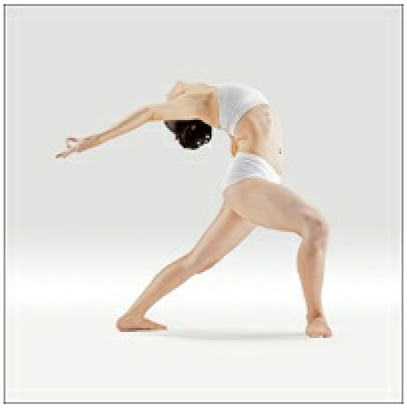

Open Heart Chakra Hands Bound Warrior

Anahata Chakra Baddha Hasta Virabhadrasana

(uh-NAH-huh-tuh CHUHK-ruh BUH-duh HUH-suh vee-ruh-buh-DRAHS-uh-nuh)

Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

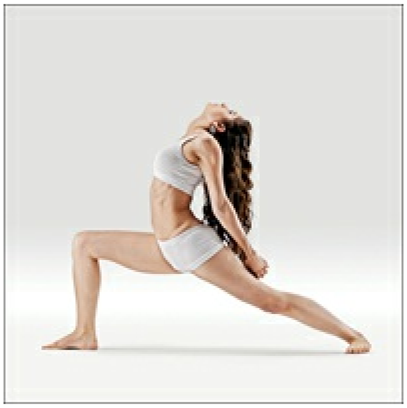

Raised Bound Hands in Warrior 1

Urdhva Baddha Hastasana in Virabhadrasana 1

(OORD-vuh BUH-duh huh-STAHS-uh-nuh in vee-ruh-buh-DRAHS-uh-nuh)

Modification: fingers interlocked deep backbend
Pose Type: standing, backbend
Drishti Point: Hastagrai or Hastagre (hands)
ल व र य ह

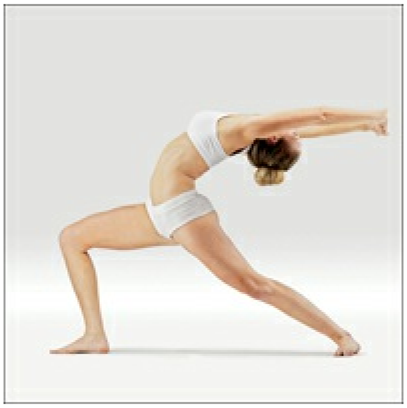

WARRIOR 1: FORWARD BEND—TO THE SIDE

Spread Out Hands Side Angle Pose

Prasarita Hasta Parshva Konasana

(pruh-SAH-ri-tuh HUH-stuh pahrsh-vuh-ko-NAHS-uh-nuh)

Also Known As: Down Dog Lunge Pose
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व ह

Hand Position of the Pose Dedicated to Garuda in Side Angle Pose

Hasta Garudasana in Parshva Konasana

(HUH-stuh guh-ru-DAHS-uh-nuh in pahrsh-vuh-ko-NAHS-uh-nuh)

Pose Type: standing, forward bend

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

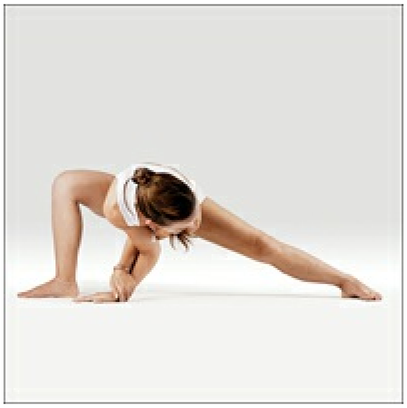

Hand Position of Cow Face Pose in Side Angle Pose

Hasta Gomukhasana in Parshva Konasana

(HUH-stuh go-muk-AHS-uh-nuh in pahrsh-vuh-ko-NAHS-uh-nuh)

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

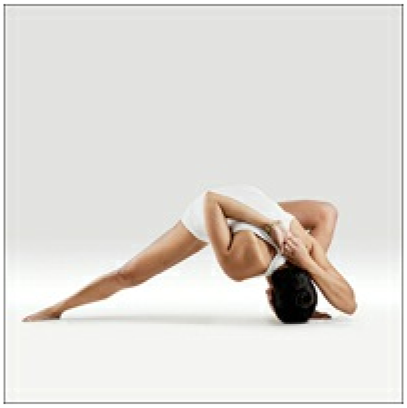

WARRIOR 1: FORWARD BEND—ARMS BEHIND

Bowing Warrior with Hands in Prayer

Nama Virabhadrasana Namaskar

(NUH-muh vee-ruh-buh-DRAHS-uh-nuh nuh-muhs-KAHR)

Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

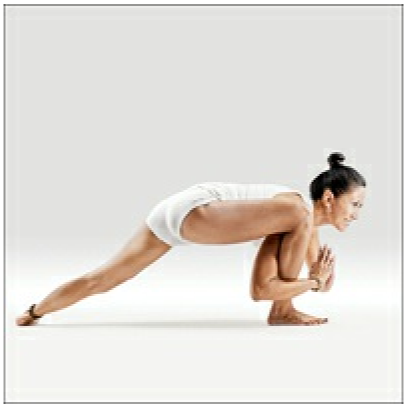

Bowing Hands Bound Warrior

Nama Baddha Hasta Virabhadrasana

(NUH-muh BUH-duh HUH-stuh vee-ruh-buh-DRAHS-uh-nuh)

Modification: grabbing onto the triceps behind the back
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

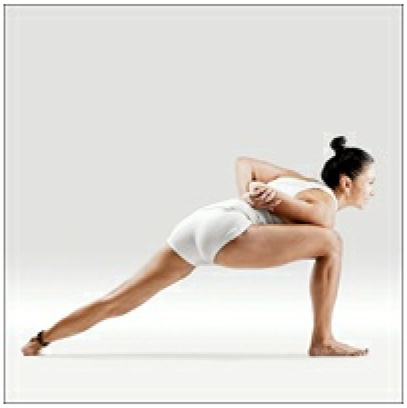

Bowing Reverse Prayer Warrior

Nama Viparita Namaskar Virabhadrasana

(NUH-muh vi-puh-REE-tuh nuh-muhs-KAHR vee-ruh-buh-DRAHS-uh-nuh)

Also Known As: Bowing Back of the Body Prayer Warrior (Nama Paschima Namaskara Virabhadrasana)
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व ह

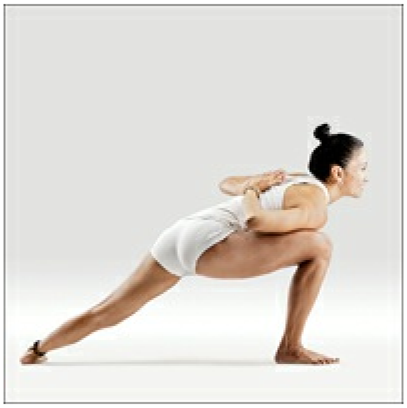

Bowing Hands Bound Warrior Pose

Nama Baddha Hasta Virabhadrasana

(NUH-muh BUH-duh HUH-stuh vee-ruh-buh-DRAHS-uh-nuh)

Also Known As: Warrior Veerastambana Pose (Veerastambanasana), Bound Hands Side Angle Pose (Baddha Hasta Parshva Konasana)
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व य ॐ

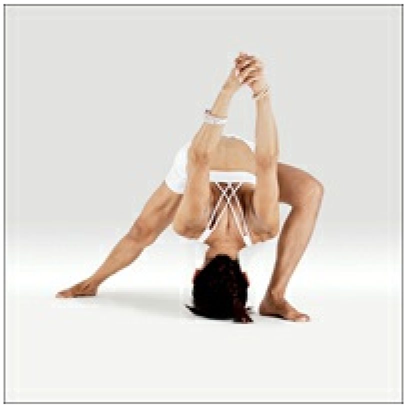

Bowing Hands Bound Warrior Pose

Nama Baddha Hasta Virabhadrasana

(NUH-muh BUH-duh HUH-stuh vee-ruh-buh-DRAHS-uh-nuh)

Also Known As: Bound Hands Side Angle Pose (Baddha Hasta Parshva Konasana)
Modification: arms bound around the front leg
Pose Type: standing, forward bend, binding
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)
ल व य ह

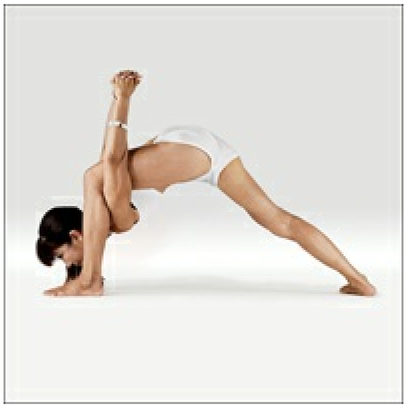

SIDE ANGLE POSE: HEAD BEHIND LEG

One Foot Behind the Head Side Angle Pose

Eka Pada Shirsha Parshva Konasana

(EY-kuh PUH-duh SHEER-shuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: arms crossed in front of the chest
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व

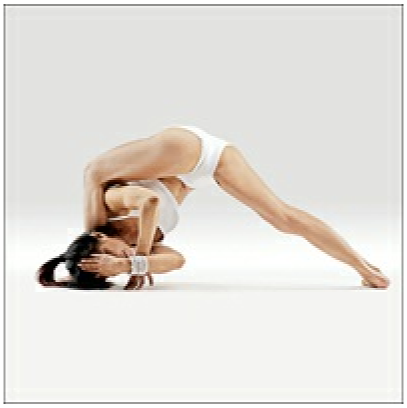

One Foot Behind the Head Side Angle Pose

Eka Pada Shirsha Parshva Konasana

(EY-kuh PUH-duh SHEER-shuh pahrsh-vuh-ko-NAHS-uh-nuh)

Modification: palms together, arms straight on the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

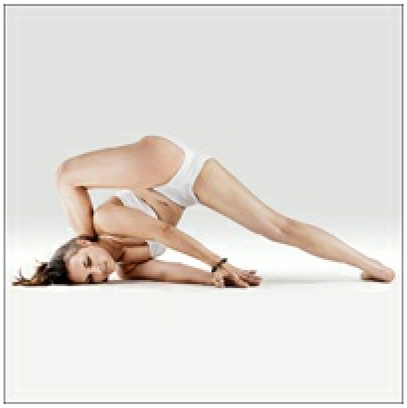

One Foot Behind the Head Bound Side Angle Pose

Eka Pada Shirsha Baddha Parshva Konasana

(EY-kuh PUH-duh SHEER-shuh BUH-duh pahrsh-vuh-ko-NAHS-uh-nuh)

Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

LUNGE: BACK KNEE ON THE FLOOR

Equestrian Riding Horse Lunge Pose

Ashva Sanchalanasana

(UHSH-vuh suhn-chuh-luh-NAHS-uh-nuh)

Modification: back knee on the floor, hands to the floor on either side of the front foot, looking up to the sky
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र ह ॐ

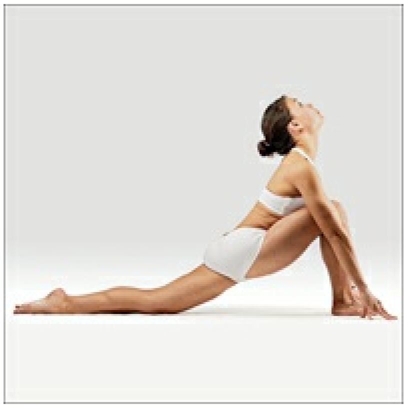

Equestrian Riding Horse Lunge Pose Prep.

Ashva Sanchalanasana Prep.

(UHSH-vuh suhn-chuh-luh-NAHS-uh-nuh)

Also Known As: Son of Anjani (Lord Hanuman) Lunge Pose Prep. (Anjaneyasana Prep.)
Modification: back knee on the floor, both hands on the front knee
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

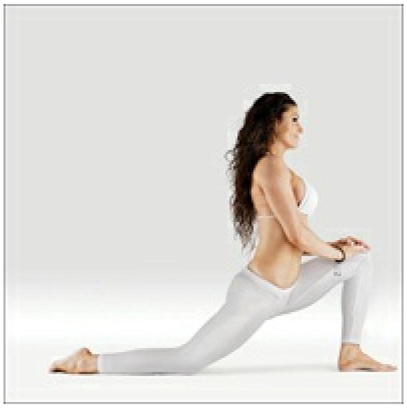

LUNGE: BACK KNEE ON THE FLOOR—FORWARD BEND

Hands Bound Extended Lizard Tail Lunge Pose

Baddha Hasta Uttana Pristhasana

(BUH-duh HUH-stuh ut-TAHN-uh prish-TAHS-uh-nuh)

Also Known As: Bowing Warrior Pose Prep. (Nama Virabhadrasana Prep.)
Modification: back knee on the floor, forehead to the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

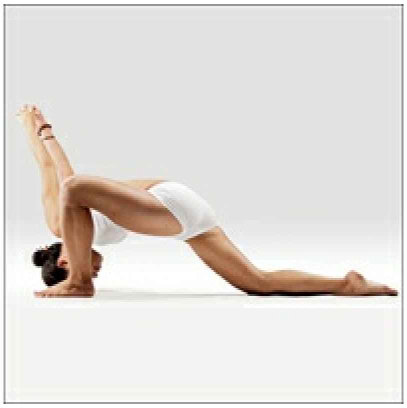

Tip Toe Extended Lizard Tail Lunge Pose

Prapada Uttana Pristhasana

(PRUH-puh-duh ut-TAHN-uh prish-TAHS-uh-nuh)

Modification: back knee on the floor, shoulder to the back of the knee, one arm straight out in front, other arm straight to the back, forehead to the floor
Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

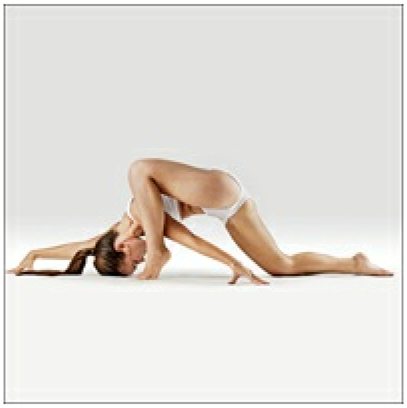

Extended Lizard Tail Lunge Pose

Uttana Pristhasana

(ut-TAHN-uh prish-TAHS-uh-nuh)

Modification: back knee on the floor
1. forearms to the floor on the inside of the foot
2. forearms to the floor on either side of the foot
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

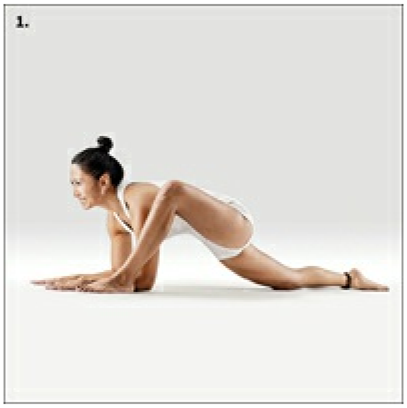

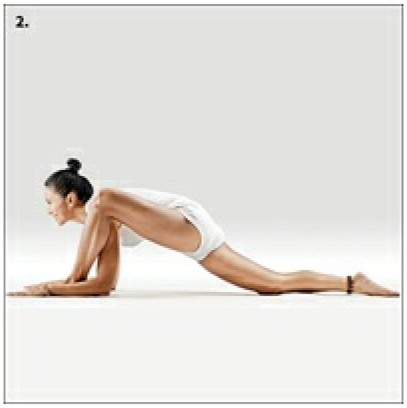

Unsupported Extended Lizard Tail Lunge Pose

Niralamba Uttana Pristhasana

(nir-AH-luhm-buh ut-TAHN-uh prish-TAHS-uh-nuh)

Also Known As: God Favour Seeking Sacrifice Ritual Pose (Yajnasana), Christ’s Cross Pose
Modification: arms straight out to the sides
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

लं वं हं

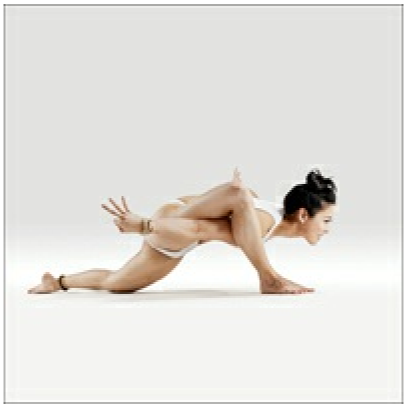

Bound Extended Lizard Tail Lunge Pose

Baddha Uttana Pristhasana

(BUH-duh ut-TAHN-uh prish-TAHS-uh-nuh)

Also Known As: Bound God Favor Seeking Sacrifice Ritual Pose (Baddha Yajnasana), Bound Christ’s Cross Pose)
Modification: back knee on the floor
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

लं वं हं

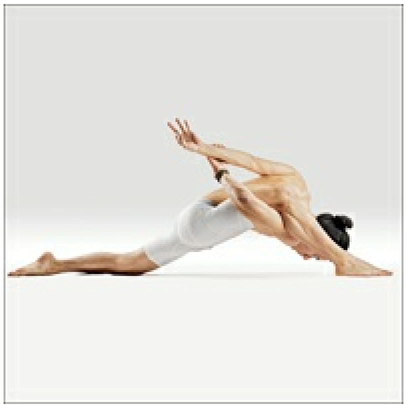

Hands Spread Out Extended Lizard Tail Lunge Pose

Prasarita Hasta Uttana Pristhasana

(pruh-SAH-ri-tuh HUH-stuh ut-TAHN-uh prish-TAHS-uh-nuh)

Also Known As: God Favour Seeking Sacrifice Ritual Pose (Yajnasana), Christ’s Cross Pose
Modification: back knee on the floor, shin to the tricep
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

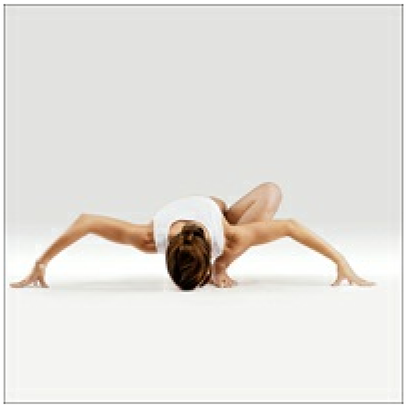

LUNGE: BACK KNEE ON THE FLOOR—BACKBEND—ARMS BEHIND THE BACK

Bound Hands Son of Anjani (Lord Hanuman) Lunge Pose

Baddha Hasta Anjaneyasana *(BUH-duh HUH-stuh uhn-juh-ney-AHS-uh-nuh)*

Also Known As: Open Heart Chakra Hands Bound Equestrian Riding Horse Pose (Anahata Chakra Baddha Hasta Ashva Sanchalanasana)
Modification: back knee on the floor, backbend, hands to the floor on the inside of the back knee
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

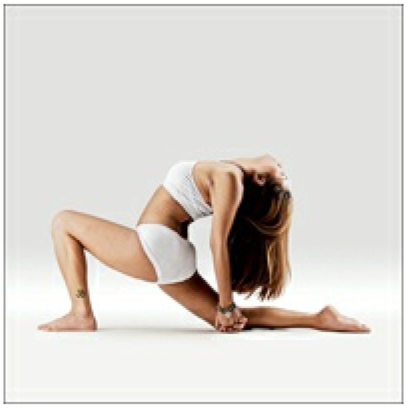

Son of Anjani (Lord Hanuman) Lunge Pose

Anjaneyasana *(uhn-juh-ney-AHS-uh-nuh)*
Also Known As: Open Heart Chakra Equestrian Riding Horse Pose (Anahata Chakra Ashva Sanchalanasana)
Modification: back knee on the floor, backbend, fingertips to the floor on either side of the back knee
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

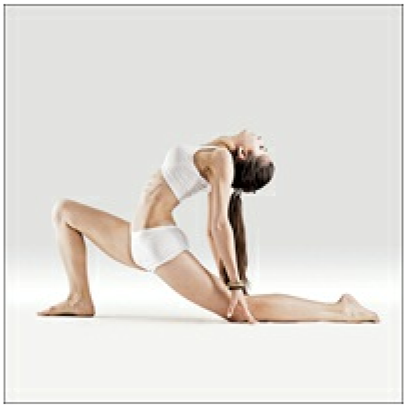

One Hand Son of Anjani (Lord Hanuman) Lunge Pose

Eka Hasta Anjaneyasana *(EY-kuh HUH-stuh uhn-juh-ney-AHS-uh-nuh)*
Also Known As: Equestrian Riding Horse Pose (Ashva Sanchalanasana)
Modification: back knee on the floor, one arm extended up over the head, other hand to the floor
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

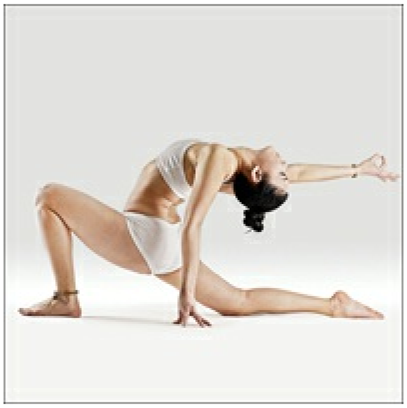

LUNGE: BACK KNEE ON THE FLOOR—BACKBEND—ARMS UP OVER THE HEAD

Son of Anjani (Lord Hanuman) Lunge Pose

Anjaneyasana *(uhn-juh-ney-AHS-uh-nuh)*
Also Known As: Crescent Lunge Pose
Modification: front heel on the floor, back knee down, back toes curled in; arms shoulder width apart, fingertips up to the sky
Pose Type: standing, mild backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

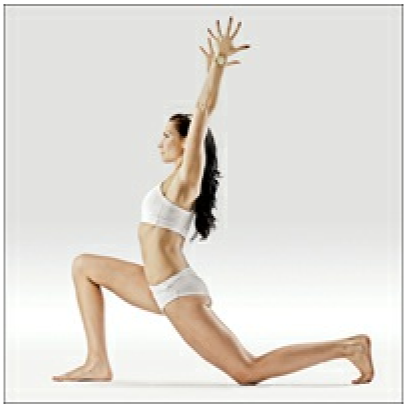

Tip Toe Son of Anjani (Lord Hanuman) Lunge Pose

Prapada Anjaneyasana *(PRUH-puh-duh uhn-juh-ney-AHS-uh-nuh)*
Also Known As: Tip Toe Crescent Lunge Pose
Modification: front heel off the floor, back knee down, back toes curled in; arms shoulder width apart, fingertips up to the sky
Pose Type: standing, mild backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

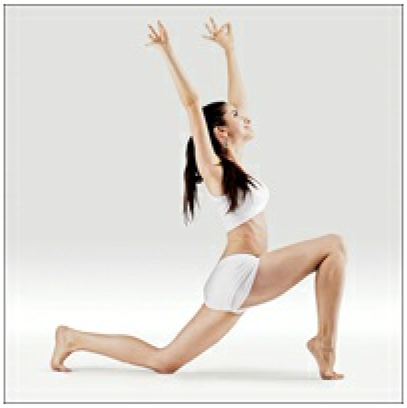

Son of Anjani (Lord Hanuman) Lunge Pose

Anjaneyasana *(uhn-juh-ney-AHS-uh-nuh)*
Also Known As: Monkey Pose (Kapyasana), Crescent Lunge Pose
Modification: arms up over the head, back knee on the floor, palms pressed together
Pose Type: standing, mild backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angusthamadhye or Angustha Ma Dyai (thumbs)

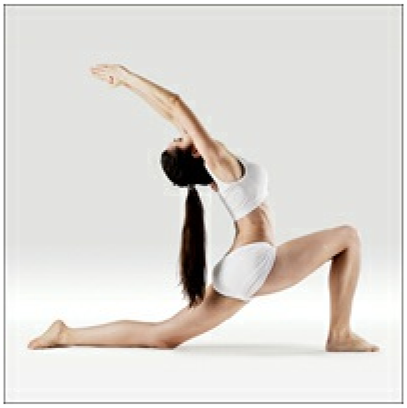

Standing Pose of the Heavenly Spirits Prep.

Stiti Valakhilyasana Prep. *(STI-ti VAH-luh-kil-YAHS-uh-nuh)*
Also Known As: Nindra Valakhilyasana; Son of Anjani (Lord Hanuman) Lunge Pose (Anjaneyasana), Crescent Lunge Pose)
Modification: back knee on the floor, arms up over the head, palms open to the sky, arms parallel to the floor
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

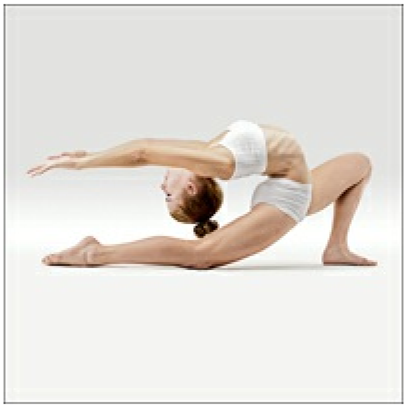

Standing Pose of the Heavenly Spirits

Stiti Valakhilyasana *(STI-ti VAH-luh-kil-YAHS-uh-nuh)*
Also Known As: Nindra Valakhilyasana
Modification: back knee to the floor, grabbing the back heel with overhead grip
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

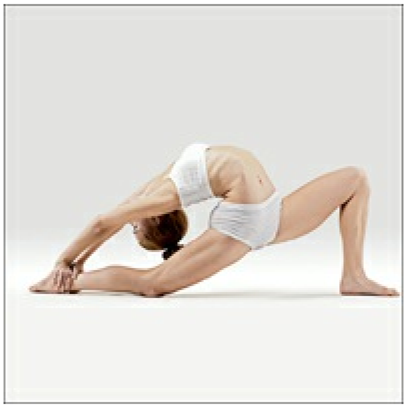

LUNGE: BACK KNEE ON THE FLOOR—TWISTS AND SIDE BENDS

Sideways Son of Anjani (Lord Hanuman) Lunge Pose

Parshva Anjaneyasana
(PAHRSH-vuh uhn-juh-ney-AHS-uh-nuh)
Modification: back knee on the floor; one hand to the inside of the opposite knee, other arm up over the head
Pose Type: standing, side bend, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व र य

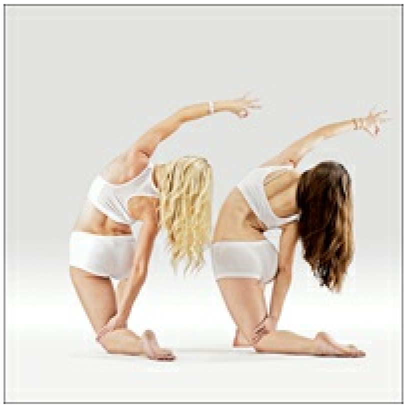

Revolved Paying Homage Lunge Pose

Parivritta Namasyasana

(puh-ri-VRIT-tuh nuh-muh-SYAHS-uh-nuh)

Also Known As: Curtsey Lunge
Modification: front leg crossed over to the side, opposite elbow over the front knee, back knee on the floor
Pose Type: standing, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
लं वं रं

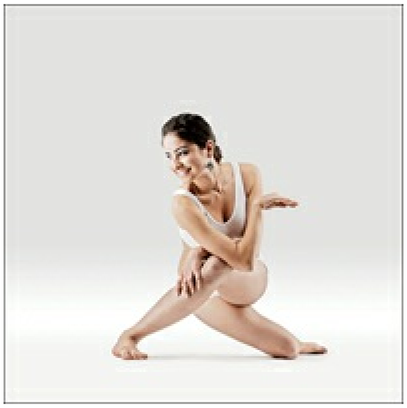

Revolved Son of Anjani (Lord Hanuman) Lunge Pose

Parivritta Anjaneyasana

(puh-ri-VRIT-tuh uhn-juh-ney-AHS-uh-nuh)

Modification: back knee on the floor, toes curled in; one hand to the floor, other arm up in the sky; twisting to the inside of the front knee
Pose Type: standing, twist
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
लं वं

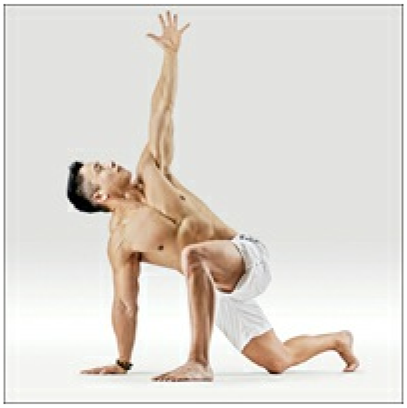

Tip Toe Sideways Son of Anjani (Lord Hanuman) Lunge Pose

Prapada Parshva Anjaneyasana

(PRUH-puh-duh PAHRSH-vuh uhn-juh-ney-AHS-uh-nuh)

Modification: back knee on the floor, toes pointed to the back, one hand to the floor, other arm up in the sky, twisting to the outside of the front knee
Pose Type: standing, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

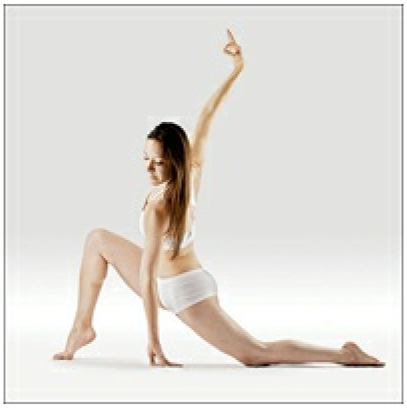

Revolved Son of Anjani (Lord Hanuman) Lunge Pose

Parivritta Anjaneyasana

(puh-ri-VRIT-tuh uhn-juh-ney-AHS-uh-nuh)

Also Known As: Revolved Monkey Pose (Parivritta Anjaneyasana)
Modification: back knee on the floor
Pose Type: standing, twist
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

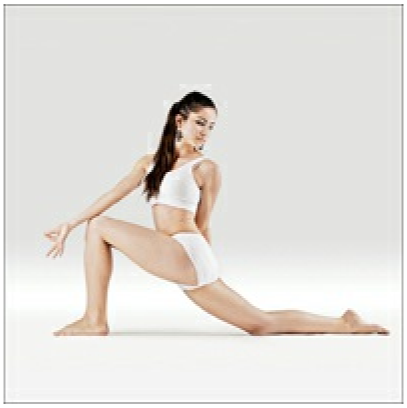

Tip Toe Revolved Son of Anjani (Lord Hanuman) Lunge Pose

Prapada Parivritta Anjaneyasana

(PRUH-puh-duh puh-ri-VRIT-tuh uhn-juh-ney-AHS-uh-nuh)

Pose Type: standing, twist

Drishti Point: Nasagrai or Nasagre (nose)

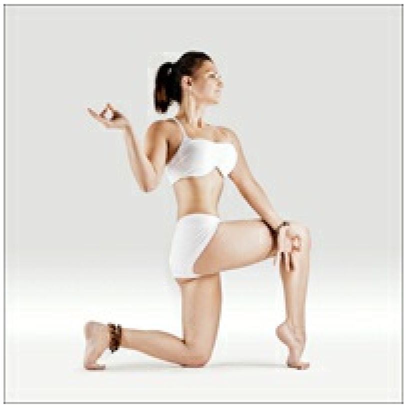

Horse Pose Prep.

Vatayanasana Prep.

(vah-tah-yuh-NAHS-uh-nuh)

Modification: back knee and foot on the floor

Pose Type: standing

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

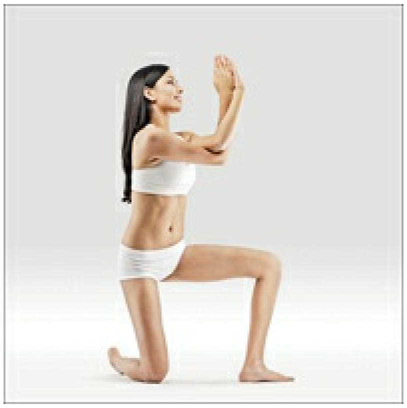

Fighting Warrior Pose 1

Yudhasana 1

(yu-DAHS-uh-nuh)

Pose Type: standing

Drishti Point: Hastagrai or Hastagre (hands)

ल व

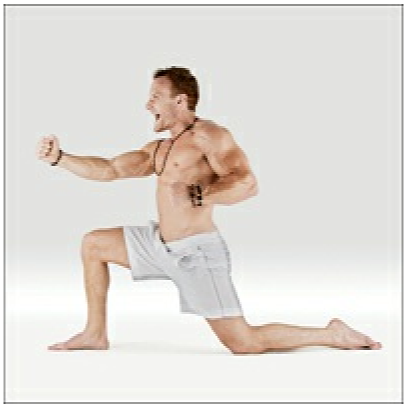

Fighting Warrior Pose 2

Yudhasana 2

(yu-DAHS-uh-nuh)

Modification: back knee on the floor

Pose Type: standing

Drishti Point: Hastagrai or Hastagre (hands)

ल व

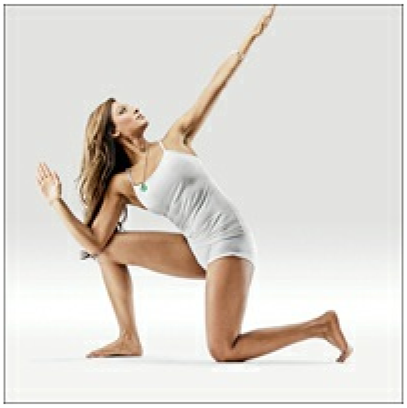

LUNGE: BACK KNEE ON THE FLOOR UNDER THE HIP SOCKET—FORWARD BEND

Half Equestrian Riding Horse Lunge Pose

Ardha Ashva Sanchalanasana

(UHR-duh UHSH-vuh suhn-chuh-luh-NAHS-uh-nuh)

Modification: spine in Cat Tilt, back knee down, toes pointing away from the head
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

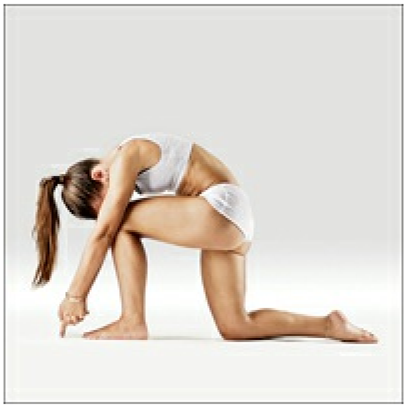

Tip Toe Bound Son of Anjani (Lord Hanuman) Lunge Pose

Prapada Baddha Anjaneyasana

(PRUH-puh-duh BUH-duh uhn-juh-ney-AHS-uh-nuh)

Modification: front leg bound, back knee on the floor, toes pointing away from the head
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

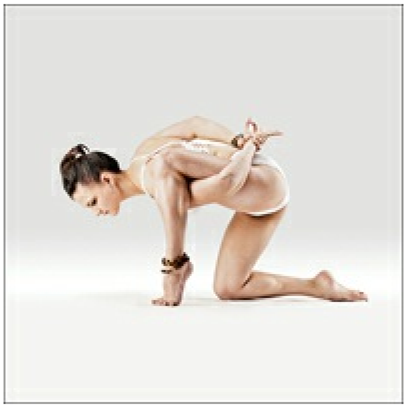

One Foot Behind the Head Extended Lizard Tail Lunge Pose

Eka Pada Shirsha Uttana Pristhasana

(EY-kuh PUH-duh SHEER-shuh ut-TAHN-uh prish-TAHS-uh-nuh)

Modification: back knee under the hip, toes pointing away from the head
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

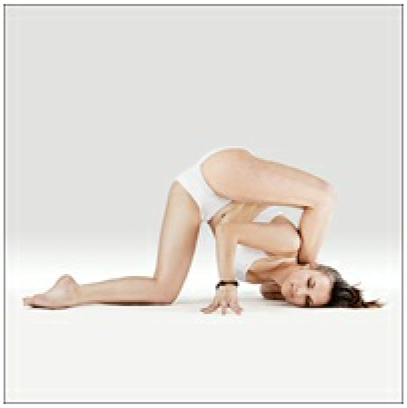

One Foot Behind the Head Extended Lizard Tail Lunge Pose

Eka Pada Shirsha Uttana Pristhasana

(EY-kuh PUH-duh SHEER-shuh ut-TAHN-uh prish-TAHS-uh-nuh)

Modification: legs working toward full vertical splits, toes of the back leg curled in
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

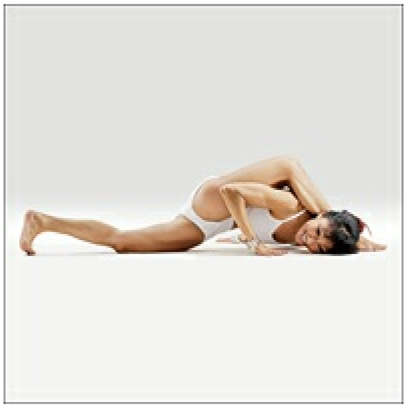

LUNGE: BACK KNEE ON THE FLOOR UNDER THE HIP SOCKET—BACKBEND

Son of Anjani (Lord Hanuman) Lunge Pose

Anjaneyasana

(uhn-juh-ney-AHS-uh-nuh)

Modification: deep backbend; chest open, arms to the side, elbows bent; back knee on the floor, toes pointing away from the head
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

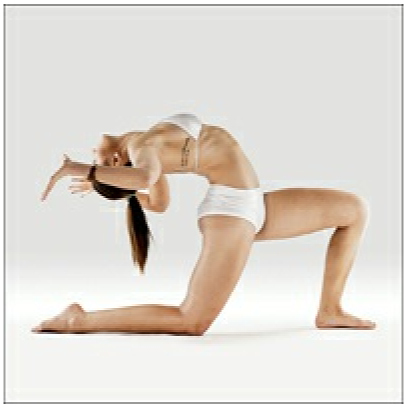

Tip Toe One Hand to Foot Son of Anjani (Lord Hanuman) Lunge Pose

Prapada Eka Hasta Pada Anjaneyasana

(PRUH-puh-duh EY-kuh HUH-stuh PUH-duh uhn-juh-ney-AHS-uh-nuh)

Modification: deep backbend; hand to the ankle on the same side
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

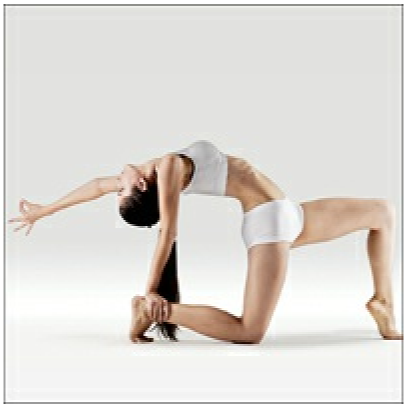

Both Hands to Foot Son of Anjani (Lord Hanuman) Lunge Pose

Dwi Hasta Pada Anjaneyasana

(DWI-huh-stuh PUH-duh uhn-juh-ney-AHS-uh-nuh)

Modification: backbend
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

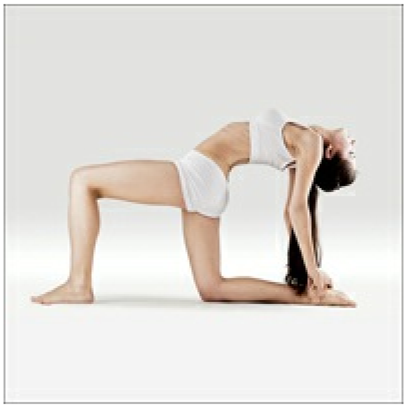

Tip Toe Both Hands to Foot Son of Anjani (Lord Hanuman) Lunge Pose

Prapada Dwi Hasta Pada Anjaneyasana

(PRUH-puh-duh DWI-huh-stuh PUH-duh uhn-juh-ney-AHS-uh-nuh)

Modification: deep backbend, back knee on the floor, back toes curled in
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

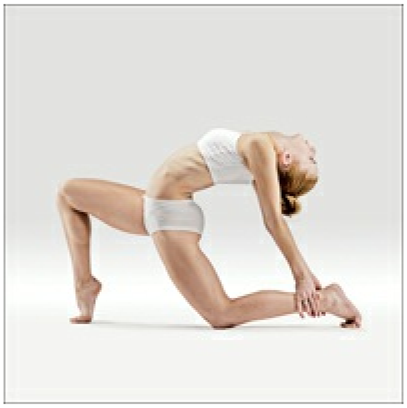

LUNGE: BACK KNEE ON THE FLOOR—FOOT OFF THE FLOOR

One-Legged King Pigeon Pose 2 Prep.

Eka Pada Raja Kapotasana 2 Prep.

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: arms open wide and raised
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

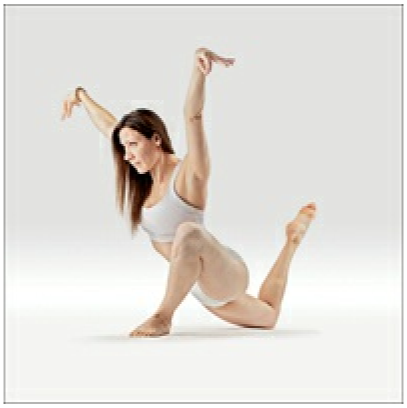

Half Bound Revolved One-Legged King Pigeon Pose 2 Prep.

Ardha Baddha Parivritta Eka Pada Raja Kapotasana 2 Prep.

(UHR-duh BUH-duh puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: standing, twist, binding

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

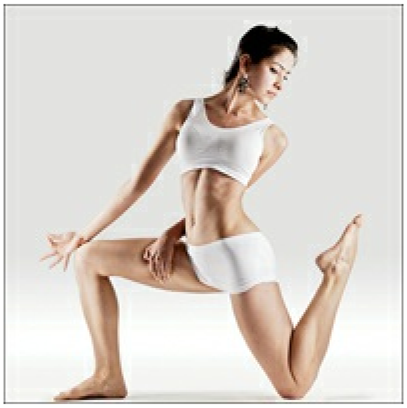

Tip Toe Extended Lizard Tail Lunge Pose Prep.

Prapada Uttana Pristhasana Prep.

(PRUH-puh-duh ut-TAHN-uh prish-TAHS-uh-nuh)

Modification: arms straight to semi-straight, back foot lifted off the floor

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

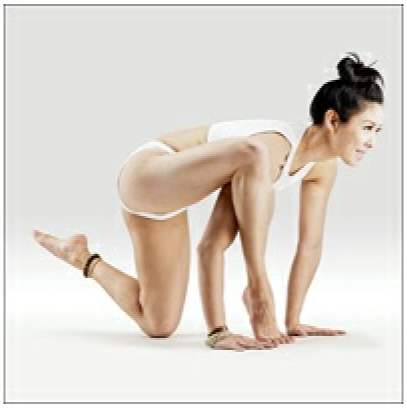

Bound Revolved Side Angle Pose Prep.

Baddha Parivritta Parshva Konasana Prep.

(BUH-duh puh-ri-VRIT-tuh pahrsh-vuh-ko-NAHS-uh-nuh)

Also Known As: Bound Revolved Son of Anjani (Lord Hanuman) Lunge Pose Prep. (Anjaneyasana Prep.)
Modification: back knee bent, heel toward the sitting bone
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

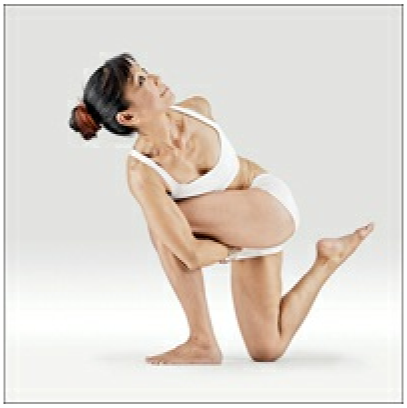

Foot to Knee Revolved Son of Anjani (Lord Hanuman) Lunge Pose with Hands in Prayer

Janu Pada Parivritta Anjaneyasana Namaskar

(JAH-nu PUH-duh puh-ri-VRIT-tuh uhn-juh-ney-AHS-uh-nuh nuh-muhs-KAHR)

Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

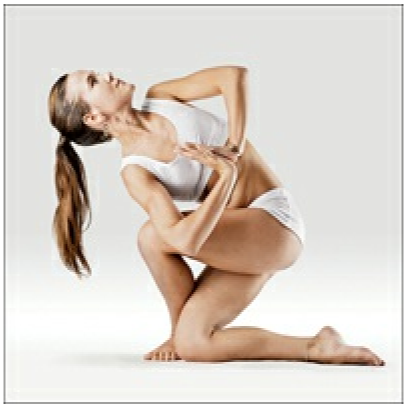

LUNGE: BACK KNEE ON THE FLOOR—GRABBING ONTO THE BACK FOOT

Tip Toe One-Legged King Pigeon Pose 2 Prep.

Prapada Eka Pada Raja Kapotasana 2 Prep.

(PRUH-puh-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: both hands grabbing onto the back foot with under-head grip
Pose Type: standing, mild backbend
Drishti Point: Nasagrai or Nasagre (nose)

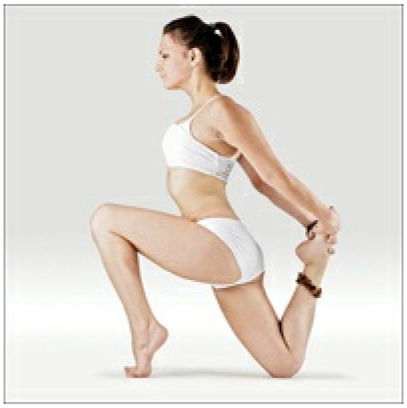

Revolved One-Legged King Pigeon Pose 2 Prep.

Parivritta Eka Pada Raja Kapotasana 2 Prep.

(puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: grabbing onto the back foot with the opposite hand using an under-head grip, twisting toward the front knee
Pose Type: standing, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

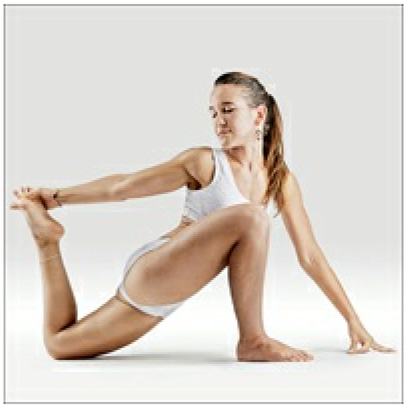

One Hand in Prayer Revolved One-Legged King Pigeon Pose 2 Prep.

Eka Hasta Namaskar Parivritta Eka Pada Raja Kapotasana 2 Prep.

(EY-kuh HUH-stuh nuh-muhs-KAHR puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: opposite hand to the back foot, other hand to the heart
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

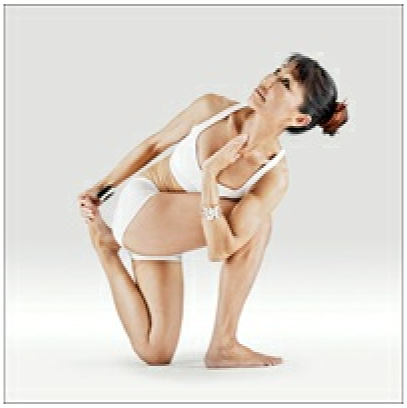

One Hand to Foot Extended Lizard Tail Lunge Pose

Eka Hasta Pada Uttana Pristhasana

(EY-kuh HUH-stuh PUH-duh ut-TAHN-uh prish-TAHS-uh-nuh)

Modification: forward bend, elbow to the floor, grabbing onto the back foot with opposite hand, heel to the glutes
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

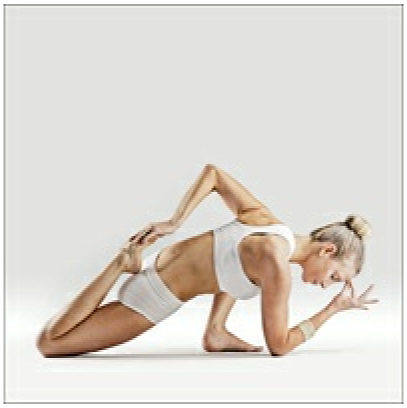

LUNGE: BACK KNEE ON THE FLOOR—FOOT TO THE HIP SOCKET

Spear Pose Prep.

Kuntasana Prep.

(kun-TAHS-uh-nuh)

Modification: grabbing onto the foot with one hand, arm crossed over in front, other arm parallel to the floor, foot away from the hip
Pose Type: standing, twist
Drishti Point: Hastagrai or Hastagre (hands)
ल व य

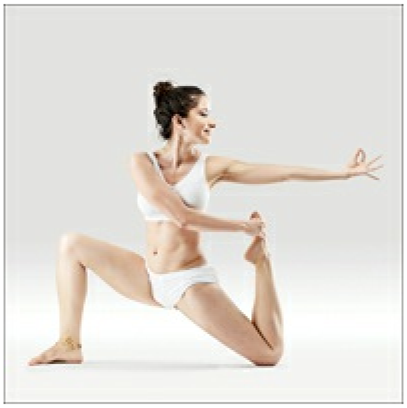

Spear Pose

Kuntasana

(kun-TAHS-uh-nuh)

Modification: grabbing onto the foot with one hand, arm crossed over in front, other hand to the knee, foot close to the hip
Pose Type: standing, twist
Drishti Point: Hastagrai or Hastagre (hands)

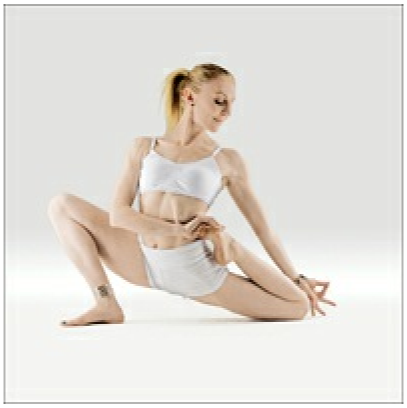

Both Hands Bound Revolved Spear Pose Prep.

Dwi Hasta Baddha Parivritta Kuntasana Prep.

(dwi HUH-stuh BUH-duh puh-ri-VRIT-tuh kun-TAHS-uh-nuh)

Pose Type: standing, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

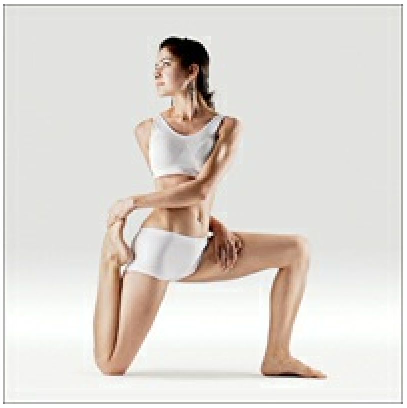

Spear Pose with Hands in Prayer

Kuntasana Namaskar

(kun-TAHS-uh-nuh nuh-muhs-KAHR)

Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

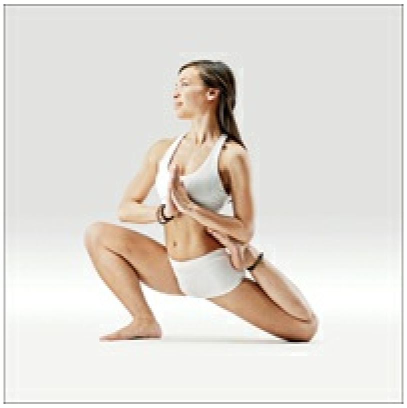

Half Bound Revolved Spear Pose Prep.

Ardha Baddha Parivritta Kuntasana Prep.

(UHR-duh-BUH-duh puh-ri-VRIT-tuh kun-TAHS-uh-nuh)

Modification: foot to the inside of the elbow
Pose Type: standing, twist, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

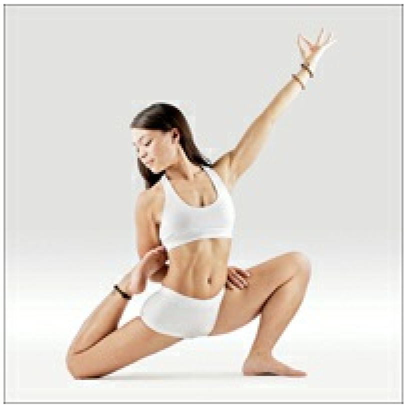

Spear Pose

Kuntasana

(kun-TAHS-uh-nuh)

Pose Type: standing
Drishti Point: Hastagrai or Hastagre (hands)

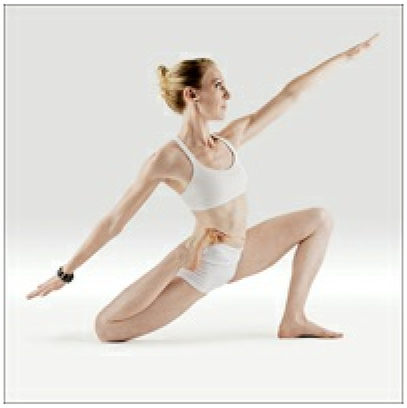

Hands Bound Ankle Stretch Pose

Baddha Hasta Kulpasana

(BUH-duh HUH-stuh kul-PAHS-uh-nuh)

Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

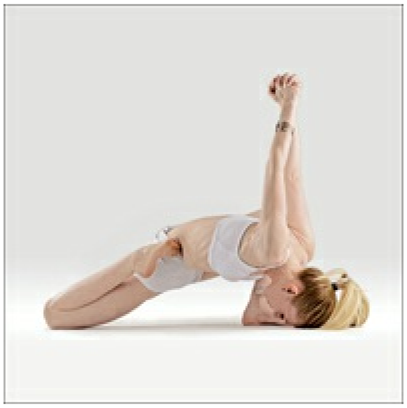

LUNGE: BACK KNEE ON THE FLOOR—HEAD BEHIND THE FRONT LEG, GRABBING ONTO THE BACK FOOT

One Foot Behind the Head Extended Lizard Tail Lunge Pose

Eka Pada Shirsha Uttana Pristhasana

(EY-kuh PUH-duh SHEER-shuh ut-TAHN-uh prish-TAHS-uh-nuh)

Modification: grabbing onto the foot with the opposite hand, back knee under the hips
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

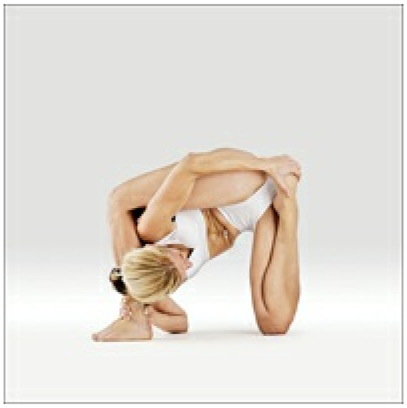

One Foot Behind the Head Extended Lizard Tail Lunge Pose

Eka Pada Shirsha Uttana Pristhasana

(EY-kuh PUH-duh SHEER-shuh ut-TAHN-uh prish-TAHS-uh-nuh)

Modification: grabbing onto the foot with the opposite hand, chest and quadriceps on the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

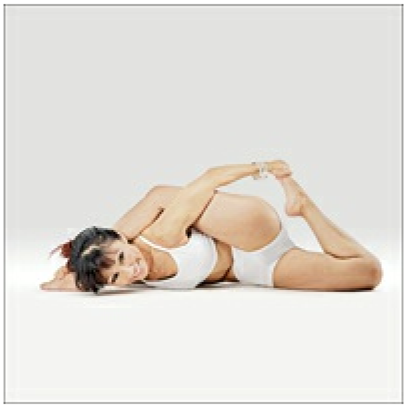

LUNGE: BACK KNEE ON FLOOR—ONE-LEGGED KING PIGEON 2

One-Legged King Pigeon Pose 2 Prep.

Eka Pada Raja Kapotasana 2 Prep.

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: back knee bent toward the glutes, front heel down; arms on the sides, spine straight
Pose Type: standing
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

Tip Toe One-Legged King Pigeon Pose 2 Prep.

Prapada Eka Pada Raja Kapotasana 2 Prep.

(PRUH-puh-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: back knee bent toward the glutes, heel of the top foot up; hands on the hips, backbend
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

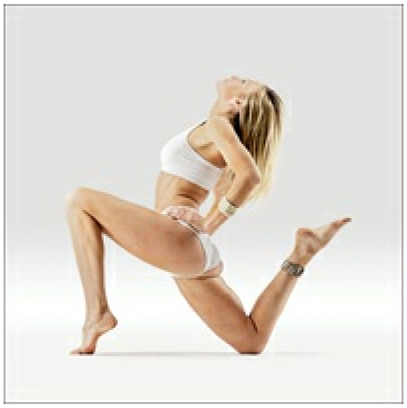

One-Legged King Pigeon Pose 2

Eka Pada Raja Kapotasana 2

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: prep.—arms up over the head, back knee bent toward the glutes
Pose Type: standing, mild backbend
Drishti Point: Nasagrai or Nasagre (nose)
ल व य

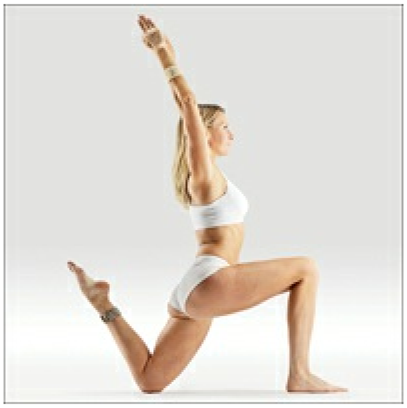

Tip Toe One-Legged King Pigeon Pose 2 Prep.

Prapada Eka Pada Raja Kapotasana 2 Prep.

(PRUH-puh-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: fingertips to the floor, front heel lifted

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

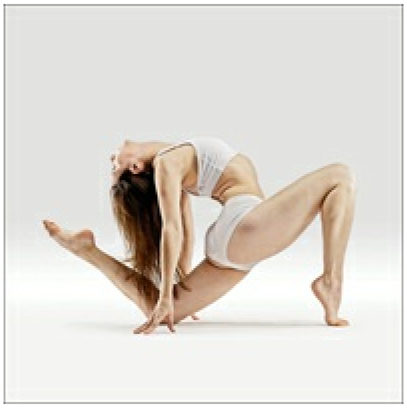

Intense Ankle Stretch One-Legged King Pigeon Pose 2

Uttana Kulpa Eka Pada Raja Kapotasana 2

(ut-TAH-nuh kul-puh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: prep.—fingertips to the floor, front toes curled under

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

LUNGE: BACK KNEE ON FLOOR—ONE-LEGGED KING PIGEON 2

One-Legged King Pigeon Pose 2 Prep.

Eka Pada Raja Kapotasana 2 Prep.

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: 1. grabbing onto the foot on the same side
2. grabbing onto the foot with the hand on the same side, foot toward the hip
3. grabbing onto the foot with the opposite hand, foot twisted toward the opposite glute
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं रं यं हं ॐ

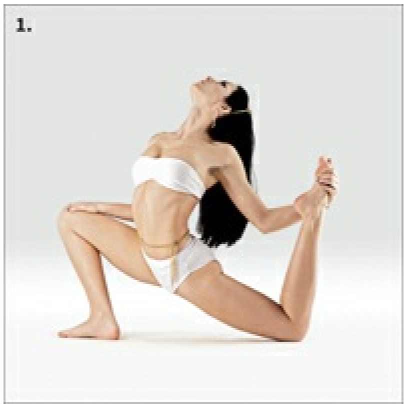

1.

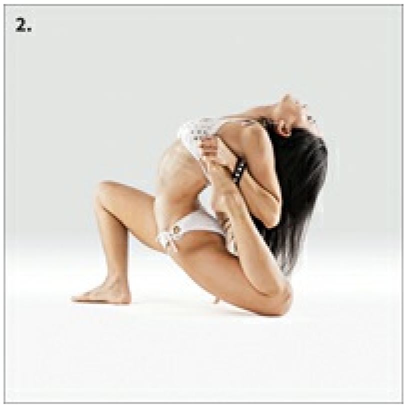

2.

Revolved One-Legged King Pigeon Pose 2

Parivritta Eka Pada Raja Kapotasana 2

(puh-ree-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: standing, forward bend, twist

Drishti Point: Parsva Drishti (to the right), Parsva Drishti (to the left)

● ●

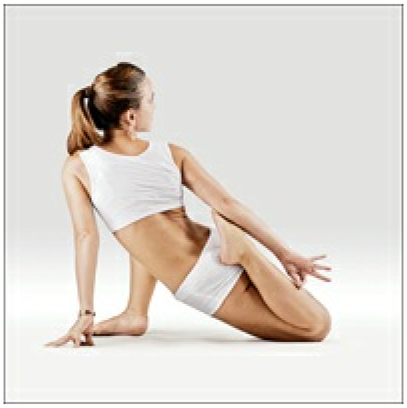

Mermaid Pose 2

Naginyasana 2

(nuh-gin-YAHS-uh-nuh)

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ● ● ● ● ●

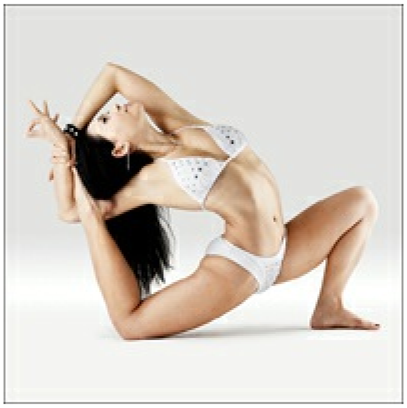

LUNGE: BACK KNEE ON FLOOR—ONE-LEGGED KING PIGEON 2—OVERHEAD GRIP

One-Legged King Pigeon Pose 2

Eka Pada Raja Kapotasana 2

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: grabbing onto the foot with the opposite hand, palm of the other hand to the floor on the side
1. foot to the head
2. heel to the forehead
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

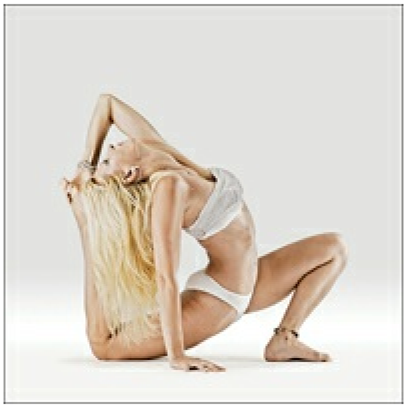

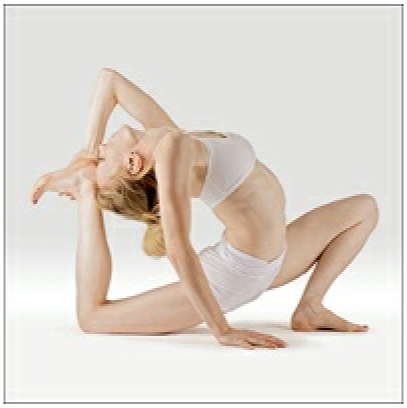

One-Legged King Pigeon Pose 2

Eka Pada Raja Kapotasana 2

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: both hands grabbing onto the foot
1. foot to the head
2. heel to the forehead

Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

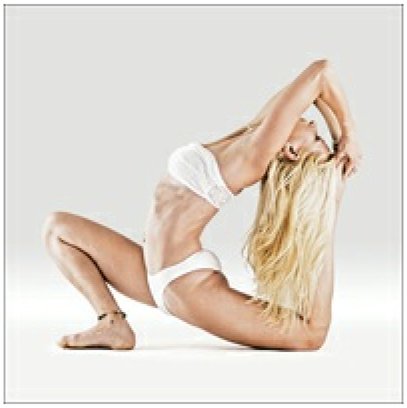

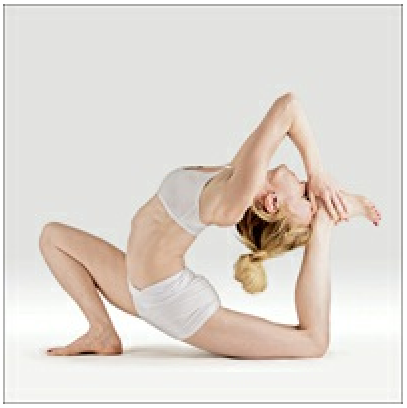

Half Bound One-Legged King Pigeon Pose 2

Ardha Baddha Eka Pada Raja Kapotasana 2

(UHR-duh BUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: standing, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

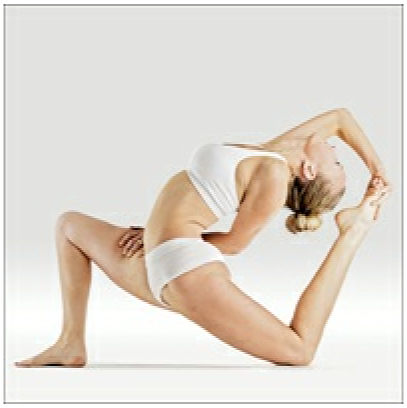

LUNGE: BACK KNEE ON FLOOR—ONE-LEGGED KING PIGEON 2—FOOT TO THE BACK OF THE HEAD

Necklace Pose Prep.

Graivasana Prep.

(gr-eye-VAHS-uh-nuh)

Also Known As: Chain Pose (Gaivasana)
Modification: back foot to the back of the head, hands to the floor
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

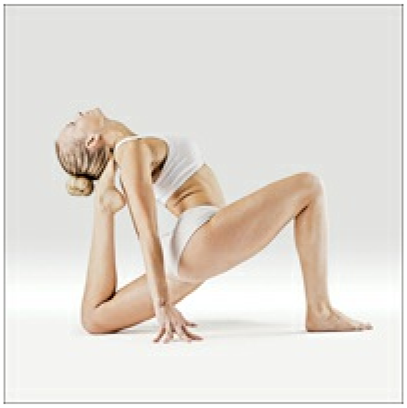

Necklace Pose Prep.

Graivasana Prep.

(gr-eye-VAHS-uh-nuh)

Also Known As: Chain Pose (Gaivasana)
Modification: back foot to the back of the head, hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

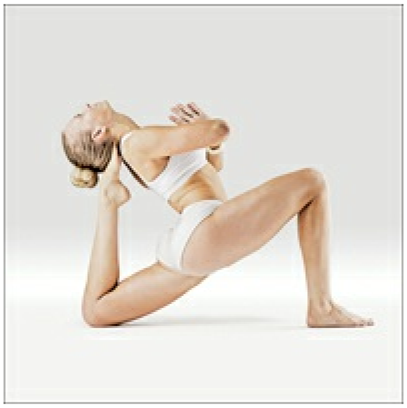

LUNGE: BACK KNEE ON FLOOR—ONE-LEGGED KING PIGEON 2—ARM CROSSED OVER IN FRONT

Bound Hand to Foot Supported One-Legged King Pigeon Pose 2

Baddha Hasta Pada Salamba Eka Pada Raja Kapotasana 2

(BUH-duh HUH-stuh PUH-duh SAH-luhm-buh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: arm crossed in front of the neck, other hand to the floor
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

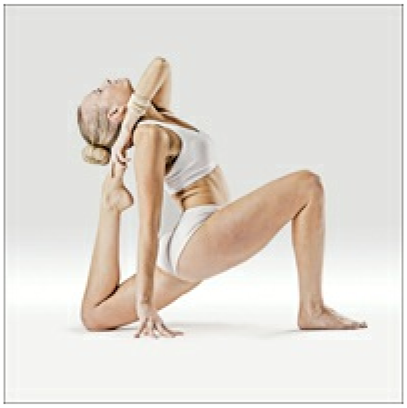

Bound Hand to Foot Unsupported One-Legged King Pigeon Pose 2

Baddha Hasta Pada Niralamba Eka Pada Raja Kapotasana 2

(BUH-duh HUH-stuh PUH-duh nir-AH-luhm-buh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: arm crossed in front of the neck, other arm straight out in front
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

LUNGE: FRONT LEG STRAIGHT—BACKBEND

Both Hands to Foot Son of Anjani (Lord Hanuman) Lunge Pose

Dwi Hasta Pada Anjaneyasana

(DWI-huh-stuh PUH-duh uhn-juh-ney-AHS-uh-nuh)

Modification: front leg straight
1. back knee on the floor
2. back knee off the floor

Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

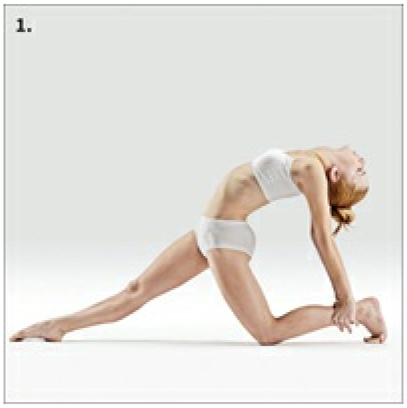
1.

2.

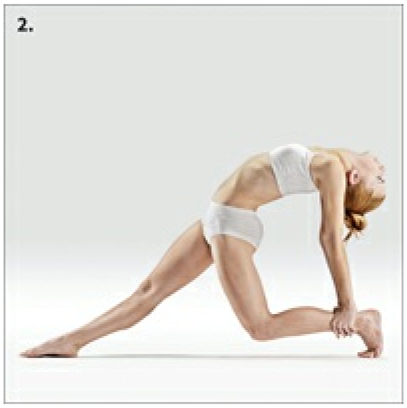

Gate Pose

Parighasana

(puh-ri-GAHS-uh-nuh)

Pose Type: standing, side bend

Drishti Point: Hastagrai or Hastagrahe (hands)

How to Perform the Pose:

1. Begin by standing on your knees with sitting bones lifted off the feet. (Make sure your knees are comfortable by rolling up the yoga mat a couple of times.) Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale and extend your right leg straight out to the side with the right foot on the floor. You can either point the toes to the right (Pose #1) or point your toes straight ahead, lifting the arch and pressing the outside edge of the right foot to the floor (Pose #2).
3. Inhale as you expand your chest and hold your arms straight out to the sides, parallel to the floor.
4. On your next exhale, side bend to the right and bring your right hand onto your right shin (Pose #2), while extending your left arm over your head to the right with the palm facing down. Make sure your chest is rotated to face forward and not collapsing toward the floor. Look toward your right hand. You can also experiment with dropping your right palm to the floor on the inside of your right leg (Pose #1).
5. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch. Inhale as you come up and bring your right knee back to the floor by your left knee. Repeat on the other side.

Modification: bending to the opposite side of the bent knee on the floor
1. palm to the floor, toes of the straight leg pointing to the side
2. hand on the shin, toes of the straight leg pointing straight ahead

parigha = iron bar used for locking a gate

1.

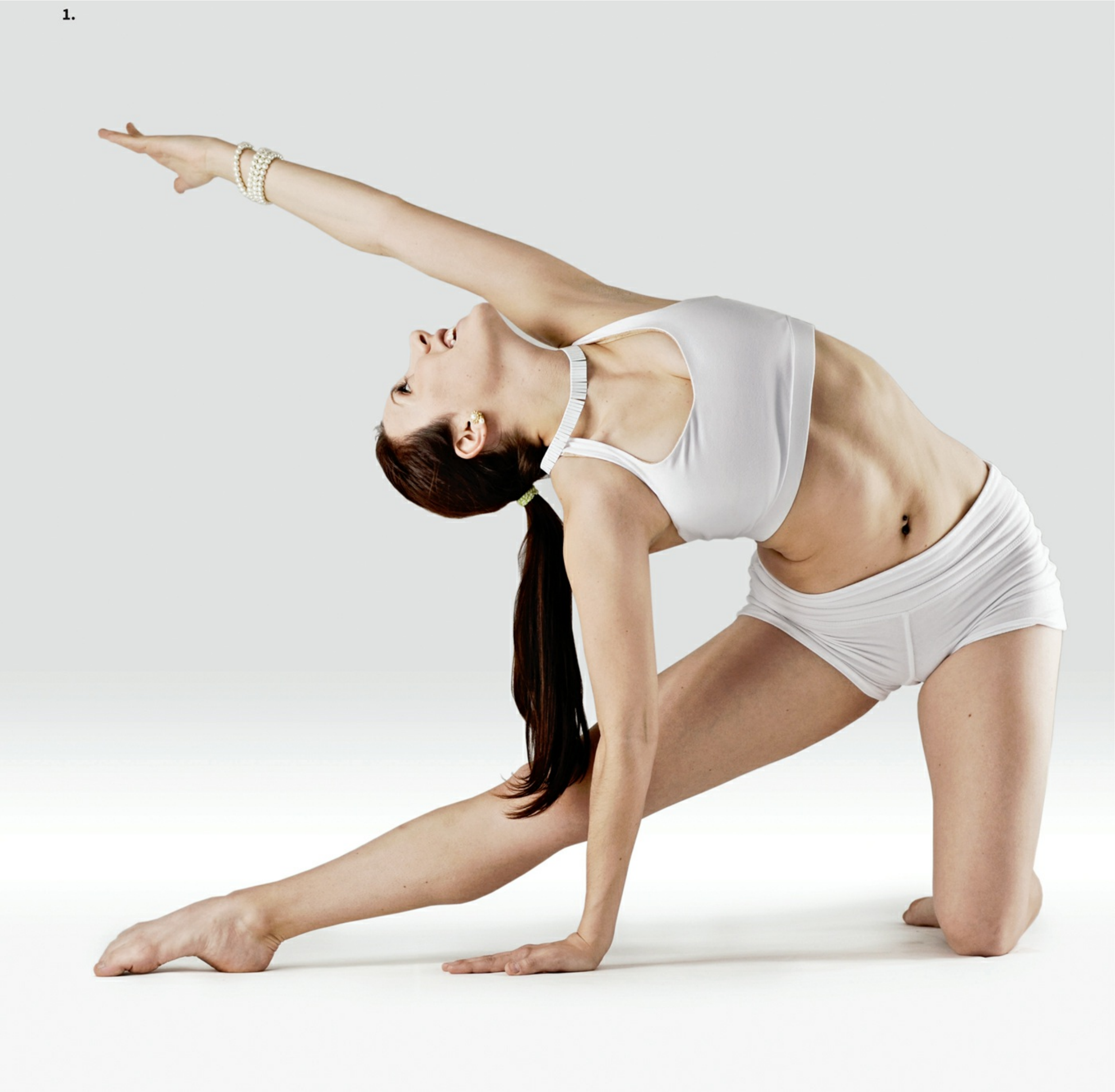

2.

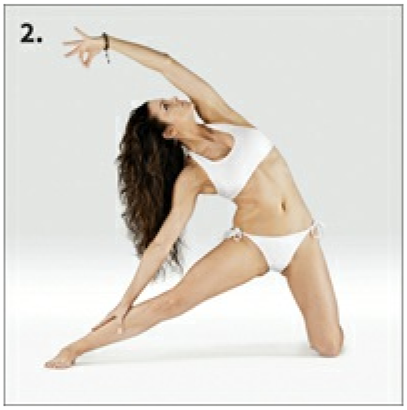

Gate Pose

Parighasana

(puh-ri-GAHS-uh-nuh)

Modification: bending to the side of the bent knee on the floor, hand to the floor on the outside of the bent leg, toes of the straight leg pointing straight ahead
Pose Type: standing, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व

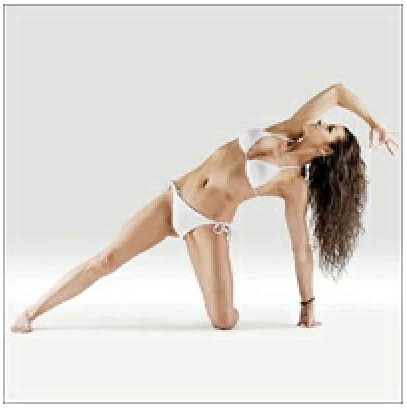

Half Bound Gate Pose

Ardha Baddha Parighasana

(UHR-duh BUH-duh puh-ri-GAHS-uh-nuh)

Modification: bending to the side of the bent knee on the floor, forearm to the floor on the side of the bent leg, toes of the straight leg pointing straight ahead
Pose Type: standing, side bend, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व य

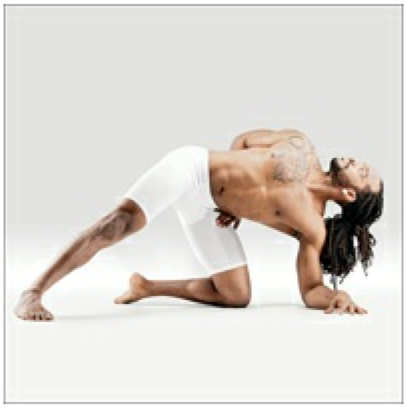

Gate Pose

Parighasana

(puh-ri-GAHS-uh-nuh)

Modification: bending to the opposite side of the bent knee on the floor, forearm to the floor, toes of the straight leg pointing to the side

Pose Type: standing, forward bend, side bend

Drishti Point: Hastagrai or Hastagre (hands)

ल व य ह

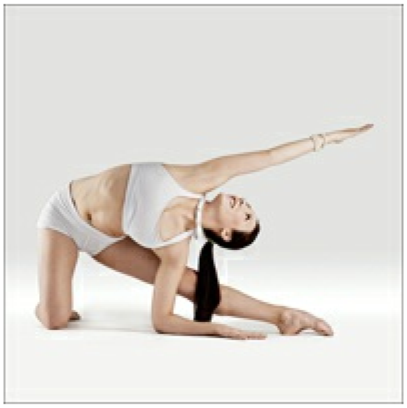

ONE LEG STRAIGHT OTHER KNEE ON THE FLOOR: SIDE BEND & BINDING

Gate Pose

Parighasana

(puh-ri-GAHS-uh-nuh)

Modification: bending to the opposite side of the bent knee on the floor, palm to the top of the foot of the straight leg, toes of the straight leg pointing to the side

Pose Type: standing, forward bend, side bend

Drishti Point: Hastagrai or Hastagre (hands)

ल व

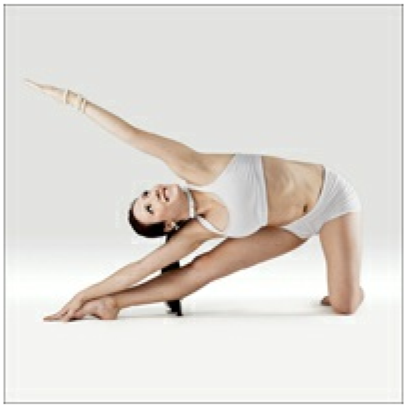

Gate Pose

Parighasana

(puh-ri-GAHS-uh-nuh)

Modification: bending to the opposite side of the bent knee on the floor, both hands to the top of the foot of the straight leg, toes of the straight leg pointing to the side
Pose Type: standing, forward bend, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

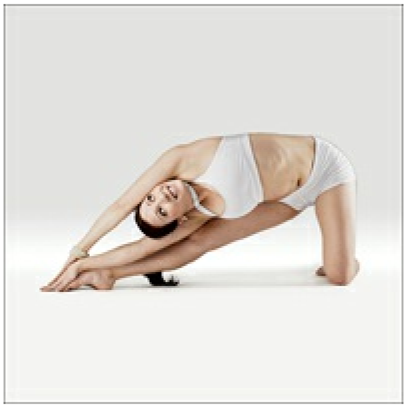

Revolved Half Bound Gate Pose

Parivritta Ardha Baddha Parighasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh puh-ri-GAHS-uh-nuh)

Modification: hand to the floor on the outside of the straight leg, rotating the chest to the sky, toes of the straight leg pointing to the side
Pose Type: standing, side bend, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

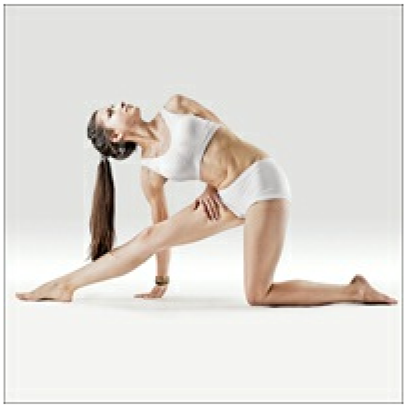

Revolved Bound Gate Pose

Parivritta Baddha Parighasana

(puh-ri-VRIT-tuh BUH-duh puh-ri-GAHS-uh-nuh)

Modification: toes of the straight leg flexed in, toes of the bent leg pointing away from the head
Pose Type: standing, forward bend, side bend, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व य

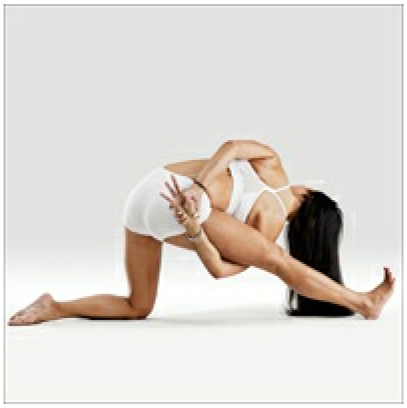

Revolved Bound Half Lotus Gate Pose

Parivritta Baddha Ardha Padma Parighasana

(puh-ri-VRIT-tuh BUH-duh UHR-duh PUHD-muh puh-ri-GAHS-uh-nuh)

Modification: toes of the straight leg pointing to the side
Pose Type: standing, side bend, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व य

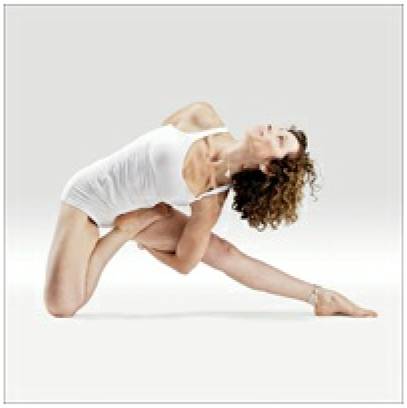

ONE LEG STRAIGHT OTHER KNEE ON THE FLOOR: TWISTS

Revolved Gate Pose

Parivritta Parighasana

(puh-ri-VRIT-tuh puh-ri-GAHS-uh-nuh)

Modification: forearm to the floor, hand grabbing onto the opposite foot, other arm straight up to the sky, toes of the straight leg pointed to the side, toes of the bent leg pointing away from the head
Pose Type: standing, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

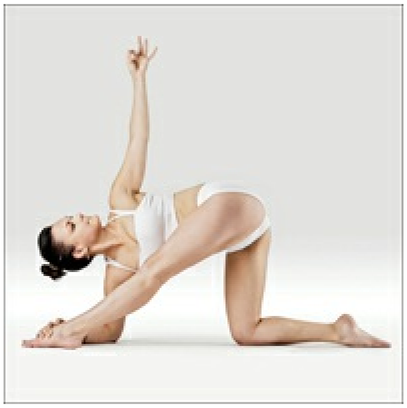

Revolved Half Bound Gate Pose

Parivritta Ardha Baddha Parighasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh puh-ri-GAHS-uh-nuh)

Modification: forearm to the floor, twisting toward the straight leg, toes of the straight leg pointing to the side
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

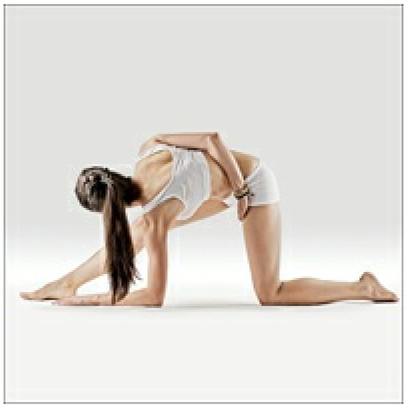

Half Pose Dedicated to Lord Hanuman

Ardha Hanumanasana

(UHR-duh huh-nu-mahn-AHS-uh-nuh)

Also Known As: Pose Dedicated to Lord Hanuman Prep. (Hanumanasana Prep.)
Modification: back toes pointed away from the head, fingertips to the floor on the outsides of the straight leg
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

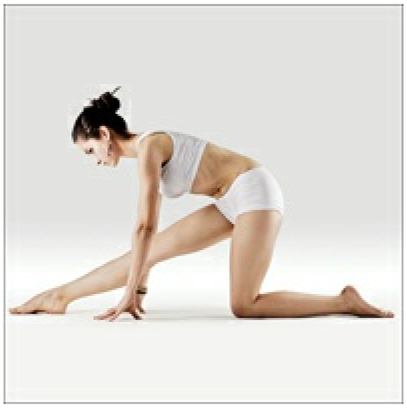

Half Pose Dedicated to Lord Hanuman

Ardha Hanumanasana

(UHR-duh huh-nu-mahn-AHS-uh-nuh)

Also Known As: Pose Dedicated to Lord Hanuman Prep. (Hanumanasana Prep.)
Modification: back toes curled in, toes of the straight leg flexed in, arms extended to the front, fingertips to the floor
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

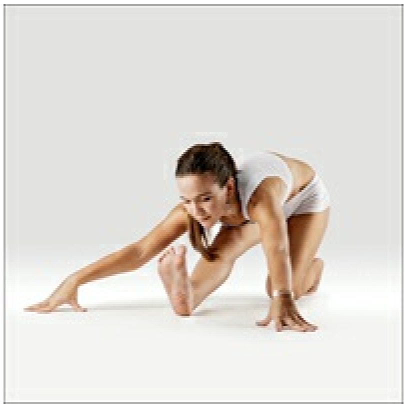

Half Bound Half Pose Dedicated to Lord Hanuman

Ardha Baddha Ardha Hanumanasana

(UHR-duh BUH-duh UHR-duh huh-nu-mahn-AHS-uh-nuh)

Also Known As: Pose Dedicated to Lord Hanuman Prep. (Hanumanasana Prep.)
Modification: toes of the bent leg pointed away from the head
Pose Type: standing, forward bend, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

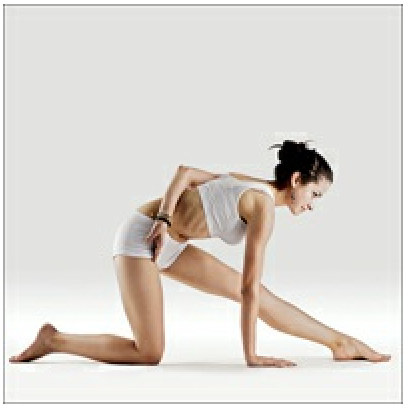

Half Pose Dedicated to Lord Hanuman

Ardha Hanumanasana

(UHR-duh huh-nu-mahn-AHS-uh-nuh)

Also Known As: Pose Dedicated to Lord Hanuman Prep. (Hanumanasana Prep.)
Modification: toes of the straight leg flexed in; hand grabbing onto the back foot on the same side, heel to the sitting bone
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

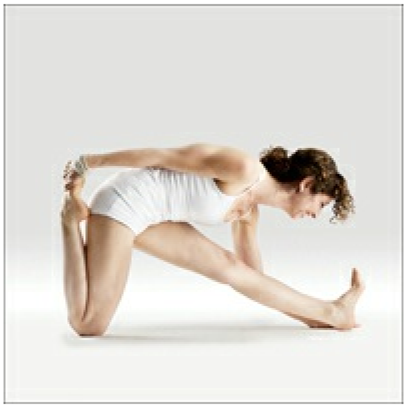

Half Lotus Half Pose Dedicated to Lord Hanuman

Ardha Padma Ardha Hanumanasana

(UHR-duh PUHD-muh UHR-duh huh-nu-mahn-AHS-uh-nuh)

Also Known As: Pose Dedicated to Lord Hanuman Prep. (Hanumanasana Prep.)
Modification: fingertips of both hands on the sides of the straight leg
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व

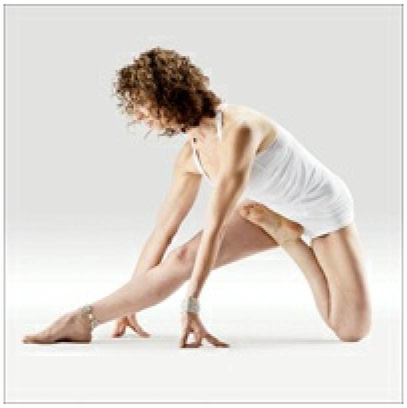

HORSE POSE: FORWARD BEND, BINDING & TWIST

Revolved Leg Position of Horse Pose with Hands in Prayer

Parivritta Pada Vatayanasana Namaskar

(puh-ri-VRIT-tuh PUH-duh vah-tah-yuh-NAHS-uh-nuh nuh-muhs-KAHR)

Pose Type: standing, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व र

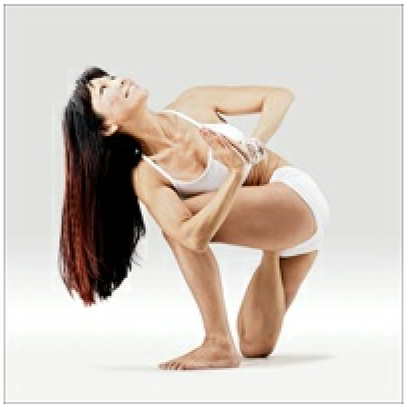

Half Lotus Extended Lizard Tail Lunge Pose

Ardha Padma Uttana Pristhasana

(UHR-duh PUHD-muh ut-TAHN-uh prish-TAHS-uh-nuh)

Modification: 1. palms to the floor, elbows bent 90 degrees
2. forearms to the floor
3. one forearm to the floor; other elbow to the floor, hand to the face
Pose Type: standing, forward bend
Drishti Point: 1 & 3. Nasagrai or Nasagre (nose)
2. Padhayoragrai or Padayoragre (toes/feet) and Angusthamadhye or Angushta Ma Dyai (thumbs)
ल व ह

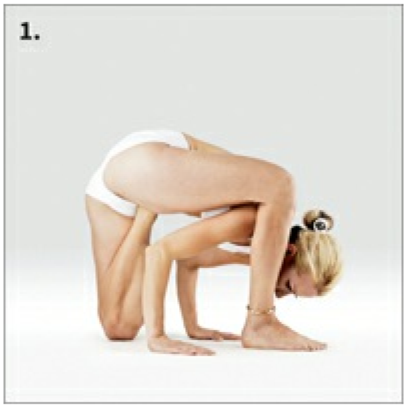

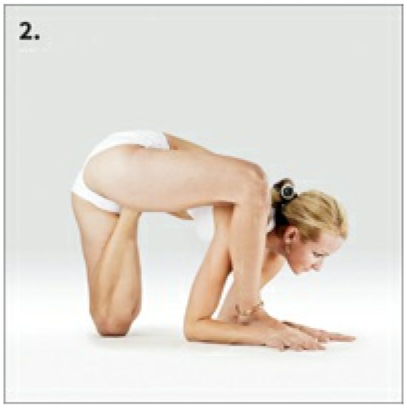

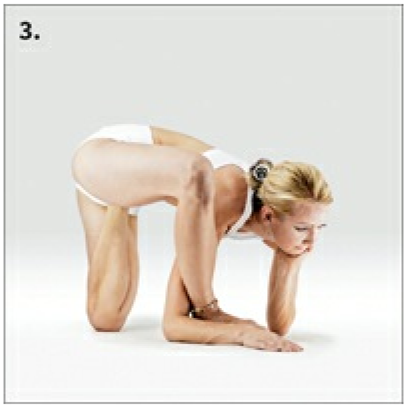

Bound Half Lotus Extended Lizard Tail Lunge Pose

Baddha Ardha Padma Uttana Pristhasana

(BUH-duh UHR-duh PUHD-muh ut-TAHN-uh prish-TAHS-uh-nuh)

Pose Type: standing, forward bend, binding

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

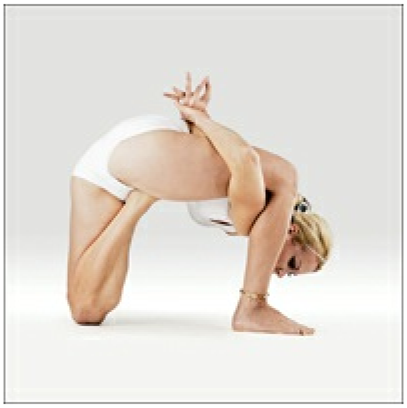

One Foot Behind the Head Bound Half Lotus Extended Lizard Tail Lunge Pose

Eka Pada Shirsha Baddha Ardha Padma Uttan Pristhasana

(EY-kuh PUH-duh SHEER-shuh BUH-duh UHR-duh PUHD-muh ut-TAHN-uh prish-TAHS-uh-nuh)

Pose Type: standing, forward bend, binding

Drishti Point: Urdhva or Antara Drishti (up to the sky)

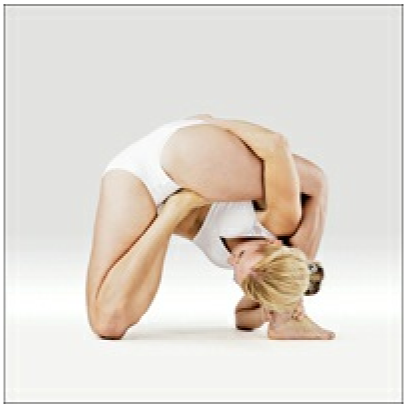

HORSE POSE: SPINE STRAIGHT

Half Bound Revolved Leg Position of Horse Pose

Ardha Baddha Parivritta Pada Vatayanasana

(UHR-duh BUH-duh puh-ri-VRIT-tuh PUH-duh vah-tah-yuh-NAHS-uh-nuh)

Modification: wrist to the knee of the front leg
Pose Type: standing, binding
Drishti Point: Hastagrai or Hastagre (hands)

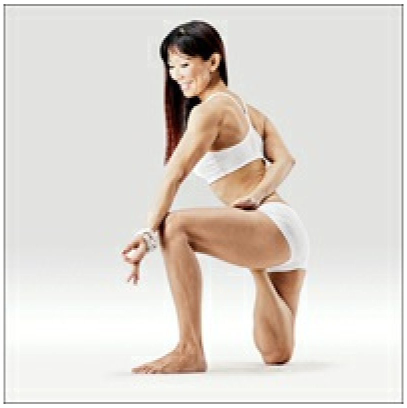

Horse Pose

Vatayanasana

(vah-tah-yuh-NAHS-uh-nuh)

Also Known As: Sea Horse Pose
Modification: 1. Front View
2. Side View
Pose Type: standing
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

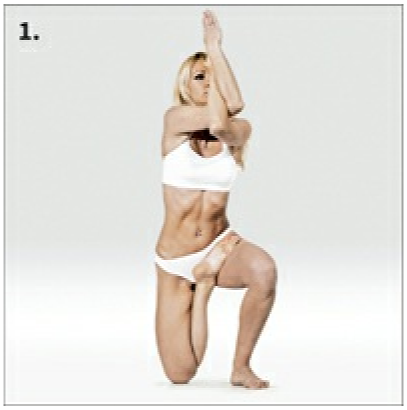
1.

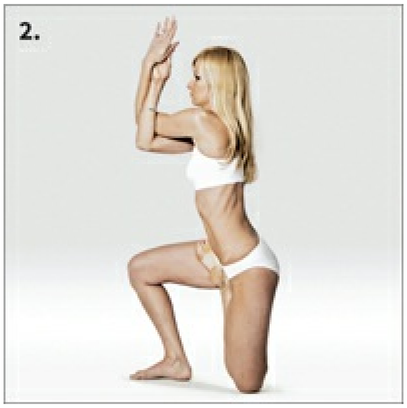
2.

ONE KNEE BENT TOWARD THE HIP: SPINE STRAIGHT

One-Legged Frog Pose in Camel Pose

Eka Pada Bhekasana in Ushtrasana

(EY-kuh PUH-duh bey-KAHS-uh-nuh in oosh-TRAHS-uh-nuh)

Modification: spine straight

Pose Type: standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

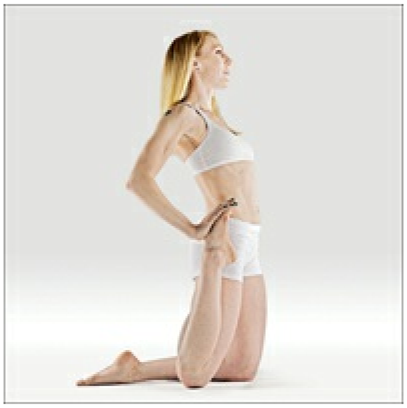

Mermaid Arm Position in Half Camel Pose

Hasta Naginyasana in Ardha Ushtrasana

(HUH-stuh nuh-gin-YAHS-uh-nuh in UHR-duh oosh-TRAHS-uhna)

Modification: spine straight
Pose Type: standing (on the knees), mild backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ॐ

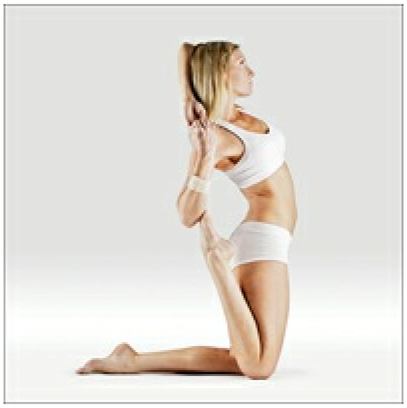

Revolved One-Legged Frog Pose in One-Legged King Pigeon 1 Version B

Parivritta Eka Pada Bhekasana in Eka Pada Raja Kapotasana 1 B

(puh-ri-VRIT-tuh EY-kuh PUH-duh bey-KAHS-uh-nuh in EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: fingertips to the floor in front
Pose Type: standing (on the knees), twist, binding
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)
ल व य

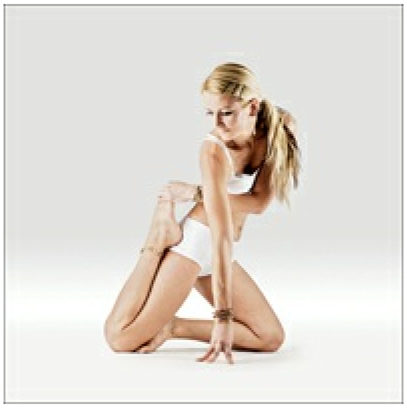

Half Bound Revolved One-Legged Frog Pose in One-Legged King

Pigeon 1 Version B

Ardha Baddha Parivritta Eka Pada Bhekasana in Eka Pada Raja Kapotasana 1 B

(UHR-duh BUH-duh puh-ri-VRIT-tuh EY-kuh PUH-duh bey-KAHS-uh-nuh in EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: standing (on the knees), twist, binding

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

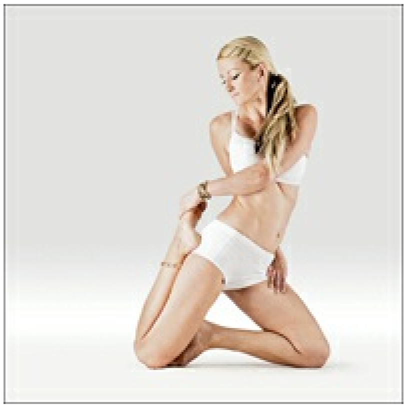

COW FACE POSE AND GARUDA ON THE KNEES

Leg Position of Cow Face Pose on the Knees

Janu Pada Gomukhasana

(JAH-nu PUH-duh go-moo-KAHS-uh-nuh)

Modification: hands on the hips

Pose Type: standing (on the knees)

Drishti Point: Nasagrai or Nasagre (nose)

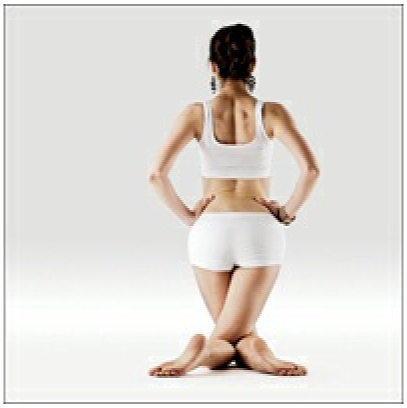

Leg Position of Cow Face Pose on the Knees

Janu Pada Gomukhasana

(JAH-nu PUH-duh go-moo-KAHS-uh-nuh)

Modification: 1. one hand on the hip, other arm straight over the head, side bend
2. one arm crossed in front of the body, other arm over the head, elbow bent, side bend

Pose Type: 1. standing (on the knees)
2. standing (on the knees), side bend

Drishti Point: Padayoragrai or Padayoragre (toes/feet), Hastagrai or Hastagre (hands)

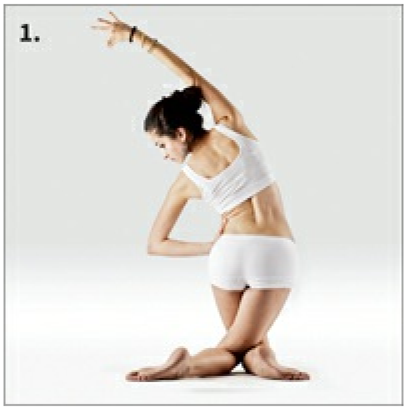

1.

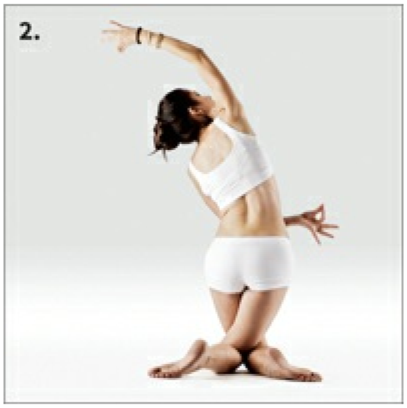

2.

KNEES ON THE FLOOR: LOTUS POSE

Half Pose Dedicated to Sage Goraksha

Ardha Gorakshasana

(UHR-duh go-rak-SHAHS-uh-nuh)

Modification: spine straight, one hand to the heart, other hand to the side of the thigh

Pose Type: standing (on the knees), balance

Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

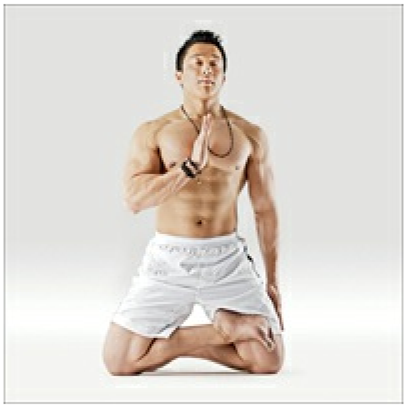

Pose Dedicated to Sage Goraksha

Gorakshasana

(go-ruhk-SHAHS-uh-nuh)

Also Known As: Yogic Seat Pose A (Yogapithasana A)
Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing (on the knees), balance
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

LUNGE: BACK KNEE OFF THE FLOOR

Fighting Warrior Pose 2

Yudhasana 2

(yu-DAHS-uh-nuh)

Modification: back knee off the floor
Pose Type: standing
Drishti Point: Hastagrai or Hastagre (hands)

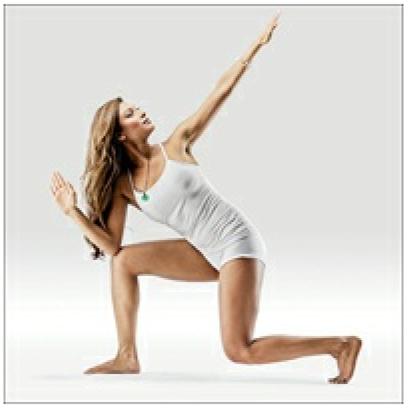

Son of Anjani (Lord Hanuman) Lunge Pose

Anjaneyasana

(uhn-juh-ney-AHS-uh-nuh)

Modification: arms open wide, forward bend, back knee bent and off the floor

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

लं वं रं हं

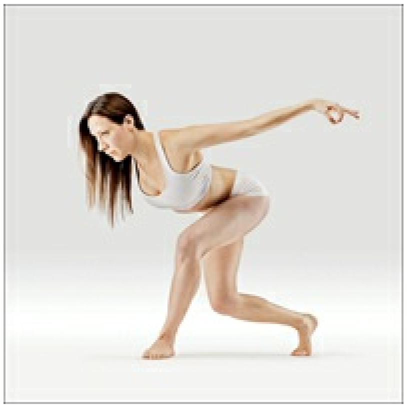

Son of Anjani (Lord Hanuman) Lunge Pose

Anjaneyasana

(uhn-juh-ney-AHS-uh-nuh)

Also Known As: Warrior 1 Modification (Virabhadrasana 1 Modification)

Modification: hands on the hips, back knee off the floor

Pose Type: standing

Drishti Point: Nasagrai or Nasagre (nose)

लं वं

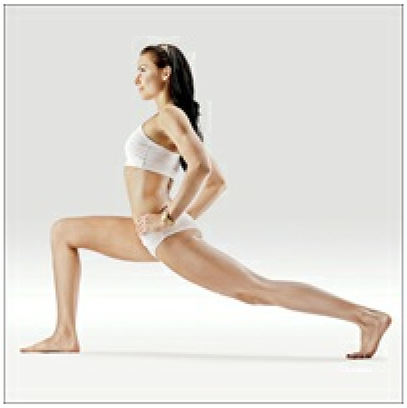

Equestrian Riding Horse Pose

Ashva Sanchalanasana

(UHSH-vuh suhn-chuh-luh-NAHS-un-nuh)

Pose Type: standing, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Padhayoragrai or Padayoragre (toes/feet)

How to Perform the Pose:

1. Begin in Intense Stretch Pose (*Uttanasana*), also known as Full Forward Bend Pose. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Inhale and bend your knees enough for your fingertips to touch the floor on the sides of the feet.
3. Exhale as you step your left foot to the back, keeping your toes curled in, your left knee off the floor and keep your left leg strong and straight. Make sure your right knee is on top of your right ankle to prevent wear and tear on your knee joint.
4. Rest your torso on your right thigh. Inhale as you lengthen your spine, stretching your body in two opposing directions: crown of the head moving to the front and the heel of the right foot pushing to the back. You can look at the toes of the right foot.
5. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
6. Exhale as you step your left foot to the front in line with your right foot, and repeat on the other side.

Modification: back knee off the floor, fingertips off the floor on either side of the front foot

ashva sanchala = horse, riding posture

Reverse Prayer Son of Anjani (Lord Hanuman) Lunge Pose

Viparita Namaskar Anjaneyasana

(vi-puh-REE-tuh nuh-muhs-KAHR uhn-juh-ney-AHS-uh-nuh)

Also Known As: Back of the Body Prayer Son of Anjani (Lord Hanuman) Lunge Pose (Paschima Namaskara Anjaneyasana)
Modification: back knee off the floor
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

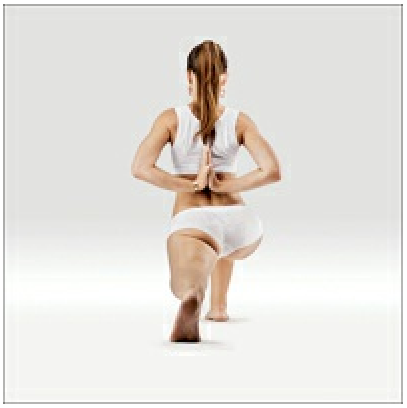

Son of Anjani (Lord Hanuman) Lunge Pose

Anjaneyasana

(uhn-juh-ney-AHS-uh-nuh)

Modification: palms together, thumbs to upper back, back knee off the floor
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

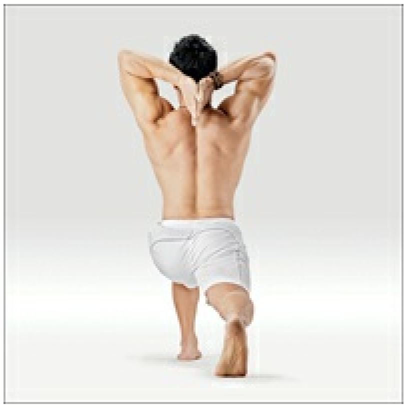

Tip Toe One Hand Son of Anjani (Lord Hanuman) Lunge Pose

Prapada Eka Hasta Anjaneyasana

(PRUH-puh-duh EY-kuh HUH-stuh uhn-juh-ney-AHS-uh-nuh)

Modification: deep backbend, one arm up over the head; fingertips of the other hand to the floor, back knee off the floor

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

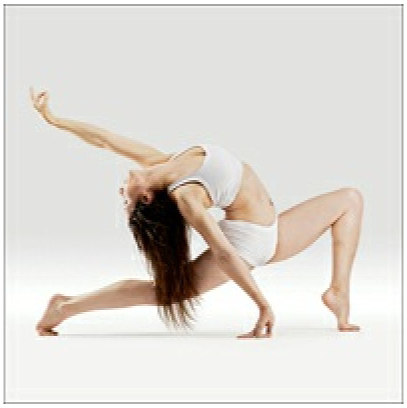

Son of Anjani (Lord Hanuman) Lunge Pose

Anjaneyasana

(uhn-juh-ney-AHS-uh-nuh)

Also Known As: Warrior 1 Modification (Virabhadrasana 1 Modification), Crescent Lunge Pose

Modification: hands up over the head and shoulder width apart, back knee off the floor

Pose Type: standing, mild backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

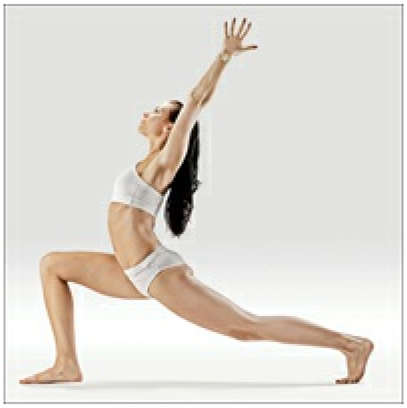

One Hand Rising Standing Pose of the Heavenly Spirits

Eka Hasta Utthita Stiti Valakhilyasana

(EY-kuh HUH-stuh UT-ti-tuh STI-ti vah-luh-khil-YAHS-uh-nuh)

Also Known As: Eka Hasta Utthita Nindra Valakhilyasana; Son of Anjani (Lord Hanuman) Lunge Pose (Anjaneyasana)

Modification: grabbing the back ankle overhead on the same side, other hand to the floor, back knee off the floor

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

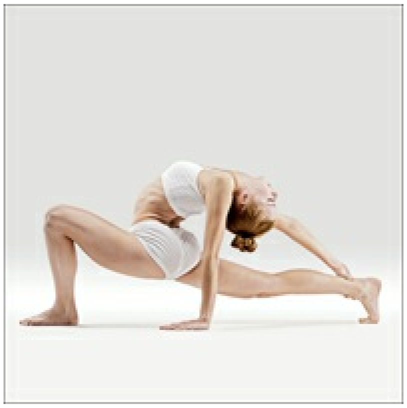

Rising Standing Pose of the Heavenly Spirits

Utthita Stiti Valakhilyasana

(UT-ti-tuh STI-ti vah-luh-khil-YAHS-uh-nuh)

Also Known As: Utthita Nindra Valakhilyasana; Son of Anjani (Lord Hanuman) Lunge Pose (Anjaneyasana)

Modification: grabbing the back ankle overhead with both hands, back knee off the floor

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

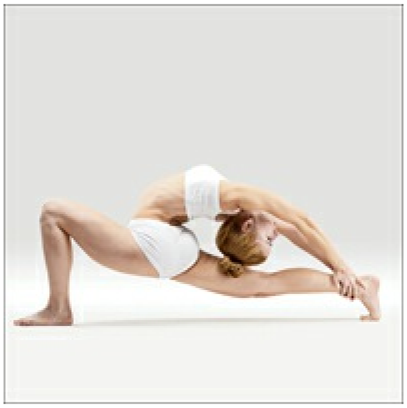

Extended Lizard Tail Lunge Pose

Uttana Pristhasana

(ut-TAH-nuh prish-TAHS-uh-nuh)

Modification: elbows to the floor on the inside of the front leg, back knee off the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

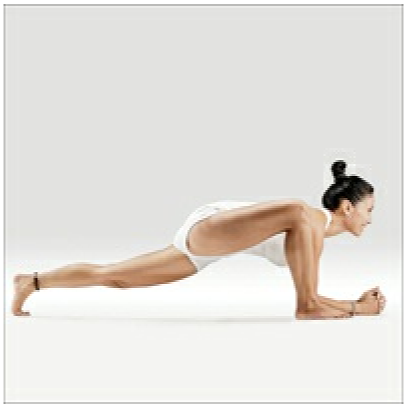

Pose Dedicated to Sage Punnakeesar Lunge

Punnakeesarasana

(pu-nuh-kee-suh-RAHS-uh-nuh)

Also Known As: Bound God Favor Seeking Sacrifice Ritual Pose (Baddha Yajnasana), Bound Christ’s Cross Pose
Modification: back knee off the floor
Pose Type: standing, forward bend, binding
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)
लं वं हं

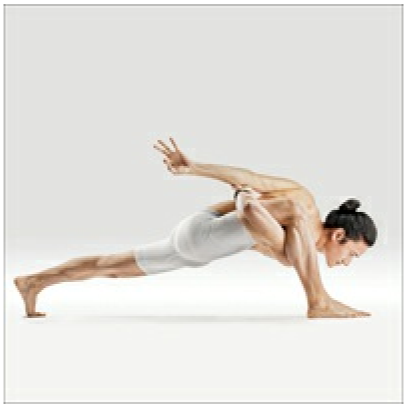

Extended Lizard Tail Lunge Pose

Uttana Pristhasana

(ut-TAH-nuh prish-TAHS-uh-nuh)

Modification: elbows bent at 90 degrees, palms to the floor, back knee off the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

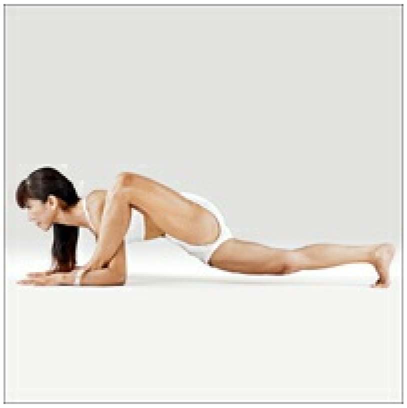

Extended Lizard Tail Lunge Pose

Uttana Pristhasana

(ut-TAH-nuh prish-TAHS-uh-nuh)

Modification: forearms to the floor on either side of the front foot, back knee off the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

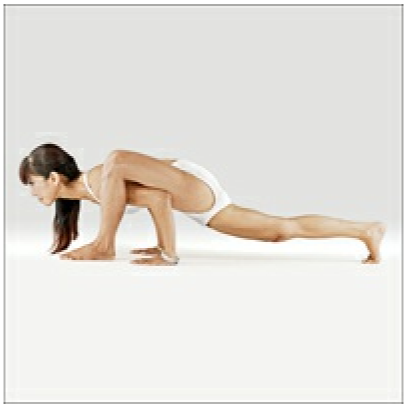

Tip Toe Extended Lizard Tail Lunge Pose

Prapada Uttana Pristhasana

(PRUH-puh-duh ut-TAH-nuh prish-TAHS-uh-nuh)

Modification: arms straight and reaching in the opposite directions, heel of the front leg up, top of the back foot to the floor, back knee off the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

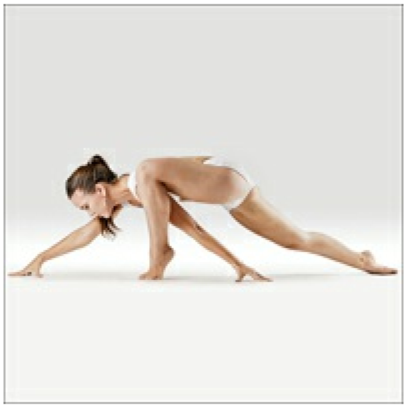

LUNGE: BACK KNEE OFF THE FLOOR, BACK LEG STRAIGHT—TWISTS

Sideways Son of Anjani (Lord Hanuman) Lunge Pose

Parshva Anjaneyasana

(PAHRSH-vuh uhn-juh-ney-AHS-uh-nuh)

Modification: top of the back foot on the floor, back knee off the floor
Pose Type: standing, side bend, backbend
Drishti Point: Hastagrai or Hastagre (hands)

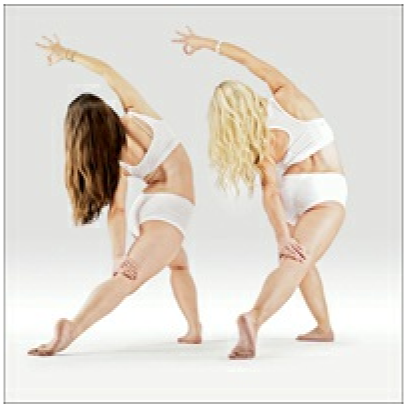

Revolved Son of Anjani (Lord Hanuman) Lunge Pose with Hands in Prayer

Parivritta Anjaneyasana Namaskar

(puh-ri-VRIT-tuh uhn-juh-ney-AHS-uh-nuh nuh-muhs-KAHR)

Modification: elbow over knee, back knee off the floor
Pose Type: standing, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

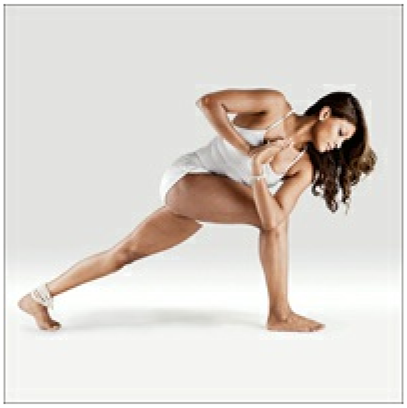

Revolved Son of Anjani (Lord Hanuman) Lunge Pose with Hands in Prayer

Parivritta Anjaneyasana Namaskar

(puh-ri-VRIT-tuh uhn-juh-ney-AHS-uh-nuh nuh-muhs-KAHR)

Modification: arm under the knee, back knee off the floor
Pose Type: standing, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

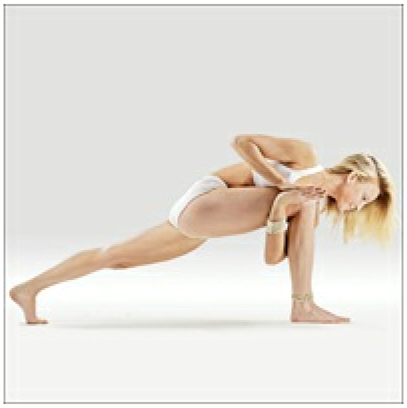

Bound Revolved Son of Anjani (Lord Hanuman) Lunge Pose

Baddha Parivritta Anjaneyasana

(BUH-duh puh-ri-VRIT-tuh uhn-juh-ney-AHS-uh-nuh)

Modification: back knee off the floor
Pose Type: standing, twist, binding
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)
ल व र

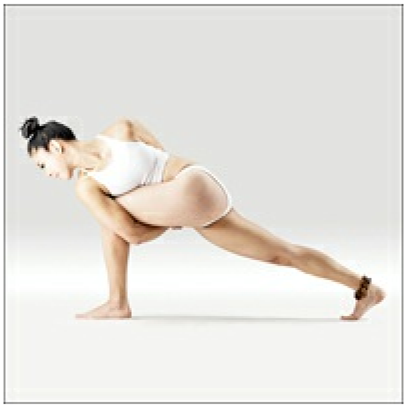

Revolved Son of Anjani (Lord Hanuman) Lunge Pose

Parivritta Anjaneyasana

(puh-ri-VRIT-tuh uhn-juh-ney-AHS-uh-nuh)

Modification: one hand to the floor, other arm up in the sky, twisting to the inside of the front knee, back knee off the floor
Pose Type: standing, twist
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व य

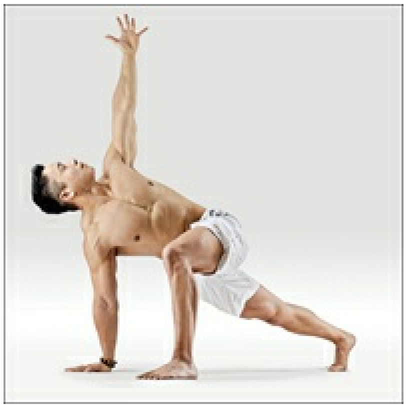

Tip Toe Sideways Son of Anjani (Lord Hanuman) Lunge Pose

Prapada Parshva Anjaneyasana

(PRUH-puh-duh PAHRSH-vuh uhn-juh-ney-AHS-uh-nuh)

Modification: one hand to the floor, other arm up in the sky, twisting to the outside of the front knee, back knee off the floor
Pose Type: standing, twist
Drishti Point: Hastagrai or Hastagre (hands)

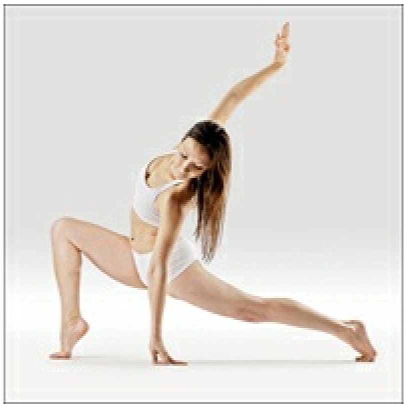

Revolved Paying Homage Lunge Pose

Parivritta Namasyasana

(puh-ri-VRIT-tuh nuh-muhs-YAHS-uh-muh)

Also Known As: Curtsey Lunge
Modification: front leg crossed over to the side, opposite elbow over the front knee, back knee off the floor
Pose Type: standing, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

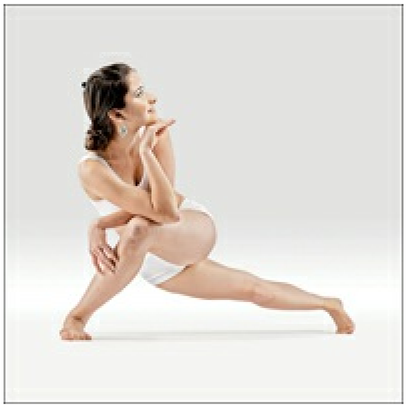

SIDE LUNGE: SITTING BONES LIFTED—ARMS BELOW HEAD

Extended Leg to the Side Squat Pose

Utthita Parshva Pada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Also Known As: Plyo Side Lunge
Modification: high stance, fingers interlocked in front of the chest
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

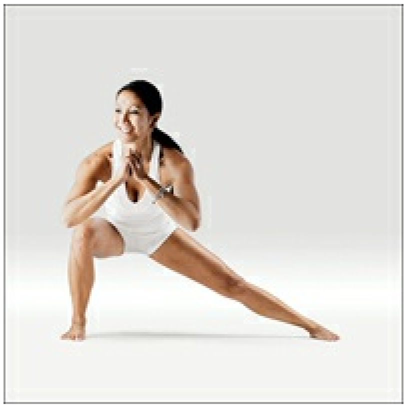

Tip Toe Extended Leg to the Side Squat Pose

Prapada Utthita Parshva Pada Upaveshasana

(PRUH-puh-duh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: high stance; arms straight out to the sides, toes of the straight leg flexed in
Pose Type: standing
Drishti Point: Hastagrai or Hastagre (hands)

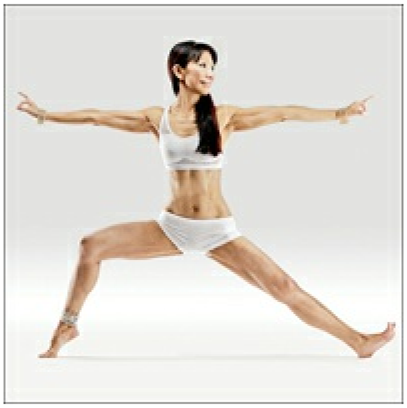

Extended Leg to the Side Squat Pose

Utthita Parshva Pada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: arms crossed in front of the chest, forward bend
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

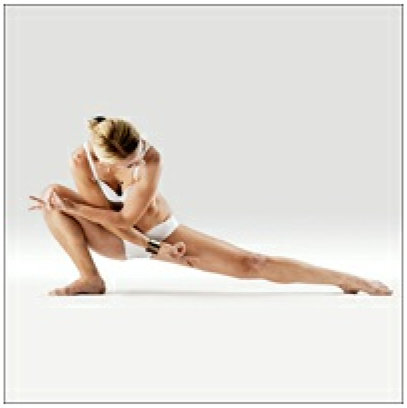

Extended Leg to the Side Squat Pose

Utthita Parshva Pada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: leaning toward the bent leg, one arm straight to the side, fingertips flexed in; other elbow to the side, fingertips to the back of the head
Pose Type: standing, side bend
Drishti Point: Hastagrai or Hastagre (hands)

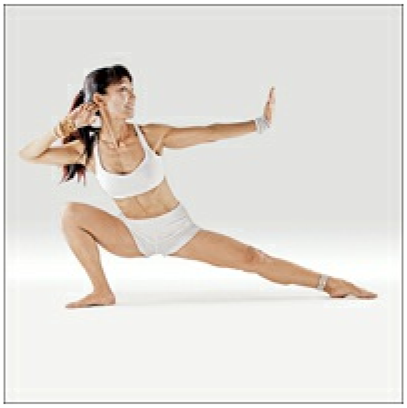

SIDE LUNGE: SITTING BONES LIFTED—ARMS ABOVE THE HEAD

Extended Leg to the Side Squat Pose

Utthita Parshva Pada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: side bend toward the straight leg, one arm parallel to the floor, other arm extended to the sky

Pose Type: standing

Drishti Point: Hastagrai or Hastagre (hands)

ल व य

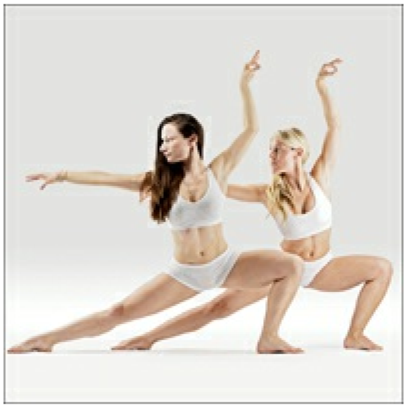

SIDE LUNGE: SITTING BONES LIFTED—HANDS TO THE FLOOR

Tip Toe Extended Leg to the Side Squat Pose

Prapada Utthita Parshva Pada Upaveshasana

(PRUH-puh-duh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: one palm to the floor

Pose Type: standing, forward bend

Drishti Point: Hastagrai or Hastagre (hands)

ल व

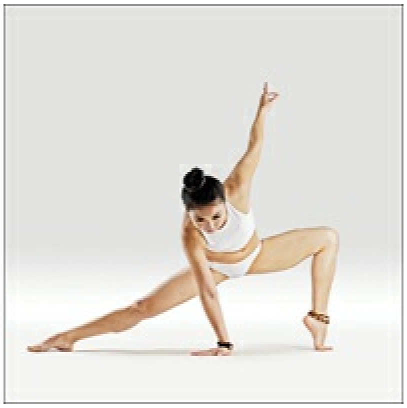

Extended Leg to the Side Squat Pose

Utthita Parshva Pada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: both hands to the floor, side bending toward the straight leg

Pose Type: standing, forward bend, side bend

Drishti Point: Hastagrai or Hastagre (hands)

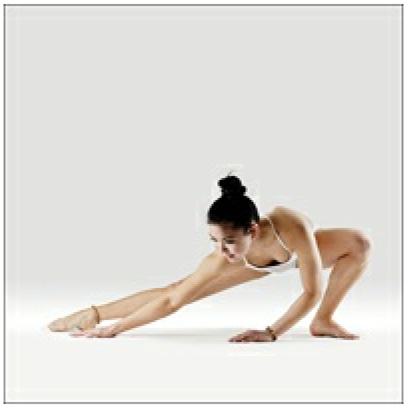

Intense Ankle Stretch Extended Leg to the Side Squat Pose

Uttana Kulpa Utthita Parshva Pada Upaveshasana

(ut-TAH-nuh kul-puh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: both palms to the floor, arms crossed

Pose Type: standing, forward bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

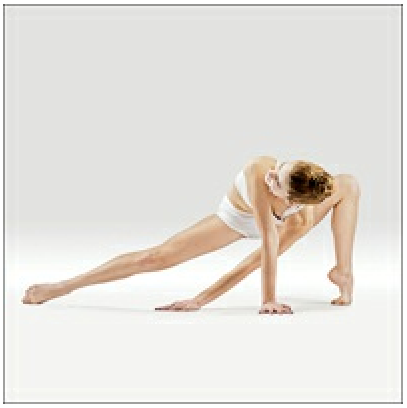

SIDE LUNGE: SITTING BONES ON THE FLOOR

Lotus Hand Seal Extended Leg to the Side Squat Pose

Padma Mudra Utthita Parshva Pada Upaveshasana

(PUHD-muh MU-druh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: forward bend, forehead to the floor, arms extended in front
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व ह

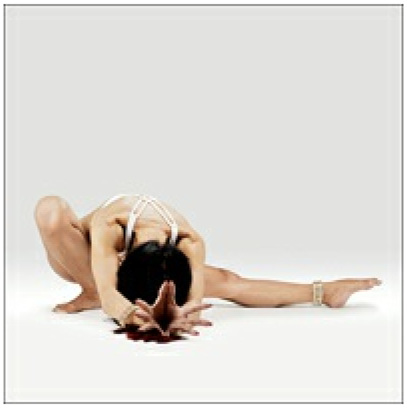

Uneven Arms Extended Leg to the Side Squat Pose

Vishama Hasta Utthita Parshva Pada Upaveshasana

(VISH-uh-muh HUH-stuh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: forward bend, one forearm to the floor, other hand to the face
Pose Type: standing, seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व

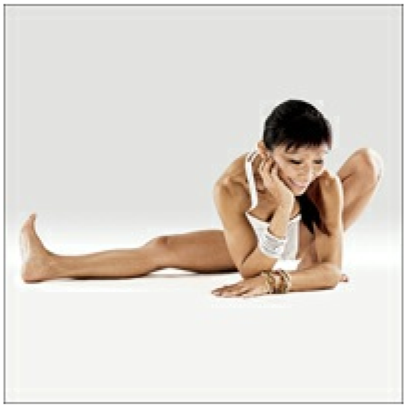

Revolved Extended Leg to the Side Squat Pose 1

Parivritta Utthita Parshva Pada Upaveshasana 1

(puh-ri-VRIT-tuh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: both hands to the foot of the straight leg

Pose Type: standing, seated, forward bend, side bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

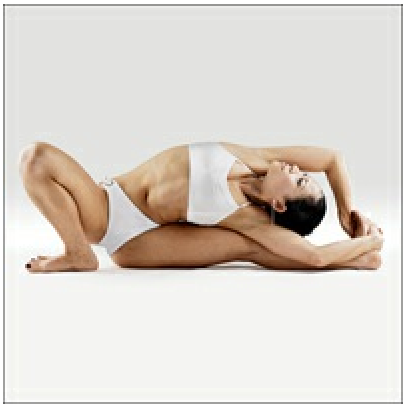

Revolved Extended Leg to the Side Squat Pose 2

Parivritta Utthita Parshva Pada Upaveshasana 2

(puh-ri-VRIT-tuh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)
Modification: grabbing onto both feet
Pose Type: standing, seated, forward bend, side bend, twist
Drishti Point: Hastagrai or Hastagre (hands), Padayoragrai or Padayoragre (toes/feet)

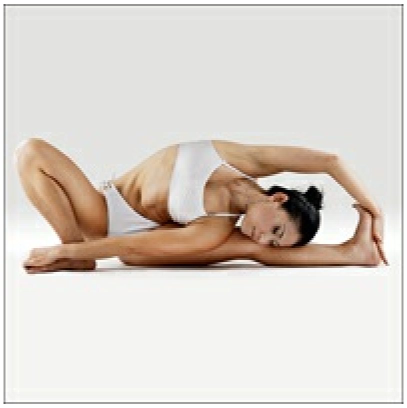

SIDE LUNGE: SITTING BONES LIFTED—TWISTS AND BINDING

Extended Leg to the Side Squat Pose

Utthita Parshva Pada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)
Modification: fingertips to the floor, toes of the straight leg flexed in
Pose Type: standing, forward bend

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

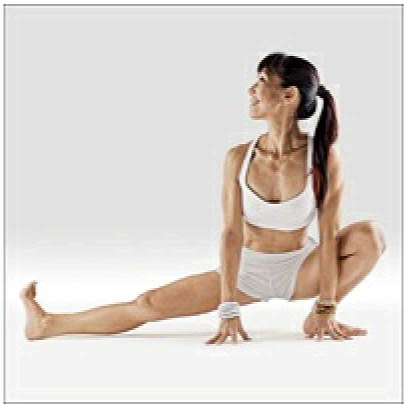

Knee to Shoulder Extended Leg to the Side Squat Pose

Janu Bhuja Utthita Parshva Pada Upaveshasana

(JAH-nu BHU-juh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: toes of the straight leg pointed
Pose Type: standing, forward bend, side bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

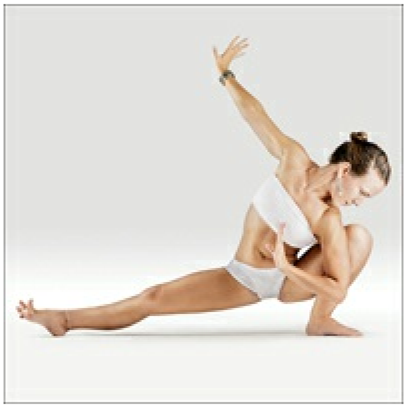

Bound Extended Leg to the Side Squat Pose

Baddha Utthita Parshva Pada Upaveshasana

(BUH-duh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: binding around the shin of the bent leg, toes of the straight leg flexed in
Pose Type: standing, side bend, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

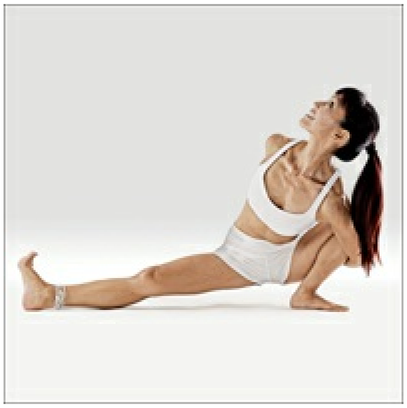

Revolved Half Bound Extended Leg to the Side Squat Pose

Parivritta Ardha Baddha Utthita Parshva Pada Upaveshasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh UT-ti-tuh PAHRSH-vuh PUH-duh u-puh-vey-SHAHS-uh-nuh)

Modification: toes of the straight leg pointed

Pose Type: standing, side bend, binding

Drishti Point: Urdhva or Antara Drishti (up to the sky)

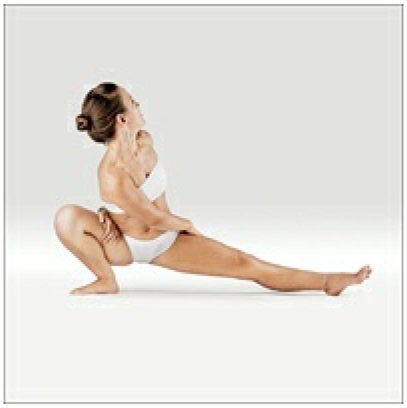

SIDE LUNGE: SITTING BONE TO THE HEEL—ARMS HIGHER THAN HEAD LEVEL

Extended Leg to the Side Tip Toe Squat Pose

Utthita Parshva Pada Prapada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: toes of the straight leg to the floor

Pose Type: standing, forward bend

Drishti Point: Hastagrai or Hastagre (hands)

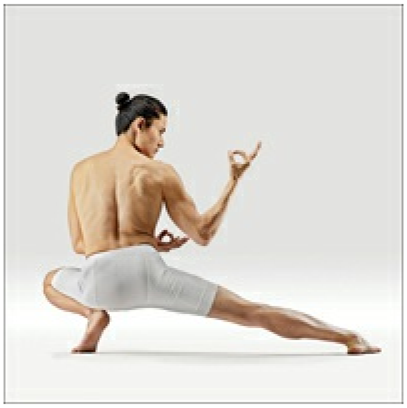

Extended Leg to the Side Tip Toe Squat Pose

Utthita Parshva Pada Prapada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: elbow to the knee of the bent leg, toes of the straight leg to the floor
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands)
लं वं

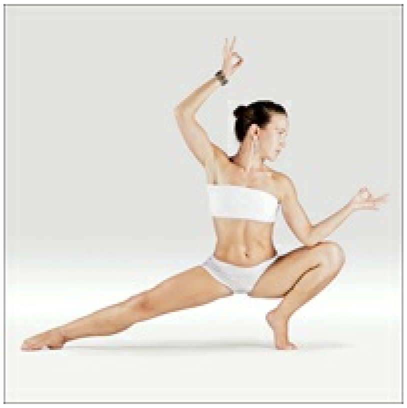

Extended Leg to the Side Tip Toe Squat Pose

Utthita Parshva Pada Prapada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: side bend toward the straight leg, one arm parallel to the floor, other arm extended to the sky, toes of the straight leg pointed
Pose Type: standing, side bend
Drishti Point: Hastagrai or Hastagre (hands)
लं वं यं हं

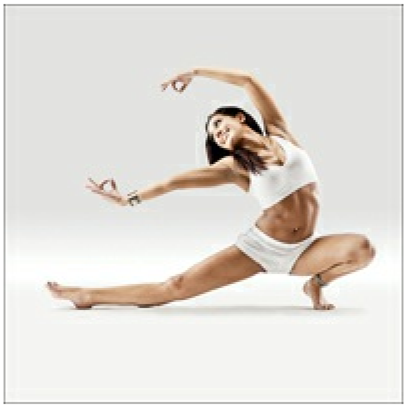

Extended Leg to the Side Tip Toe Squat Pose

Utthita Parshva Pada Prapada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: side bend toward the straight leg, arms extended to the side and straight, toes of the straight leg to the floor

Pose Type: standing, side bend

Drishti Point: Hastagrai or Hastagre (hands), Padayoragrai or Padayoragre (toes/feet)

ल व य

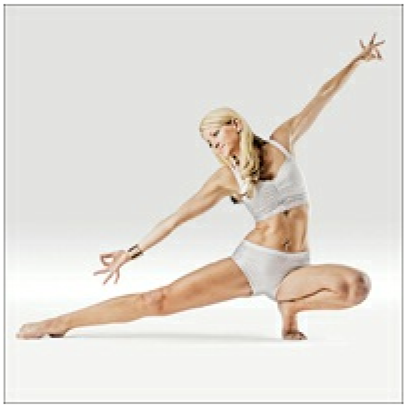

Both Arms Extended Leg to the Side Tip Toe Squat Pose

Dwi Hasta Utthita Parshva Pada Prapada Upaveshasana

(DWI-huh-stuh UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: side bend toward the bent leg, toes of the straight leg to the floor

Pose Type: standing, side bend

Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व य

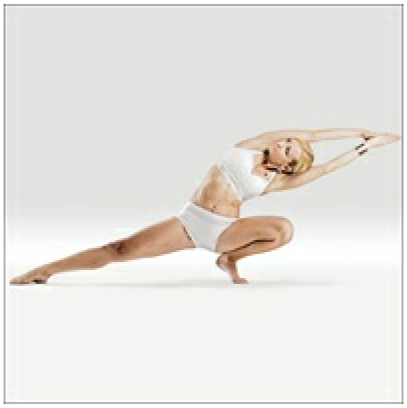

SIDE LUNGE: SITTING BONE TO THE HEEL—HAND TO THE FLOOR

Extended Leg to the Side Tip Toe Squat Pose

Utthita Parshva Pada Prapada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: one hand on the knee, other hand to the floor, toes of the straight leg pointed
Pose Type: standing, forward bend
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

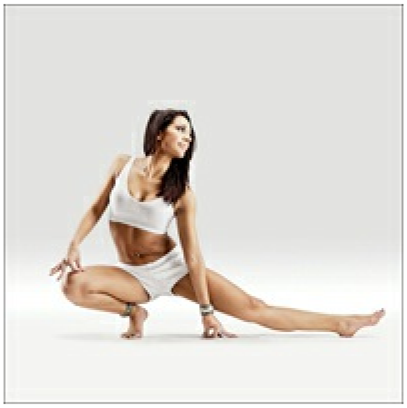

Extended Leg to the Side Tip Toe Squat Pose

Utthita Parshva Pada Prapada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: one palm to the floor, other arm straight and parallel to the straight leg, toes of the straight leg to the floor
Pose Type: standing
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Hastagrai or Hastagre (hands)

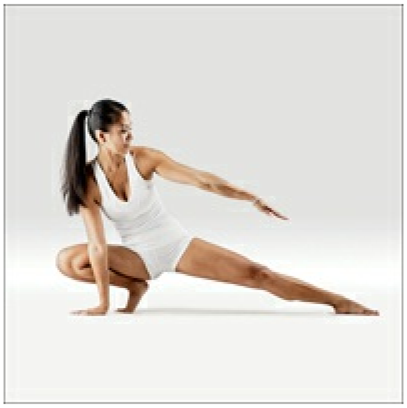

SIDE LUNGE: SITTING BONE TO THE HEEL—1 ARM OVERHEAD, OTHER HAND TO THE FLOOR

Extended Leg to the Side Tip Toe Squat Pose

Utthita Parshva Pada Prapada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: side bending toward the bent knee, one hand to the floor, other arm extended up over the head, toes of the straight leg to the floor
Pose Type: standing, side bend
Drishti Point: Hastagrai or Hastagre (hands)
ल व य ह

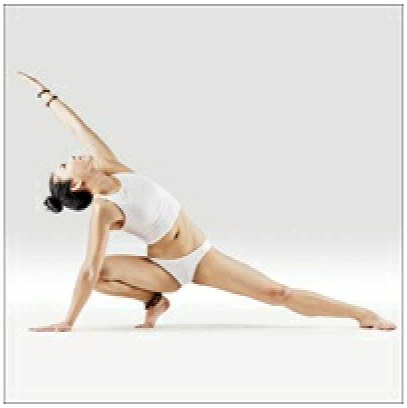

Extended Leg to the Side Tip Toe Squat Pose

Utthita Parshva Pada Prapada Upaveshasana

(UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: side bending toward the straight leg, one hand to the floor, other arm extended up over the head, toes of the straight leg to the floor
Pose Type: standing, side bend
Drishti Point: Hastagrai or Hastagre (hands)

ल व य ह

SIDE LUNGE: SITTING BONE TO THE HEEL—TWISTS & BINDING

Revolved Extended Leg to the Side Tip Toe Squat Pose with Hands in Prayer

Parivritta Utthita Parshva Pada Prapada Upaveshasana Namaskar

(puh-ri-VRIT-tuh UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh nuh-muhs-KAHR)

Modification: toes of the straight leg to the floor

Pose Type: standing, forward bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व

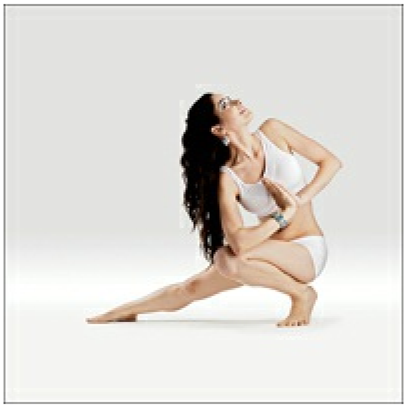

Revolved Bound Extended Leg to the Side Tip Toe Squat Pose

Parivritta Baddha Utthita Parshva Pada Prapada Upaveshasana

(puh-ri-VRIT-tuh BUH-duh UT-ti-tuh PAHRSH-vuh PUH-duh PRUH-puh-duh u-puh-vey-SHAHS-uh-nuh)

Modification: toes of the straight leg to the floor

1. knee on the floor

2. knee off the floor

Pose Type: standing, twist, binding

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

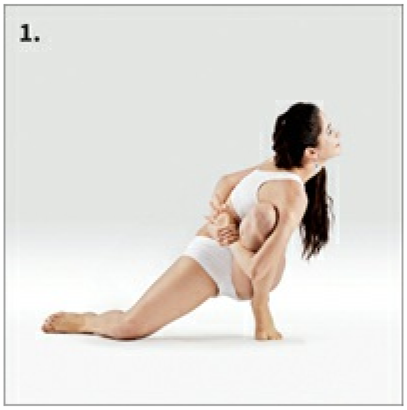

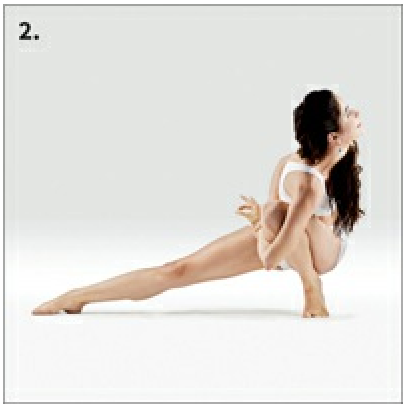

CHAIR SQUAT: HIGH STANCE

Fierce Pose

Utkatasana

(OOT-kuh-TAHS-uh-nuh)

Also Known As: Chair Pose

Modification: palms pressed together, high squat

Pose Type: standing, mild backbend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

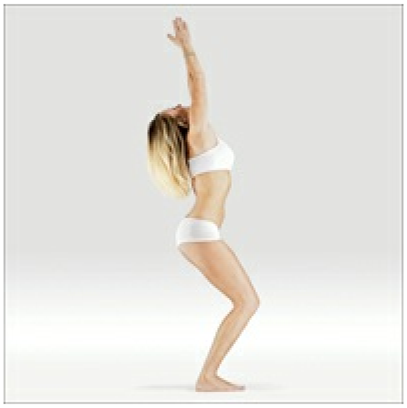

Fierce Pose

Utkatasana

(OOT-kuh-TAHS-uh-nuh)

Also Known As: Chair Pose
Modification: hands shoulder width apart, low squat
Pose Type: standing
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
🔴 🔵

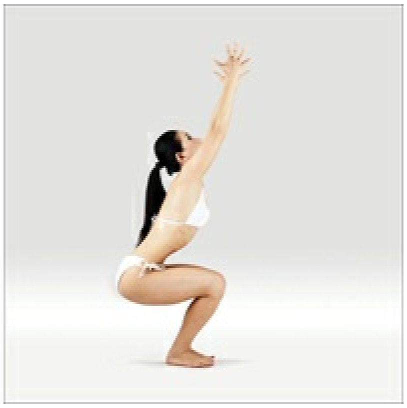

CHAIR SQUAT: HIGH STANCE—HANDS TO THE FLOOR

Wagtail Bird Pose Prep.

Khanjanasana Prep.

(kuhn-juh-NAHS-uh-nuh)

Modification: hands to the floor, palms facing up, feet together
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye/Ajna Chakra (third eye, between the eyebrows)

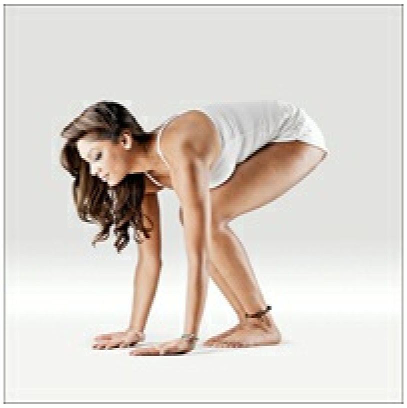

Intense Wrist Stretch Fierce Pose

Uttana Manibandha Utkatasana

(ut-TAH-nuh muh-ni-BUHN-duh OOT-kuh-TAHS-uh-nuh)

Also Known As: Intense Wrist Stretch Chair Pose
Modification: palms to the floor, fingertips facing the toes
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

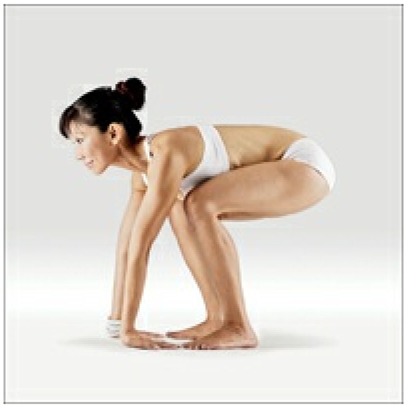

Wagtail Bird Pose

Khanjanasana

(kuhn-juh-NAHS-uh-nuh)

Also Known As: Shoulder Pressure Pose Prep. (Bhujapidasana Prep.)
Modification: shoulders to the back of the knees, fingertips to the heel
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye/Ajna Chakra (third eye, between the eyebrows)

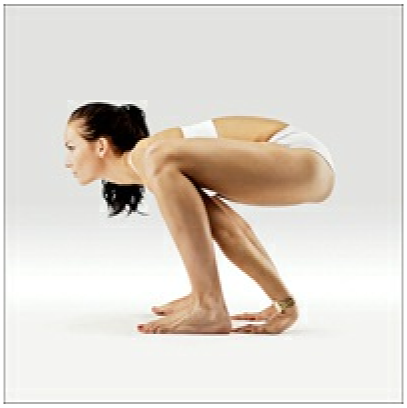

Intense Wrist Stretch Half Bound

Fierce Pose

Uttana Manibandha Ardha Baddha Utkatasana

(ut-TAH-nuh muh-ni-BUHN-duh UHR-duh OOT-kuh-TAHS-uh-nuh)

Also Known As: Intense Wrist Stretch Half Bound Chair Pose

Pose Type: standing, forward bend, binding

Drishti Point: Hastagrai or Hastagrahe (hands)

How to Perform the Pose:

1. Begin by standing in Mountain Pose (*Tadasana*). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale, bend your knees until your thighs are parallel to the floor, and drop both your hands to the floor. Rest your torso on top of your thighs.
3. Inhale as you externally rotate your right arm to point the fingertips of your right hand toward your feet.
4. On your next inhale, bring your left arm behind your back and your left hand to the inside of your right thigh. Try to seal the gap between your ribcage and your elbow.
5. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
6. Exhale as you release your left arm and bring your left hand to the floor. Repeat on the right side.

Modification: hand to the floor, fingertips pointing toward the toes

ut = intense

tan = to stretch, to extend

manibandha = wrist

ardha = half

baddha = bound

utkata = fierce

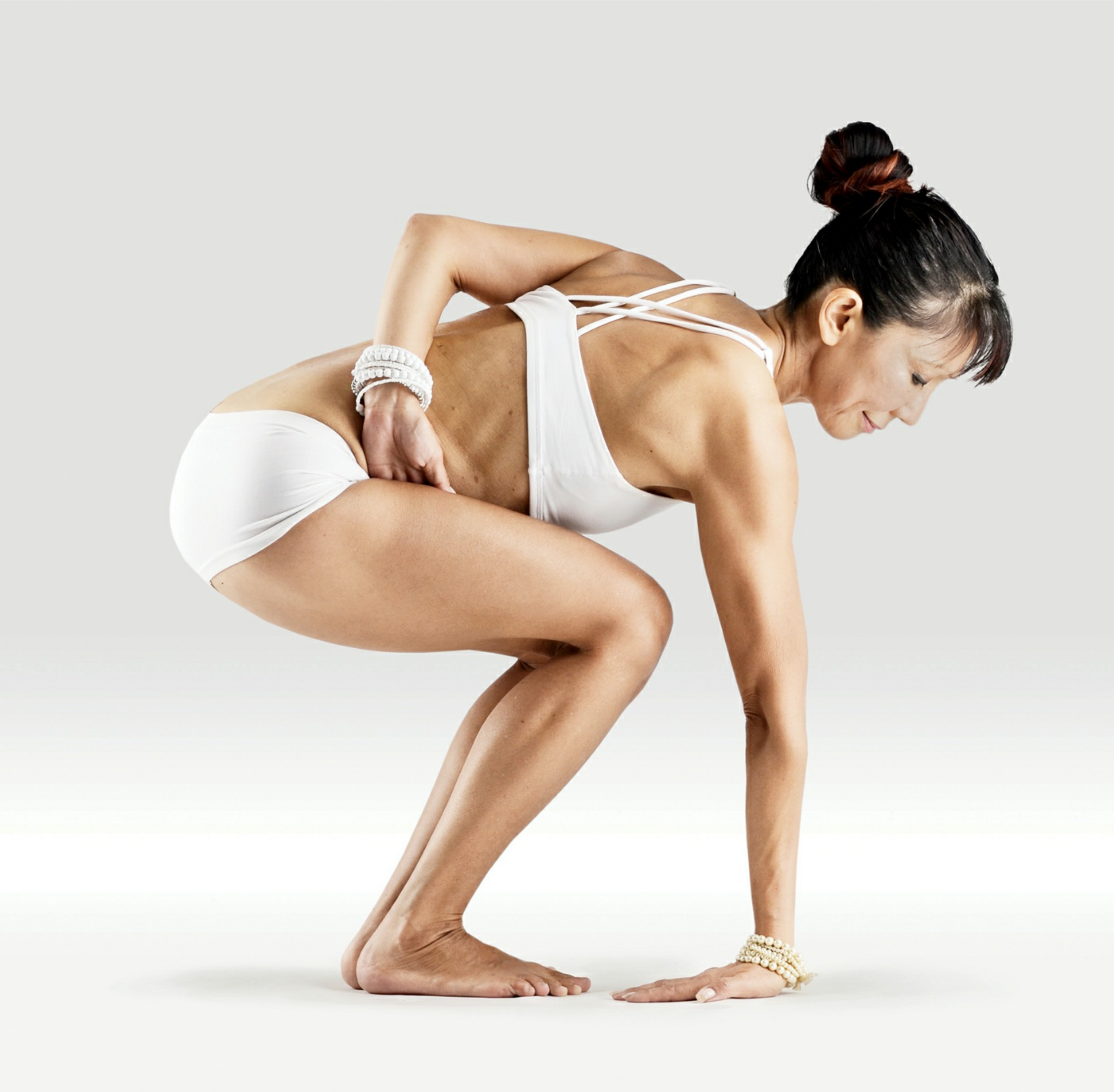

Elephant Old Form

Gaja Vadivu

(GUH-juh vah-dee-vuh)

Pose Type: standing, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

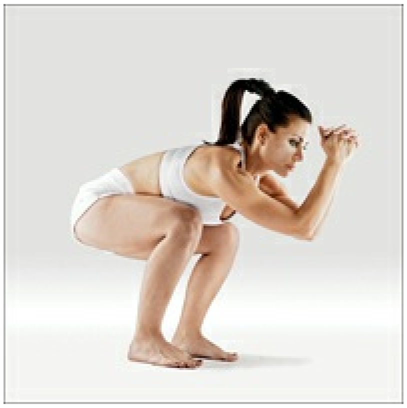

Hand Position of the Pose Dedicated to Garuda in Half Intense Stretch Pose

Hasta Garudasana in Ardha Uttanasana

(HUH-stuh guh-ru-DAHS-uh-nuh in UHR-duh ut-tahn-AHS-uh-nuh)

Also Known As: Hand Position of the Pose Dedicated to Garuda in Half Forward Bend
Modification: knees bent, feet hip width apart
Pose Type: standing, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

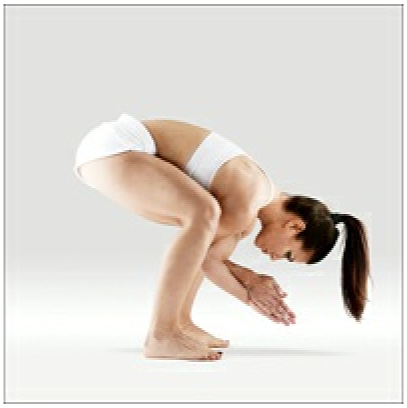

Hands Bound Ear Pressure Fierce Pose

Baddha Hasta Karnapida Utkatasana

(BUH-duh HUH-stuh kahr-nuh-PEE-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Hands Bound Ear Pressure Chair Pose
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

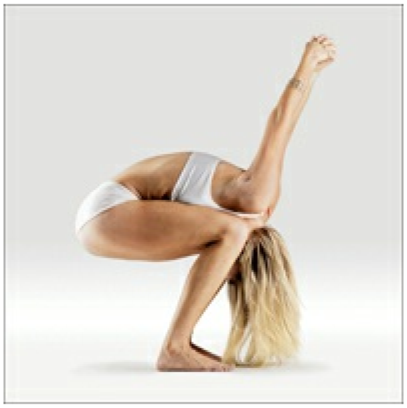

Hands Bound Revolved Fierce Pose

Baddha Hasta Parivritta Utkatasana

(Buh-duh HUH-stuh puh-ri-VRIT-tuh ut-kuh-TAHS-uh-nuh)

Also Known As: Hands Bound Revolved Chair Pose
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

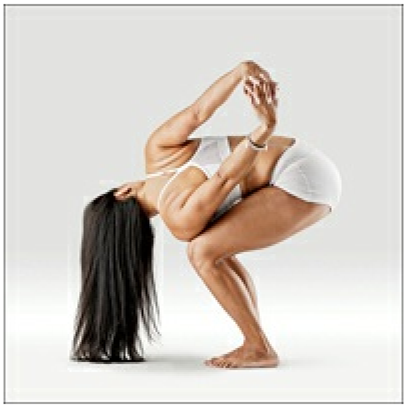

CHAIR SQUAT: HIGH STANCE—TWIST, ONE HAND TO THE FLOOR

Intense Wrist Stretch Revolved Half Bound Pose Dedicated to Yogi Shankara

Uttana Manibandha Parivritta Ardha Baddha Shankarasana

(ut-TAHN-uh muh-ni-BUHN-duh puh-ri-VRIT-tuh UHR-duh BUH-duh shuhng-kuhr-AHS-uh-nuh)

Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

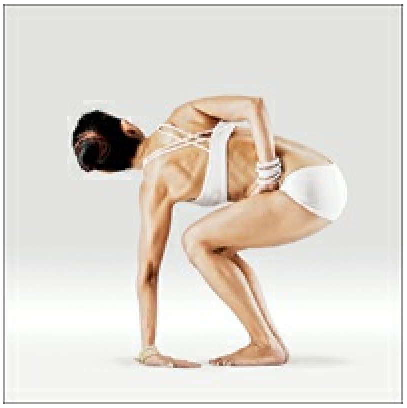

Revolved Pose Dedicated to Yogi Shankara

Parivritta Sankarasana

(puh-ri-VRIT-tuh shuhng-kuhr-AHS-uh-nuh)

Pose Type: standing, forward bend, side bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

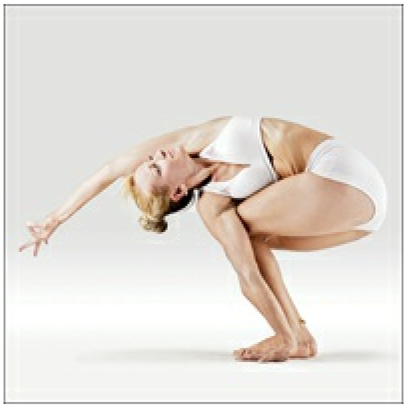

CHAIR SQUAT: HIGH STANCE—TWIST AND ELBOW OVER THE KNEE

Revolved Fierce Pose

Parivritta Utkatasana

(puh-ri-VRIT-tuh ut-kuh-TAHS-uh-nuh)

Also Known As: Revolved Chair Pose
Modification: feet wide apart, hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

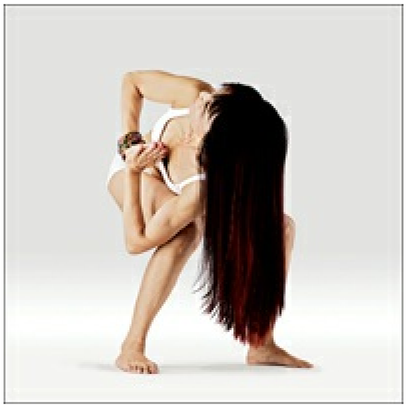

Revolved Fierce Pose

Parivritta Utkatasana

(puh-ri-VRIT-tuh ut-kuh-TAHS-uh-nuh)

Also Known As: Revolved Chair Pose
Modification: feet together, hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

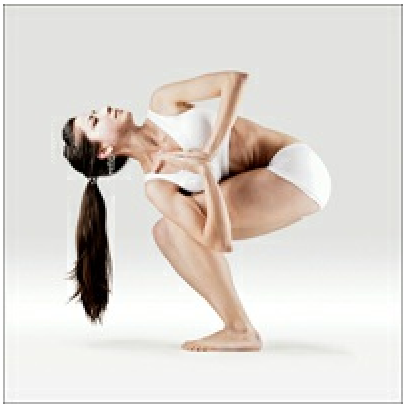

Revolved Fierce Pose

Parivritta Utkatasana

(puh-ri-VRIT-tuh ut-kuh-TAHS-uh-nuh)

Also Known As: Revolved Chair Pose
Modification: hands behind the head
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
●

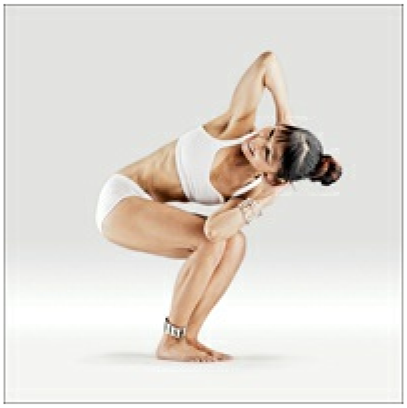

Revolved Fierce Pose

Parivritta Utkatasana

(puh-ri-VRIT-tuh ut-kuh-TAHS-uh-nuh)

Also Known As: Revolved Chair Pose
Modification: arm wrapped under the knee, hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
● ●

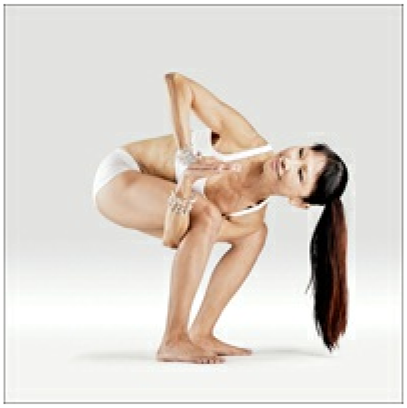

One Hand Bound Revolved Fierce Pose

Eka Hasta Baddha Parivritta Utkatasana

(EY-kuh HUH-stuh BUH-duh puh-ri-VRIT-tuh ut-kuh-TAHS-uh-nuh)

Also Known As: One Hand Bound Revolved Chair Pose
Modification: 1. one hand to the heart
2. hand grabbing onto the biceps, other elbow over the opposite knee
Pose Type: standing, forward bend, twist, binding
Drishti Point: 1. Urdhva or Antara Drishti (up to the sky)
2. Padhayoragrai or Padayoragre (toes/feet)

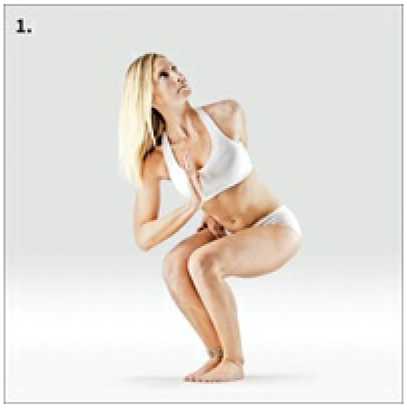

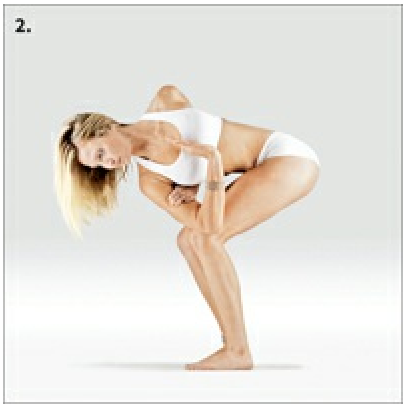

Revolved One Leg Bound Fierce Pose

Parivritta Eka Pada Baddha Utkatasana

(puh-ri-VRIT-tuh EY-kuh PUH-duh BUH-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Revolved One Leg Bound Chair Pose
Pose Type: standing, forward bend, twist, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

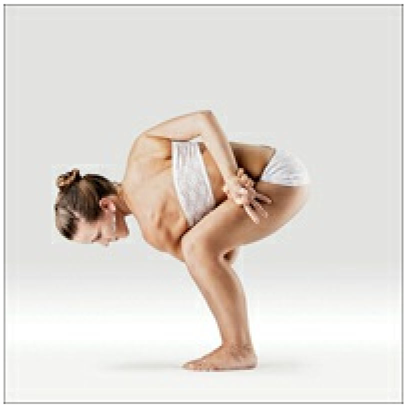

Revolved One Leg Bound Fierce Pose

Parivritta Eka Pada Baddha Utkatasana

(puh-ri-VRIT-tuh EY-kuh PUH-duh BUH-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Revolved One Leg Bound Chair Pose
Modification: binding around one leg, hands bound overhead
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

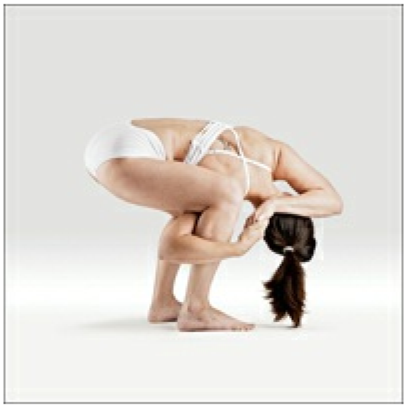

Revolved Both Legs Bound Fierce Pose

Parivritta Dwi Pada Baddha Utkatasana

(puh-ri-VRIT-tuh DWI-puh-duh BUH-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Revolved Both Legs Bound Chair Pose
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

CHAIR SQUAT: HIGH STANCE HEELS UP—CHEST AWAY FROM QUADRICEPS

Tip Toe Fierce Pose 1

Prapada Utkatasana 1

(PRUH-puh-duh ut-kuh-TAHSuh-nuh)

Also Known As: Tip Toe Chair Pose 1
Modification: arms up over the head, fingers interlocked
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

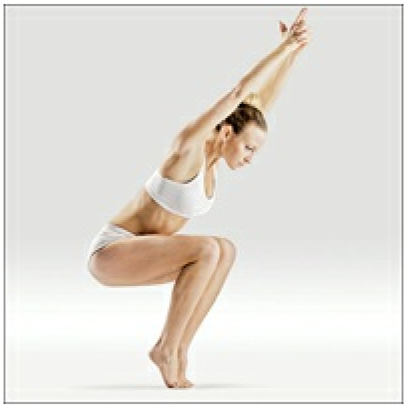

Tip Toe Fierce Pose 3

Prapada Utkatasana 3

(PRUH-puh-duh ut-kuh-TAHSuh-nuh)

Also Known As: Tip Toe Chair Pose 3
Modification: both elbows bent, one arm up in front, one arm down to the back
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

CHAIR SQUAT: HIGH STANCE HEELS UP—KNEES TO THE ARMPITS, VARIOUS HAND POSITIONS TO THE FLOOR

Crane Pose Prep.

Bakasana Prep.

(buhk-AHS-uh-nuh)

Modification: hands to the floor in front of the feet

1. fingertips to the floor
2. palms flat to the floor, arms shoulder width apart, fingertips pointing away from the head

Pose Type: standing, forward bend

Drishti Point: Hastagrai or Hastagre (hands), Nasagrai or Nasagre (nose)

● ●

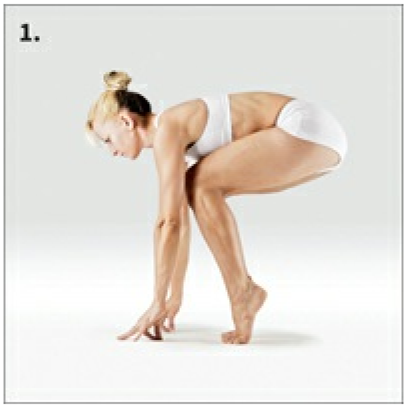

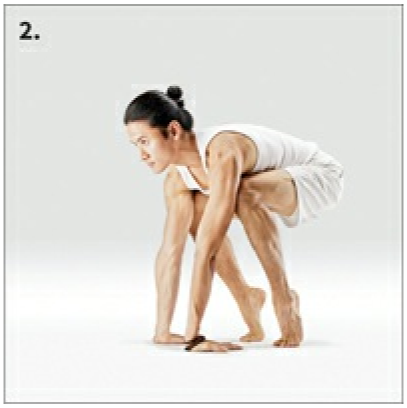

Tip Toe Fierce Pose

Prapada Utkatasana

(PRUH-puh-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Awkward Pose (Utkatasana)
Modification: arms out in front and parallel to the floor, sitting bones lifted off the heels
Pose Type: standing
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
लं

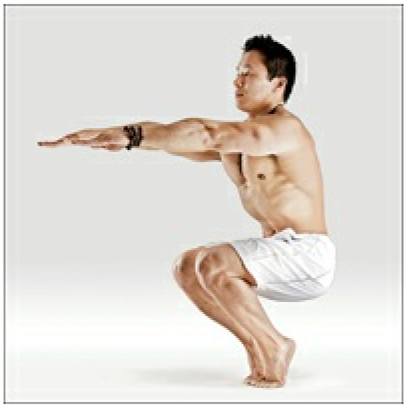

Tip Toe Fierce Pose 2

Prapada Utkatasana 2

(PRUH-puh-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Tip Toe Chair Pose 2 and Tip Toe Diver’s Pose
Modification: arms extended to the back, chest to the quadriceps, palms facing down
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं यं हं ॐ

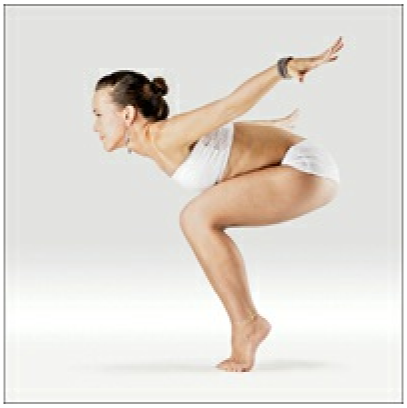

Head to Knees Tip Toe Fierce Pose 2

Janu Shirsha Prapada Utkatasana 2

(JAH-nu SHEER-shuh PRUH-puh-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Head to Knees Tip Toe Chair Pose 2 and Head to Knees Tip Toe Diver’s Pose
Modification: knees together, arms straight out to the sides, forehead to the knees
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं रं यं हं ॐ

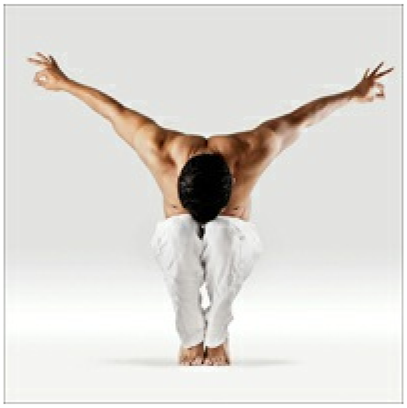

CHAIR SQUAT: HIGH STANCE HEELS UP—BACK OF THE KNEE OVER THE SHOULDER

Unsupported Tip Toe Wagtail Pose 1

Niralamba Prapada Khanjanasana 1

(nir-ah-LUHM-buh PRUH-puh-duh kuhn-juh-NAHS-uh-nuh)

Modification: elbows bent, hands reaching to the front
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

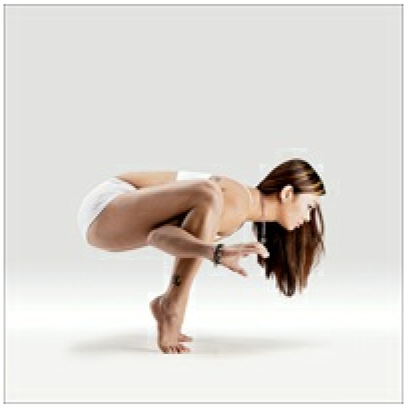

Unsupported Bound Tip Toe Wagtail Pose

Niralamba Baddha Prapada Khanjanasana

(nir-ah-LUHM-buh BUH-duh PRUH-puh-duh kuhn-juh-NAHS-uh-nuh)

Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

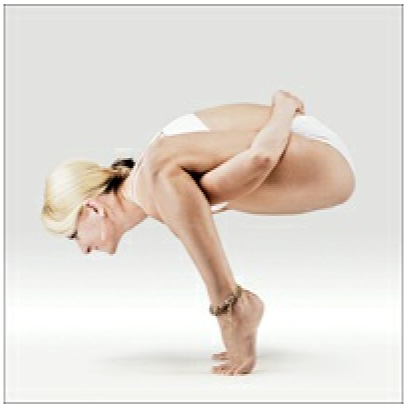

Supported Tip Toe Wagtail Pose

Salamba Prapada Khanjanasana

(sah-LUHM-buh PRUH-puh-duh kuhn-juh-NAHS-uh-nuh)

Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

Unsupported Tip Toe Wagtail Pose 2

Niralamba Prapada Khanjanasana 2

(nirsah-LUHM-buh PRUH-puh-duh kuhn-juh-NAHS-uh-nuh)

Modification: fingertips reaching away from the head, palms pressed together, high stance
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं हं ॐ

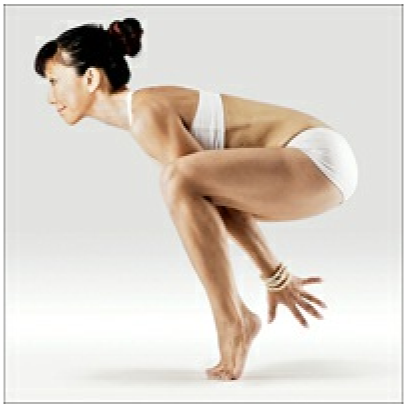

Intense Ankle Stretch Tip Toe Fierce Pose 3

Uttana Kulpa Prapada Utkatasana 3

(ut-TAH-nuh KUL-puh PRUH-puh-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Intense Ankle Stretch Tip Toe Chair Pose 3

Modification: both elbows bent, one arm up in front, one arm down to the back

Pose Type: standing, balance

Drishti Point: Hastagrai or Hastagre (hands)

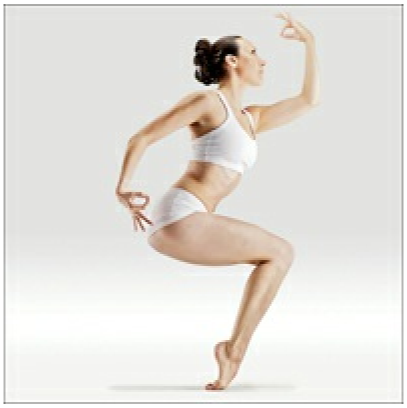

Revolved Intense Ankle Stretch Tip Toe Uneven Legs Half Intense Stretch Pose

Parivritta Uttana Kulpa Prapada Vishama Pada Ardha Uttanasana

(puh-ri-VRIT-tuh ut-TAH-nuh KUL-puh PRUH-puh-duh VISH-uh-muh PUH-duh UHR_duh ut-tahn-AHS-uh-nuh)

Also Known As: Revolved Intense Ankle Stretch Tip Toe Uneven Legs Half Forward Bend

Modification: fingertips off the floor

Pose Type: standing, balance, forward bend, twist

Drishti Point: Hastagrai or Hastagre (hands)

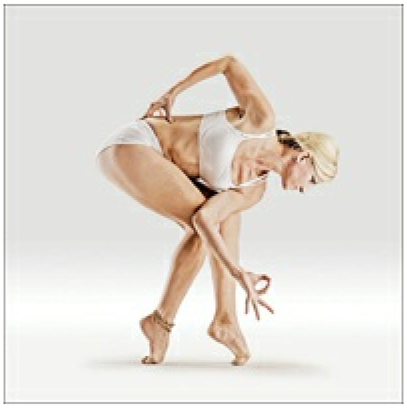

Intense Ankle Stretch Tip Toe Fierce Pose Prep.

Uttana Kulpa Prapada Utkatasana Prep.

(ut-TAH-nuh KUL-puh PRUH-puh-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Intense Ankle Stretch Tip Toe Chair Pose Prep.
Modification: fingertips to the floor
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands)
ल र

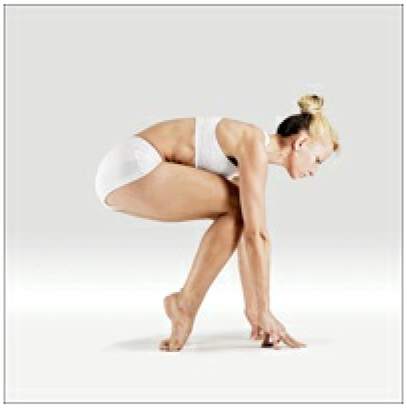

Intense Ankle Stretch Tip Toe Fierce Pose 2

Uttana Kulpa Prapada Utkatasana 2

(ut-TAH-nuh KUL-puh PRUH-puh-duh ut-kuh-TAHS-uh-nuh)

Also Known As: Intense Ankle Stretch Tip Toe Chair Pose 2 and Intense Ankle Stretch Tip Toe Diver’s Pose
Modification: arms extended to the back, chest to the quadriceps
Pose Type: standing, balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल र य

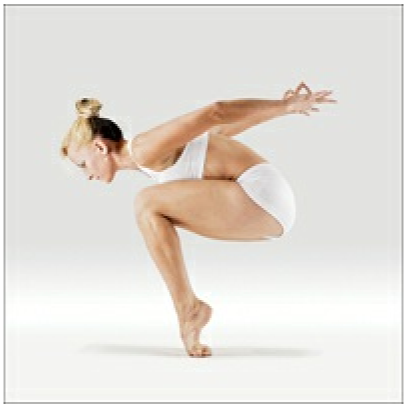

CHAIR SQUAT: HIGH STANCE HEELS UP—ANKLES CROSSED, HANDS TO THE FLOOR

Revolved Tip Toe Pose

Parivritta Prapadasana

(puh-ri-VRIT-tuh pruh-puh-DAHS-uh-nuh)

Modification: ankles crossed, one arm to the back, other hand toward the floor in front of the feet
Pose Type: standing, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

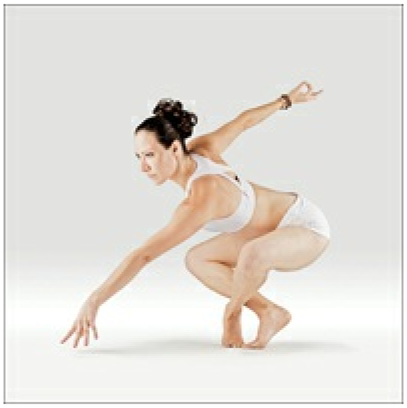

Tip Toe Pose

Prapadasana

(pruh-puh-DAHS-uh-nuh)

Modification: ankles crossed, arms crossed, fingertips to the floor in front of the feet
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

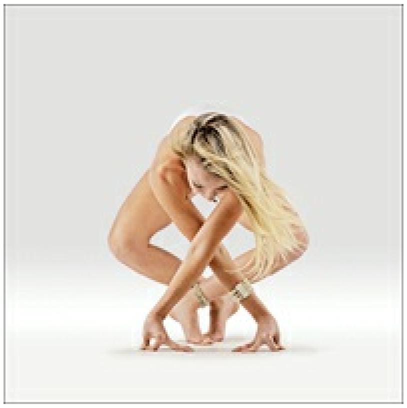

Pendant Pose Prep.

Lolasana Prep.

(lo-LAHS-uh-nuh)

Modification: ankles crossed
Pose Type: standing, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

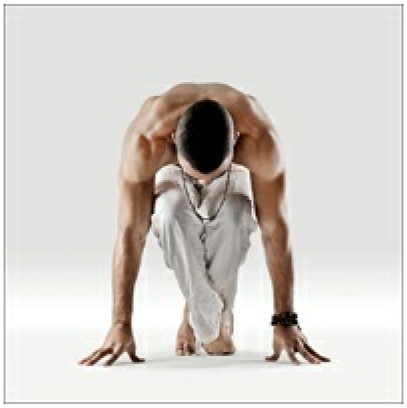

SQUAT: HEELS UP—LEGS CROSSED

Easy Leg Position of the Pose Dedicated to Garuda in Yoga Squat with Hands in Prayer

Sukha Pada Garudasana in Upaveshasana Namaskar

(SUK-kuh PUH-duh guh-ru-DAHS-uh-nuh in u-puh-veysh-AHS-uh-nuh nuh-muhs-KAHR)

Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagrahe (hands)

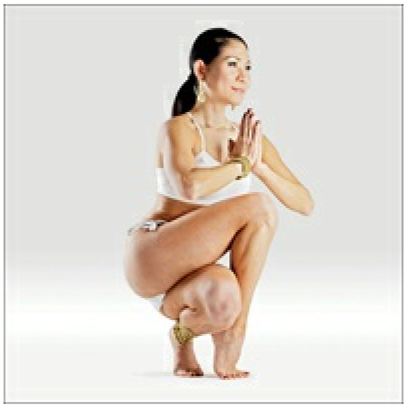

Easy Leg Position and Complete Arm Position of the Pose Dedicated to Garuda in Yoga Squat

Sukha Pada Paripurna Hasta Garudasana in Upaveshasana

(SUK-kuh PUH-duh puh-ri-POOR-nuh HUH-stuh guh-ru-DAHS-uh-nuh in u-puh-veysh-AHS-uh-nuh)

Pose Type: standing, forward bend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

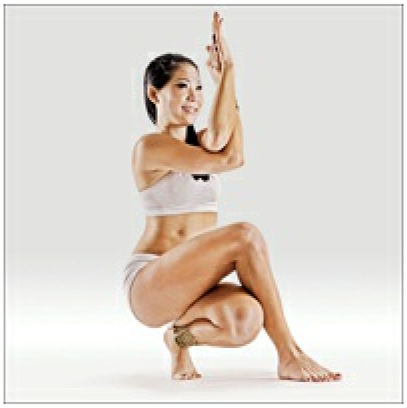

Revolved Easy Leg Position of the Pose Dedicated to Garuda in Yoga Squat

Parivritta Sukha Pada Garudasana in Upaveshasana

(puh-ri-VRIT-tuh SUK-kuh PUH-duh guh-ru-DAHS-uh-nuh in u-puh-veysh-AHS-uh-nuh)

Modification: both knees off the floor, arms open wide and straight

Pose Type: standing, forward bend, twist

Drishti Point: Hastagrai or Hastagrahe (hands)

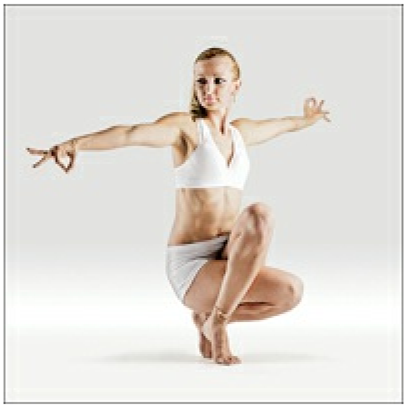

Revolved Easy Leg Position of the Pose Dedicated to Garuda in Yoga Squat

Parivritta Sukha Pada Garudasana in Upaveshasana

(puh-ri-VRIT-tuh SUK-kuh PUH-duh guh-ru-DAHS-uh-nuh in u-puh-veysh-AHS-uh-nuh)

Modification: bottom knee on the floor, arms open wide and straight
Pose Type: standing, forward bend, twist
Drishti Point: Hastagrai or Hastagrahe (hands)
लं वं यं

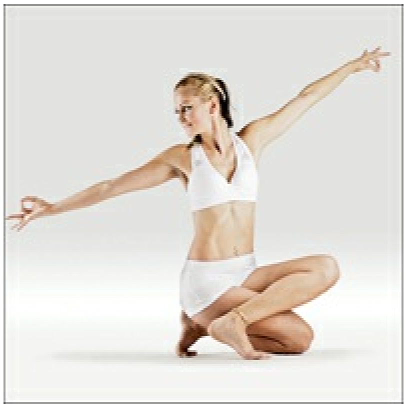

Revolved Easy Leg Position of the Pose Dedicated to Garuda in Yoga Squat

Parivritta Sukha Pada Garudasana in Upaveshasana

(puh-ri-VRIT-tuh SUK-kuh PUH-duh guh-ru-DAHS-uh-nuh in u-puh-veysh-AHS-uh-nuh)

Modification: elbow over the opposite knee, other arm straight out to the side
Pose Type: standing, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
लं वं रं

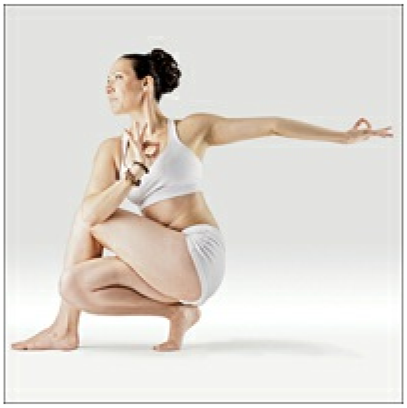

SQUAT: HEELS UP—KNEES APART

Hand Position of the Pose Dedicated to Garuda in Half Tip Toe Hero Pose

Hasta Garudasana in Ardha Prapada Virasana

(HUH-stuh guh-ru-DAHS-uh-nuh in UHR-duh PRUH-puh-duh veer-AHS-uh-nuh)

Modification: one knee on the floor, one foot to the floor on the inside of the knee
Pose Type: standing, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

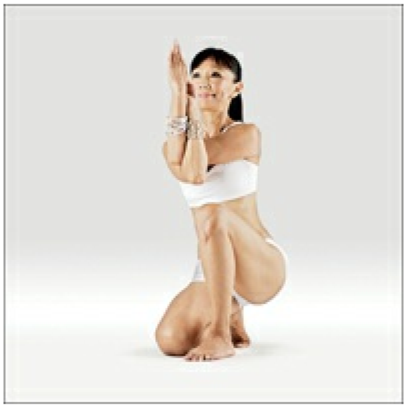

Half Bound Hand to Foot Uneven Tip Toe Pose

Ardha Baddha Hasta Pada Vishama Prapadasana

(UHR-duh BUH-duh HUH-stuh PUH-duh VISH-uh-muh pruh-puh-DAHS-uh-nuh)

Modification: one hand crossed in front, grabbing onto the heel; other hand to the inner thigh on the opposite side
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

Hand Position of the Pose Dedicated to Garuda in Half Lotus Tip Toe Pose

Hasta Garudasana in Ardha Padma Prapadasana

(HUH-stuh guh-ru-DAHS-uh-nuh in UHR-duh PUHD-muh pruh-puh-DAHS-uh-nuh)

Modification: 1. Both knees on the floor
2. One knee on the floor
Pose Type: standing
Drishti Point: 1. Angusthamadhye or Angustha Ma Dyai (thumbs)
2. Nasagrai or Nasagre (nose)

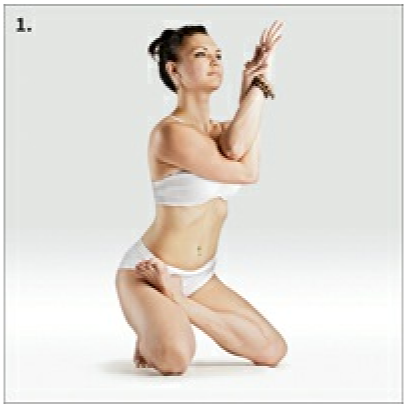
1.

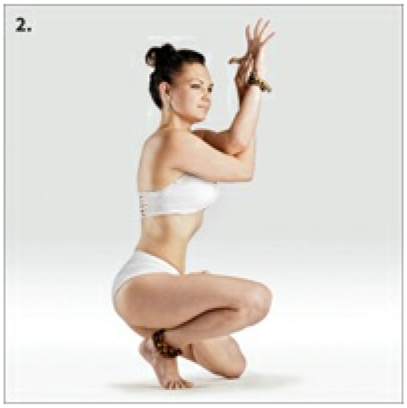
2.

Tip Toe Pose

Prapadasana

(pruh-puh-DAHS-uh-nuh)

Also Known As: Noose Pose Prep. (Pasasana Prep.)
Modification: 1. fingertips to the floor on the outside of the knees
2. one hand to the heart, other hand to the floor
3. hands on the hips
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

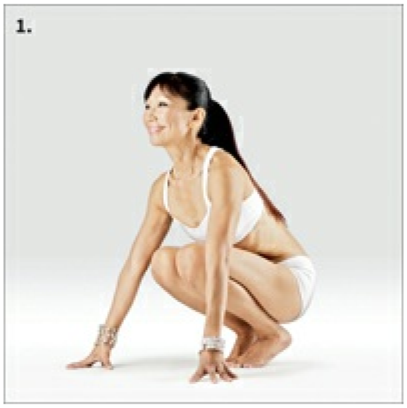

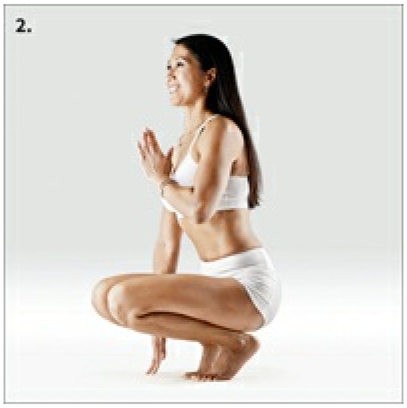

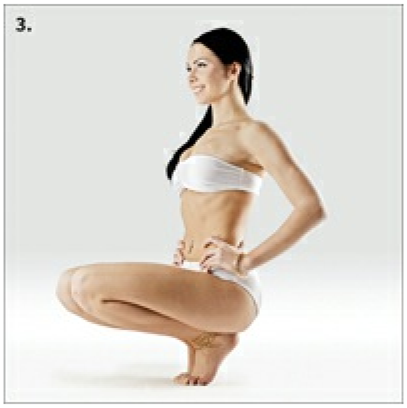

Hands Bound Tip Toe Pose

Baddha Hasta Prapadasana

(BUH-duh HUH-stuh pruh-puh-DAHS-uh-nuh)

Modification: arms straight out in front, palms facing out, chest to the quadriceps
Pose Type: standing, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

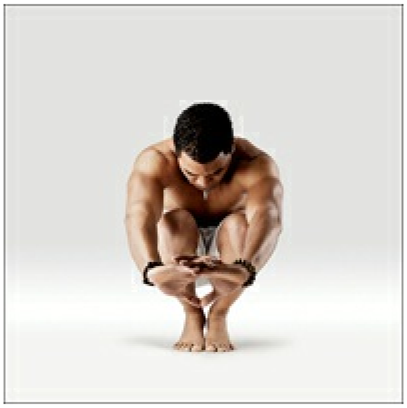

Tip Toe Pose with Hands in Prayer

Prapadasana Namaskar

(pruh-puh-DAHS-uh-nuh nuh-muhs-KAHR)

Modification: knees together, chest to the quadriceps
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

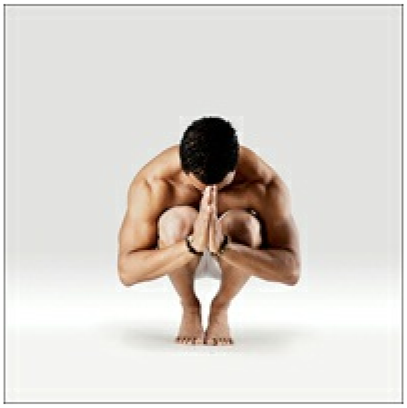

Tip Toe Pose

Prapadasana

(pruh-puh-DAHS-uh-nuh)

Also Known As: Garland Pose Prep. (Malasana Prep.), Full Squat Pose (Purna Utkatasana)
Modification: 1. arms straight out in front and parallel to the floor
2. arms up over the head, palms pressed together, elbows bent
Pose Type: standing
Drishti Point: Hastagrai or Hastagre (hands)
Drishti Point: 1. Hastagrai or Hastagrahe (hands)
2. Nasagrai or Nasagre (nose)

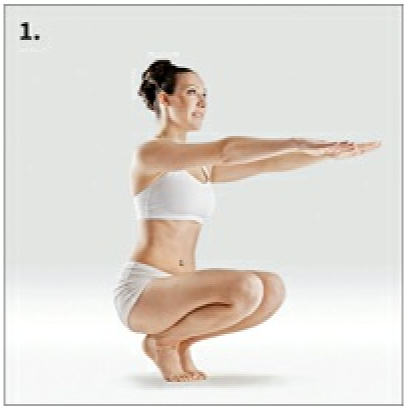

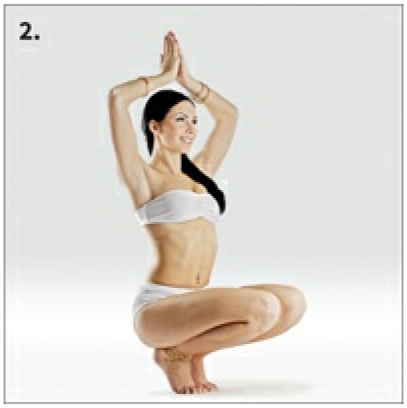

Half Bound Tip Toe Pose

Ardha Baddha Prapadasana

(UHR-duh BUH-duh pruh-puh-DAHS-uh-nuh)

Modification: one hand binding to the opposite hip behind the back, other arm straight and extended out to the side
Pose Type: standing, binding
Drishti Point: Nasagrai or Nasagre (nose)

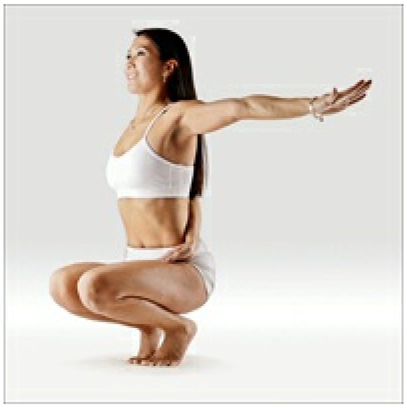

SQUAT: HEELS UP—KNEES TOGETHER—KNEES SHOULDER LEVEL

Seated Tip Toe Pose

Upavishta Prapadasana

(u-puh-VISH-tuh pruh-puh-DAHS-uh-nuh)

Also Known As: Crouching Tip Toe Pose; Upavistha Prapadasana
Modification: elbows resting on the knees, fingers interlocked in front of the face
Pose Type: standing, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

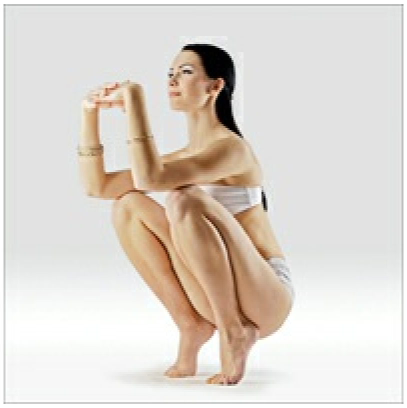

Hands Bound Seated Tip Toe Pose

Baddha Hasta Upavishta Prapadasana

(BUH-duh HUH-stuh u-puh-VISH-tuh pruh-puh-DAHS-uh-nuh)

Also Known As: Hands Bound Crouching Tip Toe Pose; Baddha Hasta Upavistha Prapadasana
Modification: arms away from the body
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

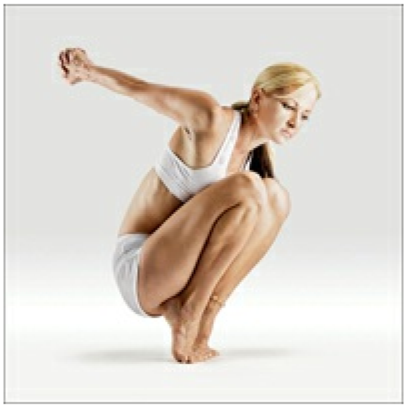

Knees to the Shoulders Seated Tip Toe Pose

Janu Bhuja Upavishta Prapadasana

(JAH-nu BHU-juh u-puh-VISH-tuh pruh-puh-DAHS-uh-nuh)

Also Known As: Knees to the Shoulders Crouching Tip Toe Pose; Janu Bhuja Upavistha Prapadasana
Modification: 1. palms together to the floor
2. palms apart and to the floor, elbows bent
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

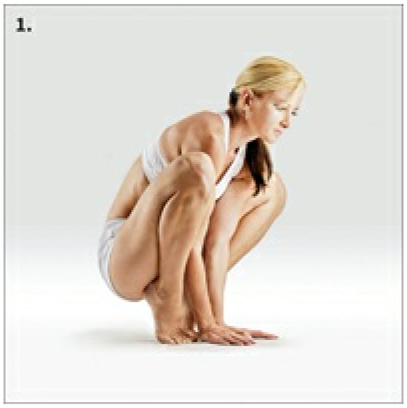

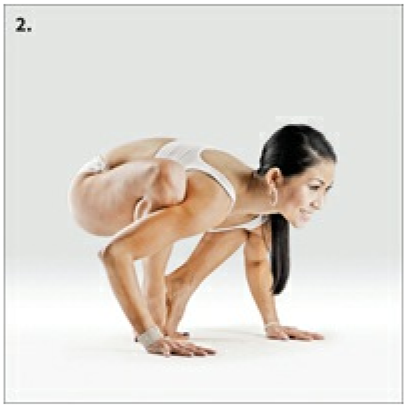

Supported Seated Tip Toe Pose

Salamba Upavishta Prapadasana

(sah-LUHM-buh u-puh-VISH-tuh pruh-puh-DAHS-uh-nuh)

Also Known As: Supported Crouching Tip Toe Pose; Salamba Upavistha Prapadasana
Modification: fingertips to the floor in front of the feet
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

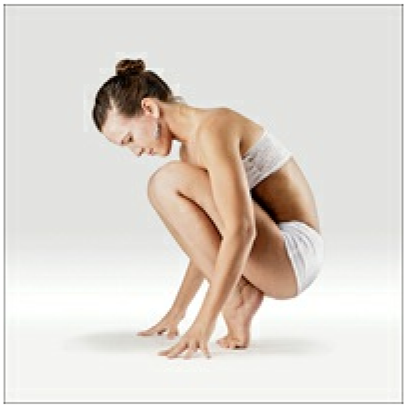

Supported/Unsupported Intense Ankle Stretch Seated Tip Toe Pose

Salamba/Niralamba Uttana Kulpa Upavishta Prapadasana

(sah-LUHM-buh-/nir-ah-LUHM-buh ut-TAHN-uh KUL-puh u-puh-VISH-tuh pruh-puh-DAHS-uh-nuh)

Also Known As: Supported/Unsupported Intense Ankle Stretch Crouching Tip Toe Pose; Salamba/Niralamba Uttana Kulpa Upavistha Prapadasana
Modification: 1. hands to the floor by the hips
2. arms wrapped around the shins, grabbing onto the triceps
Pose Type: 1. standing, forward bend
2. standing, balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

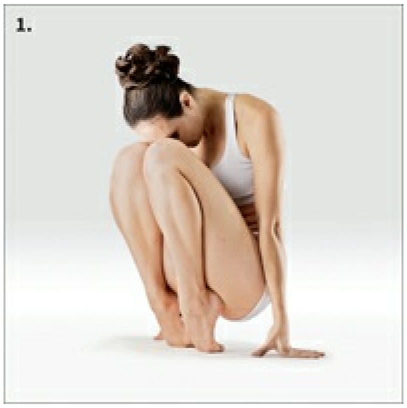

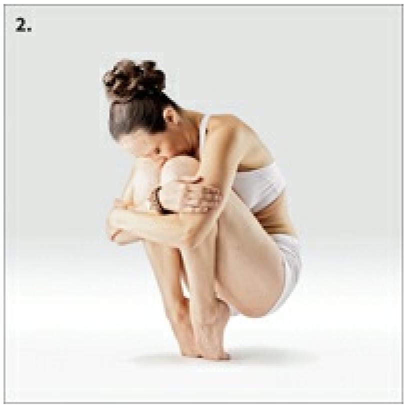

SQUAT: HEELS UP—TWISTS

Tip Toe Easy Noose Pose Prep.

Prapada Sukha Pashasana Prep.

(PRUH-puh-duh SUK-kuh puh-SHAHS-uh-nuh)

Modification: elbow on the inside of the thigh, arms open wide
Pose Type: standing, twist
Drishti Point: Hastagrai or Hastagre (hands)
ल य

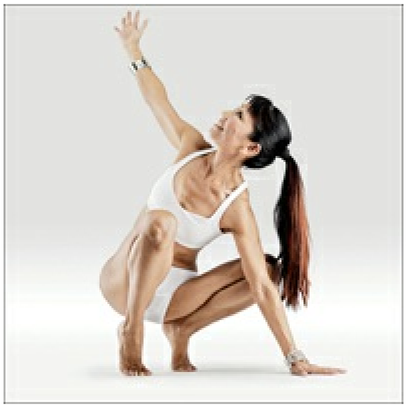

Uneven Legs One Hand to Foot Noose Pose

Vishama Pada Eka Hasta Pada Pashasana

(VISH-uh-muh PUH-duh EY-kuh HUH-stuh PUH-duh puh-SHAS-uh-nuh)

Modification: one heel up, one heel down
Pose Type: standing, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल

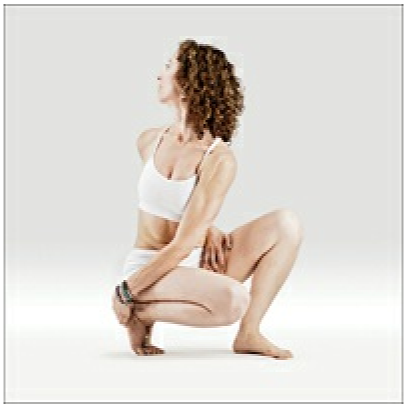

Extended Side Tip Toe Noose Pose

Utthita Parshva Prapada Pashasana

(UT-ti-tuh PAHRSH-vuh PRUH-puh-duh puh-SHAHS-uh-nuh)

Modification: one hand to the floor, other arm extended up over the head
Pose Type: standing, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

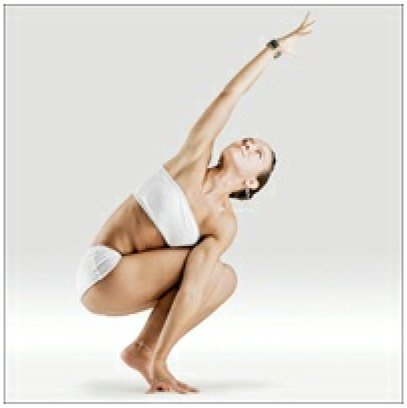

Tip Toe Noose Pose with Hands in Prayer

Prapada Pashasana Namaskar

(PRUH-puh-duh puh-SHAHS-uh-nuh nuh-muhs-KAHR)

Pose Type: standing, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

One Hand to the Foot Tip Toe Noose Pose

Eka Hasta Pada Prapada Pashasana

(EY-kuh HUH-stuh PUH-duh PRUH-puh-duh puh-SHAHS-uh-nuh)

Modification: one hand to the opposite heel, other hand binding to the opposite hip behind the back
1. right side view
2. left side view

Pose Type: standing, twist, binding

Drishti Point: Urdhva or Antara Drishti (up to the sky)

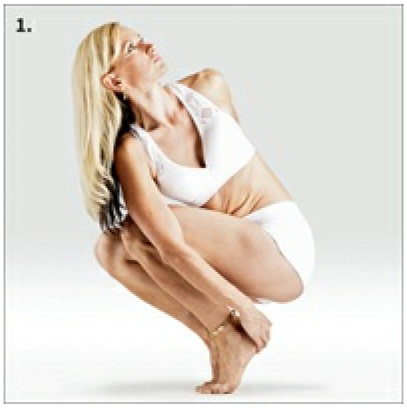
1.

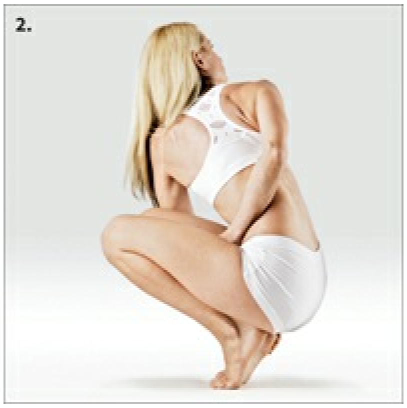
2.

Tip Toe Noose Pose Prep.

Prapada Pashasana Prep.

(PRUH-puh-duh puh-SHAHS-uh-nuh)

Modification: 1. binding around one leg with a strap
2. binding around one leg without a strap
Pose Type: standing, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

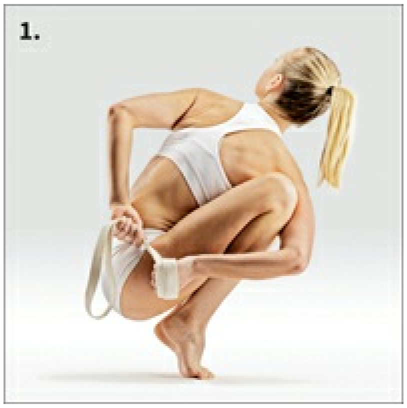

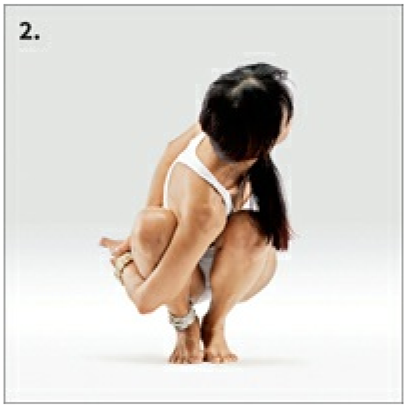

Tip Toe Noose Pose

Prapada Pashasana

(PRUH-puh-duh puh-SHAHS-uh-nuh)

Modification: binding around both legs
Pose Type: standing, forward bend, twist, binding
Drishti Point: Nasagrai or Nasagre (nose)

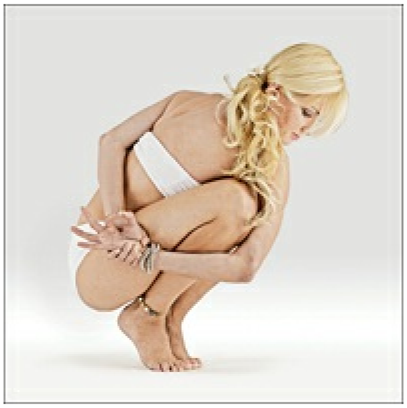

SQUAT: HEELS DOWN—ARMS BELOW HEAD

Intense Wrist Stretch Garland Pose Prep.

Uttana Manibandha Malasana Prep.

(ut-TAH-nuh muh-ni-BUHN-duh mah-LAHS-uh-nuh)

Modification: heels down, knees together, palms to the floor, fingers facing the toes, chest to the quadriceps
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
● ● ●

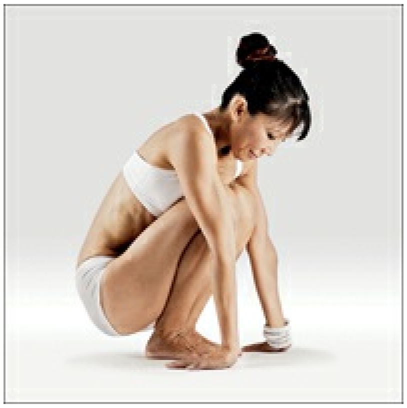

Garland Pose with Hands in Prayer Prep.

Malasana Namaskar Prep.

(mah-LAHS-uh-nuh nuh-muhs-KAHR)

Modification: heels down
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

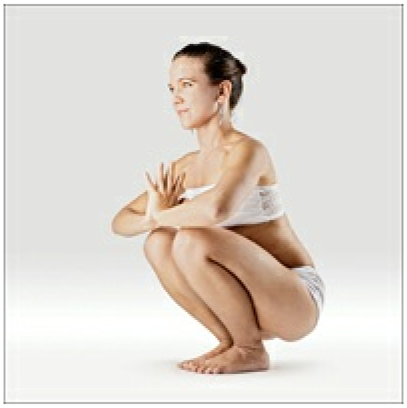

Garland Pose Prep.

Malasana Prep.

(mah-LAHS-uh-nuh)

Also Known As: Full Squat Pose (Purna Utkatasana)
Modification: knees together, heels down; arms straight out to the front, palms down
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

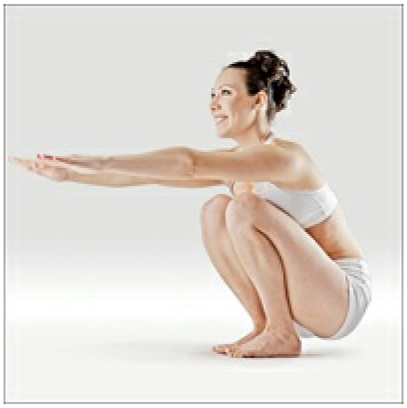

Arms Spread Out Garland Pose Prep.

Prasarita Hasta Malasana Prep.

(pruh-SAH-ri-tuh HUH-stuh mah-LAHS-uh-nuh)

Also Known As: Full Squat Pose (Purna Utkatasana), Flying Bird Pose (Khagasana)
Modification: knees together, heels down, arms straight out to the side
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

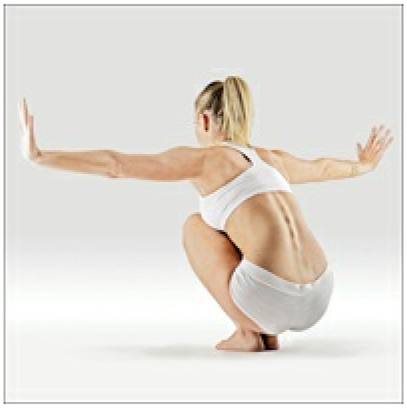

SQUAT: HEELS DOWN—ARMS ABOVE HEAD

Upward Hands Garland Pose Prep.

Urdhva Hasta Malasana Prep.

(OORD-vuh HUH-stuh mah-LAHS-uh-nuh)

Also Known As: Full Squat Pose (Purna Utkatasana)
Modification: knees together, hands to the sky
Pose Type: standing, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

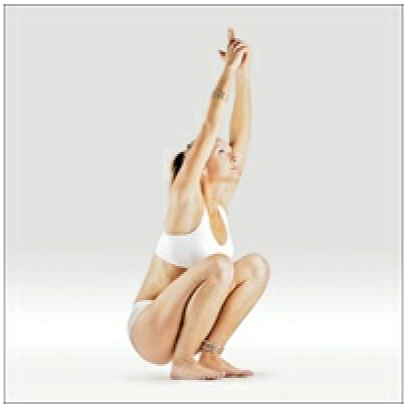

Easy Noose Pose

Sukha Pashasana

(SUK-kuh puh-SHAHS-uh-nuh)

Also Known As: Revolved Yoga Squat (Parivritta Upaveshasana)
Modification: heels down, shoulder between the knees, one palm flat on the floor, other arm straight out to the back
Pose Type: standing, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

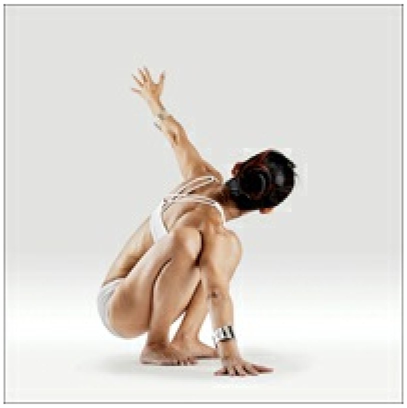

Extended Side Noose Pose

Utthita Parshva Pashasana

(UT-ti-tuh PAHRSH-vuh puh-SHAHS-uh-nuh)

Modification: heels down, one hand to the floor, other arm extended up over the head
Pose Type: standing, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

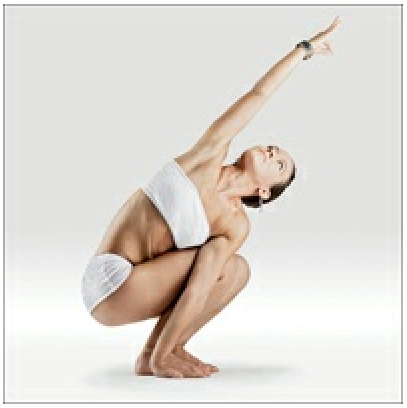

SQUAT: HEELS DOWN—TWISTS

Noose Pose with Hands in Prayer

Pashasana Namaskar

(puh-SHAHS-uh-nuh nuh-muhs-KAHR)

Modification: heels down
Pose Type: standing, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

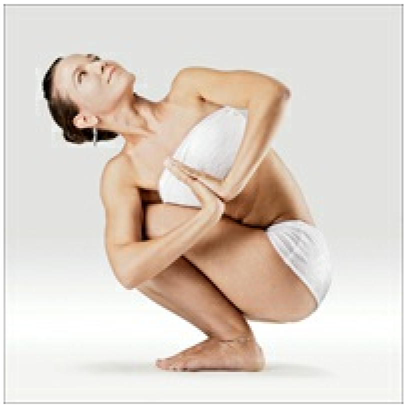

Half Bound Noose Pose

Ardha Baddha Pashasana

(UHR-duh BUH-duh puh-SHAHS-uh-nuh)

Modification: heels down
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

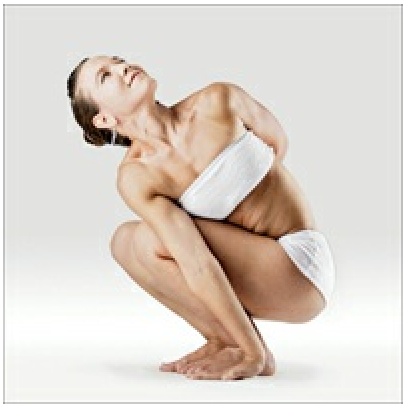

Noose Pose

Pashasana

(puh-SHAHS-uh-nuh)

Modification: heels down, binding with a strap
1. binding around one leg
2. binding around both legs
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

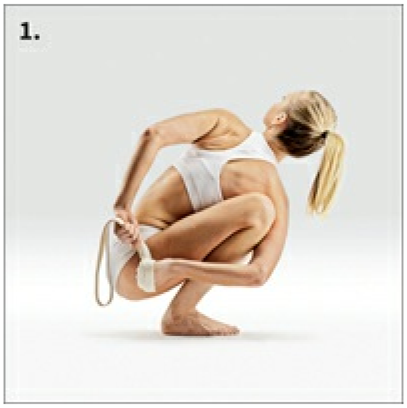

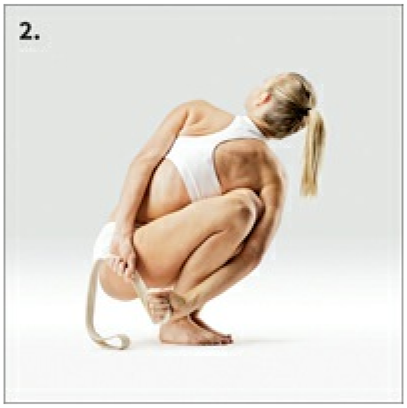

Noose Pose

Pashasana

(puh-SHAHS-uh-nuh)

Modification: heels down, binding around both legs
1. side view
2. front view
Pose Type: standing, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

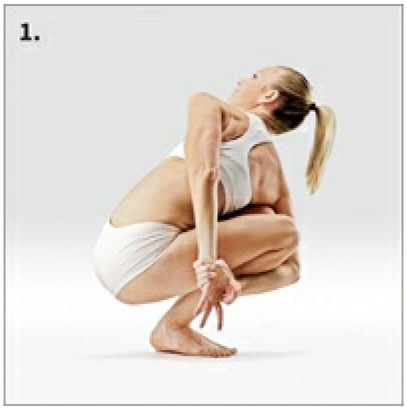

2.

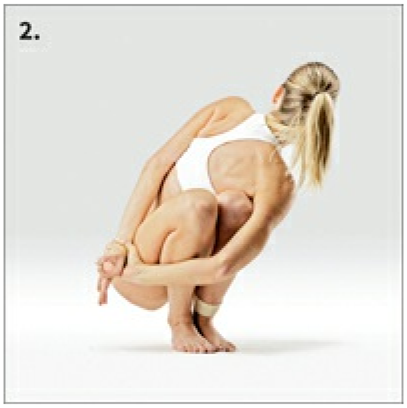

ONE-LEGGED SQUAT: FOOT TO KNEE

One-Legged Tip Toe Pose

Eka Pada Prapadasana

(EY-kuh PUH-duh pruh-puh-DAHS-uh-nuh)

Also Known As: Half Bound Lotus Tip Toe Pose Prep. (Ardha Baddha Padma Padangushtasana Prep.)
Modification: 1. both hands to the floor on the sides
2. hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)

1.

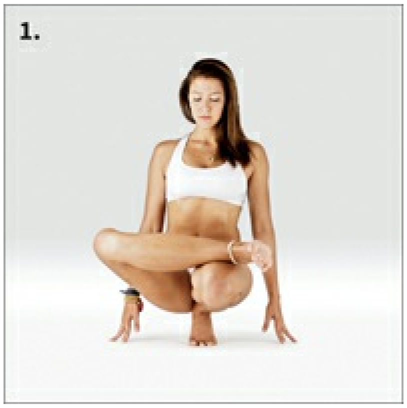

2.

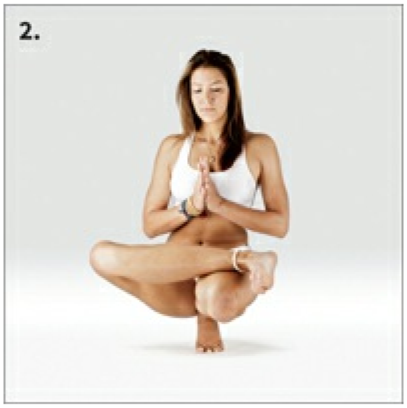

One-Legged Tip Toe Pose

Eka Pada Prapadasana

(EY-kuh PUH-duh pruh-puh-DAHS-uh-nuh)

Also Known As: Half Bound Lotus Tip Toe Pose Prep. (Ardha Baddha Padma Padangushtasana Prep.)
Modification: 1. one hand on the hip, other hand to the floor
2. one hand on the hip, one arm up over the head
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)
ल व

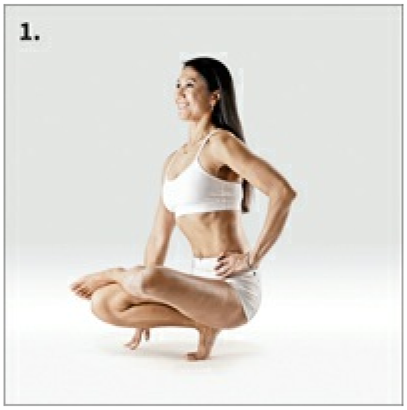

1.

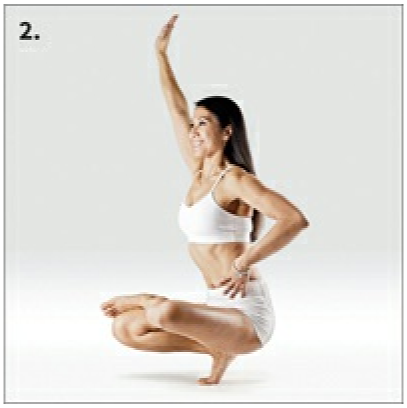

2.

Baby Cradle Pose with Hands in Prayer in One-Legged Garland Pose

Hindolasana Namaskar in Eka Pada Malasana

(hin-do-LAHS-uh-nuh EY-kuh PUH-duh mah-LAHS-uh-nuh)

Also Known As: Baby Cradle Pose with Hands in Prayer in One-Legged Yoga Squat (Hindolasana Namaskar in Eka Pada Upavesashana)
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

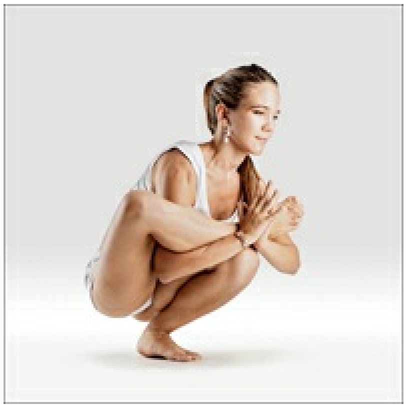

Baby Cradle Pose in One-Legged Tip Toe Pose 1

Hindolasana in Eka Pada Prapadasana 1

(hin-do-LAHS-uh-nuh in EY-kuh PUH-duh pruh-puh-DAHS-uh-nuh)

Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

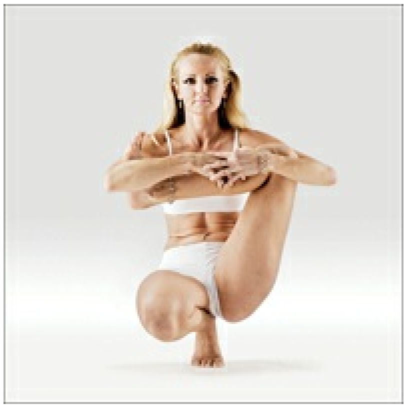

Half Yogic Staff Pose Prep. in One-Legged Tip Toe Pose

Ardha Yogadandasana Prep. in Eka Pada Prapadasana

(UHR-duh yo-guh-duhn-DAHS-uh-nuh in EY-kuh PUH-duh pruh-puh-DAHS-uh-nuh)

Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

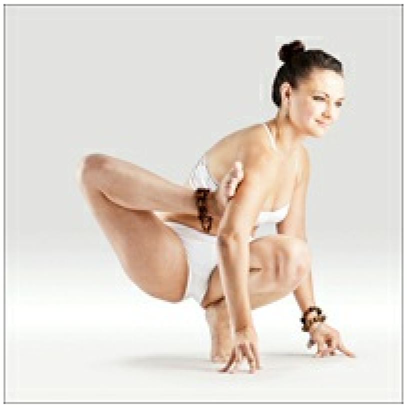

ONE-LEGGED SQUAT: FOOT UNDER THE KNEE

Baby Cradle Pose in One-Legged Tip Toe Pose 2

Hindolasana in Eka Pada Prapadasana 2

(hin-do-LAHS-uh-nuh in EY-kuh PUH-duh pruh-puh-DAHS-uh-nuh)

Pose Type: standing one-legged balance, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

लं वं रं यं

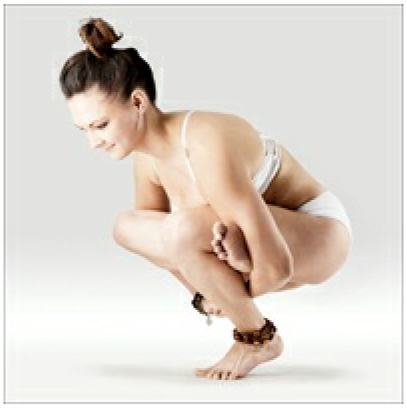

Supported Revolved One-Legged Tip Toe Pose

Salamba Parivritta Eka Pada Prapadasana

(sah-LUHM-buh puh-ri-VRIT-tuh EY-kuh PUH-duh pruh-puh-DAHS-uh-nuh)

Modification: foot under the knee

Pose Type: standing one-legged balance, forward bend, twist

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

लं वं

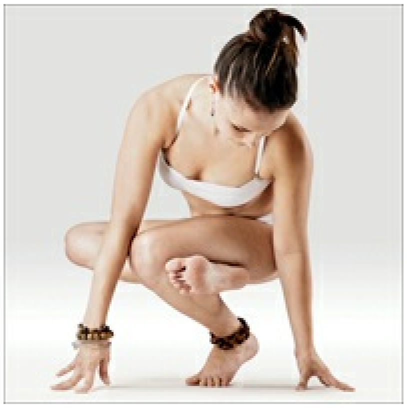

ONE-LEGGED SQUAT: FOOT TO KNEE & ONE KNEE OVER THE SHOULDER

Tip Toe Lord of the Fishes Pose

Prapada Matsyendrasana

(PRUH-puh-duh muhts-y-eyn-DRAHS-uh-nuh)

Pose Type: standing one-legged balance, twist

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

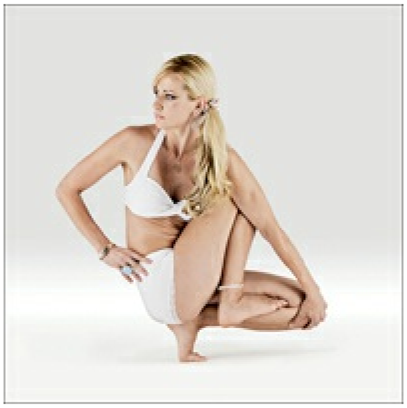

Supported One-Legged Tip Toe Pose

Salamba Eka Pada Prapadasana

(sah-LUHM-buh EY-kuh PUH-duh pruh-puh-DAHS-uh-nuh)

Modification: leg wrapped around the tricep, fingertips to the floor

Pose Type: standing one-legged balance

Drishti Point: Nasagrai or Nasagre (nose)

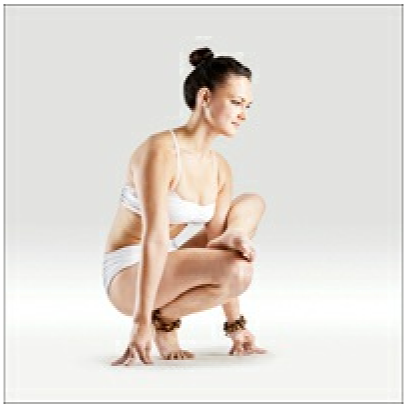

Supported One-Legged Tip Toe Pose

Salamba Eka Pada Prapadasana

(sah-LUHM-buh EY-kuh PUH-duh pruh-puh-DAHS-uh-nuh)

Modification: hands to the floor, shoulder to the back of the knee, other foot wrapped around the forearm
Pose Type: standing one-legged balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

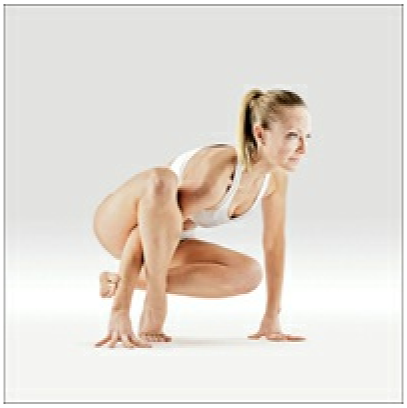

ONE-LEGGED SQUAT: HALF LOTUS

Half Bound Lotus Tip Toe Pose

Ardha Baddha Padma Prapadasana

(UHR-duh BUH-duh PUHD-muh pruh-puh-DAHS-uh-nuh)

Also Known As: Half Bound Lotus Tip Toe Pose (Ardha Baddha Padma Padangushtasana)
Modification: 1. fingertips to the floor, other hand to the heart
2. binding, other hand to the heart
3. binding, other arm out to the side, parallel to the floor
Pose Type: 1. standing one-legged balance

2 & 3. standing one-legged balance, binding

Drishti Point: Nasagrai or Nasagre (nose)

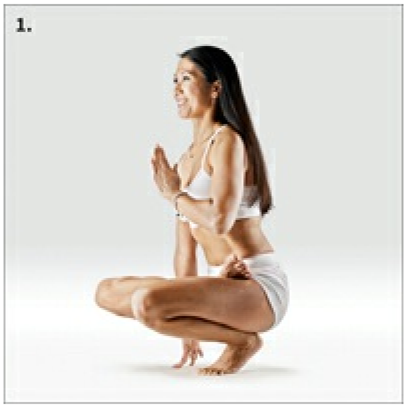

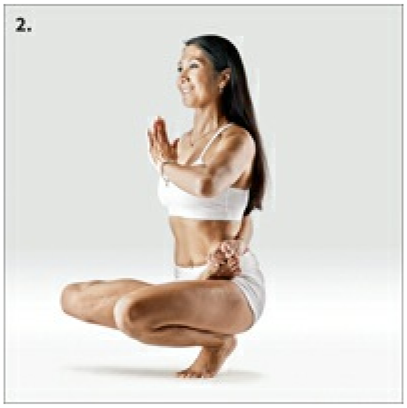

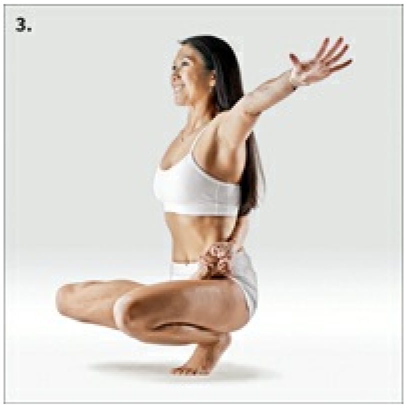

Hand Position of the Pose Dedicated to Garuda in Half Lotus Tip Toe Pose

Hasta Garudasana in Ardha Padma Prapadasana

(HUH-stuh guh-ru-DAHS-uh-nuh in UHR-duh-PUHD-muh pruh-puh-DAHS-uh-nuh)

Modification: knee on the floor, forward bend

Pose Type: standing one-legged balance, forward bend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

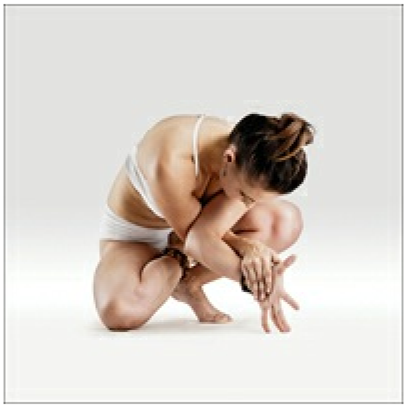

ONE-LEGGED SQUAT: HALF LOTUS—HEEL DOWN

Half Lotus One-Legged Garland Pose

Ardha Padma Eka Pada Malasana

(UHR-duh PUHD-muh EY-kuh PUH-duh mah-LAHS-uh-nuh)

Also Known As: Half Lotus One-Legged Yoga Squat (Ardha Padma Eka Pada Upaveshasana)
Modification: arms parallel to the floor
Pose Type: standing one-legged balance
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

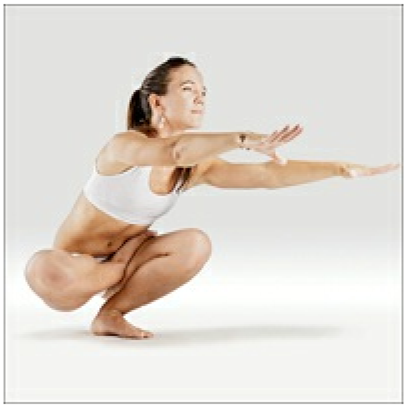

Half Lotus Bound One-Legged Garland Pose

Ardha Padma Baddha Eka Pada Malasana

(UHR-duh PUHD-muh BUH-duh EY-kuh PUH-duh mah-LAHS-uh-nuh)

Also Known As: Half Lotus Bound One-Legged Yoga Squat (Ardha Padma Baddha Eka Pada Upavesasana)
Pose Type: standing one-legged balance, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

ONE-LEGGED SQUAT: ONE LEG STRAIGHT—HEEL UP & DOWN

Hand to Foot Pose Squatting Prep.

Hasta Padasana Prep.

(HUH-stuh puh-DAHS-uh-nuh)

Modification: leg straight and lifted off the floor, arms parallel to the floor
Pose Type: standing one-legged balance
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

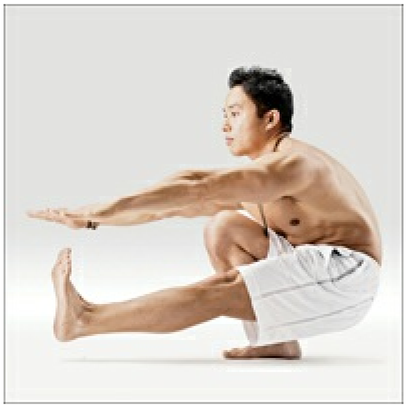

Hands to Foot Squatting Pose 1

Hasta Padasana 1

(HUH-stuh puh-DAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi Modification (Marichyasana Modification)
Modification: both hands to the foot, heel down
Pose Type: standing one-legged balance
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

Drishti Point: Nasagrai or Nasagre (nose) or Padhayoragrai or Padayoragre (toes/feet)

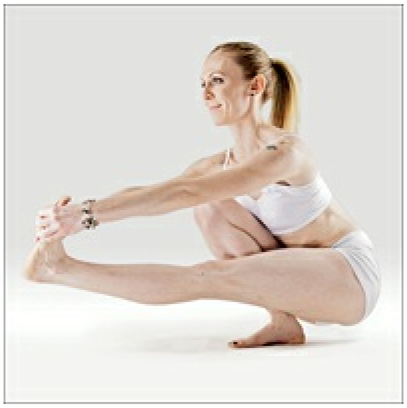

Revolved Supported Hand to Foot Pose Squatting

Parivritta Salamba Hasta Padasana

(puh-ri-VRIT-tuh SAH-luhm-buh HUH-stuh puh-DAHS-uh-nuh)

Modification: one hand to the floor
Pose Type: standing one-legged balance, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

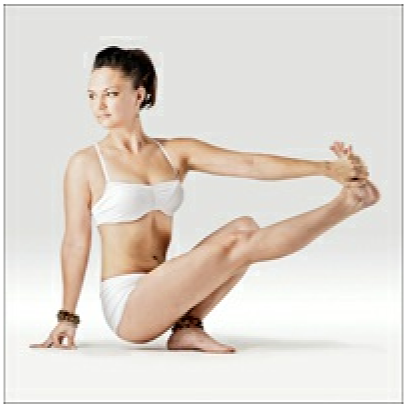

Hand to Foot to the Side Tip Toe Pose

Hasta Pada Parshva Prapadasana

(HUH-stuh PUH-duh PAHRSH-vuh pruh-puh-DAHS-uh-nuh)

Pose Type: standing one-legged balance
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

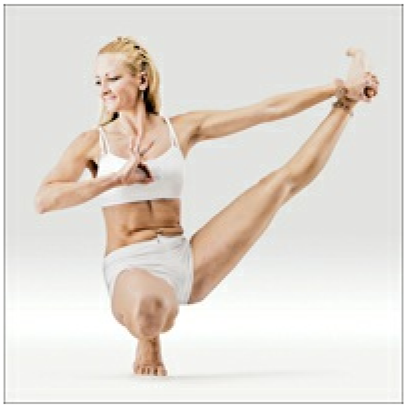

Sundial Pose in One-Legged Garland Pose

Surya Yantrasana in Eka Pada Malasana

(SOOR-yuh yuhn-TRAHS-uh-nuh in EY-kuh PUH-duh mah-LAHS-uh-nuh)

Also Known As: Sundial Pose in One-Legged Yoga Squat (Surya Yantrasana in Eka Pada Upavesasana)
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

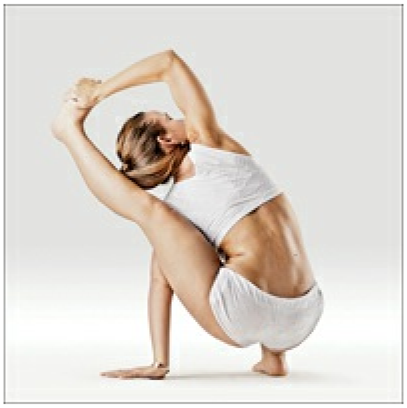

Extended Hand to Big Toe Pose in Half Tip Toe Hero Pose

Utthita Hasta Padangushtasana in Ardha Prapada Virasana

(UT-ti-tuh HUH-stuh puhd-ahng-goosh-TAHS-uh-nuh in UHR-duh PRUH-puh-duh veer-AHS-uh-nuh)

Pose Type: standing one-legged balance, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

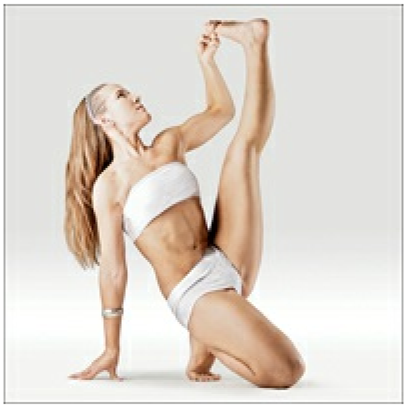

ONE-LEGGED SQUAT: ONE LEG STRAIGHT—HEEL DOWN

Baby Grasshopper Pose

Bala Shalabhasana

(BUH-luh shuh-luh-BAHS-uh-nuh)

Also Known As: Baby Locust Pose
Modification: chest facing the floor, forearm on the floor
Pose Type: standing one-legged balance, forward bend, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व

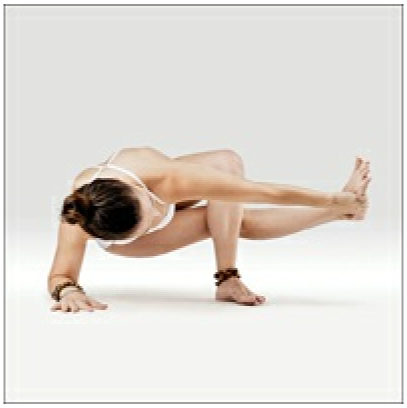

Hand to Foot Pose 2

Hasta Padasana 2

(HUH-stuh puh-DAHS-uh-nuh)

Modification: leg crossed over, one hand to the foot, other hand to the heart
Pose Type: standing one-legged balance, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Angushtamadhye or Angushta Ma Dyai (thumbs)

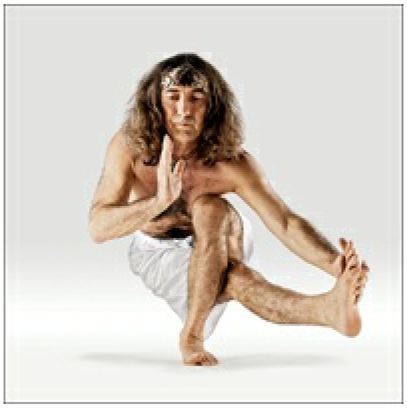

ONE-LEGGED SQUAT: LEG BEHIND THE HEAD

Supported One Foot Behind the Head Tip Toe Pose Prep. 1

Salamba Eka Pada Shirsha Prapadasana Prep. 1

(SAH-luhm-buh EY-kuh PUH-duh SHEER-shuh pruh-puh-DAHS-uh-nuh)

Modification: heel up, sitting bones lifted
Pose Type: standing one-legged balance, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

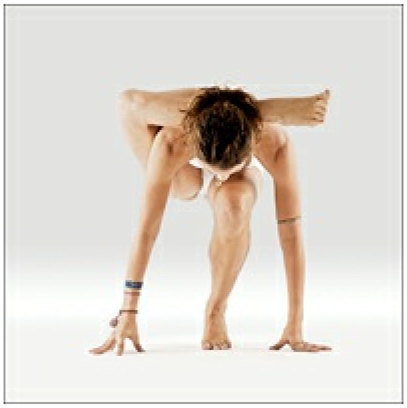

Supported One Foot Behind the Head Tip Toe Pose Prep. 2

Salamba Eka Pada Shirsha Prapadasana Prep. 2

(SAH-luhm-buh EY-kuh PUH-duh SHEER-shuh pruh-puh-DAHS-uh-nuh)

Modification: hands to the floor
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)

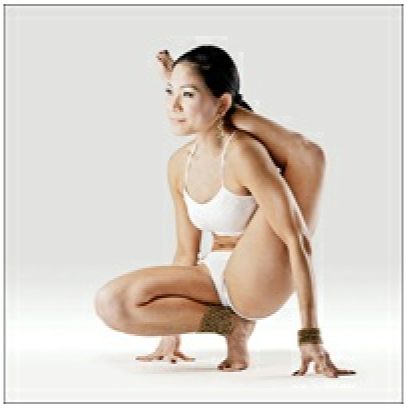

One Foot Behind the Head Tip Toe Pose

Eka Pada Shirsha Prapadasana

(EY-kuh PUH-duh SHEER-shuh pruh-puh-DAHS-uh-nuh)

Also Known As: Tip Toe Balance Foot Behind The Head Pose (Padangushta Eka Pada Shirshasana), One Foot Behind the Head Tip Toe Pose (Eka Pada Shirsha Padangushtasana)
Pose Type: standing one-legged balance
Drishti Point: Nasagrai or Nasagre (nose)

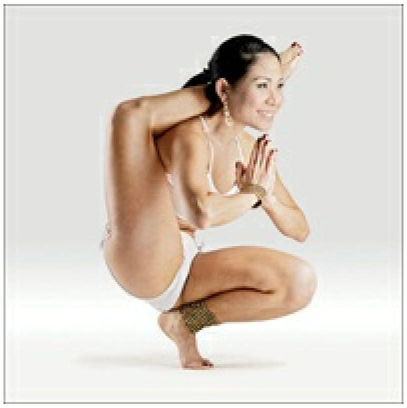

SQUAT: HEELS DOWN—KNEES WIDE

Hand Position of the Pose Dedicated to Garuda in Garland Pose

Hasta Garudasana in Malasana

(HUH-stuh guh-ru-DAHS-uh-nuh in mah-LAHS-uh-nuh)

Also Known As: Hand Position of the Pose Dedicated to Garuda in Yoga Squat (Hasta Garudasana in Upaveshasana)
Pose Type: standing, forward bend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

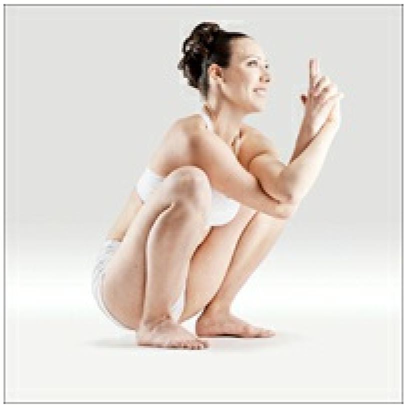

Lion Pose Dedicated to an Avatar of Lord Vishnu in Garland Pose

Narasimhasana in Malasana

(nuh-ruh-sim-HAHS-uh-nuh in mah-LAHS-uh-nuh)

Also Known As: Lion Pose in Yoga Squat (Simhasana in Upaveshasana)
Modification: heels down
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

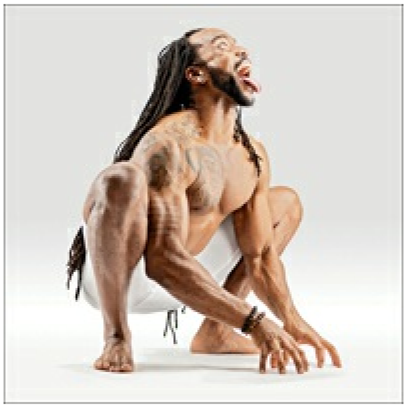

Half Bound Revolved Garland Pose

Ardha Baddha Parivritta Malasana

(UHR-duh BUH-duh puh-ri-VRIT-tuh mah-LAHS-uh-nuh)

Also Known As: Half Bound Revolved Yoga Squat (Ardha Baddha Parivritta Upaveshasana)
Modification: forearm on the knee
Pose Type: standing, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

Bound Garland Pose

Baddha Malasana

(BUH-duh mah-LAHS-uh-nuh)

Also Known As: Bound Yoga Squat (Baddha Upaveshasana)
Pose Type: standing, binding
Drishti Point: Nasagrai or Nasagre (nose)

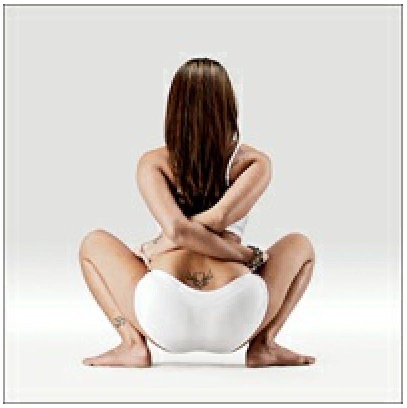

One Leg Bound Garland Pose

Eka Pada Baddha Malasana

(EY-kuh PUH-duh BUH-duh mah-LAHS-uh-nuh)

Also Known As: One Leg Bound Yoga Squat (Eka Pada Baddha Upaveshasana)
Modification: heels down, binding around one of the legs, knees open wide
Pose Type: standing, binding, forward bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

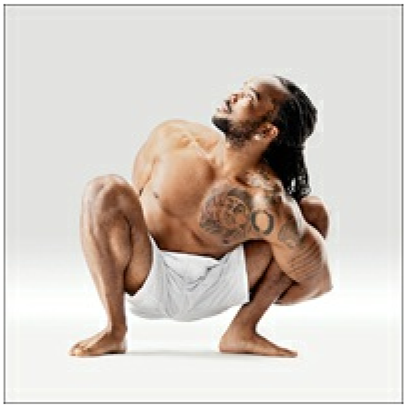

SQUAT: ONE HEEL UP, ONE HEEL DOWN—KNEES WIDE

Uneven Legs Garland Pose

Vishama Pada Malasana

(VISH-uh-muh PUH-duh mah-LAHS-uh-nuh)

Also Known As: Uneven Legs Yoga Squat (Vishama Pada Upaveshasana)
Modification: one knee higher than the other; fingertips of one hand to the floor, other elbow on the knee
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands)
● ●

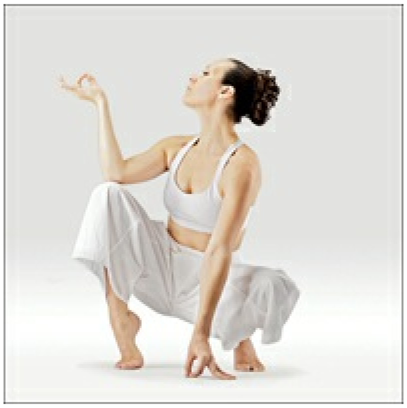

Sideways Uneven Legs Garland Pose

Parshva Vishama Pada Malasana

(PAHRSH-vuh VISH-uh-muh PUH-duh mah-LAHS-uh-nuh)

Also Known As: Sideways Uneven Legs Yoga Squat (Parshva Vishama Pada Upaveshasana)
Modification: one heel up, one heel down, one arm up over the head into a side bend, back of the other hand on the floor
Pose Type: standing, forward bend, side bend

Drishti Point: Hastagrai or Hastagre (hands)

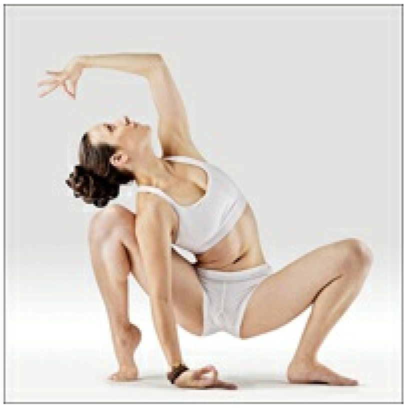

SQUAT: HEELS DOWN—KNEES WIDE, FORWARD BENDS

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Also Known As: Yoga Squat (Upaveshasana)
Modification: 1. hands in Anjali Mudra (Hands in Prayer)
2. arms straight out in front; palms down, head up
3. arms straight out in front; palms down, head down
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

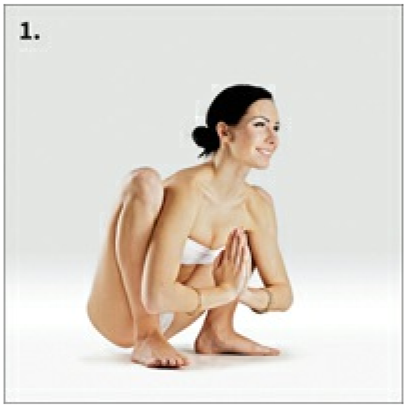

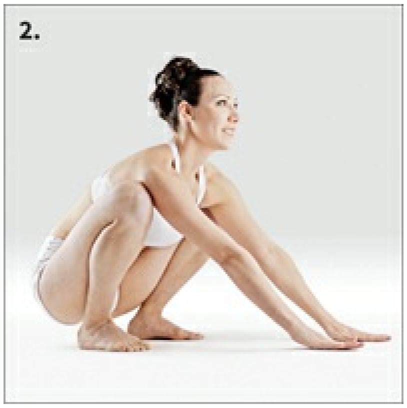

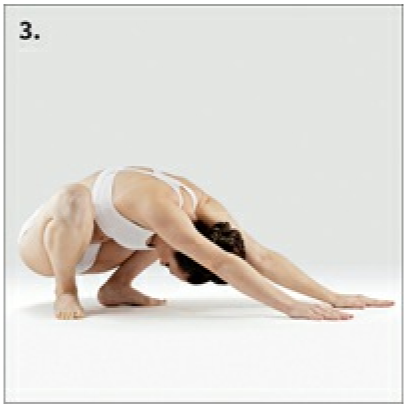

Hands Bound Garland Pose

Baddha Hasta Malasana

(BUH-duh HUH-stuh mah-LAHS-uh-nuh)

Also Known As: Hands Bound Yoga Squat (Baddha Hasta Upaveshasana)
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व य ह

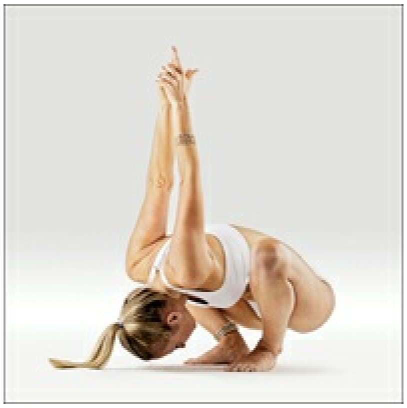

SQUAT: HEELS DOWN—KNEES WIDE, FORWARD BENDS—BINDING

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Also Known As: Tortoise Pose Squat (Kurmasana Modification)
Modification: grabbing onto heels
1. head off the floor
2. forehead to the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

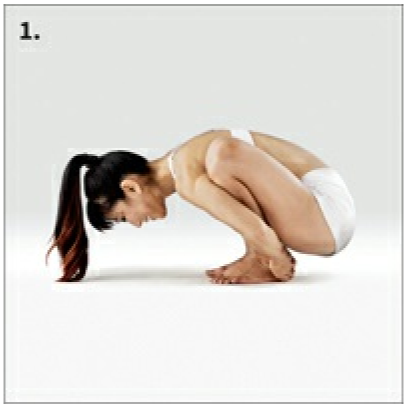
1.

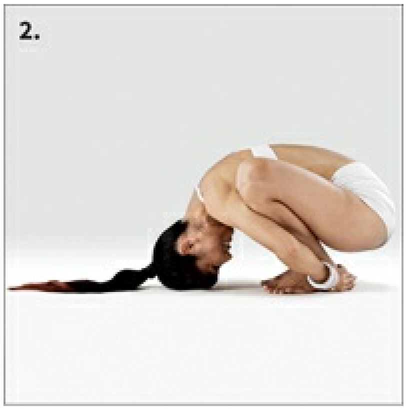
2.

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Also Known As: Golden Belt Pose (Kanchyasana)
Modification: hands bound, heels down
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

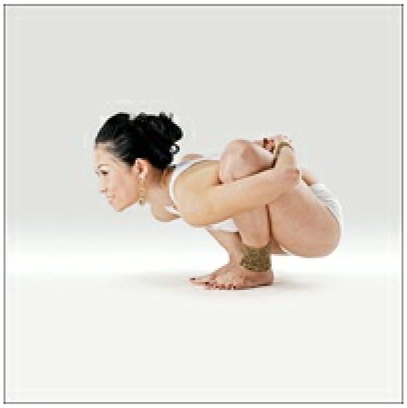

SQUAT: HEELS UP—KNEES WIDE, FORWARD BENDS—BINDING

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Modification: heels up, arms stretched out to the back, palms facing up, forehead to the floor
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

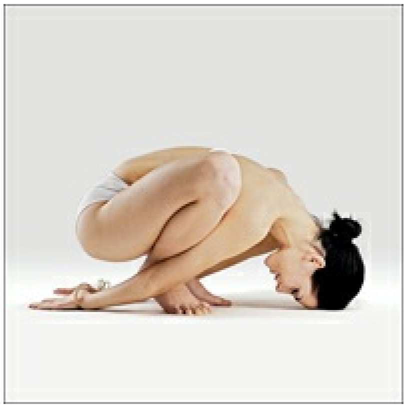

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Also Known As: Golden Belt Pose (Kanchyasana)
Modification: heels up, hands bound
1. front view
2. side view
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

1.

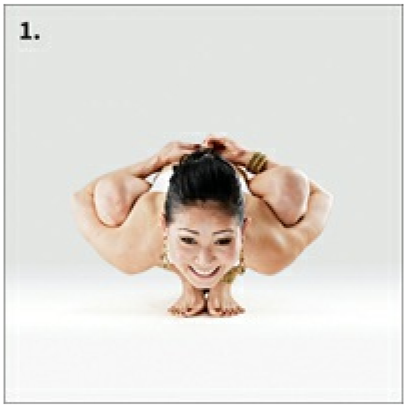

2.

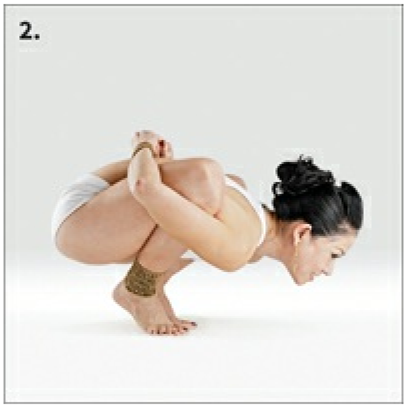

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Modification: heels up, arms binding around the shins, hands between thighs and rib cage
Pose Type: standing, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

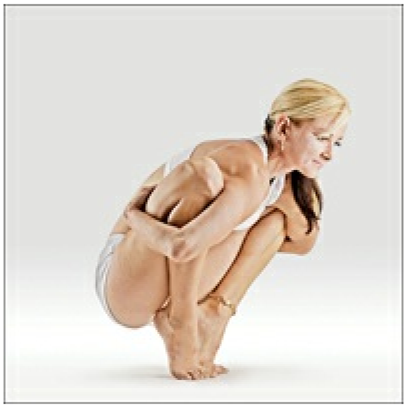

SQUAT: HEELS UP—KNEES WIDE, FORWARD BENDS—FOREARMS & ELBOWS ON THE FLOOR

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Modification: forehead to the floor, thumbs to the upper back
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

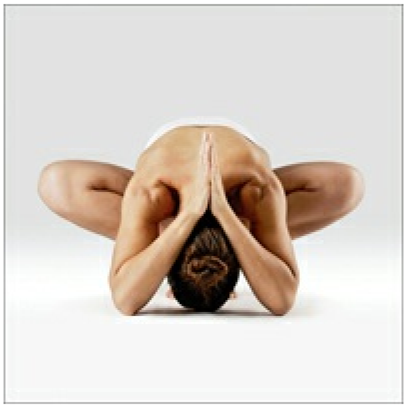

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Modification: forearms to the floor
1. grabbing onto the triceps
2. forearms on the floor, palms pressed together
Pose Type: standing, forward bend
Drishti Point: 1. Nasagrai or Nasagre (nose)
2. Angusthamadhye or Angustha Ma Dyai (thumbs)

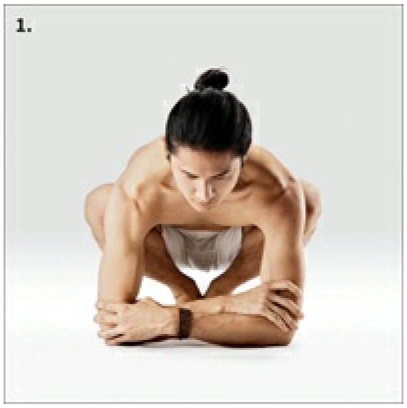
1.

2.

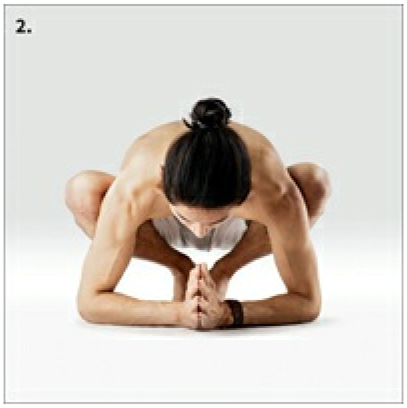

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Pose Type: standing, forward bend

Drishti Point: 1 & 2. Angusthamadhye or Angustha Ma Dyai (thumbs) 3 & 4. Nasagrai or Nasagre (nose)

How to Perform the Pose:

1. Begin by standing in Mountain Pose (*Tadasana*). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale as you bend your knees, lift your heels, and drop your sitting bones to your heels.
3. Inhale as you open your knees out to the sides. Exhale as you drop your chest between your knees and lower your forearms to the floor. Your knees should end up on the outsides of your shoulders.
4. You can experiment with various arm positions. Start with your palms flat on the floor (Pose #1), then bring either one hand to your face (Pose #2) or both hands to your face (Pose #3) and then spread your fingers out (Pose #4).
5. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
6. To come out of the pose, inhale as you lift your chest and bring your knees together. Exhale as you press into your feet, straighten your legs, and come back to Mountain Pose (*Tadasana*).

Modification:

1. forearms and palms to the floor
2. one forearm and palm to the floor, other elbow to the floor hand to the face
3. both hands to the face, elbows on the floor
4. both hands to the face, fingers spread open in a Padma Mudra - Lotus Hand Seal elbows, on the floor

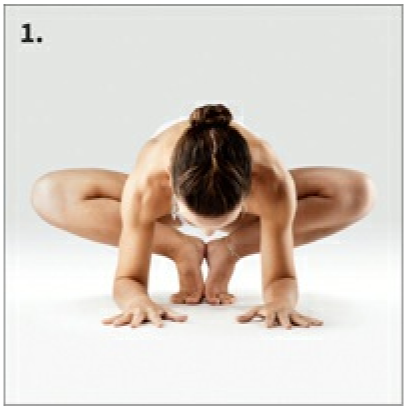

2.

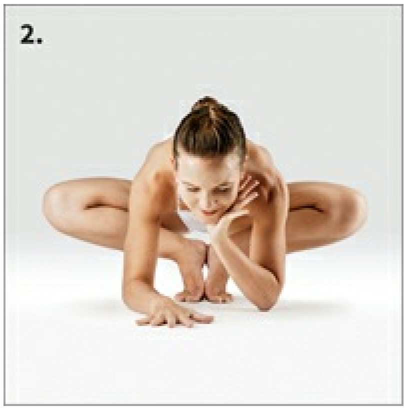

3.

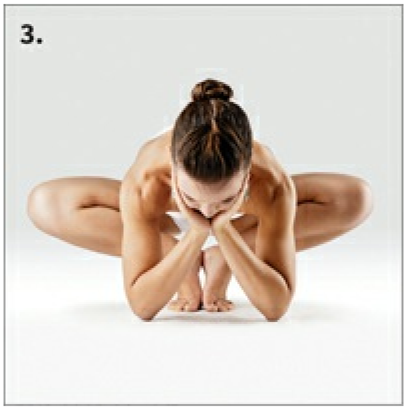

4.

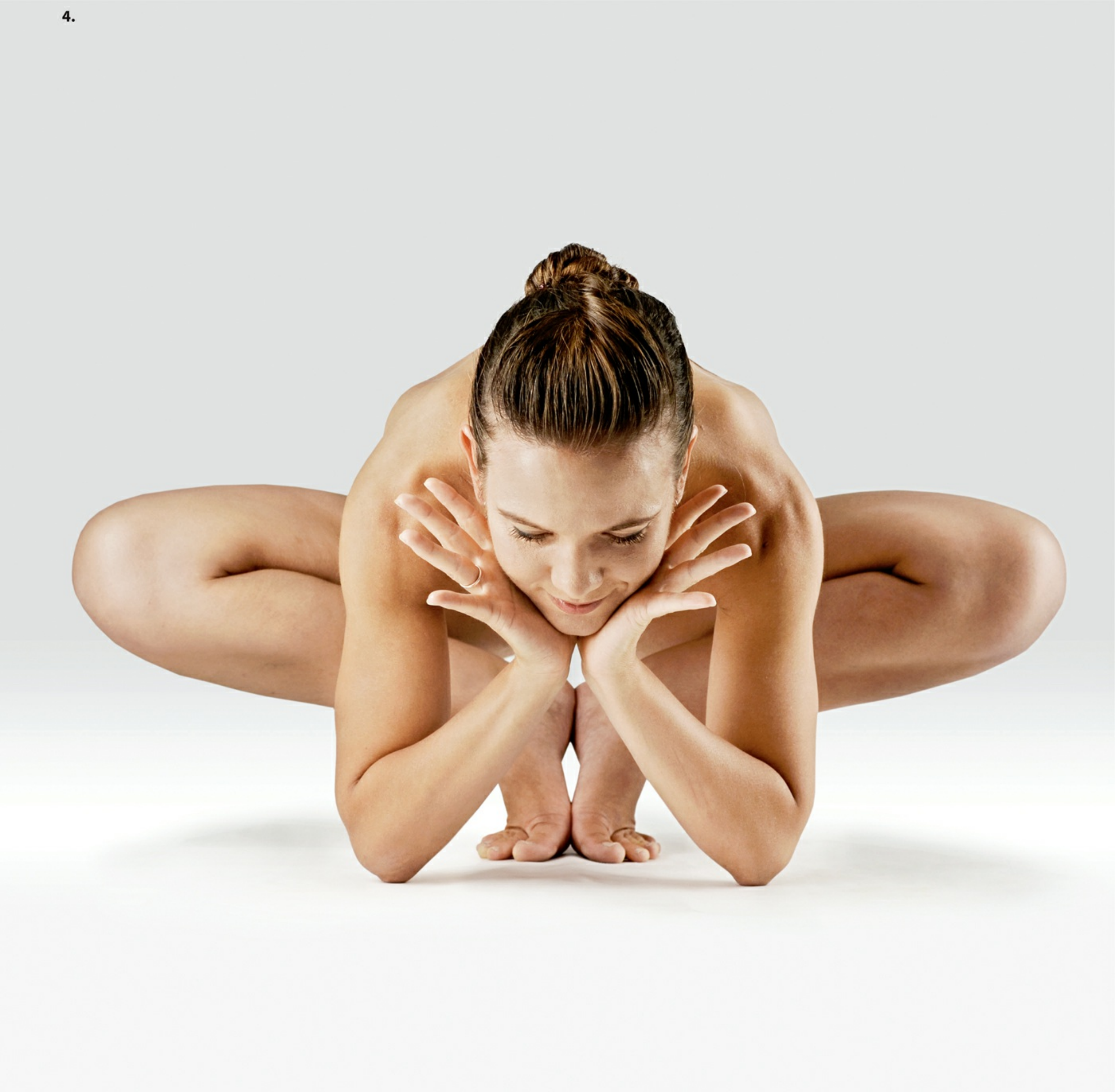

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Also Known As: Full Squat Pose (Kunthasana)

Modification: hands wide, knees on top of the triceps

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

ल व य ह

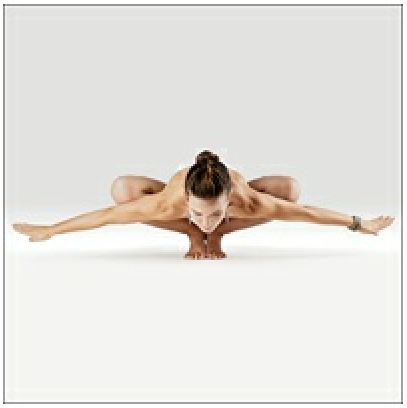

Garland Pose

Malasana

(mah-LAHS-uh-nuh)

Modification: heels up

1. knees to the armpits; hands out to the sides, fingertips to the floor
2. forehead to the floor; elbows bent at 90 degrees, fingertips to the floor
3. forehead to the floor; hands off the floor, arms out to the sides

Pose Type: standing, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

ल व य ह ॐ

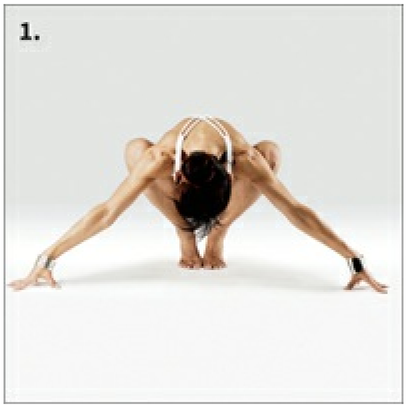
1.

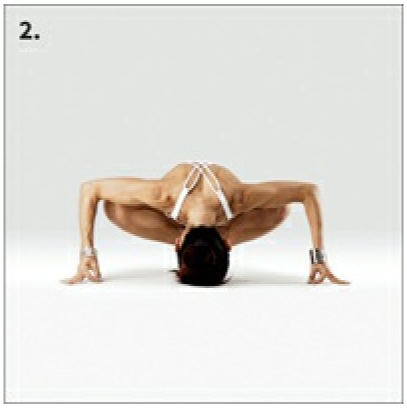
2.

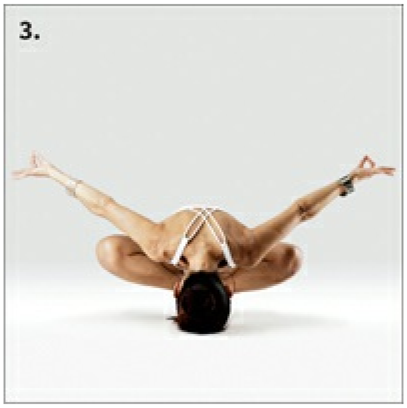
3.

SQUAT: HEELS UP—KNEES WIDE, ARMS TO THE FLOOR & BALLET TOES

Supported Tip Toe Pose

Salamba Prapadasana

(SAH-luhm-buh pruh-puh-DAHS-uh-nuh)

Modification: knees open, elbows and wrists together, fingertips touching the floor in front of the feet
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

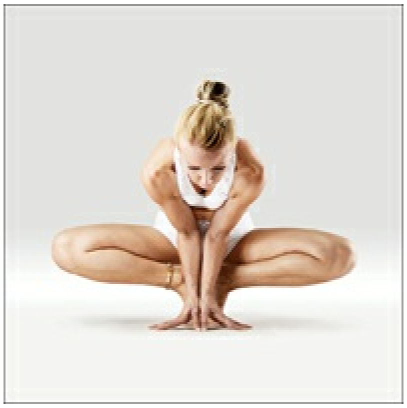

Intense Wrist Stretch Supported Tip Toe Pose

Uttana Manibandha Salamba Prapadasana

(ut-TAH-nuh muh-ni-BUHN-duh SAH-luhm-buh pruh-puh-DAHS-uh-nuh)

Modification: knees open wide; palms on the floor, fingertips toward the feet
Pose Type: standing, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

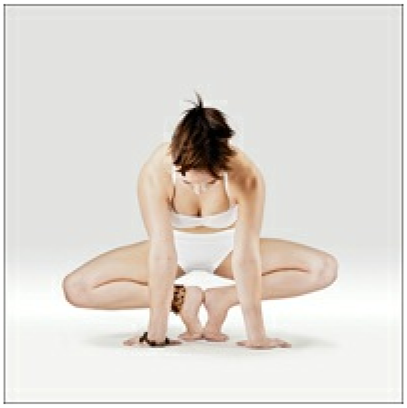

Lion Pose Dedicated to an Avatar of Lord Vishnu in Tip Toe Pose

Narasimhasana in Prapadasana

(nuh-ruh-sim-HAHS-uh-nuh in pruh-puh-DAHS-uh-nuh)

Also Known As: Lion Pose in Tip Toe Pose (Simhasana in Prapadasana)
Modification: feet wide apart
Pose Type: standing, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

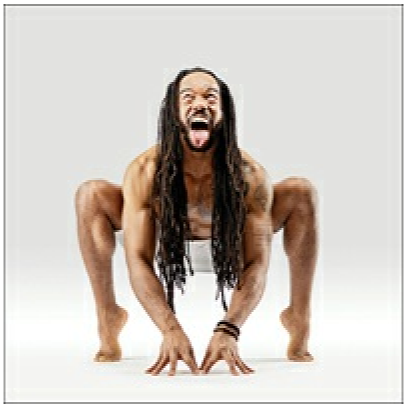

Unsupported Intense Ankle Stretch Tip Toe Pose

Niralamba Uttana Kulpa Prapadasana

(nir-AH-luhm-buh ut-TAH-nuh KUL-puh pruh-puh-DAHS-uh-nuh)

Modification: hands on the knees
Pose Type: standing, balance
Drishti Point: Nasagrai or Nasagre (nose)
लं वं

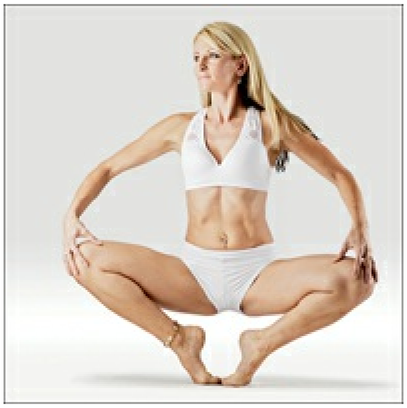

Supported Intense Ankle Stretch Tip Toe Pose

Salamba Uttana Kulpa Prapadasana

(SAH-luhm-buh ut-TAH-nuh KUL-puh pruh-puh-DAHS-uh-nuh)

Modification: fingertips to the floor
Pose Type: standing, balance
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

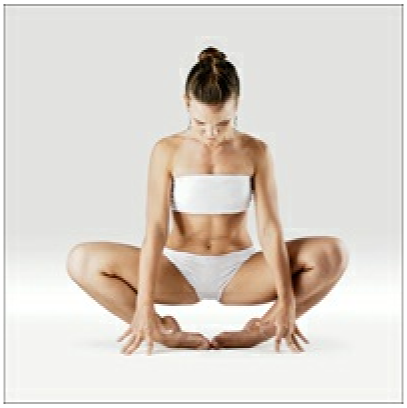

SQUAT: HEELS UP—KNEES WIDE, ARM MODIFICATIONS

Tip Toe Pose

Prapadasana

(pruh-puh-DAHS-uh-nuh)

Modification: knees open wide; palms together, elbows together
Pose Type: standing
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

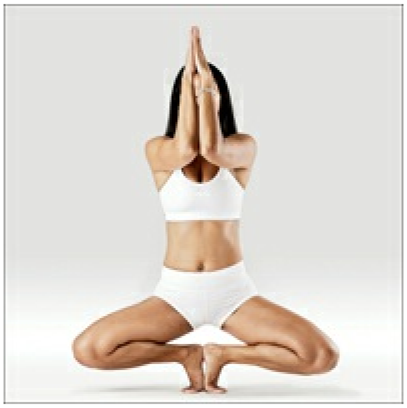

Tip Toe Pose

Prapadasana

(pruh-puh-DAHS-uh-nuh)

Modification: knees open wide, hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)

One Hand in Reverse Prayer Tip Toe Pose

Eka Hasta Viparita Namaskar Prapadasana

(EY-kuh HUH-stuh vi-puh-REE-tuh nuh-muhs-KAHR pruh-puh-DAHS-uh-nuh)

Modification: knees open wide; one hand to the knee
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)
ल व

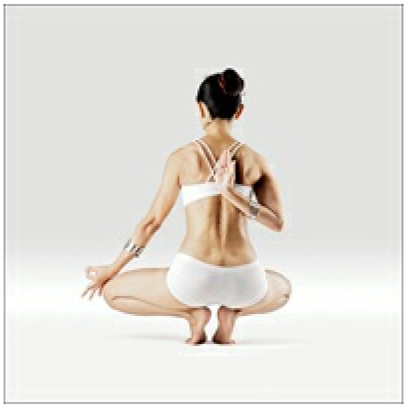

Upward Hands Tip Toe Pose

Urdhva Hasta Prapadasana

(OORD-vuh HUH-stuh pruh-puh-DAHS-uh-nuh)

Modification: knees open wide; arms extended up to the sky, head rolling back
Pose Type: standing, mild backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व य

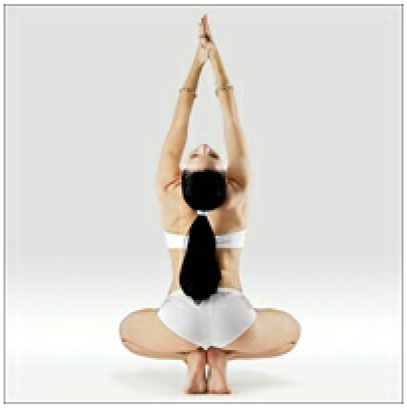

SQUAT: HEELS UP—KNEES WIDE, KNEES ON THE FLOOR—ARM MODIFICATIONS

Reverse Prayer Root Lock Pose Prep.

Viparita Namaskar Mulabhandasana Prep.

(vi-puh-REE-tuh nuh-muhs-KAHR moo-luh-buhn-DAHS-uh-nuh)

Also Known As: Back of the Body Prayer Tip Toe Pose (Paschima Namaskara Prapadasana)
Modification: knees open wide; knees to the floor
Pose Type: standing
Drishti Point: Nasagrai or Nasagre (nose)
ल व

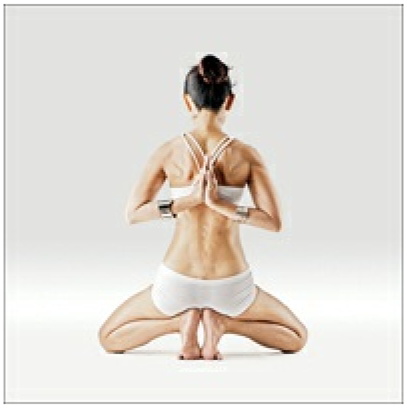

Root Lock Pose Prep.

Mulabhandasana Prep.

(moo-luh-buhn-DAHS-uh-nuh)

Also Known As: Tip Toe Pose Modification (Prapadasana Modification)
Modification: knees open wide; palms and elbows together; knees to the floor
Pose Type: standing
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

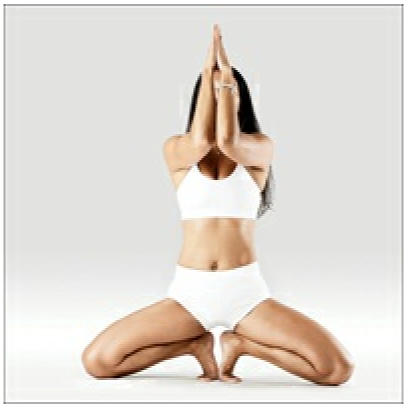

Root Lock Pose Prep.

Mulabhandasana Prep.

(moo-luh-buhn-DAHS-uh-nuh)

Also Known As: Tip Toe Pose Modification (Prapadasana Modification)
Modification: knees open wide; elbows and ankles crossed; knees to the floor, fingertips to the floor
Pose Type: standing, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

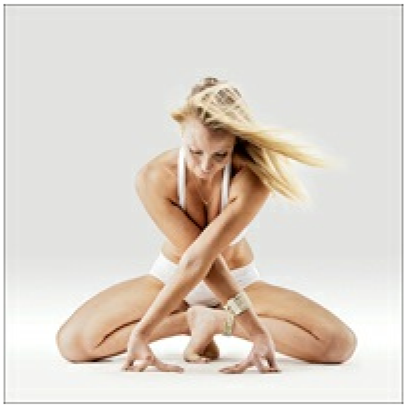

ROOT LOCK POSE

Root Lock Pose Prep.

Mulabhandasana Prep.

(moo-luh-buhn-DAHS-uh-nuh)

Also Known As: Twisted Feet Bound Angle Pose (Parivritta Pada Baddha Konasana)
Modification: fingertips to the floor behind the hips, sitting bones off the floor
Pose Type: standing, mild backbend

Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व ह

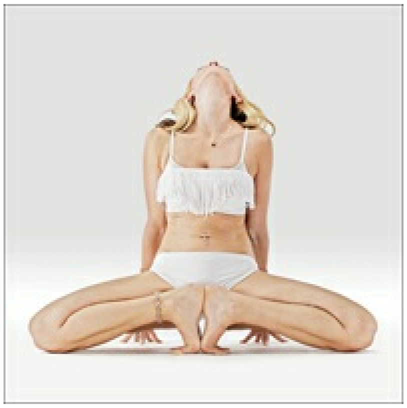

Womb Staff Pose

Yoni Dandasana

(YO-ni duhn-DAHS-uh-nuh)
Also Known As: Root Lock Pose Modification (Mulabandhasana Modification)
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)
ल व

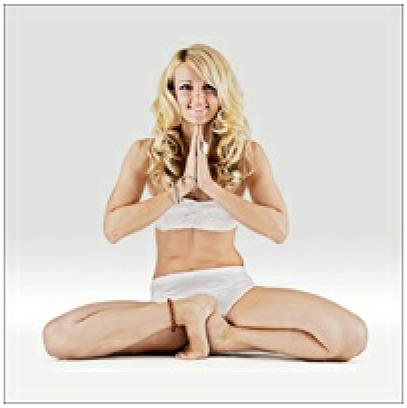

Root Lock Pose

Mulabhandasana

(moo-luh-buhn-DAHS-uh-nuh)
Also Known As: Perineal Contraction Pose (Mula Bandhasana)
Modification: palms up on the knees, head rolled back
Pose Type: seated, mild backbend
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व ह

Pose Dedicated to Sage Vamadeva 1

Vamadevasana 1

(vah-muh-dey-VAHS-uh-nuh)

Modification: both hands binding to the top foot
Pose Type: seated, twist, binding
Drishti Point: Parsva Drishti (to the right), Parsva Drishti (to the left)

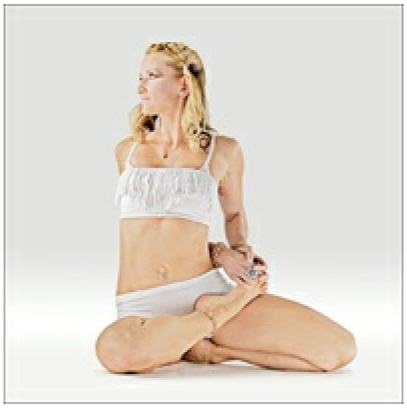

Seated Poses

EASY POSE

ACCOMPLISHED ONE POSE

FIRE LOG POSE

HALF LOTUS AND LOTUS POSE

BOTH KNEES BENT

BOUND ANGLE POSE

STAR POSE

LEGS IN FRONT

HORIZONTAL SPLITS

LEGS STRAIGHT IN FRONT

GARUDA LEGS

LEG CRADLE

HEAD TO KNEE

HERO POSE

THUNDERBOLT POSE

ONE KNEE BENT (VARIATIONS)

COW FACE POSE

SAGE VAMADEVA’S POSE

ONE-LEGGED KING PIGEON POSE

VERTICAL SPLITS

Easy Pose

Sukhasana

(suk-AHS-uh-nuh)

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

How to Perform the Pose:

1. Begin by sitting on the floor with both your legs straight out in front of you. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale as you bring your feet toward you and cross them at the ankles. Keep the soles of your feet flat on the floor and your knees in line with your shoulders.
3. Inhale, stretch your arms up to the sky to lengthen your spine, and exhale as you bring your elbows together on the inside of your knees.
4. Inhale as you open your palms up to sky with your fingertips pointing to the outsides. Exhale as you bring you thumbs and pointer fingers together.
5. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
6. Exhale as you release the pose, coming back to starting position with both your legs straight out.

Modification: knees in line with the shoulders, elbows to the inside of the knees

sukha = easy, lightness

Easy Pose

Sukhasana

(suk-AHS-uh-nuh)

Modification: knees to the chest, elbows to the outside of the knees, palms covering the face
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं हं ॐ

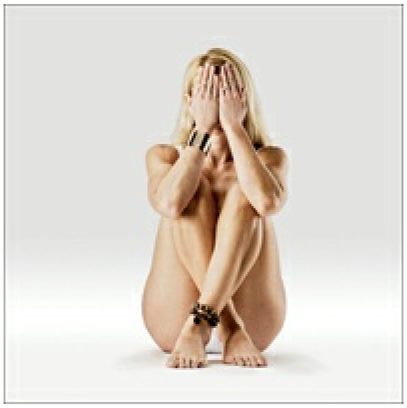

Easy Pose

Sukhasana

(suk-AHS-uh-nuh)

Modification: knees high, forearms to the shins
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं

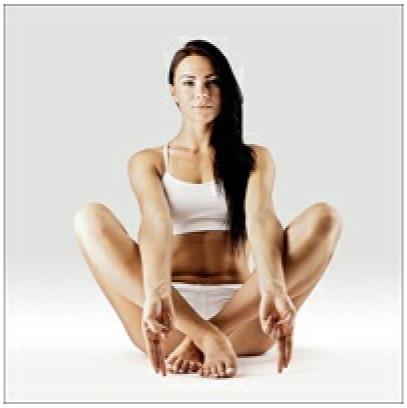

Easy Pose

Sukhasana

(suk-AHS-uh-nuh)

Modification: elbows crossed, backs of both hands to the knees

Pose Type: seated, mild backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

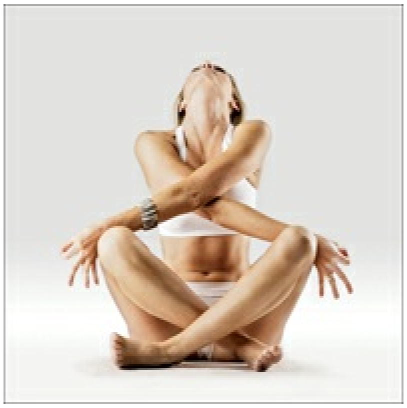

EASY POSE: SPINE STRAIGHT

Easy Pose

Sukhasana

(suk-AHS-uh-nuh)

Modification: on the yoga block

Pose Type: seated

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

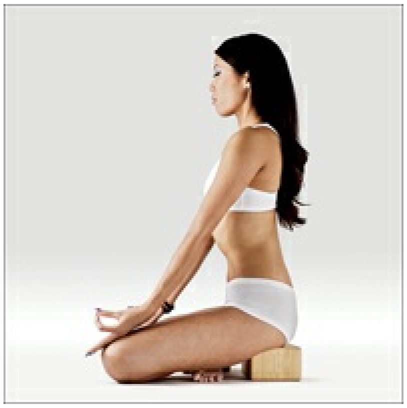

Easy Pose

Sukhasana

(suk-AHS-uh-nuh)

Pose Type: seated

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व

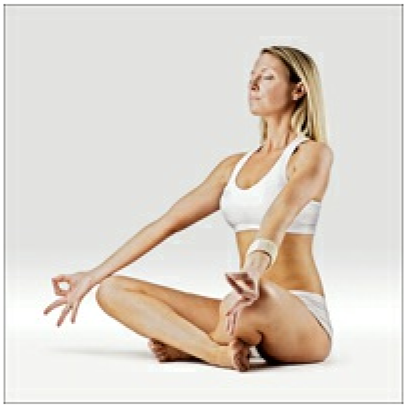

Bound Hands Easy Pose

Baddha Hasta Sukhasana

(BUH-duh HUH-stuh suk-AHS-uh-nuh)

Modification: arms in front, arms straight

Pose Type: seated

Drishti Point: Hastagrai or Hastagre (hands), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व

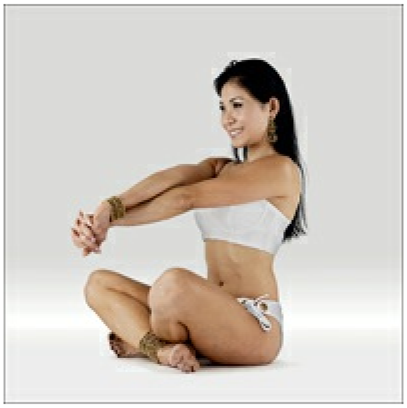

Embryo in the Womb Pose in Easy Pose

Garbha Pindasana in Sukhasana

(GUHR-buh pin-DAHS-uh-nuh in suk-AHS-uh-nuh)

Modification: grabbing onto the outside edges of the feet

Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

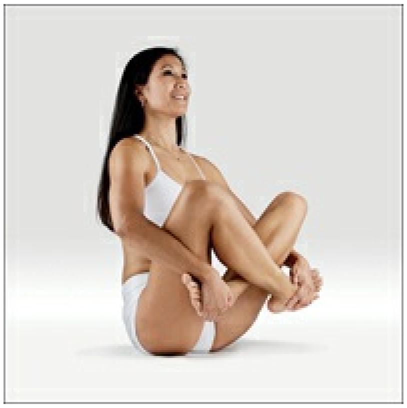

EASY POSE: TWISTS & SIDE BENDS

Easy Pose

Sukhasana

(suk-AHS-uh-nuh)
Modification: neck stretch
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

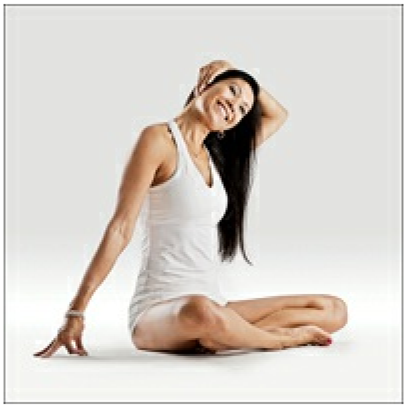

Revolved Easy Pose

Parivritta Sukhasana

(puh-ri-VRIT-tuh suk-AHS-uh-nuh)
Pose Type: seated, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

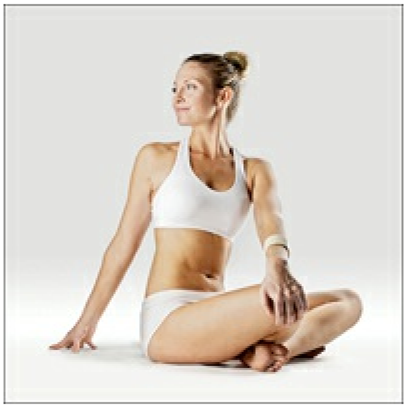

Sideways Easy Pose

Parshva Sukhasana

(PAHRSH-vuh suk-AHS-uh-nuh)

Modification: one forearm to the floor, other arm extended up over the head
Pose Type: seated, side bend
Drishti Point: Hastagrai or Hastagre (hands)

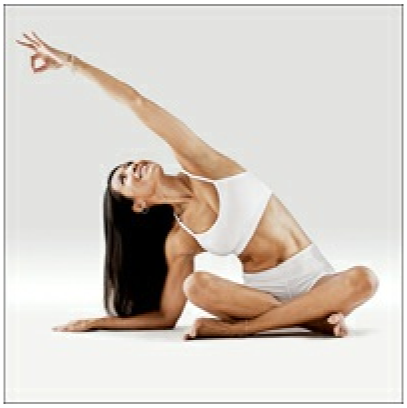

Revolved Bound Easy Pose

Parivritta Baddha Sukhasana

(puh-ri-VRIT-tuh BUH-duh suk-AHS-uh-nuh)

Modification: elbow to the floor
Pose Type: seated, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

ACCOMPLISHED ONE POSE

Accomplished One Pose

Siddhasana

(sid-DAHS-anna)

Modification: 1. one heel in front of the other
2. one foot tucked between the hamstring and the calf
Pose Type: seated
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

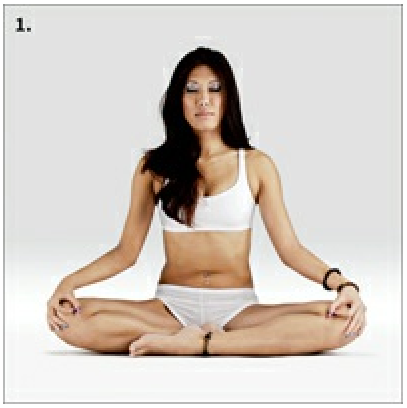

1.

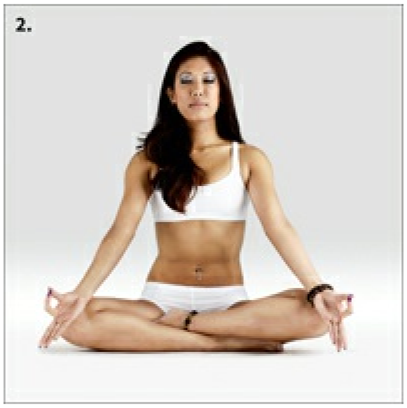

2.

Fire Log Pose

Agnistambhasana

(uhg-ni-stuhm-BAHS-anna)

Modification: 1. spine straight, fingertips to the floor behind the hips
2. leaning forward, palms on the floor by the thighs
Pose Type: 1. seated
2. seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

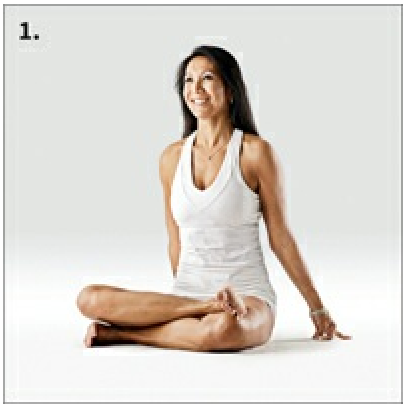
1.

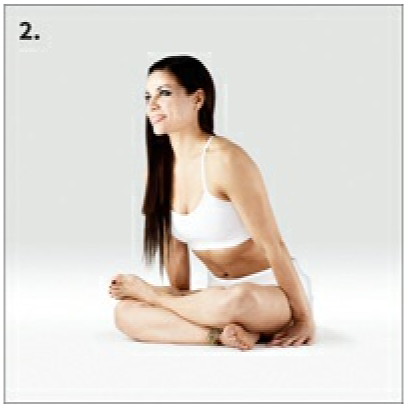
2.

Revolved Fire Log Pose

Parivritta Agnistambhasana

(puh-ri-VRIT-tuh uhg-ni-stuhm-BAHS-uh-nuh)

Modification: hands in Anjali Mudra (Hands in Prayer)
1. elbow to the knee
2. elbow to the foot
Pose Type: seated, forward bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky); Parshva Drishti (to the right), Parshva Drishti (to the left)

1.

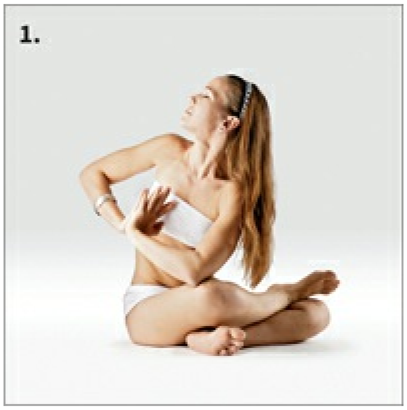

2.

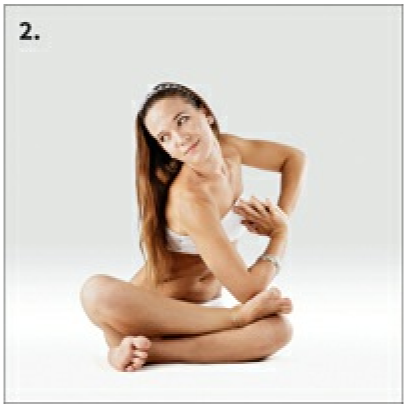

Revolved Bound Fire Log Pose

Parivritta Baddha Agnistambhasana

(puh-ri-VRIT-tuh BUH-duh uhg-ni-stuhm-BAHS-uh-nuh)

Modification: one elbow to the sole of the top foot, other arm behind the back, hand to the inside of the thigh
Pose Type: seated, forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

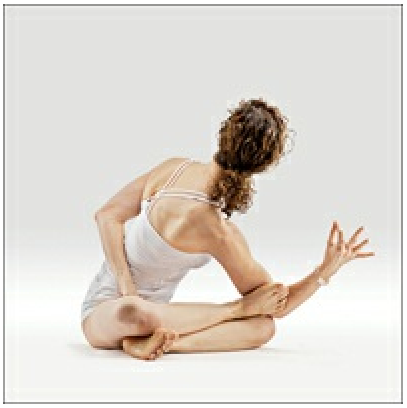

Revolved Fire Log Pose

Parivritta Agnistambhasana

(puh-ri-VRIT-tuh uhg-ni-stuhm-BAHS-uh-nuh)

Modification: forearms to the floor
Pose Type: seated, forward bend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

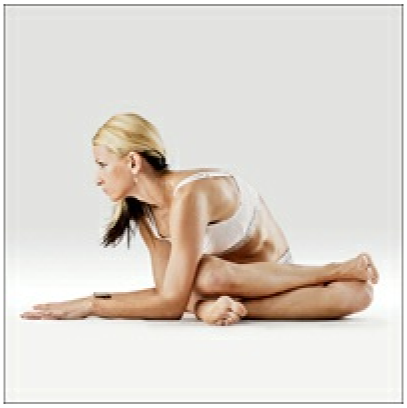

Fire Log Pose

Agnistambhasana

(uhg-ni-stuhm-BAHS-uh-nuh)

Modification: forward bend, palms together, fingers spread wide; elbows and forehead on the floor
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

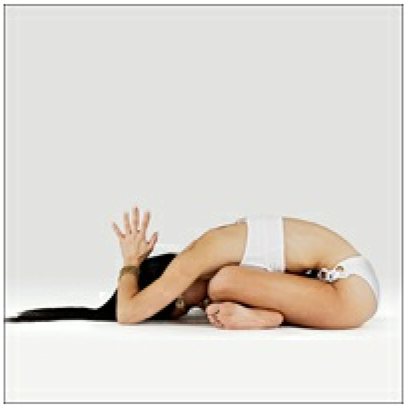

HALF LOTUS POSE

Half Lotus Pose

Ardha Padmasana

(UHR-duh puhd-MAHS-uh-nuh)

Modification: back of the hands on the knees

Pose Type: seated

Drishti Point: Nasagrai or Nasagre (nose)

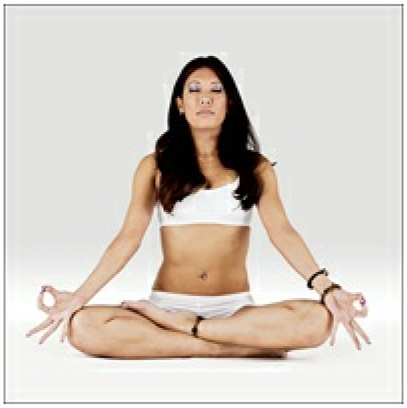

Half Lotus Pose

Ardha Padmasana

(UHR-duh puhd-MAHS-uh-nuh)

Modification: Salute to the Buddha Mudra Hand Position: one palm facing up, fingertips of the other hand to the floor

Pose Type: seated

Drishti Point: Nasagrai or Nasagre (nose)

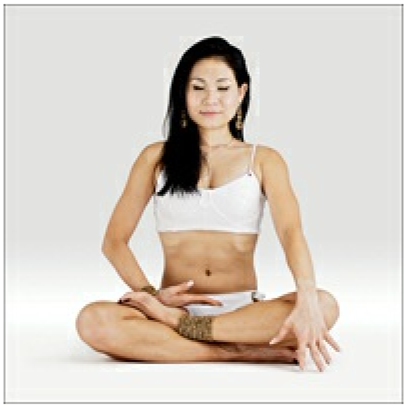

Reverse Prayer Revolved Half Lotus Pose

Viparita Namaskar Parivritta Ardha Padmasana

(vi-puh-REE-tuh nuh-muhs-KAHR puh-ri-VRIT-tuh UHR-duh puhd-MAHS-uh-nuh)

Also Known As: Back of the Body Prayer Revolved Half Lotus Pose (Paschima Namaskara Parivritta Ardha Padmasana)

Modification: top foot resting on the bottom knee

Pose Type: seated, twist

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

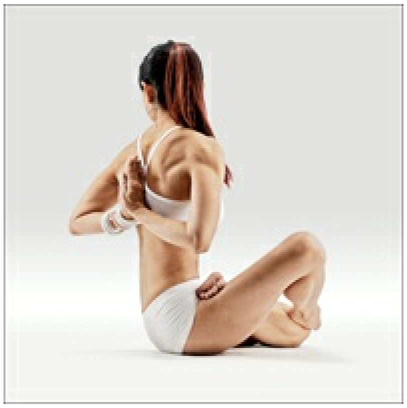

Embryo in the Womb Pose in Half Lotus Pose

Garbha Pindasana in Ardha Padmasana

(guhr-buh-pin-DAHS-uh-nuh in UHR-duh puhd-MAHS-uh-nuh)

Modification: knees lifted toward the chest

Pose Type: seated, forward bend, core

Drishti Point: Nasagrai or Nasagre (nose)

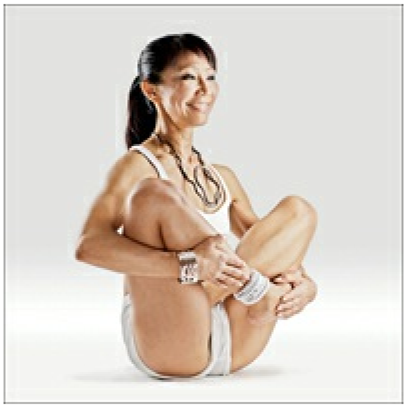

LOTUS POSE: ARMS IN FRONT

Lotus Pose

Padmasana

(puhd-MAHS-uh-nuh)

Modification: back of the hands on the knees
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

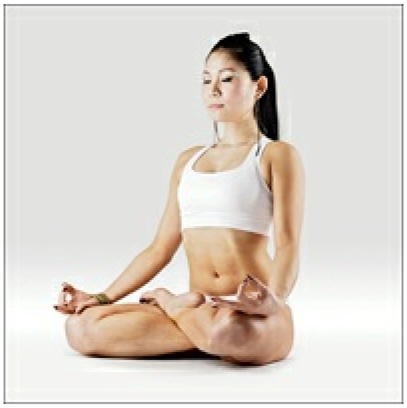

Lotus Pose

Padmasana

(puhd-MAHS-uh-nuh)

Modification: fingers interlocked, hands resting on the lap
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

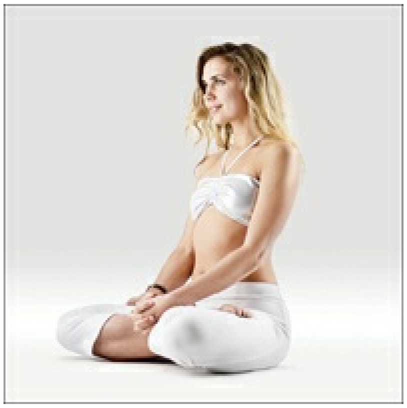

LOTUS POSE: ARMS BEHIND

Reverse Prayer Lotus Pose

Viparita Namaskar Padmasana

(vi-puh-REE-tuh nuh-muhs-KAHR puhd-MAHS-uh-nuh)

Also Known As: Back of the Body Prayer Lotus Pose (Paschima Namaskara Padmasana)
Modification: 1. fingertips pointing down
2. fingertips pointing up
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

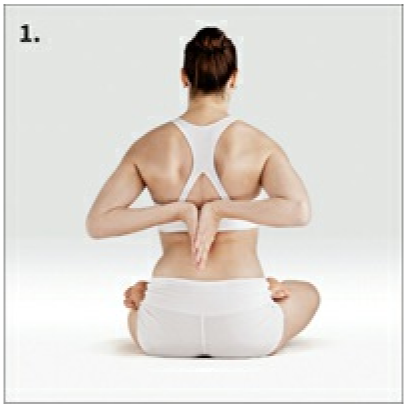
1.

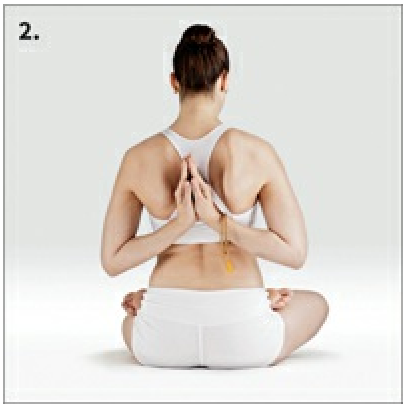
2.

Bound Lotus Pose

Baddha Padmasana

(BUH-duh puhd-MAHS-uh-nuh)

Modification: both hands to one foot, one arm behind the back
Pose Type: seated, mild backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

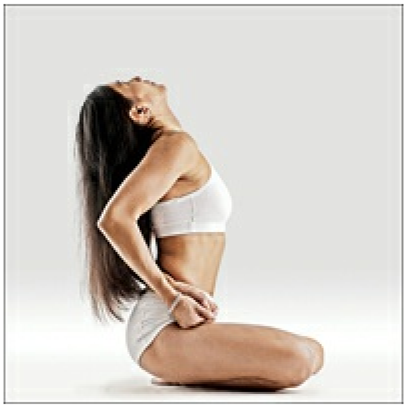

Bound Lotus Pose

Baddha Padmasana

(BUH-duh puhd-MAHS-uh-nuh)

Modification: both arms behind the back
Pose Type: seated, mild backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

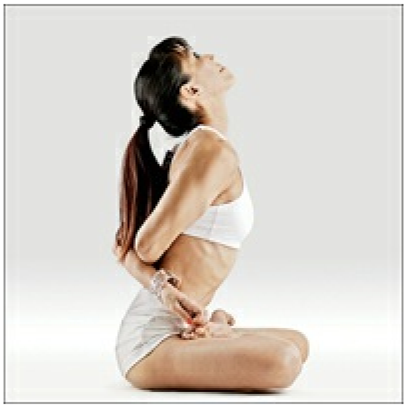

Hands Bound Lotus Pose

Baddha Hasta Padmasana

(BUH-duh HUH-stuh puhd-MAHS-uh-nuh)

Modification: backbend
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

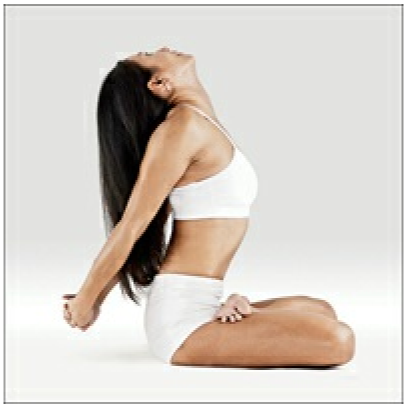

Lotus Pose

Padmasana

(puhd-MAHS-uh-nuh)

Modification: palms to the floor behind the hips; fingertips facing forward, backbend
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

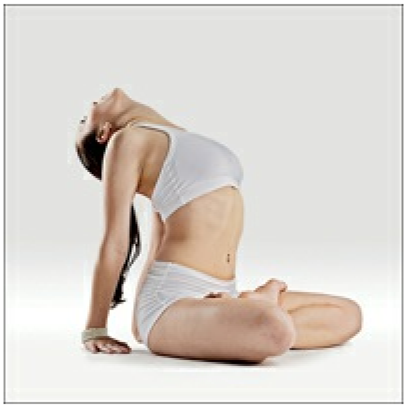

LOTUS POSE: FINGERS BOUND - BICEPS BY THE EARS

Raised Bound Hands Lotus Pose

Urdhva Baddha Hasta Padmasana

(OORD-vuh BUH-duh HUH-stuh puhd-MAHS-uh-nuh)

Also Known As: Seated Mountain Pose A (Parvatasana A)
Pose Type: seated, mild backbend
Drishti Point: Angushtamadhye (Thumbs)

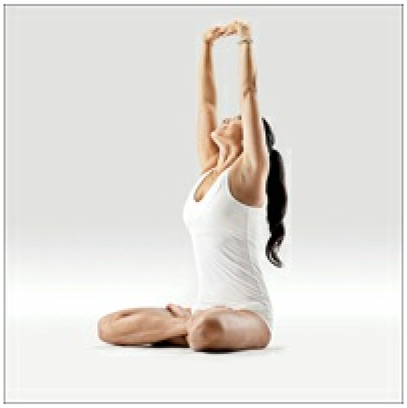

Simple Yoga Seal

Laghu Yoga Mudra

(LUH-gu YO-guh MU-druh)

Also Known As: Seated Mountain Pose B (Parvatasana B)
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

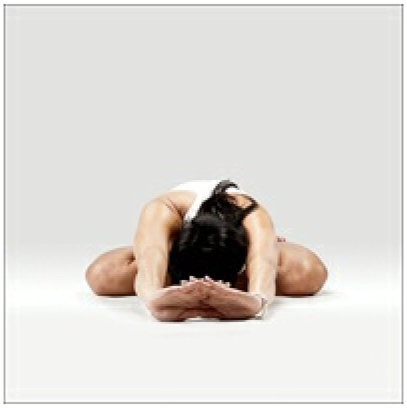

Sideways Simple Yoga Seal

Parshva Laghu Yoga Mudra

(PAHRSH-vuh LUH-gu YO-guh MU-druh)

Modification: fingers interlocked, palms facing out; arms straight to the side, forward bend
Pose Type: seated, forward bend, side bend
Drishti Point: Nasagrai or Nasagre (nose)

LOTUS POSE: TWISTS

Pose Dedicated to Siddhar Vaasamuni—Easy Version

Sukha Vaasamunvasana

(SUK-uh vah-sah-moo-NYAHS-uh-nuh)

Pose Type: seated, side bend, forward bend, twist

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

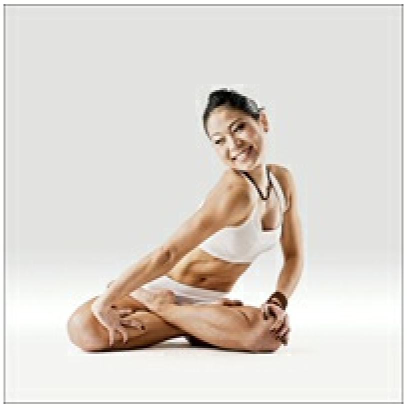

Arms of Twist Dedicated to Sage Bharadvaja 2 in Revolved Lotus Pose

Hasta Bharadvajasana 2 in Parivritta Padmasana

(HUH-stuh buh-ruhd-vuhj-AHS-uh-nuh in puh-ri-VRIT-uh puhd-MAHS-uh-nuh)

Pose Type: seated, twist, binding

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

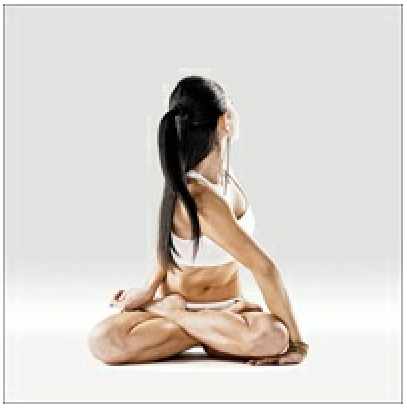

Revolved Lotus Pose

Parivritta Padmasana

(puh-ri-VRIT-uh puhd-MAHS-uh-nuh)

Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

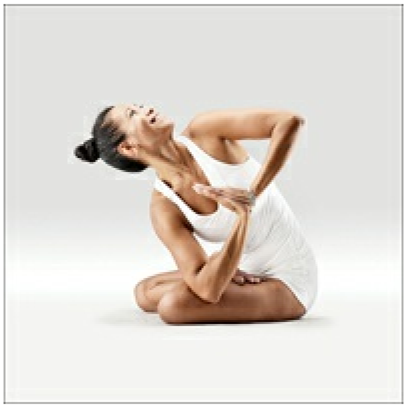

LOTUS POSE: SIDE BEND & TWISTS

Sideways Lotus Pose

Parshva Padmasana

(PAHRSH-vuh puhd-MAHS-uh-nuh)

Modification: forearm to the floor
Pose Type: seated, side bend
Drishti Point: Hastagrai or Hastagre (hands)

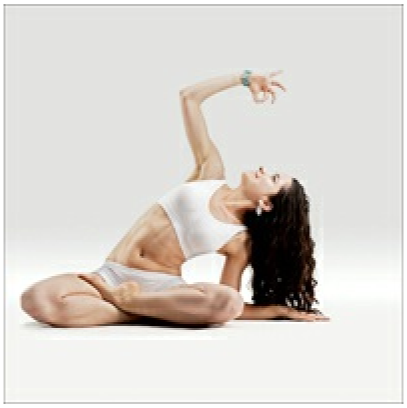

Revolved Western Intense Stretch Pose Lotus Pose

Parivritta Paschimottana Padmasana

(puh-ri-VRIT-uh puhsh-chi-mo-TAHN-nuh puhd-MAHS-uh-nuh)

Also Known As: Revolved Forward Bend Lotus Pose
Modification: shoulder to the floor, other arm extended to the sky
Pose Type: seated, forward bend, twist
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व

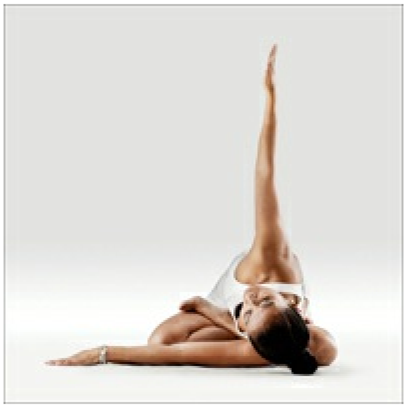

Pose Dedicated to Siddhar Vaasamuni—One Hand Modification

Eka Hasta Vaasamunvasana

(EY-kuh HUH-stuh [vah-sah-moo]-NYAHS-uh-nuh)

Pose Type: seated, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व ह

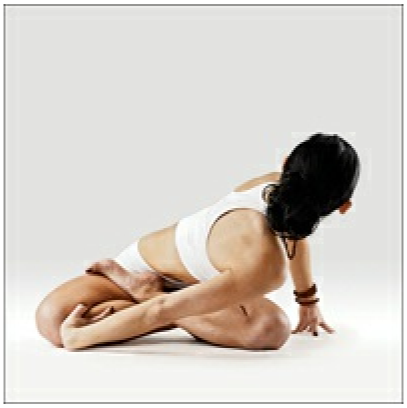

Pose Dedicated to Siddhar Vaasamuni

Vaasamunvasana

([vah-sah-moo-NYAHS-uh-nuh)

Pose Type: seated, forward bend, twist, binding

Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व य ह

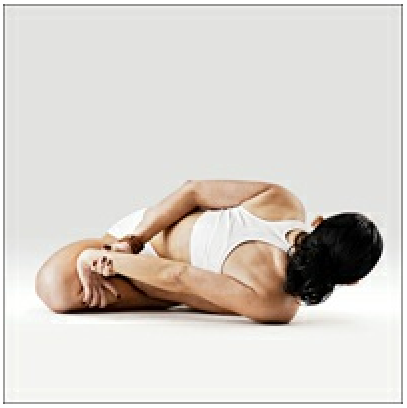

LOTUS POSE: FORWARD BEND

Yogic Seal Pose

Yoga Mudrasana

(YO-guh mu-DRAHS-uh-nuh)

Also Known As: Yoga Pose (Yogasana)

Modification: grabbing onto the wrist behind the back

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose) or Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

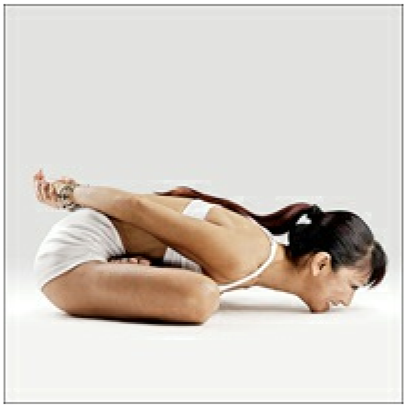

Hands Bound Lotus Pose

Baddha Hasta Padmasana

(BUH-duh HUH-stuh puhd-MAHS-uh-nuh)

Also Known As: Yoga Seal (Yoga Mudra)
Modification: chin to the floor, forward bend
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose) or Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

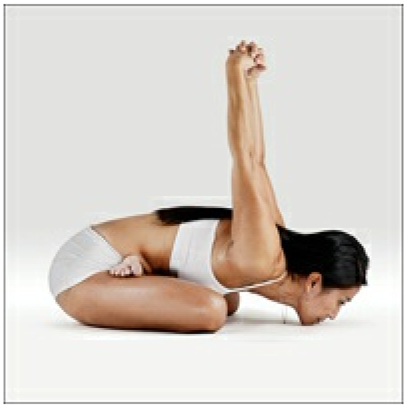

Reverse Prayer Yogic Seal Pose

Viparita Namaskar Yoga Mudrasana

(vi-puh-REE-tuh nuh-muhs-KAHR YO-guh mu-DRAHS-uh-nuh)

Also Known As: Back of the Body Prayer Yogic Seal Pose (Paschima Namaskara Yoga Mudrasana)
Modification: forehead to the floor
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र

Yogic Seal Pose Prep.

Yoga Mudrasana Prep.

(YO-guh mu-DRAHS-uh-nuh)

Modification: 1. using a yoga strap
2. wrists crossed behind the back, reaching toward the feet
Pose Type: seated, forward bend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

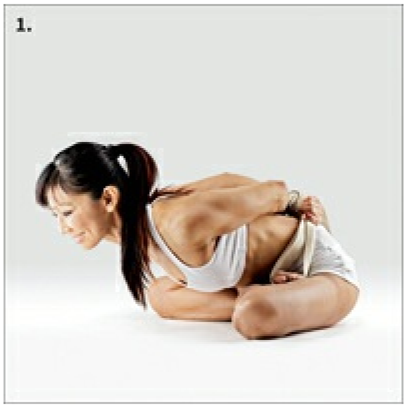

1.

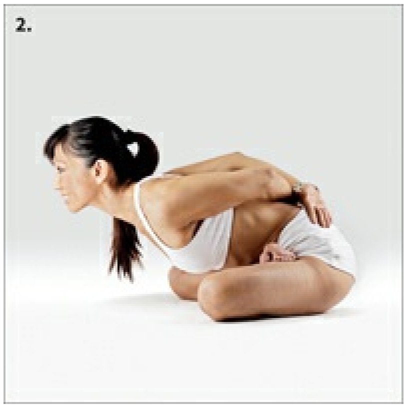

2.

Sideways Yogic Seal Pose

Parshva Yoga Mudrasana

(PARHSH-vuh YO-guh mu-DRAHS-uh-nuh)

Also Known As: Sideways Bound Lotus Pose (Parshva Baddha Padmasana)
Modification: chin to the knee
Pose Type: seated, forward bend, binding, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य

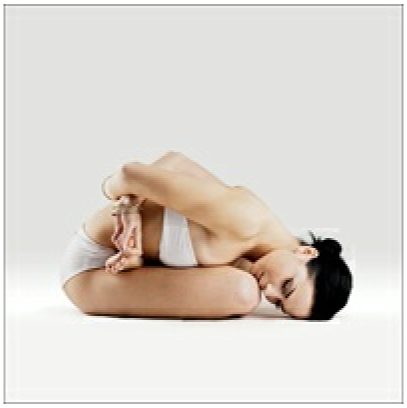

Yogic Seal Pose

Yoga Mudrasana

(YO-guh mu-DRAHS-uh-nuh)

Also Known As: Bound Lotus Pose (Baddha Padmasana)
Modification: 1. side view
2. front view
Pose Type: seated, forward bend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

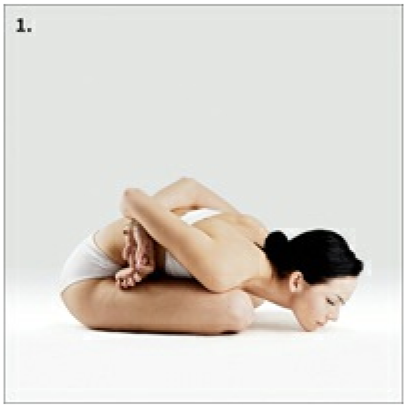

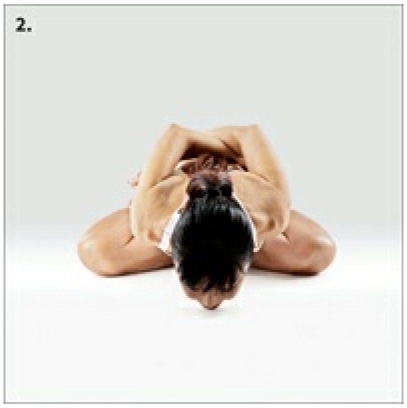

LOTUS POSE: EMBRYO IN THE WOMB POSE

Embryo in the Womb Pose

Garbha Pindasana

(guhr-buh pin-DAHS-uh-nuh)

Modification: arms wrapped around the legs
Pose Type: seated, forward bend, core
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

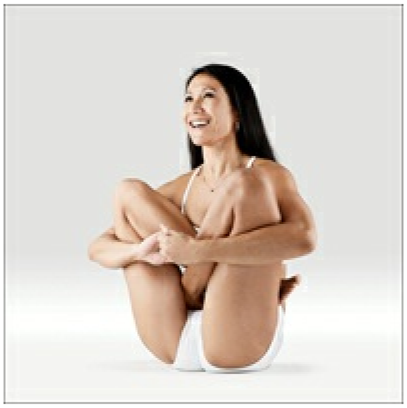

Hand Position of the Pose Dedicated to Garuda in Yoga Pose A

Hasta Garudasana in Yogasana A

(HUH-stuh guh-ru-DAHS-uh-nuh in yo-GAHS-uh-nuh)

Pose Type: seated, forward bend, core
Drishti Point: Angushtamadhye or Angustha Ma Dyai (thumbs)

Embryo in the Womb Pose

Garbha Pindasana

(guhr-buh pin-DAHS-uh-nuh)

Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Nasagrai or Nasagre (nose)

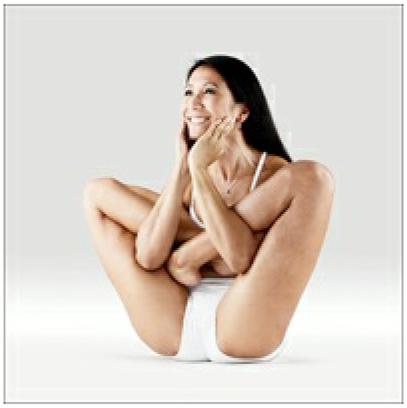

BOTH KNEES BENT: BINDING & TWISTS

Bound Sundial Pose

Baddha Surya Yantrasana

(BUH-duh SOOR-yuh yuhn-TRAHS-uh-nuh)

Modification: both knees bent
Pose Type: seated, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

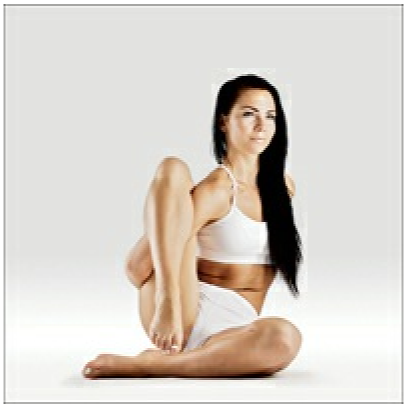

Revolved Pose Dedicated to Sage Marichi

Parivritta Marichyasana

(puh-ri-VRIT-tuh muh-ree-CHYAHS-uh-nuh)

Modification: twisting to the inside, foot to the thigh, binding under the leg
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व य

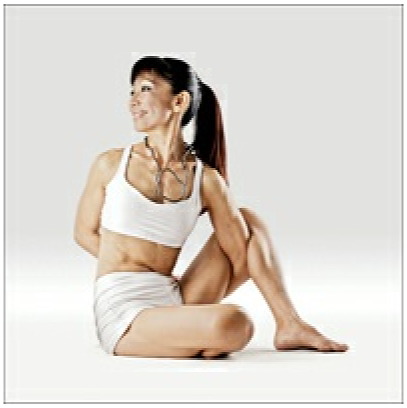

Half Lord of the Fishes Pose 3

Ardha Matsyendrasana 3

(UH-ruh muhts-yeyn-DRAHS-uh-nuh)

Modification: looking to the back
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व य

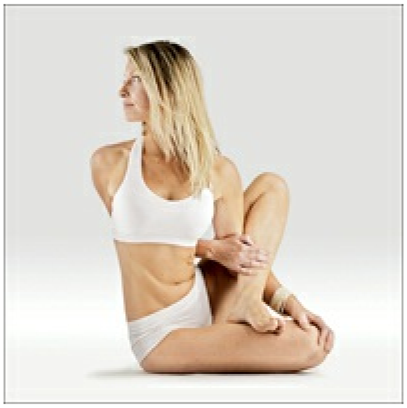

Pose Dedicated to Sage Marichi 4

Marichyasana 4

(muh-ree-CHYAHS-uh-nuh)

Modification: binding under the knee
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

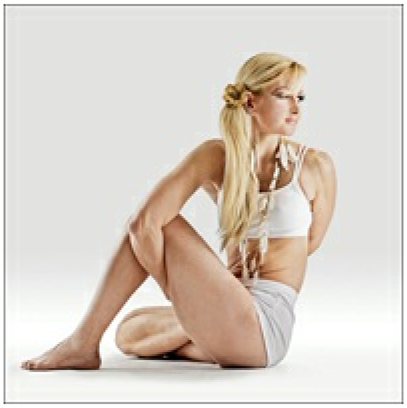

BOTH KNEES BENT: BINDING & TWISTS—ARM THREADED UNDER

Half Pose Dedicated to Sage Agasthiyar

Ardha Agasthiyarasana

(UHR-duh uh-guhs-ti-yahr-AH-suh-nuh)

Modification: 1. one arm threaded through
2. full expression of the pose
Pose Type: 1. seated, forward bend, twist, binding
2. seated, forward bend, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky), Nasagrai or Nasagre (nose), or Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

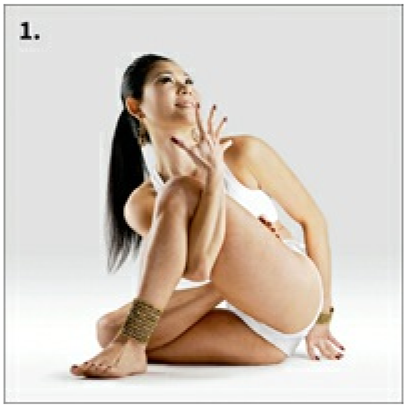

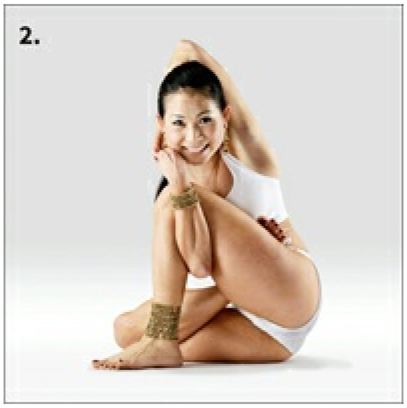

Revolved Half Pose Dedicated to Sage Agasthiyar

Parivritta Ardha Agasthiyarasana

(puh-ri-VRIT-tuh UHR-duh uh-guhs-ti-yahr-AH-suh-nuh)

Pose Type: seated, forward bend, binding, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

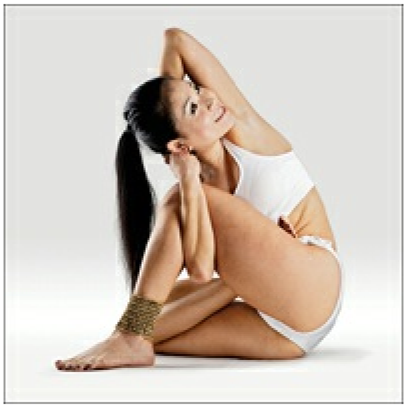

Pose Dedicated to Sage Agasthiyar

Agasthiyarasana

(uh-guhs-ti-yahr-AH-suh-nuh)

Modification: 1. prep. one arm threaded under the knee, fingertips of the other hand to the floor
2. full expression of the pose
Pose Type: seated, forward bend, binding
Drishti Point: 1. Hastagrai or Hastagre (hands)
2. Nasagrai or Nasagre (nose)

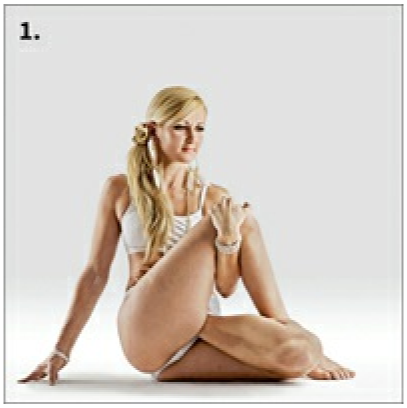

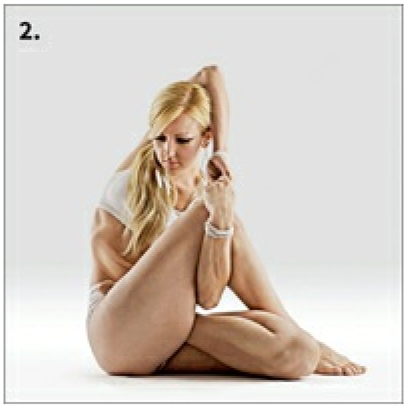

BOTH KNEES BENT: BINDING & TWISTS—KNEES TOGETHER

Seated Noose Pose Prep.

Upavishta Pashasana

(u-puh-VISH-tuh puh-SHAHS-uh-nuh)

Also Known As: Half Noose Pose (Ardha Pashasana), Upavistha Pashasana
Modification: 1. grabbing onto the knees
2. elbow over the opposite knee
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

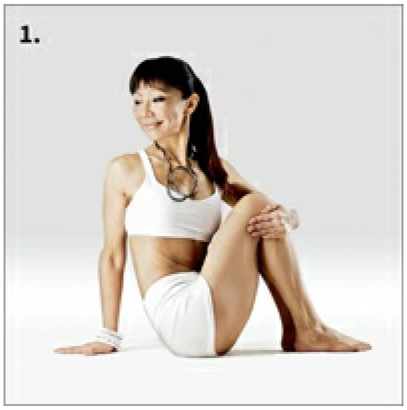

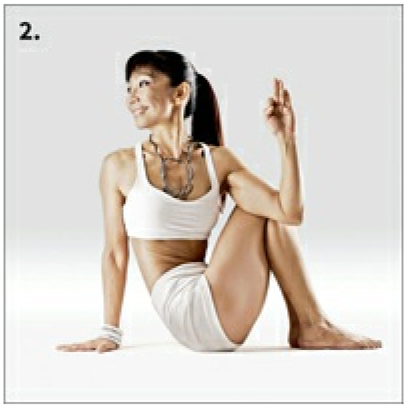

Seated Noose Pose Prep.

Upavishta Pashasana

(u-puh-VISH-tuh puh-SHAHS-uh-nuh)

Also Known As: Upavistha Pashasana
Modification: heels lifted; elbow over the opposite knee, hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated, forward bend, twist, core
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

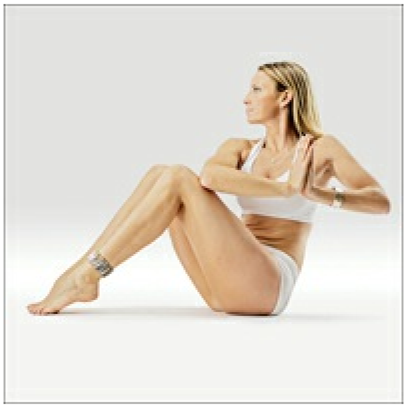

Seated Noose Pose

Upavishta Pashasana

(u-puh-VISH-tuh puh-SHAHS-uh-nuh)

Also Known As: Upavistha Pashasana
Modification: binding around both legs under the knees
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

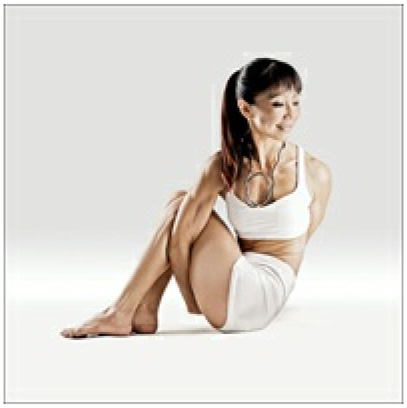

Seated Noose Pose

Upavishta Pashasana

(u-puh-VISH-tuh puh-SHAHS-uh-nuh)

Also Known As: Upavistha Pashasana
Modification: binding around both shins
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

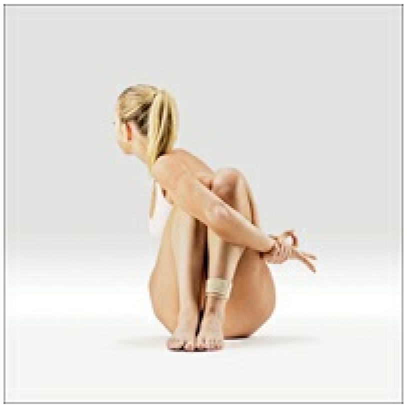

Half Lord of the Fishes Pose

Ardha Matsyendrasana

UHR-duh muhts-yeyn-DRAHS-uh-nuh

Pose Type: seated, forward bend, side bend, twist

Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

How to Perform the Pose:

1. Begin by sitting on the floor with both your legs straight out in front of you. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale as you bend the right knee toward your right shoulder and then bring it over your left leg so that your right foot is flat on the floor on the outside of your left shin.
3. Exhale and bend your left knee, bringing your left heel toward your right sitting bone. Keep the left knee and both your sitting bones evenly on the floor.
4. Exhale as you bring your right elbow to your right knee. Inhale as you bring your left arm up over your head and look towards your left foot (Pose #1).
5. To execute pose #2, exhale as you bring your left elbow over the right knee. Inhale as you bring your right arm up and exhale as you bend your right elbow and look over toward your right foot.
6. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
7. Inhale as you release the twist and exhale as you bring both your legs out in front of you. Repeat on the other side.

Modification: side bend twist:
1. elbow to the knee on the same side
2. elbow to the opposite knee

ardha = half

Matsyendra = a Hindu sage and one of the first teachers of Hatha yoga, a legend, king of the fish

1.

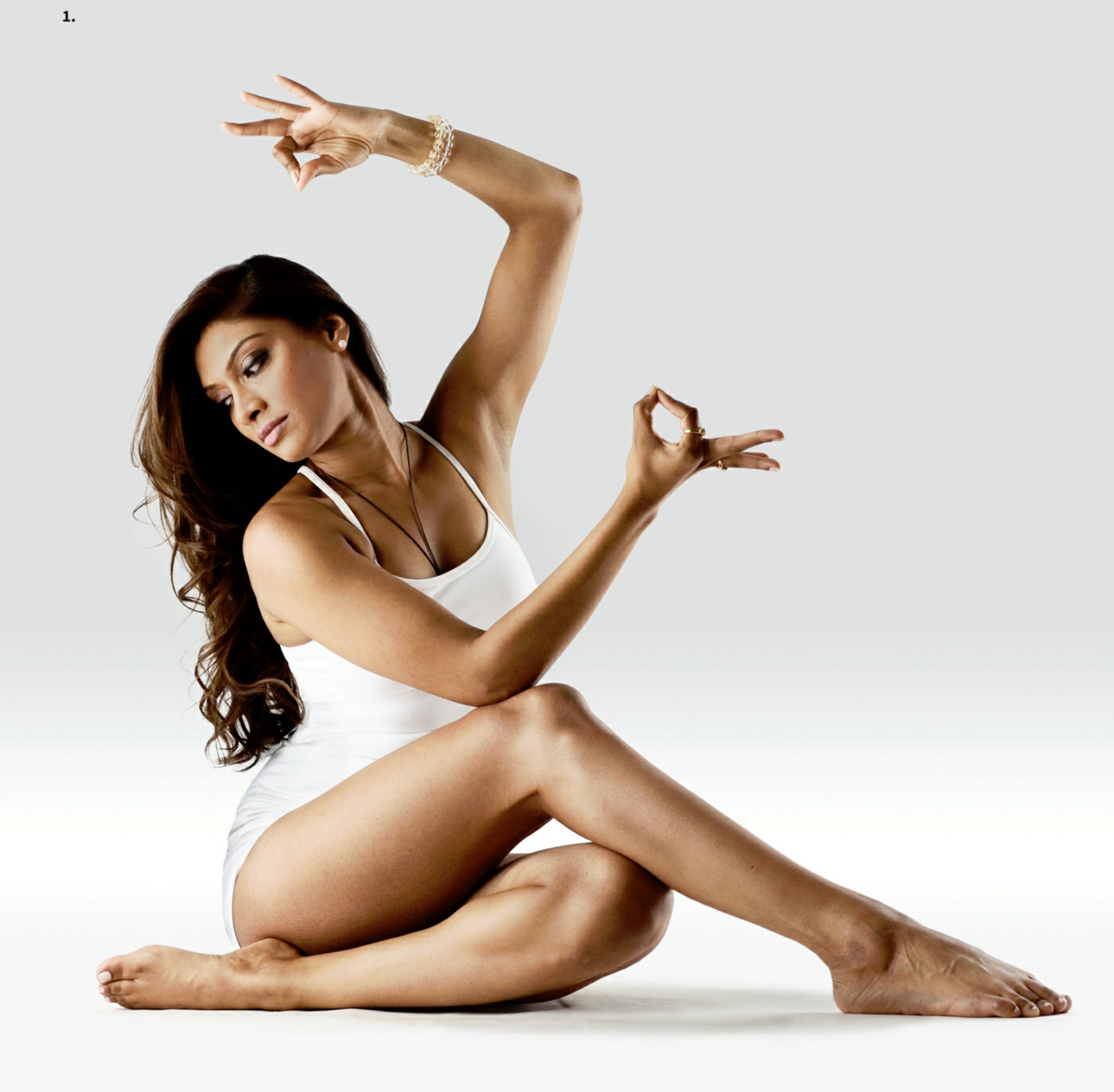

2.

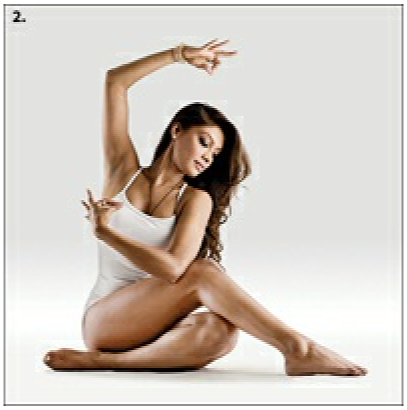

Half Lord of the Fishes Pose

Ardha Matsyendrasana

(UHR-duh muhts-yeyn-DRAHS-uh-nuh)

Modification: backbend
Pose Type: seated, backbend, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

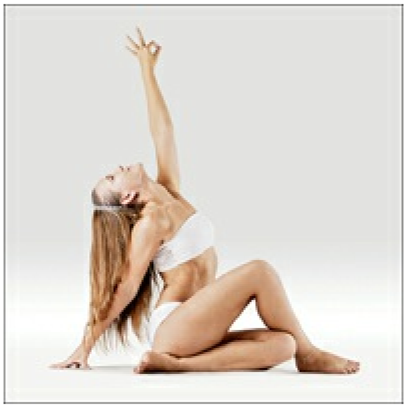

Half Lord of the Fishes Pose

Ardha Matsyendrasana

(UHR-duh muhts-yeyn-DRAHS-uh-nuh)

Modification: twisting to the inside
1. grabbing onto the ankle
2. elbow over the knee
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

1.

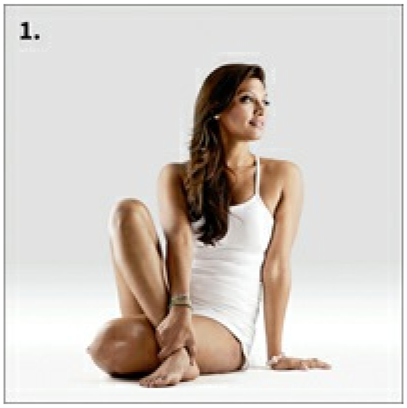

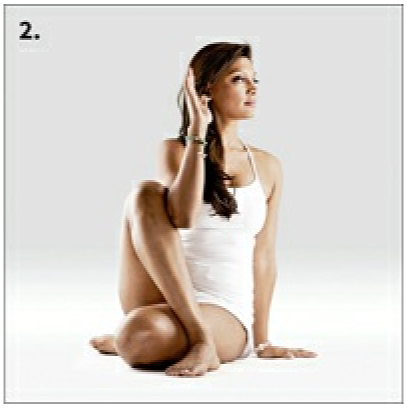

Half Lord of the Fishes Pose

Ardha Matsyendrasana

(UHR-duh muhts-yeyn-DRAHS-uh-nuh)

Modification: twisting to the outside, elbow over the knee; palm of the other hand to the floor
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

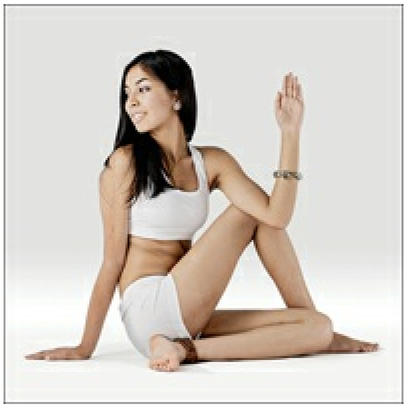

Half Lord of the Fishes Pose 1

Ardha Matsyendrasana 1

(UHR-duh muhts-yeyn-DRAHS-uh-nuh)

Modification: grabbing onto the ankle, other arm behind the back, hand to the inside of the thigh
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

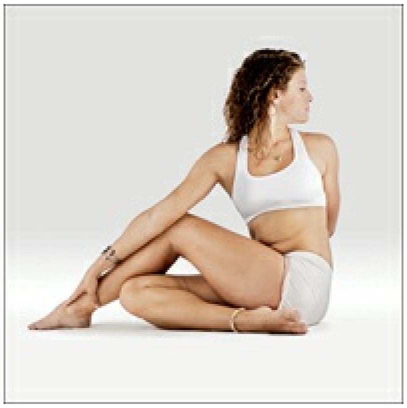

BOTH KNEES BENT: BINDING FORWARD BEND & TWIST

Pose Dedicated to Sage Marichi 3 Prep.

Marichyasana 3 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi B (Marichyasana B)
Modification: top foot to the bottom thigh, binding around the shin, nose to the knee
Pose Type: seated, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)
ल व र

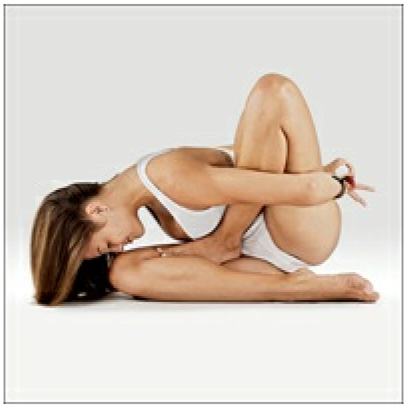

Pose Dedicated to Sage Marichi 4 Prep.

Marichyasana 4 Prep.

(muh-ree-CHYAHS-uh-nuh)

Modification: sitting in a cross-legged position, one knee dropped to the side, foot by the sitting bones; other knee bent toward the shoulder, binding around the shin on the outside of the leg
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

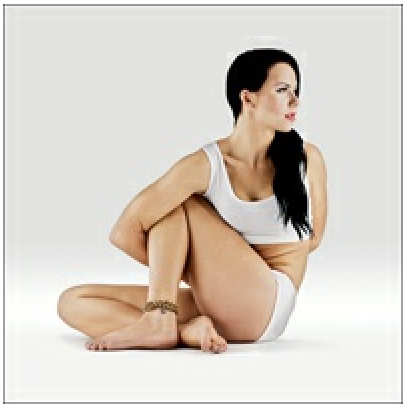

BOTH KNEES BENT: BINDING & TWIST

Pose Dedicated to Sage Marichi 3 Prep.

Marichyasana 3 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi B (Marichyasana B)
Modification: spine straight, bottom knee wrapped around the foot, fingers interlocked on the shin
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

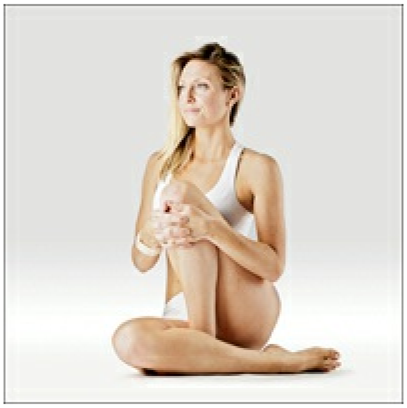

Pose Dedicated to Sage Marichi 3 Prep.

Marichyasana 3 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi B Modification (Marichyasana B Modification)
Modification: bottom knee wrapped around the foot, binding under the thigh

Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व य

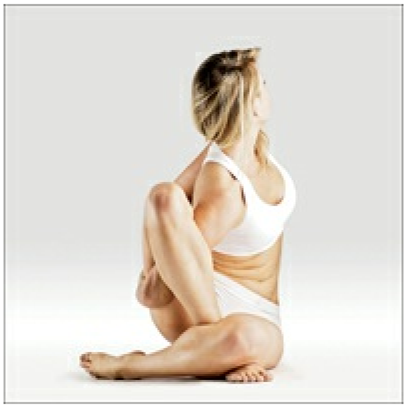

Pose Dedicated to Sage Marichi 3 Prep.

Marichyasana 3 Prep.

(muh-ree-CHYAHS-uh-nuh)
Also Known As: Pose Dedicated to Sage Marichi B Modification (Marichyasana B Modification)
Modification: knee wrapped around the foot, binding over the shin
Pose Type: 1. seated, forward bend, binding
2. seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व र य

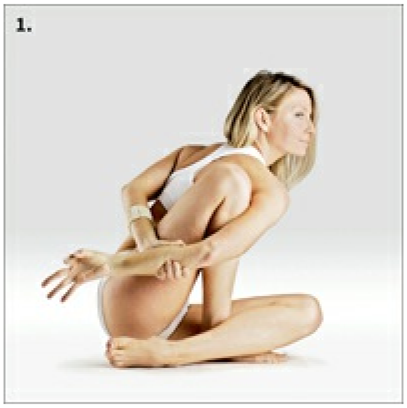

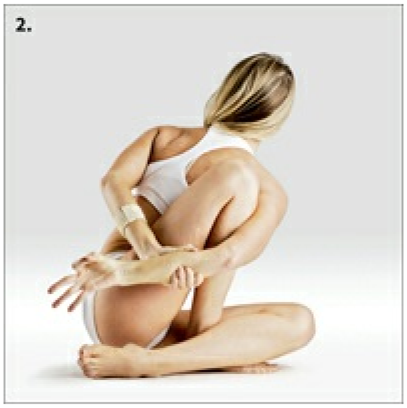

Half Lord of the Fishes Pose 1 Prep.

Ardha Matsyendrasana 1

(UHR-duh muhts-yeyn-DRAHS-uh-nuh)

Modification: foot under the knee, twisting to the outside
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

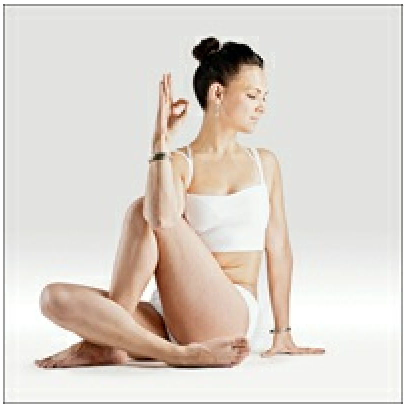

HALF LOTUS POSE: KNEE BENT TOWARD THE CHEST

Pose Dedicated to Sage Marichi Prep.

Marichyasana Prep.

(muh-ree-CHYAHS-uh-nuh)

Modification: prep.-—elbow to the knee on the same side, palm to the floor behind the hips
Pose Type: seated, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

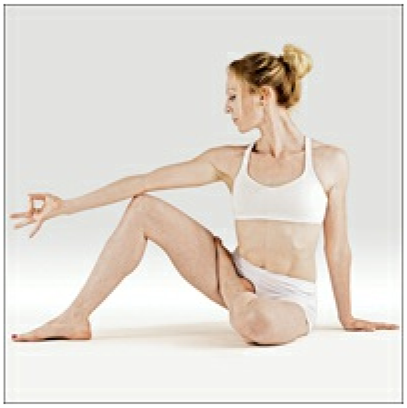

Pose Dedicated to Sage Marichi Prep.

Marichyasana Prep.

(muh-ree-CHYAHS-uh-nuh)

Modification: prep.—arms wrapping around the leg
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

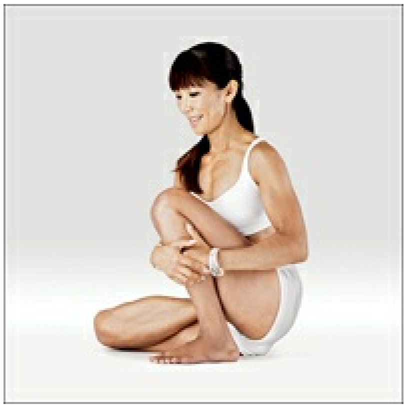

HALF LOTUS POSE: BINDING & TWISTS—KNEE BENT TOWARD THE CHEST

Pose Dedicated to Sage Marichi 3

Marichyasana 3

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi B (Marichyasana B)
Modification: sitting up straight
Pose Type: seated, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)
ल व

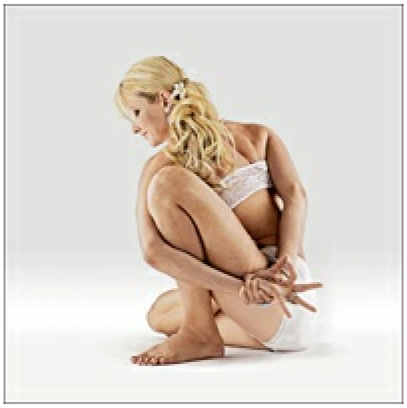

Pose Dedicated to Sage Marichi 3

Marichyasana 3

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi B (Marichyasana B)
Modification: full forward bend
Pose Type: seated, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)
ल व र

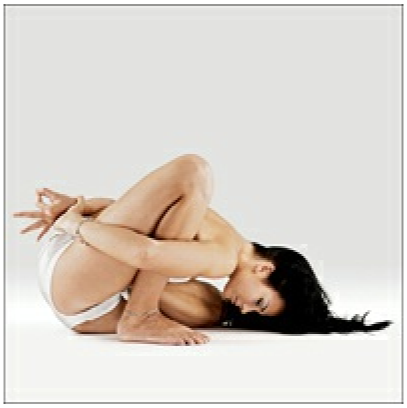

Pose Dedicated to Sage Marichi 4

Marichyasana 4

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi D (Marichyasana D)
Modification: sitting bones on the floor
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व र

LOTUS POSE: BINDING & TWISTS—KNEE BENT TOWARD THE CHEST

Full Lord of the Fishes Pose Prep.

Paripurna Matsyendrasana Prep.

(puh-ri-POOR-nuh muhts-yeyn-DRAHS-uh-nuh)

Modification: one palm to the floor behind the hips, grabbing onto the knee with the opposite hand
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

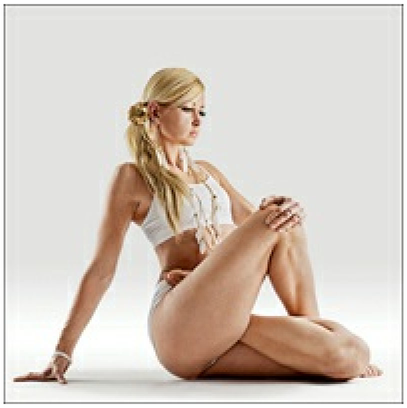

Full Lord of the Fishes Pose

Paripurna Matsyendrasana

(puh-ri-POOR-nuh muhts-yeyn-DRAHS-uh-nuh)

Also Known As: Full Lord of the Fishes Pose (Poorna Matsyendrasana)
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

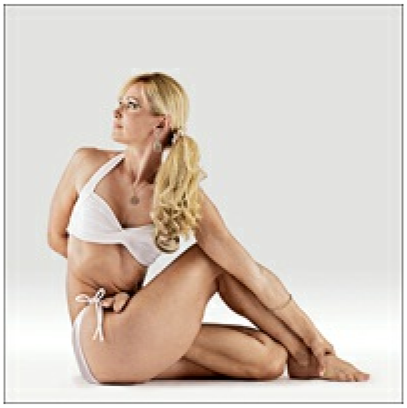

FOOT TO THE BACK: TWISTS—KNEE BENT TOWARD THE CHEST

Root Lord of the Fishes Pose Prep.

Mula Matsyendrasana Prep.

(MOOL-uh muhts-yeyn-DRAHS-uh-nuh)

Modification: heel to the perineum, other foot resting on the knee, fingertips to the floor by the hips
Pose Type: seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
●

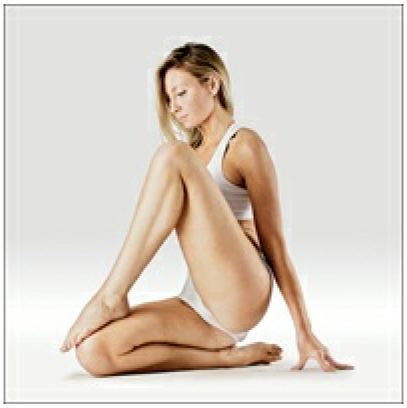

Half Root Lord of the Fishes Pose

Ardha Mula Matsyendrasana

(UHR-duh MOOL-uh muhts-yeyn-DRAHS-uh-nuh)

Modification: heel to the perineum, other foot resting on the knee, fingertips of one hand to the floor by the hips, other hand grabbing onto the opposite knee
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
● ●

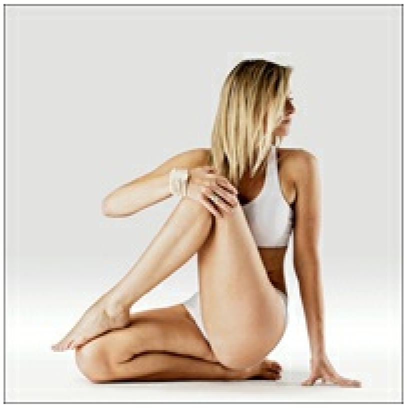

Root Lord of the Fishes Pose Prep.

Mula Matsyendrasana Prep.

(MOOL-uh muhts-yeyn-DRAHS-uh-nuh)

Modification: backbend, heel to the perineum, other foot resting on the knee, fingertips to the floor behind the hips
Pose Type: seated, backbend, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

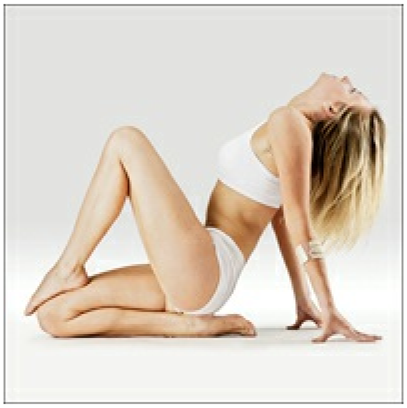

Pose Dedicated to Sage Marichi 6 Prep.

Marichyasana 6 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi F (Marichyasana F)
Modification: opposite hand grabbing onto the foot, heel resting on the knee, other arm behind the back to the inside of the thigh, looking over the shoulder to the back
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

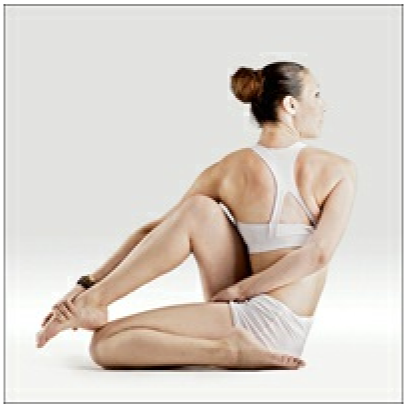

Pose Dedicated to Sage Marichi 6 Prep.

Marichyasana 6 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi F (Marichyasana F)
Modification: opposite hand grabbing onto the foot, foot flat to the floor, other arm behind the back to the inside of the thigh, looking toward the foot
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

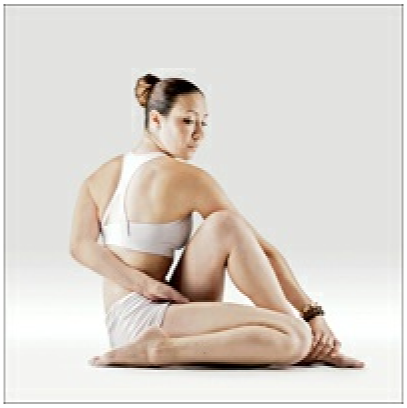

BOUND ANGLE POSE: ARMS IN FRONT

Bound Angle Pose with Hands in Prayer

Baddha Konasana Namaskar

(BUH-duh ko-NAHS-uh-nuh nuh-muhs-KAHR)

Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose) or Hastagrai or Hastagre (hands)

Bound Angle Pose

Baddha Konasana

(BUH-duh ko-NAHS-uh-nuh)

Modification: fingertips touching the floor
Pose Type: seated
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

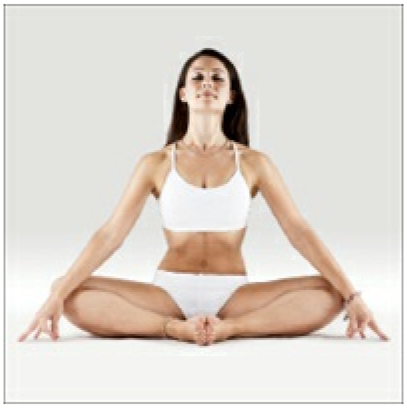

Bound Angle Pose

Baddha Konasana

(BUH-duh ko-NAHS-uh-nuh)

Modification: half forward bend, opening the soles of the feet up to the sky
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

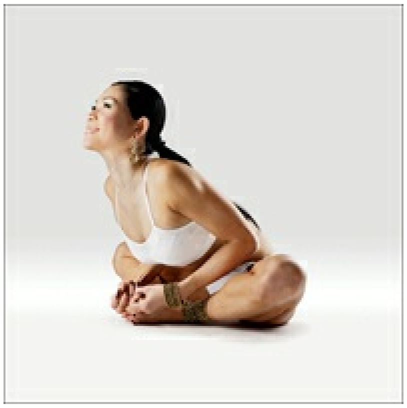

BOUND ANGLE POSE: ARMS BEHIND

Bound Angle Pose

Baddha Konasana

(BUH-duh ko-NAHS-uh-nuh)

Modification: one hand to the knee; other hand behind, grabbing onto the bicep
Pose Type: seated, binding
Drishti Point: Nasagrai or Nasagre (nose)
ल व य

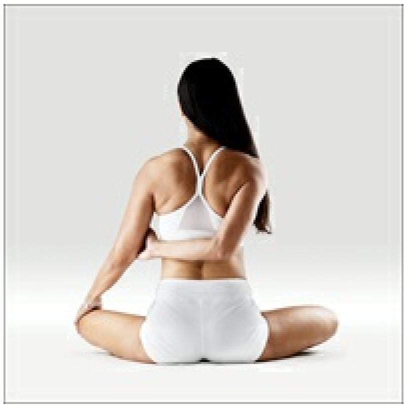

Bound Hands Bound Angle Pose

Baddha Hasta Baddha Konasana

(BUH-duh HUH-stuh BUH-duh ko-NAHS-uh-nuh)

Modification: arms behind, grabbing onto the elbows
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)
ल व य

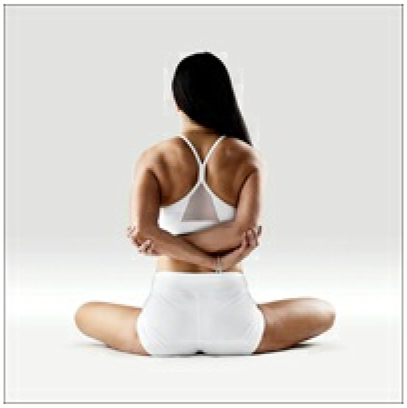

Reverse Prayer Bound Angle Pose

Viparita Namaskar Baddha Konasana

(vi-puh-REE-tuh nuh-muhs-KAHR BUH-duh ko-NAHS-uh-nuh)

Also Known As: Back of the Body Prayer Bound Angle Pose (Paschima Namaskara Baddha Konasana)
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

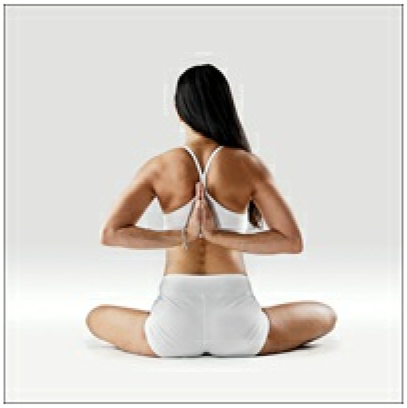

Hand Position of Cow Face Pose in Bound Angle Pose

Hasta Gomukhasana in Baddha Konasana

(HUH-stuh go-muk-AHS-uh-nuh in BUH-duh ko-NAHS-uh-nuh)

Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

BOUND ANGLE POSE: SIDE BEND & BINDING

Sideways Bound Angle Pose

Parshva Baddha Konasana

(PAHRSH-vuh BUH-duh ko-NAHS-uh-nuh)

Pose Type: seated, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

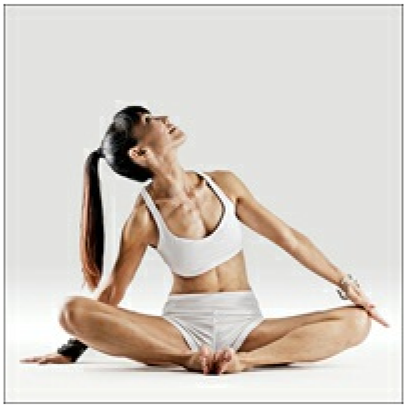

Sideways Star Pose

Parshva Tarasana

(PAHRSH-vuh tahr-AHS-uh-nuh)

Pose Type: seated, side bend
Drishti Point: Hastagrai or Hastagre (hands)

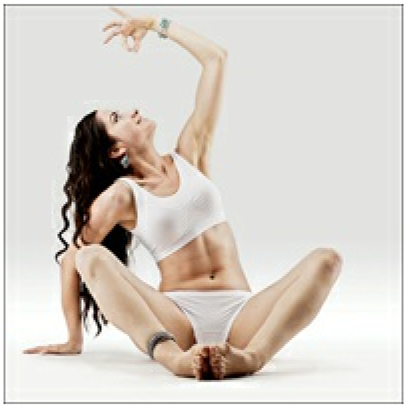

Sideways Bound Leg Bound Angle Pose

Parshva Baddha Pada Baddha Konasana

(PAHRSH-vuh BUH-duh PUH-duh BUH-duh ko-NAHS-uh-nuh)

Modification: binding around one leg, arm wrapped around the shin
Pose Type: seated, forward bend, side bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

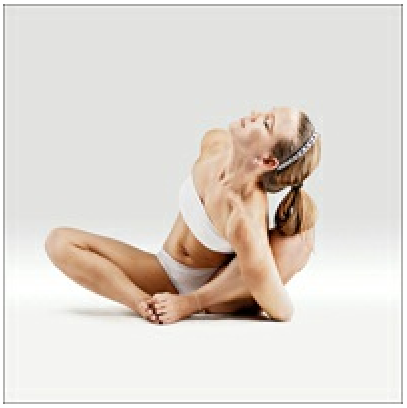

BOUND ANGLE POSE: HEELS CLOSE, FEET LIFTED, SITTING BONES LIFTED AND FORWARD BEND

Pose Dedicated to Sage Goraksha

Gorakshasana

(go-rahk-SHAHS-uh-nuh)

Modification: feet in Bound Angle Pose (Baddha Konasana), sitting on the heels
1. one hand to the centre of the chest, other hand to the floor behind the hips
2. hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated, balance
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

Equilibrium Bound Angle Pose 1

Tulya Baddha Konasana 1

(TUL-yuh BUH-duh ko-NAHS-uh-nuh)

Modification: feet lifted off the floor

Pose Type: seated, forward bend, core

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

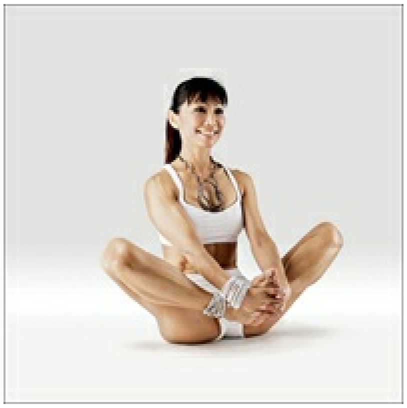

Equilibrium Bound Angle Pose 2

Tulya Baddha Konasana 2

(TUL-yuh BUH-duh ko-NAHS-uh-nuh)

Modification: sitting bones lifted off the floor
Pose Type: seated, forward bend, core
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व

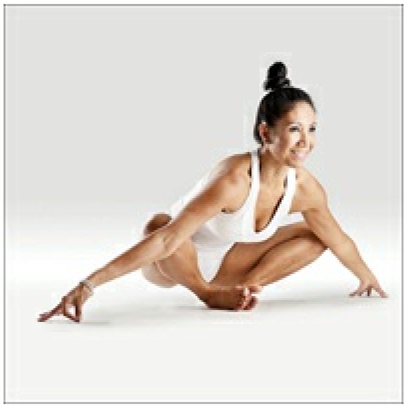

Bound Angle Pose

Baddha Konasana

(BUH-duh ko-NAHS-uh-nuh)

Modification: forward bend; arms straight to the front, palms down
Pose Type: seated, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs) or Nasagrai or Nasagre (nose)
ल व

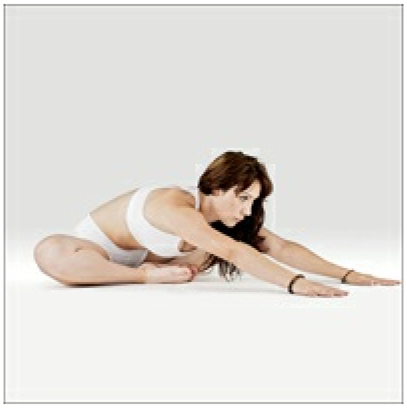

BOUND ANGLE POSE: FORWARD BEND

Bound Angle Pose

Baddha Konasana

(BUH-duh ko-NAHS-uh-nuh)

Modification: 1. half forward bend
2. full forward bend, chin to the floor

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose) or Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

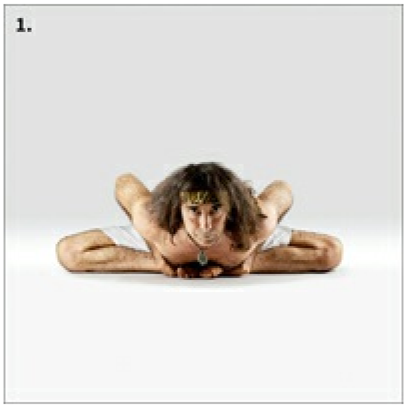

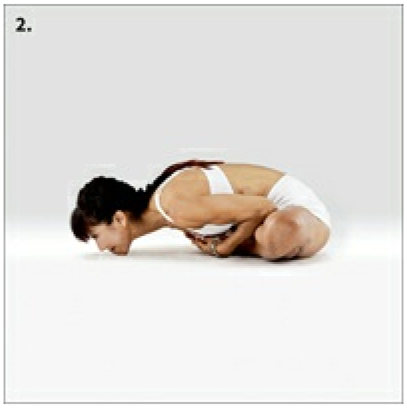

Bound Angle Pose

Baddha Konasana

(BUH-duh ko-NAHS-uh-nuh)

Modification: elbows bent and to the floor, forehead to the floor, palms together, fingers spread wide

Pose Type: seated, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Nasagrai or Nasagre (nose)

STAR POSE

Easy Embryo in the Womb Pose

Sukha Garbha Pindasana

(SUK-uh GUHR-buh pin-DAHS-uh-nuh)

Modification: arms under the legs
1. hands to the feet
2. fingertips to the temples
3. hands to the sides; heels touching, feet to the sides
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

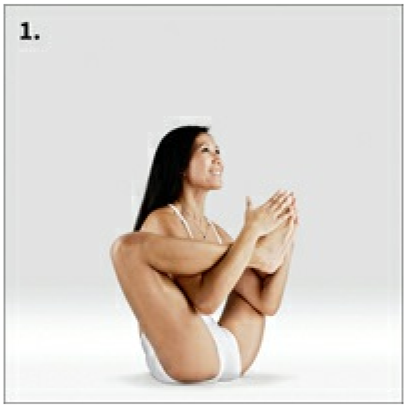

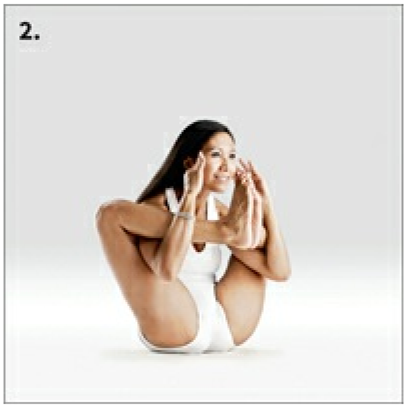

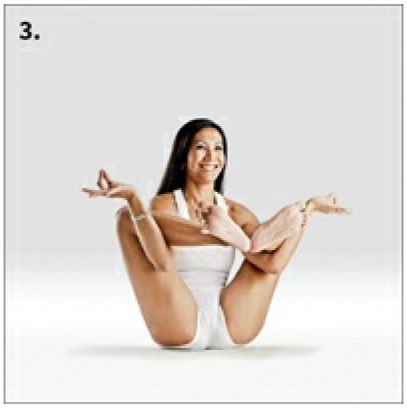

Easy Embryo in the Womb Pose

Sukha Garbha Pindasana

(SUK-uh GUHR-buh pin-DAHS-uh-nuh)

Also Known As: Yoga Pose A Prep. (Yogasana A Prep.)
Modification: shins into armpits, hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

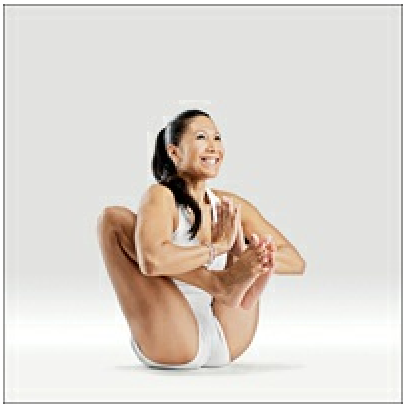

Star Pose

Tarasana

(tahr-AHS-uh-nuh)

Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

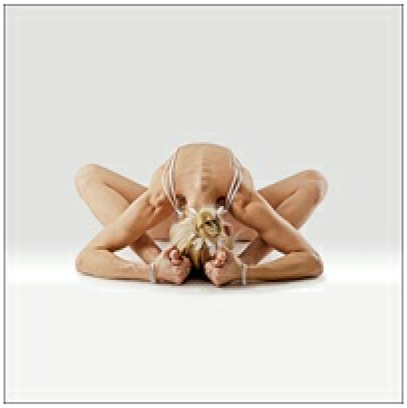

Upward Star Pose

Urdhva Tarasana

(OORD-vuh tahr-AHS-uh-nuh)

Pose Type: seated, forward bend, core

Drishti Point: Padhayoragrai or Padayoragre (toes/feet) or Nasagrai or Nasagre (nose)

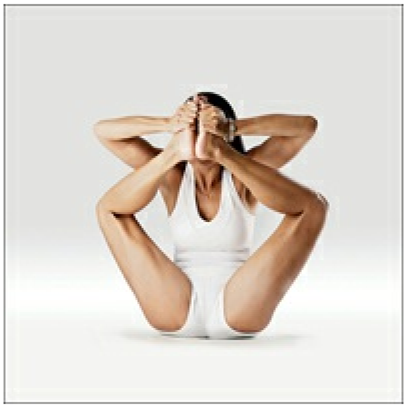

LEGS IN FRONT: KNEES BENT

Seated Eastern Intense Stretch Pose

Upavishta Purvottanasana

(u-puh-VISH-tuh poor-vo-TAHS-uh-nuh)

Also Known As: Reverse Plank Prep.; Upavistha Purvottanasana

Pose Type: seated, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

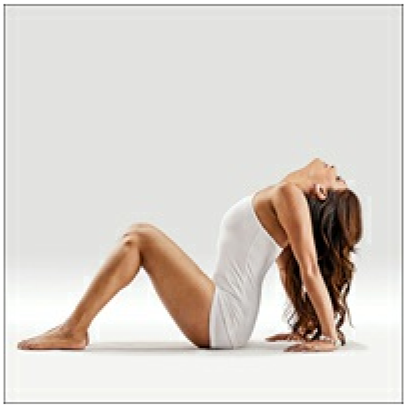

One Hand Seated Eastern Intense Stretch Pose

Eka Hasta Upavishta Purvottanasana

(EY-kuh HUH-stuh u-puh-VISH-tuh poor-vo-TAHS-uh-nuh)

Also Known As: One Hand Reverse Plank Prep.; Eka Hasta Upavistha Purvottanasana
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)
ल र य ह ॐ

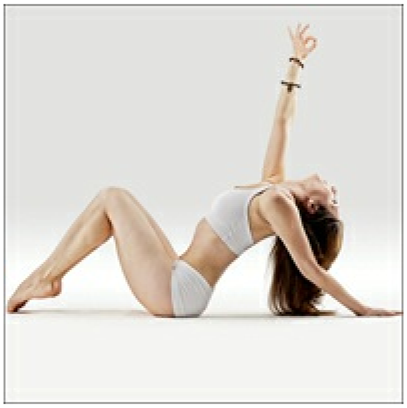

LEGS IN FRONT: KNEES BENT—LEGS OUTSIDE OF THE SHOULDERS

Seated Knee to Shoulder Pose

Upavishta Janu Bhujasana

(u-puh-VISH-tuh JAH-nu buj-AHS-uh-nuh)

Also Known As: Upavistha Janu Bhujasana
Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

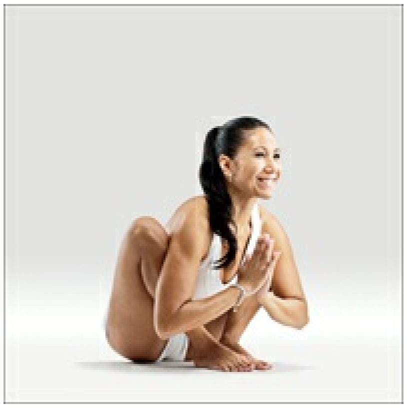

Broken Wing Pose

Avabhinna Pakshakasana

(uh-vuh-BIN-uh puhk-shuh-KAHS-uh-nuh)

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

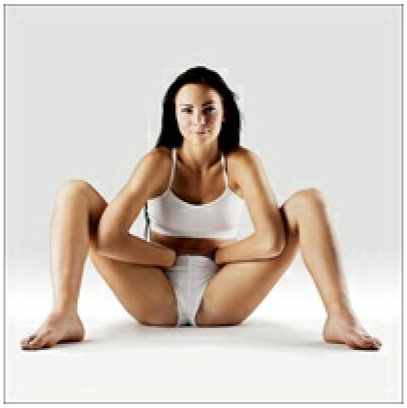

Tortoise Pose Prep.

Kurmasana Prep.

(koor-MAHS-uh-nuh)

Modification: grabbing onto the ankles

Pose Type: seated, forward bend

Drishti Point: Padhayoragrai or Padayoragre (toes/feet) or Nasagrai or Nasagre (nose)

Tortoise Pose Prep.

Kurmasana Prep.

(koor-MAHS-uh-nuh)

Modification: 1. arms straight
2. elbows bent
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose,) Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व

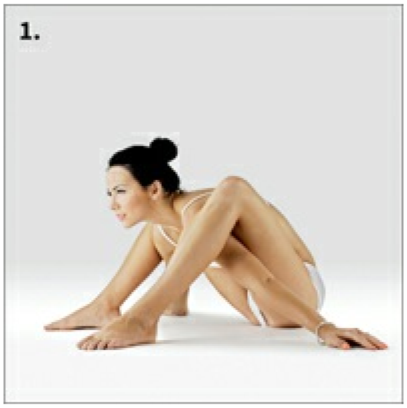
1.

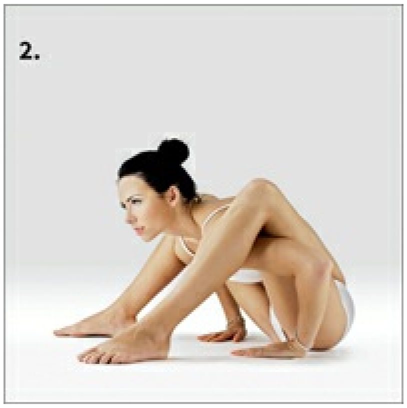
2.

Sleeping Tortoise Pose Prep.

Supta Kurmasana Prep.

(SUP-tuh koor-MAHS-uh-nuh)

Modification: feet unhooked

Pose Type: seated, forward bend, binding

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ● ● ●

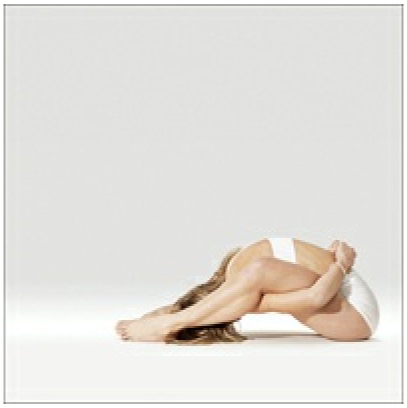

Sleeping Tortoise Pose

Supta Kurmasana

(SUP-tuh koor-MAHS-uh-nuh)

Pose Type: seated, forward bend, binding

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ● ●

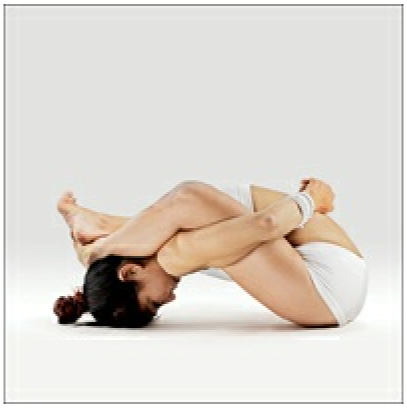

HORIZONTAL SPLITS: LEGS LIFTED—KNEES BENT

Half Upward Seated Angle Pose Prep.

Ardha Urdhva Upavistha Konasana Prep.

(UHR-duh OORD-vuh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Modification: shoulders to the back of the knees, palms to the floor, knees bent
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

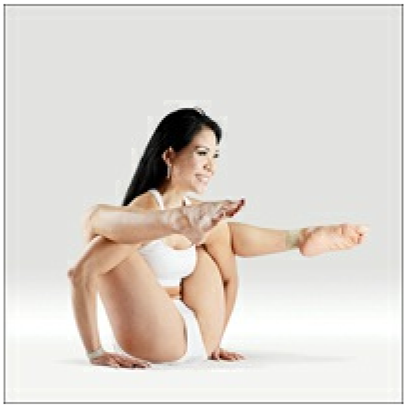

Bound Legs in Half Pose Dedicated to Sage Koormamuni

Baddha Pada Ardha Koormamunyasana

(BUH-duh PUH-duh UHR-duh koor-muh-mun-YAHS-uh-nuh)

Pose Type: seated, forward bend, core, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

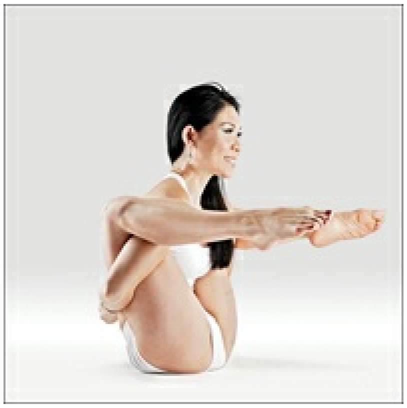

HORIZONTAL SPLITS: LEGS LIFTED

Half Upward Seated Angle Pose Prep. in One Hand to Foot Boat Pose

Ardha Urdhva Upavishta Konasana Prep. in Eka Hasta Pada Navasana

(UHR-duh OORD-vuh u-puh-VISH-tuh ko-NAHS-uh-nuh in EY-kuh HUH-stuh PUH-duh nuh-VAHS-uh-nuh)

Also Known As: Ardha Urdhva Upavistha Konasana Prep. in Eka Hasta Pada Navasana
Modification: one leg straight, back of the other knee to the shoulder
Pose Type: seated, forward bend, core
Drishti Point: Hastagrai or Hastagre (hands) or Padhayoragrai or Padayoragre (toes/feet)

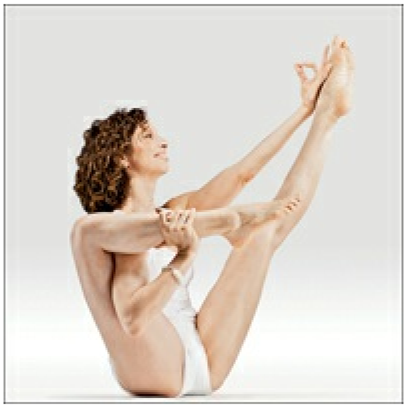

Half Upward Seated Angle Pose

Ardha Urdhva Upavishta Konasana

(UHR-duh OORD-vuh u-puh-VISH-tuh ko-NAHS-uh-nuh)
Also Known As: Ardha Urdhva Upavistha Konasana
Modification: palms to the floor, one leg straight up to the sky; other foot to the floor, knee to the chest
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

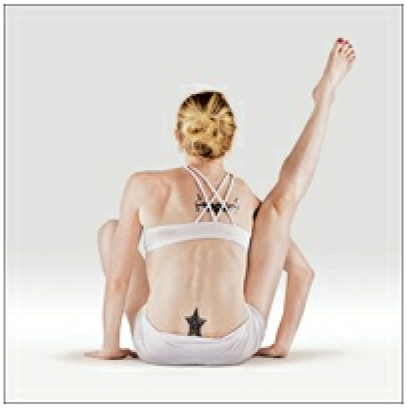

Upward Seated Angle Pose

Urdhva Upavishta Konasana

(UHR-duh OORD-vuh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Urdhva Upavistha Konasana
Modification: arms open to the sides, triceps to the back of the knees
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

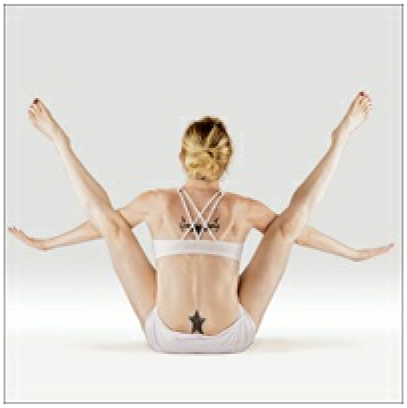

HORIZONTAL SPLITS: LEGS LIFTED & STRAIGHT

Seated Firefly Pose

Upavishta Tittibhasana

(u-puh-VISH-tuh ti-ti-BAHS-uh-nuh)

Also Known As: Upward Seated Angle Pose Prep. (Urdhva Upavishta Konasana Prep.), Pose Dedicated to Sage Koormamuni Prep. (Koormamunyasana Prep.), Upavistha Tittibhasana
Modification: palms to the floor, shoulders to the back of the knees
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

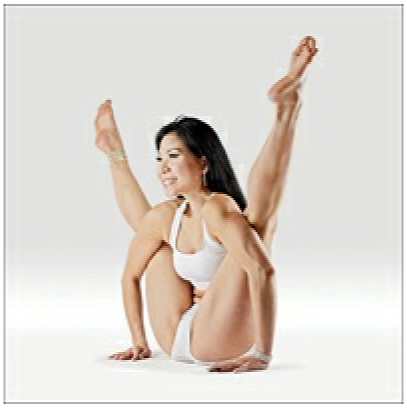

Pose Dedicated to Sage Koormamuni

Koormamunyasana

(koor-muh-mun-YAHS-uh-nuh)

Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

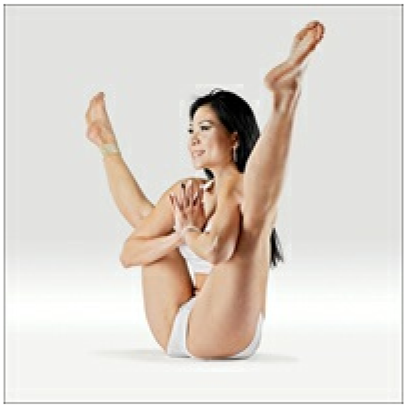

Pose Dedicated to Sage Koormamuni—Bound Legs

Baddha Pada Koormamunyasana

(BUH-duh PUH-duh koor-muh-mun-YAHS-uh-nuh)

Modification: legs straight
Pose Type: seated, forward bend, core, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

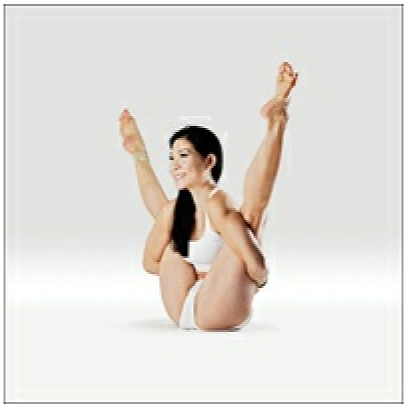

Hands to Feet Upward Seated Angle Pose

Pada Hasta Urdhva Upavishta Konasana

(PUH-duh HUH-stuh OORD-vuh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Pada Hasta Urdhva Upavistha Konasana
Modification: grabbing onto the outside edges of the feet
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Urdhva or Antara Drishti (up to the sky)

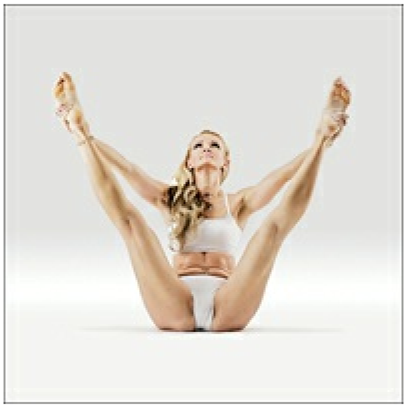

Big Toe Upward Seated Angle Pose

Padangushta Urdhva Upavishta Konasana

(puhd-ahng-GOOSH-tuh OORD-vuh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Padangushta Urdhva Upavistha Konasana
Modification: grabbing onto the big toes, arms parallel to the floor
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Urdhva or Antara Drishti (up to the sky)

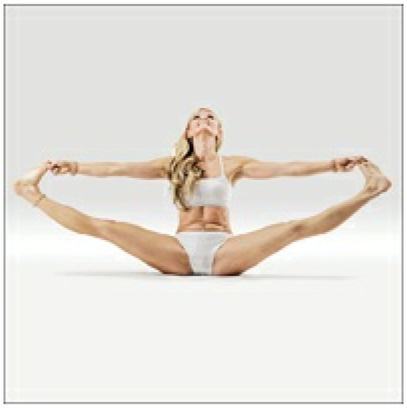

Upward Seated Angle Pose

Urdhva Upavishta Konasana

(OORD-vuh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Urdhva Upavistha Konasana
Modification: palms to the floor, heels of the palms touching, fingertips pointing to the sides; feet lifted, toes pointed
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Urdhva or Antara Drishti (up to the sky)
लं वं हं ॐ

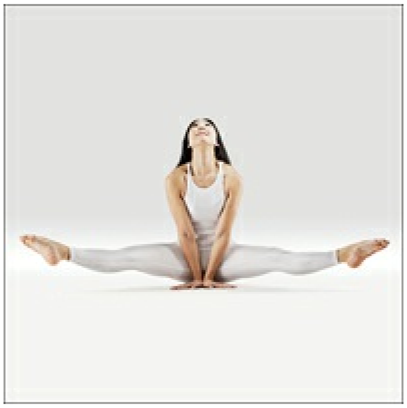

HORIZONTAL SPLITS: FORWARD BEND

Tortoise Pose

Kurmasana

(koor-MAHS-uh-nuh)

Modification: 1. knees bent
2. legs straight
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं हं ॐ

1.

2.

Equal Angle Pose

Samakonasana

(suh-muh-ko-NAHS-uh-nuh)

Modification: forward bend, forearms to the floor
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

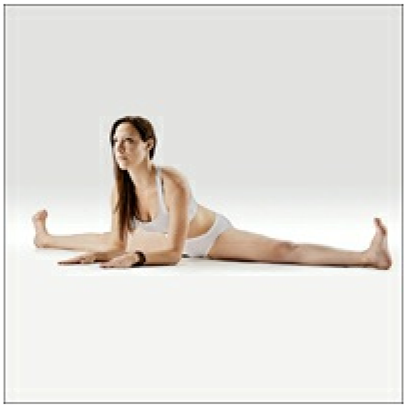

Big Toe Seated Angle Pose

Padangushta Upavishta Konasana

(puhd-ahng-GOOSH-tuh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Padangushta Upavistha Konasana

Modification: forward bend, chin to the floor

Pose Type: seated, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

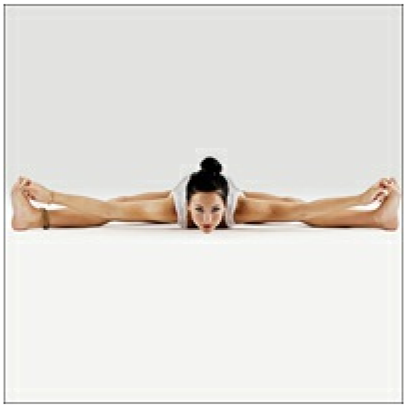

Equal Angle Pose

Samakonasana

(suh-muh-ko-NAHS-uh-nuh)

Modification: forward bend, arms crossed in front

Pose Type: seated, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

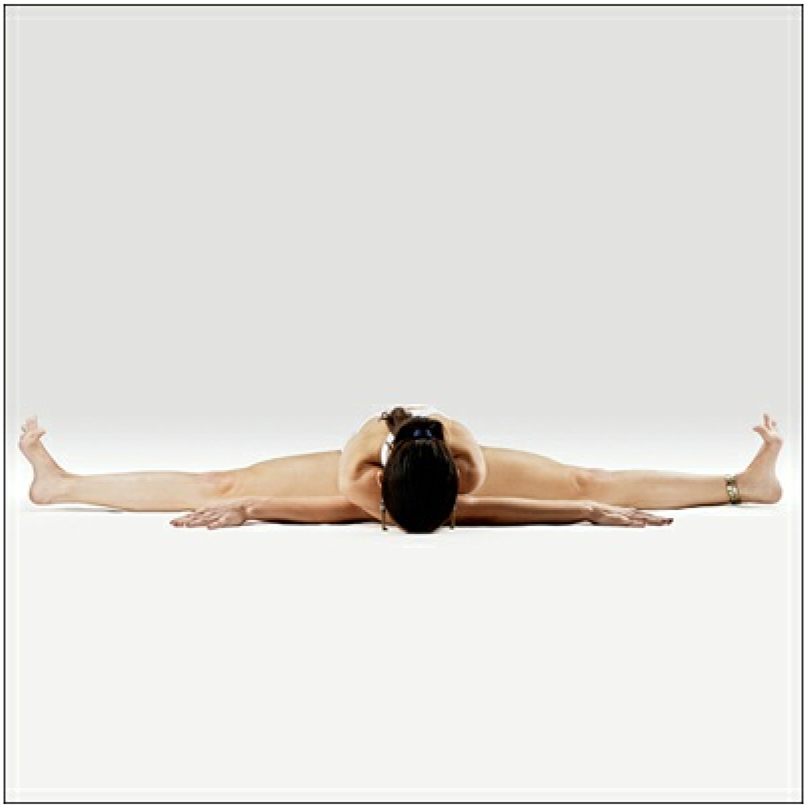

Equal Angle Pose

Samakonasana

(suh-muh-ko-NAHS-uh-nuh)

Modification: forward bend, arms extended to the front
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य ॐ

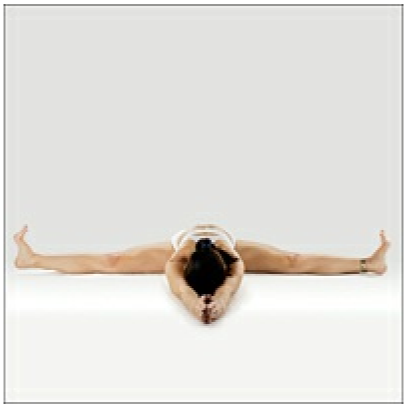

Reverse Prayer Equal Angle Pose

Viparita Namaskar Samakonasana

(vi-puh-REE-tuh nuh-muhs-KAHR suh-muh-ko-NAHS-uh-nuh)

Also Known As: Back of the Body Prayer Equal Angle Pose (Paschima Namaskara Samakonasana)
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

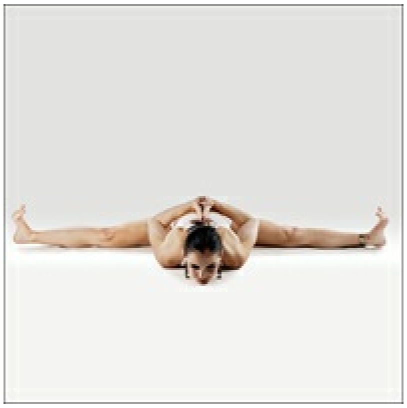

Hands Bound Equal Angle Pose

Baddha Hasta Samakonasana

(BUH-duh HUH-stuh suh-muh-ko-NAHS-uh-nuh)

Modification: forward bend
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

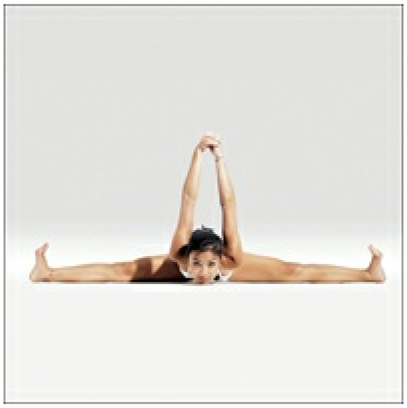

HORIZONTAL SPLITS: SPINE STRAIGHT

Seated Angle Pose

Upavishta Konasana

(u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Upavistha Konasana

Modification: mild version, backbend
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं रं यं हं ॐ

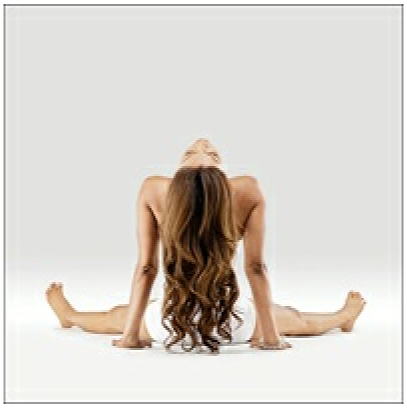

Seated Angle Pose

Upavishta Konasana

(u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Upavistha Konasana
Modification: 1. hands in Anjali Mudra (Hands in Prayer), toes flexed in
2. hands in reverse prayer, toes flexed in
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose) or Hastagrai or Hastagre (hands)
लं वं

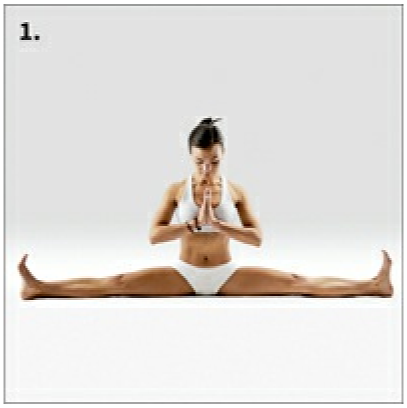

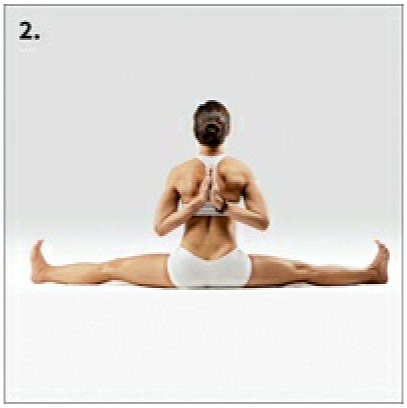

Seated Angle Pose

Upavishta Konasana

(u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Upavistha Konasana
Modification: palms to the floor in front of the hips, heels of the palms touching, fingertips pointing to the side, straight spine, toes pointed
Pose Type: seated
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

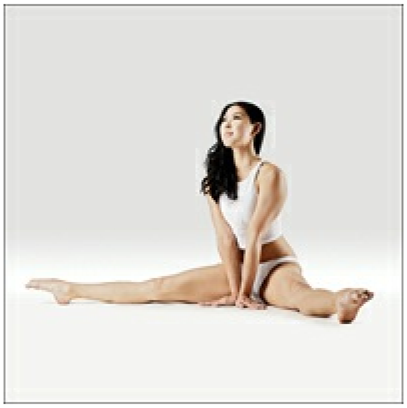

Equal Angle Pose

Samakonasana

(su-muh-ko-NAHS-uh-nuh)

Modification: hands on the knees, spine straight
Pose Type: seated
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

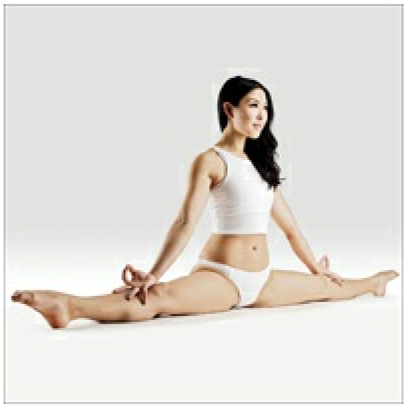

Bound Equal Angle Pose

Baddha Samakonasana

(BUH-duh su-muh-ko-NAHS-uh-nuh)

Pose Type: seated, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)
ल व य

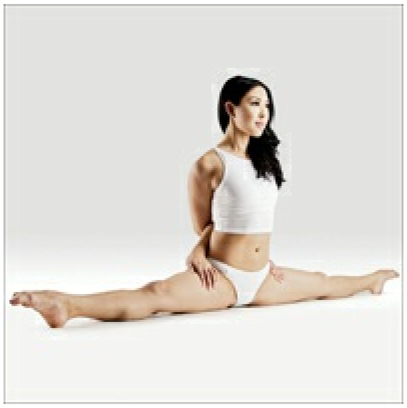

HORIZONTAL SPLITS: TWISTS

Half Bound Seated Angle Pose

Ardha Baddha Upavishta Konasana

(UHR-duh BUH-duh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Ardha Baddha Upavistha Konasana
Modification: toes pointed, one hand binding behind the back, fingertips of the other hand to the floor in front of the hips
Pose Type: seated, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

Revolved Half Bound Equal Angle Pose

Parivritta Samakonasana

(puh-ri-VRIT-tuh suh-muh-ko-NAHS-uh-nuh)

Modification: 1. toes pointed, both hands on one leg
2. toes flexed in, one hand to the leg, other hand to the floor
Pose Type: seated, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

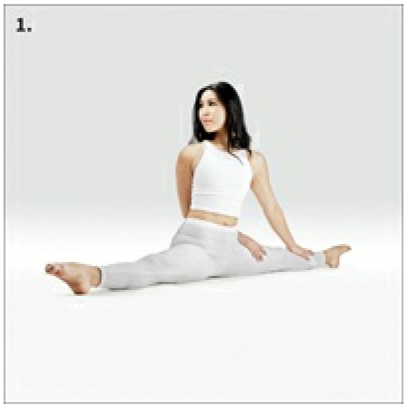

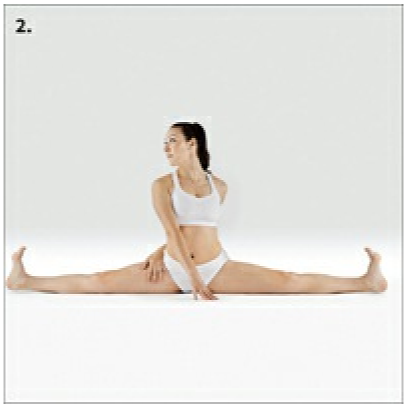

Revolved Bound Equal Angle Pose

Parivritta Baddha Samakonasana

(puh-ri-VRIT-tuh BUH-duh suh-muh-ko-NAHS-uh-nuh)

Modification: 1. toes pointed
2. toes flexed in
Pose Type: seated, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

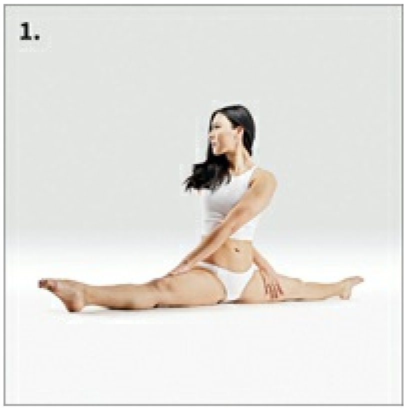

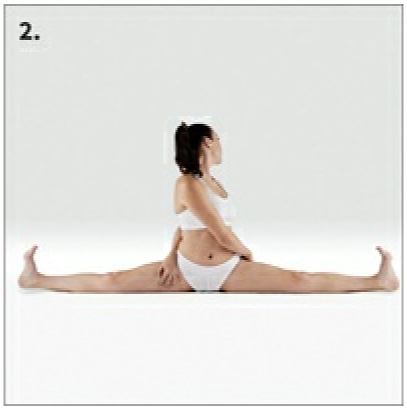

HORIZONTAL SPLITS: SIDE BENDS

Sideways Equal Angle Pose

Parshva Samakonasana

(PAHRSH-vuh suh-muh-ko-NAHS-uh-nuh)

Modification: side bend, fingertips to the back of the head; elbow to the floor
Pose Type: seated, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

Both Hands to Foot Revolved Seated Angle Pose

Dwi Hasta Pada Parivritta Upavishta Konasana

(DWI-huh-stuh PUH-duh puh-ri-VRIT-tuh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Dwi Hasta Pada Parivritta Upavistha Konasana
Pose Type: seated, side bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
● ● ● ●

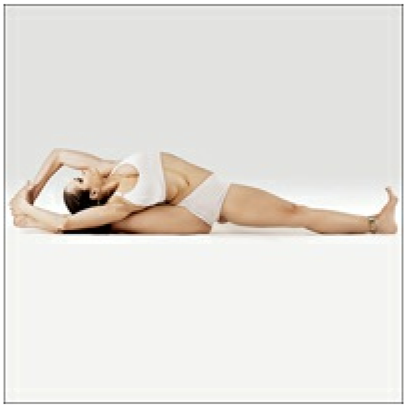

LEGS STRAIGHT IN FRONT: SPINE STRAIGHT & BACKBEND

Staff Pose

Dandasana

(duhn-DAHS-uh-nuh)

Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose), Padhayoragrai or Padayoragre (toes/feet)
●

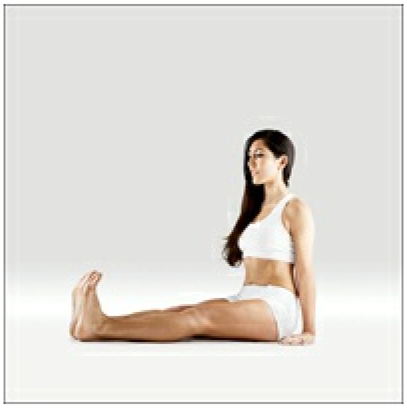

Reverse Prayer Staff Pose

Viparita Namaskar Dandasana

(vi-puh-REE-tuh nuh-muhs-KAHR duhn-DAHS-uh-nuh)

Also Known As: Back of the Body Prayer Staff Pose (Paschima Namaskara Dandasana)
Pose Type: seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

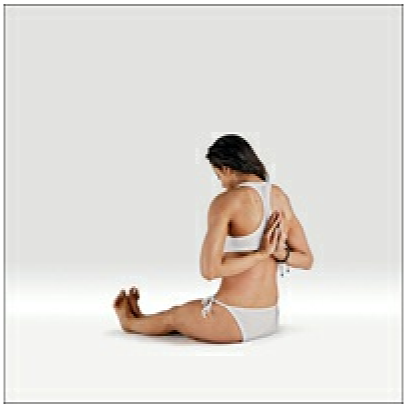

Staff Pose

Dandasana

(duhn-DAHS-uh-nuh)

Modification: 1. fingertips pointing to the heels
2. fingertips pointing away from the heels, head rolling back
Pose Type: seated, mild backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

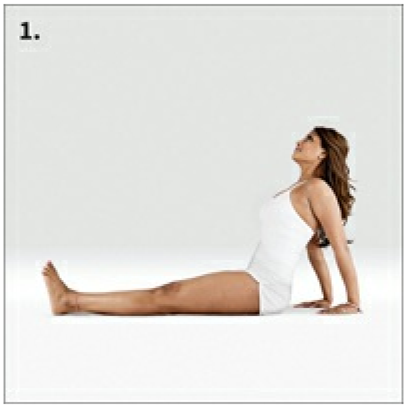

2.

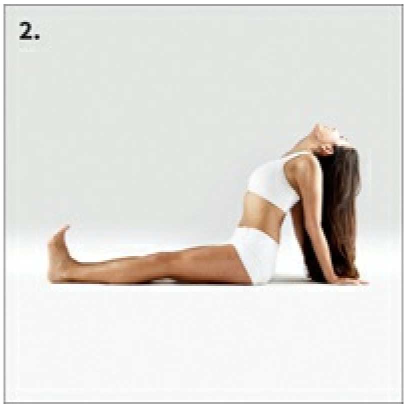

LEGS STRAIGHT IN FRONT: FORWARD BEND

Western Intense Stretch Pose

Paschimottanasana

(puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Seated Forward Bend
Modification: grabbing onto the big toes
Pose Type: seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)
ल र य

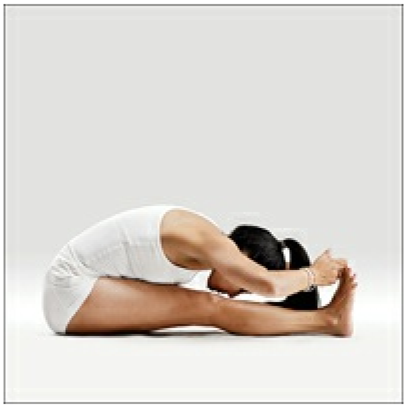

Western Intense Stretch Pose

Paschimottanasana

(puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Seated Forward Bend
Modification: grabbing onto the balls of the feet
Pose Type: seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

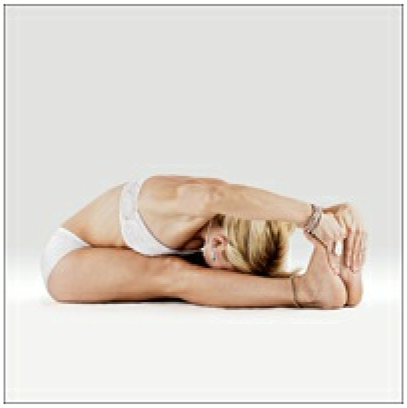

Western Intense Stretch Pose

Paschimottanasana

(puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Seated Forward Bend
Modification: grabbing onto the wrist
Pose Type: seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

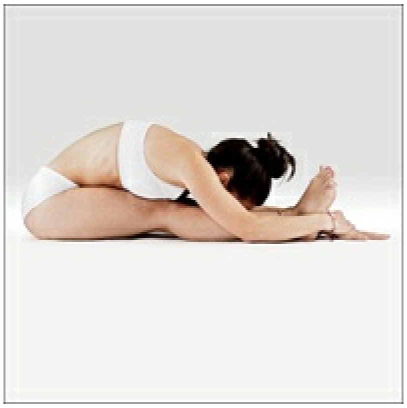

Western Intense Stretch Pose

Paschimottanasana

(puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Seated Forward Bend
Modification: palms down
Pose Type: seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)
लं रं यं

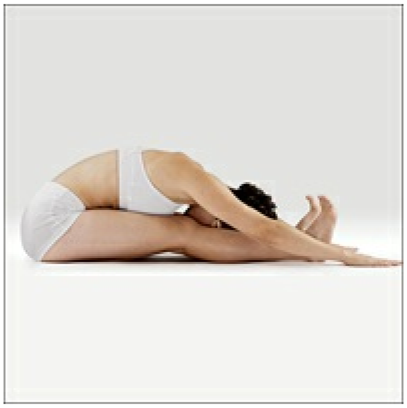

Bound Hands Western Intense Stretch Pose

Baddha Hasta Paschimottanasana

(BUH-duh HUH-stuh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Seated Forward Bend
Pose Type: seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)
लं रं यं

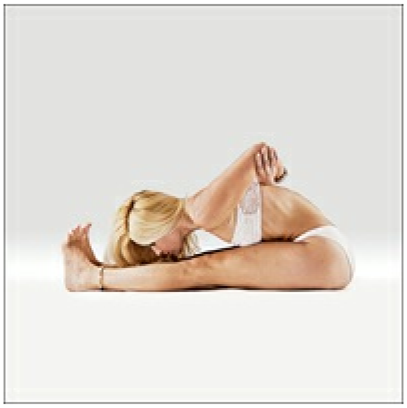

Reverse Prayer Western Intense Stretch Pose

Viparita Namaskar Paschimottanasana

(vi-puh-REE-tuh nuh-muhs-KAHR puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Back of the Body Prayer Western Intense Stretch Pose (Paschima Namaskara Paschimottanasana), Reverse Prayer Forward Bend

Pose Type: seated, forward bend

Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

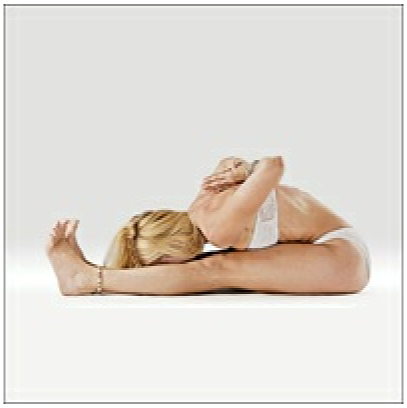

LEGS STRAIGHT IN FRONT: TWISTS

Revolved Staff Pose

Parivritta Dandasana

(puh-ri-VRIT-tuh duhn-DAHS-uh-nuh)

Pose Type: seated, twist

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

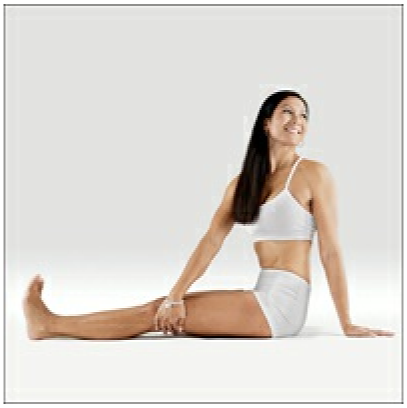

Revolved Upward One Arm Extended Hand to Foot Staff Pose

Parivritta Urdhva Eka Hasta Utthita Hasta Pada Dandasana

(puh-ri-VRIT-tuh OORD-vuh EY-kuh HUH-stuh UT-tu-tuh HUH-stuh PUH-duh duhn-DAHS-uh-nuh)

Pose Type: seated, forward bend, twist

Drishti Point: Hastagrai or Hastagre (hands)

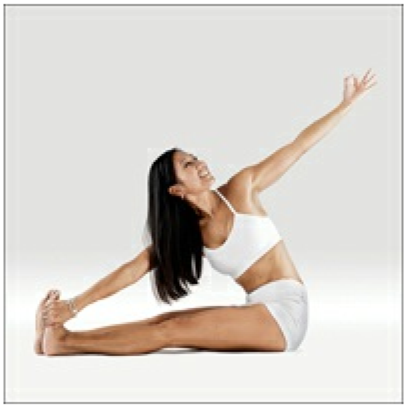

One Hand Revolved Western Intense Stretch Pose

Eka Hasta Parivritta Paschimottanasana

(EY-kuh HUH-stuh puh-ri-VRIT-tuh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: One Hand Revolved Seated Forward Bend

Modification: one hand to the foot

Pose Type: seated, forward bend, side bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

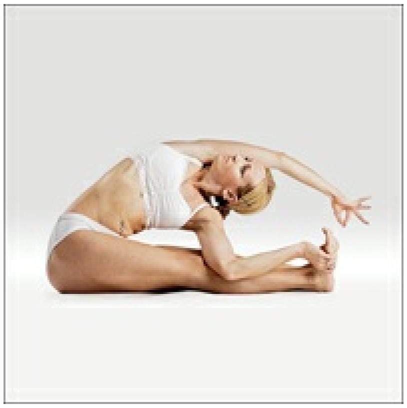

Two Hands Revolved Western Intense Stretch Pose

Dwi Hasta Parivritta Paschimottanasana

(DWI-huh-stuh puh-ri-VRIT-tuh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Two Hands Revolved Forward Bend

Pose Type: seated, forward bend, side bend, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

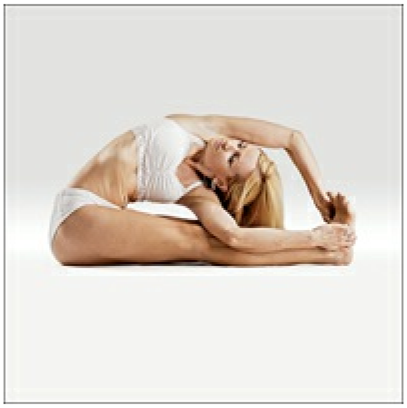

LEGS STRAIGHT: ON THE SIDE & SCISSOR LEGS

Upward Side Infinity Pose

Urdhva Parshva Anantasana

(OORD-vuh PARSH-vuh uh-nuhn-TAHS-uh-nuh)

Pose Type: seated, side bend

Drishti Point: Urdhva or Antara Drishti (up to the sky)

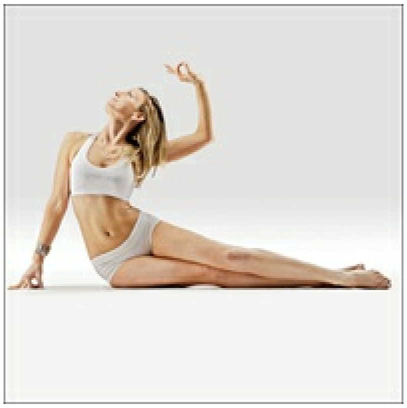

One Leg Upward Hand to Knee Revolved Staff Pose

Urdhva Eka Pada Janu Hasta Parivritta Dandasana

(OORD-vuh EY-kuh PUH-duh JAH-nu HUH-stuh puh-ri-VRIT-tuh duhn-DAHS-uh-nuh)

Pose Type: seated, forward bend, twist

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

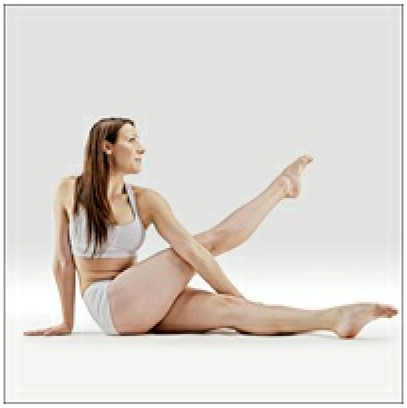

One Leg Upward Hand to Foot Revolved Staff Pose

Urdhva Eka Pada Hasta Pada Parivritta Dandasana

(OORD-vuh EY-kuh PUH-duh HUH-stuh PUH-duh puh-ri-VRIT-tuh duhn-DAHS-uh-nuh)

Pose Type: seated, forward bend, twist
Drishti Point: Parsva Drishti (to the right), Parsva Drishti (to the left)

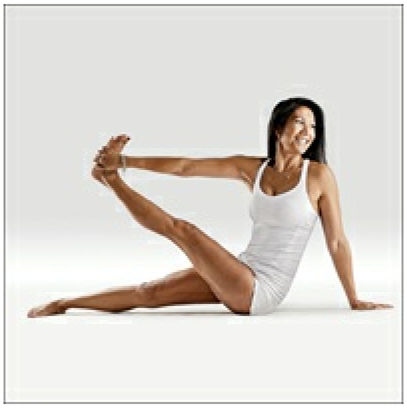

One Hand to Foot Archer’s Pose

Eka Hasta Pada Akarna Dhanurasana

(EY-kuh HUH-stuh PUH-duh AH-kuhr-nuh duh-nur-AHS-uh-nuh)

Modification: both legs straight, knee behind the shoulder, other palm to the floor
Pose Type: seated, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

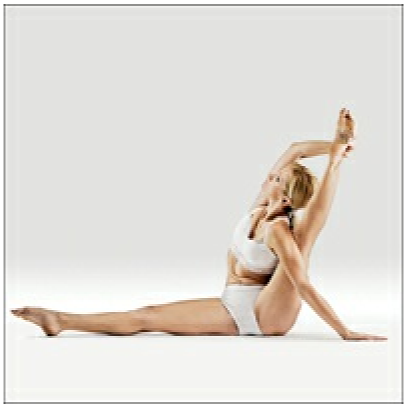

Bound Leg Archer’s Pose

Baddha Pada Akarna Dhanurasana

(BUH-duh PUH-duh AH-kuhr-nuh duh-nur-AHS-uh-nuh)

Pose Type: seated, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)
लं वं यं

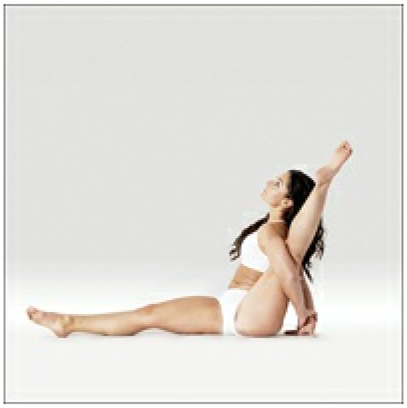

ONE LEG STRAIGHT, ONE LEG BENT

Archer’s Pose Prep.

Akarna Dhanurasana Prep.

(AH-kuhr-nuh duh-nur-AHS-uh-nuh)

Modification: knee wrapping around the arm
Pose Type: seated, mild backbend
Drishti Point: Urdhva or Antara Drishti (up to the sky)
लं वं हं

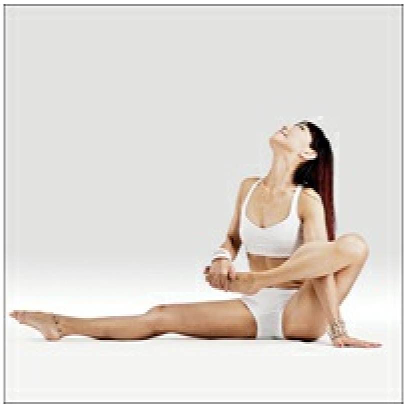

Archer’s Pose

Akarna Dhanurasana

(AH-kuhr-nuh duh-nur-AHS-uh-nuh)

Modification: foot to the ear
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

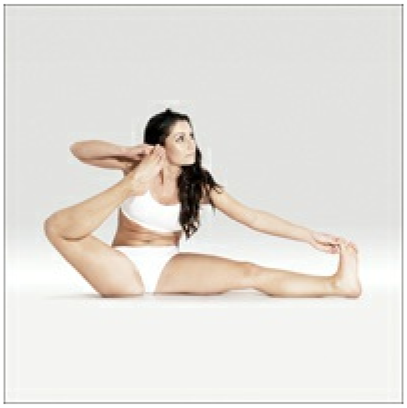

Foot Behind the Head Pose

Eka Pada Shirshasana

(EY-kuh PUH-duh sheer-SHAHS-uh-nuh)

Also Known As: Foot Behind the Head Pose A (Eka Pada Shirshasana A)
Modification: leg straight
1. palms to the floor by the hips
2. hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

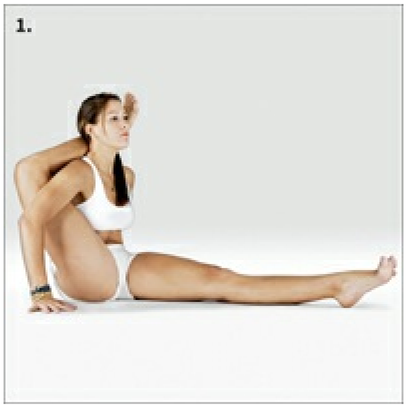
1.

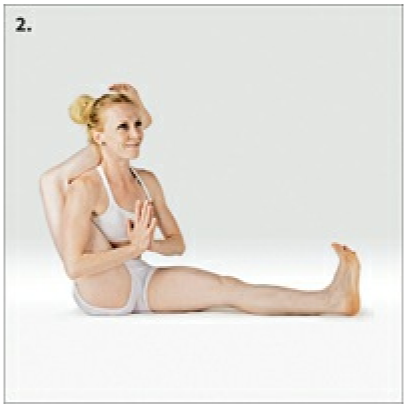
2.

Pose Dedicated to Skanda

Skandasana

(skuhn-DAHS-uh-nuh)

Also Known As: Foot Behind the Head Pose B (Eka Pada Shirshasana B)
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

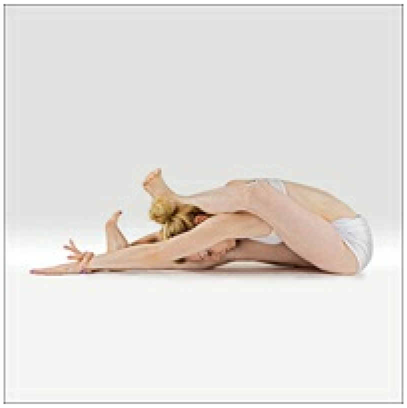

ONE LEG STRAIGHT, ONE LEG BENT: TWISTS

Easy Lord of the Fishes Pose Prep.

Sukha Matsyendrasana Prep.

(SUK-uh muhts-yeyn-DRAHS-uh-nuh)

Modification: one leg straight, arm wrapped around the bent knee
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व र

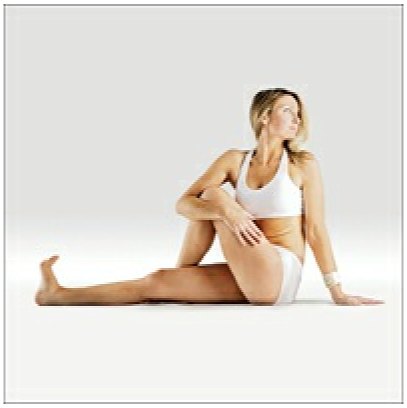

Easy Lord of the Fishes Pose

Sukha Matsyendrasana

(SUK-uh muhts-yeyn-DRAHS-uh-nuh)

Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व र

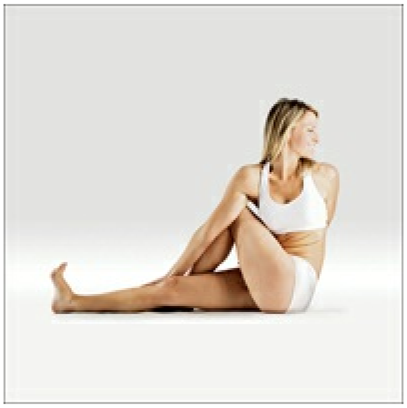

ONE LEG STRAIGHT, ONE LEG BENT: KNEE TO THE CHEST

Pose Dedicated to Sage Marichi 1 & 2 Prep.

Marichyasana 1 & 2 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi A & C Prep. (Marichyasana A & C Prep.)
Modification: arms parallel to the floor in front of the chest
Pose Type: seated, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

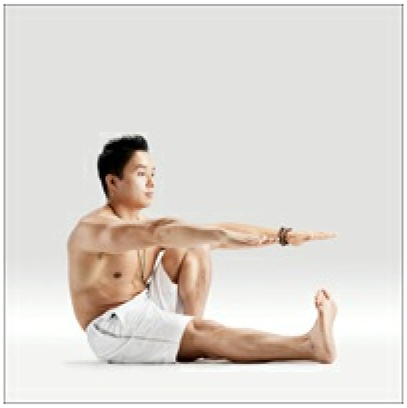

Pose Dedicated to Sage Marichi 1 & 2 Prep.

Marichyasana 1 & 2 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi A & C Prep. (Marichyasana A & C Prep.)
Modification: both hands grabbing onto the foot of the straight leg
Pose Type: seated, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

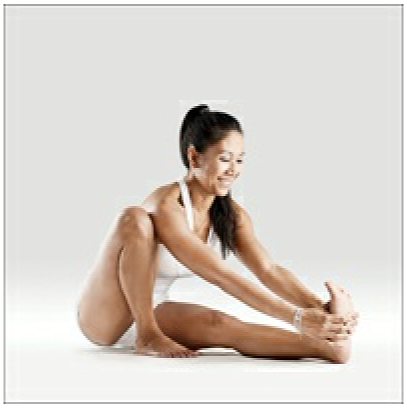

Revolved Pose Dedicated to Sage Marichi 1 Prep.

Parivritta Marichyasana 1 Prep.

(puh-ri-VRIT-tuh muh-ree-CHYAHS-uh-nuh)

Modification: twisting to the inside of the bent knee, elbow to the knee

Pose Type: seated, forward bend, twist

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

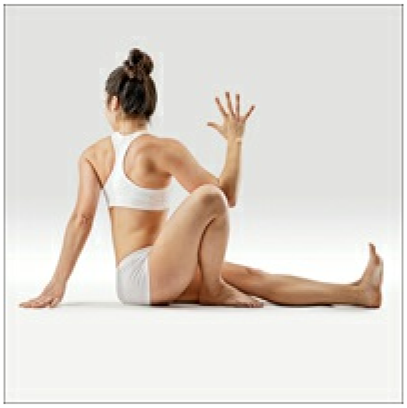

ONE LEG STRAIGHT, ONE LEG BENT: KNEE TO THE CHEST—BINDING & TWIST

Pose Dedicated to Sage Marichi 1

Marichyasana 1

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi A (Marichyasana A)

Modification: 1. spine straight
2. half forward bend
3. full forward bend, nose to the shin

Pose Type: seated, forward bend, binding

Drishti Point: 1. Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
2 & 3. Padhayoragrai or Padayoragre (toes/feet) or Nasagrai or Nasagre (nose)

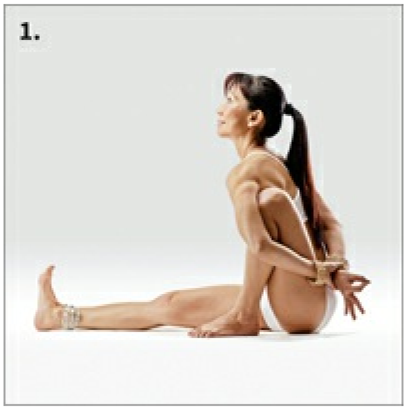

1.

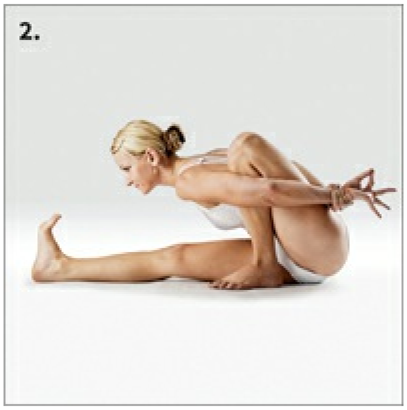

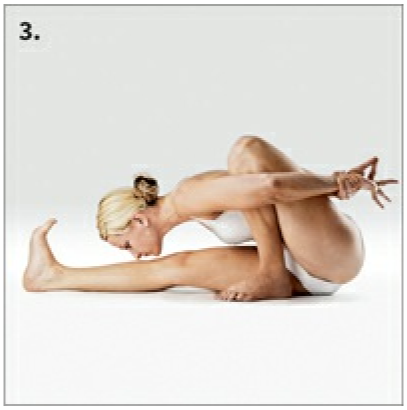

Pose Dedicated to Sage Marichi 2 Prep.

Marichyasana 2 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi C (Marichyasana C)
Modification: elbow over the bent knee
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

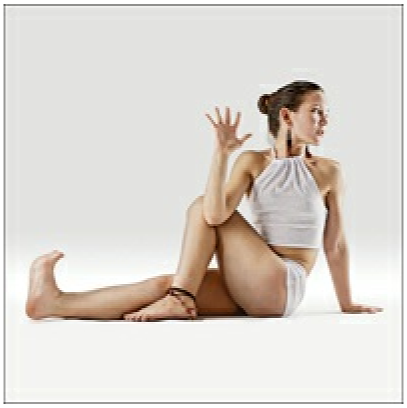

Pose Dedicated to Sage Marichi 2 Prep.

Marichyasana 2 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi C (Marichyasana C)
Modification: grabbing onto the knee of the straight leg
Pose Type: seated, forward bend, twist, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Padhayoragrai or Padayoragre (toes/feet)

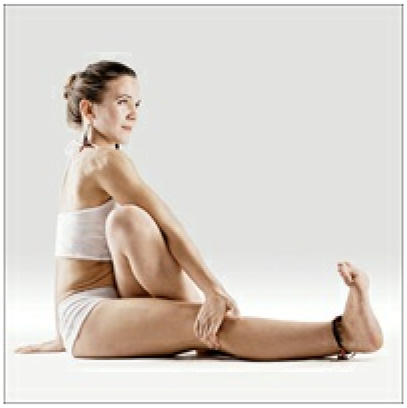

Pose dedicated to Sage Marichi 2

Marichyasana 2

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi C (Marichyasana C)
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

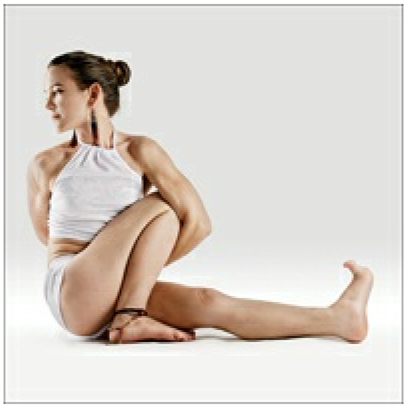

ONE LEG STRAIGHT & LIFTED, ONE LEG BENT

Seated Revolved Upward One Hand to Foot Pose

Upavishta Parivritta Urdhva Eka Pada Hastasana

(u-puh-VISH-tuh puh-ri-VRIT-tuh OORD-vuh EY-kuh PUH-duh huh-STAHS-uh-nuh)

Also Known As: Upavistha Parivritta Urdhva Eka Pada Hastasana
Modification: straight leg lifted, grabbing the foot with the opposite arm, other forearm on the floor
Pose Type: seated, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

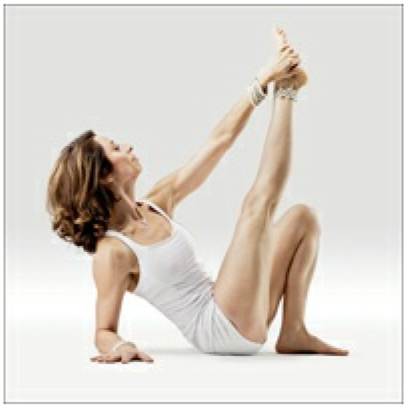

One Leg Extended Full Lord of the Fishes Pose

Utthita Eka Pada Paripurna Matsyendrasana

(UT-ti-tuh EY-kuh PUH-duh puh-ri-POOR-nuh muhts-yeyn-DRAHS-uh-nuh)

Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

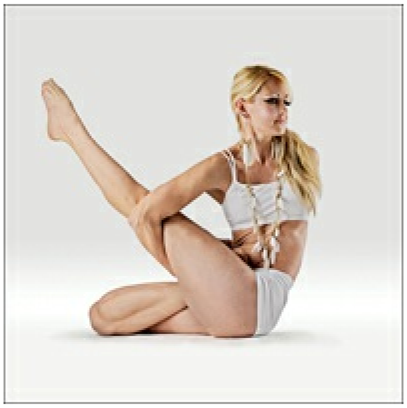

ONE LEG STRAIGHT, OTHER LEG CROSSED OVER: FOREARM & ELBOW TO THE FLOOR

Revolved Western Intense Stretch Pose Hand to Foot Side Infinity Pose

Parivritta Paschimottana Hasta Pada Parshva Anantasana

(puh-ri-VRIT-tuh puhsh-chi-mo-TAHN-uh HUH-stuh PUH-duh PAHRSH-vuh uh-nuhn-TAHS-uh-nuh)

Also Known As: Revolved Seated Forward Bend Hand to Foot Side Infinity Pose
Pose Type: seated, forward bend, twist
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

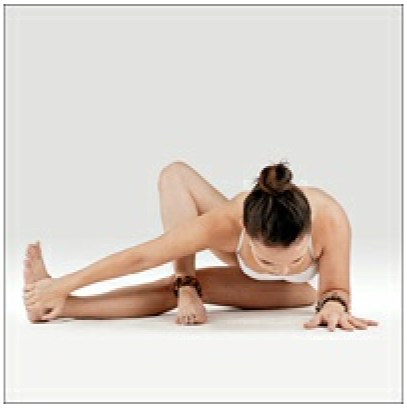

Revolved Western Intense Stretch Pose Uneven Arms Upward Side Infinity Pose

Parivritta Paschimottana Vishama Hasta Urdhva Parshva Anantasana

(puh-ri-VRIT-tuh puhsh-chi-mo-TAHN-uh VI-shuh-muh HUH-stuh OORD-vuh PAHRSH-vuh uh-nuhn-TAHS-uh-nuh)

Also Known As: Revolved Seated Forward Bend Uneven Arms Side Infinity Pose
Pose Type: seated, forward bend, twist
Drishti Point: Angushtamadhye or Angustha Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

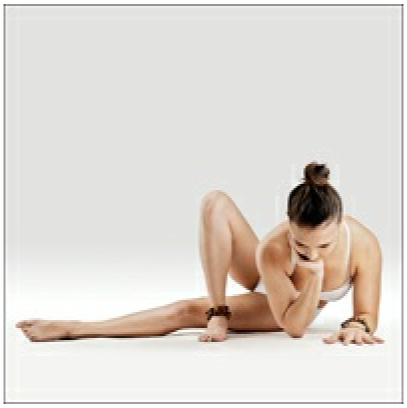

Revolved Western Intense Stretch Pose Bound Arms Upward Side

Infinity Pose

Parivritta Paschimottana Baddha Hasta Urdhva Parshva Anantasana

(puh-ri-VRIT-tuh puhsh-chi-mo-TAHN-uh BUH-duh HUH-stuh OORD-vuh PAHRSH-vuh uh-nuhn-TAHS-uh-nuh)

Also Known As: Revolved Seated Forward Bend Bound Arms Upward Side Infinity Pose
Modification: arms wrapped around the foot, elbows toward the floor
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व

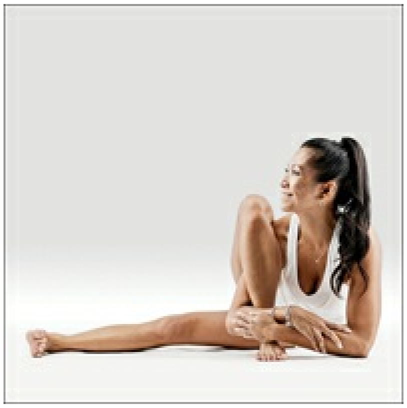

ONE LEG STRAIGHT, OTHER LEG CROSSED OVER: SIDE & BACKBEND

Upward Side Infinity Pose

Urdhva Parshva Anantasana

(OORD-vuh PAHRSH-vuh uh-nuhn-TAHS-uh-nuh)

Modification: top leg crossed over, outside edge of the foot to the floor
1. looking to the side
2. head rolling back
Pose Type: 1. seated, side bend, twist
2. seated, side bend, twist, backbend
Drishti Point: 1. Hastagrai or Hastagre (hands), Padhayoragrai or Padayoragre (toes/feet)
2. Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
1. ल व ह
2. ल व ह ॐ

1.

2.

HALF COW FACE POSE: FORWARD BEND & SIDE BEND

Half Cow Face Western Intense Stretch Pose Prep.

Ardha Gomukha Paschimottanasana Prep.

(UHR-duh GO-muk-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: One-Legged Cow Face Western Intense Stretch Pose (Eka Pada Gomukha Paschimottanasana), Half Cow Face Seated Forward Bend
Modification: half forward bend
Pose Type: seated, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

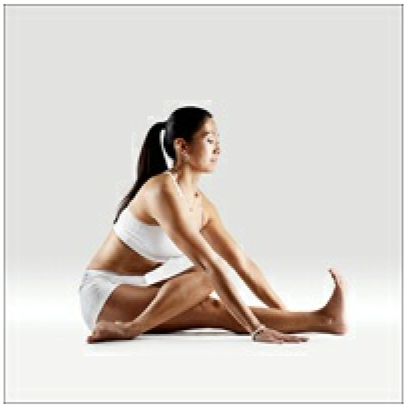

Half Cow Face Western Intense Stretch Pose

Ardha Gomukha Paschimottanasana

(UHR-duh GO-muk-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: One-Legged Cow Face Western Intense Stretch Pose (Eka Pada Gomukha Paschimottanasana), Half Cow Face Seated Forward Bend
Modification: 1. grabbing onto the wrist, chin to the knee
2. chin to the knee, arms straight, palms down in front of the foot
Pose Type: seated, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)
ल व र य

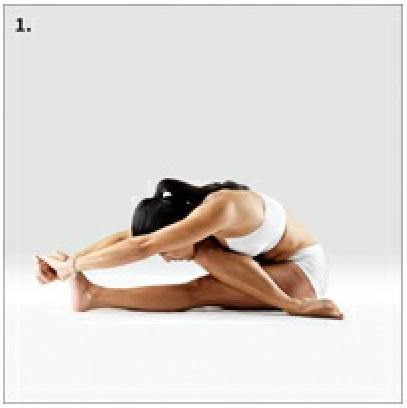

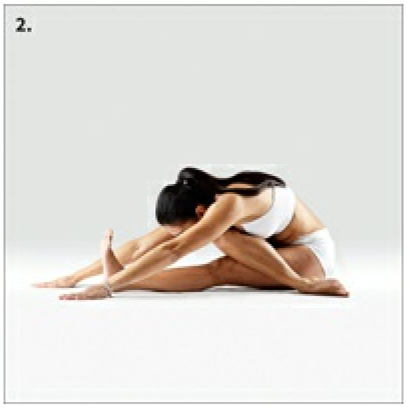

Revolved Sideways Half Cow Face Western Intense Stretch Pose

Parivritta Parshva Ardha Gomukha Paschimottanasana

(puh-ri-VRIT-tuh PAHRSH-vuh EY-kuh PUH-duh GO-muk-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Revolved Sideways One-Legged Cow Face Western Intense Stretch Pose (Parivritta Parshva Eka Pada Gomukha Paschimottanasana), Revolved Sideways Half Cow Face Seated Forward Bend
Modification: top leg straight
Pose Type: seated, forward bend, side bend, twist
Drishti Point: Padhayoragrai or Padayoragre (toes/feet), Hastagrai or Hastagre (hands), Parshva

Drishti (to the right), Parshva Drishti (to the left)

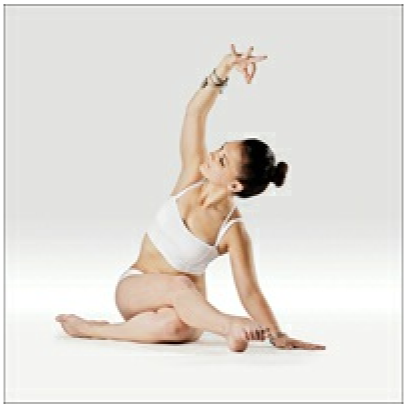

Revolved Hand to Foot Sideways Half Cow Face Western Intense Stretch Pose

Parivritta Hasta Pada Parshva Ardha Gomukha Paschimottanasana

(puh-ri-VRIT-tuh HUH-stuh PUH-duh PAHRSH-vuh UHR-duh GO-muk-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Revolved Hand to Foot Sideways One-Legged Cow Face Western Intense Stretch Pose (Parivritta Hasta Pada Parshva Eka Pada Gomukha Paschimottanasana), Revolved Hand to Foot Sideways Half Cow Face Seated Forward Bend
Modification: bottom leg straight, looking up to the sky
Pose Type: seated, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

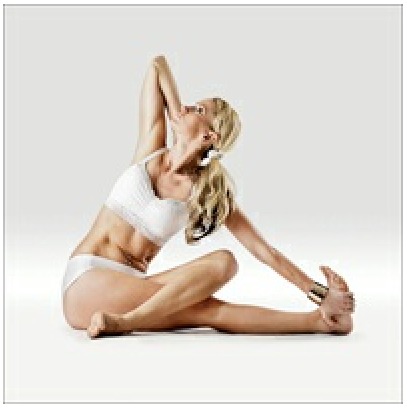

HALF COW FACE POSE: LEG LIFTED

Upward Half Cow Face Western Intense Stretch Pose

Urdhva Ardha Gomukha Paschimottanasana

(OORD-vuh UHR-duh GO-muk-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward One-Legged Cow Face Western Intense Stretch Pose (Urdhva Eka Pada Gomukha Paschimottanasana), Upward Half Cow Face Seated Forward Bend
Modification: bottom leg straight, grabbing onto the foot
Pose Type: seated, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Urdhva or Antara Drishti (up to the sky)

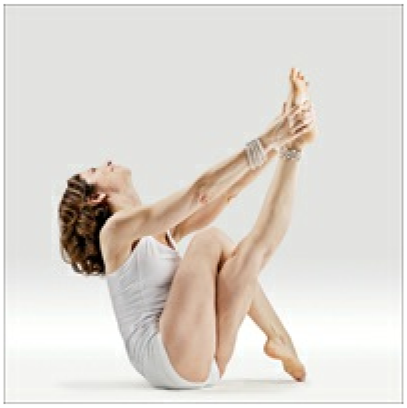

Upward Half Cow Face Western Intense Stretch Pose

Urdhva Ardha Gomukha Paschimottanasana

(OORD-vuh UHR-duh GO-muk-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward One-Legged Cow Face Western Intense Stretch Pose (Urdhva Eka Pada Gomukha Paschimottanasana), Upward Half Cow Face Seated Forward Bend
Modification: top leg straight, grabbing onto both feet
Pose Type: seated, forward bend, core
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

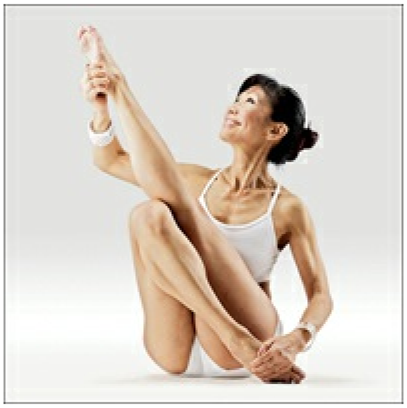

Sundial Pose 2

Surya Yantrasana 2

(SOOR-yuh yuhn-trahn-AHS-uh-nuh)

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व य

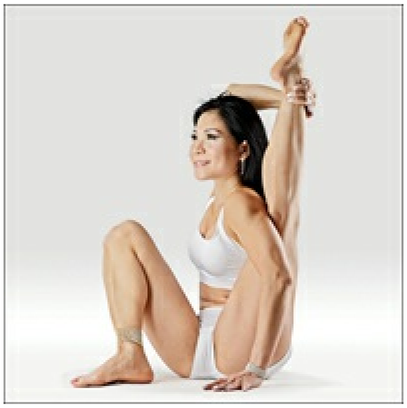

Sundial Pose 1

Surya Yantrasana 1

(SOOR-yuh yuhn-trahn-AHS-uh-nuh)

Pose Type: seated, twist

Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व य ह

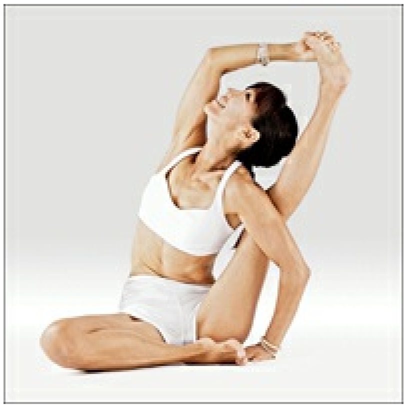

Upward Bound Hands Sundial Pose 1

Urdhva Baddha Hasta Surya Yantrasana 1

(OORD-vuh BUH-duh HUH-stuh SOOR-yuh yuhn-trahn-AHS-uh-nuh)

Pose Type: seated, twist, binding

Drishti Point: Urdhva or Antara Drishti (up to the sky)

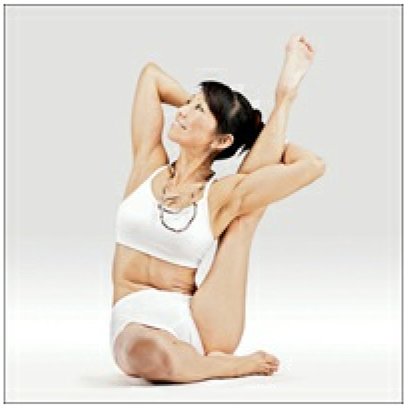

Downward Bound Hands Sundial Pose 1

Adho Baddha Hasta Surya Yantrasana 1

(uh-DO BUH-duh HUH-stuh SOOR-yuh yuhn-trahn-AHS-uh-nuh)

Pose Type: seated, forward bend, binding

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

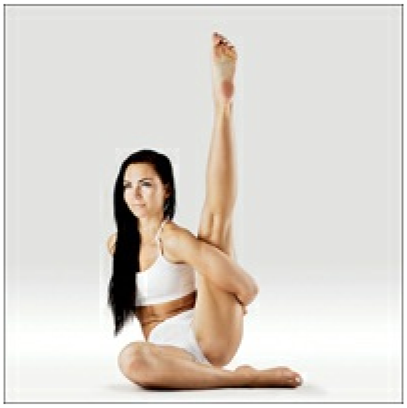

ONE LEG STRAIGHT, ONE LEG BENT: HALF LOTUS & BINDING

Half Bound Lotus Western Intense Stretch Pose Prep.

Ardha Baddha Padma Paschimottanasana Prep.

(UHR-duh BUH-duh PUHD-muh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Half Bound Lotus Seated Forward Bend Prep.
Modification: arm extended to the sky
Pose Type: seated, binding
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

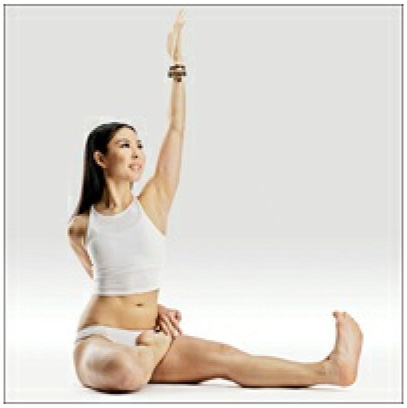

Half Bound Lotus Western Intense Stretch Pose

Ardha Baddha Padma Paschimottanasana

(UHR-duh BUH-duh PUHD-muh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Half Bound Lotus Seated Forward Bend
Modification: 1. half forward bend
2. full forward bend
Pose Type: seated, forward bend, binding
Drishti Point: Padhayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

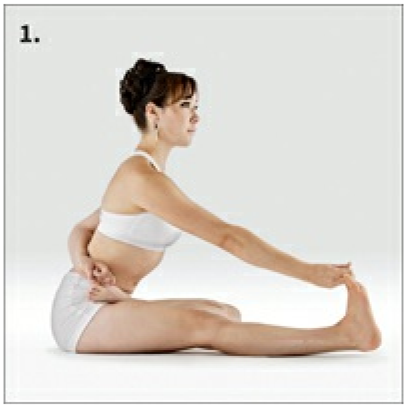

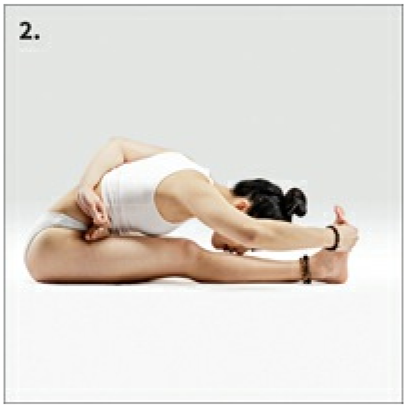

ONE LEG STRAIGHT, ONE LEG BENT: HALF LOTUS, TWISTS & BINDING

Revolved Half Bound Lotus Western Intense Stretch Pose

Parivritta Ardha Baddha Padma Paschimottanasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh PUHD-muh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Revolved Half Bound Lotus Seated Forward Bend and Half Lord of the Fishes Pose (Ardha Matsyendrasana)
Modification: grabbing onto the big toe
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व य

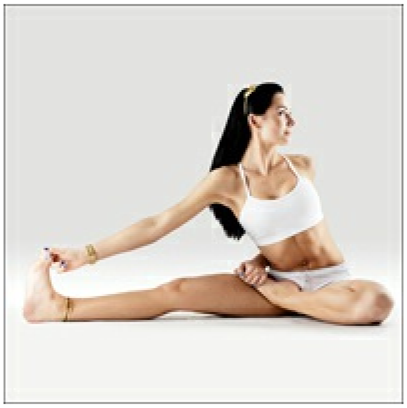

Half Lord of The Fishes Pose 2

Ardha Matsyendrasana 2

(UHR-duh muhts-yeyn-DRAHS-uh-nuh)

Also Known As: Sage Warrior Bharadvaja Prep. (Bharadvajasana Prep.)
Modification: 1. looking toward the foot
2. looking over the shoulder
Pose Type: seated, forward bend, twist, binding

Drishti Point: 1. Padhayoragrai or Padayoragre (toes/feet)
2. Parshva Drishti (to the right), Parshva Drishti (to the left)
लं वं यं

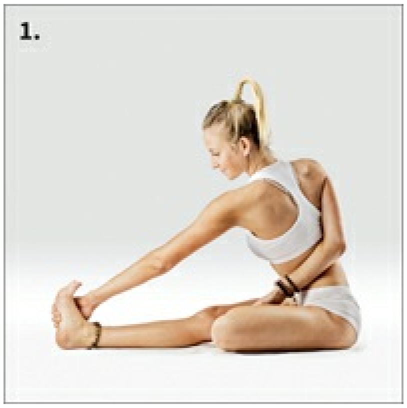
1.

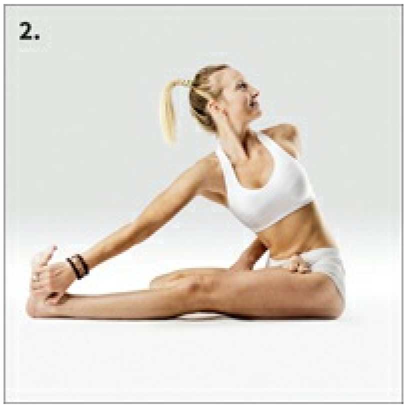
2.

ONE LEG STRAIGHT, ONE LEG BENT: HALF LOTUS—LEG LIFTED

Half Lotus Extended Hand to Big Toe Pose in Infinity Pose

Ardha Padma Utthita Hasta Padangusthasana in Anantasana

(UHR-duh PUHD-muh UT-ti-tuh HUH-stuh puhd-ahng-goosh-TAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Pose Type: seated, mild backbend
Drishti Point: Urdhva or Antara Drishti (up to the sky), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं हं ॐ

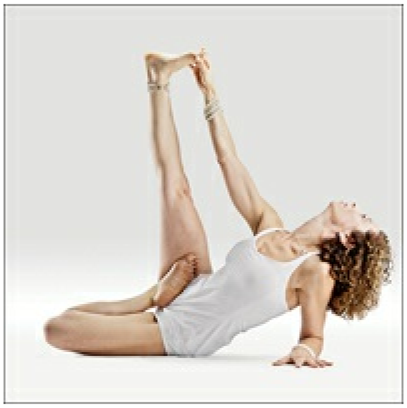

Half Lotus Extended Revolved Hand to Foot Pose in Infinity Pose

Ardha Padma Utthita Parivritta Pada Hastasana in Anantasana

(UHR-duh PUHD-muh UT-ti-tuh puh-ri-VRIT-tuh PUH-duh huh-STAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Modification: grabbing onto the foot on the opposite side
Pose Type: seated, mild backbend, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

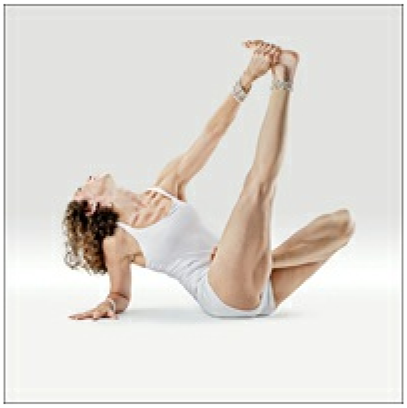

ONE LEG STRAIGHT, ONE LEG BENT: HALF LOTUS & BACKBEND—BINDING

Half Lotus Upward Infinity Pose

Ardha Padma Urdhva Anantasana

(UHR-duh PUHD-muh OORD-vuh uhn-uhnt-AHS-uh-nuh)

Pose Type: seated
Drishti Point: Urdhva or Antara Drishti (up to the sky), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

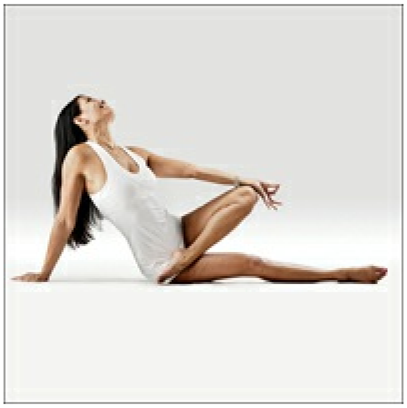

Half Bound Lotus Infinity Pose

Ardha Baddha Padma Anantasana

(UHR-duh BUH-duh PUHD-muh uhn-uhnt-AHS-uh-nuh)

Modification: Arm 1: forearm to the floor
Arm 2: grabbing onto the foot
Pose Type: seated, mild backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

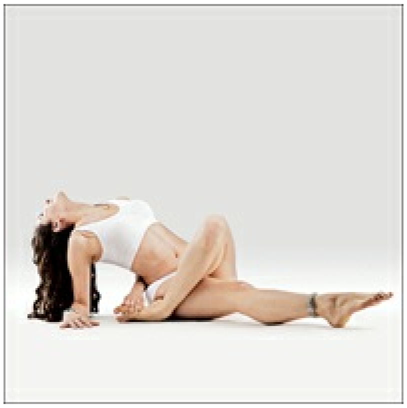

Half Bound Lotus Infinity Pose

Ardha Baddha Padma Anantasana

(UHR-duh BUH-duh PUHD-muh uhn-uhnt-AHS-uh-nuh)

Modification: Arm 1: elbow to the floor
Arm 2: hand to the hip socket
Pose Type: seated, mild backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

ONE LEG STRAIGHT, ONE KNEE BENT TO THE BACK: FORWARD BEND & TWIST

Three Limbed Face to Foot Western Intense Stretch Pose

Trianga Mukhaikapada Paschimottanasana

(tri-UHNG-uh muk-EYE-kuh-puh-duh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Three Limbed Face to Foot Seated Forward Bend
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

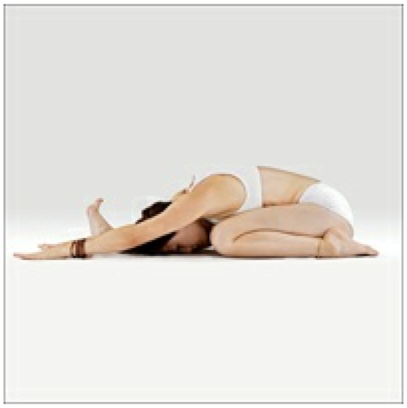

Churning Pose

Chalanasana

(chuh-luh-NAHS-uh-nuh)

Pose Type: seated, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

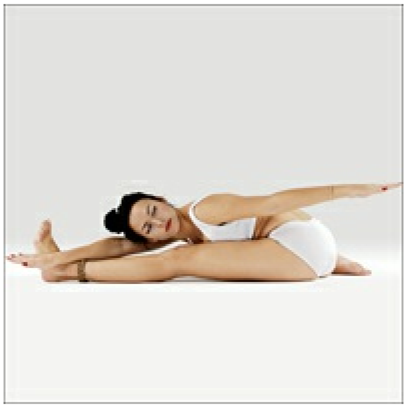

ONE LEG STRAIGHT, ONE KNEE BENT TO THE BACK: TWIST, SIDE BEND & FORWARD BEND

Revolved Half Bound Half Hero Half Seated Angle Pose

Parivritta Ardha Baddha Ardha Vira Ardha Upavishta Konasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh UHR-duh VEER-uh UHR-duh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Parivritta Ardha Baddha Ardha Vira Ardha Upavistha Konasana
Konasana Modification: spine straight
Pose Type: seated, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व

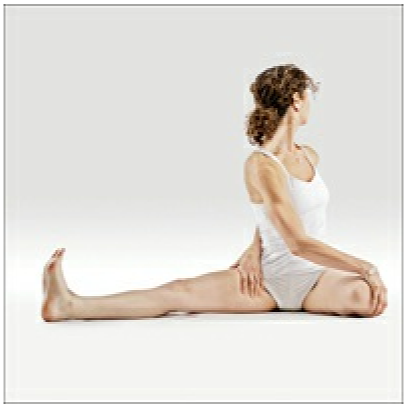

Gate Pose

Parighasana

(puh-ri-GAHS-uh-nuh)

Modification: side bend toward the straight leg; grabbing onto the heel of the bent leg with both hands
Pose Type: seated, side bend, forward bend, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व

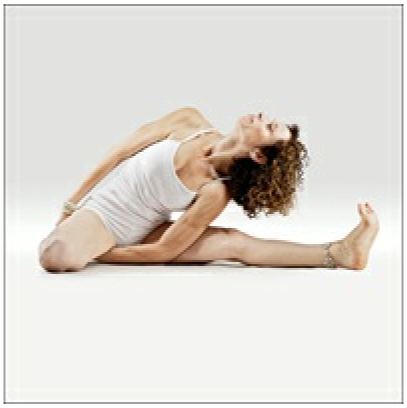

Gate Pose

Parighasana

(puh-ri-GAHS-uh-nuh)

Modification: seated: 1. palm to the floor by the ankle, other arm up over the head
2. elbow to the floor by the knee
3. both hands grabbing onto the foot, chest rotated up to the sky
Pose Type: 1 & 2. seated, side bend
3. seated, side bend, twist
Drishti Point: 1. Hastagrai or Hastagre (hands)
2 & 3. Urdhva or Antara Drishti (up to the sky)
ल व य

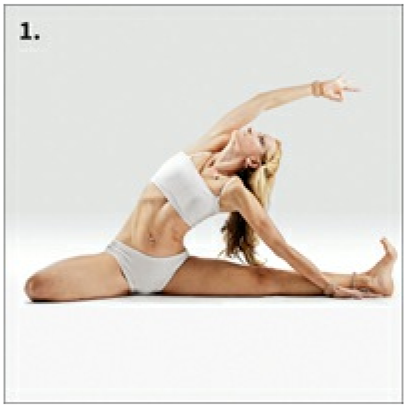

2.

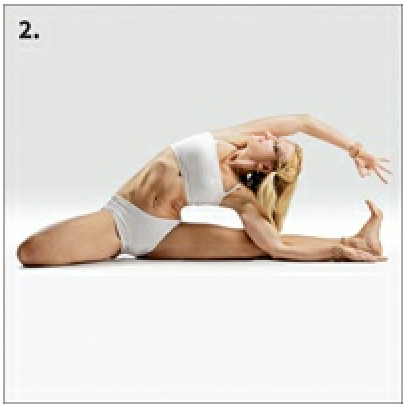

3.

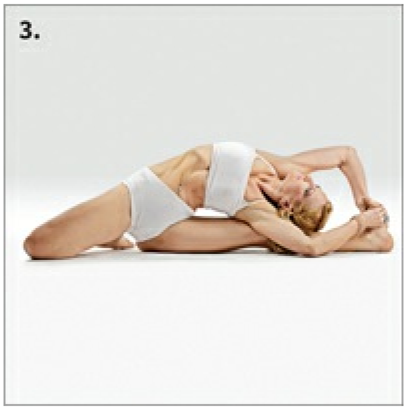

Benu Bird Pose 1

Benvasana 1

(ben-VAHS-uh-nuh)

Modification: seated; sitting bone on the heel, forward bend; arms to the side and to the back
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य ह

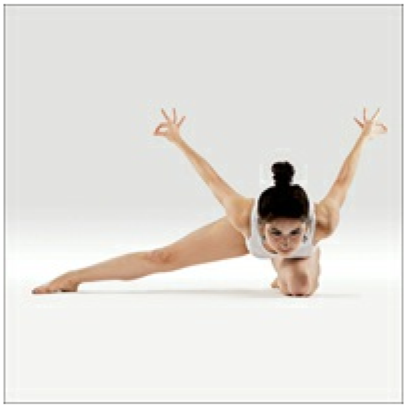

ONE LEG STRAIGHT, ONE KNEE BENT TO THE BACK: LEG LIFTED—BACKBEND

Reclined Half Hero Extended Hand to Foot Pose

Supta Ardha Vira Utthita Hasta Padasana

(SUP-tuh UHR-duh VEER-uh UT-ti-tuh HUH-stuh PUH-duhs-uh-nuh)

Modification: 1. right side view
2. left side view
Pose Type: seated, mild backbend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

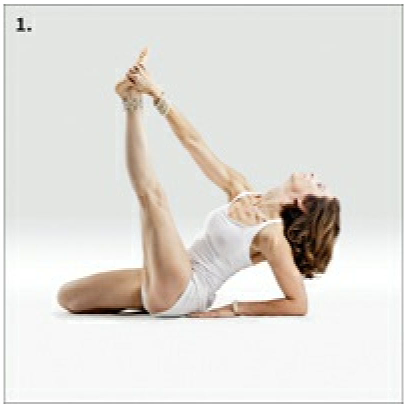

2.

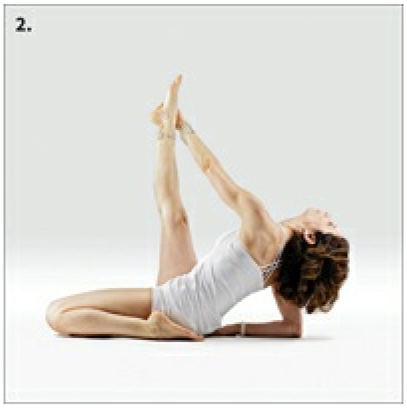

ONE LEG STRAIGHT, ONE KNEE BENT TO THE BACK: LEG LIFTED

Heron Pose

Krounchasana

(crown-CHAHS-uh-nuh)

Modification: grabbing onto the wrist, toes pointed
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

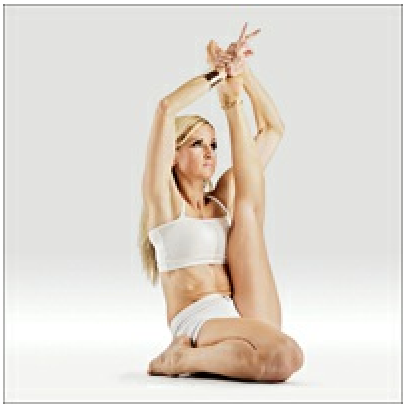

Sundial Pose 3

Surya Yantrasana 3

(SOOR-yuh yuhn-TRAHS-uh-nuh)

Modification: foot turned to the back
Pose Type: seated, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

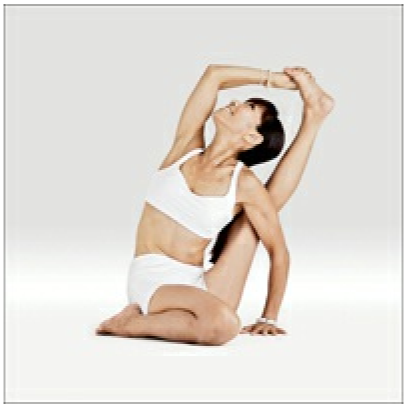

Revolved Heron Pose

Parivritta Krounchasana

(puh-ri-VRIT-tuh crown-CHAHS-uh-nuh)

Also Known As: Revolved Sundial Pose 3 (Parivritta Surya Yantrasana 3)
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

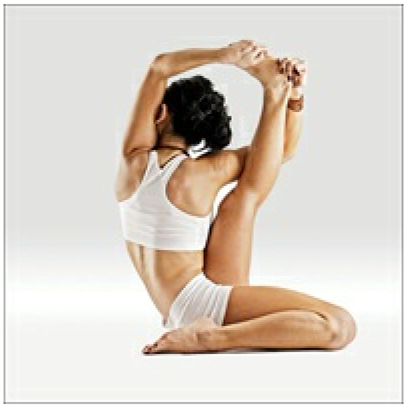

GARUDA LEGS: BACKBEND, TWIST, FORWARD BEND & BINDING

Seated Leg Position of the Pose Dedicated to Garuda

Upavishta Pada Garudasana

(u-puh-VISH-tuh PUH-duh guh-ru-DAHS-uh-nuh)

Also Known As: Upavistha Pada Garudasana
Modification: backbend, one hand to the sky
Pose Type: seated, backbend
Drishti Point: Hastagrai or Hastagre (hands)

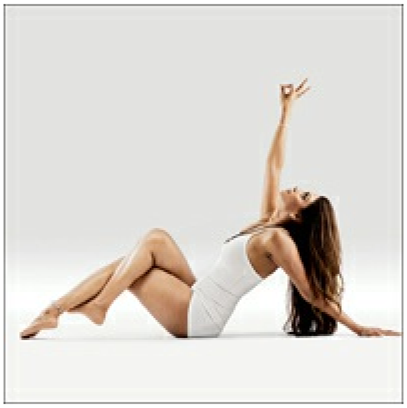

Revolved Hand to Foot Seated Leg Position of the Pose Dedicated to Garuda

Parivritta Hasta Pada Upavishta Pada Garudasana

(puh-ri-VRIT-tuh HUH-stuh PUH-duh u-puh-VISH-tuh PUH-duh guh-ru-DAHS-uh-nuh)

Also Known As: Parivritta Hasta Pada Upavistha Pada Garudasana
Pose Type: seated, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

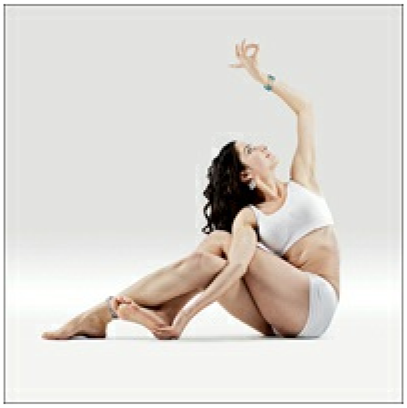

Seated Leg Position of the Pose Dedicated to Garuda

Upavishta Pada Garudasana

(u-puh-VISH-tuh PUH-duh guh-ru-DAHS-uh-nuh)

Also Known As: Upavistha Pada Garudasana
Modification: forehead to the knee
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

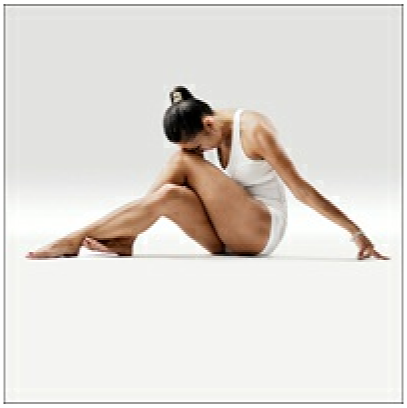

Revolved Seated Leg Position of the Pose Dedicated to Garuda

Parivritta Upavishta Pada Garudasana

(puh-ri-VRIT-tuh u-puh-VISH-tuh PUH-duh guh-ru-DAHS-uh-nuh)

Also Known As: Parivritta Upavistha Pada Garudasana
Modification: hands in Anjali Mudra (Hands in Prayer), one arm threaded through the legs
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

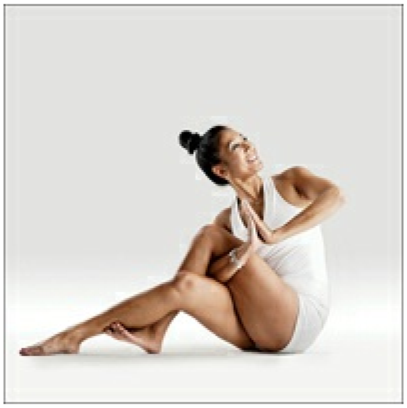

Bound Revolved Seated Leg Position of the Pose Dedicated to Garuda

Baddha Parivritta Upavishta Pada Garudasana

(BUH-duh puh-ri-VRIT-tuh u-puh-VISH-tuh PUH-duh guh-ru-DAHS-uh-nuh)

Also Known As: Baddha Parivritta Upavistha Pada Garudasana
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

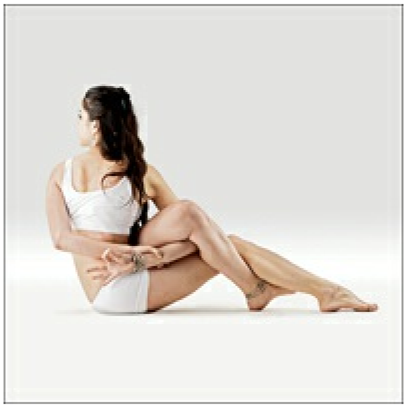

GARUDA LEGS: KNEES TO THE FLOOR ON THE SIDE

Seated Leg Position of the Pose Dedicated to Garuda in Upward Side Infinity Pose

Upavishta Pada Garudasana in Urdhva Parshva Anantasana

(u-puh-VISH-tuh PUH-duh guh-ru-DAHS-uh-nuh in OORD-vuh PAHRSH-vuh uhn-uhnt-AHS-uh-nuh)

Also Known As: Upavistha Pada Garudasana in Urdhva Parshva Anantasana
Modification: one hand to the floor, one arm extended out
Pose Type: seated, side bend
Drishti Point: Hastagrai or Hastagre (hands)

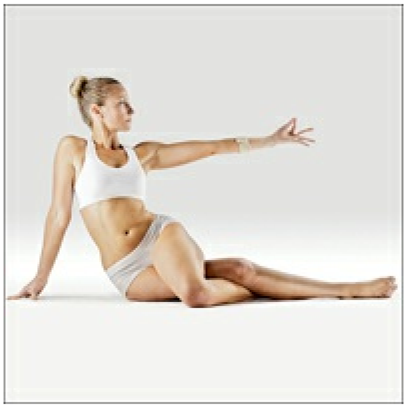

Seated Leg Position of the Pose Dedicated to Garuda in Upward Side Infinity Pose

Upavishta Pada Garudasana in Urdhva Parshva Anantasana

(u-puh-VISH-tuh PUH-duh guh-ru-DAHS-uh-nuh in OORD-vuh PAHRSH-vuh uhn-uhnt-AHS-uh-nuh)

Modification: palms flat on the floor, fingertips pointing away from the heels, backbend

Pose Type: seated, side bend, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

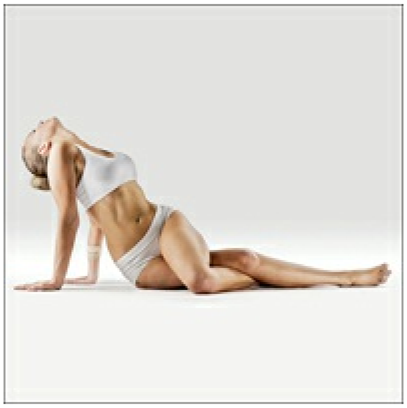

BOTH KNEES BENT: ANKLES HOOKED—TWIST & SIDE BEND

Twist Dedicated to Sage Bharadvaja 3

Bharadvajasana 3

(buh-ruhd-vahj-AHS-uh-nuh)

Pose Type: seated, forward bend, twist

Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

Sideways Twist Dedicated to Sage Bharadvaja 3

Parshva Bharadvajasana 3

(PAHRSH-vuh buh-ruhd-vahj-AHS-uh-nuh)

Modification: forearm to the floor
Pose Type: seated, mild backbend, side bend
Drishti Point: Hastagrai or Hastagre (hands)

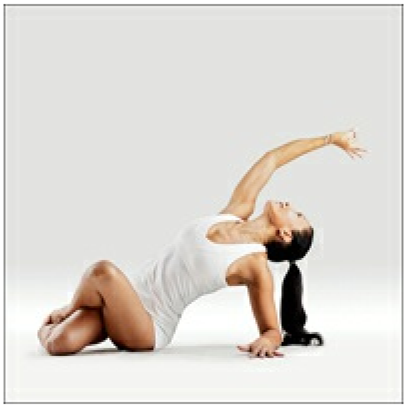

LEG CRADLE: ONE LEG STRAIGHT

Baby Cradle Pose

Hindolasana

(hin-do-LAHS-uh-nuh)

Also Known As: One Foot Behind the Head Pose A & B Prep. (Eka Pada Shirshasana A & B Prep.)
Modification: leg straight
1. spine straight, front view
2. half forward bend, side view

Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Padhayoragrai or Padayoragre (toes/feet), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

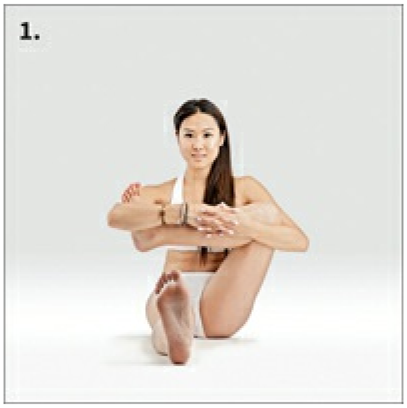

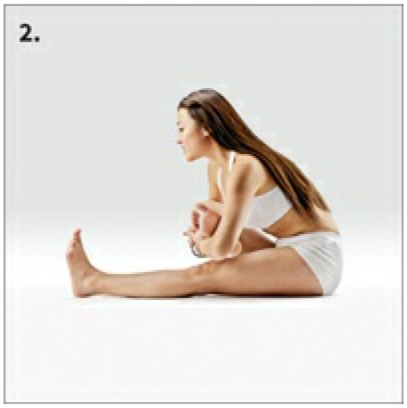

Baby Cradle Pose

Hindolasana

(hin-do-LAHS-uh-nuh)

Also Known As: One Foot Behind the Head Pose A Prep. (Eka Pada Shirshasana A Prep.)
Modification: leg straight, hands to the face
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

HEAD TO KNEE POSE: FORWARD BEND

Head to Knee Pose Prep.

Janu Shirshasana Prep.

(JAH-nu sheer-SHAHS-uh-nuh)

Modification: palms to the floor by the hips, spine straight
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose), Padhayoragrai or Padayoragre (toes/feet)

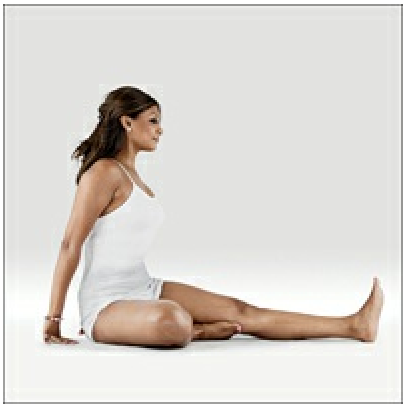

Both Hands to Ankle Head to Knee Pose

Dwi Hasta Kulpa Janu Shirshasana

(DWI-huh-stuh KUL-puh JAH-nu sheer-SHAHS-uh-nuh)

Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Padhayoragrai or Padayoragre (toes/feet)

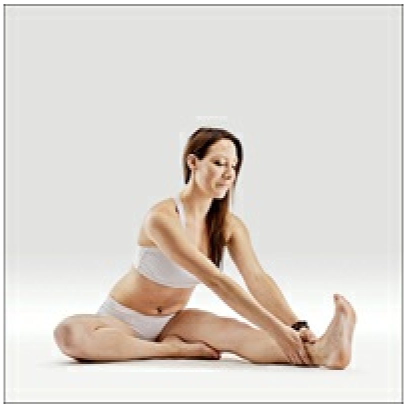

Both Hands to Foot Head to Knee Pose

Dwi Hasta Pada Janu Shirshasana

(DWI-huh-stuh PUH-duh JAH-nu sheer-SHAHS-uh-nuh)

Modification: 1. grabbing onto the foot; fingers interlocked, half forward bend
2. grabbing onto the wrist, half forward bend
Pose Type: seated, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)
ल व

1.

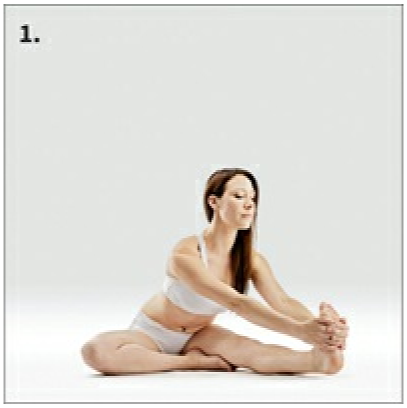

2.

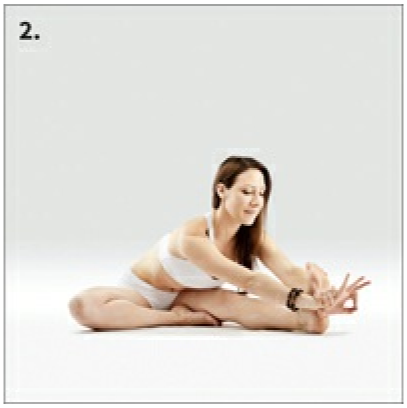

HEAD TO KNEE POSE: HEEL UP—FORWARD BEND

Head to Knee Pose C Prep.

Janu Shirshasana C Prep.

(JAH-nu sheer-SHAHS-uh-nuh)

Also Known As: Half Root Lock Pose (Ardha Mula Bandhasana)
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose), Padhayoragrai or Padayoragre (toes/feet), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

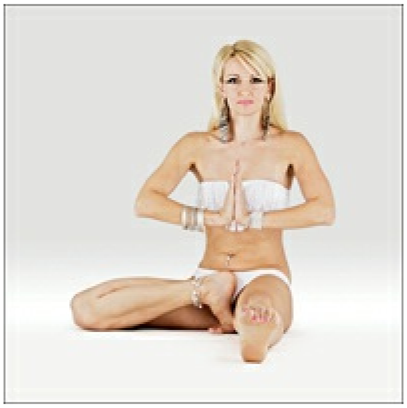

Head to Knee Pose C

Janu Shirshasana C

(JAH-nu sheer-SHAHS-uh-nuh)

Modification: 1. half forward bend
2. full forward bend
Pose Type: seated, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

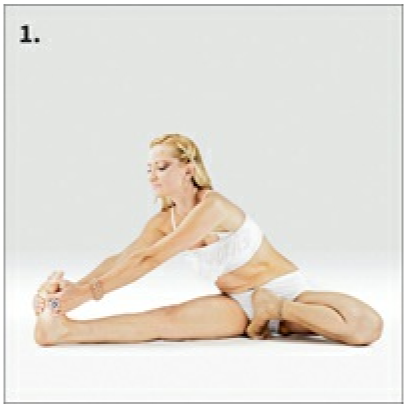

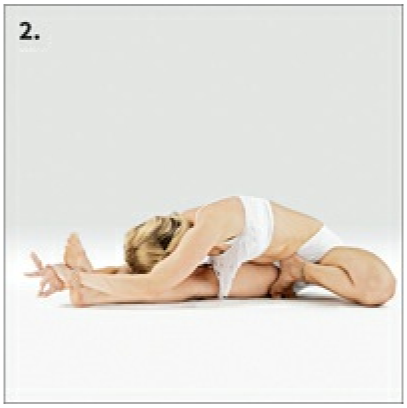

HEAD TO KNEE POSE: TWISTS

Revolved Half Bound Angle Pose

Parivritta Ardha Baddha Konasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh ko-NAHS-uh-nuh)

Modification: hand to the knee, other hand to the floor behind the hips for support

Pose Type: seated, twist

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

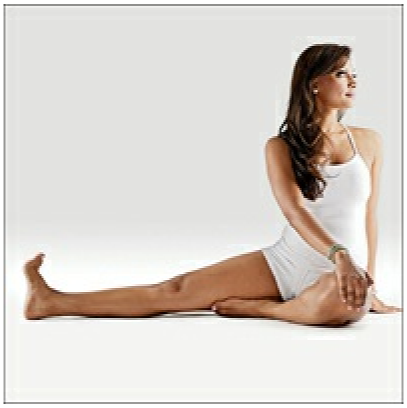

Revolved Half Bound Angle Pose

Parivritta Ardha Baddha Konasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh ko-NAHS-uh-nuh)

Pose Type: seated, twist, binding

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

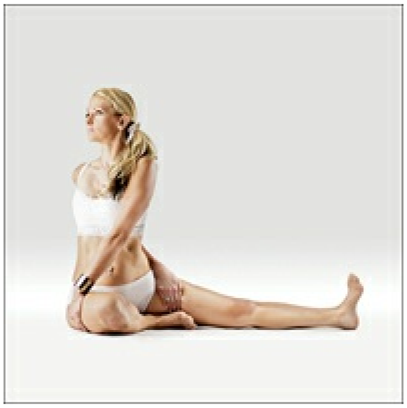

Revolved Sideways Half Bound Angle Pose

Parivritta Parshva Ardha Baddha Konasana

(puh-ri-VRIT-tuh PAHRSH-vuh UHR-duh BUH-duh ko-NAHS-uh-nuh)

Modification: grabbing onto the big toe of the bent leg
Pose Type: seated, twist, side bend, binding
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

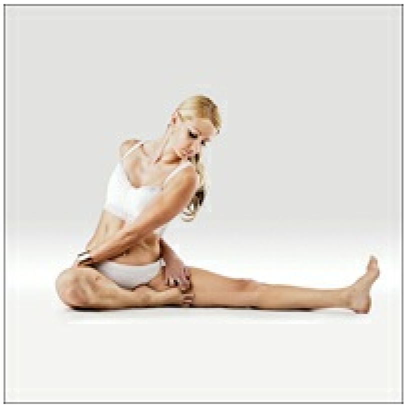

Revolved Half Bound Angle Pose

Parivritta Ardha Baddha Konasana

(puh-ri-VRIT-tuh UHR-duh BUH-duh ko-NAHS-uh-nuh)

Modification: looking back Arm 1: grabbing onto the outside edge of the opposite foot
Arm 2: grabbing onto the shin of the opposite leg behind the back
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

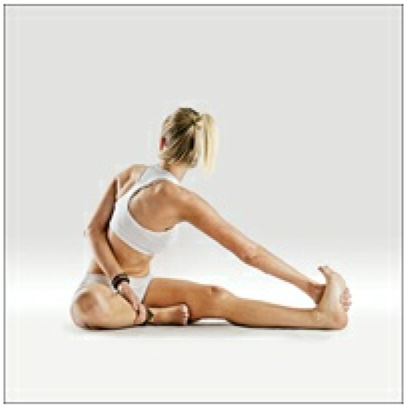

HEAD TO KNEE POSE—SIDE BENDS

Half Bound Angle Pose in Infinity Pose

Ardha Baddha Konasana in Anantasana

(UHR-duh BUH-duh ko-NAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Modification: elbow to the floor, head rolling back
Pose Type: seated, mild backbend, side bend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

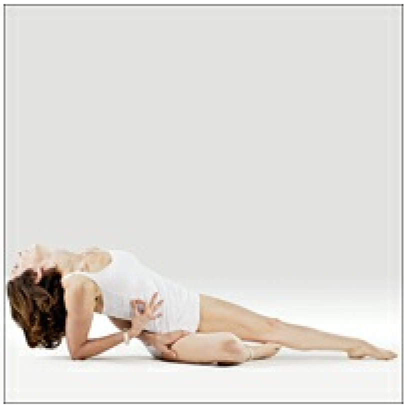

Revolved Head to Knee Pose Prep.

Parivritta Janu Shirshasana Prep.

(puh-ri-VRIT-tuh JAH-nu sheer-SHAHS-uh-nuh)

Modification: grabbing onto the foot, other hand to the knee
Pose Type: seated, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)
ल व

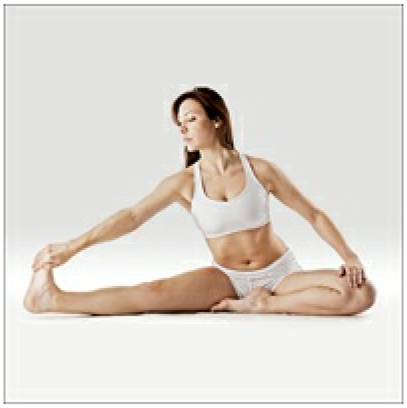

Sideways Head to Knee Pose

Parshva Janu Shirshasana

(PAHRSH-vuh JAH-nu sheer-SHAHS-uh-nuh)

Also Known As: Sideways Half Bound Angle Pose (Parshva Ardha Baddha Konasana)
Pose Type: seated, side bend
Drishti Point: Hastagrai or Hastagre (hands)

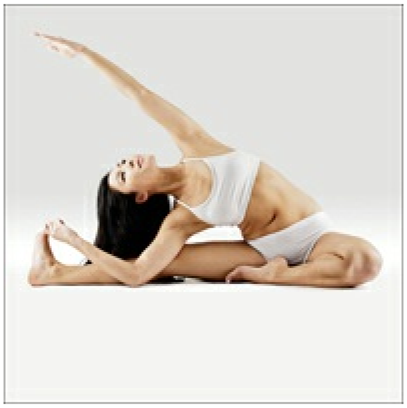

Sideways Head to Knee Pose

Parshva Janu Shirshasana

(PAHRSH-vuh JAH-nu sheer-SHAHS-uh-nuh)

Also Known As: Sideways Half Bound Angle Pose (Parshva Ardha Baddha Konasana)
Modification: elbow to the floor on the inside of the leg, other hand on the hip
Pose Type: seated, forward bend, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

Half Bound

Half Seated Angle Pose

Ardha Baddha Ardha Upavishta Konasana

(UHR-duh BUH-duh UHR-duh u-puh-VISH-tuh ko-NAHS-uh-nuh)

Also Known As: Ardha Baddha Ardha Upavistha Konasana

Pose Type: seated, side bend, binding

Drishti Point: Nasagrai or Nasagre (nose)

How to Perform the Pose:

1. Begin by sitting on the floor with both your legs straight out in front of you. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale as you bend your left knee and bring the sole of your left foot toward your right thigh.
3. On the next exhale, bring your right leg out to the right side, keeping it strong and straight by pulling up the kneecap and engaging the thigh muscles (quadriceps). Press your right toes and your right heel to the floor, lifting the right leg slightly on off the floor.
4. Inhale as you expand your chest and hold your arms straight out to the sides, parallel to the floor.
5. Exhale as you bend your left arm and reach your left hand behind the back to the inside of your right thigh to bind.
6. Inhale as you reach your right arm up to the sky. Exhale as you bend your right elbow and look toward your left knee.
7. Hold the pose for at least 30, and up to 90, seconds in order to receive full benefits of the stretch.
8. Inhale as you release the bind. Exhale as you bring both your legs straight out in front of you. Repeat on the other side.

Modification: side bend toward the bent knee

ardha = half

baddha = bound

ardha = half

upavishta = seated

kona = angle

Half Bound Sideways Head to Knee Pose

Ardha Baddha Parshva Janu Shirshasana

(UHR-duh BUH-duh PAHRSH-vuh JAH-nu sheer-SHAHS-uh-nuh)

Also Known As: Sideways Half Bound Angle Pose (Parshva Ardha Baddha Konasana)
Modification: 1. toes pointed
2. toes flexed in
Pose Type: seated, forward bend, side bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व

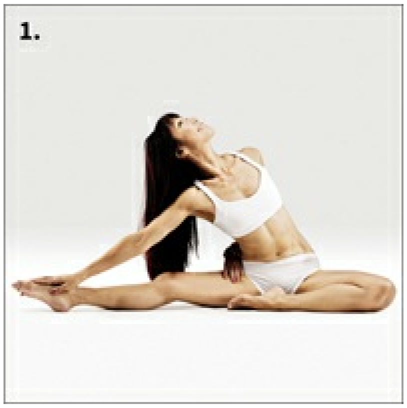
1.

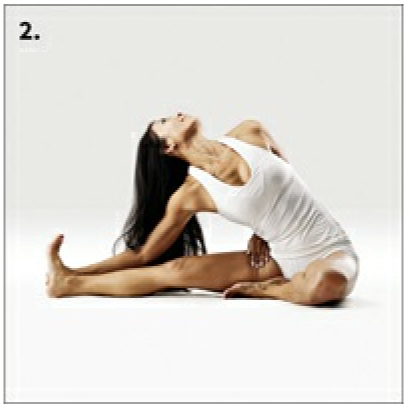
2.

Revolved Head to Knee Pose

Parivritta Janu Shirshasana

(puh-ri-VRIT-tuh JAH-nu sheer-SHAHS-uh-nuh)

Modification: elbow to the floor
1. forearm to the floor, chest to the side
2. head on the knee, chest to the sky
Pose Type: seated, side bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व र य

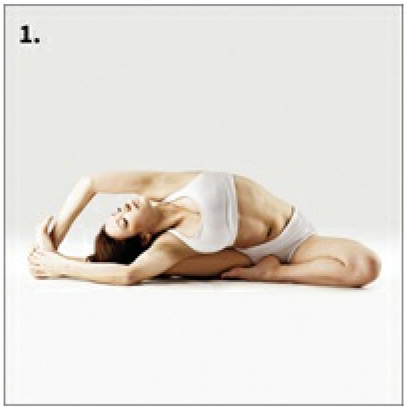

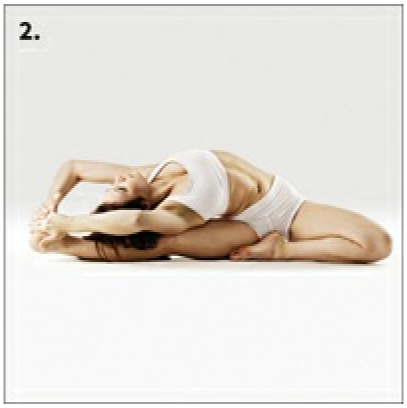

Both Hands to the Head Sideways Head to Knee Pose

Dwi Hasta Shirsha Parshva Janu Shirshasana

(DWI-huh-stuh SHEER-shuh PAHRSH-vuh JAH-nu sheer-SHAHS-uh-nuh)

Also Known As: Sideways Half Bound Angle Pose (Parshva Ardha Baddha Konasana)
Modification: 1. slight side bend
2. elbow to the knee
3. elbow to the floor
Pose Type: seated, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

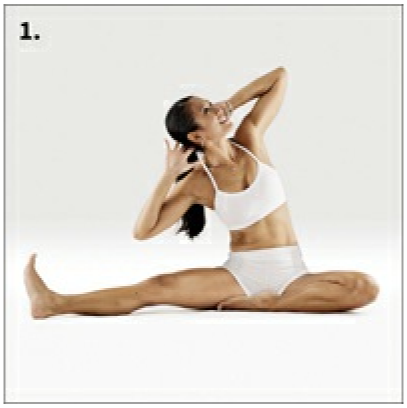

2.

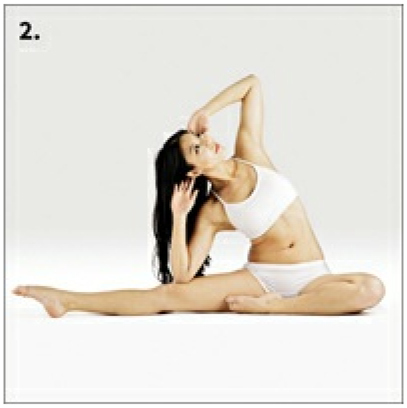

3.

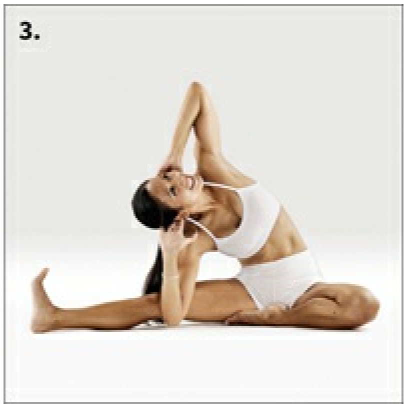

HEAD TO KNEE POSE: LEG LIFTED

Hands Free Sundial Pose 1

Mukta Hasta Surya Yantrasana 1

(MUK-tuh HUH-stuh SOOR-yuh yuhn-TRAHS-uh-nuh)

Modification: foot to thigh, fingertips of both hands to the floor
Pose Type: seated, forward bend, side bend, core
Drishti Point: Hastagrai or Hastagre (hands)

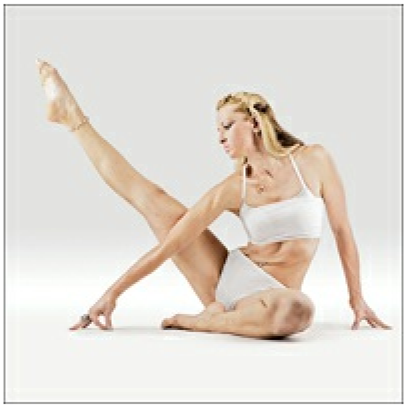

Sundial Pose 1

Surya Yantrasana 1

(SOOR-yuh yuhn-TRAHS-uh-nuh)

Modification: foot to thigh, grabbing onto the foot on the same side

Pose Type: seated, forward bend, side bend

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

ल व

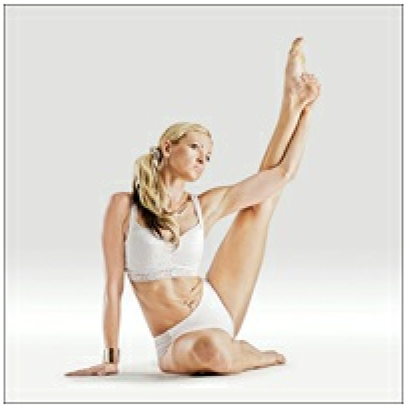

Unsupported Seated Revolved Upward One Hand to Foot Pose

Niralamba Upavishta Parivritta Urdhva Eka Pada Hastasana

(nir-AH-luhm-buh u-puh-VISH-tuh puh-ri-VRIT-tuh OORD-vuh EY-kuh PUH-duh huh-STAHS-uh-nuh)

Pose Type: seated, forward bend, twist, core

Drishti Point: Hastagrai or Hastagre (hands)

ल व र

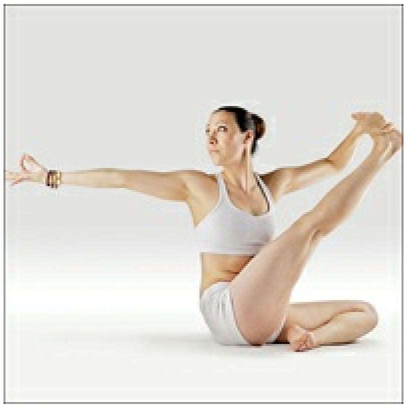

HERO POSE: TOES CURLED IN

Tip Toe Hero Pose

Prapada Virasana

(PRUH-puh-duh veer-AHS-uh-nuh)

Also Known As: Thunderbolt Pose (Vajrasana)
Modification: toes curled in, hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated
Drishti Point: Hastagrai or Hastagre (hands) or Nasagrai or Nasagre (nose)
●

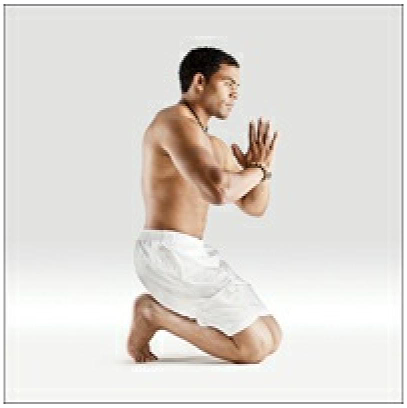

Tip Toe Hero Pose

Prapada Virasana

(PRUH-puh-duh veer-AHS-uh-nuh)

Also Known As: Thunderbolt Pose (Vajrasana)
Modification: head down, arms open wide
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
● ●

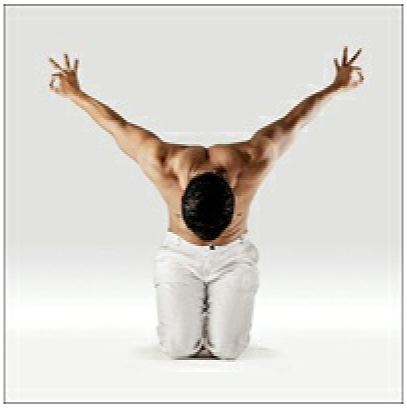

HERO POSE: SITTING ON THE HEELS

Hero Pose

Virasana

(veer-AHS-uh-nuh)

Also Known As: Thunderbolt Pose (Vajrasana)
Modification: arms to the sides; modified version for tight ankles
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

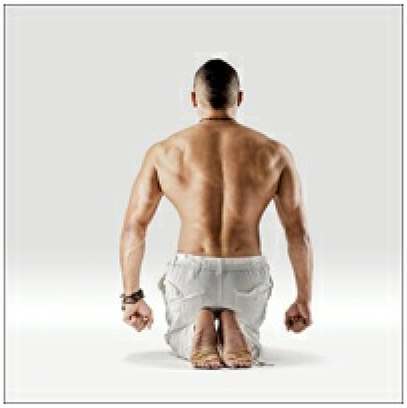

HERO POSE: SITTING ON THE HEELS—HANDS ON THE KNEES

Hero Pose

Virasana

(veer-AHS-uh-nuh)

Also Known As: Thunderbolt Pose (Vajrasana)
Modification: hands on the knees, palms up
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

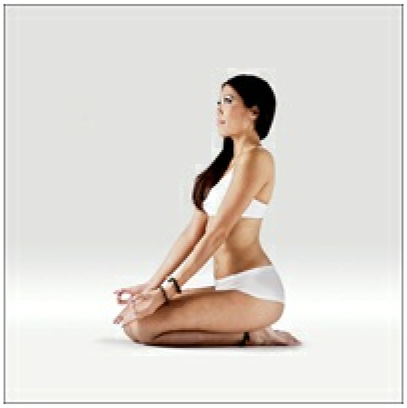

Lion Pose Dedicated to an Avatar of Lord Vishnu in Hero Pose

Narasimhasana in Virasana

(nuh-ruh-sim-HAHS-uh-nuh in veer-AHS-uh-nuh)

Also Known As: Lion Pose in Thunderbolt Pose (Simhasana in Vajrasana)
Modification: sitting on the heels
1. Cat Tilt
2. Dog Tilt
Pose Type: seated
1. forward bend
2. mild backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

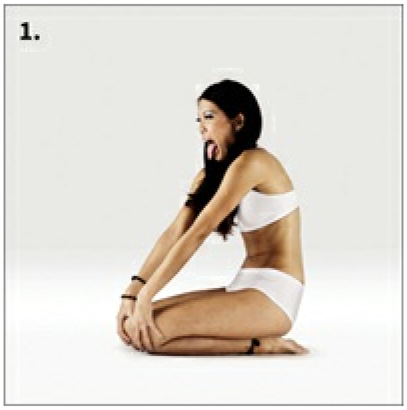

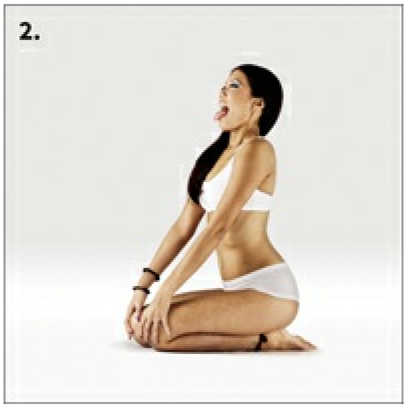

HERO POSE: SITTING ON THE HEELS—ARMS UP & FORWARD BEND

Hero Pose—Raised Bound Hands

Virasana Urdhva Baddha Hastasana

(veer-AHS-uh-nuh OORD-vuh BUH-duh huh-STAHS-uh-nuh)

Also Known As: Thunderbolt Pose Raised Bound Hands (Vajrasana Urdhva Baddha Hastasana)

Modification: palms facing up
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

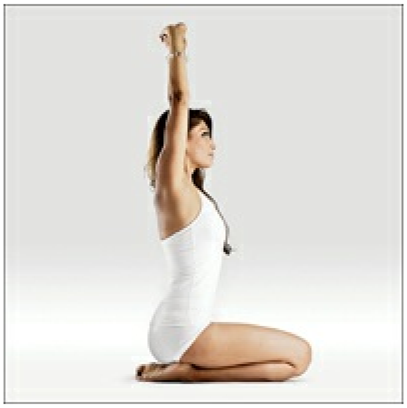

Hero Pose

Virasana

(veer-AHS-uh-nuh)

Also Known As: Thunderbolt Pose (Vajrasana)
Modification: forward bending, ankles crossed, palms to the floor by the knees
Pose Type: seated, forward bend
Drishti Point: Angushtamadhye or Angustha Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

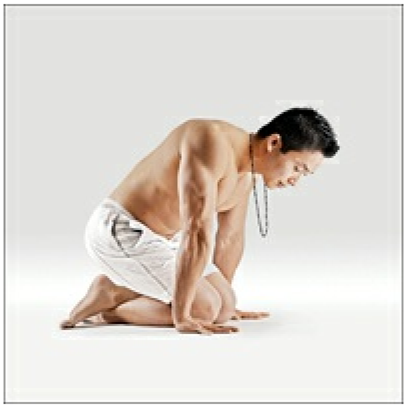

Hand Position of the Pose Dedicated to Garuda in Hero Pose

Hasta Garudasana in Virasana

(HUH-stuh guh-ru-DAHS-uh-nuh in veer-AHS-uh-nuh)

Also Known As: Hand Position of the Pose Dedicated to Garuda in Thunderbolt Pose (Hasta Garudasana in Vajrasana)
Modification: 1. half forward bend

2. spine straight

Pose Type: 1. seated, forward bend

2. seated

Drishti Point: Angushtamadhye or Angustha Ma Dyai (thumbs)

1. लं रं हं
2. लं

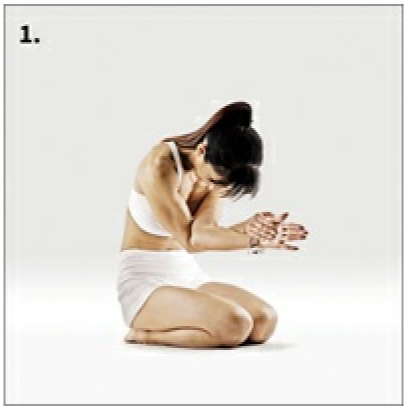

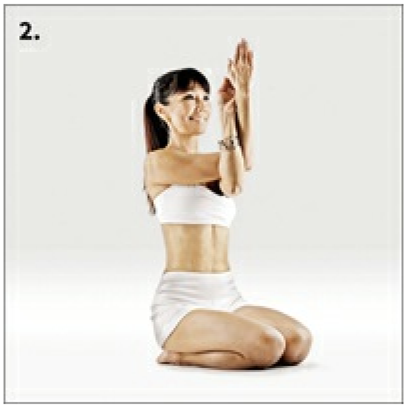

HERO POSE: SITTING ON THE HEELS—KNEES OFF THE FLOOR

Hero Scale Pose with Hands in Prayer

Vira Tolasana Namaskar

(VEER-uh to-LAHS-uh-nuh nuh-muhs-KAHR)

Modification: hands in Anjali Mudra (Hands in Prayer)

Pose Type: seated, balance, core

Drishti Point: Nasagrai or Nasagre (nose) or Hastagrai or Hastagre (hands)

लं

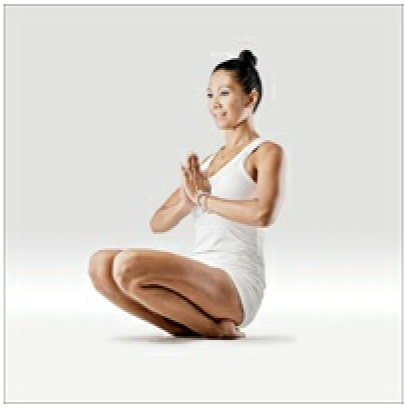

Hero Scale Pose

Vira Tolasana

(VEER-uh to-LAHS-uh-nuh)

Modification: grabbing onto the knees, Cat Tilt
Pose Type: seated, balance, core
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

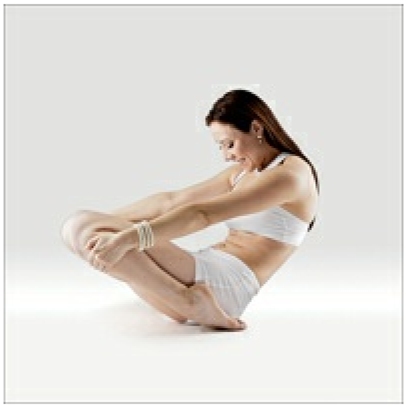

THUNDERBOLT POSE: SITTING BONES ON THE FLOOR—VARIOUS ARM POSITIONS

Hero Pose

Virasana

(veer-AHS-uh-nuh)

Also Known As: Thunderbolt Pose (Vajrasana)
Modification: sitting bones to the floor
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

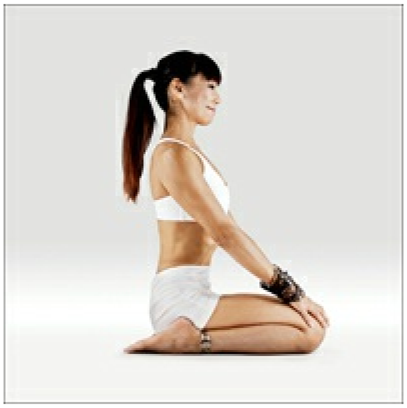

Both Hands to Feet Hero Pose

Dwi Hasta Pada Virasana

(DWI-huh-stuh PUH-duh veer-AHS-uh-nuh)

Also Known As: Both Hands to Feet Thunderbolt Pose (Dwi Hasta Pada Vajrasana)
Modification: sitting bones to the floor, arms crossed behind the back
1. back view
2. front view, mild backbend
Pose Type: 1. seated, binding
2. seated, mild backbend, binding
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

1.

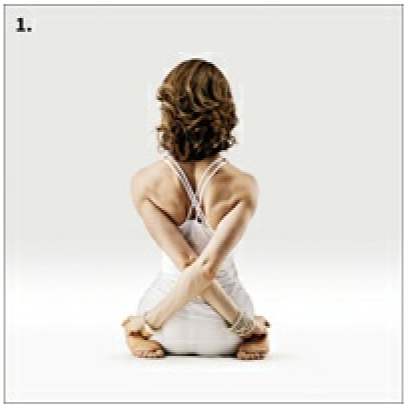

2.

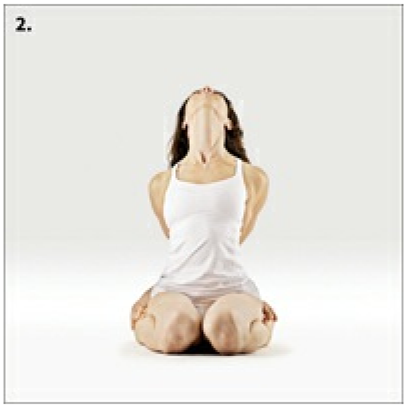

One Arm Upward Hero Pose

Eka Urdhva Hasta Virasana

(EY-kuh OORD-vuh HUH-stuh veer-AHS-uh-nuh)

Also Known As: One Arm Upward Thunderbolt Pose (Eka Urdhva Hasta Vajrasana)
Pose Type: seated
Drishti Point: Hastagrai or Hastagre (hands)

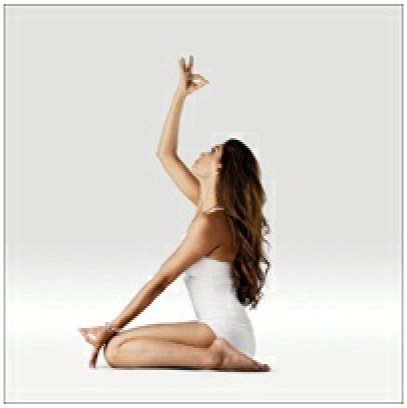

Hero Pose—Raised Bound Hands

Virasana Urdhva Baddha Hastasana

(veer-AHS-uh-nuh OORD-vuh BUH-duh huh-STAHS-uh-nuh)

Also Known As: Thunderbolt Pose Raised Bound Hands (Vajrasana Urdhva Baddha Hastasana)
Modification: palms facing up
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

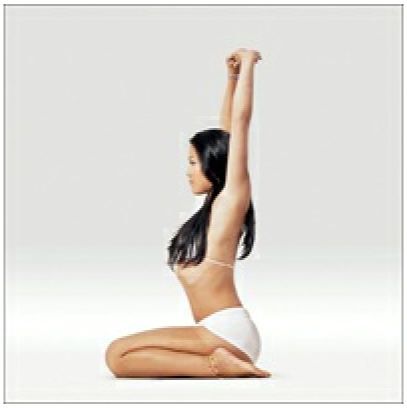

Revolved Hero Pose

Parivritta Virasana

(puh-ri-VRIT-tuh veer-AHS-uh-nuh)

Also Known As: Revolved Thunderbolt Pose (Parivritta Vajrasana)
Pose Type: seated, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
लं

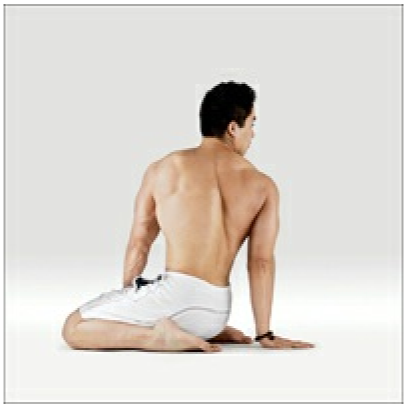

Pose Dedicated to Sage Bharadvaja 1

Bharadvajasana 1

(buh-ruhd-vuhj-AHS-uh-nuh)

Modification: ankles crossed, grabbing onto the bicep
Pose Type: seated, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
लं यं

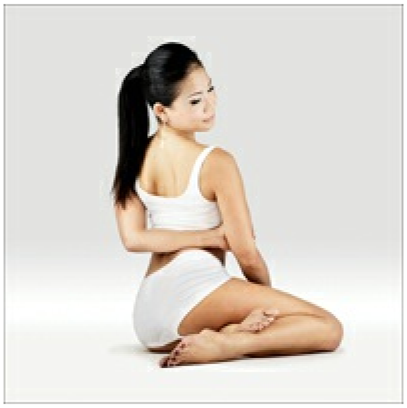

Sideways Pose Dedicated to Sage Bharadvaja 1

Parshva Bharadvajasana 1

(PAHRSH-vuh buh-ruhd-vuhj-AHS-uh-nuh)

Modification: one arm crossed in front; other arm over the head, elbow bent

Pose Type: seated, side bend

Drishti Point: Hastagrai or Hastagre (hands)

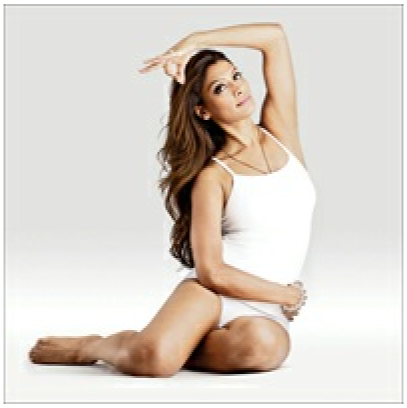

Pose Dedicated to Bharadvaja 1

Bharadvajasana 1

(buh-ruhd-vuhj-AHS-uh-nuh)

Modification: 1. grabbing onto the inside of the hip with opposite hand, other hand on the knee, looking over the shoulder

2. grabbing onto the tricep of the opposite arm, looking over the shoulder

Pose Type: seated, twist, binding

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

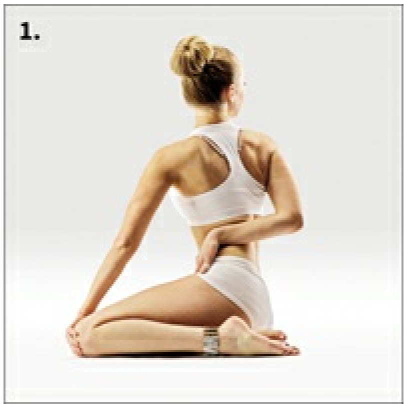

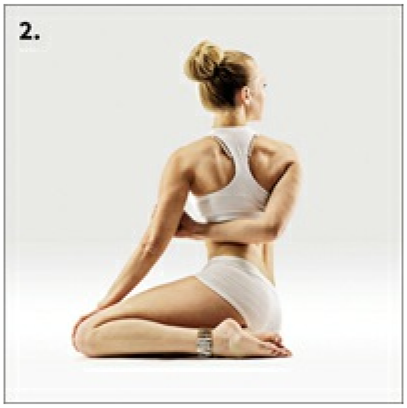

2.

THUNDERBOLT POSE: SITTING BONES ON THE FLOOR—KNEES OPEN WIDE

Lion Pose Dedicated to an Avatar of Lord Vishnu in Knees Spread Wide Hero Pose

Narasimhasana in Prasarita Janu Virasana

(nuh-ruh-sim-HAHS-uh-nuh in pruh-SAH-ri-tuh JAH-nu veer-AHS-uh-nuh)

Also Known As: Lion Pose (Simhasana)
Modification: sitting bones on the floor
Pose Type: seated, mild backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं रं यं हं ॐ

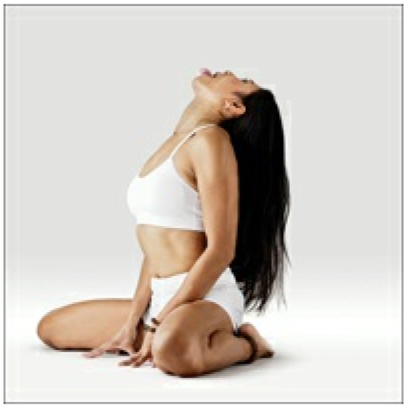

Knees Spread Wide Hero Pose

Prasarita Janu Virasana

(pruh-SAH-ri-tuh JAH-nu veer-AHS-uh-nuh)

Also Known As: Frog Pose (Mandukasana)
Modification: sitting bones off the floor
Pose Type: seated

Drishti Point: Nasagrai or Nasagre (nose)

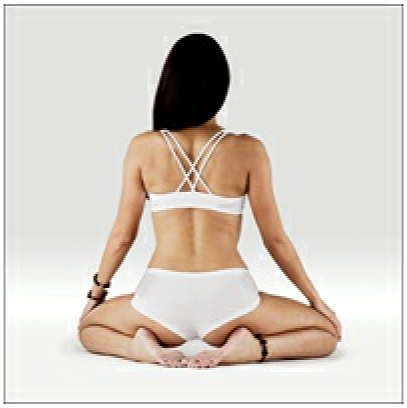

Half Bound Revolved Knees Spread Wide Hero Pose

Ardha Baddha Parivritta Prasarita Janu Virasana

(UHR-duh BUH-duh puh-ri-VRIT-tuh pruh-SAH-ri-tuh JAH-nu veer-AHS-uh-nuh)

Pose Type: seated, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

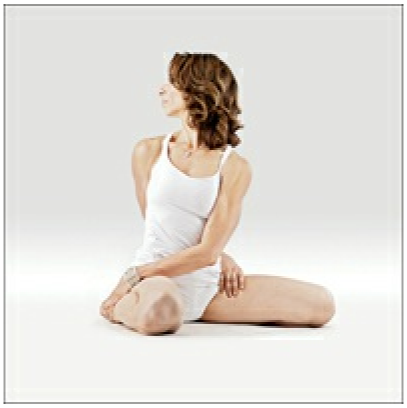

THUNDERBOLT POSE: SITTING BONES ON THE FLOOR—FEET TURNED OUT

Complete Thunderbolt Pose

Paripurna Vajrasana

(puh-ri-POOR-nuh vuhj-RAHS-uh-nuh)

Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

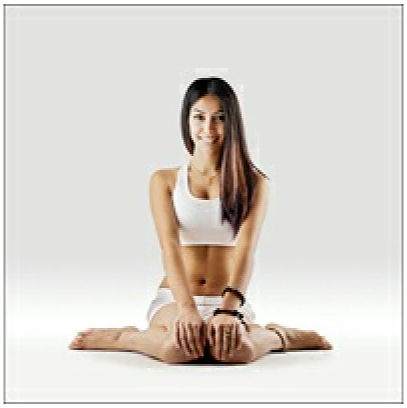

Revolved Complete Thunderbolt Pose

Parivritta Paripurna Vajrasana

(puh-ri-VRIT-uh puh-ri-POOR-nuh vuhj-RAHS-uh-nuh)

Pose Type: seated, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
लं

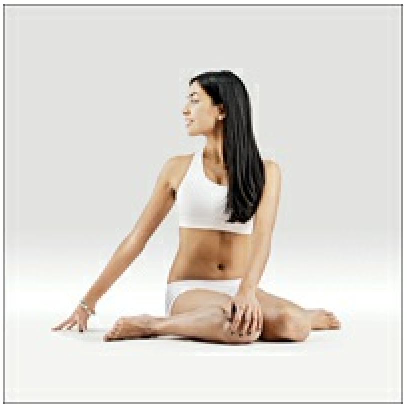

Western Intense Stretch Pose in Complete Thunderbolt Pose

Paschimottanasana in Paripurna Vajrasana

(puhsh-chi-mo-tahn-AHS-uh-nuh in puh-ri-POOR-nuh vuhj-RAHS-uh-nuh)

Also Known As: Seated Forward Bend in Complete Thunderbolt Pose
Modification: forward bend, toes turned out
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं रं

Half Pose Dedicated to Sage Marichi in Revolved Half Thunderbolt Pose

Ardha Marichyasana in Parivritta Ardha Vajrasana

(UHR-duh muh-ree-CHYAHS-uh-nuh in puh-ri-VRIT-uh UHR-duh vuhj-RAHS-uh-nuh)

Modification: 1. looking straight ahead
2. looking over the shoulder
Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व र

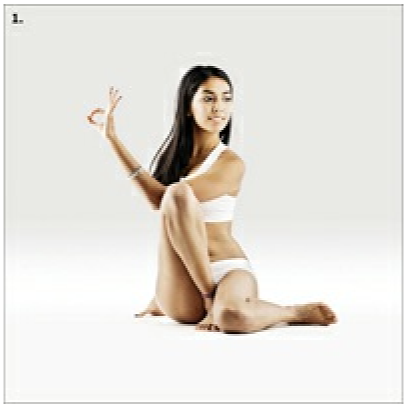

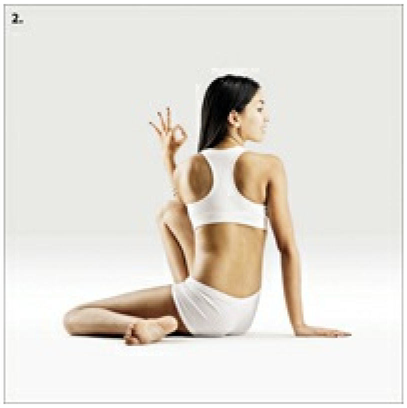

Half Bound Pose Dedicated to Sage Marichi in Half Thunderbolt Pose

Ardha Baddha Marichyasana in Ardha Vajrasana

(UHR-duh BUH-duh muh-ree-CHYAHS-uh-nuh in UHR-duh vuhj-RAHS-uh-nuh)

Pose Type: seated, forward bend, twist, binding

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

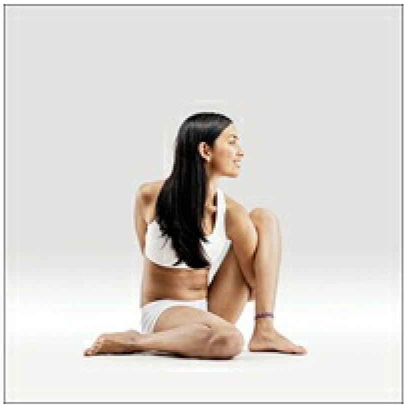

THUNDERBOLT POSE: SITTING BONES ON THE FLOOR—FEET TURNED OUT—ARMS UP OVER THE HEAD

Complete Thunderbolt Pose—Raised Bound Hands

Paripurna Vajrasana—Urdhva Baddha Hastasana

(puh-ri-POOR-nuh vuhj-RAHS-uh-nuh OORD-vuh BUH-duh huh-STAHS-uh-nuh)

Modification: 1. looking straight ahead
2. head rolling back

Pose Type: 1. seated
2. seated, mild backbend

Drishti Point: 1. Nasagrai or Nasagre (nose)
2. Angushtamadhye (thumbs)

1.
2.

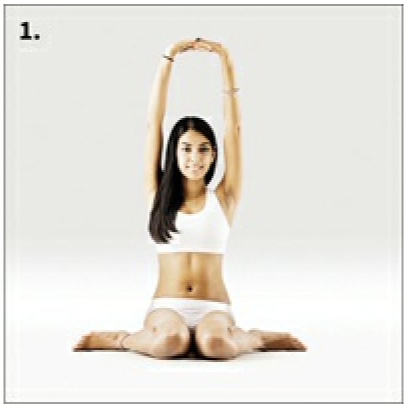

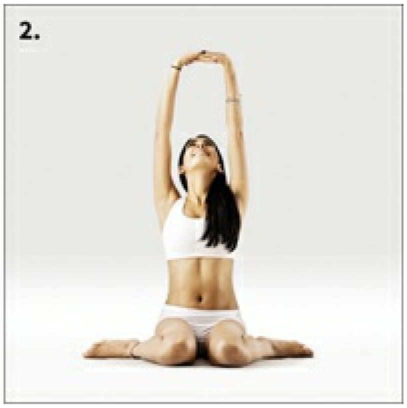

Revolved Complete Thunderbolt Pose—Raised Bound Hands

Parivritta Paripurna Vajrasana—Urdhva Baddha Hastasana

(puh-ri-VRIT-tuh puh-ri-POOR-nuh vuhj-RAHS-uh-nuh OORD-vuh BUH-duh huh-STAHS-uh-nuh)

Pose Type: seated, twist
Drishti Point: Angushtamadhye (thumbs)

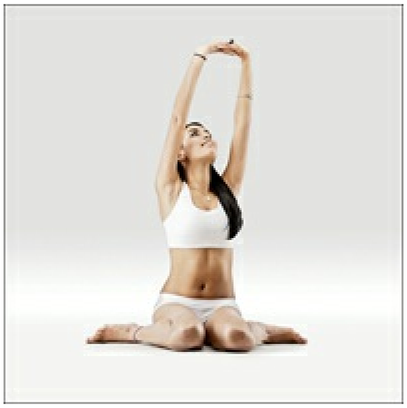

Sideways Complete Thunderbolt Pose—Raised Bound Hands

Parshva Paripurna Vajrasana—Urdhva Baddha Hastasana

(PAHRSH-vuh puh-ri-POOR-nuh vuhj-RAHS-uh-nuh OORD-vuh BUH-duh huh-STAHS-uh-nuh)

Pose Type: seated, side bend
Drishti Point: Angushtamadhye (thumbs)

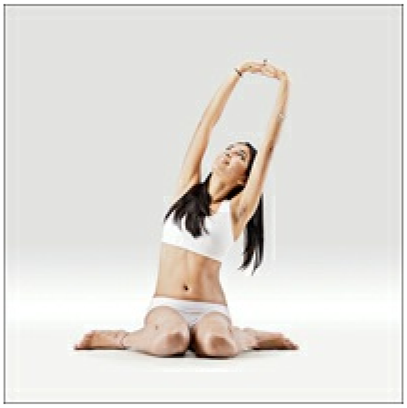

One Hand Sideways Complete Thunderbolt Pose

Eka Hasta Parshva Paripurna Vajrasana

(EY-kuh HUH-stuh PAHRSH-vuh puh-ri-POOR-nuh vuhj-RAHS-uh-nuh)

Pose Type: seated, side bend
Drishti Point: Hastagrai or Hastagre (hands)

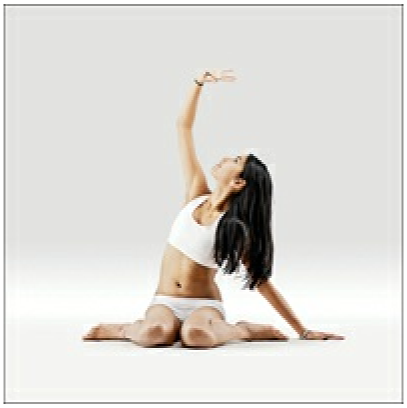

THUNDERBOLT POSE: SITTING BONES ON THE FLOOR—FEET TURNED OUT—ONE-LEGGED & FORWARD BENDS

Half Pose Dedicated to Sage Marichi in Half Thunderbolt Pose

Ardha Marichyasana in Ardha Vajrasana

(UHR-duh muh-ree-CHYAHS-uh-nuh in UHR-duh vuhj-RAHS-uh-nuh)

Modification: grabbing onto the knee with both hands
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

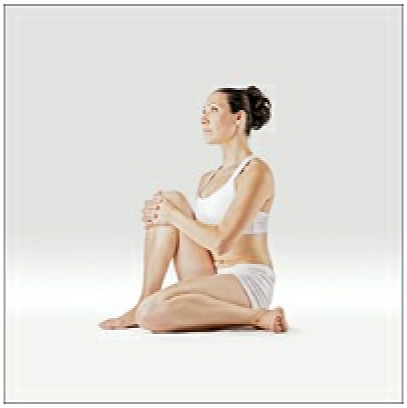

Half Pose Dedicated to Sage Marichi in Half Thunderbolt Pose

Ardha Marichyasana in Ardha Vajrasana

(UHR-duh muh-ree-CHYAHS-uh-nuh in UHR-duh vuhj-RAHS-uh-nuh)

Modification: palms to the floor in front of the hips
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

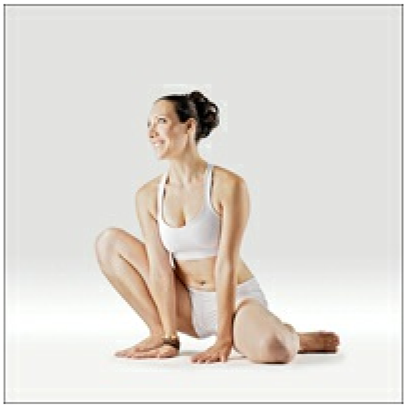

Half Star Pose in Half Frog Pose

Ardha Tarasana in Ardha Mandukasana

(UHR-duh tahr-AHS-uh-nuh in UHR-duh muhn-doo-KAHS-uh-nuh)

Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

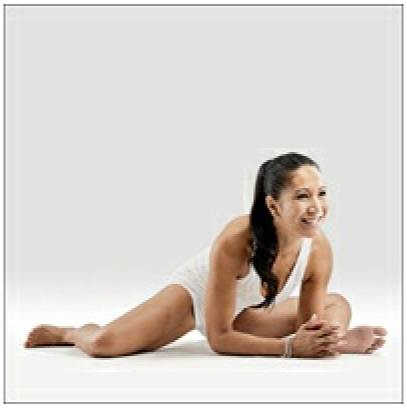

Pose Dedicated to Virancha (Brahma) 1 Prep.

Viranchyasana 1 Prep.

(vir-uhn-CHYAHS-uh-nuh)

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

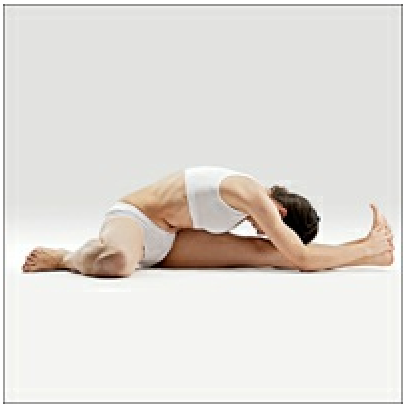

ONE KNEE BENT TO THE BACK: OTHER FOOT TO THE THIGH—SPINE STRAIGHT, TWIST & SIDE BEND

Easy Pose Dedicated to Sage Bharadvaja 2 Prep.

Sukha Bharadvajasana 2 Prep.

(SUK-uh buh-ruhd-vuhj-AHS-uh-nuh)

Modification: neutral spine, back of the hands on the knees

Pose Type: seated

Drishti Point: Nasagrai or Nasagre (nose)

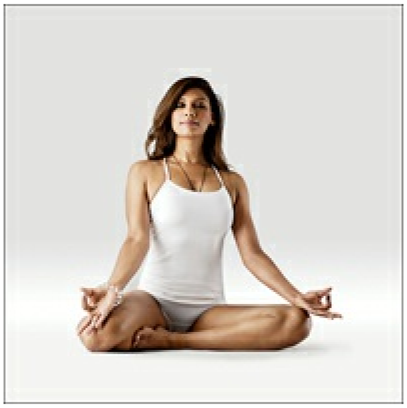

Easy Pose Dedicated to Sage Bharadvaja 2

Sukha Bharadvajasana 2

(SUK-uh buh-ruhd-vuhj-AHS-uh-nuh)

Modification: foot to the inside of the thigh
Pose Type: seated, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)
ल व

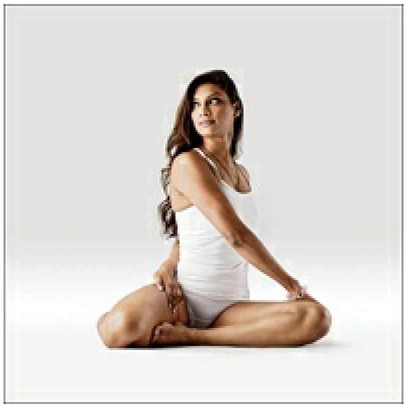

Sideways Easy Pose Dedicated to Sage Bharadvaja 2

Parshva Sukha Bharadvajasana 2

(PAHRSH-vuh SUK-uh buh-ruhd-vuhj-AHS-uh-nuh)

Modification: foot to the inside of the thigh, side bend
Pose Type: seated, side bend
Drishti Point: Hastagrai or Hastagre (hands)
ल व

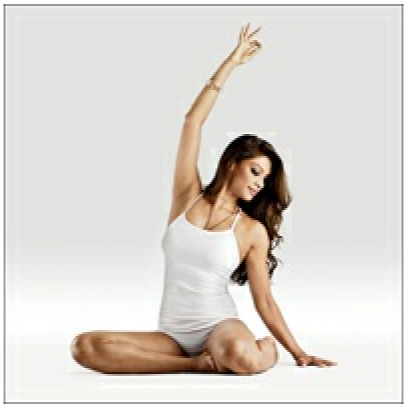

ONE KNEE BENT TO THE BACK: OTHER FOOT TO THE THIGH—SIDE BEND

Easy One-Legged King Pigeon Pose 1

Sukha Eka Pada Raja Kapotasana 1

(SUK-uh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: seated, side bend

Drishti Point: Hastagrai or Hastagre (hands)

ल व य ह

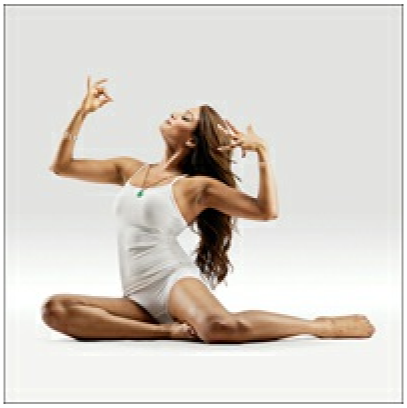

Sideways Half Hero Half Bound Angle Pose

Parshva Ardha Vira Ardha Baddha Konasana

(PAHRSH-vuh UHR-duh VEER-uh UHR-duh BUH-duh ko-NAHS-uh-nuh)

Modification: forearm to the floor

Pose Type: seated, side bend

Drishti Point: Hastagrai or Hastagre (hands)

ल व

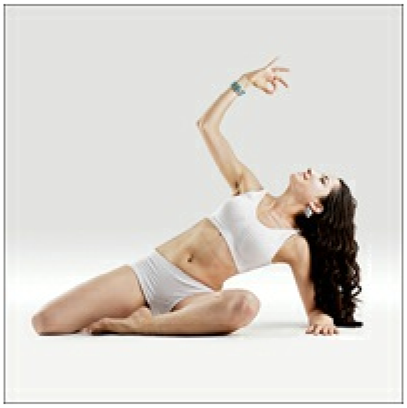

Half Hero Half Bound Angle Pose

Ardha Vira Ardha Baddha Konasana

(UHR-duh VEER-uh UHR-duh BUH-duh ko-NAHS-uh-nuh)

Modification: arm behind the back, hand to the hip; backbend
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

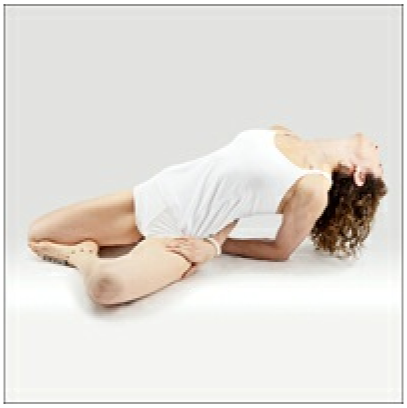

Twist Dedicated to Sage Bharadvaja 2

Bharadvajasana 2

(buh-ruhd-vuhj-AHS-uh-nuh)

Pose Type: seated, twist, binding

Drishti Point: Parsva Drishti (to the right), Parsva Drishti (to the left)

How to Perform the Pose:

1. Begin by sitting on the floor with both your legs straight out in front of you. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Inhale as you lean over to the left side, bending your right knee and bringing your right knee, shin, and ankle to the floor. Your right heel should be close to your right hip to protect the knee and your right knee should be open slightly to the side.
3. Exhale as you bend your left knee and bring your left foot to your right hip socket in the Half Lotus (*Ardha Padmasana*), rotating the sole of your left foot up to the sky. Try to keep both knees on the floor.
4. Inhale as you expand your chest and hold your arms straight out to the sides, parallel to the floor.
5. Exhale as you twist to the left, bringing your left arm behind your back and grabbing your left foot with your left hand to bind. Grab onto the left knee with your right hand and look over your left shoulder (Pose #2).
6. Inhale as you lengthen your spine. Exhale as you bring the back of your left hand under your right knee (Pose #1). On the next exhale, look over your right shoulder to deepen the twist (Pose #3).
7. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
8. Inhale as you release the bind and look forward. Exhale as you bring both your legs straight out in front of you. Repeat on the other side.

Modification:

1. palm under the knee, twisting to the inside of the body
2. hand on the knee, twisting to the inside of the body
3. palm under the knee, twisting to the outside of the body

Bharadvaja = Pindola Bharadvaja was one of four Arhats asked by Buddha to stay on earth to propagate Buddhist law or Dharma

1.

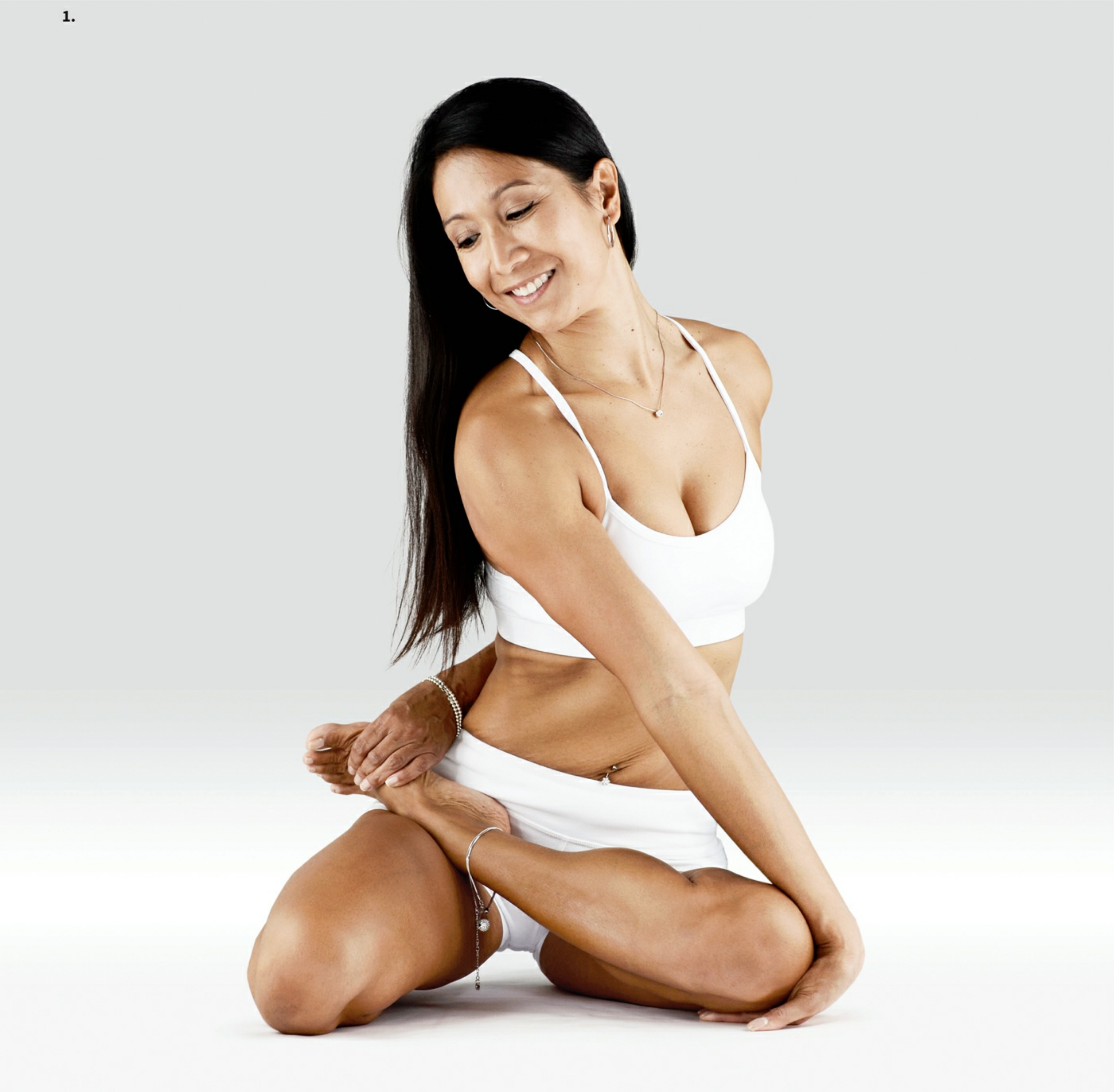

2.

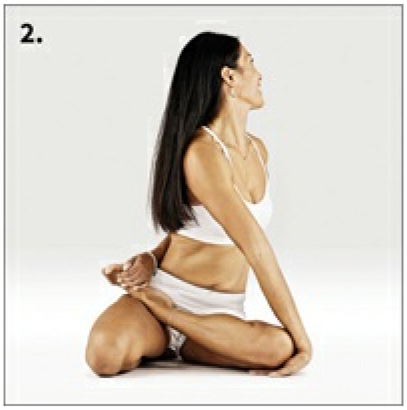

3.

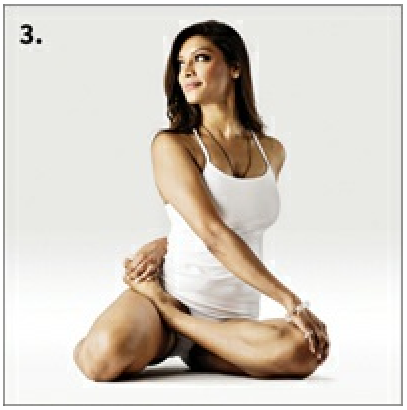

Half Fire Log Pose in Half Bound Hero Pose

Ardha Agnistambhasana in Ardha Baddha Virasana

(UHR-duh uhg-ni-stuhm-BAHS-uh-nuh in UHR-duh BUH-duh veer-AHS-uh-nuh)

Modification: 1. grabbing onto the hip
2. grabbing onto the heel
Pose Type: seated, binding
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

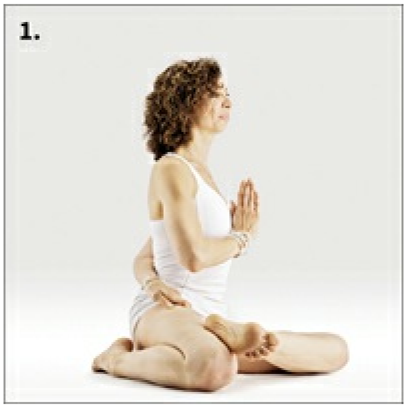

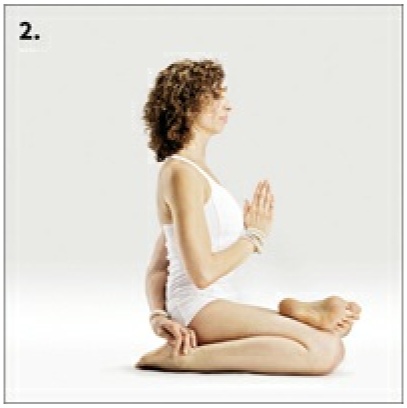

Pose Dedicated to Sage Bharadvaja 2 Prep.

Bharadvajasana 2 Prep.

(buh-ruhd-vuhj-AHS-uh-nuh)

Modification: palms together, heels of the palms resting on the crown of the head, neutral spine
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

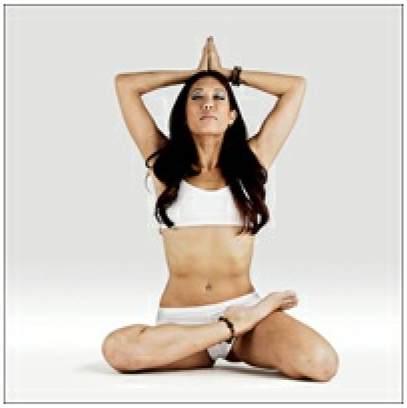

Pose Dedicated to Sage Bharadvaja 2 Prep.

Bharadvajasana 2 Prep.

(buh-ruhd-vuhj-AHS-uh-nuh)

Also Known As: Root Lock Pose (Mula Bandhasana)
Modification: sitting on the heel, both hands grabbing onto the foot, neutral spine
Pose Type: seated, binding
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व

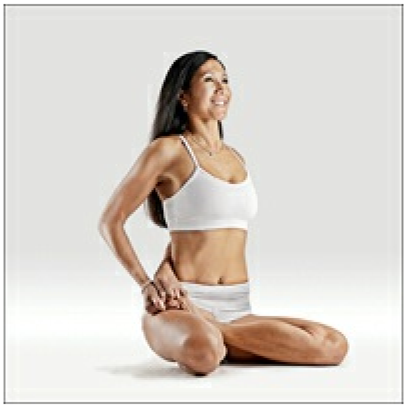

ONE KNEE BENT TO THE BACK: KNEE TO THE CHEST, BINDING, TWIST & BACKBEND

Pose Dedicated to Sage Marichi 5 & 6 Prep.

Marichyasana 5 & 6 Prep.

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi E & F Prep. (Marichyasana E & F Prep.)
Modification: spine straight
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

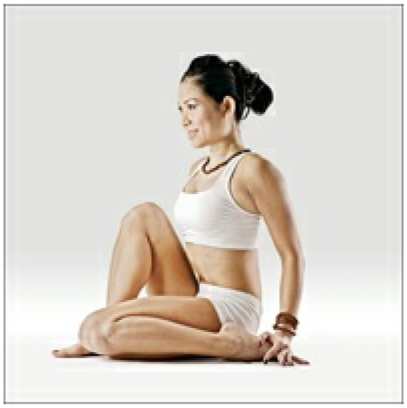

Pose Dedicated to Sage Marichi 5

Marichyasana 5

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi E (Marichyasana E)
Pose Type: seated, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

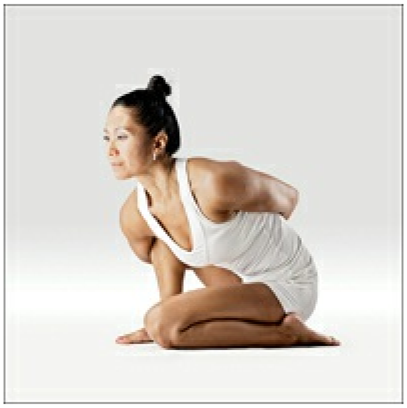

Pose Dedicated to Sage Marichi 6

Marichyasana 6

(muh-ree-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Sage Marichi F (Marichyasana F)
Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

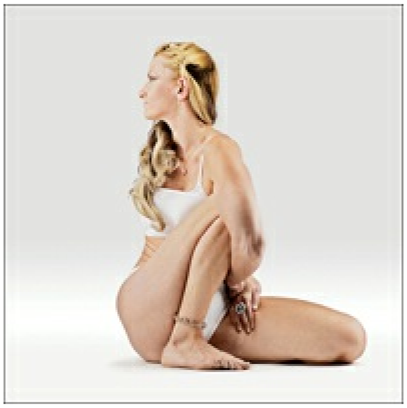

One-Legged King Pigeon Pose 3

Eka Pada Raja Kapotasana 3

(EY-kuh PUH-duh RAHJ-uh kuh-po-TAHS-uh-nuh)

Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

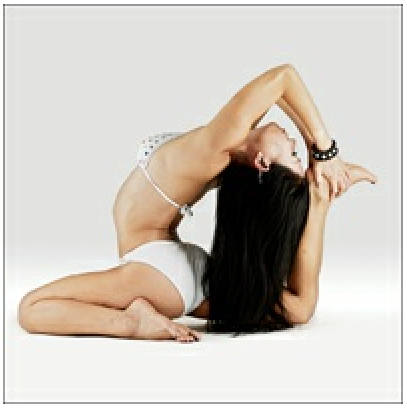

LEG CRADLE & FOOT TO THE KNEE

Baby Cradle Pose in Half Hero Pose

Hindolasana in Ardha Virasana

(hin-do-LAHS-uh-nuh in UHR-duh veer-AHS-uh-nuh)

Also Known As: Baby Cradle Pose in Half Thunderbolt Pose (Hindolasana in Ardha Vajrasana)
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व

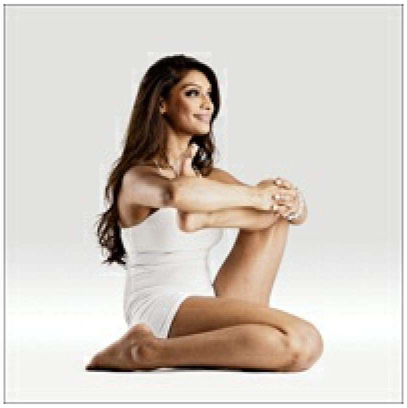

Baby Cradle Pose Prep.

Hindolasana Prep.

(hin-do-LAHS-uh-nuh)

Modification: palms to the floor behind the hips, heel of the bottom foot to the sitting bone
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

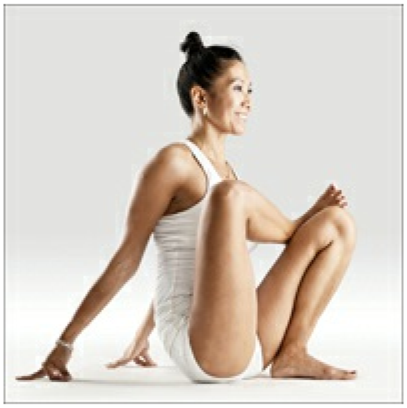

Half Fire Log Pose in Half Yogic Staff Pose Prep.

Ardha Agnistambhasana in Ardha Yogadandasana Prep.

(UHR-duh uhg-ni-stuhm-BAHS-uh-nuh in UHR-duh yo-guh-duhn-DAHS-uh-nuh)

Pose Type: seated, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

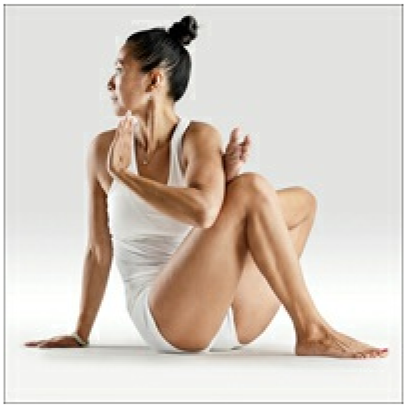

Yogic Staff Pose Prep.

Yogadandasana Prep.

(yo-guh-duhn-DAHS-uh-nuh)

Modification: grabbing onto the foot with one hand
Pose Type: seated, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

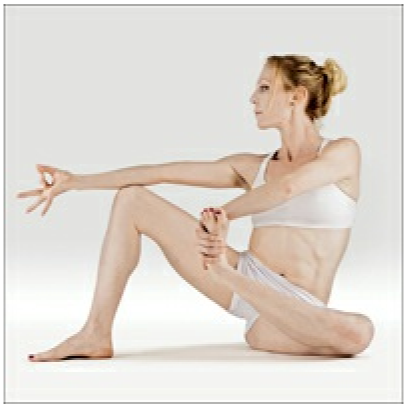

Yogic Staff Pose Prep.

Yogadandasana Prep.

(yo-guh-duhn-DAHS-uh-nuh)

Modification: grabbing onto the foot with both hands, foot to the chest
Pose Type: seated, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

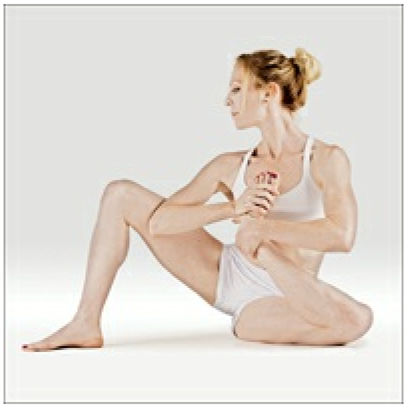

LEG CRADLE: FOOT TO THE HIP & HALF LOTUS

Baby Cradle Pose Prep.

Hindolasana Prep.

(hin-do-LAHS-uh-nuh)

Modification: both hands grabbing onto the foot, fingers interlocked
Pose Type: seated, forward bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

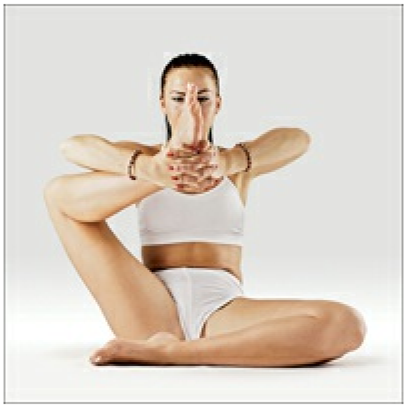

Baby Cradle Pose

Hindolasana

(hin-do-LAHS-uh-nuh)

Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

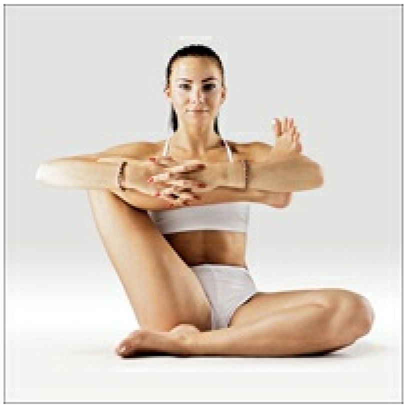

Four Corner Pose

Chatushkonasana

(chuh-tush-ko-NAHS-uh-nuh)

Pose Type: seated, forward bend, binding
Drishti Point: Nasagrai or Nasagre (nose)

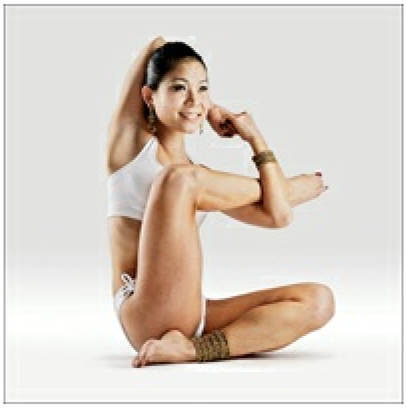

Half Lotus Baby Cradle Pose

Ardha Padma Hindolasana

(UHR-duh PUHD-muh hin-do-LAHS-uh-nuh)

Also Known As: Pose Dedicated to Virancha (Brahma) 1 or A Prep. (Viranchyasana 1 or A Prep.)
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

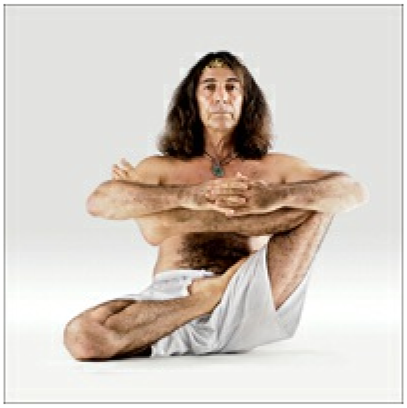

FOOT TO THE ARMPIT: BOTH KNEES BENT

Yogic Staff Pose with Hands in Prayer

Yogadandasana Namaskar

(yo-guh-duhn-DAHS-uh-nuh nuh-muhs-KAHR)

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

ल व

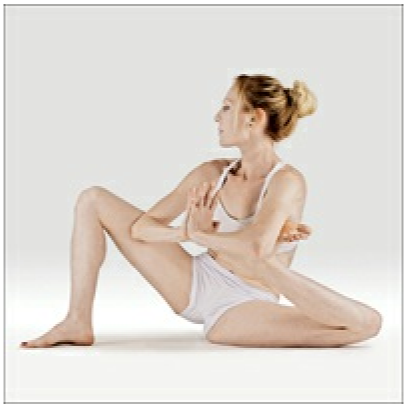

Yogic Staff Pose

Yogadandasana

(yo-guh-duhn-DAHS-uh-nuh)

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

ल व

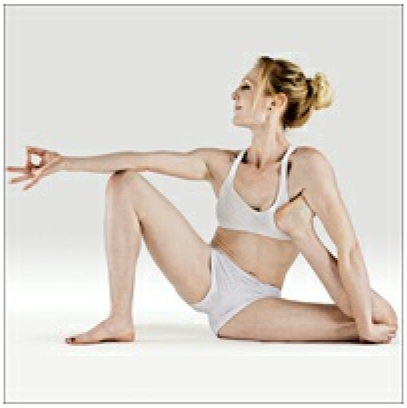

Half Bound Fire Log Pose in Half Yogic Staff Pose Prep.

Ardha Baddha Agnistambhasana in Ardha Yogadandasana Prep.

(UHR-duh BUH-duh uhg-ni-stuhm-BAHS-uh-nuh in UHR-duh yo-guh-duhn-DAHS-uh-nuh)

Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

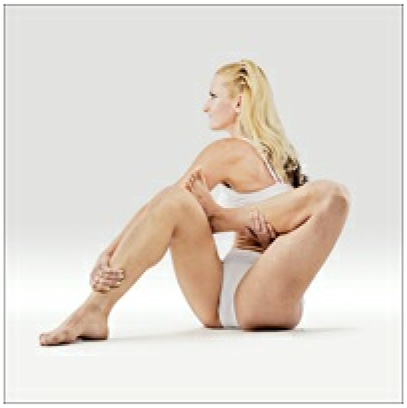

Bound Fire Log Pose in Half Yogic Staff Pose Prep.

Baddha Agnistambhasana in Ardha Yogadandasana Prep.

(BUH-duh uhg-ni-stuhm-BAHS-uh-nuh in UHR-duh yo-guh-duhn-DAHS-uh-nuh)

Pose Type: seated, forward bend, twist, binding
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

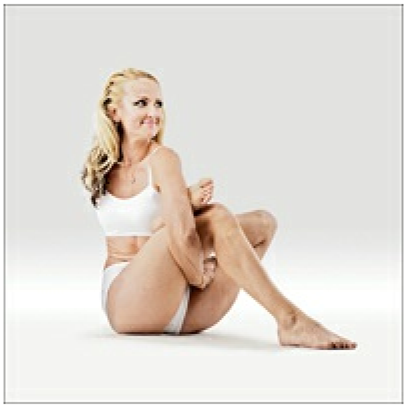

Hands Bound Half Lotus Pose in Half Yogic Staff Pose Prep.

Baddha Hasta Ardha Padmasana in Ardha Yogadandasana Prep.

(BUH-duh HUH-stuh UHR-duh puhd-MAHS-uh-nuh in UHR-duh yo-guh-duhn-DAHS-uh-nuh)

Pose Type: seated, twist, binding

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

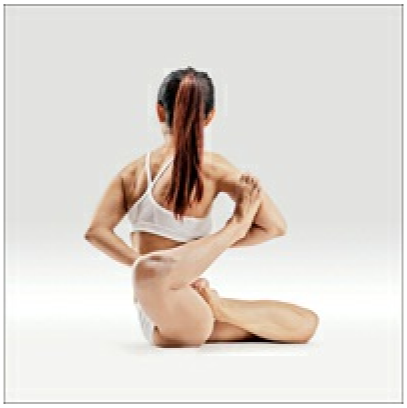

Hands Bound Half Hero Pose in Half Yogic Staff Pose Prep.

Baddha Hasta Ardha Virasana in Ardha Yogadandasana Prep.

(BUH-duh HUH-stuh UHR-duh veer-AHS-uh-nuh in UHR-duh yo-guh-duhn-DAHS-uh-nuh)

Pose Type: seated, twist, binding

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

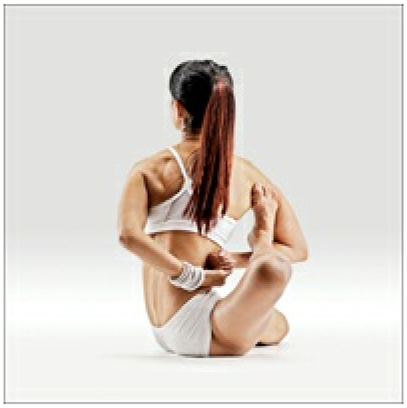

ONE LEG OVER THE SHOULDER: BOTH KNEES BENT

Sundial Pose 2 Prep.

Surya Yantrasana 2 Prep.

(SOOR-yuh yuhn-TRAHS-uh-nuh)

Modification: shoulder to the back of the knee, knee bent
Pose Type: seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

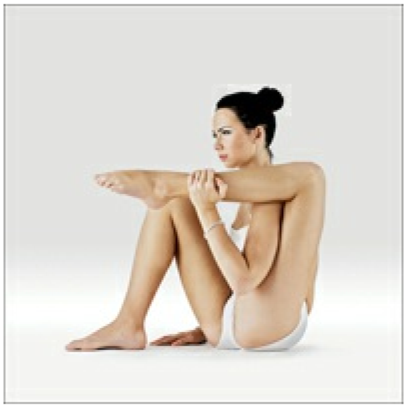

FOOT BEHIND THE HEAD

Pose Dedicated to Virancha (Brahma) 1

Viranchyasana 1

(vir-uhn-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Virancha (Brahma) A (Viranchyasana A)
Pose Type: seated, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

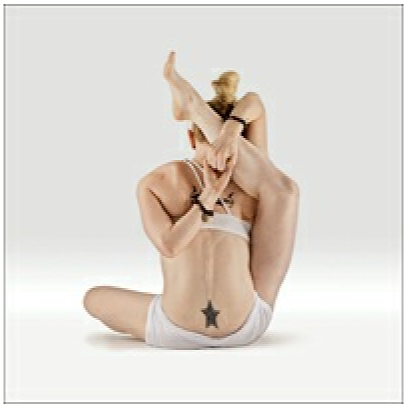

Seated Moonbird Pose

Upavishta Chakorasana

(u-puh-VISH-tuh chuh-kor-AHS-uh-nuh)

Also Known As: Upavistha Chakorasana
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

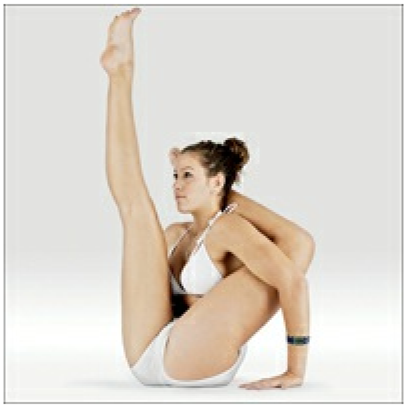

One Foot Behind the Head Pose

Eka Pada Shirshasana

(EY-kuh PUH-duh sheer-SHAHS-uh-nuh)

Modification: palms on the floor by the hips, knee bent toward the chest
Pose Type: seated, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

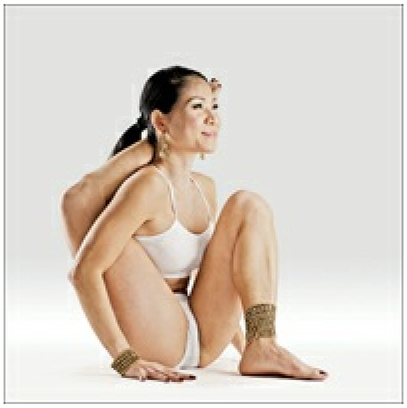

Pose Dedicated to Virancha (Brahma) 1

Viranchyasana 1

(vir-uhn-CHYAHS-uh-nuh)

Also Known As: Pose Dedicated to Virancha (Brahma) A (Viranchyasana A)
Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: seated
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

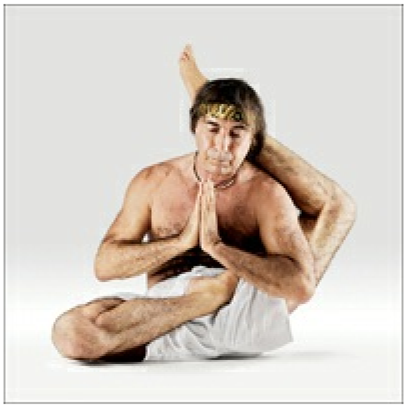

COW FACE POSE: SPINE STRAIGHT

Leg Position of Cow Face Pose

Pada Gomukhasana

(PUH-duh go-muk-AHS-uh-nuh)

Modification: hands on the top knee, one palm on top of the other
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)

Ganesh Seal in Leg Position of Cow Face Pose

Ganesh Mudra in Pada Gomukhasana

(guh-NEYSH MU-druh in PUH-duh go-muk-AHS-uh-nuh)

Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)
● ●

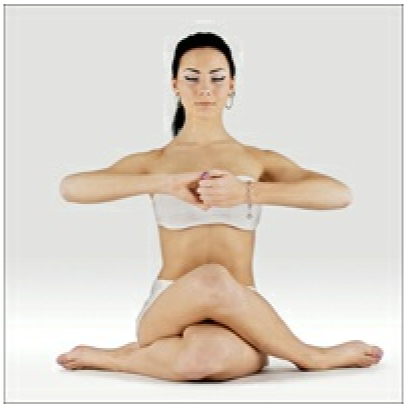

Both Hands to Feet Leg Position of Cow Face Pose

Dwi Hasta Pada Gomukhasana

(DWI-huh-stuh PUH-duh go-mu-KAHS-uh-nuh)

Modification: arms crossed behind the back, grabbing onto the feet
1. back view
2. front view
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ● ● ●

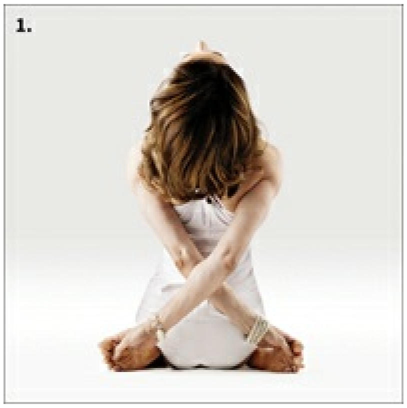

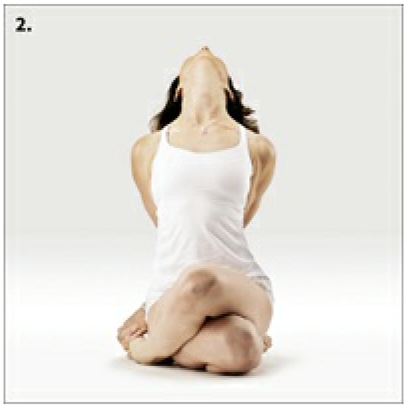

Cow Face Pose

Gomukhasana

(go-mu-KAHS-uh-nuh)

Pose Type: seated

Drishti Point: Nasagrai or Nasagre (nose)

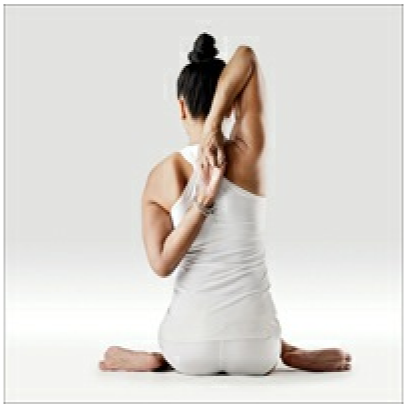

Leg Position of Cow Face Pose

Pada Gomukhasana

(PUH-duh go-mu-KAHS-uh-nuh)

Modification: elbows touching behind the head
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose)
ल व ह

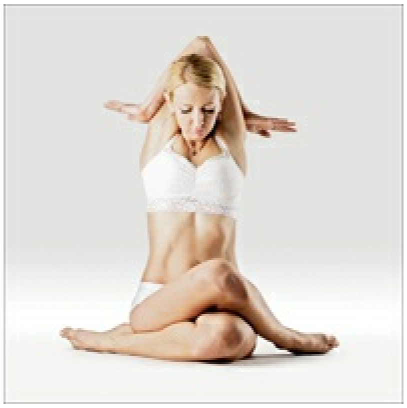

COW FACE POSE: FORWARD BEND

Leg Position of Cow Face Pose

Pada Gomukhasana

(PUH-duh go-mu-KAHS-uh-nuh)

Modification: hands in Anjali Mudra (Hands in Prayer), thumbs to third eye; elbows together
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose) or Angushtamadhye or Angustha Ma Dyai (thumbs)
ल व ॐ

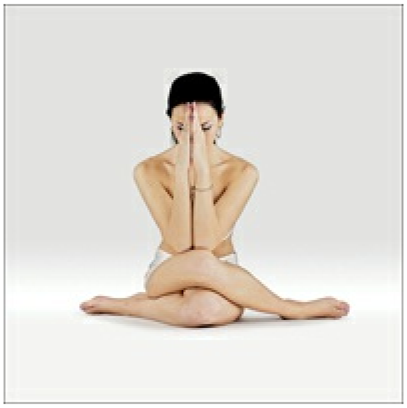

Western Intense Stretch Pose in Leg Position of Cow Face Pose

Paschimottanasana in Pada Gomukhasana

(puhsh-chi-mo-TAHS-uh-nuh in PUH-duh go-mu-KAHS-uh-nuh)

Also Known As: Seated Forward Bend in Leg Position of Cow Face Pose and Long Horn Pose

(Dighasrngasana)
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

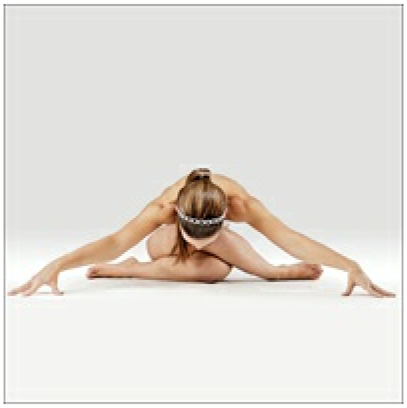

Hand Position of the Pose Dedicated to Garuda in Leg Position of Cow Face Pose

Hasta Garudasana in Pada Gomukhasana

(HUH-stuh guh-ru-DAHS-uh-nuh in PUH-duh go-mu-KAHS-uh-nuh)
Modification: forward bend
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

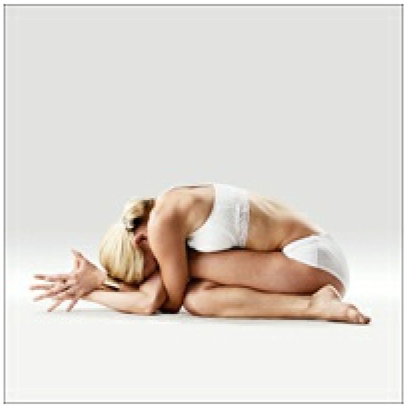

COW FACE POSE: SIDE BEND

Sideways Leg Position of Cow Face Pose

Parshva Pada Gomukhasana

(PAHRSH-vuh PUH-duh go-mu-KAHS-uh-nuh)

Pose Type: seated, side bend
Drishti Point: Hastagrai or Hastagre (hands)
ल व

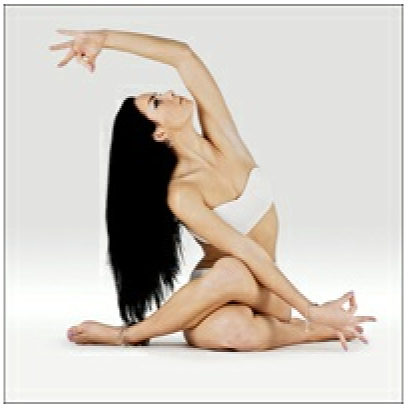

Hand to Foot Sideways Leg Position of Cow Face Pose

Hasta Pada Parshva Pada Gomukhasana

(HUH-stuh PUH-duh PAHRSH-vuh PUH-duh go-mu-KAHS-uh-nuh)

Pose Type: seated, side bend
Drishti Point: Hastagrai or Hastagre (hands)
ल व

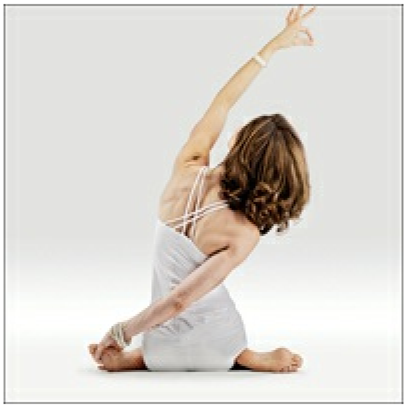

Sideways Leg Position of Cow Face Pose

Parshva Pada Gomukhasana

(PAHRSH-vuh PUH-duh go-mu-KAHS-uh-nuh)

Modification: intense shoulder stretch
Pose Type: seated, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

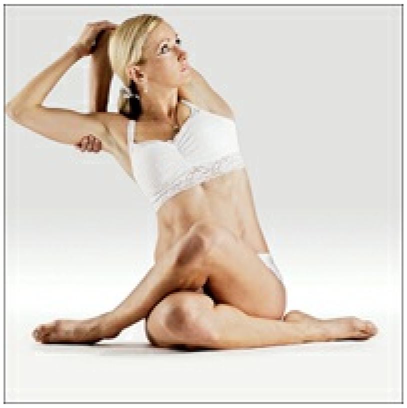

Hand to Foot Revolved Leg Position of Cow Face Pose

Hasta Pada Parivritta Pada Gomukhasana

(HUH-stuh PUH-duh puh-ri-VRIT-tuh PUH-duh go-mu-KAHS-uh-nuh)

Modification: grabbing onto the big toe, arm behind the back; other arm in front of the body, hand on the hip
Pose Type: seated, twist, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

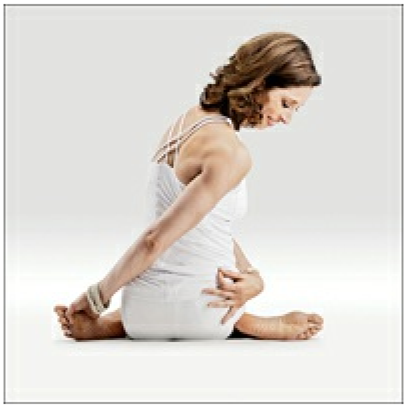

COW FACE POSE: KNEES LIFTED TO THE CHEST

Upward Leg Position of Cow Face Pose

Urdhva Pada Gomukhasana

(OORD-vuh PUH-duh go-mu-KAHS-uh-nuh)

Modification: arms under the legs, knees lifted to the chest
Pose Type: seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Hastagrai or Hastagre (hands)

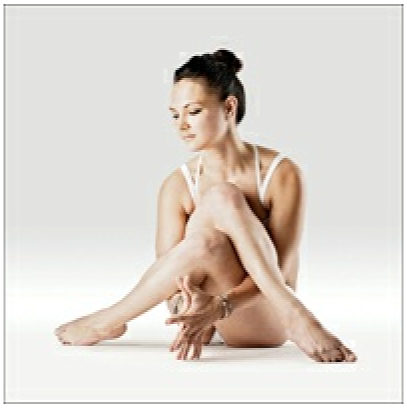

Upward Hands Upward Leg Position of Cow Face Pose

Urdhva Hasta Urdhva Pada Gomukhasana

(OORD-vuh HUH-stuh OORD-vuh PUH-duh go-mu-KAHS-uh-nuh)

Pose Type: seated, forward bend, core

Drishti Point: Hastagrai or Hastagre (hands), Nasagrai or Nasagre (nose)

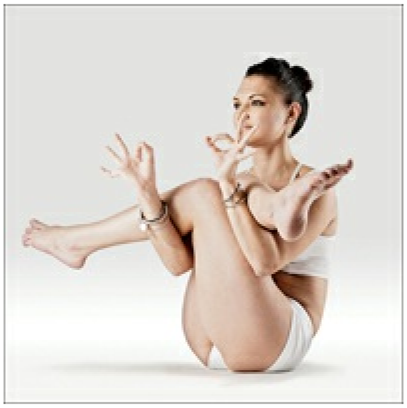

Both Hands to Feet Upward Leg Position of Cow Face Pose

Dwi Hasta Pada Urdhva Pada Gomukhasana

(DWI-huh-stuh PUH-duh OORD-vuh PUH-duh go-mu-KAHS-uh-nuh)

Modification: grabbing the outside edges of both feet

Pose Type: seated, forward bend, core

Drishti Point: Nasagrai or Nasagre (nose)

COW FACE POSE: ELBOW & FOREARM ON THE FLOOR

Half Bound Sideways Easy Leg Position of Cow Face Pose

Ardha Baddha Parshva Sukha Pada Gomukhasana

(UHR-duh BUH-duh PAHRSH-vuh SUK-uh PUH-duh go-mu-KAHS-uh-nuh)

Modification: forearm to the floor, sitting on the bottom heel
1. front view
2. back view
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

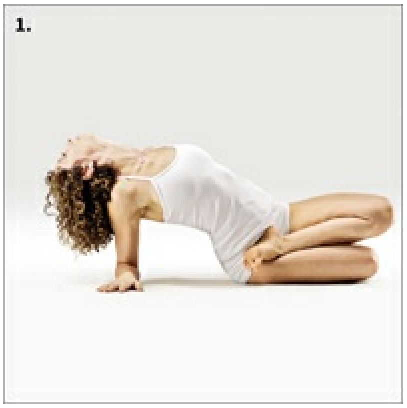

2.

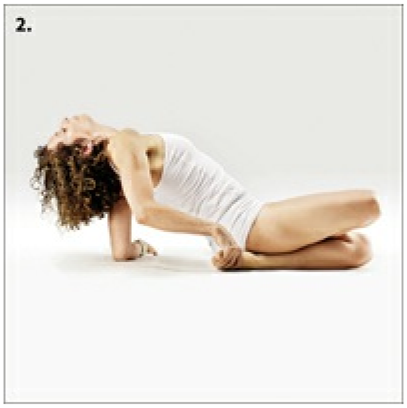

Half Bound Sideways Easy Leg Position of Cow Face Pose

Ardha Baddha Parshva Sukha Pada Gomukhasana

(UHR-duh BUH-duh PAHRSH-vuh SUK-uh PUH-duh go-mu-KAHS-uh-nuh)

Modification: elbow to the floor
Pose Type: seated, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

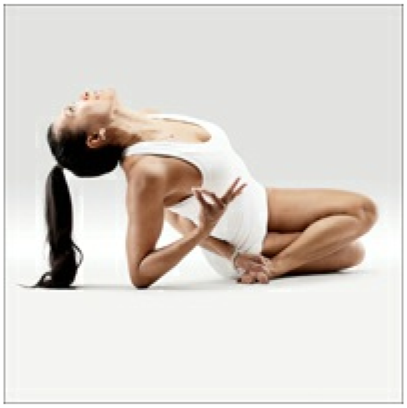

SAGE VAMADEVA’S POSE: LEANING TO THE SIDE

Pose Dedicated to Sage Vamadeva Prep.

Vamadevasana Prep.

(vahm-uh-dey-VAHS-uh-nuh)

Modification: 1. palm flat on the floor, chest rotating up
2. one arm crossed over in front, hand grabbing onto the back foot; top hand to the head
Pose Type: seated, side bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व

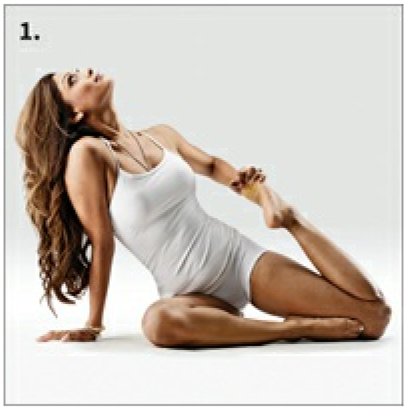

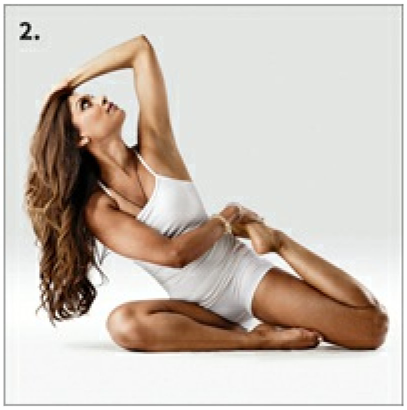

Pose Dedicated to Sage Vamadeva

Vamadevasana

(vahm-uh-dey-VAHS-uh-nuh)

Modification: binding:
1. hand to the hip socket
2. hand to the knee

Pose Type: seated, side bend, twist, binding

Drishti Point: 1. Urdhva or Antara Drishti (up to the sky)
2. Padayoragrai or Padayoragre (toes/feet)

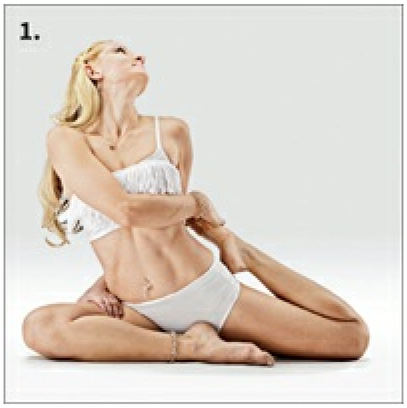

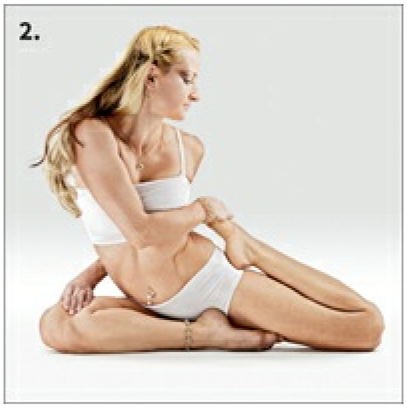

Revolved One-Legged King Pigeon Pose 1 Prep.

Parivritta Eka Pada Raja Kapotasana 1 Prep.

(puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: grabbing onto the back foot

1. one arm crossed over, other arm to the side, elbow bent
2. arms crossed in front of the body, free hand to the opposite knee

Pose Type: seated, side bend, twist

Drishti Point: Hastagrai or Hastagre (hands), Parshva Drishti (to the right), Parshva Drishti (to the left)

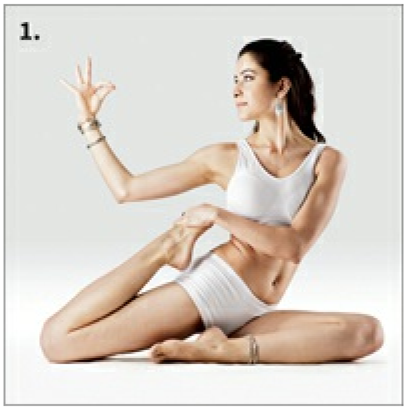

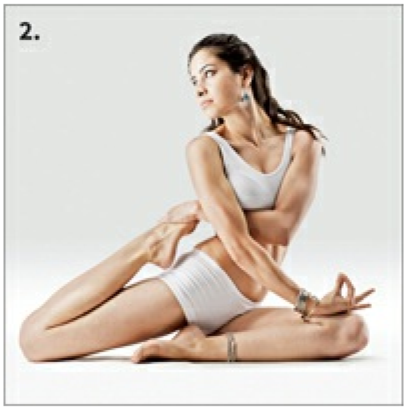

Pose Dedicated to Sage Vamadeva 2

Vamadevasana 2

(vahm-uh-dey-VAHS-uh-nuh)

Pose Type: seated, twist

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

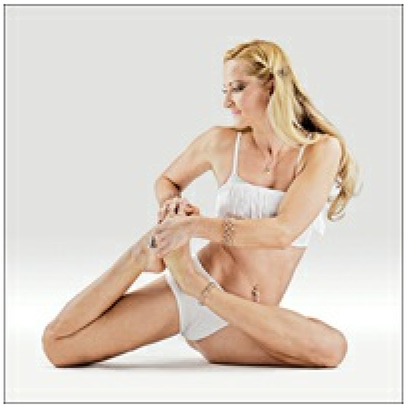

ONE-LEGGED KING PIGEON: HIPS OFF THE FLOOR

One-Legged King Pigeon Pose 1 Version B

Eka Pada Raja Kapotasana 1 B

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: standing (on the knees), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

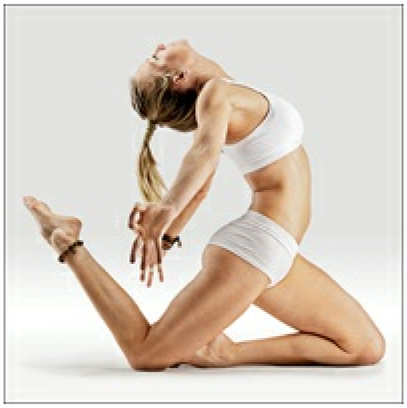

Easy One-Legged King Pigeon Pose 1

Sukha Eka Pada Raja Kapotasana 1

(SUK-uh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: hips lifted off the floor, grabbing onto the foot on the opposite side with overhead grip
1. free arm extended to the front
2. free hand to the floor by the opposite knee
Pose Type: standing (on the knees), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

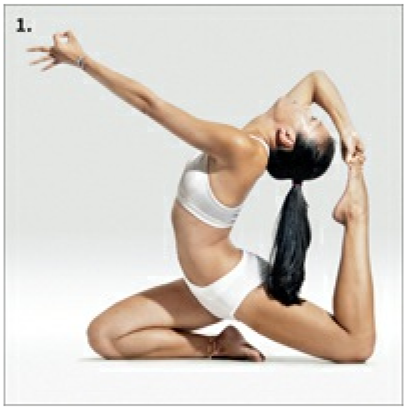

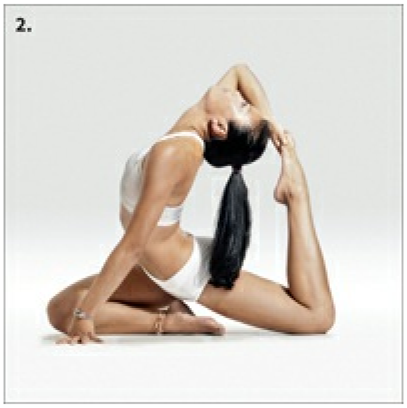

Easy One-Legged King Pigeon Pose 1

Sukha Eka Pada Raja Kapotasana 1

(SUK-uh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: hips off the floor; fingertips to the floor by the hips, foot touching the head
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

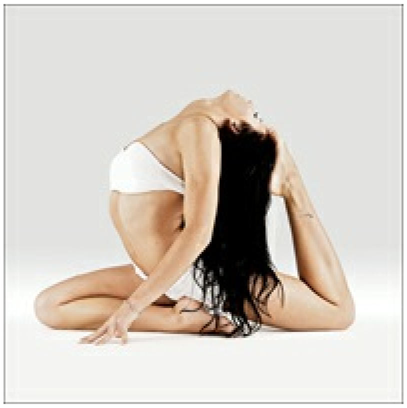

ONE-LEGGED KING PIGEON: UNDER-HEAD GRIP

One-Legged King Pigeon Pose 1 Prep.

Eka Pada Raja Kapotasana 1 Prep.

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: 1. back knee bent
2. grabbing onto the back foot with one arm on the same side
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ॐ

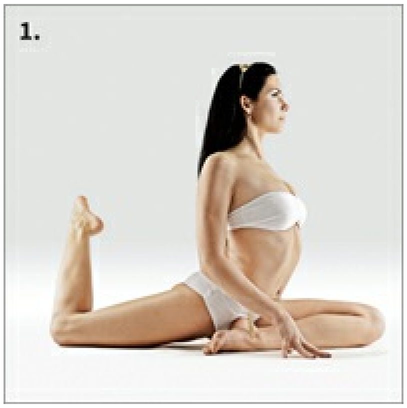

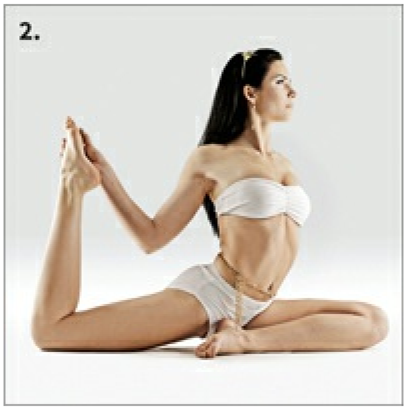

One-Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: one hand on the front knee on the same side, other hand pressing the back foot toward the hip
Pose Type: seated, backbend
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)
ल व र य

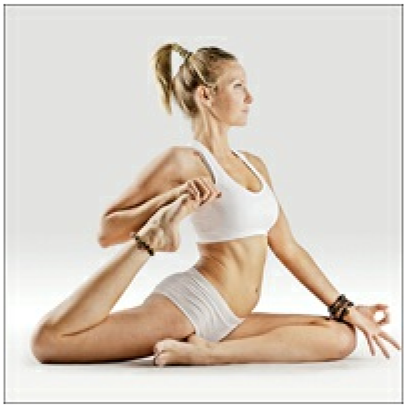

One-Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: one hand on the back knee, one hand on the front knee; foot to the back of the head
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

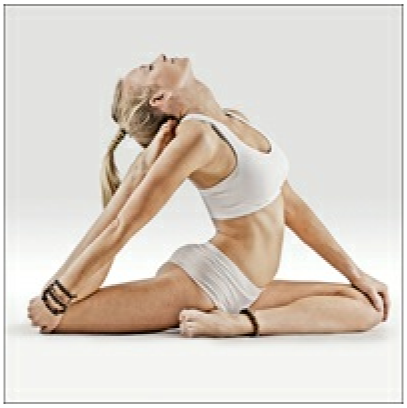

One-Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: foot to the opposite armpit
Pose Type: seated, backbend, twist
Drishti Point: Nasagrai or Nasagre (nose) or Hastagrai or Hastagre (hands)

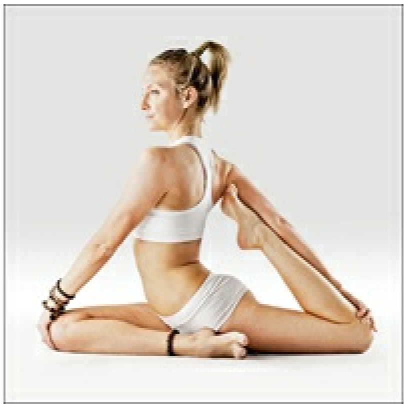

Revolved One-Legged King Pigeon Pose 1

Parivritta Eka Pada Raja Kapotasana 1

(puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: foot to the armpit on the same side

Pose Type: seated, backbend, twist

Drishti Point: Hastagrai or Hastagre (hands)

ल व र य

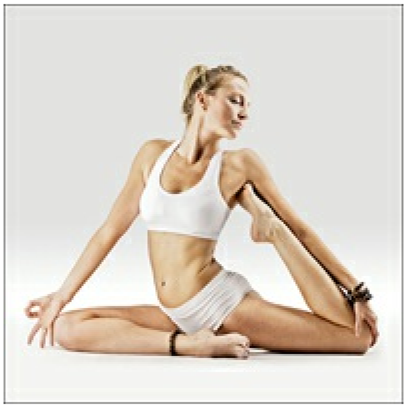

Revolved One-Legged King Pigeon Pose 1

Parivritta Eka Pada Raja Kapotasana 1

(puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: foot under the chin

Pose Type: seated, backbend, twist

Drishti Point: Hastagrai or Hastagre (hands)

ल व र य ह

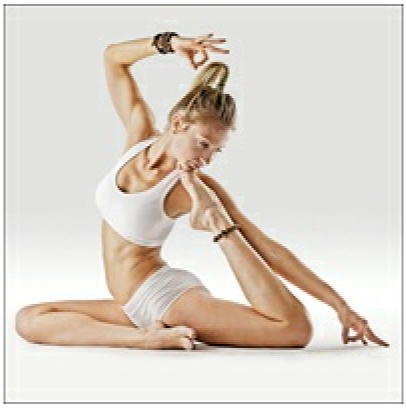

Pose of the Heavenly Spirits Prep.

Valakhilyasana Prep.

(vah-luh-kil-YAHS-uh-nuh)

Modification: under-head grip
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

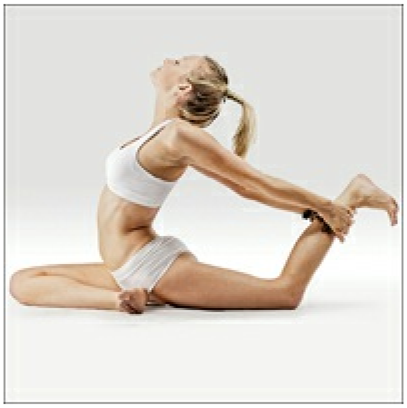

ONE-LEGGED KING PIGEON: OVERHEAD GRIP

One-Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: grabbing onto the foot with one hand on the same side, overhead grip; other hand resting on the front knee
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

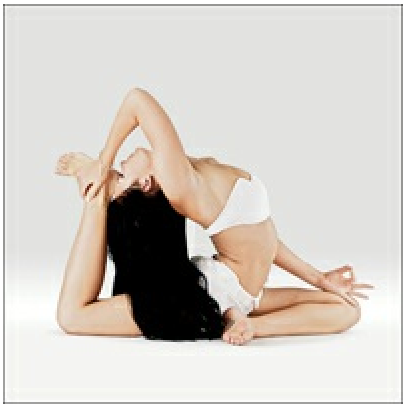

One-Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: grabbing onto the foot with both hands, overhead grip; heel of the back foot to the forehead

Pose Type: seated, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

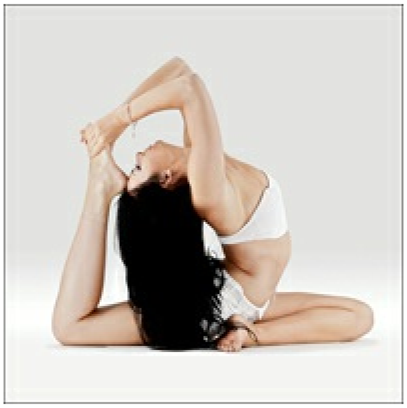

ONE-LEGGED KING PIGEON: ASYMMETRICAL ARMS—CROSS OVER & UNDER-HEAD/OVERHEAD GRIP

One Hand to Foot Bound One-Legged King Pigeon Pose 1

Eka Hasta Pada Baddha Eka Pada Raja Kapotasana 1

(EY-kuh HUH-stuh PUH-duh BUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: arm crossed over in front of the neck, other hand on the knee

Pose Type: seated, backbend

Drishti Point: Nasagrai or Nasagre (nose)

ल व र य ह

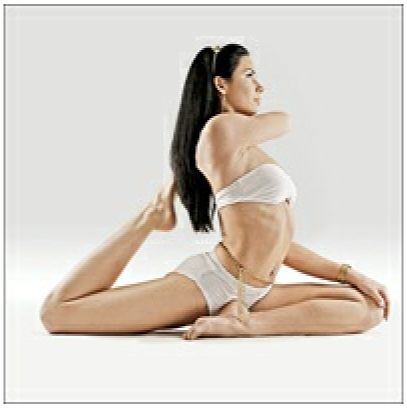

Both Hands to Feet Bound One-Legged King Pigeon Pose 1

Dwi Hasta Pada Baddha Eka Pada Raja Kapotasana 1

(DWI-huh-stuh PUH-duh BUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: arm crossed over in front of the neck
Pose Type: seated, backbend, binding
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य ह

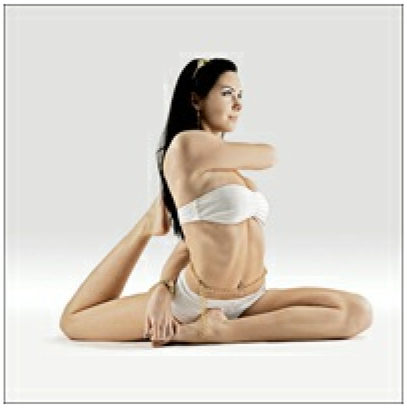

Mermaid Pose 1

Naginyasana 1

(nah-gin-YAHS-uh-nuh)

Pose Type: seated, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य

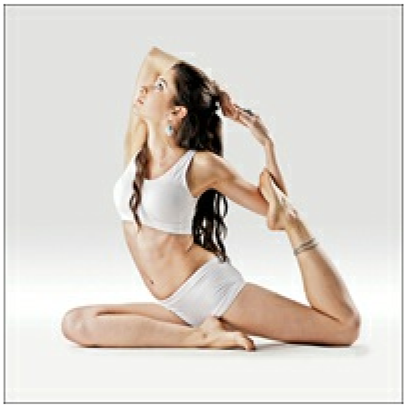

ONE-LEGGED KING PIGEON: ASYMMETRICAL ARMS—CROSS OVER & UNDER-HEAD/OVERHEAD GRIP

Both Hands to Feet Bound One-Legged King Pigeon Pose 1

Dwi Hasta Pada Baddha Eka Pada Raja Kapotasana 1

(DWI-huh-stuh PUH-duh BUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: grabbing onto the opposite foot with overhead grip

Pose Type: seated, backbend, binding

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

लं वं रं यं हं ॐ

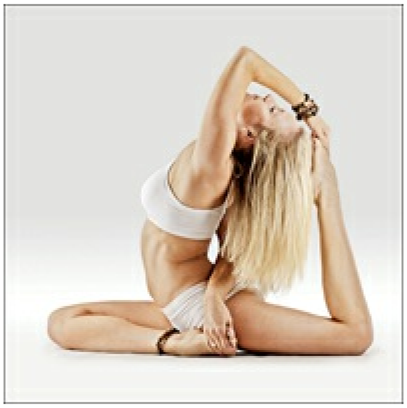

One-Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: one hand on the back knee, one hand grabbing onto the opposite foot with overhead grip

Pose Type: seated, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

लं वं रं यं हं ॐ

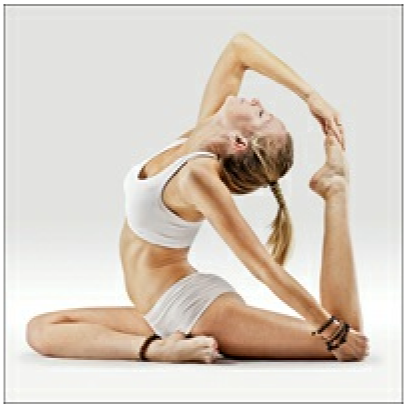

ONE-LEGGED KING PIGEON: FOOT TO THE BACK OF THE HEAD

One-Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: grabbing onto the back knee with both hands under-head grip; foot to the back of the head
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

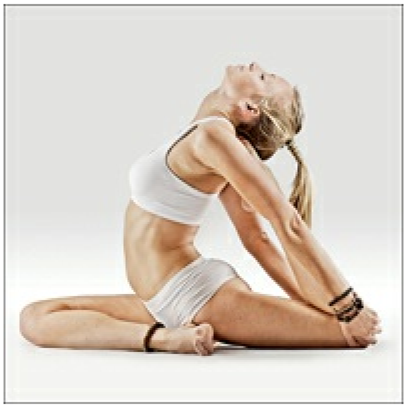

One Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: seated, backbend

Drishti Point: Angusthamadhye or Angustha Ma Dyai (thumbs)

How to Perform the Pose:

1. Begin by sitting on the floor with both your legs straight out in front of you. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale and bend your left knee, keeping the knee on the floor, and slide the left heel toward the inside of your right thigh.
3. Inhale as you lean forward. Exhale, extend the right leg straight out behind you with your right thigh, knee, shin, and front of the right foot on the floor. Square your hips by pulling the left hip to the back and right hip to the front.
4. Inhale as you lengthen the spine, pressing both hands to the floor by your hips for support. Exhale as you let your head roll back into a backbend.
5. Exhale as you bend your right knee, bringing your right foot toward your head. Once your right foot touches your head, slide your right foot under the back of your head and fix it in place.
6. Inhale as you stretch both arms up over your head. Exhale as you bring your palms together, keeping your arms straight and strong (Pose #1). Push through your chest. You can experiment with interlocking your fingers and letting your pointer fingers and thumbs be straight. You can also bring your foot to the side of your head (Pose #2).
7. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
8. Exhale as you release the right foot from under your head and lower it to the floor. Inhale and straighten your spine. Exhale as you bring both your legs straight out in front of you. Repeat on the other side.

Modification: both arms extended to the sky
1. foot to the back of the head, palms together
2. foot to the side of the head, fingers interlocked, pointer fingers free

eka = one

pada = foot or leg

raja = king, royal

kapota = pigeon

1.

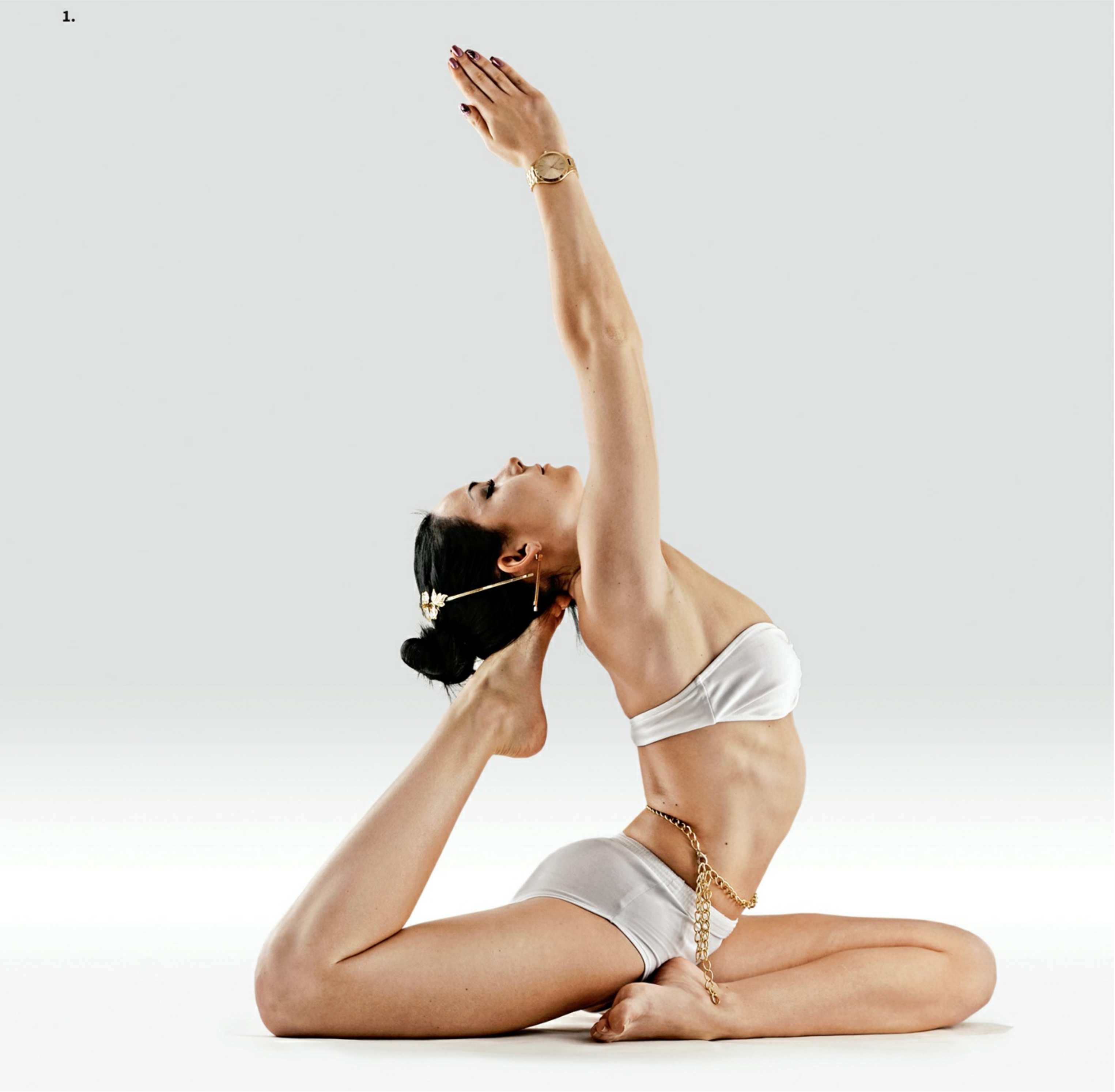

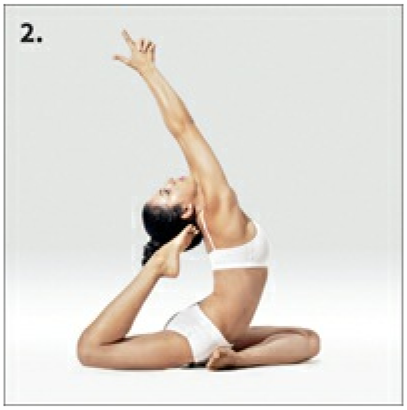

2.

Easy One-Legged King Pigeon Pose 1

Sukha Eka Pada Raja Kapotasana 1

(SUK-uh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: hips lifted, back leg straight, arms open wide

Pose Type: standing (on the knees), backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

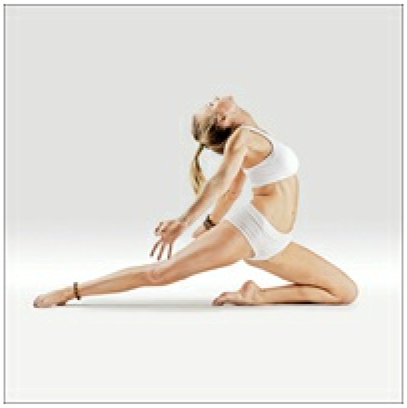

One-Legged King Pigeon Pose 1 Prep.

Eka Pada Raja Kapotasana 1 Prep.

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: hips off the floor, back leg straight; hands on the hips

Pose Type: seated, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

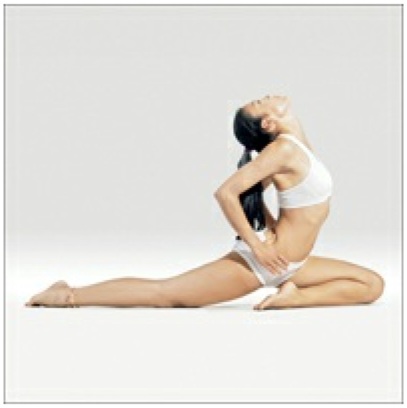

Upward Hands One-Legged King Pigeon Pose 1

Urdhva Hasta Eka Pada Raja Kapotasana 1

(OORD-vuh HUH-stuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: hips off the floor, back leg straight; arms extended to the sky
Pose Type: seated, backbend
Drishti Point: Angushtamadhye or Angustha Ma Dyai (thumbs)
ल व र य ह

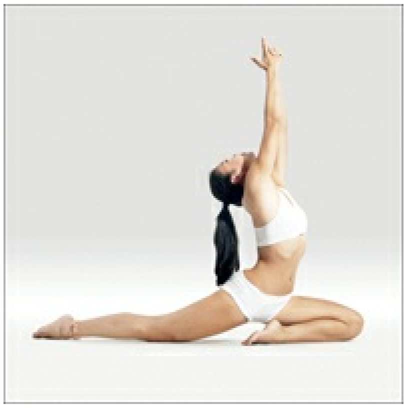

One-Legged King Pigeon Pose 1 Prep.

Eka Pada Raja Kapotasana 1 Prep.

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: back leg straight; palms facing up, back of the hand on the foot and the knee
Pose Type: seated, backbend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य

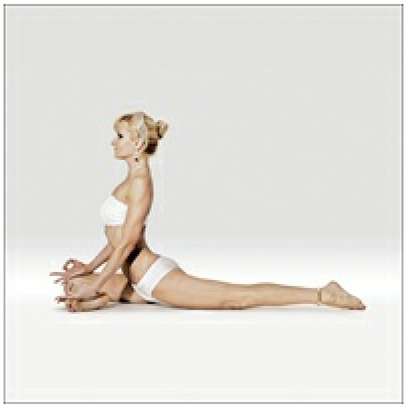

ONE-LEGGED KING PIGEON: BACK LEG STRAIGHT—FORWARD BEND & TWIST

One-Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: back leg straight; forward bend, forearms to the floor, foot and knee to the elbow creases

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

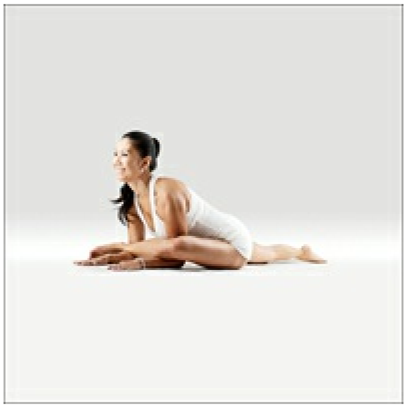

One-Legged King Pigeon Pose 1

Eka Pada Raja Kapotasana 1

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: back leg straight; forehead to the floor, palms to the floor, elbows on top of the wrists

Pose Type: seated, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

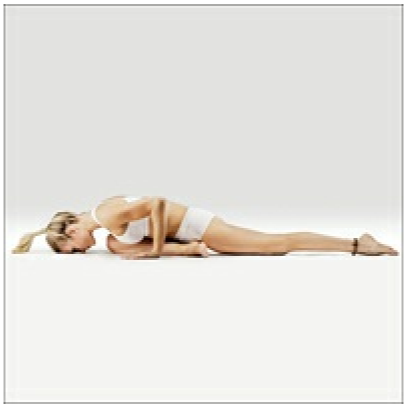

Hand Position of Cow Face Pose in One-Legged King Pigeon Pose 1

Hasta Gomukhasana in Eka Pada Raja Kapotasana 1

(HUH-stuh go-muk-AHS-uh-nuh in EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: back leg straight, forward bend with Gomukhasana Arms
Pose Type: seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य ॐ

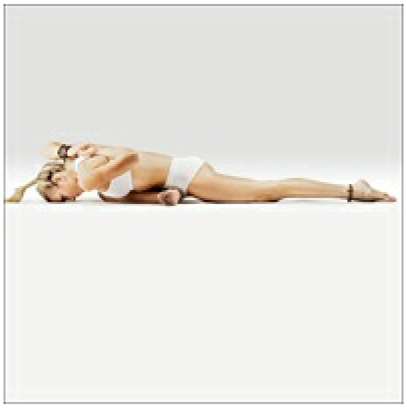

Revolved One-Legged King Pigeon Pose 1

Parivritta Eka Pada Raja Kapotasana 1

(puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)
Modification: twisting to the inside of the body, back leg straight; foot to the armpit
Pose Type: seated, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व

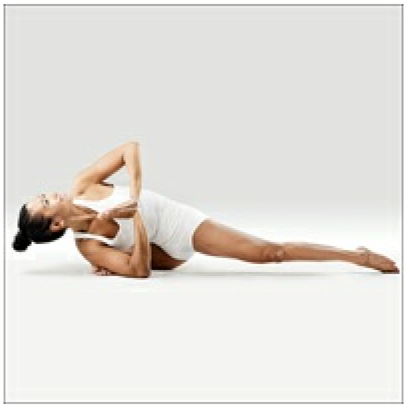

Revolved One-Legged King Pigeon Pose 1

Parivritta Eka Pada Raja Kapotasana 1

(puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)
Modification: twisting to the outside of the body, back leg straight; knee to the armpit
Pose Type: seated, forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

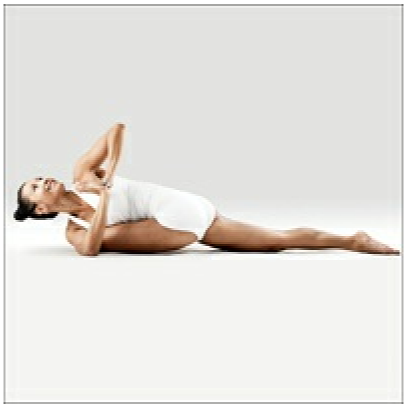

VERTICAL SPLITS: SPINE STRAIGHT

Pose Dedicated to Hanuman

Hanumanasana

(huh-nu-mahn-AHS-uh-nuh)

Also Known As: Pose Dedicated to Hanuman A (Hanumanasana A)
Modification: hands in Anjali Mudra (Hands in Prayer)
1. side view
2. front view
Pose Type: seated
Drishti Point: Nasagrai or Nasagre (nose) or Hastagrai or Hastagre (hands)

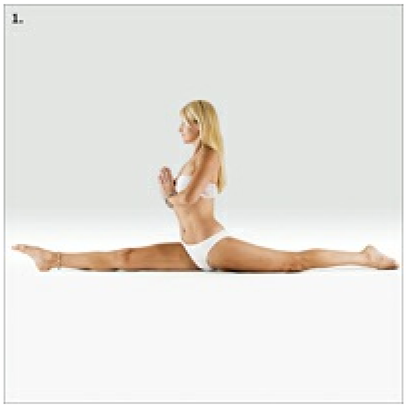

2.

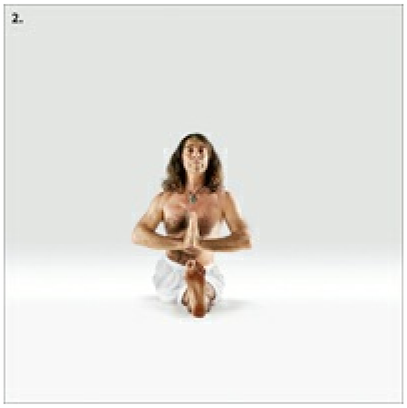

Upward Hands in

Pose Dedicated to Hanuman

Urdhva Hasta Hanumanasana

(OORD-vuh HUH-stuh huh-nu-mahn-AHS-uh-nuh)

Also Known As: Pose Dedicated to Hanuman B (Hanumanasana B)

Pose Type: seated, backbend

Drishti Point: Angusthamadhye or Angustha Ma Dyai (thumbs)

How to Perform the Pose:

1. Begin by standing on your knees with sitting bones lifted off the heels. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale, step your left foot to the front into a lunge. Keep your left knee on top of your left ankle and your right knee under your right hip.
3. Inhale as you slide your left foot out to the front, lifting your left toes off the floor until your left leg is straight.
4. Inhale and stretch your arms up to lengthen your spine. Exhale as you hinge forward from your hips, bringing your hands to the floor on either side of the left leg.
5. On your next exhale, start sliding your left foot out to the front and your right knee to the back, reaching your legs in two opposite directions, coming into a full vertical split. Keep your hips square by pulling your left hip back and your right hip forward.
6. Inhale and reach both your arms up to the sky. Exhale as you deepen the backbend by letting your head roll back and your arms reach in the direction of your back foot.
7. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
8. Inhale as you come out of the backbend. Exhale and place both your palms to the floor on the inside of the left leg. Inhale. As you exhale, swing your left leg back and bring your left knee to the floor by your right knee. Inhale and come up to the starting position with both knees on the floor. Repeat on the other side.

Modification: deep backbend

urdhva = upward

hasta = hand

Hanuman = Hindu Deity, Lord of the Monkeys

Hands Bound Pose Dedicated to Hanuman

Baddha Hasta Hanumanasana

(BUH-duh HUH-stuh huh-nu-mahn-AHS-uh-nuh)

Modification: forward bend

1. side view
2. front view

Pose Type: seated, forward bend

Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

ल व र य ह

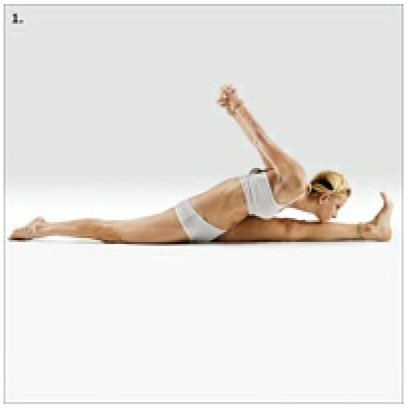

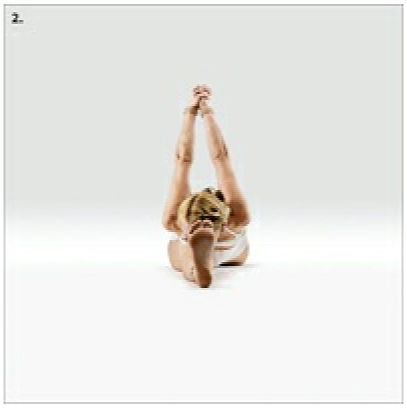

VERTICAL SPLITS: TWISTS & SIDE BENDS

Revolved Pose Dedicated to Hanuman

Parivritta Hanumanasana

(puh-ri-VRIT-uh huh-nu-mahn-AHS-uh-nuh)

Pose Type: seated, twist

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

ल व

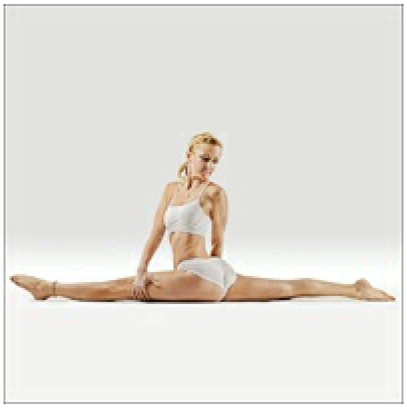

Revolved Sideways Pose Dedicated to Hanuman

Parivritta Parshva Hanumanasana

(puh-ri-VRIT-uh PAHRSH-vuh huh-nu-mahn-AHS-uh-nuh)

Modification: twisting to the inside of the body
Pose Type: seated, forward bend, twist, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

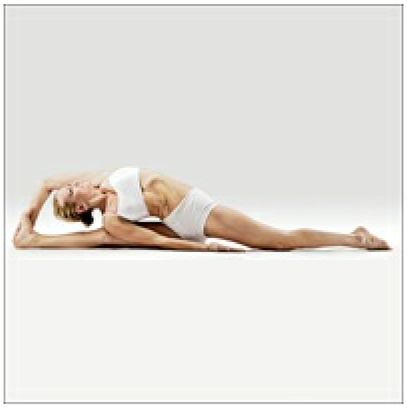

Revolved Sideways Pose Dedicated to Hanuman

Parivritta Parshva Hanumanasana

(puh-ri-VRIT-uh PAHRSH-vuh huh-nu-mahn-AHS-uh-nuh)

Modification: twisting to the outside of the body
Pose Type: seated, forward bend, twist, side bend
Drishti Point: Urdhva or Antara Drishti (up to the sky)

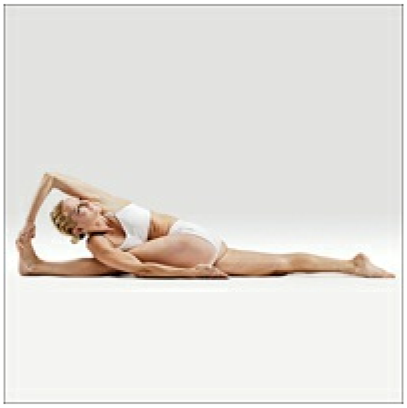

VERTICAL SPLITS: BACK FOOT TOWARD THE HIP & MERMAID POSE

Revolved One-Legged King Pigeon Pose 4 Prep.

Parivritta Eka Pada Raja Kapotasana 4 Prep.

(puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: 1. hand to the floor by the hip, other hand pressing the foot toward the hip
2. grabbing onto the big toe of the front foot on the same side, other hand pressing the back foot toward the hip
Pose Type: seated, backbend, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Parshva Drishti (to the right), Parshva Drishti (to the left)

1.

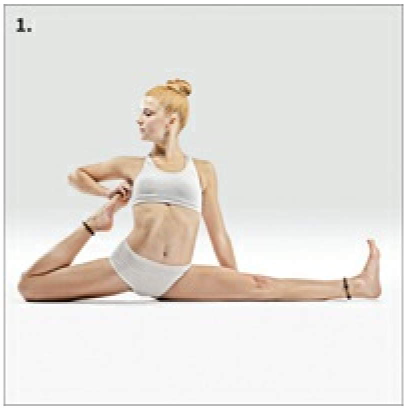

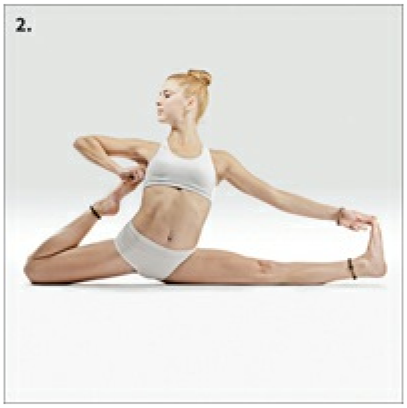

Mermaid Pose 4

Naginyasana 4

(nuh-gin-YAHS-uh-nuh)

Pose Type: seated, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ॐ

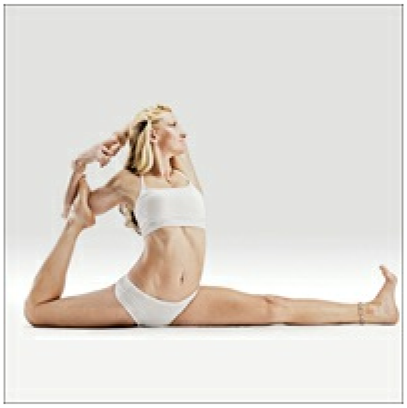

VERTICAL SPLITS: BACK KNEE BENT—OVERHEAD GRIP & FOOT TO THE BACK OF THE HEAD

One-Legged King Pigeon Pose 4

Eka Pada Raja Kapotasana 4

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: one hand grabbing onto the back foot on the same side with overhead grip
Pose Type: seated, backbend
Drishti Point: Hastagrai or Hastagre (hands)
ल व र य ह

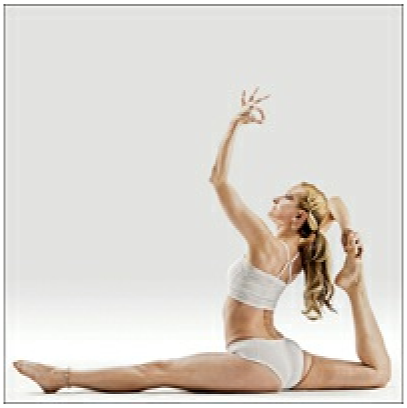

One-Legged King Pigeon Pose 4

Eka Pada Raja Kapotasana 4

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: one hand grabbing onto the back foot on the opposite side with overhead grip, foot reaching toward the head; other hand on the knee of the front leg
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

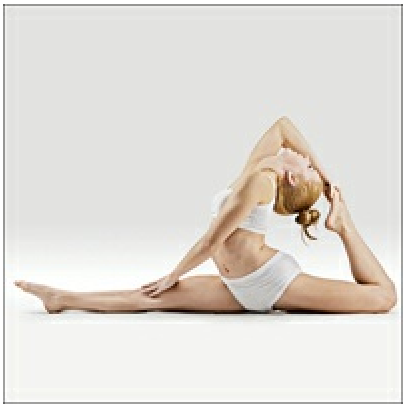

One-Legged King Pigeon Pose 4

Eka Pada Raja Kapotasana 4

(EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Also Known As: One-Legged Stretched Out King Pigeon Pose (Prasarita Pada Raja Kapotasana), Pose of the Lord Kailasha (Kailashasana), Saw Pose A (Kroukachasana A)
Modification: both hands grabbing onto the back foot with overhead grip
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

Foot to the Head One-Legged King Pigeon Pose 4

Shirsha Pada Eka Pada Raja Kapotasana 4

(SHEER-shuh PUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)

Also Known As: Hanuman Salutation Pose (Hanumana Namaskara)
Modification: both arms extended to the sky, back foot to the back of the head
Pose Type: seated, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व र य ह

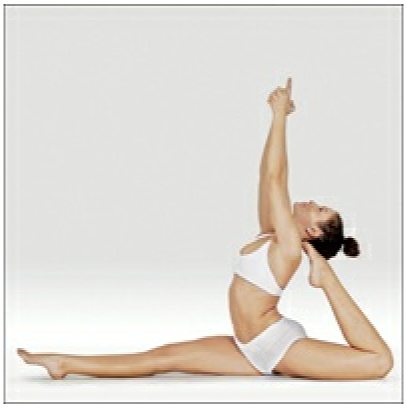

Core Poses

BOAT POSE

SIX TRIANGLES POSE

SUPINE POSES

BOAT POSE: BOTH KNEES BENT

Easy Boat Pose

Sukha Navasana

(SUK-uh nah-VAHS-uh-nuh)

Modification: knees bent, arms wrapped around the shins
Pose Type: core, seated
Drishti Point: Nasagrai or Nasagre (nose)

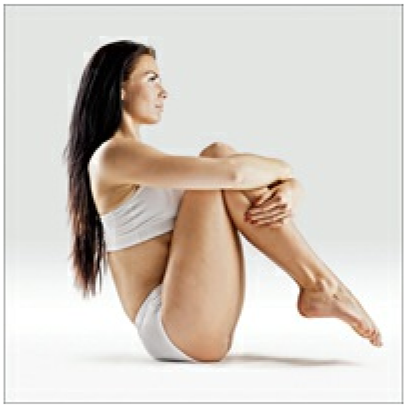

Easy Boat Pose

Sukha Navasana

(SUK-uh nah-VAHS-uh-nuh)

Modification: knees bent, grabbing the back of the thighs, toes touching the floor
Pose Type: core, seated
Drishti Point: Nasagrai or Nasagre (nose)

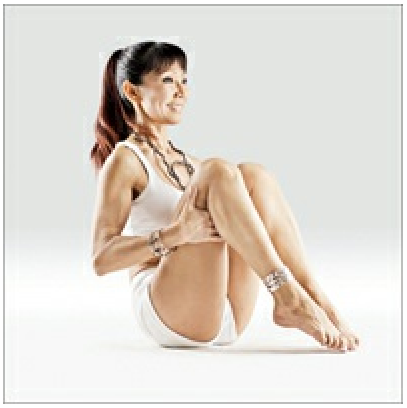

Easy Boat Pose

Sukha Navasana

(SUK-uh nah-VAHS-uh-nuh)

Modification: knees bent, arms wrapped around the back of the thighs, shins parallel to the floor
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

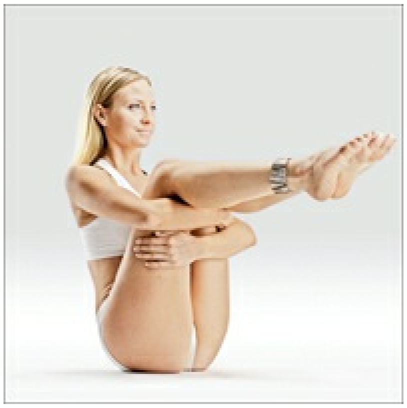

Easy Boat Pose

Sukha Navasana

(SUK-uh nah-VAHS-uh-nuh)

Modification: grabbing onto the heels, toes touching the floor
Pose Type: core, seated
Drishti Point: Nasagrai or Nasagre (nose)

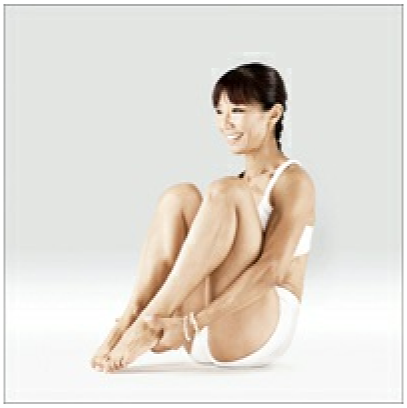

Easy Boat Pose

Sukha Navasana

(SUK-uh nah-VAHS-uh-nuh)

Modification: grabbing onto the outside edges of the feet, shins parallel to the floor

Pose Type: core, seated
Drishti Point: Nasagrai or Nasagre (nose)

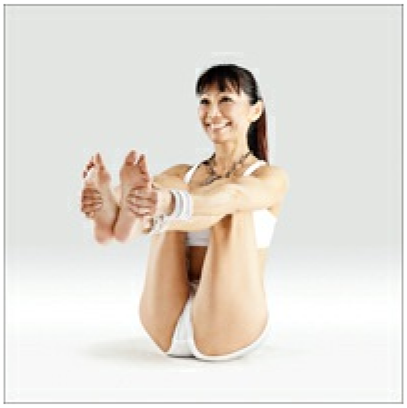

Pendant Pose

Lolasana

(lo-LAHS-uh-nuh)
Modification: knees in front of chest, ankles crossed
Pose Type: core, arm balance
Drishti Point: Nasagrai or Nasagre (nose)

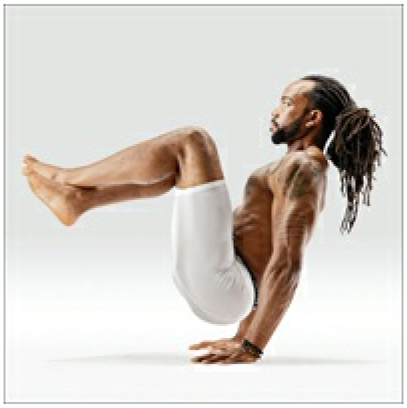

Boat Pose

Navasana

(nah-VAHS-uh-nuh)

Pose Type: core, seated

Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

How to Perform the Pose:

1. Begin by sitting on the floor with both your legs straight out in front of you. Keep your hands on the floor at the sides of your hips for support. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale, bend your knees, and slide your feet toward your sitting bones. Keep your feet and knees together.
3. Inhale as you lengthen your spine. Engage your core and lean back. Make sure your spine is straight and your lower back is off the floor.
4. Exhale and lift your feet off the floor until your shins are parallel to the floor.
5. Once you have your balance, exhale and lift your arms off the floor, keeping them parallel to the floor.
6. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of pose.
7. Inhale as you release the pose by lowering your hands and feet to the floor. Exhale and slide your feet out in front until your legs are straight to come back to the starting position.

Variation: arms straight and parallel to the floor, knees bent, shins parallel to the floor

nava = boat

Easy Boat Pose

Sukha Navasana

(SUK-uh nah-VAHS-uh-nuh)

Modification: knees bent, ankles crossed, arms straight and parallel to the floor
Pose Type: core, seated
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
लं रं

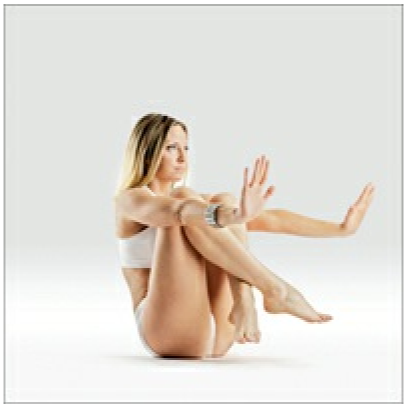

Easy Boat Pose

Sukha Navasana

(SUK-uh nah-VAHS-uh-nuh)

Modification: knees bent, fingertips to the floor, toes touching the floor
Pose Type: core, seated
Drishti Point: Nasagrai or Nasagre (nose)
लं रं यं

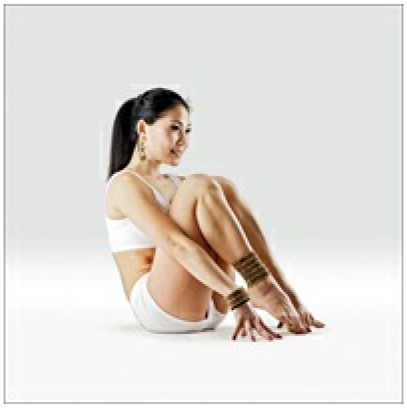

Easy Boat Pose

Sukha Navasana

(SUK-uh nah-VAHS-uh-nuh)

Modification: knees bent, fingertips to the floor, shins parallel to the floor
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल र य

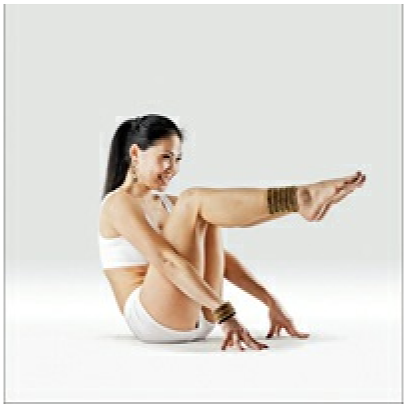

BOAT POSE: BOTH LEGS STRAIGHT

Supported Boat Pose

Salamba Navasana

(SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: palms to the floor behind the hips
1. both legs straight, fingertips pointing toward the hips
2. both legs straight, both arms straight, fingertips pointing to the back
Pose Type: core, seated
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)
ल ॐ

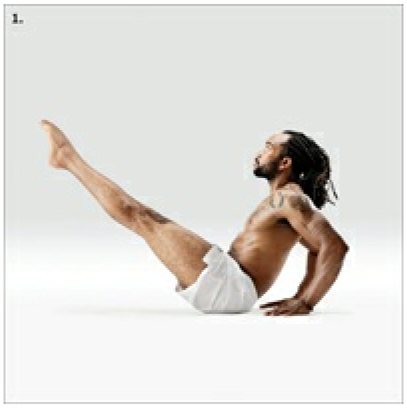

Supported Boat Pose

Salamba Navasana

(SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: both arms wrapped around both legs, legs straight
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

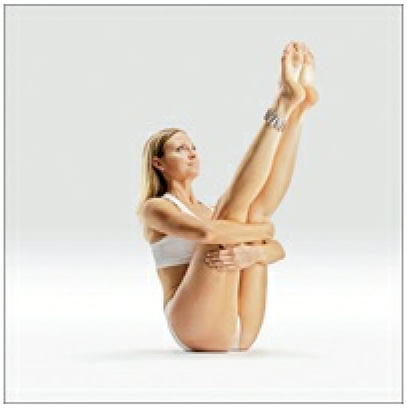

Supported Boat Pose—One-Handed

Eka Hasta Salamba Navasana

(EY-kuh HUH-stuh SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: both legs straight; one arm straight, fingertips to the sky; other arm wrapping behind the ankles of both legs
Pose Type: core, seated
Drishti Point: Urdhva or Antara Drishti (up to the sky) or Hastagrai or Hastagre (hands)

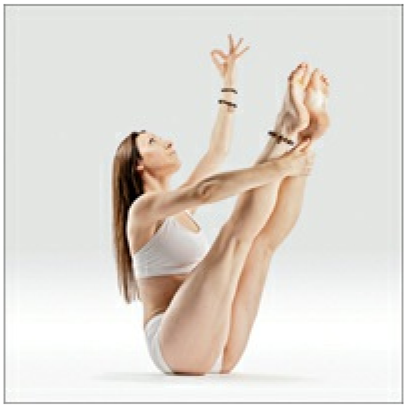

Supported Boat Pose

Salamba Navasana

(SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: fingertips on the floor, pointing to the front
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

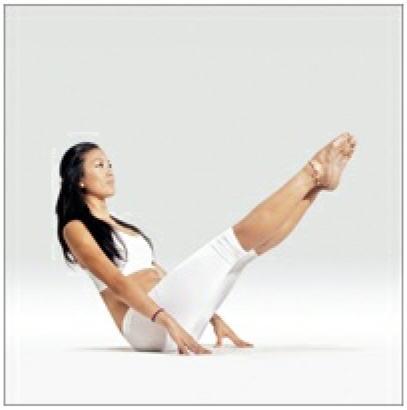

Supported Boat Pose

Salamba Navasana

(SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: grabbing behind the knees, legs straight
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

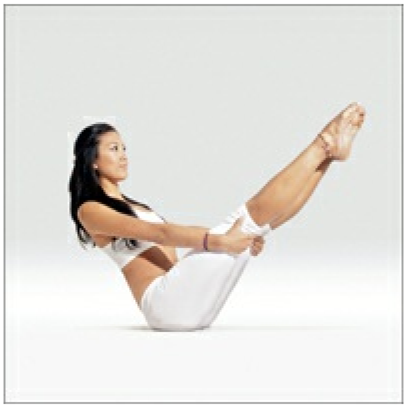

Complete Boat Pose

Paripurna Navasana

(puh-ri-POOR-nuh nah-VAHS-uh-nuh)

Modification: arms straight and parallel to the floor, legs straight
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल

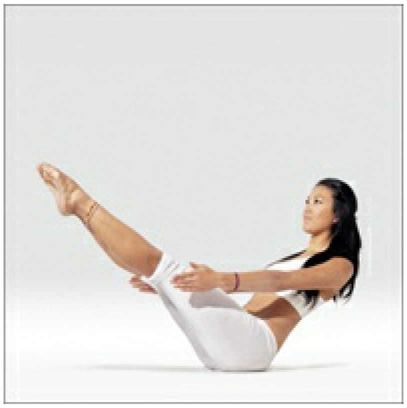

Half Boat Pose

Ardha Navasana

(UHR-duh nah-VAHS-uh-nuh)

Modification: fingers interlocked behind the head
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल र ह

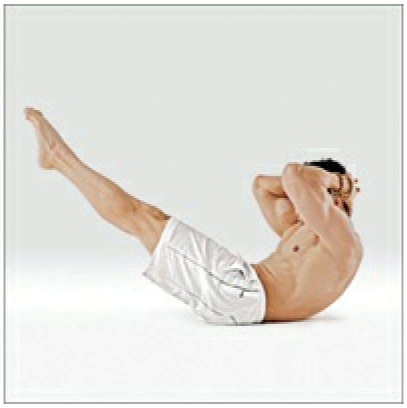

Boat Pose

Navasana

(nah-VAHS-uh-nuh)

Modification: palms and elbows together, elbows bent
Pose Type: core, seated
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

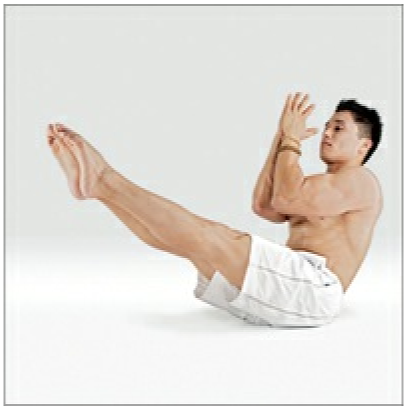

Unsupported Boat Pose

Niralamba Navasana

(nir-AH-luhm-buh nah-VAHS-uh-nuh)

Modification: arms shoulder width apart
Pose Type: core, seated
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

BOAT POSE: TWIST

Revolved Boat Pose with Hands in Prayer

Parivritta Navasana Namaskar

(puh-ri-VRIT-tuh nah-VAHS-uh-nuh nuh-muhs-KAHR)

Pose Type: core, seated, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

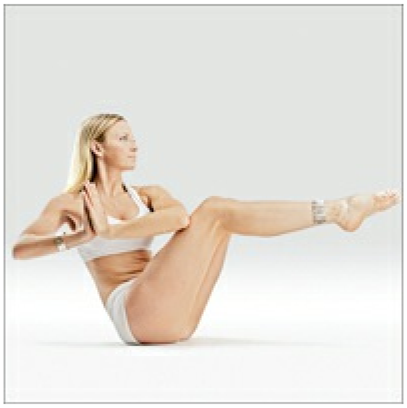

Supported Boat Pose—One-Handed

Eka Hasta Salamba Navasana

(EY-kuh HUH-stuh SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: both legs straight, ankles crossed; one arm straight, palm to the floor behind the hips; other arm wrapping behind the ankles of both legs
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

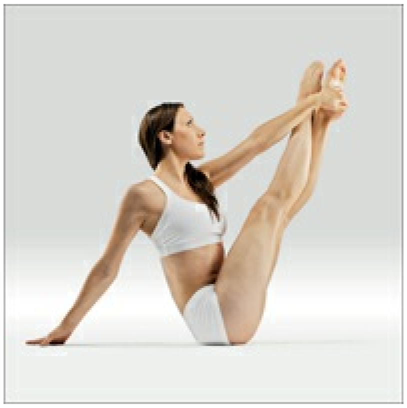

Revolved Supported Boat Pose—One-Handed

Parivritta Eka Hasta Salamba Navasana

(puh-ri-VRIT-tuh EY-kuh HUH-stuh SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: both legs straight, ankles crossed; arms straight, fingertips to the floor behind the hips, one hand grabbing onto the heel
Pose Type: core, seated, twist
Drishti Point: Hastagrai or Hastagre (hands)

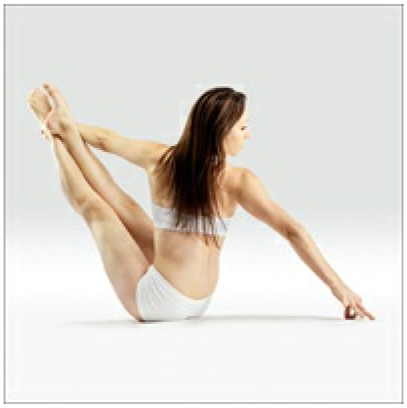

Revolved Supported Boat Pose—One-Handed

Parivritta Eka Hasta Salamba Navasana

(puh-ri-VRIT-tuh EY-kuh HUH-stuh SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: one foot under the knee, other leg straight; one arm threaded under the bent leg, fingertips of the other hand on the floor
Pose Type: core, seated, twist
Drishti Point: Hastagrai or Hastagre (hands)

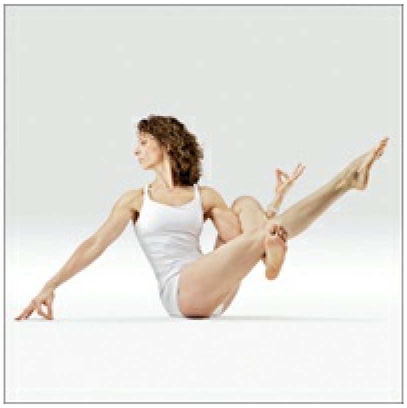

Revolved Supported Boat Pose

Parivritta Salamba Navasana

(puh-ri-VRIT-tuh SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: ankles crossed, legs wrapped around the forearm
Pose Type: core, seated, twist
Drishti Point: Parshva (far to the right), Parshva (far to the left)

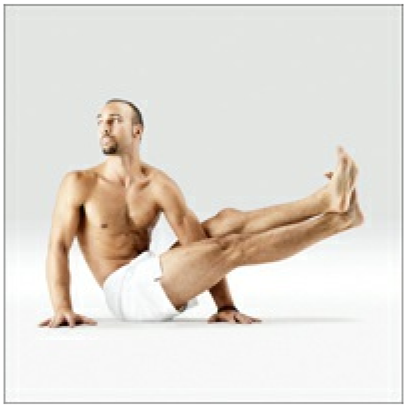

BOAT POSE: TWIST, SCISSOR LEGS

Revolved Boat Pose

Parivritta Navasana

(puh-ri-VRIT-tuh nah-VAHS-uh-nuh)

Modification: one leg over the shoulder, toes of the other leg to the floor
1. front view
2. back view
Pose Type: core, seated, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

1.

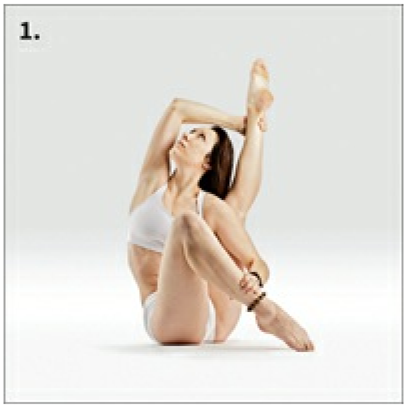

2.

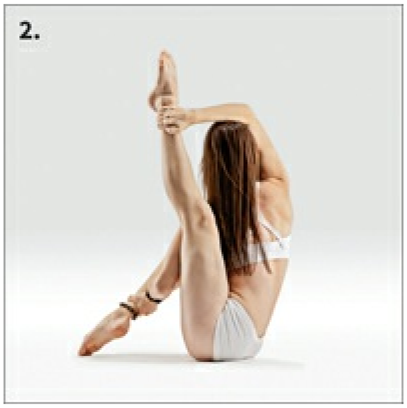

Revolved Boat Pose

Parivritta Navasana

(puh-ri-VRIT-tuh nah-VAHS-uh-nuh)

Modification: one leg over the shoulder, both legs straight
1. front view
2. back view
Pose Type: core, seated, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

1.

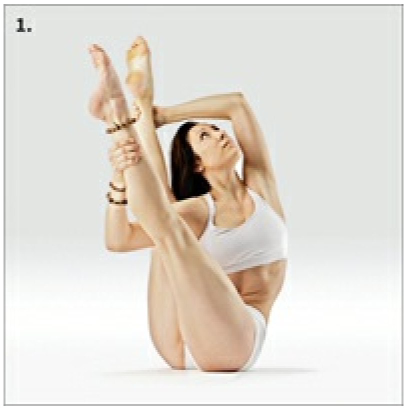

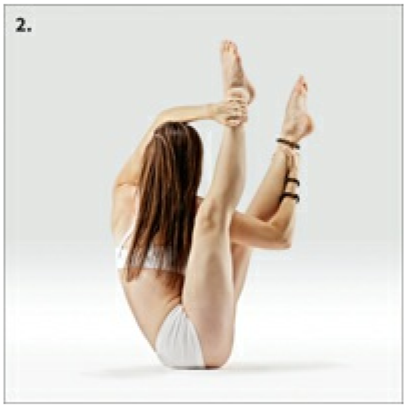
2.

BOAT POSE: ONE LEG STRAIGHT, ONE LEG BENT

One-Legged Supported Boat Pose

Eka Pada Salamba Navasana

(EY-kuh PUH-duh SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: one leg straight, one knee bent and dropped to the side for support
Pose Type: core, seated
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Padayoragrai or Padayoragre (toes/feet)
● ● ●

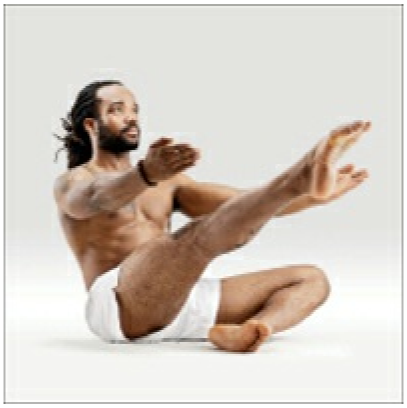

One-Legged Supported Boat Pose

Eka Pada Salamba Navasana

(EY-kuh PUH-duh SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: grabbing onto the back of the thighs for support; one leg straight, one leg bent—toes touching the floor
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
●

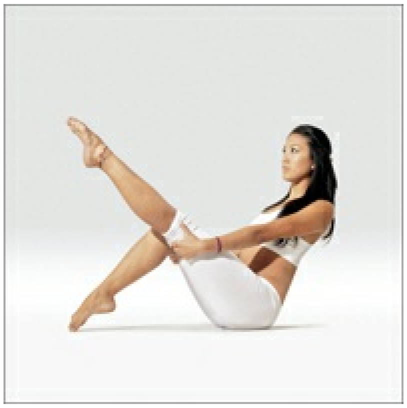

One-Legged Unsupported Boat Pose

Eka Pada Niralamba Navasana

(EY-kuh PUH-duh nir-AH-luhm-buh nah-VAHS-uh-nuh)

Modification: one leg straight; one knee bent toward the chest, foot off the floor
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

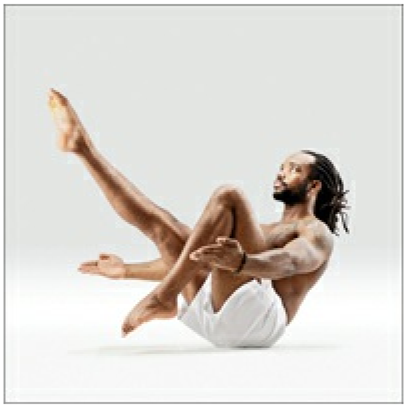

BOAT POSE: SCISSOR LEGS

One-Legged Supported Boat Pose—One-Handed

Eka Pada Eka Hasta Salamba Navasana

(EY-kuh PUH-duh EY-kuh HUH-stuh SAH-luhm-buh nah-VAHS-uh-nuh)

Modification: grabbing the outside edge of the foot with the opposite hand for support, leg straight, other arm up to the sky
1. other leg bent, toes touching the floor
2. other leg straight, foot off the floor
Pose Type: core, seated
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or

Padayoragre (toes/feet)

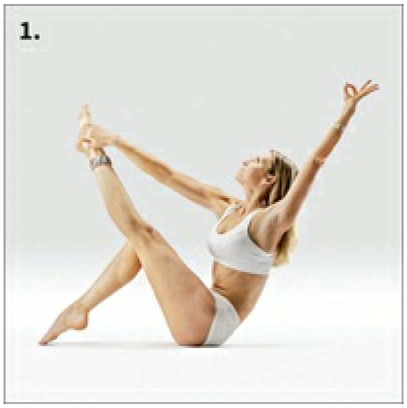
1.

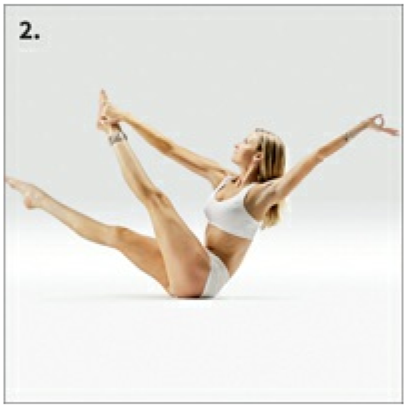
2.

BOAT POSE: ONE FOOT TO KNEE

Upward Half Firelog Western Intense Stretch Pose

Urdhva Ardha Agnistambha Paschimottanasana

(OORD-vuh UHR-duh uhg-ni-STUHM-buh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward Half Firelog Seated Forward Bend
Modification: 1. forehead to the shin
2. forehead away from the shin
Pose Type: core, seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Padayoragrai or Padayoragre (toes/feet)

Baby Cradle Pose in Boat Pose

Hindolasana in Navasana

(hin-do-LAHS-uh-nuh in nah-VAHS-uh-nuh)

Modification: foot of the bent leg to the back of the knee of the straight leg
Pose Type: core, seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
लं वं

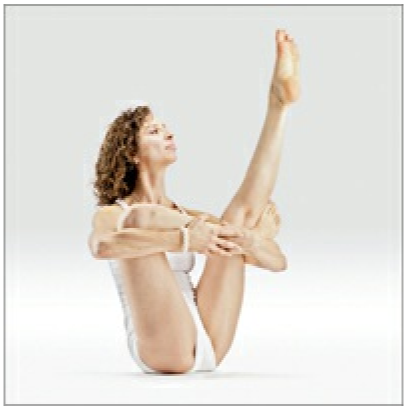

Revolved Yogic Staff Boat Pose

Parivritta Yogadanda Navasana

(puh-ri-VRIT-tuh yo-guh-DUHN-duh nah-VAHS-uh-nuh)

Modification: both knees bent; one arm behind the back, grabbing onto the ankle; other arm grabbing the outside edge of the foot

Pose Type: core, seated, twist, binding

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

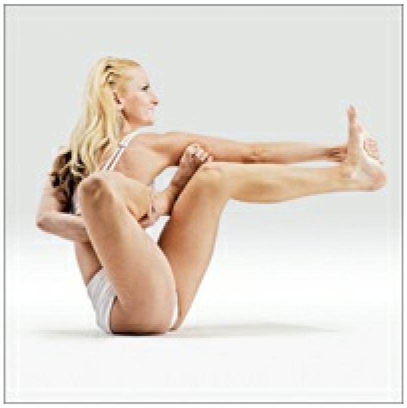

Revolved Bound Yogic Staff Boat Pose

Parivritta Baddha Yogadanda Navasana

(puh-ri-VRIT-tuh BUH-duh yo-guh-DUHN-duh nah-VAHS-uh-nuh)

Modification: binding around the straight leg

Pose Type: core, seated, twist, binding

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

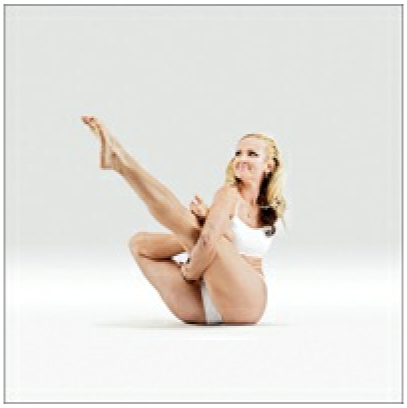

Revolved Yogic Staff Boat Pose

Parivritta Yogadanda Navasana

(puh-ri-VRIT-tuh yo-guh-DUHN-duh nah-VAHS-uh-nuh)

Modification: foot to the triceps; one arm grabbing onto the heel of the straight leg, other arm open to

the side
Pose Type: core, seated, twist
Drishti Point: Hastagrai or Hastagre (hands)
ल व

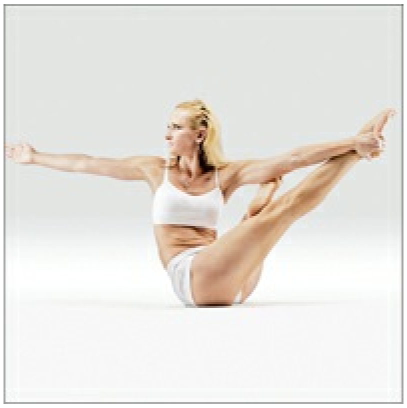

BOAT POSE: ONE FOOT TO HIP SOCKET

Upward One-Legged Half Lotus Western Intense Stretch Pose

Urdhva Eka Pada Ardha Padma Paschimottanasana

(OORD-vuh EY-kuh PUH-duh UHR-duh PUHD-muh puhsh-chi-mo-tahn-AHS-uh-nuh)
Also Known As: Upward One-Legged Half Lotus Seated Forward Bend
Modification: one hand grabbing onto the big toe, one hand on the hip
Pose Type: core, seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व

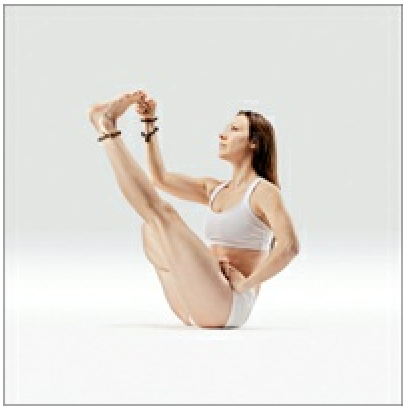

Upward One-Legged Half Bound Lotus Western Intense Stretch Pose

Urdhva Eka Pada Ardha Baddha Padma Paschimottanasana

(OORD-vuh EY-kuh PUH-duh UHR-duh BUH-duh PUHD-muh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward One-Legged Half Bound Lotus Seated Forward Bend
Pose Type: core, seated, forward bend, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

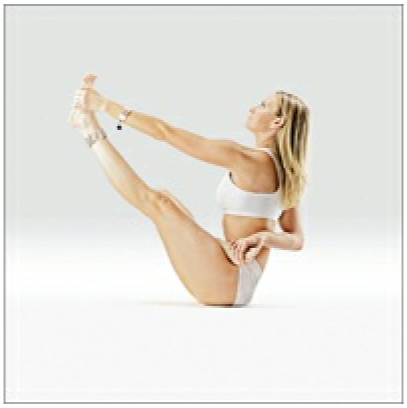

Upward One-Legged Half Bound Lotus Western Intense Stretch Pose

Urdhva Eka Pada Ardha Baddha Padma Paschimottanasana

(OORD-vuh EY-kuh PUH-duh UHR-duh BUH-duh PUHD-muh puhsh-chi-mo-tahn-AHS-uh-nuh)
Also Known As: Upward One-Legged Half Bound Lotus Seated Forward Bend
Modification: grabbing onto the foot, hand between the calf muscle and the hamstring
Pose Type: core, seated, forward bend, binding
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

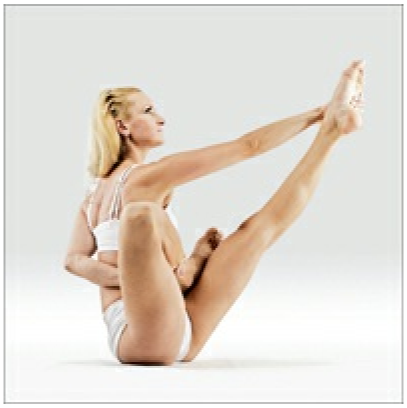

Upward One-Legged Half Lotus Western Intense Stretch Pose

Urdhva Eka Pada Ardha Padma Paschimottanasana

(OORD-vuh EY-kuh PUH-duh UHR-duh PUHD-muh puhsh-chi-mo-tahn-AHS-uh-nuh)
Also Known As: Upward One-Legged Half Lotus Seated Forward Bend
Modification: 1. grabbing onto the wrist, hands over the foot

2. grabbing onto the foot, nose to the shin
Pose Type: core, seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet) or Nasagrai or Nasagre (nose)
ल व र

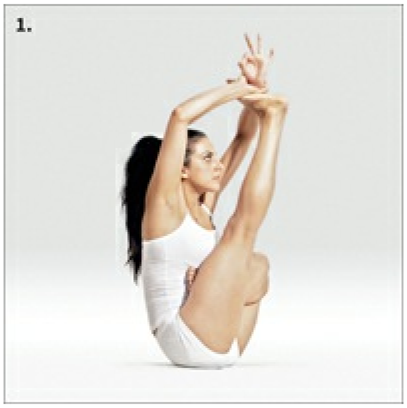
1.

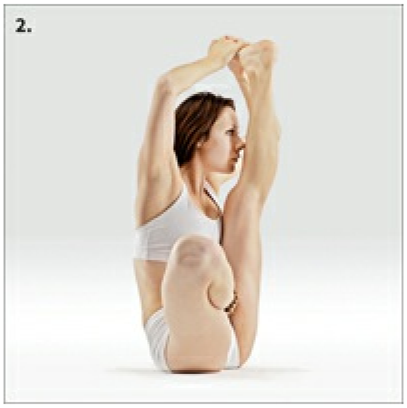
2.

BOAT POSE: GRABBING ONTO FEET/TOES

Big Toes Pose

Ubhaya Padangusthasana

(u-beye-uh puhd-ahng-goosh-TAHS-uh-nuh)
Modification: grabbing onto the big toes, toes flexed in
Pose Type: core, seated
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल

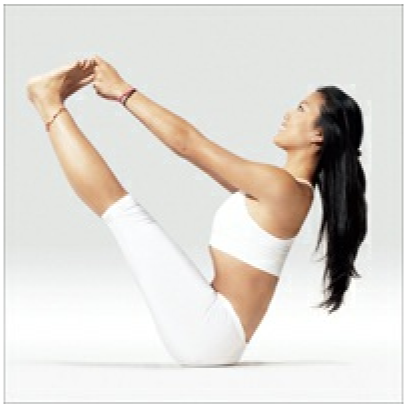

Upward Facing Western Intense Stretch

Urdhva Mukha Paschimottanasana

(OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward Facing Seated Forward Bend
Modification: grabbing onto the heels, arms straight
Pose Type: core, seated, forward bend, quadriceps to the chest
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

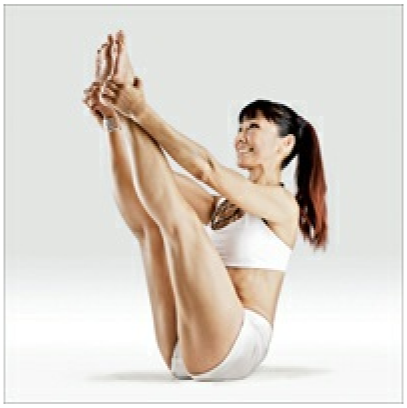

Upward Facing Western Intense Stretch

Urdhva Mukha Paschimottanasana

(OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward Facing Seated Forward Bend
Modification: grabbing onto the heels
1. forehead away from the shins
2. forehead to the shins
Pose Type: core, seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

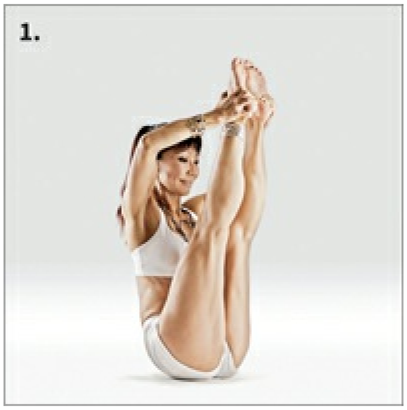

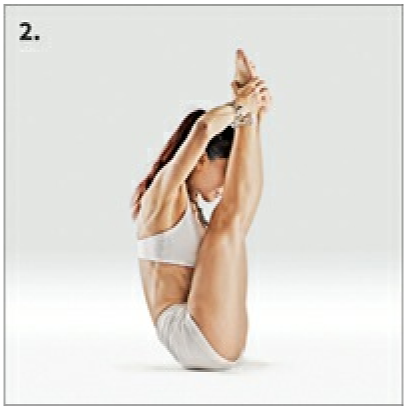

Upward Facing Supported Western Intense Stretch

Urdhva Mukha Salamba Paschimottanasana

(OORD-vuh MUK-uh SAH-luhm-buh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward Facing Seated Forward Bend
Modification: arms reaching to the front, fingertips to the floor, 1. forehead away from the shins 2. forehead to the shins
Pose Type: core, seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet) or Nasagrai or Nasagre (nose)

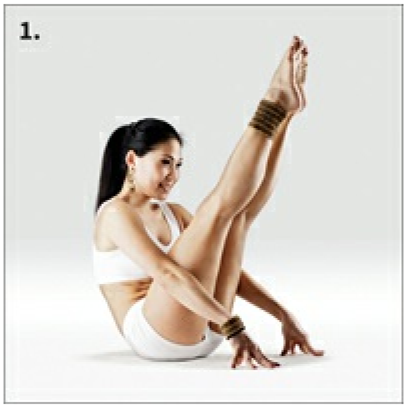

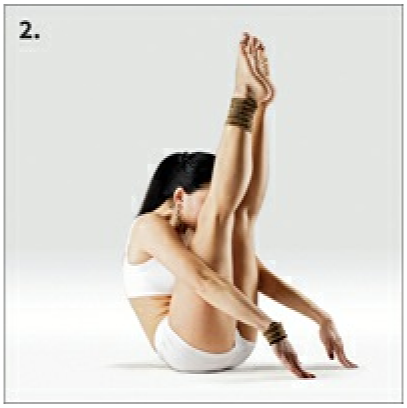

Bound Hands Upward Facing Western Intense Stretch

Baddha Hasta Urdhva Mukha Paschimottanasana

(BUH-duh HUH-stuh OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Bound Hands Upward Facing Seated Forward Bend
Pose Type: core, seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

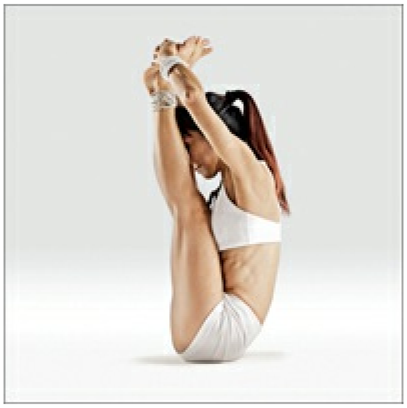

Upward Facing Western Intense Stretch

Urdhva Mukha Paschimottanasana

(OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward Facing Seated Forward Bend
Modification: grabbing onto the ankles, forehead to the shins
Pose Type: core, seated, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

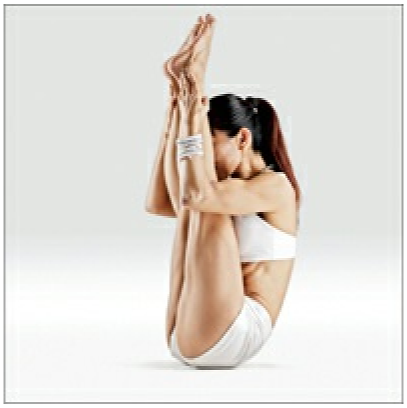

Upward Facing Western Intense Stretch

Urdhva Mukha Paschimottanasana

(OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward Facing Seated Forward Bend
Modification: grabbing onto the outside edges of the feet, feet hip distance apart, toes pointed
Pose Type: core, seated, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

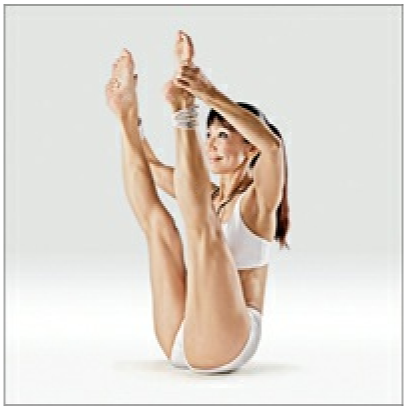

ARM BALANCE: SITTING BONES & FEET OFF THE FLOOR

Celibate Pose

Brahmacharyasana

(bruh-muh-chahr-YAHS-uh-nuh)

Also Known As: Lifted Up Staff Pose (Dandasana Utpluti)
Modification: fingers to the floor, thumbs pointing to the back
Pose Type: core, arm balance
Drishti Point: Nasagrai or Nasagre (nose)

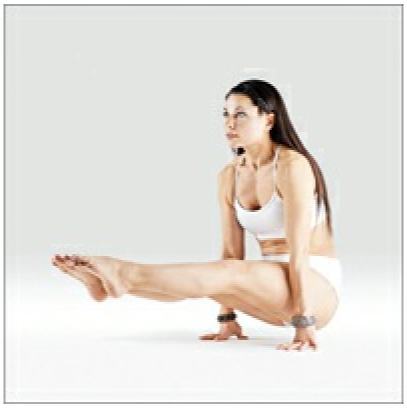

SIX TRIANGLES POSE

Six Triangles Pose

Shatkonasana

(shuht-ko-NAHS-uh-nuh)

Modification: top foot on the bottom knee
1. fingertips to the temples
2. one forearm to the floor, other arm wrapped around the top knee
3. fingertips to the temples, elbow toward the top knee
4. one elbow to the floor, other elbow to the top knee on the same side

Pose Type: core, seated, twist

Drishti Point: 1. Urdhva or Antara Drishti (up to the sky)
2. Nasagrai or Nasagre (nose)
3. Nasagrai or Nasagre (nose)
4. Hastagrai or Hastagre (hands)

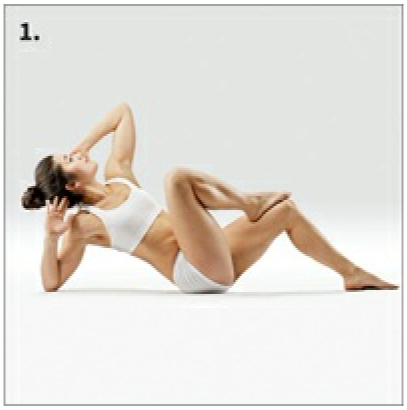

2.

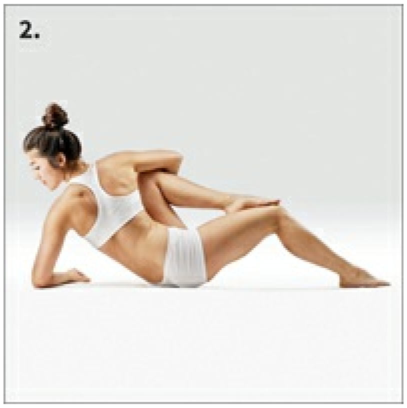

3.

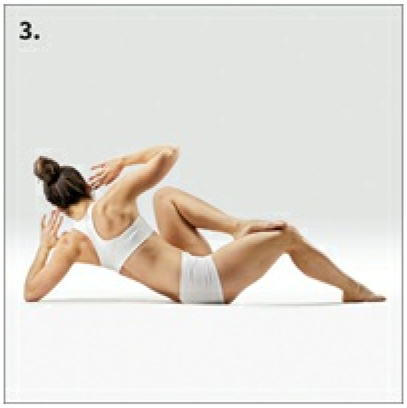

4.

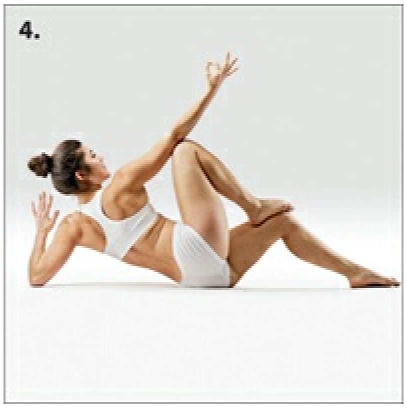

SUPINE: ONE LEG STRAIGHT, ONE LEG BENT

Half Wind Relieving Pose

Ardha Vayu Muktyasana

(UHR-duh VAH-yu muk-TYAHS-uh-nuh)

Also Known As: Wind-Relieving Pose (Pavana Muktasana)
Pose Type: core, supine
Drishti Point: Nasagrai or Nasagre (nose)
ल व र य ह

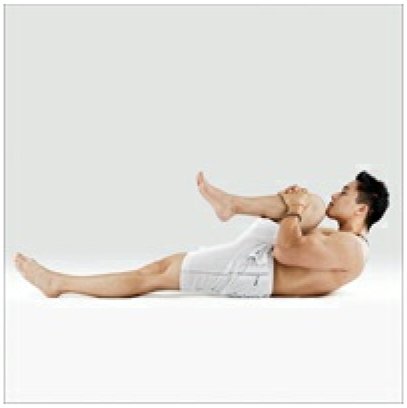

Revolved Half Wind Relieving Pose

Parivritta Ardha Vayu Muktyasana

(puh-ri-VRIT-tuh UHR-duh VAH-yu muk-TYAHS-uh-nuh)

Modification: one knee bent, one leg straight
Pose Type: core, supine, twist
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye (third eye, between the eyebrows)

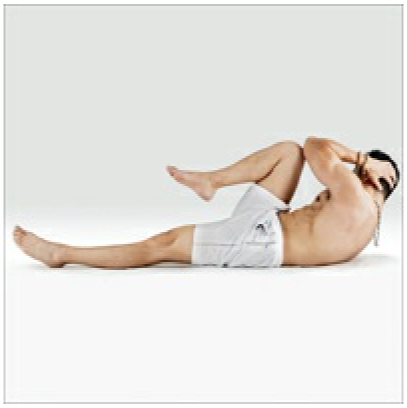

SUPINE: BOTH LEGS STRAIGHT “SCISSORS”

One-Legged Upward Extended Legs Pose

Eka Pada Urdhva Prasarita Padasana

(EY-kuh PUH-duh OORD-vuh pruh-SAH-ri-tuh puh-DAHS-uh-nuh)

Modification: 1. one arm extended to sky, other palm to the floor, head on the floor
2. one arm extended to sky, other palm to the floor, head and shoulders off the floor
3. fingers interlocked behind the back of the head, head and shoulders off the floor
Pose Type: core, supine
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

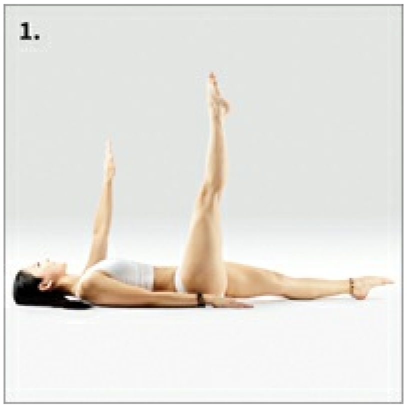

1.

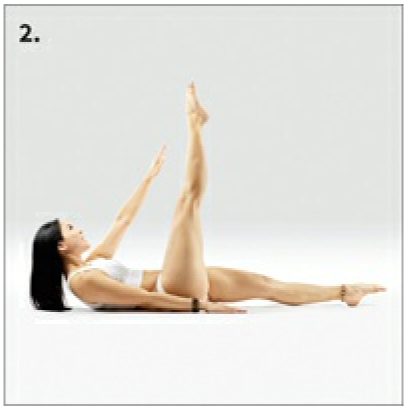

2.

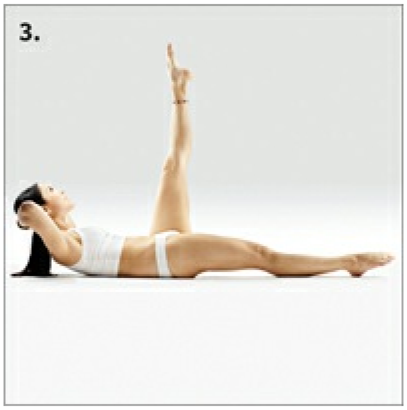

3.

SUPINE: BOTH LEGS STRAIGHT

Abdominal Lift Pose

Jatharasana

(jah-tuh-RAHS-uh-nuh)

Modification: head on the floor, palms down by the hips
Pose Type: core, supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

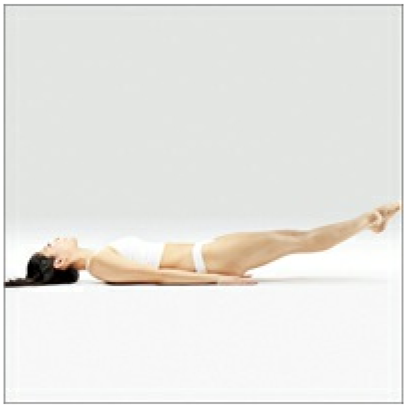

Upward Extended Legs Pose

Urdhva Prasarita Padasana

(OORD-vuh pruh-SAH-ri-tuh puh-DAHS-uh-nuh)

Pose Type: core, supine

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

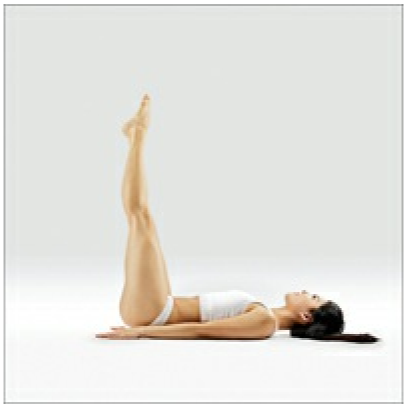

Abdominal Lift Pose

Jatharasana

(jah-tuh-RAHS-uh-nuh)

Modification: head and shoulders off the floor, feet off the floor

Pose Type: core, supine

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

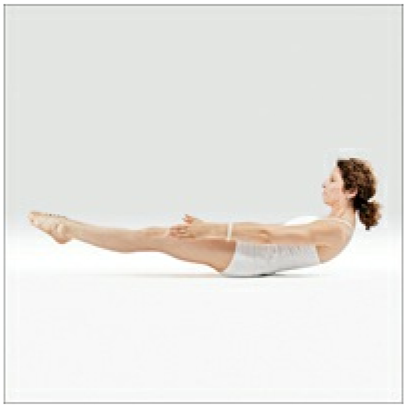

SUPINE: BOTH LEGS STRAIGHT

Two-Legged Abdominal Lift Upward Extended Legs Pose

Dwi Pada Jathara Urdhva Prasarita Padasana

(DWI-puh-duh JAH-tuh-ruh OORD-vuh pruh-SAH-ri-tuh puh-DAHS-uh-nuh)

Modification: with a yoga block
Pose Type: core, supine
Drishti Point: Urdhva or Antara Drishti (up to the sky), Padayoragrai or Padayoragre (toes/feet)
● ●

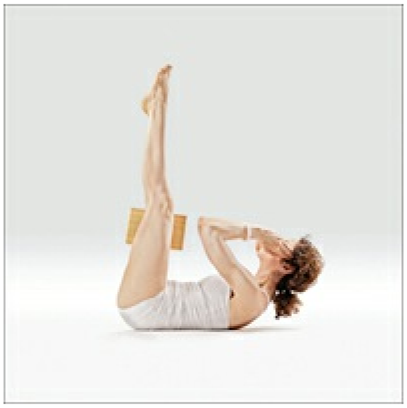

Abdominal Lift Upward Facing Western Intense Stretch

Jathara Urdhva Mukha Paschimottanasana

(JAH-tuh-ruh OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Abdominal Lift Upward Forward Bend
Modification: head lifted off the floor, forehead to the shins
Pose Type: core, supine
Drishti Point: Nasagrai or Nasagre (nose)
● ● ● ● ●

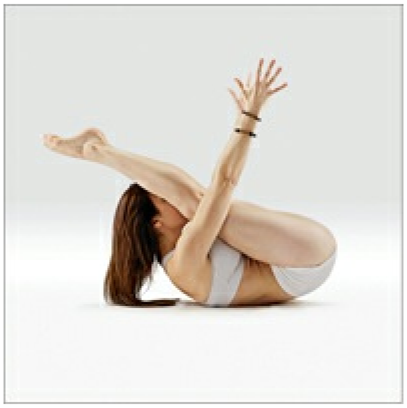

Abdominal Lift Upward Facing Western Intense Stretch

Jathara Urdhva Mukha Paschimottanasana

(JAH-tuh-ruh OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Abdominal Lift Upward Forward Bend
Modification: head lifted off the floor, legs crossed modification
Pose Type: core, supine
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
● ● ●

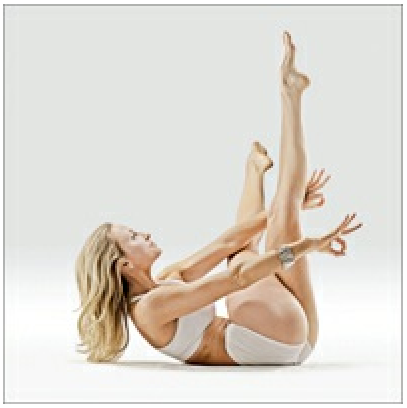

SUPINE: ONE LEG STRAIGHT, ONE KNEE BENT

Revolved Stomach One Leg Extended Pose

Parivritta Jathara Utthita Eka Padasana

(puh-ri-VRIT-tuh JAH-tuh-tuh UT-ti-tuh EY-kuh puh-DAHS-uh-nuh)

Modification: foot to the elbow
Pose Type: core, supine, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
● ● ●

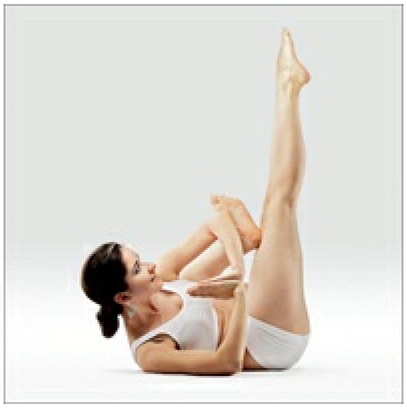

SUPINE: GARUDA LEGS ON THE BACK

Reclining Abdominal Lift Pose Dedicated to Garuda

Supta Jathara Garudasana

(SUP-tuh JAH-tuh-ruh guh-ru-DAHS-uh-nuh)

Pose Type: core, supine
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

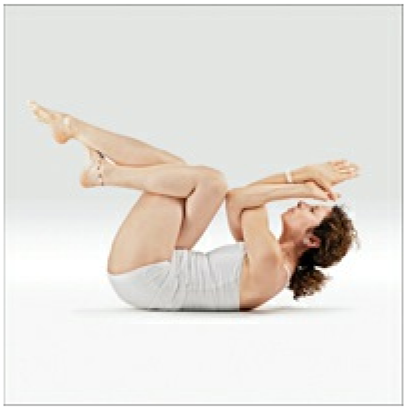

Reclining Abdominal Lift Pose Dedicated to Garuda

Supta Jathara Garudasana

(SUP-tuh JAH-tuh-ruh guh-ru-DAHS-uh-nuh)

Modification: fingers to the temples, elbows to the knees
Pose Type: core, supine
Drishti Point: Urdhva or Antara Drishti (up to the sky), Padayoragrai or Padayoragre (toes/feet)

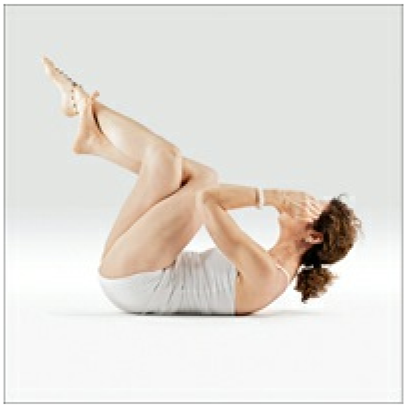

Abdominal Lift Pose Dedicated to Garuda

Jathara Garudasana

(JAH-tuh-ruh guh-ru-DAHS-uh-nuh)

Modification: palms to the floor, arms straight, sitting bones lifted off the floor
Pose Type: core, supine
Drishti Point: Urdhva or Antara Drishti (up to the sky), Padayoragrai or Padayoragre (toes/feet)

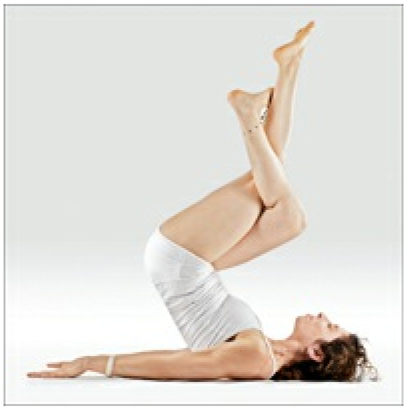

Revolved Stomach Reclining Pose Dedicated to Garuda

Parivritta Jathara Supta Garudasana

(puh-ri-VRIT-tuh JAH-tuh-ruh SUP-tuh guh-ru-DAHS-uh-nuh)

Modification: fingers to the temples, elbows out to the sides
Pose Type: core, supine, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

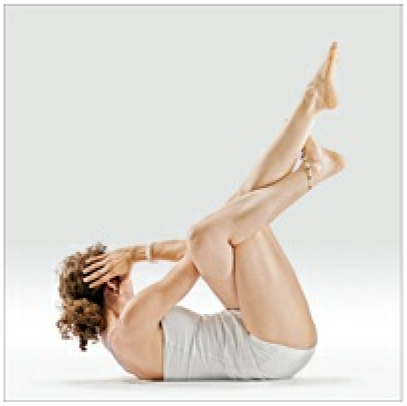

SUPINE: FULL LOTUS

Reclining Abdominal Lift Lotus Pose

Supta Jathara Padmasana

(SUP-tuh JAH-tuh-ruh puhd-MAHS-uh-nuh)

Modification: fingers to the temples, knees up to the sky
Pose Type: core, supine
Drishti Point: Nasagrai or Nasagre (nose)
लं वं रं हं

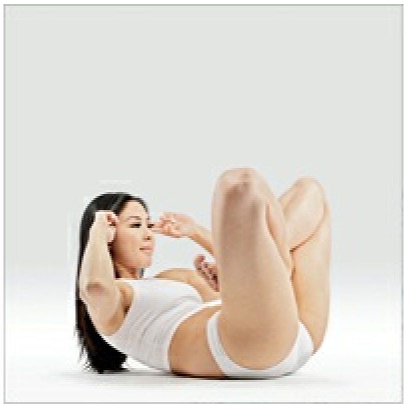

Revolved Stomach Reclining Lotus Pose

Parivritta Jathara Supta Padmasana

(puh-ri-VRIT-tuh JAH-tuh-ruh SUP-tuh puhd-MAHS-uh-nuh)

Modification: fingers to the temples, elbows out to the sides
1. back view
2. front view
Pose Type: core, supine, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

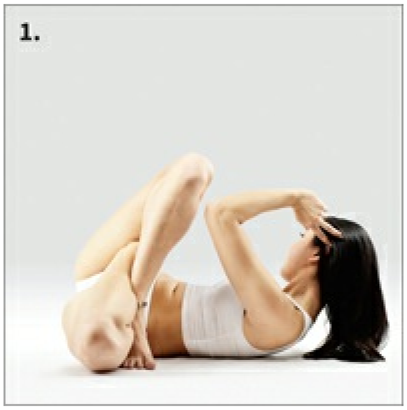
1.

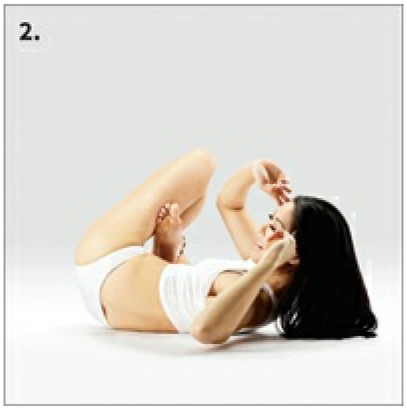
2.

Abdominal Lift Lotus Pose

Jathara Padmasana

(JAH-tuh-tuh puhd-MAHS-uh-nuh)

Modification: palms on the floor, arms straight, sitting bones lifted off the floor
Pose Type: core, supine
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

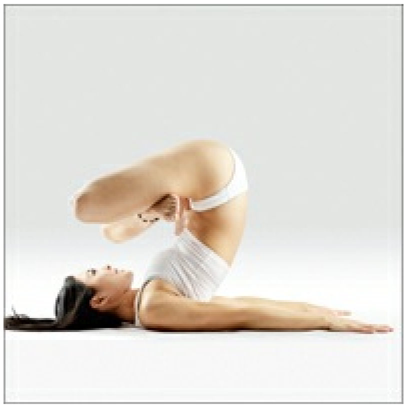

Unsupported Abdominal Lift Lotus Pose

Niralamba Jathara Padmasana

(nir-AH-luhm-buh JAH-tuh-tuh puhd-MAHS-uh-nuh)

Modification: arms extended in front, knees to the forearms

Pose Type: core, supine

Drishti Point: Nasagrai or Nasagre (nose)

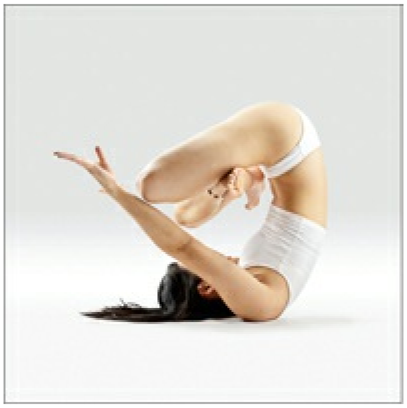

Quadruped Poses

- CHILD’S POSE
- RABBIT POSE
- COW POSE
- ALL FOURS (VARIATIONS)
- PLANK POSE
- PLANK ON FOREARMS
- FOUR LIMBED STAFF POSE
- SIDE PLANK

Child’s Pose

Balasana

(bah-LAHS-uh-nuh)

Also Known As: Child’s Pose (Garbhasana)
Modification: arms extended to the back, palms up; toes pointed to the back
Pose Type: forward bend
Drishti Point: Nasagrai or Nasagre (nose)

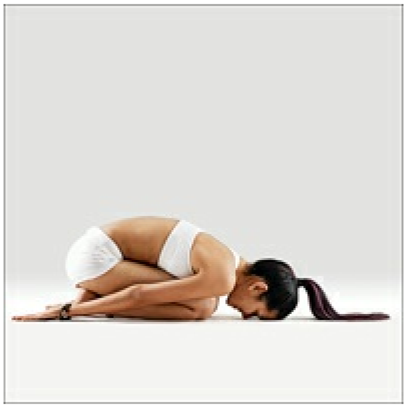

Child’s Pose

Balasana

(bah-LAHS-uh-nuh)

Also Known As: Child’s Pose (Garbhasana)
Modification: arms extended to the front, palms down; toes pointed to the back
Pose Type: forward bend
Drishti Point: Nasagrai or Nasagre (nose)

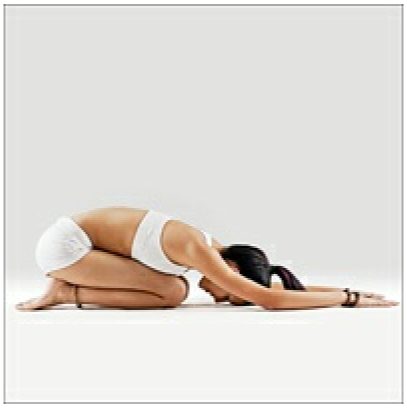

Child’s Pose

Balasana

(bah-LAHS-uh-nuh)

Modification: arms extended to the front, palms down; toes curled in

Pose Type: forward bend

Drishti Point: Nasagrai or Nasagre (nose)

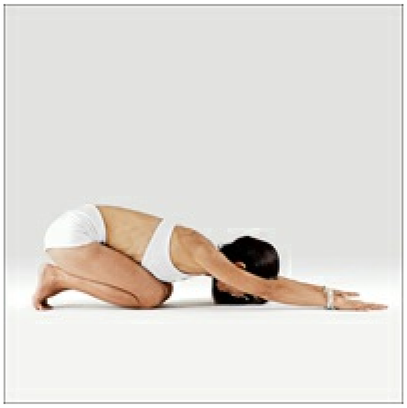

Hands to the Side Child’s Pose

Parshva Hasta Balasana

(PAHRSH-vuh HUH-stuh bah-LAHS-uh-nuh)

Pose Type: forward bend, side bend

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

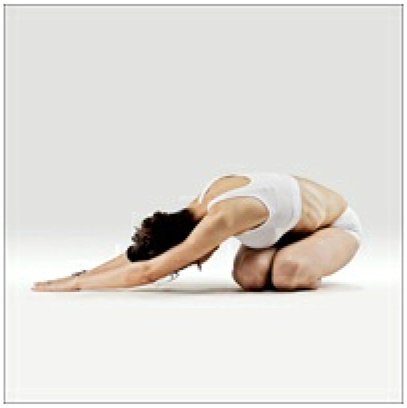

Hand Position of the Pose Dedicated to Garuda in Child’s Pose

Hasta Garudasana in Balasana

(HUH-stuh guh-ru-DAHS-uh-nuh in bah-LAHS-uh-nuh)

Pose Type: forward bend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

लं रं यं हं

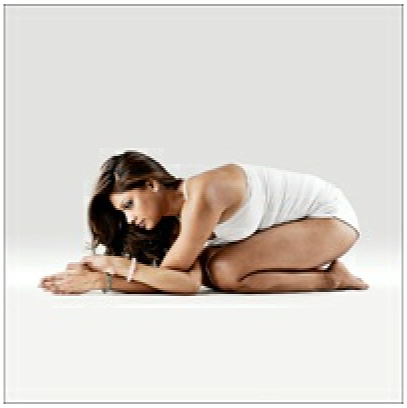

Child's Pose

Balasana

(bah-LAHS-uh-nuh)

Modification: knees open wide; palms pressed together, fingers spread wide, thumbs to the upper back

Pose Type: forward bend, mild backbend

Drishti Point: Nasagrai or Nasagre (nose)

लं यं ॐ

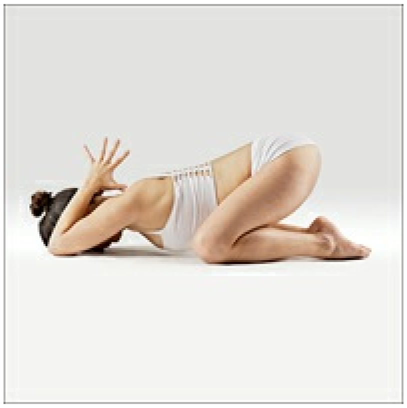

CHILD'S POSE: TWISTS

Revolved Side Child's Pose

Parivritta Parshva Balasana

(puh-ri-VRIT-tuh PAHRSH-vuh bah-LAHS-uh-nuh)

Pose Type: forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
● ●

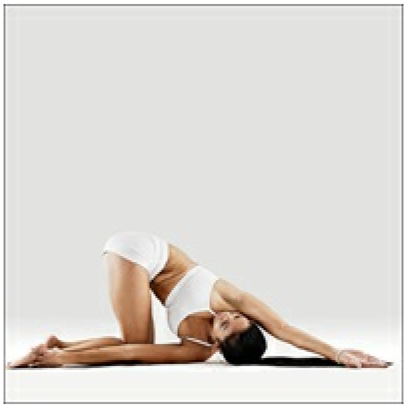

Both Hands to Legs Bound Revolved Child’s Pose

Dwi Hasta Pada Baddha Parivritta Balasana

(DWI-huh-stuh PUH-duh BUH-duh puh-ri-VRIT-tuh bah-LAHS-uh-nuh)
Pose Type: forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)
● ●

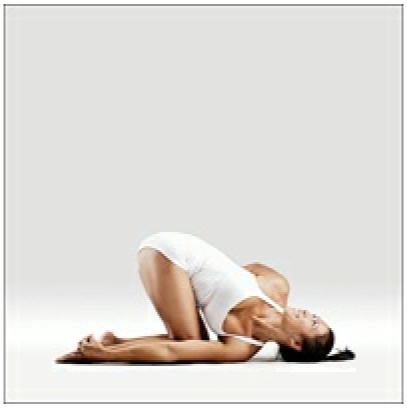

Leg Position of the Pose Dedicated to Garuda in Revolved Child’s Pose

Pada Garudasana in Parivritta Balasana

(PUH-duh guh-ru-DAHS-uh-nuh in puh-ri-VRIT-tuh bah-LAHS-uh-nuh)
Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
● ● ●

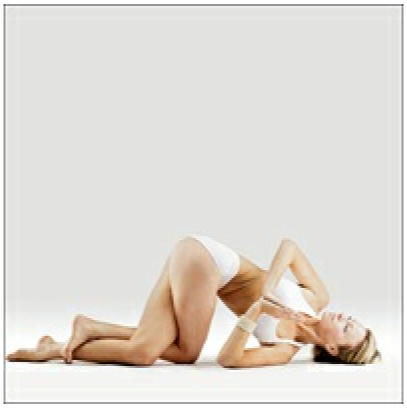

Revolved Child’s Pose

Parivritta Balasana

(puh-ri-VRIT-tuh bah-LAHS-uh-nuh)

Also Known As: Turned Child’s Pose (Parshva Balasana)
Modification: bottom arm straight out to the side; top arm bent at 90 degrees, palm to the floor overhead
Pose Type: forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
● ●

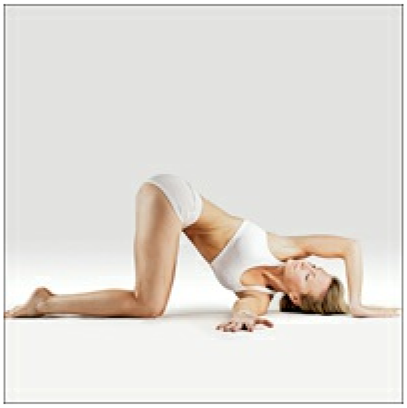

One Hand Bound Revolved Child’s Pose

Eka Hasta Baddha Parivritta Balasana

(EY-kuh HUH-stuh BUH-duh puh-ri-VRIT-tuh bah-LAHS-uh-nuh)

Modification: grabbing onto the quadricep
Pose Type: forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)
● ●

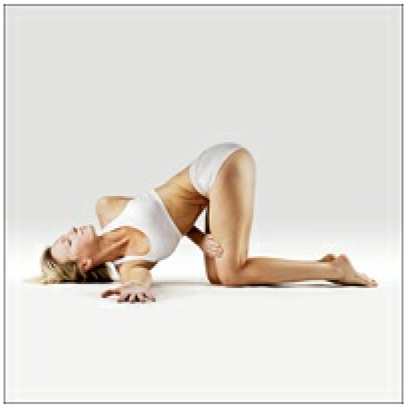

CHILD'S POSE: TWISTS & HALF LOTUS BINDING

Hand to Foot Revolved Side Child's Pose

Hasta Pada Parivritta Parshva Balasana

(HUH-stuh PUH-duh puh-ri-VRIT-tuh PAHRSH-vuh bah-LAHS-uh-nuh)

Modification: foot resting on the back of the knee
Pose Type: forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

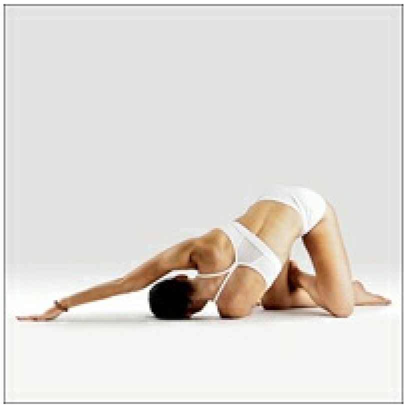

Both Hands to Feet Half Bound Lotus Revolved Child's Pose

Dwi Hasta Pada Ardha Baddha Padma Parivritta Balasana

(DWI-huh-stuh PUH-duh UHR-duh BUH-duh PUHD-muh puh-ri-VRIT-tuh bah-LAHS-uh-nuh)

Pose Type: forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

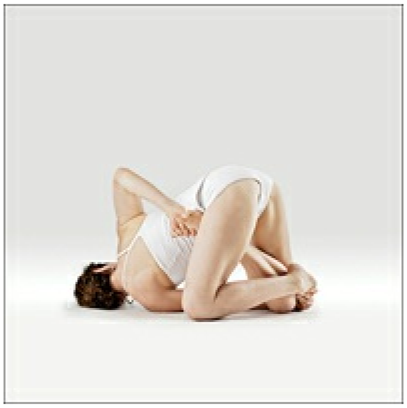

Half Bound Lotus Revolved Child’s Pose

Ardha Baddha Padma Parivritta Balasana

(UHR-duh BUH-duh PUHD-muh puh-ri-VRIT-tuh bah-LAHS-uh-nuh)

Modification: torso twisting to the outside of the body, hand binding to the foot
Pose Type: forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

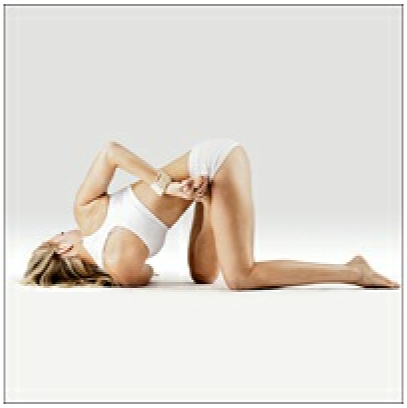

Half Bound Lotus Revolved Child’s Pose

Ardha Baddha Padma Parivritta Balasana

(UHR-duh BUH-duh PUHD-muh puh-ri-VRIT-tuh bah-LAHS-uh-nuh)

Modification: torso twisting to the inside of the body, hand binding to the shin
Pose Type: forward bend, twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

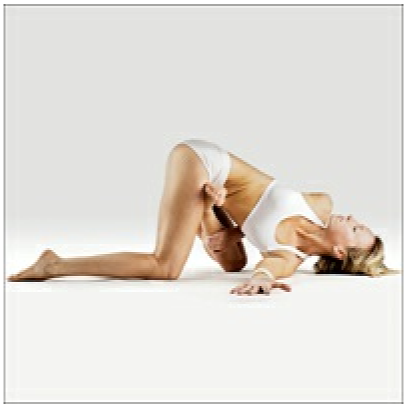

CHILD’S POSE: TWISTS—ONE-LEGGED—BACK FOOT IN THE AIR

Equilibrium One-Legged Revolved Child’s Pose with Hands in Prayer

Tulya Eka Pada Parivritta Balasana Namaskar

(TUL-yuh EY-kuh PUH-duh puh-ri-VRIT-tuh bah-LAHS-uh-nuh nuh-muhs-KAHR)

Modification: knee of the top leg bent
Pose Type: forward bend, twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

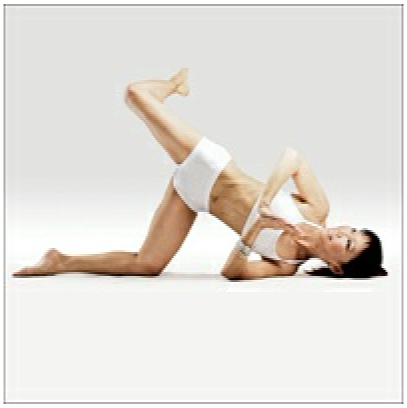

Half Bow Pose in Equilibrium One-Legged Revolved Child’s Pose

Ardha Dhanurasana in Tulya Eka Pada Parivritta Balasana

(UHR-duh duh-nur-AHS-uh-nuh in TUL-yuh EY-kuh PUH-duh puh-ri-VRIT-tuh bah-LAHS-uh-nuh)

Pose Type: forward bend, twist, backbend
Drishti Point: Hastagrai or Hastagre (hands)

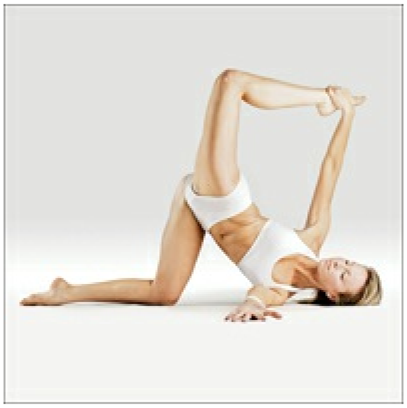

Revolved Half Bow Pose in Equilibrium One-Legged Revolved Child’s Pose

Parivritta Ardha Dhanurasana in Tulya Eka Pada Parivritta Balasana

(puh-ri-VRIT-tuh UHR-duh duh-nur-AHS-uh-nuh in TUL-yuh EY-kuh PUH-duh puh-ri-VRIT-tuh bah-LAHS-uh-nuh)

Pose Type: forward bend, twist, backbend

Drishti Point: Hastagrai or Hastagre (hands)

ल व र य

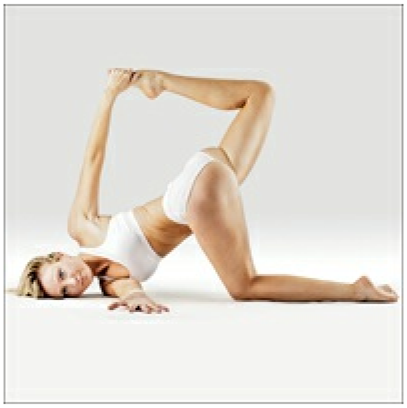

Half Bow Pose in Equilibrium One-Legged Revolved Child’s Pose

Ardha Dhanurasana in Tulya Eka Pada Parivritta Balasana

(UHR-duh duh-nur-AHS-uh-nuh in TUL-yuh EY-kuh PUH-duh puh-ri-VRIT-tuh bah-LAHS-uh-nuh)

Modification: back heel toward the sitting bone

Pose Type: forward bend, twist, backbend

Drishti Point: Hastagrai or Hastagre (hands)

ल व र य ॐ

Equilibrium One-Legged Revolved Child’s Pose

Tulya Eka Pada Parivritta Balasana

(TUL-yuh EY-kuh PUH-duh puh-ri-VRIT-tuh bah-LAHS-uh-nuh)

Modification: one leg straight, other knee bent, heel toward the sitting bone
Pose Type: forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

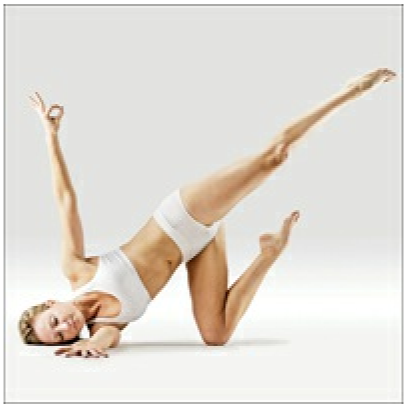

Hands Bound

Rabbit Pose

Baddha Hasta Sasangasana

(BUH-duh HUH-stuh shuh-shahng-AHS-uh-nuh)

Pose Type: forward bend

Drishti Point: Nasagrai or Nasagre (nose) or Nabhi, Nabhicakre, or Nabi Chakra (belly button)

How to Perform the Pose:

1. Begin by sitting on your heels in Hero Pose (*Virasana*). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale and bend forward, letting the crown of your head touch the floor, and lift your sitting bones off your heels. Grab onto your heels with both your hands, keeping your arms on the outsides of your shins.
3. Exhale as you rock forward, bringing your forehead toward your knees. Feel the stretch in the upper back between your shoulder blades.
4. Inhale as you lift your hands off your heels and interlock your fingers behind your back. Try to press your palms together. Exhale and stretch your arms up to the sky behind your back. Feel the stretch in your front shoulder heads and your chest (Pose #1). You can experiment by leaving your sitting bones on the heels and moving your forehead away from your knees (Pose #2).
5. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
6. Inhale as you release the arms. Exhale as you straighten your spine, coming back to Hero Pose (*Virasana*).

Modification:

1. forehead to the knees, sitting bones lifted
2. forehead away from the knees, sitting bones to the heels

baddha = bound

hasta = hand

sasanga = rabbit

1.

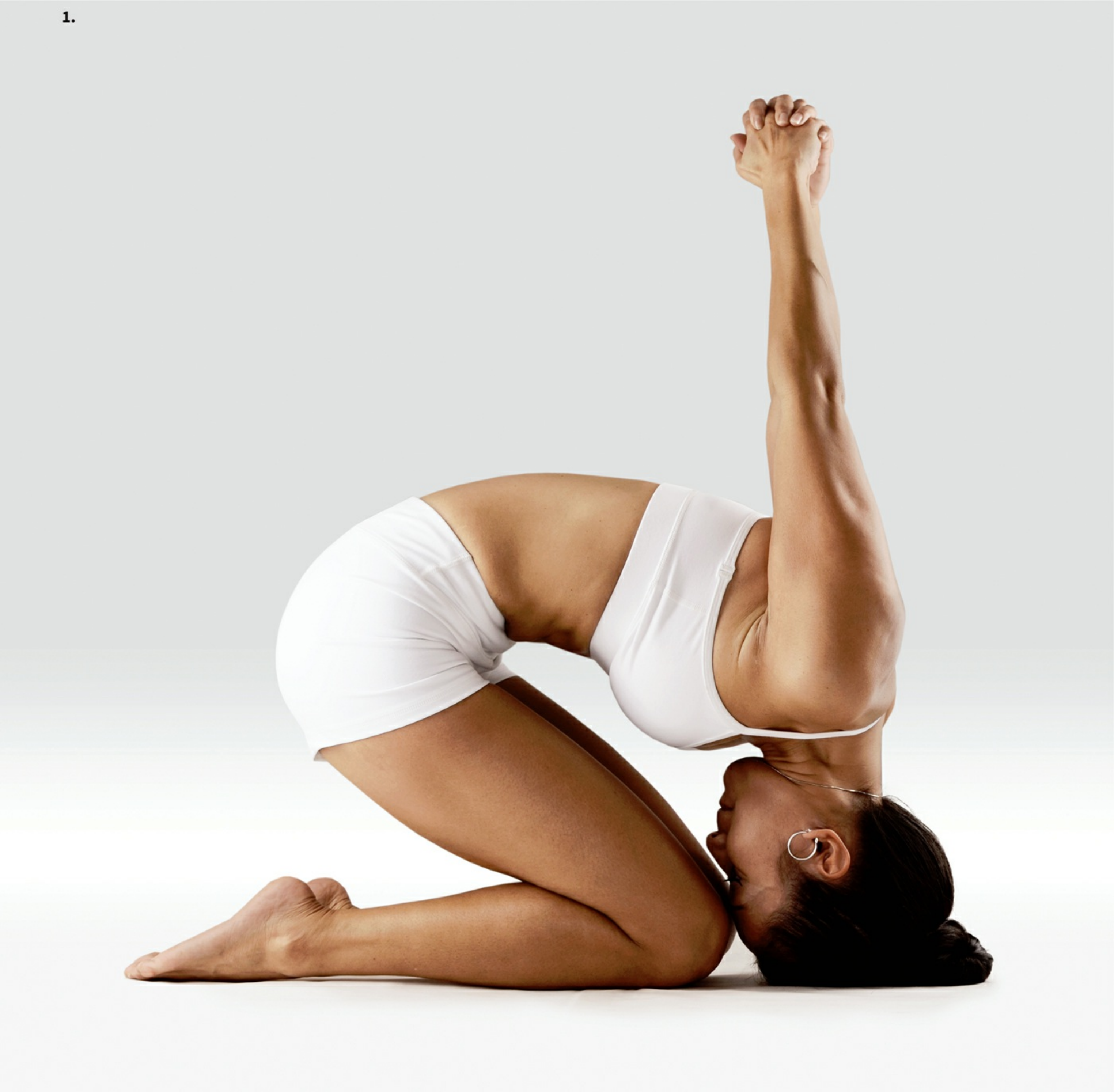

2.

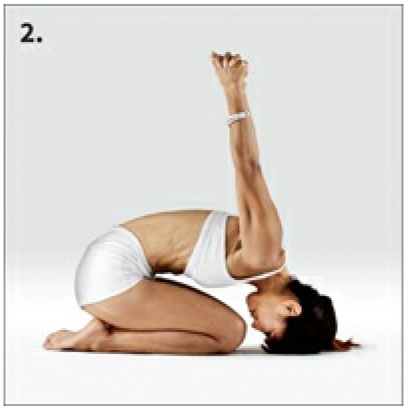

ल र य ह ॐ

Rabbit Pose

Sasangasana

(shuh-shahn-GAHS-uh-nuh)

Also Known As: Rabbit Foot Pose
Pose Type: forward bend, inversion
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)
● ● ● ● ●

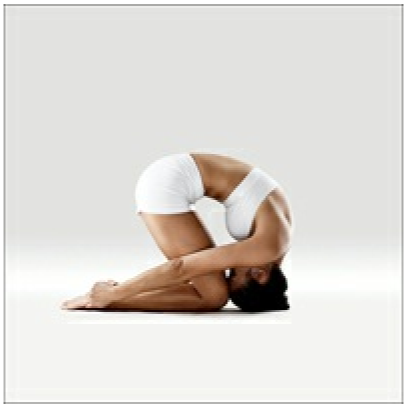

COW POSE ON THE HEAD: HALF LOTUS & GARUDA LEGS

Half Bound Lotus Cow Pose on the Head

Ardha Baddha Padma Shirsha Bitilasana

(UHR-duh BUH-duh PUHD-muh SHEER-shuh bee-til-AHS-uh-nuh)

Also Known As: Ardha Baddha Padma Shirsha Gavasana[*]
Modification: crown of the head to the floor
Pose Type: forward bend, inversion, binding
Drishti Point: Nasagrai or Nasagre (nose), Nabhi, Nabhicakre, or Nabi Chakra (belly button)
● ● ● ●

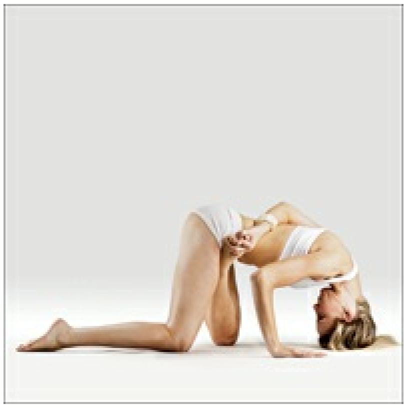

Reverse Prayer in Leg Position of the Pose Dedicated to Garuda in Cow Pose on the Head

Viparita Namaskar in Pada Garudasana in Shirsha Bitilasana

(vi-puh-REE-tuh nuh-muhs-KAHR in PUH-duh guh-ru-DAHS-uh-nuh in SHEER-shuh bee-til-AHS-uh-nuh)

Also Known As: Back of the Body Prayer Leg Position of the Pose Dedicated to Garuda in Cow Pose on the Head (Paschima Namaskara Pada Garudasana in Shirsha Bitilasana)
Pose Type: forward bend, inversion
Drishti Point: Nasagrai or Nasagre (nose), Nabhi, Nabhicakre, or Nabi Chakra (belly button)

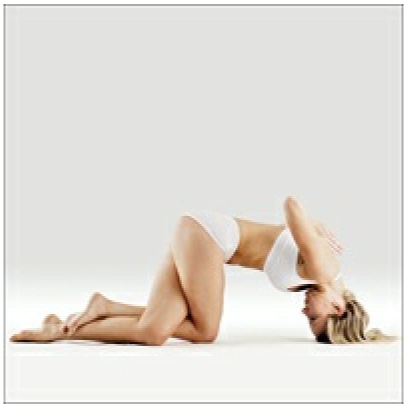

COW POSE ON THE HEAD: HEELS TO THE SITTING BONES & HEELS TO THE OUTSIDE OF THE HIPS

Both Hands to the Legs Cow Pose on the Head

Dwi Hasta Pada Shirsha Bitilasana

(DWI-huh-stuh PUH-duh SHEER-shuh bee-til-AHS-uh-nuh)

Modification: heels toward the sitting bones
Pose Type: forward bend, inversion
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

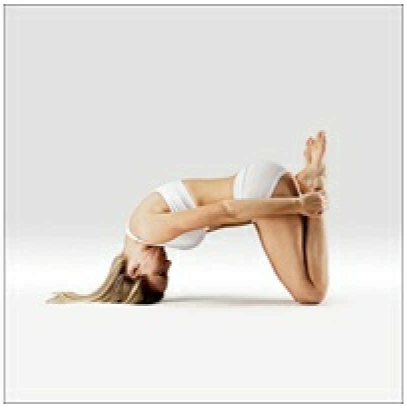

Both Hands to the Feet Cow Pose on the Head

Dwi Hasta Pada Shirsha Bitilasana

(DWI-huh-stuh PUH-duh SHEER-shuh bee-til-AHS-uh-nuh)

Modification: heels to the outside of the hips
Pose Type: forward bend, inversion
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

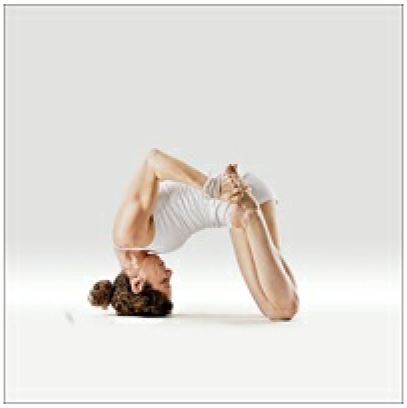

ALL FOURS: NEUTRAL SPINE

Cow Pose

Bitilasana

(bee-til-AHS-uh-nuh)

Also Known As: Table Pose (Bharmanasana)
Pose Type: standing (hands and knees)
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

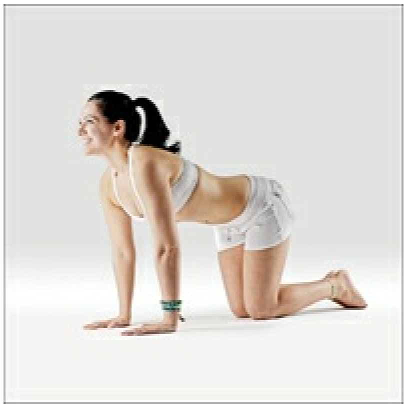

Intense Wrist Stretch Cow Pose

Uttana Manibandha Bitilasana

(ut-TAHN-uh muh-ni-BUHN-duh bee-til-AHS-uh-nuh)

Also Known As: Table Pose Modification (Bharmanasana)
Pose Type: standing (hands and knees)
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

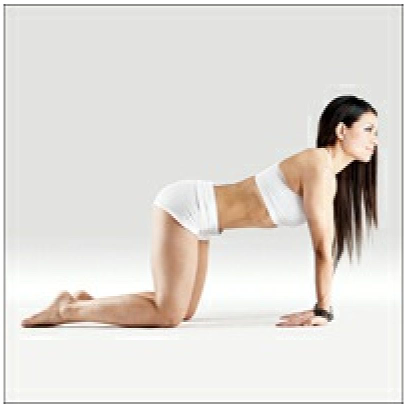

One Hand Cow Pose

Eka Hasta Bitilasana

(EY-kuh HUH-stuh bee-til-AHS-uh-nuh)

Also Known As: Table Pose (Bharmanasana)
Pose Type: standing (hands and knees)
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

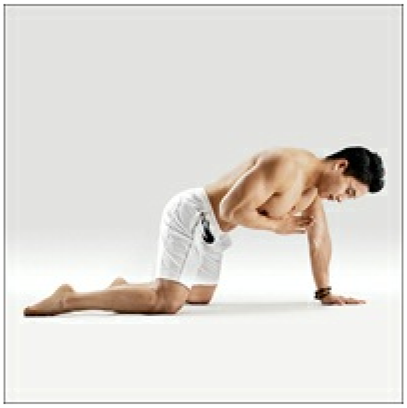

Leg Position of the Pose Dedicated to Garuda in Cow Pose

Pada Garudasana in Bitilasana

(PUH-duh guh-ru-DAHS-uh-nuh in bee-til-AHS-uh-nuh)

Also Known As: Table Pose Modification (Bharmanasana)
Pose Type: standing (on hands and knees)
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

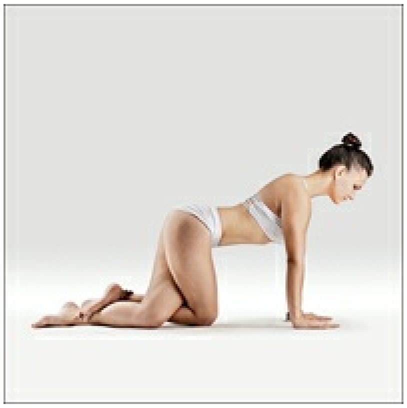

Lotus Pose in Cow Pose

Padmasana in Bitilasana

(puhd-MAHS-uh-nuh in bee-til-AHS-uh-nuh)

Also Known As: Table Pose Modification (Bharmanasana)
Pose Type: standing (on hands and knees)
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

ALL FOURS: NEUTRAL SPINE—ONE LEG EXTENDED TO THE BACK

Tiger Pose

Vyaghrasana

(vyah-GRAHS-uh-nuh)

Also Known As: Table Pose Modification (Bharmanasana)
Modification: neutral spine
Pose Type: standing (hands and knees)
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

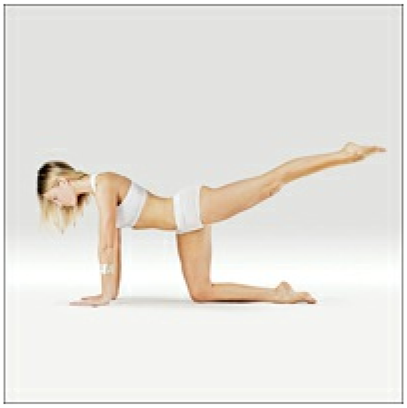

One Hand One-Legged Cow Pose

Eka Hasta Eka Pada Bitilasana

(EY-kuh HUH-stuh EY-kuh PUH-duh bee-til-AHS-uh-nuh)

Also Known As: Cat Pose (Marjarasana), Balancing Table Pose Modification (Dandayamana Bharmanasana)
Modification: neutral spine

1. arm reaching to the front, opposite leg reaching to the back
2. arm reaching to the back, opposite leg reaching to the back

3. arm reaching to the back, leg reaching to the back on the same side

Pose Type: standing (hands and knees)

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

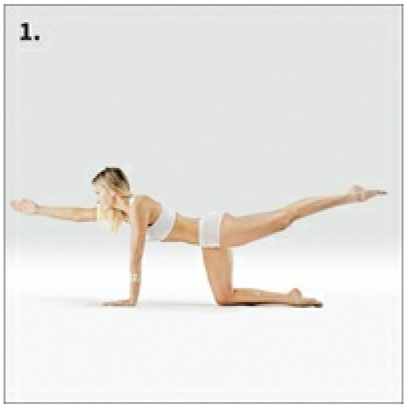

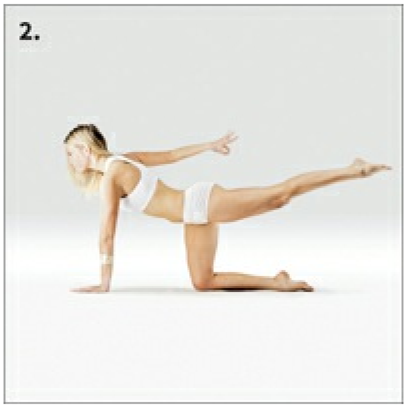

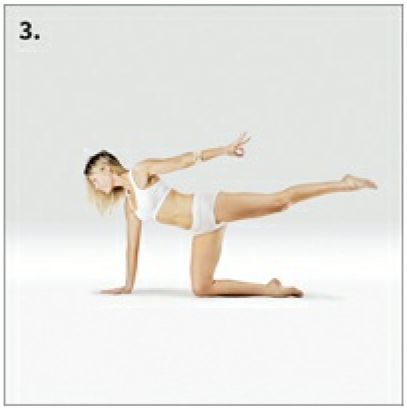

One Knee Cow Pose

Eka Janu Bitilasana

(EY-kuh JAH-nu bee-til-AHS-uh-nuh)

Also Known As: Unsupported Balancing Table Pose Modification (Niralamba Dandayamana Bharmanasana)

Pose Type: standing (on the knee), balance

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

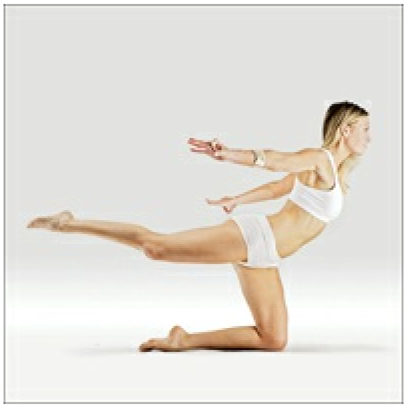

ALL FOURS: NEUTRAL SPINE—ONE LEG EXTENDED TO THE BACK—HALF LOTUS

Downward Facing Pose Dedicated to Kashyapa

Adho Mukha Kashyapasana

(uh-DO MUK-uh kah-shyuh-PAHS-uh-nuh)

Modification: Leg 1: knee on the floor Leg 2: foot on the floor
Pose Type: arm balance, binding
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

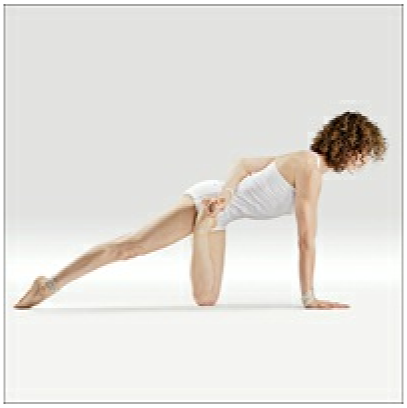

Downward Facing Pose Dedicated to Kasyapa

Adho Mukha Kasyapasana

(uh-DO MUK-uh kah-shyuh-PAHS-uh-nuh)

Modification: Leg 1: knee on the floor Leg 2: foot off the floor
Pose Type: arm balance, binding
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

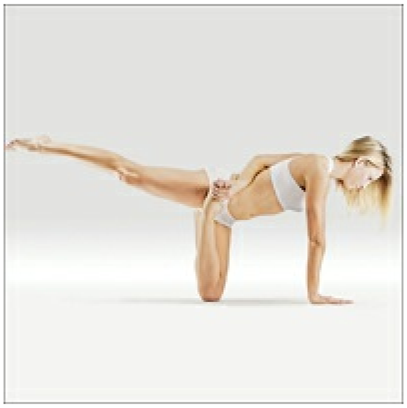

ALL FOURS: CAT TILT

Downward Facing Cat Pose

Adho Mukha Marjarasana

(uh-DO MUK-uh mahr-jah-RAHS-uh-nuh)

Modification: toes pointing away from the head
Pose Type: standing (on hands and knees), forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

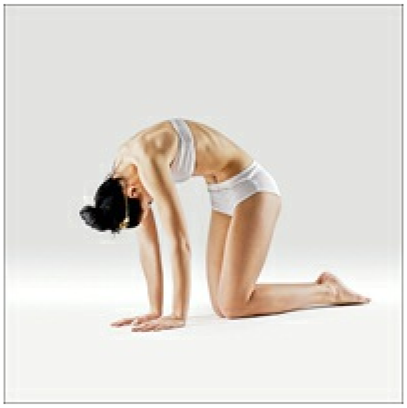

Downward Facing Cat Pose

Adho Mukha Marjarasana

(uh-DO MUK-uh mahr-jah-RAHS-uh-nuh)

Modification: arms straight to the front, fingertips to the floor; rocking back, toes curled in
Pose Type: standing (on hands and knees), forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

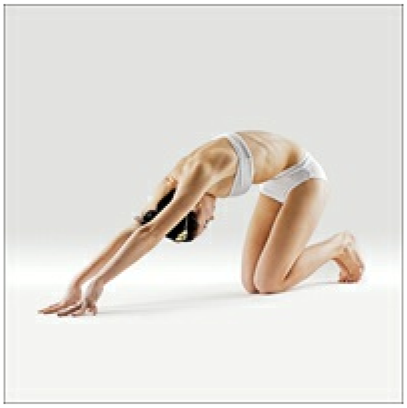

Tiger Pose

Vyaghrasana

(vyah-GRAHS-uh-nuh)

Also Known As: Cow Pose Modification (Bitilasana)
Modification: knee to the forehead
Pose Type: standing (on hands and knees), forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

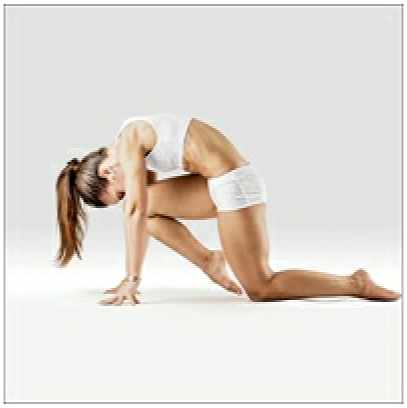

Tiger Pose

Vyaghrasana

(vyah-GRAHS-uh-nuh)

Also Known As: Cow Pose Modification (Bitilasana)
Modification: knee to the shoulder on the same side
Pose Type: standing (on hands and knees), forward bend
Drishti Point: Nasagrai or Nasagre (nose)

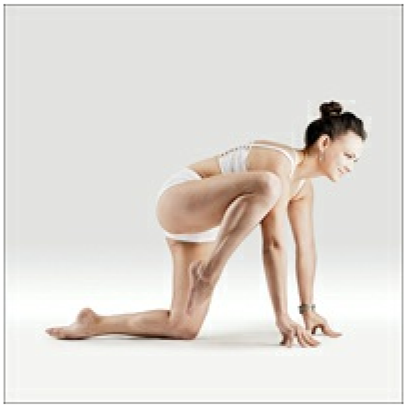

Tiger Pose

Vyaghrasana

(vyah-GRAHS-uh-nuh)

Also Known As: Cow Pose Modification (Bitilasana)
Modification: knee to the opposite tricep
Pose Type: standing (on hands and knees), forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose)

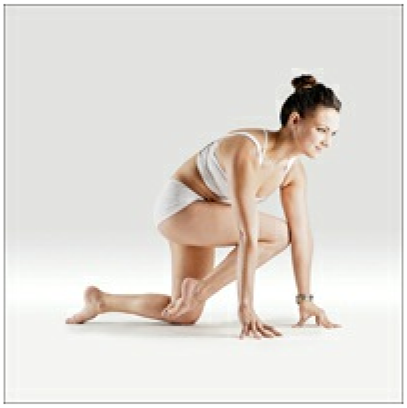

Tiger Pose

Vyaghrasana

(vyah-GRAHS-uh-nuh)

Also Known As: Cow Pose Modification (Bitilasana)
Modification: shoulder to the back of the knee; back leg, toes curled in
Pose Type: standing (on hands and knees), forward bend
Drishti Point: Nasagrai or Nasagre (nose), Padayoragrai or Padayoragre (toes/feet)

Tiger Pose

Vyaghrasana

(vyah-GRAHS-uh-nuh)

Also Known As: Cow Pose Modification (Bitilasana)
Modification: one hand to the heart
1. back of the knee to the shoulder, leg bent
2. back of the knee to the shoulder, leg straight
Pose Type: standing (on hands and knees), forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

1.

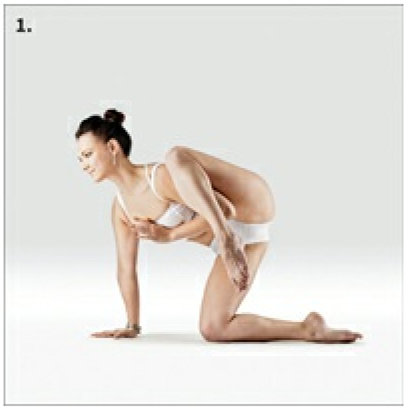

2.

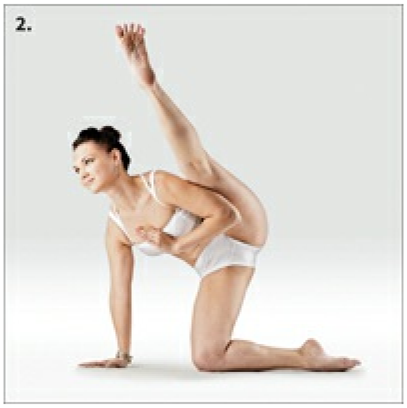

Upward Facing Dog Pose Tilt

Urdhva Mukha Shvanasana Tilt

(OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Pose Type: standing (on hands and knees), backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows

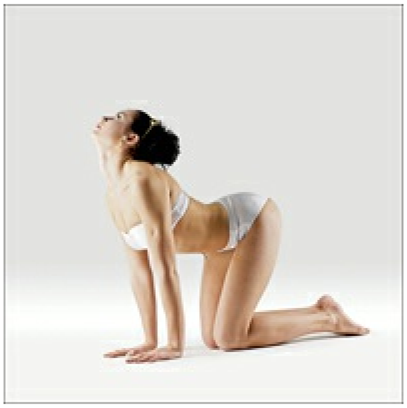

Tiger Pose

Vyaghrasana

(vyah-GRAHS-uh-nuh)

Also Known As: Balancing Table Pose Modification (Niralamba Dandayamana Bharmanasana)

Modification: one leg straight out, toes pointing away from the head

Pose Type: standing (on hands and knees), backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

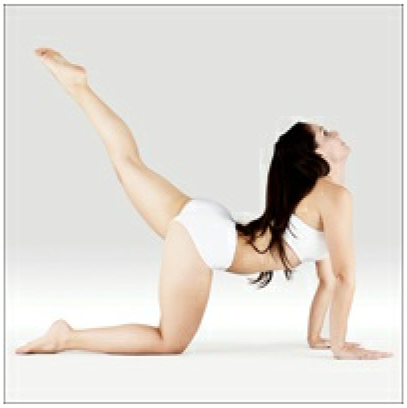

Unsupported Tiger Pose

Niralamba Vyaghrasana

(nir-AH-luhm-buh vyah-GRAHS-uh-nuh)

Also Known As: One-Legged Bow Pose from All Fours (Eka Pada Dhanurasana), Balancing Table Pose Modification (Dandayamana Bharmanasana)
Modification: grabbing onto the foot with under-head grip on the same side
Pose Type: standing (on hands and knees), backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व र य

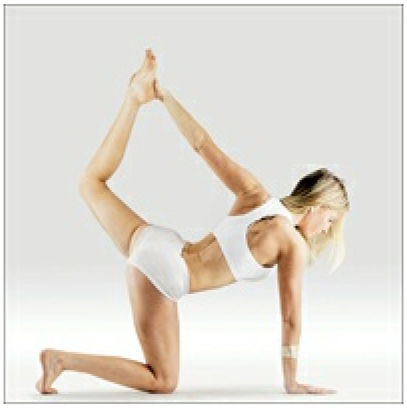

Unsupported Tiger Pose

Niralamba Vyaghrasana

(nir-AH-luhm-buh vyah-GRAHS-uh-nuh)

Also Known As: One-Legged Bow Pose from All Fours (Eka Pada Dhanurasana), Balancing Table Pose Modification (Dandayamana Bharmanasana)
Modification: grabbing onto the shin with under-head grip on the opposite side
Pose Type: standing (on hands and knees), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

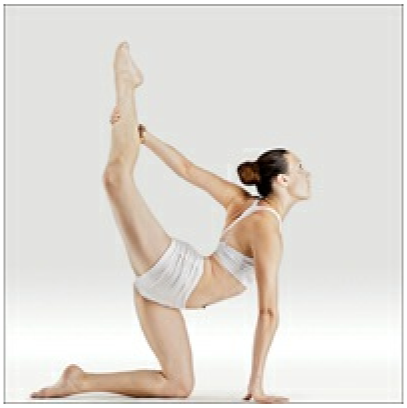

Unsupported Tiger Pose

Niralamba Vyaghrasana

(nir-AH-luhm-buh vyah-GRAHS-uh-nuh)

Also Known As: Balancing Table Pose Modification (Dandayamana Bharmanasana)
Modification: grabbing onto the foot with overhead grip on the same side, foot toward the head
1. toes pointing away from the head
2. toes curled in
Pose Type: standing (on hands and knees), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

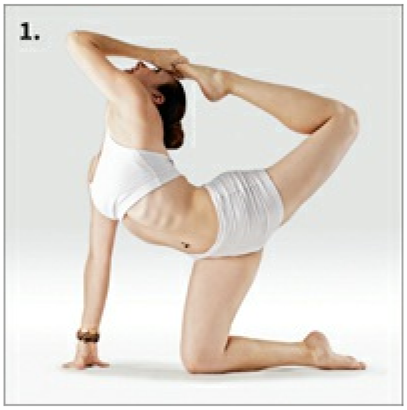

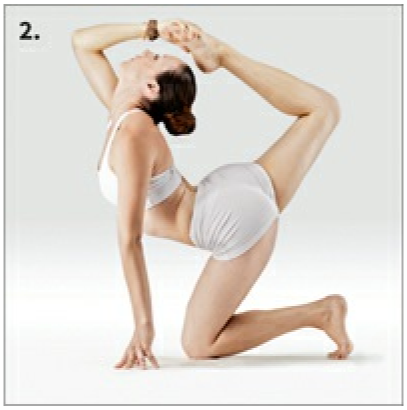

Unsupported Tiger Pose

Niralamba Vyaghrasana

(nir-AH-luhm-buh vyah-GRAHS-uh-nuh)

Also Known As: Balancing Table Pose Modification (Dandayamana Bharmanasana)
Modification: grabbing onto the foot with overhead grip on the opposite side, foot away from the head, toes curled in
Pose Type: standing (on hands and knees), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ALL FOURS: FOREARMS

Cow Pose

Bitilasana

(bee-til-AHS-uh-nuh)

Also Known As: Table Pose Modification (Bharmanasana)
Modification: forearms to the floor
Pose Type: standing (on forearms and knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

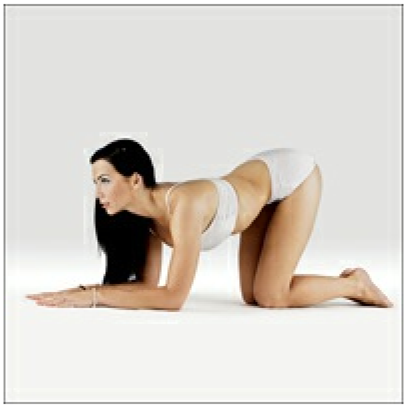

Frog Pose in Cow Pose

Mandukasana in Bitilasana

(muhn-doo-KAHS-uh-nuh in bee-til-AHS-uh-nuh)

Also Known As: Frog Pose in Table Pose Modification (Mandukasana in Bharmanasana)
Modification: forearms to the floor, knees open wider than hips
Pose Type: standing (on forearms and knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

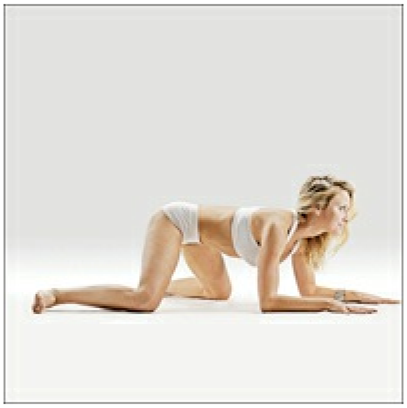

Uneven Wounded Duck Pose Prep.

Vishama Pungu Karandavasana Prep.
(VISH-uh-muh pung-u kahr-uhn-duh-VAHS-uh-nuh)
Modification: one forearm to the floor, toes of both feet on the floor
Pose Type: standing (on forearms and feet), forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

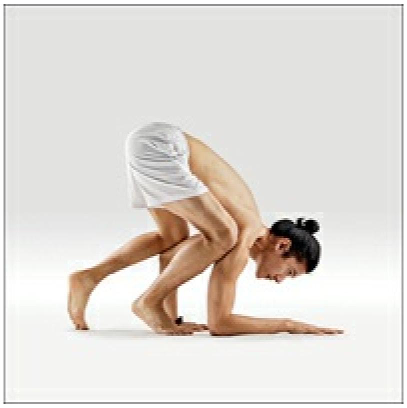

Supported Tiger Pose

Salamba Vyaghrasana *(SAH-luhm-buh vyah-GRAHS-uh-nuh)*
Also Known As: Cow Pose Modification (Bitilasana)
Modification: forearm on the floor, one hand to the heart; back of the knee to the shoulder, leg straight
Pose Type: standing (on forearms and knees), forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

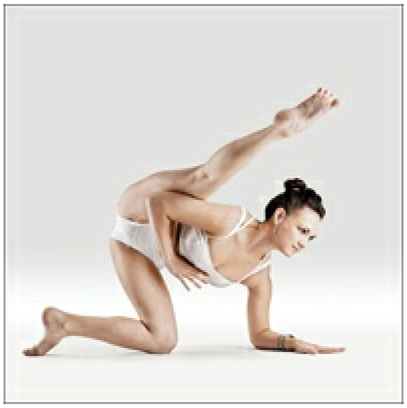

Supported Tiger Pose

Salamba Vyaghrasana *(SAH-luhm-buh-nuh vyah-GRAHS-uh-nuh)*
Also Known As: Cow Pose Modification (Bitilasana)
Modification: forearm on the floor, grabbing onto the foot with under-head grip on the opposite side
Pose Type: standing (on forearms and knees), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)
लं वं रं यं हं ॐ

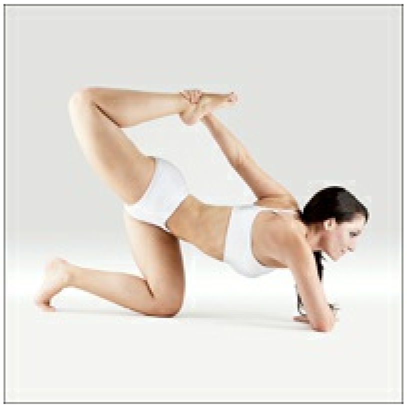

ALL FOURS: FOREARMS—TWISTS

Hand to Leg Sideways Leg Position of Cow Face Pose in Cow Pose

Hasta Pada Parshva Pada Gomukhasana in Bitilasana
(HUH-stuh PUH-duh PAHRSH-vuh PUH-duh go-mu-KAHS-uh-nuh in bee-til-AHS-uh-nuh)
Modification: forearm to the floor
Pose Type: standing (on forearms and knees), twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
लं वं

Half Bound Sideways Leg Position of Cow Face Pose in Cow Pose

Ardha Baddha Parshva Pada Gomukhasana in Bitilasana

(UHR-duh BUH-duh PAHRSH-vuh PUH-duh go-mu-KAHS-uh-nuh in bee-til-AHS-uh-nuh)

Modification: forearm to the floor
Pose Type: standing (on forearms and knees), twist, binding
Drishti Point: Urdhva or Antara Drishti (up to the sky)

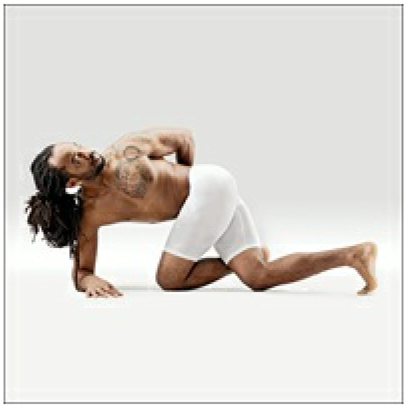

ALL FOURS: FOREARMS—HALF LOTUS

Half Lotus Cow Pose

Ardha Padma Bitilasana

(UHR-duh PUHD-muh bee-til-AHS-uh-nuh)

Also Known As: Half Lotus Table Pose Modification (Ardha Padma Bharmanasana)
Modification: forearms to the floor
Pose Type: standing (on forearms and knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

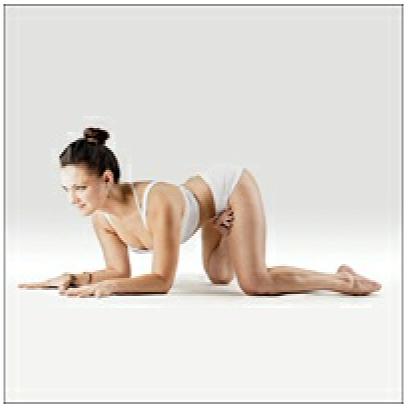

Revolved Half Lotus Cow Pose

Parivritta Ardha Padma Bitilasana

(puh-ri-VRIT-uh UHR-duh PUHD-muh bee-til-AHS-uh-nuh)

Also Known As: Revolved Half Lotus Table Pose Modification (Parivritta Ardha Padma Bharmanasana)
Modification: forearm on the floor, other hand behind the back
Pose Type: standing (on forearms and knees), twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)

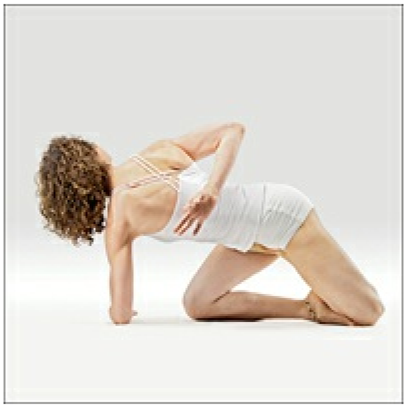

Half Bound Lotus Cow Pose

Ardha Baddha Padma Bitilasana

(UHR-duh PUHD-muh bee-til-AHS-uh-nuh)

Also Known As: Half Bound Lotus Table Pose Modification (Ardha Baddha Padma Bharmanasana)
Modification: forearm to the floor
Pose Type: standing (on forearms and knees), binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

Revolved Half Lotus Cow Pose

Parivritta Ardha Padma Bitilasana

(puh-ri-VRIT-uh UHR-duh PUHD-muh bee-til-AHS-uh-nuh)

Also Known As: Revolved Half Lotus Table Pose Modification (Parivritta Ardha Padma Bharmanasana)
Modification: forearm on the floor, top arm up to the sky
1. back view
2. front view
Pose Type: standing (on forearms and knees), twist
Drishti Point: Hastagrai or Hastagre (hands)

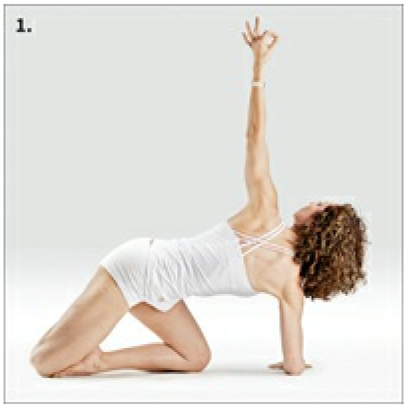

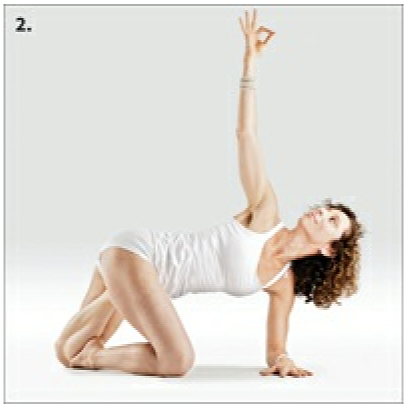

PLANK POSE: ON THE FINGERTIPS & ARMS CROSSED

Extended Four Limbs Staff Pose

Utthita Chaturanga Dandasana

(UT-ti-tuh chuh-tur-UHNG-guh-uh duhn-DAHS-uh-nuh)

Also Known As: High Plank
Modification: on the fingertips
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

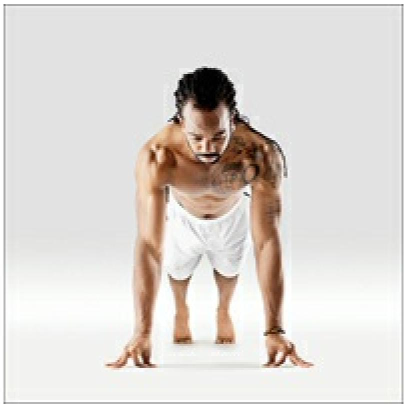

Extended Four Limbs Staff Pose

Utthita Chaturanga Dandasana

(UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: High Plank
Modification: arms crossed
1. knees on the floor
2. knees off the floor
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

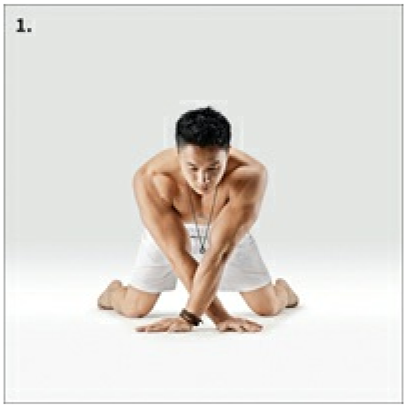

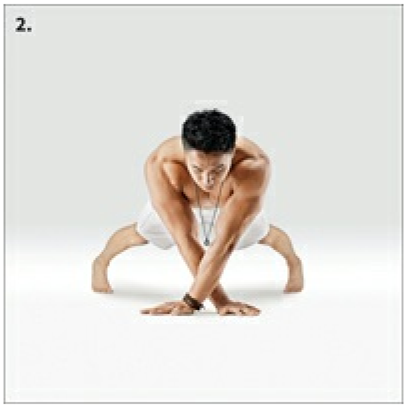

PLANK POSE: LEG MODIFICATIONS

Extended Four Limbs Staff Pose

Utthita Chaturanga Dandasana

(UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: High Plank Modification
Modification: 1. knees on the floor
2. knees off the floor
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

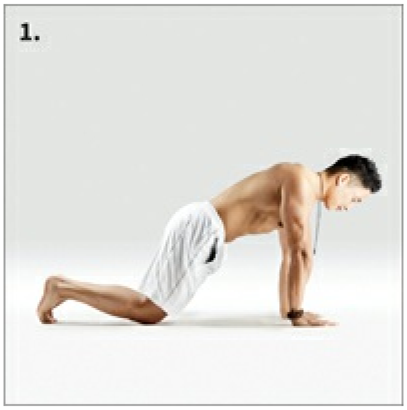

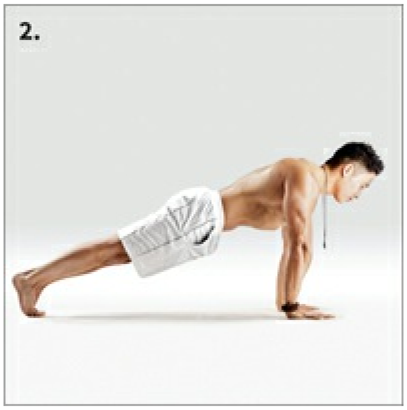

One-Legged Extended Four Limbs Staff Pose

Eka Pada Utthita Chaturanga Dandasana

(EY-kuh PUH-duh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: One-Legged High Plank Modification
Modification: knee to the nose
Pose Type: arm balance, forward bend, core
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

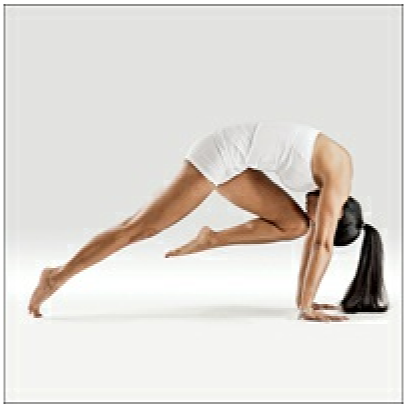

One-Legged Extended Four Limbs Staff Pose

Eka Pada Utthita Chaturanga Dandasana

(EY-kuh PUH-duh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: One-Legged High Plank Modification
Modification: knee to the shoulder on the same side
Pose Type: arm balance, forward bend, core
Drishti Point: Nasagrai or Nasagre (nose)

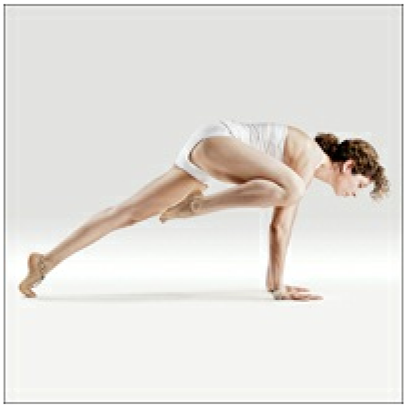

One-Legged Extended Four Limbs Staff Pose

Eka Pada Utthita Chaturanga Dandasana

(EY-kuh PUH-duh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: One-Legged High Plank Modification
Modification: foot wrapping around the calf muscle
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

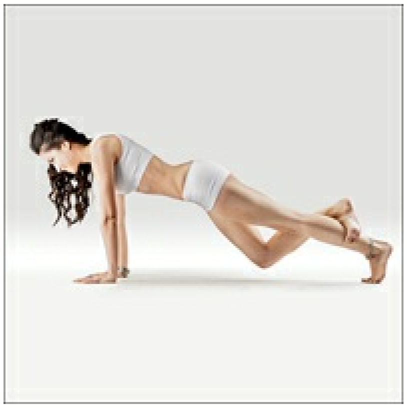

Revolved Extended Four Limbs Staff Pose

Parivritta Utthita Chaturanga Dandasana

(puh-ri-VRIT-tuh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: Revolved High Plank Modification
Pose Type: arm balance, twist, core
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

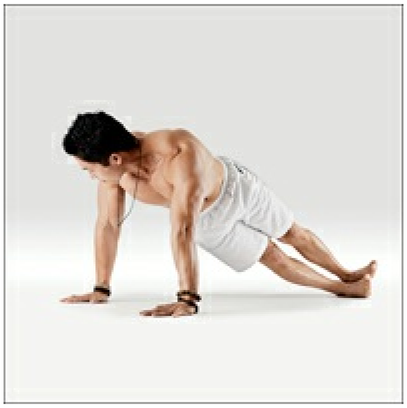

PLANK POSE: ARM AND LEG MODIFICATIONS

One-Legged Extended Four Limbs Staff Pose

Eka Pada Utthita Chaturanga Dandasana

(EY-kuh PUH-duh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: One-Legged High Plank Modification
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

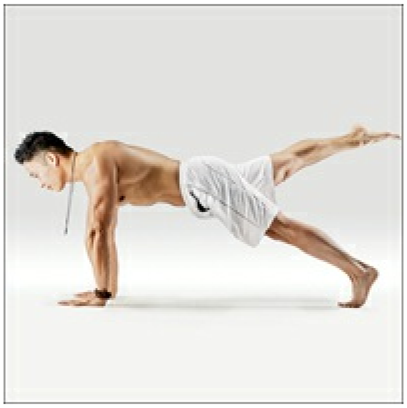

One Hand Extended Four Limbs Staff Pose

Eka Hasta Utthita Chaturanga Dandasana

(EY-kuh HUH-stuh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: One Hand High Plank Modification
Pose Type: arm balance, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

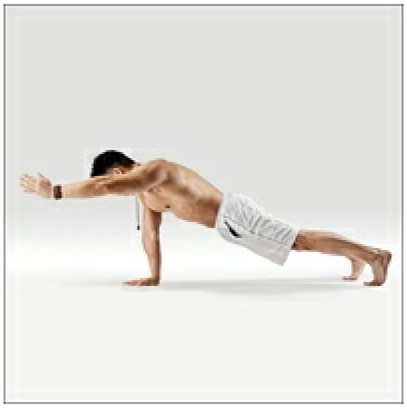

One Hand Extended Four Limbs Staff Pose

Eka Hasta Utthita Chaturanga Dandasana

(EY-kuh HUH-stuh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: One Hand High Plank Modification
Modification: hand to the heart
Pose Type: arm balance, core
Drishti Point: Angushtamadhye or Angushtha Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

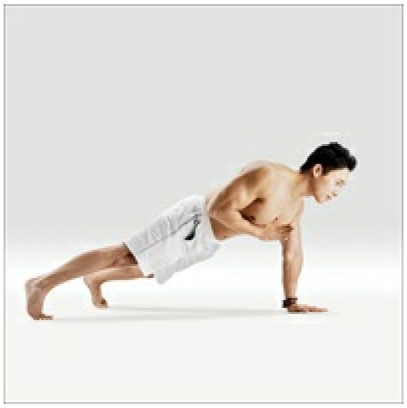

One Leg One Hand Extended Four Limbs Staff Pose

Eka Pada Eka Hasta Utthita Chaturanga Dandasana

(EY-kuh PUH-duh EY-kuh HUH-stuh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: One Hand High Plank Modification
Pose Type: arm balance, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

Tree Pose in One-Legged Extended Four Limbs Staff Pose

Vrikshasana in Eka Pada Utthita Chaturanga Dandasana

(vrik-SHAHS-uh-nuh in EY-kuh PUH-duh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: Tree Pose in One-Legged High Plank Modification
Modification: 1. knees on the floor
2. knees off the floor
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

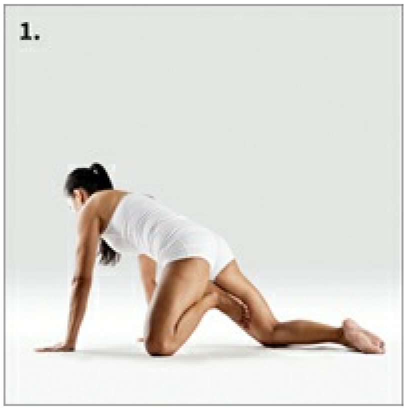

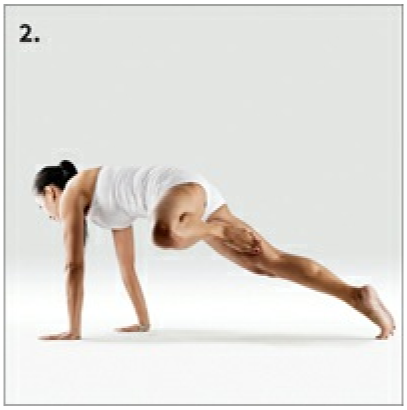

Extended Four Limbs Staff Pose

Utthita Chaturanga Dandasana

(UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: High Plank Modification
Modification: arms extended to the front, tops of the feet on the floor
1. knees on the floor
2. knees off the floor
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

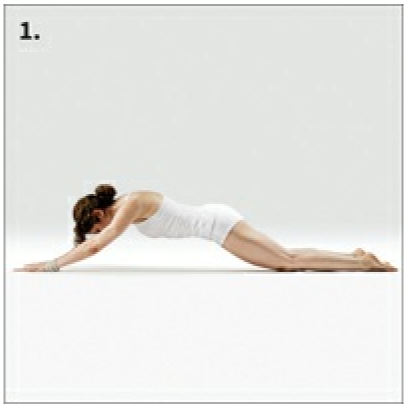

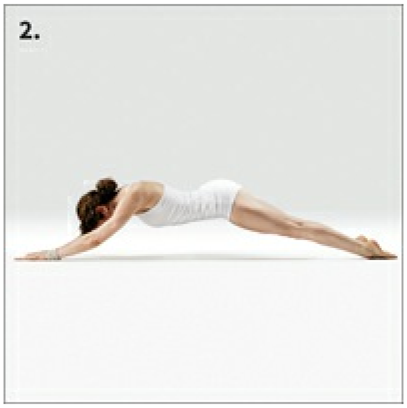

PLANK POSE: TWISTS

Leg to the Side Revolved One Hand Extended Four Limbs Staff Pose

Parshva Pada Parivritta Eka Hasta Utthita Chaturanga Dandasana

(PAHRSH-vuh PUH-duh puh-ri-VRIT-tuh EY-kuh HUH-stuh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: Leg to the Side Revolved One Hand High Plank Modification
Modification: 1. back knee on the floor
2. back knee off the floor, knee bent
3. back knee off the floor, both legs straight
4. back knee off the floor, knee bent, toes of the straight leg pointed
Pose Type: arm balance, twist, core

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

ल व य

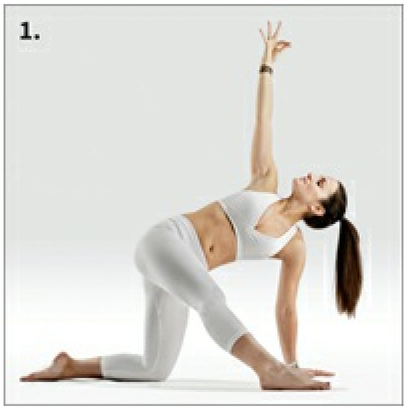

1.

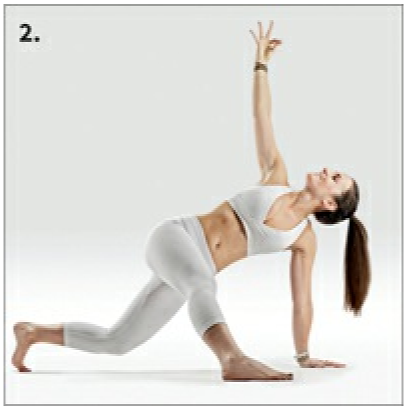

2.

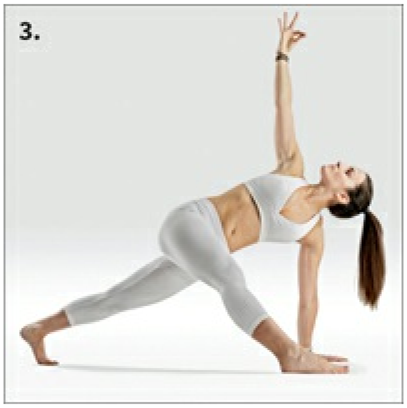

3.

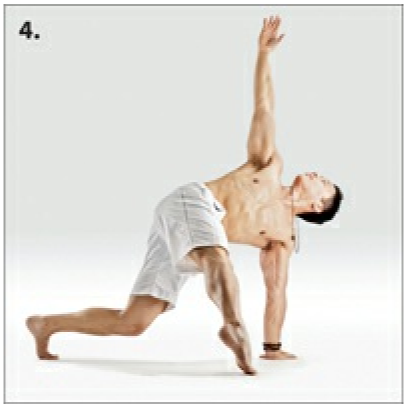

4.

Revolved One Hand Extended Four Limbs Staff Pose

Parivritta Eka Hasta Utthita Chaturanga Dandasana

(puh-ri-VRIT-tuh EY-kuh HUH-stuh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: Revolved One Hand High Plank Modification

Modification: 1. knees on the floor
2. knees off the floor

Pose Type: arm balance, twist, core

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

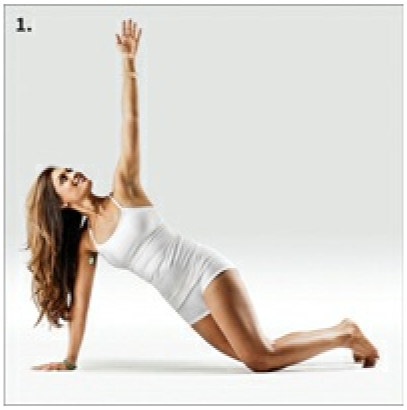

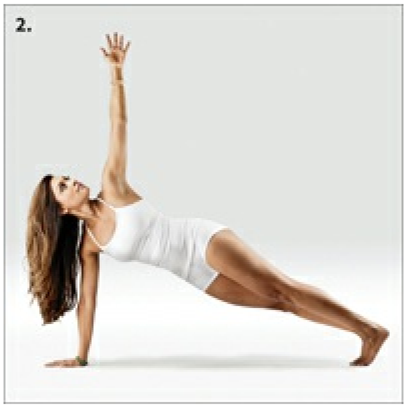

Revolved Leg to the Side Extended Four Limbs Staff Pose

Parivritta Parshva Pada Utthita Chaturanga Dandasana

(puh-ri-VRIT-tuh PAHRSH-vuh PUH-duh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: Revolved Leg to the Side High Plank Modification

Modification: leg sweeping under the torso

Pose Type: arm balance, twist, core

Drishti Point: Nasagrai or Nasagre (nose)

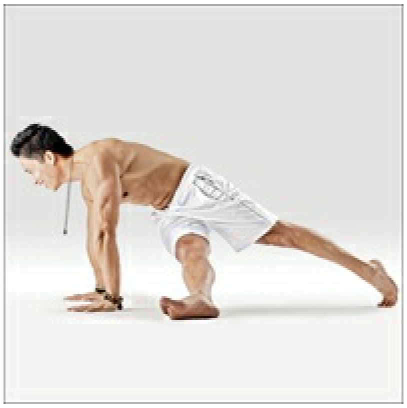

PLANK POSE: ONE LEG TO THE SIDE

Leg to the Side Extended Four Limbs Staff Pose

Parshva Pada Utthita Chaturanga Dandasana

(PAHRSH-vuh PUH-duh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: Leg to the Side High Plank Modification

Modification: 1. back knee on the floor

2. back knee off the floor

Pose Type: arm balance, core

Drishti Point: Nasagrai or Nasagre (nose)

1.

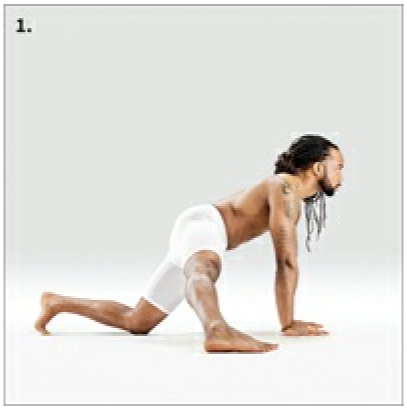

2.

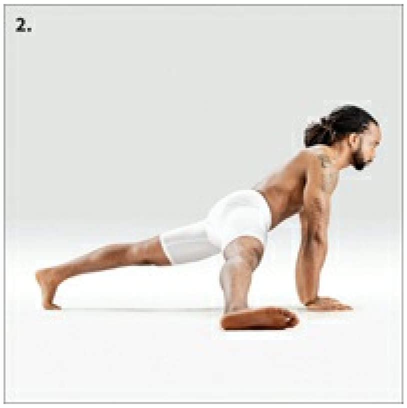

Upward One Leg to the Side Extended Four Limbs Staff Pose

Urdhva Parshva Eka Pada Utthita Chaturanga Dandasana

(OORD-vuh PAHRSH-vuh EY-kuh PUH-duh UT-ti-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: Leg to the Side One-Legged High Plank Modification
Modification: back knee off the floor, other leg off the floor
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

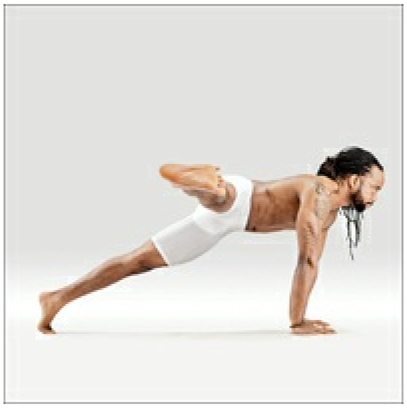

PLANK ON FOREARMS

Staff Pose Dedicated to Makara

Makara Dandasana

(MUH-kuh-ruh duhn-DAHS-uh-nuh)

Also Known As: Dolphin Plank Modification
Pose Type: forearm balance, core
Drishti Point: Nasagrai or Nasagre (nose), Angushtamadhye or Angushta Ma Dyai (thumbs)

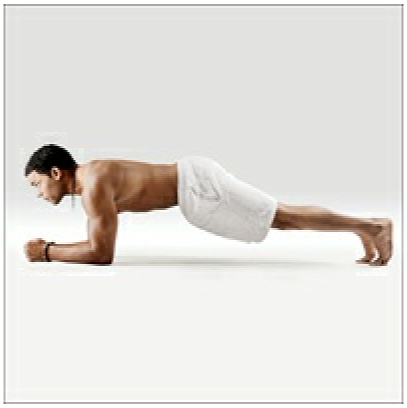

One Hand Staff Pose Dedicated to Makara

Eka Hasta Makara Dandasana

(EY-kuh HUH-stuh MUH-kuh-ruh duhn-DAHS-uh-nuh)

Also Known As: One Hand Dolphin Plank Modification
Pose Type: forearm balance, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

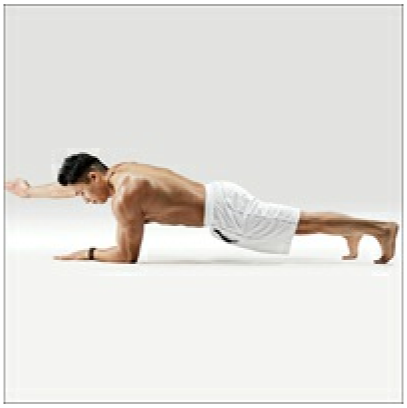

One-Legged Staff Pose Dedicated to Makara

Eka Pada Makara Dandasana

(EY-kuh PUH-duh MUH-kuh-ruh duhn-DAHS-uh-nuh)

Also Known As: One-Legged Dolphin Plank Modification
Pose Type: forearm balance, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

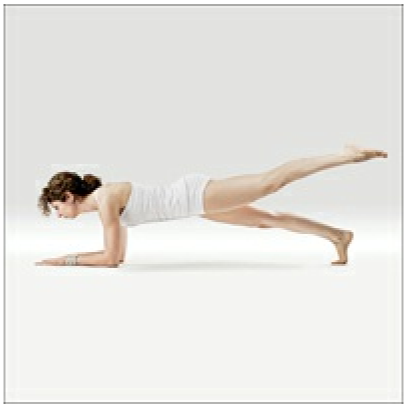

PLANK ON ELBOW

Uneven Repose Staff Pose

Vishama Sayana Dandasana

(VISH-uh-muh SHUH-yuh-nuh duhn-DAHS-uh-nuh)

Also Known As: Uneven Elbow Plank Pose Modification
Modification: one hand to the face, one forearm to the floor
Pose Type: forearm/elbow balance, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

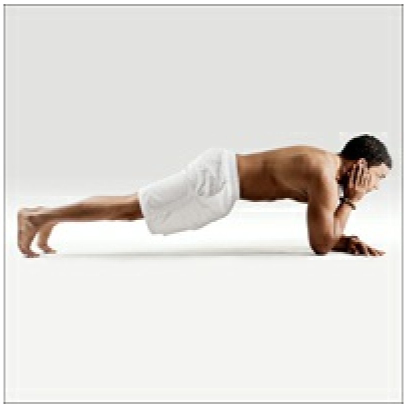

Repose Staff Pose

Sayana Dandasana

(SHUH-yuh-nuh duhn-DAHS-uh-nuh)

Also Known As: Elbow Plank
Pose Type: elbow balance, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

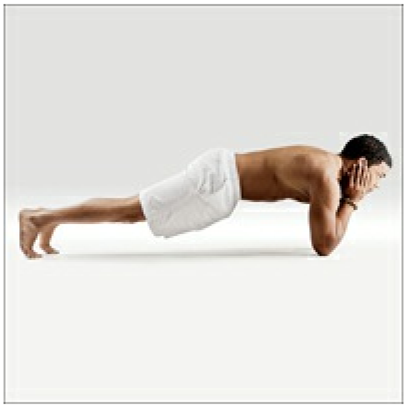

Leg Position of One-Legged King Pigeon 1 Version B in Repose Staff

Pose

Pada Eka Pada Raja Kapotasana 1B in Sayana Dandasana

(PUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh in SHUH-yuh-nuh duhn-DAHS-uh-nuh)

Pose Type: elbow balance, mild backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र ह ॐ

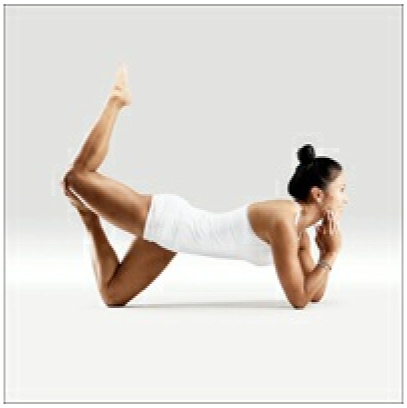

PLANK ON FOREARMS: TWIST

Revolved One Hand Staff Pose Dedicated to Makara

Parivritta Eka Hasta Makara Dandasana

(puh-ri-VRIT-tuh EY-kuh HUH-stuh MUH-kuh-ruh duhn-DAHS-uh-nuh)

Also Known As: Revolved Dolphin Plank Modification

Modification: 1 & 2. knees on the floor, front and back view
3 & 4. knees off the floor, front and back view

Pose Type: forearm balance, twist, core

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

ल व य

1.

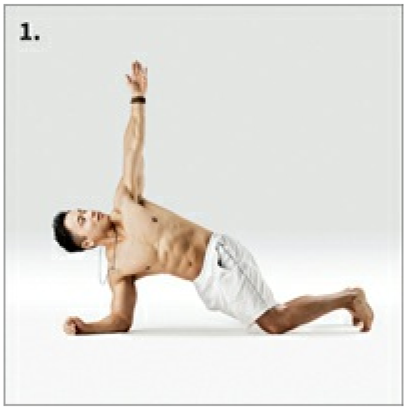

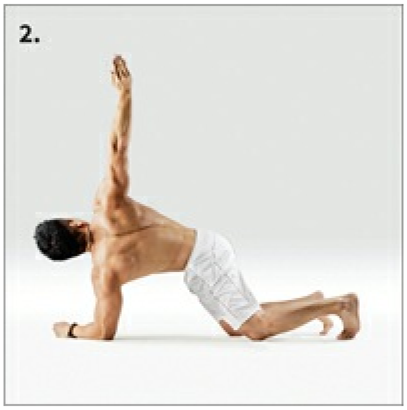
2.

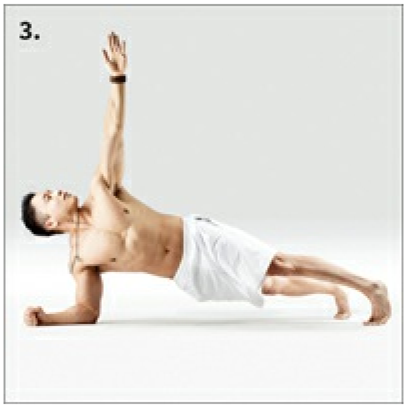
3.

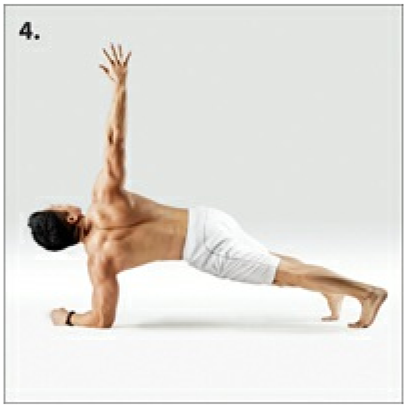
4.

PLANK ON FOREARMS: UNEVEN LEG AND ARM POSITIONS

Uneven Staff Pose Dedicated to Makara

Vishama Makara Dandasana

(VISH-uh-muh MUH-kuh-ruh duhn-DAHS-uh-nuh)

Also Known As: Uneven Dolphin Plank Modification
Modification: one elbow bent at 90 degrees, other forearm to the floor, knees on the floor
Pose Type: arm/forearm balance, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

Hand to the Leg Sideways Uneven Legs Staff Pose Dedicated to Makara

Hasta Pada Parshva Vishama Pada Makara Dandasana

(HUH-stuh PUH-duh PAHRSH-vuh VISH-uh-muh PUH-duh MUH-kuh-ruh duhn-DAHS-uh-nuh)

Also Known As: Hand to the Leg Sideways Dolphin Plank Modification
Pose Type: forearm balance, twist, core
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

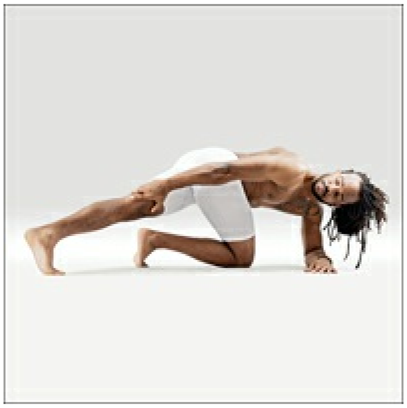

Uneven Staff Pose Dedicated to Makara

Vishama Makara Dandasana

(VISH-uh-muh MUH-kuh-ruh duhn-DAHS-uh-nuh)

Also Known As: Uneven Dolphin Plank Modification
Modification: one elbow bent at 90 degrees, other forearm to the floor, knees off the floor 1. left forearm on the floor, right arm at 90 degrees
2. right forearm on the floor, left arm at 90 degrees
Pose Type: arm/forearm balance, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

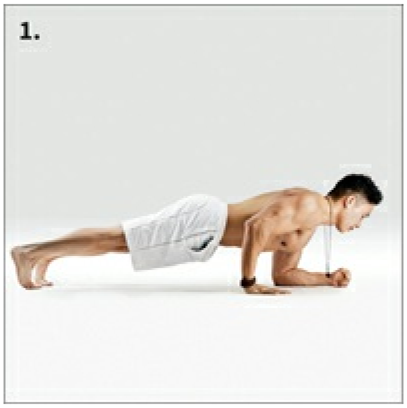

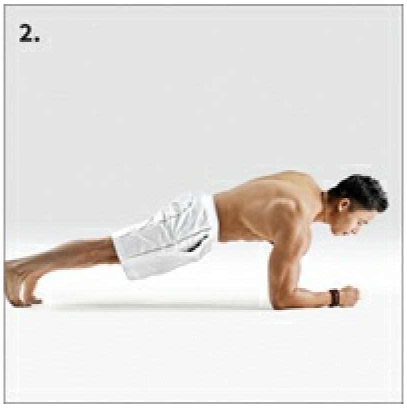

FOUR LIMBED STAFF POSE: KNEES ON THE FLOOR

Four Limbed Staff Pose

Chaturanga Dandasana

(chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: knees on the floor
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

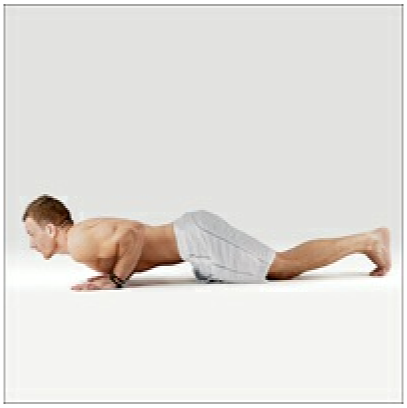

One Hand Four Limbed Staff Pose

Eka Hasta Chaturanga Dandasana

(EY-kuh HUH-stuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: knees on the floor
1. starting position
2. end position

Pose Type: arm balance, core

Drishti Point: Nasagrai or Nasagre (nose)

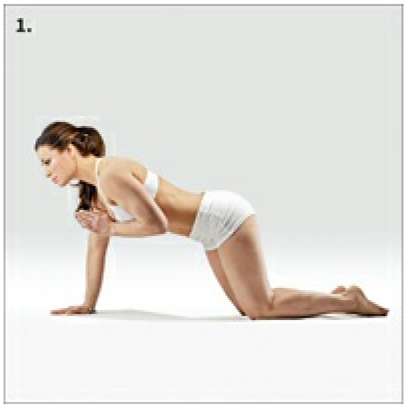

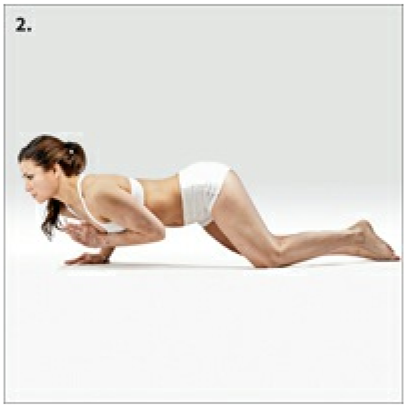

Lion Pose Dedicated to an Avatar of Lord Vishnu in Four Limbed Staff Pose

Narasimhasana in Chaturanga Dandasana

(nuh-ruh-sim-HAHS-uh-nuh in chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Also Known As: Lion Pose (Simhasana)

Modification: 1. on the fingertips
2. on the fists

Pose Type: arm balance, core

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

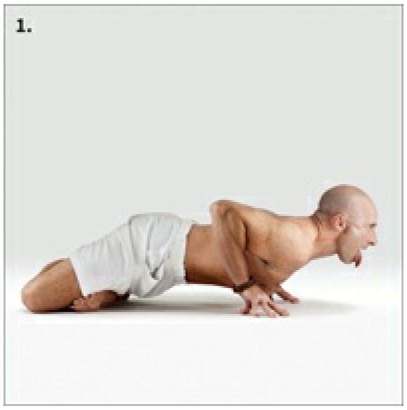

1.

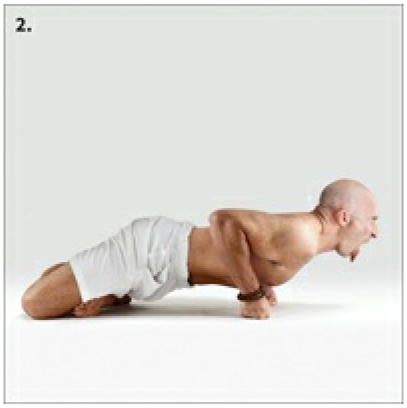

2.

Leg Position of One-Legged King Pigeon 1 Version B in Four Limbed Staff Pose

Pada Eka Pada Raja Kapotasana 1 B in Chaturanga Dandasana

(PUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh in chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: bottom knee bent and on the floor; top leg straight—knee resting on the sole of the bottom foot

Pose Type: arm balance, core

Drishti Point: Nasagrai or Nasagre (nose)

ल व ह

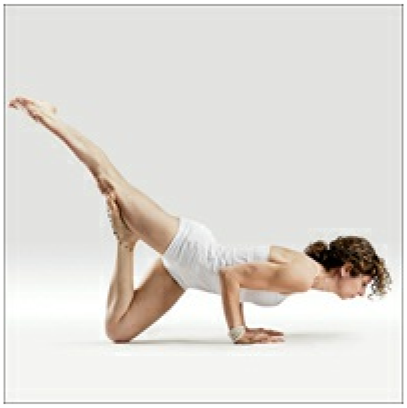

Four Limbed Staff Pose

Chaturanga Dandasana

(chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

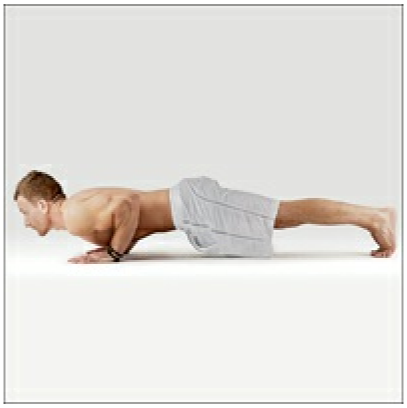

Crocodile Pose

Nakrasana

(nuh-KRAHS-uh-nuh)

Modification: feet and hands off the floor and up in the air
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

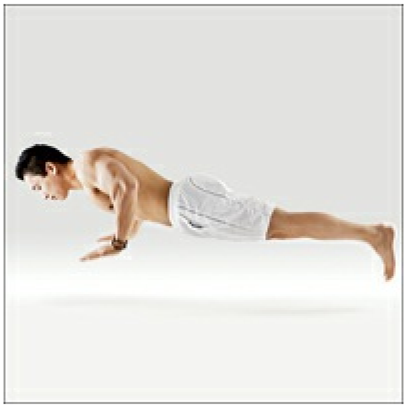

Revolved Four Limbed Staff Pose

Parivritta Chaturanga Dandasana

(puh-ri-VRIT-tuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Pose Type: arm balance, twist, core

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

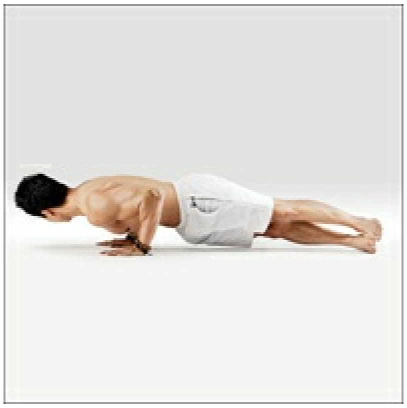

One-Legged Four Limbed Staff Pose

Eka Pada Chaturanga Dandasana

(EY-kuh PUH-duh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Pose Type: arm balance, core

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

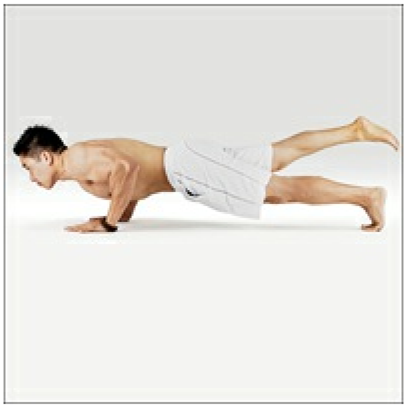

Upward One Leg to the Side Four Limbed Staff Pose

Urdhva Parshva Eka Pada Chaturanga Dandasana

(OORD-vuh PAHRSH-vuh EY-kuh PUH-duh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Pose Type: arm balance, core

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

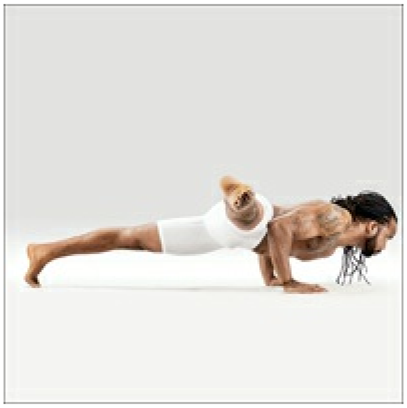

FOUR LIMBED STAFF POSE: KNEE TO THE ARM

One-Legged Four Limbed Staff Pose

Eka Pada Chaturanga Dandasana

(EY-kuh PUH-duh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: knee to the forearm
Pose Type: arm balance, forward bend, core
Drishti Point: Nasagrai or Nasagre (nose)

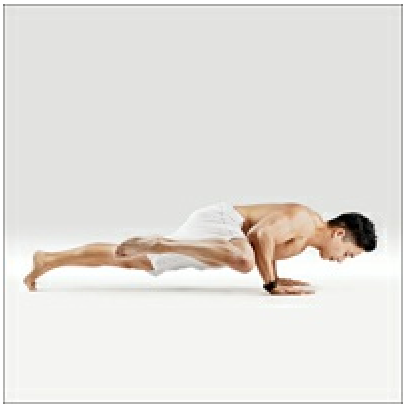

One-Legged Four Limbed Staff Pose

Eka Pada Chaturanga Dandasana

(EY-kuh PUH-duh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: knee on top of the tricep
Pose Type: arm balance, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

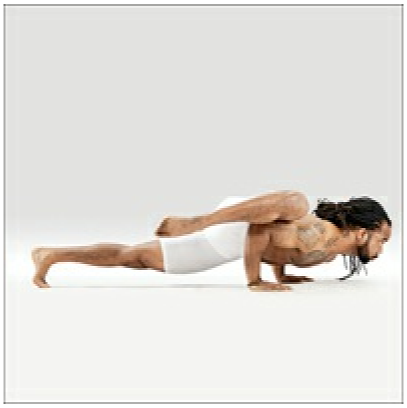

FOUR LIMBED STAFF POSE: ONE ARMED

One Hand Four Limbed Staff Pose

Eka Hasta Chaturanga Dandasana

(EY-kuh HUH-stuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: hand to the heart
Pose Type: arm balance, core
Drishti Point: Hastagrai or Hastagre (hands) or Nasagrai or Nasagre (nose)

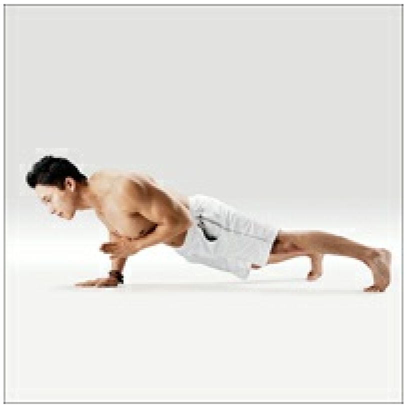

Hand to the Side Four Limbed Staff Pose

Parshva Hasta Chaturanga Dandasana

(PAHRSH-vuh HUH-stuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: 1. knees on the floor
2. knees off the floor
Pose Type: arm balance, core
Drishti Point: Hastagrai or Hastagre (hands) or Nasagrai or Nasagre (nose)

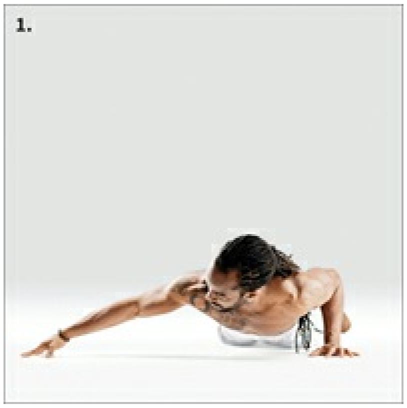

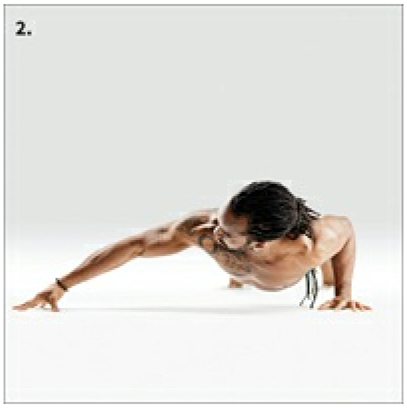

One Hand to the Side Four Limbed Staff Pose

Parshva Eka Hasta Chaturanga Dandasana

(PAHRSH-vuh EY-kuh HUH-stuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: one knee on the floor, other knee off the floor
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

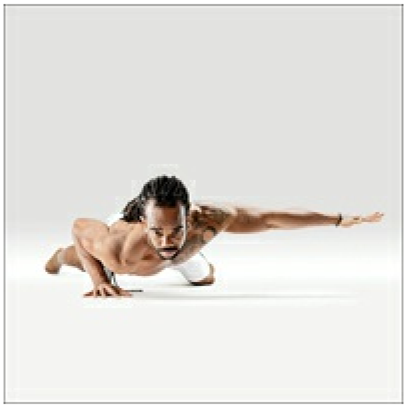

One Hand to the Side Four Limbed Staff Pose

Parshva Eka Hasta Chaturanga Dandasana

(PAHRSH-vuh EY-kuh HUH-stuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: knees off the floor
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

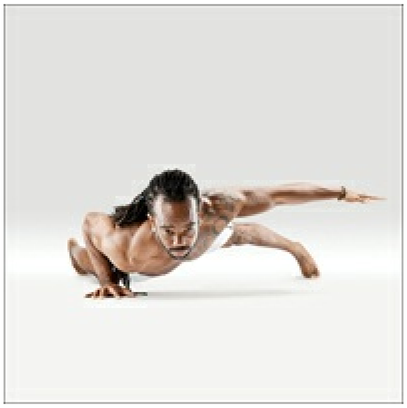

One Hand to the Side One-Legged Four Limbed Staff Pose

Parshva Eka Hasta Eka Pada Chaturanga Dandasana

(PAHRSH-vuh EY-kuh HUH-stuh EY-kuh PUH-duh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: one knee on the floor
Pose Type: arm balance, core
Drishti Point: Hastagrai or Hastagre (hands)

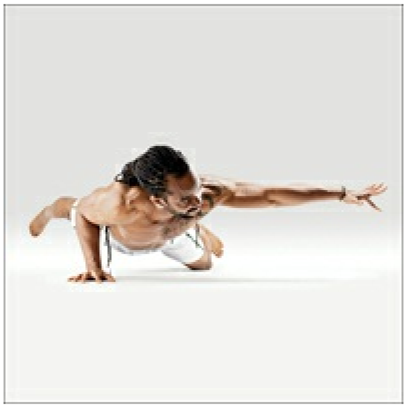

FOUR LIMBED STAFF POSE: PUSH UPS

Diamond Four Limbed Staff Pose

Vajra Chaturanga Dandasana

(VAHJ-ruh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: knees on the floor

1. start position
2. end position

Pose Type: arm balance, core

Drishti Point: Nasagrai or Nasagre (nose)

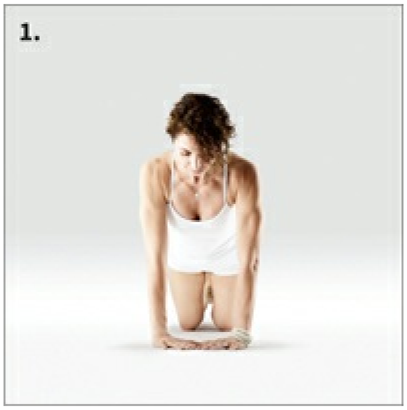

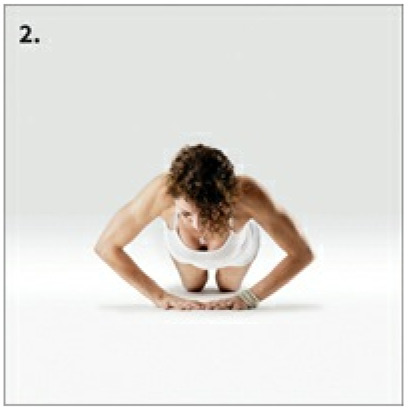

Diamond Four Limbed Staff Pose

Vajra Chaturanga Dandasana

(VAHJ-ruh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: knees off the floor

1. start position
2. end position

Pose Type: arm balance, core

Drishti Point: Nasagrai or Nasagre (nose)

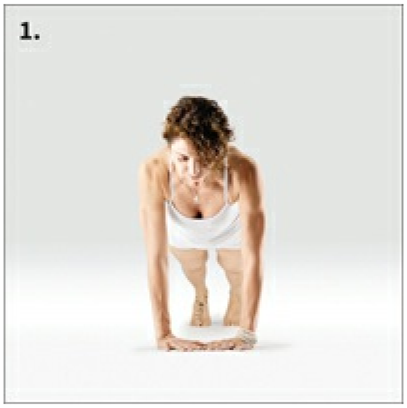

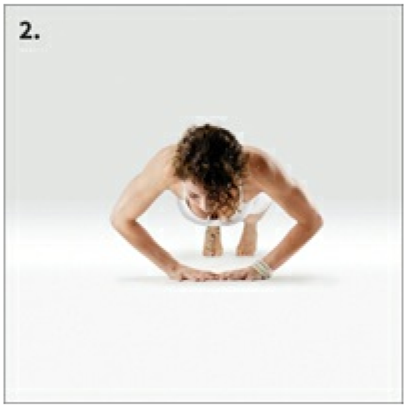

Hands Spread Wide Four Limbed Staff Pose

Prasarita Hasta Chaturanga Dandasana

(pruh-SAH-ri-tuh HUH-stuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: knees on the floor
1. start position
2. end position
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose)

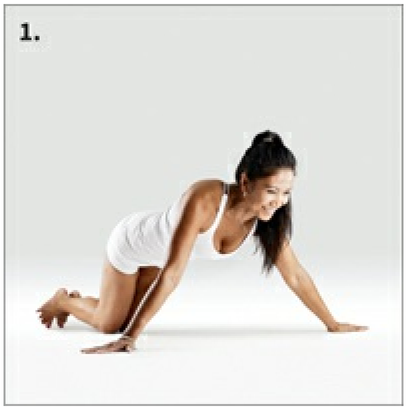

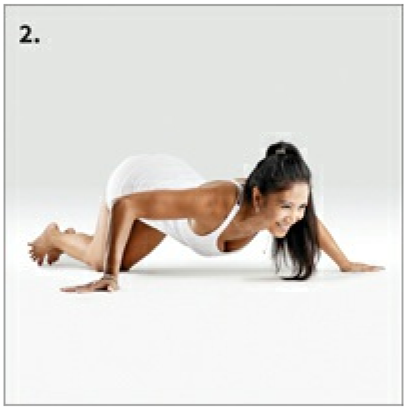

Hands Spread Wide Four Limbed Staff Pose

Prasarita Hasta Chaturanga Dandasana

(pruh-SAH-ri-tuh HUH-stuh chuh-tur-UHNG-guh duhn-DAHS-uh-nuh)

Modification: knees off the floor

1. start position

2. end position

Pose Type: arm balance, core

Drishti Point: Nasagrai or Nasagre (nose)

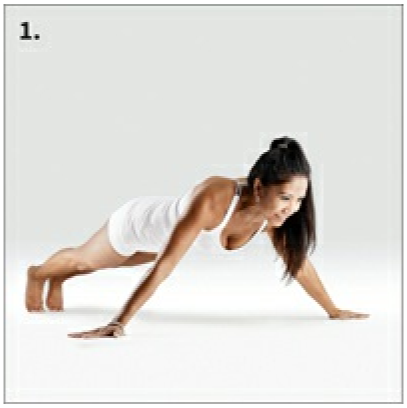

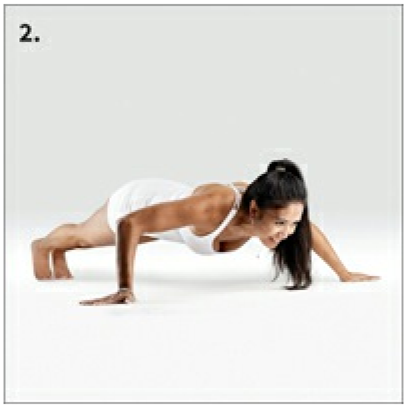

SIDE PLANK: TOP LEG CROSSED OVER

Pose Dedicated to Sage Vasistha

Vasisthasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Side Plank Modification
Modification: top leg crossing over
1. arm alongside of the torso
2. arm extended up
Pose Type: arm balance, core
Drishti Point: Nasagrai or Nasagre (nose), Angushtamadhye or Angushta Ma Dyai (thumbs)

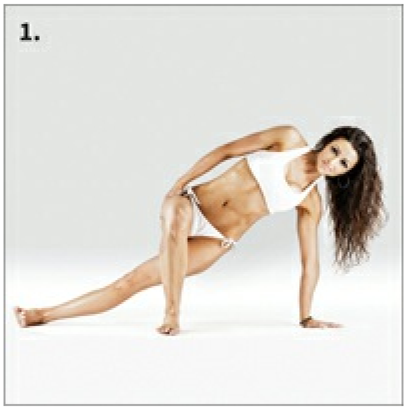

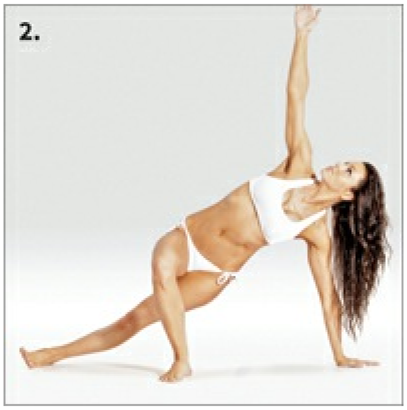

Pose Dedicated to Sage Vasistha

Vasisthasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Side Plank Modification
Modification: bottom knee to the floor
1. arm alongside of the torso
2. arm extended up
Pose Type: arm balance, core
Drishti Point: Angushtamadhye or Angushtha Ma Dyai (thumbs)

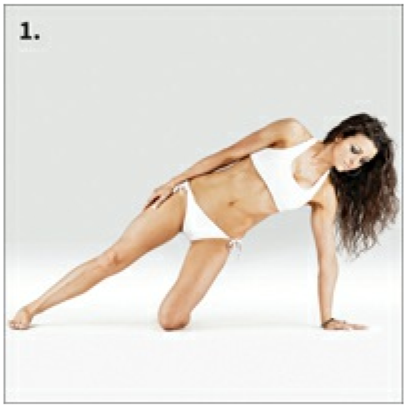

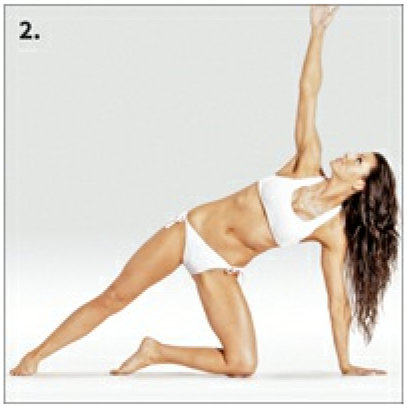

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Side Plank Modification

Modification: both knees bent, shin of the top leg on the inside of the knee of the bottom leg

1. bottom leg bent at 90 degrees
2. bottom leg slightly bent

Pose Type: arm balance, core

Drishti Point: Urdhva or Antara Drishti (up to the sky)

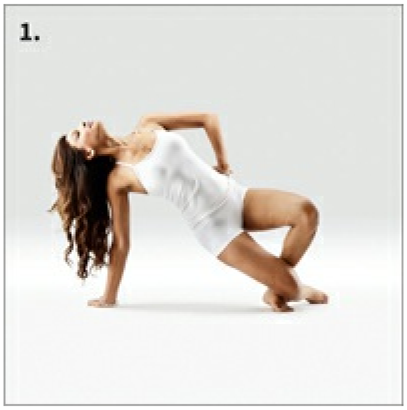

2.

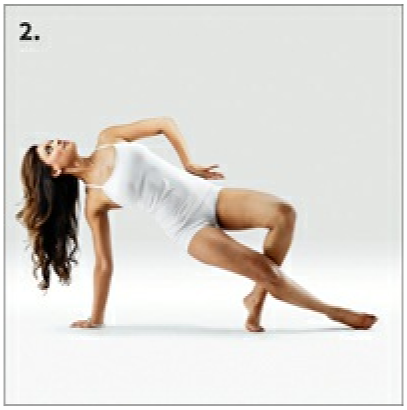

SIDE PLANK: LEGS STRAIGHT

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Side Plank Modification
Modification: both legs straight
1. arm alongside of the torso
2. arm extended up
Pose Type: arm balance, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

1.

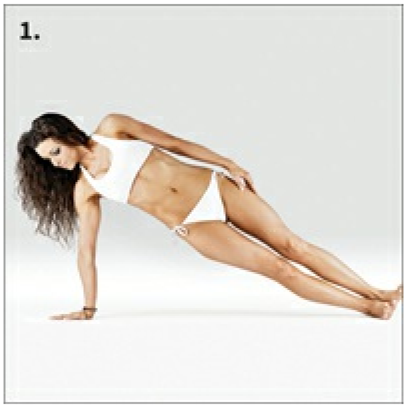

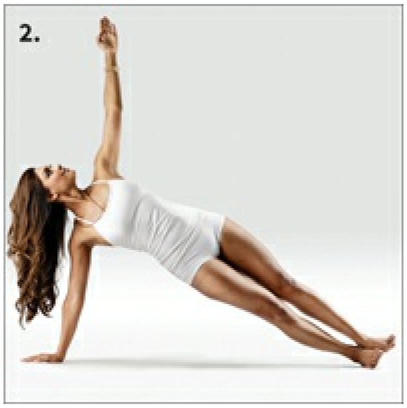

SIDE PLANK: BOTTOM LEG OFF THE FLOOR

Revolved Pose Dedicated to Sage Vasistha

Parivritta Vasisthasana

(puh-ri-VRIT-tuh vuh-sish-TAHS-uh-nuh)

Also Known As: Revolved Side Plank Modification
Modification: bottom knee bent, foot resting on the thigh of the top leg; the wrist of the top arm to the knee of the bottom leg
Pose Type: arm balance, twist, forward bend, core
Drishti Point: Hastagrai or Hastagre (hands)

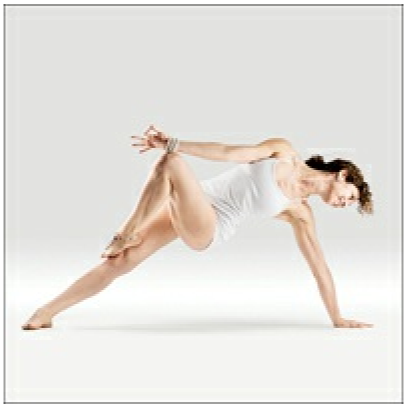

Revolved Leg to the Side Pose Dedicated to Sage Vasistha

Parivritta Parshva Pada Vasisthasana

(puh-ri-VRIT-tuh PAHRSH-vuh PUH-duh vuh-sish-TAHS-uh-nuh)

Also Known As: Revolved Leg to the Side Plank Modification
Modification: both legs straight, bottom foot lifted off the floor, leg crossing over in front of the body
Pose Type: arm balance, twist, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Padayoragrai or Padayoragre

(toes/feet)

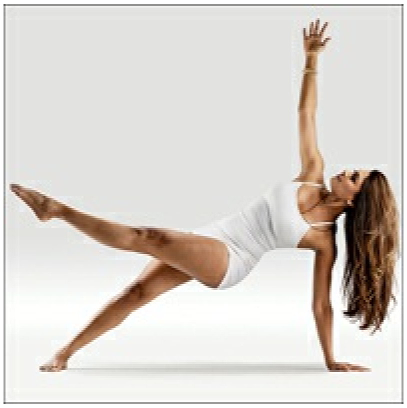

Hand to Foot Pose Dedicated to Sage Vasistha

Hasta Pada Vasisthasana

(HUH-stuh PUH-duh vuh-sish-TAHS-uh-nuh)

Also Known As: Hand to Foot Side Plank Modification

Modification: both legs straight; bottom leg crossing over in front of the body, foot in line with the center of the chest

Pose Type: arm balance, forward bend, core

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

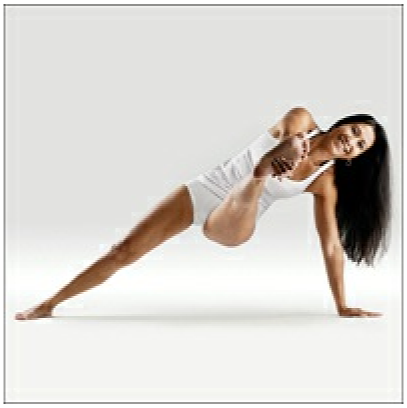

SIDE PLANK: FOOT TO THE THIGH & HALF LOTUS

Tree Pose in Pose Dedicated to Sage Vasista

Vrikshasana in Vasishtasana

(vrik-SHAHS-uh-nuh in vuh-sish-TAHS-uh-nuh)

Also Known As: Tree Pose in Side Plank Modification

Modification: top foot to the inner thigh
Pose Type: arm balance, core
Drishti Point: Urdhva or Antara Drishti (up to the sky)
लं वं

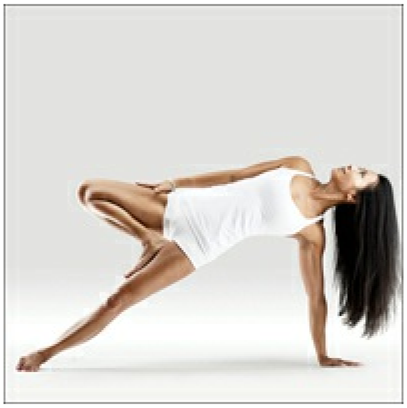

Half Lotus Pose in Pose Dedicated to Sage Vasishta

Ardha Padmasana in Vasishtasana

(UHR-duh puhd-MAHS-uh-nuh in vuh-sish-TAHS-uh-nuh)
Also Known As: Half Lotus Pose in Side Plank Modification
Modification: bottom knee on the floor
Pose Type: arm balance, core
Drishti Point: Hastagrai or Hastagre (hands)
लं वं हं

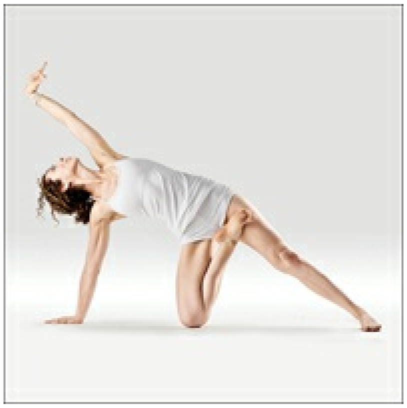

SIDE PLANK: TOP LEG OFF THE FLOOR

One Big Toe Pose Dedicated to Sage Vasishta

Eka Padangushta Vasisthasana

(EY-kuh puhd-ahng-GOOSH-tuh vuh-sish-TAHS-uh-nuh)

Also Known As: Big Toe Side Plank Modification
Modification: 1. bottom knee on the floor—leg bent, toes pointing to the back; on the fingertips
2. bottom knee off the floor; palm flat to the floor
Pose Type: arm balance, forward bend, core
Drishti Point: Urdhva or Antara Drishti (up to the sky), Padayoragrai or Padayoragre (toes/feet)
ल व य

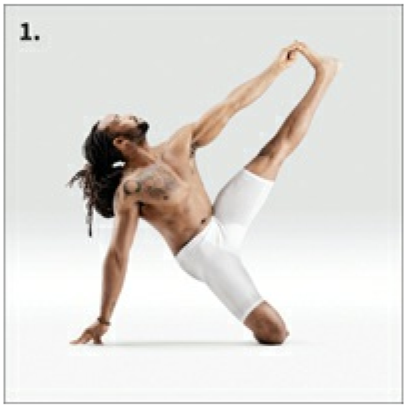

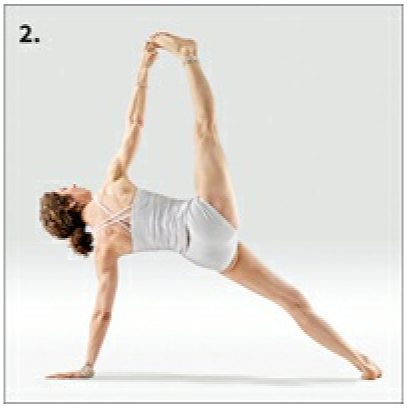

VISHVAMITRA'S POSE

Pose Dedicated to Vishvamitra

Vishvamitrasana

(vish-vuh-mi-TRAHS-uh-nuh)
Modification: knee on the floor, grabbing onto the outside edge of the foot
Pose Type: arm balance, side bend, twist, core
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व र य ह ॐ

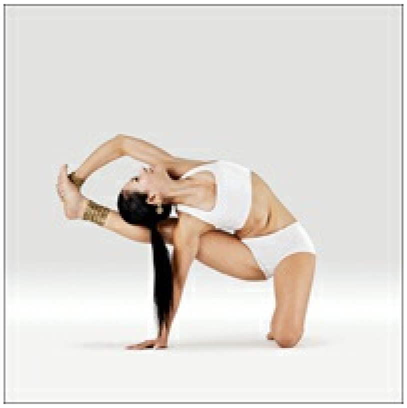

Pose Dedicated to Vishvamitra

Vishvamitrasana

(vish-vuh-mi-TRAHS-uh-nuh)

Modification: arm alongside of the torso
Pose Type: arm balance, core
Drishti Point: Urdhva or Antara Drishti (up to the sky)
ल व

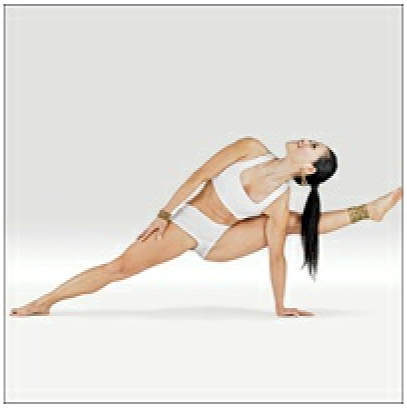

Pose Dedicated to Vishvamitra

Vishvamitrasana

(vish-vuh-mi-TRAHS-uh-nuh)

Modification: knee off the floor, top arm straight and extended up to the sky
Pose Type: arm balance, side bend, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व य

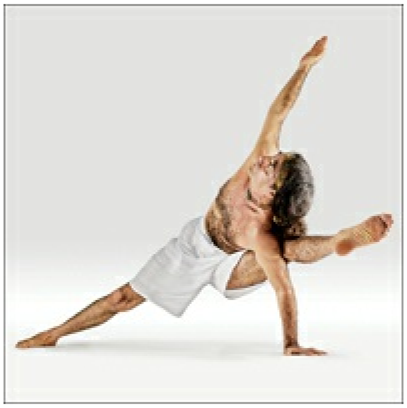

Pose Dedicated to Vishvamitra

Vishvamitrasana

(vish-vuh-mi-TRAHS-uh-nuh)

Modification: knee off the floor
1. grabbing onto the outside edge of the top foot
2. grabbing onto the ankle of the top leg
Pose Type: arm balance, side bend, twist, core
Drishti Point: Urdhva or Antara Drishti (up to the sky)

ल व र य

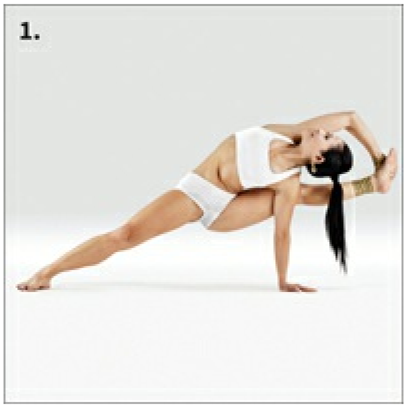

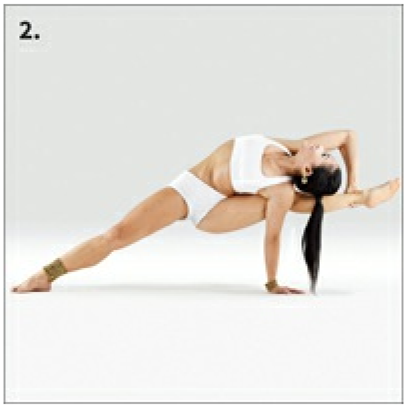

KALA BHAIRAVA’S POSE

Pose Dedicated to Shiva the Destroyer

Kala Bhairavasana

(KAH-luh beye-ruh-VAHS-uh-nuh)

Modification: 1. arm extended up
2. top hand to the inside of the bottom thigh
Pose Type: arm balance, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Padayoragrai or Padayoragre (toes/feet)

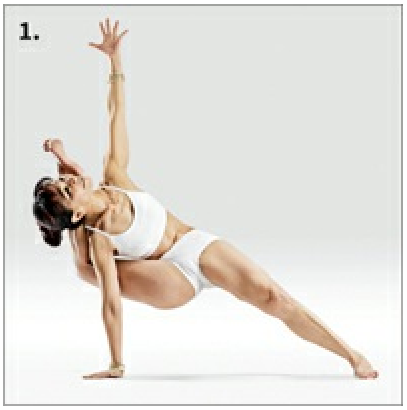

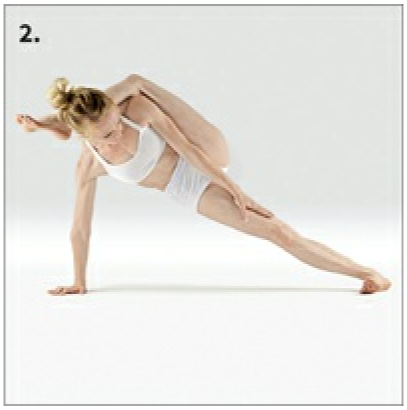

SIDE PLANK: TOP LEG OVERHEAD GRIP

Partridge Pose

Kapinjalasana

(kuh-pinj-uh-LAHS-uh-nuh)

Modification: overhead grip
1. bottom knee on the floor
2. bottom knee off the floor
Pose Type: arm balance, backbend, core

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

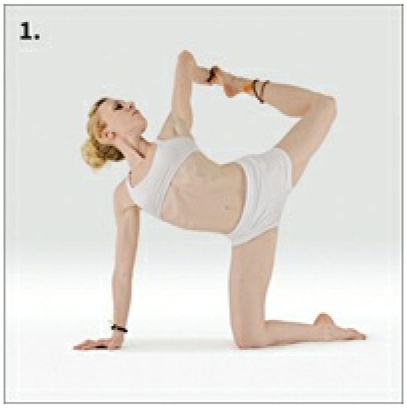
1.

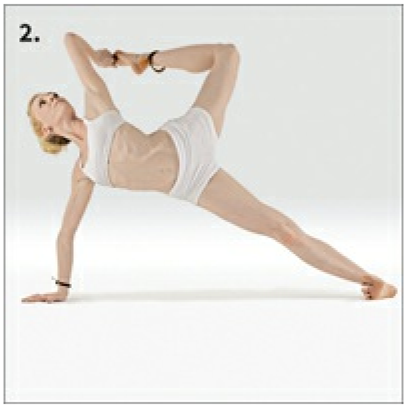
2.

SIDE PLANK ON FOREARMS: KNEES ON THE FLOOR

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification
Modification: forearm on the floor, both knees to the floor
1. arm up to the sky
2. arm alongside the torso
Pose Type: forearm balance, core
Drishti Point: 1. Angushtamadhye or Angushta Ma Dyai (thumbs)
2. Nasagrai or Nasagre (nose)

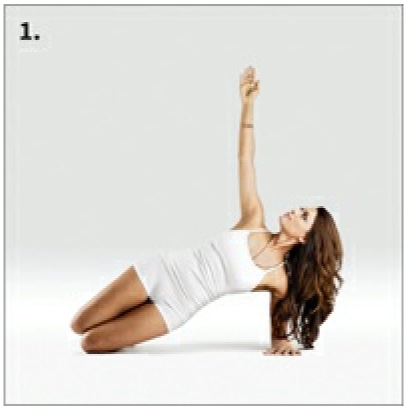

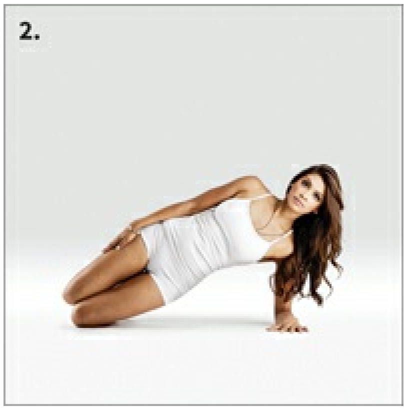

SIDE PLANK ON FOREARMS: KNEES OFF THE FLOOR

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification
Modification: forearm on the floor, arm crossed over in front of the chest, looking down
Pose Type: forearm balance, twist, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

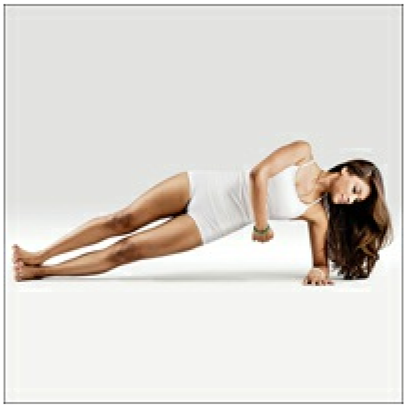

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification
Modification: forearm on the floor, top hand on the hip
Pose Type: forearm balance, core
Drishti Point: Nasagrai or Nasagre (nose)
●

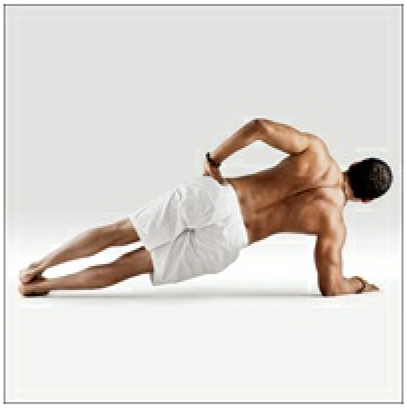

Pose Dedicated to Sage Vasistha

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification
Modification: forearm on the floor, top arm up to the sky
Pose Type: forearm balance, core
Drishti Point: Hastagrai or Hastagre (hands)
● ●

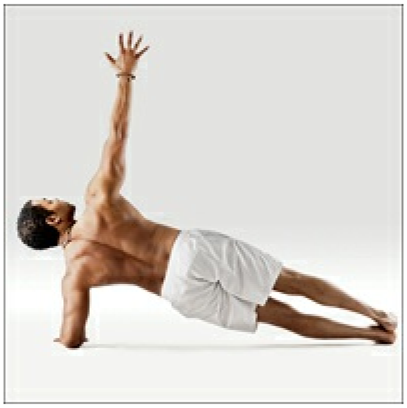

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification

Modification: forearm on the floor, top hand to the back of the head

Pose Type: forearm balance, core

Drishti Point: Nasagrai or Nasagre (nose)

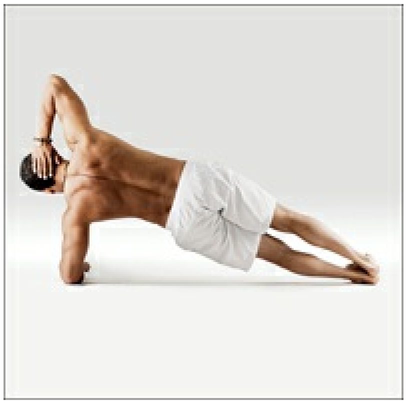

SIDE PLANK ON FOREARMS: ONE LEG CROSSED OVER

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification

Modification: forearm on the floor; top leg crossed over, foot flat on the floor; top hand by the bottom hip socket

Pose Type: forearm balance, core

Drishti Point: Urdhva or Antara Drishti (up to the sky)

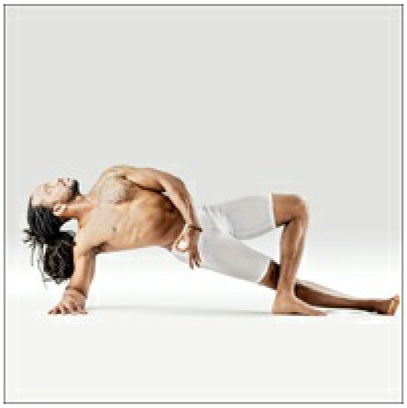

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification

Modification: forearm on the floor; top leg crossed over, heel lifted; top arm extended up to the sky

Pose Type: forearm balance, core

Drishti Point: Hastagrai or Hastagre (hands)

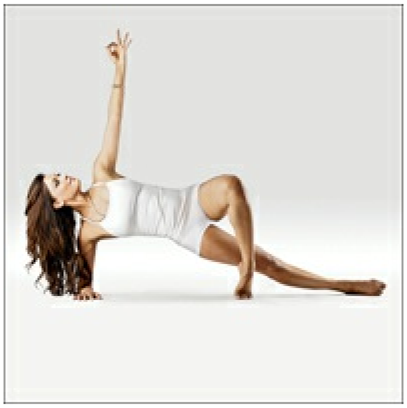

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification

Modification: forearm on the floor; top leg crossed over, foot flat on the floor, bottom leg lifted off the floor, top hand on the hip

Pose Type: forearm balance, core

Drishti Point: Urdhva or Antara Drishti (up to the sky)

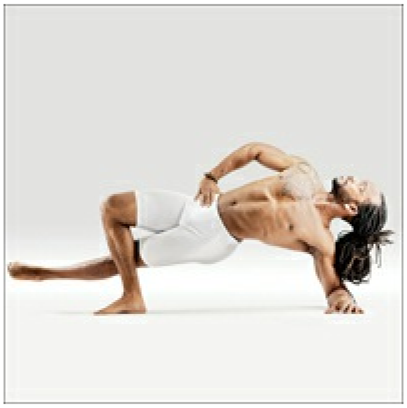

Half Lotus Pose in Pose Dedicated to Sage Vasishta

Ardha Padmasana in Vasishtasana

(UHR-duh puhd-MAHS-uh-nuh in vuh-sish-TAHS-uh-nuh)

Also Known As: Half Lotus Pose in Dolphin Side Plank Modification

Modification: forearm on the floor, top arm extended up to the sky

Pose Type: forearm balance, core

Drishti Point: Hastagrai or Hastagre (hands)

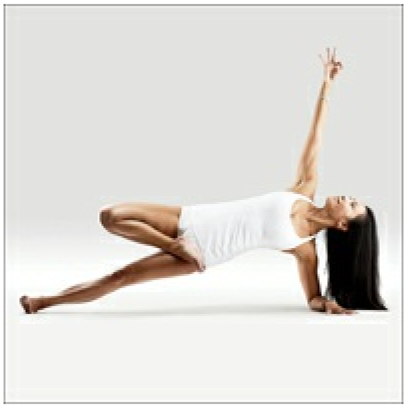

SIDE PLANK ON FOREARMS: TOP LEG LIFTED

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification

Modification: forearm on the floor; top knee toward the top elbow, bottom knee on the floor; leg bent, toes pointing to the back

Pose Type: forearm balance, forward bend, core

Drishti Point: Hastagrai or Hastagre (hands)

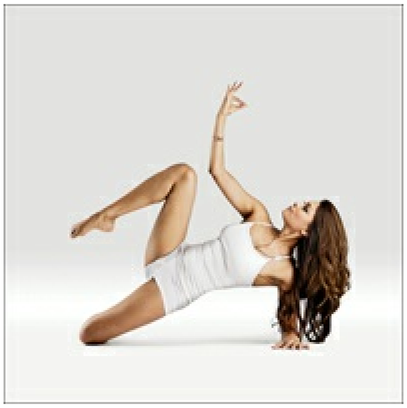

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Dolphin Side Plank Modification
Modification: forearm on the floor; top knee toward the top elbow; bottom leg straight, knee off floor
Pose Type: forearm balance, forward bend, core
Drishti Point: Hastagrai or Hastagre (hands)
लं वं

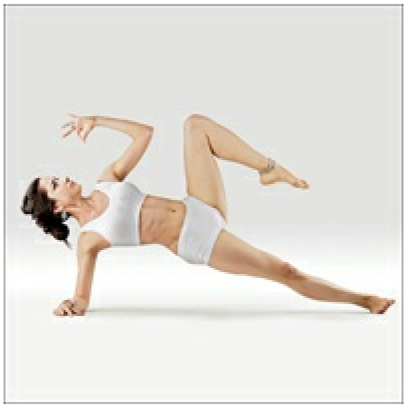

One Big Toe Pose Dedicated to Sage Vasishta

Eka Padangushta Vasishtasana

(EY-kuh puhd-ahng-GOOSH-tuh vuh-sish-TAHS-uh-nuh)

Also Known As: Big Toe Dolphin Side Plank Modification
Modification: forearm on the floor
1. knee bent
2. leg straight
Pose Type: forearm balance, forward bend, core
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
लं वं यं

1.

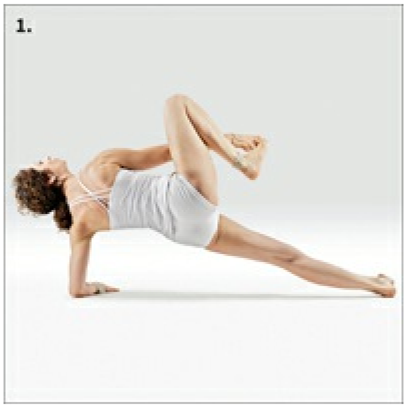

2.

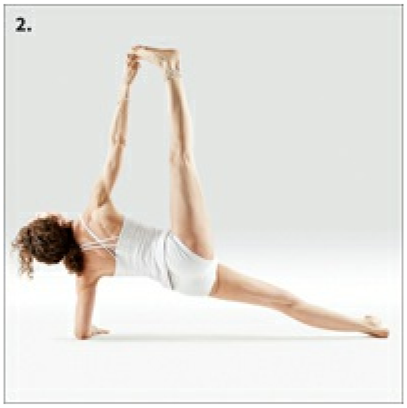

SIDE PLANK ON FOREARMS: BOTTOM LEG LIFTED

Revolved Leg to the Side Pose Dedicated to Sage Vasishta

Parivritta Parshva Pada Vasishtasana

(puh-ri-VRIT-tuh PAHRSH-vuh PUH-duh vuh-sish-TAHS-uh-nuh)

Also Known As: Revolved Leg to the Side Dolphin Side Plank Modification
Modification: forearm on the floor; both legs straight, bottom leg lifted in front of the body
Pose Type: forearm balance, twist, core
Drishti Point: Hastagrai or Hastagre (hands)

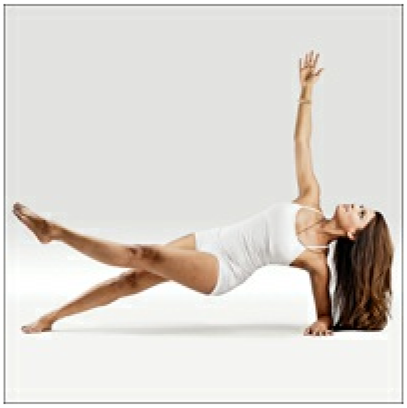

Revolved Pose Dedicated to Sage Vasishta

Parivritta Vasishtasana

(puh-ri-VRIT-tuh vuh-sish-TAHS-uh-nuh)

Also Known As: Revolved Dolphin Side Plank Modification
Modification: forearm on the floor; bottom knee bent, foot resting on the thigh of the top leg; the elbow of the top arm to the knee of the bottom leg
Pose Type: forearm balance, forward bend, twist, core
Drishti Point: Urdhva or Antara Drishti (up to the sky)

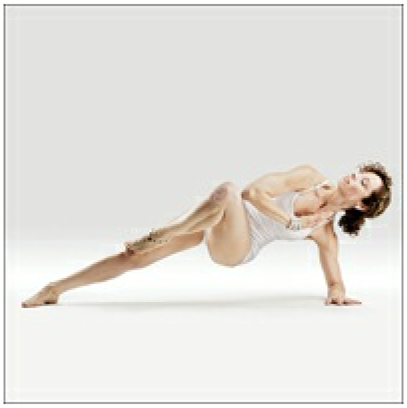

SIDE PLANK ON FOREARMS: TOP LEG UNDER-HEAD GRIP

Partridge Pose

Kapinjalasana

(kuh-pinj-uh-LAHS-uh-nuh)

Modification: forearm on the floor, under-head grip:
1. knee on the floor
2. knee off the floor

Pose Type: forearm balance, backbend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

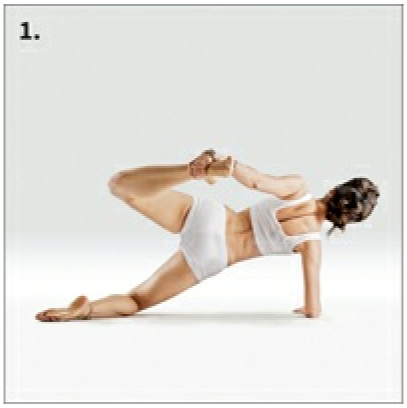

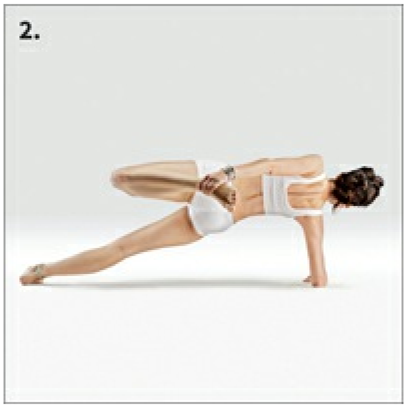

SIDE PLANK ON FOREARMS: HALF LOTUS & LOTUS

Half Lotus Pose in Pose Dedicated to Sage Vasistha

Ardha Padmasana in Vasishtasana

(UHR-duh puhd-MAHS-uh-nuh in vuh-sish-TAHS-uh-nuh)

Also Known As: Half Lotus Dolphin Side Plank Modification
Modification: forearm on the floor, bottom leg in half lotus, top knee bent, toes to the floor
Pose Type: forearm balance, core
Drishti Point: Hastagrai or Hastagre (hands)
ल व य ह

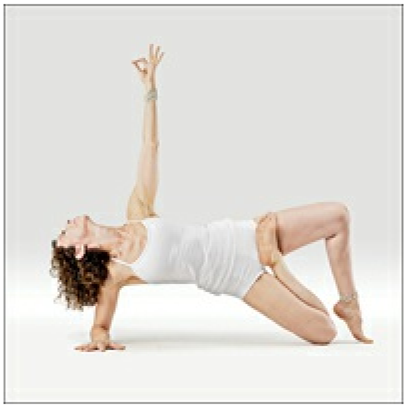

Half Lotus One Big Toe Pose Dedicated to Sage Vasistha

Ardha Padma Eka Padangushta Vasisthasana

(UHR-duh PUHD-muh EY-kuh puhd-ahng-GOOSH-tuh vuh-sish-TAHS-uh-nuh)

Also Known As: Half Lotus Big Toe Dolphin Side Plank Modification
Modification: forearm on the floor
Pose Type: forearm balance, forward bend, core
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Urdhva or Antara Drishti (up to the sky)
ल व य ह

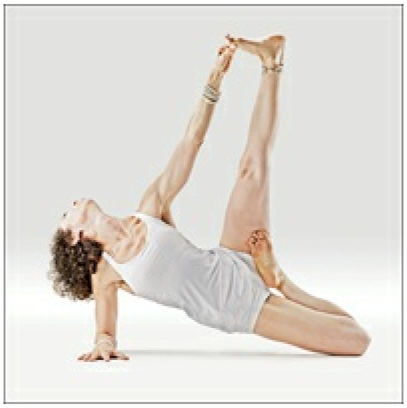

Half Lotus Half Bow Pose in Pose Dedicated to Sage Vasishta

Ardha Padma Dhanurasana in Vasishtasana

(UHR-duh PUHD-muh duh-nur-AHS-uh-nuh in vuh-sish-TAHS-uh-nuh)

Also Known As: Half Lotus Half Bow Pose in Dolphin Side Plank Modification
Modification: forearm on the floor, bottom leg in half lotus, grabbing onto the top foot with under-head grip
Pose Type: forearm balance, backbend, core
Drishti Point: Hastagrai or Hastagre (hands)
ल व र य

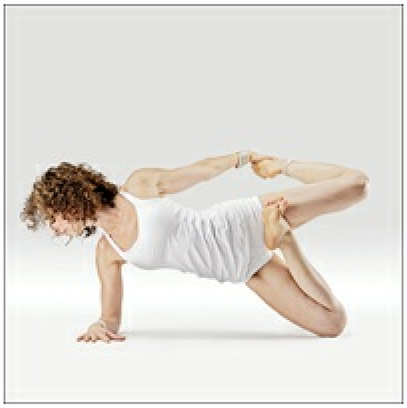

Lotus Pose in Pose Dedicated to Sage Vasishta

Padmasana in Vasisthasana

(puhd-MAHS-uh-nuh in vuh-sish-TAHS-uh-nuh)

Also Known As: Lotus Pose in Dolphin Side Plank Modification
Modification: forearm on the floor
Pose Type: forearm balance, core
Drishti Point: Hastagrai or Hastagre (hands)

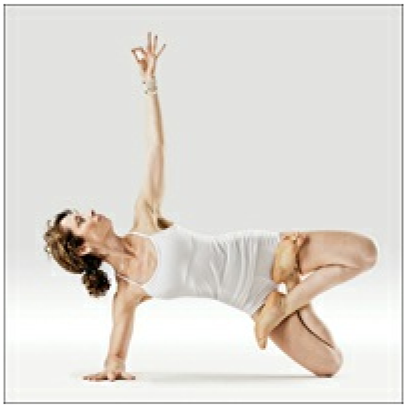

SIDE PLANK ON THE ELBOWS

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Elbow Side Plank Modification
Modification: on the elbow; both legs straight, ankles crossed
Pose Type: elbow balance, core
Drishti Point: Urdhva or Antara Drishti (up to the sky)

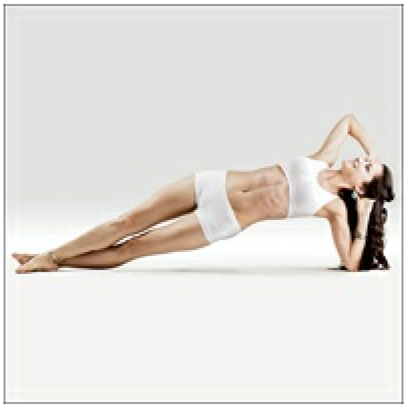

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Elbow Side Plank Modification
Modification: on the elbow; bottom leg straight; top leg crossed over, foot flat on the floor; fingertips of the top arm to the floor in front of the body
Pose Type: elbow balance, core
Drishti Point: Urdhva or Antara Drishti (up to the sky)

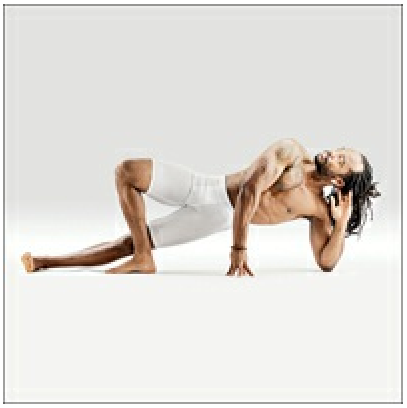

Pose Dedicated to Sage Vasishta

Vasishtasana

(vuh-sish-TAHS-uh-nuh)

Also Known As: Elbow Side Plank Modification
Modification: on the elbow, top foot to the side of the bottom leg
Pose Type: elbow balance, core
Drishti Point: Nasagrai or Nasagre (nose)

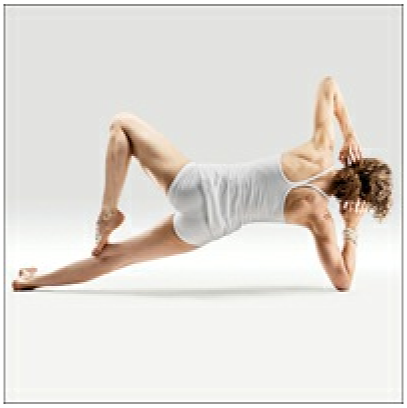

Half Lotus Pose in Pose Dedicated to Sage Vasistha

Ardha Padmasana in Vasisthasana

(UHR-duh puhd-MAHS-uh-nuh in vuh-sish-TAHS-uh-nuh)

Also Known As: Half Lotus Pose in Elbow Side Plank Modification
Modification: on the elbow
Pose Type: elbow balance, core
Drishti Point: Hastagrai or Hastagre (hands)

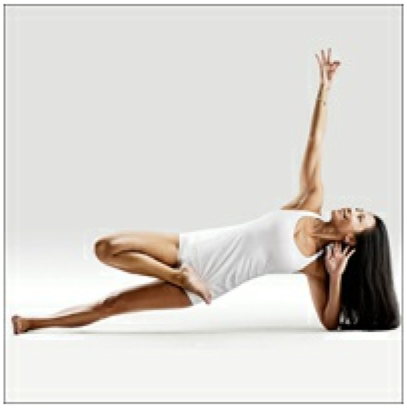

Backbends

COW FACE POSE AND GARUDA ON KNEES

KNEES ON THE FLOOR

BACKBEND ON THE KNEES

BACKBEND

UPWARD BOW POSE

WILD THING POSE

Pose Dedicated to Garuda on the Knees

Janu Garudasana

(JAH-nu guh-ru-DAHS-uh-nuh)

Pose Type: standing (on the knees), backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

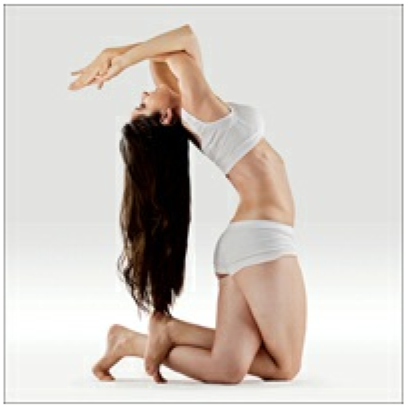

Leg Position of Cow Face Pose in Camel Pose

Pada Gomukhasana in Ushtrasana*

(PUH-duh go-moo-KAHS-uh-nuh in oosh-TRAHS-uh-nuh)

Modification: Gomukhasana legs modification
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

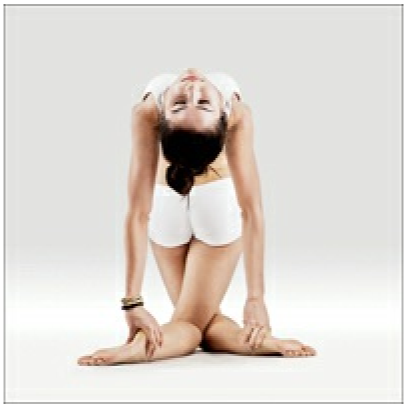

KNEES ON THE FLOOR: LOTUS POSE

Half Pose Dedicated to Siddhar Kamalamuni

Ardha Kamalamunyasana

(UHR-duh KUH-muh-luhmoo-nyAHS-uh-nuh)

Also Known As: Upward Facing Lotus Pose (Urdhva Mukha Padmasana)
Modification: fingertips pointing to the back, thumbs pointing to the front, knees close to the hands
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

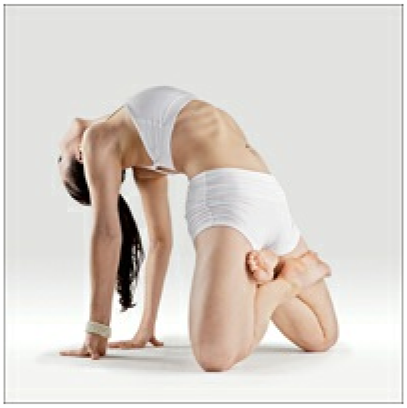

ONE KNEE TO THE FLOOR: ONE LEG BENT & ONE LEG STRAIGHT UP TO THE SKY

One-Legged Camel Pose

Eka Pada Ushtrasana

(EY-kuh PUH-duh oosh-TRAHS-uh-nuh)

Modification: Leg 1: knee to the floor Leg 2: foot to the floor in semi-lunge position
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह

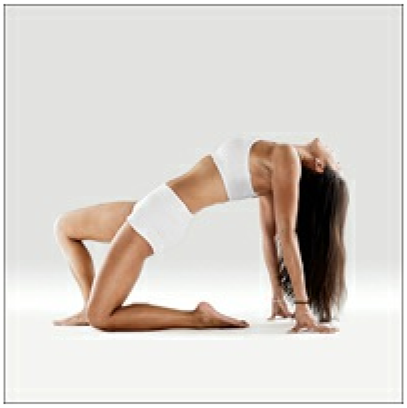

One-Legged Camel Pose

Eka Pada Ushtrasana

(EY-kuh PUH-duh oosh-TRAHS-uh-nuh)

Modification: Leg 1: knee to the floor
Leg 2: sole of the foot to the quadricep
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

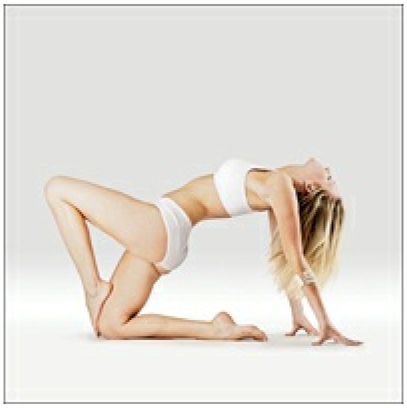

Extended Hand to Big Toe Pose in Half Camel Pose

Utthita Hasta Padangushtasana in Ardha Ushtrasana

(UT-ti-tuh HUH-stuh puhd-AHNG-goosh-TAHS-uh-nuh in UHR-duh oosh-TRAHS-uh-nuh)

Modification: 1. toes pointed to the back
2. toes curled in
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

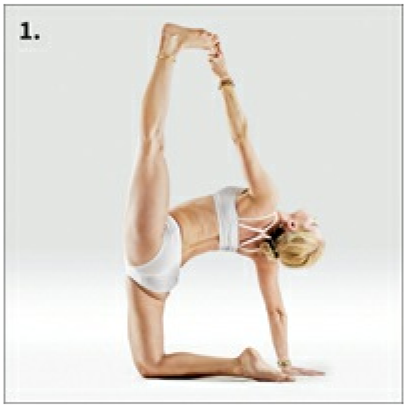

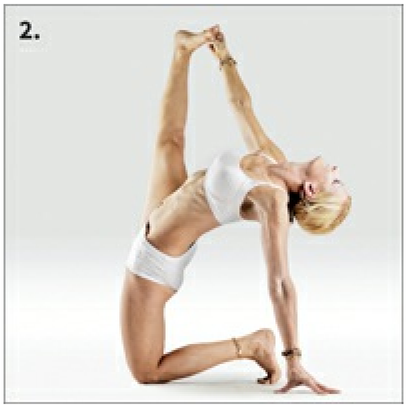

2.

KNEES TO THE FLOOR: BACKBEND

Camel Pose Prep.

Ushtrasana Prep.

(oosh-TRAHS-uh-nuh)

Modification: 1. hands to the lower back
2. hands in Anjali Mudra (Hands in Prayer)
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

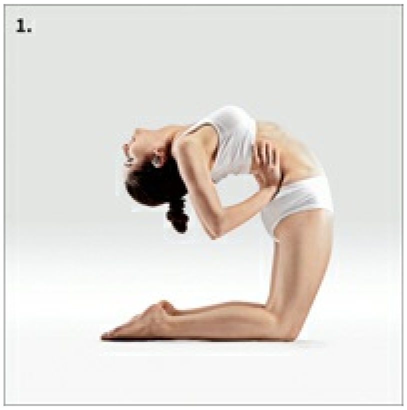

1.

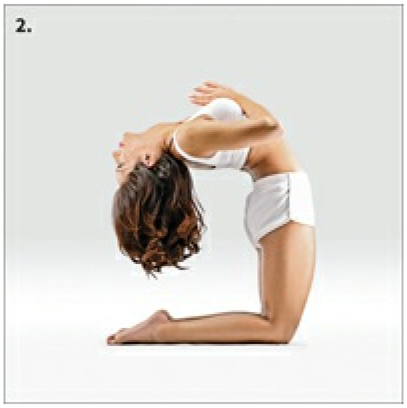

2.

Upward Salute Pose in Camel Pose

Urdhva Hastasana in Ushtrasana

(OORD-vuh huh-STAHS-uh-nuh in oosh-TRAHS-uh-nuh)

Also Known As: Pigeon Pose Prep. (Kapotasana Prep.)
Modification: palms together
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)

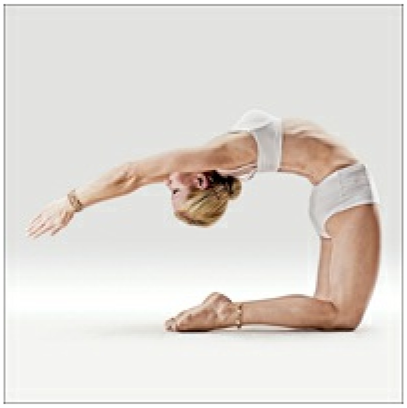

Raised Bound Hands in Camel Pose

Urdhva Baddha Hastasana in Ushtrasana

(OORD-vuh BUH-duh huh-STAHS-uh-nuh in oosh-TRAHS-uh-nuh)

Modification: 1. hips pushed back
2. hips pushed forward
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

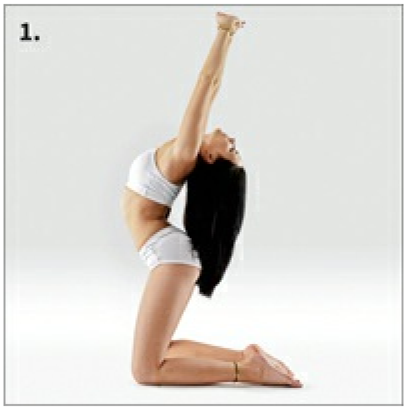

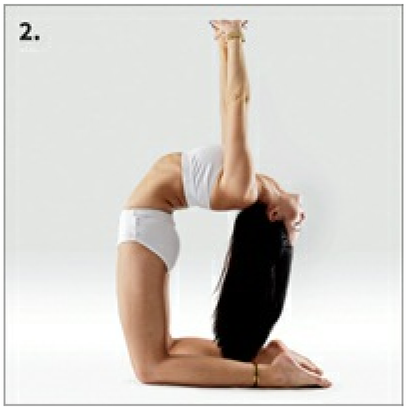
2.

KNEES ON THE FLOOR: BACKBEND—HANDS TO THE HEELS AND TO THE FLOOR

Camel Pose

Ushtrasana

(oosh-TRAHS-uh-nuh)

Modification: 1. grabbing onto both ankles
2. palms to the heels, fingertips pointing to the back
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

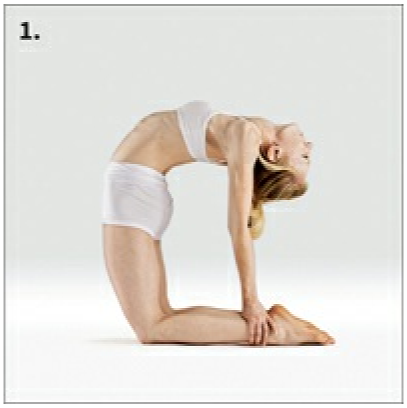
1.

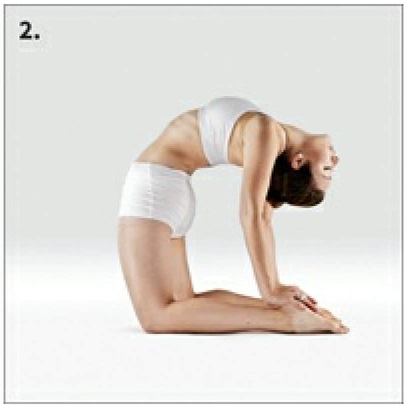
2.

Pose Dedicated to King Nahusha

Nahushasana

(nuh-hu-SHAHS-uh-nuh)

Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

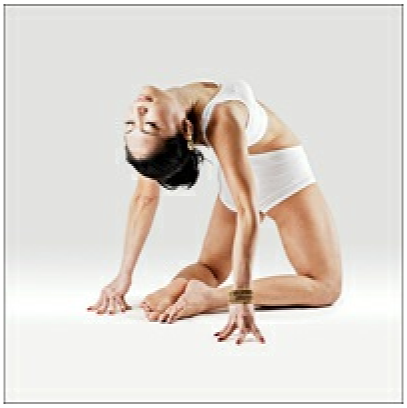

One Hand Graceful Thunderbolt Pose

Eka Hasta Laghuvajrasana

(EY-kuh huh-stuh luh-gu-vuhj-RAHS-uh-nuh)

Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

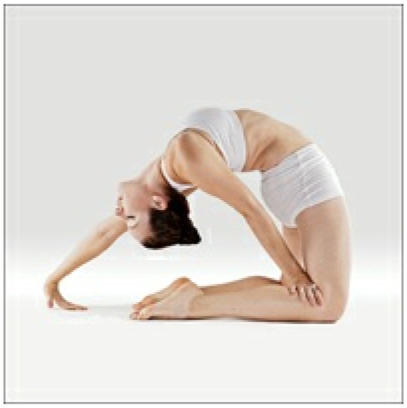

Pigeon Pose

Kapotasana

(kuh-po-TAHS-uh-nuh)

Also Known As: Sage Korakar Pose (Korakarasana), Graceful Thunderbolt Pose (Laghuvajrasana)
Modification: arms over the head, palms to the floor, feet toward the head
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

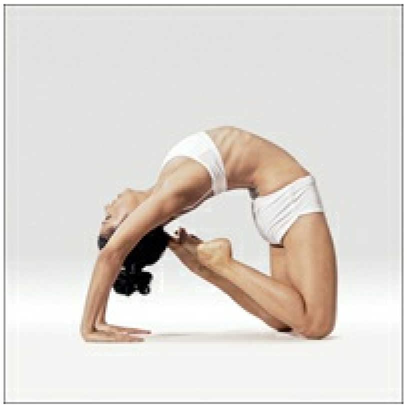

KNEES ON THE FLOOR: ONE ARM EXTENDED

Half Camel Pose

Ardha Ushtrasana

(UHR-duh oosh-TRAHS-uh-nuh)
Also Known As: One Arm Camel Pose (Eka Hasta Ustrasana)
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)

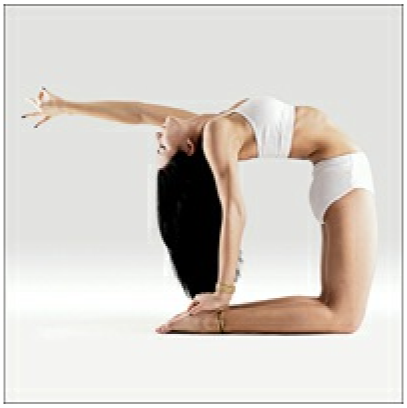

Half Camel Pose

Ardha Ushtrasana

(UHR-duh oosh-TRAHS-uh-nuh)

Modification: arm crossed behind the back to the opposite leg

1. hand to the outside of the hip
2. hand to the inside of the knee

Pose Type: backbend, standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)

● ● ● ● ●

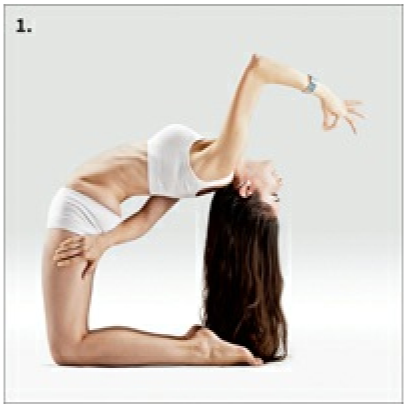

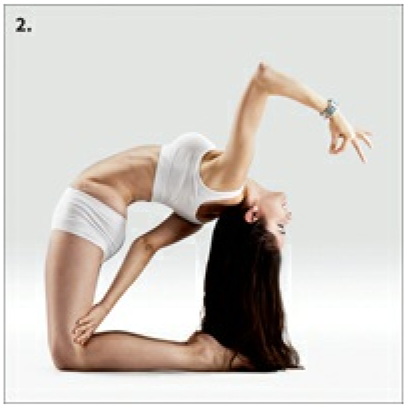

Leg Position of the Pose Dedicated to Garuda in Half Camel Pose

Pada Garudasana in Ardha Ushtrasana

(PUH-duh guh-ru-DAHS-uh-nuh in UHR-duh oosh-TRAHS-uh-nuh}

Modification: one hand to the floor; other arm up, elbow bent

Pose Type: backbend, standing (on the knees)

Drishti Point: Hastagrai or Hastagre (hands), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ● ● ● ●

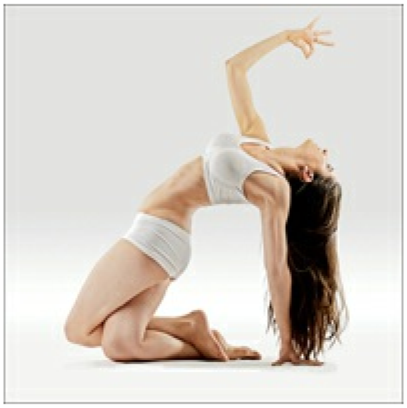

Half Camel Pose

Ardha Ushtrasana

(UHR-duh-oosh-TRAHS-uh-nuh)

Modification: one foot to the armpit on the same side; opposite arm straight up over the head
Pose Type: backbend, standing (on the knees)
Drishti Point: Hastagrai or Hastagre (hands), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

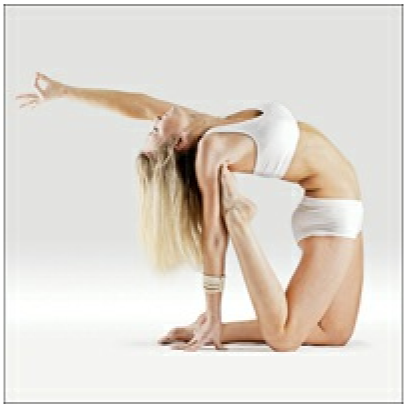

Half Bound Lotus Camel Pose

Ardha Baddha Padma Ushtrasana

(UHR-duh BUH-duh PUHD-muh oosh-TRAHS-uh-nuh)

Modification: free arm straight up over the head
Pose Type: backbend, standing (on the knees), binding
Drishti Point: Hastagrai or Hastagre (hands) or Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

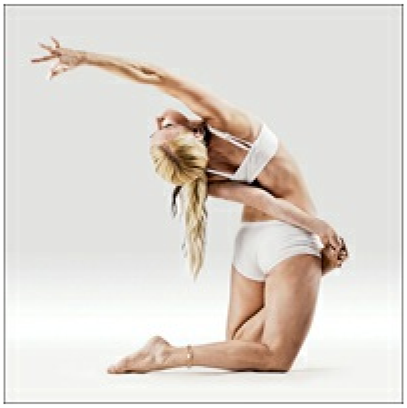

ONE KNEE BENT TOWARD THE HIP: BACKBEND

One-Legged Frog Pose in Camel Pose

Eka Pada Bhekasana in Ushtrasana

(EY-kuh PUH-duh bey-KAHS-uh-nuh in oosh-TRAHS-uh-nuh)

Pose Type: backbend, standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ● ● ● ●

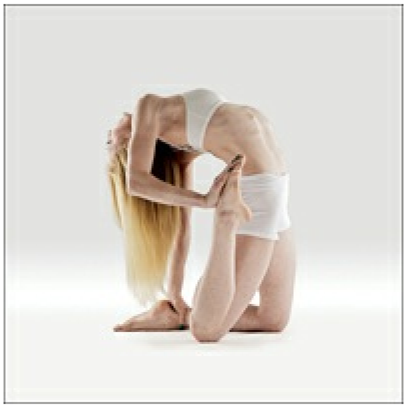

Mermaid Arm Position in Camel Pose

Hasta Naginyasana in Ushtrasana

(HUH-stuh nuh-gin-YAHS-uh-nuh in oosh-TRAHS-uh-nuh)

Pose Type: backbend, standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ● ● ●

BACKBEND ON THE KNEES, HEAD AND SITTING BONES: BED POSE

Bed Pose

Paryankasana

(puhr-yuhng-KAHS-uh-nuh)
Also Known As: Reclined Thunderbolt Pose (Supta Vajrasana)
Modification: 1. head off the floor
2. crown of the head on the floor
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

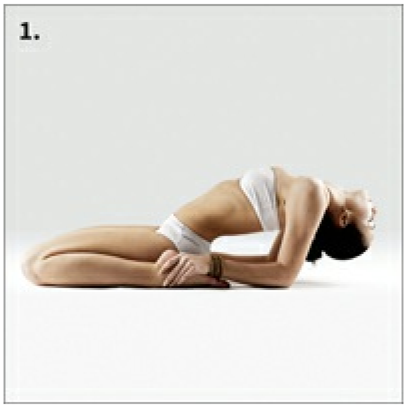
1.

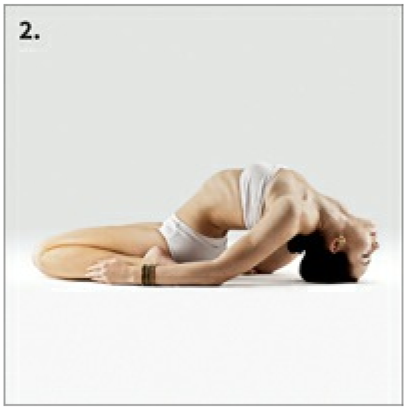
2.

Bed Pose

Paryankasana

(puhr-yuhng-KAHS-uh-nuh)

Also Known As: Reclined Thunderbolt Pose (Supta Vajrasana)
Modification: 1. hands in Anjali Mudra (Hands in Prayer), fingertips pointing to the sky
2. grabbing onto the forearms, arms to the floor up over the head
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

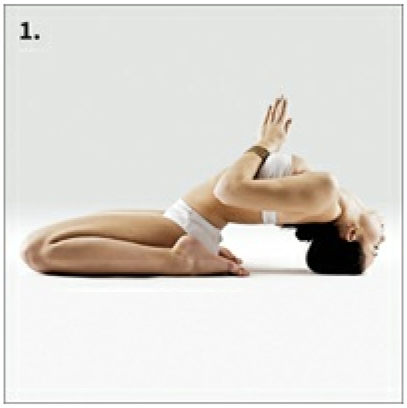
1.

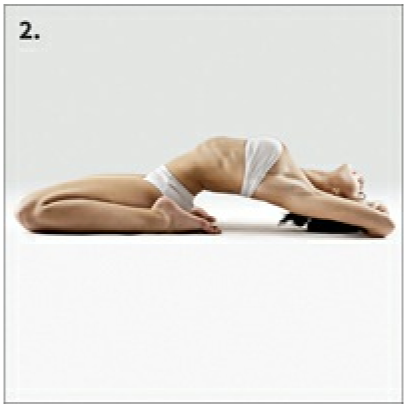
2.

BACKBEND ON THE KNEES AND HEAD: LITTLE THUNDERBOLT POSE

Little Thunderbolt Pose

Laghuvajrasana

(lluh-goo-vuhj-RAHS-uh-nuh)

Modification: 1. grabbing onto the knees, head off the floor
2. forehead to the floor, hands on the quadriceps
3. head to feet, grabbing onto the knees
Pose Type: backbend, standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल र य ह ॐ ॐ

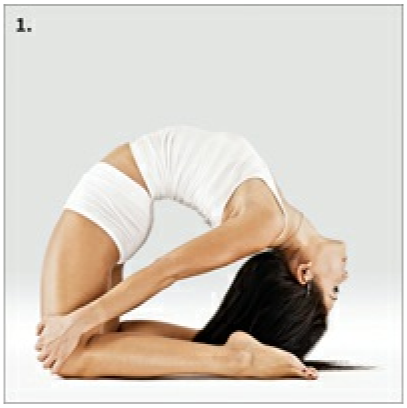

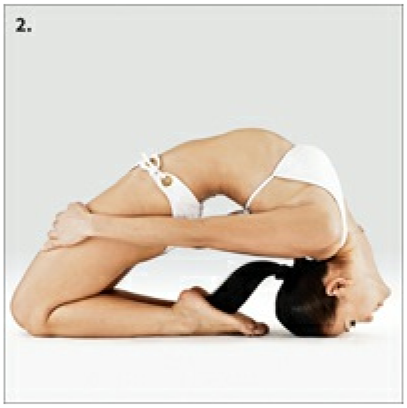

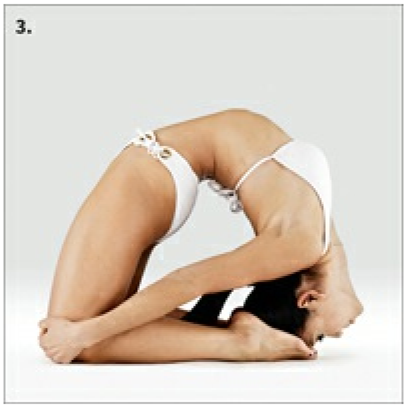

Little Thunderbolt Pose

Laghuvajrasana

(luh-goo-vuhj-RAHS-uh-nuh)

Modification: back of the head on the floor, grabbing onto the knees

Pose Type: backbend, standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल र य ॐ

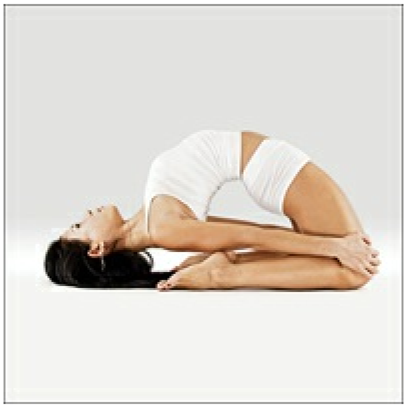

BACKBEND ON THE KNEES AND HEAD: PIGEON POSE

Little Thunderbolt Pose Prep.

Laghuvajrasana Prep.

(luh-goo-vuhj-RAHS-uh-nuh)

Modification: crown of the head to the floor, arms in Headstand 5 (see here) position
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
🔴 🟡 🟢 🔵 🟣

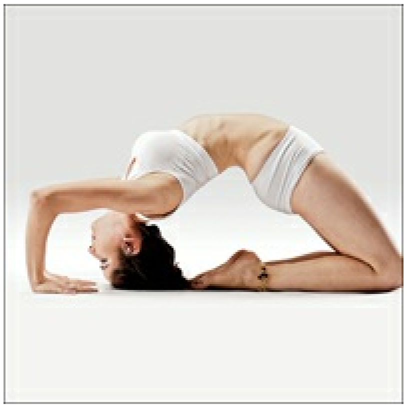

One Hand Pigeon Pose

Eka Hasta Kapotasana

(EY-kuh HUH-stuh kuh-po-TAHS-uh-nuh)

Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
🔴 🟡 🟢 🔵 🔵 🟣

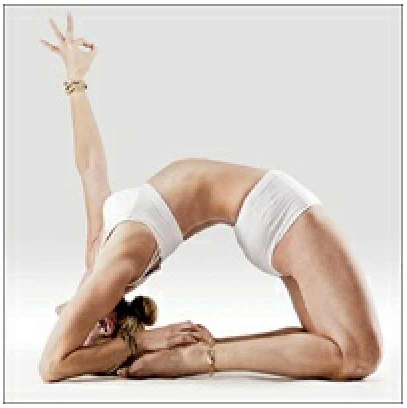

Pigeon Pose

Kapotasana

(kuh-po-TAHS-uh-nuh)

Modification: grabbing onto the ankles
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

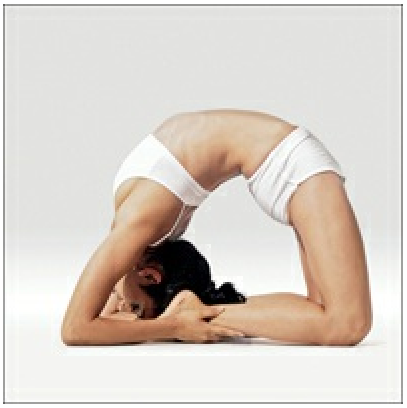

Pigeon Pose

Kapotasana

(kuh-po-TAHS-uh-nuh)

Modification: fingertips toward the knees, palms flat down
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

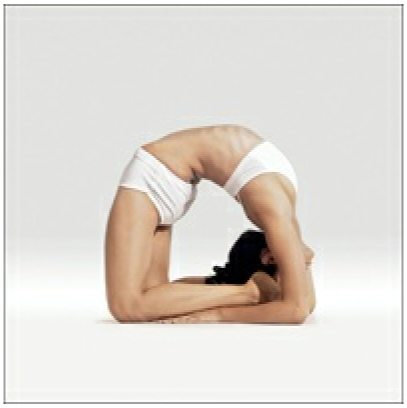

PIGEON POSE: ONE-LEGGED

One-Legged Pigeon Pose Prep.

Eka Pada Kapotasana Prep.

(EY-kuh PUH-duh kuh-po-TAHS-uh-nuh)

Modification: arms over the head palms to the floor, head away from the foot
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ● ●

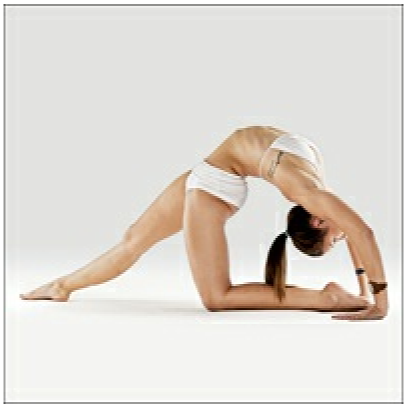

One-Legged Pigeon Pose Prep.

Eka Pada Kapotasana Prep.

(EY-kuh PUH-duh kuh-po-TAHS-uh-nuh)

Modification: 1. knee bent, foot to the floor
2. leg straight, foot to the floor
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ● ● ● ●

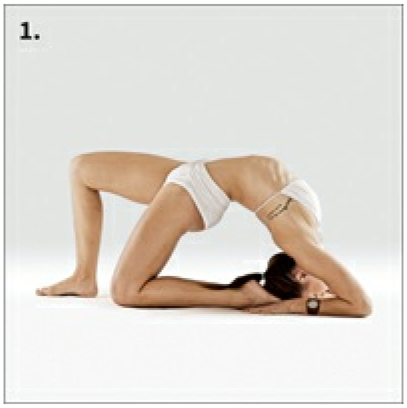
1.

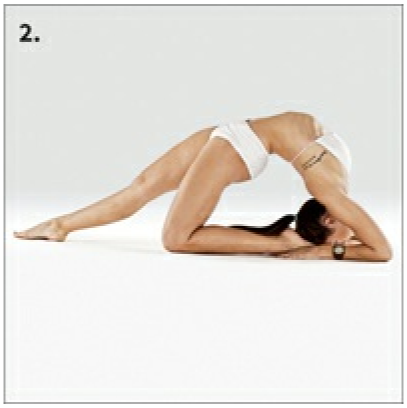
2.

One-Legged Pigeon Pose

Eka Pada Kapotasana

(EY-kuh PUH-duh kuh-po-TAHS-uh-nuh)

Modification: foot to the quadriceps
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ● ● ● ●

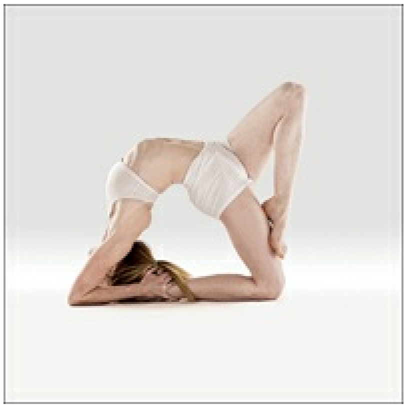

One-Legged Pigeon Pose

Eka Pada Kapotasana

(EY-kuh PUH-duh kuh-po-TAHS-uh-nuh)

Modification: leg straight

Pose Type: backbend, standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

🔴 🟠 🟡 🟢 🔵 🔵 🟣

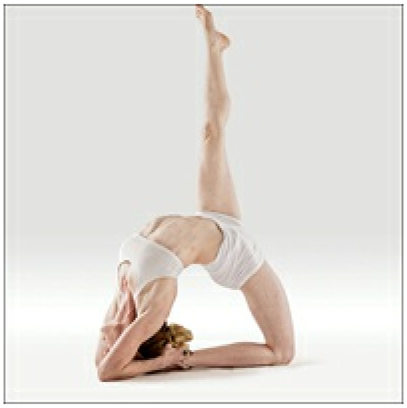

TOES CURLED IN: BACKBEND—SHOULDERS ON THE FLOOR & HEAD ON THE FLOOR

One-Legged Tip Toe Bridge Pose

Eka Pada Prapada Setu Bandhasana

(EY-kuh PUH-duh PRUH-puh-duh SEY-too buhn-DAHS-uh-nuh)

Modification: back of the head on the floor, arms up to the sky

1. knee bent, foot resting on the knee
2. leg straight and extended to the sky

Pose Type: supine, backbend

Drishti Point: Angushtamadhye or Angustha Ma Dyai (thumbs)

🔴 🟠 🟡 🟢

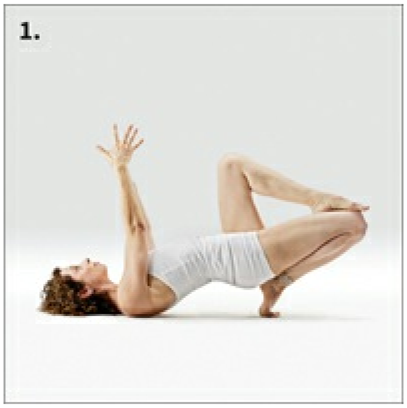

2.

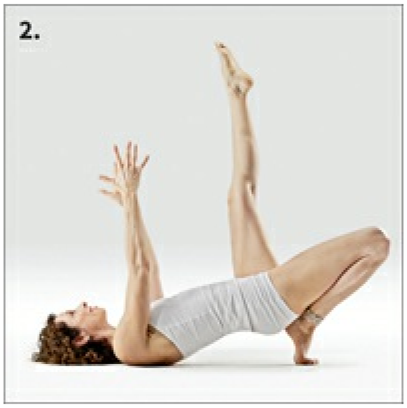

Tip Toe Bridge Pose

Prapada Setu Bandhasana

(PRUH-puh-duh SEY-too buhn-DAHS-uh-nuh)

Also Known As: Tip Toe Bed Pose (Prapada Paryankasana), Big Toe Bridge Pose (Padangushta Setu Bandhasana)
Modification: 1. grabbing onto the shins, elbows to the floor
2. grabbing onto the shins, arms straight
3. hands on the quadriceps, arms straight
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

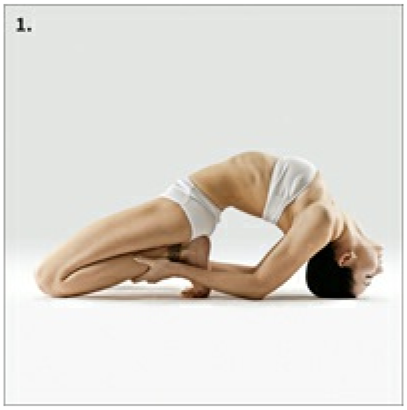

2.

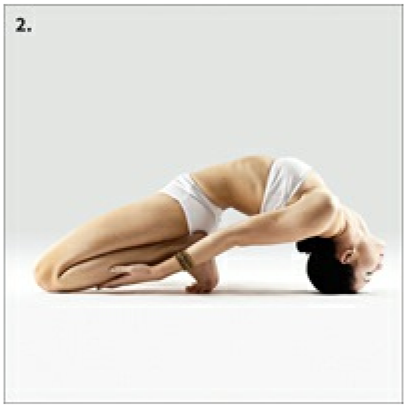

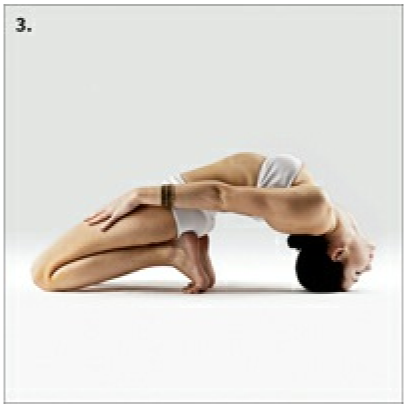

TOES CURLED IN: BACKBEND—HANDS TO THE FLOOR & TOE BALANCE

Thunderbolt Pose

Vajrasana

(vuhj-RAHS-uh-nuh)

Also Known As: Tip Toe Bed Pose Prep. (Prapada Paryankasana Prep.)
Modification: toes curled in, fingertips pointing away from the feet, intense backbend
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

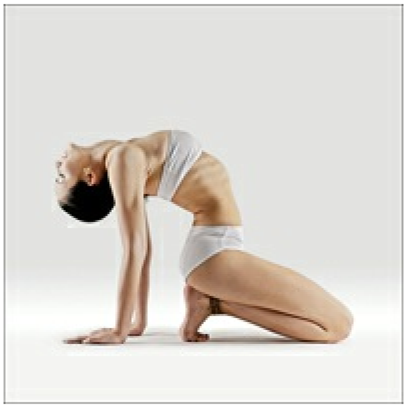

Tip Toe Camel Pose

Prapada Ushtrasana

(PRUH-puh-duh oosh-TRAHS-uh-nuh)

Modification: palms flat to the floor, fingertips pointing to the front
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

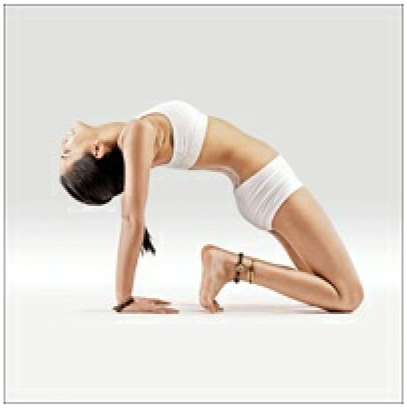

Unsupported Tip Toe Camel Pose

Niralamba Prapada Ushtrasana

(neer-AH-luhm-buh PRUH-puh-duh oosh-TRAHS-uh-nuh)

Modification: knees lifted off the floor, hands in Anjali Mudra (Hands in Prayer)
Pose Type: standing, balance, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

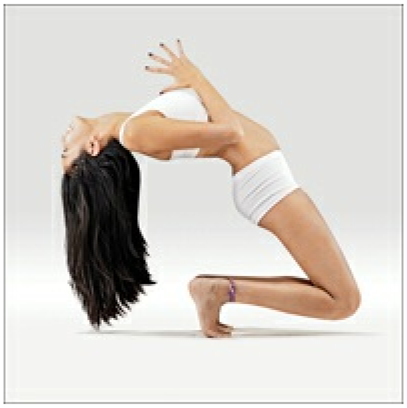

TOES CURLED IN: BACKBEND—GRABBING ONTO THE HEELS & KNEES, HANDS ON THE FLOOR & HEAD TO THE HEELS

Tip Toe Camel Pose

Prapada Ushtrasana

(PRUH-puh-duh oosh-TRAHS-uh-nuh)

Modification: hands in Anjali Mudra (Hands in Prayer)
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

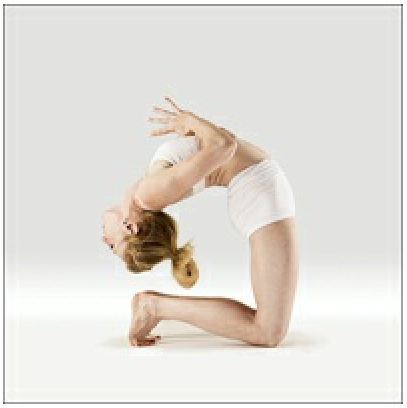

Half Tip Toe Camel Pose

Ardha Prapada Ushtrasana

(UHR-duh PRUH-puh-duh oosh-TRAHS-uh-nuh)

Pose Type: backbend, standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)

लं रं यं हं ॐ

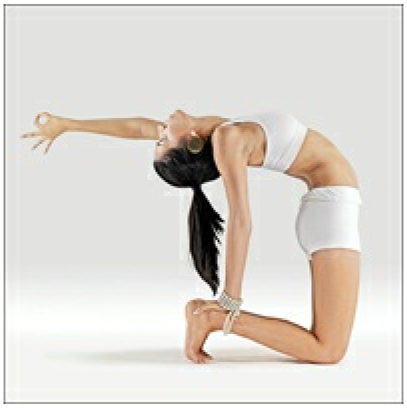

Tip Toe Camel Pose

Prapada Ushtrasana

(PRUH-puh-duh oosh-TRAHS-uh-nuh)

Pose Type: backbend, standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

लं रं यं हं ॐ

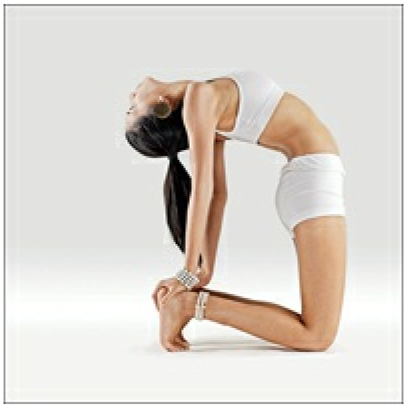

Tip Toe Camel Pose

Prapada Ushtrasana

(PRUH-puh-duh oosh-TRAHS-uh-nuh)

Modification: hands on the knees
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

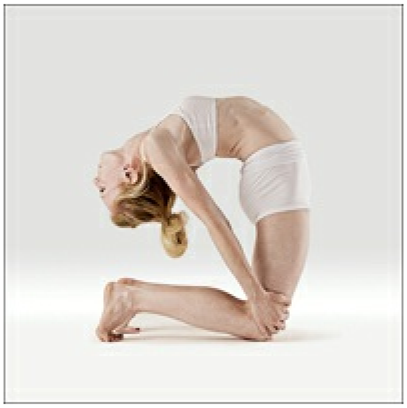

Tip Toe Pigeon Pose

Prapada Kapotasana

(PRUH-puh-duh kuh-po-TAHS-uh-nuh)

Also Known As: Big Toe Pigeon Pose (Padangushta Kapotasana)
Modification: arms straight up over the head, palms to the floor
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)

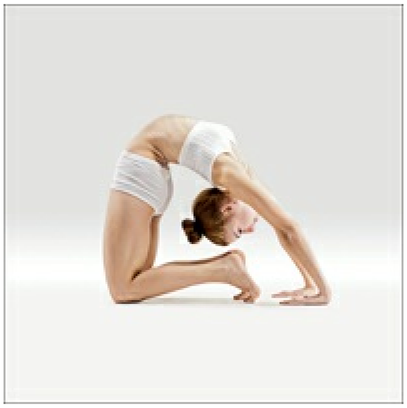

Tip Toe Pigeon Pose

Prapada Kapotasana

(PRUH-puh-duh kuh-po-TAHS-uh-nuh)

Also Known As: Big Toe Pigeon Pose (Padangustha Kapotasana), Little Wheel Pose (Laghu Chakrasana)
Modification: arms up over the head, elbows bent, palms to the floor, forehead to the heels
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

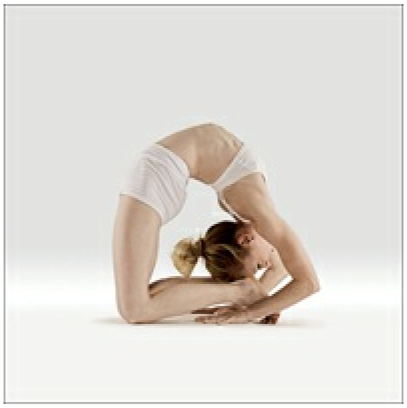

TOES CURLED IN: BACKBEND—ARMS STRAIGHT, PALMS TOGETHER

Raised Bound Hands in Tip Toe Camel Pose

Urdhva Baddha Hastasana in Prapada Ushtrasana

(OORD-vuh BUH-duh huh-STAHS-uh-nuh in PRUH-puh-duh oosh-TRAHS-uh-nuh)

Modification: hips pushing back
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)

Tip Toe Pigeon Pose

Prapada Kapotasana

(PRUH-puh-duh kuh-po-TAHS-uh-nuh)

Also Known As: Big Toe Pigeon Pose (Padangushta Kapotasana), Full Camel Pose (Purna Ushtrasana), or Little Wheel Pose (Laghu Chakrasana)
Modification: 1. arms extended to the sky
2. arms parallel to the floor
Pose Type: backbend, standing (on the knees)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

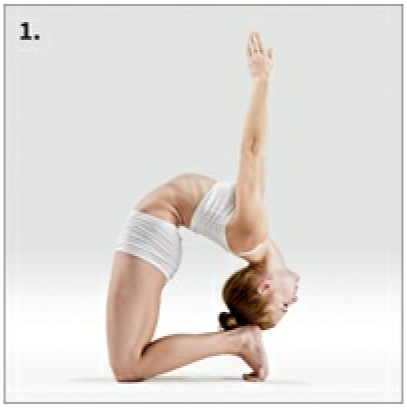

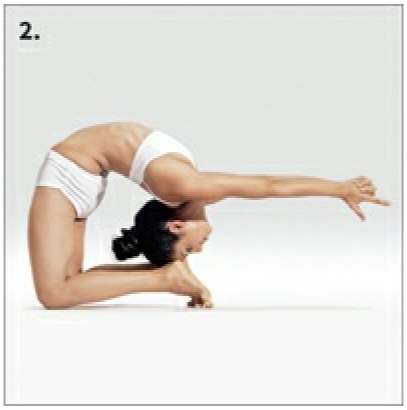

BACKBEND: HEADSTAND 5 ARM POSITION

Upward Bow Pose in Headstand 5

Urdhva Dhanurasana in Shirshasana 5

(OORD-vuh duh-nur-AHS-uh-nuh in sheer-SHAS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

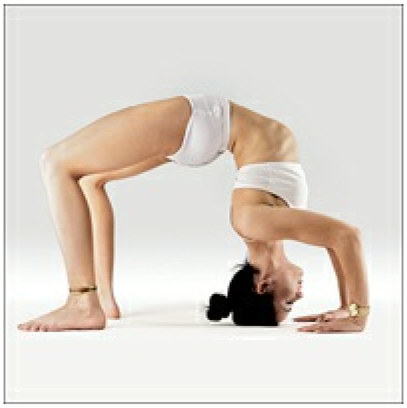

Upward Bow Pose with Leg Position of the Pose Dedicated to Garuda in Headstand 5

Urdhva Dhanurasana Pada Garudasana in Shirshasana 5

(OORD-vuh duh-nur-AHS-uh-nuh PUH-duh guh-ru-DAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

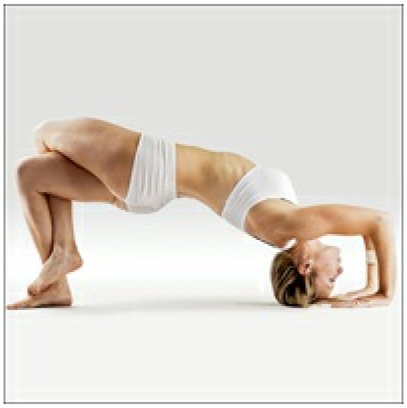

Upward Tip Toe Bow Pose in One Hand Headstand 5

Urdhva Prapada Dhanurasana in Eka Hasta Shirshasana 5

(OORD-vuh PRUH-puh-duh duh-nur-AHS-uh-nuh in EY-kuh HUH-stuh sheer-SHAHS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

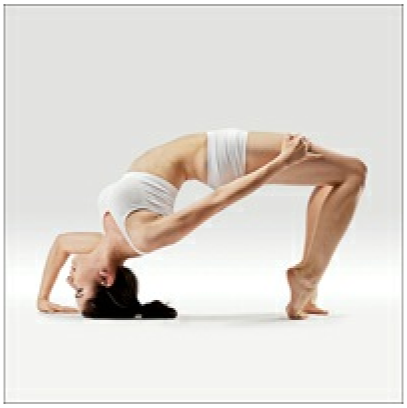

Inverted Tip Toe Bow Pose

Viparita Prapada Dhanurasana

(vi-puh-REE-tuh PRUH-puh-duh duh-nur-AHS-uh-nuh)

Also Known As: Headstand Bow Pose (Shirsha Dhanurasana)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

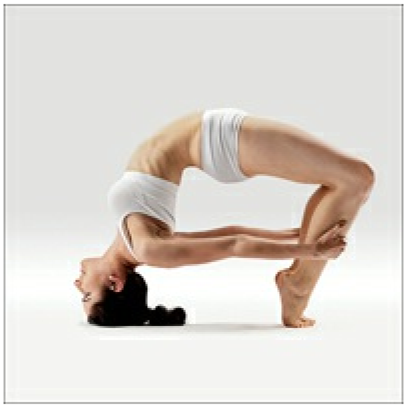

BACKBEND: HEADSTAND 1 ARM POSITION—KNEE/KNEES BENT

Upward Bow Pose in Headstand 1

Urdhva Dhanurasana in Shirshasana 1

(OORD-vuh duh-nur-AHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

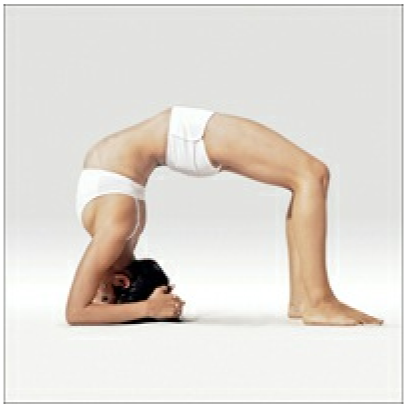

Upward Tip Toe Bow Pose in Headstand 1

Urdhva Prapada Dhanurasana in Shirshasana 1

(OORD-vuh PRUH-puh-duh duh-noor-AHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

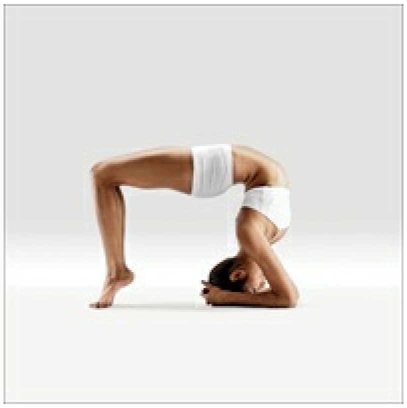

One-Legged Inverted Staff Pose

Eka Pada Viparita Dandasana

(EY-kuh PUH-duh vi-puh-REE-tuh duhn-DAHS-uh-nuh)

Modification: knee of the bottom leg bent
Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

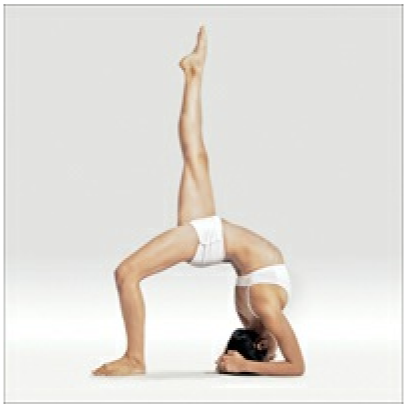

BACKBEND: HEADSTAND 1 ARM POSITION—LEGS STRAIGHT

One-Legged Inverted Staff Pose

Eka Pada Viparita Dandasana

(EY-kuh PUH-duh vi-puh-REE-tuh duhn-DAHS-uh-nuh)

Also Known As: Inverted Staff Pose B (Viparita Dandasana B)
Modification: both legs straight
Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

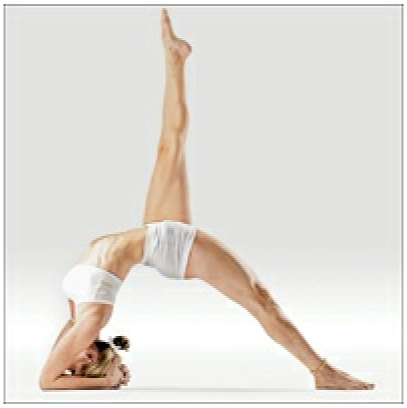

Two-Legged Inverted Staff Pose

Dwi Pada Viparita Dandasana

(DWI-puh-duh vi-puh-REE-tuh duhn-DAHS-uh-nuh)

Also Known As: Inverted Staff Pose A (Viparita Dandasana A)
Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

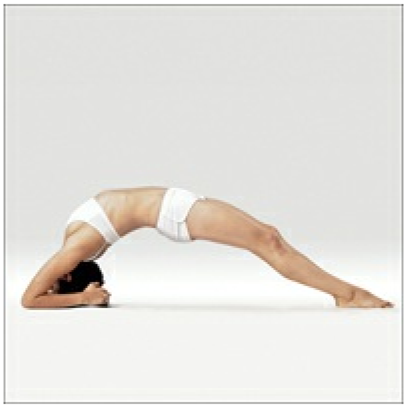

BACKBEND: FOREARMS ON THE FLOOR, BOUND HANDS

Bound Arms Tip Toe Bound Wheel Pose

Baddha Hasta Prapada Chakra Bandhasana

(BUH-duh HUH-stuh PRUH-puh-duh CHUHK-ruh buhn-DAHS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

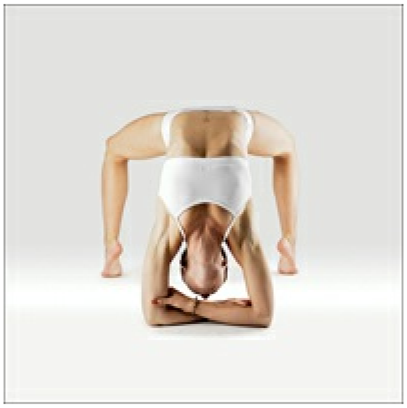

Bound Arms Both Legs Inverted Staff Pose

Baddha Hasta Dwi Pada Viparita Dandasana

(BUH-duh HUH-stuh DWI-puh-duh vi-puh-REE-tuh duhn-DAH-suh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

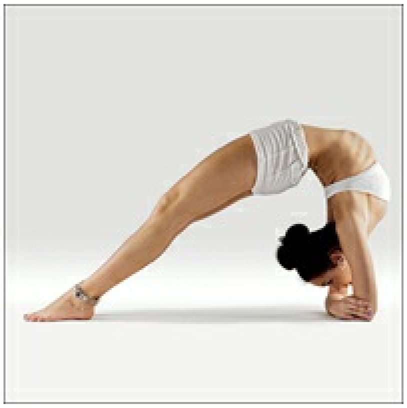

BACKBEND: FOREARMS ON THE FLOOR, HEELS DOWN

Bound Wheel Pose

Chakra Bandhasana

(CHUHK-ruh buhn-DAHS-uh-nuh)

Modification: forearms to the floor, heels down, fingertips touching the heels
Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

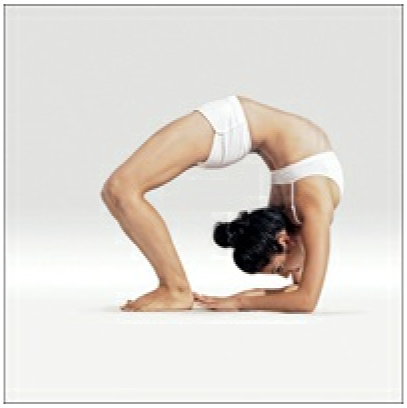

Bound Wheel Pose

Chakra Bandhasana

(CHUHK-ruh buhn-DAHS-uh-nuh)

Also Known As: Bound Wheel Pose (Bandha Chakrasana)
Modification: grabbing onto ankles
1. looking straight ahead
2. head rolling back
Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

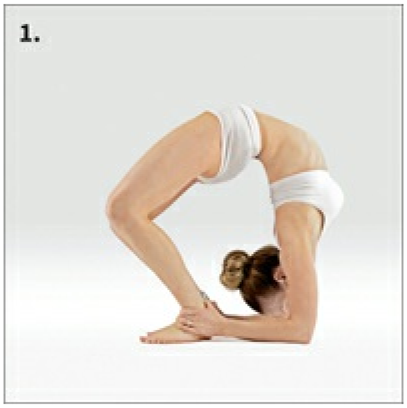

2.

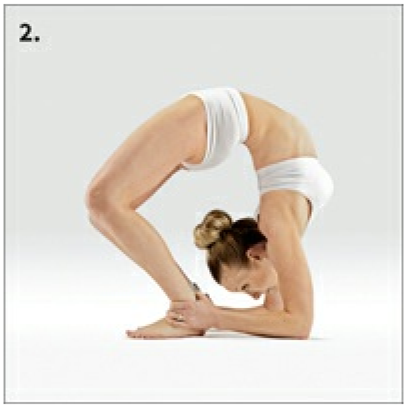

BACKBEND: FOREARMS ON THE FLOOR, HEELS LIFTED

Tip Toe Bound Wheel Pose

Prapada Chakra Bandhasana

(PRUH-puh-duh CHUCK-ruh buhn-DAHS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

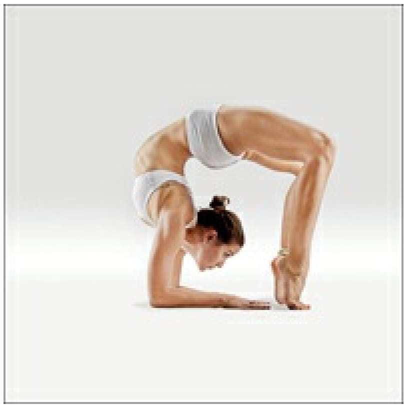

Uneven Tip Toe Bound Wheel Pose

Vishama Prapada Chakra Bandhasana

(VI-shuh-muh PRUH-puh-duh CHUCK-ruh buhn-DAHS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल र य ह ॐ

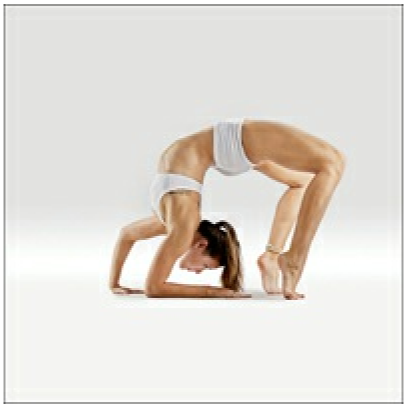

BACKBEND: FOREARMS ON THE FLOOR, ONE-LEGGED

One-Legged Inverted Staff Pose

Eka Pada Viparita Dandasana

(EY-kuh PUH-duh vi-puh-REE-tuh duhn-DAHS-uh-nuh)

Modification: forearms to the floor, head off the floor

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

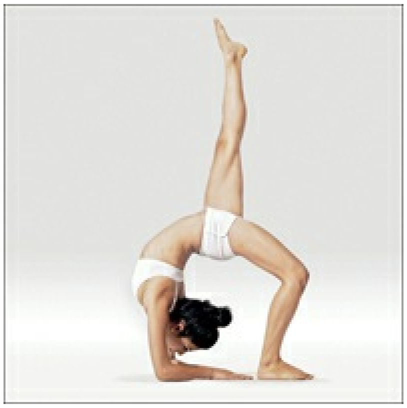

Bound Foot One-Legged Inverted Staff Pose

Baddha Pada Eka Pada Viparita Dandasana

(BUH-duh PUH-duh EY-kuh PUH-duh vi-puh-REE-tuh duhn-DAHS-uh-nuh)

Modification: forearms to the floor, head off the floor

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

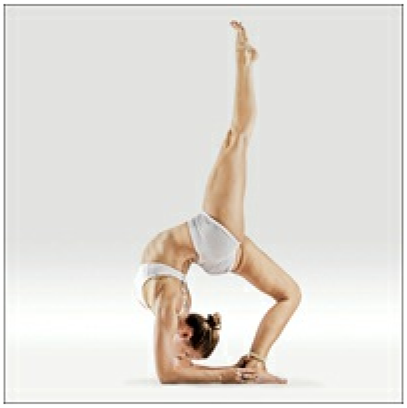

Tip Toe Bound Foot One-Legged Inverted Staff Pose

Prapada Baddha Pada Eka Pada Viparita Dandasana

(PRUH-puh-duh BUH-duh PUH-duh EY-kuh PUH-duh vi-puh-REE-tuh duhn-DAHS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

How to Perform the Pose:

1. Begin by lying flat on your back. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing. Inhale and bring your hands under your shoulders with elbows bent and close to your head, with the fingertips pointing away from your head.
2. Exhale and bend your knees, keep your feet flat on the floor and slide your heels toward your sitting bones. Keep your feet in line with your sitting bones and parallel to each other with your toes pointing straight ahead.
3. Exhale as you lift your sitting bones and your back off the floor. Place the crown of your head onto the floor, balancing on your hands, head, and feet.
4. Exhale as you slide your right palm in the direction of your feet until your forearm is flat on the floor. On the next exhale, slide your left palm to meet your right. Interlock your fingers at the back of your head.
5. Exhale and lift your head off the floor.
6. On the next exhale, walk your feet toward your hands. Grab onto your left ankle with both hands.
7. Exhale as you lift your right foot off the floor, keeping your right leg strong and straight with toes pointing up to the sky.
8. Inhale and lift your left heel, coming onto the toes of the left foot, while holding onto the left toes with both hands. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch on the right side.
9. Inhale and lower the left heel to the floor, followed by the right foot. Exhale and grab onto the right ankle with both hands. On the following exhale, lift your left foot off the floor, keeping your left leg strong and straight with toes pointing up to the sky.
10. Inhale and lift your right heel, coming onto the toes of the right foot while holding onto the right toes with both hands. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch on the left side.
11. Inhale and lower your right heel to the floor, followed by your left foot. Exhale and lower your head to the floor. Exhale and tuck your chin to your chest and lower your spine to the floor, coming

out of the pose.

Modification: forearms to the floor, head off the floor

prapada = tip of the feet

baddha = bound

pada = foot or leg

eka = one

pada = foot or leg

viparita = inverted

danda = stick or staff

Foot to the Head One-Legged Inverted Staff Pose

Shirsha Pada Eka Pada Viparita Dandasana

(SHEER-shuh PUH-duh EY-kuh PUH-duh vi-puh-REE-tuh duhn-DAHS-uh-nuh)

Modification: forearms to the floor, head off the floor
Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

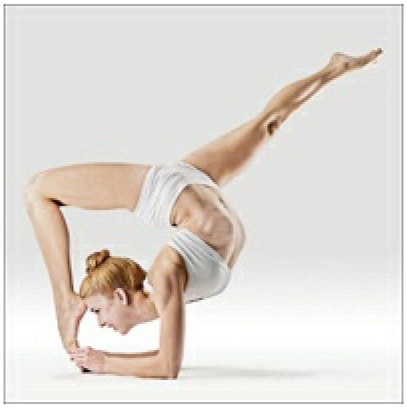

UPWARD BOW POSE

Upward Bow Pose

Urdhva Dhanurasana

(OORD-vuh duh-nur-AHS-uh-nuh)

Modification: heels down
Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

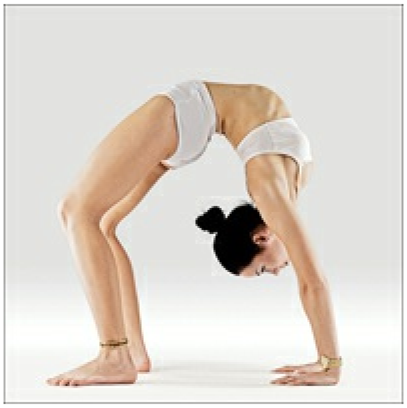

Tip Toe Upward Bow Pose

Prapada Urdhva Dhanurasana

(PRUH-puh-duh OORD-vuh duh-nur-AHS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

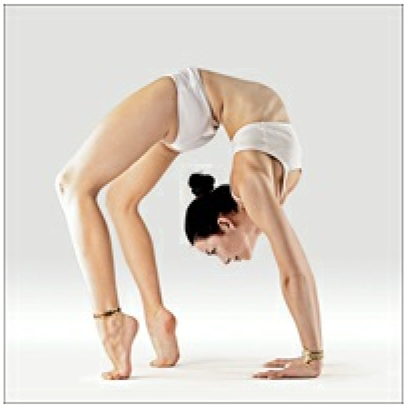

Elevated Both Legs Inverted Staff Pose

Utthita Dwi Pada Viparita Dandasana

(UT-ti-tuh DWI-puh-duh vi-puh-REE-tuh duhn-DAHS-uh-nuh)

Also Known As: Upward Bow Pose (Urdhva Dhanurasana)

Modification: head rolling back

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

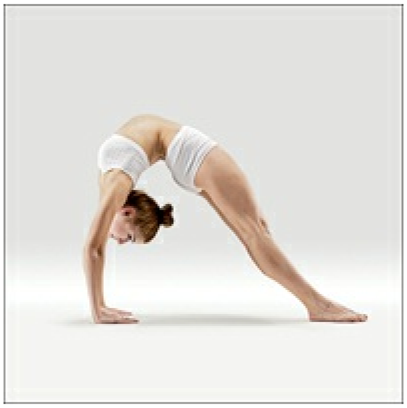

Tip Toe One-Legged Upward Bow Pose

Prapada Eka Pada Urdhva Dhanurasana

(PRUH-puh-duh EY-kuh PUH-duh OORD-vuh duh-nur-AHS-uh-nuh)

Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

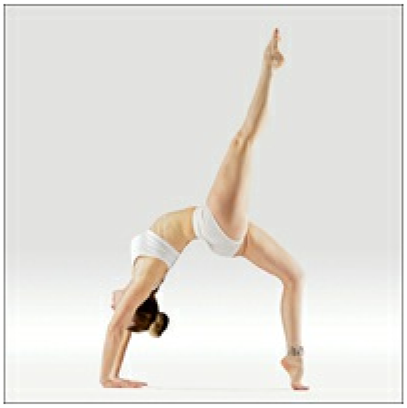

One-Legged Upward Bow Pose

Eka Pada Urdhva Dhanurasana

(EY-kuh PUH-duh OORD-vuh duh-nur-AHS-uh-nuh)

Also Known As: Two Limb Upward Bow Pose (Dwi Anga Urdhva Dhanurasana)
Modification: one hand to the thigh
Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

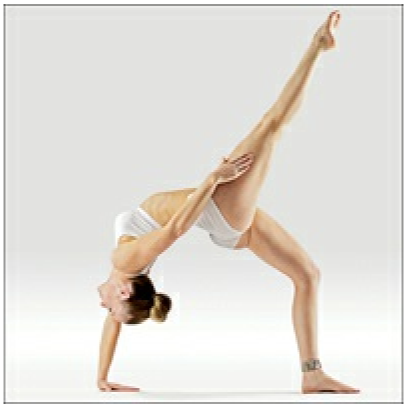

Tip Toe One Hand Upward Bow Pose

Prapada Eka Hasta Urdhva Dhanurasana

(PRUH-puh-duh EY-kuh HUH-stuh OORD-vuh duh-nur-AHS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

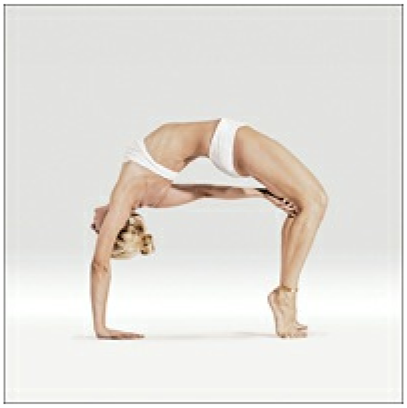

WILD THING POSE

Wild Thing Pose

Chamatkarasana

(kuh-muht-kar-AHS-uh-nuh)

Modification: forearm to the floor
Pose Type: backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Hastagrai or Hastagre (hands)
See the glossary for a more precise translation of chamatkarasana.

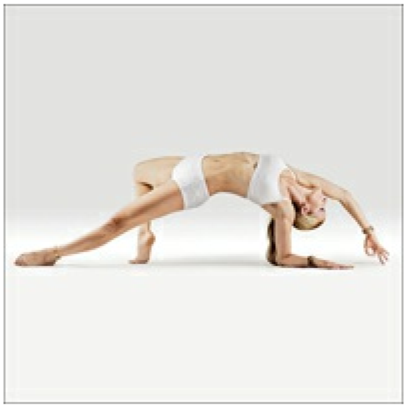

Wild Thing Pose

Chamatkarasana

(kuh-muht-kar-AHS-uh-nuh)

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Hastagrai or Hastagre (hands)

ल र य ह ॐ

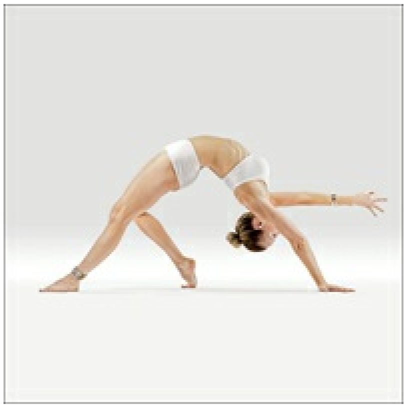

Wild Thing Pose

Chamatkarasana

(kuh-muht-kar-AHS-uh-nuh)

Modification: foot to inside of the thigh
Pose Type: backbend
Drishti Point: Hastagrai or Hastagre (hands)

ल व र य

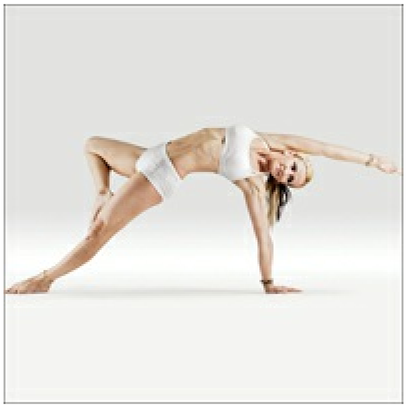

Wild Thing Pose

Chamatkarasana

(kuh-muht-kar-AHS-uh-nuh)

Modification: grabbing the ankle with overhead grip on the same side
Pose Type: backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Hastagrai or Hastagre (hands)

Wild Thing Pose

Chamatkarasana

(kuh-muht-kar-AHS-uh-nuh)

Modification: grabbing the ankle with overhead grip on the same side; other foot off the floor—knee bent

Pose Type: backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Hastagrai or

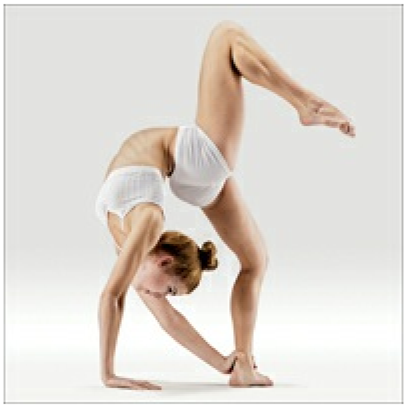

Arm Balances

LOTUS POSE

PENDANT POSE

ONE LEG OVER THE SHOULDER

CRANE POSE

FIREFLY POSE

SCISSOR LEGS

GARUDA LEGS

PEACOCK AND SWAN POSES

Half Fire Log Celibate Pose

Ardha Agnistambha Brahmacharyasana

(UHR_duh uhg-ni-STUHM-buh bruh-muh-chahr-YAHS-uh-nuh)

Pose Type: core, arm balance
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

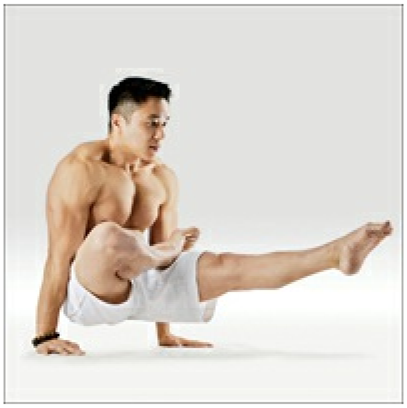

One Leg Over Shoulder Pose

Eka Hasta Bhujasana

(EY-kuh HUH-stuh buj-AHS-uh-nuh)

Also Known As: Comfortable Bird Pose (Sukha Chakorasana)
Pose Type: core, arm balance
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

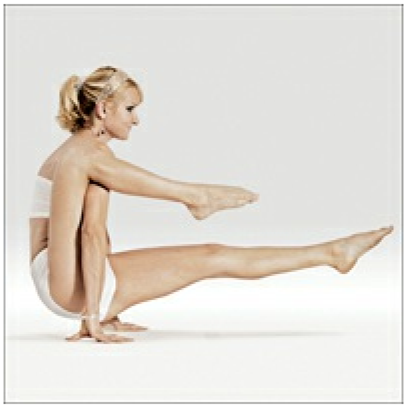

Moonbird Pose

Chakorasana

(chuh-kor-AHS-uh-nuh)

Pose Type: core, arm balance

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

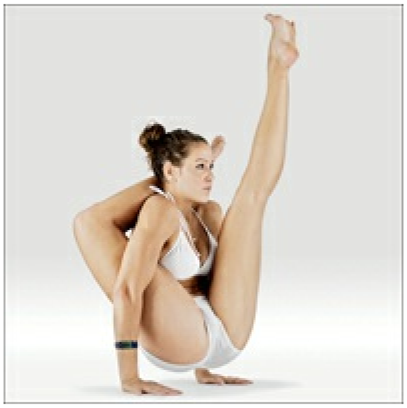

ARM BALANCE: LOTUS

Rooster Pose

Kukkutasana

(ku-ku-TAHS-uh-nuh)

Pose Type: arm balance, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

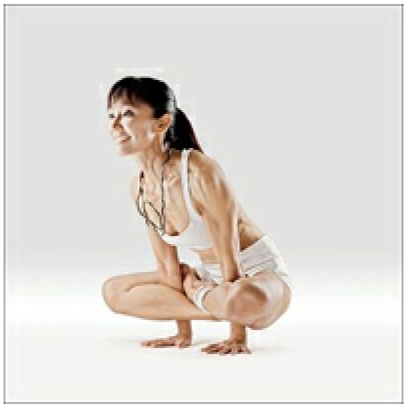

Scales Pose

Tolasana

(to-LAHS-uh-nuh)

Modification: 1. fingertips pointing to the back, thumbs pointing to the front
2. palms flat on the floor, fingertips pointing to the front
Pose Type: arm balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

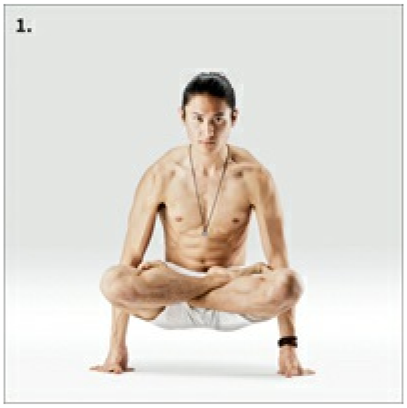

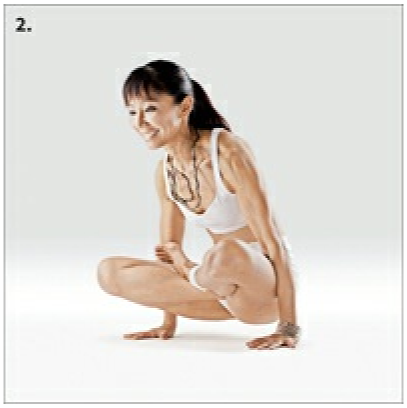

Pose Dedicated to Sage Galava

Galavasana

(gah-luh-VAHS-uh-nuh)
Modification: 1. hips low
2. hips high
Pose Type: arm balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

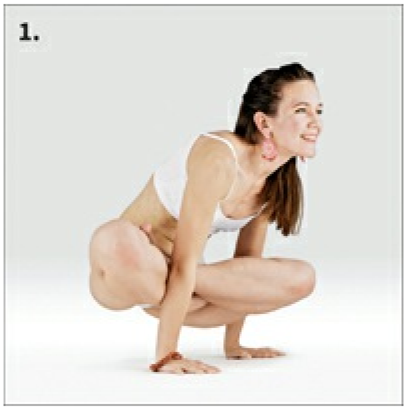

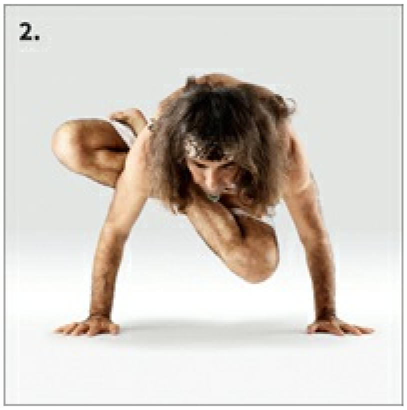

Upward Rooster Pose

Urdhva Kukkutasana

(OORD-vuh ku-ku-TAHS-uh-nuh)

Modification: 1. hips at shoulder height
2. hips higher than the shoulders
Pose Type: arm balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)

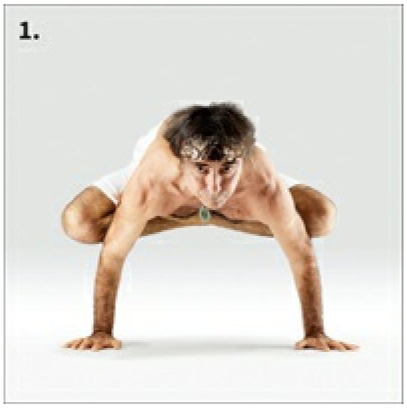

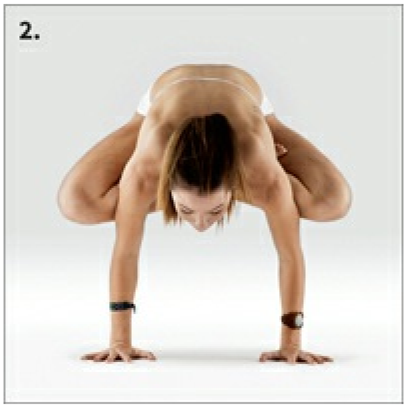
2.

Side Rooster Pose

Parshva Kukkutasana

(PAHRSH-vuh ku-ku-TAHS-uh-nuh)

Also Known As: Revolved Rooster Pose (Parivritta Kukkutasana), Wounded Rooster Pose (Pungu Kukkutasana)
Modification: 1. elbows bent
2. arms straight
Pose Type: arm balance, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose)

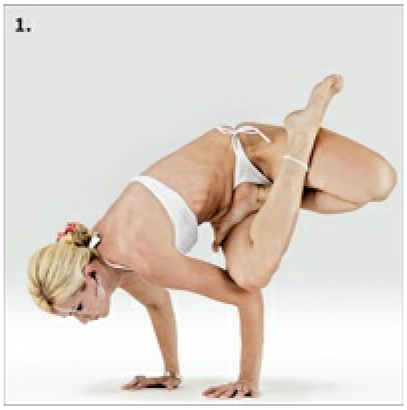
1.

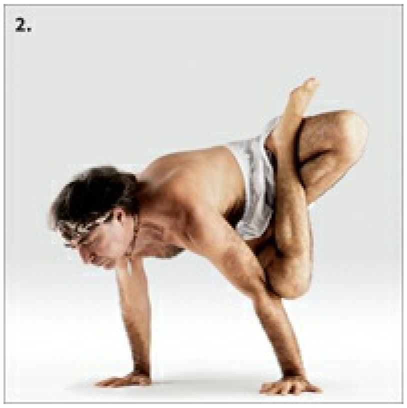
2.

Pendant Pose

Lolasana

(lo-LAHS-uh-nuh)

Modification: ankles crossed
Pose Type: arm balance, forward bend, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs) or Nasagrai or Nasagre (nose)

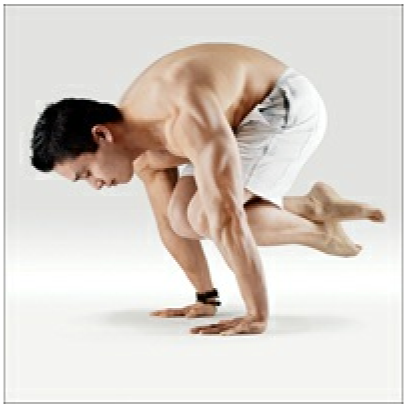

Pendant Pose

Lolasana

(lo-LAHS-uh-nuh)

Pose Type: arm balance, forward bend, core

Drishti Point: Angusthamadhye or Angustha Ma Dyai (thumbs) or Nasagrai or Nasagre (nose)

How to Perform the Pose:

1. Begin by sitting on the floor with both your legs straight out in front of you. Keep your fingertips on the floor on the sides of your hips. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale and lean forward, lifting your sitting bones off the floor. Bend your right knee, sliding your foot back. On the next exhale, bend your left knee, sliding your left foot back to meet your right. Your left shin should end up on top of your right calf muscle. Sit on your heels with ankles crossed under your sitting bones.
3. Inhale and rock forward. Exhale, engage your core, pull your quadriceps toward your chest, and lift your knees and feet off the floor, balancing on your fingertips. Make sure your arms are strong and straight and your shoulders are on top of your fingertips.
4. Aim to hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the pose.
5. Inhale and lower your feet and knees to the floor. Exhale and bring both legs straight out in front of you. Repeat on the opposite side.

Modification: on the fingertips; ankles crossed

lola = pendant

Pendant Pose

Lolasana

(lo-LAHS-uh-nuh)

Modification: fingertips pointing to the back, thumbs pointing to the front
Pose Type: arm balance, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

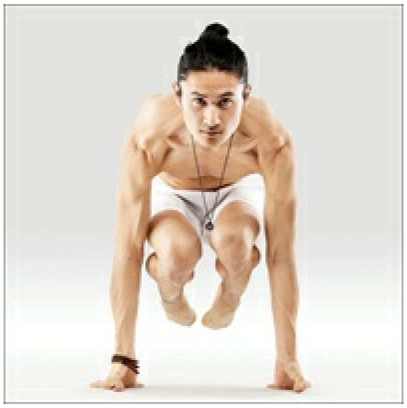

Leg Position of Cow Face Pose in Pendant Pose

Pada Gomukhasana in Lolasana

(PUH-duh go-muk-AHS-uh-nuh in lo-LAHS-uh-nuh)

Modification: fingertips pointing to the back, thumbs pointing to the front
Pose Type: arm balance, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

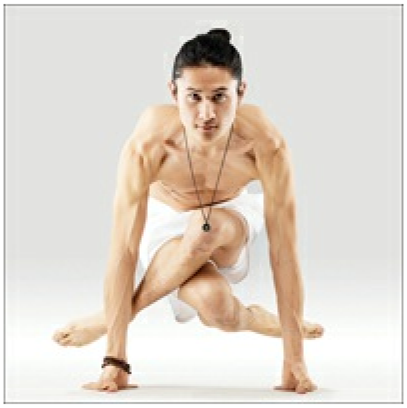

Pose Dedicated to Viranchi (Brahma) 1 Prep.

Viranchyasana 1 Prep.

(vir-uhn-CHYAHS-uh-nuh)

Modification: back of the knee toward the shoulder
Pose Type: arm balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं रं

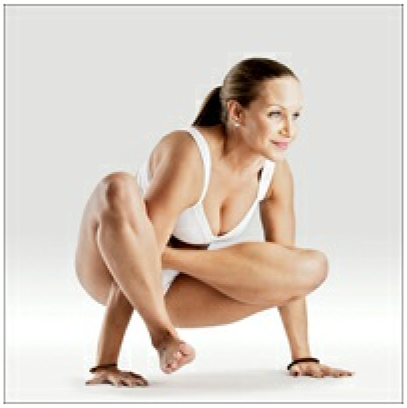

One-Legged Crane 2 Prep.

Eka Pada Bakasana 2 Prep.

(EY-kuh PUH-duh buh-KAHS-uh-nuh)

Modification: both knees bent, back of one knee over the shoulder, other knee toward the chest
Pose Type: arm balance, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
लं वं

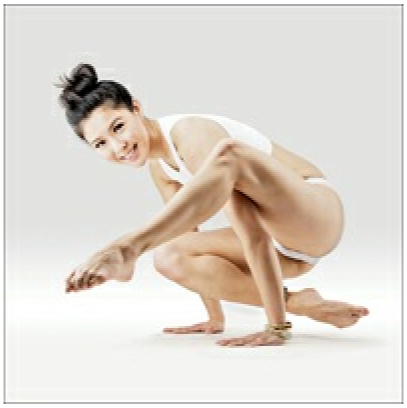

One Foot Behind Head Crane Pose

Eka Pada Shirsha Bakasana

(EY-kuh PUH-duh SHEER-shuh buh-KAHS-uh-nuh)

Modification: 1. hips at shoulder height
2. hips at elbow height, head lifted
Pose Type: arm balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

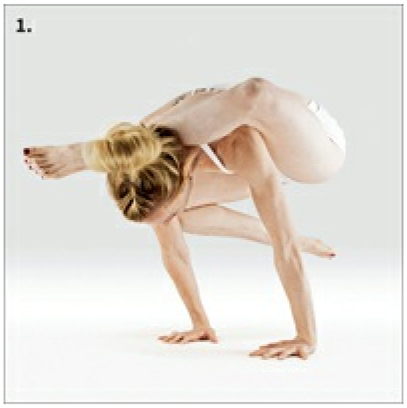
1.

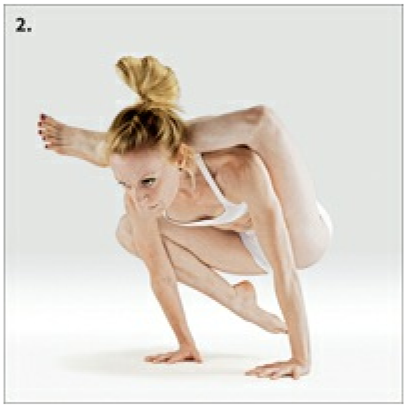
2.

ARM BALANCE: CRANE POSE

Crane Pose

Bakasana

(buh-KAHS-uh-nuh)

Also Known As: Crow Pose (Kakasana)
Modification: elbows bent
Pose Type: arm balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

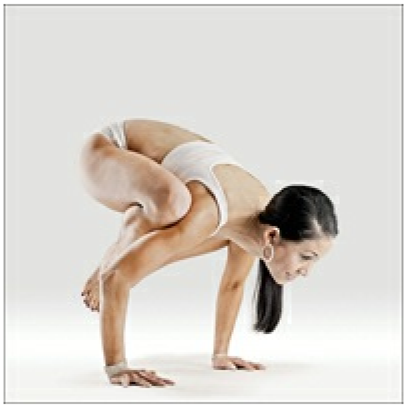

Crane Pose

Bakasana

(buh-KAHS-uh-nuh)

Modification: 1. knees off the triceps
2. knees on the triceps
Pose Type: arm balance, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Nasagrai or Nasagre (nose)

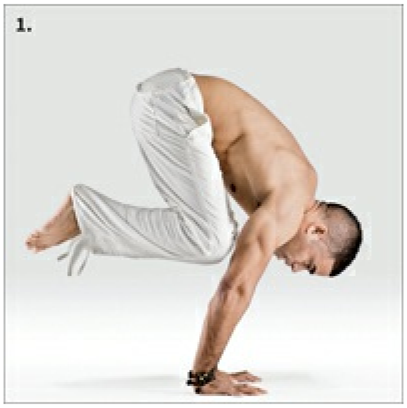

1.

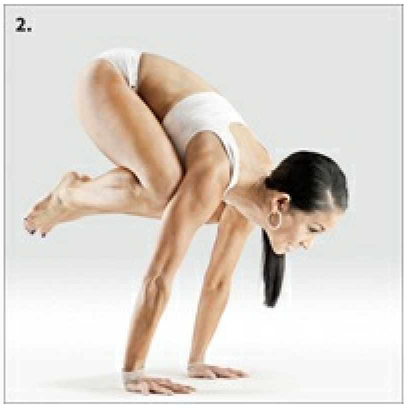

2.

ARM BALANCE: SHOULDER PRESSURE POSE & FIREFLY POSE

Two-Handed Arm Balance

Dwi Hasta Bhujasana

(DWI-huh-stuh buj-AHS-uh-nuh)

Modification: 1. feet off the floor
2. toes touching the floor
Pose Type: arm balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Padayoragrai or Padayoragre (toes/feet)

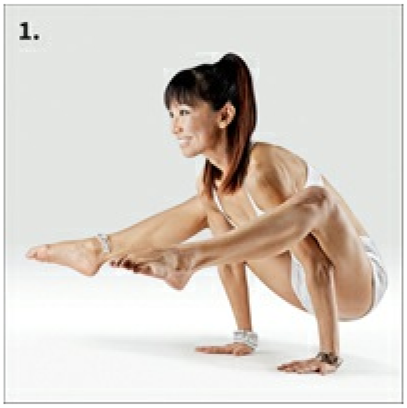

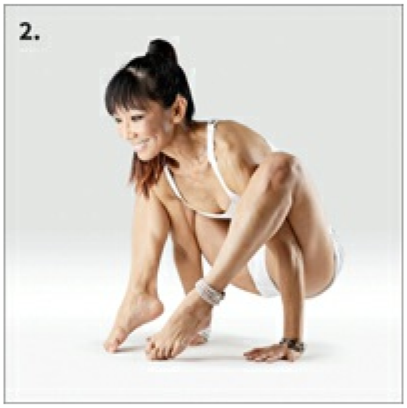

Shoulder Pressure Pose

Bhujapidasana

(buj-uh-peed-AHS-uh-nuh)

Also Known As: Shoulder Pressure Pose (Bhujapidasana A)
Modification: head off the floor
Pose Type: arm balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Padayoragrai or Padayoragre (toes/feet)

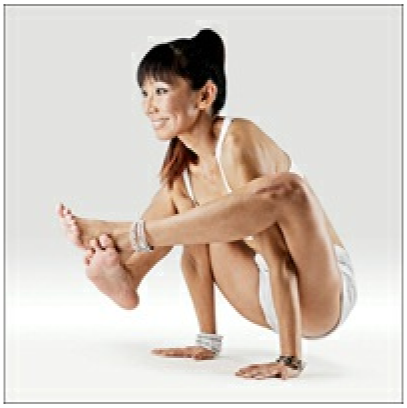

Firefly Pose 1

Tittibhasana 1

(ti-ti-BAHS-uh-nuh)

Also Known As: Firefly Pose A (Tittibhasana A)
Modification: elbows bent, legs on top of the shoulders
Pose Type: arm balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), or Nasagrai or Nasagre (nose)

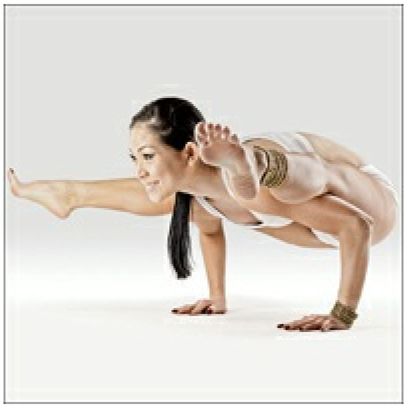

Firefly Pose 1

Tittibhasana 1

(ti-ti-BAHS-uh-nuh)

Also Known As: Firefly Pose A (Tittibhasana A)
Modification: arms straight, feet extended to the sky
Pose Type: arm balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Nasagrai or Nasagre (nose)

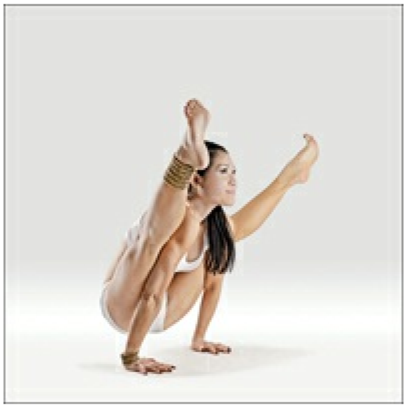

Firefly Pose 1

Tittibhasana 1

(ti-ti-BAHS-uh-nuh)

Also Known As: Firefly Pose A (Tittibhasana A), Raised Tortoise Pose (Utthita Kurmasana)
Modification: arms straight, legs parallel to the floor
Pose Type: arm balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
ल व ह

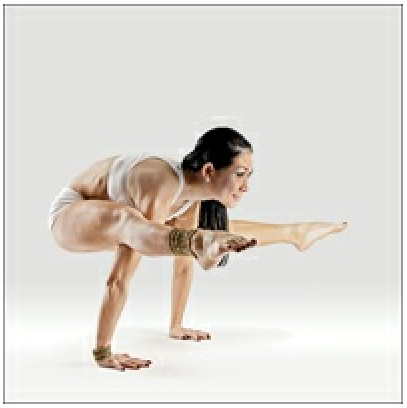

ARM BALANCE: FIREFLY POSE

Firefly Pose

Tittibhasana

(ti-ti-BAHS-uh-nuh)

Also Known As: Raised Up Feet Spread Out Resting on the Arms Pose (Utthita Dwi Pada Vrishtasana)
Modification: elbows bent, legs on triceps, wide legged
1. side view
2. front view

Pose Type: arm balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं हं ॐ

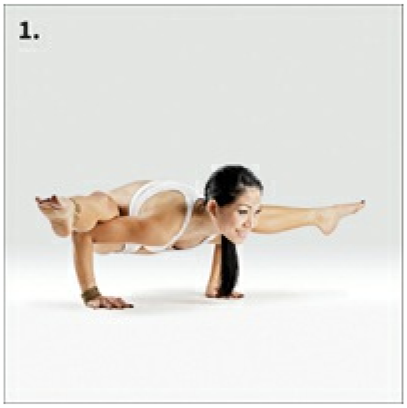
1.

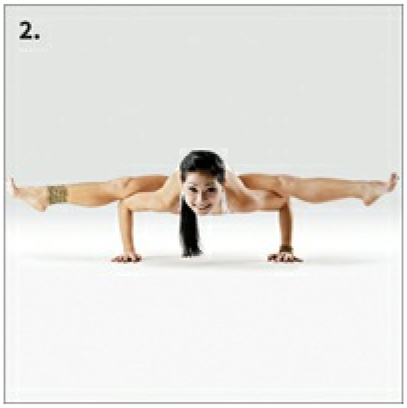
2.

Lifted Feet Spread Out Pose

Utthita Dwi Pada Vrishtasana

(UT-ti-tuh DWI-puh-duh vrish-TAHS-uh-nuh)
Modification: arms straight, heels of the palms touching, fingertips pointing to the sides
Pose Type: arm balance, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

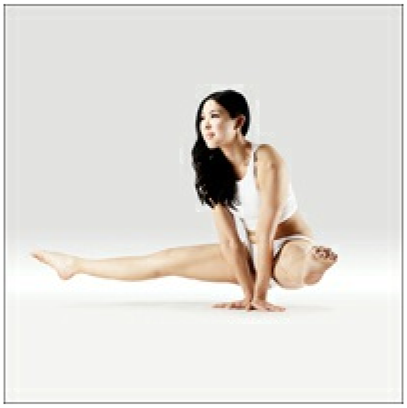

One-Legged Crane 1

Eka Pada Bakasana 1

(EY-kuh PUH-duh buh-KAHS-uh-nuh)

Modification: knee to the tricep
Pose Type: arm balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

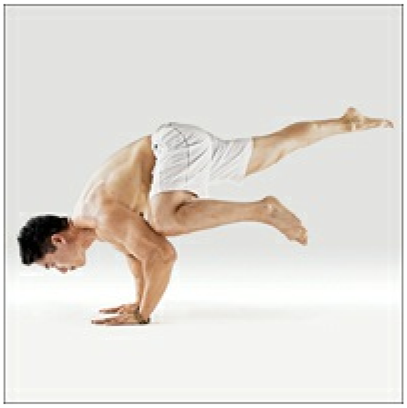

One-Legged Crane 1

Eka Pada Bakasana 1

(EY-kuh PUH-duh buh-KAHS-uh-nuh)

Modification: knee to the outside of the shoulder
Pose Type: arm balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

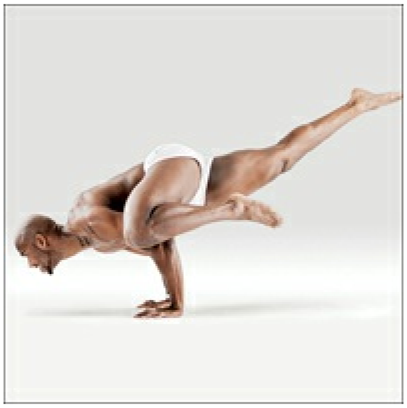

One-Legged Crane 2

Eka Pada Bakasana 2

(EY-kuh PUH-duh buh-KAHS-uh-nuh)

Modification: shin of the bent leg to the tricep

Pose Type: arm balance, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Nasagrai or Nasagre (nose)

लं वं हं ॐ

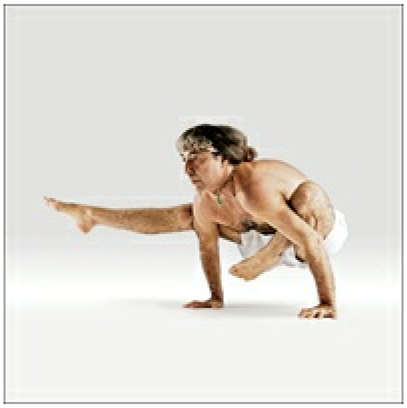

Pose Dedicated to Galava, One-Legged Modification

Eka Pada Galavasana

(EY-kuh PUH-duh gah-luh-VAHS-uh-nuh)

Pose Type: arm balance, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

लं वं हं

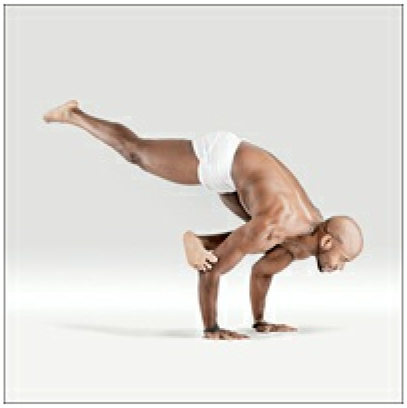

Dragonfly Pose 1

Maksikanagasana 1

(muhk-shi-kah-nah-GAHS-uh-nuh)

Also Known As: Stick Arm to the Side Grasshopper Pose (Parshva Bhuja Danda Salabhasana)
Pose Type: arm balance, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose)
लं वं हं

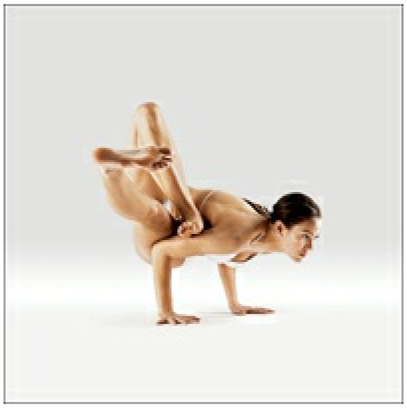

Moonbird Pose

Chakorasana

(chuh-kor-AHS-uh-nuh)

Pose Type: arm balance, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं रं ॐ

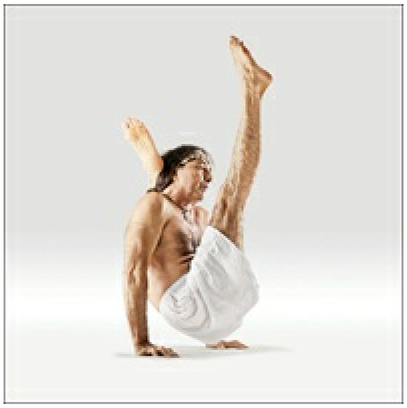

Uneven One-Legged Crane 1

Vishama Eka Pada Bakasana 1

(VISH-uh-muh EY-kuh PUH-duh buh-KAHS-uh-nuh)

Modification: one forearm to the floor, other elbow bent at 90 degrees
Pose Type: arm/forearm balance, forward bend, inversion

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

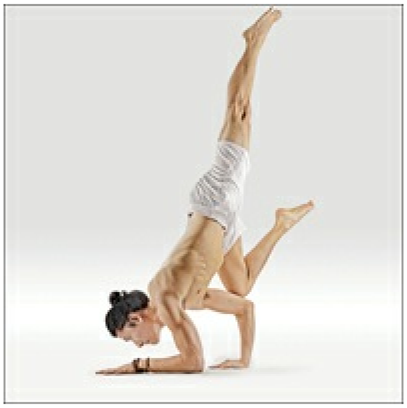

Arm to the Side Pose Dedicated to Sage Koundinya One-Legged Version 1

Parshva Hasta Eka Pada Koundinyasana 1

(PAHRSH-vuh HUH-stuh EY-kuh PUH-duh kown-din-YAHS-uh-nuh)

Modification: bottom knee bent
Pose Type: arm balance, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

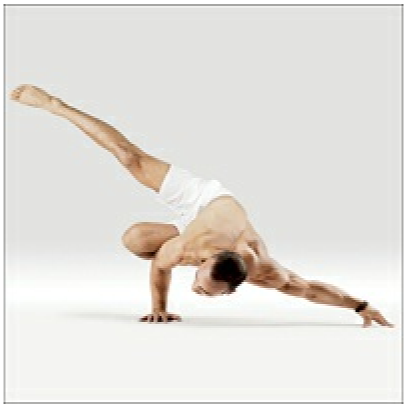

ARM BALANCE: SCISSOR LEGS

Pose Dedicated to Sage Koundinya One-Legged Version 1 Prep.

Eka Pada Koundinyasana 1 Prep.

(EY-kuh PUH-duh kown-din-YAHS-uh-nuh)

Modification: ear to the floor, bottom knee bent

Pose Type: arm balance, forward bend, twist

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

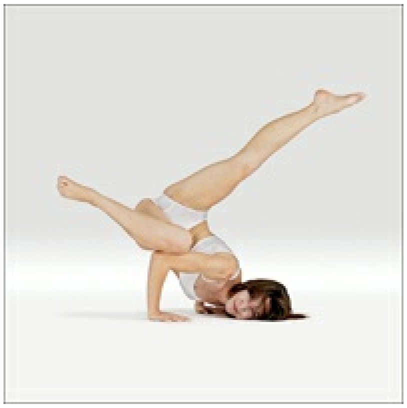

Pose Dedicated to Sage Koundinya One-Legged Version 1 Prep.

Eka Pada Koundinyasana 1 Prep.

(EY-kuh PUH-duh kown-din-YAHS-uh-nuh)

Modification: head on the floor

Pose Type: arm balance, forward bend, twist

Drishti Point: Nasagrai or Nasagre (nose)

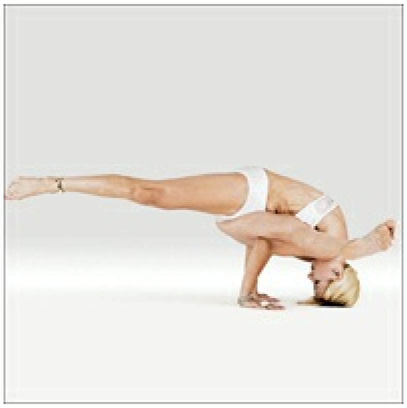

Pose Dedicated to Sage Koundinya One-Legged Version 2

Eka Pada Koundinyasana 2

(EY-kuh PUH-duh kown-din-YAHS-uh-nuh)

Modification: 1. prep. back foot on the floor, toes curled in
2. back foot lifted

Pose Type: arm balance, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre

(nose)

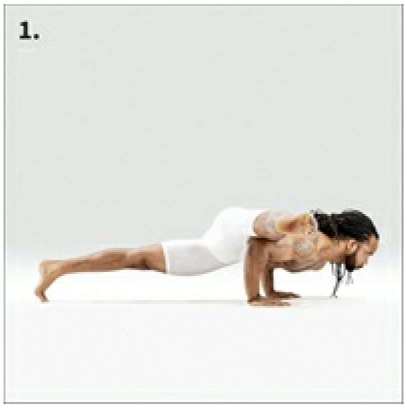

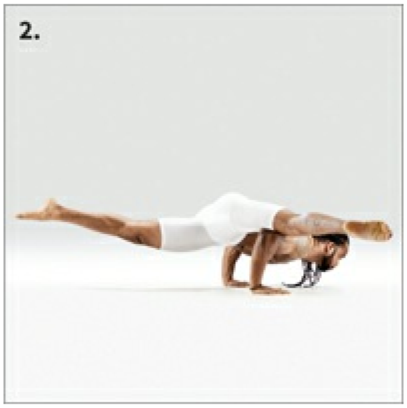

ARM BALANCE: GARUDA LEGS

Revolved Leg Position of the Pose Dedicated to Garuda in Swan Pose

Parivritta Pada Garudasana in Hamsasana

(puh-ri-VRIT-tuh PUH-duh guh-ru-DAHS-uh-nuh in huhms-AHS-uh-nuh)

Pose Type: arm balance, forward bend, twist

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

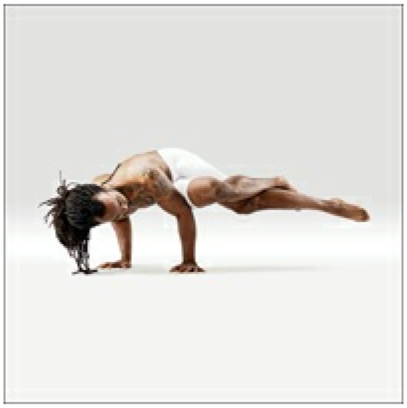

Revolved Leg Position of the Pose Dedicated to Garuda in Uneven Swan Pose

Parivritta Pada Garudasana in Vishama Hamsasana

(puh-ri-VRIT-tuh PUH-duh guh-ru-DAHS-uh-nuh in VISH-uh-muh huhms-AHS-uh-nuh)

Modification: forearm to the floor
Pose Type: arm balance, forward bend, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

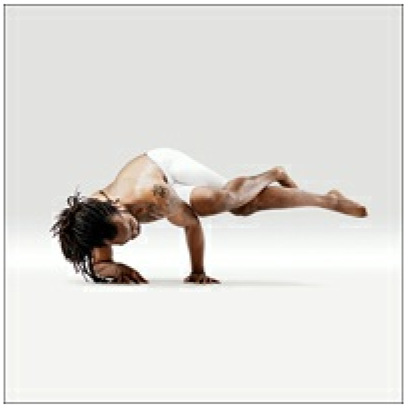

Revolved Leg Position of the Pose Dedicated to Garuda in Uneven Swan Pose

Parivritta Pada Garudasana in Vishama Hamsasana

(puh-ri-VRIT-tuh PUH-duh guh-ru-DAHS-uh-nuh in VISH-uh-muh huhms-AHS-uh-nuh)

Modification: elbow to the floor
Pose Type: arm balance, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose) or Padayoragrai or Padayoragre (toes/feet)

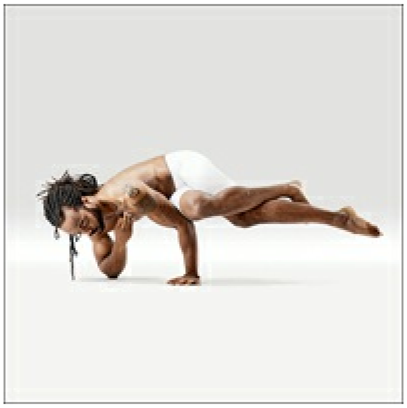

Two-Legged Pose Dedicated to Koundinya

Dwi Pada Koundinyasana

(DWI-puh-duh kown-din-YAHS-uh-nuh)

Pose Type: arm balance, forward bend, twist

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

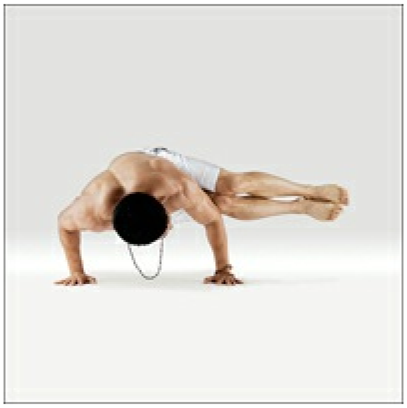

Uneven Two-Legged Pose Dedicated to Koundinya

Vishama Dwi Pada Koundinyasana

(VISH-uh-muh DWI-puh-duh kown-din-YAHS-uh-nuh)

Modification: one forearm to the floor, knees bent

Pose Type: arm balance, forward bend, twist

Drishti Point: Padayoragrai or Padayoragre (toes/feet)

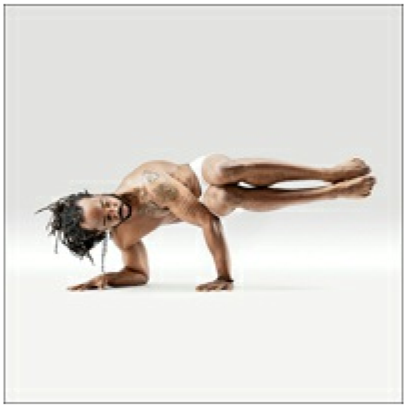

Pose Dedicated to Ashtavakra Prep.

Ashtavakrasana Prep.

(uh-shtuh-vuh-KRAHS-uh-nuh)

Also Known As: Eight Angle Pose Prep.
Modification: feet unhooked
Pose Type: arm balance, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose)
ल ह

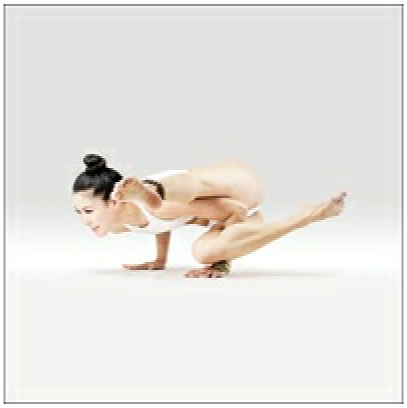

ARM BALANCE: BOTH LEGS TO THE SIDE—ASHTAVAKRA’S POSE

Pose Dedicated to Ashtavakra

Ashtavakrasana

(uh-shtuh-vuh-KRAHS-uh-nuh)

Also Known As: Eight Angle Pose
Pose Type: arm balance, forward bend, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व

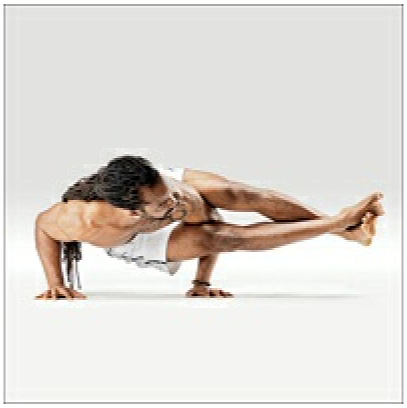

Uneven Pose Dedicated to Ashtavakra

Vishama Ashtavakrasana

(VISH-uh-muh uh-shtuh-vuh-KRAHS-uh-nuh)

Also Known As: Uneven Eight Angle Pose
Modification: one forearm to the floor
Pose Type: arm/forearm balance, forward bend, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

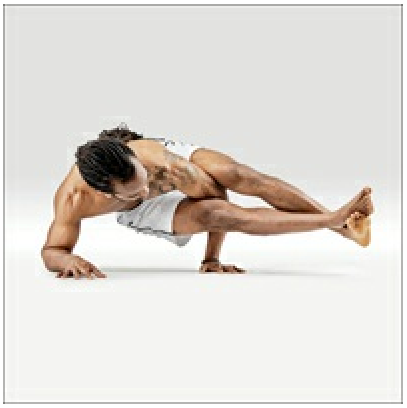

Uneven Half Repose Pose Dedicated to Ashtavakra

Vishama Ardha Shayana Ashtavakrasana

(VISH-uh-muh UHR-duh shuh-yuh-nuh uh-shtuh-vuh-KRAHS-uh-nuh)

Also Known As: Uneven Half Repose Eight Angle Pose
Modification: one elbow to the floor, hand to the face
Pose Type: arm/elbow balance, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose)

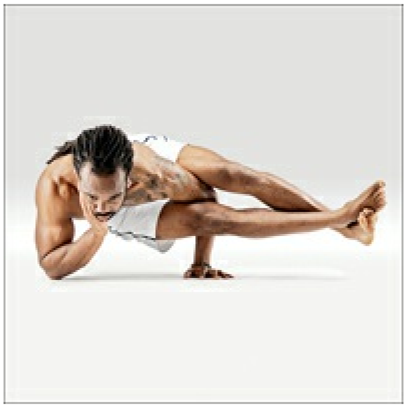

ARM BALANCE: BOTH LEGS TO THE SIDE—KNEES BENT AND TOGETHER

Side Crane Pose

Parshva Bakasana

(PAHRSH-vuh buh-KAHS-uh-nuh)

Also Known As: Revolved Crane Pose (Parivritta Bakasana), Sideways Crow Pose (Parshva Kakasana)
Modification: 1. elbows bent
2. arms straight
Pose Type: arm balance, forward bend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

1.

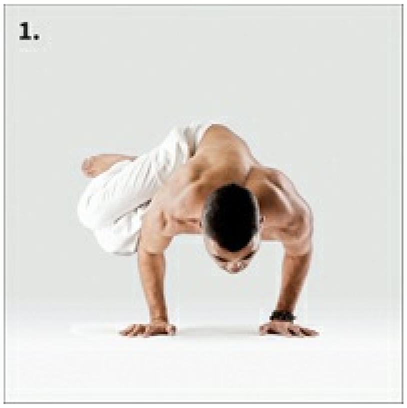

2.

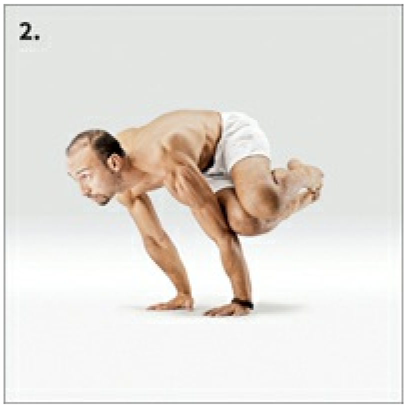

Two-Legged Pose Dedicated to Koundinya Modification on the Fists

Mushti Dwi Pada Koundinyasana

(mush-ti DWI-puh-duh kown-din-YAHS-uh-nuh)

Modification: ankles crossed
Pose Type: arm balance, forward bend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

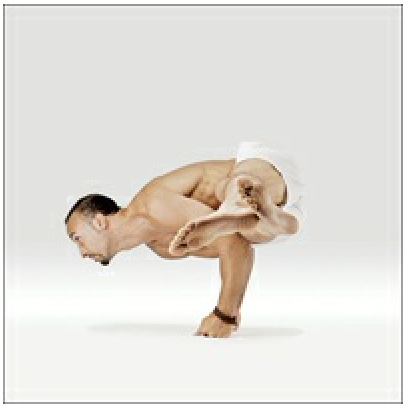

Uneven Arms Side Crane Pose

Vishama Hasta Parshva Bakasana

(VISH-uh-muh HUH-stuh PAHRSH-vuh buh-KAHS-uh-nuh)

Pose Type: arm balance, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose)

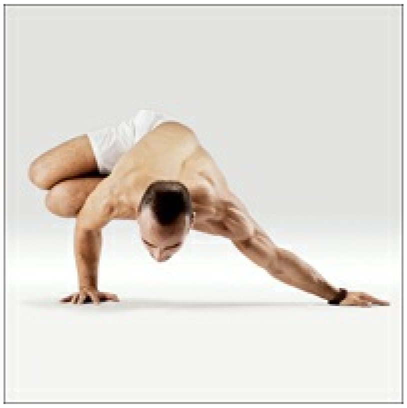

ARM BALANCE: LOTUS PEACOCK

Bound Angle Pose in Peacock Pose

Baddha Konasana in Mayurasana

(BUH-duh ko-NAHS-uh-nuh in muh-yoor-AHS-uh-nuh)

Pose Type: arm balance
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

Lotus Pose in Peacock Pose

Padmasana in Mayurasana

(puhd-MAHS-uh-nuh in muh-yoor-AHS-uh-nuh)

Pose Type: arm balance, hip opener

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

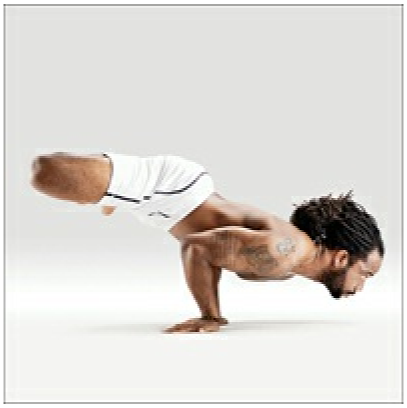

Fist Lotus Peacock Pose

Mushti Padma Mayurasana

(mush-ti PUHD-muh muh-yoor-AHS-uh-nuh)

Pose Type: arm balance

Drishti Point: Nasagrai or Nasagre (nose)

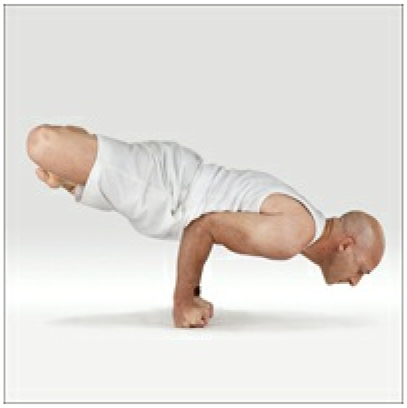

ARM BALANCE: PEACOCK AND SWAN

Peacock Pose

Mayurasana

(muh-yoor-AHS-uh-nuh)

Modification: 1. feet lifted higher than the hip level
2. body parallel to the floor
Pose Type: arm balance
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

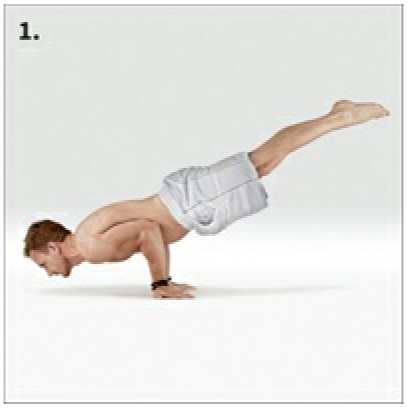

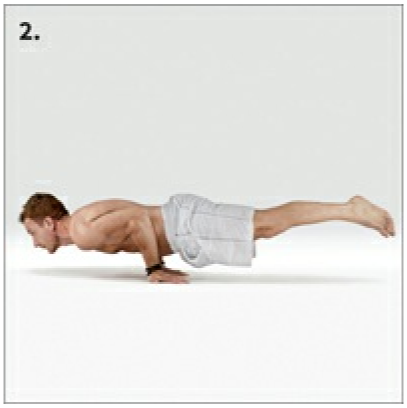

Swan Pose

Hamsasana

(huhms-AHS-uh-nuh)

Also Known As: Swan Pose (Hansasana)
Pose Type: arm balance
Drishti Point: Nasagrai or Nasagre (nose)

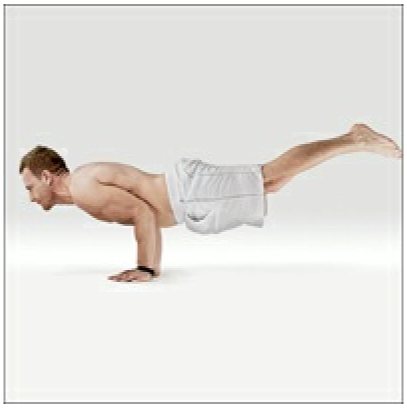

One Hand Peacock Pose

Eka Hasta Mayurasana

(EY-kuh HUH-stuh muh-yoor-AHS-uh-nuh)

Also Known As: Wounded Peacock Pose (Pungu Mayurasana)
Pose Type: arm balance
Drishti Point: Nasagrai or Nasagre (nose)

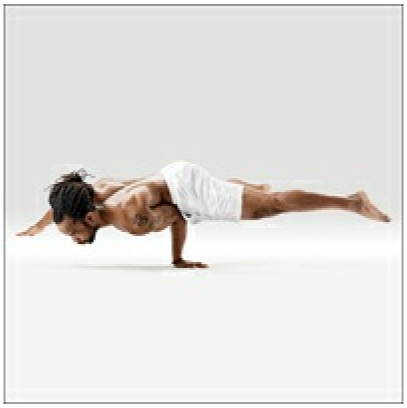

Inversion Poses

- DUCK POSE
- DOLPHIN POSE
- PEACOCK FEATHER POSE
- BACKBENDS ON FOREARMS
- HANDSTANDS (VARIATIONS)
- HEADSTANDS (VARIATIONS)
- SHOULDERSTAND (VARIATIONS)
- PLOW POSE

Duck Pose

Karandavasana

(kahr-uhn-duh-VAHS-uh-nuh)

Also Known As: Baby Duck Pose
Modification: thumbs grabbing onto the biceps, fingers grabbing onto the triceps; knees to the armpits, feet together
Pose Type: forearm balance, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

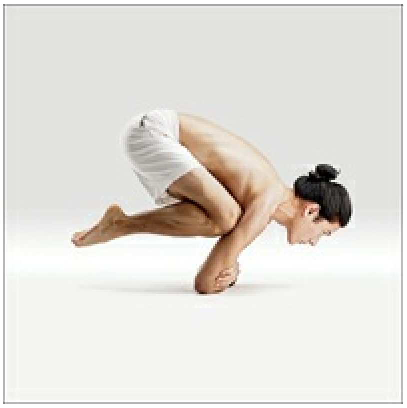

Duck Pose

Karandavasana

(kahr-uhn-duh-VAHS-uh-nuh)

Also Known As: Baby Duck Pose
Modification: forearms to the floor, knees to the armpits, feet together
Pose Type: forearm balance, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

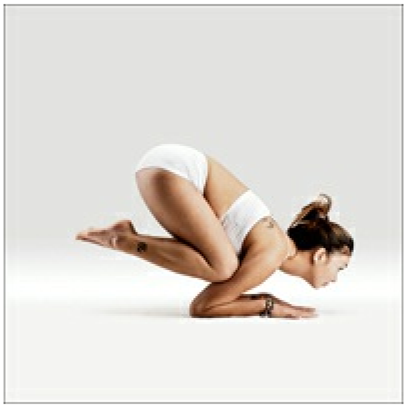

Downward Facing Pose Dedicated to Makara

Adho Mukha Makarasana

(uh-DO MUK-uh muh-kuh-RAHS-uh-nuh)

Also Known As: Dolphin Pose
Pose Type: forward bend, inversion, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

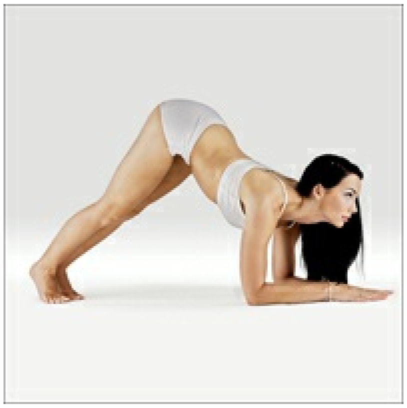

One-Legged Uneven Peacock Feather Pose

Eka Pada Vishama Picha Mayurasana

(EY-kuh PUH-duh VISH-uh-muh pich-chuh muh-yoor-AHS-uh-nuh)

Also Known As: Eka Pada Vishama Pincha Mayurasana
Pose Type: forearm balance, inversion, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

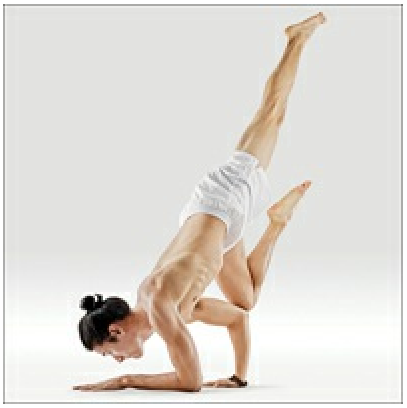

Half Lotus Downward Facing Pose Dedicated to Makara

Ardha Padma Adho Mukha Makarasana

(UHR-duh PUHD-muh uh-DO MUK-uh muh-kuh-RAHS-uh-nuh)

Also Known As: Half Lotus Dolphin Pose
Modification: heel down
Pose Type: forward bend, inversion, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

ल व य ह

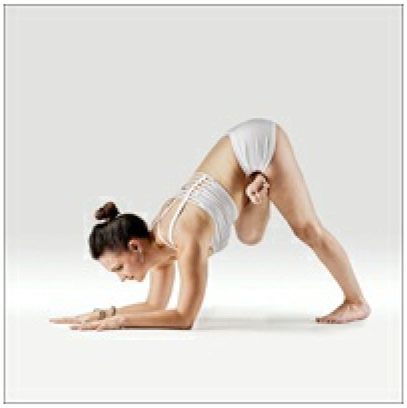

Leg Position of the Pose Dedicated to Garuda in Downward Facing Pose Dedicated to Makara

Pada Garudasana in Adho Mukha Makarasana

(PUH-duh guh-ru-DAHS-uh-nuh in uh-DO MUK-uh muh-kuh-RAHS-uh-nuh)

Also Known As: Leg Position of the Pose Dedicated to Garuda in Dolphin Pose
Pose Type: forward bend, inversion, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

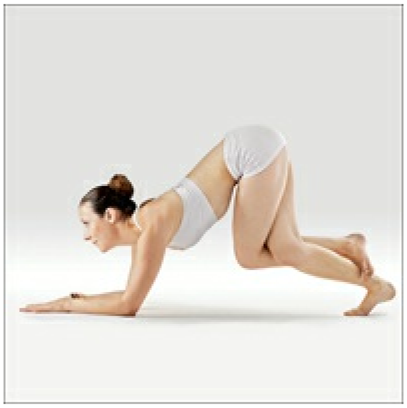

PEACOCK FEATHER POSE

Peacock Feather Pose

Picha Mayurasana

(pich-chuh muh-yoor-AHS-uh-nuh)

Also Known As: Upright Scorpion Pose (Avakra Vrishchikasana), Pincha Mayurasana
Pose Type: forearm balance, inversion
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल य ह

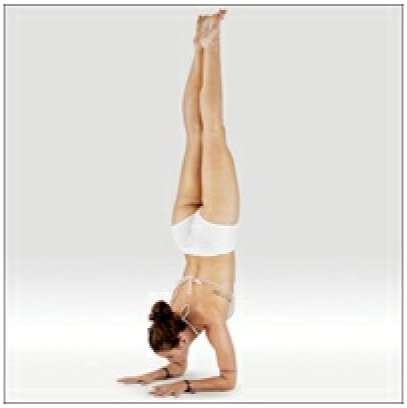

One-Legged Peacock Feather Pose

Eka Pada Picha Mayurasana

(EY-kuh PUH-duh pich-chuh muh-yoor-AHS-uh-nuh)

Also Known As: Eka Pada Pincha Mayurasana
Pose Type: forearm balance, inversion, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व य ह

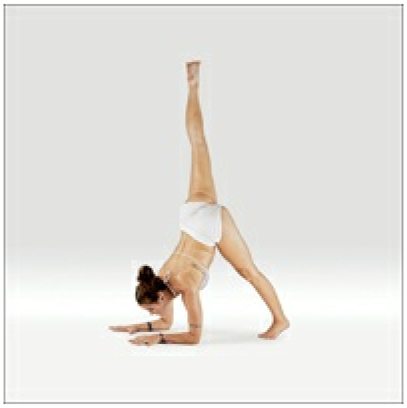

One-Legged Scorpion Pose

Eka Pada Vrishchikasana

(EY-kuh PUH-duh vrish-chi-KAHS-uh-nuh)

Modification: biceps in line with the ears

Pose Type: forearm balance, inversion, backbend
Drishti Point: Nasagrai or Nasagre (nose)

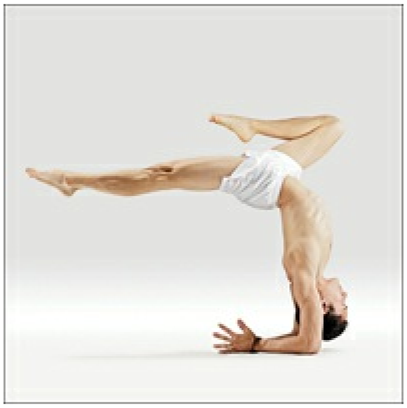

Pose Dedicated to Lord Hanuman in Peacock Feather Pose

Hanumanasana in Picha Mayurasana

(huh-nu-mahn-AHS-uh-nuh in pich-chuh muh-yoor-AHS-uh-nuh)

Also Known As: Hanumanasana in Pincha Mayurasana
Pose Type: forearm balance, inversion, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

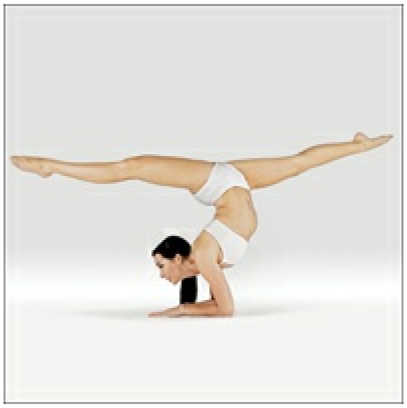

One-Legged Stretched Out Scorpion Pose

Eka Pada Paripurna Vrishchikasana

(EY-kuh PUH-duh puh-ri-POOR-nuh vrish-chi-KAHS-uh-nuh)

Pose Type: forearm balance, inversion, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

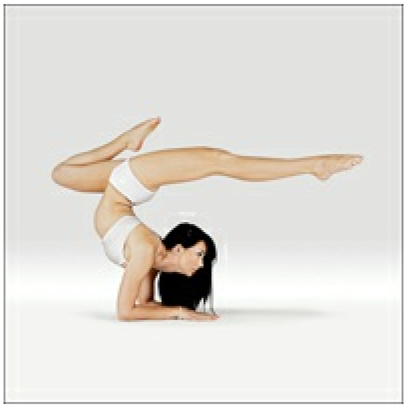

One-Legged Scorpion Pose

Eka Pada Vrishchikasana

(EY-kuh PUH-duh vrish-chi-KAHS-uh-nuh)

Modification: forearms to the floor, foot away from the head
Pose Type: forearm balance, inversion, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

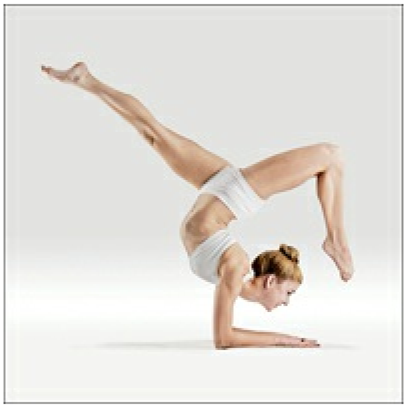

BACKBENDS ON FOREARMS, LOTUS ON FOREARMS, AND ELBOW BALANCES

Inverted Puppy Dog Pose in Peacock Feather Pose

Viparita Shvanakasana in Picha Mayurasana

(vi-puh-REE-tuh shvuh-nuh-KAHS-uh-nuh in pich-chuh muh-yoor-AHS-uh-nuh)

Also Known As: Viparita Shvanakasana in Pincha Mayurasana
Modification: knees bent, backbend
Pose Type: forearm balance, inversion, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

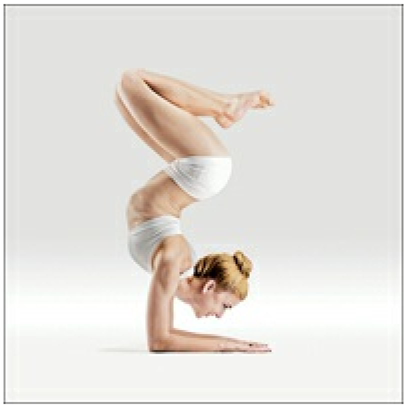

Scorpion Pose

Vrishchikasana

(vrish-chi-KAHS-uh-nuh)

Modification: forearms to the floor, feet to the head
Pose Type: forearm balance, inversion, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

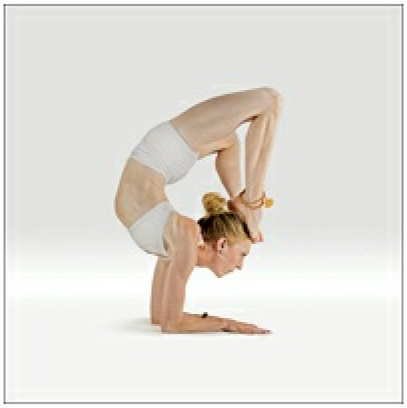

One-Legged Scorpion Pose

Eka Pada Vrishchikasana

(EY-kuh PUH-duh vrish-chi-KAHS-uh-nuh)

Modification: forearms to the floor, foot to the head, other leg parallel to the floor
Pose Type: forearm balance, inversion, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

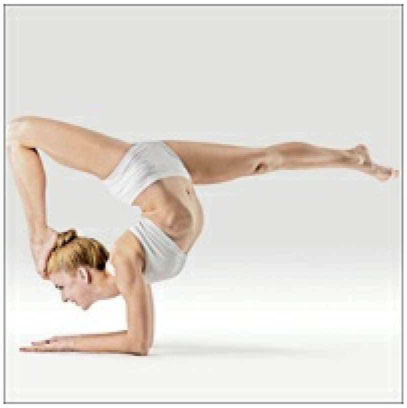

Upward Lotus Pose in Peacock Feather Pose

Urdhva Padmasana in Picha Mayurasana

(OORD-vuh puhd-MAHS-uh-nuh in pich-chuh muh-yoor-AHS-uh-nuh)

Also Known As: Raised Lotus Scorpion Pose (Urdhva Padma Vrishchikasana), Duck Pose (Karandavasana)
Pose Type: forearm balance, inversion
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
लं वं हं

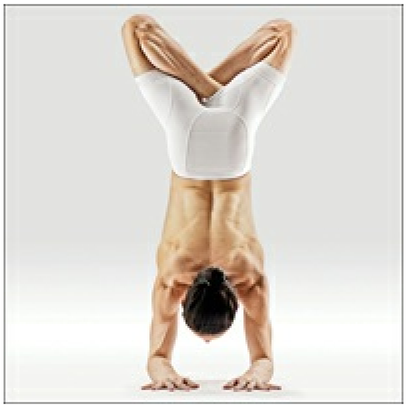

Uneven Repose Pose

Vishama Shayanasana

(VISH-uh-muh shuh-yuh-NAHS-uh-nuh)

Modification: both knees bent
Pose Type: forearm/elbow balance, inversion, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं यं हं ॐ

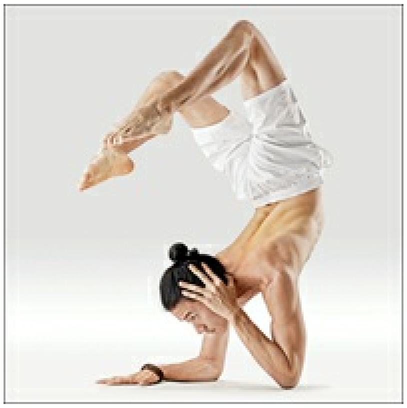

Repose Pose

Shayanasana

(shuh-yuh-NAHS-uh-nuh)

Modification: both knees bent
Pose Type: elbow balance, inversion, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

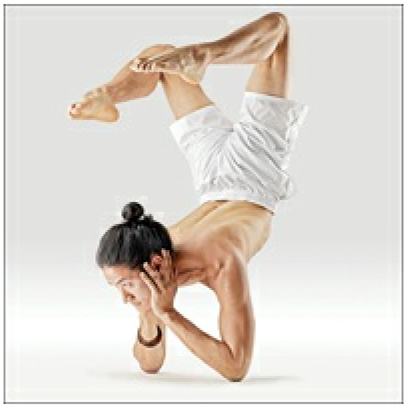

HANDSTANDS: BOTH LEGS STRAIGHT

Feet Spread Intense Stretch Pose in Downward Facing Tree Pose

Prasarita Padottanasana in Adho Mukha Vrikshasana

(pruh-SAH-ri-tuh puh-do-tahn-AHS-uh-nuh in uh-DO MUK-uh vriks-SHAHS-anna)

Also Known As: Feet Spread Full Forward Bend Pose in Downward Facing Tree Pose, Prasarita Padottanasana in Adho Mukha Vrksasana
Pose Type: inversion, arm balance, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

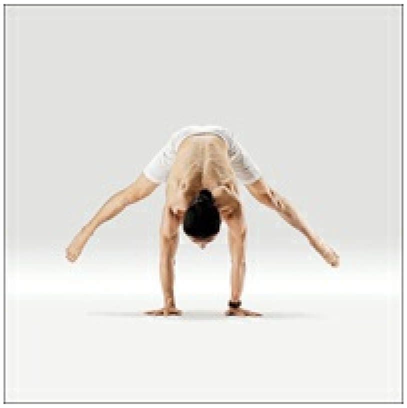

One-Legged Downward Facing Tree Pose

Eka Pada Adho Mukha Vrikshasana

(EY-kuh PUH-duh uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Eka Pada Adho Mukha Vrksasana
Pose Type: inversion, arm balance, forward bend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

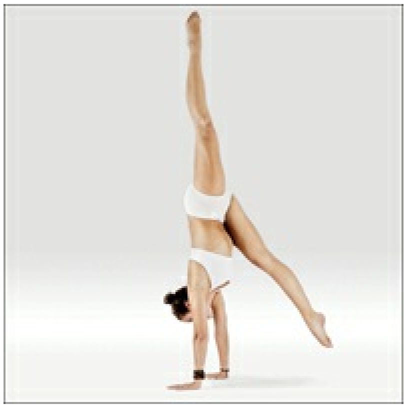

Downward Facing Tree Pose

Adho Mukha Vrikshasana

(uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Adho Mukha Vrksasana
Modification: legs crossed
Pose Type: inversion, arm balance
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

Seated Angle Pose in Downward Facing Tree Pose

Upavishta Konasana in Adho Mukha Vrikshasana

(u-puh-VISH-tuh ko-NAHS-uh-nuh in uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Upward Spread Feet Pose A (Urdhva Prasarita Padasana A), Upavista Konasana in Adho Mukha Vrikshasana
Pose Type: inversion, arm balance
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व य ह

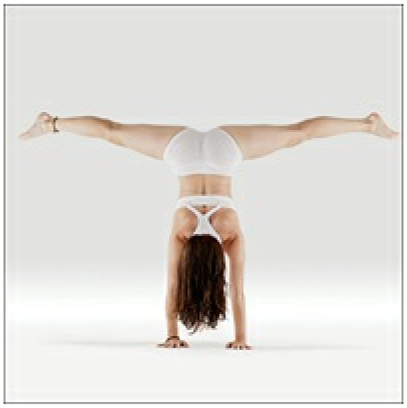

Leg Position of the Pose Dedicated to Garuda in

Downward Facing Tree Pose

Pada Garudasana in Adho Mukha Vrikshasana

(PUH-duh guh-ru-DAHS-uh-nuh in uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Pada Garudasana in Adho Mukha Vrksasana

Pose Type: inversion, arm balance

Drishti Point: Angusthamadhye or Angustha Ma Dyai (thumbs)

How to Perform the Pose:

1. Begin by standing in Mountain Pose (*Tadasana*). Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale and hinge from the hips, coming into a forward bend, placing the palms on the floor on the outsides of your feet. Your hands should be shoulder width apart or slightly wider. Make sure your arms are straight and your shoulders are on top of your fingertips.
3. There are many ways to come into a handstand. When you start practicing handstands, make sure you can balance on your hands against the wall for at least 60 seconds. Then you can experiment with jumping into a handstand or lifting your legs using your core. Press strongly into your hands.
4. Once you find your balance in a Downward Facing Tree Pose (*Adho Mukha Vrikshasana*), also known as Handstand, exhale and bend your knees; cross your right leg over your left leg, hooking your right foot around your left calf muscle. Hold for 30, and up to 90, seconds to receive the full benefits of the stretch.
5. Inhale and bring your legs back to the straight position. Exhale and switch legs as you bend your knees and cross your left leg over your right leg, hooking your left foot around your right calf muscle. Hold for 30, and up to 90, seconds to receive the full benefits of the stretch.
6. Inhale and bring your legs back to the straight position. On the following inhale, lower your feet to the floor. Inhale as you come back to Mountain Pose (*Tadasana*).

pada = foot or leg

Garuda = Hindu Deity, half-man half-eagle, carrier of Lord Vishnu

adho = downward

mukha = facing

vriksha = tree

Frog Pose in Downward Facing Tree Pose

Mandukasana in Adho Mukha Vrikshasana

(muhn-doo-KAHS-uh-nuh in uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Mandukasana in Adho Mukha Vrksasana
Pose Type: inversion, arm balance
Drishti Point: Nasagrai or Nasagre (nose) or Angusthamadhye or Angustha Ma Dyai (thumbs)

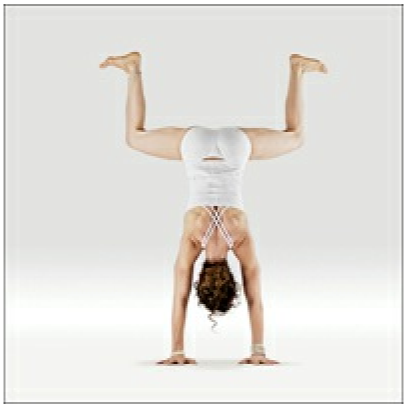

Bound Angle Pose in Downward Facing Tree Pose

Baddha Konasana in Adho Mukha Vrikshasana

(BUH-duh ko-NAHS-uh-nuh in uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Baddha Konasana in Adho Mukha Vrksasana
Pose Type: inversion, arm balance
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

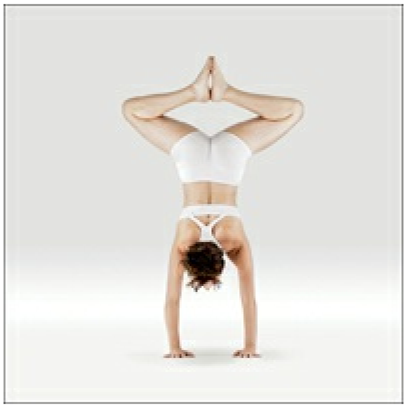

Svastika Legs in Downward Facing Tree Pose

Pada Svastikasana in Adho Mukha Vrikshasana

(PUH-duh svuh-sti-KAHS-uh-nuh in uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Pada Svastikasana in Adho Mukha Vrksasana

Modification: both knees bent, one knee bent toward the chest, other foot pointing to the back

Pose Type: inversion, arm balance, mild backbend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

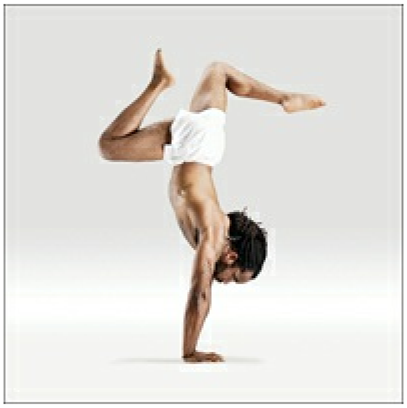

Scorpion Pose Prep.

Vrishchikasana Prep.

(vrish-chi-KAHS-uh-nuh)

Also Known As: Pose Dedicated to the Demon Taraka A (Tarakasana A)

Modification: knees bent, feet away from the head

Pose Type: inversion, arm balance, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

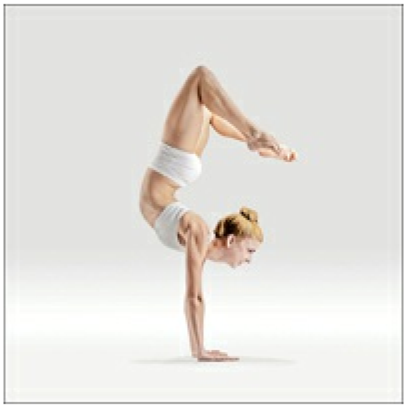

Half Lotus Pose in Downward Facing Tree Pose

Ardha Padmasana in Adho Mukha Vrikshasana

(UHR-duh puhd-MAHS-uh-nuh in uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Ardha Padmasana in Adho Mukha Vrksasana
Pose Type: inversion, arm balance
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

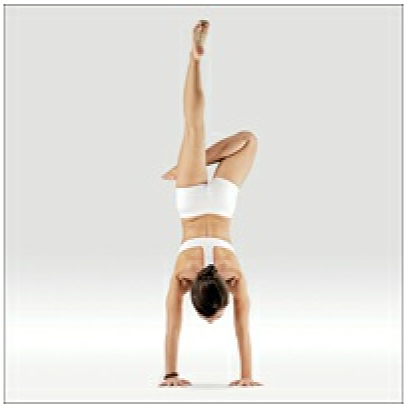

HANDSTAND, LEGS CROSSED: GARUDA LEGS & LOTUS

Leg Position of the Pose Dedicated to Garuda in Downward Facing Tree Pose

Pada Garudasana in Adho Mukha Vrikshasana

(PUH-duh guh-ru-DAHS-uh-nuh in uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Pada Garudasana in Adho Mukha Vrksasana
Modification: arms bent at 90 degrees
Pose Type: inversion, arm balance
Drishti Point: Nasagrai or Nasagre (nose)

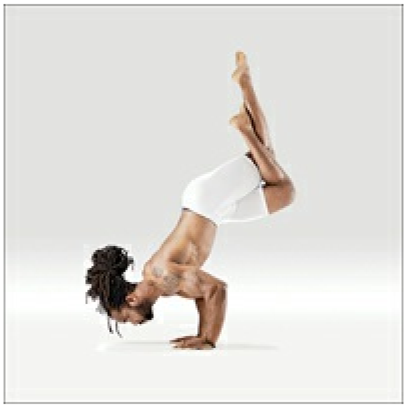

Upward Lotus Pose in Downward Facing Tree Pose

Urdhva Padmasana in Adho Mukha Vrikshasana

(OORD-vuh puhd-MAHS-uh-nuh in uh-DO MUK-uh vrik-SHAHS-uh-nuh)

Also Known As: Urdhva Padmasana in Adho Mukha Vrksasana

Pose Type: inversion, arm balance

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

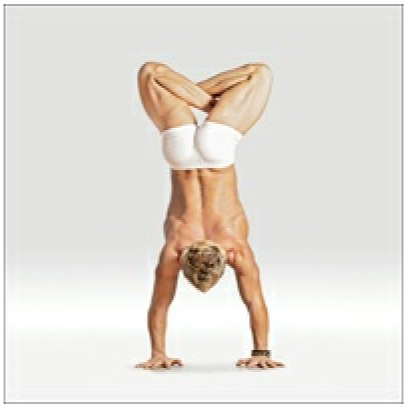

HEADSTAND 1 ARM POSITION: LEG MODIFICATIONS

Headstand 1

Shirshasana 1

(sheer-SHAHS-uh-nuh)

Also Known As: Supported Headstand (Salamba Shirshasana[*]), Bound Hands Headstand A (Baddha Hasta Shirshasana A)

Modification: 1. back view

2. side view

Pose Type: inversion

Drishti Point: Nasagrai or Nasagre (nose)

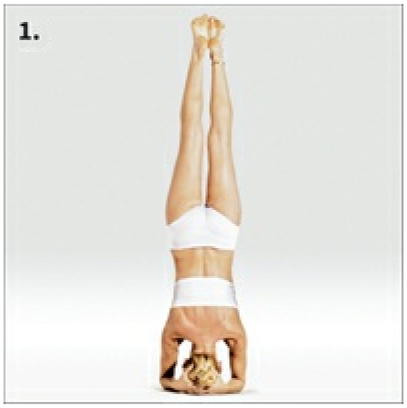

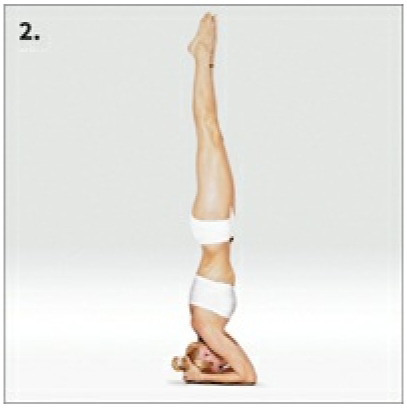

2.

One-Legged Headstand 1

Eka Pada Shirshasana 1

(EY-kuh PUH-duh sheer-SHAHS-uh-nuh)

Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Padayoragrai or Padayoragre (toes/feet)
● ● ●

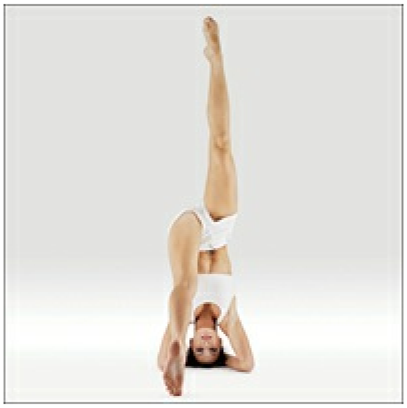

Leg Contraction Knee Bend Pose in Headstand 1

Pada Akunchanasana in Shirshasana 1

(PUH-duh uh-kunch-uh-NAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: one knee bent toward the chest
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
● ● ●

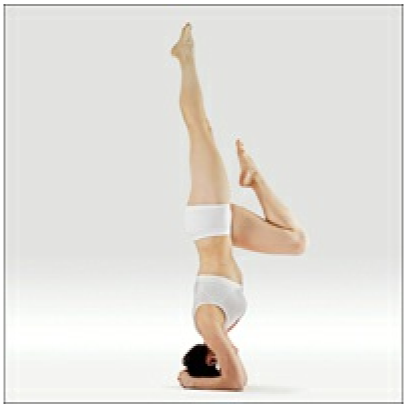

Upward Staff Pose in Headstand 1

Urdhva Dandasana in Shirshasana 1

(OORD-vuh duhn-DAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Also Known As: Half Headstand Pose (Ardha Shirshasana), Headstand B (Shirshana B)
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

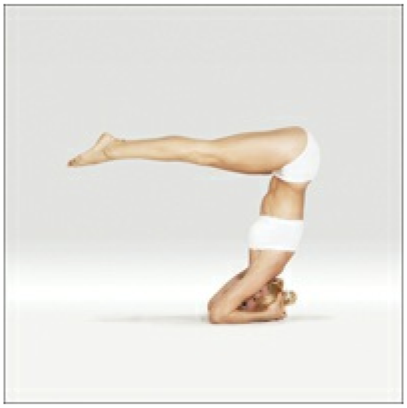

HEADSTAND 1 ARM POSITION: VARIOUS LEG POSITIONS

Svastika Legs in Headstand 1

Pada Svastikasana in Shirshasana 1

(PUH-duh svuh-sti-KAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: inversion, mild backbend
Drishti Point: Nasagrai or Nasagre (nose)

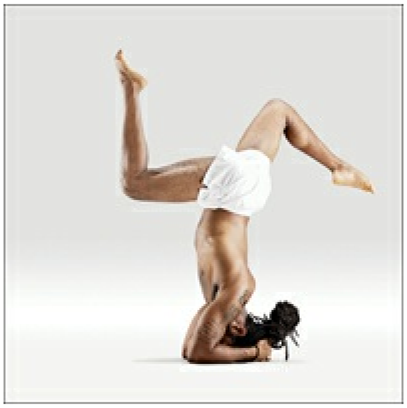

Leg Position of One-Legged King Pigeon 1 Version B in Headstand 1

Pada Eka Pada Raja Kapotasana 1 B in Shirshasana 1

(PUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: both knees bent, one knee to the sky, other foot to the knee
Pose Type: inversion, mild backbend
Drishti Point: Nasagrai or Nasagre (nose)
लं वं ॐ

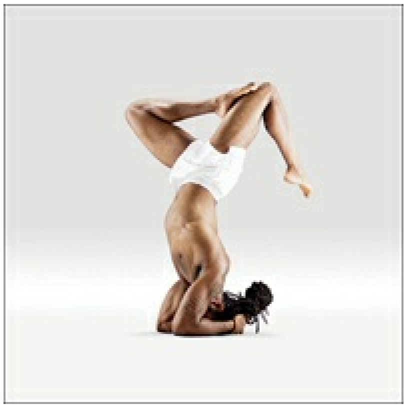

Feet to Head Pose

Shirsha Padasana

(SHEER-shuh puh-DAHS-uh-nuh)

Pose Type: inversion, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं यं हं ॐ ॐ

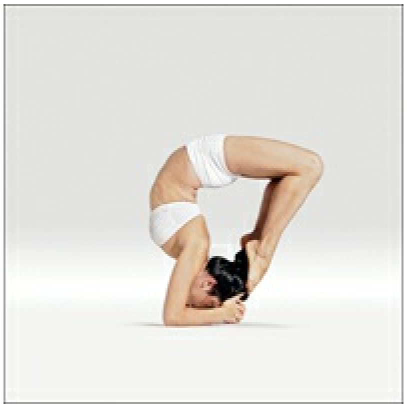

Bound Angle Pose in Headstand 1

Baddha Konasana in Shirshasana 1

(BUH-duh ko-NAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

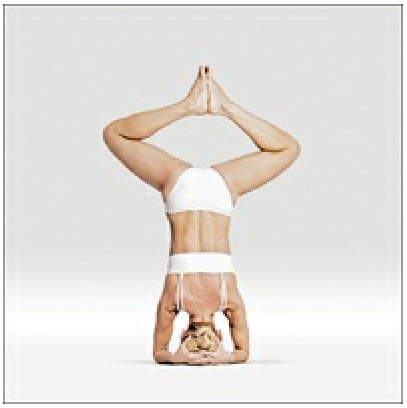

Sideways Bound Angle Pose in Headstand 1

Parshva Baddha Konasana in Shirshasana 1

(PAHRSH-vuh BUH-duh ko-NAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: inversion, side bend
Drishti Point: Nasagrai or Nasagre (nose)

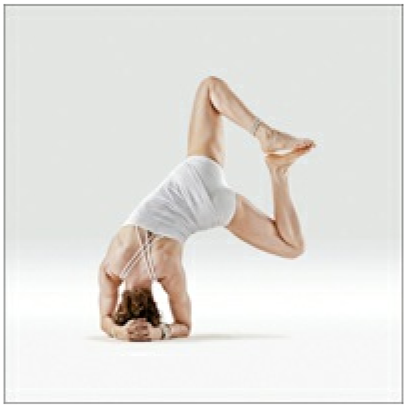

Pose Dedicated to Hanuman in Headstand 1

Hanumanasana in Shirshasana 1

(huh-nu-mahn-AHS-uh-nuh in sheer-SHAH-suh-nuh)

Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

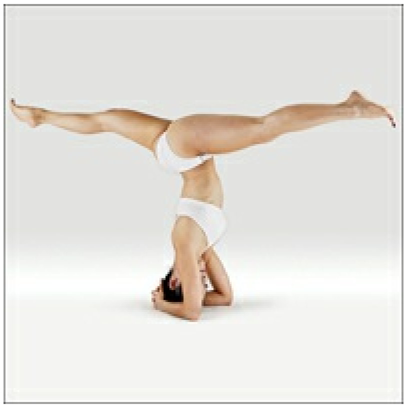

Seated Angle Pose in Headstand 1

Upavishta Konasana in Shirshasana 1

(u-puh-VISH-tuh ko-NAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Also Known As: Upavista Konasana in Adho Mukha Vrikshasana, Upavista Konasana in Adho Mukha Vrksasana
Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

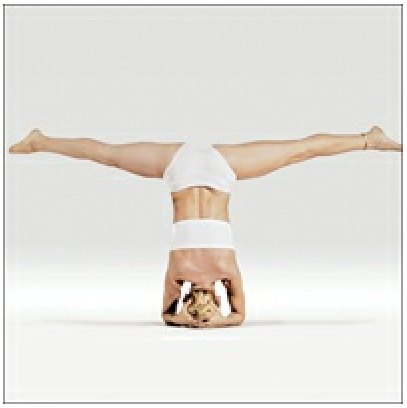

Leg Position of the Pose Dedicated to Garuda in Headstand 1

Pada Garudasana in Shirshasana 1

(PUH-duh guh-ru-DAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

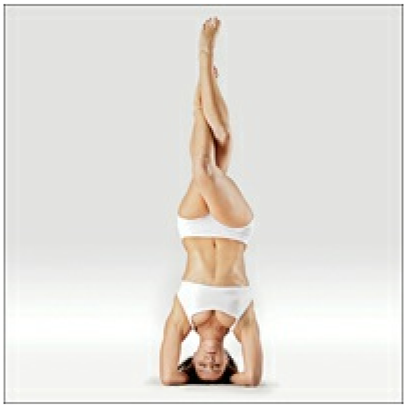

HEADSTAND 1 ARM POSITION: LOTUS LEG POSITION

Upward Lotus in Headstand 1

Urdhva Padmasana in Shirshasana 1

(OORD-vuh puhd-MAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: 1. back view
2. side view
Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

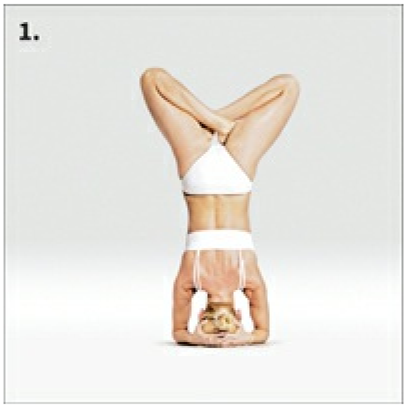

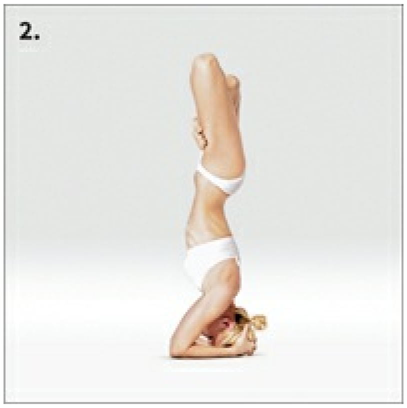

Sideways Upward Lotus in Headstand 1

Parshva Urdhva Padmasana in Shirshasana 1
(PAHRSH-vuh OORD-vuh puhd-MAHS-uh-nuh in sheer-SHAHS-uh-nuh)
Pose Type: inversion, twist
Drishti Point: Nasagrai or Nasagre (nose)
● ● ●

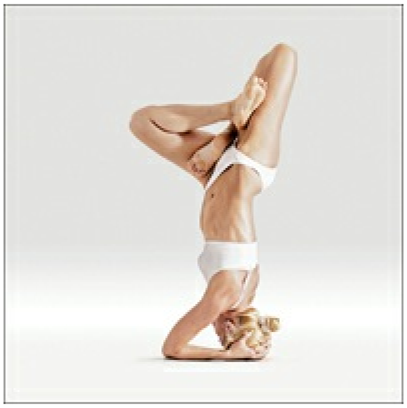

Embryo Pose in Headstand 1 Prep.

Pindasana in Shirshasana 1 Prep.
(pin-DAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: thighs parallel to the floor
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

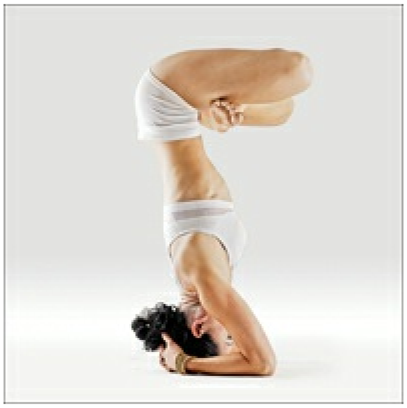

Embryo Pose in Headstand 1

Pindasana in Shirshasana 1

(pin-DAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: 1. knees to the triceps
2. knees to the chest
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

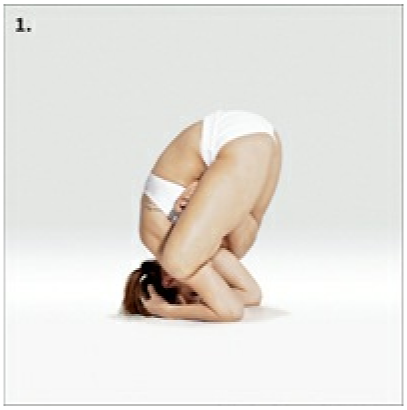

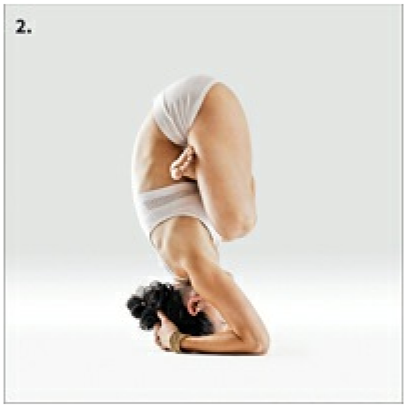

HEADSTAND 1 & 3 ARM POSITIONS: VARIOUS ARM AND LEG POSITIONS

One Hand Headstand 1

Eka Hasta Shirshasana 1

(EY-kuh HUH-stuh sheer-SHAHS-uh-nuh)

Modification: both knees bent toward the chest, one forearm on the floor
1. other arm straight, fingertips to the sky
2. other arm behind the back, fingertips pointing to the upper back
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
● ●

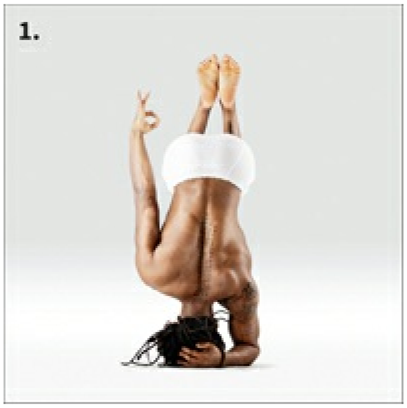

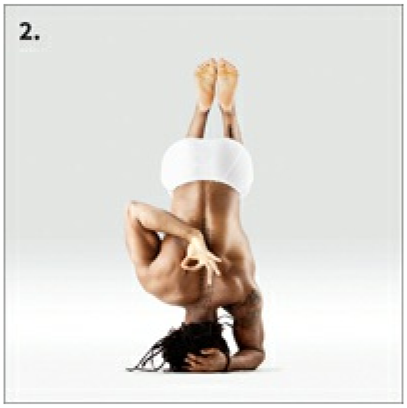

Sideways Uneven Upward Staff Pose

Parshva Vishama Urdhva Dandasana

(PAHRSH-vuh VISH-uh-muh OORD-vuh duhn-DAHS-uh-nuh)

Modification: forearm to the floor in front of the face, palm down; legs parallel to the floor
Pose Type: inversion, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

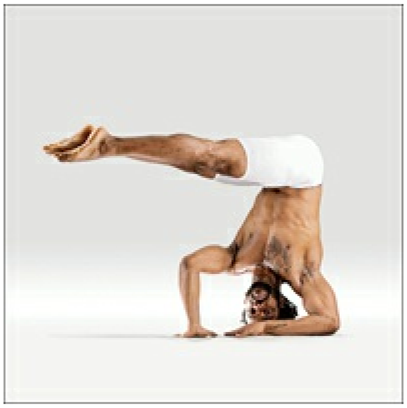

Leg Contraction Pose in Headstand 3

Pada Akunchanasana in Shirshasana 3

(PUH-duh uh-kunch-AHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: one leg straight and out to the side, other knee to the chest
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

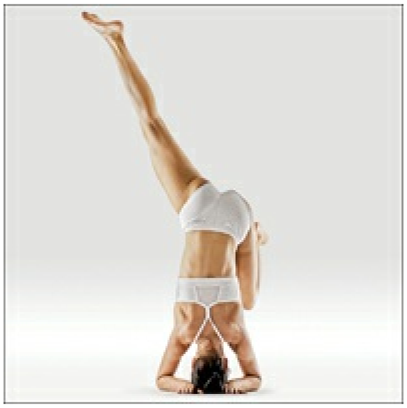

Leg Contraction Pose in Uneven Headstand (Fusion of Headstand 5 and Headstand 3 Arm Position)

Pada Akunchanasana in Vishama Shirshasana

(PUH-duh uh-kunch-AHS-uh-nuh in VISH-uh-muh sheer-SHAHS-uh-nuh)

Modification: one forearm to the floor, other elbow on top of the wrist, knee to the tricep on the same side

Pose Type: inversion, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

● ● ● ●

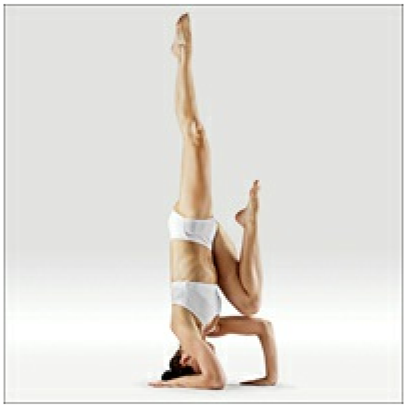

HEADSTAND 5 ARM POSITION: VARIOUS LEG POSITIONS

Leg Position of the Pose Dedicated to Garuda in Headstand 5

Pada Garudasana in Shirshasana 5

(PUH-duh guh-ru-AHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: inversion

Drishti Point: Nasagrai or Nasagre (nose)

● ● ●

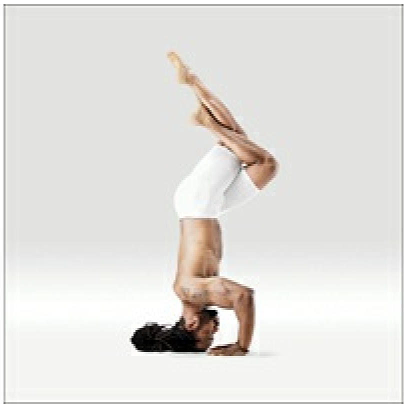

Leg Position of Cow Face Pose in Headstand 5

Pada Gomukhasana in Shirshasana 5

(PUH-duh go-muk-AHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: bottom knee to the opposite tricep
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

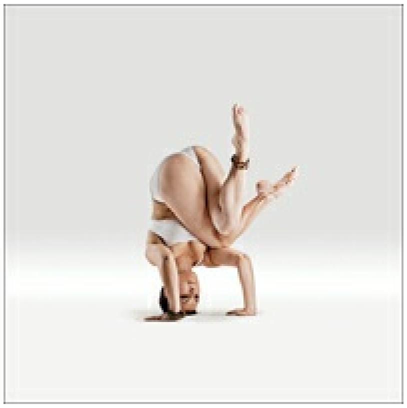

Headstand 5

Shirshasana 5

(sheer-SHAHS-uh-nuh)

Also Known As: Tripod Headstand
Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

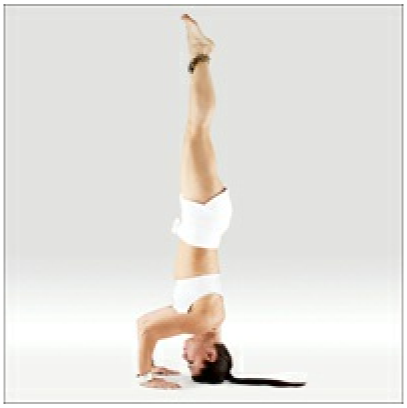

Frog Pose in Headstand 5

Mandukasana in Shirshasana 5

(muhn-doo-KAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: inversion

Drishti Point: Nasagrai or Nasagre (nose)

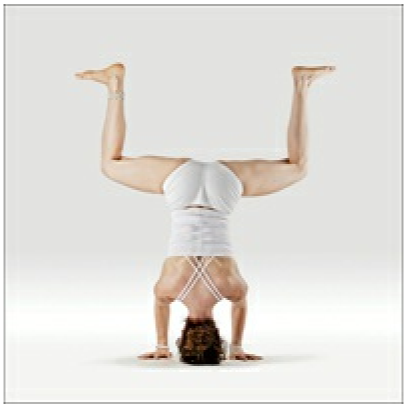

Seated Angle Pose in Headstand 5

Upavishta Konasana in Shirshasana 5

(u-puh-VISH-tuh ko-NAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Also Known As: Tripod Headstand Same Angle Pose (Utripada Shirsha Samakonasana), Upavista Konasana in Shirshasana
Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

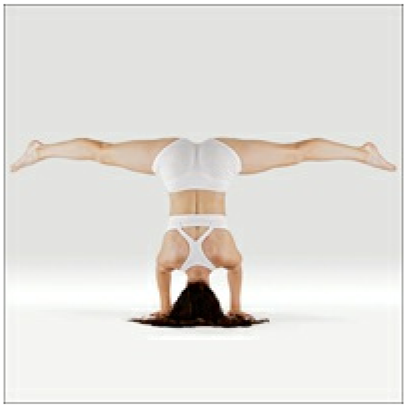

Leg Position of One-Legged King Pigeon 1 Version B in Headstand 5

Pada Eka Pada Raja Kapotasana 1 B in Shirshasana 5

(PUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: knee to the tricep on the same side, other knee resting on the foot
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

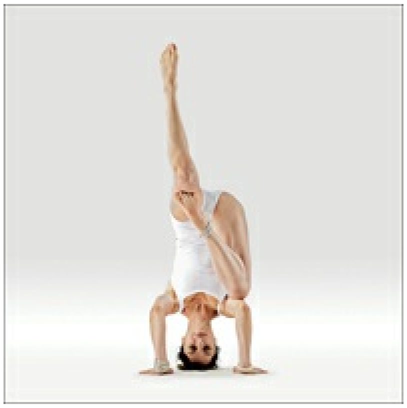

Revolved Leg Position of One-Legged King Pigeon 1 Version B in Headstand 5

Parivritta Pada Eka Pada Raja Kapotasana 1 B in Shirshasana 5

(puh-ri-VRIT-tuh PUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: knee resting on the foot
1. top leg straight
2. top knee bent at 90 degrees
Pose Type: inversion, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose)

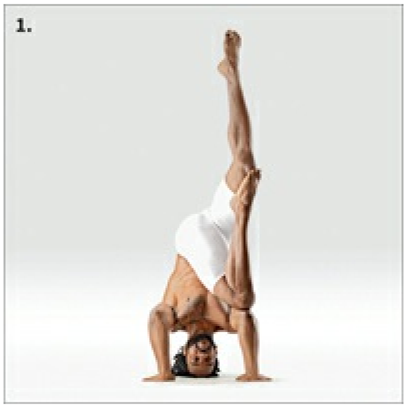
1.

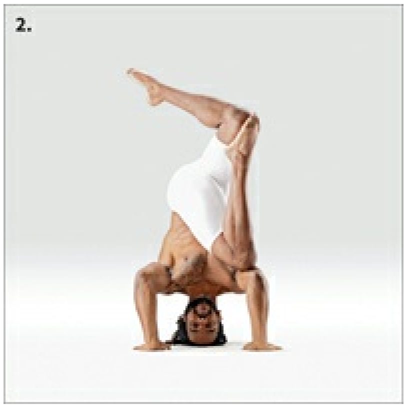
2.

Headstand 5 Prep.

Shirshasana 5 Prep.

(sheer-SHAHS-uh-nuh)

Modification: one knee to the tricep, other foot to the floor on the side
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
● ● ●

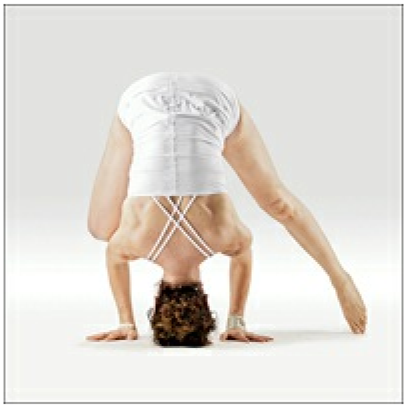

Crane Pose in Headstand 5 Prep.

Bakasana in Shirshasana 5 Prep.

(buh-KAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: knees on the triceps
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)
● ●

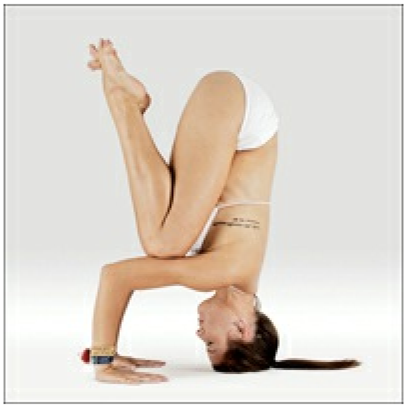

Side Crane Pose in Headstand 5

Parshva Bakasana in Shirshasana 5

(PAHRSH-vuh buh-KAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: knees together, knee to the opposite tricep

Pose Type: inversion, forward bend, twist

Drishti Point: Nasagrai or Nasagre (nose)

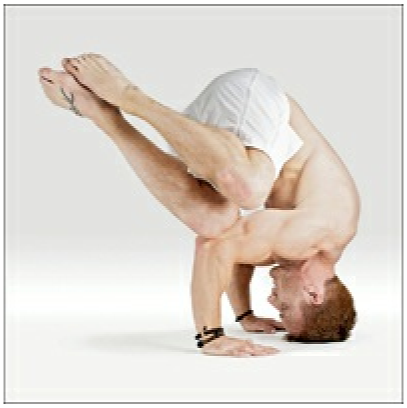

Embryo in the Womb Pose in Headstand 5

Garba Pindasana in Shirshasana 5

(guhr-buh-pin-DAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Pose Type: inversion, forward bend

Drishti Point: Nasagrai or Nasagre (nose)

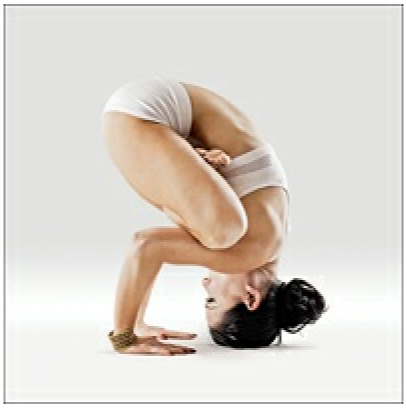

Side Rooster Pose in Headstand 5 Prep.

Parshva Kukkutasana in Shirshasana 5 Prep.

(PAHRSH-vuh ku-ku TAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: crown of the head on the floor

Pose Type: inversion, forward bend, twist

Drishti Point: Nasagrai or Nasagre (nose)

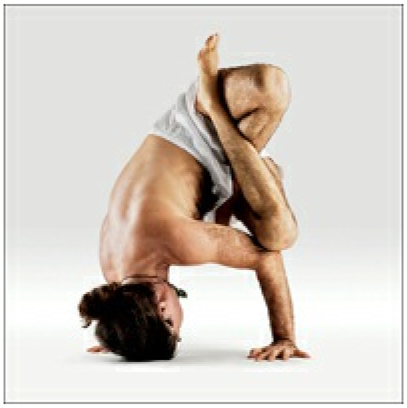

HEADSTAND 5 ARM POSITION: VARIOUS ARM POSITIONS BASED ON HEADSTAND 5 & VARIOUS LEG POSITIONS

Baby Cradle Pose in Headstand 5

Hindolasana in Shirshasana 5

(hin-do-LAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: Side 1: ankle to the inside of the elbow joint of the opposite arm, knee bent to the inside of the same arm, toes touching the floor.
Side 2: other knee bent to the side, heel toward the sitting bone
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

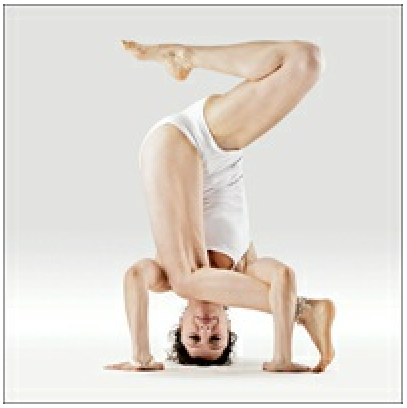

Baby Cradle Pose in Headstand 5

Hindolasana in Shirshasana 5

(hin-do-LAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: Side 1: arm bent at 90 degrees, fingertips to the floor, knee to the elbow on the same side.
Side 2: arm bent at 90 degrees, palm flat on the floor, ankle to the inside of the elbow joint, toes lifted off the floor
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

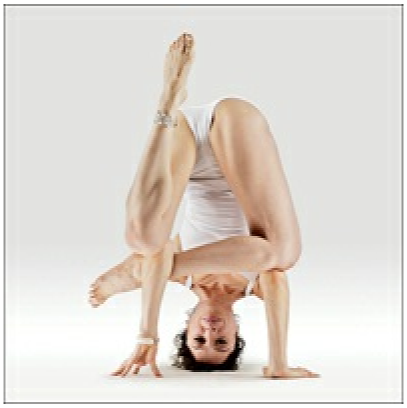

Extended Hand to Big Toe Pose in One Hand Headstand 5

Utthita Hasta Padangushtasana in Eka Hasta Shirshasana 5

(UT-ti-tuh HUH-stuh puhd-ahng-goosh-TAHS-uh-nuh in EY-kuh HUH-stuh sheer-SHAHS-uh-nuh)

Modification: Side 1: arm and leg straight out, grabbing onto the big toe.
Side 2: arm bent 90 degrees, knee resting on the tricep on the same side
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

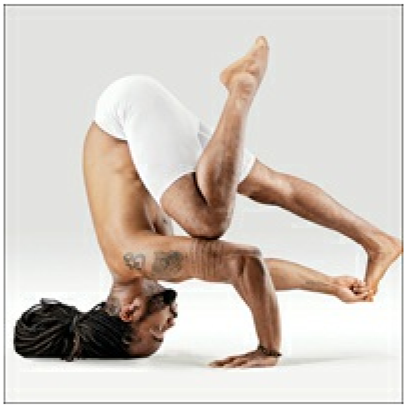

One-Legged Crane Pose 1 in One Hand Headstand 5

Eka Pada Bakasana 1 in Eka Hasta Shirshasana 5

(EY-kuh PUH-duh buh-KAHS-uh-nuh in EY-kuh HUH-stuh sheer-SHAHS-uh-nuh)

Modification: Side 1: arm and leg straight out.
Side 2: arm bent 90 degrees, knee resting on the tricep on the same side
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

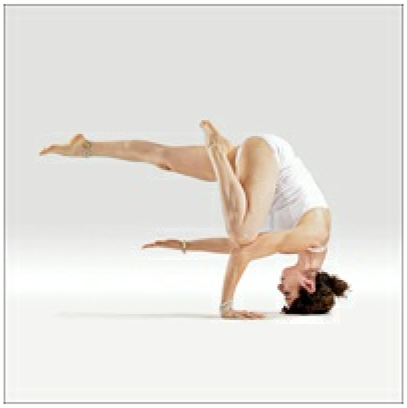

HEADSTAND 6 ARM POSITION: VARIOUS ARM POSITIONS BASED OFF HEADSTAND 6 & VARIOUS LEG POSITIONS

Extended Hand to Big Toe Pose in Uneven Headstand (Fusion of Headstand 5 and Headstand 6 Arm Position)

Utthita Hasta Padangushtasana in Vishama Shirshasana 5 & 6

(UT-ti-tuh HUH-stuh puhd-ahng-goosh-TAHS-uh-nuh in VISH-uh-muh sheer-SHAHS-uh-nuh)

Modification: Side 1: arm and leg straight out, grabbing onto the big toe
Side 2: arm bent 90 degrees, knee bent to the tricep on the same side
Pose Type: inversion, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Hastagrai or Hastagre (hands)

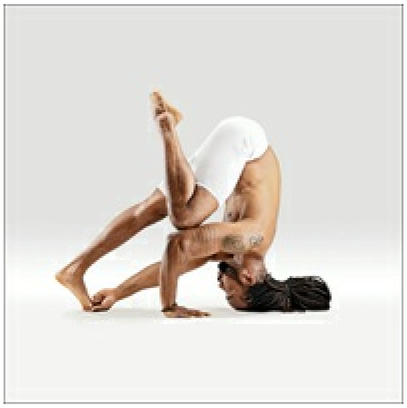

Crane Pose in Uneven Headstand (Fusion of Headstand 5 and Headstand 6 Arm Position)

Bakasana in Vishama Shirshasana 5 & 6

(buh-KAHS-uh-nuh in VISH-uh-muh sheer-SHAHS-uh-nuh)

Modification: Side 1: arm straight out, fingertips to the floor, knee to the tricep.
Side 2: arm bent 90 degrees, knee bent to the tricep on the same side
Pose Type: inversion, forward bend
Drishti Point: Hastagrai or Hastagre (hands)

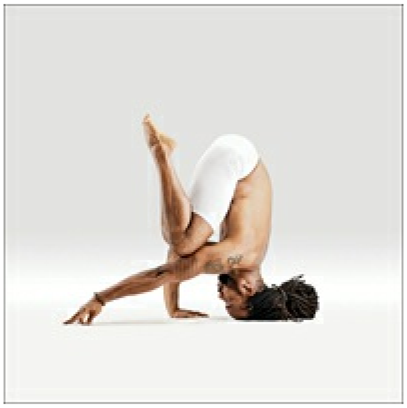

Headstand 6

Shirshasana 6

(sheer-SHAHS-uh-nuh)

Also Known As: Hands Free Headstand (Mukta Hasta Shirshasana)
Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

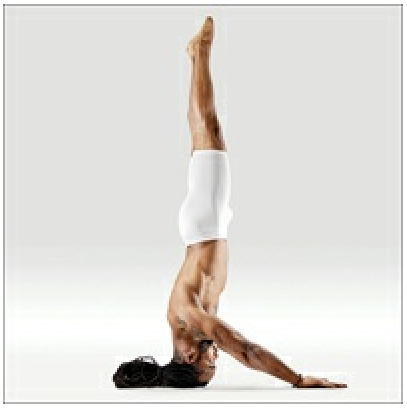

Upward Staff Pose in Headstand 6

Urdhva Dandasana in Shirshasana 6

(OORD-vuh duhn-DAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Modification: on the fingertips
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

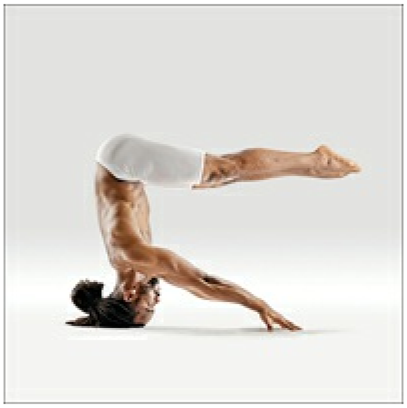

HEADSTAND 7 ARM POSITION: VARIOUS ARM POSITIONS BASED OFF HEADSTAND 7 & VARIOUS LEG POSITIONS

Headstand 7A

Shirshasana 7A

(sheer-SHAHS-uh-nuh)

Also Known As: Spread Hands Headstand (Prasarita Hasta Shirshasana), Hands Free Headstand C (Mukta Hasta Shirshasana C)
Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

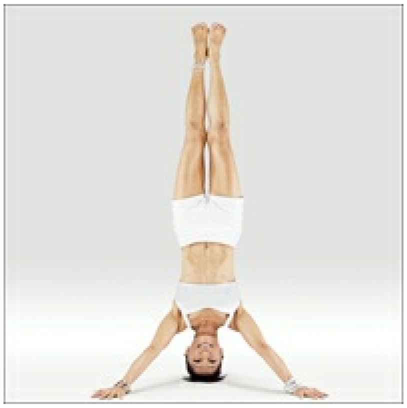

Bound Angle Pose in Headstand 7B

Baddha Konasana in Shirshasana 7B

(BUH-duh ko-NAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Also Known As: Bound Angle Pose in Free Hands Headstand (Baddha Konasana in Mukta Hasta Shirshasana)
Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

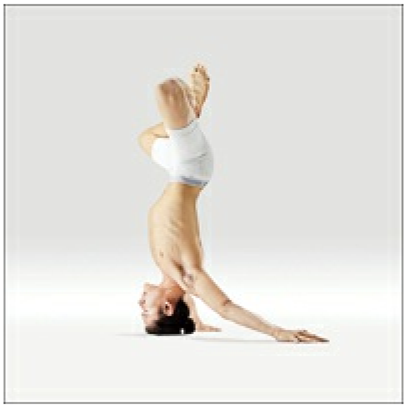

Leg Position of the Pose Dedicated to Garuda in Uneven Headstand (Fusion of Headstand 5 and Headstand 7 Arm Positions)

Pada Garudasana in Vishama Shirshasana 5 & 7A

(PUH-duh guh-ru-DAHS-uh-nuh in VISH-uh-muh sheer-SHAHS-uh-nuh)

Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

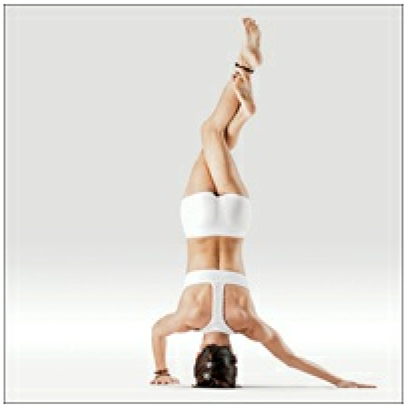

Upward Lotus Pose in Headstand 7B

Urdhva Padmasana in Shirshasana 7B

(OORD-vuh puhd-MAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Also Known As: Supported Lotus Headstand Pose (Salamba Padma Shirshasana), Upward Lotus Pose in Spread Hands Headstand (Urdhva Padmasana in Prasarita Hasta Shirshasana), Upward Lotus in Hands Free Headstand (Urdhva Padmasana in Mukta Hasta Shirshasana)

Modification: 1. neutral spine
2. backbend

Pose Type: 1. inversion
2. inversion, backbend

Drishti Point: Nasagrai or Nasagre (nose)

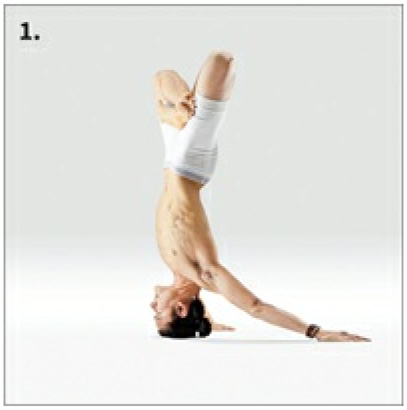

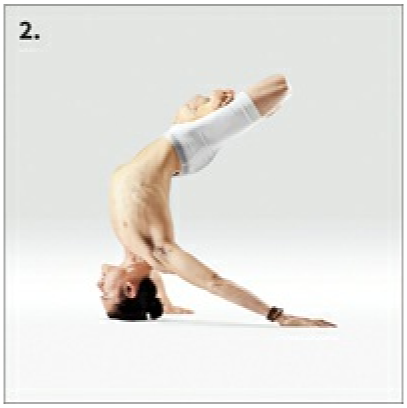

HEADSTAND 8 & 5 ARM POSITIONS: LEG STRAIGHT OUT TO THE SIDE

Leg to the Side Revolved Uneven Headstand (Fusion of Headstand 5 and Headstand 8 Arm Positions)

Parshva Pada Parivritta Vishama Shirshasana 5 & 8

(PAHRSH-vuh PUH-duh puh-ri-VRIT-tuh VISH-uh-muh sheer-SHAHS-uh-nuh)

Modification: Side 1: arm bent at 90 degrees, palm flat on the floor, knee resting on the opposite elbow
Side 2: arm bent at 90 degrees, fingertips to the floor, leg straight out to the side

Pose Type: inversion, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

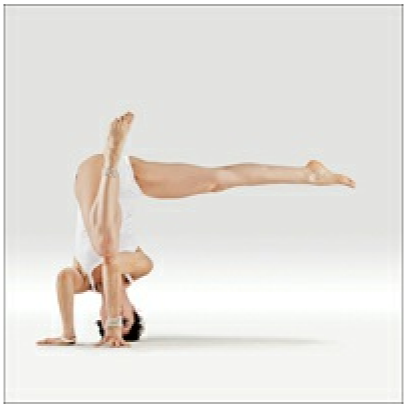

Leg to the Side One Hand Headstand 5

Parshva Pada Eka Hasta Shirshasana 5

(PAHRSH-vuh PUH-duh EY-kuh HUH-stuh sheer-SHAHS-uh-nuh)

Modification: Side 1: arm bent at 90 degrees, palm flat on the floor, knee resting on the same elbow
Side 2: arm straight, grabbing onto the shin of the leg, leg straight out to the side
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

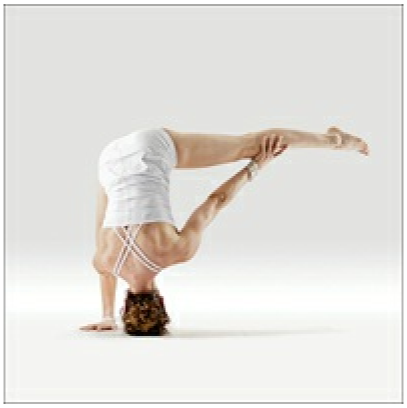

Leg to the Side Revolved One Hand Headstand 5

Parshva Pada Parivritta Eka Hasta Shirshasana 5

(PAHRSH-vuh PUH-duh puh-ri-VRIT-tuh EY-kuh HUH-stuh sheer-SHAHS-uh-nuh)

Modification: Side 1: arm bent at 90 degrees, palm flat to the floor, leg straight out to the side
Side 2: arm straight up to the sky, knee resting on the opposite tricep
Pose Type: inversion, forward bend, twist

Drishti Point: Hastagrai or Hastagre (hands), Nasagrai or Nasagre (nose)

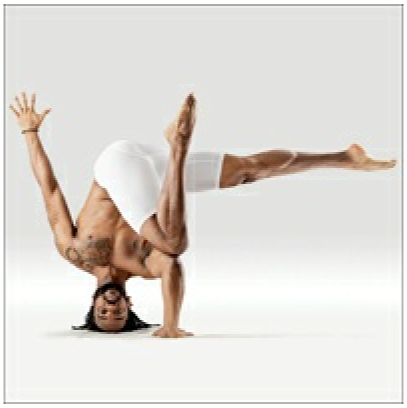

Leg to the Side Revolved One Hand Headstand 8

Parshva Pada Parivritta Eka Hasta Shirshasana 8

(PAHRSH-vuh PUH-duh puh-ri-VRIT-tuh EY-kuh HUH-stuh sheer-SHAHS-uh-nuh)

Modification: Side 1: arm bent at 90 degrees, fingertips to the floor, leg straight out to the side
Side 2: arm straight out to the side, knee resting on the opposite tricep
Pose Type: inversion, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands), Nasagrai or Nasagre (nose)

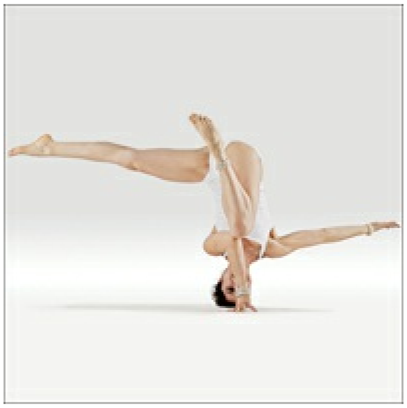

HEADSTAND 8 ARM POSITION: VARIOUS LEG POSITIONS

Feet Spread Intense Stretch Pose in Headstand 8

Prasarita Padottanasana in Shirshasana 8

(pruh-SAH-ri-tuh puh-do-tahn-AHS-uh-nuh in sheer-SHAHS-uh-nuh)

Also Known As: Feet Spread Full Forward Bend in Headstand 8
Modification: fingertips on the floor

Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

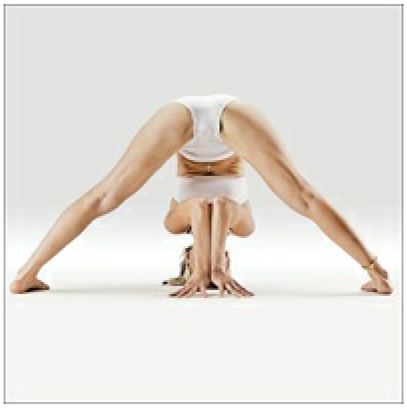

Revolved Leg Position of One-Legged King Pigeon 1 Version B in Headstand 8

Parivritta Pada Eka Pada Raja Kapotasana 1 B in Shirshasana 8

(puh-ri-VRIT-tuh PUH-duh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh in sheer-SHAHS-uh-nuh)
Modification: palms flat on the floor
Pose Type: inversion, forward bend, twist
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

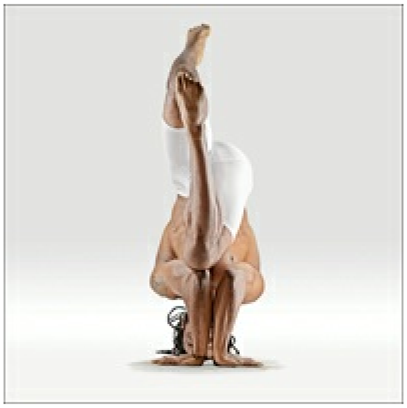

Headstand 8

Shirshasana 8

(sheer-SHAHS-uh-nuh)
Modification: knees resting on the triceps, ankles crossed
1. hands apart, elbows touching, fingertips to the floor
2. hands together, elbows together, palms flat on the floor

Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Hastagrai or Hastagre (hands)

1.

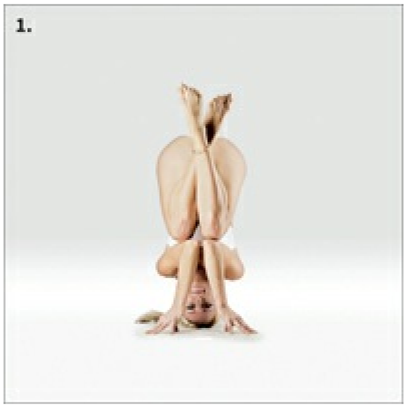

2.

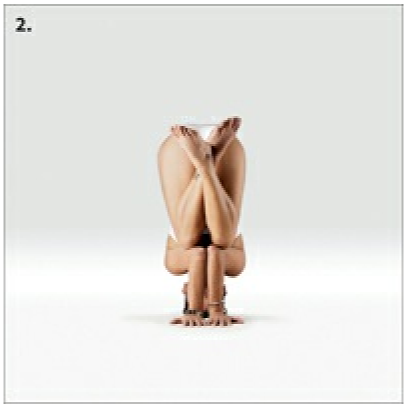

HEADSTAND 2 ARM POSITION: LOTUS

Upward Lotus Pose in Headstand 2

Urdhva Padmasana in Shirshasana 2

(OORD-vuh puhd-MAHS-uh-nuh in sheer-SHAHS-uh-nuh)

Also Known As: Upward Lotus in Hands Bound Headstand (Urdhva Padmasana in Baddha Hasta Shirshasana)
Pose Type: inversion
Drishti Point: Nasagrai or Nasagre (nose)

SHOULDERSTAND: BOTH LEGS STRAIGHT—DIFFERENT ARM POSITIONS

Inverted Pose Dedicated to Mythological Khimi Karani Pond

Viparita Khimi Karanyasana

(vi-puh-REE-tuh kuh-HEE-mee kuh-ruh-NEE-uhs-uh-nuh)

Also Known As: Inverted Lake Seal (Viparita Karani Mudra), Half Whole Body Pose (Ardha Sarvangasana)
Pose Type: inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

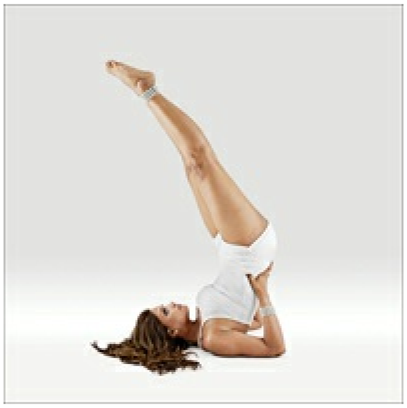

Supported Whole Body Pose

Salamba Sarvangasana

(SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Pose Type: inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

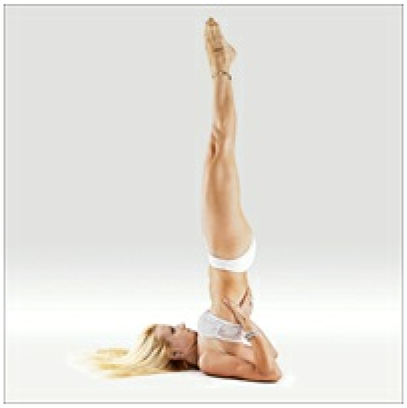

Hands Bound Supported Whole Body Pose

Baddha Hasta Salamba Sarvangasana

(BUH-duh HUH-stuh SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Pose Type: inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं यं हं ॐ

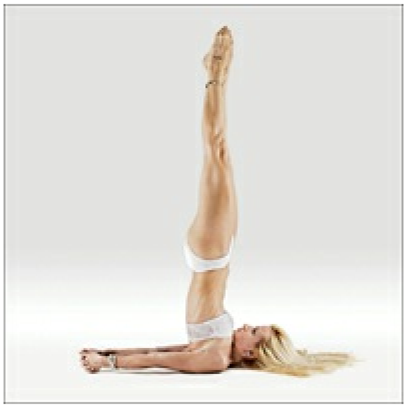

Unsupported Whole Body Pose

Niralamba Sarvangasana

(nir-AH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Pose Type: inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nabhi, Nabhicakre, or Nabi Chakra (belly button)
लं हं ॐ

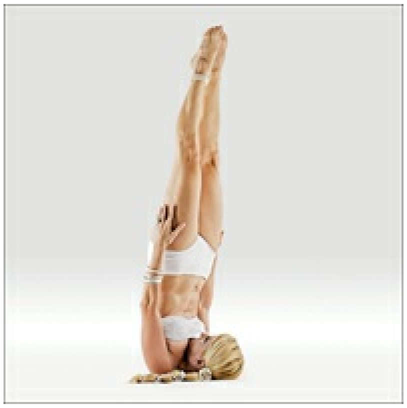

SHOULDERSTAND: ONE LEG UP, ONE LEG DOWN

Leg Contraction Pose in Supported Whole Body Pose

Pada Akunchanasana in Salamba Sarvangasana

(PUH-duh uh-kunch-AHS-uh-nuh in SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Modification: one knee bent toward the forehead
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

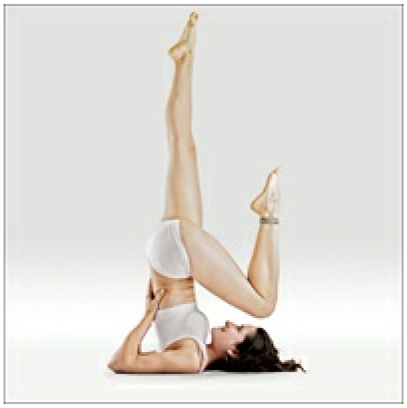

One-Legged Supported Whole Body Pose

Eka Pada Salamba Sarvangasana

(EY-kuh PUH-duh SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: One-Legged Plow Pose (Eka Pada Halasana), Shoulderstand
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

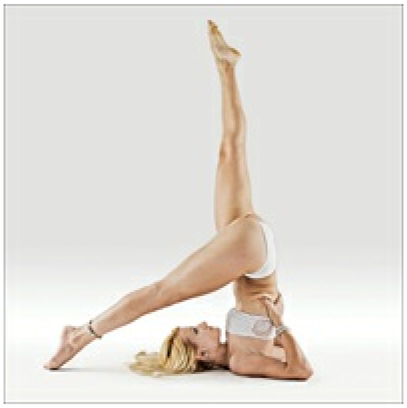

One-Legged Sideways Supported Whole Body Pose

Parshva Eka Pada Salamba Sarvangasana

(PAHRSH-vuh EY-kuh PUH-duh SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

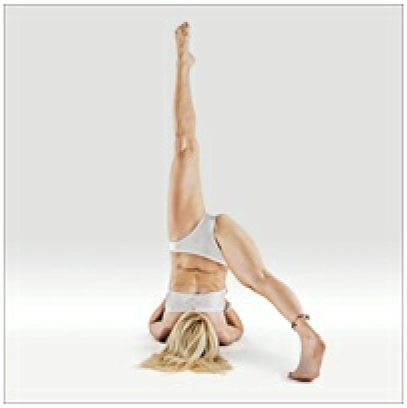

SHOULDERSTAND: ONE LEG UP, ONE LEG DOWN—DIFFERENT ARM POSITIONS

One-Legged Unsupported Whole Body Pose

Eka Pada Niralamba Sarvangasana

(EY-kuh PUH-duh nir-AH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Modification: both hands to the calf of the bottom leg
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

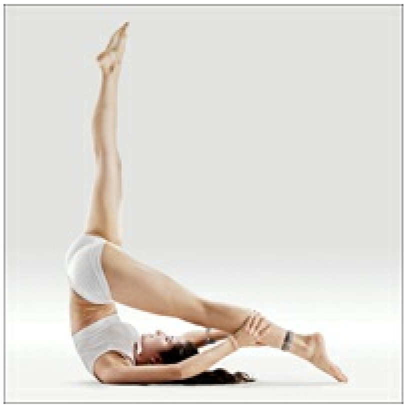

Extended Hand to Big Toe Pose in Unsupported Whole Body Pose

Utthita Hasta Padangushtasana in Niralamba Sarvangasana

(UT-ti-tuh HUH-stuh puhd-ahng-goosh-tahn-AHS-uh-nuh in nir-AH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं हं ॐ

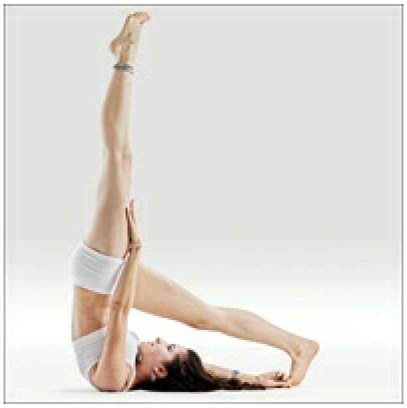

One-Legged Unsupported Whole Body Pose

Eka Pada Niralamba Sarvangasana

(EY-kuh PUH-duh nir-AH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Modification: both arms along the sides of the torso
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं हं ॐ

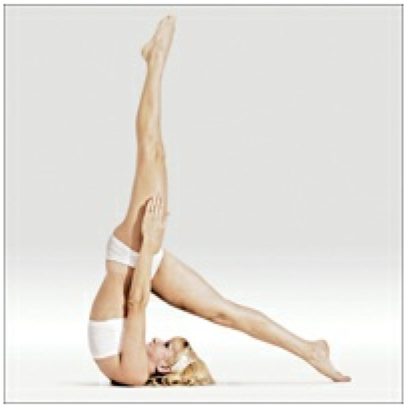

Half Bound Inverted Tortoise Pose

Ardha Baddha Viparita Kurmasana

(UHR-duh BUH-duh vi-puh-REE-tuh koor-MAHS-uh-nuh)

Pose Type: inversion, forward bend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

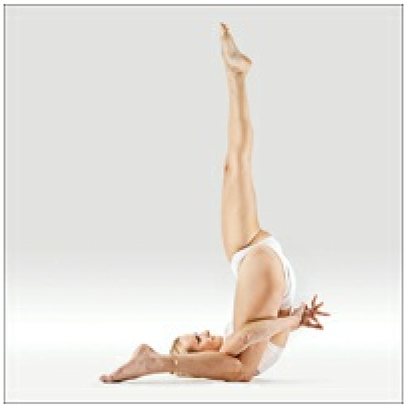

SHOULDERSTAND: GARUDA LEGS

Leg Position of the Pose Dedicated to Garuda in Hands Bound Supported Whole Body Pose

Pada Garudasana in Baddha Hasta Salamba Sarvangasana

(PUH-duh guh-ru-DAHS-uh-nuh in BUH-duh HUH-stuh SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Modification: knees toward the forehead
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

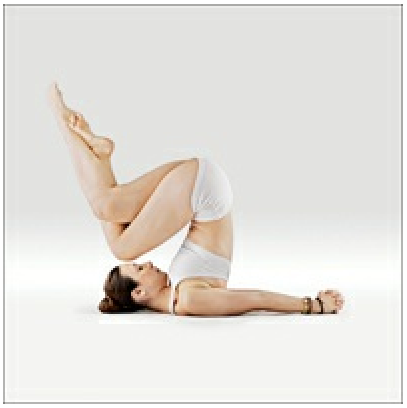

Leg Position of the Pose Dedicated to Garuda in Hands Bound Supported Whole Body Pose

Pada Garudasana in Baddha Hasta Salamba Sarvangasana

(PUH-duh guh-ru-DAHS-uh-nuh in BUH-duh HUH-stuh SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Modification: legs extended toward the sky
Pose Type: inversion, forward bend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

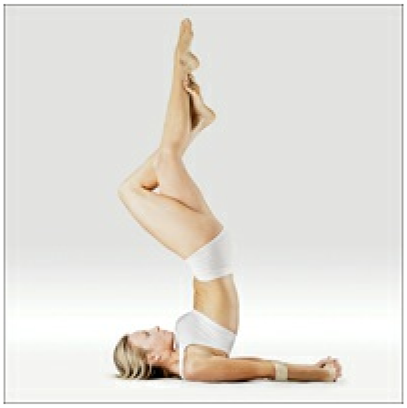

Leg Position of the Pose Dedicated to Garuda in Supported Whole Body Pose

Pada Garudasana in Salamba Sarvangasana

(PUH-duh guh-ru-DAHS-uh-nuh in SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Pose Type: inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

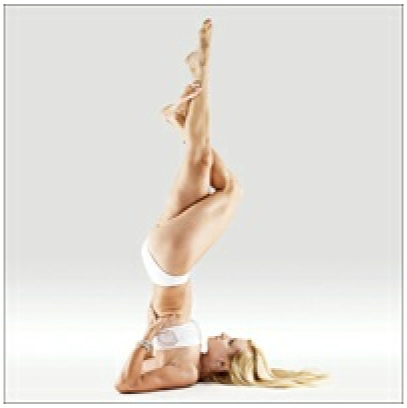

SHOULDERSTAND: BOUND ANGLE & LOTUS

Bound Angle Pose in Whole Body Pose

Baddha Konasana in Sarvangasana

(BUH-duh ko-NAHS-uh-nuh in suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Pose Type: inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

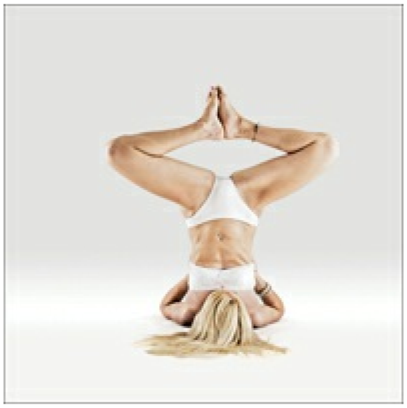

Upward Lotus in Whole Body Pose

Urdhva Padmasana in Salamba Sarvangasana

(OORD-vuh puhd-MAHS-uh-nuh in SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Pose Type: inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

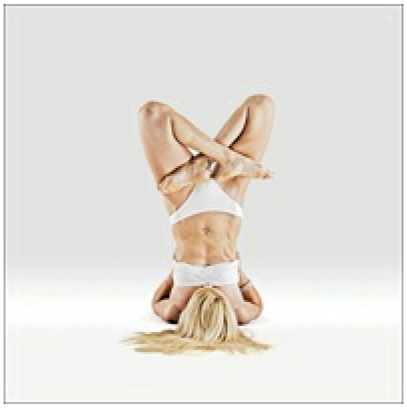

Upward Lotus in Whole Body Pose

Urdhva Padmasana in Salamba Sarvangasana

(OORD-vuh puhd-MAHS-uh-nuh in SAH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Modification: arms straight to the back, palms on the floor
Pose Type: inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

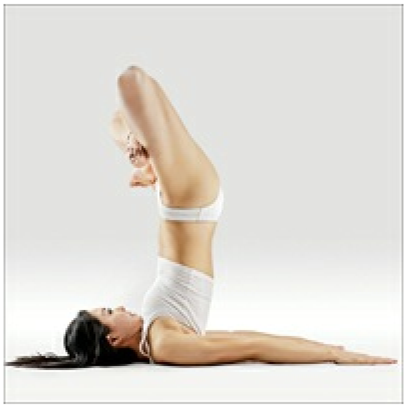

Upward Lotus in Unsupported Whole Body Pose

Urdhva Padmasana in Niralamba Sarvangasana

(OORD-vuh puhd-MAHS-uh-nuh in nir-AH-luhm-buh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Pose Type: inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

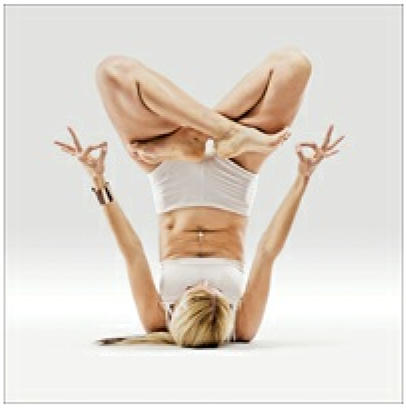

Inverted Ear Pressure Raised Up Lotus Pose

Viparita Karnapida Urdhva Padmasana

(vi-puh-REE-tuh kuhr-nah-PEED-uh OORD-vuh puhd-MAHS-uh-nuh)

Pose Type: inversion

Drishti Point: Nasagrai or Nasagre (nose)

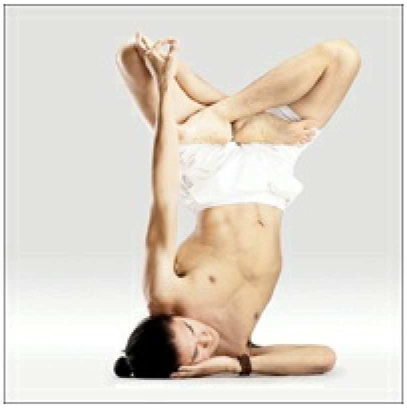

SHOULDERSTAND: LOTUS—BACKBEND & FORWARD BEND

Intense Lotus Peacock Prep.

Uttana Padma Mayurasana Prep.

(ut-TAHN-uh PUHD-muh muh-yoor-AHS-uh-nuh)

Also Known As: Intense Front Body Stretching and Rejuvenating Lotus Pose (Purvottana Padma Sarvangasana)

Modification: knees off the floor, fists to the lower back

Pose Type: inversion, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

Upward Lotus Pose in Whole Body Pose

Urdhva Padmasana in Sarvangasana

(OORD-vuh puhd-MAHS-uh-nuh in suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulderstand
Modification: 1. thighs parallel to the floor, elbows bent
2. arms straight
Pose Type: inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

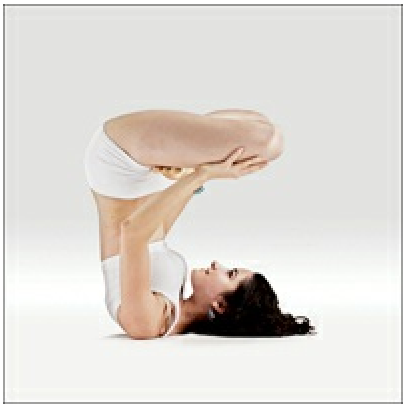

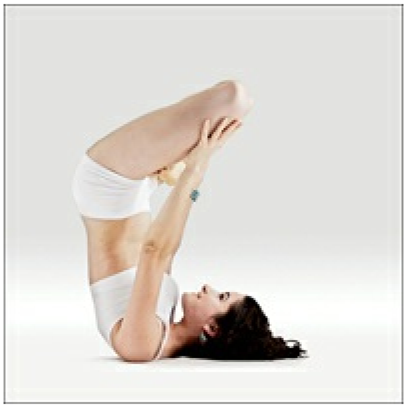

Embryo Pose in Whole Body Pose

Pindasana in Sarvangasana

(pin-DAHS-uh-nuh in suhr-vuhng-GAHS-uh-nuh)

Also Known As: Inverted Embryo Pose (Viparita Pindasana), Embryo Pose in Plow Pose (Pindasana in Halasana), Shoulderstand

Pose Type: inversion, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

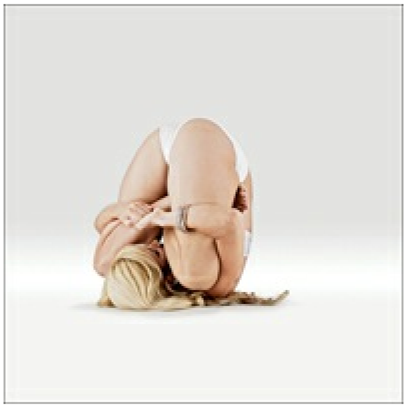

SHOULDERSTAND: LOTUS—KNEES ON THE FLOOR

Lotus Pose in Plow Pose

Padmasana in Halasana

(puhd-MAHS-uh-nuh in hul-AHS-uh-nuh)

Pose Type: inversion, forward bend

Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

ल व र य ह

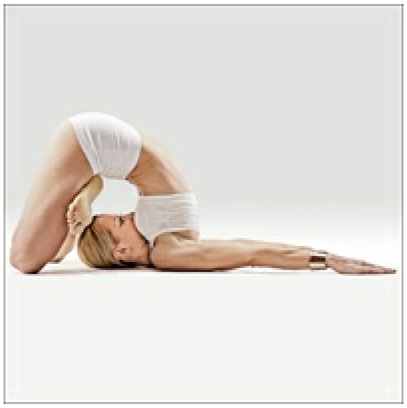

Sideways Embryo Pose in Whole Body Pose

Parshva Pindasana in Sarvangasana

(PAHRSH-vuh pin-DAHS-uh-nuh in suhr-vuhng-GAHS-uh-nuh)

Also Known As: Side Embryo Pose (Parshva Pindasana), Embryo Pose in Plow Pose (Pindasana in Halasana), Shoulderstand

Pose Type: inversion, forward bend, twist

Drishti Point: Nasagrai or Nasagre (nose)

ल व र य ह

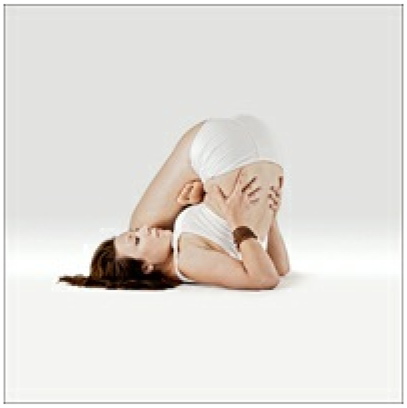

SHOULDERSTAND: EAR PRESSURE POSE—FEET ON THE FLOOR

Inverted Tortoise Pose

Viparita Kurmasana

(vi-puh-REE-tuh koor-MAHS-uh-nuh)

Modification: 1. grabbing onto the feet

2. palms to the floor, fingertips facing to the back

Pose Type: inversion, forward bend

Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

ल व र ह

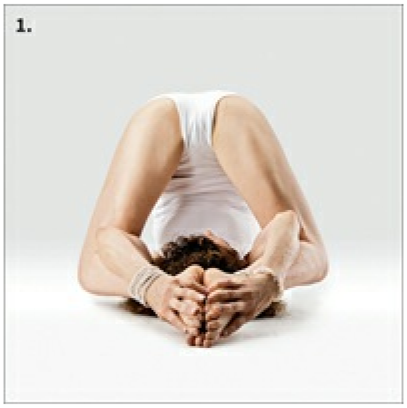

2.

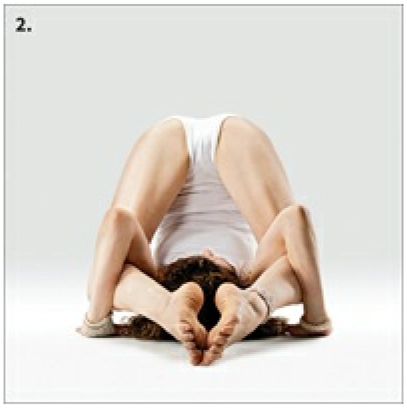

Bound Inverted Tortoise Pose

Baddha Viparita Kurmasana

(BUH-duh vi-puh-REE-tuh koor-MAHS-uh-nuh)

Pose Type: inversion, forward bend, binding
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)
लं वं रं हं

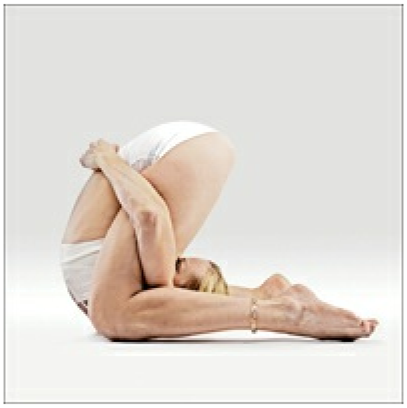

SHOULDERSTAND: EAR PRESSURE POSE—FEET OFF THE FLOOR

Ear Pressure Pose Prep.

Karnapidasana Prep.

(kuhr-nah-pee-DAHS-uh-nuh)

Modification: knees to the temples, palms to the lower back
Pose Type: inversion, forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)
लं रं हं ॐ

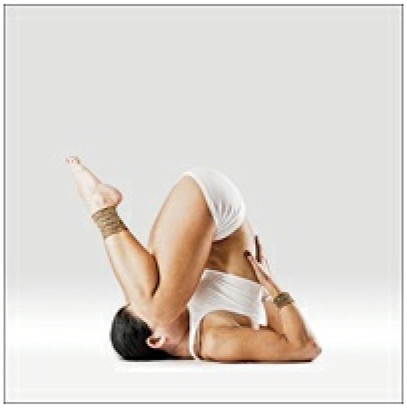

Ear Pressure Pose Prep.

Karnapidasana Prep.

(kuhr-nah-pee-DAHS-uh-nuh)

Modification: one hand to the lower back; other arm up over the head, elbow bent, fingertips to the floor; both knees bent, one knee to the elbow of the front arm, heel of the other foot toward the sitting bone
Pose Type: inversion, forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)
लं रं हं

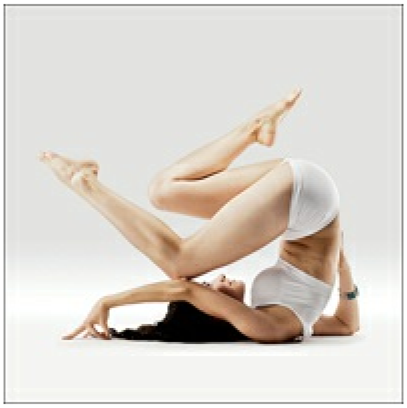

Ear Pressure Pose

Karnapidasana

(kuhr-nah-pee-DAHS-uh-nuh)

Modification: both knees bent, feet off the floor; arms straight behind the back and off the floor, palms up
Pose Type: inversion, forward bend, balance
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)
लं रं हं

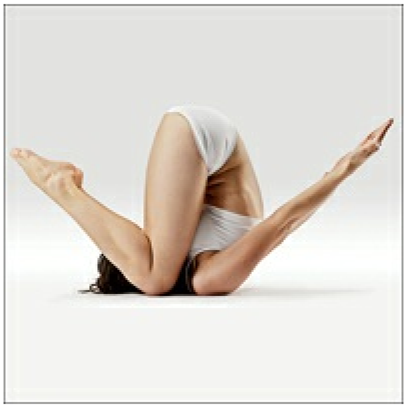

SHOULDERSTAND: EAR PRESSURE POSE—HEELS TO THE SITTING BONES, GRABBING ONTO THE SHINS

Ear Pressure Pose

Karnapidasana

(kuhr-nah-pee-DAHS-uh-nuh)

Modification: both heels to the sitting bones, grabbing onto the ankles, elbows bent
Pose Type: inversion, forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

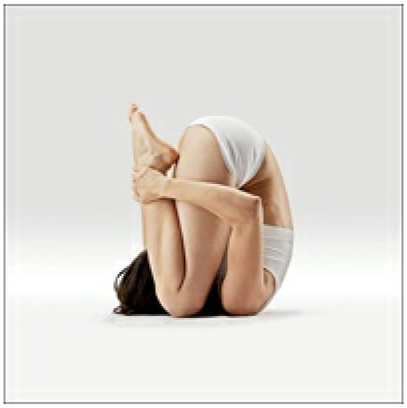

SHOULDERSTAND: EAR PRESSURE POSE—KNEES ON THE FLOOR—DIFFERENT ARM POSITIONS

Ear Pressure Pose

Karnapidasana

(kuhr-nah-pee-DAHS-uh-nuh)

Modification: palms to the lower back
Pose Type: inversion, forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

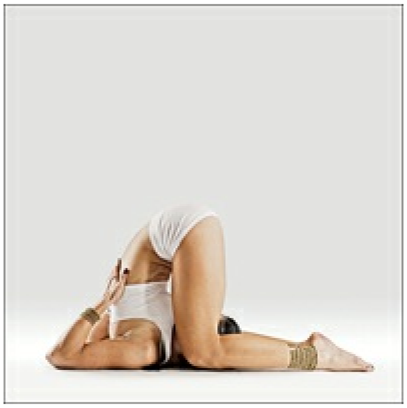

Sideways Ear Pressure Pose

Parshva Karnapidasana

(PAHRSH-vuh kuhr-nah-pee-DAHS-uh-nuh)

Also Known As: Side Contraction Pose (Parshva Akunchanasana)
Modification: palms to the lower back
Pose Type: inversion, forward bend, twist
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

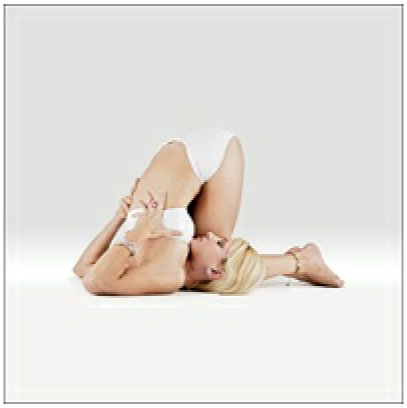

Hands Bound Ear Pressure Pose

Baddha Hasta Karnapidasana

(BUH-duh HUH-stuh kuhr-nah-pee-DAHS-uh-nuh)

Modification: fingers interlocked
Pose Type: inversion, forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

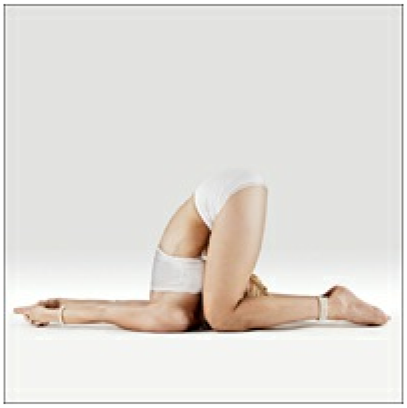

Shivalinga Pose

Lingasana

(ling-GAHS-uh-nuh)

Also Known As: Ear Pressure Pose (Karnapidasana)
Pose Type: inversion, forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

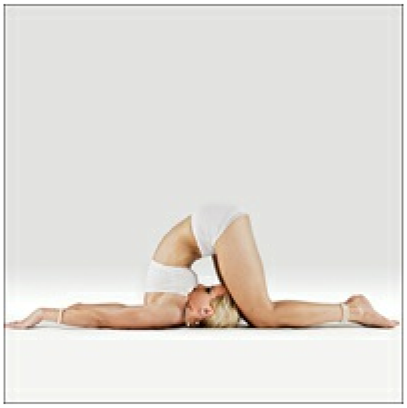

Bound Hands Shivalinga Pose

Baddha Hasta Lingasana

(BUH-duh HUH-stuh ling-GAHS-uh-nuh)

Also Known As: Ear Pressure Pose (Karna Pidasana)
Pose Type: inversion, forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

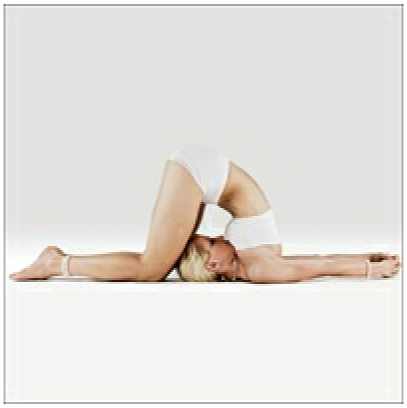

Ear Pressure Pose

Karnapidasana

(kuhr-nah-pee-DAHS-uh-nuh)

Modification: grabbing onto the heels
Pose Type: inversion, forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)
लं रं हं

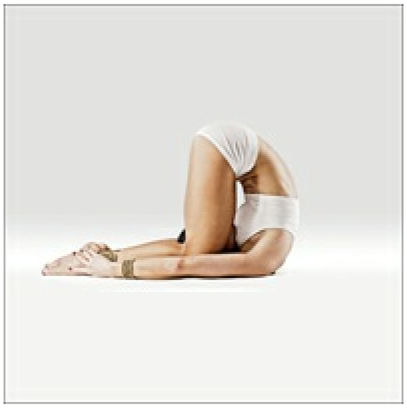

PLOW POSE: LEGS STRAIGHT AND TOGETHER—DIFFERENT ARM POSITIONS

Plow Pose

Halasana

(huh-LAHS-uh-nuh)

Modification: palms to the lower back, toes pointed to the front
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं यं हं ॐ

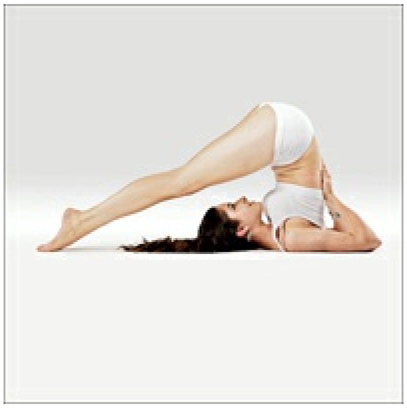

Plow Pose

Halasana

(huh-LAHS-uh-nuh)

Modification: fingers to the toes, palms up
Pose Type: inversion, forward bend
Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)
लं रं हं

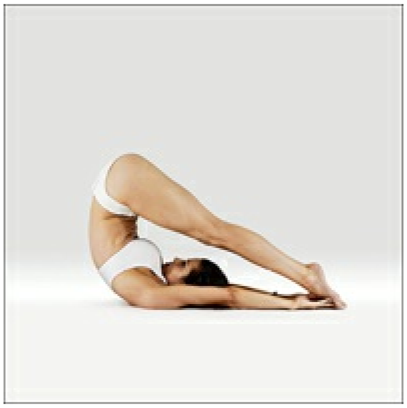

Plow Pose

Halasana

(huh-LAHS-uh-nuh)

Modification: arms straight to the back, palms on the floor
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं यं हं ॐ

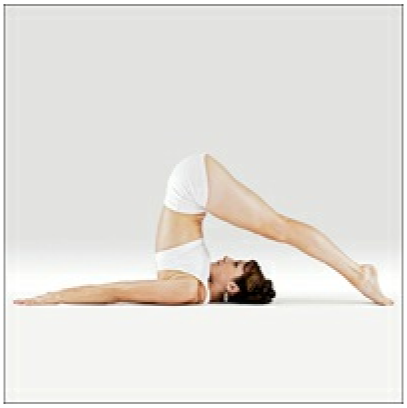

Bound Hands Plow Pose

Baddha Hasta Halasana

(BUH-duh HUH-stuh huh-LAHS-uh-nuh)

Pose Type: inversion, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ● ● ●

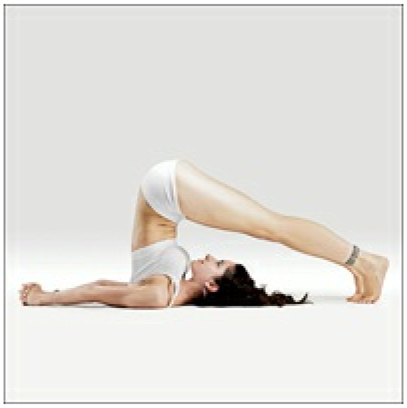

PLOW POSE: LEGS TO THE SIDE & LEGS WIDE APART

Sideways Plow Pose

Parshva Halasana

(PAHRSH-vuh huh-LAHS-uh-nuh)

Modification: palms to the lower back

Pose Type: inversion, forward bend, twist

Drishti Point: Nabhi, Nabhicakre, or Nabi Chakra (belly button)

● ● ● ●

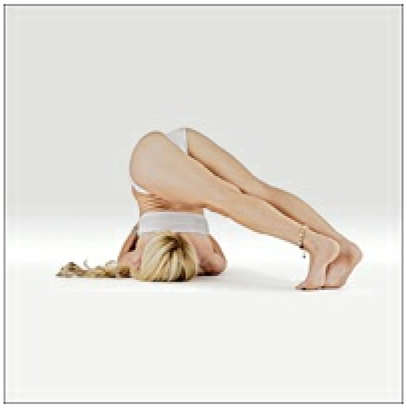

Hands Bound Feet Spread Wide Intense Stretch Pose in Plow Pose

Baddha Hasta Prasarita Padottanasana in Halasana

(BUH-duh HUH-stuh pruh-SAH-ri-tuh puh-do-tahn-AHS-uh-nuh in huh-LAHS-uh-nuh)

Also Known As: Hands Bound Feet Spread Wide Full Forward Bend in Plow Pose
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

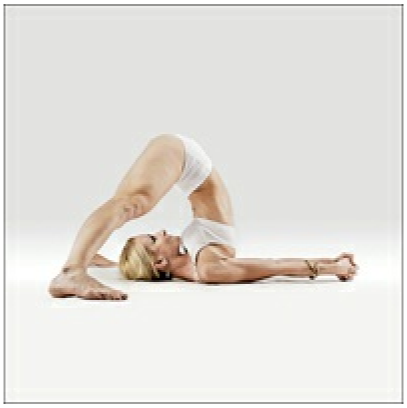

PLOW POSE: ONE LEG STRAIGHT, ONE LEG BENT

One-Legged Elbow to Knee Plow Pose

Eka Pada Kurpara Janu Halasana

(EY-kuh PUH-duh kuhr-PAH-ruh JAH-nu hul-AHS-uh-nuh)

Modification: one hand to the lower back; other arm up over the head, elbow bent, fingertips to the floor; knees together, one leg straight; knee to the elbow, other knee bent, heel to the sitting bone
Pose Type: inversion, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

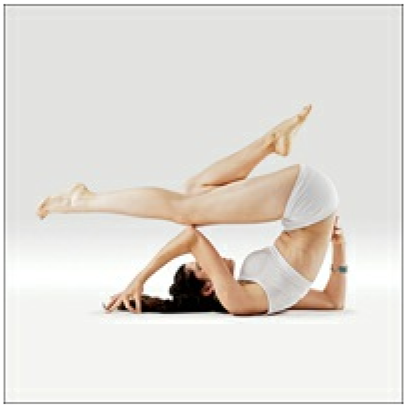

Reclined Upward Foot Thunderbolt Pose

Supta Urdhva Pada Vajrasana

(SUP-tuh OORD-vuh PUH-duh vuhj-RAHS-uh-nuh)

Also Known As: Half Bound Lotus Pose in Whole Body Pose (Ardha Baddha Padmasana in Sarvangasana), Shoulderstand
Modification: 1. front side view
2. back side view
Pose Type: inversion, forward bend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

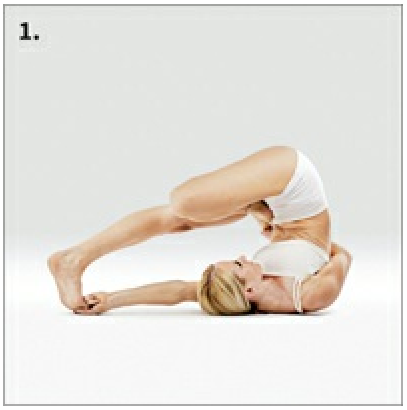
1.

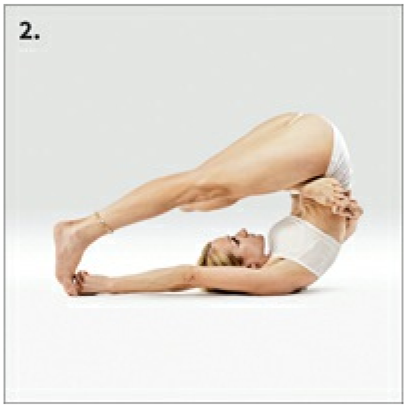
2.

Prone Poses

STOMACH ON THE FLOOR

CHEST FACING THE FLOOR

LOCUST POSE

SPHINX POSE

COBRA POSE

KING PIGEON POSE

UPWARD DOG POSE

COBRA POSE 2

BOW POSE

HALF LOTUS AND LOTUS BACKBENDS

SAGE GHERANDA’S POSE

BACKBENDS

CHIN AND CHEST ON THE FLOOR

CHEST AND KNEES ON THE FLOOR

Pose Dedicated to Makara—Prone Modification

Makarasana

(muh-kuh-RAHS-uh-nuh)

Also Known As: Crocodile Pose
Modification: whole body flat on the floor; arms straight out in front, palms together
Pose Type: prone
Drishti Point: Nasagrai or Nasagre (nose)
🔴 🟠 🟡 🟢 🔵

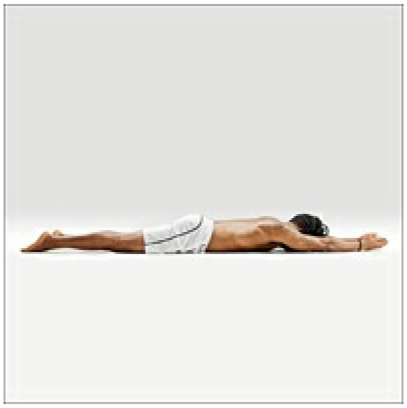

Side Corpse Pose

Parshva Shavasana

(PAHRSH-vuh shuh-VAHS-uh-nuh)

Pose Type: prone
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
🔴 🟠 🟡 🟢

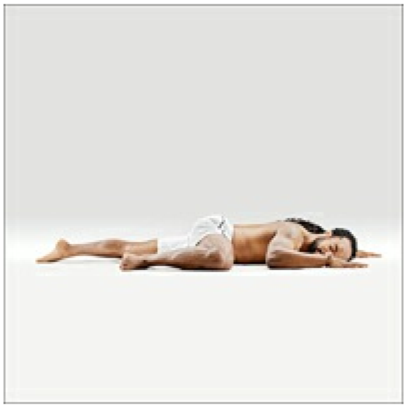

Downward Facing One Leg to the Side Pose

Adho Mukha Parshva Eka Padasana

(uh-DO MUK-uh PAHRSH-vuh EY-kuh puh-DAHS-uh-nuh)

Pose Type: prone

Drishti Point: Nasagrai or Nasagre (nose)

ल व र य ह

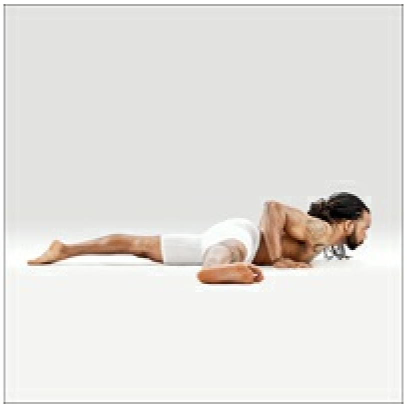

STOMACH ON THE FLOOR: ONE LEG STRAIGHT, ONE KNEE BENT—FOOT TO THE FLOOR ON THE SIDE

Intense Stretch Slithering Lizard Lunge Pose

Utthana Sarpa Godhasana

(ut-TAHN-uh SUHR-puh go-DAHS-uh-nuh)

Also Known As: Extended Lizard Tail Lunge Pose (Uttana Pristhasana)

Modification: foot to the elbow crease

Pose Type: prone

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

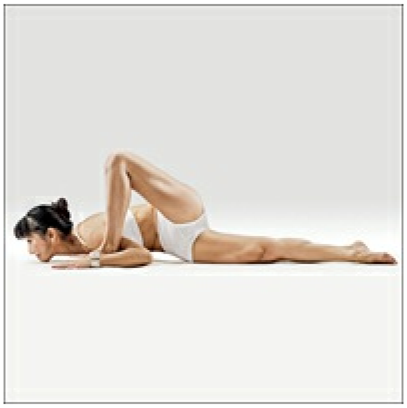

Intense Stretch Slithering Lizard Lunge Pose

Utthan Sarpa Godhasana

(ut-TAHN-uh SUHR-puh go-DAHS-uh-nuh)

Also Known As: Extended Lizard Tail Lunge Pose (Uttana Pristhasana)
Modification: foot to the outside of the rib cage, arm binds around the leg
Pose Type: prone
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

ल व र य ह ॐ

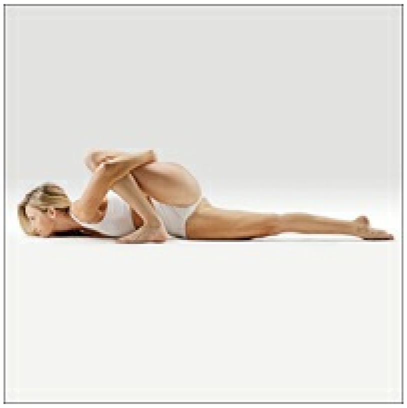

STOMACH ON THE FLOOR: FROG POSE & LOTUS POSE

Frog Pose

Mandukasana

(muhn-doo-KAHS-uh-nuh)

Also Known As: Thavaliasana
Pose Type: prone
Drishti Point: Nasagrai or Nasagre (nose)

ल व र य ह

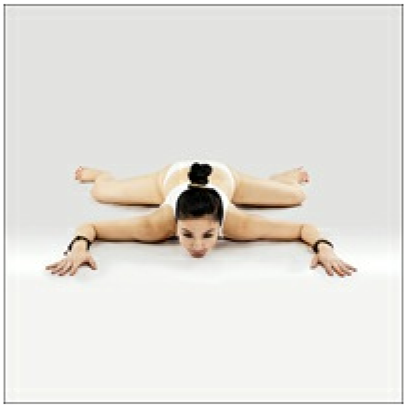

Lotus Staff Surrender Salutation Pose

Padma Danda Namaskarasana

(PUHD-muh DUHN-duh nuh-muhs-kahr-AHS-uh-nuh)

Modification: arms straight in front, palms down to the floor

Pose Type: prone

Drishti Point: Nasagrai or Nasagre (nose)

ल व र य ॐ

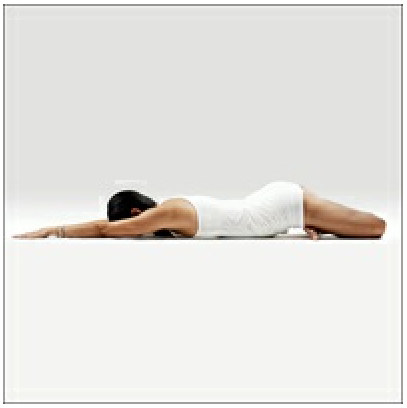

Hidden Lotus Pose

Gupta Padmasana

(GUP-tuh puhd-MAHS-uh-nuh)

Pose Type: prone, mild backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

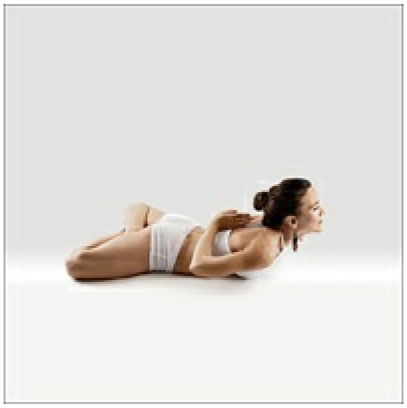

CHEST FACING THE FLOOR: TWISTS

Downward Facing Twisted Stomach Pose

Adho Mukha Jatara Parivartanasana

(uh-DO MUK-uh JAHT-uh-ruh puh-ri-vuhr-tuh-NAHS-uh-nuh)

Modification: knees together, legs bent
Pose Type: prone, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

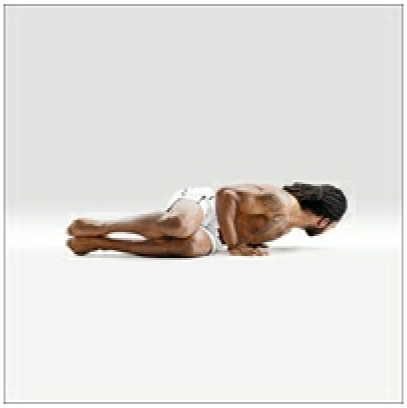

Downward Facing Twisted Stomach Pose

Adho Mukha Jatara Parivartanasana

(uh-DO MUK-uh JAHT-uh-ruh puh-ri-vuhr-tuh-NAHS-uh-nuh)

Modification: top leg straight, bottom knee bent
Pose Type: prone, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

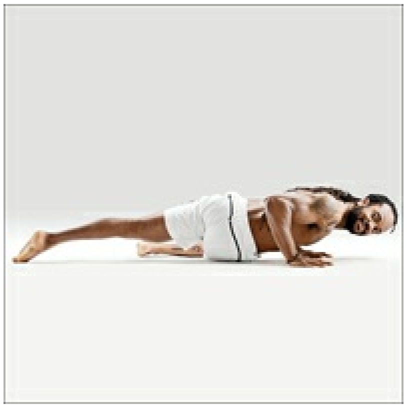

Downward Facing Twisted Stomach Pose

Adho Mukha Jatara Parivartanasana

(uh-DO MUK-uh JAHT-uh-ruh puh-ri-vuhr-tuh-NAHS-uh-nuh)

Modification: legs straight
Pose Type: prone, twist

Drishti Point: Nasagrai or Nasagre (nose)

ल

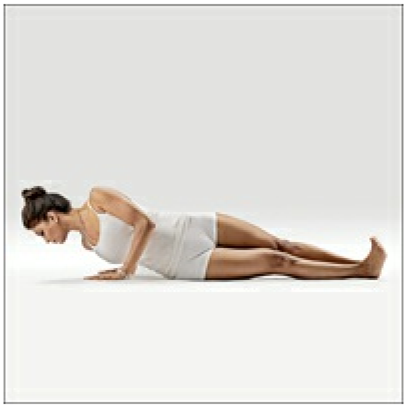

Pose Dedicated to St. Brighid of Kildare

Brighidasana

(bree-gid-AHS-uh-nuh)

Modification: bottom leg crossed under, looking straight ahead

Pose Type: prone, twist

Drishti Point: Nasagrai or Nasagre (nose)

ल व ह

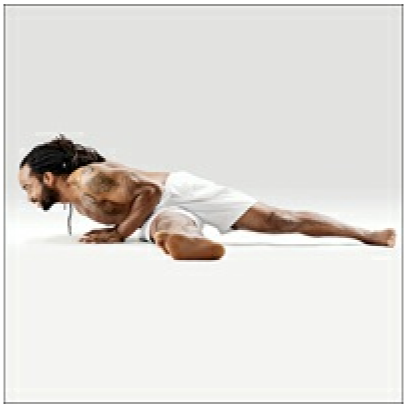

CHEST FACING THE FLOOR: TWISTS—LEGS TO THE SIDE

Pose Dedicated to Sage Koundinya One-Legged Version 1—Prone Modification

Eka Pada Koundinyasana 1

(EY-kuh PUH-duh kown-din-YAHS-uh-nuh)

Modification: one arm straight to the side, bottom knee bent

Pose Type: arm balance, forward bend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

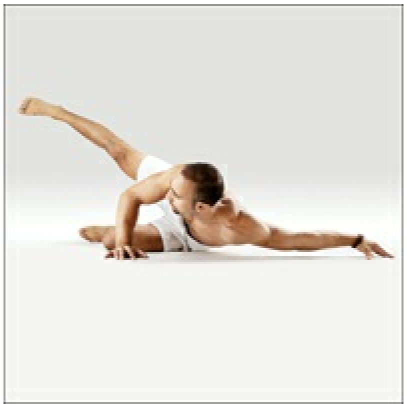

Pose Dedicated to Ashtavakra—Prone Modification

Ashtavakrasana

(uh-shtuh-vuh-KRAHS-uh-nuh)

Also Known As: Eight Angle Pose—Prone Modification
Modification: legs unhooked, bottom knee wrapped around the opposite forearm, top leg straight
Pose Type: arm balance, forward bend, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

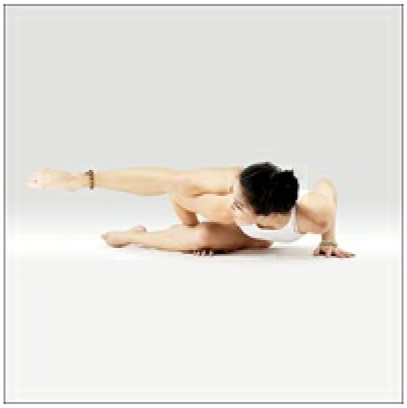

Pose Dedicated to Ashtavakra—Prone Modification

Ashtavakrasana

(uh-shtuh-vuh-KRAHS-uh-nuh)

Also Known As: Eight Angle Pose—Prone Modification
Pose Type: arm balance, forward bend, twist
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

ल व

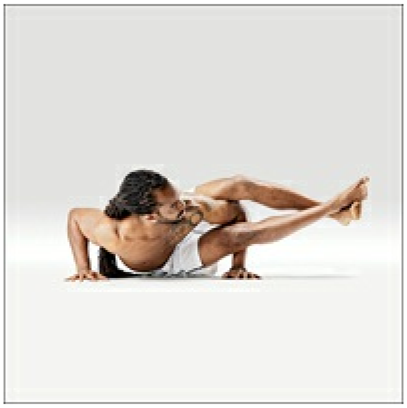

CHEST FACING THE FLOOR: TWIST & VERTICAL SPLITS

Svastika Legs in Downward Facing Twisted Stomach Pose

Pada Svastikasana in Adho Mukha Jatara Parivartanasana

(PUH-duh svuh-sti-KAHS-uh-nuh in uh-DO MUK-uh JAHT-uh-ruh puh-ri-vuhr-tuh-NAHS-uh-nuh)

Modification: one arm straight out in front, palm down to the floor; other elbow bent, temple resting on the forearm of the bent arm

Pose Type: prone, twist

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

ल व र य

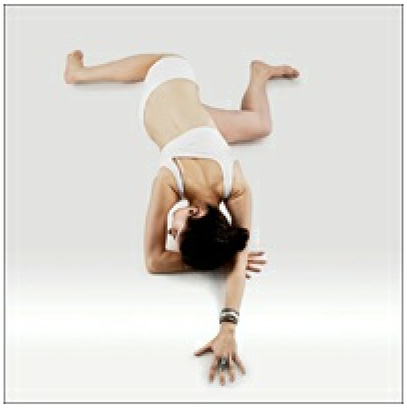

Downward Facing Pose Dedicated to Trivikrama—Prone Modification

Adho Mukha Trivikramasana

(uh-DO MUK-uh tri-vi-kruhm-AHS-uh-nuh)

Pose Type: prone

Drishti Point: Hastagrai or Hastagre (hands)

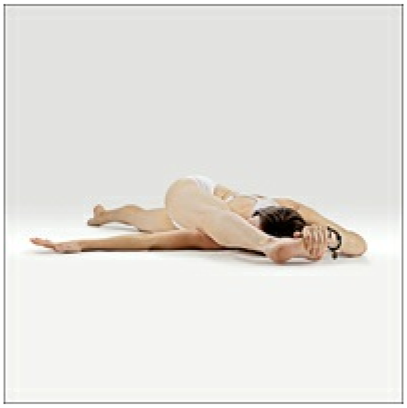

Bound Downward Facing Pose Dedicated to Trivikrama—Prone Modification

Baddha Adho Mukha Trivikramasana

(BUH-duh uh-DO MUK-uh tri-vi-kruhm-AHS-uh-nuh)

Modification: back leg bent—toes pointing to the sky
Pose Type: prone, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

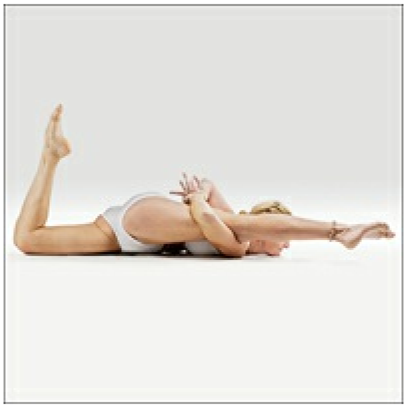

LOCUST POSE: LEGS ON THE FLOOR, ARMS BEHIND

Cobra Pose 1

Bhujangasana 1

(buj-uhng-GAHS-uh-nuh)

Modification: palms lifted off the floor, elbows bent
Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

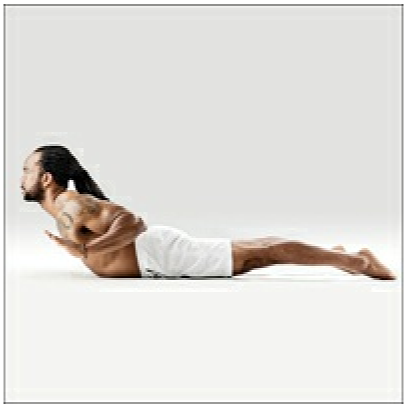

Locust Pose

Shalabhasana

(shuh-luh-BAHS-uh-nuh)

Modification: arms straight, palms facing up by the hips; feet on the floor
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

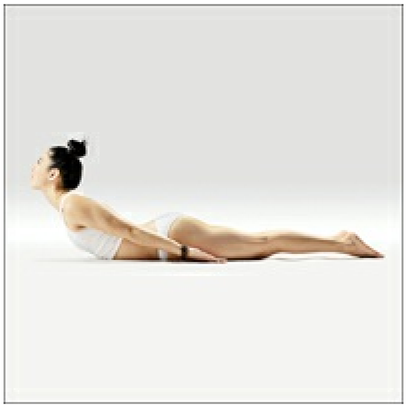

LOCUST POSE: LEGS AND CHEST OFF THE FLOOR—ARMS BEHIND ON THE FLOOR

Locust Pose

Shalabhasana

(shuh-luh-BAHS-uh-nuh)

Also Known As: Locust Pose B (Shalabhasana B)
Modification: palms to the floor by the bottom of the ribs, elbows bent at 90 degrees
Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

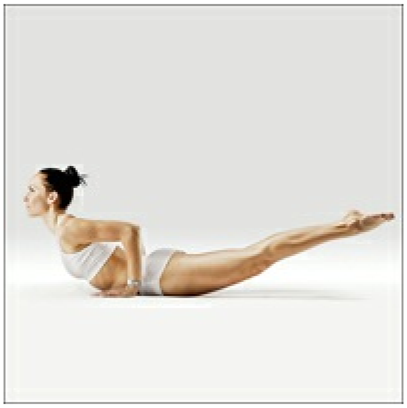

Locust Pose

Shalabhasana

(shuh-luh-BAHS-uh-nuh)

Also Known As: Locust Pose A (Salabhasana A)
Modification: palms to the floor by the hips, arms straight, legs and chest lifted
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

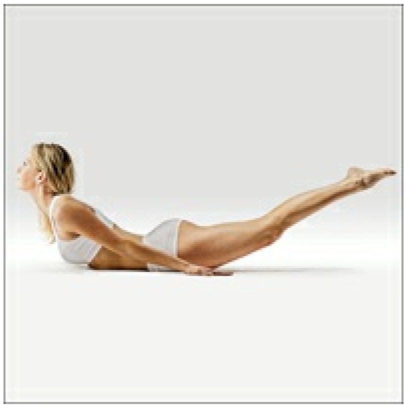

LOCUST POSE: LEGS AND CHEST OFF THE FLOOR—ARMS BEHIND

Fingers to Head Reverse Prayer Locust Pose

Anguli Shirsha Viparita Namaskar Shalabhasana

(UHNG-goo-lee SHEER-shuh vi-puh-REE-tuh nuh-muhs-KAHR shuh-luh-BAHS-uh-nuh)

Also Known As: Back of the Body Prayer Fingers to Head Locust Pose (Paschima Namaskara Anguli Shirsha Shalabhasana)

Modification: feet on the floor
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

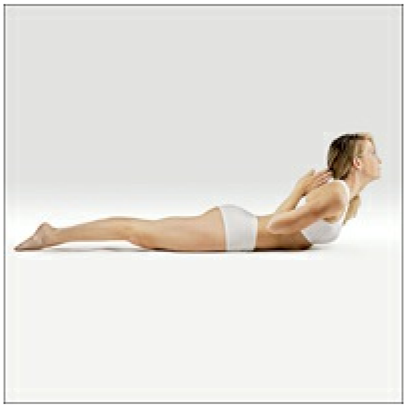

Pose Dedicated to Makara

Makarasana

(muh-kuh-RAHS-uh-nuh)
Also Known As: Crocodile Pose
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

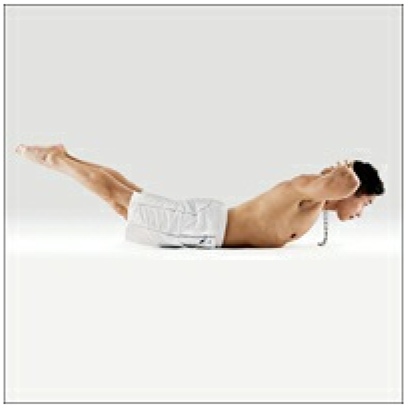

Locust Pose

Shalabhasana

(shuh-luh-BAHS-uh-nuh)
Modification: arms open wide, fingertips pointing to the toes
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

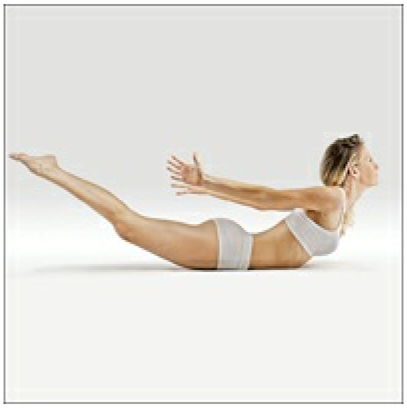

Hands Bound Locust Pose

Baddha Hasta Shalabhasana

(BUH-duh HUH-stuh shuh-luh-BAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

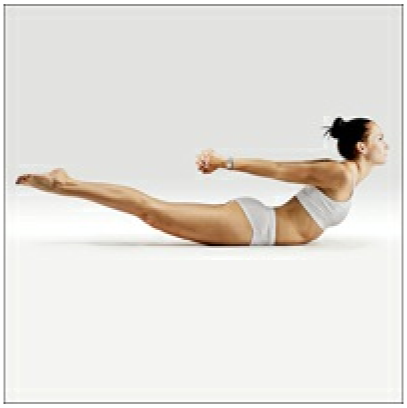

LOCUST POSE: LEGS AND CHEST OFF THE FLOOR—ARMS IN FRONT

Locust Pose

Shalabhasana

(shuh-luh-BAHS-uh-nuh)

Also Known As: Crocodile Pose (Makarasana), Boat Pose (Navasana)
Modification: both arms straight in front, palms facing down
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Angushtamadhye or Angushta Ma Dyai (thumbs)

Hand Position of the Pose Dedicated to Garuda in Locust Pose

Hasta Garudasana in Shalabhasana

(HUH-stuh guh-ru-DAHS-uh-nuh in shuh-luh-BAHS-uh-nuh)

Pose Type: prone, backbend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

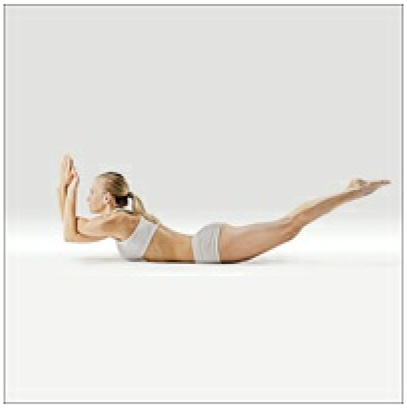

Inverted Locust Pose

Viparita Shalabhasana

(vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Modification: chin to the floor, legs lifted

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

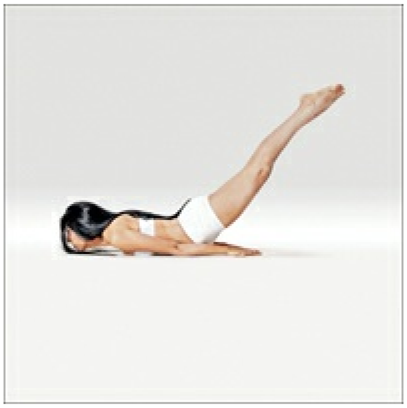

LOCUST POSE: ONE LEG

Half Locust Pose

Ardha Shalabhasana

(UHR-duh shuh-luh-BAHS-uh-nuh)

Modification: hands to the floor by the hips, palms facing up, arms straight
Pose Type: prone, backbend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

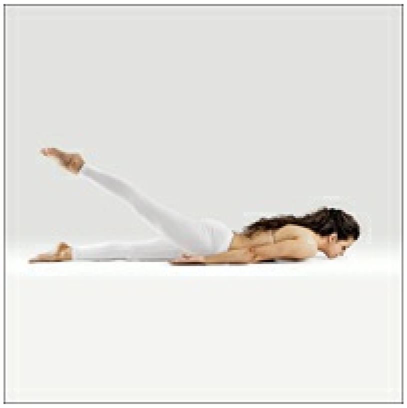

Hand Position of the Pose Dedicated to Garuda in Half Locust Pose

Hasta Garudasana in Ardha Shalabhasana

(HUH-stuh guh-ruh-DAHS-uh-nuh in UHR-duh shuh-luh-BAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व र य ह

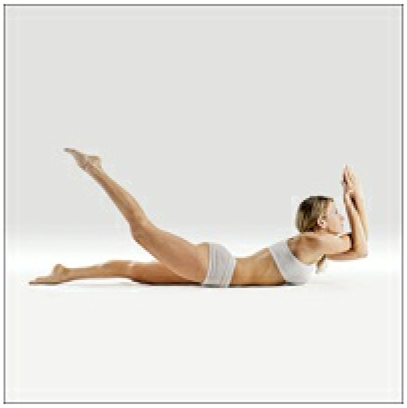

Half Locust Pose

Ardha Shalabhasana

(UHR-duh shuh-luh-BAHS-uh-nuh)

Also Known As: Crocodile Pose (Makarasana)
Modification: one arm straight to the front—palm facing down, other hand to the floor by the hip—palm down
Pose Type: prone, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
लं वं रं यं हं

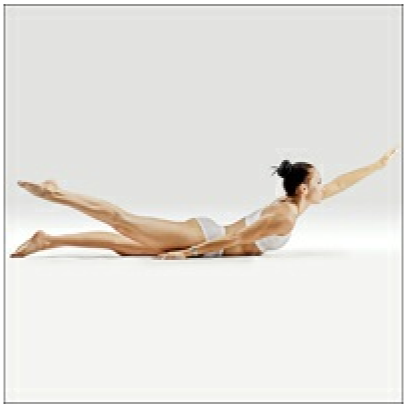

One-Legged Frog Pose in Locust Pose

Eka Pada Bhekasana in Shalabhasana

(EY-kuh PUH-duh bey-KAHS-uh-nuh in shuh-luh-BAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Hastagrai or Hastagre (hands)
लं वं रं यं हं ॐ

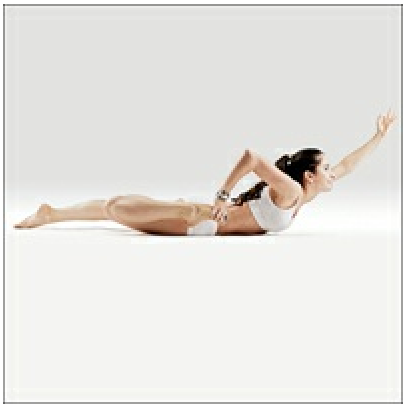

Mermaid Pose in Locust Pose

Naginyasana in Shalabhasana

(nuh-gin-YAHS-uh-nuh in shuh-luh-BAHS-uh-nuh)

Pose Type: prone, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

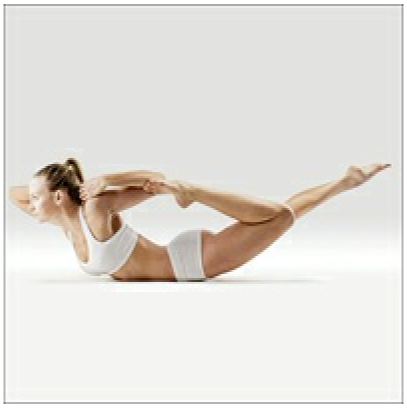

SPHINX POSE: BOTH AND ONE LEG STRAIGHT

Supported Cobra Pose

Salamba Bhujangasana

(SAH-luhm-buh buj-uhng-GAHS-uh-nuh)

Also Known As: Sphinx Pose or Crocodile Pose (Makarasana)
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

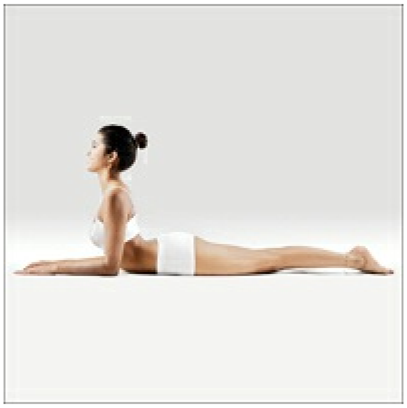

One-Legged Frog Pose

Eka Pada Bhekasana

(EY-kuh PUH-duh bey-KAHS-uh-nuh)

Modification: open ankles modification, heel to the floor by the hip socket
Pose Type: prone, backbend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ॐ

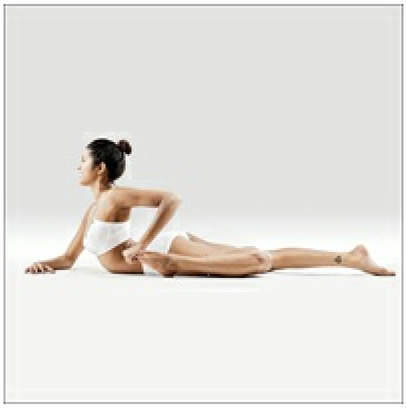

Leg Extended Uneven Supported Cobra Pose

Utthita Pada Vishama Salamba Bhujangasana

(UT-ti-tuh PUH-duh VISH-uh-muh SAH-luhm-buh buj-ung-GAHS-uh-nuh)

Also Known As: Uneven Sphinx Pose
Modification: 1. knee bent 2. leg straight
Pose Type: prone
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व ॐ

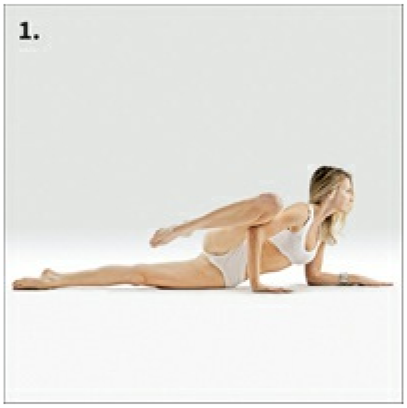
1.

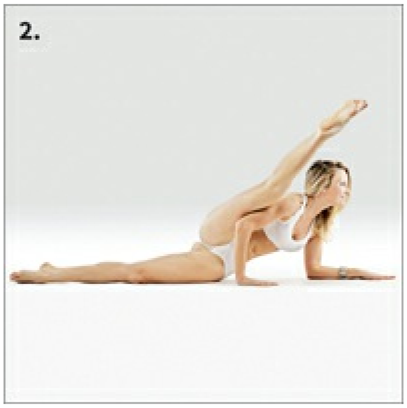
2.

SPHINX POSE: BOTH KNEES BENT

Supported Cobra Pose

Salamba Bhujangasana

(SAH-luhm-buh buj-uhng-GAHS-uh-nuh)

Also Known As: Sphinx Pose or Crocodile Pose (Makarasana)
Modification: knees bent
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ● ●

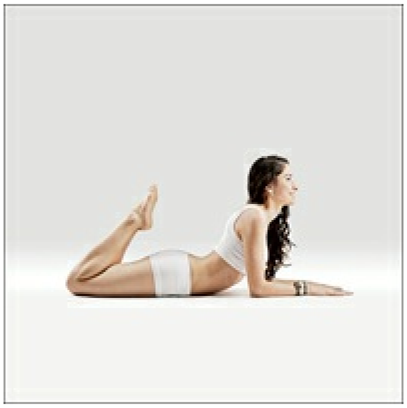

Feet to the Head Supported Cobra Pose

Shirsha Pada Salamba Bhujangasana

(SHEER-shuh PUH-duh SAH-luhm-buh buj-uhng-GAHS-uh-nuh)

Also Known As: Feet to the Head Sphinx Pose
Modification: toes to the forehead
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

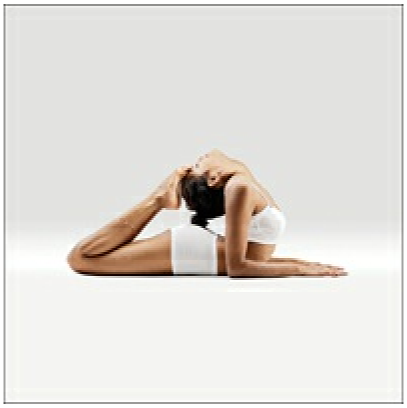

Feet to the Face Supported Cobra Pose

Mukha Pada Salamba Bhujangasana

(MUK-uh PUH-duh SAH-luhm-buh buj-uhng-GAHS-uh-nuh)

Also Known As: Feet to the Face Sphinx Pose
Modification: heels to the forehead
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)
ल व र य ह ॐ

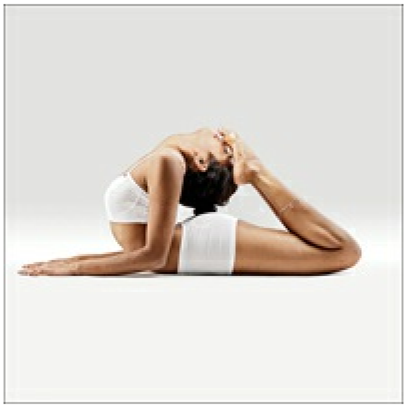

Downward Facing Bound Angle Pose

Adho Mukha Baddha Konasana

(uh-DO MUK-uh BUH-duh ko-NAHS-uh-nuh)

Modification: forearms to the floor

1. top view
2. side view

Pose Type: prone, backbend

Drishti Point: Nasagrai or Nasagre (nose)

ल व र य

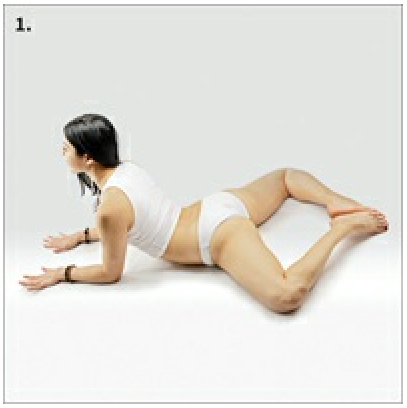

1.

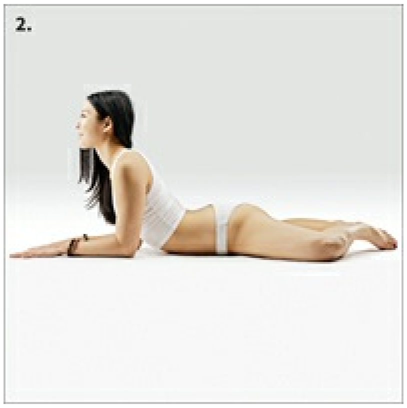

2.

Hidden Lotus Pose

Gupta Padmasana

(GUP-tuh puhd-MAHS-uh-nuh)

Modification: hands to the face on the elbows

Pose Type: prone, backbend

Drishti Point: Nasagrai or Nasagre (nose)

ल व र य

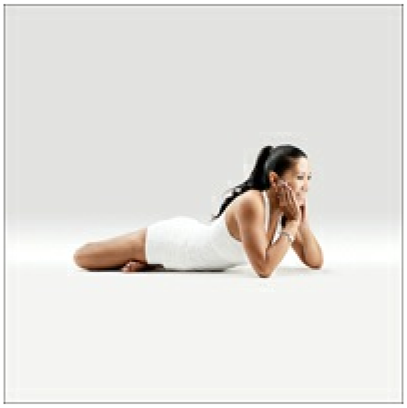

COBRA POSE

Cobra Pose

Bhujangasana

(buj-uhng-GAHS-uh-nuh)

Modification: palms on the floor, elbows bent
1. mild backbend
2. intense backbend

Pose Type: prone, backbend

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ● ● ●

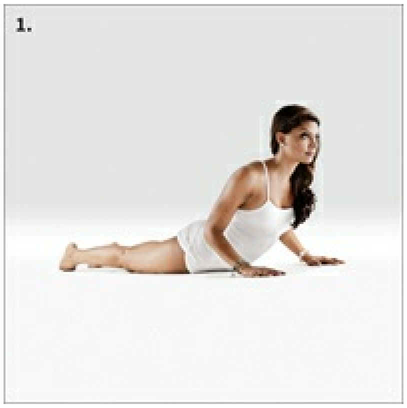

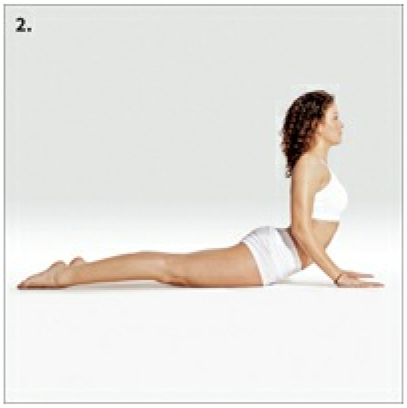

Cobra Pose

Bhujangasana

(buj-uhng-GAHS-uh-nuh)

Modification: 1. elbows bent, fingertips to the floor 2. arms straight, palms to the floor

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)

1.

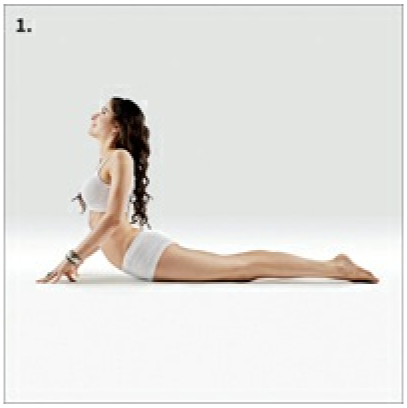

2.

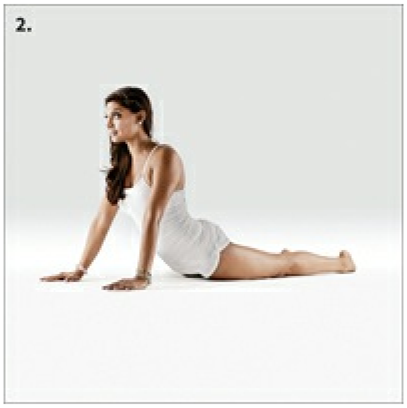

Cobra Pose

Bhujangasana

(buj-uhng-GAHS-uh-nuh)

Modification: palms on the floor, arms straight, head to the glutes

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

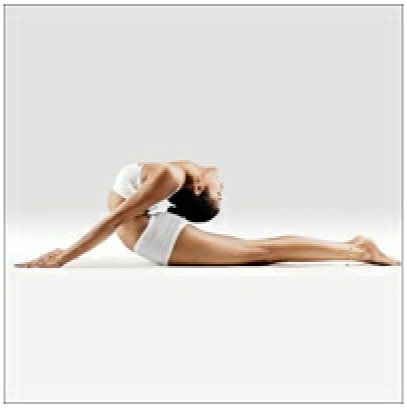

KING PIGEON POSE: HANDS TO THE FLOOR

Half King Pigeon Pose

Ardha Raja Kapotasana

(UHR-duh RAH-juh kuh-po-TAHS-uh-nuh)

Modification: palms to the floor, one foot to the head
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

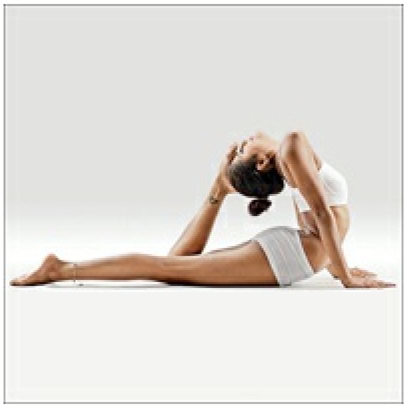

One-Legged One Hand King Pigeon Pose

Eka Pada Eka Hasta Raja Kapotasana

(EY-kuh PUH-duh EY-kuh HUH-stuh RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)

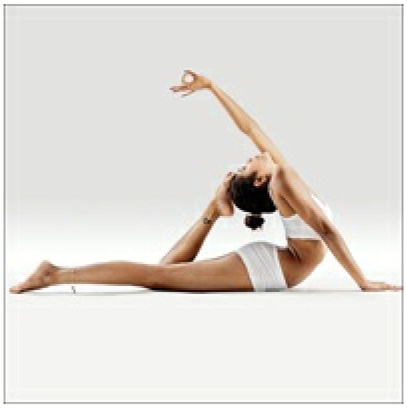

King Pigeon Pose

Raja Kapotasana

(RAH-juh kuh-po-TAHS-uh-nuh)

Modification: palms to the floor, feet away from the head
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

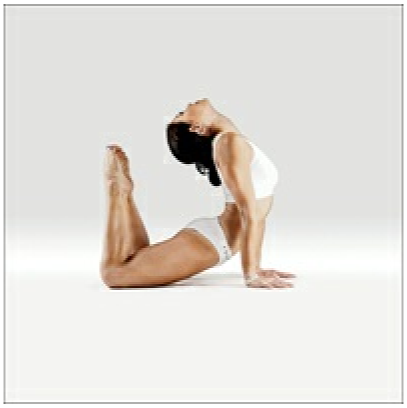

King Pigeon Pose

Raja Kapotasana

(RAH-juh kuh-po-TAHS-uh-nuh)

Modification: head touching the glutes, knees bent, feet away from the head
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

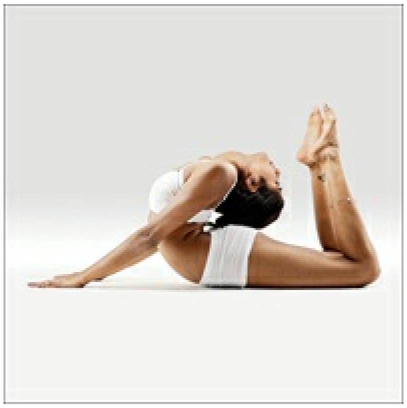

King Pigeon Pose

Raja Kapotasana

(RAH-juh kuh-po-TAHS-uh-nuh)

Modification: palms to the floor, feet to the head
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

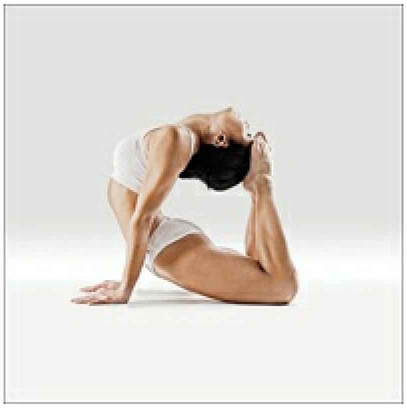

King Pigeon Pose

Raja Kapotasana

(RAH-juh kuh-po-TAHS-uh-nuh)

Modification: palms to the floor, feet to the shoulders
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

King Pigeon Pose

Raja Kapotasana

(RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

How to Perform the Pose:

1. Begin by lying flat on your stomach with your legs straight out behind you. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale and bring your forearms to the floor with your shoulders on top of your elbows. Inhale, lengthen your neck, roll your shoulder blades down, and feel the backbend in your upper back.
3. On the next exhale, press strongly into your hands and straighten your arms. Walk your hands as close as you can toward you.
4. Exhale and bend both your knees, reaching your feet toward your head.
5. Inhale and shift your weight to your left hand. Exhale, bring your right arm behind you, and grab onto your right knee. On the next exhale, bring your left arm behind you and grab onto your left knee.
6. Exhale as you push through your chest, keeping the backbend in the upper back and pressing your feet to the crown of your head (Pose #1).
7. To deepen the pose, exhale and bring your feet to your shoulders (Pose #2).
8. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
9. Inhale as you release the pose, letting go of your knees. Bring your hands to the front and lower your chest and feet to the floor.

Modification: grabbing onto the knees
1. feet to the head
2. feet to the shoulders

raja = king, royal

kapota = pigeon

1.

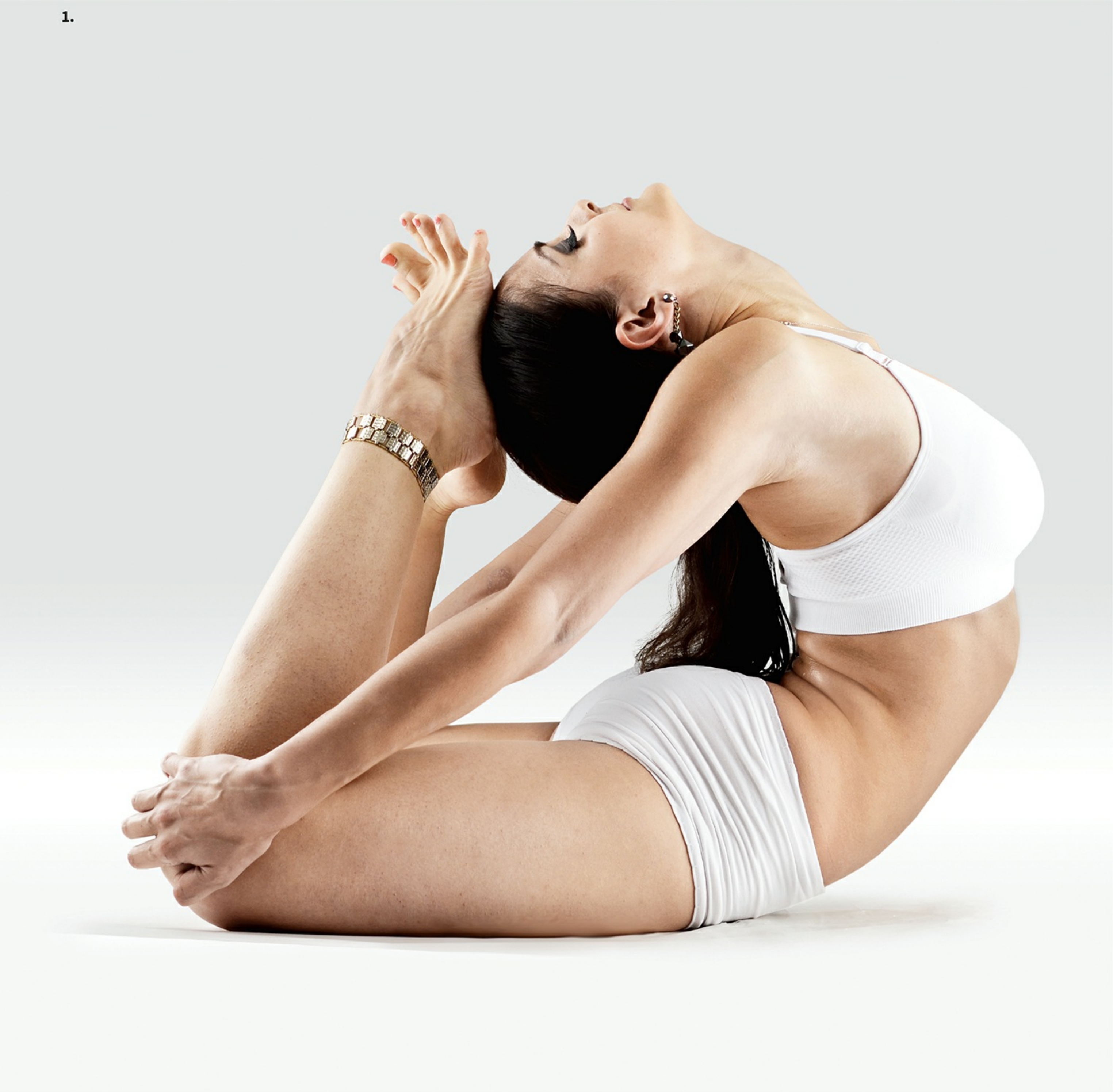

2.

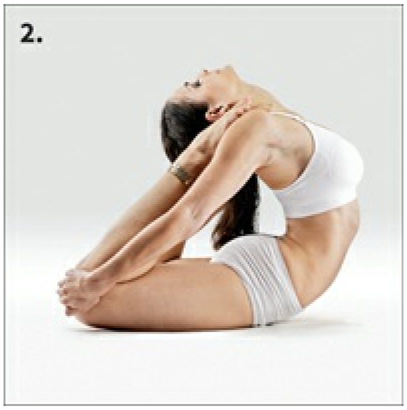

लं वं रं यं हं ॐ

King Pigeon Pose Prep.

Raja Kapotasana Prep.

(RAH-juh kuh-po-TAHS-uh-nuh)

Modification: one hand to the floor, other hand grabbing onto the knee

Pose Type: prone, backbend

Drishti Point: Nasagrai or Nasagre (nose)

ल व र य

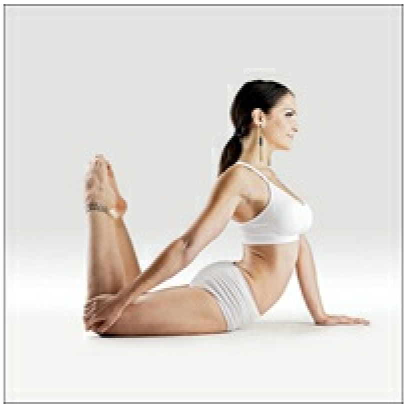

King Pigeon Pose Prep.

Raja Kapotasana Prep.

(RAH-juh kuh-po-TAHS-uh-nuh)

Modification: both hands grabbing onto the knees

1. looking straight ahead
2. head rolling back

Pose Type: prone, backbend

Drishti Point: 1. Nasagrai or Nasagre (nose)

2. Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

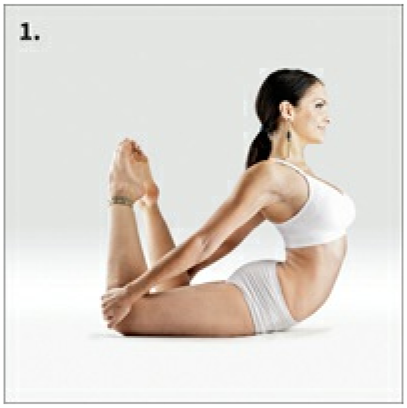

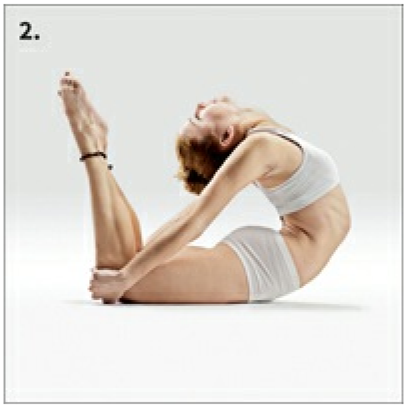

Feet to the Back King Pigeon Pose

Pada Paschima Raja Kapotasana

(PUH-duh PUHSH-chi-muh RAH-juh kuh-po-TAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं रं यं हं ॐ

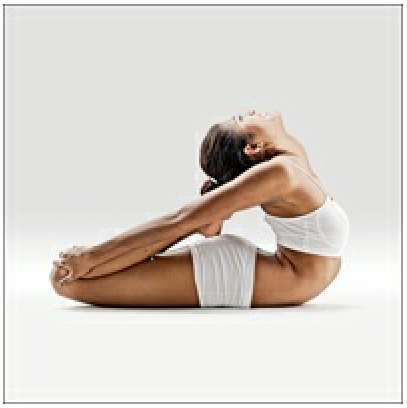

UPWARD DOG POSE: BOTH LEGS STRAIGHT

Upward Facing Dog Pose

Urdhva Mukha Shvanasana

(OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: toes pointed to the back, knees on the floor
1. looking straight ahead
2. head rolling back
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं यं हं ॐ

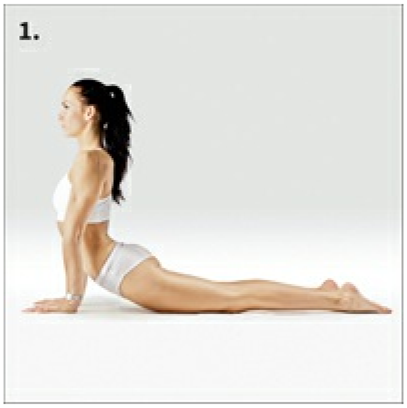

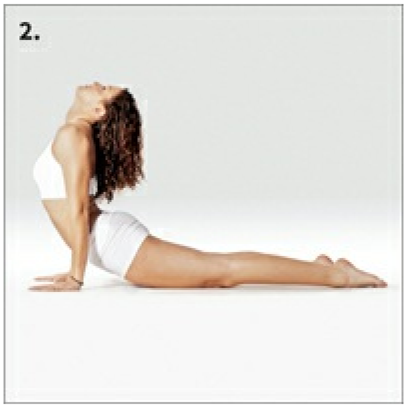

Upward Facing Dog Pose

Urdhva Mukha Shvanasana

(OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: toes pointed to the back, knees off the floor
1. looking straight ahead
2. head rolling back
Pose Type: prone, backbend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

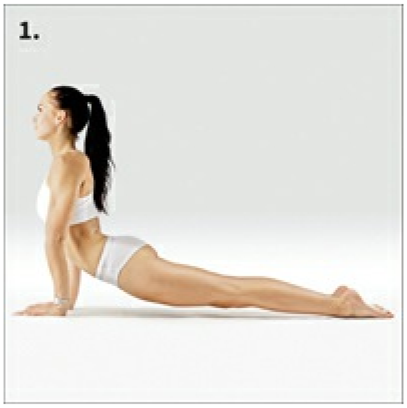

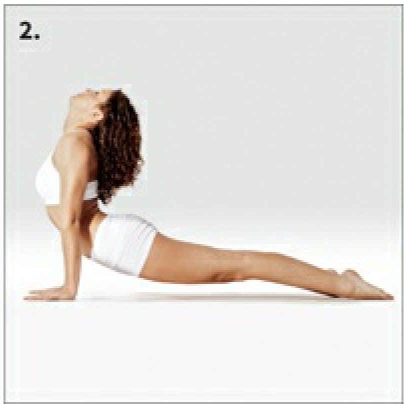

UPWARD DOG POSE: BOTH LEGS STRAIGHT—ONE LEG, ONE ARM, & TO THE SIDE

Uneven One-Legged Upward Facing Dog Pose

Vishama Eka Pada Urdhva Mukha Shvanasana

(VISH-uh-muh EY-kuh PUH-duh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: both toes pointed
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
🔴 🟡 🟢 🔵

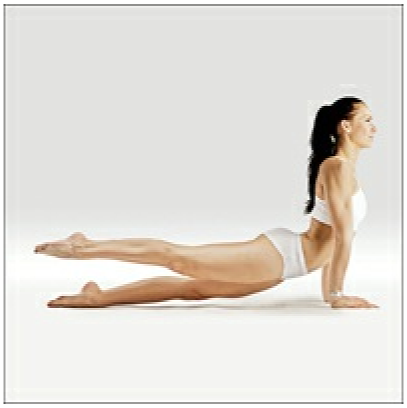

Sideways Upward Facing Dog Pose

Parshva Urdhva Mukha Shvanasana

(PAHRSH-vuh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Pose Type: prone, backbend, side bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
🔴 🟡 🟢 🔵

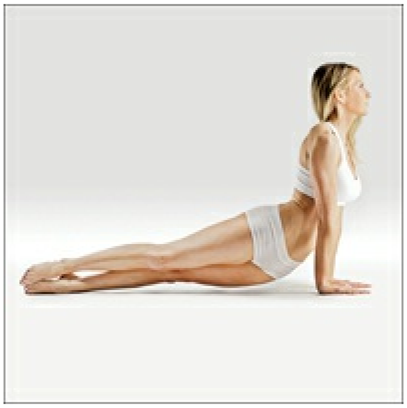

One Hand Upward Facing Dog Pose

Eka Hasta Urdhva Mukha Shvanasana

(EY-kuh HUH-stuh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: toes pointed to the back

1. knees on the floor
2. knees off the floor

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

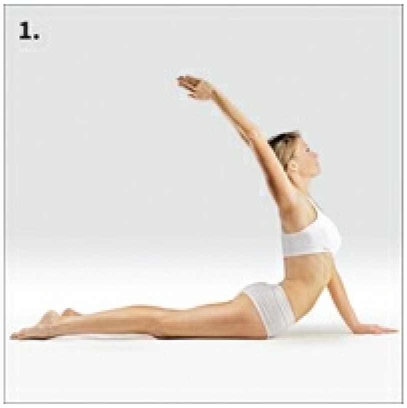
1.

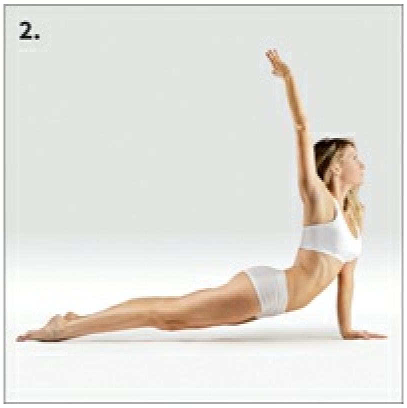
2.

UPWARD DOG POSE: TOES CURLED IN

Upward Facing Dog Pose

Urdhva Mukha Shvanasana

(OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: toes curled in
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

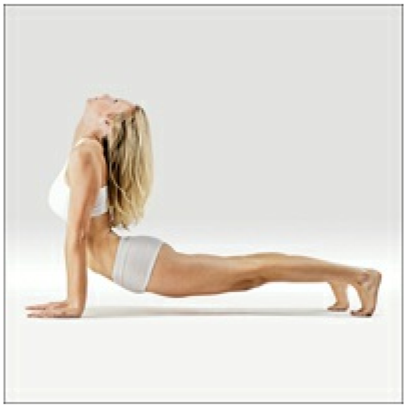

Sideways One-Legged Upward Facing Dog Pose

Parshva Eka Pada Urdhva Mukha Shvanasana

(PAHRSH-vuh EY-kuh PUH-duh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: toes curled in
Pose Type: prone, backbend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ॐ

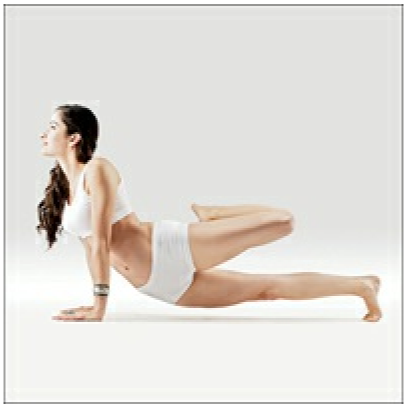

One-Legged Upward Facing Dog Pose

Eka Pada Urdhva Mukha Shvanasana

(EY-kuh PUH-duh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: toes of the straight leg curled in; knee of the other leg bent, toes pointing up to the sky

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

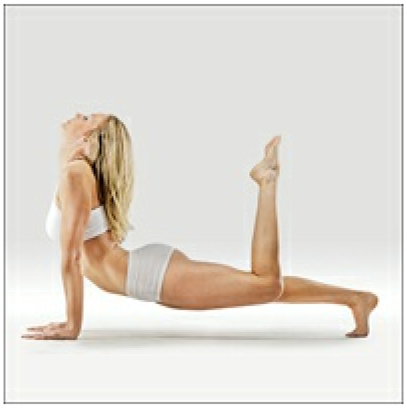

Half Big Toe Bow Pose in One-Legged Upward Facing Dog Pose

Ardha Padangushta Dhanurasana in Eka Pada Urdhva Mukha Shvanasana

(UHR-duh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh in EY-kuh PUH-duh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: grabbing onto the foot with overhead grip on the same side

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

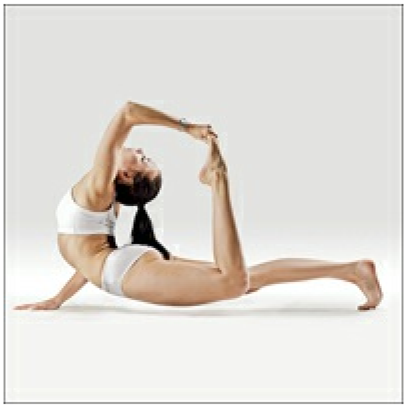

UPWARD DOG POSE: TOES CURLED IN—LEGS CROSSED

Upward Facing Dog Pose

Urdhva Mukha Shvanasana

(OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: toes curled in, ankles crossed

1. knees on the floor
2. knees off the floor, looking straight ahead
3. knees off the floor, head rolling back

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

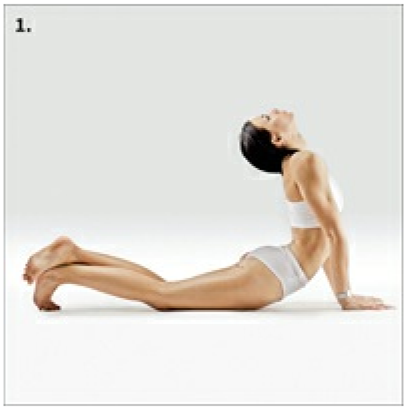

2.

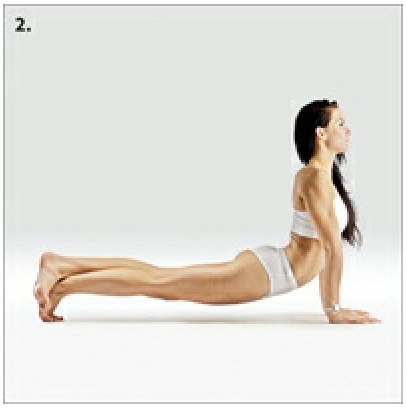

3.

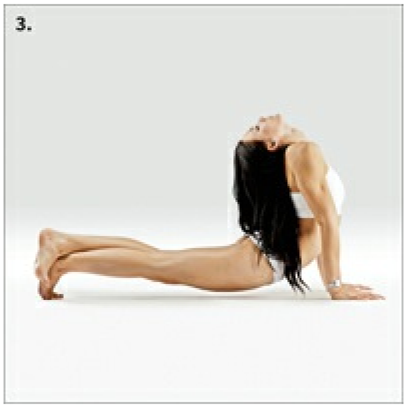

One-Legged Upward Facing Dog Pose

Eka Pada Urdhva Mukha Shvanasana

(EY-kuh PUH-duh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: toes curled in, foot of the bent leg to the inside of the knee of the straight leg
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं रं यं ॐ

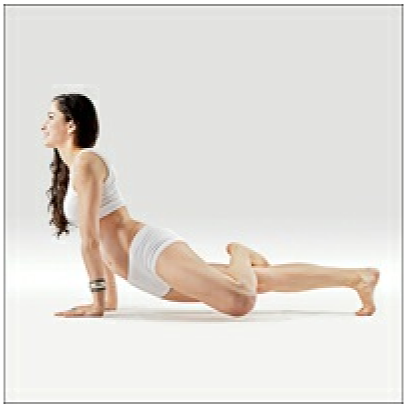

UPWARD DOG POSE: ONE LEG BENT

Uneven One-Legged Upward Facing Dog Pose

Vishama Eka Pada Urdhva Mukha Shvanasana

(VISH-uh-muh EY-kuh PUH-duh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)
Modification: Leg 1: knee bent and on the floor, toes pointed to the sky Leg 2: straight and off the floor, toes pointed away
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं यं ॐ

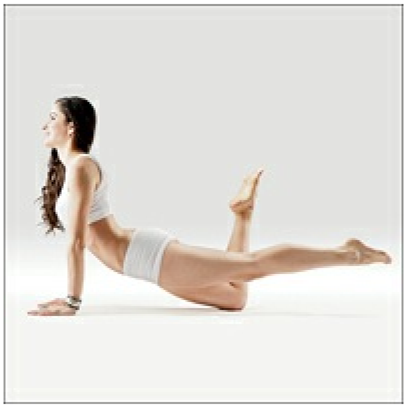

Half Bow Pose in One-Legged Upward Facing Dog Pose

Ardha Dhanurasana in Eka Pada Urdhva Mukha Shvanasana

(UHR-duh duh-nur-AHS-uh-nuh in EY-kuh PUH-duh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)

Modification: grabbing onto the opposite foot with under-head grip
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ॐ

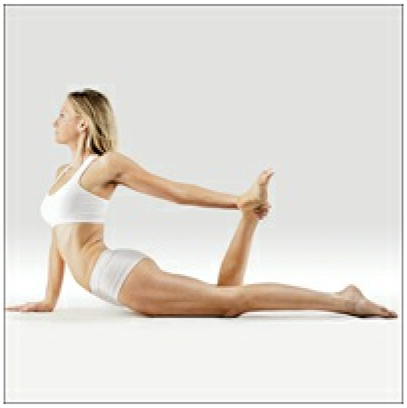

Half Bow Pose in One-Legged Upward Facing Dog Pose

Ardha Dhanurasana in Eka Pada Urdhva Mukha Shvanasana

(UHR-duh duh-nur-AHS-uh-nuh in EY-kuh PUH-duh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)
Modification: under-head grip on the same side
Pose Type: prone, backbend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ॐ

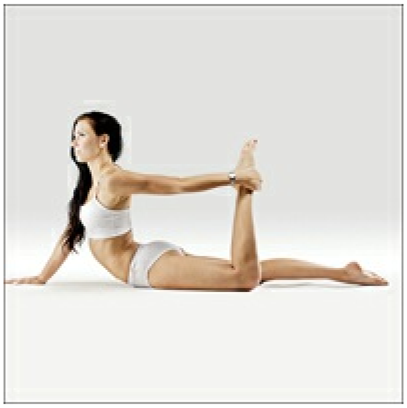

Half Pose Dedicated to Siddhar Konganar in One-Legged Upward Facing Dog Pose

Ardha Konganarasana in Eka Pada Urdhva Mukha Shvanasana

(UHR-duh kong-guh-nuh-RAHS-uh-nuh in EY-kuh PUH-duh OORD-vuh MUK-uh shvuh-NAHS-uh-nuh)
Modification: grabbing onto the knee with under-head grip on the same side—foot to the armpit

Pose Type: prone, backbend

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

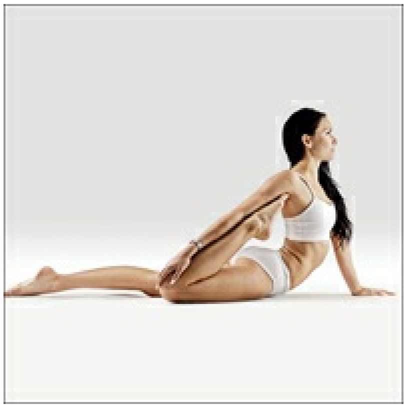

COBRA POSE: UNDER-HEAD GRIP

Half King Pigeon Pose in One-Legged Big Toe Bow Pose

Ardha Raja Kapotasana in Eka Pada Padangushta Dhanurasana

(UHR-duh RAH-juh kuh-po-TAHS-uh-nuh in EY-kuh PUH-duh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Modification: Hand 1: grabbing onto the knee on the same side.
Hand 2: grabbing onto the big toe of the opposite foot with overhead grip

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

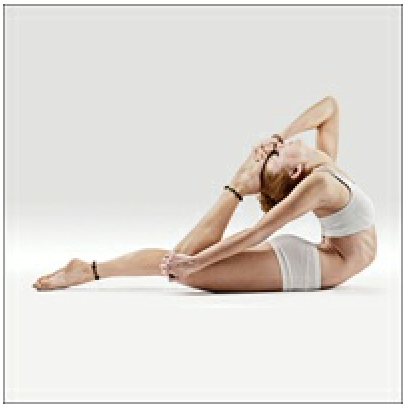

Cobra Pose

Bhujangasana

(buj-uhng-GAHS-uh-nuh)

Also Known As: Supported Cobra Pose (Alamba Bhujangasana), King Pigeon Pose Prep. (Raja Kapotasana Prep.)
Modification: grabbing onto the knees with under-head grip
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

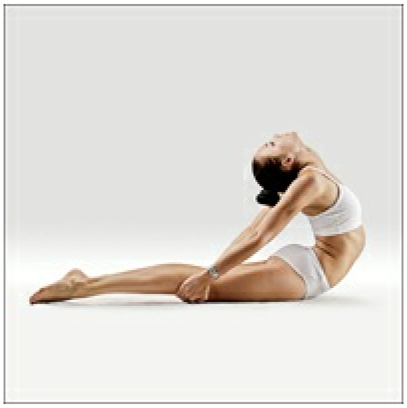

COBRA POSE: OVERHEAD GRIP

Upward Facing Unsupported Cobra Pose

Urdhva Mukha Niralamba Bhujangasana

(OORD-vuh MUK-uh nir-AH-luhm-buh buj-uhng-GAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

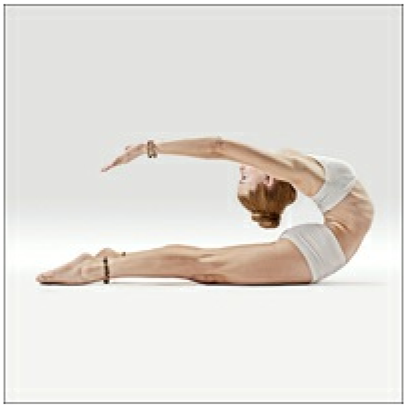

Unsupported Complete Cobra Pose

Niralamba Paripurna Bhujangasana

(nir-AH-luhm-buh puh-ri-POOR-nuh buj-uhng-GAHS-uh-nuh)

Pose Type: prone, backbend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

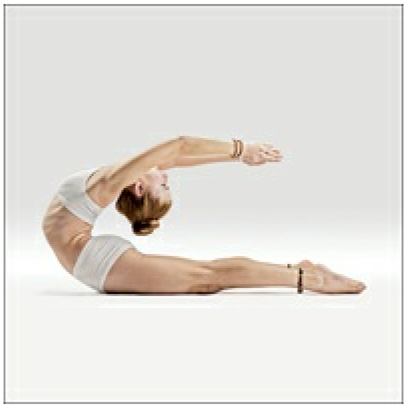

Complete Cobra Pose

Paripurna Bhujangasana

(puh-ri-POOR-nuh buj-uhng-GAHS-uh-nuh)

Modification: 1. grabbing onto the shins with overhead grip
2. grabbing onto the knees with overhead grip
3. grabbing onto the shins with overhead grip, toes curled in

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

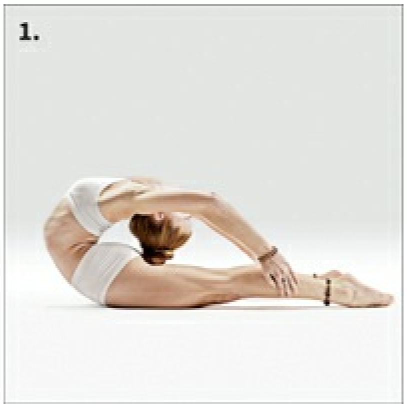

2.

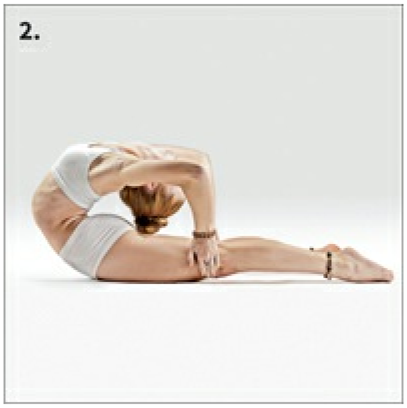

3.

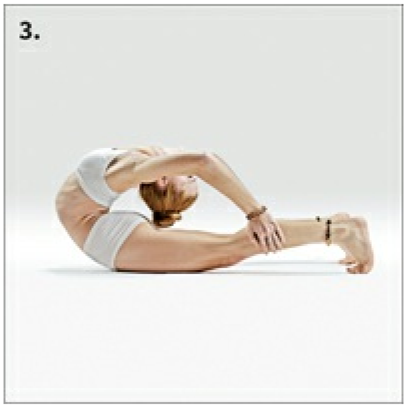

BOW POSE: UNDER-HEAD GRIP—BOTH KNEES BENT

Bow Pose

Dhanurasana

(duh-nur-AHS-uh-nuh)

Modification: hands grabbing onto the ankles

1. on the inside of the ankles
2. on the outside of the ankles

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

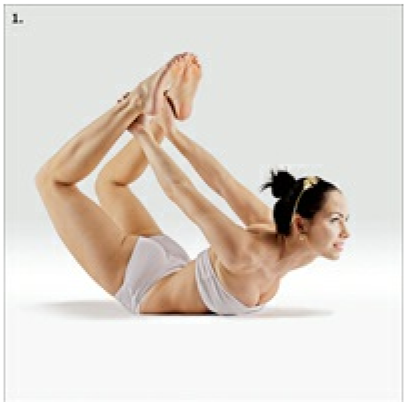

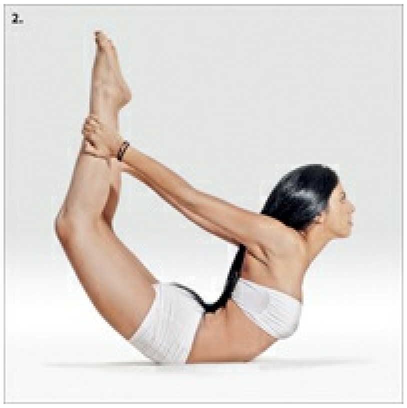

Bow Pose

Dhanurasana

(duh-nur-AHS-uh-nuh)

Modification: legs bent at 90 degrees; ankles and knees together, toes pointed to the sky
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ॐ

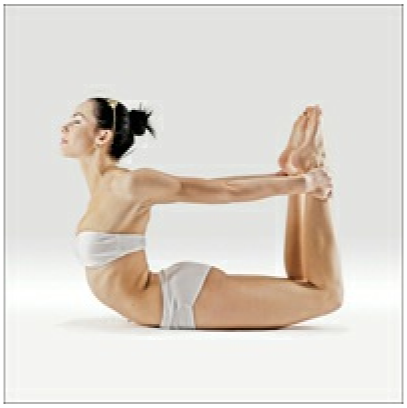

Bow Pose

Dhanurasana

(duh-nur-AHS-uh-nuh)

Modification: legs bent at 90 degrees; ankles and knees together, toes flexed away from the head
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)
ल व र य ॐ

Sideways Bow Pose

Parshva Dhanurasana

(PAHRSH-vuh duh-nur-AHS-uh-nuh)

Pose Type: supine, prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल र य ह ॐ

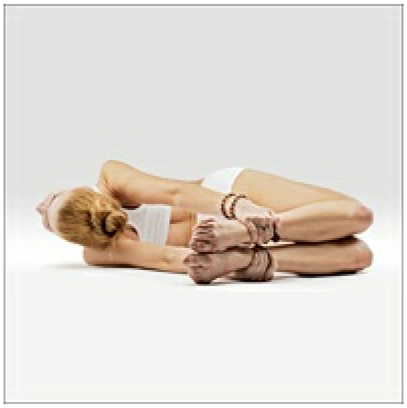

One-Legged

Pose Dedicated to Siddhar Konganar

Eka Pada Konganarasana

(EY-kuh PUH-duh kong-guh-nuh-RAHS-uh-nuh)

Pose Type: prone, backbend

Drishti Point: Nasagrai or Nasagre (nose) or Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

How to Perform the Pose:

1. Begin by lying flat on your stomach with your legs straight out behind you. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale and bring your forearms to the floor with your shoulders on top of your elbows. Inhale, lengthen your neck, roll your shoulder blades down, and feel the backbend in your upper back.
3. On the next exhale, press strongly into your hands and straighten your arms.
4. Exhale and bend both your knees, reaching your feet toward your head.
5. On the next exhale, reach your right hand behind you and grab onto your right shin. Keep sliding your right hand until your right foot reaches your right armpit and your right hand rests on your right knee.
6. Inhale and reach your left arm up and over your head. Exhale and grab onto your left foot with your left hand, keeping your left elbow close to your ear.
7. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
8. Inhale, let go of your legs one by one, and lower your chest and feet to the floor to come back to the starting position. Repeat on the opposite side.

eka = one

pada = foot or leg

Konganar = one of the Siddhars

Little Bow Pose

Laghu Dhanurasana

(LUH-gu duh-nur-AHS-uh-nuh)

Pose Type: prone, backbend

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

लं रं यं ॐ

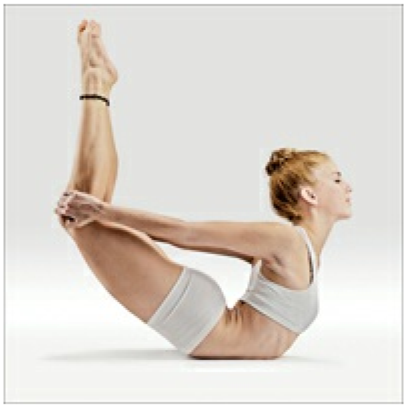

Feet to the Head Little Bow Pose

Shirsha Pada Laghu Dhanurasana

(SHEER-suh PUH-duh LUH-gu duh-nur-AHS-uh-nuh)

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

लं रं यं हं ॐ ॐ

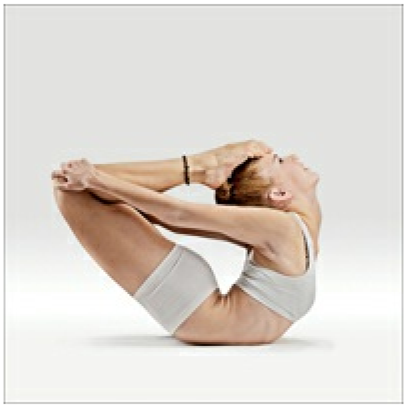

Pose Dedicated to Siddhar Konganar

Konganarasana

(kong-guh-nuh-RAHS-uh-nuh)

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

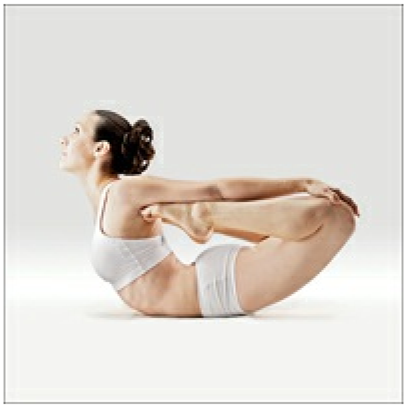

BOW POSE: OVERHEAD GRIP—BOTH KNEES BENT

One Hand Both Feet Big Toe Bow Pose

Eka Hasta Dwi Pada Padangushta Dhanurasana

(EY-kuh HUH-stuh DWI-puh-duh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Modification: grabbing onto the toes of both feet with one hand, other palm to the floor

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

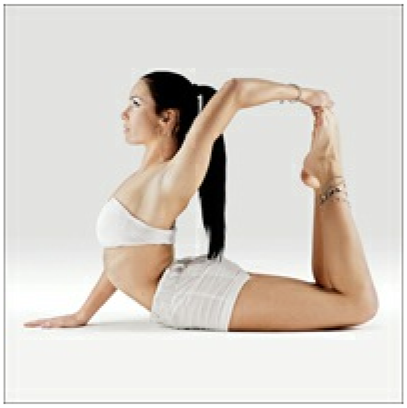

Big Toe Bow Pose

Padangushta Dhanurasana

(puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Modification: head rolling back

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

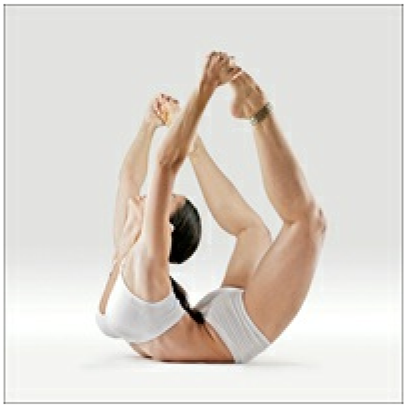

Big Toe Bow Pose

Padangushta Dhanurasana

(puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)
Modification: 1. feet to the face
2. heels to the forehead
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

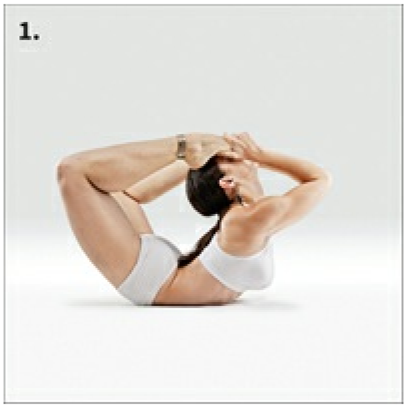
1.

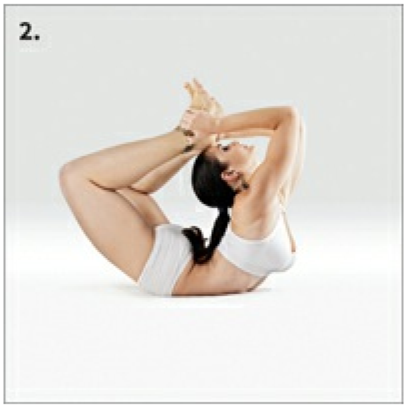
2.

One Hand One-Legged Big Toe Bow Pose

Eka Hasta Eka Pada Padangushta Dhanurasana

(EY-kuh HUH-stuh EY-kuh PUH-duh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Modification: 1. one hand grabbing onto the foot on the opposite side with overhead grip
2. with backbend
Pose Type: prone, backbend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

1.

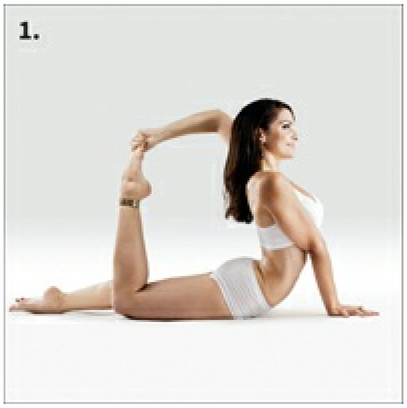

2.

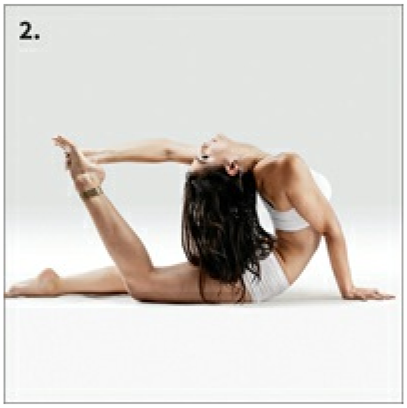

One Hand One-Legged Big Toe Bow Pose

Eka Hasta Eka Pada Padangushta Dhanurasana

(EY-kuh HUH-stuh PUH-duh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Modification: one hand grabbing onto the foot on the same side with overhead grip
Pose Type: prone, backbend
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

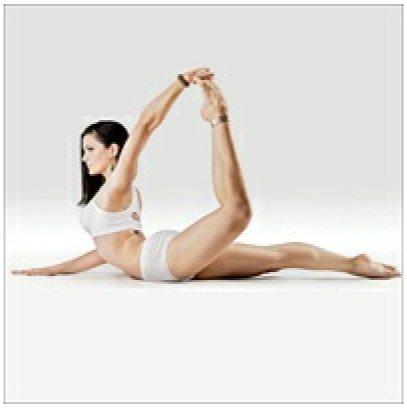

Both Hand One-Legged Big Toe Bow Pose

Dwi Hasta Eka Pada Padangushta Dhanurasana

(DWI-huh-stuh EY-kuh PUH-duh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Modification: both hands grabbing onto the foot with overhead grip

Pose Type: prone, backbend

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

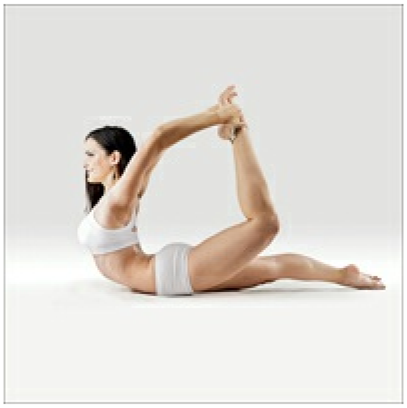

BOW POSE: OVERHEAD GRIP—ONE KNEE BENT, ONE LEG STRAIGHT AND OFF THE FLOOR

Pose Dedicated to Goddess Kamala

Kamalasana

(kuh-muh-LAHS-uh-nuh)

Also Known As: Upward Facing One-Handed Big Toe Boat Pose (Urdhva Mukha Eka Hasta Padangushta Navasana)

Modification: 1: grabbing onto the ankle

2. foot to the shoulder

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

1.

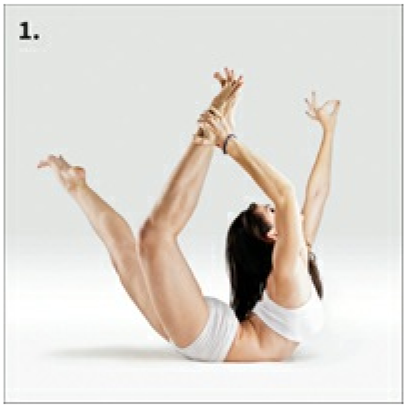

2.

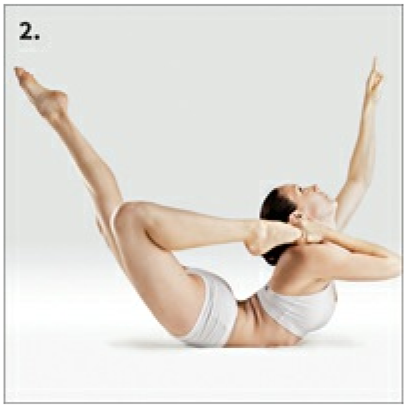

BOW POSE: CRISS-CROSS LEGS AND ARMS

Revolved Bow Pose

Parivritta Dhanurasana

(puh-ri-VRIT-tuh duh-nur-AHS-uh-nuh)

Pose Type: prone, backbend, twist

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padhayoragrai or Padayoragre (toes/feet)

ल व र य ॐ

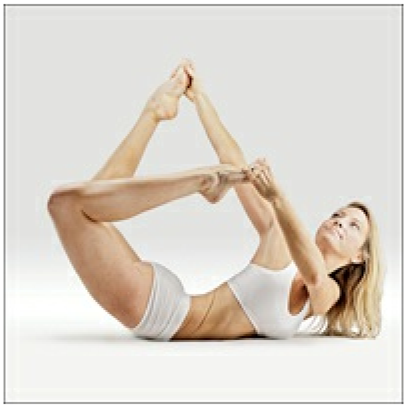

Bow Pose

Dhanurasana

(duh-nur-AHS-uh-nuh)

Modification: legs crossed

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

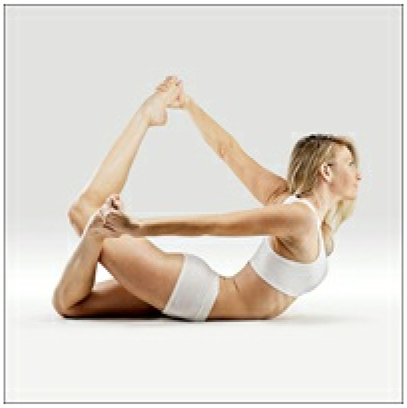

Big Toe Bow Pose

Padangushta Dhanurasana

(puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Also Known As: Difficult Bow Pose (Dur Dhanurasana)

Modification: one hand grabbing opposite ankle with under-head grip, other hand grabbing the opposite foot with overhead grip

1. one knee on the floor
2. both knees off the floor

Pose Type: prone, backbend

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

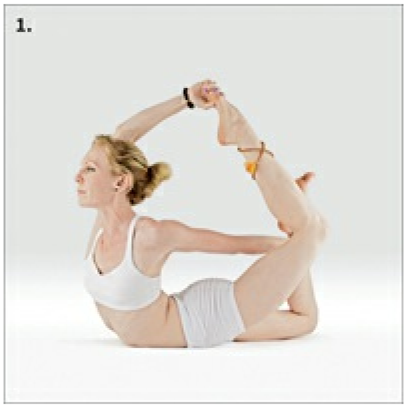

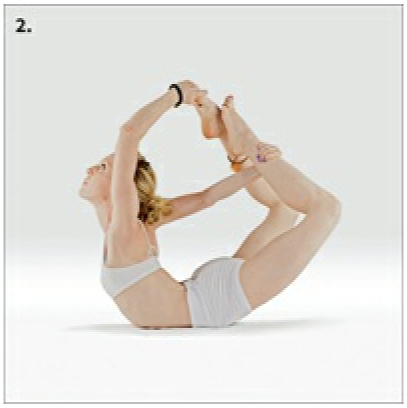

Big Toe Bow Pose

Padangushta Dhanurasana

(puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Modification: one foot toward the shoulder
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

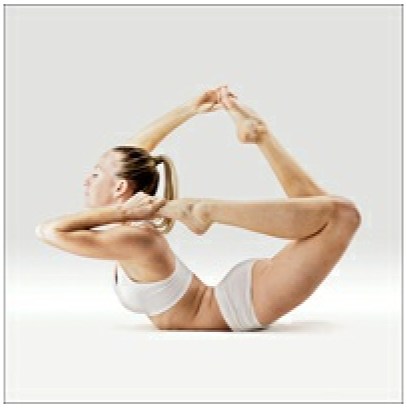

HALF LOTUS BACKBENDS

Half Lotus Frog Pose

Ardha Padma Bhekasana

(UHR-duh PUHD-muh bey-KAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

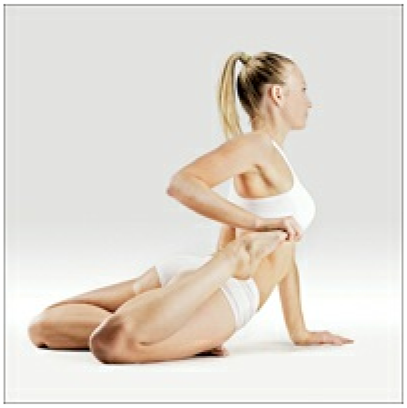

Half Lotus Bow Pose

Ardha Padma Dhanurasana

(UHR-duh PUHD-muh duh-nur-AHS-uh-nuh)

Modification: both hands grabbing onto the ankle with under-head grip
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

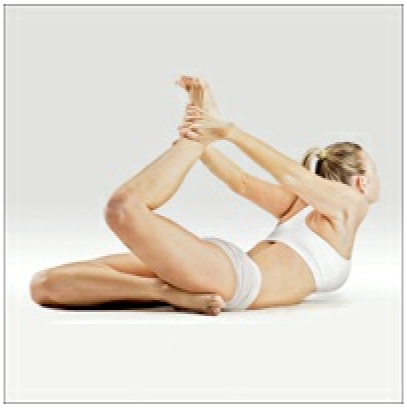

Half Lotus One Hand One-Legged Big Toe Bow Pose

Ardha Padma Eka Hasta Eka Pada Padangushta Dhanurasana

(UHR-duh PUHD-muh EY-kuh HUH-stuh EY-kuh PUH-duh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Modification: grabbing onto the foot with overhead grip on the same side, other hand to the floor

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

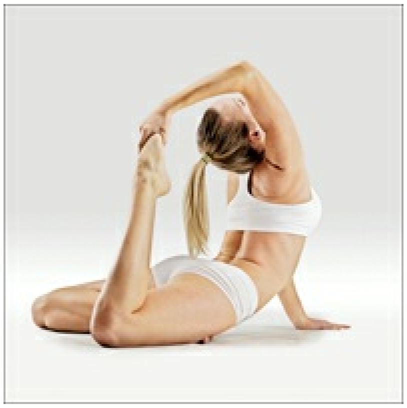

LOTUS BACKBENDS

Lotus Cobra Pose

Padma Bhujangasana

(PUHD-muh buj-uhng-GAHS-uh-nuh)

Modification: both hands to the floor, hips off the floor, arms straight

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ॐ

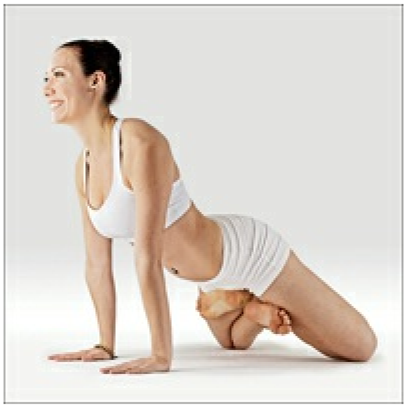

Lion Pose Dedicated to an Avatar of Lord Vishnu

Narasimhasana

(nuh-ruh-sim-HAHS-uh-nuh)

Also Known As: Lion Pose (Simhasana)
Modification: full lotus leg position
1. elbows bent, fingers in lion claw
2. arms straight, palms flat on the floor
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

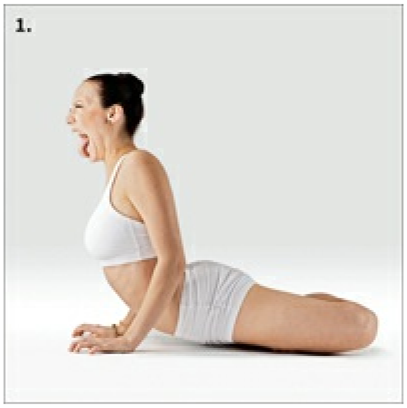
1.

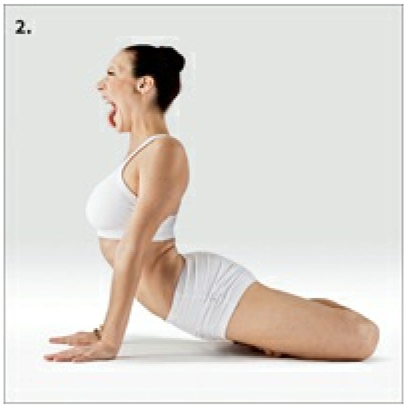
2.

One Hand Lotus Cobra Pose

Eka Hasta Padma Bhujangasana

(EY-kuh HUH-stuh PUHD-muh buj-uhng-GAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

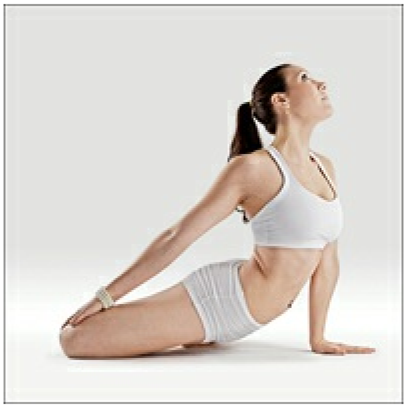

Lotus Cobra Pose

Padma Bhujangasana

(PUHD-muh buj-uhng-GAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

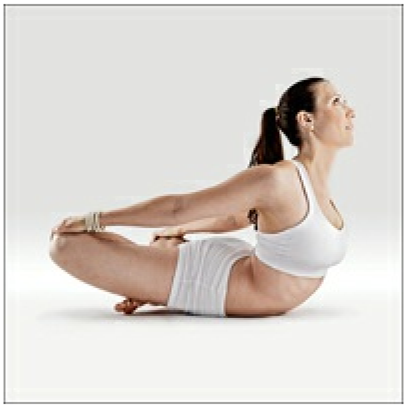

SAGE GHERANDA’S POSE: UNDER-HEAD GRIP

Revolved Pose Dedicated to Sage Gheranda 1

Parivritta Ardha Gherandasana 1

(puh-ri-VRIT-tuh UHR-duh gey-ruhn-DAHS-uh-nuh)

Pose Type: prone, backbend, twist
Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

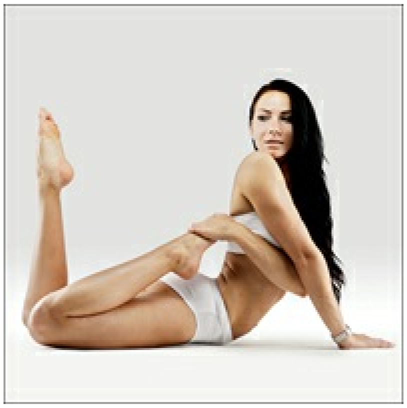

Half Pose Dedicated to Sage Gheranda 1

Ardha Gherandasana 1

(UHR-duh gey-ruhn-DAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Nasagrai or Nasagre (nose)

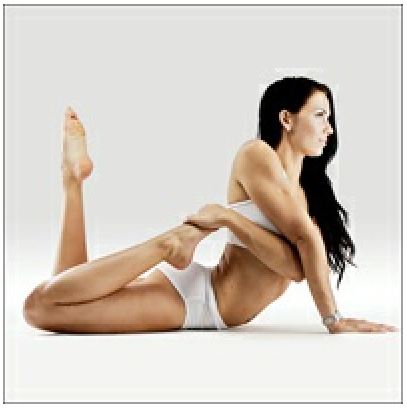

Pose Dedicated to Sage Gheranda 5B

Gherandasana 5 B

(gey-ruhn-DAHS-uh-nuh)

Modification: under-head grip, grabbing onto the ankle, elbow on the floor instead of the forearm
Pose Type: prone, backbend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

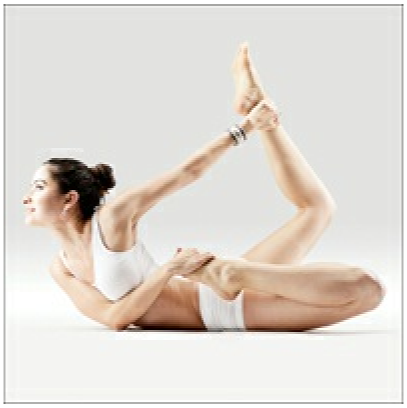

Pose Dedicated to Sage Gheranda 1B

Gherandasana 1 B

(gey-ruhn-DAHS-uh-nuh)

Modification: under-head grip

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

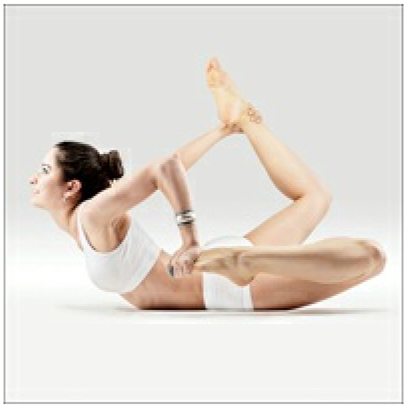

Pose Dedicated to Sage Gheranda 6

Gherandasana 6

(gey-ruhn-DAHS-uh-nuh)

Modification: under-head grip

Pose Type: prone, backbend, binding

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

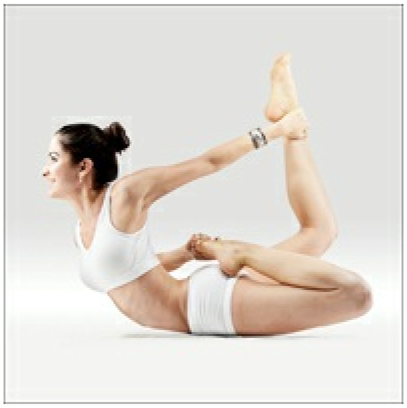

SAGE GHERANDA’S POSE: OVERHEAD GRIP

Pose Dedicated to Sage Gheranda 5

Gherandasana 5

(gey-ruhn-DAHS-uh-nuh)

Pose Type: prone, backbend, binding

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

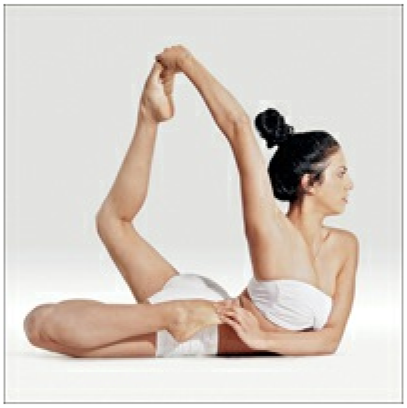

Pose Dedicated to Sage Gheranda 1

Gherandasana 1

(gey-ruhn-DAHS-uh-nuh)

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

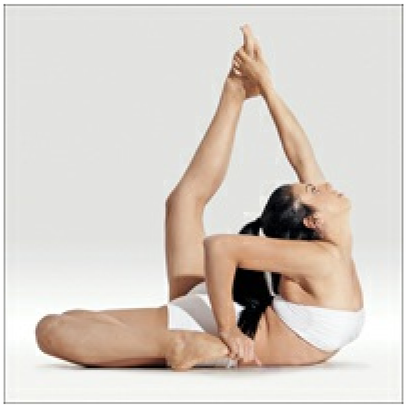

Pose Dedicated to Sage Gheranda 2

Gherandasana 2

(gey-ruhn-DAHS-uh-nuh)

Pose Type: prone, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

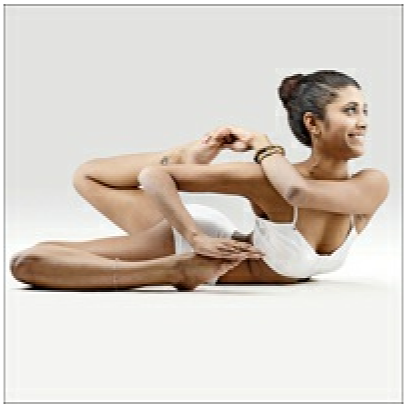

Pose Dedicated to Sage Gheranda 3

Gherandasana 3

(gey-ruhn-DAHS-uh-nuh)

Pose Type: prone, backbend, binding
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

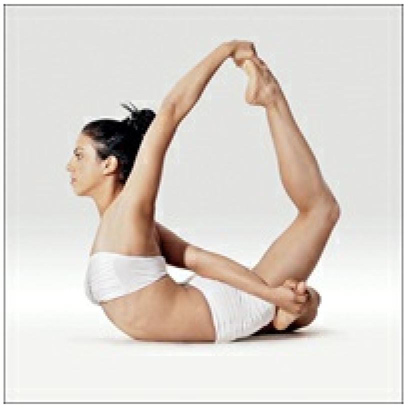

Pose Dedicated to Sage Gheranda 4

Gherandasana 4

(gey-ruhn-DAHS-uh-nuh)

Modification: 1. back view
2. front view
Pose Type: prone, backbend, binding, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह

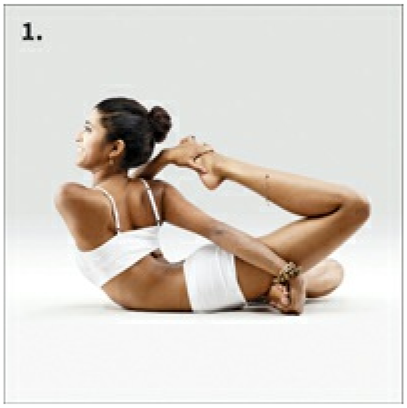

1.

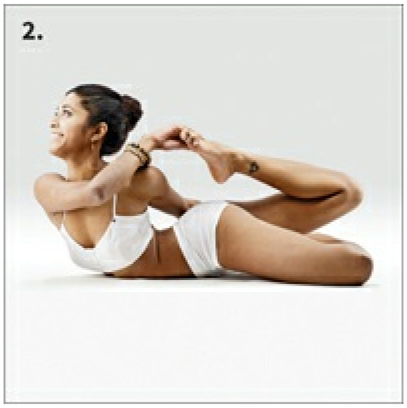

2.

BACKBENDS: GARUDA LEGS: UNDER-HEAD GRIP

Twined Legs Cobra Pose

Parivitta Pada Bhujangasana

(puh-ri-VRIT-tuh PUH-duh buj-uhng-GAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं रं यं हं ॐ

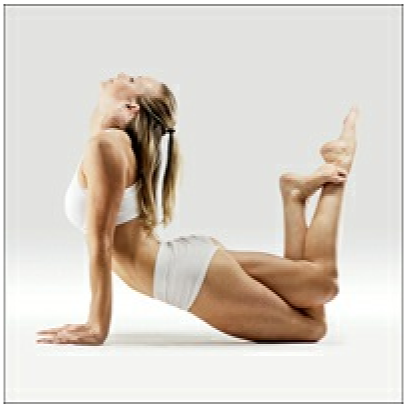

Supported Twined Legs Cobra Pose

Salamba Parivid Pada Bhujangasana

(SAH-luhm-buh PUH-ri-vid PUH-duh buj-uhng-GAHS-uh-nuh)

Also Known As: Supported Twined Legs Sphinx Pose
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं रं यं ॐ

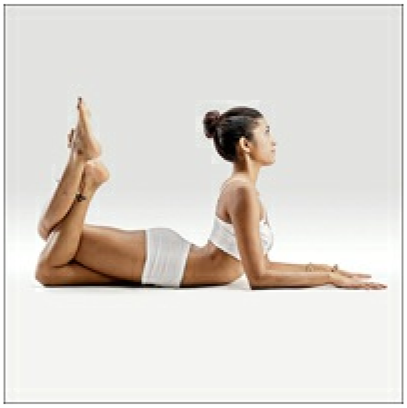

Leg Position of the Pose Dedicated to Garuda in One Hand Bow Pose

Pada Garudasana in Eka Hasta Dhanurasana

(PUH-duh guh-ru-DAHS-uh-nuh in EY-kuh HUH-stuh duh-nur-AHS-uh-nuh)

Modification: one forearm to the floor, one hand to the ankle, under-head grip

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

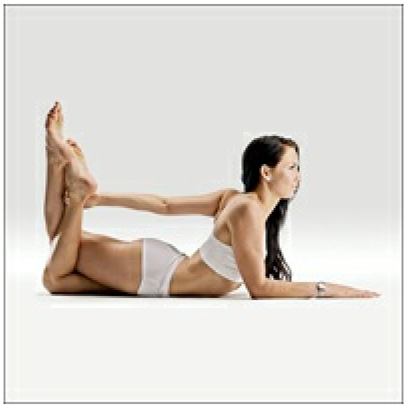

BACKBENDS: GARUDA LEGS: OVERHEAD GRIP

Leg Position of the Pose Dedicated to Garuda in One Hand Big Toe Bow Pose

Pada Garudasana in Eka Hasta Padangushta Dhanurasana

(PUH-duh guh-ru-DAHS-uh-nuh in EY-kuh HUH-stuh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Modification: forearm to the floor

1. grabbing onto the top foot
2. grabbing onto the bottom foot

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

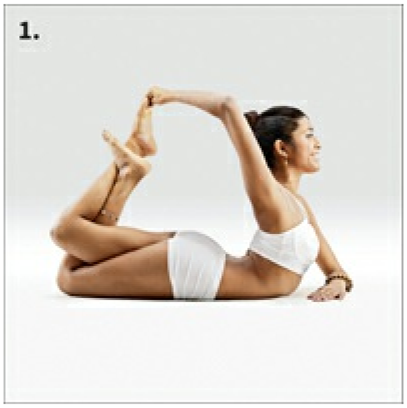

2.

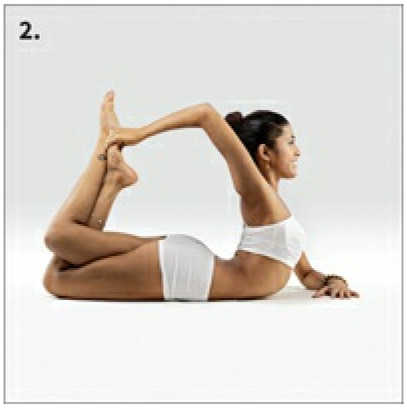

Unsupported Leg Position of the Pose Dedicated to Garuda in One Hand Big Toe Bow Pose

Niralamba Pada Garudasana in Eka Hasta Padangushta Dhanurasana

(nir-AH-luhm-buh puh-ri-vid-PUH-duh guh-ru-DAHS-uh-nuh in EY-kuh HUH-tuh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Also Known As: One Handed Twined Legs Bow Pose (Eka Hasta Parivid Pada Dhanurasana)

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

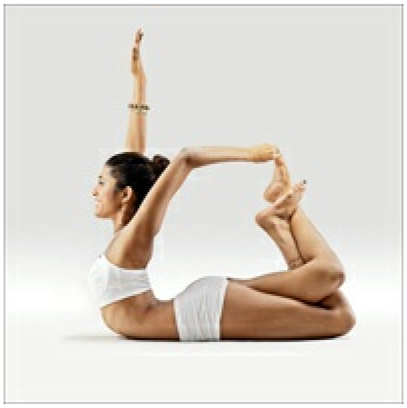

Leg Position of the Pose Dedicated to Garuda in Big Toe Bow Pose

Pada Garudasana in Padangushta Dhanurasana

(PUH-duh guh-ru-DAHS-uh-nuh in puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)

Also Known As: Difficult Bow Pose (Dur Dhanurasana)

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

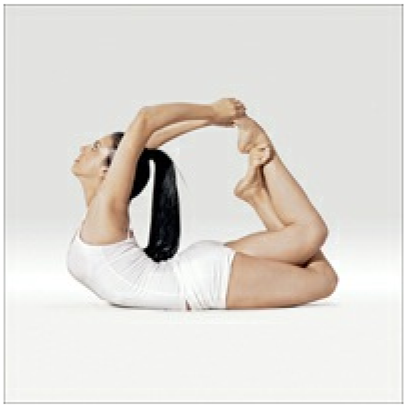

CHIN ON THE FLOOR & EAR ON THE FLOOR

One-Legged Inverted Locust Pose

Eka Pada Viparita Shalabhasana

(EY-kuh PUH-duh vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Pose Type: prone, backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

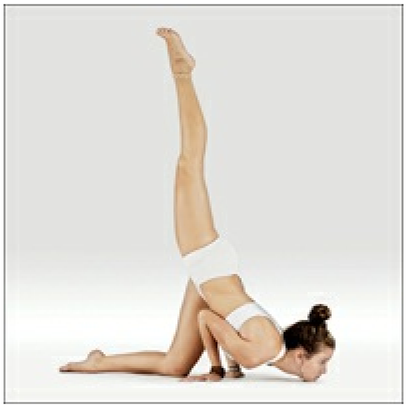

Ear Pressure Staff Formidable Face Pose

Karnapida Danda Ganda Bherundasana

(kuhr-nah-PEED-uh DUHN-duh GUHN-duh bey-run-DAHS-uh-nuh)

Pose Type: prone, backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ॐ

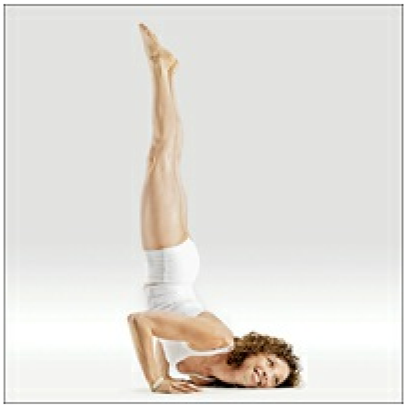

CHEST ON THE FLOOR: BOTH LEGS STRAIGHT

Inverted Locust Pose

Viparita Shalabhasana

(vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Modification: 1. chin to the floor
2. chest to the floor, chin lifted off the floor
Pose Type: prone, backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

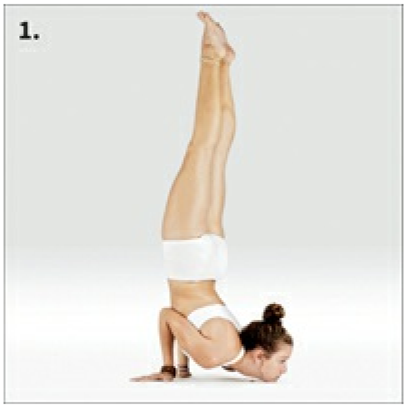
1.

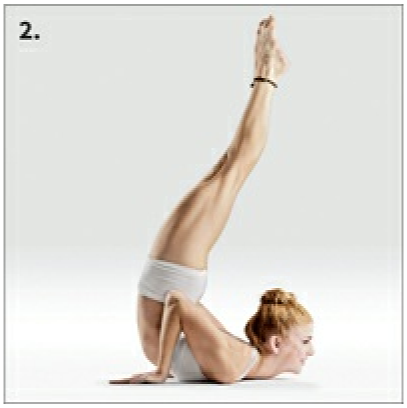
2.

Formidable Face Staff Pose

Danda Ganda Bherundasana

(DUHN-duh GUHN-duh bey-run-DAHS-uh-nuh)

Also Known As: Intense Stretch Locust Pose A (Uttana Shalabhasana A)
Pose Type: prone, backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

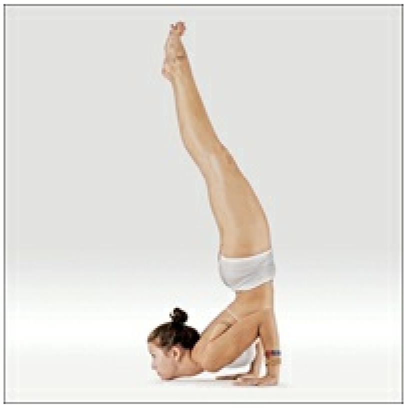

Inverted Locust Pose

Viparita Shalabhasana

(vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Also Known As: Raised Locust Pose—Urdhva Shalabhasana)
Modification: arms straight on the floor, palms down, chin lifted off the floor
Pose Type: prone, backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

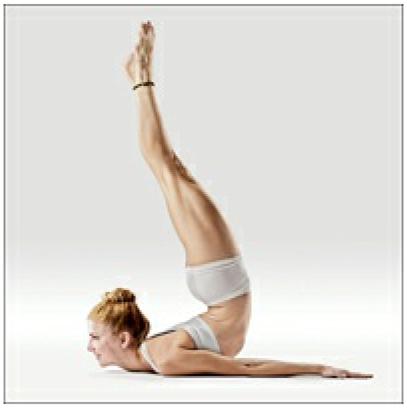

CHEST ON THE FLOOR: BOTH FEET TO HEAD

Formidable Face Pose

Ganda Bherundasana

(GUHN-duh bey-run-DAHS-uh-nuh)

Also Known As: Full Locust Pose (Purna Shalabhasana), Intense Stretch Locust Pose B (Uttana Shalabhasana B)

Modification: fingers pointing to the front of the body, feet to the head, chest on the floor 1. side view 2. front view

Pose Type: prone, backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

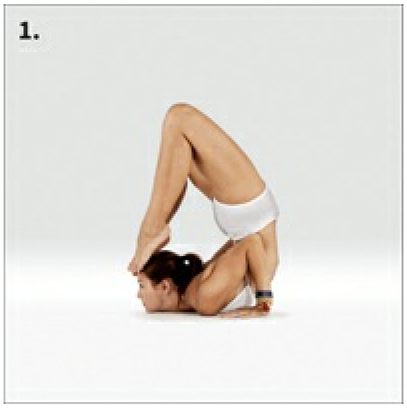

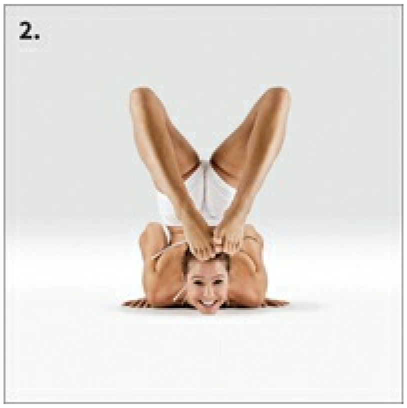

Formidable Face Pose

Ganda Bherundasana

(GUHN-duh bey-run-DAHS-uh-nuh)

Modification: fingers pointing to the back

1. knees apart

2. knees together, chin lifted off the floor

Pose Type: prone, backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

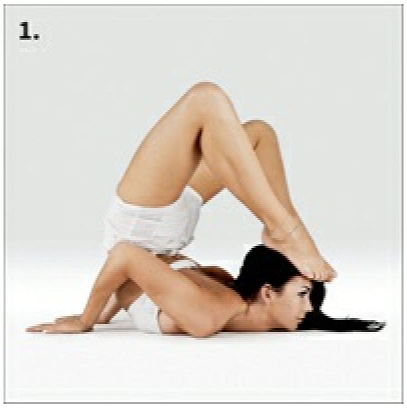

2.

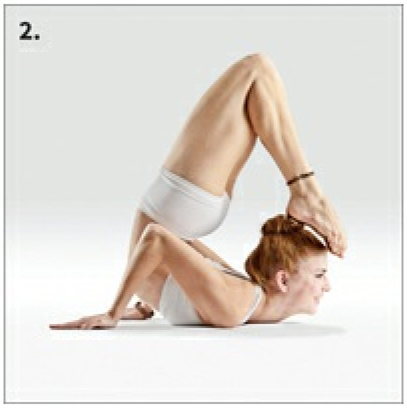

Inverted Locust Pose

Viparita Shalabhasana

(vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Also Known As: Full Locust Pose (Purna Shalabhasana)
Modification: arms straight on the floor, palms down; feet to the head
Pose Type: prone, backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

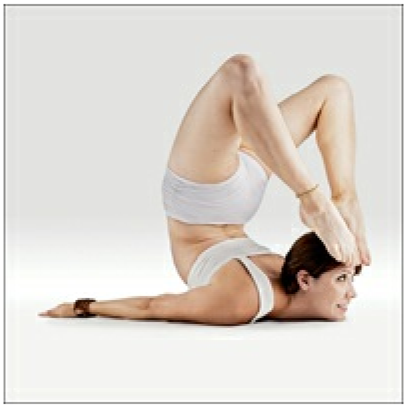

Formidable Face Pose

Ganda Bherundasana

(GUHN-duh bey-run-DAHS-uh-nuh)

Modification: feet to the head, grabbing onto both feet, chin off the floor

Pose Type: prone, backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल र य ह ॐ

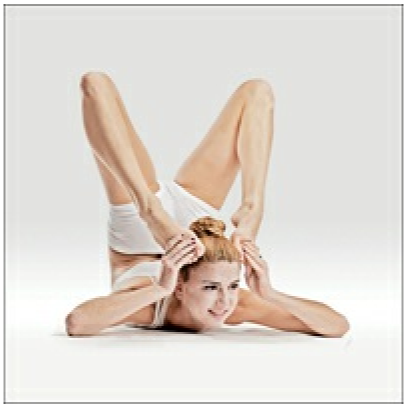

CHIN ON THE FLOOR: FULL LOTUS

Inverted Lotus Peacock Pose

Viparita Padma Mayurasana

(vi-puh-REE-tuh PUHD-muh muh-yoor-AHS-uh-nuh)

Pose Type: prone, backbend, inversion

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

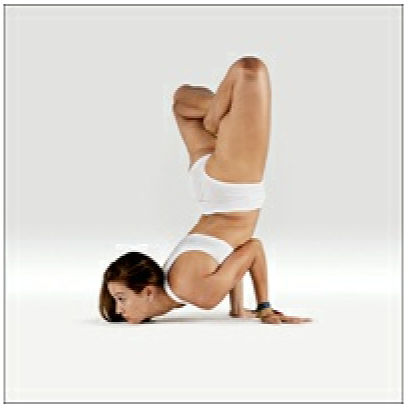

Inverted Locust Pose with Lotus Legs

Padma Viparita Shalabhasana

(PUHD-muh vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Pose Type: prone, backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

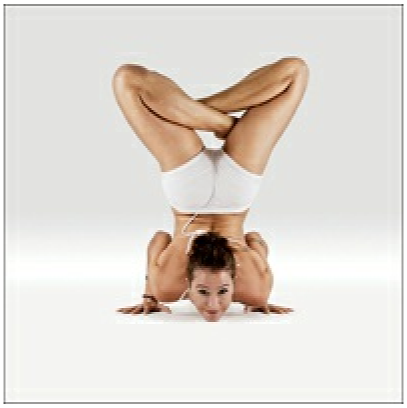

CHEST ON THE FLOOR: FEET TO THE FLOOR

Inverted Locust Pose

Viparita Shalabhasana

(vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Modification: arms straight on the floor, palms down; feet toward the floor over the head, legs bent
1. heels up
2. heels down

Pose Type: prone, backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

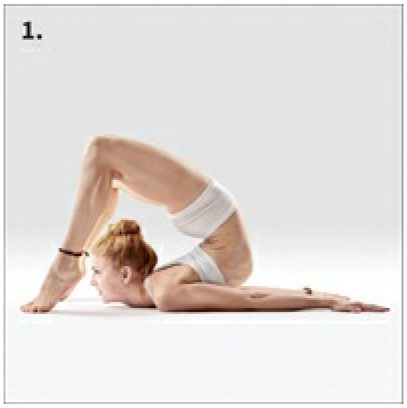

1.

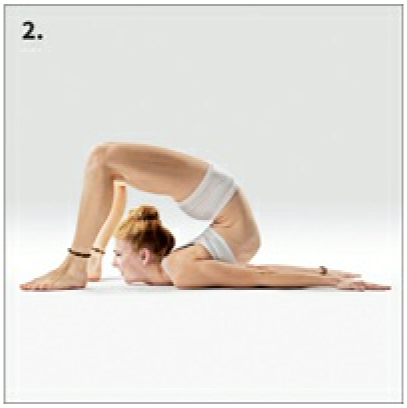

Formidable Face Pose

Ganda Bherundasana

(GUHN-duh bey-run-DAHS-uh-nuh)

Modification: feet on the floor, grabbing onto both ankles, chin off the floor
Pose Type: prone, backbend, inversion
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

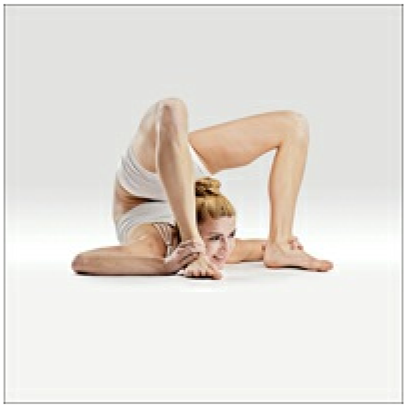

CHEST AND KNEES ON THE FLOOR: KNEES UNDER THE HIPS

Intense Extended Puppy Dog Pose

Uttana Shvanakasana

(ut-TAHN-uh shvuh-nuh-KAHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification), Extended Dog Pose (Utthita Shvanasana)
Modification: 1. palms down to the floor 2. grabbing onto the heels
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

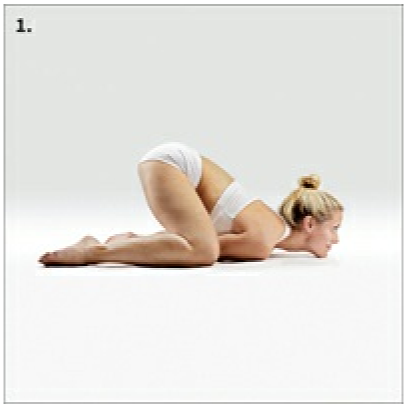

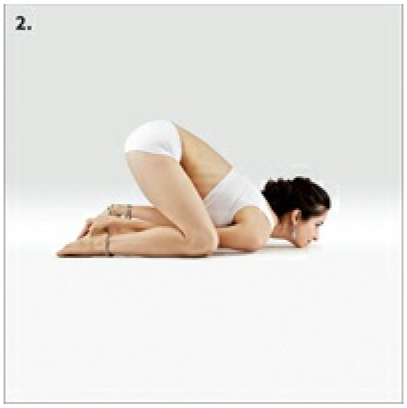

Intense Extended Puppy Dog Pose

Uttana Shvanakasana

(ut-TAHN-uh shvuh-nuh-KAHS-uh-nuh)

Also Known As: Prayer Heart Pose (Namaskar Anahatasana), Extended Dog Pose (Utthita Shvanasana)
Modification: thumbs to the upper back
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

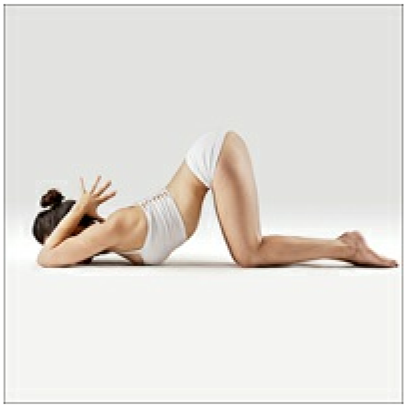

Water Grove Pose

Nirakunjasana

(neer-uh--kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)
Modification: arms in front, elbows bent; palms together; chin off the floor; heels toward the sitting bones
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)

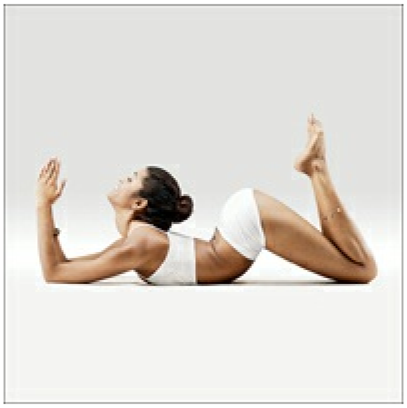

Hands Bound Water Grove Pose

Baddha Hasta Nirakunjasana

(BUH-duh HUH-stuh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)
Modification: heels toward the sitting bones
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

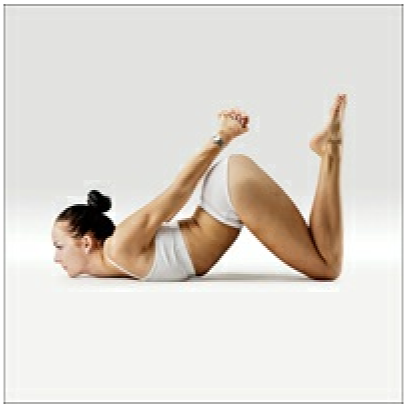

Pose Dedicated to Goddess Arani Prep.

Aranyasana Prep.

(uh-rah-NYAHS-uh-nuh)

Also Known As: Dragon Pose

Modification: 1. hands underneath the shoulders, elbows tucked in

2. arms reaching toward the knees, palms up

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

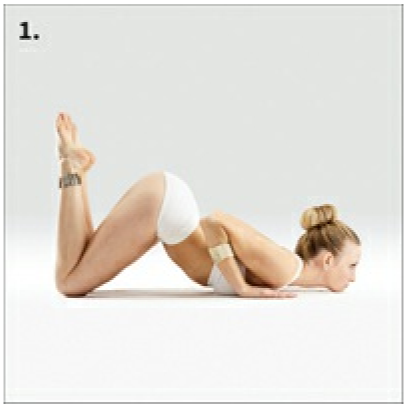

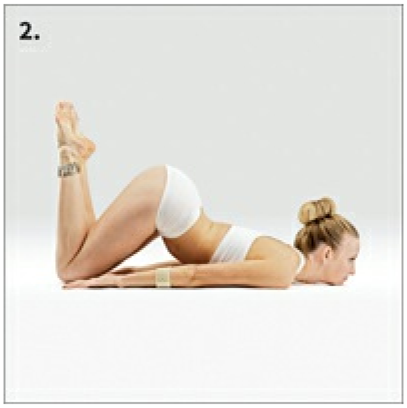

Pose Dedicated to Goddess Arani

Aranyasana

(uh-rah-NYAHS-uh-nuh)

Also Known As: Dragon Pose

Modification: 1. grabbing onto the shins

2. grabbing onto the hamstrings

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

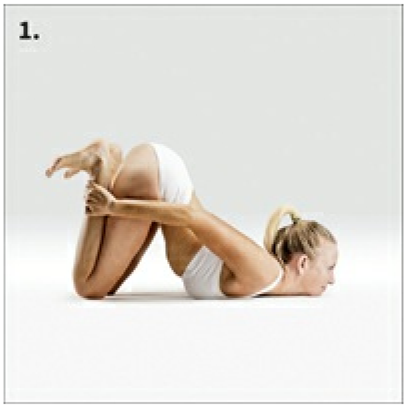

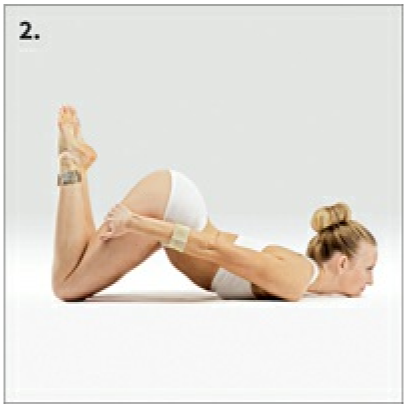

CHEST AND KNEES ON THE FLOOR: KNEES UNDER THE HIPS—HALF LOTUS

Intense Extended Puppy Dog Pose

Uttana Shvanakasana

(ut-TAHN-uh shvuh-nuh-KAHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification), Extended Dog Pose (Utthita Shvanasana)
Modification: arms extended to the front
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)

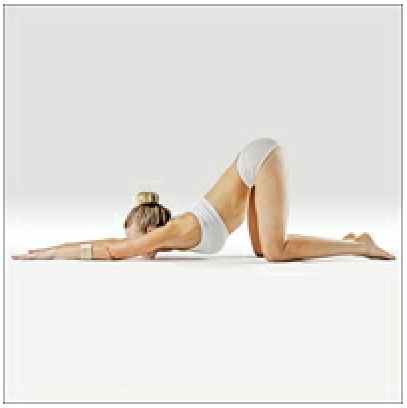

Half Lotus Intense Extended Puppy Dog Pose

Ardha Padma Uttana Shvanakasana

(UHR-duh PUHD-muh ut-TAHN-uh shvuh-nuh-KAHS-uh-nuh)

Also Known As: Half Lotus Heart Pose (Ardha Padma Anahatasana), Half Lotus Extended Dog Pose (Ardha Padma Utthita Shvanasana)
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)
लं वं रं यं हं ॐ

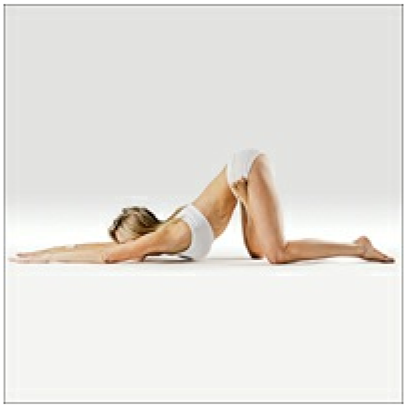

Half Bound Lotus Intense Extended Puppy Dog Pose

Ardha Baddha Padma Uttana Shvanakasana

(UHR-duh BUH-duh PUHD-muh ut-TAHN-uh shvuh-nuh-KAHS-uh-nuh)

Also Known As: Half Bound Lotus Heart Pose (Ardha Baddha Padma Anahatasana), Half Bound Lotus Extended Dog Pose (Ardha Baddha Padma Utthita Shvanasana)
Pose Type: prone, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)
लं वं रं यं हं ॐ

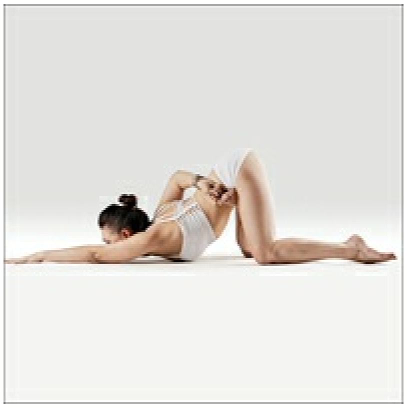

Lotus Intense Extended Puppy Dog Pose

Padma Uttana Shvanakasana

(PUHD-muh ut-TAHN-uh shvuh-nuh-KAHS-uh-nuh)

Also Known As: Lotus Heart Pose (Padma Anahatasana), Lotus Extended Dog Pose (Padma Utthita Shvanasana)
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)

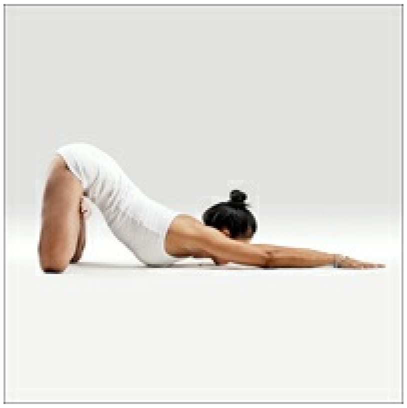

Upward Hands Bound Lotus Pose in Pose Dedicated to Goddess Arani

Urdhva Baddha Hasta Padmasana in Aranyasana

(OORD-vuh BUH-duh HUH-stuh puhd-MAHS-uh-nuh in uh-rah-NYAHS-uh-nuh)

Also Known As: Dragon Pose, Upward Bound Hand Lotus Heart Pose (Urdhva Baddha Hasta Padma Anahatasana)
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

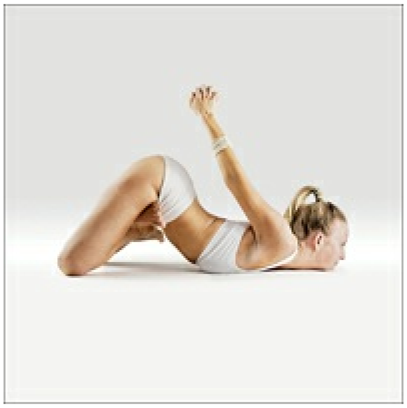

Feet Spread Wide Pose in Inverted Locust Pose

Prasarita Padottanasana in Viparita Shalabhasana

(pruh-SAH-ri-tuh puh-do-tahn-AHS-uh-nuh in vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

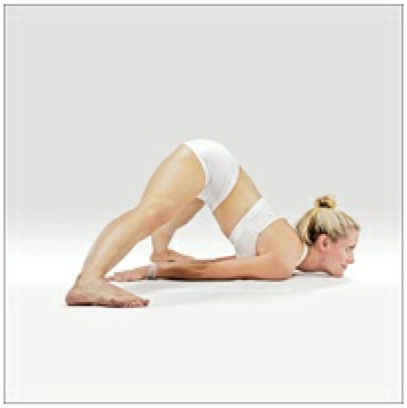

CHEST AND KNEES ON THE FLOOR: KNEES BEHIND HIPS—ARMS IN FRONT AND BEHIND THE BACK

Water Grove Pose

Nirakunjasana

(neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)
Modification: arms straight and crossed in front
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

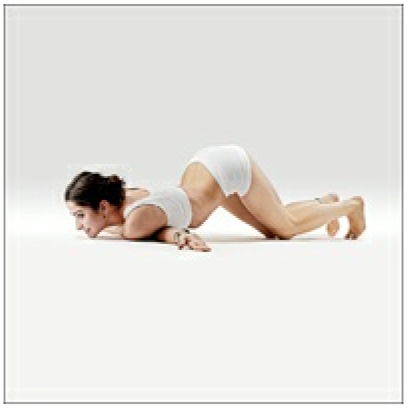

Water Grove Pose

Nirakunjasana

(neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)
Modification: chin and feet on the floor
1. knees on the floor
2. knees off the floor
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

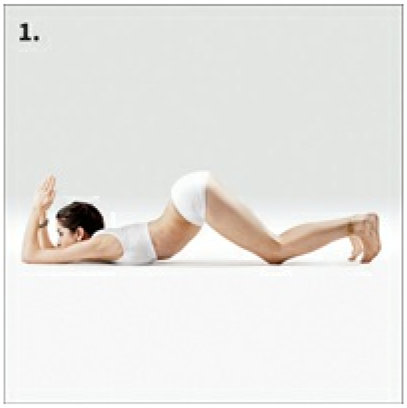
1.

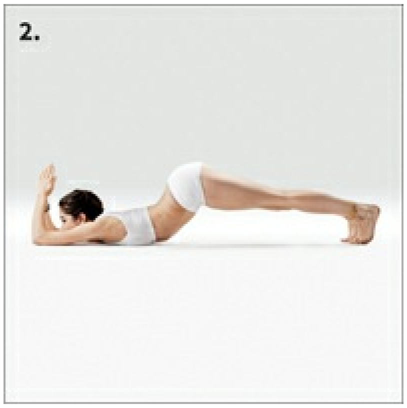
2.

Eight Limbs Pose

Ashtangasana

(uhsh-tahng-AHS-uh-nuh)

Also Known As: Eight Point Bow Pose (Ashtanga Namaskara)
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

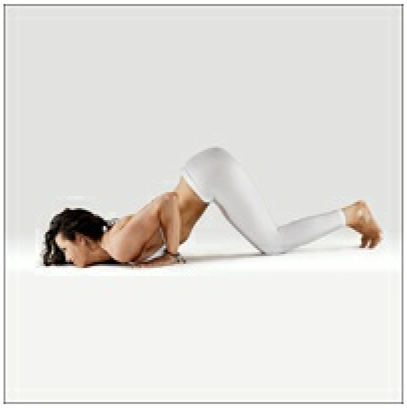

Reverse Prayer Water Grove Pose

Viparita Namaskar Nirakunjasana

(vi-puh-REE-tuh nuh-muhs-KAHR neer-uh-kunj-AHS-uh-nuh)

Also Known As: Back of the Body Prayer Water Grove Pose (Paschima Namaskara Nirakunjasana), Heart Pose Modification
Modification: toes curled in
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

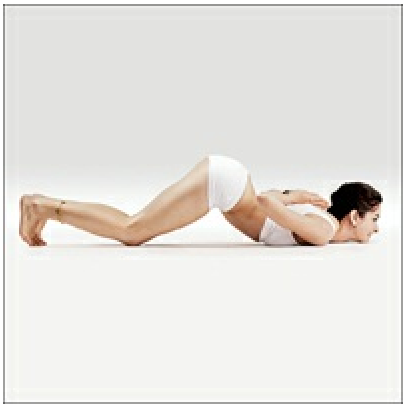

Hands Bound Water Grove Pose

Baddha Hasta Nirakunjasana

(BUH-duh HUH-stuh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)
Modification: 1. knees on the floor, toes pointed to the back
2. knees on the floor, toes curled in
3. knees off the floor
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

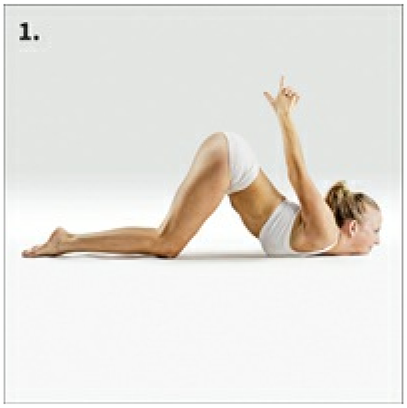

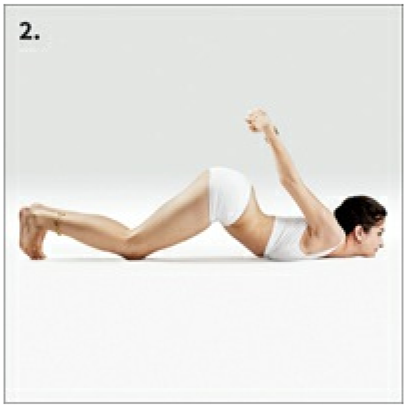

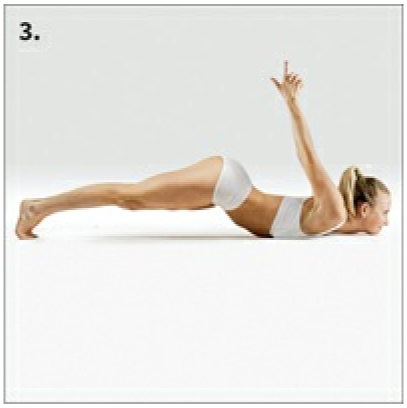

One-Legged Water Grove Pose

Eka Pada Nirakunjasana

(EY-kuh PUH-duh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)

Modification: 1. forearms on the floor, toes pointed

2. palms to the floor by the rib cage, elbows bent, toes curled in

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ● ● ● ●

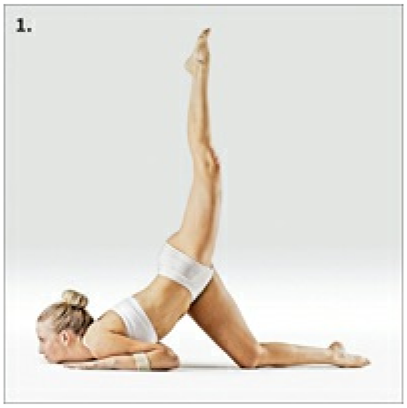
1.

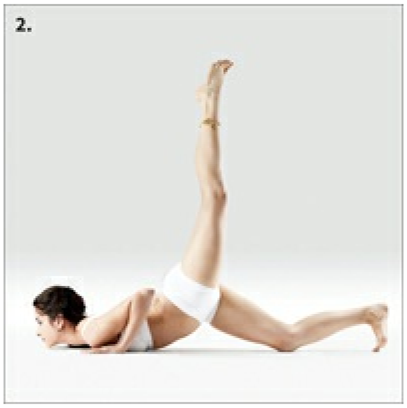
2.

Hands Bound One-Legged Water Grove Pose

Baddha Hasta Eka Pada Nirakunjasana

(BUH-duh HUH-stuh EY-kuh PUHD-uh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

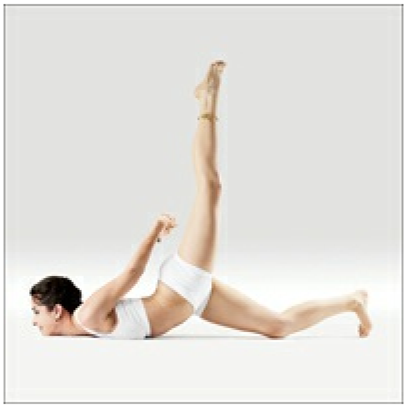

Reverse Prayer One-Legged Water Grove Pose

Viparita Namaskar Eka Pada Nirakunjasana

(vi-puh-REE-tuh nuh-muhs-KAHR EY-kuh PUH-duh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Back of the Body Prayer One-Legged Water Grove Pose (Paschima Namaskara Eka Pada Nirakunjasana), Heart Pose
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

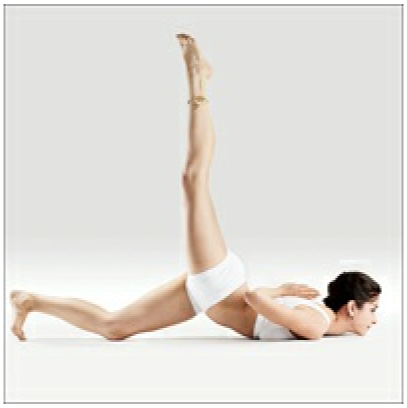

CHEST AND KNEES ON THE FLOOR: KNEES BEHIND HIPS—ONE LEG STRAIGHT AND LIFTED—BOTTOM KNEE BENT

One-Legged Inverted Locust Pose Prep.

Eka Pada Viparita Shalabhasana Prep.

(EY-kuh PUH-duh vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Modification: forearms on the floor, knee of straight leg resting on the foot of the bottom leg
Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ह ॐ

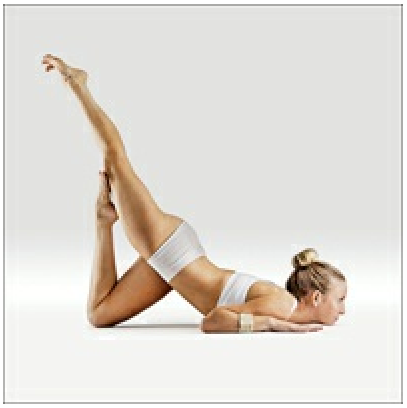

Revolved One-Legged Inverted Locust Pose Prep.

Parivritta Eka Pada Viparita Shalabhasana Prep.

(puh-ri-VRIT-tuh EY-kuh PUH-duh vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)
Modification: bottom leg bent, top leg straight, top knee resting on the sole of the bottom foot
Pose Type: forward bend, twist
Drishti Point: 1. Urdhva or Antara Drishti (up to the sky)
2. Hastagrai or Hastagre (hands)
ल व र य

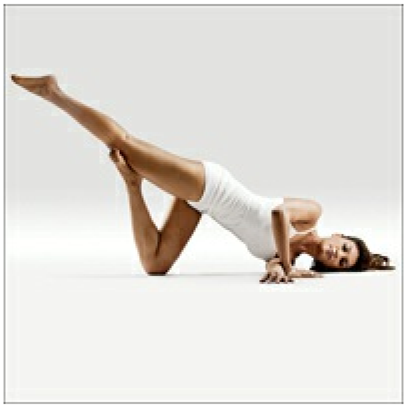

Upward Bound Hands One-Legged Inverted Locust Pose

Urdhva Baddha Hasta Eka Pada Viparita Shalabhasana

(OORD-vuh BUH-duh HUH-stuh EY-kuh PUH-duh vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)
Also Known As: Flying Locust Pose (Uddayate Shalabhasana)
Modification: 1. knee of the straight leg resting on the foot of the bottom leg
2. bottom foot away from the top leg
Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

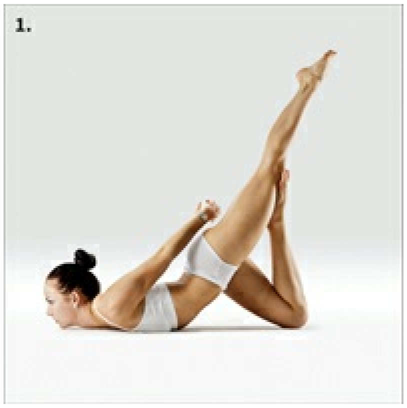
1.

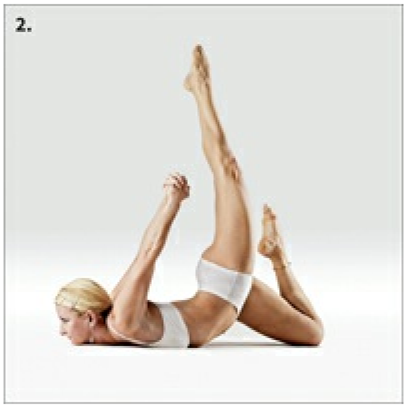
2.

CHEST AND KNEES ON THE FLOOR: KNEES BEHIND HIPS—BOTH LEGS BENT—ARMS IN FRONT

One-Legged Water Grove Pose

Eka Pada Nirakunjasana

(EY-kuh PUH-duh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)
Modification: both knees bent; arms straight out in front, palms down
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)

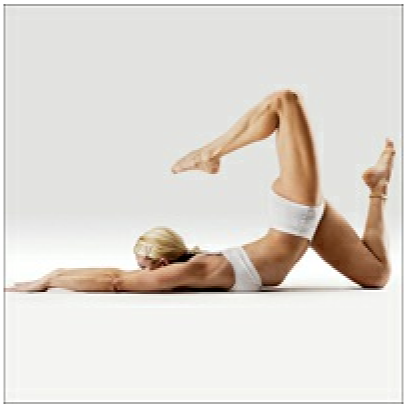

Half Bow Pose in One-Legged Water Grove Pose

Ardha Dhanurasana in Eka Pada Nirakunjasana

(IUHR-duh duh-nur-AHS-uh-nuh in EY-kuh PUH-duh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)
Modification: 1. right front side view
2. right back side view
Pose Type: backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Angushtamadhye or Angushta Ma Dyai (thumbs)
ल व र य ह ॐ

1.

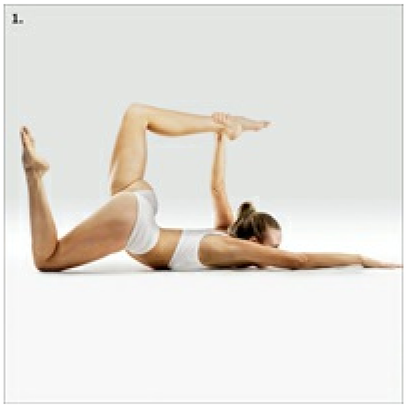

2.

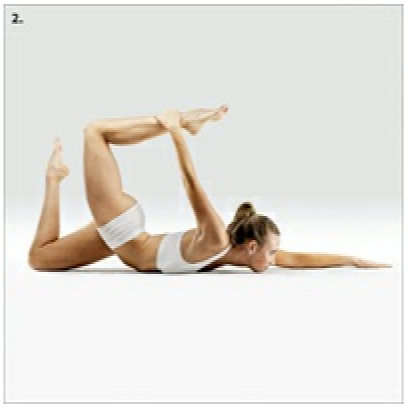

One-Legged Water Grove Pose

Eka Pada Nirakunjasana

(EY-kuh PUH-duh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)

Modification: both knees bent

1. arms straight to the back, palms down
2. arms lifted off the floor

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

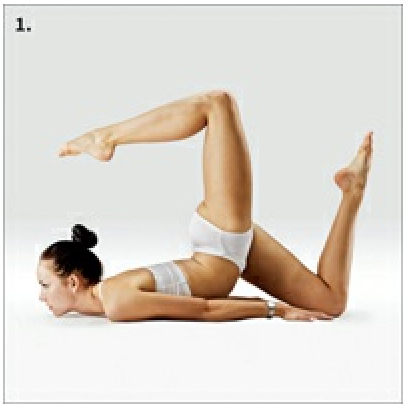

1.

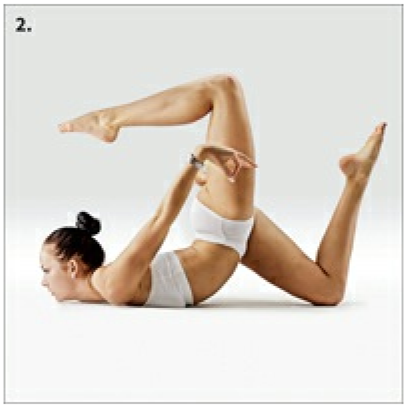

2.

Bow Pose in One-Legged Water Grove Pose

Dhanurasana in Eka Pada Nirakunjasana

(duh-nur-AHS-uh-nuh in EY-kuh PUH-duh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)

Modification: both hands grabbing the top foot with under-head grip; the knee of the bottom leg on the floor, toes pointing to the sky

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

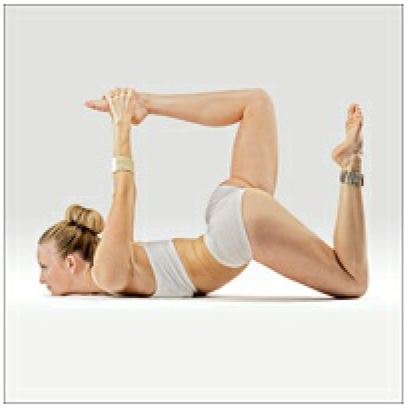

Half Bow Pose in One-Legged Water Grove Pose

Ardha Dhanurasana in Eka Pada Nirakunjasana

(UHR-duh duh-nur-AHS-uh-nuh in EY-kuh PUH-duh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)

Modification: grabbing onto the foot on the opposite side, under-head grip; other hand to the lower back, palm up

Pose Type: prone, backbend, binding

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

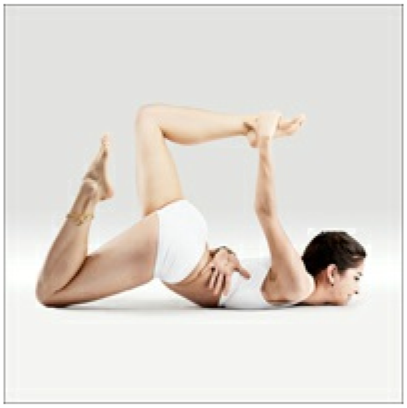

CHEST AND KNEES ON THE FLOOR: KNEES BEHIND HIPS—BOTTOM KNEE BENT—OVERHEAD GRIP

Big Toe Bow Pose in One-Legged Water Grove Pose

Padangushta Dhanurasana in Eka Pada Nirakunjasana

(puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh in EY-kuh PUH-duh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)

Pose Type: prone, backbend

Drishti Point: Padayoragrai or Padayoragre (toes/feet), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

लं वं रं यं हं ॐ

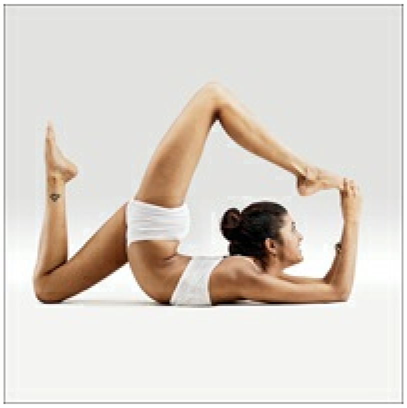

One Leg to the Head Water Grove Pose

Eka Pada Shirsha Nirakunjasana

(EY-kuh PUH-duh SHEER-shuh neer-uh-kunj-AHS-uh-nuh)

Also Known As: Heart Pose (Anahatasana Modification)
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

लं वं रं यं हं ॐ ॐ

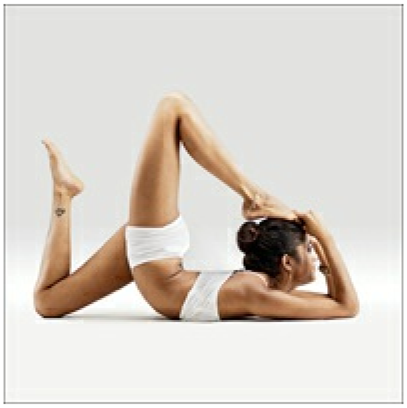

CHEST AND KNEES ON THE FLOOR: KNEES BEHIND HIPS—BOTTOM KNEE BENT—UNDER-HEAD & OVERHEAD GRIP

One-Legged Inverted Locust Pose

Eka Pada Viparita Shalabhasana

(EY-kuh PUH-duh vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Modification: one palm to the floor at the bottom of the ribs, other arm straight, palm down; both knees bent, knee of the top leg resting on the foot of the bottom leg

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

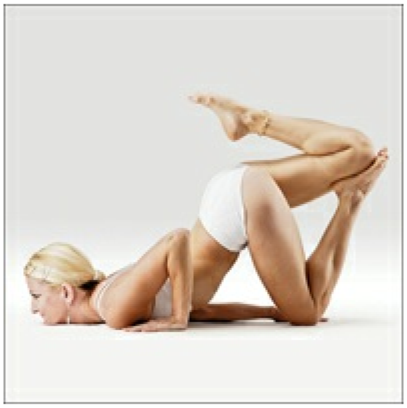

One Hand to Foot Inverted Locust Pose

Eka Hasta Pada Viparita Shalabhasana

(EY-kuh HUH-stuh PUH-duh vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Modification: grabbing onto the top foot with the opposite hand; other arm straight, palm down to the floor; both knees bent, knee of the top leg resting on the foot of the bottom leg

Pose Type: prone, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ॐ

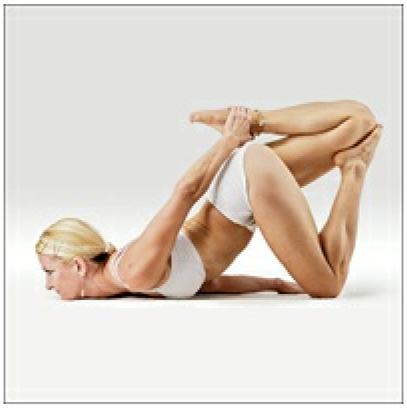

Both Hands to One Foot Inverted Locust Pose

Dwi Hasta Eka Pada Viparita Shalabhasana

(DWI-huh-stuh EY-kuh PUH-duh vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)

Modification: both hands grabbing the top foot with under-head grip; top knee resting on the sole of the bottom foot

Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

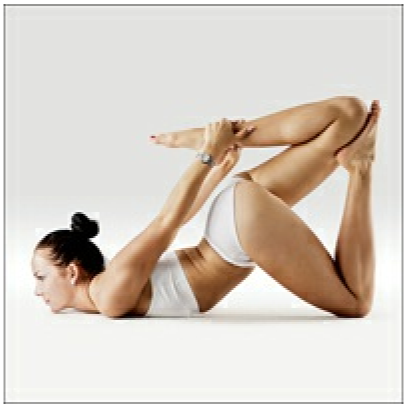

One Foot to the Shoulder Inverted Locust Pose

Eka Pada Bhuja Viparita Shalabhasana

(EY-kuh PUH-duh buj-uh vi-puh-REE-tuh shuh-luh-BAHS-uh-nuh)
Modification: grabbing onto the toes of the top foot; the quadriceps of the top leg resting on the sole of the bottom foot
Pose Type: prone, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

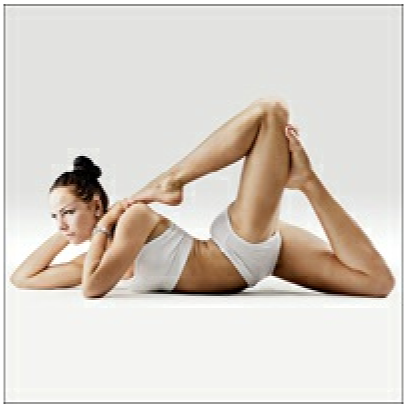

Supine Poses

SHOULDER ON THE FLOOR
TRICEPS ON THE FLOOR
ELBOW ON THE FLOOR
FOREARM ON THE FLOOR
EASTERN INTENSE STRETCH
BRIDGE POSES
HEAD AND SITTING BONES ON THE FLOOR
BOUND ANGLE POSE
KNEES BENT
HAPPY BABY POSE
ONE AND BOTH FEET BEHIND HEAD
HORIZONTAL SPLITS
ONE KNEE BENT TO THE BACK (VARIATIONS)
VERTICAL SPLITS
BOTH LEGS STRAIGHT AND TOGETHER
SUPINE LOTUS AND CORPSE POSE

Sideways Wind Relieving Pose

Parshva Vayu Muktyasana

(PAHRSH-vuh VAH-yu muk-TYAHS-uh-nuh)

Modification: palms together in front of the face, knees together
Pose Type: supine (on the side)
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

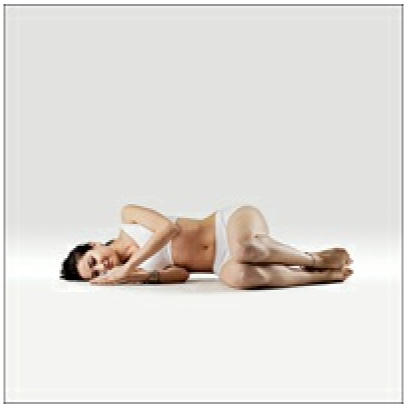

Arms Raised Sideways Wind Relieving Pose

Urdhva Hasta Parshva Vayu Muktyasana

(OORD-vuh HUH-stuh PAHRSH-vuh VAH-yu muk-TYAHS-uh-nuh)

Pose Type: supine (on the side)
Drishti Point: Urdhva or Antara Drishti (up to the sky)

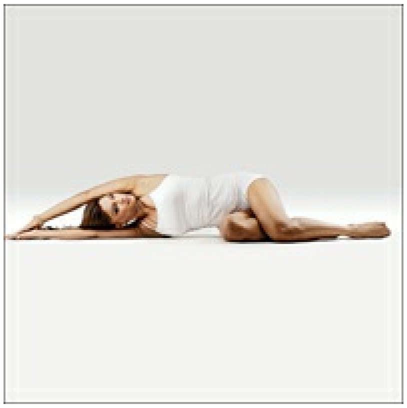

Sideways Wind Relieving Pose

Parshva Vayu Muktyasana

(PAHRSH-vuh VAH-yu muk-TYAHS-uh-nuh)

Modification: bottom leg straight; top knee bent, foot to the knee of the straight leg

Pose Type: supine (on the side)

Drishti Point: Nasagrai or Nasagre (nose)

ल व

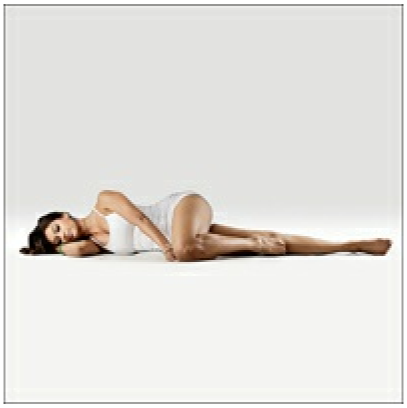

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose

Modification: top leg lifted in front of the body, bottom leg straight

Pose Type: supine (on the side), forward bend

Drishti Point: Nasagrai or Nasagre (nose), Padayoragrai or Padayoragre (toes/feet)

ल व

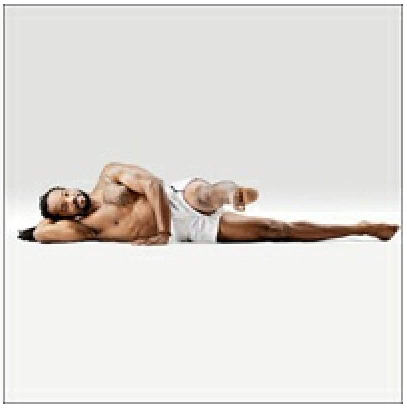

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose

Modification: top leg straight and lifted; bottom leg bent, toes pointing to the back
Pose Type: supine (on the side)
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

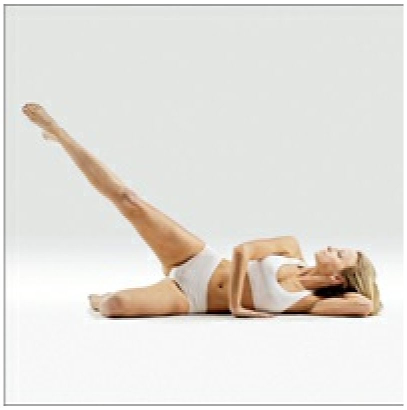

SHOULDER ON THE FLOOR: TWISTS AND BINDING

Svastika Legs in Twisted Stomach Pose

Pada Svastikasana in Jatara Parivartanasana

(PUH-duh svuh-sti-KAHS-uh-nuh in JAH-tuh-ruh puh-ri-vuhrt-ahn-AHS-uh-nuh)
Modification: grabbing onto the back foot, heel toward the glutes
Pose Type: supine (on the side), twist
Drishti Point: Hastagrai or Hastagre (hands)

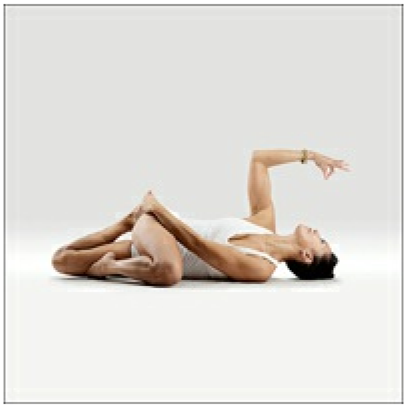

Revolved One-Legged King Pigeon Pose 1

Parivritta Eka Pada Raja Kapotasana 1

(puh-ri-VRIT-tuh EY-kuh PUH-duh RAH-juh kuh-po-TAHS-uh-nuh)
Modification: back knee bent, both hands to the back foot; front foot to hip socket

Pose Type: supine (on the side), twist
Drishti Point: Urdhva or Antara Drishti (up to the sky)
● ● ● ●

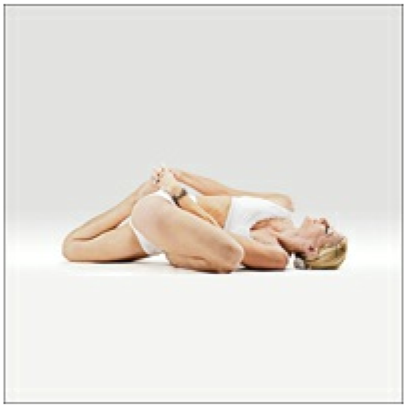

One-Legged Bound Infinity Pose

Eka Pada Baddha Anantasana

(EY-kuh PUH-duh BUH-duh uhn-uhnt-AHS-uh-nuh)
Also Known As: One-Legged Bound Sleeping Vishnu Pose
Modification: bottom leg bent, heel toward the glutes; top leg bent
Pose Type: supine (on the side), twist, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ● ●

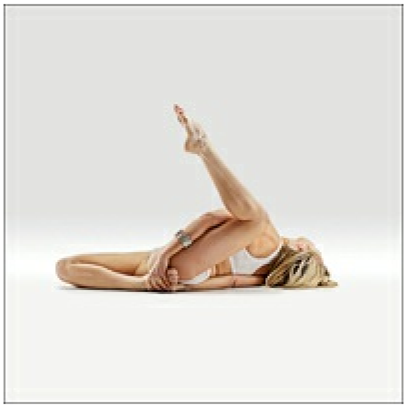

SHOULDER ON THE FLOOR: HALF LOTUS BACKBEND—UNDER-HEAD & OVERHEAD GRIP

Sideways Half Lotus Bow Pose

Parshva Ardha Padma Dhanurasana

(PAHRSH-vuh UHR-duh PUHD-muh duh-nur-AHS-uh-nuh)
Modification: grabbing onto the bottom foot with both hands with under-head grip

Pose Type: supine (on the side), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ● ● ●

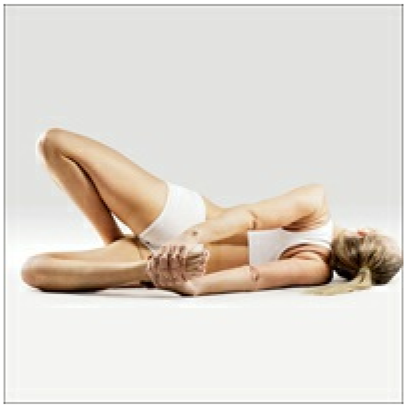

Sideways Half Lotus One-Handed Big Toe Bow Pose

Parshva Ardha Padma Eka Hasta Padangushta Dhanurasana

(PAHRSH-vuh UHR-duh PUHD-muh EY-kuh HUH-stuh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)
Modification: grabbing onto the bottom foot with overhead grip on the same side, top arm straight, fingertips to the sky
Pose Type: supine (on the side), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ● ● ● ●

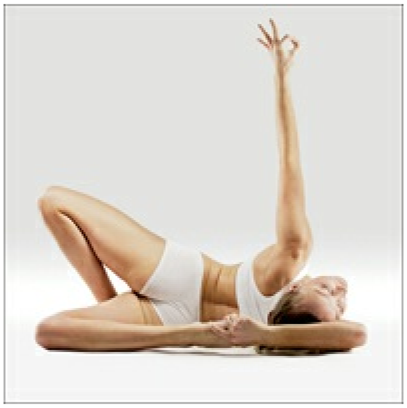

Sideways Half Lotus Two-Handed Big Toe Bow Pose

Parshva Ardha Padma Dwi Hasta Padangushta Dhanurasana

(PAHRSH-vuh UHR-duh PUHD-muh DWI-huh-stuh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh)
Modification: overhead grip, both hands grabbing onto the foot
Pose Type: supine (on the side), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

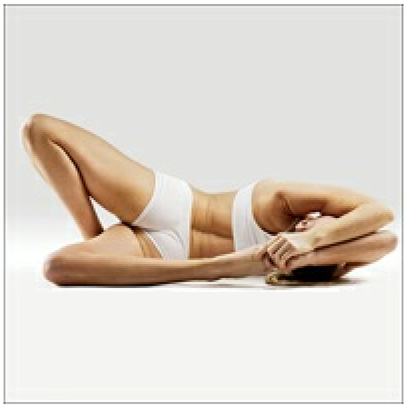

SHOULDER ON THE FLOOR: FOOT TO THE THIGH, HALF LOTUS & LOTUS

Mermaid Pose in Infinity Pose

Naginyasana in Anantasana

(nuh-gin-YAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Mermaid Pose in Sleeping Vishnu Pose
Pose Type: supine (on the side), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

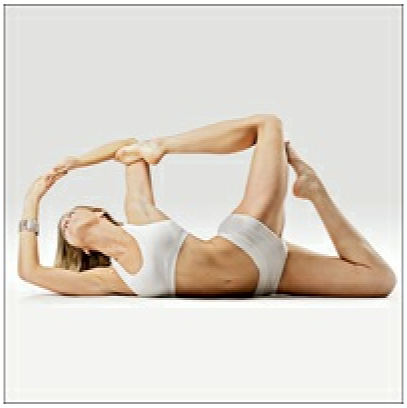

Sideways Half Lotus Bow Pose

Parshva Ardha Padma Dhanurasana

(PAHRSH-vuh UHR-duh PUHD-muh duh-nur-AHS-uh-nuh)

Modification: grabbing onto the top foot with the bottom hand; grabbing onto the bottom foot with top hand
1. front view
2. back view

Pose Type: supine (on the side), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

1.

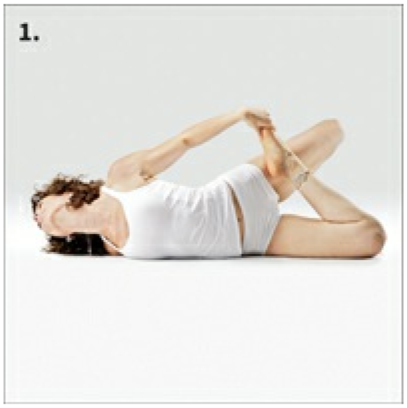

2.

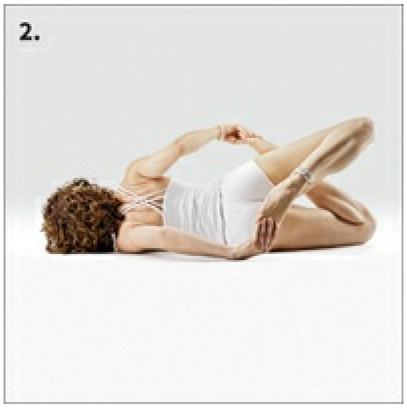

Hands Bound Lotus Pose in Infinity Pose

Baddha Hasta Padmasana in Anantasana

(BUH-duh HUH-stuh puhd-MAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)
Also Known As: Hands Bound Lotus Pose in Sleeping Vishnu Pose
Pose Type: supine (on the side), mild backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

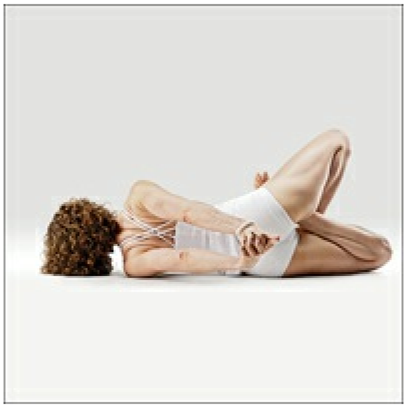

Hands Bound Infinity Pose

Baddha Hasta Anantasana

(BUH-duh HUH-stuh uhn-uhnt-AHS-uh-nuh)

Also Known As: Hands Bound Sleeping Vishnu Pose

Modification: 1. bottom leg bent, top leg straight and lifted; resting on the arms and the bent knee for support
2. bottom leg straight, top leg straight and lifted; arms lifted off the floor

Pose Type: supine (on the side), backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

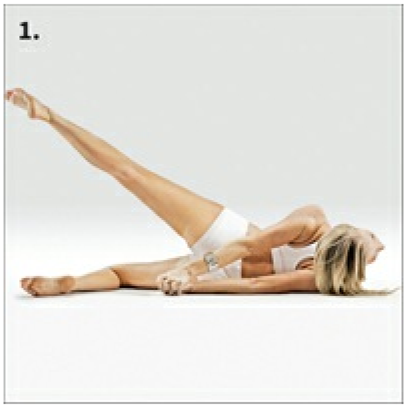

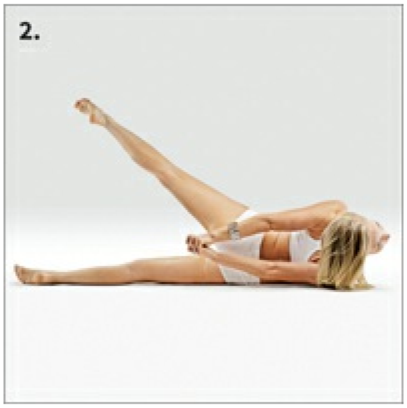

Hands Bound Half Lotus Pose in Infinity Pose

Baddha Hasta Ardha Padmasana in Anantasana

(BUH-duh HUH-stuh UHR-duh puhd-MAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Hands Bound Half Lotus Pose in Sleeping Vishnu Pose

Modification: 1. front view
2. back view

Pose Type: supine (on the side), backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

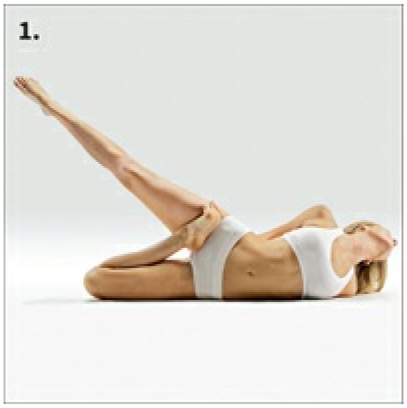

2.

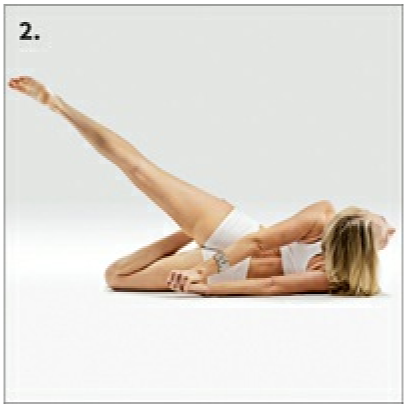

TRICEPS ON THE FLOOR: HAND TO THE HEAD—LEGS TOGETHER & LEGS APART

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose
Modification: both legs straight and together; bottom leg on the floor
Pose Type: supine (on the side), side bend
Drishti Point: Padhayoragrai or Padayoragre (toes/feet)

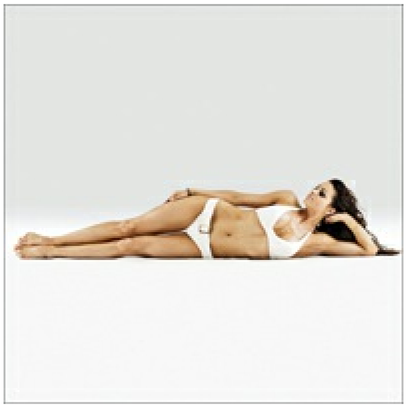

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose
Modification: legs straight, together and lifted off the floor
Pose Type: supine (on the side), core, side bend
Drishti Point: Nasagrai or Nasagre (nose)
ल

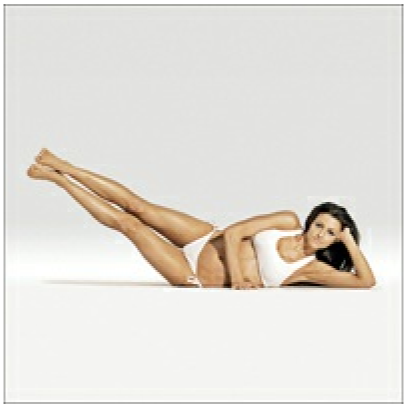

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose
Modification: both legs straight, top leg lifted, bottom leg on the floor
Pose Type: supine (on the side), core, side bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Urdhva or Antara Drishti (up to the sky)
ल व

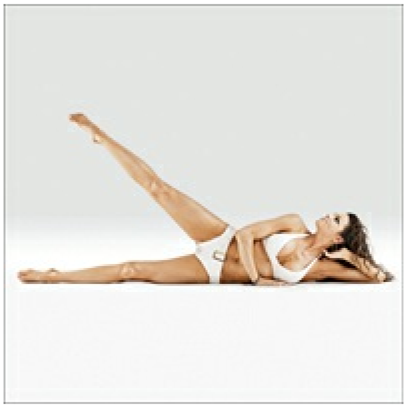

Extended Hand to Big Toe Pose in Infinity Pose

Utthita Hasta Padangushtasana in Anantasana

(UT-ti-tuh HUH-stuh puhd-ahng-goosh-TAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Extended Hand to Big Toe Pose in Sleeping Vishnu Pose
Pose Type: supine (on the side), forward bend, side bend
Drishti Point: Nasagrai or Nasagre (nose)

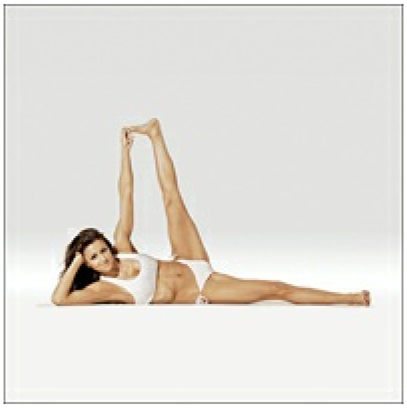

Sundial Pose in Infinity Pose

Surya Yantrasana in Anantasana

(SOOR-yuh yuhn-TRAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Sundial Pose in Sleeping Vishnu Pose
Modification: 1. grabbing onto the heel on the outside of the calf
2. grabbing onto the heel on the inside of the calf
Pose Type: supine (on the side), forward bend, side bend
Drishti Point: Nasagrai or Nasagre (nose)

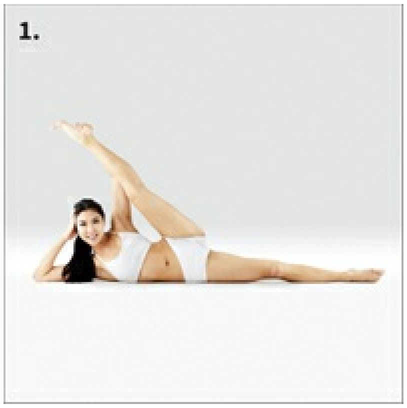

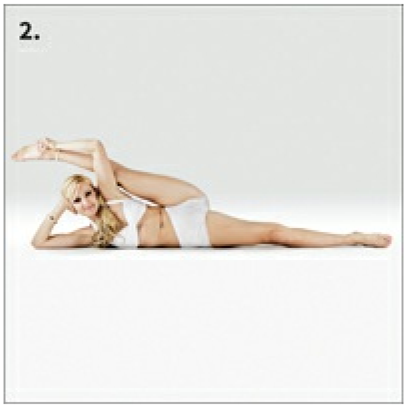

TRICEPS ON THE FLOOR: BOTH KNEES BENT

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose
Modification: top knee to the top shoulder; both knees bent
Pose Type: supine (on the side), forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Padhayoragrai or Padayoragre (toes/feet)
ल व

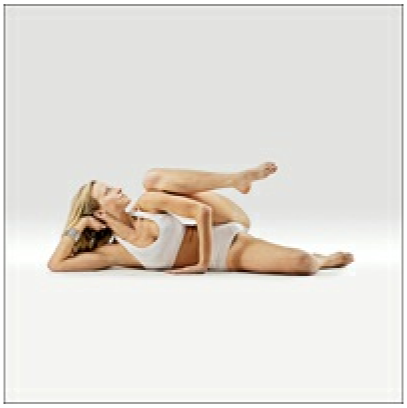

Bound Angle Pose in Infinity Pose

Baddha Konasana in Anantasana

(BUH-duh ko-NAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Bound Angle Pose in Sleeping Vishnu Pose
Modification: 1. top wrist on the top knee
2. top hand grabbing onto the top foot
Pose Type: supine (on the side), side bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Hastagrai or Hastagre (hands)

1.

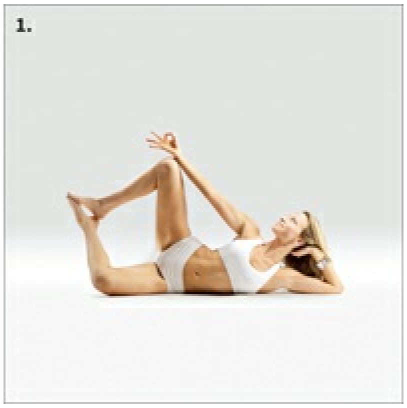

2.

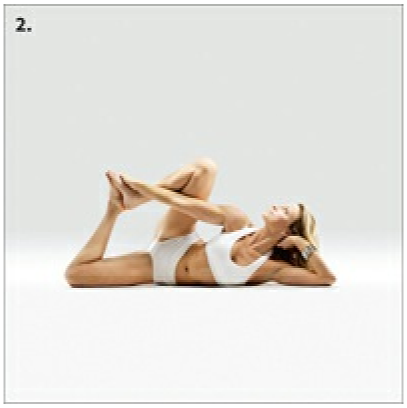

TRICEPS ON THE FLOOR: BOTTOM LEG STRAIGHT—BACKBENDS

Half Frog Pose Prep. in Infinity Pose

Ardha Bhekasana in Anantasana

(UHR-duh bey-KAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Half Frog Pose Prep. in Sleeping Vishnu Pose
Modification: hand in beginner modification; neutral spine, top leg moving into a half-frog leg position

Pose Type: supine (on the side), mild backbend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व र य

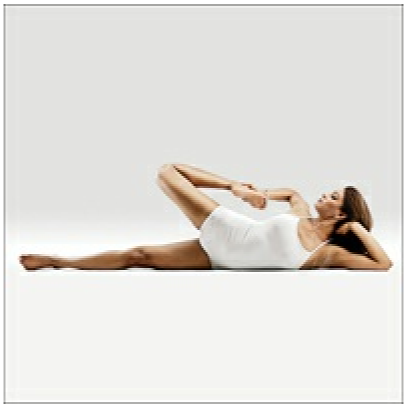

Half Bow Pose in Infinity Pose

Ardha Dhanurasana in Anantasana

(UHR-duh duh-nur-AHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)
Also Known As: Half Bow Pose in Sleeping Vishnu Pose
Pose Type: supine (on the side), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

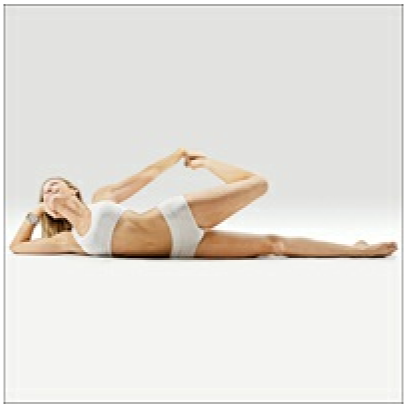

Half Big Toe Bow Pose in Infinity Pose

Ardha Padangushta Dhanurasana in Anantasana

(UHR-duh puhd-ahng-GOOSH-tuh duh-nur-AHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)
Also Known As: Half Big Toe Bow Pose in Sleeping Vishnu Pose
Pose Type: supine (on the side), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

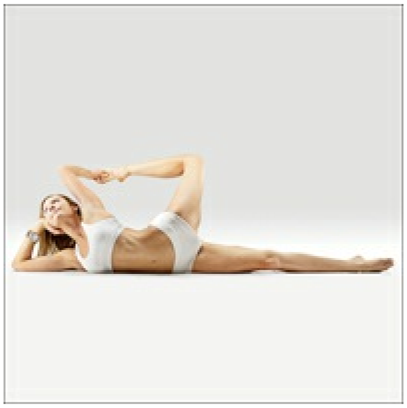

TRICEPS ON THE FLOOR: ONE LEG STRAIGHT, OTHER FOOT TO THE HIP SOCKET

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose
Modification: top foot to the hip socket, toes touching the floor; bottom leg straight on the floor
Pose Type: supine (on the side), side bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)
ल व

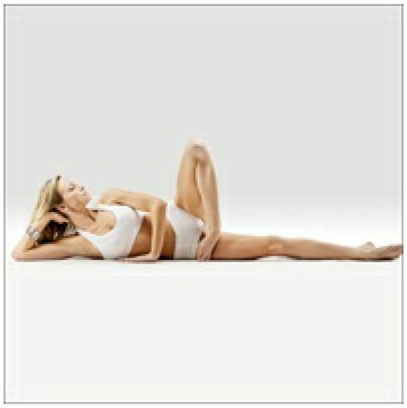

Half Lotus Pose in Infinity Pose

Ardha Padmasana in Anantasana

(UHR-duh puhd-MAHS-uh-nuh in anan-TAHS-annauhn-uhnt-AHS-uh-nuh)

Also Known As: Half Lotus Pose in Sleeping Vishnu Pose
Modification: bottom foot in half lotus, top leg lifted and straight
Pose Type: supine (on the side), side bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

ल व

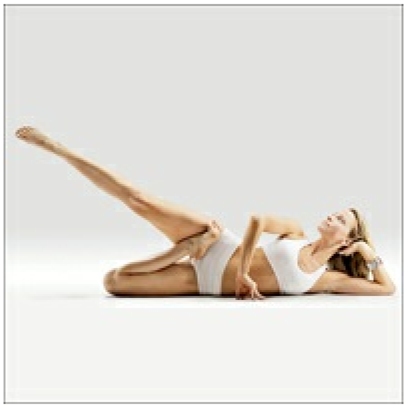

TRICEPS ON THE FLOOR: HALF LOTUS & FULL LOTUS

Half Lotus Mermaid Pose in Infinity Pose

Ardha Padma Naginyasana in Anantasana

(UHR-duh PUHD-muh nuh-gin-YAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Half Lotus Mermaid Pose in Sleeping Vishnu Pose
Pose Type: supine (on the side), backbend, side bend
Drishti Point: Hastagrai or Hastagre (hands)

ल व र य

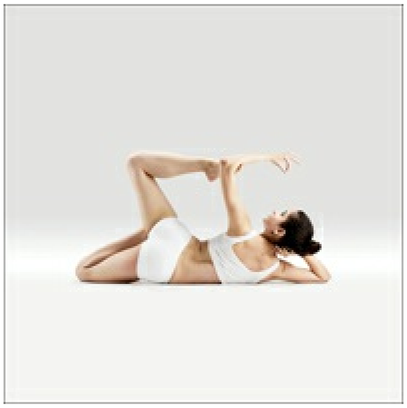

Lotus Pose in Infinity Pose

Padmasana in Anantasana

(puhd-MAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Lotus Pose in Sleeping Vishnu Pose
Pose Type: supine (on the side), side bend, knee, ankle opener
Drishti Point: Hastagrai or Hastagre (hands)

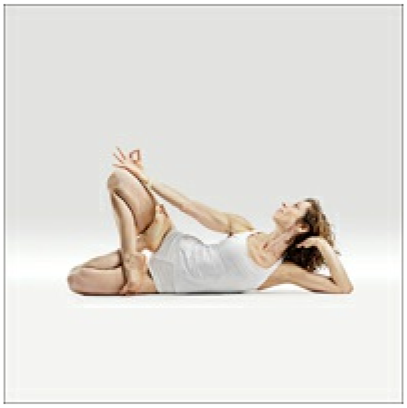

ELBOW ON THE FLOOR: BOTTOM LEG STRAIGHT, TOP LEG CROSSED OVER

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose

Modification: elbow to the floor, top knee bent, leg crossed over in front of the body; bottom leg straight

1. palm of the top hand on the floor
2. top hand on the top knee

Drishti Point: Urdhva or Antara Drishti (up to the sky)

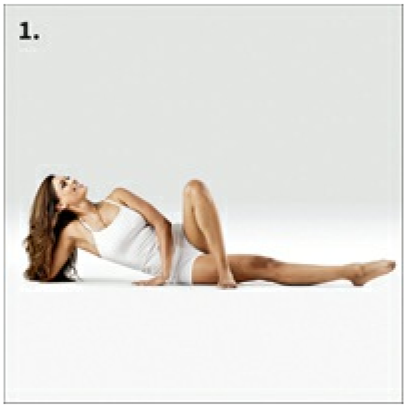

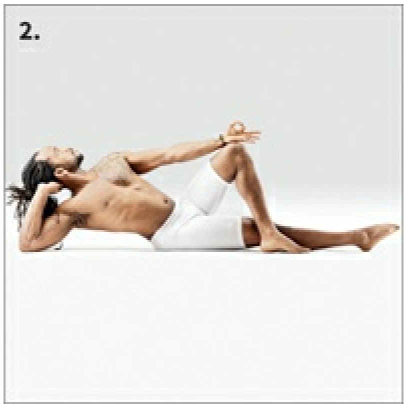

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose

Modification: elbow to the floor, hands in Anjali Mudra (Hands in Prayer); top knee bent, foot by the hip, bottom leg straight

1. bottom leg on the floor
2. bottom leg lifted off the floor

Pose Type: 1. supine (on the side), side bend

2. supine (on the side), side bend, core

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

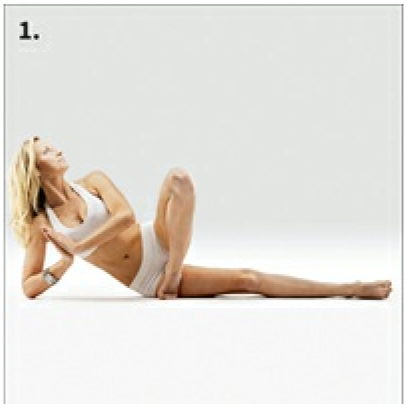

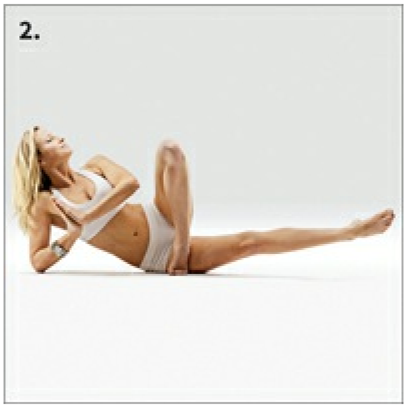

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose
Modification: top leg crossed over in front, toes on the floor; bottom leg bent, toes to the back; elbow to the floor, other arm straight fingertips to the sky
Pose Type: supine (on the side)
Drishti Point: Hastagrai or Hastagre (hands)
ल व य

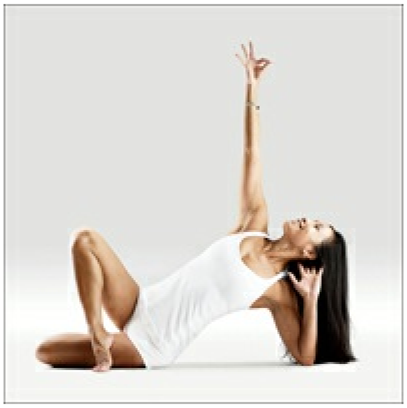

Half Lotus Pose in Infinity Pose

Ardha Padmasana in Anantasana

(UHR-duh puhd-MAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Half Lotus Pose in Sleeping Vishnu Pose
Modification: top leg in half lotus; bottom leg bent, heel to the sitting bone; elbow to the floor, other arm straight fingertips to the sky
Pose Type: supine (on the side)
Drishti Point: Hastagrai or Hastagre (hands)
ल व य

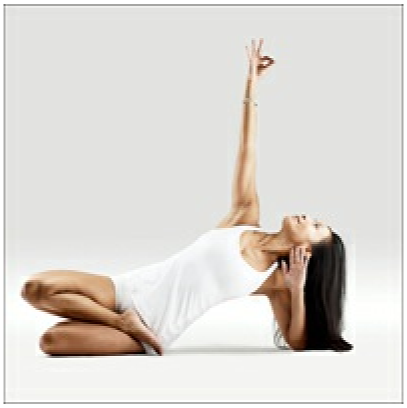

Leg Position of the Pose Dedicated to Garuda in Infinity Pose

Pada Garudasana in Anantasana

(PUH-duh guh-ru-DAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Leg Position of the Pose Dedicated to Garuda in Sleeping Vishnu Pose
Modification: elbow to the floor, hands behind the head
Pose Type: supine (on the side)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ●

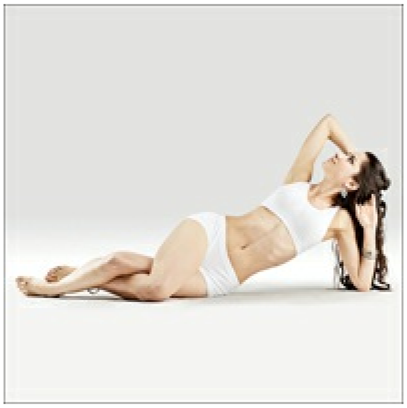

FOREARM ON THE FLOOR: ONE LEG CROSSED OVER & HALF LOTUS

Infinity Pose

Anantasana

(uhn-uhnt-AHS-uh-nuh)

Also Known As: Sleeping Vishnu Pose
Modification: top leg crossed over; forearm to the floor, top elbow resting on the top knee
Pose Type: supine (on the side)
Drishti Point: Hastagrai or Hastagre (hands)

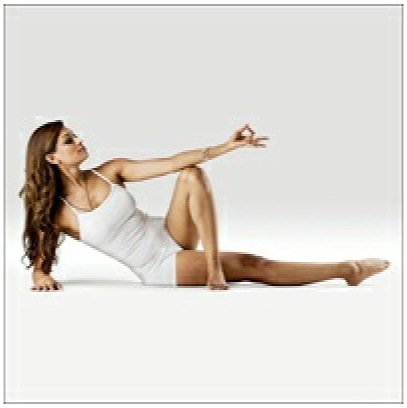

Extended Hand to Big Toe Half Cow Face Western Intense Stretch in Infinity Pose

Utthita Hasta Padangushta Ardha Gomukha Paschimottanasana in Anantasana

(UT-ti-tuh HUH-stuh puhd-ahng-GOOSH-tuh UHR-duh go-MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Extended Hand to Big Toe Half Cow Face Forward Bend in Infinity Pose, Extended Hand to Big Toe Half Cow Face Western Intense Stretch in Sleeping Vishnu Pose

Modification: forearm to the floor; top leg crossed over; grabbing onto the big toe of the bottom foot, leg off the floor

Pose Type: supine (on the side), forward bend

Drishti Point: Hastagrai or Hastagre (hands)

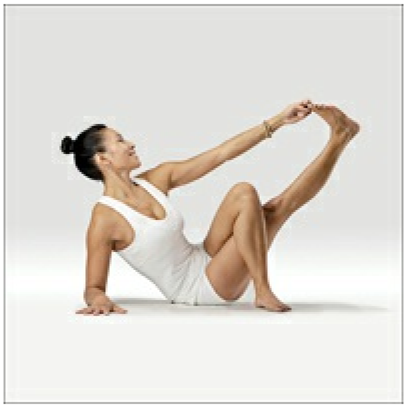

Half Lotus Pose in Infinity Pose

Ardha Padmasana in Anantasana

(UHR-duh puhd-MAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Half Lotus Pose in Sleeping Vishnu Pose

Modification: top foot to the bottom knee; top forearm to the top knee

Pose Type: supine (on the side), mild backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

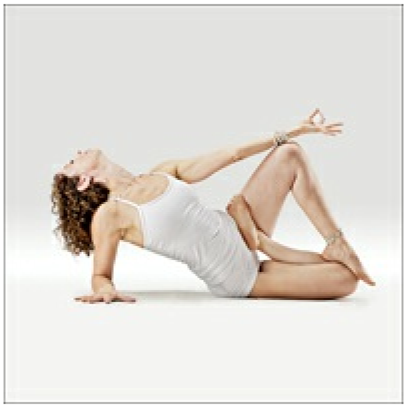

Half Lotus Extended Hand to Leg Pose in Infinity Pose

Ardha Padma Utthita Pada Hastasana in Anantasana

(UHR-duh PUHD-muh UT-ti-tuh PUH-duh huh-STAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Half Lotus Extended Hand to Leg Pose in Sleeping Vishnu Pose
Pose Type: supine (on the side), forward bend, mild backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

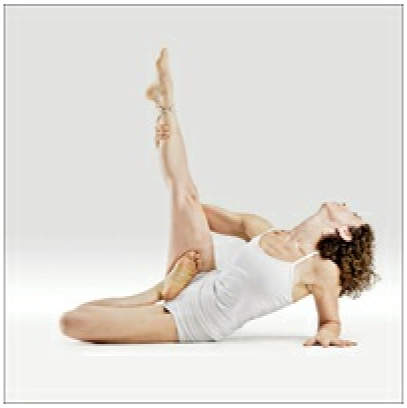

FOREARM ON THE FLOOR: TOP LEG STRAIGHT, BOTTOM FOOT TO THE THIGH

Half Bound Angle Pose in Infinity Pose

Ardha Baddha Konasana in Anantasana

(UHR-duh BUH-duh ko-NAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Also Known As: Half Bound Angle Pose in Sleeping Vishnu Pose
Modification: forearm to the floor

1. finger tips of the top arm to the floor in front of the abdomen, elbow bent
2. top arm straight and parallel to the top leg

Pose Type: supine (on the side), side bend

Drishti Point: Padayoragrai or Padayoragre (toes/feet), Hastagrai or Hastagre (hands)

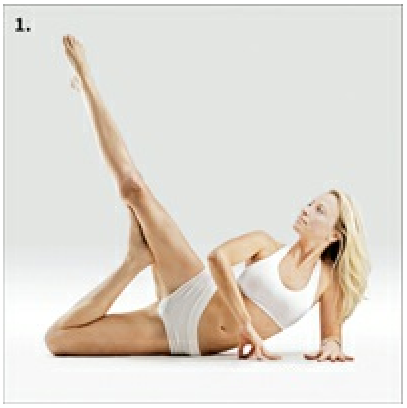

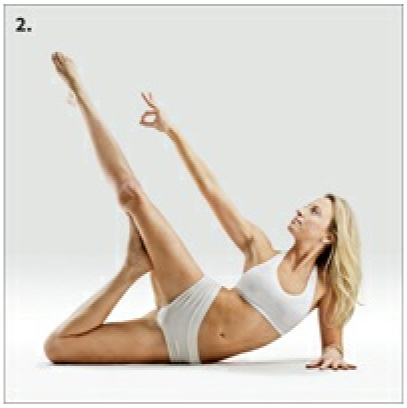

EASTERN INTENSE STRETCH: SITTING BONES BELOW KNEES—ONE LEG OFF FLOOR & TWISTS

Tip Toe Half Eastern Intense Stretch Pose

Prapada Ardha Purvottanasana

(PRUH-puh-duh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Tip Toe Half Reverse Plank Pose

Modification: prep—sitting bones below the knee level, on the fingertips

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

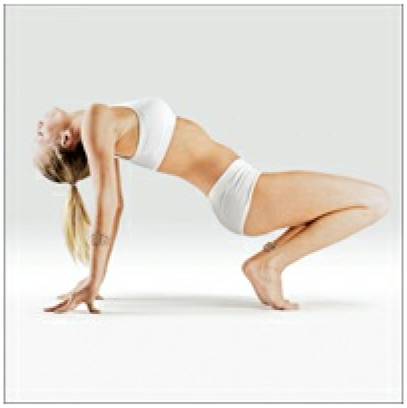

Half Eastern Intense Stretch Pose

Ardha Purvottanasana

(UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Half Reverse Plank

Modification: mild prep—sitting bones slightly off the floor, feet flat on the floor; palms flat on the floor, fingertips pointing to the heels

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

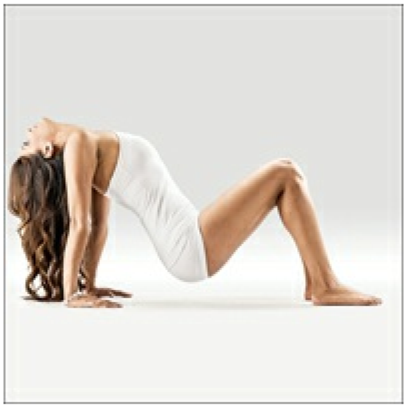

Half Fire Log Pose in Half Eastern Intense Stretch Pose

Ardha Agnistambhasana in Ardha Purvottanasana

(UHR-duh uhg-ni-stuhm-BAHS-uh-nuh in UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Half Firelog Pose in Half Reverse Plank

Modification: prep—sitting bones slightly off the floor; elbows bent, fingertips pointing to the heels

Pose Type: standing, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

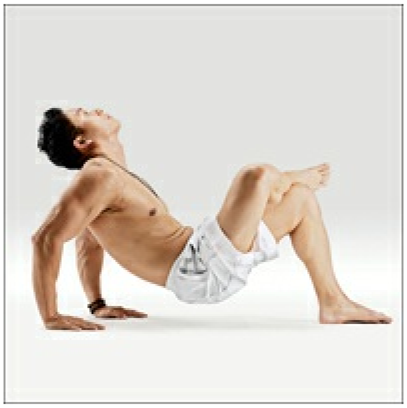

Intense Ankle Stretch Pose in One-Legged Half Eastern Intense Stretch Pose

Uttana Kulpasana in Eka Pada Ardha Purvottanasana

(ut-TAHN-uh kul-PAHS-uh-nuh in EY-kuh PUH-duh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Intense Ankle Stretch Pose in One-Legged Half Reverse Plank
Modification: knee bent toward the chest, fingertips pointing away from the heels
Pose Type: standing, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

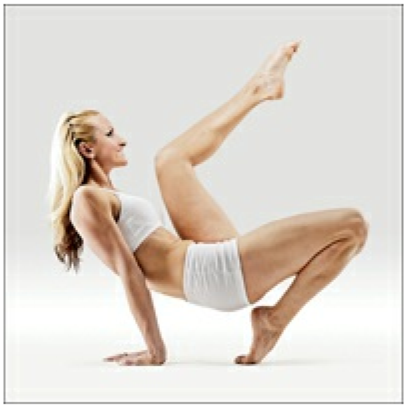

Revolved Half Eastern Intense Stretch Pose

Parivritta Ardha Purvottanasana

(puh-ri-VRIT-tuh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Revolved Half Reverse Plank Pose
Modification: opposite elbow to knee of the straight leg
Pose Type: arm balance, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

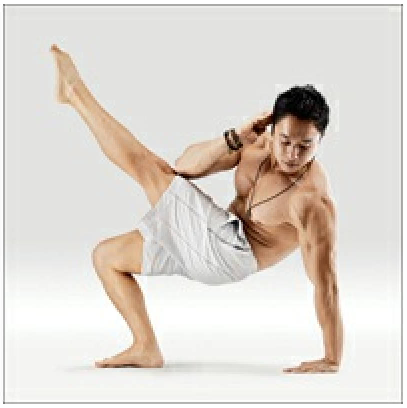

EASTERN INTENSE STRETCH: KNEES BENT

Half Eastern Intense Stretch Pose

Ardha Purvottanasana

(UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Reverse Table Top and Half Reverse Plank Pose
Modification: prep—knees bent and on top of the ankles, fingertips pointing away from the heels
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

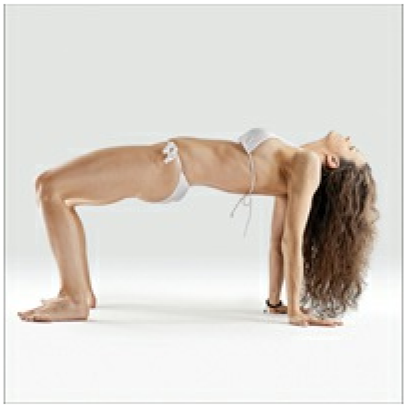

Uneven Leg Position of Cow Face Pose in Half Eastern Intense Stretch Pose

Vishama Pada Gomukhasana in Ardha Purvottanasana

(VISH-uh-muh PUH-duh go-muk-AHS-uh-nuh in UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Uneven Leg Position of Cow Face Pose in Reverse Table Top and Uneven Leg Position of Cow Face Pose in Half Reverse Plank
Modification: knees bent, fingertips pointing to the heels

Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

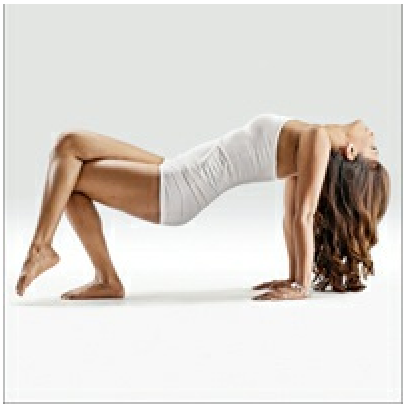

One-Legged Half Eastern Intense Stretch Pose

Eka Pada Ardha Purvottanasana

(EY-kuh PUH-duh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: One-Legged Reverse Table Top and One-Legged Half Reverse Plank Pose
Modification: prep—one knee bent, foot on the floor; other leg straight; knees in line, fingertips pointing to the heels
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

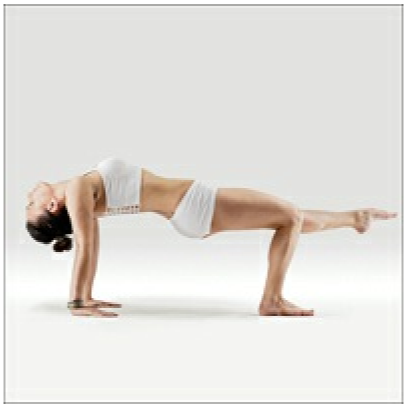

One-Legged Half Eastern Intense Stretch Pose

Eka Pada Ardha Purvottanasana

(EY-kuh PUH-duh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: One-Legged Reverse Table Top and One-Legged Half Reverse Plank
Modification: prep—both knees bent, top foot to the knee, fingertips pointing to the heels

Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

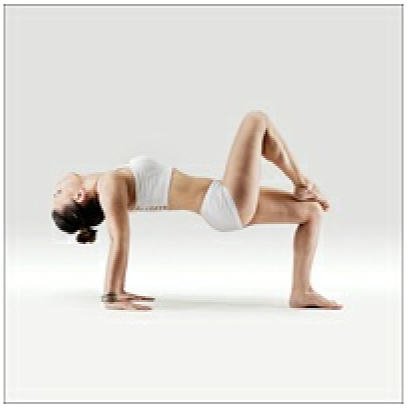

One-Legged Half Eastern Intense Stretch Pose in Pose Dedicated to Sage Vasistha on the Forearm

Eka Pada Ardha Purvottanasana in Vasishtasana

(EY-kuh PUH-duh UHR-duh poor-vo-tahn-AHS-uh-nuh in vuh-sish-TAHS-uh-nuh)

Also Known As: One-Legged Reverse Table Top in Pose Dedicated to Sage Vasistha on the Forearm, One-Legged Half Reverse Plank Top in Pose Dedicated to Sage Vasistha on the Forearm
Modification: both knees bent, foot to the knee
Pose Type: standing (on the forearm and the foot)
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ॐ

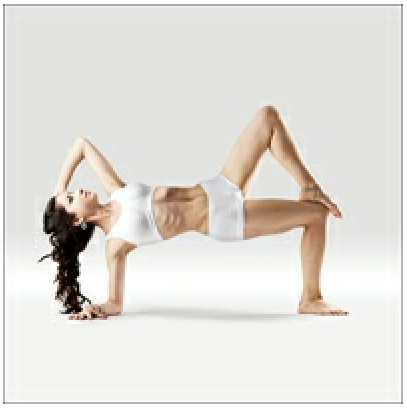

One Hand Half Eastern Intense Stretch Pose

Eka Hasta Ardha Purvottanasana

(EY-kuh HUH-stuh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: One Hand Reverse Table Top or One Hand Half Reverse Plank

Modification: knees bent, fingertips pointing away from the heels
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल य ॐ

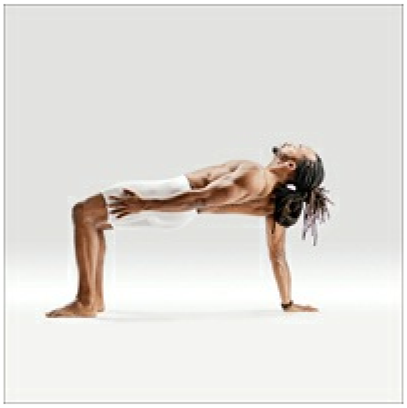

EASTERN INTENSE STRETCH: LEGS STRAIGHT

Eastern Intense Stretch Pose

Purvottanasana

(poor-vo-tahn-AHS-uh-nuh)
Also Known As: Reverse Plank and Upward Plank Pose
Modification: legs straight, fingertips pointing to the heels
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

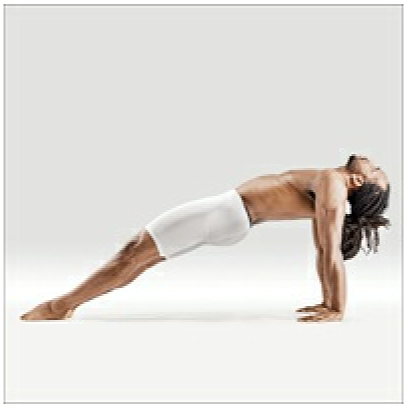

Foot to the Side

Half Eastern Intense Stretch Pose

Parshva Eka Pada Ardha Purvottanasana

(PAHRSH-vuh EY-kuh PUH-duh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Foot to the Side Reverse Table Top, Foot to the Side Half Reverse Plank and Foot to the Side Half Upward Plank Pose

Pose Type: standing, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

How to Perform the Pose:

1. Begin by sitting on the floor with both your legs straight out in front of you. Keep your hands on the floor on the sides of your hips with fingertips pointing toward your feet. Engage your *mula bandha*, *uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale, bend your knees, and slide your feet toward your sitting bones. Keep your feet and knees hip-width apart.
3. On the next exhale, press into your hands and feet and lift your sitting bones off the floor. Your shoulders should be on top of your wrists, your knees on top of your ankles, and your torso and thighs parallel to the floor.
4. Exhale, straighten your right leg, and bring it out to the side, flexing the toes of your right foot out to the side. Hold the pose for at least 30, and up to 90, seconds in order to receive full benefits.
5. Inhale and bring your right leg back to the center in line with your left leg. Repeat on the other side. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch.
6. Inhale and lower your hips to the floor, coming back to the starting position.

Modification: one knee bent, other leg straight out to the side, fingertips pointing to the heels

parshva = side

eka = one

pada = foot or leg

ardha = half

purva = east, the front of the body

ut = intense

tan = to stretch, to extend

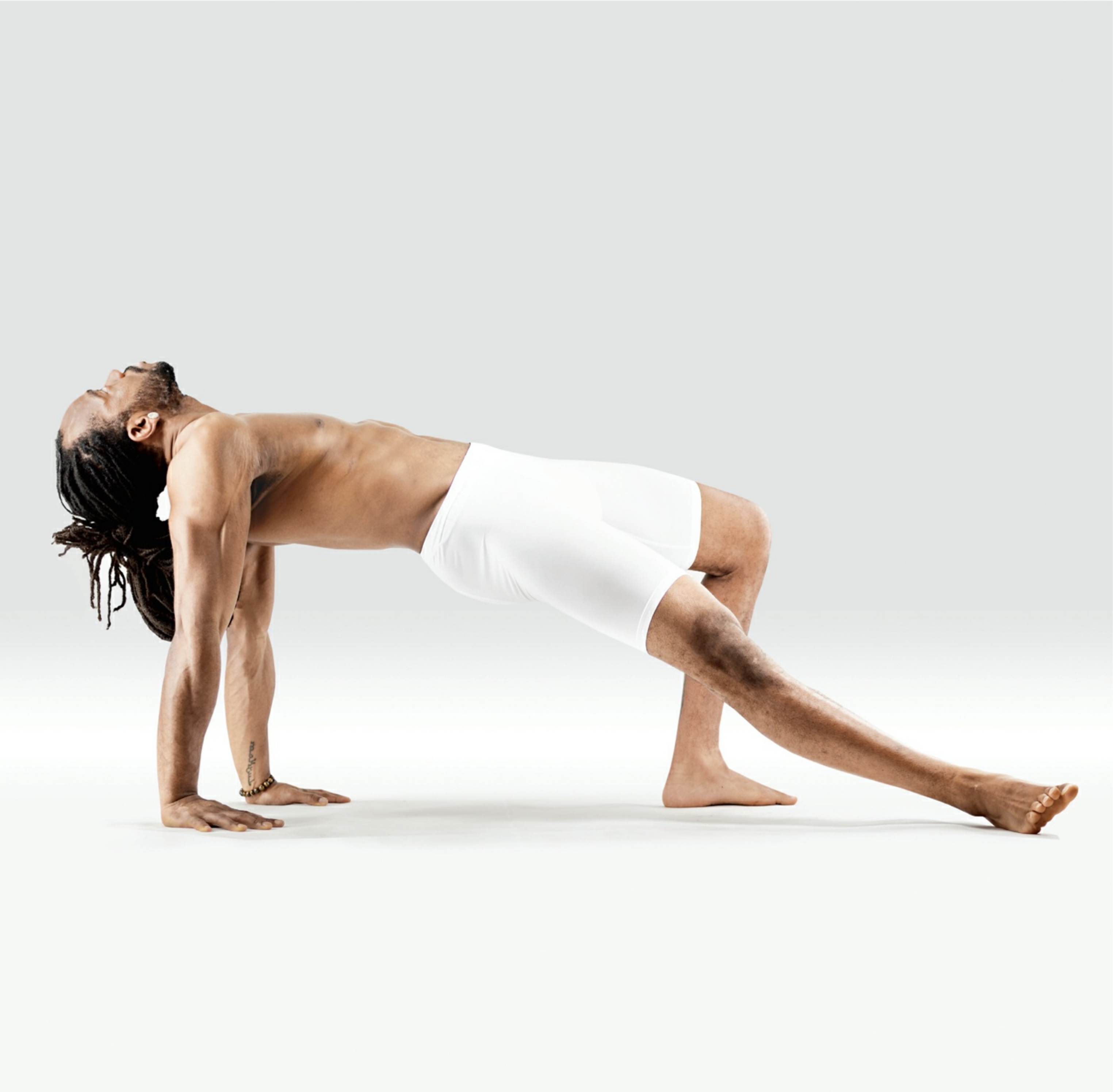

Feet Wide Eastern Intense Stretch Pose

Prasarita Pada Purvottanasana

(pruh-SAH-ri-tuh PUH-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Feet Wide Reverse Plank and Feet Wide Upward Plank Pose
Modification: legs straight, toes flexed in, fingertips pointing to the heels
1. side view
2. front view
Pose Type: standing, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

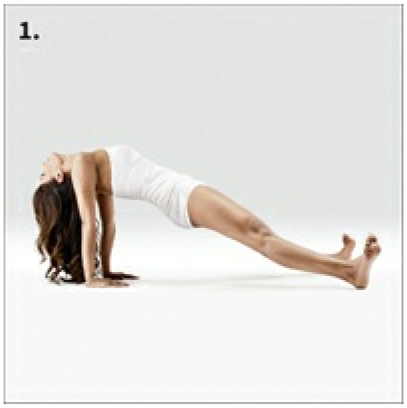

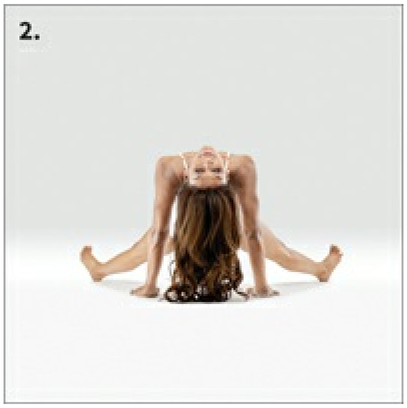

EASTERN INTENSE STRETCH: FOREARMS ON THE FLOOR

Easy Fish Pose

Sukha Matsyasana

(SUK-uh muhts-YAHS-uh-nuh)

Also Known As: Eastern Intense Stretch Pose Prep. (Purvottanasana Prep.), Reverse Plank Prep.
Modification: head off the floor, forearms on the floor, legs straight
Pose Type: seated, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

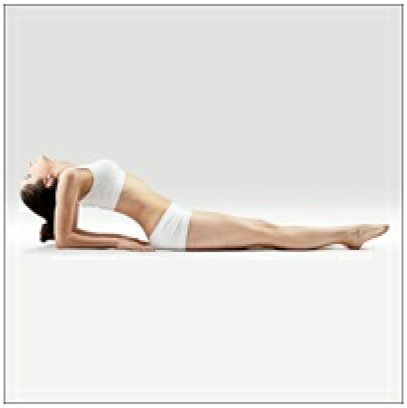

Eastern Intense Stretch Pose

Purvottanasana

(poor-vo-tahn-AHS-uh-nuh)

Also Known As: Reverse Plank on the Forearms and Upward Plank Pose on the Forearms
Modification: legs straight, forearms on the floor
Pose Type: standing (on the forearms), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

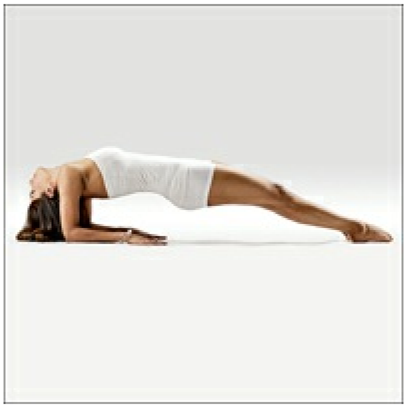

Easy Fish Pose

Sukha Matsyasana

(SUK-uh muhts-YAHS-uh-nuh)

Also Known As: Eastern Intense Stretch Pose Prep. (Purvottanasana Prep.), Reverse Plank Prep.
Modification: head off the floor, forearms to the floor; one leg straight, other knee bent, foot flat on the floor
1. straight leg on the floor

2. straight leg off the floor, toes flexed in

Pose Type: seated, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

1.

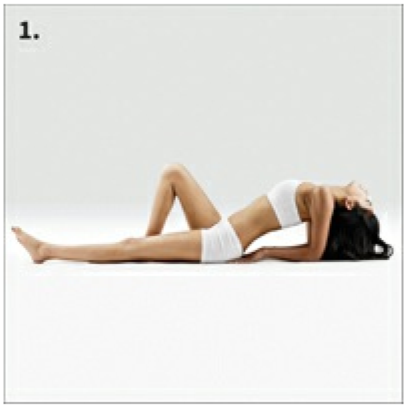

2.

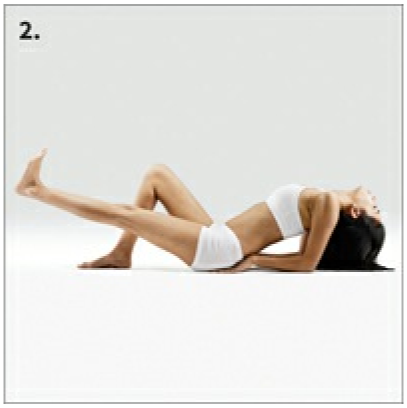

EASTERN INTENSE STRETCH: FOREARMS TO FLOOR—ONE LEG OFF FLOOR

Half Fire Log Pose in Easy Fish Pose

Ardha Agnistambhasana in Sukha Matsyasana

(UHR-duh uhg-ni-stuhm-BAHS-uh-nuh in SUK-uh muhts-YAHS-uh-nuh)

Also Known As: Eastern Intense Stretch Pose Prep. (Purvottanasana Prep.), Reverse Plank Prep.

Modification: forearms to the floor, head off the floor

Pose Type: seated, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व य ह ॐ

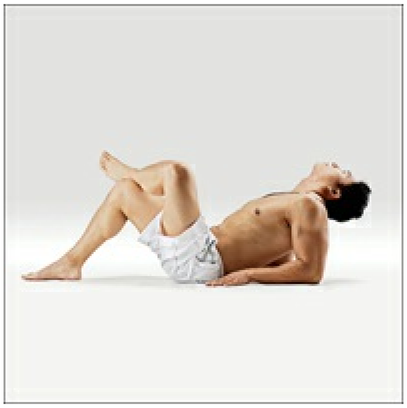

Easy Fish Pose

Sukha Matsyasana

(SUK-uh muhts-YAHS-uh-nuh)

Also Known As: Eastern Intense Stretch Pose Prep. (Purvottanasana Prep.), Reverse Plank Prep.
Modification: forearms to the floor, head off the floor, sole of the foot to the opposite knee
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल र य ह ॐ

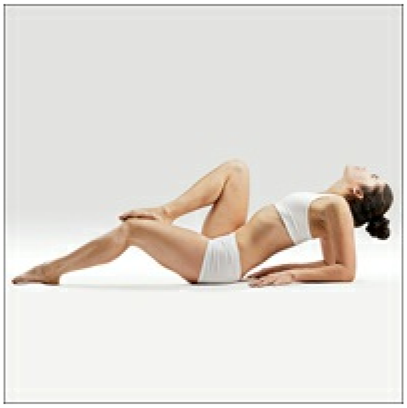

Revolved Hand to Foot Pose in Half Eastern Intense Stretch Pose on the Forearm

Parivritta Hasta Padasana in Ardha Purvottanasana

(puh-ri-VRIT-tuh HUH-stuh puh-DAHS-uh-nuh in UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Revolved Hand to Foot Pose in Reverse Table Top, Revolved Hand to Foot Pose in Half Reverse Plank
Modification: forearm on the floor; grabbing onto the outside edge of the foot, leg straight
Pose Type: standing (on the forearm and the foot), twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व य ह ॐ

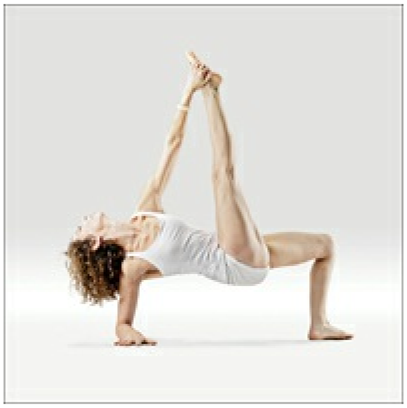

EASTERN INTENSE STRETCH: FOREARMS TO FLOOR—GRABBING ONTO TRICEPS—KNEES BENT

Hands Bound Easy Fish Pose

Baddha Hasta Sukha Matsyasana

(BUH-duh HUH-stuh SUK-uh muhts-YAHS-uh-nuh)

Modification: knees bent, feet to the floor
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

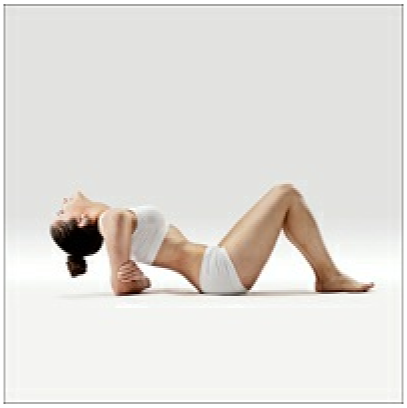

Hands Bound Half Eastern Intense Stretch Pose

Baddha Hasta Ardha Purvottanasana

(BUH-duh HUH-stuh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Hands Bound Reverse Table Top, Hands Bound Half Reverse Plank
Pose Type: standing (on the forearms), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

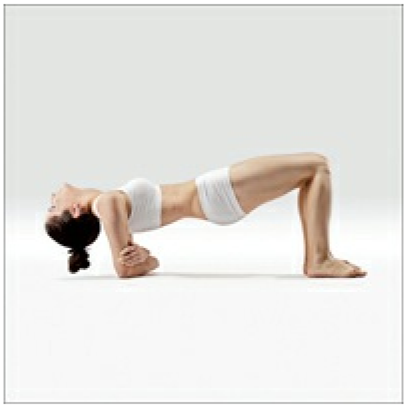

Reverse Prayer One-Legged Easy Fish Pose

Viparita Namaskar Eka Pada Sukha Matsyasana

(vi-puh-REE-tuh nuh-muhs-KAHR EY-kuh PUH-duh SUK-uh muhts-YAHS-uh-nuh)

Also Known As: Back of the Body Prayer One-Legged Easy Fish Pose (Paschima Namaskara Eka Pada Sukha Matsyasana)
Modification: both knees bent, heel to the knee
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

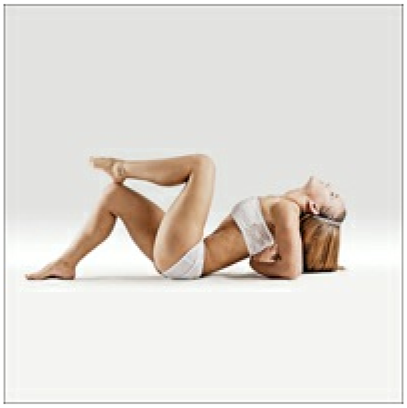

Hands Bound One-Legged Half Eastern Intense Stretch Pose

Baddha Hasta Eka Pada Ardha Purvottanasana

(BUH-duh HUH-stuh EY-kuh PUH-duh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Hands Bound One-Legged Reverse Table Top, Hands Bound One-Legged Half Reverse Plank
Modification: heel to the knee; grabbing onto the triceps
Pose Type: standing (on the forearms), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

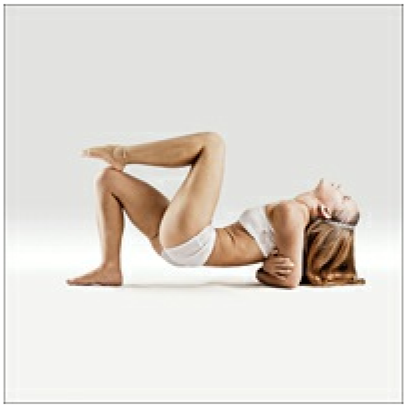

EASTERN INTENSE STRETCH: FOREARMS TO FLOOR—GRABBING ONTO THE TRICEPS

Hands Bound Easy Fish Pose

Baddha Hasta Sukha Matsyasana

(BUH-duh HUH-stuh SUK-uh muhts-YAHS-uh-nuh)

Modification: one leg straight to the floor; other knee bent, foot to the floor
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

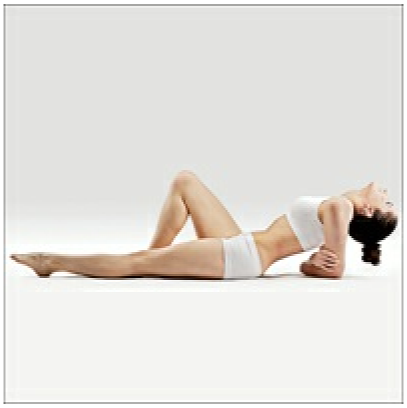

Hands Bound Easy Intense Leg Stretch

Baddha Hasta Sukha Uttana Padasana

(BUH-duh HUH-stuh SUK-uh ut-TAHN-uh puh-DAHS-uh-nuh)

Modification: knees bent at 90 degrees, feet off the floor
Pose Type: seated, backbend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

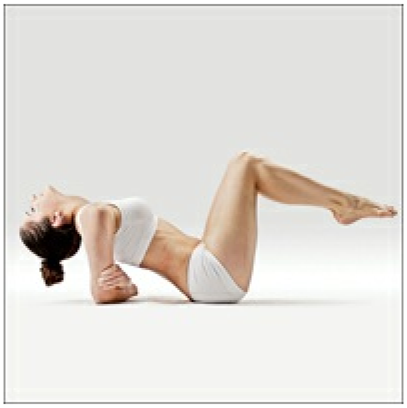

EASTERN INTENSE STRETCH: ELBOWS ON THE FLOOR—HALF LOTUS

Half Lotus Easy Fish Pose

Ardha Padma Sukha Matsyasana

(UHR-duh PUHD-muh SUK-uh muhts-YAHS-uh-nuh)

Also Known As: Eastern Intense Stretch Pose Prep. (Purvottanasana Prep.), Reverse Plank Prep.
Modification: elbows to the floor, palms to the lower back; one leg in half lotus, other knee bent; heel up
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

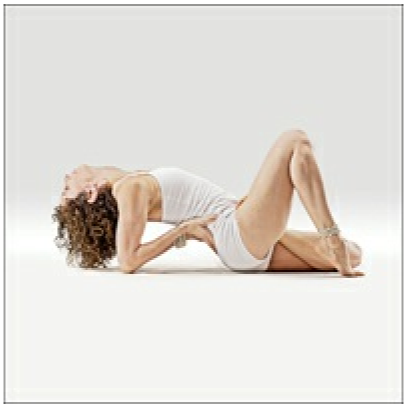

Half Lotus Half Eastern Intense Stretch Pose

Ardha Padma Ardha Purvottanasana

(UHR-duh PUHD-muh UHR-duh poor-vo-tahn-AHS-uh-nuh)

Also Known As: Half Lotus Reverse Table Top, Half Lotus Half Reverse Plank
Modification: elbows to the floor, palms to the lower back; one leg in half lotus, other knee bent; heel up
1. right side view

2. left side view

Pose Type: standing (on the elbows, knee and foot), backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ह ॐ

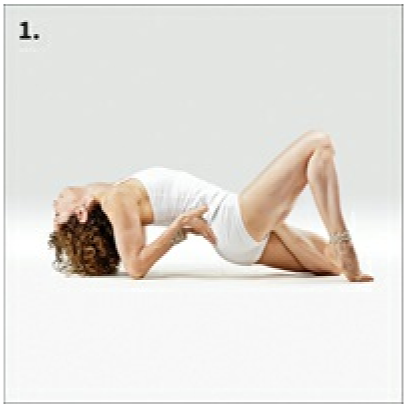

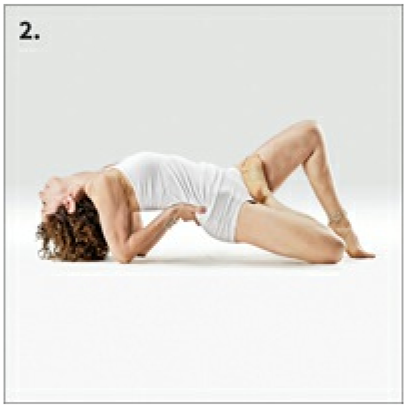

BRIDGE: HEELS DOWN, SPINE ON THE FLOOR

Bridge Whole Body Pose Prep.

Setu Bandha Sarvangasana Prep.

(SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: spine flat on the floor, palms down by the hips, heels down

Pose Type: supine

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल

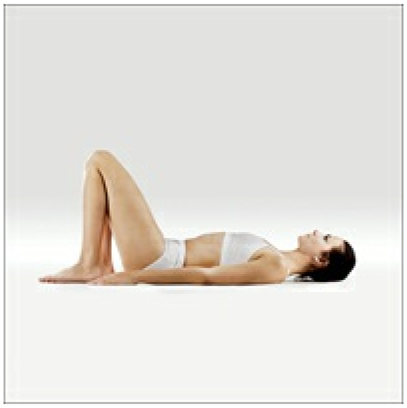

Reclined Hands Bound Twisted Stomach Pose in Bridge Whole Body Pose Prep.

Supta Baddha Hasta Jatara Parivartanasana in Setu Bandha Sarvangasana Prep.

(SUP-tuh BUH-duh HUH-stuh JAH-tuh-ruh puh-ri-vuhrt-uhn-AHS-uh-nuh in SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: ear resting on a yoga block

Pose Type: supine, twist

Drishti Point: Parshva Drishti (to the right), Parshva Drishti (to the left)

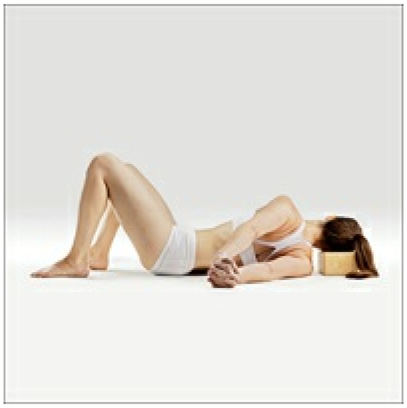

Reclined Hands Bound One Leg Extended Pose in Bridge Whole Body Pose Prep.

Supta Baddha Hasta Utthita Eka Padasana in Setu Bandha Sarvangasana Prep.

(SUP-tuh BUH-duh HUH-stuh UT-ti-tuh EY-kuh puh-DAHS-uh-nuh in SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: arms straight in front, palms facing up

Pose Type: supine

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

Reclined Half Bound Lotus Pose in Bridge Whole Body Pose Prep.

Supta Ardha Baddha Padmasana in Setu Bandha Sarvangasana Prep.

(SUP-tuh UHR-duh BUH-duh puhd-MAHS-uh-nuh in SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: grabbing onto the ankle
Pose Type: supine, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

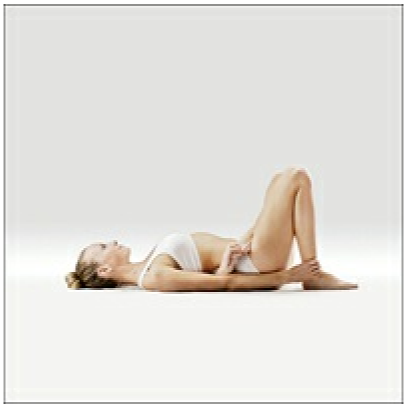

BRIDGE: HEELS DOWN, ARMS BEHIND THE BACK

Bridge Whole Body Pose

Setu Bandha Sarvangasana

(SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms to the floor, arms straight; feet in front of the knees
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

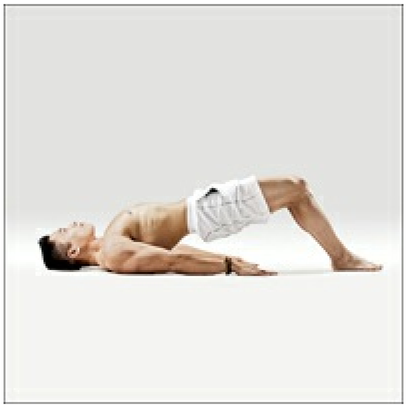

Uneven Bridge Whole Body Pose

Vishama Setu Bandha Sarvangasana

(VISH-uh-muh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Desk Pose Modification (Dwipadapitam)
Modification: palms down by the heel of the bent leg, other leg straight
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ॐ

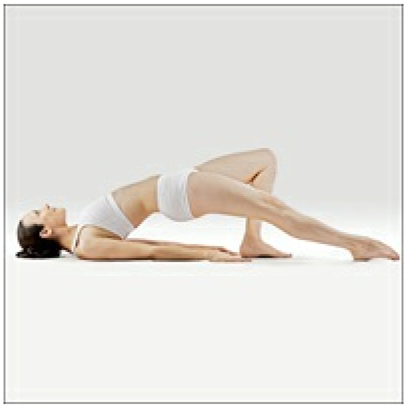

Bridge Whole Body Pose

Setu Bandha Sarvangasana

(SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Shoulder Pose (Kandharasana)
Modification: palms down by the heels
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह

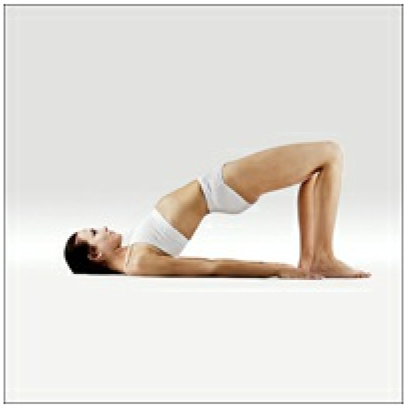

Bridge Whole Body Pose

Setu Bandha Sarvangasana

(SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms to the lower back
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

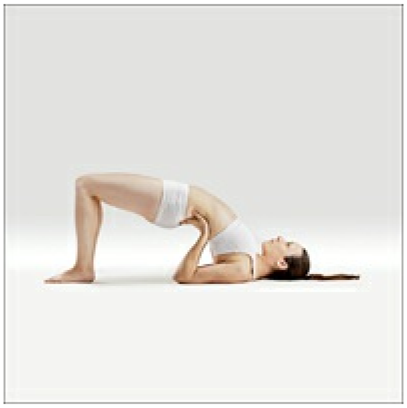

BRIDGE: HEELS DOWN, ARMS OVERHEAD & BEHIND THE BACK

Bridge Whole Body Pose

Setu Bandha Sarvangasana

(SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Desk Pose Modification (Dwipadapitam)
Modification: arms straight to the floor over the head, palms facing up
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

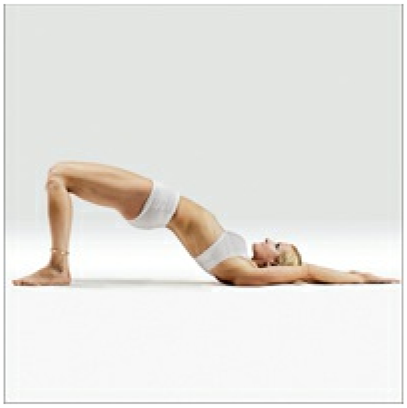

Bridge Whole Body Pose

Setu Bandha Sarvangasana

(SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Upward Bow Pose Prep. (Urdhva Dhanurasana Prep.)
Modification: palms under the shoulders, fingertips pointing to the heels
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

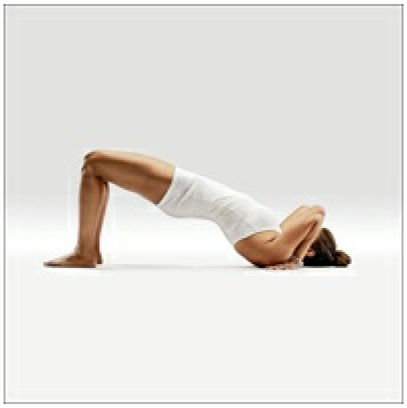

Upward Hands Bound Bridge Whole Body Pose

Urdhva Baddha Hasta Setu Bandha Sarvangasana

(OORD-vuh BUH-duh HUH-stuh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Desk Pose Modification (Dwipadapitam)
Pose Type: supine, backbend
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)
ल र य ह

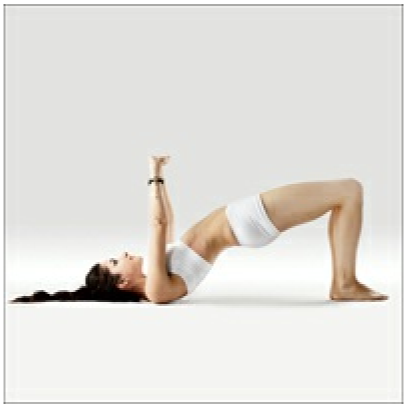

Hands Bound Bridge Whole Body Pose

Baddha Hasta Setu Bandha Sarvangasana

(BUH-duh HUH-stuh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

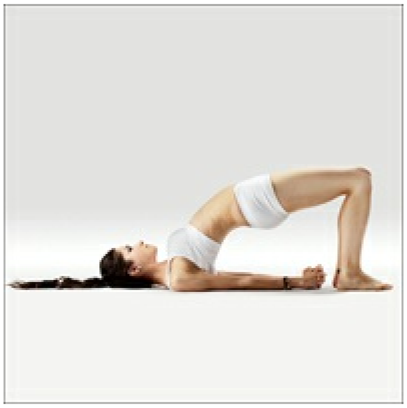

BRIDGE: HEELS DOWN, GRABBING ONTO THE ANKLES

Hand to Ankle Bridge Whole Body Pose

Hasta Kulpa Setu Bandha Sarvangasana

(HUH-stuh KUL-puh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Desk Pose Modification (Dwipadapitam)

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

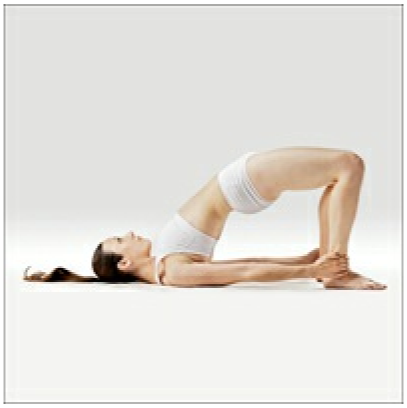

Hand to Ankle Bridge Whole Body Pose

Hasta Kulpa Setu Bandha Sarvangasana

(HUH-stuh KUL-puh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Desk Pose Modification (Dwipadapitam)
Modification: arms crossed
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं रं यं हं ॐ

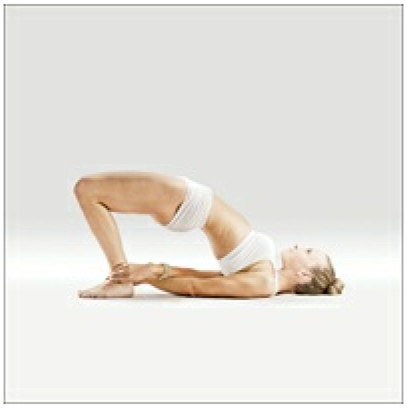

BRIDGE: HEELS DOWN—ONE FOOT OFF THE FLOOR

One Hand to Ankle Bridge Whole Body Pose

Eka Hasta Kulpa Setu Bandha Sarvangasana

(EY-kuh HUH-stuh KUL-puh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: yoga block under the sitting bones; grabbing onto the ankle on the same side, top of the foot to the floor
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

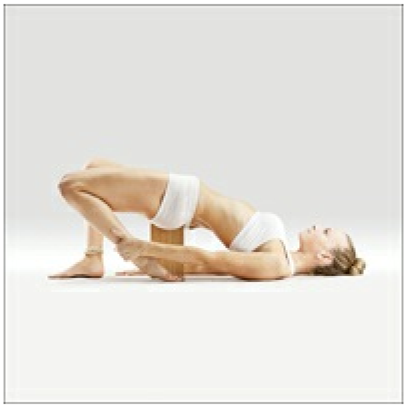

One Hand to Ankle Bridge Whole Body Pose

Eka Hasta Kulpa Setu Bandha Sarvangasana

(EY-kuh HUH-stuh KUL-puh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: Arm 1: grabbing onto the ankle of the opposite leg Arm 2: grabbing onto the opposite arm

Pose Type: supine, backbend, binding

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

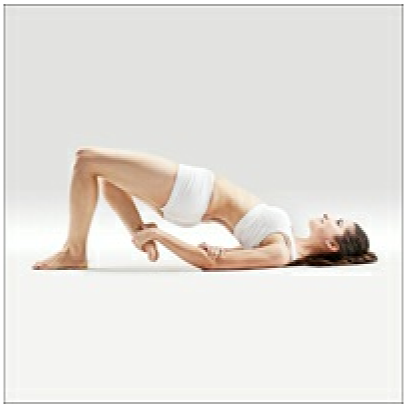

Half Fire Log Pose in One-Legged Hands Bound Bridge Whole Body Pose

Ardha Agnistambhasana in Eka Pada Baddha Hasta Setu Bandha Sarvangasana

(UHR-duh uhg-ni-stuhm-BAHS-uh-nuh in EY-kuh PUH-duh BUH-duh HUH-stuh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: ankle on top of the knee heel down, toes of the top foot pointed

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

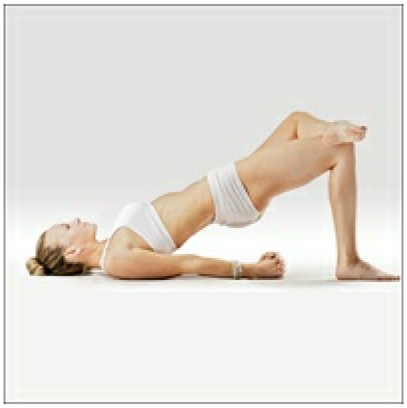

Half Bound Lotus One Hand to Ankle Bridge Whole Body Pose

Ardha Baddha Padma Eka Hasta Kulpa Setu Bandha Sarvangasana

(UHR-duh BUH-duh PUHD-muh EY-kuh HUH-stuh KUL-puh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Desk Pose Modification (Dwipadapitam)
Modification: heel down
Pose Type: supine, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

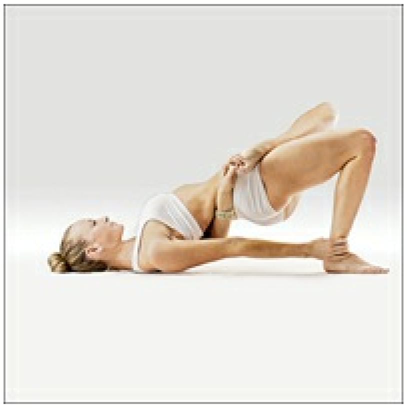

BRIDGE: HEELS DOWN—ONE LEG TO THE SKY

One-Legged Bridge Whole Body Pose

Eka Pada Setu Bandha Sarvangasana

(EY-kuh PUH-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms down, arms straight on the floor; both knees bent
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र य ॐ

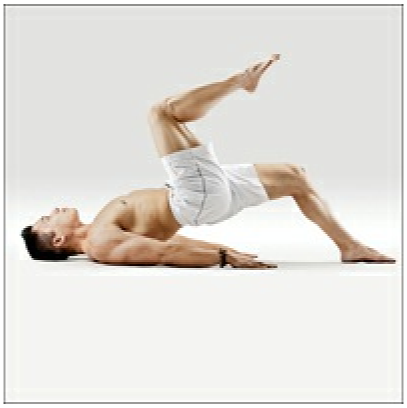

One-Legged Bridge Whole Body Pose

Eka Pada Setu Bandha Sarvangasana

(EY-kuh PUH-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms down, arms straight on the floor; one leg up and straight, knees toward each other
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

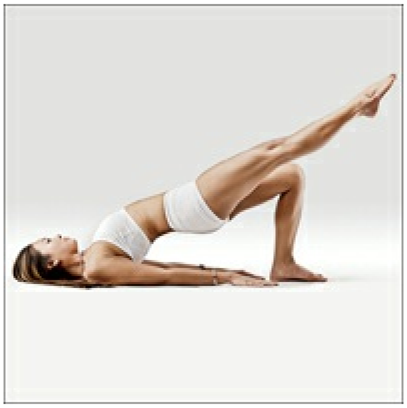

One-Legged Bridge Whole Body Pose

Eka Pada Setu Bandha Sarvangasana

(EY-kuh PUH-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Desk Pose Modification (Dwipadapitam)
Modification: palms down by the heel
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

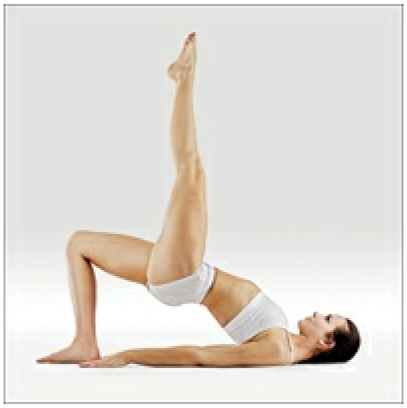

One-Legged Bridge Whole Body Pose

Eka Pada Setu Bandha Sarvangasana

(EY-kuh PUH-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Also Known As: Peacock Stretch Pose (Uttana Mayurasana)
Modification: palms to the lower back
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

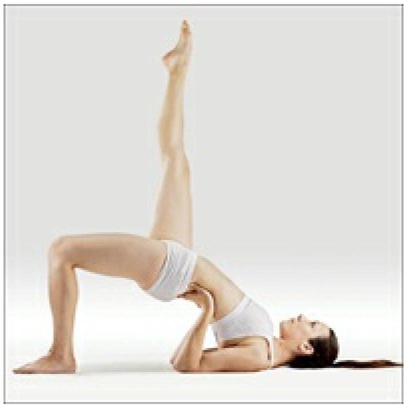

Both Hands to Ankle One-Legged Bridge Whole Body Pose

Dwi Hasta Kulpa Eka Pada Setu Bandha Sarvangasana

(DWI-huh-stuh KUL-puh EY-kuh PUH-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: grabbing onto the ankle with both hands
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

One-Legged Upward Hands Bound Tip Toe

Bridge Whole Body Pose

Eka Pada Urdhva Baddha Hasta Prapada Setu Bandha Sarvangasana

(EY-kuh PUH-duh OORD-vuh BUH-duh HUH-stuh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Pose Type: supine, backbend

Drishti Point: Angusthamadhye or Angustha Ma Dyai(thumbs)

How to Perform the Pose:

1. Begin by lying with your back flat on the floor. Keep your arms by the sides of your torso, hands by the sides of the hips, palms facing down. Engage your *mula bandha, uddhiyana bandha*, and *ujjayi* breathing.
2. Exhale, bend your knees, and slide your feet toward your sitting bones. Keep your feet flat on the floor, parallel to each other and in line with your sitting bones.
3. On the next exhale, push into your arms and feet and lift your hips off the floor until your thighs are parallel to the floor.
4. Inhale, interlock your fingers behind your back, and press your palms together. Exhale as you push through your chest and straighten the arms.
5. Exhale, bring your ankles and feet together. On the next exhale, lift your right leg, pointing your right toes up to the sky (Pose #2).
6. Exhale, bring your left heel up. Inhale, release your arms and bring them out in front of your chest. Exhale as you interlock your fingers, rotate your palms up to the sky and straighten the arms (Pose #1).
7. Hold the pose for at least 30, and up to 90, seconds in order to receive the full benefits of the stretch. Inhale, lower your right leg, and repeat on the opposite side.
8. On the next inhale, lower your spine to the floor and stretch the legs out in front of you, coming back to the starting position.

Modification: arms extended up to the sky, fingers interlocked, palms facing up, heel up

eka = one

pada = foot or leg

urdhva = upward

baddha = bound

hasta = hand

prapada = tip of the feet

setu = bridge, dam or dike

bandha = lock

sarvanga = the whole body

1.

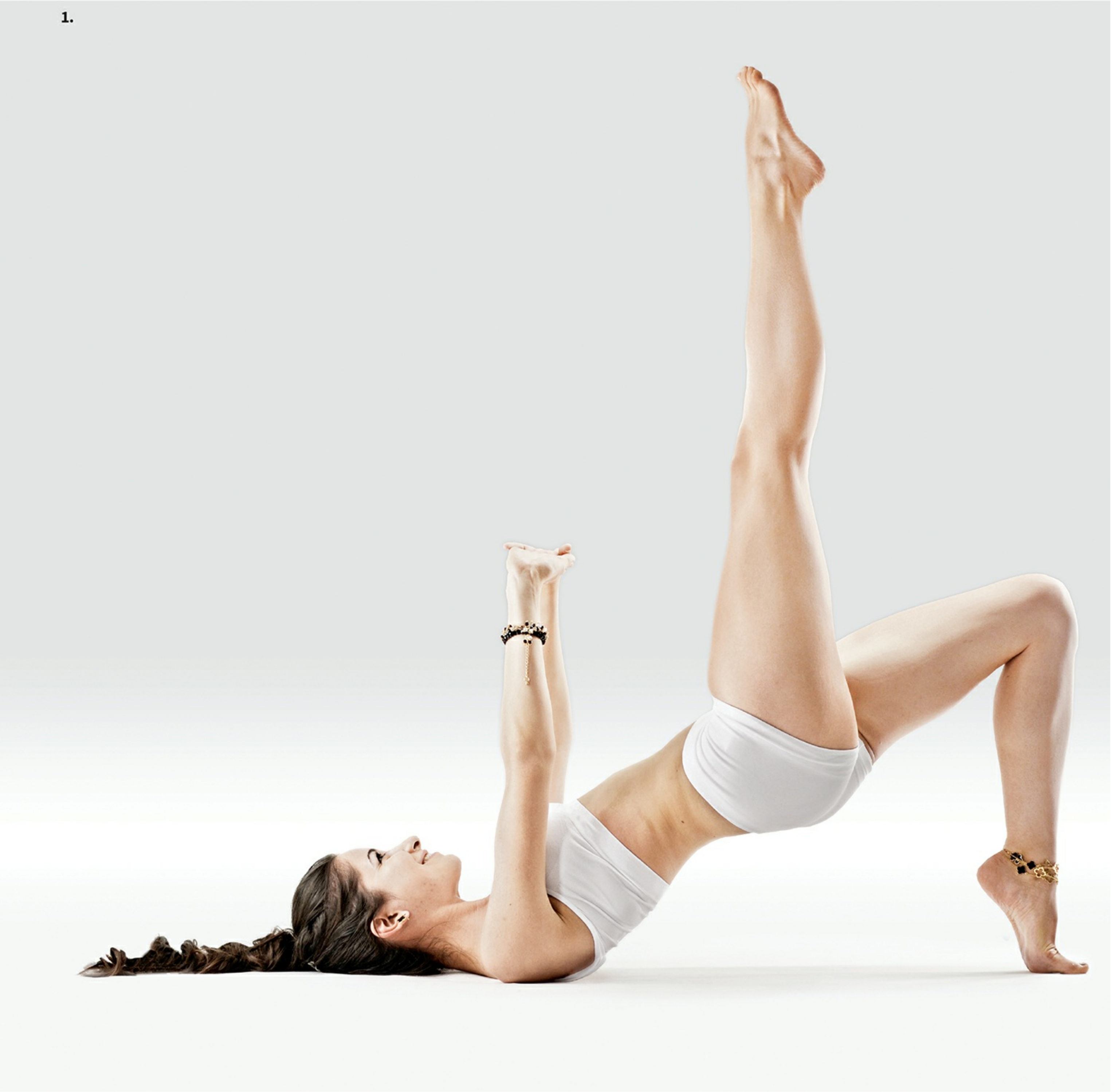

2.

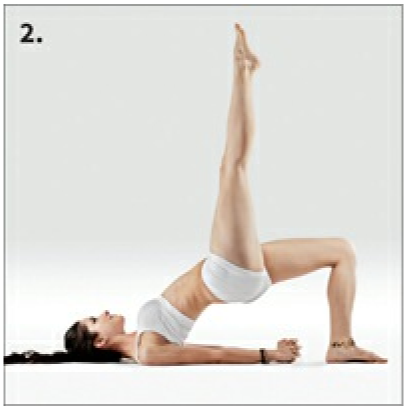

Tip Toe Bridge Whole Body Pose Prep.

Prapada Setu Bandha Sarvangasana Prep.

(PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: spine flat on the floor, palms down by the hips, heels up

Pose Type: supine

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

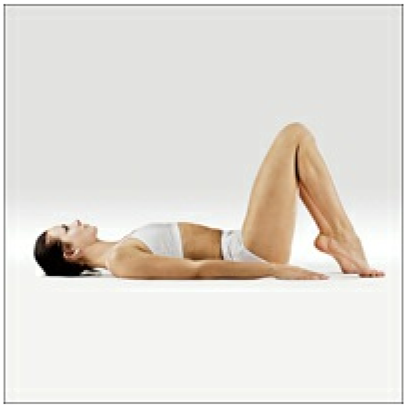

Tip Toe Bridge Whole Body Pose

Prapada Setu Bandha Sarvangasana

(PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms down

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

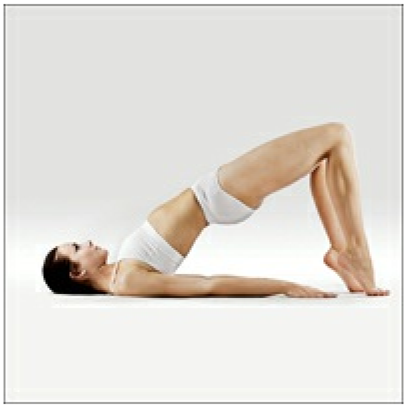

Tip Toe Bridge Whole Body Pose

Prapada Setu Bandha Sarvangasana

(PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms to the lower back

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

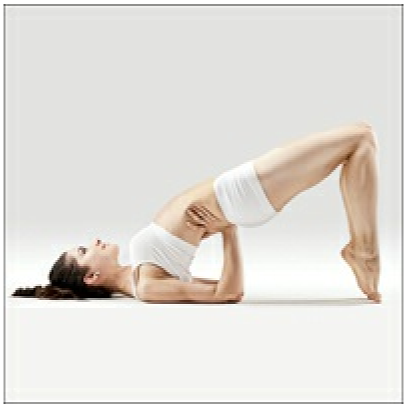

Hands Bound Tip Toe Bridge Whole Body Pose

Baddha Hasta Prapada Setu Bandha Sarvangasana

(BUH-duh HUH-stuh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

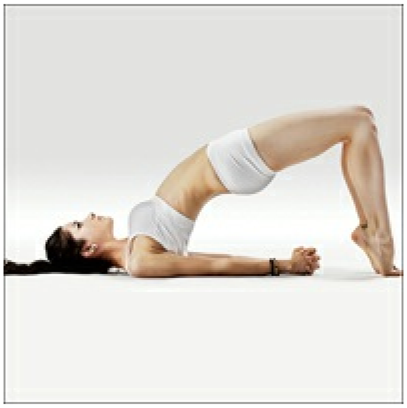

Upward Hands Bound Tip Toe Bridge Whole Body Pose

Urdhva Baddha Hasta Prapada Setu Bandha Sarvangasana

(OORD-vuh BUH-duh HUH-stuh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Pose Type: supine, backbend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

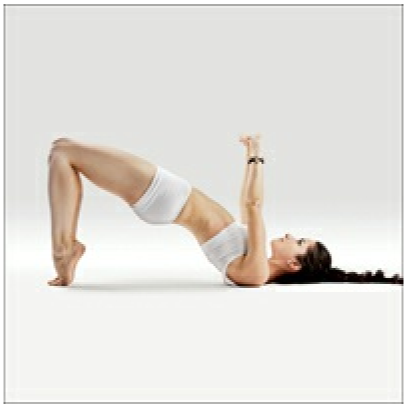

BRIDGE: HEELS UP—ONE LEG CROSSED OVER

Half Fire Log Pose in One-Legged Hands Bound Tip Toe Bridge Whole Body Pose

Ardha Agnistambhasana in Eka Pada Baddha Hasta Prapada Setu Bandha Sarvangasana

(UHR-duh uhg-ni-stuhm-BAHS-uh-nuh in EY-kuh PUH-duh BUH-duh HUH-stuh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: heel up; ankle on top of the knee, toes flexed in
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

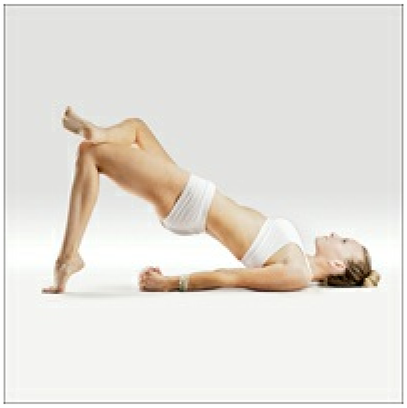

Uneven Leg Position of Cow Face Pose in One-Legged Tip Toe Bridge Whole Body Pose

Vishama Pada Gomukhasana in Eka Pada Prapada Setu Bandha Sarvangasana

(VISH-uh-muh PUH-duh go-muk-AHS-uh-nuh in EY-kuh PUH-duh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: Arm 1: grabbing onto the foot of the opposite leg Arm 2: grabbing onto the opposite arm

Pose Type: supine, backbend, binding

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

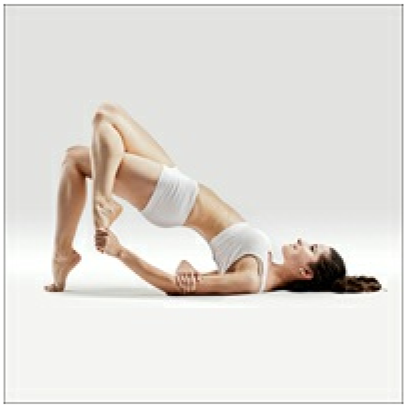

Half Bound Lotus One Hand to Ankle Tip Toe Bridge Whole Body Pose

Ardha Baddha Padma Eka Hasta Kulpa Prapada Setu Bandha Sarvangasana

(UHR-duh BUH-duh PUHD-muh EY-kuh HUH-stuh KUL-puh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: heel up

Pose Type: supine, backbend, binding

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

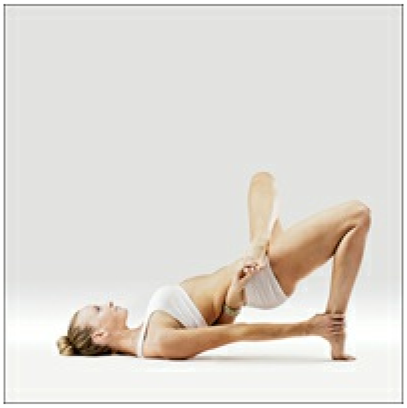

BRIDGE: HEELS UP—ONE LEG EXTENDED TO THE SKY

One-Legged Tip Toe Bridge Whole Body Pose

Eka Pada Prapada Setu Bandha Sarvangasana

(EY-kuh PUH-duh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms down

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

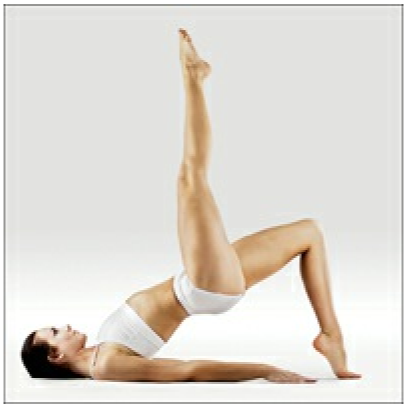

One-Legged Tip Toe Bridge Whole Body Pose

Eka Pada Prapada Setu Bandha Sarvangasana

(EY-kuh PUH-duh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms to the lower back

1. fingertips pointing to the heel
2. fingertips pointing to the head

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

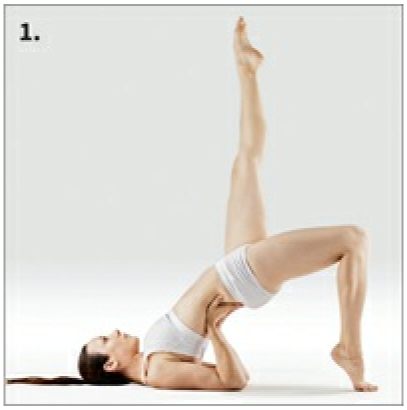

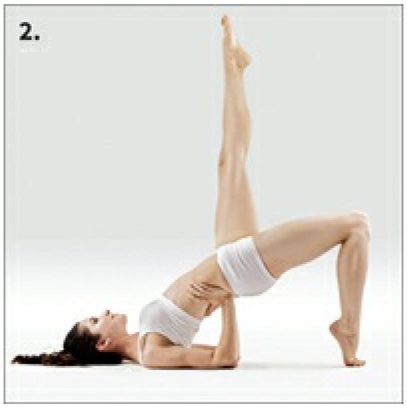

One-Legged Hands Bound Tip Toe Bridge Whole Body Pose

Eka Pada Baddha Hasta Prapada Setu Bandha Sarvangasana

(EY-kuh PUH-duh BUH-duh HUH-stuh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: leg straight and extended to the sky, heel up

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

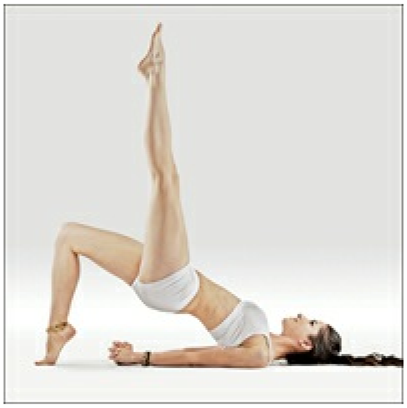

BRIDGE: HEELS UP—ONE LEG EXTENDED TO THE SKY—“BALLET TOES”

Intense Ankle Stretch One-Legged Bridge Whole Body Pose

Uttana Kulpa Eka Pada Setu Bandha Sarvangasana

(ut-TAHN-uh KUL-puh EY-kuh PUH-duh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms to the lower back; both knees bent

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

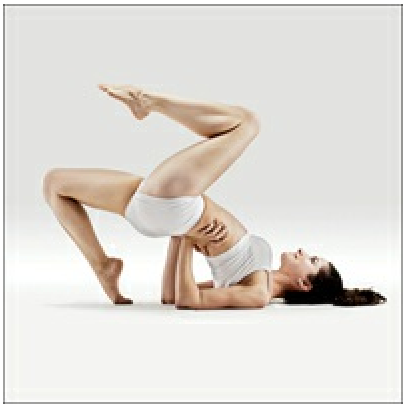

Intense Ankle Stretch One-Legged Bridge Whole Body Pose

Uttana Kulpa Eka Pada Setu Bandha Sarvangasana

(ut-TAHN-uh KUL-puh EY-kuh PUH-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms to the lower back; top leg straight

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

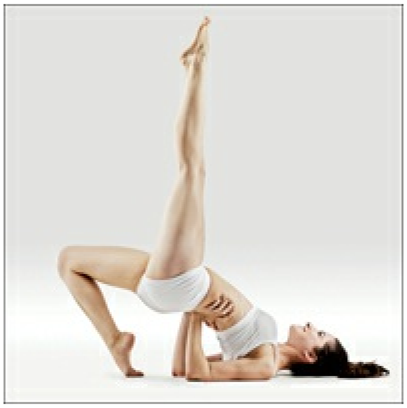

One Hand One-Legged Tip Toe Bridge Whole Body Pose

Eka Hasta Eka Pada Prapada Setu Bandha Sarvangasana

(EY-kuh HUH-stuh EY-kuh PUH-duh PRUH-puh-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: lifted leg bent; palm to the lower back, heel up

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

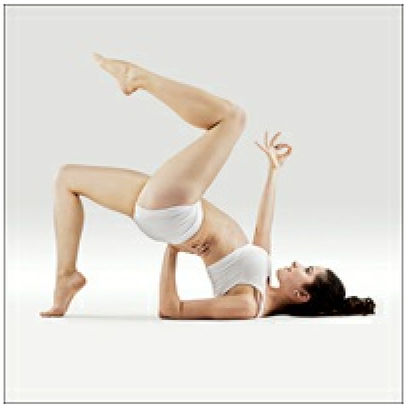

BRIDGE: LEGS STRAIGHT—BACK OF THE HEAD TO THE FLOOR

Extended Hand to Big Toe Pose Intense Ankle Stretch One-Legged Bridge Whole Body Pose

Utthita Hasta Padangushtasana in Uttana Kulpa Eka Pada Setu Bandha Sarvangasana

(UT-ti-tuh HUH-stuh puhd-ahng-goosh-TAHS-uh-nuh in ut-TAHN-uh KUL-puh EY-kuh PUH-duh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Pose Type: supine, backbend, forward bend
Drishti Point: Padayoragrai or Padayoragre (toes/feet)

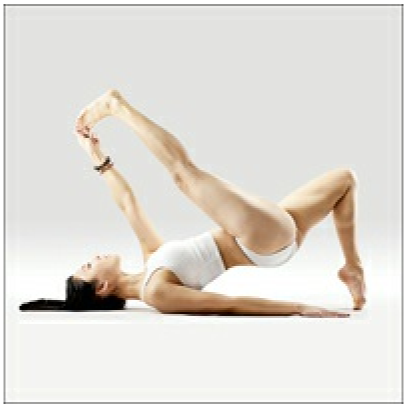

BRIDGE: LEGS STRAIGHT—BACK OF THE HEAD TO THE FLOOR

Complete Bridge Whole Body Pose

Paripurna Setu Bandha Sarvangasana

(puh-ri-POOR-nuh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms to the lower back
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

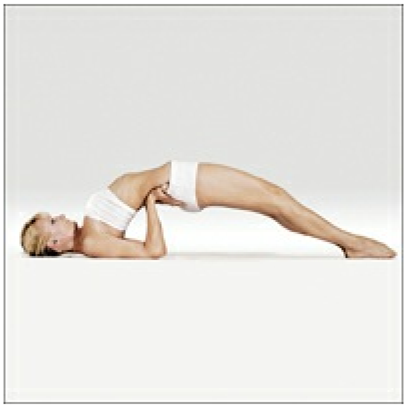

One-Legged Complete Bridge Whole Body Pose

Eka Pada Paripurna Setu Bandha Sarvangasana

(EY-kuh PUH-duh puh-ri-POOR-nuh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)
Modification: palms to the lower back
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

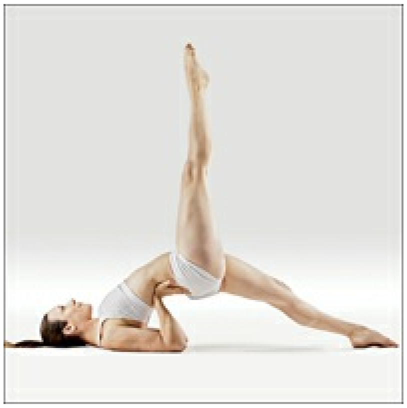

HEAD AND SITTING BONES ON THE FLOOR: KNEES BENT

Half Fish Pose

Ardha Matsyasana

(UHR-duh muhts-YAHS-uh-nuh)
Modification: knees bent; feet flat on the floor, heels toward the sitting bones; palms down to the

floor by the heels, elbows to the floor
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

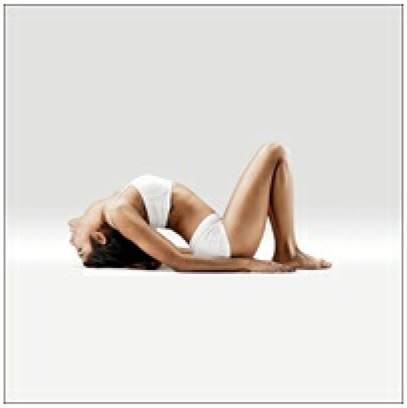

Half Fish Pose

Ardha Matsyasana

(UHR-duh muhts-YAHS-uh-nuh)
Modification: knees bent
1. arms over the head, feet flat on the floor
2. hands by the navel, palms pressed together, fingertips pointing up, both feet in the tip toe position
3. hands by the navel, palms pressed together, fingertips pointing up; one foot in the tip toe position, other leg raised
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

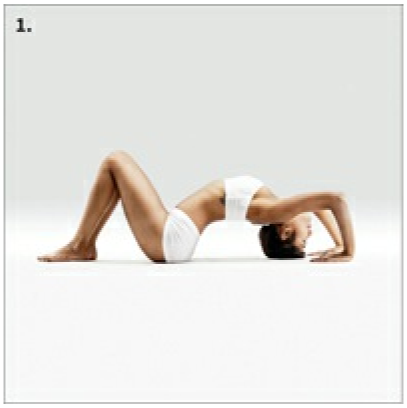
1.

2.

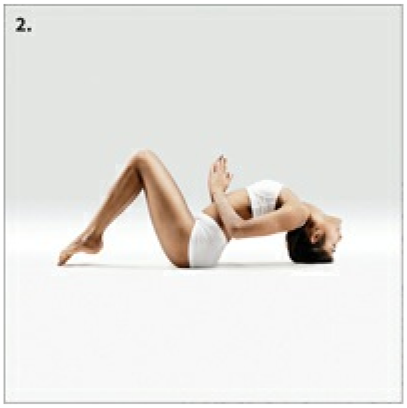

3.

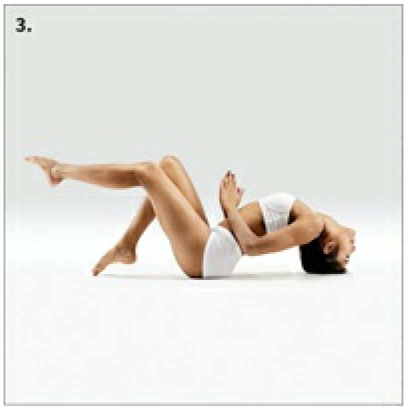

HEAD AND SITTING BONES ON THE FLOOR: LEGS STRAIGHT

Fish Pose

Matsyasana

(muhts-YAHS-uh-nuh)

Modification: legs straight on the floor, elbows to the floor, palms down to the floor by the hips
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

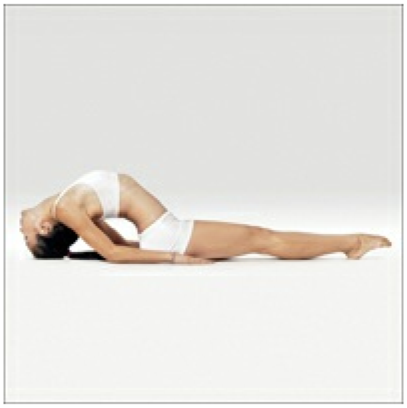

Reverse Prayer Fish Pose

Viparita Namaskar Matsyasana

(vi-puh-REE-tuh nuh-muhs-KAHR muhts-YAHS-uh-nuh)

Also Known As: Back of the Body Prayer Fish Pose (Paschima Namaskara Matsyasana)
Modification: legs straight to the floor
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

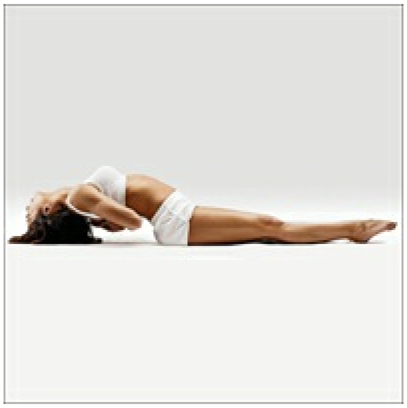

Intense Leg Stretch

Uttana Padasana

(ut-TAHN-uh puh-DAHS-uh-nuh)

Modification: legs straight on the floor, arms straight in front of the chest, palms pressed together
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

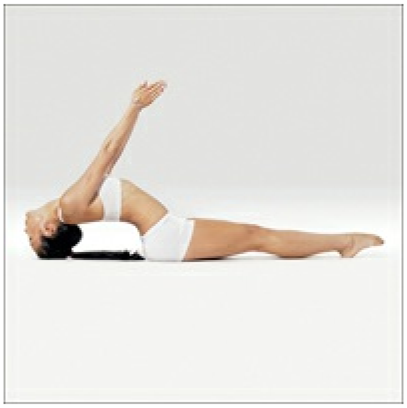

HEAD AND SITTING BONES ON THE FLOOR: ONE LEG STRAIGHT, ONE KNEE BENT

Fish Pose

Matsyasana

(muhts-YAHS-uh-nuh)

Modification: 1. one leg straight, other knee bent
2. one leg straight, other knee bent, bent leg crossed over
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

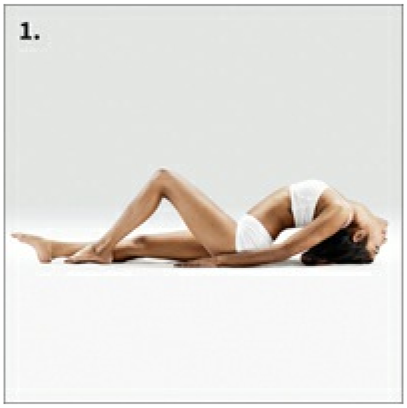

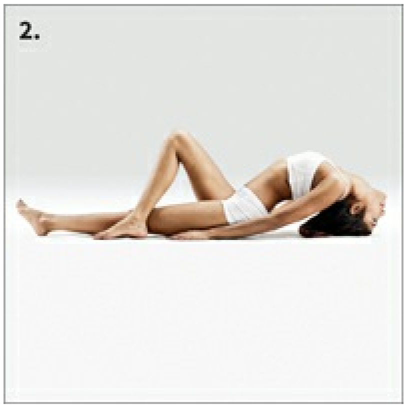

Half Bound Lotus Fish Pose

Ardha Baddha Padma Matsyasana

(UHR-duh BUH-duh PUHD-muh muhts-YAHS-uh-nuh)

Modification: sitting bones on the floor, palm to the floor by the hip
Pose Type: seated, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

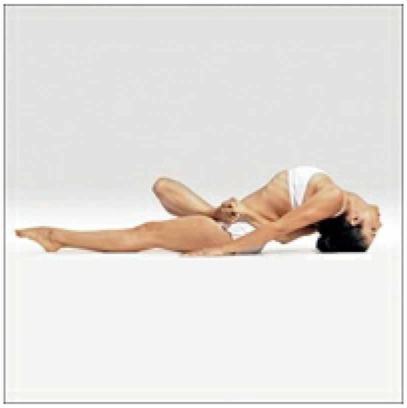

Half Bound Lotus Fish Pose

Ardha Baddha Padma Matsyasana

(UHR-duh BUH-duh PUHD-muh muhts-YAHS-uh-nuh)

Modification: one arm straight up to the sky, other arm grabbing onto the inside of the thigh behind the back; leg straight
Pose Type: seated, backbend, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

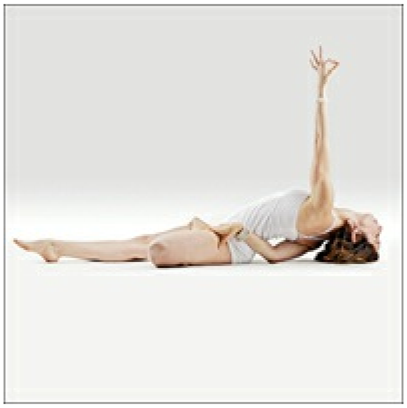

HEAD AND SITTING BONES ON THE FLOOR: HANDS AND FEET ON FLOOR—BINDING & GARUDA ARMS AND LEGS

Half Bound Lotus Half Fish Pose

Ardha Baddha Padma Ardha Matsyasana

(UHR-duh BUH-duh PUHD-muh UHR-duh muhts-YAHS-uh-nuh)

Modification: grabbing onto the foot with arm behind the back, fingertips of the other arm touching the floor overhead, elbow bent; knee of the free leg bent

1. heel down
2. heel up

Pose Type: seated, backbend, binding

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)

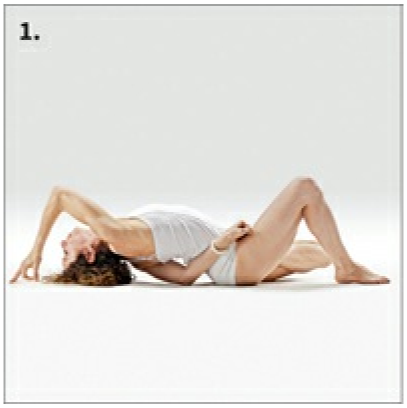
1.

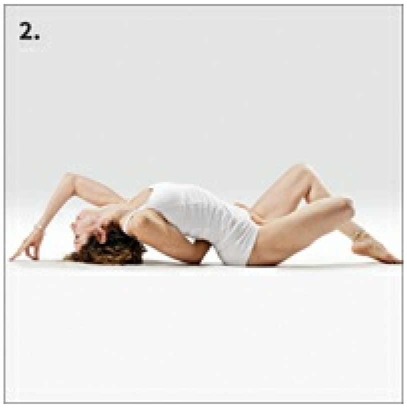
2.

Intense Leg Stretch Pose Dedicated to Garuda

Uttana Pada Garudasana

(ut-TAHN-uh PUH-duh guh-ru-DAHS-uh-nuh)

Modification: fingertips and bottom foot to the floor

Pose Type: seated, backbend

Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

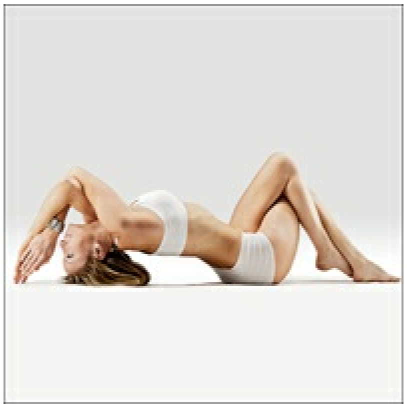

RECLINED BOUND ANGLE POSE & FISH POSE

Reclined Bound Angle Pose

Supta Baddha Konasana

(SUP-tuh BUH-duh ko-NAHS-uh-nuh)

Modification: hands to the inner thighs, palms up
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ●

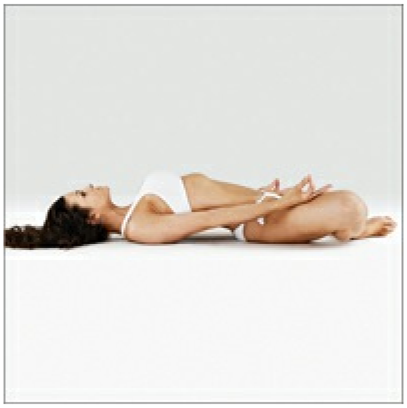

Reclined Bound Angle Pose

Supta Baddha Konasana

(SUP-tuh BUH-duh ko-NAHS-uh-nuh)

Also Known As: Bound Legs Fish Pose (Baddha Pada Matsyasana)
Modification: crown of the head to the floor, grabbing onto the shins, arms on top of the legs
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
● ● ● ● ● ● ●

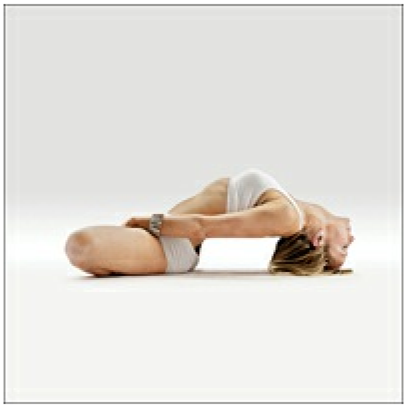

Fish Pose

Matsyasana

(muhts-YAHS-uh-nuh)

Modification: 1. palms up on the inside of the thighs
2. grabbing onto the big toes, elbows to the floor
3. grabbing onto the forearms, arms up over the head
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

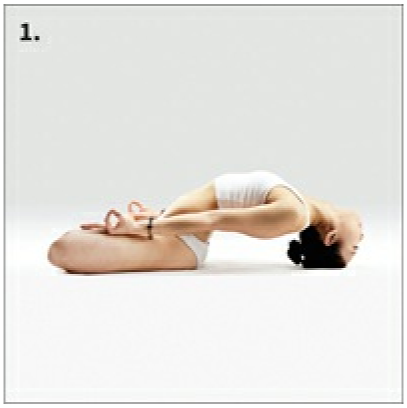

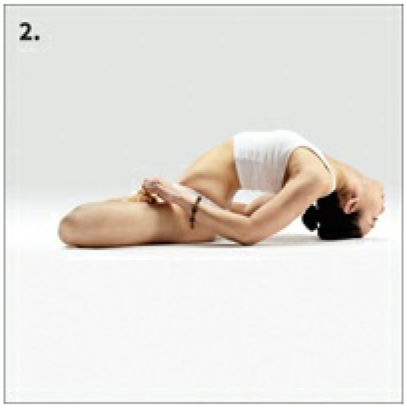

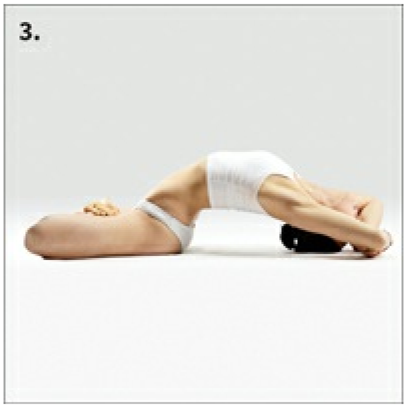

Reverse Prayer Intense Leg Stretch

Viparita Namaskar Uttana Padasana

(vi-puh-REE-tuh nuh-muhs-KAHR ut-TAHN-uh puh-DAHS-uh-nuh)

Also Known As: Back of the Body Prayer Intense Leg Stretch (Paschima Namaskara Uttana Padasana)
Modification: legs straight and off the floor, ankles crossed
Pose Type: seated, backbend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

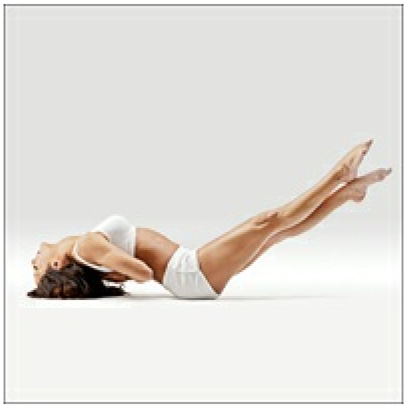

Intense Leg Stretch

Uttana Padasana

(ut-TAHN-uh puh-DAHS-uh-nuh)

Modification: legs straight and off the floor; arms straight in front of the chest, palms pressed together
Pose Type: seated, backbend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

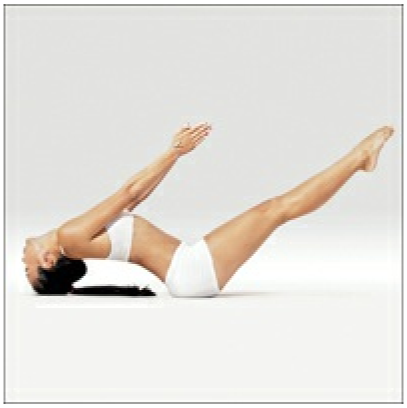

Intense Leg Stretch Pose Dedicated to Garuda

Uttana Pada Garudasana

(ut-TAHN-uh PUH-duh guh-ru-DAHS-uh-nuh)

Modification: legs and arms lifted off the floor
Pose Type: seated, backbend, core
Drishti Point: Angushtamadhye or Angushta Ma Dyai (thumbs)

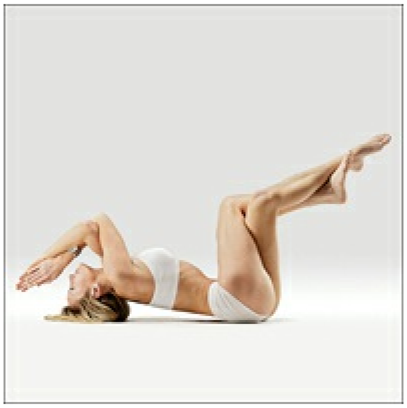

BRIDGE POSE

Complete Bridge Whole Body Pose

Paripurna Setu Bandha Sarvangasana

(puh-ri-POOR-nuh SEY-tu BUHN-duh suhr-vuhng-GAHS-uh-nuh)

Modification: palms to the floor under the shoulders, fingertips pointing to the heels; elbows tucked in, shoulder width apart
Pose Type: supine, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

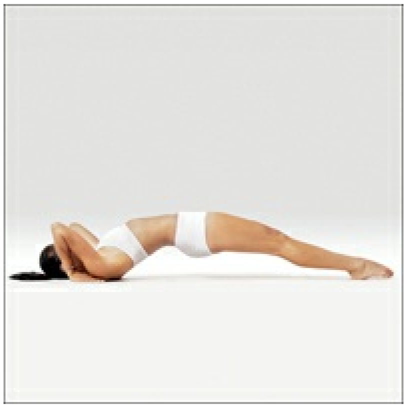

Vajra Cutter Sutra Bridge Pose

Vajracchedika Prajnaparamita Sutra Setu Bandhasana

(vahj-ruh-CHEY-di-kuh pruhj-NAH-puh-ruh-mi-TAH SU-truh SEY-tu buhn-DAHS-uh-nuh)

Modification: forehead to the floor, elbows splaying to the side, arms in “diamond” shape

Pose Type: standing (on the head and feet), backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल र य ह ॐ

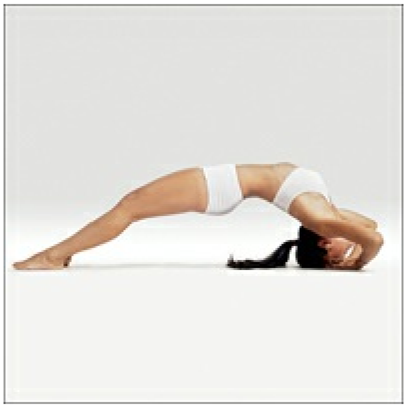

Bridge Pose

Setu Bandhasana

(SEY-tu buhn-DAHS-uh-nuh)

Modification: fingertips to the floor and pointing toward the heels

Pose Type: standing (on the head and feet), backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल र य ह ॐ

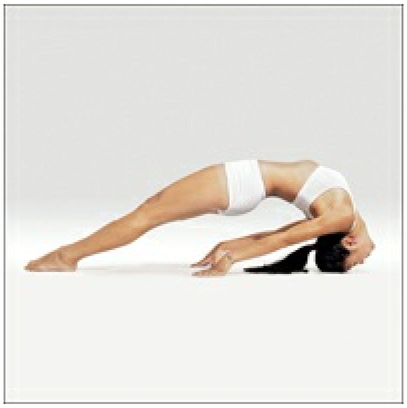

Bridge Pose

Setu Bandhasana

(SEY-tu buhn-DAHS-uh-nuh)

Modification: arms crossed in front of the chest, hands to the shoulders

Pose Type: standing (on the head and feet), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल र य ह ॐ

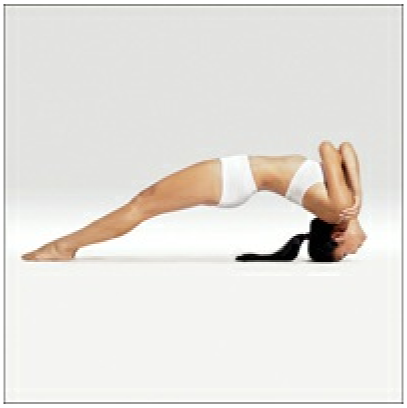

One-Legged Bridge Pose

Eka Pada Setu Bandhasana

(EY-kuh PUH-duh SEY-tu buhn-DAHS-uh-nuh)
Modification: fingertips pointing toward the heels, palms flat on the floor
Pose Type: standing (on the head and feet), backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र य ह ॐ

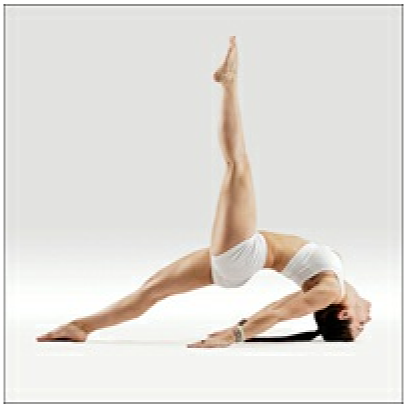

BOUND ANGLE POSE & LOTUS: BACKBEND

Reclined Bound Angle Pose

Supta Baddha Konasana

(SUP-tuh BUH-duh ko-NAHS-uh-nuh)
Also Known As: Bound Legs Fish Pose (Baddha Pada Matsyasana)
Modification: crown of the head to the floor, grabbing onto the shins under the legs

1. elbows to the floor
2. arms straight, elbows off the floor

Pose Type: backbend, standing (on the knees)

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

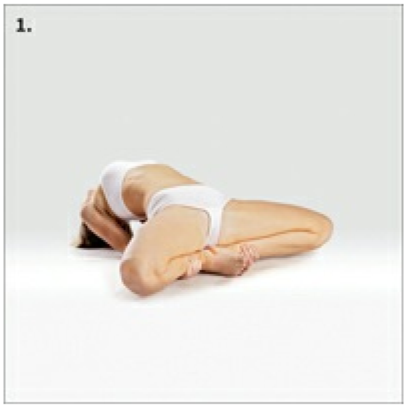
1.

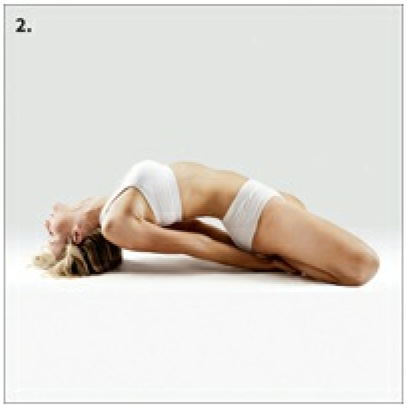
2.

Intense Lotus Peacock

Uttana Padma Mayurasana

(ut-TAHN-uh PUHD-muh muh-yoor-AHS-uh-nuh)

Also Known As: Intense Front Body Stretching and Rejuvenating Lotus Pose (Purvottana Padma Sarvangasana), All Body Parts Lotus Pose (Sarvangasana Padmasana)

Modification: knees on the floor

Pose Type: supine, backbend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

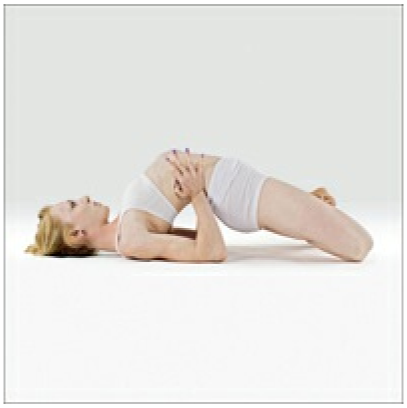

BOTH KNEES BENT: ONE LEG CROSSED OVER

Eye of the Needle Pose

Sucirandrasana

(soo-chi-ruhn-DRAHS-uh-nuh)

Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

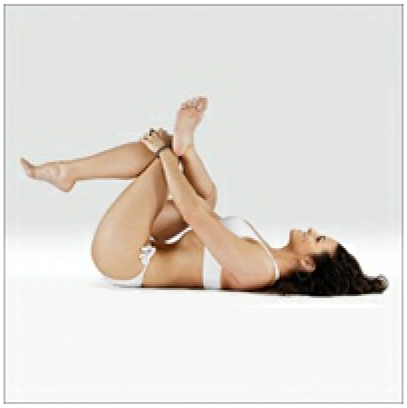

Reclined Leg Position of Cow Face Pose

Supta Pada Gomukhasana

(SUP-tuh PUH-duh go-muk-AHS-uh-nuh)

Modification: knees to the chest, grabbing onto both feet
1. front view
2. side view
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

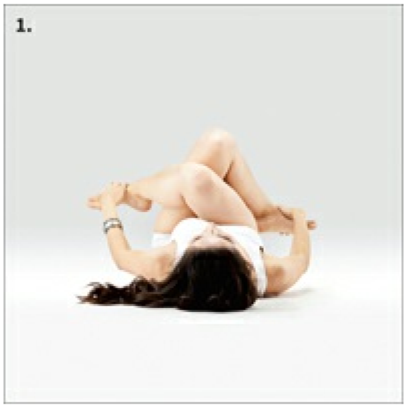

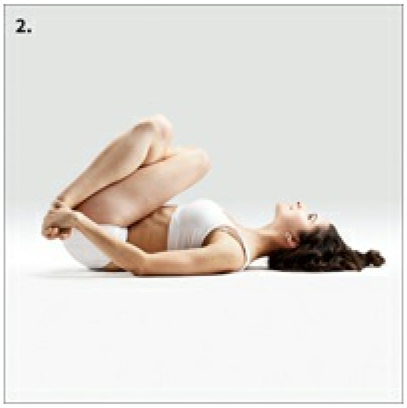

Reclined Twisting Fire Log Pose

Supta Parshva Agnistambhasana

(SUP-tuh PAHRSH-vuh uhg-ni-stuhm-BAHS-uh-nuh)

Pose Type: supine, forward bend, twist

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

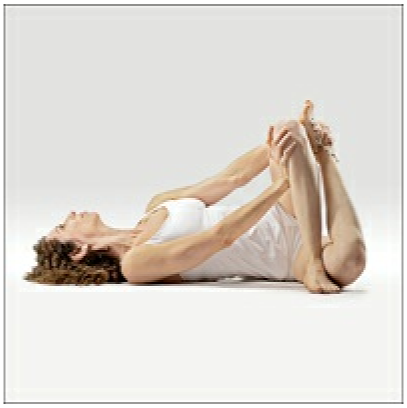

Reclined Leg Position of Half Cow Face Pose

Supta Pada Ardha Gomukhasana

(SUP-tuh PUH-duh UHR-duh go-muk-AHS-uh-nuh)

Modification: knees apart, top foot on the floor

1. one hand to the forehead, other hand on the top knee
2. grabbing onto the shoulders, arms crossed in front of the neck
3. one hand to the top ankle, other hand to the bottom foot

Pose Type: supine, forward bend, twist

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

1.

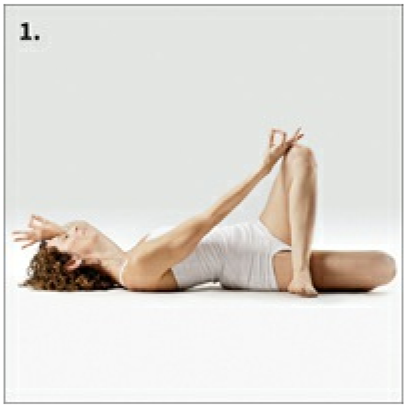

2.

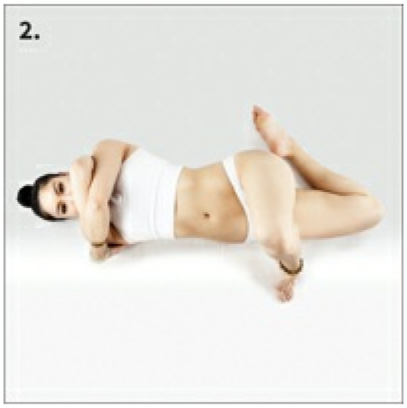

3.

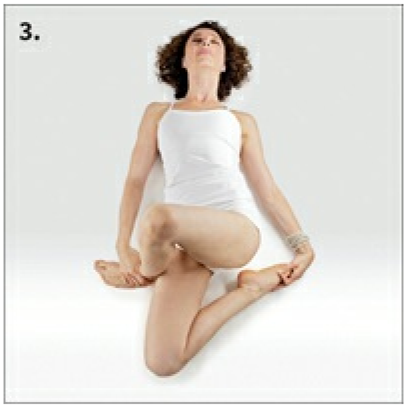

BOTH KNEES BENT: ONE LEG CROSSED OVER—KNEES TO THE FLOOR

Reclined Leg Position of Cow Face Pose

Supta Pada Gomukhasana

(SUP-tuh PUH-duh go-muk-AHS-uh-nuh)

Modification: knees to the floor

Pose Type: supine

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

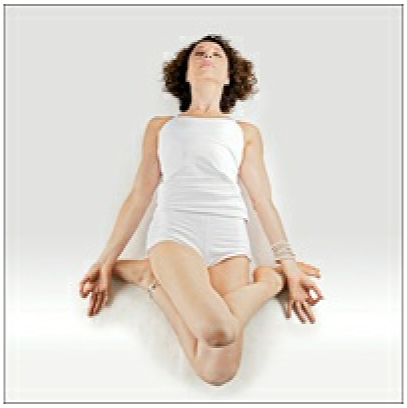

Universal All-Encompassing Diamond Pose

Vishvavajrasana

(vish-vuh-vuhj-RAHS-uh-nuh)

Also Known As: Double Diamond Pose

Modification: foot on top of the knee; hand to the sole of the back foot, other hand resting on the thigh of the front leg, looking straight ahead

Pose Type: supine, twist

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

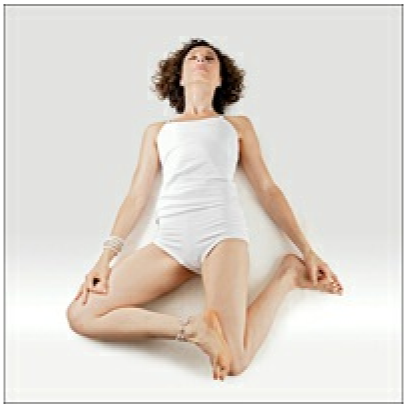

Revolved Universal All-Encompassing Diamond Pose

Parivritta Vishvavajrasana *(puh-ri-VRIT-tuh vish-vuh-vuhj-RAHS-uh-nuh)*

Also Known As: Revolved Double Diamond Pose

Modification: hand grabbing onto the foot; other hand grabbing onto the opposite knee, looking up to the sky
Pose Type: supine, forward bend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

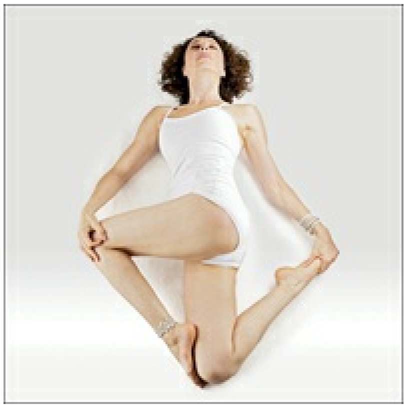

Revolved Universal All-Encompassing Diamond Pose

Parivritta Vishvavajrasana *(puh-ri-VRIT-tuh vish-vuh-vuhj-RAHS-uh-nuh)*
Also Known As: Revolved Double Diamond Pose
Modification: hand grabbing onto the ankle, arm under the thigh; other hand grabbing onto the ankle of the bottom leg, looking straight ahead
Pose Type: supine, forward bend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

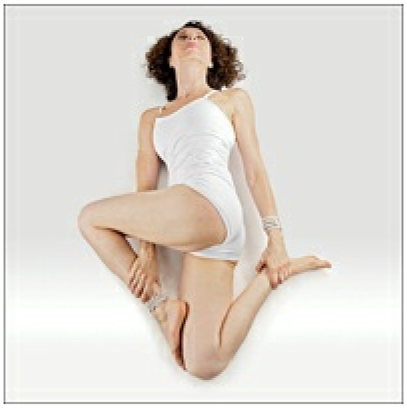

BOTH KNEES BENT: KNEES TOGETHER

Reclined Child’s Pose

Supta Balasana

(SUP-tuh bah-LAHS-uh-nuh)

Also Known As: Pelvic Pose (Apanasana)
Modification: binding arms, grabbing onto the triceps
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

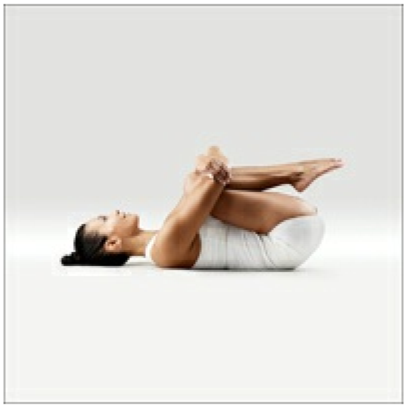

Wind-Relieving Pose

Vayu Muktyasana

(VAH-yuh muk-TYAHS-uh-nuh)

Also Known As: Wind Releaser Pose (Pavana Muktasana)
Modification: fingers interlocked on top of the shins
Pose Type: supine, forward bend, core
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

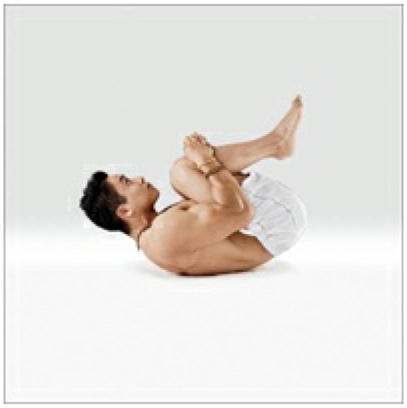

Twisted Stomach Pose

Jatara Parivartanasana

(JAH-tuh-ruh puh-ri-vuhrt-uhn-AHS-uh-nuh)

Also Known As: Reclining Waist Pose Prep.(Supta Madhyasana), Sideways Wind Releaser Pose (Pavana Muktasana)
Modification: knees bent, arms straight out to the sides
Pose Type: supine, forward bend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं यं ॐ

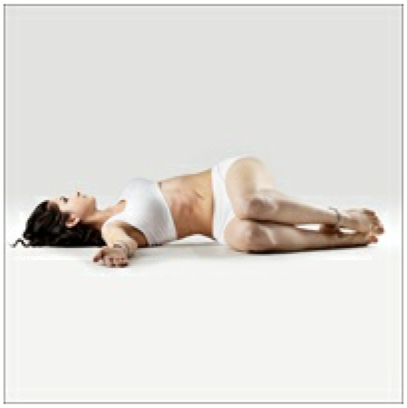

ONE LEG STRAIGHT, ONE KNEE BENT

One-Legged Wind Relieving Pose

Eka Pada Pavana Muktasana

(EY-kuh PUH-duh puh-vuh-nuh muk-TAHS-uh-nuh)

Also Known As: Reclined Big Toe Pose A Prep. (Supta Padangushtasana A Prep.)
Modification: head on the floor, fingers interlocked on top of the shin
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं रं ॐ

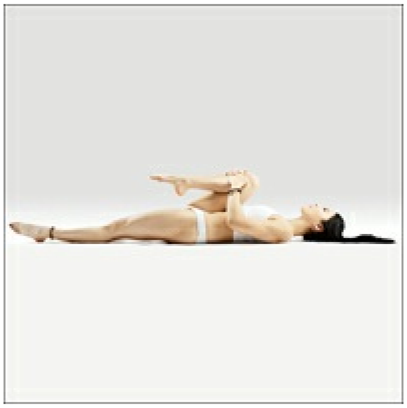

Reclining Tree Pose

Supta Vrikshasana

(SUP-tuh vrik-SHAHS-uh-nuh)

Also Known As: Reclined Big Toe Pose B Prep. (Supta Padangushtasana B Prep.)

Modification: hand to the bent knee, other hand along side of the torso, toes flexed in

Pose Type: supine

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

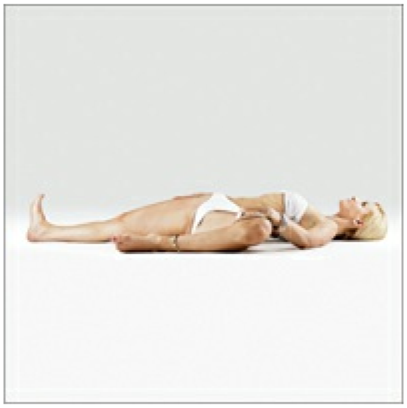

BOTH LEGS STRAIGHT: ONE LEG TO THE SIDE

Revolved Reclining Hand to Foot Pose

Parivritta Supta Hasta Padasana

(puh-ri-VRIT-tuh SUP-tuh HUH-stuh puh-DAHS-uh-nuh)

Modification: legs straight, grabbing onto the inside arch of the foot

Pose Type: supine, forward bend, twist

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

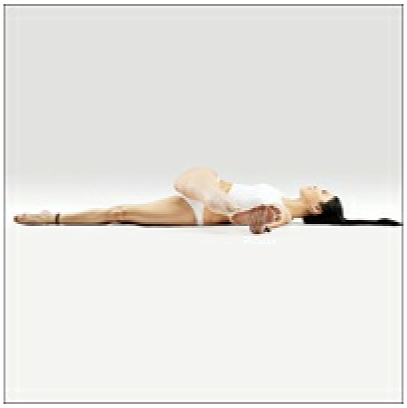

Reclining Big Toe Pose 2

Supta Padangushtasana 2

(SUP-tuh puhd-ahng-goosh-TAHS-uh-nuh)

Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

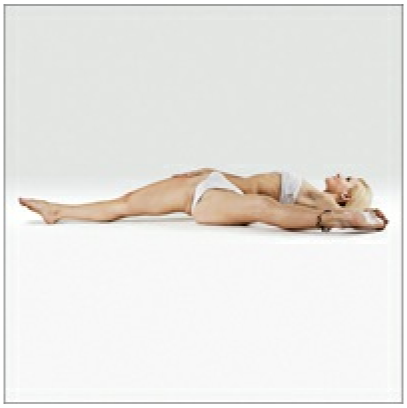

ONE LEG STRAIGHT, ONE KNEE BENT: TWIST

Reclined Lord of the Fishes Pose

Supta Matsyendrasana

(SUP-tuh muhts-yeyn-DRAHS-uh-nuh)

Also Known As: Revolving Reclined Big Toe Pose Prep. (Parivritta Supta Padangushtasana Prep.), Revolved Reclining Tree Pose (Parivritta Supta Vrikshasana)
Modification: knee bent at 90 degrees
Pose Type: supine, forward bend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

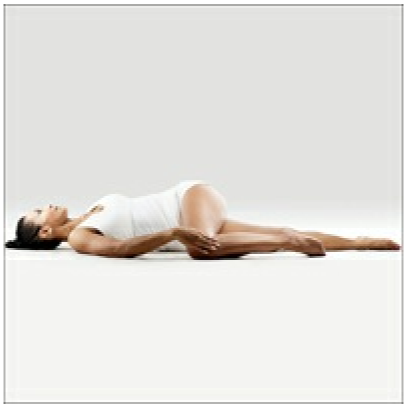

GARUDA LEGS: TWISTS

Sideways Reclining Leg Position of the Pose Dedicated to Garuda

Parshva Supta Pada Garudasana

(PAHRSH-vuh SUP-tuh PUH-duh guh-ru-DAHS-uh-nuh)

Modification: hand resting on the top knee, other arm to the floor along side the torso, looking to the side; arms straight out to the side, looking straight ahead
Pose Type: supine, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre (hands)
लं वं यं ॐ

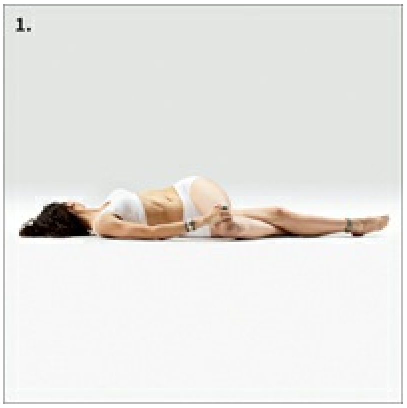
1.

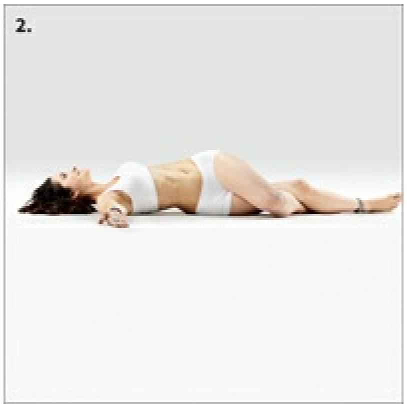
2.

BOTH LEGS STRAIGHT AND TOGETHER: TWIST

Twisted Stomach Pose

Jathara Parivartanasana

(JAH-tuh-ruh puh-ri-vuhrt-uhn-AHS-uh-nuh)

Also Known As: Reclining Waist Pose (Supta Madhyasana)
Modification: legs and arms straight
Pose Type: supine, forward bend, twist
Drishti Point: Hastagrai or Hastagre (hands)

ल य

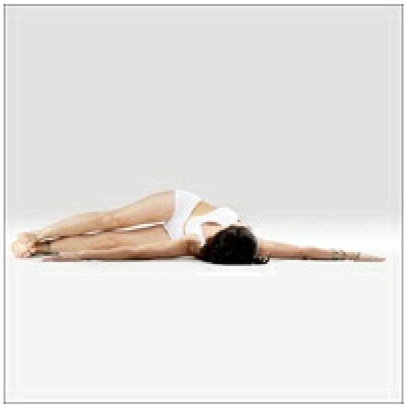

ONE LEG STRAIGHT, ONE KNEE BENT: LEG CRADLE & HALF HAPPY BABY POSE

Reclined Baby Cradle Pose

Supta Hindolasana

(SUP-tuh hin-do-LAHS-uh-nuh)

Modification: back leg straight
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र ॐ

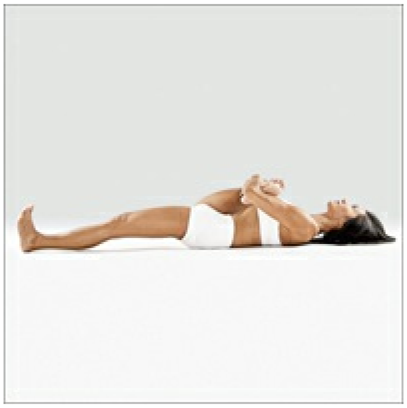

Reclined One Hand to Foot Pose

Supta Eka Hasta Padasana

(SUP-tuh EY-kuh HUH-stuh puh-DAHS-uh-nuh)

Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व ॐ

HAPPY BABY POSE & ONE FOOT BEHIND THE HEAD

Happy Baby Pose

Sukha Balasana

(SUK-uh bahl-AHS-uh-nuh)
Also Known As: Happy Baby Pose (Ananda Balasana)
Modification: grabbing onto the big toes of the feet
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

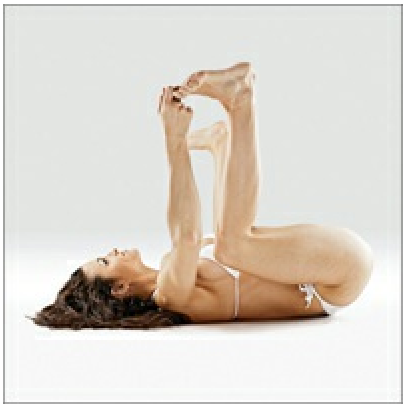

Reclined Star Pose

Supta Tarasana

(SUP-tuh tah-RAHS-uh-nuh)
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

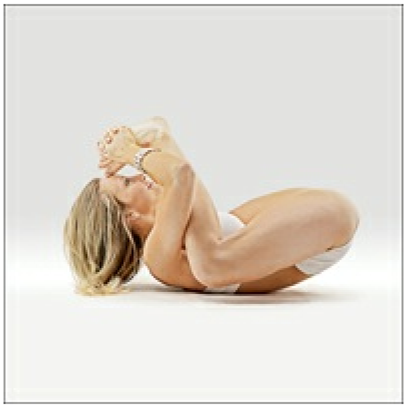

Happy Baby Pose

Sukha Balasana

(SUK-uh bahl-AHS-uh-nuh)

Also Known As: Happy Baby Pose (Ananda Balasana)
Modification: arms crossed, grabbing onto the outside edges of the feet
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ●

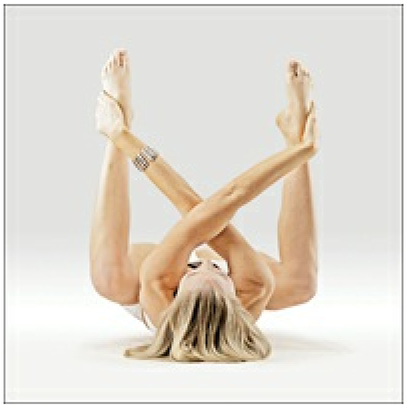

Reclined One Hand to Foot Pose

Supta Eka Hasta Padasana

(SUP-tuh EY-kuh HUH-stuh puh-DAHS-uh-nuh)

Modification: grabbing onto the back foot, heel to the sitting bone, toes pointed to the floor
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

● ● ●

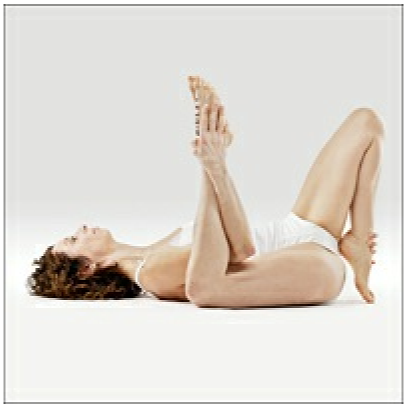

Reclined One Hand to Foot One Leg Behind the Head Pose in Infinity Pose

Supta Eka Hasta Pada Eka Pada Shirshasana in Anantasana

(SUP-tuh EY-kuh-HUH-stuh PUH-duh EY-kuh PUH-duh sheer-SHAS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Modification: bottom knee to the floor, toes pointing up to the sky
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

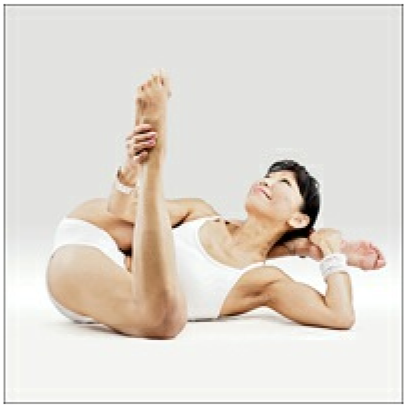

One Leg Behind the Head in Infinity Pose

Eka Pada Shirshasana in Anantasana

(EY-kuh PUH-duh sheer-SHAHS-uh-nuh in uhn-uhnt-AHS-uh-nuh)

Modification: bottom heel toward the hip
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

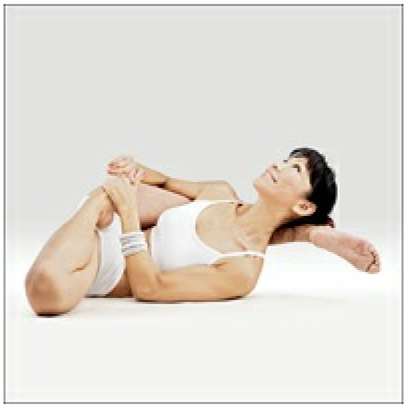

ONE AND BOTH FEET BEHIND THE HEAD

Pose Dedicated to Bhairava in Half Upward Facing Western Intense Stretch 2

Bhairavasana in Ardha Urdhva Mukha Paschimottanasana 2

(b-eye-ruh-VAHS-uh-nuh in UHR-duh OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Pose Dedicated to Bhairava in Half Upward Facing Forward Bend 2
Modification: 1. front view
2. back view
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

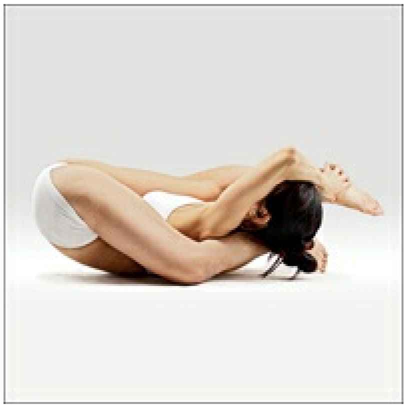

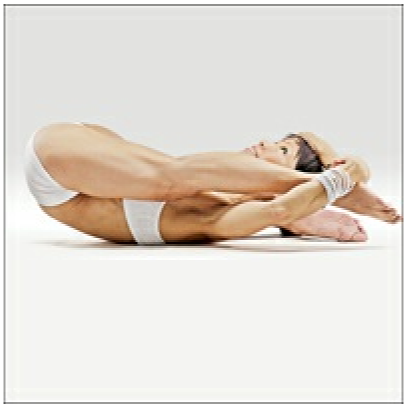

Hands Bound Yogic Sleep Pose

Baddha Hasta Yoganidrasana

(BUH-duh HUH-stuh yo-guh-ni-DRAHS-uh-nuh)

Modification: hands bound
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं ॐ

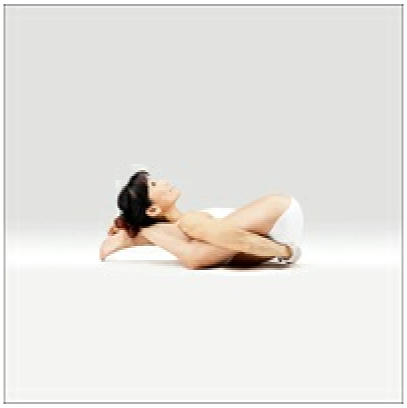

HORIZONTAL SPLITS

Reclined Angle Pose

Supta Konasana

(SUP-tuh ko-NAHS-uh-nuh)

Modification: back flat on the floor, feet lifted off the floor, grabbing onto the big toes
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
लं वं ॐ

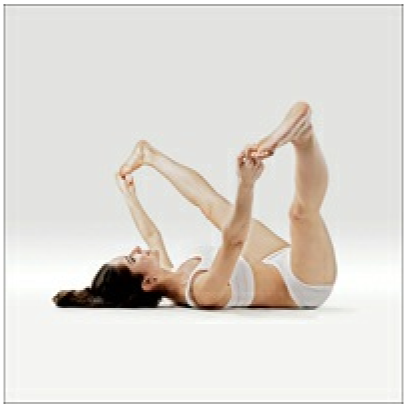

Reclined Angle Pose

Supta Konasana

(SUP-tuh ko-NAHS-uh-nuh)

Modification: back flat on the floor, toes to the floor, grabbing onto the outside edges of the feet
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

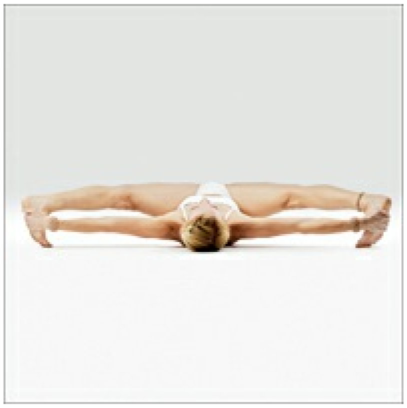

Reclined Angle Pose

Supta Konasana

(SUP-tuh ko-NAHS-uh-nuh)

Modification: back flat on the floor, feet lifted off the floor, arms straight out to the sides, palms down to the floor
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

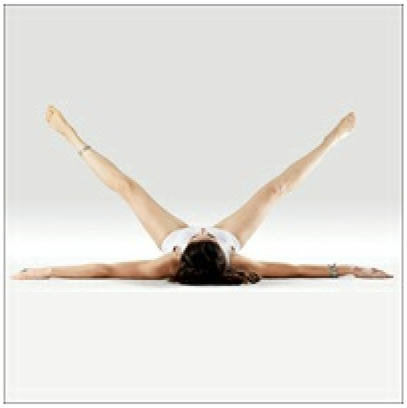

Sideways Reclined Angle Pose

Parshva Supta Konasana

(PAHRSH-vuh SUP-tuh ko-NAHS-uh-nuh)

Modification: back flat on the floor, head turning to side Side 1: leg in the air; arm straight on the floor, palm facing up
Side 2: leg on the floor, grabbing onto the big toe
Pose Type: supine, twist
Drishti Point: Hastagrai or Hastagre (hands)

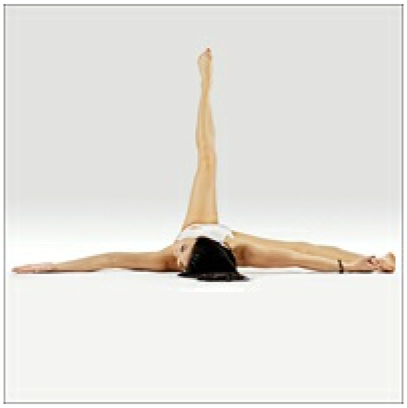

Revolved Half Bound Reclined Angle Pose

Parivritta Ardha Baddha Supta Konasana

(puh-ri-VRIT-uh UHR-duh BUH-duh SUP-tuh ko-NAHS-uh-nuh)

Modification: both legs straight Side 1: leg pointing to the sky; arm wrapping around the back, hand to the inside of the hip on the opposite side
Side 2: leg on the floor, opposite hand grabbing onto the foot
Pose Type: supine, twist, binding
Drishti Point: Hastagrai or Hastagre (hands), Padayoragrai or Padayoragre (toes/feet)

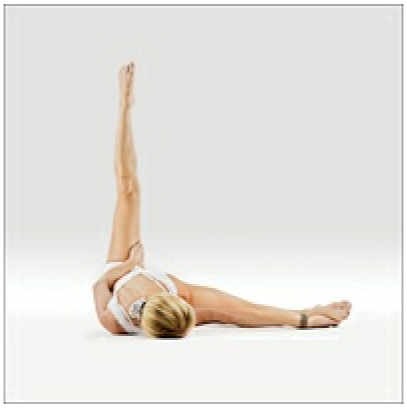

ONE LEG STRAIGHT TO THE SKY, ONE LEG CROSSED OVER

Upward Facing Western Intense Stretch

Urdhva Mukha Paschimottanasana

(OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward Facing Forward Bend
Modification: legs crossed, one arm threaded through the legs, other hand to the center of the chest
Pose Type: supine, forward bend, twist
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

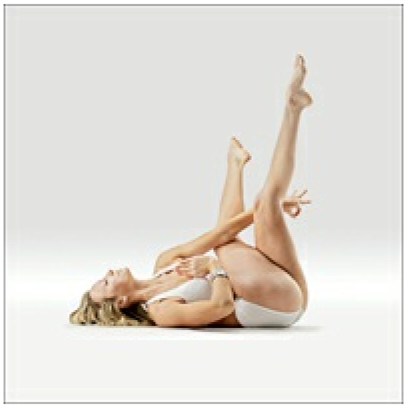

Reclined Hands Bound Pose Dedicated to Astavakra

Supta Baddha Hasta Ashtavakrasana

(SUP-tuh BUH-duh HUH-stuh uhsh-tuh-vuh-KRAHS-uh-nuh)

Also Known As: Reclined Hands Bound Eight Angle Pose
Modification: ankles crossed
Pose Type: supine, forward bend, twist, binding
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

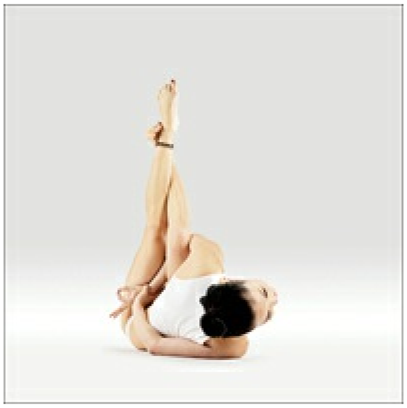

ONE KNEE BENT TO THE BACK, OTHER LEG STRAIGHT

One-Legged Reclined Hero Pose

Eka Pada Supta Virasana

(EY-kuh PUH-duh SUP-tuh veer-AHS-uh-nuh)

Modification: straight leg lifted, fingers interlocked on the back of the thigh
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

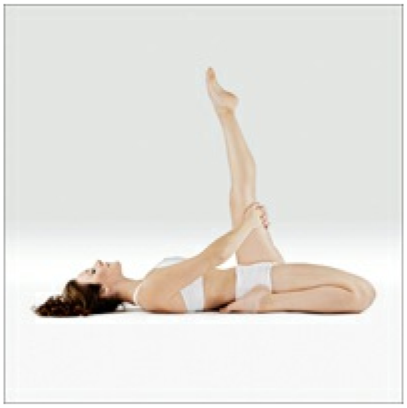

One-Legged Reclined Hero Pose

Eka Pada Supta Virasana

(EY-kuh PUH-duh SUP-tuh veer-AHS-uh-nuh)

Modification: leg straight out to the side, grabbing onto the outside edge of the foot
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

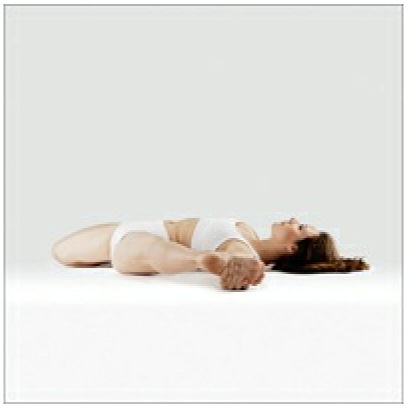

One-Legged Reclined Hero Pose

Eka Pada Supta Virasana

(EY-kuh PUH-duh SUP-tuh veer-AHS-uh-nuh)

Modification: leg straight on the floor, toes flexed in; arms straight out to the sides, palms facing up
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

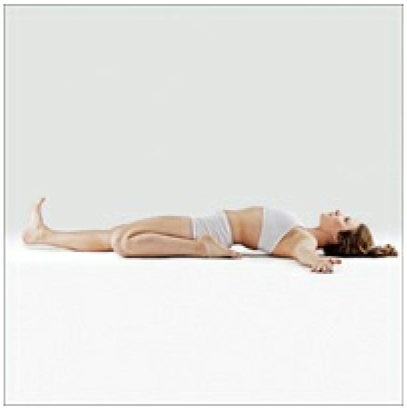

ONE KNEE BENT TO THE BACK, OTHER LEG BENT

One-Legged Reclined Hero Pose

Eka Pada Supta Virasana

(EY-kuh PUH-duh SUP-tuh veer-AHS-uh-nuh)

Modification: Leg 1: in Hero Pose (Virasana)
Leg 2: knee up to the sky, heel to the sitting bones; arms up over the head, palms facing up
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व य ॐ

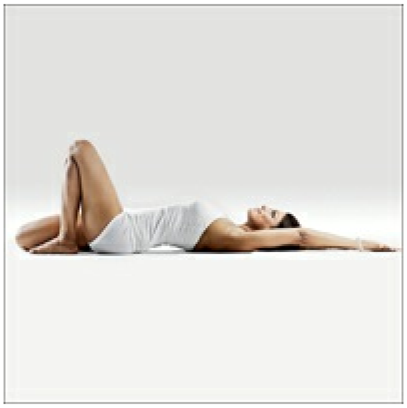

One-Legged Reclined Hero Pose

Eka Pada Supta Virasana

(EY-kuh PUH-duh SUP-tuh veer-AHS-uh-nuh)

Modification: Leg 1: in Hero Pose (Virasana)
Leg 2: knee toward the chest, toes flexed in; fingertips interlocked on top of the knee
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व र ॐ

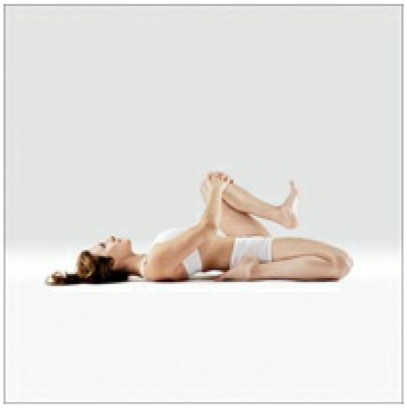

Half Bound Angle Pose in One-Legged Reclined Hero Pose

Ardha Baddha Konasana in Eka Pada Supta Virasana

(UHR-duh BUH-duh ko-NAHS-uh-nuh in EY-kuh PUH-duh SUP-tuh veer-AHS-uh-nuh)

Modification: arms up over the head, palms facing upward
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

ल व य ॐ

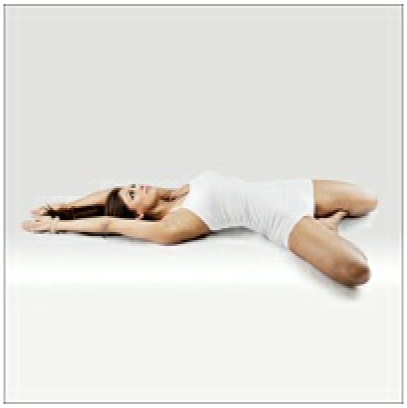

Reclined Hero Pose

Supta Virasana

(SUP-tuh veer-AHS-uh-nuh)

Also Known As: Bed Pose B (Paryankasana B)
Modification: arms up over the head, palms facing upward
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ॐ

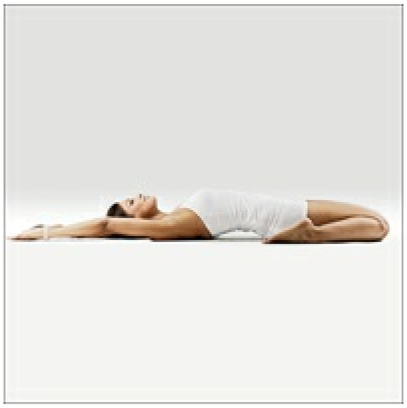

TOES TURNED OUT TO THE SIDE: ONE KNEE BENT TO THE BACK, OTHER LEG BENT

Reclined One-Legged Thunderbolt Pose

Supta Eka Pada Vajrasana

(SUP-tuh EY-kuh PUH-duh vuhj-RAHS-uh-nuh)

Modification: Leg 1: in Thunderbolt Pose (Vajrasana)
Leg 2: knee up to the sky, heel to the sitting bones; arms up over the head, fingers interlocked, palms facing out
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

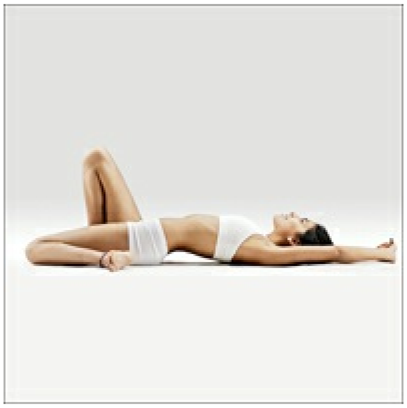

Reclined One-Legged Thunderbolt Pose

Supta Eka Pada Vajrasana

(SUP-tuh EY-kuh PUH-duh vuhj-RAHS-uh-nuh)

Modification: Leg 1: in Thunderbolt Pose (Vajrasana)
Leg 2: knee up to the sky, sole of the foot to the thigh; arms up over the head, fingers interlocked, palms facing out
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

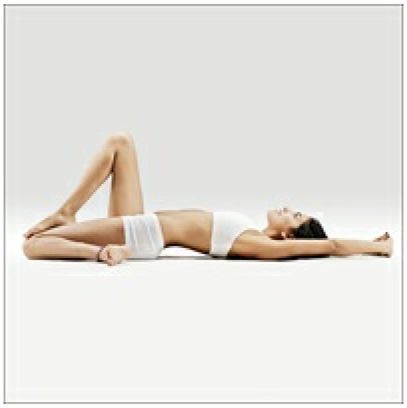

Reclined One-Legged Thunderbolt Pose

Supta Eka Pada Vajrasana

(SUP-tuh EY-kuh PUH-duh vuhj-RAHS-uh-nuh)

Modification: Leg 1: in Thunderbolt Pose (Vajrasana)
Leg 2: knee bent at 90 degrees, heel off the floor; arms up over the head, fingers interlocked, palms facing out
Pose Type: supine

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ॐ

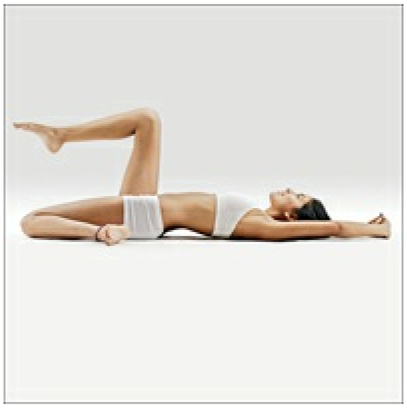

Reclined Complete Thunderbolt Pose

Supta Paripurna Vajrasana

(SUP-tuh puh-ri-POOR-nuh vuhj-RAHS-uh-nuh)
Modification: arms up over the head, fingers interlocked, palms facing out
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व य ॐ

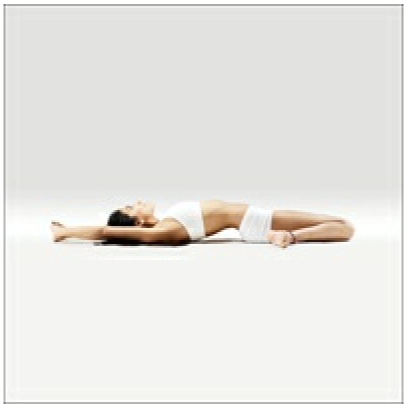

TOES TURNED OUT TO THE SIDE: ONE KNEE BENT TO THE BACK, OTHER LEG STRAIGHT & FOREARM POSES

Reclined One-Legged Thunderbolt Pose

Supta Eka Pada Vajrasana

(SUP-tuh EY-kuh PUH-duh vuhj-RAHS-uh-nuh)
Modification: leg straight to the floor, toes flexed in; arms up over the head, fingers interlocked, palms facing out

Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

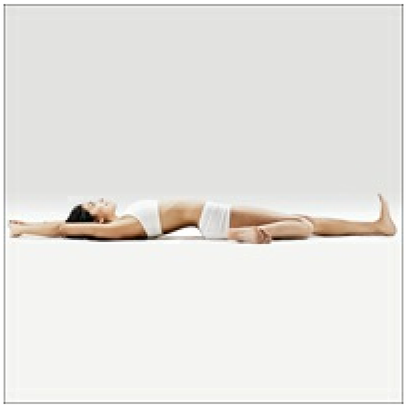

Reclined One-Legged Thunderbolt Pose

Supta Eka Pada Vajrasana

(SUP-tuh EY-kuh PUH-duh vuhj-RAHS-uh-nuh)
Modification: leg straight and lifted off the floor by using a Yoga strap
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

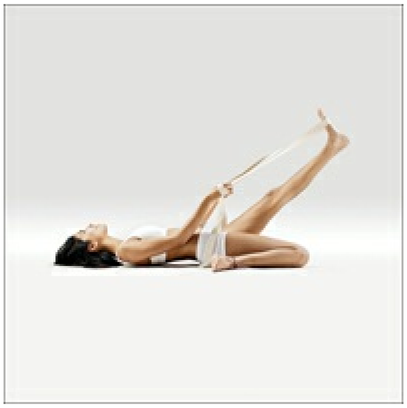

Reclined One-Legged Thunderbolt Pose

Supta Eka Pada Vajrasana

(SUP-tuh EY-kuh PUH-duh vuhj-RAHS-uh-nuh)
Modification: forearms to the floor in a backbend, one leg straight to the floor, toes flexed in, other leg in Thunderbolt Pose (Vajrasana)
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

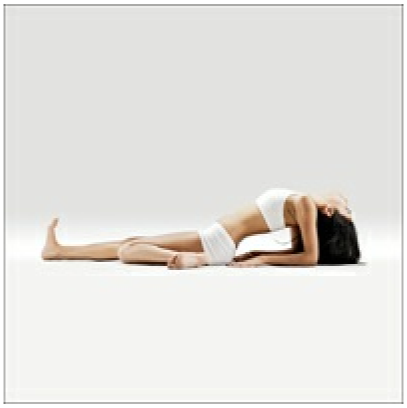

Reclined Complete Thunderbolt Pose

Supta Paripurna Vajrasana

(SUP-tuh puh-ri-POOR-nuh vuhj-RAHS-uh-nuh)

Modification: forearms to the floor in a backbend
Pose Type: seated, backbend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

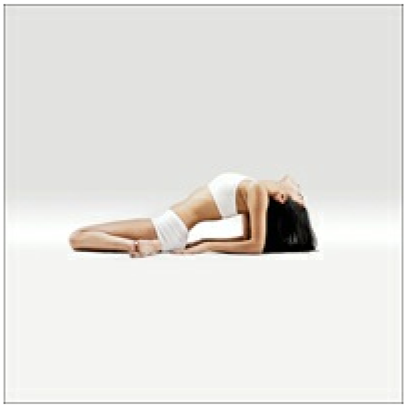

VERTICAL SPLITS: SCISSOR LEGS—BOTH KNEES BENT & ONE KNEE BENT, ONE LEG STRAIGHT

Reclining Both Hands to the Leg Pose

Supta Dwi Hasta Padasana

(SUP-tuh DWI-huh-stuh puh-DAHS-uh-nuh)

Modification: elbows bent, triceps on the floor, fingers interlocked behind the back of the thigh Leg 1: slight knee bend, foot flat on the floor Leg 2: knee bent at 90 degrees, knee toward the chest
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows) or Padayoragrai or

Padayoragre (toes/feet)

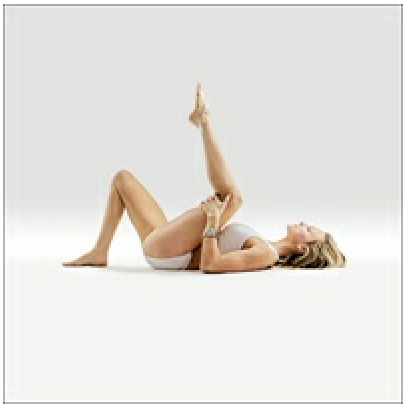

Reclining Half One Leg Extended Pose

Supta Ardha Utthita Eka Padasana

(SUP-tuh UHR-duh UT-ti-tuh EY-kuh puh-DAHS-uh-nuh)

Modification: palms flat to the floor on the sides.

Leg 1: straight up to the sky Leg 2: knee bent, foot flat to the floor, heel to the sitting bones

Pose Type: supine, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

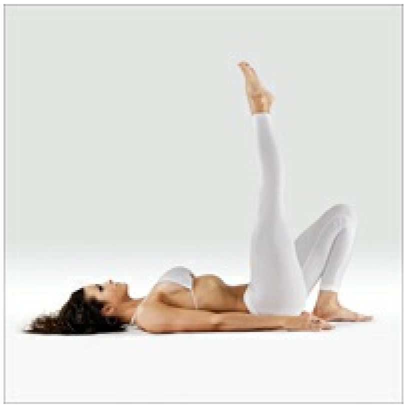

Reclining Both Hands to the Leg Pose

Supta Dwi Hasta Padasana

(SUP-tuh DWi-huh-stuh puh-DAHS-uh-nuh)

Modification: grabbing onto the calf muscle, elbows bent Leg 1: straight and toward the chest Leg 2: knee bent, foot flat to the floor

Pose Type: supine, forward bend

Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

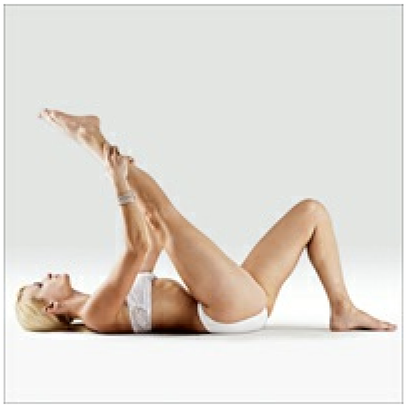

Reclining Both Hands to the Leg Pose

Supta Dwi Hasta Padasana

(SUP-tuh DWI-huh-stuh puh-DAHS-uh-nuh)

Modification: grabbing onto the calf muscle, elbows bent at 90 degrees, nose toward the shin Leg 1: straight and toward the chest Leg 2: knee bent, foot flat to the floor
Pose Type: supine, forward bend, core
Drishti Point: Padayoragrai or Padayoragre (toes/feet), Nasagrai or Nasagre (nose)

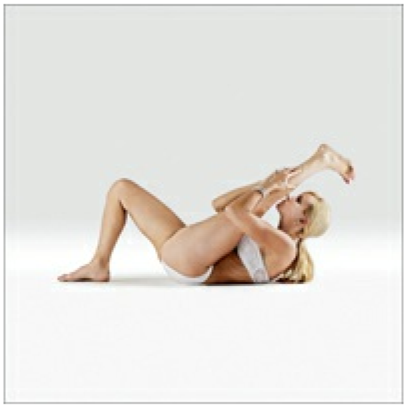

VERTICAL SPLITS: SCISSOR LEGS—BOTH LEGS STRAIGHT

Reclining One Leg Extended Pose

Supta Utthita Eka Padasana

(SUP-tuh UT-ti-tuh EY-kuh puh-DAHS-uh-nuh)

Modification: palms flat to the floor on the sides Leg 1: straight up to the sky Leg 2: straight out on

the floor, toes flexed in
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

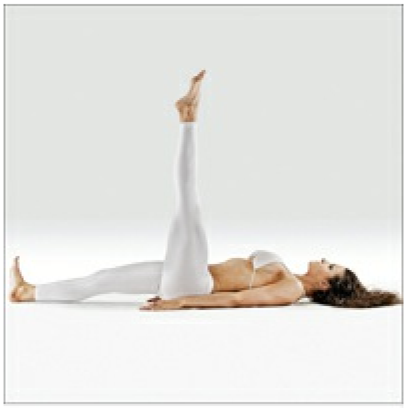

Reclining Big Toe Pose 1

Supta Padangushtasana 1

(SUP-tuh puhd-uhng-goosh-TAHS-uh-nuh)
Modification: back flat on the floor
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

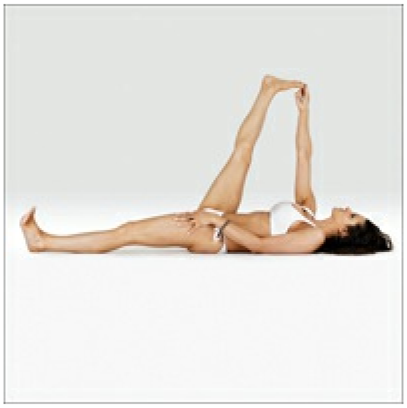

Reclining Big Toe Pose 1

Supta Padangushtasana 1

(SUP-tuh puhd-uhng-goosh-TAHS-uh-nuh)
Modification: head and shoulders off the floor, nose to the shin

Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Nasagrai or Nasagre (nose)
ल व र ह ॐ

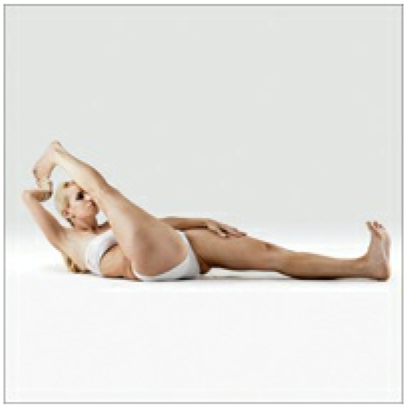

Reclining Both Hands to the Leg Pose

Supta Dwi Hasta Padasana

(SUP-tuh DWI-huh-stuh puh-DAHS-uh-nuh)
Modification: head and shoulders off the floor, nose to the shin, toes pointed
Pose Type: supine, forward bend, core
Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)
ल व र ह ॐ

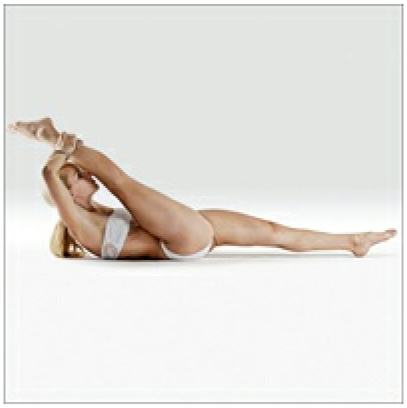

Reclining One Hand to the Leg Pose

Supta Eka Hasta Padasana

(SUP-tuh EY-kuh HUH-stuh puh-DAHS-uh-nuh)
Modification: hand to opposite leg, head and shoulders off the floor, nose to the shin, toes flexed in
Pose Type: supine, forward bend, core

Drishti Point: Nasagrai or Nasagre (nose), Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

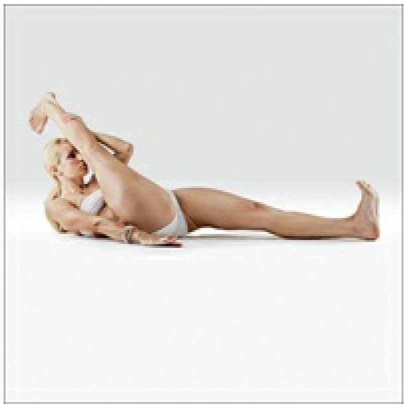

Reclining One Leg Extended Pose

Supta Utthita Eka Padasana

(SUP-tuh UT-ti-tuh EY-kuh puh-DAHS-uh-nuh)

Modification: palms flat to the floor on the sides, head and shoulders off the floor, nose to the knee, bottom toes flexed in
Pose Type: supine, forward bend
Drishti Point: Nasagrai or Nasagre (nose)

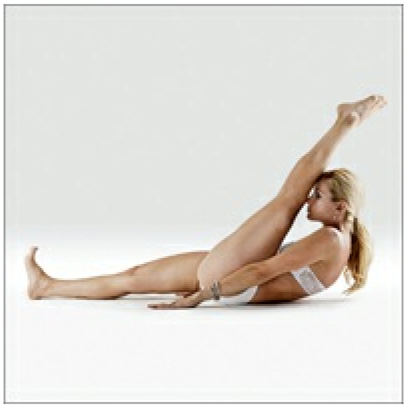

BOTH LEGS STRAIGHT AND TOGETHER

Reclining Legs Extended Pose

Supta Utthita Padasana

(SUP-tuh UT-ti-tuh puh-DAHS-uh-nuh)

Modification: arms crossed in front of the chest

Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

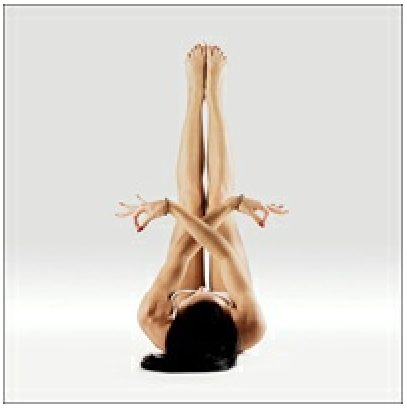

Reclining Legs Extended Pose

Supta Utthita Padasana

(SUP-tuh UT-ti-tuh puh-DAHS-uh-nuh)

Modification: arms straight up to the sky
1. arms and legs perpendicular to the floor
2. arms and legs slightly lowered toward the floor
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet), Hastagrai or Hastagre (hands)

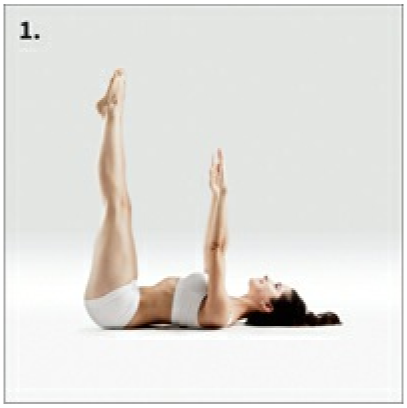

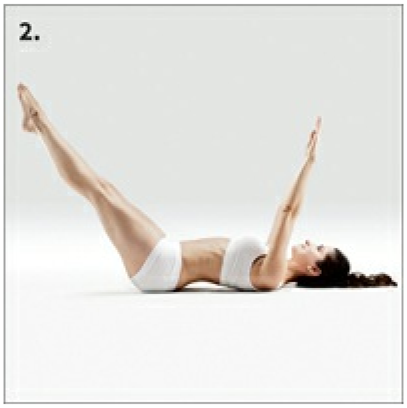

Half Upward Facing Western Intense Stretch 2

Ardha Urdhva Mukha Paschimottanasana 2

(UHR-duh OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Half Upward Facing Forward Bend, Upward Facing Western Intense Stretch Pose 2 Prep. (Urdhva Mukha Paschimottanasana 2 Prep.)
Modification: grabbing onto the calf muscles
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

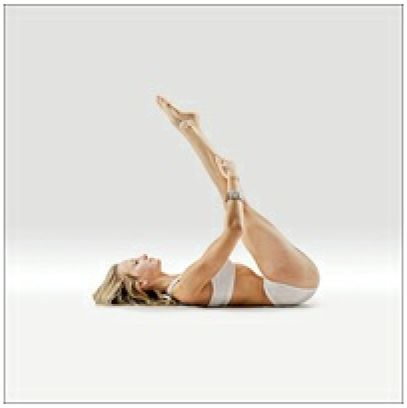

Upward Facing Western Intense Stretch 2

Urdhva Mukha Paschimottanasana 2

(OORD-vuh MUK-uh puhsh-chi-mo-tahn-AHS-uh-nuh)

Also Known As: Upward Facing Forward Bend 2
Pose Type: supine, forward bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Padayoragrai or Padayoragre (toes/feet)

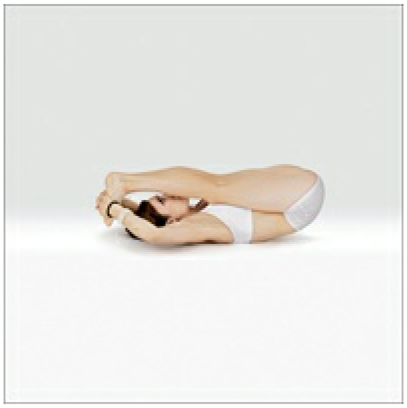

SUPINE LOTUS & CORPSE POSE

Reclining Lotus Pose

Supta Padmasana

(SUP-tuh puhd-MAHS-uh-nuh)

Also Known As: Fish Pose (Matsyasana)
Modification: back flat on the floor; arms up over the head, palms facing up
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

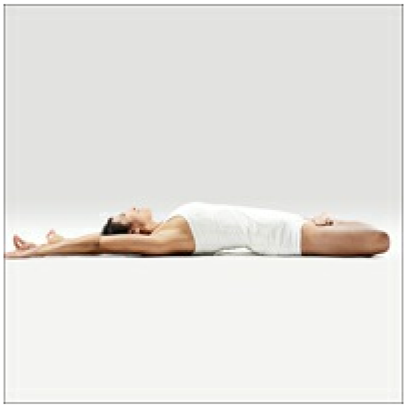

Side Corpse Pose

Parshva Shavasana

(PAHRSH-vuh shuh-VAHS-uh-nuh)

Also Known As: Belly Twist (Jataraparivritti)
Modification: 1. head turned to the opposite side of feet
2. looking straight ahead
Pose Type: supine, side bend
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows), Hastagrai or Hastagre

(hands)

1.

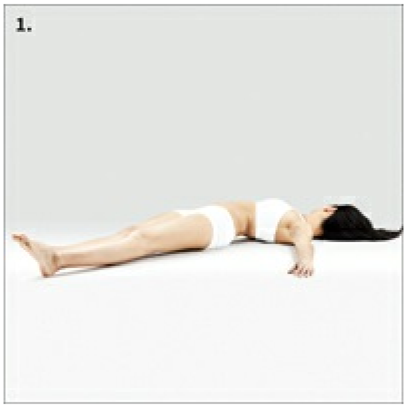

2.

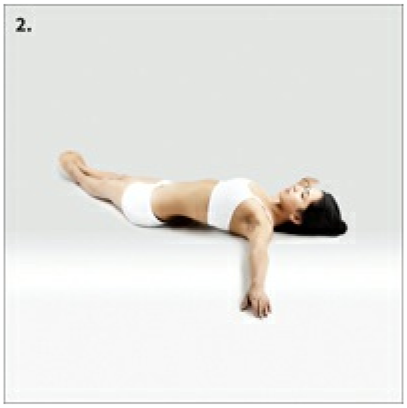

Corpse Pose

Shavasana

(shuh-VAHS-uh-nuh)

Also Known As: Dead Pose (Mrtasana)
Pose Type: supine
Drishti Point: Bhrumadhye or Ajna Chakra (third eye, between the eyebrows)

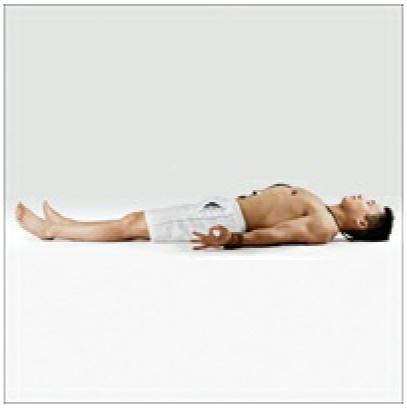

ACKNOWLEDGMENTS

Meet the talented people of Mr. Yoga, Inc. who have contributed their energy, passion, and skills to this project.

Mr. Yoga

Victoria B.

Tatiana U.

Aggie M.

Susan M.

Olga Q. B.

Laura Lisa

Vanessa D. S.

Snow W.

Paul M.

Kat D.

Paula O.

Venessa S.

Additional special thanks to: Adrienne Yau, Silvia Bogers, Antoine Bogers, Evelyn Lacerda, Chris Morino, Scott Mendel, Black Dog & Leventhal Publishers, Hachette Book Group

Welcome to the Mr. Yoga, Inc. global family. All of the models selected to appear in this book are students of Daniel Lacerda, aka Mr. Yoga.

Frances L.

Cristine C.

Carolyn L.

Tatiana U.

Elena B.

Stephanie M. & Iva M.

Matthew B.

Shriya M. D.

Mariah A.

Erin B.

Valeryia G.

Michael C.

Brent J.

Kristen S.

Christina D. F.

Laura M.

Victoria B.

Eva M.

Jemma B.

Menaka I.

Hagar E.

Kellys E.

Diana M.

Alessandra F.

Neelam P.

Katerina D.

Crio C.

Paula O.

Dasha C.

Elissa R.

Eli M.

Josiah B.

Nicki B.

Rachel M.

Hue N.

Deanna D.

Aggie M.

Corrado R. M.

Taylor K.

Susan M.

Neil F.

Justin L.

Thank you to special guest, Danny Paradise.

GLOSSARY

A

adho mukha = having the face downward

agni = fire

ahimsa = nonviolence; the word has not merely the negative and restrictive meaning of “nonkilling or nonviolence,” but the positive and comprehensive meaning of “love embracing all creation”; one of the *yamas*

ajna = to command

ajna chakra = energy or command *chakra*, energy center/the nerve plexus located between the eyebrows, the third eye, the seat of command, the sixth *chakra*

akarna = near the ear

akuncha = contraction or bend

alamba = a prop or support

anahata = unstruck

anahata chakra = spiritual heart *chakra*, energy center situated near the heart, the nervous plexus situated in the cardiac region, the fourth *chakra*

ananda = joy, happiness, bliss, ecstasy

ananta = infinite, without end; a name of Vishnu and also of Vishnu’s couch, the serpent Sesa

anga = limb, points, step, the body; a limb or a part of the body; a constituent part

angushta (angula) = finger or digit, the thumb

anjali = hands held together as in prayer

anjali-mudra = the gesture of *anjali*

Anjaneya = son of Anjani (Hanuman’s mother’s maiden name is Anjani)

antara = within, interior

apanasana = pelvic floor yoga pose

aparigraha = non-greediness, freedom from greed, desire, hoarding or collecting; one of the *yamas*

Arani = Hindu Goddess of Fire

aranya = wild animal

asana = a physical posture, the third limb or stage of yoga, originally this meant “meditation posture” or “seat”

ashta (asta, astau) = eight

Ashtavakra = one having eight bends (crooked in eight places), in reference to a Hindu Sage who was born with eight physical deformities in his body, and went on to become a spiritual preceptor of King Janaka of Mitila

ashva (asva) = horse

ashva sanchala = horse, riding posture

asteya = nonstealing, freedom from avarice; one of the *yamas*

atman = individual soul, the true self, consciousness; the term Vedanta uses instead of *purusha*

avabhinna = broken

avatara = divine manifestation, descent, advent or incarnation of God. There are ten avataras o Vishnu: Matsya (the Fish), Kurma (the Tortoise), Varaha (the Boar), Narasimha (the Man-Lion) Vamana (the Dwarf), Parasurama, Rama (hero of the epic *Ramayana*), Krishna (hero of the epic *Mahabharata*, who related the *Bhagavad Gita*), Balarama, and Kalki

B

baddha = bound
baka = crane, heron, a wading bird
bala = young, childish, not fully grown
bandha = a bond, tying, energetic lock, contraction, bondage, or fetter; a posture in which certain organs or body parts are contracted and controlled
Benu = the mythological bird of ancient Egypt symbolic of rebirth and creation, also associated wit the Sun
bhaga = strength
Bhagavad Gita = one of India’s most beloved and sacred texts, the divine song of the Lord, the most influential of all *shastras*; the epic story of Arjuna, a warrior prince who confronts moral dilemmas through sacred dialogues with Krishna (one of Lord Vishnu’s Avatars) and is led to a better understanding of reality by learning the teachings of Samkhya, Yoga, and Vedanta
Bhagavata Purana = also called Shrimad Bhagavatam, a *purana* that deals with devotion to the Supreme Being in the form of Lord Vishnu and describes some of the avatars of Vishnu, includin Krishna
bhairava = terrible, gruesome, formidable; one of the fierce manifestations of Shiva
bhakti = devotion, worship or love; from *bhaj*, (to divide), the belief that there is an eternal divide between the Supreme Being and the world that cannot be overcome through knowledge, hence the Supreme Being must be met with an attitude of devotion
bhangi = position
Bharadvaja (Bharadwaja) = a Vedic *rishi*, great warrior described in the *Mahabharata*, Pindola Bharadvaja was one of four *arhats* asked by Buddha to stay on earth to propagate Buddhist law, or *dharma*
bharman = load, nourishment, care, burden, maintenance load
bheka = a frog
bherunda = terrible, frightful; it also means a species of a bird or a name of a yogi
bhuja = arm or shoulder
bhujanga (bhujagga) = serpent, snake
bhuja-pida = pressure on the arm or shoulder
bidala = cat
bija-mantra = a mystical syllable with a sacred prayer repeated mentally during *pranayama*, and the seed thus planted in the mind germinates into one-pointedness
bindu (bindhu) = seed, point, dot, the creative potency of anything where all energies are focused, the third eye
bitila = cow
Brahma = a five-headed first deity of the Hindu Trinity; the Supreme Being, the creator; responsibl

for the creation of the world, he is the first being to appear at the dawn of each universe to create it based on its subconscious conditioning: the Brahma of the present universe is called Prajapati (progenitor), the predecessor of humankind
brahmachari = a religious student vowed to celibacy and abstinence; one who is constantly moving (*charin*) in *brahman* (The Supreme Spirit); one who sees divinity in all
brahmacharya = chastity or teacher of the soul, abstinence, a life of celibacy, religious study and self-restraint, recognition of Brahma in everything; one of the *yamas*
Brahman = the absolute, or divinity itself, infinite consciousness, universal soul, deep reality, the reality that cannot be reduced to a deeper layer; the Supreme Being, the cause of the Universe, the all-pervading spirit of the Universe
Buddha = enlightened one

C

chakora = a type of bird like a partridge (Greek partridge), moonbeam bird, said to feed on moonbeams
chakra = literally, a wheel or circle, the wheel of a wagon; metaphorically, psycho-energetic subtle centers of the subtle body in which energy flows, located along the spine, believed to transform cosmic energy into spiritual energy when activated. Energy (*prana*) is said to flow in the human body through three main channels (*nadis*), namely, *sushumna*, *pingala,* and *ida*; *sushumna* is situated inside the spinal column. *Pingala* and *ida* start respectively from the right and the left nostrils, move up to the crown of the head, and course downward to the base of the spine. These two *nadis* intersect with each other and also the *sushumna*. These junctions of the *nadis* are known as *chakras* of the flywheels, which regulate the body mechanism. The important *chakras* are: (a) *muladhara* (*mula* = root, source; *adhara* = support, vital part) situated in the pelvis, above the anus; (b) *svadhishtana* (*sva* = vital force, soul; *adhishtana* = seat of abode) situated above the organs of gestation; (c) *manipuraka* (*manipura* = navel) situated in the navel; (d) *manas* (mind) and (e) *surya* (the sun) which are situated between the navel and the heart; (f) *anahata* (= unbeaten) situated in the cardiac area; (g) *vishuddha* (= pure) situated in the pharyngeal region; (h) *ajna* (= command) situated between the eyebrows; (i) *sahasrara* (= a thousand) which is called a thousand-petaled lotus in the cerebral cavity; and (j) *lalata* (= forehead) which is at the top of the forehead.
chalana = to churn
chamatkara = delight or savoring; the refined pleasure that a connoisseur takes in a lovely poem, painting, or fine wine. Chamatkarasana is translated as “Wild Thing Pose” probably because the posture looks like someone who has just seen something so amazing and beautiful that they are “bowled over” or “blown away.”
chandra (candra) = moon
chapa = bow, rainbow, arc
chatur (chatuari, chatura) = four
chatush = four times
chatushpada = quadruped
chikitsa = therapy

chitta = consciousness that comprises mind, intellect, the restraint of consciousness; a mind in its total or collective sense, being composed of three categories: (a) mind, having the faculty of attention, selection, and rejection; (b) reason, the decisive state which determines the distinction between things; and (c) ego, the I-maker, the aggregate of intellect (*buddhi*), egoity (*ahamkara*) and thinking agent (*manas*)

chittavritti (chitta-vritti, chitta vritti) = an imbalance of the mental state, fluctuations of the mind, movement of the consciousness; a course of behavior, mode of being, condition, or mental state

cibi = chin

D

dakshina = the right side

danda = stick, staff (refers to the spine)

dasha = ten

dhanu (dhanura) = bow

dharana = concentration, sixth limb of Ashtanga yoga, orienting the mind toward a single point

dhyana = generally translated as meditation, freedom from attachments, an ongoing stream of awareness from the meditator toward the object of meditation, and of information from the object toward the meditator; the seventh limb of Ashtanga yoga

dhyana-yoga = yoga of meditation

Diti = the mother of the demons, also called Daityas

drishti = focal point, perception or looking place, “view” or “sight”: yogic gazing, such as at the tip of the nose or the spot between the eyebrows

dur = difficult

Durvasa = a very irascible, notoriously angry sage

dvija = twice-born

dwi (dve) = two

dwihasta = two hands

dwipada = two feet or legs

E

eka (ekam) = one

eka-pada (ekapada) = one leg, one-legged, one-footed

G

gaja = elephant

Galava = the pupil or son of Viswamitra

ganda = the cheek or side of the face including the temple

gandha = subtle earth element; quantum of (*tanmatra*) earth; smell

gandha-bherunda = a species of bird, also a two-headed mythological bird that embodies immense powers and destructive forces

garbha = an infant, womb, fetus

garbha kosha = uterus

garbha pinda (garbha-pinda) = fetus, embryo in the womb

Garuda = Hindu deity, half-man half-eagle, fierce bird of prey, vehicle (*vahana*) of Lord Vishnu, king of birds; Garuda is represented as a vehicle of Vishnu and as having a white face, an aquiline beak, red wings, and a golden body

gava = cow

Gheranda = author of *Gheranda Samhita*, an important text on Hatha Yoga which he taught to Chanda Kapali

Gheranda Samhita = a Tantric treatise describing Hatha Yoga written by the sage Gheranda in the 15th century

Gitananda = a well-known *yogi*, living in the 20th century CE

go = cow

godha = iguana

gomukha = cow face, face resembling a cow, cow head; it is also a kind of musical instrument, narrow at one end and broad at the other, like the face of a cow

Goraksha (Goraksha, Gorakshanath) = an 11th to 12th Century Hindu Nath*yogi*, one of Matsyendranath's two most important disciples; tending to or breeding cattle, cowherd

graiva = a chain worn around the neck of an elephant; necklace or collar

guru = "he who is heavy, weighty," a spiritual teacher or preceptor, one who illuminates the darkness of spiritual doubt, one who hands down a system of knowledge to a disciple; heavy one or dark/light, dispeller of darkness, one who helps to gain knowledge

guru-shishya parampara = the tradition of teaching dating back centuries, where a guru imparts his knowledge to his students

H

ha = first syllable of the word *hatha*, which is composed of the syllables *ha* (= sun) and *tha* (= moon); the object of Hatha Yoga is to balance the flow of solar and lunar energy in the human system

hala = plow

hamsa (hansa) = a swan; a metaphor for the soul; a vehicle of Lord Brahma; the name of the mantra by which *prakriti* permeates the universe; also refers to the breath as it moves within the body

Hanuman = a powerful monkey chief, a mythological entity, of extraordinary strength and prowess, whose exploits are celebrated in the epic *Ramayana*; he was the son of Anjana and Vayu, the god of wind, monkey-god, hero of *Ramayana*, egoless superhero and perfect devotee, who resembles a monkey leaping

hasta = hand

hastasana = forward stretch of the arms

hatha = force; the word *hatha* is used adverbially in the sense of "forcibly" or "against one's will"; Hatha Yoga is so called because it prescribes rigorous discipline in order to find union with the Supreme

hatha-vidya = the science of Hatha Yoga

Hatha Yoga = "Forceful Yoga," a major branch of yoga, developed by Goraksha and other adepts c.

1000–1100 CE, emphasizing the physical aspects of the transformative path, notably postures (*asana*), cleansing techniques (*shodhana*) and breath (*pranayama*); literally, sun/moon yoga (*ha* = sun, *tha* = moon), it emphasizes balancing the solar and lunar energy channels in the body. Hatha Yoga shifted the focus away from the mysticism and philosophy of the older Upanishadic types of yoga toward using the body as a tool; combines opposing forces to achieve balance, sighting the soul through the restraint of energy, yoga concerned with mastering control over the physical body as a path to enlightenment (self-realization)

Hatha Yoga Pradipika = a celebrated treatise on yoga compiled in the 12th century by the sage Svatmarama

himsa = violence, killing

Hindola = cradle or swing; also Hindu Religious Festival associated with baby Krishna being rocked in a decorated swing

I

ida = a *nadi*, a channel through which *prana* moves, starting from the left nostril, then moving to the crown of the head and then descending to the base of the spine on the left side; in its course it conveys lunar energy and so is called *chandra nadi,* “channel of the lunar energy”; associated with pale or blue (left/feminine)

Indra = ruler, lord of thunder, king of the heavens

indudala = crescent moon

Ishvara = God, the Supreme Being, Brahman, with form

ishvara pranidhana = one of the *niyamas*; centering on the divine, devotion or surrender to God, dedication to the Lord of one’s actions and one’s will

J

jalandhara bandha (jalandharabandha) = a *bandha* that locks the throat, chin lock; straightening the back of the neck by keeping your head straight while slightly receding your chin; a yoga pose where the neck and throat are contracted and the chin is rested in the notch between the collarbones at the top of the breast-bone

janu = knee

jatara = stomach, belly, or the interior of anything

jatara-parivartana = an action of an *asana* (yoga pose) in which the abdomen is made to move to and fro

Jnana Yoga = the emphasis is on questioning, contemplation, and meditation as a path to enlightenment, yoga that seeks to teach the identity of the individual self (*atman*) and the infinite consciousness (*brahman*)

K

kaka = crow

Kala Bhairava = Shiva in his terrible or gruesome form as Destroyer of the Universe

Kali = Hindu Goddess of Time and Change

Kamala = Hindu Goddess of Wealth, “One of the Lotus”

Kamalamuni = one of the 18 Siddhars believed to be over 4,000 years old

kanda = a bulbous root, a knot, egg, stem, stalk, trunk; the *kanda* is a round shape of about four inches situated twelve inches above the anus and near the navel, where the three main *nadis* (*sushumna*, *ida*, and *pingala*) unite and separate; it is covered as if with a soft white piece of cloth

kapala = skull

kapalabhati = bellow-like breathing technique with sharp, quick inhalations and exhalations; a cleansing ritual for the respiratory tract, lungs, and sinuses; skull shining

kapila = a sage or *rishi*, the founder of the Samkhya system, one of the six orthodox systems of Hindu philosophy, noted in the *Bhagavad Gita* and *Bhagavata Purana* as a manifestation of the Supreme Being

kapinjala = a kind of partridge, the *chataka* bird, which is supposed to drink only raindrops

kapota = dove, pigeon

kapya = monkey

karanda = duck

karani = making, doing

karma (karma law) = action, activity of any kind, including ritual acts; said to be binding only so long as engaged in a self-centered way; the law of cause and effect, or the movement toward balanced consciousness: everything that you do, say, or think has an immediate effect on the universe that will reverberate back to you in some way

karma yoga (karma-yoga) = yoga of action, path to enlightenment is through selfless acts and service to others, the achievement of union with the Supreme Universal Soul through action; in its original Vedic sense, Karma Yoga is any yoga that employs ritualistic action, such as *asana*, meditation, or *mantra*, to produce spiritual gain. The term excludes Jnana Yoga and Bhakti Yoga, which are thought to operate beyond spiritual gain

karna = the ear; also one of the heroes in the *Mahabharata*

karna-pida = pressure around the ear, blocked ears

Kashyapa = an ancient Hindu sage, husband of Aditi and Diti; he is one of the lords or progenitors of living things

khaga = bird

khanjana = a wagtail bird

Khimi Karani = mythological pond of milk in which Garuda drowned a snake to give birth to the Shami Tree

kona = angle

Konganar = one of the Siddhars, student of Siddhar Bogar

Koormamuni = Hindu Sage

Korakar = one of the 18 Siddhars, a well-known sage, author of works on philosophy, medicine, and alchemy

Koundinya (Kaundinya) = Hindu sage, Vedic scholar, and a descendant of Vasishta

kriya = act, action, cleansing

kriya yoga = the yoga of action and participation in life, preliminary yoga consisting of simplicity (*tapas*), the reading of sacred texts (*svadhyaya*), and acceptance of the existence of a Supreme Being (*ishvarapranidhana*); also, a Tantric mode of yoga using breath, *mantra*, and visualization.

krakacha = a saw
krounch (krouncha, krauncha) = heron
kukkuta = rooster, cock
kulpa = ankle
kundalini = a coiled female serpent; the divine cosmic energy, the obstacle that closes the mouth of *sushumna*; the rising of *shakti* in the *sushumna*; this force or energy is symbolized as a coiled and sleeping serpent lying dormant in the lowest nerve center at the base of the spinal column, the *muladhara-chakra*. This latent energy has to be aroused and made to ascend the main spinal channel, the *sushumna* piercing all the *chakras* right up to the *sahasrata*, the thousand-petaled lotus in the head. Then the *yogi* is in union with the Supreme Universal Soul.
Kundalini Yoga = a mode of yoga that focuses on the raising of the life force
kunja = grove, alcove
kunta = spear, lance
kurma (koorma) = a tortoise, it is also the name of one of the subsidiary vital airs whose function is to control the movements of the eyelids to prevent foreign matter or light that is too bright going into the eyes

L

laghu = little, small, simple; it also means “handsome”
lasya = beauty, happiness, and grace; also a dance performed by Goddess Parvati in response to her husband Shiva’s *tandava*
Linga (Lingam, Shivalinga) = symbol of union and origin of all life associated with Lord Shiva and Goddess Shakti
lola = tremulous, dangling, pendant; charm, swing; swinging like a pendulum, moving to and fro

M

madhya (madya) = middle of the body, central
maha = great, mighty, powerful, lofty, noble
maha bandha = the great lock
maha mudra = the great seal
Mahabharata = the celebrated epic, the largest piece of literature created by humankind, authored by Rishi Vyasa and containing the *Bhagavad Gita*, dating to the first century BCE; *dharma shastra* (scripture dealing with right action), which comes to the conclusion that however hard you try, you can never be completely right
makara = a mythological sea creature, who is the vehicle of the river Goddess Ganga; a crocodile
makshika = fly
mala = a garland or a wreath, often of prayer beads or flowers
manas-chakra = nervous plexus situated between the navel and the heart
mandala = a circle ambulation, a circular drawing or design that exemplifies sacred geometry that draws your eye to the center and is used as a focal point while meditating; it also means a “collection,” a division of Rig Veda
manduka = frog

manipuraka = a nervous plexus situated in the region of the navel; the third *chakra*, the navel *chakra*, the fire energy center, site of the sense of fear and apprehension

mantra = a mystical syllable designed to create and alter reality by influencing the vibrational patterns that make up creation; a sacred sound or phrase that has a transformative effect

Marichi = a sage, son of Brahma, the great-grandfather of Manu, the Vedic Adam and the father of humanity, the creator of the universe, and the father of Kasyapa

marjarai = cat

matsya = fish

Matsyendra = a Hindu sage and one of the first teachers of Hatha Yoga; a legend, king or lord of the fish

mayura = peacock

moksha (moksha) = liberation from bondage, final emancipation of the soul from recurring births

mrita (mrta, mritra) = dead, corpse

mudra = a seal; a pleasant hand gesture or seal posture; directs the life current (life energy) through the human body, usually a combination of *asana*, *pranayama*, and *bandha*

mukha = face

mukta = free, unbound, liberated

mukti = release, liberation, final absolution of the soul from the chain of birth and death

mula = root, foundation, bottom; a yoga pose where the body from the anus to the navel is contracted and lifted toward the spine

mula bandha (mula-bandha) = rectal lock, root lock; contraction of the pubococcygeus, a yoga pose where the body from the anus to the navel is contracted and lifted up and toward the spine

muladhara = root foundation, the name of the first *chakra*

muladhara chakra = the first *chakra*, the base *chakra*, the earth energy center situated at the root of the spine, nervous plexus situated in the pelvis above the anus at the base or root of the spine, the main support of the body that controls sexual energy

mushti = fist

mutra kosa = bladder

N

nabi = navel

nadi = river; nerve or conduit, channels that distribute energy from the *chakras* throughout the body, a tubular organ of the subtle body through which energy flows, subtle vibratory passages of psychospiritual energy; it consists of three layers, one inside the other, like insulation of an electric wire. The innermost layer is called the *sira* and the middle layer, *damani*. The entire organ as well as the outer layer is called *nadi*; they connect at special points of intensity (*chakras*).

naga = great mythological snake; one of the subsidiary vital airs that relieves abdominal pressure, causing one to belch

Nahusha = Hindu King of Aila Dynasty

nakra = crocodile

namaskar = greeting, worship, salutation with hands in prayer

namaste mudra = a *mudra* in which the hands are placed together in a prayerlike fashion to honor the

inner light
nantum = to bow with respect
nara = a man
Narasimha = an avatar of the Hindu god Vishnu in his fourth incarnation, often visualized as half-man, half-lion
nasika = nose
nata = actor, dancer, mime
Nataraj = name of Shiva as a cosmic dancer, the lord of the dancers
natya = dancing
nauka = boat
nava = boat
nava = nine
nidra = deep, dreamless sleep; the third state listed in the *Mandukya Upanishad*. The others are waking state (*jagrat*), dream (*susupt*), and consciousness (*turiya*); also the fourth fluctuation of the mind listed by Patanjali in *Yoga Sutra* I.6 (the others are correct cognition, wrong cognition, perceptualization, and memory).
nindra (nantra) = sage, praise, wonder
nindra (nitara) = standing firm, standing
nir = without
nira = water
niralamba = self-supported, independent, without support
niyama = self-restraint, personal observances, self-purification by discipline, the Vedic system of logic; the second stage or limb of Ashtanga Yoga mentioned by Patanjali; five personal disciplines, as defined by Patanjali in his *Yoga Sutras*: *shaucha, santosha, tapas, svadhyaya,* and ishvarapranidhana

O

om (aum) = the original *mantra* symbolizing the ultimate reality, the sacred syllable emitted by the Supreme Being, the sound that produces all other sounds and into which all other sounds return. Like the Latin word *omne*, the Sanskrit work *aum* means “all” and conveys concepts of Omnipotence, Omnipresence, and Omniscience.

P

pada = foot or leg; also a part of a book or text
pada-hasta = hand(s)-to-feet
padangushta = big toe
padma = lotus
pakshaka = wing
pakshi = bird
pancha = five
parampara = tradition, uninterrupted series, convention, a succession
parigha = iron bar used for locking, bolting, or shutting a gate

parigraha = hoarding, possessiveness

paripurna = full, entire, complete

parivartana = turning around, revolving

parivartana-pada = with one leg turned around

parivid = twined, twisted around

parivritta = revolved, turned around

parivritti = crossed or with a twist, turning, rolling

parshva = the side, flank, lateral

parshvaika = *parsva* (side) + *eka* (one)

parshvaika-pada = with one leg turned sideways

parvata = mountain

paryanka = a bed, a couch

pasa (pasha) = snare, trap, noose, fetter

paschima = the back of the whole body from head to heels, west side, western

paschimatana = intense stretch of the back side of the body from the nape to the heels

pakshya = being in or belonging to the wings

patan = to collapse

Patanjali = a sage, the author of the *Yoga Sutras*, a treatises on Sanskrit and *Ayurveda*; the founder of yoga, most likely lived between 200 BCE and 300 CE; a manifestation of the serpent of infinity

pavanamuktasana = wind release pose

perineum = the area between the thighs, behind the genitals, and in front of the anus

pichamayura = peacock with stretched feathers

pid = squeeze

pida = pain, discomfort, pressure

pincha = a feather of a tail, the chin

pinda = a fetus, an embryo in an early stage of gestation, ball, the body

pingala = a channel on the right side of the spine through which *prana* moves, associated with reddish color, a *nadi* or channel of masculine energy starting from the right nostril, then moving to the crown of the head and then downward to the base of the spine; as the solar energy flows through it, it is also called *surya-nadi*

pitam = stool, chair

plavana = jump through

pliha = the spleen

poorna = full

prana = breath, life, vitality, wind, energy, strength; connotes the soul, life force or inner breath; sometimes refers to anatomical or outer breath; vital upward energy current

pranayama = breath control, energy control through breathing, consisting of conscious inhalation (*puraka*), retention (*kumbhaka*), and exhalation (*rechaka*), breath extension, breathing exercises to harmonize the flow of life force; the fourth stage or limb of Ashtanga Yoga

prapada = tip of the feet

prasarita = spread, stretched out

pratyahara = internalization of the senses, independence from sensory stimuli; the fifth stage or limb of Ashtanga Yoga, withdrawal and emancipation of the mind from the domination of the senses and

sensual objects; withdrawal of the mind, mental detachment from the external world
prishta = back
Punakeesar (Punnakeesar) = one of the Siddhars, a *guru* of Machamuni (also referred to as Matsyendra)
pungu = Telugu word for “wounded”
purna = complete
purva = eastern
purvottana = intense stretch of the front side of the body

R

raja = king, royal
raja yoga (raja-yoga) = royal yoga; a term generally applied to the three higher limbs of Ashtanga Yoga, that is *dharana, dhyana,* and *samadhi*; the royal road to self-realization through the control of the mind. The achievement of union with the Supreme Universal Spirit by becoming the ruler o one’s own mind by defeating its enemies, sighting the soul through a restraint of consciousness
raja-kapota (rajakapota) = king pigeon
Ramayana = literally, Rama’s way; a famous ancient epic (*itihasa*) authored by Sage Valmiki that describes the life of Rama, an *avatar* of Lord Vishnu
Rig Veda = literally “Knowledge of Praise,” it consists of 1,028 hymns and is the oldest known reference to *yoga* and possibly the oldest known text in the world
rishi = a Vedic seer, a liberated sage or saint, one who through suspension of the mind can see to the bottom of his heart
Ruchika (Richika, Ruschika) = name of a Hindu sage, dedicated to the grandfather of an incarnation of Vishnu

S

sa = with
sahaja = easy, natural
sahasrara chakra = energy center situated at the crown of the head, the thousand-petaled lotus in the cerebral cavity, the most important seventh *chakra* which, when uncoiled, brings the seeker to freedom
St. Brighid = known for establishing numerous monasteries
sakti = power
salamba = with support
sama = same, equal, even, upright
samadhi = “putting together”: the ecstatic or state in which the mediator becomes one with the object of meditation, forgetting him/herself completely (the Supreme Spirit pervading the universe) where there is a feeling of unutterable joy and peace, absorption, ecstasy, enlightenment, self-realization; the state of meditation in which ego disappears and all becomes one; a state of absolute bliss; the eighth limb or stage of Ashtanga Yoga
samadhi yoga = yoga of absorption
samastiti = a state of balance

Sankarar = yogi from the eighth century BC
Sanskrit = the programming language used to write the operating system of the subtle body; the language of the gods
santosha (santosa) = contentment; one of the *niyamas*
sapta = seven
sarpa = serpent, snake
sarva = all, whole
sarvanga = all parts, the whole body
sasanga = rabbit
satya = truth; one of the yamas
Shesha = a celebrated serpent, said to have a thousand heads; Sesa is represented as the couch of Vishnu, floating on the cosmic ocean, or as supporting the world on his hoods; other names of Shesha are Ananta and Vasuki
setu = a bridge, dam, dike
setu-bandha = the construction of a bridge; name of an *asana* in which the body is arched
shalabha = grasshopper, locust
Shankara, Adi = world teacher, yoga master, propounder of Jnana Yoga and Advaita Vedanta; author of commentaries on the *Brahma Sutra*, the *Upanishads*, the *Bhagavad Gita* and thirty other texts; founder of ten monk orders and four large monasteries whose abbots today still carry the title Shankaracharya. His dates are disputed. Western academics often place him at 800 CE. Tradition places him at 1800 BCE. Also known as Shankaracharya or Shankara Bhagavatpada
shanti = peace
shat = six
shaucha = purity or inner and outer cleanliness; one of the *niyamas*
shava (shava) = corpse
shayana = bed, couch, sleeping
shirsha = head
Shiva (Siva) = the most Powerful God in Hinduism, the Destroyer, a name of the Supreme Being pure consciousness, Brahman with form
shvana (swana) = dog; inspiration
shvnaka = puppy dog
siddha = accomplished, fulfilled, perfected; a sage, seer, or prophet; also a semi-divine being of great purity and holiness, a perfected being, a yoga master who has become an immortal, ethereal being
siddhi = divine attribute, perfection, supernatural power, proof
simha = lion
Skanda = a name of Kartikeya, the god of war, general of the celestial army, Lord of War, second son of Lord Shiva and Godmother Uma Parvat
stamba = transition
steya = theft, robbery
sucirandra = threading the needle
sukha = ease, lightness, comfort, happiness, delight, joy, pleasure; literally, agreeable mental space
Sundaranandar = one of the eighteen Siddhars, author of numerous works on medicine

supta = lying down or sleeping, reclining, supine

surya = sun

surya yantra = sun dial

surya-chakra = nervous plexus situated between the navel and the heart

surya-nadi = the *nadi* of the Sun; another name for *pingala-nadi*

sushumna = the main *nadi* channel situated inside the spinal column, a hollow passageway between *pingala nadi* and *ida nadi* that runs through the spinal cord, and through which *kundalini* can travel once it is awakened

sutra = thread: a work consisting of aphoristic statements such as Patanjali's *Yoga Sutra*

sva = one's own, innate, vital force; soul, self

svadhyaya = education of self by study of divine literature, self-study, to study one's body, mind, intellect, and ego; one of the *niyamas*

svarga = heaven

svastika = good fortune

Svatmarama = the author of *Hatha-yoga-pradipika*, a classical textbook on Hatha Yoga

swadhishtana chakra = site of worldly desires, energy center located above the organ of generation

Swami Sivananda = a well-known *yogi* of the 20th century, founder of Sivananda Yoga

Swami Vishnu Devananda = a close disciple of Swami Sivananda

T

taal-vrksa = palm tree

tada = mountain, straight tree

tadasana-samasthiti = a state of balance; an even distribution of weight while standing

tan = stretch, lengthen out or extend

tana = to stretch out, extend

Tandava (Thandava) = sacred frantic dance representing the cosmic cycles of creation and destruction performed by the Hindu Deity Lord Shiva

tantra = thread on the loom

Tantra (Tantric) Yoga = this yoga is characterized by certain rituals designed to awaken the *kundalini*

tap = to burn, to blaze, to shine, to suffer pain, to be consumed by heat

tapa = austerity

tapas = a burning effort, glow, heat; austerity gained through the committed practice of yoga, self-discipline; practice with discipline, devotion and religiosity; one of the *niyamas*

tara = star

Taraka = a demon slain by Kartikeya, the god of war

tha = the second syllable of the word *hatha*; the first syllable *ha* stands for the sun, while the second syllable *tha* stands for the moon; the union of these two is represented in Hatha Yoga

thavali = frog in Tamil

tirieng = horizontal, oblique, transverse, reverse, upside down

tiryak= horizontally, sideways, obliquely, across

tiryang-mukha = backward facing

tittibha = a small bird living along the coastline; firefly; or insect

tola = balance, scale
tolana = weighing
tri (tri, tra) = three
tryanga = three limbs
trikona = three angle or triangle
Trivikrama = Vishnu in his fifth incarnation, The Dwarf *avatar* of Lord Vishnu, who with his three steps (*krama*) filled the earth, heaven and hell, the conqueror of three worlds
tulya = equilibrium, balance

U

ubhaya = both
uddayate = fly, soar, fly up
uddiyana = a fetter or bondage, a yogic abdominal lock; here the diaphragm is lifted high up the thorax and the abdominal organs by tilting your pelvic floor up and pulling your belly button back toward your spine; the *uddiyana-bandha,* the great bird *prana* (life), is forced to fly up through the *shushumna-nadi*; to fly up
ujjayi = a *pranayama* that produces sound in the throat with the inhalation, literally meaning "extended victory"; the lungs are fully expanded and the chest is puffed out, slow throat breathing
Upanishad = the word is derived from the prefixes *upa* (near) and *ni* (down) added to the root *shad* (to sit); it means sitting down near a *guru* to receive spiritual instruction. The *Upanishad* scriptures of ancient Hindu philosophy are the philosophical portion of the Vedas, the most ancient sacred literature of the Hindus, dealing with the nature of man and the universe and the union of the individual or self with the Universal Soul
upavishta (upavistha) = seated, sitting, with legs spread
urdhva (urdhwa) = upward, raised, elevated, inverted
urdhva-mukha = face upward
ushtra (ustra) = camel
ut = intense, a particle denoting intensity
utkata = fierce, powerful, exceeding the usual measure, excessive, squat
utpluti = lifting or pumping up
utripada = upright tripod
uttana = an intense stretch, upright
utthita = extended, risen or rising, raised up, stretched

V

Vaasamuni Siddhar = a disciple of Shiva
vadivu = old form; Gaja Vadivu is an animal posture from Kalari Yoga, a mystical tantric form of *yoga*, which stems from Kalarippayat
vajra = thunderbolt, Indra's weapon
Vajracchedika Prajnaparamita Sutra = Diamond Cutter Sutra, one of the *sutras* of Mahayana Buddhism focusing on non-attachment
vakra = bent, curved, crooked

Valakhilya = flying wise and virtuous companions, celestial beings; a class of divine personages of the size of a thumb, produced from the Creator's body, and said to precede the chariot of the sun
Valmiki = known as the father of Sanskrit classical poetry
vama = left side
Vamadeva = the name of the preserving aspect of the God Shiva
vamana = Vishnu in his fifth incarnation, when he was born as a dwarf to humble the demon king Bali
Vasishta = a celebrated sage, author of Yoga Vasistha; several Vedic hymns, most excellent, best, richest
vatayana = horse
vayu = air, vital force, wind, vital air
Vedas = the sacred scriptures of the Hindus, revealed by the Supreme Being
vibhuti = might, power, greatness
vimshati = twenty
vini = single movement
vinyasa = a steady flow of connected yoga *asanas* linked with breath, work in a continuous movement, going progressively, variation
viparita = reversed, inverted, turned
vira = a brave or eminent man, heroic, chief, hero
virabhadra = a legendary warrior, a powerful hero created out of Shiva's matted hair
Virancha (Viranchi) = one of the names of Lord Brahma
vishama = uneven, unequal
Vishnu (Vishnu, Narayana, Hari) = the second deity of the Hindu Trinity, All-Pervading essence of all beings, one who supports, Preserver God
vishuddhi chakra = seat of intellectual awareness, energy center situated behind the throat, the nervous plexus in the pharyngeal region
vishva = entire, whole
Vishvamitra = a celebrated Hindu Sage, ruler so impressed with Vasistha's knowledge and contentment that he became his disciple
vrischika = scorpion
vriksha = tree
vrishta = fallen or dropped as rain
vyaghra = tiger

Y

yajna = Hindu Sacrificial Ceremony
yama = ethical codes for daily life, self-restraint
Yama = the god of death; Yama is also the first of the eight limbs or means of attaining yoga. Yamas are universal ethical codes for daily life, self-restraint, and moral commandments or ethical disciplines transcending creeds, countries, age, and time. The five mentioned by Patanjali are: non-violence, truth, non-stealing, continence, and non-coveting
yantra = to sustain
yoga = union, communion, the path which integrates the body, senses, mind, and the intelligence with

the self, derived from “yuj”, meaning to join or to yoke, to concentrate one’s attention on. It is the union of our will to the will of God, a poise of the soul which enables one to look evenly at life in all its aspects. The chief aim of yoga is to teach the means by which the human soul may be completely united with the Supreme Spirit pervading the universe and thus secure absolution

yoga-mudra = a posture, a seal

yoga-nidra = the sleep of yoga, where the body is at rest as if in sleep while the mind remains fully conscious, though all its movements are stilled; *yoga-nidra* is also the name of an *asana*

Yogananda = a great yogi of the twentieth century

yogasana = yogic posture

Yoga Sutra (yoga-sutra) = a classical collection of aphorisms on the practice of yoga, attributed to the sage Patanjli. It consists of 185 terse aphorisms on yoga and it is divided into four parts dealing respectively with *samadhi*, the means by which yoga is attained, the powers the seeker comes across in his quest, and the state of absolution

yogi or yogini = one who follows the path of *yoga*, a student, a seeker of truth

yogic = an adjective describing things that are associated with *yoga*

yoni = the womb

yoni-mudra = womb or female seal or awakened *kundalini*, the sealing; the breeding place, and *mudra* is a seal; *yoni-mudra* is a sealing posture where the apertures of the head are closed and the aspirant’s senses are directed within to enable him to find out the source of his being

yudha = from Yudhisthira, a legendary warrior mentioned in the ancient Hindu epic *Mahabharata*

yuj = to join, to yoke, to use, to concentrate one’s attention on

yukti = union

[*] “Bitilisana” may also be translated as “Gavasana” in the following cow poses.

* “Ushtrasana” may also be spelled as “Ustrasana” in the following camel poses.

[*] “Shirshasana” may also be spelled as “Sirsasana” in the following headstand poses.

Thank you for buying this ebook, published by Hachette Digital.

To receive special offers, bonus content, and news about our latest ebooks and apps, sign up for our newsletters.

Sign Up

Or visit us at hachettebookgroup.com/newsletters

CONTENTS

Cover

Title Page

Welcome

Opening Prayer

A Message from Mr. Yoga

A History of Yoga

The Eight Limbs of Yoga

The *Ujjayi* Breath

The *Bandhas*

The *Drishtis*

How to Approach Yoga Poses

Relaxation and Meditation

A Brief Summary of *Chakras*

Benefits and Cautions for 8 Conditions

STANDING POSES

SEATED POSES

CORE POSES

QUADRUPED POSES

BACKBENDS

ARM BALANCES

INVERSION POSES

PRONE POSES

SUPINE POSES

Acknowledgments

Glossary

Newsletters

Copyright

Copyright

The information herein is not intended to replace the services of trained health professionals or be a substitute for medical advice. You are advised to consult with your health care professional with regard to matters relating to your health, and in particular regarding matters that may require diagnosis or medical attention.

Neither this exercise program nor any other exercise program should be followed without first consulting a health care professional. If you have any special conditions requiring attention, you should consult with your health care professional regularly regarding possible modification of the program contained in this book.

Copyright © 2015 Mr. Yoga, Inc.
Photographs copyright © 2015 Daniel Lacerda, Mr. Yoga, Inc.

All rights reserved. In accordance with the U.S. Copyright Act of 1976, the scanning, uploading, and electronic sharing of any part of this book without the permission of the publisher constitute unlawful piracy and theft of the author's intellectual property. If you would like to use material from the book (other than for review purposes), prior written permission must be obtained by contacting the publisher at permissions@hbgusa.com. Thank you for your support of the author's rights.

Black Dog & Leventhal Publishers
Hachette Book Group
1290 Avenue of the Americas, New York, NY 10104
hachettebookgroup.com
blackdogandleventhal.com

First ebook edition: December 2015

Management of layout and design by Tatiana Usova
Cover design by Christopher Lin
Cover photograph by Daniel Lacerda
Cover © 2015 Hachette Book Group, Inc.
Interior design by Sheila Hart Design, Inc.
Additional proofreading by: Aggie Metford, Victoria Bergin, Susan Marchese

Black Dog & Leventhal Publishers is an imprint of Hachette Books, a division of Hachette Book Group.
The Black Dog & Leventhal Publishers name and logo are trademarks of Hachette Book Group, Inc.

The Hachette Speakers Bureau provides a wide range of authors for speaking events. To find out more, go to www.HachetteSpeakersBureau.com or call (866) 376-6591.

The publisher is not responsible for websites (or their content) that are not owned by the publisher.

ISBN 978-0-316-27062-5

E3

2,100 ASANAS

THE COMPLETE YOGA POSES

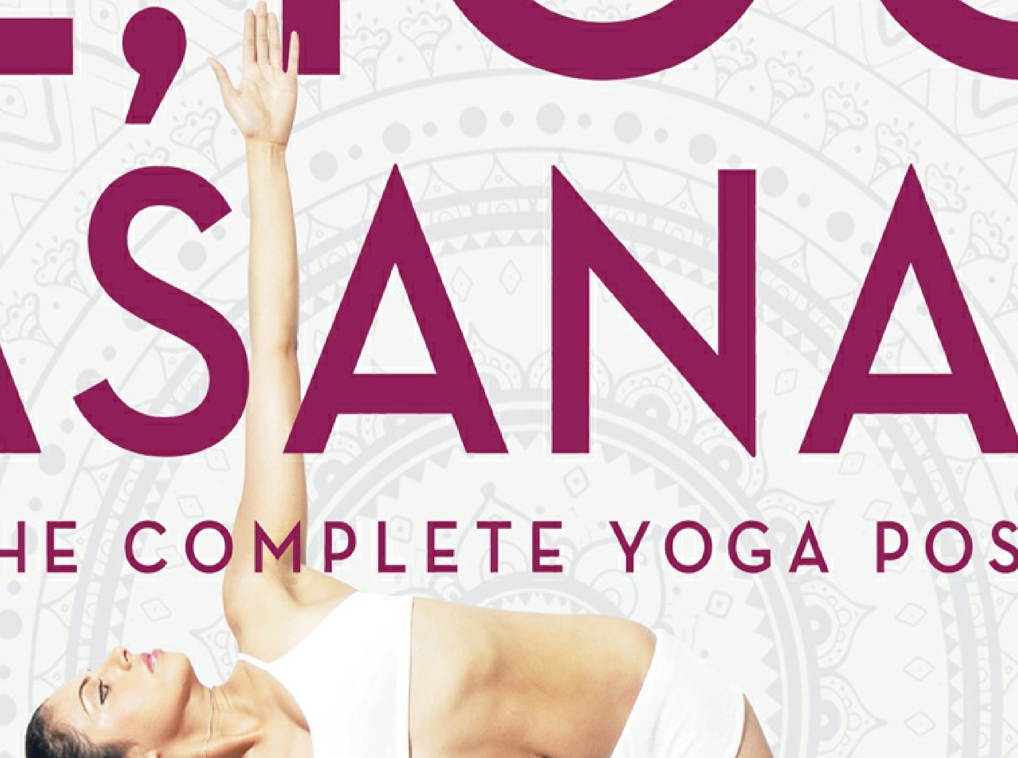

DANIEL LACERDA

FOUNDER OF MR. YOGA, INC.